## Cooperative Minds: Social Interaction and Group Dynamics

## Proceedings of the $35^{\text {th }}$ Annual Meeting of the Cognitive Science Society

Berlin, Germany, July 31-August 3, 2013

Edited by

Markus Knauff
Giessen University

Natalie Sebanz
Central European University

Michael Pauen<br>Humboldt-Universität zu Berlin

Ipke Wachsmuth
Bielefeld University


Austin, TX: Cognitive Science Society

How to cite a paper in these Proceedings:
APA formatted citation for a 6-page paper:
Author, A. \& Author, B. (2013). This is the title of the paper. In M. Knauff, M. Pauen, N. Sebanz, \& I. Wachsmuth (Eds.), Proceedings of the 35th Annual Conference of the Cognitive Science Society (pp. NUMBERS). Austin, TX: Cognitive Science Society.

APA formatted citation for a published abstract:
Author, A. \& Author, B. (2013). This is the title of the abstract [Abstract]. In M. Knauff, M. Pauen, N. Sebanz, \& I. Wachsmuth (Eds.), Proceedings of the 35th Annual Conference of the Cognitive Science Society (p. NUMBER). Austin, TX: Cognitive Science Society.

APA formatted citation for a talk (or poster) presentation:
Author, A. \& Author, B. (2013, August). This is the title of the talk or poster. Paper (or Poster) presented at the 35 th Annual Conference of the Cognitive Science Society. Berlin, Germany.

## Table of Contents

Introduction ..... 1
Acknowledgements ..... 3
Organizing Committee ..... 4
Program Committee ..... 5
Reviewers List ..... 6
Sponsors ..... 14
Conference Awards ..... 15
Glushko Dissertation Prizes ..... 18
Invited Plenary Presentations ..... 19
Workshops
Diagrammatic Cognition: Discovery and Design ..... 20
William Bechtel
Using Mechanical Turk and PsiTurk for Dynamic Web Experiments ..... 22Anna Coenen, Doug Markant, Jay B. Martin, John McDonnell
Embodied Approaches to Interpersonal Coordination: Infants, Adults, Robots, and Agents ..... 24Rick Dale, Chen Yu, Yukie Nagai, Moreno Coco, Stefan Kopp
Interactive Panel Discussion: Professional advancement through real and virtual international collaboration ..... 26Laurie Beth Feldman, Janet van Hell, Judith Kroll, Suparna Rajaram, Natasha Tokowicz
Workshop Proposal: PRE-CogSci 2013: Bridging the gap between cognitive and computational approaches to reference ..... 28Albert Gatt, Roger van Gompel, Ellen Gurman Bard, Emiel Krahmer, Kees van Deemter
Motivations and Goals in Developing Integrative Models of Human Cognition ..... 30
Glenn Gunzelmann
Deception as a Social Strategy ..... 32Swati Gupta, Tei Laine, Kayo Sakamoto
Mental Model Ascription by Language-Enabled Intelligent Agents ..... 34
Marjorie McShane
When eye see you: Gaze and joint attention in human interaction ..... 35
Maria Staudte, Ulrich Pfeiffer
Tutorials
Using Complex Network Analysis in the Cognitive Sciences ..... 37
Nicole Beckage, Michael Vitevitch, Alexander Mehler, Eliana Colunga
A General Purpose Architecture for Building Spiking Neuron Models of Biological Cognition ..... 39
Chris Eliasmith, Terrence Stewart
Virtual Humans: A New Toolkit for Cognitive Science Research ..... 41
Jonathan Gratch, Arno Hartholt, Morteza Dehghani, Stacy Marsella
Making Robust Classification Decisions: Constructing and Evaluating Fast and Frugal Trees (FFTs) ..... 43
Hansjörg Neth, Uwe Czienskowski, Lael Schooler, Kevin Gluck
Half Day Tutorial on Using Quantum Probability Theory to Model Cognition ..... 45 Emmanuel Pothos, Zheng Wang, Jerome Busemeyer
Types and states: Mixture and hidden Markov modelling for the cognitive sciences ..... 47 Maarten Speekenbrink, Ingmar Visser
Dynamic Field Theory: Conceptual Foundations and Applications in the Cognitive and Developmental Sciences ..... 49
John Spencer, Gregor Schoener, Yulia Sandamirskaya
How to Analyze Verbal Protocols ..... 51
Thora Tenbrink
Computational complexity analysis for cognitive scientists ..... 53
Iris van Rooij, Johan Kwisthout, Mark Blokpoel, Todd Wareham
Invited Symposia
Progress in Joint Action Research ..... 55
Susan E. Brennan, Daniel Richardson, Michael J. Richardson, Andreas Roepstorff, Natalie Sebanz, Günther Knoblich
Language and Gesture Evolution ..... 57
Joseph Call, Susan Goldin-Meadow, Cat Hobaiter, Katja Liebal, Michael Tomasello
New Frameworks of Rationality ..... 59
Nick Chater, Klaus Fiedler, Gerd Gigerenzer, Karl Christoph Klauer, Mike Oaksford, Keith Stenning
Mechanistic Developmental Process: Rumelhart Prize Symposium in Honor of Linda Smith ..... 61
Larissa K. Samuelson, Anthony Morse, Chen Yu, Eliana Colunga, Thomas Hills
Glushko Dissertation Prize Symposium ..... 63
Douglas Knox Bemis, Neil Cohn, George Kachergis, Andrew Lovett, Robert L. Goldstone
Symposia
Communicative Intentions in the Mind/Brain ..... 65Bruno Bara, Nick Chater, Michael Tomasello, Rosemary Varley
Exploring Cognitive Diversity Across Disciplines and Cultures ..... 67
Andrea Bender, Sieghard Beller, Daniel Haun, Cristine Legare, Asifa Majid, Bethany Ojalehto, Douglas Medin
What if? Counterfactual reasoning, pretense, and the role of possible worlds ..... 69
Daphna Buchsbaum, Caren Walker, Alison Gopnik, Nick Chater, David Danks, Christopher Lucas, Charles Kemp, Eva Rafetseder, Josef Perner
Time and Causality: Mutual Constraints; Insights from Event and Time Perception, Motor Control, and Gaming ..... 71
Marc Buehner, David Lagnado, Christos Bechlivanidis, Marc Ernst, Marieke Rohde, Michael Young
Sequential Sampling Models Representing a Unifying Framework of Human Decision Making ..... 73
Jerome Busemeyer, Adele Diederich, Andrew Heathcote, Antonio Rangel, Jörg Rieskamp, Marius Usher
Multiword Sequences as Building Blocks for Language: Insights into First and Second Language Learning ..... 75
Morten H. Christiansen, Inbal Arnon, Elena Lieven, Alison Wray
Distributed remembering in a social context: Effects of communicating a shared past on what is remembered ..... 77
Alin Coman, Gerald Echterhoff, Micah Edelson, William Hirst, Kourken Michaelian, Charles Stone
Beyond Synchrony: Complementarity and Asynchrony in Joint Action ..... 79
Rick Dale, Riccardo Fusaroli, Dorthe Døjbak Håkonsson, Patrick Healey, Dan Mønster, John McGraw, Panagiotis Mitkidis, Kristian Tylén
Computational and Cognitive Aspects of Narrative ..... 81
Mark Finlayson, Bernhard Fisseni, Dedre Gentner, Richard Gerrig, Benedikt Loewe, Jeffrey Loewenstein, Inderjeet Mani, Jan Christoph Meister, R. Michael Young
Structured cognitive representations and complex inference in neural systems ..... 83
Samuel Gershman, Joshua Tenenbaum, Alexandre Pouget, Matthew Botvinick, Peter Dayan
Cognitive Foundations of Cultural Learning ..... 85
Annette Henderson, Ben Kenward, Rachel Watson-Jones, Cristine Legare, Harvey Whitehouse
Constraints on Bayesian Explanation ..... 87
Johan Kwisthout, Iris van Rooij, Matteo Colombo, Carlos Zednik, William A. Phillips
Global Change and Cognition ..... 89
Stephan Lewandowsky, Tania Lombrozo, Gordon Brown, John Cook, Ben Newell, Ullrich Ecker
Is perception cognitively penetrable? A philosophically satisfying and empirically testable reframing ..... 91Gary Lupyan, Dustin Stokes, Fiona Macpherson, Rasha Abdel Rahman, Robert GoldstoneInformavores: Active information foraging and human cognition93Doug Markant, Todd Gureckis, Bjorn Meder, Jonathan Nelson, Peter Pirolli, Chen Yu
Social Cognition, Culture and the Self. An interdisciplinary investigation ..... 95Albert Newen, Gary Bente, Shinobu Kitayama, Shihui Han, Kai Vogeley
Scales of Cognition Evident in Action ..... 97Denis O’Hora, Nicholas D. Duran, Rick Dale, Jonathan B. Freeman, John Tomlinson
Music cognition: Bridging computation and insights from cognitive neuroscience ..... 99
Marcus Pearce, Martin Rohrmeier, Petri Toiviainen, Elvira Brattico, Psyche Loui, Edward Large, Ji Chul Kim
Thinking about norms: Epistemic, rational, and moral norms in human thinking ..... 101
Joëlle Proust, Emmanuel M. Pothos, Jerome R. Busemeyer, Ryan Miller \& Fiery Cushman, Katinka J.P. Quintelier, Shira Elqayam, Valerie A. Thompson, Jonathan St. B.T. Evans, David E. Over, Meredith R. Wilkinson
Implicit Learning Out of the Lab: Language and Music ..... 103
Patrick Rebuschat, Martin Rohrmeier, Morten Christiansen, Zoltan Dienes, Daniele Schoen, Jennifer Misyak, Clement Francois, Xiuyan Guo, Feifei Li, Richard Widdess
The Co-Existence of Naïve and Scientific Concepts in Learning and Development ..... 105
Michael Schneider, Stella Vosniadou, Andrew Shtulman, Kelsey Harrington, Ilonca Hardey, Lindsey J. Powell, Susan Carey, Stellan Ohlsson
Minimal Nativism: How does cognitive development get off the ground? ..... 107Tomer Ullman, Josh Tenenbaum, Noah Goodman, Shimon Ullman, Elizabeth Spelke
Language as a Window into the Brain and its Pathologies ..... 109
Peter Garrard, Peter beim Graben, Brita Elvevåg, Eduardo Mizraji, Juan Valle-Lisboa
The empirical investigation of semantics: Between abstract-symbolic and embodied-simulative models of meaning ..... 111
Markus Werning, Erica Cosentino
Online Education: A Unique Opportunity for Cognitive Scientists to Integrate Research and Practice ..... 113Joseph Jay Williams, Alexander Renkl, Ken Koedinger, John Stamper
Papers
Causal Reasoning with Continuous Outcomes ..... 115
Ahmad Azad Ab Rashid, Marc Buehner
A Metacognitive Stopping Rule for Problem Solving ..... 121
Rakefet Ackerman
An Information-Theoretic Account of Musical Expectation and Memory ..... 127Kat Agres, Samer Abdallah, Marcus Pearce
Effector specific response activation during word processing ..... 133Daniela Ahlberg, Carolin Dudschig, Barbara Kaup
Simulating Overall and Trial-by-Trial Effects in Response Selection with a Biologically-plausibleConnectionist Network139
Blair Armstrong, David Plaut
The dynamics of pragmatic enrichment during metaphor processing: activation vs. suppression ..... 145
John M. Tomlinson, Jr., Stavros Assimakopoulos
Object motion continuity and the flash-lag effect ..... 151Ricky K. C. Au, Katsumi Watanabe
Interface Design and Spatial Cognition: A Case of Virtual Molecule Manipulation ..... 157
Trevor Barrett, Mary Hegarty
Using the words toddlers know now to predict the words they will learn next ..... 163
Nicole Beckage, Eliana Colunga
Simultaneous unsupervised and supervised learning of cognitive functions in biologically plausible spiking neural networks ..... 169
Trevor Bekolay, Carter Kolbeck, Chris Eliasmith
Balancing Fairness and Efficiency in Repeated Societal Interaction ..... 175
Noam Ben-Asher, Christian Lebiere, Alessandro Oltramari, Cleotilde Gonzalez
Beyond Rule versus Rote? Processing of Distinctive Dative and Genitive Case Markers in German ..... 181
Christian Bentz
Social and Environmental Contributors to Infant Word Learning ..... 187
Elika Bergelson, Daniel Swingley
Grid or no grid: Distance distortion in recognizing spatial information from complex cartographic maps ..... 193
Anne-Kathrin Bestgen, Dennis Edler, Frank Dickmann, Lars Kuchinke
Evidence for cognitively controlled saccade targeting in reading ..... 197
Klinton Bicknell, Emily Higgins, Roger Levy, Keith Rayner
Fusiform Face Area in Chess Expertise ..... 203
Merim Bilalic
Analyzing Students' Metacognitive Strategies in Open-Ended Learning Environments ..... 209
Gautam Biswas, John Kinnebrew, James Segedy
"Are we still talking about the same thing?" MEG reveals perspective-taking in response to pragmatic violations, but not in anticipation ..... 215
Sara Bögels, Dale Barr, Simon Garrod, Klaus Kessler
Discovering Processing Stages by combining EEG with Hidden Markov Models ..... 221
Jelmer Borst, John Anderson
Pragmatic priming and the search for alternatives ..... 227
Lewis Bott, Emmanuel Chemla
Space is Special: A domain-specific mapping between time and nontemporal magnitude ..... 233
Roberto Bottini, Daniel Casasanto
Open-Ended Category-Based Induction: The Influence of Associative Strength and Structured Knowledge Representations ..... 239
Aimee Kay Bright, Aidan Feeney
Motor Asymmetries Predict Neural Organization of Emotion ..... 245
Geoffrey Brookshire, Cleve Graver, Daniel Casasanto
The Effect of Discourse Context on Online Sentence Processing ..... 251Juliane Burmester, Isabell Wartenburger, Katharina Spalek
Probabilistic reasoning in the two-envelope problem ..... 257
Bruce Burns
Preschoolers' ability to navigate communicative interactions in guiding their inductive inferences ..... 263
Lucas Butler, Ellen Markman
A Computational Model of General Rule Learning with Unnatural Classes ..... 269
Shira Calamaro
From individual minds to social ones. Cooperation and the structure of animal and human societies. ..... 275
Valentina Cardella, Alessandra Falzone, Antonino Pennisi
Assessing the effectiveness of older adults' spatial descriptions in a fetch task ..... 281
Laura Carlson, Marjorie Skubic, Jared Miller, Zhiyu Huo, Tatiana Alexenko
Hypothesis space checking in intuitive reasoning ..... 287
Christopher Carroll, Charles Kemp
Individuals recapitulate the proposed evolutionary development of spatial lexicons ..... 293
Alexandra Carstensen, Terry Regier
The development of predictive processes in children's discourse understanding ..... 299
Marisa Casillas, Michael Frank
English/Italian Bilinguals Switch Gesture Parameters when they Switch Languages ..... 305
Federica Cavicchio, Sotaro Kita
Modelling Graded Semantic Effects in Lexical Decision ..... 310
Ya-Ning Chang, Matthew Lambon Ralph, Steve Furber, Stephen Welbourne
Eyetracking as an Implicit Measure of Category-Based Induction ..... 316
Stephanie Chen, Brian Ross, Gregory Murphy
General Instruction Following in a Large-Scale Biologically Plausible Brain Model ..... 322
Xuan Choo, Chris Eliasmith
Understanding eye movements in face recognition with hidden Markov model ..... 328
Tim Chuk, Alvin C. W. Ng, Emanuele Coviello, Antoni B. Chan, Janet H. Hsiao
Uniquely human developmental timing may drive cerebral lateralization and interhemispheric coupling ..... 334
Benjamin Cipollini, Garrison Cottrell
Systems from Sequences: an Iterated Learning Account of the Emergence of Systematic Structure in a Non-Linguistic Task ..... 340
Hannah Cornish, Kenny Smith, Simon Kirby
On the Dynamics of Information Accumulation in Recognition ..... 346
Gregory Cox, Nicholas Lewis, Richard Shiffrin
Representing Number Sets: Encoding Statistical Properties ..... 352
Patrick Cravalho, Bradley Morris, Christopher Was, Amy Masnick
Unified Modeling of Proactive Interference and Memorization Effort: A new mathematical perspective with in ACT-R theory ..... 358
Arindam Das, Wolfgang Stuerzlinger
Integrating Cognitive Principles to Redesign a Middle School Math Curriculum ..... 364
Jodi Davenport, Yvonne Kao, Steven Schneider
Achievement Motivation and Strategy Selection during Exploratory Learning ..... 370Daniel DeCaro, Marci DeCaro, Bethany Rittle-Johnson
Cost-Based Pragmatic Inference about Referential Expressions ..... 376
Judith Degen, Michael Franke, Gerhard Jäger
Superficial, rather than true, knowledge interdependence in collaborative learning fosters individual knowledge transfer ..... 382
Anne Deiglmayr, Lennart Schalk
Analogical Reasoning with Rational Numbers: Semantic Alignment Based on Discrete Versus Continuous Quantities ..... 388
Melissa DeWolf, Miriam Bassok, Keith Holyoak
Learning to perceive coherent objects ..... 394
Nimrod Dorfman, Daniel Harari, Shimon Ullman
Removal of Information from Working Memory ..... 400Ullrich Ecker, Stephan Lewandowsky, Klaus Oberauer
Effects of Explanation and Comparison on Category Learning ..... 406
Brian Edwards, Joseph Jay Williams, Tania Lombrozo
Biologically Plausible, Human-scale Knowledge Representation ..... 412
Crawford Eric, Gingerich Matthew, Eliasmith Chris
Cross-Situational Statistical Learning of Phonologically Overlapping Words ..... 418
Paola Escudero, Karen Mulak, Haley Vlach
Intuition versus Reason: Strategies that People Use to Think about Moral Problems ..... 424
Mark Fedyk, Barbara Koslowski
Communicative biases shape structures of newly acquired languages ..... 430
Maryia Fedzechkina, T. Florian Jaeger, Elissa L. Newport
Regularization behavior in a non-linguistic domain ..... 436
Vanessa Ferdinand, Bill Thompson, Simon Kirby, Kenny Smith
Rapid Learning of Morphological Paradigms ..... 442
Sara Finley, Elizabeth Wiemers
Analogical Reinforcement Learning ..... 448James Foster, Matt JonesDevelopmental and postural changes in children's visual access to faces454
Michael Frank, Kaia Simmons, Daniel Yurovsky, Guido Pusiol
The Effect of Incremental Context on Conceptual Processing: Evidence from Visual World and Reading Experiments ..... 460
Diego Frassinelli, Frank Keller, Christoph Scheepers
The intrinsic structure of spatial configurations and the partner's viewpoint shape spatial memories and descriptions ..... 466
Alexia Galati, Marios N. Avraamides
False Consensus About False Consensus ..... 472Mirta Galesic, Henrik Olsson, Jörg Rieskamp
The Role of Feedback in Aligning Perspectives in Referential Communication ..... 477
Timothy Gann
Production of referring expressions: Preference trumps discrimination ..... 483
Albert Gatt, Emiel Krahmer, Roger van Gompel, Kees van Deemter
Bayesian vector analysis and the perception of hierarchical motion ..... 489
Samuel Gershman, Frank Jaekel, Joshua Tenenbaum
Space and Time in the Parietal Cortex: fMRI Evidence for a Neural Asymmetry ..... 495
Tom Gijssels, Roberto Bottini, Shirley-Ann Rueschemeyer, Daniel Casasanto
Modeling Dynamic Control in Normal Aging ..... 501
Brian D Glass, Magda Osman
Development of Category-Based Reasoning: Results from a Longitudinal Study ..... 507
Karrie Godwin, Anna Fisher, Bryan Matlen
'This is a wall' - Assigning Function to Objects ..... 513
Linn Gralla, Thora Tenbrink
Is Feedforward Learning more Efficient than Feedback Learning in Smart Meters of Electricity Consumption? ..... 519
Mona Guath, Peter Juslin, Philip Millroth
Fairness overrides reputation: The importance of fairness considerations in altruistic cooperation ..... 525
Sule Guney, Ben R. Newell
A Probabilistic Constraint Satisfaction Model of Information Distortion in Diagnostic Reasoning ..... 531
York Hagmayer, Olga Kostopoulou
Spatial cognition: the return path ..... 537Kai Hamburger, Lena E. Dienelt, Marianne Strickrodt, Florian Röser
Attentional Processing in Bistable Perception is Influenced by Genetic Effects Associated with Sinistrality ..... 543
Roeland Hancock
Intrusions and the Decision to Terminate Memory Search ..... 549
J. Isaiah Harbison, Eddy Davelaar, Erica Yu, Erika Hussey, Michael Dougherty
Autism, optimism and positive events: Evidence against a general optimistic bias ..... 555Adam Harris, Punit Shah, Caroline Catmur, Geoffrey Bird, Ulrike Hahn
The neural computation of scalar implicature ..... 561
Joshua Hartshorne, Jesse Snedeker, Albert Kim
Causal model and sampling approaches to reducing base rate neglect ..... 567Brett Hayes, Ben Newell, Guy Hawkiins
The Role of Familiarity, Priming and Perception in Similarity Judgments ..... 573
Laura Hiatt, Greg Trafton
Large-Scale Empricial Analyses of the Abstract/Concrete Distinction ..... 579
Felix Hill, Anna Korhonen, Christian Bentz
How Episodic and Working Memory Affect Rule- and Memory-Based Judgments ..... 585
Janina A. Hoffmann, Bettina von Helversen, Jörg Rieskamp
Preschoolers' Trust is Sensitive to Variable Intentions ..... 591
Elena Hoicka, Jessica Butcher, Felicity Malla, Paul Harris
When Is Language a Window into the Mind? Looking Beyond Words to Infer Conceptual Categories ..... 597
Kevin Holmes, Phillip Wolff
Preschoolers' Understanding of Preferences is Modulated by Linguistic Framing ..... 603
Taylor Holubar, Ellen Markman
Eye movement optimization in visual search ..... 609
Ryan M. Hope, Wayne D. Gray
More than a feeling: When emotional reactions don't predict moral judgments ..... 615
Zachary Horne, Derek Powell
Young children's developing sensitivity to discourse continuity as a cue to reference ..... 621
Alexandra Horowitz, Michael Frank
Anxiety Can Influence Analogy-Making Both Positively and Negatively Depending on the Complexity of the Mapping Task ..... 627
Penka Hristova, Yoana Petkova, Boicho Kokinov
A Neural Model of Human Image Categorization ..... 633
Eric Hunsberger, Peter Blouw, James Bergstra, Chris Eliasmith
How to improve on Quinian bootstrapping - a response to nativist objections ..... 639
Zoltan Jakab
Gestures for Thinking ..... 645
Azadeh Jamalian, Valeria Giardino, Barbara Tversky
Relational labels can improve relational retrieval ..... 651
Anja Jamrozik, Dedre Gentner
Verb bias and structural priming in non-linguistic grammar acquisition task ..... 657
Marius Janciauskas, Franklin Chang
Not so innocent: Reasoning about costs, competence, and culpability in very early childhood ..... 663
Julian Jara-Ettinger, Joshua Tenenbaum, Laura Schulz
Developmental See-Saws: Ordered visual input in the first two years of life ..... 669
Swapnaa Jayaraman, Caitlin Fausey, Linda Smith
Chinese-speaking adults' understanding of argument structure ..... 675
Lu Jiang, Etsuko Haryu
Efficient codes for multi-modal pose regression ..... 680
Leif Johnson, Joseph Cooper, Dana Ballard
A soft barrier model for predicting human visuomotor behavior in a driving task ..... 686
Leif Johnson, Brian Sullivan, Mary Hayhoe, Dana Ballard
Meaning Overrides Frequency in Idiomatic and Compositional Multiword Chunks ..... 692
Hajnal Jolsvai, Stewart M. McCauley, Morten H. Christiansen
The effects of dual verbal and visual tasks on featural vs. relational category learning ..... 698
Wookyoung Jung, John Hummel
The Effects of Repeated Sequential Context on Recognition Memory ..... 704
George Kachergis, Gregory Cox, Richard Shiffrin
More Naturalistic Cross-situational Word Learning ..... 710
George Kachergis, Chen Yu
The role of thinking-aloud instructions and prior domain knowledge in information processing and source evaluation during Web search ..... 716
Yvonne Kammerer, Peter Gerjets
Effects of numeric magnitude on the cortical valuation network ..... 722
Frank Kanayet, John Opfer, William Cunningham
The Funny Thing About Incongruity: A Computational Model of Humor in Puns ..... 728
Justine T. Kao, Roger Levy, Noah D. Goodman
What They Don't know Might Help Them: A Demonstration of Subtle Social Inference from Mimicry ..... 734
Liam Kavanagh, Giti Bakhtiari, Chris Suhler, Patricia Churchland, Rob Holland, Piotr Winkielman
Paradoxical cultural categories and paradoxical social relationships: The case of cooperation and competition ..... 740Josh Keller, Jeffrey Loewenstein, Martin Kilduff, Jin Yan
Minimally Supervised Learning for Unconstrained Conceptual Property Extraction ..... 746
Colin Kelly, Anna Korhonen, Barry Devereux
Musicality and non-native speech sound processing are linked through temporal, pitch and spectral acuity ..... 752
Vera Kempe, Dennis Bublitz, Patricia J. Brooks
Maximizing Students' Retention Via Spaced Review: Practical Guidance From Computational Models Of Memory ..... 758
Mohammad Khajah, Robert Lindsey, Michael Mozer
Spatial terms across languages support near-optimal communication: Evidence from Peruvian Amazonia, and computational analyses ..... 764Naveen Khetarpal, Grace Neveu, Asifa Majid, Lev Michael, Terry Regier
Cooperative Strategies with Incongruent Facial Expressions Cause Cardiovascular Threat ..... 770Peter Khooshabeh, Celso de Melo, Brooks Volkman, Jonathan Gratch, Jim Blascovich,Peter Carnevale
A model of population dynamics applied to phonetic change ..... 776
James Kirby, Morgan Sonderegger
Do older Gaelic-English bilinguals show an advantage in inhibitory control? ..... 782
Neil Kirk, Kenneth Scott-Brown, Vera Kempe
Bayesian Speech Production: Evidence from Latency and Hyperarticulation ..... 788
Christo Kirov, Colin Wilson
Transitive and periphrastic sentences affect memory for simple causal scenes ..... 794
Melissa Kline, Paul Muentener, Laura Schulz
Semantic Distance Modulates the N400 Event-Related Potential in Verbal Analogical Reasoning ..... 799
Matthew J. Kmiecik, Robert G. Morrison
Multiple Proposal Memory in Observational Word Learning ..... 805
Judith Koehne, John C. Trueswell, Lila R. Gleitman
Awareness in Decision-Making: Implicit Influences or Measurement Biases? ..... 811
Emmanouil Konstantinidis, David Shanks
The impact of bottom-up and top-down saliency cues on reference production ..... 817
Ruud Koolen, Emiel Krahmer, Marc Swerts
A spreading-activation model of the semantic coordination of speech and gesture ..... 823
Stefan Kopp, Kirsten Bergmann, Sebastian Kahl
Space and Time are Mutually Contagious in Sound ..... 829
Alexander Kranjec, Matthew Lehet, Anjan Chatterjee
Competition, event comprehension, and dynamic location information in sentence processing ..... 834
Anuenue Kukona, Gerry T. M. Altmann, Yuki Kamide
Evolution and modularity: the limits of mechanistic explanation ..... 840
Jaakko Kuorikoski, Samuli Pöyhönen
Incremental processing in the pragmatic interpretation of contrastive prosody ..... 846
Chigusa Kurumada, Meredith Brown, Sarah Bibyk, Daniel Pontillo, Michael Tanenhaus
Contextual inferences over speakers' pragmatic intentions: Preschoolers' comprehension of contrastive prosody ..... 852
Chigusa Kurumada
Communicatively efficient language production and case-marker omission in Japanese ..... 858
Chigusa Kurumada, T. Florian Jaeger
Snapshots of Working Memory: Using Early Eye Movements to Capture Temporal Dynamics ..... 864
Nicholas Lange, Rick Thomas, Daniel Buttaccio, Eddy Davelaar
Learning Representations of Animated Motion Sequences - A Neural Model ..... 870
Georg Layher, Martin Giese, Heiko Neumann
Modeling disambiguation in word learning via multiple probabilistic constraints ..... 876
Molly Lewis, Michael Frank
An integrated model of concept learning and word-concept mapping ..... 882
Molly Lewis, Michael Frank
Proper Sequencing and Level-Bridging Scaffolding in Learning a Chemical System with Graphical Simulations ..... 888Na Li, John Black
General and Efficient Cognitive Model Discovery Using a Simulated Student ..... 894
Nan Li, Eliane Stampfer, William Cohen, Kenneth Koedinger
Learned helplessness and generalization ..... 900
Falk Lieder, Noah D Goodman, Quentin JM Huys
Is it Time Bayes went Fishing? Bayesian Probabilistic Reasoning in a Category Learning Task ..... 906
Marcus Lindskog, Anders Winman, Peter Juslin
Extreme Expertise: Exploring Expert Behavior in Tetris ..... 912
John Lindstedt, Wayne Gray
Working memory capacity and fluid abilities: The more difficult the item, the more more is better ..... 918
Daniel R. Little, Stephan Lewandowsky, Stewart Craig
When Tuesday comes before Threesday: Cross-linguistic differences in numerical transparency of time words predicts temporal reasoning strategy and performance ..... 924
Nian Liu, Benjamin Bergen
Modeling Spatial Ability in Mental Rotation and Paper-Folding ..... 930
Andrew Lovett, Kenneth Forbus
Mnemonic Diffusion: An Agent-Based Modeling Investigation of Collective Memory ..... 936
Christian Luhmann, Suparna Rajaram
The effect of social roles on gaze cue utilisation in a real-world collaboration ..... 942
Ross Macdonald, Benjamin Tatler
Social pragmatic factors in reasoning from disjunctions of numerical estimations ..... 948
Robert Mackiewicz, Paweł Koniak
The Role of Cognitive Niches in Mediating Knowledge, Entropy \& Violence ..... 954
Lorenzo Magnani, Tommaso Bertolotti
A Computational Theory of Subjective Probability [Featuring a Proof that the Conjunction Effect is not a Fallacy] ..... 960
Phil Maguire, Philippe Moser, Rebecca Maguire, Mark Keane
Dynamic Structure of Joint-Action Stimulus-Response Activity ..... 966
MaryLauren Malone, Ramon Castillo, John Holden, Heidi Kloos, Michael RichardsonPlacing Numbers in Behavioral Space: Activity-Specific Interactions between Number and Spacewith a Single Response Button972Tyler Marghetis, Jasmeen Kanwal, Benjamin Bergen
Subgoal Labeled Worked Examples Improve K-12 Teacher Performance in Computer Programming Training ..... 978Lauren Margulieux, Richard Catrambone, Mark Guzdial
The Ambivalence of Expert Categorizers ..... 984
Jessecae Marsh, Naomi Rothman
Is it Really that Simple? The Complexity of Object Descriptions in Human-Computer Interaction ..... 990
Vivien Mast, Evelyn Bergmann
Agency Intuitions in Physical Interactions ..... 996
Ralf Mayrhofer, Michael Waldmann
Reasoning with Inconsistent Causal Beliefs ..... 1002John V McDonnell, Pedro Tsividis, Bob Rehder
Naïve Physics - the wrong theory? ..... 1008Ian P. L. McLaren, Katie Wood, Rossy McLaren
Sequential Diagnostic Reasoning with Verbal Information ..... 1014
Björn Meder, Ralf Mayrhofer
Comparative Evidence For Associative Learning In Task Switching ..... 1020
Christina Meier, Stephen E. G. Lea, Charlotte L. D. Forrest, Katharina Angerer, Ian P. L. McLaren
Relations between Body Motion and Emotion: Analysis based on Laban Movement Analysis ..... 1026
Junya Morita, Yukari Nagai, Tomoyuki Moritsu
The missing baselines in arguments for the optimal efficiency of languages ..... 1032
Fermin Moscoso del Prado
Social Roles and Category Use: A Study of Creativity Assessment ..... 1038Jennifer Mueller, Jeffrey Loewenstein, Shimul Melwani
Simultaneous Contradictory Belief and the Two-Systems Hypothesis ..... 1044
Joshua Mugg
Bounded Optimal State Estimation and Control in Visual Search: Explaining Distractor Ratio Effects ..... 1049
Christopher Myers, Richard Lewis, Andrew Howes
Is action-guiding vision cognitively impenetrable? ..... 1055
Bence Nanay
Finding hidden types: Inductive inference in long-tailed environments ..... 1061
Daniel Navarro
Grounded Spatial Belief Revision ..... 1067
Jelica Nejasmic, Leandra Bucher, Markus Knauff
Desirable Difficulty in Learning: A Computational Investigation ..... 1073
Aida Nematzadeh, Afsaneh Fazly, Suzanne Stevenson
The Individuation and Recognition of Emotion ..... 1079
Albert Newen, Anna Welpinghus
Schema-Driven, Space-Supported Random Accessible Memory Systems for Manipulation of Symbolic Working Memory ..... 1085
Nader Noori, Laurent Itti
Auditory Overshadowing in Preschoolers: A Preference for the Input, the System, or Both? ..... 1091Allison O'Leary, Vladimir Sloutsky
The interaction of different working memory mechanisms and sentence processing: A study of the P600 ..... 1097
Polly O'Rourke
Can Help Seeking Behavior in Intelligent Tutoring Systems Be Used as Online Measure for Goal Orientation? ..... 1103
Christine Otieno, Rolf Schwonke, Ron Salden, Alexander Renkl
Social categories create (biased) semantic interference during face naming ..... 1109Dominic J Packer, Padraig G O'Seaghdha, Almut Hupbach, Joseph E Bates
Changing minds by tracking eyes: Dynamical systems, gaze and moral decisions ..... 1115
Philip Parnamets, Petter Johansson, Christian Balkenius, Lars Hall, Michael J. Spivey, Daniel C. Richardson
Multimodal Networks of Interpersonal Interaction and Conversational Contexts ..... 1121
Alexandra Paxton, Rick Dale
Getting to the Point: The Influence of Communicative Intent on the Kinematics of Pointing Gestures ..... 1127
David Peeters, Mingyuan Chu, Judith Holler, Aslı Özyürek, Peter Hagoort
Comprehending with the body: Action compatibility in sign language? ..... 1133
Pamela Perniss, David Vinson, Neil Fox, Gabriella Vigliocco
Eye Gaze Affects Vocal Intonation Mimicry ..... 1139
Marie Postma-Nilsenová, Niek Brunninkhuis, Eric Postma
The Power of Words in the Brain: Systematic Sound-Meaning Associations in Novel and Existing Words ..... 1145
Marie Postma-Nilsenová, Constantijn Kaland, Leonoor Oversteegen
Justified True Belief Triggers False Recall of "Knowing" ..... 1151Derek Powell, Zachary Horne, Angel Pinillos, Keith Holyoak
Generic Priors Yield Competition Between Independently-Occurring Causes ..... 1157Derek Powell, Alice Merrick, Hongjing Lu, Keith Holyoak
Beyond backchannels: co-construction of dyadic stancce by reciprocal reinforcement of smilesbetween virtual agents.1163
Ken Prepin, Magalie Ochs, Catherine Pelachaud
Exploring word recognition with selected stimuli: The case for decorrelated parameters ..... 1169Athanassios Protopapas, Efthymia Kapnoula
Computational Models of Human Behavior in Wartime Negotiations ..... 1175David Pynadath, Ning Wang, Stacy Marsella
Using Property Induction to Evaluate Understanding of Mixing ..... 1181Connor Quinn, Michelle Ellefson, Anne Schlottmann, Keith TaberDoes the Semantic Integration of Emotion Words Depend on Emotional Empathy? N400, P600and Localization Effects for Intentional and Proprioceptive Emotion Words in Sentence Contexts 1187Natalia Rak, Jarmo Kontinen, Lars Kuchinke, Markus Werning
The Myth of Cognitive Decline ..... 1193
Michael Ramscar, Peter Hendrix, Harald Baayen
The Evolution of Rhythmic Cognition: New Perspectives and Technologies in Comparative Research ..... 1199
Andrea Ravignani, Bruno Gingras, Rie Asano, Ruth Sonnweber, Vicente Matellán, W.Tecumseh Fitch
Training Principle-Based Self-Explanations: Transfer to New Learning Contents ..... 1205
Alexander Renkl, Judith Solymosi, Michael Erdmann, Vincent Aleven
Predicting similarity change as a result of categorization ..... 1211
Irene Reppa, Emmanuel Pothos
Inferring Sensory Experiences ..... 1217
Kevin Reuter
Modeling the Emergence of Lexicons in Homesign Systems ..... 1223
Russell Richie, Charles Yang, Marie Coppola
A Bottom-up approach to the cultural evolution of bilingualism ..... 1229
Sean Roberts
Integration and Inference: Cross-Situational Word Learning Involves More than Simple Co-occurrences ..... 1235
Alexa Romberg, Chen Yu
A dynamical model of the speech perception-production link ..... 1241
Kevin D. Roon, Adamantios I. Gafos
The spatial representation of grammatical number ..... 1247
Timo Röttger, Frank Domahs
Cross-Linguistically Shared and Language-Specific Sound Symbolism for Motion: An Exploratory Data Mining Approach ..... 1253Noburo Saji, Kimi Akita, Mutsumi Imai, Katerina Kantartzis, Sotaro Kita
Polarity correspondence does not explain the SNARC effect ..... 1259
Julio Santiago, Daniël Lakens
Dependencies First: Eye Tracking Evidence from Sentence Production in Tagalog ..... 1265
Sebastian Sauppe, Elisabeth Norcliffe, Agnieszka E. Konopka, Robert D. Van Valin,Stephen C. Levinson
Learning hierarchical categories in deep neural networks ..... 1271Andrew M. Saxe, James L. McClelland, Surya Ganguli
Hierarchical Bayesian Modeling: Does it Improve Parameter Stability? ..... 1277
Benjamin Scheibehenne, Thorsten Pachur
Subjective Awareness during Cross-Language Speech Perception Attending Unattended Regions of an Acoustic Continuum ..... 1283
Jordan Richard Schoenherr, John Logan
Virtues and Vices in Monetary Tradeoffs ..... 1289
Marc Scholten, Daniel Read
Comparing Model Comparison Methods ..... 1294
Holger Schultheis, Ankit Singhaniya, Devendra Singh Chaplot
Rational preference shifts in multi-attribute choice: what is fair? ..... 1300
Pradeep Shenoy, Angela Yu
Automatic and Instructed Attention in Learned Predictiveness ..... 1306
Lauren Shone, Evan Livesey
Communication Leads to the Emergence of Sub-optimal Category Structures ..... 1312
Catriona Silvey, Simon Kirby, Kenny Smith
Encoding time and allocation of attention in analogical development ..... 1318
Nina Simms, Dedre Gentner
Parent-Child Screen Media Co-Viewing: Influences on Toddlers' Word Learning and Retention ..... 1324
Clare Sims, Eliana Colunga
Investigating the Other-Race Effect of Germans towards Turks and Arabs using Multinomial Processing Tree Models ..... 1330
Henrik Singmann, David Kellen, Karl Christoph Klauer
Semantics: Intuition, Experiment or Illusion? ..... 1336
Peter Slezak
Peter Slezak
Physical predictions over time ..... 1342
Kevin A Smith, Eyal Dechter, Joshua B Tenenbaum, Edward Vul
Linguistic structure is an evolutionary trade-off between simplicity and expressivity ..... 1348
Kenny Smith, Monica Tamariz, Simon Kirby
Modelling the Supervisory System and Frontal Dysfunction: An Architecturally Grounded Model of the Wisconsin Card Sorting Task ..... 1354
Mariam Sood, Richard Cooper
Foreign accent does not influence cognitive judgments ..... 1360Andre L. Souza, Arthur B. Markman
Picture-Word Interference with Masked and Visible Distractors: Different Types of Semantic Relatedness Inhibit Lexical Selection ..... 1366Katharina Spalek, Markus Damian
Implications of Polychronous Neuronal Groups for the Nature of Mental Representations ..... 1372
William Benjamin St. Clair, David C. Noelle
Testing theories of skill learning using a very large sample of online game players ..... 1378Tom Stafford, Michael Dewar
When seeing isn't believing: Influences of prior conceptions and misconceptions ..... 1384
Eliane Stampfer, Kenneth Koedinger
The Importance of Nouns in Text Processing ..... 1390Lauren M. Stuart, Julia M. Taylor, Victor RaskinWorking Memory, Cognitive Miserliness and Logic as Predictors of Performance on theCognitive Reflection Test1396
Edward Stupple, Maggie Gale, Christopher Richmond
Syllable Frequency and Stress Priming Interact in Reading Italian Aloud ..... 1402
Simone Sulpizio, Remo Job
How Does Prospect Theory Reflect Heuristics' Probability Sensitivity in Risky Choice? ..... 1408
Renata Suter, Thorsten Pachur, Ralph Hertwig
Scope of Real Beliefs in Belief Revision ..... 1414
Alexander Swan, Alexandra Chambers, Russell Revlin
8-month-olds Know Which Face is Reliable ..... 1420
Kristen Swan Tummeltshammer, Rachel Wu, Natasha Z. Kirkham
Using intrinsic complexity of turn-taking games to predict participants' reaction times ..... 1426
Jakub Szymanik, Ben Meijering, Rineke Verbrugge
A neural network model of working memory for episodes ..... 1432
Martin Takac, Alistair Knott
An Experimental Investigation of Adaptive Algorithm Understanding ..... 1438
Kazunori Terada, Seiji Yamada, Akira Ito
A Discussion on the Consistency of Driving Behavior across Laboratory and Real Situational Studies ..... 1444
Hitoshi Terai, Kazuhisa Miwa, Hiroyuki Okuda, Yuichi Tazaki, Tatsuya Suzuki, Kazuaki Kojima, Junya Morita, Akihiro Maehigashi, Kazuya Takeda
Is Lexical Access Driven by Temporal Order or Perceptual Salience? Evidence from British Sign Language ..... 1450
Robin L. Thompson, David P. Vinson, Neil Fox, Gabriella Vigliocco
A Natural Cure for the Pet Fish Problem: Feature Emergence as Classificatory Composition ..... 1456 Chris Thornton
Learning the language of time: Children's acquisition of duration words ..... 1462
Katharine Tillman, David Barner
Exploring the role of verbal category labels in flexible cognition ..... 1468
Jackson Tolins, Eliana Colunga
Diverse Evidence for Dissociable Processes in Inductive Reasoning ..... 1474
Eoin Travers, Aidan Feeney
Automatic and Strategic Search During Analogical Retrieval ..... 1480
Máximo Trench, Valeria Olguín, Adrián Margni, Ricardo Minervino
A Dynamic Dual-Process Model of Decision-making Under Uncertainty ..... 1486
Jennifer Trueblood
Expert marker of Chinese character recognition: Left-side bias versus holistic processing? ..... 1492
Ricky Van-yip Tso, Terry Kit-fong Au, Janet Hui-wen Hsiao
A distributional learning account of the acquisition of the locative alternation: Corpus analysis and modeling ..... 1498Katherine Twomey, Franklin Chang, Ben Ambridge
Moves in the World Are Faster than Moves in the Head: Interactivity in the River Crossing Problem ..... 1504
Frédéric Vallée-Tourangeau, Lisa Guthrie, Gaëlle Villejoubert
A Dynamical Model of Risky Choice ..... 1510
Marieke van Rooij, Luis Favela, MaryLauren Malone, Michael Richardson
Evidence for nonconscious behavior-copying in young children ..... 1516Johanna E. van Schaik, Rick van Baaren, Harold Bekkering, Sabine Hunnius
Joint Action Coordination through Strategic Reduction of Variability ..... 1522
Cordula Vesper, Laura Schmitz, Natalie Sebanz, Günther Knoblich
Incidental and Non-Incidental Processing of Biological Motion: Orientation, Attention and Life
Detection ..... 1528Peter Veto, Serge Thill, Paul Hemeren
The Role of Circadian Variations and Socially Distributed Thinking in Belief Perseverance ..... 1534
Gaëlle Villejoubert, Madiha Khan, Frédéric Vallée-Tourangeau
How does emotional content affect lexical processing? ..... 1540David Vinson, Marta Ponari, Gabriella Vigliocco
Does the Influence of Stress on Financial Risk Taking Depend on the Riskiness of the Decision? ..... 1546
Bettina von Helversen, Jörg Rieskamp
Negative observations, induction and the generation of hypotheses ..... 1552
Wouter Voorspoels, Chayenne Van Meel, Gert Storms
Explaining to Others Prompts Children to Favor Inductively Rich Properties ..... 1558
Caren M. Walker, Tania Lombrozo, Cristine H. Legare, Alison Gopnik
24-Month-Olds Engage in Relational Causal Reasoning ..... 1564
Caren M. Walker, Alison Gopnik
Programming of saccades to double-step targets in scene viewing: A test of assumptions present in the CRISP model ..... 1569R. Calen Walshe, Antje Nuthmann
New Empirical Tests of a Quantum Model for Question Order Effects ..... 1575
Zheng Wang, Tyler Solloway, Jerome Busemeyer
Identifying Predictive Collocations ..... 1581Silas Weinbach, Vera Demberg
The Cognitive Reflection Test: how much more than Numerical Ability? ..... 1587Matthew Welsh, Nicholas Burns, Paul Delfabbro
How do you know that? Sensitivity to statistical dependency in social learning ..... 1593
Andrew Whalen, Daphna Buchsbaum, Thomas Griffiths
A quantum probability perspective on the nature of psychological uncertainty ..... 1599
Lee White, Emmanuel Pothos, Jerome Busemeyer
Learning nonadjacent dependencies in thought, language, and action: Not so hard after all. ..... 1605
Jon Willits
Acquisition of Phrase Structure in an Artificial Visual Grammar ..... 1611
Sarah Wilson, Carla Hudson Kam
The Role of Scene Gist and Spatial Dependency among Objects in the Semantic Guidance of Attention ..... 1617
Chia-Chien Wu, Hsueh-Cheng Wang, Marc Pomplun
Framing effects in evaluation of accuracy of others' predictions ..... 1623
Saiwing Yeung
Cognitive Science, Aesthetics, and the Development of Taste ..... 1629
William York
An Embodied Perspective of Early Language Exposure ..... 1635
Hanako Yoshida, Joseph M. Burling
Online Processing of Speech and Social Information in Early Word Learning ..... 1641
Daniel Yurovsky, Anna Wade, Michael Frank
Cheap but Clever: Human Active Learning in a Bandit Setting ..... 1647
Shunan Zhang, Angela Yu
Knowledge tracing and cue contrast: Second language English grammar instruction ..... 1653
Helen Zhao, Kenneth Koedinger, John Kowalski
Word Reading Practice Reduces Stroop Interference in Children ..... 1659
Laoura Ziaka, Despoina Moirou, Eleni L Vlahou, Athanassios Protopapas
Context-freeness Revisited ..... 1664Willem Zuidema
The Effects of Overt Head Movements on Valenced Image Recognition and Recall in Bulgarians ..... 1670
Meryl Varadinov, Lyuben Laskin
Publication-Based Talks
Sequential Structure Suffices to Solve Nativist Puzzles ..... 1676
Rens Bod
Whorf for the 21st century: From interactive processing to linguistic relativity ..... 1678
Gary Lupyan
How many features does it take to change a lightbulb? ..... 1680
David Powers, David Powers
Experiments in dynamic group action and decision making: How crowds of people can walk a tightrope together and survive a zombie attack ..... 1682
Daniel C. Richardson, Peter Riefer, Bradley Love, Beau Lotto, Richard C. Clarke, Rick
Dale, John Rogers, James Ireland
Mental Simulation for Grounding Object Cognition ..... 1684
Haline E. Schendan
Posters
Approximating Bayesian inference with a sparse distributed memory system ..... 1686
Joshua Abbott, Jessica Hamrick, Thomas Griffiths
The Effect of Physical Load in Cognitive Process of Estimation ..... 1692
Keiga Abe
Automatic Labeling of Phonesthemic Senses ..... 1696Ekaterina Abramova, Raquel Fernández, Federico Sangati
Can Playing Portal Affect Spatial Thinking and Increase Learning in a STEM Area? ..... 1702
Deanne Adams, Richard Mayer
Dancing With Myself: The effect of majority group size on perceptions of majority and minority robot group members ..... 1708Henny Admoni, Bradley Hayes, David Feil-Seifer, Daniel Ullman, Brian Scassellati
Gesture and Language Production in Communication through Bar Graphs ..... 1714
Ozge Alacam, Christopher Habel, Cengiz Acartürk
Reactivity effects of concurrent verbalisation during a graph comprehension task ..... 1720
Nadia Ali, David Peebles
A Question of Timing: The Impact of Label Synchrony on Infants' Categorisation ..... 1726
Nadja Althaus, Kim Plunkett
Shape categorization in Autism: Does it follow the pattern of typical development? ..... 1732
Joseph Amaral, Susan Collins, Shannon O'Connor, Shealan McAlister, Heidi Kloos, Carol Luzzi
Gender and Sex: The Experiential Basis of Grammar ..... 1738
Elena Andonova
Why and How to Measure the Association between Choice Options ..... 1744Sandra Andraszewicz, Jörg Rieskamp
Probabilistic Language Modeling with Hidden Stochastic Automata ..... 1750
Mark Andrews
An evolutionary account of reactions to a wrong ..... 1756
Giulia Andrighetto, Francesca Giardini, Rosaria Conte
The Effect of Left-Hand Gestures on Metaphor Explanation ..... 1762
Paraskevi Argyriou, Sotaro Kita
Cognitive Externalism meets Bounded Rationality ..... 1768Eric Arnau, Saray Ayala, Thomas Sturm
The Role of Visual Coherence in Graphical Passwords ..... 1774
Ülkü Arslan Aydın, Cengiz Acartürk, Kürşat Çağıltay
Conceptual change in proportional reasoning: Effects of collaboration, own / partner reasoning level and hypothesis testing ..... 1780
Christa Asterhan, Baruch Schwarz, Noa Cohen-Eliyahu
Spatial Gestures Point the Way: A Broader Understanding of the Gestural Referent ..... 1786
Kinnari Atit, Ilyse Resnick, Thomas Shipley, Carol Ormand, Cathryn Manduca, Tilbe Goksun, Basil Tikoff
The Role of Contextual Repetition During Fast Mapping on Word Learning ..... 1792
Emma Axelsson, Jessica Horst
On The Possibility of an Extended, Yet Reductionist Mind ..... 1798
Saray Ayala
An Oscillator Model of Categorical Rhythm Perception ..... 1803
Rasmus Bååth, Erik Lagerstedt, Peter Gärdenfors
Measuring Cultural Relativity of Emotional Valence and Arousal using Semantic Clustering and Twitter ..... 1809
Eugene Bann, Joanna Bryson
Modeling the Emergence of an Exemplar Verb in Construction Learning ..... 1815
Libby Barak, Afsaneh Fazly, Suzanne Stevenson
Social Learning in Complex Networks: The Role of Building Blocks and Environmental Change . 1821
Daniel Barkoczi, Mirta Galesic
Spoken Words Activate Cross-Linguistic Orthographic Competitors in the Absence of Phonological Overlap ..... 1827
James Bartolotti, Natalia Daniel, Viorica Marian
Syntactic category disambiguation within an architecture of human language processing ..... 1833
Peter Baumann
Roles of Diagrams in Computational Modeling of Mechanisms ..... 1839William Bechtel, Adele Abrahamsen
Using Perlin Noise to Generate Emotional Expressions in a Robot ..... 1845
Aryel Beck, Antoine Hiolle, Lola Cañamero
Response direction and sentence-tense compatibility effects: An eye tracking study ..... 1851Raymond Becker, Bridgette DeCot, Ernesto Guerra, Pia Knoeferle, Rolf Zwaan
Word Learning in the Wild: What Natural Data Can Tell Us ..... 1857Barend Beekhuizen, Afsaneh Fazly, Aida Nematzadeh, Suzanne Stevenson
Visual Support for Instructional Analogy: Context Matters ..... 1863
Kreshnik Begolli, Richland Lindsey
Priming and Conceptual Pacts in Overhearers' Adoption of Referring Expressions ..... 1869
Marieke Behnel, Chris Cummins, Lorenz Sichelschmidt, Jan De Ruiter
Gaze strategies in object identification and manipulation ..... 1875
Anna Belardinelli, Martin V. Butz
The Role of Achievement Goal Motivation in Self-Explanation and Knowledge Transfer ..... 1881
Daniel Belenky, Timothy Nokes-Malach
The Impact of Communicative Constraints on the Emergence of a Graphical Communication System ..... 1887
Till Bergmann, Rick Dale, Gary Lupyan
Analogy and Arithmetics: An HDTP-Based Model of the Calculation Circular Staircase ..... 1893
Tarek Richard Besold, Alison Pease, Martin Schmidt
The Dynamics of Anchoring in Bidirectional Associative Memory Networks ..... 1899
Sudeep Bhatia, Shereen Chaudhry
A Neurally Plausible Encoding of Word Order Information into a Semantic Vector Space ..... 1905
Peter Blouw, Chris Eliasmith
Visualizing the Invisible: Generating Explanations of Scientific Phenomena ..... 1911
Eliza Bobek
Do You Mean What You Say? The Effect of Uncertainty Avoidance on the Interpretation of Probability Expressions - A Comparative Study between Spanish and German ..... 1917
Franziska Bocklisch, Anne Georg, Steffen Bocklisch, Josef Krems
Why 100 Once Is Worse Than 10 Times 10: Dread Risks versus "Continuous" Risks ..... 1923Nicolai Bodemer, Azzurra Ruggeri, Mirta Galesic
Benefits of Graphical and Symbolic Representations for Learning and Transfer of Statistical Concepts ..... 1928
David W. Braithwaite, Robert L. Goldstone
Entrainment on the Move and in the Lab: The Walking Around Corpus ..... 1934
Susan Brennan, Katharina Schuhmann, Karla Batres
Tracking Persons Over Time is Tracking What? ..... 1940
Andrew Brook
Plausibility and Visualizability in Relational Belief Revision ..... 1946
Leandra Bucher, Jelica Nejasmic, Sabine Bertleff, Markus Knauff
Mechanistic Diagrams as Search Organizers ..... 1952Daniel Burnston
Individual and Strategy Differences in an Allocentric-Heading Recall Task ..... 1958
Heather Burte, Mary Hegarty
Explicit awareness supports conditional visual search in the retrieval guidance paradigm ..... 1964
Daniel Buttaccio, Nicholas Lange, Sowon Hahn, Rick Thomas, Eddy DavelaarIntentionality and choice1970
Benignus Ndubuisi, Ruth Byrne
Emotional Priming of Sentence Comprehension: Effects of a Speaker's Static Emotional Expression and Listener Age ..... 1976
Maria Nella Carminati, Pia Knoeferle
How to present exemplars of several categories? Interleave during active learning and block during passive learning ..... 1982
Paulo Carvalho, Robert Goldstone
An eyetracking study of children's relational thinking: The role of labels and sustained attention ..... 1988
Paulo Carvalho, Catarina Vales, Caitlin Fausey, Linda Smith
Investing Amid Uncertainty: A Test of the Domain Specific Anchoring Hypothesis ..... 1994
Hui Yih Chai, Ben Newell
Producing gestures facilitates encoding of spatial relation ..... 2000Amy Chong, Ben Choi, Elena Kwong, Jennifer Chan, Irina Chong, Mavis Ip, ChristopherYeung, Wing Chee So
Aspectual Coercion in Non-native Speakers of English ..... 2006
Ho Leung Chan
Space, agency and word order: Evidence from Greek ..... 2012
Spyridoula Cheimariou, Sonia Loui
Whether Chinese Speakers Think about Time More Vertically Depends on their Immediate and
Lifetime Experience of Reading Horizontal or Vertical Texts: Evidence from Contextual Priming ..... 2017
Jenn-Yeu Chen, Michael Friederich, Hua Shu
Hemispheric Asymmetry in Nonconscious Processing ..... 2022
Jing Chen, Janet Hui-wen Hsiao
Generative Inferences Based on a Discriminative Bayesian Model of Relation Learning ..... 2028
Dawn Chen, Hongjing Lu, Keith Holyoak
Logical Consistency and Objectivity in Causal Learning ..... 2034Patricia Cheng, Mimi Liljeholm, Catherine Sandhofer
Inferring One's Own Prosociality Through Choice: Giving Preschoolers Costly Prosocial Choices Increases Subsequent Sharing Behavior ..... 2040
Nadia Chernyak, Tamar Kushnir
Estimating Multiple Evoked Emotions from Videos ..... 2046Wonhee Choe, Hyo-Sun Chun, Junhyug Noh, Seong-Deok Lee, Byoung-Tak Zhang
How working memory capacity constrains the learning of relational concepts ..... 2052
Adam Chuderski, Anna Chuderska
Face age and social status exert different modulatory effects on gaze following behaviour ..... 2058
Francesca Ciardo, Barbara F.M. Marino, Angela Rossetti, Rossana Actis-Grosso, Paola Ricciardelli
Towards a Statistical Model of Grammaticality ..... 2064
Alexander Clark, Gianluca Giorgolo, Shalom Lappin
Knowledge Helps: Mechanistic Information and Numeric Evidence as Cognitive Levers to Overcome Stasis and Build Public Consensus on Climate Change ..... 2070
Dav Clark, Michael Ranney, Jacqueline Felipe
Fractal Structure of the Nested Actions of Keeping the Beat ..... 2076
Charles Coey, Justin Hassebrock, Heidi Kloos, Michael Richardson
Motor experience interacts with effector information during action prediction ..... 2082
Lincoln Colling, William Thompson, John Sutton
Models of Human Category Learning: Do they Generalize? ..... 2088Nolan Conaway, Kenneth Kurtz
Enhancing Robust Learning Through Problem Solving in the Genetics Cognitive Tutor ..... 2094Albert Corbett, Ben MacLaren, Angela Wagner, Linda Kauffman, Aaron Mitchell, RyanBaker
Similarity and Strategic Effects in Recognition Memory ..... 2100
Gregory Cox, George Kachergis, Richard Shiffrin
Dimensions of specificity in musical memory: Evidence from metrical restoration ..... 2106
Sarah Creel
Reasoning with differing tasks and response formats ..... 2112Nicole Cruz de Echeverría Loebell, Markus Knauff
Artificial grammar learning of shape-based noun classification ..... 2118
Jennifer Culbertson, Colin Wilson
Moving from Levels $\xi^{3}$ Reduction to Dimensions $\S$ Constraints ..... 2124
David Danks
Proactive and Retroactive Interference Effects in Development ..... 2130
Kevin Darby, Vladimir Sloutsky
Self-Regulated Learning with Graphical Overviews: When Spatial Information Detracts from Learning ..... 2136
Sarah Davies, Kirsten Butcher, Corey Stevens
Associative strength and semantic activation in the mental lexicon: evidence from continued word associations. ..... 2142Simon De Deyne, Daniel Navarro, Gert Storms
The Index of Cognitive Activity as a Measure of Linguistic Processing ..... 2148
Vera Demberg, Evangelia Kiagia, Asad Sayeed
Pupillometry: the Index of Cognitive Activity in a dual-task study ..... 2154
Vera Demberg
Integration Costs on Auxiliaries? - a self-paced reading study using WebExp ..... 2160
Vera Demberg
Effects of Training on Categorization ..... 2166
Wei (Sophia) Deng, Vladimir Sloutsky
Information-Sharing in Three Interacting Minds Solving a Simple Perceptual Task ..... 2172
Michał Denkiewicz, Joanna Rączaszek-Leonardi, Piotr Migdał, Dariusz Plewczynski
Information Sampling, Conformity and Collective Mistaken Beliefs ..... 2177
Jerker Denrell, Gael Le Mens
Speed Facilitation In The Absence Of Enhanced Recognition For Target-Aligned But Irrelevant Stimuli Under Cross-modal Presentations ..... 2183Andrew Dewald, Scott Sinnett
The Great Deceivers: Virtual Agents and Believable Lies ..... 2189
João Dias, Ruth Aylett, Ana Paiva, Henrique Reis
One for all, all for one: Agents with social identities ..... 2195
Joana Dimas, Phil Lopes, Rui Prada
Constraining ACT-R Models of Decision Strategies: An Experimental Paradigm ..... 2201
Cvetomir Dimov, Julian Marewski, Lael Schooler
Planning ahead through space and time: from neuropsychology to motor control ..... 2207Mariama Dione, Laurent Ott, Yvonne Delevoye-Turrell
Spatial congruity effects reveal metaphors, not markedness ..... 2213Sarah Dolscheid, Cleve Graver, Daniel Casasanto
A two process account of the memory improvement due to choosing ..... 2219Charles Driver
Motion in Vision and Language: Seeing Visual Motion can influence Processing of Motion Verbs ..... 2225
Carolin Dudschig, Jan Souman, Barbara Kaup
Using Music as a Turn in Conversation in a Lesson ..... 2231
Sam Duffy, Patrick Healey
Object-based Saliency as a Predictor of Attention in Visual Tasks ..... 2237
Michal Dziemianko, Alasdair Clarke, Frank Keller
Verbal and Nonverbal Cues Activate Concepts Differently, at Different Times ..... 2243
Pierce Edmiston, Gary Lupyan
Language and cognitive load in a dual task environment ..... 2249Nikolaos Engonopoulos, Asad Sayeed, Vera Demberg
Math and Metacognition: Resolving the Paradox ..... 2255
Shanna Erickson, Evan Heit
Candle-Candle-Candle-Candy: Continuous Attraction Toward Previously Seen Phonological Competitors ..... 2261
Jesse Falke, Bodo Winter, Michael Spivey
Communication and Categorization: New Insights into the Relation Between Speech, Labels and Concepts for Infants ..... 2267
Brock Ferguson, Sandra Waxman
Stopping Rule Selection (SRS) Theory Applied to Deferred Decision Making ..... 2273
Mario Fific, Marcus Buckmann
Syntactic priming in language comprehension allows linguistic expectations to converge on the statistics of the input ..... 2279
Alex Fine, T. Florian Jaeger
How high can you jump? Children's Perception of Affordances for Self and Others with Different Abilities ..... 2285
Justin Fine, Heidi Kloos, Mona Jenkins
The Effect of Non-Linguistic Patterns on Linguistic Biases ..... 2291
Sara Finley
Using Causal Induction in Humans to Learn and Infer Causality from Video ..... 2297
Amy Fire, Song-Chun Zhu
Seeing the forest for the trees predicts accumulation decisions ..... 2303Helen Fischer, Cleotile Gonzalez
Grammatical Aspect influences Event Duration Estimations: Evidence from Dutch ..... 2309
Monique Flecken, Johannes Gerwien
Time (also) flies from left to right... if it is needed! ..... 2315
Andrea Flumini, Julio Santiago
Surprise! You've Got Some Explaining to Do ..... 2321
Meadhbh I Foster, Mark T Keane
Frequent Frames, Flexible Frames and the Noun-Verb Asymmetry. ..... 2327
Daniel Freudenthal, Julian Pine, Gary Jones, Fernand Gobet
Emotional Speech Processing and Language Knowledge ..... 2333
Conor Frye, Sarah Creel
Cooperation Decreases with Development of Number Sense ..... 2338
Ellen Furlong, John Opfer
A Computational Model of Two Cognitive Transitions Underlying Cultural Evolution ..... 2344
Liane Gabora, Wei Wen Chia, Hadi Firouzi
Completion in the Wild: Perception of $3 D$ forms from cross-sections ..... 2350
Kristin Gagnier, Thomas Shipley
Computational exploration of task and attention modulation on holistic processing and left side bias effects in face recognition: the case of face drawing experts. ..... 2356
Bruno Galmar, Janet Hui-wen Hsiao
A joint interference effect in picture naming ..... 2362
Chiara Gambi, Joris Van de Cavey, Martin Pickering
Semantic Ambiguity Resolution as a Decision Process ..... 2368Nicholas Gaylord, Colin Bannard
Individual Differences, Imagery and the Visual Impedance Effect ..... 2374Lupita Estefania Gazzo Castaneda, Markus Knauff
Playing for Us: The Influence of Joint Action on Planning in Three-year-olds ..... 2380
Sarah Gerson, Sabine Hunnius, Harold Bekkering
Back on track: Backtracking in counterfactual reasoning ..... 2386Tobias Gerstenberg, Christos Bechlivanidis, David A. Lagnado
Musical training and auditory global-local precedence ..... 2392Anne Gilman, Elizabeth Ebbets
Visual Strategies in Analogical Reasoning Development: A New Method for Classifying Scanpaths2398
Yannick Glady, Jean-Pierre Thibaut, Bob French
A ritical View on Conceptual Blending Theory ..... 2404
Vladimir Glebkin
Primacy/recency effects in infant categorisation ..... 2410
Valentina Gliozzi, Nadja Althaus, Julien Mayor, Kim Plunkett
Inventing Prepares Learning Motivationally, but a Worked-out Solution Enhances Learning Outcomes ..... 2416Inga Glogger, Julian Kappich, Rolf Schwonke, Lars Holzäpfel, Matthias Nückles, Alexander
Renkl
Development of Semantic Knowledge and Its Role in the Development of Category-Based
Reasoning ..... 2422
Karrie Godwin, Bryan Matlen, Anna Fisher
Classroom activities and off-task behavior in elementary school children ..... 2428
Karrie Godwin, Victoria Almeda, Megan Petroccia, Ryan Baker, Anna Fisher
How pitch accents and focus particles affect the recognition of contextual alternatives ..... 2434
Nicole Gotzner, Katharina Spalek, Isabell Wartenburger
The object without qualities: referring with negative properties ..... 2440
Martijn Goudbeek, Inge Haagmans, Emiel Krahmer
Lay biology in health: How adults conceptualize the benefits of exercise ..... 2446
Sarah Gripshover, Ellen Markman
Adaptive Estimation of Psychometric Slope and Threshold with Differential Evolution ..... 2452Hairong Gu, Jay Myung, Mark Pitt, Zhong-Lin Lu
Spatial meanings for function words? The link between conjunctions and spatial representations ..... 2458
Ernesto Guerra, Tyler Marghetis, Pia Knoeferle
How does social comparison affect regret and relief in children, adolescents and adults? ..... 2464
Marianne Habib, Sylvain Moutier, Grégoire Borst, Olivier Houdé, Mathieu Cassotti
Unifying Theories of Consciousness, Attention, and Conscious Attention ..... 2470
Harry Haladjian, Carlos Montemayor
Discovering Quantification and Number in a Role-Filler Model ..... 2476Aaron Hamer, Leonidas Doumas
Disambiguating NN combinations with left/right stress ..... 2482James Hampton, Daniel Heussen, Zarah Argel, Hasina Kanbi
Improving First-Year Writing Using Argument Diagramming ..... 2488
Maralee Harrell, Danielle Wetzel
Why verbalization of facial features increases false positive responses on visually-similar distractors: A computational exploration of verbal overshadowing ..... 2494
Aya Hatano, Taiji Ueno, Shinji Kitagami, Jun Kawaguchi
The effect of "Trust dynamics": Perspective taking during collaborative problem solving ..... 2500
Yugo Hayashi
How listeners respond to speaker's troubles ..... 2506Patrick Healey, Mary Lavelle, Christine Howes, Stuart Battersby, Rosemaire McCabe
A Computational Model of Systems Memory Reconsolidation ..... 2512
Peter Helfer, Thomas Shultz, Oliver Hardt, Karim Nader
Expert Blind Spot in Pre-Service and In-Service Mathematics Teachers: Task Design moderates
Overestimation of Novices' Performance ..... 2518
Katharina Hellmannn, Matthias Nückles
The Recognizability of Authenticity ..... 2524
Madeleine Henderson, Liane Gabora
Benefits for Processes Cause Decrements in Outcomes: Training Improves Tutors' Interactivity at the Expense of Assessment Accuracy ..... 2530
Stephanie Herppich, Jörg Wittwer, Matthias Nückles, Alexander Renkl
Blending and choosing within one mind: Should judgments be based on exemplars, rules or both? ..... 2536
Stefan Herzog, Bettina von Helversen
Using the letter decision task to examine semantic priming ..... 2542
Tom Heyman, Simon De Deyne, Gert Storms
Topological Similarity of Motor Coordination in Rhythmic Movements ..... 2548
Shohei Hidaka, Tsutomu Fujinami
Eye Movements Reveal Interplay Between Noun Capitalization and Word Class During Reading ..... 2554
Sven Hohenstein, Reinhold Kliegl
Here's not looking at you, kid! Unaddressed recipients benefit from co-speech gestures when speech processing suffers ..... 2560
Judith Holler, Louise Schubotz, Spencer Kelly, Peter Hagoort, Manuela Schuetze, Asli Ozyurek
Encountering Multiple Exemplars During Fast Mapping Facilitates Word Learning ..... 2566
Jessica Horst, Katherine Twomey, Samantha Ranson
Dilution Effects in Perceptual Information Integration ..... 2572
Jared Hotaling, Andrew Cohen, Jerome Busemeyer, Richard Shiffrin
Cooperation in Prisoner's Dilemma Game: Influence of Players' Social Roles ..... 2578
Evgenia Hristova, Maurice Grinberg, Iskra Georgieva, Milena Borisova
When does the majority rule? Preschoolers' trust in majority informants varies by task domain ..... 2584
Jane Hu, Daphna Buchsbaum, Thomas Griffiths, Fei Xu
Visual motion processing and perceptual decision making ..... 2590
Aziz Hurzook, Oliver Trujillo, Chris Eliasmith
What's Up can be Explained by Language Statistics ..... 2596
Sterling Hutchinson, Max Louwerse
Influences Beyond Language: Comparing Spatial Referencing in Native French Speakers from Four Countries ..... 2602
Lisa Hüther, Anne Bentz, Hans Spada, Andrea Bender, Sieghard Beller
Watching Gestures during Learning about Movements with Dynamic Visualization Activates the Human Mirror Neuron System: A fNIRS Study ..... 2608
Birgit Imhof, Ann-Christine Ehlis, Florian B. Häußinger, Peter Gerjets
A robustness approach to theory building: A case study of language evolution ..... 2614
Liz Irvine, Sean Roberts, Simon Kirby
Context Dependent Utility: Modeling Decision Behavior Across Contexts ..... 2620
Jonathan Ito, Stacy Marsella
A computational exploration on the role of semantic memory in episodic future thinking ..... 2626
Yuichi Ito, Taiji Ueno, Shinji Kitagami, Jun Kawaguchi
The Impact of Problem Space on Reasoning: Solving versus Creating Matrices ..... 2632Saskia Jaarsveld, Thomas Lachmann, Cees van Leeuwen
Visual Recognition using a Combination of Shape and Color Features ..... 2638
Sepehr Jalali, Cheston Tan, Joo-Hwee Lim, Jo-Yew Tham, Sim-Heng Ong, Paul Seekings, Elizabeth Taylor
Encoding Co-occurrence of Features in the HMAX Model ..... 2644
Sepehr Jalali, Cheston Tan, Joo-Hwee Lim, Jo-Yew Tham, Sim-Heng Ong, Paul Seekings, Elizabeth Taylor
The Assumption of Class-conditional Independence in Category Learning ..... 2650
Jana Jarecki, Björn Meder, Jonathan D. Nelson
Similarity, causality and argumentation ..... 2656
William Jimenez-Leal, Christian Gaviria
Good Decisions, Good Causes: Optimality as a Constraint on Attribution of Causal Responsibility ..... 2662
Samuel G. B. Johnson, Lance J. Rips
Learning complementary action with differences in goal knowledge ..... 2668
Jeremy Karnowski, Edwin Hutchins
Representational Shifts Towards the Prototype in Memory for Hue ..... 2674
Laura Kelly, Evan Heit
Adaptive Foraging: Effects of Resource Conditions on Search Paths in a Web-Based Foraging Game ..... 2680Bryan Kerster, Christopher Kello, Theo Rhodes, Ralph Bien-Aime
Neural-network Modelling of Bayesian Learning and Inference ..... 2686
Milad Kharratzadeh, Thomas Shultz
Uncertainty can increase explanatory credibility ..... 2692
Sangeet Khemlani, Daniel Gartenberg, Kun Hee Park, J. Gregory Trafton
Mental simulation and the construction of informal algorithms ..... 2698
Sangeet Khemlani, Phil Johnson-Laird
Channels of multimodal communication: Relative contributions to discourse understanding ..... 2704
Andrej Kibrik, Natalia Molchanova
Quantifying the Coherence of Pedestrian Groups ..... 2710
Adam Kiefer, Stephane Bonneaud, Kevin Rio, William Warren
Naturalistic Word-Concept Pair Learning With Semantic Spaces ..... 2716Brent Kievit-Kylar, George Kachergis, Michael Jones
Selection of Linker Type in Emphatic Reduplication: Speaker's Intuition meets Corpus Statistics ..... 2722
Özkan Kılıç, Cem Bozşahin
Connecting Learning Goals and Component Cognitive Skills in Digital Games ..... 2728Stephen Killingsworth, Douglas Clark
Two for one? Transfer of conceptual content in bilingual number word learning ..... 2734
Katherine Kimura, Katie Wagner, David Barner
The Effects of Personality in a Social Context ..... 2740
Kirsty Kitto, Fabio Boschetti
Comprehension cueing strategies in elderly: a window into cognitive decline? ..... 2746Vanja Kljajevic, Viviana Fratini, Aitziber Etxaniz, Elena Urdaneta, Javier. Yanguas
Returning the Ticket - Mental Time Travel Reconsidered ..... 2752
Markus Kneer
The time-course of processing discourse connectives ..... 2760
Judith Köhne, Vera Demberg
The role of morphology in spelling: Long-term effects of training ..... 2766
Kendall Kolne, Katherine Hill, Laura GonnermanHow big is the BFG? The impact of redundant size adjectives on size perception2772
Emiel Krahmer, Marret Noordewier, Martijn Goudbeek, Ruud Koolen
The case of cognitive ecology for cognitive processes in everyday life situations ..... 2778
Mattias Kristiansson
What Aspects of Cyber Cruelty are judged most distressing? An Adaptive Conjoint Study with Two Independent Samples ..... 2784
Christina Kuhlmann, Stephanie Pieschl, Torsten Porsch
Hierarchical Control and Sense of Agency: Differential Effects of Control on Implicit and
Explicit measures of Agency ..... 2790
Devpriya Kumar, Narayanan Srinivasan
Methods for Classifying Errors on the Raven's Standard Progressive Matrices Test ..... 2796
Maithilee Kunda, Isabelle Soulières, Agata Rozga, Ashok Goel
Experts' Explanations Engage Novices in Deep-Processing ..... 2802
Andreas Lachner, Matthias Nückles
Do risk-averse people lie less? A comparison of risk-taking behavior in deceptive and non-deceptive scenarios ..... 2808
Tei Laine, Kayo Sakamoto, Tomi Silander
Maximizing learning from collaborative activities ..... 2814
Rachel Lam
Measuring the separate effects of practice and fatigue on eye movements during visual search ..... 2820
Sophie Lanthier, Evan Risko, Daniel Smilek, Alan Kingstone
Applying the dynamics of post-synaptic potentials to individual units in simulation of temporally extended ERP reading data ..... 2826
Sarah Laszlo, Blair Armstrong
Movement Correlation as a Nonverbal Cue in the Judgment of Affiliation during Social Interaction ..... 2832
Nida Latif, Adriano Barbosa, Eric Vatikiotis-Batesom, Monica Castelhano, Kevin Munhall
Overcoming Screen Inferiority in Text Learning ..... 2838
Tirza Lauterman, Rakefet Ackerman
Effects of Explicit Abstract Knowledge and Simple Associations in Sequence Learning ..... 2843
Jessica Lee, Evan Livesey
Learning and representation of causative motor actions: a neural network model and its use in an embodied theory of language ..... 2849
Jeremy Lee-Hand, Alistair Knott
Coordinate-free Thinking ..... 2855
Joop Leo
Categorization and Abstract Similarity in Chess ..... 2860
Pablo León-Villagrá, Frank Jäkel
Prior Linguistic Knowledge Influences Implicit Language Learning ..... 2866
Janny H.C. Leung, John N. Williams
Reading Motor Intentions ..... 2872Daniel Lewkowicz, Yvonne Delevoye-Turrell
The Influence of Collective Opinion on True-False Judgment and Information-Sharing Decision ..... 2878
Huaye Li, Yasuaki Sakamoto
Different Strategies in Solving Series Completion Inductive Reasoning Problems: An fMRI and Computational Study ..... 2884
Peipeng Liang, Xiuqin Jia, Niels A. Taatgen, Ning Zhong, Kuncheng Li
Inductive and deductive reasoning in Obsessive-Compulsive Disorder ..... 2890
Janice Liew, Brett Hayes
The Influence of the Attention Set on Exogenous Orienting ..... 2896
Ahnate Lim, Scott Sinnett
Modeling the Relational Shift in Melodic Processing of Young Children ..... 2902
Ahnate Lim, Leonidas Doumas, Scott Sinnett
Effects of Response and Presentation Format on Measures of Approximate Number System Acuity ..... 2908
Marcus Lindskog, Anders Winman, Peter Juslin
Effects of Text Titles and the Timing of Keywording Tasks on Metacognitive Monitoring ..... 2914
Marie Lippmann, Susanne Narciss, Neil H. Schwartz, Robert W. Danielson
The relationship between blocking and inference in causal learning. ..... 2920
Evan Livesey, Jessica Lee, Lauren Shone
Modeling Continuous Representations in Visual Working Memory ..... 2926
Johannes Lohmann, Martin Butz
Sign processes in emergence of communication ..... 2932Angelo Loula, Ricardo Gudwin, João Queiroz
Spatial correspondence parameters at the basis of transfer of learning in social contexts ..... 2938
Luisa Lugli, Cristina Iani, Nadia Milanese, Sandro Rubichi
Body movements affect Counting ..... 2944Luisa Lugli, Giulia Baroni, Filomena Anelli, Anna M Borghi, Roberto Nicoletti
Won't You Think of the Children?: Traits Predicting Intergenerational Preferences ..... 2949
Christian Luhmann, Sarah Pak
Tell Us Your Story: Investigating the Linguistic Features of Trauma Narrative ..... 2955
Jeremy A. Luno, Max Louwerse, J. Gayle Beck
Language dominance modulates cross-language lexical interaction in late immersed learners ..... 2961
Barbara Malt, Ping Li, Eef Ameel, Aneta Pavlenko, Huichun Zhu
Do Language Structure or Language Proficiency Affect Critical Evaluation? ..... 2967
Emmanuel Manalo, Kyoko Watanabe, Chris Sheppard
Moral choices: The influence of the "Do not play God" principle ..... 2973
Francesco Mancini
Towards agents with human-like decisions under uncertainty ..... 2978Nuno Marques, Francisco Melo, Samuel Mascarenhas, João Dias, Rui Prada, Ana Paiva
A Computational Framework for Attentional 3D Object Detection ..... 2984
German Martin Garcia, Simone Frintrop
Effect of Positive and Negative Instances on Rule Discovery: Investigation Using Eye Tracking . 2990
Miki Matsumuro, Kazuhisa Miwa
Automatic generation of naturalistic child-adult interaction data ..... 2996Yevgen Matusevych, Afra Alishahi, Paul Vogt
Coordinating turning while walking and talking ..... 3002
Eric Mayor, Adrian Bangerter
Insight Follows Incubation In The Compound Remote Associates Task ..... 3008
Sean Mc Carthy, John Malony, Robert Morrison
Modeling a reaction time variant of the Perruchet effect in humans ..... 3014Amy McAndrew, Fayme Yeates, Frederick Verbruggen, Ian P. L. McLaren
Can US sensitization account for the electrodermal variant of the Perruchet effect? ..... 3020
Amy McAndrew, Gabrielle Weidemann, Ian P.L. McLaren
Reasoning about diamonds, gravity and mental states: The cognitive costs of theory of mind ..... 3026
Ben Meijering, Hedderik Van Rijn, Niels A Taatgen, Rineke Verbrugge
Modeling the Development of Determiner Productivity in Children's Early Speech ..... 3032Stephan Meylan, Michael Frank, Roger Levy
Spatial Reasoning in Native Speakers of Russian and German ..... 3038
Maria Mikheeva, Leandra Bucher, Jelica Nejasmic, Markus Knauff
Problem Solving Between Action-Selection and Action-Completion in a Simple Domain ..... 3044
Gareth Miles, Stephen Payne
A Mechanistic Account of Computational Explanation in Cognitive Science ..... 3050
Marcin Miłkowski
Distinguishing between Category-based and Similarity-based Induction ..... 3056
Tracey Miser, Vladimir Sloutsky
Typicality and Object Reference ..... 3062Margaret Mitchell, Ehud Reiter, Kees van Deemter
Stoic Behavior in Hint Seeking when Learning using an Intelligent Tutoring System ..... 3068
Kazuhisa Miwa, Hitoshi Terai, Nana Kanzaki, Ryuichi Nakaike
Unmet Expectations in the Comprehension of Relative Clauses in Japanese ..... 3074
Edson T. Miyamoto, Michiko Nakamura
The facilitatory role of sound symbolism in infant word learning ..... 3080
Michiko Miyazaki, Shohei Hidaka, Mutsumi Imai, H. Henny Yeung, Katerina Kantartzis, Hiroyuki Okada, Sotaro Kita
Beliefs About a Speaker Affect Feeling of Another's Knowing ..... 3086Lisette Mol, Anna Kuhlen, Ruben van der Steen, Mike Obbens
Sharing in the Dark? Target Memory and Risk Awareness in Online Communication ..... 3092
Ricarda Moll, Stephanie Pieschl, Rainer Bromme
Automatic Generation of Music for Inducing Physiological Response ..... 3098
Kristine Monteith, Bruce Brown, Dan Ventura, Tony Martinez
How Healthy Aging and Dementia Impact Memory Search ..... 3104
Ana Sofia Morais, Hansjörg Neth, Thomas Hills
Does Friendship Influence Space Perception? With Particular Reference to the Curse of the Suspicious Participants ..... 3110Nicolas Morgado, Dominique Muller, Mathieu Pinelli, Éric Guinet, Édouard Gentaz,Richard Palluel-Germain
Joint Commitment: An Analysis of Emotions and Non-Verbal Behaviors ..... 3116
Francesca Morganti, Antonella Carassa, Marco Colombetti, Paride Braibanti, Giorgio Rezzonico, Ivana Sgro
Causal determinism in toddlers ..... 3122
Paul Muentener, Laura Schulz
Spatial Distance, Availability of Means, and Moral Obligation Judgments ..... 3127
Jonas Nagel, Michael R. Waldmann
Value of a friend, a friend of your friend, and a friend of the friend of your friend:Social discounting in $n$ degrees of separation ..... 3133
Kuninori Nakamura
Managing our debt: Changing Context Reduces Misunderstanding of Global Warming ..... 3139
Ben Newell, Arthur Kary, Chris Moore, Cleotilde Gonzalez
Non-parametric estimation of the individual's utility map ..... 3145
Takao Noguchi, Adam Sanborn, Neil Stewart
$k$-Nearest Neighbor Classification Algorithm for Multiple Choice Sequential Sampling ..... 3151
Yung-Kyun Noh, Frank Chongwoo Park, Daniel D. Lee
Where What You Count is What Really Counts ..... 3157
Nader Noori, Laurent Itti
Traces of Intellectual Working Memory Tasks on Visuospatial Short-Term Memory ..... 3163
Nader Noori, Laurent Itti
Measuring the comprehension of negation in 2- to 4-year-old children ..... 3169
Ann Nordmeyer, Michael Frank
Limitations and chances of working memory training ..... 3175
Daniela Nussbaumer, Roland H. Grabner, Michael Schneider, Elsbeth Stern
The synergy of top-down and bottom-up attention in complex task: going beyond saliency models. ..... 3181
Enkhbold Nyamsuren, Niels Taatgen
Is Facial Expression Processing Holistic? ..... 3187
Akinyinka Omigbodun, Garrison Cottrell
Discovering Pronoun Categories using Discourse Information ..... 3193
Naho Orita, Rebecca McKeown, Naomi Feldman, Jeffrey Lidz, Jordan Boyd-Graber
Constraints on Theories of Serial Order Memory Revisited: The Cases of the Fill-In and Protrusion Effects ..... 3199
Adam Osth, Simon Dennis
Modeling Gain-Loss Asymmetries in Risky Choice: The Critical Role of Probability Weighting ..... 3205
Thorsten Pachur, David Kellen
Semantic Implicit Learning in Language ..... 3211
Albertyna Paciorek, John Williams
The role of verbal labels in attention to dimensional similarity ..... 3217
Lynn Perry, Larissa Samuelson
A category theory perspective on compositionality and (the development of) cognitive capacity ..... 3223
Steven Phillips
Spontaneous Analogy by Piggybacking on a Perceptual System ..... 3229
Marc Pickett, David Aha
Type-token representations in conceptual representation ..... 3235
Sandeep Prasada
Category Change in the Absence of Falsifying Feedback ..... 3239
Jared Ramsburg, Stellan Ohlsson
The 'universal' structure of name grammars and the impact of social engineering on the evolution of natural information systems ..... 3245
Michael Ramscar, Asha Halima Smith, Melody Dye, Richard Futrell, Peter Hendrix, Harald Baayen, Rebecca Starr
The Cross-Domain Re-interpretation of Artistic Ideas ..... 3251
Apara Ranjan, Liane Gabora, Brian O'Connor
A neural reinforcement learning model for tasks with unknown time delays ..... 3257Daniel Rasmussen, Chris Eliasmith
Reduced neural sensitivity to online social interactions in autism ..... 3263
Elizabeth Redcay, Rebecca Saxe
Smooth Dynamics, Good Performance in Cognitive-Agent Congestion Problems ..... 3269
David Reitter, Paul Scerri
Application of the Category Adjustment Model in Temporal, Spatial, and Abstract Magnitude at the Billions Scale ..... 3275Ilyse Resnick, Thomas F. Shipley
English Words Are Processed Like Objects ..... 3281
Gavin Revie, Yuki Kamide
SOM Cognitive Modeling of Autistic and Schizophrenic Traits Using an Oscillating Topological Neighborhood Width Function ..... 3287
Spyridon Revithis, William Wilson, Nadine Marcus
The Language of Everyday Verbal Analogies ..... 3293
Lindsey Richland, Justin Richland
Experimental insights on the origin of combinatoriality ..... 3299
Gareth Roberts, Bruno Galantucci
From Minor Mishap to Major Catastrophe: Lexical Choice in Miscommunication ..... 3303
Jennifer Roche, Alexandra Paxton, Alyssa Ibarra, Michael Tanenhaus
Effects of Target Size and Symmetry on the Structure of Variability in Precision Aiming ..... 3309
Veronica Romero, Charles A. Coey, Andrew Beach, Michael J. Richardson
The influence of perceptual and structural salience ..... 3315
Florian Röser, Antje Krumnack, Kai Hamburger
Development of Substitution Bias Sensitivity: Are Adolescents Happy Fools? ..... 3321
Sandrine Rossi, Mathieu Cassotti, Marine Agogué, Wim De Neys
Even if after If then conditionals ..... 3327
José A. Ruiz-Ballesteros, Sergio Moreno-Rios
What, When and How do the Models of Conceptual Change Explain? ..... 3331
Anna-Mari Rusanen, Otto Lappi
A PDP Model for Capturing N400 Effects in Early L2 Learners during Bilingual Word Reading Tasks ..... 3337Sepideh Sadeghi, He Pu, Matthias Scheutz, Phillip Holcomb, Katherine Midgley
The Time Course of Language Use in Multiparty Negotiations ..... 3343
Eyal Sagi, Daniel Diermeier
Interpreting Covariation in Causal Structure Learning ..... 3348Motoyuki Saito, Tsuneo Shimazaki
The Rich Who Have the Humility of the Poor: Effects of Culture and Power on Altruism ..... 3354
Niloufar Salehi, Morteza Dehghani
Intonation and positional effects in spoken serial recall ..... 3360
Michelina Savino, Andrea Bosco, Martine Grice
Is Double-Dipping an Alternative to Double-Dissociation?: Sampling Two Representational Systems Using a Single Task ..... 3366
Jordan Richard Schoenherr, Guy Lacroix
Match Me if You Can: How Smart Choices are Fueled by Competition ..... 3372
Christin Schulze, Don van Ravenzwaaij, Ben R. Newell
Learning Patterns in Noise: Environmental Statistics explain the Sequential Effect ..... 3378
Friederike Schuur, Brian P Tam, Laurence T Maloney
Infant contributions to joint attention predict vocabulary development ..... 3384
Katherine Scott, Elena Sakkalou, Kate Ellis-Davies, Elma Hilbrink, Ulrike Hahn, MeridethGattis
Intentionality, Evaluative Judgments, and Causal Structure ..... 3390
Jason Shepard, Phillip Wolff
Scientific Diagrams as Traces of Group-Dependent Cognition: A Brief Cognitive-Historical Analysis ..... 3396
Ben Sheredos
Infants’ Early Understanding of Coincidences ..... 3402
Zi Sim, Fei Xu
Exploring the Developmental Feedback Loop: Word Learning in Neural Networks and Toddlers ..... 3408
Clare Sims, Savannah Schilling, Eliana Colunga
Syntax in music and language: The role of cognitive control ..... 3414
L. Robert Slevc, Jason Reitman, Brooke Okada
Modelling the Effects of Formal Literacy Training on Language Mediated Visual Attention ..... 3420
Alastair Smith, Padraic Monaghan, Falk Huettig
Consistent physics underlying ballistic motion prediction ..... 3426
Kevin A Smith, Peter Battaglia, Edward Vul
Hear No Evil: Can Music Attenuate the Irrelevant Speech Effect? ..... 3432
Wei Jie Soh, Stephen Wee Hun Lim
Priming randomness increases the evaluation of ritual efficacy ..... 3437Andre L. Souza, Cristine H. Legare
Learning, Feedback and Information in Self-Organizing Communication Systems ..... 3442Matthew Spike, Kevin Stadler, Simon Kirby, Kenny Smith
How Do Readers Explain the Occurrence of Conflicts in Science Texts? Effects of Presentation Format and Source Expertise ..... 3448
Marc Stadtler, Lisa Scharrer, Rainer Bromme
Inductive Bias against Stem Changes as Perseveration: Experimental Evidence for an Articulatory Approach to Output-Output Faithfulness ..... 3454
Matthew Stave, Amy Smolek, Vsevolod Kapatsinski
Parsing Sequentially Presented Commands in a Large-Scale Biologically Realistic Brain Model ..... 3460
Terrence Stewart, Chris Eliasmith
Complex Mental Time Travel ..... 3468
Kurt Stocker
Representation and Criterion Differences between Men and Women in Semantic Categorization ..... 3474
Loes Stukken, Steven Verheyen, Gert Storms
A delayed discrimination task yields categorical perception of color not only in the right but also in the left visual field ..... 3480
Takashi Suegami, Bruno Laeng, Samira Aminihajibashi
Phonetic variation and the recognition of words with pronunciation variants ..... 3486
Meghan Sumner, Chigusa Kurumada, Roey Gafter, Marisa Casillas
Dyadic Cooperation Enhances Retrieval and Recall of Crossword Solutions ..... 3492
Janelle Szary, Rick Dale
A Long-Term Memory Competitive Process Model of a Common Procedural Error ..... 3498
Frank Tamborello, Greg Trafton
Implicit transfer of mirrored spatial structure in visuomotor sequence learning ..... 3504
Kanji Tanaka, Katsumi Watanabe
Inferring Subjective Prior Knowledge: An Integrative Bayesian Approach ..... 3510
Sean Tauber, Mark Steyvers
fMRI Study in Insight Problem Solving Using Japanese Remote Associates Test Based on Semantic Chunk Decomposition ..... 3516
Hitoshi Terai, Kazuhisa Miwa, Kazuaki Asami
The effect of expertise on encoding of movements and bodily indexes: a study on volleyball players ..... 3522
Alessia Tessari, Giovanni Ottoboni, Roberto Nicoletti
The Simple Advantage in Categorical Generalization of Chinese Characters ..... 3527
Khanh-Phuong Thai, Ji Son
Searching Semantic Memory as a Scale-Free Network: Evidence from Category Recall and a Wikipedia Model of Semantics ..... 3533Graham Thomposn, Christopher Kello, Priscilla MontezConstraining Bayesian Inference with Cognitive Architectures: An Updated Associative LearningMechanism in ACT-R3539
Robert Thomson, Christian Lebiere
A New Way of Linking Information Theory with Cognitive Science ..... 3545
Chris Thornton
Verifying properties from different emotions produces switching costs: Evidence for coarse-grained language statistics and fine-grained perceptual simulation ..... 3551
Richard Tillman, Sterling Hutchinson, Sara Jordan, Max Louwerse
Geographical Estimates are Explained by Perceptual Simulation and Language Statistics ..... 3557Richard Tillman, Sterling Hutchinson, Max Louwerse
Empirical investigation on spatial templates for a diagonal spatial term ..... 3563
Takenobu Tokunaga, Toshiaki Watatani, Ryu Iida, Asuka Terai
How intonation contrains pragmatic inference ..... 3569
John M. Tomlinson, Jr., Lewis Bott
The visual motion aftereffect from mental imagery depends on speed ..... 3575
Alexia Toskos Dils, Lera Boroditsky
The Multi-attribute Linear Ballistic Accumulator Model of Decision-making ..... 3581
Jennifer Trueblood, Scott Brown, Andrew Heathcote
Perceiving sounds: analytic and synthetic listening, global-local processing and possible links with empathy and self-construal ..... 3587
Olga Tsoumani, Marie Postma-Nilsenová
People, Place, and Time: Inferences from Diagrams ..... 3593
Barbara Tversky, James Corter, Jie Gao, Yuko Tanaka, Jeffrey Nickerson
Do people understand irony from computers? ..... 3598
Akira Utsumi, Yu Watanabe, Yusuke Wakayama
Learning Social Affordances and Using Them for Planning ..... 3604
Kadir F. Uyanik, Yigit Caliskan, Asil Kaan Bozcuoglu, Onur Yuruten, Sinan Kalkan, Erol Sahin
Linguistic cued attention in children: Words organize attention to shape in a visual search task . ..... 3610 Catarina Vales, Linda Smith
Reducing The Impact of Math Anxiety on Mental Arithmetic: The Importance of Distributed Cognition ..... 3615
Frédéric Vallée-Tourangeau, Miroslav Sirota, Gaëlle Villejoubert
Effects of Objects' "Embodiment" on the Acquisition of Problem-Solving Skills through Practice or Video-based Modeling Example Study ..... 3621
Tamara van Gog, Lysanne Post, Robyn ten Napel, Lian Deijkers
Using Recognition in Multi-Attribute Decision Environments ..... 3627Don van Ravenzwaaij, Ben Newell, Chris Moore, Michael Lee
The Path Less Taken: When Working Memory Capacity Constrains Insight ..... 3633
Charles Van Stockum, Marci DeCaro
The Role of Shape in Semantic Memory Organization of Objects ..... 3639
Lisanne van Weelden, Joost Schilperoord, Marc Swerts, Diane Pecher
Evaluating Two Mechanisms of Flexible Induction: Selective Memory Retrieval and Evidence Explanation ..... 3645
Nadya Vasilyeva, John Coley
An Information Foraging Model of Interactive Analogical Retrieval ..... 3651
Swaroop Vattam, Ashok Goel
Linking Cognitive Tokens to Biological Signals: Dialogue Context Improves Neural Speech Recognizer Performance ..... 3657
Richard Veale, Gordon Briggs, Matthias Scheutz
A Written Version of Sign Language can Enhance Signing Deaf Individuals' Comprehension and Learning from Texts ..... 3663
Mara Vendrame, Iaria Cutica, Monica Bucciarelli
Combinatorial structure and iconicity in artificial whistled languages ..... 3669Tessa Verhoef, Simon Kirby, Bart de Boer
Inductive Use of Semantic Word-Families to Accelerate Vocabulary Development and Reading Comprehension in Grades 3-4-5 ..... 3675
Michael Vitale, Nancy Romance
Cognitive load does not decrease pronoun use when speaker's and addressee's perspectives are dissociated ..... 3681
Jorrig Vogels, Emiel Krahmer, Alfons Maes
Rural and urban differences in language socialization and early vocabulary development in Mozambique ..... 3687
Paul Vogt, J. Douglas Mastin
Logical Patterns in Individual and General Predication ..... 3693
Momme von Sydow
The role of sampling assumptions in generalization with multiple categories ..... 3699
Wai Keen Vong, Andrew Hendrickson, Amy Perfors, Daniel Navarro
Help-Seeking As A Cause of Young Children's Collaboration ..... 3705
Christopher Vredenburgh, Tamar Kushnir
When words get physical: evidence for proficiency-modulated somatotopic motor interference during second language comprehension ..... 3711
Nikola Vukovic
Slow drift of individuals' magnitude-to-number mapping. ..... 3717
Edward Vul, David Barner, Jess Sullivan
Effects of Parafoveal Plausibility During Reading ..... 3723Laura Wakeford, Wayne Murray
Later events lie behind her, but not behind you: Compatibility effects for temporal sequences along the sagittal axis depend on perspective ..... 3729
Esther Walker, Benjamin Bergen, Rafael Nunez
A Computational Model of the Development of Hemispheric Asymmetry of Face Processing ..... 3735
Panqu Wang, Garrison Cottrell
Intentional Constraints on the Dynamics of Human Performance and Behavioral Variability in Motor Control ..... 3741
Auriel Washburn, Charles Coey, Michael Richardson
Grouping by Similarity Helps Concept Learning ..... 3747Erik Weitnauer, Paulo F. Carvalho, Robert L. Goldstone, Helge RitterModeling Bilingual Children's Acquisition of Complex Sentences in German3753
Daniel Wiechmann, Judith Steinfeld, Elma Kerz
On the Robustness of Intuitions in the two best-known Trolley Dilemmas ..... 3759Alex Wiegmann, Matthias Lippold, Robert Grigull
Diversity, Collaboration, and Learning by Invention ..... 3765
Jennifer Wiley, Olga Goldenberg, Andrew Jarosz, Michael Wiedmann, Nikol Rummel
Dual Processes in Mental State Understanding: Is Theorising Synonymous with Intuitive Thinking and is Simulation Synonymous with Reflective Thinking? ..... 3771Meredith Wilkinson, Linden Ball
Effects of Explaining Anomalies on the Generation and Evaluation of Hypotheses ..... 3777
Joseph Jay Williams, Caren Walker, Samuel Maldonado, Tania Lombrozo
Impulsivity and Overall Similarity Classification ..... 3783
Andy Wills, Chris Longmore, Fraser Milton
More is up... and right: Random number generation along two axes ..... 3789
Bodo Winter, Teenie Matlock
When groups should not imitate their most successful members ..... 3795
Jan Kristian Woike, Jean-Philippe Bonardi, Rocio Garcia-Retamero
Why we Should Not Forget About the Non-social World: Subjective Preferences, Exploratory Eye-movements, and Individual Differences ..... 3801
David W.-L. Wu, Tanya Jakobsen, Nicola C. Anderson, Walter Bischof, Alan Kingstone
The invisible hand: Toddlers infer hidden agents when events occur probabilistically ..... 3807
Yang Wu, Paul Muentener, Laura Schulz
Development of Sharing in Preschoolers in Relation to Theory of Mind Understanding ..... 3811
Zhen Wu, Yanjie Su
Worm holes in memory: Is memory one representation or many? ..... 3817
DIrk U. Wulff, Thomas T. Hills, Ralph Hertwig
Differences in Interactional Attitudes in Native and Second Languag Conversations: Quantitative Analyses of Multimodal Three-Party Corpus ..... 3823Seiichi Yamamoto, Keiko Taguchi, Ichiro Umata, Kosuke Kabashima, Masafumi Nishida
Implicit learning: A demonstration and a revision to a novel SRT paradigm ..... 3829
Fayme Yeates, Fergal W. Jones, Andy J. Wills, M.R.F. Aitken, Ian P.L. McLaren
Linguistic Variability and Adaptation in Quantifier Meanings ..... 3835
Ilker Yildirim, Judith Degen, Michael Tanenhaus, Florian Jaeger
The Role of Category Structure in Category Learning ..... 3841
Hyungwook Yim, Vladimir Sloutsky
Do walking pedestrians stabily interact inside a large group? Analysis of group and sub-group spatial structure ..... 3847
Francesco Zanlungo, Takayuki Kanda
Unstable Dynamics of Intrinsically Motivated Learning ..... 3853Arkady Zgonnikov, Ihor Lubashevsky
When to Hold and When to Fold: Detecting Structural Changes in Statistical Learning ..... 3858Benjamin Zinszer, Daniel Weiss
Member's Abstracts
Complexity Matching in Dyadic Interaction ..... 3864
Drew Abney, Alexandra Paxton, Chris Kello, Rick Dale
Measuring individual differences in cognitive biases ..... 3865Balazs Aczel, Bence Bago, Andrei Foldes, Aba Szollosi
The Effect of Test Format on Visual Recognition Memory Performance ..... 3866
Nora Andermane, Jeffrey Bowers
Belief Bias in the Perception of Sample Size Adequacy ..... 3867
Richard Anderson
Semantic interference in language production is due to graded similarity, not response relevance ..... 3868 Sabrina Aristei, Rasha Abdel Rahman
Experts in the discrimination task of the system noise increase avoidance response in the risk judgment ..... 3869
Yui Ashitani, Hyun-Min Ku, Akira Ishiguchi
Sleep deprivation accelerates delay-related loss of visual short-term memories without affecting precision ..... 3870
Christopher L. Asplund, Natalie Wee, Michael W.L. Chee
Learning Complex Scientific Concepts through Peer Argumentation: The Effect of Belief in Human Presence and Partner's Discourse Style ..... 3871
Christa Asterhan, Miriam Babichenko
Emotions in the space: arrangement and biases in preschool children and adults ..... 3872Maria Atanassova-Trifonova, Dimitar Atanasov, Pavleta Ognyanova
A Bayesian Analysis of Bias in Single Item Recognition for Continuous Word frequency ..... 3873 William Aue, Pernille Hemmer, Amy Criss
Exploring Cross-Situational Learning and Mutual Exclusivity ..... 3874Suzanne Aussems, Paul Vogt
Conditional reasoning with vertical spatial relations ..... 3875
Igor Bajšanski, Pavle Valerjev, Tanja Gulan
Does interaction matter? Testing whether fast and frugal heuristics can replace interaction in collective decision-making ..... 3876Dan Bang, Riccardo Fusaroli, Kristian Tylén, Karsten Olsen, Peter Latham, Jennifer Lau,Andreas Roepstorff, Geraint Rees, Chris Frith, Bahador Bahrami
An ACT-R model of the P300 event-related potential ..... 3877
Adrian Banks
Context-dependent memory effects on syntax ..... 3878
Brendan Barnwell
Linguistic differences in explanation requests and their effects on the evaluation of explanations: the case of English and Turkish ..... 3879
Melih Barsbey, Tania Lombrozo
Effects of orthography on second language phonology: The production of geminate consonants in speakers of English as a Second Language ..... 3880
Benedetta Bassetti
Aligning Behaviors in Everyday Environments: An Ecological Perspective ..... 3881
Lucas Bietti, Kasper Kok, Cienki Alan
Preschoolers understand the representational and communicative nature of iconic gestures ..... 3882
Simone Bijvoet-van den Berg, Ulf Liszkowski, Elena Hoicka
Modeling the genesis of a novel communicative system ..... 3883
Mark Blokpoel, Todd Wareham, Ivan Toni, Iris van Rooij
Metacognition in Decision-making - Influences of Signal Strength and Reliability ..... 3884
Annika Boldt, Vincent de Gardelle, Christopher Summerfield, Nick Yeung
Contextual Renewal and Awareness: Dissociating awareness within an electrodermal conditioning paradigm. ..... 3885
W. A. Bowditch, Rossy McLaren, A. McAndrew, Ian P. L. McLaren
Mechanisms of Active Causal Learning ..... 3886
Neil Bramley, David Lagnado, Maarten Speekenbrink
Collaborative coactivation in visual search ..... 3887
Allison Brennan, James Enns
What and When of Removing Virtual Gallbladders: Impact of Spatial and Temporal Aspects for Understanding Complex Diagrams ..... 3888
Angela Brunstein, Joerg Brunstein, Anam Waheed, Bakr Nour
Normative Dimensions of Generic Knowledge ..... 3889Dennis Bublitz, Sandeep Prasada
A hierarchical Bayesian model of conceptual knowledge transfer ..... 3890
Steven Buckingham, Alice Redfern, Christopher Summerfield
Acquiring Basic Skills in Changing Task Environments: Transfer of Working Memory Updating ..... 3891Trudy Buwalda, Aryn Pyke, Shawn Betts, Niels Taatgen, John Anderson
The representation of polysemous words in semantic and episodic memory ..... 3892
Álvaro Cabana, Camila Zugarramurdi, Juan Valle-Lisboa
Thinking in ways we don't speak: Evidence for a universal preference in semantic granularity ..... 3893Alexandra Carstensen, Grace Neveu, Lev Michael, Terry Regier
Coordinative Dynamic of Hierarchical Visual Processing ..... 3894Ramon Castillo, Heidi Kloos, John Holden, Michael Richardson
Language experience modifies the processing of visual input ..... 3895
Sarah Chabal, Viorica Marian
Understanding Emergent Phenomena with Functional Schema ..... 3896
Margaret Chan
Pedagogy and Technology: Balancing student-centric learning with teacher-centric teaching ..... 3897
Stephen Wee Hun Lim
Modelling Word and Object Naming in Pure Alexia ..... 3898
Ya-Ning Chang, Matthew Lambon Ralph, Steve Furber, Stephen Welbourne
Masked Repetition Effects of Chinese Compound Words Interact with Word Frequency because of
Properties of Compound ..... 3899
Sau-Chin Chen, Zi-Han Chen
Writing-related Visual-motor Abilities in Children with Developmental Coordination Disorder ..... 3900
Rong-Ju Cherng, Yu-Han Su, Ting-Hui Wang, Yung-Jung Chen
The construal level theory and the dual process theory in multi-attribute decision making: An empirical examination ..... 3901
Itsuki Chiba, Takashi Tsuzuki, Kikuchi Manabu
Beyond Berlyne's Conjecture: The Aesthetic Quality of Visual Patterns ..... 3902
Susan Chipman
Incremental Information Processing on Visual Search: The Critical Role of Delivery Rate ..... 3903 Eric Chiu, Michael Spivey
From Logical Positivism to Philosophical Thinking: Impacts and Implications on the Student-Learning Process ..... 3904
Dian Yi Chow, Stephen Wee Hun Lim
FOXP2 as the Genetic Basis for the Capacity to Shift Between Analytic and Associative Modes of Thought ..... 3905
Courtney Chrusch, Liane Gabora
The roles of configuration and orthography in Chinese recognition: a developmental approach ..... 3906
Yi-Ling Chung, Hsin-Hsuan Wu, Yu-Shu Chiang, Jon-Fan Hu, Chiu-Hua Huang,Hsueh-Chih Chen, Chien-Chih Tseng, Li-Yun Chang
Orthography and configuration on Chinese literacy acquisition: evidence from eye movement ..... 3907
Yi-Ling Chung, Yu-Shu Chiang, Hsin-Hsuan Wu, Jon-Fan Hu, Chiu-Hua Huang, Hsueh-Chih Chen, Chien-Chih Tseng, Li-Yun Chang
"You talkin' to me"? Understanding communicative intentions recruits the mirror and the mentalizing system ..... 3908
Angela Ciaramidaro, Cristina Becchio, Livia Colle, Bruno Bara, Henrik Walter
How does this thing work? Evaluating computational models of intervention-based causal learning3909Anna Coenen, Bob Rehder, Todd Gureckis
Individual differences in shape bias are predicted by non-linguistic perceptual ability ..... 3910
Beverly Collisson, Bernard Grela, Tammie Spaulding, Jay Rueckl, James Magnuson
Temporal motor dynamics in inductive reasoning ..... 3911
Steve Croker, Corinne Zimmerman
Where is the insight in insight problems? ..... 3912
Amory H. Danek, Michael Öllinger
Graphical Overviews Structure Online Learning: Evidence from Eye-Tracking ..... 3913
Sarah Davies, Kirsten Butcher, Anne Cook
Cultural Variation in Families' Shared Engagement ..... 3914Andrew Dayton, Barbara Rogoff
High level influences on visual action recognition ..... 3915Stephan de la Rosa, Stephan Streuber, Martin Giese, Heinrich Bülthoff, Cristobal Curio
Variables Influencing the Nature of Learned Categorical Perception Effects ..... 3916
Josh de Leeuw, Jan Andrews, Ken Livingston
Limits in Reasoning about False Beliefs in Adults: the Effect of Priming or the Curse of Knowledge? ..... 3917
Agnieszka Dębska, Krystyna Komorowska
Agency and attention influence the perceived speed of a moving stimulus ..... 3918
John Dewey, Thomas Carr
When does the causal information externally given affect causal inferences? ..... 3919
Kyung Soo Do, Rae-yeop Park
Do voices survive lexical consolidation? ..... 3920
Nicolas Dumay, Jeffrey Bowers
External Normalization: Testing a Cognitive Offloading Account ..... 3921Timothy Dunn, Srdan Medimorec, Evan Risko
Examining the transitions between decision strategies ..... 3922Gilles Dutilh, Benjamin Scheibehenne, Han L. J. van der Maas, Jörg Rieskamp
The Influence of Direct and Indirect Speech on Mental Representations ..... 3923
Anita Eerland, Jan Engelen, Rolf Zwaan
Explanation and Fractions: How Preferences for Types of Explanations Affect Learning ..... 3924
Emma H. Geller, James W. Stigler
Language and Spatial Attention: Words Evoke Independent Codes for Objects and Locations ..... 3925
Zachary Estes, Michelle Verges, James S. Adelman
The temporal dynamics of choice blindness: flat detection rates and short-term preference alterations ..... 3926Ilya Farber, Fumihiko Taya, Swati Gupta, O'Dhaniel Mullette-Gillman
Perceptual Word-Form Typicality Effects Are Modulated by Strength of Expectation During Reading ..... 3927
Thomas Farmer, Klinton Bicknell, Michael Tanenhaus
Towards identifying principles for clinical intervention in developmental language disorders:Establishing a neurocomputational foundation3928Anna Fedor, Wendy Best, Jackie Masterson, Michael S. C. Thomas
Effects of exhaustive and partial morphological segmentation vary with reading skill. ..... 3929Laurie Beth Feldman, Kit Cho, Petar Milin, Harald Baayen
Supporting Self-Regulated Learning in Conceptual Feedback Environments ..... 3930
Kirsten R. Butcher, Lisa A. Ferrara
Do Parents Adapt Descriptions of Spatial Relationships to Child Knowledge? ..... 3931Katrina Ferrara, Colin Wilson, Katherine Kelliher, Malena Silva, Barbara Landau
Using Graphical Organizers to Improve Self-Regulated Learning: Constructive Activities Enhance Deep Learning with Online Materials ..... 3932
Lisa A. Ferrara, Kirsten R. Butcher
Constant entropy rate and related hypotheses versus real language. ..... 3933
Ramon Ferrer-i-Cancho, Lukasz Debowski
The origins of analogy ..... 3934
Alissa Ferry, Susan Hespos, Yin-Juei Chang, Emily Hollenbeck, Dedre Gentner
Using Xbox Kinect to explore spatial-numerical association of arm movements in parity judgments. ..... 3935
Yariv Festman, Oliver Lindemann, Martin Fischer
From Complex to Collaborative Problem Solving ..... 3936
Andreas Fischer, Daniel Holt, Julia Hilse, Joachim Funke
Toddler's understanding of false beliefs about object identity ..... 3937
Ella Fizke, Stephen Butterfill, Hannes Rakoczy
Individual differences in 3D pointing performance between passengers and drivers ..... 3938Julia Frankenstein, Christoph Hölscher
Automatic classification of patients with mental disorders according to voice dynamics ..... 3939
Riccardo Fusaroli, Kristian Tylén, Arndis Simonsen, Ethan Weed
Conversation, Coupling and Complexity: Matching Scaling Laws Predict Performance in a Joint Decision Task ..... 3940
Kristian Tylén, Drew Abney, Bahador Bahrami, Christopher Kello, Riccardo Fusaroli
Completing the Puzzle: Online Processing during Novel Noun Generalization ..... 3941
Megan Galligan, Larissa Samuelson
Bilingualism change axonal structural network organization ..... 3942
Lorna Garcia Penton, Alejandro Perez Fernandez, Yasser Iturria-Medina, Manuel Carreiras
Infants' ability to discriminate between statements and questions ..... 3943Susan Geffen, Toben Mintz
Neural Correlates of Perceiving Dyadic Social Interactions ..... 3944Alexandra Georgescu, Bojana Kuzmanovic, Natacha Santos, Ralf Tepest, Gary Bente,Marc Tittgemeyer, Kai Vogeley
Imagined Effort Affects Object Localization and Sense of Agency ..... 3945
Devin Gill, J. Scott Jordan
Explorations in Human and Machine Learning of Decision Trees ..... 3946
Kevin Gluck, Jack Harris, Thomas Mielke, Christopher Myers, Vladislav Veksler
The Role of Conceptual and Perceptual Information on Inductive Reasoning in Early Childhood ..... 3947Karrie Godwin, Bryan Matlen, Anna Fisher
Perception and action effects on causal judgment ..... 3948
Kelly Goedert, Mengqi Guo
Effects of Brief Analogical Training after Two-Week Delay ..... 3949Micah Goldwater, Dedre Gentner
Interactions between bilingualism and non-linguistic spatial processing ..... 3950
Monica Gonzalez-Marquez, Raymond Becker, James Cutting
Conceptual Priming with Pictures and Environmental Sounds ..... 3951
Paula Goolkasian, Yongju Kim, Anne Marie Porter, Kathryn Weatherford
Taking Someone Else's Perspective: When Body "Position" is More Important than Body "Presence" ..... 3952
Michelle D. Greenwood, Justin L. Matthews, Michael J. Spivey, Teenie Matlock
Test Expectancy Effects on Metacomphrenesion, Self-regulation, and Learning ..... 3953
Thomas Griffin, Jennifer Wiley, Keith Thiede
Prisoner's Dilemma Games between Friends and Foes: Looking for Simpson's Paradox Effects in Cooperation ..... 3954
Maurice Grinberg, Stevan Tomic, Evgenia Hristova
Big Number Politics: The Effects of Training on Large Number Estimation in Fiscal Deficit-reducing Proposals. ..... 3955
Brian M. Guay, David Landy
Consistency in Degrees of Morality ..... 3956
Eoin Gubbins, Ruth Byrne, Phil Johnson-Laird
Creativity: Investigating Construct Validity across Cultures ..... 3957
C. Dominik Güss, Ma. Teresa Tuason, Yetuli de Baessa, Nila Shah, Angela Thomas
Exploration and Discovery in Children with Autism ..... 3958
Hyowon Gweon, Hannah Pelton, Rebecca Saxe, Laura Schulz
Inferring mass in complex physical scenes via probabilistic simulation ..... 3959
Jessica Hamrick, Peter Battaglia, Thomas Griffiths, Joshua Tenenbaum
Predictive processing of musical structure: effects of genre-specific expertise ..... 3960
Niels Chr. Hansen, Niels Chr. Hansen, Niels Chr. Hansen, Peter Vuust, Peter Vuust, Marcus Pearce
On the relationship between perceived foreigness, accentedness and speech comprehension under clear and adverse listening conditions ..... 3961
Adriana Hanulikova, Frank Eisner
Understanding the Role of Context in Memory for Maximally Counterintuitive Concepts ..... 3962Mary Harmon-Vukic, M. Afzal Upal, Caitlin Trainor
The effect of goal reification on error rates while solving geometric problems ..... 3963
Robert Hausmann, Annalies Vuong
Real-Time Strategy: Multi-Level Dynamics in an Uncertain Environment ..... 3964
Robert Hawkins, Robert Goldstone
Children's understanding of lying related to intention. ..... 3965
Hajimu Hayashi
Patient Knowledge and Satisfaction in Decisions About Lung Cancer Surgery ..... 3966
Josh A. Hemmerich, Mark K. Ferguson, Daniel P. Sulmasy, Arthur S. Elstein, Cindy Warnes, Masha Kochrginsky, Rita Gorawara-Bhat, Kellie Van Voorhis, William Dale
Baby physics: expectations about liquid and sand in 5-month-old infants ..... 3967
Susan Hespos, Alissa Ferry, Emily Hollenbeck, Lance Rips
The timing of turns in mother-infant interactions: A Longitudinal Study ..... 3968
Elma Hilbrink, Merideth Gattis, Elena Sakkalou, Kate Ellis-Davies, Stephen Levinson
Working Memory and Abstract Representation in the Context of Culture ..... 3969
Mark Ho, Fiery Cushman
The Atoms of Cognition: Action Learning and Problem Solving ..... 3970
Seng-Beng Ho
Measuring the Use of Recognition with the Multinomial r-Model: An Application to Aging and Individual Differences ..... 3971
Sebastian Horn, Thorsten Pachur, Rui Mata
Belief Updating in Moral Dilemmas ..... 3972Zach Horne, Derek Powell, Joseph Spino
The effects of caregivers' instruction on children's learning in easy and difficult tasks ..... 3973
Shumeng Hou, Zinnia Lam, Wing Chee So
Are beliefs biased by logic? The effect of a secondary load and complexity on belief and logic based judgments. ..... 3974
Stephanie Howarth, Simon Handley, Clare Walsh
SRNEngine: A Windows-based Neural Network Simulation Tool for the Non-programmer ..... 3975
Steven R. Howell, Suzanna Becker
Comparison and explanation in learning: Children's understanding of a basic engineering principle ..... 3976
Christian Hoyos, Dedre Gentner
The roles of L1 and markedness on Mandarin L2ers' construction of syntactic representation ..... 3977
Dong-Bo Hsu
Word Repetition Priming for Chinese Emotional Words ..... 3978Yu-Shu Chiang, Hsin-Hsuan Wu, Yi-Ling Chung, Hsueh-Chih Chen, Jon-Fan Hu
Categorization and recognition in a prototype-distortion task: ERP studies ..... 3979Yanli Huang, Zhiya Liu, Chi Shing Tse
Probabilistic negation: fine-grained preservation and distortion of truth in affirmative and negated statements ..... 3980
Stephanie Huette, Sarah Anderson, Michael Spivey
Exploring the boundary conditions of intentional forgetting ..... 3981
Almut Hupbach, Lili Sahakyan
Purity and Disgust: Same Domain, or Separate Mechanisms? ..... 3982
Gordon Ingram, Karolina Prochownik
How can an art course in photography inspire students' artistic creativity? ..... 3983
Chiaki Ishiguro, Takeshi Okada
Mental Process of Mindreading: Self and Stereotype as Anchors in Mental State Inference ..... 3984
Tatsunori Ishii, Masanori Takezawa
Eye-Movement Strategies in Multiple Target Search for Target Location and Uncertainty Reduction ..... 3985
Christian P. Janssen, Preeti Verghese
'But' how do children reason with it? ..... 3986
Leen Janssens, Stephanie Drooghmans, Sara Verbrugge, Walter Schaeken
Using SpAM to reveal decision processes and complex new context effects in similarity judgment ..... 3987
Gavin Jenkins, Larissa Samuelson, John Spencer
How Do People Judge Conjunctive Probabilities from Experience? A Hierarchical Bayesian Model Comparison ..... 3988
Mirjam Annina Jenny, Jörg Rieskamp, Håkan Nilsson
Social institutions as tools in normative cognition ..... 3989
Jeppe Sinding Jensen
Modeling spatial language acquisition as a function of lexical verb development ..... 3990
Kristen Johannes, Barbara Landau, Colin Wilson
Eye movements "steal" response time effect during imagery scanning ..... 3991
Roger Johansson, Jana Holsanova
Early Event-Related Potentials (ERPs) sensitive to animacy expectations in sentence comprehension are not overridden by context ..... 3992
Alexis R. Johns, Heather K. J. van der Lely, James S. Magnuson
Adult Category Learning Differences Predicted by a Dynamic Neural Field Theory Account of Information Sampled from the Fovea ..... 3993
Jordan Barnes, Mark Blair, Paul Tupper, R. Calen Walshe
Small Elephants and Big Needles: Can Perceptual Information Affect Memory and Judgments about the Meaning of Words? ..... 3994
Natalie A. Kacinik, Rita W. El-Haddad, Kendall J. Eskine, Lolly Starr-Glass, SamuelSalamon
Navigating the Social Environment: An Ecological Rationality Perspective on Advice Taking Behavior ..... 3995
Juliane Eva Kämmer, Hansjörg Neth, Pantelis Pipergias Analytis, Mehdi Moussaïd
Be a better multitasker. How a pause in the primary task can turn a rational into an irrational multitasker. ..... 3996
Ioanna Katidioti, Niels Taatgen
KANSEI Differences between the Child and the Caregiver in Room Arrangement Examined by the Room Arrangement Workshop for Child Room Spaces ..... 3997
Yoko Katsumata, Hideaki Kitazume, Tetsuya Yasuda, Harumi Kobayashi
The Influence of Observing Versus Imitating Gestures in L2 Phoneme Instruction ..... 3998Spencer Kelly, Yukari Hirata, Carmen Lin, Zach Zhao
Tracking Scrambled Word Order While Reasoning with Diagrams ..... 3999Özkan Kılıç
The role of Quinian bootstrapping in the acquisition of mental state terms ..... 4000
Szabolcs Kiss
An Explorative Study of Search of Model Space in Problem Solving ..... 4001
Saskia Kistner, Bruce D. Burns, Regina Vollmeyer, Ulrich Kortenkamp
Distributed Memory System Architecture based on the Analyses of Human Brain Memory ..... 4002
Muneo Kitajima, Makoto Toyota
Effects of verbalization on lie detection ..... 4003Sachiko Kiyokawa, Yoshimasa Ohmoto, Kazuhiro Ueda
Anisotropy of personal space examined in a virtual environment ..... 4004Takatsugu Kojima, Ayae Iwamoto, Masashi Sugimoto
Information flow across individuals in formation of symbol communication systems ..... 4005
Takeshi Konno, Junya Morita, Takashi Hashimoto
Are forgetting processes crucial to category learning? ..... 4006Lukasz Kopec, Brad C. Love
Words activate categorization even before the category forming task has been offered ..... 4007
Alexey Kotov, Tatyana Kotova, Elizaveta Vlasova
Influence of Ethnic Group-Membership and Gaze Direction on Emotion perception. An fMRI study. ..... 4008
Katharina Krämer, Gary Bente, Iva Barisic, Bojana Kuzmanovic, Kai Vogeley
Conceptual Art and Cognitive Science: A Case Study in Space ..... 4009
Alexander Kranjec
Efficiency of feature detection of visual stimuli influences the proportion judgments ..... 4010Hyun-Min Ku, Midori Tokita, Masami Ikeda, Akira Ishiguchi
The Effect of Set in the process of perception of unconscious information ..... 4011
Natalya Kudelkina
Navigator-driven placement of landmarks: effects on wayfinding performance in a virtual Tate Gallery. ..... 4012
Saskia Kuliga, Rul von Stülpnagel, Christoph Hölscher
Popout Attention with two foils: linear dependence and dimensional interaction ..... 4013
Kiran Kumar, Suyog Chandramouli, Richard Shiffrin
Using Relational Encoding to Promote Creative Problem Solving ..... 4014Kenneth J. Kurtz, Nuoya Zhang, Tamar Skolnick
Not Emergence as Emergence: Emergence in Artificial Intelligence and in Philosophy of Mind ..... 4015 Eliska Kvetova
Predictive coding and the Bayesian brain: Intractability hurdles that are yet to be overcome ..... 4016
Johan Kwisthout, Iris van Rooij
What working memory subcomponents are needed in the acquisition of survey knowledge? Evidences from direction estimation and shortcut tasks ..... 4017
Enia Labate, Francesca Pazzaglia, Mary Hegarty
Investigating Elementary and College Students' Development of Numerosity Judgment Using Eye-Tracking Technology ..... 4018
Meng-Lung Lai, Meng-Jung Tsai, Liang-Chen Chien, Yi-Chen Chen
Motor Resonance Effects Modulated by Perspective: A Role of Linguistic Contrual on Social Action Perception ..... 4019
Donghoon Lee, JaRang Kwak, YoungIn Lee
Psych Predicates: Subjectivity and Evidentiality ..... 4020
Chungmin Lee
Neural representations of language switching in early bilinguals: An fMRI study ..... 4021
Miaomei Lei, Hiroyuki Akama, Brian Murphy
Thinking metaphorically in Alzheimer's Disease: priming cognitive domains to understand inferential content of sentences ..... 4022
Jan Leite, Berla Moraes, Mabia Toscano, Danielly Lima
Activities of Mirror System Involved In Coordination Game ..... 4023Guanhong Li, Takashi Hashimoto, Jiro Okuda
A crosslinguistic look at how metaphor, manner, and aspect propel polititians through campaign races ..... 4024
Patricia Lichtenstein, Marcus Perlman, Teenie Matlock
What We Move to Moves Us: Biological Rhythmicity Predicts Musical Preferences ..... 4025Nicholas Jun Hao Tan, Sarah Shi Hui Wong, Stephen Wee Hun Lim
Scepticism: Genuine unbelief or implicit beliefs in the supernatural? ..... 4026
Marjaana Lindeman, Annika Svedholm
Modeling the ecological rationality of decision strategies based on internet statistics ..... 4027Daniela Link, Julian Marewski
The Roles of the Goal and Internal Models in Motor Representation: A Pragmatic Explanation . ..... 4028
Daniel Hsi-wen Liu
The interaction of the wrap-up effect and uncertainty factors in speech processing ..... 4029David J. Lobina, Josep Demestre, José E. García-Albea
The role of inflectional suffixes in lexical processing of Greek words ..... 4030
Sofia Loui, Athanassios Protopapas, Eleni Orfanidou
Using LEGO Robotics to Engage Minority Students in Science and Technology ..... 4031Carol M. Lu, John B. Black, Sorachai Kornkasem, Laura M. Lu
The Affect of Embodied Height in Perceptual Estimations ..... 4032
Jeremy A. Luno, Rick TillmanThe Less You Know, You Think You Know More; Dunning and Kruger effect in CollectiveDecision Making4033
Ali Mahmoodi, Majid Nili ahmadabadi, Bahador Bahrami
The role of semantic content of verbal categories in categorical perception: An ERP study ..... 4034
Martin Maier, Philipp Glage, Annette Hohlfeld, Rasha Abdel Rahman
Can we observe frequency-modulated syllable effects in French dyslexic children? ..... 4035
Norbert Maïonchi-Pino, Jean Écalle, Annie Magnan
Disentangling the cognitive processes underlying the testing effect ..... 4036
Simone Malejka, Edgar Erdfelder
Russian Validation Study of the International Affective Picture System ..... 4037
Olga Marchenko, Alexey Vasanov
Changes in information search strategy under "dense" hypothesis spaces ..... 4038
Doug Markant, Todd Gureckis
Analyzing Infant Attention, Interaction and Goals ..... 4039J. Douglas Mastin, Paul Vogt, Irene Claessens
Bridging the Gap Between Friends: How Presence Biases Distance Estimation ..... 4040
Justin L. Matthews, Teenie Matlock
Causal Dispositions and Transitivity in Causal Chains ..... 4041
Ralf Mayrhofer, Inesa Hildenbrand, Michael Waldmann
Cross-Linguistic Sound to Meaning Mappings in Spatial Relational Terms ..... 4042
Kelly McCormick, Lynne Nygaard
From symbols to analog magnitudes: A process model of fraction comparison, with fits to experimental data ..... 4043
Cameron Ross Lloyd McKenzie, James McClelland
Cue Competition in Human Associative Learning ..... 4044
Rossy McLaren, Fergal Jones, Fayme Yeates, Ian P. L. McLaren
Children's Production of Referring Expressions: Effects of Age and Visual Complexity on Under- and Over-Descriptions ..... 4045
Mhairi McMullon, Paul EngelhardtThe role of action prediction and control in young children's joint action coordination4046Marlene Meyer, Harold Bekkering, Rianne Haartsen, Janny C. Stapel, Sabine Hunnius
TMS provides evidence of effector-specific premotor contributions to action interpretation ..... 4047
John Michael
Running circles around symbol manipulation in trigonometry ..... 4048
Kevin W. Mickey, James L. McClelland
The Glamorgan Problem Solver (GLAM-PS): A Distributed and Embodied Production System Architecture for Cognition ..... 4049
Gareth Miles
The Influence of Selective Attention during Memory Retrieval on Subsequent Memory Performance in Young and Older Adults ..... 4050Jeremy K. Miller, Kyle L. Dixon, Tyler Young, Marianne Lloyd, Ashley Hartman
Nap-related consolidation in learning the grammar and vocabulary of a novel language ..... 4051
Jelena Mirkovic, Gareth Gaskell
How Teacher Support Children's Empathic Understandings toward Characters? -Focusing on the presentations of similar situations- ..... 4052
Satoko Miwa
The effects of the harmonization between a word's meaning and its expression style on implicit memory: Differences between typography in vision and prosody in sounds ..... 4053
Kozue Miyashiro, Etsuko T. Harada
Investigating phonological false memory effects of prior recall and recognition using the Remember-Know paradigm ..... 4054
Mohamed Shan-Rievan Mohamed Salleh, Winston Goh
Learning and task-dependent factors affect production choices ..... 4055
Jessica Montag, Maryellen MacDonald
Worry effects on visuospatial and verbal working memory ..... 4056
Andre Luiz Moreno, Juliana Avila deSouza, Gustavo Gauer
An fMRI pilot study: Altruistic Punishment in ingroup and outgroup membership ..... 4057
Rosalba Morese, Daniela Rabellino, Angela Ciaramidaro, Giovanna Carrara, Elena Prodi, Francesca M Bosco, Consuelo Valentini, Bruno G Bara
Changes in the Cognitive Control Network Associated with Adapting to Task Environment Modifications ..... 4058
Jarrod Moss, Winston Jones, Hao Bai, Stephanie Doane
Does the anticipation of a partner's reaction affect action planning? Spatial action-effect compatibility in a joint task ..... 4059
Romy Müller, Dietrich Kammer
Causal Sense-Making as Benefit in Foresight, Rather than Bias in Hindsight? ..... 4060
Edward Munnich, Jennifer Milazzo, Jade Stannard, Katherine Rainford
Reciprocal Ascription of Intentions Realized in Robot-human Interaction ..... 4061
Shoji Nagataki, Masayoshi Shibata, Takeshi Konno, Takashi Hashimoto, Hideki Ohira
When irrelevance matters! Effects of distractor-response binding in binary choices under uncertainty ..... 4062
Nadine Nett, Christian Frings
Multitasking Performance: Bound By Task Interference? ..... 4063Menno Nijboer, Niels Taatgen, Hedderik Van Rijn
Joint Action with Separate Response Sets for Location Relevant and Irrelevant Spatial Compatibility Tasks ..... 4064
Akio Nishimura, Chikashi Michimata
Temporal coordination patterns of performer and audience in vaudeville settings ..... 4065
Ryota Nomura, Takeshi Okada
Strategy flexibility in algebraic problem solving: The influence of feedback and learners' characteristics ..... 4066Daniela Nussbaumer, Michael Schneider, Elsbeth Stern
Extended Methods to Dynamically Estimate Emphasizing Points for Group Decision Making and the Evaluation ..... 4067
Yoshimasa Ohmoto, Misao Kataoka, Toyoaki Nishida
The effect of spatial layout of objects on special memory through free searching tasks ..... 4068
Kayoko Ohtsu
How expectation of context sharing with audience causes effects on writing arguments. ..... 4069
Ryosuke Onoda
Cognitive Residues of Similarity: After-Effects of Similarity Computations in Visual Search ..... 4070
Stephanie O'Toole, Mark T Keane
When is it rational to rely on heuristics? ..... 4071
Paula Parpart, Matt Jones, Bradley Love
"This could be a new approach against the flu..." - References to tentativeness of scientific knowledge and their impact on decision processes ..... 4072
Elisabeth Paus, Regina Jucks
Beyond the politeness tightrope: Message design for multiple social goals ..... 4073
David Pautler
Recognition of Common Object-Based Categories Found in Toddler's Everyday Object Naming Contexts ..... 4074
Alredo Pereira, Linda Smith
Complex functional brain network properties extracted from EEG reveal different processing patterns in late second-language learners as compared to native-speakers when performing a morphosyntactic task ..... 4075
Alejandro Pérez, Margaret Guillon Dowens, Nicola Molinaro, Yasser Iturria-Medina, Paulo Barraza, Manuel Carreiras
Vocal Charades: The Emergence of Conventions in Vocal Communication ..... 4076
Marcus Perlman, Rick Dale, Gary Lupyan
Perception of collaboration in joint musical performances ..... 4077
Ana Pesquita, Timothy Corlis, James T. Enns
I did not write you this before, did I? Destination Memory in Computer-Mediated Communication ..... 4078Stephanie Pieschl, Verena Karwinkel
Structural Alignment in Young Children's Shape Categorization: Different Roles for Learning from Comparisons and Contrasts ..... 4079
Raedy Ping, Micah Goldwater, Bryan Matlen, Linsey Smith, Susan Levine, Dedre Gentner
Fixation on Failure: Failing to Solve a Problem Hinders Subsequent Problem-Solving Ability ..... 4080Vencislav Popov
Effects of Gestures while Studying Language Animations ..... 4081
Lysanne Post, Tamara van Gog, Fred Paas, Rolf Zwaan
How Does Eye Gaze Affect Vocal Imitation in Autism? ..... 4082
Marie Postma-Nilsenová, Martijn Balsters, Mariska van Kastel
Semantic Intuitions in Statistical Causal Reasoning ..... 4083
Benjamin Quack, Ralf Mayrhofer, Michael Waldmann
Simulating the N400 ERP component as semantic network error: Insights from a feature-based connectionist attractor model of word meaning ..... 4084Milena Rabovsky, Ken McRae
Scaling-up Perception-Action Links: Evidence from Synchronization with Individual and Joint Action ..... 4085
Veronica Ramenzoni, Natalie Sebanz, Günther Knoblich
Cave-like Environments Facilitate Magical Thinking ..... 4086
Lillian Rigoli, Holley Moyes, Stephanie Huette, Michael Spivey
Approximating the Value Function in the Actor Critic Architecture using the Temporal Dynamics of Spiking Neural Networks ..... 4087
Jeffrey Rodny, David Noelle
Distractor frequency effects in language production: An ERP study ..... 4088
Lana Rohr, Rasha Abdel Rahman
Bayesian Word Learning in Multiple Languages ..... 4089
Sebastian Rolotti, Benjamin Zinszer, Elizabeth Carlson, Ping Li
Producing ambiguous messages ..... 4090
Sebastian Rose, Katharina Spalek, Rasha Abdel RahmanExplanatory Reasoning in Causal-based Categorization4091
Anselm Rothe, Ralf Mayrhofer
Learning in a different way: Interaction gestures influence category learning on multi-touch-tables ..... 4092
Susana Ruiz Fernández, Birgit Imhof, Julia Kranz, Stephan Schwan, Barbara Kaup, PeterGerjets
Social interaction and group dynamics of virtual cooperative agents ..... 4093
Alexei V. Samsonovich
Dimensional experience induces attention shifting in the dimensional change card sort (DCCS)task4094Larissa Samuelson, Sammy Perone, Stephen Molitor, Aaron Buss, John Spencer
Autonomy in Learning Sensorimotor Spaces with Dynamic Neural Fields ..... 4095
Yulia Sandamirskaya
Temporal discounting in a sequential search task ..... 4096
Ke Sang, Junyi Dai, Peter M. Todd, Robert L. Goldstone
Reducing Behaviorism and Cognitivism ..... 4097Ricardo Sanz
Modeling biconditionals: Equivalence in an uncertain world ..... 4098
Koji Sawa, Junki Yokokawa, Tatsuji Takahashi
Integrating Top-down and Bottom-up Approaches in Holistic Perceptual Categorization ..... 4099
Samer Schaat, Alexander Wendt, Dietmar Bruckner
Your mind wanders weakly, your mind wanders deeply: Objective measures reveal mindless reading at different levels ..... 4100
Daniel J. Schad, Antje Nuthmann, Ralf Engbert
Automatic processing of metaphor in right hemisphere lesion patients ..... 4101
Gwenda L Schmidt-Snoek, Marguerite McQuire, Anjan Chatterjee
Interacting with digital maps for indoor navigation ..... 4102
Verena Schnitzler
The conversational style of mothers interacting with their two- and five-year-old children ..... 4103
Tanja Schorch, Jens Brauer
Do pictures facilitate mental imagery during text processing? ..... 4104
Anne Schüler, Katharina Scheiter
We never walk alone: The influence of social interaction on wayfinding behavior ..... 4105
Sarah Schwarzkopf, Rul von Stülpnagel, Christoph Hölscher
Does the listener keeps spatial information from the speaker's cohesive gestures to comprehend subsequent sentences without gestures? ..... 4106
Kazuki Sekine, Sotaro Kita
The influence of spatial cueing on serial order visual memory ..... 4107
Rakesh Sengupta, Anvita Gopal, Prajit Basu, David Melcher, Raju Bapi
Physical Skill and Idea Interaction in the Creation of New Dance Movements ..... 4108Daichi Shimizu, Takeshi Okada
The effect of metacognitive strategies during reviewing ..... 4109
Hocheol Shin, Hongoak Yun, Jungsun Yoo, Kwangsu Cho
Foot Movement for Walking Biases the Visual Attention Allocation ..... 4110
Marie Shoda, Kazuhiko Yokosawa
CLEARPOND: Phonological and orthographic neighborhood information for multiple languages ..... 4111
Anthony Shook, Sarah Chabal, James Bartolotti, Viorica Marian
Thematic Music and Context Interact to Affect Inclusion of Schematic Elements During Story Generation ..... 4112Cynthia Sifonis
Embodied Mental Imagery ..... 4113Jan Frederik Sima
Collaborative Recall and Extended identity in Twitter communication ..... 4114Jurǵis Šķilters, Līva Brice, Krista Bērziņa, Uldis Bojārs, Monika Kreile, Marta Selecka
Probing the neural dynamics of visual working memory with dynamic fields and fMRI ..... 4115
John Spencer, Aaron Buss, Vincent Magnotta
Looking at an empty side of the world. Hemispheric specialisation in directing language-mediated eye movements. ..... 4116
Sara Spotorno, Jens Apel, Bérengère Plaza, Sylvane Faure
How common ground affects the use of gesture space ..... 4117
Kashmiri Stec, Mike Huiskes
The impact of Individual and Instructional Difference on Learning Measurement Concepts ..... 4118Nancy L Stein, Yuhtsuen Tzeng, Linda Siegel
Eye motion along a static concrete object: Why "fictive motion" is not a figurative mental phenomenon ..... 4119
Kurt Stocker
No adaptive strategy selection without outcome feedback ..... 4120
Hrvoje Stojic, Henrik Olsson
Judgments under Uncertainty: Bias, Experience, and Expectations ..... 4121
Chris N.H. Street, Daniel C. Richardson
An Attentionally Constrained Model of Statistical Word Learning ..... 4122Sumarga Suanda, Seth Foster, Linda Smith, Chen Yu
The time course of structural and number interference in sentence processing ..... 4123
Katja Suckow, Roger van Gompel
Imagining emotions: An ERP study on mental imagery of facial expressions ..... 4124
Franziska Suess, Rasha Abdel Rahman
Perspective preference and spatial learning: An eye movement investigation ..... 4125
Masashi Sugimoto, Takatsugu Kojima, Hiroyuki Tsuda, Shogo Kajimura, Ayae Iwamoto, Yuki Sato, Takashi Kusumi
Finding Structure in Space and Time: Active Trace of Patterns ..... 4126
Yanlong Sun, Hongbin Wang
Cognitive inhibition and the unbelievability of supernatural ideas ..... 4127Annika Svedholm, Marjaana Lindeman, Miika Leminen
Collaborative Memory Foraging in Categorical Recall Tasks ..... 4128
Janelle Szary, Christopher Kello, Theo Rhodes
Is Dishonesty an Automatic Tendency? ..... 4129
Maryam Tabatabaeian, Rick Dale, Nicholas D Duran
The searching effect of metaphor on text rereading: Difference by familiarity ..... 4130
Tomohiro TairaFormation of an art concept: How is visual information from photography utilized by the artist inconcept formation?4131Kikuko Takagi, Takeshi Okada, Sawako Yokochi
Bounded rationality leads to optimal decision-making and learning under uncertainty: Satisficing, prospect theory, and comparative valuation breaking the speed-accuracy tradeoff ..... 4132Tatsuji Takahashi, Kuratomo Oyo, Yu Kohno
Source-target mapping strategy in displaced communication: Communication strategy to represent absent objects ..... 4133
Kaori Tamura, Takashi Hashimoto
Conceptual Transformation in Origami ..... 4134Thora Tenbrink, Holly A. Taylor
The contribution of frequency, phonological and semantic factors in the development of the English past tense. ..... 4135
Anna Theakston, Grzegorz Krajewski, Sarah Keeble, Anna Woollams
Accommodating talker-specific phonetic detail: Influences on internal category structure ..... 4136Rachel Theodore, Emily Myers, Janice Lomibao
A Neuro-Computational Approach to Equivalence Formation in Arbitrary Categories ..... 4137
Angel E. Tovar, Gert Westermann
Coordination among edge team members during crisis management ..... 4138
Sebastien Tremblay, Isabelle Turcotte, Alexandre Labrecque, Jessica Desrochers-Pare, Vincent Rousseau
Visual landmarks and the visual impedance effect ..... 4139Kai Hamburger, Florian Roeser, Cate Marie Trillmich
Automatic and controlled processes of reasoning: insight from the matching hypothesis of syllogisms ..... 4140
Ping Ping Tse, Sergio Moreno-Rios, Juan Antonio Garcia-Madruga, Maria Teresa Bajo
Framing effects in perceptual decision-making ..... 4141Konstantinos Tsetsos, Christopher Summerfield
Differences between maximizers and satisficers in regret and counterfactual thinking during repeated versus switching decisions ..... 4142
Takashi Tsuzuki, Manabu Kikuchi, Itsuki Chiba
The effects of self-explanation on science concept learning for sixth grade students: Misconceptions matter ..... 4143Yuhtsuen Tzeng, Yu-Ping Huang, Nancy Stein
Do Japanese junior high school pupils really dislike mathematics? -Discrepancies between explicit and implicit assessment measures- ..... 4144
Akitoshi Uchida, Kazuo MoriAre you human? The engagement in real-time social interactions recruits the reward system .... 4145Ulrich Pfeiffer, Leonhard Schilbach, Bert Timmermans, Bojana Kuzmanovic, AlexandraGeorgescu, Gary Bente, Kai Vogeley
Differences in Interactional Attitudes in Second Language Conversations: From the Perspective of Expertise ..... 4146
Ichiro Umata, Seiichi Yamamoto, Kosuke Kabashima, Masafumi Nishida
Effects of Cultural Knowledge on Parallel Language Activation in German-English Bilinguals ..... 4147
Sebnem Uzuner, Henrike Blumenfeld, Marcel Giezen
Stimulating Retrieval During Reading to Improve Word Learning and Memory ..... 4148
Gesa S. E. van den Broek, Atsuko Takashima, Eliane Segers, Ludo Verhoeven
When do PDP neural networks learn localist representations? ..... 4149
Ivan Vankov, Jeffrey Bowers
Cognitive Improvement after Surgical Treatment of Moyamoya Disease ..... 4150
Thais Varzoni, C. Dominik Güss, Ricardo A. Hanel
Measuring the Degrees of Separation of a group of Minds ..... 4151
Tomas Veloz
Synchronizing to Learn and Like ..... 4152
Roy Vink, Anna Bosman, Toon Cillessen
The Influence of Motion Language and Gaze Orientation on Spatial Memory ..... 4153
David Vinson, Drew Abney, Teenie Matlock, Rick Dale
Social gaze orientation influences decision making ..... 4154
David Vinson, Rick Dale
Recognition Memory Processes in Novel Noun Generalization ..... 4155
Haley Vlach
Visitors from outer space: On picking up relevance in inductive reasoning ..... 4156
Wouter Voorspoels, Gert Storms
New methods for assessing ontological and representational changes in the conceptual system of children and adults ..... 4157
Stella Vosniadou, Despoina Lepenioti, Anna Choundala, Kalliopi Eikospendaki, Petros Papavasiliou, Emiliana Thanou
Feeling Guilt vs. Regret about Own Decisions: What are the Psychological and Neurobiological Differences? ..... 4158
Ullrich Wagner, Lisa Handke, Henrik Walter
Word Segmentation Without Statistics ..... 4159Felix Wang, Toben Mintz
Statistical Learning in Non-Chinese Speakers Exposed to a Sequence of Words without Spaces ..... 4160 Tsanyu Wang, Adam Li, Yeou-Teh Liu, Jenn-Yeu Chen
Closer than you think?: Options for efficiently approximating optimal analogies under Structure Mapping Theory ..... 4161
Todd Wareham, Tijl Grootswagers, Iris van Rooij
Robust social learning to babble in a spiking neural network model ..... 4162Anne S. Warlaumont, Sophia R. Hyatt
Emotion and Decision-Making: Modeling Strategy Selection ..... 4163Szymon Wichary, Tomasz Smolen
Television network attitudes toward political candidates implicit in lexical statistics ..... 4164 Jon Willits, Mark Seidenberg
The effect of accent on the perception of American and British English fluent speech in Chinese second language learners ..... 4165
Simpson WL Wong, Peggy Mok, Vina W.H. Leung, Kevin Kien-Hoa Chung, Bonnie Wing-Yin Chow, Michael CW Yip
The deficits of visual encoding processes in Chinese Dyslexics ..... 4166
Orieta H. Y. Wong, Michael C. W. Yip
Inflection from form versus meaning: Developmental and computational evidence ..... 4167
Anna Woollams, Joanna Moy, Sarah Keeble, Grzegorz Krajewski, Anna Theakston
Mnemonic convergence in a social network: ..... 4168
Jeremy Yamashiro, William Hirst
How to rectify the confirmation bias in Wason(1960)'s 2-4-6 task ..... 4169
Tomohito Yamazaki, Mutsumi Imai
Computational modeling of inference generation during reading comprehension ..... 4170
Meni Yeari, Paul van den Broek
Argument order as an expectation trigger in Korean ..... 4171
Hongoak Yun, Uphong Hong, Yun-Ju Nam, Hyunjung Kim
Social interaction and cognition: from L. Vygotsky and M. Bakhtin tradition to computer agents' interactivity ..... 4172
Vera Zabotkina, Elena Pozdnyakova, Artemy Kotov
Metacognitive Judgments in Category Learning ..... 4173
Valnea Zauhar, Igor Bajšanski, Drazen Domijan
Effects of Strategy and Rule Complexity on Multivariable Inductive Judgments ..... 4174
Corinne Zimmerman, Steve Croker
Difference in Single vs. Pair judgements on Deception Detection, Confidence and Bias based on the Level of Communication ..... 4175
Mircea Zloteanu, Daniel Richardson
Progressive Development of the Number Sense in a Deep Neural Network ..... 4176
Will Y. Zou, James L. McClelland
On the different types of information stored in the lexicon and their neural bases ..... 4177
Camila Zugarramurdi, Álvaro Cabana, Juan Valle-Lisboa
Author Index ..... 4178

## Dear Cognitive Scientists:

Welcome to the $35^{\text {th }}$ Annual Conference of the Cognitive Science Society! It is a great pleasure to welcome you at Humboldt-Universität zu Berlin. CogSci 2013's theme is "Cooperative Minds: Social Interaction and Group Dynamics." This theme reflects a rapidly growing interest within the cognitive science community in human sociality that has led to a move from the study of individual cognition to the social realm. The conference offers an exciting program with five plenary talks by distinguished speakers and three invited symposia with highly renowned contributors. But CogSci is also a platform for young scientists and you will likely be impressed by the high quality of research our junior colleagues will present at this conference.

The competition for oral presentations was extremely strong this year since we received more submissions than ever in the history of this conference. After the system was closed we counted 1246 submissions, including 910 full-papers, 269 member abstracts, 34 symposium proposals, 23 proposals for workshops/tutorials, and 10 publication-based talks. The quantity and - even more so - the high quality of the submissions indicate that our discipline is in good shape. After intensive reviewing we accepted 264 (28\%) of the full-paper submissions as oral presentations and $379(40 \%)$ as posters. We also accepted 247 ( $91 \%$ ) member abstracts as posters, 25 ( $73 \%$ ) symposia and 18 (82\%) workshops/tutorials. In a second round, we accepted an additional number of 77 (99 \%) member abstracts as posters. This large amount of presentations was possible only by adding parallel tracks and introducing a new program structure to accommodate more poster presentations. So: Welcome to the largest conference in the history of the Cognitive Science Society.

The outstanding interest in this year's conference might also be related to our venue. Berlin is a city with an eventful history and certainly one of the liveliest metropolises in Europe. It is a city of contrasts with a countless variety of world renowned cultural highlights and attractions that we hope you will enjoy exploring. Humboldt-Universität zu Berlin is one of Berlin's four universities. Until the Reunification of Germany in 1989, Humboldt-Universität was the most important University in East Germany; since then it has established itself as one of Germany’s leading research universities. The University is located in the heart of the city, in walking distance to many historical and contemporary sites of interest. The Brandenburg Gate, for instance, is just a few minutes away from the conference site.

We believe that a university-based conference in the urban center of Berlin will be a unique experience for all of us. Certain things might not be the way they usually are at a hotel-based conference. You may have to walk a bit more; some of the lecture rooms still have a historic aura from
the times when Helmholtz, Hegel, Einstein, and other famous scientists gave their lectures here so the seats may be a bit uncomfortable. And it might happen that not all participants will fit into the main lecture hall (Audimax) during the plenary sessions. The conference co-chairs initially planned to rent a non-university building in the neighborhood but it turned out that this would exceed the budget. So we decided to use the cinema (Kinosaal) right below the main lecture hall as an overflow area where the presentations can be seen on a video screen. The poster sessions will take place in the Maritim Hotel Berlin in the Friedrichstraße, one of the most liveliest streets in Berlin. It's just a ten minutes walk, but you may want to calculate your time to walk there generously since there is a lot to see on your way.

We are in the semester break at Humboldt-Universität but there will still be many students and faculty in the building. So you are invited to enjoy the specific atmosphere of a University that about two hundred years ago, in 1810, implemented Wilhelm von Humboldt's vision of a new type of university. Humboldt-Universität was the first to introduce the unity of research and teaching, to uphold the ideal of research without restrictions, and to provide an all-round education for its students. These principles of Wilhelm von Humboldt became general practice throughout the world. A new era of university and academic research had begun.

We hope you will enjoy this unique historic setting and wish you all an unforgettable conference at Humboldt-Universität as well as a great time in Berlin!

Berlin, Summer 2013
Markus Knauff, Michael Pauen, Natalie Sebanz, and Ipke Wachsmuth

## Acknowledgements

The conference chairs are deeply thankful to many people who helped us to make the conference a reality. First and foremost, we would like to thank the members of the Organizing Committee, the Program Committee, the prize judges, and all the many reviewers for their time and effort. The large number of submissions was a tremendous logistic challenge for all of us. We want to thank all those of you who were involved in the reviewing process for your thoughtful, comprehensive, and helpful reviews. We know many of you were deluged with requests for metareviewing or reviewing and had some trouble to find appropriate reviewers or to handle all requests for reviewing. Thank you for being so cooperative and patient. You can find the names of the program committee members on page 65 and all reviewers are listed in the conference proceedings.

We are also deeply thankful to many other people who helped us with the organization of the conference. Specifically, we want to thank Kai Hamburger for supporting the conference chairs during the submission and reviewing process and Anna Strasser for her unfailing support with all issues related to the local organization. We thank Andy Stull, the CSS conference officer, for his support with all business related issues and all matters having to do with the budget of the conference. We thank Deborah Gruber, the Cognitive Science Society Business Manager, for helping us with the bureaucracy, the Scarritt Group, Inc. for the logistics management, and James Stewart from Precision Conference Solutions for his prompt help to keep the reviewing website running. Finally, our thanks go to Humboldt-Universität and the Berlin School of Mind and Brain for hosting this conference and, last but not least, to our sponsors for giving us the possibility to support students from all over the world to attend the $35^{\text {th }}$ Annual Conference of the Cognitive Science Society.

# Organizing Committee CogSci 2013 

| Program Chairs: | Markus Knauff, Giessen University |
| :---: | :---: |
|  | Michael Pauen, Humboldt-Universität zu Berlin |
|  | Natalie Sebanz, Central European University |
|  | Ipke Wachsmuth, Bielefeld University |
| Assistance to Program Chairs: | Anna Strasser, Humboldt-Universität zu Berlin |
|  | Kai Hamburger, Giessen University |
| Sponsors Chair: | Holger Schultheis, University of Bremen |
| Awards Chairs: | David Peebles, University of Huddersfield |
|  | Richard P. Cooper, Birkbeck, University of London |
|  | Thora Tenbrink, Bangor University |
| Member Abstracts Chairs: | Paul Bello, Office of Naval Research |
|  | Cordula Vesper, Radboud University Nijmegen |
| Publicity Chair: | Ute Schmid, University of Bamberg |
| Student Volunteer Chairs: | Aidan Feeney, Queen's University Belfast |
|  | Martin Butz, University of Tübingen |
| Tutorials \& Workshops Chair: | Dongkyu Choi, University of Kansas |
| Symposia Chair: | Kai-Florian Richter, University of Melbourne |
| Publication-based Talks Chair: | Kai-Florian Richter, University of Melbourne |
| Web Chair: | Dongkyu Choi, University of Kansas |
| Conference Officer: | Andrew Stull, University of California Santa Barbara |

## Program Committee Members

| Erik Altmann | Danilo Fum | Daniel Montello |
| :---: | :---: | :---: |
| Marios Avraamides | Amelia Gangemi | Christopher Myers |
| Jerry Ball | Simon Garrod | Bernhard Nebel |
| Linden Ball | Dedre Gentner | Hansjoerg Neth |
| Thomas Barkowsky | Ashok Goel | Hartmut Neuf |
| Mike Barley | Rob Goldstone | David Noelle |
| Maren Bennewitz | Laura Gonnerman | Timothy Nokes-Malach |
| Sven Bertel | Cleotilde Gonzalez | Mike Oaksford |
| Brad Best | Geoffrey Goodwin | Marta Olivetti |
| Dorrit Billman | Stefan Th. Gries | Magda Osman |
| Stephen Blessing | Glenn Gunzelmann | John Pani |
| Jean- Francois Bonnefon | Ulrike Hahn | David Peebles |
| Henry Brighton | James Hampton | Marco Ragni |
| Andrew Brook | John-Dylan Haynes | Stephen Reed |
| Monica Bucciarelli | Mary Hegarty | Kai-Florian Richter |
| Bruce Burns | Evan Heit | Matthias Scheutz |
| Jerome Busemeyer | Sebastien Helie | Leonhard Schilbach |
| Cristina Cacciari | Joachim Hertzberg | Franz Schmalhofer |
| Angelo Cangelosi | Michael Jacobson | Ute Schmid |
| Daniel Casasanto | Christian Janssen | Mike Schoelles |
| Nicholas Cassimatis | Gary Jones | Holger Schultheis |
| Bill Clancey | Irvin Katz | Hajime Shirouzu |
| Catherine Clement | Barbara Kaup | Thomas Shultz |
| Rick Cooper | William Kennedy | Vladimir Sloutsky |
| Gary Cottrell | Sachiko Kiyokawa | David Stracuzzi |
| Anna Cox | Reinhold Kliegl | Ron Sun |
| Rick Dale | Alexander Klippel | Heike Tappe |
| David Danks | Ken Koedinger | Thora Tenbrink |
| Sidney D'Mello | Boicho Kokinov | Josh Tenenbaum |
| Scott Douglass | Stefan Kopp | Greg Trafton |
| Zachary Estes | David Landy | Barbara Tversky |
| Aidan Feeney | Brad Love | M. Afzal Upal |
| Evelyn Ferstl | George Luger | Leon Urbas |
| Ken Forbus | Lorenzo Magnani | Hedderik van Rijn |
| Mike Frank | Hanspeter Mallot | Iris van Rooij |
| Christian Freksa | Art Markman | Michael Waldmann |
| Bob French | Michael Matessa | Markus Werning |
| Daniel Freudenthal | Craig Miller | Sharon Wood |
| Wai-Tat Fu | Padraic Monaghan | Robert Youmans |

## CogSci2013 Reviewers

Abbott, Robert<br>Abe, Keiga<br>Addyman, Caspar<br>Ahmad, Sheeraz<br>Airenti, Gabriella<br>Aist, Gregory<br>Alderete, John<br>Aldrovandi, Silvio<br>Alfieri, Louis<br>Ali, Nadia<br>Alishahi, Afra<br>Allbritton, David<br>Allefeld, Carsten<br>Altman, Magda<br>Alzoubi, Omar<br>Anders, Silke<br>Andonova, Elena<br>Andrews, Janet<br>Andrighetto, Giulia<br>Anggoro, Florencia<br>Arriaga, Rosa<br>Ashley, Richard<br>Aslin, Richard<br>Athanasopoulos, Panos<br>Austerweil, Joseph<br>Azizi, Amir<br>Baillargeon, Renee<br>Bajo, Teresa<br>Bakdash, Jonathan<br>Baker, Chris<br>Baker, Ryan<br>Bambini, Valentina<br>Banks, Adrian<br>Bapi, Raju<br>Barber, Horacio<br>Bardone, Emanuele<br>Barenholtz, Elan<br>Barner, David<br>Baron, Andrew<br>Barr, Dale<br>Barry, Kevin<br>Bartolotti, James

Bassetti, Benedetta<br>Bateman, John<br>Baumann, Peter<br>Bauters, Merja<br>Bechtel, William<br>Beck, Melissa<br>Becker-Asano, Christian<br>Beckmann, Christoph<br>Beeman, Mark-Jung<br>Belardinelli, Anna<br>Belenky, Daniel<br>Bello, Paul<br>Ben-Asher, Noam<br>Bender, Andrea<br>Benitez, Viridiana<br>Bennardo, Giovanni<br>Benuzzi, Francesca<br>Bergen, Leon<br>Bergmann, Kirsten<br>Bernacki, Matthew<br>Bertenthal, Bennett<br>Berthouze, Luc<br>Berthouze, Nadia<br>Bertolotti, Tommaso<br>Bicknell, Klinton<br>Bier, Asmeret<br>Bilalic, Merim<br>Binti Abdullah, Nik Nailah Biro, Tamas<br>Blankenburg, Felix<br>Blokpoel, Mark<br>Bogart, Christopher<br>Bonato, Mario<br>Bonawitz, Elizabeth Baraff<br>Boot, Walter<br>Borovsky, Arielle<br>Borst, Gregoire<br>Borst, Jelmer<br>Bott, Lewis<br>Bottini, Roberto<br>Bowdle, Brian<br>Brennan, Susan

Breslow, Len
Briggs, Gordon
Bright, Aimee Kay
Briscoe, Erica
Brookshire, Geoffrey
Brown, Gordon
Brown, Norman
Brudzinski, Michel
Brumby, Duncan
Brunetti, Riccardo
Brunstein, Angela
Brunye, Tad
Buchanan, David
Bucher, Leandra
Buchsbaum, Daphna
Buckner, Cameron
Buechner, Simon J.
Buehner, Marc
Buhry, Laure
Bunger, Ann
Busch, Niko
Buss, Aaron
Butcher, Kirsten
Butler, Lucas
Butz, Martin
Bzdok, Danilo
Cade, Whitney
Canning, Cody
Carbone, Andrea
Cardella, Valentina
Carlson, Richard
Carter, Pelham
Casadio, Claudia
Casasola, Marianella
Casillas, Marisa
Cayton-Hodges, Gabrielle
Chan, Cheri
Chan, Joel
Chan, Margaret
Chang, Franklin
Chang, Maria
Chariker, Julia

| Chater, Nick | Cox, Charlie | Eichmann, Hanna |
| :---: | :---: | :---: |
| Chauvin, Alan | Cox, Gregory | Ellefson, Michelle |
| Chemero, Anthony | Cox, Michael | Emberson, Lauren |
| Chen, Dawn | Coyne, Emily | Engbert, Ralf |
| Chen, Jenn-Yeu | Creel, Sarah | Engelmann, Felix |
| Chen, Lang | Crick, Christopher | Engemann, Denis |
| Chen, Sau-chin | Cummins, Fred | Entin, Eileen |
| Chen, Stephanie | Cutica, Ilaria | Eren Kanat, Selda |
| Cheng, Sen | Cutting, James | Erk, Susanne |
| Chesney, Dana | Dahlbäck, Nils | Ernst, Udo |
| Chin-Parker, Seth | Dailey, Matthew | Eronen, Markus |
| Choi, Dongkyu | D'Argembeau, Arnaud | Esken, Frank |
| Christie, Stella | Dave, Kashmira | Evans, Vyv |
| Chrysikou, Evangelia | Davelaar, Eddy | Evens, Martha |
| Chu, Mingyuan | Davenport, Jodi | Fadjo, Cameron |
| Chuderski, Adam | Davies, Jim | Falzone, Alessandra Maria |
| Church, Ruth | Dawson, Colin | Fanti, Kostas |
| Cichy, Radoslaw | de Barbaro, Kaya | Farkas, Igor |
| Cipollini, Benjamin | de Boer, Miriam | Farmer, Thomas |
| Clapper, John | De La Cruz, Vivian | Fay, Nicolas |
| Clark, Andy | De Neys, Wim | Fazly, Afsaneh |
| Clark, Eve | De Palma, Paul | Feist, Michele |
| Clark, Herbert H | de Vos, Mark | Feng, Gary |
| Clark, Jason | Dehghani, Morteza | Ferlazzo, Fabio |
| Clark, Rosanne | Dell, Gary | Filliat, David |
| Claus, Berry | Delogu, Franco | Fine, Alex |
| Clausner, Timothy | Deng, Sophia | Fischer, Helen |
| Clifton, Charles | Denis, Michel | Fischer, Martin |
| Coen, Michael | Denison, Stephanie | Fisher, Anna |
| Cohnitz, Daniel | D'Errico, Francesca | Fisher, Brian |
| Colle, Livia | Dessalles, Jean-Louis | Flanagan, Robin |
| Colombo, Matteo | Destefano, Marc | Flecken, Monique |
| Coltekin, Arzu | Dewald, Andrew | Flombaum, Jonathan |
| Colzato, Lorenza | Diard, Julien | Foulsham, Tom |
| Conklin, Kathy | Do, Kyung Soo | Franco, Santiago |
| Connors, Michael | Donovan, Tim | Franconeri, Steve |
| Cook, Richard | Drolet, Matthis | Franco-Watkins, Ana |
| Corballis, Michael | Dube, Chad | Franklin, Nancy |
| Corballis, Paul | Dudschig, Carolin | Fraundorf, Scott |
| Corner, Adam | Dumas, Guillaume | Freiman, Mary |
| Corter, James | Duran, Nicholas | Friedman, Scott |
| Cosentino, Erica | Durkin, Kelley | Friedrich, Claudia |
| Couclelis, Helen | Eberhardt, Frederick | Friedrich, Jim |
| Cowell, Rosie | Egan, Suzanne | Frith, Chris |


| Frixione, Marcello |
| :---: |
| Frosch, Caren |
| Fua, Karl |
| Fujiki, Daisuke |
| Fusaroli, Riccardo |
| Gabora, Liane |
| Gadgil, Soniya |
| Gagliardi, Francesco |
| Gagne, Christina |
| Gahl, Susanne |
| Galantucci, Bruno |
| Galati, Alexia |
| Ganascia, Jean-Gabriel |
| Garcia Jurado, Raquel |
| Gattis, Merideth |
| Gauffroy, Caroline |
| Gayler, Ross |
| Geffner, Hector |
| GERBIER, Emilie |
| Gerken, LouAnn |
| Gerrig, Richard |
| Gerritsen, Charlotte |
| Gershman, Samuel |
| Gerstenberg, Tobias |
| Gianelli, Claudia |
| Giardini, Francesca |
| Giardino, Valeria |
| Giguere, Gyslain |
| Gijssels, Tom |
| Giorgio, Maurizio |
| Glady, Yannick |
| Glazek, Kuba |
| Glebkin, Vladimir |
| Gloeckner, Andreas |
| Gluck, Kevin |
| Goedert, Kelly |
| Goldwater, Micah |
| Gotts, Steve |
| Gould, Sandy |
| Graesser, Arthur |
| Graff, Peter |
| Gramann, Klaus |
| Graziano, Mario |
| Greenauer, Nathan |

Greene, Peter
Greiff, Samuel
Griffin, Harry
Griffiths, Tom
Grimm, Lisa
Grossman, Emily
Guan, Qun
Guha, Amal
Gureckis, Todd
Gust, Helmar
Gyselinck, Valerie
Habel, Christopher
Haertl, Holden
Halverson, Tim
Hammer, Barbara
Hansen, Michael
Harbison, J. Isaiah
Hare, Mary
Harris, Adam
Harris, Jack
Harris, Kevin
Harrison, Anthony
Hartshorne, Joshua
Haryu, Etsuko
Hausmann, Robert
Hayashi, Yugo
Hayes, Brett
Healey, Patrick
Hedberg, Nancy
Heim, Stefan
Heinzle, Jakob
Hekkanen, Steve
Helms, Michael
Hernández-Orallo, José
Hess, Ursula
Hettinger, Larry
Hiatt, Laura
Hicks, Jason
Hilbig, Ben
Hills, Thomas
Hilton, Denis
Hock, Michael
Hofmann, Jennifer
Hohenberger, Annette

Hoicka, Elena
Holmes, Kevin
Hölscher, Christoph
Holyoak, Keith
Honing, Henkjan
Hope, Ryan
Hörnig, Robin
Howes, Andrew
Hristova, Evgenia
Hristova, Penka
Hsiao, Janet
Huang, Chih-Mao
Hubscher, Roland
Huenefeldt, Thomas
Hund, Alycia
Hunt, Sally
Hupp, Julie
Huyck, Christian
Ippoliti, Emiliano
Israel, David
Jäkel, Frank
Jacobs, Robert
Jahn, Georg
Jameson, Jason
Jara-Ettinger, Julian Jarecki, Jana
Jarvstad, Andreas
Jasmin, Kyle
Jee, Benjamin
Jekat, Susanne
Jennett, Charlene
Jeuniaux, Patrick
Jimenez-Leal, William
Job, Remo
Jolij, Jacob
Jones, Lara
Jones, Matt
Jonsson, Martin
Jordan, Jerome Scott
Jordan, Kerry
Jordan, Kirsten
Kaastra, Linda
Kachergis, George
Kahnt, Thorsten

| Kalenine, Solene | Krems, Josef |
| :---: | :---: |
| Kaminski, Jennifer | Kriete, Trenton |
| Kanan, Christopher | Kropatsch, Walter |
| Kanayet, Frank | Krügel, André |
| Kanki, Barbara | Kuchenbrandt, Dieta |
| Kapatsinski, Vsevolod | Kucker, Sarah |
| Kaup, Barbara | Kuebler, Sandra |
| Kawamoto, Alan H. | Kuhlen, Anna |
| Kaznatcheev, Artem | Kühnberger, Kai-Uwe |
| Kearns, John | Kulms, Philipp |
| Keehner, Madeleine | Kumar, Shikhar |
| Kello, Christopher | Kunda, Maithilee |
| Kelly, Spencer | Kunene, Ramona |
| Kelty-Stephen, Damian | Kuo, Bo-cheng |
| Kempe, Vera | Kurtz, Kenneth |
| Kempter, Richard | Kuzmanovic, Bojana |
| Kershaw, Trina | Kwisthout, Johan |
| Kersten, Alan | Kwon, Mee-Kyoung |
| Khemlani, Sangeet | Lagnado, David |
| Khooshabeh, Peter | Laine, Tei |
| Khumalo, Langa | Laird, John |
| Kidd, Celeste | Lake, Brenden |
| Kim, Dahee | Lakkaraju, Kiran |
| Kinder, Annette | Lange, Elke |
| Kindsmüller, Martin Christof | Lany, Jill |
| Kiss, Szabolcs | Latif, Nida |
| Kissine, Mikhail | Laubrock, Jochen |
| Kita, Sotaro | Lawson, Chris |
| Kittredge, Audrey | Leake, David |
| Klatzky, Roberta | LeBreton, Mael |
| Klenk, Matt | Lee, N.Y. Louis |
| Kline, Melissa | Legare, Cristine |
| Kloos, Heidi | Lehman, Blair |
| Klügl, Franziska | Lemaire, Benoit |
| Knauff, Markus | Leonards, Ute |
| Knobe, Joshua | Levy, Simon |
| Knoeferle, Pia | Lewis, Mark |
| Knott, Lauren | Lewis, Molly |
| Kockler, Hanna | Li, Rui |
| Koenig, Melissa | Lim, Stephen |
| Konderak, Piotr | Lindley, Craig |
| Konstantinou, Nikos | Linhares, Alexandre |
| Körner, Christof | Linkenauger, Sally |
| Krahmer, Emiel | Little, Daniel |

Liu, Lei
Liu, Ou Lydia
Lockwood, Kate Logan, John
Lombrozo, Tania Long, Lyle
Longo, Matthew
Lonsdale, Deryle Loomis, Jack Los, Sander
Louwerse, Max
Lovett, Andrew
Lowenthal, Francis
Lucas, Christopher Lucero, Ché
Luhmann, Christian Lyons, Ian Lyre, Holger
MacDonald, Maryellen
Mack, Michael
Magerko, Brian
Magliano, Joe
Maguire, Phil
Majerus, Steve
Malave, Vicente
Manalo, Emmanuel
Manera, Valeria
Mao, Wenji
Marcovitch, Stuart
Marewski, Julian
Markant, Doug
Martin, Robert
Martinovic, Goran
Mason, Lucia
Massera, Gianluca
Matsuda, Noboru
Matsuka, Toshihiko
Matuk, Camillia
Matzen, Laura
Maurits, Luke
Mawby, Ronald
Maye, Alexander
Mayrhofer, Ralf
Mazzone, Marco

McAuley, Devin
McCormack, Teresa
McDonnell, John
McGreggor, Keith
McLure, Matthew
McShane, Marjorie
Mehlhorn, Katja
Meijering, Ben
Meilinger, Tobias
Mello, Catherine
Melnick, Robin
Mendonca, David
Meneghetti, Chiara
Mettler, Everett
Metzinger, Thomas
Minelli, Alessandro
Mirman, Daniel
Mislove, Alan
Mitchell, Robert
Moeller, Korbinian
Molinaro, Nicola
Moratz, Reinhard
Moreno-Ríos, Sergio
Morgan, Brent
Morgan, Emily
Morita, Junya
Morsanyi, Kinga
Moss, Jarrod
Moulin, Chris
Muentener, Paul
Mulatti, Claudio
Munnich, Edward
Muntanyola, Dafne
Münzer, Stefan
Murphy, Gregory
Musholt, Kristina
Mutafchieva, Milena
Nagai, Yukari
Nagel, Jonas
Nakamura, Kuninori
Neokleous, Kleanthis
Nersessian, Nancy
Newcombe, Nora
Newell, Ben

Newsam, Shawn
Nilsenova, Marie
Nüssli, Marc-Antoine
Oates, Tim
Oberauer, Klaus
Obrecht, Natalie
Odgaard, Eric
O'Donnell, Timothy
Ohlsson, Stellan
Olney, Andrew
Oltramari, Alessandro
Oppenheimer, Danny
Orenes, Isabel
Ormel, Ellen
O'Seaghdha, Padraig
Osherson, Daniel
Oshima, Jun
Ostendorf, Florian
Osth, Adam
Otto, Ross
Ouyang, Long
Pachur, Thorsten
Pado, Ulrike
Pajak, Bozena
Palmiero, Massimiliano
Panagiotou, Georgia
Pantelis, Peter
Papagno, Costanza
Paradowski, Michal B.
Park, Woosuk
Pauen, Michael
Pautler, David
Pauwels, Pieter
Pavlik, Philip
Pazzaglia, Francesca
Pearl, Lisa
Pearlmutter, Neal
Pecher, Diane
Pelachaud, Catherine
Penner-Wilger, Marcie
Pennisi, Antonino
Pereira, Alfredo F.
Peressotti, Francesca
Perfors, Amy

Perruchet, Pierre
Person, Natalie
Pesciarelli, Francesca Pessa, Eliano
Petersen, Kathrine
Petersen, Wiebke
Peterson, Eric
Petkov, Georgi
Pezzulo, Giovanni
Pfeiffer, David
Pfeiffer, Ulrich
Phillips, Brenda
Pietarinen, Ahti
Pievani, Dietelmo
Pino, Bernardo
Pipa, Gordon
Pipergias Analytis, Pantelis
Pirog Revill, Kathleen
Pitt, Benjamin
Pompe, Ulrike
Pontier, Matthijs
Portrat, Sophie
Pothos, Emmanuel
Prabhakar, Janani
Prasada, Sandeep
Protopapas, Athanassios
Putnam, Cynthia
Putzar, Lisa
Pyers, Jennie
Pyka, Martin
Pynadath, David
Queen, Jennifer
Quinn, Paul
Quirin, Markus
Raczaszek-Leonardi, Joanna
Ralph, Jason
Rami, Gabriel
Ramscar, Michael
Read, Stephen
Redeker, Gisela
Rein, Jonathan
Reiss, Jason
Reiter, Dorothea
Reitter, David

| Renkl, Alexander |
| :---: |
| Reuter, Kevin |
| Reverberi, Carlo |
| Revlin, Russell |
| Rhodes, Rebecca |
| Rhodes, Theo |
| Richardson, Daniel |
| Richardson, Mike |
| Richey, Elizabeth |
| Riesbeck, Chris |
| Riggio, Lucia |
| Rips, Lance |
| Ritter, Steven |
| Robering, Klaus |
| Robledo, Marybel |
| Roche, Jennifer |
| Rodgers, Stuart |
| Rohde, Hannah |
| Rohlfing, Katharina |
| Rolison, Jonathan |
| Roll, Ido |
| Rosenbloom, Paul |
| Rothe, Annelie |
| Rothganger, Fred |
| Rottman, Benjamin |
| Rowat, Peter |
| Rowland, Caroline |
| Roy, Brandon |
| Rubichi, Sandro |
| Saalbach, Henrik |
| Sadakata, Makiko |
| Sadeghipour, Amir |
| Sakamoto, Maki |
| Sakamoto, Yasuaki |
| Salay, Nancy |
| Salden, Ron |
| Sanborn, Adam |
| Saner, Lelyn |
| Santamaria, Amy |
| Santiago, Julio |
| Sanz, Ricardo |
| Saraschandra, Karanam |
| Scassellati, Brian |
| Schachner, Adena |


| Scheepers, Christoph | Southwood, Frenette |
| :---: | :---: |
| Scheutz, Matthias | Spanoudis, George |
| Schlobinski, Peter | Speed, Ann |
| Schloegl, Christian | Speekenbrink, Maarten |
| Schnepp, Jerry | Spenader, Jennifer |
| Schommer, Christoph | Sperber, Dan |
| Schooler, Lael | Spiering, Brian |
| Schreiber, Susanne | Spivey, Michael |
| Schubotz, Ricarda | Spranger, Michael |
| Schultz, Johannes | Sprenger, Simone |
| Schurger, Aaron | Spychalska, Maria |
| Schurz, Gerhard | Sreekumar, Vishnu |
| Schwering, Angela | Stachnis, Cyrill |
| Scott, Kimberly | Stadtler, Marc |
| Self, Matt | Stafura, Joe |
| Serrano, J. Ignacio | Steedman, Mark |
| Settle, Amber | Stefanucci, Jeanine |
| Shafto, Carissa L. | Stein, Courtney |
| Shafto, Meredith | Stein, Nancy L. |
| Shafto, Michael | Stent, Amanda |
| Shafto, Patrick | Stewart, Neil |
| Shah, Priti | Stewart, Terrence |
| Shelton, Michael | Steyvers, Mark |
| Shepherd, Stephen | Stirling, Lesley |
| Shimi, Andria | Stjernfelt, Frederik |
| Shimizu, Yuki | Stocco, Andrea |
| Shimojima, Atsushi | Stolterfoth, Britta |
| Shook, Anthony | Strain, Amber |
| Si, Mei | Street, Christopher |
| Siebers, Michael | Stromsten, Sean |
| Siegel, Max | Sullivan, Jess |
| Sieroka, Norman | Sutton, John |
| Silbert, Noah | Suzuki, Hiroaki |
| Sima, Jan Frederik | Tabor, Whitney |
| Simione, Luca | Tack, Werner |
| Simms, Nina | Takema-Jimenez, Ana |
| Sims, Chris | Tamburelli, Maroro |
| Sims, Clare | Tamietto, Marco |
| Sinnett, Scott | Tanaka, Daisuke Joshua |
| Slessor, Gillian | Tanaka, James |
| Smith, Kenny | Taylor, Eric |
| Smith, Linsey | Smith, Nathaniel |
| Sobel, David |  |
|  |  |
| Smill |  |


| Taylor, Julia | Vankov, Ivan | Wölfl, Stefan |
| :---: | :---: | :---: |
| Teller, Virginia | vanMarle, Kristy | Worthy, Darrell |
| ten Bosch, Louis | Varakin, Donald Alexander | Wu, Anna |
| Terai, Hitoshi | Vatakis, Argiro | Wu, Rachel |
| Thibaut, Jean-Pierre | Veale, Richard | Xiao, Danqing |
| Thompson, Clarissa | Veksler, Bella | Xu, Yang |
| Thompson, Dominic | Veksler, Vladislav | Yaghoubzadeh, Ramin |
| Thompson, Valerie | Verheyen, Steven | Yanowitz, Karen |
| Timmis, Matthew | Vesper, Cordula | Yap, Defu |
| Tinti, Carla | Vespignani, Francesco | Yates, Mark |
| Tirassa, Maurizio | Vinson, David | Yeap, Wai |
| Toivonen, Ida | Visser, Ingmar | Yechiam, Eldad |
| Tomai, Emmett | Vogeley, Kai | Yi, Sheng Kung |
| Tomko, Martin | Voiklis, John | Yip, Michael |
| Tomlinson, John | Vollmeyer, Regina | Yoshimi, Jeffrey |
| Tomuro, Noriko | Von der Malsburg, Titus | Young, Joseph |
| Tong, Matthew | von Stülpnagel, Rul | Yu, Alfred |
| Toscano, Joseph | von Sydow, Momme | Yu, Angela |
| Towle, Brendon | Vosgerau, Gottfried | Yun, Jiwon |
| Trémolière, Bastien | Vosoughi, Soroush | Yurovsky, Dan |
| Trickett, Susan | Votsis, Ioannis | Yurovsky, Daniel |
| Trueblood, Jennifer | Vul, Ed | Zachariou, Valentinos |
| Trukenbrod, Hans | Wade, Kim | Zacks, Jeff |
| Tsakiris, Manos | Walker, Erin | Zapata, Diego |
| Tso, Ricky Van Yip | Walsh, Matthew | Zarcone, Alessandra |
| Tsuchida, Tomoki | Warlaumont, Anne | Zatorre, Robert |
| Tummolini, Luca | Warren, Caleb | Zednik, Carlos |
| Tupper, Paul | Warwick, Walter | Zeijlstra, Hedde |
| Twomey, Katherine | Watson, Duane | Zeller, Jochen |
| Tzelgov, Joseph | Webber, Frederick | Zetsche, Christoph |
| Tzneg, Yuhtsuen | Weber, Andrea | Zheng, Robert |
| Uengoer, Metin | Welsh, Matthew | Zimmerman, Corinne |
| Ullah, Shoukat | Westermann, Gert | Zipitria, Iraide |
| Ullman, Tomer | Weygandt, Martin | Zompi, Simona |
| Ulrich, Rolf | White, Sarah | Zourou, Filio |
| Urminsky, Oleg | Wiegmann, Alex | Zuidema, Willem |
| Utsumi, Akira | Wierda, Stefan |  |
| Vallee-Tourangeau, Frederic van der Wal, Natalie | Wilkinson, Meredith Windt, Jennifer |  |
| van der Wel, Robrecht | Winkelholz, Carsten |  |
| van Riel, Raphael | Witt, Jessica |  |
| van Rooy, Bertus | Wittmann, Bianca |  |
| van Vugt, Marieke | Wojtowicz, Ralph |  |
| Van Welbergen, Herwin | Wolff, Phillip |  |

## Sponsors

We sincerely thank the sponsors of the 35th Annual Meeting of the Cognitive Science Society for their support of the conference awards and the tutorials, and for supporting student participation through student travel grants.

The Robert J. Glushko and Pamela Samuelson Foundation

German Society for Cognitive Science (GK)

## Deutsche

Forschungsgemeinschaft (DFG)

National Science Foundation (NSF)

Humboldt-Universität zu
Berlin

European Network for the Advancement of Artificial Cognitive Systems, Interaction and Robotics (EUCog)

Artificial Intelligence Journal (AIJ)

DFG Priority Programme "New Frameworks of Rationality"

Wiley-Blackwell

Berlin School of Mind and
Brain

## CogSci 2013 Awards

## Marr Prize

The Marr Prize, named in honor of the late David Marr, is awarded to the best student paper at the conference. All student first authors were eligible for the Marr Prize for the best student paper. The Marr Prize includes an honorarium of $\$ 1,000$ and is sponsored by The Cognitive Science Society. The winner of the 2013 Marr Prize for Best Student Paper is:

Nimrod Dorfman, Daniel Harari, and Shimon Ullman: Learning to perceive coherent objects (Thursday, 16:50, Session 31)

## Computational Modeling Prizes

Four prizes worth $\$ 1,000$ each are awarded for the best full paper submissions to CogSci 2013 that involve computational cognitive modeling. The four prizes represent the best modeling work in the areas of perception/action, language, higher-level cognition, and applied cognition. These prizes are all sponsored by The Cognitive Science Society. The winners of the 2013 Computational Modeling Prizes are:

## Perception/Action

Georg Layher, Martin Giese, and Heiko Neumann: Learning representations of animated motion sequences - A neural model (Saturday, 16:30, Session 95)

## Language

Russell Richie, Charles Yang, and Marie Coppola: Modeling the emergence of lexicons in homesign systems (Thursday, 16:50, Session 29)

## Higher-Level Cognition

John V McDonnell, Pedro Tsividis, and Bob Rehder: Reasoning with inconsistent causal beliefs (Thursday, 16:50, Session 32)

## Applied Cognition

Mohammad Khajah, Robert Lindsey, and Michael Mozer: Maximizing students' retention via spaced review: Practical guidance from computational models of memory (Saturday, 15:30, Session 76)

## NSF Funded Conference Grants

In association with the Cognitive Science Society, the US National Science Foundation (NSF) has funded four conference grants to US citizens who are enrolled as students at a US institution to attend the $35^{\text {th }}$ Annual Conference of the Cognitive Science Society (CogSci2013). The awardees were:

Deanne Adams, David Braithwaite, Heather Burte, and Richard Veale

## Student Travel Awards

Travel awards have been provided to students whose submissions were accepted as full papers with the highest reviewer rankings, and who indicated a need for travel funding. A total of 38 students were selected to receive a student travel award.

## Robert J. Glushko and Pamela Samuelson Foundation

The Robert J. Glushko and Pamela Samuelson Foundation generously sponsored $\$ 10,000$ for student travel awards. The 2013 Travel Awards went to:
Ahmad Azad Ab Rashid (Cardiff University, UK)
Daniela Ahlberg (Eberhard Karls Universität Tübingen, Germany)
Shira Calamaro (Yale University, USA)
Benjamin Cipollini (University of California, San Diego, USA)
Roeland Hancock (University of Arizona, USA)
Ryan Hope (Rensselaer Polytechnic Institute, USA)
Chigusa Kurumada (Stanford University, USA)
Marcus Lindskog (Uppsala University, Sweden)
Ross Macdonald (University of Dundee, UK)
MaryLauren Malone (University of Cincinnati, USA)
Tyler Marghetis (University of California, San Diego, USA)
Philip Pärnamets (Lund University, Sweden)
Alexandra Paxton (University of California, Merced, USA)
Eliane Stampfer (Carnegie Mellon University, USA)
Eoin Travers (Queen's University Belfast, UK)
Andrew Whalen (University of California, Berkeley, USA)

## GK - German Society for Cognitive Science

The German Society for Cognitive Science (GK) generously sponsored $3,000 €$ for student travel awards plus a one-year complementary subscription to GK. The 2013 Travel Awards went to:
Eric Arnau (Universitat Autònoma de Barcelona, Spain)
Martin Garcia (Rheinische Friedrich-Wilhems-Universität Bonn, Germany)
Mona Guath (Uppsala University, Sweden)
Christina Meier (University of Exeter, UK)
Falk Lieder (ETH Zurich, Switzerland)
Andrea Ravignani (University of Vienna, Austria)

## EUCog - European Network for the Advancement of Artificial Cognitive Systems, Interaction and Robotics

The European Network for the Advancement of Artificial Cognitive Systems, Interaction and Robotics (EUCog) generously sponsored $4,000 €$ for student travel awards. The 2013 Travel Awards went to:

Ulku Arslan Aydin (Capital Markets Board of Turkey)
Rasmus Bååth (Lund University Cognitive Science, Sweden)
Meadhbh Foster (University College Dublin, Ireland)
Eric Hunsberger (University of Waterloo, Canada)
Matthew J. Kmiecik (Loyola University Chicago, USA)
Yevgen Matusevych (Tilburg University, The Netherlands)
Daniel Rasmussen (University of Waterloo, Canada)
R. Calen Walshe (University of Edinburgh, UK)

## AIJ - Artificial Intelligence Journal

The Artificial Intelligence Journal (AIJ) generously sponsored 4,000 € for student travel awards. The 2013 Travel Awards went to:
Trevor Bekolay (University of Waterloo, Canada)
Chen, Dawn (University of California, Los Angeles, USA)
Choo, Feng-Xuan (University of Waterloo, Canada)
Tim Chuk (University of Hong Kong)
Arindam Das (York University, Toronto, Canada)
Molly Lewis (Stanford University, USA)
Russell Richie (University of Connecticut, USA)
Andrew Saxe (Stanford University, USA)

## Awards Committee

Richard P. Cooper (co-chair), David Peebles (co-chair), Thora Tenbrink (co-chair), Erik Altmann, Berit Brogaard, Jerome R. Busemeyer, Nancy Cooke, Sharon Goldwater, Stefan Kopp, Stephan Lewandowsky, Max M. Louwerse, Brad Love, Padraic Monaghan, Richard Shiffrin, David Uttal, Rineke Verbrugge.

## Robert J. Glushko Dissertation Prizes

The Cognitive Science Society and the Glushko-Samuelson Foundation will award up to five outstanding dissertation prizes in cognitive science each year. The goals of these prizes are to increase the prominence of cognitive science, and encourage students to engage in interdisciplinary efforts to understand minds and intelligent systems. The hope is that the prizes will recognize and honor young researchers conducting ground-breaking research in cognitive science. The eventual goal is to aid in efforts to bridge between the areas of cognitive science and create theories of general interest to the multiple fields concerned with scientifically understanding the nature of minds and intelligent systems. Promoting a unified cognitive science is consistent with the belief that understanding how minds work will require the synthesis of many different empirical methods, formal tools, and analytic theories. 2011 was the inaugural year of this prize, and a new competition is held annually.

## Robert J. Glushko Dissertation Prize Recipients

The 2013 recipients of the Robert J. Glushko Prizes for Outstanding Doctoral Dissertations / theses in Cognitive Science are:

Dr. Douglas Knox Bemis - 2012 PhD thesis "Simple Composition During Language Processing: An MEG Investigation" from New York University

Dr. Neil Cohn - 2012 PhD thesis "Structure, Meaning, and Constituency in Visual Narrative Comprehension" from Tufts University

Dr. George Kachergis - 2012 PhD thesis "Mechanisms for Cross-Situational Learning of WordReferent Mappings: Empirical and Modeling Evidence" from Indiana University

Dr. Andrew Lovett - 2012 PhD thesis "Spatial Routines for Sketches: A Framework for Modeling Spatial Problem Solving" from Northwestern University

Dr. Liad Mudrik - 2011 PhD thesis "Processing Visual Context Violations: The Roles of Attention and Awareness" from Tel Aviv University
http://www.cognitivesciencesociety.org/about awards glushko recipients.html

A special Glushko Dissertation Prize Symposium showcases the PhD research projects of the first four prize recipients listed above (Saturday, 10:30-12:10, Session 68).

## Invited Plenary Presentations

## Heineken Prize Lecture

A Core Brain System in Assembly of Cognitive Episodes
John Duncan, Cambridge (introduced by Michael Tomasello)
Thursday August 1st, 09:00

## Rumelhart Prize Lecture

It's all Connected in Developmental Process: The Body, the Statistics, Visual Object Recognition, and Word Learning
Linda Smith, Indiana University (introduced by Peter Dayan)
Friday August 2nd, 18:00

## Keynote Talks

Cooperative Machines: Coordinating Minds and Bodies Between People and Social Robots Cynthia Breazeal, MIT (introduced by Ipke Wachsmuth)
Thursday August 1st, 18:00
Shared Agency
Michael E. Bratman, Stanford (introduced by Natalie Sebanz)
Friday August 2nd, 09:00
Core Social Cognition
Elizabeth Spelke, Harvard (introduced by Michael Pauen)
Saturday August 3rd, 09:00

# Diagrammatic Cognition: Discovery and Design 

## William Bechtel (bechtel@ucsd.edu)

Department of Philosophy and Interdisciplinary Program in Cognitive Science, University of California, San Diego, La Jolla, CA 92093-0119 USA

Keywords: diagrams, visuospatial representations; external representations; problem solving; diagrammatic reasoning; scientific reasoning; cognitive tools; design

## Introduction

External representations of thought-maps, diagrams, sketches, and the like-are ancient inventions that serve thought and communication in numerous ways. A number of cognitive scientists have investigated roles these representations play in cognition (see, e.g., Donald, 1991; Larkin \& Simon, 1987; Norman, 1993; Schön, 1983). They are created and used by school students, by architects and designers, by mathematicians and scientists, by musicians, dancers, and artists. People design and use diagrams to spatialize thought and make it public, to work through ideas and clarify thinking, to reduce working memory load, to communicate ideas to others, to promote collaborative work by providing an external representation that can be pointed to and animated by gestures and collectively revised. Considerable research has shown that well-designed diagrams promote thought, creativity, discovery, and communication. Diagrams can map abstract thought to space, allowing spatial reasoning to promote abstract reasoning.

Just as many diverse groups create and use diagrams, many diverse groups are actively studying their creation and use. Educators study ways to design effective diagrams and ways to educate students to use them. Psychologists study how diagrams are perceived, comprehended, and created. Both educators and psychologists study ways to promote the spatial thinking skills underlying comprehension and creation of diagrams. Designers study their use in design, artists their use in art. Historians and philosophers of science describe case studies of the insightful development of diagrams by scientists and the insights those diagrams have provided to others. Philosophers analyze formal properties of diagrams. Mathematicians explore the diagrammatic thinking that underlies mathematical thought and discovery. Computer scientists study ways computers can understand and process diagrams. Other computer scientists develop displays that will effectively analyze and convey Big Data. Journalism schools now teach data visualization and diagram narratives, as these are increasingly important in journalism. The proliferation of digital tools have proliferated the use of diagrams.

## Goals and Plan of the Workshop

The goal of the workshop is dual: a) to bring together a diverse set of researchers working on various aspects of diagrammatic reasoning; b) to bring the issues and research to a broader audience in Cognitive Science. To these ends,
the workshop will have presentations from many disciplines: psychology, philosophy, computer science, education, design, and more. There will be two kinds of presentations: i) overview papers ( 30 minutes) by established researchers and ii) blitz talks (5 minutes) presenting specific current research projects. Blitz presentations have been highly successful in previous workshops, and are standard and excellent at large computer science meetings. The blitz presentations will allow broad participation from the cognitive science community and stimulate discussion around specific findings.

## Morning Session: Creating and Coordinating Diagrams

Barbara Tversky, "Creating Diagrams"
Professor Emerita of Psychology at Stanford University and a Professor of Psychology and Education at Teachers College, Columbia University, USA

Patrick Healey, "Coordinating Graphical Languages"
Professor for Human Interaction, School of Electronic Engineering and Computer Science, Queen Mary University of London, UK

David Kirsh, "Thinking with Illustrations"
Professor of Cognitive Science, University of California, San Diego, USA

Blitz Talks: Peter Coppin, University of Toronto; James Corter, Teachers College, Columbia University; Valeria Giardino, Free University Berlin; Azadeh Jamalian, Teachers College, Columbia University; Maithilee Kunda, Georgia Institute of Technology; David Peebles, University of Huddensfield; Anne Schüler, Knowledge Media Research Center; and others

## Afternoon Session: Diagrams in Science

William Bechtel, "Ways Scientists Reason with Diagrams" Professor of Philosophy, University of California, San Diego, USA

Peter Cheng, "Re-discovering diagrams and re-codifying knowledge in science"
Professor of Cognitive Science, Department of Informatics, University of Sussex, UK

Mary Hegarty, "Cognition and Metacognition in Reasoning with Diagrams"
Professor of Psychology, University of California, Santa Barbara, USA

Blitz Talks: Trevor Barrett, Univ. of California-Santa Barbara; Daniel Burnston, Univ. of California-San Diego; Jeff Nickerson, Stevens Institute of Technology; Benjamin Sheredos, Univ. of California-San Diego; Andy Stull, Univ. of California-Santa Barbara; and others

## Research Contributions of the Speakers

The presenters of the overview papers have each made distinctive contributions to cognitive science research on diagrams. The following are selective highlights.

Tversky $(2005,2011)$ has emphasized how people design diagrams by abstracting and schematizing contents, taking advantage of their spatial properties. She emphasizes how diagrams overcome limitations of internal information processing capacities, organize thought, and promote inference and discovery. Turning the focus to the interindividual coordination in the production and comprehension of visuospatial representations Healey (Healey, Swoboda, Umata, \& King, 2007) investigates the parallels between talking and drawing as modes of communication and the factors affecting the evolution of graphical dialects. Kirsh (2010) has identified a variety of ways external representations enhance cognitive power, including by providing a shareable object of thought, reducing the cognitive cost of inference, and coordinating thinking.

The afternoon session turns to uses of diagrams in relation to science. Bechtel has focused on how diagrams function to represent phenomena to be explained (Bechtel \& Abrahamsen, 2012), guide the search for the parts and operations of mechanisms (Sheredos, Burnston, Abrahamsen, \& Bechtel, in press) and direct the recomposition of mechanisms in computational models and synthetic organisms. Hegarty $(2010,2011)$ has employed experimental techniques to identify the cognitive abilities that underlie intelligent use of spatial representations and to address how spatial intelligence facilitates learning by students in the natural sciences. Cheng $(2002,2011)$ has explored how developing appropriate novel diagrammatic formats can enhance student learning about electric circuits and probability.

## References

Bechtel, W., \& Abrahamsen, A. (2012). Diagramming phenomena for mechanistic explanation. Proceedings of the 34th Annual Conference of the Cognitive Science Society (pp. 102-107). Austin, TX: Cognitive Science Society.
Cheng, P. C.-H. (2002). Electrifying diagrams for learning: principles for complex representational systems. Cognitive Science, 26, 685-736.
Cheng, P. C.-H. (2011). Probably good diagrams for learning: Representational epistemic recodification of probability theory. Topics in Cognitive Science, 3, 475-498.

Donald, M. (1991). Origins of the modern mind: Three stages in the evolution of culture and cognition. Cambridge, MA: Harvard University Press.
Healey, P. G., Swoboda, N., Umata, I., \& King, J. (2007). Graphical language games: Interactional constraints on representational form. Cognitive Science, 31, 285-309.
Hegarty, M. (2010). Components of spatial intelligence. Psychology of Learning and Motivation: Advances in Research and Theory, Vol 52, 265-297.
Hegarty, M. (2011). The cognitive science of visual-spatial displays: Implications for design. Topics in Cognitive Science, 3, 446-474.
Kirsh, D. (2010). Thinking with external representations. AI Soc., 25, 441-454.
Larkin, J. H., \& Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. Cognitive Science, 11, 65-99.
Norman, D. A. (1993). Things that make us smart: Defending human attributes in the age of the machine. Reading, MA: Addison-Wesley Pub. Co.
Schön, D. A. (1983). The reflective practitioner: How professionals think in action. New York: Basic Books.
Sheredos, B., Burnston, D., Abrahamsen, A., \& Bechtel, W. (in press). Why do biologists use so many diagrams. Philosophy of Science.
Tversky, B. (2005). Prolegomenon to Scientific Visualizations Visualization in Science Education. In J. Gilbert (Ed.), Visualization in Science Education (pp. 29-42). Dordrecht, Netherlands: Springer.
Tversky, B. (2011). Visualizing thought. Topics in Cognitive Science, 3, 499-535.

# Using Mechanical Turk and PsiTurk for Dynamic Web Experiments 

Anna Coenen (anna.coenen@nyu.edu)<br>Doug Markant (doug.markant@nyu.edu)<br>Jay B. Martin (jbmartin@nyu.edu) John McDonnell (john.mcdonnell@ nyu.edu)

Department of Psychology, New York University<br>New York, NY 10003

Keywords: Amazon Mechanical Turk; PsiTurk; Online Experiments; Crowdsourcing

## Objectives

This half-day workshop will demonstrate how to build custom web-based experiments that rely on participants from Amazon Mechanical Turk (AMT). Attendees will learn how to deploy web-based experiments using PsiTurk, a Pythonbased platform that simplifies the process of setting up experiments and interacting with AMT.

Workshops discussing the AMT marketplace have been offered at previous Cognitive Science Society meetings (e.g., Mason \& Suri, 2011). This workshop will complement those by stepping through a working demo that attendees can use to follow along and run on their personal computers. Importantly, the demo will illustrate how AMT can be used with dynamic, externally-hosted experiments, rather than the basic survey templates currently offered on AMT.

The workshop will have two parts. First, we will outline some of the general advantages and principles of using AMT for online behavioral experiments, including a basic introduction to the AMT website and the data collection process more generally. Second, we will show participants how to use the PsiTurk platform to run any web-based experiment on AMT. This portion of the workshop will emphasize "handson" training in AMT and PsiTurk that will teach attendees how to deploy their own web-based experiments.

## Outline of the Workshop

Throughout the workshop we will use both slides and live demonstrations of how to use AMT and PsiTurk for running web experiments.

## Introduction to Mechanical Turk

We will start by introducing the basic structure behind AMT and demonstrate how to run a simple project.

AMT is the largest online service in the US that offers a marketplace for tasks that need to be solved by human rather than machine intelligence. Human Intelligence Tasks (HITs) are submitted by requesters, such as corporations, researchers, organizations, or individuals in need for human participants. They can be completed by workers in exchange for a reimbursement that is set by the requester. Workers can also be awarded bonuses or have their payment rejected based on how they completed a HIT. We will walk attendants
through a simple example of how to post a HIT, oversee the data collection, and reimburse participants on the AMT website.

## Benefits and drawbacks of online experiments

Next, we will cover some of the advantages and pitfalls associated with using AMT for behavioral research.

For cognitive psychologists the appeal of using AMT lies in running computer experiments that would otherwise be completed in the lab, typically by undergraduate students. Online experiments have several advantages:

1. Data from a large number of participants can be collected quickly and at low costs. A few hours are typically sufficient for recruiting a full set of participants in a standard cognition or perception experiment.
2. Since the data collection is anonymous, using AMT minimizes experimenter effects and problems with contaminated subject pools at research departments.
3. For the same reason, experimental results become more replicable. Because subjects do not interact with an experimenter, there is no possibility for experimenter confound. If one researcher runs the code for another's experiment, it is, in principle, a pure replication: there is no source of systematic experimental deviation.
4. In general, web-based experiments are easier to share with other researchers since they are designed to run in standard web browsers and do not require any additional software. This facilitates the re-use of experimental code either for the purposes of direct replication or the design of new experiments.

Potential disadvantages of the method concern the quality of the data, including the possibility that comparatively low reimbursement might lower incentives to engage in a task. To address these questions, several authors have used AMT to replicate classic findings in their field. Paolacci, Chandler, and Ipeirotis (2010), for example, replicated a number of well-known cognitive biases using AMT data. Germine et al. (2012) found no systematic differences in the results of some widely-used perceptual paradigms using laboratory and online data. Rand (2012) also conducted an extensive study into the reliability of AMT workers' demographic data and verified that self-reported demographic information is highly
reliable. At NYU's Cognition and Computation lab we have successfully replicated the main findings of multiple classic studies in the concept learning literature (reported in Crump, McDonnell, \& Gureckis, in press, as Experiments 8-10), but found that it was critical to test participants for comprehension of the experimental instructions. We also manipulated the monetary incentives of one of these tasks and found it had little effect on the performance in the task, but did affect the dropout rate. In addition to these experimental replications, researchers have addressed the objective reliability of AMT data. Our workshop will delve into the findings of the literature so far on what sorts of experiments do and do not work on AMT.

## Running AMT experiments using PsiTurk

Finally, we will demonstrate how researchers can run experiments from their own website using AMT and PsiTurk.

Mechanical Turk offers some basic templates for simple online studies that can be built directly on the website. However, it can also be used to run any web-based experiment programmed directly by the researcher via the External Question type. To facilitate this process, John McDonnell and Todd Gureckis from NYU's Cognition and Computation lab coauthored and continue to maintain a Python-based platform that allows users to create HITs for experiments with minimal effort. It provides a back-end framework, handling interaction with Amazon's servers to credit participants, and logging participants' data and identifying information in a database. This allows researchers to build a user-facing front-end providing their own experimental code without having to write software to handle these logistical issues. The platform is available at http://github.com/NYUCCL/PsiTurk.

Over the course of the workshop, we will introduce the platform and show how attendants can run their own experiments on AMT. We will do so using a demo experiment coded in JavaScript that will be turned into a HIT. The code for this demo will be available for attendants to easily adapt to their own experimental needs.

## Audience

This workshop will appeal to cognitive science researchers who are conducting behavioral experiments in a wide number of areas. For those who are unfamiliar with AMT, the lecture portion of the workshop will explain the mechanics of AMT and review methods for designing and delivering experiments to participants. The interactive portion of the workshop will be particularly informative for scientists who wish to use AMT to run dynamic experiments that go beyond simple surveys, for example involving timing of stimulus presentation, collection of reaction times, or interactions with complex stimuli.

The workshop may also be of use to researchers who are unsure whether online research can accommodate their needs. For example, neuroscientists might be interested in using AMT as a platform for piloting experimental paradigms and online experiments in general for reducing dropout rates for
follow-up tasks, but may be unsure whether their paradigms can be easily translated into online experiments. One important theme of the workshop will be the capabilities and limitations of online experiments in general.

Although AMT is currently only available to requesters in the United States, we believe that researchers from other parts of the world could still benefit from the workshop. They might be able to use AMT through collaborations with laboratories in the United States, for example. Also, PsiTurk offers a general framework for running web experiments which can also be helpful for users of other online services.

## Preparation

We suggest that participants download the PsiTurk platform before attending the workshop and attempt to set it up before attending the workshop. If they do so, they will be able to follow along during the demonstration segment in which we launch an experiment on AMT.

## Presenters

All presenters of the workshop have used AMT and PsiTurk extensively to collect data, and have expertise in writing webbased experiments in JavaScript. John McDonnell is the coauthor and maintainer of the open-source PsiTurk framework for behavioral experiments on AMT. He has also validated AMT as a platform for studying learning using Turkers as participants (Crump et al., in press). The other speakers have several projects in preparation based on AMT data collected using PsiTurk. All of the speakers will be available throughout the workshop to assist attendants in setting up PsiTurk and using the AMT platform.

## References

Crump, M. J. C., McDonnell, J. V., \& Gureckis, T. M. (in press). The promise of mechanical turk: How online labor markets can help theorists run behavioral experiments. PLoS One.
Germine, L., Nakayama, K., Duchaine, B., Chabris, C., Chatterjee, G., \& Wilmer, J. (2012). Is the web as good as the lab? comparable performance from web and lab in cognitive/perceptual experiments. Psychonomic bulletin \& review, 1-11.
Mason, W., \& Suri, S. (2011). How to use mechanical turk for cognitive science research. In L. Carlson, C. Hölscher, \& T. Shipley (Eds.), Proceedings of the 33rd annual conference of the cognitive science society (pp. 66-67). Austin, TX: Cognitive Science Society.
Paolacci, G., Chandler, J., \& Ipeirotis, P. (2010). Running experiments on amazon mechanical turk. Judgment and Decision Making, 5(5), 411-419.
Rand, D. (2012). The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. Journal of Theoretical Biology, 299, 172-179.

# Embodied Approaches to Interpersonal Coordination: Infants, Adults, Robots, and Agents 

Rick Dale (rdale@ucmerced.edu)<br>Cognitive and Information Sciences<br>University of California, Merced<br>Merced, CA 95343 USA<br>Yukie Nagai (yukie@ams.eng.osaka-u.ac.jp)<br>Department of Adaptive Machine Systems<br>Osaka University<br>2-1 Yamada-oka, Suita, Osaka, 565-0871 Japan

Chen Yu (chenyu@indiana.edu)<br>Department of Psychological and Brain Sciences<br>Indiana University<br>Bloomington, IN 47401 USA

Moreno Coco (mcoco@staffmail.ed.ac.uk)<br>Institute of Language, Cognition, and Computation<br>University of Edinburgh<br>10 Crichton Street, Edinburgh EH8 9AB UK

Stefan Kopp (skopp@techfak.uni-bielefeld.de)<br>Sociable Agents Group, Cognitive Interaction Technology (CITEC)<br>Technische Fakultät, Universität Bielefeld<br>Morgenbreede 3933615 Bielefeld, Germany

Keywords: human interaction; language learning; human-agent interaction; dynamics; robotics.

## Workshop Background and Relevance

Humans interact with other humans. They do so frequently, in a wide variety of circumstances, to accomplish many different goals. This interpersonal interaction, especially in face-to-face circumstances, requires coordination (Clark, 1996). This involves many subtle behaviors, controlled carefully in the context of another person, from eye movements and gestures, to choice of words. The characteristics of the cognitive system that give way to this coordination have been a matter of debate recently in the cognitive sciences. Yet there remain many open questions about how the cognitive system functions in human interactions. How does interpersonal coordination emerge in the dyad? What behaviors are coordinated between persons, and in what manner? How can we model dyads and their interactions?

One challenge to advance our understanding of how human participants utilize social-cognitive cues in everyday communication is that the empirical evidence is based on macro-level behaviors in constrained unnatural contexts and tasks. To truly understand mechanisms of interpersonal coordination, however, we may need to focus on more micro-level behaviors as they unfold in real time, and in free flow interaction, for example, changes in eye gaze and shifts in body position as they are linked to objects, events, and actions of the social partner. Several new directions have pursued this microstructure of interpersonal interaction.

First, with advances in sensing and computing techniques, now we have the capabilities to process visual, audio and other sensory data collected from real-world interactions. This data-intensive approach provides a unique opportunity for new discoveries from various advanced data analysis techniques. These methods have leveraged visualization techniques to mine the temporal relationships between behaviors of two people (Yu et al. , 2009; see Fig. 1). This has shed light on the timing of interpersonal interaction, and how two individuals adapt to each other, both in infant-adult dyads (e.g., Smith et al., 2010; Yu \&
semi-automatic extraction $\longrightarrow$ event alignment and exploration


Figure 1: Visualization software for extracting, aligning and mining large multivariate time series of behaviors to uncover coordination (adapted from Yu et al., 2009).

Smith, 2012; Nagai et al., 2012), and in two adults (Coco et al., 2012; Richardson \& Dale, 2005).

Second, researchers in developmental robotics have investigated mechanisms of interpersonal coordination, to model and implement social systems. In developmental robotics, recent progress has been achieved in developing robots that elicit human scaffolding (Nagai, Nakatani, \& Asada, 2010). This progress has been possible by implementing underlying processes that could be involved in the dynamic control of interpersonal coordination. For example, implementing a model of a mirror neuron system can help basic skills in robots like self-other discrimination, and can support more complex abilities, such as imitation (Nagai et al., 2011; see Fig. 2, left). By grounding highlevel theories into robotic systems, we can address different aspects of how social-cognitive capabilities, such as gaze following and face preference, can be learned through sensorimotor interactions.

Third, research on virtual agents has developed new models of embodied human-agent interaction. This offers new ways to explore processes of interpersonal coordination. This has included, for example, the role of gesture and nonverbal behavior (Sadeghipour \& Kopp, 2011), attentive speaking (Buschmeier \& Kopp, 2011), and feedback (Kopp et al., 2008). Virtual embodied agents provide a foundation for testing theories of adult-adult interaction, and developing exciting social tools to support interpersonal coordination (see Fig. 2, right).


Figure 2. Left: Human-robot interaction; robot equipped with emergent mirror-neuron system (adapted from Nagai et al., 2011). Right: Humanagent interaction to explore models of gesture (adapted from Sadeghipour \& Kopp, 2011).

Together these strands of research offer new insight into human social dynamics, and the means to implement and test theories in robotics and virtual agents. Bringing them together in one workshop is an opportunity to convey these new methods, and find shared interests and synergies among different approaches and different fields, These are the primary goals of the workshop.

## Objective and Overview

The aim of this workshop is to introduce these approaches in an integrative fashion, and offer some basic demonstrations of relevant software, data analysis, and development.
Broad audience. Given the international and interdisciplinary composition of the workshop, we expect to attract broad interest from several domains, from cognitive and language development, to language processing and discourse; from human cognition to artificial intelligent systems; and from human babies, to adult, and to both physical social robots and virtual agents.
Activities. The organizers of the workshop will first offer a series of presentations on relevant research projects (see Schedule). These topics form a coherent collection of new approaches to interpersonal interaction, shown below in Table 1. Talks will include concrete details regarding data collection, system design, and so on; where appropriate, source code or software will be demonstrated and distributed to attendees (e.g., Coco \& Dale's R toolbox for recurrence). The workshop organizers will together lead a discussion with the attendees on limitations, future directions, and so on.

Table 1: Thematic organization of workshop organizers covering domains of interpersonal coordination.

|  | infant | adult |
| :---: | :---: | :---: |
| human | Nagai, Yu | Kopp, Coco, Dale |
| robot / agent | Nagai, Yu | Kopp |

Outcomes. Attendees will gain a basic understanding of human data analysis in the case of large-scale multivariate behavioral data mining (Yu et al., 2009, 2012), and the application of a particular technique referred to as cross recurrence analysis (Dale et al., 2011) which serves as a simple quantification over behavioral channels ( R toolbox
developed by Coco \& Dale, in preparation). Nagai and Kopp will offer details of developing robotics and artificial agents.

## Schedule

| Duration | Topic (Speaker) |
| ---: | :--- |
| $0-5$ minutes | Introduction to the workshop (Dale) |
| 40 minutes | Infant-caregiver coordination through software visualization (Yu) |
| 40 minutes | Adult coordination and cross recurrence analysis (Dale \& Coco) |
| 5 minutes | Break |
| 40 minutes | Social and developmental robotics and interpersonal interaction (Nagai) |
| 40 minutes | Virtual social agents, human-agent interaction, and coordination (Kopp) |
| 20 minutes | Discussion |
| Total: | $\sim 3$ hours |

## Further Materials

The first author of the workshop will maintain a website to distribute publications and software for attendees.

## References

Buschmeier, H., \& Kopp, S. (2011). Towards conversational agents that attend to and adapt to communicative user feedback. In Intelligent Virtual Agents (pp. 169-182).
Clark, H. H. (1996). Using language. Cambridge University Press.
Coco, M. I., Dale, R., \& Keller, F. (2012, September). Cognitive dynamics of alignment in dialogue games. Talk presented at Architectures and Mechanisms for Language Processing (AMLaP), Riva Del Garda, Italy.
Kopp, S., Allwood, J., Grammer, K., Ahlsen, E., \& Stocksmeier, T. (2008). Modeling embodied feedback with virtual humans. Modeling communication with robots and virtual humans, 18-37.
Nagai, Y., Kawai, Y., \& Asada, M. (2011). Emergence of mirror neuron system: Immature vision leads to self-other correspondence. In Development and Learning (ICDL), 2011 IEEE International Conference On (Vol. 2, pp. 1-6).
Nagai, Y., Nakatani, A., \& Asada, M. (2010). How a robot's attention shapes the way people teach. In Proceedings of the 10th International Conference on Epigenetic Robotics.
Nagai, Y., Nakatani, A., Qin, S., Fukuyama, H., MyowaYamakoshi, M., \& Asada, M. (2012). Co-Development of Information Transfer within and between Infant and Caregiver. In Proceedings of the IEEE ICDL-EpiRob.
Richardson, D. C. \& Dale, R. (2005). Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. Cognitive Science, 29, 39-54.
Sadeghipour, A., \& Kopp, S. (2011). Embodied gesture processing: motor-based perception-action integration in social artificial agents. Cognitive Computation, 3, 419-435.
Smith, L. B., Yu, C., \& Pereira, A. F. (2011). Not your mother's view: the dynamics of toddler visual experience. Developmental Science, 14(1), 9-17. doi:10.1111/j.14677687.2009.00947.x

Yu, C., \& Smith, L. B. (2012). Embodied attention and word learning by toddlers. Cognition, 125(2), 244-262. doi:10.1016/j.cognition.2012.06.016
Yu, C., Zhong, Y., Smith, T., Park, I., \& Huang, W. (2009). Visual data mining of multimedia data for social and behavioral studies. Information Visualization, 8(1), 56-70. doi:10.1057/ivs.2008.32

# Interactive Panel Discussion: Professional advancement through real and virtual international collaboration 

Laurie Beth Feldman* (lfeldman@albany.edu)<br>Department of Psychology; SS 369<br>The University at Albany, SUNY<br>Albany, NY 12222, USA

Judith Kroll (jfk7@psu.edu)
Moore Building
Department of Psychology
The Pennsylvania State University
University Park, PA 16802 USA
Natasha Tokowicz (Tokowicz@pitt.edu)
LRDC Room 634
3939 O'Hara St.
University of Pittsburgh
Pittsburgh, PA 15260-5169 USA

Keywords: cross cultural diversity; networking; participation; professional development; research collaboration.

Women in Cognitive Science (WICS) has a history of conducting panels at yearly meetings of several professional societies. Their goal is to increase attention to the situation of women cognitive scientists, to better understand the reasons for existing problems of under representation in key positions, and to provide a forum for professional development that encourages both junior and senior scientists to consider the ways in which they might work with their own home institutions to effect change. Specific topics have addressed networking and collaboration, best practices for institutional transformation, and issues of family and academic careers. Speakers and panelists have included both women and men who represented senior and junior scientists and topics have focused on the experience of both faculty and administration in negotiating these issues and in developing policies that are likely to support women's success. Its history demonstrates that WICS is in a unique position to address the concerns of junior as well as senior scientists as they pursue their professional careers in the sciences.

The Interactive Panel Discussion: Professional advancement through international collaborations seeks to bring together American and European researchers in Cognitive Science at the 2013 meeting of the Cognitive Science Society in Berlin, Germany. Speakers will discuss national and international collaborations as a tool towards professional advancement and visibility.

Janet van Hell (jgv3@psu.edu)<br>Moore Building<br>Department of Psychology<br>The Pennsylvania State University<br>University Park, PA 16802 USA<br>Suparna Rajaram (srajaram@notes.cc.sunysb.edu)<br>Department of Psychology<br>Stony Brook University<br>Stony Brook, NY 11794-2500 USA


#### Abstract

The central theme will be how to develop new research collaborations outside of one's primary institution, including international collaborations and why this is important to one's career development. All acknowledge that this is not a simple process and often evolves slowly, out of more social networking connections. While such solutions generally occur on an ad hoc basis and vary tremendously across individuals and settings, the aim of the WICS workshop is to enable discussion of this and related activities that enhance productivity and thus visibility. Discussion will include the role of virtual collaborations and virtual research networks in enhancing professional enhancement, even for women with less opportunity/money/time to travel.


At the time of publication, R. Harald Baayen (Eberhard Karls University, Tübingen), Melody Dye (Indiana University, Bloomington), Laurie Feldman (SUNY, Albany \& Haskins Labs), Lael Schooler (Max Planck Institute for Human Development, Berlin), and Anne Warlaumont, (University, of California, Merced) have been invited to describe and contrast varied formats and goals of their international collaborations. These people represent American and European junior and senior researchers, some who have held positions as university administrators and department heads and chairs. Each will share his/her perspective.
The theme of virtual as well as real collaboration will be of relevant not only to those who have funds to attend the meeting in Berlin but to those with more limited travel budgets. Materials from the speakers will be posted on the Women in Cognitive Science website. Because the networking function of our workshop differs from the goal
of most other workshops, it has been scheduled in a special time slot that conflicts only minimally with other workshops.

* Contact person


## Acknowledgments

Women in Cognitive Science (WICS) was founded in 2001 by Judith Kroll (Penn State), Randi Martin (Rice University), and Suparna Rajaram (Stony Brook) with NSF ADVANCE Funds. From 2007 onwards, Laurie Feldman (Albany) and Janet van Hell (Penn State) have assumed a leadership role within the group. In 2012, Natasha Tokowitz joined the group.
Partial funding for this event comes from NSF Award BCS-1049764.

# Workshop Proposal: PRE-CogSci 2013 - Bridging the gap between cognitive and computational approaches to reference 

Albert Gatt

Roger van Gompel
Ellen Gurman Bard
Emiel Krahmer
Kees van Deemter

## Outline

This full-day workshop explores the Production of Referring Expressions (PRE) from different perspectives. It follows two earlier workshops on the same theme, which were associated with two earlier CogSci conferences. The first of these, PRE-CogSci 2009, focussed on the interplay between computational and empirical methods. ${ }^{1}$ The second, PRECogSci 2011, broadened this theme to include work on dialogue and linguistic theory. ${ }^{2}$ Both events were highly successful, with each containing over 20 presentations and an audience of over 60 participants. Following the first workshop, the Topics in Cognitive Science journal published an issue containing 8 peer-reviewed articles selected from 28 full paper submissions (Krahmer et al. 2012). Following the second workshop, we are finalising a Special Issue of the journal Language and Cognitive Processes; accepted papers are expected to appear in 2013-14. A survey of research, by 2 of the organisers, on the generation of referring expressions that came out in the journal Computational Linguistics in March 2012 has since been downloaded over 700 times.

This third PRE workshop is part of CogSci 2013. It will be different from its predecessors in two ways. Firstly, a range of new intellectual themes will be explored that have emerged in recent years. Second, the workshop will be coupled with a satellite event, which will be separately funded, and which will be devoted to the role of reference in practical applications. We expand on these issues below.

## Workshop overview

Interest and significance As the earlier workshops demonstrated, there is abundant scope for cross-disciplinary research on reference. Recent advances (see below) have caused the interest in PRE-CogSci 2013 to increase further.

Importance The boundaries between the different disciplines working on reference are starting to shift, with computational methods (e.g., stochastic and Bayesian methods; ACT-R) permeating theoretical and experimental work. Increased collaboration between computational linguists and psycholinguists is bound to be mutually beneficial: computational work will benefit from a greater awareness of psycholinguistic methods and findings; psycholinguistic models (which often underdetermine the phenomena) have the potential to benefit from precise and explicit algorithmic models. The importance of this cross-disciplinary collaboration has been acknowledged though the award of two recent projects, namely the (Dutch)

[^0]NWO-VICI project Bridging the gap between psycholinguistics and computational linguistics: the case of Referring Expressions (Krahmer, 2009 - 2013), and the (British) EPSRC project REFNET An Interdisciplinary Network Focussing on Reference, (Bard and van Deemter, 2012-2015). Four new themes have recently emerged in this area, which will be explored during the workshop:

- Collaborative reference. Referring expressions are often produced collaboratively, as when one speaker says "The hill just north of us" and the other adds "You mean with the tower on top?" We believe the time has come to address collaborative reference using the combination of computational and empirical methods that have become mainstream in the study of reference. Research that takes off from a view of language as a social and collaborative phenomenon has the potential to shift the goal of production models away from the aim of identifying a referent uniquely towards effective communication.
- Nondeterminism in production. Evidence suggests that a given speaker, confronted with a given situation, will not always produce the same utterance (just like a marksman doesn't always hit the same area of a target). Recent models have started to model algorithmically what variation is found in the production of referring expressions.
- Interaction between comprehension and production. Work on comprehension (i.e., interpretation) and production (i.e., generation) has typically proceeded separately. Recent work on reference, however, has started to study comprehension and production jointly, seeking to exploit their common mechanisms and discover their differences.
- Combinations with research on vision. Recently, researchers in Natural Language Generation and researchers in Computer Vision have joined forces (e.g., EPSRC's Vision \& Language Net in the UK, Belz and Makris 20102013) for example to allow computers to describe a picture and refer to objects in it. This has raised new questions concerning the way in which visual information is represented in the human brain, and how this may be modelled computationally.

Relevance to CogSci 2013 Apart from its relevance to language comprehension and production, the workshop's focus on collaborative reference - in which a sequence of actions by two or more agents achieves an effect that none of them might have achieved on her own - will make it directly relevant to the modelling of social interaction. Our satellite event (see below) will lend further emphasis to social aspects of reference, focussing on real-life situations.

Suitability of organisers Apart from their success in organising the first two PRE-CogSci workshops, the organisers are leading researchers in the field of reference, and their backgrounds range from psycholinguistics to theoretical and applied computing science.
Proposed format and funding For the main event, we envisage a full-day workshop, consisting of two half-sessions, each starting with an address by a keynote speaker. We envisage 8-10 oral presentations and a poster session. Travel expenses of the keynote speakers will be split between the abovementioned projects NWO-VICI and EPSRC-REFNET.

As a satellite, probably on the day before the workshop, we envisage a Cross-pollination meeting in which researchers will interact with practitioners in Human-Computer Interaction, Robotics, and Geographical Information Systems, to discuss challenges involving reference that come from real applications. The satellite event will not be part of the workshop and will be financed by the RefNet project (see above), but we expect the two events to dovetail well, thus adding to the number of people likely to want to come to CogSci.
Likely audience and attendees The workshop will be of interest to a variety of cognitive scientists, including psycholinguists with an interest in computational modelling; computational linguists with an interest in experimental methods; and a limited number of others who will be attracted to the event because of the (satellite) Cross-pollination event; these will probably include a number of representatives from industry. Based on the experience of the first two editions of the workshop, we expect 50-70 participants.
Two keynote speakers have agreed to speak at the workshop: Professor Herb Clark, and Dr Noah Goodman. Professor Clark (Stanford) is a psycholinguist whose longstanding experimental work on reference has been more influential than anyone else's; his work has often stressed collaboration and other social aspects of communication. Dr Goodman (MIT, then Stanford) works in a new area where computer science overlaps with psychology; he has done extensive work in concept learning and has recently turned his attention to the modelling of reference using stochastic algorithms, integrating models of interpretation with models of production. Both speakers' work is highly relevant to the workshop. (See the themes mentioned under the header Importance.) In addition, the following people have agreed to form part of the programme committee:

- Mira Ariel, Tel Aviv University, Israel;
- Jennifer Arnold, University of North Carolina, USA;
- Adrian Bangerter, Univ. of Neuchâtel, Switzerland;
- Dale Barr, University of Glasgow, UK;
- Eva Belke, Ruhr-University Bochum, Germany;
- Holly Branigan, University of Edinburgh, UK;
- Susan Brennan, Stony Brook University, USA;
- Sarah Brown-Schmidt, Univ. Illinois, Urbana-Champaign
- Herb Clark, Stanford University, USA;
- Victor Ferreira, University of California, USA;
- Jeanette Gundel, University of Minnesota, USA;
- Martijn Goudbeek, Tilburg University, The Netherlands;
- Markus Guhe, University of Edinburgh, UK;
- Daphna Heller, University of Toronto, Canada;
- John Kelleher, Dublin Institute of Technology, Ireland;
- Frank Keller, University of Edinburgh, UK;
- Ralf Klabunde, Ruhr University Bochum, Germany;
- Danielle Matthews, University of Sheffield, UK;
- Margaret Mitchell, Johns Hopkins University, USA;
- Paul Piwek, Open University, UK;
- Massimo Poesio, University of Trento, Italy;
- Ehud Reiter, University of Aberdeen, UK;
- Amanda Stent, AT\&T Labs and Stony Brook Univ., USA;
- Matthew Stone, Rutgers University, USA;
- Takenobu Tokunaga, Tokyo Inst. of Technology, Japan;
- Mariët Theune, Twente University, Netherlands.

Publicity and impact A dedicated website, http://pre2013.uvt.nl/, links to the sites of the previous editions of PRE-CogSci, and will (eventually) contain the electronic workshop proceedings, in the form of extended abstracts. The workshop has been advertised on several emailing lists in experimental psychology, computational linguistics and theoretical linguistics. As in the case of the previous two workshops, we envisage a special issue in an appropriate scholarly journal.
List of Requirements The only requirements are a data projector and computer, preferably with internet connection.

## Contact details

Kees van Deemter ${ }^{3}$
Professor in Computing Science
Computing Science department
University of Aberdeen
Aberdeen AB24 3UE, Scotland, UK
email: k.vdeemter@abdn.ac.uk
Names and affiliations of additional authors:
Albert Gatt, Universities of Malta and Tilburg ${ }^{4}$
Ellen Gurman Bard, University of Edinburgh ${ }^{5}$
Roger van Gompel, University of Dundee ${ }^{6}$
Emiel Krahmer, Tilburg University ${ }^{7}$

## References

Krahmer, Van Gompel, Gatt and Van Deemter (2012). Topic "Bridging the Gap between Computational and Empirical Approaches to Reference", TopiCS in Cognitive Science 4 (2), pp.165-329. (Contains 8 peer-reviewed articles.)

[^1]
# Motivations and Goals in Developing Integrative Models of Human Cognition 

Glenn Gunzelmann (glenn.gunzelmann@us.af.mil)<br>Cognitive Models \& Agents Branch, Air Force Research Laboratory<br>711 HPW/RHAC, 2620 Q St., Building 852<br>Wright Patterson Air Force Base, OH 85212-6061 USA


#### Abstract

There has been tremendous growth recently in theories that attempt to provide more comprehensive accounts of the foundational mechanisms of human cognition. Such theories have taken a variety of forms, and have focused on different levels of analysis. The diversity is important and necessary, but can serve as a barrier to interaction, comparison, and integration, even at venues like the Annual Meeting of the Cognitive Science Society that should foster such dialogue. This workshop is intended to bring together individuals working on integrative models of human cognition, to emphasize shared motivations and goals. Ultimately, building scientific communities that bridge levels of analysis, methodologies, and theoretical approaches to work toward more comprehensive theories will be critical to the addressing the central goal of the Cognitive Science Society understanding the nature of the human mind.


Keywords: Integrated Models; Unified Theories of Cognition; Cognitive Models; Cognitive Architectures; Neural Architectures.

## Introduction

The motivations for integrative models of human cognition have their roots in the origins of cognitive science as a scientific discipline. Even before cognitive psychology emerged, ideas about unifying principles to explain cognition were expressed in the scientific literature (e.g., Newell, Shaw, \& Simon, 1958; Rosenblatt, 1961), including so-called grand psychological theories proposed during the first half of the $20^{\text {th }}$ century. The call for more comprehensive theories was explicitly made by Newell (1973), who expressed concern about the prospect that traditional, phenomenon-driven cognitive psychology would, by itself, lead to the kind of integrative understanding of the human mind that is the goal of cognitive science.

In the decades since, integrative theories of human cognition have become increasingly prevalent in cognitive science. These theories now represent an exciting diversity of theoretical approaches and levels of analysis, better reflecting the diversity of the cognitive science community as a whole. As noted by McClelland (2009), this growth has been tied in important ways to sustained increases in computing power that enable cognitive modeling at a scale and resolution that was unimaginable half a century ago.

The participants in the workshop have been selected to capture much of this theoretical diversity. The current state of the art in this area makes this workshop a timely and important contribution to the Annual Meeting of the Cognitive Science Society and the broader cognitive science community.

## Goals and Scope

The need to bring together this community of researchers was expressed by Newell (1990). Newell explicitly and deliberately referred to Unified Theories of Cognition in the plural, noting that multiple implementations are important for progress in the science. More recently, McClelland (2009) emphasized that "different simplifications are required to explore different issues." (p. 12). Interactions among cognitive scientists from different methodological and theoretical backgrounds are crucial to identifying common foundations and interconnections among levels of analysis and theoretical perspectives.

To reinforce and further develop the identity of this scientific community, this workshop will create an important opportunity for interaction and discussion amongst researchers working toward more integrative theories of cognition. It will not focus on a debate about the merits of developing integrative models of human cognition. The participants share, in general, an appreciation of the value of developing such theories, which provides the unifying theme for the event.

In addition, the workshop will not focus on the claims of particular integrative models. That is an important scientific activity, but the goal here is to build broader appreciation of shared motivations and goals, despite sometimes very different approaches and theories. All of the presenters seek unifying mechanisms that cut through the complexity of human cognition and enhance our understanding. Complementary perspectives and opportunities for integration will be highlighted to emphasize connections. In addition, contemporary challenges in this pursuit will be discussed, which will facilitate future scientific debates regarding particular claims and mechanisms.

## Workshop Organization

The workshop will be organized around a set of presentations and opportunities for discussion. The focus will not be on theoretical overviews. Instead, contributors will comment on the role of integrative models in cognitive science, including understanding the fundamental principles of human cognition broadly, and integrating across components of cognition to perform complex tasks. Presenters will highlight links to alternative approaches and methodologies, and discuss current challenges in developing integrative models of the human mind.

Speakers will be given approximately 25 minutes, with no more than 15 minutes of presentation material. This will allow significant opportunity for questions, comments, and discussion. In addition, the closing session of the workshop will consist of a panel discussion with the goal of
identifying and highlighting common themes and perspectives that have emerged across the day.

## Workshop Organizer

Dr. Glenn Gunzelmann is the Science and Technology Advisor for the United States Air Force Research Laboratory's Cognitive Models and Agents Branch. His research attempts to expand the explanatory breadth of integrative theories, including theories of human spatial competence (e.g., Gunzelmann \& Lyon, 2011) and fatigue (e.g., Gunzelmann, Gross, Gluck, \& Dinges, 2009).

## Target Audience

This workshop targets the central purpose of the Cognitive Science Society - it "brings together researchers from many fields who hold a common goal: understanding the nature of the human mind" (CSS Website). The presenters represent a range of disciplines in cognitive science, who will focus on important challenges associated with creating more comprehensive theories. Because this steps back from the theoretical and technical details of the theories, this workshop should be approachable and of interest to a broad audience at CogSci 2013 , and will make an important contribution to the event.

## Participants

## Presenters

## Joscha Bach

Berlin School of Mind and Brain
Humboldt University

## Jerome Busemeyer

Cognitive Science Program
Department of Psychological and Brain Sciences
Indiana University

## Chris Eliasmith

Canada Research Chair in Theoretical Neuroscience
Departments of Philosophy \& Systems Design Engineering University of Waterloo

## Noah Goodman

Department of Psychology
Stanford University

## Rick Granger

Psychological and Brain Sciences Department
Dartmouth University

## Andrew Howes

School of Computer Science
University of Birmingham, UK
Rick Lewis
Departments of Psychology and Linguistics University of Michigan
Randy O'Reilly
Department of Psychology and Neuroscience University of Colorado Boulder

## Ron Sun

Cognitive Science Department
Rensselaer Polytechnic University
Niels Taatgen
Department of Artificial Intelligence
University of Groningen

## Josh Tenenbaum

Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology

## Discussion Panel

Wayne D. Gray
Cognitive Science \& Computer Science Departments Rensselaer Polytechnic Institute

## Jay McClelland

Department of Psychology
Stanford University
Frank E. Ritter
College of Information Sciences and Technology
The Pennsylvania State University

## Archiving

In conjunction with the workshop, a proposal is being submitted to Topics in Cognitive Science. Presenters at the workshop will be invited to submit brief papers to complement the presentations at the workshop. Discussants will be encouraged to submit commentaries to the Topic on Integrative Models of Human Cognition, based on both the papers and workshop presentations. All workshop attendees, and others in the cognitive science community, will also be invited to contribute to the discussion in the journal.

## References

Gunzelmann, G., Gross, J. B., Gluck, K. A., \& Dinges, D. F. (2009). Sleep deprivation and sustained attention performance: Integrating mathematical and cognitive modeling. Cognitive Science, 33(5), 880-910.
Gunzelmann, G., \& Lyon, D. R. (2011). Representations and processes of human spatial competence. Topics in Cognitive Science, 3(4), 741-759.
McClelland, J.L. (2009). The place of modeling in cognitive science. Topics in Cognitive Science, 1(1), 11-38.
Newell, A. (1990). Unified Theories of Cognition. Cambridge, MA: Harvard University Press.
Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W. G. Chase (Ed.), Visual information processing (pp. 283-308). New York: Academic Press.
Newell, A., Shaw, J. C., \& Simon, H. A. (1958). Elements of a theory of human problem solving. Psychological Review, 65(3), 151-166.
Rosenblatt, F. (1961). Principles of neurodynamics. Washington, DC: Spartan Books.

# Deception as a Social Strategy 

Swati Gupta (guptas@ihpc.a-star.edu.sg), Tei Laine (lainet@ihpc.a-star.edu.sg), Kayo Sakamoto (sakamotok@ihpc.a-star.edu.sg)<br>Programme in Computational Social Cognition, Institute of High Performance Computing, 1 Fusionopolis Way, \#16-16 Connexis North, Singapore 138632

Keywords: deception; lying; social interaction; interpersonal relationships; decision making; goals; computational modeling; risk taking; incentives; motivation; message design


#### Abstract

Deception can be advantageous to a deceiver when the truth conflicts with his or her goals - be they personal or social, selfless or selfish. Thus it is that people regularly use deception to avoid conflict with others, to avoid punishment or embarrassment, to fit into a group, to harm, protect or help others, and for material or non-material benefit to themselves. While there are many ways in which people can deceive, for example, by choosing to fabricate rather than to tell a half-truth, there are always cost-benefit trade-offs, regardless of the strategy a person chooses. Understanding why people deceive in everyday life situations, how they do it, why they choose one strategy over another, and why sometimes they might choose not to deceive at all, even in the absence of any serious anticipated cost, will enable us to build richer models of socially intelligent behavior--models that could be employed in computational systems designed to facilitate enterprises such as elder care, tutoring, and professional training. In this workshop, we aim to address three basic questions: (1) what factors lead people to deceive? (2) what makes them decide to deceive one way rather than another? and (3) how can we model these factors computationally?


## What we aim to achieve

We aim to highlight research on how people use deception as a social strategy, and provide an overview of how issues related to the questions raised above have been addressed by various disciplines. Even though there has been research in several fields including communication studies, philosophy, linguistics, psychology, economics, and neuroscience, on the question of why and how people deceive one another, there is no comprehensive integration of the different views and observations found in the literature. The confirmed speakers represent many of these disciplines, and this workshop would be a good opportunity to explore together similar research questions from different disciplinary perspectives.

## Why is it important

Most of the research on deception has focused on deception detection, rather than generation. We believe there is a need to move in the direction of rectifying this imbalance, for the reasons mentioned in the abstract above. This workshop will provide an opportunity for people interested in deception to focus more on the generation aspect, and will help in making a stronger case of more research in this area.

## Relevance to CogSci 2013

The goals of this workshop, which focuses on an emerging and a cross-disciplinary research topic, are directly linked to the theme of CogSci 2013, which is "Cooperative Minds: Social Interaction and Group Dynamics." We treat deception as a social strategy, which often is used to enhance social interactions and cooperation within groups.

## The Organizers

We are an interdisciplinary team of scientists, with a background in computer science, cognitive science, computational linguistics, and psychology. For the past two years, we have been working on a project on this topic, in which we have brought together literature from diverse fields, and have investigated issues related to verbal deception. We proposed a theory of verbal deception which demonstrated the interaction between a taxonomy of verbal deception types, and the communicative goals of deception. Focusing on the decision making aspect of deception, we also analyzed factors that determine whether a person would deceive or not, and found that these factors differ depending on whether the person is facing potential losses or gains. Following are our most relevant publications.

Gupta, S., Sakamoto, K., \& Ortony, A. (2012).
Telling it like it isn't: a comprehensive approach to analyzing verbal deception. In F. Paglieri, L. Tummolini, R. Falcone \& M. Miceli (Eds.), The goals of cognition: Festschrift for Cristiano Castelfranchi. London, College Publications.

Sakamoto, K., Laine, T., and Farber, I. (in press). Deciding whether to deceive: Determinants of the choice between deceptive and honest communication. Journal of Economic Behavior \& Organization.

## Expected audience

This is a topic of general interest. We expect it to attract audience from a variety of disciplines like psychology, linguistics, philosophy, as well as artificial intelligence, economics, and neuroscience. People working in interpersonal / strategic communication, decision making, unethical behavior, morality, truth, and business ethics, would find it particularly relevant. It would also be of interest to people working on real world domains like law, insurance, tutoring, professional training, and elder care.

Our estimate for the number of people in the audience is approximately 40.

## Confirmed Speakers

We are currently awaiting responses from a few more people, and plan to have around 9-10 speaker slots. Below is the list of confirmed speakers.

## Steve McCornack

467 Communication Arts and Sciences Building
Michigan State University
East Lansing, MI 48824, USA
Email: mccornac@msu.edu
Phone: (517) 355-3478
Fax: (517) 432-1192
Timothy R. Levine
482 Communication Arts and Sciences Building
Michigan State University
East Lansing, MI 48824, USA
Email: levinet@msu.edu
Phone: (517) 432-1124
Fax: (517) 432-1192

## Giorgio Ganis

School of Psychology, Cognition Institute
University of Plymouth
Drake Circus, Plymouth, Devon, PL4 8AA, UK
Email: giorgio.ganis@plymouth.ac.uk
Phone: (44)1752584812
Shaul Shalvi
Room 108 Building 98, Psychology Department
Ben Gurion University of the Negev
POB 653, Beer Sheva 84105, Israel
Email: sshalvi@bgu.ac.il
Phone: (972) 086472049

## Shahar Ayal

School of Psychology
Interdisciplinary Center (IDC) Herzliya
P.O.Box 167, Herzliya, 46150, Israel

Email: s.ayal@idc.ac.il
Phone: (972) 99602799

## Aaron C. Elkins

Intelligent Behaviour Understanding Group, Department of Computing
Imperial College London
180 Queen’s Gate, London SW7 2AZ, UK
Email: a.elkins@imperial.ac.uk
Phone: (44)2075948195
Fax: (44)2075818024

## Swati Gupta

Programme in Computational Social Cognition Institute of High Performance Computing 1 Fusionopolis Way, \#16-16 Connexis North Singapore 138632
Email: guptas@ihpc.a-star.edu.sg
Phone: (65) 64191503
Tei Laine
Programme in Computational Social Cognition
Institute of High Performance Computing
1 Fusionopolis Way, \#16-16 Connexis North
Singapore 138632
Email: lainet@ihpc.a-star.edu.sg
Phone: (65) 64191305
Kayo Sakamoto
Programme in Computational Social Cognition
Institute of High Performance Computing
1 Fusionopolis Way, \#16-16 Connexis North
Singapore 138632
Email: sakamotok@ihpc.a-star.edu.sg
Phone: (65) 64191285

## Publicity

Some of the relevant mailing lists that we plan to advertize on are JDM, SPSP, IU's cogsci list, and Cognitive Science Student Society. We will utilize the social networking platform to publicize on the Facebook pages of relevant organizations like the Cognitive Science Society, International Communication Association, Association For Business Communication, and AAAI, to name a few. We also plan to advertise at the Joint Action Meeting, which is collocated with CogSci.

We would also like to document the workshop outcomes in the form of an article or a special issue in the journal of a related discipline.

# Mental Model Ascription by Language-Enabled Intelligent Agents 

Marjorie McShane (marge@umbc.edu)<br>Department of Computer Science and Electrical Engineering<br>University of Maryland Baltimore County<br>Baltimore, MD, 21250, USA

## Topic and Goal

Mental model ascription can be defined as inferring features of another human or artificial agent that cannot be directly observed, such as that agent's beliefs, plans, goals, intentions, personality traits, mental and emotional states, and knowledge about the world. This capability is an essential functionality of intelligent agents if they are to engage in sophisticated collaborations with people. The computational modeling of mental model ascription offers an excellent opportunity to explore the interaction of traditionally separate modules of cognitive architectures, such as language understanding, plan- and goal-oriented reasoning, and memory management.

The study of mental model ascription can benefit from advances in fields as disparate as machine reasoning, social interaction, developmental psychology, robotics, emotion, philosophy and computational linguistics, to name just a few. ${ }^{1}$ The common thread of this workshop will be the computational modeling of unobservable features by intelligent agents using language input as at least one of their modes of perception. Topics of interest include but are not limited to:

1. Developing computational treatments of language phenomena (e.g., indirect speech acts, irony, paraphrase, humor, coercion) that require or give rise to mental model ascription.
2. Applying computational models of other cognitive capabilities (dialog, emotion, agent collaboration/competition and plan- and goaloriented reasoning) to mental model ascription.
3. Modeling agent decisions about what to learn about other agents' unobservable features, considering that attempting to learn everything in every context would incur a heavy cognitive load.
4. Modeling how agents measure their confidence in the results of mental model ascription, which will be affected by their confidence in their understanding of contributing linguistic (or other) percepts as well as their ability to make valid inferences.
5. Modeling dynamic belief modification, including overriding a previous belief and managing memories with respect to modified beliefs.
${ }^{1}$ As a comparison, CogSci 2012 featured a workshop, "Modeling the Perception of Intention" that treated intention recognition with an emphasis on visual perception.

The main goal of the workshop is to foster mutual learning, discussion and future collaboration among researchers pursuing agent-oriented mental model ascription in integrative cognitive architectures.

## Program Committee (confirmed)

Ron Artstein (USC)<br>Jerry Ball (Air Force Research Laboratory)<br>Paul Bello (Office of Naval Research)<br>Graeme Hirst (University of Toronto)<br>Eva Hudlicka (Psychometrix Associates, Inc.)<br>Pat Langley (University of Auckland, NZ and CMU)<br>Marjorie McShane, Chair (UMBC)<br>Sergei Nirenburg (UMBC)<br>Massimo Poesio (University of Essex)<br>Chris Potts (Stanford University)<br>Yorick Wilks (IHMC).

## Organizational

This will be a full day workshop that will include invited talks, talks selected by abstract submission, a round table discussion, and, optionally, a poster session. Talks will be grouped by similarity of theme and approach, and the schedule will allow for extended discussion of each group of presentations, best exploiting the workshop genre.

We expect $30-40$ participants that include students and researchers with broad interests in the computational modeling of cognition and/or psychologically-inspired natural language processing.

The final session of the day will be devoted to planning a special journal issue (for Advances in Cognitive Systems) of papers inspired by the workshop. There are no special requirements for participants in the workshop.

The workshop website is http://ilit.umbc.edu/Workshop/MentalModelCogSci2013.html. The contact email is mentalmodel2013@gmail.com.

The workshop organizer, Marjorie McShane, has been working in the field of AI-NLP for the past fifteen years, with recent work focusing on the development of cognitive simulations of virtual patients to support clinician training. For a brief CV and list of publications, see http://ilit.umbc.edu/PubMcShane.htm.

# When eye see you: Gaze and joint attention in human interaction 

Maria Staudte (masta@coli.uni-sb.de)<br>Department of Computational Linguistics<br>Saarland University

Ulrich Pfeiffer (ulrich.pfeiffer@uk-koeln.com)<br>Department of Psychiatry<br>University Hospital Cologne

Keywords: joint attention; gaze; social interaction; language processing; dialog; grounding.

## Topic and Goals

Gaze behavior provides fundamental mechanisms for sharing mental states such as goals and desires and helps to ground communicative content. In order to establish common ground in verbal and non-verbal interactions, interlocutors often need to acquire knowledge about their interaction partners' focus of visual attention by following their gaze and, in turn, have to direct their partners attention to their own target object or location. Responding to or leading someone's gaze to a location or an object of interest results in a situation of joint attention - a referential triad between two individuals and some entity in the environment. As people often look at what they attend to and where they intend to act, joint attention is considered fundamental to an understanding of other minds and the interaction with other individuals.

Joint attention plays an important role in numerous socialcognitive processes, including Theory-of-Mind (Tomasello, 1995), perspective taking (Moll \& Meltzoff, 2011), and processes relating to learning and memory from early infancy throughout adulthood (Kim \& Mundy, 2012). However, despite extensive research in virtually all areas of cognitive science aiming at an understanding of behavioral functions, cognitive processes, and neural mechanisms of joint attention, there is a plethora of unresolved questions. The interplay of the development of joint attention and language during infancy (Baldwin, 1995) or the relation between joint attention and the perception of other persons (Frischen, Bayliss, \& Tipper, 2007) are among those. Finally, the neural circuits subserving our ability to engage in joint attention have been investigated only recently because appropriate methods to study gaze-based face-to-face interaction in real-time have only recently been made available (Redcay et al., 2010; Schilbach et al., 2010).

In addition to its role in social cognition, seeing and following or directing someone else's gaze is crucial for effective language learning (Morales et al., 2000) and language processing in adults (Clark \& Krych, 2004; Hanna \& Brennan, 2007; Staudte \& Crocker, 2011). Monitoring each other's gaze behavior supports the understanding of what the interlocutor is saying or understanding (Richardson \& Dale, 2005; Hanna \& Brennan, 2007) and fosters the synchronization of interlocutors in discourse (Garrod \& Pickering, 2004). Thus, initiating or establishing joint attention at a chosen point during dialog can be a powerful means to augment and modulate linguistic content.

Finally, severe impairments in multiple aspects of social
cognition and communication are among the core symptoms of autism spectrum disorders (ASD). Due to the broad impact of joint attention on social and communicative skills, its study has become a major focus in the empirical research on ASD. The majority of this research is dedicated to understanding the implications of mutual and triadic gaze for the development of skills related to communication among typically developing individuals and those with ASD (Mundy, Gwaltney, \& Henderson, 2010; Redcay et al., 2012).

Overall, this workshop aims to explore how traditionally separate research areas such as social cognition/neuroscience, psycholinguistics, human-computer interaction and developmental psychology contribute to an understanding of the general phenomenon of gaze-following and joint attention from all these different perspectives - and how these fields can benefit and learn from each other, e.g. by comparing different approaches and methodologies.

## Relevance to the the CogSci Conference

Recently, there has been an increased interest in psycholinguistics and human-computer interaction as well as in social cognition and ASD research to investigate human communication processes in more interactive settings. In particular, scientists have tried to extend their theories and experimental designs by the visual presence and the induced dynamics of an interaction partner in order to accommodate the complex non-verbal behavior that typically accompanies and greatly influences linguistic interaction. The domain of gaze has aroused particular interest as single acts of looking combine perception and action in social encounters. By establishing joint attention, for instance, gaze behavior guides the exchange of goals and desires which are critical motivations for communication. However, experiments incorporating such complex and dynamic yet crucial aspects of human interaction are difficult to implement and a challenge to traditionally very controlled procedures. In this workshop, we would like to gather researchers from related fields and bring them closer together by providing a platform for exchanging theories and approaches as well as methodology that is suited for investigating the use and effect of gaze in human interaction (e.g. Redcay et al., 2010; Saito et al., 2010; Wilms et al., 2010). All related fields are core areas of cognitive science and our research questions are currently of high interest in the field (as partly reflected also by the invited symposium Joint Action).

## Suitability of the Organizers

Dr. Maria Staudte has a background in psycholinguistics and human-agent-interaction and has published in established journals and conferences such as Cognition, HRI, and the An-
nual Conference of the Cognitive Science Society. Her interests have focused on studying how humans use (each other's or an artificial agent's) eye-movements in order to ground references in a shared environment, to infer (referential) intentions, and to predict upcoming action. Ulrich Pfeiffer has a background in linguistics, psychology, and social neuroscience and studies the behavioral functions and neural correlates of gaze behavior in real-time social interactions using a combination of novel and innovative interactive eyetracking and neuro-imaging methods. He has published in established journals such as Frontiers in Psychology, Social Cognitive and Affective Neuroscience, and PLoS One and wrote a book chapter on eye-tracking methodology. He has further co-edited a Research Topic Issue in Frontiers in Neuroscience titled Towards a Neuroscience of Social Interaction.

## Target Audience \& Participants

The target audience of this workshop are researchers from all subfields of cognitive science that have an interest in the study of gaze behavior in interaction and communication. We expect a large audience of approximately 30-40 participants.

## (Invited/Keynote) Speakers

The following high-profile researchers have confirmed to give keynote lectures:

- Dr. Andrew Bayliss, University of East Anglia, on "Gaze cueing: The influence of observing averted gaze on attention and affective evaluations"
- Prof. Susan Brennan, SUNY Stony Brook, tbd
- Prof. Peter Mundy, University of California at Davis, on "The Interaction of Joint Attention and Communication: Cognitive and Neurocognitive Factors"
- Prof. Elizabeth Redcay, University of Maryland, on "Brain systems supporting joint attention behaviors in typical development and autism"

In addition to the keynote lectures, we solicit submissions of abstracts (around 350 words) related to the workshop topic from all areas of the cognitive sciences. We intend this workshop to last a full day of which four $50-\mathrm{min}$ slots would be dedicated to the keynote lectures. Besides an opening session, we expect to have six more $20-\mathrm{min}$ slots to be filled by speakers based on abstract selection.

## Publicizing and Documentation

We aim to make use of our large network of collaborators and colleagues in order to personally publicize this workshop and solicit submissions from specific individuals and labs. Further, we will use social media like Facebook as well as mailing lists such as AMLaP, LINGUIST or ESAN for general advertisement. We have no plans for documentation of the workshop outcome at this point since we truly view this event as a kick-off event that should help to start discussions, form synergies, and initiate new collaborations.

## Contact Information

Maria Staudte, Saarland University / SUNY Stony Brook, 100 Bleecker St, Apt 2f, New York City, 10012 NY. Email: masta@coli.uni-sb.de

## References

Baldwin, D. A. (1995). Understanding the link between joint attention and language. In C. Moore \& P. Dunham (Eds.), Joint attention: Its origins and role in development. Lawrence Erlbaum Associates, Inc.
Clark, H. H., \& Krych, M. A. (2004). Speaking while monitoring addressees for understanding. Journal of Memory and Language, 50(1), 62-81.
Frischen, A., Bayliss, A. P., \& Tipper, S. P. (2007). Gaze cueing of attention: Visual attention, social cognition, and individual differences. Psychological Bulletin, 133(4), 694-724.
Garrod, S., \& Pickering, M. J. (2004). Why is conversation so easy? Trends in Cognitive Sciences, 8(1), 8-11.
Hanna, J., \& Brennan, S. (2007). Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. Journal of Memory and Language, 57, 596-615.
Kim, K., \& Mundy, P. (2012). Joint attention, social-cognition, and recognition memory in adults. Frontiers in Neuroscience, 6, 172.

Moll, H., \& Meltzoff, A. (2011). Perspective-taking and its foundation in joint attention. In N. Eilan, H. Lerman, \& J. Roessler (Eds.), Perception, causation, and objectivity. issues in philosophy and psychology (p. 286-304). OUP Oxford.
Morales, M., Mundy, P., Delgado, C., Yale, M., Messinger, D., Neal, R., \& Schwartz, H. (2000). Responding to joint attention across the 6-through 24-month age period and early language acquisition. Journal of Applied Developmental Psychology, 21(3), 283-298.
Mundy, P., Gwaltney, M., \& Henderson, H. (2010). Self-referenced processing, neurodevelopment and joint attention in autism. Autism, 14(5), 408-429.
Redcay, E., Dodell-Feder, D., Pearrow, M., Mavros, P., Kleiner, M., Gabrieli, J., \& Saxe, R. (2010). Live face-to-face interaction during fMRI: a new tool for social cognitive neuroscience. Neuroimage, 50(4), 1639-1647.
Redcay, E., O'Young, D., Slevc, L., Mavros, P., Gabrieli, J., \& Sinha, P. (2012). Gaze cues in complex, real-world scenes direct the attention of high-functioning adults with autism. In Proceedings of the 34th Annual Meeting of the Cognitive Science Society.
Richardson, D. C., \& Dale, R. (2005). Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. Cognitive Science, 29(6), 1045-1060.
Saito, D., Tanabe, H., Izuma, K., Hayashi, M., Morito, Y., Komeda, H., ... others (2010). Stay Tuned: Inter-Individual Neural Synchronization During Mutual Gaze and Joint Attention. Frontiers in Integrative Neuroscience, 4, 127.
Schilbach, L., Wilms, M., Eickhoff, S., Romanzetti, S., Tepest, R., Bente, G., ... Vogeley, K. (2010). Minds made for sharing: initiating joint attention recruits reward-related neurocircuitry. Journal of Cognitive Neuroscience, 22(12), 27022715.

Staudte, M., \& Crocker, M. W. (2011). Investigating joint attention mechanisms through human-robot interaction. Cognition, 120(2), 268-291.
Tomasello, M. (1995). Joint attention as social cognition. In C. Moore \& P. Dunham (Eds.), Joint attention: Its origins and role in development. Lawrence Erlbaum Associates, Inc.
Wilms, M., Schilbach, L., Pfeiffer, U., Bente, G., Fink, G., \& Vogeley, K. (2010). Its in your eyes - using gaze-contingent stimuli to create truly interactive paradigms for social cognitive and affective neuroscience. Social cognitive and affective neuroscience, 5(1), 98-107.

# Using Complex Network Analysis in the Cognitive Sciences 

Nicole M. Beckage (Nicole.Beckage@ Colorado.edu)<br>University of Colorado, Boulder<br>Department of Psychology and Neuroscience, 345 UCB<br>Boulder, CO 80309 USA<br>Michael S. Vitevitch (mvitevit@ku.edu)<br>University of Kansas<br>Department of Psychology, 1415 Jayhawk Blvd<br>Lawrence, KS 66045 USA<br>Alexander Mehler (mehler@em.uni-frankfurt.de)<br>Goethe University, Frankfurt<br>Department of Computer Science and Mathematics, Robert-Mayer Straße 10<br>Frankfurt am Main, 60325 Germany<br>Eliana Colunga (Eliana.Colunga@Colorado.edu)<br>University of Colorado, Boulder<br>Department of Psychology and Neuroscience, 345 UCB<br>Boulder, CO 80309 USA

Keywords: Network analysis, graph theory, complex systems, network descriptives, R , language

## Objectives and Scope of the Tutorial

- Provide an elementary introduction to network analysis as a tool within cognitive science, using examples from the domain of language.
- Demonstrate how to import, manipulate, and analyze network data using the R programming language.
- Participants who complete the tutorial will be able to perform basic network analyses, and use this powerful suite of analyses to examine relational data in their own domains of research.


## Tutorial Delivery

The general format of the tutorial will be a half-day introduction to research in the field of complex network analysis followed by a more detailed study of a specific research project on language acquisition. In the course of the more detailed study participants will have an opportunity to perform some statistical and hypothesis testing on networks while learning interpretations and meaning of network analysis techniques.

We will begin by introducing a few research findings that are specific to network analysis. These include results showing that there are structural network differences that can be quantified and compared between groups as well as examples of conclusions that readily emerge from a networking framework that would otherwise be difficult to capture. These require a fundamental understanding of a variety of network descriptives that will be defined and applied to the research questions at hand. For example, in capturing and explaining structural network differences, we introduce the idea of clustering coefficients and geodesic distance. These network descriptives have become specifically relevant to the field, as they have given rise to
the idea of small-world structure, which has been shown to allow for efficient processing and navigation of information. From there we introduce the idea that network statistics change with the size and density of a graph. That brings up concepts of randomization and statistical tests. While these will be handled initially as definitions and concepts, the second part of the tutorial will include working through the analyses that were conducted to yield the research results. We will conclude the tutorial by allowing participants to design their own hypothesis tests and help with refining individualized research goals in light of network theory. If time and interest permits we will also consider process models of networks, inference on missing data and missing link information as well as network-based algorithms.

We will teach participants how to do basic network calculations with built-in functions of R as well as help develop an intuitive understanding of network models. At the end of the session we will also introduce the idea of network process models specifically looking at preferential attachment and page rank algorithms.

## Instructor experience with Network Analysis

Nicole Beckage is a graduate student majoring jointly in Cognitive Science and Computer Science at University of Colorado Boulder. Nicole has spent most of her course work and research focusing on network analysis and language acquisition. She has helped run network tutorials at the Institute of Social Network Analysis's annual conference and has taken many classes in a variety of disciplines with network theory as a main topic. Her research has utilized and designed novel network approaches and she has been invited to give network related talks at many workshops and conferences. The focus of the methodological tutorial will be motivated by the techniques of her paper entitled 'Small worlds and semantic network growth in typical and late talkers' published in PLOS One in 2011.

Michael Vitevitch is an Associate Professor in the Department of Psychology at the University of Kansas. Prof. Vitevitch combines the analytic tools of Network Science with conventional psycholinguistic tasks to better understand the processes and representations involved in spoken word recognition. His work in this area has appeared in several mainstream Psychology journals (e.g., Cognitive Science, Journal of Memory \& Language, Journal of Experimental Psychology: Human Perception \& Performance) as well as in journals that focus on the topics of chaotic, complex, and nonlinear systems (e.g., Entropy, and International Journal of Bifurcation and Chaos). In addition to organizing a satellite conference on the topic of "Language and Network Science" at the 2012 NetSci conference in Chicago, he has been invited to present his network research at a number of international workshops and conferences.
Alexander Mehler is professor for Computational Humanities at the Goethe University Frankfurt am Main, Germany, where he heads the Text technology Lab as part of the Department of Computer Science and Mathematics. He is a member of the executive committee of the LOEWE Priority Program "Digital Humanities" at Frankfurt University. His research interests include empirical analysis and simulative synthesis of discourse units in spoken and written communication. He aims at a quantitative theory of networking in linguistic systems to enable multi-agent simulations of life cycles. He integrates models of semantic spaces with simulation models of language evolution and topological models of network theory to capture the complexity of linguistic systems. He heads several research projects on the analysis of linguistic networks. His work has appeared in several journals in the area of computational linguistics, cognitive science and complex systems (Neural Networks, Journal of Quantitative Linguistics, Computer Speech and Language, Entropy, Applied Artificial Intelligence). In 2012, he organized a conference on Modeling Linguistic Networks based on which he will coedit the first (Springer) volume on linguistic networks in cognitive science and related disciplines.
Eliana Colunga is an Associate Professor in the Departments of Psychology and Neuroscience and Computer Science and a Fellow of the Institute of Cognitive Science at the University of Colorado Boulder. Prof. Colunga studies interactions between language and cognition using crosslinguistic, developmental and computational modeling methods. Her work on computational models of language development has been published in journals such as Psychological Review, Cognition, Developmental Science, and has been funded by the John Merck Fund and the Eunice Kennedy Shriver National Institute of Child Health \& Human Development of the National Institute of Health. She received her PhD in Computer Science and Cognitive Science from Indiana University and her MS in Artificial Intelligence and BS in Computer Science from the Instituto Tecnologico y de Estudios Superiores de Monterrey, Mexico.

## Why Network Analysis at Cognitive Science?

Over the last few decades the work on network analysis has been revived and expanded with new analytical, numerical and theoretical approaches. It has become a fundamental force within a variety of fields from physics, computer science and psychology to sociology and political science. The types of questions many cognitive scientists ask, such as studying the structure of language, studying group dynamics or neuronal dynamics, can be framed within a network perspective and we hope by building a team of researchers who work with network analysis as their main framework we can excite others in the field to utilize these techniques. Further, utilizing a tutorial structure will allow for us not only to explain our research findings but also give others the tools they need in order to begin answering their own questions within this framework. With this in mind, we expect the audience of this tutorial to be interested in learning about network analysis for any purpose.

## Likely Audience for the Tutorial

Because our expertise is mostly tied to language many participants may be interested specifically in language, but the application of this method extends beyond language. This tutorial is specifically useful for types of relational data. Though the material covered in this workshop will be relatively basic, we hope to give participants a flavor for the strength and power of network analysis techniques.

## Special Requirements for the Tutorial

The participants will be asked to bring laptops to the meeting as well as have R and a few select libraries (statnet, sna package and network package) installed. A .zip file will be available with other necessary files for the completion of the tutorial material. By working through past research findings, participants will receive an overview of basic network functions in R and have the opportunity to perform cognitively meaningful network data manipulation. We will also introduce visualization techniques of network data. We perform statistical tests to understand the structure present in the observed network compared to what might happen under different conditions. This requires calculations of network descriptives (e.g., centrality scores, graph-level indices); and use of classical network analytic techniques and network specific statistical tests.

No prior experience with R is necessary and attendees do not need to have a familiarity with the basic concepts of descriptive network analysis.

## Acknowledgments

This tutorial was made possible with generous help from Carter Butts at University of California, Irvine, Ryan Acton at University of Massachusetts Amherst and Lorien Jasney at the University of California, Davis who contributed previous tutorial materials for adaption.

# A General Purpose Architecture for Building Spiking Neuron Models of Biological Cognition 

Chris Eliasmith (celiasmith@uwaterloo.ca)<br>Terrence C. Stewart (tcstewar@uwaterloo.ca)<br>Center for Theoretical Neuroscience, University of Waterloo 200 University Ave West, Waterloo, ON, N2L 3G1, Canada

Keywords: SPAUN, cognitive modeling; neural engineering; representation; decision making; working memory; cognitive architecture; cognitive control; semantic pointers

## Tutorial Objectives

We have recently created the world's largest biologically realistic brain model that is capable of performing tasks (Eliasmith et al., 2012). This model uses 2.5 million spiking neurons, takes visual input from a $28 \times 28$ pixel visual field, and controls a physically modelled arm. By presenting different visual inputs, the model can perform eight different tasks, including memorizing and writing a list of numbers, single-digit addition via counting, and flexible pattern completion in the Raven's Matrices task. This tutorial is meant to introduce the software toolkit and theoretical background that would allow other researchers to build their own models using the same architecture, allowing them to explore other tasks and brain functions. This tool supports a novel cognitive architecture (SPA; the Semantic Pointer Architecture) that directly connects neuroscience with cognitive science.
Our previous tutorials have focused on the underlying theory of the Neural Engineering Framework (NEF; Eliasmith and Anderson, 2003), a general method for implementing high-level cognitive theories using biologically realistic spiking neurons. In this tutorial, our emphasis will be on building large-scale models with our open-source toolkit Nengo ([http://nengo.ca](http://nengo.ca)). The tutorial will be the first presentation of our Semantic Pointer Architecture, a Python module for Nengo which takes a high-level description of the desired cognitive system, including (basic) visual processing, motor control, working memory, associative memory, and cognitive control. The software takes this specification and creates a biologically realistic neural model, including various cortical areas, the basal ganglia, and the thalamus.

An example model using the SPA is shown in Figure 1. It is able to follow basic commands such as "WRITE TWO" and "REMEMBER THREE <long pause> WRITE NUMBER". When run in Nengo, this creates a model with 48,000 spiking neurons and produces predictions of spike patterns, firing rates, fMRI time-courses, accuracy, and reaction times. Complete details can be found in the book How to Build a Brain (Eliasmith, 2013).
Participants will leave the tutorial having interactively used a method for constructing cognitive models with spiking neurons, and experience using that method in an intuitive software environment.

```
class Rules:
    def read_action(category='ACTION'):
        set(action=vision*2)
    def read_object(category='OBJECT'):
        set(object=vision*2)
    def do_write(vision='DONE',
            phrase='ACTION*WRITE', scale=0.5):
        set(motor=phrase*'~OBJECT')
    def do_write_remembered(vision='DONE'
            phrase='ACTION*WRITE+OBJECT*NUMBER' ) :
        set(motor=memory)
    def do_remember(vision='DONE',
            phrase='ACTION*REMEMBER', scale=0.5):
        set(memory=phrase*'~OBJECT')
    class Parser(SPA):
    vision = Vision()
    category = Buffer(feedback=0)
    action = Buffer(feedback=0)
    object = Buffer(feedback=0)
    actionC = Cleanup(mutual_inhibit=0.5)
    objectC = Cleanup(mutual_inhibit=0.5)
    phrase = Buffer(feedback=0)
    motor = Motor()
    memory = Buffer(pstc_feedback=0.1)
    flow = Flow("""
        action->actionC
        object->objectC
        actionC*1.1->action
        objectc*1.1->object
        action*ACTION->phrase
        object*OBJECT->phrase
        vision.WRITE->category.ACTION
        vision.REMEMBER->category.ACTION
        vision.ONE->category.OBJECT
        vision.TWO->category.OBJECT
        vision.THREE->category.OBJECT
        vision.NUMBER->category.OBJECT
        """)
    BG=BasalGanglia(Rules())
    thal=Thalamus(BG)
```



Figure 1: A script (top) to generate a model with 48,000 spiking neurons (bottom left) capable of simple cognitive behaviour (bottom right)

## Tutorial Structure

This full-day tutorial starts with a quick overview of the theory behind the Neural Engineering Framework, showing how we can specify models by solving for the synaptic connection weights between groups of neurons that cause particular computations to be approximated. By specifying what each neural group represents and what computations should be computed between neural groups, large-scale neural models can be created. This is paired with many hands-on example of applying these concepts using the Nengo software. Neural groups and connections can be created using a drag-and-drop interface, or using the Python scripting interface. Participants are expected to bring a laptop to follow along with these tutorials (Windows, OS X, and Linux are all supported, and software is provided).

The first half of the tutorial covers the basic principles of the NEF and using Nengo. The second half introduces the concept of semantic pointers (vectors that combine the benefits of semantic similarity measures with the compositionality of symbol structures). We show how this method provides a unified approach to many types of cognitive models, including perceptual processing, symbolic reasoning, and motor control models. In particular, we show how these representations can be used to bind and manipulate symbol-like structures.

In this second half of the tutorial, we introduce our semantic pointer architecture and its implementation within Nengo. This provides a scripting language for building cognitive models, allowing researchers to create models of different brain areas and connect them together via the cognitive control of a cortex-basal ganglia-thalamus loop. This system is a general-purpose, biologically constrained, and neurally plausible cognitive architecture implemented using spiking neurons, and is the core foundation of our large-scale brain simulation.

Variants of this tutorial without the SPA cognitive architecture were presented at ICCM 2009, CogSci 2010, Telluride 2011, CogSci 2011, and CogSci 2012.

## Tutorial Justification

The Neural Engineering Framework provides a method to bridge the gap between cognitive and neural theories. It has been used to build special-case models of serial-order recall (Choo \& Eliasmith, 2010), action selection in the basal ganglia (Stewart, Choo, \& Eliasmith, 2010), the Wason card task (Eliasmith, 2005), the Tower of Hanoi task (Stewart \& Eliasmith, 2011), and inductive rule generation (Rasmussen \& Eliasmith, 2010).

By combining these models with a general method for cognitive control (the cortex-basal ganglia-thalamus loop), we have created the beginnings of a novel biologically realistic cognitive architecture. Our tools allow researchers to quickly create large-scale brain models that combine novel models of particular brain areas with existing models of many other areas. We believe this is a powerful new tool for understanding cognition.

Given that understanding the SPA cognitive architecture requires a basic understanding of the Neural Engineering Framework, we feel that a full-day tutorial is required. The NEF provides an exciting new tool for cognitive science, as it provides a technique for producing direct neural predictions from cognitive theory. It is also general enough that the same framework can be used for category learning, memory research, linguistic processing, vision, and motor control research. Furthermore, components from existing cognitive architectures (such as ACT-R) can be directly interfaced to these neural models.

## Audience

Participants are not expected to have any previous experience with neural modeling. All participants are encouraged to bring a laptop for installing Nengo (Linux, OS X, and Windows versions are provided), allowing for hands-on interactions with the models discussed.

## Presenters

Chris Eliasmith holds a Canada Research Chair in Theoretical Neuroscience, and is director of the Centre for Theoretical Neuroscience at the University of Waterloo. His recent book, How to Build a Brain, and his earlier book, Neural Engineering, form the basis for this tutorial.

Terry Stewart is a research associate in the Centre for Theoretical Neuroscience. He developed the SPA module and has used it to create a variety of cognitive models.

## References

Choo, F., \& Eliasmith, C. (2010). A Spiking Neuron Model of Serial-Order Recall. 32 ${ }^{\text {nd }}$ Annual Conference of the Cognitive Science Society.
Eliasmith, C. (2005). Cognition with neurons: A large-scale, biologically realistic model of the Wason task. $27^{\text {th }}$ Annual Meeting of the Cognitive Science Society.
Eliasmith, C. (2013). How to build a brain: A neural architecture for biological cognition. New York, NY: Oxford University Press.
Eliasmith, C., \& Anderson, C. (2003). Neural Engineering: Computation, Representation, and Dynamics in Neurobiological Systems. Cambridge: MIT Press.
Eliasmith, C., Stewart T.C., Choo X., Bekolay T., DeWolf T., Tang Y., \& Rasmussen, D. (2012). A large-scale model of the functioning brain. Science. 338(6111), 1202-1205.
Rasmussen, D., \& Eliasmith, C. (2010). A neural model of rule generation in inductive reasoning. $32^{\text {nd }}$ Annual Conference of the Cognitive Science Society.
Stewart, T.C., \& Eliasmith, C. (2011). Neural Cognitive Modelling: A Biologically Constrained Spiking Neuron Model of the Tower of Hanoi Task. 33 ${ }^{\text {rd }}$ Annual Conference of the Cognitive Science Society.
Stewart, T.C., Choo, X., \& Eliasmith, C. (2010). Dynamic Behaviour of a Spiking Model of Action Selection in the Basal Ganglia. $10^{\text {th }}$ Int. Conf. on Cognitive Modeling.

# Virtual Humans: A New Toolkit for Cognitive Science Research 

Jonathan Gratch (gratch@ict.usc.edu), Arno Hartholt (hartholt@ict.usc.edu), Morteza Dehghani (morteza@ict.usc.edu), Stacy Marsella (marsella@ict.usc.edu)<br>Institute for Creative Technologies, University of Southern California<br>12015 Waterfront Drive, Playa Vista, CA 90094, USA

Keywords: Virtual humans, embodied cognition, social cognition, virtual confederates

## Tutorial Objectives

Virtual humans (VHs) are digital anthropomorphic characters that exist within virtual worlds but are designed to perceive, understand and interact with real-world humans. Although typically conceived as practical tools to assist in a range of application (e.g., HCI, training and entertainment), the technology is gaining interest as a methodological tool for studying human cognition. VHs not only simulate the cognitive abilities of people, but also many of the embodied and social aspects of human behavior more traditionally studied in fields outside of cognitive science. By integrating multiple cognitive capabilities (e.g., language, gesture, emotion, and the control problems associated with navigating and interacting with a simulated virtual world) and requiring these processes to support real-time interactions with people, VHs create a unique and challenging environment within which to develop and validate cognitive theories. In this tutorial, we will review recent advances in VH technologies, demonstrate examples of use of VHs in cognitive science research and provide hands on training using our Virtual Human Toolkit (http://vhtoolkit.ict.usc.edu/).

## Virtual Humans and Cognitive Science

In helping to define the field of cognitive science, Herb Simon emphasized the importance of "understanding by simulating" (Simon 1969, 17-22). From the perspective of cognitive science, VHs provide the opportunity to understand the mind by simulating the body. Although still limited in their capabilities, VHs combine a rich set of capabilities for exploring how cognitive processes manage interactions with the physical and social world. In this sense, they complement recent interest in robotics as a tool for cognitive science, and address many of the limitations of physical robots.

Embodied Cognition: Embodied theories argue that the brain and body are tightly linked: the configuration and state of one's body profoundly influence cognitive processes and vice versa. For example, posture can impact how easily we are persuaded (Petty, et al. 1983); gestures and language are closely coupled, often grounded in shared metaphors (McNeill 2005); and facial expressions can influence our emotions (Niedenthal, et al., 2010). VH technology is increasingly used to unpack this relationship between mind and body (e.g., Sprague, et al., 2007) and in the tutorial we will review research in this growing area.

Social Cognition: People interact socially through their bodies and VHs allow researchers to systematically examine and model the cognitions underlying social interaction. VHs can act as "virtual confederates" (Blascovich et al., 2002), allowing systematic manipulation of visual appearance, speech type, and contextual graphical environments. This makes VHs a convenient platform to isolate unique sociocultural characteristics and realize them through simulation. Along with enhanced experimental control, ease of manipulations, consistency and controlled measurements, these features make VHs useful and reliable tools for studying social cognitions. In the tutorial, we will review several examples, including how expressions of emotion by VHs can influence decision making in negotiations tasks and social dilemmas (e.g. de Melo et al., 2012; Dehghani et al., 2012); the role of accent in cultural cognition (Dehghani et al., 2012) and the role of rapport and gender in enhancing participants' performance (Karacora et al., 2012).

## The Virtual Human Toolkit

The University of Southern California's Institute for Creative Technologies (ICT) is recognized as a leader in the development of VH technology (Gratch et al., 2002) and in applying this research to application domains including "virtual role players" for interpersonal-skills training (e.g., Campbell et al. 2011), informal science education (e.g. Swartout et al., 2010), intelligent tutoring, (e.g. Lane et al., 2011), and as "virtual confederates" to study cognitive and social processes (e.g. de Melo et al., 2012; Dehghani et al., 2012). One goal of the institute is to foster research in VH by making this technology freely available for research purposes through the Virtual Human Toolkit.

The research underlying the toolkit draws heavily on cognitive science research. For example, VH "brains" are inspired by psychological theories of human cognition (e.g. Swartout, Gratch et al., 2006), language (e.g. Traum, 2008) and emotion (Gratch \& Marsella, 2005), VH bodies are informed by knowledge of physiological and biomechanical processes (e.g. Honglun, et al. 2007; Thiebaux et al., 2008) and the relation between the VH's brain and body is informed by social psychology research (Lee \& Marsella, 2006, Wang et al, 2013). Translating these theories and findings into working software requires the integration of advanced capabilities from a number of domains of computer science research including machine perception, artificial intelligence, cognitive modeling, graphics and animation.

The complexity of creating a VH can appear daunting. Fortunately, considerable research has focused on the de-
velopment of modular, sharable software architectures to facilitate application development. The Virtual Human Toolkit is a general-purpose collection of integrated VH capabilities, including speech recognition, natural language processing, perception, and nonverbal behavior generation \& execution. The goal of the Virtual Human Toolkit is to make creating VHs easier and more accessible, and thus expand the realm of VH research as well as other research areas, including cognitive science.

In this tutorial, participants will have the opportunity to get hands-on experience with the Toolkit, with the intent to create a basic virtual confederate. In particular, participants will be able to select a character from a library; place the character in a scene; author a set of lines for the confederate to speak; manipulate its gestures and facial expression; and create a set of experimental stimuli that they can take home.

At the conclusion of the workshop participants should have gained a basic understanding of VHs and their capabilities, of how VHs can be of value in the field of cognitive science in the form of virtual confederates, and of how to utilize several aspects of the Virtual Human Toolkit.

## Audience

The likely audience for this tutorial consists of researchers looking to incorporate VHs in their work. This includes researchers in the fields of human computer interaction, education, social cognition, embodied cognition, language and dialogue among others. All participants are encouraged to bring a laptop for installing the Virtual Human Toolkit for hands-on interactions to build a virtual human.

## Presenters

The presenters combine over 40 years of experience researching virtual human technologies, developing virtual human applications for health and training, as well as using the technology as a methodological tool in the study of human behavior.

## References

Blascovich, J., Loomis, J., Beall, A., Swinth, K., Hoyt, C., \& Bailenson, J. N. (2002). Immersive virtual environment technology as a methodological tool for social psychology. Psychological Inquiry, 13, 103-124.
Campbell, J. C., Hays, M. J., Core, M., Birch, M., Bosack, M., \& Clark, R. E. (2011). Interpersonal and Leadership Skills: Using Virtual Humans to Teach New Officers. In I/ITSEC 2011.
de Melo, C., Carnevale, P., Gratch, J. (2012). The impact of emotion displays in embodied agents on emergence of cooperation with people. Presence: Teleoperators and Virtual Environments Journal, 20(5), 449-465.
Dehghani, M., Gratch, J. \& Carnevale P. (2012). Interpersonal Effects of Emotions in Morally-charged Negotiations. In Proceedings of CogSci 2012.

Dehghani, M., Khooshabeh, P., Huang, L., Nazarian, A. \& Gratch J. (2012). Using Accent to Induce Cultural Frame-Switching. In Proceedings of CogSci 2012.
Gratch, J., Rickel, J., André, E., Cassell, J., Petajan, E., \& Badler, N. (2002). Creating Interactive Virtual Humans: Some Assembly Required. IEEE Intelligent Systems, July/August, 54-61.
Gratch, J. \& S., Marsella (2005). Lessons from Emotion Psychology for the Design of Lifelike Characters. Applied Artificial Intelligence, 19, pp 215-233.
Honglun, H., Shouqian, S., \& Yunhe, P. (2007). Research on virtual human in ergonomic simulation. Computers \& Industrial Engineering, 53(2), 350-356.
Karacora, B., Dehghani, M., Krämer-Mertens, N. \& Gratch, J. (2012). The Influence of Virtual Agents' Gender and Rapport on Enhancing Math Performance. In Proceedings CogSci 2012.
Lane, H. C., Noren, D., Auerbach, D., Birch, M. \& Swartout, W. (2011). Intelligent tutoring goes to the museum in the big city: A pedagogical agent for informal science education. In Proceedings of AIED 2012.
Lee J. and S. Marsella (2006). Nonverbal Behavior Generator for Embodied Conversational Agents, in Proceedings of Intelligent Virtual Agents, 2006, pp. 243-255.
McNeill, D. (2005). Gesture and Thought: University of Chicago press.
Niedenthal, P. M., Mermillod, M., Maringer, M., \& Hess, U. (2010). The simulation of smiles (SIMS) mode: embodied simulation and the meaning of facial expression. Behavioral and brain sciences, 33, 417-480.
Petty, R. E., Wells, G. L., Heesacker, M., Brock, T. C., \& Cacioppo, J. T. (1983). The Effects of Recipient Posture on Persuasion. Personality and Social Psychology Bulletin, 9(2), 209-222.
Simon, H. (1969). The sciences of the Artificial. Cam bridge, MA: MIT Press.
Swartout, W., J. Gratch, et al. (2006). Toward Virtual Humans. AI Magazine 27(1).
Thiebaux, M., A. Marshall, S. Marsella, and M. Kallmann (2008), SmartBody: Behavior Realization for Embodied Conversational Agents, in Proceedings of Autonomous Agents and Multi-Agent Systems.
Traum, D. (2008). Talking to Virtual Humans: Dialogue Models and Methodologies for Embodied Conversational Agents. Modeling Communication with Robots and Virtual Humans. I. Wachsmuth and G. Knoblich. Berlin, Springer: 296-309.
Sprague, N., D. Ballard, et al. (2007). Modeling embodied visual behaviors. ACM Trans. Appl. Percept. 4(2): 11.
Swartout, W., Traum, D., Artstein, R., Noren, D., Debevec, P., Bronnenkant, K., ... \& White, K. (2010). Ada and Grace: Toward realistic and engaging virtual museum guides. In Intelligent Virtual Agents (pp. 286-300). Springer Berlin/Heidelberg.
Wang, Z., Lee, J. and Marsella, S. Multi-party, multi-role comprehensive listening behavior, Journal Of Autonomous Agents and Multi-Agent Systems, 2013.

# Making Robust Classification Decisions: Constructing and Evaluating Fast and Frugal Trees (FFTs) 

Hansjörg Neth, Uwe Czienskowski, Lael J. Schooler<br>\{neth, sciencec, schooler\}@mpib-berlin.mpg.de<br>Center for Adaptive Behavior and Cognition<br>Max Planck Institute for Human Development, Berlin, Germany

Kevin Gluck<br>kevin.gluck@us.af.mil<br>Air Force Research Laboratory<br>Wright-Patterson AFB, OH, USA


#### Abstract

Fast and Frugal Trees (FFTs) are a quintessential family of simple heuristics that allow effective and efficient binary classification decisions and often perform remarkably well when compared to more complex methods. This half-day tutorial will familiarize participants with examples of FFTs and elucidate the theoretical link between FFTs and signal detection theory (SDT). A range of presentations, practical exercises and interactive tools will enable participants to construct and evaluate FFTs for different data sets.


Keywords: Fast and frugal trees; binary classifications; simple heuristics; signal detection theory; validity; robustness

## Motivation

Many real-world problems call for binary classification decisions. We may want to predict whether a partnership is promising, whether an investment is profitable, or whether a patient is in peril. Such classifications have important consequences, yet are typically made under time-pressure and uncertainty. Predictions of experts and laypeople in the real world require robust decision strategies that work swiftly and reliably on the basis of limited information.

Fast and Frugal Trees (FFTs) allow efficient and effective binary classification decisions by sequentially attending to a list of diagnostic cues (Martignon, Vitouch, Takezawa, \& Forster, 2003; Martignon, Katsikopoulos, \& Woike, 2008). FFTs are a special case of simple heuristics - simple decision processes that often perform remarkably well in comparison to more complex methods (Gigerenzer, Todd, \& the ABC research group, 1999; Gigerenzer, Hertwig, \& Pachur, 2011) - and have been linked with the theoretical framework for diagnostic classification decisions provided by signal detection theory (SDT, Luan, Schooler, \& Gigerenzer, 2011).

Figure 1 illustrates an example of a FFT that predicts whether an antibiotic prescription is indicated for some patients, particularly children. By checking only one or two cues physicians can identify patients at risk of being infected with a specific type of bacteria and prescribe an appropriate antibiotic treatment (Fischer et al., 2002). Beyond being both effective and efficient FFTs are useful by virtue of being robust (by being insensitive to perturbations due to noisy data and by providing reliable out-of-sample predictions) and communicable (e.g., they can easily be understood, learned


Figure 1: Example of a FFT that allows clinicians to prescribe treatment with macrolides (see Fischer et al., 2002).
and taught). FFTs have successfully been developed in a variety of applied domains, including medical, legal, and financial decision making (see Luan et al., 2011, for examples).

## Content, Structure, and Activities

This half-day tutorial builds upon the lectures and materials used in previous tutorials (e.g., at the International Conference on Cognitive Modeling, ICCM 2012) and workshops (e.g., at the Max Planck Research School on Adapting Behavior in a Fundamentally Uncertain World, 2012, and the ABC Summer Institute on Bounded Rationality, 2013). Through a combination of presentations and practical exercises participants will become familiar with the theoretical framework behind FFTs, contrast them with alternative classification algorithms, and learn to construct and evaluate FFTs for realworld data sets.

The half-day tutorial interleaves lecture-style presentations with practical exercises and will be structured as follows:

Theoretical background [45 min]: We briefly introduce the basic ideas behind the simple heuristics framework to explain when and why biased minds can make successful inferences. This illustrates how strategies with limited information search can yield robust classification decisions relative to computationally more complex models (Katsikopoulos,

Schooler, \& Hertwig, 2010). Theoretical notions reviewed in this part include the predictive validity of cues, speedaccuracy tradeoffs, the bias-variance dilemma, assessing classification outcomes via contingency tables (hits and correct rejections vs. false alarms and misses), as well as fundamental concepts of SDT (e.g., criterion shifts, bias $c$, the sensitivity index $d^{\prime}$, and the interpretation of ROC curves, Luan et al., 2011). The question How can we make effective and efficient classification decisions on the basis of limited and noisy data? will set the stage for the practical exercises.

Hands-on sessions [2 $\times 45 \mathrm{~min}]$ : Two practical parts will explore the consequences of specific cue and criterion choices. By using interactive software tools participants will acquire hands-on experience in constructing FFTs.

1. Spreadsheet-based FFTs: In a first practical part, participants will be guided through a series of exercises using a pre-designed MS Excel ${ }^{\mathrm{TM}}$ sheet. To facilitate the transfer from theoretical notions to applicable expertise we will examine the consequences of different cue choices, bias values, and criterion shifts on various measures of classification performance. After assessing a selection of minimal FFTs (with only one predictive cue) participants will re-construct a FFT that has been designed to help emergency-room doctors to rapidly decide whether to send a patient with severe chest pain to the coronary care unit (Green \& Mehr, 1997). Finally, particpants will explore alternative multi-cue FFTs and evaluate their performance on a variety of outcome measures.
2. Interactive software tool (FFT-builder): A second practical session will introduce a new version of FFT-builder an interactive software tool that allows rapid-prototyping, explorative learning and the visual inspection of outcome measures in the context of FFTs (see Figure 2). FFT-builder provides a range of features to create and manage environments, data sets, and corresponding FFTs. Numeric and visual analysis tools allow to quantify and compare the performance of different solutions to the same data or explore and inspect the consequences of applying FFTs to different data sets.

Validity and robustness [45 min]: In a final session we will cover two topics central to the theory and practical application of FFTs: their validity and their robustness. The point here is not to merely declare FFTs to be valid and robust, but rather to examine the evidence base and methodological options for addressing these important concerns. Results from cross-validation analyses and a formal quantification and methodological operationalization of robustness will supplement the conceptual discussion.

## Objectives

The goal of this tutorial is to provide participants with intellectual and software tools to tackle real-world classification problems. Upon completing the tutorial, participants will be familiar with theoretical criteria and practical skills for designing efficient, effective, and robust classification algorithms. By building and evaluating a variety of FFTs in an in-


Figure 2: Screenshot of the FFT-builder software tool.
teractive fashion, participants will be enabled and encouraged to apply FFTs to data sets in their own domain of expertise.

## References

Fischer, J., Steiner, F., Zucol, F., Berger, C., Martignon, L., Bossart, W., Altwegg, M., \& Nadal, D. (2002). Use of simple heuristics to target macrolide prescription in children with community-acquired pneumonia. Archives of Pe diatrics \& Adolescent Medicine, 156(10), 1005-1008.
Gigerenzer, G., Hertwig, R., \& Pachur, T. (Eds.). (2011). Heuristics: The foundations of adaptive behavior. New York, NY: Oxford University Press.
Gigerenzer, G., Todd, P. M., \& the ABC research group. (1999). Simple heuristics that make us smart. New York, NY: Oxford University Press.
Green, L., \& Mehr, D. R. (1997). What alters physicians' decisions to admit to the coronary care unit? Journal of Family Practice, 45(3), 219-226.
Katsikopoulos, K., Schooler, L. J., \& Hertwig, R. (2010). The robust beauty of ordinary information. Psychological Review, 117(4), 1259-1266.
Luan, S., Schooler, L. J., \& Gigerenzer, G. (2011). A signaldetection analysis of fast-and-frugal trees. Psychological Review, 118(2), 316-338.
Martignon, L., Katsikopoulos, K. V., \& Woike, J. K. (2008). Categorization with limited resources: A family of simple heuristics. Journal of Mathematical Psychology, 52(6), 352-361.
Martignon, L., Vitouch, O., Takezawa, M., \& Forster, M. R. (2003). Naïve and yet enlightened: From natural frequencies to fast and frugal decision trees. In D. Hardman \& L. Macchi (Eds.), Thinking: Psychological perspectives on reasoning, judgment and decision making (pp. 189-211). Chichester, England: Wiley.

# Half Day Tutorial on Using Quantum Probability Theory to Model Cognition 

Emmanuel M. Pothos (Emmanuel.pothos.1@city.ac.uk)<br>Department of Psychology, City University London, London, EC1R 0JD, UK

Zheng Wang (wang.1243@osu.edu)

School of Communication, Derby Hall,
The Ohio State University, Columbus, OH 43210 USA.

Jerome R. Busemeyer (jbusemey@indiana.edu)
Psychological and Brain Sciences, Indiana University, Bloomington 47468 Indiana, USA.

Keywords: probability theory, Bayesian probability, quantum theory, decision making, probabilistic models.

## General Purpose

This tutorial introduces why and how to build cognitive models using quantum probability (QP) theory. In the tutorial, we will show that QP is inherently consistent with deeply rooted psychological conceptions and intuitions. It offers a fresh conceptual framework for explaining some puzzling empirical findings of cognition, and provides a rich new source of alternative formal tools, compared to classical probability ( CP ) theory, for cognitive modeling.
CP models, including Bayesian models, have had an enormous influence in cognitive science (e.g., Griffiths et al., 2010). Such formal models are appealing for many reasons. First, CP theory provides an integrated, coherent, self-consistent set of principles, which can be flexibly applied in any inductive inference situation. Second, such approaches are more falsifiable. Core principles of CP theory are inter-dependent, and identifying an empirical violation of one principle in a setting could invalidate the applicability of CP theory as a whole in that setting. Third, CP principles are intuitive. In the words of Laplace (1816, cited in Perfors et al., 2011), "probability theory is nothing but common sense reduced to calculation."
However, human cognition often goes against the description and prescription from CP theory. In one of the most influential empirical traditions in cognitive psychology, Kahneman, Tversky, and colleagues have reported persistent, clear violations of CP principles in decision making (e.g., Tversky \& Kahneman, 1974). For example, consider the famous conjunction fallacy. Participants are told of a person, Linda, looking very much like a feminist and unlike a bank teller. Then, they are asked to judge probabilities of some events. Violating CP rules, people think the probability that Linda is a bank teller and a feminist is higher than the probability that she is just a bank teller. According to CP theory, it is a fallacy to think $\mathrm{P}(\mathrm{A}$ and $B)>P(A)$. Importantly, even when we become aware of our "fallacy," we cannot shake off the impression that Linda is indeed more likely to be a bank teller and a feminist, than to be just a bank teller.
Important findings like this have led to intense and extensive controversy about the mechanisms which guide human cognition and decision making. The inspiration for
exploring QP theory in cognitive modeling partly arises as a way to resolve this controversy.

The physical theory of quantum mechanics is a marriage between a framework for how to assign probabilities to events and assumptions regarding the nature of the physical world. We can call the former QP theory (or just quantum theory). Can it be applied outside of physics? The motivation for doing so is twofold. First, QP theory is a highly rigorous framework for probabilistic inference. It has been developed over several decades by some of the most brilliant scientists of all time (e.g., Bohr, Dirac, von Neumann, Planck) and has been intensely scrutinized ever since. Thus, the application of QP theory in cognitive modeling has exactly the same formal advantages as that of CP theory. Second, quantum theory allows us to consider the possible relevance in cognitive modeling of several novel concepts. For example, in quantum theory, a cognitive system can be in a superposition state. This means that relative to a question or measurement, the system is in an indefinite state, with all definite states having potential to be expressed. This provides an intrinsic formal representation of the conflict, ambiguity, or uncertainty that people experience in cognitive processes. For another example: states can be entangled, which means a change in one part of the system inexorably and instantaneously affects another part. Entanglement is a form of extreme association, which can be helpful for formalizing important cognitive processes, such as holism, cognitive dissonance, and social projection.

Fundamental quantum conceptions, such as superposition, entanglement, interference, and complementarity, have no formal counterparts in cognitive theory. We are part of a growing group of researchers who have been intensely exploring their applicability in understanding human cognition. Quantum theory reveals alternative intuitions in probabilistic models of cognition. The quantum cognition research program aims to explore whether these alternative intuitions can explain paradoxical findings in decision, memory, and other areas of cognitive processing.

The tutorial introduce the basic principles of quantum theory, in the context of well-known empirical findings in psychological literature. The basic elements of QP theory will require only some knowledge of linear algebra. No background in physics or quantum theory is assumed. The tutorial will be self-contained. It will show how probability
computations can be carried out in quantum theory, how one can build quantum cognitive models, and what the nature of probabilistic intuition is in such models. The tutorial will be useful to all researchers interested in modeling cognition.

## Previous Tutorials and Symposia

Similar tutorials have been presented regularly at the CogSci meetings in Nashville (2007), Washington DC (2008), Amsterdam (2009), and Sapporo (2012), and the Society of Mathematical Psychology meeting (2012). Around $30-50$ participants attended each of the tutorials, with an increasing number of attendees over the years. We have been invited to present short workshops at various universities, such as University of Osnabruck, university of Cincinnati, and Cornell University. At the 2011 CogSci meeting, we co-organized a symposium covering recent progress in the quantum cognition research program. Other tutorials were organized for the annual meetings of Quantum Interaction (since 2009; about 40 participants).

## Presenters

The main presenters, Pothos and Wang, have both contributed extensively to the quantum cognition research program. They both have multiple publications on quantum cognitive models in psychological journals targeting a broad audience. Their presentation will be rigorous, clear, relevant, and accessible. Notably, Pothos has recently coauthored a Behavioral \& Brain Sciences target article, summarizing progress with the quantum cognition research program. Wang has co-edited a special issue of Topics in Cognitive Science that synthesizes current research on quantum cognitive models. Also, both Pothos and Wang have good experience with traditional cognitive models and are currently associate editors for the Frontiers in Cognitive Science journal. Finally, Busemeyer is one of the pioneers of the quantum cognition research programme and has extensive relevant publication and editorial experience.

## Material to be covered

The tutorial will be organized in two parts: (1) an introduction to the key concepts and mathematical modeling tools in QP theory; and (2) an overview of successful cognitive applications, with concrete examples of cognitive models and corresponding MATLAB codes.
In the first part, we will provide a working definition of QP theory. What is it? Why should it be relevant to a cognitive scientist? What are its main characteristics in comparison to CP theory? We will then introduce the basic elements of QP theory (state vector, Hilbert spaces, how to compute simple and conjunctive probabilities) using simple illustrative models of well-known decision and judgment fallacies. We will explain the differences in how probability is computed in the classical vs. quantum way and how these differences give rise to QP theory's unique properties (superposition, incompatibility, interference).

An important question we will address is: is it possible to achieve some sort of isomorphism between (limited cases of) QP and CP models and, if yes, at what price?

We will then introduce structured representations and the idea of entanglement, another unique feature in QP. Time evolution in quantum models will be compared with time evolution in classical models and we will discuss how interference effects can arise in the former, but not the latter, correspondingly leading to violations of the law of total probability, or not.

In second part, we will review successful applications of QP to explain puzzling empirical results in human cognition and decision. We will present some simple MATLAB code illustrating the implementation of QP models in example situations. Perhaps contrary to the common impression of being mysterious and difficult, quantum cognitive models are intuitive. They can be very simple as well, based mostly on linear algebra. We will focus on recent quantum cognition work on probabilistic judgment, measurement order effects, memory, and conceptual combination. What these areas have in common is that they all led to empirical insights which have been hard to reconcile with a CP perspective. Yet, as we will discuss, the unique properties of QP have enabled natural, compelling, and falsifiable accounts of these empirical results. Finally, the tutorial will outline directions for future research.

## References

Busemeyer, J. R., Pothos, E. \& Franco, R., Trueblood, J. S. (2011). A quantum theoretical explanation for probability judgment 'errors'. Psychological Review, 118, 193-218.
Busemeyer, J. R., Wang, Z., \& Townsend, J. T. (2006). Quantum dynamics of human decision making. Journal of Mathematical Psychology, 50, 220-241.
Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., \& Tenenbaum, J. B. (2010). Probabilistic models of cognition: exploring representations and inductive biases. Trends in Cognitive Sciences, 14, 357-364.
Perfors, A., Tenenbaum, J. B., Griffiths, T. L., \& Xu, F. (2011). A tutorial introduction to Bayesian models of cognitive development. Cognition, 120, 302-321.
Pothos, E. M. \& Busemeyer, J. R. (in press). Can quantum probability provide a new direction for cognitive modeling? Behavioral \& Brain Sciences.
Pothos, E. M. \& Busemeyer, J. R. (2009). A quantum probability explanation for violations of "rational" decision theory. Proceedings of the Royal Society B, 276, 2171-2178.
Tversky, A. \& Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185, 11241131.

Wang, Z., Busemeyer, J. R., Atmanspacher, H., \& Pothos, E. (in press). The potential of using quantum theory to build models of cognition. Topics in Cognitive Science.
Wang, Z., \& Busemeyer, J. R. (in press). A quantum question order model supported by empirical tests of an $a$ priori and precise prediction. Topics in Cognitive Science.

# Types and states: Mixture and hidden Markov modelling for the cognitive sciences 

Maarten Speekenbrink (m.speekenbrink @ucl.ac.uk)<br>Cognitive, Perceptual and Brain Sciences, University College London<br>Gower Street, London WC1E 6BT, United Kingdom<br>Ingmar Visser (i.visser@uva.nl)<br>Department of Developmental Psychology, University of Amsterdam<br>Weesperplein 4, 1018 XA Amsterdam, The Netherlands

Keywords: Mixture model, hidden Markov model, latent class analysis, categories, types, states, state transitions

## Objectives and scope

There are many situations in which one may encounter distinct types of entities, such as different animal species, and different states in which these entities may exist, for example motivational states like hunger. Cognition is sometimes also best understood in terms of discrete types and states. For example, aspects of cognitive development can be characterised as the acquisition of increasingly complex rules constituting different types of reasoning (Jansen, Raijmakers, \& Visser, 2007). And rather than a gradually shifting trade-off, people may switch rapidly between distinct decision-making modes favouring either speed or accuracy (Dutilh, Wagenmakers, Visser, \& van der Maas, 2011). The idea that cognitive processes are guided by qualitatively different strategies underlies a wide range of theories of word recognition, cognitive development, categorization, and decision making, to name but a few topics (for an overview, see e.g. Scheibehenne, Rieskamp, \& Wagenmakers, 2013).

As the identity of cognitive types and states is generally not directly observable, appropriate statistical techniques are required to identify them. This tutorial will focus on mixture models (MMs) and hidden Markov models (HMMs), which are the foundation of such techniques. In MMs, a type or state (e.g., a cognitive strategy) is formalized as a probability distribution over observables. Because a dataset may contain different types, the overall distribution is a mixture of such individual component distributions. As the component distributions need not be of the same parametric family (e.g., Gaussian distributions can be mixed with other distributions), MMs allow for considerable flexibility in the definition of types and states. HMMs are a natural extension of MMs, allowing switches between states over time, making them particularly useful when people can switch between cognitive strategies during a task. In addition to identifying the different states, HMMs allow one to also focus on the process underlying state transitions.

While MMs and HMMs are widely used in fields such as computational biology (e.g., for DNA sequence analysis) and machine learning (e.g., for speech recognition and text classification), their use in the analysis of cognition and behaviour is relatively rare. This is unfortunate, as MMs and HMMs are ideally suited to test and explore important theoretical ideas in
cognitive science. The objective of this tutorial is to provide researchers in cognitive science with an accessible introduction to MMs and HMMs and provide them with the necessary skills to apply them in their own research.

## Outline of the tutorial

The tutorial is divided into two parts. The first part introduces the theory behind MMs and HMMs. The second part will be more practical, using a number of examples to show (a) how to apply MMs and HMMs with user-friendly and freely available software, (b) how to interpret these models, and (c) how the models can reveal aspects in the data which remain hidden with more traditional analyses. The first part of the tutorial will be delivered as a classroom style lecture. The second part will use a more hands-on approach with practical computer-based examples and exercises. The audience is encouraged to bring a laptop; all necessary software and material will be made available in advance.

## Part I: Theory

Introduction to mixture models. This part will introduce the basic structure of MMs and the use of graphical and other techniques to determine whether MMs might be applicable.

Estimation. This part will provide an intuitive treatment of maximum likelihood estimation and introduce numerical optimization and Expectation-Maximization (EM), the two main methods for this type of estimation of MMs and HMMs. Practical issues such as starting values and local maxima will also be discussed.

Inference. This part will discuss methods for model selection and how to determine the number of components (i.e., types, or latent classes). We will also discuss methods to test parameters for significance and the use of posterior probabilities to determine the component to which a data point belongs.

Hidden Markov models. This part will introduce hidden Markov models as a direct extension of mixture models. We will then discuss how to generalize the previously discussed methods of estimation and inference to these models.

## Part II: Practice

Introduction to depmixS4 This part will introduce depmixS4 (Visser \& Speekenbrink, 2010), a flexible package
to estimate MMs and HMMs in the R environment for statistical computing ( R Development Core Team, 2010). The examples in the remainder of the tutorial will mainly use this package.

Examples of mixture models Examples will include the use of MMs to detect developmental stages in the liquid conservation task and the use of MMs to detect multiple learning strategies in a category learning task.

Examples of hidden Markov models Examples will include the use of HHMs to analyse speed-accuracy trade-offs and the use of HMMs to model response strategies in multiple cue learning.

Extensions This part will briefly discuss some extensions to basic MMs and HMMs, including the use of covariates to predict the identity of mixture components and states. We will also briefly discuss the use of Bayesian methods to estimate MMs and HMMs.

## About the organizers

The organizers are the developers of depmixS4 (Visser, Jansen, \& Speekenbrink, 2010), a popular R package to estimate mixture and hidden Markov models. They are also the authors of an upcoming book on this topic (commissioned by Springer for their "UseR" series) and a recent tutorial on hidden Markov models (Visser, 2011). The organizers have extensive experience in the application of MMs and HMMs to research in developmental and cognitive science (e.g., Speekenbrink, Lagnado, Wilkinson, Jahanshahi, \& Shanks, 2010; Visser et al., 2010). They can draw upon this experience to provide the audience with real examples and practical advice relevant to a cognitive science audience.

## Justification

Theories which propose the existence of distinct types and states are widespread in the cognitive sciences. Traditional statistical analysis, such as t-tests and ANOVAs, or not applicable to test such ideas. MMs have been successfully used to test "toolbox models" of cognition (e.g., Scheibehenne et al., 2013) and HMMs to test discrete strategy switches (e.g. Speekenbrink et al., 2010; Jansen et al., 2007). This tutorial will provide cognitive scientists with the intuitive understanding and practical knowledge of these models necessary to apply them to their own research.

## Intended audience

This tutorial will be mainly introductory and no specific prior background knowledge is required. While basic knowledge of probability and statistics will be helpful, treatment of the theoretical concepts will largely be conceptual. Familiarity with the $R$ environment will be helpful in general, but by making the commands and code available, previous experience is no requirement to follow and replicate the results of the practical examples.

## Requirements

Participants would ideally bring a laptop to the tutorial. The required software ( R and depmixS4) is open source and freely and easily obtainable. $R$ is available for all major platforms and can be downloaded from http://www.r-project.org. The depmixs4 package can be downloaded from http://cran.r-project.org/web/packages/depmixS4/ or installed from within $R$ through the command install.packages('depmixS4').

## Contact details

Maarten Speekenbrink, Cognitive, Perceptual and Brain Sciences, University College London, Gower Street, London WC1E 6BT, England, UK. Tel: +44 207679 8548. Email: m.speekenbrink@ucl.ac.uk

Ingmar Visser, Department of Developmental Psychology. Weesperplein 4, 1018 XA Amsterdam, The Netherlands. Tel: +31 20 5256723, Email: I.Visser@uva.nl.

## References

Dutilh, G., Wagenmakers, E.-J., Visser, I., \& van der Maas, H. L. J. (2011). A phase transition model for the speedaccuracy trade-off in response time experiments. Cognitive Science, 35, 211-250.
Jansen, B. R. J., Raijmakers, M. E. J., \& Visser, I. (2007). Rule transition on the balance scale task: A case study in belief change. Synthese, 155(2), 211-236.
R Development Core Team. (2010). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org (ISBN 3-900051-07-0)
Scheibehenne, B., Rieskamp, J., \& Wagenmakers, E.-J. (2013). Testing adaptive toolbox models: A Bayesian hierarchical approach. Psychological Review, 120, 39-64.
Speekenbrink, M., Lagnado, D. A., Wilkinson, L., Jahanshahi, M., \& Shanks, D. R. (2010). Models of probabilistic category learning in Parkinson's disease: Strategy use and the effects of L-dopa. Journal of Mathematical Psychology, 54, 123-136.
Visser, I. (2011). Seven things to remember about hidden Markov models: A tutorial on Markovian models for time series. Journal of Mathematical Psychology, 55, 403415.

Visser, I., Jansen, B. R. J., \& Speekenbrink, M. (2010). A framework for discrete change. In P. C. M. Molenaar \& K. M. Newell (Eds.), Individual pathways of change: Statistical models for analyzing learning and development (pp. 109-123). Washington: APA.
Visser, I., \& Speekenbrink, M. (2010). depmixS4: An R-package for hidden Markov models. Journal of Statistical Software, 36(7), 1-21. Retrieved from http://www.jstatsoft.org/v36/i07/ (R package, current version available from CRAN)

# Dynamic Field Theory: Conceptual Foundations and Applications in the Cognitive and Developmental Sciences 

John P. Spencer (john-spencer@uiowa.edu)<br>Department of Psychology and Delta Center, University of Iowa, Iowa City, IA USA<br>Gregor Schöner (gregor.schoener@rub.de) and Yulia Sandamirskaya (sandayci@rub.de) Institut für Neuroinformatik, Ruhr-Universität Bochum, Germany

## Objectives and scope

Dynamical Systems thinking has been influential in the way psychologists, cognitive scientists, and neuroscientists think about sensori-motor behavior and its development. The initial emphasis on motor behavior was expanded when the concept of dynamic activation fields provided access to embodied cognition. Dynamical Field Theory (DFT) offers a framework for thinking about representation-in-the-moment that is firmly grounded in both Dynamical Systems thinking and neurophysiology. Dynamic Neural Fields are formalizations of how neural populations represent the continuous dimensions that characterize perceptual features, movements, and cognitive decisions. Neural fields evolve dynamically under the influence of inputs as well as strong neuronal interaction, generating elementary forms of cognition through dynamical instabilities. The concepts of DFT establish links between brain and behavior, helping to define experimental paradigms in which behavioral signatures of specific neural mechanisms can be observed. These paradigms can be modeled with Dynamic Neural Fields, deriving testable predictions and providing quantitative accounts of behavior.

One obstacle for researchers wishing to use DFT has been that the mathematical and technical skills required to make these concepts operational are not part of the standard repertoire of cognitive scientists. The goal of this tutorial is, therefore, to provide the training and tools to overcome this obstacle.

We will provide a systematic introduction to the central concepts of DFT and their grounding in both Dynamical Systems concepts and neurophysiology. We will discuss the concrete mathematical implementation of these concepts in Dynamic Neural Field models, giving all needed background and providing participants with some hands-on experience using interactive simulators in MATLAB. We will review robotic implementations to make the ideas concrete. Finally, we will take participants through a number of selected, exemplary case studies in which the concepts and associated models have been used to ask questions about elementary forms of embodied cognition and their development.

The interactive simulators will be available at the tutorial. We will take participants through the process of building and simulating models. We will use online tools available now

## Suggested Readings

(available at online, see below)

1. Spencer, J.P., Perone, S., \& Johnson, J.S. (2009). The dynamic field theory and embodied cognitive dynamics. In J.P. Spencer, M.S. Thomas, \& J.L. McClelland (Eds.) Toward a Unified Theory of Development: Connectionism and Dynamic Systems Theory Re-Considered. Oxford University Press, pages 86-118
2. Schutte, A.R. \& Spencer, J.P. (2009). Tests of the dynamic field theory and the spatial precision hypothesis: Capturing a qualitative developmental transition in spatial working memory. Journal of Experimental Psychology: Human Perception and Performance, 35:1698-1725.
3. Johnson, J. S., Spencer, J.P. \& Schöner, G. (2009): A layered neural architecture for the consolidation, maintenance, and updating of representations in visual working memory. Brain Research 1299:17-32
4. Sandamirskaya, Y. \& Schöner, G. (2010): An embodied account of serial order: How instabilities drive sequence generation. Neural Networks 23:1164-1179
5. Samuelson, L.K., Smith, L.B., Perry, L.K. \& Spencer, J.P. (2011). Grounding Word Learning in Space. PLoS One, 6, e28095.
6. Lipinski, J., Schneegans, S., Sandamirskaya, Y., Spencer, J.P. \& Schöner, G. (2012). A neuro-behavioral model of flexible spatial language behaviors. Journal of Experimental Psychology: Learning, Memory, and Cognition 38:1490-1511.
7. Sandamirskaya, Y., Zibner, S., Schneegans, S., Schöner, G. (2013): Using Dynamic Field Theory to extend the embodiment stance toward higher cognition. New Ideas in Psychology (in press, 2013)

## Target audience

No specific prior knowledge of the mathematics of dynamical systems models or neural networks is required as the mathematical and conceptual foundations will be provided during the tutorial. An interest in formal approaches to cognition is an advantage.

## Material covered in the course

1. Conceptual foundations of Dynamical Systems Thinking and Dynamical Field Theory (DFT): Embodied and situated cognition; Stability as a necessary property of embod-
ied cognitive processes; Distributions of population representation as the basis of spatially and temporally continuous neural representations;
2. Dynamical Systems and Dynamic Field Theory Tutorial: Concept of dynamical system; Attractors and stability; Input tracking; Detection, selection, and memory instabilities in discrete neuronal dynamics; Dynamical Fields and the basic instabilities: detection, selection, memory, boost-driven detection; Learning dynamics; Categorial vs. graded mode of operation; Practical implementation of DFT in simulators; Interactive simulation; Illustration of the ideas through robotic implementations;
3. Case study using DFT to understand embodied cognition and its development: visual and spatial working memory in children and adults; spatial precision hypothesis as a developmental mechanism in spatial recall, position discrimination, and change detection; mapping of DFT to functional neuroimaging with children.
4. Case study using DFT to understand brain-behavior relations in humans with functional neuroimaging: mapping of neural activation patterns in dynamic neural fields to the hemodynamic response measured with fMRI and fNIRS; case study on the neural processes that underlie visual working memory in children and adults.
5. Case study using DFT to understand how flexible action sequences can be generated: Dynamics of serial order and behavior organization; Coupling to real sensor and motor systems; Stability and flexible timing of actions in a sequence; Autonomy and executive control in neural and robotic systems.

## Lecturers

John P. Spencer is a Professor of Psychology at The University of Iowa and the founding Director of the Delta Center (Development and Learning from Theory to Application). He received a Sc.B. with Honors from Brown University in 1991 and a Ph.D. in Experimental Psychology from Indiana University in 1998. He is the recipient of the Irving J. Saltzman and the J.R. Kantor Graduate Awards from Indiana University. In 2003, he received the Early Research Contributions Award from the Society for Research in Child Development, and in 2006, he received the Robert L. Fantz Memorial Award from the American Psychological Foundation. His research examines the development of visuo-spatial cognition, spatial language, working memory, and attention, with an emphasis on dynamical systems and neural network models of cognition and action. He has had continuous funding from the National Institutes of Health and the National Science Foundation since 2001 and has been a fellow of the American Psychological Association since 2007. He will teach the tutorials on development and functional neuroimaging (numbers 3,4 below).

Gregor Schöner holds the Chair for Theory of Cognitive Systems and is the Director of the Institut für Neuroinformatik, Ruhr-Universität Bochum, Germany. Following his PhD in 1985 in theoretical physics at the University of Stuttgart, he held positions at the Center for Complex Systems of Florida Atlantic University, the Institut für Neuroinformatik, and the Center for Cognitive Neuroscience of the CNRS in Marseilles, France before returning to Bochum, Germany in 2001 to assume his current position. Dr. Schöner has received funding from different agencies in the US, Germany, France, and the European Union. He has published over to 200 scientific articles and chapters. Dr. Schöner is considered one of the world's experts on dynamic systems theory within the fields of Psychology and Cognitive Science, and is also a pioneer in the application of Dynamic Neural Fields to autonomous robotics. He will teach the conceptual and mathematical tutorials (numbers 1, 2 below).

Yulia Sandamirskaya is a Post-Doctoral researcher at the Institut für Neuroinformatik, Ruhr-Universität Bochum, Germany. She obtained her PhD (Dr.rer.nat.) in Physics for her work on embodied sequence generation within DFT. In her research, Dr. Sandamirskaya develops DNF models and robotic implementations of DNF architectures for sequence generation, behavior organization, and spatial language. Her work has been published in three journal articles, several conference proceedings, and a book chapter. Dr. Sandamirskaya will lecture on autonomy and sequence generation in DFT and present robotic implementations that demonstrate how concepts of DFT can lead to autonomous behavior in real environments (number 5 below).

## Schedule

1. Conceptual foundations of Dynamical Systems Thinking and Dynamical Field Theory (DFT): 30 minutes
2. Dynamical Systems and Dynamic Field Theory Tutorial: 90 minutes
3. Case studies using DFT to understand embodied cognition and its development: 60 minutes before and 60 minutes after the lunch break
4. Case study using DFT to understand brain-behavior relations 60 minutes
5. Case study using DFT to understand flexible action sequences 60 minutes

## Computer use

Participants who bring lab-tops with Matlab installed (student version is sufficient) will be able to follow demonstrations by actively working with the simulator during the lectures.

## Online resources

Publications, lecture material, and interactive simulators can be found at our DFT Summer School websites http://www.robotics-school.org and http://www.uiowa.edu/delta-center/research/dft

# How to Analyze Verbal Protocols 

Thora Tenbrink (t.tenbrink@bangor.ac.uk)<br>School of Linguistics and English Language<br>Room 331, 3rd floor New Arts, Bangor University, College Road, Bangor, Gwynedd<br>LL57 2DG, UK<br>Fon: +44 1248382263 ; Fax: +44 1248383267

Keywords: Think-aloud protocols; verbal data; cognitive processes; cognitive discourse analysis; linguistic structure; problem solving; complex cognition.

## Objectives and Scope

Cognitive science researchers are interested in a subject that is not directly accessible to observation: processes in the mind and brain, thoughts and thought processes. One way of addressing higher-level cognitive processes is to analyze verbal protocols produced along with cognitively complex tasks (Ericsson \& Simon, 1993), such as problem solving or decision making. Linguistic data of this kind can be seen as an external representation of some aspects of what is going on in the mind. In particular, think-aloud protocols and retrospective reports provide procedural information that complements other data, such as decision outcomes and behavioral performance results.

This tutorial explores the scope and limitations of verbal protocol analysis, and offers practical support for systematic analysis procedures. Language data can be analyzed with respect to content as well as structure. Conventionally, the focus of verbal protocol analysis lies on the content of verbal data, addressing those aspects (e.g., particular thought processes or strategies) that the speakers are themselves aware of (or 'heed', Ericsson \& Simon, 1993). The content-based inspection of verbal reports, particularly if carried out by experts in the problem domain and set against a substantial theoretical background (Krippendorff, 2004), often leads to well-founded specific hypotheses about the cognitive processes involved.

A detailed linguistic analysis can substantially support such content-based insights, but it can also offer further insights (e.g., Hölscher et al., 2011; Tenbrink et al., 2011; Tenbrink \& Seifert, 2011; Tenbrink \& Wiener, 2009). Research in cognitive linguistics, psychology, discourse analysis, and psycholinguistics indicates that patterns in language are systematically related to patterns of thought (e.g., Chafe, 1998). Drawing on these insights, one focus of the tutorial is to identify types of linguistic structure that point to specific cognitive processes. This is the main idea in the method of Cognitive Discourse Analysis (CODA) (Tenbrink, 2008; Tenbrink \& Gralla, 2009; Tenbrink, 2010).

Some aspects of language use reflect cognitive aspects that go beyond conscious reflection by individual speakers, and that are not necessarily directly observable in linguistic content. Speakers are typically unaware of the cognitive structures that are reflected in particular ways of framing a representation linguistically. Furthermore, they are not
consciously aware of the network of options (Tenbrink \& Freksa, 2009) that allows for a range of linguistic choices beside their own, which emerges more clearly by considering a larger data set collected under controlled circumstances. According to previous research in cognitive linguistics and discourse analysis (e.g., van Dijk, 2008), linguistic features such as the verbal representation of semantic domains reflected in ideational networks, specific choices of prepositions, lexical omissions and elaboration, conceptual perspectives revealed by language, presuppositions, hesitation and discourse markers, and many other linguistic features indicate certain conceptual circumstances; these are related to the current cognitive representations in ways that distinguish them from other options available in the network. In particular, the chosen linguistic options reflect what speakers perceive as sufficiently relevant to be verbalized, as well as the information status assigned to the diverse parts of the verbalization.

Besides building on established insights about the significance of particular linguistic choices, validating evidence for the relationship between patterns of language use and the associated cognitive processes can be gained by triangulation, i.e., the combination of linguistic analysis with other types of evidence such as memory or behavioral performance data, reaction times, eye movements, decision outcomes, or any other relevant data that can be collected in cognitively complex tasks.

## Format and organization

This tutorial is designed to cover a half day (three hours) and will be highly interactive. The tutorial will take the participants' current or intended projects as a starting point. It will be organized so as to cover the complete process of language data analysis (from initial ideas to evaluation of analysis results), including short presentations, discussion, and practical exercises where feasible. In particular, the following issues will be addressed:

Motivation: How (and to what extent) can language data serve as empirical resources to address research questions in cognitive science?

Data collection: What kinds of issues need to be considered in the light of actual research purposes?

CODA based analysis (main part): Systematic data annotation and interpretation informed by linguistic insights.

Triangulation: How can other types of empirical data complement the insights gained from language?

Participants who have already collected natural language data are encouraged to bring examples as handouts or on their computers. Furthermore they are encouraged to contribute a $10-\mathrm{min}$ talk related to one step of this process, and also to raise questions or issues to discuss for other steps. It is envisioned to prepare either a collection of papers or a collective paper, with authors interactively developing content based on combinations of their talks and the discussed issues.

## Target audience information

There is no prerequisite for taking this tutorial. It is open for researchers in cognitive science at any point in their career, ranging from graduate students to established experts.

Participants interested in a future publication are encouraged to submit a 300 -word abstract to propose a $10-$ minute presentation as part of the tutorial, and / or a critical issue to discuss.

## Tutor Information

Thora Tenbrink is a Lecturer in Cognitive Linguistics at Bangor University (Wales, UK), and a principal investigator in two projects in the Collaborative Research Center SFB/TR 8 Spatial Cognition (Bremen/Freiburg, Germany). Her main interest concerns complex cognitive processes and their representation in language. She is the author of "Space, Time, and the Use of Language" (Mouton de Gruyter, 2007), and co-editor of "Spatial Language and Dialogue" (Oxford University Press, 2009) and "Representing space in cognition: Interrelations of behavior, language, and formal models" (Oxford University Press, in press). Current research addresses cognitive strategies in various problem solving tasks, spatial communication in complex built environments, cognitive transformation processes, and inferences derived by problem solvers from situational clues, experience, and verbal and graphical information. See $\mathrm{http}: / / k n i r b . n e t$ for further information.

## Previous instantiations

This tutorial has previously been offered in various versions as listed below (see Tenbrink et al., 2012, for a report). The current version will focus on complex problem solving processes across all areas of cognitive science, tailored to the needs of its participants by establishing email contact in advance as far as possible.
"Understanding spatial thought through language use". Half-day tutorial at Spatial Cognition, August 31 September 03, 2012, Abbey Kloster Seeon, Germany.
"Understanding cognitive processes through language use". Half-day tutorial at ICCM 11th International Conference on Cognitive Modeling, April 12-15, 2012, Berlin, Germany.

Workshop "Language analysis in cognitive science". Cognitive Science Institute, University of Osnabrück (Germany), May 7-8, 2011.

Course "Language analysis in cognitive science" at the Cognitive Science Institute, University of Freiburg (Germany), summer semester 2009.

## References

Chafe, W. (1998). Language and the Flow of Thought. In M. Tomasello (ed.), The New Psychology of Language. Cognitive and Functional Approaches to Language Structure. Mahwah: Lawrence Erlbaum.
van Dijk, T.A. (2008). Discourse and Context. A Sociocognitive Approach. Cambridge: Cambridge University Press.
Ericsson, K.A., \& Simon, H.A. (1993). Protocol analysis: Verbal reports as data. Cambridge, MA: MIT Press.
Hölscher, C., Tenbrink, T., \& Wiener, J. (2011). Would you follow your own route description? Cognition 121, 228247.

Krippendorff, K. (2004). Content Analysis: An Introduction to its Methodology (2nd ed.). London: Sage.
Tenbrink, T. (2008). The verbalization of cognitive processes: Thinking-aloud data and retrospective reports. In W. Ramm \& C. Fabricius-Hansen (eds.), Linearisation and Segmentation in Discourse. Multidisciplinary Approaches to Discourse (MAD 08), Feb 20-23 2008, Lysebu, Oslo. Oslo: Dept. of Literature, Area Studies and Europ. Languages, Univ. of Oslo.
Tenbrink, T. (2010). CODA: Kognitive Diskursanalyse. In: E. Ruigendijk, T. Stolz, \& J. Trabant (Hrsg.), Linguistik im Nordwesten: Beiträge zum 1. Nordwestdeutschen Linguistischen Kolloquium. Bochum: Brockmeyer.
Tenbrink, T., Bergmann, E., \& Konieczny, L. (2011). Wayfinding and description strategies in an unfamiliar complex building. In L. Carlson, C. Hölscher, \& T.F. Shipley (Eds.), Proceedings of the 33rd Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.
Tenbrink, T., D'Odorico, T., Hertzberg, C., Mazman, G., Meneghetti, C., Reshoeft, N., \& Yang, J. (2012). Tutorial report: Understanding spatial thought through language use. Journal of Spatial Information Science 5:107-114. http://josis.org/index.php/josis/article/viewFile/124/90
Tenbrink, T., \& Freksa, C. (2009). Contrast sets in spatial and temporal language. Cognitive Processing 10 Supplement 2, S322-S324.
Tenbrink, T., \& Gralla, L. (2009). Accessing complex cognitive processes via linguistic protocol analysis. In U. Schmid, M. Ragni, M. Knauff (Eds.): Proceedings of the KI 2009 Workshop Complex Cognition, Paderborn, Germany, September 15, 2009.
Tenbrink, T., \& Seifert, I. (2011). Conceptual Layers and Strategies in Tour Planning. Cognitive Processing 12:1, 109-125.
Tenbrink, T., \& Wiener, J. (2009). The verbalization of multiple strategies in a variant of the traveling salesperson problem. Cognitive Processing 10:2, 143-161.

# Computational complexity analysis for cognitive scientists 

Iris van Rooij [contact person], Johan Kwisthout, Mark Blokpoel<br>Radboud University Nijmegen, Donders Institute for Brain, Cognition and Behaviour<br>Montessorilaan 3, 6525 HR Nijmegen, The Netherlands,<br>i.vanrooij@donders.ru.nl

Todd Wareham<br>Department of Computer Science, Memorial University of Newfoundland, St. John's, NL, Canada

Keywords: computational complexity theory, computational modeling, intractability, NP-hard, scalability, algorithms, fixedparameter tractability, approximation.

## Aims and Motivation

Many computational- or rational-level models of cognition postulate computations that appear to be computationally intractable (e.g., NP-hard or worse). Formally, this means that the postulated computations consume an exponential amount of time. Informally, this means that the postulated computations do not scale in any obvious way to explain how the modeled cognitive capacities can operate in the real world outside the lab. This problem of intractability is quite common in cognitive science. It is observed in practically all domains of cognition, including, for instance, perception, language, reasoning, categorization, decision-making, and motor planning. It is also not specific to any particular class of models, as it can arise for symbolic, connectionist, probabilistic (e.g. Bayesian), dynamical, logic-based, and even heuristic models of cognition.

How can cognitive scientists effectively deal with the intractability of their models? Several sophisticated and well-established concepts and techniques for computational complexity analysis have been developed in theoretical computer science over the last decades that can be directly utilized by cognitive scientists. Using these techniques cognitive scientists not only can assess whether or not a particular model is intractable, but also identify parameters of the model that are responsible for that intractability. As a result, these techniques can be used to generate hypotheses about how the models can be revised so as to make them computationally tractable, thereby improving the computational plausibility and scalability of the models. With this tutorial we aim to make these techniques for computational complexity analysis available for interested cognitive scientists.

[^2]computational complexity analysis, and (c) learn about the philosophical foundations of, and debate surrounding, the use of computational complexity theory for analyzing computational-level theories of cognition.

The tutorial will assume a basic level knowledge of cognitive psychology and an affinity with computational considerations.

## Morning session

In the morning session, participants will learn about the conceptual and mathematical foundations of computational complexity analysis in the context of cognitive modeling. The session will include a conceptual primer on several complexity-theoretic concepts (e.g., NP-hard, fixedparameter tractability) and techniques (e.g., polynomialtime and parameterized reduction). All these notions and techniques are also explained in: van Rooij, I. (2008). The tractable cognition thesis. Cognitive Science, 32, 939-984. Participants are kindly requested to read this paper prior to attending the tutorial. During the morning session, participants will have the opportunity to practice the described techniques via hands-on exercises (these can be done using paper and pencil). The lecturers will use an interactive style of instruction to help participants work through the exercises.

## Afternoon session

In the afternoon session, we will illustrate the broad applicability of the methodology. Wareham will guide participants through a detailed analysis of a model of analogy derivation based on Structure-Mapping Theory (van Rooij, Evans, Müller, Gedge, \& Wareham, 2008). Blokpoel will do the same for a Bayesian model of action understanding (Blokpoel, Kwisthout, van der Weide, \& van Rooij, 2010). Through interactive exercises, participants can see why both models are NP-hard and which parameters cause this intractability.

We then consider the important topic of approximation as an approach to dealing with intractability, with a focus on approximating Bayesian inferences. Kwisthout will present various notions of approximation and illustrate novel results on how constraining particular parameters of probability distributions may make approximation Bayesian strategies (like sampling or local search) successful. Our intent here is
to demonstrate that approximation is neither panacea nor a placebo when it comes to intractability (see also Kwisthout, Wareham, \& van Rooij, 2011; Kwisthout \& van Rooij, 2013).

We will close the tutorial with an interactive discussion session about questions, issues, objections and philosophical controversies regarding the demonstrated methodology (e.g., such as also covered by van Rooij, 2008; van Rooij, Wright, \& Wareham, 2012). Participants will be encouraged to bring in their own questions and points of discussion.

## Website and Materials

For more information about this tutorial, full details of the schedule and extra materials, please refer to our website: $\underline{\text { http://tcs.dcc.ru.nl/cogsci2013/. At the start of the tutorial, }}$ print-outs of the lecture notes will be made available to all participants in a tutorial booklet.

## References

Blokpoel, M., Kwisthout, J., van der Weide, T. \& van Rooij, I. (2010). How action understanding can be rational, Bayesian and tractable. In S. Ohlsson \& R. Catrambone (Eds.), Proceedings of the 32nd Annual Conference of the Cognitive Science Society (pp. 1643-1648). Austin, TX: Cognitive Science Society.
Kwisthout, J. \& van Rooij, I. (2013). Bridging the gap between theory and practice of approximate Bayesian inference. Cognitive Systems Research, 24, 2-8.
Kwisthout, J., Wareham, T., \& van Rooij, I. (2011). Bayesian intractability is not an ailment that approximation can cure. Cognitive Science, 35, 779-784.
van Rooij, I., Evans, P., Müller, M., Gedge, J. \& Wareham, T. (2008). Identifying sources of intractability in cognitive models: An illustration using analogical structure mapping. In B. C Love, K. McRae, and V. M. Sloutsky (Eds.), Proceedings of the 30th Annual Conference of the Cognitive Science Society, Austin, TX: Cognitive Science Society (pp. 915-920).
van Rooij, I. (2008). The tractable cognition thesis. Cognitive Science, 32, 939-984.
van Rooij, I., Wright, C. \& Wareham, T. (2012). Intractability and the use of heuristics in psychological explanations. Synthese, 187,471-487.

# Progress in Joint Action Research 

Participants<br>Susan E. Brennan (susan.brennan@stoneybrook.edu)<br>Department of Psychology, Stony Brook University<br>Stony Brook, NY 11794, USA<br>Daniel Richardson (dcr@eyethink.org)<br>Cognitive, Perceptual and Brain Sciences, University College London<br>26 Bedford Way, London, UK, WC1H 0AP<br>Michael J. Richardson (michael.richardson@uc.edu)<br>Center for Cognition, Action, and Perception, University of Cincinnati<br>Cincinnati, OH 45221, USA<br>Andreas Roepstorff (andreas@pet.au.dk)<br>Interacting Minds Centre, Aarhus University<br>8000 Aarhus C, Denmark<br>Natalie Sebanz (sebanzn@ceu.hu)<br>Cognitive Science Department, Central European University<br>Leo Frankel utca 30, 1021 Budapest<br>\section*{Organiser}<br>Günther Knoblich (knoblichg@ceu.hu)<br>Cognitive Science Department, Central European University Leo Frankel utca 30, 1021 Budapest

Keywords: Joint action; dialogue; perception and action; coordination; cooperation; shared intentions.

## A Current Trend in Cognitive Science

Joint action is an increasingly popular topic in Cognitive Science. This popularity reflects a recent theoretical trend of postulating, in one way or another, that human perception, action, and cognition are geared to enable successful coordination and communication with others. The speakers in the symposium will provide an overview of current progress in joint action research. Their contributions will address a wide range of phenomena ranging from tight temporal coordination to shared planning and discourse processes. Together, the contributions will illustrate that social constraints affect cognitive processing in a deeper sense than the more traditional notion of specific modules for social perception and social reasoning would suggest.

## Joint Action Shapes Processing in Dialogue Susan E. Brennan

Language processing is typically studied as comprehension or production, in solitary contexts. Studies of language use, on the other hand, often have pairs of people performing tasks together that require them to communicate, in which they fill the roles of speaker and addressee in rapid alternation or even simultaneously. Data from such studies have tended to take the form of transcribed dialogues from
which linguistic forms are coded. But a transcript is only an artifact of the processes that generated it. What are the effects of joint action on language processing, moment-bymoment, in dialogue? I will discuss several studies that illustrate how both global and local information have the potential to affect processing, and at what grains they do so. Global information may involve some degree of mentalizing about a partner and the partner's needs or intentions; this information is available at the start of the interaction (whether in detailed or quite rudimentary form) and may or may not be updated as the dialog unfolds. Local information includes verbal and nonverbal cues that emerge during the course of the interaction and that can be construed as evidence about the state of the task or the partner; such cues may be provided intentionally or instrumentally (as a byproduct of doing the task). Such cues can shape language processing in dialogue, whether implicitly (outside of awareness) or explicitly.

## The Reciprocity of Attention and Joint Action Daniel Richardson

Attention is shaped by joint action, and joint action is shaped by attention. When two participants have a discussion over an intercom, their gaze coordination is modulated by what they each believe the other can see and what they believe the other knows. But conversation is not required for coordination. We found that individuals looked
at photographs differently if they simply believed that another person sat elsewhere was looking at the same images. Joint activity does not always produces joint looking, however. Incorporating results from other labs, we can see that joint action can cause attention to converge or diverge depending on subtle aspects of the task, the rewards and the relationships between co-actors. So far, I have used visual attention as marker of perceptual and cognitive processing. But gaze, perhaps uniquely, has another function. The eyes take in information, but also interact with the social world. I will conclude with new experiments showing that gaze patterns are changed by looker's belief that they are being looked at. Gaze is not just a window onto the cognitive processes of joint action, but a tool used in its construction.

## Joint Action Coordination Michael Richardson

A fundamental feature of social behavior is face-to-face or co-present physical interaction. The success of such jointaction, whether measured in terms of social connection, goal achievement, or the ability of individuals to understand and predict the meaningful intentions and behaviors of others, is not only dependent on numerous neural-cognitive processes, but also on the physical and informational processes of perceptual-motor coordination. Understanding and modeling the dynamics of these coordination processes, including how they emerge and are maintained over time, as well as how differing stable states of coordination are activated, dissolved, and transformed is therefore imperative. Here I review research aimed at uncovering the dynamics of the perceptual-motor coordination that can emerge across a range of joint-action tasks and describe a dynamical modeling strategy for capturing such coordination. I further argue that as the enactment of a shared intention or task goal, the behavioral dynamics of perceptual-motor coordination not only lawfully express the physical, informational, and neural-cognitive relations that underlie successful joint-action, but also operate to control the behavioral intentions and action strategies adopted by social situated co-acting individuals.

## Coordination as Predictive and as Productive: Bootstraping from Low-Level Automaticity to Top-Top Interaction Andreas Roepstorff

An emerging body of research demonstrates how people, in direct interaction, become coupled along a number of dimensions: e.g., physiology, behavior, and semantics. However, once these methods are applied to more complex, goal-oriented settings, it seems that the patterns of coordination are not only synchrony and mimicry, but also more complex forms of complementarity such as in division of labor. This raises the issue of whether the same mechanisms may explain the simple and the more complex forms of coordination? I will explore the hypothesis that in a
predictive framework, the unfolding of more basic forms of coupling may support bootstraping into higher order coordination as shared perceptions, joint action and division of labor.

## Planning joint actions Natalie Sebanz

It has commonly been thought that to perform joint actions, individuals need to plan their actions around others'. However, recent evidence suggests that individuals not only plan actions around each other, but instead plan each others' actions. I will review studies showing that co-actors 1) form representations of and keep track of each other's tasks, 2) engage in motor simulation to predict the timing of each others' actions, and 3) form action plans that specify relations between their actions, thus enabling group-level action planning. These findings not only tell us about the cognitive mechanisms of joint action, but also challenge us to rethink the role of shared intentions in joint action.

## Acknowledgments

This symposium was supported by the ESF Eurocores program EuroUnderstanding (Roepstorff/Knoblich) and an ESF European Young Investigator award (Sebanz).

# Language and Gesture Evolution 

Josep Call (call@eva.mpg.de)<br>Department of Developmental and Comparative Psychology, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany<br>Susan Goldin-Meadow (sgm@uchicago.edu)<br>The University of Chicago, 5848 South University Avenue, Rm Green 510<br>Chicago, IL 60637, USA<br>Cat Hobaiter (clh42@st-andrews.ac.uk)<br>School of Psychology and Neuroscience, University of St. Andrews, Westburn Lane St Andrews, Fife, KY16 9JP, UK<br>Katja Liebal (katja.liebal@fu-berlin.de)<br>Freie Universität Berlin, Department of Education and Psychology, Evolutionary Psychology, Cluster Languages of Emotion, Habelschwerdter Allee 45 14195 Berlin, Germany<br>\section*{Discussant}<br>Michael Tomasello (tomasello@eva.mpg.de)<br>Department of Developmental and Comparative Psychology, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany

Keywords: gesture, language evolution, children, nonhuman primates, homesigner, sign language

## Introduction

In humans, gestural communication is closely intertwined with language: adults perform a variety of manual gestures, head movements and body postures while they are talking, children use gestures before they start to speak, and highly conventionalized sign systems can even replace spoken language. Because of this role of gestures for human communication, theories of language evolution often propose a gestural origin of language. In searching for the evolutionary roots of language, a comparative approach is often used to investigate whether any precursors to human language are also present in our closest relatives, the great apes, because of our shared phylogenetic history. Therefore, the aim of this symposium is to present recent progress in the field of language evolution from both a developmental and comparative perspective and to discuss the question if and to what extent a comparison with nonhuman primates is suitable to shed light on possible scenarios of language evolution.

## Gestural Communication in Wild Chimpanzees (Cat Hobaiter)

The gestural communication of great apes has provoked considerable interest by demonstrating striking evidence for the flexible and intentional use of a large communicative repertoire, key aspects of human language. We have recently expanded the extensive work in captivity, with the
first systematic study of gesture in a wild ape. In this presentation I will describe the gesturing of free-ranging chimpanzees in the Budongo forest, Uganda. Here we find that chimpanzee gestures are used in intentional communication by individuals of all ages, across a wide range of contexts, including evolutionarily urgent ones. I will discuss possible explanations for the combination of gestures into sequences, including: a) persistence following failed communications, and b) the idea that young chimpanzees may employ sequences as a fail-safe strategy to explore and fine-tune a very large repertoire of available forms, down to a more specific repertoire in regular use. I will examine flexibility in gestural communication, considering: a) whether or not gestures have specific meaning, b) if this is consistent across signalers, and c) whether or not the extent of flexibility in gestural repertoire varies with factors such as social context.

## What Can Gestures of Nonhuman Primates Tell us About Language Evolution? (Katja Liebal) ${ }^{1}$

Theories of language evolution usually argue for a unimodal origin of language, either gestural or vocal. Many of them draw on comparative evidence of the communicative

[^3]abilities of our closest relatives, the nonhuman primates. In this presentation, we will summarize the results of a systematic review of the literature on gestural communication in nonhuman primates covering the past 40 years and compare it with the main findings of studies on facial and vocal communication. We demonstrate that research into vocal, gestural and facial behaviors has very different theoretical and methodological approaches and as a result, comparisons of communicative patterns across modalities are problematic. We suggest that a multimodal approach to primate communication is essential to understand the complexity of nonhuman primates' communicative systems and to identify phylogenetic precursors to human language as part of a multi-modal system.

## From Homesign to Sign Language: Creating Language in the Manual Modality (Susan Goldin-Meadow)

Imagine a child who has never seen or heard any language at all. Would such a child be able to invent a language on her own? Despite what one might guess, the answer to this question is "yes". I describe congenitally deaf children who cannot learn the spoken language that surrounds them, and have not yet been exposed to sign language, either by their hearing parents or their oral schools. Nevertheless the children use their hands to communicate-they gesture-and those gestures, called homesigns, take on many of the forms and functions of language. I first describe the properties of language that we find in homesign. I next consider properties of language that homesigners can and cannot develop by comparing their linguistic systems to those developed by deaf individuals in Nicaragua. Thirty years ago large numbers of homesigners were brought together for the first time and Nicaraguan Sign Language (NSL) was born; NSL continues to develop as new waves of children enter the community and learn to sign from older peers. I end by taking an experimental approach to when gesture does and does not take on linguistic properties. I examine hearing individuals asked not to speak and instead communicate using only their hands. Although these silent gesturers can create some properties of language on the spot, they do not create all of the properties that homesigners develop over time.

# New Frameworks of Rationality 

Nick Chater (Nick.Chater@wbs.ac.uk)<br>Behavioural Science Group, Warwick Business School, University of Warwick<br>Coventry CV4 7AL, UK

Klaus Fiedler (klaus.fiedler@psychologie.uni-heidelberg.de)
Psychologisches Institut, Universität Heidelberg, Hauptstrasse 47-51
69117 Heidelberg, Germany

# Gerd Gigerenzer (gigerenzer@mpib-berlin.mpg.de) <br> Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, Lentzeallee 94 <br> 14195 Berlin, Germany 

Karl Christoph Klauer (klauer@psychologie.uni-freiburg.de)<br>Institut für Psychologie, Universität Freiburg, Engelberger Str. 41 79085 Freiburg, Germany

Mike Oaksford (m.oaksford@bbk.ac.uk)<br>Department of Psychological Sciences, Birkbeck College, University of London, Malet Street London, WC1E 7HX, UK<br>Keith Stenning (k.stenning@ed.ac.uk)<br>School of Informatics, University of Edinburgh, 10 Crichton Street<br>Edinburgh EH8 9AB, UK

Keywords: rationality; bounded rationality; reasoning; probabilistic reasoning; deductive reasoning; decision making; heuristics

## Introduction

The nature and extent of human rationality is an issue of ongoing debate. In the last two decades, this debate has been enlivened by the development and application of new theoretical frameworks. These include Bayesian notions of adjusting and using uncertain beliefs in an inductive manner as well as deductive probability-based logics as normative guidelines against which to weigh human judgments and decisions; the notion of ecological rationality based on lean and frugal heuristics well adapted to the structure of the environment; the notion of meta-cognitive myopia according to which people are accurate and sensitive in the processing the information in a given sample of observations, but are blind and naive to the history and validity of the sampled data; and game theory.

## Virtual Agreement: A Rational Framework for Joint Action and Communication (Nick Chater)

Game theory typically models interactions between agents in terms of players that are rational at the level of the individual. But when people need to coordinate their behaviour, which arises in joint action and communication, a vicious circle arises. What is rational for each player depends on what the other does; but figuring out what the other will do is no easier than figure out what one should do oneself. I will describe an
approach which assumes that players can resolve this problem by a process of "virtual agreement" - that is, the players figure out what they would agree to do, if they could discuss or bargain. Where the answer is well-defined, the agreement can be reached "virtually", i.e., without any information being exchanged. Virtual agreement requires common aims and knowledge - and can fail when players mis-estimate this common ground. Interesting, the process of reaching agreement can, in some cases, be modeled by conventional game theory.

## Cognitive Myopia (Klaus Fiedler)

What I have come to call "meta-cognitive myopia" (MM), using a term once suggested by Robyn Dawes, is the phenomenon that people are pretty accurate in utilizing even large amounts of stimulus information, whereas they are naive and almost blind regarding the history and validity of the stimulus data. This uncritical reliance on the information given is the most conspicuous when the task context makes it crystal-clear that the stimulus data should not be trusted. In the introduction, MM is located within a broader framework of meta-cognition research, and several examples are provided to illustrate the phenomenon. The central message is laid out that MM offers an alternative account of many biases in judgment and decision making, which have been traditionally explained in terms of capacity constraints, limited reasoning ability, motivational forces, or severely biased environmental input. The explanatory power of the MM construct, and its theoretical potential to predict new findings, is
then demonstrated in a major review section with reference to five paradigms: inability to discard irrelevant information; utilization of selectively sampled information; conditional inference biases; sample-size neglect; and myopia for the impact of aggregation levels. The final discussion is concerned with the learning origins of MM and the question of why evolution did not equip Homo sapiens with more effective metacognitive tools. An analysis of the costs and benefits will reveal that MM may serve important adaptive functions, and that eliminating MM may have maladaptive effects. Nevertheless, in the context of many real decision problems, the costs and irrational consequences of MM cannot be denied. The final discussion therefore focuses on possible ways to avoid and alleviate MM and its irrational consequences.

## Less Is More: Simple Solutions for Complex Problems (Gerd Gigerenzer)

In worlds of known risks, probability theory can provide the optimal course of action. In uncertain worlds, however, simple heuristics can result in smart solutions by focusing only on a few cues and ignoring the rest. The heuristics in the "adaptive toolbox" are anchored in the mind and the environment. They are embodied in the sense that they can exploit capacities of the human mind (such as recognition memory), which allow judgments to be quick. They are anchored in the environment in the sense that they can exploit statistical or social structures (such as signal-to-noise ratio). The study of the ecological rationality of heuristics and the bias-variance dilemma provides a general account to understand why and when less can be more.

## New Paradigms and Old Insights: Integrated Theories of Reasoning and Dynamic Inference (Mike Oaksford)

The new paradigm in reasoning, based on the probability conditional and dual process theory, offers new insights into human rationality. However, as with any psychological theory, there are a range of algorithmic issues concerning representations and processes that the new paradigm must address. We argue that doing so may require integrating these new insights with old insights from previous theoretical frameworks. In particular, the cognitive system needs to build small-scale models of the world which elaborate on information given in the premises and which are interrogated in reasoning. The nature of these processes has consequences for the new paradigm. For example, elaborative processes mean that conditional reasoning is most often dynamic and nonmonotonic involving changes in the probability distributions over which inference is defined. We draw out these consequences and sketch an integrative theory for conditional inference.

## What Linda Did Next: Relations Between an Interpretative Approach to Reasoning and the Judgment and Decision Literature (Keith Stenning and Michiel van Lambalgen) ${ }^{1}$

Stenning and van Lambalgen (2008) proposed that multiple logics are necessary to model human cognition, prominently a nonmonotonic logic known as Logic Programming (LP), which provides a cognitive model of fast frugal automatic reasoning from large human knowledgebases, to interpretations of current input. Much of the data from supposedly classical logical reasoning tasks (e.g. conditional reasoning, syllogisms, Wason's Selection Task) is derived from mixtures of subjects many of whom have nonmonotonic understandings and goals. This talk will position this program of research with regard to the judgment and decision literature on heuristic reasoning as exemplified by the Heuristics and Biases (H\&B) program of Kahneman and Tversky and the ecological fast and frugal heuristics of Gigerenzer and the ABC Research Group (ABC).

A multiple logics approach shares concerns with the ecological heuristic reasoning of the ABC group: with multiple methods of reasoning; with "automatic" rather than reflective reasoning; and with contextualisation. We illustrate with the well known Linda problem. The problem gives a personality description of Linda appropriate to the beginning of a story, but then asks a question ("Which is more likely? That Linda is a bank teller, or Linda is a bank teller and an active feminist?") from probability theory. Although there have been disagreements between $\mathrm{H} \& \mathrm{~B}$ and ABC , they share the assumption that this task is interpreted as calling for extensional reasoning. Of course there is no doubt it is intended as such an exercise, or that it is educationally important that students learn to recognise it as such. But we suggest that most subjects initially regard this task as an intensional reasoning problem, as readily developed within LP. This proposal points to the neglect of theories of the rationality of intensional reasoning. We sketch how such a program might look, and draw out some consequences for theories of reasoning, judgment and decision, and for theories of rationality more generally.

## Acknowledgments

This symposium was supported by the priority programme SPP1516 "New Frameworks of Rationality" by the Deutsche Forschungsgemeinschaft (DFG).

## References

Stenning, K., \& van Lambalgen, M. (2008). Human reasoning and cognitive science. Cambridge, MA: MIT Press.

[^4]
# Mechanistic Developmental Process: Rumelhart Prize Symposium in Honor of Linda Smith 

Larissa K. Samuelson (larissa-samuelson@uiowa.edu)<br>Department of Psychology and DeLTA Center<br>University of Iowa<br>Iowa City, IA, 52242<br>Anthony Morse (anthony.morse@plymouth.ac.uk)<br>Centre for Robotics and Neural Systems,<br>Plymouth University<br>Plymouth, Devon, PL4 8AA, UK<br>Chen Yu (chenyu@indiana.edu)<br>Department of Psychological and Brain Sciences<br>Indiana University<br>Bloomington, IN, 47405<br>Eliana Colunga (colunga@psych.colorado.edu)<br>Department of Psychology and Neuroscience, University of Colorado Boulder Boulder, CO, 80309

Thomas Hills (T.T.Hills@warwick.ac.uk) Department of Psychology University of Warwick<br>Coventry CV4 \& AL UK

Organizer<br>Larissa K. Samuelson (larissa-samuelson@uiowa.edu)<br>Department of Psychology and DeLTA Center<br>University of Iowa<br>Iowa City, IA, 52242

Keywords: cognitive development; word learning; computational models; robotics; dynamic systems.

## Motivation

Traditional views of cognition, cognitive development, and word learning have viewed knowledge as divorced from processes of perceiving and acting. Linda Smith has championed a dynamic, mechanistic, and processoriented view of cognition and focused on questions of development. She has shown how knowledge is embedded in, distributed across, and inseparable from the processes of perceiving and acting in the world. In so doing, she has enabled a new understanding of the nature of cognition and of how new ways of thinking come to be. This Rumelhart symposium in her honor illustrates how this focus on developmental process changes the questions asked and our resulting understanding of cognition. The five speakers will examine the developmental process of word learning from different vantage points ranging from perceptual to social to cognitive, and spanning multiple periods from the first words to rapid vocabulary growth to the building of semantic networks.

## Smart Behaviors from Simple Processes

## Author: Larissa Samuelson

Abstract: The period between16-months and 3-years of age is one of rapid vocabulary growth and diversification. Children this age are often referred to as "amazing word
learners" as they seem to know the whole category to which a new word applies after hearing one exemplar named one time. This perspective leads to theories of development couched in terms of innate knowledge structures and complex hypothesis testing. In contrast, this paper will present a view of development as the accumulation of small moments of knowing based on the specifics of the here-andnow that accumulate over longer timescales via simple associative processes. Data will illustrate how children use multiple sources of information such as the statistics of their vocabulary, associations in language structure, consistent space, and the relative novelty or familiarity of stimuli to solve language problems in a moment. Dynamic Neural Field and Hebbian Recurrent Network models will then be used to show that although children's behaviors look amazing, the processes that underlie them are not.

## A Unified View of Early Word Learning: Linking social interaction to sensory-motor Dynamics in Child-Parent Interaction

## Author: Chen Yu with Daniel Yurovsky

Abstract: Many theories of early word learning begin with the uncertainty inherent to learning a word from its cooccurrence with a visual scene. However, the relevant visual scene for infant word learning is neither from the adult theorist's view nor the mature partner's view, but is rather from the learner's personal view. To understand the mechanistic nature of early word learning, this talk focuses on micro-level behaviors as they unfold in real time in the dynamically complex interactions of child-parent interactions. We found that when infants interacted with
objects in play with their parents, they created moments in which a single object was visually dominant. If parents named the object during these moments of bottom-up selectivity, later forced-choice tests showed that infants learned the name. The sensory-motor behaviors of infants and parents were analyzed to determine how their actions on the objects may have created these optimal visual moments for learning. By studying the quality of parent-child social interactions at the sensory-motor level, the research provides a mechanistic understanding of the developmental dependencies between sensory-motor processes, social behavior, and language learning.

## Body Posture and Constraints on Word Learning in Robots and Children

Author: Anthony Morse with Viridiana Benitez

Abstract: The starting point for many theories of word learning is the co-occurrence of the visual object being attended, and the spoken word, but in reality such crosssituational learning has many problems and is unreliable in ways that simply don't match the human data. For example, in a series of recent experiments Smith, Samuelson \& colleagues demonstrate that children around 2 years old can learn object names in their absence so long as they are typically observed in consistent locations. We hypothesize that body posture is orchestrating this learning. In ongoing work with Linda Smith, we took this hypothesis literally placing the body centrally and binding ongoing multimodal experience via the body posture as a way to control the humanoid robot iCub , not only to replicate this data but to further generate predictions subsequently confirmed in further child experiments. With iCub we go beyond isolated cognitive modeling, embodying our system in real sensory and motor data, in a real interaction mirroring closely the setup of the psychology experiments. Based in spreading activation and self-organization our model tests and explores the role that the body plays in embodied cognition leading to a wider set of experiments in language learning. Finally we explore the role of competing systems (body and language), and simple learned skills in producing an explosion of word learning in an artificial robot.

## Time Considered as a Helix of Precious Words: Modeling the Emergence and Interactions of Word Learning Biases

## Author: Eliana Colunga

Abstract: Early word learning may be supported by a developmental feedback loop: the kind of words a child learns early on support the generalization of attentional biases, which in turn guides subsequent word learning. In a series of neural network simulations and a longitudinal behavioral study with toddlers in the lab, we explore the interactions between words learned and word learning biases, and argue that it is this interaction that builds the individual developmental trajectories children follow. First, we look at the development of the shape bias for solids and how its emergence is accompanied by an attentional shift in novel noun generalizations for other solidities, in both networks and toddlers. Second, we look at how the
emergence of a shape bias for solids is related to a shift in rate of learning for different kinds of words - shape-based or material-based - in networks and toddlers. Third, we look at how these interactions follow different developmental patterns in typically developing children at risk for language disorders, so-called "late talkers". Finally, we discuss the implications of this approach in increasing our understanding of language disorders, as well as our ability to improve early diagnosis and the design of individualized intervention plans.

## Growing Semantic Networks in a Sea of Words

Authors: Thomas Hills with Nicole Beckage
Abstract: Children learn language from exposure to a sea of words. The structure of this sea can be quantified using semantic space models applied to corpora of child-directed speech, which identify potentially meaningful statistical signatures from the company that words keep-including contextual diversity and associative structure (à la Saussure). At the same time, the words children learn can be characterized using computational models of growing networks of semantic information, with the edges between words based on semantic information embedded in the corpora. Our research has aimed at developing a quantitative theoretical account of early word learning based on the structure of the learning environment and the words children know-which we call the associative structure of child-directed language-and using this to predict the structure and growth of children's semantic networks over time. We have used this approach to predict the order of early word learning, to detect differences in typical and late talkers, and to predict differences in child versus adultdirected speech.

# Glushko Dissertation Prize Symposium 

## Participants

Douglas Knox Bemis (doug.bemis@gmail.com)<br>NSERM-CEA Cognitive Neuroimaging Unit CEA/SAC/DSV/DRM/Neurospin Center<br>F-91191 Gif-sur-Yvette Cedex, FRANCE

Neil Cohn (neilcohn@emaki.net)<br>Center for Research in Language<br>University of California, San Diego<br>La Jolla, CA 92093-0526, USA

George Kachergis (george.kachergis@gmail.com)<br>Cognitive Psychology Unit<br>Leiden Institute for Brain and Cognition<br>NL-2333 AK Leiden, The Netherlands

Andrew Lovett (andrew@cs.northwestern.edu)<br>Dept. of Electrical Engineering \& Computer Science<br>Northwestern University<br>Evanston, IL 60208-3118, USA

Chair
Robert L. Goldstone (rgoldsto@indiana.edu)
Department of Psychological and Brain Sciences, and Program in Cognitive Science
Indiana University
Bloomington, IN 47405

Keywords: language; narrative; statistical learning; cognitive modeling.

## Motivation

The Annual Glushko Dissertation Prize in Cognitive Science was established in 2011 as a way to promote future growth in cognitive science, and encourage students to engage in interdisciplinary efforts to understand minds. The prize is jointly sponsored by the Cognitive Science Society and the Robert J. Glushko and Pamela Samuelson Foundation, and honors young researchers conducting ground breaking research in cognitive science. The immediate goal is to recognize outstanding efforts to bridge between the areas that impinge on cognitive science and create theories of general interest to the multiple fields concerned with scientifically understanding the nature of minds and intelligent systems. Encouraging junior researchers to engage in these enterprises is one of the best ways to assure a robust future for cognitive science. The overarching goal is to promote a unified cognitive science, consistent with the belief that understanding how minds work will require the synthesis of many different empirical methods, formal tools, and analytic theories.

This symposium showcases the PhD research projects of the 2013 winners of the Glushko Dissertation Prizes. 2013 marks the first year that a symposium has been formed to assemble and showcase Glushko Prize winners' research. The prize-winning projects involve research on linguistic compositionality, understanding pictorial narratives, learning object-to-name mappings from complex environments, spatial problem solving, and visual awareness. The recruited research methods include neuroimaging, computational modeling, formal linguistic modeling, corpus analysis, psychological experiments, and theoretical analysis. Taken as a whole, the research projects strongly reinforces the view that contemporary cognitive science research is highly diverse, rigorous, creative, and fertile.

## Simple Composition During Language Processing: An MEG Investigation

Douglas Knox Bemis - 2012 PhD from New York University
Keywords: language; minimal phrases; magnetoencephalography.
Abstract: Language derives its expressive power from the ability to combine simple elements into complex ideas. To date, however, the vast majority of neurolinguistic investigations into combinatorial language processing have focused not on this transition from simple to complex, but rather on manipulations of complexity itself or measuring neural activity related to expectation violation. In this talk, I will present a novel neurolinguistic paradigm designed to isolate brain activity related to simple compositional mechanisms by combining the fine spatio-temporal resolution of MEG with the processing of minimal adjective-noun phrases (e.g. "red boat"). First I will demonstrate the ability of this paradigm to identify neural correlates of basic combinatorial processes that underlie the comprehension of such phrases. Then, I will present several experiments that probe the scope of these core processes both within language - comparing comprehension to production - and beyond - investigating combinatorial processing within both the pictorial and mathematical domain.

## Structure, Meaning, and Constituency in Visual Narrative Comprehension

Neil Cohn - 2012 PhD from Tufts University Keywords: narrative; grammar; comics.
Abstract: Narrative has been formally studied for at least two millennia, dating back to the writings of Aristotle. Contemporary research on the structure and comprehension of narratives has examined the discourse of spoken language. However, visual narratives in the form of
sequential images have also been pervasive throughout history, from cave paintings to contemporary comic books and strips. Yet, compared with the study of discourse in verbal language, the study of sequential image comprehension has been relatively impoverished. Just what are the structures motivating visual narratives and how are they processed?

I will explore this question using experiments guided by an overall theory that sequential images at the narrative level are structured and processed analogously to sequences of words at the sentence level. The main idea is that a narrative "grammar" organizes the structure of sequential images in the visual language used in comics, similar to the way that syntax organizes words into coherent sentences. We focus here on two salient parts of this analogy. First, I will explore the idea that visual narrative comprehension involves a system of narrative structure and a system of semantic coherence that contribute to comprehension. This correspondence is akin to the interaction between syntax and semantics at the sentence level. Second, I explore the idea that narrative structure is a hierarchic system that organizes images into constituents, analogous to the phrase structures of syntax in sentences. I will conclude by discussing the overall implications for the analogy between narrative structure in sequential images and syntax in sentence.

## More Naturalistic Cross-situational Word Learning

George Kachergis - 2012 PhD from Indiana University Keywords: statistical learning; cross-situational learning; language acquisition.
Abstract: Language acquisition is a ubiquitous, challenging problem involving fundamental cognitive abilities of attention, learning, and memory. Previous research has found that people can use word-object co-occurrences from ambiguous situations to learn word meanings. A recent associative model can account for a wide variety of wordlearning results using competing biases for familiar pairings and for stimuli with uncertain associates (Kachergis, PhD thesis). However, most studies of cross-situational learning present an equal number of words and objects in each learning situation, which is likely unrealistic. Moreover, displaying an equal number of words and objects may encourage learners to use assumptions such as each word going with one object, which may simplify the problem. This paper (Kachergis \& Yu, 2013) presents several conditions in which the number of words and objects do not match: either additional objects appear at random, or objects appear sometimes without their intended words. These manipulations do generally hurt learning in comparison to balanced conditions, but people still learn a significant proportion of word-object pairings. The results are explored in terms of statistics of the training trials-including contextual diversity and context familiarity-and with the uncertainty- and familiarity-biased associative model. Parametric differences between conditions hint at hidden cognitive constructs.

## Spatial Routines: A Framework for Modeling Visual Problem-Solving

Andrew Lovett - 2012 PhD from Northwestern University
Keywords: cognitive modeling; qualitative representation; visual problem-solving.
Abstract: Visual problem-solving tasks are an effective tool for evaluating cognitive abilities and predicting future performance. For example, Raven's Progressive Matrices is an intelligence test in which participants compare sequences of images to solve for a missing image. To better understand these tasks and the abilities they evaluate, I developed Spatial Routines, a general computational framework for modeling visual problem-solving. The framework is based on three psychological claims: 1) When possible, people reason about space using qualitative representations (e.g., identifying that one object is right of another), rather than absolute quantitative values. 2) Spatial representations are hierarchical. A given image might be represented as object groups, individual objects, or the parts within each object. 3) Qualitative spatial representations can be compared via structure-mapping: aligning their relational structure to find the corresponding elements.

The models generate symbolic representations from sketched input. They manipulate and compare these representations to determine an answer. They are useful for evaluating theories of perception, comparing problemsolving strategies, and identifying sources of difficulty for test-takers. Because the models must construct their own representations, they can highlight difficulties in representation-building and abstraction not identified by other computational models.

Three task models were constructed: Raven's Progressive Matrices, geometric analogy, and the visual oddity task. All three models perform as well as human adults, and problems that are difficult for the models are also difficult for people. Furthermore, by ablating a model's ability to perform certain operations and examining the resulting error pattern, one can generate new hypotheses about human reasoning.

## References

Kachergis, G. \& Yu, C. (2013). More naturalistic crosssituational word learning (this volume).

# Symposium <br> Communicative Intentions in the Mind/Brain 

Bruno G. Bara (bruno.bara@unito.it)<br>Center for Cognitive Science<br>University of Turin, Italy

Nick Chater ( Nick.Chater@wbs.ac.uk)<br>Warwick Business School, University of Warwick, UK

Michael Tomasello (tomasello@eva.mpg.de)<br>Max Planck Institute for Evolutionary Anthropology Leipzig, Germany

Rosemary Varley (rosemary.varley@ucl.ac.uk)<br>Division of Psychology and Language Sciences<br>University College London<br>London, UK

Keywords: communication; evolutionary psychology; neuropsychology; neuroscience.

The nature of intentions is a perpetual locus of interest for investigators of the human mind. Both occidental and oriental philosophical traditions treat intentions as the root of behavior; and many possible classifications have been offered in order to try to systematize the different types of intention. Moreover, recognition of intentions in others appears to be central to child development, and necessary for becoming a competent member of the society (Tomasello, 2008).

Recent work in the social neurosciences has focused, in particular, on social intentions, which may underpin the human predisposition toward joint, collaborative behavior. Communicative intentions are particularly central, yet have a puzzling recursive form (Bara, 2010). That is, given an actor's intention to convey a particular informational content, $C$, the actor must choose an act $A$, so that the partner will infer the actor's intention to communicate $C$. Yet the partner's inference itself depends on reconstructing that the actor would have chosen $A$, in order that the partner to infer the actor's intention to convey C (Grice, 1975).

The symposium on communicative intentions offers an analysis from a wide range of perspectives on these issues: evolution and development (Tomasello: Max Planck Institute for Evolutionary Anthropology in Leipzig, Germany), psychology and game theory (Chater: Warwick Business School, United Kingdom), neuropsychology (Varley: Division of Psychology and Language Science, University College London, United Kingdom), cognitive neuroscience (Bara: Center for Cognitive Science, Turin, Italy).

## Michael Tomasello: Communicating without Conventions

The evolutionary and developmental approach will provide a comparison between the gestural communication of human children and their nearest primate relatives, the great apes. This comparison reveals some of the cognitive and social-cognitive skills necessary for the human way of communicating that are present developmentally, and were probably present evolutionarily, before the emergence of conventional linguistic communication.

## Nick Chater: Virtual Bargaining as a Micro-Foundation for Communication and Joint Action

The psychological and game-theoretic approach outlines how a new theory of strategic social interaction, which extends the standard game theory of economics, can provide a rational theory of joint action; and how communication can be viewed as a special case of joint action, where both actor and partner must jointly infer the same content, $C$, given a mutually observed communicative act, $A$.

## Rosemary Varley: Communication without a Functioning Language System

Insights from neuropsychology and in particular, the relationship between language and communicative intentions, will be explored by examining the impact of severe aphasic language impairment on signaling communicative intentions and decoding the intentions of others. The evidence from acquired aphasia indicates considerable autonomy between language and communicative intentions in the established cognitive system (Willems, Benn, Hagoort, Toni \& Varley, 2011).

## Bruno Bara: The Intentionality Neural Network

The neuroscientific approach will describe a model of a dynamic intentionality network consisting of four brain regions, i.e. the right and left temporo-parietal junctions, the precuneus, and the anterior paracingulate cortex (Ciaramidaro et al., 2007). This model is based on a novel theoretical distinction among varieties of intention, which differ by the nature of an individual's pursued goal (private or social) and by the social interaction's temporal dimension (present or future). The intentionality network, which is independent from modality of expression, either linguistic or gestural (Enrici, Adenzato, Cappa, Bara \& Tettamanti, 2011), shows different activation patterns in relation to the nature of the intentions. The theoretical model of intention proposed contributes to enlarge our knowledge on the neurobiological bases of intention processing, in both healthy people and in people with impairments to the neurocognitive system that underlies intention recognition (Bara, Ciaramidaro, Walter \& Adenzato, 2011).

## References

Bara, B.G. (2010). Cognitive pragmatics: The mental processes of communication. Cambridge, MA: MIT Press.

Bara, B.G., Ciaramidaro, A., Walter, H., Adenzato, M. (2011). Intentional minds: A philosophical analysis of intention tested through fMRI experiments involving people with schizophrenia, people with autism, and healthy individuals. Frontiers in Human Neuroscience, 5:7, 1-11.

Ciaramidaro, A., Adenzato, M., Enrici, I., Erk, S., Pia, L., Bara, B.G., Walter, H. (2007). The intentional network: How the brain reads varieties of intentions. Neuropsychologia, 45, 3105-3113.

Enrici, I., Adenzato, M., Cappa, S., Bara, B.G., Tettamanti, M. (2011). Intention processing in communication: A common brain network for
language and gestures. Journal of Cognitive Neuroscience, 23, 2415-2431.

Grice, H.P. (1975). Logic and conversation: The William James Lectures, II. In P. Cole, \& J.L. Morgan (Eds.), Sintax and semantics 3: Speech acts. New York: Academic Press.

Tomasello, M. (2008). Origins of human communication. Cambridge, MA: MIT Press.

Willems, RM., Benn, Y., Hagoort, P., Toni, I., \& Varley, RA. (2011). Communication without a functioning language system. Neuropsychologia, 49, 3130-3135.

The symposium is limited to 4 participants, in order to allow 20 minutes of final discussion among the participants and with the public.

# Exploring Cognitive Diversity Across Disciplines and Cultures 

Andrea Bender ${ }^{1}$ \& Sieghard Beller ${ }^{1,2}$ (\{bender, beller\} @ psychologie.uni-freiburg.de)<br>${ }^{1}$ Department of Psychology, University of Freiburg, Germany<br>${ }^{2}$ Department of Human Sciences, University of Paderborn, Germany

Daniel Haun (haun@eva.mpg.de) Research Group "Comparative Cognitive Anthropology", MPI for Evolutionary Anthropology, Leipzig, DE

Cristine H. Legare (legare@ psy.utexas.edu) Department of Psychology, University of Texas at Austin, USA

Asifa Majid (Asifa.Majid@mpi.nl)<br>Center for Language Studies, Radboud University, \& Max Planck Institute for Psycholinguistics, Nijmegen, NL

Bethany Ojalehto (bethany.ojalehto@gmail.com)<br>Douglas L. Medin (medin@northwestern.edu)<br>Department of Psychology, Northwestern University, USA

Since the cognitive revolution, a widely held assumption has been that-whereas content may vary across cultures-cognitive processes would be universal, especially those on the more basic levels. Even if scholars do not fully subscribe to this assumption, they often conceptualize, or tend to investigate, cognition as if it were universal (Henrich, Heine, \& Norenzayan, 2010). The insight that universality must not be presupposed but scrutinized is now gaining ground, and cognitive diversity has become one of the hot (and controversial) topics in the field (Norenzayan \& Heine, 2005). We argue that, for scrutinizing the cultural dimension of cognition, taking an anthropological perspective is invaluable, not only for the task itself, but for attenuating the home-field disadvantages that are inescapably linked to cross-cultural research (Medin, Bennis, \& Chandler, 2010).

In a recent debate on the role of anthropology in and for cognitive science, obstacles that may hamper rapprochement were discussed in detail (Bender, Beller, \& Medin, 2012). In this symposium, we intend to move a step forward and showcase efforts to overcome these obstacles. The contributions to this symposium pursue a problem-driven approach to tackle specific questions of shared interest. The symposium brings together scholars from different disciplinary backgrounds (including cognitive and evolutionary anthropology, psycholinguistics, and cognitive, developmental, and comparative psychology), who are among the leading scientists in their fields. Each of them has contributed considerably to our expanding knowledge on how culture and cognition interact (e.g., Beller \& Bender, 2008; Haun et al., 2011; Legare \& Souza, 2012; Majid, Boster, \& Bowerman, 2008; Medin \& Atran, 2004). They present current research on different domains, ranging from causal cognition on the physical world through semantic categorization of olfaction and mental state understanding to processes of cultural transmission and moral reasoning in the biological domain, thus shedding new light on a field in cognitive science, in which recent years have seen an upsurge of interest and controversial debates.

## Olfactory language and cognition

> Asifa Majid

It has long been claimed "humans are astonishingly bad at odor identification and naming" (Yeshurun \& Sobel, 2010). However, recent evidence suggests exquisite elaboration of
olfactory lexicons in Aslian languages spoken in the Malay Peninsula (Burenhult \& Majid, 2011; Wnuk \& Majid, 2012). I present new data from speakers of Jahai, showing that Aslian language speakers show more agreement and shorter reaction times when free naming odors than their Western (Dutch) counterparts. This data further demonstrates that some speakers can be astonishingly good at odor naming. Furthermore, the Jahai data challenges current accounts of olfactory language and cognition, which in turn has implications for the larger language-thought debate.

# Weighing up physical causes in Germany and Tonga: A cross-cultural study on causal cognition 

Sieghard Beller, Annelie Rothe, Gregory Kuhnmünch, \& Andrea Bender

When people determine which of the entities involved in a physical interaction is responsible for its outcome, they weigh the entities differently even if the interaction is symmetric. This effect depends on various factors and also varies cross-culturally (Bender \& Beller, 2011). However, our results differ from previous research. In a replication study with participants from Germany and Tonga we investigate whether this is due to differences in the presentation of stimuli (visual vs. verbal) or to differences in answer mode (explanations vs. ratings of responsibility), and we test hypotheses on which cultural and/or linguistic factors may account for the cultural differences.

## Mental perspective taking across species and cultures

Daniel Haun, Katja Liebal \& Juliane Kaminski

Any trait claimed to define a species, needs not only be derived in that species, i.e. unique amongst its close phylogenetic kin, but also widespread across that species. Hence only concerted comparisons across related species and human cultures wield the power to identify the skills that define the human species (e.g., Haun et al., 2006). In the last years, psychologists have claimed such definitive traits in the area of social cognitive abilities such as the ability to understand others' knowledge, desires and beliefs. Here we compare individuals' abilities to understand others' mental states at dif-
ferent levels of complexity across a selected set of human cultures as well as across all non-human great ape species. In a non-verbal competitive game, participants were challenged to predict a competitor's moves, based on his/her knowledge, beliefs and desires. While children of all three cultures predicted with similar proficiency what their competitor chose, the non-human apes succeeded only in interpreting their competitor knowledge state, but showed no evidence of interpreting beliefs and desires. This data is consistent with the claim that reasoning about others' beliefs and desires is cross-culturally common and derived in humans.

## Communities of values: Moral reasoning about human-plant interactions among Indigenous Ngöbe of Panama

Bethany Ojalehto \& Douglas L. Medin

Research on sacred values often asks participants to make tradeoffs between a sacred good (e.g. acres of forest) and an instrumental incentive. As the external decision-maker, the participant decides the outcome for an insentient entity. But how might the decision-making process change if the entity is thought to be mindful? In previous research, we found that Indigenous Ngöbe adults of Panama are sensitive to signs of plant and animal sentience and may consider them agents with moral standing. Drawing on research suggesting that mind perception is key to moral reasoning (Gray, Gray, \& Wegner, 2007), the current study investigated Ngöbe reasoning about human-plant sacred value conflicts (e.g., right to life for plants versus humans). We find that Ngöbe treat plants as moral subjects whose interests must be considered. However, Ngöbe reframed tradeoffs from cases of competing interests to cases of cooperative relationships, reasoning in terms of the need for balanced reciprocity. We propose that Ngöbe treat sacred values not as absolute, objective goods which are pitted against each other, but as relational goods seen from multiple points of view (both human and nonhuman) which ultimately converge in systems-level perspective. We discuss implications for research on sacred values and morality.

## Imitative Foundations of Cultural Learning

## Cristine H. Legare

Imitation is multifunctional; it is crucial not only for the transmission of instrumental skills but also for learning social conventions such as rituals and facilitating social interaction. Thus, although children are indeed instrumental imitators (Gergely, Bekkering, \& Király, 2002), high-fidelity imitation has recently been linked to quintessentially social concerns, including the acquisition of normative behavior and affiliative motivations (Kenward, Karlsson, \& Persson, 2011; Over \& Carpenter, 2012). Despite the fact that imitation is a pervasive feature of children's behavior, there does not yet exist an integrated theoretical account of how children use imitation flexibly as a tool for cultural learning. Little is known about the kinds of information children use to determine when an event provides an opportunity for learning instrumental skills versus cultural conventions. I propose
that the cognitive systems supporting instrumental and conventional learning are facilitated by the differential activation of an instrumental stance (i.e., rationale based on physical causation) and a ritual stance (i.e., rationale based on cultural convention). I will present data demonstrating that (a) conventional framing increases imitative fidelity and the detection of differences between the performances of two actors and (b) witnessing multiple actors perform an action sequence increases imitative fidelity. The ritual stance increases imitative fidelity, a process essential for understanding cultural learning.

## References

Beller, S., \& Bender, A. (2008). The limits of counting: Numerical cognition between evolution and culture. Science, 319, 213-215.
Bender, A., \& Beller, S. (2011). Causal asymmetry across cultures: Assigning causal roles in symmetric physical settings. Frontiers in Psychology, 2:231.
Bender, A., Beller, S., \& Medin, D. L. (Eds.) (2012). Does cognitive science need anthropology? Topics in Cognitive Science, 4(3).
Burenhult, N., \& Majid, A. (2011). Olfaction in Aslian ideology and language. The Senses \& Society, 6, 19-29.
Gergely, G., Bekkering, H., \& Király, I. (2002). Rational imitation in preverbal infants. Nature, 415, 755.
Gray, H. M., Gray, K., \& Wegner, D. M. (2007). Dimensions of mind perception. Science, 315, 619.
Haun, D. B. M., Rapold, C. J., Call, J., Janzen, G., \& Levinson, S. C. (2006). Cognitive cladistics and cultural override in Hominid spatial cognition. Proceedings of the National Academy of Sciences, 103, 17568-17573.
Haun, D. B. M., Rapold, C. J., Janzen, G., \& Levinson, S. C. (2011). Plasticity of human spatial memory: Spatial language and cognition covary across cultures. Cognition, 119, 70-80.
Henrich, J., Heine, S. J., \& Norenzayan, A. (2010). The weirdest people in the world? Behavioral and Brain Sciences, 33, 61135.

Kenward, B., Karlsson, M., \& Persson, J. (2011). Over-imitation is better explained by norm learning than by distorted causal learning. Proceedings of the Royal Society B: Biological Sciences, 278, 1239-1246.
Legare, C. H., \& Souza, A. (2012). Evaluating ritual efficacy: Evidence from the supernatural. Cognition, 124, 1-15.
Majid, A., Boster, J. S., \& Bowerman, M. (2008). The cross-linguistic categorization of everyday events: A study of cutting and breaking. Cognition, 109, 235-250.
Medin, D. L., \& Atran, S. (2004). The native mind: Biological categorization, reasoning and decision making in development across cultures. Psychological Review, 111, 960-983.
Medin, D. L., Bennis, W. M., \& Chandler, M. (2010). Culture and the home-field disadvantage. Perspectives on Psychological Science, 5, 708-713.
Norenzayan, A., \& Heine, S. J. (2005). Psychological universals: What are they and how can we know? Psychological Bulletin, 131, 763-784.
Over, H., \& Carpenter, M. (2012). Putting the social into social learning: Explaining both selectivity and fidelity in children's copying behavior. Journal of Comparative Psychology, 126, 182-192.
Wnuk, E., \& Majid, A. (2012). Olfaction in a hunter-gatherer society: Insights from language and culture. In N. Miyake et al. (Eds.), Proceedings of the 34th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.
Yeshurun, Y., \& Sobel, N. (2010). An odor is not worth a thousand words: From multidimensional odors to unidimensional odor objects. Annual Review of Psychology, 16, 219-241.

# What if? Counterfactual reasoning, pretense, and the role of possible worlds 

Daphna Buchsbaum (daphnab@berkeley.edu)<br>Caren M. Walker (caren.walker@berkeley.edu)<br>Alison Gopnik (gopnik@berkeley.edu)<br>Department of Psychology, University of California, Berkeley, Berkeley, CA, USA<br>Nick Chater (nick.chater@wbs.ac.uk)<br>Behavioural Science Group, Warwick Business School, University of Warwick, Coventry, UK<br>David Danks (ddanks@cmu.edu)<br>Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA, USA<br>Christopher G. Lucas (cglucas@cmu.edu)<br>Charles Kemp (ckemp@cmu.edu)<br>Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, USA<br>Eva Rafetseder (Eva.Rafetseder@sbg.ac.at)<br>Josef Perner (josef.perner@sbg.ac.at)<br>Department of Psychology, University of Salzburg, Salzburg, Austria

Keywords: counterfactuals; pretense; causal reasoning

Symposium Moderator: Daphna Buchsbaum

## Motivation

Counterfactual thinking, where one envisions alternative possible events and their outcomes, is hypothesized to be one of the primary ways in which we reason about causal relationships (e.g., Pearl, 2000; Woodward, 2003). Recent computational and experimental work suggests that both adults and children may reason about causality in a manner consistent with probabilistic graphical models - coherent, complex representations of causal structure that allow distinctive kinds of inferences (e.g., Gopnik et al., 2004; Griffiths \& Tenenbaum, 2009). In particular, the causal models approach supports and distinguishes two types of inferences, predictions, on the one hand, and interventions, including counterfactual interventions, on the other. In predictions, we take what we think is true now as a premise and then use the model to calculate what else will be true. In counterfactuals, we take some value of the model that we currently think is not true as a premise, and calculate what would follow if it were.

Intuitively, childhood pretense bears a striking resemblance to counterfactual inference, but this relationship has not been widely explored. In general, pretend play seems paradoxical. Why should children spend so much time thinking about unreal worlds? Moreover, why would counterfactual inference itself be useful, since it is also about things that are not real? In this symposium we will explore the ways in which pretense and counterfactual thinking might be related (Buchsbaum, Walker \& Gopnik; Rafetseder \& Perner), the types of computations that might underly both kinds of thought (Lucas \& Kemp; Chater) and the ways in which both
might contribute to our causal understanding of the world, even without exposing us to new data (Chater; Danks).
Children's complex causal reasoning in pretend play
Authors: Daphna Buchsbaum, Caren M. Walker \& Alison Gopnik In causal counterfactuals and in causal interventions we take some value of a causal model that we currently think is not true as a premise, and calculate what would follow if it were. We propose that these crucially important abilities - creating possible causal interventions and testing alternative causal hypotheses - depend on the same cognitive machinery that children use when they pretend: adopting a premise that is currently not true, creating an event sequence that follows from that premise, and quarantining the result of this process from reality.

Empirical results with preschool children support these ideas. Buchsbaum, Bridgers, Weisberg, and Gopnik (2012) found a significant and specific relationship between counterfactual inference and pretense in a causal task. In a new study, we gave children a complex causal structure involving four different variables (e.g. the sun comes up, which makes the rooster crow and the birds chirp, and the rooster crowing wakes up the farmer). Children's counterfactual inferences about this complex structure paralleled their inferences about pretense and both were significantly accurate. Interestingly, children were more likely to make "backtracking" counterfactual inferences when explicitly asked to reason counterfactually. In contrast, they were significantly more likely to treat the "fixed" variable as an intervention ("non-backtracking") when asked to pretend its value.

## Representations, counterfactuals, and pretense

Author: Nick Chater Cognitive science views thought as computation. Computations are often conceived of as functions between input and output. But it may be more produc-
tive to explore the standard mantra from computer science, that a computer program consists of an algorithm operating over a data structure (see Chater \& Oaksford, in press; and related work by Pearl, 2000). The behaviour of the algorithm is well-defined even if there are interventions into the contents of the data structure, during the computation (if, for example, the contents of the Turing machine tape, or a register in a pocket calculator, are modified, the algorithm will react in a well-defined way). Thus, a computer program can be viewed as defining a rich set of counterfactuals over possible modifications to the data structure, as the computation unfolds.

It turns out that this point of view provides a natural analysis of what it means for information to be represented: that which is represented can be modified by an intervention on the data structure. From this perspective, human counterfactual thinking and pretend play may have a common basis: they may be different sources of evidence concerning the flexibility with which the cognitive system is able to modify its own data structures, to reason about how the world might have been (e.g., modifying the representation of the past and tracing the consequences), or how it might differently be conceived (in children's play acting, modifying a representation of a banana to be a representation of phone).

## Counterfactuals, causal learning, simulation, and pretense

Author: David Danks Causal structures provide information not just about what actually did occur, but about what would have occurred in various alternative scenarios. Counterfactuals are thus a key - in fact, necessary - guide for learning causal structures. Any method for learning about causal structures in the world must employ counterfactuals, whether explicitly or implicitly. The standard ways to judge counterfactuals for causal learning are through the use of interventions, or by focusing on "similar" (in relevant ways) cases. In many situations, however, these methods are too risky, too expensive, or infeasible for any number of other reasons. We must instead find other ways to judge counterfactuals.

Simulations based on one's present, uncertain causal beliefs provide a natural method for discovering surprising implications and incoherence in one's causal beliefs. We can use what we currently think about some causal structure to consider alternative possibilities, and thereby learn about our own (implicit) expectations and beliefs. Entirely mental simulation of causal relations is a challenging task, however, even for adults who have received training in it. One way to simplify the task is to ground the simulations in external, physical events that are analogous (in appropriate ways) to the underlying causal structure. That is, pretense and pretend play can help us learn about causal structures in the real world.

## A unified theory of counterfactual reasoning

Authors: Christopher G. Lucas \& Charles Kemp Bayesian networks have been used to account for many aspects of causal reasoning, including inferences about coun-
terfactual scenarios. We present a Bayes net model of counterfactual reasoning that generalizes and extends the work of Pearl (2000). The model distinguishes between counterfactual observations and counterfactual interventions, and can reason about both backtracking and non-backtracking counterfactuals. Several experiments demonstrate that our model accounts better for human inferences than Pearl's original proposal and a more recent Bayes net account developed by Rips (2009).

## Counterfactual reasoning vs reasoning counter-to-fact

Authors: Eva Rafetseder \& Josef Perner Pretense has some affinity with counterfactual reasoning. It typically contains a counterfactual supposition that something (a prop) is something other than what it really is, and like reasoning, it proceeds to further suppositions in a constrained, nonarbitrary way. Rafetseder, Cristi-Vargas, and Perner (2010) have distinguished counterfactual reasoning from hypothetical reasoning counter to fact. Reasoning counter to fact makes suppositions that may contradict known facts and then uses known regularities to draw further inferences. Counterfactual reasoning is more constrained; it has to adhere to the nearest possible world constraint, i.e., the reasoning from the counterfactual assumption has to stay as close as possible to what actually happened. Conformity to this constraint develops rather late around 6 to 12 years. In hypothetical reasoning typical regularities (e.g., If [whenever] somebody takes shoes off floors tend to be clean) are applied to counterfactual questions (e.g., If Carol had taken her shoes off, would the floor be dirty or clean?) without regard to what actually happened (e.g., that Max had also been walking across the floor with dirty shoes). The importance of this distinction for pretense is that the affinity of pretense to counterfactuality is limited to reasoning with premises counter to fact.

## References

Buchsbaum, D., Bridgers, S., Weisberg, D. S., \& Gopnik, A. (2012). The power of possibility: Causal learning, counterfactual reasoning, and pretend play. Philosophical Transactions of the Royal Society B: Biological Sciences, 367, 2202-2212.
Chater, N., \& Oaksford, M. (in press). Programs as causal models: Speculations on mental programs and mental representation. Cognition.
Danks, D. (2009). The psychology of causal perception and reasoning. In H. Beebee, C. Hitchcock, \& P. Menzies (Eds.), Oxford handbook of causation (p. 447-470). Oxford University Press.
Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T., \& Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. Psychological Review, 111, 1-31.
Griffiths, T., \& Tenenbaum, J. (2009). Theory-based causal induction. Psychological review, 116(4), 661-716.
Pearl, J. (2000). Causality. Cambridge University Press.
Rafetseder, E., Cristi-Vargas, R., \& Perner, J. (2010). Counterfactual reasoning: Developing a sense of "nearest possible world". Child Development, 81(1), 376-289.
Rips, L. J. (2009). Two causal theories of counterfactual conditionals. Cognitive Science, 1-47.
Woodward, J. (2003). Making things happen: A theory of causal explanation. Oxford University Press.

# Time and Causality: Mutual Constraints Insights from Event and Time Perception, Motor Control, and Gaming 

Marc J Buehner (BuehnerM@Cardiff.ac.uk) (Organizer/Moderator)<br>Cardiff University, School of Psychology, Tower Building, Park Place Cardiff, CF10 3AT UK

David A Lagnado \& Christos Bechlivanidis (d.lagnado@ucl.ac.uk, c.bechlivanidis@ucl.ac.uk)<br>University College London, Division of Psychology and Language Sciences, Gower Street, London WC1E 6BT UK

Marc O Ernst \& Marieke Rohde (marc.ernst@uni-bielefeld.de, marieke.rohde@uni-bielefeld.de)<br>Universität Bielefeld, Department of Cognitive Neuroscience, Universitätsstr. 25 33615 Bielefeld Germany

Michael E Young (michaelyoung@ksu.edu)<br>Kansas State University, Department of Psychology, 492 Bluemont Hall Manhattan, Kansas 66506-5302 USA

Keywords: Causality, Time, Perception, Learning, Reasoning, Binding, Event Segmentation, Sensory Integration

## General Background

The problem of how humans and other intelligent systems construct causal representations from non-causal perceptual evidence has occupied scholars in cognitive science since many decades. Most contemporary approaches agree with David Hume that patterns of covariation between two events of interest are the critical input to the causal induction engine, irrespective of whether this induction is believed to be grounded in the formation of associations (Shanks \& Dickinson, 1987), rule-based evaluation (White, 2003), appraisal of causal powers (Cheng, 1997), or construction of Bayesian Causal Networks (Pearl, 2000). Recent research, however, has repeatedly demonstrated that an exclusive focus on covariation while neglecting contiguity (another of Hume's cues) results in ecologically invalid models of causal inference. Temporal spacing, order, variability, predictability, and patterning all have profound influence on the type of causal representation that is constructed (Greville \& Buehner, 2010; Young \& Cole, 2012).

The influence of time upon causal representations could be seen as a bottom-up constraint (though current bottom-up models cannot account for the full spectrum of effects). However, causal representations in turn also constrain the perception of time: Put simply, two causally related events appear closer in subjective time than two (equidistant) unrelated events. This reversal of Hume's conjecture, referred to as Causal Binding (Buehner, 2012) is a top-down constraint, and suggests that our representations of time and causality are mutually influencing one another. At present, the theoretical implications of this phenomenon are not yet fully understood. Some accounts (e.g. Haggard, Clark, \& Kalogeras, 2002) link it exclusively to human motor planning (appealing to mechanisms of cross-modal temporal
adaptation, or forward learning models of motor control), while others adopt a broader perspective in line with models of Bayesian Evidence Integration (e.g. Buehner, 2012).

Causal beliefs influence not only time perception, but also judgments of temporal order, event segmentation, and phenomena related to multi-sensory integration. This symposium brings together researchers from various disciplines and backgrounds who all explore the interrelations between time, causality and perception, and do so applying learning theory, Bayesian approaches, physiological considerations, and high-level theories of cognition.

## Participant Abstracts

## Temporal Binding: Causality, Intentionality, or Both? (Buehner)

Since the first demonstration of temporal binding between actions and their consequences (Haggard et al., 2002) more than ten years ago, various theories have been put forward to account for it, ranging from modifications of an internal clock to sensory-specific re-adaptation. The common denominator across all demonstrations of the effect appears to be causality: binding in time and space occurs when there is a causal relation linking an action and its effect. I will discuss evidence of temporal binding in the complete absence of motor action. These results are at variance with theories based on sensuo-motor realignment and thus rule out intentionality as the basis for the effect. Instead, they can be accommodated by a Bayesian Framework of event perception. Intentionality could be included in this framework as an additional predictor.

## Constructing Time and Cause (Lagnado \& Bechlivanidis)

The notions of time and causality are intimately linked. Previous research shows that temporal information provides critical cues for causal inference (Greville \& Buehner, 2010; Lagnado \& Sloman, 2006) and that causal beliefs modulate judgments of temporal duration (Buehner, 2012). Using a novel experimental paradigm, where participants actively engage with a software-based 'physics world', we show both that temporal order guides causal judgments, and that causal judgments can themselves determine perceptions of temporal order. These findings highlight the constructive nature of causal and temporal perceptions.

## Asymmetries in Processing and Recalibration of Visuo-Motor Time Perception (Ernst \& Rohde)

If a voluntary movement event (e.g., a button press) and a sensory event (e.g., a visual flash) belong together, the button press has to happen before the flash, as a cause always comes before its effect. We investigated in a series of experiments whether this causal asymmetry also leads to a perceptual asymmetry in the perceived timing of visual and motor events. Participants had to judge the temporal order and the temporal interval of a visual flash and a button press after being trained to vision-lead and movement-lead temporal discrepancies. To be able to present visual stimuli both before and after motor events, we tracked participants' finger movement in real time and predicted the moment of a button press to time a visual flash with respect to this estimate. While the perception of temporal order is recalibrated symmetrically around the point of actual simultaneity, there are strong asymmetries in the recalibration of interval perception, which is mostly confined to the movement-lead side of the range of discrepancies. In a second study, participants had to rate simultaneity and action authorship, where again asymmetries around the point of actual simultaneity were observed. The temporal order of cause and effect thus has profound influences on human time perception of visual and voluntary motor events.

## Causal choice in the face of environmental complexity (Young)

Prior research on causal judgment and choice has focused on situations in which the events either lack temporal extent, only one candidate cause and effect are being judged, or events are presented as a series of discrete trials. All of these approaches help to manage the complexity of the interaction being judged, and evidence suggests that people engage in forward inference under these simplified conditions. I will discuss evidence from Video-game based research that suggests that greater environmental complexity appears to produce backward, rather than forward, inference.

## References

Buehner, M. (2012). Understanding the Past, Predicting the Future: Causation, Not Intentional Action, Is the Root of Temporal Binding. Psychological Science.
Cheng, P. W. (1997). From covariation to causation: A causal power theory. Psychological Review, 104, 367405.

Greville, W. J., \& Buehner, M. (2010). Temporal predictability facilitates causal learning. Journal of Experimental Psychology: General, 139(4), 756-771.
Haggard, P., Clark, S., \& Kalogeras, J. (2002). Voluntary action and conscious awareness. Nature Neuroscience, 5(4), 382-385.
Lagnado, D. A., \& Sloman, S. A. (2006). Time as a guide to cause. Journal of Experimental Psychology-Learning Memory and Cognition, 32(3), 451-460.
Pearl, J. (2000). Causality: Models, reasoning, and inference. Cambridge, England: Cambridge University Press.
Shanks, D. R., \& Dickinson, A. (1987). Associative accounts of causality judgment. In G. H. Bower (Ed.), Psychology of learning and motivation: Advances in research and theory (Vol. 21). San Diego, CA: Academic Press.
White, P. A. (2003). Making causal judgments from the proportion of confirming instances: The pCI rule. Journal of Experimental Psychology-Learning Memory and Cognition, 29, 710-727.
Young, M. E., \& Cole, J. J. (2012). Human sensitivity to the magnitude and probability of a continuous causal relation in a video game. Journal of Experimental Psychology-Animal Behavior Processes, 38(1), 11-22.

# Sequential Sampling Models Representing a Unifying Framework of Human Decision Making 

Jerome Busemeyer<br>(jbusemey@indiana.edu)<br>Department of Psychology, Indiana University<br>Bloomington, IN, USA

Antonio Rangel (rangel@hss.caltech.edu)<br>California Institute of Technology<br>Pasadena, CA, USA

Adele Diederich<br>(a.diederich@jacobs-university.de)<br>School of Humanities and Social Sciences<br>Jacobs University Bremen, Bremen, Germany

Jörg Rieskamp<br>(joerg.rieskamp@unibas.ch)<br>Department of Psychology, University of Basel<br>Basel, Switzerland

Andrew Heathcote<br>(andrew.heathcote@newcastle.edu.au)<br>School of Psychology, University of Newcastle Australia

Marius Usher<br>(marius@post.tau.ac.il)<br>Department of Psychology, University of Tel Aviv Tel Aviv, Israel

Keywords: Decision Making, Sequential Sampling Models, Evidence Accumulation, Attention Shifts, Cognitive Modeling

Sequential sampling models have been applied for describing the cognitive processes underlying various psychological processes such as memory, perception, or value-based decision making. They share the common assuming that people accumulate information stochastically over time and once the accumulated information passes a decision boundary a response is made. Depending on the cognitive domain the models differ in the specific nature of the accumulation process. In particular, they differ in their assumption about what type of information is processed, how the information is represented, and how the boundary is defined.
In this symposium we will illustrate how sequential sampling models successfully explain human behavior for various cognitive domains. For the decision making domain we show how people construct their preferences by accumulating options' attributes values over time once a decision threshold is reached. For the perceptual domain we show how gradual accumulation of sensory evidence over time explains people's perceptual decisions. Furthermore, we illustrate that sequential sampling models do not only predict the final outcome of a cognitive process but also present a description of the cognitive process itself. As such the symposium will illustrate how the analyses of response time distributions provide evidence for the dynamic accumulation process. Furthermore, the neurological basis of the dynamic accumulation process has recently also been explored. In sum, the symposium will illustrate the strength of sequential sampling model as a unifying framework for explain human cognition and behavior across many domains.

## Accumulation of Information with Attention Shifting Across Attributes Adele Diederich* \& Jerome Busemeyer*

Most applications of sequential sampling (diffusion) models to decision making assume that information is sampled from a constant or stationary, albeit noisy, source of information during the accumulation period leading up to a decision. Formally, these models usually employ a constant mean drift rate throughout the accumulation process leading up to a decision. However, many cognitive and decision tasks provide conflicting attributes that could compete for selective attention to guide the accumulation process while making a decision. For example, when choosing a consumer product, a person needs to shift attention between quality and cost; when a security agent scans a bag, the person needs to shift attention to different objects in the bag; when making a social choice, a person can attend to either implicit attitude feelings or explicit rational arguments.

Diederich (1997) developed a multiple stage diffusion model that represents this attention shifting as a Markov process that changes the drift rate across stages of the decision. This work describes the multiple stage model and reviews five major applications which compared the multiple stage model to stationary models with respect to their ability to account for perceptual decisions involving conflicting attributes. This includes behavioral studies modeling choice and response time as well as electrophysiological studies that used the model to account for the trajectory of neural activation during evidence accumulation. Based on this body of evidence, we conclude that there is substantial empirical support for an attention shifting process during multi-attribute decision making.

## A Two-Phase Theory of Choice Conflict Tasks

## Andrew Heathcote* \& Kirsty Hannah

We propose a theory of how decision processes are affected by response conflict. The theory is developed to account for the fine-grained time course of both response speed and accuracy as quantified by delta functions and conditional accuracy functions. The theory, Two-Phase Evidence Accumulation (TPEA), extends Brown and Heathcote’s (2008) Linear Ballistic Accumulator (LBA) model to provide a unified account of three tasks that have been central to the study of cognitive interference the Stroop, Simon and Flanker tasks. The theory is explicated by demonstrating that it provides a coherent parametric account of Stroop and Simon effect delta functions shown by Pratte, Rouder, Morey and Feng (2010) to be incompatible with existing theories. We then show that, without modification, TPEA is also able to account for White, Ratcliff and Starns' (2011) flanker data.

## The Attentional Drift-Diffusion-Model: Eye-tracking and Neurobiological Evidence

## Antonio Rangel*

We propose a computational model of value-based binary choice in which fixations guide the comparison process. The model is an extension of the classic Drift-Diffusion-Model to an environment in which attention matters. We provide eye-tracking evidence showing that the model can quantitatively explain complex relationships between fixation patterns and choices, as well as several fixationdriven decision biases. We also provide fMRI evidence showing that key elements of the model are consistent with the operations of the decision-making circuitry at the time of choice.

## Decision Making With Non-stationary Evidence, Adaptation and Decision-Confidence

## Marius Usher*

The integration of evidence supporting different choice options is a fundamental process underlying all of our decisions, ranging from the simplest perceptual decisions (e.g., detect the presence of an enemy-rocket signal embedded in a noisy radar stream) to complex economic ones (e.g., which apartment to buy). A limitation of most studies that examined evidence-integration, however, is that they focussed on situations in which the evidence is stationary. I will present recent computational and experimental studies that examines decision-making under non-stationary evidence, characterized by temporal uncertainty: Observers detected visual luminance "signals" embedded within longer streams of "noise" with signals varying in duration and occurring at different onset latencies. Using a computational model, we showed that optimizing performance under such conditions, requires a leaky ("forgetful") integration process, the time-scale of
which is matched to the expected signal duration. In subsequent psychophysical experiments, we tested whether human observers can indeed control their integration-time scale, such as to flexibly adapt it to the characteristic signal duration. The results provide strong support for this idea. Finally, I will discuss how the evidence-integration framework can account for data that requires the observers to report their decision-confidence.

## Comparing Perceptual and Preferential Decision Making

Gilles Dutilh \& Jörg Rieskamp*
What are the differences between perceptual and preferential decisions? In a perceptual decision the decision maker aims for a correct decision and there is an outside criterion that determines which decision is correct. In contrast, in a preferential task the decision maker's goals are subjective, so that no correct option exists. Despite these differences sequential sampling models have successfully been applied to both types of decisions. In our study we explore the overlap and the differences between perceptual and preferential decision making. To do so, we developed an experimental task that can be presented as either a perceptual or a preferential task. We show that the classic speed-accuracy trade-off and effects of stimulus difficulty are elicited in the perceptual version of this task. In the preferential version of the task, the stimulus array reflects a gamble that the participant can choose to play or not. In this gamble, the black and white dots represent potential gains or losses. We show that people behave risk and ambiguity averse in this task. The diffusion model is applied to both versions of the task for identifying the essential differences between the two types of decision making. We conclude that similar evidence accumulation processes could underlie rather different decision making processes, but that the model parameters have to be interpreted differently.

## Acknowledgments

We like to acknowledge the Australian research council professorial fellowship awarded to Andrew Heathcote and the SNSF research grant (SNSF100014_143854/1) awarded to Gilles Dutilh and Jörg Rieskamp.

## References

Brown, S.D., \& Heathcote, A. (2008) The simplest complete model of choice reaction time: Linear ballistic accumulation. Cognitive Psychology, 57, 153-178.
Diederich, A. (1997). Dynamic stochastic models for decision making with time constraints. Journal of Mathematical Psychology, 41, 3, 260-274.
Pratte, M. S., Rouder J. N., Morey R. D., \& Feng C. (2010). Exploring the Differences in Distributional Properties Between Stroop and Simon Effects Using Delta Plots. Attention, Perception, \& Psychophysics, 72, 2013-2025.
White, C.N., Ratcliff, R., \& Starns, J.J. (2011). Diffusion models of the flanker task: Discrete versus gradual attention selection. Cognitive Psychology, 63, 210-238.

# Multiword Sequences as Building Blocks for Language: Insights into First and Second Language Learning 

Moderator: Morten H. Christiansen (christiansen@cornell.edu) Department of Psychology, Cornell University, USA<br>Moderator: Inbal Arnon (inbal.arnon@gmail.com)<br>Department of Psychology, University of Haifa, Israel<br>Elena Lieven (elena.lieven@manchester.ac.uk)<br>School of Psychological Sciences, University of Manchester, UK

Alison Wray (wraya@cardiff.ac.uk)<br>School of English, Communication \& Philosophy, Cardiff University, UK

Keywords: formulaic language, language acquisition, language processing, second language learning, usage-based learning

## Introduction

Many grammatical frameworks view words and rules as the basic building blocks of language, with multiword sequences being treated as peripheral exceptions in the form of idioms, etc. (e.g., Pinker, 1999). The new millennium, however, has seen a shift toward construing multiword sequences not as linguistic rarities but as important building blocks for language acquisition and processing. Based on a growing bulk of evidence of sensitivity to multiword sequences in language learning and use (see Ellis, 2012, for a review), multiword sequences have come to figure prominently in many current approaches to language, including item-based learning (Lieven, 2010), formulaic language (Wray, 2008), usage-based language processing (Arnon \& Snider, 2010), and chunk-based learning (McCauley \& Christiansen, in preparation). This symposium brings together experts from these different approaches to language to explore the idea that first (L1) and second (L2) language learners differ with respect to their ability to use multiword building blocks to learn and process language, and that this difference affects learning strategies and outcomes.

Unlike young children, adult learners rarely reach native proficiency in pronunciation, morphological and syntactic processing, or the use of formulaic language and idioms (see Ellis, 2012, for a review). Yet adults do not have problems with all aspects of novel language learning: they seem to learn certain aspects of language (e.g., words) better than others (e.g., grammatical relations, formulaic expressions). Existing accounts of the differences between L1 and L2 language learning have tended to focus on biological, cognitive, and neural differences between children and adults. These accounts predict the general difference in proficiency between the two populations, but struggle to explain the specific patterns of language learning observed in children and adults.

Understanding the different paths and outcomes of L1 and L2 learning has wide-reaching implications for cognitive science in terms of what it means to know a language, how much of such knowledge is 'built-in', and how learning changes as a function of prior knowledge and experience. Crucially, while L1 acquisition, adult psycholinguistics, and L2 learning are often studied separately, we bring together insights from developmental psychology (Lieven), psycholinguistics (Arnon), computational investigations of language structure (Christiansen), and applied psycholinguistics (Wray) to present a diverse and rich perspective on multiword building blocks in language learning and use.

The symposium participants have all worked extensively on language acquisition and use. Lieven has been at the forefront of developing the usage-based approach to language learning and has conducted numerous studies on the nature of children's early language use and representation. Arnon has been studying both the processing of multiword sequences by adult native speakers and the way chunk-based learning can impact adult performance in artificial languages. Christiansen has conducted extensive psycholinguistic and computational work exploring the units of language learning and the way such units affect learning. Wray has worked broadly on formulaic expressions in both native and non-native speakers as well as more recently in the language of Alzheimer's patients. Together, the participants have published more than 70 papers relating to the role of multiword sequences in language.

## Lieven: Multiword Sequences in L1 Acquisition

Theoretical and empirical reasons suggest that children build their language not only out of individual words but also out of multiword strings. These are the basis for the development of schemas containing slots. The slots are putative categories which build in abstraction while the schemas eventually connect to other schemas in terms of both meaning and form. Evidence comes from the nature of
the input (Cameron-Faulkner et al., 2003); the ways in which children construct novel utterances (Lieven et al., 2009); and the computational modeling of children's grammars (Bannard et al., 2009). However, nearly all this research is on English which is unusual in its rigid word order and impoverished inflectional morphology. There has also been much less research on the development of the 'meaning pole' in the form-meaning mappings of schemas. I will address both these issues using our recent studies in English, German, Polish and Chintang.

## Arnon: Multiword Sequences in Adult Language Learning and Use

Prior studies have shown that native speakers are sensitive to the distributional properties of multiword sequences when processing language (see Ellis, 2012, for a review). Results are presented suggesting that this sensitivity also extends to language production and is not modulated by syntactic constituency: higher frequency phrases are phonetically reduced within and across syntactic boundaries indicating the prominence of sequence based information. A second study investigated whether such sensitivity to multiword sequences might be harnessed to improve L2 learning. Adult learners showed better learning of an artificial language incorporating a grammatical gender system when first exposed to larger chunks (sentences) and only then individual words (noun-labels). This result suggests that L2 learning of grammatical gender languages may be improved by initially exposing learners to multiword sequences instead of isolated words, thus mirroring the sensitivity to multiword sequences in L1 acquisition and use.

## Christiansen: Computational Investigations of Multiword Chunks in Language Learning

Computational modeling provides further means to investigate the use of multiword chunks by different types of language learners. The Chunk-Based Learner (CBL; McCauley \& Christiansen, in preparation) gradually builds an inventory of chunks-consisting of one or more wordsused for both language comprehension and production. The model learns incrementally from corpora of child-directed speech using simple distributional information and accommodating a range of developmental findings. Results are presented indicating that multiword chunks provide a useful basis for capturing children's productions across a number of different languages independent of their word order. When applied to L2 learner corpora, CBL reveals that the productions of such speakers rely less on multiword chunks compared to speech of both L1 learners and adult native speakers. Thus, these modeling results corroborate our hypothesis about the differential use of multiword building blocks by L1 and L2 learners.

## Wray: Formulaic Expressions: Further Issues

Why have we not progressed further than we have, in understanding the role of formulaic sequences in L2
learning? This presentation will consider how certain assumptions underpinning the existing body of knowledge could constrain the research questions we ask. For instance, how safe is the assertion that non-native speakers rarely achieve nativelikeness (typically attributed to not mastering formulaic sequences)? How appropriate is it to gauge the formulaic language knowledge of adult L2 learners by comparing it to what, for native speakers, is anchored in the social and cognitive experiences of childhood? To what extent can we assert that (all) L1 speakers know the same things about how words fit together? How do recent proposals by Hanks (2013), Port (2007) and Sinclair (Cheng et al., 2009) that the word is not a reliable unit of form or meaning impact on the growing evidence that multiword strings might be?

## Symposium Format

The symposium starts with a 5-minute introduction, followed by four 20 -minute presentations (including time for questions), and concludes with a 15 -minute general discussion.

## References

Arnon, I. \& Snider, N. (2010). More than words: Frequency effects for multi-word phrases. Journal of Memory and Language, 62, 67-82.
Bannard, C., Lieven, E. \& Tomasello, M. (2009) Modeling children's early grammatical knowledge, Proceedings of the National Academy of Sciences, 106, 17284-17289.
Cameron-Faulkner, T., Lieven, E. \& Tomasello, M. (2003). A construction based analysis of child directed speech. Cognitive Science, 27, 843-873.
Cheng, W., Greaves, C., Sinclair, J. M. \& Warren, M. (2009). Uncovering the extent of the phraseological tendency: Towards a systematic analysis of concgrams. Applied Linguistics, 30, 236-252.
Ellis, N.C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. Annual Review of Applied Linguistics, 32, 17-44.
Hanks, P. (2013). Lexical analysis: Norms and exploitations. Oxford: Oxford University Press.
Lieven, E. (2010). Input and first language acquisition: Evaluating the role of frequency. Lingua, 120, 2546-2556.
Lieven, E., Salomo, D. \& Tomasello, M. (2009). Two-yearold children's production of multiword utterances: A usage-based analysis. Cognitive Linguistics, 20, 481-508.
McCauley, S.M. \& Christiansen, M.H. (in preparation). Language learning as language use: A computational model of children's language comprehension and production.
Pinker, S. (1999). Words and rules. New York: Basic Books.
Port, R. F. (2007). How are words stored in memory? Beyond phones and phonemes. New Ideas in Psychology, 25, 143170.

Wray, A. (2008). Formulaic language: Pushing the boundaries. Oxford: Oxford University Press.

# Distributed remembering in a social context: Effects of communicating a shared past on what is remembered 

Alin I. Coman (acoman@princeton.edu)<br>Department of Psychology, Princeton University<br>Princeton, New Jersey 08544 USA<br>Gerald Echterhoff (g.echterhoff@uni-muenster.de)<br>Department of Psychology, University of Münster<br>Fliednerstr. 21, 48149, Münster, Germany<br>Micah Edelson (micah.edelson@weizmann.ac.il)<br>Department of Neurobiology, The Weizmann Institute of Science, Rehovot, 76100, Israel<br>William Hirst (hirst@newschool.edu)<br>Department of Psychology, New School for Social Research, 80 Fifth Avenue, NY, NY 10003 USA<br>Kourken Michaelian (kmichaelian@bilkent.edu.tr)<br>Department of Philosophy, Bilkent University, Ankara, 06800, Turkey<br>Charles B. Stone (charlie.stone@gmail.com)<br>Psychological Sciences Research InstituteDepartment of Psychology, Université catholique de LouvainCatholic University of Louvain, Place du Cardinal Mercier 10, B-13248, Louvain-la-Neuve, Belgium

Keywords: Distributed cognition, memory, conversations, agent-based modeling, collective memory

## Challenging Issue

Remembering frequently involves collaboration among two or more individuals, often taking the form of a conversation. In some conversations, one person conveys new information to another, as when a daughter announces to her mother that she is engaged. In other conversations, two people talk to each other about a shared past, as when a couple reminisce about the evening on which they became engaged. When the conversation is about a mutually experienced event or a shared body of knowledge, it has the potential to shape both what emerges in the discussion, as, for instance, when one participant scaffolds the remembering of the other. It also has the potential to reshape how participants might subsequently remember the material, with the possibility that the memories of the participants will be more similar after the conversation than before it. In other words, what people remember is, in part, the result of how they jointly recount the past with others. Yet, despite the critical contribution of joint conversational remembering to memory, detailed study of this phenomenon is only beginning to be undertaken. The objective of this symposium is to bring together several strands of research that explores conversational, or, more generally, communicative influences on memory. The speakers
offer a range of different approaches. Dr. Stone explores how public speeches can induce forgetting as well as reinforce memories across a large population and thereby promote a mnemonic convergence through a single social interaction. Dr. Echterhoff considers the role of motivation in moderating conversational influences. Dr. Coman investigates whether these conversational influences propagate across large networks of individuals and can thereby promote a mnemonic convergence. Dr. Edelson examines the neuroscience underlying memory conformity. And Dr. Michaelian engages the philosophical underpinning of the concept of collective memory. In a short summation, Dr. Hirst places these papers in the larger context of social aspects of memory and moderates further discussion.

## Charles B. Stone

A conversation can not only reinforce already existing memories and implant misleading ones, but also induce forgetting for unmentioned, but related memoriesaterial. Dr. Stone discusses how this selective remembering can induce forgetting not only in speakers, but also in listeners. In his talk, he extends the laboratory finding on induced forgetting to the effects of political speeches on public memories, showing that listening to a recent speech by the King of Belgium induced forgetting for unmentioned, but related political material, but not unmentioned, unrelated material for French-speaking

Belgians. Similar induced forgetting was not observed in French-speaking Belgians who did not listen to the speech or Dutch-speaking Belgians more generally. The finding underscores the importance of communicative influences on everyday memories.

## Gerald Echterhoff

After tuning their message to their audience's attitude during an initial retelling, communicators' subsequent memory for the topic often reflects the audience-tuned view expressed in the message. Dr. Echterhoff presents research investigating the motivational underpinnings of such audience-congruent memory alignment, specifically, speakers' motives to create a "shared reality" with their audience. Shared-reality creation, in turn, satisfies epistemic needs, that is, needs for a valid and confidently held understanding of the world. When speakers are epistemically uncertain, they should be more likely to incorporate the audience-tuned view into their mnemonic representation. Dr. Echterhoff discusses recent work that demonstrates the importance of epistemic motives in creating a "shared reality" and explores the implications for the social sharing of memories.

## Alin Coman

Do the conversational influence observed by Drs. Stone and Echterhoff, among others, propagate across a network of individuals? Are there limits to the extent of this propagation? How do these limits shape the formation of collective memories within the network? Dr. Coman addresses these questions using agent-based modeling techniques in order to investigate how macro-level social phenomena, such as mnemonic convergence, can emerge out of psychologically informed micro-level local dynamics. The findings show that communicative influences at the dyadic level, as well as network factors (network size, network topology, number of conversations) interact to drive mnemonic convergence in communities.

## Micah Edelson

The underlying brain mechanisms through which the social environment influences human memory are unknown. Dr. Edelson reports on studies of memory in individuals exposed to contradictory recollections of others. Using fMRI, the research finds that social milieu influences memory through evolutionary conserved medial temporal circuits, whereas the ability to overcome social influence is mediated by newer prefrontal circuits. The interplay between these systems determines the balance of personal memory versus conformity to the group.

## Kourken Michaelian

Much of the psychological work on communication and memory evokes the concept of collective memory, in that conversations are considered one critical means for forming a collective memory. Dr. Michaelian explores the philosophical underpinnings of the concept of collective memory. Although some scholars treat the concept as more than a metaphor that likens collective memory to individual memory, Dr. Michaelian argues that, when the process of consolidation is considered, critical differences exist between individual and collective memory that suggests that the concept of collective memory is metaphorical at best.

## William Hirst

Building on his extensive work on conversational remembering and the formation of collective memory, Hirst relates these different approaches to the study of communication and memory and articulates how they provide a foundation for understanding how conversations can influence individual and collective memory.

## Acknowledgements

Dr. Echterhoff would like to acknowledge the contributions of Drs. Olivier Klein (Free University of Brussels) and René Kopietz (University of Münster), and Sabrina Pierucci (Free University of Brussels); Dr. Stone would like to acknowledges, the contributions of Drs. Olivier Luminet, (Université catholique de LouvainCatholic University of Louvain), Olivier Klein (Free University of Brussels) and Laurent Licata (Free University of Brussels).

# Beyond Synchrony: Complementarity and Asynchrony in Joint Action 

Rick Dale (rdale@ucmerced.edu)<br>Cognitive and Information Sciences, UC Merced, 5200 N. Lake Road<br>Merced, CA 95343<br>Riccardo Fusaroli (semrf@hum.au.dk)<br>Center for Semiotics and the Interacting Minds Center, Aarhus University, Jens Chr. Skous Vej 2, 8000 Aarhus C, Denmark<br>Dorthe Døjbak Håkonsson (dod@asb.dk)<br>Interdisciplinary Center for Organizational Architecture, Fuglesangs Allé 20, Building 2635 I106<br>8210 Aarhus V, Denmark<br>Patrick Healey (ph@eecs.qmul.ac.uk)<br>Department of Computer Science and Electronic Engineering, Queen Mary University of London, Mile End Road London E1 4NS<br>Dan Mønster (danm@asb.dk)<br>School of Business and Social Sciences, Department of Economics and Business, Fuglesangs Allé 4 8210 Aarhus V, Denmark<br>John J. McGraw (iksjmc@hum.au.dk)<br>TESIS Network, Østboulevarden 11F, 2 8000 Aarhus C, Denmark<br>Panagiotis Mitkidis (mitkidispan@gmail.com)<br>Center for Advanced Hindsight, Social Science Research Institute, Duke University, 2024 West Main Street Durham, NC 27705<br>Kristian Tylén (semkt@hum.au.dk)<br>Center for Semiotics and the Interacting Minds Center, Aarhus University, Jens Chr. Skous Vej 2<br>8000 Aarhus C, Denmark

Keywords: Joint action; distributed cognition; social cognition; interpersonal coordination.

## Summary of Topic

Recent advances in social cognition and joint action reveal the social and the mutual, rather than the individual and the dichotomous aspects of cognition (Hasson, Ghazanfar, Galantucci, Garrod, \& Keysers, 2012). A widespread and powerful model of socially interactive behavior is 'synchrony' (Jirsa \& Kelso, 2004): Numerous studies have thus recently indicated how individuals through social interaction become increasingly entrained on multiple levels from physiology to syntax: through interaction people synchronize their heart rates, their subtle postural sways, their gestures and gaze behaviors, align their lexicon and their syntax (Fusaroli \& Tylén, 2012; Louwerse, Dale, Bard, \& Jeuniaux, 2012; Pickering \& Garrod, 2004). However, emerging scholarship is increasingly attending to many instances in which patterns of complementary and asynchronous actions rather than synchronous ones seem to predict high levels of interpersonal coordination and joint performance. While some activities such as expertly timed
rowing may afford interacting agents to synchronize their individual behaviours to reach high levels of joint performance, other types of joint activity - like playing a game of baseball - rather afford complementary actions: i.e. tightly coupled, reciprocal activity derived from different behaviours performed across an extended temporal sequence. Shared construction tasks as well as task-oriented dialogues, for instance, have been shown to require smooth turn-taking, and the development of interactional routines which might involve complementary roles (Dale, Fusaroli, Duran, \& Richardson, in press; Fusaroli, RaczaszekLeonardi, \& Tylén, accepted). Cultural practices dwell upon and stabilize complementary distribution of work, to make challenging task as the sailing of a warship or the construction of huge buildings possible (Hutchins, 1995; Perry, 2010).

The session will address the implications and respective roles of synchrony, complementarity and asynchrony as components of coordination. Different methods and perspectives for quantifying and assessing coordinative dynamics in language, behaviour and physiology will be presented conceptually and in their empirical application.

## Speakers

Rick Dale (moderator) is a cognitive scientist at UC Merced. He has worked and published extensively on language and social interaction developing and applying a range of novel non-linear statistical methods to assess dynamical properties of multimodal social coordination (Dale, et al., in press; Dale \& Spivey, 2006; Louwerse, et al., 2012; Tollefsen \& Dale, 2012).

Patrick Healey is a professor of human interaction and head of the Interaction Media and Communication research group at University of London. His research concerns experimental work on technology-mediated dialogical communication and - in particular - miscommunication (Healey, Howes, \& Purver, 2010; Healey \& Mills, 2006; Mills \& Healey, 2008).

John J. McGraw, cognitive anthropologist (TESIS, a Marie Curie Initial Training Network), and Panagiotis Mitkidis, cognitive psychologist (Interacting Minds Centre at Aarhus University and Center for Advanced Hindsight at Duke University) investigate the role of objects and material structures in the coordination of behavior, cognition, and the enhancement of cooperation (Xygalatas et al., accepted).

Dorthe Døjbak Håkonsson is an associate professor at the Aarhus School of Business. As an organization scientist, her research focuses on how team shared emotions influence organizational decision-making (Håkonsson, Burton, Obel, \& Laurdisen, 2012). Dan Mønster is a physicist and assistant professor in the Department of Economics and Business at Aarhus University. His current research interest is investigating interactions among team members and the effects of these interactions on team decisions and team performance.

Kristian Tylén and Riccardo Fusaroli are both post doctoral fellows at the Center for Semiotics and the Interacting Minds Center, Aarhus University, with a background in semiotics and cognitive science. They have published on experimental and dynamical systems approaches to social coordination - in particular taskoriented dialogue (Fusaroli et al., 2012; Fusaroli, et al., accepted; Fusaroli \& Tylén, 2012).

## Acknowledgments

Research was supported by NSF BCS-0826825 (RD), EPSRC (PH), TESIS, a Marie Curie Initial Training Network (JJM, PM), MINDLab, Danish Ministry of Science, Technology and Innovation (DDH, DM), EuroUnderstanding (Eurocore) and the Danish council for Independent Research (RF, KT).

## References

Dale, R., Fusaroli, R., Duran, N., \& Richardson, D. C. (in press). The Self-Organization of Human Interaction. Psychology of Learning and Motivation, xxx(xx), xxx-xxx

Dale, R., \& Spivey, M. J. (2006). Unraveling the dyad: Using recurrence analysis to explore patterns of syntactic coordination between children and caregivers in conversation. Language Learning, 56(3), 391-430
Fusaroli, R., Bahrami, B., Olsen, K., Rees, G., Frith, C. D., Roepstorff, A., \& Tylén, K. (2012). Coming to terms: an experimental quantification of the coordinative benefits of linguistic interaction. Psychological science, 23(8)
Fusaroli, R., Raczaszek-Leonardi, J., \& Tylén, K. (accepted). Dialogue as interpersonal synergy. New Ideas in Psychology, xxx(xx), xxx-xxx
Fusaroli, R., \& Tylén, K. (2012). Carving Language for Social Interaction: a dynamic approach Interaction studies, 13(1), 103-123
Håkonsson, D. D., Burton, R. M., Obel, B., \& Laurdisen, J. (2012). Strategy implementation requires the right executive style. Long range planning, 45(2-3), 182208
Hasson, U., Ghazanfar, A. A., Galantucci, B., Garrod, S., \& Keysers, C. (2012). Brain-to-brain coupling: a mechanism for creating and sharing a social world. Trends in cognitive sciences
Healey, P. G. T., Howes, C., \& Purver, M. (2010). Does structural priming occur in ordinary conversation? Paper presented at the Linguistic Evidence, Tübingen, Germany
Healey, P. G. T., \& Mills, G. (2006). Participation, Precedence and Co-ordination in Dialogue. Paper presented at the 28th Annual Conference of the Science Society, Vancouver, Canada
Hutchins, E. (1995). Cognition in the Wild. Cambridge, Mass.: MIT Press
Jirsa, V. K., \& Kelso, J. A. S. (2004). Coordination dynamics : issues and trends. Berlin ; New York: Springer
Louwerse, M. M., Dale, R., Bard, E. G., \& Jeuniaux, P. (2012). Behavior Matching in Multimodal Communication Is Synchronized. Cognitive Science, 36(8), 1404-1426
Mills, G., \& Healey, P. G. T. (2008). Semantic negotiation in dialogue: the mechanisms of alignment. Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue, 46-53
Perry, M. (2010). Socially distributed cognition in loosely coupled systems. AI \& Society, 1-14
Pickering, M. J., \& Garrod, S. (2004). Toward a mechanistic psychology of dialogue. Behavioral and Brain Sciences, 27(02), 169-190
Tollefsen, D., \& Dale, R. (2012). Naturalizing joint action: A process-based approach. Philosophical Psychology, 25(3), 385-407
Xygalatas, D., Mitkidis, P., Fischer, R., Reddish, P., Skewes, J., Geertz, A. W., . . . Bulbulia, J. (accepted). Extreme Rituals Promote Prosociality. Psychological Science, $x x(\mathrm{x})$, $\mathrm{xxx}-\mathrm{xxx}$

# Computational and Cognitive Aspects of Narratives 

Mark A. Finlayson ${ }^{1}$, Bernhard Fisseni ${ }^{2,8}$, Dedre Gentner ${ }^{3}$, Richard Gerrig ${ }^{4}$, Benedikt Löwe ${ }^{5,8}$, Jeffrey Loewenstein ${ }^{6}$, Inderjeet Mani ${ }^{7}$, Jan Christoph Meister ${ }^{8}$, R. Michael Young ${ }^{9}$<br>${ }^{1}$ Massachusetts Institute of Technology (markaf@mit.edu)<br>${ }^{2}$ Universität Duisburg-Essen (bernhard.fisseni@uni-due.de)<br>${ }^{3}$ Northwestern University (gentner@northwestern.edu)<br>${ }^{4}$ Stony Brook University (richard.gerrig@stonybrook.edu)<br>${ }^{5}$ Universiteit van Amsterdam (b. loewe@uva.nl)<br>${ }^{6}$ University of Illinois at Urbana-Champaign (jloew@illinois.edu)<br>${ }^{7}$ Children's Organization of Southeast Asia (inder jeet.mani@gmail.com)<br>${ }^{8}$ Universität Hamburg (\{bernhard.fisseni, benedikt. loewe, jan-c-meister\}@uni-hamburg. de)<br>${ }^{9}$ North Carolina State University (young@csc.ncsu.edu)


#### Abstract

Narrative, a distinctly cognitive phenomenon, has long been of interest to the disciplines that comprise cognitive science. The past decade has seen a resurgence of work using computational methods to understand, manipulate, generate, and leverage narratives. This symposium, which is held in association with the Fourth International Workshop on Computational Models of Narrative (CMN'13), a satellite event of CogSci 2013, will focus on aspects of the scientific and computational understanding of narrative that intersect with cognitive science. The speakers and moderators are drawn from diverse fields including cognitive psychology, artificial intelligence, cognitive science, computational linguistics, and the humanities, and they will focus on a variety of topics including: narrative and its role in analogy, education, and persuasion; challenges in the representation of syntax, discourse, and semantics of narrative; psychological and neuropsychological aspects of narrative; and the growing integration of computational models of narrative in humanities research.


Keywords: Computational Models of Narrative
Narratives are ubiquitous in human experience. We use them to communicate, convince, explain, and entertain. As far as we know, every society in the world has narratives, which suggests they are rooted in our psychology and serve an important cognitive function. It is becoming increasingly clear that, to truly understand and explain human intelligence, beliefs, and behaviors, we will have to understand why narrative is universal and explain the function it serves.

Cognitive science has long recognized the importance of narrative as an activity that engages diverse and important cognitive facilities, as well as a phenomenon that is worthy of study in its own right. After a long period of dormancy, the past decade has seen a resurgence of interest in the formal understanding and computational modeling of of narrative, as well as a more recent boom in cognitive, psychological, and neuroscientific studies relating to narrative. This symposium is an attempt to catalyze the interaction between the research community working on computation-compatible approaches to narrative with cognitive science community proper.

## The Computational Models of Narrative Workshop Series

The research community in question aims to advance the scientific understanding of narrative through progress across a
wide range of fields including cognitive science, psychology, computer science, artificial intelligence, sociology, anthropology, linguistics, logic, and philosophy. To foster and encourage this community, the Computational Models of Narrative ${ }^{1}(\mathrm{CMN})$ workshop series was founded in 2009 . The series was so titled because we believe that a true science of narrative must adhere to the principle espoused by Herbert Simon in his book The Sciences of the Artificial: that without computational modeling the science of a complex human phenomenon such as narrative will never be successful, and that computational models are the proper lingua franca of such a diverse, inter-disciplinary community.

Because the workshop series is relatively new, and the research community is still growing and developing, it was decided to embed the early workshops in the conferences of different intersecting communities, in order to promote cross-fertilization and a more diverse membership. In 2010, the second workshop was hosted by the Association for the Advancement of Artificial Intelligence (AAAI) as one of its Fall Symposia. In 2012, the third workshop was hosted by the Language Resources and Evaluation Conference (LREC), which is a part of the computational linguistics community.

In 2013, the Fourth Workshop on Computational Models of Narrative will be held in Hamburg, Germany, directly after CogSci 2013 as a satellite event. The symposium described here is the on-site event associated with the workshop, and is intended to highlight the intersection between cognitive science and the topics covered in the workshop.

## Speakers

The symposium will be moderated jointly by the workshop co-chairs, Mark Finlayson and Benedikt Löwe. Our speakers and moderators span multiple fields, highlighting the inter-disciplinarity of this symposium. The moderators bring expertise in artificial intelligence, computer science, philosophy, and mathematical logic. Among the speakers, Jeffrey Loewenstein and Dedre Gentner represent cognitive science proper, and they will discuss applying work on analogy to advance our understanding of business practice and psychol-

[^5]ogy, a field where narratives (cases) are of great importance. Richard Gerrig is a cognitive psychologist who has written extensively on human subject experiments regarding people's experiences on reading narrative. Inderjeet Mani is a computational linguist by training, and an expert in corpus annotation and linguistic representations for capturing narratives. Jan Christoph Meister is a humanist who seeks to apply the many advances in computational modeling of narrative to scholarly advances of relevance to his field. Michael Young is a computer scientist who pursues the scientific understanding of narrative to advance the state of the art in digital games.

Each subsection that follows outlines the topic which will be discussed by each speaker. The symposium is structured to leave time for a general discussion that includes the moderators and the audience.

## Loewenstein \& Gentner: Narrative Knowledge and "Repetition-Break" Plot Structures

Narratives convey causal, temporal, and other kinds of relational knowledge, the sort of knowledge that comprises expertise. Comparing narratives is one of the quickest and most powerful ways to develop expert knowledge, as indicated by the analogy literature. Analogy provides a means for identifying commonalities, and in particular, structural commonalities, and so provides a basis for revealing narrative structure. A further insight is the pairing of repetition with contrast to form a plot structure that serves as a recipe for surprise. Folktales (e.g., the three billy goats gruff), jokes (e.g., three guys walk into a bar. . . ), advertisements (e.g., MasterCard's priceless campaign), and other types of narratives frequently make use of these "repetition-break" plots. The repetition in the narrative structure leads audiences to draw comparisons, perhaps learning something, and forming expectations that more similar items will follow. The break or contrast in the narrative structure surprises audiences by deviating from their expectations. Narratives with repetition-break plots are prevalent, often well-liked, often socially-selected both by popular attention and expert judges, and capable of influencing audiences towards adopting the narrative's views. This is just one example of the potential for narrative structure to yield significant returns.

## Gerrig: Readers' Participation in Narrative Experiences

Gerrig will outline a participatory perspective on readers' experiences of narratives. He proposes that readers encode types of mental contents (called participatory responses) that fall beyond the ordinary scope of computational models of narrative. Readers, for example, encode responses toward characters actions and preferences for particular outcomes. Those participatory responses vary from reader to reader and structure their individual experiences toward narratives.

## Mani: Naturalness and Computability in Computational Narrative Representations

Humans have an astonishing ability to infer different facets of narrative structure from a description of events. These facets include the representation of the temporal order of events as well as the motivation behind the actions and reactions of agents based on their goals and beliefs. Computational representations of narrative time and plot can be assessed in terms of their naturalness for humans to infer as well as their computability. Such an assessment suggests a number of psychological investigations that could help provide constraints on formal aspects of these representations.

## Meister: Statistical vs. Intelligent Modelling of Human Narrative Processing

The former Google CEO's Eric Schmidt's dictum "In God we trust-all others bring data" epitomizes the neo-positivist underpinning of 'big data' approaches to complex phenomena. Search engine algorithms aim at representing such phenomena in terms of mathematical and statistical phenomena which by-pass human intelligence. Meister's talk will aim to reinterpret what looks like a purely methodological decision as an ethically problematic choice that is based on a reductionist concept of intelligent behavior. Meister will use examples from narrative to illustrate and emphasize his points.

## Young: Cognition as the Decider: Comprehension and the Next Steps to Plan-Based Narrative Generation

A range of methods for the automatic construction of narratives have been developed in the last 10 years, many of them based on automatic planning methods drawn from AI. While these techniques have benefited from extensions to the standard knowledge representations that target interesting aspects of narrative structure, they often leave out the role of narrative at the discourse level. These models speak to only part of the functional properties of narrative as a result. In this talk, Young will point to these limitations in his own work and describe new models being developed that focus on narrative as a means to prompt comprehension on the part of the reader/viewer/player. These new models attempt to produce narrative story and discourse in text, video and video games that create experiences for their users rather than focus solely on the creation of appropriate narrative structural elements.

## Acknowledgments

This symposium is supported by the John Templeton Foundation (JTF) via the project What makes stories similar? (Grant No. 20565).

# Structured Cognitive Representations and Complex Inference in Neural Systems 

Samuel J. Gershman, Joshua B. Tenenbaum (\{sjgershm,jbt\}@mit.edu)<br>Department of Brain and Cognitive Sciences, MIT<br>Cambridge, MA 02139 USA

Alexandre Pouget (Alexandre.Pouget @unige.ch)
Department of Neuroscience, University of Geneva
CH-1211 Geneva 4 Switzerland
Matthew Botvinick (matthewb@princeton.edu)
Department of Psychology, Princeton University
Princeton, NJ 08540 USA

Peter Dayan (dayan@gatsby.ucl.ac.uk)<br>Gatsby Computational Neuroscience Unit, University College London London WC1N 3AR United Kingdom

Keywords: Bayesian models, rational analysis, perception, olfaction, memory


#### Abstract

Summary The dream of cognitive neuroscience has always been a seamless integration of cognitive representations with neural machinery, but-despite decades of work-fundamental gaps remain. Part of the problem is that many contemporary theories of cognition are formulated in terms of representations and computations that are quite different from those used in computational neuroscience. Bridging this gap requires more than simply a translation between theoretical concepts in the two fields; what is needed is a more radical updating of neuroscience's theoretical vocabulary.

What should this vocabulary look like? Some important features of representations and computations used in contemporary cognitive theories are:


- Compositional, recursive and relational representations (Fodor, 1975; Smolensky, 1990; Hummel \& Holyoak, 2003; Stewart et al., 2011).
- Flexible use of different structural forms (e.g., taxonomic vs. causal knowledge; Kemp \& Tenenbaum, 2009).
- Multiple levels of abstraction (Tenenbaum et al., 2011).
- Knowledge partitioning / clustering (Lewandowsky \& Kirsner, 2000).
- Complex intuitive theories (e.g., naive physics, theory of mind; Carey, 2009).
- Algorithms that operate on these representations (e.g., dynamic programming, Monte Carlo methods; Griffiths et al., 2012).

These representations and computations are "structured" in the sense that they incorporate rich domain knowledge and strong constraints (Tenenbaum et al., 2011).

This symposium addresses the question: how do neural circuits acquire and compute with structured representations? This question is examined from a number of angles. Gershman introduces the basic issues and discusses attempts to articulate a neurally plausible theory of structured cognition. Pouget describes recent work on implementing complex probabilistic computations in neural circuits. Botvinick shows how neural circuits can be used to discover hierarchical task structure in the environment. Finally, Dayan discusses work on wedding richly structured models of semantics with representations of individual episodes. Each talk will be 20 minutes long, followed by a 20 minute panel discussion with speakers moderated by Tenenbaum.

## Gershman: from knowledge to neurons

How can neurons express the structured knowledge representations central to intelligence? This problem has been attacked many times from various angles. I discuss the history of these attempts and situate our current understanding of the problem. I then outline a new approach based on the idea of compressing structured knowledge using neurons in a way that supports probabilistic inference. I illustrate this approach using examples from motion perception and value-based decision making.

## Pouget: modeling the neural basis of complex intractable inference

It is becoming increasingly clear that neural computation can be formalized as a form of probabilistic inference. Several hypotheses have emerged regarding the neural basis of these inferences, including one based on a type of code known as probabilistic population codes or PPCs (Ma et al., 2006). PPCs have been used to model simple forms for multisensory integration, attentional search, perceptual decision making or causal inference, for which human subjects have been shown to be nearly optimal. However, most inferences performed by the brain are too complex be solved optimally in a reasonable
amount of time and must therefore involve approximate solutions. We have started to explore how neural circuits could implement a particular form of approximation, called variational Bayes, with PPCs (Beck et al., 2012). Remarkably, this approximation requires a nonlinearity known as divisive normalization which has already been found in most neural circuits. This approach can be applied to a wide range of complex inferences, such as the ones involved in olfactory processing, image processing in the primary visual cortex and other related problems.

## Botvinick: discovering hierarchical task structure

Naturalistic action displays a hierarchical structure: Simple actions cohere into subtask sequences or component skills, which in turn combine to realize overall goals. Computational models from cognitive psychology, artificial intelligence, and most recently neuroscience, have sought to characterize the representations and mechanisms underlying hierarchical action control (Botvinick, 2008). However, such models tend to neglect a fundamental question: How do hierarchical representations of action or task structure initially arise? We approach this as a learning problem, asking how useful component skills can be inferred from experience. Behavioral evidence suggests that such learning arises from a structural analysis of encountered problems, one that maximizes representational efficiency and, as a direct result, decomposes task into subtasks by 'carving' them at their natural 'joints.' A key question is how this analysis and optimization process might be implemented neurally. Recent data suggests an intriguing answer: Detection of hierarchical task structure might arise as a natural consequence of predictive representation. I'll present computational work fleshing out this possibility, along with behavioral and fMRI data that lend it considerable initial support.

## Dayan: unsupervised learning and the representation of episodic structure

The representation of hierarchically structured knowledge in systems using distributed patterns of activity is an abiding concern for the connectionist solution of cognitively rich problems. One particularly important unresolved issue concerns episodic versus semantic structure-how rich generative models of the semantics of domains can be used in the representation of particular, structured, entities. I will unpack this problem and suggest some routes to solutions.

## References

Beck, J., Heller, K., \& Pouget, A. (2012). Complex inference in neural circuits with probabilistic population codes and topic models. In Advances in neural information processing systems 25 (pp. 3068-3076).
Botvinick, M. (2008). Hierarchical models of behavior and prefrontal function. Trends in Cognitive Sciences, 12(5), 201-208.

Carey, S. (2009). The origin of concepts. Oxford University Press, USA.
Fodor, J. (1975). The language of thought. Harvard University Press.
Griffiths, T., Vul, E., \& Sanborn, A. (2012). Bridging levels of analysis for probabilistic models of cognition. Current Directions in Psychological Science, 21(4), 263-268.
Hummel, J., \& Holyoak, K. (2003). A symbolicconnectionist theory of relational inference and generalization. Psychological Review, 110(2), 220-264.
Kemp, C., \& Tenenbaum, J. (2009). Structured statistical models of inductive reasoning. Psychological Review, 116(1), 20-58.
Lewandowsky, S., \& Kirsner, K. (2000). Knowledge partitioning: Context-dependent use of expertise. Memory \& Cognition, 28(2), 295-305.
Ma, W., Beck, J., Latham, P., \& Pouget, A. (2006). Bayesian inference with probabilistic population codes. Nature Neuroscience, 9(11), 1432-1438.
Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. Artificial intelligence, 46(1), 159-216.
Stewart, T., Bekolay, T., \& Eliasmith, C. (2011). Neural representations of compositional structures: representing and manipulating vector spaces with spiking neurons. Connection Science, 23(2), 145-153.
Tenenbaum, J., Kemp, C., Griffiths, T., \& Goodman, N. (2011). How to grow a mind: Statistics, structure, and abstraction. Science, 331(6022), 1279-1285.

# Cognitive Foundations of Cultural Learning 

Annette Henderson (a.henderson@auckland.ac.nz)<br>School of Psychology, The University of Auckland<br>Private Bag 92019, Auckland 1142, New Zealand<br>Ben Kenward (ben.kenward@psyk.uu.se)<br>Department of Psychology, Uppsala University<br>Box 1225, 75142 Uppsala, Sweden<br>Rachel Watson-Jones (watsonjones@austin.utexas.edu)<br>Cristine Legare (legare@austin.utexas.edu)<br>Department of Psychology, The University of Texas at Austin<br>1 University Station \#A8000, Austin, Texas 78712, USA

Harvey Whitehouse (harvey.whitehouse@anthro.ox.ac.uk)<br>School of Anthropology and Museum Ethnography, University of Oxford<br>51 Banbury Road, Oxford OX2 6PE, England

Keywords: Ritual; imitation; collaboration; affiliation; cultural learning; normativity; convention

## Cognitive Foundations of Cultural Learning

The ability to learn from others is integral to sustaining and transmitting human culture. What are the cognitive processes that support imitative and collaborative cultural learning? How does cultural learning contribute to group dynamics, such as cohesion and conflict? Recent research has focused on how children acquire instrumental skills through causal inference (Call, Carpenter, \& Tomasello, 2005; Whiten, McGuigan, Marshall-Pescini, \& Hopper, 2009). However, children also need to acquire the norms and conventions of their culture, as well as an understanding of cooperative behavior, to become full-fledged members of their community. This acquisition begins early in ontogeny and is likely reliant on a unique mix of causal reasoning and affiliative goals, triggered by the nature of the action sequence itself and a variety of social cues. In this symposium, we consider the emerging experimental literature on the development of imitation and collaboration with the goal of applying this work to broader issues of group dynamics and the transmission of culture. Henderson will consider the understanding of collaborative goals in infancy. Kenward will consider the normative basis of young children's over-imitation. Watson-Jones will examine affiliative motivations underlying children's imitation. Whitehouse will consider how ritualized, normative behavior and cognition impacts group dynamics of coordination and social cohesion.

## Henderson: Infants' Understanding of Cooperative Action

Cooperative activities in which two (or more) individuals coordinate their independent actions to attain a common
goal are critical for human survival. Food, shelter, reproduction, protection from threats and knowledge transfer require cooperation (Barkow, Cosmides, \& Tooby, 1992; Tomasello, 2009). The ability to cooperate with others emerges early in development (for a review see Brownell, 2011) and plays a critical role in facilitating children's socio-cognitive development (Rogoff, 1990; Sommerville \& Hammond, 2007). Despite the growing body of evidence documenting the development of cooperative behaviour across the first few years of life, very little is known about infants' understanding of cooperation. In her talk, Annette Henderson will present new findings from a series of studies investigating the age at which infants understand that the actions of cooperating partners are directed towards the attainment of a common goal using an innovative visual habituation paradigm. Identifying when and how infants come to understand cooperation provides valuable information about the age at which infants possess the mental capacity to make sense out of the vast array of cooperative actions that they witness in their everyday lives.

Annette Henderson is a Senior Lecturer in Developmental Psychology and Director of the Early Learning Laboratory at the University of Auckland, New Zealand. Her research interests include the development of children's understanding of activities involving shared intentions, which include linguistic and non-linguistic cooperative activities.

## Kenward: The Mechanisms Behind Imitation of "Unnecessary" Actions

Why do children imitate actions with unclear functions? Copying behavior without understanding the function of the behavior is often beneficial - you don't need to know about the crocodile to benefit from copying avoidance of the river. This argument does not explain the mechanism of such blind copying, however. Various proposals have been put
forward, such as an automatic assumption of causal effect (Lyons, Young, \& Keil, 2007), and a motive to socially affiliate (Over \& Carpenter, 2012). Evidence will be presented demonstrating that when imitating a simple action sequence with a clear goal but including irrelevant actions, children do not believe that the irrelevant actions are causally necessary for achieving the goal. This evidence speaks against the hypothesis of causal assumption. Secondly, evidence will be presented demonstrating that children have a sufficiently strong belief that such irrelevant actions should be performed that they protest when a third party does not perform them. This shows that children have a normative belief that it is correct to copy actions performed by adults, even when their purpose is unclear. Kenward argues that children's ability to encode actions as normative without any information about what domain determines the normativity implies that children are capable of holding normative beliefs that are not anchored in a specific domain, such as convention, morality, or instrumental rationality.

Ben Kenward is a Researcher in Psychology at Uppsala University, Sweden. He focuses on development but has professional ADHD, having published recently on unconsciously motivated action, moral development, social learning, the development of decision making, and animal tool use.

## Watson-Jones \& Legare: Affiliative Motivations and the Development of Imitation

Imitation is used to acquire both instrumental skills (Call, et al. 2005; Lyons et al., 2007; Whiten et al., 2009) as well as cultural conventions, such as ritual. Legare, Whitehouse, Herrmann, and Wen (under review) have proposed that the causal opacity associated with cultural conventions results in high fidelity copying and low levels of innovation. The motivation to engage in imitation may be fundamentally related to implicit affiliation goals (Over \& Carpenter, 2012). We propose that affiliative motivations play an important role in the transmission and learning of cultural knowledge. Rachel Watson-Jones will present new research in collaboration with Cristine Legare and Harvey Whitehouse, demonstrating that affiliation goals may differentially affect the imitation of instrumental actions versus actions related to cultural conventions. Using ostracism as a conceptual lens, the findings of this research provide evidence of the affiliative basis of imitation.

Rachel Watson-Jones is a Postdoctoral Fellow at the University of Texas at Austin within the Cognition, Culture, and Development Laboratory. Her interests include the development of social cognition, cultural transmission, and the cognitive science of religion.

## Whitehouse: Ritual, Community, and Conflict

Some of the greatest atrocities have been caused by groups defending or advancing their political aspirations and sacred values. In order to comprehend and address the wanton violence of war, terrorism and genocide, it is necessary to
understand the forces that bind and drive human groups. Here I describe a five-year program of research investigating one of the most powerful mechanisms by which groups may be formed, inspired, and coordinated: ritual. The project examines the role of ritual in child development, in social behavior, and in the evolution of political systems. Studying how children learn the rituals of their communities is shedding light on the various ways in which rituals promote social cohesion within the group and distrust of groups with different ritual traditions. Qualitative field research, surveys, and controlled psychological experiments are being conducted in a number of troubled regions (including the Middle East and North Africa) to investigate the role of ritual in group bonding and intergroup competition. New databases are being constructed to explore the relationship between ritual, resource extraction patterns, and group structure and scale over the millennia.

Harvey Whitehouse is professor of social anthropology, director of the Institute of Cognitive and Evolutionary Anthropology, and a fellow of Magdalen College at the University of Oxford. His interests include recurrence and variation in religious thinking and behavior, and he has published many books and articles on this topic.

## References

Barkow, J. H., Cosmides, L., \& Tooby, J. (1992). The adapted mind: Evolutionary psychology and the generation of culture. NY: OUP.
Brownell, C. A. (2011). Early developments in joint action. Review of Philosophy and Psychology, 2, 193-211.
Call, J., Carpenter, M., \& Tomasello, M. (2005). Copying results and copying actions in the process of social learning: Chimpanzees and human children. Animal Cognition, 8, 151-163.
Legare, C. H., Whitehouse, H., Herrmann, P. A., \& Wen, N. J. (under review). The imitative foundations of cultural learning.
Lyons, D. E., Young, A. C., \& Keil, F. C. (2007). The hidden structure of over-imitation. PNAS, 104, 1975119756.

Over, H. \& Carpenter, M. (2012). Putting the social into social learning: Explaining both selectivity and fidelity in children's copying behavior. Journal of Comparative Psychology, 126(2), 182-192.
Rogoff, B. (1990). Apprenticeship in thinking: Cognitive development in social context. NY: OUP.
Sommerville, J. A., \& Hammond, A. J. (2007). Treating another's actions as one's own: Children's memory of and learning from joint activity. Developmental Psychology, 43, 1003-1018.
Tomasello, M. (2009). Why we cooperate. Cambridge: The MIT Press.
Whiten, A., McGuigan, N., Marshall-Pescini, S., \& Hopper, L. M. (2009). Emulation, imitation, over-imitation, and the scope of culture for child and chimpanzee. Philosophical Transactions of the Royal Society B, 364, 2417-2428.

# Constraints on Bayesian Explanation 

Johan Kwisthout (j.kwisthout@donders.ru.nl), Iris van Rooij (i.vanrooij@donders.ru.nl) [moderator]<br>Radboud University Nijmegen, Donders Institute for Brain, Cognition and Behaviour Montessorilaan 3, 6525 HR Nijmegen, The Netherlands

Matteo Colombo (m.colombo@uvt.nl)<br>Tilburg Center for Logic and Philosophy of Science, Tilburg University PO Box 90153, 5000 LE Tilburg, The Netherlands<br>Carlos Zednik (czednik@uos.de)<br>Institute of Cognitive Science, University of Osnabrück 49069 Osnabrück, Germany

William A. Phillips (w.a.phillips@stir.ac.uk)<br>University of Stirling, School of Natural Sciences Stirling FK9 4LA Scotland, UK, and Frankfurt Institute of Advanced Studies, Germany

Keywords: Bayesian inference, levels of explanation, constraints, tractability, coherent infomax, computational complexity, unification, philosophy of science

## Introduction

The hypothesis that human cognition may be well characterized as a set of Bayesian computations has been the topic of considerable debate over the last two decades. Recently, critics have argued that this hypothesis is either unlikely to be true or otherwise too unconstrained to be particularly useful for explaining cognition (e.g., Bowers \& Davis, 2012), whereas proponents have defended their position by stating that the Bayesian perspective has been misunderstood, is not necessarily in conflict with other perspectives on cognition, and can still be explanatorily useful as a framework for cognitive science even if underconstrained in many ways (e.g., Griffiths, Chater, Norris, \& Pouget, 2012). Our position in this debate is that both sides of this debate may be right as well as wrong: Proponents may be right that the Bayesian perspective has something uniquely useful to bring to cognitive science (and then the critics are wrong in their denial of this); yet, the critics may be right that cognitive theories are explanatorily useful only if properly constrained (and then proponents are wrong in their denial of this).

With this perspective in mind, we wish to move the debate forward in a constructive way by bringing in new perspectives and proposing novel constraints that can be exploited for purposes of improving the explanatory values and virtues of Bayesian explanations of cognition. Specifically, with this symposium we aim to focus on how constraints on Bayesian explanations can be exploited in ways that are yet underrepresented and underexplored.

The symposium brings together researchers from various disciplines, contributing a variety of perspectives on how Bayesian explanations can be fruitfully constrained, drawing on theories, analyses, and results from philosophy of
science, cognitive neuroscience, information theory, machine learning, and theoretical computer science.

## A complexity-theoretic perspective on the preconditions for Bayesian tractability

## Johan Kwisthout (joint work with Iris van Rooij)

Many Bayesian computations have been proven to be computationally intractable (NP-hard) for unconstrained input domains, even if only an approximate solution is sought. Informally, this means that computations postulated by Bayesian models can take astronomical amounts of time for their completion even for realistic sized inputs. This property seems to be in strong contrast with the ease and speed with which humans can typically make the inferences that are modeled by Bayesian models. Some critics of the Bayesian approach have taken this property of Bayesian models as a reason to reject the entire approach (e.g., Gigerenzer, 2008). In contrast, I propose that it means that tractability forms a useful constraint on Bayesian explanations of cognition. In this talk, I will elucidate the use of complexity-theoretic concepts and techniques for making Bayesian models meet the tractability constraint, building on known results from theoretical computer science (e.g., Kwisthout, 2011). I will furthermore report on recent complexity results that have lead to novel hypotheses about the conditions under which Bayesian inferences can be tractably approximated (Kwisthout \& van Rooij, 2013).

## Bayesian cognitive science, unification, and explanation

Matteo Colombo (joint work with Stephan Hartmann)
A recurrent claim is that the greatest value of studying cognitive phenomena such as perception, action, categorization, and decision-making, within the Bayesian framework consists in its unifying power. Several Bayesian
cognitive scientists, however, implicitly assume that unification is obviously linked to explanatory power. But this link is not obvious (e.g., Morrison, 2000).

A crucial feature of adequate explanations in the cognitive sciences is that they reveal aspects of the causal structure of the mechanism that produces the phenomenon to be explained. The kind of unification afforded by the Bayesian framework to cognitive science does not necessarily reveal the causal structure of a mechanism (cf. Colombo \& Seriès, 2012). Bayesian unification is the product of the mathematics rather than of a causal hypothesis concerning how different cognitive phenomena are brought about by a single type of mechanism. Nonetheless, Bayesian unification can place fruitful constraints on causal mechanical explanation, which will be elucidated in this talk.

## Bayesian modeling and heuristic strategies for model-development

## Carlos Zednik (joint work with Frank Jäkel)

It is generally agreed that Bayesian models in cognitive science operate at Marr's computational level of analysis (Marr, 1982). Unfortunately, it remains unclear exactly how the computational, algorithmic, and implementation levels are related.

This talk explicates inter-level relationships in terms of heuristic strategies for model-development (Zednik, in press). Specifically, Bayesian computational-level models play the heuristic role of suggesting possible algorithms to compute a particular function, and of suggesting particular ways of delineating and interpreting the components of a physical mechanism. In turn, algorithmic and mechanistic models specify memory, time, and resource limitations that constrain the cognitive tasks described by Bayesian models. In contrast to the view that Bayesian computational-level modeling is independent of low-level considerations, on this view the development of Bayesian models is constrained by, and at the same time itself constrains, the development of models at lower levels of analysis.

## Neuronal inference from the perspective of Jaynes's probability theory and the coherent infomax objective

William A. Phillips

In support of the 'Bayesian' perspective on cognition, I will agree that the adaptively organized complexity of life, and particularly mental life, depends on inductive inference. I will put five major caveats on this support, however. First, this perspective should be based not on Bayes theorem alone but on the logic of probability theory as developed most rigorously and extensively by the statistical physicist Edwin T. Jaynes (1998/2003). Interpreting probabilities as quantifying uncertainty he showed that optimal inference rests on a few requirements, or 'desiderata', and he developed maximum entropy methods for justifiably allocating prior
probabilities given only what is known. Second, Jaynes's desiderata can only be met in simple cases. Third, contextual modulation operates via likelihoods, not priors (Kay \& Phillips, 2010). Fourth, inferences are the common currency of feed-forward transmission, not prediction errors. Fifth, modulatory interactions within hierarchical levels are at least as crucial as those between levels.
I will also note that, in addition to the constraints imposed by Jaynes's desiderata, the use of prior event frequency is constrained by the curse-of-dimensionality. It becomes rapidly less useful as dimensionality of the event space increases because the number of possible locations within it then increases exponentially.
Finally, I will briefly outline the possibility that functional specializations combined with various cellular and localcircuit mechanisms for context-sensitive gain-control have evolved within mammalian cortex to restrict the problems to be solved by neuronal inference to what is feasible within the above constraints, as formulated using information theoretic concepts in the theory of coherent infomax (Phillips, 2012).

## References

Bowers, J.S., \& Davis, C.J. (2012). Bayesian just-so stories in psychology and neuroscience. Psychological Bulletin, 138(3), 389-414.
Colombo, M., \& Seriès, P. (2012). Bayes in the brain. On Bayesian modelling in neuroscience. The British Journal for Philosophy of Science, 63, 697-723.
Griffiths T., Chater N., Norris D., \& Pouget A. (2012). How the Bayesians got their beliefs (and what those beliefs actually are). Psychological Bulletin, 138(3), 415-422.
Gigerenzer, G. (2008). Why heuristics work. Perspectives in Psychological Science 3(1), 20-29.
Jaynes, E.T. (1998/2003). Probability theory: The logic of science. Oxford: Oxford University Press.
Kay, J., \& Phillips, W.A. (2010). Coherent infomax as a computational goal for neural systems. Bulletin of Mathematical Biology, 73, 344-372.
Kwisthout, J. (2011). Most probable explanations in Bayesian networks: Complexity and tractability. International Journal of Approximate Reasoning, 52(9), 1452-1469.
Kwisthout, J., \& van Rooij, I. (2013). Bridging the gap between theory and practice of approximate Bayesian inference. Cognitive Systems Research, 24, 2-8.
Marr, D. (1982). Vision. New York: Henry Holt \& Co.
Morrison, M. (2000). Unifying scientific theories, Cambridge: Cambridge University Press.
Phillips, W.A. (2012). Self-organized complexity and Coherent Infomax from the viewpoint of Jaynes's probability theory. Information, 3(1), 1-15.
Zednik, C. (in press). Heuristics, descriptions, and the scope of mechanistic explanation. In C. Malaterre \& P-A. Braillard (Eds.), How does biology explain? An enquiry into the diversity of explanatory patterns in the life sciences. Springer.

# Symposium: Global Change and Cognition 

Moderator: Stephan Lewandowsky (stephan.lewandowsky@bristol.ac.uk)<br>Department of Experimental Psychology, University of Bristol,<br>12a Priory Road, Bristol BS8 1TU, and<br>School of Psychology, University of Western Australia, Crawley, WA 6009

Contributors<br>Tania Lombrozo (University of California, Berkeley)<br>John Cook (University of Queensland)<br>Gordon D. A. Brown (University of Warwick)<br>Ben R. Newell (University of New South Wales)<br>Ullrich K. H. Ecker (University of Western Australia)


#### Abstract

We are living in a period of considerable global change. From climate change to peak oil we are facing multiple challenging problems that need to be managed carefully and wisely. Cognitive science has much to say about how people are likely to view those problems and how they will respond to them. This symposium will shed some light on those cognitive processes and how they can help-or indeed hinder-the problems we are facing.


Keywords: Global change; cognition of climate change; public acceptance of science; complex reasoning

## Summary of Symposium

There can be little doubt that human societies are facing numerous serious problems, ranging from food insecurity to resource depletion and, perhaps most serious of all, climate change. Although technological solutions to those problems arguably exist, to date there has been little enthusiasm among politicians and the public to tackle those problems. At least in part, this inaction has resulted from political factors. However, the inaction may also reflect factors related to the limitations of human cognition: People's reasoning is known to be subject to numerous biases and limitations, and our cognitive apparatus may be ill-matched to the magnitude of current global problems.

Nowhere is this mismatch more apparent than with respect to climate change, which challenges numerous cognitive and psychological processes. At a basic cognitive level, people have difficulty understanding that emissions are cumulative and that greenhouse gases remain in the atmosphere even if emissions are reduced (Sterman \& Sweeney, 2007). At a more abstract level, the public in some countries-in particular the U.S. but also in Australia-has become increasingly polarized in their attitudes towards science. Since the 1970's, Conservatives-unlike Liberals or Moderates-have become increasingly skeptical and distrustful of science (Gauchat, 2012), and people who embrace a laissez-faire vision of the free market are less likely to accept that anthropogenic greenhouse gas emissions are warming the planet than people
with an egalitarian-communitarian outlook (Dunlap \& McCright, 2008; Hamilton, 2011; Heath \& Gifford, 2006; Kahan, Jenkins-Smith, \& Braman, 2011; McCright \& Dunlap, 2011).

In light of the fundamental importance of science to the solution of global problems, the rejection of well-established scientific facts by large segments of the population must be of concern. How can this rejection be overcome? Even putting aside ideological barriers, how can people's reasoning about the future become better calibrated with the actual risks from global change?

This symposium surveys a broad range of research that addresses these questions and related issues. Tania Lombrozo will highlight the fragile relationship between understanding particular scientific claims and accepting them as true; John Cook will analyze the multi-faceted role of perceived scientific consensus (i.e., what the public believes scientists are thinking) and how that impacts attitudes; Gordon Brown will be presenting an agent-based simulation that is built around consensus-detection and will show how that explains attitude polarization; Ben Newell will report on how people judge temporal distances when considering future gains and losses; and Ullrich Ecker will explain how best to deal with the dissemination of misinformation that characterizes much contemporary public debate.

## Contributions

## Understanding science vs. accepting it

Tania Lombrozo (University of California, Berkeley)
Addressing many contemporary challenges-such as climate change and increasing resistance to antibiotics-will require more than scientific and technological advances; it will also require changes in people's attitudes and behaviors. To what extent are people's attitudes towards science and particular scientific claims shaped by their understanding of the science? There is a relatively fragile relationship between understanding particular scientific claims and accepting them as true. Nonetheless, there does seem to be a relationship between people's understanding of the nature of science in general, on the one hand, and their acceptance of specific scientific claims, on the other. Tania Lombrozo will present data for the case of evolution and consider implications for education, science communication, and policy.

[^6]It is well established that political ideology has a strong influence on public opinion about climate change, and on how people update their beliefs in the light of new climate information. Specifically, people who endorse an extreme view of free-market economics tend to reject findings from climate science. Providing people with information about the scientific consensus has been shown to partially neutralise this ideological bias (Lewandowsky, Gignac, \& Vaughan, 2012). Paradoxically, this is despite the fact that those most sceptical about climate change are also most distrustful of the scientific community. Data from several experiments are presented that explore the psychological mechanisms underlying the effectiveness of consensus information. The results are modeled within a Bayesian belief network.

## Social norms and polarization of attitudes

## Gordon D. A. Brown (University of Warwick)

Gordon Brown will describe an agent-based model of social norm effects and polarisation. The model will be applied to understanding attitudes towards climate change. The model assumes that agents located within a social network observe the behavior of neighbours and infer from that behavior the social distribution of particular attitudes (e.g. towards climate change). Agents are assumed to dislike behaviours that are extreme within their neighbourhood (social extremeness aversion), and hence have a tendency to conform. However, agents are also assumed to prefer choices that are consistent with their own true beliefs (authenticity preference). Behavioural choice-and expression of attitudes towards climate change-reflects a compromise between these opposing principles. The model sheds light on the role of perceived rather than actual scientific consensus, and "balanced" media coverage, on attitudes to climate change.

## How to weigh your options with the passage of time: Subjective and objective time preferences

## Ben R. Newell (University of New South Wales)

Many global challenges are difficult precisely because they involve trade-offs between immediate certain costs-e.g., increase in electricity prices to reduce carbon emissions-and uncertain future benefits-e.g., avoiding the worst and costliest effects of climate change. It has long been known that people discount the future very steeply; that is, they consider present monetary amounts to be more salient and valuable than when they are delayed into the future, even if those future amounts are objectively far greater. The functional form of people's discounting, however, is not well understood. This talk presents work on inter-temporal choice that sheds light on how people deal with trade-offs that involve a future cost. A particular focus is the difference between subjective and objective time estimates (cf. Malkoc \& Zauberman, 2006; Zauberman, Kim, Makoc, \& Bettman, 2009) and their implications for hyperbolic and exponential discount functions.

## Misinformation, disinformation, and the need for debiasing

Ullrich K. H. Ecker (University of Western Australia)

The dissemination of misleading information presents an obstacle for the success of science communication, public education, and evidence-based policy. Of particular concern is the resilience of misinformation: Even in the presence of clear corrections, misinformation often continues to influence people's memory and reasoning. Misinformation is particularly difficult to correct when it supports existing attitudes and when corrections counter those attitudes. Refutations of incorrect beliefs hence need to be well-designed to be efficient. Ullrich Ecker will discuss the effects of attitudes on the processing of misinformation and retractions, and highlight the important factors in the design of refutations.

## References

Dunlap, R. E., \& McCright, A. M. (2008). A widening gap: Republican and Democratic views on climate change. Environment: Science and Policy for Sustainable Development, 50, 26-35.
Gauchat, G. (2012). Politicization of science in the public sphere: A study of public trust in the United States, 1974 to 2010. American Sociological Review, 77, 167-187.
Hamilton, L. C. (2011). Education, politics and opinions about climate change evidence for interaction effects. Climatic Change, 104, 231-242. doi: 10.1007/s10584-010-9957-8
Heath, Y., \& Gifford, R. (2006). Free-market ideology and environmental degradation: The case of belief in global climate change. Environment and Behavior, 38, 48-71.
Kahan, D. M., Jenkins-Smith, H., \& Braman, D. (2011). Cultural cognition of scientific consensus. Journal of Risk Research, 14, 147-174.
Lewandowsky, S., Gignac, G. E., \& Vaughan, S. (2012). The pivotal role of perceived scientific consensus in acceptance of science. Nature Climate Change.
Malkoc, S. A., \& Zauberman, G. (2006). Deferring versus expediting consumption: The effect of outcome concreteness on sensitivity to time horizon. Journal of Marketing Research, 43, 618-627.
McCright, A. M., \& Dunlap, R. E. (2011). Cool dudes: The denial of climate change among conservative white males in the United States. Global Environmental Change, 21, 1163-1172.
Sterman, J. D., \& Sweeney, L. B. (2007). Understanding public complacency about climate change: adults' mental models of climate change violate conservation of matter. Climatic Change, 80, 213-238.
Zauberman, G., Kim, B. K., Makoc, S. A., \& Bettman, J. R. (2009). Discounting time and time discounting: Subjective time perception and intertemporal preferences. Journal of Marketing Research, 46, 543-556.

# Is perception cognitively penetrable? A philosophically satisfying and empirically testable reframing 

Gary Lupyan (lupyan@ wisc.edu)<br>Department of Psychology<br>University of Wisconsin-Madison<br>Madison, WI 53706

Dustin Stokes<br>(dustin.stokes@utah.edu)<br>Department of Philosophy<br>University of Utah<br>Salt Lake City, UT 84112

Rasha Adbel Rahman<br>(rasha.abdel.rahman@hu-berlin.de)<br>Department of Psychology<br>Humboldt-Universität zu Berlin<br>Berlin, Germany

Fiona Macpherson<br>(fiona.macpherson@glasgow.ac.uk)<br>Department of Philosophy<br>University of Glasgow<br>Glasgow, G12 8QQ

Robert Goldstone<br>(rgoldsto@indiana.edu)<br>Department of Psychology<br>Indiana University<br>Bloomington, IN 47405

Keywords: perception; top-down effects; perceptual learning; philosophy of cognitive science

## Introduction

The question of whether perception can be penetrated by cognition is in the limelight again. The reason this question keeps coming up is that there is so much at stake: Is it possible to have theory-neutral observation? Is it possible to study perception without recourse to expectations, context, and beliefs? What are the boundaries between perception, memory, and inference (and do they even exist)? Are findings from neuroscience that paint a picture of perception as an inherently bidirectional and interactive process relevant for understanding the relationship between cognition and perception?

We have assembled a group of philosophers and psychologists who have been considering the thesis of cognitive (im)penetrability in light of these questions (Abdel Rahman \& Sommer, 2008; Goldstone, Landy, \& Brunel, 2011; Lupyan, Thompson-Schill, \& Swingley, 2010; Macpherson, 2012; Stokes, 2011). Rather than rehashing previous arguments which appear, in retrospect, to have been somewhat ill-posed (Pylyshyn, 1999), this symposium will present a thesis of cognitive (im)penetrability that is at once philosophically satisfying, empirically testable, and relevant to the questions that cognitive scientists find most interesting.

## Dustin Stokes <br> Towards a consequentialist understanding of cognitive penetration

Philosophers of mind and cognitive scientists have recently taken renewed interest in in the cognitive penetration of perceptual experience. The question is whether cognitive states like belief influence perceptual experience in some important way. Since the possible phenomenon is an empirical one, the strategy for analysis has, predictably, proceed-
ed as follows: define the phenomenon and then, definition in hand, interpret various psychological data. However, different theorists offer different and apparently inconsistent definitions. And so in addition to the usual problems (e.g., definitions being challenged by counterexample), an important result is that different theorists apply their definitions and accordingly get conflicting answers to the question "Is this a genuine case of cognitive penetration?" This hurdle to philosophical and scientific progress can be remedied, I argue, by returning attention to the alleged consequences of the possible phenomenon. There are three: theory-ladenness of perception in contexts of scientific theory choice, a threat to the general epistemic role of perception, and implications for mental architecture. Any attempt to characterize or define, and then empirically test for, cognitive penetration should be constrained by these consequences. This is a method for interpreting and acquiring experimental data in a way that is agreeable to both sides of the cognitive penetration debate. Put crudely, the question shifts to "Is this a cognitive-perceptual relation that results in (or constitutes) one or more of the relevant consequences?" In answering this question it may turn out that there is no single unified phenomenon of cognitive penetration. But this should not matter, since it is the consequences that are of central importance to philosophers and cognitive scientists alike.

## Fiona Macpherson <br> Adjudicating between cognitive penetration and perceptual learning

Do we have good evidence that cognitive penetration occurs? There is a history of disagreement between those who think that perceptual experiences can be cognitively penetrated and those who think that they cannot. The argument has often proceeded on a case-by-case basis. Those who think that experiences can be penetrated present alleged examples. Most of these examples are cases in which it is
claimed that there are two different experiences, and that the best explanation for this difference is that one of the experiences was penetrated and the other was not or that the two experiences were penetrated by different cognitive states. Those who think that cognitive penetration does not occur try to offer alternative explanations. One kind of alternative explanation is that the experiences differ due to a non-cognitive difference-such as a difference in perceptual attention or in eye movement. In this paper, I discuss one variant of this strategy that tries to explain away a case of different colour experiences by claiming that the difference is due to perceptual learning, untainted by cognition. I discuss what evidence we would need to have to show that this case was one of penetration, rather than perceptual learning. I claim that we have actual evidence which is tantalizingly close to being the sort of evidence we require to show cognitive penetration, rather than perceptual learning, is occurring, and that a modicum of further easy-to-gather evidence would probably settle this case in favour of the existence of cognitive penetration.

## Rasha Abdel Rahman

## The influence of semantic knowledge on visual perception

The perception of complex visual stimuli such as objects and faces is determined not only by physical properties but may be affected by various sources of top-down influences such stored verbal categories or attention. I will present a series of experiments examining the influence of different types of semantic knowledge on perception, using the finegrained temporal resolution of event-related brain potentials (ERPs) to localize semantic effects on high and low-level components of visual analysis. The different types of knowledge include functional information that directly relates to the visual appearance of objects and their specific properties, thus explaining object shapes and features. Alternatively, the information may be unrelated to visual properties. This typically holds for biographical information about persons that can neither be derived from vision nor account for the visual appearance of a person's face. Likewise, affective biographical knowledge cannot directly be related to features or emotional expressions of faces. Furthermore, semantic information is also a major determinant of the meaningfulness of verbal categories, and semantic contents or the depth of information associated with verbal labels may play a critical role in explaining categorical perception. The results show that different types of semantic information that may or may not be directly related to visual stimulus properties shape the perception of objects and faces, including emotional facial expressions. These effects suggest an influence of semantic knowledge on sensory processing in the visual cortex that may be mediated by knowledge-induced attentional modulations and may reflect embodied cognition or reentrant activation form higherlevel semantic to sensory cortical areas.

## Robert Goldstone Hacking Our Own Perceptual Systems so that Cognition Improves

Training allows our perceptual processes to deliver outputs that would have otherwise required abstract or formal reasoning. Even without people having any privileged access to the internal operations of perceptual modules, these modules can be reliably altered over time so as to better subserve our high-level cognition needs. Strategic changes need not be implicated when perceptual systems adapt. However, there is also a continuum of intentional specificity, and with varying degrees of precision we are also able to intentionally alter our perceptual systems for our own purposes. We "hack" our perceptual systems by A) physically changing our perceptual equipment, B) strategically employing our existing perceptual equipment in new ways, $C$ ) making explicit efforts to accelerate our own perceptual adaptation process, D) creating new perceptual objects to emphasize task-relevant properties, and E) creating new physical tools to help us perceive better. Certainly not all of these adaptations should count as cases of cognitive penetration on perception, but there are striking parallels between these five classes. Strategic mechanisms of adaptation demonstrably present in some of the classes can be inferred to be at work in other classes. The semi-strategic nature of perceptual adaptation is well illustrated by the hybrid process of educating experts. The training of experts in medicine, sports, engineering, design, and food science has converged on a combination of frequent perceptual exposure to cases, explanation of causal mechanisms, and verbal descriptions that lead to selective attention to previously extracted features as well as organization into new perceptual features.

## Discussant: Gary Lupyan

## Toward a cognitive penetrability that we all care about: a consequentialist and empirically-testable one.

Abdel Rahman, R., \& Sommer, W. (2008). Seeing what we know and understand: How knowledge shapes perception. Psychonomic Bulletin \& Review, 15(6), 1055-1063.
Goldstone, R. L., Landy, D., \& Brunel, L. C. (2011). Improving Perception to Make Distant Connections Closer. Frontiers in Psychology, 2.
Lupyan, G., Thompson-Schill, S. L., \& Swingley, D. (2010). Conceptual penetration of visual processing. Psychological Science, 21(5), 682-691.
Macpherson, F. (2012). Cognitive Penetration of Colour Experience: Rethinking the Issue in Light of an Indirect Mechanism. Philosophy and Phenomenological Research, 84(1), 24-62.
Pylyshyn, Z. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. Behavioral and Brain Sciences, 22(3), 341-365.
Stokes, D. (2011). Perceiving and desiring: a new look at the cognitive penetrability of experience. Philosophical Studies, $1-16$.

# Informavores: Active information foraging and human cognition 

Doug Markant (Moderator) and Todd Gureckis<br>Dept. of Psychology, New York University<br>Björn Meder and Jonathan D. Nelson Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development<br>Peter Pirolli<br>Palo Alto Research Center<br>Chen Yu<br>Dept. of Psychological and Brain Sciences, Indiana University

Keywords: active learning, self-directed learning, information search, sensemaking

Just as the body survives by ingesting negative entropy, so the mind survives by ingesting information. In a very general sense, all higher organisms are informavores. Miller (1983)

Unlike a passive sponge floating in a sea of information, humans are active information foragers - informavores - who gather and consume new knowledge. From controlling the movement of our eyes to determining which sources of news to consult, judging the quality of alternative sources of information is a critical part of our behavior. The goal of this symposium is to bring together researchers who are working to understand the cognitive processes underlying active information foraging and how they interact with more general aspects of cognition.

The study of active information search is in the midst of a renaissance. Psychological research from diverse areas ranging from developmental psychology (Schulz \& Bonawitz, 2007), to higher level cognition (Nelson, 2005) to visual perception (Najemnik \& Geisler, 2005) have begun to understand information gathering strategies in terms of a common set of computational principles. Simultaneous developments in machine learning on "active vision" and "active learning" (Settles, 2009) have resulted in new algorithms that optimize their own learning by focusing on useful training data. Similarly, models from optimal foraging theory from biology are being brought to bear on cognitive search processes both within and outside the mind (Pirolli, 2007; Todd, Hills, \& Robbins, 2012).

This symposium aims to bring together leading experts in this area to discuss how active information foraging can be understood from a diverse set of perspectives within cognitive science. Key themes include how prior knowledge influences search (Markant \& Gureckis), how information and reward interact to determine choice (Meder \& Nelson), developmental patterns in information seeking behavior (Nelson et al.), information foraging in complex sensemaking tasks (Pirolli), and the allocation of attention during statistical word learning (Yu). While each represents a distinct area of research, all discussants in the symposium share a core approach of applying computational models to understand information search
in humans. The symposium should appeal to a broad set of attendees including educators, developmental psychologists, cognitive modelers, and computer scientists.
The influence of priors on sequential search decisions Doug Markant and Todd Gureckis

Normative models of information acquisition predict that people's search decisions should be strongly influenced by their prior beliefs, which capture the set of alternative hypotheses they are considering. In the present experiments we tested whether people adjusted their information search behavior in response to sequential changes in the prior. Participants played a search game in which they had to identify the shape and location of multiple hidden targets in a display (similar to the board game Battleship). During the task they were told that the set of possible shapes had changed, and the key question was whether they would adjust their search decisions according to the predictions of a normative model. Manipulations of the prior included changes in the frequency of certain classes of targets as well as the introduction of higherorder constraints (e.g., that all targets would have the same shape). The results showed that an individual's prior could be recovered from their sequences of search decisions, but that there were notable differences in their ability to adjust to certain changes in the hypothesis space, an effect that is not predicted by the normative model. We discuss the implications of these findings for how people generate and represent hypotheses during the course of information foraging.
Is people's information search behavior sensitive to different reward structures? - Björn Meder and Jonathan Nelson

In situations where humans actively acquire information for classification, information search preferentially maximizes accuracy (Nelson et al., 2010). However, the goal of obtaining information to improve classification accuracy can strongly conflict with the goal of obtaining information for improving utility when there are asymmetries in costs and benefits for classification decisions (e.g., in many medical diagnosis situations). Is people's information search behavior sensitive to such asymmetries? We addressed this experimentally via multiple-cue probabilistic category-learning and information-search experiments, where the payoffs corresponded either to accuracy, with equal rewards associated with the two categories, or to an asymmetric payoff function with different rewards associated with each cate-
gory. We found that people have difficulties identifying the reward-maximizing (rather than accuracy-maximizing) feature in search, following a neutral category learning task. Conversely, when trained to classify under asymmetric payoffs, they had difficulties conducting accuracy-maximizing queries when searching under symmetric rewards, where the accuracy-maximizing feature maximizes reward.

Finally, if words and numbers are used to convey environmental probabilities, neither reward nor accuracy consistently predicts search. These findings emphasize the necessity of taking into account peoples goals and search-and-decision processes during learning, thereby challenging current models of information search.
Sequential information search: Theoretical, developmental and psychological issues - Jonathan Nelson, Björn Meder, Bojana Divjak, Gudny Gudmundsdottir, Matt Jones, and Laura Martignon

We theoretically and empirically examine sequential search games in which the task is to identify an unknown target object by asking yes-no questions about its features. Globally optimal decision trees were identified using exhaustive search, in two task environments. This provided a benchmark for evaluating the efficiency of heuristic and stepwise optimal experimental design (OED) approaches for selecting questions. Some, but not all, OED approaches are useful for selecting queries. A heuristic strategy, the split-half heuristic, is mathematically equivalent to information gain, a stepwiseoptimal OED method. We investigated 4th-grade childrens search strategies on this task. Results show that children have good intuitions regarding questions' usefulness and search adaptively, relative to the statistical structure of the task environment. Search was especially efficient in a task environment that was representative of real-world experiences. This suggests that children can use their knowledge of real-world environmental statistics to guide their search behavior.

One issue for future work is to characterize the circumstances under which people identify efficient search strategies, especially in environments in which no stepwise strategy is optimal. A related issue is whether directed play can foment generalizable insights or intuitions.

## Some models of human information foraging and sensemaking - Peter Pirolli

Information Foraging Theory aims to explain and predict how people shape their information seeking behaviors to their information environments (e.g., the Web, Twitter, social tagging systems, etc.). Typically, the key steps in developing a model of information foraging involve: (a) a rational analysis of the task and information environment (often drawing on optimal foraging theory from biology) and (b) an ACTR computational cognitive model. I will present work on individual information seeking (e.g., on the Web), and then discuss how this work has been expanded to an ACT-R simulation of a complex sensemaking task involving geospatial intelligence analysis. This map-based task requires seeking (choosing) various types of available intelligence informa-
tion, and using that information to revise probability estimates about which insurgent groups might commit a future bombing attack. The model exhibits information-seeking patterns that are comparable to humans studied on this task and both model and people deviate from a rational model based on greedy maximization of expected information gain. The model also exhibits observed human biases in seeking and using information.
Active learning and selective attention in statistical word learning - Chen Yu

There are various kinds of statistical regularities in a realworld learning environment. Therefore, statistical learners have to be selective and actively gather just-in-time information required by internal learning processes and then update their internal learning states which will consequently influence their attention and selection in the next learning moment. The present study provides evidence for the operation of selective attention in the course of cross-situational learning with two main goals. The first was to show that selective attention is critical for the underlying mechanisms that support successful statistical learning. The second one was to test whether an associative mechanism with selective attention is sufficient to explain momentary gaze data in human learning.

Toward these goals, we collected eye movement data from participants engaged in a cross-situational statistical wordlearning task. Various gaze patterns were extracted, analyzed and compared between strong learners who acquired more word-referent pairs through training, and average and weak learners who learned fewer pairs. Fine-grained behavioural patterns from gaze data reveal how learners actively control their attention to gather statistical information after hearing a word, how they attend to individual objects which compete for attention within a learning trial, and how statistical evidence is selected and accumulated moment by moment, and integrated across words, across objects, and across wordobject mappings. Taken together, these findings provide new evidence on the real-time active learning mechanisms operating in the human cognitive system.

## References

Miller, G. (1983). Informavores. In F. Machlup \& U. Mansfield (Eds.), The study of information: Interdisciplinary messages (p. 111-113). Wiley-Interscience.

Najemnik, J., \& Geisler, W. (2005). Optimal eye movement strategies in visual search. Nature, 434(7031), 387-391.
Nelson, J. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. Psychological Review, 114(3), 677.
Nelson, J., McKenzie, C., Cottrell, G., \& Sejnowski, T. (2010). Experience Matters: Information acquisition optimizes probability gain. Psychological Science, 21(7), 960.
Pirolli, P. (2007). Information Foraging Theory: Adaptive Interaction with Information.
Schulz, L., \& Bonawitz, E. (2007). Serious fun: Preschoolers engage in more exploratory play when evidence is confounded. Developmental psychology, 43(4), 1045.
Settles, B. (2009). Active Learning Literature Survey (Tech. Rep. No. 1648). University of Wisconsin-Madison.
Todd, P., Hills, T., \& Robbins, T. (2012). Cognitive search: Evolution, algorithms, and the brain (Vol. 9). Cambridge, MA: MIT Press.

# Social Cognition, Culture and the Self 

Albert Newen (albert.newen@rub.de)<br>Ruhr-Universität Bochum, Institut für Philosophie II<br>Gary Bente (bente@uni-koeln.de)<br>Universität Köln, Dept. of Psychology<br>Shinobu Kitayama<br>University of Michigan, Dept. of Psychology<br>Shihui Han<br>Peking University, Dept. of Psychology<br>Kai Vogeley<br>Clinic of Psychiatry, Universität Köln


#### Abstract

The processing of social information belongs to the most complex cognitive capacities of humans, enabling us to live together in social communities. The symposium will focus on the everyday competence to form social impressions and understand others. This capability includes 1 . the understanding of oneself on the basis of an explicit selfconstrual, 2 . the understanding of others by processing their mental and bodily characteristics and states 3 . the understanding of social encounters by adequately interpreting actions, communicative signals and social roles. Human communication is essentially embedded in cultural contexts and is shaped by it; at the same time it constitutes the cultural background shared by the interactants. The main goal of this symposium is to investigate the role of cognitive and cultural factors influencing self-construal, person perception and understanding of others. Thus we deal with the following leading questions: How do we understand other human beings, what are the best theoretical perspectives, what can we learn from cognitive psychology and neurosciences and what is the role of culture in the process of understanding oneself and others? In the recent development of social cognition it has become clear that we not only have to account for the observational stance towards other people but that we also have to systematically consider situations of online interaction with other human beings ( $2^{\text {nd }}$ person perspective). The main aim of the symposium is to present the state of the art of some key topics of social and cultural cognition from the perspectives of philosophy of mind, cross-cultural psychology and social-cognitive neuroscience as well as to outline some paradigmatic lines for future research.


## Albert Newen: The Place of Culture and Self in Theories of Social Understanding

How do we make sense of the behaviour of other people? Theory-Theory and Simulation Theory both only account for an observational understanding of others. This motivated the development of the interaction theory (S. Gallagher/D. Hutto). I will shortly outline a main deficit of the latter
before I develop an alternative approach: the Person Model Theory. Person models are the basis for our ability to register and evaluate behaviour. I argue that there are two kinds of person models we rely on, nonconceptual person schemata and conceptual person images (and both kinds are developed for groups and for individuals). This theoretical approach accounts for two levels of understanding: intuitive and inferential understanding. Furthermore, it has the advantage to account for the difference of understanding very familiar persons (relying on person models of individuals), on the one hand, and complete strangers (reyling on person model for groups and thus understanding a person by his social role as a student), on the other. The person model theory accounts for modelling myself by presupposing self-models in addition to models of others. The theory allows to spell out the interaction of relying on modelling oneself and others and finally it explicitly accounts for the role of culture as shaping the mind. (s. Newen/Schlicht 2009; de Bruin/Newen 2011).

## Gary Bente \& T. Dratsch: The Role of Tacit Cues and Social Order in Observing Others. Analyzing Nonverbal Behaviour and Impression Formation across Cultures

There is ample evidence that visual cues including physical appearance and nonverbal behaviour play a crucial role in person perception and impression formation. A short look at a face can lead to attributions of trustworthiness, a body posture or movement can be perceived as dominant and a smile can lead to warm feelings of connectedness. Humans are highly sensitive to nonverbal signals and our responses are fast and mostly beyond awareness and conscious control. There is also evidence supporting the notion that culture plays a prominent role in molding our nonverbal behaviors. In a series of studies we investigated the role of culture in the processing of nonverbal cues in conflict laden interactions, collaborative tasks as well as economic games focusing on the dimensions of liking, power/control and trust. The studies involving German, American, Chinese and Arab participants clearly point to cultural specificities indicating that distinct cultural values
as described for Westerners and Easterners are also implemented on the micro-level of social interactions.(see Bente et al. 2008, 2010; Goergescu et al. in press)

## Shinobu Kitayama: Error-Related Brain Activity Reveals Self-Centric Motivation

To secure the interest of the personal self (versus social others) is considered a fundamental human motive, but the nature of the motivation to secure self-interest is not well understood. To address this issue, we assessed electrocortical responses of European Americans and Asians as they performed a flanker task while instructed to earn as many reward points as possible either for the self or for their same-sex friend. For European Americans, error-related negativity (ERN) -an event-related-potential component contingent on error responses-was significantly greater in the self condition than in the friend condition. Moreover, post-error slowing-an index of cognitive control to reduce errors-was observed in the self condition but not in the friend condition. Neither of these self-centric effects was observed among Asians, consistent with prior cross-cultural behavioral evidence. Interdependent self-construal mediated the effect of culture on the ERN self-centric effect. Our findings provide the first evidence for a neural correlate of self-centric motivation, which becomes more salient outside of interdependent social relations (see Markus \& Kitayama 1991; Kitayama et al. 2011).

## Shihui Han: What constitutes the self? Cultural neuroscience studies of neurocognitive representation of the self

The self is a mixture of both biological and social construction. How is the self represented in the human brain? I'll present psychological and brain imaging studies of self-face recognition and self-concept published by our group during the last years. These studies investigated how neurocognitive processing of the self undergo cultural and biological influences. Our findings have implications for our understanding of the biosocial nature of the human brain and mental health. (see Han et al. 2008; Ma et al. in press).

## Kai Vogeley: Person perception and culture. The perspective of neuroscience

Psychology and neuroscience have recently started to reintroduce culture as an independent factor into the experimental designs of empirical studies focusing on cognitive processes and neural mechanisms. On a conceptual level culture cannot be treated as a rigid body of generalized features of different cultural backgrounds as defined by nationality or language, but has to be conceptualized more adequately as a dialectic exchange between individual members and their collectives with
respect to habits, practices and belief systems. However, experimental studies require operationalized approaches. We have recently used Asian-looking and European-looking virtual characters that expressed anger and happiness while gazing at the participant or another third invisible person in a Chinese and a German population. Overall, expressions are perceived more pronounced if the participant was looked at. Direct gaze emphasized the perceived emotion of ethnic out-group members, but not of ethnic in-group members. These results suggest that social interaction supervenes or is at least as influential as culture related differences in the perception of emotions. (see Vogeley et al. 2009; Han et al. 2013).

## References

Bente, G., Senokozlieva, M., Pennig, S., Al-Issa, A., \& Fischer, O. (2008). Deciphering the secret code. A New Methodology for the Cross-Cultural Analysis of Nonverbal Behavior. Behavior Research Methods, 40(1), 269-277.
Bente, G., Leuschner, H., Al-Issa, A., \& Blascovich, J. J. (2010). The others: Universals and cultural specificities in the perception of status and dominance from nonverbal behavior. Consciousness and Cognition, 19(3), 762-777.
De Bruin, L. C., \& Newen, A. (2012): An association account of false belief understanding. Cognition doi:10.1016/j.cognition.2011.12.016
Georgescu, A.L., Kuzmanovic, B., Santos, N.S., Tepest, R., Bente, G., Tittgemeyer, M. \& Vogeley, K. (in press). Perceiving Nonverbal Behavior: Neural Correlates of Processing Movement Fluency and Contingency in Dyadic Interactions. Human Brain Mapping.
Han, S., Northoff, G. (2008). Culture-sensitive neural substrates of human cognition: A transcultural neuroimaging approach. Nature Review Neuroscience, 9, 646-654.
Han S, Northoff G, Vogeley K, Wexler BE, Kitayama S, Varnum MEW (2013): Cultural Neuroscience Approach to the Biosocial Nature of Humans. Ann Rev Psychol 64, 335-359.
Kitayama, S., \& Uskul, A. K. (2011). Culture, mind, and the brain: Current evidence and future directions. Annual Review of Psychology, 62, 419-449.
Ma, Y., Bang, D., Wang, C., Allen, M., Frith, C., Roepstorff, A., Han, S.(in press).Sociocultural patterning of neural activity during self-reflection. Social Cognitive and Affective Neuroscience.
Markus, H., \& Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. Psychological Review, 98, 224-253.
Newen, A., Schlicht, T. (2009): Understanding Other Minds: A criticism of Goldman's simulation theory and an outline of the Person Model Theory. Grazer Philosophische Studien 79, 209242
Vogeley K, Roepstorff A (2009): Contextualising Culture and Social Cognition. Trends in Cognitive Science 13, 511-516

# Scales of Cognition Evident in Action 

Denis O'Hora (denis.ohora@nuigalway.ie)<br>School of Psychology, NUI Galway<br>University Road, Galway, Ireland

Nicholas D. Duran (nduran2@ucmerced.edu) Rick Dale (rdale@ucmerced.edu)<br>Cognitive \& Information Sciences<br>University of California, Merced<br>5200 North Lake Rd., Merced, CA 95343

Jonathan B. Freeman (jon.freeman@dartmouth.edu)<br>Psychological and Brain Sciences, Dartmouth College<br>6207 Moore Hall, Hanover, NH 03755

John M. Tomlinson, Jr. (tomlinson@zas.gwz-berlin.de)<br>Zentrum für Allgemeine Sprachwissenschaft (ZAS), Berlin 10117 Germany


#### Abstract

When cognitive processes occur alongside observable actions, it is possible for characteristics of these processes to influence the ongoing performance of those actions. This satisfies everyday intuitions. For example, negotiators and poker players claim to be attuned to 'tells,' these early behavioral indicators of eventual decisions. Going beyond intuitions, however, several researchers have exploited this fine-grained source of behavior to highlight online cognitive processing. Using even a simple measure such as computer-mouse tracking can reveal a wide range of cognitive processing. Four participants in this symposium report on applications of the analysis of the action dynamics of cognition across multiple scales: (i) basic decisions, (ii) language processing, (iii) false responding, and (iv) social processes. The similarities and differences in expression of these processes in action highlight important continuities and discontinuities across cognitive and neural processes.


Keywords: action dynamics; social cognition; learning; decision making; language

## Cognition and Action

A perhaps still prevailing notion of the relationship between cognition and action is that motor movement takes place mostly near the end of a cognitive process or decision. Implicit in this notion is that the systems are relatively independent, potentially modular and encapsulated. An outcome of this attitude is that action has, to some extent, been neglected in many quarters of the cognitive sciences (for review on this issue, see Rosenbaum, 2005). In contrast, recent work suggests that action and cognition facilitate one another to such a degree that one can understand action as "part and parcel" of cognition (Freeman, Dale, \& Farmer, 2011). In this work, researchers extract the computer-mouse cursor movements of participants who carry out a cognitive task. By analyzing dynamic properties of the cursor, such as motion latencies, velocities, complexity of movement, and so on, new insights into cognitive processing are possible.

A tighter relationship between cognition and action has inspired explorations of cognition using densely-sampled be-
havioural data. Put simply, it may now be possible to investigate ongoing action for evidence of purported cognitive processes. The dynamic characteristics of ongoing behaviour provide a testbed for the comparison of models in various areas of cognitive science. The papers in this symposium will report on applications of the analysis of the action dynamics of cognition at four very different scales. The first talk will address the dynamics of basic choice processes by studying the dynamics of movement while those choices are made. The second talk addresses sentence process and pragmatic inferences. The third showcases the dynamics of false responding as a model of cognitive processes involved in deception. Finally, the fourth talk discusses the application of these dynamic techniques to social cognition. The symposium will end with 20-minute interactive discussion of the relationship between action and cognition, the impact of these measures on theoretical issues across scales, and the inevitable differences in how measures behave at these levels and their different tasks.

## (i) Mechanics of Choice and Decision-Making

Our everyday language about choice and decision making is, in English at least, replete with dynamic physical metaphors. We are 'pulled' or 'drawn' towards choices that we sometimes 'cannot resist'. When we make a decision, we might be asked, 'How did you come to this conclusion?' and we might feel were 'pushed' towards it or that we 'fell into it'. These metaphors highlight certain characteristics of decisionmaking. First, decision making takes time; our preferences gradually develop from less stable to more stable as we choose. Second, how we make our decisions tells us much about the quality of the decision, whether it was easy or difficult, fearful or hopeful, and so on. Indeed, the dominant models of preferential decision-making within cognitive science include the dynamic evolution of choice in that repeatedly sampling information biases the unfolding decision.

Within a short discrimination learning task, Denis O’Hora and colleagues manipulated the strength of attraction to available choices by changing the points available for making these choices. He will describe how they used this method to investigate choice trajectories under a range of conditions of choice conflict.

## (ii) Interplay between Pragmatic and non-Pragmatic Inference

Unlike most linguistic phenomena, the derivation of pragmatic inferences is optional. For example if a speaker is asked, "Are Todd and Sam coming to the party," and she responds, "Todds coming," the listener could interpret the utterance as 1) the speaker only knows that Todd is coming but doesnt know about Sam or 2) derive the inference that only Todd is coming and Sam is definitely not. In both cases, listeners at least know that Todds coming. Because pragmatic inferences are highly context dependent, research on the comprehension of inferences needs to tease part the likelihood of making an inference from the process of deriving on one (Bott, Bailey, \& Grodner, 2012).

John Tomlinson will present work on how action dynamics cannot only capture this important distinction, but also how action dynamics provide clearer insight into how pragmatic inferences are derived. In several of studies, listeners motor movements show strong initial preference for nonpragmatically enriched interpretations for scalar implicatures before correcting towards pragmatic interpretations. Studies will be presented on how intonation, context, and speaker information can streamline these inferences. Critical for this symposium is how action dynamics, specifically time normalized mouse movements, can provide new insights into how such factors above and beyond that of real time data analyses such as reaction times and eye-movements.

## (iii) Action Dynamics Reveal False Responses

Human beings are surprisingly adept at responding to questions with information that is in opposition to what is known to be true. There remain, however, many open questions about how deception is possible. Do we hold in mind what is known to be true, and actively inhibit it to respond falsely? Does it matter whether that information is biographically relevant, or is simply a statement that can be readily falsified in our own semantic memory? Does answering falsely get easier if our cognitive system can prepare for it?

Questions such as these pertain to the underlying cognitive processes that contribute to deception. Yet it is extremely difficult to create experimental situations that permit direct access into such mechanisms. In this presentation Dale and Duran survey a variety of experiments that utilize the mousetracking methodology to explore the dynamics of false responses. For example, in one experiment, participants were prompted to respond falsely about their personal experiences and biography (Duran, Dale, \& McNamara, 2010). In another, participants were prompted to lie to an imagined part-
ner in a game of 20-questions, thus confusing them about a target objects of the game (Duran \& Dale, 2012).

## (iv) Cognitive Processes in Social Categorization

Mere exposure to a social target has long been known to trigger spontaneous categorization along multiple dimensions (e.g., sex, race, age). Such categorizations are extremely rapid and efficient, yet also reflect the complex integration of a variety of bottom-up (e.g., facial and vocal cues) and topdown (e.g., stereotypic expectations, motivation) information sources. Jonathan Freeman will discuss recent work exploiting the tight link between cognitive and action dynamics to understand the underlying social categorization process.

In one series of studies, for example, participants were presented with sex-typical and sex-atypical faces and asked to categorize the targets sex by clicking on a male or female response on the screen (Freeman, Ambady, Rule, \& Johnson, 2008). During categorization of sex-atypical faces, hand trajectories were continuously attracted to the opposite sexcategory before settling into the correct response. These findings and others support an account of social categorization in which dynamic competition is central; perceived facial, vocal, and bodily cues (among other constraints) simultaneously weigh in on multiple partially-active category representations that dynamically evolve over time into stable categorical perceptions (Freeman \& Ambady, 2011).

## References

Bott, L., Bailey, T. M., \& Grodner, D. (2012). Distinguishing speed from accuracy in scalar implicatures. Journal of Memory and Language, 66(1), 123-142.
Duran, N. D., \& Dale, R. (2012). Increased vigilance in monitoring others' mental states during deception. In Proceedings of the 34th annual meeting of the cognitive science society. Austin, TX: Cognitive Science Society.
Duran, N. D., Dale, R., \& McNamara, D. S. (2010). The action dynamics of overcoming the truth. Psychonomic Bulletin \& Review, 17(4), 486-91.
Freeman, J. B., \& Ambady, N. (2011). A dynamic interactive theory of person construal. Psychological Review, 118(2), 247-79.
Freeman, J. B., Ambady, N., Rule, N. O., \& Johnson, K. L. (2008). Will a category cue attract you? Motor output reveals dynamic competition across person construal. Journal of Experimental Psychology: General, 137(4), 67390.

Freeman, J. B., Dale, R., \& Farmer, T. a. (2011). Hand in motion reveals mind in motion. Frontiers in Psychology, 2, 59.

Rosenbaum, D. A. (2005). The Cinderella of psychology: the neglect of motor control in the science of mental life and behavior. The American Psychologist, 60(4), 308-17.

# Music cognition: Bridging computation and insights from cognitive neuroscience 

Marcus Pearce (marcus.pearce@eecs.qmul.ac.uk)<br>Centre for Digital Music and Research Centre in Psychology, Queen Mary, University of London, E1 4NS, UK.

Psyche Loui (ploui@bidmc.harvard.edu)<br>Beth Israel Deaconess Medical Center and<br>Harvard Medical School, Boston, MA, USA.

Petri Toiviainen (petri.toiviainen@jvu.fi)<br>University of Jyväskylä, Finland

Keywords: Music cognition; cognitive neuroscience; computational modelling; processing; prediction; grammar

## Goals and Scope

In recent years, computational models have become an increasingly important part both of cognitive science and cognitive neuroscience. In tandem with these developments neuroscientific and cognitive investigations of musical experience and behaviour have been gathering pace. In this context, music cognition constitutes a rich and challenging area of cognitive science in which the processing of complex, multi-dimensional temporal sequences can be studied without interference of meaning or semantics (see Pearce \& Rohrmeier, 2012, for a review). Because of its complexity and well-defined problem-space, computational modelling of music witnessed a rapid growth of successful higher-order modelling approaches. This workshop investigates computational modelling as a bridge between cognition and the brain, with a focus on understanding the psychological mechanisms involved in perceiving and producing music.

Many approaches have been taken to modelling the large variety of different cognitive processes involved in music perception and creation involving various modules of basic structural processing, statistical learning, memory, as well as motor, emotional and social cognitive processes. Recent computational models range from hierarchical, rule-based systems for representing harmonic movement inspired by probabilistic grammars for language, through oscillator based network models for modelling metrical and tonal perception, to probabilistic methods derived from machine learning for modelling dynamic learning and predictive processing of style-specific musical structure. Turning to cognitive neuroscience, recent years have seen increasing interest in advanced computational modelling of EEG and fMRI data used to distinguish brain regions responsible for the processing of different aspects of music (e.g., rhythm, pitch, timbre, harmony) and the functional connectivity between them. The purpose of this symposium is to bring together and display current research trends towards a synthesis of these two research areas linking the parameters and subcomponents of cognitive models of musical processing to functional and anatomical properties of the brain.

Martin Rohrmeier (mr1@mit.edu)<br>MIT Intelligence Initiative, Department of Linguistics and Philosophy, Massachusetts Institute of Technology, Cambridge, MA, USA

Edward Large (large@ccs.fau.edu)
Ji Chul Kim (kim@ccs.fau.edu)
Center for Complex Systems \& Brain Sciences Florida Atlantic University

Elvira Brattico (brattico@mappi.helsinki.fi)<br>Aalto University, Finland

## Petri Toiviainen and Elvira Brattico Decoding the musical brain during naturalistic listening

Encoding, or prediction of neural activation from stimulus, is a common modeling approach in neuroscience. In our recent neuroimaging study, we applied encoding to predict brain activity during listening to different pieces of music from an extensive set of musical features computationally extracted from the pieces, and found widespread brain activation, including auditory, limbic, and motor areas (Alluri et al., Neuroimage, under review). With such complex and distributed neural activation, evaluation of different encoding models is not straightforward, because the goodness of prediction is difficult to assess. Decoding, or prediction of physical or perceived stimulus features from the observed neural activation, has the potential benefit of a more straightforward model evaluation because of easier performance characterization in terms of, for instance, correct classification rate.

In a series of experiments, our participants were measured with functional magnetic resonance imaging (fMRI) while they were listening to three different musical pieces. Subsequently, musical features were computationally extracted from the pieces, and continuous emotion ratings were collected from the participants. For decoding, the fMRI data were subjected to dimensionality reduction via voxel selection and spatial subspace projection, and the obtained projections were subsequently regressed against the musical features or the emotion ratings. To avoid overfitting, cross-validation was utilized. Different voxel selection criteria and subspace projection dimensionalities were used to find optimal prediction accuracy. The decoding results and the challenges of the approach will be discussed at the symposium.

## Psyche Loui <br> Behavioral and DTI Studies on Normal and Impaired Learning of Musical Structure

One of the central questions of cognitive science concerns how humans acquire knowledge from exposure to stimuli in the environment. In the context of music, knowledge
includes the structure of harmony and melody that govern how musical pitches are combined. While people of all cultures and ages show some knowledge of the structure of their music, people with tone-deafness (also known as congenital amusia) show a lack of behavioral sensitivity to harmony and melody. Here we combine behavioral evidence from subjective ratings, neuroimaging evidence from Diffusion Tensor Imaging, and neuropsychological evidence from tone-deaf individuals, to support the thesis that much of what we know and love about music is acquired via statistical sensitivity to the frequency and probability of occurrence of events in the auditory environment. This statistical learning mechanism relies on intact white matter connectivity between temporal and frontal lobe regions, and may subserve multiple auditory-motor functions including language as well as music.

## Edward Large and Ji Chul Kim A Universal 'Grammar' for Music

Since antiquity, science has asked whether mathematical relationships among acoustic frequencies govern the perception of musical relationships. Modern psychophysicists rejected this approach, citing evidence that the auditory system performs a linear analysis of sound. Cognitive psychologists have since relied primarily on statistical learning to explain music cognition, despite continued demonstrations of the importance of frequency relationships. Today evidence is rapidly mounting that the auditory system is highly nonlinear, inviting reevaluation of the role of frequency in constraining in the perception of music. Here, we present a dynamical systems analysis of auditory nonlinearities that predicts substantive universals in music perception and cognition. This approach explains perceptual ratings of Hindustani raga not only by encultured listeners, but also by listeners who were completely unfamiliar with the music of North India. This evidence suggests that universal properties of neural oscillation explain cross-cultural invariants in the perception of tonal music, implying neurodynamic constraints on the acquisition of musical languages.

## Martin Rohrmeier Computational Models of Musical Syntax

In order to create the variety of our rich musical experience, Western music employs a complex combination of interwoven features of pitch, rhythm, metrical structure, harmony and melody. Various computational models of music processing are based on local (event-to-event) processing of musical features (cf. Rohrmeier \& Koelsch, 2012). On the other hand, a number of theoretical approaches suggest that music involves recursive, hierarchical structures organized in ways similar to linguistic syntax. Further recent neurocognitive research provides evidence indicating that musicians as well as nonmusicians are sensitive to subtle long-distance violations resulting from the underlying syntactic structure. These insights suggest that musical processing is more complex than previously assumed and involves rich mechanisms of
structural parsing and implicit learning to deal with these syntactic features. This talk presents the picture emerging from converging evidence of theoretical approaches, recent experimental work and computational modeling. Probabilistic models of musical syntactic processing based on Hidden Markov Models and probabilistic context-free grammars underpin that the inference of complex nonlocal dependencies from mere exposure is plausible. They further predict experimental data of musical tension and expectancy showing that a variety of features of musical experience can be modeled by such approaches.

## Marcus Pearce <br> Expectation and Emotion in Music Perception: Computational Modeling of Dynamic Cognitive and Neural Processes

The idea that aesthetic experience of music is dependent on the confirmation and violation of expectation dates back at least to Hanslick. Meyer (1957) further proposed that such expectations depend on probabilistic models of musical structure, acquired through exposure. However, until recently such theories remained largely untested. Here we present evidence corroborating these proposals and filling in some of the details in terms of cognitive and neural processing. First, we show that musical expectations elicited in a range of musical styles result reflect probabilities acquired through a process of statistical learning. Subjective expectedness and uncertainty can be modeled dynamically through time using the information-theoretic concepts of information content and Shannon entropy respectively. Second, we identify time-locked electrophysiological brain responses to events differing in information content. Third, we show that variations in information content lead to distinct psychological and physiological emotional responses elicited in a live concert of music for solo flute. The results also indicate that expectations and emotion are influenced by factors other than the musical structure such as visual aspects of the performance. In summary this research suggests that musical expectations rely on dynamic probabilistic cognitive processing of musical structure, supported by corresponding neural processes, and generates characteristic physiological and psychological emotional responses.

## Moderators: Marcus Pearce and Martin Rohrmeier

## References

Pearce, M. T. \& Rohrmeier, M. (2012). Music cognition and the cognitive sciences. TopiCS in Cognitive Science, 4, 468-484.
Meyer, L. B. (1957). Meaning in music and information theory. Journal of Aesthetics and Art Criticism, 412-424.
Rohrmeier, M., Koelsch, S. (2012). Predictive information processing in music cognition. A critical review. International Journal of Psychophysiology, 83 (2), 164-175.

# Thinking about norms: Epistemic, rational, and moral norms in human thinking 

Joëlle Proust (jproust @ ehess.fr)<br>Institut Jean-Nicod (EHESS-ENS), UMR CNRS 8129, Ecole Normale Supérieure, Paris, France

Emmanuel M. Pothos ${ }^{1}$ (e.m.pothos@ gmail.com) and Jerome R. Busemeyer ${ }^{2}$ (jbusemey@indiana.edu)<br>${ }^{1}$ Department of Psychology, Swansea University, UK<br>${ }^{2}$ Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN, USA.

Ryan Miller (ryan_m_miller@brown.edu) and Fiery Cushman (fiery_cushman@brown.edu) Dept. of Cognitive, Linguistic and Psychological Sciences, Brown University, Providence, USA

Katinka J.P. Quintelier (katinka.quintelier@gmail.com)
Amsterdam Business School, University of Amsterdam, The Netherlands

Shira Elqayam (selqayam@dmu.ac.uk) ${ }^{\mathbf{1}}$, Valerie A. Thompson (vat128@campus.usask.ca) ${ }^{\mathbf{2}}$, Jonathan St.B.T. Evans (jevans@plymouth.ac.uk) ${ }^{3}$, David E. Over (david.over@durham.ac.uk) ${ }^{4}$; and Meredith R. Wilkinson (mwilkinson@dmu.ac.uk) ${ }^{1}$<br>${ }^{1}$ School of Applied Social Sciences, De Montfort University, UK; ${ }^{2}$ Department of Psychology, University of Saskatchewan, Canada; ${ }^{3}$ School of Psychology, University of Plymouth, UK; ${ }^{4}$ Department of Psychology, Durham University, UK

Discussant: Shulamith Kreitler (krit@netvision.net.il)<br>School of Psychological Sciences, Tel-Aviv University, Israel

## Organiser: Shira Elqayam

Keywords: action; epistemic norms; is-ought problem; moral norms; normative inference; rational norms

## Thinking about norms

Humans are uniquely good at inventing norms, thinking about norms, complying with norms and defeating norms. It is small wonder, then, that norms are a focus of much interest as well as debate across the cognitive sciences, encompassing such diverse issues as rationality, morality and action. The aim of the present symposium is to bring together a range of psychological and philosophical contributions to this pertinent debate. Contributors come from diverse backgrounds, including epistemology, metaethics, moral judgment, decision making, and reasoning. We will examine foundational issues in normative thinking, such as: What is the relation between norms and descriptions? What are the psychological mechanisms underlying normative thinking? How do epistemic and moral norms guide action? What, if any, are the appropriate norms for knowledge, rationality, and moral behaviour, and how can they be determined?

## Proust: The norms of acceptance

An area in the theory of action that has received little attention is how mental agency and world-directed agency interact. The purpose of the present contribution is to clarify the rational conditions of such interaction, through an analysis of the central case of acceptance. There are several problems with the literature about acceptance. First, it remains unclear how a context of acceptance is to be construed. Second, the possibility of conjoining, in acceptance, an epistemic component, which is essentially mind-to-world, and a utility component, which requires a
world-to-mind direction of fit, is merely posited rather than derived from the rational structure of acceptance. Finally, the norm of acceptances is generally seen as related to truth, which turns out to be inapplicable in a number of cases.

We will argue, first, that the specific context-dependence of acceptances is derived from their being mental actions, each embedded in a complex hierarchy of acceptances composing, together, a planning sequence. Second, that acceptances come in several varieties, corresponding to the specific epistemic norm(s) that constitute them. The selection of a particular norm for accepting answers to considerations of utility - to the association of an epistemic goal with an encompassing world-directed action. Once a type of acceptance is selected, however, the epistemic norm constitutive for that acceptance strictly applies. Third, we argue that context-dependence superimposes a decision criterion on the output of the initial epistemic acceptance. Strategic acceptance is regulated by instrumental norms of expected utility, which may rationally lead an agent to screen off her initial epistemic acceptance.

## Pothos \& Busemeyer: Implications for the rationality debate from the quantum cognition research programme

Bayesian theory has enabled an influential perspective on human rationality, partly based on such arguments as long term convergence and the Dutch book theorem. Moreover, behavioral predictions in decision making from Bayesian theory are typically supported by strong intuition. Yet, this intuition often goes against empirical findings. For example, Kahneman, Shafir, Tversky and collaborators have provided many compelling demonstrations of violations of the law of total probability or the conjunction principle. Recently,
researchers have shown that many of these violations can be naturally accounted for within quantum probability theory, a framework for formal probabilistic modeling alternative to Bayesian theory. If one accepts that quantum theory provides a more accurate framework for modeling human behavior, at least in some cases, then what are the implications regarding (or not) the debate on human rationality? After all, probabilistic inference in quantum theory can be strongly context and perspective dependent, perhaps going against an intuition that probabilistic inference is rational to the extent that it is objective (in some sense). Equally, we note that probabilistic inference in Bayesian models presupposes adherence to the, perhaps cognitively unrealistic, principle of unicity, the requirement that there is a complete joint probability distribution for all relevant possibilities. Consideration of the above issues provides us with two broad themes. First, is the Bayesian notion of normative rationality cognitively feasible, even if appropriate from an abstract perspective? Second, can a perspective about normative rationality arise from quantum cognitive models?

## Miller \& Cushman: Action, outcome and value

How can we characterize the underlying cognitive mechanisms that give rise to moral judgment? A popular approach has been to contrast "emotion" with "cognition", but this is widely regarded as a problematic distinction--even by its chief proponents. We advocate for an alternative approach motivated by the distinction between model-based and model-free reinforcement learning. A model-based system chooses actions with the greatest expected value based on a detailed causal model of their likely outcomes. A model-free system associates positive or negative feelings with particular actions intrinsically. We will present a series of studies suggesting that this distinction between outcome-based and action-based decision-making matches the dualsystem structure of moral judgment, with many benefits over the traditional distinction of emotion vs. reason. Dual system approaches in the moral domain have been used widely, although controversially, to distinguish normatively warranted and unwarranted moral judgments. The application of reinforcement learning theories to the moral domain has the potential to inform debates over the normative status of moral judgments. It allows us to state precisely the relationship between value, experience and choice. Leveraging this formal precision, we join others in arguing that psychological facts have implications for the normative status of moral judgments.

## Quintelier: The real is-ought problem in ethics

Numerous scholars have pointed out that 'is' and 'ought' should be kept separated. While valuable, this pursuit distracts from an equally important issue: In order for
empirical findings to be relevant for ethics, we need an account of how 'is' and 'ought' can be properly linked.

I illustrate this by means of the moral universalism versus relativism debate: Scholars have advocated that we should think of moral rules as universal because, among other reasons, lay people intuitively think of morality as universal. Recent studies however show a diversity of moral reasoning, including relativist moral reasoning, in the folk. Nevertheless, it is now debated how these data are relevant for ethics because, arguably, 'is' and 'ought' should be kept separated. In the moral universalism versus relativism debate though, illegitimate inferences from 'is' to 'ought' are not the problem. While it is true that previous arguments in favor of moral universalism relied on a specific relation between 'is' and 'ought', this relation is refuted by presentday scholars. However, no alternative is put in place. Moreover, at the same time, the rationale for doing empirical research on this topic is to further a normative debate. Thus, either existing empirical research is irrelevant, or researchers have to defend a link between 'is' and 'ought'.

## Elqayam, Thompson, Evans, Over, \& Wilkinson: When do we infer ought from is?

The debate on norms in cognitive science goes back at least as far as Hume's critique of what has come to be known as the is-ought problem: when, if ever, is it valid to infer normative conclusions from descriptive premises? Whereas philosophers are interested in the validity of such inference, we ask about the psychological mechanisms underlying it.

We present a new processing model of inference from 'is' to 'ought'. The relevant logic is deontic, the logic of rules and regulations. We propose that such inference is pragmatic, in the sense that it is socially rich, contextualised, probabilistic, and defeasible. Agents infer deontic, normative conclusions from descriptive premises under a set of conditions: (1) an agent; (2) a goal, or a valued outcome (3) an action causally linked to the goal. We present a set of findings to show that the direction of the deontic conclusion that people endorse matches the psychological value of the goal; that the strength of the conclusion is a function of the strength of the causal link between action and outcome; and that the inference is suppressed when additional premises present conflicting goals, triggering a utilitarian conflict; or conflicting norms, triggering a deontological conflict. We suggest that this normative inference underlies much of human epistemological and moral judgement and action.

## Acknowledgments (Elqayam et al.)

The research reported in this talk is supported by a research project grant from the Leverhulme Trust (RPG-2012-573).

## References

Hume, D. (2000). A treatise on human nature (Original publication date 1739-1740). Oxford: Clarendon Press.

# Implicit Learning Out of the Lab: Language and Music 

Patrick Rebuschat (p.rebuschat@lancaster.ac.uk)<br>Department of Linguistics and English Language Lancaster University, Lancaster, United Kingdom

Morten H. Christiansen (christiansen@cornell.edu)<br>Department of Psychology, Cornell University Ithaca, NY, USA

Clément François (cfrancois@idibell.cat)<br>Daniele Schön (daniele.schon@incm.cnrs-mrs.fr)<br>INSERM U 1106, Institut de Neuroscience des Systèmes Aix-Marseille Université, Marseille, France

## Shan Jiang (jiangs988@126.com)

School of Social Administration, Shanghai University of Political Science and Law, Shanghai, China

Jennifer B. Misyak (jennifer.misyak@wbs.ac.uk)<br>Behavioural Science Group, Warwick Business School<br>University of Warwick, Coventry, United Kingdom

Keywords: Implicit learning; statistical learning; explicit learning; music cognition; language acquisition.

## Goals and Scope

Implicit learning, essentially the ability to acquire unconscious (implicit) knowledge, is a fundamental aspect of human cognition. This symposium focuses on the acquisition of two cognitive systems that are widely regarded as prime examples of implicit learning "in the real world", namely language and music (see e.g. Rebuschat et al., 2011; Rohrmeier \& Rebuschat, 2012). This symposium brings together leading researchers from across the cognitive sciences (psychology, linguistics, cognitive neuroscience, computer science, and musicology) in order to discuss current trends in implicit learning research, to identify the progress made in recent years, and to outline future directions to take, both in terms of topics and novel methodologies.
The symposium will consist of five talks, followed by a brief general discussion. Each talk approaches the symposium topic from a highly innovative and interdisciplinary angle. Christiansen and Misyak focus on individual differences in implicit language learning, while Rebuschat concentrates on the role of implicit and explicit learning in second language (L2) acquisition. François and Schön's work demonstrates the impressive effect of musical practice on the implicit learning of linguistic structure, while the last two talks, by Dienes and colleagues and by Rohrmeier and Widdess, focus on the implicit learning of Chinese tonal poetry and of syntactic features of North Indian music, respectively. These last two studies reflect a particularly important trend in implicit learning research

Martin Rohrmeier (mr1@mit.edu)<br>Department of Linguistics and Philosophy, Massachusetts Institute of Technology, Cambridge, MA, USA

Zoltan Dienes (dienes@sussex.ac.uk)<br>School of Psychology, University of Sussex Brighton, United Kingdom

Xiuyan Guo (xyguo@psy.ecnu.edu.cn)<br>Shanghai Key Laboratory of Magnetic Resonance and School of Psychology and Cognitive Science, East China Normal University, Shanghai, China

Feifei Li (lifeifei1206@163.com)
School of Psychology and Cognitive Science, East China Normal University, Shanghai, China

Richard Widdess (rw4@soas.ac.uk)<br>Department of Music, School of Oriental and African Studies, University of London, United Kingdom

towards the use of more ecologically-valid stimuli. In addition to introducing novel and exciting subject areas, the research discussed in this symposium also reflects the strong tendency, within cognitive science, for methodological diversification. The talks will discuss data from behavioral and neurophysiological experiments as well as results of computational modeling.

## Morten H. Christiansen and Jennifer B. Misyak Individual differences in implicit statistical learning and language

Over the past decade, implicit learning under the guise of statistical learning has emerged as an important experimental paradigm with which to study mechanisms involved in language acquisition. Although few empirical studies have directly linked variation across statistical learning and language, it is generally assumed that greater sensitivity to statistical structure leads to better language performance. Here, we report the results of studies investigating the relationship of individual differences in statistical learning of adjacent and nonadjacent dependencies to variations in the processing of local and nonlocal dependencies in natural language. Together, the results indicate that individual differences in statistical learning are positively related to variations in language processing. However, the complexity of the pattern of interrelations suggests that future developmental and adult work on implicit statistical learning must incorporate careful attention to a diversity of natural dependency-structures to establish the proper relationship between adjacent and nonadjacent manifestations of statistical learning and the
extent to which they map onto similar structures in language.

## Patrick Rebuschat Implicit and explicit learning of $\mathbf{L} 2$ syntax

First language acquisition is generally characterized as a process where most learning proceeds implicitly, i.e. incidentally and in absence of awareness of what was learned. At the same time, however, there is considerable debate as to whether implicit learning plays a similarly important role in the case of adult second language (L2) acquisition. In this talk, I will review a series of experiments that investigated the implicit and explicit learning of L2 syntax by means of an artificial language paradigm. This research addressed questions such as the following: Is there implicit learning in the case of L2 acquisition? If so, how is implicit knowledge of language represented in the mind (rules, patterns, chunks...)? How do task instructions affect the acquisition of implicit and explicit knowledge? Is there an implicit-explicit interface? And what is the role of individual differences (e.g. working memory capacity), in the implicit and explicit learning of languages?

## Zoltan Dienes, Xiuyan Guo, Shan Jiang, and Feifei Li <br> Implicit learning of symmetries in tonal language

Implicit learning research has identified a number of structures that people can unconsciously learn, including chunks and fixed patterns of repetition. Language and music appear to involve structures more complex, indeed higher than finite state, for example symmetry structures that are simply generated by recursive rules (e.g. centre embedded, cross-serial dependency structures). The implicit learning of such structures presents an interesting challenge to existing models of implicit learning, such as the Simple Recurrent Network (SRN). We build on our earlier work in music and movement, by looking at symmetries in the tonal structure of Chinese poetry. We show that people can acquire unconscious knowledge of both cross-serial dependencies and centre embeddings in tonal poetry, with the former being easier than the latter. We also show that people can generalise their unconscious knowledge from being trained on strings of a certain length to test strings of a different length, indicating apparent learning of the symmetry itself rather than chunks or fixed length associations. We also show the SRN can model many of the details of this learning, exploring whether the SRN is more than a graded finite state machine.

## Clément François and Daniele Schön Implicit learning of linguistic structures and the effect of musical practice

Both speech and music involve sequences of sounds ordered in time according to complex rules. The acquisition of both
language and music require learners to engage several cognitive functions and notably the ability to sequence sound patterns. There is increasing evidence showing that the statistical regularities found in the input can play a important role in the implicit acquisition of several linguistic and musical structures. We previously showed that combining music and language into song can facilitate speech segmentation in implicit learning paradigms (Schön et al., 2008). Moreover, we recently conducted a set of experiments with adults and children showing that musical practice directly affects sensitivity to statistical regularities in speech both at the neural and behavioral levels (François \& Schön, 2011; François et al., 2012). Interestingly, our results seem to show that musical training and expertise have effects on brain plasticity that may go beyond primary auditory regions. These results also confirm that neurophysiological measures are more robust and sensitive than behavior to study implicit statistical learning processes.

## Martin Rohrmeier and Richard Widdess Implicit learning of musical grammar: The acquisition of North Indian music

Recent years have witnessed an increasing interest in the implicit and statistical learning of music (see Rohrmeier \& Rebuschat, 2012, for a review). Despite this interest, only few studies employed stimuli that resemble actual musical systems more closely, and only little research has been carried out on the acquisition of non-Western music. In this paper, we present the findings of a study that addressed this gap. The study focused on the implicit learning of modal melodic features in traditional North Indian music by Western learners who were unfamiliar with this musical system. Participants were trained on the ālāp (introduction) section of either the rāga Toṛī and Multānī and tested on novel excerpts from (later) jor sections of both rāgas featuring five distinct melodic features. Three of the five features were melodically distinctive of either rāga, whereas two were only distinctive based on other than mere pitch sequence features (for instance, emphasis). Findings indicated that Western participants unfamilar with Indian music learned to distinguish features of either rāga without intending to and after a very short exposure period. These results confirm that implicit learning constitutes a powerful mechanism that plays a fundamental role in the acquisition of highly complex, ecologically-valid musical stimuli.

## Moderators: Patrick Rebuschat and Martin Rohrmeier

## References

Rebuschat, P., Rohrmeier, M., Hawkins, J. H., \& Cross, I. (2011). Language and music as cognitive systems. Oxford: Oxford University Press.
Rohrmeier, M. \& Rebuschat, P. (2012). Implicit learning and acquisition of music. Topics in Cognitive Science (TopiCS), 4(4), 525-553.

# The Co-Existence of Naïve and Scientific Concepts in Learning and Development 

Moderator: Michael Schneider (University of Trier, Germany, m.schneider@uni-trier.de) Discussant: Stella Vosniadou (University of Athens, Greece) Participants:<br>Andrew Shtulman \& Kelsey Harrington (Occidental College, USA) Michael Schneider (University of Trier, Germany) \& Ilonca Hardy (University of Frankfurt, Germany)<br>Lindsey J. Powell \& Susan Carey (Harvard University, USA)<br>Stellan Ohlsson (University of Illinois at Chicago, USA)

Keywords: Conceptual change; conceptual development; naïve concepts; misconceptions; knowledge fragmentation.

## Motivation

Conceptual learning is sometimes described as replacement of incorrect knowledge by correct knowledge. However, a number of recent studies show that the storage of correct concepts in memory does not automatically erase related incorrect concepts from memory. As a result, naïve and scientific concepts in the same domain can coexist in a learner. The symposium aims at discussing these findings and their implications for definitions, models and empirical studies of conceptual learning and development.

From an empirical point of view, the coexistence of naïve and scientific concepts in learners raise the questions how common this phenomenon is, whether it differs between content domains and whether it changes in response to educational interventions, over the course of conceptual change, or over the life span. Intervention studies, crosssectional studies, and longitudinal studies are needed to investigate these aspects.

If learners simultaneously agree with naïve concepts and scientific concepts with various degrees, new operationalizations of conceptual knowledge are needed that adequately reflect this multidimensionality of knowledge. Researchers need to know when and how one-dimensional assessments of conceptual knowledge can bias empirical findings and what alternative methods yield more valid and reliable results.

Formal logic shows that from a contradiction one can derive any conclusion (ex falso quodlibet). According to coherence theories of truth, logical coherence is the defining characteristic of scientific theories. Thus, the coexistence of correct and incorrect conceptual knowledge in learners raises the questions whether a learners' conceptual knowledge in a domain can still be characterized as a theory, how incoherent knowledge influences learners' reasoning, and how learners evaluate the adequacy of their concepts.

The contributions to this symposium approach these problems from different theoretical and empirical angles. They use newly developed assessment tasks along with theoretical analyses, reaction time analyses and latent variable modeling. The studies compare age groups ranging
from 5-year olds over adults to the elderly and investigate scientific concepts as well we learners' theories of mind.

## Assessing the Resilience of Naïve Theories Across the Lifespan

Andrew Shtulman \& Kelsey Harrington

Three decades of research in cognitive development and science education have shown that students enter the science classroom with rich theories of everyday phenomena that often interfere with learning. Science educators are thus charged with two tasks: not only must they help students learn the correct, scientific theory at hand, but they must also help students unlearn their earlier, less accurate theories. This process has typically been characterized as a kind of radical restructuring, with scientific knowledge coming to overwrite earlier intuitions, but recent research suggests that those intuitions may never be fully overwritten. In this talk, I will present a method for assessing the resilience of early intuitions in potentially any domain of knowledge. This method entails asking participants to verify two types of statements as quickly as possible: "consistent" statements, whose truth value is consistent across both naïve and scientific theories of a particular domain (e.g., "The moon revolves around the Earth," which is true on both naïve and scientific theories of astronomy), and "inconsistent" statements, which involve the same conceptual relations but whose truth value differs across those theories (e.g., "The Earth revolves around the sun," which is true on a scientific theory but not a naïve theory).

If naïve theories continue to be represented in some form, then the latter should cause greater cognitive conflict than the former. Consistent with this hypothesis, adults have been shown to verify inconsistent statements more slowly and less accurately than consistent ones, and this effect has been documented in domains as diverse as astronomy, evolution, fractions, genetics, and mechanics. Naïve theories thus seem to survive the acquisition of a mutually incompatible scientific theory, coexisting with that theory for many years to follow. Indeed, preliminary research with an elderly population suggest that pre-scientific intuitions may persist across the lifespan.

# A Latent Transition Model of Naïve and Scientific Knowledge in Conceptual Change 

Michael Schneider \& Ilonca Hardy

Conceptual change requires learners to restructure parts of their knowledge base. Prior research has raised the questions to what extent misconceptions, everyday conceptions, and scientific concepts can co-exist during the course of conceptual change, whether this extent is stable over time, and how it changes in response to educational interventions. To address these questions we assessed 161 third-graders’ knowledge about floating and sinking of objects in liquids at three measurement points by means of multiple-choice tests. The tests assessed how strongly the children agreed with commonly found but mutually incompatible misconceptions, everyday conceptions, and scientific concepts about floating and sinking.

A latent profile transition analysis of the test scores revealed five profiles, some of which indicated the coexistence of inconsistent pieces of knowledge in learners. The majority of students (63\%) were on one of seven developmental pathways between these profiles. Children's knowledge profiles at a point in time were useful predictors of their further knowledge development. The extent of co-existence of misconceptions, everyday conceptions, and scientific concepts decreased on some individual developmental paths, increased on others, and remained stable on still others. The study demonstrates the usefulness of explicit quantitative models of conceptual change. The results support a constructivist perspective on conceptual development, in which developmental changes of a learner's knowledge base result from idiosyncratic, yet systematic knowledge construction processes.

## Using Executive Function Depletion to Assess Conflict between Advanced and Naïve Theories

Lindsey J. Powell \& Susan Carey

Demonstrations that children's executive function abilities (EF) correlate with conceptual development in diverse domains, including theory of mind, math, biology, and physical reasoning, have lead researchers to propose that EF plays a role in acquiring new concepts and theories. However, it may also be the case that this relationship partly reflects a critical role for EF in deploying new knowledge after its acquisition, especially if people persist in representing naïve theories that compete with their more sophisticated or scientific understanding of a given domain. We will discuss research that develops and deploys a methodological tool that can help disentangle questions of acquisition and expression by asking whether EF resources are necessary for children to use newer theories in place of older, naïve ones.

Adult research on executive function (or "ego") depletion has shown that deploying EF resources in one context decreases the ability to draw upon further EF resources immediately thereafter. An initial experiment demonstrated that the same is true for 5-year-old children. Subsequently,
we asked whether EF depletion would impair 5-year-olds' performance on a standard false belief task. While the performance of a control group suggested that children at this age have successfully acquired an explicit understanding of how beliefs impact actions, the EF depletion manipulation significantly impaired children's ability to use this understanding to guide their predictions of others' actions. A follow-up study asking children to explain rather than predict actions based on false beliefs suggests the role of EF in belief reasoning is not limited to suppressing an egocentric point of view. Even when presented with an outcome only consistent with their mature understanding of beliefs, children subjected to EF depletion were impaired in their ability to use that knowledge to generate an explanation of others' actions. Although applied here to theory of mind development, I will also discuss how this EF depletion method could be used to look at the role that EF plays in adjudicating between coexisting naïve and scientific theories in other domains.

## Cognitive Utility as the Arbiter Among CoExisting Knowledge Structures

## Stellan Ohlsson

There is little doubt that several different views of a topic can co-exist in a person's memory. For example, a science historian might be able to reason about a chemical reaction both from the perspective of the phlogiston theory and the perspective of the oxygen theory of combustion. Laboratory data from a re-categorization study in support of coexistence will be summarized briefly. If the individual components of knowledge structures are conceived as beliefs that are true or false, co-existence becomes problematic: What does the person 'really' believe? The belief-centered view also requires a theory of how people evaluate the relative strength of the evidence for alternative beliefs, a notoriously difficult problem.

Neither the history of science, nor social psychology, nor cognitive psychology has produced a widely accepted theory of how people decide on the strength of the evidence for or against a particular belief. For example, in both philosophical and political discourse, adherents of opposing views sometimes exchange arguments and other types of evidence for decades, even centuries, without resolving their disagreements, casting doubt on the idea that the evaluation of evidence is a real cognitive process. In this talk, I will develop the alternative idea that the quantity that people estimate is not the strength of evidence but the cognitive utility of a knowledge structure. The utility-based view dissolves some of the difficulties generated by the beliefcentered view, while raising some question of its own. An explanation will be offered why it seems as if (some) people are engaged in the evaluation of evidence. The utility-based view supports the notion of hands-on science instruction, but also explains why such instruction might fail under certain circumstances. The utility-centered view was anticipated by William James, Charles Sanders Peirce, and other pragmatist philosophers.

# Minimal Nativism: How does cognitive development get off the ground? 

Tomer D. Ullman and Joshua B. Tenenbaum<br>Department of Brain and Cognitive Sciences, MIT<br>Cambridge, MA USA 02139<br>Noah D. Goodman<br>Department of Psychology, Stanford University<br>Stanford, CA USA 94350<br>Shimon Ullman<br>Department of Computer Science and Applied Mathematics, Weizmann Institute of Science<br>Rehovot 76100 Israel<br>Elizabeth Spelke<br>Department of Psychology, Harvard University<br>Cambridge, MA USA 02138

Keywords: Cognitive development; Core knowledge; Learning; Computational modeling

When constructing a mind, what are the basic materials, structures and blueprints a young child has to work with? Are most of the structures already in place, with children merely working to embellish them? Do children begin with several buildings already in place (the Physics Building, the Social Building, the Number building, etc.), and only decorate a bit as they get older, perhaps building bridges between them using language? Such a view might describe a strong innate core hypothesis (Spelke et al., 1994). Or does the child begin with more of an empty plain, and an ability to construct whatever is necessary out of whatever materials are at hand at the time? Such a view might be more along the lines of classic empiricism (Quine, 1964).

Many other views are possible, lying somewhere between the extremes of positing that the child starts with everything, and positing that the child starts with nothing. For example, perhaps the child begins with a powerful general-purpose learning mechanism and a general blueprint for how to organize the world's entities into core domains, but no detailed, specific understanding of how these domains operate. Or perhaps the child begins with a powerful learning mechanism and a general blueprint for cognitive architecture, but no abstract concepts - only raw sensory experience. Yet if her sensory experience can be structured by a few crucial 'proto-concepts' - low-level input analyzers that tug her learning apparatus in certain appropriate directions - that minimal scaffolding could be sufficient.

Of course metaphors for cognitive development will only take us so far. In the last few years, a number of stimulating proposals for how cognitive development might get off the ground have been framed by computational modeling researchers, and these models offer to bring greater precision, clarity and subtlety to classic "nature versus
nurture" debates. At the same time, recent empirical work with young children offers striking new data that both motivates and challenges these computational accounts. Our symposium brings together some of the researchers who have contributed to these developments from both computational and empirical perspectives (Goodman, Ullman, \& Tenenbaum, 2011; Spelke \& Kinzler, 2006; Tenenbaum, Kemp, Griffiths, \& Goodman, 2011; Ullman, Harari, \& Dorfman, 2012; Ullman, Goodman \& Tenenbaum, 2010; Xu \& Kushnir, 2012). Our goals are to survey the landscape of developmental possibilities across multiple domains of physics, psychology, number, geometry, and language; to bring recent models and empirical work into closer contact; and to confront, honestly and clearly, the deep challenges that remain unaddressed.

Our plan is to have four 15 -minute talks, followed by a 30-40 minute discussion. T. Ullman will speak first, sketching out the space of potential approaches to a "minimal scaffolding" for cognitive development, and touching briefly on his own work modeling the development of intuitive physics, intuitive psychology, and the interface between these domains. N. Goodman will then present the "probabilistic language of thought" view - that an innate, abstract, domain-general, language-like ability for composing and manipulating conceptual representations is the minimal structure necessary for learning, potentially supplemented with specific 'named-functions' or inputanalyzers for certain domains. S. Ullman will then expand on the notion of innate perceptual input analyzers, illustrating with a case study drawn from his recent work on computer vision systems that learn to identify and reason about agents and actions in real-world video. E. Spelke will approach these issues from the standpoint of her recent work on the development of space, number and other mathematical concepts. She will also provide a more general critical perspective on the various computational perspectives presented earlier. This will set the stage for our
discussion, to be facilitated by Tenenbaum and Spelke, with the active involvement of audience participants as well as all our speakers.

## Tomer Ullman: The theoretical landscape, and a case study in the origins of physical and psychological knowledge

Cognition can be viewed as a program, albeit an incredibly complex one. Cognitive development then is the process by which the mind moves from one program to another. I will introduce a range of approaches to modeling cognitive development as different takes on the problem of "program induction" or "program synthesis". I will argue for the value of beginning with abstract templates that can capture deep patterns common to the explanatory structure of theories in many domains. I will show how this approach provides insight into the development of children's first physical and psychological concepts, such as force and utility, as well as the interface between these domains. I will briefly speculate on how these templates might arise or grow over the course of development or evolution.

## Noah Goodman: Minimal nativism and the language of thought

How much must be built into the language of thought? Universal formal languages can be built with a very small number of primitive operations, yet adult humans have a large number of conceptual operations ready-to-go for new situations. Indeed, developmental psychologists have argued that a significant and rich subset of these are innate primitives. I will argue that a universal language of thought together with a powerful learning mechanism is able to construct many of the needed concepts very quickly. However, I will find that some basic concepts can be learned more easily when supported by low-level modules that transform the perceptual input -- input analyzers. This combination cuts a middle road between strongly nativist and strongly empiricist view -- a minimal nativism.

## Shimon Ullman: Bootstrapping from domainspecific 'proto-concepts'

Already in their first months of life, infants rapidly learn to recognize complex objects and events in their visual input. Two striking examples are the detection of agents' hands and their direction of gaze, properties which play an important part in understanding actions and goals (Woodward 1998, Flom et al. 2007). In computational schemes, these problems are notoriously difficult. In contrast, detecting hands and gaze direction, and using them to make inferences and predictions, are natural for humans, and appear early in development. I will briefly describe how these problems can be solved using a learning scheme guided by an empirically motivated innate mechanism - the detection of 'mover' events in dynamic images, which are the events of a moving image region causing a stationary region to move or change after contact. The implications go
beyond the specific tasks, by showing how domain-specific 'proto concepts' can guide the system to acquire meaningful concepts, which are significant to the observer, but statistically inconspicuous in the sensory input.

Such proto-concepts may exist in other domains, forming a bridge between the notion of innate conceptual knowledge and that of learning mostly from sensory experience.

## Elizabeth Spelke: The origins of spatial and numerical thinking

When children begin to learn arithmetic, measurement, and geometric symbols such as maps, what cognitive systems support this learning process? I propose that this process is supported by four domain-specific cognitive systems: two core systems of number and two core systems of geometry. These systems are present and functional at the time that a child or animal first encounters the entities on which they operate: in this strong sense, they are innate. The systems also remain functional throughout life and support mathematical reasoning in adults as well as children: in this sense, they are foundations of mature mathematical reasoning. But the systems are far less general or powerful than the formal mathematical systems that children come to acquire, including the systems of natural number and Euclidean geometry. Powerful, domaingeneral systems for representing the information delivered by core systems, and for forming new concepts from this information, therefore constitute a fifth foundation for mathematics.

This may be the general scheme for much of later conceptual knowledge: combining core domains that have isolated innate concepts using later maturing domaingeneral systems.

## References

Flom R, Lee K, Muir D (2007) Gaze-Following: Its Development and Significance (Lawrence Erlbaum Associates, Mahwah, NJ).
Goodman, N. D., Ullman, T. D., \& Tenenbaum, J. B. (2011). Learning a theory of causality. Psychological review, 118(1), 110.

Quine, W. V. O. (1964). Word and Object. MIT Press.
Spelke, E.S., Breinlinger, K., Macomber, J., and Jacobson, K. (1992). Origins of knowledge. Psychological Review 99(4), 605.

Spelke, E. S., \& Kinzler, K. D. (2006). Core knowledge. Developmental science, 10(1), 89-96.
Tenenbaum, J. B., Kemp, C., Griffiths, T. L., \& Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. Science 331(6022), 1279-1285.
Ullman, S., Harari, D., \& Dorfman, N. (2012). From simple innate biases to complex visual concepts. Proceedings of the National Academy of Sciences.
Ullman, T. D., Goodman, N. D., \& Tenenbaum, J. B. (2012). Theory Acquisition as Stochastic Search in the Language of Thought. Cognitive Development 27(4), 455-480.
Woodward, A. L. (1998) Infants selectively encode the goal object of an actor's reach. Cognition 69:1-34.
Xu, F., \& Kushnir, T. (Eds.). (2012). Rational Constructivism in Cognitive Development (Vol. 43). Academic Press.

# Language as a Window into the Brain and its Pathologies 

Peter Garrard (pgarrard@sgul.ac.uk)<br>St George's, University of London \& St George's Stroke and Dementia Research Centre, UK.

Peter beim Graben (peter.beim.graben@hu-berlin.de)
Institut für deutsche Sprache und Linguistik, HumboldtUniversität zu Berlin, Germany.

Brita Elvevåg (brita@elvevaag.net)<br>Department of Clinical Medicine, University of Tromsø, \& University Hospital of North Norway Tromsø, Norway.

Eduardo Mizraji (mizraj@fcien.edu.uy)<br>Sección Biofísica, Facultad de Ciencias Universidad de la República, Uruguay.

Juan C. Valle Lisboa (juancvl@fcien.edu.uy)

Facultad de Ciencias, Universidad de la República, Uruguay.

Keywords: Language; Psychiatry; Schizophrenia; Model; Neural Networks

## Introduction

Traditionally, psycholinguistics and neuropsychology have been informed by conspicuous pathologies such as aphasia, which revealed the localization of some of the processes involved in language comprehension and production, in particular of those related to lexical access and morphological and syntactic processing. One of the main objectives of this symposium is to explore whether psychiatric pathologies are informative of the processes involved in meaning construction and comprehension, in the same way that aphasia research has contributed to our knowledge of the neurobiology of other aspects of language (Elvevåg, Helsen, De Hert, Sweers, \& Storms, 2011).

The kind and type of language disturbances displayed by patients can shed some light on the underlying pathologies, hopefully suggesting tractable lines for further research. Conversely, an improved understanding of the mechanism of psychiatric diseases can promote the understanding of some intricate aspects of non-pathological language production. In this Symposium we bring together different perspectives to the study of language in pathology and as a manifestation of the underlying neural networks' workings. The symposium is thus centered around language, both from an empirical perspective and a modeler's point of view. We describe this two aspects in the following sections.

## Language analysis in pathology

The last decades have seen a tremendous increase in the development of techniques to study the physiological and pathological processes in the brain. Among them, the study of language production and comprehension has been recognized as a central research topic. The disturbances in language vary between different brain pathologies. For instance Garrard and coworkers (Garrard, Maloney, Hodges \& Patterson, 2005) have shown that Alzheimer's disease can
notoriously affect the way lexical items are selected and used by a writer, even before the symptoms of the disease are apparent. Several measures of language comprehension and production have been used to assess the presence and course of schizophrenia and mania. Classically, for patient state evaluation, there is a variety of fine-grained rating scales of the coherence of speech and communication, such as the Scale for the Assessment of Thought, Language and Communication (Andreasen \& Grove, 1986). The use of these scales requires extensive training and the results remain open to variance across raters. Several recent approaches attempt to devise automatic or semiautomatic methods aimed at diagnosis or with the purpose of understanding the cognitive deficits involved (Cabana,Valle-Lisboa, Elvevåg, \& Mizraji, 2011). The description of the underlying causes of these alterations is specially relevant when these methods are combined with imaging and modeling data. We turn next to models.

## From data to models

The complex patterns of linguistic productions, with the associated generation of measurable neurobiological responses, can only be interpreted in light of models and theories. The other aspect of the symposium is the use of neural models and theories in order to interpret the data, to propose new experiments and to suggest new therapeutic avenues for research. The models of choice are based on connectionist approaches. There is an increasing tendency to develop connectionist models of psychiatric pathologies (e.g. Hoffman, Grasemann, Gueorguieva, Quinlan, Lane, \& Miikkulainen, R., 2011). At the symposium two types of models will be discussed. One type is that of coarse-grained models that when deteriorated can lead to alterations of language processing and production that mimic some of the properties displayed by language produced by patients (Valle-Lisboa, Pomi, Cabana, Elvevåg \& Mizraji, 2013 ). This type of model is aimed at the level of brain networks as
derived from fMRI. The other type of model is concerned with the syntactic level of description and its expression through brain potentials as measured through EEG (beim Graben \& Potthast, 2012). Interestingly, both type of models can be related through their basic assumptions (i. e. multiplicative synaptic interactions).

The symposium will consist of a 10 minute introduction by the chairman followed by 20 minute talks by each speaker and followed by a round of discussion.

## Speakers

## Dr. Peter Garrard (pgarrard@sgul.ac.uk)

Dr Garrard is a Neurologist with vast experience in the study of semantic dementia and Alzheimer's disease, in particular the linguistic manifestations of those pathologies. He has developed several diagnostic procedures and is currently exploring the symptoms and neuroimaging manifestations of patients suffering atypical dementias. He is Reader in Neurology at St George's, University of London and Honorary Consultant Neurologist, St George's Stroke and Dementia Research Centre, UK.

Dr. Peter beim Graben (peter.beim.graben@hu-berlin.de).
Dr beim Braben is a Physicist who works in Cognitive Neurodynamics, Computational Psycholinguistics and Computational Neuroscience. His recent work is related both to the processing of sentences by neural networks and the electrical potential measurements of neural activity. He is DFG Heisenberg Fellow for Cognitive Neurodynamics, at the Institut für deutsche Sprache und Linguistik, HumboldtUniversität zu Berlin, Germany.

## Dr. Brita Elvevåg (brita@elvevaag.net).

Dr Elvevåg is a research psychologist interested in cognitive neuropsychiatry, cognitive neuroscience, and the cognitive, neural and genetic basis of language in healthy individuals and those with clinical conditions that affect the brain, especially psychosis and dementia. She is Professor of Psychiatry in the Department of Clinical Medicine at the University of Tromsø, and also at the Norwegian Centre for Integrated Care and Telemedicine at the University Hospital of North Norway Tromsø, Norway.

## Dr. Eduardo Mizraji (mizraj@ffien.edu.uy)

Dr. Mizraji is a Biophysicist who studies neural network theory and the implementation of high level cognitive activities through the use of neural models. The part of his recent works more related to the symposium are concerned with discourse generating neural models, and the emergence of symbolic activities in a neural network. He is Professor
of Biophysics in the School of Sciences, Universidad de la República, Uruguay.

## Chairman and Organizer

## Dr. Juan Valle Lisboa (juancvl@fcien.edu.uy)

Dr Valle Lisboa is a Biophysicist and Neurobiologist interested in theoretical neuroscience and the neural basis of high level cognition. His latest works include the development of language production models and their deterioration in pathologies as well as neural models of lexical representation. Dr Valle Lisboa is at the School of Sciences and the School of Psychology at the Universidad de la República, Uruguay.

## Acknowledgments

JCVL and EM are supported by PEDECIBA and CSIC (UDELAR).

## References

Andreasen NC and Grove WM. (1986) Thought, language and communication in schizophrenia: diagnosis and prognosis. Schizophrenia Bulletin, 12:348-359.
beim Graben, P. \& Potthast, R. (2012). A dynamic field account to language-related brain potentials. In Rabinovich, Friston, \& Varona (Eds.) Principles of Brain Dynamics: Global State Interactions, MIT Press: Cambridge (MA).

Cabana Á, Valle-Lisboa JC, Elvevåg B, and Mizraji E. (2011) Detecting order-disorder transitions in discourse: Implications for schizophrenia. Schizophrenia Research, 131:157-164.

Elvevåg, B., Helsen, K., De Hert, M., Sweers, K. \& Storms, G. (2011). Metaphor interpretation and use: a window into semantics in schizophrenia. Schizophrenia Research, 133, 205-211

Garrard P, Maloney L, Hodges J, Patterson K. (2005) The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author. Brain 128: 250-260.

Hoffman, R, Grasemann, U., Gueorguieva, R., Quinlan, D., Lane, D., and Miikkulainen, R. (2011) Using Computational Patients to Evaluate Illness Mechanisms in Schizophrenia. Biological Psychiatry 69:997-1005.

Valle-Lisboa, J.C. , Pomi, A., Cabana, A, Elvevåg, B and Mizraji, E. (2013) A modular approach to language production: Models and facts. Cortex (in press).

# The empirical investigation of semantics: Between abstract-symbolic and embodied-simulative models of meaning 

Markus Werning (markus.werning@rub.de)<br>Department of Philosophy, Ruhr University Bochum 44780 Bochum, Germany<br>Erica Cosentino (ericacosentino@libero.it)<br>Department of Philosophy, University of Calabria 87036 Rende (CS), Italy

Keywords: embodiment; abstract symbol theories, EEG, emulative semantics

## Introduction

Classical theories of meaning in the field of linguistics and psycholinguistics assume that meaning arises from the combination of symbols for which a substring or other partwhole relation is defined. According to this perspective, symbols are abstract, amodal (i.e., neither perceptual, nor motoric) and only contingently related to entities in the external world.

For a long time, a convincing case for classical models has been the absence of alternatives. However, more recently, several theories subsumed under the notions of "embodied" or "grounded" theories have challenged the fundamental assumptions of classical models (e.g., Barsalou, 1999; Glenberg, 2010; Pecher \& Zwaan, 2005). From the point of view of embodied theories, cognition is grounded in modal representations which simulate actual objects, properties and situations. Such a claim carries theoretical, empirical and methodological repercussions that also change the way linguistic processes are conceived of (Ferretti et al., 2013; Zwaan \& Radvansky, 1998). The goal of the symposium is to explore which repercussions these issues have on the nature of linguistic meaning and its neural and cognitive realization or representation.

A main motivation for the symposium is that, in spite of the relevance of the issue, to this day the relationship between classical and embodied models of meaning is still not clear. More to the point, it is not clear whether classical and embodied models describe different aspects of meaning (and are then compatible) or whether they are mutually exclusive explanations of the same phenomena. As the nature and representation of meaning is a topic of increasing cross-disciplinary interest, the symposium aims to encourage an in-depth discussion among scholars interested in the problem, providing a cross-disciplinary forum of dialogue.

A further motivation for the symposium is that researchers in many specialized fields of cognitive sciences have been providing results which seem to support at least
some form of embodiment (Meteyard et al., 2012; Vigliocco et al., 2011). However, due to the specialization of competences, the circulation of these results among scholars in different arenas has not always been easy. This symposium will provide an opportunity to bring together philosophers, linguists, psychologists and neuroscientists joined by a common interest in the application of experimental methods to the analysis of the nature of meaning.

Given the interdisciplinary nature of the symposium, we would like to stress that the speakers are not only specialized in the field of semantics, but that they are also able to work at the crossroad between several disciplines. The symposium organizers share a background in philosophy with an active interest in neurolinguistics, at both a theoretical and an experimental level (Ferretti et al., 2013; Werning, 2012).

## Speakers and abstracts

## Erica Cosentino

Classical theories of meaning are two-step models, according to which contextual information is considered only after establishing phrase or sentence local meaning. In this perspective, local semantics cannot initially be overruled by the wider context. In this study we tested this prediction analyzing the effect of discourse context on affordances. Two-steps models predict that a verb-object violation, as in "She uses the funnel to hang her coat" will always be considered inappropriate, regardless of the wider discourse. In the current study we found that when this anomalous combination is embedded in a neutral context it elicits a typical N400, indicating that the subject is experiencing interpretative problems. However, when preceded by a supportive context, the very same sentence becomes perfectly acceptable, as reflected by the absence of an N400 effect. This finding challenges the classical approach to meaning suggesting that affordances are immediately integrated in the construction of meaning and that contextual information is immediately taken into account.

## Lars Kuchinke

The embodied-simulative view proposes that linguistic meaning is grounded in memory traces in modality-specific brain regions as distributed neural representations of previously experienced internal and external states. This view also subsumes emotional information linked to words, and recent empirical evidence from emotional word recognition supports this assumption. Electrocortical findings point to a very early locus of these emotional effects preceding or at least altering the onset of lexical access. These effects are modulated by emotional valence, leading back to the 'semantic cohesiveness' hypothesis that proposes differences in the amount of semantic associations of valenced words. We recently proposed an associative read-out model based on co-occurence statistics that
correctly predicts higher recognition rates for words with a greater amount of associations to other stimuli. Based on this model, we show that effects of positive words can be explained by their semantic cohesion, whereas negative words explain variance beyond their associations.

## Gabriella Vigliocco

Theories of semantic representation ought to account for how we use this information, how we learn it and how it breaks down after brain damage. I will set the stage locating existing theories along a continuum from disembodied to fully embodied and presenting evidence from behavioural, imaging and patients' studies that limit the viable theories to those that incorporate some degree of embodiment but also include information from other sources such as language. I will then give a bird eyes overview of one such theories that we have developed in the past few years in which all concepts (concrete and abstract) are grounded in our sensory motor and affective experience but also statistical information from language contributes to the learning and representing meaning.

## Markus Werning

In the first part of the talk, the central tenets of Emulative Semantics will be outlined. In the second part an EEG-based case study on the understanding of linguistic emotion contexts will be presented. (1) Emulative Semantics (Werning, 2012) is a naturalist theory of meaning. It claims that linguistic meaning consists in patterns of neuroemulations. Unlike rival naturalist theories of meaning, Emulative Semantics is a non-symbolic, but still compositional theory of meaning. Neuro-emulations are abstractly described dynamical states of the brain's sensorymotor regions that are partially isomorphic to modeltheoretical structures. Emulative Semantics thus inherits many formal features of model-theoretical semantics, which has been very successful as a formal account of meaning. (2) One prediction of Emulative Semantics is that the understanding of linguistic contexts about emotional scenarios should involve the emulation of emotions. Since the emulation of emotions is also thought to be a basis for the human capacity of empathy, Emulative Semantics predicts a correlation between empathy with emotions and the comprehension of linguistic emotion contexts. In a recent ERP study we could in fact show that, in linguistic emotion contexts, the N400 effect, which indicates violations of semantic expectations, depends on empathy as measured by the Multifaceted Empathy Test.

## Rolf A. Zwaan

Language comprehension involves the construction of mental representations. This seems an uncontroversial statement in most of cognitive science. Much research has focused on the nature of these representations: are they textbased or situational (or both) are they abstract or grounded in perception and action? My goal here is to propose and
describe an integrative view. I will do this by discussing recent research from my lab.

## Symposium program committee

Prof. Dr. Markus Werning
Dr. Erica Cosentino

## References

Barsalou, L.W. (1999). Perceptual symbol systems. Behavioral and Brain Sciences, 22, 577-609.
Ferretti F., Adornetti I., Cosentino E., Marini A. (2013). Keeping the route and speaking coherently: the hidden link between spatial navigation and discourse processing, Journal of Neurolinguistics, 26(2), 327-334.
Glenberg, A. M. (2010). Embodiment as a unifying perspective for psychology. Wiley Interdisciplinary Reviews: Cognitive Science, 1, 586-596
Meteyard, L., Cuadrado, S.R., Bahrami, B., Vigliocco, G. (2012). Coming of age: a review of embodiment and the neuroscience of semantics. Cortex, 48 (7), 788-804.
Pecher, D., Zwaan R (eds.) (2005). Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thought. New York: Cambridge Univ. Press.
Vigliocco, G., Vinson, D.P., Druks, J., Barber, H., Cappa, S.F. (2011). Nouns and verbs in the brain: a review of behavioural, electrophysiological, neuropsychological and imaging studies. Neuroscience \& Biobehavioral Reviews, 35(3), 407-426.
Werning, M. (2012). Non-symbolic Compositional Representation and Its Neuronal Foundation: Towards an Emulative Semantics. In Werning, M., Hinzen, W., \& Machery, M. (Eds.), The Oxford Handbook of Compositionality. Oxford University Press, Oxford (pp. 633-654).
Zwaan, R. A. and G. A. Radvansky (1998). Situation models in language comprehension and memory. Psychological Bulletin, 123, 162-185.

# Online Education: A Unique Opportunity for Cognitive Scientists to Integrate Research and Practice 

Joseph Jay Williams (joseph_williams@berkeley.edu)<br>Department of Psychology, University of California at Berkeley

Ken Koedinger (koedinger@cmu.edu)<br>Human Computer Interaction Institute, Carnegie Mellon University

Alexander Renkl (renkl@psychologie.uni-freiburg.de)<br>Psychological Institute, University of Freiberg<br>John Stamper (jstamper@cs.cmu.edu)<br>Human Computer Interaction Institute, Carnegie Mellon University


#### Abstract

That there is a rapid expansion of online education is much better understood than what its consequences will be. This symposium considers that one key feature of "real-world" education that takes place on the Internet is that it provides a high level of experimental control and automatic data collection \& analysis, which can support cognitive science research that was previously only possible in laboratory settings and small-scale educational environments. The presenters discuss the unprecedented opportunities online learning provides for conducting research in ecologically valid contexts: linking existing laboratory experiments to relevant online contexts, personalizing adaptive instruction, embedding in vivo research studies of education, and using the vast amount of high quality data available. The product of such work is not only theories and empirical discoveries that better characterize learning, but also the opportunity to directly translate these into practical benefits to students through concrete improvements to educational resources.


Keywords: learning; education; technology; online education; online learning; e-learning; intelligent tutors; educational data mining;

Laboratory experiments and classrooms rarely overlap in the physical world, much to the chagrin of educational psychologists. But researchers increasingly use computers and the Internet to run experiments, and the recent explosion of online education now brings student learning into the very same digital medium. Hundreds of thousands of students take Massive Open Online Courses (MOOCs from Coursera, Udacity, EdX) at the university level, use websites like www.mathtutor.cmu.edu, www.khanacademy.org or www.mathalicious.com that are populated with K-12 videos \& interactive exercises, not to mention a host of supplementary online educational resources that can be delivered over devices as ubiquitous as smartphones.

These web-based resources offer the potential for extensive novel research on learning (Anderson, 2008; Ally, 2004; Linn et al, 2004; Pea, 2003). One distinctive feature is the possibility of embedding in vivo randomized experimental comparisons (Koedinger et al, 2012) into these (now) ecologically valid online educational environments. Furthermore, unlike educational environments in bricks-and-mortar education, in a digital medium there can be precise control over materials and instructions, and systematic collection of data from large samples (Stamper et al, 2012). In addition to investigating learning processes in authentic educational contexts, studying online learning
provides an unprecedented opportunity to simultaneously carry out basic and applied research. When experimental manipulations correspond to comparisons of instructional effectiveness, stimuli are educational materials, and dependent measures are students' learning outcomes, experimental methodology from cognitive science can be used to iteratively improve the pedagogical principles incorporated into online educational resources. Rather than years of laboratory research that "suggest" instructional principles or are eventually followed by a laborious classroom study, the steps from basic to translational research are greatly simplified.

Moreover, the product of research programs that investigate online learning is not only new scientific knowledge, but specific products that concretely instantiate theories and learning principles. These proven and iteratively improved resources can be provided directly to students for use as they exhibit such great fidelity to research context. And using the Internet, they can be disseminated to hundreds of thousands of students over an extended period of time, all across the world - a clear contribution of cognitive scientists to public education.

## Mapping laboratory studies to online educational settings

In the context of his research on the role of explanation in learning, Joseph Jay Williams presents a perspective from basic experimental psychology on finding fruitful connections between lab experiments and experimental manipulations embedded in authentic online educational resources. He discusses the interplay and transitions between typical lab experiments (research on explaining membership in artificial categories, Williams \& Lombrozo, 2010), to online studies using convenience samples from Amazon Mechanical Turk (in which explaining promotes learning of mathematics, Williams et al, 2012), and experiments implemented using identical mathematics exercises on Khan Academy's educational platform, but with real students who visit the site for genuine help with authentic schoolwork.

This approach blends rigorous experimentation in contexts with different levels of control, rapid iterative improvement, and the development of an ecologically valid educational resource. Such web resources serve a basic research goal, as (for example) the structure and dynamics of an interactive video or exercise reflects a concrete and empirically supported instantiation of theoretical principles
and empirical discoveries that would otherwise be verbally communicated - reminiscent of the benefits of computational models. They also serve a clear practical goal, as the products of research are empirically evaluated and multiply revised resources that have been shown to work in learning online. They can therefore be disseminated using the massive scale of the internet, to improve education for thousands of students over an extended period of time.

## Using Eye-Tracking and Rapid Assessment to Detect and Address Knowledge Gaps During Learning

Alexander Renkl has conducted extensive research in computer environments, such as investigating how learning from worked examples can be enhanced by the use of scaffolding, fading and other instructional design features. He reports a project that bears on the ability of online environments to personalize instruction.

The project develops and researches an adaptive approach to closing gaps in students' knowledge that remain after initially studying the learning materials. Rapid assessment tasks are interspersed in the learning environment to identify the knowledge deficits in individual students, which can then be targeted by prompting learners to engage in remedial activities. In the first experiment, university students $(\mathrm{N}=71)$ learned about mitosis in a multimedia learning environment. When rapid assessment tasks indicated gaps in students' knowledge, the experimental manipulation was to randomly assign them to one of three different types of prompts, hypothesized to be optimal for the particular deficit. However, we found that each type of prompt was equally effective in closing the knowledge gaps identified by rapid assessment.

In the second experiment we obtained further results as to the effects of different prompts - finding that ostensibly "enriched" prompts even led to sub-optimal effects. Comparing our results to the final test of learning outcomes, we identified our rapid assessment tasks as failing to detect important knowledge gaps. Rather than risk disturbing learning by increasing the number of rapid assessment tasks, we now integrate eye-tracking data to improve our assessment. We use this data to select rapid assessments to verify the presence of a knowledge gap. Pilot data suggests such a combined approach is more effective for learning.

## In-vivo experiments with cognitive tutors

As director of the Pittsburgh Science of Learning Center and one of the pioneers in the development of intelligent tutoring systems, Ken Koedinger has helped set the standards of rigorous research in realistic educational contexts. He discusses how cognitive tutors can be used to conduct in vivo experiments in classroom environments, collecting sophisticated learning measures and giving personalized feedback to learners. He considers how cognitive tutors and in vivo experiments can be integrated with online education platforms to take advantage of
complementary strengths and disseminate these benefits on a large scale.

## Automated improvement of instructional systems using educational data mining

John Stamper runs Datashop, the largest openly available repository of detailed student learning data, with learning and instructional events logged as often as every 10 seconds, scored, and tagged based on models of student learning. He will discuss how educational data mining and statistical models of students' learning can be used to infer improvements to instructional systems, automatically develop intelligent tutoring systems, and shed light on many of the problems that typically are solvable only through extensive human investment.

## References

Anderson, T. (2008). The theory and practice of online learning. Au Press.
Ally, M. (2004). Foundations of educational theory for online learning. Theory and practice of online learning, 331.

Koedinger, K. R., Aleven, V., Roll, I., \& Baker, R. (2009). In vivo experiments on whether supporting metacognition in intelligent tutoring systems yields robust learning. In D. J. Hacker, J. Dunlosky, \& A. C. Graesser (Eds.), Handbook of Metacognition in Education (pp. 897-964). The Educational Psychology Series. New York: Routledge.
Linn, M. C., Davis, E. A., \& Bell, P. (Eds.). (2004). Internet environments for science education. Mahwah, NJ: Lawrence Erlbaum Associates.
Means, B., Toyama, Y., Murphy, R., Bakia, M., \& Jones, K. (2010). Evaluation of evidence-based practices in online learning: A meta-analysis and review of online learning studies.
Pea, R. D. (Ed.). (2003). Planning for two transformations in education and learning technology: Report of a workshop. National Academy Press.
Renkl, A. (2011). Instruction based on examples. In RE Mayer \& PA Alexander (Eds.), Handbook of research on learning and instruction (pp. 272-295). New York, NY: Routledge.
Stamper, J., Lomas, D., Ching, D., Linch, K., Ritter, S. (2012) Internet Scale Experimental Design and Deployment for Educational Games using BrainPOP. In Proceedings of the 8th Games + Learning + Society Conference (GLS 2012). Madison, WI. pp. 275-281.
Williams, J.J. \& Lombrozo, T. (2010). The role of explanation in discovery and generalization: evidence from category learning. Cognitive Science, 34, 776-806.
Williams, J. J., Walker, C. M., \& Lombrozo, T. (2012). Explaining increases belief revision in the face of (many) anomalies. In N. Miyake, D. Peebles, \& R. P. Cooper (Eds.), Proceedings of the 34th Annual Conference of the Cognitive Science Society (pp. 1149-1154). Austin, TX: Cognitive Science Society.

# Causal Reasoning with Continuous Outcomes 

Ahmad Azad Ab Rashid (abrashidaa@cardiff.ac.uk)<br>School of Psychology, Tower Building Park Place<br>Cardiff, CF10 3AT, UK

Marc J Buehner (buehnerm@cardiff.ac.uk)<br>School of Psychology, Tower Building Park Place<br>Cardiff, CF10 3AT, UK


#### Abstract

We describe an attempt to understand causal reasoning in situations where a binary cause produces a change on a continuous magnitude dimension. We consider established theories of binary probabilistic causal inference $-\Delta P$ and Power PC - and adapt them to continuous non-probabilistic outcomes. While $\Delta \mathrm{P}$ describes causal strength as the difference of effect occurrence between the presence and absence of the cause, Power PC normalizes this difference with the effect base-rate to obtain a proportional measure of causal power, relative to the maximum possible strength. Two experiments compared the applicability of each approach by creating scenarios where binary probabilistic scenarios were directly mapped onto inference problems involving continuous magnitude dimensions. Results from counterfactual judgments tentatively indicate that people reason about causal relations with continuous outcomes by adopting a proportional approach when evaluation preventive causal powers, and a difference approach in generative scenarios.


Keywords: causal learning; continuous outcomes; reasoning; counterfactual.

## Background

The capacity to learn about and represent causal knowledge is a fundamental aspect of cognition without which humans lose the ability to not only make predictions and decisions, but also to forecast, prepare and direct their behaviours towards achieving goals and fulfilling desires. Current research mostly focuses on causal relations involving binary events. Outside the lab, however, people do not only encounter binary events. In fact, we are more likely to be dealing with continuous variables: How much faster could I run if I lose 20 pounds of weight? How much weight would I gain if I ate cheeseburger everyday? How much sugar do I need to add to avoid over sweetening? These questions are daily examples of people's involvement with causal relations entailing continuous variables.

Binary causal relations involve a state change of a binary event (cause present/absent) to produce a change in another binary event (effect present/absent), but such simplicity is not the case for continuous variables. In a continuous causal scenario, a magnitude change of a continuous variable is produced by a magnitude change of another continuous variable. For example, in a binary relation, a state change of a cause could be flicking a switch from off to on which changes the status of a bulb from off to on. On the other
hand, a continuous relation involves a change of a dial position to cause a change of luminosity from dimmer to brighter. Despite many daily-life examples of continuous variables, very few studies have been investigating causal judgment involving continuous variables (White, 2001). Here we are trying to find out how people acquire causal knowledge involving continuous variables?

## Learning Framework: Difference or Proportion

Most theories of binary causal learning are rooted in Hume's empiricism (1739/1888): Causal knowledge is not explicitly available via sensory modalities but instead is inferred using the input received via them. One of Hume's cues to causation is contingency - i.e. the frequency of an effect and a cause co-occurring.

A longstanding model formalising contingency as an indicator of causal belief is $\Delta \mathrm{P}$, which calculates the difference of the probabilities of the effect in the presence vs. the absence of the cause (Jenkins \& Ward, 1969):

$$
\Delta P=P(e \mid c)-P(e \mid \neg c)
$$

Consider these hypothetical scenarios involving the study of skin rash as a side effect of a new group of medicines on a group of forty patients. In scenario 1 , none of them had a rash before taking medicine A , but 20 of them had rash after taking the medicine. In scenario 2 , also none of them had rash before taking the medicine, but only 10 of them reported rash after taking medicine B . $\Delta \mathrm{P}$ computes causal strength by considering the difference in relative frequencies of patients before and after taking the medicines, giving $\Delta P$ values of 0.50 and 0.25 respectively; hence concluding that medicine A has higher causal strength than medicine $B$ to cause skin rash.

Consider another scenario 3 in which 20 of 40 patients already had skin rash even before taking medicine C , but the number of patients suffering with rash increased to 30 after taking the medicine. Applying $\Delta \mathrm{P}$ in scenario 3 results in medicine C having a causal strength index of 0.25 , which is similar to medicine B. However, studies have shown that despite having the same $\Delta \mathrm{P}$ values, people tend to conclude that medicine C is more effective than medicine B in causing the rash (Cheng, 1997; Buehner, Cheng, \& Clifford, 2003). This discrepancy is captured by another influential theory on causal learning: Power PC (Cheng, 1997).

Power PC argues that in addition to the difference causal strength is also influenced by the base-rate, $\mathrm{P}(\mathrm{e} \mid \neg \mathrm{c})$. Power

PC normalizes the difference with the base rate to obtain a proportional measure of causal power.

$$
p_{g e n}=\frac{\Delta P}{1-P(E \mid \neg C)} \quad p_{\text {pre }}=\frac{-\Delta P}{P(E \mid \neg C)}
$$

Power PC has also been used to parameterise Bayesian models of causal learning (Griffiths \& Tenenbaum, 2005) and is generally recognized as a rational account of causal strength.

Applying Power PC onto scenarios 2 and 3 results in having causal strength indexes of 0.25 and 0.50 for medicine B and C respectively. Unlike $\Delta \mathrm{P}$, this model therefore captures people's ability to provide normative responses. The key difference between $\Delta \mathrm{P}$ and Power PC is that the former considers the absolute difference the cause makes to the occurrence of the effect, while the latter calculates the difference relative to the maximum causal change possible, and thus provides a proportional index of causal strength.

In the earlier scenarios, medicine B had the opportunity to cause skin rash in all 40 patients, and did so in 10 of them; in contrast, in the scenario involving medicine C , the medicine only had the opportunity to cause skin rash in 20 patients because the other 20 already had rash even before taking the medicine. From these 20 unaffected patients, medicine C managed to affect 10 of them to have skin rash. Therefore, Power PC suggests that for medicine B, the causal strength index is 0.25 because 10 out of 40 patients had rashes whereas for medicine C it is 0.50 because it caused rashes in 10 out of 20 (i.e. the initially unaffected) patients.

Moreover, the Power PC theory also tackles ceiling and floor effects. In another scenario where all 40 of the patients already had skin rash before taking medicine D , and all 40 still had skin rash after taking the medicine, $\Delta \mathrm{P}$ for this scenario would be zero, suggesting that medicine D makes no difference to the occurrence of rash. A rational judgment, however, would be that the experiment is inconclusive with respect to generative causal power because medicine D had no opportunity to demonstrate its potential effectiveness, and thus the causal status of D is unknown. Wu and Cheng (1999) showed that reasoners indeed follow this logic, and withhold judgment in cases where causal power is unknowable. If Power PC is applied to this scenario, the equation is undefined (due to division by 0 ), which is consistent with both rational assessment and empirical results.

We highlighted the contrast between the difference and proportional perspectives of both theories because they will be relevant when considering approaches to continuous causation. Proportions can only be computed with respect to a reference limit. In binary probabilistic causation, the relevant limits are $\mathrm{P}(e)=0$ (the effect never happens) and $\mathrm{P}(e)=1$ (the effect always happens). These probabilities provide the upper limit of maximal causal effectiveness for preventive and generative causation, respectively, in a binary probabilistic framework: The maximum impact a preventor could have would be to reduce the probability of
the effect to 0 , while the maximum impact of a generator would be to raise it to 1 . When considering causal changes to continuous outcome magnitudes, such natural limits are not necessarily present. While the maximum impact a preventor could have would still be to reduce the quantity of the effect to 0 magnitude, the maximum impact a generator could have might be unknown because it could keep on increasing the magnitude unless there is a known upper limit.

## Study Scope

The central idea of this study was to investigate whether people reason about causal relations involving nonprobabilistic continuous outcomes within a difference or proportional framework. Because of the wealth of prior works assessing the suitability of these approaches with respect to binary probabilistic causation, we wanted to create scenarios that afford a similar comparison between the two accounts. To this end, and as a first step on our quest, we only considered situations where a binary cause can produce a (deterministic) magnitude change on a continuous variable. This allowed us to set up situations that are one-to-one mappings of binary probabilistic causation to scenarios involving continuous outcomes. More specifically, in both cases the cause is still either present or absent, but instead of it resulting in a change of probability of the outcome, it now affects the magnitude of the outcome.
In probabilistic causation the (binary) cause results in a binary state-change across a group of entities; aggregating these state-changes across a sample results in an assessment of the change of probability of the effect brought about by the presence of the cause, which is of course a continuous variable bound between 0 and 1 . In contrast, we considered changes of a continuous outcome magnitude in a single entity. This allowed us to preserve exactly the same structure as in probabilistic causal inference tasks. For example, a probability condition of $\mathrm{P}(\mathrm{e} \mid \mathrm{c})=0.75$, which indicates that skin rash is present in 75 out of 100 patients given that all of them took the medicine, was mapped onto a quantity condition of $\mathrm{Q}(\mathrm{e} \mid \mathrm{c})=7.5 \mathrm{~cm}^{2}$, indicating that 7.5 $\mathrm{cm}^{2}$ of skin from an area of $10 \mathrm{~cm}^{2}$ where the ointment was applied broke out with a rash.

In order to maximize comparability to binary probabilistic causation and preserve structural identity, our studies employed an artificial upper limit on a continuous scenario to serve as a reference for maximum causal effectiveness (see Method). Imposing such a limit allowed us to derive predictions not only for a difference based, but also for a proportional approach. Moreover, it afforded the opportunity for a more stringent test of the two approaches, by using different counterfactual scenarios to elicit causal judgments. More specifically, we asked one counterfactual question where the upper limit of causal effectiveness corresponded to the artificial limit in the learning phase, while another made reference to a higher limit, not previously experienced in the learning phase. If reasoners
approach causal inference problems involving continuous outcome magnitudes with a difference-based approach, changing the reference limit should have no impact on their predictions for causally induced magnitude change: All that would matter is the difference the cause made in the learning phase, regardless of the upper limit of causal effectiveness. In contrast, according to a proportional approach, reasoners would relate that difference to the maximum possible difference, and scale their predictions accordingly in the presence of a different limit.

Imagine that a government wants to test the efficacy of a 20 mph speed limit on traffic fatalities in residential areas. Community A serves as a pilot and fatalities are reduced from 20 per year before the trial to 10 per year after the trial. What would we predict if community $B$, which is larger, has more roads, and suffers from 50 fatalities a year, were to adopt the same program? According to a difference-based approach, we would predict that the program results in the same absolute reduction by 10 , to result in 40 fatalities per year. The proportional approach would consider the maximum change possible in A (20) and would recognize that 10 corresponds to half of that. Consequently, it would predict a reduction from 50 to 25 . We used a similar logic to compare difference to proportion based approaches.

## Experiments

## Participants

Thirty different undergraduates from Cardiff University's School of Psychology participated in each preventive and generative experiment in exchange for course credit.

## Design and Procedure

Each participant worked on 15 conditions directly adapted from the binary probabilistic design of Experiment 1 in Buehner et. al. (2003). Each condition consisted of a pair of quantities of an effect in the presence vs. absence of the cause (see Table 1).

The generative experiment used a cover story that asked participants to imagine they were pharmaceutical consultants researching the side effects (skin rash) of synthetic substances in cosmetic creams. Fifteen different fictitious cosmetic creams corresponded to the 15 causal conditions in Table 1.

The cover story also described that the size of skin rash was measured before and after the application of the cream, and that some patients may develop skin rash even in the absence of any cosmetic products. Instructions stressed that each cream was applied to cover $10 \mathrm{~cm}^{2}$ of a patient's back and that the base rate (rash before cream application) was also expressed with reference to this $10 \mathrm{~cm}^{2}$ area. This served to impose an artificial limit of maximum causal efficacy - the cream could only create rash so as to cover the entire $10 \mathrm{~cm}^{2}$ area.

A similar cover story was used for the preventive experiment, this time introducing ointments that relieve skin rash. Again, adopting the same 15 conditions, the story
described a proper motivation on how allergic reaction would cause the skin rash to occur up to $10 \mathrm{~cm}^{2}$ without any preventive measure, and on how the ointment would reduce the skin rash.

Table 1: Fifteen causal conditions for both experiments

| $\mathrm{Q}(\mathrm{e} \mid \mathrm{c})$ | $\mathrm{Q}(\mathrm{e} \mid \neg \mathrm{c})$ | $\|\Delta \mathrm{Q}\|$ | Causal Power |  |
| :---: | :---: | :---: | :---: | :---: |
|  | 1.00 | 0.00 | - | 0.00 |
| 1.00 | 1.09 | Pre $^{1}$ |  |  |
| 0.75 | 0.75 | 0.00 | 0.00 | 0.00 |
| 0.50 | 0.50 | 0.00 | 0.00 | 0.00 |
| 0.25 | 0.25 | 0.00 | 0.00 | 0.00 |
| 0.00 | 0.00 | 0.00 | 0.00 | - |
| 1.00 | 0.75 | 0.25 | 1.00 | 0.25 |
| 0.75 | 0.50 | 0.25 | 0.50 | 0.33 |
| 0.50 | 0.25 | 0.25 | 0.33 | 0.50 |
| 0.25 | 0.00 | 0.25 | 0.25 | 1.00 |
| 1.00 | 0.50 | 0.50 | 1.00 | 0.50 |
| 0.75 | 0.25 | 0.50 | 0.67 | 0.67 |
| 0.50 | 0.00 | 0.50 | 0.50 | 1.00 |
| 1.00 | 0.25 | 0.75 | 1.00 | 0.75 |
| 0.75 | 0.00 | 0.75 | 0.75 | 1.00 |
| 1.00 | 0.00 | 1.00 | 1.00 | 1.00 |
| ${ }^{1}$ Values of $\mathrm{Q}(\mathrm{e} \mid \mathrm{c})$ and $\mathrm{Q}(\mathrm{e} \mid \neg \mathrm{c})$ are switched in preventive |  |  |  |  |

After going through the cover story, participants were presented with 15 visual stimuli to correspond to the 15 conditions in a random order (see Figure 1). They then had to judge how strong the cause generates/prevents the effect by answering two counterfactual questions - one at a time. The two counterfactual questions were presented to correspond to two limits - a limit that was consistent with the cover story, and a higher limit.


Figure 1: Sample Stimuli from the generative component

$|\Delta Q|: \bullet: 0.00$ ■: $0.25 \star: 0.50 *: 0.75 *: 1.00$ Difference:- Proportion:---

Figure 2: Power PC and $\Delta \mathrm{P}$ Predictions of Causal Ratings.
The counterfactual question for the generative experiment was: Now imagine a new patient who does not have any skin rash. If we applied this cream on the back of this patient to cover an area of $10 \mathrm{~cm}^{2}$, how big would the area of skin rash be on this patient? The exact same sentence was used for the second question except that the area (i.e. the limit) was changed to $50 \mathrm{~cm}^{2}$.

The counterfactual question for the preventive experiment was: Now imagine a new allergy patient suffering from a rash of $10 \mathrm{~cm}^{2}$. If we apply the ointment, how large would the area of rash be? Similarly, the second question was exactly the same except for substituting the area with 50 $\mathrm{cm}^{2}$. Participants provided numerical responses using the keyboard.

## Predictions

Figure 2 shows causal strength prediction plots for the 15 conditions, derived from difference based $(\Delta \mathrm{P})$ and proportional (Power PC) approaches (solid and dashed lines respectively). Causal conditions that have identical $\Delta \mathrm{Q}$ values are linked together and plotted against the base-rate.

To allow comparisons both with previous literature, and across the two limit scenarios, these predictions were plotted with respect to the value of the limits tested. Since the maximum area of skin rash is $10 \mathrm{~cm}^{2}$ in the consistent-limit scenario, the maximum power in the prediction has been set to 10 as well. In contrast, in the scaled-up limit scenario, the maximum power in the prediction has been set to be at 50 to match up with the maximum rash area of $50 \mathrm{~cm}^{2}$.

Participants' judgments were analogously converted: For instance, an area judgment of $10 \mathrm{~cm}^{2}$ in the consistent-limit scenario was converted into a causal rating of 10 in the generative, and a causal rating of 0 in the preventive experiment.

$|\Delta Q|: \bullet: 0.00$ ■: 0.25 *:0.50 *:0.75 *: 1.00

Figure 3: Medians of Counterfactual Responses.
More specifically, we subtracted the counterfactual response given by the participant from the relevant upper limit. This conversion was made on the judgments to reflect that an increase of affected skin area would indicate an increase of causal power when considering generative causes, while larger predicted skin areas would indicate weaker causal powers when considering preventive relations.

The absolute difference approach predicts that causal strength is unaffected by increments of base-rate, and that causal ratings vary only as a function of $\Delta \mathrm{P}$. Furthermore, a strict interpretation of difference approach would suggest that the same difference is then applied to a different context, involving a higher upper limit. Consequently, prediction plots for the difference approach remain within the range of 0 to 10 , across both the consistent-limit and scaled-up limit scenarios.

The proportional approach, on the other hand, predicts a consistent influence of base-rate onto causal ratings in both limit cases, which varies depending on whether generative or preventive powers are assessed. Despite having the same non-zero difference values (i.e. $\Delta \mathrm{P}$ ), in the generative scenario causal ratings should increase as the base rate increases. The reverse pattern is predicted in preventive scenarios. These influences of base rate, however, are not predicted for when the difference value is zero, and causal ratings should remain at zero for both generative and preventive cases. In addition, the proportional approach also dictates that counterfactual causal ratings are scaled up in line with a higher limit.

## Results

Kolmogorov-Smirnov test showed that judgments were nonnormally distributed. Consequently, Figure 3 plots median
judgments, and statistical analysis was based on nonparametric tests.

A qualitative inspection of the generative results in Figure 3 suggests that judgments correspond more to difference than proportional approach predictions. In the consistentlimit scenario, apart from the conditions involving $\Delta \mathrm{Q}=$ 0.25 , the judgments for other $\Delta \mathrm{Q}$ values are relatively flat at the predicted difference values, suggesting a minimal influence of base-rate.

This minimal influence of base-rate is also evident on causal judgments in the scaled-up limit scenario. In this scenario, judgements from conditions involving identical values of $\Delta \mathrm{Q}$ are also relatively consistent at the difference values, even though a small indication of a positive trend is observed in the $\Delta \mathrm{Q}=0.25$ case. Even though the minimal influence of base-rate influence is in line with a difference account, generative judgments violate its other significant property: They vary from 0 up to 50 , instead of 10 . We will discuss this in the next section.

Qualitatively inspecting the preventive results in Figure 3 suggests they fit well with proportional approach. In both limit scenarios, the contingent cases indicate the influence of the base-rates. Instead of remaining constant at the difference values, the judgments decrease as the base-rate increases. Moreover, for the non-contingent cases, judgments also follow proportional predictions, in that they stay at zero despite a change of the base-rate. Even though there is an indication of a non-normative trend in the consistent-limit scenario when $\Delta \mathrm{Q}=0.25$, in general, the preventive judgments seem to have followed proportional predictions, both with a consistent and inconsistent limit.

Statistical Analysis (Generative) Nonparametric Friedman's ANOVA was used to determine the main effect of the base-rate for every $\Delta \mathrm{Q}$ value.

Analysis of ratings from the consistent-limit case found a significant effect of base-rate when $\Delta \mathrm{Q}=0, X_{\mathrm{F}}^{2}(14)=$ $14.750, p<.05$ and $\Delta \mathrm{Q}=0.25, X_{\mathrm{F}}^{2}(14)=10.545, p<.05$. The analysis does not show any significant effect of baserate when $\Delta \mathrm{Q}=0.50, X_{\mathrm{F}}^{2}(14)=0.347, p>.05$ and $\Delta \mathrm{Q}=$ $0.75, X_{\mathrm{F}}^{2}(14)=1.190, p>.05$.

Unlike in the consistent-limit case, analysis of the scaledup limit scenario shows a significant effect of base-rate only when $\Delta \mathrm{Q}=0.25, X_{\mathrm{F}}^{2}(14)=7.978, p<.05$. No significant effects of base-rate are found when $\Delta \mathrm{Q}=0, X_{\mathrm{F}}^{2}$ (14) $=$ $6.681, p>.005 ; \Delta \mathrm{Q}=0.50, X_{\mathrm{F}}^{2}(14)=1.357, p>.005$; and $\Delta \mathrm{Q}=0.75, X_{\mathrm{F}}^{2}(14)=1.087, p>.005$.

Surprisingly, the statistical test indicates an effect of the base rate in the non-contingent case of consistent-limit scenario, despite an observation of a flat line in Figure 3. Inspection of the data distribution in these conditions (Figure 4) reveals three noteworthy points: i) the modal response is 0 in all cases, ii) a minority of participants give a non-normative non-zero response, iii) this minority of participants appears to exhibit an outcome density bias (Buehner, Cheng, \& Clifford, 2003). Because the Friedman

Test ignores ties, the significant result in $\Delta \mathrm{Q}=0$ condition is thus driven by this minority of participants.

Statistical Analysis (Preventive) In the consistent-limit scenario, no significant effect of base-rate was found when $\Delta \mathrm{Q}=0, X_{\mathrm{F}}{ }^{2}(14)=4.500, p>.05$. However, significant effects of base-rate were obtained when $\Delta \mathrm{Q}=0.25, X_{\mathrm{F}}{ }^{2}$ (14) $=57.854, p<.05 ; \Delta \mathrm{Q}=0.50, X_{\mathrm{F}}^{2}(14)=15.892, p<.05$; and $\Delta \mathrm{Q}=0.75$ as well, $X_{\mathrm{F}}^{2}(14)=9.783, p<.05$.

Similar trends were observed in the scaled-up limit scenario. The analysis shows no significant effect of baserate when $\Delta \mathrm{Q}=0, X_{\mathrm{F}}^{2}(14)=1.222, p>.05$. Again, significant base-rate effects are found when $\Delta \mathrm{Q}=0.25, X_{\mathrm{F}}^{2}$ $(14)=27.931, p<.05 ; \Delta \mathrm{Q}=0.50, X_{\mathrm{F}}^{2}(14)=12.302, p<$ $.05 ; \Delta \mathrm{Q}=0.75, X_{\mathrm{F}}^{2}(14)=3.846, p<.05$.

As with the generative scenario, non-contingent conditions uniformly elicited a median and modal response of zero. While there was also a minority of participants who deviated from this normative assessment, judgments from these participants did not display any systematic patterns. More specifically, unlike in the generative scenario, there was no evidence of an outcome density bias, even in the minority of non-normative judgments.


Figure 4: Judgment Distributions of non-contingent Conditions $(\Delta \mathrm{Q}=0)$ in the consistent-limit scenario

## Discussion

Overall, our results seem to suggest that when people reason about continuous outcomes, they do so within a proportional framework, if the context is one of preventive causation, i.e. the goal is to reduce the outcome magnitude. However, if the context involves increasing the outcome magnitude (generative causation), people seem to focus on the difference the cause makes, without normalizing this difference to an upper limit, even when the task clearly implies such a limit. Interestingly, people then do not adhere to the absolute difference a cause makes in a given context, but instead scale up this difference, where appropriate, in different scenarios.

For instance, in the condition when $\mathrm{Q}(\mathrm{e} \mid \mathrm{c})=1.00$ and $\mathrm{Q}(\mathrm{e} \mid \neg \mathrm{c})=0.25$, participants learned that a skin area of 2.5 $\mathrm{cm}^{2}$ was covered with of rash before the application of the cream, and that applying the cream to an area of $10 \mathrm{~cm}^{2}$ resulted in that entire area breaking out with rash. They considered the difference the cream made, and concluded that its application increases the area of rash by $7.5 \mathrm{~cm}^{2}$ when applied to $10 \mathrm{~cm}^{2}$ of skin of a patient who does not yet suffer from rash. Had they taken the proportional approach, they would have concluded that this cream is maximally
effective in producing rash, and applying it to an area of $10 \mathrm{~cm}^{2}$ of healthy skin would lead it all of it to break out with rash. When they were asked to transfer their knowledge to a different scenario, where the cream was applied to $50 \mathrm{~cm}^{2}$ of healthy skin, they took the difference ( $7.5 \mathrm{~cm}^{2}$ ) and scaled it up to this new area, concluding that $33 \mathrm{~cm}^{2}$ (i.e. nearly $37.5 \mathrm{~cm}^{2}$ ) of the $50 \mathrm{~cm}^{2}$ will break out with rash.

Inspection of Figure 3 shows that participants were relatively consistent in scaling up their counterfactual judgments across all the generative conditions: a factor of approximately 5 emerges. This suggests that participants indeed scaled up their judgments from one context to the other, rather than merely considering the difference, as suggested by a strict interpretation of a difference-based approach. It appears then that people were aware of the upper limit we imposed on our scenarios, and scaled their judgments up accordingly in both preventive and generative situations. However, the judgments they formed were based on proportions only for preventive contexts, and on differences in generative contexts.

One tempting conclusion might be that perhaps our generative cover story might simply have failed to instill a clear sense of an upper limit in the learning phase, despite our best efforts to do so. After all, even when cream is applied to only to $10 \mathrm{~cm}^{2}$, it is still feasible for a rash to occur in a larger area than that. In contrast, the preventive scenarios were not hampered this way - the natural upper limit of preventive causation is always 0 : No treatment could reduce rash to less than an area of $0 \mathrm{~cm}^{2}$. However, we have conducted studies with other generative contexts, involving continuous outcome magnitudes that definitely do have clear and unambiguous upper limits (such as relative humidity in the atmosphere), and the results mirror those reported here: People largely adopt a difference-based approach when evaluating generative causal influence.

## Conclusions

The work reported here represents the beginning of a quest to chart the waters of continuous causal inference. We have taken a cautious approach and created situations that are structurally identical to conventional binary probabilistic causal inference. We knew that doing so would limit the ecological validity of our results. After all, most causes are continuous variables themselves, influencing continuous outcome magnitudes. However, our goal here was a proof of concept: We wanted to measure people's inferences about causal change to continuous outcomes under ideal conditions and with clear explicit upper limits (which are not always present in the world). If under these conditions, inferences followed patterns similar to those observed in probabilistic causal inference, this might suggests that a fruitful avenue to pursue might be to try and adapt theories and models from binary probabilistic causal inference to inference about continuous causation.

Tentatively, we would conclude that people's inference patterns do correspond to what we know about probabilistic causal inference. Deviations from normative models are found frequently also in probabilistic causal inference (e.g. Lober \& Shanks, 2000), although sometimes such deviations seem to reflect ambiguities in the task demands. And indeed perhaps the non-normative results of our generative experiment may be due to such ambiguities. We are currently addressing this with follow-up studies. For example, we have not considered the reliability of the information on which participants base their judgments. Bayesian models of causal inference (e.g. Griffiths \& Tenenbaum, 2005) consider both the strength of a causal relation (as indexed by power PC), as well as the reliability of the information (as indexed by the sample size, or the effective sample size). For simplicity, and to ensure the one-to-one mapping to probabilistic causation, our study involved only single entities (i.e. one patient per treatment).

In future work, we hope to consider not only multiple instances of continuous outcome change from the same cause, but also to begin working with causes that are in themselves continuous variables.

## Acknowledgments

This work has been carried out as part of a PhD by the first author, supervised by the second author. We acknowledge funding from Malaysian Institute of Road Safety Research (MIROS) to support this study.

## References

Buehner, M. J., Cheng, P. W., \& Clifford, D. (2003). From covariation to causation: A test of the assumption of causal power. Journal of Experimental Psychology: Learning, Memory, and Cognition, 29(6), 1119.
Cheng, P. W. (1997). From covariation to causation: A causal power theory. Psychological Review, 104, 367405.

Griffiths A, Tenenbaum J.B. (2005). Structure and strength in causal induction. Cognitive Psychology. 51, 334-84.
Hume, D. (1888). Hume's treatise of human nature (L. A. Selby-Bigge, Ed.). Oxford, UK: Clarendon Press. (Original work published 1739).
Jenkins, H., \& Ward, W. (1965). Judgment of contingencies between responses and outcomes. Psychological Monographs, 7, 1-17.
Lober, K., \& Shanks, D. R. (2000). Is causal induction based on causal power? Critique of Cheng (1997). Psychological Review, 107(1), 195-212.
White, P. A. (2001). Causal judgment about relations between multilevel variables. Journal of Experimental Psychology: Learning, Memory, \& Cognition, 27, 499513.

Wu, M., \& Cheng, P. W. (1999). Why causation need not follow from statistical association: Boundary conditions for the evaluation of gener- ative and preventive causal powers. Psychological Science, 10, 92-9.

# A Metacognitive Stopping Rule for Problem Solving 

Rakefet Ackerman (Ackerman@ie.technion.ac.il)<br>Faculty of Industrial Engineering and Management, Technion - Israel Institute of Technology<br>Technion City, Haifa, 3200003 Israel


#### Abstract

Although people expect to improve by investing effort in solving a problem, several studies have found negative timeconfidence correlations in various problem-solving tasks. The present study employed the metacognitive approach to illuminate why, despite lengthy thinking, people provide solutions in which they have only low confidence. According to the proposed Diminishing Criterion Model (DCM), as people invest longer in a problem, their confidence in their solution increases in a goal-driven manner, in accordance with the common belief. Nevertheless, the process ends up with a negative time-confidence correlation, because people tend to find lower confidence levels as satisfactory as they invest longer in solving a problem, reflecting a compromise in their stopping criterion. The hypotheses derived from the DCM were supported with two problem types. Even when the participants were allowed to submit a "don't know" response, they still provided low confidence solutions after lengthy thinking, suggesting that they found these low confidence solutions to be satisfactory. The study offers reconciliation between beliefs and empirical findings and explains why people end up offering solutions with low confidence rather than continuing attempts to improve or admitting failure (via the "don't know" option).


Keywords: Metacognition; problem solving; dual-process theory; stopping rule; time allocation.

## Introduction

Solving a problem requires representations of the relevant components, rules (or constraints), and goal, followed by a sequence of inferences or calculations. However, beyond the cognitive process per se, solving a problem also involves regulation of the cognitive effort. Regulation of effort is at the heart of metacognitive theories. According to this approach, to achieve a cognitive goal, people constantly judge, or monitor, the state of their performance relative to the goal they pursue and decide whether to continue to invest further effort or cease (Koriat, Ma'ayan, \& Nussinson, 2006; Nelson \& Narens, 1990). The metacognitive approach is commonly used for learning research, mainly memorizing, but as yet is rarely employed for problem solving. Although metacognitive considerations have been mentioned in some discussions (e.g., Payne \& Duggan, 2011), establishing metacognitive monitoring as a causal link in regulating problem solving has only started to emerge, in particular with regard to dual-process theories.

According to the dual-process theories (Kahneman, 2003), System 1 or Type 1 processes (T1) are responsible for suggesting a quick solution that comes to mind based on default procedures. System 2 or Type 2 processes (T2)
execute more deliberate and lengthy analytic reasoning. However, Evans (2009) identified a third type of processes (T3). These T3 processes are responsible for (a) identifying the need for T 2 intervention, and (b) examining whether a given model is satisfactory (see also the reflective mind suggested by Stanovich, 2009a). Thompson (2009) proposed that metacognitive processes underlie identifying the need for T2 intervention. Indeed, Thompson and her colleagues (Thompson, Prowse Turner, \& Pennycook, 2011; Thompson et al., in press) asked participants to provide an initial answer and their Feeling of Rightness (FOR) about it. Subsequently, they were allowed to reconsider their answer. As expected, lower FORs were associated with more reconsideration time and with a higher likelihood of providing an alternative for the initial answer. These findings support the role of metacognitive monitoring as bridging T1 and T2 processes. This relates to the first aspect mentioned by Evans (2009). The present paper deals with the second aspect he mentioned; namely, examining whether a given model is satisfactory for deciding whether to stop investing effort in a particular problem.

## Metacognitive Stopping Rules

In the experimental examinations of metacognitive regulation of memorizing, the fact that people invest more time in studying the more difficult items has led to the development of the discrepancy reduction model (Nelson \& Narens, 1990). According to this model, people set a target level according to their motivation for the given scenario and study each item until they consider their knowledge to be satisfactory. This process seems to also be applied to a problem-solving task: people set a criterion for their confidence level, and continue to search for better solutions until they judge their chance for success to be satisfactory (Evans, 2006). Thus, for both memorizing words and solving problems, the discrepancy reduction model suggests that lengthy processing positively correlates with the chance for success. In line with this model, Koriat et al. (2006) associated goal-driven effort with a positive time-judgment correlation.

Considering the final form of time-judgment correlation, if people progress in their goal pursue until they reach the judgment level they consider as satisfactory, we would expect to find no correlation between time and judgment. This is because people are expected to stop investing effort when their judgment passes the preset goal regardless of the time it takes to reach this perceived knowledge level. However, studies of both memorizing and problem solving
repeatedly found that negative correlations dominated timejudgment relationships (e.g., Begg, Duft, Lalonde, Melnick, \& Sanvito, 1989; Koriat et al., 2006). In particular, in problem solving, this negative time-confidence correlation was found even when reaction times were not valid as predictors of success (Ackerman \& Zalmanov, 2012; Topolinski \& Reber, 2010). In these studies the participants were more confident when they provided the solutions quickly than when they provided the solutions after lengthy thinking, regardless of the actual chance for success. These findings are puzzling: Why do people stop investing effort when they knowingly provide solutions with low confidence? Is investment of time perceived as a waste of time, with no progress in the assessed chance for success?

Previous explanations for the consistent negative timejudgment correlation were based on bottom-up fluency (Koriat, Ackerman, Adiv, Lockl, \& Schneider, in press; Koriat et al., 2006), Ackerman and Zalmanov (2012) included. In light of the goal-oriented nature of problemsolving tasks (Evans, 2006), the present study offers a topdown explanation for the findings. According to the proposed Diminishing Criterion Model (DCM), as people invest longer in a problem, their confidence in the currently considered solution option increases in a goal-driven manner, aiming to improve the chance for success, as is also derived from the discrepancy reduction model. However, the stopping criterion does not remain constant along the solving process, but diminishes, reflecting an increased compromise as more time is invested. That is, if an immediate solution option comes to mind with high confidence, higher than the initial stopping criterion, this solution would be provided. If the confidence regarding the initial solution is lower than the criterion, more consideration time is invested, until reaching a satisfying level of confidence. Importantly, a confidence level which may not satisfy the solver regarding a quickly produced solution may satisfy him or her after lengthy consideration.

But what if the confidence after lengthy thinking remains very low and no way to find a better solution is found? In this case, people may prefer to respond with "don't know", stemming from their desire to provide solutions with reasonable confidence. In this case, a "don't know" response may be more socially acceptable than a solution accompanied by a very low confidence (Ackerman \& Goldsmith, 2008). The question is whether the "don't know" option would eliminate the negative time-confidence slope and lead respondents to provide only high confidence solutions. The prediction by the DCM is that it would not, because if a great deal of time is already invested, people compromise and refer to quite low confidence levels as satisfactory.

Two hypotheses derived from the DCM were examined in two experiments, one with regular problems and the other with misleading problems often used in studies of the dualprocess approach. The first hypothesis was that judgments are positively correlated with time while the final timeconfidence correlation is negative. The second hypothesis
was that the opportunity to respond with "don't know" would not eliminate this pattern of results.

## Experiment 1

The task in Experiment 1 was the Compound Remote Associate (CRA) test. In this test, participants are presented with a word triplet and their task is to find a fourth word that forms a compound word or two-word phrase with each cue word separately. In an attempt to solve these problems, immediate associations for each word are expected to come up (Wiley, 1998). However, an association that fits only one or two of the cue words does not satisfy the requirements. For example, for the triplet PINE/CRAB/SAUCE, the word PINE might initially elicit PINECONE rather than the correct PINEAPPLE. Recognition that the initial solution option does not fit should trigger a search for a better solution (Thompson et al., 2011; Thompson et al., in press). In Ackerman and Zalmanov (2012) these problems yielded a strong negative time-confidence correlation.

In the present study there were two groups. The intermediate ratings group was asked to provide ongoing confidence ratings regarding the solution options they considered at each point in time (see also Ackerman \& Goldsmith, 2011; Metcalfe \& Weibe, 1987; Vernon \& Usher, 2003). The "don't know" group provided intermediate ratings as well, but also had the option to respond with "don't know".

## Method

Participants. Forty-four undergraduates participated in the experiment for course credit or for payment $\left(\mathrm{M}_{\text {age }}=24.8\right.$; $50 \%$ females). They were randomly assigned to working with or without the "don't know" option, with 22 participants in each group.

Materials. Thirty-four CRA problems were used. Two problems were used for demonstration and two for selfpractice. Pretesting verified that all problems were solvable by the target population.

Procedure. The experiment was conducted on two to eight participants in parallel, in a small computers lab. The instruction booklet detailed the procedure, explained what constituted a valid solution, and illustrated the procedure using two problems. Pressing the "Start" button brought up a problem. Respondents had to type the solution and press the "Continue" button. Response time was measured from when participants pressed "Start" to when they pressed "Continue". This exposed the question, "How confident are you that your answer is correct?", and a horizontal scale ( $0 \%-100 \%$ ). Pressing the "Next" button cleared the screen for the next problem.

The participants were asked to report on intermediate confidence ratings interspersed with solving each problem. The ends of each scale were marked as "I still have no idea", and "I'm sure I found it". The first scale, appeared
three seconds after the problem's presentation. Later on, an additional scale appeared every 15 seconds and the previous scale became inactive, even if no rating was entered. This way the repeated request to enter a rating was clearly noticeable. The screen could present up to five intermediate scales. The participants could enter the answer at any time, rate their final confidence, and move on to the next problem. The times for entering the intermediate confidence ratings were documented. The only difference for the "don't know" group was that adjacent to the space for answer entry, there was a "don't know" button. Pressing this button deactivated the confidence rating scale.

After demonstration with two problems, the two other practice problems appeared first, and the rest were randomly ordered for each participant. The session lasted 30 minutes.

## Results and Discussion

The participants provided meaningful solution words (rather than "XXX", for example) for $97 \%$ of the problems. Overall, the results were highly similar to those of the group reported in Ackerman and Zalmanov (2012), which was drawn from the same population and solved the problems without intermediate ratings. In the "don't know" group, 19 participants of 22 used the "don't know" option. Percent correct (with "don't know": $M=56.5 \%, S D=20.3$; without a "don't know" option: $M=48.0 \%, S D=16.8$ ) and confidence ratings were somewhat higher with the "don't know" option than without it, but the differences were not significant; both were $p s>.13$. The mean response time was shorter with the "don't know" option (with "don't know": $M$ $=29.0 \mathrm{sec}$., $S D=13.1$; without a "don't know" option: $M=$ $41.6 \mathrm{sec} ., S D=12.6), t(42)=3.27, p<.01$. This finding may indicate that the "don't know" option allowed participants to avoid providing the results of their lengthy solving processes. Indeed, the "don't know" responses ( $M=$ 56.9, $S D=26.6$ ) were provided after more time than the provided solutions, $t(18)=5.90, p<.0001$. This finding suggests that the participants provided the "don't know" response after deliberation, rather than for moving quickly to the next problem.

To examine the ongoing progress of the confidence ratings, the data was split for each participant for his/her own quarters of final response times, with seven or eight problems in each quarter. The points on the black lines in Figure 1 represent the mean final times and confidence for each quarter, with (panel A) and without (panel B) the "don't know" option. A two-way Analysis of Variance (ANOVA) examining the effects of the Group (2) and Quarter (1-4) on final confidence ratings, revealed only the main effect of the quarter, $F(3,126)=136.49, M S E=$ 218.32, $p<.0001$. No difference was found among the groups, $F<1$, and the interaction was not significant, $F(3$, 126) $=1.64, M S E=218.32, p=.18$. Thus, confidence ratings were higher for the quickly provided solutions than for the lengthy solutions, and this pattern did not differ among the groups.

The progress of the problem-solving process exposed by the intermediate confidence ratings is also plotted in Figure 1. Because there was no data on all points for all participants, we used the initial confidence (by the first intermediate scale) and final confidence to statistically examine the progress in the ratings. The analysis was based on participants who provided initial confidence under all four quarters $(N=26,59 \%)$. A mixed three-way ANOVA Group (2) $\times$ Quarter (1-4) $\times$ Rating (initial vs. final confidence) yielded no main effect of the group, $F<1$. The main effect of the quarter was significant, $F(3,72)=90.92$, $M S E=352.64, p<.0001$, reflecting that the ratings fell from the first to the fourth quarters. The main effect of the rating was also significant, $F(1,24)=90.78, M S E=404.03$, $p<.0001$, supporting the increase from the initial to the final confidence ratings. Importantly, the triple interaction was insignificant, $F<1$, suggesting a similar pattern of results with and without the "don't know" option.


Figure 1: Experiment 1 - The intermediate and final confidence ratings on the timeline of solving the problems, divided by final response time quarters (1-4). Each panel presents the results of one group.

Overall, the results of Experiment 1 support the hypotheses derived from the DCM. There was a positive relationship between the time in which each rating was provided and the progress of the confidence ratings and a negative relationship between the total invested time and final confidence ratings. Importantly, this was the case even with the "don't know" option, which suggests that the participants found low confidence solutions provided after lengthy thinking as satisfactory for that point in time. It is also clear from Figure 1 that confidence levels that were not considered satisfactory in initial stages of the problemsolving process (e.g., the mean of the FOR ratings in the second quarter in panel A , which is 33 ), were provided if a similar level of confidence was reached after lengthy deliberation (e.g., the mean final confidence rating at the fourth quarter in panel A, which is also 33).

## Experiment 2

Misleading problems are commonly used in the literature related to dual-process theories to differentiate between the fast intuitive (T1) solutions and the results of more deliberate processing (T2). For example: "A bat and a ball cost $\$ 1.10$ in total. The bat costs $\$ 1$ more than the ball. How
much does the ball cost? $\qquad$ cents" (Kahneman, 2003). The immediate solution that comes to mind is 10 cents, while the correct solution is 5 cents. From a metacognitive point of view, these problems allow dissociation between the confidence and accuracy in their relationship with response time, since the very first solutions tend to be accompanied by high confidence but a low chance for being correct, in particular when presented in an open-ended format (Ackerman \& Zalmanov, 2012).

What should we expect with regard to final confidence in solutions provided after lengthy thinking? In cases where the respondent reaches the correct solution, he or she may be aware of the successful processing and be highly confident of the found solution. High confidence after deliberate processing can also be expected to accompany wrong solutions in cases such as over-generalized rules without appropriate exceptions, or investing effort in finding support for the initial and wrong solution (Stanovich, 2009a). Indeed, Ackerman and Zalmanov (2012) found higher confidence ratings regarding lengthy solutions with misleading problems than regarding CRA problems. However, despite this finding, a negative correlation between time and confidence was found even with the misleading problems. This might indicate that even with these problems, people see relatively low confidence levels as satisfying after lengthy thinking, as suggested here. To examine this possibility, Experiment 2 examined whether the "don't know" option allows the participants to avoid the low confidence solutions, with the hypothesis that it will not. Like in Experiment 1, all participants provided intermediate confidence ratings. One group worked with and one without the "don't know" option.

## Method

Participants. The 40 participants were drawn from the same population $\left(\mathrm{M}_{\text {age }}=25.2 ; 36 \%\right.$ females $)$. The participants were randomly assigned to the "don't know" conditions, with 20 participants in each group.

Materials. The problems used by Ackerman and Zalmanov (2012) were used for this experiment. They included twelve experimental problems and a practice problem for demonstrating the procedure. The experimental problems included the three problems used by Frederick (2005; the bat and ball, water lilies cover half a lake, and machines that produce widgets at a certain rate), the drinks version of Wason's selection task (Beaman, 2002), the A-is-looking-at-B problem (Stanovich, 2009b), and a conditional probability problem (Leron \& Hazzan, 2009). The other problems were misleading problems adapted from preparation booklets for the Graduate Management Admission Test (GMAT).

Procedure. The procedure was highly similar to that used in Experiment 1. The practice problem appeared first, and the rest were randomly ordered for each participant.

## Results and Discussion

The participants provided meaningful solution words (rather than "XXX", for example) for all the problems. In the "don't know" group, only six participants of 20 utilized the "don't know" option. As in Experiment 1, percent correct (with "don't know" option: $M=47.2 \%, S D=16.4$; without the "don't know" option: $M=43.9 \%, S D=15.5$ ) and confidence ratings were equivalent in both groups, both $t \mathrm{~s}<$ 1. In this case, no difference was found also for response time, $t<1$. Like in Experiment 1, the results were highly similar to those found by Ackerman and Zalmanov (2012), where there were no intermediate confidence ratings.

In this experiment, there were only 12 problems, so they were divided into thirds rather than quarters, with four problems in each third. As can be seen in Figure 2, the overall pattern of results remained with (Panel A) and without (Panel B) the "don't know" option, although confidence levels were higher than in Experiment 1.


Figure 2: Experiment 2 - The initial feeling of rightness (FOR), and intermediate and final confidence ratings on the timeline of solving the problems, divided by final response latency thirds (1-3). Each panel presents the results of one group.

A two-way ANOVA examining the effects of the Group (2) and Third (1-3) on final confidence ratings, revealed only a main effect of the third, $F(2,76)=60.37, M S E=$ $67.82, p<.0001$. No difference was found among the groups, $F<1$, and no interactive effect, $F(2,76)=1.73$, $M S E=67.82, p=.18$. Thus, confidence ratings were higher for the quickly provided solutions than for the lengthy solutions, but this pattern did not differ among the groups. A mixed three-way ANOVA Group (2) $\times$ Third (1-3) $\times$ Rating (FOR vs. Final confidence) was based on participants who provided FORs under all thirds $(N=20,50 \%)$. The main effect of the group was not significant, $F(1,18)=2.29, M S E$ $=1399.72, p=.15$. The main effect of the third was significant, $F(2,36)=29.25, M S E=163.62, p<.0001$, reflecting that the ratings dropped from the quickly provided to the slowly provided solutions. The main effect of the rating was also significant, $F(1,18)=95.80, M S E=948.31$, $p<.0001$, supporting the increase from the initial FORs to the final confidence ratings. The triple interaction was again insignificant, $F(2,36)=1.84, M S E=180.92, p=.17$ suggesting a similar pattern of results with and without the "don't know" option.

The results of Experiment 2 support the two hypotheses derived from the DCM as well. This experiment generalizes the results of Experiment 1 with a different type of problems, in which one may expect to find high confidence ratings after lengthy thinking. Even with these problems, the time-confidence relationship is neither positive nor flat. It is consistently negative, even when participants could avoid low confidence solutions by utilizing the "don't know" option.

## General Discussion

The motivation for the present study stemmed from the puzzling inconsistency between the goal-driven nature of the problem-solving task - which leads to the expectancy of positive or no correlation between time and confidence - and the empirical findings of persistent negative correlation between them, even when people are free to regulate their solving time (Ackerman \& Zalmanov, 2012; Koriat et al., 2006). The proposed DCM suggests that the cognitive process indeed progresses in a goal-driven manner, with a positive correlation between time and confidence. It also suggests that people stop investing effort when their metacognitive monitoring passes their stopping criterion. Until this point the process accords with the wellknown discrepancy reduction models (Nelson \& Narens, 1990). The unique characteristic of the proposed DCM is the suggestion that the negative correlation stems from the willingness of people to compromise on the satisfactory level of their chance for success. These predictions were supported by the two experiments, with two task types: nonmisleading and misleading problems. It was found that although the final time-confidence relationship is negative, the process progresses with a positive correlation between them. The "don't know" procedure was used to ensure that the negative correlation does not stem from the desire to move on to the next problem, even if a satisfactory solution was not yet found. The results suggest that people find the relatively low confidence they experience after lengthy thinking to be satisfactory, even though the same confidence levels were not acceptable if reached earlier in the process.

The present study suggests a difference between the stopping rules for T 1 and for T 2 . Thompson and her colleagues (Thompson et al., 2011; Thompson et al., in press) suggested that FOR is the basis for the stopping rule for T1 and for triggering T2. If FOR is high enough, people provide the first answer that comes to mind. Otherwise they activate T2. The present study extends this idea to the decision to stop T2. While time allocation for T1 was explained to be based on fluency in which the first solution option comes to mind, in a bottom-up manner, the stopping rule for T 2 is explained here as stemming from a goaldriven, top-down, effort investment. Importantly, this does not rule out fluency effects on final confidence ratings as well, but suggests that the goal-driven decision dominates the process.

To conclude, metacognitive studies traditionally focus on memorizing word lists. Investigating more complex tasks
brings to the fore additional factors that may have broader ecological validity. The present study evolved from considering problem-solving tasks, which are generally understudied from the metacognitive point of view, and which highlight puzzling aspects of time investment and its relationship to metacognitive regulation. By proposing the DCM, this paper aims to shed light on the processes that lead people to end up with low confidence in their success, even when they can potentially avoid it by continuing improvement attempts or admitting failure.

## Acknowledgments

The study was supported by a grant from the Israel Foundation Trustees (2011-2013).

## References

Ackerman, R., \& Goldsmith, M. (2008). Control over grain size in memory reporting-With and without satisficing knowledge. Journal of Experimental Psychology: Learning, Memory, and Cognition, 34(5), 1224-1245.
Ackerman, R., \& Goldsmith, M. (2011). Metacognitive regulation of text learning: On screen versus on paper. Journal of Experimental Psychology: Applied, 17(1), 1832.

Ackerman, R., \& Zalmanov, H. (2012). The persistence of the fluency-confidence association in problem solving. Psychonomic Bulletin \& Review, 19(6), 1189-1192.
Beaman, C. P. (2002). Why are we good at detecting cheaters? A reply to Fodor. Cognition, 83, 215-220.
Begg, I., Duft, S., Lalonde, P., Melnick, R., \& Sanvito, J. (1989). Memory predictions are based on ease of processing. Journal of Memory and Language, 28(5), 610-632.
Evans, J. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. Psychonomic Bulletin \& Review, 13(3), 378-395.
Evans, J. (2009). How many dual-process theories do we need? One, two, or many. In J. Evans \& K. Frankish (Eds.), In two minds: Dual processes and beyond. Oxford, UK: Oxford University Press.
Frederick, S. (2005). Cognitive reflection and decision making. Journal of Economic Perspectives, 19(4), 2542.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. American Economic Review, 93(5), 1449-1475.
Koriat, A., Ackerman, R., Adiv, S., Lockl, K., \& Schneider, W. (in press). The effects of goal-driven and data-driven regulation on metacognitive monitoring during learning: A developmental perspective. Journal of Experimental Psychology: General.
Koriat, A., Ma'ayan, H., \& Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. Journal of Experimental Psychology: General, 135(1), 36-68.

Leron, U., \& Hazzan, O. (2009). Intuitive vs analytical thinking: four perspectives. Educational Studies in Mathematics, 71(3), 263-278.
Metcalfe, J., \& Weibe, D. (1987). Metacognition in insight and noninsight problem solving. Memory \& Cognition, 15, 238-246.
Nelson, T. O., \& Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. Bower (Ed.), The Psychology of learning and motivation: Advances in research and theory (Vol. 26). San Diego, CA: Academic Press.
Payne, S. J., \& Duggan, G. B. (2011). Giving up problem solving. Memory \& Cognition, 39(5), 902-913.
Stanovich, K. E. (2009a). Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a triprocess theory? In J. Evans \& K. Frankish (Eds.), In two minds: Dual processes and beyond. Oxford, UK: Oxford University Press.
Stanovich, K. E. (2009b). Rational and Irrational Thought: The Thinking that IQ Tests Miss. Scientific American Mind, 20(6), 34-39.
Thompson, V. A. (2009). Dual process theories: A metacognitive perspective. In J. Evans \& K. Frankish (Eds.), In two minds: Dual processes and beyond. Oxford, UK: Oxford University Press.
Thompson, V. A., Prowse Turner, J., \& Pennycook, G. (2011). Intuition, reason, and metacognition. Cognitive Psychology, 63(3), 107-140.
Thompson, V. A., Turner, J. A. P., Pennycook, G., Ball, L. J., Brack, H., Ophir, Y., \& Ackerman, R. (in press). The role of answer fluency and perceptual fluency as metacognitive cues for initiating analytic thinking. Cognition.
Topolinski, S., \& Reber, R. (2010). Immediate truthTemporal contiguity between a cognitive problem and its solution determines experienced veracity of the solution. Cognition, 114(1), 117-122.
Vernon, D., \& Usher, M. (2003). Dynamics of metacognitive judgments: Pre-and postretrieval mechanisms. Journal of Experimental Psychology: Learning, Memory, and Cognition, 29(3), 339-346.
Wiley, J. (1998). Expertise as mental set: The effects of domain knowledge in creative problem solving. Memory \& Cognition, 26(4), 716-730.

# An Information-Theoretic Account of Musical Expectation and Memory 

Kat Agres (kathleen.agres@ eecs.qmul.ac.uk)<br>Centre for Digital Music and Cognitive Science Research Group<br>School of Electronic Engineering \& Computer Science<br>Queen Mary, University of London<br>London E1 4NS, United Kingdom

Samer Abdallah (samer.abdallah@eecs.qmul.ac.uk)
Centre for Digital Music
School of Electronic Engineering \& Computer Science
Queen Mary, University of London
London E1 4NS, United Kingdom

Marcus Pearce (marcus.pearce@eecs.qmul.ac.uk)<br>Cognitive Science Research Group, Centre for Digital Music and Centre for Research in Psychology<br>School of Electronic Engineering \& Computer Science<br>Queen Mary, University of London<br>London E1 4NS, United Kingdom


#### Abstract

When listening to music, we form implicit expectations about the forthcoming temporal sequence. Listeners acquire knowledge of music through processes such as statistical learning, but how do different types of statistical information affect listeners' learning and memory? To investigate this, we conducted a behavioral study in which participants repeatedly heard tone sequences varying within a range of informationtheoretic measures. Expectedness ratings of tones were collected during three listening sessions, and a recognition memory test was given after each session. This enabled us to examine how statistical information affects expectation and memory for tone sequences over a period of increasing exposure. We found significant correlations between listeners' expectedness ratings and measures of information theory (IT), and although listeners demonstrated poor overall memory performance, the IT properties significantly impacted on musical memory. Generally, simple sequences yielded increasingly better memory performance. High-information sequences, for which making accurate predictions is difficult, resulted in consistently poor recognition memory.


Keywords: Music cognition; information theory; computational approach; predictive models.

## Introduction

Music is a fruitful domain for exploring the mechanisms responsible for learning structured temporal sequences, a type of learning that subserves a wide range of human behaviors. Research by Krumhansl (1990), Pearce \& Wiggins (2006), Huron (2006), and others shows that listeners implicitly acquire knowledge about the statistical structure of music. But is this implicit learning influenced by the information contained in the musical signal and, if so, how? Using computational methods, the pitch structure of music can be manipulated systematically to help reveal the ways in which various information-theoretic properties of melody interact and influence human learning and memory.

This paper examines the process of learning novel music over time, with a focus on mental anticipatory processing and musical structure. By using carefully constructed tone sequences, we are able to test how the statistical structure of music, as measured using information theory, affects the expectedness of tones, as well as memory for specific exemplars, over a period of increasing exposure.

## Information Theory and Music

Information theory has contributed to fields as diverse as engineering and linguistics by describing and quantifying the information contained in a signal. This is especially useful for clarifying how the brain processes temporal signals; and indeed, information-theoretic measures such as entropy, a measure of uncertainty, have successfully described and predicted how the human brain anticipates forthcoming sensory input, such as music and language (e.g., Manning \& Schutze, 1999; Abdallah \& Plumbley, 2009). Within the domain of music, there has been a longstanding interest in anticipation and prediction, and statistical and probabilistic approaches to learning have been influential for decades (consider Krumhansl \& Kessler, 1982; and Saffran, Johnson, Aslin, \& Newport, 1999). Computational models such as IDyOM (Pearce, 2005) derive information-theoretic properties of music that accurately reflect and predict listeners' expectations during music listening (Pearce \& Wiggins, 2006; Pearce et al., 2010).

While statistical and computational approaches have modeled human performance on a variety of music perception tasks, these approaches have not yet been extended to modeling the learning trajectory of listeners: we do not yet know how information-theoretic measures capture musical learning over increasing exposure to musical exemplars, and how much exposure is necessary to
learn the statistical regularities of novel music. The following research addresses these questions.

## Behavioral Experiment

In the present study, computational techniques were used to create a set of tone sequences varying systematically across several information-theoretic measures. Varying the sequences' statistical structure allows us to assess which factors have the greatest impact on listeners' musical expectations and memory for tone. We focused on testing the relative influence of three information-theoretic factors based on the information theoretic concepts of entropy rate, multi-information rate (a kind of redundancy), and predictive information rate (see Abdallah \& Plumbley, 2009). These measures are defined for a random process with a known probability distribution, and hence thought of as 'objective'. However, listeners cannot know these probability distributions; they can only estimate them from observations, and so we defined variants of each measure appropriate for an observer processing events sequentially as they happen, updating its estimated probability model as it goes along: they are dynamic information measures based on an adaptive probabilistic model. Since they depend only on the actual observed sequences (rather than a theoretical statistical ensemble) and any prior expectations built into the listener model (which we may think of as summarizing the listener's previous musical experience), we can usefully think of these as 'subjective' information measures.

In our experiments, the listener model was an adaptive first-order Markov chain, as described by Abdallah and Plumbley (2009), which assumes that notes are sampled from a Markov chain with an unknown transition matrix, and tries to estimate the transition matrix from the observations. The model is supplied with an initial expectation (a Bayesian prior) that the transition matrix is similar to a first-order transition matrix derived from a large corpus of Western tonal music in a major key.

The three information measures examined in this paper are Information Content (IC), Coding Gain, and Predictive Information. IC is a measure of the subjective unexpectedness of an observation. Coding Gain measures how much temporal structure or pattern there is in the sequence. And Predictive Information quantifies how much the current observation improves the listener's predictions about future observations (assuming knowledge of the previous observations). High predictive information is also associated with temporal structure or pattern, but of the sort that has more variation, requiring the observer to continually pay attention in order to follow the pattern.

These three measures are defined in the Markov model as follows: at any integer time $t$, let $x_{t}$ be the note occurring at that time, and $\theta_{t}$ be the estimated transition matrix using information available before $t$. Then, the IC at time $t$ is the negative $\log$ probability of $x_{t}$ given the context and the estimated model: $-\log p\left(x_{t} \mid x_{t-1}, \theta_{t}\right)$, where the relevant transition probability is extracted from the matrix $\theta_{t}$. Coding Gain at time $t$ quantifies how much the model's
ability to predict the current observation depends on having observed the preceding observations, and is a difference of $\log$ probabilities: $\log p\left(x_{t} \mid x_{t-1}, \theta_{t}\right)-\log p\left(x_{t} \mid \theta_{t}\right)$, where the latter term is derived from the stationary distribution of the transition matrix. Predictive Information is quantified as the distance between two probability distributions over the next symbol $x_{t+1}$, representing the observer's probabilistic beliefs about $x_{t+1}$ before and after the observation of $x_{t}$. The average of each of these three measures was computed for every tone sequence in the present study, henceforth referred to as sequence statistics.

To investigate the processes underlying musical learning, listeners were exposed to tone sequences and tested on recognition memory over several listening sessions. In each listening session, participants heard tone sequences and rated the expectedness of a tone (termed the "probe tone") within each sequence. Probe tones varied in terms of information content (representing unexpectedness) across sequences. A recognition memory test followed each listening session. This format enabled us to compute information-theoretic measures for every tone sequence, and compare the effect of these measures on probe tone ratings.

We also examined how IT measures impacted on recognition performance in the test sessions. We hypothesized that sequences featuring generally highentropy would be difficult to remember, and probe tones would be rated with lower expectedness. Because each tone sequence was presented in every listening session, we also aimed to clarify the learning trajectory for the different classes of tone sequence; that is, how music represented in short-term memory gradually becomes more richly encoded in long-term memory, and how musical information and complexity, as measured using IT, influence this process over time.

## Method

## Participants

Twenty-three students ( 12 female and 11 male; mean age $=21.0 \mathrm{yrs}$ ) at Cornell University participated in this study for extra credit in a psychology course. The participants had an average of 1.61 years ( $S D=1.88 \mathrm{yrs}$ ) playing music in the previous five years, and an average of 5.82 years $(S D=$ 4.54 yrs ) of lifetime experience playing an instrument.

## Materials and Procedure

After receiving written and verbal instructions, participants listened to tone sequences in three sessions, each lasting approximately 15 minutes and followed by a brief test session. In the listening sessions, participants heard each of the 24 tone sequences (presented in a different order in each session) and were asked to rate the expectedness of a particular tone (the probe tone) within each sequence. This tone was identified visually on the computer screen via a clock counting down on the subsequent tones of the sequence. When the clock returned to midnight, participants rated the expectedness of the concurrently sounding tone on
a scale from 1 to 5 , where ' 1 ' represented highly unexpected and ' 5 ' represented highly expected.

Each listening session was followed by a test session. Sixteen test stimuli were presented in each of the three test sessions, where 8 sequences were Old (had been presented previously) and 8 were New. After each test sequence, participants responded "Yes" or "No" to whether they had heard the sequence before. Upon responding, the listener made a confidence rating on a scale from 1 to 5 where ' 1 ' represented not confident and ' 5 ' represented very confident.

The 24 sequences of the listening sessions each comprised 24 isochronous tones, played in a piano timbre. Each tone was 500 ms in duration, yielding sequences that were 12-seconds-long each. The sequences were generated with an alphabet of 7 pitches (representing one octave of the diatonic scale). A first-order Markov transition matrix was derived (Pearce, 2005) from the scale degrees of Canadian folk songs/ballads, Chorale melodies, and German folk songs in a major key (the same corpus described in Table 2 of Pearce and Wiggins, 2006). To construct the tone sequences, many transition matrices were generated randomly using a process biased towards the tonal transition matrix. From each matrix, one sequence of 24 notes was sampled. A subset of these was then selected manually to ensure a good spread in the 3 -dimensional subjective information space formed by the information theoretic measures described above.

A distinct 500 ms white noise clip was played after every tone sequence in the listening and test sessions as a perceptual "reset" to ensure that expectedness ratings and memory judgments were based only on the current trial. The study was administered on a MacBook Pro laptop, and stimuli were presented and responses collected using Psychophysics Toolbox (Version 3) within the programming environment of MATLAB 2010a (MathWorks, Inc). Participants listened to stimuli over headphones set to a comfortable listening volume.

## Results and Discussion

## Whole-Sequence IT measures and Expectedness During Listening Sessions

To examine how the information-theoretic properties of each sequence influenced the expectedness of probe tones, correlations were analyzed between the IT factors and Average Expectedness Ratings. In terms of whole-sequence statistics, both Sequence IC and Sequence Predictive Information were significant predictors of Average Expectedness Ratings. As shown in the top graph of Figure 1, Sequence IC was correlated with Average Expectedness Ratings such that more predictable sequences (lower Sequence IC values) yielded higher expectedness ratings of probe tones, $R^{2}=.29, F=28.87, p<.01$. The second graph of Figure 1 displays the correlation between Sequence Predictive Information and Average Expectedness Ratings, $R^{2}=.34, F=36.23, p<.01$. The third graph shows

Sequence Coding Gain and Average Expectedness Ratings, also significant in this analysis, $R^{2}=.37, F=41.54, p<.01$.


Figure 1: The main effects of Sequence IC, Sequence Predictive Information, and Sequence Coding Gain (in nats, where 1 nat $=1.44$ bits) on average expectedness ratings during the listening sessions.

Sequences with high average IC values contain unexpectedness; the tones comprising these sequences have high average Information Content. Therefore, it is logical that sequences containing many unexpected, unpredictable tones would yield lower expectedness ratings as shown below.
Regarding the effects of Sequence Predictive Information, information is inextricably associated with unexpectedness: an event cannot be informative if the observer knew it was going to happen, because it will not change the observer's beliefs about the future. (Mathematically, Predictive

Information is upper-bounded by the Information Content.) Hence, sequences with higher average Predictive Information will necessarily have moderately high average information content and thus we would expect the probe tones to see relatively lower expectedness ratings.

Coding Gain is a measure of how much information was gained about the current observation from the preceding context. Therefore, the greater the average Coding Gain of the sequence, the greater the predictability of the sequence and so we would predict higher expectedness ratings in such cases.

Expectedness and Probe Tone IC To examine which factors in the listening sessions had the greatest impact on expectation, a multiple regression analysis was performed with Probe Tone IC, Sequence IC, Sequence Coding Gain, Sequence Predictive Information, and Listening Session as independent measures, and Expectedness Ratings as the dependent measure (note that all expectedness ratings were used, not the average rating for each stimulus). Listeners were included as a random effect in the analysis. There was a significant main effect of Probe Tone $I C, F=181.74, p<$ .001, with high-IC tones rated as less expected. As for the whole-sequence IT measures, there were also main effects of Sequence IC, $F=3.92, p<.05$, and Sequence Predictive Information, $F=9.67, p<.01$. In addition to these main effects, there were also significant interactions between Probe Tone IC and all three of the IT measures of sequences statistics: Sequence IC X Probe Tone IC, F = 22.34, p< .001, and Sequence Coding Gain X Probe Tone IC, F = 35.72, $p<.001$, and Sequence Predictive Information X Probe Tone IC, $F=91.65, p<.001$. Listening Session did not contribute significantly to the results indicating that pitch expectation remained constant overall during the study.


Figure 2: Probe Tone IC as a predictor of average expectedness ratings of probe tones.

Probe Tone IC had the largest effect in the listening sessions, with a significant linear relationship with Expectedness Ratings, $R^{2}=.69, F=154.20, p<.01$. In Figure 2, the average expectedness rating for each melody is shown to display more clearly the main effect on a continuous rather than discrete scale. Low IC tones do
receive reliably higher expectedness ratings than high-IC probe tones over the course of listening.

## Recognition Memory in Test Sessions

Data from the test sessions are reported in Table 1 as Proportion Correct Response. Chance performance is 0.5 , and the similarity of performance for Old and New items indicates little bias towards either response.

Table 1: Recognition memory test performance (proportion correct) for Old and New sequences across listening sessions.

| Listening <br> Session | Old/ <br> Correct <br> (Hits) | Old/ <br> Incorrect <br> (Misses) | New/Correct <br> (Correct <br> Rejections) | New/Incorrect <br> (False Alarms) |
| :---: | :---: | :---: | :---: | :---: |
| Session 1 | 0.67 | 0.33 | 0.64 | 0.36 |
| Session 2 | 0.63 | 0.37 | 0.65 | 0.35 |
| Session 3 | 0.70 | 0.30 | 0.65 | 0.35 |

Despite little evidence for an increase in overall memory performance over the course of the experiment, we investigated whether certain types of statistical information were being learned, and examine whether performance differed depending on the properties of the individual sequences. Therefore, to examine the effects of the IT measures on recognition scores across listening sessions, a logistic regression was performed with Sequence IC, Sequence Coding Gain, Sequence Predictive information, Familiarity (Old or New stimulus), and Listening Session as factors, and Correct Response as the binary dependent variable.

All three whole-sequence statistics showed significant main effects: Sequence $I C, \chi^{2}=16.21, p<.01$; Sequence Predictive Information, $\chi^{2}=12.09, p<.01$; and Sequence Coding Gain, $\chi^{2}=4.27, p<.05$. Listening Session interacted with each of the whole-sequence IT measures: Sequence IC X Listening Session, $\chi^{2}=6.14, p<.05$, Sequence Predictive Information X Listening Session, $\chi^{2}=7.98, p<.05$, and Sequence Coding Gain X Listening Session, $\chi^{2}=6.53, p<$ .05, were all significant interactions.
The only significant interaction including Familiarity was with Sequence Predictive Information, $\chi^{2}=12.15, p<.01$. As shown in the top plot of Figure 3 below, New sequences that are high in Predictive Information yield more correct responses than those with low Predictive Information. Conversely, Old sequences show the opposite trend, with worse recognition memory performance on high Predictive Information sequences. Note that Proportion Correct Response is used in Figure 3 rather than the categorical variable Correct Response for clarity of illustration.


Figure 3: The differential effect of Sequence Predictive Information on Proportion Correct Response during recognition memory tests for New and Old sequences.

Confidence Ratings Confidence ratings of recognition memory judgments were collected after every test sequence; responses were made on a $1-5$ scale where on a where ' 1 ' represented not confident and ' 5 ' represented very confident. A logistic regression was performed with the same factors as those used above: Sequence IC, Sequence Coding Gain, Sequence Predictive Information, Familiarity (Old or New stimulus), and Listening Session. This analysis yielded significant effects of Sequence IC, $\chi^{2}=16.44, p<.01$, and Sequence Coding Gain, $\chi^{2}=15.33, p<.01$, and interactions of these two factors with Listening Session: Sequence IC X Listening Session, $\chi^{2}=21.94, p<.01$, and Sequence Coding Gain X Listening Session, $\chi^{2}=23.10, p<.01$.

As expected, listeners made more confident memory judgments when sequences had lower IC and higher Coding Gain. For Sequence IC, there was a decrease in confidence (fewer 4 and 5 responses) over the course of the experiment, which was especially noticeable for low-IC sequences (because high-IC sequences rarely received 5 responses throughout the study). Similarly, there was also a decrease in highly confident ratings (4 and 5 responses) for Sequence Coding Gain over the course of the experiment, which was more apparent in the high-Coding Gain sequences (lowCoding Gain sequences elicited few 5 responses).

## Conclusion

The analyses above highlight the significant roles that measures of entropy and predictability have on musical learning and memory. The three information-theoretic measures examined here, Sequence IC, Sequence Predictive Information, and Sequence Coding Gain, all impacted on learning over time (as evinced by their significant interactions with Listening Session). In the first memory test, Sequence IC had little effect on the correctness of participants' responses. In the subsequent listening sessions, a trend was displayed between increasing Sequence IC and number of incorrect responses ( $p<.01$ ). Similarly, Sequence Coding Gain did not have a significant effect on response in the first listening session, but was positively correlated ( $p<.01$ ) with correct response in the second and third listening sessions. Sequences with high average Coding Gain were more likely to yield correct responses in the memory tests. In addition, Sequence Predictive Information did not impact on memory performance initially, but by the third listening session, this measure was negatively correlated with Correct Response such that greater Predictive Information led to fewer correct responses ( $p<.05$ ). Again, Predictive Information is upper-bounded by Information Content (unexpectedness); therefore, high Predictive Information sequences sound relatively unpredictable. To summarize, these results suggest that the global statistical properties of the tone sequences had little bearing on recognition memory judgments initially, but over repeated listenings, sequences higher in information and entropy (those that sounded less predictable) produced both lower expectedness ratings and poorer recognition memory.

As displayed by the interaction between Familiarity and Sequence Predictive Information, New sequences that are high in Predictive Information tend to yield more correct responses (Correct Rejections) compared with Old sequences that are high in Predictive Information, which yield fewer correct responses (Misses). We suggest that sequences with high Predictive information are surprising but also distinctive, making them easier to correctly reject on New trials but harder to remember on Old trials. Listeners display poor recognition memory performance for individual sequences, and appear to respond based on the statistical properties of the sequence. Follow-up studies need to be conducted to explore these complex information dynamics, but it is clear that the information-theoretic measures investigated in this study interact dynamically with both expectedness and learning over a period of increasing exposure to novel tone sequences.

## General Discussion

Information-theoretic approaches have elucidated various aspects of music perception, such as melodic expectation (e.g., Pearce et al., 2010). In the IT study described above, three subjective information-theoretic factors, Sequence IC, Sequence Predictive Information, and Sequence Coding Gain, all significantly influenced expectedness ratings of probe tones during the listening sessions. This reveals that
the perceived expectedness of events is influenced not only by properties of the event itself, but also by properties of the sequence within which it is embedded. These factors also impacted on nuanced memory performance during the recognition tests. It was also interesting to discover a significant interaction between Familiarity and Sequence Predictive Information for memory performance. The increasing effect of IT measures on recognition accuracy may result from listeners gradually learning the underlying Markov model: Upon gleaning the basic information structure of the melodies, Predictive Information has a greater effect on recognition memory. Additionally, sequences that have high average IC can also vary in Predictive Information; that is, tones may be perceived as unexpected, but they can be surprising in either a way that increases listeners' predictions of forthcoming tones, or in a way that is surprising but does not increase predictive accuracy. The significant interaction between Familiarity and Sequence Predictive Information, but not Familiarity and Sequence IC, demonstrates that it is not simply the high-information content of sequences, but rather the Predictive Information of these sequences that listeners can successfully use when making memory judgments.

Generally, sequences that were more difficult to predict (higher IC/Predictive Information) gave rise to worse memory performance. There was also an increasing impact of these factors on memory as exposure increased. The effect of Sequence IC became more pronounced as listeners repeatedly heard melodies (e.g., sequences with low average IC were more likely to be remembered by the third listening session). To our knowledge, this research is the first investigation of the time course of music learning using an information-theoretic approach.

Although listeners struggled with the difficulty of the recognition memory task, they responded differentially based on the statistical properties of the sequences. Listeners may be more adept at learning the statistical rules underlying musical sequences than the specific exemplars themselves, especially with non-stylistic music such as the sequences used in this study (see Saffran et al., 1999; Loui, Wessel, \& Kam, 2010; Halpern \& Bartlett, 2010). Listeners are capable of learning a vast number of songs and themes, therefore more ecological stimuli may lead to better learning and memory performance. Language research (a domain in which listeners have been shown to be proficient in statistical learning of phonological sequences) has historically revealed that people tend to remember the semantics of what is said, not a verbatim account (e.g., Bartlett, 1932). Therefore, it may prove more insightful to test listeners' learning of semantics (musical structure and underlying statistics) across exemplars rather than the individual exemplars themselves.

We see from this IT study that learning individual sequences is possible, but challenging. Because we see effects of the IT properties of the stimuli but no significant effect of Listening Session, it is likely that participants were
learning the rules describing the underlying transition matrices rather than the particular exemplars themselves.

Because it is impossible to perform an exhaustive behavioral investigation into which exemplars and rules listeners learn, future work will develop computer models to simulate and predict the process of musical learning. Computational models can offer insight into this process by analyzing information-theoretic measures to predict human listeners' performance. Future work will also test memory differences between ecological melodies and experimentally controlled tone sequences with an expectation that stylistic, ecological exemplars will aid memory performance.

## Acknowledgements

This research was supported by EPSRC via grant number EP/D038855/1: Neural and Information Dynamics in the Processing of Musical Structure.

## References

Abdallah, S. \& Plumbley, M. (2009). Information dynamics: patterns of expectation and surprise in the perception of music. Connection Science, 21, 89-117.
Bartlett, F. (1932). Remembering: A study in experimental and social psychology. Cambridge, UK: Cambridge University Press.
Halpern, A. \& J. Bartlett (2010). Memory for melodies. In M.R. Jones, A. Popper, \& R. Fay (Eds.), Music Perception. New York: Springer-Verlag.
Huron, D. (2006). Sweet anticipation: Music and the psychology of expectation. Cambridge, MA: MIT Press.
Krumhansl, C. (1990). Cognitive foundations of musical pitch. Oxford, UK: Oxford University Press.
Krumhansl, C. \& Kessler, E. (1982). Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. Psychological Review, 89, 334-368.
Loui, P., Wessel, D.L., \& Kam, C. (2010). Humans rapidly learn grammatical structure in a new musical scale. Music Perception, 27, 377-388.
Manning, C. \& Schutze, H. (1999). Foundations of statistical natural language processing. Cambridge, MA: MIT Press.
Pearce, M. T. (2005). The construction and evaluation of statistical models of melodic structure in music perception and composition. Doctoral Dissertation, Department of Computing, City University, London, UK.
Pearce, M. \& Wiggins, G. (2006). Expectation in melody: The influence of context and learning. Music Perception, 23, 377-405.
Pearce, M. T., Ruiz, M. H., Kapasi, S., Wiggins, G. A., \& Bhattacharya, J. (2010). Unsupervised statistical learning underpins computational, behavioural and neural manifestations of musical expectation. NeuroImage, 50, 302-313.
Saffran, J.R., Johnson, E.K., Aslin, R.N., \& Newport, E.L. (1999). Statistical learning of tone sequences by human infants and adults. Cognition, 70, 27-52.

# Effector specific response activation during word processing 

Daniela K. Ahlberg (daniela.ahlberg@uni-tuebingen.de)<br>Department of Psychology, University of Tübingen,<br>Schleichstr. 4, 72076 Tübingen, Germany

Carolin Dudschig (carolin.dudschig@uni-tuebingen.de)<br>Department of Psychology, University of Tübingen,<br>Schleichstr. 4, 72076 Tübingen, Germany

Barbara Kaup (barbara.kaup@uni-tuebingen.de)<br>Department of Psychology, University of Tübingen, Schleichstr. 4, 72076 Tübingen, Germany


#### Abstract

Theories of embodied cognition suggest that sensorimotor processes are involved in language comprehension processes. Recent studies suggested that sentences referring to actions that involve a typical effector (e.g. "He kicks the ball") can systematically activate motor cortex areas that are involved in performing such actions (Hauk, Johnsrude \& Pulvermüller, 2004). In behavioral studies, there is mixed evidence regarding the effects of effector-specific words on corresponding actions. In the current study, we investigated the effect of four word groups on subsequent motor responses involving the hand or the foot. The four word groups were (a) action verbs (e.g., kick, grasp) (b) nouns containing the lexeme 'hand' or 'foot' (e.g., handball, football) (c) nouns referring to objects that are typically manipulated by hand or foot (e.g., cup, shoe), and (d) as control items, nouns that have a spatial association with the upper or lower space (e.g., eagle, root) and which are known to activate locational information in paradigms where no reading is required. We found strong effector-specific compatibility effects revealing a facilitation effect in all noun-groups. Surprisingly, this effect was not present for the action-verbs. Implications of these findings will be discussed.


Keywords: Embodied Cognition; Language Comprehension; Effectors

## Introduction

Many of our daily activities involve language. We speak, we listen to people speaking, we read or we write at various occasions every day. However, in research on language processing there is still no agreement on theoretical assumptions concerning the processes and representations that are involved in language processing. For a long time, the propositional, amodal theory of language comprehension was the predominant view (Kintsch \& van Dijk, 1978; Kintsch 1988; McKoon \& Ratcliff, 1992). According to this view, the result of language comprehension is a meaning representation in an amodal propositional format that captures the content of the linguistic input and integrates it with the reader's background knowledge which is also available in this format. Typically, embodied models of language understanding are viewed as the counterpart to these amodal theories of language processing (Barsalou,
1999). The main assumption of this approach is that language processing is closely connected to other cognitive systems, such as perception and action. There is a tremendous number of empirical studies providing evidence for the embodied view of language comprehension (for an overview, see Jirak, Menz, Buccino, Borghi \& Binkofski, 2010). However, in many cases the individual results are somewhat inconsistent, and cannot be integrated into a coherent processing model. As a result, important theoretical questions concerning the embodied view are still unsolved, as for instance the question whether all kinds of sensorimotor activations are functionally relevant for comprehension or in contrast sometimes constitute a kind of epi-phenomenon. Before turning to these important issues, research first needs to investigate in more detail the individual phenomena, with the goal of arriving at more definite conclusions concerning the boundary conditions for the observed effects. In the present study, we aim to address the question, whether motor activation occurs in a specific manner when processing action related verbs (e.g., kick, grasp), or nouns referring to objects that are typically manipulated with the hand or the foot (e.g., brush, shoe).

Evidence for an embodied view of language understanding has been reported in behavioral and neuroimaging paradigms. In the behavioral domain, observed interactions between language and visualprocessing, and between language and motor processing are typically taken as strong evidence for an embodied model of language comprehension. For example, Zwaan, Stanfield and Yaxley (2002) reported that sentence processing can activate very specific visual representations. In their study, participants had to process sentences such as "The girl saw the egg in the frying pan" and subsequently respond to pictures of the target entity (egg). The pictures could either match the shape of the entity described in the sentences (e.g., a fried egg sunny side up) or mismatch the shape (e.g., an unbroken egg). Responses were faster in the matching than in the mismatching conditions, suggesting that readers had available a visual representation of an egg in the frying pan when reading the corresponding sentence.

Evidence for the reactivation of motor representations during language comprehension was reported by Glenberg
and Kaschak (2002) in the so-called action sentence compatibility effect. In their study, participants had to read sentences and judge the sensibility by moving their arm away or towards their body. Responses were faster if the movement direction implied by the sentence matched the response movement (e.g., "You opened the drawer" and a movement toward the participants) compared to when there was a mismatch (e.g., "You closed the drawer" and a movement away from the participants). These results suggest that participants reactivated the described movements when processing the sentences and thus primed the response movements in the matching conditions.
Interestingly, if language understanding indeed relies on motor activation and the reactivation of experiential traces, these language-action compatibility effects should occur in a very specific manner. In other words, reading sentences such as "He throws the ball" versus "He kicks the ball" should result in rather distinct activation in the motor cortex. That is, when processing sentences such as "He throws the ball" hand related motor areas should be active, whereas when processing sentences such as "He kicks the ball" foot related motor areas should be involved. Indeed, Hauk, Johnsrude and Pulvermüller (2004) reported in an fMRI study such effector-specific motor activation during language understanding (for an EEG study, see Pulvermüller, Härle \& Hummel, 2001). They compared the brain areas activated while performing finger, feet and tongue movements with the activation in a passive reading task of face-, foot- and arm-related sentences. This study revealed clear activation in the motor cortex and in the primary motor cortex during language processing, with the activation being similar to the conditions where the participants actually performed the corresponding actions.
In addition to neuropsychological studies, some behavioral studies reported evidence for effector-specific motor activation during language processing. Marino, Gough, Gallese, Riggio and Buccino (2011) investigated the effects of words on hand movements. Their stimuli consisted of Italian nouns referring to concrete objects, which were both hand- or foot-related, and abstract entities. Participants had to decide whether a presented word referred to a concrete object (e.g., pencil) or whether the word referred to an abstract content (e.g., jealousy). Only in case of concrete objects, participants had to press the response key with their index finger. In case of abstract words, they had to withhold responses. Additionally, participants had to wait with their response until a go-signal was delivered, and this could occur early or late after word presentation. The results showed that participants (all right-handed) responded slower with their right-hand to hand-related words compared to foot-related words. In contrast, with their left hand, they were faster to hand-related words than to footrelated words. Those effects were only found in the early go-signal condition. Marino et al. (2011) explained those results with a left hemispheric specialization for language processing. In case of right hand responses, interference took place due to the left hemisphere being activated by
both, language processing and motor response activation, whereby they compete for common resources. The authors argued that this kind of interference was not present for left hand responses because the motor activation took place in the right hemisphere and thus did not overlap with activation from language processing. The authors themselves state that this explanation cannot account for the facilitation effect of the left hand, because no difference between hand and foot-related words would be predicted.

Scorolli and Borghi (2007) also reported influences of sentence understanding on effector-specific behavioral responses for sentences that imply the usage of a specific effector (e.g. hand vs. mouth). Their participants had to judge the plausibility of sentences with nouns and verbs that refer to objects and actions associated with specific effectors, e.g., to unwrap vs. to suck the sweet. In the first block, hand and mouth sentences, and in the second block, hand and foot-sentences were tested. Half of the participants had to react by saying "yes" into a microphone and the other half had to press a foot-pedal. As predicted, they found match effects with mouth- and foot-responses for mouthand foot-sentences, but not for hand sentences However, Scorolli and Borghi (2007) did not differentiate between word-based and sentence-based effects, thus leaving open whether the reported compatibility effects were triggered by single words or the processing of the whole sentence.

In summary, there is evidence supporting the hypothesis that language can activate effector-specific motor processes, but at the same time it seems difficult to come up with a consistent explanation regarding the underlying mechanisms causing facilitation or interference. We therefore consider it worthwhile to investigate this issue in a very basic behavioral paradigm with the focus being on the influence of single words on responding.

In the current study, we will investigate whether processing of single nouns and verbs with an association to the effectors hand and foot, will result in effector-specific compatibility effects if implemented in a task that does not require active reading. This task will be an alternated version of the original color-naming experiment conducted by Stroop (1935). In our experiment, all words will be presented in a color, and the color will determine the response effector, either hand or foot. According to the Stroop literature any meaningful word can potentially cause interference in a color-response task because wordprocessing is seen as more automatic and faster than responding to the word color (for a review, see MacLeod, 1991). Importantly, interference should only be found if automatic word processing interacts with the required response and can result in response conflict (Botvinick, Braver, Barch, Carter, \& Cohen, 2001). For example, in case of action verbs (e.g., kick), this word might automatically activate effector-specific responses (e.g., foot responses) and thus facilitate compatible responses or interfere with incompatible responses (e.g. hand responses).

Indeed, there is a debate regarding the automaticity of effects reflecting interactions between language processing
and motor responses. For example Bub and Masson (2008) only find an effect of words on subsequent reaching tasks if the words have to be actively processed by the participants.

In our study, we compare four different kinds of word categories: First, action-verbs, using the stimulus-set of Pulvermüller et al. (2001). Second, nouns directly related to one specific effector, involving the lexeme 'hand' or 'foot' (e.g., handball vs. football). Third, nouns referring to objects that are typically manipulated by hand or foot (e.g., paint brush vs. stirrup). Forth, nouns referring to objects with a typical location in the vertical space (e.g., bird, root). The fourth word category consists of a reduced set of up/down words used by Lachmair et al. (2011) and Dudschig et al. (2012, 2013). We will use those words as a control in our paradigm, because they can be expected to show strong compatibility effects with responses differing with respect to the vertical dimension (here: hand $=u p$, foot $=$ down) even in tasks that do not demand active reading. As described above, words will be presented in a color, and the color determines the response effector (hand vs. foot). We predict compatibility effects for all word categories, with possibly stronger effects for the verbs, as well as possibly for the nouns involving the lexemes 'hand' and 'foot'.

## Method

Participants were presented with words displayed in one of four colors in each trial. Their task was to respond to the font color. Depending on the color, they either pressed a key with their hand which was located at chest height or a pedal with their foot which was located at the ground.

## Participants

A total of 30 students ( 9 males) of the University of Tübingen, aged from 19 to 34 years ( $M=22.8$ years, $S D=$ 3.5) participated in this study. Twenty-three of the students were right-handed and 7 were left-handed. All participants had normal or corrected-to-normal vision; none of them had impaired color-vision. They were asked to fill in a form of consent before doing the experiment and received course credit for their participation.

## Apparatus and stimuli

Stimuli were presented in center position on a CRTMonitor in size 12 Courier New bold. Responses were recorded via a PST Serial Response Box, Model Number 200A with a foot pedal. The experiment was programmed with E-Prime® (Psychology Software Tools Inc., www.pstnet.com/E-Prime/e-prime.htm).

The participants stood in front of a height-adjustable table with the possibility of leaning against a wall with their backs. Prior to the experiment, the height of the table and with it the monitor was adjusted such that stimulus words were presented at eye-level of the participants. The foot pedal was also adjusted and fixed in a proper distance to the participant. The response box was situated on the table. Every participant reacted with their dominant body side.

We used 192 German nouns and verbs as stimuli, which could be subdivided into four groups: (a) The verbs were adapted from Pulvermüller et al. (2001) ( $\mathrm{N}=64$ ) consisted of 32 hand- and 32 foot-related action-verbs (e.g., grasping vs. kicking). (b) The "explicit nouns" group ( $\mathrm{N}=32$ ) consisted of 16 nouns containing the lexeme 'hand' and 16 nouns containing the lexeme 'foot' (e.g.; handbag, footprint). (c) The "associated nouns" group consisted of 16 nouns referring to an object that was typically manipulated with the hand and 16 nouns referring to an object that is typically manipulated with the foot (e.g., cup, stirrup). (d) Finally, the "up/down nouns" group ( $\mathrm{N}=64$ ) consisted of a shortened set of up/down words from the study of Lachmair et al. (2011) with 32 words referring to an entity typically located in the upper part of the world and 32 referring to an entity typically located in the lower part of the world (e.g., root, roof). See Table 1 for mean frequencies and mean length (number of characters) of the two sets in each group. Frequencies were retrieved from the "Wortschatz Portal" of the University of Leipzig (http://wortschatz.uni-leipzig.de). Words were presented in the colors blue (rgb, $0,0,255$ ), orange (rgb, 255, 128, 0) brown (rgb, 140, 80, 20) and lilac (rgb, 150, 0, 255) on a white background.

Table 1: Mean length and mean frequency of the two sets of words in each word group.

| Word group | Length | Frequency <br> (SE) |
| :--- | :--- | :--- |
| verbs (hand) | 6.88 | $1073(266)$ |
| verbs (foot) | 6.91 | $8385(3313)$ |
| explicit (hand) | 9 | $1811(1458)$ |
| explicit (foot) | 7.4 | $1682(1186)$ |
| associated(hand) | 8.1 | $442(133)$ |
| associated (foot) | 8 | $367(152)$ |
| up-down (up) | 6.25 | $2734(880)$ |
| up-down (down) | 6.13 | $2747(972)$ |

## Procedure and design

Each trial started with a fixation cross, displayed in the center of the screen for 800 ms . Then the stimulus word was presented until response. Between trials a white screen was shown for 1000 ms . Each word was presented 4 times, resulting in a total amount of 768 trials, which were subdivided into 4 experimental blocks. The experiment started with a practice block, in which 8 practice words were presented two times in different colors. Participants received feedback on speed and accuracy in the practice block (but not in the experimental trials). Reaction times were measured.

Participants were instructed to respond as quickly and accurately as possible to the font color of the word. The mapping of colors to response direction was balanced across participants: All possible color pairs occurred equally often.

For the analyses, we collapsed across the two compatible conditions in each group and the two incompatible
conditions in each group. For hand-related words and upwords, compatible conditions consisted of trials in which the correct response involved a key press with the hand at chest height. For foot-related words and down-words, compatible conditions consisted of trials in which the correct response involved pressing the foot pedal on the ground.
The design thus was a 4 (word group) x 2 (compatibility of the response) design with repeated measurement on both variables.

## Results

Responses faster than 200 ms or slower than 2500 ms , as well as errors were excluded from further analyses. This reduced the data by less than $5 \%$. Mean error rate was $3.9 \%$. Mean RTs are displayed in Figure 1.


Figure 1: Mean response times of correct responses as a function of response compatibility and word group.

The analyses revealed a main effect of group, $F(3,87)=$ 5.3, $p=.002$, a main effect of compatibility, $F(1,29)=5.30$, $p=.006$, and a compatibility-by-group interaction, $F(3,87)$ $=3.32, p=.024$. Separate analyses for the four word groups revealed significant compatibility effects for the three noun groups (explicit nouns: $F(1,29)=7.42, p<.05$; associated nouns: $F(1,29)=4.64, p<.05$; up/down nouns: $F(1,29)$ $=8.56, p<.01$ ) but no significant compatibility effect for the verb group $(F<1)$.
To investigate the different compatibility effects in more detail we compared the different pairs of word groups with each other. The verb group differed significantly from all other groups with respect to the size of the compatibility effect (verbs vs. up/down nouns: $F(1,29)=6.41, p<.05$, verbs vs. explicit nouns: $F(1,29)=8.46, p<.01$, verbs vs. associated nouns: $F(1,29)=4.63, p<.05)$. The analysis revealed no significant difference between the different
noun groups with respect to the size of the compatibility effect (all $F$ s < 1.04).

## Discussion

In the present study, we investigated whether processing of single words with an association to the hand or foot results in effector-specific motor activation if implemented in a task that does not require active reading. Our results clearly show such effector-specific compatibility effects which is in line with the view that effector-specific information is automatically activated during word processing. Our results therefore fit well with the idea that readers re-activate experiential traces during word processing that stem from interactions with the respective referents of these words.

In our experiment we compared four different word groups. As predicted we observed compatibility effects for all noun groups. For the up/down nouns we had predicted such an effect because the two responses in this paradigm (hand vs. foot) differed with respect to the location of the response key in vertical space (up vs. down). In previous studies involving hand responses with up- versus down-keys strong compatibility effects were also observed, even in tasks that did not require active reading (Lachmair et al., 2011). Finding a compatibility effect with these nouns in the present paradigm shows that the up/down-effect generalizes to an experimental situation involving responses with different effectors. As such these results provide further evidence for the stability of this effect.

For the remaining two noun groups we had predicted compatibility effects on the basis of the view that readers activate experiential traces during word processing that stem from interacting with the objects these words refer to. If these objects are typically manipulated with one of the respective effectors, then this effector should be primed during processing and a response involving this effector should be facilitated. Interestingly, the observed compatibility effects were equally strong in these two groups, although the association with the effectors was linguistically specified in the explicit noun group but not in the associated noun group. Thus, if linguistic representations had played a prominent role during processing one could have expected stronger compatibility effects for the explicit noun group compared to the associated noun group. The fact that this was not observed can be taken as further evidence for an experiential-trace view of language understanding.

In contrast to our predictions, we did not observe compatibility effects for the action verbs. This is surprising for several reasons. First, in our view the association between these words and the two effectors seems particularly strong, and for this reason we had even expected to find the strongest compatibility effects in this group. Second, and more importantly, neuro-scientific studies involving the exact same set of stimuli repeatedly found evidence for an effector-specific activation during the processing of these words (e.g., Pulvermüller, et al. 2001). What then may be the reason for not finding effector-
specific compatibility effects with verbs referring to actions that are typically performed with the hands or the feet? Several possibilities come to mind.
First, in German, nouns begin with a capital letter, but verbs do not. To present the words in their natural appearance, we presented the nouns with a beginning capital letter, and the verbs in small letters throughout. In principle it seems possible that this difference could account for the different results obtained with nouns and verbs in the present experiment, especially if one considers that active reading was not required by the experimental task. Maybe participants were more successful at ignoring the words when they homogeneously involved small letters than when there was a capital letter at the beginning. But this seems unlikely as the standard Stroop effect and several variants thereof have been observed in different sets of tasks, involving various types of word displays. Thus, we regard it as unlikely that the pure difference in word display, specifically concerning the first letter, can account for the differences between nouns and verbs in the present paradigm.
Another explanation for the missing influence of verbs on responding might be a different time course of verb processing in contrast to noun processing. We consider it conceivable that verbs require more processing effort than nouns, for instance due to differences in breadth of meaning (Gentner, 1981). If it takes relatively long to process verbs, then the information in the verb that potentially triggers the conflict may become active only after the response decision (responding with hand vs. foot) has already been made. This would explain why verbs did not influence responding. If indeed differences in processing times are responsible for the missing effects in verbs, we might find compatibility effects for verbs if a lexical-decision task was being employed instead of a Stroop color-response task (e.g., Mirabella, Iaconelli, Spadacenta, Federico, Gallese, 2012). Critically, in studies using verbs referring to upward or downward directed motion (e.g., rise, fall) similar effects were observed as in a study implementing nouns (e.g., bird, ground), even if implemented in a Stroop-like colorresponse paradigm (e.g., Dudschig, Lachmair, De Filippis, de la Vega \& Kaup, 2012). Thus, a general difference between verb and noun processing cannot fully account for our findings.
Finally an alternative explanation may be that verbs referring to actions are associated with very specific motor plans. Maybe no compatibility effect was observed for the verb group because the overlap between the motor activation involved in understanding these words and pressing a key with the hand or the foot simply was not large enough. After all, a movement such as kicking is a quite different foot movement than pressing a foot pedal, and grasping is quite different from pressing a key with the index finger. If this hypothesis is correct, compatibility effects should be observed with verbs that refer to actions that are similar to the response actions in the experiment. Indeed, the embodied language processing account predicts
that words become associated with experiential traces (e.g., Zwaan \& Madden, 2005). Such an account would predict that experiential traces are rather specific, such that activating a specific effector might not be sufficient for facilitating language understanding.

In summary, our results provide clear evidence for effector-specific activation during single word processing in a very basic color-response paradigm. Especially, participants did not have to actively process the words' meaning and nevertheless subsequent responses were affected. However, this compatibility effect is limited to certain word categories and does not seem to occur in case of verb processing (e.g., kick, grasp). Our results suggest that conflict between words and effector-specific motor responses is not a general effect between words referring to foot or hand related entities and actions, but is rather specific to certain word categories. These results are of interest to both, the embodied language processing models (e.g., Glenberg \& Kaschak, 2002) but also for the conflict monitoring model (e.g., Botvinick et al., 2001). Future studies will be needed to investigate whether difference in processing times are responsible for these effects, or whether the missing effect of action verbs on subsequent responses is due to the fact that very specific motor-plans are activated by the verbs.

## Acknowledgments

This project was supported by a Margarete-von-Wrangell Fellowship appointed to Carolin Dudschig (European Social Fund and the Ministry Of Science, Research and the Arts Baden-Württemberg) and by the SFB833/B4 project of Barbara Kaup (German Research Foundation).

## References

Barsalou, L. W. (1999). Perceptual symbol systems. Behavioral and Brain Sciences, 22, 577-660.
Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., \& Cohen, J. D. (2001). Conflict monitoring and cognitive control. Psychological Review, 108, 624-652.
Bub, D. N., Masson, M. E. J. \& Cree, G., S. (2008). Evocation of functional and volumetric gestural knowledge by objects and words. Cognition, 106, 27-58.
Dudschig, C., Lachmair, M., de la Vega, I., De Filippis, M., \& Kaup, B. (2012). From top to bottom: spatial shifts of attention caused by linguistic stimuli. Cognitive processing, 13, 151-154.
Dudschig, C., Lachmair, M., de la Vega, I., De Filippis, M., \& Kaup, B. (2012). Do task-irrelevant directionassociated motion verbs affect action planning? Evidence from a Stroop paradigm. Memory \& Cognition, 40, 10811094.

Dudschig, C., Souman, J., Lachmair, M., de la Vega, I., \& Kaup, B. (2013). Reading "Sun" and Looking Up: The Influence of Language on Saccadic Eye Movements in the Vertical Dimension. PloS ONE, 8, e56872.

Gentner, D. (1981). Some interesting differences between verbs and nouns. Cognition and brain theory, 4, 161-178.
Glenberg, A. M. \& Kaschak, M. P. (2002). Grounding language in action. Psychonomic Bulletin \& Review, 9, 558-565.
Hauk, O., Johnsrude, I. \& Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. Neuron, 41, 301-307.
Jirak, D., Menz, M. M., Buccino, G., Borghi, A. M., \& Binkofski, F. (2010). Grasping language-A short story on embodiment. Consciousness and Cognition, 19, 711-720.
Kintsch, W. (1988). The role of knowledge in discourse comprehension: a construction-integration model. Psychological Review, 95, 163.
Kintsch, W., \& Van Dijk, T. A. (1978). Toward a model of text comprehension and production. Psychological Review, 85, 363.
Lachmair, M., Dudschig, C., De Filippis, M., de la Vega, I. \& Kaup, B. (2011). Root versus roof: Automatic activation of location information during word processing. Psychonomic Bulletin \& Review, 18, 11801188.

McKoon, G., \& Ratcliff, R. (1992). Inference during reading. Psychological Review, 99, 440.
Marino, B. F. M., Gough, P.M., Gallese, V., Riggio, L. \& Buccino, G. (2011). How the motor system handles nouns: a behavioral study. Psychological Research. DOI 10.1007/s00426-011-0371-2

MacLeod, C. M. (1991). Half a century of research on the stroop effect: An integrative review. Psychological Bulletin, 109, 163-203.
Mirabella, G., Iaconelli, S., Spadacenta, S., Federico, P.\& Gallese, V. (2012). Processing of hand-related verbs specifically affects the planning and execution of arm reaching movements. PLoS ONE 7: e35403. doi:10.1371/journal.pone. 0035403
Pulvermüller, F., Härle, M. \& Hummel, F. (2001). Walking or talking? Behavioral and neurophysiological correlates of action verb processing. Brain and Language, 78, 143168.

Scorolli, C. \& Borghi, A. M. (2007). Sentence comprehension and action: Effector specific modulation of the motor system. Brain Research, 1130, 119-124.
Stroop, J. R. (1935). Studies of interference in serial verbal reactions. Journal of Experimental Psychology, 18, 643662.

Yeung, N., Botvinick, M. M., \& Cohen, J. D. (2004). The Neural Basis of Error Detection: Conflict Monitoring and the Error-Related Negativity. Psychological Review, 111, 931-959.
Zwaan, R. A. \& Madden, C. J. (2005). Embodied Sentence Comprehension. In D. Pecher \& R. A. Zwaan (Eds.) Grounding cognition: The role of perception and action in memory, language, and thinking (pp. 224-245). Cambridge University Press, Cambridge.

Zwaan, R. A., Stanfield, R. A., \& Yaxley, R. H. (2002). Language comprehenders mentally represent the shapes of objects. Psychological Science, 13, 168-171.

# Simulating Overall and Trial-by-Trial Effects in Response Selection with a Biologically-plausible Connectionist Network 

Blair C. Armstrong (b.armstrong@bcbl.eu)<br>Basque Center on Cognition, Brain, and Language<br>Paseo Mikeletegi 69, San Sebastian, 20009 Spain<br>David C. Plaut (plaut@cmu.edu)<br>Department of Psychology and Center for the Neural Basis of Cognition, Carnegie Mellon University<br>5000 Forbes Avenue, Pittsburgh, PA 15213 USA


#### Abstract

Ratcliff, Van Zandt, and McKoon (1999, Psych. Rev.) claim that connectionist models fail to simulate many aspects of how individuals select one of two possible responses. Here, these claims are re-evaluated via computational and behavioral investigations of an extended version of the original numerosity judgment task. The results of the experiment indicate that some of the empirical effects that the models failed to capture do not generalize and were likely due to idiosyncratic aspects of the original methodology. The simulations show that a more biologically-plausible model captures the bulk of the new effects, including some trial-by-trial adaptive effects that are outside the scope of models tested against aggregate data, and emergent asymptotic stability that has previously required an explicit leak parameter.


Keywords: response selection, decision making, connectionism, numerosity judgment, overall and trial-by-trial effects

Understanding how one of multiple candidate responses is selected in a given task is a long-standing and critical issue in cognitive science, and is one of the earliest domains to have been investigated with computational models. To date, much of the work has focused on the sub-issue of how individuals perform in tasks in which they must rapidly select one of two possible responses (i.e., speeded two-alternative forcedchoice tasks; 2AFC tasks). This has led to the development of several models that can be fit to data from 2AFC tasks with a high degree of precision (e.g., the diffusion model; Ratcliff, 1978). One of several key limitations of these models, however, is that they are highly domain-specific and are not naturally extendable to studying other intuitively related issues, such as 'closed-set' response selection tasks involving three or more pre-specified candidate responses, or 'open-set' response selection tasks which require the production of novel responses such as nonword naming. These models are also often fit to aggregated data and do not explain how the decision system adapts over time based on its past experiences.

One possible avenue for addressing these limitations is the development of a connectionist model of response selection, given the connectionist framework's grounding in domaingeneral learning, representation, and processing principles that are drawn from systems and cellular neuroscience. Not only might such a model be able to explain the overall and adaptive effects in 2AFC tasks, it should also be readily extendable to the other response tasks described previously. Moreover, insofar as connectionist models fail in these endeavors, this can serve to guide the development of improved
principles which can, by virtue of the domain-general nature of the framework, have wide-spread implications for domains well beyond response selection (e.g., semantic cognition).

Past work by Ratcliff, Van Zandt, and McKoon (1999) provides some initial insight into the performance of connectionist models of 2AFC tasks relative to that of the diffusion model in simulating performance in a numerosity judgment task. In this task, participants were presented with a $10 \times 10$ array which was filled with a number of asterisks sampled from two overlapping distributions with 'low' and 'high' mean numbers of asterisks, and made responses indicating which distribution they believed had been sampled from to generate the stimulus. The model comparisons revealed that the connectionist models failed to capture important aspects of the behavioral data (e.g., latency-accuracy functions, trial-by-trial adaptive effects).

To address some of these limitations, Usher and McClelland (2001) introduced a revised connectionist formalism in the leaky, competing accumulator model. Changes in this model included explicit constraints on the sign of the weights between competing units and from the underlying source of evidence that drives the response units, and the use of a threshold-linear activation function that is not differentiable at all points in time. A critical implication of the latter change is that it violates the mathematical principles that underlie standard gradient descent learning algorithms such as backpropagation (Hinton, 1989). Collectively, these changes rendered the accumulator functionally analogous to the diffusion model, and generally showed identical or superior fits to that model. This notwithstanding, a fundamental issue with this type of domain-specific connectionist model is what strengths of the standard connectionist framework were given up during model development. In particular, the disconnect between these models and standard connectionist learning algorithms prevents these models from being effortlessly extended to other response selection tasks-let alone cognitive processing and learning in other domains.

An alternative approach to developing improved connectionist models of response selection is to focus, instead, on improving the domain-general assumptions of the framework. One way to do this that is independent of the particular constraints needed to simulate response selection is to more accurately instantiate the known connectivity and processing characteristics of the brain. For instance, neurons
are either excitatory or inhibitory-not both, as is the case in standard connectionist models. There are also more excitatory than inhibitory neurons, which biases the type of information that can be encoded by each sub-population: inhibitory neurons serve primarily to regulate overall activation in the information-content-rich excitatory neurons. Connections between brain regions are also typically only excitatory and relatively sparse (or functionally weak), whereas there is dense (or functionally strong) connectivity among both excitatory and inhibitory neurons within a brain region. The activation dynamics of individual neurons are also better approximated by activation functions that do not possess an upper non-linearity as is the case for sigmoidal functions (Usher \& McClelland, 2001), but that are nevertheless differentiable at all points in time, such as the hybrid sigmoid-linear activation function that is presented in detail later.

To date, models that are constrained by the aforementioned characteristics of systems and cellular neuroscience have been found to capture a wide range of empirical effects such as the temporal dynamics of ambiguous word comprehension (Armstrong, 2012) and the ERP correlates of word and nonword processing and of behavioral lexical decision (Laszlo \& Plaut, 2012). The present work extends these investigations by evaluating whether a more biologically-plausible connectionist model can simulate overall and adaptive effects in a simple perceptual task, without abandoning key principles such as learning or adopting ad hoc connectivity constraints.

Prior to the computational work, however, an appropriate set of benchmark data must be identified. Ratcliff et al. (1999) argued that their numerosity judgment task data were representative of results of many tasks and could therefore be treated as a gold standard for model comparison. However, detailed inspection of their results suggests that idiosyncratic and esoteric aspects of their methods may have led to atypical results. For instance, participants were potentially able to adopt sophisticated response strategies beyond those that are incorporated into simple models of response selection (notwithstanding that ultimately, more complex models should account for these data). For instance, participants received extensive experience with the task $(\approx 12,000$ trials). This may have interacted with the fact that participants were also explicitly told that the 'low' and 'high' distributions overlapped and therefore that it would not always be beneficial to adjust their performance following responses that were labeled as 'incorrect' (in contrast to the definitional behavior of error-driven learning and to the behavior observed in Armstrong, Joordens, \& Plaut, 2009). To explore adaptation in the decision system, Ratcliff and colleagues also repeatedly and somewhat predictably manipulated the likelihood of sampling from the 'low' and 'high' distributions ( $\approx 30$ times) effectively changing the optimal threshold for making 'low' or 'high' responses-which may have allowed participants to utilize sophisticated cognitive control mechanisms. Consequently, we first investigated the representativeness of the original data in an extension of the original study.

## Numerosity Judgment Experiment

The experiment assessed the generality of the empirical findings reported by Ratcliff et al. (1999) using a slightly modified version of the original methodology. This involved (a) splitting the distribution of number of asterisks using a fixed threshold rather than two overlapping 'low' and 'high' distributions, so that the feedback provided in the task was perfectly accurate within a mega-block ( 400 trials split into multiple blocks) and could, in principle, provide a basis for learning to respond perfectly, (b) eliminating or minimizing participants' prior experience with potential changes in the threshold value across mega-blocks that would allow them to develop sophisticated response strategies by changing the threshold value at most two times, and (c) only presenting participants with 3 mega-blocks totaling 1,200 trials (vs. 12,000 in the original study, of which the first 1,200 were dropped) to study initial learning when adaptive effects may be larger and more readily detectable. Insofar as these modifications produce divergent results, the data from this experiment may be a superior gold standard for model evaluation.

## Methods

Participants. A total of 121 right-handed Carnegie Mellon undergraduates participated in exchange for course credit.
Apparatus. The experiment was implemented using PsychoPy 1.71.01 (Peirce, 2007). Responses were recorded on standard computer keyboard using the ' $z$ ' and ' $/$ ' keys.
Stimuli and design. The stimuli used in the experiment consisted of a variable number of asterisks in a $10 \times 10$ grid. The number of asterisks was sampled from a trimmed normal distribution (mean $=50, \mathrm{SD}=14, \min =28, \max =72$ ). These stimuli were divided into 'low' and 'high' categories on the basis of whether the number of asterisks fell above or below a threshold value. The threshold that delineated a 'low' response from a 'high' response could be either 4.5 points below or above the mean of the distribution in a given megablock. These parameters are similar to those employed in the original experiment, with the critical difference that there was no overlap between the 'low' and 'high' distributions.

Participants were presented with 10 practice trials followed by three mega-blocks of 400 experimental stimuli. The threshold for making 'low' or 'high' responses could potentially change across each of these blocks. The full set of 16 different combinations of thresholds were run across the three mega-blocks (e.g., 'low/low/low' vs. 'low/low/high' vs. 'low/high/low,' etc.). Preliminary analyses indicated that the data could be grouped based on whether the threshold remained the 'same' between adjacent blocks or 'changed' between adjacent blocks. This allowed the critical number of conditions to be reduced to four for the analyses (same-same, same-change, change-same, change-change). Trials were presented across 25 blocks of 50 trials, except for the first and last blocks, which contained 25 trials.

The frequency with which each number of asterisks would be presented was a multiple of the probability density func-
tion. The stimuli were divided into 'low' and 'high' groups on the basis of the response threshold for the mega-block. Given the threshold levels used in the experiment, $72 \%(36 \% \times 2)$ of trials fell on the tails of the distribution and were always either 'low' or 'high.' The correct response for the remaining $28 \%$ of trials in the center of the distribution depended on the threshold for the mega-block. The positioning of the asterisks within the array and the order of the stimuli were random, with the constraint that no more than five of the same type of stimuli could be presented sequentially.

Procedure. Participants were instructed that they would have to decide whether the number of presented asterisks was either 'low' or 'high.' They would have to learn what constituted a 'low' or 'high' number of asterisks by making responses and learning from the feedback that was provided. Note that in contrast to Ratcliff et al. (1999), participants were not instructed that a given number of asterisks could be produced by either the 'low' or 'high' distributions (because the feedback in the present experiment was accurate within each mega-block), nor were they informed that the threshold that delineated between the 'low' and 'high' distribution might change during the experiment. This was predicted to increase the likelihood that participants would adapt to the changes in the characteristics of the stimuli using simple statistical learning mechanisms based on the feedback that was provided.

Participants were instructed to respond as quickly as they could without making many mistakes. To operationalize this instruction, after each block participants received a message to "try to go faster, even if it means making a few more errors", or to "try to make a few less errors even if it means slowing down." The message that a participant received depended on whether their accuracy was above or below $90 \%$, although this was not known to them. Following the instructions, participants were presented with the practice trials followed by the 25 experimental blocks.

Each trial consisted of (1) a fixation stimulus (+) for 500700 ms , (2) a blank screen for 50 ms , and (3) a number of asterisks, which remained on the screen until participants responded or for a maximum of 5000 ms . If the response was incorrect (4) "INCORRECT" appeared on the screen for 400 ms . The next trial began automatically.

## Results

Initial proficiency. The instructions alone were sufficient for participants to configure their response system in line with the general demands of the task, as assessed via a binomial test on the accuracy of the first practice trial relative to chance (mean accuracy $=63 \% \mathrm{SE}=0.05, p=0.008, n=104$ ).
Overall performance for the same-same condition. Overall performance in the experiment was slightly less accurate and slower than that reported by Ratcliff et al. (1999). This notwithstanding, the qualitative similarity between the studies on a number of metrics was quite high, including accuracy and latency as a function of distance from the response threshold, the latency distribution for correct and in-
correct responses, and the latency-accuracy functions. Figures 1 and 2 plot a subset of these data (the omitted figures, which are not included because of space constraints, showed similar qualitative matches to the model data). Note that because fewer stimuli were presented as distance increased from the response threshold, the data from both the experiment and the simulation becomes increasingly unreliable as distance increases (particularly for incorrect response latency), so later comparisons between the model and the simulation focus on distance from the threshold of 10 asterisks or less.

Sequential effects for the same-same condition. Sequential effects in the same-same condition were examined as a function of the number of blocks for stimuli of different distances from the threshold. In contrast to Ratcliff et al. (1999), the data plotted in Figure 3 showed a continuous decrease in latency as a function of practice, particularly for large distances from the threshold. Similar trends were observed in the accuracy data (not shown), although performance reached an asymptotic state within the first five blocks.

Mixed-effect regression models (Baayen, Davidson, \& Bates, 2008) were used to further explore the effects of a number of characteristics of the preceding trial on the current trial's accuracy and latency. Due to space constraints, only the effects of previous trial accuracy, stimulus type, and response are reported. Significant effects have p-values less than .05 . For the dependent measure of accuracy, previous trial accuracy did not predict significant variance, repetitions of the same stimulus increased accuracy, and repetitions of the same response decreased accuracy. For the dependent measure of correct latency, a previous accurate response and a repetition of the same response both decreased latency, and there was no effect of stimulus repetition. Additionally, the effects of prior accuracy decreased as a function of practice.
Adaptive effects following threshold changes. Figure 4 plots correct latency as a function of trial number for different numbers of asterisks and combinations of constant or changing response thresholds across mega-blocks (participant's accuracy data, not shown, showed similar dynamics). Three groups of asterisks are presented: one fell just below the initial low threshold (37-45), one fell just above the initial low threshold and just below a later high threshold in conditions in which a threshold change occurred later (46-54), and one was well above the initial low threshold but was immediately above a high threshold if the threshold value changed (5563). The results indicated that in both the accuracy and the latency data the adaptation that followed a threshold change occurred over an extended number of trials. Specifically, performance generally did not approach an asymptotic level until approximately 100 trials after the threshold change.

## Discussion

Despite the methodological similarities, many of the effects observed in this experiment diverge from those reported by Ratcliff et al. (1999). Moreover, the analyses of early performance reported here question some of the modeling as-


Figure 1: Latency distributions. [Left] Expt. [Right] Sim.


Figure 2: Latency as a function of accuracy.


Figure 3: Correct latency as a function of block number for bins of asterisks of varying distance from the threshold.



Figure 4: Correct latency as a function of trial number for the different conditions.
sumptions reported in that paper that could have contributed to the poor performance of the connectionist models that were tested. For instance, the above-chance initial accuracy during the practice blocks and the consistent performance improvements throughout the experiment suggest that a connectionist model should begin simulating the task with a basic level of proficiency which continues to improve with practice. This contrasts with the simulations conducted by Ratcliff and colleagues, for which either a trained model for which learning had been disabled or an untrained network for which learning was enabled were assessed.

Similarly, the significant effects of the accuracy of the previous trial on the accuracy and the latency of the subsequent trial that were observed in the experimented reported here were not in line with those reported by Ratcliff et al. (1999). These effects are, however, in line with the expected behavior that would result from using an error-driven learning algorithm and with the results observed in other studies (Armstrong et al., 2009). One possible reason for the discrepancy is that the participants used their explicit knowledge of the inconsistent nature of the feedback provided in the original experiment to develop adaptation strategies that override the effects of simple error-driven learning algorithms (but that may be captured by more sophisticated algorithms that do consider such factors; Hinton, 1989).

The adaptive effects observed in the present experiment following the change in the threshold also differed from those reported by Ratcliff et al. (1999). In particular, the rate of adaptation following a threshold change was relatively slow and approximately 100 additional trials were necessary to reach a new asymptotic level of performance. this contrasts with the results of the original experiment, which showed that the new asymptotic level was reached in an order of magnitude fewer trials (5-15). This discrepancy is likely due to participants' extensive experience with threshold changes at semi-predictable intervals in the original task.

Taken together, these discrepancies undermine prior claims about the representativeness of the original task and data as a gold standard for model comparison, and suggest that the present data are a more appropriate gold standard for evaluating the performance of simple models of response selection.

## Numerosity Judgment Simulation

The simulation work evaluates whether a more biologicallyplausible connectionist model produces the same patterns of effects that were observed in the experiment.

## Methods

Participants. Two simulated participants completed each of the main conditions in the experiment (same-same, samechange, change-same, change-change).

Network architecture. The model architecture, based on the biologically-plausible connectivity principles described in the introduction, is presented in Figure 5. The visual inputs were divided into two groups of 100 units, the first of which coded for the presence of an asterisk in a given location, whereas the second coded for the absence of an asterisk in a particular location. This normalizes the overall amount of activity in a similar fashion to on-center/off-surround and off-center/on-surround visual neurons. One response unit coded for 'low' responses and the other for 'high' responses.

Arrows indicate full connectivity from one pool of units to another, with the exception that units were not connected to themselves. Outgoing connections from excitatory units were constrained to be positive and were initialized to a mean value of 0.15 . Outgoing connections from inhibitory units


Figure 5: Network Architecture. $\mathrm{I}=$ Inhibitory unit. Solid arrows = excitatory connections. Dashed arrows = Inhibitory connections.
were constrained to be negative and were initialized to a mean value of -0.4. All of the units received a bias connection with a mean value of -2.19 . Weights were sampled from a uniform distribution centered on the mean and with a range of 1.0 , with the condition that weights below zero for excitatory units and above zero for inhibitory units were clipped at zero. Furthermore, so that the network would not need to learn positional invariance (i.e., that the same amount of excitation should arrive at the hidden layer regardless of where in the array the asterisk was presented) the weights from the units in each visual sub-group were constrained to have the same values. Finally, to reduce the difference in terms of total activation across the different pools, the output from each visual sub-group was normalized to range between 0 and 1 .

All of the hidden and response units integrated their inputs over time $(d t=0.2)$. A unit's output, $o$, was a sigmoid-linear function of its net input, $i$, and of normally-distributed output variability (error), $\varepsilon$, per the following equation:

$$
o(i)= \begin{cases}\frac{1}{1+e^{-i}}+\varepsilon, & \text { if } i<=0 \\ 0.25 i+\varepsilon, & \text { if } i>0\end{cases}
$$

This equation approximates a threshold linear function as a continuous transition from a relatively low and stable activation state regardless of the specific amount of input, to a state wherein the activation of the unit varies linearly as a function of the input (the 0.25 value was selected because it is the derivative of the sigmoid for $i=0$; the equation is therefore continuous and differentiable despite being defined in two parts). The error reflects the variability inherent to neural processing and had a standard deviation of 0.025 for all but the visual units. For those units, error was also intended to capture the uncertainty in an individual's estimate of the number of presented asterisks, and was set to 0.1 .

Representations. The input patterns for the network were generated in the same manner as in the behavioral experiment. The target outputs for the response units were set such that a 'low' number of asterisks had a target of 1.0 for the 'low' unit and 0.0 for the 'high' unit; the complementary pattern was used for presentations of 'high' stimuli. Two sets of 1,200 patterns were created: pre-training patterns and task simulation patterns. The number of pre-training patterns was determined in pilot simulations that found that after approximately 1,200 trials, the model was about as accurate as the human participants at the beginning of the practice trials.

Pre-training and task simulation. The processing of pretraining and task-simulation trials was identical. On each trial, the net input and output of the hidden and response units
were set to -2.19 and 0.1 respectively. The network was then presented with the input pattern and trained for 100 unit updates. Error was calculated for the last 95 unit updates. The error was scaled by a factor of 3.0 for the units that were supposed to be off to encourage the simulation to make slower but mostly-accurate responses. A unit's target activation was adjusted such that it was considered to be correct once it was either below 0.1 or above 0.9 for units that were supposed to be off and on, respectively. Error was calculated using a two-piece error function: cross-entropy error was used for activation values below 0.5 and sum-squared error was used for activation values above 0.5 (Hinton, 1989). Weights were adjusted after each trial using a steepest gradient descent algorithm and a learning rate of 0.01 . Units were considered to have made a response when a response unit's activation exceeded 0.5 . The network's response latency was how many unit updates had occurred prior to responding.

## Results

Overall performance for the same-same condition. The model showed the same qualitative effects (and reasonable quantitative similarity) for accuracy and latency as a function of distance from the response threshold and the hazard functions (not shown), the latency distribution for correct responses (Figure 1), and the latency-accuracy functions up to approximately 10 asterisks, beyond which the both empirical and simulation data are not very reliable (Figure 2).

Sequential effects for the same-same condition. The model showed similar increases in accuracy (not shown) and decreases in latency (Figure 3) as a function of practice, including differential latency decreases for the slowest stimuli that were closest to the response threshold.

Mixed-effect regression analyses of the effects of the characteristics of the previous trial did show some weak patterns of disagreement with the behavioral data, however. For the dependent measure of accuracy, prior accurate responses and repetitions of the previous stimulus decreased the accuracy on the subsequent response, whereas repetitions of the response increased accuracy. For the dependent measure of latency, prior accurate responses non-significantly increased accuracy, repetitions of the previous stimulus type significantly decreased latency, and repetitions of the response nonsignificantly increased latency.
Adaptive effects following threshold changes. Figure 4 shows that the simulation recapitulated the main effects in the behavioral experiment following a threshold change: gradual increases in the latencies for stimuli that were suddenly closer to the new threshold, and gradual decreases in the latencies of stimuli that were further from the new threshold.
Activation trajectories. While running the simulations, an additional emergent property of this architecture was observed: despite employing an activation function for which there is no explicit upper bound, the units tended to settle to stable asymptotic activation levels (Figure 6). This was true both if the input corresponded to a number of asterisks


Figure 6: Activation trajectories for the 'low' (top line) and 'high' (bottom line) response units for a 'low' stimulus (32 asterisks).
that was near or far from the response threshold, although the asymptotic level of activation did differ across these two cases. This property is of interest because the leaky accumulator model required an explicit leak current to avoid runaway activation if a stimulus was presented for an extended period.

## Discussion

In contrast to the claims of Ratcliff et al. (1999) the connectionist model succeeded in capturing a substantial portion of the effects observed behaviorally. The main disagreement, in terms of model-behavior mismatch, was in terms of the specific sequential effects of previous trial accuracy, stimulus repetition, and response repetition. Two causes of these discrepancies are currently being investigated: First, the current model may trade off speed and accuracy slightly differently to the human participants, which may be addressed by adjusting the scaling of the error for incorrectly activating a response unit. Second, the current simulation only instantiated memory of prior experience in the form of weight adjustments, whereas these effects may be more accurately captured in a slightly more sophisticated model wherein residual activation from processing the prior stimulus influences subsequent performance (Plaut \& Booth, 2000). Nevertheless, the model did succeed in capturing the overall rates of sequential and trial-by-trial adaptation-effects that are usually outside the scope of models that are typically only evaluated by fitting aggregate data (e.g., the diffusion model) and that are not used to understand how and why model parameters (the weights in the present model) are gradually derived by learning from trial-by-trial experience.

The use of the more biologically-plausible framework also had the effect of inheriting many of the properties of the leaky accumulator model, or at least close approximations thereof, that has been tailored to response selection (e.g., bottom-up excitation, indirect competition between response units via lateral inhibition). Thus, this model gains parsimony and independent support and validity from neurobiology, while also being domain-general and suitable for studying other phenomena such as the ERPs and ambiguous word comprehension (Armstrong, 2012; Laszlo \& Plaut, 2012). Moreover, as shown by the activation trajectories in Figure 6, this framework shows initial promise at generating stable asymptotic states that were not present in the leaky accumulator model without the addition of an explicit leak parameter.

## General Discussion

The simulation's ability to capture most of the effects in the new behavioral data has several important implications. First,
it highlights the importance of careful task analysis and a consideration of the mechanisms that drive performance in identifying an appropriate gold standard task. More generally, it also suggests that a better method for assessing model performance is through the use of a broad range of tasks that generate multiple gold standards, to avoid discrediting certain frameworks on the basis of what may ultimately be established as somewhat atypical effects (notwithstanding that more complex models with appropriate mechanisms should capture those effects). Furthermore, it highlights the value of independently-motivated biological constraints on developing an improved set of domain-general computational principles that can be readily extended to other phenomena (e.g., selecting an appropriate response when naming a nonword). Finally, this work highlights the inherent trade-off between developing tailored quantitatively-precise models versus developing domain-general models: the latter may (at least initially) produce less precise quantitative fits, but they can be extended to a much broader set of phenomena and provide not only an existence proof of model plausibility via data fitting, but a principled explanation of how and why the model performs the way it does. Thus, although much work remains in refining the biologically-plausible framework and the model of response selection, this approach promises to be of broad value to cognitive science.

## Acknowledgments

The bulk of the reported work was conducted as part of BCA's PhD thesis and was supported by the CMU Psychology Department and the Center for the Neural Basis of Cognition (CNBC). We thank Kevin Mickey for discussion and for developing a preliminary simulation as part of the CNBC Summer Research Program.

## References

Armstrong, B. C. (2012). The Temporal Dynamics of Word Comprehension and Response Selection: Computational and Behavioral Studies. Unpublished doctoral dissertation, Carnegie Mellon University Psychology Department.
Armstrong, B. C., Joordens, S., \& Plaut, D. C. (2009). Yoked criteria shifts in decision system adaptation: Computational and behavioral investigations. In N. A. Taatgen \& H. van Rijn (Eds.), Proceedings of the $31^{\text {st }}$ Annual Conference of the Cognitive Science Society (pp. 2130-2135). Austin, TX: Cognitive Science Society.
Baayen, R. H., Davidson, D. J., \& Bates, D. M. (2008). Mixedeffects modeling with crossed random effects for subjects and items. Journal of Memory and Language, 59(4), 390-412.
Hinton, G. E. (1989). Connectionist learning procedures. Artificial Intelligence, 40(1-3), 185-234.
Laszlo, S., \& Plaut, D. (2012). A neurally plausible parallel distributed processing model of event-related potential word reading data. Brain and Language, 120(3), 271-281.
Peirce, J. W. (2007). Psychopy-psychophysics software in python. Journal of Neuroscience Methods, 162(1-2), 8-13.
Plaut, D. C., \& Booth, J. R. (2000). Individual and developmental differences in semantic priming: Empirical and computational support for a single-mechanism account of lexical processing. Psychological Review, 107(4), 786.
Ratcliff, R. (1978). A theory of memory retrieval. Psychological Review, 85(2), 59-108.
Ratcliff, R., Van Zandt, T., \& McKoon, G. (1999). Connectionist and diffusion models of reaction time. Psychological Review, 106(2), 261-300.
Usher, M., \& McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. Psychological Review, 108(3), 550-592.

# The dynamics of pragmatic enrichment during metaphor processing: activation vs. suppression. 

John M. Tomlinson, Jr. (tomlinson@zas.gwz-berlin.de)<br>Zentrum für allegemeine Sprachwissenschaft (ZAS)<br>Berlin 10117 Germany

Stavros Assimakopoulos (stavros.assimakopoulos@um.edu.mt)<br>Institute of Linguistics, University of Malta<br>Msida MSD 2080 Malta


#### Abstract

In this paper, we test between suppression and activation accounts of metaphor processing by means of a novel metaphor interference paradigm that makes use of mouse-tracking. The goal is to understand how context influences the activation of salient and non-salient features of a concept during the on-line processing of a metaphor. In two mouse-tracking experiments, we examine the activation and availability of conceptual features that were either irrelevant or relevant for understanding a metaphor across various contexts. Our findings support the conclusion that context works primarily by rapidly suppressing salient features of a concept that are not relevant for the particular metaphorical interpretation. What is more, it seems that even further contextual manipulation does not facilitate the activation of non-salient metaphor relevant features.


Keywords: Figurative language, lexical pragmatics, psycholinguistics, mouse-tracking.

## Introduction

Among the most intriguing aspects of language are the creative ways in which we can use it. We often use words and sentences in novel ways that create nonce interpretations of a stable linguistic repertoire for various social and pragmatic purposes. One particularly contentious debate surrounding creative language use is how we understand figurative language. While much research has sought to understand the various linguistic and conceptual mechanisms needed to understand figurative meaning (Glucksberg, 1998, Lakoff \& Johnson, 1980), less attention has been paid to the pragmatic enrichment mechanisms needed to rapidly integrate contextual information into figurative interpretations (Gibbs, 2002).

In this paper, we address this question by using a novel mouse-tracking paradigm in an attempt to better understand how context affects the on-line processing of figurative language. Specifically, we examine metaphorical interpretations, which have received the greatest amount attention in the psycholinguistics literature, and report the results from two experiments designed to test between two contrasting views of how context can aid the comprehension of metaphors; either by suppressing information not relevant for the metaphorical interpretation or by increasing the accessibility of information needed to reach it..

## Metaphors and language processing

Many early experimental studies have demonstrated that metaphor processing is incompatible with the standard "literal-first" Gricean view of pragmatics (Gibbs, 2002). This is because figurative speech is understood just as quickly, and in some cases even more quickly, than literal speech (see Glucksberg, 2001 \& Glucksberg, 2003 for extensive reviews). Processing delays reported in the literature are more likely to occur for unfamiliar metaphors, due to the relative difficulty needed to integrate contextual information into the novel interpretation at hand (Shinjo \& Meyers, 1987), while reaction time differences might ultimately stem from speed-accuracy trade-off issues. McElree \& Nordlie (1999), for example, observed that literal and figurative meanings might be derived equally fast, but the overall accuracy of figurative speech interpretation seems to be lower than that of literal speech at both early and late processing times. These findings suggest that strong contexts should not decrease the processing time for figurative language, but only increases the likelihood that the intended interpretation will be understood.

The finding that both literal and metaphorical speech is understood with the same speed and facility has been a central tenet of numerous theoretical accounts in psychology. One example is the class inclusion model (Hampton, 1988; Glucksberg \& Keysar, 1990), according to which, understanding a metaphor such as "his job is a jail" amounts to updating the topic's category structure by integrating select features of the vehicle 'jail' (e.g. "confinement", "no exit", etc.) into it. Experimental work seems to suggest that this is done by first suppressing basic level meanings of a category (Gernsbacher, Keysar, Robertson, \& Werner, 2001) and then creating ad hoc categories from the combination of topic and vehicle features (Barsalou, 1983).

Another prominent account that incorporates the lack of processing time difference between literal and metaphorical speech is that of Cognitive Linguistics (Lakoff \& Johnson, 1983), which suggests that literal and metaphorical interpretation do not rely on different processing mechanisms. According to this approach, understanding
literal and metaphorical speech involves the recognition of patterns as opposed to complex mappings assumed in other theories; in the case of "my job is a jail", for example, both "job" and "jail" share the "container" relationship.

## Figurative language and lexical pragmatics

In linguistics, researchers have sought to integrate the findings mentioned above into both more formal and processing-oriented theories. For example, Relevance Theory currently offers an account of metaphor as a case of conceptual broadening and narrowing (Sperber \& Wilson 2008). Relying on the aforementioned idea of ad-hoc categories (Barsalou 1983), relevance theorists also hold that nonce conceptual categories are constructed for each contextually specific metaphorical use. For example, in a metaphor such as "the goalie is a spider", the vehicle 'spider' is assumed to encompass encyclopaedic information ("insect", "has 6 legs", "fangs", "catches prey in web", etc.), which is selectively combined into a contextually relevant ad-hoc category SPIDER* (denoting something like "catches things near its web") after certain features of the lexically encoded concept SPIDER are suppressed.

Other accounts seem to rely more on processes of activation when it comes to explaining how context helps process figurative language on-line. The direct access view, for instance, holds that context primes the relevant features needed to interpret the figurative meaning to the extent that figurative meanings are interpreted in an almost identical manner to literal meanings (Gibbs, 2002). This account differs from the relevance-theoretic one in that it predicts that context affects the way in which conceptual/semantic information is accessed from the start rather than further processed once a lexical meaning is retrieved during some decoding process. This view is consistent with Recanati (1995)'s theory of truth-conditional pragmatics, which holds that when interpreting a sentence such as "my job is jail", a strong biasing context would make the properties of the concept JAIL that are needed for the figurative interpretation directly available (e.g. "confined space" as opposed to "punishment").

Prior work investigating suppression and activation in metaphor processing has relied on various priming and lexical decision paradigms (see Glucksberg, 2003 for a review). For example, Rubio-Fernadez (2007) tested the idea of active suppression in a cross modal priming paradigm and found that after reading metaphorical utterances, metaphor relevant properties remained active at longer intervals than metaphor inconsistent properties. This suggests that all properties of a given concept are activated and then properties not relevant for the metaphor are suppressed. One limitation of such studies, however, is that reaction time measurements might mask the continuous changes in the accessibility of properties. In other words, similar reaction times at a given interval might not necessarily reflect equal activation levels. In the present
study, we make use of a novel paradigm to further test how suppression and activation mechanisms are implemented during on-line metaphor comprehension. We use mousetracking because via motor movements it can provide a window into the way in which listeners' access conceptual information during processing, since it "breaks up" the button press (Freeman \& Ambady, 2010). More specifically, we examine how both metaphor relevant and metaphor irrelevant features can interfere with mouse-trajectories toward metaphorical interpretation.

## Overview of Experiments

In the following two experiments, a metaphor interference paradigm (Glucksberg, Gildea, \& Bookin, 1982; Wolff \& Gentner, 2000) was combined with mouse-tracking. This paradigm offers a novel test of the on-line availability and competition of salient and contextually relevant features during the interpretation of a metaphor. Participants read metaphors such as "the goalie is a spider" along with filler items such as "the apple is red". They then clicked on either one of two pictures that best captured the overall meaning of the utterance. In the critical trials, one of these pictures, the correct target, depicted the topic in its metaphorical state (goalie making a save), while the other picture, a competitor image, was either an attribute from the vehicle available in the metaphorical state (spider web) or one only available in its non-figurative meaning (spider close-up). The interference of competitor target images on participants' mouse trajectories was compared across both literal and figurative utterances (Experiment 1) and then across strong and weak contexts (Experiment 2).
Prior to both of these experiments, participants completed a picture norming experiment to determine the relative salience of the various attributes for the metaphorical vehicle and become familiar with the pictures used in the main experiment. For this, they were shown the topic of the metaphor in isolation (SPIDER) and were asked to choose between the two competitor target pictures (web or closeup). This allowed for the experimental items (metaphorical utterances) to be separated into two groups:

1) Salient (metaphor) relevant feature group - the figurative attribute of the vehicle has more baseline salience than the literal attribute.
2) Salient (metaphor) irrelevant feature group - the literal attribute of the vehicle has more baseline salience than the figurative attribute.

## Experiment 1

28 participants at Cardiff University took part in both the picture-norming and main experiment in exchange for course credits. Participants completed both parts of the experiment within 30 minutes.
Stimuli Forty metaphorical sentences were adapted from Jones \& Estes (2006), in which both the topics and vehicles of which had been already been normed for aptness and conventionality. In Jones \& Estes (2006), aptness was
defined as the extent to which the vehicle's figurative meaning expresses an important feature of the topic. Conventionality was defined as the strength of the association between the metaphor vehicle and its figurative meaning. This allowed us to use a variety of metaphors that differed along these parameters.

Various pictures that depicted either relevant (spider web) or irrelevant (spider eyes/fangs) features of the metaphorical topic in this particular context were collected using Google Search. During the collection, multiple candidate pictures expressing the metaphorical meaning of the vehicle were collected (e.g. an outstretched goalie making a diving save, a tall goalie, a goalie stretching, etc.), and several research assistants independently decided which one best expressed the metaphorical meaning of the utterance at hand. The same pictures were used for both the norming and the main parts of experiment.

## Norming experiment

In order to establish the relative salience of the feature pictures prior to the main experiment, participants were presented with the metaphor topics (spider) and had to click on the feature picture that best represented the word. This allowed us to distinguish a relative baseline for which feature picture was more salient for the topic. For example, participants read the word VOLCANO (from "his anger is a volcano") and clicked on either a picture of an inactive volcano (metaphor irrelevant feature) or a picture of hot lava (metaphor relevant feature). In order for an item to be assigned to a group (salient metaphor relevant feature group or salient metaphor irrelevant feature group), the average ratings for a picture response for a given topic had to be above $66 \%$. Two items that were towards chance, i.e. 50/50, were excluded from the analysis of the main experiment. 18 of the 40 metaphor vehicles were rated as having a salient (metaphor) relevant feature and 20 of the 40 metaphor vehicles were rated as having a salient (metaphor) irrelevant feature. This part of the experiment also served as a way of familiarizing participants with the pictures used in the main experiment.

## Main experiment

After completing the norming study, participants were instructed to read a sentence and choose the picture that best corresponds to its overall meaning. For metaphors, participants had to click on the correct target, e.g. a picture of the metaphor topic in its figurative state (goalie making a save). Different feature pictures (metaphor relevant or metaphor irrelevant) were used as competitors to test 1) the amount of interference with the correct target and 2) the stage in processing (early or late) during which the interference occurs. Early interference would suggest that a feature is available during lexical access, whereas later interference would suggest that this feature becomes available at a later stage, i.e. during pragmatic enrichment.

Design and Procedure In the main experiment, participants read metaphors such as "The goalie is a spider" as well as literal filler items such as "The apple is red". The filler items were included to make sure that participants listened until the end of the sentence. Without these sentences, listeners could have made their decisions just after hearing the metaphor topic. In the filler sentences participants would choose between two competing images, which were only distinguished by the final word in the utterance ("The apple is red" vs. "The apple is green"). For the 40 metaphor items, participants saw either one of three target-competitor picture versions for metaphorical items. In the relevant feature condition, participants chose between a picture of the metaphor relevant feature (spider web) and a picture of the topic in its metaphorical state (the outstretched goalie making a save). In the irrelevant feature condition, the same picture of the topic was used, but in conjunction with a metaphor irrelevant feature picture (frontal close up of a spider). In the control condition, the same topic picture was also used, but the competitor picture had no relationship to the topic or vehicle (e.g. an apple). Three lists were created so that each participant saw only one version of each item. To start a trial, the participant would click on the START button at the bottom center of the screen. Each item was presented word by word at an interval of 350 ms per word. Participants were allowed to move the mouse towards a picture target, located at the top corners of the screen, only at the onset of the final word in the utterance. The trial ended once a target was clicked on.

Predictions If participants first access the salient meaning of the concept at hand, then salient picture features should interfere at an earlier point during the response. This should happen for both relevant and irrelevant features, however irrelevant features should not interfere later on in the response, as these features are not part of the figurative meaning, i.e. they become suppressed. Put differently, if a metaphor irrelevant feature is indeed the most salient, this should interfere with responses towards the correct target (picture of topic in metaphorical state) early on during the response because it would be "active" during lexical access. Similarly, if a metaphor relevant feature is not salient, later interference in responses should occur. When metaphor relevant features have higher salience than metaphor irrelevant features, both early and late interference should be noticed because this feature would be active both during lexical access as well as during the construction of the figurative meaning.

## Results and Discussion

Figures 1 and 2 show the mouse paths for correct responses for the relevant and irrelevant features groups, across salient and non-salient items respectively. Control groups were not included because mouse paths went directly to the target. To examine the relative interference of competitor pictures on participants' mouse paths to the correct targets, the average x-coordinates for mouse paths across the feature conditions
and saliency groups were compared. The time points for the x-coordinates from the normalized mouse paths (101 time stamps) were collapsed into 10 groups (or time bins) in order to better operationalize early vs. late processing. A mixed-model was used to analyze the x-coordinates of the mouse paths, which used time-bin, competitor (relevant vs. irrelevant), and feature salience as fixed effects. Subjects and items were used as random effects.
Interaction terms for feature competitors and feature salience were significant at time bins 30 to $40, t=2.83, p<$ .05 and time bins 50 to $60, t=2.51, p<.05$. Figure 2 shows the mouse paths for the relevant and irrelevant feature competitors only for items for which the irrelevant feature was rated as more typical of the metaphor vehicle. A crossover pattern for x -coordinate position is observed, in which metaphor irrelevant features interfere early on in the response (time bins 30-40), whereas the metaphor relevant feature interfere later on (time bins 50-60). This replicates several findings from different paradigms in that salient features seem to be more active, i.e. interfere more, during early processing and less salient features are accessed later, i.e. interfere later during the response. In the next experiment, we sought to replicate the cross-over pattern for feature interference and then investigate how priming versus neutral contexts mediate the availability of features when interpreting metaphors.

## Experiment 1 - Competitor x Likelihood



Figure 1. Time- normalized mouse paths for all conditions (competitor feature type vs. salience groups).

## Experiment 1 - IF High salience



Figure 2. Time- normalized mouse paths for relevant vs. irrelevant features in the salient irrelevant feature group.

## Experiment 2

In this experiment, prior to each metaphor, participants read either a "strong" and "neutral" preceding context. This allowed us to test whether contextual information can rapidly adjust the relative salience of a feature of a concept, as activation accounts would predict. Specifically, "strong" contexts should explicitly promote the status/activation of metaphor relevant features, whereas "neutral", albeit felicitous ones should not. The direct access view would predict that the (late) interference effect found for nonsalient metaphor relevant features should occur at an earlier time bin when preceded by strong contexts because of the increased activation of these features by the context. Suppression accounts, on the other hand, would predict that the early interference account for irrelevant features should be more diminished in strong when compared to weak contexts.

## Norming experiment

Prior to the main experiment, the same norming experiment as Experiment 1 was repeated. The two excluded items from Experiment 1 were rated as having high salience metaphor relevant features, however 1 of the remaining items was now excluded because it did not meet the $66 \%$ threshold. This resulted in 20 items in the high salience relevant feature group and 19 items in the high salience irrelevant feature group.

## Main experiment

The main experiment was identical to Experiment 1 except that items were preceded by either neutral or strong contexts. Details are discussed below.

Stimuli The stimuli were the same used in Experiment 1. The only difference was that items were preceded by a strong or weak context. For example, in our example item "the goalie is a spider", the strong context was: "We had many opportunities to score, but the ball was always
stopped" while the neutral one was: "We had many opportunities to score but couldn't convert our chances".

Design and Procedure The design and procedure were the same as in Experiment 1. The only difference was that contexts were presented prior to the onset of the metaphor or literal utterance. Participants were able to read each context at their own pace and press the enter key in order for the word-by-word presentation of the item to start.

## Results and Discussion

Figures 3 and 4 show mouse paths for the two conditions. Using the same time bins from Experiment 1, a mixed regression model was used to examine the interaction of context, feature competitor, and feature salience. When collapsed across contexts, the same interaction terms for feature competitor and feature salience were significant at time bins 30 to $40, t=1.89, p<.05$ and at time bins 50 to $60, t=3.07, p<.01$. When context was included in the interaction terms, they were not significant; time bins 30 to $40, t=95, p=.69$ and time bins 50 to $60, t=.34, p=71$. Context was, however, a significant predictor at time bins 30-40 for salient irrelevant features when the irrelevant features were the competitor, $t=2.61, p<.03$, but not at time bins 50-60, $t=.75, p=53$.
Experiment 2 replicated the findings from Experiment 1 in that early interference effects were observed for irrelevant features and late interference effects were observed for relevant features when the irrelevant feature was the most salient. Both neutral and strong contexts showed the crossover interaction, however strong contexts only reduced the early interference effects of irrelevant features for items for which irrelevant features were the most salient. Taken together, these findings provide evidence for suppression accounts because relevant features still showed late interaction effects even with strong contexts. This is not predicted by the direct access view.

## Experiment 2 - Weak Context



Figure 3. Time- normalized mouse paths for relevant vs. irrelevant features in the salient irrelevant feature groups in weak (neutral) contexts.

## Experiment 2 - Strong context



Figure 4. Time- normalized mouse paths for relevant vs. irrelevant features in the salient irrelevant feature groups in strong contexts.

> Exp 2 - IF competitor IF high salience


Figure 5. Time- normalized mouse paths for irrelevant features in the salient irrelevant feature groups across weak and strong contexts.

## Conclusion

In two mouse-tracking experiments, we tested the on-line availability of conceptual information when interpreting figurative language. Our main question was whether context acts primarily by suppressing context-independent features during lexical access for metaphorical interpretation or whether it increases the availability of non-salient features needed to understand the metaphor under question. Experiment 1 showed that salient metaphor irrelevant features of the vehicle provided early interference in participants' mouse paths towards correct targets. Nonsalient features of the vehicle relevant to the interpretation interfered much later on in the participants' mouse paths towards the correct target. Experiment 2 tested whether the
presence of a biasing context primes the activation of the salient feature that is relevant to the metaphor earlier on in processing, but the same cross-over pattern found in Experiment 1 was replicated in both the neutral context and the strong context conditions. What context seemed to do was reduce the early interference effect for metaphor irrelevant features when these same features were salient features of the vehicle. Taken together, the findings from Experiment 2 provide support for suppression accounts.

While our paradigm did not find increased activation for non-salient metaphor relevant features in strong contexts, one possibility is that our strong contexts might not have been strong enough to adequately test direct access accounts. We used one sentence introductory context, when classical studies, such as Swinney's (1979) "bug" task, use an entire paragraph of context. In this vein, Noveck, Bianco, \& Castry (2001) showed that longer contexts speed up metaphor processing in comparison to shorter ones. It may well be the case that the priming effects associated with activation accounts operate on a larger time scale than suppression mechanisms. We feel that this is a worthy topic for future research, for which our mouse-tracking paradigm is well suited to test.

## Acknowledgments

The second author was supported by ESRC grant RES-062-23-240 and all of the authors were supported by the ESF EuroXPrag grant "Varieties of Lexical Adjustment". We would like to thank Elena Chepucova for helping with the preparation of the pictures and audio files.

## References

Barsalou, L. W. (1983). Ad hoc categories . Memory \& Cognition,11, 211-27.
Gernsbacher, M. A., Keysar, B., Robertson, R. R. W., \& Werner, N. K. (2001). The role of suppression and enhancement in understanding metaphors. Journal of Memory and Language, 45, 433-450.
Gibbs, R. (2002). A new look at literal meaning in understanding what is said and implicated. Journal of Pragmatics, 34, 457-486.
Glucksberg, S. Gildea, P. \& Bookin, H. E. (1982). On Understanding nonliteral speech: Can people ignore metaphors? Journal of Verbal Learning and Verbal Behavior, 21, 85-98.
Glucksberg, S. \& Keysar, B. (1990). Understanding metaphorical comparisons: Beyond similarity. Psychological Review, 97(1), 3-18.
Glucksberg, S. (1998). Understanding Metaphors. Current Directions, 7, 39-43
Glucksberg, S. (2001). Understanding Figurative Language: From Metaphors to Idioms. New York: Oxford University Press.
Glucksberg, S. (2003). The Psycholinguistics of Metaphor. Trends in Cognitive Science, 7, 92-96

Hampton, J. (1981). An investigation of the nature of abstract and concepts. Memory \& Cognition, 9, 149-I 56.
Lakoff, G., \& Johnson, M. (1980). Metaphors we live by. Chicago: University of Chicago
McElree, B., \& Nordlie, J. (1999). Literal and figurative interpretations are computed in equal time. Psychonomic Bulletin \& Review, 6, 486-494.
Noveck, I., Bianco, M. \& Castry, A. (2001). The costs and benefits of metaphor. Metaphor and Symbol 16, 109-121.
Recanati, F. (1995). On the alleged priority of literal interpretation. Cognitive Science 19, 207-222.
Rubio Fernández, P. (2007). Suppression in metaphor interpretation: Differences between meaning selection and meaning construction. Journal of Semantics, 24, 345-371.
Shinjo, M. \& Myers, J., (1987). The role of context in metaphor comprehension. Journal of Memory and Language 26, 226-241.
Sperber, D., \& Wilson, D. (2008) A deflationary account of metaphors. In R. Gibbs The Cambridge Handbook of Metaphor and Thought. New York: Cambridge University Press.
Swinney, D. (1979). Lexical access during sentence comprehension: ( Re ) consideration of context effects. Journal of Verbal Learning and Verbal Behavior, 18, 645659
Wolff, P., \& Gentner, D. (2011). Structure-mapping in metaphor comprehension. Cognitive Science, 35. 14561448.

# Object motion continuity and the flash-lag effect 

Ricky K. C. Au ${ }^{1,2}$ (ricky@fennel.rcast.u-tokyo.ac.jp)<br>Katsumi Watanabe ${ }^{1}$ (kw@fennel.rcast.u-tokyo.ac.jp)<br>${ }^{1}$ Research Center for Advanced Science and Technology, The University of Tokyo, Tokyo, Japan<br>${ }^{2}$ Japan Society for the Promotion of Science, Tokyo, Japan


#### Abstract

When a visual object is briefly flashed, it appears to lag behind another moving object (flash-lag effect; FLE). Previous studies show that a sudden change to the moving object at the time of the flash presentation can eliminate the FLE. We examined whether the FLE is eliminated when a moving object alternates in color as it moves. Observers viewed a moving disc, the color of which did not change at all, changed only once when another object flashed, or alternated between two colors as it moved before the flash presentation. The results showed that although the magnitude of the FLE was reduced compared with the no-change condition, the FLE observed with the moving object that changed color during motion was significantly stronger than the FLE in the onechange condition. The results are discussed in relation to the object updating account of the FLE.


Keywords: Flash-lag effect; Motion continuity; Object updating

## Introduction

Humans depend heavily on the perceptual system to collect information about the surrounding environment, but the perceptual system is sometimes prone to illusions that lead to inaccurate judgments. In the domain of object localization, one extensively studied illusion is the flash-lag effect (FLE), a perceptual phenomenon where a briefly-flashed stationary object appears to lag behind another moving object even though the two objects are physically aligned when the flash occurs (MacKay, 1958; Nijhawan, 1994). Studies on the FLE have found that this effect occurs in various conditions. For example, the FLE has been reported in objects with continuously changing features (Sheth, Nijhawan, \& Shimojo, 2000), in objects moving in depth (Harris, Duke, \& Kopinska, 2006; Ishii, Seekkuarachchi, Tamura, \& Tang, 2004), in audition, and across modalities (Alais \& Burr, 2003). In addition, the FLE was also found to depend on observers' eye movements (Nijhawan, 2001) and the perceptual organization of the moving object (Watanabe, 2004; Watanabe, Nijhawan, Khurna, \& Shimojo, 2001). Putting the effect in a two-dimensional context, Watanabe and Yokoi (2006) found that the perceived position of the flash is not uniformly displaced, but appears to converge towards a single point behind the position of the moving object.
Ever since Nijhawan (1994) revitalized interest in the FLE within the psychology community, various explanations have been formulated to account for the effect, including motion extrapolation (Nijhawan, 1994, 1997), differential latency in processing for the flashed object and the moving object (Kanai, Carlson, Verstraten, \& Walsh,

2009; Whitney \& Murakami, 1998; Whitney, Murakami, \& Cavanagh, 2000), motion integration and postdiction (Eagleman \& Sejnowski, 2000), and attention (Baldo \& Klein, 1995; Krekelberg \& Lappe, 2000). However, this ongoing debate has not yet been settled.

Moore and Enns (2004) proposed a relatively new explanation of the FLE. They view the effect as the result of an ongoing object updating process based on the principle of object substitution (Enns \& Di Lollo, 1997). They proposed that due to the ongoing updating process, positional information of the moving object acquired immediately after the flash presentation overwrites (replaces) that acquired at the time of the flash presentation, resulting in the illusory perception that the moving object overshoots the flash. In the case where the moving object stops at the time of the flash presentation, since there is no new information about the moving object after the flash presentation that can replace (update) previous information, the alignment of the two objects can be accurately perceived. In the same study, Moore and Enns (2004) further reported that when the visual features of the moving object, such as size and color, changed abruptly at the moment of the flash presentation and changed back immediately after the flash (we refer to this as the "One Change" motion stream), observers tended to perceive that the moving object appeared at two positions (one object with the changed color and aligned with the flash, and the other with the original color located in front of the flash) when asked about the perception at the moment of the flash presentation. The authors explained that the disruption of motion continuity by a large, transient change leads the visual system to interpret the scene as containing two separate objects. When the original object reappears at a new position after the flash, its position and color information is updated, while the information acquired at the moment of the flash presentation (which is interpreted as a different object) is spared from the overwriting process. However, if a scene-based reason is provided for the discontinuity, the object updating process is spared from disruption, preserving the representation of the original object, and thus, the FLE is observed (Moore, Mordkoff, \& Enns, 2007).

According to the idea above, whether object motion continuity is preserved depends on whether only a single (i.e., the same) object is identified throughout the motion scene. The nature of object persistence has been widely studied based on object file theory (Kahneman, Treisman, \& Gibbs, 1992). According to this theory, episodic representations (object files) keep track of the individual entities in a scene over space and time, and are updated
based on spatiotemporal information (i.e., location at different moments). Object files store the representations of persistent objects and mediate conscious perception, informing the observer about "which went where" (Mitroff, Scholl, \& Wynn, 2005), and object identity information can be stored on a scale of seconds (Noles, Scholl, \& Mitroff, 2005). Empirical evidence has suggested that object files encode identity information rather than semantic or precise physical information (i.e., physical features) about objects, and that object file representations are flexible (Gordon \& Irwin, 1996, 2000). Although Mitroff and Alvarez (2007) showed that spatiotemporal information, but not surface features, effectively determines object persistence (as measured by standard object-specific preview benefits; Kahneman et al., 1992), Moore, Stephens, and Hein (2010) demonstrated that abrupt changes in surface features disrupt preview benefits, and an object feature alone could determine object persistence under some conditions. It is therefore still unclear what role object surface features play in the establishment and maintenance of object files.
An interesting question derived from the study of Moore and Enns (2004) is what would be observed if a stream of events consisted of an object moving in a uniform trajectory while its surface feature (e.g., color) keeps changing? This would represent a case in which spatiotemporal continuity suggests only a single object moving throughout the journey, but the information from surface features suggests that multiple units exist. In the present experiment, we investigated this question by introducing two conditionsAlternating stream (in which the color of the moving object alternates between two colors) and Random stream (in which the color of the moving object changes randomly between two colors) - in addition to the One Change and No Change conditions employed in the original study by Moore and Enns (2004). Based on previous work on object file theory, if spatiotemporal information dominates the formation and updating of episodic object files (so that the visual system identifies only one object in the stream), we would expect the FLE to occur even in the Alternating and Random stream conditions. This would also mean that the unexpected and highly salient change at the moment of flash presentation in the One Change stream is a necessary condition for breaking motion continuity (leading the visual system to identify multiple objects in the stream) which eliminates the FLE. In contrast, if object surface features play a significant role in maintaining object files, the history of color change in conjunction with motion would cause the visual system to conclude that multiple objects exist in the motion stream. In this case, the FLE might be eliminated because the process of overwriting previous information at each instant is largely disrupted by the color change.

## Method

To examine the effect of object motion continuity on the magnitude of the FLE, we compared performance across three motion stream conditions (No Change, One Change,
and Alternating or Random) in two separate sessions with two different groups of observers.

## Participants

Twenty-four paid volunteers recruited at The University of Tokyo participated as observers in the experiment. Twelve observers were assigned to the session with the Alternating stream condition, and twelve were assigned to the session with the Random stream condition. All were naïve as to the purpose of the study and had normal or corrected-to-normal vision. Informed consent was obtained from the observers prior to the experiment.

## Stimuli and procedures

The stimuli used in the experiment were developed based on the previous study by Moore and Enns (2004; Part 2), and were programmed in MATLAB R2012b (MathWorks, USA) using the Psychophysics Toolbox extension (version 3.0.8; Brainard, 1997; Pelli, 1997). The stimuli were displayed on a CRT monitor with a refresh rate of 100 Hz (resolution $=800 \times 600$ pixels), controlled by a personal computer running the Windows 7 operating system. Observers viewed the stimuli at a distance of 60 cm in a dark and quiet environment.

All experimental stimuli were presented on a black background (luminance $=0.022 \mathrm{~cd} / \mathrm{m}^{2}$ ). The observer initiated each trial by pressing the space bar on the keyboard. After the space bar was pressed, a white fixation cross consisting of one horizontal line and one vertical line (length $=0.317^{\circ}$, width $=0.0453^{\circ}$ ) appeared at the center of the screen and remained throughout the trial until a response was made. Observers were required to fixate on the fixation cross throughout the trial. When the trial was initiated, a circular target stimulus (diameter $=0.907^{\circ}$ ) in either red or green (luminance $=0.47 \mathrm{~cd} / \mathrm{m}^{2}$ ) appeared either just above or below the central fixation cross at a distance of $4.171^{\circ}$ and remained there for 500 ms . Then, the target stimulus started to move in clockwise or counter-clockwise direction on an imaginary circle (radius $=4.171^{\circ}$ ) around the fixation cross for a random angular distance of $105^{\circ}, 195^{\circ}, 285^{\circ}$, or $375^{\circ}$ at an angular speed of $15^{\circ} /$ frame. Each frame was displayed for 70 ms , and thus, the duration of the motion stream was $490 \mathrm{~ms}, 910 \mathrm{~ms}, 1330 \mathrm{~ms}$, or 1750 ms . One of the following three possible motion streams was presented on each trial: (i) No Change, (ii) One Change, and (iii) Alternating or Random (depending on session assignment). In the No Change stream, the color of the target remained unchanged throughout the trial. In the One Change stream, the target color changed to the other color during the second last frame of the motion (which corresponds to the position just above, below, to the left, or to the right of fixation, and thus is always aligned with fixation), and then changed back to its original color in the last frame of the motion. In the Alternating stream, the color of the target alternated between red and green in each frame of the motion. In the Random stream, the color of the target changed randomly (either red or green) in each frame of its motion.

The flash stimulus was a white disc (diameter $=0.544^{\circ}$, luminance $=2.89 \mathrm{~cd} / \mathrm{m}^{2}$ ) presented at the position just above, below, to the left, or to the right of fixation (i.e., always aligned with fixation) at a distance of $2.901^{\circ}$. The flash was presented at either the third last, second last, or last frame of the motion for a duration of one frame. These three flash conditions resembled the "behind," "aligned," and "ahead" conditions in Moore and Enns (2004; Fig. 1a and 1b). In addition to these three flash conditions, there were also two baseline flash conditions for each stream condition. In the previous study, when the flash appeared, the target disc was presented at the second last position of the motion in the No Change condition, and was presented at the second last and last position of the motion in the One Change condition (Moore \& Enns, 2004; Fig. 1c). However, in the present study, we included both of these baseline conditions in all stream conditions to reduce any possible difference or bias in the magnitude of the FLE elicited by the different baseline conditions in the No Change and One Change streams, thus allowing a better comparison across different stream conditions. Specifically, in the Baseline 1 condition, the target stimulus stream was identical to the "aligned" flash condition, except that the target disappeared along with the flash; in the Baseline 2 condition, the target stream was the same as the Baseline 1 condition, except that an additional target was also presented in the second last frame and disappeared along with the flash. This additional target was presented at the position where the target should appear in the last frame in a non-baseline condition (see the "small change" and "large change" conditions in Fig. 1c of Moore \& Enns, 2004). Therefore, in the two baseline conditions, the target discs were presented up to the second last frame of the motion stream, and only the central fixation cross was displayed in the last frame.


Figure 1: The four motion stream conditions employed in the present study.

Observers were required to judge, upon the disappearance of the target disc, whether the target disc was aligned with the flash (and also the fixation) at the moment when the flash occurred. They were also instructed to respond "aligned" if they saw two target discs and either one of them was aligned with the flash. There were a total of 480 trials (3 Streams conditions $\times 5$ Flash conditions $\times 2$ Starting Positions $\times 4$ Travel Distances $\times 2$ Starting Colors $\times 2$ Motion Directions). Observers were instructed to take a five-minute break halfway through the experiment. The experimental session took about 35 minutes to complete.

## Results

Following Moore and Enns (2004), we plotted the average proportion of trials where the observers reported that the target disc and the flash were aligned for each stream condition. The data are plotted separately for the sessions with Alternating and Random streams (Figure 2; only data for the two baseline flash conditions and the flash condition where the target disc and the flash were physically aligned are shown).

Separate omnibus repeated measures ANOVAs were conducted on the data in the Alternating and Random sessions of the experiment. In the Alternating session, the main effect of Flash condition $[F(2,22)=25.777, p<.001]$, the main effect of Stream condition $[F(2,22)=33.997, p$ $<.001]$, and the Flash $\times$ Stream interaction $[F(4,44)=9.685$, $p<.001]$ were all statistically significant. The main effect of Travel Distance was not significant $[F(3,33)=2.630, p$ $=.066]$. Specific comparisons revealed that when the target disc and the flash were physically aligned (i.e., Aligned in Figure 2), there was a significantly lower proportion of "aligned" responses [i.e., P("aligned")] in the No Change condition compared to the Alternating condition, while there was a significantly higher proportion of "aligned" responses in the One Change condition compared to the Alternating condition (both at $p<.01$, adjusted for multiple comparisons). No significant difference in proportion was found between the three stream conditions in the Baseline 1 condition; a significant difference in the proportion of "aligned" responses was found between the No Change vs. One Change, and between the No Change vs. Alternating conditions (both at $p<.01$ ) in the Baseline 2 condition.

Similar results were found in the Random stream session. The main effect of Flash condition $[F(2,22)=11.581, p$ $<.001]$, the main effect of Stream condition $[F(2,22)=$ 14.137, $p<.001]$, and the Flash $\times$ Stream interaction $[F(4,44)=6.795, p<.001]$ all reached statistical significance. The main effect of Travel Distance was marginally significant $[F(3,33)=2.927, p=.048]$, while pairwise comparisons showed that the four Travel Distance conditions did not differ significantly from each other. Specific comparisons showed that when the target disc and the flash were physically aligned (i.e., Aligned in Figure 2), the proportion of "aligned" responses was significantly lower in the No Change condition compared to the Random condition, whereas there was a significantly higher
proportion of "aligned" responses in the One Change condition compared to the Random condition. Similar to the Alternating session, no significant difference in response proportion was found among the three stream conditions in the Baseline 1 condition; there was a significant difference between the No Change vs. One Change, and between the No Change vs. Random stream conditions (both at $p<.01$ ) in the Baseline 2 condition.
The Baseline 1 condition appeared to more strongly eliminate the FLE than the Baseline 2 condition did. One possible reason for this difference is that, since the experiment was mixed with both baseline conditions, observers were aware that there was a condition where the target disc and the flash were obviously aligned and disappeared together (Baseline 1), possibly leading to lower confidence reporting alignment in the Baseline 2 condition, where there were two discs in different positions.
To summarize, the two sessions of the experiment replicated the finding that inserting a single change in an object's appearance during motion (i.e., the One Change stream) eliminated (or greatly attenuated) the FLE compared with the No Change stream. Furthermore, our experiments demonstrated that a motion stream where the object alternates colors or changes color randomly elicits some degree of FLE. These results imply that (a) the weakened FLE in the Alternating and Random streams may be due to impaired perceptual smoothness of motion compared to the No Change stream, and (b) elimination of FLE in the One Change stream may be due to the exceptionally high salience of the target disc during the second last frame of the motion; in the Alternating and Random streams, the disc may no longer be salient at the moment of flash presentation (cf. the One Change stream) because the surface feature is continuously changing throughout the disc's motion, leading to survival of FLE under these conditions.

To verify these two hypotheses, we conducted short control experiments with five additional observers, where they were requested to judge the smoothness of the motion stream or the salience of the target disc during the second last frame of the motion. In each trial in the sessions where smoothness of motion was evaluated, one No Change stream and one Alternating stream (or a Random stream in a separate session) were presented sequentially in a random order, and observers were asked to indicate which of the two motion streams exhibited greater smoothness in motion. There were 24 trials in each session. In most of the trials, the observers reported that the No Change stream was more smooth than either the Alternating and Random streams (average percentage of trials in which the No Change stream was judged as more smooth in comparison to the Alternating stream $=84.2 \%$, Random stream $=84.2 \%$ ). The sessions testing target disc salience during the second last frame of the motion were conducted in a similar manner, but a One Change stream was presented instead of a No Change stream. Observers were asked to judge which of two sequentially presented streams showed a more salient target disc during the second-last frame of the motion. The
observers judged the target disc to be more salient in the One Change stream compared to the Alternating (85.8\%) and Random (90.8\%) streams. The control experiments therefore suggest that both hypotheses (a) and (b) contribute to explain the reduced, but not eliminated, FLE in the Alternating and Random conditions.


Figure 2: The average proportion of trials the observers reported alignment of the target disc and the flash stimuli for the Alternating (upper graph) and the Random (lower graph) sessions of the experiment; error bars represent the standard error of the mean.


Figure 3: The average proportion of trials that the observers reported the motion of the No Change stream appears to be more smooth than the Alternating/Random stream (left panel), and that the target disc looks more salient at the second last frame in the One Change stream than the
Alternating/Random stream (right panel); error bars represent the standard error of the mean.

## Discussion

The results of the present study showed that under conditions where the target object kept changing color while moving in a uniform trajectory (i.e., Alternating and Random streams), a significant FLE was observed, although it was somewhat attenuated compared to the No Change condition. Furthermore, the results of the control experiments suggested that (a) the attenuation of the FLE under those conditions might be due to lower perceived motion smoothness compared to the No Change stream, and (b) the high salience of the target disc at the moment when the flash occurred might be responsible for the elimination of the FLE in the One Change stream. These results therefore suggest that smooth motion defined by unchanged physical surface features is not a necessary condition for the FLE. As long as the visual system identifies a single entity throughout motion, without a salient transient change (i.e., in the Alternating and Random streams), the observer can still perceive the FLE as in physically smooth motion. A highly salient change that occurs unexpectedly (i.e., in the One Change stream) is required to break continuity and cause the visual system to perceive multiple objects in the stream.
In the context of the FLE, the present results support the notion that spatiotemporal continuity dominates surface feature in processing object persistence (Mitroff \& Alvarez, 2007). Although under some conditions, surface features can guide the mapping and updating of individual objects (Moore et al., 2010), spatiotemporal information is weighted more strongly in the computation of object persistence when both types of information are available (Tas, Dodd, \& Hollingworth, 2012). A brain imaging study by Yi et al. (2008) also provides strong evidence that discontinued spatiotemporal trajectories can cause visually identical faces to be represented as different individual objects, in which the brain area involved was the most staunchly "featural" area of the ventral visual cortex. The determination of object persistence during object motion involves identifying the correspondence between objects over short periods. This is similar to how the visual system computes motion correspondence in the apparent motion phenomenon, in which solutions are sometimes needed to map multiple objects at one instance to multiple objects at other locations at the next instance; in such a case, spatiotemporal information plays an important role in assisting the visual system to arrive at an appropriate solution (Dawson, 1991).

From the results of the control experiments, we infer that the perceived smoothness of object motion and the salience of the transient change during motion mediate the magnitude and determine the survival of the FLE. Our results suggest that observers' subjective perception of smoothness was related to the magnitude of the FLE. In the Alternating and the Random conditions, observers reported less motion smoothness compared to the No Change condition, and the results of the main experiment indicated a significantly smaller FLE in the Alternating and Random conditions compared to the No Change condition. This is
consistent with previous findings that perceived motion smoothness (i.e., sampling rate of the motion trajectory) and the magnitude of the FLE are highly correlated (Khurana, Nijhawan, \& Watanabe, 1998). Such a relationship between motion smoothness and the magnitude of the FLE implies that the maintenance of object files that give rise to the FLE may be associated with smoothness of motion. In the context of the present study, the rapid change in physical features in the Alternating and Random streams impaired perceived motion smoothness, and the maintenance of object files was thus degraded, leading to a smaller FLE. Although the maintenance of object files was interrupted, the visual system still perceived only one object in the motion stream. In terms of the salience of the transient change at the time of the flash, our results are consistent with the proposal of Moore and Enns (2004) that the FLE depends on such a salient and unexpected change in smooth motion, as abrupt changes in object features may disrupt object representations (Moore et al., 2007). One possibility is that the salient and unexpected change in the One Change stream captured observers' attention. At the moment of flash onset, the abrupt change in the moving object increases attention and allows the moving object to be associated with the flash onset at its veridical position, sparing it from the FLE. In the Alternating and Random streams, since the color change was ongoing, any change would become less salient and less able to capture attention, thus preserving the FLE.

To summarize, the present study extended the results of previous FLE experiments (e.g., Moore and Enns, 2004) and showed that the FLE can occur in motion streams where the physical features of the moving object continuously change during motion. The magnitude and survival of the FLE was determined by perceived motion smoothness and the salience of the moving object at the time of the flash. We propose that a rapid change in a physical feature partially degrades the maintenance of the object file, but does not eliminate the overall percept of only one object in the motion stream. At the same time, it mostly reduces the salience of the disc at the moment of flash presentation. Future studies should focus on how attention at the moment of the flash influences the FLE.

## Acknowledgments

This work was supported by Grant-in-Aid for Scientific Research (KAKENHI) from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan; and the Japan Society for the Promotion of Science (JSPS).

## References

Alais, D., \& Burr, D. C. (2003). The "Flash-Lag" effect occurs in audition and cross-modally. Current Biology, 13, 59-63.
Baldo, M. V., \& Klein, S. A. (1995). Extrapolation or attention shift? Nature, 378, 565-566.
Brainard, D. H. (1997). The Psychophysics Toolbox. Spatial Vision, 10, 433-436.

Dawson, M. R. W. (1991). The how and why of what went where in apparent motion: Modeling solutions to the motion correspondence problem. Psychological Review, 98, 569-603.
Eagleman, D. M., \& Sejnowski, T. J. (2000). Motion integration and postdiction in visual awareness. Science, 287, 2036-2038.
Enns, J. T., \& Di Lollo, V. (1997). Object substitution: A new form of masking in unattended visual locations. Psychological Science, 8, 135-139.
Gordon, R. D., \& Irwin, D. E. (1996). What's in an object file? Evidence from priming studies. Perception and Psychophysics, 58, 1260-1277.
Gordon, R. D., \& Irwin, D. E. (2000). The role of physical and conceptual properties in preserving object continuity. Journal of Experimental Psychology: Learning, Memory, and Cognition, 26, 136-150.
Harris, L. R., Duke, P. A., \& Kopinska, A. (2006). Flash lag in depth. Vision Research, 46, 2735-2742.
Ishii, M., Seekkuarachchi, H., Tamura, H., \& Tang, Z. (2004). 3D flash lag illusion. Vision Research, 44, 19811984.

Kahneman, D., Treisman, A., \& Gibbs, B. J. (1992). The reviewing of object files: Object-specific integration of information. Cognitive Psychology, 24, 175-219.
Kanai, R., Carlson, T. A., Verstraten, F. A. J., \& Walsh, V. (2009). Perceived timing of new objects and feature changes. Journal of Vision, 9(7):5, 1-13, http://journalofvision.org/9/7/5/, doi:10.1167/9.7.5.
Khurana, B., Nijhawan, R., \& Watanabe, K. (1998). Extrapolation of stroboscopically moving objects. Presented at the Association for Research in Vision and Ophthalmology Annual Meeting, Fort Lauderdale, Florida, USA.
Krekelberg, B., \& Lappe, M. (2000). A model of the perceived relative positions of moving objects based upon a slow averaging process. Vision Research, 40, 201-215.
MacKay, D. M. (1958). Perceptual stability of a stroboscopically lit visual field containing self-luminous objects. Nature, 181, 507-508.
Mitroff, S. R., \& Alvarez, G. A. (2007). Space and time, not surface features, guide object persistence. Psychonomic Bulletin and Review, 14, 1199-1204.
Mitroff, S. R., Scholl, B. J., \& Wynn, K. (2005). The relationship between object files and conscious perception. Cognition, 96, 67-92.
Moore, C. M., \& Enns, J. T. (2004). Object updating and the flash-lag effect. Psychological Science, 15, 866-871.
Moore, C. M., Mordkoff, J. T., \& Enns, J. T. (2007). The path of least persistence: Object status mediates visual updating. Vision Research, 47, 1624-1630.
Moore, C. M., Stephens, T., \& Hein, E. (2010). Features, as well as space and time, guide object persistence. Psychonomic Bulletin and Review, 17, 731-736.
Nijhawan, R. (1994). Motion extrapolation in catching. Nature, 370, 256-257.

Nijhawan, R. (1997). Visual decomposition of colour through motion extrapolation. Nature, 386, 66-69.
Nijhawan, R. (2001). The flash-lag phenomenon: Object motion and eye movements. Perception, 30, 263-282.
Noles, N. S., Scholl, B. J., \& Mitroff, S. R. (2005). The persistence of object file representations. Perception and Psychophysics, 67, 324-334.
Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. Spatial Vision, 10, 437-442.
Sheth, B. R., Nijhawan, R., \& Shimojo, S. (2000). Changing objects lead briefly flashed ones. Nature Neuroscience, 3, 489-495.
Tas, A. C., Dodd, M. D., \& Hollingworth, A. (2012). The role of surface feature continuity in object-based inhibition of return. Visual Cognition, 20, 29-47.
Watanabe, K. (2004). Visual grouping by motion precedes the relative localization between moving and flashed stimuli. Journal of Experimental Psychology: Human Perception and Performance, 30, 504-512.
Watanabe, K., Nijhawan, R., Khurna, B., \& Shimojo, S. (2001) Perceptual organization of moving stimuli modulates the flash-lag effect. Journal of Experimental Psychology: Human Perception and Performance, 27, 879-894.
Watanabe, K., \& Yokoi, K. (2006). Object-based anisotropies in the flash-lag effect. Psychological Science, 17, 728-735.
Whitney, D., \& Murakami, I. (1998). Latency difference, not spatial extrapolation. Nature Neuroscience, 1, 656657.

Whitney, D., Murakami, I., \& Cavanagh, P. (2000). Illusory spatial offset of a flash relative to a moving stimulus is caused by differential latencies for moving and flashed stimuli. Vision Research, 40, 137-149.
Yi, D.-J., Turk-Browne, N. B., Flombaum, J. I, Kim, M.-S., Scholl, B. J., \& Chun, M. M. (2008). Spatiotemporal object continuity in human ventral visual cortex. Proceedings of the National Academy of Sciences of the United States of America, 105, 8840-8845.

# Interface Design and Spatial Cognition: A Case of Virtual Molecule Manipulation 

Trevor J. Barrett (trevor.barrett@psych.ucsb.edu)<br>Department of Psychological and Brain Sciences, UCSB Santa Barbara, CA 93106 USA<br>Mary Hegarty (mary.hegarty@ psych.ucsb.edu)<br>Department of Psychological and Brain Sciences, UCSB<br>Santa Barbara, CA 93106 USA


#### Abstract

Virtual models are a common instructional tool used in chemistry education to help students learn about the 3D structure of molecules. The present study examined effects of two interface design features on participant performance during a molecule orientation task. The features examined were 1) colocation of the visual and haptic workspace and 2) stereoscopic viewing. The results indicate that colocating the interface increased participant accuracy, while providing stereo did not. Neither factor affected response time. The effects of colocation were also reflected in subjective ratings of task demand measured by the NASA-TLX. Spatial ability was predictive of task performance but did not interact with interface effects. The findings are discussed in the context of spatial cognition and interface design for manipulating virtual objects.


Keywords: spatial cognition; interface; virtual; rotation; stereoscopic; colocation; organic chemistry

Computer-based virtual models are becoming an increasingly common instructional medium in science, technology, engineering, and mathematics (STEM) education (Trindade, Fiolhais \& Almeida, 2002). Virtual learning environments have shown promise in fostering meaningful learning, but virtual models vary considerably in the perceptual cues and interfaces that they provide so there is still much to be understood regarding how to best design and implement these technologies. For example, current stereoscopic displays are more expensive and less available, so it is important to know whether they provide a benefit to performance and learning outcomes, or whether monoscopic displays are as effective.

The present study aims to understand the relative value of two factors on which virtual displays vary, (1) colocating the hand-held interface and the displayed virtual image (colocation), and (2) providing stereoscopic 3D viewing (stereo) for a representation matching task in organic chemistry. Stull, Barrett, \& Hegarty (2012), found participants performed this task with greater efficiency using a virtual model system (with stereoscopic display and colocation of visual and haptic workspaces) than when using standard concrete models. Given this efficiency advantage, a goal of this study is to investigate the relative importance of providing stereo and colocation during a virtual object manipulation task.

Klatzy, Wu, \& Stetten (2008) suggest that more perceptually mediated interfaces allow for better performance over cognitively mediated interfaces. Perceptually mediated interfaces decrease demand on spatial working memory, thereby freeing up cognitive resources to allocate to performance or learning. If stereo and colocation increase perceptual mediation, decrease spatial cognitive load, and allow for additional cognitive recourses to be devoted to performance, then participants should show faster and more accurate performance when these cues are provided.

## Stereo and Colocation Technology

Both stereo and colocation technologies have been shown to increase speed and accuracy in virtual object manipulation tasks (Ware \& Rose, 1999, Arsenault \& Ware, 2004, Klatzky et al., 2008), however some studies have shown no significant effect of stereo (Khooshabeh \& Hegarty, 2010) or colocation (Liere, 2005). It is important to note that the majority of studies investigating performance with these technologies have used object translation tasks, rather than a rotation task as in the current study. In general, we should be cautious in generalizing specific interface design effects across various tasks, as different perceptual cues may be important for supporting rotation and translation.

To our knowledge, no study has investigated the effects of varying both stereo and colocation during a virtual object rotation task in the same experiment. The results of the present study will help to elucidate the importance of these cues for increasing perceptual mediation and decreasing cognitive load in a task that involves virtual object rotation.

## Organic Chemistry as a Test-bed

Organic chemistry is a domain rich in spatial representation; diagrams and models of 3D molecular structures are ubiquitous in instruction as well as in cutting edge research environments. Understanding molecular structure is an essential skill all organic chemists must have in order to learn, research, and communicate their science. Diagrams and models serve as a language of spatial connections and structures and therefore are vital in developing understanding of structures and making advancements in the field (Kozma \& Russell, 2005).

Organic chemistry representations serve as an excellent test-bed for studying spatial cognition in virtual environments for these reasons. Virtual models are commonly employed in chemistry education (Barnea \& Dori, 2000, Limniou, Roberts \& Papadopoulos, 2008) to teach students about the 3D structure of molecules and to introduce them to other representations (diagrams, equations etc.). However, little work has investigated interface design for virtual models in chemistry. One aim of the present study is to inform the design of virtual models for chemistry education, by identifying which visual cues and interface functionality best support usability and learning with these models.

The task used in the present study requires that the participant understand the spatial structure of a 3D molecular model, then manipulate the model in order to match to the orientation of a simultaneously displayed diagram of the same molecule. This task is relevant as it is a commonly employed activity for teaching about molecular structure, and also shares similarities with the virtual object orientation matching tasks used in the human-computer interaction literature. Grounding the study in the real-world domain of organic chemistry allows for simultaneous investigation of applied interface design issues as well as theories of small scale spatial cognition and virtual object manipulation.

In addition to performance measures, we assessed selfreports of usability of the virtual models. We predicted that participants who received stereo and colocation would rate the interfaces as more usable. Given the spatial nature of the task and known sex differences in spatial ability (Voyer, Voyer \& Bryden, 1995), we also investigated possible interactions between aspects of the virtual models, spatial ability, and sex. If colocation and stereo displays increase perceptual mediation, we might expect an interaction with spatial ability such that lower spatial ability participants should receive a greater benefit from the additional cues than higher spatial ability participants, who can presumably better handle more cognitively mediated interfaces. We might also expect females to benefit more from the perceptually mediated interface, because they tend to have lower spatial ability (Voyer et al, 1995) and less experience with computers (Waller, 2000).

## Virtual Model System

A 'fishtank' virtual reality system was constructed to allow for colocated naturalistic manipulation of a virtual molecular model in stereoscopic 3D (Earnst \& Banks, 2002). The display was mounted horizontally above the user and faced downward onto a mirror mounted at $45^{\circ}$, which projected the virtual image to the viewer. This configuration allowed the participant to manipulate the input device in the same location as the perceived virtual image of the model, giving an experience similar to direct manipulation of a concrete model. In the displaced condition, the input device
was located to the left and below the image in the natural computer mouse location ( 15 " total displacement). Stereoscopic viewing was provided by Nvidia 3D Vision Wireless Glasses Kit.

The interface was composed of a cylinder that was roughly the same dimensions as the virtual models, and consisted of two halves that freely rotated about the long axis of the interface. One half contained a 3-degree of freedom motion sensor to track yaw, pitch, and roll of the interface, and was used to control global rotations of the virtual models. The opposite half was attached via an optical encoder that tracked twisting rotations of the interface halves, and was used to control local rotations of a bond within the molecule itself (as was necessary on some of the experimental trials). Please refer to Stull, Barrett \& Hegarty (2012) for a more detailed description of the system design.


Figure 1: a) The hand-held interface workspace was colocated with the displayed virtual image. b) The motion sensor is depicted in blue, and the optical encoder is depicted in red. The cords for the two devices emerged at the junction between the two halves.

## Method

## Design

The study had a two (colocation vs. displaced) by two (stereo vs. mono) between subjects design. Dependent variables include accuracy as measured by angular error and response time. Subjective experience ratings, spatial ability, and computer use were also measured.

## Participants

One hundred twenty college students (65 Female) (age: $M=$ 18.7, $S D=1.8$ ) from the psychology subject pool at a research university participated in the study in return for course credit. None of the participants had studied organic chemistry. All participants had normal, or corrected to normal vision. Participants were randomly assigned to each condition.

## Materials

The study materials included an informed consent sheet, a video tutorial, a sheet with descriptions of the task and diagrams, a set of diagram orientation matching task problems, a measure of task load, a measure of computer beliefs and attitudes, two measures of spatial ability, and a post-task questionnaire.

A 10-minute instructional video explained the conventions of the models and diagrams, how to find and understand important features of the model (e.g., central carbon-carbon bond), how to write the chemical formula for each molecular subgroup (e.g., $\mathrm{CH}_{3}$ for a methyl group made up of a carbon atom and three hydrogen atoms), the color conventions for the different atoms (e.g., black for carbon, red for oxygen etc.), and how to structurally align the models to each of the three diagram types.

The diagram problems required rotation of the virtual model to match one of two commonly used target diagram types, dash-wedge (side-view) and Newman (end-view). There were 24 problems total, half with dash-wedge target diagrams and half with Newman target diagrams. The starting orientation of the model was such that it maximized the global angular distance to each of the target diagrams. Half of the trials involved a conformation change (local rotation) of the molecular model (i.e., changing the spatial configuration of substituents by rotating the bond between the molecule's two chiral carbons). Six different molecules were used in the 24 trial problems; and were systematically varied with the target diagram type, and local rotation trials. All participants received the trials in the same order, in which two consecutive trials never showed the same target diagram or molecule.


Figure 2: Participants manipulated the 3D molecular model to match the orientation depicted by either a Newman (left) or Dash-Wedge (right) diagram.

Items from the NASA Task Load Index (TLX) (Hart \& Staveland, 1988) were administered to assess participants' subjective experience of the task with regard to six criteria: mental demand, physical demand, temporal demand, own performance, effort, and frustration. Participants rated each
of these criteria on a scale from 0 to 100 with 0 being the lowest and 100 being the highest rating.

Participants were administered two tests of spatial ability, a mental rotation test (MRT) (Vandenburg \& Kuse, 1978), and a three dimensional perspective taking test, Visualization of Viewpoints (VoV) (Guay \& McDaniels, 1976).

Items from Waller's (2000) computer use questionnaire were administered to assess participants' attitudes and experience with computers. Participants rated 10 statements on a scale of 1 (completely disagree) through 7 (completely agree).

## Results

The following results include 108 participants ( 56 female). Data from 12 students were excluded from the analyses as they had much lower accuracy (angular errors of $>30^{\circ}$ ) suggesting they did not understand the task or were unmotivated. The four interface condition groups had approximately equal numbers of males and females. Response times that were greater than 2.5 standard deviations from a participant's mean response time were replaced with their mean response time. The groups did not significantly differ on the MRT, $F(3,104)=0.8, p=.56$, VoV, $F(3,104)=1.7, p=.17$, computer experience, $F(3$, $104)=0.43, p=.73$, or attitudes toward computers, $F(3$, 104) $=1.0, p=.42$.

The mean angular error for the different experimental groups is shown in Figure 3. Overall, participants had an average angular error of $13.7^{\circ}(S D=6.6)$. A significant effect of colocation was found on error, $F(1,104)=6.6, p$ $=.01, \eta_{p}{ }^{2}=.06$. Marginal means showed that participants provided with the colocated interface had a lower average angular error (i.e., greater accuracy) of $12.1^{\circ}(S D=5.7)$, than those using the displaced interface $15.4^{\circ}(S D=7.1)$. No significant effect of stereo was observed, $F(1,104)=$ $0.6, p=.46$. There was no observed interaction between colocation and stereo $F(1,104)=2.5, p=.12$.


Figure 3: Effects of providing stereo and colocation on participant accuracy $(M \pm S E)$.

Overall, participants had an average response time of 33.8 s ( $S D=15.2$ ). No significant effect of colocation was found on response time, $F(1,104)=1.5, p=.22$. Also, no significant effect of stereo was observed, $F(1,104)=1.2, p$
$=.28$. There was no observed interaction between colocation and stereo, $F(1,104)=0.6, p=.43$.


Figure 4: Effects of providing stereo and colocation on participant response time $(M \pm S E)$.

NASA-TLX ratings are shown in Table 1. On average, participants provided with a colocated interface rated their experience as having significantly less physical demand and frustration than those with the displaced interface. Further, participants with colocated interfaces rated their own performance to be significantly greater than those with displaced interfaces. No main effects of stereo were found on any of the six task demand ratings.

Significant interactions between stereo and colocation were observed on ratings of effort, $F(1,104)=4.3, p=.04$, $\eta_{\mathrm{p}}{ }^{2}=.04$, and frustration, $F(1,104)=8.3, p=.005 \eta_{\mathrm{p}}{ }^{2}=$ .06. When colocation was provided, participants using stereoscopic displays reported less task effort ( $M=58.1$, $S E$ $=4.6$ ) than participants using monoscopic displays ( $M=$ $72.3, S E=4.8$ ). Further, when provided stereo, participants using colocated interfaces reported less frustration ( $M=$ 19.4, $S E=5.1$ ) than those using displaced interfaces ( $M=$ $45.6, S E=5.3$ ).

Table 1: Effect of stereo and colocation on NASA-TLX ratings.

| Task Demand | Colocated | Displaced | ANOVA |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $M(S E)$ | $M(S E)$ | $d f$ | F | $p$ | $\eta_{\mathrm{p}}{ }^{2}$ |
| Mental | $48.1(3.3)$ | $53.8(3.3)$ | 104 | 1.5 | .23 | - |
| *Physical | $24.4(3.3)$ | $34.2(3.4)$ | 104 | 4.2 | $.04^{*}$ | .04 |
| Temporal | $44.6(3.1)$ | $46.9(3.1)$ | 104 | 0.3 | .60 | - |
| *Performance | $82.2(2.3)$ | $75.2(2.3)$ | 104 | 4.6 | $.03^{*}$ | .04 |
| Effort | $65.2(3.3)$ | $65.7(3.4)$ | 104 | 0.0 | .92 | - |
| *Frustration | $25.4(3.7)$ | $36.5(3.8)$ | 104 | 4.3 | $.04^{*}$ | .04 |
|  | Stereo | Mono |  |  |  |  |
| Mental | $48.9(3.2)$ | $52.9(3.4)$ | 104 | 0.7 | .40 | - |
| Physical | $30.9(3.3)$ | $27.8(3.4)$ | 104 | 0.4 | .52 | - |
| Temporal | $49.5(3.1)$ | $42.1(3.2)$ | 104 | 2.9 | .09 | - |
| Performance | $76.7(2.3)$ | $80.8(2.4)$ | 104 | 1.5 | .22 | - |
| Effort | $63.3(3.3)$ | $67.6(3.4)$ | 104 | 0.8 | .36 | - |
| Frustration | $32.6(3.7)$ | $29.2(3.8)$ | 104 | 0.4 | .52 | - |
| $p<.05 ; \mathrm{N}=108$ |  |  |  |  |  |  |

[^7]Table 2: Correlations between dependent measures, spatial ability, and computer use.

|  | RT | Ang. <br> Error | MRT | VoV | Comp. <br> Att. | Comp. <br> Exp. |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| RT | 1 | - | - | - | - | - |
| Ang. | $.36^{* *}$ | 1 | - | - | - | - |
| Error | $-.25^{*}$ | $-.25^{* *}$ | 1 | - | - | - |
| MRT | $-.32^{* *}$ | $-.36^{* *}$ | $.45^{* *}$ | 1 | - | - |
| VoV | $-.32^{* *}$ |  |  |  |  |  |
| Comp. <br> Att. <br> Comp. <br> Exp. | -.17 | -.07 | $.30^{* *}$ | $.34^{* *}$ | 1 | - |

* $p<.05, * * p<.01$, two-tailed; $\mathrm{N}=108$

As shown in Table 2, MRT scores showed small but significant correlations, while VoV scores showed somewhat higher correlations with response time and accuracy. In order to investigate possible interactions between spatial ability and aspects of the interface and display, scores from the MRT and VoV were standardized and averaged to produce a combined spatial ability score. A median split was used to separate high and low spatial ability participants, and was then used as a factor in the analysis. As expected, high spatial ability participants completed the task faster and had greater accuracy than low spatial ability participants $F(1,100)=7.9, p=.006, \eta_{p}{ }^{2}=$ .07. High spatial ability participants had an average response time of $29.3 \mathrm{~s}(S D=12.2)$ and low spatial ability participants $38.1 \mathrm{~s}(S D=16.6)$. For accuracy, high spatial ability participants had a lower average angular error of $11.2^{\circ}(S D=5.6)$ than low spatial ability participants $16.2^{\circ}$ $(S D=6.5)$, this finding was also significant $F(1,100)=$ $13.2, p=<.001, \eta_{p}^{2}=.12$. Interactions of spatial ability with stereo and / or colocation did not reach significance.

Participants' level of experience with computers showed a small significant correlation with accuracy, but no significant correlation with response time. As shown in Table 2, there were moderate correlations between computer attitude and experience with both the MRT and VoV scores. A median-split was used to separate participants into high and low groups for computer attitude and computer experience. When used as a factor in the analysis, there was no main effect of either attitude or experience on performance. Further, there were no significant interactions of computer attitude or experience with stereo and / or colocation.

Overall, males were significantly more accurate than females, $F(1,106)=5.1, p=.025, \eta_{\mathrm{p}}{ }^{2}=.05$. Males had an average angular error of $12.3^{\circ}(S D=6.6)$ and females $15.1^{\circ}$ ( $S D=6.3$ ). Males and females did not significantly differ in response time, $F(1,106)=0.5, p=.50$. However, a significant interaction of gender with colocation is evident in response time, $F(1,104)=5.0, p=.028, \eta_{p}{ }^{2}=.05$.

Pairwise analysis revealed that female participants performed significantly faster with a colocated interface ( $M$ $=29.9 \mathrm{~s}, S E=2.8$ ) than with a displaced interface $(M=$ $39.8 \mathrm{~s}, S E=2.8), F(1,104)=6.2, p=.014, \eta_{\mathrm{p}}{ }^{2}=.06$. Response times for males were not affected by colocation, $F(1,104)=0.4, p=.49$. There were no other significant interactions with gender.

## Discussion

The main purpose of this study was to examine the importance of providing stereoscopic 3D viewing and colocation of the haptic interface and virtual image during a virtual object orientation task using molecular models. The results demonstrate that providing colocation had a small but significant impact on accuracy. The experiment failed to demonstrate an effect of stereo on accuracy. The results failed to show an effect of colocation or stereo on task completion speed. Overall, these findings suggest that colocation of haptic and visual information enabled perceptual mediation of the task to some degree, whereas stereo did not significantly increase perceptual mediation or decrease cognitive load for this particular task.

Our results can be compared with previous studies investigating the effect of colocation on virtual object rotation. Ware and Rose (1999) found that colocation led to $35 \%$ faster performance during an object rotation task, however no effect on accuracy was found. In a later study, Ware and Arsenault (2006) found that colocation led to faster response times, however they did not have a measure of accuracy because trials were automatically terminated when the manipulated object was within $5^{\circ}$ of the target orientation. Other studies investigating effects of colocation are difficult to compare with the present study; the majority of tasks used involve object translation rather than pure object rotation, as in our task.

The findings of the present study demonstrate an accuracy advantage, rather than the speed advantages demonstrated in the previous studies. In regard to the Ware and Rose (1999) result, it is important to note the authors' task involved repeatedly rotating a single simple shape and did not require any local manipulation of the object, as in our task. It is possible that when task demands are greater, providing colocation benefits accuracy more than response time. In regard to Ware and Arsenault (2006), the accuracy advantage from colocation found in the present study would translate into a response time benefit had the trials required a minimum angular error for completion; thus it is likely that the findings are complementary. Despite the relatively small effect on accuracy, this study adds to the body of literature demonstrating a performance benefit from colocating visual and haptic workspaces for virtual object rotation tasks. Further, this study shows that when rotating different complex structures to match a given orientation, providing colocation of haptic and visual information may benefit precision more than speed.

Numerous studies have demonstrated performance advantages from providing stereoscopic 3D viewing in virtual object manipulation tasks (Ware \& Franck, 1996; Hu, Hellen, 2000; Arsenault \& Ware, 2004; Liere, Kok, Martens, 2005). However, many of these studies involve virtual object translation tasks, rather than rotation tasks, as in the present study. One must use caution in generalizing effects of interface design on translation tasks to rotation tasks, as the two processes are independent (Wang, MacKenzie, Summers, \& Booth, 1998; Ware \& Rose, 1999). Other studies comparing stereo and mono displays that involved tasks other than translation often find no beneficial effect of stereoscopic viewing (Hoffmeister, Frank, Cuschieri, \& Wade 2001, Kooshabeh \& Hegarty, 2010).

Another possible explanation for the null result of providing stereo is that the task had low demand on depth perception. The task used in the present study was purely an object rotation task, the models manipulated were regular structures that rotated around a fixed origin in space, and the task did not require making difficult judgments about relative distances in depth. The tetrahedron structure of the molecules may have allowed for necessary judgments of depth to be made via monocular depth cues such as occlusion, motion, linear perspective, and shadowing. It is possible that the depth perception demands of the task could be supported by monocular cues alone. Despite the growing excitement surrounding 3D stereoscopic displays, performance on certain tasks and applications may not benefit from providing the latest display technology. Future research is needed to elucidate the specific task qualities and learning situations under which providing stereo actually benefits the user.

Results from the NASA-TLX measure of participants' subjective experience were consistent with the performance data, and further demonstrated the importance of providing colocated visual and haptic workspaces and the null effect of viewing the display in stereo. Participants with the colocated interface reported the task to be significantly less physically demanding, less frustrating, and rated their perceived performance higher than participants who used the displaced interface. Participants were more comfortable and confident when the interface was colocated; this provides further evidence that although the performance effects were small, they were meaningful in that they were associated with the perceived task demands of the users. Stereo did not affect ratings of subjective experience, further demonstrating its unimportance for this task.

Males performed the task more accurately than females. An interesting result is that females performed trials about 10 seconds faster when provided with colocation. This result suggests that females have a more difficult time dealing with visual and haptic mismatches, which might be attributed to differences in spatial ability, experience with computers, or both.

This study demonstrated that providing colocated haptic and visual workspaces had a small beneficial impact on accuracy during a virtual object orientation matching task, while providing stereo had no significant effect on accuracy. Further, neither factor affected overall response time. It will be important to examine whether results found on representation matching performance generalize to meaningful learning of the spatial structures. Future studies will investigate how specific interface design features relate to students' ability to understand concepts regarding 3D molecular structure and whether this learning can be maintained and utilized during novel situations in which models are no longer available. In addition to providing basic information regarding the perceptual cues that facilitate virtual object manipulation, this research will inform the design of virtual models for science learning.

## Acknowledgements

This research was supported by the National Science Foundation, grant \#1008650.

## References

Arsenault, R., \& Ware, C. (2004). The importance of stereo and eye-coupled perspective for eye-hand coordination in fish tank VR. Presence: Teleoperators \& Virtual Environments, 13(5), 549-559.
Barnea, N., \& Dori, Y.J. (2000). Computerized Molecular Modeling - The New Technology for Enhancing Model Perception Among Chemistry Educators and Learners. Chemistry Education: Research and Practice in Europe, 1, 109-120.
Guay, R., \& Mc Daniels, E. (1976). The visualization of viewpoints. West Lafayette, IN: The Purdue Research Foundation (as modified by Lippa I., Hegarty M., \& Montello DR, 2002).
Hart, S. G., \& Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. Human mental workload, 1, 139-183.
Hofmeister, J., Frank, T. G., Cuschieri, A., \& Wade, N. J. (2001). Perceptual aspects of two-dimensional and stereoscopic display techniques in endoscopic surgery: review and current problems. Surgical Innovation, 8(1), 12-24.
Hu, Helen H., et al. "Visual cues for imminent object contact in realistic virtual environment." Proceedings of the conference on Visualization'00. IEEE Computer Society Press, 2000.
Khooshabeh, P., \& Hegarty, M. (2010). Inferring crosssections: When internal visualizations are more important than properties of external visualizations. HumanComputer Interaction, 25(2), 119-147.
Klatzky, R. L., Wu, B., \& Stetten, G. (2008). Spatial Representations From Perception and Cognitive Mediation The Case of Ultrasound. Current directions in
psychological science, 17(6), 359-364.
Kok, A. J., \& Van Liere, R. (2004, March). Co-location and tactile feedback for 2D widget manipulation. In Virtual Reality, 2004. Proceedings. IEEE (pp. 233-234). IEEE.
Kozma, R. B., \& Russel, J. (2005). Multimedia Learning in Chemistry. In R. E. Mayer (Ed.) The Cambridge Handbook of Multimedia Learning. New York, Cambridge University Press.
Limniou, M., Roberts, D., \& Papadopoulos, N. (2008). Full Immersive Virtual Environment CAVE ${ }^{\text {TM }}$ in Chemistry Education. Computers \& Education, 51, 584-593.
Stull, A. S., Barrett, T., Hegarty, M. (2012) Design of a virtual reality system for the study of diagram use in organic chemistry. Proceedings of the Workshop on Technology-Enhanced Diagrams Research (pp. 33-41). Canterbury, UK.
Trindade, J., Fiolhais, C., \& Almeida, L. (2002). Science learning in virtual environments: a descriptive study. British Journal of Educational Technology, 33(4), 471-488.
van Liere, R., Kok, A., Martens, J. B., \& van Tienen, M. H. A. V. (2005). Interacting with molecular structures: user performance versus system complexity. In Eurographics Symposium on Virtual Environments, (Aalborg University, Denmark) (pp. 147-156).
Vandenberg, S. G., \& Kuse, A. R. (1978). Mental rotations, a group test of three-dimensional spatial visualization. Perceptual and motor skills, 47(2), 599.
Voyer, D., Voyer, S., \& Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: a meta-analysis and consideration of critical variables. Psychological bulletin, 117(2), 250.
Waller, D. (2000). Individual differences in spatial learning from computer-simulated environments. Journal of Experimental Psychology: Applied, 6(4), 307.
Wang, Y., MacKenzie, C. L., Summers, V. A., \& Booth, K. S. (1998, April). The structure of object transportation and orientation in human-computer interaction. In Conference on Human Factors in Computing Systems: Proceedings of the SIGCHI conference on Human factors in computing systems (Vol. 18, No. 23, pp. 312-319).
Ware, C., \& Arsenault, R. (2004, August). Frames of reference in virtual object rotation. In Proceedings of the 1st Symposium on Applied perception in graphics and visualization (pp. 135-141). ACM.
Ware, C., \& Franck, G. (1996). Evaluating stereo and motion cues for visualizing information nets in three dimensions. ACM Transactions on Graphics (TOG), 15(2), 121-140.
Ware, C., \& Rose, J. (1999). Rotating virtual objects with real handles. ACM Transactions on Computer-Human Interaction (TOCHI), 6(2), 162-180.

# Using the words toddlers know now to predict the words they will learn next 

Nicole M. Beckage (Nicole.Beckage@Colorado.edu)<br>Department of Psychology and Neuroscience, 345 UCB<br>Boulder, CO 80309 USA

Eliana Colunga (Eliana.Colunga@Colorado.edu)<br>Department of Psychology and Neuroscience, 345 UCB<br>Boulder, CO 80309 USA


#### Abstract

We set forth to show that lexical connectivity plays a role in understanding early word learning. By considering words that are learned in temporal proximity to one another to be related, we are able to better predict the words next learned by toddlers. We build conditional probability models based on data from the growing vocabularies of 77 toddlers, followed longitudinally for a year. This type of conditional probability model outperforms the current norms based on baseline probabilities of learning given age alone. This is a first step to capturing the interaction between a child's productive vocabulary and their learning environment in order to understand what words a child might learn next. We also test different types of variants of this conditional probability and find that not only is there information in words that are learned in proximity to one another but that it matters how models integrate this information. The application of this work may provide better cognitive models of acquisition and perhaps allow us to detect children at risk for enduring language difficulties earlier and more accurately.


Keywords: word learning, semantics, language acquisition, co-occurrence, development, longitudinal data, CDI

## Introduction

Do children learn words systematically? There is a lot of evidence that words are not all learned equally. Perhaps not surprisingly, for example, parents' vocabulary is related to their children's vocabulary (e.g., Weizman \& Snow, 2001; Veen, et al., 2009). That is, the child will learn the words in his or her environment. In addition, some concepts, and therefore the words that name them, may be easier to learn than others. For example, concrete nouns are learned earlier than verbs and adjectives (e.g., Sandhofer, Smith, \& Luo, 2000; Gentner, 2006). Furthermore, the child may bring some preferences and constraints to the task of word learning. For example, children may become particularly interested in dinosaurs or construction equipment or even tea sets (DeLoache, Simcock, \& Macari, 2007). That is, in characterizing the forces that guide word learning, there is evidence that at least three distinct but not necessarily mutually exclusive sources of information can come to bear: a) the structure and composition of the linguistic environment, b) the structure of the concepts and categories being named, and c) the characteristics of the learner itself. In this paper we focus mostly on this third source of variability by constructing conditional probability models from longitudinal trajectories of word learning that make predictions at the word level, for individual children. That
is, we ask: can we use the words a child knows now to predict the words that a child will learn next?

## Measuring the developing lexicon

One well-established way to characterize toddlers' lexicons is to use vocabulary checklists, such as the MacArthur-Bates Communicative Development Inventory (CDI; Fenson et al., 1994). These parent-reported measures have been shown to be effective in evaluating children's communicative skills up to 30 months of age (e.g., Thal et al., 1999). The CDI: Words \& Sentences Toddler form is a checklist of over 700 early words that at least $50 \%$ of children typically say at 30 months of age. By pooling data over thousands of children, the CDI provides norms of the percentage of children who say each of these words at a given age from 16 to 30 months of age, month by month. Aside from being shown to be a valid measure of communicative skills for this age group, the CDI has been recently shown to be an effective tool for sorting toddlers at the low and high end of the acquisition distribution into late talkers and typically developing children (Heilmann, et. al, 2005). This might allow us to see universality in learning but it also masks some of how the process works-the aggregate cannot explain individual differences but models of learning necessarily must.

The CDI norms can be used to build models of growth. For example, Hills and colleagues (2009) have used CDI norms to build growth models based on networks of words connected by feature similarity or associative strength. Beckage, Smith and Hills (2011) used the connectivity of language within the vocabulary of young learners and showed that there are differences in the structure of the vocabularies of children at risk for language impairments and those of typically developing children.

Note that these approaches presuppose that there is information in the relationships between words. If this is the case, there should be predictive power in looking at the between-word dependencies over time. We do this by exploiting the statistical regularities present in the developing vocabularies of 77 children, followed longitudinally for a year, at monthly intervals.

## Rationale

We propose a simple way of uncovering the interaction between the language environment and learning and thus uncovering more of the systematicity of word learning. Instead of just considering the frequency of production for a
given word conditioned on the age of a child, as with the CDI norms, we suggest that there might be additional information in the structure of the language knowledge itself, in the set of words that are known. To pursue this claim, we build the most naïve notion of relatedness and leverage this information in order to predict what words a child might learn next. We define relatedness to be the conditional probability of a word given the child knows another word. For the sake of this paper, we consider words learned in temporal proximity to be related. We build up these values from the longitudinally collected CDIs by considering a connection between words that are learned at the same time (within the same month). We then compute the conditional probability as follows:

$$
\begin{equation*}
\operatorname{Pr}(a \mid b)=\frac{P(a \bigcap b)}{\operatorname{Pr}(b)} \tag{1}
\end{equation*}
$$

For example, we compute the probability of knowing "cat" given "dog" by calculating the probability of a child learning "cat" and "dog" in the same month, normalized by the probability of knowing "dog" in the population as a whole. We can then use a variety of methods to combine these conditional probabilities into a single probability of learning word $i$ given that they know a set of words $\mathbf{J}$. That is, for each not-yet-known word, we can calculate the probability that the child will learn that word next, given the set of words the child already knows.

In order to combine the conditional probabilities given the set of known words, we need to integrate over the conditional probability given each of the words known. Here we test three different models of this: the Additive model assumes that every conditional probability contributes equally. In the additive model we simply sum up the conditional probability of $i$ given every $j$ in the set of known words. This gives us a proportion of learning for every word not yet learned. The issue with this model is that it requires a large amount of information and storage. One rudimentary simplification would be to assume that only the maximum conditional probability was used. This model we call the Maximal model because we use only the strongest conditional probability between $i$ and some $j$ from the set of known words $\mathbf{J}$. Finally in the Threshold model we compare a model that considers conditional probabilities in an additive fashion but considers links only as present or absent. The link is determined as present when the conditional probability strength is above a certain threshold (in our case the median of all conditional probability values) and absent otherwise. We compare these conditional probability models to two population-based models, one based on the CDI norms (norm-baseline), another based on the observed frequencies (observed-baseline), as well as the null model (the assumption that all words will be learned with equal chance). We evaluate the conditional probability models by comparing their predictive power to the population-based models, and use 5 -fold crossvalidation to evaluate the model's performance in predicting untrained trajectories.

## Methods

## Vocabularies and Co-occurrences

We utilize CDI measured vocabularies collected at the University of Colorado Boulder. Seventy-seven toddlers between 15.7 and 18.6 months (mean starting age 17 months) were recruited as part of a year-long longitudinal study. These participants completed monthly behavioral tasks as well as vocabulary assessments. The vocabulary assessments were conducted through parent report using the CDI Words \& Sentences toddler forms. These CDIs were collected for 12 consecutive months with the majority of parents completing the forms each month. On average we have 9.8 months for each child.

In our study, we include a total of 650 words from the full form, marking duplicate words with parts of speech (such as "orange" as a noun and "orange" as an adjective). We included words that were both on the full form and had norms available online (http://www.sci.sdsu.edu/cdi/). Words that were not part of our modeling included words like above, after, on and off. All together we have 77 children and a total of 684 CDI forms. For the sake of this paper we consider each month to be independent of every other month. That is, we build associative structure only from words that are learned during the same month (or words that are known at the beginning of the study.) This limits the co-occurrence measure to capture only short-term dependencies. In the future we plan to extend this work to include cross-sectional vocabularies as well, which will allow us to capture long-term dependencies.

To derive the strength of connectivity, we simply take a count of the number of times two words appear in the same vocabulary (i.e. are learned in the same month) normalized by the population level knowledge as measured from our sample for the words. This provides the basic counts that are then used to compute an 'activation level' that will then give rise to predictions of the next word learned. We then calculate the probability of learning word $i$ given that a child already knows word $j$. This is then compared to the models based only on population level data as well as a model that assumes uniform learning.

## Models

We compute two population-based measures. The first normed model is based on the CDI norms where we consider the likelihood of a child learning the specific set of words we observe to be a function of the population level age of acquisition (AoA) norms (Dale \& Fenson 1996). The second measure is calculated analogously, but computing the likelihood according to the AoA as observed in our own sample. We also compare these and all other models to a straw-man baseline measure (the null model) that gives every unlearned word equal probability of being learned.

We compare these population level models to conditional probability models. For the additive model we calculate the probability that each word is learned as proportional to the sum of the conditional probability of all known words.

$$
\begin{equation*}
\operatorname{Pr}(a \mid \mathbf{B}) \propto \sum_{b_{i} \in \mathbf{B}} \frac{\operatorname{Pr}\left(a \bigcap b_{i}\right)}{\operatorname{Pr}\left(b_{i}\right)} \tag{2}
\end{equation*}
$$

In effect, the probability of learning word $a$ is computed such that we sum across the conditional probability of each known word $b$ in the set of known words $\boldsymbol{B}$. For example if a child knows words "cat" and "dog" then the probability that the child will learn word "pet" is proportional to the sum of the probability of learning "pet" given "cat" plus the probability of learning "pet" given "dog". This assumes a level of independence that does not exist in language itself.

Using similar methods, in the maximal model we consider only the maximum conditional probability to be proportional to the probability of learning a given word. We test this model because it requires only one point of information per word as opposed to considering all possible combinations of known and unknown words.

$$
\begin{equation*}
\operatorname{Pr}(a \mid \mathbf{B}) \propto \arg \max _{b_{i} \in \mathbf{B}}\left\{\frac{\operatorname{Pr}\left(a \bigcap b_{i}\right)}{\operatorname{Pr}\left(b_{i}\right)}\right\} \tag{3}
\end{equation*}
$$

This simplification may still capture much of the variance if the maximal connection dominates the additive model or if strong connections in the learning environment really do highlight words to be learned next.

We also consider thresholded conditional probability. The idea here is that the learner has access to most of the conditional probability space but only at a coarse level. The learner is considered to maintain only the strongest conditional probabilities and that these are considered as present or not. Mathematically, the threshold model is:

$$
\begin{equation*}
\operatorname{Pr}(a \mid \mathbf{B}) \propto \sum_{b_{i} \in \mathbf{B}} \mathbb{I}_{\left\{\operatorname{Pr}\left(a \mid b_{i}\right)>c\right\}} \tag{4}
\end{equation*}
$$

Here $\mathbf{I}$ is the indicator function and is valued only when the conditional probability is greater than some constant c . For this analysis we let c be the median conditional probability across all children. This adds an additional variable to our model but it is set a priori and thus has little significance on the complexity of the model. From an information processing point of view, this model may take less effort since we consider only the presence and absence of a link and not the weight, reducing the complexity of the space.

While we do not consider it here, even these very simple models can quickly be extended to other types of more sophisticated models. First, conditional probability is at best a first order approximation to the full complexity that language embodies. The model can only be as good as the measure used to inform it which in this case is simple cooccurrences. Further, these co-occurrences are based on one month time slots, we could consider data at other time scales, or multiple timescales. Finally, we have chosen to integrate the conditional probability information here in the simplest ways. These models can be seen as network models which allows us to consider not only what words are predicted to be learned next but how these mechanisms for learning transform the semantic structure present in the network. We mention this to highlight the implications of testing these basic assumptions on the larger word learning models (e.g., Vitevitch, 2008; Hills et al, 2009).

## Evaluation

We begin by looking at the percent of vocabularies better fit than the null model and the percent likelihood improvement. This tells us some information about the general variability of the input and the ability of the model to account for this variability but little about which model is best. Thus we consider the percent of vocabularies that are better than each of the population based models. This tells us the proportion of vocabularies better fit by the model but not how much better (or worse) of a fit the models give us for our sample. Thus we include the total likelihood of the test data given each of the models. We then compare the likelihood fits across models in order to understand how a model is performing in comparison to other models. We also look at the percent of vocabularies best fit by a given model.

This gives us a good deal of information about the performance of the models and the ability of the models to utilize and combine information in order to predict words learned next at the level of an individual child's vocabulary. To be sure that we are capturing actual signal we want to calculate the conditional probabilities based on a different set of vocabularies than those on which we test the models. Thus, we use cross validation and iteratively build up the necessary associations that the models require from $80 \%$ of the vocabularies and then test on the remaining $20 \%$. We do this at the child level since sequential vocabularies are not independent of the child (an issue we've ignored up to this point). We randomly select the $20 \%$ test group and repeat this 5 times such that every observed set of vocabularies for a given child is in the test set once. This allows us to test how well the model can predict the vocabulary growth of a child it has no direct information about. We compare the average performance on the five different test sets.

## Results

We know what specific words a child learned in a given month, and we use our model to calculate the probability of a given set of words. Some models give a zero probability to learning certain words and thus we first want to look at what percentage of our population cannot be fit by a specified model. This will give us information about how constrained the models are in their ability to fit the wide range of data present in our sample. Column 1 of Table 1 shows the results. In general the models are able to capture the learned words fairly well. The worst model is the Threshold model which is due to the fact that many words are assigned a zero probability under this model since connections that are not above the median strength are considered absent-this results in about 4\% of the observed vocabularies not being explainable under the strictest definition of the model. The model based on the CDI norms also has some difficulty accounting for some ( $2 \%$ ) of the vocabularies seen. In practice, this means that some children learned words earlier than the normed CDI measures would have predicted-that is, in the vocabularies used to build up the norms, there were no children in the sample that learned some words that children in our study did. This is even more extreme for our
observed CDI norms-which is probably an effect of sample size. In general a large majority of the 684 vocabularies could be fit under the models we constructed; a total of 633 overlapped across all models. We constrain our model comparisons to this subset in all further evaluations.

The next major question is whether or not the constructed models outperform the null model in which each word is given equal probability of being learned. The answer, in short, is that all models perform better than the null model when we consider the total likelihood across all vocabularies. Further, the minimum number of vocabularies better fit by our models than the null model was $82 \%$. This suggests that there is systematicity to the order in which children learn words. In fact we don't only fit the observed data better we get a fairly substantial improvement in overall likelihood when we utilize these models. We have at last $8 \%$ improvement and at most $19 \%$ improvement.

Table 1: Model performance compared to null model. We consider \% of vocabularies not fit, improvement over null and $\%$ of vocabularies better fit by a given model.

| Model | vocabs not <br> fit (\%) | improvement <br> over null (\% llk) | vocabs. better <br> than null (\%) |
| :--- | :--- | :--- | :--- |
| Normed | 2.37 | 14.54 | 81.99 |
| Observed | 3.97 | $\mathbf{1 9 . 0 4}$ | $\mathbf{9 0 . 5 2}$ |
| Additive | 0 | 18.22 | 89.10 |
| Maximal | 0 | 8.39 | 86.41 |
| Threshold | 4.05 | 18.66 | 82.94 |

However, showing that words are not acquired randomly does not answer the question of how individual children build up a vocabulary. Returning to the ideas from the introduction, this does not rule out the effect of the structure of the environment. Children learning words proportional to the frequency they encounter them in the environment could explain these results. This would maintain independence between the words a child knows and the words the child is going to learn next. The two baseline models maintain this independence as well: the model based on the normed CDIs and the model fitting to the observed CDIs. In contrast, the other models assume conditional probability plays a role in prediction of vocabulary growth and uses this to link known words to what words will be learned next. Thus, to get at our original question we want to compare these population level models to the other models that require conditional information. We already have a bit of information about the overall model performance when we look at the total likelihood across all vocabularies. We see that we get the largest improvement in likelihood when we utilize the observed CDIs. And we also see that this model gives us the most vocabularies that are fit better than random acquisition.

The gains resulting from using conditional probability are clearer when we consider which vocabularies were best fit rather than looking at the overall likelihood which could be easily inflated by isolated vocabularies that are particularly difficult for a given model to fit. With cross-validation, the
threshold model outperforms all others, as shown in Table 2. This improvement is non-trivial as it accounts for the best model in over $50 \%$ of vocabularies. This is maintained when we look across children as well-most children are best fit by the threshold model. The observed norm model does provide the best fit for $22 \%$ of the data suggesting that there is some predictive powers in the population level rate and time of acquisition. When we look across the population level models we see that over $70 \%$ of vocabularies are better fit by a conditional probability model than by a population level normed model. Critically, this suggests that there is some added information in conditional probabilities.

Table 2: Performance with cross-validation. Overall ability to account for the data as well as percent of vocabularies best fit by a given model. For comparison the model performance is directly compared to population models.

| Model | \% vocabs <br> best fit | \% better <br> normed | \% better <br> observed |
| :--- | :--- | :---: | :---: |
| Normed | 7.28 |  | 30.82 |
| Observed | 22.04 | 69.17 |  |
| Additive | 10.23 | 75.20 | 54.29 |
| Maximal | 10.31 | 25.63 | 17.67 |
| Threshold | $\mathbf{5 0 . 1 1}$ | $\mathbf{7 7 . 3 8}$ | $\mathbf{6 0 . 5 2}$ |

To show the extent of improvement offered by conditional probabilities, we consider the percent of vocabularies better fit by a given model and the CDI data. In Table 2, column 2 and 3, we see that most of the models perform much better than the normed model with roughly $75 \%$ of children being better fit under a given model than the published norms and further that many vocabularies are better fit when compared to the norms based on our particular population of toddlers in boulder. This suggests that the norms may be predictive for some children but that in general accounting for the words that are learned previously as well as the relationship of words that are learned together may help us predict what word a child will learn next. Further, the way we combine the type of joint information about word learning may influence our ability to capture vocabulary growth.

## Discussion

These results suggest that conditional probabilities do aid in accounting for word learning trajectories. That is, the words that a child already knows can help predict the words that they are going to learn in the future. This implies that there is some sort of systematicity in word learning and that it is not explainable by structure of the environment alone or by conceptual complexity but rather by the interaction of the structure of concepts and meaning within the knowledge of the individual child. The two models that are based on normed data can be seen as independent of individual variation. That is, for these models to perform well at predicting what words a child will learn next, children across a variety of settings and in a variety of learning
environments would be expected to learn words in similar proportion and at a similar rate. This could suggest that the input is structured in a systematic way or that the learning strategy is the same across all children and not dependent on the child's productive vocabulary at any point in time. We did see that these models in general can be fairly predictive of word learning and in fact the total likelihood of the data was minimized under the model that built norms from CDIs collected in our lab. This suggests that these models capture some important aspect of learning. However if we are interested in understanding the different styles of how children learn and capturing the variability across children, these models, inherently, cannot help us with these types of questions as they average out variability.

Looking more closely at the overall likelihood of the models, we see a strong trend that the population models are not as able to adapt to new data. When we fit the models on the full data (that is we included the test set in the training set) the observed CDI norms had a much better total likelihood. However, this model took a big hit in the crossvalidation method (results not shown in this paper) suggesting that the observed CDI norms may have overfit the data. The fact that conditional probability models performed better than the population level models in predicting unseen data suggests that the whole story is not in the input alone, but that there is an interaction between a child's productive vocabulary and what words the child will learn next. Even with very simple models of conditional probability we were able to increase our ability to predict and account for the ways vocabularies expand. Thus, if we were to refine our models to include other types of relationships (or more meaningful semantic relationships) between known words and words learned we might be able to understand how children take in their language environment and combine this with their individual vocabulary knowledge to learn new words. The work presented here only begins to look at this by testing models that combine the relationship of co-learned words in different ways, but refinement on these types of models could provide a way for us to uncover not just how children learn new words but also how they integrate a variety of information in order to develop representations of the world. For example, here we considered the median and as our cutoff in the threshold models, but in theory this could be a free parameter fit at the level of individual children (or at the population level conditioned on age) and could hold added information about how children interact with the learning environment. It is true that this threshold model has an additional variable but by setting this before looking at the data we have dealt with any issues in comparing this model to the other models. In the future we plan to do more extensive parameter fits as well as extend the basic models in complexity. For example, we would like to allow the number of maximal values included in our maximal model to be n instead of just 1 , where n is a free parameter itself.

The first model (the additive model) tested combined conditional probabilities by maintaining connections and
weights and summing up all of the conditional probabilities between the word candidate and all known words. This resulted in a model that was able to fit much of the data and often better than the population level models. But this was not the best fitting model suggesting that this model might have required too much information, accumulating a ratio that included significant noise in addition to the signal. A huge simplifying assumption that led to our next model was one that suggested that children would maintain only the strongest relationship between a word candidate and known words. This model performed poorly-returning a total likelihood significantly worse than the normed models and the closest to the null model. However, the children's vocabularies that were better fit by this model than the CDI norm models were vocabularies that were often best fit overall by this model. The best model is the model that forces a threshold on the conditional probability matrix. This suggests that strong connections may be the important ones and that the weight of the connection is not important just that it is present.

We do not only gain insight from looking at what models succeed but also what models failed and how. The CDI norming data had difficulty capturing individual vocabularies. It is important to note that in some way this model was handicapped from the beginning. None of the observed data was used in building up the norms. On the other hand, the frequencies noted in the norms were accrued over thousands of children, as opposed to our much smaller sample. Nonetheless, even when the other models were handicapped in the same fashion, the discrepancy in performance still exists. This highlights one of the major weaknesses in utilizing normed data in order to help predict future vocabulary progression. First, it fails to exploit the temporal dependencies available when using longitudinal data. Second, it fails to utilize the dependencies between subsets of words. Of course the poor performance of the norm baseline could be due to a variety of other reasons which would plague any attempt to characterize universals from individual data, and which pose problems to the traditional norming studies. For example, geographic changes between where the norms are collected and Boulder, CO, where our vocabularies were collected could produce variation thus restricting the generalizability of the norms. Or there could be cohort differences due to the fact that the world in which our current children are growing up has a different underlying structure in small but significant ways than the world of the children who contributed data to the norms 20 years ago. This suggests a need for us to consider other tools and methods in order to build up a robust and predictive measure of infant word learning.

## Conclusions and Further Directions

Altogether, our findings demonstrate that the conditional probabilities contain information that captures the relationship between the words known by a child at a specific time point and the words that child will learn next. Further, our results show that it matters how we integrate
these probabilities. For example, the maximal model is utilizing only minimal information from the conditional probability (the strongest conditional probability between known and candidate words only) and this model performs very poorly. This suggests that, even though conditional probabilities do contain useful information, not every use of it improves predictive power. The fact that the threshold model does best, suggests that understanding how to combine information can increase fit of the model and allow us to make more accurate future predictions. Interestingly, the model that integrated over the complete conditional probability matrix did not perform better than the model with less information. This result is not atypical for the world of child language acquisition and suggests that perhaps taking into account memory or other cognitive constraints may be useful, if not necessary, in capturing early learning (e.g. Phillips \& Pearl, 2012).

This work offers evidence that word learning is affected by a combination of forces and understanding these forces may allow us to predict words that a child would be likely to learn next. We would like to extend these results. Specifically we would like to more closely examine what types of relationships might exist and ways to measure them. If we understand the language environment where a child is learning as well as the way in which the child might be integrating this information with their current vocabulary we should be able to predict which words a child may learn next. This matters because this may allow us to capture children who have learning strategies leading to language difficulty or impairment. These types of models could let us diagnose such children earlier and may allow us to provide effective and child specific interventions.

Another potential direction is the development of tools and techniques that allow us to understand temporal dependencies at different time scales other than a month. Time series analysis combined with graph clustering on the semantics may allow us to expand this work from joint probability to a more complex probability space giving us better temporal resolution as well as more predictive models. Along those same lines, we may be able to finetune these models with cognitive theory (which are not included at all in these models, see Hills et al, 2009 for a paper that does consider this) to test generative and process motivated theories of word learning. This would allow us not only to build new computational tools but to refine and expound upon theories of word learning.

At the onset of this paper we asked whether it would be possible to predict the words a child will learn next from the words she knows now. Our findings, even with this simple set of models, suggest that the answer to that question is yes. Significantly, this opens up doors that have far-reaching implications. If we understand how children utilize their environment, conceptual understanding and semantic connectivity as they interact with the world and build up a vocabulary, we can design individualized teaching paradigms that may allow us to build upon, or compliment, what the child already knows aiding in language acquisition.

## References

Beckage, N., Smith, L., \& Hills, T. (2011). Small worlds and semantic network growth in typical and late talkers. PloS One, 6(5), e19348.
Dale, P.S., \& Fenson, L. (1996). Lexical development norms for young children. Behavior Research Methods, Instruments, \& Computers, 28, 125-127.
DeLoache, J.S., Simcock, G., \& Macari, S. (2007). Planes, Trains, Automobiles - and Tea Sets: Extremely Intense Interests in Very Young Children. Developmental Psychology, 43(6), 1576-1586.
Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., \& Stiles, J. (1994). Variability in early communicative development.Monographs of the society for research in child development.
Gentner, D. (2006). Why verbs are hard to learn. In K. Hirsh-Pasek, \& R. Golinkoff, (Eds.) Action meets word: How children learn verbs (pp. 544- 564). Oxford University Press.
Heilmann, J., Ellis Weismer, S., Evans, J., \& Hollar, C. (2005). Utility of the MacArthur-Bates Communicative Inventory in identifying language abilities of late talking and typically developing toddlers. American Journal of Speech-Language Pathology, 14, 40-51.
Hills, T., Maouene, M., Maouene, J., Sheya, A., \& Smith, L. (2009). Longitudinal analysis of early semantic networks: Preferential attachment or preferential acquisition? Psychological Science, 20,
Phillips, L., \& Pearl, L. (2012). "Less is More" in Bayesian word segmentation: When cognitively plausible learners outperform the ideal. In Proceedings of the 34th Annual Conference of the Cognitive Science Society.
Sandhofer, C.M., Smith, L.B., \& Luo, J. (2000). Counting nouns and verbs in the input: differential frequencies, different kinds of learning? J. Child Lang., 27, 561- 585.
Thal, D. J., O'Hanlon, L., Clemmons, M., \& Fralin, L. (1999). Validity of a parent report measure of vocabulary and syntax for preschool children with language impairment. Journal of Speech, Language and Hearing Research, 42(2), 482.
van Veen, R., Evers-Vermeul, J., Sanders, T.J.M. \& Bergh, H. van den (2009). Parental input and connective acquisition: a growth curve analysis. First Language, 29(3), 267-289.
Vitevitch, M.S. (2008). What can graph theory tell us about word learning and lexical retrieval? Journal of Speech, Language, and Hearing Research, 51, 408-422
Weizman, Z. O. \& Snow, C. E. (2001). Lexical output as related to children's vocabulary acquisition: Effects of sophisticated exposure and support for meaning. Developmental Psychology, 37, 265-279.

# Simultaneous unsupervised and supervised learning of cognitive functions in biologically plausible spiking neural networks 

Trevor Bekolay (tbekolay@uwaterloo.ca)<br>Carter Kolbeck (ckolbeck@uwaterloo.ca)<br>Chris Eliasmith (celiasmith@uwaterloo.ca)<br>Center for Theoretical Neuroscience, University of Waterloo<br>200 University Ave., Waterloo, ON N2L 3G1 Canada


#### Abstract

We present a novel learning rule for learning transformations of sophisticated neural representations in a biologically plausible manner. We show that the rule, which uses only information available locally to a synapse in a spiking network, can learn to transmit and bind semantic pointers. Semantic pointers have previously been used to build Spaun, which is currently the world's largest functional brain model (Eliasmith et al., 2012). Two operations commonly performed by Spaun are semantic pointer binding and transmission. It has not yet been shown how the binding and transmission operations can be learned. The learning rule combines a previously proposed supervised learning rule and a novel spiking form of the BCM unsupervised learning rule. We show that spiking BCM increases sparsity of connection weights at the cost of increased signal transmission error. We also demonstrate that the combined learning rule can learn transformations as well as the supervised rule and the offline optimization used previously. We also demonstrate that the combined learning rule is more robust to changes in parameters and leads to better outcomes in higher dimensional spaces, which is critical for explaining cognitive performance on diverse tasks.


Keywords: synaptic plasticity; spiking neural networks; unsupervised learning; supervised learning; Semantic Pointer Architecture; Neural Engineering Framework.

In this paper, we demonstrate learning of cognitively relevant transformations of neural representations online and in a biologically plausible manner. We improve upon a technique previously presented in MacNeil and Eliasmith (2011) by combining their error-minimization learning rule with an unsupervised learning rule, making it more biologically plausible and robust.

There are three weaknesses with most previous attempts at combining supervised and unsupervised learning in artificial neural networks (e.g., Backpropagation [Rumelhart, Hinton, \& Williams, 1986], Self-Organizing Maps [Kohonen, 1982], Deep Belief Networks [Hinton, Osindero, \& Teh, 2006]). These approaches 1) have explicit offline training phases that are distinct from functional use of the network, 2) require many layers with some layers connected with supervised learning and others with unsupervised learning, and 3) use non-spiking neuron models. The approach proposed here

[^8]overcomes these limitations. Our approach 1) remains functional during online learning, 2) requires only two layers connected with simultaneous supervised and unsupervised learning, and 3) employs spiking neuron models to reproduce central features of biological learning, such as spike-timing dependent plasticity (STDP).

Online learning with spiking neuron models faces significant challenges due to the temporal dynamics of spiking neurons. Spike rates cannot be used directly, and must be estimated with causal filters, producing a noisy estimate. When the signal being estimated changes, there is some time lag before the spiking activity reflects the new signal, resulting in situations during online learning in which the inputs and desired outputs are out of sync. Our approach is robust to these sources of noise, while only depending on quantities that are locally available to a synapse.

Other techniques doing similar types of learning in spiking neural networks (e.g., SpikeProp [Bohte, Kok, \& Poutre, 2002], ReSuMe [Ponulak, 2006]) can learn only simple operations, such as learning to spike at a specific time. Others (e.g., SORN [Lazar, Pipa, \& Triesch, 2009], reservoir computing approaches [Paugam-Moisy, Martinez, \& Bengio, 2008]) can solve complex tasks like classification, but it is not clear how these approaches can be applied to a general cognitive system. The functions learned by our approach are complex and have already been combined into a general cognitive system called the Semantic Pointer Architecture (SPA). Previously, the SPA has been used to create Spaun, a brain model made up of 2.5 million neurons that can do eight diverse tasks (Eliasmith et al., 2012). Spaun accomplishes these tasks by transmitting and manipulating semantic pointers, which are compressed neural representations that carry surface semantic content, and can be decompressed to generate deep semantic content (Eliasmith, in press). Semantic pointers are composed to represent syntactic structure using a "binding" transformation, which compresses the information in two semantic pointers into a single semantic pointer. Such representations can be "collected" using superposition, and collections can participate in further bindings to generate deep structures. Spaun performs these transformations by using the Neural Engineering Framework (NEF; Eliasmith \& Anderson, 2003) to directly compute static connection weights between populations. We show that our approach can learn to transmit, bind, and classify semantic pointers.

## Theory

## Cognitive functions with spiking neurons

In order to characterize cognitive functions at the level of spiking neurons, we employ the methods of the Semantic Pointer Architecture (SPA), which was recently used to create the world's largest functional brain model (Spaun; Eliasmith et al., 2012), able to perform perceptual, motor, and cognitive functions. Cognitive functions in Spaun include working memory, reinforcement learning, syntactic generalization, and rule induction.

The SPA is implemented using the principles of the Neural Engineering Framework (NEF; Eliasmith \& Anderson, 2003), which defines methods to 1) represent vectors of numbers through the activity of populations of spiking neurons, 2) transform those representations through the synaptic connections between those populations, and 3) incorporate dynamics through connecting populations recurrently. In the case of learning, we exploit NEF representations in order to learn transformations analogous to those that can be found through the NEF's methods.

Representing semantic pointers in spiking neurons Representation in the NEF is similar to population coding, as proposed by Georgopoulos, Schwartz, and Kettner (1986), but extended to $n$-dimensional vector spaces. Each population of neurons represents a point in an $n$-dimensional vector space over time. Each neuron in that population is sensitive to a direction in the $n$-dimensional space, which we call the neuron's encoder. The activity of a neuron can be expressed as

$$
\begin{equation*}
a=G\left[\mathbf{\alpha e} \cdot \mathbf{x}+J_{b i a s}\right], \tag{1}
\end{equation*}
$$

where $G[\cdot]$ is the nonlinear neural activation function, $\alpha$ is a scaling factor (gain) associated with the neuron, $\mathbf{e}$ is the neuron's encoder, $\mathbf{x}$ is the vector to be encoded, and $J_{\text {bias }}$ is the background current of the cell.

The vector (i.e., semantic pointer) that a population represents can be estimated from the recent activity of the population. The decoded estimate, $\hat{\mathbf{x}}$, is the sum of the activity of each neuron, weighted by an $n$-dimensional decoder.

$$
\begin{equation*}
\hat{\mathbf{x}}(t)=\sum_{i} \mathbf{d}_{i} a_{i}(t) \tag{2}
\end{equation*}
$$

where $\mathbf{d}_{i}$ is the decoder and $a_{i}$ is the activity of neuron $i$.
Neural activity is interpreted as a filtered spike train; i.e.,

$$
\begin{equation*}
a_{i}(t)=\sum_{s} h\left(t-t_{s}\right)=\sum_{s} e^{-\left(t-t_{s}\right) / \tau_{P S C}} \tag{3}
\end{equation*}
$$

where $h(\cdot)$ is the exponential filter applied to each spike, and $s$ is the set of all spikes occurring before the current time $t$.

The decoders are found through a least-squares minimization of the difference between the decoded estimate and the actual encoded vector.

$$
\begin{equation*}
\mathbf{d}=\Upsilon^{-1} \Gamma \quad \Gamma_{i j}=\int a_{i} a_{j} d x \quad \Upsilon_{j}=\int a_{j} \mathbf{x} d x \tag{4}
\end{equation*}
$$

Transforming semantic pointers through connection weights The encoders and decoders used to represent semantic pointers also enable arbitrary transformations (i.e., mathematical functions) of encoded semantic pointers. If population A encodes pointer $X$, and we want to connect it to population B , encoding pointer $Y$, a feedforward connection with the following connection weights transmits that semantic pointer, such that $Y \approx X$.

$$
\begin{equation*}
\omega_{i j}=\alpha_{j} \mathbf{e}_{j} \mathbf{d}_{i} \tag{5}
\end{equation*}
$$

where $i$ indexes the presynaptic population A , and $j$ indexes the postsynaptic population B . Other linear transformations are implemented by multiplying $\mathbf{d}_{i}$ by a linear operator. Nonlinear transformations are implemented by solving for a new set of decoding weights. This is done by minimizing the difference between the decoded estimate of $f(\mathbf{x})$ and the actual $f(\mathbf{x})$, rather than just $\mathbf{x}$, in Equation (4).

## Supervised learning: PES rule

MacNeil and Eliasmith (2011) proposed a learning rule that minimizes the error minimized in Equation (4) online.

$$
\begin{align*}
\Delta \mathbf{d}_{i} & =\kappa \mathbf{E} a_{i} \\
\Delta \omega_{i j} & =\kappa \alpha_{j} \mathbf{e}_{j} \cdot \mathbf{E} a_{i} \tag{6}
\end{align*}
$$

where $\kappa$ is a scalar learning rate, $\mathbf{E}$ is a vector representing the error we wish to minimize, and other terms are as before.

When put in terms of connections weights $\left(\omega_{i j}\right)$, the rule resembles backpropagation. The quantity $\alpha_{j} \mathbf{e}_{j} \cdot \mathbf{E}$ is analogous to the local error term $\delta$ in backpropagation (Rumelhart et al., 1986); they are both a means of translating a global error signal to a local error signal that can be used to change an individual synaptic connection weight. The key difference between this rule and backpropagation is that the global-tolocal mapping is done by imposing the portion of the error vector space each neuron is sensitive to via its encoder. This limits flexibility, but removes the dependency on global information, making the rule biologically plausible. We will refer to Equation (6) as the Prescribed Error Sensitivity (PES) rule.

## Unsupervised learning: Spiking BCM rule

A Hebbian learning rule widely studied in the context of the vision system is the BCM rule (Bienenstock, Cooper, \& Munro, 1982). This rule has been used to explain orientation selectivity and ocular dominance (Bienenstock et al., 1982). Theoretically, it has been asserted that BCM is equivalent to triplet-based STDP learning rules (Pfister \& Gerstner, 2006).

The general form is

$$
\begin{equation*}
\Delta \omega_{i j}=a_{i} a_{j}\left(a_{j}-\theta\right) \tag{7}
\end{equation*}
$$

where $\theta$ is the modification threshold. When the postsynaptic activity, $a_{j}$, is greater than the modification threshold, the synapse will be potentiated; when the postsynaptic activity is less than the modification threshold, the synapse will be depressed.

The modification threshold reflects the expectation of a cell's activity. It is typically calculated as the temporal average of the cell's activity over a long time window (on the order of hours). The intuition behind BCM is that cells driven above their expectation must be playing an important role in a circuit, so their afferent synapses become potentiated. Cells driven less than normal have synapses depressed. If either of these effects persists long enough, the modification threshold changes to reflect the new expectation of the cell's activity.

However, BCM as originally formulated is based on nonspiking rate neurons. We implement BCM in biologically plausible spiking networks by interpreting neural activity as spikes that are filtered by a postsynaptic current curve.

$$
\begin{align*}
\Delta \omega_{i j} & =\kappa \alpha_{j} a_{i} a_{j}\left(a_{j}-\theta\right) \\
\theta(t) & =e^{-t / \tau} \theta(t-1)+\left(1-e^{-t / \tau} a_{j}(t)\right), \tag{8}
\end{align*}
$$

where $a$, the activity of a neuron, is interpreted as a filtered spike train, as in Equation (3), and $\tau$ is the time constant of the modification threshold's exponential filter.

With our spiking BCM implementation, we aim to test the claim that BCM is equivalent to triplet STDP rules. Functionally, we hypothesize that unsupervised learning will have a small detrimental effect on the function being computed by a weight matrix, but will result in weight sparsification.

## Simultaneous supervised and unsupervised learning: hPES rule

The PES rule gives us the ability to minimize some provided error signal, allowing a network to learn to compute a transformation online. However, biological synapses can change when no error signal is present. More practically, transformation learning may be easier in more sparse systems. For these reasons, we propose a new learning rule that combines the error-minimization abilities of the PES rule with the biological plausibility and sparsification of the spiking BCM rule. The rule is a weighted sum of the terms in each rule.

$$
\begin{equation*}
\Delta \omega_{i j}=\kappa \alpha_{j} a_{i}\left(S \mathbf{e}_{j} \cdot \mathbf{E}+(1-S) a_{j}\left(a_{j}-\theta\right)\right) \tag{9}
\end{equation*}
$$

where $0 \leq S \leq 1$ is the relative weighting of the supervised term over the unsupervised term. Note that this rule is a generalization of the previously discussed rules; if we set $S=1$, this rule is equivalent to PES, and if we set $S=0$, this rule is equivalent to spiking BCM.

We hypothesize that the unsupervised component helps maintain homeostasis while following the error gradient defined by the supervised component. Because of this, we call the rule the homeostatic Prescribed Error Sensitivity (hPES) rule. We hypothesize that this rule will be able to learn the same class of transformations that the PES rule can learn, and that this class includes operations critical to cognitive representation, such as semantic pointer binding.

Note that a similar combination can be done with other supervised learning rules that modify connection weights. A more general form of Equation (9) would replace the local error quantity $\mathbf{e}_{j} \cdot \mathbf{E}$ with $\delta$, which could be determined through
any method. However, for clarity, we will use hPES to refer to the specific form in Equation (9).

## Methods

We performed three experiments in order to test our hypotheses about the hPES rule. Experiments were implemented in the Nengo simulation environment, in which we implemented the PES, spiking BCM, and hPES learning rules. All experiments use leaky integrate-and-fire neurons with default parameters. The scripts used to implement and analyze these experiments are MIT licensed and available at http://github.com/tbekolay/cogsci2013.
Experiment 1: Unsupervised learning We constructed a network composed of two populations connected in a feedforward manner such that one population provides input to the other. The network can be run in a "control" regime, in which the weights between the two populations are solved for with the NEF's least-squares optimization and do not change, or in a "learning" regime, in which the weights are the same as the control network, but change over time according to the hPES rule (9) with $S=0$ (i.e., according to the spiking BCM rule [8]). This experiment tests the hypothesis that the unsupervised component of the hPES rule increases sparsity of the connection weight matrix.
Experiment 2: Supervised learning We constructed a network composed of two populations connected in a feedforward manner, and one population that provides an error signal to the downstream population. The network can be run in a "control" regime, in which the weights between the two populations are computed to transmit a three-dimensional semantic pointer, or to bind two three-dimensional semantic pointers into one three-dimensional pointer. The network can be run in a "learning" regime, in which the weights between the two populations are initially random and are modified over time by the hPES rule. This experiment tests the hypothesis that the hPES rule can learn to transmit and bind semantic pointers as well as the control network and the supervised learning rule (i.e., hPES with $S=1$ ).
Experiment 3: Digit classification In order to investigate how simultaneous supervised and unsupervised learning scales in higher dimensional situations, we constructed a network similar to that in Experiment 2, but whose input is handwritten digits.* In order to be computationally tractable, the 28-by-28 pixel images were compressed to a 50dimensional semantic pointer using a sparse deep belief network that consists of four feedforward Restricted Boltzmann Machines trained with a form of contrastive divergence (full details in Tang \& Eliasmith, 2010). Those 50-dimensional pointers were projected to an output population of 10 dimensions, where each dimension represents the confidence that the input representation should be classified into one of the 10 possible digits. The classified digit is the one corresponding

[^9]to the dimension with the highest activity over 30 ms when a 50 -dimensional input is presented. 60,000 labeled training examples were shown to the network while the hPES rule was active. The network was then tested with 10,000 training examples in which the label was not provided. The results are compared to an analogous control network, in which the 50-dimensional pointers are classified with a cleanup memory whose connection weights are static, as in Eliasmith et al. (2012). This experiment examines how well the hPES rule scales to high-dimensional spaces.
Learning parameters While there are many hundreds of parameters involved in each network simulation, the vast majority are randomly selected within a biologically plausible range without significantly affecting performance. Some parameters, especially those affecting the learning rule, can have a significant performance effect. These significant parameters and the values used in specific simulations are listed in Tables 1 and 2. These parameters were optimized with a tree of Parzens estimators approach, using the hyperopt package (Bergstra, Yamins, \& Cox, 2013).

Table 1: Parameters used for transmitting semantic pointers

| Parameter | Description | Value |
| :--- | :--- | :---: |
| $N / D$ | Neurons per dimension | 25 |
| $\kappa$ | Learning rate | $3.51 \times 10^{-3}$ |
| $S$ | Supervision ratio | 0.798 |
| $\kappa$ for $S=1$ | Learning rate (PES) | $2.03 \times 10^{-3}$ |

Table 2: Parameters used for binding and classifying SPs

| Parameter | Description | Value |
| :--- | :--- | :---: |
| $N / D$ | Neurons per dimension | 25 |
| $\kappa$ | Learning rate | $2.38 \times 10^{-3}$ |
| $S$ | Supervision ratio | 0.725 |
| $\kappa$ for $S=1$ | Learning rate (PES) | $1.46 \times 10^{-3}$ |

## Results

## hPES replicates STDP results

Previously, Pfister and Gerstner (2006) have theorized that BCM and STDP are equivalent. Our experiments support this theory. Varying the amount of time between presynaptic and postsynaptic spikes results in an STDP curve extremely similar to the classical Bi and Poo (2001) STDP curve (Figure 1).

However, these STDP curves do not capture the frequency dependence of STDP. In order to capture those effects, modellers have created STDP rules that take into account triplets and quadruplets of spikes, rather than just pre-post spike pairings (Pfister \& Gerstner, 2006). These rules are able to replicate the frequency dependence of the STDP protocol. Figure 2 shows that, despite being a much simpler rule, the hPES rule with $S=0$ also exhibits frequency dependence.


Figure 1: STDP curve replicated by two neurons connected with the hPES rule, $S=0$. Solid and dashed lines are best fits of the curve $a e^{-x / \tau}$ for the experimental and simulated data, respectively. Experimental data from Bi and Poo (2001).

## hPES encourages sparsity while increasing signal transmission error

When the hPES rule is applied to a network that has been optimized to implement semantic pointer transmission, the hPES rule with no supervision ( $S=0$ ) increases signal transmission error at a rate proportional to the learning rate. Figure 3 (top) shows the gradual decrease in accuracy over 200 seconds of simulation with an artificially large learning rate. Figure 3 (bottom) shows that the sparsity of the weight matrix, as measured by the Gini index, increases over time. Therefore, the hPES rule with $S=0$ increases network sparsity at the cost of an increase in signal transmission error.

## hPES can learn cognitive functions

The error in the learned networks relative to the mean of 10 control networks can be seen decreasing over time in Figure 4. The parameters used for Figure 4 are listed in Tables 1 and 2. The transformations are learned quickly (approximately 25 seconds for transmission, 45 seconds for binding). Therefore, the central binding and transmission operations of the SPA are learnable with the hPES rule.

Critically, while hPES without error sensitivity introduces error while increasing sparsity (see Figure 3), with error sensitivity, this error can be overcome. Interestingly, binding, the intuitively more complex operation, is more reliably learned than transmission. This is due to how effectively the NEF can optimize weights to perform linear transformations like transmission in the control networks.

As a proof of concept that the hPES rule scales to highdimensional spaces, Table 3 shows that the learned handwritten digit classification network classifies digits more accurately than Spaun's cleanup memory (Eliasmith et al., 2012).


Figure 2: STDP frequency dependence replicated by two neurons connected with the hPES rule, $S=0$. This also demonstrates the effect of different $\theta$ values on the frequency dependence curve. Experimental data from Kirkwood et al. (1996).

This supports the suggestion that the hPES rule scales to high-dimensional spaces. While the hPES rule with $S<1$ achieved higher classification accuracy than hPES with $S=1$, not enough trials were attempted to statistically confirm a benefit to combined unsupervised and supervised learning for classifying handwritten digits.

Table 3: Classification accuracy of handwritten digits

| Classification technique | Accuracy |
| :--- | :--- |
| Cleanup memory (Spaun) | $94 \%$ |
| hPES learning, $S=1$ | $96.31 \%$ |
| hPES learning | $98.47 \%$ |

## hPES is less sensitive to parameters

The parameters of the hPES rule were optimized with $S$ fixed at 1 for 50 simulations of binding and transmission. A separate parameter optimization that allowed $S$ to change was done for 50 simulations of binding and transmission.

Surprisingly, despite optimizing over an additional dimension, when $S$ was allowed to change, error rates were lower during the optimization process for the binding network but not the transmission network. In both cases, the interquartile range of the hPES rule's performance when $S$ was allowed to change is lower. Figure 5 summarizes the performance of all 200 networks generated for parameter optimization. While in all four cases parameters were found that achieve error rates close to the control networks, hPES was more robust to changes in parameters when $S$ was allowed to change. This suggests that unsupervised learning may be beneficial in high-dimensional nonlinear situations.


Figure 3: Data gathered from 50 trials of Experiment 1; filled region represents a bootstrapped $95 \%$ confidence interval. (Top) Accuracy of signal transmission. Accuracy is proportional to negative mean squared error, scaled such that accuracy of 1.0 denotes a signal identical to that transmitted by the NEF optimal weights with no learning, and accuracy of 0.0 represents no signal transmission (i.e., the error is equal to the signal). (Bottom) Sparsity of the connection weight matrix over time, as measured by the Gini index. This demonstrates the expected tradeoff between sparsity and accuracy.

## Discussion

In this paper, we have presented a novel learning rule for learning cognitively relevant transformations of neural representations in a biologically plausible manner. We have shown that the unsupervised component of the rule increases sparsity of connection weights at the cost of increased signal transmission error. We have also shown that the combined learning rule, hPES, can learn transformations as well as the supervised rule and the offline optimization done in the Neural Engineering Framework. We have demonstrated that the combined learning rule is more robust to changes in parameters when learning nonlinear transformations.

However, it is still the case that the parameters of the learning rule were optimized for each transformation learned. This is a challenge shared by all learning rules, but in the context of biologically plausible simulations, there is the additional question of the biological correlate of these parameters. It could be the case that these parameters are a result of the structure of the neuron, and therefore act as a fixed property of the neuron. However, it could also be the case that these parameters are related to the activity of the network, and are modified by each neuron's activity, or by the activity of some external performance monitoring signal. Examining these possibilities is the subject of future work.


Figure 4: Error of learning networks over time compared to the mean error of 10 control networks. Each type of network is generated and simulated 15 times. For the binding network, every 4 seconds, the learning rule is disabled and error is accumulated over 5 seconds. For the transmission network, every 0.5 seconds, the learning rule is disabled, and error is accumulated over 2 seconds. Filled regions are bootstrapped 95\% confidence intervals. Time is simulated time, not computation time.

## Acknowledgments

This work was supported by the Natural Sciences and Engineering Research Council of Canada, Canada Research Chairs, the Canadian Foundation for Innovation and the Ontario Innovation Trust. We also thank James Bergstra for creating the hyperopt tool and assisting with its use.

## References

Bergstra, J., Yamins, D., \& Cox, D. D. (2013). Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In 30th international conference on machine learning.
Bi, G.-Q., \& Poo, M.-M. (2001). Synaptic modification by correlated activity: Hebb's postulate revisited. Annual Review of Neuroscience, 24(1), 139-66.
Bienenstock, E. L., Cooper, L. N., \& Munro, P. (1982). Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. Journal of Neuroscience, 2(1), 32.
Bohte, S. M., Kok, J. N., \& Poutre, H. L. (2002). SpikeProp: Backpropagation for networks of spiking neurons. Neurocomputing, 48(1-4), 17.
Eliasmith, C. (in press). How to build a brain: A neural architecture for biological cognition. New York, NY: Oxford University Press.
Eliasmith, C., \& Anderson, C. H. (2003). Neural engineering: computation, representation, and dynamics in neurobiological systems. Cambridge, MA: MIT Press.
Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., Tang, C., \& Rasmussen, D. (2012). A large-


Figure 5: Summary of error rates for networks used to optimize learning parameters. Each column summarizes 50 experiments of the labeled condition. Boxes indicate the median and inter-quartile range, and whiskers indicate the inner fence (i.e., $Q_{1}-1.5 I Q R$ and $Q_{3}+1.5 I Q R$ ). Outliers are not shown.
scale model of the functioning brain. Science, 338(6111), 1202-5.
Georgopoulos, A. P., Schwartz, A. B., \& Kettner, R. E. (1986). Neuronal population coding of movement direction. Science, 233(4771), 1416-1419.
Hinton, G., Osindero, S., \& Teh, Y. (2006). A fast Learning Algorithm for Deep Belief Nets. Neural computation, 1554, 1527-1554.
Kirkwood, A., Rioult, M. G., \& Bear, M. F. (1996). Experience-dependent modification of synaptic plasticity in visual cortex. Nature, 381(6582), 526-528.
Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. Biological Cybernetics, 43(1).
Lazar, A., Pipa, G., \& Triesch, J. (2009). Sorn: a selforganizing recurrent neural network. Frontiers in Computational Neuroscience, 3(23).
MacNeil, D., \& Eliasmith, C. (2011). Fine-tuning and the stability of recurrent neural networks. PloS One, 6(9).
Paugam-Moisy, H., Martinez, R., \& Bengio, S. (2008). Delay learning and polychronization for reservoir computing. Neurocomputing, 71(7), 1143-1158.
Pfister, J.-P., \& Gerstner, W. (2006). Triplets of spikes in a model of spike timing-dependent plasticity. Journal of Neuroscience, 26(38), 9673-82.
Ponulak, F. (2006). Supervised learning in spiking neural networks with ReSuMe method. Phd thesis, Poznan University of Technology.
Rumelhart, D. E., Hinton, G. E., \& Williams, R. J. (1986). Learning representations by back-propagating errors. Na ture, 323(6088), 533-536.
Tang, Y., \& Eliasmith, C. (2010). Deep networks for robust visual recognition. In 27th international conference on machine learning. Haifa, Il: ICML.

# Balancing Fairness and Efficiency in Repeated Societal Interaction 

Noam Ben-Asher (noamba@cmu.edu) ${ }^{1}$<br>Christian Lebiere (cl@cmu.edu) ${ }^{2}$<br>Alessandro Oltramari (aoltrama@andrew.cmu.edu) ${ }^{2}$<br>Cleotilde Gonzalez (coty@cmu.edu) ${ }^{1}$<br>${ }^{1}$ Dynamic Decision Making Laboratory, Department of Social and Decision Sciences, Carnegie Mellon University Pittsburgh, PA 15213 USA<br>${ }^{2}$ Department of Psychology, Carnegie Mellon University Pittsburgh, PA 15213 USA


#### Abstract

Fairness and efficiency are important aspects that influence cooperation in social dilemmas. During a repeated interaction, they have the potential to serve as competing goals for the decision maker. The ability to balance between fairness and efficiency depends, among other things, on available information regarding mutual accountability for the outcomes in an interaction. In this paper, we examine how information regarding mutual interdependencies influences the interplay between fairness and efficiency in repeated Chicken Game. We distinguish between three possible types of fair behavior: mutual cooperation, alternating cooperation, and mutual destruction. Our results show that the first two types of fairness are positively correlated with the availability of social information. In contrast, mutual destructive fairness is not sensitive to the availability of information and is generally avoided. We also find that without information regarding mutual interdependencies, unfairness increases in parallel with efficiency. When social information is available, however, increases in fairness is coupled with a decrease in efficiency, and the best compromise between fairness and efficiency is reached when mutual interdependencies are learned through repeated experiences. We highlight the significance of our results for fair and efficient interaction in repeated social interactions.


Keywords: Efficiency; Fairness; Cooperation; Game Theory; Interdependence; Information; Social Interaction

## Introduction and Background

In our daily lives, we constantly face situations in which our well-being and success depends on the actions of others. Whether the interactions occur between individuals or between organizations, there is mutual accountability for the outcomes. For example, the interaction between two toddlers that learn to share a toy by taking turns holds some similarities with the interaction between companies that adopt a competitive brinkmanship pricing policy while trying to gain control over a certain market. If the toddlers are not willing to behave in a fair manner they will both end up screaming and none of them will play with the desired toy. An alternative behavior is that both toddlers decide simultaneously to switch interest to other toys and thus eliminate the conflict. A more mutually beneficial behavior is where the toddlers will share the toy so one can play with
it for a while and then the other will have the joy of playing with the desired toy as well. However, if only one of the toddlers gets to play with the toy and the other does not get the same opportunity, feelings of unfairness and frustration that might lead to aggressive behavior when facing similar conflicts in the future can arise.

Such social conflicts are well captured in Game Theory using the Chicken Game (CG), as introduced by Russell (1959). According to this game, two drivers are heading towards each other on a single lane road from opposite directions at full speed. Just before colliding, each of the drivers has to choose simultaneously and independently between driving straight towards a possible collision (i.e., Dare) or turning the steering wheel (i.e., Swerve) and avoiding the accident. As represented in the game's payoff matrix (see Table 1), this is a prototypical dangerous game, because a player has to risk the lowest payoff [-10] to have a chance of winning the highest payoff [10]. Under reasonable assumptions for single-trial CG, the best outcome is for a player that Dares while the other player Swerves [10,-1]; the second-best for each is if both Swerve [1,1]; the third-best is for a player that Swerves while the other player Dares [$1,10]$; and the worst for each is if both Dare [-10,-10], because then the outcome is mutually destructive. Thus, the best strategy in single-trial CG depends on the opponent's expected behavior and a player can maximize the outcome by doing the opposite of what the other player does. However, for repeated CG (infinitely repeated in theory; finitely repeated with unknown endpoint in practice), successful alternations, where one player wins the highest payoff in one round and then the other player wins the highest payoff in the next round, is the best strategy to obtain a joint maximum outcome. This type of cooperation corresponds well to the situation where, over time, the wellbeing of one player depends of the well-being of the other.

The need to consider the well-being of the other challenges traditional economic theories, which assume that people act selfishly. In contrast, there is growing experimental evidence that actual behavior is also shaped by factors inconsistent with pure selfishness (e.g., Dufwenberg \& Kirchsteiger, 2004; Fehr \& Schmidt, 1999; Roth, 1995). The psychology literature also provides evidence for other
factors that individuals consider, beyond their own wellbeing, and acknowledges the importance of reciprocity by incorporating it into models of human behavior. Heider (1958) introduced the idea that causal inference, where one takes into account another person's motives and situational constraints, as an important cognitive process for perceiving social contexts. Similarly, Game Theory incorporates these ideas by considering altruism and fairness (Bolton \& Ockenfels, 2000; Fehr \& Schmidt, 1999). Altruism can be simply defined as an interaction in which people care not only about their own well-being but also about others. This over-simplified definition is extended through fairness into different directions by incorporating distributive concerns (Bolton \& Ockenfels, 2000), inequity aversion (Fehr \& Schmidt, 1999), and reciprocity theories (Rabin, 1993).

CG is particularly suitable to studying different aspects of fairness in social interaction. In its basic form, fairness is achieved when one player's outcome is identical to the other player's. Considering the payoff matrix presented in Table 1, fairness can happen when both players make the same decision in a given round. It is possible that both players selected Dare (i.e., mutual Dare) and therefore received an outcome of $[-10,-10]$, or when both players selected Swerve (i.e., mutual Swerve) and received the outcome of $[1,1]$. Repeated CG adds another kind of fair interaction where the players alternate for consecutive rounds, $[10,-1]$ followed by $[-1,10]$ and so on, resulting in alternating cooperation. This type of fairness is also the optimal strategy in repeated CG (the one that results in the highest long-term outcome). It is possible that fair coordination like [1, 1] would be easier to achieve than alternating coordination because the latter require a more complex coordination of choices. However, once a state of alternating coordination is achieved, it might be more stable compared to simple coordination (Rapoport, Guyer, \& Gordon, 1976).

One common but unrealistic assumption in research on strategic social interaction is that players possess full information about their interdependence, usually presented using a payoff matrix. However, several studies demonstrated the value of certain types of information in well-known social games, such as the repeated Prisoner's Dilemma (Camerer, 2003; Rapoport \& Chammah, 1965). Recently, Gonzalez and Martin (2011) proposed the Hierarchy of Social Information (HSI), a theoretical framework for conceptualizing and organizing the major categories of interpersonal information that may play a role in social interactions. Martin, Gonzalez, Juvina, and Lebiere (2012) used this framework to examine the effects of information on cooperation in repeated Prisoner's Dilemma. Their findings reveal a generally positive impact of information on joint performance and satisfaction. They showed that an increase in cooperation with an increase in the availability of social information was driven in part by players' greater willingness to reciprocate the other player's prior cooperation, and concluded that players possessing more interdependence information were more likely to enforce social norms of reward and retribution. Such social
norms depend heavily on perceptions of fairness and how it is expressed in repeated interaction. Furthermore, the availability of interdependence information can influence awareness of fair and unfair aspects of the interaction. Thus, it is possible to assume that having enough information to compare between one's own and another's payoffs might be a minimal requirement for fairness. However, it is currently unclear how more information systematically influences fairness, and more specifically the impact of different types of fairness that can occur as part of the interaction between two interdependent players. Furthermore, there is a need to understand the costs of maintaining fairness during a social interaction. Some social settings emphasize the tradeoff between fairness and performance. For example, in one-shot Ultimatum Game, a proposer maintains fairness and cooperation with a responder by decreasing her own personal gain. Other social settings completely disentangle such relationship between fairness and performance like in one-shot Dictator Game, where there is no need to maintain fairness and the proposer can keep all the gains for herself. Forsythe and colleagues (1994) demonstrated that on average offers in the Ultimatum game were higher than in the Dictator Game, confirming the tradeoff between efficiency of cooperation and fairness. However, people do care about fairness and in some situations are willing to sacrifice some of their own to maintain fairness (Bolton \& Ockenfels, 2000, Güth, Kliemt, \& Ockenfels). Thus, it is possible that a fair distribution of payoffs is appealing from a social perspective, but it might have some negative influence on performance in repeated interaction.

Table 1: Chicken game payoff matrix, with Action A denoting Dare and Action B denoting Swerve. The cells show a pair of outcomes $(x, y)$ where $x$ is the payoff to Player 1 and y is the payoff to Player 2.

|  | Player 2 Action |  |  |
| :--- | :---: | :---: | :---: |
|  |  | $\mathrm{A}($ Dare $)$ | B (Swerve) |
| Player 1 | A (Dare) | $-10,-10$ | $10,-1$ |
|  | Bction | B (Swerve) | $-1,10$ |

The current paper presents an experiment to examine how information regarding mutual interdependencies can influence fairness in CG. We start by providing background information on a repeated CG game that was used to collect the behavioral data, and present the four levels of information proposed in the HSI framework (Gonzalez \& Martin, 2011). Then, we distinguish between three types of fair outcomes in repeated CG and examine the interactions between fairness and the availability of social information. Following that, we focus on the relations between fairness and efficiency in alternating cooperation. Finally, we discuss how awareness of interdependence may encourage efficient and fair behavior in real-world interactions and describe potential future directions of this research.

## Experiment

## Participants and Procedure

Participants ( $\mathrm{N}=240 ; 120$ pairs, $45 \%$ of whom were women, $\mathrm{M}_{\text {age }}=22.8, \mathrm{SD}_{\text {age }}=4.53$ ) were recruited to a computer laboratory at Carnegie Mellon University, and randomly paired to play 200 rounds of repeated CG over the laboratory's network. Players were not told the number of rounds in the game, and the rounds were not numbered in the course of play. In each round, the two anonymous members of a pair (seated in different rooms without having met one another) chose simultaneously between buttons labeled Action A and Action B with payoffs as in Table 1. These payoffs were converted to incentive pay (one cent per point) beyond $\$ 10$ base pay. One pair was excluded from the analysis due to communication failure during the experiment.

## Information Conditions

All participants saw their own action and payoff in each round. Thirty pairs ( 60 participants) were assigned to each of the four conditions that determined the amount of information available about their interdependence. These conditions were modeled after the layers of the Hierarchy of Social Information (HSI) outlined by Gonzalez and Martin (2011). We briefly describe each of the information conditions here.

In the "Individual" condition, players were not informed that they were interacting with another player, so the selection of an action in each round was most likely perceived as an independent binary choice between two options with probabilistic payoffs. Participants may have realized that the outcome probabilities for each of the two actions were not static, as they in fact varied with the other player's actions, but this could more easily be attributed to a computerized process that shifted exogenously or in response to their own actions.

In the "Minimal" condition, players knew that their outcomes depended on the actions of another player and vice versa, yet they still did not know the other's specific actions and payoffs. With this information, individuals may have been able to speculate about the other's motivations, but it would remain difficult to infer the other's actions and payoffs.

Next, pairs in the "Experiential" condition saw each other's actions and outcomes in each round. This information allowed players to reason about the mutual interdependencies through repeated experiences.

Finally, in the "Descriptive" condition, in addition to seeing each other's actions and outcomes, players were shown the complete payoff matrix (as in Table 1) from the outset and throughout the repeated interaction.

## Results

We analyzed the three different fair outcomes described above: mutual Swerve, where both players receive a small
positive payoff of +1 with the risk of experiencing a moderate loss of -1 if the other player deviates from fairness; mutual Dare where both players lose 10 points with the potential of winning +10 if the other player deviates; and the alternating cooperation, which requires that both players successfully alternate between Swerve and Dare in a way that the payoffs in one round are $[+10,-1]$ and players then get $[-1,+10]$ in the following round.

## General Fairness Preferences

Overall, the joint proportion of fair interactions increased with more information given: .38 in the Individual condition, .50 in the Minimal condition, .71 in the Experiential condition, and .74 in the Descriptive condition. However, as illustrated in Figure 1, behavior varied for the different types of fairness. Next, we examine each of the fair interactions separately.

## Alternating Cooperation

The average proportion of alternating cooperation varied significantly across the four information conditions and generally increased with greater availability of information, $F(3,115)=7.086, p<.001$. As shown in
Figure 1, the average proportion of alternating cooperation in the Individual condition was extremely low ( $\mathrm{M}=.03$, $\mathrm{SD}=.05$ ). Knowing that the payoffs depend on the decisions of another human player provided in the Minimal condition ( $\mathrm{M}=.1, \mathrm{SD}=19$ ) increased the alternating cooperation significantly, $t(57)=2.106, p=.040$. The average proportion of alternating cooperation increased substantially more in the Experiential condition ( $\mathrm{M}=.29, \mathrm{SD}=.34$ ), where each player observed the actions and payoffs of the other player, and was significantly higher than the Individual and Minimal conditions, $t(58)=4.202, p<.001$ and $t(57)=2.575$, $p=.012$, respectively. Similarly, in the Descriptive condition ( $\mathrm{M}=.22, \mathrm{SD}=.28$ ) the average proportion was significantly higher than the Individual condition and marginally higher than the Minimal condition, $t(58)=3.765, p<.001$ and $t(57)=1.880, p=.065$, respectively. No significant differences were found between the Descriptive and Experiential conditions, $t(58)=.847, p=\mathrm{ns}$.

## Mutual Swerve

The average proportion of mutual Swerve varied significantly across the four information conditions, $F(3,115)=2.78, p=.044$. As shown in Figure 1, on average, the proportion of fair rounds where both players swerved increased with the availability of information: $\mathrm{M}=21$ ( $\mathrm{SD}=.23$ ) in the Individual condition, $\mathrm{M}=.21$ ( $\mathrm{SD}=.17$ ) in the Minimal condition, $\mathrm{M}=.27(\mathrm{SD}=.24)$ in the Experiential condition, and $\mathrm{M}=0.36$ ( $\mathrm{SD}=.27$ ) in the Descriptive condition. The average proportion of mutual Swerve in the Descriptive condition was significantly higher than in the Individual and Minimal conditions, $t(58)=2.341, p=.023$ and $t(57)=2.466, p=.017$, respectively. These results suggest that participants in the Descriptive condition tended to prefer the less risky option that could yield a small gain but also a
small loss compare to the risky option that could yield a large gain or a large loss. Also, the ability to see the interdependencies as they are represented by the payoff matrix highlights the fairness of mutual Swerve. Thus, it is possible that some synergy between these two interpretations of the social information can explain the increased proportion of mutual Swerve.

## Mutual Dare

When both players select Dare in the same round, they both receive the same negative payoff. As seen in Figure 1, the average proportion of mutual Dare was not affected by the availability of information, $F(3,115)=.37, p=\mathrm{ns}$. The average proportion of rounds where both players dared remained similar while the available information increased: $\mathrm{M}=.15$ ( $\mathrm{SD}=.08$ ) in the Individual condition, $\mathrm{M}=.18$ ( $\mathrm{SD}=.10$ ) in the Minimal condition, $\mathrm{M}=.15$ ( $\mathrm{SD}=.11$ ) in the Experiential condition, and $\mathrm{M}=.16(\mathrm{SD}=.21)$ in the Descriptive contrition. The SD in the Descriptive condition is relatively higher compared to the other conditions, mainly due to two pairs in the Descriptive condition who mutually dared in more than $80 \%$ of the 200 rounds.


Figure 1. Average proportion of fair outcomes for four levels of social information across 200 rounds.

## Efficiency and Fairness tradeoff in Alternating Cooperation

Rapoport et al. (1976) suggested the alternating cooperation index $(k)$ as a measure suitable for evaluating cooperation between two players in games where the optimal collective strategy is coordinated alternations. This measure uses the frequencies of the asymmetric payoffs in the following form: $\mathrm{k}=(\mathrm{DS})+(\mathrm{SD})-|(\mathrm{DS})-(\mathrm{SD})|$

Where DS and SD refer to the number of times that the asymmetric payoffs occurred. The maximal value of $k$ is achieved when DS and SD occur exclusively and with equal frequencies, which would correspond to perfect alternation.

Considering the setting of the study presented here, the maximal value of $k$ is 200 and it is achieved when $\mathrm{DS}=\mathrm{SD}=100$. The minimal value of $k$ is 0 and it represents a
game in which one player dominated the other player throughout the 200 rounds, resulting in $\mathrm{DS}=200$ or $\mathrm{SD}=200$.

The sum of DS and SD reflects the exclusiveness of the asymmetric payoffs compared to other possible payoffs. Thus, it represents the efficiency of the alternating cooperation. On the other hand, the absolute value of the difference between DS and SD reflects the balance or fairness between the two players and serves as a penalty for unfair behavior. The magnitude of the penalty determines whether one player dominated the other. For the sake of simplicity and clarity, we decompose $k$ to its terms (i.e., efficiency and unfairness penalty) and converted them to proportions by dividing each term by the number of rounds (i.e., 200). Furthermore, we use 1 - the proportion of unfairness penalty to calculate the proportion of fairness.

Similar to alternating cooperation, $k$ varied significantly across the information conditions. We observed a general increase of $k$ with more information: $\mathrm{M}=.25(\mathrm{SD}=.18)$ in the Individual condition, $\mathrm{M}=.32$ ( $\mathrm{SD}=.23$ ) in the Minimal condition, $\mathrm{M}=.46$ ( $\mathrm{SD}=.28$ ) in the Experiential condition, and $\mathrm{M}=.39$ ( $\mathrm{SD}=.28$ ) in the Descriptive condition.

To gain a better understanding of how the availability of social information influenced the alternating cooperation index ( $k$ ) and especially the relations between its efficiency and fairness components, we analyzed each of the terms that construct $k$ separately. As seen in Figure 2, fairness increased significantly with the availability of social information, $F(3,115)=12.778, p<.001$. The average fairness in the Descriptive condition ( $\mathrm{M}=.91, \mathrm{SD}=.11$ ) was significantly higher than the average fairness in the Individual ( $\mathrm{M}=.61, \mathrm{SD}=.26$ ) and Minimal ( $\mathrm{M}=.72, \mathrm{SD}=.28$ ) conditions, $t(58)=5.826, p<.001$ and $t(57)=3.601, p<.001$, respectively. Similarly, the average fairness in the Experiential condition ( $\mathrm{M}=.88, \mathrm{SD}=.18$ ) was significantly higher than in the Individual and Minimal conditions, $t(58)=4.817, p<.001$ and $t(57)=2.869, p=.005$.

As shown in Figure 2, the analysis of the efficiency term also indicated that efficiency varied significantly across the information conditions, $F(3,115)=3.030, p=.032$. However, we found an trend opposite of fairness, where efficiency in the Descriptive condition ( $\mathrm{M}=.48, \mathrm{SD}=.25$ ) was significantly lower compared to the Individual ( $\mathrm{M}=.64$, $\mathrm{SD}=.21$ ) and Minimal ( $\mathrm{M}=.61, \mathrm{SD}=.17$ ) conditions, $t(58)=-$ 2.756, $p=.008, t(57)=2.296, p=.025$.

The analysis above indicates that the availability of social information influenced the alternating cooperation index in general and differentially influenced the efficiency and fairness terms that compose it. It seems that the relatively high fairness in the Experiential and Descriptive came at the cost of decreased efficiency. However, the decrease in efficiency when moving from the Minimal condition to the Experiential condition is not significant, while the increase in fairness is. This suggests that best compromise between fairness and efficiency was reached in the Experiential condition.


Figure 2: Average proportion of efficient and fair alternating cooperation for four levels of social information across 200 rounds.

To gather a better understanding of the relations between fairness and efficiency, we analyze efficiency and fairness at the pair level, for each of the information conditions (see Figure 3). Results indicate that efficiency and fairness were strongly and negatively correlated in the Individual and Minimal conditions, $r=-.73, p<.001$ and $r=-.56, p=.002$, respectively, in contrast to the Experiential and Descriptive conditions, where no significant correlations were found, $r=.10, p=\mathrm{ns}$. and $r=-.02, p=\mathrm{ns}$. This finding suggests that as the efficiency of alternating cooperation increased in the Individual and Minimal conditions, fairness between the two players decreased. This means that one player dominated the other more often, resulting in an unfair and heedless behavior.

## Discussion and Conclusion

We tested how fairness is influenced by the availability of social information regarding mutual accountability in a social dilemma, as represented by the Chicken Game. Drawing upon Gonzalez and Martin's (2011) Hierarchy of Social Information, we find that fairness in such social dilemmas depended on the availability of information and in general increased when more information was available to the decision makers. Thus, information availability moderates unfairness and can increase fair social behaviors.

First, the overall high proportion of fair outcomes in the Descriptive and Experiential conditions implies that the availability of detailed information regarding interdependencies with others elicit fair behavior. This holds mainly for the two constructive interactions (mutual cooperation and alternating cooperation), which leads to positive outcomes for both sides of the conflict. On the other hand, when there is no awareness of the counterpart's conditions, unfairness increases on the expense of fairness. This results in an antisocial interaction where one exploits the other, even unintentionally and without being aware of


Figure 3: Correlations between proportions of efficiency and fairness in alternating cooperation for each pair, for four levels of social information, across 200 rounds.
the nature of the interaction.
Mutual destructive fairness stands out from the other two types of fairness as it was not influenced by the availability of information. Even without the knowledge of interdependence, players managed to avoid the mutually destructive escalation of the conflict which leads to negative outcomes for both sides. However, individuals possessing interdependence information were more likely to behave in a social manner and preferred other fair interactions. In contrast, individuals that did not possess such information were more likely to exhibit unfair and unsocial behavior where one exploits the other.

We also find that when only experiential information is available, the proportions of mutual cooperation and alternating cooperation are relatively similar. In contrast, when descriptive information is also provided, there is a greater preference towards the fair outcome resulting from mutual cooperation, compared to alternating cooperation. This might relate to the increased complexity of coordination that alternating cooperation requires, compared to mutual cooperation (Rapoport et al., 1976). Alternatively, it is possible that players concluded from the descriptive information (i.e., the payoff matrix) that there is a relatively low risk in choosing to Swerve compared to Dare and thus preferred this option more.

The alternating cooperation index ( $k$ ) provided us with important insights regarding the relations between fairness and efficiency within this type of interaction. Both fairness and efficiency were influenced by the availability of social information. While fairness increased with more information regarding mutual accountability, efficiency decreased. Moreover, increases in fairness between the information conditions were steeper and more drastic, compared with decreases in efficiency. This finding suggests that under certain conditions, increases in fairness
might have some cost, and lead to a minor decrease in the interaction's efficiency. This finding in repeated interaction is consistent with the tradeoff between efficiency and fairness in one-shot Ultimatum Game (Forsythe et al., 1994; Güth et al., 2003). However, it seems that for repeated interactions, fairness is more sensitive to the changes in the availability of social information compared to efficiency. The best compromise between fairness and efficiency was achieved when mutual interdependencies were learned only through repeated experiences. In this condition, fairness increased significantly while any decrease in efficiency was insignificant. This suggests that the availability of complete information, as in the Descriptive condition, do elicit fairness, but not necessarily the most efficient kind of fairness.

Examining the relations between fairness and efficiency at the pair level provided supporting evidence for these ideas. Where there is no or limited information regarding mutual accountability (i.e., the Individual and Minimal conditions), we find a decrease in fairness as the efficiency of a pair increases. This stands in contrast to the Experiential and Descriptive conditions, where we find no correlation between fairness and efficiency at the pair level. This demonstrates how overall the availability of information serves as a guard for fairness, and the efficiency of the interaction within the boundaries of fair behavior, depends on the specific interaction between individuals.

A key implication of this study is the importance of information to the fair resolution of conflicts. We show that social fairness is sensitive to the availability of information regarding mutual interdependencies between the members of the community. It is illuminating to see that the mere knowledge of interdependence with another person given in the Minimal condition is insufficient to promote alternating cooperation compared to the Individual condition. Further availability of social information increases cooperation and fairness. Thus, a preliminary requirement of fair conflict resolution should involve sharing information that sheds light on the interdependencies between the different entities involved in the conflict.

Even though it seems that overall fair behaviors in the Descriptive and Experiential conditions were somehow similar and less sensitive to the different way in which social information was conveyed, studies of individual decision making distinguish between these two sources of information (descriptive and experiential), and demonstrate how decisions from experience and description differ (e.g., Hertwig \& Erev, 2009). Recently, a similar distinction has been made in social dilemmas involving more than one decision maker (Martin et al., 2012). We believe that descriptive and experiential information influences the way fair behavior evolves over time. Thus, further analysis should carefully examine and compare the dynamics of fair interaction over time for different levels of social information.

## Acknowledgments

This research was supported by a Defense Threat Reduction Agency (DTRA) grant to Cleotilde Gonzalez and Christian Lebiere (HDTRA1-09-1-0053); and by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on. We are also thankful for the help and insightful comments of Hau-yu Wong.

## References

Bolton, G., \& Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. American Economic Review, 90(1), 166-193.
Camerer, C.F. (2003). Behavioral game theory: Experiments in strategic interaction. Princeton, NJ: Princeton University Press.
Dufwenberg, M., \& Kirchsteiger, G. (2004). A theory of sequential reciprocity. Games and Economic Behavior, 47(2), 268-298.
Fehr, E., \& Schmidt, K. (1999). A theory of fairness, competition, and cooperation. The Quarterly Journal of Economics, 114(3), 817-868.
Forsythe, R., Horowitz, J. L., Savin, N. E., \& Sefton, M. (1994). Fairness in simple bargaining experiments. Games and Economic Behavior, 6(3), 347-369.
Gonzalez, C., \& Martin, J. M. (2011). Scaling up instancebased learning theory to account for social interactions. Negotiation and Conflict Management Research, 4(2), 110-128.
Güth, W., Kliemt, H., \& Ockenfels, A. (2003). Fairness versus efficiency: An experimental study of (mutual) gift giving. Journal of Economic Behavior \& Organization, 50(4), 465-475.
Heider, F., (1958). The psychology of interpersonal relations. New York City: Wiley.
Hertwig, R., \& Erev, I. (2009). The description-experience gap in risky choice. Trends in Cognitive Sciences, 13(12), 517-523.
Martin, J. M., Gonzalez, C., Juvina I., \& Lebiere, C. (2012). A description-experience gap in social interactions: Information about interdependence and its effects on cooperation. Manuscript under review.
Rabin, M. (1993). Incorporating fairness into game theory and economics. The American Economic Review, 83(5), 1281-1302.
Rapoport, A., \& Chammah, A. M. (1965). Prisoner's dilemma: A study in conflict and cooperation. Ann Arbor, MI: University of Michigan Press.
Rapoport, A., Guyer, M. J., \& Gordon, D. G. (1976). The 2 $x 2$ game. Ann Arbor, MI: University of Michigan Press.
Roth, A. E., (1995). Bargaining experiments. In J. Kagel \& A. E. Roth (Eds.), Handbook of experimental economics. Princeton, NJ: Princeton University Press.
Russell, B., (1959). Common sense and nuclear warfare. New York City: Simon \& Schuster.

# Beyond Rule versus Rote? Processing of Distinctive Dative and Genitive Case Markers in German 

Christian Bentz (cb696@cam.ac.uk)<br>Department of Theoretical and Applied Linguistics, Sidgwick Avenue<br>Cambridge, CB3 9DA UK


#### Abstract

The rule versus rote distinction is one of the most debated issues in recent psycholinguistics. Dual route accounts hold that words can either be stored whole in the mental lexicon or computationally derived by simple combinatorial rules such as stem+affix. Within this framework, response latencies in lexical decision tasks have been applied to point out the difference between rote memorization, on the one hand, and combinatorial rule manipulation, on the other. However, this paper argues that there may be alternatives to this distinction. It will be shown that German nouns, which can be distinctively marked for number, case or both number and case, do elicit differing reaction times. Crucially, this effect can neither be explained by surface frequency effects nor by internal morphological structure. Rather, it seems to be triggered by the degree of embedding into usage-based units.


Keywords: Rule versus rote; lexical decision; German case marking; usage-based units.

## Introduction

The rule versus rote distinction in psycholinguistic theories of lexical access has been fiercely debated (see Pinker \& Ullman, 2002 as well as McClelland \& Patterson, 2002 for a review). Lexical decision tasks (LDT), priming studies, event related potentials and fMRI studies (see Clahsen, 1999 for a review) have been applied to answer the question whether lexical processing of morphologically simplex and complex items is rule-governed or associative, or both. It has been argued that lexical decision latencies can help us to distinguish processes involving abstract rule manipulation from mere memorization effects (Pinker \& Ullman, 2002; Taft, 2004; Clahsen, 1999, Clahsen, Eisenbeiss \& Sonnenstuhl-Henning, 1997; Marslen-Wilson \& Tyler, 2007, Sonnenstuhl \& Huth, 2002). In this context, the absence of frequency effects for regularly derived forms has been explained by abstract rule manipulation, whereas the occurrence of frequency effects was associated with rote memorization of irregular forms (see for example Clahsen 1999: 998, but also Hahn \& Nakisa 2000 for critical remarks). If these assumptions hold, then processing difficulties in lexical decision tasks must stem from:
a) The low frequencies of test items (in the case of memorization);
b) The difficulty of parsing by means of grammatical rules applied to derive the internal structure of a morphologically complex word (symbol manipulation).

However, the study presented here suggests that the 'grammatical load' of inflections is another potential factor
relevant for processing difficulty, depending on word external rather than word internal factors. Along those lines, it will be argued that a usage-based account of lexical access can provide an alternative explanation of the processing difficulties reflected in lexical decision tasks.

To this end, a lexical decision experiments was designed which involved German words with $-(e) n$ and $-s$ plural marking, which can additionally encode dative and genitive case. It will be shown that forms with more grammatical load, i.e. forms encoding both case and plural meaning, elicited significantly longer response latencies than unmarked forms. Crucially, these prolonged latencies can neither be explained by token frequency effects nor by word-internal parsing, rather, the participants seemed to have invoked redundantly case marked articles or prepositional phrases triggering case marking. This way they could decide whether the case marked word is a possible word form in German. This strategy prolongs reaction times (RTs) for morphologically complex forms.

Therefore, this paper will argue that the distinction between rule governed processes and memorization effects in LDT research lacks an important aspect of language processing: the embedding of items in phrases and sentences, i.e. usage-based units. In the following, the case marking and plural paradigms of nouns in German will be sketched in section 1. In section 2 the methods and results of the LDT will be presented and discussed in section 3.

## 1. German Dative and Genitive Inflections

German has four distinct case marking paradigms: nominative, accusative, genitive and dative (Engel, 1991: 505; Griesbach, 1986: 294; Kempe \& MacWhinney 1998: 549). However, since there is a fair amount of syncretism between case markers and singular/plural markers across different noun classes, the only markers that are distinctive inflectional case markers ${ }^{l}$ are the -(e)s genitive marker for a subset of singular masculine and neuter nouns as well as the -(e)n dative marker in the plural for all genders (Griesbach, 1986: 294; Engel, 1991: 505). Hence, distinctively case marked forms are restricted to these $-(e) s$ and $-n$ inflections for some nouns.

For example, the high frequent noun Haus (house) is inflected as Häus-er (houses) for all plural forms except for the dative, for which Häus-er-n (houses.DAT) is the

[^10]grammatically correct form. Likewise, the singular form Haus is the same for all cases except for the genitive: Hauses (house's).

Now, with regards to the design of a lexical decision task, two groups of target words were distinguished: Words ending in $-n$ and words ending in $-s$ (N-Group and SGroup). Furthermore, these two groups were then split up according to the 'grammatical load' of the suffixes, which renders three subgroups each (N1, N2-PL, N3-PL-DAT, S1, S2-SG-GEN, S3-PL-GEN) as depicted in table 1.

Table 1: Dative and genitive groups with grammatical load indicated by colors.

|  | Grammatical load | Example |
| :---: | :--- | :---: |
| Group N1 | -n part of stem (low) | Zah $\underline{n}$ <br> (tooth) |
| Group N2- <br> PL | -n denoting plural for all <br> cases (medium) | Rabe- $\underline{n}$ <br> (ravens) |
| Group N3- <br> PL-DAT | -n as distinctive dative <br> plural marker (high) | Stiefel- $\underline{n}$ <br> (boots.DAT) |
| Group S1 | -s part of stem (low) | Gleis <br> (platform/track) |
| Group S2- | -s as genitive singular <br> marker (medium) | Pferde- $\underline{s}$ <br> (horse's) |
| SG-GEN |  |  |
| Group S3- | -s as genitive singular and <br> PL-GEN <br> plural marker for all cases <br> (high) | Zoos, zoo's) |

*Umlaut was avoided, except for Ästen (branches)

As can be seen in Table 1, the groups are put together according to different functions of the final $-n$ and $-s$. They might not have any grammatical function (groups N1 and S1), they can have one specific function, namely denoting the plural (group N2-PL) or the genitive singular (S2-SGGEN), or they can represent two different grammatical functions - both plural and case marking - as in groups N3-PL-DAT and S3-PL-GEN.

In order to also control for potential frequency effects, the WEBCELEX ${ }^{2}$ database was used to select 20 target words for each of the 6 groups. These 120 target words were matched for surface frequency (ranging from 20-1 per $\sim 5$ million) and length in letters (ranging from 3-10 letters per word). Additionally, data on other frequency measures such as stem frequency, type frequency, family size and family frequency ${ }^{3}$ was also included.

[^11]
## 3. Lexical Decision Experiment

### 3.1 Methods

Participants. A lexical decision task was performed with 26 participants volunteering to participate in the study, all of them native speakers of German with a mean age of $\sim 27$ (14 females, 12 males).

Materials. The aforementioned 120 target words - split up into 6 groups (N1-S3) - were selected from the WEBCELEX database and matched for surface frequency and length in letters within groups. Additionally, 120 random filler words were selected from WEBCELEX, as well as 240 non-words of which 120 were produced by manually changing two or three letters of the stem (of other words in WEBCELEX), and 120 by changing potential affixes. This way, subjects were prevented from relying solely on recognition of stems for their lexical decision. All non-words adhered to the phonotactic rules of German. All filler words and non-words were chosen to reduce possible priming effects with regards to the target words. Overall the number of words and non-words added up to 480 items.

Items were presented by using the SuperLab 4.5.2 stimulus presentation software (Abboud, Heller, Matsak, Schultz \& Zeitlin, 2011). To present the stimuli, the item list was split up into three blocks with 160 items each, which all contained roughly the same ratio of target words, filler words and non-words. Items were presented as black Tahoma letters in font size 20 against a light turquoise background. They were preceded by a black fixation point in the center of the screen for 500 ms before stimulus onset. There was no time limit for responses. Participants responded to stimuli by using a Cedrus response pad (model RB-730) with green and red buttons for word and non-word decisions.

For statistical analyses and data plotting the software $R(\mathrm{R}$ Development Core Team, 2012) was used. Additionally, the software packages lme4 (Bates \& Maechler, 2010) and languageR (Baayen, 2010; cf. Baayen, 2008) as well as ggplot2 (Wickham \& Chang, 2012) were used to construct linear mixed-effects models and for plotting.

Procedure. In the instructions participants were told to decide as quickly and accurately as possible whether the
been shown to play a role in reaction time experiments (Nagy, Anderson, Schommer, Scott \& Stallman, 1989; Alegre \& Gordon, 1999). Moreover, the family size of a word is the stem frequency + the number of derived words (e.g. health/health-y) and the number of compounds (e.g. table/tablecloth) (Schreuder \& Baayen, 1997; Bertram, Baayen \& Schreuder, 2000). Finally, the family frequency of a word is the sum of frequencies of all the forms belonging to the same morphological family.

Besides this class of token frequencies, which are used to predict RTs for lexical entries and lemmas of words, there is the concept of type frequencies, too, which captures the number of different words inflected with a particular marker (e.g. the number of verbs which are inflected with regular -ed versus the number of irregular verbs) (Bybee, 2007; Marcus, Brinkmann, Clahsen, Wiese \& Pinker, 1995: 212).
presented items are German words or not. They were explicitly told that forms with plural and case inflections can be part of the stimulus set. Then they were presented with a test trial containing 8 words and 8 non-words. Both dative and genitive marked words were represented in this set of test items. In the test trial items remained on the screen until the participant had pressed the correct button. The instructor remained in the room during the test trail and participants were able to ask questions. After that the instructor left the room and participants were presented with the three blocks of 160 items each (with one minute pauses in between). The testing took 15-20 minutes.

After finishing the main experiment, participants were presented with a questionnaire to clarify 1) whether they had guessed what the exact purpose of the experiment is; 2) whether they had issues with specific items; 3) whether they had used any specific strategy to decide on words with dative and genitive marking. Participants could use the keyboard to type their answers, but they were also told that they can just type "no" if they did not want to answer the questions.

### 3.2 Results

A pre-analysis of the data revealed that 4 participants had to be excluded from the dataset because they had guessed the purpose of the experiment. Also, three of the items ${ }^{4}$ were excluded because their per item error rate exceeded $50 \%$. The error rates per subject ranged from $1.6 \%$ to $21 \%$. No further subjects were excluded. This left 22 subjects and 117 items to be analyzed. Furthermore, RTs were cleaned by excluding all RTs of less than 300 ms for reasons of lower processing bounds (Baayen, 2008: 243). Also, all RTs longer than 3000 ms were excluded because both inspection of quantile-quantile plots (Baayen, 2008: 243) as well as considering 2-3 standard deviations from the overall mean (mean: 959 ms ; SD: 934 ms ) as a cut-off point suggested that 3000 ms are a realistic upper bound for RTs. Moreover, for the analysis of reaction times all incorrect responses were excluded from the sample. These cleaning procedures caused an additional data loss of $\sim 8 \%$.

In the following, the RTs for the N -Groups and S-Groups are analyzed separately. Plotting the subgroups and logarithmically transformed RTs for each group reveals that there are differences in mean reaction times (see figure 1a and $1 b$ ).

In order to check the significance of these results linear mixed-effects models (Baayen, 2008.; Baayen, Davidson \& Bates, 2008; Barr, Levy, Scheepers and Tily, 2013) with RTs (logarithmically transformed) as dependent variable and group as predictor variable (fixed effect) as well as subjects and items as crossed random effects were used. In accordance with Barr et al. (2013) random intercepts for subjects and items as well as random slopes for subjects were included. P-values are based on likelihood ratio tests

[^12]for comparisons of the original models with null models (no fixed effects).



Figure 1: Boxplots for $\log (\mathrm{RTs})$, inflectional categories (x-axes), and grammatical load indicated by color.

Model validation was performed by checking homoscedasticity and normality for plots of residuals versus fitted values.

These models reveal that subgroup membership for both dative (N1, N2-PL, N3-PL-DAT) and genitive (S1, S2-SGGEN, S3-PL-GEN) is a significant predictor of RT (dative: $\chi^{2}(2)=8.6, p=0.01$; genitive: $\left.\chi^{2}(2)=13.12, p=0.001\right)$, longer RTs being associated with subgroups of higher grammatical load.

Now, in order to contrast these results with the predictive power of frequency effects on RTs, two more mixed-effects models were designed. This time surface frequency, stem
frequency and family frequency ${ }^{5}$ were added as fixed effects besides group, again with random intercepts for subjects and items and random slopes for groups by subjects. The likelihood ratio tests for the full models versus the null models (without the frequency measures but with group as predictor) rendered a non-significant result for both the dative data $\left(\chi^{2}(3)=4.19, p=0.24\right)$ and the genitive data $\left(\chi^{2}(3)=2.72, \mathrm{p}=0.43\right)$. This suggests that adding different token frequencies as predictors does not render a better model. Note that these results are not affected by potential multicollinearity effects, since the variance inflation factor (VIF) was < 2 for all predictors in both models.

Finally, it should be noted that all the linear mixed-effects models presented in this section are more or less "stressed" for longer response latencies. This follows logically from the fact that RT distributions are somewhat skewed, exhibiting longer right tails. However, as will be argued in the following section, it is exactly the occurrence of nonnormally prolonged response latencies that is interesting for the overall interpretation of the data.

## 4. Discussion

The results reported for the lexical decision task suggest that there are systematic differences between nouns for which the $-n$ and $-s$ suffixes are grammatically meaningless (N1 and S1 subgroups in table 1) and nouns which are grammatically highly loaded (N3-PL-DAT, S3-PL-GEN). Moreover, subgroups which are inflected for plural or case only (subgroups S2-SG-GEN and N2-PL) lie somewhere in between the unmarked nouns and the heavily marked nouns in terms of reaction times. Interestingly, the observed patterns of reaction times per subgroup are not predicted by measures of token frequency. Token frequencies could not be shown to be significant predictors of RTs in post-hoc regression analyses.

However, it is important to be aware of the fact that type frequencies are tied with subgroups N1-S3 since they reflect the 'inflectional status' of a word, which is in turn the grouping factor for further divisions of the N -Group and S Group. For example, all the words in N1 have a type frequency of $15926 / 35315$ ( $45 \%$ of all the nouns in WEBCELEX), whereas all the words in N3 have a type frequency of $3140 / 35315$ (8.9\%). Likewise, all the nouns in N 1 share the inflectional status of being unmarked for case or plural and all the nouns in N3 share the inflectional status of being marked for plural and case. These were basically the search criteria for finding appropriate nouns in WEBCELEX. Hence, type frequency and subgroup membership are two sides of the same coin.

At this point the question arises what actually causes the longer response latencies. According to dual route accounts there are two possible explanations: a) Differences in token frequencies have an impact via the direct lexical access

[^13]route - this has been ruled out by controlling for surface frequency in the experiment and by including other measures of token frequencies in a post-hoc multiple regression model; or b) The differences in RTs stem from parsing difficulty for complex morphological structures within the words (see parsing example in figure 2 ).


Figure 2: Potential word internal structure for the morphologically complex noun Häusern (houses.DAT) with both plural and dative marking.

However, according to this rationale we would not expect the groups N2-PL and N3-PL-DAT as well as S2-SG-GEN and S3-PL-GEN to exhibit differing reaction times. This is because we chose words that do inflect for both number and case by simply adding either -n or -s (see table 1). Hence the word internal parsing difficulty and the decision latencies should be the same for all these groups. However, the RTs actually differ most between these groups.

This requires an alternative explanation: A third possibility is that the differences in RTs are due to the additional grammatical and conceptual load that these suffixes carry. This means, rather than analyzing structures within the word, it would be more interesting to analyze the context these words are typically embedded in. See, for example, a typical sentence involving the noun Häusern in German (figure 3).

This figure illustrates the grammatical relationships between the word internal and word external structure. The dative marking is triggered by a preposition hinter (behind) (i.e. lexical case). Moreover, the plural form needs to agree with the DAT.PL of the article die.SG, i.e. den.DAT.PL. Hence the word Häusern is embedded into a construction that involves a preposition and a case marked article. We could think of more such examples with other prepositions (e.g. auf (on top of), in (in), mit (with)).

"The wood behind the houses"
Figure 3: Grammatical relationships between elements of a sentence involving dative marking.

Crucially, note that the type frequency of this dative plural marker, i.e. the range of words it is applied to, hinges upon the productivity of such prepositional constructions (plus the productivity of dative forms in other contexts). This would suggest that increased processing difficulty in the LDT for grammatically loaded words stems from the strength of embedding into common or uncommon constructions.

Of course, there needs to be further research with and beyond LDTs to further elaborate this hypothesis. However, first hints suggesting that this explanation is along the right lines can be found in the questionnaire.

### 4.1 Questionnaire

When the first participant came across the German word Messers (knife's) in the trial set, he kept pressing the 'nonword' button several times, although this is a grammatically correct form and the item kept occurring on the screen. When the instructor noted that this is a genitive form of the word Messer, the participant said: "... auf Messers Schneide!" A German idiom directly translated as: "on knife's blade", meaning: "to be on a knife-edge".

Evaluating the post-test questionnaire revealed that this spontaneous associative reaction might not have been a single coincidence. When asked (question 2) whether they had particular problems with specific items, 10 ( $45 \%$ ) of the participants answered "no", 6 (27\%) of the participants had problems with either dative, genitive or plural forms, and the rest ( $28 \%$ ) named non-words and potential foreign words as problematic. Most intriguingly, when subjects were more specifically asked (question 3 ) whether they had problems with case marked words (by giving them some examples of the target set) 13 ( $52 \%$ ) answered with "no", 6 ( $24 \%$ ) had imagined the correct articles to take a decision, and $5(20 \%)$ had even used phrases like "die Spitze des Doms" (the cathedral's spire) or prepositional phrases "wegen des Kochs" (because of the cook) to take their decision.

To test whether the strategies named here might have prolonged reaction times, participants were post-hoc divided into two groups: one group (no-context group) for subjects that had negatively answered questions 2 and 3 (or who had named other difficulties like non-words and foreign words), and another group for subjects that had answered affirmative and noted that they used context related strategies to take lexical decisions (context group). Interestingly, for these two groups the mean RTs for S3 and N3 taken together differ: For the context group the mean RTs for words in S3 and N3 is higher ( 956 ms ) than for the non-context group ( 939 ms ), although this difference is not significant ( $\mathrm{p}=0.33$ ).

However, the fact that 12 (55\%) of the participants either had problems with dative and genitive markers or used "minimal phrases" as disambiguation strategy suggests that this is at least partly the reason for prolonged response latencies. Note that the rest of the participants (10, 45\%) did not necessarily use some other strategy or no strategy at all.

Rather, participants could just type "no" if they did not want to bother with the questionnaire in the first place. Overall, the insights from the questionnaire suggest that there are systematic reasons for prolonged response latencies, namely, whether forms are more or less embedded into usage-based units.

## 5. Conclusion

In the past, lexical decision tasks have been invoked to find out whether certain lexical items are processed as a whole or decomposed into stem+affix. In this context, it has been argued that for units stored whole in the lexicon there should be surface frequency or other token-related frequency effects observable, whereas for morphologically complex and regular items symbolic rules will be applied. These are not sensitive to frequency effects (Marcus et al., 1995; Clahsen, 1999; Pinker \& Ullman, 2002). However, the results reported in this paper suggest an alternative to this binary distinction.

First of all, it has been shown that token frequencies are not a significant predictor when it comes to morphologically complex nouns in German, whereas grammatical load and type frequencies do still correctly predict longer reaction times for these forms. Thus, instead of trying to explain response latencies by analyzing morphological structures within the lexical items, this study suggests that the relevant factor is the embedding of these items in more or less frequent phrases. This is in line with accounts arguing that statistical learning and frequency effects are not only relevant for "lexical entries" but also for whole constructions (Ellis \& O'Donnell, 2012).

In conclusion, there are measurable processing differences between grammatically marked and unmarked nouns in German. Hence, it is correct, on principal, to distinguish between words that are perceived as "basic" or "default" and words which are perceived as grammatically complex. However, this does not necessarily entail that such morphologically complex forms are composed out of simpler units by means of symbolic rule manipulation. Rather, such forms carrying more 'grammatical load' are more likely to be associated with whole phrases and sentences even in isolation. And this embedding in redundant and disambiguating structures is what makes them belong to the grammatical rather than the lexical domain in the first place.

## References

Abboud, H., Heller, K., Matsak, E., Schultz, H., \& Zeitlin, V. (2011). SuperLab. Stimulus presentation software. San Pedro, California: Cedrus Corporation.
Alegre, M., \& Gordon, P. (1999). Frequency effects and the representational status of regular inflections. Journal of Memory and Language, 40, 41-61.
Baayen, H. R. (2010). languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics". Retrieved from http://CRAN.R-project.org/ package $=$ languageR.

Baayen, H. R. (2008). Analyzing linguistic data: A practical introduction using $R$. Cambridge: Cambridge University Press.
Baayen, H. R., Davidson, D. J. \& Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language, 59, 390-412.
Barr, D. J., Levy, R., Scheepers, C., \& Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language, 68, 255-278.
Bates, D. \& Maechler, M. (2010). Ime4: Linear mixedeffects models using S4 classes. Retrieved from http://cran.r-project.org/package=lme4.
Bertram, R., Baayen, R. H., \& Schreuder, R. (2000). Effects of family size for complex words. Journal of Memory and Language, 42, 390-405.
Bybee, J. (2007). Frequency of use and the organization of language. Oxford: Oxford University Press.
Clahsen, H. (1999). Lexical entries and rules of language: a multidisciplinary study of German inflection. Behavioral and Brain Sciencesrain sciences, 22(6), 991-1060. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/ 11301574
Clahsen, H., Eisenbeiss, S., \& Sonnenstuhl-Henning, I. (1997). Morphological structure and the processing of inflected words. Theoretical Linguistics, 23(3), 201-250.
Ellis, N. C., \& O'Donnell, M. B. (2012). Statistical construction learning. In J. N. Williams \& P. Rebuschat (Eds.), Statistical learning and language acquisition. Berlin: Mouton de Gruyter.
Engel, U. (1991). Deutsche Grammatik. Heidelberg: Julius Groos.
Griesbach, H. (1986). Neue deutsche Grammatik. Berlin: Langenscheidt.
Hahn, U., \& Nakisa, R. C. (2000). German inflection: single route or dual route? Cognitive Psychology, 41(4), 313-60. doi:10.1006/cogp. 2000.0737
Kempe, V., \& MacWhinney, B. (1998). The acquisition of case marking by adult learners of Russian and German. Studies in Second Language Acquisition, 20(4), 543-587. doi:10.1017/S0272263198004045
Marcus, G. F., Brinkmann, U., Clahsen, H., Wiese, R., \& Pinker, S. (1995). German inflection: The exception that proves the rule. Cognitive Psychology, 29, 189-256.
Marslen-Wilson, W. D., \& Tyler, L. K. (2007). Morphology, language and the brain: The decompositional substrate for language comprehension. Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 362, 823-836. doi:10.1098/rstb.2007.2091
McClelland, J. L., \& Patterson, K. (2002). Rules or connections in past-tense inflections: What does the evidence rule out. Trends in Cognitive Sciences, 6(11), 456-463.
Nagy, W., Anderson, R. C., Schommer, M., Scott, J. A., \& Stallman, C. (1989). Morphological families in the
internal lexicon. Reading Research Quarterly, 24(3), 262-282.
Pinker, S., \& Ullman, M. T. (2002). The past and the future of the past tense. TRENDS in Cognitive Science, 6(11), 456-463.
R Development Core Team. 2012. R: A language for statistical computing. R foundation for Statistical Computing, Vienna, Austria. Downloadable at http://www.r-project.org/.
Sonnenstuhl, I., \& Huth, A. (2002). Processing and representation of German -n plurals: A dual mechanism approach. Brain and Language, 81, 276-290.
Taft, M. (2004). Morphological decomposition and the reverse base frequency effect. The Quarterly Journal of Experimental Psychology, 57A(4), 745-765.
Wickham, H., \& Chang, W. (2012). ggplot2: An implementation of the grammar of graphics. Retrieved from http://cran.r-project.org/web/packages/ggplot2/index .html

# Social and Environmental Contributors to Infant Word Learning 

Elika Bergelson (elikab@psych.upenn.edu)<br>Department of Psychology, 3401 Walnut St, 400a<br>Philadelphia, PA 19104 USA<br>Daniel Swingley (swingley@psych.upenn.edu)<br>Department of Psychology, 3401 Walnut St, 400a<br>Philadelphia, PA 19104 USA


#### Abstract

Infants demonstrate comprehension of early nouns (e.g. "hand") around six months, and comprehension of early non-nouns (e.g. "eat") around 10 months. In two experiments, we explore the reasons for this lag. Expt. 1 is a gaze-following study, the results of which suggest an improvement in point-following around ten months, and reveal correlations between pointing and both overall and non-noun vocabulary. Expt. 2 is a set of corpus analyses, the results of which suggest that word frequency does not explain the difference between noun and non-noun age of acquisition, while suggesting that the co-presence of words and their referents may play an important role. The results of these experiments contribute to our understanding of word-learning across word classes, and lend support to environmental and social factors as having an impact on the trajectory of word learning in the first year of life.


Keywords: language acquisition; word learning; cognitive development; infancy; psycholinguistics; corpus analysis

## Introduction

Infants learn words by taking in the environment around them and, over time, creating links between bits of language and bits of the world, with abstraction at both ends. Not all words are learned with equal ease, a phenomenon having the potential to help explain how word learning works. Diary studies and databases of parental checklists show that infants' early comprehension and production vocabularies, while quite broad, numerically favor nominals over action words, modifiers, and social expressions (Benedict, 1979; Dale \& Fenson, 1996).

These findings agree with comprehension studies in which infants who were asked to look at referents of nouns like "apple" and "hand" succeeded by 6 months of age (Bergelson \& Swingley, 2012a; Tincoff \& Jusczyk, 1999, 2012), but did not reliably show understanding of non-nouns like 'uh-oh' and 'eat' until around 10 months of age (Bergelson \& Swingley, 2013; See table 1).

The reasons for this developmental lag could rest within the child, within the nature of the linguistic elements, or within the environment. Here, we examine several broad hypotheses about the source of this lag, which we consider in light of new behavioral research (Expt.1) and corpus analyses (Expt. 2). These hypotheses are not new,
and have been the focus of a great deal of research, primarily among children older than the infants we consider here. Studies show that each of the factors we examine is very likely to be important at some point in development. What we begin to address here is the extent to which they might explain the developmental course we have found in infants' word understanding.

## Frequency

The Frequency hypothesis maintains that nouns are more frequent than non-nouns in infants' early experience, and that this leads to their being learned earlier, once a sufficient mass of exposure has occurred. Frequency may aid learning because learning is incremental over exposures, or because with more tokens the likelihood of a very useful exposure instance increases (Medina, Snedeker, Trueswell, \& Gleitman, 2011).

This hypothesis has several forms. The simplest is that infants just hear the non-nouns less often than the nouns. A more specific hypothesis is that infants hear nouns in isolation (in one-word utterances) more often than nonnouns, which might lead to earlier learning of nouns (Brent \& Siskind, 2001). We can evaluate whether frequency differences might explain infants' relatively late learning of non-nouns by measuring whether there are corpus frequency differences between the tested nouns and non-nouns.

## Environment

The Environment hypothesis we consider here maintains that nouns and non-nouns differ in the degree to which the contexts of their typical use support learning, where "support" refers to environmental conditions that prior research suggests are relevant to word learning. For example, non-nouns may appear in a broader number of situation-settings (e.g. playing, eating) than nouns, or may involve different amounts of attention-getting movement. Parents' use of nouns may go along with tactile support or clear signs of visual attention. Nouns may be used more often while the referent is present than is the case for nonnouns. All of these features are reasonable candidates as factors that support word learning (e.g., Kersten \& Smith, 2003; Meyer, Hard, Brand, McGarvey, \& Baldwin, 2011; Tomasello \& Kruger, 1992).

## Social Skill

The Social Skill hypothesis maintains that between six and ten months infants gain skills of social cognition that underlie the capacity for learning more abstract wordsskills that might not be criterial for learning nouns but which, by hypothesis, may be imperative for learning non-nouns. Existing research points to important changes in social-cognitive skills in the second half of the first year. For example, gaze-following improves substantially over this period, and may facilitate word learning (Morales, Mundy, \& Rojas, 1998). Evidence of increasing social skills around $9-10$ months indirectly supports the possibility that a social skill that was not necessary for learning nouns may be necessary for learning non-nouns. Prior experiments testing gaze-following showed that $10-$ and 11-month-olds but not 9-month-olds were more likely to follow the gazer's regard when his eyes were open than when they were closed, and that this skill correlated with language scores at 18 months (Brooks \& Meltzoff, 2005; Brooks \& Meltzoff, 2008).

## Conceptual Difficulty

The Conceptual hypothesis proposes that non-nouns are harder to learn because of the nature of the concepts and categories involved. Instances of a word like 'uh-oh' vary more and thus may be harder to recognize as having a common semantic core than instances of 'hand'. This hypothesis can be expressed as stemming from higherlevel differences in the kinds of linguistic roles played by nouns in contrast with adjectives, exclamations, verbs, and social greetings. It can also be thought of as a lowlevel difference in what 'features' must be summed over: in the noun case, visual features such as shape, size, and color, in the case of e.g., banana, which may be easily graspable from the environment, while non-nouns require more abstract (perhaps second-order) features that are harder to posit or grasp. A related hypothesis concerns biases in word-learning; it could be the case that in the absence of further evidence, infants choose to posit that a new content word they have isolated from the speech stream refers to a noun before they consider that it may refer to another part of speech.

## Overview of the Present Research

The hypotheses laid out above overlap; for example, conceptual difficulty might cause a need for social skills in learning non-nouns. Still, evidence can be brought to bear that favors or disfavors these hypotheses. Our two concrete research questions are:

1) Do we find evidence that gaze- and/or pointfollowing correlate with early word comprehension in laboratory tasks and/or vocabulary checklists?
2) Are there frequency-based or environmental differences between nouns and non-nouns when examining naturalistic interactions between infants and their caregivers?

In Expt. 1 we tested 6-14 month olds ( $\mathrm{n}=37$ ) in a gazefollowing task. Parents also completed vocabulary checklists (MCDI, Dale \& Fenson, 1996). Most of these subjects ( $\mathrm{n}=25$ ) also participated in a noun comprehension eyetracking study on the same day. The goal of this experiment was to look for specific mappings between social behaviors and word understanding. To date we have tested individual infants on nouns and on gaze following, with the intent to examine non-nouns and gaze following as well; this study speaks to the Social Skill Hypothesis.

In Expt. 2 we investigated how nouns and non-nouns appear in infants' environment, through analyses of audio and video corpora of infants interacting with their caregivers, with the goal of gaining a better understanding of whether the environment of these two types of words differs in relevant ways. If frequency and environmental factors affect noun and non-noun learning differentially, we expect to find differences in these measures across the word types in the corpus. This study speaks to the Frequency and Environment hypotheses.

In both studies we used a set of nouns and non-nouns tested in previous eyetracking studies (Bergelson \& Swingley, 2012a, 2013). See Table 1.

For both sets of items a corpus of 16 mothers speaking to their infants (Brent \& Siskind, 2001), and a database of vocabulary checklists (MCDI, Fenson et al., 1994) were used to select a pool of items that were used often by most mothers and reportedly understood by a large percentage of 12-18 m.o. infants. Items were then selected among candidates based on imageability, and phonological properties.

The nouns are all foods and body parts, as the authors had a secondary interest in these abstract categories as such. , Given that infants early vocabularies are indeed a smorgasbord of different parts of speech (Benedict, 1979), and that is it not always easy to determine that the word class an item belongs to for a young infant corresponds to its word class in adults' vocabulary, the non-nouns were simply the most common imageable words heard by infants that were not concrete objects

| Non- | all gone, bye, dance, drink, eat, hi, hug, kiss, <br> Nouns <br> more, sleeping, smile, splash, uh-oh,wet, |
| :--- | :--- |
| Nouns | apple, banana, bottle, cookie, ear, eye/s, face, <br> foot/feet, hair, hand/hands, leg/s, milk, mouth, <br> nose, spoon, yogurt |

Table 1: Previously Tested Nouns and Non-Nouns

## Experiment 1

## Methods

## Participants

The gaze-following study tested 37 infants ( $\mathrm{R}=6.0$ 14.9 mo., $\mathrm{M}=9.8 \mathrm{mo}$.), of which 25 infants also participated in a word-comprehension study just prior to the gaze-following study ( $\mathrm{R}=6.6-12.8 \mathrm{mo}$., $\mathrm{M}=9.2 \mathrm{mo}$.).

All participants were healthy, typically-developing monolingual English-exposed full-term infants with normal hearing and vision, recruited in the Philadelphia area. 8 additional infants were excluded from the gazefollowing study (technical error, $\mathrm{n}=5$; fussiness, $\mathrm{n}=1$; parental interference $n=1$, premature birth status $n=1$ ); 10 additional infants were excluded from the wordcomprehension study (technical error, $\mathrm{n}=6$; fussiness, $\mathrm{n}=2$; parental interference $\mathrm{n}=1$, premature birth status $\mathrm{n}=1$ ).

## Design

All parents first completed consent, background, and vocabulary checklist forms. Infants sat on their parent's lap and watched short video clips on a computer outfitted with a remote eyetracker (Eyelink, SR Research). Parents wore a visor to block their view of the screen, but not of their child. Infants wore a small sticker on their forehead to facilitate tracking. After calibration, infants saw a series of 16 test trials. In each test video, infants saw an actress with two toys, one on either side of her. At the beginning of each clip, the actress looked down at the table (2s), looked up at the camera and smiled (3s), and then turned her head to the left or right and gazed (gaze trials, $\mathrm{n}=8$ ) or gazed and pointed (point trials, $\mathrm{n}=8$ ) to one of two toys ${ }^{1}$. She kept her gaze there until the trial ended ( 5 s ; this was the time window of interest, hereafter "postturn window"). Each video was 10 s long, and side of look, point, and toy were counterbalanced within and across subjects.

For the infants in the word comprehension study, the design was identical to that in Bergelson \& Swingley, (2012a), except that the experimenter spoke the words in lieu of the parent (as in Bergelson \& Swingley, 2012b). Briefly, infants were presented with images of foods and body-parts, one of which was named by the experimenter, while their eyegaze was monitored.

## Data Analysis

We quantified infants' performance in the gaze-study as proportion target looking. For each subject, we computed a difference score consisting of the proportion of infant gaze to the target, minus the proportion to the distracter, averaged over trials. This measure ignores gaze at the actor's face or hand and compares correct and incorrect looks. Indiscriminate looking at target and distracter would yield, on average, a score of zero.

Given our interest in infants' increased comprehension of non-nouns at ten months, we split subjects into two groups around this age.
We also correlated infants' performance in the gazefollowing study with two different vocabulary measures: MacArthur CDI scores and noun-comprehension subject

[^14]means. MCDI scores were calculated based on the number of words parents said their child understood or said on the MCDI; we looked at MCDI scores overall, and at the specific sets of nouns and non-nouns for which we found a developmental lag (Bergelson \& Swingley, 2012a; 2013, see Table 1). Noun-comprehension subject means were calculated from infants' performance in the noun-comprehension experiment that preceded the gazefollowing study. ${ }^{2}$


Figure 1: Expt. 1 Subject Means by Age in the Gaze- and Point-Following Task. Each dot represents each subject's proportion of target looking averaged over trials, for each condition. The symbol used for the dot represents infants' vocabulary size, binned into three groups (see legend). Infants performed above chance in the Gaze \& Point condition in the over-10 month age group (the right side of the right graph).

## Results

Across both age groups, infants failed to look more to the side the actress looked at on gaze trials ( $<10 \mathrm{mo}$.: $12 / 26$ infants with positive performance, $\mathrm{M}=.012$, $\mathrm{Mdn}=-$ $.00019, \mathrm{p}=1 ;>10 \mathrm{mo} .: 8 / 11$ infants, $\mathrm{M}=.031$, $\mathrm{Mdn}=.041$ $\mathrm{p}=.10) .^{3}$ For point trials, infants under 10 months performed at chance ( $14 / 26$ infants, $\mathrm{M}=.023$, $\mathrm{Mdn}=.018$, $\mathrm{p}=.35$ ), while infants over 10 months succeeded ( $9 / 11$ infants, $\mathrm{M}=.13$, $\mathrm{Mdn}=.15 \mathrm{p}=.008$ ).

We next examined the correlations between infants' vocabulary size, as reported by their parents, and their gaze-following behavior. Infants' MCDI scores correlated with proportion target looking on point trials, but not on gaze trials (point trials: (Kendall's) $\tau=.26, \mathrm{p}=.024$; gaze trials: $\tau=.09, \mathrm{p}=.43$ ). Looking at the MCDI subset containing nouns and non-nouns that were tested in previous eyetracking studies (Bergelson \& Swingley, 2012a, 2013), we found correlations between point trials and non-noun vocabulary $(\tau=.28, \mathrm{p}=.019)$ and a marginal

[^15]correlation between point trials and noun vocabulary ( $\tau=.21, \mathrm{p}=.080$ ). Noun and non-noun correlations with gaze trials were not significant (non-nouns: $\tau=.022$, $\mathrm{p}=.89$, nouns: $\tau=.11 ; \mathrm{p}=.37$ ). For the subset of subjects in the noun-comprehension study ( $\mathrm{n}=25$ ), performance in the gaze-following study and noun-comprehension study were not significantly correlated (gaze trials: $\tau=.11$, $\mathrm{p}=.44$, point trials: $\tau=-.05, \mathrm{p}=.73$ ). See figure 1 .

## Discussion

These results suggest that around ten months, when infants begin to show comprehension of non-nouns, they show an increase in their ability to follow pointing and gaze, but not gaze alone. This ability is correlated with non-noun and overall vocabulary, assessed by parental checklist.

While previous work has found links between infants' gaze-following and vocabulary size at 18 months (Brooks \& Meltzoff, 2005), here we find a correlation between point-following and current vocabulary size, and in particular, with knowledge of non-nouns tested in previous work.

The finding that infants' noun comprehension task behavior did not correlate with point-following is in keeping with the marginal correlation we found on the MCDI noun subset. The possibility remains that although noun comprehension is evidently not strongly correlated with point-following ability, non-noun comprehension may be, a hypothesis we cannot address directly yet.

Vocabulary size and age are correlated ( $\tau=.46$, $\mathrm{p}<.0001$ ). This makes it difficult to untangle their relationship to point-following ${ }^{4}$. However, the results suggest that point-following seems to be a categorical ability attained around ten months (See figure 1). Before ten months, performance is variable and centered around zero; after ten months virtually all children are above zero. This, coupled with the result that age and residualized vocabulary both predict pointing suggests that age and vocabulary are not redundant predictors.

Thus, in answer to our first question, the timing of point-following success and correlations between MCDI and point-following data provide indirect evidence for the Social Skills Hypothesis. Thus, point-following is a strong candidate social skill that might be useful in nonnoun learning but which is apparently not necessary for noun learning, inasmuch as our somewhat exaggerated pointing materials test point-following in natural contexts. An alternate possibility is that a third skill emerges around ten months, and that this skill mediates both pointfollowing and non-noun learning.

Thus, with evidence for the social hypothesis garnered in Expt. 1, we turn to a set of corpus analyses that will allow us to assess the roles of frequency and

[^16]environmental factors in the lag between noun and nonnoun learning.

## Experiment 2

## Methods

We examined mothers' use of the set of nouns and nonnouns that we have tested in eyetracking experiments in both the Brent Corpus (an audio corpus of 16 mothers interacting with their $9-15$ m.o. infants; Brent \& Siskind, 2001), and in 20 videos of the Providence Video Corpus (5 mothers interacting with their young children; we selected a subset in which children ranged from 11 to 18 mo; Demuth, Culbertson, \& Alter, 2006). In the Brent Corpus we compared frequency counts in isolation (i.e., in one-word utterances) and overall. In the Providence corpus we extracted 919 utterances in which both the caregiver and child were clearly visible, and in which one of our words of interest was said.

These utterances were coded for a number of features, including whether the referent of the word was present (e.g. is there an apple when 'apple' is said, is someone eating when 'eat' is said, did something fall accidentally when 'uhoh' was said, etc.), what the parent was looking at/touching, what the child was looking at/touching, the situation the word was used in, what (if anything) was moving, whether the word was said before, during, or after attention to the relevant referent transpired, and what was present in the room. In the case of body-parts, coders noted 'presence' only when the relevant part was involved in the interaction in any important sense: e.g., if the mother was looking at a child who had yogurt all over her mouth and said "look at your messy face!" this counted as 'presence' of the word 'face'; in contrast, if the mother was holding her crying child while singing "if you're happy and you know it clap your hands", this did not count as an instance in which 'hands' were considered 'present'.

## Results

A series of analyses was conducted to test whether the difference in eyetracking-task performance between nonnouns and nouns might be due to higher frequency of the nouns rather than something more fundamental about the words' meanings. Frequency was estimated using the Brent corpus. There was not a significant difference in the frequency of the nouns and non-nouns ${ }^{5}$. Descriptively, each noun occurred $45-562(\mathrm{M}=262$, $\mathrm{Mdn}=244)$ times within the corpus while each non-noun occurred 33-1292 ( $\mathrm{M}=453$, $\mathrm{Mdn}=219$ ) times. Across each set of words, the total number of usages did not vary significantly (244 vs. $219, \mathrm{p}=.98$ by Wilcoxon test). Given that previous research supports a link between word learning and

[^17]frequency of isolated word tokens (Brent \& Siskind, 2001), we also examined this variable here. The sets of words were not differentially likely to occur in isolation either: nouns occurred 2-92 ( $M=26$ ) times and non-nouns occurred 0-1091 times $(\mathrm{M}=152)$; this difference was not significant (noun $\mathrm{Mdn}=19$, non-noun $\mathrm{Mdn}=11 ; \mathrm{p}=.95$ by Wilcoxon test.).

Analyses of the Providence Corpus subset revealed that there too, our nouns and non-nouns occurred with similar frequency: each non-noun occurred 1-94 times ( $M=37$, Mdn=23), there were 523 non-nouns total. Nouns occurred 5-46 times ( $\mathrm{M}=21$, $\mathrm{Mdn}=19$ ), with 396 nouns tokens total (estimated difference per word type: 7 words; $\mathrm{p}=.29$ by Wilcoxon test over words). Similarly, nouns and non-nouns as a group did not differ in number of isolated occurrences ( 72 isolated non-noun tokens total, $R=1-7$ over words; 35 isolated noun tokens total, $R=1-3$ over words; estimated difference 1.8 words; $\mathrm{p}=.13$ by Wilcoxon test over words).

Hand-coding of interactional features during parental use of the tested words revealed a large word-type (noun versus non-noun) difference in whether the referent of the word was present as part of the interaction. Non-nouns were said much more often than nouns when their referent was not present-e.g., saying "hi!" when no-one was newly on the scene, or "kiss" when there were no evident attempts at kissing. By contrast, nouns ("a banana!") were more often spoken in the presence of the referent (an actual banana, or a picture of one). For non-nouns the referent was not present $39 \%$ of the time; for nouns, $15 \%$. This pattern held for $5 / 5$ children in the corpus, and was significant over words (estimated difference $=.24$, $\mathrm{p}<.012$ by Wilcoxon test over words). See figure 2.

No significant difference between word-types was found in what mothers or children were touching or looking at, the number of situation-types that the word occurred in (e.g. playing, eating, interacting, bookreading), what in the scene was moving (e.g. child or mom, their hands, other objects, etc), whether the word was said before, during, or after attention to the relevant referent transpired, nor what was present in the room (all ps $>.05$ by Wilcoxon tests, and not significant predictors in logistic regressions of word-type). In short, on most coded variables, nouns and non-nouns did not differ in various features of the learning environment-except whether the referent was present or not. ${ }^{6}$

## Discussion

Expt. 2 showed that nouns are used in speech to infants when their referents are present; non-nouns are used when their intended referents are present about $60 \%$ of the time. However, nouns and non-nouns appear to be said at equivalent rates both in sentence context and as one-word utterances.

[^18]These findings fail to support the Frequency Hypothesis, and lend support to the Environmental Hypothesis, insofar as we found non-nouns were said more often in less ideal learning environments (i.e. when the referent was not present). While it remains possible that other environmental factors varied across the words as well, of those we coded this was the only one that differed significantly across word groups.

One limitation of this study is that the videos were of infants older than ten months, leaving open the possibility that at younger ages there are other word-type differences. Such an account would stipulate that parents interact differently with children once children know some nonnouns.

Thus, in response to our second question, when examining naturalistic interactions between infants and their caregivers, we did not find support for frequencybased differences between nouns and non-nouns, but did find support for environmental differences.


Figure 2: Referent Presence as a function of Word-type in the Providence Corpus Subset. This figure depicts the counts of instances of nouns and non-nouns of interest (see table 1). Color distinguishes whether the referent was present as an image (top), present (middle) or not present

## General Discussion

In two studies we have explored the underpinnings of early word learning, seeking to explain why infants are able to understand different types of words at different ages. More specifically, through a gaze-following experiment and corpus analyses, we have sought to explain infants' ability to understand nouns around six months and non-nouns around ten months.

We found some support for the Social Hypothesis, in that point-following emerged around ten months and was correlated with overall vocabulary and, more strongly, non-noun vocabulary. We found support for the Environmental Hypothesis in that nouns were used in the presence of their referent with much greater regularity than was the case for non-nouns. We did not find support for the Frequency Hypothesis, or for other versions of the

Environmental hypothesis related to situation, attention, or motion-related factors.

Both of these experiments leave open the Conceptual Hypothesis, whereby non-nouns are learned later because they are more complicated, either due to the inherent linguistic word-class that the words belong to, or due to the nature of the visual features that need to be tracked to learn non-nouns as opposed to nouns. One way to examine this hypothesis is to teach infants novel nouns and non-nouns in very similar linguistic and environmental conditions to examine whether with equivalent exposure some types of words may be, simply put, harder to learn. Such work is ongoing in our lab.

The hypotheses we tested here are not new to this research, and so our evidence supporting these notions (in some cases) corroborates prior work. Perhaps most surprising, then, are the null effects: little evidence for the importance of frequency (among these already frequent words), and little evidence for a broad range of environmental variables that would, a priori, seem reasonable predictors of infants' success in learning. Of course, it is always possible that more sensitive measures would reveal influences that did not emerge here; however, a strength of the present approach is that we are testing learning that takes place in infants' ordinary, daily-life experience.

It is unlikely that any one cause is responsible for infants' later learning of non-nouns than nouns. It is probable that skills underlying pointing and the environmental conditions in which words appear play a role in how easily and at what age words are learned. But a simple frequency-based account does not seem viable for accounting for this difference. Further work is needed to examine the interactions of conceptual, environmental, and social factors in early word-learning of different word types.

Our findings about the developmental timeline of noun and non-noun acquisition suggest that it takes infants nearly as long to learn their first non-nouns as it took to learn their first nouns. Understanding the mechanisms underlying this lag, which we have begun to explore here, requires understanding the intertwined roles of social development, the structure of the world, and the structure of concepts as expressed in natural language. This line of research, in turn, has implications for word-learning, language acquisition, and cognitive development on a much broader scale.

## Acknowledgments

This work was funded by an NSF GRFP \& IGERT grant to EB, and NIH Grant R01-HD049681 to DS.

## Reference

Benedict, H. (1979). Early Lexical Development Comprehension and Production. Journal of Child Language, 6(2), 183-200.

Bergelson, E., \& Swingley, D. (2012a). At 6-9 months, human infants know the meanings of many common nouns. Proceedings of the National Academy of Sciences, 109(9), 3253-3258.
Bergelson, E., \& Swingley, D. (2012b). Talkerindependence in word comprehension before 12 months. Paper presented at the Boston University Conference on Language Development.
Bergelson, E., \& Swingley, D. (2013). The Acquisition of Abstract Words by Young Infants. Cognition, 127(3), 391-397.
Brent, M. R., \& Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. Cognition, 81(2), B33-B44.
Brooks, R., \& Meltzoff, A. N. (2005). The development of gaze following and its relation to language. Developmental Science, 8(6), 535-543.
Brooks, R., \& Meltzoff, A. N. (2008). Infant gaze following and pointing predict accelerated vocabulary growth through two years of age: a longitudinal, growth curve modeling study. Journal of Child Language, 35(01), 207-220.
Dale, P. S., \& Fenson, L. (1996). Lexical development norms for young children. Behavior Research Methods Instruments \& Computers, 28(1), 125-127.
Demuth, K., Culbertson, J., \& Alter, J. (2006). Wordminimality, epenthesis and coda licensing in the early acquisition of English. Language and Speech, 49(2), 137-173.
Kersten, A. W., \& Smith, L. B. (2003). Attention to novel objects during verb learning. Child Development, 73(1), 93-109.
Medina, T. N., Snedeker, J., Trueswell, J. C., \& Gleitman, L. R. (2011). How words can and cannot be learned by observation. Proceedings of the National Academy of Sciences, 108(22), 9014-9019.
Meyer, M., Hard, B., Brand, R. J., McGarvey, M., \& Baldwin, D. A. (2011). Acoustic packaging: Maternal speech and action synchrony. Autonomous Mental Development, IEEE Transactions on, 3(2), 154-162.
Morales, M., Mundy, P., \& Rojas, J. (1998). Following the direction of gaze and language development in 6-month-olds. Infant Behavior \& Development, 21(2), 373-377.
Tincoff, R., \& Jusczyk, P. W. (1999). Some beginnings of word comprehension in 6-month-olds. Psychological Science, 10(2), 172-175.
Tincoff, R., \& Jusczyk, P. W. (2012). Six-month-olds comprehend words that refer to parts of the body. Infancy, 17(4), 432-444.
Tomasello, M., \& Kruger, A. C. (1992). Joint attention on actions: Acquiring verbs in ostensive and non-ostensive contexts. Journal of Child Language, 19(2), 311-333.

# Grid or no grid: Distance distortion in recognizing spatial information from complex cartographic maps 

Anne-Kathrin Bestgen (Anne-Kathrin.Bestgen@rub.de)<br>Department of Psychology, Universitätstr. 150<br>44780 Bochum, Germany<br>Dennis Edler (Dennis.Edler@rub.de)<br>Department of Geography, Universitätstr. 150<br>44780 Bochum, Germany

Frank Dickmann (Frank.Dickmann@rub.de)
Department of Geography, Universitätstr. 150
44780 Bochum, Germany
Lars Kuchinke (Lars. Kuchinke @rub.de)
Department of Psychology, Universiẗttstr. 150
44780 Bochum, Germany


#### Abstract

Mental representations of environments are embodied in cognitive maps. Cognitive maps enclose spatial and distance distortions, which appear due to transcription errors based on processing of map information. Participants processed complex cartographical maps of varying amounts of visual details like topography, boundaries and grid to examine their effects on recall and orientation performance. The results indicate that the presentation of boundaries, topographies and a square grid significantly reduced distortion errors compared to a blank map, whereas a presentation of more than one visual element did not further reduce the distortions.


Keywords: Cognitive Maps; Complexity; Visual Boundaries; Square Grid; Distance Distortions; Spatial Cognition

## Introduction

A map, as a symbolic two-dimensional image of spatial relations, is a complex display of different kinds of stimuli, like object names, shapes, colors, spatial relationships and distances (Thorndyke \& Stasz, 1980). Orientation on a map, thus, must be seen as a complex skill, which involves encoding a broad set of verbal and spatial information as well as combining and constructing a mental representation of the visual stimulus. Furthermore, all information is given simultaneously, which forces the map reader to evaluate and process the information in procedures running parallel.

The mental representations of environments are embodied in cognitive maps (Tversky, 1993). In a cognitive map, elements are structured and can be mentally inspected (Tversky, 1992). However, cognitive maps are not miniaturized models of the reality, but rather a derivation of the reality (Barkowsky, 2002). Examining the mental representation of knowledge acquired from simple maps, it can be shown that these representations are systematically distorted affecting both the participants' recognition and orientation performance (McNamara, 1986). Such
distortions have been shown to follow the principles of perceptual grouping. For example, cities that are connected by lines are found to be recalled as lying closer together (Klippel, Knuf, Hommel \& Freska, 2004; McNamara, Ratcliff, \& McKoon, 1984). Moreover, omissions and additions that cannot be retrieved in the reality are present (Mark et al., 1999). To examine the representation of cognitive maps and distortions of real maps, previous research has focused on subsets of cartographical tools. Okabayashi and Glynn (1984), for instance, analyzed straight and curved boundaries on simple white maps and came to the conclusion that participants who studied the curved boundary maps made more distortion errors than those who studied the straight boundary maps and noboundary maps. Furthermore, Sadalla and Magel (1980) showed that landmarks also distort the space around them. The hierarchical organization of landmarks was examined by Hirtle and Jonides (1985), who found a clustering of landmarks on the basis of non-spatial attributes. Moreover, they were able to point out that the clusters have consequences for participants' performance in distance estimations. In addition, Hommel, Gehrke and Knuf (2000) and Hurts (2005) showed a clustering due to boundaries or rivers.

As introduced above, previous research mainly focused on the examination of simple maps, and it still remains open whether such distortions are also observed when using higher complex cartographic material. Moreover, based on cognitive principles, square grids are commonly used in cartography as an artificial tool to guide a map reader's attention and to reduce spatial distortions, but this hypotheses has not been tested empirically yet. Therefore, the maps presented in this study were created at different levels of visual complexity, varying in the amount of visual details displayed at three dimensions: (1) territorial
boundaries: no boundaries vs. boundaries, (2) topography: no topographic details vs. maps containing topographic information (e.g., mountains), (3) grid: no grid vs. overlaid square grid. The goal is to evaluate the orientation and memory performance when participants learn the location of unknown cities at the different types of cartographic maps.

## Methods

## Participants

Sixty-two participants ( 30 male, 32 female) aged between 19 to 34 years [ $M=23,1 ; S D=3,4$ ] participated voluntarily in the study. All participants gave written informed consent before inclusion in the study.

## Materials

Eight different maps ( $680 \mathrm{px} \times 510 \mathrm{px}$ ) were created as study material. The first map was a blank white map, the second map included a network of red boundaries and the third included a colored topography. The fourth map comprised both the topography and boundary lines. Maps 5-8 copied these maps but additionally were overlaid by an artificial square grid (170 px x 170 px ) (see Figure 1). The other design parameters of the grid, such as the color and width of the grid line, are based on the map grids standardly used in printed German topographic maps (scale $1: 25,000$ ). In addition, a pool of 28 newly created city names of eight-letter-length was established. Based on this pool, seven cities were pseudorandomly overlaid on each of these maps illustrated as red dots (so that no city was presented twice for any participant).

## Design

The study consisted of a three-factorial $2 * 2 * 2$ mixed design comprising the within-subjects factors boundaries and topography, and the between-subjects factor grid (see Figure 1). Participants were randomly divided into two groups. Both groups received the same maps with the same arrangement of cities, boundaries and topography, but one group with an overlaid square grid and the other without. All participants were told to study the seven cities on each map. For each city they had to remember the localization on the map as well as the name.

Each trial began with a study phase of two minutes, which was followed by a two minute interval with filler tasks. After the interval participants had to recall the seven cities on the same map, which now only included the background but not the red dotted cities. Participants were instructed to complete the task as accurate and as fast as possible by placing the cities on the map, using the cursor. After the placement of each city, participants had to type in the cityname. The trial was confined to three minutes. Furthermore, the presentation order of the different map types was randomized for each participant. The cartographic visualization of the maps was made with Adobe ${ }^{\circledR}$

Illustrator® CS5. The final maps (RGB color model) were exported as vector data sets and, in a further step, imported into Adobe ${ }^{\circledR}$ Flash ${ }^{\circledR}$ CS5. Here, based on the objectoriented language ActionScript 3.0, the maps were imbedded into an animated application used to execute the computer-based trial.


Figure 1: Schematic visualization of the four maps of the paradigm. Maps 2, 6, 4 and 8 include boundary lines. Maps $3,7,4$ and 8 include a colored topography. Maps in the second row are overlaid by an artificial square grid.

## Statistics

The orientation and memory performance was assessed by measuring the distance between the recalled place location compared to the original location of the city from the studyphase map. The Euclidian distance was measured in pixels. A city was rated correct in case the recalled location fell within a distance smaller than $28,4 \mathrm{px}(=1 \mathrm{~cm})$ from the study location (Okabayashi \& Glynn, 1984). A mixed threeway $2 * 2 * 2$ ANOVA comprising the within-subjects factors boundaries and topography and the between-subjects factor grid was computed for the number of correctly recalled cities and the mean distances of correctly positioned cities within the 1 cm radius at a given significance threshold of $\mathrm{p}=.05$.

## Results

The repeated measures ANOVA for mean distances of correct placements with grid versus no-grid as a between subject factor revealed a significant main effect of grid $\left[F(1,60)=6.359, p=.014, \eta_{p}{ }^{2}=.096\right]$, but no significant main effects of topography ( $p=.69$ ) and boundaries ( $p=$
.21). Moreover, a significant interaction between topography and grid was observed $[F(1,60)=6.464, p=$ $\left..014, \eta_{p}{ }^{2}=.097\right]$. Post-hoc $t$-tests showed that distances significantly decreased with the presentation of grids when no topography was present $[t(122)=3.631, p<$. 001].Distances for maps without topography and without grids are higher compared to those with topography and either without or with grids $[t(122)=2.069, p<.043 ; t(122)$ $=2.287, p<.024]$ whereas the presentation of the grid has


Figure 2: Mean distances between original and correctly recalled cities (a city was rated correct, when the recalled location fell within a distance smaller than $28,4 \mathrm{px}$ ) on maps with topography or no topography and overlaid with a square grid or without. $*=p<.05 ; * * *=p<.001$
no effect when a topography is present $(p=.770)$ (Figure 2).
The ANOVA on distances also revealed a significant interaction between boundaries and grid $[F(1,60)=5.821$, $\left.p=.019, \eta_{p}{ }^{2}=.088\right]$. The results of the post-hoc $t$-tests indicate that distances significantly decrease with the presentation of grids when no boundary is present $[t(122)=$ 3.676, $p<.001$ ] and distances are higher compared to maps with boundaries and either without or with grids $[t(63)=$ 2.471, $p<.016 ; t(122)=2.874, p<.005]$, whereas the presentation of the grid has no effect when boundaries are present ( $p=.560$ ) (Figure 3). The interactions between topography and boundary ( $p=.40$ ) and between grid, topography and boundary ( $p=.21$ ) were not significant (for means and standard deviations see Table 1).

Table 1: Distances in pixels between original and correctly recalled cities [mean $(S D)$ ] for the different map types.

|  | no topography |  | topography |  |
| :---: | :---: | :---: | :---: | :---: |
|  | no boundary | boundary | no boundary | boundary |
| no grid | $18.1(5.0)$ | $14.8(5.8)$ | $15.0(4.6)$ | $14.2(4.6)$ |
| grid | $12.6(4.6)$ | $13.4(5.3)$ | $14.2(4.7)$ | $14.6(4.5)$ |

The repeated measures ANOVA for mean percentages of correctly recalled cities with grid versus no-grid as a between subject factor revealed a significant main effect of topography $\left[F(1,60)=4.976, p=.029, \eta_{p}^{2}=.077\right]$. No other main effect reached significance (grid, $p=.12$; boundaries,


Figure 3: Mean distances between original and correctly recalled cities (a city was rated correct, when the recalled location fell within a distance smaller than $28,4 \mathrm{px}$ ) on maps including boundaries or without and overlaid with a square grid or without. $*=p<.05 ; * * *=p<.001$
$p=.06)$. Moreover, significant interactions between grid and topography $\left[F(1,60)=10.762, p=.002, \eta_{p}^{2}=.152\right]$, and between grid, topography and boundary $[F(1,60)=10.245$, $\left.p=.002, \eta_{p}{ }^{2}=.146\right]$ were observed. Although the presentation of a grid did not show any effect if a topography was present, the recall of city-locations was significantly better among maps without topography if a grid was present $\left[F(1,60)=8.717 p=.004, \eta_{p}{ }^{2}=.127\right]$ (Table 2). The interactions between grid and boundaries ( $p=.17$ ), and between topography and boundaries ( $p=.44$ ) were not significant (for means and standard deviations see Table 2).

Table 2: Percent of correctly recalled cities [mean $(S D)$ ] on the different map types.

|  | no topography |  | topography |  |
| :---: | :---: | :---: | :---: | :---: |
|  | no boundary | boundary | no boundary | boundary |
| no grid | $38.8(24.1)$ | $42.9(25.1)$ | $57.1(24.3)$ | $49.6(29.5)$ |
| grid | $63.8(23.6)$ | $44.3(25.3)$ | $51.9(22.0)$ | $51.4(21.7)$ |

## Discussion

The inclusion of visual boundaries, topographic information and of a square grid significantly modulate distance distortions and recall performance of complex cartographic maps, probably due to altered cognitive representations of the spatial relations. Following the analysis of the distance
distortions a clear picture emerges in that adding of visual elements significantly reduces distances relative to the original locations. Of particular note is that no further additive effects of these visual elements were observed. For example, no further advantage of boundaries or the displayed topography is visible in orientation performance when a square grid is already present - but also no disadvantage. Participants seem to set different additional visual information apart from an other to enhance their orientation performance and thus to reduce distortions (Figures 2 and 3), but the transfer to combination of visual information is missing.

In general these result support the notion of structured cognitive maps derived from environmental (visual) information - as the present paradigm rules out effects of previous knowledge (Barkowsky, 2002; Tversky, 1992). A likely way to use such complex visual information is to build visual hierarchies (Eastman, 1985; Hirtle \& Jonides, 1985) forming frames of representation that clusters the visual elements. Thus, cartographic details seem, in accordance with our initial hypothesis to guide the processes of cognitive map formation. It will be a question of future research to examine why no additive effects of the visual elements were visible, a first hypothesis would need to consider a kind of visual overkill, which if it can be replicated would affect the designing of cartographic material.

Regarding the recall performance of cities, topography had an overall effect, whereas grids again only improved the orientation on simple maps. This matches the results of the distance distortion analyses. Future studies should include different grid types to examine this result and to improve the use of grids, but also the presentation of topographic details and boundary information as assistant cartographic elements. Topographic details and boundaries should be modified in density and spacing to receive an impression of the degree in which the adding of visual elements effects distance distortion. Additionally, different grid types in combination with the assistant cartographic elements should be analyzed in terms of a coordination of cartographic elements to support the map reader. Furthermore, general investigations in navigation performances should be pursued based on these results.

## Acknowledgments

The study was funded by the German Research Foundation (DFG, DI 771/7-1) to F.D. and L.K.

## References

Barkowsky, T. (2002). Mental Representation and Processing of Geographic Knowledge- A Computational Approach. Berlin: Springer.
Eastman, J.R., (1985). Graphic Organization and Memory Structures for Map Learning. Cartographica, 22(1), 1-20.
Hirtle, S, Jonides, J. (1985): Evidence of hierarchies in cognitive maps. Memory and Cognition, 13, 208-217.
Hurts, K. (2005). Common region and spatial performance using map-like displays. Proceedings of the Human Factors and Ergonomic society, USA, 49, 1593-1597.
Hommel, B., Gehrke, J., \& Knuf, L. (2000). Hierarchical coding in the perception and memory of spatial layouts. Psychological research, 64(1), 1-10.
Klippel, A., Knuf, L., Hommel, B., \& Freska, C. (2004). Perceptually Induced Distortion in Cognitive Maps. Spatial Cognition IV: Reasoning, Action, Interaction, 204-213
Mark, D.M., Freksa, C., Hirtle, S.C., Lloyd, R., \& Tversky, B. (1999). Cognitive models of geographical space. International Journal of Geographical Information Science, 13(8), 747-774.
McNamara, T.P. (1986). Mental Representations of Spatial Relations. Cognitive Psychology, 18, 87-121.
McNamara, T.P., Ratcliff, R., \& McKoon, G. (1984). The mental representation of knowledge acquired from maps. Journal of Experimental Psychology: Learning, Memory, and Cognition, 10(4), 723-732.
Okabayashi, H., \& Glynn, S. M. (1984). Spatial Cognition: Systematic Distortions in Cognitive Maps. The Journal of General Psychology, 111, 271-279.
Sadalla, E., \& Magel, S. (1980). The Perception of Traversed Distance. Environment and Behavior, 12(1), 65-79.
Thorndyke, P. W., \& Stasz, C. (1980). Individual differences in procedures for knowledge acquisition from maps. Cognitive psychology, 12(1), 137-75.
Tversky, B. (1992). Distortions in Cognitive Maps. Geoforum, 23(2), 131-138.
Tversky, B. (1993). Cognitive Maps, Cognitive Collages and Spatial Mental Models. Lecture Notes in Computer Science, 716, 14-24.

# Evidence for cognitively controlled saccade targeting in reading 

Klinton Bicknell ${ }^{1}$ (kbicknell@ucsd.edu), Emily Higgins ${ }^{1}$ (ehiggins@ ucsd.edu), Roger Levy ${ }^{2}$ (rlevy@ucsd.edu), \& Keith Rayner ${ }^{1}$ (krayner@ucsd.edu)<br>${ }^{1}$ Department of Psychology, UC San Diego, La Jolla, CA, USA<br>${ }^{2}$ Department of Linguistics, UC San Diego, La Jolla, CA, USA


#### Abstract

It is generally assumed that the character position targeted within a particular word is not under direct cognitive control, but is rather determined by oculomotor processes sensitive only to word length and distance. An alternative view is that readers target more distant characters in words when they have parafoveally processed these words more. These possibilities are difficult to distinguish because the actual landing site within a word has large effects on subsequent word processing measures. In two experiments, we decoupled the targeted location from the actual landing site by shifting the text 3 characters during the saccade into a target word. Results show that subsequent word processing time given a particular landing site was lower/higher when the eyes would have landed further forward/backward in the word. This effect remains significant in some cases when controlling for saccade launch site. These data provide evidence against the oculomotor theory and support a cognitive account of saccade targeting.


Keywords: eye movements; reading; display change

## Introduction

Reading is a complex process that requires the combination of language processing with visual information to make decisions about when and where to move the eyes. These decisions are made very rapidly: saccades in reading take around 150 ms to program (Rayner, Slowiaczek, Clifton, \& Bertera, 1983), yet fixation durations in reading are around 200250 ms , leaving only $50-100 \mathrm{~ms}$ to decide when and where to send the eyes next. Given these temporal requirements, one central question of reading research is the extent to which these decisions are made by the cognitive system - and thus are sensitive to ongoing linguistic processing - or made by faster, low level oculomotor heuristics. Much of this debate has focused on how readers decide when to make a saccade, e.g., investigating the sensitivity of the distribution of fixation durations to the linguistic properties of a fixated word such as its frequency or predictability (Staub, White, Drieghe, Hollway, \& Rayner, 2010; Staub, 2011; Feng, 2009b). It is generally assumed, however, by researchers on both sides of this debate that it is via oculomotor heuristics that readers decide where within a word to target their eyes. ${ }^{1}$ In this paper, we provide evidence against this view, suggesting that characterlevel saccade targeting decisions are under cognitive control, and thus supporting a view in which even the fine details of eye movements are sensitive to ongoing linguistic processing.

## Character-level saccade targeting

It has been known since Rayner (1979) that the eyes' modal landing position in (medium and long) words is slightly left

[^19]of the center, and Rayner initially suggested that readers may intentionally send their eyes to this position because it is the most efficient location from which to process the word (cf. O'Regan, 1981). However, Rayner, Well, Pollatsek, and Bertera (1982) found evidence from a display change paradigm in which they controlled the amount of preview - visual information available about the next word - by replacing some letters with Xs that readers send their eyes further into the following word when they had received more preview. They suggested a cognitive account of character-level saccade targeting, in which readers target a position further into a word when they have already processed more of the word. For example, if readers are able to identify initial letters in a word, they no longer need visual information about those letters, and it is an efficient reading strategy to target the eyes at the latter, still-unidentified part of the word (Rayner, McConkie, \& Zola, 1980). However, it is possible that Rayner et al.'s (1982) results do not reflect normal reading behavior, and may instead reflect an experiment-specific strategy, e.g., making shorter saccades when the next word contains more Xs.

McConkie, Kerr, Reddix, and Zola (1988) investigated this issue with an analysis of the effect of preview on saccade targeting in a corpus of naturalistic reading. To assess the effect of preview, McConkie et al. investigated the effect on landing position of launch site, the distance of the previous fixation from the beginning of the word. Because the quality of visual information rapidly decreases away from the fovea, nearer launch sites would be expected to yield more preview, and - under Rayner and colleagues' cognitive account - landing positions further into the word. McConkie and colleagues' results confirmed this prediction, showing that the modal landing position was more rightward for nearer launch sites. However, McConkie et al. presented analyses suggesting that this result was not best explained by the cognitive account. Specifically, they presented evidence that the relationship between launch site and modal landing position was linear, and argued that an account that explains the shift in modal landing position in terms of parafoveal preview should predict a non-linear relationship. Because readers only obtain significant information about letter identities from 7 or so characters away (Underwood \& McConkie, 1985), they argued that a preview account would predict that the effect of launch site on landing site should asymptote by launch sites of 7 characters. McConkie and colleagues presented evidence that the shift in modal landing position was well modeled as a linear function of launch sites from 1 to 7 characters, with no evidence of becoming smaller near 7 characters. Neverthe-
less, because they did not analyze launch sites past 7 characters, this is not strong evidence against the cognitive account. Based on their evidence, however, McConkie et al. (1988) proposed an oculomotor account of character-level saccade targeting, in which the functional target of the eyes is always the center of the word, but in which systematic error biases saccade lengths toward 7 characters (and happens to do so linearly). They further suggested that this systematic error is related to range error found in other saccadic (Kapoula, 1985) and manual (Poulton, 1981) tasks, which biases saccades toward the mean saccade length. This oculomotor account of saccade targeting has since become the dominant theory, and is encoded in all major models of eye movement control in reading (e.g., Reichle, Pollatsek, Fisher, \& Rayner, 1998; Engbert, Nuthmann, Richter, \& Kliegl, 2005).

## Goals

In the present work, we tease apart the cognitive and oculomotor accounts by testing their predictions for word processing that occurs after landing on a new word. This is difficult to disentangle in natural reading, because there are large effects of the actual landing site on eye movement measures that indicate word processing time, such as gaze duration and refixation rates (e.g., O'Regan, 1981; McConkie, Kerr, Reddix, Zola, \& Jacobs, 1989; Rayner, Sereno, \& Raney, 1996). Here, we experimentally decouple intended landing site from actual landing site using a sentence shift paradigm, which allows us to investigate the relationship between target word processing and where a reader would have landed in the word. The cognitive account, in which readers direct their eyes to later character positions in upcoming words when they have performed more parafoveal processing of the beginnings of these words already, predicts that - when controlling for actual landing site - readers will require less time to finish processing a word when they had targeted a later character. The oculomotor account, which holds that where a reader lands in a word is purely a function of launch site, does not obviously make this prediction. However, because parafoveal preview should be larger when the eyes are closer to the word (i.e., for closer launch sites), the oculomotor account may also make this prediction, because it predicts that landing sites are correlated with launch site. Crucially, though, in the oculomotor account, all effects of original landing position (i.e., the position at which the eyes would have landed had we not shifted the sentence) must be mediated by launch site. The cognitive account by contrast, under the assumption that the amount of parafoveal processing performed is variable even for a constant launch site, predicts that word processing times will be smaller when the eyes would have landed further into the word, even when controlling for effects of launch site. (Note, however, that a large amount of the amount of parafoveal preview obtained is likely to be correlated with launch site even on this account). We test these predictions in the following two experiments.

## Experiment 1

We use a sentence-shifting paradigm (McConkie, Zola, \& Wolverton, 1980; O’Regan, 1981; Inhoff, Weger, \& Radach, 2005; Nuthmann, 2006; Feng, 2009a) to tease apart effects of intended landing site from actual landing site. In our first experiment, we shift the sentence to the right during the saccade into a target word, as described below. This paradigm allows us to align actual landing sites and compare two cases: (1) when the actual site was the intended landing site (when no shift occurred), and (2) when the intended landing site was, instead, further into the target word (when the sentence shifted to the right). The cognitive account predicts that the latter case - when more distant locations were targeted - is more likely to reflect instances in which readers had parafoveally processed the word to a greater extent. Thus, the account predicts that word processing times on the target word should be reduced compared with the control, no shift condition. The oculomotor account may make this same prediction, but only as mediated by effects of launch site, since under this view landing position is strictly a function of launch site, and closer launch sites may yield more parafoveal preview. We thus seek to answer two questions: (1) whether any information about upcoming word processing can be recovered from original landing site, as measured by whether there is an effect of shift on subsequent eye movement measures aligned by actual landing site, and (2) whether this effect is completely mediated by launch site.

## Method

Subjects All subjects were students at the University of California, San Diego who received course credit for participation. All were naive to the purpose of the experiment and reported that they were native speakers of English with normal or corrected-to-normal vision. Data from 40 subjects were included in our analyses. Five additional subjects participated in the experiment but were excluded from analysis for reasons discussed below.
Apparatus Eye movements were monitored with an SR Eyelink 2000 eye tracker (SR Research Ltd., Kanata, Ontario, Canada) sampling at 1000 Hz . The system was configured in 'tower mode' and equipped with a chin rest. While subjects read binocularly, only one eye (the right eye by default) was tracked. Sentences were displayed on an HP p1230 20 in. CRT monitor with refresh rate set to 150 Hz and resolution set to $1024 \times 768$ pixels. Viewing distance was approximately 60 cm . Approximately 2.4 characters were encompassed by $1^{\circ}$ of visual angle. We used custom software (EyeTrack, developed at the University of Massachusetts, Amherst) to present and update the display.
Materials One hundred and sixty experimental sentences were included in this study. Eye movement measures were obtained from a single, pre-selected target word (always a 7letter verb) within each sentence, which was immediately preceded by a 3- or 4-letter noun. Each sentence appeared alone

```
    Pre-shift: The seasoned fig! couples well with goat cheese.
    No shift: The seasoned fig couples well with goat cheese.
Right shift: The seasoned fig couples well with goat cheese.
    Left shift: The seasoned fig couples well with goat cheese.
```

Figure 1: Example sentence. The first line depicts an experimental sentence and a boundary (invisible to subjects but shown here after the pre-target word 'fig') that, when crossed, will trigger a display change. The target word is 'couples'. The second line shows the control condition in which the sentence remains in place after the boundary has been crossed. The final two lines depict rightward (Experiments 1 and 2) and leftward shifts (Experiment 2 only) respectively.
on a single line of the screen in Courier New 14 pt. font.
Procedure After giving informed consent and receiving experimental instructions, subjects placed their heads in the chin rest and performed a 3-point horizontal calibration. Subjects then read 6 practice sentences, all without display changes, before beginning the experiment. Subjects read each experimental sentence silently for comprehension. For each subject, the order of sentence presentation was randomly and independently selected. After one third of trials, a simple comprehension question was presented to encourage attentive reading. Breaks were offered approximately halfway through the experiment and were available at any other time upon request. We used the gaze-contingent boundary technique (Rayner, 1975) to update the display when subjects' eyes crossed an invisible boundary placed after the last letter of the pre-target word. When this boundary was crossed, on half of the trials the display was re-drawn so that the entire sentence was shifted 3 characters to the right (the Right Shift condition, see Figure 1). In the remaining half of trials, the sentence was simply re-drawn in its original location (No Shift). The assignment of items to shift conditions was counterbalanced.
Analysis Data were processed using a suite of custom software developed at the University of Massachusetts, Amherst and the University of California, San Diego. Fixations shorter than 80 ms that occurred within a single character width (11 pixels) of an adjacent fixation were combined, and those that did not were removed. Trials containing a fixation longer than 1000 ms or a blink on or immediately preceding or following the target word region were also excluded. ${ }^{2}$ Trials were also excluded if the display change completed more than 9 ms after the beginning of the following fixation. Subjects were excluded from analysis for excessive data-loss, defined as $25 \%$ or more of trials being excluded for blinks or $50 \%$ or more of trials being excluded for late display changes. Trials were also excluded if the eyes (1) would not have landed on the target had no shift taken place, or (2) would have landed on the target under natural circumstances but were 'thrown off' by the shift. This requirement meant that all data from the shift condition was limited to actual landing positions $1-4$. In our statistical analysis of the effect of shift, we thus com-

[^20]pared the two shift conditions at only these four positions. ${ }^{3}$ Finally, in order to increase the probability that all fixations were intended for the target word, and not mislocated fixations intended for the previous word, we also excluded cases in which the previous word was skipped. Note that it is possible that two classes of unintentional fixations of the target word remain in the data: (a) fixations intended to be refixations of the previous word and (b) fixations that were intended to skip over the target word and fixate a subsequent word. However, each of these possibilities is unlikely to represent a substantial portion of the dataset, as the refixation probability for words of length 3-4 is very low (about $13 \%$ for words of length 4 and even lower for words of length 3; McConkie et al., 1989) and the probability of skipping over a 7-letter word is only about 10\% (Drieghe, Brysbaert, Desmet, \& De Baecke, 2004). ${ }^{4}$

We analyzed two measures of word processing: (1) gaze duration, defined to be the summed duration of all fixations made on a region prior to leaving it and (2) refixation probability, defined as the probability of making more than one fixation on a region prior to leaving it. We analyzed the effect of shift on gaze duration with linear mixed-effects regression (Pinheiro \& Bates, 2000) and on refixation probability with logistic mixed-effects regression (Agresti, 2002). In addition to a fixed effect of shift, all models included random intercepts and random slopes for shift for both subjects and items. As a control variable, the actual (post-shift) landing site was included as an unordered categorical fixed effect, and random slopes for landing site were included for subjects and items. In cases of nonconvergence, we iteratively removed random slopes of landing site until the model converged. We do not report control variable effects. Outlier gaze durations were excluded by removing all gaze durations more than 2 standard deviations from a subject's mean, without respect to experimental condition.

We report two analyses to answer the two questions described above. The first analysis seeks to determine whether any information about upcoming word processing can be recovered from original landing position by testing for an ef-

[^21]

Figure 2: Effect of shift on gaze duration and refixation rates by actual (post-shift) landing site for Experiment 1. Error bars show the standard error of the mean, computed after aggregation by subjects. Note that effects mediated by launch site have not been parceled out of this figure.
fect of shift on word processing controlling for actual landing site. The second tests whether this effect is completely mediated by launch site (as predicted by the oculomotor account) by including launch site as an unordered, categorical control predictor. To assess significance for the linear gaze duration models, we report the $t$ statistic. For datasets of this size, this statistic will be approximately normally distributed (Baayen, Davidson, \& Bates, 2008), meaning that $|t|>1.96$ indicates a significant effect $(p<.05)$ and $1.64<|t|<1.96$ indicates a marginal one (. $10<p<.05$ ). For logistic refixation models, we report the $z$ statistic, which has the same interpretation, and also give effect sizes in logits, which is the difference in the log-odds of making a refixation between conditions (Agresti, 2002).

## Results

The effects of shift are plotted in Figure 2, aligned by actual, post-shift landing position. There is an effect of shift on gaze duration: gaze durations are estimated to be significantly faster $(-19 \mathrm{~ms}, t=-3.1)$ when the eyes would originally have landed further into the word (i.e., after a rightward shift). Refixations are estimated to be 0.2 logits less likely after a rightward shift, but this is not significant $(z=-1.1)$. In analyses including launch site as a control predictor, the effect of shift on gaze durations was reduced to an insignifi-
cant $10 \mathrm{~ms}(t=-1.4)$, and the effect on refixations remained similar $(-0.2$ logits, $z=-1.1)$.

## Discussion

The results of Experiment 1 confirmed that original landing position does provide some information about upcoming word processing, as cases in which the eyes would have landed further forward in the word result in 19 ms shorter gaze durations. While this result is predicted by the cognitive account, the oculomotor account can also predict it, but only to the extent that it is completely mediated by launch site. In analyses controlling for launch site, the results from this experiment were unclear, however, and there was only an insignificant trend for gaze durations to be 10 ms shorter when the eyes would have landed further forward in the word. Thus, the results of this experiment are consistent with both models.

## Experiment 2

One limitation of the design of Experiment 1 is that it is possible that the effects we saw on gaze duration and refixation rate were merely low-level responses to shifting the sentence rather than true effects of prior processing of the target word. To allay this concern, in Experiment 2, we tested both right and left shifts of the sentence. While a simple, low-level response to the detection of a shift may be expected to affect eye movement measures similarly for leftward and rightward shifts, the cognitive account of saccade targeting makes opposite predictions for these two conditions. By the same logic as described for Experiment 1, this account predicts that gaze duration and refixation rates should be reduced in the rightward shift condition relative to the no shift condition when aligning on actual landing site. This is because the saccades in the rightward shift condition were directed further into the word, which on this account is caused by readers having performed more parafoveal processing. Analogously, this account predicts that these measures should be increased in the leftward shift condition relative to the no shift condition when aligning on actual landing site, since the leftward shift saccades were directed further back in the word than those in the no shift condition. The oculomotor account once again makes the same predictions as the cognitive account, but again requires that these effects be solely mediated by launch site. Experiment 2 thus allows us to test two predictions. First, if the simple, low-level shift effect is correct, we should find similar patterns of data for leftward and rightward shifts. Second, if we instead find opposite patterns of data for leftward and rightward shifts (as outlined above), examining whether these effects are solely driven by launch site will allow us to distinguish between the oculomotor and cognitive accounts of saccade targeting.

## Methods

Experiment 2 was identical to Experiment 1 with two exceptions. First, while 40 subjects were again included in our analysis, 7 were excluded (for reasons explained above). Second, while sentences remained static, once again, in half of trials,


Figure 3: Effect of shift on gaze duration and refixation rates by actual (post-shift) landing site for Experiment 2. Error bars show the standard error of the mean, computed after aggregation by subjects. Note that effects mediated by launch site have not been parceled out of this figure.
they shifted 3 characters to the right in one quarter of trials and shifted 3 characters to the left in the remaining quarter. An example is given in Figure 1.
Analysis Analysis was similar to that in Experiment 1 except that we now separately analyze the effect of left and right shift, comparing each to the no shift condition. Because we again exclude cases in which the eyes would have skipped the target or were 'thrown off' the target by the shift, this means that the data from the right shift condition are at actual landing sites $1-4$ and the data from the left shift condition are at actual landing sites $4-7$. We thus analyzed data from only these landing site ranges in each analysis.

## Results

The effects of shift are plotted in Figure 3, aligned by actual, post-shift landing position. For right shifts, the effect on gaze durations and refixation rates is again estimated to be in the predicted direction ( $-7 \mathrm{~ms},-0.3$ logits), but neither effect is significant ( $t=-0.7 ; z=-1.5$ ). For left shifts, the effects are in the opposite direction, and are larger ( $20 \mathrm{~ms}, 0.6$ logits) and significant $(t=3.7 ; z=6.1)$.

As before, we also performed analyses in which launch site is a control predictor, to determine whether these effects are exclusively mediated by launch site. For right shifts, this
analysis revealed insignificant effects on gaze duration ( 3 ms , $t=0.3)$ and refixation rates $(-0.3$ logits, $z=-1.4)$. The effect of left shifts controlling for launch site was estimated to be slightly smaller than when not controlling for launch site ( $18 \mathrm{~ms}, 0.5$ logits), but still robust $(t=2.9 ; z=4.5$ ).

To gain more power to assess the possible effects of rightward shifts, we performed a further, post-hoc analysis on the pooled data from for landing positions 1-4 from Experiments 1 and 2. This analysis revealed a significant effect of rightward shifts for gaze duration $(-16 \mathrm{~ms}, t=-3.2)$ and a marginal trend for refixation rate ( -0.2 logits, $z=-1.8$ ). An analysis controlling for effects of launch site revealed an insignificant 6 ms trend on gaze durations $(t=-1.0)$ and a marginal effect on refixation rates $(-0.2$ logits, $z=-1.8)$.

## Discussion

This experiment revealed, first, that leftward and rightward shifts produced opposite patterns of results, contrary to the predictions of the simple, low-level shift detection account: gaze duration and refixation rate were lower and higher in the rightward and leftward shift conditions respectively as compared with the static control condition, although this was only significant for the leftward shift condition. Because both the cognitive and the oculomotor accounts predicted this pattern of data, we also analyzed the data when controlling for launch site, a factor that should, according to the oculomotor view, entirely account for these results. These analyses revealed results more consistent with the cognitive account than the oculomotor account. In the leftward shift condition, gaze duration and refixation rate were significantly elevated even when controlling for launch site. For rightward shifts, pooling data across the two experiments also provided suggestive evidence in favor of the cognitive account of saccade targeting, suggesting that the effect was not entirely driven by launch site.

## Conclusion

In summary, we described two alternative accounts of how readers decide where, precisely, to aim their eyes when planning a saccade to an upcoming word. According to the cognitive account, readers send their eyes further into a word after having parafoveally processed it more. According to the oculomotor account, readers always target the center of a word, but are subject to systematic error, which is a function of launch site. We presented evidence in favor of the cognitive account from two sentence-shift experiments. As predicted by the cognitive account, the word processing measures of gaze duration and refixation rate suggested that readers perform less subsequent processing of a word when they would have landed further into it, and more subsequent processing of a word when they would have landed further back, controlling for actual landing site. This was a significant effect for rightward shifts in Experiment 1 and for leftward shifts in Experiment 2. Crucially, we found evidence that this effect was not fully mediated by launch site, as required by the oculomotor account. When controlling for launch site, in Experiment 2 , the effect of leftward shifts was fully reliable, and when
pooling data across Experiments 1 and 2, the effect of rightward shifts was marginal for refixation rate. This evidence thus suggests that readers decide where to target their eyes within a word based on how much processing of the word they have accomplished, and not just based on the current position of their eyes. Such an account requires that the details of saccade targeting are sensitive to ongoing cognitive, linguistic processing.

There is, however, one way in which the oculomotor account may still be able to accommodate these findings. As was pointed out above (see Expt. 1, Analysis), some trials included in our analysis may represent unintentional fixations on the target word: failed attempts (a) to refixate the pretarget word and (b) to skip the target word. If these trials represent a substantial portion of our data, the oculomotor account could also predict our findings, since (a) failed refixations would tend to land at the beginning of the target word and represent cases in which the target word was not yet the focus of processing and (b) failed skips would tend to land at the end of the target word and represent cases in which the target word was already processed. Further analyses will be required to determine whether the likely rates of such possibilities would be sufficient to render this account of our data plausible.

The data are certainly consistent, however, with the view that character-level saccade targeting is under cognitive control. Specifically, these results are predicted by an account in which readers send their eyes further into a word when they have obtained more parafoveal preview of it. Since this effect is not mediated by launch site, this means that where a reader's eyes land in a word provides information about how much they processed the word on that particular trial, which is not only a function of the location of their eyes on the previous fixation. If this account is correct, it would support the notion that fine-grained eye movements decisions in reading are tightly linked to the details of ongoing linguistic processing, suggesting that readers do not merely rely on heuristic strategies to guide their eyes. More generally, our results support a view in which humans optimize the fine details of their behavior to maximize their efficiency in linguistic tasks such as reading (Bicknell \& Levy, 2010; Lewis, Shvartsman, \& Singh, in press) and in cognition more broadly.

## Acknowledgments

This research was supported by NIH grant T32-DC000041 from the Center for Research in Language at UC San Diego to K.B., NSF grant 0953870 to R.L., and NIH grant R01HD065829 to R.L. and K.R. We thank Araceli Cervantes, Georgina Chen, Tiffany Lai, Hannah Johansen, and Vinny Uy for assistance with data collection and stimuli development.

## References

Agresti, A. (2002). Categorical data analysis (Second ed.). New York, NY: John Wiley \& Sons.
Baayen, R., Davidson, D., \& Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language, 59, 390-412.
Bicknell, K., \& Levy, R. (2010). A rational model of eye movement control in reading. In J. Hajič, S. Carberry, S. Clark, \& J. Nivre
(Eds.), Proceedings of the 48th annual meeting of the Association for Computational Linguistics (ACL) (pp. 1168-1178). Uppsala, Sweden: Association for Computational Linguistics.
Drieghe, D., Brysbaert, M., Desmet, T., \& De Baecke, C. (2004). Word skipping in reading: On the interplay of linguistic and visual factors. European Journal of Cognitive Psychology, 16, 79-103.
Engbert, R., Nuthmann, A., Richter, E. M., \& Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. Psychological Review, 112, 777-813.
Feng, G. (2009a). Mixed responses: Why readers spend less time at unfavorable landing positions. Journal of Eye Movement Research, 3, 1-26.
Feng, G. (2009b). Time course and hazard function: A distributional analysis of fixation duration in reading. Journal of Eye Movement Research, 3, 1-23.
Inhoff, A. W., Weger, U. W., \& Radach, R. (2005). Sources of information for the programming of short- and long-range regressions during reading. In G. Underwood (Ed.), Cognitive processes in eye guidance (pp. 33-52). New York: Oxford University Press.
Kapoula, Z. (1985). Evidence for a range effect in the saccadic system. Vision Research, 25, 1155-1157.
Lewis, R. L., Shvartsman, M., \& Singh, S. (in press). The adaptive nature of eye-movements in linguistic tasks: How payoff and architecture shape speed-accuracy tradeoffs. Topics in Cognitive Science.
McConkie, G. W., Kerr, P. W., Reddix, M. D., \& Zola, D. (1988). Eye movement control during reading: I. the location of initial eye fixations on words. Vision Research, 28, 1107-1118.
McConkie, G. W., Kerr, P. W., Reddix, M. D., Zola, D., \& Jacobs, A. M. (1989). Eye movement control during reading: II. frequency of refixating a word. Perception \& Psychophysics, 46, 245-253.
McConkie, G. W., Zola, D., \& Wolverton, G. S. (1980). How precise is eye guidance in reading? Paper presented at the 1980 meeting of the American Educational Research Association, Boston, MA.
Nuthmann, A. (2006). The "where" and "when" of eye fixations in reading. Unpublished doctoral dissertation, University of Potsdam.
O'Regan, J. K. (1981). The "convenient viewing position" hypothesis. In D. F. Fisher, R. A. Monty, \& J. W. Senders (Eds.), Eye movements: cognition and visual perception (pp. 289-298). Hillsdale, New Jersey: Lawrence Erlbaum Associates.
Pinheiro, J. C., \& Bates, D. M. (2000). Mixed effects models in $S$ and S-Plus. New York: Springer Verlag.
Poulton, E. C. (1981). Human manual control. In Handbook of physiology (Vol. II, pp. 1337-1389). Bethseda, MD: American Physiology Society.
Rayner, K. (1975). The perceptual span and peripheral cues in reading. Cognitive Psychology, 7, 65-81.
Rayner, K. (1979). Eye guidance in reading: fixation locations within words. Perception, 8, 21-30.
Rayner, K., McConkie, G. W., \& Zola, D. (1980). Integrating information across eye movements. Cognitive Psychology, 12, 206226.

Rayner, K., Sereno, S. C., \& Raney, G. E. (1996). Eye movement control in reading: a comparison of two types of models. Journal of Experimental Psychology: Human Perception and Performance, 22, 1188-1200.
Rayner, K., Slowiaczek, M. L., Clifton, C., Jr., \& Bertera, J. H. (1983). Latency of sequential eye movements: Implications for reading. Journal of Experimental Psychology: Human Perception and Performance, 9, 912-922.
Rayner, K., Well, A. D., Pollatsek, A., \& Bertera, J. H. (1982). The availability of useful information to the right of fixation in reading. Perception \& Psychophysics, 31, 537-550.
Reichle, E. D., Pollatsek, A., Fisher, D. L., \& Rayner, K. (1998). Toward a model of eye movement control in reading. Psychological Review, 105, 125-157.
Staub, A. (2011). The effect of lexical predictability on distributions of eye fixation durations. Psychonomic Bulletin \& Review, 18, 371-376.
Staub, A., White, S. J., Drieghe, D., Hollway, E. C., \& Rayner, K. (2010). Distributional effects of word frequency on eye fixation durations. Journal of Experimental Psychology: Human Perception and Performance, 36, 1280-1293.
Underwood, N. R., \& McConkie, G. W. (1985). Perceptual span for letter distinctions during reading. Reading Research Quarterly, 20, 153-162.

# Fusiform Face Area in Chess Expertise 

Merim Bilalić (merim.bilalic@med.uni-tuebingen.de)<br>Department of Neuroradiology, Hoppe-Seyler Str. 2<br>Tübingen, 72076, Germany


#### Abstract

The ability to recognize faces is arguably one of the most important and most practiced skills. The possible functions of the fusiform face area (FFA), generally believed to be responsible for face recognition, also feature these two characteristics. On the one hand, there are claims that the FFA has evolved into a face specific module due to great importance of face processing. On the other, the FFA is seen as a general visual expertise module that distinguishes between individual examples within a single category. The previous studies used experts and novices on stimuli such as cars, birds or butterflies with ambiguous results. Here this research stream is extended to the game of chess, which does not share visible features with faces. The first study shows that chess expertise modulates the FFA activation when complex multi-object chess positions were presented. In contrast, isolated single chess objects did not produce different activation patterns among experts and novices. The second study confirmed that even a couple of isolated objects do not differently engage the FFA among experts and novices. The two studies provide support for the general expertise view of the FFA function, but also extend the scope of our understanding about the function of the FFA. The FFA does not merely distinguish between different exemplars. It also seems to engage into parsing complex multi-object stimuli that contain numerous functional and spatial relations.


Keywords: face perception; expertise; pattern cognition; chess; fMRI.

## Introduction

Faces are arguably the most important and most practiced stimuli. We start practicing face perception from our early moments and we are highly dependent on correctly distinguishing individual faces. It is fitting that the proposed functions of the fusiform face area (FFA), possibly the most important brain area in face perception, center on these two characteristics: importance and practice. On the one side, we have researchers who believe the FFA, due to, among other things, its importance in our lives, has evolved into a brain module exclusively specialized for perception of faces (Kanwisher, McDermott, \& Chun, 1997; Kanwisher \& Yovel, 2006). On the other side, in contrast to this facespecificity hypothesis, we have researchers that advance the claim that the FFA is a general expertise module (Gauthier, Skudlarski, Gore, \& Anderson, 2000; Gauthier, Tarr, Anderson, Skudlarski, \& Gore, 1999). According to the expertise hypothesis, the FFA is responsible for perceptual processes associated with differentiating among different objects within a single category stimulus (e.g., visual individuation), without regard of the type of stimuli. We
investigated the expertise hypotheses using the game of chess as a model for visual expertise.

The neural basis of face perception has been extensively investigated (for a review, see Kanwisher \& Yovel, 2006). Different category types have been used to investigate the expertise hypothesis, ranging from birds (Gauthier et al., 2000), cars (Gauthier et al., 2000; Gilaie-Dotan, Harel, Bentin, Kanai, \& Rees, 2012; Grill-Spector, Knouf, \& Kanwisher, 2004; McGugin, Gatenby, Gore, \& Gauthier, 2012; Xu, 2005), butterflies (Rhodes, Byatt, Michie, \& Puce, 2004), to novel object types (Gauthier et al., 1999). The results are mixed and their interpretation has been the focus of an extensive debate (Bukach, Gauthier, \& Tarr, 2006; Nancy Kanwisher \& Yovel, 2006; Op de Beeck \& Baker, 2010). Among factors complicating the interpretation is the visual similarity between faces and other categories employed - cars, birds, and even butterflies all have facelike features (Kanwisher \& Yovel, 2006).

The game of chess offers a way around the resemblance problem. Chess entails both individual objects and complex "chess positions" made out of individual objects. None of chess objects resemble faces and chess positions do not have much in common with face, at least not at the superficial perceptual level. Individual chess objects can be, however, differentiated just like individual faces. Expert chess players have accumulated extensive knowledge about chess objects and are quicker in recognizing them as well as their relations than novice chess players (Bilalić, Kiesel, Pohl, Erb, \& Grodd, 2011; Kiesel, Kunde, Pohl, Berner, \& Hoffmann, 2009; Saariluoma, 1995). The real (chess) expertise lies, however, in using knowledge to quickly size the gist of chess positions (Bilalić, Langner, Erb, \& Grodd, 2010; Bilalić, Turella, Campitelli, Erb, \& Grodd, 2012; Gobet \& Simon, 1996). This expertise process of automatically parsing complex multi-object environment bears similarity to that found in face perception. Both processes are automatic, quick, and efficient in binding individual features into meaningful units.

These characteristics make chess a suitable domain for investigation of the FFA expertise hypothesis. A recent study showed that expertise in chess is negatively correlated with the performance on face perception (Boggan, Bartlett, \& Krawczyk, 2012). One possible interpretation would be that both skills engage similar processes that compete for the resources in the same brain areas. Indeed, we (Bilalić, Langner, Ulrich, \& Grodd, 2011) recently showed that chess expertise mediates the activation in the FFA regardless of the task (domain specific or not) as long as the stimuli
feature naturalistic chess positions (but see Krawczyk, Boggan, McClelland, \& Bartlett, 2011). However, it is unclear how the FFA will behave with individual chess objects instead of multi-object positions. Here I present two studies that test the FFA responses with individual chess objects and complex chess positions. In the first study we presented participants with single isolated chess objects and chess positions during a 1 -back task. In the second study we used only a couple of isolated chess objects in chess specific tasks (see Bilalić et al., 2011).

## Study 1

## Method

The first study involved a 1-back task where participants indicated whether the current stimulus was the same as the previous one.
Participants Table 1 presents the information about the number of experts and novices, their mean age (with SD), and their chess ability score [mean Elo rating with SD; available only for experts] in both studies. All participants were male and right-handed. The sample size is relatively small, but it reflects the rarity of the studied group and is comparable to recent behavioural studies using chess experts (e.g., Bilalić et al., 2008a, 2008b, 2009; Brockmole et al., 2008; Kiesel et al., 2009). The small sample size is offset by the large differences between experts and novices. We also used exclusively male participants as they outnumber female chess players and we were not interested in gender differences. Written informed consent was obtained in line with the study protocol as approved by the Ethics Committee of Tübingen University.

Table 1: Participants - Overview

| Study | Group | Age $\pm$ SD | Elo $\pm$ SD | $n$ |
| :--- | :--- | :---: | :---: | :---: |
|  |  |  |  |  |
| 1 | Expert | $24 \pm 8$ | $2116 \pm 125$ | 12 |
|  | Novice | $27 \pm 6$ | ---- | 14 |
|  |  |  |  |  |
| 2 | Expert | $29 \pm 7$ | $2130 \pm 147$ | 8 |
|  | Novice | $29 \pm 5$ | --- | 8 |
|  |  |  |  |  |

Localizer Participants first passively watched pictures of faces and objects to localize face related areas (for more details, see Bilalic et al., 2011). The area used in further analysis, the right FFA and the right posterior superior temporal sulcus ( pSTS ) were then identified (contrast faces vs. objects; $p<.0001$ uncorr.) and isolated for the use in Study 1 and 2 - see Figure 1.
Stimuli and procedure The 1-back task in Study 1 featured the following stimuli: faces, isolated chess objects (pieces), and chess positions (see Figure 2). The stimuli of
each category were blocked in 12 second units that featured 6 individual stimuli (each stimuli taking 1.75 s with a break of 0.25 s between them). There were ten blocks of each category spread over two different runs. Baseline ( 18 s of black screen with a cross in the middle) was presented between the blocks of stimuli.
MRI acquisition and data analysis fMRI data were acquired using a 3-T scanner (Siemens Trio) with a 12channel head coil. All measurements covered the whole brain using a standard echo-planar-imaging (EPI) sequence with the following parameters: TR, 2 s ; FOV, $192 \times 192$; TE, 30 ms ; matrix size, $64 \times 64$; 32 slices with thickness of 3.2 and 0.8 mm gap resulting in voxels with the resolution of $3 \times 3 \times 3.2 \mathrm{~mm} 3$. Finally, anatomical images covering whole brain with 176 sagittal slices were obtained after the functional runs using an MP-RAGE sequence with a voxel resolution of $1 \times 1 \times 1 \mathrm{~mm} 3$ (TR, 2.3 s ; TI, $1.1 \mathrm{~s} ; \mathrm{TE}, 2.92$ ms ). SPM software package was used for analysis. All functional data were first preprocessed using standard SPM routines for realignment, coregistration, normalization and smoothing ( 8 mm ). Blocks of individual stimulus categories as conditions of interest were then modeled using the standard canonical response function. The ROI analysis was performed on the mean percentage signal change extracted using Marsbar SPM Toolbox from all the voxels within the selected region - FFA and pSTS.


Figure 1: FFA (upper picture - inferior view) and pSTS (lower picture - lateral view) used as regions of interest (ROI) in the studies.

## Results and discussion

The faces unsurprisingly activated the FFA more than the two chess categories, but chess positions also elicited more activation than chess objects (ANOVA for main effect of stimulus category $-F(2,48)=79, p=.001)-$ see Figure 2.

There was no overall effect of expertise $(F(1,24)=0.3$, ns.) but the expertise modulated activation depending on the stimulus category (ANOVA for interaction expertise x stimulus category $-F(2,48)=4.5, p=.016)$. While there were no differences between experts and novices on chess objects and faces, experts' FFA was more activated on the chess positions than the FFA of novices $(t(24)=2.2, p=$ $0.039)$.


Figure 2: FFA activation pattern among experts (blue) and novices (red) on faces, chess positions, and chess objects in Study 1.

Unlike with the FFA, in the pSTS there were no expertise effects $(F(1,24)=1.6, n s)$ nor there was interaction between expertise and stimulus categories $(F(2,48)=0.2$, $n s$ ) - see Figure 3. Faces again elicited most activation, which resulted in the significant main effect of stimulus categories $(F(2,48)=11.6, p=.001)$.

This is the first time both isolated and complex chess stimuli were used in a single study. The results confirm the previous study on chess expertise (Bilalic et. al., 2011) and its finding of FFA sensitivity to expertise on complex chess positions. Here it is shown that the same pattern of activation does not generalize to single isolated objects. When isolated chess pieces were presented, expertise did not modulate the FFA activity.

## Study 2

## Method

The second study again used chess stimuli in chess specific tasks but this time they were neither completely isolated they always featured two objects. The study has been published (Bilalić et al., 2011) but here we use the unpublished ROI analysis on the FFA and pSTS.
Participants Information about participants is presented in Table 1.
Task, stimuli and procedure There were three tasks (Figure 4). In the check task, participants had to indicate if


Figure 3: pSTS activation pattern among experts (blue) and novices (red) on faces, chess positions, and chess objects in Study 1.
the white king is given check (one of the most important aspects in the game of chess) by the present black piece. In the identity task, the participants were presented with the same stimuli as in the check task, but this time they had to identify the black piece presented. In the control task, chess pieces were changed for geometrical shapes and the participants had to indicate the identity of the shape (diamond or square). We again used block design (for more details, see (Bilalić et al., 2011) with blocks of 13.5 s containing 4 trials.
MRI acquisition and data analysis This part of the study was the same as the previous study, except that this time a different EPI sequence was used: TR, 2.5 s ; FOV, $192 \times 192$; TE, 35 ms ; matrix size, $64 \times 64 ; 36$ slices with thickness of 3.20 .8 mm gap resulting in voxels with the resolution of $3 \times 3 \times 4 \mathrm{~mm} 3$. We again specified condition of interest as blocks, convolved it with HRF and analyzed responses in the selected ROIs using MarsBar toolbox.

## Results and discussion

Unlike in the previous study, there were no differences among experts and novices in the FFA activity in none of the three tasks (ANOVA for expertise, $F(1,14)=0.1$, $n s$ ) see Figure 4. There were no differences between the tasks (ANOVA for task, $\mathrm{F}(2,28)=1.1, \mathrm{~ns}$ ) nor there were differences between the task among the groups (ANOVA for task x expertise interaction, $\mathrm{F}(2,28)=0.9$, ns).

Similarly, the pSTS also did not produce different responses among experts and novices in all three tasks ( $\mathrm{F}(1$, $14)=1.8, \mathrm{~ns})$ and there was no main effect of task $(\mathrm{F}(2,28)$ $=5.1, \mathrm{~ns})$ nor interaction with expertise $(\mathrm{F}(2,28)=0.04$, ns).


Figure 4: FFA activation pattern among experts (blue) and novices (red) on control task (identifying geometrical shapes), identity task (identifying chess objects), and check task (identifying check relations among objects) in Study 2.


Figure 5: pSTS activation pattern among experts (blue) and novices (red) on control task (identifying geometrical shapes), identity task (identifying chess objects), and check task (identifying check relations among objects) in Study 2.

## General Discussion

Our previous study (Bilalić et al., 2011) showed that the FFA is sensitive to expertise as long as chess positions were present, even when the task at hand did not require specific chess activity. Here this result is extended to other kind of chess stimuli - isolated chess objects. Study 1 showed that chess positions, stimuli featuring several chess objects, produced an expertise effect, confirming our previous study. There were no, however differences when isolated chess objects were presented. The lack of expertise modulation with isolated chess objects was further confirmed in Study 2. Even when two objects formed a relation, the FFA was not responding differently in experts and novices.
Chess objects (as featured in Study 1) and chess relations (as featured in Study 2) are main building blocks of chess positions and the very same stimuli that consistently elicit
expertise effects in the FFA. It is thus surprising to find a lack of expertise effect in the FFA when it comes to isolated chess objects and their relations. One reason could be that Study 1 did not use explicit individuation between chess objects. Study 2, however, did use the differentiation between chess objects (based on which the tasks could be only done), not to mention that individuation processes are assumed to be implicit and automatic. It is, of course, possible that the lack of expertise effects in the FFA was due to low power of the studies. After all, the studies featured dozen participants in each group at most and the non-significant results should not be confused with a complete absence of effects. It is nevertheless the case that chess positions produced significant expertise effects in FFA in this and previous study, although both studies did not have large samples.
The FFA seems to be the only face area involved in chess perception. Here it was again shown that the pSTS does not differentiate between experts and novices on chess stimuli. As with the previous non-significant effect, one needs to be careful with conclusions. It seems reasonable, however, to conclude that the role of pSTS in chess expertise is arguably not as pronounced as that of FFA.
Although visually different, chess positions are essentially rather similar to faces. They are also made out of different individual parts (chess objects and relations between them). These parts are perceived as such only by beginners. Experienced chess players perceive chess positions rather as meaningful units, not unlike most of us perceive faces. The stored knowledge structures in memory (Gobet \& Simon, 1996) that enable them to quickly recognize situations on the board. In that sense, processes involved into parsing chess positions are much closer to those involved in face perception that are the processes involved in recognition of chess individual objects.
The exact role of FFA in chess expertise remains to be determined. Our previous study (Bilalic et al., 2011) demonstrated that experts' FFA reacts to chess positions without regard of the executed task. Even task that were not chess related (e.g., counting the number of all chess objects on the board) elicited expertise effects in the FFA. This indicates that the chess related processes in the FFA are automatic and stimulus, not task, dependent. In contrast, the other chess areas identified in our studies (Bilalić et. al., 2010; 2012), such as a part of the collateral sulcus and retrosplenial cortex, are also sensitive to task demands in addition to stimuli. How these regions are connected and how and to what extend they enable chess expertise remains an important question for future research.

The results also revise the expertise hypothesis by providing evidence against individuation as the primary function of the FFA. Study 1 did not involve explicit individuation as the individual chess objects were only passively observed. Study 2, however, involved explicit identification of a single chess objects (Identity task) and there were still no expertise-modulated response in the FFA. Only chess positions, consisting of numerous chess objects
and relations between those objects, produced different activation in the FFA of experts and novices. These results support previous studies demonstrating the importance of holistic parsing of individual parts of faces as the main FFA function (Arcurio, Gold, \& James, 2012; Gold, Mundy, \& Tjan, 2012), and put under questions is role in individuation.
They also revise the expertise hypothesis by providing evidence that complexity of stimuli and the processes that enable their fast and efficient perception are at the heart of the FFA function, and not only individuation.
These two chess studies, together with the previous work on the similarities between face and chess perception (Bilalić, et al., 2011; Boggan, 2012), underline the suitability of chess as an exploration vehicle in cognitive neuroscience.

## Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (DFG BI 1450/1-2). I am grateful for the help of all expert and novice chess players, I would also like to thank to Robert Langner, Michael Erb, Wolfgang Grodd and Rolf Ulrich for the help with the presented project.

## References

Arcurio, L. R., Gold, J. M., \& James, T. W. (2012). The response of face-selective cortex with single face parts and part combinations. Neuropsychologia, 50, 2458-2459
Bilalić, M., Kiesel, A., Pohl, C., Erb, M., \& Grodd, W. (2011). It Takes Two-Skilled Recognition of Objects Engages Lateral Areas in Both Hemispheres. PLoS ONE, 6(1), e16202.
Bilalić, M., McLeod. P., \& Gobet, F. (2008a). Why good thoughts block better ones: the mechanism of the pernicious Einstellung (set) effect. Cognition, 108, 652661.

Bilalić, M., McLeod. P., \& Gobet, F. (2008b). Inflexibility of experts: reality or myth? Quantifying the Einstellung effect in chess masters. Cognitive Psychology, 56, 73102.

Bilalić, M., McLeod. P., \& Gobet, F. (2009). Specialization effect and its influence on memory and problem solving in expert chess players. Cognitive Science, 33, 1117-1143.
Bilalić, M., Langner, R., Erb, M., \& Grodd, W. (2010). Mechanisms and neural basis of object and pattern recognition: A study with chess experts. Journal of Experimental Psychology: General, 139(4), 728-742.
Bilalic, M., Langner, R., Ulrich, R., \& Grodd, W. (2011). Many Faces of Expertise: Fusiform Face Area in Chess Experts and Novices. Journal of Neuroscience, 31(28), 10206-10214.
Bilalić, M., Turella, L., Campitelli, G., Erb, M., \& Grodd, W. (2012). Expertise modulates the neural basis of context dependent recognition of objects and their relations. Human Brain Mapping, 33(11), 2728-2740.

Boggan, A. L., Bartlett, J. C., \& Krawczyk, D. C. (2012). Chess masters show a hallmark of face processing with chess. Journal of Experimental Psychology: General, 141(1), 37.
Brockmole, J. R., Hambrick, D. Z., Windisch, D. J., \& Henderson, J. M. (2008). The role of meaning in contextual cueing: evidence from chess expertise. Quarterly Journal of Experimental Psycholology, 61, 1886 -1896.
Bukach, C. M., Gauthier, I., \& Tarr, M. J. (2006). Beyond faces and modularity: the power of an expertise framework. Trends in cognitive sciences, 10(4), 159-166.
Gauthier, I., Skudlarski, P., Gore, J. C., \& Anderson, A. W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. Nature Neuroscience, 3(2), 191-7.
Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., \& Gore, J. C. (1999). Activation of the middle fusiform "face area" increases with expertise in recognizing novel objects. Nature Neuroscience, 2(6), 568-73.
Gilaie-Dotan, S., Harel, A., Bentin, S., Kanai, R., \& Rees, G. (2012). Neuroanatomical correlates of visual car expertise. NeuroImage, 62(1), 147-153.
Gobet, F., \& Simon, H. A. (1996). Templates in Chess Memory: A Mechanism for Recalling Several Boards. Cognitive Psychology, 31(1), 1-40.
Gold, J. M., Mundy, P. J., \& Tjan, B. S. (2012). The perception of a face is no more than the sum of its parts. Psychological Science 23, 427-434.
Grill-Spector, K., Knouf, N., \& Kanwisher, N. (2004). The fusiform face area subserves face perception, not generic within-category identification. Nature neuroscience, 7(5), 555-562.
Kanwisher, N, McDermott, J., \& Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 17(11), 4302-11.
Kanwisher, Nancy, \& Yovel, G. (2006). The fusiform face area: a cortical region specialized for the perception of faces. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 361(1476), 210928.

Kiesel, A., Kunde, W., Pohl, C., Berner, M. P., \& Hoffmann, J. (2009). Playing chess unconsciously. Journal of experimental psychology. Learning, memory, and cognition, 35(1), 292-298.
Krawczyk, D. C., Boggan, A. L., McClelland, M. M., \& Bartlett, J. C. (2011). The neural organization of perception in chess experts. Neuroscience letters, 499(2), 64-69.
McGugin, R. W., Gatenby, J. C., Gore, J. C., \& Gauthier, I. (2012). High-resolution imaging of expertise reveals reliable object selectivity in the fusiform face area related to perceptual performance. Proceedings of the National Academy of Sciences, 109(42), 17063-17068.

Op de Beeck, H. P., \& Baker, C. I. (2010). The neural basis of visual object learning. Trends in Cognitive Sciences, 14(1), 22-30.
Rhodes, G., Byatt, G., Michie, P. T., \& Puce, A. (2004). Is the fusiform face area specialized for faces, individuation, or expert individuation? Journal of cognitive neuroscience, 16(2), 189-203.
Saariluoma, P. (1995). Chess Players' Thinking: A Cognitive Psychological Approach. Routledge.
$\mathrm{Xu}, \mathrm{Y}$. (2005). Revisiting the role of the fusiform face area in visual expertise. Cerebral cortex (New York, N.Y.: 1991), 15(8), 1234-1242.

# Analyzing Students' Metacognitive Strategies in Open-Ended Learning Environments 

Gautam Biswas (gautam.biswas@ vanderbilt.edu)<br>Department of EECS/ISIS, $102516^{\text {th }}$ Ave South<br>Nashville, TN 37212 USA

John S. Kinnebrew (john.s.kinnebrew@ vanderbilt.edu)<br>Department of EECS/ISIS, $102516^{\text {th }}$ Ave South<br>Nashville, TN 37212 USA

James R. Segedy (james.segedy@ vanderbilt.edu)<br>Department of EECS/ISIS, $102516^{\text {th }}$ Ave South<br>Nashville, TN 37212 USA


#### Abstract

Novices often lack metacognition and self-regulation skills that are important for effective learning. Betty's Brain, an open-ended computer-based learning environment helps students practice and develop metacognitive strategies as they learn science topics. We extend previous work on sequence mining methods to discover students' frequentlyused behavior patterns from their activity sequences. Our results show that it is possible to interpret aspects of students' learning strategies and their effectiveness by taking into account the context of their activities in the system.


Keywords: open-ended learning environments, metacognition, measuring metacognition, scaffolding, sequence mining.

## Introduction

Cognitive scientists have established that metacognition and self-regulation are essential for developing effective learning strategies in the classroom and beyond (Bransford et al, 2000; Zimmerman, 2001). However, novice learners often have ineffective self-regulation profiles, which may be attributed to their lacking the well-organized domain knowledge structures of experts. This affects their ability to break down their learning and problem solving into distinct task understanding, planning and solution generation, monitoring and evaluation phases, leading them to use suboptimal learning and problem solving strategies (Chi et al, 1988; VanLehn, 1996).

Our research group has developed Betty's Brain, an openended learning environment (OELE), to study how students develop metacognitive strategies that include constructing information and monitoring as they learn science topics (Leelawong and Biswas, 2008). Our approach utilizes trace methodologies derived from students' actions and activity patterns in the environment to infer aspects of their metacognitive abilities (Aleven et al, 2006; Azevedo, et al., 2012; Hadwin et al, 2007). This is based on a metacognition as events hypothesis, which theorizes that the use of metacognitive strategies manifests as continually unfolding events that can be inferred from learners' behaviors.

In this paper, we extend our previous work on using sequence mining methods to discover students' frequently-
used behavior patterns from their activity sequences as they work in the Betty's Brain system (Kinnebrew \& Biswas, 2012). In particular, we extend our techniques for analyzing students' action sequences by (i) interpreting and characterizing behavior patterns using a cognitive/metacognitive model of the task, (ii) mapping students' frequently observed cognitive and metacognitive process patterns back into their overall activity sequences, and (iii) using metrics to evaluate the effectiveness of these processes. The results in this paper represent a post hoc analysis of student behaviors, but our longer term goal is to use such results to monitor and measure students' cognitive and metacognitive processes online as they work on their learning and problemsolving tasks, and use these results to develop adaptive scaffolding mechanisms that support student learning.

## Background

Metacognition is often described as being made up of two constituent parts (Flavell et al, 1985; Veenman, 2012): (1) Metacognitive knowledge, which is declarative and deals with the interplay between knowledge of one's abilities to perform tasks, the nature of the task, and the strategies one can employ to successfully perform the task; and (2) Metacognitive monitoring and regulation, which includes activities related to planning, monitoring, and evaluating one's cognitive processes in order to better regulate those processes in the future.

Researchers have established strong links between learners' metacognitive abilities and their effectiveness in executing cognitive processes. Winne (1996) characterizes cognition as dealing with knowledge of objects and operations on objects (the object level) while characterizing metacognition as the corresponding meta-level that contains information about cognitive processes. Metacognitive monitoring brings the two levels together, as it describes the process of observing one's own execution of cognitive processes at the object level and exerting control over the object level using metacognitive knowledge and strategies.

An important implication of the interplay between cognition and metacognition relates to the dependence of metacognition on cognition (Land, 2000). In other words, metacognitive knowledge may not be sufficient for achieving
success in learning and problem solving, especially for learners who lack the cognitive skills and background knowledge necessary for interpreting, understanding, and organizing critical aspects of the problem under study (Bransford et al, 2000). Learners may also lack knowledge of effective strategies (e.g., the ability to extract relevant information when reading a science text), and, therefore, resort to suboptimal strategies in performing their tasks (Azevedo, 2005; Kinnebrew \& Biswas, 2012). Poor selfjudgment abilities result in difficulties for monitoring and evaluating one's own effectiveness and progress, which can be a significant stumbling block in selecting and implementing relevant strategies in a timely manner.

However, research studies have shown that with proper scaffolding, middle school students can improve their metacognitive awareness and develop effective metacognitive strategies (Kramarski \& Mevarech, 2003). Our system, Betty's Brain is designed to help middle school students develop metacognitive knowledge and strategies as they learn about science topics. Other systems with similar goals include MetaTutor (Azevedo, et al., 2012) and Crystal Island (Rowe, et al., 2011).

## Betty's Brain

Betty's Brain (Figure 1) is an open-ended learning environment (Land, 2000) that provides students with a learning context and a set of tools for pursuing authentic and complex learning tasks. Students teach a virtual agent, Betty, about science topics by constructing a causal map. The goal for students using Betty's Brain is to teach Betty a map, whose correctness is determined in relation to a hidden, expert causal map.

The students' learning and teaching tasks are organized around three activities: (1) reading hypertext resources to learn the domain material, (2) building and refining a causal map, which represents the domain material, and (3) asking Betty to take a quiz. Students explicitly teach Betty by constructing a causal map. For example, they may draw a causal link between garbage and landfills and methane to represent the relationship garbage and landfills increase methane (a greenhouse gas). Students can check what Betty knows by asking her questions, e.g., if garbage and landfills decrease, what effect does it have on polar sea ice? To answer questions, Betty uses qualitative reasoning that operates through chains of links from the source concept to the target concept (Leelawong \& Biswas, 2008). The learner can further probe Betty's understanding by asking her to explain her answer. Betty illustrates her reasoning by explaining her thinking and animating her explanation by highlighting concepts and links on the map as she mentions them.

Learners can assess Betty's (and, therefore, their own) progress in two ways. After Betty answers a question, learners can ask Mr. Davis, a pedagogical agent that serves as a mentor, to evaluate the answer. Learners can also have Betty take a quiz on one or all of the sub-topics in the resources. Quiz questions are selected dynamically to reflect the current state of the student's map; questions are chosen (in pro-
portion to the completeness of the map) for which Betty will generate correct answers. The remaining questions produce incorrect answers, and they direct the student's attention to incorrect and missing links.

After Betty takes a quiz, her results, including the causal map she used to answer the questions appear on the screen as shown in Figure 1. The quiz questions, Betty's answer, and the Mentor's assigned grade, i.e., correct, correct but incomplete, or incorrect appear on the top of the window. Clicking on a question will highlight the causal links that Betty used to answer that question. To help students keep track of correct and incorrect links, the system allows students to annotate them with a green check-mark (correct), a red X (incorrect), or a gray question mark (not sure).


Figure 1: Betty's Brain Interface with Quiz Window

## Cognitive/Metacognitive Process Model

To interpret students learning behaviors on the system, we have developed a model that takes into account the tight connection between the cognitive and metacognitive processes needed to address the learning task effectively. Overall, this model includes four primary processes that students are expected to engage in while using Betty's Brain: (1) Goal Setting \& Planning, (2) Knowledge Construction (KC), (3) Monitoring (Mon), and (4) Help Seeking. In this work we focus on the KC and Mon process models.

Knowledge construction includes metacognitive strategies for (1) information seeking, i.e., determining when and how to locate needed information in the resources, and (2) information structuring, i.e., organizing one's developing understanding of the domain knowledge into structural components (e.g., causal links). In executing these metacognitive processes, learners have to apply relevant cognitive processes listed under information seeking and structuring. Seeking information, for example, requires that students to identify the causal information by reading the resources and making sense of the content. Similarly, information structuring captures the process of successfully converting the acquired information into causal links and adding them to the causal map.

Monitoring processes include (1) model assessment, i.e., assessing the correctness of all or a part of the causal model, and (2) progress recording, i.e., making explicit annotations to mark parts of the causal model as correct, which makes it easier to focus on parts of the map that need more work. Successful execution of monitoring metacognitive processes relies on students' abilities to execute cognitive processes for assessing the causal model (via questions, explanations, quizzes, and question evaluations) and recording progress (via note taking and annotating links with correctness information). The cognitive and metacognitive process model provides a framework for interpreting students learning activities and behaviors (activity sequences) on the system.

## Measuring Cognition and Metacognition

We have developed a set of data mining methods for analyzing students' learning activity sequences and assessing their learning processes as they work in Betty's Brain. In addition, we have developed visualization methods for measuring how student behaviors evolve during the course of the intervention depending on the type of feedback and support that they received from the Mentor agent. In particular, we were interested in studying whether students' suboptimal behaviors were replaced by more optimal strategies as the intervention progressed.

To assess student activities with respect to our cognitive/metacognitive model, we calculate four measures: map edit effectiveness, map edit support, monitoring effectiveness, and monitoring support. Map edit effectiveness is calculated as the percentage of causal link additions, removals, and modifications that improve Betty's causal map. Map edit support is defined as the percentage of causal map edits that are supported by previous reading of pages in the resources that discuss the concepts connected by the manipulated causal link. Monitoring effectiveness is calculated as the percentage of quiz questions and explanations that generate specific correctness information about one or more causal links. For example, all of the links used in a quiz question whose answer is marked correct, must be correct. If the answer to a question is incorrect, at least one of the links used in the answer must be incorrect. Finally, monitoring support is defined as the percentage of causal link annotations that are supported by previous quiz questions and explanations. For support metrics, a further constraint is added: an action can only support another action if both actions occur within the same time window, and we calculated support for a ten minute time window.

The information for calculating the measures and deriving student behavior using sequence mining is extracted from $\log$ files. For example, if a student accesses a page in the resources, this is logged as a Read action that includes additional information, e.g., the page accessed. In this work, students' activity sequences contain six categories of actions: (1) Read, (2) Link Edit, (3) Query, (4) Quiz, (5) Explanation, and (6) Link Annotation. Actions were further distinguished by context details, such as the correctness of a link edit. Sequence mining techniques are applied to discov-
er frequent behavior patterns for students in a given group (Kinnebrew, et al., 2013; Kinnebrew \& Biswas, 2012). Students' use of metacognitive processes was determined by interpreting the patterns using the cognitive and metacognitive model.

## Method

The present analysis used data from a recent classroom study with Betty's Brain in which students learned about the greenhouse effect and climate change. The study tested the effectiveness of two support modules designed to scaffold students' understanding of cognitive and metacognitive processes important for success in Betty's Brain. The knowledge construction (KC) module provided support on how to identify causal relations in the resources, and the monitoring (Mon) support module helped students understand how to use Betty's quizzes to identify correct and incorrect causal links on the causal map. Participants were divided into three treatment groups. The KC group (KC-G) used a version of Betty's Brain that included the KC support module and a causal link tutorial that they could access at any time during learning. The tutorial allowed students to practice identifying causal relations in short text passages. The Mon group (Mon-G) used a version of Betty's Brain that included the Mon support module and a marking links correct tutorial that they could access at any time during learning. The tutorial presented practice problems in which students used the results of graded quiz questions and the causal map used to answer those questions to select the links that could be marked as correct. Finally, the control group (Con-G) used a version of Betty's Brain that included neither the tutorials nor the support modules.

The KC module was activated when three out of a student's last five map edits were incorrect, at which point Mr. Davis would begin suggesting strategies for identifying causal links during reading. Should students continue to make incorrect map edits despite this feedback, the KC module activated a second tier of support: guided practice. During guided practice, students were moved to the causal link tutorial where they read short text passages and expressed the primary idea in the passage as a causal relation. When they worked on the tutorial, students were not permitted to access any other portion of the program. Students completed the tutorial session once they solved five problems correctly without making a mistake.

The Mon module was activated after the third time students did not use evidence from quizzes and explanations to annotate links on their map. At this time, Mr. Davis began suggesting strategies for using quizzes and explanations to identify and keep track of which links were correct. Additionally, Mr. Davis discouraged students from annotating links as being correct without using the suggested strategies. Should students continue to use quizzes and explanations without annotating links correctly, the Mon module provided students with guided practice. Like the KC tutorial, students had to complete five problems correctly on the first try to complete the tutorial session.

Seventy-three seventh grade students from four middle Tennessee science classrooms, taught by the same teacher, participated in the study. Because use of Betty's Brain relies on students' ability to independently read and understand the resources, the system is not suited to students with limited English proficiency or cognitive-behavioral problems. Therefore, data from English as a Second Language (ESL) and special education students were not analyzed. Similarly, we excluded the data of students who missed more than two class periods of work on the system. Our experimental analysis used data collected from fifty-two students who participated in the study.

Learning was assessed using a pre-post test design. Each written test consisted of five questions that asked students to consider a given scenario and explain its causal impact on climate change. Scoring was based on the causal relations that students used to explain their answers to the questions, which were then compared to the chain of causal relations used to derive the answer from the expert map. One point was awarded for each causal relationship in the student's answer that came from or was closely related to an expert causal link. The maximum combined score for the five questions was 16 . Two coders independently scored a subset of the pre- and post-tests with at least $85 \%$ agreement, at which point the coders split the remaining tests and individually coded the answers and computed the scores.

Performance on the system was assessed by calculating a score for the causal map that students created while teaching Betty. This score was computed as the number of correct links (the links in the student's map that appeared in the expert map) minus the number of incorrect links in the student's final map. We also used the log data collected from the system to derive students' behavior patterns, interpret them using our cognitive/metacognitive model, and study the temporal evolution of the observed KC and Mon strategies over the period of the intervention.

Study duration was 9 school days. During the first 60 minute class period, students completed the pre-test. During the second and third class periods, researchers introduced students to causal modeling and reasoning with causal models, and how to identify causal relations in text passages. During this time, students completed paper-and-pencil group exercises involving causal reasoning and identifying causal relations. During the fourth class period, students were provided with hands-on system training by the researchers. Students then spent four class periods using their respective versions of Betty's Brain with minimal intervention by the teachers and the researchers. On the ninth day, students completed the post-test.

## Results

Figure 2 presents the overall learning and performance results for each condition in the intervention. Repeated measures ANOVA performed on the data revealed a significant effect of time on test scores ( $\mathrm{F}=28.66, \mathrm{p}<0.001$ ). Pairwise comparison of the three groups revealed that the MonG had marginally better learning gains than KC-G, which
had better learning gains than the Con-G group. The Mon-G learning gains were significantly better than the Con-G gains at the 0.1 significance level ( $\mathrm{p}<.075$ ), indicating the two interventions may have resulted in better understanding of the science content. The small sample size and the large variations in performance within groups made it difficult to achieve statistical significance in these results. However, one positive aspect of this finding is that while students in the Mon-G and KC-G spent an average of $10 \%$ and $17 \%$ of their time in guided practice, respectively, they learned, on average, just as much, if not more, than the Con-G students.


Figure 2: Pre-post Test Results (mean (std dev)) and Final Map Score

To assess students' overall behaviors, we calculated the effectiveness and support measures, which are illustrated in Table 1. The KC-G students had the highest scores on both map editing effectiveness and support, suggesting that the KC feedback did help students more effectively and systematically read and construct their causal maps (however, only the map edit support showed a statistically significant difference, KC-G > Con-G, $p=0.02$, and the map edit effectiveness illustrated a trend, KC-G > Con-G, $p=0.08$ ). However, the monitoring support did not help the Mon-G students do better than the other two groups for monitoring effectiveness or support. The Mon-G students did have the highest monitoring effectiveness, but it was not statistically significant. Further, the Con-G students had the monitoring support average ( $p<0.10$, when comparing with other groups). It is not clear why the Mon or KC support and tutorials resulted in students performing less supported monitoring activities tan the Con-G students.

Table 1: Effectiveness \& Support Measures
((mean (std dev)) by Group

| Measure | Con-G | KC-G | Mon-G |
| :--- | :--- | :--- | :--- |
|  | 0.46 | 0.52 | 0.5 |
| Map edit effectiveness | $(0.13)$ | $(0.07)$ | $(0.12)$ |
|  | 0.43 | 0.64 | 0.55 |
| Map edit support | $(0.25)$ | $(0.19)$ | $(0.23)$ |
|  | 0.3 | 0.32 | 0.4 |
| Monitoring effectiveness | $(0.22)$ | $(0.21)$ | $(0.20)$ |
|  | 0.61 | 0.32 | 0.33 |
| Monitoring support | $(0.30)$ | $(0.4)$ | $(0.32)$ |

In order to investigate student learning behavior in more detail, we employed sequence mining analyses to identify 143 different action patterns that were observed in the majority of students. Table 2 lists the 10 most frequent patterns that employed at least two actions and could be interpreted as a metacognitive strategy in our cognitive/metacognitive model. Each pattern is defined by two or more primary actions, and each action is qualified by one or more attributes. For example, a [Read] $\rightarrow$ [Add correct link, relevant to recent actions] pattern describes a KC behavior, where the student added a correct causal link to the map after a [Read] action where the student read a page that discussed the added link. In contrast, the action labeled [Read] $\rightarrow$ [Add incorrect link, relevant to recent actions] implies the student added an incorrect link even after reading a page that contained information about the link. The $\rightarrow$ symbol implies that the action to the left of the arrow preceded the action to the right of the arrow.

The average frequency represents the average number of times students used a particular behavior pattern when they worked on the system. These numbers are broken down for the three conditions. The last column represents our interpretation of the type of strategy a particular behavior represents. In this study, the strategy corresponding to a behavior was determined by the category of the cognitive process ( KC or Mon) implied by the individual actions that made up the behavior Therefore, some behaviors, e.g., pattern \#3: [Quiz] $\rightarrow$ [Remove incorrect link], span KC and Mon (KC+Mon) strategies.

The frequency numbers indicate that for almost all of the top 10 behaviors the CON-G showed a higher frequency of use than the two experimental groups. This may be partly attributed to the time the KC-G and Mon-G groups spent in tutorials, therefore reducing the amount of time they spent on the map building task. However, an equally likely reason may be that the CON-G students used more trial-and-error approaches, spending less time editing and checking the correctness of their maps in a systematic way. This is further supported by looking at the highest average frequency behaviors for each of the groups. The top five behavior strategies for the Mon-G students are primarily Mon or $\mathrm{KC}+\mathrm{Mon}$ related (patterns $1,3,5,7$, and 9 ), involving quizzes, map editing, and explanations. KC-G students, on the other hand, more often employed KC strategies related to adding and removing links along with a couple of strategies that combine KC and Mon activities. The Con-G students seem to have employed KC and Mon strategies in about equal numbers, but they were less effective in using these strategies.

An interesting strategy is pattern \#10: [Add incorrect link (AIL)] $\rightarrow$ [Quiz (Q)] $\rightarrow$ [Remove incorrect link (RIL)]. This may represent a strategy where a student first adds a link (which happens to be incorrect) and then takes a quiz to determine if the quiz score changes. Depending on the outcome (in this case, the score likely decreased), the student determines that the link added was incorrect, and, therefore, removes it. This represents a trial-and-error strategy. While
students in all three groups used this strategy, the Mon-G group used it with lower frequency than the other two groups, and this may be attributable to the effectiveness of the Monitoring scaffolding. To study this pattern further we developed two measures: (1) a measure of cohesiveness of the pattern, i.e., in what percentage of the AIL $\rightarrow \mathrm{Q} \rightarrow$ RIL patterns was the delete action supported by the quiz result; and (2) a support measure, i.e., in what percentage of the AIL $\rightarrow \mathrm{Q} \rightarrow$ RIL patterns was the addition of the link supported by recent actions. The MON group had higher cohesiveness ( 41.9 to 38.0 and 37.3 for the CON and KC groups) and support ( 27.7 to 20.3 and 187.7 for the CON and KC groups) measures, implying that they used this pattern in a more systematic way than the other two groups.

## Discussion and Conclusions

The results presented in the previous section provide evidence that a combination of theory-driven measures and da-ta-driven mining techniques can be successfully employed to produce a more complete description of the metacognitive strategies use in their learning and problem-solving tasks. In our work on investigating cognitive and metacognitive processes in Betty's Brain, we had to carefully instrument the system to collect rich data on the students' activities and the context associated with those activities. Post hoc mining and analysis of this data revealed a number of interesting results. Perhaps most important, the results show (i) that it is possible to infer aspects of students' use of strategies as they learn through these data mining and analysis techniques combined with a cognitive/metacognitive model of the task, and (ii) that tracking student performance and related context information with respect to their activities allows us to better characterize these strategies as suboptimal versus optimal.

Our analyses in this study focused on students' knowledge construction and monitoring strategies. Knowledge construction strategies include seeking out information, thinking deeply about the material to develop a sufficient understanding to apply it to model building and problem solving tasks. In particular, information structuring strategies in Betty's Brain help students with their map-building activities, which include understanding the structure of the causal model, the ability to construct it in parts, the ability to add links correctly to an existing structure, and also the ability to reason (e.g., answer questions, formulate hypotheses) with the evolving structure. The primary monitoring strategies relate to determining when and how to check the correctness of the current causal map, and then, in more detail, using the quiz (assessment) results to determine the correctness of individual links, and what parts of the map are incomplete or still need work.

In summary, the analysis presented in this paper successfully employed our metacognition measurement framework to evaluate the effects of scaffolding support for metacognitive and cognitive processes important for success in Betty's Brain. In particular, we applied these analyses to a comparison of different versions of Betty's Brain, a version that pro-

Table 2: Comparison of Pattern Frequencies across Conditions

| Rank | Pattern | Avg. Frequency |  |  | Model <br> Category |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | CON | KC | MON |  |
| 1 | [Add incorrect link] $\rightarrow$ [Quiz] | 11.20 | 7.35 | 8.24 | KC+Mon |
| 2 | [Add incorrect link] $\rightarrow$ [Remove incorrect link] | 6.00 | 12.65 | 3.71 | KC |
| 3 | [Quiz] $\rightarrow$ [Remove incorrect link] | 7.87 | 6.10 | 6.29 | KC+Mon |
| 4 | [Add concept] $\rightarrow$ [Add correct link] | 7.53 | 6.75 | 4.94 | KC |
| 5 | [Quiz] $\rightarrow$ [Explanation] | 8.40 | 3.80 | 5.35 | Mon |
| 6 | [Remove incorrect link] $\rightarrow$ [Add incorrect link] | 4.53 | 9.20 | 3.41 | KC |
| 7 | [Add correct link] $\rightarrow$ [Quiz] | 5.87 | 4.05 | 5.06 | KC+Mon |
| 8 | [Remove incorrect link] $\rightarrow$ [Quiz] | 5.93 | 4.45 | 4.12 | KC+Mon |
| 9 | [Explanation] $\rightarrow$ [Explanation] | 5.67 | 2.95 | 4.88 | Mon |
| 10 | [Add incorrect link] $\rightarrow$ [Quiz] $\rightarrow$ [Remove incorrect link] | 5.20 | 4.40 | 3.88 | KC+Mon |

Bransford J, Brown A, Cocking R (eds) (2000) How people learn. National Academy Press Washington, DC, Washington.
Chi M, Glaser R, Farr M (1988) The nature of expertise. Lawrence Erlbaum Associates, Inc.
Flavell J, Miller P, Miller S (1985) Cognitive development. Prentice-Hall Englewood Cliffs, NJ Hadwin A, Nesbit J, Ja-mieson-Noel D, Code J, Winne P (2007) Examining trace data to explore self-regulated learning. Metacognition and Learning, 2(2):107-124.
Kinnebrew J.S., Biswas G. (2012) Identifying learning behaviors by contextualizing differential sequence mining with action features and performance evolution. In: Proc. 5th International Conference on Educational Data Mining (EDM 2012), Chania, Greece
Kinnebrew J.S., Loretz KM, Biswas G (2013) A contextualized, differential sequence mining method to derive students' learning behavior patterns. Journal of Educational Data Mining
Kramarski B., Mevarech Z. (2003) Enhancing mathematical reasoning in the classroom: The effects of cooperative learning and metacognitive training. American Educational Research Journal 40(1):281-310
Land S. (2000) Cognitive requirements for learning with open-ended learning environments. Educational Technology Research and Development, 48(3):61-78
Leelawong K, Biswas G (2008) Designing learning by teaching agents: The Betty's Brain system. International Journal of Artificial Intelligence in Education, 18(3):181208.

Rowe, J., Shores, L., Mott, B., and Lester, J. (2011). Integrating learning, problem solving, and engagement in nar-rative-centered learning environments. Intl Jour of Artificial Intelligence in Education, 21(1-2):115-133.
VanLehn K (1996) Cognitive skill acquisition. Annual review of psychology 47(1):513-539.
Veenman M (2012) Metacognition in science education: Definitions, constituents, and their intricate relation with cognition. Metacognition in Science Education, pp 21-36.
Winne P (1996) A metacognitive view of individual differences in self-regulate learning. Learning and individual differences 8(4):327-353.
Zimmerman B (2001) Theories of self-regulated learning and academic achievement: An overview and analysis. In: Zimmerman B, Schunk D (eds) Self-regulated learning and academic achievement: Theoretical perspectives, Erlbaum, Mahwah, NJ, pp 1-37.

# "Are we still talking about the same thing?' MEG reveals perspective-taking in response to pragmatic violations, but not in anticipation 

Sara Bögels ${ }^{1}$ (Sara.Bogels@mpi.nl), Dale Barr ${ }^{2}$ (Dale.Barr@ glasgow.ac.uk), Simon Garrod ${ }^{2}$ (Simon.Garrod@glasgow.ac.uk), \& Klaus Kessler ${ }^{2,3}$ (Klaus.Kessler@glasgow.ac.uk)

${ }^{1}$ MPI for Psycholinguistics, P.O. Box 310, 6500 AH Nijmegen, The Netherlands<br>${ }^{2}$ Institute of Neuroscience and Psychology, University of Glasgow, UK<br>${ }^{3}$ Dept. of Psychology, University of Essex, UK


#### Abstract

The current study investigates whether mentalizing, or taking the perspective of your interlocutor, plays an essential role throughout a conversation or whether it is mostly used in reaction to misunderstandings. This study is the first to use a brain-imaging method, MEG, to answer this question. In a first phase of the experiment, MEG participants interacted "live" with a confederate who set naming precedents for certain pictures. In a later phase, these precedents were sometimes broken by a speaker who named the same picture in a different way. This could be done by the same speaker, who set the precedent, or by a different speaker. Source analysis of MEG data showed that in the 800 ms before the naming, when the picture was already on the screen, episodic memory and language areas were activated, but no mentalizing areas, suggesting that the speaker's naming intentions were not anticipated by the listener on the basis of shared experiences. Mentalizing areas only became activated after the same speaker had broken a precedent, which we interpret as a reaction to the violation of conversational pragmatics.


Keywords: Language; Pragmatics; Precedents; Common Ground; Conversation; MEG.

## Introduction

Humans have the special capacity to think about what others are thinking or feeling (Tomasello et al., 2005) and employ such "mentalizing" in everyday conversations. In the current study, we employ Magneto encephalography (MEG) to investigate the neural basis of mentalizing and how this process interacts with other brain areas during interactive language use. We focus on the critical question of whether mentalizing plays an immediate and constant role in conversation, or whether it mainly comes into play to enable interlocutors to detect and recover from misunderstandings.

In a conversation, it is important to know what is in "common ground" between you and your interlocutor. For example, you need to make sure that you are talking about the same thing, when referring to an object. One of the strategies interlocutors use in this situation is to establish conceptual pacts (Brennan \& Clark, 1996) or conversational precedents (Barr \& Keysar, 2002). This entails that, once they have agreed on (or "grounded", Clark \& Brennan, 1991) a certain referential expression for the object (e.g., 'salami' for a particular piece of meat), interlocutors in a certain conversation will continue referring to that same referent with the same term. Listeners generally expect speakers to adhere to this strategy. Thus, if a speaker
"breaks" the precedent and suddenly uses a different term (e.g., 'sausage') for the same referent, this is expected to confuse the listener; perhaps the speaker now refers to a different object? Such confusion, resulting in longer latency times to look at objects, has indeed been attested in eyetracking studies (Brown-Schmidt, 2009; Metzing \& Brennan, 2003; Kronmüller \& Barr, 2007). However, to test whether this confusion was not just related to "egocentric" processing, because this was the last name the listener heard for this object, a second speaker was introduced. This second speaker, unaware of the precedent that the first speaker had set, also breaks the precedent. This situation should be much less confusing to the listener as there has never been a common ground with the second speaker. All studies cited above consistently showed evidence of perspective-taking at some point after breaking the precedent; listeners were slower to look at the intended object when the same speaker broke a precedent than when a different speaker did so. However, the eye-tracking studies did not conclusively distinguish between two different theoretical accounts. First, a shared perspective could be maintained throughout the conversation. In that case, listeners can easily anticipate speaker's naming intentions, if the object and its name are in common ground. Thus, breaking a precedent would immediately lead to confusion, but only by the same speaker. This was supported by some research (Metzing \& Brennan, 2003), whereas other research suggests that, in first instance, breaking a precedent leads to confusion regardless of the speaker (Kronmüller \& Barr, 2007). This latter result implicates that listeners might not use common ground in the form of a shared perspective with the speaker by default, instead, only when it becomes necessary due to a pragmatic violation. According to this view, the speaker's referential intentions would not be anticipated based on the previously shared perspective (i.e., mentalizing), but rather, perspective-taking would only be engaged after a violation has taken place.

Using MEG within a "precedents" paradigm offers unique insights into whether mentalizing is used to interpret language throughout a conversation, or whether it is only engaged in reaction to pragmatic violations. Like eyetracking research, MEG has an excellent temporal resolution that allows effects to be localized in time, while also localizing these effects in the brain. We devised a paradigm that allowed for live interactive dialog between a participant in the MEG scanner with two different (confederate)
speakers. The speakers and the participant/listener viewed pictures of everyday objects on separate computer monitors (see Figure 1). The experiment was divided into a series of blocks, each consisting of an "interactive" phase in which precedents were established for various pictures, and a later "test" phase in which some of these pictures were named once again by either the same or a different speaker. For example, in the test phase, the listener might see a picture of a piece of meat (see Figure 1, right) that had been called 'salami' during the interactive phase, but now hears the (same or different) speaker say 'sausage'. Based on how the test-phase speaker named their object, listeners had to decide whether or not the speaker saw the same object as them. This task provided a cooperative reason why the (same) speaker might break the precedent, namely to signal that this was a different sausage than the salami that they saw together before. This "broken precedent" condition was contrasted with a baseline "no precedent" condition in which the picture named in the test phase had been referred to by its location (i.e., not named), during the interactive phase (see methods). Importantly, in the test phase, listeners viewed the picture for 800 milliseconds before the object was named. Also, the speaker for a particular block of testtrials was announced in advance. Thus, listeners could anticipate the referential expression based on the picture and the speaker, before the name was actually given.

We hypothesized that listeners would engage episodic memory areas (medial temporal lobe, e.g., Baddeley, 2000; lateral prefrontal cortex, Sakai \& Passingham, 2004; Kessler \& Kiefer, 2005) together with language areas (e.g., temporal pole, Imaizumi et al., 1997) in this anticipation period and possibly after naming. Especially objects that had been named during the interactive phase (in contrast to objects without a precedent) should engage these areas, since the picture would serve as a retrieval cue for retrieving the precedent. Episodic memory might be activated even more for objects that had been named by the same speaker, if speaker identity is used as a further retrieval cue.

For our research question, the involvement of mentalizing networks (i.e., temporo-parietal junction: TPJ, ventromedial prefrontal cortex: vmPFC, and possibly precuneus: PC; e.g., van Overwalle \& Baetens, 2009) and the timing of such involvement was of particular importance. The "anticipation" account would predict that episodic retrieval of common ground information while viewing the picture would lead to anticipation of the speaker's naming intentions, by employing mentalizing areas, already before naming. Crucially, such anticipation of naming intentions is only very meaningful when the current speaker has named this picture before. In contrast, the "egocentric" account would predict that, despite episodic retrieval of common ground information, mentalizing areas would not be employed to anticipate naming preferences based on these memories. Instead, this account would predict late, deliberate, post-naming activation of the mentalizing network, suggesting that it is only called upon to make sense of the experienced violation.

## Methods

## Participants

Seventeen British students from Glasgow University (8 males) with English as their native language participated in the MEG experiment, with approval of the local ethics committee. They were paid for their participation and gave their informed consent. One female participant was excluded from the analyses because she clicked the wrong picture too often in the interactive phase (22 times).

## Materials and Apparatus

Materials consisted of 320 experimental pictures that could be named in two roughly equivalent ways (based on a pilot study) and 640 filler pictures. The names for the test phase (320 experimental names and 190 filler names) were recorded beforehand, divided equally among the two confederates. Some of the filler names were presented with a hesitation, to make naming on-the-spot more plausible. The experimental names were always preceded by 800 ms of recording noise ( 600 to 1200 ms for the fillers).

MEG data were acquired using a 248 -channel 4DNeuroimaging magnetometer system, sampled at 508.63 Hz and band-pass filtered between 0.1 and 400 Hz .


Figure 1: Example of the participants' screen in the interactive phase (left) and the test phase (right).

## Procedure

Participants were first prepared for the MEG, including head digitization, in about 45 minutes. They were introduced to the confederates ("the speakers"), who would talk to them from separate rooms and were not able to hear each other. Participants had the role of listener. After two practice blocks in which the participants had the role of listener and speaker, 20 experimental blocks followed, all consisting of an interactive and a test phase, divided into 5 parts of 20 minutes each, with breaks in between. In the interactive phase, participants saw 9 pictures on the screen at a time (see Figure 1, left). In each interactive phase, one speaker asked them to click on one of these pictures, for a total of 42 times. Each critical picture was clicked on twice by the participants, using a trackball. Eight critical pictures were named and another eight were referred to by their location (e.g., 'top left'; participants were told that the speaker saw a question mark in place of these pictures). Participants could
freely interact with speakers in this phase. In the test phase, the speaker was the same as for the preceding interactive phase in half of the cases and different in the other half. Participants saw one picture at a time on the screen (see Figure 1, right) and were told that the speaker also saw one picture and had to name that picture. Participants had to indicate whether the speaker saw the same or a different picture than they, with a button press. The listeners were instructed not to interact with the speaker in this phase as they were unaware of hearing recorded utterances. Each phase consisted of 8 broken precedent trials (named differently than in the interactive phase), 8 no precedent trials (indicated by their location in the interactive phase), 4 maintained precedent fillers (named the same way in the interactive phase) and some new fillers.

## Data analysis

Pre-processing and statistical analysis of MEG data was conducted using the Fieldtrip Matlab® toolbox (Oostenveld, Fries, Maris, \& Schoffelen, 2011). We extracted epochs from 500 ms before the picture was shown until 500 ms after the response for all test trials. These epochs were detrended, denoised, and subjected to ICA to remove eye, heart, and movement artefacts. For evoked responses (ERF), trials of the same condition were averaged per participant, with a baseline of 200 ms prior to picture onset and a bandpass filter between 0.5 and 35 Hz . For these averages, planar gradient representations were calculated prior to sensor level analysis. For time-frequency representations, the power of frequencies between 2 and 30 Hz was calculated over time using a Hanning taper with a window of 4 cycles. For statistical analysis, we used the cluster-based approach implemented in the Fieldtrip toolbox (Maris \& Oostenveld, 2007), to circumvent the multiple-comparisons problem. We employed 2-step analyses for emulating the interaction between two factors. We first calculated a t-statistic for the difference between two conditions per participant and then included these $t$-values into a group statistic that compared a second difference. To identify sources underlying the sensor-level effects, individual single-shell head models were generated based on the individual MRI ( 6 mm voxel size) aligned with the MEG sensor array, subsequently normalized to a standard brain. A Linearly Constrained Minimum Variance (LCMV) beam former (van Veen and Buckley, 1988), common for all conditions (to increase SNR), was used for ERFs to transform individual conditions into source space for comparisons between conditions. Dynamic Imaging of Coherent Sources (DICS) beam formers (Gross et al., 2001) were used for theta source analysis. In this case we used condition-specific spatial filters to reveal qualitative differences between conditions.

## Results

## Behavioural results

As shown in Figure 2, the proportion of "different picture" responses was increased when a precedent was broken by
both same $(\mathrm{t}(15)=-5.63, \mathrm{p}<.001)$ and different speakers $(\mathrm{t}(15)=-3.97, \mathrm{p}=.001)$, but this was more pronounced for same speakers $(\mathrm{F}(1,15)=21.15, \mathrm{p}<.001)$. Participants were slower when an established precedent was broken, but only when the same speaker broke the precedent $(t(15)=3.47$, p $=.003$ ), resulting in an interaction $(\mathrm{F}(1,15)=8.43, \mathrm{p}=$ .011). Thus, listeners experienced greater confusion when a speaker broke his or her own precedent than when a speaker broke another's precedent (confirming Metzing \& Brennan, 2003; Kronmüller \& Barr, 2007). Note that this does not imply that common ground is considered by default or in anticipation, since it might also mean that listeners still experience conflict when a different speaker breaks the precedent, but resolves the conflict more quickly and/or in a different way than when the same speaker does so.


Figure 2: Behavioural responses in the test phase: proportion of "different" responses (left) and RTs (right).

These behavioural data confirmed our approach to look at the interaction effect in the MEG data (using the two-step analysis, see Data analysis), which we did for the theta source analysis (see Theta Oscillatory Results). Furthermore, on the basis of these behavioural data, it is most likely that mentalizing occurred for trials in which a precedent was broken by the same speaker and where participants responded to have seen a different picture. By this behavioural response, they show that they are aware of the conflict between the precedent and the new term and have resolved this by deciding that the speaker probably sees a different picture now. Next to that, especially slow responses, probably reflecting confusion upon hearing the new term, could also reflect the engagement of mentalizing processes. In contrast, in trials that elicited a quick "same picture" response, listeners might not even be aware of the conflict. Thus, we selected the "different"-responses plus one third of the slowest "same"-responses for the same speaker/broken precedent condition and refer to them as "deliberation trials". We used separate analyses of these trials as corroborative evidence when necessary (see ERF Results), next to analysing all trials of this condition.

## ERF Results

In a cluster analysis between 300 and 500 ms after naming (on the basis of visual inspection), a significant cluster was found for same speaker between 318-454 ms (p $=.009$ ) and a marginally significant cluster for different speaker between $300-415 \mathrm{~ms}(\mathrm{p}=.048)$ (see Figure 3, panel A). The topography of the effects was slightly more anterior (left) for the same speaker and more posterior (left) for the
different speaker. However, a direct comparison between the two contrasts did not reveal significant clusters. We also analysed the deliberation trials between -800 and 1000 ms . This revealed a cluster with a similar topography and timing of peak activity as the previous analysis ( $67-680 \mathrm{~ms}$ after naming, $\mathrm{p}<0.00001$, Figure 3, panel B right) plus two clusters in an early time interval before naming, when the target object was visually presented, ( 550 to 23 ms before naming onset, $\mathrm{p}=.004$, and 306 to 0 ms before naming onset, $\mathrm{p}=.004$; Figure 3, panel B left). This strongly suggests anticipatory processing in deliberation trials.


Figure 3: Sensor level ERF. A: comparisons for "broken vs. no precedent" for same speaker (left column) and different speaker (right column). B: comparisons for the deliberation trials (dotted dark blue line) compared to the no precedent trials (light blue line), within the same speaker condition. Significant clusters for this contrast are indicated pre-naming (left) and post-naming (right).

Source analysis was employed for the comparison between deliberation and no-precedent trials within the same speaker for both time intervals. The post-naming analysis between 300 and 500 ms revealed one significant, spatially distributed cluster ( $\mathrm{p}<0.00001$ ) and the prenaming analysis between -350 and -150 ms also revealed one significant, spatially distributed cluster, ( $\mathrm{p}<0.00001$ ). Table 1 lists the brain areas included in the pre- and postnaming clusters. See Bögels et al. (submitted), for figures and a more detailed description of these source-level results.

In both intervals, we found activation related to episodic memory processing, suggesting that participants were continuously retrieving the episodic context for a particular target object. Parahippocampal gyrus has been associated primarily with episodic encoding of the visuo-spatial context (e.g., Epstein \& Kanwisher, 1998), but more
recently also with integration of social, communicative, and paralinguistic context (e.g., Rankin et al., 2009). Thus, this activation might reflect retrieval of the episodic context of the interaction with the target object, including information about the speaker and the used name. We also found language-related areas, possibly indicating retrieval of the referent established during the interaction (in the prenaming interval, cf. Duff \& Brown-Schmidt, 2012) and semantic matching processes between the object and the used name (in the post-naming interval, e.g., Grabowski, Damasio et al., 2001; Pobric et al., 2007). We found differences in visual (attention) areas in the early interval which could suggest more visually detailed episodic retrieval of previously named objects, when anticipating a naming by the same speaker. In the late interval, we found activation of motor areas which could reflect more intense or more conflicting motor preparation. Anterior cingulate cortex activity was found in both intervals, suggesting an anticipation of conflict in the early and monitoring of conflict during the late interval. Most importantly for our research question, we found activation of so-called "mentalizing" areas only after naming. These areas have been found to be part of a mentalizing network, for example involved in social judgments (van Overwalle \& Baetens, 2009), in visuo-spatial perspective taking tasks (Blanke et al., 2005), and in reasoning about other's beliefs (Samson et al., 2004).

Table 1: Brain areas involved in the pre- and post-naming interval for ERFs in the deliberation trials vs. no precedent same speaker comparison ( $1 / \mathrm{r}$ : left and right hemispheres).

| Brain areas | Pre-naming | Post-naming |
| :--- | :--- | :--- |
| Episodic <br> memory | Parahippocampal <br> gyrus (l), <br> dorsolateral <br> prefrontal cortex (r) | Parahippocampal <br> gyrus (l) |
| Language | Temporal cortex (r) | Temporal pole (l) |
| Visual <br> (attention) | occipital cortex (r), <br> occipital temporal <br> cortex (r), parietal <br> occipital cortex (r) |  |
| Motor <br> (conflict) | Anterior cingulate <br> cortex (ACC) $(\mathrm{r})$ | ACC(l), <br> premotor cortex(l), <br> supplementary <br> motor area(l) |
| Mentalizing |  | TPJ (r), vmPFC (l/r) |

## Theta Oscillatory Results

Time-frequency analysis ( -800 to 1000 ms ) revealed a significant cluster ( $\mathrm{p}=.012$ ) in the theta range $(4-6 \mathrm{~Hz})$ for "same speaker, broken precedent" compared to "same speaker, no precedent" in a time window around 350-650 ms after naming onset. No results were found in the corresponding comparison for different speaker.

We localised the sources of this theta effect in a postnaming time-window ( $200-800 \mathrm{~ms}$ ) for 3 to 7 Hz , using a two-step analysis to look at the speaker by precedent
interaction ( $\mathrm{p}<.008$; see Bögels et al. (submitted) for more details and figures). In Figure 4, results are displayed of another two-step analysis within same speaker, comparing anticipatory and reactive intervals ( $\mathrm{p}<.01$ ).


Figure 4: Theta sources comparing same speaker (broken vs. no precedent) effects before and after naming.

Again, both analyses show episodic (working) memory areas (parahippocampal gyrus: blue circle in Figure 4; prefrontal cortex). Together with stronger activation of visual (attention) areas (occipital cortex: yellow circles in Figure 4) in the early than the late interval, this suggests stronger episodic retrieval in the right hemisphere along with stronger visual reactivation in response to the naming. In both analyses, we also see language (green circles in Figure 4) and motor areas (black circles in Figure 4) again, as in the ERF analysis. Importantly, mentalizing areas in right TPJ, right precuneus and vmPFC showed stronger theta effects for the same speaker contrast than the different speaker contrast, but also within the same speaker contrast in the post- vs. the pre-naming interval (pink circles, Figure 4).

## Discussion

Our results indicate that brain areas related to language, vision, episodic (working) memory, and mentalizing are dynamically and jointly involved in encountering and resolving conflict after a previously negotiated precedent is broken by the same speaker, and more so than when it is broken by a different speaker.

Episodic memory (together with language and vision) areas were engaged already in anticipation, suggesting a retrieval of the circumstances in which this picture was encountered before. Specifically, seeing a picture that was named before by the same speaker resulted in a stronger involvement of episodic memory (Theta localisation results), suggesting that the name was retrieved based on the picture and the identity of the speaker. We found especially strong anticipatory episodic memory activation for the trials in which listeners later on decided that the speaker saw a different picture ("deliberation trials", only for the time
domain). Successful retrieval of the name and speaker on the basis of the picture probably allowed listeners to notice the conflict with the actual name that was given by the speaker and decide that the speaker might see a different picture.

With regards to our main research question, we found (both in the time and the frequency domain) that mentalizing areas were clearly more engaged in response to the violation than in response to the other conditions (in accordance with Kronmüller \& Barr, 2007; Metzing \& Brennan, 2003). Crucially for our research question, however, mentalizing was not engaged more strongly in the anticipatory time interval in the condition in which the same speaker was going to name an object he or she had named before. Thus, while retrieval of the precedent and probably the speaker associated with that precedent took place on the basis of the picture, this did not lead to inferences of the current speaker's naming intentions using mentalizing. One might argue that listeners will try to infer the speaker's intentions in every condition. However, only in the case in which the same speaker has named this object before, does the listener really have grounds to use the speaker's perspective for this anticipation. Therefore, we argue that it would be expected, under the anticipation view, that mentalizing areas should be involved more strongly in precisely that condition, but this is not corroborated by our findings. Even when focussing only on "deliberation" trials (only for the time domain), where mentalizing was most likely to occur, since listeners show that they are aware of the conflict, we found no evidence for anticipation of the speaker's referential intentions (mentalizing).

These findings are in accordance with earlier eye-tracking results by Kronmüller and Barr (2007), showing that common ground is not taken into account by default since a broken precedent at first leads to confusion regardless of the speaker. In this context, recent approaches involving a second-person perspective (e.g., Schilbach et al., 2012) could also be of interest. A different processing "mode" and differential activation of brain areas is assumed for observation of others or "third person perspective" (as was used in most previous research) and for direct interaction with others, or "second person perspective". The latter type of processing (which might involve the posterior temporal sulcus; Tylén et al., 2012), related to fine temporal coordination during interaction, might be involved throughout a conversation, dealing for example with building up common ground by setting up new precedents. In the current study, all conditions probably involve such processes, resulting in no differential effects. In contrast, the pragmatic violations listeners encounter in the current study could invoke a mode that resembles the third person perspective, since listeners try to infer why the speaker breaks the precedent, involving the "classic" mentalizing areas (e.g., vmPFC; Tylén et al., 2012).

In conclusion, anticipating the speaker's referential intentions based on previously established common ground does not seem to be a default process. In contrast,
anticipation seems to rely only on episodic retrieval of visual and linguistic associations without any inference of the speaker's current mental states. Mentalizing about the other's perspective seems to be engaged "on demand" once a pragmatic violation or misunderstanding has occurred.

## Acknowledgments

SB was funded by NWO (Rubicon: 446-10-020) and ERC (269484 INTERACT), SG by ESRC/MRC (RES-060-250010), and KK by ESRC (RES-000-22-4325). We thank our two confederates, Louis Renn and Hellen McNair, and Laura Steenhuis for help with stimulus preparation.

## References

Baddeley, A. (2000). The episodic buffer: a new component of working memory? Trends in Cognitive Sciences, 4, 417-423.
Barr, D. J., \& Keysar, B. (2002). Anchoring comprehension in linguistic precedents. Journal of Memory and Language, 46, 391-418.
Blanke, O., Mohr, C., Michel, C. M., Pascual-Leone, A., Brugger, P., Seeck, M., \& Thut, G. (2005). Linking out-of-body experience and self processing to mental ownbody imagery at the temporoparietal junction. The Journal of Neuroscience, 25, 550-557.
Bögels, S., Barr, D., Garrod, S., \& Kessler, K. (submitted). Do I see what you mean? Mentalizing during conversational language processing as revealed by MEG.
Brennan, S. E., \& Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. Journal of Experimental Psychology: Learning, Memory, and Cognition, 22, 1482.
Brown-Schmidt, S. (2009). Partner-specific interpretation of maintained referential precedents during interactive dialog. Journal of Memory and Language, 61, 171-190.
Clark, H. H., \& Brennan, S. E. (1991). Grounding in communication. Perspectives on Socially Shared Cognition, 13, 127-149.
Duff, M. C., \& Brown-Schmidt, S. (2012). The hippocampus and the flexible use and processing of language. Frontiers in Human Neuroscience, 6.
Epstein, R., \& Kanwisher, N. (1998). A cortical representation of the local visual environment. Nature, 392, 598-601.
Grabowski, T. J., Damasio, H., Tranel, D., Ponto, L. L. B., Hichwa, R. D., \& Damasio, A. R. (2001). A role for left temporal pole in the retrieval of words for unique entities. Human Brain Mapping, 13, 199-212.
Groß, J., Kujala, J., Hämäläinen, M., Timmermann, L., Schnitzler, A., \& Salmelin, R. (2001). Dynamic imaging of coherent sources: studying neural interactions in the human brain. Proceedings of the National Academy of Sciences, 98, 694-699.
Imaizumi, S., Mori, K., Kiritani, S., Kawashima, R., Sugiura, M., Fukuda, H., \& Nakamura, K. (1997). Vocal identification of speaker and emotion activates different brain regions. Neuroreport, 8, 2809-2812.

Kessler, K., \& Kiefer, M. (2005). Disturbing visual working memory: electrophysiological evidence for a role of the prefrontal cortex in recovery from interference. Cerebral Cortex, 15, 1075-1087.
Kronmüller, E., \& Barr, D. J. (2007). Perspective-free pragmatics: Broken precedents and the recovery-frompreemption hypothesis. Journal of Memory and Language, 56, 436-455.
Maris, E., \& Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. Journal of Neuroscience Methods, 164, 177-190.
Metzing, C., \& Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. Journal of Memory and Language, 49, 201-213.
Oostenveld, R., Fries, P., Maris, E., \& Schoffelen, J. M. (2011). FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Computational intelligence and neuroscience.
Pobric, G., Jefferies, E., \& Ralph, M. A. L. (2007). Anterior temporal lobes mediate semantic representation: mimicking semantic dementia by using rTMS in normal participants. Proceedings of the National Academy of Sciences, 104, 20137-20141.
Rankin, K. P., Salazar, A., Gorno-Tempini, M. L., Sollberger, M., Wilson, S. M., Pavlic, D., \& Miller, B. L. (2009). Detecting sarcasm from paralinguistic cues: anatomic and cognitive correlates in neurodegenerative disease. NeuroImage, 47, 2005-2015.
Sakai, K., \& Passingham, R. E. (2004). Prefrontal selection and medial temporal lobe reactivation in retrieval of short-term verbal information. Cerebral Cortex, 14, 914921.

Samson, D., Apperly, I. A., Chiavarino, C., \& Humphreys, G. W. (2004). Left temporoparietal junction is necessary for representing someone else's belief. Nature neuroscience, 7, 499-500.
Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., \& Vogeley, K. (2012). Toward a second-person neuroscience. Behavioral and Brain Sciences.
Tomasello, M., Carpenter, M., Call, J., Behne, T., \& Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. Behavioral and Brain Sciences, 28, 675-690.
Tylén, K., Allen, M., Hunter, B. K., \& Roepstorff, A. (2012). Interaction vs. observation: distinctive modes of social cognition in human brain and behavior? A combined fMRI and eye-tracking study. Frontiers in human neuroscience, 6.
Van Overwalle, F., \& Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: a meta-analysis. Neuroimage, 48, 564-584.
Van Veen, B. D., \& Buckley, K. M. (1988). Beamforming: A versatile approach to spatial filtering. ASSP Magazine, IEEE, 5, 4-24.

# Discovering Processing Stages by combining EEG with Hidden Markov Models 

Jelmer P. Borst (jelmer@cmu.edu)<br>John R. Anderson (ja+@cmu.edu)<br>Dept. of Psychology, Carnegie Mellon University


#### Abstract

A new method is demonstrated for identifying processing stages in a task. Since the 1860 s cognitive scientists have used different methods to identify processing stages, usually based on reaction time ( RT ) differences between conditions. To overcome the limitations of RT-based methods we used Hidden Markov Models (HMMs) to analyze EEG data. The HMMs indicate for how many stages there is evidence in the data, and how the durations of these stages vary with experimental condition. This method was applied to an associative recognition task in which associative strength and target/foil type were manipulated. The HMM-EEG method identified six different processing stages for targets and repaired foils, whereas four similar stages were identified for new foils. The duration of the third, fifth and sixth stage varied with associative strength for targets and re-paired foils. We present an interpretation of the identified stages, and conclude that the method can provide valuable insight in human information processing.


Keywords: EEG; HMM, processing stages.

## Introduction

One of the main goals of cognitive science is to understand how humans perform tasks. To this end, scientists have long tried to identify different processing stages in human information processing. The first to do this in a systematic manner was probably Franciscus Donders. Almost 150 years ago, Donders proposed a method to measure the duration of cognitive stages (1868). By subtracting the RTs of two tasks that were hypothesized to share all but one processing stage, the duration of that stage could be calculated. A strong - and often problematic - assumption of Donders' subtractive method is the idea that it is possible to add an entire stage without changing the duration of other stages. To test whether different stages exist in the first place, Sternberg proposed the additive-factor method (1969). Although Sternberg overcame a limitation of Donders' method, the additive-factors method has its own drawbacks: it can only indicate the minimum number of stages in a task and it does not yield duration estimates of the stages. To improve on these inherent problems of RTbased methods and get better insight in stage existence and duration we propose a new method that uses HMMs (e.g., Rabiner, 1989) to analyze EEG data.

The basic idea of our method is to fit HMMs with different numbers of states to the EEG data (note that we use 'processing stages' and 'HMM states' interchangeably throughout the paper). The optimal number of states can then be determined by comparing the log-likelihoods of the fitted HMMs. Subsequently, the durations of the different states can be inspected, as well as how these durations vary
with condition. Using this information, and by comparing EEG signatures between states and experimental conditions, one can interpret the functional characteristics of the identified processing stages.

Our approach is based on a similar method that was used to analyze fMRI data (Anderson \& Fincham, in press; Anderson et al., 2010). For instance, Anderson and Fincham (in press) applied the method to mathematical problem solving, and discovered four stages: encoding the problems, planning a solution strategy, solving the problems, and entering a response. Although these results were promising, the temporal resolution of fMRI is severely limited, both by having scans that typically last one to two seconds and by the sluggish nature of the hemodynamic response. EEG, on the other hand, has a millisecond resolution, allowing for the discovery of processing stages in fast-paced tasks.

We applied the HMM-EEG analysis to an associative recognition task. During the study phase of this task, subjects were asked to learn word pairs. In a subsequent test phase - during which EEG data were collected - subjects were again presented with word pairs, which could be the same pairs as they learned previously (targets), rearranged pairs (re-paired foils), or pairs consisting of novel words (new foils). Subjects had to decide whether they had seen the pair during the study phase or not. Successful discrimination required remembering not only that the words were studied (item information), but also how the words were paired during study (associative information).

A conventional EEG analysis and a classifier analysis of this study were reported elsewhere (Borst et al., submitted). Currently, we are interested in finding out how many stages the subjects went through while determining a correct response.

## Methods

## Subjects

Twenty individuals from the Carnegie Mellon University community participated in a single 3-hr session for monetary compensation ( 9 males and 11 females, ages ranging from 18 to 40 years with a mean age of 26 years). All were right-handed and none reported a history of neurological impairment.

## Design

The experiment consisted of a study phase in which subjects learned word pairs and a test phase in which they were tested on these word pairs. In addition to probe type (targets, re-paired foils, or new foils), we manipulated word length and associative strength. Words could either be short (4 or 5
letters) or long (7 or 8 letters). Associative strength was manipulated by varying the number of word pairs a particular word occurred in. This is referred to as associative fan, and is known to have a strong effect on RT and accuracy (for a review, see Anderson \& Reder, 1999). Words in our experiment could have a fan of 1 or 2 , that is, they could occur in one or two word pairs. Both words in a word pair always had the same associative fan. New foils (foils consisting of words that were not presented in the study phase) always had an associative fan of 1 , they only appeared in a single word pair. Thus, there were 10 conditions: 2 (Probe: target or re-paired foil) $\times 2$ (Word Length: short or long) $\times 2($ Fan: 1 or 2$)+$ short and long new foils.

## Materials

Word pairs were constructed from a pool of 464 words selected from the MRC Psycholinguistic Database (Coltheart, 1981). Half of the words were nouns of 4 or 5 letters and composed the short word list. The other half of the words were nouns of 7 or 8 letters and composed the long word list. Word frequency and imageability ratings were matched between those lists. The 232 words of each length were divided randomly into two lists - a 24 -word study list and a 208 -word new foil list - such that the lists were matched on word frequency, imageability, and word length according to $t$-tests (all $p \mathrm{~s}>.1$ ).

The lists were used to create three sets of probes: targets, re-paired foils, and new foils. A set of 32 target word pairs was constructed from the study lists such that there were eight word pairs for each combination of length (short or long) and fan (1 or 2). Both words in short pairs were 4 or 5 letters and both words in long pairs were 7 or 8 letters. Each word in a fan 1 pair appeared only in that pair, whereas each word in a fan 2 pair appeared in two pairs. A corresponding set of 32 re-paired foil pairs was constructed in a similar manner by combining words from different target pairs of the appropriate length and fan. A set of 208 new foil word pairs was constructed from the new foil lists such that there were 104 word pairs for each length (all fan 1). The randomization of words and their assignment to conditions were unique for each subject.

## Procedure

The study phase started with each target word pair presented onscreen for 5000 ms , followed by a $500-\mathrm{ms}$ blank screen. Subjects were instructed to read each pair and make an initial effort to memorize it. Following target presentation, subjects completed a cued recall task designed to help them learn the word pairs. On each trial they were presented with a randomly selected target word and had to recall the word(s) paired with it (two-word responses were required for fan 2 words). The self-paced responses were typed and feedback (in the form of the correct response) was provided for 2500 ms following errors. If a target word elicited an error, it was presented again after all other target words had been presented. A block of trials concluded when all 48
target words had elicited a correct response. Subjects completed a total of three blocks of cued recall.

After the study phase, subjects entered the EEG recording chamber and completed the test phase. Each trial began with a centrally presented fixation cross for a duration sampled from a uniform distribution ranging from 400 to 600 ms . Following fixation, a probe word pair appeared onscreen (one word above the other) until the subject responded with a keypress to indicate whether the probe had been studied during the training phase. The probe was either a target, repaired foil, or new foil. Targets required "yes" responses (indicated by pressing the J key with the right index finger) and foils required "no" responses (indicated by pressing the K key with the right middle finger). Subjects were instructed to respond quickly and accurately. Following the response, accuracy feedback was displayed for 1000 ms , after which a blank screen appeared for 500 ms before the next trial began. Subjects completed a total of 13 blocks with 80 trials per block. All 10 conditions occurred equally often in random order in each block, resulting in 104 trials per condition during the test phase. Targets and re-paired foils were repeated during the test phase (they each appeared once per block), but each new foil appeared only once in the entire experiment.

## EEG recording

Subjects sat in an electromagnetically shielded chamber. Stimuli appeared on a CRT monitor placed behind radiofrequency shielded glass and set 60 cm from the subjects. The electroencephalogram was recorded from $32 \mathrm{Ag}-\mathrm{AgCl}$ sintered electrodes (10-20 system). Electrodes were also placed on the right and left mastoids. The right mastoid served as the reference electrode, and scalp recordings were algebraically re-referenced offline to the average of the right and left mastoids. The vertical electrooculogram (EOG) was recorded as the potential between electrodes placed above and below the left eye, and the horizontal EOG was recorded as the potential between electrodes placed at the external canthi. The EEG and EOG signals were amplified by a Neuroscan bioamplification system with a bandpass of 0.1 to 70.0 Hz and were digitized at 250 Hz . Electrode impedances were kept below $5 \mathrm{k} \Omega$.

## EEG preprocessing

Recording artifacts in the EEG data were removed based on visual inspection. Following artifact rejection, the data were decomposed into independent components. Components associated with eye blinks were visually identified and projected out of the EEG recordings. A $0.5-30 \mathrm{~Hz}$ band-pass filter was applied to attenuate high-frequency noise. Trials were extracted from the continuous recording and baselinecorrected using a linear baseline, such that the 200 ms before stimulus onset and $80-160 \mathrm{~ms}$ after the response were on average 0 (visual inspection showed no condition difference at this interval after the trial). Incomplete trials due to artifact rejection were excluded, as well as trials containing voltages above $+75 \mu \mathrm{~V}$ or below $-75 \mu \mathrm{~V}$. In
addition, all incorrect trials and correct trials with RTs exceeding three standard deviations (SDs) from the mean per condition per subject were removed. For the HMM-EEG analysis we also removed trials with RTs longer than 3000 ms . In total, $16.1 \%$ of the trials was excluded.

For efficiency, the EEG data were down-sampled to 50 Hz . Every four samples were then combined into a single 'super-sample', by quadrupling the number of channels. That is, from four $20-\mathrm{ms}$ samples with 32 channels we created one $80-\mathrm{ms}$ super-sample with 128 channels. A super-sample contained information about the mean voltage in each channel, as well as about whether this voltage increased or decreased over the 80 ms interval. Next, we normalized each channel to a mean of 0 and a SD of 1 , and applied a principle component analysis (PCA) to the 128 channels. The results of the PCA were again normalized; the first 20 PCA components were used for the HMM-EEG analysis.

## The HMM-EEG Analysis

The HMM-EEG analysis consists of two main parts: (1) determining the optimal number of states and (2) computing the properties of the identified states. Both parts of the analysis depend on fitting HMMs to the preprocessed EEG data. We will therefore first discuss the structure and parameter estimation procedure that was used for the HMMs. We then explain how these HMMs were used to find the optimal number of states and how we computed the properties of these states.

## HMM structure and parameter estimation

An HMM simulates a system that is at any given time in one of a set of distinct states, between which it transitions at certain times (e.g., Rabiner, 1989). In our analysis, each state represents a processing stage in the task (e.g., encoding the stimulus, executing a response). A state is associated with a brain signature $M_{i}$ that represents the average EEGactivation pattern during this processing stage, and with a gamma distribution $G_{i}$ that represents the state's durations over the trials in the experiment. For current purposes, we only consider HMMs with a linear structure, that is, state 1 always transitions to state 2 , state 2 to state 3 , etc.

An example of a four-state HMM is shown in Figure 1. At the top of the figure EEG data is shown for three channels over three trials of the experiment, at the bottom the HMM with associated brain signatures and gamma distributions.


Figure 1. Overview of the HMM-EEG analysis. EEG data comes in at the top and is preprocessed into PCA components. At the bottom a fitted 4-state HMM is shown, with state signatures and gamma distributions. The center graph shows the probability of each sample $j$ for each state given this HMM. The connections between sample likelihoods and states are shown for states 1 and 3.

HMM algorithms can be used to find parameters $M_{i}$ and $G_{i}$ that yield the optimal interpretation of the data given an HMM with $r$ states (see Anderson \& Fincham, in press, for a more detailed explanation for the kind of HMMs that are used in this paper; Rabiner, 1989; Yu \& Kobayashi, 2006). To calculate the solutions we adapted software that minimizes the summed log-likelihood of the HMM over all trials (Yu \& Kobayashi, 2006).

Figure 1 shows the result of such an optimization procedure for a 4 -state HMM. Given the optimal state signatures and gamma distributions, the probability that each sample $j$ belongs to a state is depicted in the center of the figure. As expected, the first samples in each of the three trials probably belong to state 1 (blue), the next samples to state 2 (green), etc. In addition, state 1 is always two samples long in the three trials in the figure, matching the gamma distribution of this stage. State 3, on the other hand, is much more variable in duration.

For clarity the explanation above assumes a gamma distribution for each state. In the actual analysis we used separate gamma distributions for each condition and state, allowing for different duration estimates per condition.


Figure 2. State signatures and gamma distributions.

## Number of states and state properties

Above we explained how an $r$-state HMM can be determined that gives an optimal interpretation of the data. However, what we are really interested in is finding the optimal number of states to describe the data. A simplistic approach would be to compare the log-likelihoods of HMMs with different numbers of states. However, because HMMs with more states have more parameters to fit the data, they will typically yield a better fit. What we want to know is if the extra parameters explain enough extra variance to be warranted. To this end we applied leave-oneout cross validation (LOOCV).

Our LOOCV method estimated state signatures for $n-1$ subjects, and calculated the log-likelihood of the $n^{\text {th }}$ subject given these signatures while allowing for different state durations for the $n^{\text {th }}$ subject (to accommodate speed differences between subjects, unlike Anderson \& Fincham, in press). This process was repeated for all subjects.

The LOOCV procedure was repeated for HMMs with different numbers of states. To select the best model we used a sign-test: if a $k$-state model fitted the data of $x$ out of $n$ participants better than all $(l<k)$-state models we choose it as the winner. The underlying idea is that while a $(k+1)$ state model will fit the data of $n-1$ subjects better in the estimation phase than a $k$-state model, it is at least as likely to fit the $n^{\text {th }}$ subject worse (Anderson \& Fincham, in press). According to a sign-test, a significant increase is reached when 15 out of 20 subjects improve ( $p=.04$ ).

After determining the optimal number of states, we computed the properties of the identified states. First, we estimated an optimal HMM on the data of all subjects. We used the state signatures of this model to estimate optimal gamma distributions for each subject. These gamma distributions were used to calculate the average state duration for each subject and condition, which were used in subsequent ANOVAs to determine which states change in duration with condition. In addition, the subject-specific
models give us the probability for each sample in the data to be in a certain state (center of Figure 1). This was used to calculate differences in EEG activation between conditions.

## Results

For reasons of brevity we do not report behavioral results separately. RTs can be inferred from Figure 3. For targets and re-paired foils, $\operatorname{Fan}(F(1,19)=65.42, p<.001)$, Probe $(F(1,19)=45.10, p<.001)$, and the interaction between Fan and Probe $(F(1,19)=31.40, p<.001)$ had a significant effect on RT, as indicated by a repeated measure ANOVA. In addition, new foils were responded to much faster than the other probe types, which was expected given that no associative information has to be retrieved for new foils.

## Number of stages

Because new foils are very different from the other probe types - no associative information has to be retrieved for new foils - we decided to run separate analyses for new foils and targets/re-paired foils. For targets and re-paired foils a 6 -state HMM turned out to be the winner. It was better for at least 16 subjects than HMMs with fewer states, and no HMM with more states had a higher log-likelihood for more than 9 subjects. The new foils also showed evidence for 6 states: 17 subjects fitted better with a 6 -state HMM than with 4 states. However, the 4 -state solution compares better to the 6 -state solution of targets and repaired foils. ${ }^{1}$ Although there might be more stages in the data, we can be secure in the assumptions that there are at least 6 states for targets and re-paired foils and 4 states for new foils and in whatever conclusions these assumptions lead to. Thus, we will focus on the 4 -state solution for new foils in this paper.

## Stage properties

Figure 2 shows the gamma distributions and state signatures of the 6 -state HMM for the targets/re-paired foils and the 4 state HMM for the new foils. Interestingly, the first two states of both solutions seem very similar. Correlations between the state signatures confirm that stage 1 and stage 2 in both HMMs resemble each other closely: 0.98 and 0.97 .

The estimated gamma distributions in Figure 2b are averaged over conditions and subjects. They show that stage 1 has a very fixed duration, of two samples or 160 ms . The other stages are more variable. A duration of 0 means that the state is skipped, which happens most often (in $50 \%$ of the trials) for stage 4 of the targets/re-paired foils. For the other stages these percentages are under $30 \%$.

Figure 3 shows the state durations in more detail, split out for conditions. We will only list major effects with $p$-values $<.01$ (repeated measure ANOVAs), as these are used below to interpret the results.

[^22]

Figure 3. State durations. A shows how the state durations add up to form a complete trial, whereas B illustrates the effect of condition on state duration.

State 1 and 2 seem stable over the different conditions, even between the two different HMMs. This matches the observation that their state signatures are very similar. Stage 3 is longer for fan 2 items than for fan $1(F(1,19)=15.14, p$ $<.001$ ). Stage 4 seems to be an intermediate stage that is often skipped for the targets/re-paired foils, and it does not change with condition. For the new foils, stage 4 is the final stage. It does not change in duration with word length. Stage 5 varied strongly in duration with both Fan $(F(1,19)=$ 16.12, $p<.001$ ) and $\operatorname{Probe}(F(1,19)=20.32, p<.001)$. Stage 6, the final stage for targets/re-paired foils, is longer for fan 2 items than for fan 1 items $(F(1,19)=21.55, p<$ .001). In addition, there is an interaction between Fan and Probe $(F(1,19)=16.75, p<.001)$, with the fan effect being stronger for re-paired foils than for targets.

The HMM-EEG analysis aims to find states with similar brain signatures in the different conditions of the experiment. Although that is the case, there might still be differences between conditions within a stage. Figure 4A shows the differences between conditions for the 6 -state HMM for targets/re-paired foils; Figure 4B for the 4 -state HMM for new foils. These differences were calculated by estimating brain activity for each state, condition, and subject. The resulting values were subjected to t-tests for each electrode.


Figure 4. Differences between conditions in states for (A) targets/repaired foils and (B) new foils. The maps show $t$ values for FDR-corrected $p$-values $<.05$.

Figure 4A shows that long words resulted in less activity than short words in state 1 over left prefrontal electrodes, and in state 6 over central electrodes. Fan 2 items resulted in more activity than fan 1 items in states 3 and 4 over midline electrodes, whereas they showed less activity in state 6 over parietal regions. Finally, targets elicited a little less activity than re-paired foils in state 3 , and more activity in states 5 and 6 over parietal and occipital sites. The largest effect for new foils was in state 2 , where long words resulted in less activity than short words for frontal electrodes.

## Interpretation of the Processing Stages

The underlying reason for wanting to identify processing stages is explaining how tasks are performed. In this section we will give our interpretation of the processing stages discovered by the HMM-EEG method.

The first two stages seem to reflect visually perceiving the two words on the screen. Both stages hardly varied with condition, and are very similar between targets/re-paired foils and new foils - implying that the words are not processed yet in relation to the experimental task. In addition, there are effects of word length on brain activity in stage 1 for targets/re-paired foils and in stage 2 for new foils. Although word length effects are typically strongest in occipital regions, Hauk et al. (2009) showed a left prefrontal effect that appears to match our observation.

We hypothesize that stage 3 reflects item retrieval, to determine whether the presented words were learned during the study phase of the experiment. First, the duration of stage 3 varies strongly with fan and there is also a strong effect of fan on brain activity in stage 3. Existing models of the fan effect assume that the effect originates in declarative memory, implying that this stage is memory related (e.g., Anderson \& Reder, 1999). Second, for new foils this is the stage where information has to be retrieved about whether the words were studied or not. After the third stage there is only a short response stage, which is similar to stage 6 of the targets/re-paired foils. Given the matching time course, we assumed that stage 3 reflects an item retrieval stage for targets/re-paired foils as well.

The idea of an early item retrieval stage and a later associative retrieval stage (stage 5) resembles dual-process theories of recognition (e.g., Rugg \& Curran, 2007). To judge whether a stimulus was experienced before, dualprocess theories assume an early 'familiarity' process, followed by a functionally distinct recollection process. With respect to our experiment, the familiarity phase could correspond to stage 3 - in which it is determined whether the items are familiar - whereas stage 5 could correspond to the recollection stage in which associative information is retrieved.

Familiarity and recollection processes have been related to different ERP components (Rugg \& Curran, 2007). Familiarity elicits a negative response between $300-500 \mathrm{~ms}$ over mid-frontal electrodes, with new items being more negative than studied items. This matches the observation that new foils in our experiment have a more negative brain signature over mid-frontal electrodes than targets/re-paired foils in stage 3. Recollection has been linked to the parietal old/new effect, which is more positive for old than for new items. If our stage 5 reflects recollection of associative information, it should show a parietal positivity for targets versus re-paired foils, which it does.

Stage 4 is skipped in $50 \%$ of the trials. We tentatively hypothesize that it reflects working memory consolidation of the items that are retrieved from memory in stage 3. This is not a necessary process, which might explain why it is skipped in $50 \%$ of the trials.

As explained above, we assume that stage 5 reflects associative retrievals. Not only does it show the parietal old/new effect, but it also varied in duration both with fan and probe type, which are known to influence the length of associative retrievals.

Stage 6 of the targets/re-paired foils and stage 4 of the new foils are the final stages in the task. We assume that they reflect response stages. The duration of stage 6 changes with fan, and shows an interaction between fan and probe type. For new foils this last stage is shorter than for the other conditions. We interpreted these duration differences as an effect of response confidence. Subjects responded faster and more accurate to new foils than to targets/repaired foils, and faster and more accurate to fan 1 items than to fan 2 items - indicating they might have been more confident in those responses.

The effects on brain activity support this interpretation. There were differences over parietal electrodes between targets and re-paired foils (targets being more positive), between fan 1 and 2 items (fan 1 items being more positive), and between new foils and targets/re-paired foils (the signature of new foils is slightly more positive). We hypothesize that these effects resemble a P300, which is known to increase with response confidence (Wilkinson \& Seales, 1978).

## Discussion

In this paper we have presented a new method for identifying processing stages in a task, which uses HMMs to
analyze EEG data. For the associative recognition task, the method yielded a 6 -state solution for targets and re-paired foils and a 4 -state solution for new foils. These solutions seem to be reasonable, and could be interpreted by using information about how the stages varied in length, and how the brain activity within stages differed between conditions. The results matched dual-process theories of recognition, both in expected stage duration and brain activity.

Naturally, other interpretations of these results are also conceivable. For instance, the duration of the last two stages could be explained with an accumulator model, which samples faster for the easier conditions.

That being said, especially in combination with earlier promising effects on fMRI data (e.g., Anderson \& Fincham, in press; Anderson et al., 2010), we think that the HMMEEG method shows great promise for investigating human information processing.

## References

Anderson, J. R., Betts, S., Ferris, J. L., \& Fincham, J. M. (2010). Neural imaging to track mental states while using an intelligent tutoring system. PNAS USA, 107(15), 7018-7023. doi:10.1073/pnas. 1000942107
Anderson, J. R., \& Fincham, J. M. (in press). Uncovering the Sequential Structure of Thought. Cognitive Science.
Anderson, J. R., \& Reder, L. M. (1999). The fan effect: New results and new theories. Journal of Experimental Psychology General, 128(2), 186-197.
Borst, J. P., Schneider, D. W., Walsh, M. M., \& Anderson, J. R. (submitted). Journal of Cognitive Neuroscience.

Coltheart, M. (1981). The MRC psycholinguistic database. Quarterly Journal of Experimental Psychology, 33A, 497-505.
Donders, F. C. (1868). De snelheid van psychische processen (On the speed of mental processes).
Hauk, O., Pulvermüller, F., Ford, M., Marslen-Wilson, W. D., \& Davis, M. H. (2009). Can I have a quick word? Early electrophysiological manifestations of psycholinguistic processes revealed by event-related regression analysis of the EEG. Biological psychology, 80(1), 64-74.
Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2), 257-286.
Rugg, M. D., \& Curran, T. (2007). Event-related potentials and recognition memory. Trends in Cognitive Sciences, 11(6), 251-257.
Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. Acta psychologica, 30, 276-315.
Wilkinson, R. T., \& Seales, D. M. (1978). EEG eventrelated potentials and signal detection. Biological psychology, 7(1), 13-28
Yu, S. Z., \& Kobayashi, H. (2006). Practical implementation of an efficient forward-backward algorithm for an explicit-duration hidden Markov model. IEEE Transactions on Signal Processing, 54, 1947-51.

# Pragmatic priming and the search for alternatives 

Lewis Bott (Bottla@Cardiff.ac.uk)<br>School of Psychology, Cardiff University<br>Park Place, CF10 3AT, Cardiff, UK

Emmanuel Chemla (Chemla@ens.fr)<br>Ecole Normale Supérieure, LSCP<br>29 rue d'Ulm, Paris, France


#### Abstract

Meanings of basic expressions can be enriched by considering what the speaker could have said, but chose not to, that is, the alternatives. We report three experiments testing whether there is a single enrichment procedure that stretches across diverse linguistic phenomena. Participants were primed to understand either the basic meaning or the enriched meaning of a sentence. We found that the enrichment mechanism could be primed across some expressions but not others, arguing against a universal enrichment mechanism. Our results have implications for understanding the processing of implied meaning and how linguistic phenomenon should be grouped together.


Keywords: Pragmatics; psycholinguistics; alternatives; implications; structural priming.

Interpreting a sentence requires taking into consideration the combination of words that have been uttered, but also words that have not been uttered, that is, the alternatives. For instance, if a speaker says, "John has read some of the books", the listener can use the alternatives to derive the meaning that John has read some, but not all, of the books. The derivation procedure would be something like (i) accept that John read some (or all) of the books; (ii) identify "John read all the books" as a relevant alternative; and (iii) select this alternative to negate. In this paper we consider how the processor derives the alternatives. Our approach is to apply a structural priming technique (e.g., Raffray \& Pickering, 2010) to test whether the search for alternatives can be primed across and within different inferences.
Enrichment by negation of alternatives is a large phenomenon. There is always some phrase or item that may generate an alternative, and furthermore, there is an infinite range of potentially relevant alternatives. Consider the example above again. In the right context, "John has read some of the books" could imply that Bill/Helen/Mary etc. have not read some of the books, or that John has read the books but not seen the films, or even that John has not written some of the books. The wide variety of possible alternatives raises a serious processing question, however: How does the processor know which alternatives to negate? There have been several theories in the linguistics literature that provide partial answers to this question. Horn (1972) suggested that certain expressions are grouped together in the lexicon to form semantic scales, and the alternatives for a given item are its scale mates. With some for example, some, many, and all form a semantic scale, and the
alternatives for some would be many and all. More liberally, Rooth's (1992) work on focus suggests that alternatives can be any item in the same semantic category as the target (type <e,t> etc.). Intermediately, Katzir (2007) has suggested that alternatives are any items that are less than or equally complex than the trigger, or that are particularly salient in the context. In our study we take a slightly different approach, however. Rather than identifying a set of structural principles for defining the alternatives, we ask whether there is a single procedure that enriches the basic meaning in different linguistic contexts. In doing so, we also seek to find evidence that the seemingly diverse linguistic phenomena share a common root in how they are derived.
Table 1 shows the set of phenomena that we used in our experiments. All of them involve a basic meaning that can then be enriched by negating alternatives. The first column refers to the name of the phenomenon (the expression), the second to the basic (or weak) meaning, the third to a plausible alternative, and the fourth to the result of enriching the basic meaning with the alternative.

| Expression | Semantic | Alternative | Result |
| :--- | :--- | :--- | :--- |
| Some | some or all | All | some but not all |
| Number $n$ | at least $n$ | Number $n+1$ | $n$ but not $n+l=$ <br> exactly $n$ |
| Plural <br> morphology | vacuous | singular <br> Ad horphology | not singular <br> =plural |
|  | There is an <br> A | There is an <br> A and a B | There is an A <br> and not a B. |

Table 1. Experimental phenomena. The semantic form of each expression can be enriched using the negation of alternatives.

The table summarizes the following cases:
(i) some, which trigger the archetypical scalar implicature and for which there exist arguments for the alternatives to be stored in the lexicon (see Horn, 1984, or Levinson, 2000).
(ii) Numbers, which are claimed by some authors to operate in a similar manner to the some cases (e.g., Horn, 1989, van Rooy \& Schulz, 2006) but not by others (e.g., Breheny, 2008). According to the former group, when a speaker says, e.g., "Dave has three children", the weak meaning of the expression, "three," is at least three, but this meaning can be enriched to negate the alternatives, (at least)
four, (at least) five, etc. to form the exactly three reading.
(iii) plural morphology, which have been contentiously linked to the some case, by providing arguments showing that plural morphemes are semantically vacuous, surprisingly, and that the plural reading is obtained via negation of the singular alternatives (e.g., Spector 2007).
(iv) ad hoc implicatures, for which the alternatives are only specifiable given an appropriate context. For example, if a speaker says, "There is an elephant," and the context suggests that it would have been relevant for him to say, "There is an elephant and a lion," the listener is licensed to infer the alternative is not true, suggesting that the speaker meant that there was an elephant but no lion.

In summary, the phenomena are of diverse kinds: the root alternative trigger may be lexical (some and numerals), morphological (plurals) or contextual (ad hoc), and the motivation for these claims may be more or less intuitive and debated, as we described above. They are nonetheless similar in a way that is important for our experiments. The enrichments shown in the Results column are optional. In each case, the listener must derive the basic meaning, but then has a choice about whether to enrich the statement and interpret the meaning with the negated alternatives.

Our experiments test whether the enriched meanings shown in Table 1 are all computed by a single, universal mechanism, or whether separate, individual procedures are applied in each case. There is good reason to suppose either of these possibilities might be true. First, in favour of a universal mechanism, all of the cases shown in Table 1 are arguably derived using the same negation-of-alternative procedures. Grice's Quantity maxim, for instance, could be invoked to generate reasoning along the lines of, "well, if the speaker had meant [alternative], they would have said so," (see Chierchia, 2004, for a wider range of views) and the linguistic contexts that give rise to the enrichment are similar across phenomena. The most simple processing view would be that if these phenomena can be grouped together linguistically, then they should share similar processing mechanisms. Conversely, there are also arguments for different enrichment procedures for different phenomena. For example, numbers may behave differently to quantifiers (see e.g., Breheny, 2008), and Katzir's (2007) theory distinguishes the alternatives involved in ad hoc implicatures from those involved in the some and number cases (as we describe in more detail in Experiment 2). Furthermore, different types of expressions vary in how frequently they are enriched. For example, Zevakhina \& Geurts (2011), show that adjectives in scalar implicatures, such as, "John's cake is ok" (implying John's cake is not delicious) are less likely to undergo enrichment than quantifiers, such as "Some of John's cakes were eaten" (implying not all of John's cakes were eaten). The variation in enrichment could be because different enrichment mechanisms are involved across different cases.

## Overview of Experiments

To test between universal and individual enrichment procedures, we used a structural priming paradigm (see, e.g., Pickering \& Ferreira, 2008). We reasoned that if there were a universal mechanism, it should be possible to prime the procedure, that is, make it more likely that the enriched meaning would be derived. Priming across different phenomena would provide support for a universal procedure for searching alternatives, but priming restricted to particular expressions would support individual enrichment procedures.

There are many different versions of structural priming but we modeled our experiment on Raffray and Pickering (2010), who tested priming of scopal relations. In their experiments, participants had to match one of two pictures to a sentence. The sentence always involved every and $a$, but let the scopal relation free to create an ambiguous sentence, as in Every child climbed a hill (which can be interpreted either as there being a single hill that every child climbed, or multiple hills where every child climbed a separate hill). In the prime trials, the pictures were consistent with only one reading of the sentence. For example, for the Every child climbed a tree sentence, one of the pictures was of a single hill with multiple children climbing it, and the other picture was of something unrelated like cows in a field. In the trial that immediately followed it, the probe trial, participants saw a different every sentence and a further two pictures. One picture was consistent with wide scope reading and one with the narrow scope reading. Participants chose which interpretation best matched the sentence. Raffray and Pickering hypothesized that wide scope prime trials would prime a wide-scope reading in the probe trials and vice versa, which is exactly what they found.

Our experiments were very similar except that we used sentences that involved the constructions shown in Table 1, rather than every. Just like the every sentences, our sentences were ambiguous because implicatures are optional. We hoped to be able to prime whether participants interpreted the sentence with or without the enriched meaning. If either a universal enrichment or an individual enrichment procedure can be primed, we would expect priming within each of the expressions in Table 1. More interestingly, if there is a universal enrichment process, we should observe priming across the different expressions.

## Experiment 1

Participants saw a sentence and had to match the sentence with one of two pictures. All of the sentences referred to the presence of letters in a set, such as "All of the letters are As." In the experimental trials, the sentences invited enrichment, as shown in Table 1. However, because the enrichment was optional, participants could choose to interpret the sentence in its basic form. This meant that the sentences could have either a weak meaning (without enrichment) or a strong meaning (with enrichment). For a
given sentence, three types of pictures were possible: (a) false pictures, that made both readings false, (b) weak pictures, that made the weak reading true but the strong reading false, and (c) strong pictures that made both readings true.

There were two types of prime trials. First, weak primes, which displayed a false picture and a weak picture, so that participants would click on the weak picture and access the weak reading. Second, strong primes, which displayed a weak picture and a strong picture. We reasoned that participants would access the strong reading (the one that makes the two pictures different in a relevant way) and click on the strong picture. An example of the weak some prime is shown in the upper-panel of Figure 1.


Figure 1. Weak some prime (upper-panel), and some probe (lower-panel).

In the probe trials, participants read another experimental sentence and saw two more pictures. One of the pictures was a weak picture, and the other picture was a box with "Better Picture?" written inside it. Participants were instructed that the "Better Picture" option should be selected if they did not feel that the other picture sufficiently captured the sentence meaning. The lower panel of Figure 1 shows the probe trials. We expected that participants should click on the weak picture if they accessed the weak reading, and opt for the "Better Picture" option if they accessed the strong reading. Probe trials immediately followed prime trials. Consequently, priming of the enriched meaning would be observed when a participant selected the weak interpretation option more often after the weak prime than after the strong prime (and vice versa).

In Experiment 1 we used the first three expressions shown in Table 1: (1) some sentences, for which the weak interpretation picture was a box in which pictures were all one type of letter, namely As, and the strong interpretation picture involved a set filled partly with A's and partly with B's. (2) Number sentences, such as "Three of the letters are As," where the target enriched meaning was no more than 3. The weak picture was a box in which 6 letters were A's, and the strong picture was a box in which exactly 3 of the letters were A's. (3) Plural sentences, such as, "There are As," in
which the target enriched meaning was there is more than one $A$. The weak picture involved a single A, whereas the strong picture involved multiple A's.

If (a) all three expressions are related, (b) involve alternatives, and (c) there is a universal enrichment procedure, then priming should be observed within and between the three expressions. If the enrichment mechanism depends on specific structures, priming should be observed only within each expression. We tested within-expression priming by presenting sequences of trials in which the prime from one phenomenon was followed by a probe from the same phenomenon. For example, a some prime, such as "Some of the letters are As" would be followed by a some probe, such as "Some of the letters are Bs". We tested between-expression priming by presenting the prime from one expression followed by the probe from a different expression. For example a plural prime, such as "There is an A" might be followed by a some probe, such as, "Some of the letters are Bs".

## Method

Participants. In each experiment reported in this paper, we used 50 participants, all recruited online using Amazon Turk and all claiming to be native speakers of English. A different set of participants was used for each experiment. They were paid for their participation.

Design and materials. All trials were either probe trials, prime trials, or bias trials. An experimental sequence of trials was two prime trials followed by one probe trial, i.e., prime-prime-probe. We thought the effect of the prime would be greater if the prime trial was doubled. The prime and probe trials were completely crossed so that each participant saw prime-probe sequences of all possible combinations, e.g., some (weak)-number; number (strong)some; some (weak)-plural. This meant that for each expression, there were 6 possible sequences ( 2 withinexpression trials, and 4 between-expression trials), and hence 18 sequences in total. We replicated this set 4 times and so there were 72 probe trials and 144 prime trials.

We also added bias trials to encourage participants to (a) select the "better picture" box, and (b) consider appropriate alternatives to the experimental sentences. For example, we included all trials so that participants would realise the speaker sometimes said all instead of some. There were 12 bias trials per set, and 48 in total.

Sentences were all statements about letters, as shown in Figure 1. The particular letters were randomly chosen for each experimental sequence. Each experimental sequence was presented in a random order.

In the prime trials the expected answer (weak or strong) was on the right for half the trials. In the probe trials, the "better picture" box was always on the right.

## Results and Discussion

Figure 2 shows the proportion strong interpretations during probe trials. The within-expression effect is shown on the left, and the between-expression effect is shown in
the right. A large within-expression priming effect can be seen by the difference between the weak and strong primes, such that there were more strong interpretations after the strong primes than after the weak primes, $F(1,46)=63.25, p$ $<.001$. There was no significant difference in the size of the effect across expressions, however, $F<1$. More interesting was the priming effect between expressions. The betweenpriming effect was marginally significant using an ANOVA with probe type and interpretation (weak vs strong), $F(1,50)$ $=3.063, \mathrm{p}=.086$, and fully significant using a nonparametric bootstrapping test, $p<.05$ (we also replicate a similar between-expression priming effect in Experiments 2 and 3).


Figure 2. Experiment 1 results.
Our findings suggest that for these sorts of linguistic expressions, there is a universal mechanism that enriches the basic meaning. If the mechanism were tied to individual phenomena, we would not have observed betweenexpression priming. In Experiment 2, we consider the enrichment mechanism in more detail.

## Experiment 2

The enrichment mechanism involves two procedures: (1) identifying the appropriate alternatives, and (2) negating them. The results of Experiment 1 are consistent with either (or both) of these: we could have primed the search for alternatives, or the procedure that negates them. To investigate this in more detail we tested the ad hoc expressions shown in Table 1. The ad hoc expressions are similar to those used in Experiment 1 in that they involve negation of alternatives (see Hirschberg, 1991). However, they are different in that the alternatives for the ad hoc expressions must be determined with reference to the context, whereas the alternatives for the other expressions can be determined lexically (they are context-free). It follows that if we were to observe priming within the ad hoc expressions, but not between the ad hoc expressions and the lexical expressions, we could conclude that the effects of Experiment 1 were at least partly due to priming of the search for alternatives and not priming of the negation process. We would also conclude that there were separate processes computing the alternatives for the ad hoc expressions compared to the lexical expressions. Of course, if we found priming across all of the expressions, as we did
in Experiment 1, we could only draw conclusions about the general enrichment process, not the negation of the alternatives.
In Experiment 2, we introduced ad hoc sentences into the priming design from Experiment 1. The ad hoc sentences were sentences like There is an $A$, which, given the visual context, invited enrichments like There is an $A$ but not a $B$. The weak ad hoc prime and an ad hoc probe are shown in Figure 3. The design was exactly the same as Experiment 1 except that we replaced the plural expressions with the ad hoc expressions, which meant that we had some, number, and ad hoc expressions.


Figure 3. Weak ad hoc prime (upper-panel), and ad hoc probe (lower panel).

## Results and Discussion

Proportions of strong responses to the probe are shown in Figure 4. The upper panel shows within-expression priming. For each expression, there were more strong interpretations after the strong prime than after the weak prime, all $t$ 's(49) $>4.11, p$ 's $<.001$, but there was a marginal interaction between probe expression and interpretation, $p=.087$, suggesting less within-expression priming in the ad hoc expressions. The lower-panel shows between-expressions priming. Here, there was robust priming between some and number expressions, $F(1,47)=5.58, p=.022$, but not between some and ad hoc expressions, $F<1$, or number and ad hoc sentences, $F<1$. There was also an interaction between the degree of between-expression priming for ad $h o c$ expressions and the other combinations, $F(2,98)=$ 4.42, $p=.015$.

We did not observe between-expression priming for the $a d$ hoc expressions in Experiment 2. However, the withinexpression priming effect was marginally smaller for the ad hoc expressions than for the others, and there were far fewer strong interpretation responses. Participants might therefore have had more difficulty identifying the alternatives for the ad hoc sentences and consequently, even if betweenexpression priming was occurring, priming effects would have been smaller and more difficult to observe.


Figure 4. Experiment 2 results. The figure shows responses for each probe expression for within-expression priming (upper panel) and each combination of betweenexpression priming (lower panel).

## Experiment 3

In Experiment 3 we hoped to remedy the low rate of strong interpretations in the ad hoc condition by introducing additional items to increase the salience of the alternatives. We reasoned that participants were always selecting the weak interpretation because they were unsure what might make a "better picture" (i.e., whether the alternative would have any relevance). We therefore introduced 20 extra bias trials at the start of the experiment of the form, "There is an A," with the target picture being a single letter.

## Results and Discussion




Figure 5. Experiment 3 results.
Within-expression priming is shown in the upper-panel of Figure 5. The proportion of strong interpretations for the ad hoc expressions is much higher than in Experiment 2, and
the degree of within-expression priming appears larger. As in Experiment 2, significant within-expression priming is observed in each expression, all $t(49)$ 's $>4.76, p$ 's $<.001$, but we found no evidence that the within-expression priming varied among expressions, $\mathrm{F}<1$. The lower-panel of Figure 5 shows the between-expression priming trials. As in Experiments 1 and 2, there was robust priming between some and number expressions, $t(50)=2.12, p=.036$. Critically however, there was no between-expression priming for combinations involving the ad hoc expressions, $t ' s<1$, and there was a significant interaction between the three combinations, $F(2,100)=3.39, \mathrm{p}<.05$.
In this experiment we observed the same sized withinexpression priming effect for ad hoc expressions compared to the other expressions. This suggests that the alternatives for ad hoc expressions were just as available too. Yet we failed to observe any between-expression priming effects involving the ad hoc sentences. These findings suggest that between-expression priming was due to priming of the search for alternatives, and not priming of the mechanism that negates the alternatives. Furthermore, our results suggest that there are separate mechanisms for determining context-free alternatives (the lexical expressions) and context-dependent alternatives.

## General Discussion

Our studies investigated how the processor enriches basic meanings with negated alternatives. An intuitive and parsimonious processing prediction was that there is a single, universal mechanism across diverse linguistic forms. After all, the enriched meanings that we used could all arguably be derived using the same reasoning. Contrary to this prediction, however, we found that whilst enriched readings of some, number, and plural morphology expressions can prime each other, they cannot prime enriched readings of ad hoc expressions, even though all the expressions can prime enriched meanings of their own form. This suggests that there are multiple procedures for enrichment based on alternatives, and that these are split between context-free and context-dependent expressions.
We conceive of context-free enrichment procedures as an instruction to the processor to search a part of the lexicon for the appropriate alternatives. For example, with some, the instruction would be to retrieve appropriate alternatives, such as all, which could then be negated. The contextdependent procedure is different in that it does not involve instructions to search the lexicon, but to search out plausible alternatives from the context. While there might be some overlap between these procedures, the failure to observe priming between the different expressions also provides a robust test that distinguish them.

## What is being primed?

We explain our results by referring to the priming of the search for alternatives. Here we consider other explanations.
One that we can eliminate is that we have primed a
general acceptance of weak statements. Our items are
constructed in such a way that there were informationally weak interpretations (e.g., some and possibly all) and informationally strong interpretations (e.g., some but not all), and a possible explanation of Experiment 1, therefore, is that participants were primed to accept the weaker/stronger interpretation (and also a potential explanation for Raffray \& Pickering, 2010). Experiments 2 and 3 rule out this explanation, however, because the $\mathrm{ad} h o c$ sentences also had the weak/strong distinction but were not primed by the other expressions. More generally, the failure to observe between-expression priming effects with the ad hoc sentences eliminates any explanation that would apply across all of the expressions.
Another possibility is that our findings could be explained by the priming of alternatives, rather than priming of the search for alternatives. Our within-expression priming effects could indeed be explained in this way. For example, the strong primes of some could make the alternative all more salient to the participant, and therefore when presented with the probe, such as "some of the letters are As," the participant might have been more likely to realise that the sentence could have read, "all of the letters are As." This would then have led them to choose the "better option" box more often. This cannot be the whole story, however, because we also observed between-expression priming. Here, the alternatives were different across expressions and so the salience of alternatives from one expression should not influence the rate of strong responding from other expressions. For example, a some strong prime might make all more salient, but the salience of all should not influence the salience of the exactly $N$ reading of the numbers. The between-expression effect cannot be explained by priming of alternatives; it requires priming of the search for the alternatives more abstractly.

## Similarities and differences between phenomena

Our results show that some, the numbers, and plural morphology can be grouped together, but that ad hoc expressions behave differently. This pattern can be related to the linguistic literature that investigates how these particular phenomena are derived.

First, consider the similarity between some, the numbers and plurals. These phenomena are very diverse. For instance, some theorists have argued that the numbers and some should be considered different phenomena (e.g. Breheny, 2008). Our results provide evidence against an extreme form of this view (that the numbers and some are completely unrelated) or the idea that numbers would not involve alternatives. Furthermore, the claim that plurals may be related to the others is an audacious one, which relies on the fully counterintuitive hypothesis that plural morphology is semantically vacuous. Our results provide further striking evidence in favor of this counterintuitive view of plural morphology.

Second, our results also distinguish different phenomena in a meaningful way. Katzir (2007) provides the most precise and complete implementation of alternative
generation. In essence, Katzir argues that there are two separate procedures for calculating alternatives. The first involves replacing a phrase by a simpler, related phrase, e.g., some $=>$ all, or ate a lot $=>$ ate, and the second involves replacing a phrase by a contextually salient phrase which may or may not be simpler $(A=>A$ and $B)$. Interestingly, Katzir specifies different procedures for the computation of ad hoc alternatives and alternatives related to the other three phenomena. This is exactly how our results split the landscape as well (see also Fox 2012 for converging developmental data). Hence, we obtain both a confirmation of the theory, and a natural interpretation of our results.

## Conclusion

We set out to investigate whether there are abstract procedures for enriching basic meanings considering words that were not pronounced, much like the structural priming research has investigated whether there are abstract representations of syntactic structure. Our results show that the scope of the enrichment procedure is wide (e.g., affecting the interpretation of numbers as well as of the plural morphology), and confirm finer-grained properties of the system by distinguishing contextual and non-contexutal alternatives.

## References

Breheny, R. (2008) A new look at the semantics and pragmatics of numerically quantified noun phrases Journal of Semantics 25(2): 93-140.

Katzir, R (2007). Structurally-defined alternatives. Linguistics and Philosophy 30:669-690.

Fox, D. (2012) Children's Interpretation of Disjunctive Sentences and the Theory of Scalar Implicature. Joint work with Singh, Wexler, Astle, Kamawara. Talk presented at MIT colloquium.

Horn, L. (1989) A Natural History of Negation. University of Chicago Press. Chicago.

Hirschberg, J (1991). A Theory of Scalar Implicature. New York: Garland Press.

Pickering, M.J., \& Ferreira, V. S. (2008). Structural priming: A critical review. Psychological Bulletin, 134, 427-459.

Raffrey, C. \& Pickering, M.J. (2010). How do people construct logical form during language comprehension? Psychological Science, 21, 1090-197.

Spector, B (2007). Aspects of the pragmatics of Plural Morphology: On Higher-Order Implicatures Published in U. Sauerland \& P. Stateva (eds.), Presuppositions and Implicatures in Compositional Semantics, PalgraveMacmillan, 2007.
van Rooy, R. \& K. Schulz (2006). Pragmatic meaning and non-monotonic reasoning: The case of exhaustive interpretation. Linguistics and Philosophy 29:205-250.

Rooth, (1992). A theory of focus interpretation. Natural Language Semantics 1, 75-116.

Zevakhina, N. and B. Geurts. (2011). Scalar diversity. Manuscript.

# Space is Special: A domain-specific mapping between time and nontemporal magnitude 

Roberto Bottini (bottinir@newschool.edu)<br>Christopher Guarino (guarc637@newschool.edu)<br>Daniel Casasanto (casasand@newschool.edu)<br>Department of Psychology, $805^{\text {th }}$ Avenue, New York, NY 10011 USA


#### Abstract

Across different domains the magnitude of a stimulus is positively correlated with its perceived duration: bigger, brighter or louder stimuli are usually perceived to last longer than smaller, dimmer or softer ones. According to A Theory of Magnitude (ATOM), temporal and nontemporal magnitudes are linked in the human mind by virtue of sharing a common metric. This claim has been challenged by studies in the domains of brightness and loudness suggesting that it is not the difference in magnitude between stimuli, but rather their degree of change from background that modulates duration judgments. But do the same relationships hold between perceived duration and all prothetic dimensions? We tested the influence of stimulus magnitude and relative change on temporal judgment in the domain of space. We found that, unlike brightness and loudness, spatial length can influence duration judgments independently of the degree of change from a common background, and that this effect is context dependent. Thus, an approach based exclusively on the degree of change between stimulus and background is not sufficient to account for the effect of magnitude on temporal judgments. Our results suggest that space has a privileged link with temporal representations compared to other prothetic domains, challenging the hypothesis that space-time relationships are the product of a domain-general magnitude system.


Keywords: ATOM; Metaphor; Space; Time; ATOC

## Introduction

Judgments of duration can be influenced by non-temporal aspects of events such as stimulus magnitude (Walsh, 2003; Xuan, Zhang, He, \& Chen, 2007). Bigger stimuli are judged to last longer than smaller ones (Xuan et al., 2007), brighter stimuli longer than dimmer ones (Xuan et al., 2007; Goldstone et al., 1978), and louder sounds longer than softer ones (Goldstone et al., 1978). Magnitude Effects have often been interpreted as the effect of absolute magnitude on stimulus duration: more intense stimuli seem to last longer (Xuan et al. 2007; Bueti \& Walsh, 2009). It has been suggested that temporal and non-temporal magnitudes are positively correlated in the human mind by virtue of sharing a common metric (Xuan et al. 2007; Walsh, 2003). On this view, magnitudes across different prothetic domains (i.e. domains that can be experienced as 'more than' or 'less than'; Walsh, 2003) are represented in the brain by a generalized magnitude system. Duration and other prothetic domains are linked by a monotonic "more A - more B" mapping (Bueti \& Walsh, 2009) such that "bigger, faster, brighter, further in one domain should correlate with bigger, faster, brighter, further in another" (Bueti \& Walsh, 2009, p.1832).

In spite of a large body of supporting evidence, the hypothesis that stimulus magnitude and its perceived duration are positively correlated has been challenged. It has been suggested (Matthews, Stewart, \& Wearden, 2011) that it is the relative difference between the stimuli and a common background, rather than the absolute magnitude of the stimuli, that modulates the subjective experience of duration. In one experiment (Matthews et al., 2011) participants judged the duration of two successive stimuli that varied both in duration and brightness. When the stimuli were presented on a dark background, brighter stimuli were judged to last longer than dimmer ones, on average. Yet, when the same stimuli were presented on a white (brighter) background, the opposite effect was found: dimmer (less intense) stimuli were judged to last longer than brighter ones. The same results were obtained when louder and softer sounds were presented against quiet or noisy backgrounds. Further experiments also support the hypothesis that duration judgments are proportional to the difference between the stimulus and its background. When an "oddball" stimulus is presented within a sequence of repeated presentations of a standard stimulus, the perceived duration of the oddball is exaggerated compared to the standards (Tse, Intriligator, Rivest, \& Cavanagh, 2004; Ulrich, Nitschke, \& Rammsayer, 2006). In the Oddball Effect the relative difference between the stimulus (the oddball) and the background (the repeated standard stimulus) has been found to predict subjective temporal dilation: The larger the difference between oddball and standard, the larger the temporal dilation (Schindel, Rowlands, \& Arnold, 2011).

Moreover, the absolute magnitude of the oddball appears to be irrelevant for duration judgments. Schindel and colleagues (Schindel et al. 2011, Exp. 2) presented their participants with a series of gray disks (the standard stimulus) that was unpredictably interrupted by an odd disk that was either brighter (more intense) or dimmer (less intense). If the subjective duration of events is positively correlated with the absolute magnitude of these events, then dimmer oddballs should be judged to last less time compared to the brighter standard. At minimum, the absolute brightness of the oddballs should modulate their effect: Even if both brighter and dimmer oddballs were judged to last longer than the standard, the temporal dilation should be more pronounced for the brighter, more intense oddballs. Contrary to these predictions, however, stimulus magnitude had no effect whatsoever on duration judgments: Dimmer and brighter oddballs led to equivalent temporal expansion. Once again, it was the relative difference
between the stimuli and the background that determined changes in subjective duration, and not the absolute magnitude of the stimuli (Shindel et al., 2011).

Altogether, these studies suggest a reinterpretation of "more A - more B" Magnitude Effects (Bueti \& Walsh, 2009). We propose that these effects may be a special case of Stimulus-Background Effects. The subjective duration of a stimulus is proportional to the difference between the stimulus and its background. Magnitude Effects, then, are simply Stimulus-Background Effects for which the relative change happens to be in magnitude. The relative variation does not need to occur in prothetic domains: Similar effects have been found for variation in shape (Tse et al. 2004), complexity (Schiffman, H. R., \& Bobko, 1974), color (Tse et al. 2004) and orientation (Schindel et al. 2011), which are qualitative (metathetic) domains of experience. If indeed Magnitude Effects are a species of Stimulus-Background Effects, they need not depend on any neurocognitive mechanisms that are specific for magnitude representations, but rather on mechanisms that support comparison of values along prothetic and metathetic continuums, alike (see Eagleman \& Pariyadath, 2009 for a similar proposal).

To summarize, we can distinguish two theoretical approaches that seek to explain the effect of non-temporal magnitudes on temporal judgments. A Theory of Magnitude (ATOM; Walsh, 2003) posits that duration is positively correlated with other prothetic domains in the mind and brain by virtue of sharing the same magnitude-specific representational basis (Bueti \& Walsh, 2011; Xuan et al. 2007; Walsh, 2003). Under this assumption, temporal distortions induced by variation in metathetic (qualitative) domains such as color or shape exploit different cognitive and neural mechanisms compared to similar effects induced by variation in prothetic (quantitative) domains. The alternative approach, which we will call A Theory of Change (ATOC), suggests instead that Magnitude Effects are particular cases of Stimulus-Background effects.

The experimental evidence reviewed above favors ATOC over ATOM. There is no special representational link between duration and prothetic dimensions, and no positive correlation between the magnitude of a stimulus and its duration: Temporal illusions attributed to the absolute magnitude of the stimuli can be explained by the relative difference between stimulus and background (Matthews et al., 2011; Shindel et al., 2011).

In this study we seek to investigate whether ATOC can fully explain the relationship between perceived duration and non-temporal magnitudes. Both ATOM and ATOC have in common the assumption that all prothetic domains influence temporal judgments in the same way. On the basis of metaphor theory (Lakoff \& Johnson, 1999; Casasanto \& Boroditsky, 2008), however, we predict that different nontemporal domains will influence temporal judgments differentially, depending on the relationships between these domains in our experience. Specifically, the relationship between perceived duration and non-temporal magnitude should be different for spatial magnitude than it is for other
prothetic domains. That is, the relationship between space and time is special.

## Space and time: an experiential link

Compared to other prothetic domains, space and time seem to be linked in the human mind by a special relationship. Across languages and cultures, spatial expressions are widely recruited to talk metaphorically about time (Lakoff \& Johsnon 1999). These patterns in language have motivated non-linguistic experiments supporting the hypothesis that people use spatial conceptual structures to think about time. Across studies, stimuli that extend farther in space are judged to last longer in time (e.g., Casasanto \& Boroditsky, 2008). This relationship between duration (i.e., temporal magnitude) and length (i.e., a kind of spatial magnitude) has been found in the judgments of children (Casasanto, Fotakopoulou, \& Boroditsky, 2010) and infants (Srinivasan \& Carey, 2010), as well as adults.

Why do many of the world's languages metaphorize duration in terms of length (e.g., a long time), instead of some other prothetic domain such as brightness or loudness (e.g., a bright time; a loud time)? Perhaps this is because space and time are correlated in our experience of the world in a way that brightness and time and loudness and time are not. As a moving object travels farther though space, more time passes. This positive correlation between magnitudes in space and time does not seem to exist between duration and other prothetic domains. Brighter things do not necessarily last a longer time than dimmer things (in fact the opposite may be true), and louder events do not necessarily last longer than softer ones.

Implicitly linking space and time in our minds may be useful because these domains are linked in the world. Knowing that "more space" is generally correlated with "more time" can provide a useful heuristic, facilitating interactions with our physical environment. By contrast, there does not appear to be any analogous link between duration and other prothetic domains in the world. As such, a representational link between temporal and spatial magnitudes in the human brain/mind is functionally motivated, and reflects regularities in our physical experience. But an analogous link between temporal magnitude and brightness or temporal magnitude and loudness would not have the same functional motivation, since these links would not have any clear basis in experiential regularities.

## Testing for a special link between time and space

In this study, we compared the effect of spatial magnitude (specifically spatial length) and relative degree of change on duration judgments. In previous studies (Casasanto \& Boroditsky, 2008; Xuan et al., 2007) the relative degree of change was positively correlated with magnitude (e.g. line length). This correlation made it impossible to tell which of the two factors were driving the effect. We designed the current experiments so that absolute magnitude of the stimuli and relative difference from the
standard/background were orthogonal to each other. If the effect of spatial magnitude on time is due to the relative amount of change, we should expect that when the difference from the background is the same, stimuli with different spatial magnitudes will be perceived as having the same duration. This outcome would support ATOC, and indicate that space stands in the same relation to time as other prothetic domains (e.g., Shindel et al., 2011). Alternatively, if spatial and temporal magnitudes, per se, are linked in the mind, we should observe a magnitude effect yielding longer duration judgments for spatially longer stimuli than for spatially shorter ones. This outcome would suggest the relationship between space and time differs from the relationship between time and other prothetic domains.

## Experiment 1. Long and short oddballs

In this first experiment we used a classic oddball paradigm to test the influence of stimulus magnitude on perceived duration. The standard stimulus was a 5 cm gray line, while the oddballs were lines of either shorter or longer length. If perceived duration is affected by the relative difference between the standard and the oddball (Schindel et al. 2011; Matthews et al. 2011) rather than spatial magnitude per se, both large and small oddballs should lead to the same effect of temporal expansion (i.e. a classic Oddball Effect). On the other hand, if stimulus spatial magnitude influences perceived duration (Casasanto \& Boroditsky, 2008), we should expect that large oddballs should lead to a greater temporal expansion compared to smaller ones.

## Methods

Participants 12 participants were recruited in the NYC area. All had normal or corrected-to-normal vision.

Stimuli. Stimuli consisted of lines of different sizes centered on a black background. The standard stimulus was a 5 cm gray (RGB 128, 128, 128) line. The oddballs were a 2.5 and 10 cm line of the same color. The width of all lines was fixed at 2 mm .

Procedure. Participants were seated in a darkened room and viewed stimuli from a distance of approximately 60 cm . For each trial 9 lines appeared sequentially in the middle of the computer screen. The standard lines $(5 \mathrm{~cm})$ were presented eight times in each trial with the remaining stimulus being the odd line (either 2.5 or 10 cm ). Each oddball appeared unpredictably between the 5th and 8th stimuli. Oddball position was determined randomly on a trial-by-trial basis. Each stimulus was followed by a blank screen during a 300 ms ISI. Standard stimuli were presented for 500 ms , whereas oddballs were presented for $300,400,450,500$, 550,600 or 700 ms . At the end of each trial, a fixation cross appeared in the middle of the screen and participants had to indicate whether the oddball had remained on the screen for more or less time than the standards. Each of the seven oddball durations was presented 10 times for each of the 2 oddballs, for a total of 140 trials. Participants completed the
experiment in three blocks of 42, 42 and 56 trials. Participants responded by pressing a key with the left index finger for "less time" and a key with the right index finger for "more time" or vice versa, with key position counterbalanced between subjects.

Proportions of "more time" responses to each oddball duration were fitted using a Weibull function for individual data sets. The point of subjective equality (PSE), which is the point at which the duration of the oddball is on average judged equal to the duration of the standard, was calculated graphically as the duration corresponding to $50 \%$ of "more time'" responses.

## Results

Long oddballs led to significant temporal expansion: PSE: 480, $t(11)=2.66, p=.02$; while short oddballs led to nonsignificant temporal contraction, PSE: 521, $\mathrm{t}(11)=1.39, \mathrm{p}=$ .19 (Fig.1, left panel).
To examine the effects of spatial length on duration judgment, we fitted a generalized linear model with binomial distribution for time judgments using the seven oddball durations and the two oddball lengths as predictors of "more time" and 'less time" responses. We found that oddball spatial length influenced the oddball effect, with longer oddballs leading to a greater temporal expansion than shorter oddballs, Wald $\chi 2(1)=14.53, \mathrm{p}<.001$.


Fig1. Effect of Long and Short oddballs in Experiment 1 and 2. Error bars depict SEM (corrected for within subjects comparisons).

## Discussion

In Experiment 1 the repetitive presentation of a gray line was unpredictably interrupted by the presentation of either a spatially longer or shorter line of different duration. Even though the two oddball lines had the same relative difference from the standard, spatially longer lines induced a greater oddball effect than spatially shorter ones. Moreover, while longer oddballs led to a significant temporal expansion (compared to the Point of Objective Equality (POE)), shorter oddballs led to a non-significant temporal contraction. These results suggest that the magnitude of the stimulus does influence the subjective experience of duration independently of the relative difference between stimulus and background.

Nevertheless, a different interpretation of the data is possible. Perhaps the difference between the long oddball
and the standards was perceived to be greater than the difference between the short oddball and the standards. Since magnitude judgments for prothetic dimensions, including space, follow Weber's Law, we selected the three values of spatial length according to a logarithmic scale in which the central value was the geometric mean of the two extreme values. Thus, the long and short oddballs should be psychologically equidistant from the standards. Nevertheless, participants may have noticed that the difference between the long line and the standard was numerically greater than the difference between the short line and the standard. This "difference of differences" might have lead to the asymmetric results reported above.

## Experiment 2. Linear scaling

Experiment 2 was designed to rule out the possibility that the difference between the longer oddball and the standard was perceived as greater than the difference between shorter oddball and standard. In this test the differences between the standard line and the longer oddball and the standard line and the shorter oddball were numerically the same.

## Methods

Participants 20 participants were recruited in the NYC area. All had normal or corrected-to-normal vision.

Stimuli and Procedure Stimuli and procedure for Experiment 2 were the same as those in Experiment 1 with the following exception: Long oddballs were 7.5 cm long.

## Result

Longer oddballs led to significant temporal expansion, PSE: $471, \mathrm{t}(19)=2.73, \mathrm{p}=.01$, while shorter oddballs led to a small, non-significant temporal expansion, PSE: 491, t(19)= $0.82, \mathrm{p}=.42$ (Fig.1, right panel).

To examine the effects of spatial length on duration judgments, we fitted a generalized linear model with binomial distribution for time judgments using the seven oddball durations and the two oddball lengths as predictors of " more time" and 'less time" responses.

We found that oddball length influenced the oddball effect, with longer oddballs leading to greater temporal expansion than shorter oddballs, Wald $\chi 2(1)=6.00, p=.01$.

## Discussion

In this experiment the relative difference between oddballs and standards was the same numerically for both longer and shorter oddballs. Yet, long oddballs led to a greater temporal expansion than did smaller ones. The Magnitude Effect observed in both experiments one and two seems to be proportional to the magnitude of the stimuli, independent of the degree of change (relative difference) between the oddball and the standards. These findings again suggest that the spatial magnitude of a stimulus modulates its perceived duration independently from the difference between stimulus and background. An approach based exclusively on the degree of change between stimulus and background is
not sufficient to account for the temporal modulation observed.

Experiments 1 and 2 provided an interesting additional piece of evidence. Long oddballs always led to significant temporal expansion, whereas short oddballs led to nonsignificant contraction (Exp.1) or expansion (Exp.2). In the case of the shorter line, the effect of relative change would lead to temporal expansion (the shorter line is different from the standard), while the effect of magnitude would lead to temporal contraction (the shorter line is indeed shorter than the standard). We can hypothesize that when the two factors are in opposition they cancel each other out, leading to neither temporal expansion nor temporal contraction. That is, both the degree of change and the absolute magnitude of the stimuli contribute to the judgment of duration, and their relative weight seems to be roughly equal. However, the answer may not be so simple.

Seifried and Urlich (Seifried \& Ulrich, 2010) report an experiment in which a smaller stationary disk was presented as an oddball among repetitive presentations of a bigger disk. Even though the oddball had a smaller size compared to the standard it led to a significant effect of temporal expansion (Seifried and Urlich, 2010, Exp. 3, footnote on page 97). This result is inconsistent with the hypothesis of an equal and opposite influence of magnitude and relative change, and with the effect of magnitude reported here. But in Seifried and Urlich's experiment there was only one oddball type, the smaller size-disk, instead of both a larger and smaller one. It is possible that the pattern of interaction between stimulus magnitude and degree of change is context dependent. The weight of each factor in influencing duration judgments depends on the salience of each factor in a given context. In a classic oddball paradigm, with only one kind of oddball, the direction of change (more/less) may be overshadowed by the fact that the oddball is simply different from the standard. The oddness of the oddball is a more salient feature of the event compared to its absolute magnitude. In this context the relative change is a more weighted factor than stimulus magnitude, and even small oddballs would lead to temporal expansion. Conversely, when two oddballs with different sizes are included in the design, absolute magnitude may become more salient: both oddballs are different from the standard but they differ in different ways. The oddballs aren't just odd, but either longer or shorter. The polarity of the magnitude continuum becomes more salient, leading to an increased effect of stimulus magnitude over relative change in influencing duration judgments. We designed Experiment 3 to test this hypothesis.

## Experiment 3: Short oddballs only

In Experiment 3 the only oddball presented was the shorter line. If the interaction between stimulus magnitude and relative difference is context dependent, modulated by the relative salience of each factor, we should expect to see a significant oddball effect (i.e., subjective temporal expansion).

## Methods

Participants. 20 participants were recruited in the NYC area. All had normal or corrected-to-normal vision.

Stimuli and procedure. In this experiment there was only one type of oddball, a gray line (same color as the standard) 2.5 cm long. Each of the seven oddball durations was presented 12 times, for a total of 84 trials. Participants completed the experiment in two blocks of 42 trials each. Otherwise, the stimuli and procedure were the same as those in experiments 1 and 2.

## Results

The subjective duration of the oddballs was exaggerated compared to the standard, PSE: 480, $\mathrm{t}(19)=2.10, \mathrm{p}=.05$. To examine the effects of context on duration judgments, we conducted a generalized linear model with binomial distribution to compare the effect of short oddballs across Experiment 1 and Experiment 3. We found that short oddballs led to a greater temporal expansion in Experiment 3 than in Experiment 1, Wald $\chi 2(1)=6.34, p=.01$ (Fig.2).
The same comparison between Experiment 2 and Experiment 3 didn't produce a significant result (Wald $\chi 2(1)=1.76, p=.28)$.


Fig2. Effect of the short oddball in Experiment 1 and 3. Error bars depict (uncorrected) SEM

## Discussion

Shorter oddballs unpredictably appearing among longer standard stimuli led to a classic expansion of subjective duration, as reported by a previous study (Seifried \& Urlich, 2010). These results support the hypothesis that the effect of stimulus magnitude and degree of change on duration judgment is context dependent. The more relevant one factor is made by contextual features, the more it will contribute to shaping the subjective experience of duration of a given event.

The oddball effect elicited by the short oddball in experiment 3 was significantly greater than the effect produced by the same stimulus in experiment 1, but not in experiment 2 . This outcome can be explained by the fact that, in experiment 2 , the difference between the standard and the shorter oddball was probably perceived as bigger compared to the difference between the standard and the longer oddball, due to linear scaling. This perceptual
asymmetry may have inflated the oddball effect produced by the short-oddball. For this reason Experiment 1, in which geometric scaling was used and the relative difference between oddballs and the standard was equated, constitutes a better basis for comparison.

## General Discussion

The main finding of the current research is that the absolute spatial magnitude of stimuli can influence duration judgments independently from the relative amount of change between stimuli and background. When the difference from the standard (background) was the same, oddballs that were spatially longer than the standard led to a greater subjective temporal expansion than oddballs that were spatially shorter than the standard. Space and time seem to be linked in the human mind by a positive correlation according to which objects that extend farther in space are judged to last longer (Casasanto \& Boroditsky, 2008). Such a positive correlation between absolute magnitude and duration does not hold for other prothetic domains like brightness or loudness: The apparent correlation between duration and brightness and duration and loudness has been explained in terms of the relative difference from stimuli and background that modulates duration judgments, independent of stimulus magnitude (Matthews et al. 2011, Schindel et al. 2011).

Our results support the hypothesis that space and time share a special link in the human mind (Casasanto \& Boroditsky, 2008; Lakoff \& Johnson, 1999). This link is experientially motivated, since space and time are correlated in our everyday experience, in a way that brightness and time and loudness and time are not. The domain specificity of the link we observe between space and time is inconsistent with a domain-general magnitude metric as hypothesized by ATOM: Not all prothetic domains are represented the same way in the human mind.

Moreover, our results cannot be explained entirely as effects of the degree of change between stimuli and background. Therefore, ATOC cannot completely account for the pattern of temporal distortions observed in our experiments, either. Rather, both the spatial magnitude of the stimuli and the relative difference between stimuli and background play a role in shaping duration judgments.

The relative weight of these two factors is context dependent. When there was only one kind of oddball (Exp. 3), which was shorter than the standard, the oddball led to a classic temporal expansion. Yet, when a longer oddball was added to the design (Exp. 1 and 2), shorter oddballs were judged, on average, to have the same duration as the standard. That is, the same oddball embedded in the same sequence of standard stimuli produced different patterns of temporal distortion depending on the context in which it was presented. Such contextual variability is consistent with the hypothesis that the Oddball Effect is not mediated by low level perceptual processes like visual adaptation, but rather depends on higher-level comparison (Schindel et al. 2011) and on the contextual salience of the oddballs (Van

Wassenhove, Buonomano, Shimojo, \& Shams, 2008). When oddness is the salient feature (Exp. 3) a "more change more time" mapping is evident, whereas when the polarity of the magnitude continuum becomes salient (Exp. $1 \& 2$ ), a "more space - more time" mapping is also evident.

There is now considerable evidence that humans' representations of time are grounded in their nontemporal experience as well as in their temporal experience. Why should people systematically incorporate certain kinds of non-temporal information into their temporal thinking? Some non-temporal aspects of events are often good proxies for time, and they may be easier to perceive or remember than time, per se. For instance, the domain general "more change - more time" mapping, which is at the basis of ATOC, is consistent with our experience that greater changes occur over greater durations (see Fraisse, 1984). Often, amount of change may provide a perceptible basis for duration judgments: We cannot see time passing, but we can see physical objects changing (e.g., containers filling, leaves changing color, children growing). Likewise, people may rely on the domain-specific mapping between spatial extent and time because spatial aspects of our experience are generally more perceptible than the associated temporal aspects (e.g., it is possible to see how far a ball rolls (distance) but not to see how long it takes (duration); Casasanto \& Boroditsky, 2008; Casasanto et al. 2010; Lakoff \& Johnson, 1999).

## Conclusions

Spatial magnitude and duration share a representational link that does not extend to other prothetic domains such as brightness and loudness. This domain specificity is inconsistent with a domain-general magnitude metric as hypothesized by ATOM. Our results are also only partly explained by ATOC as effects of a change in the magnitude of a nontemporal aspect of the stimulus. Results are best understood as supporting both ATOC and metaphor theory, in combination.

The pattern of subjective temporal expansion predicted by ATOC was observed most clearly when only one type of oddball stimulus was included, which highlighted the simple fact of a difference (i.e., change) between the oddball and the standard. The pattern predicted by metaphor theory was found most clearly when two oddballs that varied in spatial length were included, which highlighted their magnitudes.

Grounding representations of temporal magnitude in our experiences of relative amount of change (ATOC) and in spatial magnitude (metaphor theory) are both functionally motivated: As objects change, or as they travel farther through space, more time passes. As such, nontemporal aspects of events that correlate reliably with time can serve as perceptible indices of temporal change, which is imperceptible. Grounding representations of temporal magnitude in other prothetic magnitudes, however, would not be functionally motivated: The absolute magnitudes of brightness and loudness, for example, do not appear to be correlated with duration in our everyday experience. From
this functional-experiential perspective, it is unsurprising that our data support ATOC and metaphor theory, but not ATOM. ATOC and metaphor theory appear to be functionally and experientially motivated in a way that ATOM is not.

## References

Bueti, D., \& Walsh, V. (2009). The parietal cortex and the representation of time, space, number and other magnitudes. Phil. Trans. of the Royal Society of London. 364(1525), 1831-40.
Casasanto, D., Fotakopoulou, O., \& Boroditsky, L. (2010). Space and Time in the Child's Mind: Evidence for a Cross-Dimensional Asymmetry. Cognitive science, 34(3), 387-405.
Eagleman, D. M., \& Pariyadath, V. (2009). Is subjective duration a signature of coding efficiency? Phil. Trans. of the Royal Society of London. 364(1525), 1841-51.
Fraisse, P. (1984). Perception and estimation of time. Annual review of psychology, 35(1), 1-37.
Goldstone, S., Lhamon, W. T., \& Sechzer, J. (1978). Light intensity and judged duration. Bulletin of the Psychonomic Society, 12, 83-84.
Lakoff, G., \& Johnson, M. (1999). Philosophy in the flesh: The embodied mind and its challenge to Western thought. Chicago: University of Chicago Press.
Matthews, W. J., Stewart, N., \& Wearden, J. H. (2011). Stimulus intensity and the perception of duration. $J E P: H P P, 37(1), 303-13$.
Ornstein, R. (1969). On the experience of time. Hammondsworth: Penguin.
Schiffman, H. R., \& Bobko, D. J. (1974). Effects of stimulus complexity on the perception of brief temporal intervals. JEP:G, 103, 156-159.
Schindel, R., Rowlands, J., \& Arnold, D. (2011). The oddball effect: Perceived duration and predictive coding. Journal of Vision, 11, 1-9.
Seifried, T., \& Ulrich, R. (2010). Does the asymmetry effect inflate the temporal expansion of odd stimuli? Psychological research, 74(1), 90-8.
Srinivasan, M. and Carey, S. (2010). The long and the short of it: On the nature and origin of functional overlap between representations of space and time. Cognition, 116, 217-241.
Tse, P. U., Intriligator, J., Rivest, J., \& Cavanagh, P. (2004). Attention and the subjective expansion of time. Perception \& psychophysics, 66(7), 1171-89.
Ulrich, R., Nitschke, J., \& Rammsayer, T. (2006). Perceived duration of expected and unexpected stimuli. Psychological research, 70(2), 77-87.
Van Wassenhove, V., Buonomano, D. V, Shimojo, S., \& Shams, L. (2008). Distortions of subjective time perception within and across senses. PloS one, 3(1).
Walsh, V. (2003). A theory of magnitude: common cortical metrics of time, space and quantity. $\mathrm{TiCS}, 7(11), 483-488$.
Xuan, B., Zhang, D., He, S., \& Chen, X. (2007). Larger stimuli are judged to last longer. Journal of Vision, 7, 1-5.

# Open-Ended Category-Based Induction: The Influence of Associative Strength and Structured Knowledge Representations 

Aimée Kay Bright (aimee.bright@northumbria.ac.uk)<br>Department of Psychology, Northumberland Road, Newcastle upon Tyne, NE1 8ST, UK

Aidan Feeney (a.feeney@qub.ac.uk)<br>School of Psychology, Queen's University Belfast<br>Belfast, BT7 1NN, UK


#### Abstract

Accounts of category-based inductive reasoning can be distinguished by the emphasis they place on associative retrieval processes versus structural knowledge representation. Using an open-ended category-based induction task with a secondary task manipulation, we explored whether the relative importance of these two processes in determining the reasoning output depends upon the availability of mental resources. Regressing indices of strength of association and measures of structured relation against reasoners' inferences showed that people's inductions generated under cognitive load were more strongly predicted by associative strength between base and conclusion category. In contrast, inferences made under no load were best predicted by the measure of the existence of structural relations between base and conclusion category. This suggests that people make use of associative processes and recruit structured knowledge to make inductive inferences, and that the relative importance of these two forms of reasoning is determined by the availability of mental resources.


Keywords: Category-Based Induction; Knowledge; Categorical Inferences; Reasoning.

## Knowledge and Category-Based Induction

Generalizing properties from one category to another is known as category-based inductive reasoning. If people learn that carrots have a certain disease, they might infer that rabbits could also be affected. However, if people are reasoning about shared cells, they might prefer to generalize from carrots to parsnips rather than to rabbits. Understanding how people select relevant knowledge has become central to explaining the mental processes that underlie category-based inductive reasoning (Shafto, Baldwin and Coley, 2007). But is knowledge selection based on a single process, such as the activation of automatic associations in semantic memory (Rogers \& McCllelland, 2004) and the calculation of similarity (Sloman, 1993; Sloutsky \& Fisher, 2004) or the explicit representation of structural relations between categories (Osherson et al., 1990; Tenenbaum, Griffiths \& Kemp, 2007), or does it depend on an interplay between such processes? In this paper, we argue that how people reason is determined by more domain-general factors, such as available cognitive resources.

## Associative Processes in Inductive Reasoning

Associative processes can explain a host of phenomena in category-based inductive reasoning. For example, Sloman's (1993) feature-based induction model assumes that similarity represented by the degree to which premise and conclusion categories activate common features determines the strength of the conclusion. Similarly, Roger and McClelland's application of the parallel-distributed processing model to category-based inductions assumes that generalizations from one instance to another will be strong to the extent that the activated distributed representations of the two instances overlap via their shared attributes. Several predictions follow from the way in which the connectionist model acquires semantic knowledge and makes generalizations. As it acquires knowledge gradually based on experiential input, the internal representations should mirror the structure of the learning environment. For example, if one repeatedly encounters two species in the same context, the internal representations ought to reflect this statistical co-occurrence. Inductive inferences between categories should be stronger to the extent that the categories have repeatedly been simultaneously activated in semantic memory, forming strong associations

## Structural Knowledge Representations

In contrast to models that emphasize associative processes, structured knowledge representations might be necessary to draw accurate inferences where the categories have complex ecological, causal or taxonomic relations. The seminal study by Heit and Rubinstein (1994) demonstrated that people recruited differential knowledge depending upon the type of property they were asked to generalize. Similarly, structured Bayesian models (e.g., Kemp \& Tenenbaum 2009; Shafto et al., 2008) successfully explain phenomena that arise from paying attention to the higherorder interrelationships between categories. For example, reasoning about causal transmission is best predicted by inferences computed over a theoretical model of food web relations, whereas inferences about physiological properties seem to be based on an understanding about taxonomic interrelationships. Use of structural representations can also explain phenomena such as the causal asymmetry effect.

Thus, people believe that diseases are more likely to be transmitted from prey to predator than vice versa (Medin, Coley, Storms \& Hayes, 2003; Shafto et al., 2008).

An interesting question is what factors determine the use of such structural representations. Cross-cultural work and research on experts (e.g. Lopez, Atran, Coley, Medin \& Smith, 1997; Proffitt, Coley \& Medin, 2000) suggests that to some extent, use of structural representations in inductive reasoning depends upon having the appropriate background knowledge. For example, Shafto and Coley (2003) compared commercial fishermen's inductive inferences about marine life to those of US undergraduates. When reasoning about novel diseases, only the fishermen drew on causal/ecological relations between premise and target categories to inform their inference. The undergraduates tended to base all their inferences on similarity. However, while such studies illustrate that people vary in the sophistication of the structural representations they have across different domains, the underlying mental processes that prompt people to draw on these or instead fall back on simple similarity during the reasoning process remain unclear.

## Two Types of Reasoning?

One possibility is that drawing on structural representations is an effortful process, whereas the use of simple associations and similarity requires fewer mental resources. The relative importance of each strategy might be determined by domain-general factors, such as available time and mental resources. Support for this position comes from a study looking at reasoning in music experts (composers and musicians) and novices (Baraff \& Coley, 2003; Coley \& Barraff, 2003). Compared to novices, experts tended to use more elaborate context-dependent relational knowledge. However, when the induction task was carried out under time pressure, thus decreasing available cognitive processing time, experts' reasoning was indistinguishable from novice reasoning. This change in expert responding suggests that drawing on structured knowledge representations during reasoning is slow and effortful. Thus, under time pressure, experts had to rely more on associative similarity, the default for novices who lack relevant structural knowledge representations.

A study by Bright and Feeney (2010) lends further support to the suggestion that people reason differently depending upon available mental resources. Thus, when people made speeded inferences, argument strength was predicted by associative strength between the two categories, whereas causal and biological knowledge predicted inference strength when people were not under time pressure.

However, Coley et al. (2005) have argued that some phenomena may be task-specific, especially if people are unaware of the nature of the relation between categories. Most findings are based on experimental paradigms in which people evaluate the strength of an inductive argument (Rabbits have property X, therefore, Foxes have property
X), evaluate a series of conclusions (Rabbits have property X. How likely is that Foxes have property X? Eagles? Hares?), or are forced to choose between two alternative conclusion categories (Rabbits have property X. Is it more likely that Hares or Foxes have property $X$ ?). When people are presented with pre-determined base and conclusion categories, lack of structural knowledge representations that highlight relevant relations between categories might force people into adopting a default associative reasoning strategy that they wouldn't normally use. In contrast, open-ended methodologies allow people to use background knowledge in a more flexible manner. For example, Baker and Coley (Baker \& Coley, 2005; Coley \& Baker, 2004) gave their participants two related category pairs and asked them to make inferences about which other categories might also have a novel property. People tended to make inferences based on complex ecological relations rather than on taxonomic similarity, suggesting that they were recruiting whatever relevant structural knowledge representations were available to them.

In the following experiment, people were told that a base category had a property and were asked to infer which other category was most likely to also have that property. We predicted that people who generated categories under cognitive load would use a strategy that placed less demand on cognitive resources, such as similarity or strength of association. In contrast, we expected people to make use of diverse structural knowledge representations when they were not under cognitive load. Previous work using a speeded response paradigm (Shafto, Coley \& Baldwin, 2007) has suggested that taxonomic knowledge is more available to reasoning processes than is ecological knowledge. When they have sufficient time, people tend to bring taxonomic knowledge to bear when reasoning about intrinsic properties such as cells and ecological knowledge to bear when reasoning about extrinsic properties such as diseases (see Shafto et al., 2007). So that we could attempt to replicate Shafto et al's finding that taxonomic knowledge is more available to reasoning processes, we asked people to reason about cell and disease properties expecting that only under light load would there be evidence of use of ecological relations when people reasoned about diseases.

## Methods

The experiment had three phases, the induction generation phase 1 , the associative rating phase 2 and the structured relation rating phase 3 .

## Induction Generation Phase

The first phase had a 2 (load: heavy or light) by 2 (property: infection or cells) mixed design, with load as the between-subjects manipulation.

Twenty-three students ( $M$ age $=24.2$ years) from Durham University (the reasoners) were presented with 20 base categories and told that each category had a novel property,
either an infection (e.g. has infection 5 y 5 u ) or cells (e.g. has 3 -yu-cells). There were equal numbers of each property type and the combination of property type and base category was counterbalanced. Participants were then asked to generate ONE other category that they believed was most likely to also have the property. For example, people would read the following generative induction problem:

Weasels have 4Ou-cells / infection 4Ou.
Which other category is most likely to also have 4Oucells/ infection 4OU?

Once people had written down their response, they rated how likely they thought it was that the two categories shared the property on a scale from 1 (very unlikely) to 9 (highly likely).

Preceding each of the induction trials was a secondary memory task. People were presented with a $4 * 4$ dot matrix with 4 randomly placed black dots for 2000 ms . Participants remembered the location of the dots, completed the induction task and then recalled the location of the dots in an empty matrix. The configuration of the dots was different for each of the 20 trials.

In the heavy load condition, the dots were completely randomly placed, with the restriction that they could never appear in a straight or diagonal line. In the light load control condition, the dots always appeared in a straight or diagonal line, placing minimal burden on working memory.

## Association Rating Phase

In the second phase each individual reasoner's 20 category pairs were transcribed onto an association rating sheet and interspersed with 15 weakly associated distracter items. A group of 92 participants (the raters) who had not taken part in the first phase received one of 23 different sheets (approximately 4 participants per sheet) and were asked to rate the strength of association on a scale from 1 (unrelated) to 9 (very highly associated) between the 35 category pairs. They were instructed to respond as fast as possible, based on the first intuitive answer that came to mind.

## Structured Relation Ratings

In order to determine the underlying structural relations between the base and conclusion categories generated by reasoners in phase 1 , the experimenter and a second blind coder rated whether there was a taxonomic and/or interaction-based relationship between the 20 category pairs. Table 1 below contains examples of the different types of relation.

Thus, category pairs were awarded 0 if there was no discernible link between the base and the generated category (e.g. alligator $\rightarrow$ soil), 1 if they were taxonomically related (e.g. zebra $\rightarrow$ horse), 1 if they were related through a causal link or ecological interaction (e.g. hawk $\rightarrow$ mouse) and 2 if there was both a taxonomic and interaction-based relation
between the categories (e.g. cod $\rightarrow$ shark). Concordance rate across the two primary coders was $67 \%$. Disagreements were resolved through discussion with two further colleagues.

## Table 1: Coding Scheme for Structured Relations

| Taxonomic Relationship |  |
| :--- | :--- |
| Category <br> Membership | Both categories belong to the same class or <br> category (e.g. carrot \& parsnip) |
| Physiological <br> Similarity | Both categories are similar with respect to <br> specific organs or systems (e.g. bat \& bird) |
| Interaction-Based Relationship |  |
| Similar <br> Habitat | Both categories share similar or the same <br> habitat (e.g. trout \& shrimp) |
| Behavioural <br> Interaction | Both Categories interact via some aspect <br> of behaviour (e.g. monkey \& tree) |
| Food Chain <br> Interaction | Both categories interact with respect to <br> diet or eating, i.e., one category eats or is <br> eaten by the other (e.g. heron \& fish) |

## Results

## Association Ratings

We averaged association ratings made by raters in phase 2 across all 20 category pairs generated by reasoners in phase 1. One rater in phase 2 failed to complete more than $50 \%$ of the association ratings and was excluded from the analysis.
The mean association scores were analyzed with a 2 (load: heavy or light) by 2 (property: cells or infection) mixed-design ANOVA, with load as the between-subjects variable.

There was no main effect of property, $\mathrm{F}_{(1,89)}=1.08, \mathrm{p}=$ .30 , effect size $\mathrm{d}=.22$. People gave a mean association rating of $6.23(\mathrm{SE}=0.12)$ for category pairs which had been generated about shared cells, and a mean association rating of $6.15(\mathrm{SE}=0.13)$ for category pairs generated about infections.

As predicted though, there was a main effect of load, $\mathrm{F}_{(1,}$ ${ }_{89}=4.03, \mathrm{p}=.048$, effect size $\mathrm{d}=.42$, such that categories generated under conditions of heavy load ( $\mathrm{M}=6.42$, SE $=.16$ ) were rated as more strongly associated than categories generated by reasoners whose resources were minimally taxed $(\mathrm{M}=5.96, \mathrm{SE}=.16)$.

Finally, there was no interaction between property and load condition $\mathrm{F}_{(1,89)}=.55, \mathrm{p}=.46$, effect size $\mathrm{f}=0.08$.

## Types of Structured Relations

We summed the taxonomic relationship ratings and the interaction-based relationship ratings across the categories for which reasoners had made inferences about diseases, and likewise across the 10 category pairs that were generated for
shared cells. These were analyzed with a 2 (type of relationship: taxonomic or interaction-based) by 2 (load: heavy or light) by 2 (property: cells or infections) mixeddesign ANOVA, with load as the between subjects variable. The crucial result was a three-way interaction between property, relation and load, $\mathrm{F}_{(1,21)}=5.43, p=.03$, effect size $\mathrm{f}=0.51$, illustrated in Figure 1 below.


Figure 1: Number of interaction-based and taxonomic relations (and standard errors) across the two types of property for heavy and light load conditions

## Heavy Load

In the heavy load condition, there was a main effect of property, $\mathrm{F}_{(1,10)}=26.94, p<.0005, \mathrm{~d}=3.3$. Thus, when people reasoned about cells, the conclusion categories they generated shared more structural relations with the base categories ( $\mathrm{M}=5.50, \mathrm{SE}=0.19$ ) than when they generated categorical inferences about infections $(\mathrm{M}=4.64, \mathrm{SE}=$ $0.18)$.

There was also a significant main effect of type of relation, $\mathrm{F}_{(1,10)}=68.15, p<.0005, \mathrm{~d}=5.3$. Thus, people seemed to generate more taxonomically-related categories $(\mathrm{M}=8.05, \mathrm{SE}=0.41)$ than conclusion categories which were related via an interaction $(M=2.09, S E=0.39)$.

Finally, there was a significant two-way interaction between property and relation $\mathrm{F}_{(1,10)}=8.27, p=.017$, effect size $\mathrm{f}=0.91$. Bonferroni post-hoc tests showed that there were no property effects for interaction-based responses. Interaction-based responses were similarly low when reasoning about cells ( $\mathrm{M}=1.82, \mathrm{SE}=0.38$ ) and infections ( $\mathrm{M}=2.36, \mathrm{SE}=0.53, p=.29$ ). In contrast, property effects arose for taxonomic responses, showing that people gave more taxonomic responses when reasoning about cells ( $\mathrm{M}=$ $9.18, \mathrm{SE}=0.30$ ) than when reasoning about infections ( $\mathrm{M}=$ $6.91, \mathrm{SE}=0.63, p=.002$ ).

## Light Load

The pattern of results was different in the light load condition. Here, the only significant effect was for property, $F_{(1,11)}=33.0, p<.0005$, effect size $\mathrm{d}=3.5$. When people reasoned about cells, the conclusion categories they generated shared more structural relations with the base
categories $(M=6.58, S E=0.20)$ than when they generated categorical inferences about infections $(\mathrm{M}=5.33$, $\mathrm{SE}=$ 0.09 ).

Although the pattern of means suggests that people generate more taxonomically related conclusion categories ( $\mathrm{M}=7.54, \mathrm{SE}=0.77$ ) than interaction-based conclusions categories ( $M=4.38, S E=0.89$ ), this main effect of type of relation was not statistically significant, $\mathrm{F}_{(1,11)}=3.69, p=$ .08 , effect size $\mathrm{d}=0.4$. Finally, the most important difference compared to the heavy load condition was an absence of an interaction between property and type of relation, $\mathrm{F}_{(1,11)}=0.128, p=.73$, effect size $\mathrm{f}=0.38$.

## Generative Inductive Strength Ratings

Inductive strength ratings for the categories the reasoners had generated were analyzed with a 2 (load) by 2 (property) mixed-design ANOVA with load as a between-subjects variable. Inductive strength ratings did not differ between the load conditions, $\mathrm{F}_{(1,21)}<.001, p=.99$, effect size $\mathrm{d}<$ .01. Reasoners under heavy load gave a mean inductive strength rating of 5.55 ( $\mathrm{SE}=.40$ ) whereas those under minimal load rated the strength of their induction at 5.56 ( $\mathrm{SE}=.42$ ).
There was also no main effect of property, $\mathrm{F}_{(1,21)}=2.1, p$ $=.16$, effect size $\mathrm{d}=.63$. Inferences about cells $(\mathrm{M}=5.68$, $\mathrm{SE}=.32$ ) were rated as strong as inferences about infections ( $\mathrm{M}=5.42, \mathrm{SE}=.29$ ).
The interaction between load and property was not statistically significant, $\mathrm{F}_{(1,21)}=2.68, p=.12$, effect size $\mathrm{f}=$ . 35 .

## Relations between Inductive Strength Ratings, Structured Relations and Associative Strength

To explore whether reasoners place different emphasis on associative processes and reasoning based on structured knowledge representation in the two load conditions we used an associative strength measure and the index of structured relations described above to predict their inductive strength ratings.

To create the associative strength measure we averaged the mean strength of association scores attached to each reasoner's 20 category pairs across the four raters from phase 2. We then calculated Cronbach's Alpha for each of the 23 reasoners across the association ratings. The mean Cronbach's Alpha across all reasoners was .71 ( $\mathrm{SD}=.13$ ), showing that the association ratings had good inter-rater reliability.
To create the structured relation measure, the experimenter and a second blind coder assessed in how many ways the generated target could be related to the base. 0 was attached if there was no obvious structured link, 1 if there was either a taxonomic or an interaction-based connection, and 2 if they were related in more than one way.
For each reasoner who had taken part in phase 1, we used the associative strength and structured relation measures to predict his/her inductive strength ratings. The beta weights were then subjected to a 2 (load: heavy or light) by 2 (type
of beta weight: associative versus structured relation) mixed-design ANOVA, with type of beta weight as the repeated-measures variable.

There was no significant main effect of type of beta weight, $\mathrm{F}_{(1,21)}=.068, \mathrm{p}=.80$, effect size $\mathrm{d}=.11$ and no main effect of load, $\mathrm{F}_{(1,21)}=3.22, \mathrm{p}=.09$, effect size $\mathrm{d}=$ .78. However, there was a significant interaction between beta weight type and load, $\mathrm{F}_{(1,21)}=6.53, \mathrm{p}=.018$, effect size $d=1.1$. This is illustrated below in Figure 2.


Figure 2: Beta weights (and standard errors) across the two load conditions

Bonferroni posthoc tests showed that when reasoners were under heavy memory load, the associative strength beta weight ( M beta $=.20, \mathrm{SE}=.07$ ) was larger than the structured relation beta weight $(\mathrm{M}$ beta $=.02, \mathrm{SE}=.02)$, although this difference was not quite statistically significant due to the small number of participants in this condition ( $\mathrm{p}=.065$, effect size $\mathrm{d}=1.2$ ). A one-sample t -test confirmed that associative strength beta weight was significantly above zero, $\mathrm{t}_{(10)}=3.79, p=.004$, but that the structured relation beta weight was not significantly above zero, $\mathrm{t}_{(10)}=0.45, p=.66$.

The pattern was reversed when reasoners were not under a heavy memory load. Thus, the structured relation beta weight ( M beta $=.28, \mathrm{SE}=.05$ ) was slightly but not significantly larger in magnitude than associative strength beta weight $(\mathrm{M}$ beta $=.11, \mathrm{SE}=.07, p=.11$, effect size $\mathrm{d}=$ .62). However, the one-sample t-test showed that whereas the structured relation beta weight was significantly above zero, $\mathrm{t}_{(11)}=5.24, \mathrm{p}<.0005$, the associative strength beta weight was not statistically different from zero, $\mathrm{t}_{(11)}=1.56$, $p=.15$.

Across the two load conditions, the associative strength beta weight was slightly but not significantly larger for reasoners who generated their inferences under load compared to those who were not cognitively compromised ( $p=.44$, effect size $\mathrm{d}=.32$ ). In contrast, the mean structured relation beta weights were significantly larger for reasoners who generated their inferences under minimal cognitive load compared to reasoners who were cognitively burdened
by the complex dot matrix task ( $p=.001$, effect size $\mathrm{d}=$ 1.6).

The results suggest that the reasoning process used to arrive at a particular inference depends to some extent on available cognitive resources. Whereas structured knowledge representations were influential when reasoners were only under minimal cognitive load, associations seemed to be more important to reasoners under a heavy cognitive load.

## Discussion

Our results suggest that the process people adopt to generate category-based inductive inferences depends upon available cognitive resources. Categories produced under heavy load were rated as more strongly associated than categories generated under a light load. Furthermore, under heavy load conditions, those association ratings were better predictors of inductive strength than an index of structured relations. In contrast, in the light load condition the index of structured relations was the better predictor of inductive strength ratings. Furthermore, under heavy cognitive load, people were less likely to generate categories that shared more complex interaction-based relationships, whereas generating taxonomically related categories was unaffected by cognitive load.

The advantage of the open ended paradigm is that we can be sure that participants possess knowledge about structured relations between base and conclusion categories. Despite using this more flexible reasoning paradigm, people who were under cognitive load seemed less able to make use of complex structural representations, and instead relied more strongly on associative processes. This suggests that while people do seem motivated to base their reasoning on domain-specific knowledge representations such as ecological/ causal structures, this comes at a cognitive cost. If necessary, people can shift towards a more associative strategy that might result in an inference that is different to the one that would have been generated through the activation of more complex structural knowledge representations. Sloutsky and colleagues (2008) suggest that structured knowledge can arise from simple associative processes and co-occurrence. Thus, it is conceivable that even once people possess more elaborate knowledge structures, they may use associative strength as a useful heuristic short-cut during reasoning, especially when time and/or cognitive resources are sparse.

Interestingly, we found no differences in people's inductive strength ratings across the two load conditions, suggesting that people were equally confident about inferences generated using associative reasoning or a reasoning strategy based on more complex structural knowledge.

As well as allowing us to disentangle the relative effects of associative and structured knowledge on reasoning, our results replicate Shafto et al's finding that taxonomic knowledge is more available to reasoning processes than is
ecological knowledge. Although our results clearly show that associative knowledge was a better predictor of reasoning when participants were under load, they also show that participants were much more likely to generate conclusion categories that were taxonomically rather than ecologically related to the base category when under load. This was regardless of the property they were asked to reason about. Large numbers of ecologically related conclusion categories were seen only when participants reasoned about disease properties under light load.

## Conclusion

Our results suggest that apparently contradictory theories of category-based inductive reasoning best explain inference strategies under different domain-general processing conditions. People's reasoning might best be explained by associative approaches such as parallel-distributed processing connectionist accounts (Rogers \& McClelland, 2004) and featural similarity (Sloman, 1993) when they do not have time or available mental resources to engage in more elaborate reasoning.. In contrast, people with plenty of time and cognitive capacity might recruit complex structural knowledge representations (e.g. Kemp \& Tenenbaum, 2009) to derive their inferences.

## Acknowledgments

This research was funded by an ESRC postgraduate research studentship awarded to A. K. Bright.

## References

Baker, A., \& Coley, J. D. (2005). Taxonomic and ecological relations in open-ended induction. Paper presented at the 27th Annual Conference of the Cognitive Science Society.
Baraff, L., \& Coley, J. D. (2003). Thinking about music: Novice and expert inductive reasoning. Paper presented at the 25 th Annual Conference of the Cognitive-Science-Society, Jul 31-Aug 02, Boston, MA.
Bright, A.K. \& Feeney, A. (2010). The Effects of Domain and Type of Knowledge on Category-Based Inductive Reasoning. Paper presented at the Annual Conference of the Cognitive-Science-Society, Portland.
Coley, J. D., \& Baker, A. (2004). Taxonomic and Ecological Relations in Open-Ended Folk Biological Induction. Paper presented at the Fifth International Conference on Thinking.
Coley, J. D., \& Barraff, E. (2003). Effects of Time Pressure on Expert and Novice Category-Based Induction. Paper presented at the Annual Meeting of the Psychonomic Society.
Coley, J. D., Shafto, P., Stepanova, O., \& Barraff, E. (2005). Knowledge and Category-Based Induction. In W. K. Ahn, Goldstone, R. L., Love, B. C., Markman, A. B. and Wolf, P. (Ed.), Categorization inside and outside the laboratory: Essays in honor of Douglas L. Medin. Washington, D.C.: American Psychological Association
Feeney, A. (2007). How many processes underline category-based induction? Effect of conclusion specificity and cognitive ability. Memory \& Cognition, 35, 1830-1839.

Fisher, A. V., \& Sloutsky, V. M. (2005). When induction meets memory: Evidence for gradual transition from similarity-based to category-based induction. Child Development, 76(3), 583597.

Heit, E., \& Rubinstein, J. (1994). Similarity and Property Effects in Inductive Reasoning Journal of Experimental PsychologyLearning, Memory and Cognition, 20(2), 411-422.
Heylighen, F. (2001). Mining Associative Meanings from the Web: from Word Disambiguation to the Global Brain. Standaard Publishers.
Kemp, C., \& Tenenbaum, J. B. (2009). Structured Statistical Models of Inductive Reasoning. Psychological Review, 116(1), 20-58.
Lopez, A., Atran, S., Coley, J. D., Medin, D. L., \& Smith, E. E. (1997). The tree of life: Universal and cultural features of folkbiological taxonomies and inductions. Cognitive Psychology, 32(3), 251-295
Medin, D. L., Coley, J. D., Storms, G., \& Hayes, B. K. (2003). A relevance theory of induction. Psychonomic Bulletin \& Review, 10(3), 517-532
Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., \& Shafir, E. (1990). Category-based induction. Psychological Review, 97(2), 185-200.
Proffitt, J. B., Coley, J. D., \& Medin, D. L. (2000). Expertise and category-based induction. Journal of Experimental PsychologyLearning Memory and Cognition, 26(4), 811-828.
Rehder, B. (2009). Causal-Based Property Generalization. Cognitive Science, 33(3), 301-344
Rips, L. J. (1975). Inductive judgments about natural categories. Journal of Verbal Learning and Verbal Behavior, 14(6), 665681.

Rogers, T. T., \& McClelland, J. L. (2004). Semantic cognition: a parallel distributed processing approach. London: MIT.
Shafto, P., \& Coley, J. D. (2003). Development of categorization and reasoning in the natural world: Novices to experts, naive similarity to ecological knowledge. Journal of Experimental Psychology-Learning Memory and Cognition, 29(4), 641-649.
Shafto, P., Coley, J. D., \& Baldwin, D. (2007). Effects of time pressure on context-sensitive property induction. Psychonomic Bulletin \& Review, 14(5), 890-894.
Shafto, P., Kemp, C., Bonawitz, E. B., Coley, J. D., \& Tenenbaum, J. B. (2008). Inductive reasoning about causally transmitted properties. Cognition, 109(2), 175-192.
Sloman, S. A. (1993). Feature-Based Induction. Cognitive Psychology, 25(2), 231-280.
Sloutsky, V. M., \& Fisher, A. V. (2004). Induction and categorization in young children: A similarity-based model. Journal of Experimental Psychology-General, 133(2), 166-188.
Sloutsky, V. M., \& Fisher, A. V. (2008). Attentional learning and flexible induction: How mundane mechanisms give rise to smart behaviors. Child Development, 79(3), 639-651.
Smith, E. R., \& DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. Personality and Social Psychology Review, 4(2), 108-131.
Tenenbaum, J. B., Griffiths, T. L., \& Kemp, C. (2006). Theorybased Bayesian models of inductive learning and reasoning. Trends in Cognitive Sciences, 10(7), 309-318.

# Motor Asymmetries Predict Neural Organization of Emotion 

Geoffrey Brookshire ${ }^{1}$, Cleve Graver ${ }^{1}$, and Daniel Casasanto ${ }^{1,2}$<br>\{broog731; gravc243; casasand\} @ newschool.edu<br>${ }^{1}$ Department of Psychology, The New School for Social Research, New York<br>${ }^{2}$ Donders Center for Brain, Cognition, and Behaviour, Nijmegen, NL


#### Abstract

According to decades of research in affective neuroscience, approach and avoidance motivation are supported by the left and right hemispheres, respectively. With the Sword and Shield Hypothesis (SSH), we challenge this conclusion, and propose a novel principle underlying the organization of emotion in the brain: the hemispheric lateralization of motivation depends on the neural locus of motor control for the dominant hand (used preferentially for approach actions) and the non-dominant hand (used preferentially for avoidance actions). The SSH predicts that the laterality of approach motivation should vary continuously with the laterality of circuits used for planning and executing approach-related actions. To test this prediction, we measured mood before and after 5 sessions of tDCS applied bilaterally to DLPFC in right- and lefthanders. Results in right-handers show that positive emotions increased after left-excitatory stimulation, but decreased after right-excitatory stimulation. In non-right-handers, however, the opposite pattern was found: Positive emotions decreased after left-excitatory stimulation, but increased after right-excitatory stimulation. These findings reveal continuous covariation between the neural systems for action and emotion, supporting the SSH.


Keywords: Emotion; motivation; motor control; handedness; hemispheric specialization, tDCS.

## Introduction

A cornerstone of cognitive-affective neuroscience is the robust finding that the left hemisphere is specialized for approach motivation and the right hemisphere for avoidance motivation (reviewed in Harmon-Jones, Gable, \& Peterson, 2010). Although temporal and parietal areas have been implicated in affective motivation (Amodio, Master, Yee, \& Taylor, 2008; Brookshire \& Casasanto, 2012), studies suggest that this asymmetry centrally involves dorsolateral prefrontal cortex (DLPFC): Approach motivation recruits left DLPFC, and avoidance motivation recruits right DLPFC (Berkman \& Lieberman, 2010).

Approach/avoidance asymmetries can also be observed in behavior. People tend to perform approach actions with the dominant hand, and avoidance actions with the non-dominant hand (Casasanto, 2009). The dominant hand, for example, is preferred when eating. In contrast, people reflexively protect their faces with the non-dominant hand when startled (Coren, 1992). Sword-fighters in centuries past approached their enemies with a sword held in the dominant hand, and avoided incoming blows with a shield in the non-dominant hand (Harris, 2010).

In right-handers, therefore, the left hemisphere is involved both in approach motivation and in coordinating actions with the hand preferred for approach actions. Casasanto (2009) proposed that this correspondence may result from a functional relationship between affective motivation and manual
motor control for approach and avoidance actions. We call this the Sword and Shield Hypothesis (SSH; Brookshire \& Casasanto, 2012). The SSH suggests that the lateralization of approach/avoidance motivation is functionally linked to the way we use our hands to perform approach and avoidance actions, and offers a new principle to predict and explain the neural organization of emotion.

The SSH predicts that differences in how people use their hands for approach and avoidance actions should correspond to differences in the neural organization of affective motivation. Left-handers tend to perform approach actions with their left hands and avoidance actions with their right hands (Coren, 1992; Harris, 2010). Accordingly, approach motivation in left-handers should be lateralized to the right hemisphere, the reversal of the pattern found in right-handers.

To test this prediction, Brookshire and Casasanto (2012) examined resting activation asymmetries in EEG as a function of manual motor asymmetries and trait approach motivation. As in previous studies (Sutton \& Davidson, 1997), higher approach motivation in right-handers correlated with greater leftward activation asymmetries. This pattern reversed in left-handers, however. Consistent with the SSH, increased approach motivation in left-handers correlated with greater rightward activation asymmetries (Brookshire \& Casasanto, 2012).

## Causal role of frontal asymmetry

Frontal asymmetries are widely believed to play a causal role in determining emotional experience (Harmon-Jones et al., 2010). Supporting this idea, patients with left hemisphere lesions are prone to depression, whereas those with right hemisphere lesions are prone to mania and indifference to their injuries (Robinson, Boston, Starkstein, \& Price, 1988). Similarly, deactivating the right hemisphere with sodium amobarbital causes laughter and elation, whereas deactivating the left hemisphere causes crying and negative affect (Lee, Loring, Meader, \& Brooks, 1990). These data underscore the necessity of the two hemispheres in emotion, but they are limited by low spatial resolution (constrained to the level of hemisphere), and by the extremity of processing disruptions used (completely deactivating a neural area). Would subtler, more spatially restricted manipulations of activation asymmetries influence emotional processing?

Allen, Harmon-Jones, and Cavender (2001) used biofeedback in EEG to train participants to induce rightward or leftward frontal activation asymmetries. Participants trained to produce leftward asymmetries experienced greater positive, approach-oriented emotions than those with rightward train-
ing. However, it is possible that participants used strategies to complete the biofeedback task (e.g. rehearsing approachmotivated memories), complicating inferences about a causal role of frontal activation in producing emotions. Activation asymmetries may have been a consequence-not a cause-of changes in emotional state.

Subsequent work has addressed the causal role of frontal asymmetries using transcranial direct current stimulation (tDCS). In tDCS, a weak, constant electrical current is passed between two conductive electrodes on the scalp. After 5-20 minutes of stimulation, neurons beneath the anodal electrode are transiently excited, and those beneath the cathode are inhibited (Nitsche \& Paulus, 2000).

Several studies have used tDCS to intervene on activation asymmetries in DLPFC, but many of these fail to find support for a causal involvement of frontal asymmetries in emotional experience. Specifically, tDCS applied in a single session to DLPFC had no effect on self-reported mood or trait motivational tendencies (Koenigs, Ukueberuwa, Campion, Grafman, \& Wassermann, 2010; Plazier, Joos, Vanneste, Ost, \& De Ridder, 2011; Nitsche et al., 2012).

Studies of the causal role of frontal asymmetries on more implicit emotional processing have yielded mixed results. Nitsche et al. (2012) found that anodal tDCS over left DLPFC facilitated identification of both positive and negative facial expressions, but that the effect was stronger for positive expressions. Penolazzi et al. (2010) examined memory for emotional pictures after bilateral tDCS over DLPFC. Anodal tDCS almost invariably improves memory performance (for review see Jacobson, Koslowski, \& Lavidor, 2011). Surprisingly, however, left-anodal/right-cathodal stimulation facilitated recall of unpleasant pictures, and right-anodal/leftcathodal stimulation facilitated recall of pleasant pictures. Although this experiment seems to suggest a causal role of frontal asymmetries in emotional memory, the fact that leftexcitatory tDCS improved recall of negative pictures, and right-excitatory of positive pictures, is inconsistent with a great deal of research in right-handers linking positive emotions with the left hemisphere (Harmon-Jones et al., 2010).

Researchers noting the clinical potential of neurostimulation have begun using repeated sessions of tDCS to treat major depressive disorder, with somewhat more consistent results than the single-session studies reviewed above. Left-anodal tDCS often ameliorates symptoms of depression (reviewed in Murphy, Boggio, \& Fregni, 2009). However, these treatment-oriented studies did not include all of the experimental conditions needed to support the conclusion that induced activation asymmetries play a causal role in emotional experience: Researchers delivered only left-excitatory and sham stimulation, but never right-excitatory stimulation. Thus, it is not possible to conclude that increasing activity in the left hemisphere relative to the right was responsible for the positive effects of neurostimulation of mood. Perhaps multiple sessions of tDCS may boost positive mood regardless of stimulation montage. To determine whether frontal
activation asymmetries play a functional role in determining emotional states, experiments must apply both left- and rightexcitatory stimulation. Furthermore, in order for the findings to be generalizable beyond a clinical population, relationships between lateralized stimulation and mood would need to be shown in healthy participants.

## Parametric covariation in brain and behavior

According to the SSH, there is a functional relationship between the neural circuits for motivation and manual motor asymmetries. In addition to predicting a reversal in approach/avoidance lateralization in left-handers, the SSH predicts that parametric variation in manual motor asymmetries should correlate with graded differences in the lateralization of motivation. Strong right-handers, that is, should show stronger left-lateralization of approach motivation than weak right-handers. Previous work has not tested this prediction.

## The present experiment

In this study, we measure mood before and after 5 sessions of tDCS applied bilaterally to DLPFC. We analyze changes in emotional state as a function of participants' manual motor asymmetries, and whether they received left-anodal/rightcathodal or right-anodal/left-cathodal stimulation. In doing so, we test for a causal role of frontal activation asymmetries in determining emotional state. Furthermore, we test for the graded relationship between motor control and the lateralization of affective motivation predicted by the SSH.

## Method

## Participants

Participants $(\mathrm{N}=30)$ were recruited from the New School community, postings to the website www.craigslist.org/, and a database of participants who have taken part in other studies in our lab. To ensure that the sample included participants with the full range of handedness asymmetries, we selectively contacted left-handed and ambidextrous participants from the database. These participants were not aware that they were being contacted based on their handedness.

Several exclusion criteria were followed to ensure participants' safety. Respondents were not included in the study if they indicated that they were pregnant, had ever experienced an epileptic seizure, had ever sustained a stroke or other brain injury, or were taking any psychoactive drugs or medications. Additionally, we did not test anyone who reported ever having been diagnosed with depression, bipolar disorder, anxiety disorder, or schizophrenia.

One participant was canceled during the first session when a low impedance could not be obtained. Four additional participants did not complete the study (Right-excitatory stimulation, $\mathrm{N}=2$; Left-excitatory stimulation, $\mathrm{N}=2$ ), one of whom returned to complete the final day of data collection. Data were analyzed from the remaining 25 participants (Right-handers, EHI $\geq 40: \mathrm{N}=17$; Non-right-handers, EHI $<40: \mathrm{N}=8$ ). Demographics such as age and gender were not collected.

## Procedure

This study took place over five consecutive days (MondayFriday). Informed consent was obtained at the beginning of each session, and participants were payed at the end of every session. On day 1, participants completed an untimed, computerized version of the Positive and Negative Affect Scale (PANAS; Watson \& Clark, 1994). Emotion words appeared on the screen one at a time, and participants rated the degree to which they had experienced that emotion "during the past few days" on a scale of one ("very slightly or not at all") to five ("extremely") by pressing the numbers $1-5$ on a computer keyboard. To assess handedness, participants completed the Edinburgh Handedness Inventory (EHI; Oldfield, 1971). This scale offers a continuous measurement of handedness, in which scores vary from strongly left-handed ( -100 ) to strongly right-handed (100).

On days 2-4, tDCS was applied after ensuring that participants had not experienced any discomfort after the previous sessions. After applying tDCS on day 5, the same tests were performed as on day 1. After the first cohort of 7 participants, we began collecting EHI at day 5. Participants also completed a brief adverse effects questionnaire. Upon completing the study, participants were debriefed and encouraged to contact the experimenter if they had any further questions or experienced any discomfort. ${ }^{1}$

## Transcranial Direct Current Stimulation

Direct current stimulation was delivered using a batterypowered stimulator (Soterix Medical, New York) with two $5 \times 7 \mathrm{~cm}$ saline-soaked sponges covering the electrodes. New sponges were used for each session. In each session, a current was applied at 2 mA for 20 min . To minimize discomfort, the current slowly ramped between 0 and 2 mA when powering on and off. Stimulation was delivered bilaterally above DLPFC at F3-4 in the 10-20 system (DaSilva, Volz, Bikson, \& Fregni, 2011). An experimenter was in the room with the participant at all times to ensure that stimulation remained comfortable.

Stimulation was delivered double-blindly in two betweensubjects conditions. Before beginning the study, a confederate set a polarity-blinding box to either reverse the polarity of the outgoing wires, or leave polarity unchanged, and then sealed the box. This allowed both the experimenter and the participant to remain blind to the stimulation condition. Participants were randomly assigned to one of the two conditions.

In one condition, the anode was placed above F3 and the cathode above F4, exciting left frontal areas while inhibiting right frontal areas (Left-excitatory). In the second condi-

[^23]tion, the anode was placed above F4 and the cathode above F3, exciting right while inhibiting left frontal areas (Rightexcitatory). Stimulation condition remained the same across all 5 sessions. Of the participants retained in the final analysis, $\mathrm{N}=10$ were given right-excitatory stimulation, and $\mathrm{N}=15$ left-excitatory stimulation.

## Results

## Adverse effects

One participant canceled the study due to a persistent headache, and three further participants requested that the intensity be reduced for several minutes in one session. Of the four participants reporting discomfort, two had received leftexcitatory stimulation, and two right-excitatory stimulation. No other subjects reported significant discomfort.

## Manual motor asymmetries

To examine whether tDCS altered manual motor asymmetries, we compared EHI scores before and after stimulation. EHI scores on days 1 and 5 were strongly correlated ( $r=.98$ ). Change in EHI scores did not significantly depend on tDCS polarity (Welch's $t(12.3)=-1.50, p=.16)$.

## Emotional state

PANAS responses were analyzed using linear mixed-effects regressions fit by maximum likelihood in R ( R Core Team, 2012) with the lmer () function in the lme 4 library (Baayen, Davidson, \& Bates, 2008). Change in each emotion (day 5 day 1) was modeled as a function of valence (Positive; Negative), tDCS polarity (Left-excitatory; Right-excitatory), and handedness (entered continuously using EHI score collected on day 1). Random intercepts were included for Subjects and Items (i.e. emotion words). All categorical predictors were entered using deviation coding. Unless otherwise noted, pvalues and 95\% Highest Posterior Density intervals (HPD) of the parameter estimates were estimated using Markov chain Monte Carlo (MCMC) sampling with 20,000 samples using the pvals.fnc () function in the languageR library.

Of primary interest, handedness, valence, and tDCS polarity interacted to predict change in PANAS ratings ( $\beta=$ $-0.015, \mathrm{HPD}=[-0.022,-0.009], p=.0001)$. As evident from Figure 1, this interaction was driven primarily by strong effects of handedness and polarity on positive emotions, but not on negative emotions. Separate mixed-effects regressions with positive and negative items support this conclusion.

Change in negative emotions did not significantly depend on handedness, tDCS polarity, or their interaction (all $p \mathrm{~s}>.5$; Fig. 1a). In contrast, handedness significantly interacted with tDCS polarity to predict change in the intensity of positive emotions $(\beta=-0.016$, HPD $=$ [ $-0.021,-0.010], p=.0001$; Fig. 1b). In participants receiving left-excitatory stimulation, stronger right-handedness correlated with greater increases in positive emotions ( $\beta=$ $0.011, \mathrm{HPD}=[0.0064,0.016], p=.0001)$. In those receiving right-excitatory stimulation, the opposite pattern


Figure 1: Change in (a) negative and (b) positive emotion from day 1 to day 5 as a function of manual motor asymmetries (EHI). Each point illustrates Z-transformed average change across all PANAS items for a single participant. Best-fit lines are plotted for each stimulation condition. Left-excitatory $=$ anode left, cathode right; Right-excitatory $=$ cathode left, anode right.
was observed: stronger right-handedness correlated with decreases in positive emotions $(\beta=-0.0044$, HPD $=$ [ $-0.0080,-0.0006], p=.02$ ).

By defining handedness categorically, we examined differences between right- and non-right-handers in the effects of tDCS on emotional state. For right-handers, left-excitatory tDCS led to more positive emotions than with rightexcitatory tDCS $(\beta=-0.21, \mathrm{HPD}=[-0.41,-0.012], p=$ $.04)$. For non-right-handers, the opposite pattern emerged: left-excitatory tDCS caused decreases in positive emotions compared with right-excitatory tDCS $(\beta=0.55, \mathrm{HPD}=$ [0.19, 0.89], $p=.006$ ).

These regression analyses leave open the question of whether parametric variation in handedness corresponds to graded differences in the hemispheric lateralization of emotion; significant parameter estimates in linear regressions can be caused by either continuous covariation or a stepfunction. Rank-order tests, however, can be used to discriminate between categorical and continuous relationships. A significant Spearman's correlation revealed that stronger right-handedness was continuously related to greater increases in positive emotions in participants who received leftexcitatory stimulation $(\rho(13)=0.71, p=.003)$. In those who received right-excitatory stimulation, this correlation was marginally significant in the opposite direction: stronger right-handedness continuously predicted greater reductions in positive emotions ( $\rho(8)=-0.56, p=.09$ ).

## Discussion

The effects of tDCS on mood depend upon the hemisphere to which excitatory stimulation is applied and the handedness of the participant. In right-handers, five sessions of left-excitatory (left-anodal, right-cathodal) tDCS led to
increased positive emotions, whereas right-excitatory (leftcathodal, right-anodal) tDCS led to decreased positive emotions. Non-right-handers, by contrast, showed the opposite pattern, with right-excitatory tDCS increasing positive emotions and left-excitatory tDCS decreasing them. Furthermore, we find graded, parametric variation between manual motor asymmetries and emotion in the brain. Stronger motor asymmetries correlate with more strongly lateralized circuits for emotion. These results demonstrate a functional relationship between activation asymmetries in the frontal lobes and the experience of positive emotions, and show that the laterality of positive emotion covaries continuously with the laterality of manual motor control.

According to the motivational model of hemispheric specialization for emotion, approach motivation is lateralized to the left hemisphere and avoidance to the right (Harmon-Jones et al., 2010). In conflict with this model, we show that neural regions specializing in approach motivation are co-lateralized with circuits that control the dominant hand. This finding is consistent with the SSH, which proposes a functional relationship between motivation and manual action.

Manual motor asymmetries predict the way approach motivation is distributed across the two hemispheres. For the right-handed majority, this appears as left-lateralized approach motivation. Does this mean that the classic motivational model is mostly correct-that it is right for the approximately $90 \%$ of people who are right-handed, and only wrong for the other $10 \%$ ? We suggest the answer is no: As a field, we have arrived at incorrect generalizations about the cortical basis of emotions. It is not the case that "anterior regions of the left and right hemispheres are specialized for approach and withdrawal processes, respectively" (Davidson, 1992, p. 127). It is only incidentally true that the left hemi-
sphere is specialized for approach motivation in most of the people who have been tested. This specialization is not due to any functional properties of the left hemisphere, per se. It appears that any theory that assigns a privileged role to the left hemisphere in processing approach motivation is incorrect.

These findings may help to elucidate an enduring mystery in affective neuroscience: What role do activation asymmetries play in motivation? Although no clear consensus has emerged, some researchers believe that leftward asymmetries may reflect "expression of approach-related emotions" (Harmon-Jones, 2004, p. 55) or "approach-related, goaldirected action planning" (Davidson, 2004, p. 225). By highlighting the close connection between action and emotion, our findings suggest that leftward asymmetries are closely linked to performance of approach actions.

## Causal links between frontal asymmetries, motivation, and hand action

To our knowledge, these findings provide the first unequivocal evidence that frontal activation asymmetries casually influence emotional experience in healthy participants. However, this study leaves open the question of the causal relationship between neural circuits for motivation and for motor control of the hands. We consider three possibilities.

First, handedness could determine the laterality of motivation. In this case, handedness is assumed to be set by some combination of genetic and environmental influences. If approach actions require greater dexterity than avoidance actions, then habits could develop in which approach actions are performed by the more adept dominant hand. These habits could then stabilize on an evolutionary or a developmental timescale, causing cortical areas involved in planning actions with the dominant hand to specialize in approach actions.

Second, the laterality of motivation could determine handedness. In this case, the laterality of motivation is assumed to be determined by genetic and environmental factors. Manual action circuits ipsilateral to regions specializing in approach motivation may subsequently come to be used preferentially for approach actions. If approach actions are more frequent or require more skill than avoidance actions, then dexterity may be enhanced in the hand used to perform them.

Third, handedness and the laterality of motivation could be determined by a common factor. In this case, there would be no direct causal link between the lateralization of neural circuits of motivation and manual motor control. Any proposed third factor would need to account for the close covariation we observe.

## Valence and motivational direction

We find that positive emotions are strongly modulated by induced frontal activation asymmetries, whereas negative emotions are unaffected. If negative emotions are assumed to be the mirror image of positive emotions, this result seems incongruous. This apparent contradiction resolves when examining the motivational direction of the words in the positive and negative PANAS subscales.

The left and right hemispheres appear to be differentially specialized for motivational direction, not valence (Berkman \& Lieberman, 2010; Harmon-Jones et al., 2010). Although these dimensions are highly correlated, they can also be dissociated. Induced frontal asymmetries, then, should alter the motivational direction-but not necessarily the valence-of participants' mood.

The emotions comprising the positive PANAS subscale uniformly involve strong approach motivation (active, alert, attentive, determined, enthusiastic, excited, inspired, interested, proud, strong). The negative subscale, on the other hand, is more varied. Some items seem to involve avoidance motivation (afraid, scared, ashamed), some approach motivation (hostile, irritable), and some do not have any clear motivational direction (nervous, jittery, guilty, upset, distressed).

In summary, we find that an emotion category with a consistent motivational direction (the positive subscale) is influenced by manipulations of frontal activation asymmetries, whereas a more heterogenous emotion category (the negative subscale) is not affected. Further studies must determine if avoidance motivation can be similarly modulated by induced activation asymmetries.

## Clinical implications

Neurostimulation techniques such as tDCS and transcranial magnetic stimulation (TMS) are currently in use as treatments of major depressive disorder (Murphy et al., 2009). By increasing activation in left frontal areas, clinicians hope to augment positive, approach-oriented emotions, alleviating depression. This treatment is predicated on the assumption that the left hemisphere is specialized for approach motivation. We provide evidence against this assumption. Hemispheric specialization for motivation reverses in many people, including left-handers (see also Brookshire \& Casasanto, 2012), who are at increased risk for depression (Denny, 2009).

Systematic differences in the neural organization of motivation may have urgent consequences for the success and safety of neurostimulation therapies. We show that positive affect is reduced after anodal tDCS to the hemisphere that controls the non-dominant hand. This result suggests that FDA-approved treatments involving anodal tDCS to leftDLPFC may exacerbate depression in non-right-handers.

## Conclusions

Accepted theories of emotion in the brain hold that the left hemisphere is specialized for approach motivation, and the right for avoidance motivation. We provide evidence against this "motivation model" and in support of the SSH. Hemispheric lateralization for emotion covaries with manual motor asymmetries, consistent with a causal relationship between motivation and motor control. The SSH proposes a principle by which the hemispheres become specialized for approach and avoidance states, and may lead not only to a better understanding of how motivation is organized in the cerebral cortex, but also of why it is organized that way.

## Acknowledgments

Research was funded by a James S. McDonnell Foundation Scholar Award (\#220020236; http://www.jsmf.org) to D.C.

## References

Allen, J., Harmon-Jones, E., \& Cavender, J. (2001). Manipulation of frontal EEG asymmetry through biofeedback alters self-reported emotional responses and facial EMG. Psychophysiology, 38(4), 685-693.
Amodio, D. M., Master, S. L., Yee, C. M., \& Taylor, S. E. (2008). Neurocognitive components of the behavioral inhibition and activation systems: implications for theories of self-regulation. Psychophysiology, 45(1), 11-9.
Baayen, R. H., Davidson, D. J., \& Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language, 59(4), 390-412.
Berkman, E. T., \& Lieberman, M. D. (2010). Approaching the Bad and Avoiding the Good: Lateral Prefrontal Cortical Asymmetry Distinguishes between Action and Valence. Journal of Cognitive Neuroscience, 22(9), 1970-1979.
Brookshire, G., \& Casasanto, D. (2012). Motivation and Motor Control: Hemispheric Specialization for Approach Motivation Reverses with Handedness. PLoS ONE, 7(4), e36036.
Carver, C. S., \& White, T. L. (1994). Behavioral Inhibition, Behavioral Activation, and Affective Responses to Impending Reward and Punishment: The BIS/BAS Scales. Journal of Personality and Social Psychology, 67(2), 319-333.
Casasanto, D. (2009). Embodiment of abstract concepts: good and bad in right- and left-handers. Journal of Experimental Psychology: General, 138(3), 351-67.
Coren, S. (1992). Handedness, Traffic Crashes, and Defensive Reflexes. Public Health, 82(8), 1176-1177.
DaSilva, A., Volz, M., Bikson, M., \& Fregni, F. (2011). Electrode positioning and montage in transcranial direct current stimulation. Journal of visualized experiments: JoVE(51).
Davidson, R. (1992). Anterior cerebral asymmetry and the nature of emotion. Brain and cognition, 20(1), 125-151.
Davidson, R. (2004). What does the prefrontal cortex do in affect: perspectives on frontal eeg asymmetry research. Biological psychology, 67(1), 219-234.
Denny, K. (2009). Handedness and depression: evidence from a large population survey. Laterality, 14(3), 246-255.
Harmon-Jones, E. (2004). Contributions from research on anger and cognitive dissonance to understanding the motivational functions of asymmetrical frontal brain activity. Biological psychology, 67(1-2), 51-76.
Harmon-Jones, E., Gable, P. A., \& Peterson, C. K. (2010). The role of asymmetric frontal cortical activity in emotionrelated phenomena: A review and update. Biological Psychology, 84(3), 451-462.
Harris, L. J. (2010). In fencing, what gives left-handers the edge? Views from the present and the distant past. Laterality, 15(1/2), 15-56.

Jacobson, L., Koslowski, M., \& Lavidor, M. (2011). tDCS polarity effects in motor and cognitive domains: a metaanalytical review. Experimental Brain Research.
Koenigs, M., Ukueberuwa, D., Campion, P., Grafman, J., \& Wassermann, E. M. (2010). Bilateral Frontal Transcranial Direct Current Stimulation: Failure to Replicate Classic Findings in Healthy Subjects. Clinical Neurophysiology, 120(1), 80-84.
Lee, G. P., Loring, D. W., Meader, K. J., \& Brooks, B. B. (1990). Hemispheric specialization for emotional expression: A reexamination of results from intracarotid administration of sodium amobarbital. Brain and Cognition, 12(2), 267-280.
Murphy, D. N., Boggio, P. S., \& Fregni, F. (2009). Transcranial direct current stimulation as a therapeutic tool for the treatment of major depression: insights from past and recent clinical studies. Current Opinion in Psychiatry, 22(3), 306-311.
Nitsche, M. A., Koschack, J., Pohlers, H., Hullemann, S., Paulus, W., \& Happe, S. (2012). Effects of frontal transcranial direct current stimulation on emotional state and processing in healthy humans. Frontiers in psychiatry, 3(58).
Nitsche, M. A., \& Paulus, W. (2000). Excitability changes induced in the human motor cortex by weak transcranial direct current stimulation. The Journal of physiology, 527(3), 633-639.
Oldfield, R. C. (1971). The Assessment and Analysis of Handedness: The Edinburgh Inventory. Neuropsychologia, 9, 97-113.
Penolazzi, B., Domenico, A. D., Marzoli, D., Mammarella, N., Fairfield, B., Franciotti, R., et al. (2010). Effects of Transcranial Direct Current Stimulation on Episodic Memory Related to Emotional Visual Stimuli. PloS one, 5(5), e10623.
Plazier, M., Joos, K., Vanneste, S., Ost, J., \& De Ridder, D. (2011). Bifrontal and bioccipital transcranial direct current stimulation (tDCS) does not induce mood changes inhealthy volunteers: a placebo controlled study. Brain stimulation, 25-30.
R Core Team. (2012). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from http://www.R-project.org/
Robinson, R., Boston, J., Starkstein, S., \& Price, T. (1988). Comparison of mania and depression after brain injury: causal factors. American Journal of Psychiatry, 145(2), 172-8.
Sutton, S. K., \& Davidson, R. J. (1997). Prefrontal Brain Asymmetry: A Biological Substrate of the Behavioral Approach and Inhibition Systems. Psychological Science, 8(3), 204-210.
Watson, D., \& Clark, L. A. (1994). The PANAS-X: Manual for the Positive and Negative Affect Schedule - Expanded Form. University of Iowa.

# The Effect of Discourse Context on Online Sentence Processing 

Juliane Burmester (juliane.burmester@uni-potsdam.de)<br>Isabell Wartenburger (isabell.wartenburger@uni-potsdam.de)<br>University of Potsdam, Department of Linguistics, Center of Excellence Cognitive Science, Karl-Liebknecht-Str. 24-25, 14476 Potsdam, Germany<br>Katharina Spalek (katharina.spalek@staff.hu-berlin.de)<br>Humboldt-Universität zu Berlin, Department of German Language and Linguistics, Dorotheenstr. 24, 10117 Berlin, Germany


#### Abstract

Information structure describes how the information is packaged within a discourse to optimize information transfer. We addressed the question if and how a discourse context modulates the online processing of German declaratives. Native speakers of German read fictitious stories that depicted a simple action scene of two characters while we recorded event-related brain potentials (ERPs). Two types of discourse contexts (topic vs. neutral) were compared with regard to the processing of declarative canonical subject-before-object (SO) and non-canonical object-before-subject (OS) sentences. The preceding topic context only modulated the processing of OS sentences. This was indicated by a less pronounced positivity around 500 to 900 ms for the topic compared to the neutral context. As supported by previous research we argue that this context-induced effect in the processing of noncanonical sentences reflects reduced processing costs for the integration of the discourse relevant topic information into the current discourse model.


Keywords: information structure; discourse context integration; topic; sentence processing; word order variation; ERP; P600

## Introduction

In our everyday-life communicative utterances are typically linked to the discourse environment of the interlocutors. Previous evidence suggests that contextual information (e.g., from prior discourse) plays a crucial role in sentence comprehension. Information structure is concerned with the question how the information is structured and packaged within a discourse to optimize information transfer (e.g., Chafe, 1976). If and to which degree information structure interacts with syntax and other linguistic domains is still under debate (e.g., Büring, 2007; Fanselow \& Lenertová, 2011)

Information structure research in the domain of syntax (and in particular word order) addressed the question how word order variation might be affected by information structural concepts such as topic-comment, focusbackground, or the given-new distinction (Lambrecht, 1994; Rizzi, 1997). Topic (also aboutness topic) refers to the entity the sentence is about (Gundel, 1988; Reinhart, 1981). Mostly topics are introduced by the previous discourse (e.g., Skopeteas et al., 2006). It has been proposed that topics, independent of their grammatical function, are preferably placed in a specific syntactic position which is sentence-
initial (i.e., prefield) for German main clauses (Büring, 1999; Rosengren, 1993).

German is a language with relatively flexible word order because morphological features such as case marking allow the reordering of constituents without changing the grammatical function of the constituents. ${ }^{1}$ Therefore, German is ideal to study the effect of information structure on word order. In German, subject-before-object (SO) is the canonical word order which is preferred to object-beforesubject (OS) sentences (e.g., Gorrell, 2000). If presented without a felicitous discourse context, OS sentences lead to lower acceptability ratings and longer latencies in reading (e.g., Meng, Bader \& Bayer, 1999) compared to SO sentences. It has been pointed out that the processing of OS sentences might require contextual licensing (e.g., Bornkessel \& Schlesewsky, 2006; Höhle, 1982; Hörnig, Weskott, Kliegl \& Fanselow, 2006). Specific contextual information (e.g., object given in prior discourse or object in contrastive whole-part relation to a referent in the context) has been found to improve acceptability ratings and shorten reading times of OS sentences (Meng et al., 1999; Weskott, Hörnig \& Fanselow, 2009).

However, acceptability ratings and reading times do not tell us which underlying mechanisms of information structure help to optimize information transfer. Eventrelated potentials (ERPs) are a promising tool to shed more light on the effect of context on sentence processing (e.g., Van Berkum, 2008). For instance, a context-induced N400 effect for inferable vs. given referential expressions across different sentence positions was attributed to discourselinking processes (Schumacher \& Hung, 2012). A late positivity was associated with processing costs induced by updating the current discourse model and the integration of a new referent into discourse (e.g., Burkhart, 2007; Schumacher \& Hung, 2012).

Although many studies tested the online processing of OS sentences in isolation and indicated higher processing difficulty compared to SO (Matzke, Mai, Nager, Rüsseler \&

[^24]Münte, 2002; Rösler, Pechmann, Streb, Röder \& Hennighausen, 1998), previous behavioral and neurophysiological research has shown that contextual information (e.g., focus, givenness) is of central relevance for the processing of canonical and non-canonical sentences (e.g., Bornkessel, Schlesewsky \& Friederici, 2003; Schumacher \& Hung, 2012). However, it remains unclear if a topic introduced by the discourse context reveals an immediate processing advantage at the sentence-initial position of OS sentences.

The present ERP experiment aims to characterize the differential effect of a preceding topic context on the processing of German declaratives. We compare if and how a preceding topic context which assigns topic status to one of two characters of a scene modulates the processing of the following topic-first OS or SO sentence. The effect of the topic context is compared to a neutral context which induces a wide scope on the scene and serves as a baseline. Based on previous research we expect that the processing of the canonical SO sentences is not modulated by the preceding context information. Instead, for the non-canonical OS sentences we predict a processing advantage if they are preceded by a topic context as compared to a neutral context such that the topic status of the sentence-initial object leads to a less effortful linking and integration of information into the current discourse model. This type of processing advantage might be reflected in a reduced P600 component (e.g., Burkhart, 2007).

## Methods and Material

## Participants

Twenty-one native speakers of German participated in the present ERP experiment. All participants were right-handed according to the Edinburgh Handedness Inventory (Oldfield, 1971). None of them reported any neurological disorder. All had normal or corrected-to-normal vision. The participants were reimbursed or received course credits for participation. Due to a low behavioral accuracy ( $<60 \%$ correct) in the sentence-picture-verification task (see Procedure) the data of two participants were excluded from further analysis. Thus the ERP analysis is based on 19 participants (11 female, mean age 25 years, age range 19-30 years).

## Material

Each trial consisted of a description of a fictitious scene with two animals. Each animal was both a plausible agent and a plausible patient of the depicted action. Each trial comprised three parts starting with a lead-in context (1) in which both animals were introduced plus the instrument of the action that is going to be performed (see (1) in Table 1 for an example). Thus both referents were discourse-given in terms of information structure (Prince, 1981). The action was inferable from the instrument mentioned. The lead-in context was followed by one of two context types presented in form of a context question (2) that was either neutral indicating a wide scope on the scene ('What exactly is going on?`) or assigned topic status to one of the two animals ('What about the owl?`). The context question was followed by the target sentence (3) in SO or OS order revealing the answer to the context question.

Participants were presented with 80 SO and 80 OS target sentences that were either preceded by a neutral context or a topic context (i.e., 40 trials per condition). The different scenes for the 160 trials were created by combining two of the 40 nouns (animals, monomorphemic, masculine, 1-syllabic $(\mathrm{n}=18)$ to 2 -syllabic $(\mathrm{n}=22)$ ) with one of 10 action verbs (monomorphemic, transitive, 2-syllabic) with its corresponding instrument. The nouns and verbs were controlled for written lemma frequency, type frequency and normalized $\log _{10}$ familiarity values (dlex database: Heister et al., 2011). To avoid lexical-semantic effects of certain nouns in the first and second noun phrase position of the target sentence, each noun occurred once in each of the four conditions at both sentence positions, respectively. Thus each animal served four times as the agent and four times as the patient of the target sentence, respectively, always with a different action. For the lead-in context the first and second mention of the potential agent and patient of the action was counterbalanced across conditions. All animal pairs in the trials always differed in the initial phonemes. The 160 trials were pseudo-randomized such that maximal two consecutive trials were of the same condition or had the same word order in the target sentence to minimize possible effects of structural priming (e.g., Scheepers \& Crocker, 2004). To avoid any preferences of thematic role or topic assignment due to the previous trial at least five trials separated the repetition of an animal and at least two trials the repetition of an action.

## Procedure

Each participant was seated in a sound-attenuated cabin 90 cm in front of a computer screen and a button box (Cedrus response pad model RB-830). The trials were presented visually in the center of the screen by means of the software Presentation (www.neurobs.com). Each trial began with the presentation of a red asterisk for 1000 ms to indicate the beginning of a new scene. Before and after the lead-in context a blank screen was displayed for 200 ms . Lead-in-context and context question were presented as a whole in a self-paced reading manner with a minimum reading time of 3350 ms and 1400 ms , respectively. The participant had to press a button with the left thumb for further reading. The context question was followed by a fixation cross for 500 ms in the center of the screen. Then the target sentence was presented phrase-wise with 500 ms for each determiner phrase (DP) and prepositional phrase (PP) and 450 ms for the verb with an ISI of 100 ms (as used in previous studies, e.g., Bornkessel et al., 2003).

In $20 \%$ of the trials a sentence-picture-verification task followed the target sentence. The pictures depicted the scene of the preceding target sentence with correct or exchanged thematic roles (the owl painting the hedgehog vs. the hedgehog painting the owl). For each of the four conditions there was the same number of matching/mismatching

Table 1: Example of experimental trial for each condition (vertical bars in target sentence indicate phrase-wise presentation, Abb.: NOM = nominative case, $\mathrm{ACC}=$ accusative case, $\mathrm{S}=$ subject, $\mathrm{V}=$ verb, $\mathrm{O}=$ object, $\mathrm{PP}=$ prepositional phrase, $\mathrm{SO}=$ subject-before-object, OS = object-before-subject).

\begin{tabular}{|c|c|c|c|}
\hline (1) Lead-in context \& (2) Context question \& (3) Target sentence \& Condition <br>

\hline \multirow[t]{4}{*}{\begin{tabular}{l}
Der Uhu und der Igel haben eine Staffelei im Park aufgebaut. <br>
`The owl and the hedgehog have set up an easel in the park.

} \& 

Was ist denn genau los? <br>
‘What exactly is going on?
\end{tabular} \& \multirow[t]{2}{*}{Der Uhu | malt | den Igel | im Park. [the ${ }_{[\mathrm{NOM}]}$ owl $\left._{[\text {NOM }]}\right]_{\mathrm{s}}$ [paints] $]_{\mathrm{V}}\left[\right.$ the $_{\text {[ACC] }}$ hedgehog $\left.{ }_{[A C C]}\right]_{\mathrm{O}}[\text { in the park }]_{\mathrm{PP}}$. `The owl paints the hedgehog in the park.} \& NEUTRAL-SO <br>

\hline \& Was ist mit dem Uhu? `What about the owl?` \& \& TOPIC-SO <br>
\hline \& Was ist denn genau los? 'What exactly is going on? \& \multirow[t]{2}{*}{Den Uhu | malt | der Igel | im Park. $\left[\text { the }{ }_{[A C C]} \mathrm{owl}_{[A C C]}\right]_{0}[\text { paints] }]_{V}\left[\right.$ the ${ }_{[\text {Nom }]}$ hedgehog $\left.{ }_{[\text {Nom }}\right]_{\text {s }}$ [in the park] ${ }_{\text {pp }}$. `In the park the owl is painted by the hedgehog.`} \& NEUTRAL-OS <br>

\hline \& | Was ist mit dem Uhu? |
| :--- |
| ‘What about the owl?` | \& \& TOPIC-OS <br>

\hline
\end{tabular}

probes. The picture was presented for two seconds before the participant had to press the corresponding button for match (yes) or mismatch (no) within a time window of two seconds.
Participants were instructed to read each scene attentively and silently and to answer the sentence-picture-verification task after some of the scenes as accurately and fast as possible. The assignment of the response buttons to the right fore and middle finger was counterbalanced across participants. Participants were asked to avoid any movements during the time of sentence reading. To become familiar with the procedure participants performed three practice trials. The whole experiment included pauses after each 40 trials and lasted approximately 30 minutes.

## EEG Recording

EEG was recorded by means of a 32 channel active electrode system (Brain Products, Gilching) with a sampling rate of 1000 Hz . The electrode configuration included the following 29 scalp sites according to the international 10-20 system: F7/8, F5/6, F3/4, FC3/4, C5/6, C3/4, CP5/6, P3/4, P7/8, PO3/4, FPz, AFz, Fz, FCz, Cz, CPz, Pz, POz, Oz. To detect blinks and vertical eye movements an electrooculogram was monitored by one electrode under and one electrode above the right eye. The left mastoid served as the reference electrode but the recording was re-referenced to bilateral mastoids offline. The ground electrode was placed at FP1. Impedance was kept below 5 kOhm .

## ERP data analysis

The raw data were filtered by applying the Butterworth zero phase filter with a 0.3 Hz low cutoff and 70 Hz high cutoff (slope: $12 \mathrm{~dB} / \mathrm{oct}$ ) and a Notch Filter of 50 Hz . An automatic artifact rejection was applied to reject blinks and drifts in the time window of -200 to 1700 ms before and after the onset of the target sentence (rejection criteria: max. voltage step of $30 \mu \mathrm{~V} / \mathrm{ms}$, max. $200 \mu \mathrm{~V}$ difference of values in interval). On average $5.43 \%$ of the trials per condition had to be excluded from the analysis. The rejections were equally distributed across the conditions. For the correction of vertical eye movements the algorithm by Gratton, Coles
\& Donchin (1983) was used. Baseline correction was applied for 200 ms before the onset of the target sentence.
Time locked to the onset of the target sentence, mean amplitude values of the ERPs per condition were analyzed within three time windows ( $100-300 \mathrm{~ms}, 300-500 \mathrm{~ms}$ and $500-900 \mathrm{~ms}$ ) based on visual inspection and according to the current literature on language related ERP components (i.e., N400, P600). The following regions of interest (ROIs) were analyzed via mean amplitudes of the three appropriate electrodes: left anterior (F5, F3, FC3), left central (C5, C3, CP5), left posterior (P3, P7, PO3), right anterior (F6, F4, FC4), right central (C6, C4, CP6), right posterior (P4, P8, PO4), anterior midline ( $\mathrm{FPz}, \mathrm{AFz}, \mathrm{Fz}$ ), central midline ( FCz , $\mathrm{Cz}, \mathrm{CPz}$ ), posterior midline ( $\mathrm{Pz}, \mathrm{POz}, \mathrm{Oz}$ ). For statistical ERP analysis a repeated measures ANOVA was applied for the three within-subject factors WORD ORDER (SO, OS), CONTEXT TYPE (TOPIC, NEUTRAL) and ROI (nine levels) using SPSS Statistics (version 21). The correction according to Greenhouse and Geisser (1959) was applied. We report the corrected F - and p -values but the original degrees of freedom. Only significant main effects and interactions ( $\mathrm{p}<.05$ ) including the factors CONTEXT TYPE and/or WORD ORDER are reported and resolved by using paired T-Tests. Note that we only compare context effects on the very same sentence structures, that is, we compare SO with SO sentences and OS with OS sentences, depending on their preceding CONTEXT TYPE (TOPIC vs. NEUTRAL).

## Results

Figure 1 displays the grand average ERPs for the factor CONTEXT TYPE (TOPIC vs. NEUTRAL) at the onset of the SO and OS target sentences. The statistical analysis of the ERPs in three different time windows revealed the following results:

## Time window $100-300 \mathrm{~ms}$

Statistical analysis in the time window 100 to 300 ms after onset of the target sentence revealed a statistically significant main effect of CONTEXT TYPE
$[F(1,18)=5.29, \mathrm{p}<.05]$ reflected by less positive going ERP amplitudes following the topic context relative to the neutral context (see Figure 1, panel A and B).

## Time window 300-500 ms

The ANOVA in the time window of 300 to 500 ms after the onset of the target sentence neither revealed statistically significant main effects nor interactions.


Figure 1: Grand average ERPs of electrode FC4 as an example from the right anterior ROI time-locked to the onset of the target sentences showing the effect of CONTEXT TYPE (TOPIC vs. NEUTRAL) for SO and OS sentences. For presentation purposes the displayed ERP-plots are 7 Hz low-pass filtered. Negativity is plotted upwards.

## Time window 500-900 ms

In the late time window of 500 to 900 ms after the onset of the target sentence the ANOVA revealed a statistically significant interaction of WORD ORDER $x$ ROI $[\mathrm{F}(8,144)=4.09, \mathrm{p} \leq .01]$, WORD ORDER x CONTEXT TYPE $[F(1,18)=4.84, \mathrm{p}<.05]$ as well as of WORD ORDER x CONTEXT TYPE x ROI $[F(8,144)=4.29$, $\mathrm{p} \leq .01]$. Follow-up analysis of the three-way interaction reached significance for OS sentences in the right anterior ROI $[t(18)=-2.20, p=<.05]$ : OS sentences revealed a less
pronounced positivity in case of the preceding TOPIC context compared to the NEUTRAL context. SO sentences did not reveal a statistically significant difference depending on the CONTEXT TYPE in any of the nine ROIs.

## Summary of ERP results

The statistical analysis showed that the factor CONTEXT TYPE significantly interacted with WORD ORDER. In the late time window ( 500 to 900 ms ) the preceding context affected the processing of otherwise identical OS sentences: The topic context lead to a less pronounced positivity as compared to the neutral context. For SO sentences no such difference was induced by the preceding context. Besides, the analysis revealed a main effect of the preceding CONTEXT TYPE in the early time window ( 100 to 300 ms after target sentence onset) such that the topic context induced a reduced positivity compared to the neutral context.

## Discussion

The present ERP experiment addressed the question if and how a preceding topic context modulates the online processing of German declaratives. In line with previous research the preceding topic context did not affect the processing of SO sentences because SO is the canonical word order in German sentences. Importantly, for the processing of OS sentences we found an impact of the topic context reflected in a reduced late positivity ( 500 to 900 ms ) in comparison to a neutral context.

Besides the late positivity, the early positivity (100 to 300 ms ) was modulated by the context type independent of the word order of the target sentence: Sentences following the topic context showed a reduced positivity compared to the neutral context. Note that in the topic context condition the sentence-final noun of the context sentence is repeated in the sentence-initial position of the target sentence, whereas no such repetition occurred in the neutral context condition. As rather early ERP components have commonly been linked to basic visual processes (e.g., Dunn, Dunn, Languis \& Andrews, 1998), the reduced P2-like response in our study might be attributable to a pure word repetition effect (e.g., Van Petten, Kutas, Kluender, Mitchiner \& McIsaac, 1991). Although recent findings report modulations of an early positivity by contextual information in terms of the integration of semantic information (e.g., Federmeier \& Kutas, 2001; Lee, Liu \& Tsai, 2012) or the integration of discourse relevant information comparing a focus vs. neutral context (Bornkessel et al., 2003) such an interpretation of the early positivity in our study is not eligible due to the chosen experimental design.
As expected, the late positivity in the time window of 500 to 900 ms (P600) was reduced only for OS sentences following the topic context relative to the neutral context. In line with recent data this reduced P600 might indicate lower processing efforts for updating the current discourse model (e.g., Burkhart, 2007; Schumacher \& Hung, 2012) or structural re-analysis (as suggested for instance by the
neurocognitive model of sentence processing by Friederici (2002)). Thus, the chosen topic context elicited a processing advantage for the non-canonical OS sentences such that the integration of the sentence-initial object was facilitated compared to the neutral context in which a wide scope on the scene with the discourse-given referents and action was induced.
Our interpretation of the reduced P600 in OS sentences as reflecting lower processing efforts for the sentence-initial topic compared to a preceding neutral context is in line with the results of a follow-up study using the same experimental material and design as in the ERP experiment combined with a categorical judgment task on the comprehensibility of each story instead of the sentence-picture-verification task. Across 28 participants the mean percentage of stories judged as easily comprehensible was $90.71 \%$ for the condition NEUTRAL-SO, 88.93 \% for TOPIC-SO, 34.82 \% for NEUTRAL-OS and 51.79 \% for TOPIC-OS. Statistical analysis using a linear mixed effects model revealed a significant main effect of CONTEXT TYPE ( $z=3.13$, $\mathrm{p}<.01$ ) and WORD ORDER ( $\mathrm{z}=-7.41, \mathrm{p}<.001$ ) and a significant CONTEXT TYPE x WORD ORDER interaction ( $\mathrm{z}=-2.53, \mathrm{p}<.01$ ). Post-hoc comparisons showed that stories containing OS target sentences were significantly more likely to be judged as easily comprehensible if presented together with the topic context ( $\mathrm{z}=3.22, \mathrm{p}<.01$ ), whereas the context type did not affect the comprehensibility of the canonical SO sentences $(z=0.40$, p > .05).

Notably, other than expected we did not see a modulation of the N400 component in our data, neither in SO nor in OS sentences (see e.g., Schumacher \& Hung, 2012). Moreover, the effect of canonicity which was reported in form of a negativity at around 400 ms for OS vs. SO sentences in some (e.g., Matzke et al., 2002) but not in other studies (e.g., Frisch, Schlesewsky, Saddy \& Alpermann, 2002) was not present in our experiment. This might be due to the rather simple sentences used or due to the fact that $50 \%$ of the presented sentences were OS sentences, so the initial preference for SO might have been "overwritten" by our experimental design.

In summary, our findings suggest that the topic assigning contextual information used in the present experiment seemed to play a crucial role just in the processing of noncanonical OS sentences. The processing of OS sentences was modulated in terms of lower processing costs for the integration of discourse relevant information induced by the preceding topic context relative to a preceding neutral context. For the processing of canonical SO sentences that have been known to be felicitous in the absence of a supportive context no impact of the preceding discourse context was found. Hence our data indicate that the online processing of a sentence-initial object is enhanced by a topic assigning contextual discourse through an easier integration of discourse relevant information into the current discourse model.

## Acknowledgments

This research was supported by the German Research Foundation (DFG) as part of the Collaborative Research Center SFB 632 "Information structure". IW is supported by the Stifterverband für die Deutsche Wissenschaft (Claussen-Simon-Stiftung). We thank Jan Ries and Franziska Machens for assistance in data acquisition and analysis.

## References

Bornkessel, I., \& Schlesewsky, M. (2006). The role of contrast on local licensing of scrambling in German: Evidence from online comprehension. Journal of Germanic Linguistics, 18(1), 1-43.
Bornkessel, I., Schlesewsky, M., \& Friederici, A. D. (2003). Contextual information modulates initial processes of syntactic integration: The role of inter- vs. intra-sentential predictions. Journal of Experimental Psychology: Learning, Memory, and Cognition, 29, 871-882.
Burkhart, P. (2007). The P600 reflects cost of new information in discourse memory. Neuroreport, 18, 18511854.

Büring, D. (1999). Topic. In P. Bosch \& R. van der Sandt (Eds.), Focus: Linguistic, cognitive and computational perspectives. Cambridge: Cambridge University Press.
Büring, D. (2007). Semantics, intonation and information structure. In G. Ramchand \& C. Reiss (Eds.), The Oxford Handbook of Linguistic Interfaces, Oxford: University Press.
Chafe, W. L. (1976). Givenness, contrastiveness, definiteness, subjects, topics and point of view. In Charles, N. L. (Eds.), Subject and topic. Academic Press, New York.
Dunn, B. R., Dunn, D. A., Languis, M., \& Andrews, D. (1998). The relation of ERP components to complex memory processing. Brain and Cognition, 36(3), 355-376.
Fanselow, G., \& Lenertová, D. (2011). Left peripheral focus: Mismatch between syntax and information structure. Natural Language \& Linguistic Theory, 29(1), 169-209.
Federmeier, K. D., \& Kutas, M. (2001). Meaning and modality: Influences of context, semantic memory prganization, and perceptual predictability on picture processing. Journal of Experimental Psychology: Learning, Memory, and Cognition, 27(1), 202-224.
Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. Trends in Cognitive Sciences, 6(2), 78-84.
Frisch, S., Schlesewsky, M., Saddy, D., \& Alpermann, A. (2002). The P600 as an indicator of syntactic ambiguity. Cognition, 85(3), B83-92.
Gorrell., P. (2000). The subject-before-object-preference in German clauses. In B. Hemforth \& L. Konieczny (Eds.), German sentence processing. Dordrecht: Kluwer.
Gratton, G., Coles, M. G. H., \& Donchin, E. (1983). A new method for the off-line removal of ocular artefact. Electroencephalography and Clinical Neurophysiology, 55(4), 468-484.

Greenhouse, S. W., \& Geisser, S. (1959). On methods in the analysis of profile data. Psychometrika, 24(2), 95-112.
Gundel, J. K. (1988). Universals of topic-comment structure. In M. Hammond, E. Moravczik \& J. R. Wirth (Eds.), Studies in Syntactic Typology. Amsterdam: Benjamins.
Heister, J., Würzner, K.-M., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A., \& Kliegl, R. (2011). dlexDB - eine lexikalische Datenbank für die psychologische und linguistische Forschung. Psychologische Rundschau, 62(1), 10-20.
Höhle, T. (1982). Explikation für 'normale Betonung' und 'normale Wortstellung'. In W. Abraham (Ed.), Satzglieder im Deutschen. Tübingen: Gunter Narr.
Hörnig, R., Weskott, T., Kliegl, R., \& Fanselow, G. (2006). Word order variation in spatial descriptions with adverbs. Memory \& Cognition, 34(5), 1183-1192.
Lambrecht, K. (1994). Information structure and sentence form. Topic, focus, and the mental representation of discourse referents. Cambridge: Cambridge University Press.
Lee, C.-Y., Liu, Y.-N., \& Tsai, J.-L. (2012). The time course of contextual effects on visual word recognition. Frontiers in Psychology, 3, 1-13.
Matzke, M., Mai, H., Nager, W., Rüsseler, J., \& Münte, T. (2002). The costs of freedom: An ERP - study of noncanonical sentences. Clinical Neuophysiology, 113(6), 844-852.
Meng, M., Bader, M., \& Bayer, J. (1999). Die Verarbeitung von Subjekt-Objekt Ambiguitäten im Kontext. In I. Wachsmuth \& B. Jung (Eds.), Proceedings der 4. Fachtagung der Gesellschaft für Kognitionswissenschaft, (pp. 244-249). St. Augustin, Germany: Infix Verlag.
Prince, E. F. (1981). Toward a taxonomy of given/new information. In P. Cole (Ed.), Radical Pragmatics. New York: Academic Press.
Reinhart, T. (1981). Pragmatics and linguistics. An analysis of sentence topics. Philosophica, 27, 53-94.
Rizzi, L. (1997). The fine structure of the left periphery. In L. Haegeman (Ed.), Elements of Grammar: Handbook in Generative Syntax. Dordrecht: Kluwer Academic Publishers.
Rosengren, I. (1993). Wahlfreiheit mit Konsequenzen Scrambling, Topikalisierung und FHG im Dienste der Informationsstrukturierung. In M. Reis (Ed.), Wortstellung und Informationsstruktur. Tübingen: Niemeyer.
Rösler, F., Pechmann, T., Streb, J., Röder, B., \& Henninghausen, E. (1998). Parsing of sentences in a language with varying word order: Word-by-word variations of processing demands are revealed by eventrelated brain potentials. Journal of Memory and Language, 38(2), 150-176.
Scheepers, C., \& Crocker, M. W. (2004). Constituent order priming from reading to listening: A visual-world study. In M. Carreiras \& C. Clifton (Eds.), The on-line study of
sentence comprehension: Eyetracking, ERP, and beyond. New York: Psychology Press.
Schumacher, P. B., \& Hung, Y.-C. (2012). Positional influences on information packaging: Insights from topological fields in German. Journal of Memory and Language, 67(2), 295-310.
Skopeteas, S., Fiedler, I., Hellmuth, S., Schwarz, A., Stoel, R., Fanselow, G., Féry, C., \& Krifka, M. (2006). Questionnaire on Information Structure (QUIS). Interdisciplinary Studies on Information Structure 4. Working Papers of the SFB 632. Potsdam: University of Potsdam.
Van Berkum, J. J. A. (2008). The electrophysiology of discourse and conversation. In M. Spivey, M. Joanisse \& K. McRae (Eds.), The Cambridge Handbook of Psycholinguistics. Cambridge: Cambridge University Press.
Van Petten, C., Kutas, M., Kluender, R., Mitchiner, M., \& McIsaac, H. (1991). Fractioning the word repetition effect with event-related potentials. Journal of Cognitive Neuroscience, 3(2), 131-150.
Weskott, T., Hörnig, R., \& Fanselow, G. (2009). Strong contextual licensing of German marked OVS word order. In Proceedings of AMLaP, Barcelona, Spain, 2009. University of Barcelona.

# Probabilistic reasoning in the two-envelope problem 

Bruce D. Burns (bruce.burns@sydney.edu.au)<br>School of Psychology, The University of Sydney<br>NSW 2006 Australia


#### Abstract

In the two-envelope problem a reasoner is offered two envelopes, one containing exactly twice the money in the other. After observing the amount in one envelope it can be traded for the unseen contents of the other. Until recently it was argued that it did not matter whether the envelope was traded or not, but Abbott, Davis \& Parrondo (2010) showed that gains could be made if trading was a probabilistic function of amount observed. Three experiments varied where the observed and maximum amounts fell in a possible distribution and tested whether this affected choices. Trading was less likely for lower observed amount than higher, but this effect differed depending on the stated distribution. This suggests that participants' trade decisions were affected by where observations fit in the distribution, and thus their probabilities. The modeling tools used here may be applicable to other reasoning phenomena.


Keywords: Probabilistic reasoning, two-envelope problem, mathematical modelling, decision making.

## Introduction

The overwhelming evidence of heuristics and biases affecting people's reasoning has often been seen as evidence of irrationality in human thought (Stanovich, 1999). Stanovich pointed out that this conclusion relies on the apparent gap between normatively correct decisions and actual behaviour. However that such a gap indicates irrationality has been challenged by those suggesting that such norms are inappropriate. For example, Oaksford and Chater (1994) suggested that what is seen as an error in the well-known Wason's 4-card selection task is not an error in terms of information gain if you make appropriate assumptions about the distribution of the relevant variables in the environment. Such probabilistic reasoning approaches have gained increasing acceptance (Oaksford \& Chater, 2007). Setting normative standards against which to judge rationality is especially difficult when formal analysis of a problem is difficult, such as has been the case for the twoenvelope problem (Zabell, 1988). However a recent analysis (Abbott, Davis \& Parrondo, 2010; McDonnell \& Abbott, 2009) supported by simulations suggests that distributions are critical to analysing that problem, so it is reasonable to ask whether people act rationally by showing sensitivity to distributions when faced with what has sometimes been considered a paradox.

## The two-envelope problem

Versions of the two-envelope problem were proposed by Kaitchik (1953, pp. 133-134) and attributed by the mathematician Littlewood to the physicist Erwin Schödinger (Littlewood, 1953/1986, p. 26). Although he
does not claim authorship of it, Zabell (1988) stated a twoenvelope version with the following characteristics: the contents of the two envelopes are $x$ and $2 x$; no distribution or limit is given for $x$; the reasoner is handed an envelope (randomly) and opens it; however then the reasoner is given a choice: keep the amount observed, or trade it for the contents of the other envelope. Before the envelope is opened the expected outcome is:

$$
\begin{equation*}
\mathrm{E}=1 / 2(x+2 x)=3 x / 2 \tag{1}
\end{equation*}
$$

Opening an envelope cannot change the amounts in the envelopes so it should not matter whether you keep or trade envelopes because to trade is equivalent to changing your initial random choice. However, opening an envelope containing $y$ means that trading yields either $2 y$ or $1 / 2 y$. If each is a $50 \%$ possibility then trading appears to result in an expected outcome equal to $5 y / 4$. Worse, if the two envelopes were held by two different people (as Zabell proposed), then after opening their own envelopes both would expect to gain from trading. This cannot be true so the problem has sometimes been called a paradox. As Zabell and others have pointed out, the resolution of this paradox is that the envelopes contain two possible pairs of amounts $[2 y, y]$ or $[y, 1 / 2 y]$ but they are not equally likely. The $p(y \mid$ pair ) is not equal to $p$ (pair|y); the first probability is known but it is the second that the reasoner needs. Analyzing what that probability is, and thus what the reasoner should do, has defied a satisfactory mathematical solution (Albers, Kooi, \& Schaafsma, 2005). So the paradox was resolved but the problem of whether to trade remained.

McDonnell and Abbott (2009) point out that the envelope problem has attracted wide interest in game theory and probability theory, and that it is paradigmatic of recent problems in physics, engineering and economics which involve probabilistic switching between two states. For example, it has been shown in stochastic control theory that random switching between two unstable states can result in a stable state (Allison \& Abbott, 2001). Maslov and Zhang (1998) modelled how switching between volatile assets and non-performing cash reserves can produce a net gain.

There is only one published paper on how people respond to the envelope problem. Butler and Nickerson (2008) presented participants with six different versions of the problem. They were told that Envelope 1 (E1) contained a random amount between $\$ 1$ and $\$ 100$, and Envelope 2 (E2) contained either twice or half that amount depending on the result of a coin toss. They varied whether the participant was given E1 or E2, whether the participant knew which it was, and whether they opened the envelope. If participants


Figure 1: Markov model based on Abbott el al's (2010) analysis. $P(x)$ representing the probability of trading if the value in the opened envelope is $x$, and $P(2 x)$ representing the probability of trading if the observed value is $2 x$.
observed the amount then Butler and Nickerson asked them what they would do if it had various values (\$1, \$20, \$40, $\$ 60, \$ 80, \$ 100$ ). Nickerson and Falk's (2006) analysis of these different versions showed whether it was optimal to always trade, trade depending on the observation, always keep, or to be indifferent. For example, if you know you have been given E1 then it is optimal to trade because E2 was generated from E1 with a 50\% chance of two outcomes. They found that if participants observed the amount then there was a tendency towards trading when the amount was less than $\$ 50$ and keeping when it was above $\$ 50$, but this was irrespective of whether the conditions should influence their decision. Consistent optimal decision making was rare, so Butler and Nickerson concluded that participants were largely insensitive to the logical structure of the problem; instead they applied simple heuristics or folk wisdom.

## A general mathematical solution

Different predictions regarding human performance in the two-envelope problem may arise if there was an accepted mathematical analysis of it. Recently McDonnell and Abbott (2009) and Abbott, et al (2010) propose a strategy that can increase the expected outcome above that in Equation 1. The key to their approach is to recognize that once an envelope is opened the information of what it contains breaks the symmetry that leads to Equation 1. Their
starting point was Cover’s (1987) switching function used to solve the pick the largest number problem and the analysis of Parrondo's games in which two losing strategies can be combined to produce a winning strategy if the current state of the problem is used as a criterion (Harmer \& Abbott, 1999). Solving these types of problems requires probabilistic switching between states.
Abbott et al (2010) supposed that opening the envelope reveals $y$ dollars, and the player then trades envelopes with a probability $\mathrm{P}(y) \in[0,1]$. Figure 1 converts their analysis to a Markov model. From the model it can be seen that the expected return (E) when $x$ represents the smaller of the two amounts and $2 x$ the larger, will be:

$$
\begin{align*}
\mathrm{E}=1 / 2[2 x & \mathrm{P}(x)+x[1-\mathrm{P}(x)]+x \mathrm{P}(2 x)+2 x[1-\mathrm{P}(2 x)]]  \tag{2}\\
& =1 / 2[3 x+x \mathrm{P}(x)-x \mathrm{P}(2 x)] \\
& =3 x / 2+x / 2[\mathrm{P}(x)-\mathrm{P}(2 x)]
\end{align*}
$$

Equation 2 shows that probabilistic trading as a function of $x$ can raise the expected value above that expected from either trading or keeping regardless of the observed amount (i.e., Equation 1). Returns can be only be improved if $\mathrm{P}(x)>$ $\mathrm{P}(2 x)$, that is, when the trading function is such that trading is less likely the higher the observed amount is (i.e., the more likely it is to be $2 x$ rather than $x$ ). Abbott et al (2010) show that a monotonically decreasing function will increase
the expected outcome, and that this does not presuppose any particular probability density function for $x$. Calculating the optimal trading function requires knowing the probability density function, but their analysis demonstrates that a simple negative monotonic tendency to trade as a function of observed amount can increase expected outcomes.

## Goals

Abbott et al's (2010) model shows that the higher an observed amount sits within the distribution of amounts, the less likely trading should be. Thus adaptive behavior for people faced with the two-envelope problem would be to be less likely to trade the higher the observed is within the distribution of possibilities. This prediction was tested in two experiments. A third tested whether the distribution itself was critical. Participants may be acting more rationally than Butler and Nickerson (2008) suggested.

## Experiment 1

Where the observed contents of an envelope sit in a distribution of possible amounts depends both on what the amount is and what are the upper and lower limits of possible amounts. So in Experiment 1 both the observed amount ( $\$ 10$ or $\$ 100$ ) and the limit ( $\$ 200$ or no limit) were manipulated. It was predicted that trading rates would be affected by the interaction of the observed and limit factors, such that they would be least likely to trade when the observed was $\$ 100$ and the upper limit was $\$ 200$.

## Method

Participants. A total of 160 senior psychology students at the University of Sydney participated during a practical class focused on reasoning.
Materials and Procedure. Participants read and responded to the following scenario on paper (the italicized text in the squared brackets replaced the underlined text in the relevant condition):
Imagine that you given a choice between two envelopes each containing a sum of money. You are told that neither envelope could hold more than $\$ 200$ [You are told that the envelopes could contain any amount of money], but one envelope contains exactly twice as much money as the other. You randomly choose one of the envelopes and open it, revealing that it contains $\$ 100$ [ $\$ 10]$. You are told that you can either keep the $\$ 100$ [\$10] or take whatever is in the other envelope. What would you do?
Participants circled whether they would keep the $\$ 100$ [\$10] or trade it for whatever was in the other envelope.

## Results \& Discussion

Table 1 presents the proportion of participants choosing to trade in each condition. A logistic regression analysis (using the "Logistic Regression" procedure in SPSS) was performed on choice ( $0=$ keep, $1=$ trade) entering the factors of limit, observed amount, and their interaction. This yielded the following equation for trading:
$\log ($ odds $)=1.355+-0.385 *$ limit $+-2.128 *$ observed + 1.39*limit*observed

The parameter for limit was not significant, Wald $\chi 2(1)=$ $0.525, p=.469$, but that for observed was, Wald $\chi 2(1)=$ $16.224, p<.001$, and so was the interaction, Wald $\chi 2(1)=$ 3.885, $p=.049$.

As predicted, these results showed that participants' choices were affected by the observed contents of the envelope, in that overall there was a strong effect of observed amount. However there was also a significant interaction in that trading was least likely if the highest and observed amounts were such that the largest amount possible was at the limit. This suggests that people's responses were affected by where they saw the possible amounts as falling in the distribution of amounts.

Table 1: Proportion of participants in each condition of Experiment 1 choosing to trade (with sample sizes).

|  | \$10 in opened <br> envelope | \$100 in opened <br> envelope |
| :--- | :---: | :---: |
| $\$ 200$ limit | $.80(\mathrm{n}=39)$ | $.32(\mathrm{n}=38)$ |
| unlimited | $.73(\mathrm{n}=40)$ | $.56(\mathrm{n}=43)$ |

## Experiment 2

An alternative explanation for the interaction between observed amount and limit in Experiment 1 could be that the observed is perceived as worth less in the context of a limit that it is close to. Butler and Nickerson (2008) alluded to such a context effect. So in Experiment 2 participants were directly asked to judge the prior probability of the amount in the envelope. These probabilities should also be lower when the observed amount is half the limit, but such a pattern could not be due to perceptions of monetary value.

Other changes were also made to help generalize the results of Experiment 1. Having a definite limit changes some analyses of the two-envelope problem, so instead of "no limit" a large limit $(\$ 10,000)$ was used. It is unlikely this makes much difference to participants but it removes a potential difference between the two limit conditions. Another possibility is that using such a small amount (\$10) for the lower observed amount may have led to trading because it was perceived as a trivial amount. So in Experiment 2 the lower observed amount was set to $\$ 50$.

The $2 x 2$ design of Experiment 2 was the similar to that for Experiment 1, with factors for limit ( $\$ 200$ or $\$ 10,000$ ) and observed ( $\$ 50$ or $\$ 100$ ). Again I predicted an interaction between trading and observed such the lowest rate should be when the observed amount was close to the limit.

## Method

Participants. A total of 235 senior psychology students participated during practical classes focused on reasoning.
Materials and Procedure. Unlike Experiment 1, the task was presented on a computer. Participants read on-screen instructions that were the same as in Experiment 1 (with appropriate variations for the condition) except that now the
envelope they opened was referred to as "Envelope A" and the unopened as "Envelope B".

Participants were asked the following four questions (\$50 replaces $\$ 100$ in the appropriate condition):

QUESTION 1. First, to check if you understand the instructions correctly, can you type what is the MAXIMUM dollars that Envelope B could contain: \$

QUESTION 2. What would you do? (click one)
Keep the $\$ 100$ [\$50] in Envelope A
Take whatever is in Envelope B
QUESTION 3. Approximately what do you think is the percentage chance that Envelope A (the one you FIRST opened) contains the LARGER amount of money?
$\qquad$ \%

QUESTION 4. In this situation, before any envelopes had been opened, what do you think would have been the probability that the first envelope opened contained $\$ 100$ [\$50] or more? $\qquad$ \%

## Results \& Discussion

Question 1 was designed to check that participants had correctly understood the problem. Most participants (84.3\%) gave the correct answer (either $\$ 100$ or $\$ 200$, depending on condition), but rates of correctness were not affected by condition. It was decided that participants who did not answer this question correctly either misinterpreted the instructions or were not paying attention. Either way their responses could not be relied on, so only the 198 participants who answered correctly were analysed.

Table 2: Proportion of participants in each condition of Experiment 2 choosing to trade. Samples sizes are in parentheses.

|  | $\$ 50$ in opened <br> envelope | $\$ 100$ in opened <br> envelope |
| :--- | :---: | :---: |
| $\$ 200$ limit | $.65(\mathrm{n}=51)$ | $.30(\mathrm{n}=61)$ |
| $\$ 10,000$ limit | $.55(\mathrm{n}=38)$ | $.60(\mathrm{n}=48)$ |

Table 2 shows the proportion of participants in each condition choosing to trade envelopes in response to Question 2. (Sample sizes varied because participants were randomly assigned to a condition by their individual computer.) A logistic regression analysis was performed on choice ( $0=$ keep, $1=$ trade) entering the factors limit, observed amount, and their interaction. This yielded the following equation for trading:

```
Log(odds) = 0.534 + -0.483*limit + -1.374*observed +
1.728*limit*observed
```

The parameter for limit was not significant, Wald $\chi 2(1)=$ 1.190, $p=.275$, but that for observed was, Wald $\chi 2(1)=$ $11.190, p=.001$, and so was the interaction, Wald $\chi 2(1)=$
$8.420, p=.004$. So the Experiment 1 interaction pattern was replicated despite changing the lower observed amount, the specification of the higher limit, and mode of presentation.

In response to Question 3 most participants (92.2\%) thought there was exactly a $50 \%$ chance that the other envelope would contain more money. The overall mean response was $49.49 \%$, and there were no effects of condition. Thus, despite choosing to keep or trade their envelope, very few participants seemed to think the odds of the other envelope containing more was other than $50 \%$. Even if participants act as though sensitive to a distribution, this does not necessarily mean they are aware of it (e.g., Bargh \& Ferguson, 2000).

Table 3: Mean judgments (with standard deviations) of prior probabilities (percentages) for each condition.

|  | \$50 in opened <br> envelope | \$100 in opened <br> envelope |
| :--- | :---: | :---: |
| \$200 limit | $57.25(\mathrm{sd}=17.3)$ | $46.70(\mathrm{sd}=14.1)$ |
| $\$ 10,000$ limit | $54.41(\mathrm{sd}=29.1)$ | $58.22(\mathrm{sd}=25.5)$ |

In response to Question 4 most participants thought that there was about a $50 \%$ probability that their envelope could have contained an equal or higher amount before it was observed, but Table 3 shows that this varied with condition. A $2 x 2$ ANOVA found no main effects of limit, $F(1,194)=$ $1.980, p=.161$, no effect of amount observed, $F(1,194)=$ 1.190, $p=.277$, but a significant interaction, $F(1,194)=$ 5.409, $p=.021$. Thus consistent with the observed $\$ 100$ and limit $\$ 200$ condition being the one least likely to lead participants to favour trading envelopes, participants in this condition were also least likely to think that their envelope could have contained more a priori. Why Question 4 but not 3 showed a difference may be because it does not so starkly ask participants to contradict their intuition that two coinflip like choices should mean $50 \%$ each.

By replicating the interaction found in Experiment 1, Experiment 2 further supported the hypothesis that participants are less likely to trade when the higher amount would be at the end of the distribution. Adding support to the claim that this was because of where they felt the observed amount fell in the distribution the manipulations had a similar effect on a direct measure of how likely they thought that the observed amount could have been higher.

## Experiment 3

Abbott et al's (2010) solution suggests that people may be less likely to trade when the observed amount is higher in the distribution, but working out the optimal trading strategy would depend on knowing the details of the distribution of amounts. If people act consistent with this analysis, then people’s tendency to trade should be affected by what they believe about the distribution. So far the results suggest that that people's responses reflect a sensitivity to the distribution of amounts, so explicitly stating a different distribution could affect their choices.

In Experiment 3 participants were told that the envelope amounts had either a flat or a bimodal distribution. It is likely that many participants assumed an essentially flat distribution in the previous experiments, in which case explicitly stating that the distribution is flat should produce similar results to Experiments 1 and 2. However explicitly stating that there was a bimodal distribution could lead to a different pattern of results. By increasing the chances of high amounts in envelopes this distribution should increase trading when the other envelope potentially contains an amount at the top of the distribution. A 2 x 2 design was used with factors for distribution (flat or bimodal) and observed ( $\$ 50$ or $\$ 100$ ). The limit was always $\$ 200$.


Figure 2: Diagrams accompanying the instructions for the bimodal (Panel A) and flat (Panel B) distributions.

## Method

Participants. One hundred and three first-year psychology students completed the experiment for partial course credit.

Materials and procedure. Materials and procedure were identical to Experiment 2 except for the addition of the distribution manipulation. In the flat condition participants read that "the probability of any amount is equal to any other" and saw the graph in Panel B of Figure 2. In the bimodal condition they read "the probability of any amount is not equal, in that amounts closer to the minimum or maximum amounts are more likely" and saw the graph in Panel A of Figure 2. These graphs were intentionally vague in order to give a general shape to the distribution rather than provide a precise way to calculate the probabilities.

The observed amount in the opened envelope was either \$50 or $\$ 100$, but the maximum possible was always $\$ 200$.

## Results \& Discussion

Most (83.5\%) participants correctly identified the maximum amount the unopened envelope could contain, but as in Experiment 2 only these 86 were analysed. The proportion in each condition choosing to trade is presented in Table 4.

Table 4: Proportion of participants in Experiment 3 choosing to trade. Maximum amount was always $\$ 200$.

|  | \$50 in opened <br> envelope | \$100 in opened <br> envelope |
| :--- | :---: | :---: |
| Bimodal distribution | $.25(\mathrm{n}=20)$ | $.42(\mathrm{n}=24)$ |
| Flat distribution | $.58(\mathrm{n}=19)$ | $.26(\mathrm{n}=23)$ |

For the flat distribution the trading proportions were similar to the same conditions in Experiment 2 in which no distribution was specified, with more trading when the observed amount was $\$ 50$ than when $\$ 100, \chi 2(1)=4.37, p$ $=.037$. In the bimodal condition, the direction of the effect of revealed amount was the opposite, but this effect was not significant, $\chi 2(1)=1.35, p=.246$. A logistic regression analysis was performed on choice entering the factors distribution ( $0=$ bimodal, $1=$ flat), revealed amount, and their interaction. This yielded the following equation for trading:

Log(odds) $=-1.099+1.417 *$ distribution $+0.762 *$ observed + -2.122*distribution*observed

The parameter for distribution was significant, Wald $\chi 2(1)$ $=4.161, p=.041$, but not that for observed, Wald $\chi 2(1)=$ 1.326, $p=.250$. The interaction parameter was significant, Wald $\chi 2(1)=5.120, p=.024$.

These results indicated that people were sensitive to the distributions of amounts when deciding whether to trade. For the same amount with the same limit their propensity to trade was influenced by what they were told about the distribution of amounts. When the distribution was flat they responded similarly to how they did in Experiment 2, suggesting that participants had previously assumed a flat distribution. However a bimodal distribution changed the pattern of their responses implying that they took into account the prior probabilities of different amounts.

It should be noted that the Figure 2 distributions are only possible for either the higher or the lower amounts, not the sum of their distributions. Given that participants do not know if they observe the higher or the lower amount they may have been confused as to what exactly was the distribution represented by their diagram. However the main point of the experiment was to test whether the distribution plays a role in participants' choices, and confusion about the distribution should not affect their choices unless they see the distribution as important.

## General Discussion

Abbott et al's (2010) analysis suggests that a probabilistic strategy for trading can lead to gains in the two-envelope problem unobtainable by a pure strategy. In general, such a probabilistic strategy can increase expected outcome over an absolute strategy if the probability of trading is a monotically decreasing function of the observed amount. This suggests that people given the two envelope problem may have a tendency to trade that is sensitive to the distribution of amounts. The results of Experiments 1-3 support the claim that people do this when faced with the two-envelope problem. Participants were consistently least likely to trade when the higher alternative would be at the top of the distribution, except in Experiment 3 when the bimodal distribution increased the likelihood of such an amount. Furthermore, in Experiment 2 it was found that participants' assessments of the prior probabilities of amounts had the same pattern. Thus the results suggest that people are consistent with what Abbott et al's model suggests optimizes responses to the two-envelope problem: trading as a function of the observed amount and being sensitive to the distributions. Experiment 3 is critical in showing that not just the size of the observed amounts but their perceived distribution affected choices. However this conclusion is weakened by possible limitations of its methodology, therefore more research is required.

These experiments did not systematically manipulate the amount in the revealed envelope to see what shape there might be to any monotonic function to trade. Inspection of Butler and Nickerson's (2008) data suggests that there is a trend within the large effect of greater/lesser than \$50 towards less trading as observed amounts increase. However their sample size is not large enough to expect a post-hoc analysis to show a significant effect. Overall, the results do not dispute Butler and Nickerson's finding that participants often make fundamental errors in analysing the twoenvelope problem. The errors they revealed were in understanding the logical implications of the details of different versions, and in this way they are analogous to Wason's (1968) finding that people were poor at understanding the logical implications of his selection task. However Oaksford and Chater's (1994) analysis showed that people's responses may make sense if seen in terms of how information is distributed in the world. Thus my results fit with a more general trend of finding that people are poor at applying formal logic but can be sensitive to the implications of probability distributions. Applying probabilistic inference may be seen as the computational goal of cognition.

McDonnell and Abbott (2009) saw the two-envelope problem as interesting because it embodies a phenomenon that comes up in many domains, that of probabilistic switching between two states. Their analysis demonstrates that an appropriate probabilistic function may improve outcomes even when important characteristics of the distributions are unknown. A number of decision making tasks require a choice between functions whose properties
are uncertain, for example, choices between different market options. The demonstration here that the mathematical analysis of such choices can lead to supportable behavioural predictions suggests that these mathematical tools may have value for analysing other types of decisions.

## References

Abbott, D., Davis, B. R., \& Parrondo, J. M. R. (2010). The two-envelope problem revisited. Fluctuation and noise letters, 9, 1-8.
Albers, C. J., Kooi, B. P., \& Schaafsma, W. (2005). Trying to solve the two-envelope problem. Synthese, 145, 89109.

Allison, A., \& Abbott, D. (2001). Control systems with stochastic feedback. Chaos, 11, 715-724.
Bargh, J. A., \& Ferguson, M. J. (2000). Beyond behaviorism on the automaticity of higher mental processes. Psychological Bulletin, 126, 925-945
Butler, S. F., \& Nickerson, R. S. (2008). 'Keep or trade? An experimental study of the exchange paradox. Thinking \& Reasoning, 14, 365 - 394.
Cover, T. (1987). Pick the largest number. In T. Cover \& B. Gopinath (Eds.), Open Problems in Communication and Computation (p. 152). New York: Springer-Verlag.
Harmer, G. P., \& Abbott, D. (1999). Losing strategies can win by Parrondo's paradox. Nature, 402, 864.
Kraitchik, M. (1953). Mathematical Recreations (2nd edition). New York: Dover.
Littlewood, J. E. (1953/1986). Littlewood’s Miscellany (B. Bollobas, Ed.). Cambridge, UK: Cambridge University Press.
Maslov, S., \& Zhang, Y. (1998). Optimal investment strategy for risky assets. International Journal of Theory and Applications of Finance, 1, 377-387.
McDonnell, M. D., \& Abbott, D. (2009). Randomized switching in the two-envelope problem Proceedings of the Royal Society A, 465, 3309-3322.
Nickerson, R. S., \& Falk, R. (2006). The exchange paradox: Probabilistic and cognitive analysis of a psychological conundrum. Thinking and Reasoning, 12, 181-213.
Oaksford, M., \& Chater, N. (1994). A rational analysis of the selection task as optimal data selection. Psychological Review, 101, 608-631.
Oaksford, M., \& Chater, N. (2007). Bayesian rationality: The probabilistic approach to human reasoning. Oxford, UK: Oxford University Press.
Stanovich, K. E. (1999). Who is rational?: Studies in individual differences in reasoning. Mahwah, NJ: Lawrence Erlbaum Associates.
Wason, P. C. (1968). Reasoning about a rule. Quarterly Journal of Experimental Psychology, 20, 273-281.
Zabell, S. L. (1988). Symmetry and its discontents. In B. Skyrms \& W. L. Harper (Eds.), Causation, chance, and credence (Vol. I, pp. 155 - 190). Dordrecht, the Netherlands: Kluwer Academic Publishers.

# Preschoolers' ability to navigate communicative interactions in guiding their inductive inferences 

Lucas P. Butler (lucas_butler@eva.mpg.de) ${ }^{1,2} \&$ Ellen M. Markman (markman@stanford.edu) ${ }^{2}$<br>${ }^{1}$ Department of Developmental and Comparative Psychology<br>Max Planck Institute for Evolutionary Anthropology<br>Deutscher Platz 6, 04103 Leipzig, Germany<br>${ }^{2}$ Department of Psychology, Stanford University<br>Jordan Hall, Building 420, 450 Serra Mall<br>Stanford, CA 94305 USA


#### Abstract

Evaluating whether information is generalizable, essential knowledge about a novel category is a critical component of conceptual development. In previous work (Butler \& Markman, 2012) 4-year-old children were able to use their understanding of whether information was explicitly communicated for their benefit to guide such reasoning, while 3-year-olds were not. In two experiments, we further investigate this finding. Four-year-olds were adept at navigating pedagogical interactions, judiciously identifying which specific actions in an ongoing interaction were meant as communicative demonstrations for their benefit, while 3-year-olds did not distinguish between the manners of demonstration even in a simpler context. Taken together, these experiments illustrate that this powerful learning mechanism for facilitating children's conceptual development is under construction during the preschool years.


Keywords: Social cognition; inductive inference; generalization; pedagogy; communication.

## Introduction

A foundational developmental process is the acquisition of generic knowledge about kinds and categories that supports the construction of a coherent conceptual understanding of the world (Gelman, 2003; Gelman \& Wellman, 1991; Keil, 1989; Markman, 1989). But acquiring such knowledge often requires making inductive inferences on the basis of limited information. For example, imagine that a person learns a new fact about an individual animal, say that a bird has a particular shape of feather. Should this person infer that all birds of this kind have similar feathers? That all birds have similar feathers? Or alternatively, that these feathers are unusual and idiosyncratic to only this individual, or to a limited set of related birds? One could arguably make each of these generalizations with equal legitimacy based on the given evidence (cf., Goodman, 1965).

Determining the scope of a generalization is a challenge for young children, who experience a flood of new information and must rapidly construct a conceptual framework for understanding the world (Lopez, Gelman, Gutheil, \& Smith, 1992). Although in many cases children can tackle this problem by relying on linguistic cues that mark generic knowledge (e.g., Cimpian \& Markman, 2009,

2011; Gelman, Star, and Flukes 2002; Hollander, Gelman, \& Raman, 2009), there are many potential non-linguistic contexts in which children observe others' actions and must evaluate whether the information those actions produce is generic. How do children carry out this process?

Previous research with infants suggests that even infants are sensitive to cues that someone is deliberately communicating information for their benefit (see Cisbra, 2010), and that this sensitivity appears to change how infants process that information. Specifically, they appear to encode information conveyed communicatively as more kind-relevant and stable than information produced noncommunicatively (Gergely, Egyed, \& Király, 2007; Futó, Téglás, Csibra, \& Gergely, 2010; Yoon, Johnson, and Csibra, 2008). Does this early sensitivity to communicative cues also play a role in older children's ability to assess whether information is generic?

In recent research, Butler and Markman (2012) demonstrated that by age 4 children utilize cues that someone is deliberately communicating information for their benefit to guide such inductive inferences. Three- and 4-year- old children were first taught a label ("blicket") for a novel object. Children then observed perceptually identical evidence that this object was magnetic, but produced with subtly different actions: the experimenter accidentally used the object as a magnet; did so intentionally; or did so while conveying that they were acting communicatively and pedagogically for the child's benefit.

Importantly, after being taught the object's name, children engaged in a short, unrelated distractor task. This was done in an attempt to create a clear break between the pedagogical word-learning phase and the experimental manipulation so that children would not interpret every action as meant "for them," simply because the adult had previously been teaching them something about this kind.

After seeing this evidence, children were given 10 additional blickets that were identical, but which upon exploration turned out not to be magnetic. To assess the strength of children's inductive inferences, we measured their continued exploration of the inert objects when they discovered that they failed to have the novel property (after Schulz, Standing, \& Bonawitz, 2008): specifically how long and how many attempts children made to try get the inert
blickets to pick up paperclips. This persistence in the face of such negative evidence is an index of how strongly children inferred that the property should generalize.

Four-year-olds showed significantly more such exploration in the pedagogical condition, suggesting they made stronger inductive generalizations about the property when it was demonstrated communicatively. So by age 4 preschoolers are able to use communicative cues to guide their reasoning about whether information represents generic knowledge about a kind. Interestingly, 3-year-olds also based their inferences on the intentions of the adult, but did so purely on the basis of whether the evidence was produced intentionally rather than accidentally. That is, 3 year olds treated intentional and pedagogical actions as warranting similarly strong inferences, more so than an accidental action. Thus there appears to be a developmental difference in how children identify and make use of communicative acts in guiding their inductive inferences.

This developmental difference is intriguing. It seems unlikely that 3 -year-old children fail to recognize the communicative cues that signal that an action is meant for their benefit. Indeed, infants as young as 10 months appear to recognize such cues, and having information conveyed with those communicative cues does appear to impact how they treat new information. (Csibra, 2010; Csibra \& Gergely, 2009). The hypothesis we consider here is whether there may be a developmental difference in the ability to assess, in real time, which actions within a given context are pedagogical. Even within a pedagogical interaction adults may perform a number of actions that are not meant to carry meaningful information, but which could potentially be misinterpreted as acts of teaching. Imagine that a child is watching her mother or father preparing a snack. The adult is interacting with a number of kitchen tools, occasionally interrupting an action to attend to a forgotten item or ingredient, and then returning to the task at hand. In such a dynamic, flowing context, if a child is going to utilize her sensitivity to whether or not an action is meant for her, she needs to be able to identify which actions are truly meant as demonstrations, and which are merely incidental, unnecessary, or part of a different embedded event.

The current research aims to shed initial light on this issue, taking two complementary approaches. First, we ask what factors might play a role in driving the effect seen previously in older children. How best can we characterize 4 year olds' use of communicative cues to guide their inferences, and how nuanced is this learning mechanism? Second, what might be preventing younger children from using this distinction to guide their inferences? Does a manipulation that might make it easier for children to draw the distinction between the conditions reveal a similar pattern of inferences at a younger age?

## Experiment 1

In utilizing their sensitivity to communicative cues to guide their inductive inference, children need to be able to do so while navigating ongoing interactions with adults,
discriminating those actions that are truly meant for their learning benefit from those that are not.

As discussed earlier, in Butler and Markman's (2012) task, in which children learned that an object was called a "blicket," and then saw that it was magnetic, they were given an unrelated distractor task between the word learning and evidence phases, in an attempt to distance the pedagogical teaching of the word from the experimental manipulation of how the key evidence was produced. Thus the distractor task was meant to provide a clear interruption of the ongoing pedagogical interaction, potentially implying that the subsequent actions were not necessarily meant for the child's benefit, unless they were clearly marked as communicative acts. This opens up a question about the nature of 4-year-olds' ability to selectively use communicative cues to guide their inferences about novel information. How adept are they at identifying an action meant for their benefit when the target action is embedded within a still-continuing pedagogical interaction?

## Method

The procedure for Experiment 1 was closely modeled after Butler and Markman (2012), in which children were presented with evidence that a novel object had a novel function, in one of three subtly different ways: communicatively, intentionally, or accidentally. In order to directly address the question of how adept 4 -year-olds are at applying their sensitivity to communicative cues in an ongoing context, we manipulated the the position of the distractor task within the continuous stream of events (see Figure 1), thus varying whether or not the ongoing pedagogical interaction was clearly interrupted prior to the evidence phase. If 4 -year-olds are truly conducting a nuanced, moment-by-moment analysis of an adult's communicative intentions at each particular point in time, they may distinguish between pedagogical and intentional actions even within an ongoing pedagogical interaction, regardless of the position of the distractor task.

Materials The materials in Experiment 1 were 11 PVC pipes, 5 cm in diameter and 7.5 cm tall. The active object had a noisemaker inside that made an animal noise when it was flipped upside down. The 10 inert objects had several pebbles taped to the inside of the PVC piping in order to give them an equivalent weight and feel as the active object. All 11 objects were covered with blue duct tape, on one end and around the sides, and had yellow duct tape covering one end. The objects were perceptually indistinguishable.

Participants The participants were 96 children from a university preschool ( $M_{\text {age }}=4$ years, 8 months). An additional 9 children were excluded because of experimenter error, or because they did not attend to or finish the procedure. Participants were randomly assigned to one of two orders and three experimental conditions, with the constraint that all six condition-order groups were matched for gender and age.

Procedure The procedure was modeled after that used in Butler and Markman (2012) and described above, but the order of the procedure was manipulated (see Figure 1). In the interruption order, the sequence of events was identical to that used previously. The experimenter brought out the target object and four distractor objects, and proceeded to teach the child a novel word ("femo") for the target object. All children successfully picked out the target object when asked for it by name on two successive trials. The experimenter then said, "Now let's do something really fun! We can make a boat out of colored paper!" and then proceeded with the distractor task. In the no interruption order, the placement of the word learning and distractor phases was switched. Children first participated in the boatmaking task as a warm-up game. The experimenter then brought out the target object and the four distractors and conducted the word learning procedure. In both orders, after the second phase of the procedure the experimenter then said, "I'm going to put a few of my things away" and proceeded to put away each of the distractor objects, finally picking up the target object.

In the evidence phase, the experimenter moved the object from one hand into the other, flipping it upside down in the process and placing it yellow-side-down on the table, all in one continuous action. This action was identical in the pedagogical and intentional conditions, except that in the pedagogical condition the experimenter made eye contact with the child and said, "Look, watch this." The action in the accidental condition was nearly identical, but the experimenter appeared to lose her grip on the object as she was picking it up and moving it, saying, "Oops!" and then appeared to catch it with her other hand and place it on the table. In all three conditions the experimenter then said, "Wow!" after placing the object on the table.


Figure 1: Overview of task structure in the two orders of Experiment 1.

Coding \& Data Analysis Two independent judges watched only the exploration phase of each video, and coded both the
amount of time children spent exploring the inert objects, and how many times they attempted to elicit the property from those inert objects.

Children's exploration violated assumptions of normality and homogeneity of variance. Thus we used non-parametric Mann-Whitney tests to conduct pairwise comparisons between conditions, and ordinal logistic regressions (see Cimpian \& Cadena, 2010; Cimpian \& Markman, 2009) to assess the overall impact of condition and order.

## Results

In order to assess the overall impact of condition and order on children's exploration, we conducted ordinal logistic regressions (OLR) on our two principal measures exploration (number of attempts to elicit the property from the inert objects and time exploring the inert objects), with condition and order as predictors. There was a significant effect of condition on both number of attempts (Wald $\chi^{2}=$ 16.06, $p<0.001$ ) and time exploring (Wald $\chi^{2}=16.67, p<$ 0.001 ). There was no effect of order of the distractor task on number of attempts (Wald $\chi^{2}=.064, p=.801$ ), nor was their an interaction between order and condition (Wald $\chi^{2}=$ $.345, p=.841)$. There was a marginal effect of order on time exploring (Wald $\chi^{2}=3.52, p=.061$ ), but there was no interaction between order and condition (Wald $\chi^{2}=.255, p$ $=.880$ ). Overall, these analyses suggest that the experimental condition (pedagogical, intentional, or accidental) had a significant effect on 4 year olds' exploration, and that while the order might have had an effect, it did not alter the effect of condition. To explore these results further, we now compare the results across condition for each order.

Interruption Order The patterns of exploration across conditions in the interruption order clearly replicated the findings of Butler and Markman (2012).

Four-year-olds made significantly more attempts to elicit the property from the inert objects in the pedagogical condition $\left(M_{\text {pedagogical }}=9.37, \mathrm{SD}=7.20\right)$ than in either the intentional $\left(M_{\text {intentional }}=2.69, \mathrm{SD}=3.77\right.$; Mann-Whitney $Z=$ 3.07, $p=.002$ ) or accidental $\left(M_{\text {accidental }}=3.94, \mathrm{SD}=5.20\right.$; Mann-Whitney $Z=2.38, p=.017$ ) conditions (see Figure 2). There was no difference between the intentional and accidental conditions (Mann-Whitney $Z=.218, p=.828$ ).


Figure 2: Mean number of attempts in each condition in the interruption order of Experiment $1 . N=64$ (16 per condition). Error bars represent $+/-1$ SEM.

Four-year-olds also spent significantly longer exploring the inert objects in the pedagogical condition $\left(M_{\text {pedagogical }}=\right.$ $13.94 \mathrm{~s}, \mathrm{SD}=10.54$ ) than in either the intentional ( $M_{\text {intentional }}$ $=3.87 \mathrm{~s}, \mathrm{SD}=5.05$; Mann-Whitney $Z=2.97, p=.003$ ) or accidental $\left(M_{\text {accidental }}=5.44 \mathrm{~s}, \mathrm{SD}=9.03\right.$; Mann-Whitney $Z$ $=2.59, p=.010)$ conditions. There was no difference between the intentional and accidental conditions (MannWhitney $Z=.276, p=.783$ ).

No-Interruption Order Despite the marginal overall effect of order on time exploring seen in the OLR analyses, the patterns of exploration by children who saw the distractor first, and then learned the word, followed immediately by the evidence phase with no interruption, were nearly identical to those who saw a clear interruption between these phases.

Four-year-olds made significantly more attempts to elicit the property from the inert blickets in the pedagogical condition ( $M_{\text {pedagogical }}=10.37, \mathrm{SD}=10.12$ ) than in either the intentional $\left(M_{\text {intentional }}=3.94, \mathrm{SD}=4.22\right.$; Mann-Whitney $Z=$ 1.97, $p=.048)$ or accidental $\left(M_{\text {accidental }}=3.67, \mathrm{SD}=4.98\right.$; Mann-Whitney $Z=2.07, p=.038$ ) conditions (see Figure 3 ). There was no difference between the intentional and accidental conditions (Mann-Whitney $Z=.270, p=.787$ ).


Figure 3: Mean number of attempts in each condition in the no interruption order of Experiment $1 . N=64$ (16 per condition). Error bars represent $+/-1$ SEM.

Four-year-olds also spent significantly longer exploring the inert blickets in the pedagogical condition $\left(M_{\text {pedagogical }}=\right.$ $24.50 \mathrm{~s}, \mathrm{SD}=20.65)$ than in either the intentional $\left(M_{\text {intentional }}\right.$ $=10.31 \mathrm{~s}, \mathrm{SD}=11.72$; Mann-Whitney $Z=2.11, p=.035$ ) or accidental $\left(M_{\text {accidental }}=8.07 \mathrm{~s}, \mathrm{SD}=9.92\right.$; Mann-Whitney $Z$ $=2.35, p=.019$ ) conditions. There was no difference between the intentional and accidental conditions (MannWhitney $Z=.353, p=.724$ ).

## Discussion

The results of Experiment 1 help clarify how older preschoolers use communicative cues to guide their inductive inferences. Regardless of whether or not they were given a clear interruption in the overarching pedagogical interaction, a deliberate break between engaging in an overarching pedagogical interaction and seeing the target action producing the evidence, children selectively
modulated the strength of their inductive generalizations on the basis of whether or not the experimenter explicitly demonstrated that action for their benefit. This suggests that 4-year-olds' inferences are driven a moment-by-moment analysis of whether or not each individual action or series of actions is meant for them.

A judicious application of this learning mechanism is important for the accuracy of young children's developing conceptual representations. If children simply made a broad inference about whether an adult is currently teaching them or not, they might be misled about the importance of various pieces of information that they might witness in such contexts. If children did not engage in a moment-bymoment analysis of which actions were deliberately meant for them, they might mistakenly generalize even an incidental action action as what one does with this kind of object, even though in fact this is an idiosyncratic, unusual way to use this kind of object. Simply put, it is helpful for children to use their understanding of others' communicative intentions to guide the inferences if they can do so selectively. The results of this experiment suggest that, at least by age 4, children appear to be capable of using communicative cues to guard against such over interpretation, ensuring that the information that does make it into their representations is likely to be important, generic information about the kind.

## Experiment 2

The results of Experiment 1 shed light on the sophistication of 4-year-olds' ability to navigate ongoing interactions in order to identify which actions are meant as communicative acts for their benefit, and thus what information ought to be taken as important and generalizable. Even when children have to conduct a more nuanced moment-by-moment analysis of the interaction, 4-year-olds show a consistent pattern of making stronger inferences about evidence that is demonstrated communicatively for their benefit.

This contrasts with the findings with 3-year-olds from Butler and Markman (2012). In three experiments, 3-yearolds consistently showed an analogous effect for pedagogical demonstration compared to seeing the same action done accidentally. However, 3-year-olds did not distinguish between the pedagogical and intentional actions, making similarly strong inferences in both conditions. What might account for this developmental difference? Moreover, as we have suggested this inference requires that children not only recognize that an action is for them, but have to be able to navigate an ongoing interaction, assessing moment-by-moment whether or not particular actions are indeed meant for them, even when they occur within a communicative context. If younger children struggle with identifying individual pedagogical actions within an overarching stream of actions, it might take a more explicit demarcation of the specific actions that produce the relevant evidence in order to elicit the pattern of reasoning seen in older children.

In Experiment 2, we take a first step at asking whether presenting this manipulation in a context that might make it easier for children to navigate the ongoing interaction and identify which actions are truly made for their benefit, might lead 3-year-olds to be more successful at selectively using communicative cues to guide their inferences.

## Method

Recall that in Experiment 1, we investigated the importance of the distractor task in potentially facilitating children's ability to navigate the ongoing interaction and identify which actions were and were not meant for them. For older children this did not seem to have any tangible impacteven without any clear interruption between the word learning and evidence phases, 4-year-olds were readily able to selectively use the communicatively demonstrated evidence to make a stronger inference about the novel property. But this interruption seems like a logical place to start in asking whether manipulating the complexity of the context might help facilitate 3 -year-olds' ability to engage in the same inference process.

In Experiment 2, we attempted to boost the clarity of the break provided by the distractor task by making it clearly non-pedagogical, non-communicative, and even noninteractive. We hypothesized that 3 -year-olds have a more global sense of whether or not they are engaged in a pedagogical interaction, and that this may be leading them to over interpret everything that occurs in this context as likely pedagogical unless otherwise marked. If so, then establishing a clearer break between the pedagogical word learning and the evidence phase might help them distinguish between the pedagogical and intentional actions.

Materials Having established in Experiment 1 that previous findings were not an artifact of the materials used in those studies, in Experiment 2 we returned to using materials identical to those used in Butler and Markman (2012): 1 target magnetic object, 10 identical inert objects, and metal paperclips.

Participants The participants were an additional 24 children from a university preschool ( $M_{\text {age }}=3$ years, 5 months). Participants were randomly assigned to one of two experimental conditions (Pedagogical or Intentional), with the constraint that the conditions be equated for gender and age.

Procedure As in Butler \& Markman (2012) and the Interruption order of Experiment 1, children were first explicitly taught a label for the target object. After children were taught the novel word and had successfully indicated the target object on two successive trials, the experimenter brought out the paperclips, colored pencils, and a sheet of paper with a simple triangle outline on it, and said, ""And here's a picture to color! Why don't pick out your favorite color to color the triangle with, and then I can write your name on your picture!" The experimenter then let the child
color for 60 s while she pretended to write something down, not making eye contact or otherwise engaging with the child during this distractor task. The experimenter then said "I'm going to put these away" and began to clean up each of the distractor objects, finally picking up the target object and using the object to magnetically pick up the paperclips. This evidence was produced one of two three subtly different ways, as in Experiment 1: pedagogically or intentionally (but non-communicatively). As our main interest was in whether these younger children would distinguish between the pedagogical and intentional conditions (and not whether patterns of exploration in these two conditions would differ from the accidental condition), only these two conditions were run..

The key change from previous studies was the use of this non-interactive distractor task. This was done as an attempt to provide children with a clearer interruption in the ongoing pedagogical interaction. If 3-year-olds' failure to distinguish between the intentional and pedagogical conditions in previous studies was due to an overall sense of being in a pedagogical interaction, this change might help them disengage from this and discriminate whether or not the target action is truly meant as an act of communication.

## Results

Replicating the findings from Butler \& Markman (2012), 3-year-olds did not appear to make any distinction between the pedagogical and intentional conditions. Children were equally likely to explore the inert objects in both the pedagogical ( 8 children, $75 \%$ ) and intentional ( 7 children, $58 \%$ ). Moreover, children made similar numbers of attempts to elicit the property form the inert objects in both the pedagogical $\left(M_{\text {pedagogical }}=3.17, \mathrm{SD}=3.49\right)$ and intentional $\left(M_{\text {intentional }}=4.08, \mathrm{SD}=4.80\right.$, Mann Whitney $\mathrm{Z}=.240, p=$ .810). Children also spent a similar amount of time exploring the inert objects in both the pedagogical ( $M_{\text {pedagogical }}=8.67 \mathrm{~s}, \mathrm{SD}=10.31$ ) and intentional ( $M_{\text {intentional }}$ $=8.42 \mathrm{~s}, \mathrm{SD}=11.76$, Mann Whitney $\mathrm{Z}=.090, p=.928$ ).

## Discussion

Providing younger children with a clearer break between the pedagogical word learning and evidence phases did not facilitate their ability to selectively use communicative cues to guide their inferences. Even when we had a 60 s break in which the experimenter did not interact with the children, these children still failed to distinguish the pedagogical demonstration and the intentional, but non-pedagogical, action, making similarly strong inferences in both cases. Although we cannot compare children's inferences in these conditions to an accidental condition, across three studies in Butler \& Markman (2012), 3-year-olds consistently distinguished between intentional or pedagogical and accidental conditions, and there is little reason to expect them not to do so in this experiment. Moreover, the key point for our conclusion is that children failed to distinguish between a pedagogical and identical, but non-pedagogical, action, even given a clearer break.

However, it should be noted that there was still an overarching pedagogical context. That is, children still came to the room with a knowledgeable adult and engaged in a brief pedagogical interaction with them. If the distinction between pedagogical and intentional is more global for these younger children, happening on the level of the overarching context rather than moment-by-moment actions, this may explain why having the ongoing pedagogical interaction interrupted by a clearly noncommunicative, non-interactive distractor task was not enough to enable them to fully disengage from the pedagogical expectation and evaluate whether the individual instrumental action was deliberately done for their benefit.

## General Discussion

To return to our original research question, how best to we characterize preschoolers' ability to use communicative cues to guide their inductive inferences? Our perspective is that although younger children most certainly are capable of recognizing communicative cues (Cisbra, 2010), they may be less adept at navigating pedagogical interactions in order to identify which actions are meant for their benefit. On this account, younger children may have a more global sense of whether or not they are currently engaged in a pedagogical interaction with a knowledgeable adult, and may interpret, or even over-interpret, a variety of intentional actions as for them, even if they are not clearly meant as such, and only disengage from this interpretation when an action is clearly marked otherwise, for example as accidental.

The results of these experiments are consistent with this interpretation. Although 4-year-olds judiciously identified which actions were meant for them even when embedded in an ongoing pedagogical interaction, even when 3-year-olds were given a clear break between being taught the word and seeing the evidence produced either communicatively or non-communicatively, they did not use that distinction to guide their inferences. Thus simply interrupting the pedagogical interaction does not seem to be enough to disengage children's overall pedagogical interpretation of the situation.

More broadly, the learning mechanism explored in this paper clearly has powerful implications. Preschoolers are highly sensitive to communicative cues that indicate a particular action is meant for their benefit, and use this to guide stronger and inferences about novel information. But important questions remain about how best to characterize this learning mechanism, what inferential processes are children engaged in, and how this develops over the preschool years.

## Acknowledgments

This research was based on a doctoral dissertation submitted to Stanford University, and was supported by a NSF Graduate Research Fellowship.. Special thanks to the parents, children, teachers, and staff at the Bing Nursery School. We also thank Kimmy Scheible for assistance with data collection and coding, and Andrei Cimpian and
members of the Markman Lab for helpful discussion of this work.

## References

Butler, L. P., \& Markman, E. M. (2012). Preschoolers use intentional and pedagogical cues to guide inductive inferences and exploration. Child Development, 83, 14161428.

Cimpian, A., \& Markman, E. M. (2009). Information learned from generic language becomes central to children's biological concepts: Evidence from their openended explanations. Cognition, 113, 14-25.
Cimpian, A., \& Markman, E. M. (2011). The generic/nongeneric distinction influences how children interpret new information about social others. Child Development, 82, 471-492.
Csibra, G. (2010). Recognizing communicative intentions in infancy. Mind \& Language, 25, 141-168.
Csibra, G., \& Gergely, G. (2009). Natural pedagogy. Trends in Cognitive Sciences, 13, 148-153.
Futó, J., Téglás, E., Csibra, G., \& Gergely, G. (2010). Communicative Function Demonstration induces kindbased artifact representation in preverbal infants. Cognition, 117, 1-8.
Gelman, S. A. (2003). The essential child: Origins of essentialism in everyday thought. New York: Oxford University Press.
Gelman, S. A., Star, J., \& Flukes, J. (2002). Children's use of generics in inductive inferences. Journal of Cognition and Development, 3, 179-199.
Gelman, S. A., \& Wellman, H. M. (1991). Insides and essences: Early understandings of the non-obvious. Cognition, 38(3), 213-244.
Gergely, G., Egyed, K., \& Király, I. (2007). On pedagogy. Developmental Science, 10, 139-146.
Goodman, N. (1965). Fact, fiction, and forecast. Indianapolis, IN: Bobbs-Merrill Company.
Hollander, M. A., Gelman, S. A., \& Raman, L. (2009). Generic language and judgements about category membership: Can generics highlight properties as central? Language and Cognitive Processes, 24, 481-505.
Keil, F. C. (1989). Concepts, kinds, and cognitive development. Cambridge, MA: MIT Press.
Lopez, A., Gelman, S. A., Gutheil, G., \& Smith, E. E. (1992). The development of category-based induction. Child Development, 63, 1070-1090.
Markman, E. M. (1989). Categorization and naming in children: Problems of induction. Cambridge, MA: MIT Press.
Schulz, L. E., Standing, H. R., \& Bonawitz, E. B. (2008). Word, thought, and deed: The role of object categories in children's inductive inferences and exploratory play. Developmental Psychology, 44, 1266-1276.
Yoon, J. M. D., Johnson, M. H., \& Csibra, G. (2008). Communication-induced memory biases in preverbal infants. Proceedings of the National Academy of Sciences of the USA, 105, 13690-13695.

# A Computational Model of General Rule Learning with Unnatural Classes 

Shira Calamaro (shira.calamaro@ yale.edu)<br>Department of Linguistics, 370 Temple Street<br>New Haven, CT 06511 USA


#### Abstract

This paper presents the results of a computational model of generalized phonological rule learning (Calamaro and Jarosz, 2012), which is used to model experimental studies on the learning of phonotactic patterns governed by natural and unnatural classes. I focus on two papers with conflicting results on the learnability of natural and unnatural rules. Saffran and Thiessen (2003) find that a phonotactic pattern of positional voicing restrictions governed by a natural class of segments is learned by infants, but a similar pattern governed by an unnatural class is not learned. In contrast, Chambers, Onishi, and Fisher (2003) find that infants can learn a phonotactic pattern governed by an unnatural class of segments. The computational model presented in this paper is able to account for these seemingly conflicting results, explaining both the learnability and unlearnability of rules governed by unnatural classes.


Keywords: Linguistics; Phonology; Language Acquisition; Computational Model; Statistical Learning

## Introduction

Many artificial-language experiments have explored the learnability of sound patterns in acquisition and how these may reflect biases in the phonology. The interpretation of experimental results can often be attributed to a number of different theoretical models. In this paper, I explore the results of experimental studies on the learnability of unnatural rules and provide an analysis in a computational model.

Experiments in language acquisition have found conflicting results in the learnability of unnatural sound patterns. In one study, Saffran and Thiessen (2003) found that infants were able to learn phonotactic voicing restrictions governed by a natural class of segments, but were unable to learn the same pattern when governed by an unnatural class. In contrast, Chambers, Onishi, and Fisher (2003) have shown that phonotactic patterns governed by unnatural classes may be learnable.

In this paper, I present a computational model of generalized rule learning (Calamaro and Jarosz, 2012), which offers an account of the results found by Saffran and Thiessen (2003) and Chambers et al. (2003). This model uses statistical regularities in the input, as well as linguistic filters, to learn phonological alternations. It encodes these patterns as generalized rules over natural classes of segments, which can explain the inability to generalize certain patterns that do not fall into a natural class.

The results of these computational experiments will help to further clarify the nature of the results of the acquisition studies. The preference for patterns governed by natural
classes over unnatural classes may be explained by the inability to generalize over certain classes of segments. This preference is realized in the model through a generalization bias, or preference for general rules. The learnability of some types of unnatural rules is also explained by the model, which can identify the robust patterns present in the data and distinguish between them through the interaction of complexity and competition.

## Background

In their artificial-language experiments on phonological acquisition, Saffran and Thiessen (2003) attempt to find the types of patterns that are learnable by infants and identify the types of pattern that are more difficult to learn. 9-monthold infants were trained on a set of language data exhibiting the specified pattern. They were then tested using the headturn preference procedure, in which listening times for familiar and novel words were measured. A significant difference in listening times would indicate which patterns had been learned by the infants after a brief training period.

In one experiment, they looked at the learning of voicing restrictions in different positions of a syllable. Using two conditions, they restricted the types of consonants that could appear in the onset, the position preceding the vowel, and the coda, the position following the vowel. In one condition, the onset position was restricted to the set of voiceless stops $[\mathrm{p}, \mathrm{t}, \mathrm{k}]$, while the coda was restricted to voiced stops $[\mathrm{b}, \mathrm{d}, \mathrm{g}]$. For example, words of the form pibtad were permitted, but not *bipdat. In the second condition, the restrictions were reversed, with voiced stops in the onset and voiceless stops in the coda. The sets $[p, t, k]$ and $[b, d, g]$ each form a natural class of stop consonants because they can be distinguished using a single feature, [voice]. The results showed that infants were indeed capable of learning this distinction, with a significant difference in looking times between familiar and novel words. In this experiment, the infants were able to learn a phonotactic pattern governed by a natural class of segments.

The next experiment investigated the learning of voicing restrictions of unnatural classes in different prosodic positions. Unlike the previous experiment, in which the sets $[\mathrm{p}, \mathrm{t}, \mathrm{k}]$ and $[\mathrm{b}, \mathrm{d}, \mathrm{g}]$ could be distinguished by the [voice] feature, the sets used in the second experiment cannot be distinguished by any feature, making them unnatural. In one condition, [p,d,k] appeared in the onset while $[b, t, g]$ appeared in the coda. The reverse was true in the second condition, with $[\mathrm{b}, \mathrm{t}, \mathrm{g}]$ appearing in the onset and $[\mathrm{p}, \mathrm{d}, \mathrm{k}]$ occurring in the coda. The experimental results differed, with no significant difference in looking times between the familiar and novel words. In the experiment, the infants
failed to learn a phonotactic pattern governed by an unnatural class of segments.

Overall, the Saffran and Thiessen (2003) results show that infants are capable of learning patterns over a set of segments that form a natural class and can be described by a minimal number of features, and it is more difficult to learn a pattern over an unnatural class of segments which cannot be described by any set of features.

In contrast, Chambers et al. (2003) have shown that infants are capable of learning phonotactic patterns governed by an unnatural class of segments. In this experiment, 16.5 -month-old infants were tested using a head-turn preference test. The training data consisted of artificial CVC words in which the set of segments [b, k, m, $\mathrm{t}, \mathrm{f}]$ and $[\mathrm{p}, \mathrm{g}, \mathrm{n}, \mathrm{t}$, s] were restricted by position, appearing in either the onset or coda. There is no combination of features that can be used to define these segments, so these sets of segments constitute an unnatural class. In the testing phase of the experiment, infants were able to distinguish between legal and illegal words, meaning they had learned the phonotactic pattern they had been trained on.

The results found by Chambers et al. (2003) seem to be in conflict with the results found by Saffran and Thiessen (2003) on the learnability on rules governed by unnatural classes. In addition to the distinction between natural and unnatural classes, a learning model should also be able to account for these different results on the learnability of unnatural classes. In the next section, I present such a model to account for these results.

## Generalized Rule Learning Model

The Generalized Rule Learning model (GRL: Calamaro and Jarosz 2012) presented here is used to test the learning of the acquisition data in a computational setting. The GRL is a statistical model with linguistic constraints and generalized rule learning. The generalization component of the model is motivated by experimental evidence showing that infants are able to generalize rules using features (Maye, Weiss, and Aslin 2008; Cristiá and Seidl 2008). Given a set of segmented data, the model learns general rules for alternations in the data at the contexts in which they occur, as well as a score reflecting the strength of the rule. The original goal of the GRL model was the learning of alternations, such as word-final devoicing in Dutch, but in this paper it is applied to static phonotactic patterns. The GRL model is based on an earlier model (PLND: Peperkamp, Le Calvez, Nadal, and Dupoux, 2006) for learning pairs of alternating segments by calculating their statistical distribution in the data with an application of KLdivergence (Kullback and Leibler, 1951) and linguistic filters.

The GRL model maintains the use of the two linguistic filters ${ }^{1}$ from PLND, which remove spurious pairs that should not be considered as alternating segments for linguistic reasons. The first filter removes pairs which have

[^25]an intervening segment in the phonetic space based on their features, which are represented by a vector with values for place, sonority, voicing, nasality, rounding, and vocalic. ${ }^{2}$ The second filter removes pairs in which the allophone is not more similar to its context than the default segment. Overall, these filters are able to introduce phonological knowledge not available to a purely statistical model.
The GRL model also maintains use of KL-divergence, though the formulation is somewhat changed, with the new calculation shown in (1) :
(1) $\operatorname{Score}_{(\mathrm{c}, \mathrm{s} 1, \mathrm{~s} \mathbf{2})}=D\left(c, s_{1}, s_{2}\right) * \mathbf{Z}^{2}$
\[

$$
\begin{aligned}
& \text { Where: } \\
& D\left(c, s_{1}, s_{2}\right)=P\left(c \mid s_{1}\right) \log \frac{P\left(c \mid s_{1}\right)}{P\left(c \mid s_{2}\right)}+P\left(c \mid s_{2}\right) \log \frac{P\left(c \mid s_{2}\right)}{P\left(c \mid s_{1}\right)} \\
& \text { and } Z=\frac{K L\left(c, s_{1}, s_{2}\right)-\mu}{\sigma}
\end{aligned}
$$
\]

The equation in (1) is used to calculate scores for pairs of alternating pairs of segments at a context $c$, defined as the following segment. The use of KL-divergence to find alternating pairs captures the intuition that segments which have highly distinct distributions in the data are likely to governed by some phonological or phonotactic rule.
The model creates general rules by merging alternating pairs which undergo an identical structural change, as represented by a feature vector. For example, the alternating pair ( $\mathrm{d}, \mathrm{t}$ ) has a structural change of $[0,-1,-1,0,0,0]$, calculated as the difference between the feature vector of segments $t$ : $[4,1,0,0,0,0]$ and $d:[4,2,1,0,0,0]$. This difference vector represents the devoicing pattern of the ( $\mathrm{d}, \mathrm{t}$ ) pair.
The scores of alternating pairs as calculated in (1) are summed for all pairs whose change in features is the same, giving a contextualized rule score. Each contextualized rule is represented by a structural change vector, the context in which it occurs, and a rule measuring its strength. The calculation of contextualized rule scores is shown in (2):
(2)

$$
\operatorname{Score}_{(c, \vec{f})}=\sum_{(s 1, s 2)} \operatorname{Score}_{(c, s 1, s 2)}
$$

The output of the formula in (2) is a set of rules which each apply at a single context. Many phonological rules apply at multiple, related contexts, such as a vowel nasalization rule that applies in the context of all nasal segments. The contextualized rule scores can be further generalized, by merging rules whose contexts are phonologically related to each other and the change undergone by the rule. The formal calculation of the rule merging is defined in (3):
(3)

$$
\text { Score }_{(\{C\}, \vec{f})}=\sum_{\mathbf{c} \in \mathrm{C} \text { where SCC and SvC hold }} \text { Score }_{(\mathrm{c}, \vec{f})}
$$

[^26]Shared Change Condition (SCC): To merge, contexts must share feature values for any non-zero values in $\overrightarrow{\boldsymbol{f}}$.

Shared Values Condition (SVC): To merge, contexts must not differ along more than one feature.

The formula in (3) is used to calculate the score of a generalized rule as the sum of all rules whose contexts meet two conditions: the Shared Change Condition (SCC) and the Shared Values Condition (SVC). Like the linguistic filters from Peperkamp et al (2006), the merging conditions provide linguistic information in assigning classes of sounds that pattern together. The SCC requires that contexts must be related to the rule change in the same way by restricting merging to contexts which share non-zero values of the rule vector. The SVC requires that contexts be related to each other by restricting merging to contexts which only differ along a single feature dimension, thus approximating a natural class. The merging of contextualized rules into generalized rules can capture generalizations about the data as well as assign increased scores to more robust rules occurring in a set of related contexts.

This model learns generalized rules as a difference vector of features, a set of contexts of application, and a score indicating the goodness of the rule. The rules learned by the model will need to be interpreted somewhat differently from the results of the Saffran and Thiessen (2003) and Chambers et al. (2003) experiments, which measured successful learning by significant differences in looking times in a head-turn test. Instead, this model will need to look for rules which reflect the regularities found in the training data. Additionally, the model looks at alternations conditioned by contexts defined as following segments and does not have access to syllabic structure. Due to these limitations of the model, this discussion will focus on the results as they relate to the learning of the pattern in coda position, which is defined by the following segment. With these restrictions in mind, successful replication of results in the model will mean the learning of a word-medial and word-final voicing/devoicing rule in Experiment 1, no successful learning of any such a rule in Experiment 2, and the learning of meaningful rules in Experiment 3.

## Experiment 1: Learning rules governed by natural classes

In Experiment 1, I replicate the results of an experiment by Saffran and Thiessen (2003), in which infants were able to learn voicing restrictions by position.

## Method

The Generalized Rule Learning Model, as described in the previous section, was used.

## Data

The same training data from Saffran and Thiessen (2003) was used. Each condition in the training data consisted of 30 unique CVCCVC words for each condition, made from an
alphabet with four vowels $[\mathrm{a}, \mathrm{i}, \mathrm{o}, \mathrm{u}]$, three voiceless stops $[\mathrm{p}, \mathrm{t}, \mathrm{k}]$, and three voiced stops [b, d, g]. In condition a, voiced stops were restricted to coda position and voiceless stops were restricted to onset position. The opposite was true for condition $b$, with voiceless stops occurring in coda position and voiceless stops occurring in onset position.

While the model does not specifically reference syllable structure, successful learning of this data would find a rule of voicing/devoicing in the word-final context and before voiceless/voiced consonants.

## Results

The results from experiment 1 are shown in Figure 1, reflecting the highest scoring rules found by the model.


Figure 1: Exp. 1 results
Each bar in Figure 1 shows the score of a generalized rule. In Condition a, the highest scoring rule is the wordfinal voicing rule, (\# [0, 1, 1,0,0,0]), where \# represents the word-final context and $[0,1,1,0,0,0]$ represents the structural change vector. The two non-zero values in the vector indicate a change in the sonority and voicing features in pairs such as $(\mathrm{t}, \mathrm{d}) .{ }^{3}$ The reverse rule is found in Condition b, with a structural change vector $[0,-1,-1,0,0,0]$ indicating devoicing in pairs such as $(\mathrm{d}, \mathrm{t})$.

In each of the two conditions, the highest scoring rule is the desired voicing or devoicing rule. This rule reflects the change in voicing of the stops in coda position in the training data. The voicing/devoicing rule is quite robust in each of the two conditions, scoring much higher than the next highest scoring rule. A similar rule for word-medial codas is also found, which is the voicing/devoicing rule occurring in $\{\mathrm{p}, \mathrm{t}, \mathrm{k}\}$ or $\{\mathrm{b}, \mathrm{d}, \mathrm{g}\}$ contexts.

A number of spurious rules were also found by the model. These rules reflect a change in place of articulation, shown as fronting and backing rules. While these rules are not desired, they do reflect a generalization in the data, namely, a possible alternation between pairs like [ $\mathrm{p}, \mathrm{t}]$ or $[\mathrm{t}, \mathrm{k}]$, in which the segments differ only in place of articulation. These spurious rules are likely an artifact of the small

[^27]segment inventory, making a minor statistical regularity appear to reflect a possible alternation. In artificial language learning, these types of spurious statistical regularities have the potential to affect the results to a greater extent than in natural language learning, as will be seen in Experiment 3.

Overall, the results in Experiment 1 show learning of the phonotactic pattern, aligning with the results found by Saffran and Thiessen (2003). The model successfully learned the voicing restrictions when they were governed by a natural class of segments.

## Experiment 2: Failure to learn rules governed by unnatural classes

In experiment 2, I replicate the results of a second experiment from Saffran and Thiessen (2003), in which infants were not able to learn phonotactic restrictions of unnatural classes of segments which are specified by voice and place of articulation.

## Method

The Generalized Rule Learning Model, as used in the previous experiment.

## Data

The same training data from the Saffran and Thiessen (2003) experiment was used. The training data consisted of 30 CVCCVC words in each condition with the same alphabet as experiment 1 . In condition a, the set of coda consonants was $[\mathrm{b}, \mathrm{t}, \mathrm{g}]$ and the set of onset consonants were [ $\mathrm{p}, \mathrm{d}, \mathrm{k}]$. In condition b the voicing specifications were reversed, with codas [p, d, k] and onsets [b, t, k].

## Results

In Exp. 2, the model failed to learn voicing restrictions governed by unnatural classes. These results are shown in Figure 2, with the highest scoring rules represented.


Figure 2: Exp. 2 results
The voicing and devoicing rules are no longer learned as the highest scoring rules, as seen in Figure 2. The highest scoring rule is now a spurious rule reflecting a change in the
place of articulation. This rule would account for a possible alternation between pairs such as ( $\mathrm{p}, \mathrm{t}$ ) or ( $\mathrm{d}, \mathrm{g}$ ), which can be generalized from statistical regularities in the data.

The desired rules of voice alternations receive lower scores than some of the spurious rules. The overall strength of these desired rules has decreased, with the weight of each voicing/devoicing rule being split into two lower weighted rules. The reason for this decrease in the score is that the patterns cannot be fully generalized because they belong to an unnatural class. In experiment 1 , the desired voicing rules were supported by three pairs of segments, one for each place of articulation. In this experiment, the scores were split between two separate rules, each supported by one or two pairs of segments, $(\mathrm{t}, \mathrm{d})$ or $(\mathrm{b}, \mathrm{p})$ and $(\mathrm{g}, \mathrm{k})$.

Both factors of decrease in rule rank and loss of rule strength contribute to the increased difficulty of learning the phonotactic pattern in experiment 2 . This difficulty in learning is a desired result because infants failed to learn this same pattern in an experimental setting (Saffran and Thiessen 2003).

In this case, the unnatural voicing pattern was not the most robust pattern in the data. The model found other patterns which were generalizable from the given data, obscuring the desired patterns. From this result, a prediction of the model is that it would be able to learn a rule governed by unnatural classes, if the data did not contain any other patterns which could be inferred. Such a case is used by Chambers et al. (2003), which will be shown in the following experiment.

## Experiment 3: Learning rules governed by unnatural classes

In a final experiment, I run the GRL model on the data from Chambers et al. (2003), in which infants were able to learn phonotactic patterns governed by an unnatural class of segments.

## Method

The Generalized Rule Learning Model, as used in the previous experiments.

## Data

The data used in this experiment were replicated from Chambers et al. (2003). A set of CVC words were creating using two groups of consonants belonging to an unnatural class: $[\mathrm{b}, \mathrm{k}, \mathrm{m}, \mathrm{t}, \mathrm{f}]$ and $[\mathrm{p}, \mathrm{g}, \mathrm{n}, \mathfrak{t}, \mathrm{s}]$. The onsets were drawn from one group and codas from another, creating a phonotactic pattern governed by an unnatural class of segments.

## Results

While the data could not be generalized, the patterns were learnable as separate rules, as shown in Figure 3:


Figure 3: Exp. 3 results
The rules shown in Figure 3 are striking due to the uniformity of the data. While the segments could not be generalized by position, the model was able to find relationships among between-group segments. For example, the ( $\mathrm{b}, \mathrm{p}$ ) pair is reflected by the word-final devoicing rule (\# $[0,-1,-1,0,0,0]$ ), while the ( $k, \mathrm{~g}$ ) pair is reflected by the word-final voicing rule (\# [0,1,1,0,0,0]).

With a one-to-one mapping of segments to learned rules, we would expect five rules, but instead find eight. While each segment belongs to at least one rule, some segments are learned as multiple rules. For example, ' $p$ ' is found in both the devoicing rule (\# [0,-1,-1,0,0,0]) with the pair (b,p), but also in the fortition rule (\# [-1,-2,0,0,0,0]) with (f,p).

While some of these are the same rules which were unlearnable in Experiment 2, namely voicing and devoicing, a potential difference here is the lack of interference from spurious rules. While in the case demonstrating unlearnability, the desired rules were dominated by spurious rules. In this experiment, the desired rules were the highest scoring rules.

## Discussion

The computational experiments presented in this paper seek to address two fundamental questions about the learnability of phonotactic patterns: Why are patterns governed by natural classes easier to learn than those governed by unnatural ones? How can we explain results in which unnatural patterns are learnable? The first question is addressed by comparing the results of Experiments 1 and 2, and the second by comparing the results of Experiments 2 and 3.

## Natural vs. Unnatural Classes

In Experiments 1 and 2, the GRL replicated the results found by Saffran and Thiessen (2003), that a phonotactic pattern governed by a natural class is learned, while one
governed by an unnatural class is not. Specifically, infants can learn patterns which occur over a set of segments that all agree in voicing and differ in place, they cannot learn patterns which occur over a set of segments that differ in both place and voicing.

The GRL finds an asymmetry in the learning of natural and unnatural classes due to an inherent bias in the generalization mechanism. Generalized rules receive higher scores from the model because they have support from more pairs of segments. The strength of general rules is computed by summing the scores of rules governing alternations between a single pair. Therefore, the more pairs of segments contributing to a general rule, the higher its score will be. In the case of the Saffran and Thiessen (2003) data, the rules governed by natural classes are supported by more segments than the unnatural ones. This inherent generalization bias assigns higher scores to the natural rules in Exp. 1 than the unnatural rules in Exp. 2.

The asymmetry in the learning of natural and unnatural rules has previously been explained by a Complexity Bias (Moreton and Pater, 2011). Under this account, the more complex set of features needed to describe unnatural classes makes the learning of unnatural patterns more difficult. Natural classes, which can be described with fewer features, can be learned more easily.

The generalization mechanism in the GRL accounts for the same patterns as the Complexity Bias, but for a different reason. While the Complexity Bias asserts that unnatural rules are more difficult to learn because they require the encoding of additional feature values, the GRL attributes this asymmetry to weaker statistical regularities due to the more complex data. This prediction of the GRL can be seen by the difference in rules scores in Exp. 1 versus Exp. 2.

The GRL model has an additional property that interacts with complexity: competition. In the results from Exp.1, the desired pattern was learned because of the high score relative to other rules. In Exp. 2, the lower scoring unnatural rules were dominated by competing spurious rules, interfering with their learnability. This interaction between complexity and competition allows the GRL to make additional predictions beyond complexity alone, which will play a role in the learning of different types of unnatural classes.

## Unnatural vs. Unnatural Classes

In Experiments 2 and 3, the learning data contained phonotactic patterns governed by unnatural classes. In the original experimental setting, infants did not learn the unnatural pattern in Experiment 2 (Saffran and Thiessen 2003), but did learn the pattern in Experiment 3 (Chambers, et al. 2003). Likewise, the GRL found a similar difference in the learnability of the two unnatural patterns, as shown in this paper. The distinction to be made between these two unnatural patterns lies in the nature of the data.

Both experiments presented artificial data in which syllable positions were restricted to a specific set of consonants. In Saffran and Thiessen (2003) the sets were [p,
$\mathrm{d}, \mathrm{k}]$ and $[\mathrm{b}, \mathrm{t}, \mathrm{g}]$; in Chambers et al. (2003) they were [b, k, $\mathrm{m}, \mathrm{t}, \mathrm{f}]$ and $[\mathrm{p}, \mathrm{g}, \mathrm{n}, \mathrm{t}$, s]. While both sets of data are unnatural to some extent, there is a striking difference in the segment inventories of the two experiments.

While the pattern presented in Saffran and Thiessen (2003) is unnatural, the segment inventory is well-balanced among the feature set it uses, with a voicing distinction present for each place of articulation. In contrast, the segment inventory of Chambers et al. (2003) is not as balanced, with a mix of voicing, place and sonority distinctions that do not apply across all pairs of segments. For example, there is a voicing distinction for the pairs ( $\mathrm{p}, \mathrm{b}$ ) and ( $\mathrm{k}, \mathrm{g}$ ), but there exists no pair ( $\mathrm{t}, \mathrm{d}$ ).

The effects of the overall naturalness of the data are seen directly in the computational results of Experiments 2 and 3. In Experiment 2, the more balanced data allowed the GRL to make a number of spurious generalizations, obscuring the robustness of the desired unnatural rules. In Experiment 3, the less balanced data could not be generalized by the model, leaving the set of desired unnatural rules as the most robust in the data.

In the distinction between these two sets of unnatural patterns, the GRL is better able to predict these results than a model using complexity alone. The Complexity Bias (Moreton and Pater 2011) predicts difficulty in the learning of both types of unnatural patterns, but would predict even greater difficulty in Exp. 3 due to the greater number of features needed to describe the unrelated set of segments. However, the experimental evidence shows the opposite is true, with the data in Exp. 3 being learned more easily. The predictions of the GRL align with the experimental evidence due to the interaction of competition and complexity in the model. The unnatural pattern in Exp. 3 is learned more easily than that in Exp. 2 because the desired rules are not in competition with any high scoring spurious rules as is the case in Exp. 2.

## Conclusion

The GRL is able to model the results of experimental data showing the learning of phonotactic patterns by infants. It can account for the preference for learning natural rules over unnatural ones, as well as the distinction between the learnability of different patterns of unnatural classes. This preference for natural classes is an inherent property of the model, due to the rule generalization component. While the generalization component of the model does facilitate the learning of natural rules, it does not exclude the learning of rules governed by unnatural classes. Indeed, rules governed by unnatural classes were learned by the model, when there were no other more robust rules in the data.

These experiments provide some promising results for the GRL model, with its ability to account for attested cases of phonological learning. While there remains a possibility that differences in infant learning can be attributed to differences in experimental methodologies, these results show compelling evidence for further exploration of this topic.

Future work will explore other predictions made by the model and extensions needed to account for additional data.

## Appendix

## Linguistic filters (Peperkamp et al. 2006)

Allophonic distributions of $\mathrm{s}_{\mathrm{a}}$ and $\mathrm{s}_{\mathrm{d}}$ are spurious if:

$$
\exists s\left[\begin{array}{c}
\forall i \in\{1, \ldots, 5\}, v_{i}\left(s_{a}\right) \leq v_{i}(s) \leq v_{i}\left(s_{d}\right) \\
\operatorname{or} v_{i}\left(s_{d}\right) \leq v_{i}(s) \leq v_{i}\left(s_{a}\right)
\end{array}\right]
$$

With $v_{i}(s)$ the ith component of the vector representation of s.

Allophonic distributions of $\mathrm{s}_{\mathrm{a}}$ and $\mathrm{s}_{\mathrm{d}}$ are spurious if:

$$
\exists i \in\{1, \ldots, 5\},\left|\sum_{s \in C\left[s_{a}\right]}\left(v_{i}\left(s_{a}\right)-v_{i}(s)\right)\right|>\left|\sum_{s \in C\left[s_{a}\right]}\left(v_{i}\left(s_{d}\right)-v_{i}(s)\right)\right|
$$

## Feature values

Segments are represented as feature vectors with the following values:

Place: bilabial 1, labio-dental 2, dental 3, alveolar 4, postalveolar 5, palatal 6, velar 7, uvular 8, glottal 9

Sonority: voiceless stop 1, voiced stop 2, voiceless fricative 3, voiced fricative 4 , nasal 5, lateral 6 , rhotic 7 , glide 8 , high vowel 9 , mid vowel 10, low vowel 11

Voicing: voiceless 0 , voiced 1
Nasality: oral 0, nasal 1
Rounding: unrounded 0, rounded 1
Vocalic: non-vowel 0, vowel 1

## Acknowledgments

Many thanks for feedback from Gaja Jarosz, CLAY Lab, audiences at LSA 2013, and an anonymous reviewer.

## References

Calamaro, S. and Jarosz, G. (2012). A computational model of general rule learning for phonological alternations. Ms.
Chambers, K.E., Onishi, K.H., and Fisher, C. (2003). Infants learn phonotactic regularities from brief auditory experience. Cognition 87 B69-B77.
Christiá, A. and Seidl, A. (2008). Is infants' learning of sound patterns constrained by phonological features? Language Learning and Development 4(3), 203-227.
Kullback, S. and Leibler, R. (1951). On information and sufficiency. Annals of Mathematical Statistics 22, 76-86.
Maye, J., Weiss, D.J., and Aslin, R.N. (2008) Statistical phonetic learning in infants: Facilitation and feature generalization. Developmental Science(11)1, 122-134.
Moreton, E. and Pater, J. (2011). Learning artificial phonology: A review. Ms.
Peperkamp, S., Le Calvez, R., Nadal, J.P., and Dupoux, E. (2006). The acquisition of allophonic rules: Statistical learning with linguistic constraints. Cognition 101(3), B31-B41.
Saffran, J. and Thiessen, E. (2003). Pattern Induction by Infant Language Learners. Developmental Psychology Vol. 39, No. 3, pp. 484-494.

# From individual minds to social ones. Cooperation and the structure of animal and human societies. 

Valentina Cardella (vcardella@unime.it)<br>Department of Cognitive Sciences, 6 via Concezione<br>Messina, Italy

Alessandra Falzone<br>Department of Cognitive Sciences, 6 via Concezione<br>Messina, Italy

Antonino Pennisi<br>Department of Cognitive Sciences, 6 via Concezione<br>Messina, Italy


#### Abstract

The aim of this paper is to highlight the role of the social factors among the different forces which influence natural selection. To do this, we'll start analyzing an example of highly complex society, like that of the baboons, in order to show that the building-up of a society depends on extremely flexible and continually negotiated social relationships, rather than on features that are genetically determined. However, in our view, when we shift from animal societies to the human ones, we have to recognize the central role of language. In fact, even if the role of social influence appears to be relevant in other animal species, in humans it was language that provided the way to make this social influence much more important, to redefine the roles, to reverse the genetic dominances even more, and to make human cooperation something unique.


Keywords: cooperation; social influence; evolution; language.

## Introduction

The concept of cooperation can be particularly useful to throw a new light into the building-up of societies both in animals and humans. A more in-depth analysis of the ways in which cooperation works in different animal societies, can allow us to assign the right role, in the evolutionary perspective, to social influences and non-deterministic factors like chance, history and the single choices of the individuals belonging to one group. Our framework is the one suggested by Weiss and Buchanan, which provides "an alternative view of natural selection in which there is more slippage and tolerance, multiple solutions with larger acceptability spaces, and the possibility that an adaptive fit will be 'good enough' rather than seamless" (Weiss and Buchanan 2009, p. 305). In this perspective, we'll start analyzing an example of highly complex society, like that of the baboons, in order to show that the buildingup of a society depends on extremely flexible and continually negotiated social relations, rather than on features that are genetically determined. Baboons' society seems to be a good example of the way
cooperation can modify, and even reverse, the genetic dominances. Our examination of the role played by social influence in these non human primates will bring us to look to the interaction between society and biology in a different way, both in animals and humans.

## The survival of the fittest. Or not?

The law of evolution is that the strongest survives. This seems to be what Darwin (1859) has taught us. The fittest do survive, and produce offspring, the others simply don't. At first glance, baboon males, with their aggressive anatomy, seem to be the perfect image of the law of the strongest, and the perfect match between anatomy and behavior (Washburn and DeVore, 1961; DeVore and Hall, 1965): remarkable canines, big mantle of hair, large body size, all seem to suggest that the social rank, and therefore the reproductive success, is determined by the males' fighting ability. But then, why, on a closer inspection, the social structure turns up to be matrilinear, where females have political functions (besides motherhood) that are not defined in terms of their relationships with males? Why the male hierarchies, based on the physical structure, can't predict the priority of access to resources (including the females), and not even the winner in a fight?
Clearly, in Darwinian (and also in NeoDarwinian) terms fittest is not necessarily the strongest, but often the smartest, the swiftest, the most patient, and so on. Darwin would agree that not always the strongest survives (as claimed by hyperdarwinism). Various authors recently focus on a more general "multilevel selection" theory, and on the compatibility between individual selection, kin selection and group selection (Pievani in press). They identify the level of the population structure as the key factor, offering a
pluralistic perspective within the theory of evolution. Here, we suggest to focus on the relationship among individuals (including the echological constraints) rather than on the single member of a species. And, in the case in question, if we want not to miss these factors, we must take into account something else, apart from the role of competition; we have to refer to the role played by cooperation if we want to have a clearer idea of the way the baboons' society works. Baboons have some alternatives to aggression, social strategies that are less risky and that settle competition and defense. These alternatives are based on "special relationships". For example, a potential loser can come back to the aggressor with a baby on his belly, and thus leading the aggressor to back off. In order to be successful, however, a special relationship between the male and the baby is needed, and this relation must be grounded on trust. Thanks to this special relationship, thus, the infant will start screaming when becoming closer to the aggressor, making him the object of mobbing by the group. But without trust, the infant could scream at "the false friend" rather than at the opponent, and the male could run the risk of himself becoming the object of mobbing.
Baboons use their social abilities in other different ways. For example, an older baboon can get copulation thanks to his social skills; he can monitor a male from a distance, manipulate aggression between this male and his closer followers, and then, while he's engaged on a fight, he can dash to his consort, copulate with her and become her new partner. In general, one can notice the same trajectory through a male's life history, "from socially unskilled maturing male with growing physical powers (using aggressive strategies) to the socially skillful mature male whose success depends entirely on social expertise when he is old (using social strategies of competition and defense)" (Strum 2012, p. 6). The possibility to use different alternatives to aggression makes baboons' society more complex and less predictable. Among baboons, it is not always that the strongest wins.

## Evolution in action

Strum (1987; 2012) had the opportunity to "test" natural selection. Because of the increasing encounters with humans, Strum and his colleagues had to translocate three groups of baboons to the arid savanna of Eastern Laikipia Plateau. This was a unique opportunity to test the baboons' adaptability and to look at the moment where a behavior becomes crucial for survival and reproductive success, rather than infer it ex-post, as scientists usually do. Strum notes that "if
ever there was a situation where survival of the fittest should operate, where competition should have had an upper hand, this was it. Instead [...] the translocation, pointed out how evolutionary principles get embedded in a specific time and place and why context matters. Each group's 'natural history' illustrated a variety of different paths" (Strum 2012, p. 8).
The process of adaptation was actually much more complicated than expected, and chance and individuals' social abilities played an important role. Relationships between the immigrants and the indigenous troops were also crucial for survival, for examples an immigrant baboon could follow an indigenous one in order to "adjust" his diet or to find water.
The possibility to observe the process of adaptation in action allows us to take a hard look at the natural selection, which can't be reduced to the survival of the fittest. More precisely, when observing the behavior of a group, one can notice the large amount of flexibility (including the possibility of mistakes that don't lead to extinction) depending on the group's social complexity. Thus, we agree with Strum, when she claims that "evolutionary time speculations of cause and effect assume rather than clarify how evolutionary principles are situated by chance, contingency, and history. This is because they either lack the information or ignore the importance of context and 'black-box' the relationships between real behaviors and evolutionary outcomes" (Strum 2012, p. 10).

## Natural selection is not enough

The idea according to which the survival of the fittest is not the unique force to guide evolution is not a new one. There are in fact many elements, coming from different fields of studies, which contribute to "soften" the adaptationist aspect that Darwin's theory (especially in some of its contemporary versions) risk to endorse.
For example, the evo-devo researchers highlighted the role of structural constraints. As a matter of fact, some forms are not present in nature even if their absence can't be explained through natural selection, because they seem to be very small variations of other forms that are actually produced. These unexpected absences seem to suggest that natural selection can't help but choose among the possible variations, for it can't interfere with the fixed ways that are prescribed by the laws of development (Minelli 2009). The modern molecular biology, from its corner, seems to confirm the fundamental role played by the constraints, because the properties of very different living beings
are grounded on the same mechanism of molecular conservation. Genes involved in behavior are highly conserved during evolution (Gehring and Ikeo 1999, Carroll 2000) and one can find genetic and behavioral homologies even between traits that evolved convergentely (i.e. not from an immediate common ancestor, cf. Shubin et al. 1997). As claimed by Gould (1997, p. 36) "most biologists feel that such stability acts primarily as a constraint upon the range and potentiality of adaptation, for if organisms of such different function and ecology must build bodies along the same basic pathways, then limitation of possibilities rather than adaptive honing to perfection becomes a dominant theme in evolution. At a minimum, in explaining evolutionary pathways through time, the constraints imposed by history rise to equal prominence with the immediate advantages of adaptation".
Moreover, we must look at the role of adaptations from the perspective of long-term evolutionary trends. These trends show that evolution can't be described as a gradual accumulation of adaptations within a population. The actual picture is much more similar to the punctuated equilibrium described by Gould and Eldredge (1977), long periods of stability (even million years of stasis) punctuated by the brunching of new species, that only in limited cases can be described as adaptative improvements of some individuals within a species.
In this long-term perspective, the adaptations with immediate effect can even play no role in survival. In the North American Pleistocene megafauna extinctions the species that went extinct did not have characteristics that set them apart from those that survived; so these processes are not deterministic but they are contingent and complex because the individual's behavioral responses are often unpredictable (cf. Gilbert and Epel, 2009 and their eco-evo-devo).
But what we want to stress here is that, among the different forces which influence natural selection, the social factors have to be taken into account. Thus, to come back to this influence, we have to mention another kind of selection, the sexual one.
The idea that the stronger the male is, the more reproductive success he has, comes from Darwin's theory of sexual selection. According to this theory, males and females obey to different behavioral patterns: the former would be passionate, the latter coy. "Males of almost all animals have stronger passions than females [...] The female, on the other hand, with the rarest exception is less eager than the
male" (Darwin, 1871, pp. 272-3). Males would fight over the possession of females, and females would be agents of selection, choosing only the most virile and showy males, because these features would guarantee many healthy descendants.
Evolutionary psychology, that applies sexual selection also to humans, gave us a very simplified (and quite grotesque) picture of the relationship between male and female (and therefore between man and woman). Males would be more promiscuous because sperm are supposed to be cheap and they would therefore spend all the time looking for females to fertilize. Females, on the other hand, would be more "choosy" because pregnancy is expensive and they would have to guard themselves from wrong investments into "bad" genes. But, as explained by Roughgarden in her book Rainbow (2004), what actually happens in nature proves these theories wrong. Males are not always passionate and females are far from being always coy. Once again, baboons' society shows us that females often present to males, but are refused by them. Why should males refuse females in heat when sperm is supposed to be cheap? Moreover, females don't select always the best genes. They can prefer mating with males who turn out to be good fathers (e.g., in protecting the eggs) rather than males that are dominant in competition with other males. For example, whether a male of sand gobies is dominant in competition for nests does not correlate with whether he is a good father in protecting the eggs. And the goby female knows it; in fact, an experiment reported by Roughgarden (2004) showed that the females of this species regularly select the males that would protect the eggs, and that they don't care if the male they preferred won some fights against other males. These cases show us that male success rests on more than the male dominance hierarchy, and that it depends on different factors, including the relationships based on collaboration that he is able to build (as baboons have taught us). So, it seems that, as Strum says in her paper (2012, p. 15): "there are multiple ways for individuals to succeed, not just one optimal evolutionary path".

## From a complex society to a complicated one. Language and cooperation

Nature provides us with various examples of strongest males that are not successful. But we would like to stress above all that it provides us also with examples of extremely complex societies, with constant negotiations about roles, where the strongest doesn't always win over the weaker, and where a large amount
of intertwined relationships reverses the genetic hierarchy. These data show that society is nature, and that there's no opposition between the two terms.
But there is no difference in complexity between the baboon's society and the human one, then? We could say that baboons are pretty much human? It is evident that we won't find, in the animal kingdom, something that compares to the social and technological complexity of Homo sapiens. Nevertheless, baboons' society reveals the shared roots of social complexity which, in our continuistic framework, found our society, too. But what made human sociality all the more complex than baboons' one?
We believe that what makes our sociality so complex is the presence of language. First of all, the faculty of language seems to be one of the variables that allowed that jump which 40.000 years ago gave rise to the first symbolic uses and to the sudden and explosive variation of human tools. As claimed by Spelke (2009), the technological abilities specific to humans probably descend from the possibility to combine representations of objects and representation of actions in a rapid and flexible way, and this ability depends on the combinatoriality typical of human language.
Moreover, language gave us the opportunity to shift from a linear transmission to a cumulative one. Horner and coll. (2006) described in an experiment that chimpanzees are able to spread learned techniques in a linear way. The authors examined a diffusion chain paradigm, whereby a behavior is passed from one individual to the next in a linear sequence: different methods used to obtain food were accurately transmitted along two chains of chimpanzees, with a remarkable fidelity of transmission within each chain. Faithful transmission is not therefore a distinctive feature of human species. But there is a significant difference between linear and cumulative transmission (Tomasello, 1999; Richerson and Boyd, 2008). If humans had only the possibility to faithfully spread a technique across multiple generations, the different practices of tool making would have been almost unchanged through the hominids' evolution. Humans, on the contrary, are able to refine the acquired techniques and to transmit these improvements, so that we won't have to start all over each time. According to our vision the cumulative transmission, which is possible thanks to the cognitive abilities specific to humans, changed its nature with the emergence of language. And this doesn't mean that language simply "helped" to transmit the traditions and made this transmission more rapid. The simple fact that, thanks to language, it is possible to learn how to use a tool in
its absence, inaugurated a completely new way to transmit knowledge. Moreover, a real technology implies the ability to plan and design combination of tools that are more and more complex, and it is also language that develops and improves this planning ability in a significant way.
At the end of her paper, Strum asked herself what would happen to baboons' society if they wore hats (Strum 2012). If they had the possibility to symbolize their roles this way, these roles would be stable in time and space and baboons wouldn't need to point them out or continually negotiate them. But baboons don't use these kinds of symbols, and therefore they can't simplify their negotiations and they have to spend a lot of time making them evident and visible. In other words, they can't manage more complexity and they can't build a complicated large-scale society, which we sapiens can do thanks to language.
But, in our view, language isn't just a set of labels that gave us the opportunity to spare time. Or, in other words, it is not just a tool of communication, as claimed by the vast majority of evolutionary linguistic theories (cf. e.g. Dunbar 1998, Bickerton 1990). Language is a kind of technology, that manages the manipulation of perceived elements into representations and can conversely express inner needs by linking them to the external world (Pennisi and Falzone, 2011). Language is species-specific to sapiens for it is an auditory-vocal technology (speech making) applied to symbolic needs and highly specialized, just like the manual technology (tool making). We don't need to refer to a human specialty or uniqueness: the auditory-vocal technology relies on the evolution of structures and functions which come from a long evolutionary history, but that in nonhuman primates acquired an adaptive role thanks to the social organization (and to other aspects like passage to bipedalism, parental care or social learning).
The thesis according to which language is a modality to represent the world rather than just a tool of communication doesn't belong only to the philosophical and phenomenological framework today. PET and ERP studies showed that language is heavily involved in the building-up of our knowledge about the world, that kind of knowledge which we build because we belong to a social and linguistic specific community. Hagoort (2005) found in the Broca's area the neural locus where the "checking" between linguistic expressions and social reality would occur. These data seem to confirm our hypothesis according to which language is a technology that serves to
represent reality.
Since we have language, it's not possible for us to choose another technological modality to express our symbolic needs, apart from the auditory-vocal technology. Humans can't help but use language to represent the world. So the possibilities that modern sapiens' biological features gave us granted the development and use of a function which worked as a catalyst for the others that were already present and phylogenetically inherited. These abilities have been improved thanks to the interaction with the possibilities of segmentation which language gave us. Finally, language provides a cue to social preferences, which in turn have an effect on cooperation. Experiments with children from five months to five years provided evidence for an early developing social preference for members of one's native language group; even infants who have not begun to produce or understand speech show this preference. This implies that children prefer to play or work together with people of their native language, and that "infants and young children are selective social learners and cooperators, and language provides one basis for this selectivity" (Kinzler et al., 2012, p. 2). Thus, language isn't simple a way to improve communication skills, but it conditions the way we perceive the others (familiar or not, friend or not) and the possibility to collaborate with them. This is a clear example of how language can have a cognitive function in humans, rather than simply a communicative one. In our view, this cognitive role is linked to the articulated nature (speech-making technology) of our language (Pennisi, in press).
Moreover, linguistic development seems to have an effect on the ability to cooperate with peers: children who are more skilled in talking about their own and others' actions and internal states and who refer to themselves and others using personal pronouns are also better at cooperating with their peers (Brownell et al. 2006).
Thus, language seems to be the key factor of human sociality. Even if we can't underestimate the role of social influence in other animal species, because it appears to be huge also without language, we have to consider that, in our species, language provided the way to make this social influence much more important, to redefine the roles, to reverse the genetic dominances even more, to create societies grounded on values, like money or institutions, which have been negotiated once for all, and, finally, it was language that made human cooperation something unique.

## References

Bickerton, D. (1990). Language and species. Chicago: University of Chicago Press.
Brownell, C. A., Ramani, G. B., Zerwas, S. (2006). Becoming a social partner with peers: cooperation and social understanding in one- and two-year-olds. Child Development, 77, 4, 803-821.
Carroll, S. B. (2000). Endless forms: the evolution of gene regulation and morphological diversity. Cell, 101, $577-$ 580.

Darwin, C. R. (1871). The descent of man, and selection in relation to sex. London: John Murray. Vol. 1.
DeVore, I., Hall, K. R. L. (1965). Baboon ecology. In: DeVore I, editor. Primate behavior: field studies of monkeys and apes. New York: Holt, Rinehart and Winston.
Dumbar, R. (1998). Grooming, gossip and the evolution of language. Cambridge: Harvard University Press.
Gehring, W. J., Ikeo, K. (1999). Pax 6: mastering eye morphogenesis and eye evolution. Trends in Genetics, 15, 371-377.
Gilbert, S., Epel, D. (2009). Ecological developmental biology: integrating epigenetics, medicine, and evolution. Sunderland: Sinauer Associates.
Gould, S. J., Eldredge, N. (1977). Punctuated equilibria: the tempo and mode of evolution reconsidered. Paleobiology, 3 (2), 115-151.
Gould, S. J. (1997). Darwinian fundamentalism. The New York Review of Books, june 12, 34-37.
Hagoort, P. (2005). On Broca, brain, and binding: a new framework. Trends in Cognitive Sciences vol.9,n.9.
Horner, V., Whiten, A., Flynn, E., de Waal, F. B. M. (2006). Faithful replication of foraging techniques along cultural transmission chains by chimpanzees and children. Proceedings of the National Academy of Sciences, vol.103, n.37.
Kinzler, K. D., Dupoux, E., Spelke, E. S. (2012). "Native" objects and collaborators: infants' object choices and acts of giving reflect favour for native over foreign speakers. Journal of Cognition and Development, 3:1, 67-81.
Minelli, A. (2009). Forms of becoming: the evolutionary biology of development. Princeton: Princeton University Press.
Pennisi, A., Falzone, A. (2010) Il prezzo del linguaggio. Bologna: Il Mulino.
Pennisi, A. (in press) Per una tecnologia dello speech making: scienze cognitive e specie-specificità del linguaggio umano. Atti del I Convegno SLI (nuova serie).
Pievani, D. (in press). Individuals and groups in evolution: Darwinian pluralism and the multilevel selection debate. Journal of biosciences.
Richerson, P. J., Boyd, R. (2008). Not by genes alone. Chicago: University of Chicago Press.
Roughgarden, J. (2004). Evolution's rainbow. Berkeley-Los Angeles-London: University of California Press.
Shubin, N. H., Tabin C., Carroll S. (1997). Fossils, genes, and the evolution of animal limbs. Nature, 388, 639-648.

Spelke, E. S. (2009). Forum: Elizabeth S. Spelke. In M. Tomasello, Why we cooperate. Cambridge, MA: The MIT Press.
Strum, S. C. (1987). Almost human. New York: Norton \& Company.
Strum, S. C. (2012). Darwin's monkey: why baboons can't become human. Yearbook of Physical Anthropology, 55, 3-23.
Tomasello, M. (1999). The cultural origins of human cognition. Cambridge: Harvard University Press.
Washburn, S. L., DeVore, I. (1961). Social behavior of baboons and early man. In Washburn S.L., editor. The social life of early man. New York: Wenner-Gren Foundation for Anthropological Research.
Weiss, K. M., Buchanan, A. V. (2009). The mermaid's tale: four billion years of cooperation in the making of living things. Cambridge, MA: Harvard University Press.

# Assessing the effectiveness of older adults' spatial descriptions in a fetch task 

Laura A. Carlson ${ }^{1}$ (lcarlson@nd.edu) Marjorie Skubic ${ }^{2}$ (skubicm@missouri.edu) Jared Miller ${ }^{1}$ (jmille39@nd.edu) Zhiyu Huo ${ }^{2}$ (zhiyuhuo@mail.missouri.edu) Tatiana Alexenko ${ }^{3}$ (ta7cf@mail.missouri.edu)<br>${ }^{1}$ Department of Psychology, 119-D Haggar Hall, University of Notre Dame, Notre Dame, IN 46556 USA<br>${ }^{2}$ Electrical and Computer Engineering Department, University of Missouri, Columbia, MO<br>${ }^{3}$ Computer Science Department, University of Missouri, Columbia, MO


#### Abstract

The current paper examines spatial descriptions provided by older adults in the context of a fetch task in a virtual house environment that mimics an eldercare setting. Sixty-four older adults provided directions for how to find a target or where to find a target to a robot or human (named Brian) avatar. There were systematic differences in the form and structure of the descriptions based on the communicative task. Specifically, how descriptions were longer, contained more detail, and were dynamically structured as compared to where descriptions. However, where descriptions were found to be more effective in conveying the target location, as assessed with a subsequent target selection task. Implications for the development of robot algorithms for the comprehension of naturalistic spatial language across these two communicative tasks are discussed.


Keywords: Human-robot interaction (HRI); spatial language; dynamic and static; how and where; effectiveness; fetch task; assistive robotics; eldercare.

## Introduction

An emerging line of research in human-robot interaction involves the development of assistive devices for use in eldercare settings, either as social companions (e.g., Heerink et al., 2008; Kidd, Taggart, \& Turkle, 2006; Libin \& CohenMansfield, 2004; Shibata, Kawaguchi, \& Wada, 2011; Wada et al., 2003) or as task-oriented robots assisting with navigation (Montemerlo et al., 2002), managing medication (Tiwari et al., 2011)], and providing reminders (Pollack et al., 2002). Older adults also report wanting help with tasks such as cleaning, heavy lifting, and fetching objects (Beers et al., 2012). They also prefer to speak naturally to these assistive devices, rather than use a more constrained interface (Scopelliti, Guiliani, \& Fornara, 2005).

To accommodate these preferences, recently we gathered a corpus of spatial descriptions from older adults who interacted with an avatar within a virtual house setting in the context of a fetch task. Our primary goal in this project is to develop robot algorithms for the online comprehension of these natural language spatial descriptions and to test these algorithms in an analogous physical environment with a physical robot. In working toward this goal, on the basis of the corpus, we have identified key components that need to be developed for the robot including speech recognition for
older adults (Alexenko et al., 2013), parsing the natural language descriptions and coding them into chunks that can be converted into robot commands, recognizing key furniture items within a cluttered environment that are included in the descriptions, and identifying spatial relations within the horizontal plane (e.g., behind the couch) and the vertical plane (e.g., on top of the table) (Skubic et al, 2012).

Given that the robot algorithms are driven by the properties of the spatial descriptions, in the current paper we examine how the communicative task of the speaker impacts the features of the descriptions, and present data that reflect the effectiveness of the descriptions.

## Spatial Directions and Spatial Descriptions

A fetch task is one in which a speaker specifies the location of a desired target for an addressee whose goal is to retrieve the target. There are two ways in which the location can be indicated by the speaker. The speaker could provide directions that tell how to get to the target location or the speaker could provide descriptions that specify information about where a given target location is. Research has shown systematic differences in the type and structure of the language that is used for each of these communicative tasks. For example, Plumert et al. (1995) found that written directions on how to find a target in a hierarchically organized doll-house environment were more likely to provide more detailed messages and contain more spatial units that tended to be organized in a descending sequence (floor $\rightarrow$ room $\rightarrow$ reference object. e.g., The keys are on the first floor in the living room on the table.) as compared to written descriptions of where to find a target that were less detailed and organized in an ascending sequence (reference object $\rightarrow$ room $\rightarrow$ floor, e.g., The keys are on the table in the living room on the first floor).

This distinction between how and where has also been characterized as dynamic and static (Wahlster. 1995, Fasola and Mataric 2012) spatial language, respectively, with dynamic stepping the addressee through the environment in a point by point fashion and static offering spatial information that does not embed the addressee in the environment. Dynamic spatial directions are also inherently sequential, while static descriptions are not. Nevertheless, static descriptions are often overlooked or treated the same
as dynamic directions by other researchers (Tellex et al, 2011), perhaps due to the focus on two-dimensional route instructions or the assumption that dynamic descriptions are better or more prevalent (Kollar et al., 2010, MacMahon et al., 2006, Vogel and Jurafsky 2010, Shimizu and Haas 2009).

In the current work, we assess two questions related to this how/where distinction: First, we ask whether there are consistent differences in the type and form of the spoken spatial language that is produced by older adults in response to how and where instructions that might echo Plumert et al's (1995) findings with written spatial language. We focus on the type of language included in the descriptions, and the amount of detail, and ignore the hierarchical sequencing that Plumert et al. (1995) measured, because our environment consists of a single floor, as intended for mimicking an eldercare setting. Second, we ask whether these differences are associated with differences in the relative effectiveness of the descriptions.

## Corpus of descriptions from older adults in a virtual fetch task

Our corpus consists of 512 spatial descriptions collected from 64 older adults (mean age $=76$ years) who specified the location of 8 targets embedded in the virtual house environment shown in Figure 1. Targets were placed in the living room (on the left in Figure 1) and bedroom (on the right in Figure 1) on tables that also contained two other objects that could potentially serve as reference objects. On each trial, older adults explored the virtual house with the assistance of an experimenter and found a designated target. They were then positioned in the central hallway (marked by "Start" in Figure 1), and provided a description


Figure 1: Overview and screen shots of the virtual house
of the target location to either a robot or human avatar (named Brian) who faced them (as indicated by the arrow in Figure 1), such that their perspectives were misaligned by 180 degrees. Previous research has shown a preference for speakers to use an addressee perspective when perspectives between speaker and addressee are misaligned (Mainwaring et al., 2003; Schober, 1993), with such preference also observed for robot addressees (Tenbrink et al., 2002). However, given that older adults have shown negative emotional responses to robots (Scopelliti et al, 2005), we included the addressee manipulation to assess whether older adults in particular would be more likely to adopt their own perspective rather than the perspective of the robot. These perspective results are presented in Carlson et al. (2013). A second manipulation was related to the task instructions. Specifically, older adults were instructed to either provide directions for how to find the target or to provide descriptions of where the target was located. Both the addressee manipulation (robot or Brian) and the task instruction manipulation (how or where) were between subject manipulations, with the consequence that 16 older adults each provided 8 descriptions (128) for each of the 4 addressee X instruction combination (128 X $4=512$ descriptions in total).

A full report of the older adult corpus can be found in Carlson et al. (2013). We focus here on the how versus where differences. Figure 2 provides the task instructions (adapted from Plumert et al., 1995), and examples from the corpus.

|  | How | Where |
| :--- | :--- | :--- |
| $\begin{array}{c}\text { Instructions to } \\ \text { participants }\end{array}$ | $\begin{array}{l}\text { Please tell the Brian/the robot how to } \\ \text { find the target. }\end{array}$ | $\begin{array}{c}\text { Please tell Brian/the robot } \\ \text { where the target is. }\end{array}$ |
| $\begin{array}{l}\text { Addressee: Brian } \\ \text { Target 1 = Book }\end{array}$ | $\begin{array}{l}\text { Brian turn to your right. Go to the room } \\ \text { on your right and go straight ahead and } \\ \text { the book is right there before you. } \\ \text { Trian please take 1 or 2 steps forward } \\ \text { Target 2 = Cell phoner the room on your left. Upon } \\ \text { entering the room take about 3 steps } \\ \text { forward and you'll see a small brown } \\ \text { table with a black top. The cell phone } \\ \text { that you are looking for is on that table. }\end{array}$ | $\begin{array}{l}\text { Brian the book is in the room to } \\ \text { your right and its on a table at } \\ \text { the far side of the room. }\end{array}$ |
| The cell phone is in the bedroom |  |  |
| on the bedside table. |  |  |$\}$

Figure 2: Instructions and sample descriptions by how/where and addressee

As shown in Table 1, how descriptions contained more words overall per description, and included more spatial terms (such as "on", "to" and "right") and more hedges (such as "immediately" and "slightly"). In contrast, where descriptions contained more house units (such "room", "door" and "wall"). Descriptions often contained large furniture items in the rooms (such as "bed" and "couch"), and rarely contained reference objects that were collocated
on the tables (such as "lamp"), with the incidence of these categories not varying across how and where descriptions. Finally, how descriptions were more likely to have a dynamic form than the where descriptions.

|  | How | Where |
| :--- | :--- | :--- |
| Words per description | $\mathrm{M}=27.0, \mathrm{SE}=1.8$ | $\mathrm{M}=19.4, \mathrm{SE}=1.5$ |
| Spatial terms per description | $\mathrm{M}=4.1, \mathrm{SE}=.26$ | $\mathrm{M}=2.3, \mathrm{SE}=.29$ |
| House units per description | $\mathrm{M}=1.1, \mathrm{SE}=.10$ | $\mathrm{M}=1.4, \mathrm{SE}=.12$ |
| Hedges per description | $\mathrm{M}=.22, \mathrm{SE}=.05$ | $\mathrm{M}=.09, \mathrm{SE}=.02$ |
| \% dynamic descriptions | $\mathrm{M}=95.3, \mathrm{SE}=3.3$ | $\mathrm{M}=35.8, \mathrm{SE}=7.9$ |

Table 1: Significant differences between how and where older adult descriptions

These results are consistent with Plumert et al. (1995) who also found that how descriptions contained more spatial units and were more detailed than where descriptions. What remains unclear is whether these differences are associated with any differences in effectiveness. That is, these descriptions were all collected by older adult speakers in the context of a fetch task in which accurately specifying the location of the target is critical for the success of the task. We ask next whether how or where descriptions are more effective in identifying the location of the target.

## Differences in how vs. where effectiveness

To assess effectiveness, we randomly selected from the corpus of older adult descriptions two from each speaker's set of 8 descriptions (with half of the speakers addressing Brian and half the robot), with the constraint that the location of each target was specified an equal number of times across the set of descriptions that we were assessing. These descriptions were then provided to sixtyfour younger adults to assess effectiveness. Their task was to listen to a description without the target, navigate through the house in accordance with the description, and then guess the identity of the target. 4 targets were placed on tables in the living room and 4 targets were placed on tables in the bedroom. Each table contained a target and two distractor objects. Each participant performed two trials (one in the living room and one in the bedroom). Before the trials began, the younger adults were shown a video tour of the house that did not include the targets. This was to familiarize them with the house environment and the relative locations of the rooms and their contents. On each trial, participants started in the hallway of the house, standing in the location and at the orientation of the avatar, as specified by the label "robot or avatar" shown in Figure 1 , and facing the original speaker's location (which is marked in Figure 1 with the label "Start" and with an orientation specified by the arrow). They were therefore facing the position of the participants from which the descriptions were gathered. A participant was given a description from the corpus with the target item removed, and they navigated through the house until they thought
they found the target, and then named it. The key dependent measure was their accuracy in selecting the target.

As shown in Figure 3, we examined two indicators of this accuracy: selection of the correct target, and selection of the correct table on which the target appeared. This latter measure is important because two potential reference objects appeared on the tables next to the targets, and often the descriptions did not provide enough information to identify which object on the table was the target (see example descriptions in Figure 2). The infrequent use of the reference objects that appeared next to the target is consistent with Plumert et al. (1995) who found that such reference objects were only consistently used when the target was located on the reference object as opposed to beside it.


Figure 3: Selection accuracy for correct table (top) and correct target (bottom) as a function of how/where and addressee. Dotted line indicates chance selection.

With respect to selection of the correct table, performance in all conditions was significantly above chance performance of 12.5 , based on 8 possible tables in the environment. In addition, significantly better performance was observed for "where" descriptions $(M=61 \%)$ than for "how" descriptions $(\mathrm{M}=42 \%), \mathrm{F}(1,60)=4.51, \mathrm{p}<.05$. In addition, there was a significant effect of addressee, with more accurate performance for descriptions provided to Brian $(M=67 \%)$ than to the robot $(M=36 \%), F(1,60)=$
$12.55, \mathrm{p}<.01$. The interaction between instruction and addressee was marginal, $\mathrm{F}(1,60)=3.13, \mathrm{p}=.08$.

With respect to the selection of the correct object, performance in all conditions was significantly above chance performance of 4.2 (based on 24 possible targets in the environment ( 3 on each of 8 tables)). For this analysis, there was only a main effect of addressee: $F(1,60)=7.10$, $p$ $<.05$; the effect of instruction and the interaction were not significant ( $\mathrm{Fs}<1.6$, $\mathrm{ps}<.21$ ).

For the object selection measure, we also assessed how likely it was that participants selected the correct object, given that they selected the correct table. Chance performance in this case is $33 \%$, given that there are three objects (target and two reference objects on each table).

Figure 4 shows that in all conditions, accuracy was significantly above chance. We expect that this is because the target objects were generally smaller than the reference objects on the tables. Clark, Schreuder and Buttrick (1983) argue that when a reference is under-determined by a speaker, the addressee will select an object from a group of


Figure 4: Selection accuracy for the correct object, conditional on correct table selection. Dotted line indicates chance selection.
objects that offers the most contrast from the others along a given dimension. For example, imagine a speaker tells an addressee to pick up a ball and refers to a collection of three balls (a golf ball, a squash ball and a basketball) that are placed on a table in front of them. Clark et al. argue that it is likely that the addressee will select the basketball because it is the most unique item in the set, standing out in terms of size.

Finally, we also examined whether there were differences in accuracy for the individual targets. Given that each of the targets appeared in a given location (and location was not counterbalanced across targets), this serves an indicator as to whether any of the target locations were particularly difficult to describe and find. Figure 5 shows accuracy as a function of the targets, both as indicated by the correct selection of the table and correct selection of the object.


Figure 5: Selection accuracy for table and target selection as a function of target. Dotted lines indicate chance selection.

Chi-square analyses revealed no significant differences among targets, for either the correct table accuracy or the correct target accuracy. This indicates that there were not any targets or locations that were particularly difficult to describe and/or find. This is likely due to the simple layout and relatively impoverished contents of the rooms in the virtual house.

Overall, the results for the assessment of the older adults spatial descriptions indicate that the where descriptions allowed participants to more easily select the target and its table than the how descriptions. We suspect that the differences in accuracy for table selection as a function of addressee that were observed are likely due to other properties of the descriptions, such as the perspective adopted by the speaker. For a full report of the older adult corpus, see Carlson et al. (2013).

## Conclusions

Together, the detailed analysis of the corpus and the results of the experiment assessing the effectiveness of the descriptions point to an interesting contrast. On the one hand, the corpus analysis reveals that how descriptions are longer, offer more detail, include more spatial terms, and are dynamic, as compared to the where descriptions that are shorter, include more references to house structures, and are often static. On the other hand, these same how descriptions are not as effective in communicating the location of the target, as assessed by the accuracy for selecting the target and its table. We are currently comparing the effectiveness of these older descriptions with the effectiveness of a corpus of descriptions collected from younger adults within the same virtual environment. Moreover, we are also recording the paths that participants take to the target in response to these descriptions, with the idea that the paths may offer an additional online measure of effectiveness. Metrics we are
examining include path length, navigation speed, number of pauses, and changes in heading.

These results also have several interesting implications for the development of robot algorithms in this task. For example, it may be beneficial for the robot algorithm to initially classify a description as one that is conveying directions or one that is conveying location, given that the form and content of the descriptions vary as a function of communicative task. In a natural setting, of course, the speaker may not be explicit about whether he or she is providing directions or specifying location (that is, the speaker is not assigned a how or where task per se, as in our current work). This classification would need to be based on the properties of the descriptions themselves.

In addition, the robot algorithms will need to take into account the differential effectiveness of the two types of descriptions. The "how" descriptions may provide a more explicit approach to allow direct translation into robot commands; however, varying viewing perspectives will complicate the interpretation. To follow the directions of the "how" descriptions, a robot does not rely as much on perception, which may improve the efficiency of the fetch navigation in some static environments but not necessarily the effectiveness. In contrast, the "where" descriptions provide more hints using reference structures and objects so that the robot can navigate to the target using perception. The "how" descriptions may be easier to interpret but have a lower probability to navigate the robot to the specified target, especially given a dynamic environment in which reference furniture items have been moved. The "where" descriptions require the challenge of translating them into navigation commands but may provide more reliable fetch results, even in the case of moved reference items.

## Acknowledgments

This research was supported with funding by the National Science Foundation Grant IIS-1017097. We thank Xiao Ou Li for constructing the virtual house environment and Erin Gibson, Oscar Gonzalez, Diane Garritson, Ashley Herrmann, Mary Johnson, Kevin Kimberly, Kathleen Loftus, Angelique Laboy-Capparopa, Elliott Mitchem, and Joe Wernke for help collecting and coding the data.

## References

Beer, J., Smarr, C.A., Chen, T.L., Prakash, A., Mitzner, T.L., Kemp, C.C., and Rogers, W.A. (2012). "The Domesticated Robot: Design Guidelines for Assisting Older Adults to Age in Place," Proc., IEEE Conference on Human Robot Interaction - Session: Living and Working with Service Robots, Boston, MA, pp. 335-342.
Carlson, L. A., Skubic, M., Miller, J., Huo, Z., \& Alexenko, T. (2013). Strategies for human-driven robot comprehension of spatial descriptions by older adults in a robot fetch task. Revised manuscript submitted for publication in special issue of Topics.

Clark, H. H., Schreuder, R., \& Buttrick, S. (1983). Common ground and the understanding of demonstrative reference. Journal of Verbal Learning and Verbal Behavior, 22, 245-58.
Fasola, Juan, and Maja J. Matarić. "Using Spatial Language to Guide and Instruct Robots in Household Environments." (2012), AAAI Technical Report, FS-1207.

Heerink, M., Kröse, B., Wielinga, B., and Evers, V. (2008). "Enjoyment, Intention to Use and Actual Use of a Conversational Robot by Elderly People," Proc., Intl. Conf. on Human-Robot Interaction, pp. 113-119.
Kidd, C., Taggart, C.W., and Turkle, S. (2006). "A Sociable Robot to Encourage Social Interaction among the Elderly," Proc., IEEE Intl Conf on Robotics and Automation, pp. 3972-3976.
Kollar, T., Tellex, S., Roy, D., and Roy, N. (2010). "Toward Understanding Natural Language Directions," Proc., 5th ACM/IEEE International Conference on Human-Robot Interaction, pp. 259.
Libin, A. and Cohen-Mansfield, J. (2004). "Therapeutic Robocat for Nursing Home Residents with Dementia: Preliminary Inquiry," American Journal of Alzheimer's Disease and Other Dementias, 19(2):111-116.
MacMahon, M., Stankiewicz, B., and Kuipers, B. (2006). "Walk the Talk: Connecting Language, Knowledge, and Action," Route Instructions, pp 1475-1482.
Mainwaring, S., Tversky, B., Ohgishi, M. \& Schiano, D. (2003). Descriptions of simple spatial scenes in English and Japenese. Spatial Cognition and Computation, 3, 3-42
Montemerlo, M., Pineau, J., Roy, N., Thrun S., and Verma, V. (2002). "Experiences with a Mobile Robotic Guide for the Elderly," Proc., AAAI-02, pp. 587-592.
Plumert, J. M., Carswell, C., DeVet, K., and Ihrig, D. (1995). "The Content and Organization of Communication about Object Locations," Journal of Memory \& Language, 34:477-498.
Pollack, M.E., Brown, L., Colbry, D., Orosz, C., Peintner, B., Ramakrishnan, S., Engberg, S., Matthews, J.T., Dunbar-Jacob, J., and McCarthy, C.E. (2002) "Pearl: A Mobile Robotic Assistant for the Elderly," Proc., AAAI Workshop on Automation as Eldercare.
Schober, M. F. (1993). Spatial perspective-taking in conversation. Cognition, 47(1), 1-24.
Scopelliti, M., Giuliani, V., and Fornara, F. (2005). "Robots in a Domestic Setting: A Psychological Approach," Universal Access in the Information Society, 4, 146-155. Scopellite, Guiliani, \& Fornara, 2005
Shibata, T., Kawaguchi, Y., and Wada, K. (2011). "Investigation on People Living with Seal Robot at Home - Analysis of Owners' Gender Differences and Pet Ownership Experience," International Journal of Social Robotics, 4(1):53-63.
Shimizu, N. and Haas, A. (2009). "Learning to Follow Navigational Route Instructions," Proc., Intl. Joint Conf. on Artificial Intelligence, pp.1488-1493.

Skubic, M., Alexenko, T., Huo, Z., Carlson, L., and Miller, J. (2012). "Investigating Spatial Language for Robot Fetch Commands," Grounding Language for Physical Systems, AAAI Technical Report, WS-12-07, pp. 39-45.
Tellex, S., Kollar, T., Dickerson, S., Walter, M., Banerjee, A., Teller, S. and Roy, N. (2011). "Understanding Natural Language Commands for Robotic Navigation and Mobile Manipulation," Proc., Conf. on Artificial Intelligence (AAAI).
Tenbrink, T., Fischer, K., \& Moratz, R. (2002). Spatial strategies of human-robot communication." KI \#4 [Online]. Available: http://citeseerx.ist.psu.edu/ viewdoc /summary? doi=10.1.1.58.2151
Tiwari, P., Warren, J., Day, K.J., and Datta, C. (2011). "Comprehensive Support for Self-Management of Medications by a Networked Robot for the Elderly" Proc., Health Informatics New Zealand 10th Annual Conference and Exhibition, Aukland.
Vogel, A. and Jurafsky, D. (2010). "Learning to Follow Navigational Directions," Proc., 48th Annual Meeting of the Association for Computational Linguistics, pp. 806814.

Wada, K., Shibata, T., Saito, T. and Tanie, K. (2003). "Psychological and social effects of robot assisted activity to elderly people who stay at a health service facility for the aged," in Proc., Intl. Conf. on Robotics and Automation, pp. 3996-4001.
Wahlster, W., Statiques Et Dynamiques, and Gerd Herzog. "Coping with Static and Dynamic Spatial Relations." (1995).

# Hypothesis Space Checking in Intuitive Reasoning 

Christopher D. Carroll (cdcarroll@gmail.com)<br>Department of Psychology, Carnegie Mellon University<br>5000 Forbes Ave., Pittsburgh, PA 15213<br>Charles Kemp (ckemp@cmu.edu)<br>Department of Psychology, Carnegie Mellon University<br>5000 Forbes Ave., Pittsburgh, PA 15213


#### Abstract

The process of generating a new hypothesis often begins with the recognition that all of the hypotheses currently under consideration are wrong. While this sort of falsification is straightforward when the observations are incompatible with each of the hypotheses, an interesting situation arises when the observations are implausible under the hypotheses but not incompatible with them. We propose a formal account, inspired by statistical model checking, as an explanation for how people reason about these probabilistic falsifications. We contrast this account with approaches such as Bayesian inference that account for hypothesis comparison but do not explain how a reasoner might decide that the hypothesis space needs to be expanded.


Keywords: hypothesis testing; model checking; surprise

## Introduction

Many modern scientific disciplines are characterized by strange and unintuitive theories that previous generations of scientists never would have imagined. On a less dramatic scale, people often generate inventive explanations in their everyday lives. The existence of these unintuitive theories and inventive explanations raises an interesting question: how are these new theories and explanations discovered?

In many cases, the process of generating a new hypothesis starts when the reasoner decides that all of the hypotheses currently under consideration are wrong. In some cases, the available evidence is incompatible with every hypothesis under consideration, and this decision is straightforward. In other cases, however, the available evidence is implausible under, but not strictly incompatible with, the hypotheses. In cases like these, a reasoner may engage in hypothesis space checking to decide whether the hypothesis space is adequate or needs to be expanded.

Although psychologists have explored many approaches to hypothesis testing, most of these approaches are unable to account for hypothesis space checking. Bayesian accounts, for instance, are able to specify the relative strength of a hypothesis within the hypothesis space, but they do not provide criteria for evaluating the hypothesis space itself.

Statisticians, however, have developed various measures that quantify the extent to which observations are surprising under a given hypothesis or hypothesis space. In this paper, we investigate the possibility that formal measures of this kind can help to explain how people decide that all of the hypotheses in their current hypothesis space are probably wrong.

## Hypothesis space checking

Figure 1 illustrates the kind of situation where hypothesis space checking may be required. There is a universe $U$ of possible explanations for the given observations, but the hypotheses available to the reasoner fall within a hypothesis space $H$ that is a proper subset of $U$. It is possible, of course, that the true explanation is not in $H$; the ability to determine whether this is the case would be useful.

U


Figure 1: The universe $U$ includes all possible hypotheses, and hypothesis $H$ is the subset of these hypotheses that are currently available to the reasoner.

In principle, the adequacy of $H$ could be evaluated by computing whether the available observations are better explained by hypotheses that lie within or outside $H$. Bayesian inference provides one way to formalize this sort of comparative hypothesis testing. Bayes' theorem establishes that given the observed data $d$, the odds that $H$ contains the true explanation are:

$$
\begin{equation*}
\frac{P(H \mid d)}{P(\bar{H} \mid d)}=\frac{P(d \mid H) P(H)}{P(d \mid \bar{H}) P(\bar{H})} \tag{1}
\end{equation*}
$$

Equation 1 shows that the probability that $H$ contains the true explanation depends on (1) the relative probabilities of the data under $H$ and under its complement $\bar{H}$ and (2) the relative prior probabilities of $H$ and $\bar{H}$.

Although Equation 1 is appealing in principle, it is impossible to apply. Given that $\bar{H}$ consists of hypotheses that are unavailable to the reasoner, the term $P(d \mid \bar{H})$ will be impossible to compute (Earman, 1990, Ch. 7; Salmon, 1990). Consider the problem faced by a Newtonian physicist attempting to explain the anomalous precession of Mercury's perihelion. Although the physicist might be able to estimate $P(d \mid H)$ by considering various Newtonian explanations, estimating $P(d \mid \bar{H})$ has a paradoxical flavor:
how would the physicist compute probabilities with respect to theories he cannot currently imagine?

The paradox just described applies to any account (Bayesian or otherwise) that uses comparative hypothesis testing to address the problem defined by Figure 1. We therefore propose that this problem can only be addressed by non-comparative accounts of hypothesis testing in which the current hypothesis space is evaluated not in relation to specific competitors but on its own merits. In statistical practice, this sort of evaluation is often referred to as model checking or goodness-of-fit testing, and it typically involves comparing the actual observations to the expected distribution of the observations given the current hypothesis space. To the extent that the actual observations seem surprising in this context, there is an incentive to search for new hypotheses.

Comparative and non-comparative hypothesis testing seem to address distinct problems in that comparative hypothesis testing seems most useful for selecting among the hypotheses in $H$ and non-comparative hypothesis testing seems most useful for checking $H$ itself (for similar proposals, see Bayarri \& Berger, 1999; Gelman \& Shalizi, 2013; Gillies, 2007). We propose that both kinds of hypothesis testing are represented among people's intuitive inferences, but in this paper we deliberately focus on a situation that calls for non-comparative hypothesis testing.

## A model of non-comparative hypothesis testing

We propose that intuitive hypothesis space checking resembles the process specified in Figure 2. Specifically, we propose that people extract the salient or important features of the available observations, assess the extent to which those individual features are surprising under $H$, and then compute a global measure of surprise. This global measure of surprise provides a criterion for deciding whether to initiate the search for new hypotheses.


Figure 2: The reasoner extracts the salient features of the observations $d$, calculates a measure of surprise for each feature, and combines the surprise values into a global measure $S_{H}$ that captures the extent to which the data are surprising given the current hypothesis space $H$.

Statisticians have proposed various measures of surprise (e.g., Bayarri \& Berger, 1998; Weaver, 1948), but we focus on statistical null hypothesis testing, which is the bestknown statistical procedure that can be used for hypothesis space checking. To investigate null hypothesis testing in the simplest possible setting, we focus on situations where the
hypothesis space contains a single focal hypothesis $h$ (i.e., where $H=\{h\}$ ), but various generalizations of our approach are applicable to composite hypothesis spaces (e.g., Bayarri \& Berger, 1999; Gelman, Meng, \& Stern, 1996). In null hypothesis testing, the statistician first defines a real-valued test statistic $T(d)$ that measures some property of the data; this test statistic can be viewed as one of the features in Figure 2. To evaluate the surprise of the observed value of the test statistic, the statistician then considers the expected distribution of $P\left(d^{\text {rep }}\right)$ given $h$, where $d^{\text {rep }}$ is a random variable representing the data that might be observed if one were to replicate the "experiment" that produced the data. By comparing the observed value of $T(d)$ to the expected distribution of $T\left(d^{\text {rep }}\right)$ under $h$, the statistician can assess whether $T(d)$ is surprising under $h$. If the test statistic is defined such that greater values represent greater deviations from $h$, the surprise of $T(d)$ can be summarized by a $p$-value:

$$
\begin{equation*}
p_{T}(d)=P\left[T\left(d^{\text {rep }}\right) \geq T(d) \mid h\right] . \tag{2}
\end{equation*}
$$

Intuitively, the p-value represents the probability that the test statistic in the imagined replications would be at least as extreme as what was actually observed. Small p-values correspond to surprising results where the observations are unusually extreme.

In the final step of Figure 2, the reasoner combines the surprise measures for each feature into a global measure of surprise. To avoid making assumptions about how people integrate surprise ratings across different features, we focus on situations where there is a single surprising feature. In such situations, it seems reasonable to adopt the surprise value for that feature as the global measure of surprise.

## Method

To evaluate our proposed model of non-comparative hypothesis testing, we conducted an experiment in which participants learned about the ancient burial sites found on a remote island chain. The burial sites were marked by "cairns" (rock piles), and each island had been occupied by one of two cultures that constructed the cairns using different procedures: the "Chaotics" placed a random number of boulders in each cairn and the "Numerologists" placed a number of boulders in accordance with a mathematical function. The instructions explained that the Numerologists used different mathematical functions on different islands but that the mathematical function was always based on the number of people buried at the site. The participants were asked to infer which cultural group occupied an island from information about the burial sites on the island.

Because the number of possible mathematical functions is infinite, we expected that participants would not be able to assess every possible explanation for the observations. We expected that when faced with this impossible task, participants would consider the hypothesis that the Chaotics occupied the island as well as various hypotheses where the

Numerologists occupied the island and used some simple mathematical function. Because the materials were designed so that no simple mathematical function would explain the observed number of boulders at the burial sites, we expected that most participants would end up with a hypothesis space that contained a single viable hypothesis: the hypothesis that the Chaotics occupied the island. We expected that participants would check this hypothesis through a procedure resembling the one depicted in Figure 2. We predicted that when the observations were sufficiently surprising, participants would be willing to attribute occupancy to the Numerologists. Critically, we expected that participants would sometimes make this attribution even when they could not identify a single mathematical function that the Numerologists might have used. As we discuss later, this finding would be difficult to explain as a consequence of comparative hypothesis testing.

The experimental materials were based on three "test statistics" that reflected the salient numerical concepts of equality and magnitude (see Table 1). The equality test statistic, for example, was defined as the count of the burial sites that had the same number of people and boulders, and we expected participants to be surprised when many of the burial sites had the same number of people and boulders.

Table 1: Test statistics

| Name | Definition |
| :--- | :--- |
| Equality | number of burial sites where $b=p$ |
| Minimum | smallest observed value of $b$ |
| Repetition | frequency count for the most frequent $b$ |

Note. $p=$ the number of people at a burial site; $b=$ the number of boulders at a burial site.

## Participants

Sixty-one undergraduates participated in the experiment for course credit.

## Materials

Table 2 displays the observations presented to the participants. In the table and in the rest of the paper, we represent burial sites by two numbers separated by a dash, with the first and second numbers representing the number of people buried and boulders, respectively. Each row of the table corresponds to a different island. Twelve of the islands were designed to be surprising according to exactly one of the test statistics in Table 1 and four additional islands were designed to have no surprising features (the "None" islands). All of the islands contained either three or six burial sites, and the surprising islands were designed so that the coincidence involving the test statistic would be either moderately (. $01<p<.15$ ) or highly ( $p<.01$ ) surprising, as calculated from Equation 2.

The rightmost column shows the p -values for each island; these p -values summarize how surprising the observations would be if the Chaotics occupied the island. The p-values were calculated under the assumption that the number of boulders at a burial site could range from 1 to 100 (the
instructions informed participants that this was the case). To illustrate, consider the calculation of the p-value for the first island. The observed value of the equality test statistic for this island was one: there was exactly one burial site that had the same number of people and boulders. If the Chaotics occupied the island, then the equality test statistic would follow a binomial distribution with a probability parameter of .01 . Consequently, the probability that at least one of the burial sites on a three-site island has the same number of people and boulders is approximately .0297.

For the equality and minimum statistics, the four islands represented the four possible combinations of surprise condition and island size. For the repetition test statistic, we did not include a high-surprise island with three burial sites; instead, we included two moderate surprise islands with three burial sites. The reason for this was that creating a high-surprise "repetition" island with three burial sites necessitated selecting an island where each burial site had the same number of boulders. Because we were interested in situations where the participants would not be able to find a mathematical function to explain the observations, we chose not to present such an island.

Table 2: Experimental materials

| Feature | Srprs. | Sz. | Burial sites | $p$-value |
| :---: | :---: | :---: | :---: | :---: |
| Equality | M | 3 | 20-94, 39-39, 85-78 | . 0297 |
| Equality | H | 3 | 16-16, 65-65, 49-12 | . 0003 |
| Equality | M | 6 | $\begin{aligned} & 7-62,33-85,40-1, \\ & 53-26,59-59,94-18 \end{aligned}$ | . 0585 |
| Equality | H | 6 | $\begin{aligned} & 12-100,19-42,21-21, \\ & 32-14,75-75,93-56 \end{aligned}$ | . 0015 |
| Minimum | M | 3 | 15-86, 63-98, 84-75 | . 0176 |
| Minimum | H | 3 | 16-92, 42-97, 93-90 | . 0013 |
| Minimum | M | 6 | $\begin{aligned} & 5-67,24-81,35-72, \\ & 52-68,57-93,83-54 \end{aligned}$ | . 0108 |
| Minimum | H | 6 | $\begin{aligned} & 13-75,32-95,35-98 \\ & 37-80,72-85,96-94 \end{aligned}$ | . 0003 |
| Repetition | M | 3 | 3-19, 27-84, 74-19 | . 0299 |
| Repetition | M | 3 | 11-75, 39-28, 80-75 | . 0299 |
| Repetition | M | 6 | $\begin{aligned} & 2-5,6-97,31-69 \\ & 59-38,62-52,75-52 \end{aligned}$ | . 1404 |
| Repetition | H | 6 | $\begin{aligned} & 12-98,15-98,26-4 \\ & 45-73,60-53,77-98 \end{aligned}$ | . 0020 |
| None | - | 3 | 23-18, 40-69, 93-55 | - |
| None | - | 3 | 31-46, 80-24, 94-87 | - |
| None | - | 6 | $\begin{aligned} & 1-78,43-61,45-12 \\ & 52-35,83-87,91-46 \end{aligned}$ | - |
| None | - | 6 | $\begin{aligned} & 1-26,8-92,14-36 \\ & 35-20,40-11,63-45 \end{aligned}$ | - |

Note. Srprs. = surprise condition; Sz. = island size; M = moderate surprise; $\mathrm{H}=$ high surprise .

All of the observations were designed so that at most one of the test statistics in Table 1 would be surprising at a level greater than $p=.30$. In addition, we controlled for the distribution of even and odd numbers and for the correlation between the number of people and number of boulders.

## Procedure

Participants were provided with a cover story that described their task and the Chaotics and Numerologists. Participants then completed a familiarization trial. On both the familiarization and experimental trials, the burial sites were represented by "cards" on a computerized display. Each card listed one number next to a stick figure and another number next to an illustration of a boulder pile. These numbers represented the number of people buried at the site and the number of boulders in the cairn, respectively. Participants were encouraged to re-arrange the cards by clicking and dragging them. The interface also provided buttons to automatically sort the cards according to either the number of people buried or the number of boulders. For the practice trial, the three burial sites were 31-1, 48-5, and $90-4$, and participants were told that the island was occupied by Numerologists who placed a number of boulders equal to the number of prime factors of the number of people buried at the site (e.g., because $48=3 * 2 * 2 * 2 * 2$, the burial site with 48 people had 5 boulders). This rule was intended to establish that the Numerologists were sophisticated mathematicians who had access to a wide variety of mathematical properties and rules. In doing so, our goal was to establish a universe of possible explanations that would be too large to consider in full.

Participants reported their inferences about which cultural group had occupied the island using a seven-point rating scale where the leftmost point was labeled "definitely Chaotics", the rightmost point was labeled "definitely Numerologists", and the middle point was labeled "not sure". Responses were coded from -3 ("definitely Chaotics") to 3 ("definitely Numerologists"). Participants who indicated that the Numerologists were more likely to have occupied the island than the Chaotics were also asked to indicate whether they had "discovered ANY function that the Numerologists might have used to determine the number of boulders." Participants answering affirmatively were asked to describe the function. Finally, at the end of each trial, participants were asked to list "any features, coincidences, or patterns in the burial sites that would have been surprising if the Chaotics occupied the island." Participants were provided with three text input fields and could identify up to three features, coincidences, or patterns. The responses to this prompt were intended to measure whether participants noticed the relevant features or any other features of the observations.

After completing the familiarization trial, the participants completed experimental trials for each of the 16 islands listed in Table 2. The presentation order was randomized.

## Results

A preliminary analysis confirmed that participants frequently noticed the relevant features. For each feature, a majority of the participants listed the feature as surprising on at least one of the relevant trials; the proportions were .59 for the minimum feature, .72 for the equality feature and .66 for the repetition feature. A second preliminary analysis
confirmed that the islands did not contain many surprising features other than the intended relevant features. Participants listed other features on only $16.8 \%$ of the experimental trials. The proportion of participants listing other features was similar across the experimental conditions: a logistic regression with categorical predictors corresponding to the surprise conditions, the relevant features, and the island sizes did not explain a significant proportion of the variance in the probability that participants noticed other features, $R^{2}=.26, F(5,8)=0.55, p=.74$.

To evaluate our formal approach we compared the modelderived p-values and the mean culture ratings. Because people often evaluate probabilities on a logarithmic scale (e.g., Gonzalez \& Wu, 1999), we adopted the logit (i.e., the log-odds) of the island p-values as the model's measure of surprise (lesser values corresponded to greater surprise). When calculating the mean culture ratings for each condition, we excluded any culture ratings for which the participant who provided the rating failed to identify the relevant feature as surprising at any point during the experiment. The rationale for this exclusion is that a participant who did not notice the relevant feature could not have been surprised by it.


Figure 3: The culture rating as a function of the logit of the p-value. Different features are represented by different marker shapes, different island sizes are represented by different marker sizes (large markers correspond to islands with six burial sites), and different surprise conditions are represented by different shadings (black markers correspond to high-surprise islands).

Figure 3 shows the comparison of the p -values and the mean culture ratings. A linear regression confirmed that the logit of the p-values explained a significant proportion of the variance in the ratings, $R^{2}=.82, F(1,10)=44.2, p<$ .001. This relationship was essentially unchanged even when culture ratings were included for participants who failed to notice the relevant features, $R^{2}=.77, F(1,10)=$ $33.83, p<.001$. Inspection of Figure 3 also suggests that the islands with six burial sites may have been viewed as less surprising than the islands with three burial sites. The statistical significance of this finding was confirmed by a multistep regression that showed that island size predicts variance in the culture ratings above and beyond the variance explained by the logit of the p -values, $\Delta R^{2}=.14$,
$F(1,9)=25.7, p=.001$. This finding may reflect a general tendency to underestimate the extent to which deviations from the mean become increasingly surprising for larger samples (Kahneman \& Tversky, 1972).

Although our participants compared hypotheses in the sense that they reported whether the Chaotics or Numerologists occupied an island, it seems difficult to explain their inferences as the product of what we have called comparative hypothesis testing. Consider, for example, the difficulties that arise in explaining the culture ratings by appealing to Equation 1, which in the context of our experiment involves the comparison of $P(d \mid$ Chaotics $)$ and $\quad P(d \mid$ Numerologists $)$. Note that $\quad P(d \mid$ Chaotics $)$ depends only on the number of burial sites on the island: for any island with three burial sites, for example, $P(d \mid$ Chaotics $)$ is $(1 / 100)^{3}$. Thus, if the participants' inferences were indeed based on Equation 1, then the differences in the culture ratings must have arisen primarily because of differences in $P(d \mid$ Numerologists $)$.

If $P(f \mid$ Numerologists $)$ is a prior distribution over specific functions $f$, then

$$
\begin{equation*}
P(d \mid \text { Numerologists })=\int_{f}(d \mid f) P(f \mid \text { Numerologists }) \tag{3}
\end{equation*}
$$

The integral in Equation 3 will be large to the extent that there are many functions that are plausible a priori ( $P(f \mid$ Numerologists $)$ is high) and consistent with the data $(P(d \mid f)>0)$. Approximating this integral using sampling or some other standard method would involve identifying one or more functions $f$ for which $P(d \mid f)>0$. Our participants, however, rarely identified even a single function $f$ for which $P(d \mid f)>0$. Recall that participants who claimed that the Numerologists occupied an island were asked whether they had found any mathematical function to explain the observations. Participants reported finding a function on only $15.5 \%$ of these occasions. Furthermore, the "functions" that these participants reported were often not fully specified functions at all. One representative participant claimed to have found a function but then wrote that "I don't have a function, but when put roughly on a graph it almost-kinda-sorta forms a wave." Summarizing his inference, the same participant later added, "I'm grasping at straws though." Figure 4, furthermore, shows that the relationship between function finding and the culture ratings is weak and, according to a linear regression, non-significant, $R^{2}=.016, F(1,12)=.19, p=.67$.

Could participants have estimated $P(d \mid$ Numerologists $)$ without identifying a single specific function $f$ that might have been used by the Numerologists? Might participants, for example, have used some computational procedure that approximates the integral in Equation 3 without actually identifying any specific functions? We cannot exclude this possibility, but we do not know of any such procedure. In the absence of a known procedure that approximates the integral in Equation 3 given some plausible specification of the prior, it seems reasonable to conclude that our participants did not rely on comparative hypothesis testing.


Figure 4: The mean culture ratings as a function of the proportion of participants who claimed to have identified a mathematical function that explained the observations.

As a final test of our model, we investigated whether the model-derived surprise predicted the culture ratings even after excluding trials in which participants claimed to have found a mathematical function. To do so, we recalculated the mean culture ratings while excluding any culture rating where either (1) the participant reported finding a mathematical function or (2) the participant never noticed the relevant feature (as in previous analyses). A linear regression on these recalculated culture ratings confirmed that the logit of the p-values remained strongly predictive of the culture ratings, $R^{2}=.87, F(1,10)=69.43, p<.001$.


Figure 5: Culture ratings as a function of surprise condition and the number of burial sites. The error bars show standard errors.

Our analyses so far have focused on the islands that were designed to be of "moderate" or "high" surprise. We compared these islands to the unsurprising "None" islands by analyzing the mean culture ratings as a function of surprise condition. Figure 5 shows that participants were much more willing to attribute island occupancy to the Numerologists in the high-surprise condition. The similarity between the culture ratings for the unsurprising and moderately-surprising condition was not expected, but it may be that the culture ratings are only influenced by observations once the surprise exceeds a certain threshold. Figure 5 also suggests that island size might have influenced
the culture ratings, either on its own or in an interaction with the surprise condition. A within-subjects ANOVA showed that the culture ratings were influenced by both surprise condition, $F(2,100)=27.80, p<.001$, and island size, $F(1$, $50)=6.89, p=.01$; the interaction between surprise condition and island size was marginally significant, $F(2$, 100) $=2.89, p=.06$. In post-hoc analyses, we confirmed that the culture ratings in the unsurprising and moderatelysurprising conditions were not significantly different, $t(50)$ $=.80, p=.43$, and that the culture ratings in the highsurprise condition were significantly different from those in the unsurprising, $t(50)=5.66, p<.001$, and moderatelysurprising, $t(50)=6.74, p<.001$, conditions.

## Discussion

The experimental findings suggest that people perform hypothesis space checking using an intuitive version of noncomparative hypothesis testing. The findings are not naturally explained by comparative hypothesis testing. This is not to say that comparative hypothesis testing is never useful: recall that our experiment was deliberately designed so that comparative hypothesis testing would be of limited relevance, and comparative hypothesis testing undoubtedly plays an important role in other settings. Moreover, although comparative hypothesis testing cannot explain our main experimental findings, there are reasons to believe that it influenced our participants' thinking to some extent. Participants were often reluctant to fully commit to the idea that the Numerologists occupied the island even after observing very surprising observations: even in the most surprising condition ( $p \approx .0001$ ), the mean culture rating was only 0.73 . One interpretation of this finding is that people are often unwilling to fully reject a hypothesis space until a better explanation is discovered (see also Griffiths \& Tenenbaum, 2007).

Other researchers have proposed that people employ methods such as sampling to approximate Bayesian inference in situations where it is impossible for them to evaluate the entire hypothesis space (e.g., Sanborn, Griffiths, \& Navarro, 2010). These methods, however, do not address the problem posed by Figure 1. Sampling from $H$ may be useful if this hypothesis space is large, but this approach does not explain how a reasoner might decide that the true hypothesis lies outside $H$. Supporters of sampling might respond that the problem of hypothesis space checking never arises because the space of available hypotheses is always equivalent to $U$. This position, however, seems incompatible with the intuition that scientists and others are sometimes able to generate hypotheses and explanations that are genuinely new.

The justifications for comparative and non-comparative testing remain controversial among statisticians and philosophers (e.g., Howson \& Urbach, 1989/1996; Mayo, 1996; see also Gigerenzer et al., 1990, Chapter 3), but both kinds of hypothesis testing seem necessary to account for the inferences that people make. Non-comparative hypothesis testing is especially notable for the role it plays
in the discovery of new hypotheses. These discovery processes, while often mysterious and difficult to explain, are involved in many of the most interesting inferences that people make. Statistical model checking does not explain where new hypotheses come from, but it can explain why people initiate the search for new hypotheses.

## Acknowledgements

This work was supported in part by NSF grant CDI0835797 and by the Pittsburgh Life Sciences Greenhouse Opportunity Fund. We thank Angela McCarthy for assistance with data collection.

## References

Bayarri, M. J., \& Berger, J. O. (1999). Quantifying surprise in the data and model verification. Bayesian statistics, 6, 53-82.
Earman, J. (1992). Bayes or bust?: A critical examination of Bayesian confirmation theory. MIT Press.
Gelman, A., \& Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. British Journal of Mathematical and Statistical Psychology, 8-38.
Gelman, A., Meng, X. L., \& Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. Statistica Sinica, 6, 733-759.
Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., \& Kruger, L. (1990). The empire of chance: How probability changed science and everyday life. Cambridge University Press.
Gillies, D. (2001). Bayesianism and the fixity of the theoretical framework. In D. Corfield, \& J. Williamson (Eds.), Foundations of Bayesianism. Dordrecht: Kluwer.
Gonzalez, R., \& Wu, G. (1999). On the shape of the probability weighting function. Cognitive psychology, 38 (1), 129-166.

Griffiths, T. L., \& Tenenbaum, J. B. (2007). From mere coincidences to meaningful discoveries. Cognition, 103, 180-226.
Howson, C., \& Urbach, P. (1996). Scientific reasoning: The Bayesian approach (3rd ed.). Open Court Publishing Co. (Original work published 1989).
Kahneman, D., \& Tversky, A. (1972). Subjective probability: A judgment of representativeness. Cognitive Psychology, 3, 430-454.
Mayo, D. (1996). Error and the growth of experimental knowledge. Chicago: University of Chicago Press.
Salmon, W. (1990). Rationality and objectivity in science or Tom Kuhn meets Tom Bayes. In C. W. Savage (Ed.), Scientific Theories. University of Minnesota Press.
Sanborn, A. N., Griffiths, T. L., \& Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. Psychological Review, 117 (4), 1144-1167.
Weaver, W. (1948). Probability, rarity, interest, and surprise. The Scientific Monthly, 67 (6), 390-392.

# Individuals recapitulate the proposed evolutionary development of spatial lexicons 

Alexandra Carstensen (abc@berkeley.edu)<br>Department of Psychology, University of California, Berkeley, CA 94720 USA<br>Terry Regier (terry.regier@berkeley.edu)<br>Department of Linguistics, Cognitive Science Program, University of California, Berkeley, CA 94720 USA


#### Abstract

When English speakers successively pile-sort colors, their sorting recapitulates an independently proposed hierarchy of color category evolution during language change (Boster, 1986). Here we extend that finding to the semantic domain of spatial relations. Levinson et al. (2003) have proposed a hierarchy of spatial category evolution, and we show that English speakers successively pile-sort spatial scenes in a manner that recapitulates that proposed evolutionary hierarchy. Thus, in the spatial domain, as in color, proposed universal patterns of language change based on crosslanguage observations appear to reflect general cognitive forces that are available in the minds of speakers of a single language.


Keywords: Language and thought; spatial cognition; color naming; semantic universals.

## Language as a mirror of the mind

A core question in cognitive science is whether the structure of language reflects the structure of the human mind. Languages vary widely, both in their formal structure and in their semantic categorization of the experienced world (Evans \& Levinson, 2009). At the same time, similar structures and categories appear in unrelated languages, and many logically possible linguistic structures and categories are not attested. A natural question is whether this constrained variation in language reflects universals of human cognition.

One means of pursuing this question concerns language change. One may observe or infer general patterns in the ways languages evolve over historical time, and ask whether these patterns of change, based on observation across languages and across time, are also evident at a given moment in the minds of individuals who speak a single language.

Such a demonstration has already been made in the semantic domain of color (Boster, 1986), and here we present an analogous demonstration in the semantic domain of spatial relations. In what follows, we first describe the Boster (1986) study on color. We then describe recent work on spatial language (Levinson et al., 2003) that proposes a hierarchy for the evolution of spatial categories over historical time. We next present our study, which closely follows Boster's in design. Our central finding is that English speakers successively pile-sort spatial scenes in accordance with Levinson et al.'s (2003) proposed evolutionary hierarchy. We conclude from this finding that generalizations concerning language change may reflect
cognitive forces in the mind of speakers of a single language, in the domain of space as well as in that of color.

## Color categories in language and cognition

Boster (1986) asked speakers of English to successively pile-sort colors. He initially instructed participants to sort a set of eight colors into two "natural groupings" on the basis of similarity, imagining that they spoke a language with only two color terms. He then asked them to subdivide either of those two groups, making three groups total-and so on until each color was in a group by itself. Finally, he tested whether these hierarchical pile-sorts matched a linguistic hierarchy that had been proposed to represent the historical evolution of color categories across languages (Kay \& McDaniel, 1978, elaborating a proposal by Berlin \& Kay, 1969). That hierarchy of color term evolution is shown in Figure 1. ${ }^{1}$ The top split of this hierarchy represents the claim that a two-term color naming system will tend to group BLUE, PURPLE, GREEN, and BLACK into one category, while grouping white, red, orange, and yellow into the other-as in the language Dani (Heider, 1972). Splits lower in the tree represent claims about finer-grained linguistic divisions, which also tend to match cross-language synchronic and diachronic data (e.g. Dougherty, 1977; Kay, 1975).


Figure 1: Kay and McDaniel's (1978) proposed evolutionary hierarchy of color terms.

Boster (1986) found that there was a significant tendency for successive pile-sorts by English speakers to follow the "successive differentiation" (Kay \& McDaniel, 1978: 640)

[^28]of this linguistic evolutionary hierarchy. This finding suggests that, at least in the semantic domain of color, the forces that produce language change over time may be present in the mind of an individual at a given moment.

## An evolutionary hierarchy for spatial language

We wished to further test this claim in a different semantic domain: spatial relations. For this, we required an evolutionary hierarchy of spatial terms, to play the same role in our analysis that Kay and McDaniel's (1978) color hierarchy played in Boster's. Levinson et al. (2003) have suggested such a spatial hierarchy, based on cross-language observations of spatial systems, and drawing an explicit analogy with the above-cited work on color. ${ }^{2}$ They hypothesized that spatial topological categories in the world's languages evolve such that "large categories will tend...to be split into [smaller] categories over time under particular functional pressures" (Levinson et al., 2003: 512), as shown below in Figure 2, to be interpreted as the color hierarchy in Figure 1 was interpreted. ${ }^{3}$


Figure 2: Levinson et al.’s (2003) proposed evolutionary hierarchy of topological spatial concepts.

## The present study

The present study examines successive pile-sorting of spatial scenes by speakers of English, and asks whether these pile-sorts recapitulate the evolutionary spatial category hierarchy proposed by Levinson et al. (2003). If so, that result would generalize the central claim of Boster (1986) to a new semantic domain.

[^29]
## Methods

Following Boster (1986), we performed an experiment with two conditions in which participants sorted spatial stimuli. In both conditions, participants were instructed to sequentially subdivide the eight stimuli-either the line drawings of Figure 3 (scene sorting condition) or corresponding verbal labels (label sorting condition)-into partitions with $2,3,4,5,6$, and finally 7 groups, at which point there were no further decisions to make about which group to split next.

## Participants

A total of 60 participants took part in the two conditions, with 30 participants in each. The population in our study was a convenience sample of the UC Berkeley community; the majority were undergraduate or graduate students, and received either course credit or monetary compensation for their participation. Of the 60 people who completed the task, data from 15 participants were excluded from analysis: 10 due to missing data or failure to follow instructions, 3 because they were not native speakers of English, and 2 who reported familiarity with the color or spatial relational hierarchies proposed by Berlin and Kay (1969), Kay and McDaniel (1978), and/or Levinson et al. (2003). Data from the remaining participants were included in all analyses. Accordingly, 24 participants were included in the scene sorting condition ( 5 female, mean age $=25.6$ ) and 21 participants in the label sorting condition ( 12 female, mean age $=21.3$ ), all of whom had learned English by age 4 (although a number were bilingual), and were naïve to the research hypothesis and related findings.

## Spatial scene sorting

Participants were presented with eight scenes from Bowerman and Pederson's Topological Relations Picture Series (TRPS; 1992). The scenes were arranged linearly on a tabletop in a randomly shuffled order and participants were instructed to successively divide them based on the similarity of the depicted spatial relationships. Each of the eight scenes-shown in Figure 3-depicts an orange figure object located relative to a black background, representing the following spatial relations: NEAR (TRPS scene 37), ON (59), IN (60), ATTACHED (38), UNDER (31), INSIDE (54), ON TOP (34), and OVER (36). These particular scenes were chosen to represent focal "attractors" in spatial semantics (Levinson et al., 2003), analogous to the focal colors proposed by Berlin and Kay (1969) and used in Boster's (1986) color chip sorting task. Each focal spatial scene was selected based on (1) consistency with Levinson et al.'s (2003) characterization of focal attractors within the core spatial categories named above, and (2) the preferences of native English speakers in a pilot study.

Instructions were adapted from Boster (1986) and asked participants to imagine they spoke a language with only two spatial words, and accordingly, to divide up the relations shown in the scenes to make two natural groupings. After participants initially split the eight scenes into two groups,
they were instructed to successively subdivide their categories until all scenes were separated, and each subdivision was recorded to create a full ordered hierarchy of divisions for each participant (see Figure 4 below for an example).


Figure 3: Focal scenes from the Topological Relations Picture Series used in the sorting tasks.

## Spatial label sorting

The spatial label sorting task was identical to spatial scene sorting, except that in this task, participants were presented with the written English spatial expressions NEAR, ON, IN, ATTACHED, UNDER, INSIDE, ON TOP, and OVER. The labels were presented on paper in a randomly shuffled order, and again, participants were instructed to successively divide the stimuli based on the similarity of the spatial relations they describe. As in Boster (1986), the images from the visual sorting task were made available to participants for reference, although they were instructed to base their partitions on the meanings of the spatial phrases themselves, rather than any specific components of the reference scenes.


Figure 4: Example hierarchy from a participant in the scene sorting condition.

## Analysis

Following Boster (1986), we first measured the similarity between Levinson et al.'s (2003) hierarchy (which we refer to as the model) and the empirical data. We then compared this observed similarity to that between the model and random permutations of the empirical data, to determine whether the observed similarity was significant. Finally, we asked whether there was a significant amount of residual data left unaccounted for by the model.

## Similarity metric

In order to compare the empirical color hierarchies made by participants in his experiment to Kay and McDaniel's (1978) theoretical hierarchy representing the diachronic stages of color lexicon evolution, Boster (1986) converted each hierarchy to a similarity matrix. For each pair of colors, he determined the earliest stage in the hierarchy at which those two colors were separated into different groups, and took this to be the similarity between them. Thus, each non-identical pair had a minimal similarity of 1 , meaning they were grouped together only when all eight colors were grouped together, and a maximal similarity of 7, meaning that they were the last pair to be separated, only after the other 6 colors were fully partitioned into groups of 1 each.

We applied the same analysis to the spatial hierarchies produced in this experiment, creating an $8 \times 8$ matrix representing the similarities across all pairs of spatial relations for each participant. Following Boster (1986), we then averaged across corresponding cells in the matrices from all participants in a given condition to create two group similarity matrices-one based on scene sorting and the other on label sorting. As in the color study, we used Pearson correlations to measure the similarity between matrices, where correlations were calculated based on all corresponding pairs of off-diagonal cells.

## Model comparison

Given the empirical similarity matrices from each condition and Pearson correlations as a metric of similarity between such matrices, we ask whether the English speakers in our experiment created hierarchies that were systematically consistent with the cross-linguistic evolution of spatial lexicons as hypothesized by Levinson et al. (2003).

As with the empirical hierarchies, we created similarity matrices based on the Levinson et al. hierarchy which models "successive fractionation of composite concepts." ${ }^{4}$ Like the Kay and McDaniel model (1978), Levinson et al.'s hierarchy includes some variability in the relative order with which certain categories emerge. For instance, the authors leave intentional variability in whether UNDER or a cluster of ON-like relations (i.e. ON, ON TOP, ATTACHED, OVER) are

[^30]split from a more general composite locative concept first. In keeping with Boster's treatment of such variability in the Kay and McDaniel model, we created two model-consistent hierarchies expressing both alternatives (one of which is shown in Figure 2). ${ }^{5}$ Thus, the similarity matrix representing the Levinson et al. model was created by averaging the similarities derived from these two model-consistent hierarchies.

We assessed the alignment of our empirical and model similarity matrices using Pearson correlations, so in order to determine whether these observed correlations were significantly greater than expected by chance, we used Monte Carlo simulations to create a distribution of comparison correlations. To do this, we randomly permuted the labels on our empirical similarity matrices, creating 1,000 permuted variants of each. Each permuted variant was comparable to the original in that all similarity values were preserved in the matrix, but simply re-assigned to different pairings of spatial foci. We then measured the correlation between each of these permuted matrices and the model matrix to determine whether the correlation between the model and the actual empirical data was greater than chance, i.e. that the actual data was more strongly correlated with the model than $95 \%$ of random permutations derived from it.

## Residual analysis

Because our model comparison was based on correlations, it is difficult to assess how well the model explains the observed data beyond testing whether it does so to a significant degree. To this end-and again following Boster's (1986) methods-we employed an analysis designed to determine whether a significant portion of the observed similarity matrix data was left unexplained by the model (Hubert \& Golledge, 1981). The model similarity matrix and two empirical similarity matrices were standardized by subtracting the mean of all values for each matrix from each cell in that matrix, and dividing the result by the standard deviation of the original values in that matrix. The values in each cell of the now standardized model matrix were then subtracted from corresponding cells in the standardized empirical matrices to determine the residual empirical data left unexplained by the model. We measured the Pearson correlations between these residual matrices and their corresponding empirical counterparts.

If the residual matrices no longer bear significant similarity to their full empirical counterparts, we take that to

[^31]mean that the Levinson et al. (2003) model has accounted for the explainable empirical variation. In order to test the significance of the correlation between the residual and observed data, we again create a set of 1,000 simulated matrices by randomly permuting the labels on each of the residual matrices. We measure the correlations between these permuted simulations of the residual matrices and the original empirical matrix and compare this distribution of correlations to that between the actual residual matrices and their empirical counterparts. As before, we take the observed correlation to be significant only if it is greater than that of $95 \%$ of the randomly permuted variants.

## Results

Our similarity analysis found strong correlations between the Levinson et al. (2003) model matrix and the empirical matrices derived from spatial scene sorting ( $r=$ 0.638 ) and spatial term sorting ( $r=0.664$ ), as well as between the two empirical matrices themselves ( $r=0.861$ ). These correlations are presented in Table 1 below alongside the corresponding correlations from Boster (1986). The model and empirical matrices themselves are shown in Tables 2-4.

Table 1: Pearson correlations compared to Boster (1986).

| Correlation | Present study | Boster |
| :--- | :--- | :--- |
| Image sorting vs. model | 0.64 | 0.84 |
| Label sorting vs. model | 0.66 | 0.81 |
| Image vs. label sorting | 0.86 | 0.87 |

Table 2: Similarity matrix from Levinson et al. (2003) hierarchy of topological concepts.

|  | IN |  | INS | UND | NR | OVR | TOP | ATT |
| ---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| ON |  |  |  |  |  |  |  |  |
| IN | 8.0 | 7.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| INSIDE | 7.0 | 8.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| UNDER | 1.0 | 1.0 | 8.0 | 2.5 | 2.0 | 2.0 | 2.0 | 2.0 |
| NEAR | 1.0 | 1.0 | 2.5 | 8.0 | 2.5 | 2.5 | 2.5 | 2.5 |
| OVER | 1.0 | 1.0 | 2.0 | 2.5 | 8.0 | 4.0 | 4.0 | 4.0 |
| ONTOP | 1.0 | 1.0 | 2.0 | 2.5 | 4.0 | 8.0 | 5.0 | 5.0 |
| ATTACHED | 1.0 | 1.0 | 2.0 | 2.5 | 4.0 | 5.0 | 8.0 | 6.0 |
| ON | 1.0 | 1.0 | 2.0 | 2.5 | 4.0 | 5.0 | 6.0 | 8.0 |

Table 3: Similarity matrix from spatial scene sorting.

|  | IN | INS | UND | NR | OVR |  | TOP | ATT |
| ---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| ON |  |  |  |  |  |  |  |  |
| IN | 8.00 | 3.92 | 2.08 | 1.38 | 1.67 | 1.67 | 1.71 | 2.08 |
| INSIDE | 3.92 | 8.00 | 2.29 | 1.29 | 1.04 | 1.54 | 2.08 | 1.67 |
| UNDER | 2.08 | 2.29 | 8.00 | 2.58 | 2.46 | 1.29 | 2.21 | 1.83 |
| NEAR | 1.38 | 1.29 | 2.58 | 8.00 | 3.29 | 1.88 | 2.29 | 1.54 |
| OVER | 1.67 | 1.04 | 2.46 | 3.29 | 8.00 | 3.08 | 1.58 | 2.54 |
| ONTOP | 1.67 | 1.54 | 1.29 | 1.88 | 3.08 | 8.00 | 2.42 | 5.08 |
| ATTACHED | 1.71 | 2.08 | 2.21 | 2.29 | 1.58 | 2.42 | 8.00 | 2.13 |
| ON | 2.08 | 1.67 | 1.83 | 1.54 | 2.54 | 5.08 | 2.13 | 8.00 |

Table 4: Similarity matrix from spatial label sorting.

|  | IN |  | INS | UND | NR | OVR | TOP | ATT |
| ---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| ON |  |  |  |  |  |  |  |  |
| IN | 8.00 | 5.76 | 1.90 | 1.71 | 1.10 | 1.38 | 2.48 | 1.57 |
| INSIDE | 5.76 | 8.00 | 1.90 | 1.38 | 1.33 | 1.43 | 2.19 | 1.38 |
| UNDER | 1.90 | 1.90 | 8.00 | 2.62 | 3.29 | 1.76 | 1.67 | 1.52 |
| NEAR | 1.71 | 1.38 | 2.62 | 8.00 | 1.95 | 1.76 | 2.38 | 1.67 |
| OVER | 1.10 | 1.33 | 3.29 | 1.95 | 8.00 | 3.24 | 1.24 | 2.67 |
| ONTOP | 1.38 | 1.43 | 1.76 | 1.76 | 3.24 | 8.00 | 1.86 | 5.33 |
| ATTACHED | 2.48 | 2.19 | 1.67 | 2.38 | 1.24 | 1.86 | 8.00 | 2.24 |
| ON | 1.57 | 1.38 | 1.52 | 1.67 | 2.67 | 5.33 | 2.24 | 8.00 |

Our permutation analysis found that the randomly permuted variants of the empirical scene matrix were more strongly correlated with the Levinson et al. (2003) model predictions than was the empirical scene matrix itself in only 5 out of 1000 simulations, corresponding to a 1 -tailed $p$-value of .005 . Similarly, only 3 out of 1000 permuted versions of the empirical spatial label matrix were more strongly correlated with the model than was the empirical label matrix itself, corresponding to a 1 -tailed $p$-value of .003. These results (pictured in Figures 5-6) confirm that the observed correlations represent a significant degree of similarity between the empirical matrices and that of the spatial hierarchy model.


Figure 5: Pearson correlations between permuted spatial scene matrices and model matrix.


Figure 6: Pearson correlations between permuted spatial label matrices and model matrix.

The correlation between the empirical scene sorting data and the corresponding residual data after subtracting out the model-explained variation is negligible and not significant ( $r=-0.072$; Monte Carlo 1-tailed $p=0.674$ ). Results are comparable for tests of the correlation between empirical and residual data in the label sorting task $(r=0.073 ; p=$ 0.340 ), which may be interpreted as suggesting that the Levinson et al. (2003) model accounts for all of the explainable observed variation.

## Discussion and conclusions

We find substantial evidence in support of the hypothesis that English speakers synchronically recapitulate Levinson et al.'s (2003) proposed cross-linguistic patterns in the diachronic evolution of spatial lexicons. Our finding in the spatial domain directly parallels that of Boster (1986) in the color domain. Taken together, our finding and his suggest that, at least in these two semantic domains, proposed patterns of language change may be reflected in the minds of individuals at a given moment.

At the same time, there are at least two grounds for caution. First, as we have noted, the Levinson et al. (2003) hierarchy was intended as a tentative diachronic hypothesis, based on synchronic cross-language observation-not as a firm diachronic claim. Direct assessment of that hierarchy using historical data has to our knowledge not yet been conducted, and would be needed before our account can be considered to concern actual, rather than merely proposed, patterns of spatial language change. Second, our analyses, like Boster's (1986), were based on a comparison between model predictions and an aggregate measure of all participants' sorting. When viewed in this way, the evidence does support the recapitulation claim. However, no participants either in Boster's (1986) study or in ours actually recapitulated the model predictions exactly. This may not be surprising given the large number of hierarchical pile-sorts that are possible, some of which are only minimally different from model predictions. Still, in future research it would be informative to analyze such data in a way that separately measures how close each participant came to the model prediction, rather than rely solely on an aggregate measure of all participants' behavior. Such analyses may support a more precise picture of the extent to which individuals recapitulate broad proposed generalizations concerning language change. The present study, like Boster's (1986), has nonetheless demonstrated that such recapitulation is clearly present as a general shared tendency-and that in this sense at least, the character of language change may reflect the structure of the mind.

## Acknowledgments

We thank Joshua Abbott and two anonymous reviewers for helpful comments. This work was supported by NSF under grant SBE-1041707, the Spatial Intelligence and Learning Center (SILC).

## References

Berlin, B., \& Kay, P. (1969). Basic color terms: Their universality and evolution. Berkeley: University of California Press.
Boster, J. (1986). Can individuals recapitulate the evolutionary development of color lexicons? Ethnology, 26(1), 61-74.
Bowerman, M. \& Pederson, E. (1992). Cross-linguistic studies of spatial semantic organization. In Annual Report of the Max Planck Institute for Psycholinguistics 1992 (pp. 53-56).
Dougherty, J.W. D. (1977). Color categorization in West Futunese: Variability and change. In M. Sanchez (Ed.), Sociocultural dimensions of language change. New York: Academic.
Evans, N., \& Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for
cognitive science. Behavioral and Brain Sciences, 32, 429-492.
Heider, E. R. (1972). Probabilities, sampling, and ethnographic method: The case of Dani colour names. Man, 7, 448-466.
Hubert, L., \& Golledge, R. (1981). A heuristic method for the comparison of related structures. Journal of Mathematical Psychology, 23, 214-226.
Kay, P. (1975). Synchronic variability and diachronic change in basic color terms. Language in Society 4: 25770.

Kay, P., \& McDaniel, C.K. (1978). The linguistic significance of the meanings of basic color terms. Language, 54, 610-646.
Levinson, S. C. \& Meira, S. (2003). Natural concepts in the spatial topological domain-adpositional meanings in crosslinguistic perspective. Language, 79, 485-516.

# The development of predictive processes in children's discourse understanding 

Marisa Casillas<br>middyp@stanford.edu<br>Department of Linguistics<br>Stanford University

Michael C. Frank<br>mcfrank@stanford.edu<br>Department of Psychology<br>Stanford University


#### Abstract

We investigate children's online predictive processing as it occurs naturally, in conversation. We showed 1-7 year-olds short videos of improvised conversation between puppets, controlling for available linguistic information through phonetic manipulation. Even one- and two-year-old children made accurate and spontaneous predictions about when a turn-switch would occur: they gazed at the upcoming speaker before they heard a response begin. This predictive skill relies on both lexical and prosodic information together, and is not tied to either type of information alone. We suggest that children integrate prosodic, lexical, and visual information to effectively predict upcoming linguistic material in conversation.


Keywords: Prediction; online comprehension; turn-taking; timing; child language; prosody; eye-tracking

## Introduction

Conversation is the primary way we use language. It is most often spoken face-to-face with two or more speakers, and is deeply embedded in our current interactional context. Participants in conversation don't just listen; given that inter-turn gaps are so brief, speakers must be simultaneously planning at least part of their response while the current speaker is still talking (Sacks et al., 1974; Stivers et al., 2009). So under normal conditions, listeners deal with critically different processing pressures during conversation than they do in rigidly controlled experiments. For children especially, experience and skill in processing and conversation can be critical to later language development (e.g., Weisleder, 2012). The current study seeks to draw a link between language processing in the lab and language processing in broader contexts by tracing predictive processing during conversation across a broad developmental sample.

Timing is critical in conversation because, in addition to parsing the linguistic signal for its parts and meanings, conversational participants are interested in the upkeep of the ongoing interaction. For example, if someone asks you, "What are your plans for dinner?" you are obligated to do more than just parse the linguistic signal; you must respond. You can't just respond at your convenience either-especially for an implied invitation like this one, a slight hesitation might communicate that you will turn the offer down. This is a substantial cognitive load to bear since speakers must quickly figure out what was said and how to respond. Predictive processes can help maintain the flow of conversation by allowing us to plan for what is likely to happen next in the interaction.

When listening to a single utterance, we make predictions about what the speaker will say next. Many studies have shown that we can use a wide variety of linguistic and nonlinguistic cues to incrementally update our expectations about
what linguistic material will follow (e.g., Altmann \& Kamide, 1999; Ito \& Speer, 2008; Snedeker \& Yuan, 2008).

In multi-utterance contexts, like conversation, speakers must also coordinate their ongoing actions, and so our predictive prowess is even more to our advantage than it is in the lab. There is both naturally-occurring and experimental evidence that adults effortlessly anticipate when a speaker switch will occur during conversation. In order to respond with brief gaps, they need to accurately predict when to begin speaking (Sacks et al., 1974). There is also experimental evidence that adults can anticipate upcoming turn-structurewhen asked to press a button when they think a speaker will finish her turn, adult listeners demonstrate incredible timing accuracy ( $M=168 \mathrm{~ms}$ from the offset of speech; de Ruiter et al., 2006). They also spontaneously track turn-timing and anticipate upcoming speakers with their gaze when watching videos of conversation (Tice \& Henetz, 2011).

Here we ask: do children also make online predictions about conversation? We focus on how children process the multi-utterance speech around them because, as mentioned, children's conversational skill and experience can influence their later language learning.

Children learn language in the context of conversation, and their conversational skills allow them to practice comprehending and using language with others. Children begin to take turns in early infancy, but their coordination of turn-timing with others takes several years to develop. By four months, infants regularly engage in coordinated back-and-forth interactions with their caregivers (Masataka, 1993). Twelve-month-old infants who are watching two-person conversations can (1) track who is speaking, and (2) expect speech to be responded to verbally (Thorgrímsson et al., 2011). Despite this, even at $5 ; 0$, children's timing is significantly delayed compared to adults'-their response delay at $3 ; 0$ is up to 10 times slower (Casillas et al., in prep)—leading some to believe that children can't or don't perform the same predictive processing that adults do (Garvey \& Berninger, 1981). ${ }^{1}$

We propose that, on the contrary, children develop their predictive turn-taking skill early in life, and that their apparent delay is due to the time needed to plan and execute a response. Thus, when children simply observe an ongoing interaction, they show predictive timing similar to adult norms. Casillas and Frank (2012) found that when children and adults watched videos of conversation in a language they didn't speak, they were able to use the available information

[^32](prosodic, temporal, and visual) to track and anticipate the ongoing turn structure with their gaze. Because some linguistic units are more informative than others in predicting turnboundaries (e.g., words > intonation; de Ruiter et al., 2006), we also hypothesize that, like adults, children's online predictions about turn-taking are more heavily influenced by lexical information than they are by prosodic information. By testing these proposals we can (1) track children's development of predictive turn processing during discourse while (2) also beginning to tease apart which linguistic cues children attend to in making their predictions about turn-structure. We measured children's online anticipation about who will speak next in conversation and found that children use multiple linguistic cues to make accurate predictions about what will come next-and they do so even at 1-2 years old.

## Method

We tracked children's eye movements as they watched short videos of conversation to measure their predictive gaze to upcoming speakers at points of speaker-transfer. We controlled the audio signal to limit children's access to either prosodic information or lexical information, making comparisons to their gaze behavior in normal audio conditions and conditions without any linguistic information. We focus here on effects of linguistic information, so we eliminated visual cues to turn-taking by using videos of puppets to replace the original videos of our speakers.

## Participants

We recruited 129 children ages $1 ; 0-7 ; 0$ from the Children's Discovery Museum in San Jose, CA, to participate in the current study. We collected data from 20-23 children for each of the six 1-year age groups. All participants were native English speakers, though some parents reported that their child heard a second (and sometimes third) language at home. ${ }^{2}$

## Materials: Puppet videos

Audio-recordings We recorded six 20-25 second twoperson conversations for use in the puppet videos. Each of the six conversations featured a native English-speaking male and female talker. Talkers were directed to improvise a short conversation on a given topic (one of: 'riding bikes', 'pets', 'breakfast', 'birthday cake', 'rainy days', 'the library'). We asked talkers to talk "as if they were on a children's television show" to establish a child-friendly style. We gave talkers approximately five minutes to work out a basic conversation and then perform it with minimal practice. We edited each conversation to a 20-25 second clip for use in the final video stimuli.

[^33]

Figure 1: The six puppet pairs (and associated audio conditions). Each pair was linked to three distinct conversations from the same condition across the three experiment versions.

Audio Manipulation To control for the linguistic information available in the final puppet videos, we phonetically manipulated the recordings to fall into four conditions: Normal, Words-only, Prosody-only, and No Discernible Speech. Normal videos simply used the 20-25 second audio recording. Words-only videos featured manipulated speech in which intonation was flattened to each talker's average pitch (F0) and every syllable nucleus and coda duration were set to each talker's average nucleus and coda duration. ${ }^{3}$ To do this we used PSOLA resynthesis in Praat (Boersma \& Weenink, 2012). The resulting audio signal was devoid of pitch and durational cues to turn-boundary, so we referred to this audio as 'robot' speech when talking to children. Prosody-only videos also featured manipulated speech, in which the original audio recording was low-pass filtered at 500 Hz with a 50 Hz Hanning window (following de Ruiter et al., 2006). Low-pass filtering removes the phonetic information used to distinguish between phonemes, and so the resulting audio has no identifiable words, but retains the original intonational and rhythmic qualities of the conversation. Low-pass filtered audio sounds muffled, like voices under water, so we referred to this audio as 'merperson' speech. To create Non-discernible speech audio, we overlaid eight different child-oriented conversations (not including the original one) to create multi-talker babble. This is sometimes referred to as 'cocktail party' speech, but we referred to it as 'birthday party' speech. Finally, the Prosody-only audio sounded much quieter than the other conditions because it lacked acoustic energy above 500 Hz , so all other audio conditions were adjusted to match its lower volume.
Video-recordings We then created puppet videorecordings to match the final audio signals. The puppets were designed to be minimally expressive so that the experimenter could only control the opening and closing of their mouths. There were three Normal condition puppet pairs-'red',

[^34]

Figure 2: Proportion gaze to the answerer during the first 333 ms of the answer. Age in years is plotted on the x -axis for each of the four conditions (Question-Answer switches = dark gray; Non-Question-Answer switches = light gray). The vertical bars show the $95 \%$ confidence intervals around each point.

| Condition | Current | Non-current | Elsewhere |
| :--- | :--- | ---: | ---: |
| No discernible speech | 0.51 | 0.15 | 0.35 |
| Prosody only | 0.55 | 0.14 | 0.31 |
| Words only | 0.65 | 0.14 | 0.21 |
| Normal speech | 0.68 | 0.14 | 0.18 |

Table 1: Overall proportion gaze, averaged across all participants, to the current and non-current speakers (and elsewhere) during utterances.
'blue' and 'yellow' ones-and one puppet pair for each of the other conditions: 'robots', 'merpeople', and 'party-goers' (Figure 1). Three conversation topics ('birthday cake', 'pets', and 'breakfast') were used for the Normal conversations, and three ('riding bikes', 'rainy days', and 'the library') were used for the other three conditions. We created three versions of the experiment so that each of the six puppet pairs was associated with at least three different conversation topics. We then hand-aligned the final audio to the puppet video recordings and ensured that half of the videos in each version were female-left-male-right and vice-versa by flipping the video and audio channels as needed. ${ }^{4}$

## Procedure

We seated children in front of a large screen with speakers placed below and at the sides of the screen. Mounted beneath the screen was an SMI 120 Hz remote infrared eye-tracker that continuously recorded their eye movements throughout the experiment. Children then watched a series of short videos comprising six brief puppet conversations and five engaging filler videos (e.g., running puppies and music). The filler videos were inserted between the puppet videos, which were ordered randomly for each participant. The six puppet videos fell into four audio conditions: Normal (3), Words only (1), Prosody only (1), and No Discernible Speech (1). The entire experiment took less than five minutes for most children.

[^35]
## Data analysis

For each participant in the study, we only included data from those video segments in which the participant gazed at the video for more than $75 \%$ of its duration. In prior work (Casillas \& Frank, 2012; Tice \& Henetz, 2011) adults and children $3 ; 0$ and older made anticipatory gaze shifts to upcoming talkers while watching videos of conversation. The shifts sometimes began before the prior turn ended, within the final 300 ms of speech. To determine whether children $1 ; 0-7 ; 0$ in our data made similar anticipatory shifts, we conducted our analyses contingent on looks to the prior listener. Specifically, we only included children who were looking at the prior speaker 333 ms before the prior turn ended. This follows contingent looking analyses in other child language work (Fernald et al., 2008) and guarantees that the children in our analyses were prepared make a gaze switch to the upcoming speaker. We then averaged gaze to the upcoming speaker during the first 333 ms of the answer. ${ }^{5}$ Since each child in our analysis started by looking at the prior speaker, looks to the upcoming speaker at the answer onset will represent the magnitude of children's anticipatory gaze shifts. Because prior work has found that children shift their gaze more quickly after hearing a question than non-question (Casillas \& Frank, 2012), we separated these in our analysis. When gaps are too long they can signal a troubled speaker transition or a disfluency that might need conversational repair (Jefferson, 1974). For this reason we excluded all turn-transitions longer than 550 ms in our stimuli.

## Results

In all conditions, participants were nearly three times more likely to keep their eyes on a talker when that person was speaking, rather than when they were silent (Table 1). Par-

[^36]

Figure 3: Proportion gaze to the answerer during the last 1 second of the prior turn and the first 1 second of the upcoming turn, broken down by participant age and linguistic condition (Question-Answer switches = dark gray; Non-Question-Answer switches $=$ light gray). Error bars indicate the standard error of the mean. The inter-turn gap is represented by the blank area along the trajectory. Included speaker switches had gaps ranging from $3-497 \mathrm{~ms}(M=308 \mathrm{~ms})$. The gap shown above is 300 ms . Gaps in the stimuli varied in length, so looking data during this period isn't plotted.
ticipants looked away from the talkers 18-35\% of the time. This closely matches our prior results (Casillas \& Frank, 2012), though children looked elsewhere more often in the No Discernible Speech and Prosody Only conditions than in the other two. Children's consistent looks to the current, rather than the non-current, talker suggest that the participants were tracking basic turn-taking with their gaze by using information from the audio, video, or (most likely) both. Participants most consistently looked at the current speaker (and looked away least) in the Normal Speech condition.

Children of all ages and in all conditions made anticipatory shifts to upcoming speakers (Figures 2 and 3). Even in the No Discernible Speech condition-in which the children saw puppets mouthing words to unrelated multi-talker babblechildren shifted their gaze toward upcoming speakers by the time the response began. This anticipatory shift was much
smaller in magnitude than what we found for Normal Speech with the same children ( $\sim 25 \%$ vs. $\sim 40 \%$ ).

Perhaps surprisingly, when children only had access to prosodic or lexical information, they performed similarly to when they had no linguistic information at all with slight, if any, improvement with age (Figures 2 and 3). In the Words Only condition, looks to the answerer showed a small, but consistently greater magnitude for Question-Answer turn switches than for non-Question-Answer switches. Switch in gaze to the upcoming speaker was strongest in the Normal Speech condition, in which older children also clearly distinguished their gaze to Question-Answer and non-QuestionAnswer switches.

To test the reliability of the differences in anticipation between conditions and switch types, we fit a linear mixedeffects model (Gelman \& Hill, 2007) to participants' aver-
age gaze at the upcoming speaker during the first 333 ms of the response. We included turn-switches and subjects as random effects, using maximal random effects structure (Barr et al., 2013) to control for variability between participants on the switch type (Question-Answer vs. non-Question-Answer) and linguistic condition. We also included a three-way interaction term for age, switch type, and condition with two-way interaction terms for age and gap duration and condition and gap duration. ${ }^{6}$

Model coefficients suggest that there were two significant effects in the gaze data. First, duration is a significant predictor of anticipation; longer gap durations result in more anticipation $(\beta=0.73$, s.e. $=0.28, t=2.61)$. Second, there is a highly significant three-way interaction between age, switch type, and condition ( $\beta=0.11$, s.e. $=0.03, t=3.16$ ) in predicting anticipation. This derives from the Normal Speech condition, in which children's differential looking behavior for Question-Answer vs. non-Question-Answer switches clearly diverges with age. No other coefficients reached significance.

Are children simply reacting to turn-ends and then looking to the other puppet, or are they instead anticipating the end of the prior speaker's turn and looking early on? To test this we fit a second model on turn-transitions that lasted less than 200 ms. Anticipatory looks in this subset of the data must have been planned before the prior turn ended. Model coefficients suggest anticipation still occurs with 2 two-way interactions between age and condition for Words only and Normal speech ( $\beta=0.07$, s.e. $=0.03, t=2.18$ and $\beta=0.08$, s.e. $=0.06, t=$ 2.31 , respectively). ${ }^{7}$

## General Discussion

Children's looking patterns suggest that they reach at least two developmental benchmarks for predictive processing for discourse. First, children recognize that turn-taking requires immediate responses, and they quickly integrate linguistic and non-linguistic cues to shift their gaze in anticipation of a response. We saw this behavior from all children in our data set. As children grow older, they become more sensitive to linguistic cues, using them to distinguish between different conversational actions (questions vs. non-questions) and make earlier and swifter predictive shifts.

Since children in this age range still appear quite delayed in their own turn-taking, these results are strong evidence that children's apparent delays in everyday conversation are not due to the ability to predict when a turn-switch will occur. We propose that these delays are instead due to the cost of planning a response. Children's turn-timing during conversation is most delayed when they must make a complex response, and so a three-year-old's timing during conversation may appear to be slower than a one-year-old's (Casillas et al., in prep). But in our task, when the cognitive load was lightened so children were only required to perform comprehen-

[^37]sion, we saw that children's skill in predicting turn-structure develops early on and becomes more sensitive to discourse distinctions with age, using linguistic information to distinguish between different conversational acts (e.g., questions).

Children made their earliest and most consistent predictive looks in the condition where they had all linguistic information available to them. These results strengthen claims from previous work that young children spontaneously anticipate what is coming next in conversation (Casillas \& Frank, 2012). By testing a broad age range, we found that children show greater anticipation, with a greater advantage for question- over non-question switches as they get older. Children in our study could effectively make predictions about normal speech by age $1 ; 0$, but that they begin distinguishing between different types of conversational actions (questions vs. non-questions) by the time they are $3 ; 0$ (Figure 3). Question effects are strongest when both prosodic and lexical cues are present, contrary to prior findings with adult listeners that found lexical information sufficient to predict upcoming turnend boundaries (de Ruiter et al., 2006).

Children's performance was significantly downgraded by phonetically controlled stimuli such that their predictive eye movements were comparable to conditions in which they had no linguistic information at all. We suggest that children were able to make anticipatory shifts without linguistic information because they simply waited for one puppet to stop talking before looking to the next. Rather than anticipating the end of the ongoing turn, these children are likely anticipating the start of the next speaker's turn, which explains the significant effect of longer inter-turn gaps. In contrast, anticipation in the Normal Speech and Words only conditions still occurs when gaps are shorter than 200 ms , in which case children do not have time to simply react to the end of the prior turn and make significant shifts by the start of the response-in these cases they must have instead anticipated the end of the prior turn.

One limitation of the current study is that, by using puppets for the visual signal, we removed all visual cues to turn-taking except mouth movement. We did this to focus our analysis on linguistic cues, but visual cues are culturally variable and important indicators of conversational timing and coordination (Kendon, 1967; Stivers et al., 2009). In related work (Casillas \& Frank, 2012), we asked 3-5 year-old children to watch short clips of conversation in languages they didn't speak. We saw larger and earlier-initiated anticipatory shifts in that experiment even though children in that study had no access to lexical information, only non-native prosodic and visual cues. Since children in the current study have smaller shifts, even in the Normal Speech condition, we suspect that visual cues play a large role in helping children guess what will come next, and that children integrate these cues with linguistic information when given the chance. Further work will be required to test this hypothesis. Also, children rarely hear phoneticallycontrolled speech, and may not have been able to process it as efficiently as normal speech, though they still were able to make small anticipatory shifts.

## Conclusion

Just as children must learn to break into the linguistic stream and segment it into words, they must also learn to break into the interactional stream of conversation and segment it into turns. Using children's spontaneous gaze behavior while watching improvised conversations, this study has attempted to link online predictive processes with naturalistic, conversation-based stimuli. We have focused here on children's predictive skill in conversation because children's conversational skills can impact the form of their linguistic input and may be critical to understanding what children hear and how they practice language. Children's turn-taking skills help them become active interactants who have control over the linguistic input and practice they receive.

The implications of conversation-specific skills for language development are likely important (e.g., Weisleder, 2012), but are still largely unknown. Within single- and multi-utterance sequences, children's ability to predict what's coming next can aid in their uptake of new information (Fernald et al., 2008). By being able to predict what upcoming turn-structure will look like and anticipating the type of response needed for different types of actions (e.g., questions vs. non-questions), children are developing conversational skills that affect their input more globally: they can become more successful participants in multi-party interaction. Their skill in prediction within and across utterances then affects the type and quality of linguistic practice that children get during development.

Our findings indicate that rapid turn-timing is one of the earliest properties of organized interaction that children acquire, and that over the first seven years of life, children come to rely on their linguistic knowledge to refine and build on their predictions about what to expect next in conversation. So while children learn about language, they can use their linguistic knowledge online to take turns more effectively, and as children learn to take turns, they can use language more effectively in conversation with others.

## Acknowledgments

We thank C. Kurumada, M. Lewis, and E. V. Clark for their helpful feedback, also the John Merck Scholars Program \& the Hellman Faculty Fellows Program. This research is supported by an NSF dissertation grant to M. Casillas.

## References

Altmann, G., \& Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. Cognition, 73(3), 247-264.
Barr, D. J., Levy, R., Scheepers, C., \& Tily, H. J. (2013). Random-effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language, 68(3), 255-278.
Boersma, P., \& Weenink, D. (2012). Praat: doing phonetics by computer [Computer software manual]. Available from
http://www.praat.org ([Computer program] Version 5.3.16)

Casillas, M., Bobb, S. C., \& Clark, E. V. (in prep). Turntaking, timing, and access in early language acquisition.
Casillas, M., \& Frank, M. C. (2012). Cues to turn boundary prediction in adults and preschoolers. In Proceedings of SemDial 2012 (SeineDial): The 16th workshop on the semantics and pragmatics of dialogue.
de Ruiter, J., Mitterer, H., \& Enfield, N. (2006). Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. Language, 82, 515-535.
Fernald, A., Zangl, R., Portillo, A., \& Marchman, V. (2008). Looking while listening: Using eye movements to monitor spoken language comprehension by infants and young children. In I. Sekerina, E. Fernandez, \& H. Clahsen (Eds.), Language acquisition and language disorders (Vol. 44, pp. 97-135). Amsterdam/Philadelphia: John Benjamins.
Garvey, C., \& Berninger, G. (1981). Timing and turn taking in children's conversations. Discourse Processes, 4(1), 2757.

Gelman, A., \& Hill, J. (2007). Data analysis using regression and multilevel/hierarchical models (Vol. 1). Cambridge University Press New York.
Ito, K., \& Speer, S. (2008). Anticipatory effects of intonation: Eye movements during instructed visual search. Journal of Memory and Language, 58(2), 541-573.
Jefferson, G. (1974). Error correction as an interactional resource. Language in Society, 2, 181-199.
Kendon, A. (1967). Some functions of gaze-direction in social interaction. Acta Psychologica, 26, 22-63.
Masataka, N. (1993). Effects of contingent and noncontingent maternal stimulation on the vocal behaviour of threeto four-month-old Japanese infants. Journal of Child Language, 20(02), 303-312.
Sacks, H., Schegloff, E., \& Jefferson, G. (1974). A simplest systematic for the organization of turn-taking for conversation. Language, 50, 696-735.
Snedeker, J., \& Yuan, S. (2008). Effects of prosodic and lexical constraints on parsing in young children (and adults). Journal of Memory and Language, 58(2), 574-608.
Stivers, T., Enfield, N., Brown, P., Englert, C., Hayashi, M., Heinemann, T., et al. (2009). Universals and cultural variation in turn-taking in conversation. PNAS, 106, 1058710592.

Thorgrímsson, G., Fawcett, C., \& Liszkowski, U. (2011). 12-month-olds expect speech to provoke a response in others' interactions. Poster presented at the Biennial Meeting of the Society for Research in Child Development, Montréal.
Tice, M., \& Henetz, T. (2011). Turn-boundary projection: Looking ahead. In Proceedings of the 33rd annual meeting of the Cognitive Science Society.
Weisleder, A. (2012). Richer language experience leads to faster understanding: Links between language input, processing efficiency, and vocabulary growth. Unpublished doctoral dissertation, Stanford University.

# English/Italian Bilinguals Switch Gesture Parameters when they Switch Languages 

Federica Cavicchio (federica.cavicchio@unitn.it)<br>Center for Mind/Brain Science, Corso Bettini 31<br>Rovereto (Tn), 38068 IT<br>and<br>School of Psychology, University of Birmingham<br>Edgbaston, B15 2TT UK<br>Sotaro Kita (kita.s @bham.ac.uk)<br>School of Psychology, University of Birmingham<br>Edgbaston, B15 2TT UK


#### Abstract

We investigated gestural communication in early bilinguals. In particular, we tested which aspects of gestures were "transferred" from a language to another. Though transfer in spoken languages has been studied extensively, transfer in gesture is understudied. Gesture transfer can provide useful information on the cognitive architecture in bilingualism. In this study our focus is on gesture rate and gesture space in Italian/English bilinguals. Contrary to previous findings, we have no evidence of transfer. When bilinguals switch language, their gesture parameters switch accordingly. The switch of gesture (cultural) parameters such as rate and salience show that language and gesture are tightly linked. This suggests that a language and the corresponding gesture parameters might be selected in a high level processing stage at which verbal and nonverbal aspects of communication are planned together.


Keywords: bilingualism; gesture rate; gesture space; linguistic transfer; gesture transfer; lexical access.

## Introduction

Different languages and cultures use gestures differently. For example, Italian is reported as a high gesture frequency language (Barzini, 1964; Kendon, 1992, 1995), as opposed to (British) English, described as a low gesture frequency language (Graham and Argyle, 1975).

It has been claimed that bilinguals' gesture use is linked to their proficiency in the two spoken languages. A common measure of gesture use is gesture rate (the number of gestures performed over the number of words uttered). Sherman and Nicoladis (2004) found no differences between bilinguals' gesture rate when participants have an equal proficiency in both their languages (Canadian English and Spanish, where Spanish is supposed to be a high frequency gesture rate language). Those studies (Nicoladis et al., 1999; Pika, et al., 2006), however, are not very informative about whether or gestural transfer occurs due to the limitation in the design; for example they lack one of the monolingual control groups (see Nicoladis, 2007).

The evidence for gestural transfer in the literature is mixed. In a study on English/French bilingual children in Canada Nicoladis and colleagues (2005) found that bilinguals tend to gestures more than both the two monolingual control groups, but no evidence for gestural
transfer was found. Nicoladis and colleagues explained their results claiming that bilinguals have more "choices" about how to package verbal messages with respect to monolinguals. Therefore, bilinguals gesture more than monolinguals (see also Pika et al., 2006). On the other hand, a study by So (2010) found gesture transfer between American English and Mandarin Chinese in EnglishMandarin bilinguals in Singapore. American English monolinguals gestured significantly more than Mandarin speaking monolinguals. Bilinguals gestured more when speaking Mandarin than the Mandarin monolingual control group, and when speaking English, they gestured at about the same rate as English monolinguals.

Another gesture parameter that varies across cultures is gesture size. Since the seminal study of Efron (1941/1972) comparing Jewish and Italian immigrants' gestures, we know that in different cultures gestures differ in how they are performed in the space. In particular, Efron observed that Italian immigrants' gestures were spatially expansive, moving the entire arm from the shoulder joint, and tended to occupy the lateral (transversal) plane. More recently, Müller (1998) compared the gesture space of native Spanish and German speakers involved in a naturalistic conversation task with a language matching confederate. She found that Spanish speakers produced more gestures in the space above their shoulder than German speakers. Interestingly, Müller did not find difference in gesture rates between German and Spanish. She suggested that the difference in gesture salience create an 'illusion' that Mediterranean region cultures gesture more frequently than north European cultures.

Gesture size is an interesting variable to consider for gesture transfer in bilinguals. First, gesture size varies crossculturally: bigger in Mediterranean cultures than in northern European cultures. Second, gesture size is determined by different psychological processes than gesture rates (Chu, Meyer, Foulkes \& Kita, under review). Thus, gestural transfer or lack of transfer for gesture rates and gesture size may shed light on the relationship between speech and gesture production processes. However, no previous studies have investigated gesture size in bilinguals.

Because the evidence for transfer of gesture rates in the
literature is mixed and there are no studies on transfer of gesture space, we investigate transfer of gesture rates and gesture space in Italian-English bilinguals. We tested two monolingual control groups so that we can properly address the question whether parameters of gesture production transfer in bilinguals and whether bilinguals gesture differently from monolinguals. The two monolingual control groups of English and Italian speakers matched with the bilinguals for gender, age and education background. We focused on highly proficient Italian/English early bilinguals (i.e. they learned both languages before age 6) who had a very similar fluency in both languages. Bilinguals and monolinguals described the exact same stimuli in each language to a confederate language matching speaker. Differently from previous studies using long cartoons that were edited in shorter scenes, our stimuli consisted of 10 single-scene cartoons.

## Method

## Participants

30 participants (10 English native speakers, Females $=8$ Males=2, age mean=22.3 years, recruited at the University of Birmingham; 10 Italian native speakers, Females=8, Males $=2$ age mean $=23.1$, recruited at the University of Trento; and 10 English/Italian bilinguals, Females=8, Males=2; age mean=23.8, recruited at the University of Birmingham and Trento) took part to the experiment. All the participants took a test to assess their linguistic background (Gullberg \& Indefrey, 2003) and the Controlled Word Association Test (COWAT; see Loonstra et al., 2001 for a review) in English and Italian. The COWAT scores ensured that participants were equally fluent in both languages. The mean fluency score for bilinguals was 62.8 words in English and 62.2 words in Italian. The mean fluency score for Italian native speakers was 63.4 words and for English native speakers was 61.7 words.

The bilinguals enrolled in this study started speaking both languages before age 6 , while the native speakers of Italian and English did not learn any other language before age 11 and were not fluent respectively in English or Italian. They were all students enrolled at university bachelor or master degrees.

## Materials

Participants watched 10 Tomato man stimuli (Özyürek, Kita, \& Allen, 2001) depicting two characters (i.e. Tomato man and the green Triangle) performing some actions (Fig. 1, left panel). The goal of these stimuli was eliciting the description of manner and path in the verbal and gesture modality as the participants described Tomato and Triangle actions. The stimuli were presented on a 13 -inch TFT monitor at a resolution of $800 \times 600$. Stimulus presentation was controlled by a PC running Power Point. The participants were audio and video recorded with a Sanyo Xacti HD2000 camera at a medium shot (i.e. they were shot from up their head to their knees, Fig. 1, right panel).


Figure 1: On the left panel, an example of the Tomato man cartoons used to elicit participants' gestures. In this movie Tomato man is "rolling down the hill". On the right panel, a participant describes the cartoon. The two dotted concentric squares define the gesture space: centre (the inner square) and periphery (the outer square).

## Procedure

Participants were seated at approximately 40 cm from the computer screen. An assistant pressed the mouse button to start the experiment. After the participants saw the first stimulus they turned toward a listener sitting near the camera and described what they had just seen. The monolingual participants repeated twice the task in the same language to two listeners who are native speakers of the relevant language. The bilingual participants repeated the task once in Italian, talking to a native speaker of Italian, and once in English to a native speaker of English.

The order of the stimuli was counterbalanced. In particular the stimuli run from clip 1 to clip 10 for the forward order and from 10 to 1 for the backward one. For bilinguals, the order of the task repetition was counterbalanced by language across participants.

## Data Transcription and Analysis

## Transcriptions

Two native speakers of Italian and English transcribed the videotapes following the instruction manual. Disfluencies, repetitions and laughter were transcribed with special fonts. The transcriptions were checked for accuracy by a second fluent speaker. All the transcriptions were reported in Elan 4.3.3 to ensure a correct time alignment with coverbal gestures.

Gesture were transcribed and aligned with videos and transcriptions.

## Gesture Coding

We coded the gestures produced by participants when telling the whole cartoon to the listeners.
In this paper we focus on two main aspect of gesture production:
Gesture Rate was calculated as the number of gesture produced by each participant describing each cartoon over
the number of words produced in each cartoon description (Ngestures/Nwords).
Gesture Salience: Gesture salience was coded for the target gesture performed during the cartoon description (e.g. rolls up, tumble down etc.). To code salience we followed McNeill (1992), who divided the gesture space into sectors using a system of concentric squares. Our annotation coding scheme reflects this notation dividing the gesture space in 2 sectors (see Fig. 1, left panel): "centre" and "periphery". When the gesture stroke was produced in the central sector, the gesture was annotated with 0 (not salient), whereas when the gesture stroke was produced in the periphery sector, the gesture was annotated with 1 (salient).
To ensure the reliability of the adopted coding scheme, a subset of the corpus ( 659 gesture tokens) was annotated by three independent coders. For gesture salience we found a high agreement above the chance level (Kappa $=0.89$ ).

## Results

We analysed our data in a linear (for gesture frequency) and a general (for gesture salience) mixed-effect model, as implemented in the statistical package, R. The analysis was run in R 2.15 using the package lme4, version 0.999999-0 (the function glmer was used for the gesture salience analysis).

## Gesture Rate

A linear mixed model was performed on Gesture Rate (observations $\mathrm{n}=390$ ). We fit the linear mixed model on gesture rate using a "maximum model random slopes" approach, i.e. calculating random (slopes and intercepts) effects for subject and item as well as the following fixed effects: language (Italian vs. English) and language status (bilingual vs. monolingual) and the interaction between language and language status. Because of the high correlation in the random effects (and the consequent danger of over fitting the data), we used a "backward algorithm" to set for the model that best described the variance in the data without over fitting them. Starting from the maximal random slopes and intercepts model, we first tested for the exclusion of random slopes. In this way we set, step by step, for the simpler model that better described the variance of the data. To ensure that the models described the same amount of variance, in each step we confronted the fitting of the simpler model with the previous "more random" ones. The model that better described the variance of the data had random intercepts for subjects, random intercepts for items (cartoons) and Language (Italian or English) varying by subjects random slopes.

We found a significant effect for Language (Est. $=0.06$, S.E. $=0.01, \mathrm{p}<0.001$ ), such that the gesture rate is higher in Italian than in English, but no significant effect for Language Status (monolingual or bilingual; Est.=-0.02, $\mathrm{S} . \mathrm{E}=0.02, \mathrm{p}=0.19$ ). Interaction between the fixed effects (Language status and Language) was investigated but not found ( $\mathrm{Est}=0.007$, S.E. $=0.03, \mathrm{p}=0.84$ ).
p values were calculated from the t values obtained in the
linear mixed effect model output. We treated the $t$ values as they were draw from a normal distribution, using the pnorm function in R. A post hoc power analysis through simulation ( n simulations=1000) revealed that 27 participants per group (81 participants in total) would be needed to obtain statistical power at .80 level.

In Fig. 2 we report the mean values of gesture rate for each group (monolingual or bilingual) in each language (Italian and English).


Figure 2: Mean values of gesture rate for Language Status (Bilingual, dotted line or Monolingual, solid line) in each Language (English and Italian).

## Gesture Salience

A generalized mixed linear model was performed on gesture salience (sample size $n=390$ ). Following the same procedure described for gesture rate, we set for the model that had by items (cartoons) random intercepts, by subjects random intercepts and Language (English or Italian) varying by subjects random slopes. We found a significant effect for both Language ( $\mathrm{Est}=1.85$, S.E. $=0.38$, p $<0.001$ ) and Language Status (Est. $=0.98$,S.E. $=0.39, \mathrm{p}=0.01$ ). Interaction between the fixed effects (Language status and Language) was investigated but not found (Est=0.33, S.E. $=0.76$, $\mathrm{p}=0.66$ ). That is, gestures were more salient in Italian than in English and bilinguals' gestures were more salient than monolinguals'.
p values were automatically calculated from z scores by glmer function. In Fig. 3 we report the probability of producing salient gestures in each Language Status (bilingual and monolingual) and each Language (Italian and English). A post hoc power analysis conducted with data simulation ( n simulations=1000) revealed that 22 participants per group (monolingual English, monolingual Italian, Bilinguals= 66 participants overall) would be needed to obtain statistical power at .80 level.


Figure 3: Probability of producing a salient gesture by Language Status (Bilingual, dotted line or Monolingual, solid line) in each Language (English and Italian).

## Discussion

The aims of this study were to investigate gesture frequency and gesture space in Italian/English bilinguals and the relationship between gesture and language in bilinguals. In addressing this question, the data from a bilingual and two monolingual control groups have been collected and analysed.

A first result was that Italian speakers gestured more frequently and that their gestures were more salient than English speakers. As observed by Kendon (1992) and Efron (1972) Italian is indeed a "high gesture culture". With regard to gesture rate, we found no evidence of transfer when bilinguals switch between Italian and English. With regard to salience, we found, again, no evidence of transfer but, overall, bilinguals' gestures were more salient with respect to the gestures performed by the two control groups. From our results we can conclude that when English/Italian bilinguals switch language, their gesture parameters switch accordingly with the language they talk.

Whether or not one finds gestural transfer in bilinguals may depend upon many variables. First of all, the societal context for bilingualism and the bilingualism level of the participants can affect transfer. Unlike the current study, So (2010) found evidence of transfer for representational gestures only from American English (high gesture rate) to Mandarin-Chinese (lower gesture rate) in Singapore. In Singapore multilingualism is a long established and prominent feature of the society, encouraged by laws. The bilinguals who took part in the present study mostly grew up in non bilingual communities (in Italy or the UK) where one of the two languages was mostly spoken with parents, family members and friends. Although bilingual participants
in this study reported in the Linguistic Background questionnaire that to them it was important to speak well both languages and they equally liked to speak in both, it might be that it is easier for the bilinguals tested in our study to "keep apart" the two linguistic systems.

In contrast, bilinguals in Singapore might have been much more exposed to two or more languages in daily life and it has been documented that transfer of words occurred together with gesture frequency transfer (So, 2010). The bilinguals in this study had some tip-of-the-tongue phenomena but did always choose to talk in the target language. The societal and linguistic context may account for the lack of gestural transfer found in Nicoladis et al. (2005) and for the lack of difference in gesture frequency between the two monolingual control groups. Their EnglishFrench bilingual children were recruited in Alberta, which is an English speaking province of Canada. Thus, just like our English-Italian bilinguals, one of the two languages (French) was mostly spoken with parents, family members and friends. Differently, the French monolingual group was recruited in Quebec, a bilingual area of Canada where French Canadians are highly exposed to English too.

One of the most interesting findings of this study is that bilinguals' gestures were overall more salient than monolinguals' gestures. One possible explanation is that bilinguals may often be in a communicative situation where some people are weak in Italian and others are weak in English. In such situations, bilinguals may make their gestures more salient in order to facilitate communication. This might become a habitual feature of bilinguals' gestures. This speculation though needs to be substantiated by future studies.
Our results indicate that language and gesture, even gesture "cultural" parameters such as frequency and salience, are tightly linked. In addition to that, our results suggest that the selection of those parameters happens at a pre-linguistic level, as these parameters have no strictly communicative meaning. The features specifying a language and the corresponding gesture parameters might be selected at a high level processing stage in which verbal and nonverbal aspects of communication are planned together. This is compatible with the idea that bilinguals specify the language at a conceptual level, as suggested by La Heij's concept selection hypothesis (2005). La Heij stated that the semantic system directly activates target-language lexical nodes over lexical nodes in the non-target language. Thus, the intended language is selected at the conceptual stage after a series of communicative aspects have been taken into account (e.g. who is the interlocutor, in which communicative situation we are etc.).

## Acknowledgments

Dr. Cavicchio gratefully acknowledges the financial support of the European Union and the Provincia Autonoma di Trento (7th Framework Program "People", Marie Curie Action - COFUND-TRENTINO, project CoByGest).

## References

Barzini, L. (1964). The Italians. New York: Atheneum.
Chu, M., Meyer, A., Foulkes, L., \& Kita, S. (under review). Individual differences in frequency and saliency of speech-accompanying gestures: The role of cognitive abilities and empathy. Journal of Experimental Psychology: General.
Efron, D. (1972). Gesture, race, and culture. The Hague: Mouton. (Original work published as Gesture and environment, 1941.) Graham, J. A., \& Argyle, M. (1975). A cross-cultural study of the communication of extraverbal meaning by gestures. International Journal of Psychology, 10, 57-67.
Gullberg, M., \& Indefrey, P. (2003). Language Background Questionnaire. Developed in The Dynamics of Multilingual Processing. Nijmegen, Max Planck Institute forPsycholinguistics.http://www.mpi.nl/departments/other -research/research-projects/the-dynamics-of-multilingual-processing/toolshttp://www.mpi.nl/departments/other-research/research-projects/the-dynamics-of-multilingualprocessing/tools
Loonstra A. S., Tarlow, A.R., \& Sellers, A. H. (2001). COWAT metanorms across age, education, and gender. Applied Neuropsychology, 8, 3, 161-166.
Kendon, A. (1992). Some recent work from Italy on quotable gestures ('emblems'). Journal of Linguistic Anthropology, 21, 72-93.
Kendon, A. (1995). Gestures as illocutionary and discourse structure markers in southern Italian conversation. Journal of Pragmatics, 23, 247-279.
Kita, S. (2000). How representational gestures help speaking. In D. McNeill (Ed.), Language and gesture (pp. 162-185). Cambridge: Cambridge University Press.
Krauss, R. M., \& Hadar, U. (1999). The role of speechrelated arm/hand gestures in word retrieval. In R. Campbell \& L. Messing (Eds.), Gesture, speech, and sign (pp. 93-116). Oxford: Oxford University Press.
La Heij, W. (2005). Selection processes in monolingual and bilingual lexical access. In J. F. Kroll \& A. M. B. De Groot (eds.), Handbook of bilingualism: Psycholinguistic approaches, (pp. 289-307). Oxford: Oxford University Press.
Levy, R. (2012) Probabilistic Models in the Study of Language.http://idiom.ucsd.edu/~rlevy/pmsl_textbook/ch apters/pmsl_8.pdf
McNeill, D. (1992). Hand and mind: What gestures reveal about thought. Chicago: Chicago University Press
Müller, C. (1998). Redebegleitende Gesten: Kulturegeschichte -Theorie -Sprachvergleich. Berlin: Berlin Verlag.
Nicoladis, E., Mayberry, R., \& Genesee, F. (1999). Gesture and early bilingual development. Developmental Psychology, 35, 514-526.
Nicoladis, E. (2007). The effect of bilingualism on the use of manual gestures. Applied Psycholinguistics, 28, 441454
Özyürek, A., Kita, S., \& Allen, S. (2001). Tomato Man
movies: Stimulus kit designed to elicit manner, path and causal constructions in motion events with regard to speech and gestures. Nijmegen, The Netherlands: Max Planck Institute for Psycholinguistics, Language and Cognition group.
Pika, S., Nicoladis, E., \& Marentette, P. (2006). A crosscultural study on the use of gestures: Evidence for crosslinguistic transfer? Bilingualism: Language and Cognition, 9, 319-327.
Sherman, J.,\& Nicoladis, E. (2004). Gestures by advanced Spanish-English second-language learners. Gesture, 4, 143-156.
So, W. C. (2010). Cross-cultural transfer in gesture frequency in Chinese-English bilinguals. Language and Cognitive Processes, 25, 10, 1335-1353.

# Modelling Graded Semantic Effects in Lexical Decision 

Ya-Ning Chang (ya-ning.chang@manchester.ac.uk)<br>Neuroscience and Aphasia Research Unit (NARU), University of Manchester, Brunswick Street, Manchester, M13 9PL, UK<br>Matthew Lambon Ralph (matt.lambon-ralph@manchester.ac.uk)<br>Neuroscience and Aphasia Research Unit (NARU), University of Manchester, Brunswick Street, Manchester, M13 9PL, UK<br>Steve Furber (steve.furber@manchester.ac.uk)<br>Advanced Processor Technologies Group, University of Manchester, Oxford Road, Manchester, M13 9PL, UK<br>Stephen Welbourne (stephen.welbourne@manchester.ac.uk)<br>Neuroscience and Aphasia Research Unit (NARU), University of Manchester, Brunswick Street, Manchester, M13 9PL, UK


#### Abstract

Recent studies have shown that the involvement of semantic information in visual lexical decision depends on the nature of nonword foils with semantic effects increased as nonwords become more word-like (Evans, Lambon Ralph \&Woollams, 2012). Given that most models of lexical decision focus on orthographic information (Coltheart, Rastle, Perry, Langdon \& Ziegler, 2001; Grainger \& Jacobs, 1996; Seidenberg \& McClelland, 1989), the role of semantics and its interactions with vision, orthography, and phonology has been overlooked. We developed a recurrent connectionist model of single word reading including visual, orthographic, phonological, and semantic processing. The model differentiated words from nonwords by integrating measures of polarity across four key processing layers. The contribution of semantics depended on the type of nonword foils. The model was more reliant on semantic information when the nonword foils were pseudowords and pseudohomophones rather than consonant strings. The results support the view that semantic involvement in lexical decision is graded by the difficulty of the decision task.


Keywords: semantic effects; lexical decision; reading; computational modelling; visual word recognition.

## Introduction

Lexical decision (LD) has been widely used to study the cognitive processes involved in visual word recognition. Subjects are asked to judge whether a letter string is a word or not. Measures of accuracy and response time are thought to reflect the differences in lexical-semantic processing of words and nonwords. There seems to be consistent evidence that vision, orthography and phonology play roles in visual lexical decision (Coltheart, Davelaar, Jonasson, \& Besner, 1977; Grainger \& Jacobs, 1996; Meyer, Schvanev, \& Ruddy, 1974), however the extent of the involvement of semantics in lexical decision remains debateable (James, 1975; Joordens \& Becker, 1997; Lupker \& Pexman, 2010). James (1975) showed a reliable concreteness effect during lexical decision when using pseudoword and pseudohomophone foils, while the effect disappeared when testing with consonant strings. He suggested subjects might
be able to exploit semantic information to support efficient LD. Although some subsequent studies have found reliable semantic influences on lexical decision under different foil conditions (Joordens \& Becker, 1997), others have failed to find such effect (Lupker \& Pexman, 2010). Evans, Lambon Ralph and Woollams (2012) demonstrated that semantic involvement in lexical decision was graded by the difficulty of the decision task as indexed by the word-likeness of the foil. There were stronger semantic effects with pseudohomophones than with pseudowords, and the effects were stronger with pseudowords than with consonant strings. Apart from the behavioural data, there is also evidence of semantic involvement in lexical decision from neuroimaging studies. Woollams, Silani, Okada, Patterson and Price (2011) revealed that left anterior temporal activation, increased for atypical relative to typical strings when lexical decisions were made more difficult in the context of pseudohomophone foils. The left anterior temporal lobe has been considered as a region for combining various types of sensory and motor information to form amodal semantic representations (Patterson, Nestor, \& Rogers, 2007). The orthographic typicality effect in the left anterior temporal lobe has also been found in a previous electrophysiological (EEG) study. In a speeded lexical decision task, atypical words were found to elicit stronger source currents than did typical words at around 160 msec in the left anterior temporal lobe (Hauk, Patterson, Woollams, et al., 2006). These effects are consistent with what has been observed in the neuropsychological studies of patients with semantic dementia (SD), who have asymmetrically bilateral atrophy degeneration of the anterior temporal lobes. These patients show a progressive degeneration of semantic knowledge (Hodges, Patterson, Oxbury, \& Funnell, 1992). When patients are asked to perform two-alternative forced-choice visual lexical decision, they can correctly choose orthographically typical words from the relatively atypical nonwords but have difficulty in the reverse condition (Rogers, Lambon Ralph, Hodges, \& Patterson, 2004). Taken together, this evidence supports the view that semantic processing is involved in
lexical decision in particular when the words are orthographically atypical and the foils are pseudohomophones.

## Models Based on Localist Views

In the literature, several theories of visual word recognition have been proposed to explain the underlying mechanisms of lexical decision (Coltheart, et al., 1977; Coltheart, et al., 2001; Grainger \& Jacobs, 1996; Plaut, 1997; Seidenberg \& McClelland, 1989). Some researchers argue that lexical decision relies upon the orthographic lexicon (Coltheart, et al., 1977). If there is a match, subjects would give a positive response, otherwise, the negative response is made. On this view, the locus of lexical decision is based on activation within the orthographic lexicon. The involvement of phonology is a relatively late process after the mental lexicon search while the semantic system is generally not involved in the recognition processes unless the discrimination becomes extremely difficult (Coltheart, et al., 2001). This orthographically based approach is shared with Grainger and Jacobs (1996), who developed a computational model of lexical decision. In their multiple read-out model (MROM), a word response could be made either when the particular word unit activation reached a local criterion, $M$, or the overall activity in the word layer reached a global criterion, $\Sigma$, before the temporal deadline as $T$. The RT was based on the earliest moment where either of criteria was met. If neither of the activation criteria was met, a nonword response was given and the RT was the value of the deadline criterion. Grainger and Jacobs (1996) assumed that the $M$ criterion should be fixed as a normal recognition level and was set corresponding to individual word units. While the global criterion $\Sigma$ and the temporal deadline $T$ would vary according to the lexical frequency status of the stimulus. The higher probability the stimulus was a word, the lower global criterion and the longer temporal deadline were used. By this, the MROM model was able to simulate several standard effects seen in lexical decision including the frequency effects, the orthographic neighbourhood size effects, and their interactions (Grainger and Jacobs, 1996). Other models of visual word recognition such as the dual-route cascaded (DRC) model (Coltheart, et al., 2001) and the connectionist dual process (CDP+) model (Perry, Ziegler, \& Zorzi, 2007) share similar decision mechanisms to the MROM model.

## Models Based on Distributed Views

An alternative theory of visual word recognition argues that there is no mental lexicon for the store of word knowledge in the recognition system (Dilkina, McClelland, \& Plaut, 2010; Plaut, 1997; Seidenberg \& McClelland, 1989). On this view, the decision can be made on the basis of the differential activations elicited by familiar words and unfamiliar nonwords. When presenting a word, strong activations are expected because the mappings between the visual or orthographic representation of the word and its phonological and semantic representations have been
learned. Conversely, relatively weaker activations would be expected for a nonword representation as it is a novel stimulus. One important model of lexical decision was developed by Plaut in 1997, who proposed that the measure of how strongly units were activated, called stress or polarity, could be used as a basis for making lexical decisions. He built a feedforward model which consisted of orthographic, phonological and semantic components and demonstrated that words tended to produce higher stress than nonwords at the semantic layer. With the proper decision criteria, over 95 percent of words in the training corpus could be discriminated from nonwords. In addition, the network tended to produce higher semantic stress for pseudohomophones than for pseudowords in line with the behavioural data.

## Accumulated Information for Lexical Decision

There are also other models which have emphasised the use of accumulated information within the system for making decisions. One of these is the diffusion model, developed by Ratcliff, Gomez and Mckoon (2004). The central idea of the diffusion model was that the speed (drift rate) at which information was accumulated over time was affected by the lexical status of the stimuli. They hypothesized that the drift rate had a positive correlation with a measure of how wordlike a stimulus was. In their model, the decision was then made when a random walk process driven by the drift rate reached either a word criterion or nonword criterion. Another model is the Bayesian reader model developed by Norris (2009). The basic premise of this model was to assume subjects would consistently compute the probability of the stimulus being a word or a nonword on the basis of its lexical status. In the simulations conducted in Norris (2009), the recognition of a letter string being a word was made on the basis of the sum of the probabilities of all possible letter strings and this value was expected to be 1.0. Therefore, the nonword likelihood could be computed simply by using 1 minus summed probability of letter strings corresponding to words.

In summary, data from behavioural, neuroimaging and patient studies, all point to the involvement of semantic processing in lexical decision. Previous models either postulate an exclusive role for semantics (Dilkina, McClelland, \& Plaut, 2010; Plaut, 1997) or no role for semantics (Coltheart et al. 2001; Grainger \& Jacobs, 1996; Norris, 2009). Importantly none of the previous models would be able to account for the data from Evans et al. (2012), which indicates that the degree of semantic involvement is flexible and can be modulated by the nature of the nonwords foils. The goal of this paper was to use a novel model of reading to explore to what extent semantics is involved in lexical decision and how it interacts with other processing layers. In addition we aimed to be able to simulate the data from Evans et al. illustrating how changes to the nature of the nonwords foils can bias lexical decision tasks. Based on earlier work (Chang, Furber, \& Welbourne, 2012a), we developed a fully implemented recurrent model
of visual word recognition. The model included a visual processing stage along with the orthographic, phonological and semantic processing stages. Importantly, the orthographic representations were allowed to learn during the training.

## Method

## Network Architecture

The architecture of the model is shown in Figure 1. The model had two separate pathways for recognising words from visual input: a phonological pathway and a semantic pathway. The H0 layer was functionally responsible for visual processing while the OH layer was equivalent to the orthographic layer in the triangle model except that the orthographic representations were learned through the course of training rather than being supplied as inputs. This mimics the situation in human development where orthographic representations emerge to support reading acquisition in children. The word recognition process started from the visual input layer and moved progressively to the orthographic layer, and then progressed in separate pathways to the phonological and semantic layers. The phonological component consisted of 61 phonological units which were all connected to a set of 20 clean up units. These clean up units projected back onto the phonological units, forming an attractor. Similarly, the semantic component consisted of 200 semantic units. These units were all connected to another set of 80 clean up units, which projected back onto the semantic units. The context component consisted of 3 units, which were used to provide additional contextual information for discriminating between homophones. The numbers of hidden units for each layer were determined by pilot trials to ensure the model
was trainable and that the performance of the model was good on the production, comprehension and reading tasks.

There were also control units for each layer except input and output layers. These acted to flexibly inhibit the activation of the layer they were connected to. The control units were important because they allowed the model learn to manage its own temporal dynamics. In particular they allowed the units at the latter layers to be suppressed until the input to them had had time to ramp up to values that reflected the influence of the visual input to the model.

The training corpus consisted of 2,971 words. The visual representations used here were adapted from those used in Chang et al.'s (2012a) study. The network was trained on 12-point lower case words in Arial font. Each word was positioned with its vowel aligned on a fixed slot of the image. Ten slots were used in all and the size of each slot was $16 \times 16$ pixels. The scheme of phonological representations was the same as that used in the Plaut et al.'s (1996) model. The context units were used to differentiate the meanings of homophones, which have same pronunciations but different meanings. For those pronunciations with only one possible word meaning, the context units were all set to zero. For other pronunciations corresponding to more than one word meanings, the context units were all set to 0 for the first meaning; and one of the context units from right to left was set to 1 to represent the second, third and fourth meaning accordingly. The semantic representations were generated using the same scheme as in Chang, Furber, and Welbourne (2012b). The meaning of each word was represented by a 200 -dimensional semantic vector. Each vector had 5 active units in the first half of the vector converted from the top positive attributes and 15 active units in the second half of the vector converted from the top negative attributes.


Figure 1. The architecture of the model. The dashed lines indicate inhibitory connections.

## Training Procedure

The training was separated into two phases. In phase 1 only the phonology-semantics mappings were trained while in phase 2 the full reading model was trained starting from the trained weights obtained in phase 1. In phase 1, the phonology-semantics model was first subdivided into two parts: the production model learning the mappings from semantics to phonology and context, and the comprehension model learning the mappings from phonology and context to semantics. The production and comprehension model were trained separately. The presentation of each example lasted for 6 intervals of time and each interval of time was divided into 3 ticks. In each presentation, the input pattern of a word was clamped onto the input units for the full 6 intervals of time and the task was to produce the correct target representation. For the last 2 intervals, the activations of output units were compared to their targets. Error score, the difference between the units' outputs and their targets, was used to calculate weight changes. No error was recorded if the output unit's activation and target were within 0.1 of each other. At the end of phase 1 the accuracy rates of the production and comprehension model were $99.97 \%$ and $99.43 \%$ for the phonological level and semantic level respectively.

In phase 2, the weights obtained from the end of training the phonology-semantics model were embedded and frozen into the full reading model. The weight connections from the visual layer to both phonological and semantic layers were updated through training. There were local control units for each layer except input and output layers. The initial output of each control unit was set to 1 . The weight connections from its previous layer to each control unit were free to be updated. The weight connections from each control unit to those units that it was controlling were trainable, but the values were limited to between -4 and 0 . The negative boundaries used here were to ensure that the control unit acted to inhibit activation. The model was allowed to update for 30 ticks of time. The visual representation of a word was presented at the input units for all 30 ticks. The task was to produce correct phonological and semantic patterns. For the last 2 intervals, the output units were compared with their corresponding phonological or semantic targets and errors were computed. To encourage more accurate learning, no error was computed when the output unit's activation and target were within 0.001 .The model was trained to produce $99.3 \%$ correct phonological and $97.4 \%$ correct semantic patterns in the word reading task.

## Polarity Measures and Decision Criteria

Plaut (1997) proposed that parallel distributed models can perform the lexical decision task based on the measure of polarity, which is whether the units in the model have learned to adopt a binary representation. To capture this phenomenon, Plaut (1997) introduced a formula to compute the index of unit binarization which was termed unit polarity
as follows:

$$
y=x \cdot \log _{2}(x)+(1-x) \cdot \log _{2}(1-x)+1
$$

where $x$ is the unit activation ranging from 0 to $1 ; \log _{2}(\cdot)$ is the logarithmic function with the base of $2 ; y$ is the polarity measure. When known words are presented, the units tend to become binary, leading to high polarity values. However, when nonwords are presented, the activities of the units tend to be low and closer to 0.5 , resulting in generally low polarities. Two criteria were used for the model to make word-nonword decisions: (1) word boundary: the 3 standard deviation line above the average nonword polarity; (2) nonword boundary: the 3 standard deviation line below the average word polarity. The polarity for an item was computed by combining the measures of polarity for that item at the H 0 (visual processing), OH (orthographic processing), phonological, and semantic layers. If an item polarity crossed over the word boundary the item was classified as a word. By contrast, if the item polarity crossed over the nonword boundary, the item was determined as a nonword. There were, however, a few item polarities that remained between the two boundaries. In this case, responses were made based on which boundary the polarity was closest to at the last time tick. The response time was the time tick when an item polarity first crossed over either word or nonword boundary. In the situation where neither boundary was crossed the response time was taken as 30 ticks.

## Inverse Efficiency

To control for potential differences in speed-accuracy tradeoff caused by the arbitrary selection of standard deviation lines, we adopted a measure of inverse efficiency, which is considered to be a corrected reaction time (Roder, Kusmierek, Spence, \& Schicke, 2007). Inverse efficiency is a combination of both reaction and accuracy (i.e., dividing reaction time by accuracy). The lower the score, the more efficiently the model performed the task.

## Results

## Semantic influences on lexical decision

Evans et al. (2012) suggested that the subjects needed to access semantic information in the lexical decision task particularly when words were tested with more word-like nonwords such as pseudowords and pseudohomophones. They showed a graded imageability effect in lexical decision depending on the difficulty of the task. The imageability effect was larger when words were tested against with pseudohomophones than with pseudowords. The effect disappeared in the context of consonant strings. We tested the model to see whether it could produce the similar pattern as seen in Evans et al.'s data. After removing those words which were not in the training exemplars and their matched nonword items, there were 70 words, consisting of 35 high- and 35 low-imageability words. Their matched nonword pairs for the three different foil conditions, consonant string (CS), pseudoword (PW), and
pseudohomophone ( PH ) were also used in the current test. To compare with Evans et al (2012)'s data, the scores of inverse efficiency were normalised by the value obtained from the low imageability pseudohomophone condition. The same procedure was applied to Evans et al.'s (2012) data. The results are shown in Figure 2. It is clear that the simulation results (Figure 2, left) follow the pattern of Evans et al.'s data (Figure 2, right). A $2 \times 3$ repeated measures ANOVA was conducted with imageability (High/Low) and foil condition (CS/PW/PH) as within subject factors and the scores of inverse efficiency were used as a dependent variable. There was a reliable main effect of imageability, $F(1,19)=9.88, \mathrm{p}<.01$. The main effect of foil condition was also significant, $\mathrm{F}(1.31,24.85)=59.75$, $\mathrm{p}<.001$ (with a Greenhouse-Geisser adjustment).

Importantly, there was a significant interaction between imageability and foil condition, $\mathrm{F}(2,38)=3.60$, $\mathrm{p}<.05$, showing that the size of imageability effect increased along with the word-likeness of the foils. Note that we also ran the statistical tests on the unnormalised scores with the same pattern of results. This is what would be expected based on Evans et al.'s (2012) data. The post-hoc analyses showed that the imageability effect was not significant with consonant strings ( $p>.05$ ) while there were significant imageability effects in the contexts of pseudowords, $\mathrm{F}(1$, $19)=6.76, \mathrm{p}<.05$, and pseudohomophones, $\mathrm{F}(1,19)=15.06$, $\mathrm{p}<.01$. The results were consistent with the findings in Evans et al.’s (2012) study, suggesting semantic effects vary in lexical decision, depending on the foil type.



Figure 2. Data are from simulation (Left) and from Evans et al. (2012). Normalised scores were computed by equating two results based on the low imageability pseudohomophone condition.

## General Discussion

The primary aim of this paper was to develop a large-scale recurrent reading model containing visual, orthographic, phonological, and semantic processing to support lexical decision tasks. The model was used to explore the involvement of semantics in lexical decision with other processing components implemented in the system. This approach is different to most existing models of lexical processing which have focused on activity within a single processing layer. Based on the measure of polarities at four core processing layers ( $\mathrm{H} 0, \mathrm{OH}$, phonology and semantics), the model was able to perform the lexical decision tasks and account for the graded semantic effects found by Evans et al. (2012), as shown in Figure 2. The magnitude of semantic effects increased as nonwords became more word-like, where the semantic effect was stronger with pseudohomophones than with pseudowords and then with consonant strings. This provides evidence supporting the distributed view of lexical decision which proposes that semantic access is important and automatic in lexical decision (Plaut, 1997). The actual use of semantic information is flexible and is largely dependent on the
difficulty of the tasks (Evans, et al., 2012). That is in contrast with the localist view arguing for no or little involvement of semantics in lexical decision (Coltheart, et al., 2001).

There are some existing lexical decision models developed on the basis of the localist view of lexical decision including the MROM model (Grainger \& Jacobs, 1996) and the DRC model (Coltheart, et al., 2001) and the CDP+ model (Perry, et al., 2007). These models can simulate several effects in lexical decision and the strategic influences on lexical decision by flexibly adjusting decision criteria. However, their results are almost all based on orthographic processing with little attention to other processing components in particular the semantic system. Thus the questions as to how these models implement the involvement of semantics in lexical decision, which presumably requires some feedback connections from semantics to their orthographic lexicon (Coltheart, et al., 2001) remain unclear. In particular, these localist models would find it difficult to account for the graded changes in the involvement of semantics depending on foil type. In the current model this graded effect emerges naturally as a consequence of increasing task difficulty.

In this paper we have followed Evans et al. (2012) by talking the size of the imageability effect as an index of semantic involvement, but future work could extend this in the model by developing additional metrics to quantify the involvement of semantics including a direct comparison of performance with and without the contribution from the semantic layer.

To summarise, this paper uses a model of human visual word recognition to explore the role of semantics in lexical decision. Crucially, the model was able to account for the graded semantic influences on lexical decision corresponding to the various types of foils, providing evidence for semantic influences on lexical decision.

## Acknowledgments

This research was supported by grants under the Cognitive Foresight Initiative (jointly funded by EPSRC, MRC and BBSRC - EP/F03430X/1) and the Neuroscience Research Institute at the University of Manchester.

## References

Chang, Y. N., Furber, S., \& Welbourne, S. (2012a). "Serial" Effects in Parallel Models of Reading. Cognitive Psychology, 64(4), 267-291.
Chang, Y. N., Furber, S., \& Welbourne, S. (2012b). Generating Realistic Codes for Use in Neural Network Models. In N. Miyake, D. Peebles, \& R. P. Cooper (Eds.), Proceedings of the 34th Annual Confrerence of the Cognitive Science Society (pp. 198-203). Austin, Tx: Cognitive Science Society.
Coltheart, M., Davelaar, E., Jonasson, J. T., \& Besner, D. (1977). Access to the Internal Lexicon. In S. Dornic (Ed.), Attention and Performance Vi. Hillsdale, NJ: Lawrence Erlbaum Associates.
Coltheart, M., Rastle, K., Perry, C., Langdon, R., \& Ziegler, J. (2001). Drc: A Dual Route Cascaded Model of Visual Word Recognition and Reading Aloud. Psychological Review, 108(1), 204-256.
Dilkina, K., McClelland, J. L., \& Plaut, D. C. (2010). Are There Mental Lexicons? The Role of Semantics in Lexical Decision. Brain Research, 1365, 66-81.
Evans, G. A. L., Lambon Ralph, M. A., \&Woollams, A. M. (2012). What's in a Word? A Parametric Study of Semantic Influences on Visual Word Recognition. Psychonomic Bulletin \& Review, 19(2), 325-331.
Grainger, J., \& Jacobs, A. M. (1996). Orthographic Processing in Visual Word Recognition: A Multiple Readout Model. Psychological Review, 103(3), 518-565.
Hauk, O., Patterson, K., Woollams, A., Watling, L., Pulvermuller, F., \& Rogers, T. T. (2006). Q: When Would You Prefer a Sossage to a Sausage? A: At About 100 Msec. Erp Correlates of Orthographic Typicality and Lexicality in Written Word Recognition. Journal of Cognitive Neuroscience, 18(5), 818-832.
Hodges, J. R., Patterson, K., Oxbury, S., \& Funnell, E. (1992). Semantic Dementia - Progressive Fluent Aphasia with Temporal-Lobe Atrophy. Brain, 115, 1783-1806.

James, C. T. (1975). Role of Semantic Information in Lexical Decisions. Journal of Experimental PsychologyHuman Perception and Performance, 104(2), 130-136.
Joordens, S., \& Becker, S. (1997). The Long and Short of Semantic Priming Effects in Lexical Decision. Journal of Experimental Psychology-Learning Memory and Cognition, 23(5), 1083-1105.
Lupker, S. J., \& Pexman, P. M. (2010). Making Things Difficult in Lexical Decision: The Impact of Pseudohomophones and Transposed-Letter Nonwords on Frequency and Semantic Priming Effects. Journal of Experimental Psychology-Learning Memory and Cognition, 36(5), 1267-1289.
Meyer, D. E., Schvanev, R. W., \& Ruddy, M. G. (1974). Functions of Graphemic and Phonemic Codes in Visual Word-Recognition. Memory \& Cognition, 2(2), 309-321.
Norris, D. (2009). Putting It All Together: A Unified Account of Word Recognition and Reaction-Time Distributions. Psychological Review, 116(1), 207-219.
Patterson, K., Nestor, P. J., \& Rogers, T. T. (2007). Where Do You Know What You Know? The Representation of Semantic Knowledge in the Human Brain. Nature Reviews Neuroscience, 8(12), 976-987.
Perry, C., Ziegler, J. C., \& Zorzi, M. (2007). Nested Incremental Modeling in the Development of Computational Theories: The Cdp+ Model of Reading Aloud. Psychological Review, 114(2), 273-315.
Plaut, D. C. (1997). Structure and Function in the Lexical System: Insights from Distributed Models of Word Reading and Lexical Decision. Language and Cognitive Processes, 12(5-6), 765-805.
Plaut, D. C., McClelland, J. L., Seidenberg, M. S., \& Patterson, K. (1996). Understanding Normal and Impaired Word Reading: Computational Principles in QuasiRegular Domains. Psychological Review, 103(1), 56-115.
Ratcliff, R., Gomez, P., \& McKoon, G. (2004). A Diffusion Model Account of the Lexical Decision Task. Psychological Review, 111(1), 159-182.
Roder, B., Kusmierek, A., Spence, C., \& Schicke, T. (2007). Developmental Vision Determines the Reference Frame for the Multisensory Control of Action. Proceedings of the National Academy of Sciences of the United States of America, 104(11), 4753-4758.
Rogers, T. T., Lambon Ralph, M. A., Hodges, J. R., \& Patterson, K. (2004). Natural Selection: The Impactof Semantic Impairment on Lexical and Object Decision. Cognitive Neuropsychology, 21(2-4), 331-352.
Seidenberg, M. S., \& McClelland, J. L. (1989). A Distributed, Developmental Model of Word Recognition and Naming. Psychological Review, 96(4), 523-568.
Woollams, A. M., Silani, G., Okada, K., Patterson, K., \& Price, C. J. (2011). Word or Word-Like? Dissociating Orthographic Typicality from Lexicality in the Left Occipito-Temporal Cortex. Journal of Cognitive Neuroscience, 23(4), 992-1002.

# Eyetracking as an Implicit Measure of Category-Based Induction 

Stephanie Y. Chen (syc341@nyu.edu)<br>Department of Psychology, 6 Washington Place<br>New York, NY 10003 USA

Brian H. Ross (bhross@illinois.edu)<br>Department of Psychology, 603 E. Daniel St. Champaign, IL 61820 USA

Gregory L. Murphy (gregory.murphy@nyu.edu)<br>Department of Psychology, 6 Washington Place<br>New York, NY 10003 USA


#### Abstract

Category information is used to predict unknown properties of category members. Previous research has found that when categorization is uncertain, property predictions do not reflect integration of information across categories as normative principles and Bayesian models would suggest. Rather, people often base their predictions on only the most likely category and disregard information from less likely ones Research in category-based induction tends to elicit explicit, verbal responses which may not readily allow for integration of information across categories. This paper explores whether changing response mode can promote more normative use of category information in induction. Experiment 1 used an implicit measure of prediction: eye movements. The results suggest that when making predictions implicitly people integrate information across categories. The results of Experiment 2 suggest that the integration of information found in Experiment 1 were not a result of explicit strategies.


Keywords: category-based induction; reasoning; implicit processes.

## Introduction

The ability to use category-level information to infer information about novel objects aids our reasoning, social interactions, communication and predictions. By placing an object into a category, we can make predictions about it even though we have never encountered that particular object before. Because you know about the category of Chinese food in general, when you see some Chinese food cartons in your refrigerator you know that there is some chance that the food is spicy, but it's likely not. For our purposes, category-based induction refers to a process like the one described above (the extension of category information to a new item in that category). This process becomes more complicated when you are unsure what category an item belongs to. Imagine that your roommate has left unmarked cartons of leftover food in the refrigerator, and you can't tell whether they hold bland Chinese or spicy Indian food. Do you take an acid reducer before eating? You must make a prediction about the food's spiciness based on the characteristics you can observe.

To decide whether the food will be spicy, you should take into account both the possibility that it is Chinese food and the possibility that it is Indian food. This type of reasoning
is consistent with Bayesian approaches to classification and prediction in which people weight different possibilities by their prior likelihoods. Anderson (1991) proposed such a model of category-based induction ${ }^{1}$ in which the probability that an object with observed features, $F$, has an unobserved feature, $j$, is the weighted sum of the probabilities across all categories, $k$ (assuming they are mutually exclusive):

$$
\begin{equation*}
P(j \mid F)=\sum_{k} P(k \mid F) \times P(j \mid k) . \tag{1}
\end{equation*}
$$

Thus, if you were a Bayesian food thief you would take the probability that the unknown food is Chinese food and multiply that by the probability that Chinese food is spicy. Next you would take the probability that the food is Indian food and multiply that by the probability that Indian food is spicy. The sum of the two products is the probability that the food is spicy. This appears normatively correct, since it takes into account your uncertainty and weighs the strength of the prediction accordingly. If very certain that the food is Chinese food you should make a moderate prediction about the likelihood of it being spicy; if uncertain, you should make a stronger prediction. Surprisingly, however, previous research on induction with uncertain categories has provided evidence using both real-life and artificial categories that people usually base their induction on only a single category (Hayes \& Newell, 2009; Malt, Ross, \& Murphy, 1995; Murphy, Chen, \& Ross, 2012; Murphy \& Ross, 1994).

These findings are in contrast to those of perception and motor control research that often find that people integrate information across possibilities in a Bayesian manner (Kersten, Mamassian, \& Yuille, 2004; Tassinari, Hudson, \& Landy, 2006; Trommershäuser, Landy, \& Maloney, 2006; Trommershäuser, Maloney, \& Landy, 2008). In perception, Bayesian models are used to explain how the visual system takes ambiguous inputs and returns percepts that are most

[^38]likely. People use knowledge about prior probabilities of states of the world and the likelihood of each state given the visual stimulus to arrive at the most probable interpretation of the stimulus (Kersten et al., 2004). In motor control, one action may be best suited to achieve a goal, given the state of the world. But since perception is not perfect, the state of the world is uncertain. Models of action propose that people integrate information about the likelihood of the possible states of the world to make near optimal actions (Haruno, 2001). These actions are sensitive to the payoff structure of the task: Subjects make motor decisions that minimize costs, given the uncertainty of different motor outcomes and the costs and benefits associated with each action (Trommershäuser, Landy, \& Maloney, 2006; Trommershäuser, Maloney, \& Landy, 2008).

Why might people be unable, or unwilling, to combine information about two categories in category-based induction tasks, but are able to integrate across possibilities and weigh costs and benefits in seemingly more complex perception-action tasks? We suggest that this discrepancy can, in part, be explained by the distinction between implicit and explicit processes (Sloman, 1996). Explicit processes are conscious and rule-based, while implicit processes are unconscious and associative. Explicit reasoning is subject to a reasoning heuristic called the singularity principle, which states that people generally only consider one possibility at a time (Evans, 2007). More specifically, we suggest that response mode is critical to whether information is integrated across categories. In category-based induction tasks, subjects often explicitly report what category they think an item belongs to prior to making a prediction. In contrast, perception and motor control experiments tend to depend on implicit responses. Subjects in these experiments are not asked to explicitly consider the potential possibilities (states of the world) but are instead prompted to act on this information (often, but not always, with a motor response).

Chen, Ross, \& Murphy (in press) provided evidence that implicit and explicit responding lead to different use of category information during induction. In one experiment, subjects learned artificial categories of moving geometric figures defined by two features: shape and direction. At test, subjects were presented with a shape and asked to predict its direction either implicitly or explicitly. The implicit test was a novel, game-like motor task that elicited a speeded prediction, and the explicit test was a formally identical verbal task that elicited a conscious, unspeeded prediction.

The categories consisted of eight moving geometric figures (see Table 1). There were two critical shapes of interest: squares and hearts. Each of these shapes belonged to one of two categories, the target or secondary categories. The target category is the category that the shape is most likely to be in given its distribution in the categories. For example, there was a $66 \%$ chance that a square belonged to Category 1, the target category, and a $33 \%$ chance that it belonged to Category 2, the secondary category (that is, there were eight squares in Category 1 and four in Category 2). In the target category, half of the squares moved in the 1
o'clock direction and half moved in the 5 o'clock direction. In the secondary category, the critical shapes moved in only one direction. In Condition 1 , the squares moved to 1 o'clock; in Condition 2, which served to counterbalance the direction of the secondary category, they all moved to 5 o'clock. Therefore, if people only attend to the target category in predicting the direction of a new square, they should be indifferent between predicting movement toward 1 and 5 o'clock, and thus their average prediction should be around 3 o'clock. If they attended to both the target and alternative categories, they should have a preference, because the alternative category (Category 2) would break the tie (in different directions in the two conditions).

This design was replicated for another stimulus and other directions: For hearts, the target category was Category 4, and half of the hearts moved in the 11 o'clock direction and half moved in the 7 o'clock direction. The secondary category was Category 3, and its hearts moved either toward 11 or 7 o'clock, depending on condition (see Table 1). Thus, if people integrated information across categories they would shift their predictions depending on what condition they are in, that is, depending on the less likely, secondary category. All subjects went through an identical learning phase in which they learned all four categories, based on the objects' shapes and direction of movement.

For the implicit test, subjects saw each shape presented briefly in the center of the screen before it rapidly moved off the screen in one of the learned directions. The subjects' task was to catch the shape with their cursor before it disappeared from the screen. Subjects were unable to catch the shapes in the middle of the screen, so they had to place their cursor towards the edge of the screen. Subjects controlled cursor placement and movement with the mouse.

For the explicit test, subjects were presented with static shapes and asked three questions about them: what category the shape was most likely to belong to, the probability their categorization was correct, and what direction the shape was most likely to travel in.

Subjects performed both the implicit and explicit induction tasks (order of tasks was counterbalanced). The results revealed that the exact same category knowledge led to significantly different inductions. Implicit inductions were, on average, shifted towards the secondary category, showing evidence of integration of information across categories. Explicit inductions showed no evidence of normative integration across categories. This pattern of results suggests that response mode is critical in determining how category information is used in induction. This is not to say that all things that make categories implicit lead to integration across categories. In Experiment 4 of Chen et al. (in press), subjects learned categories implicitly and made predictions explicitly. These predictions showed no evidence of integration of information across categories.

While these results suggest that implicit response promotes integration of information across categories, they are in contrast to much research on category-based induction under uncertainty which has consistently found

Table 1: Category Structure used in Experiments 1 and 2 (and Chen et al., in press)

|  | Category 1 <br> (target for squares) |  | Category 2 <br> (secondary for squares) |  | Category 3 <br> (secondary for hearts) |  | Category 4 (target for hearts) |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Exemplar | Shape | Direction | Shape | Direction* | Shape | Direction* | Shape | Direction |
| 1 | Square | 1 | Square | 1/5 | Heart | 7/11 | Heart | 7 |
| 2 | Square | 1 | Square | 1/5 | Heart | 7/11 | Heart | 7 |
| 3 | Square | 1 | Square | 1/5 | Heart | 7/11 | Heart | 7 |
| 4 | Square | 1 | Square | 1/5 | Heart | 7/11 | Heart | 7 |
| 5 | Square | 5 | Rectangle | 1/5 | Diamond | 7/11 | Heart | 11 |
| 6 | Square | 5 | Rectangle | 1/5 | Diamond | 7/11 | Heart | 11 |
| 7 | Square | 5 | Rectangle | 1/5 | Diamond | 7/11 | Heart | 11 |
| 8 | Square | 5 | Rectangle | 1/5 | Diamond | 7/11 | Heart | 11 |

Note. The direction entries are clock directions ( $1=1$ o'clock, etc.).
*The first number refers to the direction in condition 1 , the second to condition 2.
that most people based their inferences on only a single category. In Experiment 1, we seek to replicate this result with a different implicit measure of induction: eye movements. In Experiment 2, we provide evidence that subjects are not consciously aware of the strategies used in this implicit induction task.

## Experiment 1

To examine whether subjects would integrate information across categories when making predictions implicitly, Experiment 1 used a cover task, in which predicting movement was incidental. Subjects learned the four categories of moving shapes used in Chen et al. (in press). During test they performed a cover task (same/different task) in which they saw the shapes appear in the center of the screen. The shapes were the same as the ones subjects had learned, except they now had diagonal stripes that were either tilted right or left. After their initial presentation in the center, the shapes moved towards the edge of the computer screen but momentarily disappeared behind an annulus that was on the test screen such that subjects were unable to tell which direction the shape was going to move. Shapes briefly reappeared from behind the annulus and then disappeared off the edge of the screen. When the shapes reappeared from behind the annulus, their stripes may have reversed their tilt (e.g., from left to right). Subjects' task was to report whether the tilt of the stripes was the same or different from when it appeared in the center of the screen.

Thus, subjects were never asked to predict direction or category as they were only questioned about the stripes. However, since the shapes only reappeared briefly, looking close to where they reappeared improved performance (e.g, for squares, it would be beneficial to look near 1 o'clock or 5 o'clock depending on where you thought it would go). Position of eye gaze just prior to the shape's reappearance is the dependent measure as it is a proxy for subjects' prediction of shape direction. If subjects integrate information across categories, fixations should, on average, be shifted towards the direction of the secondary category.

## Method

Design Subjects were randomly assigned to one of two between-subjects conditions. The conditions served to counterbalance the direction of the secondary categories.
Participants Subjects were 32 undergraduates at New York University who participated for course credit. Data from eight subjects were dropped for not fixating prior to the shape's reappearance on at least five trials. One subject was dropped for not reaching the performance criterion during learning.
Materials Stimuli for each category were 8 black shapes approximately 1.75 to 2.5 cm in length, as shown in Table 1. The same shapes were used during test except they had stripes (see Figure 1). The category structure was the same as that used in Chen et al. (in press). See Table 1 for details.

All stimuli were presented on the background of a light gray circle 30 cm in diameter centered on a black computer screen. Stimuli started in the center of the screen and then moved off the screen disappearing once they moved beyond the border of the circle. Eye movements were monitored with the SR Research (Ontario, Canada) EyeLink 1000.
Procedure The experiment consisted of three phases: 1) observation, 2) learning, and 3) test. A Macintosh computer presented the instructions and controlled all three phases. Eye movements were recorded during the test phase only.

Subjects were told that they would view four categories of moving shapes and were to learn what combination of shapes and directions belonged to each category for a memory test. During observation, all shapes from each category were presented singly. Each shape appeared in the center of the screen for 1 s , then moved horizontally (towards 3 o'clock for shapes in Categories 1 and 2, towards 9 o'clock for Categories 3 and 4) for .4 s , and then moved towards its assigned clock direction for .95 s until it disappeared off the edge of the gray circle (see Table 1 for directions). Each shape's category name appeared in the center of the screen for the entire time it was on the screen. All exemplars from Category 1 were presented, then all exemplars from Category 2, and so on.

Subjects were next told that they would see the same items as in the observation phase. They were to classify
each shape into one of the four categories by pressing a number key on the keyboard. At the beginning of each trial, a white fixation cross appeared in the center of the screen for 1 s . The shape then moved as they did in the observation phase. There was no time limit on responding. After answering, the correct answer appeared for 1.25 s . After an error, subjects viewed a repeat display (without responding) of the moving shape with the correct category displayed. There were four learning blocks in which each of the 32 items was tested in random order. Because of the category uncertainty of the critical items (e.g., a square could be in Categories 1 or 2 ), subjects could get no more than $75 \%$ correct, assuming they chose the most likely category for all presented stimuli. In all experiments subjects had to reach at least $50 \%$ correct during the final block of learning to be included in analysis.

The final phase of the experiment consisted of a 64-trial test in which subjects had to perform the same/different task while their eye movements were tracked by the EyeLink 1000. Subjects saw the same items they had seen in the previous phases except that the shapes would now move a little bit faster and have diagonal stripes on them. These shapes would appear in the center of the screen (for 1 s ) and continue to move along the same path as in previous phases. However, there was now a black annulus on the screen such that the shape would move horizontally (for .25 s ) and then disappear behind the annulus for .7 s . The shape would then reappear from behind the annulus just before it disappeared from the screen. (After the shape's reappearance from behind the annulus it was visible for .15 s before it disappeared.) Recall that all stimuli were presented on a gray circle 30 cm in diameter. The annulus $(24 \mathrm{~cm}$ in diameter) was centered on this image. Its center hole had a diameter of 8 cm (see Figure 2).

The stripes on a test object were either tilted left or right when the shape initially appeared (see Figure 1). The subjects' task was to report whether the direction of the stripes was the same or different when it reappeared. The direction of stripes remained the same for half of the trials and changed for the other half. Subjects saw a 1.25 s feedback message. There were five practice trials prior to the test phase. As shapes only briefly reappeared from behind the annulus, looking close to where shapes reappeared was beneficial. Thus, fixation location just prior to the shape's reappearance was used as a proxy for prediction of direction and as the dependent measure. (Recall that horizontal movement for the critical shapes did not indicate its category as the horizontal direction was the same for Categories 1 and 2, and Categories 3 and 4.)


Figure 1: Example of stimuli used in the test phase of Experiments 1 and 2.


Figure 2: Illustration of the implicit induction task. The shape appeared in the center of the screen for 1 s . It then moved horizontally for .25 s and disappeared behind the annulus while traveling on its path (learned in phase 1). Subjects reported whether the diagonal lines had changed when it reappeared. Arrows indicate the shape's path when
it was visible and did not appear in the experiment.
Data Analysis Responses for critical shape trials were coded such that a position exactly in between the two possible directions of the shape was 0 degrees, and a shift from that point towards the direction reinforced by the secondary category was coded as positive. For example, for the squares in Condition 1 (which might move to 1 o'clock or 5 o'clock), the 3 o'clock position was 0 degrees, the 1 o'clock position (the direction of the secondary category) was 60 degrees, and the 5 o'clock position was -60 degrees. In Condition 2, the latter values were reversed. We obtained the mean fixation position for each subject by averaging the mean fixation position for squares and hearts. Thus, use of a single category (i.e., use of only the target category) is evidenced by an average prediction of 0 deg. Normative use of categories is evidenced by a positive average prediction, as this represents a shift from 0 deg in the direction of the secondary category.

Trials in which the fixation position was greater than 100 degrees or less than -100 degrees were not included in the analysis because the subject was fixated on the opposite side of the screen from where the shape traveled, indicating that the subject either forgot where the shapes went, or did not see the shape correctly prior to its movement. Additionally, trials where fixation was within the hole of the annulus were excluded from analysis. When subjects looked at the center of the screen while doing the task, they were effectively not making a prediction about direction.

## Results \& Discussion

Subjects were on average $66.4 \%$ correct (chance $=25 \%$ ) during their last training block, suggesting that they learned the categories quite well. (Recall that maximum performance was $75 \%$, if subjects always classified
ambiguous items into the most likely category.) Performance on the same/difference task averaged $72 \%$.

As explained above, integration of information across categories is evidenced by a shift from 0 deg in the direction of the secondary category, which we coded as positive. This is indeed what we found. The mean fixation position for the critical shapes, $(M=7.5 \mathrm{deg}, S D=8.9)$, was significantly greater than $0 \mathrm{deg}, t(23)=4.1, p<.01, d=.84$, indicating that people's predictions of direction were integrated across the two categories. The mean fixation position was positive for 21 of the 24 subjects. These results are consistent with those of Chen et al. (in press) and suggest that implicit induction promotes integration of information across categories. A question for future research will be to examine how categories are used during implicit induction. The multiple category use found in Experiment 1 may be a result of a feature-level strategy (e.g., using information about only squares when making a prediction about where a square will go) rather than a category-level strategy like that described in Eq 1 (see Griffiths et al., 2011, for similar ideas).

Perhaps subjects did not truly induce the objects' direction but learned to change their eye movements via practice in doing the task. To examine this possibility we compared the mean fixation position for the first and second blocks of testing. The difference between the mean fixation positions for the first and second blocks was not significant $(M s=6.2$ and $8.8 \mathrm{deg}, S D s=9.1$ and 11.9), $t(23)=1.0, p$ $>.05, d=.25$ suggesting that subjects' normative use of categories was not a result of learning during test. The positive shift in eye movements was significant in block 1, $t(23)=3.3, p<.01, d=6.8$, and in block $2, t(23)=3.6, p<$ $.01, d=.76$.

## Experiment 2

Experiment 1 revealed that people use information from multiple categories when making inductions implicitly. However, it is possible that the placement of eye fixation was not the result of implicit processes but instead the result of a conscious strategy (i.e., after practice subjects could have realized that they would perform better when they looked closer to the direction reinforced by the secondary category). Experiment 2 tested this explanation. Subjects completed the full learning procedure of Experiment 1. They then saw a few example trials of the same/different task and then reported (using feedback from the eyetracker) where they would look to best perform the task. This question sampled subjects' explicit beliefs about where they would look. If the results match those of Experiment 1, this would suggest that the fixations were the result of an explicit strategy.

## Method

Participants Subjects were 21 New York University undergraduates who participated for course credit. Data from four subjects were dropped for not fixating on at least
three trials. One more subject was dropped for not reaching the performance criterion during learning.
Materials and Design Identical to Experiment 1.
Procedure The procedures of the observation and learning phases were identical to those used in Experiment 1. As with Experiment 1, eye movements were only recorded during the test phase. The test phase consisted of a 16-trial test in which subjects were asked to report where they would look in order to best do the same/different task that subjects in Experiment 1 performed. Subjects saw the same five practice trials used in Experiment 1 and then were told that they would not be doing the task but rather reporting where they would look just prior to the shape's reappearance from behind the annulus to best do the task. In order to keep the dependent measures of the two experiments similar, we used eye position to indicate this prediction. A white dot on the display indicated where the subjects were looking. The task was to look at the location on the screen that they thought would be best to do the same/different task they had just observed. They then saw a test screen (gray circle with the annulus) and were instructed to look around the screen to get a sense of how the white dot corresponded to their eye gaze.

The test phase consisted of four blocks in which each shape was tested once in random order (except that shapes were not queried twice in a row). Each test trial started with the presentation of the shape in the center of the screen for 1 s. It then moved horizontally for .25 s until it disappeared behind the annulus (the shape never reappeared). Subjects then saw the white dot that marked their eye gaze on the screen. To report their location, subjects moved their eyes until they were satisfied with the location of the white dot and then pressed the enter key. The white dot stayed on the screen for 1.25 s so that the subjects could see their answer.

## Results \& Discussion

Subjects were on average $68.2 \%$ correct (chance $=25 \%$ ) during their last training block, near the $75 \%$ maximum, suggesting that they learned the categories quite well.

As in the analysis of Experiment 1, subjects' responses for the critical shapes were coded such that the time corresponding to the point exactly in between the two possible directions of the shape was 0 degrees ( 3 o'clock for squares and 9 o'clock for hearts, and a shift towards the direction reinforced by the secondary category was positive). To find the mean prediction (the amount of shift from 0 deg towards the secondary category) for each subject, we calculated the mean prediction for each shape and took the average of the two. The mean prediction ( $M=$ $0.2 \mathrm{deg}, S D=2.9 \mathrm{deg}$ ) was not significantly different than the average observed direction for the shapes in their target category only ( 0 deg ), $t(15)=0.2, p>.05, d=.07$, suggesting that subjects were not basing their responses on multiple categories. Subjects chose locations around 0 deg the majority of the time. In fact, $84 \%$ of all responses were with within 10 deg of 0 deg. In contrast, in Experiment 1, only $25 \%$ were in this range.

These predictions from this experiment show no integration of information across categories. This suggests that the integration of information across categories found in Experiment 1 was not the result of a conscious decision or strategy and provides further evidence that response mode is critical to how category information is used in induction.

## General Discussion

The results of Experiment 1 suggest that people integrate information across categories when making inductions implicitly. The results of Experiment 2 revealed that explicit prediction of eye fixation position in the same/different task showed no evidence of integration of information, suggesting that subjects were unaware of the strategies used to perform the task. Taken together, these results suggest that response mode is critical in determining when people integrate information across categories when making inductions and that the single category focus found in previous research on category-based induction may result from conscious reasoning strategies. These results are consistent with the findings of Chen et al. (in press), that speeded catching of a stimulus also showed integration across categories, but verbal predictions did not. These results also help explain the discrepancy between studies of induction in reasoning vs. perception and action.

Our findings suggest that implicit responses can, at least sometimes, lead to greater use of available information than our conscious, explicit responses do. This is particularly important because many everyday predictions are about items whose categorizations we may be unsure of. Doctors may have to predict which treatment is most likely to work even though they are not certain what the correct diagnosis is. A person who is walking alone at night and sees an unknown person approaching may have to decide whether to avoid the person despite being unsure whether that person belongs to the category of mugger or pedestrian. The results of the present experiments help in understanding which situations and contexts people are most likely to consider alternative possibilities and make predictions based on relevant information from them.

Additionally, many of these inferences can be made either implicitly or explicitly (e.g., one might run upon seeing an unknown person approaching, but given more time, one may exclude less likely possibilities and act as if certain that the unknown person is a pedestrian). In fact, in social psychology, a similar distinction has been made between automatic and controlled processes in prejudice. Automatic processes are often associated with stereotype activation (a type category-based induction) which, in low-prejudice people, conflicts with explicit attitudes and is inhibited in favor of explicit beliefs (Devine, 1989). Thus, the explicit system's bias to disregard or avoid information from alternative categories (that made it less normative in our task) could, in other cases, lead to more normative responses. Our research shows that this distinction is crucial for understanding when category-based predictions are more likely to be accurate or inaccurate.

## Acknowledgments

We thank Rebecca Bainbridge for her help with data collection and analysis. The research was supported in part by NSF grants BCS-1128769 and 1128029.

## References

Anderson, J. R. (1991). The adaptive nature of human categorization. Psychological Review, 98, 409-429.
Chen, S. Y., Ross, B. H., \& Murphy, G. L. (in press). Implicit and explicit processes in category-based induction: Is induction best when we don't think? Journal of Experimental Psychology: General.
Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. Journal of Personality and Social Psychology, 56, 5-18.
Evans, J. S. B. (2007). Hypothetical thinking: Dual processes in reasoning and judgment. New York: Psychology Press.
Griffiths, O., Hayes, B. K., Newell, B. R., \& Papadopoulos, C. (2011). Where to look first for an explanation of induction with uncertain categories. Psychonomic Bulletin \& Review, 18, 1212-1221.
Haruno, M., Wolpert, D., \& Kawato, M. (2001). Mosaic model for sensorimotor learning and control. Neural Computation, 13, 2201-2220.
Hayes, B. K., \& Newell, B. R. (2009). Induction with uncertain categories: When do people consider the category alternatives? Memory \& Cognition, 37, 730-743.
Kersten, D., Mamassian, P., \& Yuille, A. L. (2004). Object perception as Bayesian inference. Annual Review of Psychology, 55, 271-304.
Malt, B. C., Ross, B. H., \& Murphy, G. L. (1995). Predicting the features for members of natural categories when categorization is uncertain. Journal of Experimental Psychology: LMC, 21, 646-661.
Murphy, G. L., Chen, S. Y., \& Ross, B. H. (2012). Reasoning with uncertain categories. Thinking \& Reasoning, 18, 81-117.
Murphy, G. L., \& Ross, B. H. (1994). Predictions from uncertain categorizations. Cognitive Psychology, 24, 148193.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. Psychological Bulletin, 119, 3-22.
Tassinari, H., Hudson, T. E., \& Landy, M. S. (2006). Combining priors and noisy visual cues in a rapid pointing task. Journal of Neuroscience, 26, 10154-10163.
Trommershäuser, J., Landy, M. S., \& Maloney, L. T. (2006). Humans rapidly estimate expected gain in movement planning. Psychological Science, 17, 981-988.
Trommershäuser J., Maloney L. T., \& Landy M. S. (2008). Decision making, movement planning and statistical decision theory. Trends in Cognitive Science 12, 291-297.

# General Instruction Following in a Large-Scale Biologically Plausible Brain Model 

Xuan Choo (fchoo@uwaterloo.ca)<br>Chris Eliasmith (celiasmith@uwaterloo.ca)<br>Center for Theoretical Neuroscience, University of Waterloo<br>Waterloo, ON, Canada N2L 3G1


#### Abstract

We present a spiking neuron brain model implemented in 318,870 LIF neurons organized with distinct cortical modules, a basal ganglia, and a thalamus, that is capable of flexibly following memorized commands. Neural activity represents a structured set of rules, such as "If you see a 1, then push button A, and if you see a 2 , then push button B". Synaptic connections between these neurons and the basal ganglia, thalamus, and other areas cause the system to detect when rules should be applied and to then do so. The model gives a reaction time difference of 77 ms between the simple and two-choice reaction time tasks, and requires 384 ms per item for sub-vocal counting, consistent with human experimental results. This is the first biologically realistic spiking neuron model capable of flexibly responding to complex structured instructions.


Keywords: neural engineering; spiking neuron model; instruction following; instruction processing; cognitive control; cognitive architectures

## Introduction

One of the hallmarks of complex cognition is the ability to perform a multitude of tasks using the same underlying architecture. When given an instruction, the human brain is capable of processing and executing the instruction without the need for extensive rewiring of the underlying neural connections. As far as we are aware, no neural model to date has been shown to exhibit this ability.

Eliasmith et al. (2012) describes what is currently the world's largest functional brain model. While the model, called Spaun (for Semantic Pointer Architecture Unified Network), is able to perform 8 different cognitive tasks without necessitating changes to its architecture, the knowledge needed to complete these 8 tasks is hard-coded into the action selection mechanism (the basal ganglia) of the model, making it unable to perform any task other than the predefined 8. In this paper, we propose an extension to the Spaun action selection component making it capable of processing generic instructions.

## Terminology

Four key concepts are discussed in this paper: states, actions, rules, and instructions.

States are internal variables that the action selection system monitors to figure out what is the best action to perform. States can be both internal (e.g. goal memories, working memories (WM)) and external (e.g. visual input) to the system.

Actions are atomic commands within the architecture, and are typically motor commands (e.g. "write the number $X$ ", "push the $X$ button") or cognitive commands (e.g. "remember the word $X$ ", "route information from WM area $X$ to WM area
$Y$ ", "add 1 to the value in WM area $X$ "). Apart from motor and cognitive commands, actions can also be utilized to change the values of the model's states.

Rules are conditional statements typically of the form "IF $X$, THEN $Y$ " (e.g. "If you see a 1, then push button A") where $X$ is a set of conditions which have to be met for the set of actions $Y$ to be executed. More generally, in Spaun, rules are statistical maps between cortical states and actions.

An instruction is a combination of rules or actions that can be executed sequentially (e.g. "Remember the number 1 ; add 1 to that number; write the result") or in any order (e.g. "If you see a 1 , then push button A; If you see a 2 , then push button B").

## Spaun

The architecture of Spaun (the Semantic Pointer Architecture, or SPA) is composed of 9 distinct but interconnected modules (see Figure 1A). Of interest to this paper is how the action selection module interacts with the rest of the model. Fundamentally, the action selection module of Spaun is identical to the basal ganglia (BG) based production system described in (Stewart, Bekolay, \& Eliasmith, 2012), and functions similarly to the action selection component of production system models (e.g. (Anderson, 1996)).

In these systems, action selection is hard-coded by a predefined set of rules. To select an action, the BG monitors internal cortical state variables and executes a rule whose antecedent best matches the values of the internal state variables (see Figure 1B). Critically, to encode instructions, the transitions between each rule in the instruction has to be hardcoded into the BG as well. For example, if the instruction was to perform ACTION-A followed by ACTION-B, and then $A C T I O N-C$, the following rules would have to be encoded into the BG :

$$
\begin{aligned}
& \text { IF } \text { INIT, THEN } \text { state }=A C T I O N-A \\
& \text { IF } \text { state }=A C T I O N-A, \text { THEN } \text { state }=A C T I O N-B \\
& \text { IF } \text { state }=A C T I O N-B, \text { THEN } \text { state }=A C T I O N-C
\end{aligned}
$$

Several ACT-R models (e.g. (Taatgen \& Lee, 2003), (Taatgen, 1999)) able to follow instructions, however no neural implementation has been previously discussed.

Aside from its architecture, Spaun is also unique in the way information is represented. Information is encoded and represented using semantic pointers (Eliasmith, In Press). These representations are used in the SPA to define a type of vector symbolic architecture (VSA). In typical VSAs, the vector


Figure 1: A) High-level architecture of Spaun. B) Method by which Spaun chooses an action. The action selection system monitors cortical state variables (solid arrows), selects an action that bests matches these states, and effects the action on efferent modules (dotted arrows).
that represents the number ONE and the vector that represents the number TWO would be chosen from a random distribution and thus have no direct relation to each other. In the SPA however, the semantic pointer for the number TWO is computed as the bound combination of the semantic pointer ONE with a vector that represents the concept ADD1, thus imparting semantic meaning to each vector:

$$
\mathbf{T W O}=\mathbf{O N E} \circledast \mathbf{A D D 1}
$$

Similarly, the semantic pointer for the number THREE can be computed as follows:

$$
\begin{aligned}
\mathbf{T H R E E} & =\mathbf{T W O} \circledast \mathbf{A D D 1} \\
& =\mathbf{O N E} \circledast \mathbf{A D D 1} \circledast \mathbf{A D D 1}
\end{aligned}
$$

## Vector Symbolic Architectures

Vector symbolic architectures have four core properties. First, information is represented by high-dimensional vectors usually chosen from a random distribution.

Second, vectors can be combined using a superposition operation (denoted with $\mathrm{a}+$ ). Of note, the vector result of the superposition operation is similar to the original vector operands, where similarity is measured by a dot product.

Third, vectors can be bound together using a binding operation (denoted with a $\circledast$ ). Unlike the superposition operator, the vector result of the binding operation is dissimilar to the original vector operands.

Last, an approximate inverse operator (denoted with *, such that $A^{*}$ is the approximate inverse of $A$ ) is defined such
that binding $A$ with $A^{*}$ results in approximately the identity vector $\mathbf{I}\left(A \circledast A^{*} \approx \mathbf{I}\right)$. This property of the approximate inverse can be used to unbind previously bound vectors.

Both the superposition and binding operations are analogous to addition and multiplication in scalar mathematics, and are often associative, commutative, and distributive.

In the SPA, vector addition is used for superposition, and circular convolution is used for binding, bearing close similarity to the Holographic Reduced Representation (Plate, 2003).

## Encoding Instructions

Instructions are encoded using a positional encoding schema similar to that used in Spaun and in the Ordinal Serial Encoding model of serial working memory (Choo, 2010). Each rule in the instruction is tagged (bound) to a position vector to indicate its relative order within the instruction. For example, the instruction "1. RULE1; 2. RULE2" is encoded as

$$
\mathbf{I N S T R}=\mathbf{P} 1 \circledast \text { RULE1 }+\mathbf{P} 2 \circledast \text { RULE2 }
$$

where P1 and P2 are the position vectors. Importantly, since the position vectors are also semantic pointers, the position vectors have some relation. That is to say $\mathbf{P 2}=\mathbf{P} 1 \circledast$ ADD1, and likewise for subsequent position vectors.

Individual rules in the instruction are encoded as a superposition of the conditions that make up the antecedent and the actions that make up the consequence of the rule. For example, the rule "IF STATEA THEN ACTIONB" is encoded as

$$
\mathbf{R U L E}=\operatorname{ant}(\mathbf{S T A T E A})+\mathbf{A C T I O N B}
$$

where $\operatorname{ant}()$ is a randomly generated linear operator applied to the STATEA vector that serves to disambiguate the antecedent and consequent components of the rule.

State conditions are encoded by binding vectors that describe the state being monitored with the state value required for the rule to be executed. Thus, the state condition "state $=$ $\mathbf{A}^{"}$ is constructed as

## $\boldsymbol{S T A T E A}=\boldsymbol{S T A T E} \circledast \mathbf{A}$.

Other conditions can also be combined in this state representation. For example, if the state conditions was "vision $=\mathbf{3}$ and state $=\mathbf{A}$ " (i.e. looking at a 3 while in state A), then the state representation would be

$$
\text { VIS3\&STATEA }=\text { VISION } \circledast 3+\text { STATE } \circledast \text { A. }
$$

Actions are encoded by combining the bound result of an "action" descriptor with the specific action to be performed with an optional bound result of a "data" descriptor with the specific data to be used with the action. A "write the number 2 " action is thus represented as

$$
\mathbf{W R I T E} 2=\text { ACTION } \circledast \mathbf{W R I T E}+\text { DATA } \circledast 2 .
$$

Combining all of the representations above, the full encoding of an instruction can be demonstrated. As an example the instruction:

> 1. IF vision $=0$, THEN push button A
> 2. IF vision $=1$, THEN push button B
is encoded as

$$
\begin{array}{r}
\mathbf{I N S T R}=\mathbf{P 1} \circledast[\operatorname{ant}(\mathbf{V I S I O N} \circledast \mathbf{0})+ \\
\text { ACTION } \circledast \mathbf{P U S H}+\text { DATA } \circledast \mathbf{B T N A}]+ \\
\mathbf{P 2} \circledast[\operatorname{ant}(\mathbf{V I S I O N} \circledast \mathbf{1})+ \\
\text { ACTION } \circledast \mathbf{P U S H}+\mathbf{D A T A} \circledast \mathbf{B T N B}] \tag{1}
\end{array}
$$

It is important to note that at the end of this computation, the instruction is encoded as a single vector with the same dimensionality as the original atomic components.

## Decoding Instructions

With the instruction encoding schema presented above, the instructions can be decoded in one of two ways: by using positional information, and by using the values of the states in the system.

Sequential Decoding of Instructions A rule associated with a specific position within the instruction can be retrieved by binding the instruction vector with the inverse of the position vector.

$$
\begin{align*}
\text { rule } & =\mathbf{I N S T R} \circledast \mathbf{P} \mathbf{1}^{*}  \tag{2}\\
& =\mathbf{P} \mathbf{1}^{*} \circledast \mathbf{P} 1 \circledast \mathbf{R U L E} 1+\mathbf{P} \mathbf{1}^{*} \circledast \mathbf{P} \mathbf{2} \circledast \mathbf{R U L E} 2 \\
& =\mathbf{I} \circledast \mathbf{R U L E} 1+\mathbf{P} \mathbf{1}^{*} \circledast \mathbf{P} \mathbf{2} \circledast \mathbf{R U L E} 2 \\
& \approx \mathbf{R U L E} 1
\end{align*}
$$

Given the rule vector, it is possible to retrieve information related to the consequent by binding it with the inverse of the "action" descriptor or the "state" descriptor.

$$
\begin{align*}
\text { action }= & \text { rule } \text { ACTION }  \tag{3}\\
= & {[\operatorname{ant}(\text { VISION } \circledast \mathbf{0})+\text { ACTION } \circledast \mathbf{P U S H}+} \\
& \quad \text { DATA } \circledast \mathbf{B T N A}] \circledast \boldsymbol{A C T I O N}^{*} \\
\approx & \mathbf{I} \circledast \mathbf{P U S H}=\mathbf{P U S H}
\end{align*}
$$

Likewise,

$$
\begin{aligned}
\text { data }= & \operatorname{rule} \text { DATA } \boldsymbol{D A T}^{*} \\
= & {[\operatorname{ant}(\text { VISION } \circledast \mathbf{0})+\boldsymbol{\text { ACTION }} \circledast \mathbf{\text { PUSH }}+} \\
& \quad \text { DATA } \circledast \text { BTNA }] \circledast \text { DATA }^{*} \\
\approx & \mathbf{I} \circledast \text { BTNA }=\mathbf{B T N A}
\end{aligned}
$$

After the rule has been executed, the next rule can be computed by incrementing the position vector ( $\mathbf{P 2}=\mathbf{P} 1 \circledast$ ADD1)) and repeating Equations 2, $3 \& 4$ with this new position vector.

Conditionally Responsive Decoding of Instructions An instruction can also be decoded using the values of the state conditions. In order to do so, the value of the state condition(s) is bound to its associated state descriptor(s), and the inverse of this result is bound to the instruction vector to yield the position of the rule that best matches the state condition(s). Using Equation 1 as an example, if the vision state condition had a value of 1 , the position of the rule that best matches this can be found like so:

$$
\begin{align*}
p o s= & \mathbf{I N S T R} \circledast\left(\text { ant }(\text { state }) \circledast \text { state_val }^{*}\right)^{*}  \tag{5}\\
= & \mathbf{I N S T R} \circledast(\operatorname{ant}(\mathbf{V I S I O N}) \circledast \mathbf{1})^{*} \\
= & \mathbf{P 1} \circledast[\operatorname{ant}(\boldsymbol{V I S I O N} \circledast \mathbf{0}) \ldots] \circledast(\operatorname{ant}(\text { VISION }) \circledast \mathbf{1})^{*}+ \\
& \mathbf{P} 2 \circledast[\operatorname{ant}(\boldsymbol{V I S I O N} \circledast \mathbf{1}) \ldots] \circledast(\operatorname{ant}(\boldsymbol{V I S I O N}) \circledast \mathbf{1})^{*} \\
\approx & \mathbf{P} \mathbf{2} \circledast[\mathbf{I}+\ldots] \approx \mathbf{P} \mathbf{2}
\end{align*}
$$

Once the position vector has been retrieved, the sequential instruction decoding equations can then be used to obtain the action and data associated with the rule.

## The Model

With the ability to encode and decode general instructions, modifying the existing Spaun action selection module to take advantage of this is straightforward. It only entails the addition of a instruction processing module that implements the instruction decoding equations (Eq $2-5$ ) above. The output of this module then become new state variables which the action selection system monitors when selecting an appropriate action (see Figure 2).


Figure 2: Proposed modification to Spaun's action selection system with the addition of an instruction processing module (italicized). As in Figure 1, state monitoring is indicated with a solid arrow, and action effects with a dotted arrow.

Validation of the model comes in the form of behavioural analysis as well as matching the model dynamics to human timing data. The model is implemented with spiking neurons and biologically realistic synaptic time constants in order to generate realistic temporal dynamics.

## Neural Representation

Fundamental to the SPA is the vector-based representation of information. We use methods of the Neural Engineer-
ing Framework (NEF) to accomplish this in spiking neurons (Eliasmith \& Anderson, 2003). Georgopoulos et al. (1986) demonstrated that motor neurons are well characterized as having responses driven by their preferred direction to movement in two dimensions. The NEF generalizes this notion to suggest that neurons can represent any number of dimensions, and the neuron's preferred direction determines its activity with regards to its input in a given vector space. Mathematically, the current $J$ flowing into a neuron can be calculated using as

$$
\begin{equation*}
J(\boldsymbol{x})=\alpha(\mathbf{e} \cdot \boldsymbol{x})+J^{\text {bias }} \tag{6}
\end{equation*}
$$

where $\alpha$ and $J^{\text {bias }}$ are neuronal scaling terms, e is the neuron's preferred direction (or encoding vector), and $\boldsymbol{x}$ is the vector to be represented. The inner product computes the similarity between the encoding and input vector and determines how much current is being fed to the neuron. The leaky integrate-and-fire (LIF) neuron model equation is then used to convert this current into a firing rate.

$$
\begin{equation*}
a(\boldsymbol{x})=G[J(\boldsymbol{x})]=\frac{1}{\tau^{r e f}-\tau^{R C} \ln \left(1-\frac{J^{t h}}{J(\boldsymbol{x})}\right)} \tag{7}
\end{equation*}
$$

In the equation above, $\tau^{r e f}$ is the neuron refractory time constant, $\tau^{R C}$ is the neuron RC time constant, and $J^{t h}$ is the neuron threshold firing current. With a population of neurons, it then possible to derive optimal decoding vectors that can be used to convert the neural activity back into the high dimensional vector space. Eliasmith and Anderson (2003) demonstrate how these decoders $\mathbf{d}$ can be computed.

$$
\begin{align*}
\mathbf{d} & =\Gamma^{-1} \Upsilon, \text { where } \\
\Gamma_{i j} & =\int a_{i}(x) a_{j}(x) \mathrm{d} x \quad \Upsilon_{i}=\int a_{i}(x) x \mathrm{~d} x \tag{8}
\end{align*}
$$

An estimate of the original vector $x$ can then be generated by multiplying each neuron's decoding vector with its activity.

$$
\begin{equation*}
\hat{x}=\sum_{i} a_{i}(x) \mathbf{d}_{i} \tag{9}
\end{equation*}
$$

The encoding and decoding vectors can also be used to determine the optimal connection weights between two neural populations.

$$
\begin{equation*}
w_{i j}=\alpha_{j} \mathbf{e}_{j} \mathbf{d}_{i} \tag{10}
\end{equation*}
$$

Taking into account a specific function while solving for the decoding vectors yields the set of connection weights that will cause the neurons in the post-synaptic population to compute said function. For example,

$$
\begin{equation*}
\widehat{f(x)}=\sum_{i} a_{i}(x) \mathbf{d}_{i}^{f} \tag{11}
\end{equation*}
$$

where $\mathbf{d}^{f}$ are the decoding vectors solved with the function $f$ incorporated into Equation 8. In other words, these equations allow us to build a spiking neuron model that performs arbitrary specified computations. See Eliasmith and Anderson (2003) for additional details.

## Neural Implementation

The model proposed here relies on two key functions: the binding operation and working memory.

The binding operation is performed by a two step process. First the Fourier transform (FT) of both input vectors is computed, and these are multiplied element-wise. Performing an inverse Fourier transform (IFT) on this result provides the desired answer. That is,

$$
\begin{equation*}
\mathbf{A} \circledast \mathbf{B}=\operatorname{IFT}(F T(\mathbf{A}) \odot F T(\mathbf{B})), \tag{12}
\end{equation*}
$$

where $\odot$ is the element-wise multiplication operation. The FT, IFT and element-wise multiplication are functions that can be computed by spiking neurons using the methods discussed in the previous section (see Equation 11).

The working memory component is identical to that used in Spaun. This component is implemented by a recurrent network that is able to stably store information over time. The storage and retrieval of information is determined by gates controlled by the basal ganglia.

## Response Timing

In this section we compare the behaviour of the model to two different tasks: the choice reaction time task, and a sub-vocal counting task. The choice reaction time task demonstrates the model's ability to perform unordered instructions, while the sub-vocal counting task demonstrates the model's ability to perform sequential instructions. Note that for both of these tasks, the architecture of the model remains the same, with the only difference being the instruction vector and visual stimuli that it is required to process.

## Conditionally Responsive Decoding - Two-Choice and Simple Reaction Time Task

To test the model's ability to account for human instruction processing time, it was tested with the two-choice (CRT) and simple reaction time (SRT) tasks described in Grice, Nullmeyer, \& Spiker (1982). Since the input stimuli and motor action performed are similar in both of these tasks, any difference in reaction time can be attributed to the speed at which the different instructions are processed.

In the two-choice reaction time task, the subject is instructed to push one of two buttons, the identity of which is indicated by some sort of visual stimuli. To simulate this with the general instruction following model, it is given the instruction:

$$
\begin{aligned}
& \text { 1. IF } \text { vision }=\text { ZERO, THEN } \text { state }=\text { Push, } \text { motor }=\mathbf{A} \\
& \text { 2. IF vision }=\mathbf{O N E}, \text { THEN } \text { state }=\text { Push, } \text { motor }=\mathbf{B}
\end{aligned}
$$

Figure 3 demonstrates the model performing this instruction. In the simple reaction time task, the subject is instructed to push a single button in response to a single stimulus. This task requires no instruction processing so the rule:

$$
\text { IF vision }=\mathbf{T W O}, \text { THEN } \text { state }=\text { Push, } \text { motor }=\mathbf{C}
$$



Figure 3: Neural response data for the two-choice reaction time task. Shown are the decoded representations for two neural populations (an internal state memory, and the motor output), and the visual stimulus provided. Also displayed is the spiking neural data associated with each of the neural populations. Note that only the cognitive components (i.e. no input stimuli processing lag nor motor lag) of the reaction time task are being simulated in this model.
is encoded directly in the basal ganglia. By doing so, the model is able to execute the desired action when presented with the appropriate stimuli without requiring any additional processing in the instruction processing module.

The model reports a reaction time difference of $77 \pm 34$ ms between the two tasks, while Grice, Nullmeyer, \& Spiker report a reaction time difference of $81 \pm 72 \mathrm{~ms}$ for human subjects.

## Sequential Decoding - Sub-vocal Counting

For this task, the model was given a sub-vocal counting instruction. This instruction is formatted as sequence of actions, and thus have no antecedent.

$$
\begin{aligned}
& \text { 1. } \text { memory }=\text { Store, data }=X \\
& \text { 2. state }=\text { Add } 1 \\
& \text { 3. state }=\text { Write, } \text { motor }=\text { memory }
\end{aligned}
$$

In the instruction above, the variable $\boldsymbol{X}$ is a vector representing a digit from 0 to 9 . Instructions requiring more than one count (e.g. add 1 twice), have the second action repeated the appropriate number of times (and appropriately renumbered). Figure 4 illustrates the model peforming the sub-vocal counting task for one count.

The mean reported count time per item is $384 \pm 29 \mathrm{~ms}$ which falls well between the reported human range of $344 \pm$ 135 ms (Landauer, 1962), and provides a much better match to the human data than Spaun's reported count time per item time of $419 \pm 10 \mathrm{~ms}$ (Eliasmith et al., 2012).

## Simulation Details

In total the model is made up of 318,870 spiking LIF neurons, and uses 256 -dimensional semantic pointers. It should be noted that Spaun utilizes semantic pointers with 512 dimensions, and this was reduced for this model to decrease the amount of time required to simulate the experiments. It takes $275 \pm 25$ seconds of CPU time to simulate 1 second of simulation time on a machine with a 3.40 GHz Core i7-3770 quad-core CPU and 16 GB of RAM.

## Discussion

The model presented in this paper demonstrates the ability to process and execute generic instructions without needing any changes to the underlying architecture. It is also able to reproduce response times in human reported ranges based purely on the temporal dynamics of the underlying neural implementation - without the need for data fitting of any kind.

Because the model utilizes semantic pointers to represent information, it is also highly scalable. The maximum number of concepts the model is able to represent is dependent on the dimensionality of the semantic pointer used, and not on the number of knowledge nodes present in the model. Crawford, Gingerich and Eliasmith (Crawford, Gingerich, \& Eliasmith, 2013) demonstrate that the entirety of WordNet (117,659 concepts) can be represented using 512 dimensional semantic pointers. Increasing the proposed model to utilize 512 dimensional semantic pointers would add an additional 287,488 neurons to the model.

One major limitation to this model, however, is its inability to learn frequently executed instructions. In essence, even


Figure 4: Neural response data for the sub-vocal counting task. Shown are the decoded representations for three neural populations (an internal state memory, working memory (WM), and the motor output), and the visual stimulus provided. Also displayed is the spiking neural data associated with each of the neural populations. Note that the ADD value for the state variable indicate both the start and end of the number addition action.
if it is presented with multiple instances of the same instruction, it is unable to form an expert action for that instruction. This issue is currently being investigated and integrating this ability in the proposed model remains as future work.

This paper also makes no mention of how the model could construct a new instruction vector given purely a visual stream of words or symbols. Concurrent work done by Stewart and Eliasmith (Stewart \& Eliasmith, 2013) provides insight on how this issue can be made tractable.

## Acknowledgments

This research was supported by the Natural Sciences and Engineering Research Council of Canada, the Canadian Foundation for Innovation, and the Ontario Innovation Trust.

## References

Anderson, J. R. (1996). Act: A simple theory of complex cognition. American Psychologist, 51, 355-365.
Choo, X. (2010). The ordinal serial encoding model: Serial memory in spiking neurons. Unpublished master's thesis, The University of Waterloo.
Crawford, E., Gingerich, M., \& Eliasmith, C. (2013). Biologically plausible, human-scale knowledge representation. In Proceedings of the 35th annual conference of the cognitive science society.
Eliasmith, C. (In Press). How to build a brain: A neural architecture for biological cognition. New York, NY: Oxford University Press.
Eliasmith, C., \& Anderson, C. H. (2003). Neural engineer-
ing: computation, representation, and dynamics in neurobiological systems. Cambridge, MA: The MIT Press.
Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., Dewolf, T., Tang, Y., et al. (2012). A large-scale model of the functioning brain. Science, 338(6111), 1202-1205.
Georgopoulos, A. P., Schwartz, A., \& Kettner, R. E. (1986). Neuronal population coding of movement direction. Science, 233, 1416-1419.
Grice, G. R., Nullmeyer, R., \& Spiker, V. A. (1982). Human reaction time: Toward a general theory. Journal of Experimental Psychology: General, 111(1), 135-153.
Landauer, T. K. (1962). Rate of implicit speech. Perceptual and Motor Skills, 15, 646.
Plate, T. A. (2003). Holographic reduced representations: distributed representations for cognitive structures. Stanford, CA: CSLI Publications.
Stewart, T. C., Bekolay, T., \& Eliasmith, C. (2012). Learning to select actions with spiking neurons in the basal ganglia. Frontiers in Decision Neuroscience, 6(2).
Stewart, T. C., \& Eliasmith, C. (2013). Parsing sequentially presented commands in a large-scale biologically realistic brain model. In Proceedings of the 35th annual conference of the cognitive science society.
Taatgen, N. A. (1999). A model of learning task-specific knowledge for a new task. In Proceedings of the twentyfirst annual conference of the cognitive science society (pp. 730-735). Mahwah, NJ: Erlbaum.
Taatgen, N. A., \& Lee, F. J. (2003). Production composition: A simple mechanism to model complex skill acquisition. Human Factors, 45(1), 61-76.

# Understanding eye movements in face recognition with hidden Markov model 

Tim Chuk (u3002534@connect.hku.hk) ${ }^{1}$ Alvin C. W. Ng (asangfai@gmail.com) ${ }^{2}$ Emanuele Coviello (ecoviell@ucsd.edu) ${ }^{3}$<br>Antoni B. Chan (abchan@cityu.edu.hk) ${ }^{2}$ Janet H. Hsiao (jhsiao@hku.hk) ${ }^{1}$<br>${ }^{1}$ Department of Psychology, The University of Hong Kong, Pokfulam Road, Hong Kong<br>${ }^{2}$ Department of Computer Science, City University of Hong<br>${ }^{3}$ Department of Electrical and Computer Engineering, University of California San Diego, La Jolla, CA, USA

Kong, Tat Chee Avenue, Kowloon, Hong Kong


#### Abstract

In this paper we propose a hidden Markov model (HMM)based method to analyze eye movement data. We conducted a simple face recognition task and recorded eye movements and performance of the participants. We used a variational Bayesian framework for Gaussian mixture models to estimate the distribution of fixation locations and modeled the fixation and transition data using HMMs. We showed that using HMMs, we can describe individuals' eye movement strategies with both fixation locations and transition probabilities. By clustering these HMMs, we found that the strategies can be categorized into two subgroups; one was more holistic and the other was more analytical. Furthermore, we found that correct and wrong recognitions were associated with distinctive eye movement strategies. The difference between these strategies lied in their transition probabilities.


Keywords: Hidden Markov Model (HMM); eye movement; scan path; holistic processing; face recognition.

## Introduction

In the late $19^{\text {th }}$ century, soon after Edmund Huey's invention of the world's first eye tracker, researchers discovered that in many daily life activities, eye movements were rapid, discontinuous, and interrupted by temporary fixations (Wade \& Tatler, 2011). Nowadays, this finding has been widely accepted and described as the 'saccade and fixate' strategy (Land, 2011). Eye movements were found to facilitate face learning and recognition. For instance, Henderson et al. (2005) showed that when participants were restricted to view face images only at the center of the images, their recognition performances were significantly lowered than when they were allowed to view the images freely. Autistic patients, who could not judge facial expressions correctly, were found to have abnormal eye fixations patterns (Pelphrey et al, 2002).

Empirical studies on the relationship between eye movement and face recognition have primarily been focusing on identifying the regions of interest (ROIs). A ROI is a region on the face which people frequently fixate in, such as the two eyes. Early studies often divided a face into several regions and then identified the ROI through comparing the frequencies of each region being fixated in. However, this approach suffered from the lack of an objective manner to divide faces. For instance, Barton et al. (2006) defined the two eyes as two irregularly shaped

ROIs, while Henderson et al. (2005) defined the two eyes as one ROI. Another problem is that the predefined ROIs may not really represent the data because different individuals have different saccade patterns. More recent studies attempted to discover ROIs directly from data. A commonly adopted way was to generate statistical fixation maps. A fixation map can be created by identifying the location of fixations and convolving a Gaussian kernel on each fixation. Two fixation maps can be compared by Pixel test, which discovers statistically significant differences in pixels (Caldara \& Miellet, 2011). Using fixation maps, it was found that the upper center (i.e. the nose) and the upper left (i.e. the left half of the nose and the left eye) parts of a face were the two most frequently viewed areas (Hsiao \& Cottrell, 2008). This result was consistent with an earlier study which used the Bubbles technique in discovering regions with diagnostic features in face recognition (Gosselin \& Schyns, 2001). Fixation maps also showed that children from different cultural backgrounds demonstrated different eye fixation patterns (Kelly et al, 2011).

The use of fixation maps in face recognition studies had been fruitful. However, as discussed earlier, eye movements combine saccades and fixations. The fixations recorded in eye movement studies should be considered as time-series data that are collected over time. The eyes fixate at a location shortly, before a saccade brings them to the next location. Many studies showed that saccades can be influenced by top-down expectations as well as bottom-up inputs. Yarbus's (1965) well-known eye movement studies showed that depending on what people expect to see, they exhibited different saccade patterns when looking at the same target image. Mannan et al. (1997) discovered that saccades were more likely to be driven to the more 'informative' areas of an image, such as the edges and the high-spatial-frequency areas. These findings imply that the target location of a saccade could be a variable that has a set of possible values; different values could be associated with different probabilities. In this sense, eye movements may be considered as a stochastic process, which could be better understood using time-series probabilistic models. The fixation maps, however, do not contain temporal information.
Currently, there are two methods for describing the temporal information in eye movement data. One is the string-editing method. It requires an image to be divided
into several ROIs, each labeled with a letter, so that a sequence of eye fixations can be described by a string. Two strings are then compared by measuring their Levenshtein distance (Goldberg \& Helfman, 2010). This method does not capture the temporal information very precisely because the measure of Levenshtein distance does not precisely represent the sequential differences between two strings. For instance, the strings CAT and BAT differ in their first element, while the strings CAB and CAT differ in their last element. In both cases, however, the Levenshtein distance is one. The other method is to generate fixation maps by fixation and compare between conditions (Caldara \& Miellet, 2011). For instance, if an experiment has two conditions, all the first fixations in each condition can be used to generate a fixation map. A comparison between the two fixation maps will show whether the two groups differ significantly in their first fixations. However, the problem associated with this method is that the significant areas are likely to be scattered so that the pattern could be hard to interpret. In this paper, we propose to use a time-series statistical model, the hidden Markov model (HMM) with Gaussian emission densities, to analyze eye movement data. We show that HMMs can 1) summarize a person's general eye movement strategy, including person-specific ROIs and saccade patterns, 2) reveal between-subject similarities and differences of eye movement patterns, and 3) discover the association between recognition performance and eye movement strategies. In the next section, we will 1) briefly describe the experiment in which we collected the data, and 2) explain the HMMbased method in more length.

## Method

## Experiment

A total of 32 Chinese participants were recruited at the University of Hong Kong. The experiment was divided into a training phase and a testing phase. In the training phase, participants were shown a total of 20 frontal face images. In the testing phase, participants were shown 40 frontal face images; 20 were new images and 20 were the ones appearing in the training phase. They were asked to judge whether they had seen the faces before. Their responses in the testing phase were recorded together with the fixations they made before the response. Eye movements were tracked and recorded using the Eyelink II eye-tracking system. On average, participants made 2.5 fixations per trial, ranged from one to three (this average was 1.8 fixations in Hsiao \& Cottrell, 2008).

## Model

HMMs are widely used to model data generated from Markov processes (Barber, 2012). A Markov process is a process whose present state is determined only by its previous state. The states in an HMM are not directly observable, so that the current state of the process can only be inferred from 1) the association between the assumed hidden state and the observed data, and 2) the likelihood of transiting to
the assumed state from the previous state. The association among the observable data and the hidden states are summarized using probability distributions; each distribution represents the likelihood of a hidden state generating the data. The probabilities of transiting from one state to other states are summarized in a transition matrix; each element represents the probability of that transition. An HMM also has a vector of prior values; each value indicates the probability of the HMM starting from the corresponding state.

For instance, natural language processing is one area in which HMM has been widely applied. The observable data are the words in a corpus, and the hidden states are the word-class tags, such as nouns, verbs, and adjectives. An HMM cannot directly observe the word-class tags of the words, but can infer them from the observed words and the likelihood of transiting from one word-class to another.

In the context of face recognition, the HMM contains a number of hidden states, which each represents a different ROI of the face. The directly observable data is the fixation location, which belongs to a particular hidden state (ROI). The distribution of fixations in each ROI is modeled as a two-dimensional Gaussian distribution in a Cartesian space. Over time, the transition from the current hidden state to the next state represents the saccade pattern, i.e., movement between ROIs, which is modeled by the transition matrix of the HMM. In summary, the hidden states of the HMM correspond to the ROIs of the face, where each is observable through a two-dimensional Gaussian emission density of fixations, and the transitions between hidden states represent the saccade patterns.

Given a set of chains of fixations, we estimated the parameters of the HMM using a two-stage procedure. We first learned the ROIs on the face from the fixation data. Ignoring the temporal information, the ROIs can be seen as a mixture of two-dimensional Gaussian distributions, i.e., a Gaussian mixture model (GMM). In this study, we used the variational Bayesian framework for Gaussian mixture models (VBGMM) to estimate the Gaussian parameters, as well as the number of GMM components (Bishop, 2006). This Bayesian hierarchical method puts prior distributions on the GMM parameters, and uses approximation methods to find the maximum a posteriori (MAP) estimate. One important feature of VBGMM is that it can automatically estimate the optimal number of ROIs and 'deactivate' the redundant ones. After discovering the GMM components, or the ROIs, we next estimated the transition probabilities and prior probabilities of the hidden states, using the forwardbackward algorithm (Bishop, 2006).

In this study, we aim to use HMMs to address two questions. Firstly, we wanted to discover the eye movement strategy of each individual in order to reveal the common strategies shared by a subgroup of the participants. Secondly, we wanted to explore whether accuracy at face recognition was related to eye movements. To address the first question, we trained one HMM per subject, using fixations collected from all the trials of the subject, in order to represent the general eye movement pattern of that subject. To
cluster the subjects' HMMs, we used the variational hierarchical EM algorithm (VHEM) for HMMs (Coviello et al, 2012). The VHEM algorithm takes HMMs as inputs, separates the inputs into subgroups, and estimates a representation HMM for each subgroup.

To address the second question, we trained two HMMs per subject, using fixation sequences collected from all the correct trials (i.e., correct HMM) and all the wrong trials (i.e., wrong HMM), respectively, to represent two eye movement strategies that led to different performances. We compared the correct HMMs to the wrong HMMs using subject analysis, based on the differences in log-likelihoods of the observed data, in order to examine whether eye movement strategies that lead to correct or wrong responses have significantly different patterns. Specifically, for the fixation sequences of a participant leading to correct responses, we calculated the log-likelihoods of observing the sequences from the correct HMM, and then computed the mean. We also calculated the mean log-likelihood from the wrong HMM using the same sequences. Doing this on all the 32 participants yielded two vectors of mean loglikelihoods, one represented the mean log-likelihoods of the correct HMMs generating the correct eye movements, and one represented the mean log-likelihoods of the wrong HMMs generating the correct eye movements. The differences between the mean log-likelihoods for each subject is an approximation to the Kullback-Leibler (KL) divergence between the correct HMM and the wrong HMM, which is a measure of difference between two distributions (Bishop, 2006). Similarly, we also calculate the mean log-likelihoods of the fixation sequences leading to incorrect responses under the wrong and correct HMMs.

## Results

## Section 1.1- Summary of all eye movement patterns

In order to model a participant's eye movement patterns, we pooled all the fixations that a participant made, regardless of their sequential order, and applied the VBGMM to discover a mixture of Gaussian distributions. We then used the found Gaussian components and the fixations in the forward-backward algorithm to estimate the transition probabilities and the prior values of the Gaussian components. The fixations put into the forward-backward algorithm were in their sequential orders. Each participant's eye movements were modeled by an HMM. Using the VHEM to group all HMMs into one cluster, the VHEM generated a representation of the cluster which summarized the eye movement patterns of all the participants in one HMM. Figure 1 below shows the representation HMM and the fixation map of all the fixations combined. Figure 2 below shows the fixation maps per each fixation.

The left image in figure 1 shows the HMM model. For instance, the prior value of the red region suggests the probability of a first fixation belonging to that region. The probability of the next fixation transits from the red into the green region is 0.07 .


Figure 1: The image on the left shows the three GMM components of the HMM. Each colored region represents a ROI (red, green, or blue). The transition probabilities and the prior values are summarized in the table beneath. The image on the right shows the fixation map of all the fixa-
tions.


Figure 2: From the left to the right, the three images show the first, second, and third fixations that all subjects made.

From the comparison between the VHEM output and the fixation map of all the fixations combined, it can be seen that the VHEM output was spatially similar to the fixation map. The fixation map showed that most fixations landed in the middle of the face, with some slightly to the right. The three Gaussian components found using the VHEM demonstrated a similar tendency. One advantage that the VHEM output has over the fixation map is that on top of the spatial distributions, it provides the temporal information of the eye movement data in the forms of the prior values and the transition probabilities.

The transition probabilities and the prior values suggested that in general, fixations were more likely to start from the red and the blue regions and to remain in or shift between the two regions. The chance of beginning from the green region was lower. However, these fixations were more likely to stay in the same region than moving to the other regions. The fixation maps are shown in Figure 2. While there appears to be some movement between fixations, the fixation maps carry no information about the actual saccade pattern. However, using the results from the HMM analysis, we can better interpret the fixation maps. The higher probabilities of remaining in the same regions and the lower probability of starting from the green region may have re-
sulted in the fixations forming three separate clusters at the third fixation; the cluster corresponded to the green region was less compacted.

## Section 1.2-Two general strategies

Another advantage of using the HMM-based method is that the VHEM can group the input HMMs into several subgroups and generate a representation HMM for each subgroup. These would reveal the eye movement patterns shared by the participants in the same subgroup. The VHEM adopts a bottom-up, data-driven approach. It estimates the distance between an input HMM and a representation HMM. The distance between an input HMM and all the representation HMMs are then normalized, which gives a probability-based measure of how likely the input HMM belongs to a subgroup.

Using the VHEM, we discovered two subgroups, as shown in Figure 3 below.

|  | Holistic strategy | Analytic strategy |  |
| :---: | :---: | :---: | :---: |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| Holistic blue |  |  |  |
| prior values | 0.33 | 0.39 | 0.11 |
| from red | 0.65 | 0.24 | 0.17 |
| from green | 0.22 | 0.61 | 0.63 |
| from blue | 0.12 | 0.25 | to blue |
| Analytic | to red | to green | 0.47 |
| prior values | 0.06 | 0.47 | 0.39 |
| from red | 0.39 | 0.22 | 0.22 |
| from green | 0.03 | 0.75 | 0.70 |
| from blue | 0.05 | 0.25 |  |

Figure 3: The two representation HMMs of the two subgroups are shown in the left and the right images respectively.
It can be seen that the representation HMM on the left was more 'condensed'. The three Gaussian components were relatively small in size and were squeezed toward the center of the face. This pattern was similar to the "Eastern pattern" found in a previous study (Kelly et al., 2011) that was argued to represent a more holistic strategy. The HMM representation on the right, on the other hand, was more 'spread'. The three Gaussian components were large and more separated from one another. This pattern could be loosely associated with the "Western pattern" (Kelly et al., 2011) that represented a more analytic way of perceiving a face.

The table below shows the probabilities of the 32 HMMs belonging to the two subgroups. Each HMM was a model of a participant's eye movement patterns, so that the two numbers of each participant can be conceptualized as the degree to which the participant was biased to holistic or analytic eye movement strategies. Overall, 10 participants used holistic pattern, while 22 used the analytic strategy.

Table 1: Summary of the normalized log-likelihoods of the subjects belonging to the two subgroups.

| ID | Holistic | Analytic | ID | Holistic | Analytic |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 01 | 0 | 1 | 17 | 0 | 1 |
| 02 | 0 | 1 | 18 | .04 | .96 |
| 03 | 1 | 0 | 19 | 0 | 1 |
| 04 | 1 | 0 | 20 | 1 | 0 |
| 05 | 0 | 1 | 21 | 1 | 0 |
| 06 | 0 | 1 | 22 | 0 | 1 |
| 07 | 1 | 0 | 23 | 0 | 1 |
| 08 | 0 | 1 | 24 | 0 | 1 |
| 09 | 0 | 1 | 25 | 0 | 1 |
| 10 | 1 | 0 | 26 | 1 | 0 |
| 11 | 0 | 1 | 27 | 0 | 1 |
| 12 | 0 | 1 | 28 | 1 | 0 |
| 13 | 1 | 0 | 29 | 1 | 0 |
| 14 | 0 | 1 | 30 | .02 | .98 |
| 15 | 0 | 1 | 31 | 0 | 1 |
| 16 | 0 | 1 | 32 | 0 | 1 |

The log-likelihoods suggested that the two subgroups were very distinctive from each other. To confirm whether they really represented two distinctive eye movement patterns, we randomly created 50 pseudo-data chains; each was a sequence of three pseudo fixations. We measured the loglikelihoods of the two HMMs generating the pseudo-data. Paired t-test showed that the log-likelihoods generated by the two HMMs were significantly different, $\mathrm{t}(49)=-12.81$, $\mathrm{p}<.001$; mean log-likelihood difference was 13.84. The finding further confirmed that the two eye movement patterns were distinctive from each other.

## Section 2 - Association between performance and eye movement patterns

To investigate whether the differences in recognition performance are associated with different eye movement patterns, we trained per participant an HMM on all the fixations collected from the correctly responded trials (correct HMM), and an HMM on all the fixations collected from the incorrectly responded trials (wrong HMM). We compared the mean log-likelihoods of the data being generated by the two HMMs.

Paired t-test showed that the mean log-likelihoods of correct data being generated by the correct HMMs ( $\mathrm{M}=$ 18.13) were significantly higher than the mean loglikelihoods of correct data being generated by the wrong HMMs ( $\mathrm{M}=-18.42$ ), $\mathrm{t}(31)=-2.58, \mathrm{p}=.01$. The mean loglikelihoods of the wrong data being generated by the wrong

HMMs ( $\mathrm{M}=-17.9$ ) was also significantly higher than the mean log-likelihoods of correct data being generated by the wrong HMMs $(\mathrm{M}=-18.53), \mathrm{t}(31)=-4.58, \mathrm{p}<.001$. The results suggested that the two sets of HMMs were significantly different from each other. Figure $4-7$ below illustrate the HMMs and the fixation maps of a few subjects.


Figure 4: The correct and wrong HMMs of subject 1.


Figure 5: The correct and wrong HMMs for subject 2.


Figure 6: The correct and wrong HMMs of subject 3.


Figure 7: From the left to the right, the three images show the difference between the fixation maps of correct and the wrong responses of the three subjects shown in Figure 4-6.

From the figures above, we see that in some cases, the key difference between the wrong and correct HMMs can be discovered from the temporal rather than the spatial domain of the data. For instance, for subject 1 , the correct and the wrong HMMs were spatially similar, but the wrong HMM had a different set of prior values and transition probabilities. If the subject started looking at the image from the right eye, the response is more likely to be incorrect. ${ }^{1}$

One disadvantage of comparing fixation maps between correct and wrong responses can be seen from figure 7 above. The pixel test in each case discovered many significantly different regions. These regions are all over the face, which make them very hard to be qualitatively explained.

## Discussion

[^39]In this paper, we have proposed an HMM-based method to analyze eye movement data and demonstrated several advantages.

Firstly, our method can learn the ROIs for each person from the data together with their temporal information. This provides the information for describing and inferring the scan paths. Although fixation maps can be generated by fixations, such that the maps could be used to show the distributional difference of fixations over time, they do not contain transition information so that describing and inferring scan paths are impossible.

Secondly, using VHEM, the HMMs can be grouped into clusters based on their similarities. Our finding of this clustering showed that participants demonstrated either a holistic strategy or an analytic strategy. The two strategies were significantly different from each other.

Lastly, by comparing the correct and the wrong HMMs, we showed that the 'correct' eye movements were significantly different from the 'wrong' eye movements, and that the difference to a considerable extent can be attributed to the transition differences instead of spatial distribution differences. Comparison of the fixation maps of correct and wrong responses also showed the differences between the 'correct' and 'wrong' eye movements, but the differences were too spread so that the results lacked identifiable patterns. Also, the fixation map method was not able to show the difference in transition probability between eye movements in correct and wrong trials.
The lack of empirical findings to support the scan path theory caused eye movement researchers' lack of interest in sequential information (Henderson, 2003). Our findings, however, suggest that sequential information could be associated with performance. Theoretically, given a chain of fixations, using the two HMMs, the accuracy of the response can be predicted. This further justifies using HMMs to describe and analyze eye movement patterns. Future work will test this hypothesis.

In the current study, we pooled all the fixations together to find the ROIs because we assumed that the ROIs are the same across fixations. An alternative approach that does not rely on this assumption is to train the GMMs by fixation, so that at each fixation, there are a unique set of ROIs. An HMM in this case will have time-dependent states. For future research, we attempt to investigate this further.

In summary, here we show that eye movements can be better studied and understood using HMMs. With HMMs, we can describe both the spatial and the sequential aspects of eye movements. We also show that clustering the HMMs can yield interesting between-group differences. The two subgroups roughly correspond to more holistic and more analytic strategies. We further show that correct and wrong recognitions have different eye movement patterns and that the differences can be found in the transition probabilities.

## Acknowledgements

A.B. Chan acknowledges support from the Research Grants Council of the Hong Kong SAR, China (CityU
110610). We thank Tina Liu for collecting the data for the study. We are also grateful to the Research Grant Council of Hong Kong (project code HKU 745210H to J.H. Hsiao) and the HKU Seed Funding Program for Basic Research (project code 201011159124 to J.H. Hsiao).

## References

Baber, D. (2012). Bayesian reasoning and machine learning. Cambridge: Cambridge University Press.
Barton, J. J. S., Radcliffe, N., Cherkasova, M. V., Edelman, J., \& Intriligator, J. M. (2006). Information processing during face recognition: the effects of familiarity, inversion, and morphing on scanning fixations. Perception 35, 1089-1105.
Bishop, C. M. (2006). Pattern recognition and machine learning. New York: Springer.
Caldara, R. \& Miellet, S. (2011). iMap: a novel method for statistical fixation mapping of eye movement data. Behavior research methods 43, 864-878.
Coviello, E., Chan, A. B., \& Lanckriet, G. R. G. (2012). The variational hierarchical EM algorithm for clustering hidden Markov models. In Neural Information Processing Systems (NIPS).
Goldberg, J. H. \& Helfman, J. I. (2010). Scanpath clustering and aggregation. Proceedings of the 2010 Symposium on Eye-Tracking Research \& Applications, 227-234.
Gosselin, F. \& Schyns, P. G. (2001). Bubbles: a technique to reveal the use of information in recognition tasks. Vision Research 41, 2261-2271.
Henderson, J. M. (2003). Human gaze control during realworld scene perception. TRENDS in Cognitive Sciences 7, 498-504.
Henderson, J. M., Williams, C. C., \& Falk, R. J. (2005). Eye movements are functional during face learning. Memory \& Cognition 33, 98-106.
Hsiao, J. \& Cottrell, G. (2008). Two fixations suffice in face recognition. Psychological Science 19, 998-1006.
Kelly, D. J., Jack, R. E., Miellet, S., De Luca, E., Foreman, K., \& Caldara, R. (2011). Social experience does not abolish cultural diversity in eye movements. Frontiers in Psychology 2, 1-11.
Land, M. F. (2011). Oculomotor behavior in vertebrates and invertebrates. In S. Liversedge, I. Gilchrist, \& S. Everling (Eds.), Oxford handbook of eye movements. Oxford: Oxford University Press.
Mannan, S. K., Ruddock, K. H., \& Wooding, D. S. (1997). Fixation patterns made during brief examination of two dimensional images. Perception, 26, 1059-1072.
Pelphrey, K. A., Sasson, N. J., Reznick, S., Paul, G., Goldman, B. D., \& Piven, J. (2002). Visual scanning of faces in autism. Journal of Autism and Developmental Disorders 32, 249-261.
Wade, N. J., \& Tatler, B. W. (2011). Origins and applications of the eye movement research. In S. Liversedge, I. Gilchrist, \& S. Everling (Eds.), Oxford handbook of eye movements. Oxford: Oxford University Press.

# Uniquely human developmental timing may drive cerebral lateralization and interhemispheric collaboration 

Ben Cipollini (bcipolli@cogsci.ucsd.edu)<br>Department of Cognitive Science, 9500 Gilman Dr<br>La Jolla, CA 92093 USA<br>Garrison Cottrell (cottrell@eng.ucsd.edu)<br>Department of Computer Science and Engineering, 9500 Gilman Dr<br>La Jolla, CA 92093 USA


#### Abstract

Cerebral lateralization is intertwined with virtually every cognitive function that we think makes us human. Yet a clear dichotomy has never been explained: lateralized processing suggests independent, local development of neural circuits, but the complementary nature of lateralized functions and extremely strong functional coupling between homologous areas suggest robust interhemispheric interactions. Here, we review literature and present modeling evidence that this dichotomy can be explained by the uniquely steep trajectory of human post-natal brain growth. This drastic volumetric change cause most long distance, interhemispheric connections to be more unreliable than shorter, intrahemispheric connections, leading to lateralization. Strong interhemispheric collaboration is enabled by the later maturation and myelination of long-distance callosal connections. We also review and reanalyze a well-cited modeling paper (Ringo, Doty, Demeter, and Simard (1994)) thought to show a relationship between the degree of hemispheric coordination and length of conduction delays, showing that previous claims have a clear alternative explanation.


Keywords: corpus callosum; lateralization; asymmetry; conduction delays;

## Introduction

A single concept, supported by a single paper, has dominated thought as to the origins of cerebral lateralization. The modeling work of Ringo et al. (1994) has been exclusively and extensively cited to support the notion that large magnitude conduction delays, due to the large human brain size, enable cerebral lateralization. This delay magnitude hypothesis has intuitive appeal, as it supports another long-held notion: that some combination of large brains and functional lateralization have made us human.

There is no denying the importance of functional lateralization in human cognitive abilities; we are functionally lateralized in virtually all cognitive functions that we think are special to our species, including language, high-precision manual use of tools, spatial processing abilities, and even our emotional processing (Gazzaniga, 2000; Craig, 2005). There is also no denying, however, that the lateralized hemispheres are also tightly coupled in terms of both their complementary abilities (Gazzaniga, 2000; Hellige, 2006) and their functional coupling (Stark et al., 2008). This dichotomy is simply not captured by the delay magnitude hypothesis. Nor does the hypothesis account for the anatomical and functional asymmetries that appear throughout the animal kingdom (Rogers \& Andrew, 2002; Rogers, 2009) in organisms with small
brains. Whether or not the delay magnitude hypothesis is correct, it certainly is not complete.

We hypothesize that functional lateralization is not caused by the magnitude of conduction delays. First, we'll review literature that supports our hypothesis. We'll present a reanalysis of Ringo et al. (1994) that severely restricts the scope of their results. We'll propose a new hypothesis that the developmental trajectory of human brains enables functional lateralization-specifically, that the vast and accelerated postnatal expansion of brain size and delayed maturation of the corpus callosum causes unreliable timing of interhemispheric information in pre-adult humans. We'll review literature supporting this hypothesis, then we'll present our own model supporting the plausibility of our developmentalhypothesis. Finally, we'll summarize our results and discuss implications of our findings to the general phenomena of lateralization, asymmetry, and cognition.

## The failure of conduction delay magnitude

Callosal axons are especially long in humans, due to their need to traverse through our large, highly gyrified brains to connect to the opposite hemisphere. Because the average conduction velocity of axons does not sufficiently compensate for the additional axon lengths when compared to smallerbrained animals, the resulting interhemispheric transmission delay over the majority of callosal axons is longer in human brains. The delay magnitude hypothesis suggests that this increased delay would cause less interhemispheric collaboration and therefore enable cerebral asymmetry.

While the anatomy and physiology of callosal axons is well-established, their seemingly intuitive effects on interhemispheric collaboration is supported by a single model in a single paper (Ringo et al., 1994). Here, we argue against the delay magnitude hypothesis in two parts. First, we present 4 results from the literature that are inconsistent with the delay magnitude hypothesis. Second, we show that the model itself does not support the hypothesis.

## 1. Increased interhemispheric collaboration is associated with an increase in slow fibers

Larger corpus callosum size is associated with less lateralization. This is true for regions of the corpus callosum, as well as the corpus callosum as a whole. The midbody of the callosum, which carries fibers to and from motor
cortex, is larger for individuals with less lateralization in handedness (Witelson, 1989; Luders et al., 2010). Callosal cross-sectional area is proportionally larger for left-handers (Witelson, 1985), who show less functional lateralization than right-handers.

Within humans, larger corpus callosum size is associated with a larger number of thin fibers, not with the thickness of fibers (Aboitiz, Scheibel, Fisher, \& Zaidel, 1992). This suggests that those with more interhemispheric collaboration have significantly more slow fibers-just the opposite of what the delay magnitude hypothesis would predict.

## 2. Homotopic areas show functional coupling

The corpus callosum largely connects corresponding (homotopic) areas between left and right cerebral hemispheres. Thus, according to the delay magnitude hypothesis, homotopic areas connected with slow, thin fibers (Aboitiz \& Montiel, 2003) should show weak functional connectivity. In fact, this is not the case at all. For example, when examining interhemispheric correlations through resting-state fMRI, Stark et al. (2008) found very strong interhemispheric correlations between association areas. Reduced interhemispheric coherence (measured with EEG) at locations away from primary sensory/motor cortices has been measured in mental disabilities or diseases, such as dyslexia (Dhar, Been, Minderaa, \& Althaus, 2010) and schizophrenia (Hoptman et al., 2012).

Hellige (2006) points out that functional specializations tend to be complementary. For example, visual processing of the left hemisphere seems biased towards high frequency processing, while the right hemisphere seems biased towards low-frequency processing (Sergent, 1982; Ivry \& Robertson, 1998). If there is less interhemispheric integration due to more independent processing, then why would the two hemispheres show any type of relationship at all? The delay magnitude hypothesis offers no answer.

## 3. Longer delays may support coordination

The corpus callosum in larger brains doesn't simply have longer conduction delays; it also has a broader range of conduction delays. Innocenti (2011) reviewed data suggesting that a broader range of conduction delays supports a broader range of oscillations across the corpus callosum (Caminiti, Ghaziri, Galuske, Hof, \& Innocenti, 2009), which may increase the stability of those oscillations (Roberts \& Robinson, 2008). The current belief is that these oscillations are necessary for binding of information between two distant cortical areas (Fries, 2005); stabilization of inter hemispheric oscillations would presumably enhance interhemispheric communication. Thus, longer delays may be associated with improved ability to coordinate interhemispheric integration.

## 4. Shorter delays are detrimental in development

Many ideas of how the human brain may be unique have been debunked, including suggestions that the human brain is specially gyrified or has a unique fundamental asymmetry. Human brains are clearly unique developmentally-as precocial
mammals (born with our eyes open), we are the only species known to extend the accelerated rate of prenatal brain growth well beyond birth (Martin, 1983). This means that the rate of brain growth is especially high in humans.

Lewis and Elman (2008) used a version of Ringo et al.'s model to show that, the steeper the developmental brain growth curve, the more detrimental interhemispheric connections are to learning. This is due to the fact that, as brain size changes more quickly, the conduction delays change more as well, and those larger changes are more detrimental to learning. As their model "matured", even though the magnitude of delays were longer, because they were more stable, they promoted interhemispheric collaboration.

The delay magnitude hypothesis only addresses mature, adult brains. We suggest that taking a developmental angle to this problem may give more general results.

## 5. Delays only affect the onset of communication



Figure 1: (a) The model architecture of Ringo et al. (1994). Information flows from bottom to top; left model hemisphere is to the left, and right model hemisphere is to the right. Arrows represent full connections between pre- and post-synaptic units. All delays are 1 time-step, except the interhemispheric ("callosal") connections, whose delay were varied across conditions. Note the shared output nodes, which allow an (unintended) path for fast interhemispheric coordination independent of the "callosal" connections.
(b) The model architecture of Lewis and Elman (2008) simplifies the structure and splits the inputs and outputs.

## The model failed to control all interhemispheric transfer

 Although Ringo et al. aimed to separate interhemispheric communication through long conduction delays, their model setup failed to do so (Fig. 1a). In addition to their "callosal" connections that were varied with short and long delays, their model also had converging connections from the hemispheres to a shared bank of output nodes, whose delays were always short. Thus, even if they re-trained their models without any "callosal" connections, the hemispheres would still show interhemispheric dependence; one hemisphere would not be able to complete the task without the other. ${ }^{1}$ This issue is[^40]an important confound in interpreting their results.
This issue was addressed in the only paper to follow-up the Ringo et al. study, by simply splitting the output nodes into two separate banks (Lewis \& Elman, 2008), as depicted in Fig. 1b. All modeling work in this paper uses this same split-output architecture.

## The results are often misunderstood and misinterpreted

Fig. 3 describes the Ringo et al. methods; Fig. 3a shows the original results. Citations to this paper are often made to support the notion that functional lateralization is inevitable, given the human brain size. This is a misrepresentation of the Ringo et al. results. In the paper, the authors only claim that lateralization at short settling times is caused by long delays. Tasks that allow "multiple passes" across the callosum were interpreted to show indistinguishable results across delays.

In fact, the original Ringo et al. results do not show anything except a static delay in interhemispheric coordination, of exactly the value of a single pass across the model corpus callosum. When we transpose the results from the two models by this value ( 9 time steps), the two models are indistinguishable (see Fig. 2 for details). This suggests that the only change in interhemispheric interactions found by their model is a simple, static delay in onset of interhemispheric communication, of a value equal to the time it takes for information to move from one hemisphere to the other. Note that this onset of activity may be mediated by "gigantic" callosal fibers-the largest $0.1 \%$ of fibers that do vary with brain size (Olivares, Montiel, \& Aboitiz, 2001; Wang, 2008), an effect not captured in this model.

The model does not provide any evidence of qualitatively reduced interhemispheric interaction, only weak evidence for a (slightly) delayed one.

## Our hypothesis: changes in timing reliability

The delay magnitude hypothesis fails to explain the basic dichotomy of how functional areas become both lateralized and functionally coupled. The developmental time-course of the thin callosal fibers suggest a two stage process to us: an initial stage where all associative / pre-frontal white-matter connections are immature, favoring local processing, and then a later stage were white-matter connections mature and come on-line. This pattern is well-supported in white matter in general, with some support in the corpus callosum as well (e.g. the anterior, frontal portion) (Jernigan, Baar, Stiles, \& Madsen, 2011). Current imaging technologies can only detect particular types of maturational changes, with those that we outline below (relatively small changes in fiber diameter) currently excluded.

As mentioned above, previous work by Lewis and Elman suggested that unreliability in conduction delays could affect interhemispheric processing. However, though human postnatal brain growth is fast compared to other species, it is still quite slow compared to the time-scale of plasticity in the brain. Based on their work, we suggest that any factor that


Figure 2: (a) Original data from Ringo et al. (1994), showing performance of networks after lesioning interhemispheric fibers, for two networks with different interhemispheric delays ( 1 time-step vs. 10 time-steps). Different networks were required to process across a range of times (x-axis; 15-75 time-steps), while they were trained to output binary strings that were associated with particular input binary strings. After training, "callosal" connections were lesioned, and network performance was measured. The network with the shorter interhemispheric delays ( $\mathrm{D}=1$; empty triangles) shows poorer performance on networks running for fewer time-steps ( x -axis $=15-30$ time-steps); this was interpreted as indicating less interhemispheric interaction.
(b) We expect a network with delay=1 and delay=10 to have a difference of 9 time-steps to the onset of hemispheric interaction. We shifted the $\mathrm{D}=1$ curve by 9 time-steps later (right on the x -axis) to allow us to visualize any qualitative difference in the interhemispheric interaction outside of this difference in onset. The overlap of the curves suggest that there is no other variation in interhemispheric communication besides this simple static delay.
disrupts the reliability of timing between cortical areas will have a detrimental effect on their coordination.

In reviewing the developmental literature, we did find one source of variability in the coordination of timing that is relevant to the corpus callosum: unreliable conduction delays in unmyelinated fibers with a thickness less than $0.5 \mu \mathrm{~m}$ (Wang, 2008). Interestingly, in all adult animal species, callosal fibers are rarely found with a thickness below $0.5 \mu \mathrm{~m}$ (Aboitiz \& Montiel, 2003). However, neonates have a preponderance of such fibers ${ }^{2}$ (Berbel \& Innocenti, 1988; LaMantia \& Rakic, 1990). While some of these small-diameter, unmyelinated fibers persist into adulthood, many become myelinated and all become more reliable by increases in their diameter (Aboitiz \& Montiel, 2003).

Thus, we have a mechanism that fits all of our criteria: initially thin axon diameters decrease the reliability of the timing of information on a timescale relevant to neural processing (individual spikes), and follow a developmental trajectory that would initially support more independence (through less

[^41]reliable timing), with developmental maturation tending towards interhemispheric collaboration (more reliable timing).

We hypothesize that association areas develop with decreased interhemispheric contributions, due to the unreliability of interhemispheric signals through small, unreliable fibers. This allows for the hemispheres to develop independently, which enables asymmetries to develop. Developmental changes in these fibers (diameter increases and myelination) make them reliable, and interhemispheric coordination comes online. In the discussion section, we'll expand this hypothesis to show exactly how it can address each of the questions laid out above.

## Methods

We implemented a version of Lewis and Elman's model (Fig. 1b), which uses rate-coded leaky-integrator units. We suggest this is plausible: asymmetries are linked to higher-order cortical areas (Sergent, 1982; Schenker, Sherwood, Hof, \& Semendeferi, 2007) which tend to interconnect over the corpus callosum using slow fibers (Aboitiz \& Montiel, 2003) that are suggested to use rate-coding, rather than spike-time coding employed by thicker, faster fibers (Wang, 2008).

All connections in the model carry a delay; in all simulations cited and implemented, intrahemispheric delays are set to 1 time-step. Each hemisphere consists of 5 input units, fully connected to 15 hidden units. The hidden units have full recurrent self-connections, as well as full feed-forward connections to 5 output units. 3 hidden units from each hemisphere connected fully and reciprocally to each other as a model "corpus callosum"; these were the only shared connections between the hemispheres ${ }^{3}$. For all simulations, only these interhemispheric connections were manipulated.

As in the previous studies, the task for the network was to learn associations between input binary strings and output binary strings. We used a version of backpropagation through time appropriate for learning with conduction delays (Pearlmutter, 1989), for calculating our error gradients, and used resilient backpropagation for computing our gradient updates (Riedmiller \& Braun, 1993). We used the sumsquared error function function ${ }^{4}$.

We found that learning in the networks with published learning rates was slow; we also found that the degree of interhemispheric communication was dependent on parameters that were not varied in each study. We chose parameters to optimize learning speed, while balancing between interhemispheric and intrahemispheric dependencies ${ }^{5}$. Since the purpose of this study was to examine changes in intrahemispheric and interhemispheric processing, and not their actual magnitude, this seemed a reasonable approach.

[^42]For this study, we set a fixed total time (30 time-steps) and interhemispheric delay ( 10 time-steps).

## Experimental Setup

We measured two values for performance: the classification error was the percentage of output nodes that were not within 0.5 of their expected output value ( +1 or -1 ), and the training error was the average (sum-squared) error at each output node. For each of these values, the lesion-induced error was computed as difference between performance of the intact and lesioned network, as was done in Lewis and Elman. Again following that paper, our dataset contained both intrahemispheric patterns ( $50 \%$ of the input patterns), one hemisphere could determine its output without receiving any information from the other, as well as interhemispheric patterns ( $50 \%$ of the input patterns), one hemisphere had to receive information from the other hemisphere to choose between 4 possible output strings. This allowed a more nuanced analysis of network performance differences.

Importantly, we operationalized unreliable conduction delays as Gaussian noise of the activity (instantaneous firing rate) transmitted over fixed (reliable) delays. In a rate-coding system, variation in the arrival of individual spikes, or a missing spike, leads to jitter in the instantaneous firing rate. We implemented this jitter directly in our rate-coded network as Gaussian noise on the activity, on a per-synapse basis. The jitter was a function of the delay at each synapse ${ }^{6}$.

We ran 25 no-noise networks first (without any noise introduced on interhemispheric connections), to establish baseline measures. The networks were trained until they had zero classification error or until 1000 training epochs elapsed. Every 100 epochs, we measured error in the in-tact networks, as well as lesion-induced error. We then duplicated this procedure for 25 noise networks, which were identical except for having random Gaussian noise ( $2 \%$ of average unit activity) injected on the interhemispheric connections.

We thought that introducing Gaussian noise would cause interhemispheric information to be less reliable than intrahemispheric information, causing intra-hemispheric patterns to be learned more independently of the other hemisphere, and delaying learning of interhemispheric patterns. Therefore, we predicted that (1) the learning trajectory of the network with interhemispheric noise would be more gradual (i.e. have a smaller slope), and asymptote earlier. We also predicted that lesion-induced error would be lower in the noise vs. no-noise networks.

## Results

Fig. 3a shows learning trajectories of classification error, for no-noise and noise networks on both in-tact and lesioned conditions. As predicted, learning in the noise networks was slower and reached asymptote at a higher error than the nonoise networks. We show these results for comparison to pre-

[^43]vious papers. Fig. 3b shows the learning trajectories for learning error, on the same set of networks. Notice that, consistent with our predictions, lesion-induced error (the difference between corresponding intact and lesioned curves) is smaller for the noise networks vs. the no-noise networks on both measures. Interestingly, for learning error (a more nuanced measure of network performance), noise networks had less error in the lesioned networks than the no-noise networks.


Figure 3: Changes in (a) classification error and (b) training error over training epochs, for noise and no-noise networks in-tact and lesioned conditions.

In order to examine these results more closely, we computed lesion-induced error on for the noise and no-noise networks for training error, then separated them into intrahemispheric and interhemispheric patterns (Fig. 4a). The noise networks showed less lesion-induced error for both interhemispheric and intrahemispheric patterns. Looking more closely at the differences between noise and no-noise networks (Fig. 4b), we find the surprising result that noise networks had much less lesion-induced error for intrahemispheric patterns than the non-noise networks, while the two had relatively equal levels of lesion-induced error for interhemispheric patterns.

These results indicate that the network with noisy interhemispheric fibers tried to accomplish the task (as much as possible) intrahemispherically, particularly in cases where both hemispheres are necessary to complete a task.

## Discussion

We argued that current thought on interhemispheric integration fails to explain a basic dichotomy: how interhemispheric segregation may be necessary for developing asymmetries, but must be overcome to produce interhemispheric coupling found in adults. We showed that ideas based on the magnitude of conduction delays cannot explain interhemispheric segregation, nor can they explain strong interhemispheric coupling. We suggested that the especially steep developmental gradient of humans may hold the key. Our literature review revealed that two properties of these fibers might cause such a pattern: their changing length and their changing reliability in timing. The latter is relevant on the necessary timescale, so we focused our work here on examining the effects of changes in the reliability of the timing of information.


Figure 4: (a) Lesion-induced error for both noise and nonoise networks, split into interhemispheric and intrahemispheric patterns. (b) Difference between noise and no-noise networks for lesion-induced error; positive values mean more lesion-induced error in no-noise networks.

The results of our computational experiments showed that timing unreliability, in the form of Gaussian noise of our instantaneous firing rate, could induce more independent development of the cerebral hemispheres. This effect need not be related to the magnitude of conduction delay-but in mammalian brains it is: longer fibers have both longer delays and more jitter in their timing. These effects are also temperaturedependent (Wang, 2008); perhaps a mechanism leading to the large number of asymmetries found in cold-blooded species (Rogers \& Andrew, 2002).

Our simulations also hinted at how asymmetry may emerge. We found that noisy callosal fibers led networks to try and use local, intrahemispheric processing, even for patterns requiring interhemispheric information. This type of early local processing is an indication that these networks may encourage developmental asymmetries.

Finally, we note a few recent papers on the benefits of noise in learning (Ermentrout, Galn, \& Urban, 2008; Faisal, Selen, \& Wolpert, 2008; Vincent, Larochelle, Lajoie, Bengio, \& Manzagol, 2010). We intend to investigate whether initially noisy interhemispheric interactions facilitate both generalization and specialization of the hemispheres. Shared processing of highly salient features may allow each hemisphere to select secondary features that it is more specialized to process.

## Acknowledgments

This work was partly funded by a Center for Academic Research and Training in Anthropogeny (CARTA) fellowship and by NSF grant SMA 1041755 to the Temporal Dynamics of Learning Center, an NSF Science of Learning Center.

## References

Aboitiz, F., \& Montiel, J. (2003). One hundred million years of interhemispheric communication: the history of the corpus callosum. Brazilian journal of medical and biological research, 36(4), 409-20.
Aboitiz, F., Scheibel, A. B., Fisher, R. S., \& Zaidel, E. (1992).

Fiber composition of the human corpus callosum. Brain Research, 598(1-2), 143-153.
Berbel, P., \& Innocenti, G. M. (1988). The development of the corpus callosum in cats: a light- and electronmicroscopic study. The Journal of comparative neurology, 276(1), 132-156.
Caminiti, R., Ghaziri, H., Galuske, R., Hof, P. R., \& Innocenti, G. M. (2009). Evolution amplified processing with temporally dispersed slow neuronal connectivity in primates. Proceedings of the National Academy of Sciences, 106(46), 19551-19556.
Craig, A. B. (2005). Forebrain emotional asymmetry: a neuroanatomical basis? Trends in Cognitive Sciences, 9(12), 566-571.
Dhar, M., Been, P. H., Minderaa, R. B., \& Althaus, M. (2010). Reduced interhemispheric coherence in dyslexic adults. Cortex, 46(6), 794-798.
Ermentrout, G. B., Galn, R. F., \& Urban, N. N. (2008). Reliability, synchrony and noise. Trends in neurosciences, 31(8), 428-434.
Faisal, A. A., Selen, L. P. J., \& Wolpert, D. M. (2008). Noise in the nervous system. Nature Reviews Neuroscience, 9(4), 292-303.
Fries, P. (2005). A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. Trends in Cognitive Sciences, 9(10), 474-480.
Gazzaniga, M. S. (2000). Cerebral specialization and interhemispheric communication: Does the corpus callosum enable the human condition? Brain, 123(7), 1293-1326.
Hellige, J. B. (2006). Evolution of brain lateralization in humans. Cognition, Brain, Behavior, 10(2), 211-234.
Hoptman, M. J., Zuo, X.-N., D'Angelo, D., Mauro, C. J., Butler, P. D., Milham, M. P., et al. (2012). Decreased interhemispheric coordination in schizophrenia: A resting state fMRI study. Schizophrenia Research, 141(1), 1-7.
Innocenti, G. M. (2011). Development and evolution: two determinants of cortical connectivity. Progress in brain research, 189, 65-75.
Ivry, R. B., \& Robertson, L. C. (1998). The two sides of perception. The MIT Press.
Jernigan, T. L., Baar, W. F. C., Stiles, J., \& Madsen, K. S. (2011). Postnatal brain development: structural imaging of dynamic neurodevelopmental processes. Progress in brain research, 189, 77-92.
LaMantia, A. S., \& Rakic, P. (1990). Axon overproduction and elimination in the corpus callosum of the developing rhesus monkey. The Journal of neuroscience, 10(7), 21562175.

Lewis, J. D., \& Elman, J. L. (2008). Growth-related neural reorganization and the autism phenotype: a test of the hypothesis that altered brain growth leads to altered connectivity. Developmental Science, 11(1), 135-155.
Luders, E., Cherbuin, N., Thompson, P. M., Gutman, B., Anstey, K. J., Sachdev, P., et al. (2010). When more is less: associations between corpus callosum size and hand-
edness lateralization. NeuroImage, 52(1), 43-49.
Martin, R. D. (1983). Human brain evolution in an ecological context. American Museum of Natural History.
Olivares, R., Montiel, J., \& Aboitiz, F. (2001). Species differences and similarities in the fine structure of the mammalian corpus callosum. Brain, Behavior and Evolution, 57(2), 98-105.
Pearlmutter, B. A. (1989). Learning state space trajectories in recurrent neural networks. Neural Computation, 1(2), 263-269.
Riedmiller, M., \& Braun, H. (1993). A direct adaptive method for faster backpropagation learning: the RPROP algorithm. In, IEEE international conference on neural networks, 1993 (pp. 586-591). IEEE.
Ringo, J. L., Doty, R. W., Demeter, S., \& Simard, P. Y. (1994). Time is of the essence: A conjecture that hemispheric specialization arises from interhemispheric conduction delay. Cereb. Cortex, 4(4), 331-343.
Roberts, J. A., \& Robinson, P. A. (2008). Modeling distributed axonal delays in mean-field brain dynamics. Physical review. $E$, 78(5 Pt 1), 051901.
Rogers, L. J. (2009). Hand and paw preferences in relation to the lateralized brain. Phil. Transactions of the Royal Society B: Biological Sciences, 364(1519), 943-954.
Rogers, L. J., \& Andrew, R. (2002). Comparative vertebrate lateralization (1st ed.). Cambridge University Press.
Schenker, N. M., Sherwood, C. C., Hof, P. R., \& Semendeferi, K. (2007). Microstructural asymmetries of the cerebral cortex in humans and other mammals. In Special topics in primatology (Vol. 5, pp. 92-118).
Sergent, J. (1982). The cerebral balance of power: confrontation or cooperation? Journal of Experimental Psychology. Human Perception and Performance, 8(2), 253-272.
Stark, D. E., Margulies, D. S., Shehzad, Z. E., Reiss, P., Kelly, A. M. C., Uddin, L. Q., et al. (2008). Regional variation in interhemispheric coordination of intrinsic hemodynamic fluctuations. The Journal of neuroscience, 28(51), 1375413764.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., \& Manzagol, P.-A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res., 11, 33713408.

Wang, S. S.-H. (2008). Functional tradeoffs in axonal scaling: Implications for brain function. Brain, Behavior and Evolution, 72(2), 159-167.
Witelson, S. F. (1985). The brain connection: the corpus callosum is larger in left-handers. Science, 229(4714), 665668.

Witelson, S. F. (1989). Hand and sex differences in the isthmus and genu of the human corpus callosum a postmortem morphological study. Brain, 112(3), 799-835.

# Systems from Sequences: an Iterated Learning Account of the Emergence of Systematic Structure in a Non-Linguistic Task 

Hannah Cornish (hannah@ling.ed.ac.uk)<br>Kenny Smith (kenny@ling.ed.ac.uk)<br>Simon Kirby (simon@ling.ed.ac.uk<br>University of Edinburgh, School of Philosophy, Psychology and Language Sciences, Language Evolution and Computation Research Unit, 3 Charles Street Edinburgh, EH8 9AD, UK


#### Abstract

Systematicity is a basic property of language and other culturally transmitted behaviours. Utilising a novel experimental task consisting of initially independent sequence learning trials, we demonstrate that systematicity can unfold gradually via the process of cultural transmission.


Keywords: Iterated Learning; Cultural Evolution; Sequence Learning

## Introduction

Language, like many other culturally-transmitted aspects of human behaviour, works as a system: individual words or phrases do not behave independently of one another but instead form part of a mutually reinforcing system of conventions. This is at the heart of what we mean when we talk about the grammar of language - grammar exists only to the extent that individual utterances are non-independent, and related to one another in systematic ways. This is so obviously true of language, and other cultural systems such as music, that it hardly seems to stand in need of explanation. But where does this basic fundamental property come from? Why are individual utterances not independent? After all, they arguably are for all other systems of communication in nature. Even when we discount the fact that non-human animals can only convey a finite set of meanings, many animal communication signals are gradable and related to one another only as much as they are produced by the same vocal apparatus (Fitch, 2010). This stands in stark contrast to human language, resting as it does on a system of infinitely reusable discrete signals ${ }^{1}$.

We propose that the answer lies in the nature of cultural transmission. In systems like language which are transmitted by iterated learning ${ }^{2}$, the most transmissible behaviours are those that are most learnable (Smith, Kirby, \& Brighton, 2003). Cultural evolution tends therefore to maximise learnability (Christiansen \& Chater, 2008; Kirby, Cornish, \& Smith, 2008). One way to increase learnability of a set of behaviours is for those behaviours to behave systematically, so

[^44]that learning one will increase the ease with which others will be learned. In other words, we should expect cultural evolution to create systems of dependence between previously independent learned behaviours.

In this paper we present an experimental paradigm in which we can observe the cultural evolution of such systematicity in a task which involves many initially independent learning trials. Our task is purposefully non-linguistic, but designed to have relevant similarities with language. Specifically, it is a simple immediate sequence-recall task based around the Simon Game. This was a children's electronic game developed by Milton-Bradely in 1978 with four coloured illuminated buttons arranged on its surface in a circle. These buttons lit up to display a random sequence and the player's goal was to repeat this sequence back immediately. This task has a number of useful properties for our purposes. Firstly, although it is clearly non-linguistic, thereby making it unlikely that participants will bring any language-specific biases to it, the task nevertheless involves sequence learning, which is highly relevant to the linguistic domain (Misyak, Christiansen, \& Tomblin, 2010; Christiansen, Conway, \& Onnis, 2012). Secondly, the task is overtly one in which each sequence acts as an independent task. The player can be scored on their learning of each sequence immediately after recall. This is in contrast to a typical artificial language learning task (e.g., Gomez \& Gerken, 2000) which might involve learning a set of sequences for recall at a later stage.

Our question is the following: given this kind of simple independent sequence recall task, will cultural transmission nevertheless lead to the evolution of systematicity in the set of sequences? In effect, can an implicit system-wide learning effect exert influence on the evolution of the set of sequences? To test this, we create an iterated version of the Simon Game in which the sequences produced by one participant in the task become the sequences that the next participant in the experiment is exposed to. We start with a set of 60 random sequences, and observe whether these sets evolve in such a way to make learning easier, and whether they do so by becoming more systematically structured.

## Methods

The experiment utilises a diffusion chain paradigm, a technique used widely amongst researchers investigating cultural
transmission (e.g., Mesoudi \& Whiten, 2008), whereby all learners (apart from those in the initial generation) are trained on the output of previous learners.

## Participants

In total, 40 participants (mean age $21 \mathrm{y}, 11 \mathrm{~m}$; females $=25$ ) were recruited from the University of Edinburgh's graduate employment service, to take part in a visual memory experiment involving sequences of flashing coloured lights. Each participant was allocated at random into one of four different chains (A, B, C, D), consisting of ten generations each. All learners received $£ 7$ remuneration for taking part.

## Procedure

The task itself was simple: participants were shown a light sequence on a touch-screen tablet device, and then asked to immediately reproduce it by tapping the sequence back. The layout of the Simon Game is shown in Figure 1. Once a complete sequence has been input (Figure 1.a), immediate accuracy feedback is given (Figure 1.b). Participants could then request another sequence.


Figure 1: A diagram showing the layout of the Simon Game: (a) participants see a sequence on the screen, and are asked to immediately reproduce it; (b) feedback is then given on the task.

In all, participants were trained and tested on 60 different sequences, seeing each sequence once in each of two rounds in random order, making 120 exposures in total. In order to catch obvious mistakes in sequence entry, if any participant submitted a sequence of length 6 or shorter, this was rejected by our software, and the target sequence would reappear at a random point later in the player's round for them to reattempt. The 60 sequences produced in the second round were collected to be used as training stimuli for the next learner in the chain.

## Initial Sequences

Although subsequent learners were trained on the output of the previous learner, the four initial participants were trained on a set of sequences that adhered to the following properties: (i) the length of each sequence was 12 ; (ii) each sequence consisted of 3 flashes of each colour (red, blue, green, yellow); (iii) these colours appeared in random order. This resulted in a set of 60 sequences which had no structure.

## Results

The sequences were analysed in order to determine (1) whether the individual sequences would adapt to become easier to learn over time, and (2) whether individual sequences would co-evolve together to form a collective system. In order to assess these effects, we look at quantitative measures of learnability and structure, along with an additional measure examining the degree of divergence between the four chains into specific lineages. We also qualitiatively examine some of the evolved sequences at the ends of the chains, and note some striking structural regularities.

## Learnability

To determine the learnability of a sequence set at a given moment in time, we first need a measure of how accurately each sequence is reproduced. For this we calculated the intergenerational error using the Levenshtein (1966) edit-distance between each target sequence and response from the participant, normalised for length of sequence (Kirby et al., 2008) ${ }^{3}$ : we count the minimum number of insertions, deletions and substitutions required to turn one sequence (input) into another (output), dividing this by the length of the longer sequence. From the normalised edit distance of each individual sequence, we then calculate the average error of the sequence set. For consistency with our later analysis we converted this into mean similarity (1-error), shown in Figure 2 below.


Figure 2: Graph showing the average mean similarity score of sequences in each set over generations. Error-bars represent the $95 \%$ confidence intervals across the four chains, here and throughout.

[^45]As Figure 2 indicates, the sequence sets become easier to learn over time: reproduced sequences become more similar to their targets. In order to determine whether this cumulative increase in similarity was significant, we ran Page's (1963) $L$ trend test. This reveals a significant increase in similarity over generations, both when including ( $L=1469, m=4, n=10$, $p<.0001$ ) and excluding ( $L=1074, m=4, n=9, p<.0001$ ) the initial set of sequences, which had not been produced by participants.

## Structure

One possible explanation for the increase we see in learnability could simply be that early participants are forgetting parts of each sequence, leading to the sequences eventually becoming short enough to be more easily reproduced by later learners. In order to assess this claim, we examined the average length of sequences across each chain for any signs of change. Figure 3 confirms that that there was no significant reduction in sequence length over the course of the experiment. Given that length is in fact highly stable across each generation, some other feature of the sequences must be responsible for their increase in learnability.


Figure 3: Graph showing the average length of each sequence by generation. Sequence length remains stable throughout the experiment, ruling out a simplistic explanation for the improvements to performance in Fig. 2.

The other possibility is that the sequences have become structured in some way. In order to determine whether this is the case or not, we examined the composition of the sequences in each set, using two different metrics. The first is a measure of dispersion, which looks at how similar each sequence is to other sequences within that set. This is calculated using the same distance metric as before, this time comparing the distance of each sequence from all other sequences within
that generation, rather than across generations between target and reproduction. This figure, when averaged over all pairs of sequences, returns the amount of dispersion within the set at a given generation. Figure 4 shows that over time, individual sequence sets lose variation as the sequences within them begin to resemble one another ( $L=1980, m=4, n=11$, $p<.0001$ ). This could happen if, for instance, smaller subsequences come to be shared across whole sequences within the set.


Figure 4: Graph showing the mean of the normalised dispersion score of the four sequence sets over generations. Dispersion decreases cumulatively over generations, indicating that sequence sets are becoming more self-similar over time.

In order to explore this idea further, the second measure looks at compression. This is related to the notion of Kolmogorov complexity (Kolmogorov, 1963), and is essentially a measure of how easy it is to compress data into a smaller representation. If a dataset contains repetitions (redundancy), then the algorithm can exploit that by creating a shorter representation to substitute for the larger one, and thus the size of the file can be reduced. We tested this directly by computing the compression ratio (size of the file after compression/size of the file before compression) ${ }^{4}$ in order to assess how much structure was present in each sequence set (Fig. 5).

Figure 5 demonstrates that there is a decrease in the compression ratio over time ( $L=1964.5, m=4, n=11, p<.0001$ ). This shows that the sequences are becoming structured, and further supports the idea that those sequences produced later on in the chains have become fractionated into smaller higherfrequency units which repeat within sequence sets.

[^46]

Figure 5: A graph showing the mean compression ratio over generations. Sequence sets become more compressible over time, indicating that there is more structure in later sequences than in earlier sequences.

## Identifiability

The decrease in dispersion scores and compression ratios across our four chains could be due to a universal bias pulling all of the sequences towards a similar (structured) attractor. If this were the case, we would expect to find that a given sample of sequences drawn from within a chain, at a certain generation, would look fairly similar to any other sample of sequences drawn from any other chain at that same generation. In order to to determine whether there are in fact different types of structural patternings in our data, organised across the different lineages, or just one kind of structural patterning shared amongst all chains, we used a measure of lineage divergence (referred to as identifiability) taken from Matthews, Roberts, and Caldwell (2012). This determines (for each sequence) the within-group similarity and the across-group similarity, and then calculates a proportion: [within-group similarity/(within-group similarity + across-group similarity)]. This returns a value between 0 and 1 , where values above 0.5 indicate higher overall withingroup similarity, and values below 0.5 indicate higher acrossgroup similarity.

Figure 6 shows, the initial sequences (generation 0 ) are not identifiable as coming from their particular chain. This is to be expected due to the fact that they were all randomly constructed according to the same procedure. However, sequences do begin to diverge into separate lineages. A onesample Wilcoxen test confirms that the within-group identifiability of sequences from all chains produced by participants (generations 1-10) were significantly higher than our


Figure 6: Graph showing how the average identifiability of each sequence increases over time. The dashed line represents chance levels.
expected chance level of $0.5(Z=-5.86, N=40, p<.0001)$. Furthermore, our trend analysis reveals that this effect is cumulatively increasing over time ( $L=1901, m=4, n=11, p<.0001$ ), such that sequences drawn from later generations are significantly more likely to resemble sequences from within their own lineage, than from those of any other lineage.

## Qualitiative analysis

The quantitative results all point towards the fact that our independent sequences are gradually becoming structured as a collective, and that they do so in ways which are specific to different lineages. What then might some examples of these systematic structures look like? Figure 7 shows a sample of sequences that came from chain A.

As Figure 7.a shows, the initial set of sequences contain very little obvious structure. By generation 10 however (Figure 7.b), a common pattern has emerged. In fact, of the 60 sequences in this set, just over half of them begin with an initial alternation pattern of red-yellow-red-yellow, or red-red-yellow-yellow. This is frequently followed by a cyclical pattern - moving around the Simon board in either a clock-wise or anti-clockwise direction from a given starting point (usually red in this case) - which can itself be repeated to extend the sequence. This kind of structure lends itself easily to being analysed into hierarchically arranged sub-parts, containing non-adjacent dependencies. Figure 7.c shows one such possible analysis of this kind ${ }^{5}$.

[^47]

Figure 7: Some examples of sequences from chain A: (a) a sample of six random sequences at generation 0 ; (b) those same sequences at generation 10; (c) sequences at generation 10 again, bracketed to highlight their nested hierarchical structure. This bracketing can be used to generate tree-structures (as shown in example v) which more clearly demonstrate the nature of the system.

As to be expected from the identifiability results however, the way in which the other chains are organised is noticeably different, both to the statistical measures employed earlier, and to the human eye. These different styles can be easily contrasted visually in Figure 8, again by drawing a sample of six strings from the set to illustrate general structural regularities in the final generations.


Figure 8: Some examples of sequences from the final generations of chains $\mathrm{B}, \mathrm{C}$ and D . There are clear qualitative differences between the sequences across these different chains, and from chain A in Figure 7 above.

In contrast to the distinctive alternation found in chain A , chain B favours the more cyclical patterns, and shows a very dominant tendency to begin all sequences with a red ( $92 \%$ ). Chain C also places restrictions on the identity of the first colour: green is most common ( $48 \%$ ), followed by yellow ( $30 \%$ ) and red ( $22 \%$ ), but never blue. It also contained the highest proportion of sequences with two or more of the same colour adjacent to one another (58\%: as compared to A $28 \%$, B 35\% and D 23\%). Finally, Chain D seems to prefer triplets and alternations as reusable sub-sequences.

In summary, the qualitative analysis of the chains reveals that sequence-sets can become systematic in multiple ways. Some commonalities do exist across lineages - for instance, a
strategy of repeated alternation of two colours was present in all chains. However, (i) the frequency with which a given strategy was employed, (ii) where it was employed (sequence-initially, sequence-medially, or sequence-finally), and (iii) with what particular colour combinations it was employed, all varied, contributing to the development of a unique 'profile' for each lineage.

## Discussion

We have presented an experiment in which participants attempt to immediately recall visually presented sequences. The sequences that participants produce become the sequences which subsequent participants try to recall. In this way, we create lineages of sequences in an experimental simulation of cultural evolution. These lineages are potentially independent of each other, since the initial set of sequences are generated at random and participant responses are gathered immediately after each sequence.

The effect of cultural evolution in the experiment is that the sequences become easier to recall correctly. In other words, errors introduced by participants are in the direction of easier sequences. How is this achieved? We see that the set of sequences at each generation becomes self-similar, suggesting that the sequences are not operating independently any more. This conclusion is confirmed if we look across separate chains in the experiment: the sequences are more similar within a chain, and less similar across chains. Additionally, the set of sequences at each generation in the experiment becomes more compressible, as system-wide structure starts to emerge.
The systematic structure in sequences shows tantalising evidence of hierarchy, although a deeper analysis will have to await further analytic tools being applied. For example, in some chains we see the emergence of pairs of pairs of colours. We also see a pattern in which some sequences are "doubled" versions of others in the set. So, for example, the sequence prefix "rryyrryy" in chain A matches the prefix "ryry". It is tempting to suggest that this provides evidence of a grammar with centre embedding of the form $\left(A^{n} B^{n}\right)^{m}$, although such an analysis is premature without further probing of the way in which these sequences are processed. Building on work such as Christiansen and Ellefson (2002), recent attempts have been made to tease apart the different cognitive mechanisms at work when processing non-adjacent dependencies resembling these sequences in language (Vries et al., 2012; Christiansen et al., 2012). This may provide further clues as to why these particular structures emerge in this study. Likewise, studying the process of emergence itself, using iterated sequence learning tasks in the laboratory as we have done, may help us better understand the way these learning and processing biases shape behaviour at the population level.

## Conclusion

A hallmark of complex culturally transmitted behaviours in humans, such as language and music, is their systematic
structure. Instances of behaviour do not behave independently, but form part of a system of mutually reinforcing conventions. Here we show that such systematic structure can emerge in an experimental task through the process of cultural transmission even when the task is designed to minimise the influence of domain specific biases and with no explicit reward for treating behaviours as co-dependent. We propose that this result is suggestive of a similar process operating in the origins of behaviours like language in humans. Cultural evolution favours transmissible behaviours. A solution to the challenge of becoming more transmissible is for behaviours to form part of a system, thus increasing their learnability. Language, our most systematic suite of behaviours, bears the hallmark of just such a process of cultural optimisation.

## Acknowledgments

Hannah Cornish is supported by a British Academy PostDoctoral Fellowship.

## References

Christiansen, M. H., \& Chater, N. (2008). Language as shaped by the brain. Behavioral and Brain Science, 31(5), 489-509.
Christiansen, M. H., Conway, C. M., \& Onnis, L. (2012). Similar neural correlates for language and sequential learning: Evidence from event-related brain potentials. Language and Cognitive Processes, 27, 231-256.
Christiansen, M. H., \& Ellefson, M. (2002). Linguistic adaptation without linguistic constraints: the role of sequential learning in language evolution. In A. Wray (Ed.), Transitions to language (p. 335-358). Oxford: Oxford University Press.
Feher, O., Wang, H., Saar, S., Mitra, P., \& Tchernichovski, O. (2009). De novo establishment of wild-type song culture in the zebra finch. Nature, 459, 564-568.
Fitch, W. T. (2010). The evolution of language. Cambridge: Cambridge University Press.
Gomez, R. L., \& Gerken, L. A. (2000). Infant artificial language learning and language acquisition. Trends in Cognitive Science, 4(5), 178-186.
Kirby, S., Cornish, H., \& Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. Proceedings of the National Academy of Sciences of the United States of America, 105, 10681-10686.
Kirby, S., \& Hurford, J. (2002). The emergence of linguistic structure: An overview of the iterated learning model. In A. Cangelosi \& D. Parisi (Eds.), Simulating the evolution of language (p. 121-148). Cambridge: Cambridge University Press.
Kolmogorov, A. N. (1963). On tables of random numbers. Sankhyā Series A., 25, 369-375.
Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics-Doklady, 10, 707-710.

Matthews, C., Roberts, G., \& Caldwell, C. A. (2012). Opportunity to assimilate and pressure to discriminate can generate cultural divergence in the laboratory. Evolution and Human Behavior, 33, 759-770.
Mesoudi, A., \& Whiten, A. (2008). The multiple roles of cultural transmission experiments in understanding human cultural evolution. Philosophical Transactions of the Royal Society B, 363, 3477-3488.
Misyak, J. B., Christiansen, M. H., \& Tomblin, J. B. (2010). Sequential expectations: The role of prediction-based learning in language. Topics in Cognitive Science, 2(138153).

Page, E. B. (1963). Ordered hypotheses for multiple treatments: A significance test for linear ranks. Journal of the American Statistical Association, 58(301), 216-230.
Smith, K., Kirby, S., \& Brighton, H. (2003). Iterated learning: a framework for the emergence of language. Artificial Life, 9(4), 371-386.
Vries, M. H. de, Petersson, K., Guekes, S., Zwitserlood, P., \& Christiansen, M. H. (2012). Processing multiple nonadjacent dependencies: evidence from sequence learning. Philosophical Transactions of the Royal Society B, 367, 2065-2076.

# On the Dynamics of Information Accumulation in Recognition 

Gregory E. Cox, Nicholas J. Lewis, Richard M. Shiffrin<br>(grcox, njlewis, shiffrin)@indiana.edu<br>Department of Psychological and Brain Sciences, Cognitive Science Program, Indiana University<br>1101 E. Tenth St., Bloomington, IN 47405 USA


#### Abstract

Inspired by a dynamic approach to recognition memory (Cox \& Shiffrin, 2012), we present results from a recognition memory experiment in which the time at which diagnostic information arrives is unconsciously varied. Contrary to the predictions of most models, performance improves when diagnostic information is available later, rather than earlier. These results are accounted for by a dynamic model of recognition, where the time at which information starts to be accumulated for a recognition decision can vary independently of when features are available to be sampled from the test display. The same model is shown to be able to reproduce the priming results of Jacoby and Whitehouse (1989), originally attributed to a fluency heuristic. The ability to account for such seemingly disparate results with a single model illustrates the utility of a dynamic approach to recognition.


Keywords: Episodic memory; recognition memory; memory models.

## Introduction

Recognition continues to be a rich source of evidence regarding the processes and mechanisms that underly episodic memory performance. Throughout its long history in psychology and cognitive science, recognition memory experiments have collected measures of reaction time. Despite this, most theories in recognition memory have been concerned only with accuracy. Most of the few models of recognition that also make predictions about response time (Hockley \& Murdock, 1987; Mewhort \& Johns, 2005; Malmberg, 2008; Nosofsky \& Stanton, 2006) assume nonetheless that the evidence is stationary over time (an exception is Brockdorff \& Lamberts, 2000). Thus, it would appear that much work remains to be done to better understand the fine-grained temporal aspects of the recognition process.

As a step in that direction, Cox and Shiffrin (2012) introduced a model of recognition that was based on the gradual accumulation of features over time. As features are sampled, they are added to a memory probe which is then compared to all the traces in memory (or at least those above a certain threshold level of activation), resulting in a "familiarity" value. Familiarity will move up and down over time in a noisy fashion as features get sampled; positive changes in familiarity are evidence in favor of an "old" decision, while negative changes favor a "new" decision. However, because only a finite number of features can be sampled, familiarity will eventually reach a (noisy) asymptote. Thus, the evidence for the recognition process in this model is inherently nonstationary. Furthermore, its predictions will vary greatly with experimental manipulations that affect the timing with which different information becomes available.

There is evidence that, even outside of experimental manipulations to that effect, the nature of the evidence for recognition may vary over time. Information about the "oldness" of individual items is available quite early in processing, while associative information (e.g., whether a word pair was studied in intact or rearranged order) requires approximately an additional 200 ms to become available (Gronlund \& Ratcliff, 1989). And Hintzman and Curran (1994, Experiment 3) found that, when tested with a foil that was a plural or singular form of a word that had been studied in the opposite plurality (e.g., "apple" was studied and "apples" was tested), subjects' tendency to endorse the foil initially increased but then reversed at longer response lags. These results are consistent with a recognition process that accumulates information over time, but at different rates for different kinds of information (e.g., Brockdorff \& Lamberts, 2000).

In an attempt to better understand the dynamics of the recognition process, we first present results from an experiment in which stimuli were constructed from a set of components which varied in diagnosticity as to whether the stimulus is old or new. In some conditions, components became visible at different times, allowing us to assess the effect of presenting diagnostic information later or earlier. These results are explained in the context of a dynamic model of recognition (Cox \& Shiffrin, 2012). The mechanisms employed can also be used to explain the "fluency" results of Jacoby and Whitehouse (1989).

## Experiment 1: Dynamic Presentation

In this experiment, the diagnosticity of information arriving at different times was varied unconsciously.

## Participants

55 undergraduate students from Indiana University participated in the experiment for course credit.

## Stimuli

All stimuli for a given list were generated from two prototypes consisting of random consonant triads, displayed in a triangular manner to minimize the effect of a left-to-right reading preference, e.g., $\mathrm{X}^{\mathrm{L}} \mathrm{K}_{\mathrm{K}}$ and $\mathrm{Z}_{\mathrm{J}} \mathrm{D}_{\mathrm{J}}$. An old item was made from a prototype by replacing one of its letters with another consonant. This resulted in 6 old items for each prototype ( 2 replacements each for the three letters) and a total of 12 old items for study. The prototypes were not studied. New items were generated in a similar manner by replacing a single prototype consonant with a new randomly selected consonant that did not appear at study. This structure allows the sin-


Figure 1: Observed and predicted mean accuracy (A) and mean correct RT (B) for Experiment 1. Error bars denote $95 \%$ confidence intervals.
gle unique consonant to be diagnostic as to whether an item is old and new.

## Procedure

The stimuli were presented centrally on a computer screen. During the study phase, participants were instructed to study the triads as they appeared on the screen and to remember them for a later test of memory. Individual triads were presented in random order for 1 s each with a 1 s blank screen between triads. During the test phase, the participants were instructed to respond whether the presented triad came from the previously studied list. Old and new responses were randomly mapped to the "A" or "L" key for each participant. In the static condition, all letters became visible at the same time. In the dynamic conditions, letters appeared sequentially at a rate of 30 ms (below the threshold for conscious detection) and stayed on until the end of the trial. For new triads, the unique consonant could appear as the first (diagnosticearly) or last (diagnostic-late) letter in the sequence. For old triads, the non-prototype consonant could appear second (diagnostic-early) or last (diagnostic-late). After a recognition judgment was made, the screen was cleared and the next test trial began after 1 s .

## Results

Observed hit rate (HR) and false alarm rate (FAR) are shown in Figure 1A. FAR for dynamic-late is significantly lower than for dynamic-early $(t(54)=2.73, p=0.008)$ or static presentation $(t(54)=3.42, p=0.001)$, which are not significantly different from one another. HR does not differ significantly between conditions. RT for both hits and correct rejections (CR; Figure 1B) is marginally slower in the dynamic conditions than static (for hits, $t(51)=2.04, p=0.05$; for CR, $t(53)=2.06, p=0.04$ ), but otherwise does not differ between conditions.

## A Dynamic Model for Recognition

We now provide a technical description of a model that can account for these effects of dynamic presentation. The model given here is a further development of the one described by Cox and Shiffrin (2012), although the present version is conceptually quite similar and is able to account for the same effects as the original version.

## Structure of Probe and Memory Traces

Events-for example, the study of a memory list item—result in the formation of a memory trace in long-term memory (LTM). Both a memory trace and a memory probe consists of a finite number of features, the number being determined by short-term memory capacity limitations. $N_{x}$ features arise from the context in which the event occurs, for example, the time, location, and internal state of the participant. These features are stable across all study and test trials. There are also $N_{c}$ content features which contain information about the event itself. For example, the memory trace formed from studying a word would include content features related to the word's spelling, phonology, and semantics. For the moment, we do not specify the exact nature of each feature, nor do we assume that the memory system "knows" whether a given feature is a content or context feature. For simplicity, we assume that all features are binary, e.g., " 0 " or " 1 ", with an equal prior probability for each value.

In the full model, different kinds of events can be encoded with different kinds of features. For example, the trace formed from studying a word will contain orthographic, phonological, and semantic features while the trace formed from studying a picture of a face will contain features relating to the shape of the eyes, nose, mouth, etc., and their relative positions. The low degree of featural overlap between traces of different types means that probing with, for example, a word will not tend to activate traces of faces. In this paper, all items in a given experiment are of the same type, so this aspect of the model does not come into play.

## Feature Sampling

Prior to the presentation of a test item, the only features present in the probe are context features since those are persistent in the environment. Once a test item (or prime) is presented, content features may also enter the probe. We assume that content features are sampled as a Poisson process, with sampling events occurring at exponentially distributed intervals according to $f(\tau)=\frac{1}{\rho_{T}} \exp \left(-\frac{\tau}{\rho_{T}}\right)$ at test with rate $\rho_{T}$ and rate $\rho_{S}$ at study. On each sampling event, all the available content features have an equal probability of being selected for sampling. Whichever is selected, the correct value of the feature is stored in the probe with probability $c$, otherwise a random value is stored (in this case, either 0 or 1 with equal probability). ${ }^{1}$ Note that, because all content features have an equal probability of being sampled on each sampling event, it is possible to sample a value for a feature that already has a value in the probe. In that case, the most recently sampled value replaces any previously stored value.

We assume that the same feature sampling process occurs at study. The probability that an available content feature will have a value stored, given limited study time $T_{s}$ and sampling rate $\rho_{S}$ features per second, is $1-\left(1-1 / N_{c}\right)^{\rho_{S} T_{S}}$, which in-

[^48]creases with both $\rho_{S}$ and $T_{s}$. While not all $N_{c}$ content features may end up being stored in a trace, we assume that all context features have a stored value.

## Comparison of Probe to Memory

At a given time $t$, the probe consists of a set of context features as well as whatever features of the test item have been sampled by that time. The probe is compared to each trace in LTM. These comparisons result in a set of likelihood ratios, $\lambda_{i}(t)$ for each trace $i$ in LTM, reflecting the likelihood that the probe and trace encode the same item versus the likelihood that they encode different items (c.f., Shiffrin \& Steyvers, 1997; McClelland \& Chappell, 1998).

Likelihood The features of the probe and a memory trace are aligned and compared individually. In the current restricted version of the model, the only features that affect the likelihood are those in which a value is stored in both the probe and trace, and the values either match or mismatch. For simplicity, we assume the same value of $c$ at study and test, so there are four ways a feature value might match if the probe and trace encode the same item: the value was correctly copied at both study and test (with probability $c^{2}$ ); a value was copied correctly at either study or test but not the other and matches by chance (with total probability $c(1-c)$ ); or the value was copied incorrectly at both study and test but still matches by chance $\left(2\left[\frac{1}{2}(1-c)\right]^{2}\right)$. Summing these probabilities yields the probability of a feature value match given that the probe and trace encode the same item: $\operatorname{Pr}(M \mid$ Same $)=c^{2}+c(1-c)+\frac{1}{2}(1-c)^{2}$. Similarly, if the probe and trace encode the same item, the stored values could mismatch if the value in either the probe or the trace or both were sampled incorrectly and failed to match by chance: $\operatorname{Pr}(N \mid$ Same $)=c(1-c)+\frac{1}{2}(1-c)^{2}$. If the probe and trace encode different events, then regardless of whether either value were sampled correctly, they could only match or mismatch by chance: $\operatorname{Pr}(M \mid$ Diff. $)=\operatorname{Pr}(N \mid$ Diff. $)=\frac{1}{2}$.

Since features are encoded independently of one another, the likelihood ratio across all features is the product of the likelihood ratios for the individual features. Letting $N_{M}(t)$ and $N_{N}(t)$ be the number of feature value matches and mismatches, respectively, the relative likelihood that a probe and trace encode the same versus different events is

$$
\begin{aligned}
\lambda_{i}(t) & =\left[\frac{\operatorname{Pr}(M \mid \text { Same })}{\operatorname{Pr}(M \mid \text { Diff. })}\right]^{N_{M}(t)}\left[\frac{\operatorname{Pr}(N \mid \text { Same })}{\operatorname{Pr}(N \mid \text { Diff. })}\right]^{N_{N}(t)} \\
& =\left(1+c^{2}\right)^{N_{M}(t)}\left(1-c^{2}\right)^{N_{N}(t)}
\end{aligned}
$$

Familiarity Because the number of event traces in memory is likely to be quite large, we assume that there is a threshold for activation and only those traces whose likelihood ratios are greater than this threshold contribute to familiarity. For simplicity, we set this threshold equal to 1 . The familiarity at time $t, \phi(t)$, is the average likelihood ratio among the active traces: $\phi(t)=\left\langle\lambda_{i}(t): \lambda_{i}(t)>1\right\rangle$.

## Making a Recognition Decision

The raw familiarity $\phi(t)$ is not used directly to make a recognition decision, as its absolute value can fluctuate with a variety of factors that would preclude the setting of consistent decision criteria (Cox \& Shiffrin, 2012). Rather, changes in $\log \phi(t)$ are used to make a recognition decision ${ }^{2}$. Positive changes in $\log \phi(t)$ are evidence that the test item is old while negative changes are evidence that the item is new. The evidence state at time $t$, denoted $B(t)$, is the accumulated change in $\log \phi(t)$ since a given start time. If accumulation starts at $t=0$, then $B(t)=\sum_{\tau=1}^{t} \log \phi(\tau)-\log \phi(\tau-1)=$ $\log \phi(t)-\log \phi(0)$.

When $B(t)$ reaches criterion $\beta_{O}$, an "old" response is made and if it reaches $\beta_{N}$, a "new" response is made. However, because at most $N_{c}$ content features are available for sampling, $\log \phi(t)$ will reach a noisy asymptote. As a result, criteria cannot be constant over time because, for some trajectories of $B(t)$, there is a non-zero probability that they will never reach either criterion. Thus, we allow the decision bounds to start at initial values $\beta_{O}^{0}$ and $\beta_{N}^{0}$ and gradually collapse according to a power function of time $r(t)=\left(\frac{t}{t+1}\right)^{N_{c}}$, scaled by the number of available features $N_{c}$. The resulting decision bounds are given by

$$
\beta_{O}(t)=\beta_{O}^{0}-r(t)\left(\frac{\beta_{O}^{0}-\beta_{N}^{0}}{2}\right), \beta_{N}(t)=\beta_{N}^{0}+r(t)\left(\frac{\beta_{0}^{0}-\beta_{N}^{0}}{2}\right) .
$$

## Response Time Predictions

As is standard in RT modeling, we assume that the observed response time arises from a decision component and a residual component, i.e., $T_{o b s}=T_{D}+T_{R}$. The number of samples needed to reach criterion determines the decision component of the RT. If $N_{s}$ samples are taken to reach criterion, then because sampling is a homogeneous Poisson process, the decision time is a sample from a Gamma distribution with rate $\rho_{T}$ and shape $N_{s}$ and expected value $\bar{T}_{D}=\rho_{T} N_{s}$.

The residual component of the RT is due to a number of factors, including the time needed to execute the motor actions needed to make a response. $T_{R}$ may also vary with factors that affect the ability to successfully recognize a stimulus. We do not yet model this process in detail; instead, because we only predict mean RT in the studies reported here, we only assume that the residual process has some stationary mean value such that the mean predicted RT is $\bar{T}_{o b s}=\bar{T}_{R}+\bar{T}_{D}$.

## Model fitting

To fit the model to each experiment reported here, we first selected by hand a set of reasonable values for the key memory parameters $\left(N_{c}, N_{x}, c\right.$, and $\left.\rho_{S}\right)$ and any experiment-specific parameters. The remaining parameters-principally the initial decision bounds $\beta_{o}^{0}$ and $\beta_{N}^{0}$, sampling rate at test $\rho_{T}$, and mean residual time $\bar{T}_{R}$-were fit by minimizing the sum of squared error to each available group data point (hit and FA

[^49]

Figure 2: Mean values of $\log \phi(t)$ for the conditions of Experiment 1. A) Raw profiles with accumulation starting at $t=0$; B) profiles for dynamic conditions with the start point of accumulation moved forward to $t=20$ (dashed vertical lines). Profiles for targets are in blue and foils in red; black lines show time-varying decision bounds.
rates and mean RTs for each response type in each condition) using the SIMPLEX algorithm from several different random start points. As a result, we cannot claim that the fits reported here are the best possible, but our aim is to demonstrate the qualitative behavior of the model, rather than a strict quantitative fit.

## Accounting For Experiment 1

We assume that the stimuli in Experiment 1 are represented by a set of $N_{c}$ content features, where $N_{p}$ features represent each of the three consonants and $N_{w}$ features represent their configuration/conjunction ( $N_{c}=3 N_{p}+N_{w}$ ). Within the 2 categories defined by a prototype, items share $2 N_{p}$ features (i.e., 2 letters) but differ in their "diagnostic" (unique) letter and configural features. Foils also share $2 N_{p}$ features with one of the categories of studied items, but contain a third letter and configural features that differ from all studied items. In the static conditions, all $N_{c}$ features are available to be sampled from the beginning of the trial $(t=0)$. In the dynamic conditions, when the first letter appears $(t=0)$, only the $N_{p}$ features representing it are available for sampling into the probe. When the second letter appears $(t=\delta)$, its features become available for sampling as well. When the final letter appears ( $t=2 \delta$ ), all content features- including the $N_{w}$ configural features-become available for sampling.

Because our interest is in explaining the qualitative patterns in Experiment 1, and because a wide variety of parameter values are capable of producing such patterns, we arbitrarily let $N_{p}=5, N_{w}=15$, and $\delta=10$. Study time was fixed at $T_{S}=1$ second. With these parameter values, along with others given in Table 1, the resulting mean value of $\log \phi(t)$ for each condition is shown in Figure 2. All conditions reach the same asymptotes, but take very different routes to get there. In the early diagnostic condition, foil and target profiles separate widely early on, as would be expected, while in the late diagnostic condition, both foils and targets produce increasing mean familiarity before dividing.

A static model that only used the asymptotic value of familiarity would, incorrectly, make the same predictions for all three conditions. However, most dynamic models-including

Table 1: Parameter values used in the simulations.

| Overall |  | Exp. 1 | Exp. 2 |  |  |
| ---: | ---: | ---: | ---: | ---: | ---: |
| Param. | Value | Param. | Value | Param. | Value |
| $N_{c}$ | 30 | $\rho_{s}$ | 40 | $\rho_{s}$ | 57 |
| $N_{x}$ | 30 | $\beta_{0}^{0}$ | 5.279 | $\beta_{0}^{0}$ | 6.160 |
| $c$ | 0.86 | $\beta_{N}^{0}$ | -5.291 | $\beta_{N}^{0}$ | -5.975 |
|  |  | $\rho_{T}$ | 100 | $\rho_{T}$ | 95 |
|  |  | $\bar{T}_{R}$ | 0.438 | $\bar{T}_{R}$ | 0.360 |
|  |  | $N_{p}$ | 5 | $\pi$ | 0.25 |
|  |  | $N_{w}$ | 15 | $\eta$ | 0.76 |
|  |  | $\delta$ | 10 | $K$ | 200 |
|  |  |  |  | $\bar{T}_{C}$ | 0.016 |

the one outlined above-would incorrectly predict an increase in FAs for the dynamic-late condition due to the increased probability of reaching $\beta_{O}(t)$ early on. A critical feature of our model, however, is that it accumulates changes in familiarity, not absolute familiarity. If instead of accumulating $\log \phi(t)-\log \phi(t-1)$ from $t=0$, accumulation began when all features were available (at $t=2 \delta$ ), the resulting evidence state would be $B(t)=\log \phi(t)-\log \phi(2 \delta)$, as shown in Figure 2B. This delay leads to predictions that match the data: overall greater RT in the dynamic conditions, relatively little difference in HR, and a marked decrease in FAR for the dynamic-late condition (see Figure 1). The FAR prediction arises because the first $2 \delta$ samples for a foil in the dynamiclate condition all tend to match the studied items, so ignoring those early matching samples means that the later nonmatching samples are emphasized.

Why wait to begin accumulating changes? Although the dynamic presentation was fast enough that participants could not know which letters came on in what order, they could perceive that the display was noisy or "flickery". Rationally, one would not want to risk accumulating noise and so it makes sense that participants would wait until the display was sufficiently clear to begin accumulating evidence for recognition (e.g., Smith, Ratcliff, \& Wolfgang, 2004). This kind of waiting is also analogous to discounting in short-term recognition (Huber, Shiffrin, Lyle, \& Ruys, 2001), in that evidence is down-weighted when it is attributed-perhaps erroneouslyto noise.

## Experiment 2: Fluency

It turns out that essentially the same mechanism—missing the first few samples before beginning accumulation-can explain an apparently unrelated result in the recognition literature: the so-called "fluency effect". It is based on the idea that the subjective feeling of familiarity, rather than the presence or absence of a memory trace, leads one to decide that an item is old, and that this feeling can arise from multiple sources (Jacoby \& Dallas, 1981). One such source is a "fluency heuristic" in which people detect the relative ease of perceptual processing of a test item and use this as a sign of past experience. Jacoby and Whitehouse (1989) demonstrated that old and new words preceded by a subliminal matching prime increased the probability of judging a word as old. In terms of fluency, the subliminal flash provides a head start in processing thereby increasing fluency and giving the illusion of familiarity regardless of whether the word was old or new. We
present a replication of these results and show how a dynamic model of recognition can account for them without appealing to a fluency heuristic.

## Participants

81 undergraduate students from Indiana University participated in the experiment for course credit.

## Stimuli

Stimuli consisted of concrete nouns of moderate length and frequency drawn from the Toronto word pool. 90 words were selected for study, and another 90 served as foils at test. In addition, 60 words served as different primes at test.

## Procedure

The stimuli were presented in lowercase letters in the center of a computer monitor. In the study phase, participants were instructed to read words aloud as they appeared on the screen and to remember them for a later test of memory. The study phase was divided into two blocks, one with words presented for 1 s , another with words presented for 3 s (the order of the blocks was randomized).

During the test phase, participants were instructed to respond whether the presented test word came from the study list ("old") or from the set of new words. Old and new responses were randomly mapped to the "A" or "L" key for each participant. Each recognition test word was preceded by a nondiagnostic subliminal prime: on $1 / 3$ of trials, the prime was identical to the test word, on another $1 / 3$ of trials, the prime was a different word that had not been previously seen, and on another $1 / 3$, the prime was a neutral string of characters (XOXOXO). On each trial, a pre-mask ( \&\&\&\&\&\&\&) was presented for 500 ms followed by a prime (same, different, or neutral) for 50 ms and a post-mask for an additional 500 ms . The screen went blank for 300 ms before the test word was presented. After the participant made a response, the screen was cleared for 1865 ms until the next test trial. Participants were not informed that the primes would be present.

## Results

Prior to analysis, trials with RT that were too fast (less than 200 ms ) or too slow (longer than 3 s ) were excluded ( 273 out of 14580 total trials). The observed mean probability of responding "old" in each condition is shown in Figure 3A. Replicating the original result of Jacoby and Whitehouse (1989), participants are significantly more likely to endorse an item that was preceded by an identity prime than a neutral prime $(t(80)=12.0, p<0.001)$. Surprisingly, they are also more likely to endorse an item that was preceded by a different prime than a neutral one $(t(80)=2.91, p=0.005)$, an effect also remarked on, but unexplained, in the original work of Jacoby and Whitehouse (1989). Observed mean correct RT are shown in Figure 3B. Identity primes speed hits $(t(80)=-10.7, p<0.001)$, but slow CR $(t(80)=2.68$, $p=0.009)$ relative to neutral primes. Different primes also


Figure 3: Observed and predicted mean accuracy and correct RT for Experiment 2. Predictions averaged over 2000 simulations. Error bars depict 95\% confidence intervals.


Figure 4: Mean familiarity over time for Experiment 1. Targets are in blue, foils in red, and black lines show timevarying decision bounds.
slow CR relative to neutral primes $(t(80)=3.21, p=0.002)$, but have no significant effect on RT for hits $(t(80)=0.93$, $p=0.35$ ).

## The Dynamic Account of Fluency

The core of our account of the fluency effect lies in the assumption that the prime, if it is a word, contributes some features to the probe before features begin to be sampled from the test item and accumulation begins. In the case of an identical prime, this is exactly like changing the start time of accumulation in Experiment 1, since it eliminates the effects of the first few samples, as shown in Figure 4. Notice that, for both targets and foils, the first few samples in the neutral prime condition will, on average, produce negative changes in familiarity. If the prime word is identical to the subsequent test item, "pre-loading" the first few features eliminates some of these negative changes, making it harder to reach $\beta_{N}(t)$ and increasing the probability of responding "old" for both targets and foils.

The initial negativity for targets is a consequence of how the set of activated traces changes over time as features accumulate in the probe, as outlined in Figure 5. Before any content features are sampled, the probe contains only context features and the active traces tend to be those from recent

| Features | Start <br> Context <br> only | Early <br> Context and some <br> content | Late <br> Context and all content |
| :--- | :--- | :--- | :--- |
| Traces <br> active for <br> target | List traces <br> (good <br> match) | List traces (moderate <br> match), target trace <br> (moderate match) | Some list traces (poor <br> match), target trace (good <br> match), some history traces <br> (poor match) |
| Traces | List traces | List traces (moderate | Some list traces (poor <br> match), some history traces |
| active for | (good <br> foil | match) |  |
| (poor match) |  |  |  |

Figure 5: Outline of the evolution of familiarity as a function of feature and trace activation over time.
experience, i.e., the study list. If just a few content features are sampled, most or all list traces will remain active, even though most will not match on content features. As an example, say you had studied the list "table", "moon", "parent" and were shown "table" at test. If you had only sampled features of the first letter (" t "), they would only match 1 out of 3 study items. It is only after many content features have been sampled (e.g., another several letters) that list traces that do not match the target drop below the threshold for activation and the match to the target trace takes precedence, raising the average likelihood.

This kind of priming effect also operates for different primes. Because the first few sampled features will not match most of the list traces in any case, the features that leak from a different prime will also tend to eliminate some initial negative changes, leading to an increased probability to say old to both targets and foils. However, if the test word differs from a prime word, this also impairs word recognition by introducing competition between the prime word and the test word (McClelland \& Rumelhart, 1981; Segui \& Grainger, 1990). If we assume that, as in Experiment 1, participants wait until they have a clear percept before beginning accumulation, it is reasonable to suggest that participants wait until this competition is resolved (i.e., they have a clear percept of the word) before beginning accumulation. This takes some time, during which some of the prime features-which are no longer being actively sampled or maintained-have a chance of deactivating and losing their sampled values, thereby separating the different and same prime predictions.

In sum, we assume that each of the prime's features has a probability $\pi$ of being sampled into the probe by the time accumulation begins. There is a constant mean duration $\bar{T}_{C}$ required to resolve the competition during word recognition in the different-prime condition, during which there is a probability $\eta$ that any sampled prime feature will deactivate. The features of the study and test words are assigned randomly. We also assume that, because participants have prior experience with words, there are $K$ traces of each word from life history that can be activated at test (their context features are assigned randomly; values used for these parameters are given in Table 1). As shown in Figure 3C-D, the model predicts the canonical "fluency" finding of increased $p$ ("Old") with an identical prime, as well as decreased RT for hits and increased RT for CR. It also exhibits the observed small positive priming effect for different primes.

## Discussion

We have shown how a dynamic model of recognition that allows the start time of accumulation to vary independently of feature sampling can account not just for the novel results of our dynamic presentation study, but for older findings previously attributed to a fluency heuristic. Further, this model is able to predict the observed positive priming from novel, different primes, a prediction made by no other extant recognition model. These seeming disparate predictions fall directly out of the core feature of the model, namely, that it is changes in familiarity from a given start time, not absolute familiarity, that drive recognition decisions.

The model and these results thus illustrate the importance of knowing when information for recognition becomes available. Here, we have assumed that unconscious priming and dynamic presentation result in features being sampled before changes get accumulated for recognition, but other manipulations (e.g., manipulations of salience or contrast) might also influence the start time of accumulation, which would be fruitful questions for further study. We suggested that, to account for results in the different-prime condition in Experiment 2, responses are slowed by an increased difficulty in recognizing the test word. Although this suggestion remains to be formalized in a model, it emphasizes the need to consider all the possible processes that contribute to recognition. Creating this kind of unified model requires a dynamic approach, which not only suggests new answers to issues like fluency, but new questions.

## References

Brockdorff, N., \& Lamberts, K. (2000). A feature-sampling account of the time course of old-new recognition judgments. Journal of Experimental Psychology: Learning, Memory, and Cognition, 26(1), 77-102.
Cox, G. E., \& Shiffrin, R. M. (2012). Criterion setting and the dynamics of recognition memory. Topics in Cognitive Science, 4(1), 135-150.
Gronlund, S. D., \& Ratcliff, R. (1989). Time course of item and associative information: Implications for global memory models. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15(5), 846-858.
Hintzman, D. L., \& Curran, T. (1994). Retrieval dynamics of recognition and frequency judgments: Evidence for separate processes of familiarity and recall. Journal of Memory and Language, 33, 1-18.
Hockley, W. E., \& Murdock, B. B. (1987). A decision model for accuracy and response latency in recognition memory. Psychological Review, 94(3), 341-358.
Huber, D. E., Shiffrin, R. M., Lyle, K. B., \& Ruys, K. I. (2001). Perception and preference in short-term word priming. Psychological Review, 108(1), 149-182.
Jacoby, L. L., \& Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. Journal of Experimental Psychology: General, 110(3), 306-340.
Jacoby, L. L., \& Whitehouse, K. (1989). An illusion of memory: False recognition influenced by unconscious perception. Journal of Experimental Psychology: General, 118(2), 126-135.
Malmberg, K. J. (2008). Recognition memory: A review of the critical findings and an integrated theory for relating them. Cognitive Psychology, 57, 335-384.
McClelland, J. L., \& Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. Psychological Review, 105(4), 724-760.
McClelland, J. L., \& Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. Psychological Review, 88, 375-405.
Mewhort, D. J. K., \& Johns, E. E. (2005). Sharpening the echo: An iterative resonance model for short-term recognition memory. Memory, 13(3/4), 300-307.
Nosofsky, R. M., \& Stanton, R. D. (2006). Speeded old-new recognition of multidimensional perceptual stimuli: Modeling performance at the individual-participant and individual-item levels. Journal of Experimental Psychology: Human Perception and Performance, 32(2), 314-334
Segui, J., \& Grainger, J. (1990). Priming word recognition with orthographic neighbors: Effects of relative prime-target frequency. Journal of Experimental Psychology: Human Perception and Performance, 16(1), 65-76.
Shiffrin, R. M., \& Steyvers, M. (1997). A model for recognition memory: REMretrieving effectively from memory. Psychonomic Bulletin \& Review, 4(2), 145166.

Smith, P. L., Ratcliff, R., \& Wolfgang, B. J. (2004). Attention orienting and the time course of perceptual decisions: Response time distributions with masked and unmasked displays. Vision Research, 44, 1297-1320.

# Representing Number Sets: Encoding Statistical Properties 

Patrick F. Cravalho (pcravalh@,kent.edu), Bradley J. Morris (bmorri20@kent.edu), Christopher A. Was (cwas@kent.edu)<br>Department of Educational Psychology, 405 White Hall<br>Kent, OH 44242 USA

Amy M. Masnick (Amy.M.Masnick@hofstra.edu)<br>Department of Psychology, Hauser Hall<br>Hempstead, NY 11549 USA


#### Abstract

Previous research suggests that people often recall individual items when sets are smaller than four and aggregate set features for sets larger than four. One intriguing possibility is that the process of aggregating sets creates summary representations that maintain the statistical properties of the set itself. For sets of numbers, this process might implicitly create approximate means. We report the results of two experiments investigating memory for number sets and its relation to working memory and metacognitive monitoring. In both experiments, participants were shown a series of data sets that varied in size ( 4,6 , or 8 numbers) and variance ( $10 \%$ or $20 \%$ of the mean) and were then presented with an actual value from the set and the set mean. In Experiment 1, participants were asked to select the actual value, and in Experiment 2, participants were asked to select the set mean. Results indicated that the proportion of correct selections and metacognitive confidence decreased with set size. Working memory was related to performance only when the set size was 6 . The findings suggest that participants often erroneously reported the set mean as being a member of the set and that this effect increased for sets larger than four. The findings suggest that the process of aggregating number sets results in approximate means.


Keywords: Number sets; Representation; Metacognition.

## Introduction

Data interpretation (i.e., interpreting numbers in context) is critical to everyday reasoning. Numbers are represented in two ways, as an approximate magnitude and as a symbolic value (Dehaene, 2009). Previous research suggests that people compare single numbers as analog magnitudes, meaning that numbers are represented as approximate rather than exact values. Key evidence for this argument is that comparisons are faster and more accurate as the ratio of difference between values increases (Feigenson, Dehaene, \& Spelke, 2004; Hyde, 2011). Yet little is known about how people represent numerical data sets. Previous research involving the cognitive processing of number sets suggests that people's comparisons are highly related to the statistical properties of the sets. Specifically, as the ratio of mean differences between number sets increase and variance within each set decreases, comparisons to assess which set has the higher mean become faster and more accurate (Masnick \& Morris, 2008). These findings suggest that
when asked to make number comparisons, people implicitly (i.e., without deliberate effort) represent the properties of number sets, such as means and variance, rather than information about individual numbers.

## Sets vs. Individuals

A well-replicated finding is that people individuate items (i.e., represent individual values within a set) for sets smaller than four (Scholl, 2001). For example, given a set of three dots, participants are likely to remember the individual dots rather than the properties shared by the dots (Airely, 2001). Given sets larger than four, people aggregate items, or average over sets retaining information about set features (Scholl, 2001; Masnick \& Morris, 2008). For example, given a set of a dozen dots of various sizes, participants were likely to erroneously recall a dot that represents the average size, rather than correctly recall an actual dot from the set they viewed (Airely, 2001). Earlier research suggested two separate memory stores for individual versus aggregated sets. For example, Feigenson, Dehaene, and Spelke (2004) suggested one system for individuating small values $(<4)$ and another system for aggregating larger sets. However, recent research (Hyde, 2011) suggests that the difference in representation may be functional (i.e., due to differences in strategy) rather than structural (i.e., separate stores).

The process of aggregating appears to maintain information about the statistical properties of sets. Previous research on category learning provides evidence that aggregating over individual objects results in a prototype, or a most representative set member. Researchers have suggested that the process of aggregating yields "a measure of central tendency" (e.g., an average or prototypical category member; Medin, Altom, \& Murphy, 1984, p. 334). Early research into prototypes suggested that when people aggregate across multiple exemplars, they extract a prototypical category member that represents an average of category features (Medin et al., 1984; Nosofsky, Denton, Zaki, Murphy-Knudsen, \& Unverzagt, 2012). For example, a robin is often seen as a prototypical bird because it shares most of the features common among birds.

Numbers provide an intriguing extension of this work because aggregation of sets may produce approximate
measures of central tendency. Specifically, encoding a set of numbers may implicitly provide information about the set mean. The reviewed research suggests two related effects. One, for small set sizes ( $<4$ ), participants will encode individual set members. For larger set sizes (> 4), participants will aggregate across values, losing information about individual values. Two, the process of aggregating values creates summary representations of the number sets, including approximate mean and variance values.

## Set Encoding and Working Memory

Working memory (WM) is often described as the processes and mechanisms involved with the maintenance of task relevant information necessary for performance of cognitive tasks (Miyake \& Shah, 1999). In the current study, we predict that set representation is more likely to occur as the number sets increase in size because maintenance of all items from larger sets requires greater WM resources than maintenance of items from smaller sets. Therefore, we also predict that participants with larger WM capacity should be able to maintain a greater number of items in WM and correctly identify presented items.

## Set Encoding and Metacognitive Monitoring

A secondary focus of this experiment was to investigate the influence of metacognitive monitoring on the representation of number sets. Monitoring refers to one's judgment of their current cognitive processing, specifically, the degree to which a process is deliberate and a rough estimate of one's performance (Dunlosky \& Metcalfe, 2009). We assessed monitoring through participant confidence judgments.

## Aims and Predictions

This experiment extends the work of Masnick and Morris (2008) by examining how people represent and remember number sets, and how accurately they judge their memory for number sets. Independent variables were number set size ( 4,6 , and 8 ) and coefficient of variation ( $20 \%$ and $10 \%$ of mean). Dependent variables were number selection accuracy, reaction time, confidence judgments, strategy use, and working memory capacity. We predicted that people would erroneously recall seeing set means or medians rather than actual numbers from the sets as set sizes increased. As suggested by Hyde (2011), different strategies are predicted to be associated with encoding goals. Specifically, encoding individual numbers is expected to be associated with longer reaction times than encoding set aggregates, because the latter is a relatively implicit process. Finally, we predicted that number recall and mean identification would be related to confidence judgments in terms of accuracy, or that they would increase or decrease together.

## Experiment 1

## Method

Participants. Participants ( $\mathrm{N}=51$ ) were undergraduate students at a Midwestern state university. The average age was $22.04(S D=6.84), 84 \%$ of the participants were female and most were of Caucasian descent.

Procedure. Experiment 1 consisted of two counterbalanced tasks: the number set task and a working memory task. All participant tasks were presented using E-Prime software. The number set task consisted of 126 trials. Each trial had four parts, described here in order of presentation (see Figure 1a).


Figure 1. Number set and working memory task procedures.
First, a fixation cross was presented to focus the participant's attention. Then, a number set was shown (details of the sets and presentation durations below). Next, two numbers were presented, an actual value from the set and the mean or median value from the set. The participant was asked to indicate which of the two numbers was in the set presented in the previous slide by pressing a computer key. The two numbers and indication prompt remained on the screen until the participant chose an answer. After responding, participants were asked how sure they were of their answer. Confidence judgment response options were presented as follows: 1) $0-25 \%$, 2) $26-50 \%$, 3) $51-75 \%$, \& 4) $76-100 \%$. The response options and prompt remained on the screen until the participant chose an answer.

There were four sets of stimuli, each preceded by a set of instructions explaining the process outlined above to the participant. The first six number sets were practice trials and were not analyzed. Experimental trials included 40 sets of
four numbers (each set presented for 2 s ), 40 sets of six numbers (each set presented for 2.5 s ), and 40 sets of eight numbers (each set presented for 3 s ). Each set consisted of three digit numbers. Within each set size, 10 sets were drawn from one of four variance types. Set variance was either $10 \%$ (low variance) or $20 \%$ (high variance) of the set mean. For half of the choice trials, participants were shown the set mean. For the remaining half, participants were shown the set median; although for sets of 4 the mean and median were identical.

The three blocks of experimental trials were presented sequentially, randomized within-participants. The presentation location of the actual value and mean or median value was also randomized so that the actual value and mean were presented on the left or right side of the screen in $50 \%$ of trials. After completing all number set trials, participants were surveyed about their strategy use during the task. Participants were presented nine strategy descriptions (see Table 1) and were asked to estimate how often they used each strategy during the experiment using the following scale: 1) never, 2) some trials, 3) most trials, or 4) always. The response options and prompt remained on the screen until the participant chose an answer.

Table 1: Strategy Descriptions and Examples.

| Strategy | Example |
| :--- | :--- |
| Look at the first digit. | The "1" in 125. |
| Look at the second digit. | The " "in 125. |
| Look at the third digit. | The "5" in 125. |
| Try to figure out the <br> average. | Calculate mean value. |
| Find the biggest number. | Scan set for highest value. |
| Find the smallest number. | Scan set for lowest value. |
| Just get a sense of the <br> numbers. | Scan set values. |
| Look for a number that is <br> not like other numbers. | Find any value unlike other <br> values. |
| Try to memorize specific <br> numbers. | Memorize a few numbers. |

Participants were also given the alphabet mathematics working memory task (cf. Was, Rawson, Bailey, \& Dunlosky, 2011). The general format of this task was as follows (see Figure 1b): the words "Get Ready" preceded a letter or set of nonadjacent letters (presented for 2.5 s ), followed by a prompt containing a transformation direction and number (ex. "Forward 2"). The participant then counted forward or backward from the given letter or set of letters by the number given. The transformation prompt remained on the screen until the participant was ready to choose an answer from a list of eight alternatives. Participants were instructed to solve the problem before advancing to the response alternative screen. Once a participant had advanced to the response alternative screen, a time limit of 6 s was imposed to prevent one from solving the problem while examining the alternatives in the response window.

This pattern spanned four practice trials and a set of 24 stimuli, which was preceded by a set of instructions explaining the process outlined above to the participant. Feedback of "Correct!" or "Incorrect" was given for each response. More detailed feedback was given during the practice trials. Specially, if a correct answer was given it was accompanied by a brief explanation of why that answer was correct, but if an incorrect answer was given it was accompanied by a more detailed explanation of how the correct answer would have been attained. The block of 24 stimuli was randomized within-participants.

## Results and Discussion

Participants were most accurate ( $M=.79, S D=.12$ ) in their responses for sets of four numbers. Accuracy (see Figure 2) then declined as set size increased to six ( $M=.68$, $S D=.10)$ and again when it increased to eight ( $M=.61, S D$ $=.08)$. Participants were also most confident $(M=3.05, S D$ $=.52$ ) in their responses for sets of four numbers. Confidence also declined as set size increased to six ( $M=$ 2.63, $S D=.54$ ) and again when it increased to eight ( $M=$ 2.43, $S D=.63$ ).

## Exp. 1 Accuracy



Figure 2. Mean accuracy scores for Experiment 1.
Reaction times (see Figure 3) for sets of four ( $M=$ 2269.37, $S D=585.34$ ) were faster than for sets of six $(M=$ 2338.76, $S D=620.24$ ), but not faster than for sets of eight ( $M=2197.69, S D=618.27$ ).

## Exp. 1 Reaction Time



Sets of $4 \quad$ Sets of $6 \quad$ Sets of 8

Figure 3. Mean reaction times for Experiment 1.

Repeated measures ANOVA were used to analyze all main variables. There was a significant decline in accuracy as set size increased, $F(2,104)=59.94, p=.000, \eta^{2}=.535$. More specifically, accuracy decreased significantly as set size increased from 4 to $6, F(1,52)=39.62, p=.000, \eta^{2}=$ .432, and as set size increased from 6 to $8, F(1,52)=20.98$, $p=.000, \eta^{2}=.288$.

There was a significant decrease in reaction times as set size increased, $F(2,104)=5.15, p=.007, \eta^{2}=.090$. While there was no significant difference in reaction time between sets of 4 and sets of $6, F(1,52)=2.48, p=.121, \eta^{2}=.046$, there was a significant difference in reaction time between sets of 6 and sets of $8, F(1,52)=11.38, p=.001, \eta^{2}=.180$.

There was a significant decrease in confidence in one's answer as set size increased, $F(2,104)=52.07, p=.000, \eta^{2}$ $=.500$. Contrasts showed significant decreases in confidence as set size increased from 4 to $6, F(1,52)=$ $55.80, p=.000, \eta^{2}=.518$, and as set size increased from 6 to $8, F(1,52)=12.19, p=.001, \eta^{2}=.190$. Zero-order correlations indicated a positive relation between selection accuracy and confidence judgment for set size of $4, r(52)=$ $.58, p=.000$, and set size of $6, r(52)=.45, p=.001$, but not set size of $8, r(52)=.19, p=.173$.

We compared the effect of high and low variance number sets, aggregated across set size. A paired samples t-test indicated that there was greater accuracy when variance in the number set was higher $(M=.72)$ than when it was lower $(M=.67), t(52)=4.30, p=.000, d=1.181$. However, there was no significant difference in reaction times between high ( $M=2301.25$ ) and low ( $M=2260.82$ ) variance, $t(50)=$ $1.39, p=.171, d=.393$.

No correlations were found between working memory or any of the variables of interest, so we performed a mean split of the participants based on working memory task performance and conducted independent samples $t$-tests to compare the groups in terms of accuracy. There was no significant differences between groups based on sets of 4, $t(51)=.46, d=.02, p=.65$, or sets of $8, t(51)=.95, d=.02$, $p=35$. However, the group scoring above the mean on the WM task performed better than the group scoring below the mean when set size was $6, t(51)=2.08, d=.06, p=.04$, 95\% CI [.002, .114].

The most frequently used strategies were to "just get a sense of the numbers" ( $M=3.19, S D=.90$ ) and to "try to memorize specific numbers" ( $M=2.87, S D=1.01$ ). These data reflect the general pattern in recall accuracy, as participants likely tried to memorize numbers from the sets of 4 , and then were forced to abandon that strategy in favor of "just getting a sense of the numbers" as set sizes exceeded four. This pattern supports the finding of Scholl (2001) that people individuate items for sets smaller than four. One of the least frequently used strategies was to "try and figure out the average" ( $M=1.26, S D=.62$ ). This low rating provides some evidence that the participants were not consciously calculating the mean for the number sets. This pattern is interesting because their response patterns indicate that they chose the mean or median value of the set more
often than the actual value from the set as set size increased. This pattern supports findings that people aggregate items for sets larger than four (Scholl, 2001; Masnick \& Morris, 2008).

The results demonstrate a clear effect of set size in that participants were more likely to recall the actual value given smaller set sizes (e.g., 4) than larger set sizes (e.g., 8). This finding is consistent with findings using different types of stimuli (e.g., dots) suggesting that encoding individual items is the default operation until the number of items exceeds working memory limits, demonstrated as four elements in previous research (Ariely, 2001; Feigenson et al., 2004). Once working memory limits are reached, the default operation appears to be encoding the set as an aggregate. The data also suggest that the process of aggregating numbers yields as approximate mean. As set size increased, participants were more likely to select the set mean rather than the actual value as the number they had seen in the experiment. Further, when the numbers varied less and were within a smaller range, accuracy also decreased. It became more difficult to distinguish set members from their mean when the set members were more similar.

Working memory capacity also played a role in the ability to identify presented values. When the sets presented contained six items, individuals scoring above the mean on the WM task out performed those scoring below the mean. When the set size was four, WM capacity was not taxed for either group, and when set size was eight, WM capacity was likely exceeded for both groups.

In general, the confidence level of the participants fell along with their performance, although the influence of metacognitive monitoring on number recall was greater for sets of 4 and sets of 6 than for sets of 8 . The participants showed accurate recognition that their performance was declining as set size, and task difficulty, was increasing. The lack of correlation between recall and confidence accuracy for sets of 8 may be related to the pattern of strategy use, as this set size would have been too large for one to memorize the numbers presented, as well as large enough to pose difficulty in trying to get a sense of the whole set.

## Experiment 2

One possible limitation of Experiment 1 is that the task explicitly asked participants to attend to and encode individual values. Experiment 2 provided the same experimental conditions but changed instructions to ask participants to identify the set mean. This task parallels the Deese-Roediger-McDermott (DRM) paradigm (Deese, 1959; Roediger \& McDermott, 1995), as participants will study lists that are comprised of numbers (rather than words) related to a non-presented response target (Sugrue \& Hayne, 2006). The change in instructions changes the task demands and likely changes the strategy used for the task. Specifically, if participants are asked to recall set means, they will be less likely to encode individual values. Previous research demonstrates that although people are relatively
accurate when estimating set means, estimation accuracy decreases with set size (Peterson \& Beach, 1967).

## Method

Participants. Participants ( $\mathrm{N}=27$ ) were undergraduate students at a Midwestern state university. The average age was $20.67(S D=5.88), 70 \%$ of the participants were female and most were of Caucasian descent.

Procedure. The number sets and working memory tasks for experiment two followed the same procedure as for experiment one, except that for number sets participants were asked to indicate which of the two numbers represented the average of the set of numbers (rather than which of the values was actually in the set).

## Results and Discussion

Participants were most accurate $(M=.71, S D=.19)$ in their responses for sets of four numbers. Accuracy (see Figure 4) then declined as set size increased to $\operatorname{six}(M=.62$, $S D=.14$ ) and again when it increased to eight ( $M=.58, S D$ $=.13$ ). Participants were also most confident ( $M=2.76, S D$ $=.64$ ) in their responses for sets of four numbers. Confidence also declined as set size increased to six ( $M=$ 2.37, $S D=.66$ ) and again when it increased to eight ( $M=$ $2.15, S D=.56$ ).


Figure 4. Mean accuracy scores for Experiment 2.
Reaction times (see Figure 5) for sets of four ( $M=$ 2743.02, $S D=673.50$ ) were slower than for sets of $\operatorname{six}(M=$ 2512.29, $S D=951.66$ ). Reaction times for sets of eight ( $M$ $=2238.00, S D=922.49$ ) were faster than for sets of six.

As in Experiment 1, repeated measures ANOVA were used for all main analyses. There was a significant decline in accuracy as set size increased, $F(2,52)=9.00, p=.000$, $\eta^{2}=.257$. Contrasts showed that accuracy decreased significantly as set size increased from 4 to $6, F(1,26)=$ 7.26, $p=.012, \eta^{2}=.218$, but not as set size increased from 6 to $8, F(1,26)=2.32, p=.140, \eta^{2}=.082$.


Figure 5. Mean reaction times for Experiment 2.
There was a significant decrease in reaction times as set size increased, $F(2,52)=11.01, p=.000, \eta^{2}=.298$. There was no significant difference in reaction time between sets of 4 and sets of $6, F(1,26)=3.76, p=.063, \eta^{2}=.127$, but there was a significant decrease in reaction time between sets of 6 and sets of $8, F(1,26)=13.98, p=.001, \eta^{2}=.350$.

There was a significant decrease in confidence ratings across set size, $F(2,52)=22.44, p=.000, \eta^{2}=.463$. Contrasts showed significant decreases in confidence as set size increased from 4 to $6, F(1,26)=18.52, p=.000, \eta^{2}=$ .416 , and as set size increased from 6 to $8, F(1,26)=8.55, p$ $=.007, \eta^{2}=.248$. Zero-order correlations indicated no relation between identification accuracy and confidence judgment for set size of $4, r(27)=.07, p=.709$, set size of $6, r(27)=-.07, p=.708$, or set size of $8, r(27)=.33, p=$ . 087.

We compared high and low variance, aggregated across set size. A paired samples t-test indicated that there was no difference in accuracy when variance in the number set was higher $(M=.64)$ than when the variance was lower ( $M=$ .64), $t(26)=.184, p=.855, d=.072$. However, there was a significant difference in reaction times between high ( $M=$ 2543.05) and low ( $M=2439.62$ ) variance, $t(26)=2.24, p=$ $.034, d=.879$.

No correlations were found between working memory or any of the variables of interest, so we performed a mean split of the participants based on working memory task performance and conducted independent samples t-tests to compare the groups in terms of accuracy. There was no significant differences between groups based on sets of 4 , $t(25)=-1.07, d=-.09, p=.29$, on sets of $6, t(25)=-1.50, d$ $=.15, p=.146$, or sets of $8, t(25)=-1.03, d=.31, p=.311$.

The most frequently used strategies were to "just get a sense of the numbers" $(M=2.85, S D=1.02)$ and to "look at the first digit" $(M=2.59, S D=.93)$. As with experiment one, most participants tried to get a sense of each number set. Predictably, fewer participants cited memorization ( $M=$ $2.33, S D=1.07$ ) as a frequently used strategy than those from experiment one, as this would not have been necessary due to the change in answer format. These data reflect the general pattern in identification accuracy, as participant performance stabilized once the sets grew larger than 4
numbers, the point at which one would expect participants to have a better chance at inferring set characteristics rather than trying to focus on memorizing individual numbers. This pattern supports the finding that people individuate items for sets smaller than four, but aggregate items for sets larger than four (Scholl, 2001; Masnick \& Morris, 2008).

These results reveal a similar pattern of behavior as in Experiment 1, with a few exceptions. One exception was that variance in set size did not affect accuracy in identifying the mean in the same way it did identifying members of the original set. Another exception is that working memory did not appear to play a role in accuracy, even in the largest sets, when identifying the mean. It is possible that differences between a recognition task and a slightly more inferential task are playing a role, but more data with direct comparisons will be necessary to explore this issue in more detail.

## General Discussion

Our results provide new insights into how people represent number sets. One, it appears that number sets and nonnumerical sets are aggregated similarly. Number sets of four appear to be accurately encoded as individual values, likely represented as symbolic values. Larger sets appear to be encoded as aggregates and represented as analog magnitudes. It appears that the process of aggregating over sets results in an approximate mean. In Experiment 1, as set size increased, participants were increasingly likely to erroneously select the set mean, rather than the actual value from the set. It appears that even with larger sets, participants quickly approximated means, though less accurately, than for smaller sets.

The results suggested that strategies are important in creating different representations. One intriguing piece of evidence from Experiment 1 is that reaction times for sets of four and eight were shorter than for sets of six. This suggests that participants might have been attempting to use the same individuation strategy with sets of four and six. In correspondence, we found a relation between working memory capacity and performance only for those with high capacities and only for sets of six items. This outcome was not surprising, because the individuation strategy would be more difficult for sets of six than for sets of four. This suggestion is supported by a decline in accuracy and confidence from sets of four to sets of six.

In conclusion, our results suggest that memory for number sets is similar to memory for sets of objects in that sets smaller that four are likely to be individuated while sets larger than four are likely to be aggregated. The process of aggregation appears to maintain approximate representations of the statistical properties of the sets. One possible explanation for the lack of relation between working memory and recall accuracy found in this study is that the tasks themselves load on spatial working memory, rather than verbal memory (Morris \& Masnick, 2008).

## References

Ariely, D. (2001). Seeing sets: Representation by statistical properties. Psychological Science, 12(2), 157-162.
Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. Journal of Experimental Psychology, 58, 17-22.
Dehaene, S. (2009). Origins of mathematical Intuitions: The case of arithmetic. The Year in Cognitive Neuroscience 2009: Ann. N.Y. Acad. Sci., 1156, 232-259.
Dunlosky, J., \& Metcalfe, J. (2009). Metacognition. London: Sage Publications, Inc.
Feigenson, L., Dehaene, S., \& Spelke, E. (2004). Core systems of number. Trends in Cognitive Sciences, 8(7), 307-314.
Hyde, D. C. (2011). Two systems of non-symbolic numerical cognition. Frontiers in Human Neuroscience, 5, 1-8.
Masnick, A. M., \& Morris, B. J. (2008). Investigating the development of data evaluation: The role of data characteristics. Child Development, 79(4), 1032-1048.
Medin, D. L., Altom, M. W., \& Murphy, T. D. (1984). Given versus induced category representations: Use of prototype and exemplar information in classification. Journal of Experimental Psychology: Learning, Memory, and Cognition, 10(3), 333-352.
Miyake, A. \& Shah, P. (1999). Models of Working Memory. Cambridge, UK: Cambridge University Press.
Morris, B. J., \& Masnick, A. M. (2008). Making numbers out of magnitudes. Behavioral and Brain Sciences, 31(6), 662-663.
Nosofsky, R. M., Denton, S. E., Zaki, S. R., MurphyKnudsen, A. F., \& Unverzagt, F. W. (2012). Studies of implicit prototype extraction in patients with mild cognitive impairment and early Alzheimer's disease. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15(46), 1-21.
Peterson, C. R., \& Beach, L. R. (1967). Man as an intuitive statistician. Psychological Bulletin, 68(1), 29-46.
Roediger, H. L., III, \& McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. Journal of Experimental Psychology: Learning, Memory, and Cognition, 21, 803-814.
Scholl, B. J. (2001). Objects and attention: The state of the art. Cognition, 80(1/2), 1-46.
Sugrue, K., \& Hayne, H. (2006). False memories produced by children and adults in the DRM paradigm. Applied Cognitive Psychology, 20, 625-631.
Was, C. A., Rawson, K. A., Bailey, H., \& Dunlosky, J. (2011). Content-embedded tasks beat complex span for predicting comprehension. Behavioral Research Methods. doi: 10.3758/s13428-011-0112-x

# Unified Modeling of Proactive Interference and Memorization Effort: A new mathematical perspective within ACT-R theory 

Arindam Das, Wolfgang Stuerzlinger<br>Computer Science and Engineering, York University, Toronto, Canada<br>\{arindam, wolfgang\} @cse.yorku.ca


#### Abstract

We parsimoniously model the effect of proactive interference and memorization effort in learning stable graphical layouts. We model the visual search cost, i.e. the number of distractors visually encoded while looking for a target item, as a reasonable surrogate of onscreen proactive interference. Further, we show that a novel quantity that we term "effort factor" is an acceptable estimate for comparing the memorization effort across different access cost of onscreen information during the early stages of practice.


Keywords: ACT-R declarative memory, Proactive Interference, Memorization Effort, User Interface

## Introduction

Onscreen information is an important part of daily life today - On one hand, they are prevalent in handheld devices like smart-phones and tablets; On the other hand, they can also be found in critical displays in aircraft and other machinery. The screens usually display a structured set of items for the user to interact with. When interacting, it is rare that users remember the position of every item in the set perfectly. One explanation for this forgetting effect is proactive interference caused by distractor items seen during the visual search for the desired item. Proactive interference causes loss of memory activation. In contrast, explicit memorization of item locations helps to mitigate the effect of such interference. People exert mental effort in order to accomplish such memorization.

A study in flight simulation training (Waldron et al., 2008) found that temporarily decreasing the availability of onscreen information for pilots orients pilots more towards memory-based interaction strategies. This in turn helps them better remember critical information such as the aircrafts' location. The study established that an increase in information access cost increases the perceptual-motor effort. This normally encourages users to choose the highest performance option of using fewer perceptual-motor operations but more memory operations, even if memory retrieval is imperfect.

Rowe et al. (2008) empirically suggested that "practice" and "memorization" positively influence visuo-spatial learning while "proactive interference" impacts it negatively. On the other hand, Altmann et al. (2002) proposed a theory that not only holds proactive interference but also "decay" (i.e. loss of memory activation with passage of time) responsible for forgetting. Taking into account the mutually constraining effects of "practice", "memorization" effort, "decay" and "proactive interference", an integrated, yet simple and easily applicable performance model is possible that would reflect the effect of these phenomena on visuo-spatial learning.

Following this idea, we propose a simple mathematical model of visuo-spatial learning that combines the effect of "practice" in terms of practice time, the effect of "decay" in terms of a small numeric constant, the effect of "proactive interference" in terms of visual search cost, and the effect of "memorization" effort in terms of a newly introduced model parameter, an effort factor, explained later. All these effects are expressed in a single equation of memory activation. To achieve this goal, we adapt an existing memory activation equation of ACT-R theory developed by Anderson et al. (1998). We focus on the cognitive aspects of interaction more than the perceptual-motor control complexities in our model. Therefore, we leverage the empirically proven axioms of ACT-R theory that the time cost of a visual encoding is a constant and that a motor response can be modeled as an average value, according to the task specific behavior, such as a mouse movement.

Guided by Altmann et al. (2002), we implement our mathematical model in a spreadsheet and validate it against previous empirical data collected by others.

## ACT-R Theory of Declarative Memory

The ACT-R theory by Anderson et al. (1998) describes a modular system that aims to replicate the human mind. The theory is a framework of mathematical equations that models the neural computations in order to realize human dynamic behavior.

The core of ACT-R declarative memory builds upon the notion of memory activation. It posits that memory encodings of items have different levels of activation to reflect their past use: items that have been used recently or items that are used very often receive a high activation. This activation decays over time if the item is not used. When the cognitive system needs to retrieve an item, memory returns the one with the highest activation at that instant. The job of memory retrieval is complicated by the noise in activation levels, which can temporarily make an item more active than the current one, or which can temporarily push all items below a threshold, thereby making the cognitive system transiently unable to recall information (Altmann et al., 2008; p. 604). Furthermore, the activation of an item controls its speed of retrieval. We focus on the following three equations behind the ACT-R declarative memory system that we leverage in our current work.

## ACT-R Activation Equation

The equation describing the activation, $A$, of an item in the memory is given by
$A=B+\varepsilon \quad$ Activation Equation
where $B$ is the base-level activation of the item discussed later in detail and $\varepsilon$ is the noise component. Noise is assumed to cause transient fluctuations in activation levels. Guided by Altmann et al. (2002), we implement the noise $\varepsilon$ as a constant for our modeling purposes. In the complete ACT-R memory model, environmental context and relevance to the current goal also influences the activation of an item (Gray et al., 2006, p. 481). However this component introduces additional complexity not relevant to our modeling effort in this work. Being guided by Gray et al. (2006) we have therefore omitted the component here.

## ACT-R Base-Level Activation Equation

The equation describing the base-level activation of an item in memory is given by

$$
\begin{equation*}
B=\ln \left(\sum_{j=1}^{n} t_{j}^{-d}\right) \quad \text { Base-Level Activation Equation } \tag{2}
\end{equation*}
$$

where $n$ is the number of "practices" of the item completed so far, $t_{j}$ is the age of the $j$-th practice of the item, and the negative exponent $d$ is the decay constant that controls how quickly the activation decreases. As postulated by ACT-R theory, the $d$ term thus models the loss of memory strength with the passage of time. The equation therefore represents the strength of a memory item as the sum of a number of individual memory strengthening, each corresponding to a past practice event. It implies that each time an item is practiced, the activation of the item receives an increment in strength that decays away as a power function of time.

## ACT-R Reaction Time Equation

The activation of an item discussed earlier controls its speed of retrieval. The time required for the declarative memory to respond to a request (recognition or recall) for an item is given by the following equation:
$R T=I+F * e^{(-f * A)}$ Reaction Time Equation
where $I$ is an intercept time reflecting the time cost of perceptual (visual) encoding and motor response. $F$ is the latency factor, and maps activation to time. $f$ is the latency exponent. The purpose of parameters $F$ and $f$ is only to scale the time to retrieve an item from memory. They remain fixed across all experimental conditions.

The time cost of a visual encoding is set at 185 ms which is taken from the estimate used by ACT-R (Anderson \& Lebiere, 1998, pp. 150-151) for human attention to move to an object at a location.

The time cost of a motor response is set according to the task specific behavior. The task we model involves finding a pre-cued item on a structured layout of graphical buttons presented on a computer screen and then selecting it by clicking on the appropriate button using a mouse (Ehret, 2000, 2002). Guided by Ehret (2000), Gray et al. (2006) and Card et al. (1978), we estimate the average time cost of a motor response to be 300 ms for our modeling purposes.

## The Model

We next propose our extension to the base-level activation equation in order to account for the effect of proactive interference and memorization effort. We do so largely by adapting existing cognitive constructs rather than developing entirely new ones.

## Proactive Interference Modeling

Our approach adapts ACT-R's classic model of memory strength to account for proactive interference. In other words, we account for the effects of distractors that get visually encoded or cumulated before the encoding or accumulation of the target item, during a visual search. We accomplish this by replacing the decay constant, $d$, of the base-level activation equation, with a function consisting of a constant term and a varying term. The constant term models the loss of memory strength with passage of time as before. The new varying term models the loss of memory strength due to proactive interference. Our proposal for modeling the combined effect of decay and interference on memory activation is in line with the observations of Altmann et al. $(2002,2008)$ which indict both decay and proactive interference for forgetting.

The varying term we propose is governed by the visual search cost - the number of distractors that get visually encoded prior to encoding the target item when one tries to find an item on a user interface. The encoded number of distractors during a search contributes to a measure for the proactive interference effect: The lower the number of distractors visually encoded during a search for a target item, the lower should be the "loss" of activation of the target item. Hence, the next recall of that item will be affected by its higher activation, leading to the lowering of its retrieval time. This will show an improvement in "search and selection" performance time during exploration of the interface. Our hypothesis is grounded in the primary research result of Underwood (1957) on proactive interference, namely, the effect that the number of previously learned items has on the recall of the target item: The lower the number of previously learned items is, the lower is the forgetting effect and therefore the lower is the recall latency for the target item.

We propose a decay rate, $d_{j}$, calculated for an item, after $j$ practices of the item are completed, as follows:

$$
\begin{equation*}
d_{j}=h+0.5 * X_{j-1} / N \quad \text { Decay Rate Equation } \tag{4}
\end{equation*}
$$

where $h$ represents the time-based decay constant, the fraction 0.5 is a scaling factor (our choice of 0.5 is explained in the next paragraph), $N$ is the total number of items on the layout and $X_{j-1}$ is the number of distractors visually encoded at the time of $j^{\text {th }}$ practice. Naturally, $j$ has to be larger or equal to 1. $X_{0}$ denotes the number of distractors encoded at the first practice. When $X_{j-1}$ is 0 , i.e. when the user is able to complete the task by direct recall, without going through any explicit visual search, the decay rate equation degenerates to $d_{j}=h$. This implies that, in the absence of the impact of distractors, decay in activation occurs only
with the passage of time as modeled by the classic baselevel activation equation.

We introduce the varying term $0.5 * X_{j-1} / N$ to represent the loss of memory activation due to proactive interference. It transforms the number of distractors, $X_{j-1}$, to a "decay" value suitable for ACT-R theory. We assume such values to be ranging from 0 to 0.5 : Since 0 implies no decay, it can be considered as a lower bound. The value of 0.5 is used as the default decay constant in the classic ACT-R theory (see Anderson et al., 1998). Therefore 0.5 can be considered as a valid upper bound for our work. The ratio $X_{j-1} / N$ ranges from 0 to 1 . Consequently, the varying term $0.5 * X_{j-1} / N$ results in a value in the desired interval, 0 to 0.5 . The $0.5 * X_{j-1} / N=0.5$ refers to a situation where the maximum possible number of distractors is encountered, i.e. when $X_{j-1}=N$, leading to the highest level of proactive interference effect. This, in turn, reduces the term to the maximum of 0.5 . On the other hand, $0.5 * X_{j-1} / N=0$ implies an absence of impact from distractors, and therefore no proactive interference effect as a consequence. This occurs when the user is able to complete the task by direct recall.

Our model of proactive interference is adapted from the model of Das et al. (2010). Our work is a significant improvement over their model of proactive interference because firstly, our decay rate equation contains less number of free parameters (decay constant $h$ is the only free parameter in our equation) and secondly, our equation is constrained by the total number of items, $N$, of a layout under scrutiny. Consequently, the chances for data overfitting decrease significantly in our model.

## Memorization Effort Modeling

Our modeling of memorization effort is guided by the soft constraints hypothesis of Gray et al. (2006). The soft constraints hypothesis is a rational analysis approach which proposes that the mixture of perceptual-motor and cognitive resources allocated for interactive behavior is adjusted based on temporal cost-benefit tradeoffs, such that the least-effort path of executing the visuo-spatial task at hand, gets implicitly chosen. As perceptual-motor effort increases, users will normally choose the least-effort option of fewer perceptual-motor operations and more memory operations, even if the memory retrieval is imperfect. We term the effort exhausted in carrying out the memory operations as "memorization effort".

The soft constraints hypothesis concludes that the tradeoff between selecting the perceptual-motor versus cognitive behavior minimizes the total effort (and hence performance cost) measured in the currency of time (Gray et al., 2006, p. 463). Motivated by the hypothesis, we introduce a parameter in the base-level activation equation of ACT-R (Equation 2) as a coefficient of practice time and include it inside the logarithmic term (shown later in Modified BaseLevel Activation Equation). We call this novel parameter effort factor. We hypothesize the effort factor to be the "temporal" representation of the memorization effort expended to accomplish a visuo-spatial learning task. The works of Anderson (1983, p. 277) as well as Stewart et al. (2007, p. 235), also motivate our choice for the adoption of
an effort factor, as they suggested the usage of a cost factor similar to ours, albeit in different domains.

## Modified Base-Level Activation Equation

With the decay rate equation and the effort factor parameter conceptualized, we modify the base-level activation equation (Equation 2) to

$$
\begin{equation*}
B=\ln \left(k \sum_{j=1}^{n} t_{j}^{-d_{j}}\right) \tag{5}
\end{equation*}
$$

## Modified Base-Level Activation Equation

Equation 5 is obtained by adding two new elements $d_{j}$ and $k$ to Equation 2. We explain the new elements below.
$d_{j}$ describes the new decay rate equation (Equation 4) that sums up two terms: one representing the traditional timebased "decay" constant and the other representing the "loss of activation due to proactive interference".

The element $k$ is the aforementioned effort factor parameter. We explain $k$ in the context of learning layouts that vary in the information access costs (henceforth referred to as "access cost") associated with their items. The access cost differs in terms of representativeness of item labels. Our context of learning accounts for the fact that the total practice time for learning is held constant across all layouts (i.e. for every level of access cost).
If all model parameters, except $k$, in Equations 1, 3, 4 and 5 are left at fixed values across layouts that differ in access costs, then we hypothesize two properties about $k$ while comparing layouts in terms of reaction time estimates $(R T)$ of Equation 3 as follows:
(i) First, we hypothesize that one value of $k$ corresponds to one particular layout, i.e. one particular access cost condition.
(ii) Second, a lower value of $k$ would correspond to higher memorization effort whereas a higher value of $k$ would correspond to lower memorization effort. The Appendix provides an argument for this.

Our modified base-level activation equation is therefore a hypothesis that accounts for the combined effect of "practice time", "memorization" effort, "proactive interference" and "decay" on visuo-spatial learning performance. We validate our hypotheses later in this work.

Our model of memorization effort is adapted from the work of Das et al. (2012). Their model did not account for proactive interference which is the central constraint compared to decay in learning in situations where learning is affected by distractors (Altmann et al., 2002, 2008). Moreover, they had varied the values of multiple model parameters across different conditions of access cost leaving their model vulnerable to overfitting.

## Validation

In order to validate our model, we use existing experimentally derived data sets for human performance over several practice sessions for location learning of items in a stable layout. Our goal is to focus on the novice to expert transition because of two reasons. On one hand, the effect of proactive interference is most pronounced during
this transition phase. On the other hand, the effect of memorization effort to overcome such interference is also evident in this stage. We therefore concentrate on modeling early sessions of skill development. Each data set we validate against corresponds to a certain access cost in terms of label representativeness of graphical buttons that were laid out on a computer screen. The task we model involves finding a pre-cued button and selecting it using a mouse.

We next explain the rationale behind all model parameter values that were fixed across all experimental conditions.

The time-based "decay" constant $h$ in the decay rate equation was fixed at $h=0.058$. We are motivated here by Pavlik et al. (2005, p. 572), who used it as a decay intercept albeit in a different modeling context. In the absence of any inter-trial data in the empirical study that we validate against, we assume that there have been insignificant pauses between any two consecutive trials. Hence, a relatively small value for the time-based "decay" constant is appropriate, implying that the decay due to passage of time had been minimal. Since the focus of our decay rate equation is to model the effect of proactive interference, we place greater emphasis on the role of distracting information. In this regard, we are motivated by the discourse of Altmann et al. (2002) who argues for the influential role of proactive interference in forgetting compared to the role of decay in the domain of distractoraffected learning. Our choice of a very small value of the time-based "decay" constant is therefore appropriate.

The activation noise $\varepsilon$ in the activation equation was fixed at $\varepsilon=0.28$, a value in line with other applications of this equation in the domain of graphical user interface (e.g., Gray et al., 2006).

The latency factor $F$ in the reaction time equation is left at its default value of $F=1 \mathrm{sec}$, as per classic ACT-R theory.

The latency exponent $f$ in the reaction time equation is fixed at $f=0.65$. On carrying out sensitivity analysis, we found that setting $f$ at 0.65 instead of 1 substantially reduces the root-mean-square error (RMSE) value between the human data and its corresponding model data. It has very negligible influence on the correlation between them.

As we discuss below, the effort factor $k$ of the Modified Base-Level Activation Equation is the only parameter that we varied across conditions in order to account for the relative differences in memorization effort spent in learning layouts with different access costs (conditions).

## Circle of Buttons Experiment

Knowing an object's location can reduce a user's task time, errors, and frustration. As the number of screen objects increases, so does the utility of location knowledge. Ehret (2002) carried out an experiment that tests how well users learn the location of buttons arranged in a circle on a computer screen and how the mechanisms underlying location learning interact with the level of meaningfulness of button labels. He used a "search and select" task in which, for a given trial, participants were presented a particular color and were required to find and click the button associated with that color. The correct button was one among the twelve buttons that remained in constant positions throughout the experiment. The contour and shape
of every button was always visible across all conditions (Ehret, 2000; Figure 2, p. 27). To discourage errors, when participants clicked the wrong button the computer would beep five times, a dialog box would appear, and the trial would have to be repeated (Ehret, 2002; p. 212).

Ehret's observations were point-of-gaze data collected via an eye-tracker. In order to validate our model we extracted three data sets from his observations. The data sets were mean "visual search and select" time (reaction time) from an experiment, limited to the first 10 sessions of practice, since learning plateaued off after the tenth session. In his study, Ehret (2002; 2000, p. 19) had reported two costs, the visual search cost which is the number of buttons visually encoded before the target button is found and the verification time, which is the time required to decide whether the button visually encoded is really the target or not. For a given session, we arrived at the mean human reaction time per button by multiplying the mean visual search cost with the mean verification time corresponding to that session.

The three data sets differed in the level of meaningfulness of labels associated with the buttons. The first set of data was obtained while searching for a pre-cued color in buttons labeled with the name of color written in English. The aim was to have a meaningful association between a color and the button representing the color. The second set of data was obtained while searching for buttons labeled with arbitrary icons. The aim was to reduce the meaningfulness of the association between a color and the button representing it. The third set consisted of the reaction times for searching and selecting a pre-cued color among buttons with no labels on them. The aim was to eliminate any meaningfulness of the association between a color and the button representing it. The data sets thus contain three sets of reaction times corresponding to the three different levels of difficulty in accessing information: textual label, arbitrary label and invisible label. Each condition therefore represented a certain level of access cost, the textual label condition being the lowest cost condition among them. The total practice time was held constant across all conditions. It is to be noted that for the arbitrary and invisible label conditions, a tooltip was provided for each button to aid the subject, if memory failed. Accessing the tooltip for a button revealed a small rectangle containing the color associated with it. The cost of accessing this tip was a one-second delay between moving the mouse cursor to the button and the appearance of the tooltip.

Our choice of data aligns with our modeling objective. We aim to model the combined effect of visual search cost (the surrogate of proactive interference) as well as memorization effort on reaction time, over a reasonable number of practice sessions. Ehret's data shows that for any given access cost condition, the visual search cost decreases over practice sessions implying that proactive interference decreases with practice. However, Ehret's data further shows that during the search for a pre-cued color, as the access cost increased from textual to arbitrary to invisible label conditions of buttons, so did the time to visually verify and decide (verification time) whether a button currently under scrutiny is indeed the target or not, at any given session. The verification time was observed to be the lowest
for the textual label condition and highest for the invisible label condition. In other words, the layouts with higher access cost featured higher verification time to identify the correct item, implying higher effort to learn those layouts compared to the ones with lower access cost. As posited by the soft constraint hypothesis and given the same amount of practice time across all conditions, the higher perceptual cost of arbitrary and invisible label conditions results in a higher memorization effort for those label conditions compared to the memorization effort required for the textual label condition.

For our validation, we had to make a few assumptions, as certain information was not mentioned explicitly in the work of Ehret (2002). The assumptions are the same across all conditions as follows: Each practice session took 37.5 seconds to complete - since 16 sessions took 10 minutes or 600 seconds as expressed in a related work by Ehret (2000, p. 136). We also assume the inter-session periods to be constant. Also, except for the target pre-cue, we assume that environmental context cuing is minimal and irrelevant for our purposes.

## Validating the Proactive Interference Effect

We provide an example scenario on how the effect of proactive interference on spatial learning can be modeled using our new model. Ehret (2002) had an onscreen layout of graphical buttons labeled with icons where each icon is arbitrarily associated with a color. A subject's task was to visually search for a pre-cued color among the buttons and click the appropriate button when found using a mouse. The pre-cued color always appeared at the center of the circle. In case the subject's memory failed to recall the color associated with a button, she could access the button's tooltip to know its color by moving the mouse cursor over it. The tooltip appeared after a one-second delay once the mouse cursor was moved to the button.

The mean numbers of distractors measured in Ehret's experiment in the arbitrary label condition are 5.27, 2.93, $2.58,2.34,2.31,1.61,1.49,1.31,1.36$ and 1.14 corresponding to sessions 1 to 10 . We input these numbers in the decay rate equation of our model to obtain the mean activation value per item for each session. We adjust the value of $k$ in our model to 0.068 for the experimental condition (i.e. arbitrary labeling condition). The other model parameters stay fixed at the values discussed earlier. We fit our model to the empirical reaction time for a button. We found the $\mathrm{R}^{2}$ of the fit to be .993 implying a qualitative correspondence between human and model results.

The effect of proactive interference was also evident in the textual label condition. After substituting the values of mean numbers of distractors for this condition (measured in Ehret's experiment) in the decay rate equation, we again found a close match between the human and model results with $\mathrm{R}^{2}=0.978$. Our adjusted value of $k$ was 0.500 in this condition.

As apparent from the decay rate equation, a change in the number of distractors changes the decay rate. While modeling the proactive interference, we noticed that the mean number of distractors per item, $X_{j-1}$ in the decay rate equation influences the shape of the curve at each sessionpoint. A small change in the decay rate, $d_{j}$, (at the level of
$10^{-2}$ ) is found to have noticeable impact on the reaction time estimates. This is particularly true for the first few sessions of practice.


Figure 1. Reaction times per item (button) for textual, arbitrary and invisible label conditions.

## Validating the Comparison of Memorization Effort

Figure 1 shows the fit of our model to the human data in terms of reaction times. We compared the effort factor $k$ for the invisible label condition against the textual label condition. We found $k=0.056$ for the difficult to access invisible labels, compared to $k=0.500$ for the easily accessible textual labels. Furthermore, $k$ was 0.068 for the difficult to access arbitrary labels, compared to $k$ being 0.500 for the easy to access textual labels. The comparison of $k$ in both instances thus points to lower values of $k$ for layouts with high access cost (high perceptual cost) compared to the conditions where relevant information is easily available in the environment. We therefore conclude that the comparison of memorization effort via our new effort factor $k$ follows the soft constraint hypothesis to a significant extent.

With $\mathrm{R}^{2}=0.978, \mathrm{RMSE}=0.215$ for the textual, $\mathrm{R}^{2}=$ 0.993 , $\mathrm{RMSE}=1.153$ for the arbitrary and $\mathrm{R}^{2}=0.941$, RMSE $=0.785$ for the invisible conditions, the correlation between the human and model data were good. The RMSE as a percentage was $13 \%$ for textual and $15 \%$ for invisible condition. However, the percentage RMSE for arbitrary condition being $38 \%$ was higher than the $20 \%$ mark suggested by John and Newell (1989). The RMSE for the arbitrary condition therefore implied a high error.

## Discussions

Our work in this paper introduces two mathematical terms, one to account for the effect of "proactive interference" (PI) and the other to account for the effect of "memorization effort". We add them to an existing memory activation equation of ACT-R theory that hitherto accounted for the effects of only "practice" and "decay".

In this work, we have left all but one model parameter fixed across all conditions, thereby omitting the scope of overfitting significantly. The effort factor $k$ is the only model parameter that we varied in order to reflect the
differences in the memorization effort across different accessibility conditions.

Earlier, Altmann et al. (2002) had used ACT-R theory to mathematically model the effect of PI on recall probability. On the other hand, we have mathematically modeled the effect of PI on response latency.

Our modulation of decay rate to reflect PI is motivated by the approach of previous researchers such as Pavlik et al. (2005), Cochran et al. (2006) who had modulated the decay rate to model phenomena, albeit different from PI.

Previously, Ehret $(2000,2002)$ had used ACT-R theory to model memorization effort. Unlike ours, his approach involved computer-based simulation. In this work, we provide an alternative look at Ehret's modeling endeavor. We do so through a mathematical model.

Initially, to keep our modeling endeavor simple, we started out by creating separate models of proactive interference as well as memorization effort. While developing the standalone model of proactive interference, we tried to leave the effort factor constant across all conditions. On the other hand, while developing the standalone model of memorization effort, we tried to leave the decay rate constant across all conditions. In both cases, however, we were unable to identify fixed values for model parameters. Rather, every "access cost" condition demanded a separate set of values for multiple model parameters to fit the data in a satisfactory manner. This motivated us to model proactive interference and memorization effort in a unified way.

Our mathematical model has its limitations. (i) At any given trial for searching a target location on a layout, when the number of distractors $X_{j-l}$ encountered is much less than the total number of items $N$ on the layout, we assume that proactive interference owing to that trial has been negligible. This situation may arise when $N$ is very large. Further investigation is warranted to identify a practical upper limit on $N$. (ii) Our model is restricted to comparing layouts that have the same number of items in them. (iii) We do not consider the level of similarity between distractors and target. (iv) Increased recall latency observed in high PI conditions can be caused by interference of the target with distractor activations at the time of retrieval. We have not considered that. (v) ACT-R theory has a threshold parameter that specifies a minimum activation below which an item is invisible to the cognitive system. Similar to Altmann et al. (2002), we assume no such threshold. As the threshold parameter is not a variable in the equations we use, this assumption does not impact our work directly.

Our model concentrated purely on the cognitive aspects of interaction; thus it did not model the motor control complexities involved in the spatial search and selection processes on graphical user interfaces. In reality though, these are all important factors that influence the overall user experience.

The advantages of our proposal are its simplicity and transparency. However, it is an ad hoc alternative focused at solving a specific problem in a specific way. We do not claim that we have arrived at a "generic" solution.

## References

Altmann, E. M. et al. (2002). Integrating decay and interference: A new look at an old interaction. 24th CogSci. pp. 65-70.
Altmann, E. M. \& Gray, W. D. (2008). An integrated model of cognitive control in task switching. Psychol. Rev. 115, 602-639.
Anderson, J. R. (1983). A Spreading Activation Theory of Memory, J. Verbal Learning and Verbal Behavior, 22, 261-295.
Anderson, J. R. \& Lebiere, C. (1998). The atomic components of thought. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
Card, S. K., English, W. K., \& Burr, B. J. (1978). Evaluation of mouse, rate-controlled isometric joystick, step keys and text keys for text selection on a CRT. Ergonomics, 21, 601-613.
Cochran, R. E. et al. (2006). Modeling emotion: Arousal's impact on memory. 28th CogSci. pp. 1133-1138.
Das, A. et al. (2010). Proactive Interference in Location Learning: A New Closed-Form Approximation. $10^{\text {th }}$ ICCM. pp. 37-42.
Das, A. et al. (2012). Comparing cognitive effort in spatial learning of text entry keyboards and ShapeWriters. AVI. pp. 649-652.
Ehret, B. D. (2000). Learning where to look: The acquisition of location knowledge in display-based interaction. Doctoral Dissertation. George Mason University.
Ehret, B. (2002). Learning Where to Look: Location Learning in Graphical User Interfaces, CHI 2002, 211-218.
Gray, W. D. et al. (2006). The soft-constraints hypothesis: A rational analysis approach to resource allocation for interactive behavior. Psychological Review, 113, 461-482.
John, B. E., \& Newell, A. (1989). Cumulating the science of HCI: From S-R compatibility to transcription typing. CH1 1989, pp. 109-114.
Pavlik, P. I., Jr., \& Anderson, J. R. ( 2005). Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. Cogn. Sci., 29, 559-586.
Rowe, G. et al. (2008). Age differences in visuospatial working memory. Psychology and Aging, 23, 79-84.
Stewart, T. C. and West, R. L. (2007). Deconstructing and reconstructing ACT-R: Exploring the architectural space. Cognitive Systems Research, 8, 227-236.
Underwood, B. J. (1957). Interference and forgetting. Psychol. Rev. 64, 49-60.
Waldron, S. M. et al. (2008). Designing information fusion for the encoding of visual-spatial information. Ergonomics, 775-797.

## Appendix

If the effort factor $k$ is varied while leaving other model parameters at fixed values across different accessibility conditions, then a lower value of $k$ would correspond to higher memorization effort whereas a higher value of $k$ would correspond to lower memorization effort. The reason is as follows. A lower $k$ (in Equation 5) results in a higher $R T$ (in Equation 3). Higher values of $R T$ s are typically evident in the early stages of practice for layouts with higher access costs (see the empirical data in Ehret, 2002). However according to the soft constraint hypothesis, learning a layout with higher access cost would require a higher number of memory operations compared to perceptual-motor operations. Consequently, we conclude that a lower value of $k$ refers to a higher number of memory operations and therefore reflects higher memorization effort. In contrast, a higher value of $k$ refers to a lower number of memory operations and therefore reflects lower memorization effort.

# Integrating Cognitive Principles to Redesign a Middle School Math Curriculum 

Jodi L. Davenport (jdavenp@wested.org)<br>WestEd STEM Program, 300 Lakeside Drive, $25{ }^{\text {th }}$ Floor<br>Oakland, CA 94612 USA

Yvonne S. Kao (ykao@wested.org)<br>WestEd STEM Program, 400 Seaport Court, Suite 222<br>Redwood City, CA 94063 USA

Steven A. Schneider (sschnei@wested.org)<br>WestEd STEM Program, 400 Seaport Court, Suite 222<br>Redwood City, CA 94063 USA


#### Abstract

Does a middle school mathematics curriculum that is redesigned using principles based in cognitive research improve student outcomes? To test whether research can be effectively translated into practice, the Connected Mathematics Project 2 (CMP2) curriculum was revised according to four principles 1) integrating visual with verbal information, 2) prompting for self-explanation of correct and incorrect worked examples, 3) spacing learning over time, and 4) using formative assessment. This study of $6^{\text {th }}$ grade and $8^{\text {th }}$ grade mathematics education addresses the research question: "Do students who are exposed to specific redesigned CMP2 curriculum modules (treatment) exhibit greater improvements in mathematics performance in the module-specific content area than their counterparts exposed to the regular CMP2 curriculum (control)?" Preliminary analyses show statistically significant effects of the redesigned CMP2 units in three of the four curricular units in this study.


Keywords: cognitive psychology; mathematics; math education; education; spaced learning; formative assessment; worked examples; visual representations

## Introduction

Lab-based research in cognitive and learning sciences has led to a number of recommendations for improving learning and instruction (e.g., Pashler et al., 2007). Tightly controlled experiments have shown that learning can be enhanced with strategies such as mapping between visual representations, prompting for explanation of worked examples, using quizzing to promote learning and spacing practice opportunities over time. The vast majority of studies focus on specific strategies in isolation rather than how principles may be combined. If these research findings are to be meaningfully applied to classrooms, the synergistic effects of the strategies must be tested in real-world settings. In the current paper, we describe a large-scale effort of the National Center on Cognition and Mathematics Instruction (Math Center) in the United States to bridge research and practice by applying cognitive principles to redesign an existing mathematics curriculum and testing the efficacy of these materials.

To test the synergistic effects of research-based instructional strategies, the Math Center applied four
principles to redesign Connected Mathematics Project 2 (CMP2), a widely-used middle school (grades 6-8) mathematics curriculum. The Math Center team selected cognitive-based principles shown to improve student learning: 1) integrating visual with verbal information to promote the integration of concepts, 2 ) prompting for selfexplanation of correct and incorrect worked examples, 3) carefully spacing the learning of critical content and skills over time, and 4) using quizzes to provide focused feedback and adjust instruction to the needs of students.

The CMP2 curriculum is an NSF-funded, research-based curriculum for grades 6-8 that covers topics emphasized in both national and state standards and aligns well with key ideas from the NCTM (2006) Focal Points. Key features of the curriculum are that it (1) is organized around important mathematics ideas and processes, e.g., number sense, symbolic reasoning, and probability, (2) is problemcentered, and (3) builds and connects concepts across problems, units, and grades. Each year of the curriculum is divided into eight units; each unit includes a student booklet and accompanying teacher materials to support instructional practice.

Applying the principles to revise instructional materials (e.g., the print curriculum) and instructional practice (e.g., what happens in the classroom) required expertise across many fields. Teams devoted to cognition research, mathematics, professional development, and production collaborated to ensure that the revised materials were grounded in the research findings, were mathematically accurate and appropriate (in terms of student development and curriculum standards), were clearly specified for teachers, and were produced with a high level of technical quality. The iterative, multi-layered design process that we have developed for integrating the cognitive principles with the CMP2 curriculum applies not only in the context of mathematics instruction, but also to bridging research with instructional design across content areas.

## The Principles

The following four principles were selected as they have demonstrated effectiveness in student learning, have broad
applicability to instruction, and can be readily implemented in a range of curricular materials.

Integrating Visual and Verbal Information Combining visual information with verbal descriptions serves two important functions in mathematics instruction: 1) ensuring that text for instruction and problem-solving are perceived and understood and 2) promoting fluency in mapping between representations (e.g., equations, diagrams, graphs, or tables). To maximize learning benefits, research suggests that visual and verbal information should be integrated (e.g., Clark \& Mayer, 2003; Larkin \& Simon, 1987; Moreno \& Mayer, 1999) and task-irrelevant information should be removed (e.g., Harp \& Mayer, 1998). Visual cues such as color, proximity and grouping can support integration. Removing "seductive details;" that is, representations that are engaging but only tangentially related to the topic of instruction or the problem at hand (e.g., Harp \& Mayer, 1998), helps learners focus on relevant information. To apply the visual mapping principle, researchers removed irrelevant images, added visual cues (e.g., color), and modified existing images to facilitate mapping.

Worked Examples In mathematics, students must learn to fluently carry out procedures across a variety of problem types. Interleaving problems to solve with worked examples of how to solve a problem improves student learning (Zhu \& Simon, 1987; Clark \& Mayer, 2003). Prompting students to explain worked examples further increases learning by facilitating the integration of new information. (Chi, 2000; Roy \& Chi, 2005). In worked example exercises, students see complete or partially worked out solutions (which can be correct or incorrect) and explain the rationale behind problem solving steps or the error that was made in an incorrect example. Positive effects of interleaving worked examples have been reported in a variety of courses (Clark \& Mayer, 2003; Paas \& Van Merrienboer, 1994; Sweller \& Cooper, 1985). Worked examples are more effective and more efficient for learning and transfer because they allow students to spend limited cognitive resources on understanding the ideas underlying the solutions rather than on generating solutions (Sweller, 1999). Further, explaining both correct and incorrect worked examples promotes greater learning than correct examples alone (Siegler, 2002; Siegler \& Chen, 2008; Rittle-Johnson 2006). To apply the worked examples principle, researchers modified existing homework activities to include worked examples that prompt for self-explanation of problem solving steps.

Spaced Learning and Formative Assessment Extensive research in cognitive psychology has demonstrated large retention advantages when learners have multiple opportunities over time to practice key facts, concepts, and knowledge rather than few instances of "massed" practice, a phenomenon called the spacing effect (Cepeda et al., 2006; Rohrer \& Taylor, 2007). When learners practice recalling and applying relevant information through quizzing, they
are more likely to retain that knowledge for a greater period of time. Spacing instruction and practice reinforces connections between key ideas and promotes transfer.

Periodic testing provides students with opportunities to practice retrieving knowledge, reflect on the state of their knowledge, and transfer knowledge to new problems (Butler \& Roediger, 2007; Roediger \& Karpicke, 2006; Rohrer, 2009). Cycles of feedback and reflection that allow for revision and knowledge updating can help learners master targeted concepts and skills (e.g., Pavlik et al., 2007). Evidence from classroom learning contexts shows that the formative use of assessment can enhance instructional effectiveness (e.g., Black \& Wiliam, 1998); here, formative assessment is defined as a process used by teachers and students that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes. In the revised materials, teachers were provided with quizzes and instruction on how to use feedback formatively in the classroom.

## Method

The design of this study is a within-teacher clusterrandomized trial. The primary research question of this study is: "Do $6^{\text {th }}$ and $8^{\text {th }}$ grade students who are exposed to a redesigned curricular unit (treatment) show greater pre-topost test improvements in mathematics scores than students exposed to the unmodified curricular unit (control)?"

Participants Researchers collected data from $646^{\text {th }}$ grade teachers ( 1270 students at 45 schools) and $568^{\text {th }}$ grade teachers (1180 students at 42 schools). Teachers had prior experience with the CMP2 curriculum and came from a diversity of schools across seventeen states in the United States. Background characteristics of participating teachers and demographic characteristics of their students are presented respectively in Table 1 and Table 2.

Table 1: Professional background of participating teachers.

| Characteristic | $\mathbf{6}^{\text {th }}$ Grade | $\mathbf{8}^{\text {th }}$ Grade |
| :--- | :---: | :---: |
| Majored in math or math <br> $\quad$ education | $27 \%$ | $43 \%$ |
| Advanced degree | $64 \%$ | $66 \%$ |
| Mean years of teaching | 12.3 | 13.7 |
| $\quad$ experience | $(\mathrm{SD}=8.2)$ | $(\mathrm{SD}=7.6)$ |

Table 2: Demographic characteristics of participating students.

| Characteristic | $\mathbf{6}^{\text {th }} \mathbf{G r a d e}$ | $\mathbf{8}^{\text {th }}$ Grade |
| :--- | :---: | :---: |
| Socioeconomically <br> disadvantaged | $41 \%$ | $43 \%$ |
|  | Ethnicity |  |
| White | $67 \%$ | $60 \%$ |
| Black | $10 \%$ | $13 \%$ |
| Hispanic | $14 \%$ | $13 \%$ |
| Other | $9 \%$ | $14 \%$ |

Materials Two $6^{\text {th }}$ grade units and two $8^{\text {th }}$ grade units from the CMP2 curriculum were revised according to the cognitive principles described above. The $6^{\text {th }}$ grade units used in this study were Bits and Pieces III (decimals and percents) and Covering and Surrounding (area and perimeter). The $8^{\text {th }}$ grade units were Shapes of Algebra (linear equations and coordinate geometry) and Say it with Symbols (expressions and equations). Teams of researchers were formed for each of the principles. The cognitive research teams developed rubrics to identify whether the existing materials aligned with the cognitive design principles, and if not, to specify how the materials would be altered to be in compliance. Next each team made sequential revisions to the CMP2 materials. Changes that overlapped with other principles were discussed and resolved in biweekly meetings.

1. Concy Island Park wants a bumper-car ride with 24 square meters of floor space and 22 meters of rail section.
a. Sketch some floor plans for this request.
b. Describe the bumper-car ride in terms of its area and perimeter. Report what cach measure tells you about the ride.


Figure 1: A problem from the original Covering and Surrounding unit.

1. Concy lsland Park wants a bumper-car ride with 24 square meters of floor space and 22 meters of rail section.
a. Sketch some floor plans for this request.

Dominick completed the first two floor plans in a correct way, but his third plan does not meet the requirements. Look at his work, and then answer the questions below.


Dominick's third floor plan meets one requirement, but not the other. Which one does it fail to meet? How can you tell?
b. Describe the bumper-car ride in terms of its area and perimeter. Report what each measure tells you about the ride.
Figure 2: The revised version of the problem in Figure 1. A worked example has been incorporated into part a and the park photograph has been removed.

The mathematics team reviewed the revised curricular materials to ensure mathematical accuracy and
appropriateness. Finally, the production team worked with the cognitive and math content teams to clarify design decisions as necessary. Examples of the original and revised curriculum materials are shown in Figures 1 and 2. Concurrent with the production of the materials, the professional development team met to develop measures of fidelity of implementation and to identify effective ways to communicate the underlying rationale and practical implementation of the cognitive design principles to the participating teachers.

Design This study used a within-teacher design: each teacher provided data from two units of CMP2, one revised and one control. Whether a given unit was used in its original or redesigned format was counterbalanced across participants. Teachers were randomly assigned to one of two groups, A and B, as depicted in Table 1 below. Group A served as the experimental group for one of the curriculum units and Group B served as the experimental group for the other. When multiple teachers taught at the same grade level in the same school, half the teachers at the school were assigned to group A and half to group B.

Table 1: Assignment of teachers to group.

| Group | Treatment Unit | Control Unit |
| :--- | :---: | :---: |
|  | $6^{\text {th }}$ Grade |  |
| A | Bits and Pieces III | Covering and Surrounding |
| B | Covering and Surrounding | Bits and Pieces III |
|  | $8^{\text {th }}$ Grade |  |
| A | Say it with Symbols | Shapes of Algebra |
| B | Shapes of Algebra | Say it with Symbols |

Procedure All teachers attended a two-day, online, professional development workshop to introduce them to the research-based principles and implications for instructional materials and practice. During these sessions, teachers worked as groups and in pairs to plan instruction for the treatment units. Teachers administered pre-tests for both study units immediately following the professional development. Teachers then taught CMP2 in their normal curriculum order, administering post-tests immediately upon completion of each study unit, treatment and control. Teachers completed weekly instructional logs for both the treatment and control units, in which they described their implementation of the unit, including any application of the research-based principles. This enabled researchers to measure fidelity of implementation and estimate the achieved relative strength (Hulleman \& Cordray, 2009) of the treatment intervention by comparing the degree to which teachers implemented the research-based principles in their treatment vs. their control units.

## Measures

Researcher-developed assessments were used to evaluate student learning. The content of each curriculum unit was
carefully mapped in order to assess the content areas, skills, and contexts presented to students. The same mapping was performed on the assessments to ensure they were wellaligned to the curriculum unit. All items were field-tested to establish reliability. Assessments included approximately 16 multiple-choice items and two open-ended items. Approximately half of the items were derived from existing CMP2 materials, and the remaining items were taken from state, national and international standardized tests.

For each unit, two test forms were created with approximately half of the multiple-choice and both openended items as linking items. Test forms were randomly assigned by class such that half of the classes took form A for pretest and form B for posttest, and the other half of the classes took form B for pretest and form A for posttest. Open-ended items were scored by trained raters using a standardized holistic rubric. Researchers computed weighted kappas to measure both intra-rater and inter-rater reliability. Intra-rater reliability ranged from 0.90 to 0.99 . Inter-rater reliability ranged from 0.83 to 0.94 .

## Data Analysis

Item response theory (IRT) was used to equate the test scores across forms (Cook \& Eignor, 1991). A partial credit model was used to generate item parameters, scale scores ${ }^{1}$ for students, and assessment reliabilities, which ranged from $0.55-0.74$ on pre-test and $0.77-0.82$ on post-test.

ANCOVA models were used to estimate the treatment effects, controlling for pre-test scale scores and socioeconomic status. ANCOVAs for each unit were performed on students with complete demographic information and who completed both the pre-test and the post-test for that unit. The ANCOVA sample for each unit is shown in Table 2-the ANCOVA samples do not differ statistically from the full sample in their demographic makeup.

Table 2: ANCOVA sample for each unit.

| Unit | Control | Treatment |
| :--- | :---: | :---: |
|  | $6^{\text {th }}$ Grade |  |
| Covering and Surrounding | 481 | 384 |
| Bits and Pieces III | 431 | 496 |
|  | $8^{\text {th }}$ Grade |  |
| Shapes of Algebra | 349 | 371 |
| Say it with Symbols | 386 | 435 |

## Results

## $6^{\text {th }}$ Grade

To provide context for the IRT scale scores, traditional descriptive statistics for the overall change in students' performance from pre-test to post-test are shown in Table 3.

[^50]Students made meaningful gains from pre-test to post-test on both units.

Table 3: Mean $6^{\text {th }}$ grade assessment performance, all students

| Test section | Pre-test | Post-test |
| :---: | :---: | :---: |
|  | Covering and Surrounding |  |
| Multiple-choice | $41.2 \%$ | $61.0 \%$ |
| $\%$ correct | $(\mathrm{SD}=16.1 \%)$ | $(\mathrm{SD}=21.1 \%)$ |
| Open-ended | 1.3 | 2.2 |
| out of 7 points | $(\mathrm{SD}=0.9)$ | $(\mathrm{SD}=1.3)$ |
|  | Bits and Pieces III |  |
| Multiple-choice | $47.2 \%$ | $65.1 \%$ |
| $\%$ correct | $(\mathrm{SD}=20.9 \%)$ | $(\mathrm{SD}=23.1 \%)$ |
| Open-ended | 1.6 | 2.7 |
| out of 8 points | $(\mathrm{SD}=1.7)$ | $(\mathrm{SD}=2.1)$ |

Post-test scale scores for both $6^{\text {th }}$ grade units, holding pretest scores constant, are shown in Figure 3.


Figure 3: Post-test IRT scale scores for the $6^{\text {th }}$ grade units. Error bars represent $\pm 2$ standard error.

ANCOVA results are presented in Table 4 (mean-square error is shown in parentheses). Pre-test was significantly associated with post-test scores in both units.

$$
\text { Table 4: } 6^{\text {th }} \text { grade ANCOVA results }
$$

| Source | $\boldsymbol{d} \boldsymbol{f}$ | $\boldsymbol{F}$ | $\boldsymbol{p}$ | Partial $\boldsymbol{\eta}^{\mathbf{2}}$ |
| :--- | :---: | :---: | :---: | :---: |
| Covering and Surrounding |  |  |  |  |
| Pre-test | 1 | 288.93 | $<.001$ | 0.251 |
| Socioec. disadv. | 1 | 54.50 | $<.001$ | 0.060 |
| Treatment | 1 | 12.78 | $<.001$ | 0.015 |
| Error | 861 | $(0.46)$ |  |  |
| Bits and Pieces III |  |  |  |  |
| Pre-test | 1 | 352.13 | $<.001$ | 0.276 |
| Socioec. disadv. | 1 | 80.31 | $<.001$ | 0.080 |
| Treatment | 1 | 0.40 | .528 | $<0.001$ |
| Error | 923 | $(0.62)$ |  |  |

There was a statistically significant main effect of socioeconomic status in both units, with students who are not socioeconomically disadvantaged performing better than students who are. There was also a statistically significant
effect of treatment in Covering and Surrounding, with treatment out-performing control, but no statisticallydifferent differences between groups for Bits and Pieces III.

## $8^{\text {th }}$ Grade

Traditional descriptive statistics illustrating the overall change in students' performance from pre-test to post-test is shown in Table 5.

Table 5: Mean $8^{\text {th }}$ grade assessment performance, all students

| Test section | Pre-test | Post-test |
| :---: | :---: | :---: |
|  | Shapes of Algebra |  |
| Multiple-choice | $37.2 \%$ | $51.4 \%$ |
| $\%$ correct | $(\mathrm{SD}=15.1 \%)$ | $(\mathrm{SD}=20.6 \%)$ |
| Open-ended | 1.00 | 2.8 |
| out of 8 points | $(\mathrm{SD}=1.6)$ | $(\mathrm{SD}=2.5)$ |
|  | Say it with Symbols |  |
| Multiple-choice | $43.2 \%$ | $55.0 \%$ |
| $\%$ correct | $(\mathrm{SD}=17.3 \%)$ | $(\mathrm{SD}=21.4 \%)$ |
| Open-ended | 1.5 | 2.7 |
| out of 8 points | $(\mathrm{SD}=1.8)$ | $(\mathrm{SD}=2.4)$ |

Again, students made significant gains from pre-test to post-test on both units, although the $8^{\text {th }}$ grade assessments were relatively more difficult than the $6^{\text {th }}$ grade assessments. Post-test scale scores for both $8^{\text {th }}$ grade units, holding pretest scores constant, are shown in Figure 4.


Figure 4: Post-test IRT scale scores for the $8^{\text {th }}$ grade units. Error bars represent $\pm 2$ standard error.

ANCOVA results are presented in Table 6 (mean-square error is shown in parentheses). As in the $6^{\text {th }}$ grade units, pretest was significantly associated with post-test scores in both units and there was also a statistically significant main effect of socioeconomic status in both units, with students who are not socioeconomically disadvantaged performing better than students who are. Statistically significant effects of treatment were found for both units, with treatment outperforming control. Effect sizes for Shapes of Algebra and Say it with Symbols are similar.

Table 6: $8^{\text {th }}$ grade ANCOVA results

| Source | $\boldsymbol{d f}$ | $\boldsymbol{F}$ | $\boldsymbol{p}$ | Partial $\boldsymbol{\eta}^{\mathbf{2}}$ |
| :--- | :---: | :---: | :---: | :---: |
| Shapes of Algebra |  |  |  |  |
| Pre-test | 1 | 157.57 | $<.001$ | 0.180 |
| Socioec. disadv. | 1 | 34.09 | $<.001$ | 0.045 |
| Treatment | 1 | 6.58 | .011 | 0.009 |
| Error | 716 | $(0.55)$ |  |  |
| Say it with Symbols |  |  |  |  |
| Pre-test | 1 | 434.39 | $<.001$ | 0.347 |
| Socioec. disadv. | 1 | 26.99 | $<.001$ | 0.032 |
| Treatment | 1 | 9.72 | .002 | 0.012 |
| Error | 817 | $(0.46)$ |  |  |

## Discussion

Students demonstrated large learning gains for each unit, suggesting both versions of the CMP2 curriculum were effective. Further, three of the four units in this study produced statistically significant effects of the treatment manipulation. That is, the treatment materials produced an additional boost to student learning over and above the existing materials. Why were some treatment units more effective than others? One possible explanation for this differential effect is that Covering and Surrounding and Shapes of Algebra, two of the three units showing a statistically-significant treatment effect, are both more spatially-oriented units. Covering and Surrounding addresses area and perimeter and Shapes of Algebra emphasizes coordinate geometry. While Say it with Symbols focuses on expressions and equations, students must link symbolic representations to graphs and other figures. In contrast, Bits and Pieces III more strongly emphasizes symbolic and tabular representations. The more figureoriented units may allow for a more potent treatment, as the first cognitive principle directly relates to increasing the coherence in visual representations.

The current findings suggest that research-based instructional strategies can be applied synergistically to improve student outcomes in authentic classroom settings. These findings are of particular importance as the vast majority of existing research investigates design principles in highly controlled (and artificial) lab-based studies. Ongoing analyses will provide further insight into the nature of the treatment effects. We are currently analyzing teachers' instructional logs in order to better understand when and how they implemented the cognitive principles in their teaching practice, aside from using the revised student books. We would expect larger learning gains for students when teachers integrated the principles into classroom practice in addition to giving students the revised books. Additional studies are also being carried out at the sites of the partner institutions to investigate the effects of the additive effects of the principles. The Math Center team is also conducting a cluster-randomized trial of revisions to the entire $7^{\text {th }}$ grade CMP curriculum, taking place during the 2012-2013 and 2013-2014 academic years. If the effects of
the principles are cumulative throughout the school year, we would expect greater differences in performance

## Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305C100024 to WestEd. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. We acknowledge the extensive contributions of our research partners of the National Center for Cognition and Mathematics Instruction, particularly the co-PIs of the center: Martha Alibali, Julie Booth, Susan Goldman, Neil Heffernan, Ken Koedinger, Mitchell Nathan, James Pellegrino, and the leads of the math review, production, and data collection teams: Shandy Hauk, Kimberly Vivani, and Kathleen Lepori.

## References

Black, P., \& Wiliam, D. (1998). Assessment and classroom learning. Assessment in Education: Principles, Policy \& Practice, 5(1), 7.
Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., \& Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. Psychological Bulletin, 132(3), 354-380.
Chi, M. T. H. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In R. Glaser (Ed.), Advances in instructional psychology (pp. 161-238). Mahwah, NJ: Lawrence Erlbaum Associates.
Clark, R. C., \& Mayer, R. E. (2003). e-Learning and the Science of Instruction: Proven Guidelines for Consumers and Designers of Multimedia Learning. San Francisco, California: Jossey-Bass.
Cook, L. L. \& Eignor, D. R. (1991). IRT equating methods. Educational Measurement: Issues and Practice, 10(3), 37-45.
Goldstein, H. (1987). Multilevel models in educational and social research. London: Griffin.
Harp, S. F., \& Mayer, R. E. (1998). How seductive details do their damage: A theory of cognitive interest in science learning. Journal of Educational Psychology, 90(3), 414434.

Hulleman, C. S. \& Cordray, D. S. (2009). Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. Journal of Research on Educational Effectiveness, 2, 88-110.
Karpicke, J. D., \& Roediger, H. L., III. (2010). Is expanding retrieval a superior method for learning text materials? Memory \& Cognition, 38, 116-124.
Larkin, J. H., \& Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. Cognitive Science: A Multidisciplinary Journal, 11(1), 65-100.
Moreno, R., \& Mayer, R. E. (1999). Cognitive principles of multimedia learning: The role of modality and contiguity. Journal of Educational Psychology, 91(2), 358-368.

Murray, D. M. (1998). Design and analysis of grouprandomized trials. New York: Oxford University Press.
Paas, F. G. W. C. \& Van Merrienboer, J. J. G. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. Journal of Educational Psychology, 86(1), 122133
Pashler, H., Bain, P., Bottge, B., Graesser, A., Koedinger, K., McDaniel, M., et al. (2007). Organizing instruction and study to improve student learning: IES practice guide (NCER 2007-2004). Washington, DC: National Center for Education Research.
Pavlik, P. I., Jr., Presson, N., Dozzi, G., Wu, S.-M., MacWhinney, B., \& Koedinger, K. (2007). The FaCT (fact and concept) system: A new tool linking cognitive science with educators. In Proceedings of the 29th Annual Conference of the Cognitive Science Society. Nashville, TN, USA.
Rittle-Johnson, B. (2006). Promoting transfer: Effects of self-explanation and direct instruction. Child Development, 77(1), 1-15.
Roediger, H.L., III, \& Karpicke, J.D. (2006). Test enhanced learning: Taking memory tests improves long-term retention. Psychological Science, 17, 249-255.
Raudenbush, S. W., \& Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods (2nd ed.). Thousand Oaks, Calif.: Sage Publications.
Rohrer, D. (2009). Avoidance of overlearning characterises the spacing effect. European Journal of Cognitive Psychology, 21(7), 1001-1012.
Roy, M., \& Chi, M. T. H. (2005). The self-explanation principle in multimedia learning. In R. E. Mayer (Ed.), Cambridge handbook of multimedia learning, (pp. 271286). New York: Cambridge University Press.

Siegler, R. S. (2002). Microgenetic studies of selfexplanations. In N. Granott \& J. Parziale (Eds.), Microdevelopment: Transition processes in development and learning (pp. 31-58). New York: Cambridge University.
Siegler, R.S., \& Chen, Z. (2008). Differentiation and integration: Guiding principles for analyzing cognitive change. Developmental Science, 11, 433-448.
Sweller, J. (1999). Instructional design in technical areas. Camberwell, Australia: ACER.
Sweller, J., \& Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. Cognition and Instruction, 2(1), 59-89.
Underwood, B. J. (1961). Ten years of massed practice on distributed practice. Psychological Review, 68(4), 229247.

Zhu, X., \& Simon, H.A. (1987). Learning mathematics from examples and by doing. Cognition and Instruction, 4, 137-166.

# Achievement Motivation and Strategy Selection during Exploratory Learning 

Daniel A. DeCaro ${ }^{1}$ (ddecaro@indiana.edu)<br>Ostrom Workshop in Political Theory and Policy Analysis, Indiana University<br>Bloomington, IN 47408 USA

Marci S. DeCaro ${ }^{1}$ (marci.decaro@louisville.edu)
Department of Psychological and Brain Sciences, University of Louisville Louisville, KY 40292 USA

Bethany Rittle-Johnson (bethany.rittle-johnson@vanderbilt.edu)<br>Department of Psychology and Human Development, Vanderbilt University Nashville, TN 37203 USA


#### Abstract

Exploratory learning before instruction can benefit understanding, but can also be challenging. Individual differences in response to challenge, such as achievement motivation, may therefore moderate the benefits of exploratory learning. Higher mastery orientation generally leads to increased effort in response to challenge, whereas higher performance orientation leads to withdrawal. Children ( $2^{\text {nd }}-4^{\text {th }}$ grade; $N=159$ ) were given mathematical equivalence problems to solve as either an exploratory learning activity (before instruction) or as practice (after instruction). Higher mastery orientation was associated with improved learning from exploration. In contrast, performance orientation did not lead to learning improvements-and sometimes even hurt learning. Higher mastery orientation was also associated with more sophisticated problem-solving strategies during exploration. Although exploratory activities have the potential to advance strategy selection and subsequent learning, achievement motivation may boost or hinder these benefits.


Keywords: exploratory learning, achievement motivation, mathematics, strategy selection

## Introduction

Exploratory instructional activities can increase individuals’ understanding of new concepts. By wrestling with different solution approaches or conceptual perspectives in a trial-and-error fashion, learners encounter a broader range of both correct and incorrect strategies than might normally be encountered during more traditional "tell-then-practice" methods of instruction (Bonawitz et al., 2011). As a result, learners who explore a new concept before receiving direct instruction on the topic may develop a more sophisticated appreciation for why or how a particular solution approach is better, or worse, than another. This training potentially translates into deeper understanding and better retention of the material (Schwartz, Lindgren, \& Lewis, 2009).

For example, Schwartz et al. (2011) examined the learning of eighth-grade students who explored density problems before receiving instruction. These students exhibited better understanding of the problem structure and better transfer to novel problems at a later test compared to those who received instruction before solving the density
problems. Similar findings have been observed for ninthgrade students learning descriptive statistics (Kapur, 2012; Schwartz \& Martin, 2004) and college students learning cognitive psychology (Schwartz \& Bransford, 1998).

Although exploratory learning can enhance conceptual understanding, such exploration can be challenging for the learner. Compared to more traditional tell-then-practice instruction, learners typically make more mistakes during exploratory learning activities, and they must focus on those mistakes in order to develop more sophisticated conceptualizations of the problem (Kapur, 2010). This learning process often entails considerable effort, as individuals engage in trial-and-error learning or hypothesistesting (Kirschner, Sweller, \& Clark, 2006; Rittle-Johnson, 2006). Learners also may encounter considerably more confusion about how to proceed (Dewey, 1910). In some cases, these learning challenges may pose a "desirable difficulty" (Bjork, 1994) or "productive failure" (Kapur, 2010) that encourages learners to rethink their previous conceptions and develop better understanding, thereby preparing them to learn from further instruction (Schwartz \& Bransford, 1998). In other cases, the difficulty posed by exploratory learning may be too high (Kirschner et al., 2006).

## Achievement Motivation and Challenge

In this study, we ask whether some learners may be better motivated than others to cope with the challenges posed by exploratory learning and thereby capitalize on the instructional experience. Research on achievement motivation demonstrates that individuals approach learning events with different goals and conceptions of what constitutes "ideal" learning performance. These differences influence how individuals interpret and respond to challenge during learning (Dweck \& Leggett, 1988; Elliot \& McGregor, 2001). Individuals can have both mastery and performance goals to different degrees (Barron and Harackiewicz, 2006). Individuals higher in masteryorientation desire personal growth (i.e., learning goals) and tend to view challenge as an opportunity to learn something new. Therefore, they generally seek challenge and respond

[^51]to it with increased effort and interest. Individuals higher in performance-orientation desire to prove their ability (i.e., performance goals). As such, they tend to interpret effort as a sign of incompetence, leading them to interpret difficult learning activities as a potential threat and to withdraw from challenges (Dweck, 1986).

For example, Diener and Dweck (1978) compared how mastery- versus performance-oriented $4^{\text {th }}-6^{\text {th }}$ graders reacted to failure in a difficult category-learning task. Participants first completed several solvable categorization problems matched to their age group (with accuracy feedback). Afterward, they encountered four unsolvable problems, known to be too advanced for their age group. While completing the solvable problems, children with higher performance- versus mastery-orientation exhibited equal degrees of problem-solving accuracy and positive affect. They also had equally sophisticated problem-solving approaches. However, their behavior quickly diverged during the unsolvable trials. Children with higher masteryorientation responded with increased interest and effortattributing the setback to a need for more effort. In addition, they maintained a high degree of strategy sophistication or invented more sophisticated problem-solving strategies to successfully deal with the new challenge. In contrast, children with higher performance-orientation responded with increased negative affect and disinterest-attributing failure to lack of ability. These children defensively withdrew their effort or regressed to developmentally simpler strategies that could not lead to success. Thus, children with higher mastery-orientation learned more from this difficult task and solved more developmentally challenging problems. Similar observations have been made in confusing learning conditions (Licht \& Dweck, 1984).

## Current Study

Individuals may respond to exploratory learning activities like they respond to challenge more generally. That is, based on their typically positive reaction to challenge, learners with higher mastery orientation may be better equipped to deal with the potential confusion and intellectual obstacles posed by exploration. We examined this possibility by comparing the problem-solving strategies, and subsequent learning, of individuals higher or lower in mastery versus performance orientation. Children were instructed on a novel mathematical concept. Half were given instruction, and then solved practice problems with accuracy feedback (instruct-first condition). The other half received the same materials, but in reverse order: They first completed exploratory problem-solving with accuracy feedback, and then received instruction (solve-first condition).

Second- through fourth-grade children were taught the concept of mathematical equivalence-that values on both sides of the equal sign represent the same quantity. This concept is fundamental for future conceptual development in mathematics, such as early algebra understanding (McNeil \& Alibali, 2005). Children in these grades often lack a relational understanding of mathematical
equivalence (e.g., understanding that $2+3$ is "the same as" 5). Children often demonstrate their misconceptions of the equal sign with the strategies they use for mathematical equivalence problems such as $4+5+3=++3$ (e.g., McNeil \& Alibali, 2005). In these problems, children often view the equal sign as a procedural cue. For example, they may ignore the values to the right of the equal sign and sum the numbers on the left side (resulting in the incorrect answer 12; add-to-equals strategy). Or, they may sum every number in the equation, ignoring the sides delineated by the equal sign (resulting in the incorrect answer 15; add-all strategy).

These types of incorrect strategies reflect an operational understanding of the equal sign, and indicate a developmentally immature understanding of mathematical equivalence (Rittle-Johnson et al., 2011). Such responses also resemble learning errors identified in the achievement motivation literature, in which performance-orientation leads learners (e.g., $4^{\text {th }}-6^{\text {th }}$ graders) to perseverate on disconfirmed strategies or revert to less mature (e.g., preschool level) representations of a problem following failure trials (e.g., Diener \& Dweck, 1978).

## Hypotheses

Considering the literatures on exploratory learning and achievement motivation, we predicted different learning outcomes depending on the type of knowledge assessed. We assessed learner's knowledge of mathematical equivalence both immediately after they completed an individual tutoring session, and approximately two weeks later. We also examined problem-solving strategies during the tutoring session itself. These questions were examined by reanalyzing previously-reported-data (DeCaro and RittleJohnson, 2012) to examine the role of achievement motivation.

Conceptual Knowledge Our main interest in the present research was how achievement motivation affects learners’ conceptual knowledge, their ability to grasp the underlying principles of mathematical equivalence, following exploration. Prior work suggests that exploration primarily benefits conceptual knowledge (Schwartz et al., 2009), but is mistake-prone and initially more confusing than a tell-then-practice instructional approach (e.g., Alfieri et al., 2011). Previous research also indicates that individual differences in achievement motivation influence learning and performance primarily when learners encounter challenging tasks (Dweck, 1986). Mastery orientation typically leads learners to respond to initial setbacks with increased resolve, and by maintaining or inventing more sophisticated learning strategies (e.g., Diener \& Dweck, 1978). Thus, we expected higher mastery orientation to be associated with improved conceptual knowledge, specifically in the more demanding solve-first condition.

The prediction for performance orientation in the solvefirst condition is less straightforward. Higher performance orientation often leads learners to respond to setbacks with defensive withdrawal of effort and regressive thinking (e.g.,

Diener \& Dweck, 1978). Therefore, performance orientation may be detrimental to conceptual knowledge in the solvefirst condition. Alternatively, performance orientation may not actually hurt conceptual knowledge, compared to that obtained in the instruct-first condition; instead, it may simply hinder one’s ability to profit from the exploratory learning opportunity. This prediction is supported by Barron and Harackiewicz's (2005) multiple-motive hypothesis, which suggests that mastery and performance motives represent separate signals with different degrees of relevance for conceptual versus procedural knowledge. According to this hypothesis, the mastery motive is more relevant to conceptual knowledge than the performance motive, because understanding and deeper processing of information are more clearly central to personal development and less diagnostic of ability.

Procedural Knowledge We also evaluated procedural knowledge, or the ability to execute the correct action sequences to solve problems. Procedural knowledge is strongly correlated with conceptual knowledge (RittleJohnson \& Alibali, 1999). However, problem-solving assessments provide especially diagnostic information about ability. Therefore, according to Barron and Harackiewicz's (2005) multiple-motive hypothesis, performance orientation may be more relevant to procedural knowledge than mastery orientation (cf. Grant \& Dweck, 2003). We therefore predicted a positive, but weaker, relationship between mastery orientation and procedural knowledge in the solvefirst condition. Moreover, we predicted a negative relationship between performance orientation and procedural knowledge acquisition (cf. Dweck \& Leggett, 1988; Grant \& Dweck, 2003).

Problem-Solving Strategies In addition to assessing knowledge outcomes (after tutoring), we examined children's problem-solving strategies during tutoring. Such information may reveal how achievement motivation impacts learning from exploration. Because children in the solve-first condition completed the problems as an exploratory activity, we expected them to use poorer problem-solving strategies. Specifically, they might use fewer relational strategies that evidence understanding of the equal sign as a relational symbol. Instead, they might rely more on operational strategies, in keeping with developmentally simpler views of the equal sign as an operational symbol (i.e., "add-all" or "add-to-equals").
Although we thought the solve-first condition would be more challenging, we expected mastery orientation to promote a more adaptive response to these setbacks (cf. Diener \& Dweck, 1978). Specifically, mastery orientation should be associated with increased use of relational strategies and decreased use of operational strategies. In contrast, performance orientation should be associated with increased reliance on these developmentally simpler, operational strategies (and decreased use of relational strategies).

## Method

## Participants

Participants were $2^{\text {nd }}-4^{\text {th }}$ grade children at a suburban public school. Children who scored below $80 \%$ on a pretest assessing procedural and conceptual knowledge of mathematical equivalence were selected ( $N=159,56 \%$ female, age $M=8.5$ years, range 7.3-10.8 years). Approximately $18 \%$ were ethnic minorities (10\% AfricanAmerican, 6\% Asian, and 2\% Hispanic).

## Research Design and Procedure

Consenting children first completed a pretest in their classrooms, followed by a self-report measure of their achievement motivation. Within one week following the pretest, children selected for the study participated in individual tutoring sessions on mathematical equivalence. Children were randomly assigned to the instruct-first condition ( $n=79$ ) or the solve-first condition ( $n=80$ ). Children were additionally assigned to either self-explain (i.e., explain why particular answers were correct/incorrect) or solve extra problems instead; however, this manipulation had no discernible effects and will not be discussed further. The session ended with a posttest assessing children's procedural and conceptual knowledge. Approximately two weeks later, children completed an equivalent retention test.

## Tutoring Session

Conditions The instruct-first and solve-first conditions were identical, except that the presentation order for the instruction ("instruct") and problem-solving ("solve") portions of the lesson were reversed. Thus, in the instructfirst condition, the problems served as practice after a lesson on mathematical equivalence. In the solve-first condition, these problems served as an exploratory learning activity followed by formal instruction.

Instruction During instruction (adapted from Matthews \& Rittle-Johnson, 2009), children were taught about the relational meaning of the equal sign. Five number sentences (e.g., $3+4=3+4$ ) were individually shown on the computer. The experimenter explained the structure of each number sentence (i.e., that there are two sides) and the explicit meaning of the equal sign (i.e., that the equal sign means that both sides are "equal or the same").

Problem-Solving During the problem-solving phase, children completed six mathematical equivalence problems presented individually on the computer. Problems increased in difficulty from three operands (i.e., $10=3+$ ) to five operands (e.g., $5+3+9=5+$ ). Children could use pencil and paper to solve each problem. After entering their answer on the computer, children were asked to report their problemsolving strategy. Then they were shown the correct answer.

## Learning Assessments

Problem-Solving Strategies, Children's problem-solving strategies in the tutoring session were categorized as relational, operational, or other incorrect (kappa=.80). Relational strategies evidenced a deliberate attempt to equalize the values on each side of the equation or conceptualize the values as equivalent (Rittle-Johnson et al., 2011). Operational and other incorrect strategies both evidenced an erroneous conceptualization of the equal sign. However, operational strategies represented misconceptions previously identified as developmentally less sophisticated and fundamentally inadequate (i.e., add-all and add-toequals strategies; McNeil \& Alibali, 2005).

Posttest and Retention Test We measured children's conceptual and procedural knowledge of mathematical equivalence by adapting assessments from past research (Rittle-Johnson et al., 2011). Conceptual knowledge items assessed two key concepts: the symbolic meaning of the equal sign and the structure of equations (8 items; kappas=.89-.96). Procedural knowledge items consisted of ten mathematical equivalence problems. Answers to procedural knowledge items were scored as correct if they came within one point of the correct answer, to reduce false negatives. The retention test was identical to the posttest, but also included eight far-transfer items that will not be discussed further, due to space limitations. Because we were most interested in long-term learning, and because the results of the posttest mirrored those of the retention test, we report only the results of the retention test.

## Achievement Motivation

Achievement motivation items were adapted from Elliot and Church (1997). Two items assessed mastery orientation (e.g., "I want to learn as much as possible about math, even if I have to work hard"). Two items assessed performance orientation (e.g., "In math class, it is important for me to do well compared to others in my class"). Children responded on a 6-point, Likert-type scale ranging from 1 (Strongly Disagree) to 6 (Strongly Agree). Mastery-orientation and performance-orientation scores were created by averaging the two responses on each subscale (Elliot \& Church, 1997).

## Results

We examined the relationship between mastery and performance orientation and learning in the two tutoring conditions. We also examined children's problem-solving strategies during tutoring. We used hierarchical linear regression for all analyses. The predictors in the model were mastery orientation score, performance orientation score, condition (dummy-coded), and two interaction terms (Condition $\times$ Mastery Orientation, Condition $\times$ Performance Orientation). Preliminary analyses showed no significant two-way interactions between mastery and performance orientation, or three-way interaction with condition, so they were not included in the final model. Thus, the final model
represents the independent and joint effects of achievement motivation and tutoring condition on the dependent variables (Barron \& Harackiewicz, 2001). We also included children's age and conceptual and procedural knowledge pretest scores to control for prior knowledge. Each predictor was centered. Significant interactions were explored through simple slopes analyses. Estimated means were plotted at one SD above and below the mean, to represent the effect of low versus high achievement motivation on the dependent variable as a function of condition.

No significant main effects of performance or mastery orientation emerged ( $F s<1$ ). Therefore, only the results for Condition and Condition $\times$ Achievement Motivation interactions will be reported. Children in the instruct-first and solve-first conditions did not differ at pretest by their procedural knowledge, conceptual knowledge, or achievement motivation ( $F \mathbf{s}<1$ ). Mastery and performance orientation were not correlated: $r(156)=.08, p=.151$.

## Conceptual Knowledge

At retention test, a marginally significant main effect of condition emerged ( $B=.05$, $\mathrm{SE}=.03, p=.078$ ). Learners in the solve-first condition demonstrated higher conceptual knowledge than learners in the instruct-first condition. This effect of condition was qualified by a Mastery Orientation $\times$ Condition interaction ( $B=.08, \mathrm{SE}=.04, p=.059$ ). As depicted in Figure 1, higher mastery orientation was associated with higher conceptual knowledge acquisition in the solve-first condition ( $B=.08, \mathrm{SE}=.03, p=.009$ ), indicating that higher mastery orientation helped children learn from exploration. Mastery orientation was unrelated to conceptual knowledge in the instruct-first condition $(B=0)$. There was no Performance Orientation $\times$ Condition interaction ( $B=0$ ), indicating that performance orientation did not hurt conceptual knowledge.


Figure 1. Conceptual and Procedural Knowledge

## Procedural Knowledge

At retention test, the condition term was not significant ( $B$ $=.02$; Figure 1). A Mastery Orientation $\times$ Condition interaction emerged ( $B=.12, S E=.06, p=.036$ ). Higher mastery orientation was associated with a trend towards higher procedural knowledge in the solve-first condition ( $B$ $=.07, S E=.04, p=.118$ ), whereas it was associated with a trend towards poorer procedural knowledge in the instructfirst condition ( $B=-.05, S E=.04, p=.159$ ).

A significant Performance Orientation $\times$ Condition interaction also emerged ( $B=-.11, S E=.05, p=.041$; Figure 1). Higher performance orientation was associated with lower procedural knowledge in the solve-first condition ( $B=-.09, S E=.04, p=.035$ ) but was unrelated to procedural knowledge in the instruct-first condition ( $B=.02$ ). Higher performance orientation reduced gains in procedural knowledge from exploration.

## Performance during Tutoring Intervention

To provide further insight into how the knowledge acquisition observed at retention test may have emerged, we examined the problem-solving strategies children used during the tutoring session. Doing so indicates how children responded to difficulties encountered during exploration.

There was a main effect of condition on use of both relational strategies ( $B=-.12, \quad p=.01$ ) and operational strategies ( $B=.09, p=.01$ ). On average, children in the solvefirst condition used relational strategies less than children in the instruct-first condition-reflecting the overall difficulty of exploratory learning in the solve-first condition. This effect was qualified by interactions with both mastery orientation ( $B=.15, \mathrm{p}<.01$ ) and performance orientation ( $B=-$ .16, $p<.01$ ). As shown in Figure 2, higher mastery orientation was associated with increased use of relational strategies in the solve-first condition ( $B=.11, p<.05$ ). In contrast, higher performance orientation was associated with decreased use of relational strategies in this condition ( $B=-$ .11, $p<.05$ ). Neither mastery nor performance orientation were associated with relational strategy use in the instructfirst condition ( $B=-.03$ and $B=.04$ ). In fact, children in the solve-first condition with higher mastery orientation appear to have matched their instruct-first counterparts in use of relational strategies.

Operational strategy use was consistent with these findings. As shown in Figure 2, children in the solve-first condition used operational strategies more than children in the instruct-first condition. No interaction with mastery orientation was found ( $B=-.04$ ). However, performance orientation interacted with condition ( $B=.10, p<.05$ ). In the solve-first condition, higher performance orientation was associated with increased use of operational strategies ( $B=.09, p<.01$ ). This finding suggests that the difficulty of exploratory learning leads children higher in performanceorientation to adopt developmentally immature strategies. No relationship with performance orientation was found in the instruct-first condition ( $B=-.02$ ).

## Relational Strategies

Operational Strategies


Figure 2. Strategy Use during Tutoring Session

## Discussion

As predicted, children higher in mastery orientation learned a new mathematics concept better when a problemsolving session was used as an exploratory activity, rather than practice. That is, higher mastery orientation was associated with improved conceptual knowledge acquisition (and somewhat improved procedural knowledge) in a solvefirst condition where problem-solving preceded formal instruction. Higher performance orientation, in contrast, did not facilitate learning from exploration: These children performed at normal levels on conceptual knowledge acquisition and did worse than normal on procedural knowledge acquisition (i.e., problem-solving success).

These differences in learning from exploration could be attributed to the challenge inherent in such activities. Children in the solve-first condition were less likely to use relational problem-solving strategies, which indicate a sophisticated understanding of mathematical equivalence. They relied more on operational strategies, erroneously treating the equal sign as a procedural cue (e.g., to add only the numbers to the left of the equal sign).

However, this overall effect of condition was moderated by achievement motivation. Children higher in mastery orientation tended to use relational strategies during exploration, not operational strategies. Moreover, children higher in performance orientation-who desire to prove their ability and, therefore, avoid challenge-tended to revert to developmentally simpler operational strategies. This finding is consistent with findings in the achievement motivation literature (cf. Diener \& Dweck, 1978), and may help explain why exploration was only useful to some children. The challenge and confusion associated with exploratory learning may lead some children to explore
better strategies during learning, but lead others to perseverate on poorer strategies, which impede learning.

Recent discussions on exploratory learning versus direct instruction have concluded that there may be benefits of combining aspects of both approaches (cf. Alfieri et al., 2011). The current findings demonstrate that using exploratory problem-solving activities prior to instruction can be beneficial-but namely for children who have a mastery-oriented approach to learning mathematics.

Hence, the current findings highlight the importance of considering motivational influences on learning and strategy selection. Teachers may want to emphasize mastery and promote a forgiving learning environment to help nonmastery oriented students cope better with the inherent challenge posed by exploration. Future research is needed to see if the deleterious effects of performance orientation on strategy selection can be mitigated with mastery framing.

## Acknowledgments

Research supported by NSF grant DRL-0746565 to B. Rittle-Johnson and IES, U.S. Department of Education, training grant R305B080008 to Vanderbilt University. Portions of these findings were presented at the 2011 Society for Research on Educational Effectiveness meeting.

## References

Alfieri, L., Brooks, P. J., Aldrich, N. J., \& Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? Journal of Educational Psychology, 103, 1-18.
Barron, K. E., \& Harackiewicz, J. M. (2001). Achievement goals and optimal motivation: Testing multiple goal models. Journal of Personality and Social Psychology, 80, 706-722.
Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe \& A. Shimamura (Eds.). Metacognition: Knowing about knowing. Cambrige, MA: MIT Press.
Bonawitz, E., Shafto, P., Gewon, H., Goodman, N. D., Spelke, E., \& Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. Cognition, 120, 322-330.
DeCaro, M. S., \& Rittle-Johnson, B. (2012). Exploring mathematics problems prepares children to learn from instruction. Journal of Experimental Child Psychology, 113, 552-568.
Dewey, J. (1910). How we think. Boston: Heath.
Diener, C. I., \& Dweck, C. S. (1978). An analysis of learned helplessness: Continuous changes in performance, strategy, and achievement cognitions following failure. Journal of Personality and Social Psychology, 36, 451462.

Dweck, C. S., (1986). Motivational processes affecting learning. American Psychologist, 41, 1040-1048.
Dweck, C. S., \& Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. Psychological Review, 95, 256-273.

Elliot, A. J., \& Church, M. A. (1996). A hierarchical model of approach and avoidance achievement motivation. Journal of Personality and Social Psychology, 72, 461475.

Elliot, A. J., \& McGregor, H. A. (2001). A $2 \times 2$ achievement goal framework. Journal of Personality and Social Psychology, 80, 501-519.
Grant, H., \& Dweck, C. S. (2003). Clarifying achievement goals and their impact. Journal of Personality and Social Psychology, 85, 541-553.
Kapur, M. (2012). Productive failure in learning the concept of variance. Instructional Science, 40, 651-672.
Kirschner, P. A., Sweller, J., \& Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. Educational Psychologist, 41, 75-86.
Licht, B. G., \& Dweck, C. S. (1984). Determinants of academic achievement: The interaction of children's achievement orientations with skill area. Developmental Psychology, 20, 628-636.
Matthews, P., \& Rittle-Johnson, B. (2009). In pursuit of knowledge: Comparing self-explanations, concepts, and procedures as pedagogical tools. Journal of Experimental Child Psychology, 104, 1-21.
McNeil, N. M., \& Alibali, M. W. (2005). Why won't you change your mind? Knowledge of operational patterns hinders learning and performance on equations. Child Development, 76, 883-899.
Rittle-Johnson, B. (2006). Promoting transfer: Effects of self-explanation and direct instruction. Child Development, 77, 1-15.
Rittle-Johnson, B., \& Alibali, M. W. (1999). Conceptual and procedural knowledge of mathematics: Does one lead to the other? Journal of Educational Psychology, 91, 175189.

Rittle-Johnson, B., Taylor, R., Matthews, P. G., \& McEldoon, K. (2011). Assessing knowledge of mathematical equivalence: A construct modeling approach. Journal of Educational Psychology, 103, 85104.

Schwartz, D. L., \& Bransford, (1998). A time for telling. Cognition and Instruction, 16, 475-522.
Schwartz, D. L., Lindgren, R., \& Lewis, S. (2009). Constructivist in an age of non-constructivist assessments. In S. Tobias \& T. M. Duffy (Eds.), Constructivist Instruction: Success or Failure. New York, NY: Routledge.
Schwartz, D. L., \& Martin, T. (2004). Inventing to prepare for future learning: The hidden efficacy of encouraging original student production in statistics instruction. Cognition and Instruction, 22, 129-184.
Schwartz, D. L., Chase, C. C., Oppezzo, M. A., \& Chin, D. B. (2011). Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer. Journal of Educational Psychology, 103, 759-775.

# Cost-Based Pragmatic Inference about Referential Expressions 

Judith Degen (jdegen@bcs.rochester.edu)<br>Dept. of Brain and Cognitive Sciences<br>University of Rochester

Michael Franke (m.franke@uva.nl)<br>Institute for Logic, Language and Computation<br>Universiteit van Amsterdam

Gerhard Jäger (gerhard.jaeger@uni-tuebingen.de)<br>Institute of Linguistics<br>University of Tübingen


#### Abstract

We present data from three experiments addressing how much theory of mind reasoning is involved in production and interpretation of ambiguous referential expressions in an artificial language task, and how this interacts with the cost and availability of alternative utterances. When an unambiguous alternative is not available, listeners tend to draw simple Quantity inferences reminiscent of scalar implicatures (Grice, 1975). When an unambiguous alternative is available, fewer inferences are observed, but gradiently more as the cost of unambiguous alternatives increase. We outline a novel game theoretic model of pragmatic reasoning based on probabilistic back-and-forth reasoning about interlocutors' rational choices and beliefs. The model provides a good fit to the data and raises interesting issues for future research.


Keywords: Pragmatics; Game theory; Referential Expressions; Language production; Language comprehension.

## Introduction

People are lazy: when they speak, they like to save effort. But if speakers are too lazy and say too little, their listeners will not understand them. A good example is the choice and interpretation of referential expressions. A rational speaker who wants to establish reference should choose the most economic (shortest, easiest, least effortful) description that, according to his beliefs about the listener's dispositions to interpret utterances, will allow for the listener to safely infer the correct referent. A rational listener should take the speaker's production costs ${ }^{1}$ into account and so a rational speaker should in turn take that into account, etc. But this is an idealized picture. From an empirical point of view two related questions arise: 1) How much do speakers and listeners take into account each other's perspective? 2) How much influence do economy considerations have; do listeners weigh the speaker's production costs?

When it comes to referential language use, the latter question has not been investigated thoroughly, but the former question has been addressed in a variety of ways, both theoretically (Clark \& Marshall, 1981) as well as experimentally (Hanna, Tanenhaus, \& Trueswell, 2003; Keysar, Barr, \& Brauner, 2000). This paper aims to adress both questions and to bring theoretical and empirical approaches closer together. The paper's empirical contribution is to report on

[^52]experimental data from referential language games. Its theoretical contribution is a novel probabilistic model of back-and-forth reasoning that synthesizes recent Bayesian models (Frank \& Goodman, 2012) and game theoretic approaches (Camerer, Ho, \& Chong, 2004; Rogers, Palfrey, \& Camerer, 2009; Franke, 2011; Jäger, 2013).

Our experiments probed interlocutors' perspective-taking ability in a task where an artificial language left some relevant meaning features inexpressible or made some expressions costly. When critical meaning features are inexpressible, the situation is reminiscent of scalar implicature calculation (Grice, 1975). We manipulated how many steps of such reasoning were needed for communicative success and tested both comprehension (Exp. 1) and production (Exp. 2) to investigate question 1. Our design was chosen so as to improve on previous related studies where non-linguistic pictorial messages were used (Degen \& Franke, 2012) and where the availability of alternative expressions was not explicitly controlled (Stiller, Goodman, \& Frank, 2011; Frank \& Goodman, 2012). In addition, rather than making some messages entirely unavailable, we investigated the effect of variable production costs on interpretation behavior to address question 2 (Exp. 3). Rohde, Seyfarth, Clark, Jäger, and Kaufmann (2012) showed that listeners take into account message costs when messages are assigned an explicit dollar value. Here we investigate whether these results replicate when costs, as in real language use, are implicit.

The data from our experiments is explained well by a probabilistic model of back-and-forth reasoning. The model parameterizes how deeply interlocutors reason about each other's perspective and how close they are to being rational. Parameter values that best explain our data suggest that participants were reasonably rational and applied a small but non-negligible amount of theory of mind reasoning and that they took production costs into account in comprehension.

## Referential Language Games

Referential communication can be conceived of as a signaling game (Lewis, 1969): a sender (speaker) $S$ knows which referent he wants to talk about, but a receiver (listener) $R$ does not; $S$ chooses a referential description; if $R$ can identify the intended referent, communication is successful, otherwise a failure. Different games ensue for different sets of potential referents and referential expressions. In the critical trials of our experiments the referential games were isomorphic to the


Figure 1: Target implicature conditions. Hearers choose one of the Possible Referents $T=\left\{t_{\mathrm{t}}, t_{\mathrm{c}}, t_{\mathrm{d}}\right\}$. Speakers have Message Options $M=\left\{m_{\mathrm{t}}, m_{\mathrm{c}}, m_{\mathrm{d} 1}, m_{\mathrm{d} 2}\right\}$, shown here for ease of interpretation visually (the experiment used artificial words). Trigger items are indicated with asterisks: e.g., $t_{\mathrm{t}}^{*}$ is the referent to be communicated on complex production trials.
situations in Fig. 1. There were three possible referents in the form of monsters and robots wearing hats or scarves (not depicted in the example) as accessories. Additionally, there is a fixed set of possible descriptions that are available to the sender. Messages for monsters and hats were always available and were equally costly. Messages for robots and scarves were either not available at all (Exp. 1 and 2) or more costly (Exp. 3).

Experiments 1 and 3 tested participants' choice of referent for a given trigger message (comprehension). Experiment 2 tested their choice of message for a trigger referent (production). Trigger items for the critical conditions are marked with an asterisk in Fig. 1. Indices $t, c, d$ stand for target, competitor and distractor respectively. We refer to a game as in Fig. 1a as the simple condition, because it involves one step of Quantity reasoning similar to scalar implicature calculation (Grice, 1975). Hearing trigger message $m_{\mathrm{c}}^{*}, R$ should reason that $S$ must have meant target state $t_{\mathrm{t}}$, and not competitor state $t_{\mathrm{c}}$, because if $S$ had wanted to refer to the latter he could have used an unambiguous message. Conversely, when $S$ wants to refer to trigger state $t_{\mathrm{c}}^{*}$, he should not use the true but semantically ambiguous message $m_{\mathrm{c}}$, because he has an unambiguous message $m_{\mathrm{t}}$. Similarly, we refer to a game in Fig. 1b as the complex condition, because it requires performing similar reasoning twice in sequence (see also Degen and Franke (2012) for details).

## Experiment 1 - comprehension

Exp. 1 tested participants' behavior in a comprehension task using instantiations of the signaling games just described. ${ }^{2}$

## Methods

Participants Using Amazon's Mechanical Turk, 48 selfreported native English speakers were paid $\$ 1.00$ to participate. All were naïve as to the purpose of the experiment.

Procedure and Materials Participants engaged in an artificial language referential comprehension task. The experi-

[^53]ment proceeded in two stages: a language learning stage and an inference stage. Only the inference stage was of theoretical interest. In the language learning stage, participants learned four 3-character words ( $R A V, Z U B, X E K, K O R$ ) of the alien language Zorx. The words referred to visual features: ontological kinds (one of two monster species) and accessories (red or blue hat). Each unique word-to-feature mapping (24 total) occurred twice between participants to ensure effects were not artifacts of the particular mapping.

Language learning occurred in three steps. First, participants saw each word twice with a visual representation of its meaning. They were then presented with each word alongside two choices for the meaning of the word and had to click on the correct meaning. Finally, they were presented with each meaning in succession and had to produce the correct word by clicking on characters in a two-row character array. They repeated this process until achieving $100 \%$ accuracy on the production task. They then proceeded to the inference stage.

On each trial in the inference stage, participants saw three possible referents on a display (as in Fig. 1). Each referent differed systematically along two dimensions: its ontological kind (robot or one of two monster species) and accessory (scarf or either blue or red hat). In addition to these three referents, participants saw a Zorx word that they were told was sent to them by a previous participant whose task it was to get them to pick out one of these three referents. They were told that the previous participant was allowed to send a message expressing only one feature of a given referent, and that the other participant had learned the same words of Zorx they did (i.e., they could send monster/hat messages, but not robot/scarf messages).

Participants initially completed four production trials. They saw three referents, one of which was highlighted with a yellow rectangle, and were asked to send one of the Zorx words to another Mechanical Turk worker to get them to pick out the highlighted object. They were told that the other worker did not know which object was highlighted but knew the same language they did. The four production trials contained three unambiguous and one ambiguous trial which functioned as fillers in the main experiment.


Figure 2: Proportions of target, competitor, and distractor choices in critical and filler conditions for Exp. 1 (comprehension, left) and Exp. 2 (production, right).

Participants saw 36 experimental trials, with a 2:1 ratio of fillers to critical trials. Of the 12 critical trials, 6 occurred in the simple (one iterated reasoning step) and 6 in the complex (two steps) condition as described above (see Fig. 1).

Target position was counterbalanced (each critical trial occurred equally often in each of the 6 possible orders of target, competitor, and distractor), as were the target's features and the number of times each message was sent. Of the 24 filler trials, half used the displays from the critical conditions but the target was either $t_{\mathrm{c}}$ or $t_{\mathrm{d}}$ (as identified unambiguously by the trigger message). This was intended to prevent learning associations of display type with the target. On the other 12 filler trials, the target was either entirely unambiguous or entirely ambiguous given the message. That is, there was either only one object with the feature denoted by the trigger message, or there were two identical objects that were equally viable target candidates. Trial order was pseudo-randomized such that there were two lists (reverse order) of three blocks, where critical trials and fillers were distributed evenly over blocks. Each list began with three filler trials.

## Results and Discussion

Proportions of choices are displayed in Fig. 2 (left panel). As expected, participants were close to ceiling in choosing the target on unambiguous filler trials but at chance on ambiguous ones. This confirms that participants understood the task. On critical trials, participants' performance was intermediate between ambiguous and unambiguous filler trials. On simple trials, participants chose the target $66 \%$ of the time. On complex trials, the target was chosen less often (50\%).

To test whether the observed differences in target choices above were significantly different, we fitted a logistic mixedeffects regression to the data. Trials on which the distractor was selected were excluded to allow for a binary outcome variable (target vs. competitor choice). The model predicted the log odds of choosing a target over a competitor from a Helmert-coded Condition predictor. Three Helmert con-
trasts over the four relevant critical and filler conditions were included in the model, comparing each condition with a relatively less skewed distribution against the more skewed distributions (in order: ambiguous fillers, complex, simple, unambiguous fillers). This allowed us to capture whether the differences in choice distributions for neighboring conditions suggested by Fig. 2 were significant. We included the maximal random effects structure, i.e., by-participant random intercepts, by-participant random slopes for Condition, and by-item random intercepts.

Of the three contrasts, two reached significance; there were more target choices in the unambiguous filler condition than in the simple condition ( $\beta=4.08, S E=0.41, p<.0001$ ) and there were more target choices in the simple than in the complex condition ( $\beta=1.27, S E=0.47, p<.01$ ). However, there was no significant difference in target choices between the ambiguous filler and the complex condition $(\beta=0.38, S E=$ $0.45, p<.4$ ). This suggests that participants made simple, but not complex inferences.

## Experiment 2 - production

Exp. 2 tested participants' behavior in a production task using instantiations of the signaling games described above.

## Methods

Participants Using Mechanical Turk, 48 self-reported native speakers of English were paid $\$ 1.20$ to participate.

Procedure and Materials The experiment again proceeded in two stages: the language learning stage and the production stage. The procedure for language learning was the same as in Exp. 1. The procedure for the production stage was the same as on the production trials in Exp. 1. Participants saw 36 trials with a $2: 1$ ratio of fillers to critical trials. There were 12 critical trials ( 6 simple and 6 complex situations as in Fig. 1). Half of the fillers used the same displays as the critical trials, but one of the other two objects was highlighted. This meant that the target message was either unambiguous (e.g. when the highlighted object was $t_{\mathrm{t}}$ in Fig. 1(a) the target message was $m_{\mathrm{c}}$ ) or entirely ambiguous. The remaining 12 filler trials employed other displays with either entirely unambiguous or ambiguous target messages. Two experimental lists were created and counterbalancing ensured as in Exp. 1.

## Results and Discussion

Proportions of choices are displayed in Fig. 2 (right panel). To test whether the observed differences in target choices were different, the same logistic mixed-effects regression was fit to the data as in the Exp. 1 analysis. Trials on which a distractor message was sent were excluded to allow for a binary outcome variable (target vs. competitor choice).

Of the three Helmert contrasts, again only two reached significance; there were more target choices in the unambiguous filler condition than in the simple condition ( $\beta=$ $3.84, S E=0.47, p<.0001$ ) and there were more target choices in the simple than in the complex condition ( $\beta=$
$2.81, S E=0.50, p<.0001$ ). However, there was no difference between the ambiguous filler and the complex condition ( $\beta=-0.43, S E=0.37, p<.3$ ). This suggests that, as in comprehension, participants made simple, but not complex inferences.

## Experiment 3 -comprehension with costs

Exp. 3 tested whether listeners take into account speakers' preferences for producing minimally effortful messages. To this end, we introduced messages for the robot and scarf feature but varied the implicit cost of these messages via word length measured in characters. If listeners take into account their interlocutor's perspective, their behavior should approximate the results from the simple conditions of Exp. 1 (i.e., draw more Quantity inferences) as the message becomes more costly (and thus, more dispreferred/unavailable).

## Methods

Participants A total of 240 participants were recruited over Mechanical Turk who were all self-reported native speakers of English. They were paid $\$ 0.80$ plus a $\$ 0.10$ bonus if they completed the cost estimation stage in under one minute.

Procedure and Materials The experiment proceeded in three stages: the language learning stage, the cost estimation stage, and the inference stage. The procedure in the language learning and inference stage was the same as in Experiment 1 with the following three exceptions: a) the learned language contained two extra costly words (to refer to the robot and the scarf feature) in addition to four free words (to refer to monsters and hats), b) there were no complex, only simple conditions (Fig. 1) in the inference stage, c) there were only 12 rather than 24 filler trials, of which 6 were completely ambiguous and 6 were completely unambiguous.

The cost estimation stage was introduced to estimate participants' subjective cost function. Each of the nine permutations of feature combinations \{robot, green monster, purple monster $\} \times\{$ scarf, red hat, blue hat $\}$ was presented to participants one at a time and they were asked to send one of the Zorx words they had learned to another participant to get them to pick out the presented referent. As in the previous experiments, sending a message required spelling out the word on a virtual keyboard on the screen by clicking on each character individually. In addition, participants were told that they would receive a bonus if they completed this part of the study in under one minute. We hoped these two features of the task would increase participants' subjective costs associated with the objective increase in number of characters and thus encourage a message cost effect.

There were three cost conditions. In the NO-COST condition, the costly messages were of the same length as the free messages ( 3 characters). In the LOW-COST and HIGH-COST conditions, the costly messages were one and three characters longer than the free messages, respectively. LOW-COST and HIGH-COST were manipulated within participants (we refer to this group as the COST condition, 192 participants) and the

NO-COST condition consisted of a separate group of 48 participants. Thus the languages in the different conditions:


## Results and Discussion

Proportion of choices in the cost estimation stage (messages) and in the inference stage (referents) are shown in Fig. 3a and $3 b$. We analyzed participants' performance in both stages.

In the cost estimation stage, we analyzed participants' message choices for the four referents with one costly and one free message (i.e., referents with either a robot or a scarf feature). The NO-COST condition served as the baseline in the mixed effects logistic regression predicting the log odds of a costly over a free message choice. Cost condition was dummy-coded and entered as a three-level categorical predictor (NO-COST, LOW-COST, HIGH-COST). The model additionally included random by-participant and by-item intercepts as well as by-participant slopes for feature type (scarf or robot) to account for individual variability in participants' preferences for referring to these features. There was a significant decrease in the log odds of choosing the costly message compared to the NO-COST reference level when the message was HIGH-COST $(\beta=-0.83, S E=0.30, p<.01)$. For the LOW-COST message, the difference trended in the predicted direction ( $\beta=-0.44, S E=0.30, p<.14$ ). Thus, increasing message cost led to a small, but gradient decrease in preference to send the costly message.

Next, we analyzed participants' performance in the inference stage by fitting a mixed effects logistic regression model predicting target over competitor choices. Two Helmert contrasts over the three relevant cost conditions were included in the model, comparing each condition with a relatively lower cost against the higher cost level(s) (in order: no cost, small cost, high cost). The model additionally included byparticipant and by-item random intercepts. The difference between the NO-COST and other conditions did not reach significance, though it trended in the predicted direction ( $\beta=$ $-0.08, S E=0.05, p<.14$ ). However, there were significantly more target choices in the HIGH-COST than in the LOW-COST condition ( $\beta=0.25, S E=0.12, p<.05$ ). Thus, the gradient effect of message cost on message choice is in turn reflected in listener inferences: as the cost of the unambiguous message increases, listeners make more scalar inferences and begin to approximate the behavior displayed in Exp. 1, where robot/scarf messages were entirely unavailable.

## The Iterated Quantal Response Model

The observed production and comprehension behavior can be predicted by a parameterized model that returns a quantitative description of speaker and listener behavior. The iterated quantal response (IQR) model combines key features of socalled cognitive hierarchy models from behavioral economics


Figure 3: Experiment 3 results.
(Camerer et al., 2004; Rogers et al., 2009) with game theoretic models of pragmatic reasoning (Franke, 2011; Jäger, 2013). The resulting model is also very similar to, but slightly more general than recently popular Bayesian models (Frank \& Goodman, 2012; Bergen, Levy, \& Goodman, 2012).

We consider two parameters. Parameter $\tau$ models the depth of strategic reasoning that language users engage in. Parameter $\lambda$ models how successful language users are at making rational choices. The output of the model is a prediction of probabilistic speaker and listener behavior.
Signaling Games We model our referential tasks as signaling games. For our purposes, a signaling game is just a structure $\langle T, M, B, c\rangle$ with $T=\left\{t_{1}, \ldots, t_{\mathrm{a}}\right\}$ a set of $a$ different states (referents), $M=\left\{m_{1}, \ldots, m_{\mathrm{b}}\right\}$ a set of $b$ relevant descriptions, $B$ is a Boolean $(a, b)$-matrix with $B_{i j}=1$ if description $m_{\mathrm{j}}$ is true of referent $t_{\mathrm{i}}$, and $c=\left(c_{1}, \ldots, c_{b}\right)$ a vector of costs. ${ }^{3}$

Strategies A sender strategy $\sigma$ is a row-stochastic $(a, b)$ matrix, mapping each state onto a probability distribution over messages. A sender strategy describes how likely an average speaker would choose a message given that they want to talk about a given state. Likewise, since the receiver chooses states in $T$ as interpretations of an observed message, a receiver strategy $\rho$ is a row-stochastic $(b, a)$-matrix.

Expected Utilities A sender who believes that his listener plays $\rho$ has an $(a, b)$-matrix of expected utilities $\mathrm{EU}(\rho)=$ $\mathbf{T}(\rho)-c .^{4}$ A receiver who believes that his opponent plays $\sigma$ has a unique posterior belief $\mu_{\sigma}$ derived from Bayes' formula iff $\sigma$ has at least one non-zero entry in each column, i.e., each

[^54]message is expected to be sent with some positive probability (guaranteed by the quantal response function introduced below). This unique $\mu(\boldsymbol{\sigma})$ is just $\mathbf{N}(\mathbf{T}(\boldsymbol{\sigma}))$. The receiver's expected utilities are then $\mathrm{EU}(\sigma)=\mu(\sigma)$.

Best \& Quantal Responses Generally speaking, a response function maps expected utilities to choice probabilities. Rational choices maximize expected utility. In case of ties, rational agents are indifferent. If $U$ is an expected utility matrix, the rational best response function is $\operatorname{BR}(\mathrm{U})=\mathbf{N}(\max \operatorname{row}(\mathrm{U}))$. In contrast, the quantal response function assumes that agents make small mistakes in implementing the $\operatorname{BR}(\cdot)$ function. For given $U$, quantal response $\mathrm{QR}_{\lambda}(\mathrm{U})$ is the unique row-stochastic matrix with $\mathrm{QR}_{\lambda}(\mathrm{U})_{i j} \propto$ $\exp \left(\lambda \mathrm{U}_{i j}\right)$. Here $\lambda$ is a rationality parameter, with entirely random choices for $\lambda=0$ and $\lim _{\lambda \rightarrow \infty} \mathrm{QR}_{\lambda}(\mathrm{U})=\mathrm{BR}(\mathrm{U}) .{ }^{5}$

IQR The IQR model defines a hierarchy of player types. Unsophisticated level-0 behavior is anchored in the given semantics. Level- $(k+1)$ players play quantal responses to a belief that the interlocutor is of a lower type. Concretely, level-0 senders and receivers simply try to implement the semantic meaning: $\sigma_{0}=\mathrm{QR}_{\lambda}(B-c)$ and $\rho_{0}=\mathrm{QR}_{\lambda}\left(\mathbf{T}(B) I_{a}\right)$. Level$(k+1)$ player behavior is defined as a quantal response to a belief that the other player is at most of level $k$. Following the relevant literature in behavioral game theory (Camerer et al., 2004; Rogers et al., 2009), we subscribe to the simplifying assumption that the actual distribution of strategic types is a Poisson distribution $\operatorname{Pois}(\tau, k)=\tau^{k} / k!\exp (-\tau)$ with parameter $\tau$, and that agents know this. So, level- $(k+1)$ players' beliefs are derived by conditioning the underlying population distribution by the event that the opponent is less sophisticated: $f_{\tau}^{\leq k}(l)=\operatorname{Pois}(\tau, l) / \sum_{i=1}^{k} \operatorname{Pois}(\tau, i)$ if $l \geq k$ and 0 otherwise. This yields the following definition of level- $(k+1)$ players:

$$
\begin{array}{ll}
\sigma_{k+1}=\mathrm{QR}_{\lambda}\left(\mathrm{EU}\left(\rho_{\leq k}\right)\right) & \text { with } \rho_{\leq k}=\sum_{l \leq k} f_{\tau}^{\leq k}(l) \times \rho_{l} \\
\rho_{k+1}=\mathrm{QR}_{\lambda}\left(\mathrm{EU}\left(\sigma_{\leq k}\right)\right) \text { with } \sigma_{\leq k}=\sum_{l \leq k} f_{\tau}^{\leq k}(l) \times \sigma_{l}
\end{array}
$$

[^55]Given $\lambda$ and $\tau$, the model's behavioral prediction is the pair of strategies $\sigma^{*}=\sum_{k=1}^{\infty} f_{\tau}(k) \times \sigma_{k}$ and $\rho^{*}=\sum_{k=1}^{\infty} f_{\tau}(k) \times \rho_{k}$.
Model fitting As stated above, we fitted a mixed effects logistic regression to participants' behavior in the cost estimation task to estimate the difference $x$ between the log odds of costly vs. cheap message. Assuming that $x$ is the result of a quantal choice rule, we can compute the average subjective costs for a given fixed $\lambda$ as $c=x / 2 \lambda .{ }^{6}$ Using this, we determined a pair of parameters $\lambda$ and $\tau$ separately for the data on comprehension (Exps 1 and 3) and production (Exp. 2) using a least squares regression $(\lambda=4.825, \tau=0.625, r=0.99$ for comprehension; $\lambda=8.853, \tau=0.818, r=0.99$ for production). The prediction-to-data plot is given in Figure 3c.

These results are interesting in many respects. First, they serve as a proof-of-concept that a rather general game theoretic framework can predict behavioral data on language use and interpretation rather well. Second, the small but non-negligible $\tau$ indicates that participants in our experiment were able to perform one but not necessarily more steps of best response reasoning including considerations of production costs. Third, production behavior is better explained by higher values of $\lambda$ and $\tau$. This suggests that the model of Frank and Goodman (2012), which assumes that listeners perform two steps of optimization, while speakers perform exactly one, might be too inflexible. More relevant data is pending, but at present the more general model of Bergen et al. (2012) or the IQR model seem more realistic.

## Conclusion

The empirical contributions of this paper are two-fold. First, we provided evidence that language users can draw simple $a d$ $h o c$ scalar inferences in an artificial language paradigm. Second, we showed that even when the cost of an unambiguous message is only implicit (i.e., without telling participants explicitly that a message has a certain dollar value (Bergen et al., 2012; Rohde et al., 2012)), scalar inferences were drawn with increased frequency as costs of competing unambiguous messages increased. The theoretical value of this paper lies in the proposal for a probabilistic model of pragmatic reasoning that synthesizes previous Bayesian and game theoretic approaches. The model provides a good fit to the choice distributions of both speakers and listeners; and within listeners, for message cost effects, thus constituting a powerful model of pragmatic inference.

We conclude that not only do listeners take into account available utterances a speaker could have made, they also maintain a gradient estimate of production cost and take this into account in interpretation.

## Acknowledgments

We thank M. Tanenhaus, D. Kleinschmidt, and C. Kurumada and our three anonymous reviewers. This work was partially

[^56]supported by a EURO-XPRAG grant awarded to the authors and NIH grant HD-27206 to M. Tanenhaus.

## References

Bergen, L., Levy, R., \& Goodman, N. D. (2012). That's what she (could have) said: How alternative utterances affect language use. In Proceedings of the $34^{\text {th }}$ annual meeting of the cognitive science conference.
Camerer, C. F., Ho, T.-H., \& Chong, J.-K. (2004). A cognitive hierarchy model of games. The Quarterly Journal of Economics, 119(3), 861-898.
Clark, H. H., \& Marshall, C. R. (1981). Definite reference and mutual knowledge. In A. K. Joshi, B. Webber, \& I. A. Sag (Eds.), Elements of discourse understanding (p. 10-63). Cambridge University Press.

Degen, J., \& Franke, M. (2012). Optimal Reasoning About Referential Expressions. In S. Brown-Schmidt, J. Ginzburg, \& S. Larsson (Eds.), Proceedings of SemDial (p. 2-11).

Frank, M. C., \& Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. Science, 336(6084), 998.

Franke, M. (2011). Quantity implicatures, exhaustive interpretation, and rational conversation. Semantics \& Pragmatics, 4(1), 1-82.
Grice, H. (1975). Logic and conversation. Syntax and Semantics, 3, 41-58.
Hanna, J., Tanenhaus, M. K., \& Trueswell, J. C. (2003). The effects of common ground and perspective on domains of referential interpretation. Journal of Memory and Language, 49, 43-61.
Jaeger, T. F., \& Tily, H. (2011). On language 'utility': processing complexity and communicative efficiency. Wiley Interdisciplinary Reviews: Cognitive Science, 2(3), 323335.

Jäger, G. (2013). Rationalizable signaling. Erkenntnis.
Keysar, B., Barr, D. J., \& Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. Psychological Science, 11, 32-37.
Lewis, D. (1969). Convention. a philosophical study. Harvard University Press.
Luce, D. R. (1959). Individual choice behavior: A theoretical analysis. New York: Wiley.
Rogers, B. W., Palfrey, T. R., \& Camerer, C. (2009). Heterogeneous quantal response equilibrium and cognitive hierarchies. Journal of Economic Theory, 144(4), 1440-1467.
Rohde, H., Seyfarth, S., Clark, B., Jäger, G., \& Kaufmann, S. (2012). Communicating with Cost-based Implicature: a Game-Theoretic Approach to Ambiguity. In Proceedings of SemDial 16 (pp. 107-116).
Stiller, A., Goodman, N. D., \& Frank, M. C. (2011). Ad-hoc scalar implicature in adults and children. In Proceedings of the $33^{\text {rd }}$ annual conference of the cognitive science society.
Sutton, R. S., \& Barto, A. G. (1998). Reinforcement learning. MIT Press.

# Superficial, rather than true, knowledge interdependence in collaborative learning fosters individual knowledge transfer 

Anne Deiglmayr (anne.deiglmayr@ifv.gess.ethz.ch)<br>ETH Zurich, Research on Learning and Instruction, Universitaetsstrasse 41 8092 Zurich, Switzerland<br>Lennart Schalk (lennart.schalk@ifv.gess.ethz.ch)<br>ETH Zurich, Research on Learning and Instruction, Universitaetstrasse 41<br>8092 Zurich, Switzerland


#### Abstract

We test the hypothesis that superficial knowledge interdependence is more effective in fostering individual learning from collaboration than the true knowledge interdependence often realized by jigsaw-type collaboration arrangements. Based on research on group informationprocessing, we argue for the benefits of distributing only contextual information, but not core principles between learners, establishing superficial knowledge interdependence. In a computer-supported collaborative learning environment, 78 university students learned about stochastic urn models. Knowledge interdependence was established by systematically distributing learning materials within student triads, so that students either became experts for an urn model, establishing true knowledge interdependence, or for one of the embedding cover stories, establishing superficial knowledge interdependence. Afterwards, all triads worked on the same collaborative tasks, and were exposed to all models. Results show successful learning across conditions, but superior knowledge transfer in triads collaborating under superficial knowledge interdependence. Benefits were highest for low prior knowledge learners.


Keywords: computer-supported collaborative learning; learning through comparison; knowledge interdependence; knowledge transfer

## Introduction

In this paper, we explore different ways of distributing information between collaborative learners, with the goal of promoting the interactive construction of mathematical principles during learning from collaborative comparison of worked examples. In doing so, we address the more fundamental question of what characterizes optimal knowledge interdependence in collaborative learning, as assessed by measures of individual learning and transfer.

Collaborative learning has the potential of engaging students in forms of interactive knowledge construction that yield learning outcomes beyond those within the reach of an individual learner (Chi, 2009). However, this requires a certain amount of knowledge interdependence between students, that is, the individual students should hold a certain amount of unshared (unique) knowledge, ideas, and perspectives. The deliberate creation of knowledge interdependence is an important factor in many instructional methods for fostering collaborative learning, with the jigsaw collaboration script as their prototype. In a jigsaw
collaboration script, each learner becomes an expert for a specific domain before collaborating with other learners who have studied a different domain. To ensure fruitful collaboration, the distribution of expertise within groups typically ensures that "none of the group members has enough information or knowledge to solve the task alone" (Dillenbourg \& Jermann, 2007, p. 292), establishing true knowledge interdependence.

In fact, differences in prior knowledge and perspectives can lead to fruitful knowledge co-construction, in which ideas are critically evaluated, knowledge is elaborated and restructured, and more abstract representations are derived (Andriessen, Baker, \& Suthers, 2003; Schwartz, 1995). When learners integrate and transform their complementary knowledge resources, new knowledge can be created that no individual learner would have been capable of constructing (Deiglmayr \& Spada, 2011). On the other hand, research on group information processing shows that much of students` unshared knowledge remains unshared in real group discussions. For example, Buchs, Butera, and Mugny (2004) showed that students studying with a jigsaw collaboration script learned substantially less about their partner's domain of expertise than about their own, even though they were instructed to teach one another during a face-to-face learning phase. Deiglmayr and Spada (2011) showed that students had severe difficulties integrating interdependent information that was distributed between them.

Educators face the challenge of creating knowledge interdependence in a way that ensures that learners' discussions, and the cognitive activities involved, are focused on the most relevant learning content. Establishing true knowledge interdependence, as in classical jigsaw-type collaboration scripts, may not always be the optimal way to achieve this goal. Rather, we argue that superficial knowledge interdependence is often the better solution. Superficial knowledge interdependence denotes that core structures, such as domain principles and important concepts, remain shared between learners, while only contextual information, such as illustrative examples or application contexts, is distributed between learners. The fact that all relevant structural information is given to all students from the beginning maximizes the chance that each learner becomes familiar with the relevant principles via constructive learning processes, while the distributed
context information still creates sufficient interdependence for fostering truly interactive knowledge construction (Chi, 2009). In this paper, we test this "shared structure, distributed context"-hypothesis in a schema-abstraction learning setting (learning by collaborative comparison), with a learning domain that allows for a straightforward distinction between structure and context (word problems instantiating mathematical principles within different application contexts).

## Learning by collaborative comparison

Comparing and contrasting worked examples has proven an efficient way of fostering learning and transfer (for recent reviews see Alfieri, Nokes, \& Schunn, in press, and RittleJohnson \& Star, 2012). According to this approach at least two carefully constructed worked examples, which are instantiation of the to-be-learned principle or schema, are presented simultaneously in space and time. Learners are prompted to compare and contrast the examples in order to identify commonalities and differences (e.g., Gentner, Loewenstein \& Thompson, 2003; Schalk, Saalbach, \& Stern, 2011). These activities require learners to map and structurally align aspects of the worked examples, which "leads to learning via abstraction, rerepresentation, inference-projection, and difference-detection" (Gentner, 2010, p. 753). These are higher-order learning processes in which learners need to focus on deep, structural information rather than on contextual features, and to elaborate the to-be-learnt principles. In our collaborative comparison script, students begin with slightly different sets of examples from which they have to generate joint explanations of principles. This presumably fosters principle-based comparisons and elaboration via processes of grounding (Andriessen et al., 2003; Schwartz, 1995) and knowledge co-construction (Chi, 2009). Because the to-be-learnt principles (structural information) are embedded within different cover stories (contextual information), collaborative comparison as an instructional method allows to design well-controlled tests of the "shared structure, distributed context"-hypothesis.

## The domain: Learning to reason with probability

The relevant principles that students could learn in our experiment were urn models. These models serve to describe the probability of a series of random events (i.e., multilevel random experiments) in basic probability theory and allow for differentiating precisely between structure (urn models and the principles underlying them) and context (application contexts in the form of story problems).

A sound understanding of basic probability theory is a fundamental precondition for acquiring the ability to solve problems in statistics and, as such, is required in many professions and academic disciplines. High quality teaching seems to be particularly important as reasoning about probabilities does not come naturally to most people, and biases and misconceptions are abundant (Kahneman, Slovic, \& Tversky, 1982). Basic principles of probability theory and stochastics are introduced quite early in high school
mathematics. In Switzerland, for example, the principles governing multilevel random events (the learning domain from which our learning materials were taken) is introduced as early as in eighth grade. Typical problems are, for example, finding the probability of getting twelve points when throwing two dice, or finding the likelihood of guessing the right combination of numbers in a lotto game. The ultimate goal is that mathematical/statistical knowledge acquired in school will be applied outside the classroom and in students' later work; that is, that transfer occurs (Singley \& Anderson, 1989). However, transfer does not come about naturally even for these basic probability theory principles, and even university students have difficulties with basic stochastic concepts (Gal, 2002).

## The present research

In our experiment, university students had the chance to refresh and deepen their knowledge about basic probability theory, specifically, their knowledge about multilevel random events. The most important conceptual knowledge learners need to acquire when learning about multilevel random events is the ability to differentiate between four different urn models, in which random events are modeled as balls being drawn from an urn.

We combined learning through collaboration with learning triggered by comparing and contrasting worked examples in a collaborative comparison script. The script was modeled after a prototypal jigsaw script with an individual and a collaborative learning phase, implemented within a computer supported collaborative learning (CSCL) environment. Learning materials consisted of worked examples, which embedded the to-be-learned urn models in different cover stories. We varied whether, prior to collaboration, students became experts for one urn model (MODEL experts: true knowledge interdependence) or for one cover story (STORY experts: superficial knowledge interdependence). This setting allowed us to test our hypothesis that superficial knowledge interdependence would be more effective than true knowledge interdependence in fostering students` learning.

## Method

## Participants

Participants were 87 students of universities in Zurich (Switzerland), majoring in a wide range of subjects (students of mathematics or statistics were excluded). All participants spoke German or Swiss German as a native language. They were paid for participation. Participants were randomly assigned to triads and conditions. We excluded three triads from analysis because at least one of their members did not pass the threshold of four out of six correct answers in a basic prior knowledge test. This test assessed basic skills necessary for learning about multilevel random events (e.g. finding the likelihood of single random events in story problems; adding and multiplying fractions), or because they did not follow instructions. These exclusion
criteria left a total of 78 participants ( 42 female, 33 male) in 26 triads. Their age ranged from 18 to 36 years $(M=24.4$, $\mathrm{SD}=4.0$ ).

## Materials

Four urn models from probability theory (specifically, multilevel random events) were the core learning content of our learning environment. These four models result from combinations of two principles: relevance of order (the order in which balls are drawn from an urn is relevant vs. irrelevant) and replacement (the balls are drawn with replacement vs. without replacement). We will refer to these four models as Model 1 (order relevant, without replacement), Model 2 (order relevant, with replacement), Model 3 (order irrelevant, without replacement), and Model 4 (order irrelevant, with replacement). Story problems exemplified the four urn models by embedding them in simple cover stories (see Table 1 for examples). We used three different story problems, adapted with modifications from Berthold and Renkl (2009). In the remainder of this paper, these stories will be referred to as Story 1 (random events $=$ the distribution of bicycle helmets among participants in a biking course), Story 2 (random events = ranking results in a competition among equally capable skijumpers), and Story 3 (random events $=$ the drawing of unlabeled gas bottles from cupboards in a chemist`s lab). In the learning materials, we used nine story problems that result from crossing Models 1-3 with Stories 1-3. They were presented in the form of worked examples, that is, together with an arithmetic solution approach and a final numerical solution (as in Table 1). The three problems resulting from crossing Model 4 with Stories 1-3 were used as transfer tasks in the post test. All materials were presented within a computer-based learning environment.

## Measures and Scoring

Pretest In addition to the six basic knowledge questions used for screening participants (see Participants), the pretest contained four story problems assessing learners' prior knowledge about Models 1-4. The cover stories differed from those used in the learning phase. For each problem, one point could be obtained for generating an equation that corresponded to the model underlying the story problem.

Posttest The posttest had three sections. Within each section, the order of tasks was randomized. For each task, one point could be obtained for generating an equation that corresponded to the correct model. In the first section, three familiar tasks represented Models 1-3, each embedded in one of the Stories 1-3 that students already knew from the learning environment, but with new numerical values. In the second section, six direct application tasks embedded Models 1-3 in novel cover stories (two tasks for each model). The third section comprised the three tasks that result from crossing Model 4 with Stories 1-3. These Model 4 transfer tasks were included to measure transfer of the principles underlying Models 1-3: Since the four urn models result from crossing the principles relevance of order (relevant / irrelevant) and replacement (with / without), the fourth model can be derived from the other three. Students were told that the transfer tasks constituted a new type of model, but that they would be able to solve them by combining what they had learned during the learning phase.

## Procedure

Students came to our lab in groups of up to 18 participants. After a brief introduction, they were randomly assigned to computer work stations. Each student sat in his or her own cubicle, so that there was no face-to-face contact possible between learners. Students did not know with whom they

Table 1: Three worked examples from the learning materials (translated from the original language, German) exemplifying the three models and the three cover stories used in the learning phases

| Model 1, Story 1 | Model 2, Story 2 | Model 3, Story 3 |
| :---: | :---: | :---: |
| You and your friend participate in a two day mountain bike course. Each day, the instructor brings five bicycle helmets in five different colors which are randomly distributed among the course participants in the morning, and collected again in the evening. On both days, you are the first to receive a helmet, and your friend is the second. What is the probability for you to get the red helmet on the first day and the yellow helmet on the second day? | The four ski jumpers Adam, Beat, Christoph, and Daniel test a newly build ski-jumping hill today. The four ski jumpers have all performed equally well on previous competitions, thus, it only depends on random factors (e.g., wind regime) which of them will jump the greatest distance. There are two rounds of jumps. What is the probability that Adam will be on the first rank and Daniel on the second rank after the first round of jumps? | A chemist stores noble gases in two safes. There are the same three noble gases (argon, krypton, and xenon), in three identical single bottles, in both safes. Unfortunately, her colleague forgot to label the bottles. For her experiments, the chemist needs two different gases. The chemist takes one bottle out of each safe. What is the probability for her to obtain one bottle of argon and one of xenon? |
| Approach Solution | Approach Solution | Approach Solution |
| $\frac{1}{5} * \frac{1}{5} \quad=\frac{1}{25}$ | $\frac{1}{4} * \frac{1}{3} \quad=\frac{1}{12}$ | $\frac{1}{3} * \frac{1}{3}+\frac{1}{3} * \frac{1}{3} \quad=\frac{2}{9}$ |

were collaborating, and were logged into the system with an anonymous, gender-neutral nickname. After arriving at their workstations, students filled in a questionnaire on demographic variables and worked on the pretest individually. Afterwards, and before starting the learning phase, students received an introducing to the chat tool, and the three students who had been assigned to the same triad engaged in a brief warming-up chat session. The learning phase was segmented into an individual learning phase followed by a collaborative learning phase. Table 2 gives an overview of the worked examples presented in both phases, along with the self-explanation prompts provided (abbreviated for the individual learning phase).

Table 2: Learning materials (worked examples) for both experimental conditions. Worked examples are denoted by their combination of Model (M1-3) and Story (S1-3).

|  | Individual learning phase |  |
| :---: | :---: | :---: |
|  | MODEL-experts | STORY-experts |
| Learner | M1S1-M1S2-M1S3 | M1S1-M2S1-M3S1 |
| $\Omega_{A}$ | Commonalities? Differences? | Commonalities? Differences? |
| Learner | M2S1-M2S2-M2S3 | M1S2-M2S2-M3S2 |
| $\Omega_{B}$ | Commonalities? Differences? | Commonalities? <br> Differences? |
| Learner | M3S1-M3S2-M3S3 | M1S3-M2S3-M3S3 |
| § | Commonalities? <br> Differences? | Commonalities? Differences? |


| Collaborative learning phase |  |
| :---: | :---: |
| Triad | Screen 1: M1S1-M1S2-M1S3 |
| $\Omega_{A} \Omega_{B} \Omega_{C}$ | Why are the fractions multiplied rather than added up? |
| Triad | Screen 2: M2S1-M2S2-M2S3 |
| $\Omega_{A} \Omega \Omega_{B} \Omega$ | Why is the fractions' denominator decreasing? |
| Triad | Screen 3: M3S1-M3S2-M3S3 |
| $\Omega_{A} \Omega_{B} \Omega_{C}$ | Why does the solution require both addition and multiplication? |

The experimental variation was established in the individual learning phase, in which each learner studied three worked examples that were presented side-by-side on one screen. Learners were prompted to compare the examples and to list the most important similarities and the most important differences. Each member of a triad was assigned a different set of examples, so that, among them, the three learners studied all nine examples that result from crossing Models 1-3 with Stories 1-3. In the MODEL-experts condition, each triad member became an expert for a different urn model (true knowledge interdependence), whereas in the STORYexperts condition, each triad member became an expert for a different cover story (superficial knowledge interdependence).

In the collaborative learning phase, materials and instructions were identical for all triads, regardless of experimental condition. Three sets of worked examples, corresponding to Models 1-3, were presented on three consecutive screens (see Table 2). Thus, each and every learner was exposed to all nine worked examples during the collaborative learning phase. The triads compared and contrasted the worked examples and generated collaborative self-explanations. For each set of worked examples they were prompted to focus on one specific feature of the urn model being exemplified (see Table 2 for details). Triads used the chat tool in order to discuss their answer. Once group members had agreed on a joint solution, they went on to the next screen. After the collaborative learning phase, students worked on the posttest individually. All in all, the experiment took about 100 minutes.

## Results

There were no relevant differences between experimental conditions in participants` age, final high school math grade, or performance on the basic knowledge test used for participant screening (all $t \mathrm{~s}<|1.5|$; all $p \mathrm{~s}>.15$ ). Further, conditions did not differ significantly in the proportion of females/males $\left(\chi_{(d f=l)}^{2}=.83 ; p=.36\right)$. Conditions also did not differ in the distribution of students who solved $0,1,2$, 3 or 4 of the pretest Models 1-4 tasks correctly $\left(\chi_{(d f=3)}^{2}=\right.$ $.42 ; p=.94$ ) indicating similar levels of prior knowledge (see Table 3 for mean proportions correct).

Table 3: Mean proportions correct (and standard deviations) of pre- and post-test scores (total $N=78$ )

|  | MODEL <br> Experts | STORY <br> Experts | whole <br> sample |
| :--- | :--- | :--- | :--- |
| pretest: |  |  |  |
| Models 1-4 total | $.55(.24)$ | $.54(.26)$ | $.55(.25)$ |
| Models 1-3 only | $.68(.25)$ | $.66(.25)$ | $.67(.25)$ |
| Model 4 only | $.18(.39)$ | $.21(.41)$ | $.19(.39)$ |
| posttest: |  |  |  |
| Models 1-3 familiar | $.76(.26)$ | $.79(.24)$ | $.78(.25)$ |
| Models 1-3application | $.75(.23)$ | $.76(.19)$ | $.75(.21)$ |
| Models 1-3 combined | $.75(.22)$ | $.77(.19)$ | $.76(.20)$ |
| Model 4 transfer | $.46(.44)$ | $.62(.35)$ | $.54(.40)$ |

Before analyzing the post-test scores, we calculated intraclass correlations for the members of each triad in order to test for a possible hierarchical data structure. In no case was the ICC above .05 (all $F \mathrm{~s}<1.1$; all $p \mathrm{~s}>.40$ ), indicating only unsystematic agreement in post-test scores between triad members and, thus, a non-hierarchical data structure. Therefore, we calculated all further analyses on the level of individual learners $(\mathrm{N}=78)$. Given that our data is made up by series of 0 vs. 1 (correct vs. incorrect) responses, we calculated generalized logit regression models (using SPSS`s GENLIN procedure, with a logit link function) rather than $t$-tests or ANOVAs (Jaeger, 2008). However, for ease of comparison, Table 3 gives the scores that students in
the two experimental conditions obtained as mean proportions correct. Students in both conditions achieved very similar scores on the Models 1-3 familiar and the Models 1-3 direct application tasks. The differences between these two post-test sections (as within-subjects factor), experimental condition (as between-subjects factor), and their interaction were all statistically non-significant in a generalized logistic regression (all Wald $-\chi^{2}{ }_{(d f=1)}<2.6$; all $p \mathrm{~s}>.11)$. We therefore formed a combined posttest score (Table 3: Models 1-3 combined).

We first looked at students` posttest performance on tasks representing Models $1-3$, that is, the learning content we directly taught. Table 3 shows that students in both conditions showed an overall gain in their performance from pre- to posttest. We calculated a generalized logistic regression with solution rate as the dependent variable, time (pretest: Models 1-3 vs. posttest: Models 1-3 combined) as within-subjects factor, and experimental condition as between-subjects factors. Only the effect of time was significant $\left(\right.$ Wald $\left.\left.-\chi^{2}{ }_{(d f}=1\right)=6.5 ; p=.01\right)$. These findings indicate that both conditions were effective in improving the recognition and application of the three urn models that were directly taught.

On the Model 4 transfer tasks, however, students` posttest performance was notably higher in the STORY-experts condition (Table 3). Figure 1 shows that the absolute solution rate shows a U-shaped distribution in the MODELexperts conditions, while the mode of the distribution in the STORY-experts condition is at the highest end of the distribution. This difference in distribution of scores between conditions is statistically significant $\left(\chi^{2}{ }_{(d f}=3\right)=$ 8.55; $p=.04$ ).

To further scrutinize the differential effects on transfer in both conditions, we took students' prior knowledge into account. We tested the effects of experimental condition, prior knowledge (specified as a covariate), and their interaction, on the number of correctly solved Model 4 transfer tasks in a generalized logistic regression model. We chose the combined pretest score for Models 1-4 as the most reliable and most informative predictor; however, analyses with performance on only the items for Models 1-3 yielded the same pattern of results; the same was true when the 15 students who had already mastered the Model 4 task in the pretest were excluded from analysis. All postulated
predictors in the model (experimental condition, prior knowledge, and their interaction) were shown to significantly predict performance on the Model 4 transfer tasks (for parameter estimates see Table 4; overall model likelihood ratio: $\left.\chi^{2}{ }_{(d f=3)}=39.83 ; p<.001\right)$. The significant interaction indicates that learners low in prior knowledge profited more in the STORY-experts condition than in the MODEL-experts condition: Prior knowledge showed a significant, positive correlation with transfer performance in the MODEL-experts condition (Spearman`s \(r=.56, p<\) .001) but a smaller, statistically non-significant correlation in the STORY-experts condition (Spearman`s $r=.25, p=$ .13).


Figure 1: Distribution of learners (by experimental condition) who solved 0, 1, 2, or all 3 Model 4 transfer tasks correctly

## Discussion

In the present study, we aimed at testing the hypothesis that in collaborative learning settings superficial knowledge interdependence is more effective in fostering individual learning than true knowledge interdependence. Specifically, we tested whether collaborative learning supported by a jigsaw-type collaboration script is more effective when the knowledge interdependence established between students ensures that the to-be-learnt, structural information (in our case, the three urn models) is shared from the beginning, while only contextual information (in our case, the cover

Table 4: Summary of effects in the generalized logit model with experimental condition, prior knowledge, and their interaction as predictors of students' performance on the Model 4 transfer tasks $\left(N_{\text {subjects_x_trials }}=234\right)$

| Predictor | Coefficient <br> (B) | SE | Coefficient (B): $95 \%-C I$ (Wald) | $e^{B}$ | Wald $\chi^{2}{ }_{(d f=1)}$ | $p$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Intercept | -. 37 | . 46 | [-1.28; .54] | . 69 | . 64 | . 42 |
| $\begin{gathered} \text { Experimental Condition } \\ \text { MODEL-experts }=0 \\ \text { STORY-experts }=1 \end{gathered}$ | -2.44 | . 76 | [-3.94; -.95] | . 09 | 10.29 | $<.01$ |
| Prior Knowledge <br> (Models 1-4 pretest score) | . 41 | . 20 | [.01; .81] | 1.51 | 4.11 | . 04 |
| Interaction: Experimental Condition x Prior Knowledge | . 79 | . 33 | [.15; 1.44] | 2.21 | 5.76 | . 02 |

stories) is distributed between learners (shared structure distributed context hypothesis).

The results partially support our hypothesis: Students in the STORY-experts condition (superficial knowledge interdependence) did profit more from our CSCL learning environment than students in the MODEL-experts condition (true knowledge interdependence), but only on the transfer tasks. In both conditions, students gained to a similar degree from pre- to post-test for the three models that had been trained. Since students in both conditions learned with a highly structured learning environment and with carefully constructed worked examples, this finding is reassuring. Still, STORY-experts outperformed MODEL-experts on the transfer tasks, which required them to combine the principles behind the three trained models in order to derive a solution for a fourth model that had not been introduced within the learning environment. Learners with low prior knowledge profited particularly from the superficial knowledge interdependence realized in the STORY-experts condition, that is, they were more likely to obtain a high score on the transfer tasks in this condition.
We assume that these effects arise because the superficial knowledge interdependence realized in the STORY-experts condition (1) ensures that each learner becomes familiar with all relevant principles via constructive learning processes already during the preparatory individual learning phase, while (2) the distributed context information still creates sufficient interdependence for fostering truly interactive knowledge construction (Chi, 2009). However, further fine grained analyses of individual learning (e.g. self-explanations during individual learning phase) and of collaborative processes (e.g., discourse analyses of chats) are needed to be able to precisely identify the underlying cognitive and interactive processes. Analyses currently under way include coding the quality of students` selfexplanations, as a measure of the level of expertise they gained during the individual learning phase, as well as analyses of the patterns of contributions, both qualitatively and quantitatively, made within story expert and model expert triads. Further experiments will include additional test and transfer tasks in order to increase the reliability of the pre- and post-test measures, and will be designed to enable direct comparisons with purely individual (constructive) learning conditions.

## Acknowledgments

(1) This research was partially supported by an early career scholarship to the first author from the Swiss National Science Foundation. (2) Both authors contributed equally.

## References

Alfieri, L., Nokes-Malach, T. J., \& Schunn, C. D. (in press). Learning through case comparisons: A meta-analytic review. Educational Psychologist.
Andriessen, J., Baker, M., \& Suthers, D. (Eds.) (2003). Arguing to learn: Confronting cognitions in computer-
supported collaborative learning environments. Dordrecht: Kluwer.
Berthold, K., \& Renkl, A. (2009). Instructional aids to support a conceptual understanding of multiple representations. Journal of Educational Psychology, 101(1), 70-87.
Buchs, C., Butera, F., \& Mugny, G. (2004). Resource interdependence, student interactions and performance in cooperative learning. Journal of Experimental Educational Psychology, 24(3), 291-314.
Chi, M. T. H. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. Topics in Cognitive Science, 1, 73-105.
Deiglmayr, A., \& Spada, H. (2011). Training for fostering knowledge co-construction from collaborative inferencedrawing. Learning and Instruction, 21, 441-451.
Dillenbourg, P., \& Jermann, P. (2007). Designing integrative scripts. In F. Fischer, H. Mandl, J. Haake, \& I. Kollar (Eds.), Scripting computer-supported collaborative learning -- Cognitive, computational, and educational perspectives (pp. 275-301). New York: Springer Computer-Supported Collaborative Learning Series.
Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. International Statistical Review, 70(1), 1-25.
Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. Cognitive Science, 34, 752-775.
Gentner, D., Loewenstein, J., \& Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. Journal of Educational Psychology, 95(2), 393-408.
Jaeger, T. F. (2008). Categorical data analysis: away from ANOVAs (transformation or not) and towards Logit Mixed Models. Journal of Memory and Language, 59, 434-446.
Kahneman, D., Slovic, V. M., \& Tversky, A. (Eds.) (1982). Judgment under Uncertainty: Heuristics and Biases. New York: Cambridge University Press.
Rittle-Johnson, B., \& Star, J. R. (2012). The power of comparison in learning and instruction: Learning outcomes supported by different types of comparison. In B. Ross \& J. Mestre (Eds.), Psychology of Learning and Motivation: Cognition in Education (Vol. 55). San Diego: Elsevier.
Schalk, L., Saalbach, H., \& Stern, E. (2011). Designing learning materials to foster transfer of principles. In L. Carlson, C. Hölscher, \& T. Shipley (Eds.), Proceedings of the $33^{\text {rd }}$ Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.
Schwartz, D. L. (1995). The emergence of abstract representations in dyad problem solving. Cognition and Instruction, 22(2), 129-184.
Singley, M. K., \& Anderson, J. R. (1989). Transfer of Cognitive Skill. Cambridge, MA: Harvard University Press.

# Analogical Reasoning with Rational Numbers: Semantic Alignment Based on Discrete Versus Continuous Quantities 

Melissa DeWolf (mdewolf@ucla.edu)<br>Department of Psychology, University of California, Los Angeles<br>Los Angeles, CA, USA

Miriam Bassok (mbassok@u.washington.edu)
Department of Psychology, University of Washington
Seattle, WA, USA

Keith J. Holyoak (holyoak@lifesci.ucla.edu)<br>Department of Psychology, University of California, Los Angeles<br>Los Angeles, CA, USA


#### Abstract

Non-integer rational numbers, such as fractions and decimals, pose challenges for learners, both in conceptual understanding and in performing mathematical operations. Previous studies have focused on tasks involving access and comparison of integrated magnitude representations, showing that adults have less precise magnitude representations for fractions than decimals. Here we show the relative effectiveness of fractions over decimals in reasoning about relations between quantities. We constructed analogical reasoning problems that required mapping rational numbers (fractions or decimals) onto pictures depicting either partwhole or ratio relations between two quantities. We also varied the ontological nature of the depicted quantities, which could be discrete, continuous, or continuous but parsed into discrete components. Fractions were more effective than decimals for reasoning about discrete and continuous-parsed (i.e., discretized) quantities, whereas neither number type was particularly effective in reasoning about continuous quantities. Our findings show that, when numbers serve as models of quantitative relations, the ease of relational mapping depends on the analogical correspondence between the format of rational numbers and the quantity it models.


Keywords: analogy; relational reasoning; number concepts; fractions; decimals; semantic alignment; math education

## Introduction

## Mathematical Understanding as Relational Reasoning

Mathematics is in essence a system of relations among concepts based on quantities. A core problem with math education, particularly in the United States, is that greater focus is placed on execution of mathematical procedures than on understanding of quantitative relations (Richland, Stigler \& Holyoak, 2012; Stigler \& Hiebert, 1999; RittleJohnson \& Star, 2007). An early manifestation of this problem involves teaching of non-integer rational numbers in the standard curriculum-typically, first fractions and subsequently decimals. Students often leave middle-school (and often enter community college: see Stigler, Givvin \& Thompson, 2010; Givvin, Stigler \& Thompson, 2011)
without having grasped how fractions relate to decimals, or how either number type relates to integers. This conceptual disconnection in turn contributes to a compartmentalization of mathematical operations (e.g., multiplication of fractions is treated as unrelated to multiplication of integers; Siegler et al., 2011; Siegler \& Pyke, 2012).

Although mathematical relations are typically construed as internal to the formal system of mathematics, the application of mathematics to real-world problems also depends on grasping relations between mathematical concepts and the basic ontological distinctions among the concepts to which mathematics must be applied. Rather than treating mathematical concepts as purely formal, both children and adults are naturally guided by a process of semantic alignment, which favors mapping certain mathematical concepts (and their associated operations) onto certain conceptual types. Bassok, Chase and Martin (1998) demonstrated that the basic mathematical operations of addition, subtraction, multiplication, and division are typically conceptualized within a system of relations between mathematical values and objects in the real world. Specific mathematical operators are semantically aligned with particular relationships among real-world objects. For example, addition is aligned with categorical object relations (e.g., people find it natural to add two apples plus three oranges, because both are subtypes of a common category, fruit), whereas division is aligned with functional object relations (e.g., a natural problem would be to divide ten apples between two baskets). Semantic alignment has been demonstrated with both children and adults (e.g., Martin \& Bassok, 2005), and for many adults the process is highly automatic (Bassok, Pedigo, \& Oskarsson, 2008). Although natural semantic alignments are implicitly acknowledged in the construction of textbook word problems (Bassok et al., 1998), teachers seldom discuss these alignments with their students. This gap may contribute to the difficulty of conveying how and why mathematical formalisms "matter" in dealing with realworld problems.

## Discreteness Versus Continuity

A particularly important ontological distinction relevant to mathematical modeling involves the nature of quantities, which can be viewed as either discrete or continuous. Roughly, some entities are viewed as comprising a set of individual objects (e.g., a number of apples in a basket), whereas others are viewed as a continuous mass without individuation (e.g., a bucket of water). Although continuous, as well as, discrete quantities can be subdivided, in the case of continuous quantities the divisions are arbitrary in the sense that they do not isolate conceptual parts (e.g., one could distinguish between subsets of red and green apples in a basket by saying that $2 / 3$ of the apples are red and $1 / 3$ are green, but there is no psychological difference between the water contained in $2 / 3$ of a bucket and in the complementary $1 / 3$ of the bucket). Importantly, discreteness versus continuity is a distinction based fundamentally not on physics, but on psychology. For example, a pile of sand is viewed as a continuous quantity even though we know it is composed of individual grains, because those units are too small and interchangeable to be typically viewed as "important". The impact of this basic ontological distinction has been documented both in young babies (e.g., Spelke, Breilinger, Macomber, \& Jacobson, 1992) and in adults (Bassok \& Olseth, 1995). For example, Bassok and Olseth found that college students viewed an increase in attendance at an annual conference as discrete (since it is based on a change between magnitudes associated with two discrete events well-separated in time), but viewed an annual increase in a city's population growth as continuous (since it is based on changes stemming from the psychologicallyconstant process of gaining and losing undifferentiated individual residents).

The ontological distinction between discreteness and continuity underlies the linguistic distinction between count and mass nouns, which is syntactically important in English and many other natural languages (Bloom \& Wynn, 1997). Infants and young children are able to make distinctions among continuous quantities (Clearfield \& Mix, 2001; Fiegenson, Carey \& Spelke, 2002). However school-aged children have an advantage when performing operations with discrete quantities (e.g., counting; Gelman, 1993) over performing operations with their continuous counterparts (e.g., measurement in general; Nunes, Light \& Mason, 1993). Indeed, measurement of continuous quantities depends on the introduction of arbitrary equal-sized units, which serve to parse a continuous whole into countable subparts (e.g., a continuous length can be broken down into some number of inches or centimeters). The ability to discretize continuous concepts (as contrasted with the lack of a natural operation to make discrete concepts continuous) leads to asymmetries in transfer of mathematical operations. For example, college students can transfer the equation for calculating the sum of an arithmetic progression (discrete concept) to solve a physics problem requiring solving for final velocity after constant acceleration (continuous concept), but find transfer in the opposite direction
(continuous to discrete) nearly impossible (Bassok \& Holyoak, 1989; Bassok \& Olseth, 1995).

## Fractions as Relational Representations

As the first non-integer number type introduced to elementary-school students, fractions pose particular challenges. Research indicates that children have difficulty integrating fractions into their already well-established understanding of whole numbers (Staflyidou \& Vosniadou, 2004; Vamvakoussi \& Vosniadou, 2010; Ni \& Zhou, 2005), and even adults at community colleges seem to lack fundamental understanding of how to use fractions (Stigler et al., 2010). Research on understanding fractions has primarily focused on the ability to grasp and manipulate their integrated magnitude value associated with the $a / b$ form. Although adults can compare fractions based on integrated magnitudes (Schneider \& Siegler, 2010), this process is very slow and error-prone relative to performing the same task with decimal equivalents (DeWolf, Grounds, Bassok \& Holyoak, in press). The difficulty of making magnitude comparisons with fractions presumably reflects their bipartite structure (numerator divided by denominator), which makes them both formally and conceptually distinct from integers. In contrast, decimals have a unitary structure more similar to integers (though not identical; Cohen, 2010).

However, even though the internal structure of fractions apparently hinders access to precise integrated magnitudes, this same structure may facilitate understanding of key relations. In particular, the $a / b$ form can be aligned with the concepts underlying relations such as part/whole, subset/set, ratio, and rate. When children are first taught the concept of a fraction, some type of pictorial representation is often provided, such that each of the two values in the fraction are structurally aligned with two separate elements in the picture. For example, take the very common example of cutting up a pizza pie into pieces. A child might be taught that $4 / 5$ is equivalent to 4 slices of a pizza pie that is divided into 5 slices. This type of mapping is also encouraged with verbal examples (e.g., 4 out of every 5 dentists recommend a certain toothpaste). Such instructional practices highlight the relational nature of fractions and encourage children to reason about fractions relationally.

We propose that semantic alignment will also modulate people's understanding of fractions. Fractions seem particularly appropriate as models of relations between sets of discrete elements. The representations typically used to teach fractions focus on discrete countable units that can map to the numerator and denominator values. For example, the pizza is sliced into exactly the number of pieces in the denominator before the numerator pieces are counted up. Rapp and Bassok (in preparation) reviewed a math textbook series and found that very rarely are students encouraged to think about fractions with continuous representations, such as a number line. In fact, continuous measures (e.g., length, weight) are almost exclusively represented with decimals. Rapp and Bassok also found that, consistent with this


Figure 1: Examples of types of pictures used in analogy problems.
training, college students show a preference for using fractions rather than decimals to describe relations between discrete quantities, and use decimals to describe magnitudes of continuous quantities.

## Analogical Reasoning with Quantitative Relations

Our study was designed to test the hypothesis that, due to their relational structure $(a / b)$, fractions are better suited than decimals for representing relations between countable quantities. To this end, we compared analogical reasoning with either fractions or decimals, while varying the ontological distinction between discrete and continuous concepts. Figure 1 shows examples of variations in discreteness versus continuity. The pictorial stimuli were based on discrete elements (top), continuous rectangles (bottom), or continuous rectangles parsed into discrete units (middle). We hypothesized that semantic alignment would yield higher accuracy and faster response times for solving analogies using fractions rather than decimals for the discrete and continuous-parsed pictures. The fraction advantage was predicted to disappear or even reverse for the continuous pictures, where the semantic alignment is most difficult.

## Method

## Participants

Participants were 52 undergraduates at the University of California, Los Angeles (mean age $=21 ; 30$ females) who received course credit, randomly assigned in equal numbers to the two between-subjects conditions.

## Materials and Design

The study was a 2 (number type: fractions vs. decimals) X 2 (relation type: ratios vs. part/whole fractions) X 3 (picture


Figure 2: Example of an analogy problem (part/whole fraction trial with continuous-parsed pictures).
type: continuous, continuous-parsed, discrete) design, with number type as a between-subjects factor and relation type and picture type as within-subjects factors.

The analogy problems were constructed using each of the three ontological types illustrated in Figure 1: discrete, continuous-parsed, and continuous. An example problem appears in Figure 2. These analogy problems were in the format $\mathrm{A}: \mathrm{B}:: \mathrm{C}: \mathrm{D}$ vs. D', where the source analog ( $\mathrm{A}: \mathrm{B}$ ) consisted of a picture and a number (fraction or decimal). The task required making a choice of the correct number to complete the target analog. The number type was always the same across the source and target.

Solving an analogy problem required first identifying the relationship in the A picture characterized by the number given as B . This relationship could be part-whole or a ratio between two parts. In Figure 2, the A picture indicates 4 green units out of a total of 6 , making a part-whole relation of $4 / 6$ (. 67 in a matched problem using decimals). An alternative ratio relation in Figure 2 is based on the units of red relative to green (i.e., $2 / 4$, or .50 in decimal notation). Once the higher-order relation between A and B was extracted, the solution required identifying the same relation type in target picture C , and choosing the corresponding number as D term. D' mapped to the alternative relationship.

As Figure 2 illustrates, the same two colors were used in the A and C pictures, and the color relationship was maintained, such that the same color mapped to the same part of the relation in both A and C. This constraint served to identify which part (lesser or greater) mapped to the numerator in a ratio relation. Color assignments varied across trials, so the same color might indicate the lesser subset on one trial and the greater subset on another. The actual test trials contained only red and green colors (practice trials were given that had yellow and brown colors). The discrete items were circles, squares, crosses, trapezoids, and cloud-like shapes. Continuous and continuous-parsed items differed in width, height and


Figure 3: Percent accuracy for solving analogy problems using fractions or decimals for each quantity type.
orientation (vertical or horizontal). For each trial, the source and target were randomly assigned for each participant so that the only thing that was consistent between the two was the higher-order relationship (ratio or part/whole) and the color mapping. The fractions and decimals were always less than one and decimals were shown rounded to two decimal places.

## Procedure

Stimuli were displayed with Macintosh computers using Superlab 4.5, and response times and accuracy were recorded. Extensive instructions and practice was provided prior to beginning the test trials. Participants were told that there were two different types of relations between the pictures and values. For the ratio relationship, participants were shown a picture with 1 O and 2 X 's. For the fractions condition this was explained as " $1 / 2$ amount of O's per amount of X's;" for the decimals condition it was explained as ". 50 amount of O's per amount of X's." The part/whole relationship was represented with a picture of 2 O 's and 3 $X$ 's. For the fractions condition this was explained as " $2 / 5$ of the total is the amount of O's;" for the decimal condition it was explained as ". 40 of the total is the amount of O's." The first of these explanations of the ratio and part/whole relations was shown with discrete items. The following screen showed the same values paired with continuousparsed pictures. A third screen showed the same values paired with continuous pictures.

After this introduction, participants completed an example trial in which they were shown the source ( $\mathrm{A}: \mathrm{B}$ ), asked to figure out the type of relation (ratio or part/whole) in their head, and press the space bar. After the space bar was pressed, the target (C:D vs. D') was shown on the screen below the source so that the two components were on the screen simultaneously. Participants were asked to select which of two numbers ( D or $\mathrm{D}^{\prime}$ ) shared the same relationship with the picture as the relationship provided in the source. Half of the time, D appeared on the right side of the screen. They made their selection by pressing the $z$ key


Figure 4: Response times for correctly solving analogy problems using fractions or decimals for each quantity type.
for the number shown on the left and the $m$ key for the number shown on the right. The $z$ and $m$ keys were labeled with "L" and "R", respectively, so that participants could remember which key went with each number. After completing the initial example trial, participants were shown the correct answer, with an explanation of which relationship was shared between the source and target.

Participants then completed 12 practice trials. Feedback was given for incorrect trials, in the form of a red " X " on the screen. After the practice trials had been completed, a screen was displayed informing participants that the actual test trials were beginning. For each trial, the source was shown, then the participant pressed the spacebar when they determined the relationship. The target was then shown in addition to the source. Feedback was continued for incorrect trials. There were 72 test trials ( 12 for each of the 6 within-subjects conditions). The specific pictures, numbers, and pairings used in the test trials were different from those used in practice trials. Relation types and picture types were shown in a different random order for every participant.

## Results

Accuracy and mean response time (RT) on correct trials were computed for each condition for each participant. A mixed factors ANOVA was used to compare differences in RT and accuracy. No reliable overall differences were obtained between the two relation types (part-whole and ratio) on either measure; hence all results are reported after collapsing across this variable. Figure 3 presents the pattern of performance based on accuracy, and Figure 4 presents the pattern based on mean correct RT. Both dependent measures revealed an overall advantage for solving analogies based on fractions rather than decimals, with the advantage most pronounced for pictures of discrete quantities. For accuracy, both number type, $F(1,50)=8.65, p=.005)$, and picture type, $F(2,49)=33.52, \mathrm{p}<.001$, were highly reliable, as was the interaction of the two factors, $F(2,49)=25.20, p<.001$. Planned comparisons indicated that accuracy was higher for fractions than decimals for the discrete condition ( $87 \%$ vs.
$66 \% ; t(50)=5.38, p<.001)$ and the continuous-parsed condition ( $80 \%$ vs. $67 \% ; t(50)=3.17, p=.003$ ), but did not differ for the continuous condition $(61 \%$ vs. $65 \% ; t(50)=$ $.93, p=.36$ ).

RTs were measured from the onset of the source display on the screen to the selection of the target answer. Response times for incorrect answers were excluded from analyses. In addition, outliers were trimmed to exclude any times that were greater than three standard deviations from the mean (roughly 2\% of response times). As shown in Figure 4, the RT pattern closely resembled that for accuracy. In particular, there was a reliable interaction between number type and picture type, $F(2,49)=16.19, p<.001$. Planned comparisons indicated that RTs were faster with fractions than decimals for the discrete condition ( 8.5 s vs. 12.8 s ; $t(50)=2.70, p=.01)$, with a strong trend for the continuousparsed condition ( 8.3 s vs. $11.2 \mathrm{~s} ; t(50)=1.87, p=.067$ ). RTs for fractions versus decimals did not differ reliably for the continuous condition ( 9.3 s vs. $7.7 \mathrm{~s} ; t(50)=1.45, p>$ .15).

## Discussion

The results of the current study demonstrated an overall advantage for fractions over decimals in a relational task. Moreover, this advantage was moderated by the ontological nature of the depicted quantities. Participants were better able to extract relationships for discrete and continuousparsed pictures when fractions were mapped to the quantities, rather than decimals. There was no difference in performance on the continuous pictures between fractions and decimals. This pattern suggests that fractions are semantically-aligned with relations between countable, discrete quantities. Performance with decimals was relatively flat (and generally poorer) for all picture types.

These results support two basic claims about semantic alignment for specific types of quantities. First, people can and do align quantities with numbers. Second, ease of alignment depends on two factors: the type of number format (fractions vs. decimals), and the type of quantities (countable, i.e., discrete and continuous-parsed, vs. continuous).

The central difference between fractions and decimals is that their formats provide an explicit representation of relations (fractions) or of relation magnitudes (decimals). That is, fractions have a bipartite structure $(a / b)$ that expresses a specific relationship between two natural numbers, $a$ and $b$. Decimals represent the magnitudes of such fractional relations. This difference has important implications for how people align these numbers with specific quantities. For fractions, alignment should be simple when the numerator and denominator can be readily mapped onto distinct subsets, A and B. Our results show that this is indeed the case when $A$ and $B$ are comprised of countable entities, depicted by the discrete and continuousparsed picture types. However, alignment should be difficult when $A$ and $B$ are continuous quantities, as the task
becomes more like a magnitude assessment. Despite the explicit relation $(a / b)$, it is difficult to assess the magnitude of A and the magnitude of B , which makes the overall mapping more complicated. Decimals represent magnitudes of relations without specifying the relational parts. Hence, mapping to the A and B sets is difficult irrespective of whether the sets are shown as discrete or continuous quantities.

The current pattern of results is consistent with the schooling experience of our participants (Rapp \& Bassok, in preparation). Typically, students learn about fractions from part/whole and set/subset examples (Sophian, 2007; Mack, 1993). However, the goal of such examples is not to help children understand that fractions represent relations. Rather, they are provided to help children understand the existence of values smaller than 1 . That is, as discussed earlier, the main focus of initial instruction about fractions is to convey their magnitude. This focus is problematic because, while fractions are well-suited for representation of relations, they are poorly suited for representation of magnitudes (DeWolf et al., in press; Stigler et al., 2010). Our findings also suggest that if decimals were taught prior to fractions, children might have a better opportunity to learn about magnitudes smaller than 1. Because decimals have a unitized format, like whole numbers, they might provide an easier opportunity for children to master the idea of magnitudes smaller than 1. Fractions, then, might be taught later than decimals with an emphasis on their status as a relationship between two natural numbers. The magnitude of such relational representations would not be limited to values smaller than 1 (ratios).

Interestingly, Moss and Case (1999) implemented a curriculum with $4^{\text {th }}$ graders in Canada that reorganized the order of rational number instruction. Children were first taught percentages (in the context of volumes and on number lines), then decimals, and lastly fractions. Fractions were explained simply as another way to represent a decimal. By contrast, typical curricula describe teaching decimals as another way to represent a fraction. Moss and Case found that children taught number types in this novel sequence suffered less interference from whole-number strategies when using other rational numbers, and achieved a deeper understanding of them. Though Moss and Case did not emphasize fractions in the relational context we have discussed here, it seems that introducing the idea of magnitudes less than 1 with decimals rather than fractions may be preferable.

In summary, understanding how non-integer rational numbers align to specific types of quantities, and how format can affect ease of semantic alignment, has important implications for how we conceptualize and teach fractions and decimals. It is important to foster understanding of fractions beyond simple algorithmic procedures, and to bolster conceptual understanding in order to address the difficulties children and adults face in understanding fractions.

## Acknowledgments

Preparation of this paper was supported by grant R305C080015 from the Institute of Education Sciences.

## References

Bassok, M., Chase, V. M., \& Martin, S. A. (1998). Adding apples and oranges: Alignment of semantic and formal knowledge. Cognitive Psychology, 35, 99-134.
Bassok, M., \& Holyoak, K., J. (1989). Interdomain transfer between isomorphic topics in algebra and physics. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15, 153-166.
Bassok, M., \& Olseth, K. L. (1995). Object-based representations: Transfer between cases of continuous and discrete models of change. Journal of Experimental Psychology: Learning, Memory, and Cognition, 21, 15221538.

Bassok, M., Pedigo, S. F., \& Oskarsson, A. T. (2008). Priming addition facts with semantic relations. Journal of Experimental Psychology: Learning Memory and Cognition, 34, 343-352.
Bloom, P., \& Wynn, K. (1997). Linguistic cues in the acquisition of number words. Journal of Child Language, 24, 511-533.
Clearfield, M. W., \& Mix, K. S. (2001). Infants use continuous quantity-not number-to discriminate small visual sets. Journal of Cognition and Development, 2, 243-260.
Cohen, D. J. (2010). Evidence for direct retrieval of relative quantity information in a quantity judgment task: Decimals, integers, and the role of physical similarity. Journal of Experimental Psychology: Learning, Memory, and Cognition, 36, 1389-1398.
DeWolf, M., Grounds, M. A., Bassok, M., \& Holyoak, K. J. (in press). Representation and comparison of magnitudes for different types of rational numbers. Journal of Experimental Psychology: Human Perception and Performance.
Feigenson, L., Carey, S., \& Spelke, E. (2002). Infants' discrimination of number vs. continuous extent. Cognitive Psychology, 44, 33-66.
Gelman, R. (1993). A rational-constructivist account of early learning about numbers and objects. In D. Medin (Ed.). Learning and motivation, Vol. 30 (pp. 61-96). New York: Academic Press.
Givvin, K. B., Stigler, J. W., \& Thompson, B. J. (2011). What community college developmental mathematics students understand about mathematics, Part II: The interviews. The MathAMATYC Educator, 2, 4-18.
Mack, N. K. (1993). Learning rational numbers with understanding: The case of informal knowledge. In T. P Carpenter, E. Fennema, \& T. A. Romberg (Eds.), Rational numbers: An integration of research (pp. 85105). Mahwah, NJ: Erlbaum.

Martin, S. A., \& Bassok, M. (2005). Effects of semantic cues on mathematical modeling: Evidence from wordproblem solving and equation construction tasks. Memory \& Cognition, 33, 471-478.
Moss, J., \& Case, R. (1999). Developing children's understanding of the rational numbers: A new model and an experimental curriculum. Journal for Research in Mathematics Education, 30, 122-147.
Ni, Y., \& Zhou, Y. (2005). Teaching and learning fractions and rational numbers: The origins and implications of whole number bias. Educational Psychologist, 40, 27-52.
Nunes, T., Light, P., Mason, J. (1993) Tools for thought: The measurement of length and area. Learning and Instruction, 3, 39-54.
Rapp, M., \& Bassok, M. (in preparation). Form and meaning of rational numbers. Department of Psychology, University of Washington.
Richland, L. E., Stigler, J. W., Holyoak, K. J. (2012). Teaching the conceptual structure of mathematics. Educational Psychologist, 47, 189-203.
Rittle-Johnson, B., \& Star, J. R. (2007). Does comparing solution methods facilitate conceptual and procedural knowledge? An experimental study on learning to solve equations. Journal of Educational Psychology, 99, 561574.

Schneider, M., \& Siegler, R.S. (2010). Representations of the magnitudes of fractions. Journal of Experimental Psychology: Human Perception and Performance, 36, 1227-1238.
Siegler, R. S., \& Pyke, A. A. (2012). Developmental and individual differences in understanding of fractions. Developmental Psychology http://dx.doi.org/10.1037/ a0031200
Siegler, R. S., Thompson, C. A., \& Schneider, M. (2011). An integrated theory of whole number and fractions development. Cognitive Psychology, 62, 273-296.
Sophian, C. (2007). The origins of mathematical knowledge. Mahwah, NJ: Erlbaum.
Spelke, E. S., Breinlinger, K., Macomber, K., \& Jacobson, K. (1992). Origins of knowledge. Psychological Review, 99, 605-632.
Stafylidou, S., \& Vosniadou, S. (2004). The development of student's understanding of the numerical value of fractions. Learning and Instruction, 14, 508-518.
Stigler, J. W., Givvin, K. B., \& Thompson, B. (2010). What community college developmental mathematics students understand about mathematics. The MathAMATYC Educator, 10, 4-16.
Stigler, J. W., \& Hiebert, J. (1999). The teaching gap: Best ideas from the world's teachers for improving education in the classroom. New York: Free Press.
Vamvakoussi, X., \& Vosniadou, S. (2010). Understanding the structure of the set of rational numbers: A conceptual change approach. Learning and Instruction, 14, 453- 467.

# Learning to Perceive Coherent Objects 

Nimrod Dorfman ${ }^{1}$ (nimrodd@weizmann.ac.il) Daniel Harari ${ }^{1}$ (danny.harari@weizmann.ac.il)<br>Shimon Ullman ${ }^{1}$ (shimon.ullman@weizmann.ac.il)<br>${ }^{1}$ Department of Computer Science, Weizmann Institute of Science, Rehovot, Israel


#### Abstract

Object segregation in a visual scene is a complex perceptual process that relies on the integration of multiple cues. The task is computationally challenging, and even the best performing models fall significantly short of human performance. Infants initially have a surprisingly impoverished set of segregation cues and their ability to perform object segregation in static images is severely limited. Major questions that arise are therefore how the rich set of useful cues is learned, and what initial capacities make this learning possible. Here we present a computational model that initially incorporates only two basic capacities known to exist at an early age: the grouping of image regions by common motion and the detection of motion discontinuities. The model then learns significant aspects of object segregation in static images in an entirely unsupervised manner by observing videos of objects in motion. Implications of the model to infant learning and to the future development of object segregation models are discussed.


Keywords: Visual perception; computational modeling; development; object segregation; figure-ground.

## Background and Goals

We naturally perceive the scene around us as containing coherent objects, separated from each other and from their background. Even in a complex image such as Figure 1A, we can count for example the number of distinct cars, delineate their boundaries, etc. The ability to segregate the scene into objects, delineate their boundaries, and determine occlusion relations (termed here 'object segregation'), relies on a complex set of processes, which integrate multiple cues that are only partially understood.

Infants' initial ability to segregate scenes into coherent objects is rudimentary and it does not make use of even basic salient 'Gestalt' properties such as uniformity of texture, brightness or color, the smooth continuity of boundary contours, occlusion cues and the like (Spelke et al. 1993). For instance, infants at 3 months of age do not appear to distinguish that the shape in Figure 1B is likely to be composed of two distinct components. The contrast between Figures 1A and 1B illustrates the span of learning accomplished in performing object segregation. The ability to segregate objects based on multiple cues develops quickly already in the first year of life, but the learning process continues over an extended period of time (Kovaks et al. 1999). The process of learning object segregation raises fundamental questions for cognitive development and computational modeling of vision. For cognitive development, it is of basic interest to understand the innate capacities and learning mechanisms that allow the system to


Figure 1: Object segregation, infant to adult capacity. (A): A complex scene, easily segregated by an adult. (B): At 3 months, infants do not appear to divide the figure into two components (after Spelke et al. 1993).
start from a surprisingly limited capacity for segregating the world into coherent objects, and reach the capability of the adult system. For computational modeling of vision, an intriguing possibility is to try to surpass the capabilities of current models by following a strategy similar to human development, namely, start with the appropriate set of basic capacities and learning mechanisms and allow the model to develop on its own the final segregation capabilities.

In the current study we focus on specific sub-problems within this broad domain. We develop a model that incorporates simple basic capacities, which are known empirically to already exist in young infants. It uses them to segregate familiar objects and to extract and use so-called 'boundary ownership' cues (indicating boundaries as well as figure/background direction) for static object segregation. The model initially has no ability to segregate objects in static images, but it can compute visual motion and motion discontinuities. It is exposed in an unsupervised manner to video sequences containing moving objects. It uses them to segregate familiar objects in static images and to learn local boundary ownership cues. These are used as cues for static object segregation, applied to novel objects.

In the next sections we briefly summarize relevant background from developmental studies of object segregation, followed by a presentation of the current model.

## Early Development of Object Segregation

Initial object segregation by infants is based almost exclusively on dynamic cues, which are then used to learn static object segregation. We focus below on two main
aspects of using visual motion for object segregation: grouping by common motion, and the use of motion discontinuities. We also comment briefly on the use of static cues.

Common Motion Infants use visual motion to group together adjacent regions that move together. These grouped image entities, discovered through motion, are also stored in memory and can subsequently be identified in static images (Needham 2001, Needham \& Baillargeon 1998, Needham \& Modi 1999, Spelke 1990, Spelke et al. 1989). For example, if 4.5 months old see in a static image a region A next to a second region B , their expectations are shaped by their recent experience of seeing these regions in motion. If $A$ and $B$ moved together, infants will treat them as a unit and will be surprised if they move separately, but not if they saw A or B moving alone. The grouping of regions into a single unit depends on their common motion: if two regions differ in their image motion, even if they remain in contact, they are treated as separate objects (Spelke 1990, Spelke et al. 1989). Retention in memory of the formed unit is limited in time (about 24 hours at 4.5 months of age), but grows gradually with age (Needham \& Baillargeon 1998, Needham \& Modi 1999). This use of stored object representations for segregation is termed 'object-based segregation', and it can generalize with more experience to other similar objects ('class based' segmentation), provided that the differences are initially small (Needham \& Modi, 1999). Two regions moving together can also be grouped together to form a single unit when they are non-contiguous but separated behind an occluder (Kellman \& Spelke 1983) provided that the parts are roughly aligned (Johnson \& Aslin 1996).

Motion Discontinuities In addition to region grouping based on common motion, infants are also sensitive from an early age ( 5 months or earlier) to dynamic cues created by the boundaries of moving objects (e.g., Granrud et al. 1984).

Static Cues In terms of static cues, at 3-5 months contiguous regions that are not separated by a visible gap tend to be grouped together, and are expected for example to move together rather than separately (Needham \& Baillargeon 1998, Spelke 1990, Spelke et al. 1989). At this age they show little or no evidence for using grouping principles based on uniformity of color, texture, and continuity of bounding contour in object perception. At 9 months the effect of such grouping cues is still weak (Spelke et al. 1993). The learning of static cues is gradual, and appears to depend on familiarity with many objects (Needham \& Modi 1989, Spelke 1990).

Following extended learning, perceptual organization into distinct objects and their boundaries develops into a complex process that relies on a rich set of cues. In addition to image-based, or bottom-up properties, organization into objects depends on top-down cues, based on familiarity with
specific objects and object classes. The different cues and their integration into a full segregation scheme are still a subject of active research in both human studies and computational modeling. Yearly competitions and evaluations of natural image segmentation ${ }^{1}$ show consistent improvements, but current performance is still significantly below human performance. Due to space limitations, we will not review here different modeling efforts. The closest to the current study is the SANE (segmentation according to natural examples) model by Ross et al. (2009), where, like in the current study, motion segmentation was used to guide static segmentation. However, the SANE model does not use the two main components of the current model: learning boundary-ownership cues near a boundary, and learning object-based segregation. It uses instead local binary $5 \times 5$ boundary elements, with no ownership information, and their pair-wise relationships.

## Goals of the Current Study

As reviewed above, infants are sensitive to motion cues for segregation, but lack sensitivity to most static cues for objects identity. It is therefore natural to ask how static segregation cues may be learned during development, guided by dynamic cues. We focus on two dynamic cues that are prominent in early infant perception. The first is common motion, guiding object-based segmentation. That is, infants naturally segregate adjacent image regions that share common motion, and can identify similar configurations in static images. One goal is therefore to model this learning of object-based segregation. Second, infants are sensitive to dynamic cues created by the boundaries of moving objects, and these are used by the model to learn useful static boundary cues. Although boundary ownership cues appear to play a major role in human object segregation (e.g. McDermott 2004, Ghose \& Palmer, 2010), they are not usually used in computational models, in part because it is still unclear which features are useful for assigning boundary ownership. A possible outcome of a model for the unsupervised learning of boundary ownership features could be, therefore, the extraction and use of such features in future segmentation models and algorithms.

## The Model

The current learning model has initially two 'innate' capacities for using visual motion to learn object segregation. The first is the capability to group together adjacent regions based on their common motion. A representation of the grouped shape is stored and can then be used for segregating similar shapes in novel static images. The second capacity is to extract motion discontinuities. These are used as teaching signals to extract image features located along object boundaries, together with a labeling of the figure/background sides, and

[^57]subsequently use them to locate novel object boundaries and identify the figure direction in new static images. These two components and how they are used by the learning model are described in subsequent sections, following a brief description of the training data used for learning.

## Training and Testing Data

Data consisted of 48 movies, each depicting an object (doll, banana, remote control etc.) moved by hand in front of a textured background ( 12 objects, 12 backgrounds). For each movie, there are 3 other movies showing the same object on a different background, 3 movies showing different objects on the same background; the remaining 41 have different object and a background. Each movie is one minute long ( 1500 frames), frame size varies between $520 \times 720$ pixels to $576 \times 752$ pixels.

## Object-based Segregation

The goal of object-based segregation is to learn the appearance of a specific object, such as the doll, fruit, etc., in our movies, and then find the full extent of the object and separate it from its background under new settings. The part of the model that deals with object-based segregation is based on an object detection model used, with some variations, in computer vision schemes, termed 'star model'. For the purpose of object segregation, the model is augmented with a 'back projection' stage. Since this part relies on existing object detection models it will be described here briefly.

The input to the object-based segregation is an image in a movie, together with the visual motion associated with the image. The scheme used for motion computation was an available optical flow algorithm (Sun et al. 2010) combined with background subtraction, assuming that the camera itself is stationary (as in Ross \& Kaelbling, 2009).

The motion computation divides the image into two components: a stationary one, and a set of one or more moving regions. One of the moving regions is selected for further processing. The selected region is covered by local image descriptors, each one representing the appearance of a local region. The implementation used the standard SIFT image descriptor (Lowe 2004) because computationally, it is robust and efficient, and biologically, it is similar to intermediate level units used in modeling (e.g. S2 units in the cortical H-Max model, Riesenhuber \& Poggio 1999). A single reference point C is selected at the center of the selected region, and for each image descriptor $F_{i}$, the displacement $\mathrm{V}_{\mathrm{i}}$ from its location to the center C is stored. The object defined by the moving region is therefore represented by its center C , and the set of image descriptors $\left(\mathrm{F}_{\mathrm{i}}\right)$, each one with its displacement $\mathrm{V}_{\mathrm{i}}$ from the object's center.


Figure 2: Examples of object-based segregations produced by the algorithm. Bottom right: an erroneous example.

Segregation of Static Images If the same or similar object appears in a new image, it can be detected and segregated based on the above representation, using the following algorithm. The new image is represented by its local SIFT descriptors. For each descriptor $F$ in the image, we find its $\mathrm{K}=25$ nearest neighbors among the descriptors of the stored object. Each neighbor $F_{k}$, votes for the location of the center C according to the displacement $\mathrm{V}_{\mathrm{k}}$. Votes are weighted by the similarity between F and $\mathrm{F}_{\mathrm{k}}$, and aggregated over the image. If an image location C obtains a sufficient number of total votes, an object is detected, centered at location C . The full object is then segregated by a 'back projection' step: all image descriptors that contributed their votes to the selected location are identified as components of the detected object. A final object/background decision is made by an automatically set threshold.

Results - Speed and Generalization In infants, even a few seconds of observing an object in motion already affects subsequent segregation of the same object in static images (e.g. Needham, \& Baillargeon 1998). The segmentation is effective for images of the same or similar object and generalizes gradually to less similar objects (Needham, \& Baillargeon 1998, Needham \& Modi 1999). Object-based segregation in the model showed similar characteristics. Brief ( 5 seconds) training was sufficient for learning object segregation of a specific object in subsequent parts of the movie, with some generalization to a different pose and different background. The object is often grouped by motion with the holding hand; the two can be separated when the hand is learned as an object on its own (Ullman et al. 2012). Figure 2 shows example segregations.


Figure 3: Detecting object boundaries. Left: Original image, with object-based segregation. Object is located, but boundaries are inaccurate. Center: Detection of boundary features. Warm colors indicate figure side of boundary, cold colors - ground side. Both object and background were not seen during training of boundary detector. Right: Combining object-based segregation with boundary detections. Object is detected with correct boundaries.

Results were tested by learning an object model in each movie using 5 sec and 40 sec segments, and testing on both later parts of the same movie, as well as the same object in other movies, with different backgrounds and larger variations in pose and lighting. Agreement between the true object (extracted by motion) and the model segregation were measured by the standard score $\mathrm{s}=|T \cap S| /|T \cup S|$, where T is the true object and S the segmented. Mean scores for 5 -sec training were $\mathrm{s}=0.3 \mathrm{vs} .0 .23$ on same vs. different movies, and for $40-\mathrm{sec}$ training $\mathrm{s}=0.49,0.36$ respectively. Effects of training time and generalization are highly significant (1-tailed t-test, $n=1200, p<10^{-6}$ in all comparisons).

The object-based segregation in the model segregates the general object region but it does not accurately delineate the boundaries. Since the object is represented by local appearance patches, it is sensitive to texture properties inside the object, in agreement with infant's object-based segregation (Needham \& Modi, 1999). In contrast, the model shows limited accuracy around object boundaries; it will be interesting to test this prediction in infants' vision (see discussion).

## Learning Boundary Features

The accurate delineation of boundaries is important for interacting with objects, e.g. for grabbing, finding free space to place them, etc. This is obtained in the model by a second mechanism, which uses motion discontinuities to learn static cues for occluding boundaries, as described next.

Learning Process To learn useful boundary features, motion discontinuities are used to guide the extraction of static boundary features and their figure-ground labeling. The learning procedure is simple, proceeding along the following stages. In each frame of the training movies, motion discontinuities are detected, and at each pixel along the boundary, image patches are extracted at 5 different sizes (ranging from $12 \times 12$ pixels to $60 \times 60$ pixels). Each patch is represented by a rotation invariant SIFT descriptor, producing a fixed-size descriptor regardless of original patch size. The motion signal is also used to label the figure part (which is moving in the training images) and background part (which is stationary) in each stored patch. From these, a subset of boundary patches is later selected, as described in the Results section below.

Use In Static Images The learned boundary features are then used to identify likely object boundaries in novel static images. Given a static input image, local SIFT features are extracted at the same 5 sizes, densely over the entire image. For each feature, its 25 nearest neighbors in the stored set of trained boundary features are extracted (using a fast approximation algorithm, Arya \& Mount 1993). These neighbors are used to estimate the likelihood of an object boundary at this location, and to identify the figure side of the potential boundary. Specifically, each neighbor i has a SIFT descriptor $D_{i}$ and an object direction $\theta_{i}$. For an image patch with descriptor $D$, we define the predicted object direction $\theta$ and a score $S$ as follows:


Figure 4: 25 examples of top-scoring boundary detection features, chosen by cross validation testing over 48 folds. Individual features are not reliable on their own - it takes at least 1,000 features to get good predictions (see text).

$$
\begin{gathered}
v=\sum_{i=1}^{25} e^{-\frac{\left\|D-D_{i}\right\|_{2}^{2}}{2 \sigma^{2}} v_{i},} \quad v_{i}=\left[\begin{array}{c}
\cos \theta_{i} \\
\sin \theta_{i}
\end{array}\right] \\
\theta=\operatorname{atan} 2(v), \quad S=\|v\|_{2}
\end{gathered}
$$

Where atan2 is a 4-quadrant arctangent function. $\sigma$ is set to 0.25 . $\theta, \mathrm{S}$ are then used to estimate the figure/background direction at the patch in question. Estimations of all patches are added together weighted by $S$ and smoothed by a spatial Gaussian function (positive in the figure, negative in background side). This yields a single total figure-score at each image location, where a positive score is likely to be the figure side of an object boundary.

Results Examples of boundary detection are shown in Figure 3. We used statistical testing to compare the density of boundary features in a region ( 10 pixels) around object boundaries compared with inside the object and on the background. Density was significantly higher around the boundaries compared with internal or external regions. In contrast, object-based segregation produced higher density in internal regions compared with boundary or external regions (1-tailed t-test, $n=1200 \mathrm{p}<10^{-6}$ in all comparisons).

Types And Number Of Boundary Features Psychophysical and computational studies of boundary features have suggested several types of informative boundary features, including: interposition (T-junctions), surface junctions, such as Y-junctions and arrow-junctions, and extremal edges, or folds, (Geisler et al. 2009, Ghose \& Palmer 2010) coming from the projection of an occluding edge curving smoothly in 3D, typically creating a highlight or shadow along the curving edge.

The current study used automatically labeled object boundaries, identified by motion discontinuities.


Figure 5: Object segmentation with the GrabCut algorithm. Left: Segmentation produced by the algorithm using default initialization. Right: Segmentation results with initialization by our segregation score maps.

Consequently, it became possible to extract and study a much richer set of boundary features compared with previous studies that used human annotated boundaries (Geisler et al. 2009, Fowlkes et al. 2007). The learning process produced a rich and varied set of boundary cues. Their analysis revealed the following properties. (i) Individual boundary features are probabilistic in the sense that they contribute information to the correct figure direction, but individual features are usually not definitive on their own. When training on 100,000 boundary features, the correct figure side is predicted in novel boundary features $78 \%$ of the times. (ii) Boundary features are consistent across image sets and are therefore useful for generalization to novel images. Our testing was done in 48 cross-validation folds, each time testing one movie, and excluding all movies with the same object or background from training data. (iii) There is a large set of useful boundary features, and using a restricted subset is less accurate than using the larger set. We selected the best performing features by cross-validation folds, and tested sets of different sizes, yielding $75 \%$ accuracy for 10,000 patches, $71 \%$ for $1,000,65 \%$ for $100,54 \%$ for 50 . Nonetheless, the improvement diminishes for very large sets, suggesting that saturation may be reached at some point, and there is no need to memorize every observed feature. Exploring mechanisms of feature retention is left for future work. (iv) Among the top-scoring boundary features (examples in Fig. 4) there is a significant fraction that can be labeled 'extremal edges'. These have only recently been found to play a crucial role in human vision (Ghose \& Palmer 2010), and have not been tested in infants' object segregation. Our model focuses on learning boundary features, and does not model their integration within a fully functional segregation system. To illustrate their contribution we therefore used them as input to an existing algorithm (GrabCut, Rother et al. 2004); results are
illustrated in Figure 5. The figure shows performance of the algorithm in its standard form (left), and the same algorithm when supplied with our object and boundary scores.

## Discussion

The model demonstrates how static object segregation can be learned effectively guided by two motion based mechanisms known to be innate or early learned in infants' vision: grouping by common motion and sensitivity to motion discontinuities.

These mechanisms are used by the model for two complementary goals: common motion is used for objectbased segregation, and motion discontinuities are used for learning static occlusion cues. In agreement with infants learning, the learning of object-based segregation by the model is fast, with initial sensitivity to details of the object's internal texture. It identifies well the region of the object with reduced accuracy near the boundaries. Boundary cues require more prolonged learning, but they appear to generalize broadly to novel object images. The set of useful boundary features found by the model is large and varied, including a major contribution from extremal edges, which have played a limited role in modeling so far.

The results of the study suggest a number of interesting directions for further research. In terms of infant studies, it will be of interest to test their capacity for object segregation based on extremal cues, which, to the best of our knowledge have not been tested so far. Another prediction that can be tested is whether object-based segregation by infants, which is sensitive to internal texture, will exhibit insensitivity to the object's boundary. Computationally, it will be interesting to compile a large set of useful boundary features that could be used by future segmentation algorithms. Finally, since scene segmentation in natural images is still a challenging open problem, it will be of interest to extend the current approach and examine whether following human development, by letting object segregation (including cues not considered in the current model) be guided and learned using dynamic cues, could lead to the emergence of models approaching human segregation capacities.

## Acknowledgments

The work was supported by European Research Council (ERC) Advanced Grant "Digital Baby" (to S.U.).

## References

Arya, S., \& Mount, DN (1993). Approximate nearest neighbor queries in fixed dimensions. ACM-SIAM, 271280.

Geisler, WS, Najemnik, J, Ing, AD. (2009). Optimal stimulus encoders for natural tasks. J. Vision 9(13), 1-16
Granrud, CE \& Yonas, A. (1984) Infants’ Perception of Pictorially Specified Interposition. J. Exp. Child Psych. 37, 500-511

Granrud CE, Yonas A, Smith IM, Arterberry ME, Glicksman ML, Sorknes AC. (1984) Infants' sensitivity to accretion and deletion of texture as information for depth at an edge. Child Dev. 55(4), 1630-1636.
Ghose, T, \& Palmer, S. (2010) Extremal edges versus other principles of figure-ground organization. J. Vision 10(8)3, 1-17
Johnson, SP \& Aslin, RN. (1996). Perception of Object Unity in Young Infants: The Roles of Motion, Depth, and Orientation. Cog. Dev. 11, 161-180.
Kellman, PJ, Spelke, ES, (1989). Perception of Partly Occluded Objects in Infancy. Cog. Psych. 15, 483-524.
Kovacs, I, Kozma, P, Feher, A, Benedek, G. (1999). Late Maturation of Visual Spatial Integration in Humans. PNAS 961(21), 12204-12209.
Lowe, D. (2004) Distinctive image features from scaleinvariant keypoints. Int. J. Comp. Vis., 60, 291-110.
McDermott, J. (2004) Psychophysics with junctions in real images. Perception, 33, 1101-1127.
Mount, D. \& Arya, S. (2010) ANN: A Library for Approximate Nearest Neighbor Searching. http://www.cs.umd.edu /~mount/ANN/
Needham, A,, \& Baillargeon, R. (1998). Effects of prior experience in 4.5 -month-old infants' object segregation. Inf. Behav. Dev. 21, 1-24.
Needham, A. (2001). Object Recognition and Object Segregation in 4.5-Month-Old Infants. J. Exp. Child Psychol, (78), 3-24.
Needham A, Modi A., (1999). Infants' use of prior experiences with objects in object segregation: implications for object recognition in infancy. Adv Child Dev Behav., 27:99-133.
Riesenhuber, M. \& Poggio, T. (1999), Hierarchical models of object recognition in cortex. Nat. Neurosci, 2(11), 1019-1025.
Ross, MG \& Kaelbling, LP. (2009). Segmentation According to Natural Examples: Learning Static Segmentation from Motion Segmentation. IEEE PAMI 31(4), 661-676.
Rother, C., Kolmogorov, V., Blake, A. (2004) GrabCut: Interactive Foreground Extraction using Iterated Graph Cuts. ACM Transactions on Graphics, 23 (3), 309-314.
Spelke, ES. (1990). Principles of object perception. Cognitive Science, 14, 29-56
Spelke, E.S, Breinlinger, K, Jacobson, K, Phillips, A. (1993) Gestalt relations and object perception: a developmental study. Perception, (22) 1483-1501.
Spelke, ES, von Hofsten, C, Kestenbaum, R. (1989). Object Perception in Infancy: Interaction of Spatial and Kinetic Information for Object Boundaries. Developmental Psychology, 25(2), 185-196.
Sun, D, Roth, S., Black, M.J. (2010). Secrets of optical flow estimation and their principles. IEEE CVPR, 2432-2439.
Ullman, S. Harari, D. Dorfman, N. 2012. From simple innate biases to complex visual concepts. PNAS, 109(44), 18215-18220.

# Removal of Information from Working Memory 

Ullrich K. H. Ecker (ullrich.ecker@uwa.edu.au)<br>School of Psychology, University of Western Australia

Stephan Lewandowsky (stephan.lewandowsky@bristol.ac.uk)<br>Department of Experimental Psychology, University of Bristol<br>School of Psychology, University of Western Australia

Klaus Oberauer (k.oberauer@psychologie.uzh.ch)<br>Department of Psychology, University of Zurich


#### Abstract

Standard working memory (WM) updating tasks confound updating requirements with generic WM functions. We introduce a method for isolating a process unique to WM updating, namely the removal of no-longer relevant information. In a modified version of an established updating paradigm, to-be-updated items were cued before the new memoranda were presented. Longer cue-stimulus intervals-that is, longer removal time-led to faster updating, showing that people can actively remove information from WM. Well-established effects of item repetition and similarity on updating RTs were diminished with longer removal time, arguably because representational overlap between out-dated and new information becomes less influential when out-dated information can be removed prior to new encoding. The benefit of removal time was found only for partial updating, not for complete updating of entire memory sets. We conclude that removal of out-dated information can be experimentally isolated, and that removal is a unique, active WM updating process.


Keywords: Working memory updating; Executive functions
Imagine you ask a colleague for his phone extension and he replies: "It's 3266 . No, hang on, in my new office it's 3257 ." Ideally, one should easily discard the last two digits of the outdated information (i.e., " 66 ") and replace them in working memory (WM) with the correct digits (i.e., " 57 "). However, this updating of WM content is no trivial task, and outdated information often continues to affect WM (Oberauer, 2001).

WM updating has been identified as one of three primary executive processes (Miyake, Friedman, Emerson, Witzki, \& Howerter, 2000). Updating has been claimed to be the only executive process to predict fluid intelligence (Friedman et al., 2006). However, most updating tasks used in previous research (e.g., Miyake et al., 2000) not only require WM updating but arguably also measure general WM abilities. This has led some to conclude that updating tasks constitute reliable assays of general WM capacity (Schmiedek, Hildebrandt, Lövdén, Wilhelm, \& Lindenberger, 2009).

This creates an unsatisfactory situation. If WM updating tasks measure just the same as other WM tasks such as complex span tasks, then why call them updating tasks? Both conceptually and theoretically, updating can be distinguished from maintenance and processing in WM. If updating is to be established as a non-redundant construct, it must be isolated and measured separately from other WM processes.

In a recent individual-differences study, we identified a processing component that was independent of general WM
capacity and unique to situations that demanded WM updating (Ecker, Lewandowsky, Oberauer, \& Chee, 2010). In that study we analyzed the processing components involved in widely used WM updating tasks, and we identified three separable components: retrieval, transformation, and substitution. Ecker et al. found that retrieval and transformation operations co-varied with general WM capacity, but that the substitution component did not. We thus argued that substitution is the only process that uniquely represents WM updating, without being "contaminated" by any association with WM. One implication of this analysis is that previous studies measuring WM updating did not separate variance unique to updating from the variance of generic WM processes.

In this article, we further decompose the components of WM updating. In Ecker et al. (2010), we suggested that information substitution can be further subdivided into the removal of outdated information and the encoding of new information. As encoding is a simple and generic operation involved in many cognitive tasks, we argue that it is the removal process that lies at the heart of memory updating. Accordingly, we focus on the removal of information from WM. Here we show that removal of outdated information can be separated experimentally from encoding of new information.

Our removal measure is based on the work of Kessler and Meiran (2008). In the updating paradigm they used, items (e.g., letters or digits) are presented in a set of individual frames. Items are then repeatedly updated by presenting new items in some frames. On each updating step, between one and $n$ items are updated, where $n$ is the memory set size (equal to the number of frames). Participants had to press a key at the end of each step to indicate that they finished updating.

Kessler and Meiran (2008) proposed a distinction between local and global updating. Local updating refers to changes made to individual items, whereas global updating refers to the integration of all items in the current memory set after individual items were changed. A key piece of evidence for this distinction comes from the observation (Experiment 3 in Kessler \& Meiran, 2008) that updating RTs increased with the number of to-be-updated items up to $n-1$ items, but updating was much faster again when all $n$ items were to be replaced on a given step. Thus, updating latencies depended in a nonmonotonic fashion on the number of to-be-updated items. Kessler and Meiran (2008) explained this non-monotonicity
by assuming that partial updates require a complex sequence of (1) unbinding of the integrated representation of the previous memory set, (2) substitution of some but not all items (i.e., the actual local updating), followed by (3) re-binding the new set as part of the global updating process. In contrast, when the entire set is updated, steps (1) and (2) can be omitted, the old set is simply discarded and a new memory set is encoded and globally updated.

Our interpretation of the non-monotonicity of updating latencies is a specific instantiation of the ideas of Kessler and Meiran (2008), motivated by a computational model of WM, SOB (Lewandowsky \& Farrell, 2008). SOB is a two-layer neural network in which items (represented in one layer) are associated to position markers (represented in the other layer) through Hebbian learning, which rapidly modifies the matrix of connection weights between the two layers. In the present updating paradigm, the position marker would represent the location of the item's frame on the screen. Forgetting in SOB is entirely based on interference; there is no time-based decay. To avoid overloading of the system in the absence of decay, an interference model requires a mechanism to remove outdated information; such a mechanism is implemented in the most recent version of SOB (see Oberauer, Lewandowsky, Farrell, Jarrold, \& Greaves, 2012). Removal of a specific item involves retrieving that item by cueing with its position marker, and "unlearning" the association between that item and its position. Unlearning is computationally implemented as Hebbian anti-learning. Thus, in SOB the idea of "unbinding" (cf. Kessler \& Meiran, 2008) refers specifically to the unbinding of selected items from their position markers. Both encoding and removal of individual items take time (cf. Oberauer, 2001). By contrast, wholesale removal of an entire memory set can be achieved by simply resetting the entire weight matrix, which we assume to be a very rapid process. This explains why updating the entire memory set is faster than partial updating.

In SOB, removal of old information and encoding of new information are described as two separate processing steps. It follows that updating should be facilitated if a cue about what information needs to be removed is given ahead of the to-be-encoded new information. In standard WM updating tasks-including the one used by Kessler and Meiran (2008)—removal can only begin when the new item(s) are presented. For example, when updating the telephone extension from from 3266 to 3257 , one can only begin removing the " 66 " when given the " 57 ". Hence updating times in such a task will include both time for removal and time for encoding. In our new updating task we present cues indicating which items are to be updated before presenting the new to-be-encoded stimuli. People can use the cue only to selectively remove old items from the memory set, not to encode new information. By varying the cue-stimulus interval, we vary the available time for removal. If longer available removal time leads to faster updating RTs, people must have used the cuestimulus interval for removal or unbinding. We tested this
idea in a series of four experiments.

## Experiment 1

In Ecker et al. (2010), we found that repeating (i.e., maintaining) an item during an updating task carries a benefit of nearly 400 ms . Experiment 1 tested the idea that this benefit should diminish when people are given the opportunity to remove outdated information before encoding the (identical) updated item.

## Method

Experiment 1 used a letter updating task in which each trial consisted of an encoding stage, an updating stage with multiple updating steps, and a final recall stage. Participants encoded 3 letters, presented simultaneously in individual frames. This was succeeded by an unpredictable number of updating steps; each updating step involved only a single, randomly selected letter. In most cases, the outdated letter was replaced with a new letter, but sometimes the letter repeated. Each update was cued before the new letter was presented (the respective frame turned red and bold). Presentation time of this cue-henceforth referred to as removal time-was either 200 ms or 1500 ms . The longer cue should be sufficient time for removal, whereas the shorter cue should be just enough time to focus attention on the to-be-updated frames without permitting removal. After each updating step, the frames were blanked for 500 ms or 1800 ms in the long and short removal-cue condition, respectively, ensuring equal retention intervals in both conditions. The experiment had a 2 (repetition no/yes) $\times 2$ (removal time: short/long) withinsubjects design.

Participants We tested 15 participants, mainly students from the University of Western Australia (UWA).

Apparatus \& Procedure The experiment was controlled by a MatLab program. There were 32 trials, each featuring on average 9 updating steps (the number of updating steps ranged from 1 to 21 ). The probability of item repetition during updating was set at $p=0.15$. We chose this low rate of repetition to ensure that removal of outdated information was still an attractive strategy. At each updating step, participants were required to press a key when they had finished updating (max. response time was 5 s ). This updating RT was the dependent variable of main interest. At the end of each trial, participants were prompted to recall all three letters in random order.

## Results

Updating response time data are shown in Figure 1.
A $2 \times 2$ repeated measures ANOVA on updating RTs with the factors repetition (no/yes) and removal time (short/long) yielded a main effect of repetition, $F(1,14)=15.23, M S E=$ $0.03, p<.01, \eta_{p}^{2}=0.52$, a main effect of removal time, $F(1,14)=9.78, M S E=0.01, p<.01, \eta_{p}^{2}=0.41$, and crucially, a significant interaction, $F(1,14)=8.22, M S E=0.01$, $p=.01, \eta_{p}^{2}=0.37$. The interaction indicates that the effect of
repetition on updating RTs is larger with short removal (265 ms ) than long removal time ( 86 ms ).


Figure 1: Updating response times from Experiment 1. Vertical bars denote within-subject standard errors of the mean.

## Discussion

Experiment 1 demonstrated that the benefit of item repetition during memory updating (Ecker et al., 2010) is strongly reduced if participants are given sufficient time to remove outdated information prior to the encoding of the updated information. In a sense, if there is no time for removal, a condition with repeating items does not require true memory updating-the established item representation in WM can be maintained, no information needs to be removed and substituted-hence the time advantage of repetition. In contrast, to the degree that an item is removed from WM, the time taken to encode a new item into that position will no longer depend on the identity of the removed item. This result pattern supports our notion that active removal of information is integral to WM updating under normal conditions (i.e., when there is no opportunity to "outsource" the removal process.

## Experiment 2

Experiment 2 had a similar rationale. Previous research (Lendinez, Pelegrina, \& Lechuga, 2011) had shown that updating numbers is quicker when the new to-be-remembered number is similar to the outdated number. Experiment 2 tested whether this benefit would diminish if participants were given sufficient time to remove the outdated number before encoding the new number.

## Method

The task in Experiment 2 was very similar to Experiment 1, but it used two-digit numbers. During updating, about half the updates used similar and dissimilar numbers, respectively. This was achieved by manipulating both the proximity of numbers (proximal/distant) and the repetition
of one of the digits (yes/no). This resulted in four updating conditions, which had different probabilities of occurrence: proximal/repeating (e.g., updating from 18 to 19 ; $p=.15$ ), proximal/non-repeating (e.g., updating from 20 to $18 ; p=.15$ ), distant/repeating (e.g., updating from 59 to 19 ; $p=.15$ ), and distant/non-repeating (e.g., updating from 18 to 59; $p=.50$ ). Proximal updates ranged from -3 to +3 (excluding zero); distant/repeating updates were constrained to multiples of 10 , and distant/non-repeating updates used prime numbers from 13 to 83 . Filler updates with intermediate proximity ( $\pm 4-8$ ) were used with $p=.05$. Positive and negative updates were randomly intermixed (this was not considered an experimental factor). Removal time was again an additional factor (short/long).

Participants We tested 27 UWA students.
Apparatus \& Procedure The apparatus and procedure was identical to Experiment 1, with the exception that the experiments had 36 trials.

## Results

Updating response time data are shown in Figure 2.


Figure 2: Updating response times from Experiment 2. Vertical bars denote within-subject standard errors of the mean.

For the sake of simplicity, we collapsed the three 'similar' conditions into one, and ran a $2 \times 2$ repeated measures ANOVA with the factors similarity (low/high) and removal time (short/long). We found reliable main effects of similarity, $F(1,26)=11.38, M S E=.01, p<.01, \eta_{p}^{2}=0.30$, and removal time, $F(1,26)=101.43, M S E=.02, p<.001, \eta_{p}^{2}=$ 0.80 . Most importantly, these main effects were qualified by a significant interaction: the similarity effect was larger with short removal time ( 116 ms ) than long removal time ( 38 ms ), $F(1,26)=6.04, M S E=.01, p=.02, \eta_{p}^{2}=0.19$.

## Discussion

Experiment 2 demonstrated that a well-documented similarity effect in memory updating is diminished when participants
are given time to remove a to-be-updated item from memory before encoding the new item. This supports our notion of removal: We assume that the similarity effect in memory updating arises because of representational overlap between the replaced and the new item. That is, two similar numbers share a digit and/or a region in number space, and to the degree that only new features are substituted, not entire item representations, this similarity will facilitate updating. Yet, the more an item representation is removed before the updated number can be encoded, the less facilitation there will be.

Having established some support for our notion of removal, we now turn to a test of our prediction that an active removal process is only utilized during partial updates, not global updates (which unlike partial updates can be achieved more efficiently by 'wiping' memory, or in SOB terms, by resetting the weight matrix).

## Experiment 3

Based on the distinction between slow selective removal and fast resetting of the weight matrix, we predicted that longer removal times should lead to a substantial time gain in partial updating, but little gain on updating steps replacing the entire memory set.

In Experiment 3, the letter updating task was modified such that each updating step could update the memory set either partially or entirely (i.e., 1-, 2-, or 3-frame update).

## Methods

Participants Sixty-nine UWA undergraduates participated.
Apparatus and Procedure Apparatus and procedure were identical to previous experiments except that on each step, new letter(s) were presented in 1,2 , or 3 frames, and that there were 28 trials in total.

## Results

Updating response time data are shown in Figure 3.
A two-way repeated measures ANOVA on updating response times yielded a significant main effect of the number of updated frames, $F(2,136)=64.30, M S E=0.03, p<$ .001, $\eta_{p}^{2}=0.49$, a significant main effect of removal time, $F(1,68)=281.68, M S E=0.02, p<.001, \eta_{p}^{2}=0.81$, as well as a significant interaction, $F(2,136)=109.55, M S E=0.01$, $p<.001, \eta_{p}^{2}=0.62$. Planned contrasts showed that on average it took significantly longer to update two frames compared to one ( 1.44 seconds vs. 1.26 seconds; $F(1,68)=$ $146.29, M S E=0.02, p<.001$ ) and that with 1 or 2 frames, updating took significantly longer with short as compared to long removal time ( 1.53 vs. 1.17 seconds; $F(1,68)=320.14$, $M S E=0.03, p<.001$. However, updating three frames was relatively quick ( 1.23 seconds), and removal time had a negligible impact on updating time when all three frames were updated. While the removal time effect was statistically significant (arguably due to the large sample size and power; 1.26 vs. 1.20 seconds; $F(1,68)=13.59, M S E=0.01, p<.001)$,
it was much smaller with three frames than with one or two frames ( 53 vs .362 ms ).


Figure 3: Updating response times from Experiment 3. Vertical bars denote within-subject standard errors of the mean.

## Discussion

The results of Experiment 3 support our hypothesis that partial updating of a memory set involves a process of active removal. "Bringing forward" this removal process by cueing the to-be-updated frames prior to presentation of the new items sped up partial updating substantially. This time saving of roughly $300-400 \mathrm{~ms}$ can be interpreted as the time that is required to remove information from WM.

In contrast, the opportunity to remove items from memory before presentation of new memoranda brought no substantial advantage when the entire memory set was updated. This finding supports our notion that memory can be cleared almost instantly, and that updating of an entire memory set does not require the time-consuming selective removal process.

The speed-up induced by the long removal cue did not increase with the number of to-be-updated items. This unexpected observation seems to imply that removing one item from WM takes as long as removing two items. There are two possible explanations. One is that removal of multiple items can occur in parallel. While this is a theoretical possibility, it is at odds with SOB's notion that items must be retrieved individually to be removed by anti-learning. ${ }^{1}$

The second explanation is that people use the removal time only to remove one item, even when two items are about to be updated. This might be an efficient strategy because itemspecific removal is likely to require the focus of attention, and switching the focus of attention to a new item takes time (cf. Garavan, 1998). Thus, the most efficient use of removal time

[^58]might be to focus on the first to-be-removed item and remove it, then wait. As soon as the new item(s) are presented, one of them can immediately be encoded in the currently focused frame. Only then would the focus move on to the second to-be-updated item (if there is one).

This interpretation is in line with a recent proposal by Kessler and Oberauer (2013), namely that partial updating also involves task-switching, which participants may likewise try to avoid. This notion assumes that, without substantial pre-cued removal time, participants scan the items from the beginning of the list to the end, starting in a maintenance mode (M). Hence, if the first item on the list is not updated, people might refresh that representation, then move to the next item. As soon as they encounter an item that requires updating, they switch to updating mode and actively remove that item from WM (U). Updating frames 1 and 3 would hence require 3 switches, (M)UMU; updating frames 1 and 2 two switches, (M)UUM; updating frames 2 and 3 one switch, (M)MUU. In the case of 1 -frame updates, updating frame 1, (M)UMM, and updating frame 2, (M)MUM, both require two switches, but updating frame 3 only requires one switch, (M)MMU. A strategy of scanning the list up to the first to-be-updated item, switching to updating mode, removing the item, and waiting would hence minimize both focusswitch and task-switch costs.

We applied multi-level regression analysis to the data of Experiment 3 to confirm the importance of task switching as specified above. We coded processes required at each updating step with the following parameters: The number of items to encode ( $\mathrm{E} ; 1-3$ ), the number of items to remove ( $\mathrm{R} ; 0-2$ ), and the number of task switches (SW; 0-3). Additional parameters were introduced in the modeling, including a wipe (W) parameter (coding the discarding of an entire memory set), and a refresh (RF) parameter (coding the number of non-updated, to-be-refreshed items). Importantly, coding differed for short and long removal time conditions. For example, with long removal time a two-frame update of frames 2 and 3 would involve no refreshing or switching: Participants could use the removal time to refresh the first item, switch to updating mode for frames 2 and 3 , remove the letter in frame 2, then wait to encode the new letter in frame 2 , and remove the old and encode the new letter in frame 3 ( $R F=S W=W=0 ; R=1 ; E=2$ ). In contrast, if removal time is short, participants would have to refresh the first item, switch once, remove two items and encode the two replacement items ( $W=0 ; R F=S W=1 ; R=E=2$ ).

A model including both removal and task switch parameters achieved the best fit; UpdatingRT $=704+156 * E+$ $87 * R+393 * S W-95 E * S W$, with a coefficient of determination $C O D=0.437$ and a Bayes Information Criterion $B I C=23403$. (BIC of the best-fitting model without removal parameter was 23475; according to Raftery (1996) this provides very strong evidence in favor of the removal model).

## Experiment 4

Experiment 3 suggested that people only remove one item even with long removal time. We argued this is in fact efficient behavior as it minimizes focus and task switch costs. Yet, an alternative hypothesis is that people would have removed more information had they had more time. Experiment 4 tested the idea that with sufficient removal time, people might remove more than one item, and thus show shorter updating RTs in conditions that would benefit from the removal of more than one item.

## Methods

Participants We tested 34 UWA undergraduates.
Apparatus and Procedure Apparatus and procedure were identical to Experiment 3 except for the addition of a third, very-long removal time condition (i.e., removal time conditions of 200,1500 , and 3000 ms ), and the omission of the full update condition (i.e., either 1 or 2 frames were updated with each updating step). There were 15 trials in total.

## Results

Updating response time data are shown in Figure 4.


Figure 4: Updating response times from Experiment 4. Vertical bars denote within-subject standard errors of the mean.

A $2 \times 3$ repeated measures ANOVA on updating RTs with the factors number of updated frames (1 vs. 2) and removal time (short/long/very long) yielded a main effect of frame number, $F(1,33)=71.66, M S E=0.03, p<.001$, $\eta_{p}^{2}=0.68$, a main effect of removal time, $F(2,66)=167.88$, $M S E=0.02, p<.001, \eta_{p}^{2}=0.84$, but no significant interaction, $F(2,66)=2.06, M S E=0.02, p>.10, \eta_{p}^{2}=0.06$. Results showed that it took longer to update two frames than one, and that more removal time led to faster updating RTs. However, the effect of removal time was strong when comparing the short ( 200 ms ) and long ( 1500 ms ) conditions, but negligible when comparing long and very-long ( 3000 ms ) conditions. The lack of interaction means that doubling re-
moval time did not lead to quicker updating, not even in the 2frame updates where performance could have benefitted from the removal of both to-be-updated items.

Applying the regression models (as specified in Exp. 3), we found that a simple model provided the best fit: UpdatingRT $=901+131 * R+255 * S W$, with $C O D=$ $0.38, B I C=6865$. (Note that the more complex model specified in Exp. 3 did explain more variance than this simple model when fit to the data of Exp. 4, but had a higher overall BIC. The BIC for the best removal-free model was 6880; this is strong evidence in favor of the simple removal model.)

## Discussion

Experiment 4 showed that in the present task people only removed one item in anticipation of an update, even when this update concerned more than one item. This supports our notion that people avoid focus and task switching when removing information from WM during updating.

## General Discussion

In this article we have introduced a novel measure of WM updating. Traditional WM updating tasks arguably measure general WM processes in addition to updating, whereas it is the removal of information from WM that is specific and unique to WM updating. We demonstrated that giving people preparation time to remove information from WM speeds up updating when new information is subsequently presented, but only when a subset of the memory set is updated. Updating an entire memory set does not benefit (much) from preparation time, arguably because-in line with the predictions derived from SOB (Lewandowsky \& Farrell, 2008; Oberauer et al., 2012)-the time-consuming removal process only applies to individual items, whereas an entire memory set can be removed by instant resetting of the weight matrix. Our notion of removal by unlearning item-position associations is a specific incarnation of the more general idea advanced by Kessler and Meiran (2008), who suggested that partial updating of memory sets involves "dismantling" or "unbinding" the old representations.

Whereas our research is guided by and supports the SOB model, it is important to note that other researchers have provided independent evidence for the existence of an active and attention-demanding removal process. For example, Fawcett and Taylor (2012) have shown that directed forgetting of an item (1) slows down responses on an unrelated secondary task for up to 2.6 seconds, and (2) impairs incidental memory for a subsequent distractor, in particular when the directed forgetting of the studied item is successful.

We assume removal to be crucial to maintaining a functional WM system that can efficiently focus on relevant information. Further research is needed to ascertain whether removal abilities co-vary with WM capacity, whether they predict intelligence-related variance, or whether they relate to other executive functions.

## Acknowledgments

This research was supported by Australian Research Council Discovery Grants and Research Fellowships to UE and SL.

## References

Ecker, U. K. H., Lewandowsky, S., Oberauer, K., \& Chee, A. E. H. (2010). The components of working memory updating: An experimental decomposition and individual differences. Journal of Experimental Psychology: Learning, Memory, and Cognition, 36, 170-189.
Fawcett, J. M., \& Taylor, T. L. (2012). The control of working memory resources in intentional forgetting: Evidence from incidental probe word recognition. Acta Psychologica, 139, 84-90.
Friedman, N. P., Miyake, A., Corley, R. P., Young, S. E., DeFries, J. C., \& Hewitt, J. K. (2006). Not all executive functions are related to intelligence. Psychological Science, 17, 172-179.
Garavan, H. (1998). Serial attention within working memory. Memory \& Cognition, 26, 263-276.
Kessler, Y., \& Meiran, N. (2008). Two dissociable updating processes in working memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 34, 13391348.

Kessler, Y., \& Oberauer, K. (2013). Working memory updating latency reflects the cost of switching between maintenance and updating modes of operation.
Lendinez, C., Pelegrina, S., \& Lechuga, T. (2011). The distance effect in numerical memory-updating tasks. Memory and Cognition, 39(4), 675-685.
Lewandowsky, S., \& Farrell, S. (2008). Short-term memory: New data and a model. In B. H. Ross (Ed.), The psychology of learning and motivation (Vol. 49, pp. 1-48). London, UK: Elsevier.
Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., \& Howerter, T. D., A.and Wager. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. Cognitive Psychology, 41, 49-100.
Oberauer, K. (2001). Removing irrelevant information from working memory: A cognitive aging study with the modified Sternberg task. Journal of Experimental Psychology: Learning, Memory, and Cognition, 27, 948-957.
Oberauer, K., Lewandowsky, S., Farrell, S., Jarrold, C., \& Greaves, M. (2012). Modeling working memory: An interference model of complex span. Psychonomic Bulletin \& Review.
Raftery, A. E. (1996). Approximate bayes factors and accounting for model uncertainty in generalised linear models. Biometrica, 83(2), 251-266.
Schmiedek, F., Hildebrandt, A., Lövdén, M., Wilhelm, O., \& Lindenberger, U. (2009). Complex span versus updating tasks of working memory: The gap is not that deep. Journal of Experimental Psychology: Learning, Memory, and Cognition, 35, 1089-1096.

# Effects of Explanation and Comparison on Category Learning 

Brian J. Edwards (Brian.Edwards@U.Northwestern.Edu)<br>Department of Psychology, Northwestern University, 2029 Sheridan Road-102 Swift Hall<br>Evanston, IL 60208 USA

Joseph J. Williams (Joseph_Williams@Berkeley.Edu), Tania Lombrozo (Lombrozo@Berkeley.edu)<br>Department of Psychology, University of California, Berkeley, 3210 Tolman Hall<br>Berkeley, CA 94720 USA


#### Abstract

Generating explanations and making comparisons have both been shown to improve learning. While each process has been studied individually, the relationship between explanation and comparison is not well understood. Three experiments evaluated the effectiveness of explanation and comparison prompts in learning novel categories. In Experiment 1, participants explained items' category membership, performed pairwise comparisons between items (listed similarities and differences), did both, or did a control task. The explanation task increased the discovery of rules underlying category membership; however, the comparison task decreased rule discovery. Experiments 2 and 3 showed that (1) comparing all four category exemplars was more effective than either within-category or between-category pairwise comparisons, and that (2) "explain" participants reported higher levels of both spontaneous explanation and comparison than "compare" participants. This work provides insights into when explanation and comparison are most effective, and how these processes can work together to maximize learning.


Keywords: Explanation; comparison; categorization; learning.

## Introduction

Explanation (i.e., answering "why" questions) and comparison (i.e., describing the similarities and differences between entities) are both powerful learning processes. Although they have typically been studied independently, they are often interconnected. Asking people to generate explanations can invite implicit comparison, and the patterns that people discover by comparing can motivate a search for explanations. For example, explaining why someone prefers coffee versus tea might lead one to identify similarities and differences between the two beverages, and comparing coffee and tea might provide insights into why a person would prefer one over the other. Explanation and comparison can also support similar ends: both promote abstraction and generalization, and both facilitate the discovery of patterns that are deep in a system's underlying structure (for reviews, see Gentner, 2010, on analogy and comparison; Lombrozo, 2012, on explanation).

Although explanation and comparison can generate similar effects, these two processes might rely on different cognitive mechanisms and exert different constraints on learning. Explanation has been hypothesized to improve learning through a variety of mechanisms, including an
increase in metacognitive awareness (Chi, 2010) and an increase in attention and engagement (e.g., Siegler, 2002), among others. In the context of category learning, generating explanations also enables learners to generalize beyond a specific set of observed data. In particular, Williams and Lombrozo (2010, 2013) proposed a subsumptive constraints account of how explanation impacts learning, whereby explaining leads people to interpret individual cases as part of a general pattern. As a result, explanation can help people unify multiple observations and focus on patterns with broader scope, increasing the discovery of rules that account for $100 \%$ of the data versus only $75 \%$ (Williams \& Lombrozo, 2010).

One mechanism by which comparison has been hypothesized to support learning is by promoting explicit structural alignment, leading people to focus on alignable differences between two entities (i.e., differences that are embedded in a common relational structure) (Gentner, 1983; Gentner \& Markman, 1997). Since comparison causes people to analyze these differences in the context of the common structure, comparison can illuminate deeper similarities and support the formation of an abstract relational schema, even (and especially) when the items being compared have surface differences (Gentner et al., 2009). For example, the analogy "an atom is like a solar system" highlights the fact that an atom consists of electrons orbiting around a nucleus, whereas a solar system consists of planets orbiting around the sun. Across a number of domains, comparing two examples that are superficially dissimilar but share a common relational structure supports transfer more effectively than studying the same examples separately (e.g., Kurtz, Miao, \& Gentner, 2001; Loewenstein, Thompson, \& Gentner, 2003).

Despite the abundance of research showing that explanation and comparison can (individually) enhance learning, few studies have investigated the effects of both explanation and comparison on the same experimental task. Kurtz, Miao, and Gentner (2001) found that comparing two analogous examples of heat flow helped participants discover similarities between the two examples more effectively than describing and explaining the same examples sequentially. Additionally, comparison was most effective when participants performed a task that involved listing which elements of the second scenario corresponded to specific elements of the first scenario. In another study, Nokes-Malach et al. (2012) found that introductory physics
students who explained the solutions to worked examples of physics problems achieved greater "near" transfer than participants who compared pairs of problems, but both groups performed similarly on "far" transfer and outperformed participants in a control condition. While these studies provide valuable insights into the conditions under which explanation and comparison are most effective, many questions remain open.

The present studies examine whether and how explanation and comparison interact to support learning novel categories. Previous work using similar materials (alien robots) has found that relative to control conditions, participants prompted to explain why individual robots belong to particular categories are more likely to discover a categorization rule that accounts for all cases (Williams \& Lombrozo, 2010, 2013). The present studies extend this work by investigating whether having participants compare robots also facilitates category learning, and whether participants who perform both explanation and comparison tasks are more likely to discover categorization rules than participants who perform only one of these tasks.

We hypothesize that comparison and explanation play complementary roles in category learning. Comparison may be crucial for identifying similarities among members of the same category and differences between members of different categories. In contrast, explanation should encourage learners to seek broad patterns within and across categories, potentially drawing upon the similarities and differences identified through comparison. Very broadly, these hypotheses predict that participants should be more effective in discovering categorization rules to the extent that they both compare and explain, with explanation being especially important in discovering broad patterns.

Three experiments evaluated these predictions. In Experiment 1, participants were randomly assigned to study the robots in one of four ways: (1) explain why individual robots are members of a particular category, (2) compare pairs of robots that belong to the same category, (3) perform both the explanation and comparison tasks, or (4) engage in a "free study" control task. Experiments 2 and 3 evaluated the effectiveness of different types of comparison prompts: between-category pairwise comparison and "group" comparison, respectively. We included a "group" comparison prompt to see whether it would be more effective at improving participants' ability to integrate pairwise comparisons and detect broad patterns.

## Experiment 1

## Method

Participants One-hundred-sixty-one adults participated through the Amazon Mechanical Turk marketplace. An additional 56 participants were tested, but excluded because they failed a catch trial or had previously completed a similar experiment. Participants were paid for participation.

Materials The stimuli (see Fig. 1) were eight robots adapted from Williams and Lombrozo (2010, 2013). Four robots (AD) were classified as Glorp robots and the other four robots (E-H) were classified as Drent robots.


Figure 1: Robots used in Exp. 1-3
Four rules could be used to categorize robots as either Glorp robots or Drent robots. Two rules were " $100 \%$ rules" that could be used to categorize all eight robots and two rules were " $75 \%$ rules" that could be used to categorize six of the eight robots (i.e., two robots were anomalous with respect to each $75 \%$ rule). The four rules were as follows:
(1) Foot rule (100\%): All Glorp robots have feet with pointy bottoms; all Drent robots have feet with flat bottoms.
(2) Antenna rule ( $100 \%$ ): All Glorp robots have a right antenna (from the robot's perspective) that is longer than the left antenna; all Drent robots have a left antenna that is longer than the right antenna.
(3) Elbows/knees rule (75\%): Three out of four Glorp robots (A, B, D) have elbows but no knees; three out of four Drent robots (F, G, H) have knees but no elbows. One Glorp robot (C) has knees but no elbows and one Drent robot (E) has elbows but no knees.
(4) Body shape rule (75\%): Three out of four Glorp robots (A, B, C) have a rectangular body; three out of four Drent robots (E, F, H) have a round body. One Glorp robot (D) has a round body and one Drent robot (G) has a rectangular body.

The robots also differed in body color; however, there were no systematic category differences in body color.

Procedure The procedure consisted of a study phase followed by a rule-reporting phase.

In the study phase, each participant was assigned to one of four study conditions: (1) comparison only, (2) explanation only, (3) both explanation and comparison, or (4) free study. In every condition, all eight robots appeared on screen for the duration of the study phase, as shown in Figure 1. The total study time ( 640 seconds) was equal across conditions. The study prompts and procedures for each condition were as follows.

Comparison only condition: "What are the similarities and differences between Glorp [Drent] robot X and Glorp [Drent] robot Y?" Participants were given 160 seconds to perform each comparison. The comparisons were presented
in the following order: A and $\mathrm{B}, \mathrm{F}$ and $\mathrm{H}, \mathrm{C}$ and $\mathrm{D}, \mathrm{E}$ and G . This order was chosen so that the four robots that were consistent with respect to both $75 \%$ rules were studied before the four robots that were anomalous with respect to one of those rules, making it more likely that participants would learn the $75 \%$ rules in addition to the $100 \%$ rules.

Explanation only condition: "Try to explain why robot X is a Glorp [Drent] robot." Participants were given 80 seconds to provide an explanation. The explanations were requested in the following order: $\mathrm{A}, \mathrm{B}, \mathrm{F}, \mathrm{H}, \mathrm{C}, \mathrm{D}, \mathrm{E}, \mathrm{G}$. This matched the order in the comparison condition.

Both explanation and comparison condition: Participants responded to both the explanation and comparison prompts above. To ensure that all the conditions were matched for study time, participants were given 40 seconds to respond to each explanation prompt and 80 seconds to respond to each comparison prompt. The order of the explanation and comparison prompts was counterbalanced across participants. Participants performed both tasks sequentially for each pair (e.g., explain A, explain B, compare A and B) before moving on to study the next pair of robots. The study order was otherwise the same as in the other conditions.

Free study condition: "Write out your thoughts below as you learn to categorize Glorp [Drent] robot X." Participants were given 80 seconds to study each robot. The study order was the same as in the other conditions.

At the end of each study period, the screen automatically advanced to the next robot or pair of robots. Participants could not advance before the study period had elapsed.

After each 160 seconds, participants solved a simple math exercise (e.g., " $9+7$ "). These exercises were included as a "catch trial" to verify that participants' attention was not diverted to other tasks. Response time was recorded and participants who took more than one minute to answer a question were excluded from analysis.

In the rule-reporting phase, participants listed the patterns they noticed "that might help differentiate Glorps and Drents." These responses were classified by a coder who was blind to experimental condition. Twenty-five percent of the data was independently coded for reliability by a second blind coder; agreement for each experiment exceeded $95 \%$. For each pattern that participants discovered, they also indicated (1) how many of the eight study robots could be categorized using that pattern and (2) how many new Glorp and Drent robots (out of 100) could be categorized using that pattern. Because answers to these two questions were contingent on the participant having discovered a particular rule, the sample sizes were relatively small and these data are not discussed further.

After completing the rule-reporting phase, participants answered debriefing questions regarding the extent to which they (1) generated explanations and (2) made comparisons, regardless of the task instructions, using a numerical response on a 1-7 scale, where 1 indicated "not at all" and 7 indicated "all of the time." Participants were then asked whether they had previously completed a similar study and answered a "catch trial" adapted from Oppenheimer,

Meyvisb, and Davidenkoc (2009) to find out whether they were reading the instructions. Participants who reported previously doing a similar study and participants who failed the catch trial were excluded from analysis.

## Results and Discussion

We first considered whether study task influenced the total number of rules discovered. A $2 \times 2$ ANOVA with explain prompt (yes/no) and compare prompt (yes/no) as betweensubjects factors and total number of rules discovered (0-4) as a dependent measure revealed no effects of condition ( $p$ s $>.15)$. We thus considered whether discovery of the $100 \%$ and $75 \%$ rules varied across study conditions (see Fig. 2).

A log-linear analysis of explain prompt (yes $/ n o$ ) $\times$ compare prompt (yes/no) $\times$ discovered a $100 \%$ rule (yes/no) revealed that performing the explanation task made participants significantly more likely to discover at least one of the two $100 \%$ rules, $\chi^{2}(1)=21.4, p<.001$. Performing the comparison task had the opposite effect: participants were less likely to discover a $100 \%$ rule, $\chi^{2}(1)=5.90, p=$ .015. There was no significant interaction ( $p=.67$ ). A comparable analysis on discovery of a $75 \%$ rule (yes/no) found that performing the explanation task made participants less likely to report a $75 \%$ rule, $\chi^{2}(1)=11.3, p$ $<.001$, with no effect of the comparison task, $p=.75$.


Figure 2: Rule discovery by study condition in Exp. 1, showing the percent of participants discovering at least one rule of each type.

These findings challenge our predictions in that a comparison prompt actually impaired $100 \%$ rule discovery, and that explanation and comparison did not have additive benefits. The findings do support the idea that explanation and comparison exert distinct constraints on learning, but raise an important puzzle: why didn't comparison - which has been shown to have robust and beneficial effects in other domains - improve performance on this task? We analyzed participants' self-reported explanation and comparison to better understand why the comparison task impaired performance, and in particular, whether the study prompts were effective at promoting explanation and comparison processes as intended.

A $2 \times 2$ ANOVA with explanation task (yes/no) and comparison task (yes/no) as between-subjects factors and
amount of self-reported explanation as the dependent variable showed that participants who performed the explanation task reported more explanation $(M=5.83, S D=$ 1.46) than participants who did not $(M=4.36, S D=2.09)$, $F(1,154)=27.7, p<.001$. Additionally, participants who performed the comparison task reported doing less explanation $(M=4.75, S D=2.06)$ than participants who did not $(M=5.49, S D=1.73), F(1,154)=7.26, p=.008$. Selfreported explanation was positively correlated with the number of $100 \%$ rules discovered, $r=.34, p<.001$.

A $2 \times 2$ ANOVA with explanation task (yes/no) and comparison task (yes/no) as between-subjects factors and amount of self-reported comparison as the dependent variable showed that participants who performed the explanation task reported doing more comparison ( $M=$ 5.62, $S D=1.64$ ) than participants who did not $(M=4.69$, $S D=2.05), F(1,155)=10.2, p=.002$. However, performing the comparison task did not affect the amount of reported comparison (Comparison: $M=5.13, S D=1.92$; No comparison: $M=5.21, S D=1.89$ ). Self-reported comparison was positively correlated with the number of $100 \%$ rules discovered, $r=.22, p=.006$, but the effect was not significant after controlling for reported explanation.

Two factors might help explain why the comparison task did not support discovery of the $100 \%$ rules. First, the comparison prompt failed to boost overall comparison (as reflected in self-reports), and additionally decreased selfreported explanation, which was beneficial to learning. Second, the comparison prompt may have constrained the particular types of comparisons that participants performed in unhelpful ways, restricting them to within-category, pairwise comparisons at the expense of between-category comparisons or category-wide comparisons. In particular, previous work has shown that between-category pairwise comparison can be more effective than within-category pairwise comparison for learning feature-based categories (Higgins \& Ross, 2011). The subsequent experiments evaluated these hypotheses by investigating whether between-category comparison (Experiment 2) or "group" comparison (Experiment 3) would support greater rule discovery than within-category pairwise comparison.

## Experiment 2

## Method

Participants One-hundred-sixty-one adults participated in the study through the Amazon Mechanical Turk marketplace. An additional 54 participants were tested, but were excluded because they failed a catch trial or because they had previously completed a similar experiment. Participants were paid for their participation.

Materials The stimuli were those in Experiment 1.
Procedure As in Experiment 1, the procedure consisted of a study phase followed by a rule-reporting phase.

The study phase was identical to Experiment 1 with the following changes. First, the total study time was reduced from 640 seconds to 360 seconds, with the time allotted for each study prompt reduced proportionally. Second, each participant was assigned to one of four study conditions: (1) the Experiment 1 explanation task, (2) the Experiment 1 within-category pairwise comparison task, (3) a betweencategory pairwise comparison task, or (4) an explanation task in which participants alternated explaining Glorp and Drent robots. Conditions (3) and (4) are described below.

Between-category pairwise comparison task: "What are the similarities and differences between Glorp robot X and Drent robot Y?" The comparisons were performed in the following order: A and $\mathrm{H}, \mathrm{B}$ and $\mathrm{F}, \mathrm{C}$ and $\mathrm{G}, \mathrm{D}$ and E .

Between-category explanation task: This task was identical to the Experiment 1 explanation task except that the robot study order matched the between-category pairwise comparison task.

The rule-reporting phase was identical to Experiment 1. After the rule-reporting phase, but before the debriefing questions, participants completed a recognition memory task. However, performance was very poor and did not differ across conditions; this task is not discussed further.

After completing the memory task, participants answered debriefing questions regarding the extent to which they (1) generated explanations, (2) made within-category comparisons, (3) made between-category comparisons, and (4) described the features of individual robots, all regardless of the task instructions. As in Experiment 1, participants were asked if they had previously completed a similar experiment and answered a "catch trial" question.

## Results and Discussion

We first analyzed the total number of rules discovered (0-4) in a $2 \times 2$ ANOVA with study task (explain/compare) and study order (between/within) as between-subjects factors. The explanation task resulted in a marginal increase in the total number of rules discovered, $F(1,157)=3.62, p=.059$.


Figure 3: Rule Discovery by Condition in Exp. 2

A log-linear analysis of study task (explain/compare) $\times$ study order (between/within) $\times$ discovered a $100 \%$ rule (yes/no) found that participants who performed the explanation task were more likely to discover a $100 \%$ rule
than participants who performed the comparison task, $\chi^{2}(1)$ $=10.0, p=.002$ (see Fig. 3), with no effect of study order. An equivalent analysis on discovery of a $75 \%$ rule (yes/no) found no effect of condition, $\chi^{2}(1)=.57, p=.45$.

As in Experiment 1, we found that the explain prompt was successful in boosting self-reported explanation (relative to compare), $F(1,149)=26.9, p<.001$, but that the compare prompt was not effective in boosting self-reported comparison (between-category comparison + withincategory comparison). In fact, participants prompted to explain reported significantly higher levels of total comparison that participants prompted to compare, $p=.005$.

These results suggest that the poor performance of participants prompted to compare in Experiment 1 was not due to the restriction to within-category comparisons. Experiment 3 thus considers whether a broader withincategory comparison, one that focuses on all four items at once, might lead to better learning.

## Experiment 3

## Method

Participants One-hundred-ninety-three adults participated in the study through the Amazon Mechanical Turk marketplace. An additional 60 participants were tested, but were excluded because they failed a catch trial or because they had previously completed a similar experiment. Participants were paid for their participation.

Materials The stimuli were those in Experiments 1-2.
Procedure As in Experiments 1-2, the procedure consisted of a study phase followed by a rule-reporting phase.

The study phase was identical to Experiment 2 except that the four study conditions were as follows: (1) the explanation task from Experiments 1-2, (2) the withincategory pairwise comparison task from Experiments 1-2, (3) a group comparison task in which participants simultaneously compared all four robots in each category, or (4) a group explanation task. Conditions (3) and (4) are described below. As in Experiment 2, the total study time in each condition was 360 seconds.

Group comparison task: "What are the similarities and differences between the Glorp robots (Robots A-D)?" After participants responded to this prompt, they received a similar prompt for the Drent robots.

Group explanation task: "Try to explain why robots A-D are Glorp robots." After participants responded to this prompt, they received a similar prompt for the Drent robots.

The rule-reporting phase was identical to Experiments 1 and 2. After completing the rule-reporting phase, participants received the same debriefing questions as in Experiment 2. No memory task was included in this study.

## Results and Discussion

We first analyzed the total number of rules discovered (0-4) across each of the four study conditions. A one-way

ANOVA revealed a significant difference in number of rules discovered, $F(3,189)=4.74, p=.003$. A Tukey posthoc analysis showed that participants who performed pairwise comparisons discovered significantly fewer rules than participants who performed individual explanations ( $p$ $=.013$ ) or group explanations $(p=.005)$, and marginally fewer rules than participants who performed group comparisons ( $p=.068$ ).


Figure 4: Rule Discovery by Study Condition in Exp. 3
We next analyzed whether the proportion of participants who discovered at least one $100 \%$ rule varied across conditions (see Fig. 4). A log-linear analysis of study task $\times$ discovery of at least one $100 \%$ rule (yes/no) found a significant effect of study task on whether participants discovered a $100 \%$ rule, $\chi^{2}(3)=26.4, p<.001$. Additional log-linear analyses found no difference in performance between the group comparison, group explanation, and individual explanation conditions, $\chi^{2}(1)=4.12, p=.13$; however, the pairwise-comparison prompt was significantly less effective than the other three, $\chi^{2}(1)=22.3, p<.001$, including the group-comparison condition, $\chi^{2}(1)=6.86, p=$ .009. A log-linear analysis of study task $\times$ discovered a $75 \%$ categorization rule (yes/no) found that the study task did not affect whether participants discovered a $75 \%$ rule, $\chi^{2}(3)=$ $.54, p=.91$.

These results suggest that the pairwise comparison condition was relatively ineffective not because comparison is an ineffective category learning strategy more generally, but instead because participants in the pairwise comparison condition focused on a prescribed set of comparisons involving two items at a time. When it comes to category learning, it may be important to consider the global structure of categories to effectively assess the cue and category validities of different features.

## General Discussion

The present study investigated whether generating explanations and making comparisons would improve people's ability to discover rules that could be used to categorize a set of novel objects. All three experiments found that performing an explanation task enhanced discovery of categorization rules that could account for all cases; however, the effects of the comparison tasks were
more varied. Performing either within-category or betweencategory pairwise comparisons did not support rule discovery. However, comparing all the category exemplars in each group did increase $100 \%$ rule discovery.

Our results are consistent with previous work demonstrating that engaging in explanation supports learning. In particular, we replicate the results of studies that have used similar materials (Williams \& Lombrozo, 2010, 2013). In the present study, the explanation task succeeded in helping participants discover abstract patterns that unified each of the categories. Furthermore, the explanation task stimulated spontaneous comparison, allowing participants to reap the benefits of comparison even if they were not explicitly asked to compare.

Surprisingly, we find that under some conditions engaging in a pairwise comparison task can impair learning. However, other types of comparison, such as comparing all the exemplars in each category, did promote learning, suggesting that comparison can be an effective strategy for learning novel categories. But importantly, some comparison prompts are more effective than others (see also Rittle-Johnson \& Star, 2009), and comparison prompts may be most effective when they stimulate a broad range of comparison processes. One question for future research is whether the combination of within-category and betweencategory pairwise comparisons can in fact be beneficial, or whether "group" comparison provides unique advantages.

It is also worth pointing out some of the limitations of this study. Overall, the eight robots were highly similar and easily alignable. This might explain why spontaneous comparison was so common among participants who completed the explanation task. The high rates of spontaneous comparison make it difficult to differentiate effects of explanation from effects of comparison; the question of whether explanation and comparison exert unique constraints on learning may be easier to address with a task that more effectively isolates each process.

In future work, we hope to explore whether explanation and comparison have additive effects in more difficult learning tasks, where we also anticipate benefits to comparing (to align features) before explaining (to identify patterns). More research is needed, but the present studies provide important steps towards understanding the relationship between explanation and comparison and how these processes can most effectively support learning.

## Acknowledgments

We thank Dedre Gentner and members of the Northwestern University Language and Cognition Laboratory for valuable feedback on this work and Lena Lam for assistance with coding. BJE was supported by an NSF Graduate Research Fellowship; TL was supported by an NSF grant (DRL1056712) and the McDonnell Foundation.

## References

Chi, M. T. H. (2000). Self-explaining expository tests: The dual processes of generating inferences and repairing
mental models. In R. Glaser (Ed.), Advances in Instructional Psychology (pp. 161-238). Hillsdale, NJ: Lawrence Erlbaum Associates.
Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. Cognitive Science, 7, 155-170. (Reprinted in A. Collins \& E. E. Smith (Eds.), Readings in cognitive science: A perspective from psychology and artificial intelligence. Palo Alto, CA: Kaufmann).
Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. Cognitive Science, 34, 752-775.
Gentner, D., Loewenstein, J., Thompson, L., \& Forbus, K. (2009) Reviving inert knowledge: Analogical abstraction supports relational retrieval of past events. Cognitive Science, 3, 1343-1382.
Gentner, D., \& Markman, A. B. (1997). Structure mapping in analogy and similarity. American Psychologist, 52, 4556.

Higgins, E. J., \& Ross, B. H. (2011). Comparisons in category learning: How best to compare for what. In L. Carlson, C. Holscher, \& T. Shipley (Eds.), Proceedings of the $33^{\text {rd }}$ Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.
Kurtz, K. J., Miao, C., \& Gentner, D. (2001). Learning by analogical bootstrapping. Journal of the Learning Sciences, 10, 417-446.
Loewenstein, J., Thompson, L., \& Gentner, D. (1999). Analogical encoding facilitates knowledge transfer in negotiation. Psychonomic Bulletin \& Review, 6, 586-597.
Lombrozo, T. (2012). Explanation and abductive inference. In K.J. Holyoak and R.G. Morrison (Eds.), Oxford Handbook of Thinking and Reasoning (pp. 260-276), Oxford, UK: Oxford University Press.
Nokes-Malach, T. J., VanLehn, K., Belenky, D., Lichtenstein, M., \& Cox, G. (2012). Coordinating principles and examples through analogy and selfexplanation. European Journal of Education of Psychology.
Oppenheimer, D. M., Meyvisb, T., \& Davidenkoc, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. Journal of Experimental Social Psychology, 45(4), 867-872.
Rittle-Johnson, B., \& Star, J. R. (2009). Compared with what? The effects of different comparisons on conceptual knowledge and procedural flexibility for equation solving. Journal of Educational Psychology, 101, 529-544.
Siegler, R. S. (2002). Microgenetic studies of selfexplanations. In N. Granott \& J. Parziale (Eds.), Microdevelopment: Transition processes in development and learning (pp. 31-58). New York: Cambridge University.
Williams, J. J., \& Lombrozo, T. (2010). The role of explanation in discovery and generalization: evidence from category learning. Cognitive Science, 34, 776-806.
Williams, J. J., \& Lombrozo, T. (2013). Explanation and prior knowledge interact to guide learning. Cognitive Psychology, 66, 55-84.

# Biologically Plausible, Human-scale Knowledge Representation 

Eric Crawford (e2crawfo@uwaterloo.ca)<br>Centre for Theoretical Neuroscience, University of Waterloo, Waterloo, ON, N2L 3G1<br>Matthew Gingerich (majugi@cs.ubc.ca)<br>University of British Columbia, Vancouver, BC, V6T 1Z4<br>Chris Eliasmith (celiasmith@uwaterloo.ca)<br>Centre for Theoretical Neuroscience, University of Waterloo, Waterloo, ON, N2L 3G1


#### Abstract

Several approaches to implementing symbol-like representations in neurally plausible models have been proposed. These approaches include binding through synchrony (Shastri \& Ajjanagadde, 1993), mesh binding (van Der Velde \& de Kamps, 2006), and conjunctive binding (Smolensky, 1990; Plate, 2003). Recent theoretical work has suggested that most of these methods will not scale well - that is, they cannot encode structured representations that use any of the tens of thousands of terms in the adult lexicon without making implausible resource assumptions (Stewart \& Eliasmith, 2011; Eliasmith, in press). Here we present an approach that will scale appropriately, and which is based on neurally implementing a type of Vector Symbolic Architecture (VSA). Specifically, we construct a spiking neural network composed of about 2.5 million neurons that employs a VSA to encode and decode the main lexical relations in WordNet, a semantic network containing over 100,000 concepts (Fellbaum, 1998). We experimentally demonstrate the capabilities of our model by measuring its performance on three tasks which test its ability to accurately traverse the WordNet hierarchy, as well as to decode sentences employing any WordNet term while preserving the original lexical structure. We argue that these results show that our approach is uniquely well-suited to providing a biologically plausible, human-scale account of the structured representations that underwrite cognition.


Keywords: knowledge representation; biologically plausible; scaling; neural; vector symbolic

## Introduction

One of the central challenges for contemporary cognitive modelling is scaling. As Jeff Hinton remarked in his address to the Cognitive Science Society, "In the Hitchhiker's Guide to the Galaxy, a fearsome intergalactic battle fleet is accidentally eaten by a small dog due to a terrible miscalculation of scale. I think that a similar fate awaits most of the models proposed by Cognitive Scientists" (Hinton, 2010). This observation can be taken as a challenge for cognitive modellers: Will the principles demonstrated in a small-scale cognitive model scale up to the complexity of a human-sized cognitive system? This scaling problem has often been thought to be a special challenge for biologically inspired approaches to cognitive modelling (Jackendoff, 2002). This is because the basic principles employed in such models often do not allow for a straightforward characterization of structured representations, despite the ubiquity of such representations in cognitive behaviour. This same concern is not as immediate for symbolicist approaches which typically take structured representations to be primitive (Anderson, 2007).

In this paper, we briefly review past connectionist approaches to addressing the problem of representing structure, and discuss recent criticisms of those approaches which suggest that they will not scale. We then present a new approach that we have developed that allows for the representation and manipulation of large-scale structured representations in anatomically and physiologically plausible models of brain function. In past work we have provided theoretical arguments suggesting that this approach will scale better than others (Stewart \& Eliasmith, 2011). Here, our focus is on empirically demonstrating that claim. We do so by encoding the central structural relations in WordNet into neural representations in a spiking network. We present the results of three experiments showing that 1 ) this information can be decoded for arbitrary lexical items, 2) lexical hierarchies of any depth within WordNet are successfully represented, and 3) these lexical representations can be combined to represent structured sentences with the same methods.

## Past Approaches

There have been many approaches to representing structure in connectionist networks. We consider three of the most successful: 1) binding through synchrony; 2) mesh binding; and $3)$ conjunctive binding.

The suggestion that structured cognitive representations could be constructed using binding through synchrony (Shastri \& Ajjanagadde, 1993) was imported into cognitive modelling from the earlier hypothesis that feature binding in vision can be accounted for by the synchronization of neurons in visual cortex (von der Malsburg, 1981). Recently, this approach has seen a revival in the DORA architecture (Doumas et al., 2008) and its variants, which focus on representing structures for analogical reasoning. In these models, the temporal relationships between connectionist nodes are employed to represent structured relations. Structured representations (e.g. bigger(Fido, Sarah)) are constructed out of four levels of representation, where nodes in higher levels represent more complex structures via their connections to nodes in lower layers. As has been argued in more detail elsewhere, this kind of representational scheme will not scale well (Stewart \& Eliasmith, 2011; Eliasmith, in press) because the number of nodes needed to support arbitrary structured representations over even small vocabularies (e.g. 6000 lexical items) is larger than the number of neurons in the brain.

Notably, this is not an issue with the use of synchrony per se, but rather with the the way binding has been mapped to network nodes. However, it has also been suggested that synchrony itself will not scale well to binding complex structures (Stewart \& Eliasmith, 2011; O'Reilly \& Munakata, 2000).

A different approach to structured representation has been taken by van Der Velde \& de Kamps (2006) in their work on the Neural Blackboard Architecture (NBA). To avoid the exponential growth in resources, the NBA employs "neural assemblies." These assemblies are temporarily bound to particular symbols using a mesh grid of neural circuits (e.g. bind(noun1, Fido)). Larger structures are then built by binding these assemblies to roles using a gating circuit (e.g. gate(agent1, bind(noun1, Fido))). Neural assemblies that bind roles and their gated word assemblies are used to define higher level structure assemblies which can be used to represent sentential structures. The use of temporary binding in this manner significantly reduces the resource demands of this approach compared to DORA. However, as argued in Stewart \& Eliasmith (2011) and demonstrated in more detail in Eliasmith (in press), just to represent simple sentences of the form relation(agent, theme) from a vocabulary of 60,000 terms, this approach requires about $480 \mathrm{~cm}^{2}$ of cortex, approximately one fifth of total cortical area. This is much larger than known language areas which account for both representation and processing of linguistic terms. Consequently, while the NBA has improved scalability compared to DORA, it remains implausible.

The final approach we consider is the class of proposals broadly called conjunctive coding approaches, or, more recently, Vector Symbolic Architectures (VSAs; (Gayler, 2003)). In general, these approaches propose some kind of nonlinear vector operation to bind two vectors together. The earliest and perhaps best known such approach is that proposed by Smolensky (1990), which employs the tensor product as the binding operation. The model presented in this paper employs a VSA which uses circular convolution as the binding operation (after Plate (2003)). The crucial difference between using the tensor product vs. using circular convolution is that for an n-dimensional vector, a tensor binding results in an $n^{2}$-dimensional vector, whereas the circular convolution binding results in an $n$-dimensional vector. This computational difference results in severe scaling differences when considering possible biological implementations. In particular, tensor products scale exponentially poorly as the depth of the structure increases. For example, Eliasmith (in press) shows that encoding a sentence where lexical items have hierarchical relations of depth two or greater (e.g. Sarah isA(person isA(mammal))) will require approximately 625 $\mathrm{cm}^{2}$ of cortex. Again, this is significantly larger than relevant language areas.

The above considerations suggest two main challenges for connectionist implementations of structured representations: lexical scaling and hierarchical scaling. Lexical scaling means having a lexicon that is as large as an adult hu-
man's vocabulary. Hierarchical scaling refers to being able to encode the depth of grammatical and lexical relations found in adult humans. Any method that claims to provide appropriate scaling will have to demonstrate success along both of these dimensions. The approach that we adopt in this work employs a neural implementation of a VSA which uses circular convolution for binding. The purpose of this paper is to demonstrate empirically that our approach successfully meets both of these challenges.

## Theoretical Approach and Methods WordNet

The target of our efforts - the human-scale knowledge base that we will be encoding - is WordNet, a manually constructed lexical database of the English language (Fellbaum, 1998). WordNet's design is intended to reflect the organization of concepts in a psychologically plausible way using a handful of common relationships. In WordNet words are divided into "synsets" or synonym sets of words that have the same meaning. Words that have multiple meanings are listed in multiple synsets, so the fundamental unit in WordNet is not a word but a word sense. Each synset is linked to other synsets by relations, of which there are several types. The two relation types that are of the most interest are hypernymy and holonymy. A hypernym of a concept is the general type of the concept (i.e. dog has the hypernym canine); a holonym of a concept is something that that concept is a part of (i.e. lock has the holonym door). These relations are explicitly encoded in the lexicon we employ. The inverse of the hypernym and holonym relations are also implicitly included in our encoding, although we do not test their extraction as this requires more complex control of signal flow that is beyond our present scope. The depiction of lexical relations found in WordNet is slightly simplified, though it is sufficient for our purposes; a complete description of the simplifications made can be found in Fellbaum (1998).

Each synset can be defined in terms of its relationships with other synsets. This means that a term such as dog can be defined as:

$$
\begin{equation*}
\operatorname{dog}=\text { isA }(\text { canine }) \text { and partOf }(\text { pack }) \tag{1}
\end{equation*}
$$

We will make extensive use of this type of representation in our model. We think of the relations in (1) as belonging to the dog synset, and pack and canine as the targets of the relations.

## Vector Symbolic Architectures

VSAs in general provide a means for representing structured knowledge using high-dimensional vectors. This makes VSAs amenable to neural implementation using the Neural Engineering Framework, which shows how to systematically use populations of spiking neurons to represent vectors and functions thereof (Eliasmith \& Anderson, 2003). In a VSA, each symbol to be represented is randomly assigned a high-dimensional vector. Two core operations are provided, each of which takes two vectors as input and returns a third.

Binding $(\circledast)$ two vectors returns a third vector that is disimilar to both of the original vectors. The superposition $(+)$ of two vectors returns a third vector that is similar to both of the original vectors. Here we employ a close relative of Holographic Reduced Representations (Plate, 2003), a type of VSA in which binding is implemented via circular convolution, superposition is implemented via vector addition and vectors representing symbols are randomly chosen from the D-dimensional unit sphere. The circular convolution returns a vector which has the same dimension as the two input vectors, which is a significant improvement over the tensor product VSA discussed in the Past Approaches section.

We can use these operations to encode graph-like structures such as WordNet. First we fix a dimension D for our vectors ( $\mathrm{D}=512$ in our model). Then each WordNet synset and each relation type is assigned a random vector on the D dimensional unit sphere called an ID-vector. Each synset is also assigned a second D-dimensional vector, built-up using the VSA operations, which stores the structural information about the synset. To construct this vector, for each relation belonging to a particular synset, we bind the ID-vector for the relation type to the ID-vector for the target of the relation. We then superpose the results from all the relations. The following equation demonstrates this process for the dog synset:

$$
\begin{equation*}
\operatorname{dog}=\text { isA } \circledast \text { canine }_{\text {ID }}+\text { partOf } \circledast \text { pack }_{\text {ID }} \tag{2}
\end{equation*}
$$

where all variables on the right-hand side are ID-vectors. We have disambiguated the two vectors assigned to a synset by denoting the ID-vector with the ID subscript. What makes (2) useful is that dog preserves information about its constituents; we can use a third operation, dereferencing, to determine what a given vector is bound to in dog. The dereferencing operation is performed by binding dog with the inverse of the given vector. As an example, imagine we want to extract the synset that the dog synset is related to via the isA relation type. We bind $\operatorname{dog}$ with $\overline{\mathrm{isA}}$, the inverse of the isA vector:

$$
\begin{align*}
& \operatorname{dog} \circledast \overline{\mathrm{isA}} \\
& =\left(\mathbf{i s A} \circledast \text { canine }_{\mathbf{I D}}+\text { partOf } \text {. } \text { pack }_{\text {ID }}\right) \circledast \overline{\mathbf{i s A}} \\
& =\text { isA } \circledast \text { canine }_{\text {ID }} \circledast \overline{\text { isA }}+\text { partOf } \not \text { pack }_{\text {ID }} \circledast \overline{\text { isA }} \\
& \approx \text { canine }_{\text {ID }}+\text { partOf } \circledast \text { pack }_{\text {ID }} \circledast \overline{\mathbf{i s A}} \tag{3}
\end{align*}
$$

Equation (3) shows that $\boldsymbol{\operatorname { d o g }} \circledast \overline{\mathrm{isA}}$ is canine ${ }_{\text {ID }}$ superposed with another vector which can effectively be regarded as noise. All that remains is to remove that noise, and we will discuss methods for doing so below.

We call vectors constructed in the manner of dog in (2) semantic pointers because they are a compressed representation of their constituents, and preserve similarity (i.e. semantic) relations in their compressed form. In addition, the dereferencing operation is similar to the dereferencing of pointers in programming languages. Semantic pointers have a wide range of uses; indeed, they are central to the Semantic Pointer

Architecture (Eliasmith, in press) which was used to create Spaun, currently the world's largest functional brain model, which is able to account for several perceptual, motor and cognitive behaviours (Eliasmith et al., 2012).

Now, given a semantic pointer representing a synset, we can traverse the connections between that synset and related synsets by dereferencing its semantic pointer with the IDvector of a relation type, obtaining a noisy version of the IDvector of the target of the relation, as demonstrated in (3).

After this initial dereferencing, we must still determine how to remove the noise from the vector returned by the dereferencing operation. Looking again at equation (3), we see that $\operatorname{dog} \circledast \overline{\mathbf{i s A}}$ is similar to canine ${ }_{\text {ID }}$ since it consists of canine $_{\text {ID }}$ superposed with a noise vector, and is dissimilar to the rest of the ID-vectors ( pack $_{\text {ID }}$, isA, partOf) since they are related to $\boldsymbol{\operatorname { d o g }} \circledast \overline{\mathbf{i s A}}$ via the binding operation. A cleanup memory, which returns the vector in a vocabulary which is most similar to a given input vector, is a potential solution to this denoising problem. However, we want our model to be able to traverse the full WordNet hierarchy, not just a single relation, and ID-vectors alone contain no structural information. We need to have some way to move from the ID-vector for a synset to its semantic pointer. We can perform both the denoising and mapping to semantic pointer in one step by using an associative cleanup memory rather than a pure cleanup memory. In short, we take the noisy ID-vector returned by the dereferencing operation and feed it into an associative cleanup memory mapping each synset's ID-vector to its semantic pointer, thus obtaining a clean semantic pointer which can then be used in further traversals.

## Neural Representation and Computation

Thus far we have described VSAs and how they can be used to encode structural knowledge such as WordNet, but have not yet said anything of how to implement them in neurons. For this purpose we turn to the Neural Engineering Framework (NEF), a set of methods for building biologically plausible models using principles for neural representation, computation and dynamics (Eliasmith \& Anderson, 2003). The central idea behind the NEF is that a group of spiking neurons can represent vectors over time, and that connections between groups of neurons can compute functions on those vectors. More precisely, a group of neurons represents any of a set of vectors, that is, a vector space. The NEF provides a set of methods for determining what the connections need to be to compute a given function on the vector space represented by a group of neurons. Suppose we wish to compute the function $y=f(x)$, where vector space $x$ is represented in population A , and vector space y is represented in population B. To do so, the NEF assumes that each neuron in A and B has a "preferred direction vector." The preferred direction vector is the vector (i.e. direction in the vector space) for which that neuron will fire most strongly. Consequently, the spiking activity of every neuron in a population A can be written

$$
\begin{equation*}
a_{i}(x)=G\left[\alpha_{i} e_{i} x+J_{\text {bias }}\right] \tag{4}
\end{equation*}
$$

where $a_{i}$ is the ith neuron in the population, $G$ is the spiking neural nonlinearity, $\alpha_{i}$ is the gain of the neuron, $e_{i}$ is the preferred direction (or encoding) vector, and $J_{\text {bias }}$ is a bias current to account for background activity of the neuron. The elements in the square brackets determine the current flowing into the cell, which then drives the spiking of the chosen single cell model $G$. For computational efficiency, we employ a leaky integrate-and-fire (LIF) neuron model, though the NEF can be applied for arbitrary neuron models. Equation (4) is referred to as an encoding equation because it describes how a vector space, in this case $x$, is encoded into neural spikes. The NEF assumes a least-squares optimal linear decoding to reconstruct x or any nonlinear function thereof, $\mathrm{f}(\mathrm{x})$. Thus, we must find the decoders $d_{i}^{f}$, such that

$$
\begin{equation*}
E=\frac{1}{2} \int\left[f(x)-\sum_{i} a_{i}(x) d_{i}^{f}\right]^{2} d x \tag{5}
\end{equation*}
$$

is minimized. Finding the decoders in this manner then provides us with a way to estimate any vector $\mathrm{f}(\mathrm{x})$ given the activities from the encoding equation. We can write this as the decoding equation:

$$
\begin{equation*}
\widehat{f(x)}=\sum_{i} a_{i}(x) d_{i}^{f} \tag{6}
\end{equation*}
$$

where N is the number of neurons in the group and $\widehat{f(x)}$ is the estimate of $f(x)$ where $x$ is the input driving the neurons. Recall that our purpose in defining the representation of a vector space in a neural population is to use it to compute a function between two populations. If we define the encoding and decoding for groups A and B using equations (4) and (6), we can substitute the decoding of $A$ into the encoding of $B$, thereby deriving connection weights. In addition, if the function we wish to compute is linear, we can include the relevant linear operator in the weight equation. The weight equation for computing any combination of linear and nonlinear functions is then:

$$
\begin{equation*}
\omega_{i j}=d_{i}^{f} \alpha_{j} L e_{j} \tag{7}
\end{equation*}
$$

where $i$ indexes the neurons in group A and $j$ indexes the neurons in $\mathrm{B}, \mathrm{f}$ is any nonlinear function and L is any $D_{B} \mathrm{x}$ $D_{A}$ linear operator, where $D_{A}$ and $D_{B}$ are the dimensionalities of the two vector spaces.

It is worth noting that these representations and computations can be implemented to any desired precision, by adding enough neurons. Specifically, the root mean-squared-error goes down as $1 / \mathrm{N}$ (Eliasmith \& Anderson, 2003). One of the main concerns of this paper is to demonstrate that the operations required for representing human-scale lexical structure can be done with a reasonable number of neurons.

It is straightforward to use the NEF to create networks of spiking neurons for computing the inverse and circular convolution operations (Eliasmith, 2005). However, neurally implementing an associative memory requires a specific application of these methods, which we will outline in the next section.

## Neural Associative Memory

There are several ways in which associative memories can be implemented (see (Lansner, 2009) for a review). Recently, (Stewart et al., 2010) used the NEF to construct an efficient, fast autoassociative (a.k.a. cleanup) memory out of spiking neurons, and this approach can be trivially extended to construct an associative memory. Moreover, they demonstrate that this approach significantly outperforms a linear associator, a direct function approximator and a standard multi-layer perceptron. However, that paper only considers lexicons up to 10,000 items, and does not discuss any actual lexical processing, as is our focus here.

Given a noisy version of an ID-vector as input, we want our associative memory to output a clean version of the corresponding semantic pointer. A simple algorithm that achieves this is to take the dot product of the input vector with each of the ID-vectors in the vocabulary, threshold these values (set to 0 all values below some fixed threshold), multiply each semantic pointer vector by its corresponding thresholded value, and add all the resultant vectors together to obtain a single D dimensional vector. If the input vector is only similar to one of the ID-vectors, then all of the dot products will be thresholded except for one and the output vector will be equal to the correct semantic pointer.

We can use the NEF to implement this algorithm in spiking neurons as follows. Assign each synset a small ( $\sim 20$ ) population of neurons. Then we set the preferred direction vector of each neuron equal to the ID-vector for the synset that the neuron is assigned to. Equation (4) shows the activities of each population can be seen as encoding the similarity between the input vector and the population's assigned ID-vector. To determine the weight matrices between these populations and the output population, we first minimize equation (5) with $f$ set to a thresholding function to find optimal decoders, and then substitute these into equation (7) with $L$ set to semantic pointer of the population's assigned synset. Thus, the output of a population with ID-vector $e$ and semantic pointer $s$ is a neural reconstruction of $\operatorname{threshold}\left(x^{T} e\right) \cdot s$. Summing the output of all the association populations is implicitly performed by the dendrites of the neurons in the output population.

## The Model

The core of the model is a network of spiking neurons, constructed using the techniques outlined above, which, given a semantic pointer corresponding to a WordNet synset and a query vector corresponding to a relation type, returns the semantic pointer corresponding to the target of the relation. This network can be used to traverse the WordNet hierarchy by running it recursively, with the output of the last run used as input on the next run. The tasks of moving the output into the input, controlling which relation goes into the query vector population, etc, are not investigated here as they are peripheral to our central concern of representing human-scale structured knowledge in a biologically plausible manner.

A schematic diagram of the model is depicted in Fig-


Figure 1: The network of spiking neurons that traverses the WordNet graph. Assume $S=R \circledast T_{I D}+U \circledast V_{I D}$ where R, $T_{I D}$, U , and $V_{I D}$ are all ID-vectors and T is the semantic pointer corresponding to $T_{I D}$. All nodes represent neural populations.
ure 1 . The nodes correspond to populations of spiking neurons which represent and manipulate 512-dimensional vectors. All neurons employ the leaky integrate-and-fire neuron model. Each time the model was run to perform a single edge traversal, it was simulated for 100 ms with a simulation timestep of 1 ms , after which the vector represented by the Output population was taken to be the output of the model. The Binding node, which performs a circular convolution between two vectors, contains 51,400 neurons, and each of the other 4 nodes outside the associative memory contain 25,600 neurons. The associative memory contains 117,659 populations, one for each WordNet synset, with 20 neurons each, resulting in a grand total of $2,506,980$ neurons. This is equivalent to approximately $14.7 \mathrm{~mm}^{2}$ of cortex, much smaller than previous neural approaches to structured representation, which required on the order of $500 \mathrm{~cm}^{2}$ cortex (as there are about 170,000 neurons per $\mathrm{mm}^{2}$; (Eliasmith, in press)).

## Experimental Results

We performed three experiments on the model to test different aspects of its performance. For each experiment, a trial consists of using the network to answer a single question about the WordNet graph (the question is different for each experiment). A run consists of a group of trials. For each experiment we perform some number of runs, calculate the performance on each run as the percentage of trials on which the model answered correctly, and report the mean performance over all the runs. We employ a bootstrapping method to obtain $95 \%$ confidence intervals on the mean performance. This data can be seen in Table 1. The model was perfectly successful on Experiment 1, and nearly so on Experiments 2 and 3.

## Experiment 1: Decoding Accuracy

This test investigates how many of the 117,659 concepts in the WordNet can be accurately decoded. For this experiment, we present the model with a semantic pointer corresponding to a randomly chosen synset and an ID-vector corresponding to a relation type, and see if the model returns the semantic pointer corresponding to the target of that relation. For example, we might present the network with the semantic pointer for $\mathbf{d o g}$ and the ID-vector for the relation isA and see if the
network returns the semantic pointer for canine. To be considered correct, the returned vector must have a larger dot product with the correct semantic pointer than with any incorrect semantic pointer in the vocabulary, and this dot product must pass a threshold of 0.7. We ran 20 runs, each of which consisted of 100 trials, amounting to 2000 edge traversals.

## Experiment 2: Hierarchy Traversal

This experiment is designed to test the model's ability to traverse the network to arbitrary depth. To that end, we use the model to answer the following question: given two synsets and a relation type, can the second synset be reached from the first synset solely by following links of the specified type? We present the model with the semantic pointer corresponding to the first synset as well as the ID-vector for the given relation type. Then we run the model, and compare the output vector to the semantic pointer for the second synset. If they are the same (their normalized dot product is above a fixed threshold), then the model responds with a Yes. If not, we feed the output vector back into the model as the new semantic pointer and run the model again. This process is repeated until the model returns a vector with a norm below a fixed threshold. If the model reaches this point, it responds with a No. Here it is especially important that the decoded semantic pointer be very similar to the correct semantic pointer since we recursively use the output, and large errors would build up with successive edge traversals. Our tests were performed using only the isA relation type as it is the most prominent in WordNet and permits the deepest traversals. We ran 20 runs, each consisting of 20 positive examples (the second synset could be reached in the actual WordNet graph), and 20 negative examples.

Table 1: Experimental Results

| Experiment | \% correct | 95\% CI |  |
| :--- | :--- | :--- | :--- |
|  |  | $10 w e r$ | upper |
| 1. Decoding Accuracy | 100.0 | 100.0 | 100.0 |
| 2. Hierarchy Traversal | 95.5 | 94.3 | 96.9 |
| 3. Sentence Encoding | 99.6 | 99.3 | 99.8 |

## Experiment 3: Sentence Encoding

The final experiment we performed was designed to confirm that this method of knowledge representation is flexible enough to allow concepts to bind to arbitrary roles while still encoding the thousands of relationships between themselves. To this end, we test whether the network can accurately decode a crude approximation of a sentence, consisting of synsets bound to one of six different roles. To build a sentence, we randomly choose roles for inclusion, each with a different probability, and then synsets are randomly chosen to fill the selected roles. Each role type is assigned an ID-vector, in the same way that ID-vectors are assigned to relation types. A semantic pointer for the sentence is created by binding synset ID-vectors to role ID-vectors in the usual way. We then present the network with the semantic pointer for the sentence and the ID-vector for a role, and see if the vector it outputs is the same as the semantic pointer for the concept filling that role in that sentence. To determine the correctness of a particular decoding, we used the same criteria as in Experiment 1. We ran 20 runs, each of which consisted of constructing 30 sentences and asking the model about each role therein, amounting to 2411 edge traversals.

## Conclusion

These empirical results demonstrate that our spiking neural network can accurately represent structured knowledge representations approaching the scale of those found in an adult human. Moreover, our's is the only approach with neural resource requirements that fall within the range of biological plausibility.

## Acknowledgments

Funding for this work was provided by the National Science and Engineering Research Council of Canada, Canada Research Chairs, the Canadian Foundation for Innovation and the Ontario Innovation Trust.

## References

Anderson, J. R. (2007). How can the human mind occur in the physical universe? Oxford University Press.
Doumas, L. A. A., Hummel, J. E., \& Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. Psychological Review, 115, 1-43.
Eliasmith, C. (2005). Cognition with neurons: A large-scale, biologically realistic model of the Wason task. In G. Bara, L. Barsalou, \& M. Bucciarelli (Eds.), Proceedings of the 27th annual meeting of the cognitive science society ( pp . 624-630).
Eliasmith, C. (in press). How to build a brain: A neural architecture for biological cognition. New York, NY: Oxford University Press.
Eliasmith, C., \& Anderson, C. H. (2003). Neural engineering: Computation, representation and dynamics in neurobiological systems. Cambridge, MA: MIT Press.

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., et al. (2012). A large-scale model of the functioning brain. Science, 338(6111), 1202-1205.
Fellbaum, C. (1998). Wordnet: an electronic lexical database. Cambridge, Massachusetts: MIT Press.
Gayler, R. W. (2003). Vector Symbolic Architectures answer Jackendoff's challenges for cognitive neuroscience. In P. Slezak (Ed.), Iccs/ascs international conference on cognitive science (pp. 133-138).
Hinton, G. (2010). Where do features come from? In Outstanding questions in cognitive science: A symposium honoring ten years of the david e. rumelhart prize in cognitive science. Cognitive Science Society.
Jackendoff, R. (2002). Foundations of language: Brain, meaning, grammar, evolution. Oxford University Press.
Lansner, A. (2009, March). Associative memory models: from the cell-assembly theory to biophysically detailed cortex simulations. Trends in neurosciences, 32(3), 17886.

O'Reilly, R. C., \& Munakata, Y. (2000). Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain (1st ed.). The MIT Press. Paperback.
Plate, T. A. (2003). Holographic reduced representations. Stanford, CA: CSLI Publication.
Shastri, L., \& Ajjanagadde, V. (1993). From simple associations to systematic reasoning: A connectionist representation of rules, variables, and dynamic bindings. Behavioral and Brain Sciences, 16, 417-494.
Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. Artificial Intelligence, 46, 159-217.
Stewart, T., \& Eliasmith, C. (2011). Compositionality and biologically plausible models. In W. Hinzen, E. Machery, \& M. Werning (Eds.), Oxford handbook of compositionality. Oxford University Press.
Stewart, T., Tang, Y., \& Eliasmith, C. (2010). A biologically realistic cleanup memory: Autoassociation in spiking neurons. Cognitive Systems Research.
van Der Velde, F., \& de Kamps, M. (2006). Neural blackboard architectures of combinatorial structures in cognition. Behavioral and Brain Sciences, 29(29), 37-108.
von der Malsburg, C. (1981). The Correlation Theory of Brain Function (Vol. Internal R; Tech. Rep.). Abteilung fiir Neurobiologie, MPI for Biophysical Chemistry.

# Cross-Situational Statistical Learning of Phonologically Overlapping Words 

Paola Escudero (paola.escudero@uws.edu.au)<br>MARCS Institute, University of Western Sydney<br>Locked Bag 1797, Penrith NSW 2751, Australia

Karen Mulak (k.mulak@uws.edu.au)

MARCS Institute, University of Western Sydney
Locked Bag 1797, Penrith NSW 2751, Australia

Haley Vlach (hvlach@wisc.edu)<br>Department of Educational Psychology, University of Wisconsin, Madison<br>1025 W. Johnson Street, Madison, WI 53706, USA


#### Abstract

Recent research has sought to examine how learners are able to track the co-occurrence of words and objects across moments in time, a behavior commonly termed crosssituational statistical learning. The current experiment was designed to examine if learners can simultaneously determine word-referent pairings while engaging in other cognitive processes that support language learning, such as distinguishing phonologically overlapping words. Participants were presented with a cross-situational statistical learning task with pairs of words in four categories: non-minimal pairs, near minimal pairs, vowel minimal pairs, and consonant minimal pairs. The results revealed that participants were able to simultaneously learn word-referent pairings while distinguishing all four categories of word pairings. However, learners experienced the most difficulty learning vowel minimal pairs. This work demonstrates that learners are able to simultaneously engage in multiple cognitive processes that support language learning.


Keywords: cross-situational statistical learning; statistical learning; word learning; phonologically minimal pairs; bilingualism

## Introduction

In one moment in time, the world presents learners with a seemingly infinite amount of information. Across several fields of study, including cognitive psychology, developmental psychology, computer science, and linguistics, a large research pursuit has been to characterize how it is that learners acquire, store, and later retrieve such a large data set of information. Indeed, this task has historically been characterized as theoretically impossible (e.g., Quine, 1960), but yet learners appear to acquire a great deal of information with ease.

A more recent trend in research has been to examine how it is that learners acquire, store, and later retrieve information across several moments in time. For example, in the domain of language learning and development, research has sought to determine how learners resolve
ambiguity in word-referent pairings across moments in time. This phenomenon is most commonly termed crosssituational or statistical word learning (e.g., Fazly, Alishahi, \& Stevenson, 2010; Frank, Goodman, \& Tenenbaum, 2009; Smith \& Yu, 2008; Vlach \& Sandhofer, 2011; Yu \& Smith, 2007, 2011).

In a typical experiment, learners are presented with a series of ambiguous learning events, which include multiple words and multiple objects. After a series of learning events, adult participants are presented with a forced-choice test in which they are asked to infer object-label pairings, while infants are presented with a preferential-looking task. This body of work has revealed that infants (e.g., 12- and 14-month-olds; Smith \& Yu, 2008) and adults (e.g., Yu \& Smith, 2007) are able to learn word mappings by tracking co-occurrence probabilities across learning events.

Cross-situational statistical learning research has focused on questions examining learners' ability to determine wordreferent pairings. However, in real-world language learning environments, learners are faced with the challenge of determining word-referent pairings while simultaneously engaging in other cognitive processes that support language learning. For example, learners must simultaneously determine word-referent pairings while parsing words that overlap phonologically.

To date, experiments have primarily used words that contain gross phonological differences, that is, words that differ in multiple sound segments, such as "beat" and "rule". However, many words, especially in English, contain the same sounds with the exception of one segment, either a vowel or a consonant. In other words, they form phonologically minimal pairs such us "beat"-"bit" or "bet""debt". Consequently, it is unknown whether learners are able to simultaneously learn cross-situational statistics while distinguishing phonologically minimal pairs.

Adults have difficulty in learning phonologically minimal pairs. For example, Dutch and Spanish listeners were presented with a word learning task in which they were explicitly taught twelve pseudo-words together with their corresponding visual referents (Escudero, Broersma, \& Simon, 2012). The words followed Dutch phonotactic
probabilities and were produced by a Dutch female speaker. Their visual referents were pictures of novel objects. At test, the native Dutch listeners made more errors for words that formed a minimal pair (e.g. "pax"-"pix") than when they formed a non-minimal pair (e.g. "beeptoe"-"pix"). Spanish listeners demonstrated an even greater difficulty in this task for minimal pairs that contained Dutch vowel contrasts that are not present in Spanish (e.g. "piex"-"pix", "pax"-"paax").

Can learners simultaneously learn cross-situational statistics and distinguish phonologically overlapping words? The current study examined whether phonologically overlapping words or minimal pairs can be successfully learned within a typical cross-situational statistical learning paradigm. In this experiment, learners were exposed to eight novel English words and eight picture referents with no explicit instructions. To examine the effect of word-pair similarity on word learning, the experiment presented learners with monosyllabic words such as "bon" and "deet" that when paired, formed four different levels of phonological overlap: (1) non-minimal pairs (nonMP), (2) near minimal pairs (nearMP), (3) vowel minimal pairs (vowelMP), and, (4) consonant minimal pairs (consMP).

## Method

## Participants

Participants were 71 undergraduates at the University of Western Sydney. A language background questionnaire revealed that 31 participants were monolingual English speakers, whose age range was 17.85 years to 52.19 years ( $\mathrm{M}=26.52$ years, $\mathrm{SD}=10.21$ years; 10 males), while 40 participants spoke two or more languages and ranged in age from 17.73 years to 28.94 years ( $\mathrm{M}=20.70$ years, $\mathrm{SD}=3.18$; 7 males). English was the dominant language of all participants.

## Stimuli

Eight monosyllabic nonsense words were recorded by a female native speaker of Australian English. Figure 1 shows the eight spoken words (in phonetic symbols) together with their randomly assigned picture referents. Four of the words were minimally different in their first consonant (left), while the other four differed in their vowel (right).

The novel words followed English phonotactic probabilities and were chosen from those included in previous studies with infant learners (see Curtin et al., 2009 for the words differing in vowels, and; Fikkert, 2010 for those differing in consonants). The female speaker produced a number of tokens of each word with child-directed intonation contours. These words and speech style were chosen to enable direct comparison of adult and infant responses to the same stimuli (Escudero, Mulak \& Vlach in preparation-a).

Two tokens of each of the eight spoken words were selected to be used in the experiment such that intonation contours were comparable across words. The visual
referents for the words were colorful pictures of novel items previously used in studies of cross-situational word learning (e.g., Vlach \& Sandhofer, 2011).


Figure 1. The eight novel words and their novel object referents.

Stimuli were presented in pairs, with four types of phonological overlap between the two spoken words that were the names of the pictures within a pair: (1) nonminimal pairs (nonMP), where the two words in the pair differed in all three sounds (e.g. /dit///pon/); (2) near minimal pairs (nearMP), where the words overlapped in one sound (e.g. /dっn/-/dit/); (3) vowel minimal pairs (vowelMP), where the words only differed in their vowel (e.g. /dit/-/dit/), and, (4) consonant minimal pairs (consMP), where the words differed in only their consonant (e.g. /bon/-/don/).

## Procedure

Participants were presented with the cross-situational learning tasks in two phases: a learning phase and a subsequent test phase.

During the learning phase, stimuli were presented via Tobii Studio on a 17 -inch screen, and the spoken words played from two speakers positioned below the screen. In each learning trial, two of the eight pictures of novel items appeared on the screen while two novel words for the pictures were spoken, such that pictures were either named left to right, or right to left. The pairings of the words and pictures were randomly assigned. The word for each picture was played once with 500 ms between them.

Trial 1: /ditt/, / bon /


Trial 2: /don/, / dit /


Trial 3: /bon/, / pon /


Figure 2. Examples of the word learning trials.

Participants were instructed to watch the pictures and listen to the words and were not told that the words were names for the pictures, nor were they asked to try and discover which word was associated with which picture.

Across the learning phase, there were a total of 36 learning trials, presented in a counterbalanced order. As mentioned above, stimuli were presented in pairs, with each trial consisting of two different pictures and two different words. There were 18 nonMP trials, and 6 each of nearMP, vowelMP, and consMP trials presented during the learning phase. Figure 2 shows examples of nonMP, nearMP and consMP trials, respectively.

After the learning phase, two testing phases were presented, though only one is reported here. In the first testing phase, stimuli were presented through Tobii Studio, as in the training phase. This phase followed immediately after the testing phase and participants were not given any additional instruction. Participants' eye-gaze was recorded without them having to make any overt response. This was done to later compare these adult data to infant eye-gaze to the same training and testing trials (Escudero, Mulak \& Vlach in preparation-a).

Here we report the results of the second testing phase which was performed immediately after the first. During the second test phase, participants performed a forced-choice inference test, which required learners to infer word-picture pairings by clicking on the corresponding computer key. Stimuli were presented through a laptop computer with a 15-inch monitor, which was set up next to the monitor for the training and first test phase. Stimuli presentation was controlled with E-Prime and participants listened to the stimuli through headphones.

During the test phase, participants saw a pair of pictures and heard four repetitions of the word that always co-
occurred with one of the pictures during the learning phase. The word was presented using two alternating repetitions of the same two tokens of that word used in the training phase, with a 500 ms interval between repetitions. Participants were asked to select whether the word corresponded to the left or right picture. There were 36 test trials in total with the same picture pairs as in the training, but the left/right positions of the pictures were randomized once for the test trials.

## Results

The current experiment sought to determine if learners could simultaneously acquire cross-situational statistics in order to learn word-referent pairings, and parse phonologically minimal pairs. Figure 3 shows the percentage of correct word-referent pairings chosen during the testing trials, separately for the four different types of phonologically overlapping pairs. Percentages for monolingual and multilingual participants are presented separately. This is because it has been shown that bilingualism affects language processing, especially word retrieval (Fennell, Byers-Heinlein, \& Werker, 2007; Bialystok, Craik, \& Luk, 2008)


Figure 3. Percentage of correct word-referent pairings for the different pair types in monolingual and multilingual listeners.

The first set of analyses examined learners' overall performance on the testing trials. Accuracy was above chance for all pair types and in both participant groups ( $M=65-76 \%, t=15-23$, all $p \mathrm{~s}<.001$ ). These results suggest that, despite the additional challenge of distinguishing phonologically overlapping words, learners
are able to learn and infer word-referent pairings during cross-situational statistical learning.

Further, a repeated measures ANOVA on the percentage of correct word-referent pairings chosen on testing trials, with pair type as the within-subject factor and language group as a between-subjects factor, revealed a main effect of pair type $(F(3,69)=3.009, p=.031)$, which indicates that both groups of listeners had lower performance for some pair types than for others. Neither the main effect of language group $(F(1,69=0.360, p=.550)$ nor the interaction of pair type * language group $(F(1,69=0.31$, $p=.942$ ) yielded statistical significance. These results suggest that, although there were not differences across groups of learners, overall participants' performance differed across the word pair types.

To follow up the effect of pair type, a series of planned comparisons were conducted between the pair types, with Bonferroni corrections for multiple comparisons: four comparisons (2-tailed). The results of these tests revealed that participants were less accurate on vowelMP than on nonMP $(t(70)=-2.53, p=.014)$ and consMP $(t=-2.44$, $p=.017$ ), while no difference was found between consonants and nonMP $(t(70)=0.189, p=.850)$ or nonMP and nearMP $(t(70)=1.468, p=.147)$. In sum, learners demonstrated the lowest performance on the vowel minimal pairs.

## Discussion

The results demonstrate that young adults can successfully learn monosyllabic nonsense words in a statistical crosssituational paradigm and without explicit instruction of word-referent pairings. Specifically, learners are able to simultaneously acquire cross-situational statistics and parse phonologically minimal pairs when learning novel words. Thus, the present study extends the findings of previous cross-situational studies (e.g., Fazly et al., 2010; Frank et al., 2009; Smith \& Yu, 2008; Vlach \& Sandhofer, 2011; Yu \& Smith, 2007) by demonstrating that a more challenging set of word-referent pairings can still be learned through the tracking of co-occurrence and statistical probabilities.

The current experiment also demonstrates that vowel minimal pairs are more difficult to learn because participants' accuracy for vowel minimal pairs was lower than that of non-minimal and consonant minimal pairs. This finding is consistent with the numerous studies that demonstrate that consonant information is more important than vowel information for lexical processing (e.g., Berent \& Perfetti, 1995; Lee, Rayner, \& Pollatsek, 2001; Perea \& Carreiras, 2006; Perea \& Lupker, 2004) and lexical acquisition (Bonatti, Peña, Nespor, \& Mehler, 2005; Nazzi \& New, 2007; Nazzi, 2005; Nespor, Peña, \& Mehler, 2003; Peña, Bonatti, Nespor, \& Mehler, 2002).

The above line of research has proposed that the main role of consonants is to signal word meaning, while vowels enable the identification of rhythm and syntactic structure (Nespor et al., 2003). Additionally, consonant information is
more critical in accessing the whole word form (see Berent \& Perfetti, 1995; Carreiras, Vergara, \& Perea, 2007; Lee et al., 2001; Lee, Rayner, \& Pollatsek, 2002; Perea \& Carreiras, 2006; Perea \& Lupker, 2004). For example, in an experiment using response time and electrophysiological measures, Carreiras et al. (2009) demonstrated that a delay in the presentation of consonant information is more detrimental for lexical processing than a delay in presentation of vowel information.

However, studies have also shown vowel information to be more important than consonant information when identifying words in fluent speech (Cole, Yan, Mak, Fanty, \& Bailey, 1996; Kewley-Port, Burkle, \& Lee, 2007). In Kewley-Port et al. (2007), the vowels or consonants were removed from sentences produced in fluent speech and it was found that vowel information had a $2: 1$ benefit over consonant information for both young normal-hearing listeners and elderly hearing-impaired listeners.

The authors argue that the reason why they find opposite results to those of the studies described above is because linguistic processing of monosyllables relies on sound-bysound, bottom-up information, while sentence intelligibility tasks incorporate considerable predictive information from top-down processing. Thus, in the context of fluent speech, we may have observed a different pattern of results. Future research should examine how acquiring cross-situational statistics and distinguishing minimal word pairs may differ in the context of fluent speech streams.

In that respect, Curtin et al. (2009) demonstrated that in lexical acquisition, infants can learn some vowel minimal pairs earlier than consonant minimal pairs, which suggests vowels may have a more lexical role than consonants in early word learning. However the authors explain that different task demands may cause the contradictory results. For example, Nazzi (2005) and Nazzi and New (2007), who found contrasting results, used a task in which infants were presented information from a real speaker, with multiple labels in the interactive communication. These task demands may thus be very different from the ones in the explicit word-referent association task in Curtin et al. (2009). Interestingly, Giezen, Escudero \& Baker (under review) suggest that these divergent results may have a developmental nature, since they found more successful vowel than consonant minimal pair learning in children, while adults exhibited the opposite bias.

Ongoing research (Escudero, Mulak \& Vlach, in preparation-a) examines infant word learning abilities using the same cross-situational word learning task as that of the present study. The results of this new study will likely shed light on the differential processing of vowels versus consonants across development.

The lack of group effects in the present study suggests that multilingualism does not influence cross-situational word learning and word retrieval immediately after learning. Interestingly, it runs contrary to studies that have demonstrated a negative influence of multiple language activation on word learning (Fennell et al., 2007) and
retrieval (Bialystok, Craik, \& Luk, 2008), and a positive influence on cognitive control (Bialystok et al., 2008; Bialystok \& Martin, 2004). Given that a bilingual processing advantage has been shown across a wide range of problem types, including both verbal and nonverbal domains, the null effect in the present study may come as a surprise. However, although word learning within a crosssituational paradigm involves intricate statistical computations and a high load on short-term memory (see Vlach \& Sandhofer, 2011, for a discussion), the 2x2 (two pictures and two spoken words per trial) learning condition may have not provided enough challenge to observe differences. It may be the case that, in the context of a crosssituational learning task with higher working memory demands, the differences between monolingual and multilingual learners will emerge.

Ongoing research (Escudero, Mulak \& Vlach, in preparation-b) is being conducted using tasks that present many words and objects in each learning event, in turn taxing working memory (e.g., in the context of $3 \times 3$ and $4 \times 4$ learning conditions). The results of this new study will likely reveal the influence of multiple language activation and cognitive control on the learning of phonologically overlapping word pairs.

On a final note, it is important to highlight that this study demonstrates the incredible capacity that human learners possess for learning language. Mapping new words to referents in the world has historically been characterized as a theoretically impossible task (e.g., Quine, 1960). However, the results of the current work demonstrate that learners can map words to referents in the world while simultaneously distinguishing phonologically overlapping sounds into words. Indeed, learners appear to accomplish multiple challenging cognitive tasks at the same time. Future research should continue to examine the cognitive processes that operate in parallel in order to support language learning and development.

## Acknowledgments

This research was funded by MARCS Institute start-up funds awarded to the first author. We would like to thank the staff at the MARCS BabyLab for their support and help during testing.

## References

Berent, I., \& Perfetti, C. A. (1995). A rose is a REEZ: The two-cycles model of phonology assembly in reading English. Psychological Review, 102(1), 146-184. doi:10.1037/0033-295X.102.1.146
Bialystok, E., Craik, F., \& Luk, G. (2008). Cognitive control and lexical access in younger and older bilinguals. Journal of Experimental Psychology: Learning, Memory, and Cognition, 34(4), 859-873. doi:10.1037/02787393.34.4.859

Bialystok, E., \& Martin, M. M. (2004). Attention and inhibition in bilingual children: evidence from the dimensional change card sort task. Developmental Science, $\quad 7(3)$, 325-339. doi:10.1111/j.14677687.2004.00351.x

Bonatti, L. L., Peña, M., Nespor, M., \& Mehler, J. (2005). Linguistic constraints on statistical computations: The role of consonants and vowels in continuous speech processing. Psychological Science, 16(6), 451-459. doi:10.1111/j.0956-7976.2005.01556.x
Carreiras, M., Vergara, M., \& Perea, M. (2007). ERP correlates of transposed-letter similarity effects: are consonants processed differently from vowels? Neuroscience letters, 419(3), 219-224. doi:10.1016/j.neulet.2007.04.053
Carreiras, M., Vergara, M., \& Perea, M. (2009). ERP correlates of transposed-letter priming effects: The role of vowels versus consonants. Psychophysiology, 46(1), 3442. doi:10.1111/j.1469-8986.2008.00725.x

Cole, R. A., Yan, Y., Mak, B., Fanty, M., \& Bailey, T. (1996). The contribution of consonants versus vowels to word recognition in fluent speech. In IEEE International Conference on Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings (Vol. 2, pp. 853 - 856 vol. 2). doi:10.1109/ICASSP.1996.543255
Curtin, S. A., Fennell, C., \& Escudero, P. (2009). Weighting of vowel cues explains patterns of word-object associative learning. Developmental Science, 12(5), 725-731. doi:10.1111/j.1467-7687.2009.00814.x
Escudero, P., Broersma, M., \& Simon, E. (2012). Learning words in a third language: Effects of vowel inventory and language proficiency. Language and Cognitive Processes, iFirst, 1-16. doi:10.1080/01690965.2012.662279
Escudero, P., Mulak, K. \& Vlach, H. (in preparation-a). Infants learning of phonological overlapping words through cross-situational statistics: An eye-tracking sudy.
Escudero, P., Mulak, K. \& Vlach, H. (in preparation-b). Cross-situational learning of phonologically overlapping words with different levels of cognitive demand: Monolingual versus multilingual learners.
Fazly, A., Alishahi, A., \& Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. Cognitive Science, 34(6), 1017-1063. doi:10.1111/j.1551-6709.2010.01104.x
Fennell, C. T., Byers-Heinlein, K., \& Werker, J. F. (2007). Using speech sounds to guide word learning: The case of bilingual infants. Child Development, 78(5), 1510-1525. doi:10.1111/j.1467-8624.2007.01080.x
Fikkert, P. (2010). Developing representations and the emergence of phonology: evidence from perception and production. In C. Fougeron, B. Kühnert, \& M. D'Imperio (Eds.), Laboratory Phonology 10: Variation, Phonetic Detail and Phonological Representation (Vol. 10, pp. 227-258). Berlin: De Gruyter Mouton.
Frank, M. C., Goodman, N. D., \& Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early
cross-situational word learning. Psychological Science, 20(5), 578-585. doi:10.1111/j.1467-9280.2009.02335.x
Giezen, M., Escudero, P., \& Baker, A. E. (under review). Rapid learning of minimally different words in five- to six year old children: Effects of hearing impairment and sound perception.
Kewley-Port, D., Burkle, T. Z., \& Lee, J. H. (2007). Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners. The Journal of the Acoustical Society of America, 122(4), 2365-2375. doi:10.1121/1.2773986
Lee, H.-W., Rayner, K., \& Pollatsek, A. (2001). The Relative Contribution of Consonants and Vowels to Word Identification during Reading. Journal of Memory and Language, 44(2), 189-205. doi:10.1006/jmla.2000.2725
Lee, H.-W., Rayner, K., \& Pollatsek, A. (2002). The processing of consonants and vowels in reading: Evidence from the fast priming paradigm. Psychonomic Bulletin \& Review, 9(4), 766-772. doi:10.3758/BF03196333
Nazzi, T. (2005). Use of phonetic specificity during the acquisition of new words: differences between consonants and vowels. Cognition, 98(1), 13-30. doi:10.1016/j.cognition.2004.10.005
Nazzi, T., \& New, B. (2007). Beyond stop consonants: Consonantal specificity in early lexical acquisition. Cognitive Development, 22, 271-279. doi:10.1016/j.cogdev.2006.10.007
Nespor, M., Peña, M., \& Mehler, J. (2003). On the different roles of vowels and consonants in speech processing and language acquisition. Lingue e Linguaggio, 2.
Pater, J., Stager, C., \& Werker, J. F. (2004). The perceptual acquisition of phonological contrasts. Language, 80(3), 384-402. doi:10.1353/lan.2004.0141
Peña, M., Bonatti, L. L., Nespor, M., \& Mehler, J. (2002). Signal-driven computations in speech processing. Science, 298(5593), 604-607. doi:10.1126/science. 1072901
Perea, M., \& Carreiras, M. (2006). Do transposed-letter effects occur across lexeme boundaries? Psychonomic bulletin \& review, 13(3), 418-422. doi:10.3758/BF03193863
Perea, M., \& Lupker, S. J. (2004). Can CANISO activate CASINO? Transposed-letter similarity effects with nonadjacent letter positions. Journal of Memory and Language, 51(2), 231-246. doi:10.1016/j.jml.2004.05.005
Quine, W. V. O. (1960). Word and object. Cambridge, MA: MIT Press.
Smith, L., \& Yu, C. (2008). Infants rapidly learn wordreferent mappings via cross-situational statistics. Cognition, 106(3), 1558-1568. doi:10.1016/j.cognition.2007.06.010
Stager, C. L., \& Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. Nature, 388, 381-382.
Vlach, H.A., \& Sandhofer, C. M. (2011). Retrieval dynamics of in-the-moment and long-term statistical word
learning. In L. Carlson, C. Hölscher, \& T. Shipley (Eds.), Proceedings of the 33rd Annual Conference of the Cognitive Science Society (pp. 789-794). Boston, MA: Cognitive Science Society.
Werker, J. F., Fennell, C. T., Corcoran, K. M., \& Stager, C. L. (2002). Infants' ability to learn phonetically similar words: Effects of age and vocabulary size. Infancy, 3(1), 1-30. doi:10.1207/S15327078IN0301_1
Yu, C., \& Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. Psychological Science, 18(5), 414-420. doi:10.1111/j.14679280.2007.01915.x

Yu, C., \& Smith, L. B. (2011). What you learn is what you see: using eye movements to study infant cross-situational word learning. Developmental Science, 14(2), 165-180. doi:10.1111/j.1467-7687.2010.00958.x

# Intuition versus Reason: Strategies that People Use to Think about Moral Problems 

Mark Fedyk (mfedyk@mta.ca)<br>Department of Philosophy, Mount Allison University<br>Sackville, NB, E4L 1G9, Canada

Barbara Koslowski (bmk2@cornell.edu)<br>Department of Human Development, Cornell University<br>Ithaca, NY, 14853, USA


#### Abstract

We asked college students to make judgments about realistic moral situations presented as dilemmas (which asked for an either/or decision) vs. problems (which did not ask for such a decision) as well as when the situation explicitly included affectively salient language vs. non-affectively salient language. We report two main findings. The first is that there are four different types of cognitive strategy that subjects use in their responses: simple reasoning, intuitive judging, cautious reasoning, and empathic reasoning. We give operational definitions of these types in terms of our observed data. In addition, the four types characterized strategies not only in the whole sample, but also in all of the subsamples in our study. The second finding is that the intuitive judging type comprised approximately $26 \%$ of our respondents, while about $74 \%$ of our respondents employed one of the three styles of reasoning named above. We think that these findings present an interesting challenge to models of moral cognition which predict that there is either a single, or a single most common, strategy - especially a strategy of relying upon one's intuitions - that people use to think about moral situations.


Keywords: Moral Judgment; Moral Reasoning; Verbal Behavior; Cognitive Strategies.

## Introduction

A significant amount of the recent and most influential research in adult moral psychology has focused on the intuitive processes that are frequently represented as the main sort of cognition that occurs when people are asked to make moral judgments about moral situations. This intuition-focused research frequently relies upon so-called "trolley dilemmas" to elicit this intuitive cognition in research participants. Trolley dilemmas are easy-tounderstand fictional moral situations that present an either/or choice between one of two ethically appealing courses of action. Participants are asked to decide one of the two courses of action, and this decision is customarily treated as a paradigmatic representation of a moral judgment. (Mikhail, Sorrentino, \& Spelke 1998, Mikhail 2012, Nichols \& Mallon 2006, Cushman \& Greene 2012)
In one of the better-known lines of research in this area, Mikhail (2007) proposes that an innate and subconscious 'universal moral grammar' determines the semantic content of people's moral judgments. The universal moral grammar is defined using the concepts of deontic logic augmented with a variety of psychological concepts like 'intentional
action'. Moreover, the computations that the universal moral grammar is hypothesized to perform and that generate moral judgment are entirely intuitive. Thus, evidence for the existence of the universal moral grammar comes from studies in which subject's intuitive decisions about trolley dilemmas are shown to be consistent with theoretical predictions about the content and logical structure of the moral grammar. (Mikhail 2008, Mikhail 2012)
Similarly, Waldmann and Dieterich (2007) argue that the content of moral judgments about trolley dilemmas and similar moral situations is determined by a different kind of intuitive knowledge, namely knowledge of a causal map. This map symbolizes causal pathways in the moral situation that a person is considering, dividing the causal pathways into those that involve agents and patients and those that do not. Agents play an active role in realizing a harmful effect, and patients experience the harmful effects caused by agents. Accordingly, the either/or choice at the heart of the situation is intuitively represented as choice between two actions (interventions), each of which can alter in different ways the causal relations that hold between agents and patients. Waldmann and Dieterich test whether a "focus on action may sometimes lead to what we call intervention myopia", such that people will focus primarily on interventions on causal pathways involving patients and agents, and focus less on interventions on causal pathways that do not involve agents and patients. (Waldmann and Dieterich 2007) Their data leads them to conclude, "the locus of intervention is one key factor contributing to moral intuitions." That may be, but the important point for our purposes is that, like Mikhail's theory, Waldmann and Dieterich's proposal is that judgments about moral either/or choices are computed by intuitive cognitive processes.
We mention these studies in some detail because we want to use them as evidence for the following claim: if you want to "nudge" subjects towards using intuitive cognitive processes to produce moral judgments about moral situations, one approach which seems likely to have this effect is to present subjects with a moral situation that embeds an explicit either/or choice - that is, a moral dilemma. After all, both studies described above are examples of experimental subjects producing moral intuitions in response to situations that contain explicit either/or choices, and there are many other studies like this. (Lombrozo 2009)

But there is another body of literature in recent moral psychology that suggests that different way of inducing intuitive responses is to ask subjects to engage with affectively salient moral situations, whether or not they are technically moral dilemmas. (Borg, Hynes, Van Horn, Grafton, \& Sinnott-Armstrong, 2006; Ciaramelli, Muccioli, Ladavas, \& di Pellegrino, 2007; Damasio, 1994; Greene, Nystrom, Engell, Darley, \& Cohen, 2004) Thus, for example, Jonathan Haidt's well-known social intuitionist model (SIM) holds that intuitively generated affective states almost always fix the content of people's moral judgment. (Haidt 2001) The basic idea is that things that we intuitively feel that we like are judged to be morally permissible, and things we intuitive feel dislike for are judged to be morally impermissible. So creating feelings along the like/dislike continuum is another potential way of encouraging uses of moral intuition; and it certainly should be if the social intuitionist model is correct.

We believe that the various lines of research that focus on the intuitive aspects of moral cognition have produced a number of novel and important scientific insights into the relationship between intuition, affect, and moral judgment. However, we are skeptical of the idea that - as per Haidt, Mikhail, and others - intuitive processes are the very nearly the only processes by which people form moral judgments. Accordingly, we present here the results of an experiment designed to identify some of the different cognitive strategies that people use when thinking about moral situations. And one of the questions we were most interested in answering was just how frequently people rely on their moral intuitions when responding to a moral situation. Because of this, our experiment was designed to maximize the likelihood that some of our participants use their moral intuitions to respond to the moral situations we asked them to consider. Our test conditions were Dilemma Non-Affect, Dilemma Affect, Problem Non-Affect, and Problem Affect, and all but the last condition was constructed so as to try to "push" subjects in these conditions towards the use of intuitive cognitive processes. Specifically, our Dilemma conditions were designed to replicate closely the trolley dilemmas discussed above by presenting our subjects with a clear either/or choice. Similarly, our Affect conditions which included language designed to elicit feelings of either disgust or sympathy - were designed to target the intuitively-mediated emotional processes that are posited by theories like the social intuitionist model. We provide a fuller description of our test conditions below - for now, we simply want to make the point that the rationale for our test conditions was our goal of trying to encourage subjects in some of our test conditions to use moral intuition. Thus, Dilemma Non-Affect targets intuitive systems like those posited by Mikhail, Waldmann, Dieterich and others; Problem Affect targets intuition systems like those posited by the social intuitionist model; Dilemma Affect targets both kinds of intuitive systems; and Problem Non-Affect acts as a control condition, insofar as it is not designed to target any specific intuitive process that has been described in the recent moral psychology literature.

Furthermore, we want stress that our aim was not to show that people are more likely to use intuitive cognitive processes in any of our test conditions. Although we designed our experiment to maximize the chances that participants in our dilemma and/or our affect conditions would be more likely to use moral intuition than those in the problem non-affect condition, our fundamental aim was to identify cognitive strategies that did not vary across our test conditions. In other words, we wanted if possible to identify any condition invariant cognitive strategies, while at the same time employing test conditions that worked against this end by making it more likely that participants in some of these test conditions would use different cognitive strategies.

## Method

We collected our data using person-to-person interviews rather than online interviews. We did this because in a separate experiment (Koslowski and Fedyk, in prep) we observed that in person-to-person interviews subjects produce a richer expression of the cognitive processes they use when responding to moral problems than they do in online-only sampling contexts.
In all of our conditions, participants were simply asked for a judgment about a moral decision faced by a character in a fictional vignette.

Participants. Participants were eighty-three undergraduate students ( $\mathrm{m}=43, \mathrm{f}=40$ ) at Cornell University. Participants were enlisted using the university's internet-based recruitment tool, and all participants received course credit for their participation.

Interview. Our interviews took place in a quiet lab room at Cornell University. No one but the interviewer and the participant was in the room at the time of the interview. Each interview was recorded using either a digital audio recorder or a tape recorder, and tapes from the latter were subsequently digitized. Data for 1 female participant was excluded from our analysis because of a tape-recorder failure that occurred during this participant's interview and thus prevented her interview from being transcribed.

Stimuli. Each participant was presented with 6 different vignettes that described a situation in which the main character in the vignette faces a moral choice. The moral situation described by our vignettes intentionally resembled the situations described by vignettes that have been used in previous research. For example, our "Smith" vignette is version of the classic runaway trolley case, albeit involving people trapped in a subway tunnel. We also used an updated version of Kohlberg's famous "Heinz" vignette, and a very simple vignette derived from Peter Singer's famous article about moral obligations towards people experiencing a devastating famine in a distant country. (Singer 1972)
All of our vignettes were written in plain English, and one of the ways in which they differed from other vignettes used by some other moral psychologists is that they described
situations that either have occurred or at least could likely occur. We did this in order to increase the ecological validity of the study, as some vignettes used by other researchers require subjects to engage in deeply counterfactual thinking. The "fat man" trolley case, for example, requires subjects to believe that, despite the laws of physics and human physiology, there exists a man fat enough to stop a runaway trolley car. (cf. Pinker 2008)
Our vignettes varied in their length, where the shortest vignette was 85 words long, and the longest 303 words long, with an average length of approximately 196 words. Each vignette introduced a main character with a gender-neutral name (like "Smith", "Davis", or "Parker") and described a situation faced by the main character that called for a moral decision.

Here are more explicit definitions of the four types of vignettes we used:

Dilemma Vignettes - for these vignettes the same language as for the Problem condition is used, except that a short phrase (like "Smith can either...") or sentence is added to the vignette that stipulates that the main character faces an either/or choice.
Affect Vignettes- in this condition, 1 or 2 short sentences were added to either the Dilemma or Problem vignettes. The sentences were designed to elicit mild feelings of either sympathy or disgust in our participants. Examples of these sentences are:
a. "The cancer is very painful, and the woman cries most days."
b. "Relief workers are trying desperately to treat children who are suffering a range of painful and eventually fatal illnesses caused by malnutrition."
c. "The boss is dirty and smells bad. He tells Adams that the sandwich he is eating is a horse-meat and pickles sandwich..."
d. "Lisa is one of the members of Smith's team. She works to support her two high-school aged children after her husband died of cancer several years ago."
Non-Affect Vignettes - for these vignettes, the sentence designed to elicit affect is omitted.
Problem Vignettes - in these vignettes, the language describing the "either/or" choice is omitted.

We constructed four test conditions by crossing these two variables: decision type (problem vs. dilemma) and affect type (affect vs. non-affect). Thus, the four test conditions were Dilemma Non-Affect, Dilemma Affect, Problem NonAffect, and Problem Affect. We did this so that - as explained in more detail above - Dilemma Affect was the condition that for theoretical reasons was most likely to push our participants towards using their moral intuitions.
Participants were randomly assigned to one of the four test conditions. Subjects therefore only ever responded to one type of vignette. Within each condition, the order in which the six vignettes were presented was random.

## Procedure

After participants had settled into the interview room and consented to participate, participants were given an unbound stack of 6 pieces of paper, where each piece of paper had written upon it one of the 6 vignettes. Participants were then asked to read along silently while the interviewer read the vignette out loud. After the interviewer was done reading the vignette, he or she then asked the participant, "What is your judgment about what X should do?", where " X " stands for the name of the main character in the relevant vignette. Participants were provided with no further instructions or feedback. They were not asked to produce any other judgments than a judgment about what the main character should do. Neither were they asked explicitly to reason about the vignettes they were presented with. Participants were therefore free to respond to our question however they wanted, which means the reasoning we observed (see below) was produced spontaneously. Once each participant concluded his or her response to our question, the interviewer moved on to the next vignette. This process was repeated until each subject had responded to all six of the vignettes.

## Coding and Analysis

The audio recordings from each of our interviews were transcribed by a professional transcription service that specializes in legal and academic work. The transcriptions were made using an "absolute verbatim" style, which means that every utterance, pause, "hmmm", and so on was transcribed and, importantly, done so using a standardized notation.
Two coders who were blind to our study's hypotheses, aims, and methods then coded these transcriptions independently. Disagreements between our coders were very infrequent, and were resolved through discussion. 82 interviews were coded, where each response to the question "What is your judgment about what X should do" was treated as an individual case. This means that our data set consisted of 492 discrete cases.

Coding Categories. We created 11 coding categories that describe easy to observe speech-acts or other kinds of verbal behavior. For example, one of our categories was "Subject asks at least one question about the vignette". Only one of our 11 categories explicitly referred to verbalized reasoning (see 4. below). The remaining 10 categories were derived from examining our transcripts for reoccurring speech-acts. We used this approach so that we did not render it a priori that our data analysis would find either an intuition / reason distinction, or find different types of reasoning in our cases. Thus, we had prior to running our analysis as much reason to expect that our analysis would sort responses into, for instance, questioning and non-questioning responses as we did for reasoned and intuitive responses.
Each of the coding categories was defined as a categorical variable, and no coding categories were treated as exclusive of any other. This permits our coding categories to nest within one another, and this property allows us to logically
construct the definitions of the cognitive strategies out of the definitions of our coding categories.

Analysis. We used a two-step cluster analysis algorithm to find natural clusters formed in the cases that comprise our observational data. Specifically, we looked for clusters that occurred in all of the different populations of cases that we could create by sorting according to gender or test condition. We used the two-step algorithm because it is able to find natural clusters in categorical variable data.
The algorithm looks for cases that have the same coding category values, and creates a preliminary cluster out of any set of cases that have the same values. It then scores a number of different "models" of the clusters identified in first step of the analysis according to their Bayesian Information Criteria (BIC). Importantly, this second step is able to resolve any borderline cases: if a particular case $c$ is similar to those cases in a group of cases $G$ which all have exactly the same values, and $c$ is also different than many other cases that, in the first step, the algorithm did not put into any clusters, the algorithm may then put $c$ into G if the model which places $c$ in G has the best BIC score. However, if the case data is too heterogeneous - as may occur if there are nearly as many clusters as there are cases -then the algorithm in the second step will delete some or all of the preliminary clusters. Because of this, it is also possible that two-step cluster analysis will find no natural clusters in some data sets. (cf. Norušis 2011)
We examined combinations of 11 different coding categories applied to 11 different populations of cases (see below). More explicitly, we looked for a combination of coding categories which the two-step algorithm was able to use to find clusters that (a) occurred in all 11 populations of cases we analyzed and that (b) had a silhouette coefficient greater than 0.6 . We also wanted to identify a set of clusters that (c) were the only natural clusters in all 11 populations of cases.
The 11 populations of cases we analyzed were: all participants, male / female participants, participants in each of the four test conditions (Dilemma Affect, Dilemma NonAffect, Problem Affect, Problem Non-Affect), and participants in each of the types of situations used to construct our test conditions (Dilemma, Affect, Problem, Non-Affect). Thus, we used the two-step algorithm to determine if the coding category "Subject says what the main character should do" picked out any natural clusters in the populations listed above. We also used the cluster analysis algorithm to determine if the two coding categories "subject says what the main character should do" and "subject asks for more information about the vignette" together picked out any natural clusters in the 11 populations of cases above And we also asked the cluster analysis algorithm to determine if the three coding categories consist of the previous two plus "subject uses moral language in their response" together picked out any natural clusters of cases in the 11 populations of cases listed above. And so on.
Thus, we supplied approximately 121 different
combinations of coding categories to the two-step algorithm. As we said, we were searching for clusters of cases that occurred in all 11 populations of cases and which scored a high silhouette coefficient (>0.6). Any such clusters would therefore represent types of responses that were invariant across conditions and populations.

## Results

We found 4 such clusters. Specifically, we found that a combination of 4 coding categories defined four different natural clusters that occurred in all 11 populations listed above. What's more, these 4 clusters were the only natural clusters in 9 of the 11 populations. The coding categories that define these clusters are:

1. Subject says what the main character should do.
2. Subject uses the word "might" or "probably" or a similar word to express hesitation when verbalizing their judgment.
3. Subject says something indicating that they are imagining themselves in the situation of the main character of the vignette (such as "Well, what I would do in that situation is..." or "If it was me there, I think that I...").
4. Subject expresses at least one inference when making their response (such as "if ... then ..." or ".... because ....").

Because these were treated as categorical variables, each of these coding categories can take the only the values "true" or "false". Each of our 492 cases will therefore have a value of "true" or "false" for each of these categories. This means that there are 16 logically possible clusters that the cluster analysis algorithm could have found using these coding categories, although it is also possible that the algorithm find could have found no clusters at all.
Here are the clusters that the algorithm found. Note that each cluster is operationally defined out of logical values for the four coding categories listed directly above:

Simple Reasoning $=$ subject expresses a judgment and expresses an inference, but does not use language indicative of hesitation and does not imagine themself in the position of the main character.
Intuitive Judging = subject expresses a judgment, and does not express an inference, does not use language indicative of hesitation and does not imagine themself in the position of the main character.
Cautious Reasoning $=$ subject expresses a judgment, expresses an inference, and does use language indicative of hesitation, but does not imagine themself in the position of the main character.
Empathic Reasoning = subject expresses a judgment and expresses an inference and imagines themself in the position of the main character, but does not use language indicative of hesitation.

We think that these four natural clusters - or, if you prefer, types of response - represent four cognitive strategies that people use to respond to moral situations. And just to be clear, these are not types of people; they are cognitive
strategies that occurred in all of our test conditions.
Figure 1 presents the proportions of these clusters in our total respondents - that is, for all 492 cases taken together. For this population, the cluster analysis algorithm placed every case into one of the four clusters, meaning that no case was excluded. Importantly, there was no significant variation in the relative proportion of these four clusters across all of the populations of cases that we analyzed. This means that the ratio of reasoners to intuitors was approximately $3: 1$ in all of our populations. It also means that the proportion of the four clusters in, for instance, the male population of cases looks very similar to the proportion of clusters in our total population of cases; for illustration, please see Figure 2.

## Proportion of Four Clusters (or Cognitive Strategies) in Total Population of Cases



Figure 1: The proportion of four natural clusters found in the population consisting of all 492 of our cases. Each of the 492 cases was placed into one of these four clusters. Proportion is expressed as a percentage, rounded to the nearest whole number. (Silhouette measure of cohesion and separation $=0.7$ )

## Proportion of Four Clusters (or Cognitive Strategies) in only Male Cases.



Figure 2: The proportion of four natural clusters found in the population consisting of only our male cases. As in Figure 1, each case was placed into one of these four clusters. Proportion is expressed as a percentage, rounded to the nearest whole number. (Silhouette measure of cohesion and separation $=0.7$ )

As we indicated above, we also found that these clusters were the only natural clusters that were present in 9 of the 11 case populations we analyzed. The two exceptions were the subpopulation of cases in the Problem Non-Affect condition (where these four clusters accounted for $75 \%$ of the natural clusters in the population) and the female respondents (where these four clusters accounted for $85 \%$ of
the natural clusters in the populations). Remember, the algorithm can resolve borderline cases by placing statistically similar though not identical cases together in the same cluster, but borderline cases will not always be resolved in the same way across different populations of cases. This is because the treatment of a borderline depends partly on what the statistical properties of other borderline cases in the population under analysis.
Finally, we would like to report that that the coding category "subject uses moral language" failed to figure in any of the condition invariant natural clusters. We find this result particularly intriguing.

## Impact of the Conditions and Other Objections

A natural worry with our claim that the four cognitive strategies that we observed are condition invariant is that our test conditions simply failed to have any experimentally meaningfully impact on our subjects - even though three of our four test conditions were designed on theoretical grounds to try to push subjects in those conditions towards intuitive responses.
So as a control for this possibility, we analyzed the cases in the different test conditions for any significant differences, and we found several. For instance, subjects in the two affect conditions were more likely to ask our interviewer for information about the vignette than subjects in the two nonaffect conditions ( $x^{2}=6.54, \mathrm{p}=0.0105$ ). We also found that, when we scored the coherency of the reasoning on a 7 point scale derived from a grading rubric used in a critical reasoning course, the coherency of reasoning of the subjects in the Problem Non-Affect condition was significantly higher than all other conditions (e.g., for Problem NonAffect versus Dilemma Affect, $\left.x^{2}=19.05, \mathrm{p}=0.0019\right)$. These data indicate that our test conditions did have different psychological effects, and this speaks to the strength of the clusters we found in our data.
We would also like to speak to the assumption that differences in people's verbal responses can be read without further experimental constructs as evidence of differences in the underlying cognitive processes. This assumption is often implicit in the analysis of experimental data in moral psychology, and it is most prominent in the work of researchers who have taken the view that moral cognition is largely driven by intuition processes (cf. Haidt 2001). Our position is that this assumption is warranted as a premise in an abductive inference for our conclusion - namely, that the best explanation of the differences we observed in our subject's verbal responses is that these differences reflect different underlying cognitive strategies.
In sum, we claim that the four natural clusters we defined above represent four different cognitive strategies that people use to respond to moral situations. Sometimes people are simple reasoners, intuitive judgers, cautious reasoners, or empathic reasoners - and, importantly, these four strategies are used whether or not people are asked explicitly to think about an either/or dilemma, and whether or not they read a vignette designed to induce mild feelings of either disgust or sympathy.

## Discussion

Some of the categories we used in our analysis did not yield any condition invariant clusters, and this result provides an interesting independent confirmation of some of the claims made by the social intuitionist model. The operational definition of "intuitive judger" above is nearly identical to the theoretical definition of an "intuitive judgment" given by Haidt (2001). And because in our experiment there was no a priori reason to think that the intuitive judging cluster would be one of the four clusters found in all of our test conditions, our observation that a large number of our respondents behaved in a way that very nearly exactly satisfies the social intuitionists' definition of "intuition" is therefore evidence of the accuracy of their theoretical definition for the concept. This result, moreover, comports very well with dual-process approaches to cognition.
However, we failed to observe significantly more intuitive judgers in the three conditions designed to induce intuitive moral cognition. The ratio of intuitive judgers to reasoners held steady across all of our test conditions. Remember: participants were given no explicit instruction to reason about the moral situations we read to them; we asked each participant for only a judgment about what the main character in the vignette should do. So the fact that we were unable to "push" subjects in some conditions to rely more frequently on moral intuition is a challenging result to intuition-based models of moral cognition like the social intuitions model and Mikhail's universal moral grammar. Despite our attempts to maximize the likelihood that participants would use intuitive cognitive processes in some of our test conditions, participants were in all of our conditions at least three times more likely to use some kind of reasoning than to use intuition.
This finding is relevant to the methodology of moral psychology. Moral intuition is often defined as the absence of reasoning, and reasoning is a normative ability the manifestation of which varies according to the skill and epistemic context of a subject. By setting the range of permitted reactions to an either/or choice or recording agreement with a proposition on a Likert scale, many moral psychological experiments are automatically designed not to record any reasoning. Yet, these constructs do not cause subjects to not reason during the experiment; they only proactively "screen-off" the expression of any underlying reasoning that may or may not occur. Our experiment was designed to see what subjects would do when this screen was removed, and our findings suggest that reasoning is a very common cognitive strategy used to arrive at moral judgments. But the deeper lesson implied by our findings is that there is more than a single concept of moral intuition employed in contemporary experimental moral psychology. Let an experimental intuition be any judgment recorded in an experiment where the subject is prevented from expressing any reasoning that may or may not occur in the production of the judgments, and let a psychological intuition be any judgment that is not produced on the basis of any immediately prior reasoning. Our findings suggest that many moral psychologists are studying only
experimental intuitions - for exactly the same reason that our "intuitive judging" category might capture only experimental, not psychological, intuitions.

## Acknowledgments

We would like to thank Richard Boyd, Tamar Kushnir, Luca Bonatti, Albert Costa, Jane Dryden, Robbie Moser, Roopen Majithia, and the members of the RICO group at the University Pompeau Fabra. We are grateful for the financial support of both Mount Allison University and Cornell University.

## References

Ciaramelli E, Muccioli, M., Làdavas, E, and di Pellegrino, G. (2007) Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex. Social Cognitive and Affective Neuroscience, 2, 84-92.
Cushman, F.A. and Greene, J.D. (2012) Finding faults: How Moral Dilemmas reveal cognitive structure. Social Neuroscience, 7(3), 269-279
Damasio, A (1994) Descartes' Error: Emotion, Reason, and the Human Brain. Putnam Books.
Greene, J.D., Nystrom, L.E., Engell, A.D., Darley, J.M., and Cohen, J.D. (2004) The neural bases of cognitive conflict and control in moral judgment. Neuron, 44, 389-400.
Haidt, J. (2001) The Emotion Dog and its Rational Tail. Psychological Review, 108, 814-834.
Koslowski, B, and Fedyk, M. (in prep) Online versus person-to-person surveys in moral psychology
Lombrozo, T. (2009) The role of moral commitments in moral judgment. Cognitive Science, 33, 273-286.
Mikhail, J., Sorrentino, C.M., and Spelke, E. (1998) Toward a Universal Moral Grammar. Proceedings of the $12^{\text {th }}$ Annual Conference of the Cognitive Science Society.
Mikhail, J. (2011) Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment. Cambridge University Press.
Mikhail, J. (2007) Universal Moral Grammar: Theory, Evidence, and the Future. TRENDS in Cognitive Science, 11(4), 143-152
Nichols, S. and Mallon, R. (2006) Moral Dilemmas and Moral Rules. Cognition.
Norušis, M.J. (2011) IBM SPSS Statistics 19 Guide to Data Analysis. Prentice Hall.
Schaich Borg, J., Hynes, C., Van Horn, J., Grafton, S., and Sinnott-Armstrong, W. (2006) Consequences, action, and intention as factors in moral judgments: an FMRI investigation. Journal of Cognitive Neuroscience, 18(5), 803-17
Singer, P. (1972) Famine, Affluence, and Morality. Philosophy and Public Affairs, 1, 229-243.
Pinker, S. (2008) The Moral Instinct. New York Times Magazine.
Waldmann, M.R. and Dieterich, J.H. (2007) Throwing a Bomb on a Person versus Throwing a Person on a Bomb: Intervention Myopia in Moral Intuitions. Psychological Science 18(3), 247-253

# Communicative biases shape structures of newly acquired languages 

Maryia Fedzechkina (mashaf@bcs.rochester.edu), T. Florian Jaeger (fjaeger@bcs.rochester.edu)<br>Department of Brain and Cognitive Sciences, University of Rochester<br>Rochester, NY 14627 USA

Elissa L. Newport (eln10@georgetown.edu)<br>Department of Neurology, Georgetown University, Washington, DC 20007 USA


#### Abstract

Languages around the world share a number of commonalities known as language universals. We investigate whether the existence of some recurrent patterns can be explained by the learner's preference to balance the amount of information provided by the cues to sentence meaning. In an artificial language learning paradigm, we expose learners to two languages with optional case-marking - one with fixed and one with flexible word order. We find that learners of the flexible word order language, where word order is uninformative of sentence meaning, use significantly more case-marking than the learners of the fixed word order language, where case is a redundant cue. The learning outcomes in our experiment parallel a variety of typological phenomena, providing support for the hypothesis that communicative biases can shape language structures.


Keywords: Language acquisition; learning biases; language universals; efficient communication.

## Introduction

In his seminal paper, the American linguist Joseph Greenberg (Greenberg, 1963) noticed that the vast majority of patterns that recur in apparently unrelated languages, also known as language universals, take the form of implicational statements: If a language has property $A$, then it will most likely have property $B$. Language universals point towards constraints on the space of structures possible in natural language since some of the theoretically possible feature combinations are cross-linguistically observed more frequently than others.

The nature of these recurrent patterns has sparked a debate in the cognitive sciences: Are language universals due to constraints specific exclusively to language which are not shared by other aspects of human cognitive systems (Chomsky, 1965), or are they due to general cognitive constraints such as constraints on perception, memory and learning (Hawkins, 2004; E Newport, 1981; Slobin, 1973)?

In this work, we explore the long-standing hypothesis that domain-general pressures associated with human communication can shape languages over time (Bates \& MacWhinney, 1989; Slobin, 1977; Zipf, 1949). Support for this hypothesis comes from recent studies that apply mathematical theories of communication to the study of language structures. This work has found that speech has many properties that strike an efficient balance between successful and fast communication (Jaeger, 2010). Recent cross-linguistic studies have further found that languages
across the world share many properties that facilitate such efficient information transfer (S.T. Piantadosi, H Tily, \& E Gibson, 2011; S.T. Piantadosi, H. Tily, \& E. Gibson, 2011; Qian \& Jaeger, 2012).

The study of language universals has primarily relied on typological and diachronic data, which has several major limitations. First, typological and diachronic studies suffer from data sparsity since only a small fraction of 6909 known languages (Lewis, 2009) have been sufficiently documented. This led some researchers to suggest that there is no evidence for language universals once common ancestry and geographical proximity between languages are taken into account (Dunn, Greenhill, Levinson, \& Gray, 2011). Second, typological and diachronic studies do not provide an insight into the mechanism of how the hypothesized constraints come to shape language over time.

The aim of this study is three-fold. First, we set out to provide direct behavioral evidence for the existence of cross-linguistic universals. Second, we investigate the cause of these universals. We ask, in particular, whether some typologically frequent phenomena can be explained by domain-general biases associated with considerations about human communication. Finally, we explore whether learning can provide a potential mechanism through which these biases come to shape language structures.

We employ an artificial language learning paradigm, where learners are exposed to miniature languages designed to have certain properties of interest. This method has been used to study learning biases in adults and children (Culbertson, Smolensky, \& Legendre, 2012; Hudson Kam \& Newport, 2009) and has provided behavioral evidence for typological universals (Christiansen, 2000; Culbertson et al., 2012; Finley \& Badecker, 2008; E. Newport \& Aslin, 2004). Of particular interest is a recent study by Fedzechkina, Jaeger, and Newport (2012), who have used this paradigm to investigate the impact of learners' communicative preferences on language structure and shown that language learners are biased towards efficient case systems.

Here we focus on the correlation in a language between the flexibility of word order and the presence of a case system. It has long been observed that languages with rich case-marking typically allow more word order freedom than languages with no case-marking (Blake, 2001; Sapir, 1921). Languages like Russian or Latin, which allow sentential subjects and objects to be placed in a variety of positions
with respect to each other and the verb, tend to have rich case systems. However, languages that enforce strict order of subject and object (e.g., English and French) typically have no or only rudimentary case-marking.

Additional evidence for the correlation between word order flexibility and the presence of case-marking comes from studies of language change. For instance, Old English allowed the permutation of subject and object while having rich case-marking. This relationship between word order and case-marking substantially changed during the history of English, and Modern English became a language with a fixed word order and no case-marking.

We explore whether this correlation between word order freedom and the presence of a case system in a language can result from a trade-off between the information content of a cue and a the amount of effort necessary to produce this cue. Word order is highly informative of sentence meaning (i.e., grammatical function assignment can be successfully recovered based on word order alone) in a fixed word order language. Case-marking is thus redundant in such language and can be omitted to conserve effort without hindering robust communication. In a flexible word order language, however, word order is less informative of grammatical function assignment, and situations can occur when sentence meaning cannot be successfully recovered based on the linear order of elements alone. The relative lack of informativity of word order is compensated for by casemarking, which, when present, provides crucial information about grammatical function assignment.

We expose learners to languages with optional casemarking that have either fixed or flexible word order. If learners indeed balance the amount of information provided by cues to sentence meaning, we predict that the relative lack of one cue will make it more likely that learners recruit alternative cues. In particular, we expect learners of the flexible word order language to use significantly more casemarkers than the learners of the fixed word order language.

## Experiment

## Participants

Participants in the experiment were monolingual native English speakers recruited from the undergraduate students at the University of Rochester or their age-matched peers from the surrounding community. All participants were compensated $\$ 25$ for their time. Participants were pseudorandomly assigned to one of the two languages: variable word order or fixed word order language (described below). Recruitment continued until the number of participants who successfully learned the language reached 20 in each of the two languages. 52 volunteers participated in the experiment. 12 participants were excluded from the analysis for the following reasons: failure to achieve $65 \%$ accuracy on unambiguous trials during the comprehension test (10 participants in the flexible word order language, see below), computer error (1 participant), or being a bilingual (1
participant). This left 40 participants for analysis, 20 in each of the two languages.

## The Languages

Lexicon Verbs There were four verbs (geed, kleidum, shen, zamper) that denoted simple transitive actions (HUG, KNOCKOVER, ROCK, KICK). All verbs occurred equally frequently in the input overall and with each word order variant allowed by the language.

Nouns There were six nouns (glim, flugit, bliffen, norg, spad, melnawg), all of which denoted male referents (MOUNTIE, CHEF, REFEREE, CONDUCTOR, HUNTER, BANDIT). There were no restrictions on nouns. All nouns occurred equally often as subjects and objects of each of the four verbs.

Case-marker There was one case-marker ('kah') that optionally marked the object of the action.

Grammar There were two language conditions in the experiment.

Fixed word order language did not contain word order variation - subject-object-verb (SOV) occurred in $100 \%$ of the input sentences. The language had optional casemarking $-67 \%$ of the input sentences contained a casemarker that marked the object of the action.

Since grammatical function assignment could be unambiguously identified by word order in this language, case-marking added little information to successful recovery of sentence meaning.

Flexible word order language contained word order variation - subject-object-verb (SOV) and object-subjectverb (OSV) word orders occurred equally frequently in the input. The language contained optional case-marking - 67\% of sentential objects were case-marked regardless of sentence word order.

In this language, word order was uninformative about grammatical function assignment. Case-marking, when present, provided important information about sentence meaning.

Head-final languages were chosen for both language conditions since they are cross-linguistically more likely to have case-marking systems (Greenberg, 1963).

## The Procedure

The procedure builds on the method developed by Hudson Kam and Newport (2009). Participants were trained and tested on one of the two languages during three 30-35 minute visits to the lab spread over three consecutive days, with at most one day between the visits. The same procedure was used on all three visits. During each session, participants were presented with a mixture of training and test blocks that fell into two broad categories: noun training and sentence training.

Noun Exposure and Tests During noun exposure participants were presented with pictures of each of the characters accompanied by their label in the novel language
(12 trials total). Participants were instructed to repeat the names of the characters aloud to help them learn. The initial noun presentation was followed by a noun comprehension test where participants were presented with pictures of two characters and asked to choose the correct match for the name they heard (12 trials total). Feedback was provided after each trial. After completing the noun comprehension test, participants were presented with the noun production test where they were asked to provide the name of the character shown on the screen (6 trials). Feedback on performance was provided after each trial. The three noun training blocks were repeated immediately upon completion of the noun production test. On Day 1 noun exposure and noun production blocks were also presented before the sentence production test. On Days 2 and 3 participants were only given the noun production block before the sentence production test. Noun exposure and comprehension blocks were also shorter on Days 2 and 3 ( 6 trials each).

Sentence Exposure and Tests During the sentence exposure phase, participants viewed short computergenerated videos and heard their descriptions in the novel language. Participants were asked to repeat the sentences out loud to facilitate learning. On all days, exposure sentences were presented in sets of two training blocks (24 trials each). During the first sentence exposure block on Day 1 participants were allowed to replay the videos and sounds as many times as they wanted; replay was disabled for all other exposure blocks throughout the study. Sentence exposure was followed by a sentence comprehension test ( 24 trials total). Participants were shown two videos involving the same action where the order of the doer and undergoer was reversed and were asked to choose the video best matching the sentence they heard. No feedback on performance was provided during the sentence comprehension test. After the sentence comprehension test, participants completed two more sentence exposure blocks and one more sentence comprehension block. Each experimental session ended with a sentence production test (48 trials) where participants were asked to describe a previously unseen video using the provided verb prompt. No feedback on performance was provided.

## Scoring

In the comprehension test, participants' responses were scored as 'correct' if they matched the intended sentence interpretation. This was based only on case-marked (unambiguous) trials for both languages. Participants who failed to achieve $65 \%$ accuracy were excluded from all analyses. The results reported below still hold, however, if these participants are included as well.

In the production test, we scored the word order used in the sentence, the presence of case-marking on the object as well as lexical (using the wrong word for a referent or an action) and grammatical mistakes (using a word order not allowed by the grammar or using the case-marker incorrectly). If the name of only one referent was incorrect
and it was still possible to determine sentence word order, productions were scored as overall correct but containing a lexical error. Such productions were included in the analyses below. Productions containing grammatical errors were excluded from all analyses. The results presented below still hold if productions containing lexical errors are excluded as well.

## Results

Accuracy of Acquisition Both languages were acquired with a high degree of accuracy. On the final day of training learners of the fixed word order language made less than $1 \%$ lexical mistakes and no grammatical mistakes, while learners of the flexible word order language made $1.6 \%$ lexical mistakes and $6.2 \%$ grammatical mistakes in their productions. These data suggest that the task was feasible for our participants.

Word Order Use in Production One way learners of the flexible word order language can ensure robust communication is by fixing word order and dropping casemarking. Thus we first analyzed participants' word order use in the flexible word order condition, asking whether there was a tendency to regularize word order. If learners of the flexible word order language behave just like fixed word order learners who use SOV in all their productions, there will be little reason to expect differential case-marker use between the two language conditions.

Overall, learners of the flexible word order language maintained word order flexibility: There was no word order regularization in participants' productions on any day of training (Day 1: 49\% SOV word order in production, not significantly different from the $50 \%$ input proportion [ $\left(\chi^{2}\right.$ (1)=.15, $\mathrm{p}=.7, n s]$; Day 2: $45 \%$ SOV word order in production, not significantly different from the input [ $\chi^{2}$ $(1)=2.66, \mathrm{p}=.1, n s]$; Day 3: $49 \%$ SOV word order in production, not significantly different from the input $\left[\left(\chi^{2}\right.\right.$ (1) $=.17, \mathrm{p}=.7, n s]$ ).

Case-Marker Use in Production We now turn to our main question: Do learners balance the amount of cues to the meaning of the sentence, recruiting additional cues in those cases when existing cues do not provide enough information to successfully decode the intended meaning?

We used a mixed logit model to predict the use of casemarking in participants' productions based on language condition (flexible/fixed word order language) and day of training $(1,2,3)$ as well as the interactions between these two factors. The model included the maximal random effects structure justified by the data based on model comparison. The results reported below hold when the model with the full random effects structure was used. There was a significant main effect of language (see Figure 1): Learners of the flexible word order language used significantly more case-markers in their productions than the learners of the fixed word order language throughout the experiment ( $\beta=1.45, \mathrm{z}=2.24, \mathrm{p}<.05$ ). There was a significant
interaction between Day 2 of training and language condition ( $\beta=.46, \mathrm{z}=3.4, \mathrm{p}<.001$ ) and Day 3 of training and language condition ( $\beta=.25, \mathrm{z}=2.75, \mathrm{p}<.01$ ). Simple effects test shows that learners of the flexible word order language used significantly more case-markers than the learners of the fixed word order language on Day $2(\beta=1.65, \mathrm{z}=2.5$, $\mathrm{p}<.05$ ) and Day 3 ( $\beta=1.94, \mathrm{z}=2.72, \mathrm{p}<.01$ ) of training.

As expected under our hypothesis, then, learners used significantly more case-marking when they learned a language where word order was uninformative of grammatical function assignment (flexible word order language) as compared to the language where referent-to-grammatical-function assignments were reliably identified by word order (fixed word order language).


Figure 1: Case-marker use in production
What is driving the observed difference in case-marker use between the two language conditions? Under our hypothesis, we expect learners of the fixed word order language to gradually lose case-marking, producing fewer case-markers than the input proportion, since case-marking is a redundant cue to sentence meaning in a language that does not allow word order variability. Given the design of our flexible word order language, learners could take advantage of case-marking to facilitate successful decoding of the intended meaning in two ways. First, learners could regularize case-marking in the language overall and use more case-markers in their own productions than in the input. Alternatively, learners could condition case-marking on word order and use significantly more case-markers with one word order variant than with the other. Both possibilities will increase successful recovery of the intended meaning, but the latter will minimize effort at the same time (since not all sentential objects will need to be case-marked) and would thus be a more efficient option. In the following sections, we explored these predictions in more detail.

## Case-Marker Use in the Fixed Word Order Language

Do learners of the fixed word order language deviate from the input they receive and reduce the amount of casemarking in the newly acquired language? Indeed, they showed a strong tendency to drop case-marking and used
significantly fewer case-markers in their own productions compared to the input starting on the first day of training (Day 1: 50\% case-marking in production, significantly lower than the $67 \%$ input proportion $\left[\left(\chi^{2}(1)=23.51\right.\right.$, $\mathrm{p}<.001]$; Day 2: 45\% case-marking in production, significantly lower than the input $\left[\left(\chi^{2}(1)=40.6, \mathrm{p}<.001\right]\right.$; Day 3: $41 \%$ case-marking in production, significantly lower than the input $\left[\left(\chi^{2}(1)=61.87, \mathrm{p}<.001\right]\right)$.

The behavior of the majority of individual subjects followed our prediction. Out of 20 participants, 14 participants produced fewer case-markers than the input on the final day of training; 8 of these did not use case-marking at all in their own productions; and only 3 participants produced substantially more case-markers than the input proportion (see Figure 2).


Figure 2: Individual preferences in case-marker use in the fixed word order language on the final day of training. The dashed line indicates the input proportion of case-marking.

Case-Marker Use in the Flexible Word Order Language Next we took a closer look at the learning outcomes in the flexible word order language. Do learners increase communicative success by favoring robustness but sacrificing efficiency and regularize case-marking in the language overall? Or do they favor efficiency to achieve the same goal, conditioning case-marking on sentence word order?

Learners of the flexible word order language did not use case-markers significantly more frequently than in the input language (Day 1: 55\% case-marking in production, significantly below $67 \%$ input proportion $\left[\left(\chi^{2} \quad(1)=9.5\right.\right.$, $\mathrm{p}<.01]$; Day 2: $72 \%$ case-marking in production, not significantly different from the input $\left[\left(\chi^{2}(1)=1.78, p=.18\right.\right.$, $n s]$; Day 3: $71 \%$ case-marking in production, not significantly different from the input $\left[\left(\chi^{2}(1)=.89, p=.3, n s\right]\right)$.

There was a significant tendency to condition casemarking on sentence word order throughout the experiment (see Figure 4): Learners overtly marked objects with case significantly more often if sentence word order was OSV ( $\beta=1.11, \mathrm{z}=17, \mathrm{p}<.001$ ). A significantly higher proportion of
object case-marking in OSV sentences compared to SOV sentences was observed on every individual day of training (Day 1: $(\beta=1.53, \mathrm{z}=12.6, \mathrm{p}<.001)$; Day 2 ( $\beta=.9, \mathrm{z}=8.6$, $\mathrm{p}<.001$ ); Day 3 ( $\beta=.91, \mathrm{z}=8.23, \mathrm{p}<.001)$ ).


Figure 3: Individual preferences in case-marker use in the flexible word order language on the final day of training. The dashed line indicates the input proportion of casemarking.


Figure 4: Case-marker use by sentence word order in production (flexible word order condition)

Learning outcomes in each language condition thus support our hypothesis. As expected, learners of the fixed word order language tend to gradually drop case-marking as they acquire the novel language. Learners of the flexible word order language maintain case-marking in the language they acquire and make efficient use of it by conditioning case-marking on sentence word order.

## Discussion

Our results provide experimental evidence that some of the typological properties observed cross-linguistically can stem from learners' biases towards communicatively efficient systems. We found that language learners have a bias to balance the amount of information provided by two cues to sentence meaning (case and word order) and tend to use an
additional cue significantly more often in those cases when the other cue does not provide sufficient information to successfully parse the sentence.

Importantly, our results parallel synchronic and diachronic typological data from natural languages. We presented learners with languages that contained the same amount of case-marking, but the learning outcomes consistently differed depending on the amount of word order flexibility allowed in the input language. While learners of the flexible word order language retain casemarking in their own productions, as do Russian, Latin and other free word order languages, learners of the fixed word order language tend to lose case-marking as they acquire the new language, mimicking fixed word order languages such as English and French.

The learning outcomes in our experiment also parallel diachronic phenomena such as the evolution of English from Old English to Modern English. Our data, however, does not speak to whether word order fixing was a result of the loss of case-marking or whether case-marking became a redundant cue and was lost after English word order became fixed for independent reasons. Under our hypothesis, both processes will yield the same outcome.

We found that learners of the flexible word order language did not regularize case-marking uniformly across the two possible word orders. Instead they restructured the language to make efficient use of case-marking by conditioning it on sentence word order, using significantly more case-markers when sentence word order was OSV.

Why do learners preferentially case-mark objects in sentences with OSV word order and not the other way around? This behavior could be reflective of a cognitive bias to mark the atypical, somewhat resembling the crosslinguistically common phenomenon known as 'word order freezing' (Lee, 2001). Many flexible word order languages (Turkish, Russian, German, Hindi, Japanese, etc.) enforce default word order for sentences in which case-marking is uninformative (e.g., in the absence of case-marking or in the presence of case syncretism) and require direct objects to be overtly case-marked if sentence word order deviates from the dominant one. For instance, the Russian example in (1) with case-syncretism can only be interpreted as SVO, although Russian generally allows both SVO (dominant) and OVS orders (non-dominant).

$$
\begin{align*}
& \text { Mat' ljubit doch'. }  \tag{1}\\
& \text { Mother-NOM/ACC loves daughter-NOM/ACC } \\
& \text { 'Mother loves daughter' }
\end{align*}
$$

Even though SOV and OSV are equally likely both in the input and in participants' productions in our experiment, OSV word order is typologically rare and is almost always a non-default word order cross-linguistically and thus might attract a higher proportion of case-marking.

The second possibility is that learners prefer to put more informative cues earlier in the sentence to allow for faster processing, as has been hypothesized by Hawkins (2004)
and Nichols (1986). Preliminary support for this hypothesis comes from Fedzechkina et al. (2012), who found that learners of a language with object case-marking preferred to use more case-marking in OSV sentences, whereas learners of a language with subject case-marking used significantly more case-marking in SOV sentences. That is, in both cases, learners preferred case-marking on the sentence-initial argument, thereby providing the disambiguating cue as early as possible in the sentence.

## Conclusions

We used an artificial language learning paradigm to ask whether language structures are shaped by communicative biases. We find that the cross-linguistically common correlation between word order flexibility and the presence of case-marking can be at least partially explained by domain-general learning biases stemming from a preference to balance the amount of information provided by the cues to sentence meaning.

## Acknowledgments

We thank T. Stanley and S. Krishnan for their help with data collection and data coding; C. Anzalone and G. Smith for their help with data coding; C. Donatelli, I. Minkina and A. Wood for their help with video stimuli. This work was supported in part by NSF grants BCS-0845059 and IIS1150028, as well as an Alfred P. Sloan Fellowship to TFJ, and by NIH grant DC00167 and HD037082 to ELN.

## References

Bates, E., \& MacWhinney, B. (1989). Functionalism and the competition model. The crosslinguistic study of sentence processing, 3-73.
Blake, B. J. (2001). Case. Cambridge: CUP.
Chomsky, N. (1965). Aspects of the Theory of Syntax (Vol. 119): The MIT press.

Christiansen, M. H. (2000). Using artificial language learning to study language evolution: Exploring the emergence of word order universals The Evolution of Language: 3rd International Conference.
Culbertson, J., Smolensky, P., \& Legendre, G. (2012). Learning biases predict a word order universal. Cognition, 122, 306-329.
Dunn, M., Greenhill, S. J., Levinson, S. C., \& Gray, R. D. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. Nature, 473(7345), 79-82.
Fedzechkina, M., Jaeger, T. F., \& Newport, E. (2012). Language learners restructure their input to facilitate efficient communication. Proc Natl Acad Sci USA.
Finley, S., \& Badecker, W. (2008). Substantive biases for vowel harmony languages. Paper presented at the Proceedings of WCCFL 27.
Greenberg, J. (1963). Some universals of grammar with particular reference to the order of meaningful
elements. In J. Greenberg (Ed.), Universals of human language (pp. 73-113): MIT Press.
Hawkins, J. A. (2004). Efficiency and Complexity in Grammar. Oxford: Oxford University Press.
Hudson Kam, C., \& Newport, E. (2009). Getting it right by getting it wrong: When learners change languages. Cognitive Psychol, 59(1), 30-66.
Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. Cognitive Psychol, 61, 23-62.
Lee, H. (2001). Markedness and word order freezing. In P. Sells (Ed.), Formal and empirical issues in optimality theoretic syntax (pp. 63-128). Stanford: CSLI.
Lewis, M. (Ed.). (2009). Ethnologue: Languages of the world (16th edition ed.). Dallas, TX: SIL International.
Newport, E. (1981). Constraints on learning and their role in language acquisition: Studies of the acquisition of American Sign Language. Lang Scie, 10, 147-172.
Newport, E., \& Aslin, R. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. Cognitive Psychol, 48(2), 127-162.
Nichols, J. (1986). Head-marking and dependent-marking grammar. Language, 56-119.
Piantadosi, S. T., Tily, H., \& Gibson, E. (2011). The communicative function of ambiguity in language. Cognition, 122, 280-291.
Piantadosi, S. T., Tily, H., \& Gibson, E. (2011). Word lengths are optimized for efficient communication. Proc Natl Acad Sci USA, 108(9), 3526.
Qian, T., \& Jaeger, T. F. (2012). Cue Effectiveness in Communicatively Efficient Discourse Production. Cognitive Sci, 36, 1312-1336.
Sapir, E. (1921). Language: An introduction to the study of speech. New York: Harcourt, Brace.
Slobin, D. I. (1973). Cognitive prerequisites for the development of grammar. First Language Acquisition: The Essential Readings.
Slobin, D. I. (1977). Language change in childhood and in history. In J. Macnamara (Ed.), Language Learning and Thought (pp. 185-214). New York: Acad. Press.
Zipf, G. K. (1949). Human Behavior and the Principle of Least Effort. Cambridge, MA: Addison-Wesley.

# Regularization behavior in a non-linguistic domain 

Vanessa Ferdinand (v.a.ferdinand@sms.ed.ac.uk), Bill Thompson (bill@ling.ed.ac.uk), Simon Kirby (simon@ling.ed.ac.uk), Kenny Smith (kenny@ling.ed.ac.uk)<br>Language Evolution and Computation Research Unit<br>School of Philosophy, Psychology \& Language Sciences, University of Edinburgh<br>Dugald Stewart Building, 3 Charles Street, Edinburgh, EH8 9AD, UK


#### Abstract

Language learners tend to regularize unpredictable variation and some claim that is due to a language-specific regularization bias. We investigate the role of task difficulty on regularization behavior in a non-linguistic frequency learning task and show that adults regularize variable input when tracking multiple frequencies concurrently, but reliably reproduce the variation they have observed when tracking one frequency. These results suggest that regularization behavior may be due to domain-general factors, such as memory limitations.


Keywords: frequency learning; regularization; probability matching; Bayesian models;

## Introduction

Languages contain very little unpredictable variation (Chambers et al., 2003) and language learners tend to regularize the inconsistent input they encounter (Reali \& Griffiths, 2009; Hudson Kam \& Newport, 2009, Smith \& Wonnacott, 2010). For example, English contains two forms of the indefinite article $a$ and $a n$, but a deterministic rule (based on the initial phoneme of the following noun) governs the use of these two variants. Why are languages regular, and what drives learners to eliminate free variation in language? Some have suggested that we come to the task of language learning with the expectation that languages are regular and that this expectation takes the form of a language-specific innate bias (Bickerton, 1984; DeGraaff, 1999; Lumsden, 1999; Becker \& Veenstra, 2003). Others claim that linguistic regularization can be explained by domain-general learning mechanisms, such as the effects of memory limitations on the type of variation that learners produce (Hudson Kam \& Newport, 2005, 2009; Hudson Kam \& Chang, 2009). Hudson Kam and Newport $(2005,2009)$ have shown that children tend to regularize free variation, whereas adults maintain it by probability matching, and attribute this difference to children having lower working memory capacity than adults. Newport (1990) demonstrated that children have more of a limited ability to learn from inconsistent input and Hudson Kam and Chang (2009) showed that adults probability matched more when word retrieval was made easier and regularized more when it was difficult, further corroborating their claim that memory limitations can lead to regularization, although see Perfors (2012) for an account of restricted memory encoding that does not lead to regularization.

A similar effect of memory limitations can be found in a non-linguistic tasks. In a study with adults, Kareev et al. (1997) reported an effect of individual differences in working memory capacity (as determined by a digit-span test) on participants' perception of the correlation of two probabilistic
variables. Participants with lower capacity overproduced the most common variant, whereas participants with higher capacity did not. Regularization is also modulated by the number of variables in a task; adults regularized slightly more when predicting which of three lights will flash next than when predicting for two lights (Gardner, 1957).

In this paper, we explore the effect of tracking single versus multiple frequencies on the regularization behavior of adults in a non-linguistic task. We show that participants probability match when tracking a single frequency, but regularize when tracking six frequencies concurrently. Because concurrent frequency learning is a prominent aspect of language learning (Saffran, Alin \& Newport, 1996), and also elicits regularization in a non-linguistic task, this is consistent with a domain-general account of the observed regularization bias in language, possibly attributable to limited working memory.

## Frequency learning experiment

Participants 381 participants were recruited via Amazon's Mechanical Turk crowdsourcing platform and completed our experiment online. 37 participants were excluded on the basis of the following criteria: failing a color vision test (2), self-reporting the use of a pen or pencil during the task (14), not reporting their sex or age (2), or having previously participated in any of our experiments, as determined by their user ID with MTurk (19). More participants were recruited than necessary with the expectation that many would be excluded by these criteria. Once the predetermined number of participants per condition was met, data from the last participants was excluded, totaling 24 participants across all conditions and tasks. All excluded participants received the full monetary reward for the task. The average monetary reward per participant, converted to an hourly rate, was $\$ 2.64$. Of the final 320 participants, 184 are female, and the mean age is 36 $(\min =18, \max =69)$, with a standard deviation of 12 years.
Materials The experiment was coded up as a java applet that ran in the participant's web browser in a 600x800-pixel field. Photographs of 6 different containers (a box, pouch, jar, bowl, bucket, and basket) and graphically generated images of marbles in 12 different colors (blue, orange, brown, grey, black, yellow, red, teal, olive, pink, purple, and lime) served as stimuli.

One-item task This experiment consisted of a training phase in which participants observed a series of 10 marble draws from a bag, and a testing phase in which participants were asked to produce another several likely draws from the


Figure 1: Each pane displays the percentage of participants that responded with a given output frequency of the minority marble $(m)$ during testing. Columns are the input ratio of $m: M$ during training. Dashed lines mark the input frequency of $m$. In the one-item task, participants probability matched, reproducing the input ratio with high fidelity. This task was between-subjects; each participant was trained on one input ratio only. In the six-item task, participants were more likely to regularize than to reproduce the input ratio. This task was within-subjects; each participant was trained on all six input ratios concurrently.


Figure 2: Training and testing trials for the six-item task.
same bag. In each training trial, a picture of the bag was displayed for 1000 milliseconds and then a marble (blue or orange) appeared over the bag for 2000 milliseconds. There were 10 training trials, with no break between trials. In each testing trial, the bag was displayed with the two marble colors below. Participants mouse clicked on a marble to make their choice of one draw from the bag. Their choice was displayed above the bag for 2000 milliseconds and then the next testing trial began. There were 10 testing trials with no breaks between trials. Locations (left or right) of the blue and orange marbles were held constant across test trials for each participant, but counterbalanced across participants.

A fixed ratio of blue to orange marbles was shown in the training phase. Each participant was randomly assigned to one of 6 training conditions based on this ratio. The color of the training ratio's minority marble ( m ) and majority marble $(M)$ was counterbalanced across participants. All possible ra-
tios of $m: M$ were tested and will be referred to as the $0: 10$, 1:9, 2:8, 3:7, 4:6, and 5:5 conditions. 192 participants took part in this task, with 32 in each condition.

Six-item task This task is based on the word frequency learning task from Reali and Griffiths (2009). Participants observed 10 marble draws each from six different containers, totaling 60 marble draws (see Figure 2). Each container was associated with 2 unique marble colors ( 12 unique marble colors were therefore used). Training and testing trials were identical to the one-item task. Each container was uniquely associated with one of the possible ratios specified by condition $0: 10,1: 9,2: 8,3: 7,4: 6$, and 5:5 above. Thus, the six-item task is a within-subject version of the one-item task, with the addition that training and testing trials from all six conditions are interleaved. Assignments of a ratio and marble colors (in predefined color pairs) to each container was randomized per participant. 64 participants took part in this task. Two additional versions of this experiment were also run; one where all 6 bags were in condition 0:10 (each container was mapped to one color only) and one where all 6 containers were in condition 5:5. Each of these versions was completed by 32 new participants.

## Experiment results

Participants in the six-item task were more likely to regularize their responses per container than participants in the oneitem task. Here, we refer to regularization as the production of a more extreme ratio than that observed during training,


Figure 3: Difference in mean entropy scores between tasks, for each input ratio. Each participant's sequence of marble draws during testing was converted into an entropy score. Lower scores denote greater regularity within a response. Participant responses were significantly more regular in the six-item task than in the one-item task for input ratios 3:7, 4:6, and 5:5. Error bars show the standard error of the mean.
where $0: 10$ and 10:0 are the most extreme ratios and 5:5 is the least extreme. The distributions of participant responses are shown in Figure 1. Each pane displays the percentage of participants that responded with a given output frequency of $m$, per input frequency and per task. In the one-item task, participants probability matched; the mode of the population is on the input frequency of $m$, meaning that the most common response was perfect reproduction of the ratio observed during training. In the six-item task, visual inspection suggests that participants did not reproduce the training ratios with as high fidelity. Most participants regularized by overproducing the majority marble (all mass in the bars to the left of the dotted line) and a large number of responses are fully regular, meaning the output frequency of $m$ is 0 or 10 .

To better assess the different degrees of regularization between tasks, we calculated the entropy of each participant's sequence of test choices. This quantifies the amount of variation (in bits) with a value between 0 and 1 ; where 0 denotes a completely regular sequence (i.e. a series of all blue marble draws) and 1 denotes a maximally variable sequence (i.e. a series of 5 blue and 5 orange draws, in any order). This allows us to refine our definition of regularization as the overproduction of one marble, such that the entropy of the participant's testing choices is lower than that of their training observations. The mean entropy scores of participant responses per input frequency are shown in Figure 3.

A linear mixed effects regression analysis showed a significant effect of task on entropy scores, $t(34)=-7.226, p<$ .001, and a significant effect of input frequency on entropy scores, $t(34)=-10.832, p<.001$. This means the two tasks elicited different amounts of regularity within participants’

## Six-item task <br> 

Figure 4: Distribution of participant responses for two additional versions of the six-item task, where all items contained the same input ratio of $m: M$. One group of participants was trained on all 0:10 ratios and another group was trained on all 5:5 ratios.
responses and that participants' responses were modulated by training frequencies; they noticed differences in the input frequencies and this affected their responses. A significant interaction of task and input frequency on entropy scores was also obtained, $t(34)=4.570, p<.001$; participants responded differently to different input frequencies, and this pattern of responses also differed by task.

There was a significant difference in mean entropy scores between tasks for input frequencies 3:7, 4:6, and 5:5 ( $W=$ $1427.5, p=.001 ; W=1714, p<.001 ; W=1585.5, p<$ .001 ), respectively. ${ }^{1}$ The difference in mean entropy between tasks was not significant for input frequencies $0: 10$, $1: 9$, and $2: 8(W=894, p=.228 ; W=1184.5, p=.192 ; W=$ $1264, p=.054^{2}$ ), respectively.

Two additional experiments were conducted to explore the possibility that regularization in the six-item task is due to interference between containers, such that ratios learned for one container get confused with ratios learned for another container. We eliminated this type of interference by training participants on 6 containers with identical ratios. Figure 4 shows participant responses when trained on all 0:10 ratios (left) and all 5:5 ratios (right). The average entropy for the all $0: 10$ task is significantly lower than that of the $0: 10$ condition in the six-item task ( $W=5061, p=.004$ ), but not significantly different than the $0: 10$ condition in the one-item task ( $W=2900.5, p=.466$ ). Tracking multiple 0:10 ratios is no different than tracking one 0:10 ratio, but it is different from tracking one $0: 10$ ratio concurrently with other ratios. This means interference may account for the errors participants make in the original six-item task when producing draws for the container they observed as $0: 10$. However, for the all 5:5 task, the average entropy was not significantly different from

[^59]the $5: 5$ condition in the six-item task $(W=5892.5, p=.617)$. Participants still produced 0:10 and 10:0 responses in the absence of observing these ratios during training. Therefore, interference may account for some of the differences between the one-item and six-item tasks, but this isn't the sole cause of the regularization behavior observed in the six-item task.

## Frequency learning models

What cognitive processes cause regularization? So far our analyses have quantified the difference in regularity between participants' training and testing responses. In this section, we turn our focus to an internal force that can affect a learner's behavior; an inductive bias favoring certain ratios of marbles.

## Bayesian model

Bayesian models provide a way to quantify inductive biases and understand their effect on behavior. We fit a betabinomial Bayesian sampler model to participants' responses, following Reali and Griffiths (2009), and ask what prior expectation for regularity a Bayesian rational learner would need to have in order to produce the data that our participants produced.

A Bayesian rational learner uses Bayes' rule, $P(h \mid d) \propto$ $P(d \mid h) P(h)$, to infer what proportion of marbles generated the draws that they observed. Here, each proportion is a hypothesis and the observed draws are the data. Bayes rule combines the prior probability of a hypothesis, $P(h)$, with the likelihood of the data under that hypothesis, $P(d \mid h)$, to arrive at a posterior probability of that hypothesis given the data, $P(h \mid d)$. The prior is a beta distribution over all hypotheses, $\operatorname{Beta}\left(\frac{\alpha}{2}, \frac{\alpha}{2}\right)$, where the parameter $\alpha$ determines whether the learner expects to see regular draws or variable draws. A learner with $\alpha<2$ will tend to regularize their productions, a learner with $\alpha=2$ is unbiased toward any particular proportion of draws, and a learner with $\alpha>2$ is biased towards variability in draws. The likelihood of drawing $N$ marbles in ratio $k:(N-k)$ from a container of marbles in proportions $p:(1-p)$ follows a binomial distribution (Equation 1).

$$
\begin{equation*}
P(k \mid p, N)=\binom{N}{k} p^{k}(1-p)^{N-k} \tag{1}
\end{equation*}
$$

Once the posterior probability over all hypotheses has been determined, the learner must choose a hypothesis to generate testing responses from. We take the case where learners sample a hypothesis from the posterior distribution, and then sample data from this hypothesis according to its likelihood (as if the learner were randomly drawing marbles from the hypothesized proportion, with replacement, as in Equation 1).

This model defines the probability of generating all testing proportions (output states) from all training proportions (input states) and can be visualized as a transition matrix between all possible states in the system. Because our experiment covers all possible training proportions for 10 draws from a bag, we can also construct an empirical transition matrix from participant responses in each task. From here on,
we switch to visualizing our data in terms of marble $1\left(m_{1}\right)$ and marble $2\left(m_{2}\right)^{3}$. Figure 5 (top row) shows the two empirical transition matrices and three model matrices for different values of the prior parameter $\alpha$. Each value of $\alpha$ defines a unique transition matrix, and thus a unique pattern of behavior. For example, if a Bayesian learner observes 1 draw of $m_{1}$ and 9 of $m_{2}$, and if their prior is $\alpha=0.01$, they are most likely to produce 0 draws of $m_{1}$ and 10 of $m_{2}$, regularizing their productions. If their prior is $\alpha=2$, they are most likely to produce 1 draw of $m_{1}$ and 9 of $m_{2}$, probability matching their productions. And if their prior is $\alpha=10$, they are most likely to produce 3 draws of $m_{1}$ and 7 of $m_{2}$, increasing variation in their productions. Thus, the prior used here intuitively captures a range of human behaviors in frequency learning.

The model fitting task at hand is to determine which model transition matrix most resembles the empirical transition matrix, by assigning the most likelihood to the empirical data. The prior associated with the best-fit model is the one that best explains participant behavior and gives us an idea of what biases our participants may have.

The best-fit bias in the one-item task is $\alpha=1.55$ with a $\log$ likelihood of -413 , which is equivalent to correctly predicting $20 \%$ of participant responses in this task ${ }^{4}$. This prior shows an expectation for a slight amount of regularity in the data set. For the six-item task, the best-fit bias is $\alpha=1.21$ with a log likelihood of -1186 , equivalent to $9 \%$ response prediction. This prior shows a stronger bias toward regularity in the six-item task than in the one-item task.

Prediction percentages are lower for the six-item task because participant responses are more variable in the this task than in the one-item task. Only deterministic processes (with one output per input) can be predicted with $100 \%$ accuracy. The ceiling on model prediction for each task was determined by fitting each data set to itself, yielding a maximum of $32 \%$ accuracy for the one-item task and $16 \%$ accuracy for the sixitem task. Relative to these ceilings, the best-fit models account for $61 \%$ and $56 \%$ of participant responses in the oneitem and six-item tasks, respectively.

## Bootstrap model

An input-based random sampling model was also fit to the data. This model defines the transition matrix that would be obtained if participants produced their testing responses by randomly sampling 10 draws from their training observations, with replacement. In this case, each row would be a binomial where $p$ equals the training proportion of $m_{1}$. It is important to note that this transition matrix defines the dynamics of drift in one generation and may be used as a baseline for the loss of variation that can occur in the absence of a regularization bias.

[^60]

Figure 5: Transition matrices (top row) and their associated stationary distribution (bottom row) for the experimental results of the two frequency learning tasks, and for the Bayesian model showing three example bias strengths $(\alpha=0.01,2,10)$. Transition matrices give the probability of moving from each input frequency (the number of training trials showing marble 1) to each output frequency (the number of testing trials in which participants produced marble 1) ${ }^{3}$. The stationary distribution shows how often the transition matrix will produce each output frequency of marble 1.

For this model, the log likelihood of the one-item task data is -259 , equivalent to $25 \%$ response prediction, and is a better fit than the best-fit beta-binomial sampler model ${ }^{5}$. Thus, of the models explored in this paper, drift provides the best account of our participants' probability matching behavior. However, a repeated measures Monte Carlo test shows that the standard deviation among participant output entropies in the one-item task data are significantly lower than that obtainable by drift: $p=.04, p=.03, p=.01, p=.003$, for conditions $2: 8,3: 7,4: 6,5: 5$, respectively. Although these data are wellaccounted for by the drift model, they still show a quantitative difference in standard deviation, meaning that the forces behind probability matching are not truly isomorphic to drift. As for the six-item task, the log likelihood is -1076 , equivalent to $6 \%$ response prediction. Here, the sampler model with a bias toward regularization is still the better fit.

## Null model

This model is the transition matrix that would be obtained if participants were randomly sampling from the two testing choices each trial (i.e. not engaging in the task). Here, every row would be a binomial distribution where $p=0.5$. For this model, the log likelihood of the one-item task data is -604 , equivalent to $4 \%$ response prediction. For the six-item task, the $\log$ likelihood is -1630 , equivalent to $1 \%$ response prediction. Of all models considered, this is the worst fit for both tasks, meaning that participants are not likely to be randomly sampling from their testing choices.

[^61]The results of these model fits strongly suggest that participants in the six-item condition are not just performing poorly at reproducing their training proportion, but they are regularizing their responses in a way that can not be accounted for by random errors.

## Learning biases and long-term behavior

In addition to comparing the transition matrices, which describe the behavior of one generation of learners, we can also look at the long-term behavior of the system, which is described by the stationary distribution of the transition matrix (Figure 5, bottom row). This distribution tells us what percent of the population we would expect to see in each state, after an arbitrarily large number of generations, if the output state of one learner served as the input state to another. Griffiths and Kalish (2007) have shown that the stationary distribution mirrors the prior distribution over hypotheses for the Bayesian sampler model utilized here. The stationary distributions of the empirical transition matrices are most interesting because these would be an estimate of our participants' regularization bias (the prior) if they were Bayesian sampler learners ${ }^{6}$. In line with this interpretation, the stationary distribution of the six-item task closely resembles that of its best-fit Bayesian model, which has a beta distribution $\operatorname{Beta}(0.605,0.605)$. However, the stationary distribution of the one-item task does not resemble that of its bestfit Bayesian model, which has a u-shaped beta distribution Beta( $0.775,0.775$ ). In general, the Bayesian model is a good fit to participant behavior in the six-item task, but does not account very well for participant behavior in the one-item task.

[^62]A close examination of the model's transition matrices and stationary distributions shows that probability matching behavior with a low standard deviation is not within this model's range of behavior.

## Discussion

We have shown that learning a single versus multiple frequencies modulates participants' regularization behavior in a non-linguistic task. When participants tracked the frequency associated with a single item, they probability matched; reproducing the variation they had observed with high fidelity. However, when tracking multiple frequencies concurrently, participants regularized their responses, usually by overproducing the most common variant.

A beta-binomial Bayesian sampler model was fit to the results of each task and showed a stronger prior bias toward regularization in the six-item task than in the one-item task. Strictly speaking, the prior represents the inductive bias of the learner, and participants should come to a marble-drawing task with a particular expectation about the ratios of marbles in containers, regardless of the difficultly of the task. The fact that we find different best-fit priors according to different task demands means that we are not revealing the inductive bias of our participants, per se, but a composite picture that characterizes more than one cognitive constraint. At least one constraint that is sensitive to task demands should be added to the model, such as a memory constraint that disproportionally forgets lower-frequency observations. Such an addition could free up the prior to more accurately reflect participants’ inductive bias. This raises a point of caution in comparing inductive biases across domains without controlling for task demands, since task demands can modulate bias strengths.

Our modeling results also suggest that human probability matching and regularization behavior do not lie on a simple continuum that can be captured by the prior alone. Although the Bayesian model accounted well for our participants' regularization behavior, it failed to account for the restricted variance of probability matching. Participants may be trying to produce a representative sample of draws, where the most likely response is the training ratio itself. Such a parameter might lead to high-fidelity reproduction of the training proportion under low memory constraints only.

If memory constraints are the cause of the regularization bias revealed when learning the frequencies of marbles in several containers, then this same domain-general factor may be the cause of regularization in tasks naturally characterized by concurrent frequency learning, such as language learning.

## Acknowledgements

Special thanks to Tom Griffiths and Luke Maurits for feedback. This research was supported by the University of Edinburgh's College Studentship, the SORSAS award, and the Engineering and Physical Sciences Research Council.

## References

Becker, A., \& Veenstra, T. (2003). The survival of inflectional morphology in French-related creoles. Studies in Second Language Acquisition, 25, 283-306.
Bickerton, D. (1984). The language bioprogram hypothesis. Behavioral and Brain Sciences, 7, 173-221.
Chambers, J., Trudgill, P., \& Schilling-Estes, N. (2003). The handbook of language variation and change. Blackwell, Malden, MA.
DeGraaff, M. (1999). Creolization, language change, and language acquisition: an epilogue. In M. DeGraaf (Ed.) Language creation and language change: creolization, diachrony, and development. Cambridge, MA: MIT Press.
Gardner, A. (1957). Probability-learning with two and three choices. The American Journal of Psychology, 70(2), 174-185.
Hudson, C., \& Newport, E. (2005). Regularizing unpredictable variation: the roles of adult and child learners in language formation and change. Language Learning and Development, 1(2), 151195.
Hudson Kam, C., \& Newport, E. (2009). Getting it right by getting it wrong: when learners change languages. Cognitive Psychology, 59, 30-66.
Hudson Kam, C., \& Chang, A. (2009). Investigating the cause of language regularization in adults: memory constraints or learning effects? Journal of Experimental Psychology, 35(3), 815-821.
Griffiths, T. L., \& Kalish, M. L. (2007). Language evolution by iterated learning with Bayesian agents. Cognitive Science, 31, 441480.
Kareev, Y., Lieberman, I., \& Lev, M. (1997). Through a narrow window: sample size and the perception of correlation. Journal of Experimental Psychology, 126(3), 278-287.
Lumsden, J. S. (1999). Language acquisition and creolization. In M. Degraaf (Ed.) Language creation and language change: creolization, diachrony, and development. Cambridge, MA: MIT Press.
Newport, E. (1990). Maturational constraints on language learning. Cognitive Science, 14, 11-28.
Perfors, A. (2012). When do memory limitations lead to regularization? An experimental and computational investigation. Preprint submitted to Elsevier.
Reali, F., \& Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. Cognition, 111, 317-328.
Reali, F., \& Griffiths, T. L. (2010). Words as alleles: connecting language evolution with Bayesian learners to models of genetic drift. Proc. R. Soc. B, 277, 429-436.
Saffran, J., Aslin, R., \& Newport, E. (1996). Statistical learning by 8 -month-old infants. Science, 274(5294), 19261928.

Smith, K., \& Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. Cognition 116, 444-449.

# Rapid Learning of Morphological Paradigms 

Sara Finley (sara.finley@waldorf.edu)<br>Department of Psychology, Waldorf College<br>Forest City, IA, 50436, USA

Elizabeth Wiemers (wiemerse@net.elmhurst.edu)<br>Department of Psychology, Elmhurst College<br>Elmhurst, IL, 60126, USA


#### Abstract

The present study presents a novel paradigm for testing the ability for adults to rapidly learn novel morphological categories in the wake of irrelevant information: specifically number markings intermixed with irrelevant gender cues. Using an artificial language learning paradigm, participants were exposed to picture-sound pairs in which pictures of animals varied by number (singular, dual and plural), but with irrelevant gender information intermixed with the exposure items (masculine, feminine and neuter). Auditory stimuli were presented in CVCVCV forms (e.g., [zovabu]) in which the first two syllables denoted the animal (e.g., [zova] for snail) and the final syllable denoted number. (e.g., [bu] for single). Results revealed that participants were able to learn which category the suffix endings referred to, based on a two-alternative forced-choice generalization task. Implications for the learning of complex paradigms are discussed.


Keywords: statistical learning, number, morphology.

## Introduction

Languages are governed by complex sets of rules in which sentences are formed through a systematic combination of words. The rule-governed nature of sentences relies heavily on the use of morphological rules and syntactic categories. Words of the same syntactic categories (e.g., nouns and verbs) share similarities beyond meaning. For example, while verbs typically describe actions, in English verbs frequently follow a noun and precede a prepositional phrase (as in The dog sat on the carpet). In addition, syntactic categories share morphological properties, such as tense, gender and number. These morphological indicators are often present as prefixes or suffixes. The specific phonological form of these morphemes appears to be arbitrary in many languages, as in the use of $/-\mathrm{s} /$ as plural in English. However, there is some evidence that languages with complex systems of gender and number morphology such as French, German, and Hebrew may show signs of systematicity within subcategories (Brooks, Braine, Catalano, and Brody, 1993). In this paper we explore how adults are able to learn complex systems of morphology, and whether learners are sensitive to differences in the arbitrariness of morphological patterns.

This study specifically looks at learning of morphologically complex words. Morphemes are parts of words that do not break down into smaller segments with meanings. For example, the word flying contains two morphemes: fly and -ing. Morphology plays an important role in language learning (as well as in learning linguistic categories) because morphology involves both the form of the morpheme (e.g., -ing) as well as its meaning (i.e., progressive). Because morphological forms are often bound - attached to the stems of each word in the category, learners must be able to recognize morphemes within complex words. In order to learn the morphology of one's language, the learner must be able to separate words in terms of their morphological parts. This ability is referred to as morpheme segmentation.

Because all morphemes involve form and meaning, there is a question as to when both aspects of the morpheme are learned. Given that infants are exposed to complex words in speech before they know the meaning of many words, it is likely that infants are able to segment morphological information without semantics. Studies have shown that by 15 months, infants can use distributional cues to learn nonadjacent dependencies, which are necessary for learning syntactic categories (Gomez \& Maye, 2005), and by 18 months, can begin to acquire categories (Mintz, Newport, \& Bever, 2002). Gomez and Gerken (2000) suggest that some learning of categories occurs at the very earliest stages of life. These studies suggest that morpheme segmentation can occur without knowledge of the specific meaning of the words. In addition, these results also suggest the possibility that early learners are equipped with biases to learn linguistic systems using distributional cues. Because languages tend to show broad similarities with respect to the nature of complex morphological rules, it is possible that languages evolved to accommodate biases within the learner.

The nature of learning biases in young infants raises the question as to why there are differences between child and adult language learners. Any theory of learning biases, innate or otherwise, must explain how biases change (and remain the same) over time, in order to explain why children and adults show differences in language learning strategies (Newport, 1990). One hypothesis is that adults have lost the abilities for acquisition that children have (MacWhinney, 1983), but in some cases of artificial language learning, adults outperform children. Specifically, Braine, Brody, and Brooks (1990) showed higher rates of learning for adults
compared to children when learning novel suffix endings. Another hypothesis suggests adults have learned new methods for acquisition that override the initial learning biases (McWhinney, 1983).

While there are clear differences between children and adults, there is evidence that adults still show biases towards certain morphemes. For example, several studies have shown that children learn suffixes more quickly than adults, and adults learn prefixes more quickly than children (Frigo \& McDonald, 1998; MacWhinney, 1983), but research has also shown that adults can parse suffixes with no additional distributional cues (Finley, 2010). In addition, Finley and Newport (in prep) showed that adult learners are biased against typologically infrequent morphological patterns such as infixation. By studying when adults are able to learn patterns that deviate from their native language (and when they cannot), we can better understand the biases that exist for language learning, as well as to better understand how adults can better achieve native-like competence in learning a second language. If adults show biases for particular patterns that are common, but against patterns that are rare or unattested in natural language, it suggests that biases about language learning persist into adulthood that may help to shape how languages across the world are structured.

The present study focuses on how morphological patterns are learned when the pattern itself differs from the native language, and there is information in the input that is irrelevant to the morphological parsing. The question is whether adults can easily ignore the irrelevant parsing, and learn a morphological pattern that is similar to the native language (English) but differs in important respects. For example, number marking of nouns in English follows a singular-plural distinction in which singular nouns are unmarked, but plural nouns are marked with a suffix. There are also languages that have a three-way number marking system in which singular, dual and plural are marked each with a specific suffix, as in Slovene (Greenberg, 2006). This type of system poses a specific challenge for an adult learner because the participant may enter the experiment with the assumption that number marking works exactly like English, but will have to undo these assumptions in order to learn that all numbers are marked and that there is a distinction between 'two' and 'plural' that is not found in English.

Previous research has explored how adults and children learn novel category patterns. The bulk of these studies focused on the statistical properties of the items themselves, such as the frequency of presentation, the role of immediate feedback (Braine, et al., 1990), similarity of words belonging to each class (Brooks, Braine, Catalano, \& Brody, 1993) and the density and overlap between subcategories (Reeder, Newport, \& Aslin, 2013; Reeder, Newport, \& Aslin, 2009, 2010). Finley and Newport (2010; 2011) focused on the statistical cues that allow for morpheme segmentation without semantic information. In addition, it has been shown that providing visual word cues can
enhance speech segmentation of a novel language (Cunillera, Laine, Camara, \& Rodriguez-Fornells, 2010).

The present paper extends previous research by focusing specifically on morphological paradigms that relate to a specific system of form-meaning combinations that can extend to novel words as in a wug test (Berko, 1958). While Finley and Newport $(2010,2011)$ focused on learning a novel language in which all words were systematically marked by a morpheme, the morphemes had no meaning associated with them, and so it was not clear how the morphemes worked together to form a morphological system, or paradigm. A morphological paradigm is a set of morphemes that marks specific classes (e.g., three suffixes, each marking a different number, /-bu/ 'singular', /-ke/ 'dual', /-mi/ 'plural'). In this study, we test the role of distributional information in learning novel morphological systems, thus extending Finley and Newport $(2010,2011)$ to include morphologically complex systems where both form and meaning are required to learn the language. In order to understand what aspects of the system participants learn, we measured generalization to novel items. This involved measuring responses to test items that appeared in the set of training words in addition to a new set of test words.

In addition, the present study explored whether learners can cope with irrelevant cues when learning a novel pattern. For example, when exposed to a novel label, the learner must weigh many possibilities, many of which are not part of the intention of the speaker (Medina, Snedecker, Trueswell, \& Gleitman, 2011). The same is true in learning novel morphological patterns. If the word ending has a specific morphological meaning (e.g., /-ing/ in English /running/ as opposed to /-ing/ in /string/), the learner must discern whether (and when) this ending has semantic significance, and what (if any) that semantic significance is.

In the present study, participants were exposed to a novel language in which all words were nouns that marked number (singular, plural and dual). However, gender information was provided for the nouns, simulating the problem of ambiguity in learning novel instances in a controlled manner.

## Methods

The present study used an artificial language that contained a large number of stem words and fewer suffixes, mirroring the fact that many natural languages have many more open class morphemes (stems) than closed class morphemes (affixes). Participants were exposed to a miniature language with nouns marked for number, in the form of picture-word pairings. Following exposure, participants were provided with a test in order to determine whether the participant was able to discriminate between the different suffixes and their appropriate meanings.

## Participants

All thirteen participants were adult native English speakers recruited from Elmhurst College and the surrounding community. Each participant was given a $\$ 10$
gift card for participating. Some participants may have previously participated in an artificial grammar learning experiment, however no participant had been exposed to the stimuli used during the present experiment.

## Design

The experiment was designed to test the ability of adult learners to integrate learning form and meaning when the form is arbitrary. A miniature language was developed for the study that contained only words with stems and suffixes. Stems of the words were paired with a type of animal (e.g., /befa/ denotes a ladybug). Each suffix corresponded to the number of animals. The suffix /-bu/ denoted 'singular' (e.g., /befabu/ 'one ladybug'), the suffix /-ke/ denoted 'dual', (e.g., /befake/ 'two ladybugs'), and the suffix /-mi/ denoted 'plural' (e.g.. /befami/ 'more than two ladybugs').

Exposure to the language was created via picture-word pairings in which the sound of the word was paired with a picture of the appropriate number of animals. The gender of the animal varied randomly throughout, and served as irrelevant information. The gender of the animal was denoted using a bowtie for males, purses for females and no marking for unmarked gender. Examples of the picturesound pairings can be found in Table 1.

Table 1: Examples of Picture-Naming Pairings.


All stems were of the form CVCV and all suffixes were of the form CV (/-bu/, /-ke/ and /-mi/), where C is a consonant and V is a vowel. All words were therefore CVCVCV. Consonants were all from the set $[b, d, g, k, m$, $\mathrm{n}, \mathrm{p}, \mathrm{s}, \mathrm{t}, \mathrm{v}, \mathrm{z}]$ and vowels were from the set $[\mathrm{a}, \mathrm{e}, \mathrm{i}, \mathrm{o}, \mathrm{u}]$. No words overlapped with English words, and each consonant and vowel was presented equally often in each position.

Training consisted of 12 stems (each corresponding to a different animal), combined with two of the three suffixes, creating 24 total training items. Each stem and each suffix appeared with equal frequency across the 24 training items. Exposure consisted of repeating the 24 training items eight times. While each sound could be paired with three different pictures (e.g., if /ganubu/ signified a single giraffe, the appropriate picture would be for any gender: female, male and neuter), the same picture was used for each pictureword pairing for all eight cycles of the training stimuli. The irrelevant information (e.g., gender markings on the training pictures) was distributed throughout the training items.

Participants were tested on their knowledge of the language as well as their ability to generalize the suffix information to novel stem forms using a two-alternative forced-choice test. In the test, participants matched two spoken words to a single picture. Participants were told one of the words would be from the language they had been listening to and the other word would not be from the language. Participants chose which of two words correctly corresponded to the picture shown. There were twelve items in three different test conditions, described in more detail below, with examples in Tables 2-4.

Familiar Stem-Familiar Picture The first type of test item specifically tested the learner's ability to match a picture seen in training to its corresponding word. Participants heard two words with the same bi-syllabic stem. One word was heard during training, and the other word was a word not heard in training but contained the same stem as the 'correct' test item. If the participant was able to match the picture to the correct suffix, it demonstrates that the participant had learned the suffix-picture pairings. Because the two options contained the same stems, the options were highly similar, and could thus not rely on the stem to make the correct response.

Table 2: Familiar Stem-Familiar Picture Test Items

| Correct Item | Decoy <br> Item |
| :--- | :--- |
| befabu | befake |
| sufemi | sufebu |

Familiar Stem-Novel Picture The second type of test item probed the learner's ability to generalize the suffixes to the items that were not heard in training. For every stem, there were three possible suffix pairings, but only two were heard
in training. In this test condition, the picture shown corresponded to the stem+suffix pairing that was not heard in training, and the decoy item was a stem+suffix item that was heard during training. Both options had the same stem, meaning that participants had to rely on the suffix to choose the correct response. Because the decoy item was familiar to the participant, if participants chose the item that was most familiar, they would be incorrect. Examples of these test items can be found in Table 3, below.

Table 3: Familiar Stem-Novel Picture Test Items

| Correct Item | Decoy <br> Item | Picture |
| :--- | :--- | :--- |
| befake | befami | sufemi |
| ganumi | ganubu |  |

Novel Word-Novel Picture The third type of test item probed the learner's ability to generalize to novel stem items. This served as a comprehension version of a wug test. Participants heard two stem+suffix combinations, in which the stems were identical in both conditions. The picture shown corresponded to one of the suffixes. Participants could only rely on knowledge of the suffix to get these items correct, as the participants had not seen these stem items in training.

Table 4: Novel Word-Novel Picture Test Items

| Correct Item | Decoy <br> Item | Picture |
| :---: | :--- | :---: |
| pumubu | pumumi | pazike |
| pozimi | kovebu |  |

There were 12 tokens of each of the three test sets of test items. These items were presented in a random, mixed fashion. The 12 items in each test condition were balanced such that the correct response was singular, plural and dual an equal number of times (four). The items were also balanced such that all possible suffix combinations were heard an equal number of times (e.g., in a test trial where /bedemi/ is pit against /bedeke/, the two suffix options are $/ \mathrm{ke} /$ and $/ \mathrm{mi} /$. This suffix combination occurred equally often as $/ \mathrm{ke} / \mathrm{vs}$. $/ \mathrm{bu} /$ and $/ \mathrm{mi} / \mathrm{vs}$. $/ \mathrm{bu} /$ ).

All stimuli were recorded by an adult female native speaker of English in a sound attenuated booth at $12,000 \mathrm{~Hz}$ (though participants were allowed to adjust headphones to a comfortable volume during the experiment). Stress was placed on the first syllable using standard English pronunciation, with the exception that no vowels were reduced, meaning all syllables contained partial stress (as English reduces unstressed syllables). All stimuli items were normalized for intensity (set at 70dB) using Praat (Boersma \& Weenink, 2005). All phases of the experiment were run in Psyscope X (Cohen, MacWhinney, Flatt, \& Provost, 1993). Participants were given both written and verbal instructions. The entire experiment took approximately 20 minutes.

## Results

Proportion of correct responses for all three test items are given in Figure 1. We compared each test item to $50 \%$ chance via three separate one-sample $t$-tests. All three test items were significantly above chance; the Familiar StemFamiliar Picture test items had a mean of $0.88, t(12)=7.91$, $p<0.001$, the Familiar Stem-Novel Picture test items had a mean of $0.85, t(12)=75.69, p<0.001$ and the Novel WordNovel Picture test items had a mean of $0.88, t(12)=7.37$, $p<0.001$, suggesting that the participants learned the suffixation pattern.

Figure 1: Test Item Results.


Because the novel language contained contrasts and markings for number (singular-dual-plural) that are not found in English, we divided responses by number marking, to ensure that all three number markings were learned by participants. These are presented in Figure 2.

Figure 2: Number-Marking Results.


We compared the results via a $3 \times 3$ within-subjects ANOVA. There were no main effects for Test Type, $(F(2,24)=1.29$, $p=0.29)$, Number Type $(F(2,24)=1.79, p=0.19)$, and no Interaction $(F(4,48)=1.71, p=0.16)$. This suggests that all three number markings were learned equally well.

We hypothesized that the difference between dual and plural may be the most difficult to learn because dual and plural are not distinguished in English. We therefore performed planned comparisons between dual and plural test items. There were no differences between dual and plural test items for either the Familiar Word-Familiar Picture, $t(12)=0.32, p=0.75$, or the Familiar Word-Novel Picture, $t(12)=1.10, p=0.29$ test items. There was a significant difference between the dual and the plural test items for the Novel Word-Novel Picture test items, $t(12)=2.50, p=0.028$. This suggests that if there is a difference in the difficulty of learning novel number markings, that this is most likely to appear during generalization to novel stems.

The learning rates were relatively robust across participants. Of the 13 participants, only two had overall means less than $80 \%$ ( $50 \%$ and $41.67 \%$ respectively). For these two participants, the difference in number was most pronounced. These participants were most accurate on dual test items ( $70.83 \%$ correct), around chance for singular items $(45.83 \%$ correct) and below chance for plural items ( $20.83 \%$ correct). Because so few participants scored below $80 \%$ correct, no inferential statistics can be made. However, these results may indicate that those who have difficulty learning novel morphological systems may only have trouble with specific number markers.

## Discussion

The results of the present study provide important insights into how novel complex morphological systems are learned. First, consistent endings along with consistent number cues allowed the vast majority of participants to infer that the final suffix referred to number, and that this final suffix applied to novel items, both for stems heard during exposure, as well as novel items not heard during exposure. Second, this ability is very robust in adults. Of the thirteen participants, only two showed means below $80 \%$ suggesting that these relatively complex patterns are learned with ease, without any feedback from the learning paradigm. Third, the number markings in the present experiment differed from those found in English: all different numbers were marked (as opposed to only plural in English), and a distinction was created between dual and plural (as opposed to only plural in English).

The stimuli in the present experiment included irrelevant cues to gender, which the participants were able to rule out. Because both gender and number cues were provided, the paradigm allows for future research to study both gender and number markings simultaneously. It also demonstrates that learners are able to cue into the relevant aspects of novel data, and ignore irrelevant aspects.

The results showed relatively few differences between test items that probed for knowledge of the different number markings, despite the fact that the dual number marking was novel to the English speakers. This suggests that learners are adaptable to novel number markings. Interestingly, the two participants who showed poor performance overall, seemed to show differential responses to different number categories, suggesting that problems in learning novel morphology may be specific to a specific morpheme, rather than the entire morphological paradigm. Because these trends can only be made for two participants, more research is needed to understand why some learners have difficulty learning novel morphological structures, while others have little difficulty.

The present study presented a novel paradigm for exploring how adults are able to learn novel morphological systems. The results demonstrated that adults are able to rapidly and robustly learn novel number marking systems despite irrelevant gender cues. The present paradigm provides a tool for future research to explore how complex systems of form and meaning are learned and generalized to novel items. The present paradigm allows the experimenter to control for how much information is relevant to the morphological system and how much is irrelevant.

The paradigm specifically allows the researcher to explore novel questions about how complex morphological systems are learned. In many languages, the same phonological unit is used to mark multiple morphemes. For example, $/-\mathrm{s} /$ is used in English to mark both plural as well as third person singular, present tense verbs. In German, /der/ is used to mark nominative singular case, as well as plural genitive case. In these instances, the learner must sort out when each morpheme is used. The present paradigm may help to sort out what aspects of the morphological paradigm are most helpful to learning a complex paradigm. Future research will explore how phonological regularities and semantic consistency contribute to learning a novel morphological paradigm.

The present study makes use of adult participants. While studying children is often ideal when examining language learning, adult studies are also extremely useful in terms of understanding how learning biases persist into adulthood. The present paradigm is well suited to adapt to child language studies, allowing future research to easily make adult-children comparisons in learning. However, there are many reasons that using adults in the present study has theoretical importance. Throughout life, novel stimuli are presented in a language no matter how long ago the language was learned. New words come into the language (e.g., as each new generation adds to the list of slang words). Adult studies increase the knowledge about continuing language learning in the first language and learning in general. In addition, studying adults in a second language environment will help to understand the biases that adults use in second language learning, which may provide insight into making adult second language learning easier. In addition, studies of adult second language learning often
reveal deficiencies in learning the morphology of the language (Johnson \& Newport, 1989; Newport, 1990). Thus, understanding adult learning biases for morphological learning may have direct implications for understanding these deficits (and possibly finding methods to correct them).

The present study adds to the growing number of studies that demonstrate that learners are able to make use distributional cues to learn the regular (rule-based) aspects of language. When forms (e.g., suffixes) are paired consistently with a meaning, the learner infers a general rule that can apply to items that have never been seen or heard before. This is done despite additional, irrelevant cues that could potentially disrupt the learning mechanism. The fact that learners are able to sort out which cues are relevant without any direct feedback, demonstrates the enormous inferential power of the human mind.

## Acknowledgments

The authors would like to thank Matt Goldrick, Patricia Reeder, Elissa Newport, Emily Kasman, Morgan Smith and Elizabeth Jennings, members of the Aslin-Newport Lab, as well as participants at the 2012 Psychonomics Meeting for helpful comments and discussion. In addition, we would like to thank the Elmhurst College Psychology Department and the participants at Elmhurst College. S. Finley assumes all responsibility for any errors. Funding was provided by a student-faculty collaboration grant from Elmhurst College.

## References

Berko, J. (1958). The child's learning of English morphology. Word, 15, 150-177.
Boersma, P., \& Weenink. (2005). Praat: Doing phonetics by computer.
Braine, M. D. S., Brody, R. E., Brooks, P. J., Sudhalter, V., Ross, J. A., Catalano, L., et al. (1990). Exploring language acquisition in children with a miniature artificial language: Effects of item and pattern frequency, arbitrary subclasses, and correction. Journal of Memory and Language, 29, 591-610.
Brooks, P. J., Braine, M. D. S., Catalano, L., \& Brody, R. E. (1993). Acquisition of gender-like noun subclasses in an artificial language: The contribution of phonological markers to learning. Journal of Memory and Language, 32, 76-95.
Cohen, J. D., MacWhinney, B., Flatt, M., \& Provost, J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. Behavioral Research Methods, Instruments and Computers, 25, 257271.

Cunillera, T., Laine, M., Camara, E., \& Rodriguez-Fornells, A. (2010). Bridging the gap between speech segmentation and word-to-world mapping: Evidence from an audiovisual statistical learning task. Journal of Memory and Language, 63, 295-305.
Finley, S., \& Newport, E. (2010). Morpheme segmentation from distributional information. In J. Chandleee, K.

Franich, K. Iserman \& K. L. (Eds.), BUCLD 34 Proceedings Supplement, online. Boston: BUCLD.
Finley, S., \& Newport, E. (in prep). Morpheme segmentation in artificial grammar learning.
Finley, S., \& Newport, E. N. (2011). Morpheme segmentation in school-aged children. In A. Fine (Ed.), Rochester working papers in the language sciences. Cambridge, MA: MIT Press.
Frigo, L., \& McDonald, J. L. (1998). Properties of phonological markers that affect the acquisition of gender-like subclasss. Journal of Memory and Language, 39, 218-245.
Gomez, R. L., \& Gerken, L. (2000). Infant artificial language learning and language acquisition. Trends in Cognitive Science, 4(5), 178-186.
Gomez, R. L., \& Maye, J. (2005). The developmental trajectory of nonadjacent dependency learning. Infancy, 7(2), 183-206.
Greenberg, M. (2006). A short reference grammar of standard Slovene. from http://www.theslovenian.com/articles/2008/greenberg.pdf
Johnson, J. S., \& Newport, E. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. Cognitive Psychology, 21, 60-99.
MacWhinney, B. (1983). Miniature Linguistic Systems as Tests of the Use of Universal Operating Principles in Second-Language Learning by Children and Adults. Journal of Psycholinguistic Research, 12(5), 467-478.
Medina, T. N., Snedecker, J., Trueswell, J. C., \& Gleitman, L. R. (2011). How words can and cannot be learned by observation. Proceedings of the National Academy of Sciences, 108(2), 9014-9019.
Mintz, T., Newport, E., \& Bever, T. G. (2002). The distributional structure of grammatical categories in speech to young children. Cognitive Science, 26, 393-424.
Newport, E. (1990). Maturational constraints on language learning. Cognitive Science, 14, 11-28.
Reeder, P., Newport, E., \& Aslin, R. (2013). From shared context to syntactic categories: The role of distributional information in learning linguistic form-classes. Cognitive Psychology, 66, 30-54.
Reeder, P., Newport, E., \& Aslin, R. N. (2009). The role of distributional information in linguistic category formation. In S. Ohlsson \& R. Catrambone (Eds.), Proceedings of the 32nd Annual Conference of the Cognitive Science Society (Vol. 32, pp. 2063-2068). Austin, TX: Cognitive Science Society.
Reeder, P., Newport, E., \& Aslin, R. N. (2010). Novel words in novel contexts: The role of distributional information in form-class category learning. In S. Ohlsson \& R. Catrambone (Eds.), Proceedings of the 32nd Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.

# Analogical Reinforcement Learning 

James M. Foster \& Matt Jones<br>james.m.foster@colorado.edu, mcj@colorado.edu<br>University of Colorado, Department of Psychology \& Neuroscience<br>Boulder, CO 80309 USA


#### Abstract

Research in analogical reasoning suggests that higher-order cognitive functions such as abstract reasoning, far transfer, and creativity are founded on recognizing structural similarities among relational systems. Here we integrate theories of analogy with the computational framework of reinforcement learning (RL). We propose a computational synergy between analogy and RL, in which analogical comparison provides the RL learning algorithm with a measure of relational similarity, and RL provides feedback signals that can drive analogical learning. Initial simulation results support the power of this approach.


Keywords: Analogy; Reinforcement Learning; Schema Induction; Similarity; Generalization

## Introduction

The goal of the present work is to develop a computational understanding of how people learn abstract concepts. Previous research in analogical reasoning suggests that higherorder cognitive functions such as abstract reasoning, far transfer, and creativity are founded on recognizing structural similarities among relational systems (Doumas et al., 2008; Gentner, 1983; Hummel \& Holyoak, 2003). However, we argue a critical element is missing from these theories, in that their operation is essentially unsupervised, merely seeking patterns that recur in the environment, rather than focusing on the ones that are predictive of reward or other important outcomes.

Here we integrate theories of analogy with the computational framework of reinforcement learning (RL). RL offers a family of learning algorithms that have been highly successful in machine learning applications (e.g., Bagnell \& Schneider, 2001; Tesauro, 1995) and that have neurophysiological support in the brain (e.g., Schultz et al., 1997). A shortcoming of RL is that it only learns efficiently in complex tasks if it starts with a representation (i.e., a means for encoding stimuli or states of the environment) that somehow captures the critical structure inherent in the task. We formalize this notion below in terms of similarity-based generalization (Shepard, 1987) and kernel methods from statistical machine learning (Shawe-Taylor \& Cristianini, 2004). In other words, RL requires a sophisticated sense of similarity to succeed in realistically complex tasks. Psychologically, the question of how such a similarity function is learned can be cast as a question of learning sophisticated, abstract representations.

This paper proposes a computational model of analogical RL, in which analogical comparison provides the RL learning algorithm with a measure of relational similarity, and RL provides feedback signals that can drive analogical learning. Relational similarity enables RL to generalize knowledge from past to current situations more efficiently, leading to faster
learning. Conversely, the prediction-error signals from RL can be used to guide induction of new higher-order relational concepts. Thus we propose there exists a computationally powerful synergy between analogy and RL. The simulation experiment reported here supports this claim. Because of the strong empirical evidence for each of these mechanisms taken separately, we conjecture that the brain exploits this synergy as well.

## Analogy

Research in human conceptual knowledge representation has shown that concepts are represented not just as distributions of features (cf. Nosofsky, 1986; Rosch \& Mervis, 1975) but as relational structures. This relational knowledge includes both internal structure, such as the fact that a robin's wings allow it to fly (Sloman et al., 1998), as well as external structure, such as the fact that a dog likes to chase cats (Jones \& Love, 2007). Theories of analogical reasoning represent relational knowledge of this type in a predicate calculus that binds objects to the roles of relations, for example CHASE(DOG,CAT). According to these theories, an analogy between two complex episodes (each a network of relations and objects) amounts to recognition that they share a common relational structure (Gentner, 1983; Hummel \& Holyoak, 2003).

At a more mechanistic level, the dominant theory of analogy is structural alignment (Gentner, 1983). This process involves building a mapping between two episodes, mapping objects to objects and relations to relations. The best mapping is one that maps objects to similar objects, maps relations to similar relations, and most importantly, satisfies parallel connectivity. Parallel connectivity means that, whenever two relations are mapped to each other, the objects filling their respective role-fillers are also mapped together. An example is shown in Figure 1. Parallel connectivity is satisfied here because, for each mapped pair of attack relations (red arrows), the objects filling the ATTACKER role are mapped together (knight is mapped to queen), and the objects filling the ATTACKED role are also mapped together (rook to rook and king to king). Thus structural alignment constitutes a (potentially partial or imperfect) isomorphism between two episodes, which respects the relational structure that they have in common. Importantly, if the search for a mapping gives little emphasis to object-level similarity (as opposed to relation-level similarity and parallel connectivity), then structural alignment can find abstract commonalities between episodes having little or no surface similarity (i.e., in terms of perceptual features).

We propose structural alignment is critical to learning of


Figure 1: An example of structural alignment between two chess positions. Both positions contain instances of the abstract concept of a FORK: black's piece is simultaneously attacking both of white's pieces. These attacking relations are represented by the red arrows. Cyan lines indicate the mapping between the two episodes. The mapping satisfies parallel connectivity because it respects the bindings between relations and their role-fillers.
abstract concepts for three reasons. First, perceived similarity of relational stimuli depends on structural alignability (Markman \& Gentner, 1993). Second, structural alignment is important for analogical transfer, which is the ability to apply knowledge from one situation to another, superficially different situation (Gick \& Holyoak, 1980). For example, a winning move in one chess position can be used to discover a winning move in a different (but aligned) position, by translating that action through the analogical mapping. Third, a successful analogy can lead to schema induction, which involves extraction of the shared relational structure identified by the analogy (Doumas et al., 2008; Gentner, 1983; Hummel \& Holyoak, 2003). In the example of Figure 1, this schema would be a system of relational knowledge on abstract (token) objects, including ATTACK(PIECE1,PIECE2), ATTACK(PIECE1,PIECE3), and potentially other shared information such as NOT_ATTACKED(PIECE1) and KING(PIECE2).

These three observations suggest that analogy plays an important role in learning and use of abstract relational concepts. The first two observations suggest that analogical transfer can be cast as a form of similarity-based generalization, as we elaborate in the next two sections. In brief, structural alignment offers a sophisticated form of similarity that can be used to generalize knowledge between situations that are superficially very different. The third observation suggests that analogy can discover new relational concepts (e.g., the concept of a chess fork, from Figure 1), which can in turn lead to perception of even more abstract similarities among future experiences.

One potential shortcoming of the basic theory of analogy reviewed here is that is it essentially unsupervised. In this framework, the quality of an analogy depends only on how well the two systems can be structurally aligned, and not on how useful or predictive the shared structure might be. For
example, one could list many relational patterns that arise in chess games but that are not especially useful for choosing a move or for predicting the course of the game. In previous work, we have found that implementing structural alignment and schema induction in a rich and structured artificial environment results in discovery of many frequent but mostly useless schemas (Foster et al., 2012). An alternative, potentially more powerful model of analogical learning would involve feedback from the environment, so that the value of an analogy or schema is judged partially by how well it improves predictions of reward or other important environmental variables. For example, the concept of a fork in chess is an important schema not (only) because it is a recurring pattern in chess environments, but because it carries information about significant outcomes (i.e., about sudden changes in each player's chances of winning). A natural framework for introducing this sort of reward sensitivity into theories of analogy is that of RL, which we review next.

## Reinforcement Learning

RL is a mathematical and computational theory of learning from reward in dynamic environments. An RL task is characterized by an agent embedded in an environment that exists in some state at any given moment in time. At each time step, the agent senses the state of its environment, takes an action that affects what state occurs next, and receives a continuousvalued reward that depends on the state and its action (Sutton \& Barto, 1998). This framework is very general and can encompass nearly any psychological task in which the subject has full knowledge of the state of the world at all times (i.e., there are no relevant hidden variables).

Most RL models work by learning values for different states or actions, which represent the total future reward that can be expected from any given starting point (i.e., from any state or from any action within a state). These values can be learned incrementally, from temporal-difference (TD) error signals calculated from the reward and state following each action (see Model section). There is strong evidence that the brain computes something close to TD error, and thus that RL captures a core principle of biological learning (Schultz et al., 1997).

In principle, this type of simple algorithm could be used to perfectly learn a complex task such as chess, by experiencing enough games to learn the true state values (i.e., probability of winning from every board position) and then playing according to those values. However, a serious shortcoming of this naive approach is that it learns the value of each state independently, which can be hopelessly inefficient for realistic tasks that typically have very large state spaces. Instead, some form of generalization is needed, to allow value estimates for one state to draw on experience in other, similar states.

Many variants of RL have been proposed for implementing generalization among states (e.g., Albus, 1981; Sutton, 1988). Here we pursue a direct and psychologically motivated form of generalization, based on similarity (Jones \&

Cañas, 2010; Ormoneit \& Sen, 2002). We assume the model has a stored collection of exemplar states, each associated with a learned value. The value estimate for any state is obtained by a similarity weighted average over the exemplars' values; that is, knowledge from each exemplar is used in proportion to how similar it is to the current state. This approach is closely related to exemplar-generalization models in more traditional psychological tasks such as category learning (Nosofsky, 1986). It can also be viewed as a subset of kernel methods from machine learning (Shawe-Taylor \& Cristianini, 2004), under the identification of the kernel function with psychological similarity (Jäkel et al., 2008).

A critical consideration for all learning models (including RL models) is how well their pattern of generalization matches the inherent structure of the task. If generalization is strong only between stimuli or states that have similar values or outcomes, then learning will be efficient. On the other hand, if the model generalizes significantly between stimuli or states with very different outcomes, its estimates or predictions will be biased and learning and performance will be poor. The kernel or exemplar-similarity approach makes this connection explicit, because generalization between two states is directly determined by their similarity. As we propose next, analogy and schema induction offer a sophisticated form of similarity that is potentially quite powerful for learning complex tasks with structured stimuli.

## Analogical RL

The previous two sections suggest a complementary relationship between analogy and RL, which hint at the potential for a computationally powerful, synergistic interaction between these two cognitive processes. We outline here a formal theory of this interaction. The next two sections provide a mathematical specification of a partial implementation of this theory, and then present simulation results offering a proof-inprinciple of the computational power of this approach.

The first proposed connection between analogy and RL is that structural alignment yields an abstract form of psychological similarity that can support sophisticated generalization (Gick \& Holyoak, 1980; Markman \& Gentner, 1993). Incorporating analogical similarity into the RL framework could thus lead to rapid learning in complex, structured environments. For example, an RL model of chess equipped with analogical similarity should recognize the similarity between the two positions in Figure 1 and hence generalize between them. Consequently the model should learn to create forks and to avoid forks by the opponent much more rapidly than if it had to learn about each possible fork instance individually.

The second proposed connection is that the TD error computed by RL models, for updating value estimates, can potentially drive analogical learning by guiding schema induction. Instead of forming schemas for whatever relational structures are frequently encountered (or are discovered by analogical comparison of any two states), an analogical RL model can be more selective, only inducing schemas from analogies that
significantly improve reward prediction. Such analogies indicate that the structure common to the two analogue states may have particular predictive value in the current task, and hence that it might be worth extracting as a standalone concept. For example, if the model found a winning fork move by analogical comparison to a previously seen state involving a fork, the large boost in reward could trigger induction of a schema embodying the abstract concept of a fork.

The proposed model thus works as follows (see the next section for technical details). The model maintains a set of exemplars $E$, each with a learned value, $v(E)$. To estimate the value of any state $s$, it compares that state to all exemplars by structural alignment, which yields a measure of analogical similarity for each exemplar (Forbus \& Gentner, 1989). The estimated value of the state, $\tilde{V}(s)$, is then obtained as a similarity-weighted average of $v(E)$. After any action is taken and the immediate reward and next state are observed, a TD error is computed as in standard RL. The exemplar values are then updated in proportion to the TD error and in proportion to how much each contributed to the model's prediction, that is, in proportion to $\operatorname{sim}(s, E)$.

Additionally, whenever the structural alignment between a state and an exemplar produces a sufficient reduction in prediction error (relative to what would be expected if that exemplar were absent), a schema is induced from that analogy. The schema is an abstract representation, defined on token (placeholder) objects, and it contains only the shared information that was successfully mapped by the analogy. The schema is added to the pool of exemplars, where it can acquire value associations directly (just like the exemplars do). The advantage conferred by the new schema is that it allows for even faster learning about all states it applies to (i.e., that contain that substructure). For example, rather than learning by generalization among different instances of forks, the model would learn a direct value for the fork concept, which it could immediately apply to any future instances. A consequence of the schema induction mechanism is that the pool of concrete exemplars comes to contain more and more abstract schemas. Thus the model's representation transitions from initially episodic to more abstract and conceptual.

Analogical RL thus integrates three principles from prior research: RL, exemplar generalization, and structural alignment of relational representations. Because each of these principles has strong empirical support as a psychological mechanism, it is plausible that they all interact in a manner similar to what we propose here. Thus it seems fruitful to explore computationally what these mechanisms can achieve when combined.

## Model

The simulation study presented below uses a variant of RL known as afterstate learning, in which the agent learns values for the possible states it can move into (Sutton \& Barto, 1998). This is a reasonable and efficient method for the task we use here-tic-tac-toe, or noughts \& crosses-because the
agent's opponent can be treated as part of the environment and is the only source of randomness. Our main proposal regarding the interaction between RL and analogical learning is not limited to this approach.

The operation of the model is illustrated in Figure 2. On each time step, the model identifies all possible actions and their associated afterstates. For each afterstate $s$, it computes an analogical similarity, $K$, to each exemplar, $E$, by structural alignment. Each possible mapping $M: s \rightarrow E$ is evaluated according to

$$
\begin{align*}
& \Phi(M)=\beta \cdot \sum_{o \in s} \operatorname{sim}(o, M(o)) \\
& +\sum_{r \in s} \operatorname{sim}(r, M(r)) \cdot\left[1+\sum_{i=1}^{n_{r}} I_{\left\{M\left(\operatorname{child}_{i}(r)\right)=\operatorname{child}_{i}(M(r))\right\}}\right] \tag{1}
\end{align*}
$$

This expression takes into account object similarity, by comparing each object $o$ in $s$ to its image in $E$; relational similarity, by comparing each relation $r$ in $s$ to its image in $E$; and parallel connectivity, by having similarity between mutually mapped relations "trickle down" to add to the similarity of any mutually mapped role-fillers (Forbus \& Gentner, 1989). The sim function is a primitive (object- and relationlevel) similarity function, $\beta$ determines the relative contribution of object similarity, $n_{r}$ is the number of roles in relation $r, \operatorname{child}_{i}(r)$ is the object filling the $i^{\text {th }}$ role of $r$, and $I_{\{P\}}$ is an indicator function equal to 1 when proposition $P$ is true. Analogical similarity is then defined as the value of the best mapping (here the $\theta$ parameter determines specificity of generalization):

$$
\begin{equation*}
K(s, E)=\exp \left(\theta \cdot \max _{M} \Phi(M)\right) \tag{2}
\end{equation*}
$$

The activation $a(E)$ of each exemplar is determined by normalizing the analogical similarities, and the estimated value of $s, \tilde{V}(s)$, is computed as a similarity-weighted average of the exemplar values $v(E)$ (Figure 2). Thus the estimate is based on the learned values of the exemplars most similar to the candidate state.

Once values $\tilde{V}(s)$ have been estimated for all candidate afterstates, the model uses a softmax (Luce-choice or Gibbssampling rule) to select what state to move into (here $\tau$ is an exploration parameter):

$$
\begin{equation*}
\operatorname{Pr}\left[s_{t}=s\right] \propto e^{\tilde{V}(s) / \tau} \tag{3}
\end{equation*}
$$

Learning based on the chosen afterstate $s_{t}$ follows the SARSA rule (Rummery \& Niranjan, 1994), after the model chooses its action on the next time step. This produces a TD error, which is then used to update the exemplar values by gradient descent (see Equations for $\delta$ and $\Delta v(E)$ in Figure 2).

The model also grows its representation in two ways. First, it begins with no exemplars, and on each trial adds the state it moves to as a new exemplar with probability inversely proportional to the number of exemplars already in the model. This recruitment policy leads the exemplar pool to grow with


Figure 2: Model operation. Each candidate afterstate is evaluated by analogical comparison to stored exemplars, followed by similarity-weighted averaging among the learned exemplar values. Learning is by TD error applied to the exemplar values. On some trials, especially useful analogies produce new schemas that are added to the exemplar pool. In the example here, $s$ and $E$ both have guaranteed wins for X by threatening a win in two ways. The induced schema embodies this abstract structure. Dots with red arrows indicate ternary "same-rank" relations. $r=$ reward; $\gamma=$ temporal discount parameter; $\alpha=$ learning rate; other variables are defined in the text.
the square root of time, which seems to give good performance.

The more important form of representation learning in the model is schema induction. Schema induction has not been implemented yet, but Figure 2 shows how it is expected to work. Following learning after each trial, the model determines how much each exemplar contributed to reducing prediction error, by comparing $\delta$ to what it would have been without that exemplar. If the reduction is above some threshold, the analogical mapping found for that exemplar (lower right of figure) produces a schema that is added to the exemplar pool (far right). The schema is given a value of $v$ initialized at $\tilde{V}\left(s_{t}\right)$. This schema value is updated on future trials just as are the exemplar values. Acquisition of new schemas in this way is predicted to improve the model's pattern of generalization, tuning it to the most useful relational structures in a task.

## Simulation

The model was tested on its ability to learn tic-tac-toe. Each board position was represented by treating the nine squares as objects of types 0 (blank), 1 (focal agent's), and 2 (opponent's), and defining 8 ternary "same-rank" relations for the rows, columns, and diagonals. Thus a player wins by filling all squares in any one of these relations. Object similarity was defined as 1 for matching object types and 0 otherwise. Similarity between relations was always 1 because there was only one type of relation. Reward was given only at the end
of a game, as +1 for the winner, -1 for the loser, or 0 for a draw. After the game ended, it moved to a special terminal state with fixed value of 0 . For simplicity, all free parameters of the model $(\beta, \theta, \alpha, \gamma, \tau)$ were set to a default value of 1 .

Three variations of the model were implemented, differing in their levels of analogical abstraction. The Featural model was restricted to literal mappings between states (upper-left square to upper-left square, etc.). This model still included generalization, but its similarity was restricted to the concrete similarity of standard feature-based models. The Relational model considered all 8 mappings defined by rigid rotation and reflection of the board. This scheme was used in place of searching all 9 ! possible mappings for every comparison, to reduce computation time. Finally, the Schema model extended the Relational model by starting with two hand-coded schemas, 111 and 022 . The first of these is a single same-rank relation bound to three instances of the player's own token. Thus moving into a state satisfying this schema produces an immediate win. Likewise, moving into a state satisfying the second schema risks an immediate loss. The model was given no information about these schemas (i.e., $v$ was initialized to 0 for both), but it was capable of learning values for them. The purpose of this model was to test the utility of having schemas that capture task-relevant structures. Logically this question is separate from that of how such schemas are acquired, although we have addressed that question elsewhere (Foster et al., 2012), and we plan to integrate a solution into the present model soon.

Each model variant was trained in blocks of 10 games of self-play followed by a pair of testing games against an ideal player (playing first in one game and second in the other). Learning occurred only during training. In testing games, the model was given one point for each non-losing move it made (i.e., moves from which it could still guarantee a draw), for a maximum of 9 points per pair of testing games.

Average learning curves are shown in Figure 3A for 50 independent copies of each model over 5000 blocks (50,000 training games). Figure 3B shows results for single copies of the Relational and Featural models over 30,000 blocks. These results show that the Featural model does eventually learn, but the Relational model learns an order of magnitude faster, and the Schema model learns another order of magnitude faster than the Relational model.

## Discussion

The results presented here constitute a proof-of-principle that analogy and schema induction can be productively integrated with a learning framework founded on RL and similaritybased generalization. This integration leads to a model exhibiting sophisticated, abstract generalization derived from analogical similarity, as well as discovery of new higher-order relational concepts driven by their ability to predict reward.

The basic modeling framework used here applies not just to analogical similarity and schema induction, but to other forms of representational learning as well. Kernel-based RL offers


Figure 3: Learning curves. A: 50 copies of each model. B: Single copies of the two slower models over extended training.
a powerful and general theory of representation learning, because it can be integrated with any form of representation that yields a pairwise similarity function. Its TD error signal can drive changes in representation via the objective of improving generalization. In previous work, we have applied this idea to learning of selective attention among continuous stimulus dimensions (Jones \& Cañas, 2010). The current model offers a richer form of representation learning, in that it acquires new concepts rather than reweighting existing features.

The analogical RL model also builds on other models of relational learning. Tomlinson \& Love (2006) propose a model of analogical category learning, with essentially the same similarity and exemplar generalization mechanisms adopted in the present model. Our model adds to theirs in that it applies to dynamic tasks and in that it grows its representation through schema induction. Van Otterlo (2012) has developed methods for applying RL to relational representations of the same sort used here, although the approach to learning is quite different. His models are not psychologically motivated and hence learn in batches and form massive conjunctive rules, with elaborate updating schemes to keep track of the possible combinations of predicates. In contrast, the present approach learns iteratively, behaves probabilistically, and grows its representation more gradually and conservatively. This approach is likely to provide a better account of human learning, but a more interesting question may be whether it offers any performance advantages from a pure machine-learning perspective.

In the present model, the activation of each exemplar elicited by a candidate state can be thought of as a feature of that state. The exemplar effectively has a "receptive field" within the state space, defined by the similarity function. This duality between exemplar- and feature-based representations is founded in the kernel framework (see Shawe-Taylor \& Cristianini, 2004). The present model takes advantage of this duality, producing a smooth transition from an episodic, similarity-based representation to a more semantic, featurebased representation defined by learned schemas.

The model as currently implemented does have several limitations. Foremost, it does not yet include a mechanism for inducing new schemas. We and others have shown how schema induction can be successfully deployed in an open-
ended model in a complex environment (Doumas et al., 2008; Foster et al., 2012). We hope that building this type of mechanism into the analogical RL framework will produce a bettercontrolled, directed system capable of autonomously discovering genuinely new abstract concepts.

A second limitation of the current model is its slowness to learn, due to the nature of gradient descent operating in a large weight space. In contrast, human learning often shows understanding of new concepts in as little as one trial (Maas \& Kemp, 2009). The theory of analogy via structure mapping seems like the best candidate for a process-level theory of such rapid learning, and we predict that the full analogical RL model with schema induction will show significant steps in that direction.

The present work is complementary to hierarchical Bayesian models that discover relational structure through probabilistic inference (Tenenbaum et al., 2011). Whereas our model builds up schemas from simpler representations, the Bayesian approach takes a top-down approach, defining the complete space of possibilities a priori and then selecting among them. The top-down approach applies to any learning model, because any well-defined algorithm can always be circumscribed in terms of its set of reachable states. This is a useful exercise for identifying inductive biases and absolute limits of learning, but it offers little insight into the constructive processes that actually produce the learning. These mechanistic questions are critical if the goal is to understand how the human mind discovers new, abstract concepts.

## Acknowledgments

Supported by AFOSR Grant FA-9550-10-1-0177 to MJ.

## References

Albus, J. S. (1981). Brains, Behavior and Robotics. Byte Books.
Bagnell, J. A., \& Schneider, J. C. (2001). Autonomous helicopter control using reinforcement learning policy search methods. IEEE Int Conf Robo, 1615-1620.
Doumas, L. A., Hummel, J. E., \& Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. Psychol Rev, 115, 1-43.
Forbus, K. D., \& Gentner, D. (1989). Structural evaluation of analogies: What counts. Proceedings of the 11th Annual Conference of the Cognitive Science Society, 341-348.
Foster, J. M., Cañas, F., \& Jones, M. (2012). Learning conceptual hierarchies by iterated relational consolidation. Proceedings of the 34th Annual Conference of the Cognitive Science Society, 324-329.
Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. Cognitive Sci, 7, 155-170.
Gick, M. L., \& Holyoak, K. J. (1980). Analogical problem solving. Cognitive Psychol, 12, 306-355.
Hummel, J. E., \& Holyoak, K. J. (2003). A symbolicconnectionist theory of relational inference and generalization. Psychol Rev, 110, 220-264.

Jäkel, F., Schölkopf, B., \& Wichmann, F. A. (2008). Generalization and similarity in exemplar models of categorization: Insights from machine learning. Psychon B Rev, 15, 256-271.
Jones, M., \& Cañas, F. (2010). Integrating reinforcement learning with models of representation learning. Proceedings of the 32nd Annual Conference of the Cognitive Science Society, 1258-1263.
Jones, M., \& Love, B. C. (2007). Beyond common features: The role of roles in determining similarity. Cognitive Psychol, 55, 196-231.
Maas, A. L., \& Kemp, C. (2009). One-shot learning with bayesian networks. Proceedings of the 31st Annual Conference of the Cognitive Science Society.
Markman, A. B., \& Gentner, D. (1993). Structural alignment during similarity comparisons. Cognitive Psychol, 25, 431-431.
Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. J Exp Psychol Gen, 115, 39-57.
Ormoneit, D., \& Sen, S. (2002). Kernel-based reinforcement learning. Mach Learn, 49, 161-178.
Rosch, E., \& Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. Cognitive Psychol, 7, 573-605.
Rummery, G. A., \& Niranjan, M. (1994). On-line q-learning using connectionist systems (Tech. Rep. No. CUED/FINFENG/TR 166). Cambridge University.
Schultz, W., Dayan, P., \& Montague, P. R. (1997). A neural substrate of prediction and reward. Science, 275, 15931599.

Shawe-Taylor, J., \& Cristianini, N. (2004). Kernel Methods for Pattern Analysis. Cambridge University Press.
Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. Science, 237, 1317-1323.
Sloman, S. A., Love, B. C., \& Ahn, W. K. (1998). Feature centrality and conceptual coherence. Cognitive Sci, 22, 189-228.
Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. Mach Learn, 3, 9-44.
Sutton, R. S., \& Barto, A. G. (1998). Reinforcement Learning: An Introduction. The MIT Press.
Tenenbaum, J. B., Kemp, C., Griffiths, T. L., \& Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. Science, 331, 1279-1285.
Tesauro, G. (1995). Temporal difference learning and tdgammon. Commun ACM, 38(3), 58-68.
Tomlinson, M. T., \& Love, B. C. (2006). From pigeons to humans: Grounding relational learning in concrete examples. Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06), 199-204.
Van Otterlo, M. (2012). Solving relational and first-order logical markov decision processes: A survey. Reinforcement Learning, 12, 253-292.

# Developmental and postural changes in children's visual access to faces 

Michael C. Frank, Kaia Simmons, Daniel Yurovsky, \& Guido Pusiol<br>\{mcfrank,kaias,yurovsky, pusiol\}@stanford.edu<br>Department of Psychology<br>Stanford University


#### Abstract

The faces of other people are a critical information source for young children. During early development, children undergo significant postural and locomotor development, changing from lying and sitting infants to toddlers who walk independently. We used a head-mounted camera in conjunction with a face-detection system to explore the effects of these changes on children's visual access to their caregivers' faces during an in-lab play session. In a cross-sectional sample of 4-20 month old children, we found substantial changes in face accessibility based on age and posture. These changes may translate into changes in the accessibility of social information during language learning.


Keywords: Social development; face processing; headcamera.

## Introduction

A father offers his young daughter a novel object: a bright yellow feather duster. A few moments after she accepts the toy, he remarks, "Isn't the zem funny?" Her father may still be talking about the feather duster, or he may be describing a new object. To find out she has access to a simple and reliable method: she can look to his face to infer the direction of his attention.

The ability to follow social signals like eye-gaze is an important part of early social cognition (Scaife \& Bruner, 1975) and a strong predictor of children's early language development. For example, Brooks and Meltzoff (2005) found that children who followed an experimenter's gaze better before their first birthday had larger vocabularies at 18 months. Similarly, Carpenter, Nagell, and Tomasello (1998) found that children's level of joint engagement (as well as the degree to which mothers followed the child's focus of attention in their labeling) predicted vocabulary growth in both language production and comprehension. These studies suggest that children's social environment plays a powerful supportive role in language learning.

But at the same time as children are beginning to learn their first words, their view of the world is changing radically (Adolph \& Berger, 2007). As speechless infants, they are unable to locomote independently. Before their first birthday, they begin crawling; soon after, they begin to walk independently. Infants' visual field is subject to the whims of their caregivers, but caregivers often place them in positions conducive to joint attention. In contrast, toddlers determine their own input to a much greater degree, but as a consequence they spend much of their time in a world primarily populated by knees. These postural and locomotor changes may have a profound effect on what children see.

A recent study suggests the possibility of links between motor milestones, social cognition, and language. Walle and

Campos (under review) noted robust correlations between children's ability to walk and their vocabulary, both receptive and productive. On the basis of an observational study of parent input, they speculated that the emergence of walking may change the ability of the child to access social information (because walking toddlers see more of the social world than crawling infants). Accessing more social information may in turn allow children to discover word meanings more effectively.

Recent methodological developments have the potential to provide data that would allow this hypothesis to be tested. The availability of head-mounted cameras and eye-trackers allows for the measurement of children's naturalistic environment in a way that was not previously possible. Yoshida and Smith (2008) gave the first demonstration of the radical differences between toddler and adult perspectives on the social world, with toddlers' visual field being dominated by hands and objects much more than that of adults. More recent work has used head-mounted eye-tracking methods to measure young toddlers' fixations (Franchak, Kretch, Soska, \& Adolph, 2011), also finding that children look relatively infrequently at their mothers' faces in naturalistic play.

These methods are now being applied to understand inputs to language acquisition. Work by Yu, Smith, and colleagues suggests that word learning is facilitated when parents and children create moments in which the visual field is dominated by a single object (Smith, Yu, \& Pereira, 2011; Yu \& Smith, in press). Some data even suggest that young children's restricted viewpoint may be more effective for learning words than the comparable adult perspective (Yurovsky, Smith, \& Yu, in press). Together, this body of evidence suggests that measuring infants' perspective-and how it changes in motor development-is a critical part of understanding early language learning.

In the current study we took a developmental approach to understanding the relationship between perspective and access to social information. We recorded head-camera data from a group of infants and children across a broad age range as they played with their caregivers during a brief laboratory visit. We then hand-annotated these data for the child's posture and parents' naming behavior and used face-detection algorithms to measure the frequency of faces in the child's visual field. The resulting dataset allows us to analyze changes in access to faces according to children's age, posture, and linguistic input.


Figure 1: Our light-weight, low-cost head-mounted camera with fisheye lens.

## Methods

## Participants

Participants were 20 infants and children ( $\mathrm{N}=4$ each at 4 , $8,12,16$, and 20 months, 9 females total) in an ongoing large-scale study, recruited from the surrounding community via state birth records. Participants had no documented disabilities and were reported to hear at least $80 \%$ English at home. Success rates for children wearing the camera for long enough to initiate the play session varied from $100 \%$ at 8 months to approximately $50 \%$ at 20 months.

## Head-mounted camera

Our head-mounted camera ("headcam") is composed of a small, inexpensive MD80 model camera attached to a soft elastic headband from a camping headlamp. An aftermarket fisheye lens intended for iPhones and other Apple devices is attached to increase view angle. The total cost of each camera is approximately $\$ 60$. The camera captures $720 \times 480$ pixel images at approx. 25 frames per second, and has battery life of $60-90$ minutes. Without the fisheye lens, the viewing angle for the camera is $32^{\circ}$ horizontal by $24^{\circ}$ vertical; with the fisheye, $64^{\circ}$ horizontal by $46^{\circ}$ vertical. The device is pictured in Figure 1.

The vertical field of view of the camera was considerably smaller than the child's approximate vertical field of view, which-even at 6-7 months-spans around $100-120^{\circ}$ in the vertical dimension (Mayer, Fulton, \& Cummings, 1988; Cummings, Van Hof-Van Duin, Mayer, Hansen, \& Fulton, 1988). We were therefore faced with a choice in the orientation of the camera. If we chose a lower or higher orientation, we would be at risk of truncating either the child's own hands and physically proximate objects, or the faces of the adults around the child. Yet if we chose the middle orientation, we would still be at risk of underestimating the proportion of faces viewed by the child. Thus, for the purposes of the current study-measuring visual access to faces-we chose to orient the camera towards the upper part of the visual field. ${ }^{1}$ While this orientation decreased our chances of

[^63]recording the objects being manipulated by the child, it nevertheless allowed us to capture the majority of the faces in the child's visual field.

## Procedure

After coming to the lab, families were seated in our waiting room where they signed consent documents and where children were fitted with the headcam. After a short period of play, they were escorted to a playroom in the lab where the free-play session (the focus of the current study) was conducted.

In the waiting room, the experimenter placed the headcam on children's heads after they had time to adjust to the environment. For children who resisted wearing the headcam, the experimenter used distractor techniques (presenting stimulating toys or engaging the children in hand-occupying activities) intended to keep children's focus elsewhere and prevent them from taking off the camera (Yoshida \& Smith, 2008). Once the child was wearing the camera comfortably for a period of time, child and caregiver (or caregivers: in two cases, there were two adults present) were escorted to the playroom.

In the playroom, the experimenter presented the child's parent with a box containing three labeled pairs of objects, each consisting of a familiar and a novel object (e.g. a ball and a feather duster, marked as a "zem"). Parents confirmed that the child had not previously seen the novel toys. Parents were instructed to play with the object pairs with the child, one at a time, "as they typically would" and to use the novel labels to refer to the three toys. After giving these instructions, the experimenter left the room for a period of approximately 15 minutes. During this time, a tripod-mounted camera recorded video from a corner of the room and the headcam captured video from the child's perspective.

## Data Processing and Annotation

All headcam videos were cropped to exclude the period of entry to the playroom and were automatically synchronized with the tripod-mounted videos using FinalCut Pro Software. The final sample was approx. 5 hours of headcam video (M $=12 \mathrm{~min}$, range: $2-21 \mathrm{~min}$, for a total of roughly 400,000 frames.

Posture and Orientation Annotation One major goal of our study was to understand the relationship between children's posture and their access to information from the faces of their caregiver. To investigate this relationship, we created a set of annotations for the child's physical posture (e.g. standing, sitting) and orientation of the caregiver relative to the child (e.g. in front of, behind, close, far away). For each headcam video, a coder used OpenSHAPA software to annotate both orientation and posture (Adolph, Gilmore, Freeman, Sanderson, \& Millman, 2012).

Orientation was initially categorized as having the caregiver in front, to the side, or behind the child, and close (de-

[^64]

Figure 2: Sample frames from the headcam videos for a child from each age group, selected because they featured successful face detections (green squares).
fined informally as within arm's reach) or farther away. Because of data sparsity, we consolidated this scheme into three categories: close to the caregiver with the caregiver either in front or on the side, farther from the caregiver again with caregiver either in front or on the side, and a global category of caregiver behind the child. Posture was categorized as being held/carried, lying face-up, sitting, prone (crawling or lying), standing, or other. Data from when the child was out of view of the tripod camera was marked as uncodable and excluded from these annotations.

Labeling Annotation We were also interested in the availability of social information proximate to naming events in the caregivers' speech to children. Accordingly, a human coder marked the onset time when the name of any of the six objects in the object set was used. Overall, caregivers produced a median of 35 labels in a highly skewed distribution across participants (range: 9-131).

## Face Detection

An additional goal of the study was to measure the presence of caregivers' faces in the child's field of view (as approximated by the headcam). To avoid hand-annotating the size and position of faces in every frame of video, we tested two face detection systems. Sample frames from the video with successful detections are given in Figure 2.
Face detection algorithms The first algorithm was based on freely available computer vision tools (Bradski \& Kaehler, 2008) and is described in depth in our previous work (Frank, 2012). This system had two parts. The first was the application of a set of four Haar-style face detection filters (Viola \& Jones, 2004) to each frame of the videos independently. These detectors each provide information about whether a face is present in the frame as well as size and position for any detections. In a second step, these detections are then combined via a hidden Markov model (HMM), trained on hand-annotated data (see Appendix). The HMM model (which performed nearly as well as the more complex and computationally-intensive Conditional Random Field model used in our previous work) attempted to estimate whether a face was truly present in each frame of the videos, using as its input the number of Haar detectors that were active in any given frame.

The second algorithm that we evaluated was a semiautomated adaptive tracker-by-detection (SAATD). The algorithm required manual user input (selecting a single face example per video) for its initialization, but then needed no additional training data. The tracker is based on Kalal, Mikolajczyk, and Matas (2010) which uses patches in the trajectory of an optical-flow based tracker (Lucas \& Kanade, 1981) to train and update a face detector. The optical flow tracker and the face detector work in parallel. If the face detector finds a location in a new frame exhibiting a high similarity to its stored template, the tracker is re-initialised on that location. Otherwise, the tracker uses the optical flow to decide the location of a face in the new frame. The primary advantage of the SAATD algorithm is the use of motion for face detection: Following the movement of the pixels that define a face it is possible for the algorithm to adapt to new morphologies (i.e. different face poses).

Detector evaluation To ensure that our evaluation was not biased by the relatively rare appearance of faces in the dataset, we annotated two samples, both a random sample from the data and a sample with a high-density of faces (see Appendix). We evaluated each algorithm on its precision (hits / hits + false alarms) and recall (hits / hits + misses), as well as F-score (the harmonic mean of these two measures). Results are reported in Table 1.

The HMM model obtained a relatively high level of performance for the random subsections, but performed poorly when there was a relatively high density of faces present. In contrast, SAATD performed well on both samples, giving better performance especially in cases where there was partial occlusion. Our goal in using face-detection algorithms was to provide a measurement technique that eliminated tedious and expensive hand-coding and provided acceptable results. We therefore selected the SAATD model and report detections from this algorithm as an estimate of face presence in all further analyses.

## Results

We report results from three different sets of analyses. First, we explore developmental changes in posture and orientation in our dataset. Next, we explore how these changes affect access to faces, as measured using our face-detection algorithm.


Figure 3: Proportion time in each posture, plotted by child's age (left panel). Proportion time in each orientation relative to the caregiver, again plotted by child's age (right panel). For clarity, the "other" code is not plotted in either figure. Error bars show standard error of the mean across participants.

Table 1: Model performance on gold standard generalization training set dataset. P, R, and F denote precision, recall, and F-score for each of the two samples.

|  | High-density |  |  | Random |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: |
| Model | P | R | F | P | R | F |
| HMM | .55 | .38 | .45 | .89 | .74 | .81 |
| SAATD | .86 | .78 | .81 | .93 | .76 | .83 |

Finally, we report preliminary results on the accessibility of faces during labeling.

## Changes in Posture and Orientation

Our posture coding captured typical developmental milestones (Figure 3, left). Overall, sitting was the most common posture for interactions in the caregiver play session. The youngest infants in our sample mostly sat (with parental assistance), but also lay down and were carried a significant proportion of the time. The 12-month-olds were the only group who spent a large amount of time crawling, and the 16- and 20-month-olds sat and stood in equal parts.

Similarly, our coding of orientation revealed some significant developmental changes (Figure 3, right). Younger children more frequently had the caregiver behind them, often because the caregiver was supporting the child's sitting posture (for the 4-month-olds especially). In contrast, the 12-20 month olds were able to locomote independently and so were able to spend more time further from the caregiver.

## Access to Faces

We next investigated the effects of the child's posture and orientation on the presence and size of the caregiver's face in the visual field. Figure 4 shows the proportion of frames with a positive face detection, plotted by the child's age, posture, and orientation relative to the caregiver.

Overall, there were very large differences in access to faces across age. The 4-month-olds saw almost no faces-their parents were behind them most of the time, supporting them
since they could not sit independently. In contrast, the 8-month-olds, who could sit independently, typically sat across from their caregiver and saw many faces in both the sitting and prone postures. The 12-month-olds spent a large amount of time in the prone position (typically crawling after the ball, for example) and saw almost no faces in that posture. The 16and 20-month-olds saw many faces because they were standing while their parents were sitting, putting their faces at a relatively similar level.

Across ages, the carrying and prone postures resulted in the smallest number of faces seen, while standing and sitting resulted in far more. These postures both presented opportunities for seeing faces in large part because parents were sitting or lying on the floor with children. Although far fewer faces were seen when the caregiver was behind the child, ${ }^{2}$ both the close and far positions resulted in approximately equal proportions of face detections.

## Access to Faces During Labeling

Our final analysis concerned the accessibility of caregivers' faces during labeling events. Franchak et al. (2011) found that referential speech was marginally more likely to draw toddlers' attention to mothers' faces. We were similarly interested in whether looking at faces occurred during labeling. Accordingly, we used the labeling annotations for each child to identify the 2 s before and after each labeling event. We then computed the proportion face detections within this window across ages.

The overall pattern of face accessibility closely mirrored the base rates shown in Figure 4. Although this general pattern in itself is important in assessing developmental access to social information, in the current analysis we were interested in whether there was differential access to faces around labeling instances. We thus computed difference scores between the baseline face detection rate and the rate of face detections

[^65]

Figure 4: Proportion face detections, split by age group (left panel), posture (middle panel), and caregiver's orientation (right panel). We omit the lying face-up posture due to data sparsity. Black points show individual participants and are jittered slightly on the horizontal, red lines show means and $95 \%$ confidence intervals.


Figure 5: Proportion of faces detected in a 4 s window of time centered around labeling events, plotted by age group and whether the word was familiar or novel. Error bars show $95 \%$ confidence intervals. 4-month-olds are omitted due to the limited number of total face detections for this group.
in labeling windows for each participant. Figure 5 shows the results of this analysis.

Although any conclusion must remain extremely tentative because of the small sample, we nevertheless saw an increase in label-related face access for the 20 -month-olds. This difference was robust across a variety of window sizes from 1-6 s. (8-month-olds were more variable but similarly showed some trend towards greater face access during naming.) We cannot yet make inferences about the source of these differences: They could be could be caused by children, caregivers, or a combination of the two. Nevertheless, these results converge with previous work and suggest that, in combination with face detection techniques, the headcam may be a viable
method for examining social access during language learning.

## General Discussion

Using a head-mounted camera, we explored the relationship between infants' postural and locomotor development and their visual access to social information. The use of automated annotation tools from computer vision allowed us to measure the prevalence of caregivers' faces in their children's visual field. We found systematic differences in the visual accessibility of faces based on posture, orientation, and age, as well as hints of differences in language-related changes in visual access. While these results remain preliminary given the size of our developmental sample, this work nevertheless provides an important proof-of-concept that computer vision techniques can be used as a measurement method in the developmental context.

The measures developed here have broad applicability to the study of individual and cultural differences. Since the physical circumstances of child rearing vary widely across households and across cultures, there may be important and predictable differences in children's visual experience. As suggested by the correlations between walking and vocabulary development (Walle \& Campos, under review), postural development may have substantial downstream consequences for language. For example, shifts in how infants are placed in particular postures by strollers or carriers (Zeedyk, 2008) or how their motor development is encouraged by parent practices (Bril \& Sabatier, 1986) may lead to differences in social input which in turn affect their language learning. Since our variant of the headcam method is both inexpensive and highly portable, we have been able to deploy it in children's homes with some success; it may thus be a valuable tool for investigating differences in child-rearing practices.

A deep body of work uses children's linguistic inputmeasured using audio recordings-to understand the learning mechanisms underlying vocabulary acquisition
(Huttenlocher, Haight, Bryk, Seltzer, \& Lyons, 1991; Hart \& Risley, 1995; Fernald, Perfors, \& Marchman, 2006). There have been some important initial successes in using visual input to predict language uptake ( $\mathrm{Yu} \&$ Smith, in press). Nevertheless, we have a long way to go before our knowledge about children's visual input parallels our understanding of their linguistic environment. Coming to such an understanding will require the creation of both corpus resources and automated tools such as those we have begun to develop here.

## Acknowledgments

Thanks to Ally Kraus, Kathy Woo, Aditi Maliwal, and other members of the Language and Cognition Lab for help in recruitment, data collection, and annotation. This research was supported by a John Merck Scholars grant to MCF.

## References

Adolph, K., \& Berger, S. (2007). Motor development. In Handbook of child psychology. Wiley Online Library.
Adolph, K., Gilmore, R., Freeman, C., Sanderson, P., \& Millman, D. (2012). Toward open behavioral science. Psychological Inquiry, 23(3), 244-247.
Bradski, G., \& Kaehler, A. (2008). Learning OpenCV: Computer vision with the OpenCV library. O'Reilly Media.
Bril, B., \& Sabatier, C. (1986). The cultural context of motor development: Postural manipulations in the daily life of Bambara babies (Mali). International Journal of Behavioral Development, 9(4), 439-453.
Brooks, R., \& Meltzoff, A. (2005). The development of gaze following and its relation to language. Developmental Science, 8(6), 535-543.
Carpenter, M., Nagell, K., \& Tomasello, M. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. Monographs of the Society for Research in Child Development, 63(4).
Cummings, M., Van Hof-Van Duin, J., Mayer, D., Hansen, R., \& Fulton, A. (1988). Visual fields of young children. Behavioural and Brain Research, 29(1), 7-16.
Fernald, A., Perfors, A., \& Marchman, V. (2006). Picking up speed in understanding: Speech processing efficiency and vocabulary growth across the 2nd year. Developmental Psychology, 42(1), 98-116.
Franchak, J., Kretch, K., Soska, K., \& Adolph, K. (2011). Head-mounted eye tracking: A new method to describe infant looking. Child Development.
Frank, M. C. (2012). Measuring children's visual access to social information using face detection. In Proceedings of the 33nd Annual Meeting of the Cognitive Science Society.
Hart, B., \& Risley, T. (1995). Meaningful differences in the everyday experience of young American children. Baltimore, MD: Brookes Publishing Company.
Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., \& Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. Developmental Psychology, 27(2), 236248.

Kalal, Z., Mikolajczyk, K., \& Matas, J. (2010). Forwardbackward error: Automatic detection of tracking failures. In 20th International Conference on Pattern Recognition (pp. 2756-2759).
Lucas, B., \& Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In Proceedings of the 7th International Joint Conference on Artificial intelligence.
Mayer, D., Fulton, A., \& Cummings, M. (1988). Visual fields of infants assessed with a new perimetric technique. Investigative Ophthalmology \& Visual Science, 29(3), 452-459.
Scaife, M., \& Bruner, J. (1975). The capacity for joint visual attention in the infant. Nature, 253, 265-266.
Smith, L. B., Yu, C., \& Pereira, A. F. (2011). Not your mother's view: The dynamics of toddler visual experience. Developmental Science, 14, 9-17.
Viola, P., \& Jones, D. H. (2004). Robust real-time face detection. International Journal of Computer Vision.
Walle, E. A., \& Campos, J. J. (under review). Infant language development is related to the acquisition of walking.
Yoshida, H., \& Smith, L. (2008). What's in view for toddlers? using a head camera to study visual experience. Infancy, 13, 229-248.
Yu, C., \& Smith, L. B. (in press). Embodied attention and word learning by toddlers. Cognition.
Yurovsky, D., Smith, L., \& Yu, C. (in press). Statistical word learning at scale: The baby's view is better. Developmental Science.
Zeedyk, M. (2008). What's life in a baby buggy like?: The impact of buggy orientation on parent-infant interaction and infant stress (Tech. Rep.). London, UK: National Literacy Trust.

## Appendix: Face presence annotation

We selected 1 minute of interaction for each age group, divided evenly across the four dyads at that age. For each dyad, we divided the recorded video into contiguous 1 s segments and selected 16 in accordance with two criteria. First, 8 of these segments were selected by choosing the parts of the videos highest in face detection (high density sample). To be fair to both algorithms, half of this was chosen from the segments with the most HMM detections and half were chosen from the segments with the most SAATD detections. The remaining segments were chosen by randomly sampling from segments not yet selected for coding (random sample). Segments were annotated frame-by-frame by a human coder, who marked each frame as containing a face if at least half of the face was in the child's view. Detector output for each of these frames was then compared to this gold standard. A detection was counted as correct if it overlapped a face with half of its total area. The HMM training sample was selected via the same method as this gold standard sample, but used separate set of video segments.

# The Effect of Incremental Context on Conceptual Processing: Evidence from Visual World and Reading Experiments 

Diego Frassinelli (d.frassinelli@sms.ed.ac.uk)<br>Frank Keller (keller@inf.ed.ac.uk)

Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, UK
Christoph Scheepers (Christoph.Scheepers@glasgow.ac.uk)
Institute of Neuroscience and Psychology, University of Glasgow
58 Hillhead Street, Glasgow, G12 8QB, UK


#### Abstract

The analysis of the internal structure of concepts reveals the presence of a substantial amount of contextual information. Even though this interaction is easily recognizable, it is not clear how contextual information is processed and included into concept representations. The aim of this paper is to shed light on this question by analyzing the effect that an increasing amount of context exerts on conceptual processing. We report a self-paced reading experiment and a visual world experiment to test two hypotheses about the integration of context information: the incremental activation hypothesis suggests that the degree of facilitation in concept processing increases with the amount of context available; and the immediate activation hypothesis states that once a sufficient amount of contextual support is reached, no more facilitation occurs. Our data are compatible with the latter account.


Keywords: incremental processing; context effects; eye movements; visual world; self-paced reading.

## Introduction

Contextual information plays an essential role in different cognitive domains, including language processing, visual processing, or reasoning. In this paper we investigate the effect that context exerts on conceptual processing (Murphy, 2002). A widely recognized way to analyze the internal structure of concepts is the use of semantic feature norms (Wu \& Barsalou, 2009; McRae, Cree, Seidenberg, \& McNorgan, 2005). In experiments eliciting feature norms, participants enumerate features associated with a target concept; these lists can then be used to shed light on the internal structure of concepts, including the role of context (Frassinelli \& Keller, 2012). However, feature norms are static, and do not allow to study the time-course of concept processing. Real-time data are required to investigate how conceptual representations are constructed, and how they interact with conceptual information. Relevant prior work includes a study by Huettig, Quinlan, McDonald, and Altmann (2006), which analyzed the activation of conceptual information over time using the visual world paradigm (Tanenhaus, Spivey-Knowlton, Eberhard, \& Sedivy, 1995). However, Huettig et al. (2006) were not interested in contextual effects and used contexts that were as neutral as possible. Frassinelli and Keller (2012) replicated this study, but also introduced contextual variability as a factor, comparing the effect of a neutral context and two biasing
contexts associated with the target concept in different ways. However, Frassinelli and Keller (2012) focused their analysis on the concept region, providing only indirect evidence regarding effects taking place in the previous part of the sentence, where contextual information is integrated.

The aim of the present paper is to study contextual constraints on the processing of context, focusing in particular on the question of how context is integrated with conceptual information. According to Federmeier and Kutas (1999), contextual facilitation effects can occur in different tasks (and affect, e.g., reading time, lexical decision times, pronunciation time); effects appear both at the level of lexical priming and at the level of the entire sentence. Based on this assumption, we aim to clarify the relation between single contextual words and the entire sentence. We constructed sentence materials that provide a differential number of context words that bias comprehension towards the target concept. This design allows us to determine how such facilitation (or bias) occurs, and to distinguish two possible hypotheses: the incremental activation hypothesis, which suggests that the degree of facilitation in concept processing increases with the amount of context available; and the immediate activation hypothesis, which states that once a sufficient amount of contextual support is reached, no more facilitation occurs.

We performed two experiments. In a self-paced reading experiment, we tested if the amount of context available has an effect on the ease of conceptual processing, measured as the reading time for the target concept. In a visual world study, we presented the target concept pictorially, allowing us to measure the degree of facilitation (i.e., the number of looks) that occurs while the context words are processed, giving us access to the time course of conceptual integration.

## Experiment 1: Self-Paced Reading

The aim of this experiment was to analyze the effect that context information has on participants' reading time for the target concept. As widely discussed in the reading literature (Morris, 1994), higher coherence between the context and the target word is reflected in lower reading time. We therefore predict that reading times for a target word representing a
concept are reduced in proportion to the number of context words presented that bias the reader towards the concept.

## Method

Materials We used the same 24 concepts as Huettig et al. (2006) (but we dropped their semantically related condition, as it is not relevant for the present experiment). We embedded these concepts into a sentential context using the following general structure:

## (1) location - actor - verb - object - target concept -spill-over region

The target word is in bold, the three context words in italics (see below for an example). For each target concept, we identified three context words which were highly related to it (high-biasing (HB) words) and three context words that were unrelated to it (low-biasing (LB) words). Of the resulting eight possible combinations of LB and HB context words, we chose four, illustrated by the following examples:
(2) All LB context (None): On the path, the man was holding a box full of mushrooms carefully.
(3) HB location context (Loc): In the forest, the man was holding a box full of mushrooms carefully.
(4) HB location and actor context (LocAct): In the forest, the picker was holding a box full of mushrooms carefully.
(5) All HB context (All): In the forest, the picker was holding a basket full of mushrooms carefully.

This resulted in 96 experimental sentences: four contexts for each of the 24 concepts.

Norming Studies In order to make sure that the context words we chose had the biasing effect we expected them to have, we conducted a series of norming studies on Amazon Mechanical Turk.

First, 20 participants performed a sentence plausibility judgment task: they assessed how plausible the experimental sentence was by rating it on a scale from 1 (completely implausible) to 7 (completely plausible). A sentence was considered plausible when the averaged rating was higher than 4. This process allowed us to identify those sentences that were not completely plausible; they were replaced and re-tested.

A sentence completion study then evaluated the predictability of the target concept from the sentence context. Twenty new participants had to complete each sentence (with the target concept removed) by typing in a noun. An Anova showed a statistically significant difference between the None ( $6.4 \%$ of correct answers) and the All (43\%) condition $(F(1,3)=1.84, p<.001)$. The Loc (19.2\%) and the Lo$\operatorname{cAct}(14.4 \%)$ were not significantly different $(F(1,3)=0.64$ and $F(1,3)=0.40)$.

Twenty-four further participants performed a word completion study: the three context words appeared on the screen
one after the other and then participants had to type a word related to the context word. The aim of this study was to exclude syntactic effects (word order, but also the effect of the verb) in the completion study. The outcome of this experiment was in line with the sentence completion experiment, despite the fact that the context words were presented in isolation rather than in a sentence.

Finally, sixty participants performed a word association study in order to test the associations between the six context words (three LB and three HB words) and the target word in the sentence. According to an Anova, the three HB context words were equally strongly associated with the target word. Furthermore, all of them were more associated with the target than the LB words.

Procedure The 96 experimental sentences were distributed over four lists of 24 items each according to a Latin square design. Twenty-seven fillers were added and the list randomized for each participant. Twenty yes/no questions about the sentence were also included.

Thirty-four native English speakers from the University of Edinburgh took part in the experiment after giving informed consent and were paid $£ 5$. Each saw one of the lists. We excluded one participant based on a low percentage of correct answers ( $<50 \%$ ) and another participant with a reading time averaging 2.5 standard deviations above the grand mean (as suggested in Hofmeister (2011)).

We used a moving-window self-paced reading procedure (Just, Carpenter, \& Woolley, 1982), in which participants read a sentence on the screen at their own pace. At the beginning of each trial, all the words in the sentence are masked with dashes and separated by spaces; participants had to press the space bar to uncover the next word and hide the previous one. For this experiment, we used the software package Linger (version 2.94) on Apple computers.

## Results

We analyze the reading times associated with the target word, which can be assumed to index the amount of effort associated with processing the target concept. Table 1 shows the mean reading times for the target concept across the four context conditions. The results indicate that reading time decreases in proportion with the number of HB words in the context. Table 2 reports a linear mixed-effects model (LME) with log-transformed reading time as the dependent variable. In the model, the factors Loc, LocAct, and All were contrast coded against the reference level None. Participant and Item were included as random intercepts, and as random slopes under Loc, LocAct, and All (thus implementing a maximal random effects structure, D. Barr, Scheepers, and Tily (2013)). The LME model shows that LocAct and All are statistically different from the None condition. After a HB context, participants spend less time reading the target concept compared to a LB context. The difference in reading time between None and the Loc does not reach significance.

| Condition | Reading Time |
| :--- | :---: |
| None | $390.1 \pm 25.2$ |
| Loc | $358.5 \pm 21.4$ |
| LocAct | $346.0 \pm 15.5$ |
| All | $336.7 \pm 10.5$ |

Table 1: Reading time (in ms) with standard errors for the target concept in the four contextual conditions.

| Predictor | Coefficient |
| :--- | :---: |
| (Intercept) | $12.771^{* * *}$ |
| Loc | -0.067 |
| LocAct | $-0.097^{*}$ |
| All | $-0.094^{*}$ |
| ${ }^{*} p<.05,{ }^{* *} p<.01,{ }^{* * *} p<.001$ |  |

Table 2: Coefficients for the mixed effects model for the reading time data in Table 1.

## Discussion

In this self-paced reading experiment, we looked at the relation between the amount of HB information in the context and the reading time of the target concept. The hypotheses we started with predict two distinct outcomes: the incremental activation hypothesis predicts that the reading time of the target concept is reduced in proportion to the number of HB context words that are present. On the other hand, based on the immediate activation hypothesis, we expect that there is a threshold on the amount of contextual information that is required before an effect of context on reading time occurs.

Descriptively, the reading times in Table 1 are compatible with incremental activation: each additional contextual concept results in a further reduction in reading time. However, the mixed model analysis in Table 2 shows significant differences only between LocAct and None and All and None; Loc on its own does not have a significant effect. Furthermore, post-hoc tests failed to show significant differences between the reading times in the three context conditions (Loc, LocAct and $A l l)$. This is a pattern that would be expected under the immediate activation hypothesis: the contextual threshold has been reached at LocAct, and additional context words do not significantly pre-activate the target concept any further.

Taken together, the results of this experiment are inconclusive. We therefore performed a follow-up experiment which directly tests our two hypotheses by measuring the amount of activation the target concept receives during the processing of each context word.

## Experiment 2: Visual World

The aim of this experiment was to test whether the activation of concepts happens gradually (more activation with every new context word, as predicted by incremental activation), or at once (the first context word triggers full activation, which
then declines, as predicted by immediate activation). In a visual world study, we measure the amount of activation for the target concept at each context word in terms of the proportion of looks received by the object corresponding to the target concept.

## Method

Materials We used visual scenes that consisted of four black and white line drawings extracted from the Snodgrass and Vanderwart (1980) collection: one target object and three distractors randomly arranged in four quadrants. (These were the same stimuli as in Frassinelli and Keller (2012), already normed by Huettig et al. (2006)). The sentence materials were the same as in Experiment 1, and the stimuli instantiated the same 24 target concepts. As an example, consider the visual stimulus in Figure 1, which corresponds to the sentences in (2)-(5).


Figure 1: Example of the scene for the target concept mushroom (the box is not shown to the participants).

Procedure The 96 sentences in the experiment were spoken by the speech synthesis system Festival (Clark, Richmond, \& King, 2007) using an HMM voice (Roger), so as to reduce possible effects of prosody or speaker variation.

In order to counterbalance order or position effects, we rotated the four objects on the screen. The resulting 384 items were distributed over 32 lists of 24 items each according to a Latin square design. Twenty-five fillers were added and the list randomized for each participant. Nine yes/no questions about the sentence were also presented.

Thirty-four native English speakers from the University of Edinburgh took part in the experiment after giving informed consent and were paid $£ 5$. Each saw one of the lists.

Participants were sat in front of a 21 " multi-scan monitor with a resolution of $1024 \times 768$ pixels and their eye movements were recorded using an EyeLink II head-mounted eyetracker with a sampling rate of 500 Hz . Only the dominant eye was tracked. At the beginning of the experiment and after every ten trials, the eye-tracker was recalibrated using a ninepoint randomized calibration. Before each trial, drift correction was performed. At the beginning of each trial the scene
appeared on the screen, and the sentence began to play at the same time; the scene disappeared 1500 ms after the end of the sentence. The experiment lasted approximately 20 minutes.

Data Analysis The analysis is based on the proportion of fixations on the target object across experimental conditions. We excluded out-of-screen fixations and blinks from the analysis.

In order to analyze the effects exerted by a context word before and after its acoustic offset, we aligned the fixation probabilities at that point ( 0 ms ). In order to exclude any overlap between two regions of analysis in the sentence we calculated the minimum amount of time between the onset and the offset of the context word ( 150 ms ) and between the offset of the context word and the onset of the following one ( 400 ms for location, actor, and concept; 150 ms for object). The vertical line shows the offset of the context word, while the horizontal dotted line indicates the probability of randomly fixating on one of the four objects depicted on the screen ( $25 \%$ of total fixations).

For each context word we report an LME analysis of the results. As suggested by D. J. Barr (2008), the dependent variable of our models is the empirical logit of the fixation probability calculated for each bin. We used a bin size of 5 ms . To compare the effects produced by HB and LB contexts, we included three factors in contrast coding: each factor encodes the differences between the reference level None (coded as -.5) and one of the three other conditions (Loc, LocAct, All; coded as .5). The continuous factor Time shows variations over time. In order to identify the minimal model that best fits our data, we used the best-path forward model selection procedure (recommended by D. Barr et al. (2013) if a model with full random effects structure fails to converge). We report only the coefficients and the significance levels for the minimal model, i.e., we show only the main effects and the interactions included during the selection procedure. All models included Participant and Item as random intercepts, as well as random slopes for Context and Time.

## Results

Location Word The first context word we analyze is location. The plot in Figure 2(a) shows the probability of fixating the target object at this context word. Before its offset, it is already possible to identify a general effect produced by the presence of a location (both HB and LB), as the fixation probabilities are higher than random. However, specific effects appear only 100 ms after the offset of the context word. The None (low biasing) condition shows a decrease over time, while an increase in fixation probability is observed in the Loc and LocAct conditions (compared to None, the reference level), which corresponds to the significant interactions Time: Loc and Time: LocAct (see Table 3, column 1). A similar effect is visible for All, but fails to reach significance.


Figure 2: Fixation probabilities aligned at the offset ( 0 ms ) of the context words.

| Predictor | Coefficient Location | Coefficient Actor | Coefficient Object | Coefficient Concept |
| :--- | :---: | :---: | :---: | :---: |
| (Intercept) | $-0.7548^{* * *}$ | $-0.6764^{* * *}$ | $-0.6215^{* * *}$ | -0.0138 |
| Time | 0.0005 | $0.0002^{* * *}$ | 0.0003 | $0.0003^{* * *}$ |
| Loc | 0.1019 | -0.7303 | 0.1955 | -0.1110 |
| LocAct | -0.0388 | 0.4809 | -0.0744 | -0.2919 |
| All | - | -0.1136 | 0.2204 | -0.0849 |
| Time:Loc | $0.0008^{* * *}$ | $-0.0011^{* * *}$ | - | $0.0003^{* *}$ |
| Time:LocAct | $0.0003^{* *}$ | - | $-0.0024^{* * *}$ | $-0.0007^{* * *}$ |
| Time:All | - | - | $0.0011^{* * *}$ | $-0.0004^{* * *}$ |
| $\quad{ }^{*} p<.05,{ }^{* *} p<.01,{ }^{* * *} p<.001$ |  |  |  |  |

Table 3: Coefficients for the mixed effects model for the data in Figures 2(a), 2(b), 2(c), 2(d). Empty cells indicate that the factor in question was not included during model selection.

Actor Word The plot in Figure 2(b) shows the fixations at the word encoding an actor. In the Loc condition, participants tend to fixate less on target object (compared to None), an effect that is more evident before the offset of the word. This corresponds to a significant negative interaction Time: Loc in the LME (see Table 3, column 2). The plot also seems to indicate an overall higher level of fixations in the LocAct condition (compared to $A l l$, which is identical at this point in the sentence). However, this difference is not significant (no main effect or interaction involving LocAct in the LME).

Object Word The next context word analyzed is the object of the sentence. Figure 2(c) shows that before the offset of the object, the HB conditions all show a higher overall fixation proportion compared to None. After the offset, the curves diverge: All shows a steeper increase than None (significant positive interaction All:Time, see Table 3, column 3). This is explained by the fact that All is the only condition with a HB object. The condition LocAct, shows a steep decrease, i.e., the significant negative interaction LocAct:Time, while Loc remains constant (no significant effects involving Loc).

Concept Word Figure 2(d) shows the number of fixations at the point when participants hear the concept associated with target object on the screen (note the different $y$-axis). At this point, global effects of different amounts of HB information across conditions should be visible. After the offset of the context word, there is an inverse relation between the amount of HB information and the slope of the curves in the Loc, LocAct, and All conditions. The more HB information is available, the sooner fixation proportions decrease. This is consistent with the pattern observed in the reading times of the concept word in Experiment 1.

On the other hand, fixation probability in the None condition increases, in particular after word offset, and remains high. The negative interactions Time:LocAct and Time:All (see Table 3, column 4) are consistent with this observation, indicating a significant decrease in fixations in LocAct and All compared to None. Furthermore, there is a significant positive
interaction Time: Loc, indicating an increase in fixation probability in this condition compared to None.

## Discussion

The aim of this experiment was to analyze the effect of incremental context information over time. The analysis of actor and object (see Figure 2(b) and Figure 2(c)) showed few interesting effects. The regions at which it was possible to identify a clear effect of contextual variability were location and concept. Location is the first context word participants are exposed to and it had a strong effect on driving their fixations towards the target concept. This is in line with previous results of visual world studies on language comprehension (Altmann \& Kamide, 1999), showing anticipatory eye movements towards a target as a result of predictive spoken language input. Less expected are the outcomes related to the concept area. We found that high-biasing contexts allows participants to identify and process the target object at an early stage: this effect is visible even at the first context word (location in our case). At the concept word itself, we then fail to observe a sustained increase in fixations to the target object. The opposite pattern was observed in the low-biasing context: were we see no increase in fixations at the context words, but a sustained increase once the concept word has been processed. In an HB context, the target word is contextually expected, and thus fixated less, while in the LB context, it is unexpected and thus fixated more.

One possible explanation for this pattern of results (i.e., a decrease in target fixations at the target word after a biasing context) is inhibition of return. This is a well-know effect in eye-movements, which manifests itself in a low probability of returning to a region once it has been fixated (Posner, Rafal, Choate, \& Vaughan, 1985). Our failure to find an increase in fixations at the concept region in the HB conditions could be due to inhibition of return, as the target had already been fixated at an earlier point in these conditions (i.e., during anticipatory processing while hearing biasing context words).

## General Discussion

The aim of this study was to test the effect that an incremental amount of contextual information exerts on conceptual processing. We ran two experiments: a self-paced reading and a visual world experiment. The former identified a reduction of the time required to read a target concept with increasing amount of biasing context. The visual world study showed increased looks to the target during the processing of the first context word (location) and at the concept word itself.

The results obtained in our visual world study corroborate the findings of Frassinelli and Keller (2012) who found that a biasing context leads to an early recognition and processing of the target concept. Moreover, the results show an expectation effect: over time, the biasing contexts produced an expectation of the target concept and this led to a reduced number of fixations to the target object when the corresponding concept word occurred (potentially involving inhibition of return as the driving mechanism behind this decrease in fixation probability).

The key novel contribution of the paper is to elucidate the time course of contextual integration. In the introduction, we proposed two possible hypothesis on how context interacts with concept processing: the incremental activation hypothesis (the degree of facilitation increases with the amount of context) and the immediate activation hypothesis (once sufficient contextual support has accrued, no additional facilitation occurs). The results of Experiment 2 allow us to evaluate these two hypotheses. We found that when participants are exposed to a low-biasing context, we see an increase of fixations to the target concept only when that concept is mentioned. In the high-biasing conditions, on the other hand, this increase occurs already at the first context word, with no further increases at the second or third context word. Also when at the concept word, only a small increase in fixations is observed. It seems that a single context word is sufficient to identify the concept on the screen. Additional context exerts only a confirmatory effect. This pattern of results is compatible with the immediate activation hypothesis: a certain amount of contextual information is sufficient to trigger conceptual processing; additional contextual information does not trigger an incremental increase in concept activation.

In more theoretical terms, these results enhance our understanding of conceptual representation. They indicate that the activation of a specific concept takes place when the overlap between its internal structure and the information extracted from the context reach a certain critical level.

## Acknowledgments

This work was funded by the European Research Council under award 203427 "Synchronous Linguistic and Visual Processing". We are grateful to Moreno I. Coco for comments.

## References

Altmann, G., \& Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent refer-
ence. Cognition, 73(3), 247-264.
Barr, D., Scheepers, C., \& Tily, H. (2013). Random-effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language, 68(3), 255-278.
Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. Journal of Memory and Language, 59(4), 457-474.
Clark, R., Richmond, K., \& King, S. (2007). Multisyn: Opendomain unit selection for the festival speech synthesis system. Speech Communication, 49(4), 317-330.
Federmeier, K., \& Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. Journal of memory and Language, 41(4), 469-495.
Frassinelli, D., \& Keller, F. (2012). The plausibility of semantic properties generated by a distributional model: Evidence from a visual world experiment. In N. Miyake, D. Peebles, $\&$ R. P. Cooper (Eds.), Proceedings of the 34th annual conference of the cognitive science society (pp. 1560-1565). Sapporo.
Hofmeister, P. (2011). Representational complexity and memory retrieval in language comprehension. Language and cognitive processes, 26(3), 376-405.
Huettig, F., Quinlan, P. T., McDonald, S. A., \& Altmann, G. T. M. (2006). Models of high-dimensional semantic space predict language-mediated eye movements in the visual world. Acta Psychologica, 121(1), 65-80.
Just, M., Carpenter, P., \& Woolley, J. (1982). Paradigms and processes in reading comprehension. Journal of Experimental Psychology: General, 111(2).
McRae, K., Cree, G., Seidenberg, M., \& McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. Behavior Research Methods, 37(4), 547-559.
Morris, R. (1994). Lexical and message-level sentence context effects on fixation times in reading. Journal of Experimental Psychology: Learning, Memory, and Cognition, 20(1), 92.
Murphy, G. L. (2002). The Big Book of Concepts. Cambridge, MA: MIT Press.
Posner, M., Rafal, R., Choate, L., \& Vaughan, J. (1985). Inhibition of return: Neural basis and function. Cognitive Neuropsychology, 2(3), 211-228.
Snodgrass, J. G., \& Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. Journal of Experimental Psychology. Human Learning and Memory, 6(2), 174-215.
Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., \& Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. Science, 268(5217), 1632-1634.
Wu, L., \& Barsalou, L. W. (2009). Perceptual simulation in conceptual combination: Evidence from property generation. Acta psychologica, 132(2), 173-89.

# The intrinsic structure of spatial configurations and the partner's viewpoint shape spatial memories and descriptions 

Alexia Galati ${ }^{\text {a }}$ (galati@ucy.ac.cy)<br>\&<br>Marios N. Avraamides ${ }^{\text {a, }}$ (mariosav@ucy.ac.cy)<br>${ }^{\text {a }}$ Department of Psychology, University of Cyprus<br>${ }^{\mathrm{b}}$ Centre for Applied Neuroscience, University of Cyprus<br>Nicosia, Cyprus


#### Abstract

We examined how the intrinsic orientation of spatial layouts and the conversational partner's viewpoint shape how people organize spatial information in memory and subsequently describe it. In 24 pairs, Directors first studied an array with a symmetrical structure while either knowing their Matcher's subsequent viewpoint or not. When describing the array to the Matcher, the array's intrinsic orientation was aligned with the Director, the Matcher, or neither partner. Memory tests preceding descriptions revealed that Directors misaligned with the structure organized information according to a priori knowledge, being more likely to use the structure as an organizing orientation when knowing that Matchers were aligned with it. The perspective of Directors' descriptions was also influenced both by the partners' alignment with the structure and their advance knowledge of that. Altogether, speakers are guided by converging social and representational cues to adapt flexibly the organization of their memories and perspectives of their descriptions.


Keywords: perspective-taking; spatial memory; intrinsic structure; audience design; common ground; spatial descriptions

## Introduction

When people make spatial judgments they access memory representations that maintain spatial relations around a preferred direction (e.g., Mou, McNamara, Valiquette, \& Rump, 2004). This preferred direction can be influenced by egocentric preferences for organizing information, based on one's learning perspective (Shelton \& McNamara, 2001) and on representational cues like the symmetry of the spatial configuration (Mou \& McNamara, 2002; Li et al, 2011) or the geometry of the environment (Shelton \& McNamara, 2001). However, the extent to which people take into account their conversational partner's viewpoint when organizing information in memory and communicating this information is still unclear.

A study by Shelton and McNamara (2004) addressed whether describing information from the partner's viewpoint influences speakers' resulting memory representations. Indeed, after describing an array to their partner, speakers were more accurate to make spatial judgments from perspectives aligned with the one that had been occupied by their partner (vs. other perspectives). However, since speakers learned the arrays while describing them from the partner's viewpoint (following explicit
instructions), it's possible that speakers don't spontaneously represent their partner's viewpoint in spatial memory, and instead resort to egocentric preferences for organizing information.

We recently adapted Shelton and McNamara's (2004) study to ask whether in fact speakers spontaneously represent their partner's viewpoint in memory. In Galati et al. (2013), one participant (the Director) first studied a randomly configured array, while either knowing or not knowing their partner's (the Matcher's) subsequent viewpoint, which was misaligned by $90^{\circ}, 135^{\circ}$, or $180^{\circ}$. In memory tests preceding descriptions, rather than finding facilitation for the partner's viewpoint when it was available (cf. Shelton \& McNamara, 2004), we found that speakers represented that viewpoint in memory without using it as an organizing direction. Directors took longer to imagine orienting to perspectives known to be aligned with their Matcher (at least for $90^{\circ}$ and $135^{\circ}$ ) and rotated their array drawings toward the Matcher's viewpoint. Nonetheless, these findings could indicate that, under those circumstances, speakers did not have sufficient pragmatic motivation to invest the cognitive effort to organize spatial relations around a non-egocentric viewpoint, so they simply represented it and used it as needed.

In the present study, our first goal is to elucidate whether, under different circumstances, the partner's viewpoint could be used as an organizing direction in memory. In particular, we ask whether, in collaborative tasks, a given partner's alignment with the array's intrinsic structure affords sufficient pragmatic motivation to organize spatial relations around that viewpoint. Our view is that, when selecting an organizing direction, people consider a confluence of different sources of information, including egocentric cues (e.g., their own learning viewpoint), representational cues (e.g., the array's intrinsic orientation) and social cues (e.g., the partner's viewpoint). Thus, the partner's viewpoint could be used as an organizing direction if it is reinforced by additional cues, like the array's intrinsic orientation. This prediction follows the proposal that in collaboration people try to minimize their collective effort, with one partner investing greater cognitive effort to ensure mutual understanding upon appraising that the other is likely to find the interaction difficult (e.g., Clark \& Wilkes-Gibbs, 1986). In spatial perspective-taking, attributions about the partner's
ability to contribute to the task, based on social cues, should influence whether the partner's perspective is adopted.
This is in line with findings concerning the interpretation of spatial descriptions. For instance, when people believe that their partner doesn't know their viewpoint they are more likely to interpret spatial descriptions from the partner's perspective, whereas when they believe that their partner is real (vs. simulated) they are more likely to interpret them egocentrically presumably because they shift the burden of ensuring mutual understanding to the partner (Duran, Dale, \& Kreuz, 2011). Related findings come from production tasks as well. People invest the cognitive effort to describe information from their partner's perspective when the partner does not share their viewpoint (Schober, 1993), cannot provide feedback (Shelton \& McNamara, 2004), or has worse spatial abilities than them (Schober, 2009). People are also more likely to help their partners by using available environmental features, like the intrinsic axes of objects, as the basis of their descriptions' perspective, instead of their own egocentric perspective (Tenbrink, Coventry, Andonova, \& 2011), and referring to more landmarks for orienting, and navigating along fewer, larger and more prominent streets when describing routes to a partner unfamiliar with the environment (Hölscher, Tenbrink, \& Wiener, 2011).

Thus, the second goal of our study is to examine how people adapt their spatial descriptions when faced with different cues. Specifically, we aim to clarify the extent to which they rely on their memory representations when describing information. We do so by dissociating the learning of spatial arrays from their description (cf., Shelton \& McNamara, 2004). Our earlier work suggests that speakers don't merely rely on their initial representations during descriptions, but are able to use available perceptual information (i.e., their degree of misalignment from their partners) to adapt descriptions appropriately (Galati et al., 2013). Here, we examine whether advance knowledge of the partner's viewpoint guides speakers in selecting a perspective for their descriptions, depending on whether the intrinsic structure is aligned with the speaker, their partner, or neither partner during the description. If the convergence of available cues during the description strongly biases a particular perspective, then advance knowledge of the partner's viewpoint may not influence descriptions significantly. On the other hand, advance knowledge of the partner's viewpoint and its relation to the intrinsic structure may highlight alternative perspectives for both encoding and describing the array.

## Method

## Design

Directors first studied an array with an intrinsic structure, then had their memory of the array tested, and finally described the array to a partner, their Matcher, who
reconstructed the array on the basis of the Directors' descriptions. We manipulated the alignment of the array's intrinsic structure with either partner during the description phase, as well as the partners' advance knowledge of that. In a third of the pairs, Directors studied arrays while aligned with its intrinsic structure (referred to as $0^{\circ}$, see Figure 1), and later described it to Matchers who were offset by $135^{\circ}$ measured counterclockwise (Aligned with Director condition). In another third of the pairs, Directors studied arrays from $225^{\circ}$ and later described it to Matchers who were at $0^{\circ}$ (Aligned with Matcher condition). In the final third of the pairs, Directors studied arrays again from $225^{\circ}$ and later described to Matchers who were offset by $135^{\circ}$; thus both partners were misaligned with the structure (Aligned with Neither condition). Half of the Directors in each condition studied the array while knowing where their Matcher would later be, whereas the remaining half didn't.


Figure 1: The seven-object array used, indicating $0^{\circ}, 135^{\circ}$, and $225^{\circ}$ headings.

## Participants

Forty-eight undergraduate and graduate students from the University of Cyprus participated, half of them as Directors and half as Matchers, in 24 pairs. Six were female-female pairs, 6 were male-male pairs, 6 were mixed-gender pairs with female Directors, and 6 were mixed-gendered pairs with male Directors. All pairs of participants were recruited to be friends.

## Procedure

Study phase After a practice phase during which Directors were familiarized with the Judgments of Relative Direction (JRD) task (see below), Directors studied an array with an intrinsic axis of symmetry, comprising seven common objects that lacked intrinsic front-back and left-right axes, displayed on a 70 cm -diameter circular table (Figure 1).

Directors studied the array while either aligned or misaligned with its structure (from either $0^{\circ}$ or $225^{\circ}$ ), while either knowing where their Matcher would be during the
description phase or not. When the Matcher's viewpoint was known, Matchers sat at a separate, identical table next to the Director's (see Figure 2), at the position they would occupy during the description (at $0^{\circ}$ or at $135^{\circ}$ ).
Testing phase After ensuring that Directors memorized the array, Directors moved to an adjacent room to complete the memory tasks (JRDs and the drawing task). On JRD trials, Directors were instructed to imagine being at one location facing a second, constituting an imagined heading or viewpoint, and to point to a third object, the target (e.g., Imagine being at the bucket, facing the marble. Point to the candle.). Directors first read a statement in this form (i.e., "Imagine being at $x$, facing $y$ "), pressed a button on a joystick once they adopted that heading, and then responded to the second statement ("Point to $z "$ ) by deflecting the joystick in the direction of $z$ as if they were facing $y$ and pressing a button to $\log$ in their response. Sixty-four such trials were presented individually on a computer screen. They included eight imagined headings ( $0^{\circ}, 45^{\circ}, 90^{\circ}, 135^{\circ}$, $180^{\circ}, 225^{\circ}, 270^{\circ}, 315^{\circ}$ relative to the intrinsic structure) and their order was randomized.

After the JRDs, Directors did an array drawing task. They were given 20 cm -diameter grid (with 1 mm lines) and were asked to reconstruct as accurately as possible the studied array by placing seven circular transparent markers, each labeled with a name of the array's objects, on the grid.
Description phase After their testing phase, Directors returned to the original room for the description. Pairs sat at the positions prescribed by their condition of alignment with the array's intrinsic structure. Directors described the array's configuration from memory, while the Matcher used the seven objects to reconstruct the array at their table. Instructions emphasized that participants could interact freely and that they should reconstruct the array so that, given the Director's study viewpoint, objects be translated to the Matcher's table (i.e., not rotated by the Matcher's offset). Although pairs could interact freely, Directors could not look over the barrier ( 113 cm tall) separating the two tables; they could see each other's faces but not each other's tabletops. After turning on the cameras, the experimenter left the room for the description phase. After completing the description phase, pairs were debriefed and compensated for their time, if participating for payment.

## Coding of Spatial Descriptions

Each pair's interaction during the description phase was transcribed in detail, including contributions by both Directors and Matchers. We adapted our coding scheme from Galati et al. (2013) to classify spatial expressions in the Directors' turns as:
a. Director-centered, e.g., "in front of me is the marble"
b. Matcher-centered, e.g., "the vase is to your left"
c. Structure-centered, e.g., "it's on the perpendicular line. You're supposed to be on one side on the left, and I'm on one right side of the table"


Figure 2: Set-up of a study phase in which the Director was aligned with the array's intrinsic structure (at $0^{\circ}$ ), while the Matcher was misaligned with it (at $135^{\circ}$ ).
d. Neutral, capturing inter-object relations independently of a particular viewpoint, e.g., "it's close to the bucket" or "they form a triangle"
e. Other headings, not coinciding with the Director's, the Matcher's, or the structure's intrinsic orientation, e.g., "say the candle is facing the bucket; from the bucket, it's on the left"
f. Ambiguous, when expressions could be interpreted as involving more than one of coding categories
Three more categories (both-centered, environmentcentered, and object-centered) will not be considered further since they constituted less than $1.5 \%$ of all 1609 spatial expressions.
Reliability The first author coded 20 pairs, while a second coder redundantly coded 6 pairs as well as the remaining 4 pairs. Prior to comparing their judgments, the coders discussed 52 instances for which there was disagreement over the segmentation of spatial expressions (i.e., cases where one coder identified a spatial expression while the other didn't, or parsed a phrase as two spatial expressions while the other did as one). These disagreements were resolved by discussing them until consensus was reached; the remaining, non-redundantly coded dialogues were checked for consistent application of the agreed upon criteria. For the 383 spatial expressions from the redundantly coded dialogues, the two coders made identical classifications $98 \%$ of the time, Карра $=.98, p<.001$.

## Results

## Spatial Memory

Array drawings When Directors studied the array while aligned with the intrinsic structure (from $0^{\circ}$ ), all of them used the structure as the organizing direction of their drawings, whether they knew the Matcher's viewpoint $\left(135^{\circ}\right)$ or not. On the other hand, as Table 1 shows, when
they studied the array while misaligned with its structure (from $225^{\circ}$ ), the orientation of their drawings depended on whether they knew their Matcher's viewpoint. When the Matcher's viewpoint was unavailable, they were more likely to use their learning viewpoint $\left(225^{\circ}\right)$ to draw the array. But when they had known in advance that the Matcher was aligned with the array's structure, they used the structure's axis as an organizing direction more frequently. And when they had known in advance that the Matcher would also be misaligned with the structure (at $135^{\circ}$ ), half of the Directors opted for their learning viewpoint, while half used the axis of the structure as their organizing direction. The probability that the overall distribution of the drawings' orientation was observed by chance is small ( $p=.03$, Fisher's exact test).

Table 1: Proportion of Directors who drew arrays as aligned with the intrinsic structure vs. from own viewpoint, when having studied arrays from $225^{\circ}$.

|  | Aligned with <br> intrinsic <br> structure | Aligned with <br> learning <br> viewpoint |
| :--- | :---: | :---: |
| Knows Matcher <br> is at $0^{\circ}$ | .75 | .25 |
| Knows Matcher <br> at $135^{\circ}$ | .50 | .50 |
| Does not know <br> Matcher's <br> viewpoint | .25 | .75 |

Judgments of Relative Direction Analyses of JRD performance were initially conducted while ignoring the organization suggested by the Directors' drawings. However, these results were obfuscated by the fact that, as Table 1 illustrates, when misaligned with the structure, Directors were split in their preferred orientation at any given condition of availability of the Matcher's viewpoint. For instance, although most Directors preferred the structure's axes when knowing that the Matcher would be aligned with the structure, some still preferred their learning orientation. Thus, subsequent analyses of JRD performance centered on corroborating that Directors organized object relations in memory as indicated by their drawings' orientation.

As Figure 3 illustrates, the Directors' orientation latency (time to orient to an imagined heading) was consistent with the preferred orientation of their array drawings. Directors whose drawings were aligned with the structure were generally faster to orient to the structure's canonical axes $\left(0^{\circ}, 90^{\circ}, 180^{\circ}, 270^{\circ}\right)$ than to the oblique headings $\left(45^{\circ}\right.$, $\left.135^{\circ}, 225^{\circ}, 315^{\circ}\right)$. This pattern was reversed when Directors had drawn arrays from $225^{\circ}$. Indeed, the interaction between the heading from which the array was drawn and the JRD trial's imagined heading was significant, $F(7,154)=4.96, p$ $<.001$.

We examined this sawtooth pattern of performance by fitting planned contrasts with weights: $-1.625, .875,-0.625$, $1.375,-1.625,1.375,-0.625, .875$. This contrast, with the
minimums at $0^{\circ}$ and $180^{\circ}$, adequately described the orientation latencies of Directors who drew arrays aligned with the structure, $F(1,14)=10.34, p<.01$, accounting for $88 \%$ of the variance associated with the imagined heading and leaving a non-significant amount of variance unaccounted for ( $p=.98$ ). For Directors who drew arrays from their $225^{\circ}$ study viewpoint, the sawtooth contrast with the minimums at $225^{\circ}$ and its counteralinged heading (45 ) also described performance adequately, $F(1,8)=6.43, p$ $<.05$, accounting for $62 \%$ of the variance associated with the imagined heading and leaving a non-significant amount of variance unaccounted for ( $p=.82$ ).


Figure 3: Orientation latencies (in secs) across imagined headings according to how Directors had drawn arrays

The same pattern was observed for Directors' response latency (the time to point to the target after adopting an imagined heading) and their pointing error (the unsigned angular deviation of the joystick response from the veridical response). For brevity, these analyses are not reported here.

## Spatial Descriptions

Overall, Directors produced most frequently Neutral expressions in their descriptions $(48 \%$ of all spatial expressions), with Matcher-centered expressions constituting 20\%, Director-centered 15\%, Structurecentered $8 \%$, other headings $2 \%$, and ambiguous expressions $5 \%$ of all expressions.

We will focus on the distribution of Director-centered, Matcher-centered, and Structure-centered expressions, given our manipulation of the alignment of the intrinsic structure with either partner. The distribution of these three types of spatial expressions indeed depended on the partners' alignment of the intrinsic structure during the description, $F(4,36)=3.96, p<.01$. This interaction was driven by Directors using more Matcher-centered expressions than Director-centered ones when the Matchers were aligned with the structure ( $95 \% \mathrm{CI}[-.56,-.15], p<.01$ ), whereas the reverse was true when Directors were the ones aligned with the structure (though this difference was not statistically significant: $95 \% \mathrm{CI}[-.09, .32], p=.27$ ).

On its own, the availability of the Matcher's viewpoint didn't reliably affect the distribution of these spatial
expressions; this interaction was not significant, $F(2,36)=$ $1.83, p=.18$. Nonetheless, Directors used significantly more Matcher-centered expressions than Director-centered ones ( $26 \%$ vs. $7 \%$ ) when they knew the Matcher's viewpoint in advance ( $95 \%$ CI[-.36, -.02], $p<.05$ ), whereas they used comparable proportions ( $21 \%$ vs. $20 \%$ ) when they hadn't known their Matcher's viewpoint (95\% CI[-.18, .15], n.s.).


Figure 4: Proportion of Director-centered, Matchercentered, and Structure-centered expressions when the Matcher's viewpoint was available (a) or unavailable (b) at study, across the partners' alignment with the intrinsic structure.

As Figure 4 a shows, Directors who studied arrays from $0^{\circ}$ while knowing that their Matcher would be offset by $135^{\circ}$ used comparable proportions of egocentric and Matchercentered expressions in their descriptions ( $17 \%$ vs. $18 \%$; $95 \% \mathrm{CI}[-.31, .27]$, n.s.). On the other hand, as Figure 4b shows, they tended to use more egocentric expressions (34\% vs. $10 \%$ ) when their Matcher's viewpoint wasn't available at study ( $95 \% \mathrm{CI}[-.05, .54], p=.10$ ). When the Matcher was at $0^{\circ}$ during the description, Directors used predominately Matcher-centered expressions whether this information was available in advance or not. Finally, when neither partner was aligned with the structure during the description, the distribution of expressions differed depending on whether Directors knew this in advance. As Figure 4b shows, when

Directors hadn't known the Matcher's viewpoint in advance, they used numerically more egocentric than Matcher-centered expressions ( $27 \%$ vs. $10 \%$; $95 \%$ CI: [-.13, .46], $p=.25$ ), whereas as Figure 4a shows, when they had known it in advance, they used more Matcher-centered than egocentric expressions ( $29 \%$ vs. $1 \%$; $95 \%$ CI: [ $-.58, .01]$, $p=$ .06). Moreover, as suggested by the white bars across the two figures, Directors used numerically more Structurebased descriptions when they knew in advance that Matchers would also be misaligned with the structure than when they didn't (95\% CI: [-.33, .02], $p=.08$ ).

The distribution of spatial expressions that Directors used was not influenced by the pair's gender combination or the gender of the Director; the interaction of each of these factors with the type of spatial expression was not significant: $F(6,40)=.87, p=.52$, for the pair's gender combination, and $F(6,40)=.85, \mathrm{p}=.44$ for the Director's gender.

## Discussion

Our findings suggest that people consider both representational and communicative factors when organizing spatial information in memory and when selecting the perspective from which to describe that information. The preferred direction around which people organize spatial relations in memory depends on whose viewpoint is reinforced by the configuration's intrinsic orientation. This was demonstrated by the Directors' drawings and was corroborated by their performance in the JRD task. When Directors were aligned with the intrinsic structure, they defaulted to their own viewpoint as the organizing direction, regardless of what they knew about their partner's viewpoint. On the other hand, when they were misaligned with the structure during learning, knowing that their Matchers would be aligned with the structure's orientation increased the probability of using the structure's axes as an organizing direction. Moreover, knowing that the Matcher would also be misaligned with the structure increased the probability of using the structure's axes as an organizing direction compared to not knowing the Matcher's viewpoint.

These findings suggest that a given partner's alignment with the intrinsic structure affords sufficient pragmatic motivation to organize spatial relations from that orientation: when the structure's orientation is aligned with a partner's viewpoint, these converging cues influence the preferred orientation that people use. This extends our earlier findings that, when no intrinsic structure is available, people encode the partner's viewpoint in memory but don't necessarily use it as an organizing direction, likely due to insufficient pragmatic motivation to do so when they can freely interact with their partner (Galati et al, 2013).

We propose that, when selecting the preferred orientation of their spatial memories, people combine probabilistically different sources of information. When the intrinsic structure and their own learning viewpoint converge, they
use that egocentric viewpoint; when the intrinsic structure and their partner's viewpoint converge, they opt for the partner's viewpoint. This also held for how speakers adapted their spatial descriptions. When the intrinsic structure and the Director's learning viewpoint converged, Directors tended to describe spatial information from their own perspective, with Matchers having to unpack the spatial mappings of these Director-centered descriptions. When the intrinsic structure and Matcher's viewpoint converged, Directors alleviated the Matcher's cognitive burden by describing spatial information from the Matcher's viewpoint. Speakers used the available social and representational cues to adapt their descriptions in ways that minimized their collective effort (e.g., Clark \& WilkesGibbs, 1986), with the assumption here being that a perspective supported by converging cues is optimally effective.

Moreover, speakers flexibly used information that was perceptually available in the communicative setting and didn't merely rely on the organization of the memories. For instance, when Directors who studied the array from $225^{\circ}$ without knowing their Matchers viewpoint later interacted with a Matcher at $0^{\circ}$, they used overwhelmingly Matchercentered descriptions, even though most of them had used their own viewpoint as an organizing direction in memory. This is consistent with findings that, in describing spatial information, people do not always adhere to their memory's organizing direction when it conflicts with perceptual evidence (Li et al, 2011). Altogether, the adaptation we report here underscores that people use all relevant information as soon as it becomes available (whether at study or at collaboration) to make attributions about their respective ability to contribute to the task. This is in line with the view that probabilistic constraints on information processing influence perspective-taking behavior in conversation (e.g., Hanna, Tanenhaus, \& Trueswell, 2003).

Our study offers a caveat on earlier demonstrations that the misalignment between partners influences perspectivetaking (e.g., Duran, Dale, \& Kreuz, 2011; Schober, 1993), by highlighting that misalignment interacts with other representational cues. When Directors were at $0^{\circ}$ and Matchers at $135^{\circ}$, Directors overall opted for their own perspective in descriptions, presumably because reasoning from an oblique perspective was computationally more difficult (especially when not made salient at study). However, when Matchers were at $0^{\circ}$ and Directors at $225^{\circ}$ (also a $135^{\circ}$ offset), Directors readily adopted their partner's perspective in descriptions. Thus, misalignment on its own does not determine the preferred perspective of speakers' descriptions.

In sum, in collaborative spatial tasks people adapt their memory representations and linguistic behavior in nuanced ways. They consider converging communicative and representational cues, whenever they become available, to appraise whose perspective would be optimal for efficient coordination; this influences whether they encode their
partner's available viewpoint in memory and adopt it in descriptions.

## Acknowledgments

This material is based upon work supported by the European Research Council under grant 206912-OSSMA to M.A. We are grateful to Chrystalleni Nicolaou and Margarita Antoniou for assistance with data collection and coding.

## References

Clark, H. H., \& Wilkes-Gibbs, D. (1986). Referring as a collaborative process. Cognition, 22, 1-39.
Duran, N. D., Dale, R., \& Kreuz, R. J. (2011). Listeners invest in an assumed other's perspective despite cognitive cost. Cognition, 121, 22-40.
Galati, A., Michael, C., Mello, C., Greenauer, N. M., \& Avraamides, M. N. (2013). The conversational partner's perspective affects spatial memory and descriptions. Journal of Memory and Language, 68, 140-159.
Hanna, J. E., Tanenhaus, M. K., \& Trueswell, J. C. (2003). The effects of common ground and perspective on domains of referential interpretation. Journal of Memory and Language, 49, 43-61.
Hölscher, C., Tenbrink, T., \& Wiener, J. M. (2011). Would you follow your own route description? Cognitive strategies in urban route planning. Cognition, 121, 228247.

Li, X., Carlson, L. A., Mou, W., Williams, M. R., \& Miller, J. E. (2011). Describing spatial locations from perception and memory: The influence of intrinsic axes on reference object selection, Journal of Memory and Language, 65, 222-236.
Mou, W., \& McNamara, T. P. (2002). Intrinsic frames of reference in spatial memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 28, 162170.

Mou, W., McNamara, T.P., Valiquette, C. M., \& Rump, B. (2004). Allocentric and egocentric updating of spatial memories. Journal of Experimental Psychology: Learning, Memory, and Cognition, 30, 142-157.
Schober, M.F. (1993). Spatial perspective-taking in conversation. Cognition, 47, 1-24.
Shelton, A.L., \& McNamara, T.P. (2001). Visual memories from nonvisual experiences. Psychological Science, 12, 343-347.
Shelton, A. L., \& McNamara, T. P. (2004). Spatial memory and perspective taking. Memory and Cognition, 32, 416426.

Tenbrink, T., Coventry, K. R., Andonova, E. (2011). Spatial strategies in the description of complex configurations. Discourse Processes, 48, 237-266.

# False Consensus About False Consensus 

Mirta Galesic (galesic@mpib-berlin.mpg.de)<br>Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development Lentzeallee 94, 14195 Berlin, Germany<br>Henrik Olsson (olsson@mpib-berlin.mpg.de)<br>Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development Lentzeallee 94, 14195 Berlin, Germany<br>Jörg Rieskamp (joerg.rieskamp@unibas.ch)<br>Department of Psychology, University of Basel<br>Missionsstrasse 62a, 4055 Basel, Switzerland


#### Abstract

Research on human reasoning is dominated by demonstrations of the errors people make in various judgment and decision-making tasks. The area of social cognition is not an exception: the list of apparent errors is long and includes a number of contradictory phenomena. Here we explore a prominent example of the contradictory pairs of biases: false consensus and false uniqueness. We show in an empirical study and with simulations that the consensus in the literature about the stability of these effects may be premature, as their occurrence depends on the format of questions used to measure them.


Keywords: False consensus; false uniqueness; social circle; response formats.

## Introduction

The false consensus effect (Ross, Greene, \& House 1977) or "looking glass perception" (Fields \& Schuman 1976) describes a phenomenon that people who exhibit a certain behavior or endorse a particular view ("performers") believe that this behavior or view is more common overall than do people with different behaviors or views ("nonperformers"). For example, Democrats would judge that democratic views are more spread in the general public than Republicans would. This kind of result has been documented so often that the false consensus bias has been considered an automatic response that may be "developmental vestiges of the infantile belief that all others are like us" (Krueger \& Clement, 1994, p.609). However, an opposite bias called false uniqueness has also been documented (Frable, 1993; Mullen, Dovidio, et al., 1992). People holding a particular view sometimes tend to think that their view is less popular than do people holding a different view.

At least five different explanations have been proposed to explain false consensus effects (Marks \& Miller, 1987). First, people are likely to have selective exposure to similar others, so their estimates of larger social environments are based on biased samples. Second, their preferred view may be more salient to them than a different view, which may make them think that their preferred view has a stronger
social support. Third, people may believe that situational factors that led them to hold a particular view will affect others in a similar way, leading them to adopt the same view as well. Note that this view contrasts with another popular bias, namely the fundamental attribution error, whereby people believe that their behavior is caused by situation but others' behavior is caused by dispositional factors. Fourth, believing that others share one's view may have a motivational cause, such as fulfilling the need to validate own belief and maintain self-esteem. Fifth, false consensus is in line with a Bayesian analysis that assumes a uniform prior distribution and one's own view as the only evidence (Dawes \& Mulford, 1996).

It is more difficult to explain false uniqueness. Suls and Wan (1987) extend the motivational account and propose that false uniqueness can contribute to one's self esteem when the behavior or view in question is desirable, but find inconsistent support for this view (Suls, Wan, \& Sanders, 1988). Moore and Kim (2003) show that because people rely more on information about themselves than about others when forming judgment of prevalence of their views, effects similar to both false consensus and false uniqueness can occur. However, their measure of these effects is different than that used in most other studies: they use the difference between people's judgments and true population values rather than the difference between judgments of groups of people holding different views.

Here we investigate a so far neglected possible factor that may lead to both effects: the format of the questions used to measure these effects. Most studies investigating false consensus use one of two response formats. Either they ask about both the percentages of performers and nonperformers, for example, "What \% of your peers do you estimate would carry the sandwich board around campus? \% What \% would refuse to carry it? \% (Total should be $100 \%$ )" (Ross, Greene, \& House, 1977), or they ask only about the percentages of performers, e.g., "What percentage of students do you think agreed to wear the sign?" (Krueger \& Clement, 1994). There are no studies, however, that compare how different response formats affect estimates of the false consensus effect. For example,
it is not known whether the effect would remain the same if participants were asked about nonperformers rather than performers. It is well known from survey methodology literature that response formats can have strong effects on answers independently of people's true beliefs (Tourangeau, Rips, \& Rasinski, 2000). Similarly, research on subjective probability calibration shows that people can appear overconfident, well calibrated, or underconfident depending on the response format used (Juslin, Wennerholm, \& Olsson, 1999). This motivates us to explore these effects in the case of false consensus and false uniqueness effects.

## Method

We asked 104 participants recruited from Mechanical Turk ( $43 \%$ female, mean age $34,44 \%$ with bachelor or higher degree) three groups of questions about 10 characteristics, listed in Table 1. The questions were taken from publicly available results of large national surveys (Gallup World Poll 2011 for characteristics 1-5, Pew Center 2011 for 610); full texts are available on request. In the present study, participants first gave their personal answer to each of the 10 questions. In this way we classified them as either performers or nonperformers on a particular characteristic. Thereafter they estimated the percentage of performers and/or nonperformers in their social circle (defined as adults you were in personal, face-to-face contact with at least twice this year), and in the general population of the United States. One random half of the participants answered the questions about their social circle first, and another half about the population.

For each characteristic, a random third of performers and a random third of nonperformers gave estimates of social circle and population percentages in one of the following response formats: 1) estimating only the percentage of performers, 2) estimating only the percentage of nonperformers, and 3) estimating both percentage of performers and nonperformers. Figure 1 provides an example of the three response formats for one of the characteristics, and Table 2 lists all formats. Note that in format 3 the estimates for performers and nonperformers have to sum to 100 , but there is no such check in formats 1 and 2. Estimates for social circle and for the population were given always in the same format. The same individual could have answered questions for different characteristics in different formats, depending on whether he was a performer or nonperformer himself, and to which response format group he was randomized to.

Table 1: Characteristics used in the study, along with percentage of people answering "yes" (performers) in national surveys, and percentage of such people in the present sample.

|  | Characteristic | Population \% of performers | Sample \% of performers |
| :---: | :---: | :---: | :---: |
| 1 | No money for food in past 12 months | 19 | 18 |
| 2 | Donated to charity last month | 57 | 41 |
| 3 | Experienced theft in past 12 months | 12 | 21 |
| 4 | Religion is important part of daily life | 64 | 28 |
| 5 | Attended worship in past 7 days | 47 | 14 |
| 6 | Belief in God necessary to be moral | 53 | 13 |
| 7 | Believes in God | 70 | 54 |
| 8 | Smokes tobacco at least once/day | 15 | 24 |
| 9 | Military force sometimes necessary | 77 | 84 |
| 10 | Homosexuality should not be accepted | 36 | 18 |



Figure 1. Example of the three response formats used to elicit estimates of performers and nonperformers in general population.

Table 2: Different ways in which prevalence of performers can be inferred, depending on the response format.

| Response <br> format | Estimates about <br> prevalence of | Estimates given <br> by | Abbre- <br> viation |
| :--- | :--- | :--- | :--- |
| 1 | Performers only | Performers | P.P |
| 2 | Nonperformers only | Performers | NP.P |
| 3 | Performers and | Performers | Pnp.P |
|  | Nonperformers |  | NPp.P |
| 1 | Performers only | Nonperformers | P.NP |
| 2 | Nonperformers only | Nonperformers | NP.NP |
| 3 | Performers and | Nonperformers | Pnp.NP |
|  | Nonperformers |  | NPp.NP |

False consensus and false uniqueness can be measured in different ways. The most prevalent approach in the literature is to calculate the difference between the prevalence of performers as estimated by performers (P.P) and the prevalence of performers as estimated by nonperformers (P.NP). A positive difference P.P - P.NP is interpreted as false consensus, and a negative difference as false uniqueness. In our study, separate groups of both performers and nonperformers gave estimates in 3 different formats. This enables calculating the size of false consensus in 9 different ways, listed in Table 3.

Table 3: Different ways in which false consensus effects can be calculated, depending on the response format.

| Type of false <br> consensus | Calculation |
| :--- | :--- |
| 11 | P.P - P.NP |
| 12 | P.P - (100 - NP.NP) |
| 13 | P.P - Pnp.NP |
| 21 | $(100-$ NP.P $)-$ P.NP |
| 22 | $(100-$ NP.P $-(100-$ NP.NP $)$ |
| 23 | (100 - NP.P) - Pnp.NP |
| 31 | Pnp.P - P.NP |
| 32 | Pnp.P - (100 - NP.NP $)$ |
| 33 | Pnp.P - Pnp.NP |

## Results

How stable are false consensus effects across different response formats? If response format does not play a role, estimates of prevalence of performers should be the same for all formats, consequently resulting in same direction and size of false consensus effects. However, Figure 2 shows that the effects vary depending on response formats used to estimate prevalence of performers. The most extreme example is characteristic number 1 (no money for food), where estimates exhibit false uniqueness when performers estimate prevalence of performers (types 11-13), but false consensus when performers estimate prevalence of nonperformers (types 21-23) or when they estimate prevalence of both performers and nonperformers (types 3133). Several other characteristics show similar patterns of both false consensus and false uniqueness effects.


Figure 2: False consensus effects for nine different ways of calculating false consensus (see Table 3 for details). Small numbers denote effects for different characteristics. Full line denotes mean of the effects. Dotted lines denotes difference between performers and nonperformers in participants social circles (see text).

How can these different false consensus and false uniqueness effects for the same characteristics be explained? We propose a simple model of how estimates of prevalence of performers in the population are derived. The model has two plausible assumptions. First, people derive estimates about the general population based on the samples they have in their immediate social environment, that is their social circles (see Galesic, Olsson, \& Rieskamp, 2012, for a social circle model that accounts for people's estimates of population distributions). Support for this assumption is shown in Figure 2, where dotted line represents differences
in percentages of performers and nonperformers, calculated by different methods, in participants' social circles. They parallel the population estimates ( $\mathrm{r}=.89$ ).

Second, we assume that to derive these estimates, people attempt to recall as many individuals in their social circle belonging to the required category (e.g. performers) as they can. Because of time and effort limits, they are often not able to recall all such individuals. Consequently they may underestimate the percentage of those individuals in the population relative to what they would report had they recalled all such individuals in their social circle.

The model can be formalized for each response format separately, as follows. Recall that to estimate false consensus effects, a researcher needs estimates of prevalence of performers. When a person is asked only about performers in the general population (response format 1 , see Table 2), his estimate of population prevalence of performers can be modeled as:

$$
P_{1}=S C_{P} \times \alpha,
$$

where $P_{1}$ is the performers' population prevalence estimated in response format $1, S C_{P}$ is the percentage of that person's social circle that are performers, and $\alpha$ is a memory activation level parameter ranging from 0 to 1 . Note that according to this model people are assumed to always estimate population prevalence of performers as lower or equal than in their social circle. When a person is asked only about nonperformers in the general population (response format 2), his estimate of population prevalence of performers can be inferred from his estimate of population prevalence of nonperformers. This can be modeled as:

$$
P_{2}=100-S C_{N P} \times \alpha,
$$

where $S C_{N P}$ is the percentage of that person's social circle that are nonperformers, and the meanings of other symbols are the same as above. Note that if prevalence of nonperformers is underestimated $(\alpha<1)$, then the inferred prevalence of performers in this response format will be overestimated relative to the true percentage in the social circle.

Finally, when a person is asked to estimate the percentage of both performers and nonperformers (response format 3), his estimate of population prevalence of performers can be modeled as:

$$
P_{3}=100 \times\left(\frac{S C_{P} \times \alpha}{S C_{P} \times \alpha+S C_{N P} \times \alpha}\right)=S C_{P}
$$

where meanings of symbols are the same as above. Because in this response format percentages of performers and nonperformers have to sum to 100 , the denominator serves to normalize the sum of prevalence estimates of performers and nonperformers, which would be lower than 100 if $\alpha<1$. That is, it is assumed that people recall a subset of performers and nonperformers from all performers and nonperformers in memory, and then estimate the percentage
of each group in the sum of both groups. If their recall of performers and nonperformers is unaffected by other factors (see Discussion for more comments on this possibility), then their population estimate of performers equals the percentage of performers in their social circle $\left(S C_{P}\right)$.

Note that for simplicity we do not model the fact that reports of social circles are similarly affected by response formats as the population estimates. However we believe that modeling this would only make estimated parameters larger, but would not change relative differences between estimates in different formats.

To check whether this simple model could reproduce the pattern of results in Figure 2, we simulated estimates of prevalence of performers for 10 different fictitious characteristics with social circle prevalence ranging from $1 \%$ to $91 \%$ in steps of 10 percentage points. We modeled population estimates using the formulas above and different values of $\alpha$. For all values of $\alpha$ lower than 1 the pattern of false consensus effects is very similar to the empirical results in Figure 2. Figure 3 shows an example for $\alpha=8$.


Figure 3: Simulated patterns of false consensus and false uniqueness effects, for $\alpha=.8$. Full line denotes results assuming the same social circles for performers and nonperformers. Dotted line denotes results assuming that performers know relatively more performers than do nonperformers.

As visible in Figure 3, the pattern of false consensus and false uniqueness effects in this fictitious data set is very similar to the empirical pattern shown in Figure 2. The full line represents false consensus estimates assuming that performers and nonperformers have the same percent of performers in their social circles. However, in reality each group typically knows more individuals similar to themselves. Therefore we observe stronger false uniqueness effects in the simulation than in the empirical data. However, if we assume a small difference in social circles so that nonperformers have 20 percentage points fewer performers in their social circles than do performers, all
effects shift towards stronger false consensus. This is shown as a dotted line in Figure 3.

## Discussion

The pattern of false consensus and false uniqueness effects seems to be a product, to a large extent, of the way questions are asked and the samples people take from their social environments.

If both performers and nonperformers are asked about their own groups (false consensus type 12), then false uniqueness effects are likely to occur. This is so because imperfect recall of nonperformers about the members of their own group leads to inflated estimates of the prevalence of performers. More generally, when nonperformers are asked about their own group rather than about performers (types 12, 22, and 32), we see a reduction of false consensus that in some cases turns into false uniqueness.

In contrast, when nonperformers are asked about performers (types 11, 21, and 31), the imperfect recall alone will lead to underestimation of performers' prevalence. If there is no difference between social circles of performers and nonperformers, then the false consensus effect for type 11 will be zero (see the point 11 of the full line in Figure 3). If there is a difference, then the false consensus effect will occur, as in the point 11 of the dotted line in Figure 3 and in empirical results in Figure 2). In all other conditions where nonperformers answer about performers, false consensus effects are most likely.

The two response formats that are most often used in the literature, where performers and nonperformers answer about performers only (type 11) or about both performers and nonperformers (type 33), produce very similar false consensus effects. This may contribute to the wide-spread consensus in the literature about the robustness of the false consensus effect. However as our findings show, false uniqueness and false consensus can occur for the same characteristics, depending on how the question is asked. Therefore false consensus may not be such a robust bias as previously assumed.

Note that the simple model described here neglects effects of frequency of contact on recall, and does not specify how the percentage estimates are formed in the first place. This model cannot explain the empirical fact that population estimates often resemble smoothed versions of one's social circle (Galesic et al, 2012), that is performers report smaller proportion of performers and larger proportion of nonperformers in population than in their social circle. A more elaborate model would describe how people sample from their social circle, for example based on frequency of contact, and how they estimate the percentage of performers based on that sample. The final estimates are likely to be shaped by both, effects of question format, and other sampling and estimation processes.

A common approach to explaining apparent errors in social judgments is to look at the human mind and search for motivational and cognitive processes that deviate from
normative rules of reasoning. Here we show that properties of social environments as well as of memory processes, and their interplay with the way questions are asked, can produce apparent false consensus and false uniqueness effects on its own. This work does not diminish the potential importance of other explanations of origins of these effects, but provides a baseline for the part of these effects that are artifacts rather than a cognitive bias.

## Acknowledgments

We thank Max Planck Institute for Human Development for financial support for this study.

## References

Dawes, R. M., \& Mulford, M. (1996). The false consensus effect and overconfidence: Flaws in judgment or flaws in how we study judgment? Organizational Behavior and Human Decision Processes, 65, 201-211.
Frable, D. E. S. (1993). Being and feeling unique: Statistical deviance and psychological marginality. Journal of Personality, 61, 85-110.
Galesic, M., Olsson, H., \& Rieskamp, J. (2012). Social sampling explains apparent biases in judgments of social environments. Psychological Science, 23, 1515-1523.
Juslin, P., Wennerholm, P., \& Olsson, H. (1999). Format dependence in subjective probability calibration. Journal of Experimental Psychology: Learning, Memory, and Cognition, 25, 1038-1052.
Krueger, J., \& Clement, R. W. (1994). The truly false consensus effect: An ineradicable and egocentric bias in social perception. Journal of Personality and Social Psychology, 67, 596-610.
Moore, D. A., \& Kim, T. G. (2003). Myopic social prediction and the solo comparison effect. Journal of Personality and Social Psychology, 85, 1121-1135.
Mullen, B., Dovidio, J. F., Johnson, C., \& Copper, C. (1992). In-group-out-group differences in social projection. Journal of Experimental Social Psychology, 28, 422-440.
Ross, L., Greene, D., \& House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. Journal of Experimental Social Psychology, 13, 279-301.
Suls, J., \& Wan, C. K. (1987). In search of the falseuniqueness phenomenon: Fear and estimates of social consensus. Journal of Personality and Social Psychology, 52, 211-217.
Suls, J., Wan, C. K., \& Sanders, G. S. (1988). False consensus and false uniqueness in estimating the prevalence of health-protective behaviors. Journal of Applied Social Psychology, 18, 66-79.
Tourangeau, R., Rips, L. J., \& Rasinski, K. (2000). The psychology of survey response. New York: Cambridge University Press.

# The Role of Feedback in Aligning Perspectives in Referential Communication 

Timothy M. Gann (tgann@ucmerced.edu)<br>Sierra Nevada Research Institute and Cognitive and Information Sciences, University of California, Merced 5200 North Lake Rd. Merced, CA 95343 USA


#### Abstract

Successful dialogue frequently requires that interlocutors construct and align their conceptualizations of referents. This study presents data from a referential communication experiment the manipulates contextual factors such as the availability of feedback and role constancy in order to investigate how conversational partners reconcile their perspectives in the face of mutual uncertainty about what constitutes common ground. The results show that speakers tend to incorporate information about the addressee's perspective, and that this information tends to come through direct feedback rather than through indirect channels such as turn-taking.


Keywords: Psycholinguistics; language production; audience design; perspective taking.

## Introduction

When deciding how to describe an object, people must first categorize it in a way that is useful to themselves them and ultimately for their partners. Sometimes, the taxonomic, established categorization is not well suited to the task and interlocutors may be forced to develop an ad hoc, or socially emergent category that is more suited to achieving their current conversational goals (Barr \& Kronmüller, 2007; Barsalou, 1991). For example, when two people are discussing where they wish to go out to eat, they may form a category of acceptable restaurants on the spot. This category emerges over the conversation and is neither pre-existing in memory nor particularly enduring (unless they frequently go out together). This socially emergent conceptualization then becomes the foundation for future exchanges. What this means is that early on in a conversation, taxonomic categories are going to be activated more often and attract more attention. However, as time goes by and speakers learn to put more emphasis on shared information, and other types of information are found to be in conflict with the goal of successfully referring to an object, the socially constructed categories should be used more often. These socially emergent categories would probably demonstrate some degree of being person-specific (Horton \& Gerrig, 2005), but that would not preclude their activation and propagation to new conversational partners if they were useful (Garrod \& Doherty, 1994).

The development of new conceptualizations during conversation should also lead to changes in how attention is allocated. Rehder and Hoffman (2005) conducted an experiment in which participants had to do a category learning task while having their eyes tracked. At the beginning of the experiment, participants would overtly attend to each of the features in the display and consider them all before making a decision. By the end of the experiment, participants learned to overtly attend only to those features that were necessary for the task of choosing which category was represented by the display.

Similarly, in a referential communication context involving perspective taking, as time passes and socially emergent categories become more dominant, attention should shift away from privileged competitors. By the end of a conversation they should be attended to about as little as an object unrelated to the target. This shift should also happen in situations in which the speaker does not initially have access to what the listener knows. As evidence about listener's knowledge, or lack of knowledge, becomes available they should gradually begin to focus on the stimuli that are most consistent with the listener's perspective.

The question then becomes, what factors influence the rate of learning during conversation? One possibility is the potential for feedback. Closely related to feedback are the roles of the interlocutors: Are they in a didactic situation in which the roles are fixed, or is it a fully interactive dialogue in which there is turn taking? Unrestricted feedback can come in many forms. First, speakers can receive pragmatic feedback in the form of knowledge about the success or failure of the listener in establishing joint attention to a target object. It is implicitly accepted that pragmatic feedback is necessary for learning to happen, with positive feedback leading to a reinforcement of the speaker's current conceptualization, and negative feedback encouraging re-conceptualization. Second, the listener can give signals to the speaker, who can make inferences about the listeners confidence about an interpretation. For example, Barr (2003) found that listeners will use speakers' confidence as a cue that the current referent is either a typical or atypical member of a category. In principle, speakers should also similarly use the listener's confidence in their selection of a target as a cue to how well their message was received.

Finally, listeners can give direct verbal feedback to the speaker. Schober and Clark (1989) conducted an experiment in which a participant acted as a non-interactive observer to a conversation. This participant attempted to do the same task the actual listener was doing, but without the ability to give feedback to the speaker.They found that the observers never aligned as closely to the speaker because they were unable to provide feedback and have the speaker adjust to or correct their interpretations. The lack of verbal feedback can also make the speakers prone to give more information as a hedge against possible misunderstandings (Gann \& Barr, 2012). For example, Krauss and Weinheimer (1966) found that when listeners were unable to give feedback, speakers were not inclined to simplify their utterances.Instead they kept utterances long so they could minimize the possibility of giving too little information. While it is possible for alignment to happen without it, direct verbal feedback likely serves to speed up the
process. Without it, speakers may not be able to easily settle on a shared perspective and fully acquire a socially emergent category that corresponds to the intersection of their knowledge with their partner's.

Role switching may serve a similar function. In a natural conversation, both participants are speakers and listeners. These participants both offer their respective conceptualizations of the referents they are bringing to each others attention, and both are able to interactively sculpt each others understanding. Even simply priming from ones partners utterances may serve to make dimensions more salient than the other might be from an individuals egocentric perspective (Pickering \& Garrod, 2004). In addition, in situations where one person is the primary speaker, there might be less incentive for that individual to align, and thus, the speaker might rely on the listener to align to align to speaker knowledge.

In many referential communication experiments that manipulate common ground, knowledge is asymmetric in one direction. Specifically, the speaker knows more than the listener, or the listener knows more than the speaker. This experiment focuses on a case in which the speaker and listener both must learn to account for this difference in knowledge. The major question is whether or not conversational partners in such a situation tend to reconceptualize the stimuli in a way that is consistent with a union of their perspectives, thus taking into account both their sets of privileged information, or an intersection of their perspectives, focusing only on that information that is mutually shared. If in a particular circumstance, speakers tend towards reconceptualizing the referent in terms of an intersection between their knowledge and their partner's, will speakers then begin to ignore competitors that are not relevant to that overlap? Additionally, is their tendency to adopt one scheme or another influenced by the form of feedback they experience: either explicit feedback, or implicit feedback through role switching? Direct feedback should allow addressees to clearly indicate to the speaker what their informational needs are, and should thus be associated with rapid alignment between the interlocutors. Learning about a partner's perspective in the role switching condition presumably relies on more indirect learning, and may not have the same impact as direct feedback.

## Method

## Participants

The participants were all native English speakers, drawn from a pool of undergraduates at the University of California, Riverside. A total of 32 students participated in the study, forming 16 dyads. For half of the dyads, the participants were assigned to either the role of the director or the matcher. The other half of the dyads did not have fixed roles, but rather switched roles after each trial of the experiment.

## Apparatus

An ISCAN ETL-400 table-mounted eye-tracker, sampling at a rate of 60 Hz , was used to track the director's gaze through-
out the experiment. The experimental stimuli were displayed for the participants on two LCD monitors. Each participant's screen was only visible to themselves. The director was given a set of headphones through which instructions could be given without the knowledge of the matcher.

## Materials

The experimental stimuli consisted of 32 sets of five items. Each stimulus item was a $300 \times 300$ pixel colored bitmap of an object placed on a black background. The sets were constructed such that there was a target object, three competitor objects that are typically referred to by the same name, and an unrelated filler item. The competitor objects contrasted with the target object along the dimensions of size, openness, and material (or color). The target was always consistently open and larger; each of those two dimensions were visible to only one of the participants. Material differed from object to object; this dimension was always shared between the participants. So for example, one participant may have seen a target that was a large open trashcan, a smaller identical competitor, a plastic trashcan, and an unrelated fourth object. The other participant would see the same target, the plastic trashcan, the same unrelated competitor, but would have a closed competitor that is otherwise as physically similar to the target as possible. See Figure 1 for an example.

## Procedure

Participants were assigned to either the role of director or matcher by having one of the participants choose a face down card at random that specified a role. If the dyad was participating in the condition in which their roles would switch throughout the experiment, it was explained to them that they were picking an initial role. For the switching participants, the participant who was eye-tracked was always the initial director.

The experiment was described to the participants as a simple communication game that they would be playing together in which the director would be describing a target object such that the matcher could pick out which object on their screen. They were shown a set of example stimuli demonstrating the perspective of both partners. The sample stimuli did not include competitors that varied along the privileged dimensions, only one that contrasted along the shared dimension. The participants were instructed that the target would be indicated to the current director with a red border. The matcher would listen to the director's description and select the object being referring to on their screen by selecting it with their mouse.

## Design

The experimental variables manipulated in the design of the experiment included whether the partners switched roles or not, and level of feedback ( $2 \times 2$ between-subjects design). The dependent variables are accuracy (whether the addressee chooses the right target), speech onset time, use of an adjective corresponding to the speaker's privileged perspec-


Figure 1: Each of the displays above corresponds to the perspective of one participant. In this example, the target would be the closed metal trashcan in the upper left corners. The display on the left contains a competitor based on openness. The display on the right shows a competitor based primarily on material/color. Each display has a competitor in common based on size.
tive/dimension, use of an adjective corresponding to the addressee's privileged perspective, use of an adjective corresponding to the shared perspective, and the proportion of time the speaker spent gazing at the shared and privileged competitors prior to speech onset.

## Analysis

The director's speech was recorded, transcribed, and coded for accuracy, speech onset, and the use of modifiers matching each of the three dimensions on which the competitors contrasted with the target. The eye-tracking data were coded as number of frames (sampled at 60 hz ) that the director spent looking at each object prior to speech onset. Eye-tracking data for trials in which the participants switched roles was removed from the data set due to the inability to make an apples to apples comparison across the condition due to the current director only contributing eye-data for half the trials in the switching condition.

The analyses were conducted using linear mixed effects models with random effects included for subjects and items. Models were fit within R using the lmer function within the package lme4 version 0.999375-39 (Bates \& Maechler, 2010). Appropriate link functions were chosen depending on the distribution of the outcome variable of interest. The significance of the fixed effects were assessed using a $\chi^{2}$ model comparisons approach in which a model without the variable of interest is compared with the full model (Barr, Levy, Scheepers, \& Tily, 2013).

## Results

Thirteen of 512 ( $2.5 \%$ ) trials were dropped from the analyses because of experimenter and procedural errors. Across all conditions the listener correctly identified the referent in $72.55 \%$ of trials (Figure 2 shows accuracy plotted by condition). There were significant effects of feedback $\left(\chi_{(1)}^{2}=\right.$ $30.431, p<0.01)$ and trial order $\left(\chi_{(1)}^{2}=11.709, p<0.01\right)$, with the presence of feedback and increasing trial order being
associated with greater accuracy. However, a three-way interaction of these factors with role switching $\left(\chi_{(1)}^{2}=5.214, p<\right.$ 0.05 ) showed that while the feedback effect was stable in both switch conditions, order only had an effect when there was no role switching and it interacted positively with feedback when there was role switching $\left(\chi_{(1)}^{2}=6.034, p<0.05\right)$.

## Speech Analysis

Potentially, speakers could refer to three dimensions in their descriptions of the referent corresponding to the contrasts between the referent and the shared competitor, the speaker's privileged competitor, and the listener's privileged competitor. At the beginning of the experiment speakers are much more likely to use the first two by virtue of their lack of access to the third, but as the experiment progresses and evidence for that dimension becomes available it should be seen to rise. The use of adjectives associated with these dimensions speaks to the underlying scheme the speaker has about what information is necessary to identify the target. Because the common competitor is always necessary (and indeed, is referred to in $88 \%$ of trials) its use is a useful check when it comes to assessing whether a speaker understands and is faithfully attempting to do the task. However, because the underlying hypotheses are focused on how speakers reconcile their privileged knowledge with that of their listeners, how they used modifiers related to the two privileged contrasts is more theoretically interesting.

Use of a modifier relating to the speaker's privileged competitor was influenced by whether or not the participants were switching roles $\left(\chi_{(1)}^{2}=6.500, p<0.05\right.$; see Figure 3). For the non-switchers, use of the speaker's privileged competitor stayed relatively close to ceiling over the course of the experiment whereas its use dropped over time for partners who switched roles. This may be because, for a speaker in the no-switch condition, their privileged competitor was always salient to them, whereas in the switching condition speakers experienced trials in which its presence was not linked to pro-


Trial
Trial

Figure 2: Proportion of targets correctly selected by the addressee.
duction. Because the use of the privileged competitor was not strongly associated with accuracy ( $\bar{r}_{\phi}=-0.028, \mathrm{p}=0.595$ ), there was not necessarily a strong reason to reduce its salience to speakers in the non-switching condition in the absence of another role.

On the other hand, the speaker had more substantial motivation to learn and use the listener's privileged competitor because its use was strongly associated with success in the task ( $\bar{r}_{\phi}=0.310, p<0.01$ ). In their use of modifiers that matched the listener's privileged competitor, speakers were significantly influenced by trial order (using more over time; $\left.\chi_{(1)}^{2}=26.866, p<0.01\right)$ and an interaction between the feedback and switching conditions ( $\chi_{(1)}^{2}=6.899, p<0.01$ ), such that the effect of feedback was larger when there was no role switching. It is interesting that there doesn't appear to be an additive effect between the two routes to getting partner feedback associated with the two factors, with the greatest degree of use of the listener's privileged competitor being when there was no role switching (see Figure 4). It's possible that the switching condition in this case reduced the pressure on the speaker to unilaterally integrate the perspective of their partners that is granted primarily through the channel of direct feedback. Indirect feedback about their partner's knowledge may not figure into their subsequent production in the switching conditions. This is because some of the information is coming during trials in which they do not have to integrate this information into an utterance themselves, and for which it is unnecessary from their perspective, thus reducing its salience.

Speech onset offers a window into feedback's anticipatory effect in regards to how much planning speakers have to do when they are aware of the possibility that their addressee can make a clarification request. Planning (time to onset of speech) was influenced by the opportunity for feedback $\left(\chi_{(1)}^{2}=10.235, p<0.01\right)$ with planning time decreasing with the presence of feedback. This is likely due to reduced pressure on the speaker to carefully craft their utterances due to
the lack of a chance for correction. Time needed for planning showed a significant decreasing trend over the course of an experimental session $\left(\chi_{(1)}^{2}=34.045, p<0.01\right)$, which is consistent with a general increase in familiarity with the task.

## Gaze Analysis

One of the primary claims of the hypotheses motivating this experiment is that patterns of gaze during planning may reflect the underlying categorical structure a speaker is considering. Rehder and Hoffman (2005) found that in category discrimination tasks participants learned to only attend to the features in a display that are required for discriminating category membership for the pattern as a whole. Likewise, it was hypothesized that gaze in a referential communication task would demonstrate the shifting category membership status of the objects under consideration by the speaker. For these analyses, due to the absence of eye-tracking data for the second participant in the switching group, and the inclusion of trial order as a factor in the analyses, only the first sixteen trials of the non-switching group were analyzed and compared with the performance of the eye-tracked partner in the switching group.

Attention directed to the privileged competitor (measured here as the proportion of pre-onset fixation time) does not appear to be influenced by the same factors that influence the use of the privileged competitor. In this case, the only factor of influence is feedback ( $\chi_{(1)}^{2}=4.820, p<0.05$ ), with feedback being associated with increased attention being directed to the privileged competitor. Theoretically, attention to the shared competitor should be relatively insensitive to condition due to its constant use regardless of circumstance, but it appears to be more heavily influenced by experimental factors than attention to the privileged competitor. This is demonstrated by a three-way interaction between switching, feedback, and trial order $\left(\chi_{(1)}^{2}=4.777, p<0.05\right)$, in which feedback led to more attention to the shared competitor at the beginning of the experiment, but less at the end. This may be due to feedback initially bringing a greater focus to the


Trial
Trial

Figure 3: Proportion of utterances that refer to the dimension that is seen only from the speaker's perspective.


Figure 4: Proportion of utterances that refer to the dimension that is seen only from the addressee's perspective.
shared competitor as parters in the role switching condition acclimated to the task.

## Discussion

Speakers and their addressees tend to align their conceptual and semantic representations over the course of a conversation. However, the final form that these representations takes may depend on the characteristics and constraints under which the interaction is undertaken. In this experiment, each participant had information that was privileged and the goal was to determine what effect conversational constraints, such as role constancy and the opportunity (and actuality) of feedback, had on the nature of the interlocutors' apparent representations as evidenced by the form of their utterances. These changed representations are supposed to reflect the creation of socially emergent categories that arise out of the alignment process.

Interestingly, the most successful descriptions included the listener's privileged competitor. This is why performance was lowest by far in the most feedback-impoverished condition. This is notable because the task was, in principle,
solvable if the addressee aligned to the speaker by realizing that they could not see the addressee-privileged competitor. Just as the speaker had the opportunity to learn what the listener's perspective was through feedback, the listener could infer the speaker's perspective through their speech. However, the speaker's use of their own privileged perspective had relatively little bearing on the outcome of an individual trial.

Speakers seemed to acquire information about the addressee's perspective most effectively through direct, corrective feedback. Despite the opportunity to model their partner's knowledge, or at least benefit from priming, when they occupied the directing role in the role switching condition, speakers appeared to primarily acquire this information through direct feedback such as the listener asking about their privileged dimension. However, the significant role switching by feedback interaction speaks to a boosted effect of feedback when partners did not switch roles. This is interesting because it suggests that perhaps it is more beneficial to have role constancy and consistent feedback than to experience both roles and be in their partner's shoes. A participant who is consistently the director may feel it is part of their
role to adapt to their partners needs, whereas partners who switched roles may have been less sure about which partner should be aligned to. On the other hand, a speaker's use of their own privileged competitor seemed to be sensitive only to role switching, with use staying at a constant, high rate when partners have stable roles, but declined steadily when they switch roles. One possible explanation is that addressee feedback is more likely to be about communicating the dimension that is salient to them, and is not used to elicit clarifications about the superfluous (from their perspective) dimension used by the speaker. An impoverished description is potentially more damaging, than over-specification, but is amenable to feedback correction (Gann \& Barr, 2012). Direct feedback could put an upward pressure on the use of the addressee's perspective, but be neutral in regards to the speaker's. Thus salience from the speakers perspective is a sufficient reason for a speaker to include information in a given description, even if it has no apparent communicative value to the addressee.

Thus, when speakers are role switching and have feedback, they are likely to align in a way that favors aligning to the addressee through the speaker's increasing use of the listener's privileged dimension. This suggests these participants are moving toward a conceptualization of the shared perspective that is a "union" of their perspectives: The speaker makes use of adjectives related to both partners privileged dimensions. The adaption that results in this union is driven by the structure of the task, and the quality of the feedback experienced. Additionally, as seen in previous experiments such as Wardlow Lane, Groisman, and Ferreira (2006), anything that raises the salience of a particular competitor is likely to increase mention of that competitor, which might explain why the privileged competitor continues to be used despite it only being relevant from the speaker's perspective. Because its use doesn't seem to have a negative effect, there is little pressure to reduce its use absent another factor.

The online measures paint a more mixed picture. The effect of speech onset in regards to feedback seemed to mirror that which was seen in the prior experiment. When speakers were not permitted the opportunity for direct advance feedback, they took a little longer to plan their utterances. The eye-tracking data were less clear. At the outset it was anticipated that the attention paid to the competitors would help reveal the underlying conceptual structure the speaker was consulting in order to formulate her utterances (Rehder \& Hoffman, 2005). However, because the listener's competitor is never visible to the speaker, there was no way to match up eye-movements with its consideration, except for perhaps if the speaker preferentially fixates nothing when considering their partner's knowledge (which is one potentially interesting future analytical direction). In addition, the relative salience of the competitors on the screen was under different selective pressures than the features in the categorization task of Rehder and Hoffman (2005), as additional decisions had to be made about how to describe the contrast rather than
making a simple judgment of category membership.

## Acknowledgments

A version of this work was included in Gann (2012) as part of the author's doctoral dissertation.

## References

Barr, D. J. (2003). Paralinguistic correlates of conceptual structure. Psychonomic Bulletin and Review, 10(462-467).
Barr, D. J., \& Kronmüller, E. (2007). Conversation as site of category learning and category use. In B. H. Ross \& A. Markman (Eds.), The Psychology of Learning and Motivation (Vol. 47). Academic Press.
Barr, D. J., Levy, R., Scheepers, C., \& Tily, H. J. (2013). Random effects structure in mixed-effects models: Keep it maximal. Journal of Memory and Language.
Barsalou, L. W. (1991). Deriving categories to achieve goals. In G. H. Bower (Ed.), The Psychology of Learning and Motivation: Advances in Research and Theory (Vol. 27, pp. 1-64). San Diego, CA: Academic Press.
Bates, D., \& Maechler, M. (2010). Ime4: Linear mixedeffects models using S4 classes [Computer software manual].
Gann, T. M. (2012). Toward a comprehensive theory of audience design. Unpublished doctoral dissertation, University of California, Riverside.
Gann, T. M., \& Barr, D. J. (2012). Speaking From Experience: Audience Design as Expert Performance. Language and Cognitive Processes.
Garrod, S., \& Doherty, G. (1994). Conversation, coordination and convention: an empirical investigation of how groups establish linguistic conventions. Cognition, 53, 181-215.
Horton, W. S., \& Gerrig, R. J. (2005). Conversational common ground and memory processes in language production. Discourse Processes, 40(1), 1-35.
Krauss, R. M., \& Weinheimer, S. (1966). Concurrent feedback, confirmation, and the encoding of referents in verbal communication. Journal of Personality and Social Psychology, 4, 343-346.
Pickering, M. J., \& Garrod, S. (2004). Toward a mechanistic psychology of dialogue. Behavioral and Brain Sciences, 27, 1-22.
Rehder, B., \& Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. Cognitive Psychology, 51, 1-41.
Schober, M., \& Clark, H. H. (1989). Understanding by addressees and overhearers. Cognitive Psychology, 21(2), 211-232.
Wardlow Lane, L., Groisman, M., \& Ferreira, V. S. (2006). Don't talk about pink elephants! Speakers' control over leaking private information during language production. Psychological Science, 17(4), 273-277.

# Production of referring expressions: Preference trumps discrimination 

Albert Gatt (albert.gatt@um.edu.mt)<br>Institute of Linguistics, University of Malta<br>Tilburg center for Cognition and Communication (TiCC), Tilburg University<br>Emiel Krahmer (e.j.krahmer@uvt.nl) Tilburg center for Cognition and Communication (TiCC), Tilburg University

Roger P.G. van Gompel (r.p.g.vangompel@dundee.ac.uk)
School of Psychology, University of Dundee
Kees van Deemter (k.vdeemter@abdn.ac.uk)
Department of Computing Science, University of Aberdeen


#### Abstract

When referring to an object using a description, speakers need to select properties which jointly distinguish it from any potential distractors. Previous empirical and computational work addressing this content selection process has highlighted the role of both (i) the discriminatory power of properties of a referent, i.e. how many of the distractors in a domain each property excludes; (ii) how inherently salient or preferred a property is. To date, there has been no attempt to systematically investigate the trade-off between these two potentially competing motivations. This paper investigates experimentally the extent to which speakers take discriminatory power versus preference into account during content selection for reference production. Our results suggest that discriminatory power in fact plays a relatively unimportant role. We discuss the implications of this for computational models of reference production. Keywords: Referring expressions, language production, psycholinguistics, computational modelling


## Introduction

Referring expressions such as the large bottle are an essential feature of communication. Without the ability to refer, it would be difficult to ground our communicative efforts in the physical and mental world. The processes underlying reference production have been the object of intensive study by researchers in Computational Linguistics (see Krahmer \& van Deemter, 2012, for an extensive review) and Experimental Psycholinguistics (e.g. Levelt, 1989; Arnold, 2008). Many researchers agree that the primary aim of a referring expression is to identify an object for an interlocutor, a position that is rooted in a long tradition of philosophical work on the subject (e.g. Searle, 1969).

Consider a situation in which a speaker needs to identify an object (the target referent), which has not been introduced earlier in the discourse and which is visually co-present for both speaker and listener. Here, the speaker needs to perform content selection, to determine which properties of the target referent to mention in a description. This process is non-trivial because objects have several properties to choose from; moreover, the goal of identification entails choosing a set of properties
that jointly exclude all the distractors in the domain with which a listener might confuse the target. The speaker has to tread a fine line between efficiency on the one hand and sufficient detail on the other. Thus, it would seem desirable to avoid producing an overspecified description which contains more properties than required, or an underspecified one, which does not succeed in identifying the target. Both constraints would seem to follow to the extent that speakers observe Grice's Maxim of Quantity (Grice, 1975).

To take an example, the bottle with the black border in Figure 1 has three properties that are potentially distinguishing, namely, its colour, its size and the fact that it is marked with a black diamond (hereafter referred to as its pattern). On its own, none of these properties is sufficient to distinguish it from the distractors, the other bottles in the domain. Closer inspection reveals that this target minimally requires two properties (in fact, any two of these three) for successful identification. For example, the large bottle with a diamond would do the trick without overspecifying.


Figure 1: Example domain: the target's diamond pattern excludes 4 distractors, while its green colour excludes the 3 blue bottles on the bottom row

## Models of content selection

What process would best model speakers' content selection procedure? It is widely accepted that, since speech production is incremental (cf. Levelt, 1989; Pechmann,
1989), properties would be selected one after the other. The main question is: on what basis is the choice made at each point? One possibility would be for the speaker to weigh each property in terms of its discriminatory power. For instance, looking at Figure 1, it is easy to see that starting with the target's pattern would eliminate four distractors, while either of the other two properties would eliminate only three. Hence, a possible strategy for a speaker might be to always select the most discriminatory property, given the state of the domain and the description. In this case, once a pattern is chosen, either one of the remaining two properties (colour or size) would suffice to completely distinguish the target, since their discriminatory power is equal. This strategy is embodied in a well-known computational algorithm for the automatic generation of referring expressions, the Greedy heuristic (Dale, 1989). In the psycholinguistic community, it has been proposed most explicitly in the theoretical work of Olson (1970). Olson suggested that speakers 'specify the object to the level required by the listener to differentiate the intended referent from the alternatives' (p. 244-5). One way of interpreting this, assuming an incremental procedure, is that the speaker weighs the contribution of each available property to the ultimate goal of identification, choosing the one that is most likely to help in achieving it, as the Greedy heuristic does. ${ }^{1}$

In contrast to these models, experimental work has suggested that speakers' content selection processes tend to rely on heuristics related to the inherent salience of certain properties. The primary source of evidence for this is that some properties - notably an object's colour - tend to be used even when they do not contribute to the identification of a target, leading to an overspecified description. By contrast, other properties, such as size (or, presumably, pattern in the sense being used here), tend to be used only when absolutely required. In the case of size, its relatively dispreferred status is likely due to its being a relative property, requiring comparison to other objects in the domain.

These results are extremely robust (see Pechmann,

[^66]1989; Belke \& Meyer, 2002; Engelhardt, Bailey, \& Ferreira, 2006, among many others) and appear to persist even when the colour of an object doesn't differ too starkly from that of its distractors (e.g. the target is light green, whereas a distractor is a darker shade of green; see Viethen, Goudbeek, \& Krahmer, 2012). According to Pechmann (1989), this can be explained with reference to the fact that when speakers incrementally select properties, they initiate their descriptions before having completely scanned a domain. The preference for a property like colour - which may be related to its being an inherently salient attribute of perceived objects (e.g. Pechmann, 1989) - would make it more likely for that property to be selected before others.

In short, the evidence suggests that a property's discriminatory value is not the only consideration in content selection. In the computational literature, this evidence inspired the development of the well-known Incremental Algorithm for the generation of referring expressions (Dale \& Reiter, 1995). In contrast to the Greedy heuristic, the Incremental procedure works by selecting properties one by one on the basis of their preference rather than their discriminatory value. Given an ordering of properties by their preference, the algorithm considers each in turn. If a property excludes some distractors, it is included in the description, and the distractor set is updated, before considering the next property. Like the Greedy heuristic, this algorithm terminates when the description is fully distinguishing.

In our example domain, the Incremental Algorithm would thus start with the target's colour rather than its pattern. This excludes all the blue objects, leaving two other green objects. If, in the predefined preference ordering, pattern follows colour, this is the next property that would be considered. Since pattern excludes both remaining distractors, the description generated is the green bottle with a diamond.

Thus, there are two potentially conflicting motivations underlying content selection: discriminatory power and preference. The potential trade-off is exemplified in Figure 1, where the most discriminatory property (pattern) is not the most preferred one (colour).

In spite of the evidence for preferences stemming from overspecification, there is to our knowledge no research that explicitly tests the predictions of the two models, although some of the implications of the two strategies are evident in recent computational psycholinguistic work. Gatt, van Gompel, Krahmer, and van Deemter (2011) and van Gompel, Gatt, Krahmer, and van Deemter (2012) propose a non-deterministic model of reference production called PRO, which follows one of two different paths, each of which involves the throw of a dice, loaded to reflect the degree of preference of a set of properties. Path 1 is only followed if there exists a property that rules out all distractors (the limiting case of what
we have called discriminatory power): the output of the algorithm in this case is a description containing this one property. Should several properties rule out all distractors then the (preference-loaded) dice is thrown to choose one of them. Path 2 is followed if no such property exists (as in Figure 1). Here, properties are added incrementally to the description until all distractors have been removed. Which property is chosen next is based on a throw of the dice. Once all distractors have been removed, however, the dice is thrown again to determine whether to terminate or include one more property; if the latter decision is taken, then the dice is thrown again to decide whether to terminate after that, or continue, and so on.Thus, preference does not only govern the choice between properties, it also governs the decision whether or not to over-specify.

Although PRO was found to have an excellent fit to production data, it was compared to human-produced descriptions in very simple domains in which there were only two properties available (colour and size) and one property always sufficed to distinguish the target referent. Thus, it is an open question whether speakers computed relative discriminatory power, or more simply based their strategy on the limiting case, namely, the availability of a fully distinguishing property.

More recently, Frank and Goodman (2012) proposed a Bayesian model to predict property choice ${ }^{2}$ in very simple language games in which a speaker has to choose one property (such as blue vs. circle) to describe an object in a domain. Although this work does not explicitly address identification, it is nevertheless highly relevant to the present discussion. In this model, the speaker's choice of a property given a referent is based on utility. Letting $p$ be some property of the referent, $P$ the set of available properties, and $|p|$ stand for the number of objects of which $p$ is true, the likelihood of using $p$ is

$$
\frac{|p|^{-1}}{\sum_{q \in P}|q|^{-1}}
$$

This definition approaches the notion of discriminatory power being discussed here, because the utility of $p$ increases the fewer objects it is true of (i.e. the more distractors it eliminates). However, this model does not consider preference. A more serious shortcoming is that the model assumes that a speaker can only refer using a single property; thus, it would never overspecify. Indeed, it turns out that the utility function over-estimates speakers' tendency to underspecify. Consider a case where a referent is both large and green. Assume that there is an additional green distractor, but no other large distractors (size is fully distinguishing). In this case, the probability of using the property large works out to 0.67 ,

[^67]with a probability of 0.33 for green. In the experiment reported by Gatt et al. (2011), which contained a condition precisely analogous to this one, speakers produced size-only descriptions only $17 \%$ of the time, and overspecified with both size and colour $83 \%$ of the time, a finding that tallies with figures in the literature on overspecification. Speakers never produced an underspecified description with colour only. Thus, the model of Frank and Goodman too does not satisfactorily account for the interplay between discriminatory power and preference.

In summary, the question addressed by the present paper is: To what extent do preferences trump discrimination in the process of selecting properties incrementally? We investigate the issue experimentally, using domains such as the one exemplified above. If speakers tend to prioritise properties by discriminatory power, then a property should be more often included if it is the most discriminatory one available, than if it is not. By contrast, if speakers prioritise properties by preference, then more preferred properties should be included more often than less salient properties.

## The experiment

The experiment traded off the discriminatory power of properties against their degree of preference, which was determined on the basis of previous empirical work. Our aim was to investigate which of the two heuristics outlined in the preceding discussions - one that prioritises properties based on preference, or one that does so based on discriminatory power - best matches speakers' content selection strategies. If preferences are more important, then the frequency with which properties are used should be independent of how discriminatory they are in different conditions. By contrast, if discriminatory power is more important, then a property should be used more often in case it is more discriminatory, regardless of whether it is highly preferred (as colour is) or not.

## Participants

The experiment was conducted at the Tilburg center for Cognition and Communication. 72 native speakers of Dutch (49 female, 23 male), all undergraduate students at Tilburg University, participated in return for course credit. All had normal or corrected-to-normal vision and none reported any problems with colour perception.

## Materials and design

The experimental stimuli consisted of 36 items selected from a version of the Snodgrass and Vanderwart set of line drawings with colour and texture (Rossion \& Pourtois, 2004). The items were selected on the basis of a pretest in which seven native speakers of Dutch were asked to name greyscale versions of the pictures. For the items, we selected only those pictures for which at least 5 out of the 7 speakers agreed on the name of the object. The pictures were subsequently manipulated to create a
version of each in two different sizes (large/small) and four different colours (red, blue, green and grey), with three superimposed patterns (a circle, a diamond or a square).

The rationale for using these three properties was as follows. First, there is a lot of previous work indicating that colour is highly preferred over size (see above). Second, the choice of pattern as the third property was based on its having to be realised (in Dutch, the language of the experiment) as a post-modifier, while size and colour tend to be realised as pre-modifiers, with a relatively fixed order (see e.g. Gatt et al., 2011, for previous work manipulating colour and size with similar materials). To the extent that the syntactic linearisation of properties reflects their order of selection, this would suggest that pattern would be selected after the other two. Be that as it may, however, we wanted to ensure minimal variation in syntactic ordering of the properties involved.

For each item, a visual domain was constructed, consisting of a target referent indicated by a black border, and five distractors. In each domain, all objects (target and distractors) were of the same type (e.g. all were bottles). In every domain, the target could be minimally distinguished from its distractors via any subset of two of its properties. As an example, the bottle in Figure 1 can be distinguished from its distractors by using its colour and pattern (the green bottle with a diamond), its colour and size (the large green bottle) or its size and pattern (the large bottle with a diamond). Each item was used in one of three conditions; the difference between conditions lay in which property of the target had the highest discriminatory value. One property was designated the most discriminatory property (hereafter $\mathbf{m d} \mathbf{p}$ ): this property excluded four of the distractors. The other two properties were equally discriminatory and each excluded three distractors. For example, in Figure 1, the MDP is pattern.

Note that, regardless of which property was the MDP, two properties were always minimally needed to distinguish the target. A description which mentioned all three properties would be overspecified, while one that mentioned only one property would be underspecified. As a result, there is no length confound: distinguishing descriptions are equally long in all conditions, unless they are over- or underspecified. This setup excludes another possibility, namely that speakers are biased to select a single, fully distinguishing property if one exists. This could happen, for example, because when a target has such a property (e.g. the target is the only red object), it becomes so salient that it induces a 'popout' effect (Treisman \& Gelade, 1980). While such effects have been reported in experiments on visual search, they have recently also been found to influence reference production as well (Gatt, van Gompel, Krahmer, \& van

Deemter, 2012). Here we are interested in testing a subtler notion of discriminatory power, in a more complex domain configuration.

In each trial, objects were presented in a sparse grid. For each item, the position of the target was fixed in advance and was the same in all conditions. Both items and participants were randomly divided into 3 groups. Item and participant groups were rotated through a Latin square so that each item appeared in each condition and each participant saw all conditions, but each participant saw each item only once. The experiment consisted of 36 trials, with 108 fillers. 36 of these were constructed using the objects with the same three properties as those used in the experiment. However, the type of a target sufficed to distinguish the target. The remaining 72 fillers consisted of targets that could be distinguished from their distractors using a variety of properties (such as stripes, spots etc).

## Procedure

The experiment was run using DMDx (Forster \& Forster, 2003), and used a director-matcher paradigm. Participants were divided into 36 pairs, with one randomly assigned to the role of speaker/director and the other to the role of listener/matcher. Participants did not switch roles. The director and matcher faced each other; each had a computer screen that could not be seen by the other. The speaker used a keyboard to request an item, whereupon she identified the target for the listener, who clicked on the target on his own screen. Participants were instructed to keep the interaction to a minimum, with the listener only responding by indicating to the speaker that he had finished identifying the target. The speaker's descriptions were recorded through a headset.

## Data coding

Descriptions were transcribed and coded according to which of the three properties of a target (colour, size and/or pattern) were mentioned. This classification ignored the mention of the object's type (e.g. 'bottle'), which we assumed would be included in any case and which, in our design, had no discriminatory value. A description was further classified as follows (i) Wellspecified if it contained exactly two properties (excluding type) of the target; Overspecified if it mentioned all three properties; or (iii) Underspecified if it mentioned only one property, or only the object's type. The descriptions were further coded according to whether they included the MDP or not.

## Results

In what follows, we report results from logit mixed effects (LME) analyses with Condition as fixed effect and random intercepts for participants and items.

Table 1 displays the proportion of overspecified, wellspecified and underspecified descriptions in each condi-

| MDP | Well-spec | Overspec | Underspec |
| :---: | :---: | :---: | :---: |
| Colour | 0.699 | 0.296 | 0.005 |
| Size | 0.685 | 0.310 | 0.005 |
| Pattern | 0.676 | 0.324 | 0.000 |

Table 1: Proportion of well-specified, overspecified and underspecified descriptions, by condition
tion. The number of underspecified descriptions was minimal overall (4 in total) and the rate of overspecification does not appear to differ greatly across conditions. We recoded descriptions to indicate whether each one was overspecified or not. There was no effect of condition on the likelihood of overspecification $(Z=1.02, p>.3)$, that is, the likelihood of overspecifying did not depend on which property was the MDP.

| MDP | Colour | Size | Pattern |
| :---: | :---: | :---: | :---: |
| Colour | 0.99 | 0.73 | 0.57 |
| Size | 0.99 | 0.76 | 0.55 |
| Pattern | 0.99 | 0.73 | 0.59 |

Table 2: Proportion of descriptions containing colour, size and pattern (columns) in each condition (rows)

Table 2 displays the proportion of descriptions containing colour, size or pattern, in each condition. There are two striking facts about the data: (i) the frequency with which any of the three properties was used was largely independent of the condition, that is, whether that property was the most discriminatory one or not; (ii) there is clear evidence for preferences, with colour being used more frequently than size, and size more frequently than pattern.

An LME analysis showed a highly significant effect of condition on the likelihood with which participants used the MDP $(Z=-4.33, p<.001)$; this effect was also found when the analysis was repeated focusing only on well-specified descriptions, that is, those containing two of the three properties $(Z=4.38, p<.001)$. The results show quite unambiguously that whether or not the MDP was used turned out to depend on whether it was colour, size or pattern. This is a clear indication that preference trumps discriminatory power, not the other way round.

## Discussion

Our results suggest that speakers are insensitive to subtle differences in the discriminatory power of properties, relying on preference-based heuristics. Previous work has gleaned evidence for such heuristics from overspecification data, which is further used to argue against the notion that speakers observe a strict interpretation of the Gricean Maxim of Quantity. The present experiment manipulated both property preference (by contrasting colour, size, pattern) and discriminatory power (by orthogonally manipulating whether each of these properties is most discriminatory).

The evidence shows that preference has an effect on
how frequently a property is chosen, but speakers are relatively insensitive to subtle differences in discriminatory power. We draw this conclusion from the clear tendency to make selections on the basis of which property is involved, rather than its contrastive value. Thus, colour, for example, is highly preferred and tends to be chosen irrespective of its discriminatory power. It is possible that the marked preference for colour is due to the fact that in our domains (e.g. 1), it becomes more salient since it characterises the entire object (e.g. a bottle is green in its entirety), whereas pattern, for example, does not. However, the consistency of our results in this regard with previous work (e.g. Pechmann, 1989; Belke \& Meyer, 2002) suggests that the colour preference reflects a more general tendency.

The findings directly contradict computational models such as the Greedy heuristic (Dale, 1989), as well as proposals in the psycholinguistic literature based on theoretical work such as that of Olson (1970). In contrast, it suggests that models such as the Incremental Algorithm (Dale \& Reiter, 1995) are on the right track, insofar as they make choices based on preference. On the other hand, recent work suggests that this algorithm does not give a complete picture of human reference production either. One of its limitations is that it only selects a property if it excludes some of the (remaining) distractors at a given point in the procedure, something that has been shown not to hold of human speakers (Viethen, Dale, \& Guhe, 2011). Another is that the procedure is entirely deterministic and always produces the same output for a given input and a given preference ordering among available properties. In contrast, human speakers appear to treat the notion of preference stochastically, so that a model that interprets preferences in terms of a probability distribution fits human data better (Gatt et al., 2011).

This brings us to our earlier discussion of probabilistic models. One interesting question is raised by the PRO model of van Gompel et al. (2012). This model first tries to find a property which fully distinguishes the target referent. Additional content selection decisions are carried out probabilistically based on preference. As we have discussed, this model has been shown to have a remarkably good fit to data elicited from human speakers, albeit in much simpler domains than the ones used here. Now, a possible generalisation of this model would be one that first looks for the most discriminatory property available, rather than a fully distinguishing one. The results of the present experiment, which explicitly excluded the possibility of there being a single distinguishing property for the target, suggest that this would not improve its goodness of fit. However, it should be noted that our results are based on domains in which the difference in discriminatory power between the most distinguishing property and the others is exactly 1 . Would a greater
difference motivate speakers to select the MDP, even if it was highly dispreferred? A positive answer to this question would imply that the PRO model can indeed be generalised to look for highly discriminatory properties, but only if their discriminatory value was relatively high, making them very visually salient. Thus, sensitivity to discriminatory power might fall on a continuum.

A similar point can be made about Frank and Goodman (2012)'s Bayesian model, which estimates the likelihood of a property being used for a referent as a function of the number of potential referents of that property, and the number of properties that the referent may be distinguished by. Modulo the simplification inherent in this work, namely that referents are to be distinguished using a single property, it would be interesting to investigate to what extent this notion of utility is also gradable and impacts visual salience.

## Conclusions and future work

This paper investigated content selection in reference production. It addressed the possible trade-off between (i) the discriminatory power of a property, that is, the extent to which it is likely to help in the task of distinguishing a referent from its distractors and (ii) the extent to which a property is preferred. Our results suggest that subtle differences in discriminatory power do not influence content selection choices. One question that is left open by the present work is whether larger discriminatory power differences would alter these findings, something that we plan to investigate in future work.

## Acknowledgments

Albert Gatt and Emiel Krahmer were supported by a grant from the Netherlands Organization for Scientific Research (NWO) to the project Bridging the gap between psycholinguistics and computational linguistics: The case of Referring Expressions. Thanks to Jette Viethen and Ruud Koolen for useful comments on a previous version of this paper.

## References

Arnold, J. (2008). Reference production: Productioninternal and addressee-oriented processes. Language and Cognitive Processes, 23(4), 495-527.
Belke, E., \& Meyer, A. (2002). Tracking the time course of multidimensional stimulus discrimination: Analysis of viewing patterns and processing times during same-different decisions. European Journal of Cognitive Psychology, 14 (2), 237-266.
Dale, R. (1989). Cooking up referring expressions. In Proc. ACL'89.
Dale, R., \& Reiter, E. (1995). Computational interpretation of the Gricean maxims in the generation of referring expressions. Cognitive Science, 19(8), 233263.

Engelhardt, P. E., Bailey, K., \& Ferreira, F. (2006). Do speakers and listeners observe the Gricean Maxim of Quantity? Journal of Memory and Language, 54, 554-573.
Forster, K. I., \& Forster, J. C. (2003). DmDx: A windows display program with millisecond accuracy. Behavior Research Methods, Instruments, $\mathcal{E}$ Computers, 35, 116-124.
Frank, M., \& Goodman, N. (2012). Predicting pragmatic reasoning in language games. Science, 336, 998.
Gatt, A., van Gompel, R., Krahmer, E., \& van Deemter, K. (2011). Non-deterministic attribute selection in reference production. In Proc. PreCogSci'11.
Gatt, A., van Gompel, R., Krahmer, E., \& van Deemter, K. (2012). Does domain size impact speech onset time during reference production? In Proc. CogSci'12.
Grice, H. (1975). Logic and conversation. In P. Cole \& J. Morgan (Eds.), Syntax and semantics: Speech acts. (Vol. III). New York: Academic Press.
Krahmer, E., \& van Deemter, K. (2012). Computational generation of referring expressions: A survey. Computational Linguistics, 38(1), 173-218.
Levelt, W. (1989). Speaking: From intention to articulation. Cambridge, Ma.: MIT Press.
Olson, D. R. (1970). Language and thought: Aspects of a cognitive theory of semantics. Psychological Review, 77, 257-273.
Pechmann, T. (1989). Incremental speech production and referential overspecification. Linguistics, 27, 89110.

Reiter, E. (1990). The computational complexity of avoiding conversational implicatures. In Proc. acL '89.
Rossion, B., \& Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's object databank : The role of surface detail in basic level object recognition. Perception, 33, 217-236.
Searle, J. R. (1969). Speech acts: An essay in the philosophy of language. Oxford: Oxford University Press.
Treisman, A., \& Gelade, G. (1980). A feature-integration theory of attention. Cognitive Psychology, 12(1), 97136.
van Deemter, K., Gatt, A., van der Sluis, I., \& Power, R. (2012). Generation of referring expressions: Assessing the Incremental Algorithm. Cognitive Science, 36(5), 799-836.
van Gompel, R., Gatt, A., Krahmer, E., \& van Deemter, K. (2012). Pro: A computational model of referential overspecification. In Proc. amlap'12. Trento, Italy.
Viethen, J., Dale, R., \& Guhe, M. (2011). Serial dependency: Is it a characteristic of human referring expression generation? In Proc. PreCogSci'11.
Viethen, J., Goudbeek, M., \& Krahmer, E. (2012). The impact of colour difference and colour codability on reference production. In Proc. CogSci'12.

# Bayesian Vector Analysis and the Perception of Hierarchical Motion 

Samuel J. Gershman ${ }^{1}$ (sjgershm@mit.edu), Frank Jäkel ${ }^{2}$ (fjaekel@uos.de), Joshua B. Tenenbaum ${ }^{1}$ (jbt@mit.edu)<br>${ }^{1}$ Department of Brain and Cognitive Sciences, MIT<br>${ }^{2}$ Institute of Cognitive Science, University of Osnabrück


#### Abstract

Scenes filled with moving objects are often hierarchically organized: the motion of a migrating goose is nested within the flight pattern of its flock, the motion of a car is nested within the traffic pattern of other cars on the road, the motion of body parts are nested in the motion of the body. Humans perceive hierarchical structure even in stimuli with two or three moving dots. An influential theory of hierarchical motion perception holds that the visual system performs a "vector analysis" of moving objects, decomposing them into common and relative motions. However, this theory does not specify how to resolve ambiguity when a scene admits more than one vector analysis. We describe a Bayesian theory of vector analysis and show that it can account for classic results from dot motion experiments. Our theory takes a step towards understanding how moving scenes are parsed into objects.


Keywords: motion perception; Bayesian inference; structure learning

## Introduction

Motion is a powerful cue for understanding the organization of a visual scene. Infants use motion to individuate objects, even when it contradicts property/kind information (Kellman \& Spelke, 1983; Xu et al., 1999). The primacy of motion information is also evident in adult object perception (Mitroff \& Alvarez, 2007). In addition to individuating and tracking objects, motion is used by the visual system to decompose objects into parts. In biological motion, for example, the motion of body parts are nested in the motion of the body. Object motion may be hierarchically organized into multiple layers: an arm's motion may be further decomposed into jointed segments, including the hand, which can itself be decomposed into fingers, and so on.

The hierarchical organization of motion presents a formidable challenge to current models of motion processing. It is widely accepted that the visual system balances motion integration over space and time (necessary for solving the aperture problem) and motion segmentation in order to perceive multiple objects simultaneously (Braddick, 1993). However, it is unclear how simple segmentation mechanisms can be used to build a hierarchically structured representation of a moving scene. Segmentation lacks a notion of nesting: when an object moves, its parts should move with it. To understand nesting, it is crucial to represent the underlying dependencies between objects and their parts.

The experimental and theoretical foundations of hierarchical motion perception were laid by the pioneering work of Johansson (1950), who demonstrated that surprisingly complex percepts could arise from simple dot motions. Johansson proposed that the visual system performs a "vector analysis" of moving scenes into common and relative motions between
objects. In the example of biological motion (see Johansson, 1973), the global motion of the body is subtracted from the image, revealing the relative motions of body parts; these parts are further decomposed by the same subtraction operation.

While the vector analysis theory provides a compelling explanation of numerous motion phenomena (we describe several below), it is incomplete from a computational point of view, since it relies on the theorist to provide the underlying motion components and their organization; it lacks a mechanism for discovering a hierarchical decomposition from sensory data. This is especially important in complex scenes where many different vector analyses are consistent with the scene. Various principles have been proposed for how the visual system resolves this ambiguity. For example, Restle (1979) proposed a "minimum principle," according to which simpler motion interpretations (i.e., those with a shorter description length) are preferred over more complex ones. Gogel (1974) argued for an "adjacency principle," according to which the motion interpretation is determined by relative motion cues between nearby points. However, there is still no unified computational theory that can encompass all these ideas.

In this paper, we recast Johansson's vector analysis theory in terms of a Bayesian model of motion perception. The model discovers the hierarchical structure of a moving scene, resolving the ambiguity of multiple vector analyses using a set of probabilistic constraints. We show that this model can account for several classic phenomena in the motion perception literature that are challenging for existing models.

## Bayesian vector analysis

In this section, we describe our computational model formally. We start by describing a probabilistic generative model of motion-a set of assumptions about the environment that we impute to the observer. The generative model can be thought of as stochastic "recipe" for generating moving images. We then describe how Bayesian inference can be used to invert this generative model and recover the underlying hierarchical structure from observations of moving images.

## Generative model

Our model describes the process by which a sequence of twodimensional visual element positions $\left\{\mathbf{s}_{n}(t)\right\}_{n=1}^{N}$ is generated, where $\mathbf{s}_{n}(t)=\left[s_{n}^{x}(t), s_{n}^{y}(t)\right]$ is the $x$ and $y$ position of element $n$ at time step $t .{ }^{1}$ Elements can refer to objects, parts or features;

[^68]

Figure 1: Illustration of how a moving scene is decomposed into a motion tree. Each node in the tree corresponds to a motion component. Each object in the scene traces a path through the tree, and the observed motion of the object is modeled as the superposition of motion components along its path.
in this paper we will simply refer to them as objects. The object positions are modeled as arising from a tree-structured configuration of motion components; we refer to this representation as the motion tree. Each motion component is a transformation that maps the current object position to a new position.

An illustration of a motion tree is shown in Figure 1. Each node in the tree corresponds to a motion component. The motion of the train relative to the background is represented by the top-level node. The motions of Spiderman and Dr. Octopus relative to the train are represented at the second-level nodes. Finally, the motions of each body part relative to the body are represented at the third-level nodes. The observed motion of Spiderman's hand can then be modeled as the superposition of the motions along the path that runs from the top node to the hand-specific node. The aim for our model is to get as inputs the retinal motion of pre-segmented objectsin this example, the motion of hands, feet, torsos, windows, etc.-and output a hierarchical grouping that reflects the composition of the moving scene.

The motion tree can capture the underlying motion structure of many real-world scenes, but inferring which motion tree generated a particular scene is challenging because different trees may be consistent with the same scene. To address this problem, we need to introduce a prior distribution over motion trees that expresses our inductive biases about what kinds of trees are likely to occur in the world. This prior should be flexible enough to accommodate many different structures while also preferring simpler structures (i.e., parsimonious explanations of the sensory data). These desiderata are satisfied by a nonparametric distribution over trees known as the nested Chinese restaurant process (nCRP; Blei et al., 2010). The nCRP generates a motion tree by drawing, for each object $n$, a sequence of motion components, denoted by $\mathbf{c}_{n}=\left[c_{n 1}, \ldots, c_{n D}\right]$, where $D$ is the maximal tree depth. ${ }^{2}$ The component assignments are drawn according to:

$$
P\left(c_{n d}=j \mid \mathbf{c}_{1: n-1}\right)= \begin{cases}\frac{M_{j}}{n-1+\gamma} & \text { if } j \leq J  \tag{1}\\ \frac{{ }_{\gamma}}{n-1+\gamma} & \text { if } j=J+1\end{cases}
$$

[^69]where $j$ indexes motion components, $M_{j}$ is the number of previous objects assigned to component $j$, and $J$ is the number of components currently in use (i.e., those for which $M_{j}>0$ ). The assignment at depth $d$ is restricted to a unique set of components specific to the component assigned at depth $d-1$. In this way, the components form a tree structure, and $\mathbf{c}_{n}$ is a path through the tree. The parameter $\gamma \geq 0$ controls the branching factor of the motion tree. As $\gamma$ decreases, different objects will tend to share the same motion components. Thus, the nCRP exhibits a preference for trees that use a small number of motion components.

Note that so far we have generated a path through a potentially very deep tree for each object. Each path has the same length $D$. Remember that each node in the tree will represent a motion component. We want each object $n$ to be associated with a node in the tree and its overall motion to be the sum of all the motion components above it (including itself). Hence, for each object we need to sample an additional parameter $d_{n} \in\{1, \ldots, D\}$ that determines to which level on the tree the object will be assigned. This depth specifies a truncation of $\mathbf{c}_{n}$, thereby determining which components along the path contribute to the observations. The depth assignments $\mathbf{d}=\left[d_{1}, \ldots, d_{N}\right]$ are drawn from a Markov random field:

$$
\begin{equation*}
P(\mathbf{d}) \propto \exp \left\{\alpha \sum_{m=1}^{N} \sum_{n>m}^{N} \mathbb{I}\left[d_{m}=d_{n}\right]-\rho \sum_{n=1}^{N} d_{n}\right\} \tag{2}
\end{equation*}
$$

where the indicator function $\mathbb{I}[\cdot]=1$ if its argument is true and 0 otherwise. The parameter $\alpha$ controls the penalty for assigning objects to different depths, and the parameter $\rho$ controls a penalty for deeper level assignments.

Each motion component, i.e. each node in the motion tree, is associated with a time-varying flow field, $\mathbf{f}_{j}(\mathbf{s}, t)=$ $\left[f_{j}^{x}(\mathbf{s}, t), f_{j}^{y}(\mathbf{s}, t)\right]$. We place a prior on flow fields that enforces spatial smoothness but otherwise makes no assumptions about functional form. In particular we assume that $f_{j}^{x}$ and $f_{j}^{y}$ are spatial functions drawn independently at each time discrete time step $t$ from a zero-mean Gaussian process with covariance function

$$
\begin{equation*}
k\left(\mathbf{s}, \mathbf{s}^{\prime}\right)=\tau \exp \left\{-\frac{\left\|\mathbf{s}-\mathbf{s}^{\prime}\right\|^{2}}{2 \lambda}\right\} \tag{3}
\end{equation*}
$$

where $\tau$ is a global scaling parameter and $\lambda>0$ is a lengthscale parameter controlling the smoothness of the flow field. When $\lambda$ is large, the flow field becomes rigid. Smoothness is only enforced between objects covered by the same node in the motion tree.

To complete the generative model, we need to specify how the motion tree gives rise to observations, which in our case are the positions of the $N$ objects over time. For each object, the dot position at the next time step is set by sampling a displacement from a Gaussian whose mean is the sum of the flow fields along path $\mathbf{c}_{n}$ truncated at $d_{n}$ :

$$
\begin{equation*}
\mathbf{s}_{n}(t+1)=\mathbf{s}_{n}(t)+\sum_{d=1}^{d_{n}} \mathbf{f}_{c_{n d}}\left(\mathbf{s}_{n}(t), t\right)+\varepsilon_{n}(t) \tag{4}
\end{equation*}
$$

where $\varepsilon_{n}(t) \sim \mathcal{N}\left(\mathbf{0}, \sigma^{2} \mathbf{I}\right)$.
This generative model contains a number of important special cases under particular parameter settings. When $\gamma=0$, only one motion component will be generated; in this case, the prior on flow-fields-favoring local velocities close to 0 that vary smoothly over the image-resembles the "slow and smooth" model proposed by Weiss \& Adelson (1998). When $\gamma=0$ and $\lambda \rightarrow \infty$, we obtain the "slow and rigid" model of Weiss et al. (2002). When $D=1$, the model will generate multiple motion components, but these will all exist at the same level of the hierarchy (i.e., the motion tree is flat, with no nesting), resulting in a form of transparent layered motion (Wang \& Adelson, 1993; Weiss, 1997).

## Inference

The goal of inference is to compute the posterior over the motion tree given a set of observations. ${ }^{3}$ Because we are mainly interested in the highest probability tree, we use annealed Gibbs sampling to search for the posterior mode. The algorithm alternates between holding the depth assignments fixed while sampling the node assignments, and holding the node assignments fixed while sampling the depth assignments. By raising the conditional probabilities to a power $\beta>1$, the posterior becomes peaked around the mode. We gradually increase $\beta$, so that the algorithm eventually settles on a high probability tree. We repeat this procedure 10 times (with 500 sampling iterations on each run) and pick the tree with the highest posterior probability. Below, we derive the conditional distributions used by the sampler.

The conditional distribution over $\mathbf{c}_{n}$ is given by:

$$
\begin{equation*}
P\left(\mathbf{c}_{n} \mid \mathbf{c}_{-n}, \mathbf{s}, \mathbf{d}\right) \propto P\left(\mathbf{c}_{n} \mid \mathbf{c}_{-n}\right) P(\mathbf{s} \mid \mathbf{c}, \mathbf{d}) \tag{5}
\end{equation*}
$$

where $\mathbf{c}_{-n}$ denotes the set of all paths excluding $\mathbf{c}_{n}$. The first factor in Eq. 5 is the nCRP prior (Eq. 1). The second factor in Eq. 5 is the likelihood of the data, given by:

$$
\begin{equation*}
P(\mathbf{s} \mid \mathbf{c}, \mathbf{d})=\prod_{t} \prod_{z \in\{x, y\}} \mathcal{N}\left(\mathbf{s}^{z}(t+1) ; \mathbf{s}^{z}(t), \mathbf{K}(t)+\sigma^{2} \mathbf{I}\right) \tag{6}
\end{equation*}
$$

[^70]

Figure 2: Johansson (1950) two dot experiment. (A) Veridical motion vectors. (B) Perceived motion. (C) Inferred motion vectors. Each color corresponds to a different component in the motion tree $(D)$, but note that a component will predict different vectors depending on spatial location.
where

$$
\begin{equation*}
K_{m n}(t)=k\left(\mathbf{s}_{m}(t), \mathbf{s}_{n}(t)\right) \sum_{j} \mathbb{I}\left[j \in \mathbf{c}_{m} \wedge j \in \mathbf{c}_{n}\right] . \tag{7}
\end{equation*}
$$

Intuitively, the covariance between two points counts the number of nodes shared between their paths, weighted by their proximity in space.

The conditional distribution over $d_{n}$ is given by:

$$
\begin{equation*}
P\left(d_{n} \mid \mathbf{c}, \mathbf{s}, \mathbf{d}_{-n}\right) \propto P\left(d_{n} \mid \mathbf{d}_{-n}\right) P\left(\mathbf{s} \mid \mathbf{c}, \mathbf{d}_{-n}, d_{n}\right) \tag{8}
\end{equation*}
$$

where $\mathbf{d}_{-n}$ denotes the level assignments excluding $d_{n}$ and

$$
\begin{equation*}
P\left(d_{n} \mid \mathbf{d}_{-n}\right) \propto \exp \left\{\alpha \sum_{m \neq n} \mathbb{I}\left[d_{m}=d_{n}\right]-\rho d_{n}\right\} \tag{9}
\end{equation*}
$$

To visualize the motion components that are given by a grouping through $\mathbf{d}_{n}$ and $\mathbf{c}_{n}$, we can calculate the posterior predictive mean for object $n$ at each component $j$ (shown here for the $x$ dimension):

$$
\begin{equation*}
\mathbb{E}\left[f_{j}^{x}\left(\mathbf{s}_{n}(t), t\right)\right]=\mathbf{k}_{n j}^{\top}\left(\mathbf{K}(t)+\sigma^{2} \mathbf{I}\right)^{-1}\left(\mathbf{s}^{x}(t+1)-\mathbf{s}^{x}(t)\right), \tag{10}
\end{equation*}
$$

where $\mathbf{k}_{n j}$ is the $N$-dimensional vector of covariances between $\mathbf{s}_{n}(t)$ and the locations of all the objects whose paths pass through node $j$ (if an object does not pass through node $j$ then its corresponding entry in $\mathbf{k}_{n j}$ is 0 ).

## Simulations

In this section, we show how the Bayesian vector analysis model can account for several classic experimental phenomena. These experiments all involve stimuli consisting of moving dots, so for present purposes $\mathbf{s}_{n}(t)$ corresponds to the position of dot $n$ at time $t$. In these simulations we use the following parameters: $D=3, \sigma^{2}=0.01, \tau=1, \lambda=100, \alpha=1, \rho=$ 0.1 . The interpretation of $\sigma^{2}$ and $\lambda$ depend on the spatial scale


Figure 3: Johansson (1973) three dot experiment. (A) Veridical motion vectors. (B) Perceived motion. (C) Inferred motion vectors. ( $D$ ) Inferred motion tree.
of the data; in general, we found that changing these parameters (within the appropriate order of magnitude) had little influence on the posterior. We set $\lambda$ to be large enough so that objects assigned to the same layer moved near-rigidly.

Johansson (1950) demonstrated that a hierarchical motion percept can be achieved with as few as two dots. Figure 2A shows the stimulus used by Johansson, consisting of two dots translating orthogonally to meet at a single point. Observers, however, do not perceive the orthogonal translation. Instead, they perceive the two dots translating along a diagonal axis towards each other, which itself translates towards the meeting point (Figure 2B). Thus, observers perceive the stimulus as organized into common and relative motions. This percept is reproduced by the Bayesian vector analysis model (Figure 2C); the inferred motion tree (shown in Figure 2D) represents the common motion as the top level component and the relative motions as subordinate components. The subordinate components are not perfectly orthogonal to the diagonal motion, consistent with the findings of Wallach et al. (1985); this arises in our model through a form of "explaining away"i.e., posterior coupling between the motion layers implied by Eq. 10.

Another example studied by Johansson (1973) is shown in Figure 3A. Here the bottom and top dot translate horizontally while the middle dot translates diagonally such that all three dots are always collinear. The middle dot is perceived as translating vertically as all three dots translate horizontally (Figure 3B). Consistent with this percept, the Bayesian vector analysis assigns all three dots to a common horizontal motion component, and additionally assigns the middle dot to a vertical motion component (Figure 3C-D).

Duncker (1929) showed that if a light is placed on the rim of a rolling wheel in a dark room, cycloidal motion is perceived (Figure 4A), but if another light is placed on the hub then rolling motion is perceived (Figure 4B). Simulations of these experiments are shown in Figure 5. When a light is placed only on the rim, there is strong evidence for a single cycloidal motion component, whereas stronger evidence for a


Figure 4: Duncker wheel. (A) A light on the rim of a rolling wheel produces cycloidal motion. ( $B$ ) Adding a light on the hub produces rolling motion (translation + rotation).


Figure 5: Simulations of the Duncker wheel. (Top) A single light on the rim produces one vector following a cycloidal path. (Middle) Adding a light on the hub produces two vectors: translation + rotation, giving rise to the percept of rolling motion. (Bottom) Placing the light on the interior of the wheel produces weaker rolling motion: the translational component is no longer perfectly horizontal.
two-level hierarchy (translation + rotation) is provided by the hub light. ${ }^{4}$ It has also been observed that placing a light in between the rim and the hub produces weaker rolling motion (i.e., the translational component is no longer perfectly horizontal; Proffitt et al., 1979), a phenomenon that is reproduced by Bayesian vector analysis (Figure 5, bottom).

So far, we have been considering qualitative characterizations of various motion phenomena, but one advantage of a computational model is its ability to make quantitative predictions. We illustrate the quantitative power of Bayesian vector analysis for the case of motion transparency. When two groups of randomly moving dots are superimposed, observers may see either transparent motion (two planes of motion sliding past each other) or non-transparent motion (all dots moving in the direction of the average motion of the two groups). Which percept prevails depends on the relative direction of the two groups (Braddick et al., 2002): as the direction difference increases, transparent motion becomes more perceptible. We computed the probability of transparent motion (i.e.,

[^71]

Figure 6: Simulations of transparent motion. Transparency increases as a function of direction difference between two superimposed groups of dots.


Figure 7: Motion contrast. (A) The velocity of the background (black) dots increases along the horizontal axis. Although A and B have the same velocity, A is perceived as moving faster than B . (B) Model simulation.
two layers in our model) for a range of relative directions using 20 dots. As the relative direction increases, the statistical evidence in favor of two separate layers increases, resulting in a smoothly changing probability (Figure 6).

Inferences about the motion hierarchy may interact with the spatial structure of the scene. The phenomenon of motion contrast, originally described by Loomis \& Nakayama (1973), provides an illustration: The perceived motion of a dot depends on the motion of surrounding "background dots" (the black dots in Figure 7A). If a set of dots moves on a screen such that the dots on the left move more slowly than dots on the right, they form a velocity gradient. Two "target" dots that move with the same velocity and keep a constant distance (the red dots in Figure 7A) can still be perceived as moving with radically different speeds, depending on the speed of the dots close by. In our model, most of the motion of the velocity gradient is captured by the Gaussian process on the top-level motion component. However, this top-level component does not capture all of the motion of each dot. The target dots (in red), in particular, are each endowed with their own motion component and move relative to the toplevel node. This relative motion differs depending on where along the gradient the target dot is located, resulting in motion contrast (Figure 7B).

How does our model scale up to more complex displays?

An interesting test case is biological motion perception: Johansson (1973) showed that observers can recognize human motions like walking and running from lights attached to the joints. Later work has revealed that a rich variety of information can be discriminated by observers from point light displays, including gender, weight and even individual identity (Blake \& Shiffrar, 2007). We trained our model (with the same parameters) on point light displays derived from the CMU human motion capture database. ${ }^{5}$ These displays consisted of the 3 -dimensional positions of 31 dots, including walking, jogging and sitting motions. The resulting motion parse is illustrated in Figure 8: the first layer of motion (not shown) captures the overall trajectory of the body, while the second and third layers capture more fine-grained structure, such as the division into limbs and smaller jointed body parts. Note that the model knows nothing about the underlying skeletal structure; it infers body parts directly from the dot positions. This demonstrates that Bayesian vector analysis can scale up to more complex and realistic motion patterns.

## Conclusion

How does the visual system parse the hierarchical structure of moving scenes? In this paper, we have developed a Bayesian framework for modeling hierarchical motion perception, building upon the seminal work of Johansson (1950). The key idea of our theory is that a moving scene can be interpreted in terms of an abstract graph-the motion tree-encoding the dependencies between moving objects. Bayesian vector analysis is the process of inferring the motion tree from a sequence of images. Our simulations demonstrated that this formalism is capable of capturing a number of classic phenomena in the literature on hierarchical motion perception.

Two limitations of our theory need to be addressed. First, the generative model assumes that motion components combine through summation, but this is not adequate in general. For example, a better treatment of the Duncker wheel would entail modeling the composition of rotation and translation. In its current form, the model approximates rotation by inferring motion components that are tangent to the curve traced by the rotation. We are currently investigating a version of the generative model in which motion transformation compose with one another, which would allow for nonlinear interactions. Second, although we described an algorithm for finding the optimal motion tree, Bayesian vector analysis is really specified at the computational level; our simulations are not illuminating about the mechanisms by which the vector analysis is carried out. Nor does it commit to any particular neural implementation. More work is needed to connect all these levels of analysis. Grossberg et al. (2011) have described a detailed theory of how vector analysis could be performed by the visual cortex, and their efforts offer a possible starting point.

We view hierarchical motion as a model system for study-

[^72]

Figure 8: Analysis of human motion capture data. Each color represents the assignment of a node to a motion component. All nodes are trivially assigned to the first layer (not shown). In addition, all nodes were assigned to the second layer ( $A$ ). A subset of the nodes were also assigned components in the third layer $(B)$. Unfilled nodes indicate that no motion component was assigned at that layer. The skeleton is shown here for display purposes; the model was trained only on the dot positions.
ing more general questions about structured representations in mind and brain. The simplicity of the stimuli makes them amenable to rigorous psychophysical and neurophysiological experimentation, offering hope that future work can isolate the neural computations underlying structured representations like motion trees.

## Acknowledgments

We thank Ed Vul and Peter Battaglia for helpful discussions. This work was supported by the Deutsche Forschungsgemeinschaft (DFG JA 1878/1-1), ONR MURI N00014-07-10937, IARPA ICARUS program, and the MIT Intelligence Initiative.

## References

Blake, R., \& Shiffrar, M. (2007). Perception of human motion. Annual Review of Psychology, 58, 47-73.
Blei, D., Griffiths, T., \& Jordan, M. (2010). The nested chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. Journal of the ACM, 57, 7.
Braddick, O. (1993). Segmentation versus integration in visual motion processing. Trends in Neurosciences, 16, 263268.

Braddick, O., Wishart, K., \& Curran, W. (2002). Directional performance in motion transparency. Vision Research, 42(10), 1237-1248.
Duncker, K. (1929). Über induzierte Bewegung. (Ein Beitrag zur Theorie optisch wahrgenommener Bewegung). Psychologische Forschung, 12, 180-259.
Gogel, W. (1974). Relative motion and the adjacency principle. The Quarterly Journal of Experimental Psychology, 26, 425-437.
Grossberg, S., Léveillé, J., \& Versace, M. (2011). How do object reference frames and motion vector decomposition emerge in laminar cortical circuits? Attention, Perception, \& Psychophysics, 73(4), 1147-1170.
Johansson, G. (1950). Configurations in event perception. Almqvist \& Wiksell.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. Perception, \& Psychophysics, 14(2), 201-211.
Kellman, P., \& Spelke, E. (1983). Perception of partly occluded objects in infancy. Cognitive Psychology, 15(4), 483-524.
Loomis, J., \& Nakayama, K. (1973). A velocity analogue of brightness contrast. Perception, 2(4), 425-427.
Mitroff, S., \& Alvarez, G. (2007). Space and time, not surface features, guide object persistence. Psychonomic Bulletin \& Review, 14, 1199-1204.
Proffitt, D., Cutting, J., \& Stier, D. (1979). Perception of wheel-generated motions. Journal of Experimental Psychology: Human Perception and Performance, 5, 289302.

Restle, F. (1979). Coding theory of the perception of motion configurations. Psychological Review, 86(1), 1-24.
Wallach, H., Becklen, R., \& Nitzberg, D. (1985). Vector analysis and process combination in motion perception. Journal of Experimental Psychology: Human Perception and Performance, 11, 93-102.
Wang, J., \& Adelson, E. (1993). Layered representation for motion analysis. In Computer vision and pattern recognition (pp. 361-366).
Weiss, Y. (1997). Smoothness in layers: Motion segmentation using nonparametric mixture estimation. In Computer vision and pattern recognition (pp. 520-526).
Weiss, Y., \& Adelson, E. (1998). Slow and smooth: A bayesian theory for the combination of local motion signals in human vision. AI Memo 1616, MIT.
Weiss, Y., Simoncelli, E., Adelson, E., et al. (2002). Motion illusions as optimal percepts. Nature Neuroscience, 5(6), 598-604.
Xu, F., Carey, S., Welch, J., et al. (1999). Infants' ability to use object kind information for object individuation. Cognition, 70, 137-166.

# Space and Time in the Parietal Cortex: fMRI Evidence for a Neural Asymmetry 

Tom Gijssels ${ }^{1,2,4} \quad$ Roberto Bottini $i^{1,2,4} \quad$ Shirley-Ann Rueschemeyer ${ }^{3,4} \quad$ Daniel Casasanto ${ }^{1,2,4}$<br>(gijsselt@newschool.edu) (bottinir@newschool.edu)<br>(sr852@york.ac.edu) (casasand@newschool.edu)<br>${ }^{1}$ Department of Psychology, New School for Social Research, New York, USA<br>${ }^{2}$ Max Planck Institute for Psycholinguistics, Nijmegen, NL<br>${ }^{3}$ University of York, York, UK<br>${ }^{4}$ Donders Center for Brain, Cognition, and Behavior, Nijmegen, NL


#### Abstract

How are space and time related in the brain? This study contrasts two proposals that make different predictions about the interaction between spatial and temporal magnitudes. Whereas ATOM implies that space and time are symmetrically related, Metaphor Theory claims they are asymmetrically related. Here we investigated whether space and time activate the same neural structures in the inferior parietal cortex (IPC) and whether the activation is symmetric or asymmetric across domains. We measured participants' neural activity while they made temporal and spatial judgments on the same visual stimuli. The behavioral results replicated earlier observations of a space-time asymmetry: Temporal judgments were more strongly influenced by irrelevant spatial information than vice versa. The BOLD fMRI data indicated that space and time activated overlapping clusters in the IPC and that, consistent with Metaphor Theory, this activation was asymmetric: The shared region of IPC was activated more strongly during temporal judgments than during spatial judgments. We consider three possible interpretations of this neural asymmetry, based on 3 possible functions of IPC.


Keywords: ATOM, Metaphor, Space, Time, fMRI, Parietal lobe

## Introduction

It is clear that spatial and temporal magnitudes interact in the brain and mind, but the exact nature of this interaction is a matter of debate. According to one account, A Theory Of Magnitude (ATOM; Walsh, 2003), space, time and other prothetic domains (i.e. that can be experienced as more or less in magnitude) interact because they are represented by a common metric, located in the inferior parietal cortex (IPC; Walsh 2003; Bueti \& Walsh 2009). Support for this model comes from behavioral experiments showing cross-dimensional interference between different prothetic dimensions, as well as from neuroimaging studies showing that magnitude processing in various dimensions activates overlapping areas of IPC, mainly in the right hemisphere (see Bueti \& Walsh 2009 for review). According to ATOM, these different magnitudes share a representational substrate because they need to be integrated for successful execution of actions: Bueti and Walsh (2009) note that, "There is no such thing as getting to the right place at the wrong time" (pg. 1832). Like Locke (1689/1995) before them, ATOM theorists imply that space and time are symmetrically related. Indeed, if different prothetic domains are represented by the same metric, there is no a priori reason to assume that one domain should depend asymmetrically on another.

According to a second theoretical proposal, Metaphor Theory (MT), space and time are asymmetrically related: Temporal representations depend on spatial representations, more than vice versa. This asymmetry is fundamental to MT, which posits that representations of abstract concepts depend, in part, on representations of more concrete, perceptible domains (Lakoff \& Johnson, 1980; 1999). Since time is an abstract entity that we can never see or touch, it is argued to rely on spatial representations for conceptual scaffolding. Evidence for this asymmetric relationship comes from psychophysical studies showing asymmetric cross-dimensional interference between distance and duration: Task-irrelevant spatial magnitude influences temporal judgments more than task-irrelevant temporal magnitude influences spatial judgments (Bottini \& Casasanto, 2010; Casasanto \& Boroditsky, 2008; Casasanto, Fotakopoulou, \& Boroditsky, 2010; Merritt, Casasanto, \& Brannon, 2010; see also Boroditsky, 2000).

These cross-domain asymmetries were predicted by MT but not by ATOM. Yet, the available data leave open a possibility: Perhaps spatial and temporal magnitudes are encoded symmetrically. The observed asymmetry could arise subsequently, as magnitudes are re-represented during retrieval or response planning. If so, this finding would help to reconcile ATOM and MT, suggesting that initial stages of magnitude processing may be ATOMic even if later stages are metaphoric.

To test this proposal, we used fMRI to measure neural activity during spatial and temporal magnitude reproduction tasks. First, we compared activity during the encoding of spatial and temporal magnitudes to establish whether space and time interact at this stage, and whether they do so in the IPC. Second, by defining this area of overlap as a Region of Interest (ROI) and by comparing neural activity during the encoding of space and of time, we contrasted predictions of MT and ATOM. Both theories predict that encoding space and time should activate overlapping areas: On the basis of previous findings, we assume areas within IPC. MT predicts that this common area will be activated more by time than by space, because people involuntarily encode more irrelevant spatial information during temporal encoding than vice versa. ATOM does not predict any cross-domain asymmetry in the region of overlap.

## Methods

## Participants

18 healthy native English speakers (16 right-, 1 lefthanded, 1 ambidextrous, 9 male, mean age $=23.7$, range: 20-31) took part in the current experiment. All participants provided informed consent and were compensated for their participation.

## Materials

Lines of varying lengths were presented for varying durations. Durations ranged from 1000 ms to 4000 ms in 600 ms increments. Displacements ranged from 100 to 400 pixels in 60 pixel increments. The six durations were fully crossed with the six spatial displacements, producing 36 unique lines. Half of the lines were red and half were blue. Color was randomly assigned and counter-balanced across conditions and participants. Lines grew horizontally across the screen from left to right, along the vertical midline. The starting position of the lines was, on average, at 337 pixels from the left border of the screen, with the starting point randomly jittered ( $\pm 25$ pixels) so the monitor could not be used as a reliable reference frame for spatial estimations. Participants responded with a joystick (Current Designs, Philadelphia, USA; model: HHSC-JOY-1). Participants used their right hand to control the joystick for cursor movement and their left index finger for button responses.

## Procedure

Participants engaged in four different tasks: spatial reproduction, temporal reproduction and two color identification control tasks. Each trial started with a white cue that was presented for 1 second and indicated which dimension participants would need to reproduce for this trial (an " X " for space, an hourglass for time, and different colored squares for the two control conditions). This cue was followed by a single growing line that stayed on the screen until it reached its maximum spatial and temporal extent and then disappeared. After a period of 5 sec . ( $+/$ random jitter; range $=0-1$ sec.), a response cue and a cursor appeared until participants performed the required task or until a time-out period of 12 seconds had elapsed.

In the space condition (S), the X-icon appeared in either the upper- or lower-left corner of the screen (location counterbalanced across participants). To reproduce the distance that the line had traveled, participants moved the cursor from the center of the screen to the center of the icon, clicked once, moved the cursor rightwards in a straight line and then clicked a second time. The distance between the clicks represented the estimated displacement of the line. In the time condition (T), the hourglass-icon appeared in the lower- or upper-left corner of the screen (i.e. in the corner opposite the space cue). To reproduce duration, participants moved the cursor to the center of
the icon, clicked once, waited for the amount of time the line had been on the screen, and clicked a second time in the same spot. The time between the two clicks represented the duration of the line (procedure adapted from Casasanto \& Boroditsky, 2008).

Finally, two color identification conditions were included as controls: the "color half" (CH) and "color-full" (CF) conditions provided a low level visual control for the target lines, and also allowed us to subtract out activity due to motor preparation prior to responses. Each color condition required the same motor response as one of the target conditions. In the "color-half" condition (CH), participants saw two squares appear, each consisting of a red and a blue half. The left square was presented in the same corner in which the spatial response cue was presented for that participant and the second square was presented 250 pixels to the right of the left one. Participants first moved the cursor to the half of the left square that matched the color of the line, clicked once, moved the cursor rightwards in a straight line to the half of the right square that matched the color of the line and clicked again. Both squares were identical within a given trial (e.g. 2 red-blue squares), but the order of the colored halves was counterbalanced across trials ( $50 \%$ red-blue; $50 \%$ blue-red). In the "color-full" condition (CF), participants saw a blue and a red square appear in the upper- and lower-left corners of the screen (square position was counterbalanced across trials) and clicked twice on the square that had the color of the presented line.

Before entering the scanner, participants read the instructions and performed 3 practice trials of each condition. While in the scanner, participants performed each of the four tasks for each of the 36 unique lines $(4 \times 36=$ 144 trials in total). Lines were presented randomly within condition, and the order of conditions varied pseudorandomly (maximum of 3 trials of the same condition in a row).

## fMRI Data Acquisition

fMRI data were acquired on a Siemens Avanto 1.5 T MRI system (Siemens, Erlangen, Germany) using a standard birdcage head-coil for RF transmission and signal reception. T2*-weighted BOLD-sensitive images were acquired using a gradient EPI sequence (Echo Time (TE) = 40 ms ; Repetition Time (TR) $=2.28 \mathrm{~s} ; 32$ axial slices in ascending order; voxel size $=3.3 \times 3.3 \times 3.0 \mathrm{~mm}^{3}$ ). For each subject we also acquired a T1-weighted high-resolution anatomical scan ( $\mathrm{TE}=2.95 \mathrm{~s}, \mathrm{TR}=2.25 \mathrm{~s}$, voxel size $=$ $1.0 \times 1.0 \times 1.0 \mathrm{~mm}^{3}, 176$ sagittal slices, field of view $=256$ ).

## fMRI Data Analysis

Functional data were preprocessed and analyzed with SPM8 (http://www.fil.ion.ucl.ac.uk/spm/). Preprocessing involved the removal of the first 5 volumes to allow for T1 equilibration effects. Images were spatially realigned
with rigid body registration, temporally realigned to the middle slice (slice 17), co-registered to each participant's structural scan, normalized to a standard EPI template in MNI space and resampled at an isotropic voxel size of 2 mm . To remove baseline-drifts and low frequency signal changes, a $1 / 128 \mathrm{~Hz}$ temporal high-pass filter was applied. The normalized images were then smoothed with an isotropic 8 mm FWHM Gaussian kernel.

The preprocessed data were analyzed on a subject-wise basis using an event-related approach. The time series were entered into a GLM with separate regressors for the encoding and response phase for each condition (respectively modeled at one second before stimulus offset and response onset for $\mathrm{S}, \mathrm{T}, \mathrm{CH}$ and CF ), which were then convolved with a canonical hemodynamic response function. Although response phase regressors were added to the model for completeness, they were not analyzed further and will not be discussed. Finally, nuisance regressors were added to account for disturbances caused by small head movements.

To examine neural activity specific to spatial and temporal encoding, we computed two contrast images for each participant individually ([S-CH] and [T-CF]). These were then entered into separate second level random effects analyses to compute the space- and time-specific activations on the group level. Each of these two analyses consisted of a one-sample t-test to reveal activations significantly different from zero across the contrast images from all participants. A double threshold was applied to protect against Type I errors: only voxels with a $p<.001$ (uncorrected) and a volume exceeding 41 voxels (328 $\mathrm{mm}^{3}$ ) were considered (volume sizes were defined by Monte Carlo Simulation, $\mathrm{p}<.001$, Slotnick, 2011).

To reveal the neural overlap between spatial and temporal encoding, we performed a conjunction analysis on the 2 second-level contrast images ([S-CH] $\cap$ [T-CF]). Based on our a priori hypothesis, bilateral clusters of IPC activity that emerged from the conjunction were extracted and defined as our ROI's. From these ROI's, we extracted separate contrast values for the [S-CH] and the [T-CF] contrasts for each subject, using the MarsBaR package (Brett, Anton, Valabregue, \& Poline, 2002; http://marsbar.sourceforge.net, v.042).

## Results

## Behavioral Results

First we tested whether spatial and temporal reproduction was affected by variation in the task-irrelevant stimulus dimension. The spatial and temporal extents of the stimuli and responses were normalized, so that slopes could be compared across domains ${ }^{1}$. We calculated the normalized

[^73]slopes of the effect of irrelevant spatial information on duration reproduction (ST) and the effect of irrelevant temporal information on spatial distance reproduction (TS) for each participant separately (Fig. 1). The results showed significant cross-dimensional interference effects: The spatial extent of stimuli predicted the variation in the temporal responses $\left(\right.$ Wald $\left.\mathrm{X}^{2}(1)=23.55, p=.001\right)$ and the duration of stimuli predicted the variation in spatial responses $\left(\right.$ Wald $\left.\mathrm{X}^{2}(1)=12.21, p=.001\right)$. Importantly, these effects were asymmetric: Spatial information affected duration reproduction more than temporal information affected distance reproduction (Wald $\mathrm{X}^{2}(1)=8.00, p=$ .01).


Figure 1: Cross-domain interference effects. TS (Blue): Effect of line duration on spatial distance reproduction. ST (Red): Effect of line displacement on duration reproduction. Error bars indicate SEM.

To investigate within-domain performance, we used the normalized stimulus and response values to calculate the normalized slopes of the effect of spatial variation on spatial reproduction (SS) and temporal variation on temporal reproduction (TT), for each participant separately. Although the results show strong effects both of actual space on estimated space (SS: Normalized slope $=0.98$; Wald $\left.\mathrm{X}^{2}(1)=10139, p=.001\right)$ and of actual time on estimated time (TT: Normalized slope $=0.88$ Wald $\mathrm{X}^{2}(1)=$ $1465, p=.001)$, they also show a significant difference between within-dimension effects: participants were significantly better at spatial than at temporal reproduction (Difference of normalized slopes $=0.10$ Wald $\mathrm{X}^{2}(1)=$ 20.41, $p=.001$ ).

This difference in within-domain performance between space and time is potentially problematic for the interpretation of the between-domain asymmetry. If performance in one domain is nearly perfect, estimates in this domain may be less susceptible to interference than estimates in
the other domain (Bottini \& Casasanto, 2010). To rule out this concern, we re-ran the regression model after equating for within-dimension performance. Following Casasanto, et al. (2010), we excluded the data from participants with low TT-slopes until the within-dimension performance was the same for Space and Time (i.e. until SS $=\mathrm{TT} ; \mathrm{N}=7$ ). Even after equating within-dimension performance, a strong space-time asymmetry persisted: the spatial extent of stimuli predicted the variation in the temporal responses more than vice versa (Wald $X^{2}(1)=$ 12.56, $p=.001$ ), as in previous experiments (e.g., Casasanto \& Boroditsky, 2008; Merritt et al., 2010).

## Imaging Results

Patterns of Activation As the ROI analysis was our main point of focus, we provide only a cursory overview of the whole-brain results. During spatial encoding ([S-CH]), we observed bilateral activations in parietal areas (including the IPC, intraparietal sulcus (IPS) \& superior parietal gyrus), extrastriate visual cortex (extending into the inferior temporal gyrus) and frontal areas (precentral gyri, IFG, anterior insulae (AI), middle frontal gyri, SMA \& medial superior frontal gyri). During temporal encoding ([TCF]), we found activations in the parietal cortex (bilateral IPC, IPS, \& supramarginal gyri; right angular gyrus), left dorsal extrastriate cortex and a range of bilateral frontal activations (precentral gyri, AI, DLPFC, SMA, anterior cingulate and medial superior frontal gyri). Additional activity was observed in bilateral superior temporal gyri, right inferior temporal gyrus and subcortical areas (thalamus and basal ganglia).

To reveal neural activations common to spatial and temporal encoding, we performed a conjunction analysis on both aforementioned contrasts ([S-CH] $\cap$ [T-CF]; Fig. 2, left). Overlapping activations were found in bilateral parietal cortex, ranging from lateral IPC into the IPS and including part of the supramarginal gyrus in the right hemisphere. Additional clusters of activity included the left extrastriate cortex, just anterior to the cuneus, and a posterior part of the right inferior temporal gyrus. Frontal activations included bilateral DLPFC and parts of both precentral gyri (extending ventrally into the IFG and AI). Furthermore, we observed right-lateralized dorsal activation of the posterior middle and superior frontal gyri, extending into the superior frontal sulcus. Finally, the conjunction revealed medial frontal activations comprising the SMA, the medial superior frontal gyri and the middle and anterior cingulate.

ROI Analysis To investigate whether space and time encoding activated the bilateral IPC clusters revealed by the conjunction analysis symmetrically or asymmetrically, we defined the left and right IPC clusters as our two ROI's and extracted the contrast values per subject for each con-
dition of interest. These contrast values were entered into a regression model, with Condition (Space; Time), Hemisphere (Left; Right) and their interaction (Condition*Hemisphere) as within-subject factors and Subject as a repeated random effect. The IPC was activated more strongly by temporal encoding than by spatial encoding (main effect of Condition: Wald $\mathrm{X}^{2}(1)=4.65, \mathrm{p}=.03$ ). Furthermore, both domains activated the right IPC more than the left IPC (main effect of Hemisphere: Wald $\mathrm{X}^{2}(1)$ $=6.79, \mathrm{p}=.01)$, but the relationship between spatial and temporal activation did not differ between hemispheres (Condition*Hemisphere interaction: Wald $\mathrm{X}^{2}(1)=.073$, p $=.79$ ). Since previous studies have tended to implicate the right IPC in magnitude processing, we analyzed the ROI in each hemisphere separately. The main effect of Condition was significant in the right hemisphere ( $\mathrm{T}>\mathrm{S}$, Wald $\mathrm{X}^{2}(1)=5.14, \mathrm{p}=.02$, and marginally significant in the left-hemisphere $\left(\mathrm{T}>\mathrm{S}\right.$, Wald $\left.\mathrm{X}^{2}(1)=3.49, \mathrm{p}=.06\right)$ (Fig. 2, right).


Figure 2: Left: IPC activated by both space and time. Yellow: areas activated by spatial encoding ([S-CH]); Red: areas activated by temporal encoding ([T-CF]); Orange: areas activated by both ([S-CH] $\cap$ [T-CF]).
Right: Contrast values for space and time for the left and right IPC clusters.

Although these findings demonstrate that the IPC was differentially activated by spatial and temporal encoding, we must consider a skeptical account of this neural asymmetry. Our behavioral results indicated that, for some participants, temporal encoding was less accurate than spatial encoding, and may therefore have been more effortful. In principle, a difference in cognitive effort could be responsible for the observed cross-domain asymmetry in the IPC. If the observed asymmetry were due to more effort during temporal vs. spatial encoding, then adding a measure that reflects this difference in effort to the regression model as a covariate should reduce or eliminate the main effect of Condition.

To address this possibility, we calculated the difference between the normalized slopes of within-dimension performance in the space and the time condition (SS-TT) for each participant and included this difference score in the model. Even when cognitive effort was controlled for, we observed the same robust cross-domain asymmetry effects. Both IPC clusters were still activated more by tem-
poral than by spatial encoding (main effect of Condition: Wald $\mathrm{X}^{2}(1)=4.19, p=.04$ ), with stronger overall right hemisphere activation (main effect of Hemisphere: Wald $\mathrm{X}^{2}(1)=4.24, p=.04$ ). The main effect of this behavioral measure of cognitive effort, however, was not significant (Wald $\mathrm{X}^{2}(1)=2.42, p=.12$ ), nor was its three-way interaction with Dimension and Hemisphere (Wald $\mathrm{X}^{2}(1)=$ $1.60, p=.66$ ). These analyses show that the cross-domain asymmetry in BOLD activity in the IPC ROIs cannot be attributed to the observed differences in within-domain performance (i.e., to the finding that temporal estimates were less accurate, and potentially more effortful, than spatial estimates).

## Discussion

This study investigated whether representations of space and time interact at encoding and, if so, whether their relationship is symmetric or asymmetric. Consistent with both ATOM and MT, we observed that encoding spatial and temporal magnitudes activated overlapping clusters of a widespread neural network, most notably in the bilateral IPC. Of primary interest, our behavioral and ROI data provide converging support for an asymmetric relationship between these two domains, as predicted by MT. The behavioral findings indicated that when people reproduced duration they incorporated task-irrelevant spatial information, more so than they incorporated taskirrelevant temporal information when reproducing spatial extent. Our fMRI results showed that this behavioral asymmetry corresponded to a neural asymmetry: the IPC was more active during temporal encoding than during spatial encoding. The asymmetry between space and time is already present during encoding of spatial and temporal stimuli.

Further interpretation of the asymmetric IPC activation requires addressing the question of what exactly is being represented in the IPC. On one possibility, the IPC is the locus of a domain-general magnitude metric that accumulates undifferentiated bits of information (Bueti \& Walsh, 2009, Walsh, 2003), in any prothetic domain. In our task, the activation of the IPC might reflect the degree to which this metric accumulates bits from both domains simultaneously. Our behavioral results indicated that taskirrelevant spatial information was being encoded during time trials, more than task-irrelevant temporal information was being encoded during space trials. Hence, the IPC metric would have accumulated more task-irrelevant magnitude information (along with the task-relevant magnitude information) during time trials than during space trials, resulting in the increased BOLD signal. This account can potentially reconcile ATOM with MT: It is compatible with ATOM's claim of an IPC-based general magnitude representation, and is also consistent with MT, as the asymmetric IPC activation indicates an asymmetric interaction between space and time.

A second possibility is that magnitudes from different domains are represented independently in different parts of the brain (Cohen Kadosh et al. 2008), and that the IPC hosts a mechanism by which cross-domain magnitude representations are selected and integrated according to contextual demands.

A range of empirical data supports this interpretation. Several studies have shown that, during magnitude judgments, the activation of parietal areas around the IPS is modulated by the degree of interference from irrelevant dimensions (Ansari, Fugelsang, Dhital, \& Venkatraman, 2006; Cohen Kadosh, Cohen Kadosh \& Henik, 2008; Pinel, Piazza, Le Bihan, \& Dehaene, 2004; Cohen Kadosh, Cohen Kadosh, Linden, et al., 2007; Kaufmann et al., 2005). Moreover, studies of magnitude-irrelevant visual processing find that the IPC is activated by the need to suppress task-irrelevant distractors (Friedman-Hill, Robertson, Desimone, \& Ungerleider, 2003; Marois, Chun, \& Gore, 2000; Wojciulik \& Kanwisher, 1999). These processes are mostly right-lateralized (Chun \& Marois, 2002; Marois, Chun, \& Gore, 2000; 2004) and they are independent from task difficulty (Marois et al., 2000; Wojciulik \& Kanwisher, 1999).

Our task required participants to selectively attend to the relevant dimension of the stimuli and filter out the irrelevant dimension (i.e., space or time), which varied orthogonally. The greater IPC activation we observed during temporal encoding could reflect the increased demands posed on the 'magnitude selector/integrator' to filter out the task-irrelevant spatial information. Whereas this interpretation is consistent with MT, it contradicts one of the core claims of ATOM by positing that spatial and temporal magnitude representations are distinct: What is in common is the process of selecting and integrating relevant magnitude information.

Finally, on a third possibility, encoding space and time activated nearby but separate neural populations in the IPC, but the low spatial resolution of fMRI does not allow us to separate them (see Shuman \& Kanwisher 2004 for similar arguments). On this view, spatial encoding would have mainly activated the spatial representations, whereas temporal encoding would have activated both temporal and spatial representations. Not only could this account for the BOLD asymmetry (the combined activation of the two separate neural populations in same voxels during time processing leads to a higher BOLD signal), it could also explain the behavioral pattern. If temporal encoding activates neural populations that code for space more than vice versa, due to the "source domain-target domain" link posited by metaphor theorists, there should be more opportunity for crosstalk during time encoding than during space encoding. This notion of separate but closely interacting neural representations of space and time is consistent with MT, but argues against ATOM's claim of a shared representational basis of all prothetic magnitudes.

In summary, here we show a neural asymmetry between space and time that underlies the behavioral asymmetry found here and in multiple previous studies of distance and duration estimation. We consider three possible interpretations of these behavioral and neural asymmetries, all of which are consistent with MT, but only the first of which is compatible with ATOM. If the first account is correct and the IPC is the locus of a domaingeneral magnitude metric, ATOM and MT can be reconciled and the apparent contradiction in behavioral data resolved. The two other proposals are only consistent with MT and argue directly against ATOM's main claim of a shared IPC-based magnitude metric. Rather, they suggest that space and time are represented by distinct but closely interacting neural structures, either in the IPC, or in the form of a broadly distributed network. Further studies are needed to decide among these possibilities and clarify the role of the IPC in representing or integrating magnitudes.

## Acknowledgements

This research supported in part by a James S. McDonnell Foundation Scholar Award (\#220020236) to DC.

## References

Ansari, D., Fugelsang, J., Dhital, B., \& Venkatraman, V. (2006). Dissociating response conflict from numerical magnitude processing in the brain: an event-related fMRI study. NeuroImage, 32(2), 799-805.
Boroditsky, L. (2000). Metaphoric structuring: Understanding time through spatial metaphors. Cognition, 75(1), 1-28.
Bottini, R. \& Casasanto, D. (2010). Implicit spatial length modulates time estimates, but not vice versa. In S. Ohlsson \& R. Catrambone (Eds.), Proceedings of the 32nd Annual Conference of the Cognitive Science Society (pp. 13481353). Austin, TX: Cognitive Science Society.

Brett, M., Anton, J., Valabregue, R., \& Poline, J. (2002).
Region of interest analysis using an SPM toolbox. 8th
International Conference on Functional Mapping of the Human Brain. Sendai, Japan.
Bueti, D., \& Walsh, V.(2009). The parietal cortex and the representation of time, space, number and other magnitudes. Phil. Trans. R. Soc. B.,364(1525),1831-40.
Casasanto, D., \& Boroditsky, L. (2008). Time in the mind: using space to think about time. Cognition, 106(2), 57993.

Casasanto, D., Fotakopoulou, O., \& Boroditsky, L. (2010). Space and Time in the Child's Mind: Evidence for a Cross-Dimensional Asymmetry. Cognitive Science, 34(3), 387-405.
Chun, M. M., \& Marois, R. (2002). The dark side of visual attention. Curr. Opinion in Neurobiology, 12, 184-189.
Cohen Kadosh, R., Cohen Kadosh, K., \& Henik, A. (2008). When brightness counts: the neuronal correlate of numerical-luminance interference. Cerebral Cortex, 18(2), 337-43.
Cohen Kadosh, R., Cohen Kadosh, K., Linden, D., Gevers, W., Berger, A., \& Henik, A. (2007). The brain locus of
interaction between number and size: a combined functional magnetic resonance imaging and event-related potential study. J. Cog. Neurosci., 19(6), 957-70.
Cohen Kadosh, R., Cohen Kadosh, K., Schuhmann, T., Kaas, A., Goebel, R., Henik, A., \& Sack, A. (2007). Virtual dyscalculia induced by parietal-lobe TMS impairs automatic magnitude processing. Current Biology, 17(8), 689-93.
Cohen Kadosh, R., Henik, A., Rubinsten, O., Mohr, H., Dori, H., Van de Ven, V., Zorzi, M., et al. (2005). Are numbers special? The comparison systems of the human brain investigated by fMRI. Neuropsychologia, 43(9), 1238-48.
Cohen Kadosh, R., Lammertyn, J., \& Izard, V. (2008). Are numbers special? An overview of chronometric, neuroimaging, developmental and comparative studies of magnitude representation. Progress in neurobiology, 84(2), 132-47.
Dormal, V., Andres, M., \& Pesenti, M. (2008). Dissociation of numerosity and duration processing in the left intraparietal sulcus: a transcranial magnetic stimulation study. Cortex, 44(4), 462-9.
Fias, W., Lammertyn, J., Reynvoet, B., Dupont, P., \& Orban, G. A. (2003). Parietal representation of symbolic and nonsymbolic magnitude. J. Cog. Neurosci., 15(1), 47-56.
Friedman-Hill, S., Robertson, L., Desimone, R., \& L. Ungerleider (2003). Posterior parietal cortex and the filtering of distractors. PNAS, 100(7), 4263-8.
Kaufmann, L., Koppelstaetter, F., Delazer, M., Siedentopf, C., Rhomberg, P., Golaszewski, S., Felber, S., et al. (2005). Neural correlates of distance and congruity effects in a numerical Stroop task: an event-related fMRI study. NeuroImage, 25(3), 888-98.
Lakoff, G., \& Johnson, M. (1980). Conceptual Metaphor in Everyday Language. Journal of Philosophy, 77 (8), 453486.

Lakoff, G., \& Johnson, M. (1999). Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Tought. Chicago: University of Chicago Press.
Locke, J. (1689/1995). An essay concerning human understanding. Amherst: Promethius Books.
Marois, R, Chun, M., \& Gore, J. (2000). Neural correlates of the attentional blink. Neuron, 28(1), 299-308.
Marois, R., Chun, M., \& Gore, J. (2004). A common parietofrontal network is recruited under both low visibility and high perceptual interference conditions. Journal of Neurophysiology, 92(5), 2985-92.
Merritt, D.J., Casasanto, D., \& Brannon, E.M. (2010). Do Monkeys Think in Metaphors? Representations of Space and Time in Monkeys and Humans. Cognition, 117, 191202.

Shuman, M., \& Kanwisher, N. (2004). Numerical magnitude in the human parietal lobe; tests of representational generality and domain specificity. Neuron, 44(3), 557-69.
Slotnick, S. (2011). Cluster_threshold (beta) [Software]. Available from http://www2.bc.edu/sd-slotnick
Walsh, V. (2003). A theory of magnitude: common cortical metrics of time, space and quantity. TiCS, 7(11), 483-488.
Wojciulik, E. \& Kanwisher, N. (1999). The generality of parietal involvement in visual attention. Neuron, 23(4), 747-64.

# Modeling Dynamic Control in Normal Aging 

Brian D. Glass (b.glass@qmul.ac.uk)<br>Biological and Experimental Psychology Centre, School of Biological and Chemical Sciences, Queen Mary College, University of London, Mile End, London E1 4NS, UK.<br>Magda Osman (m.osman@qmul.ac.uk)<br>Biological and Experimental Psychology Centre, School of Biological and Chemical Sciences, Queen Mary College, University of London, Mile End, London E1 4NS, UK.


#### Abstract

Complex and dynamic decision making environments are common throughout life, but little is known about how normal aging influences performance on these types of scenarios. To determine performance differences associated with normal aging, we test older and younger adults in a dynamic control task. The task involves the control of a single output variable via multiple and uncertain input controls. A computational model is developed to determine the behavioral characteristics associated with normal aging in a dynamic control task. Older adults exhibit a positivity effect, congruent with previous research. Model based analysis demonstrates a unique performance signature profile associated with normal aging.


Keywords: dynamic decision making; learning; normal aging; computational modeling

## Introduction

Normal human aging is associated with cognitive changes that lead to differences in the way older adults approach and perform in decision making tasks. Specifically, older adults appear to suffer from executive control deficits (Braver, et al., 2001; Kray, Li, \& Lindenberger, 2002; Ortega, et al., 2012). However, emerging evidence suggests that older adults can utilize compensatory strategies to return performance to or beyond baseline levels (Glass, et al., 2012; Huang, et al., 2012; Worthy, et al., 2011).

While previous research has focused on classic paradigms such as category learning, task switching, and singleresponse choice procedures, little is known about normal aging in dynamic control tasks for which the participant controls multiple input variables in an integrative and uncertain task environment. Such complex dynamic environments are analogous to many real-life situations. For example, we make several distinct health choices on a daily basis which influence our overall health and wellbeing in uncertain ways. These types of environments are often noisy and the specific influence of the various choices is often unclear or unspecified.

The present research contrasts older adult and younger adult performance in a dynamic control task designed to simulate such real-life dynamic decision making environments (Osman \& Speekenbrink, 2011). A novel computational modeling technique is developed to assess individual
behavioral characteristics and strategies in the dynamic control task.

## Method

## Procedure

In the present dynamic control task, the participant attempts to control a single outcome value towards a goal. To do so, on each trial the participant chooses values for three separate cues. These cue values are then combined via the dynamic control equation (Equation 1) then summed with the outcome value plus some normally distributed random noise (standard deviation $=8$ ). In this way, the participant's cue selections guide the outcome value. The outcome value is initialized at 178 with a goal value of 62 and a "safe range" ( $\pm 10$ around the goal value)

$$
y(t)=y(t-1)+0.65 x_{1}-0.65 x_{2}+e
$$

Equation 1.
where $y(t)$ is the outcome on trial $t, x_{1}$ is the positive cue, $x_{2}$ is the negative cue, and $e$ is an error term randomly sampled from a normal distribution with a mean of 0 and SD of 8 .

The dynamic control equation was designed such that one cue has a positive impact on the outcome value, one cue has a negative impact, and a third cue has no impact. The impact of the cue is not labeled or available to the participant, thus the participant must learn to control the outcome value based solely on the resulting movement of the outcome value on each trial. After each trial, the cue input values are reset to 0 . The participant can then freely select input values for each of the three cues before confirming the choices.

A critical feature of this control task is that the outcome value can swing below the target, meaning the participant must dynamically adapt in order to maximize performance. After an initial learning phase, the participants completed 2 Test blocks of 20 trials each. The first Test block was a Congruent Test in which the starting value and goal criterion were equivalent to the learning phase. The second Test block was a Transfer Test with a different starting value and goal value than the earlier phases. At the beginning of each block, the control task was reset to the initial state.

## Participants

27 younger participants aged 18 to 25 ( $\mathrm{M}=22.3$, $\mathrm{SD}=5.4$ ) and 15 older participants aged 61 to $75(\mathrm{M}=67.92, \mathrm{SD}=$ 5.03) participated in the dynamic control task. The younger participants were recruited from the Queen Mary, University of London undergraduate community and paid $£ 6$ (\$9.50). The older participants were recruited via the National Hospital for Neurology and Neurosurgery. The older adults were recruited as a healthy control group via the National Hospital for Neurology and Neurosurgery. To qualify for the healthy adult participation pool, the older adults completed the. Beck Depression Inventory-II (BDI-II; Beck, et al., 1996) and Mini-Mental State Examination (MMSE) (Folstein, et al., 1975). All scores fell within the normal cutoff range for both the MMSE (greater than 27) and BDI-II (less than 18). None of the HCs had a history of neurological or physical or psychiatric illness, head injury or drug or alcohol abuse.

## Computational Model

A computational model of behavior in the dynamic control task was constructed to determine behavioral characteristics of individual participants. The model is based on memory trace reinforcement learning. After each trial, a reinforcement history for each of the three cues is updated according to whether the cue choices resulted in the discrepancy between achieved outcome value and goal value increasing or decreasing. On the following trial, the reinforcement history becomes the basis for a probabilistic action selection function using Luce's choice. Previous research has found that participants often vary the value of more than one cue on each trial. Thus, the model includes an inter-cue gating mechanism which allows each cue value selection to take into account the action selection probabilities of the other two cues.

The resulting model features four free parameters: an exploitation parameter governing the action selection function, an inter-cue gating parameter, and two memoryupdating reinforcement strengths (one for successful trials, and one for unsuccessful trials). To evaluate the model, the model's probability of selecting the human participant's cue choice are combined across all trials and all three cues into a single model fit value. The model is fit to an individual participant's responses by an optimization procedure that determines the parameter values which maximize the fit value.

## Memory-Updating Reinforcement Strengths

After each trial, the computational model determines whether the cue values it selected resulted in the outcome value moving towards or away from the goal. For each cue, a Gaussian curve with a mean equal to the chosen cue is constructed (Equation 2).

$$
P_{\text {update }}(v)=\frac{1}{\sigma \sqrt{2 \pi}} e^{-\frac{1}{2}\left(\frac{v-v_{p}}{\sigma}\right)^{2}}
$$

Equation 2.
where $P_{\text {update }}(v)$ is the probability of selecting a value of $v$ when the previous selected value was $v_{p}$.

This curve is then summed (successful trial) or subtracted (unsuccessful trial) to the cue's former reinforcement history. A free parameter (one for successful trials, one for unsuccessful trials) determines the relative weight of the updating summation. For example, if the memory-updating positive reinforcement strength is 0.8 , then the reinforcement history is updated such that $80 \%$ of the new reinforcement history reflects the current cue value choice and $20 \%$ reflects the previous reinforcement history (Equation 3).

$$
\begin{aligned}
& P_{\text {History }}(v)=
\end{aligned} \quad\left[\left(1-\gamma_{s}\right) P(v)\right] \text {. } \quad+\left[\operatorname{sign}(R) \cdot \gamma_{s} \cdot P_{\text {update }}(v)\right] \text {. }
$$

Equation 3.
where $P_{\text {History }}(v)$ is the cue selection probability history for cue value $v, \gamma_{s}$ is the memory-updating reinforcement strength for feedback $s$ (positive or negative), and $R$ is the change in the outcome value's distance to the goal from the previous trial.

In summary, there are two memory-updating reinforcement strengths, one for positive outcomes and one for negative outcomes. Each strength represents the weight with which current choices impact choice history (see Figure 1).


Figure 1: Sample probability density curves of selecting a given value for a given cue. Over the course of a block, the curves will alter in various ways depending on the model parameters, trial success, and uncertainty inherent in the outcome value.

## Inter-cue Parameter

Before the final probabilistic selection of the cue value occurs, for each of the three cues, the reinforcement history of the two other cues are taken into consideration. The level of this consideration is controlled by an inter-cue parameter. This parameter determines the strength at which the reinforcement history of other two cues will influence the action selection of the cue at hand. This is done using a gating equation which weights the alternate cues using the inter-cue parameter (Equation 4).

$$
\begin{aligned}
P_{\text {Intercue }}\left(v_{c_{A}}\right)= & {\left[\left(1-\frac{2 \beta}{3}\right) P_{\text {History }}\left(v_{c_{A}}\right)\right] } \\
& +\left[\frac{\beta}{3} \cdot P_{\text {History }}\left(v_{c_{B}}\right)\right] \\
& +\left[\frac{\beta}{3} \cdot P_{\text {History }}\left(v_{c_{C}}\right)\right]
\end{aligned}
$$

## Equation 4.

where $P_{\text {Intercue }}\left(v_{c A}\right)$ is the probability of selecting value $v$ for cue $c_{A}$ (e.g., cue 1 ), $\beta$ is the inter-cue parameter, and $c_{A}$ and $c_{B}$ are the other two cues (e.g., cue 2 and 3 ). As the inter-cue parameter approaches 1 , the computational model is more likely to pick similar cue values for all three cue inputs. As the inter-cue parameter approaches 0 , the model is less likely to select an action for one cue based on the reinforcement history of the other two. In this way, the computational model can vary the strength in which cue values vary together in the action selection state of the decision process.

## Exploration Parameter

On each trial, the computational model evaluates the reinforcement history of each cue to generate the probability of selecting each of the 100 cue value options. From these options, a single value is chosen using the Softmax decision rule (Equation 5). The equation's exploitation parameter, $K$, determines the level of determinism in the choice process (Daw \& Doya, 2006). As $K$ approaches $\infty$, the process is more likely to choose the most probable option. As $K$ approaches 0 , the equation is more likely to pick a less probable option.

$$
P_{\text {Final }}\left(v_{i}\right)=\frac{e^{\left[P_{\text {Intercue }}\left(v_{i}\right) \cdot K\right]}}{\sum_{j=0}^{100} e^{\left[P_{\text {Intercue }}\left(v_{j}\right) \cdot K\right]}}
$$

Equation 5.
where $P_{\text {Final }}\left(v_{i}\right)$ is the final probability of selecting cue value $v_{i}, K$ is the exploitation parameter, and $v_{j}$ are all the cue values from 0 to 100 for given cue.

## Results

## Task Analysis

By considering the optimal cue actions that will maximize the outcome value's movement toward the target, the optimal selections can be computed for each trial (Equation 5). The difference between the optimal selections and the actual chosen selections results in an optimality score for each participant. Figures 2 and 3 shows that the Younger group tended to select more optimal responses in both Test blocks, although the difference was not statistically significant.

## Optimality - Congruent Test



Figure 2: Optimality scores for Congruent Test block

## Optimality - Transfer Test



Figure 3: Optimality scores for Transfer Test block
At first blush, it may seem that the Older group performed similarly to the Younger group. However, further analysis of the strategies used by both groups demonstrates critical differences in the way the Older adults completed the dynamic control task. The strategy analysis considered four different types of cue changes: varying none, varying one cue, varying two cues, and varying all three cues. Figures 4
and 5 illustrate the cue varying strategies for both groups on both the Congruent Test and Transfer Test. A 2 (Older, Younger) x 2 (Congruent, Transfer) x 4 (Strategy Type) repeated measures ANOVA reveals an Age by Block by Strategy interaction, $F(3,120)=2.95, p<0.05, \eta=0.07$. There was also a main effect of strategy, $F(3,120)=24.42$, $p<0.001, \eta=0.38$. No other main effects of interactions were statistically significant.


Figure 4. Cue varying strategies for Congruent Test
Cue Strategies - Transfer Test
$\square$ OlderAdults YoungerAdults


Figure 5. Cue varying strategies for Transfer Test
Not only did the Younger and Older groups differ in their cue varying strategies, they also differed in the values selected for the cues. Figures 6 and 7 report the mean cue values (between 0 and 100) selected for each of the three Cue Types. A 2 (Older, Younger) x 2 (Congruent, Transfer) x 4 (Strategy Type) repeated measures ANOVA revealed a
main effect of Cue Type, $F(2,80)=5.11, p<0.01, \eta=$ 0.11 , as well as an interaction of Age and Cue Type, $F(2$, $80)=3.51, p<0.05, \eta=0.08$. This suggests that the Older group tended to select higher values for the Positive and Null cues

## Cue Values - Congruent Test

$\square$ OlderAdults $\quad$ Younger Adults


Figure 6. Cue values selected for Congruent Test


Figure 7. Cue values selected for Transfer Test
Taken together, analyses of surface level behavior suggest the Older group differed from the Younger group in completing the dynamic control task. However, the nature of the underlying cognitive processes which led to these patterns of behavior remains elusive using basic task analysis. In order to distill psychologically relevant characteristics of the processes involved in the dynamic decision making task performance, a computational reinforcement learning model of the dynamic control task was fit to individual participant data.

## Model Based Analysis

Task behavior was fit to the computational model using an optimization procedure that attempted to minimize the difference between observed trial-by-trial cue value selections and the expected cue value selections as determined by the model. This was done by considering the probabilities given to the various cue values for each cue on a given trial. The optimization procedure attempted to determine best fitting free parameters (exploitation parameter, inter-cue parameter, positive and negative reinforcement sensitivity parameter) that maximized the probability that the model would select the same cue values as the human participant on a given trial.


Figure 8. Exploitation Parameter

Inter-Cue Parameter


Figure 9. Inter-Cue Parameter


Pos. Sensitivity Parameter
OlderAdults YoungerAdults

Figure 10. Positive Sensitivity Learning Parameter


Figure 11. Negative Sensitivity Learning Parameter
Figures 8 through 11 reports the mean best fit parameter values for the Younger and Older groups. In the Congruent Test, those in the Older groups were best fit with a lower exploitation parameter $(t[40]=-2.37, p=0.02)$, a higher positive reinforcement strength parameter $(t[40]=3.17, p=$ 0.003 ), and a lower inter-cue parameter ( $t[40]=-2.35, p=$ 0.02 ). There was no significant difference in the negative reinforcement strength parameter between the two groups, $t(40)=-0.29, p=0.85$. In the Transfer Test, the Older adults continued to be better fit by a higher positive reinforcement parameter than Younger adults, $\mathrm{t}(40)=2.74, \mathrm{p}<0.01$. In short, in the Congruent Test, the Older group's performance was better fit with parameters associated with higher exploration, higher positive feedback sensitivity, and lower inter-cue selection. In the Transfer Test, the Older group
continued to be better fit by model parameters associated with higher positive feedback sensitivity.

## Discussion

The present study examined the role of normal aging in a dynamic control task using a novel computational modeling technique. Standard behavioral analysis revealed older adults potentially utilized an alternative strategy in completing the dynamic control task than younger adults. A computational model of the task revealed specific behavioral characteristics associated with normal aging. In the Congruent block, older adults demonstrated more exploratory behavior, less inter-cue behavior, and more reliance on recent and positive success. On the Transfer block, older adults did not differ from younger adults in their exploratory and inter-cue behavior, but continued to demonstrate more reliance on recent and positive success.

One possible interpretation of this pattern of results is that older adults were able to achieve the final performance profile of younger adults (as measured by deviation from optimal responses) by relying on compensatory mechanisms to engage the task. Specifically, in the congruent goal test, the older adults were more exploratory, relied less on the reinforcement history of alternative cues when determining cue values, and were more influenced by trials on which they received positive feedback. During the transfer goal test, the older adult's compensatory strategy gave way to a closer performance signature exhibited by younger adults. However, they remained more influenced by positive feedback. This interpretation is supported by previous research which has shown that older adults are able to achieve the performance levels of younger adults via a compensatory strategy (Glass, et al., 2012;Worthy \& Maddox, 2012).

Another interpretation of the current results is that older adults approached the task by utilizing alternative mechanisms which may be enhanced in older adults. For example, older adults exhibit a positivity effect characterized by superior emotional processing of positively valenced content (Carstensen \& Mikels, 2005). This could account for the older adults' higher learning rate sensitivity parameter for positive feedback, but not for negative feedback. Thus, when older adults encountered successful trials, their learning rate parameters increased such that prior knowledge was discounted. In this interpretation, older adults differed in their overall strategy due to specific enhancements associated with normal aging. This interpretation is supported by the positive learning rate sensitivity parameter remaining higher for older adults than younger adults in both the congruent and transfer tasks. Future research should determine whether the differences in strategies used by older adults to complete the dynamic control task are simply the result of slower overall learning rates, or due to differences in underlying cognitive mechanisms associated with normal aging. Future work
should incorporate manipulations to test these interpretations, such as limiting feedback types to determine whether the aging positivity effect can account for performance differences.

## Acknowledgments

The authors thank Agata Ryterska for help with data collection for the project.

## References

Beck, A. T., Steer, R. A., Ball, R., \& Ranieri, W. F. (1996). Comparison of Beck Depression Inventories-IA and-II in psychiatric outpatients. Journal of Personality Assessment, 67, 588-597.
Braver, T. S., Barch, D. M., Keys, B. A., Cohen, J. D., Taylor, S. F., Carter, C. S., Kaye, J. A., Janowsky, J. S., Yesavage, J. A., Mumenthaler, M. S., Jagust, W. J., \& Reed, B. R. (2001). Context processing in older adults: Evidence for a theory relating cognitive control to neurobiology in healthy aging. Journal of Experimental Psychology: General, 130, 746-763.
Carstensen, L. L. \& Mikels, J. A. (2005). At the intersection of emotion and cognition. Current Directions in Psychological Science, 14, 117-121.
Daw, N. D. \& Doya, K. (2006). The computational neurobiology of learning and reward. Current Opinion in Neurobiology, 16, 199-204.
Folstein, M. F., Folstein, S. E. \& McHugh, P. R. (1975). Mini-Mental State: A practical method for grading the cognitive state of patients for the clinician. Journal of Psychiatric Research, 12, 189-198.
Glass, B. D., Chotibut, T., Pacheco, J., Schnyer, D. M., \& Maddox, W. T. (2012). Normal aging and the dissociable prototype learning systems. Psychology and Aging, 27, 120-128.
Huang, C. M., Polk, T. A., Goh, J. O., Park, D. C. (2012). Both left and right posterior parietal activations contribute to compensatory processes in normal aging. Neuropsychologia, 50, 55-66.
Kray, J., Li, K. Z. H., \& Lindenberger, U. (2002). Agespecific changes in task-switching components: The role of task uncertainty. Brain and Cognition, 49, 363-381.
Ortega, A., Gómez-Ariza, C. J., Román, P. E., \& Bajo, M. T. (2012) Memory inhibition, aging and the executive deficit hypothesis. Journal of Experimental Psychology: Learning, Memory \& Cognition, 38, 178-186.
Osman, M., \& Speekenbrink, M. (2012). Prediction and control in a dynamic environment. Frontiers in Psychology, 3:68.
Worthy, D. A., Gorlick, M. A., Pacheco, J. L., Schnyer, D. M., \& Maddox, W. T. (2011). With age comes wisdom: Decision making in younger and older adults. Psychological Science, 11, 1375-1380.
Worthy, D.A. \& Maddox, W. T. (2012). Age-based differences in strategy-use in choice tasks. Frontiers in Neuroscience, 5, 1-10.

# Development of Category-Based Reasoning: Results from a Longitudinal Study 

Karrie E. Godwin (kegodwin@andrew.cmu.edu)<br>Anna V. Fisher, (fisher49@andrew.cmu.edu)<br>Bryan J. Matlen (bmatlen@cmu.edu)<br>Carnegie Mellon University, Department of Psychology, 5000 Forbes Avenue, Pittsburgh, PA 15213 USA


#### Abstract

Prior research on the development of category-based reasoning indicates a protracted developmental course of this ability as well as a high degree of individual variability. However, the sources of this individual variability as well as the sources of developmental change remain unclear. The present study aimed to examine these issues, with a focus on the role of representational change and executive function development. Across two time points spaced approximately 7 months apart, children's category-based reasoning was assessed along with a battery of executive function and representational change measures. Results replicated prior work in that only a small proportion of children exhibited spontaneous category-based reasoning at Time1, and this proportion increased with development. In addition, both executive function and representational change were found to predict the development of category-based reasoning.


Keywords: Category-based reasoning; inductive reasoning

## Introduction

Category-based reasoning is central to mature cognition and underlies much of our learning and functioning in the world (e.g., Osherson et al., 1990; Sloman, 1993). Despite early reports that even very young children spontaneously engage in category-based reasoning (e.g., Gelman \& Markman, 1986; Gelman \& Coley, 1990; Welder \& Graham, 2001), recent evidence suggests that development of category-based reasoning follows a relatively protracted developmental course (e.g., Badger \& Shapiro, 2012; Godwin, Matlen, \& Fisher, in press; Fisher, Matlen, \& Godwin, 2011; Fisher, 2010; Fisher \& Sloutsky, 2005).

One of the hallmarks of category-based reasoning is one's ability to make inferences based on the knowledge that two (or more) items belong to similar kinds in the absence of supporting perceptual information. For example, if one is shown a picture of a rock, a sponge, and another rock and asked to predict which two items have properties in common, one could rely on perceptual similarity to make an inference. Similarly, if the pictures are ambiguous (or not presented) and labels are used to indicate category membership, one could base their inference on matching labels (e.g., rock-rock), not necessarily because one understands that labels refer to kinds, but because the labels are perceptually identical (Sloutsky \& Fisher, 2004). However, one's ability to rely on semantically-similar labels (e.g., rock-stone) to make inferences is commonly interpreted as an index of category-based reasoning (e.g., Gelman \& Markman, 1986).

Several studies have documented that the ability to spontaneously engage in category-based reasoning appears between 4 and 6 years of age (e.g., Badger \& Shapiro, 2012;

Godwin, et al., in press; Fisher, et al., 2011; Fisher 2010; Fisher \& Sloutsky, 2005). However, it remains unclear what leads to the development of spontaneous category-based reasoning during the preschool years. Two classes of explanations have been put forth to explain changes in various areas of cognitive development, namely Representational Change and Executive Function development. We briefly discuss both explanations below.

## Representational Change

Representational change is "reorganization of existing knowledge or a difference in the utilization of information, rather than the acquisition of new information" (Nelson, 1977, p. 109). Representational change has been implicated as an explanatory factor in several areas of cognitive development, including analogical reasoning (e.g., Gentner et al., 1995), problem solving (e.g., Karmiloff-Smith, 1984), and numerical development (e.g., Opfer \& Siegler, 2007).

With regards to semantic development, there are several compelling sources of evidence pointing to representational change in the multidimensional scaling literature (e.g., Howard \& Howard, 1977), free association studies (e.g., Brown \& Berko, 1960), and development of semantic priming (e.g., McCauley, Weil, \& Sperber, 1976).

Furthermore, different approaches to modeling semantic cognition suggest that early conceptual organization is fairly undifferentiated (such that penguin, trout, and alligator may start out as belonging to the same cluster) with greater differentiation emerging with development (Kemp \& Tenenbaum 2008; Rogers \& McClelland 2004). At present, there is no direct empirical evidence testing these predictions, although Carey's (1985) seminal work is largely consistent with these developmental profiles.

## Executive Functions

Executive Functions (EF) are psychological processes thought to control other (typically, higher-order) psychological processes such as planning, reasoning, and problem-solving. Most researchers distinguish the following EF processes: set shifting, active maintenance of representations (sometimes referred to as working memory), and inhibitory control (Bunge et al., 2002; Carlson et al., 2002).

The EF system is traditionally associated with prefrontal cortex, which is believed to be one of the slowest brain regions to mature (e.g., Diamond, 2002). Development of EF has been implicated in developmental accounts of category learning (Sloutsky, 2010), and there is evidence that representation maintenance and inhibitory control play a role in the development of analogical reasoning (Morrison
et al., 2011). With regards to semantic cognition, regions of the PFC (specifically, ventrolateral PFC or VLPFC) have been shown to be engaged in controlled semantic access, for instance in classification or category generation tasks. While it remains unclear whether the role of VLPFC is to bias retrieval of task-relevant semantic information through maintaining task representations or to select task-relevant representations among competing activated representations (e.g., Wagner, 2002; Kan \& Thompson-Schill, 2004), prefrontal cortex clearly is important for controlled semantic access.

## The Present Study

Prior research on the development of category-based reasoning indicates not only a protracted developmental course of this ability, but also a high degree of individual variability. Specifically, results aggregated across several studies suggest that approximately $20 \%$ of 4 -year-olds spontaneously make category-based inferences with semantically-similar labels, and this proportion increases to approximately $40 \%$ and $65 \%$ among 5 - and 6 -year-olds, respectively (Fisher, 2010; Fisher et al., 2011; Godwin et al., in press). However, the sources of this individual variability as well as sources of developmental change remain unclear. The goal of the present study was to begin the exploration of these questions, with a focus on the putative role of representational change and executive functions. A battery of assessments was administered over the course of one school year. At Timel (Fall) we collected measures of children's category-based reasoning, verbal working memory, IQ, and semantic knowledge organization. At Time2 (Spring) we collected measures of children's category-based reasoning, semantic knowledge organization, inhibitory control, non-verbal working memory, sustained attention, and category generation.

## Method

## Participants

Participants in this study were 43 four-year-old children from a local preschool (Mage $=4.32$ years, $S D=0.28$ years, 20 females, 23 males).

## Design and Procedure

Each child participated in 13 sessions over the course of the school year ( 6 sessions at Timel and 7 sessions at Time2). Children were tested individually in a quiet room adjacent to their classroom by a trained research assistant. A brief description of the task battery is provided below.

## Category-Based Reasoning Task

This task included 9 label triads, 3 of which referred to artifacts, 3 to inanimate natural kinds, and 3 to animate natural kinds (see Table 1). All triads contained a target item, category-choice, and an unrelated lure (e.g., rat-mouse-fish). Visual stimuli consisted of sets of three identical doors. Children were told that objects were hiding behind doors. The objects were never revealed in order to
encourage children to rely on the category information conveyed by the labels.

Table 1: Category-Based Reasoning Task Linguistic Stimuli

| Target | Category Choice | Lure | Property |
| :---: | :---: | :---: | :---: |
| Rock | Stone | Grass | Higa |
| Alligator | Crocodile | Butterfly | Omat |
| Rug | Carpet | Window | Koski |
| Rat | Mouse | Fish | Lignin |
| Hill | Mountain | Flower | Erwin |
| Sea | Ocean | Apple | Manchin |
| Sofa | Couch | Cup | Creighan |
| Shoe | Boot | Car | Troxel |
| Lamb | Sheep | Frog | Matlen |

Children were first told what objects were hiding behind the doors and then told about a novel property of the target item (e.g., "The rock has higa inside"). Children were asked to generalize the novel property from the target item to either the category-choice or the unrelated lure. The task was administered four times (twice within each time point) in order to obtain a more stable estimate of children's performance. The delay between task administrations within a time point was one to two weeks. The trials were administered in one of two counter-balanced orders.

## Picture Identification Task

The picture identification task is similar in format to the Peabody Picture Vocabulary Test (Dunn \& Dunn, 1997). It assessed children's familiarity with the linguistic stimuli utilized in the category-based reasoning task. The task was administered at Time 1 and Time 2.

## Intelligence Test

The Wechsler Preschool and Primary Scale of Intelligence (WPPSI) provided a measure of children's general intelligence (Full-scale IQ or FSIQ), as well as an index of children's Verbal IQ (VIQ), Performance IQ (PIQ), and Processing Speed Quotient (PSQ). The WPPSI was administered at Time 1 only.

## Semantic Space Task

The semantic space task served as a measure of children's semantic organization. The stimuli entailed a game board ( 9 x 9 grid ; see Figure 1) and 2 game pieces (1" wooden cubes). Verbal stimuli included 24 animal pairs: 6 semantically-similar dyads (e.g., chick-hen), 6 dyads that share a common setting or habitat (e.g., chick-goat), 6 dyads that are unrelated (e.g., chick-goldfish), and 6 filler dyads (see Table 2). Thus, throughout the task, the target animal was paired with 3 different test items (i.e., the categorychoice, setting/habitat match, and the unrelated item). The 3 animal trials from the Category-Based reasoning task were included in the Semantic Space task.

Children were told that they were helping Zibbo the zookeeper organize his zoo. Children were instructed to put animals of the same kind close together on the board. For each trial, the experimenter put one of the game pieces on a predetermined square on the game board and told the child that the specified location was where Zibbo put the target animal (e.g., "The zookeeper put the chick here"). The child
was then asked to identify where the test item (i.e., the second game piece) should be placed (e.g., "Where do you think the goat should go?"). After each trial the child's response was recorded in order to calculate the distance between the target and the placement of the test item. Prior to playing the game, children were given two examples where the experimenter demonstrated that animals of similar kind (e.g., bunny and rabbit) should be placed close together and animals of different kind (e.g., dog and shark) should be placed far apart.

Table 2: List of Stimuli for the Semantic Space Task

| Critical Trials |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: |
| Target | Category <br> Choice | Setting/ <br> Habitat | Unrelated |  |
| Crocodile | Alligator | Fish | Grasshopper |  |
| Chick | Hen | Goat | Goldfish |  |
| Lamb | Sheep | Horse | Swan |  |
| Whale | Dolphin | Octopus | Elephant |  |
| Monkey | Gorilla | Parrot | Chipmunk |  |
| Mouse | Rat | Pig | Hippo |  |
| Filler Pairs |  |  |  |  |
| 1. Zebra/Turkey; 2. Bear/Snake; 3. Panther/ Turtle; |  |  |  |  |
| 4. Tiger/Butterfly; 5. Frog/Lion; 6. Giraffe/Seal |  |  |  |  |



Figure 1: Schematic depiction of the semantic space game board. Red squares mark the location of critical trials; yellow squares mark the location of filler trials. Note that none of the locations were colored when the task proper was administered.

Children's score on the semantic space task was calculated by averaging habitat dyads and unrelated dyads together in order to create a composite score for non-semantically-similar dyads. Next, the average score for semantically-similar dyads was subtracted from the non semantically-similar composite score to obtain a difference score. Difference scores approaching zero indicate that children did not differentiate the placement of semanticallysimilar and dissimilar dyads. Difference scores above zero indicate that children placed semantically-similar dyads closer than dissimilar dyads. The semantic space task was administered at Time1 and Time2.

## EF Measures

Verbal Working Memory Tasks Forward and Backward word-span tasks were administered to assess children's verbal working memory capacity. Verbal stimuli consisted of 60 common count nouns selected from the MacArthur Communicative Development Inventory (Dale \& Fenson,
1996). In the Forward-word span task, children were asked to recite the words in the same order in which they were presented; in the Backward-word span task children were asked to repeat the words in the opposite order. If a child made a mistake, they were given another opportunity to recite a different list of the same word length. Children's score was determined by the longest list length the child was able to recite correctly.
Inhibitory Control Measures Two common measures of response inhibition were included in the assessment battery: the Day-Night task (Gerstadt et al., 1994) and the Flanker task (Rueda et al., 2004). In the Day-Night Task, children were shown a set of cards depicting the sun and the moon. Children were asked to provide a verbal response that conflicts with the presented image (e.g., if the child was shown a picture of the sun, the correct response would be "night". Conversely if the child was shown a picture of the moon, the correct response would be "day"). The task consisted of 16 trials (the moon and the sun were presented 8 times each). Two presentation orders were created: The trials were randomized for order 1 and the sequence was reversed for order 2.
We used the version of the Flanker Task adapted for use with young children (Rueda et al., 2004). In this version children are presented with arrays of fish on a computer screen. Children are asked to feed the center fish by pressing either the left or right button. The correct response is dependent upon the direction the center fish is facing. The center fish is surrounded by four other fish (two on each side). The surrounding fish may be congruent (e.g., swimming in the same direction as the center fish) or incongruent (e.g., swimming in the opposite direction as the center fish). Neutral trials were also presented in which the central fish appears in isolation (i.e., not flanked by other fish). A total of 48 trials were administered: 16 neutral trials, 16 incongruent trials, and 16 congruent trials. For the purposes of the analyses reported below, we used the Flanker Accuracy Difference score (calculated by subtracting each child's accuracy for the Incongruent trials from the Neutral trials) and Flanker RT Difference score (calculated by subtracting each child's reaction time for the Incongruent trials from the Neutral trials).

## Non-Verbal Working Memory \& Sustained Attention

 The Track-It Task (Fisher et al., 2012) was used as an index of non-verbal working memory; this task also provided a measure of sustained attention. In this task children watched a set of moving objects: six distractors and one target. The objects moved randomly across a computer screen for 10 seconds, and then disappeared. On each trial, children were asked to select the location where the target object disappeared; the location questions provided a measure of sustained attention. Upon answering the location question, children were shown a laminated card that contained an array of 9 objects (the target object and 8 lures). Children were asked to point to the target object that they had been tracking; children's responses to this question provided ameasure of non-verbal working memory (WM). The Trackit task included 10 experimental trials and one practice trial.

## Results

Picture Identification The results from the Picture Identification task suggested that children were familiar with the labels used in the category-based reasoning task: Children's accuracy on this task approached ceiling levels ( $M=92 \%, S D=14 \%$ and $M=96 \%, S D=8 \%$ for Time 1 and 2 respectively). As an additional precaution, children's category-based reasoning scores were adjusted for their vocabulary knowledge to ensure that children possessed the pre-requisite knowledge to perform category-based induction. Thus, if a child missed an item on the picture identification task, this trial was removed from their category-based reasoning score.

Category-Based Reasoning Task Mean category-based reasoning scores at Time 1 a and 1 b were very similar (adjusted means: 0.63 and 0.66 , respectively) and significantly correlated ( $r=.483, p=0.001$ ). Mean categorybased reasoning scores at Time 2 a and 2 b were also similar (adjusted means: 0.73 and 0.80 , respectively) and significantly correlated ( $r=.689, p=0.0001$ ). Consequently, induction scores were averaged across Time 1 a and 1 b and across Time 2 a and 2 b to yield average category-based reasoning scores for Time1 ( $M=0.64, S D=0.22$ ) and Time2 ( $M=0.76, S D=0.21$ ).

The rate of category-based responding at Time1 ( $M=.64$ ) was above chance $(t(40)=4.08, p<.001)$ and somewhat higher than in our prior studies ( $M=.54$ across Fisher et al., 2011; Godwin et al., in press; Matlen et al., under review). However, it should be noted that in the present study the sample consisted entirely of children enrolled in a laboratory campus school at a private university, and prior studies utilized more diverse community-based samples.

The proportion of category-based responding at Time2 ( $M=.76$ ) was also above chance, $(t(41)=8.01, p<.001)$ and higher than at Time1, paired-samples $t(39)=3.53, p<.001$, Cohen's $d=.56$. Note that the latter finding cannot be attributed simply to children having experience with performing the same task, as scores were not significantly different at Time 1 a (.63) and 1 b (.66). Therefore, this finding points to a developmental increase in the propensity towards category-based reasoning.

To investigate individual patterns of responses, participants were classified as either category-based or non-category-based responders. A category-based responder was defined as a participant who gave a category-based response on at least 7 out of 9 (78\%) trials (binomial $p=0.09$ ). Analysis of the individual patterns revealed that only a small percentage of children were classified as category-based responders at Timel (27\%). In contrast, the majority of children were classified as category-based responders at Time 2 ( $67 \%$ ). The association between responder type and testing point (Time1 vs. Time2) was significant, McNemar's $\chi^{2}(1)=7.22, p<.005$.

Predicting Category-Based Reasoning We performed linear stepwise regression to identify the best predictors of category-based responding at Time1 and 2. Only predictors that were significantly correlated with category-based responding were entered into the model. Thus, three predictors were included: Semantic Space scores Timel, FSIQ, and Non-Verbal WM score. Overall, the model significantly predicted children's responses on the categorybased reasoning task, $R^{2}=.211, F(1)=11.42, \quad p=.002$. However, only one predictor was found to be significantly related to children's induction performance at Time 1 : Semantic Space scores $(\beta=.481, t(1)=3.38, p=.002)$.

For predicting category-based responding at Time2, only predictors that were significantly correlated with induction performance at Time2 were included in the model. The following predictors were entered into the model: Semantic Space Time 2 scores, Forward Word-Span, Backward WordSpan, FSIQ, Non-Verbal WM score, Sustained Attention score, and Day/Night score (VIQ was excluded from the analysis due to concerns regarding colinearity based on its high correlation with FSIQ). Overall, the two-predictor model significantly predicted children's responses on the category-based reasoning task, $R^{2}=.474, \quad F(2)=19.00$, $p<.001$. However, only two predictors were found to be significantly related to children's induction performance at Time 2: Non-verbal WM ( $\beta=.522, t(2)=4.37, p<.0001)$ and Day/Night scores $(\beta=.352, t(2)=2.95, p=.005)$.

Category-Based Reasoning: What Develops? What factors play a role in the development of category-based reasoning? Since several children performed at nearly ceiling level on the category-based reasoning task at Time 1, it was not possible to address this question using gain scores from Time1 to Time2. Therefore, to address this question we split the sample into three groups based on the children's performance on the category-based reasoning task at Time 1 and Time2 (see Figure 2). Group 1 included children who were already category-based responders at Time1 ( $27 \%$ of the sample); all of these children remained category-based responders at Time2. Group 2 included children who were not yet category-based responders at Timel but became category-based responders at Time2 ( $40 \%$ of the sample). Group 3 included children who were not yet category-based responders at either Time1 or Time2 (32.5\% of the sample). Splitting the sample in this manner allowed for analyses examining potential factors that may differentiate Groups 2 and 3 (i.e., children who became category-based responders at Time2 from children who were not yet category-based). Three children were missing scores for either Time 1 or 2 and were omitted from this analysis.

Importantly, children in Groups 2 and 3 obtained comparable FSIQ scores ( $M=108, M=100$ respectively), $t(27)=1.41, n s$. This finding suggests that performance differences between the two groups on the category-based reasoning task were not simply a result of disparities in children's general intelligence. Children in Group 1 obtained FSIQ scores $(M=118)$ that were over one standard deviation above the population mean $(M=100, S D=15$;

Wechsler, 2002). FSIQ scores of children in Group 1 were also significantly higher than children in Group 3, $t(21)=3.04, p<.01$, and marginally higher than those of children in Group 2, $t(25)=1.78, p=.088$. Based on IQ scores, the group of children who were already categorybased responders on the induction task at Time1, were cognitively advanced. However, it is important to note that children in Groups 2 and 3 were not lagging behind as they exhibited average intelligence compared to the general population.


Figure 2. Proportion of category-based responses at Time 1 and Time 2 by group.

Due to space limitations we cannot describe the performance patterns for these three groups on all administered tasks. However, it should be noted that on some measures children's performance was equivalent across groups. For example, children's Flanker Accuracy Difference scores were comparable in all three groups (MGroup1=.35, MGroup2=.38, MGroup3=.30, all $t \mathrm{~s}<.75, n s$ ). Similarly, Flanker RT Difference scores were comparable in all three groups (MGroup1=-127.37, MGroup2=-116.24, $M$ Group $3=-106.48$, all $t \mathbf{s}<.21, n s)$. The biggest performance differences were found on three measures: Semantic Space, Non-Verbal WM, and the Day/Night task.

Recall that the Semantic Space task was administered twice, once at Time1 and again at Time2. Therefore, we were able to compare children's performance on this task across time. As can be seen in Figure 3, children in Group 1 exhibited equivalently high performance on the Semantic Space task at both Time1 and Time2 ( $M \mathrm{~T} 1=2.41, M \mathrm{~T} 2=3.04$, paired-sample $t(10)=.73, n s)$. In contrast, children in Group 2 significantly improved in their performance on the Semantic Space task from Time1 to Time2 ( $M \mathrm{~T} 1=1.4$, $M T 2=3.01$, paired-sample $t(15)=2.38, p=.03$ ). Children in Group 3 exhibited relatively low performance on the Semantic Space task at both Time1 and Time2 ( $M$ T1 $=.88, M$ $\mathrm{T} 2=1.15, t(12)=.59, n s)$.

Overall, these findings suggest that children who showed consistently high performance on the Semantic Space task also showed consistently high performance on the categorybased reasoning task, and children who showed consistently low performance on the Semantic Space task also showed consistently low performance on the category-based reasoning task. Only those children who showed improved performance on the Semantic Space task also showed improved category-based reasoning.


Figure 3: Mean performance on the Semantic Space task at Time 1 and Time 2 by group.

A stark difference was observed in children's non-verbal WM performance. Children in Groups 1 and 2 demonstrated similar levels of performance on the non-verbal WM task ( $M=.70$ and $M=.62$ respectively; $t(25)=.67, n s$ ) with children in both groups demonstrating better non-verbal WM than children in Group 3 ( $M=.18$ ); all $t \mathrm{~s}>4.66, p<.0001$ ). A similar pattern of results was obtained for the Day/Night task as the mean accuracy rate was superior for Groups 1 and 2 ( $M=.71$ and $M=.79$ respectively) compared to Group 3 ( $M=.47$ ); all $t \mathrm{~s}>1.84, p<.08$.

Taken together these findings corroborate the results obtained from the regression models suggesting that developmental improvement in working memory, inhibitory control, and semantic differentiation underlies the development of category-based reasoning.

## Discussion

One potential limitation of the present study is that the assessment battery did not include a direct measure of vocabulary size. Arguably, vocabulary may be a precursor to category-based reasoning. Nevertheless, there is reason to believe that the high degree of individual variability observed in preschool children's category-based reasoning performance is unlikely to be explained by differences in vocabulary. It is important to note that FSIQ is a composite measure that includes an index of children's verbal ability. Recall that FSIQ scores were comparable between children who became category-based responders at Time 2 and children who were not yet category-based. Additionally, children's high accuracy rates on the Picture Identification task suggests that children are familiar with the labels that were utilized in the category-based reasoning task.

In line with prior work (Fisher et al., 2011; Godwin et al. in press), the present study provides additional evidence demonstrating that only a small percentage of preschool-age children spontaneously engage in category-based reasoning. Additionally, this work implicates three cognitive factors in the development of young children's category-based reasoning: representational change, working memory, and inhibitory control.

First, the strong relationship between improvements in semantic differentiation scores and induction scores suggests that representational change may be one underlying
mechanism in the development of children's ability to engage in spontaneous category-based reasoning. Second, working memory may be an important cognitive factor in the development of category-based reasoning as children need to maintain and manipulate the task-relevant information in working memory. Finally, sufficiently developed inhibitory control may be required to select taskrelevant representations among competing activated representations. These findings indicate that both general cognitive advances (EF and working memory) and changes in domain-specific knowledge (representational change) contribute to the advancement of category-based reasoning.

## Acknowledgments

We thank Malika Sinha, Laura Pacilio, Alyssa Montanaro, Anna Loiterstein, Like Li, Kayoung Joung, and Rachel Walsh for their help collecting data. We thank the children, parents, and teachers for making this work possible. This work was supported by a Graduate Training Grant awarded to Carnegie Mellon University by the Department of Education (R305B090023 \& R305B040063).

## References

Badger, J. R. \& Shapiro, L. R. (2012). Evidence of a transition from perceptual to category induction in 3- to 9 -year-old children. Journal of Experimental Child Psychology, 113, 131-146.
Brown, R. \& Berko, J. (1960). Word association and the acquisition of grammar. Child Development, 31, 1-14.
Bunge, S. A., Dudukovic, N. M., Thomason, M. E., Vaidya, C. J., \& Gabrieli, J. D. E. (2002). Immature frontal lobe contributions to cognitive control in children: Evidence from fMRI. Neuron, 33, 301-311.
Carey, S. (1985). Conceptual Change in Childhood. Cambridge, MA: Bradford Books, MIT Press.
Carlson, S, M., Moses, L. J., \& Breton, C. (2002). How specific is the relation between executive function and theory of mind? Contributions of inhibitory control and working memory. Infant and Child Development, 11, 73-92.
Dale, P.S., \& Fenson, L. (1996). Lexical development norms for young children. Behavior Research Methods, Instruments, \& Computers, 28, 125-127.
Diamond, A. (2002). Normal development of prefrontal cortex from birth to young adulthood: Cognitive functions, anatomy, and biochemistry. In D.T. Stuss \& R.T. Knight (eds.), Principles of Frontal Lobe Function (p. 466-503). London, UK: Oxford University Press.
Dunn, L. M., \& Dunn, L. M. (1997). Peabody Picture Vocabulary Test-Third Edition. Circle Pines, MN: AGS Publishing.
Fisher, A.V. (2010). What's in the name? Or how rocks and stones are different from dogs and puppies. Journal of Experimental Child Psychology, 105, 198-212.
Fisher, A.V., Matlen, B.J., \& Godwin, K.E. (2011). Semantic similarity of labels and inductive generalization: Taking a second look. Cognition, 118, 432-438.
Fisher, A.V. \& Sloutsky, V.M. (2005). When induction meets memory: Evidence for gradual transition from similarity-based to category-based induction. Child Development, 76(3), 583-597.
Fisher, A. V., Thiessen, E., Godwin, K., Kloos, H., \& Dickerson, J. P. (2012). Mechanisms of focused attention in 3- to 5-year-old children: Evidence from a new object tracking task. Journal of Experimental Child Psychology, 114(2), 275-294.

Gelman, S.A. \& Coley, J.D. (1990). The importance of knowing a dodo is a bird: Categories and inferences in 2-year-old children. Developmental Psychology, 26, 796-804.
Gelman, S.A., \& Markman, E. (1986). Categories and induction in young children. Cognition, 23, 183-209.
Gentner, D., Rattermann, M. J., Markman, A. B., \& Kotovsky, L. (1995). Two forces in the development of relational similarity. In T. J. Simon \& G. S. Halford (Eds.), Developing cognitive competence: New approaches to process modeling (pp. 263-313). Hillsdale, NJ: LEA.
Gerstadt, C. L., Hong, Y. J., \& Diamond, A. (1994). The relationship between cognition and action: Performance of 3.5-7 years old on a stroop-like day-night test. Cognition, 53, 129-153.
Godwin, K., Fisher, A., \& Matlen, B. (In press). Development of category-based reasoning in 4- to 7 -year-old children: The influence of label co-occurrence and kinship knowledge. Journal of Experimental Child Psychology.
Howard, D.V. \& Howard, J.H., Jr. (1977). A multidimensional scaling analysis of the development of animal names. Developmental Psychology, 13, 108-113.
Kan, I. P. \& Thompson-Schill, S. L. (2004). Selection from perceptual and conceptual representations. Cognitive, Affective, \& Behavioral Neuroscience, 4, 466-482.
Karmiloff-Smith, A. (1984). "Children's problem solving". In M.E.Lamb, A.L.Brown \& B.Rogoff (Eds.) Advances in Developmental Psychology, Vol.III. New Jersey: Erlbaum, 39-90.
Kemp, C. \& Tenenbaum, J. B. (2008). Structured models of semantic cognition. Behavioral and Brain Sciences. 31(6), 717-718.
Matlen, B.J., Fisher, A.V., \& Godwin, K.E. (under review). Mechanisms of inductive generalization: The influence of cooccurrence probability.
McCauley, C., Weil, C. M., \& Sperber, R. D. (1976). The development of memory structure as reflected by semantic-priming effects. Journal of Experimental Child Psychology, 22, 511-518.
Morrison, R. G., Doumas, L. A. A., \& Richland, L. E. (2011). A computational account of children's analogical reasoning: Balancing inhibitory control in working memory and relational representation. Developmental Science, 14, 516-529.
Nelson, K. (1977). The syntagmatic-paradigmatic shift revisited: A review of research and theory. Psychological Bulletin, 84, 93-116.
Opfer, J. Siegler, R. (2007). Representational change and children's numerical estimation. Cognitive Psychology, 55, 169-195.
Osherson, D., Smith, E., Wilkie, O., Lopez, A. \& Shafir, A. (1990). Category-based induction. Psychological Review, 97, 185-200.
Rogers, T.T. \& McClelland, J.L. (2004). Semantic Cognition: A Parallel Distributed Processing Approach. Cambridge, MA: MIT Press.
Sloman, S.A. (1993). Feature-based induction. Cognitive Psychology, 25, 231 - 280.
Sloutsky, V.M. (2010). From perceptual categories to concepts: What develops? Cognitive Science, 34, 1244-1286.
Sloutsky, V.M., \& Fisher, A.V. (2004). Induction and categorization in young children: A similarity-based model. Journal of Experimental Psychology: General, 133, 166-188.
Wagner, A.D. (2002). Cognitive control and episodic memory: Contributions from prefrontal cortex. L. R. Squire \& D. L. Schacter (Eds.). Neuropsychology of Memory (3rd ed.), pp. 174192. New York: Guilford Press.

Welder, A.N., \& Graham, S.A. (2001). The influences of shape similarity and shared labels on infants' inductive inferences about nonobvious object properties. Child Development, 72, 1653-1673.
Wechsler, D. (2002). WPPAI-III Technical Manual. San Antonio, TX: Pearson, Inc.

# 'This is a wall' - Assigning Function to Objects 

Linn Gralla (gralla@uni-bremen.de)<br>Department of Linguistics and Literary Studies<br>Bremen University, Germany<br>Thora Tenbrink (t.tenbrink@bangor.ac.uk)<br>School of Linguistics and English Language<br>Bangor University, UK


#### Abstract

Construction tasks involve numerous demanding sub-tasks such as creating a mental model of the goal object and integrating the object parts into this model. For this purpose object parts need to be assigned a function within the overall structure. In this paper we examine the linguistic representation of this process. Participants were given 16 object parts to assemble without a manual, and were asked to think aloud while doing so. Depending on condition they were not given any specific information, or told that the goal object was a dollhouse, or shown a picture of the dollhouse. In a second study, participants were asked to instruct a partner to assemble the dollhouse. Results of our linguistic analysis of think-aloud data and instructions reveal three strategies of assigning function to objects, one of which occurred exclusively in instructions. With less specific information about the goal object, functions were more often assigned explicitly. In these cases function tended to relate to the overall structure (e.g. 'house') rather than to structural parts (e.g. 'wall').


Keywords: verbal reports, cognitive discourse analysis, function assignment, conceptual domains

## Introduction

Adults who observe children play are often amazed by their imagination. In their play a plastic cup becomes a boat that sails on the stormy sea. The existence of the boat in the children's mind can be seen as the result of a conceptual mapping process. The child performs a mapping between the domain of plastic objects (cup) and vehicles (boat). Mapping phenomena between different conceptual domains have been widely studied in research on metaphorical transfer (Lakoff \& Johnson, 1980; Croft, 1993) and analogical problem solving (Gentner, 1983). Here we address how speakers use language to assign an object's function in relation to a target domain, in a situation where this function is not self-evident and needs to be identified. This kind of conceptual domain shifting is essential for everyday reasoning. It is required whenever the 'meaning' of an object or depiction needs to be determined that represents something else, for example a symbol on a map that stands for an environmental feature. Our study sheds light on this kind of conceptual mapping by a closer look at the language used to represent it, providing both qualitative (how do speakers say this?) and quantitative (how often do they say it under different circumstances?) insights.

Our scenario concerns the construction of a dollhouse from wooden parts that bear little resemblance to their function within the dollhouse. A board, for instance, needs to be recognised as a 'wall' before it can be placed correctly to serve its function. We investigate the linguistic representation of function assignment in two set-ups: thinkaloud protocols collected while assembling the dollhouse, and verbal instructions to another person assembling the dollhouse. In particular, we analyze the referential terms that are used in reference to both of the domains involved (wooden objects with particular structural features and functional parts within the dollhouse), as well as the linguistic means by which the domains are linked. In the next section we will take a closer look at relevant previous findings on linguistic domain mapping.

## Mapping between conceptual domains

## Use of functional terminology in construction tasks

Malt et al. (1999) propose that the categorization of objects involves two levels: knowing an object vs. naming it. Object perception leads to a representation in terms of a recognition category, along with similar objects. The communication of objects, however, involves a representation in terms of linguistic categories, using conventional or new labels.

So far object categorization within construction tasks has only been studied in settings involving a real or an imagined addressee. Rieser (1996;1997), for example, studied object references in a dialogue scenario where assembler and instructor did not share the same workspace while constructing a toy airplane. As a result, they could share conceptualizations by spoken interaction only. Rieser examined the instructors' strategies of reference to facilitate identification of object parts, and found that they frequently described them in terms of their function in the conceptualized target object (i.e., the toy airplane; e.g., 'this is a horizontal stabilizer', Rieser, 1997:181). He called this phenomenon representational metonymy. The effect may be seen as a kind of reconceptualization and is based on world knowledge and the specific context. Apart from representational metonymy, the physical object itself could also be referred to in descriptive terms, based on the object's structural appearance along with conventional terms for them (e.g. 'Fünfträger' refers to a bar with five holes).

Studies of conceptual layers in construction tasks have focused on the frequency (e.g. von Stutterheim et al., 1993) or interactional purpose of usage of goal structure specific terminology (Rieser, 1997). For a task of instructing a generic addressee on assembling a TV stand, Daniel and Tversky (2012) report that lay instructors started their instruction by giving a list of objects. Although they did not analyze object reference directly, one of their examples shows that functional assignment took place at least once.

Although various authors have thus reported sporadic examples of function assignment, the phenomenon has not been studied systematically so far, and it is unclear to what extent it relates to the need to communicate. However, this kind of conceptual domain mapping is central to the ubiquitous process of recognizing what things are for. In this paper we investigate the linguistic representation of function assignment in two assembly scenarios, and we discuss the cognitive implications that are implied by different options. The next subsection gives some indications of possible linguistic forms.

## Linguistic representation of assignment of function in verbal reports

Rieser (1996; 1997) did not aim to analyze the linguistic features of representational metonymy, but the examples in these papers provide a good impression of how conceptual mapping is represented in assembly dialogues. Participants used phrases such as "das ist das Leitwerk" (this is a horizontal stabilizer) (1997:181), "diese zwei Schienen als Propeller" (those two tracks as propeller) (1997:191), and "dieses Baufix, diese Baufixschraube" (this Baufix, this Baufix screw) (1996:13).

Similar linguistic forms were found by Tenbrink and Seifert (2011) for conceptual mapping processes in a spatial problem-solving task. They analyzed written reports of participants planning holiday tours based on a map. This scenario involves two conceptual domains: that of the physically present road map, and that of the real world environment represented by the map. In their analysis of conceptual mapping processes, Tenbrink \& Seifert focused on the distribution and nature of nouns, verbs, adjectives/ adverbs, and temporal markers as indicators of either domain. If indicators for both domains were used within a single sentence, this sentence typically also contained indicators for conceptual mapping between them, such as the modal verbs could and should (e.g., 'I looked for a route that could be traveled'). Further indicators were final discourse markers (i.e. 'in order to') and the particle 'als' (as), which signaled "mapping from plan to purpose" (Tenbrink \& Seifert, 2011:116).

## Dollhouse assembly: Empirical studies

## Research goals

Gralla (in prep.) collected an explorative corpus of unconstrained language production data related to the assembly of a dollhouse from a set of wooden object parts
in various conditions (explained below). For the purposes of this paper we inspected this corpus to identify the ways in which participants spontaneously used language to assign function to objects. Based on Malt et al.'s (1999) distinction between perceiving and communicating about objects, we expected systematic differences to emerge between the two distinct discourse tasks (thinking aloud vs. instructing a partner). Furthermore, the linguistic representation of function assignment should also be affected by the amount of prior knowledge available about the goal object.

In order to systematize our insights on linguistic forms, we define a mapping phrase as consisting of the reference to a physically present object $\mathbf{x}$ that is assigned the functional term $\mathbf{y}$ by a relational term. Preliminary findings (reported in Tenbrink \& Gralla, 2009) and the examples seen in Rieser $(1996 ; 1997)$ suggest that these relational terms can vary with respect to the amount of certainty expressed. Relational verbs, as in "this is a wall", for instance, signal that the speaker is absolutely certain about the assignment, since the relation is expressed as a plain fact. The modal verb 'müssen' (must) implies a lesser but still high amount of certainty, as in "this must be a wall", whereas 'können' (can) reflects uncertainty, and 'sollen' (should) encodes a medium level (Halliday, 1985). In other cases the relational term expresses a comparison (look like, use as). This strategy implies that the speaker decides that $\mathbf{x}$ represents $\mathbf{y}$ because $\mathbf{x}$ and $\mathbf{y}$ share some features. This assignment is tentative because $\mathbf{x}$ may also be something else.

Interestingly, Tenbrink and Seifert (2011) also identified different kinds of modal verbs as markers of domain mapping for their tour planning scenario, in which certainty did not play any role as the planning process was entirely in the participants' hands. However, modal verbs also carry different connotations, which might play a role in this context. Whereas 'wollen' (want) expresses the subject's intention 'können' (can) expresses the possibility given that the subject is granted the permission (Engel, 2002), and 'sollen' (have to) signals an obligation. In our study, we aimed to shed further light on the repertory of linguistic forms used to express conceptual mapping, along with their distribution across the different conditions in the corpus.

Regarding the influence of prior knowledge, we expected that participants who were provided with unspecific information about the goal object should signal more uncertainty in their assignment of function than participants provided with a picture of the goal object. Participants asked to instruct another person (rather than think aloud while constructing the dollhouse themselves) should introduce the given objects and assign their function explicitly in mapping phrases (cf. Rieser, 1996; 1997; Daniel \& Tversky (2012).

## Methods

In the first of the studies carried out by Gralla (in prep.), think-aloud protocols were recorded while participants assembled a two-story dollhouse for themselves. 50 university students ( 22 male, 28 female; aged 19-42 years, mean age 24 years) participated for course credit or mone-
tary compensation. They were told that they would be given object parts to be assembled without a manual. Knowledge of the goal state varied between mention of "a dollhouse" (verbal goal condition), being shown a picture of the assembled dollhouse for 30 seconds (verbal and visual goal condition), and no such information (underspecified goal condition). Following the instruction, participants entered a room and saw a cardboard box, two wooden boards, and a triangular piece of wood on a table. The box contained 13 wooden parts (see picture 1 for an example of one object). Participants were instructed and reminded to think aloud while solving the task. There were no time constraints.

Picture 1: One wooden assembly piece


In the second study, verbal instructions on the assembly of the same two-story dollhouse were recorded. 16 students ( 9 male, 7 female; aged $20-28$ years, mean age 23 years) who did not participate in the previous experiment were first instructed to assemble the dollhouse for themselves. They were given the same information as in the verbal and visual goal condition but without the task to think aloud. After successful completion of the task the participants were introduced to another student (a confederate). They were asked to instruct this person to assemble the dollhouse, using a Skype-based one-way video connection that did not allow for any responses by the assembler.

## Analysis methods

Our data set contained 50 think-aloud protocols that were equally distributed between conditions (17 underspecified goal, 16 verbal goal, 17 verbal and visual goal), and 16 instructions. First, mapping phrases were identified in a qualitative analysis. They contained one object reference (either deictic or nominal, e.g. this thing) and a domain specific functional term, e.g. wall. Nouns belonging to the semantic field of the goal domain 'house' were classified as domain specific. Second, all phrases were classified either as direct mapping (use of relational verbs) or as representational mapping (use of comparison). Third, all verbs were annotated with regard to verb kind and type. Furthermore, the referential terms for the object (either deictic or nominal) as well as the functional term were annotated.

Example 1 (which refers to the object in picture 1) represents direct mapping with the verb 'sein' (be) as the relational term, and a deictic reference (this) for the object $\mathbf{x}$ that is assigned the functional term (y) wall.
(1) so das ist dann so ne Wand (so this is some wall then)

Example 2 (which refers to the roof of the dollhouse) illustrates representational mapping with the verb phrase 'aussehen wie' (look like) as the relational term, and the nominal reference red building part for the object $\mathbf{x}$ that is assigned the functional term roof.
(2) $\ldots$ das rote Baustück nen bisschen wie n Dach eines Hauses aussieht (... this red building part looks a bid like a roof of a house.)

The qualitative analysis and annotation of categories then led to the identification of quantitative frequencies in the various conditions.

## Results

131 mapping phrases were identified in the think-aloud protocols. The highest frequency ( 54 cases produced by 14 participants) was observed in the underspecified goal condition, as opposed to 44 cases produced by 14 participants in the verbal goal condition, and 33 cases produced by 11 participants in the verbal and visual goal condition. 23 cases of function assignment could be identified in 12 of the 16 instructions.

Figure 1: Distribution of mapping strategies (mean raw frequency with error bars $+/-2$ SE)


## Assignment of function in assembly

Direct mapping (108 cases) was more frequent than representational mapping ( 22 cases). The frequency of mapping strategies did not differ significantly between conditions (see Figure 1), $\chi^{2}(4, \mathrm{~N}=131)=7.12, \mathrm{p}=.130$. With respect to the linguistic representation of both mapping strategies, the verb 'be' was most frequent ( $91.51 \%$ ) in direct mapping. In representational mapping 'look like' (59.09\%) was used along with 'use as' (22.73\%).

29 modal verbs were identified in 18 protocols. They were almost equally distributed between conditions (11 in the underspecified goal condition: $\mathrm{M}=0.59, \mathrm{SD}=0.23 ; 8$ in verbal goal: $\mathrm{M}=0.50, \mathrm{SD}=0.13$; and 10 in verbal and visual goal: $\mathrm{M}=0.58, \mathrm{SD}=0.23$ ). Almost all (27) modal verbs occurred in direct mapping phrases. Three modal
verbs were used, namely could, must, and should. These were distributed between conditions as follows (figure 2). Participants in the underspecified goal condition used all three modal verbs, with could $(\mathrm{M}=0.83, \mathrm{SD}=0.31)$ and should $(\mathrm{M}=0.67, \mathrm{SD}=0.21)$ being more frequent than must $(\mathrm{M}=0.33, \mathrm{SD}=0.21)$. Participants in the verbal goal condition showed a preference for could $(\mathrm{M}=1.00, \mathrm{SD}=$ 0.26 ) whereas participants in the verbal and visual goal condition used must most frequently ( $\mathrm{M}=0.83, \mathrm{SD}=0.31$ ). However, the observed differences between conditions did not reach statistical significance, $\mathrm{L} \chi^{2}(4, \mathrm{~N}=29)=8.51, \mathrm{p}=$ .075, probably due to the low numbers and varied individual production of linguistic choices in our setting.

Figure 2: Distribution of modal verb type in mapping processes within conditions (mean raw frequency with error bars +/- 2 SE)


## Assignment of function in instructions

Besides identifying instances of direct and representational mapping, our qualitative data analysis revealed a further (unexpected) strategy of function assignment in instruction texts. Instructors used a more implicit way of assigning function, illustrated in example 3:
(3) „jeweils mit einem kurzen Stück einem ähm einem Wandelement ..." (each with a small piece a um a wall element)

In this example the object (a small piece) is assigned its function (wall element) by simply renaming it after a hesitation phase. Since there is no relational term connecting object reference and functional term, no conclusions can be drawn about the level of certainty reflected in this strategy. This type is called reframing because the speaker changes the frame of conceptualization from unspecific to specific. The distribution of mapping strategies shows that instructors used direct mapping most frequently ( $\mathrm{M}=0.63, \mathrm{SD}=0.20$ ). A closer look at the linguistic structure of direct mapping reveals that the verb 'be' was used in all cases; instructors never used modal verbs to assign function. Furthermore,
representational mapping and reframing ( $\mathrm{M}=0.31, \mathrm{SD}=$ 0.12 ) were also frequently used. The influence of discourse type on mapping strategies is significant, $\chi^{2}(3, N=141)=$ 8.37, $\mathrm{p}=.039$. Representational mapping was used more frequently in instructions $(0.44(\mathrm{SD}=0.18))$ than in think aloud protocols ( $0.40(\mathrm{SD}=0.77), \mathrm{z}=2.0, \mathrm{p}<.05$ ).

## Object reference and functional terminology in mapping phrases

A significant difference emerged between the two different discourse types (think-aloud vs. instruction) with respect to the referential term used for the object, $L \chi^{2}(6, N=154)=$ $57.73, \mathrm{p}=.000$. Participants who were thinking aloud while assembling the dollhouse mostly used deixis (e.g., this, that) rather than nominal references. Instructors, on the other hand, used nominal object references ( $0.94(\mathrm{SD}=0.85)$ ) more frequently than assemblers $(0.06(\mathrm{SD}=0.24), \mathrm{z}=7.5$, $\mathrm{p}<.001$ ), and they used deictic references ( 0.50 (SD = $0.63)$ ) less frequently than assemblers ( $2.42(\mathrm{SD}=2.54), \mathrm{z}=$ $-2.6, \mathrm{p}<.01$ ).
Five nouns or their synonyms were used most frequently to express function in the dollhouse: floor, roof, wall, story, and house. This list represents two perspectives. Some of these terms (floor, roof, and wall) highlight individual pieces of the dollhouse and are therefore part-based, while the others refer to the whole structure (house) or larger portions of it (story), which may consist of several individual pieces and are therefore structure-based, e.g. "aah das könnten auch Stockwerke sein" (aah this could also be stories). Conditions differed with respect to the distribution of these two perspectives (see figure 3). While participants in the underspecified goal condition used structure-based terminology $(\mathrm{M}=1.65, \mathrm{SD}=0.57)$ frequently, participants in all other conditions (including instruction) referred to individual parts significantly more often, $\chi^{2}(6, \mathrm{~N}=154)=$ $13.44, \mathrm{p}=.037$.

Figure 3: Functional terminology assigned to objects in mapping phrases - categorized according to perspectives (mean raw frequency with error bars $+/-2$ SE)


## Discussion

Our studies addressed the linguistic representation of function assignment in an unaided object assembly task. We presented a qualitative analysis of the different components of functional mapping phrases, i.e. relational term, object reference, and assigned functional terminology. Furthermore, we examined the relative frequencies of occurrence in different conditions, highlighting the influence of prior knowledge as well as differences between think-aloud protocols and instructions. The most striking differences were found with regard to the latter, revealing the influence of the speaker's communicative intention on the linguistic representation of function assignment. With respect to the former, lack of information appeared to lead to a more structure-based (rather than part-based) perspective, as expressed by the nouns used to refer to objects.

In our qualitative analysis, we identified three strategies of function assignment that reveal various levels of certainty. In the direct mapping strategy, the relational terms directly connect an object to a functional purpose, typically using the verb 'be', which signals a high level of certainty. Modal verbs (such as must, could) were used to modify the level of certainty by weakening it. In the representational mapping strategy, the relational terms look like, use as, etc. signal that the object is assigned to a goal-object based concept that serves a particular function. Since the relation signals representation rather than 'being', participants signal the tentative character of their assignment by choosing this strategy. In the reframing strategy, which in our study only occurred in instructions, the relational term is replaced by a hesitation marker between the reference to the object and the assigned functional term.

On the whole, participants giving instructions used mapping phrases only rarely. This can be interpreted in two ways. First, it may mean that they focused on the step-bystep procedure of the assembly, rather than attending to the goal structure as such. This explanation is supported by the finding that instructors used part-based terminology more frequently than structure-based references when assigning function. Furthermore, Daniel and Tversky (2012) found that instructors omitted explicit information on object parts and sequential order, but not on sequences of actions, when time was constrained. Second, our instructors may have believed that functional terms would not facilitate object identification, but rather result in additional cognitive load for their addressee (cf. von Stutterheim et al. 1993) because the wooden objects did not resemble their function in the goal structure in any obvious way.

Whenever instructors used mapping phrases, this happened with certainty, as reflected in the linguistic form chosen. This result straightforwardly reflects the fact that participants were already familiar with the dollhouse assembly task when they gave the instructions, and therefore did not need to assign objects to function in a tentative way. However, instructors used representational mapping more often than participants in the assembly study did. This may
be seen as an explicit strategy of emphasizing that function is assigned to an object. The instructors may have considered their addressee's situation, who could only see the wooden objects rather than their function. Explicitly highlighting the mapping process may therefore be felt as a useful supportive strategy.

The most interesting case, in our view, is the third strategy, which was found only in instructions: introducing the functional term after a marker of hesitation (reframing) rather than a relational term. Although this construction may seem accidental due to the hesitation marker, which is by its nature exclusive to spoken language, it was used by as many as 5 out of 12 participants. Arguably, Rieser's (1996:13) example cited above (this Baufix, this Baufix screw) is similarly structured, although no hesitation marker is reported.

What might lead speakers to use this function assignment strategy in instructions? In effect, the previous reference is elaborated by the first. In interactive scenarios, such an elaboration would happen frequently in response to a request for clarification, as described in the literature related to the referential communication paradigm (e.g. Clark \& Wilkes-Gibbs, 1986; Horton \& Gerrig, 2002). However, this cannot explain our observations, since our scenario did not allow any interaction between instructor and assembler. Instead, speakers apparently spontaneously felt the need to be more specific with regard to the function of the object that had just been referred to. Conceivably, the function of this particular object was so prominent in their minds that they directly reframed the sentence. Again, they started from their addressee's perception, leading over to the function of the wooden object. This suggests that the strategy is specific to the communicative goal of shared attention and object identification.

The analysis of the referential form for the physically present object provides further evidence that our participants in the instruction scenario took an addressee-centered perspective. Participants who were thinking aloud while assembling the dollhouse used deictic references more frequently than instructors, who tended to use nominal references instead. These findings highlight the influence of communicative intention on referential form and thereby support the assumption that think-aloud data reflects the speaker's thoughts that are not tailored for an addressee (Ericsson \& Simon, 1984).

A clear effect of prior knowledge emerged with respect to the perspective expressed by the functional term assigned to an object. Participants who had to construct the goal object from scratch, i.e. without prior knowledge, frequently assigned function by reference to the whole structure of the goal object, or larger portions of it, as in house or story. Participants who knew about the goal structure, on the other hand, assigned function mostly to object parts (e.g. 'wall'). This difference suggests that participants who had associations about typical parts of the goal object assigned these functions to the given objects. Participants who needed to construct a mental model of the goal object, on
the other hand, additionally seemed to assign function to the object arrangement as a whole, either to construct the model or to confirm their hypotheses about the goal object. This suggests that instances of function assignment are traces of the mental process of conceptualization of objects within the mental model of the goal object.

The consistency of using direct mapping strategies across conditions in the unaided assembly study showed that participants assigned function with a high amount of certainty in most cases, independent of the amount of prior knowledge. Some differences were observed at the level of certainty expressed by the use of modal verbs. Surprisingly, less knowledge about the goal objects appeared to result in higher certainty as expressed in mapping phrases. Specifically, participants who were provided with no specific prior knowledge or only a verbal clue tended to use mapping phrases frequently, with a preference for direct mapping using the relational term be without modification by modal verbs that signal uncertainty. This may suggest that participants who were shown the picture of the goal object may feel somewhat constrained by their expectation to match the given objects to their memory of the picture. If the provided objects do not match this memory, the matching process may be felt as less certain than in a more flexible situation where the only information given (if any) is the nature of the goal object. The verbal clue would then result in useful associations that are not too specific to constrain the participants' flexibility in assigning function to the wooden object parts.

In this paper we reported findings on explicit function assignment only (i.e., verbalizations directly assigning function to object parts). For this set of results, significant differences could be found only in the comparison of thinkaloud vs. instruction studies, but not between the three think-aloud assembly conditions. Gralla (in prep.) further considered indirect function assignment by investigating all domain-specific nominal references in the think-aloud protocols. With this larger data set, a clear influence of prior knowledge could be observed. Participants who were told about the nature of the goal object used domain specific nouns more frequently than participants in the other conditions did (Gralla, in prep.).

## Conclusion

Our study addressed for the first time how speakers assign function to objects during an explorative unaided object assembly task and in instructions. Results show influences of the situational context on the ways in which this domain mapping is made explicit, on certainty expressed in language, and on the functional term chosen for reference. Those findings encourage more controlled studies to further explore the effects of problem solving conditions on mapping processes of this kind.

## Acknowledgments

Funding by the Zentrale Forschungsförderung Universität Bremen and the SFB TR8 Spatial Cognition (Deutsche

Forschungsgemeinschaft DFG), I6-[NavTalk] is gratefully acknowledged.

## References

Clark, H.H. \& Wilkes-Gibbs, D. (1986). Referring as a collaborative process. Cognition, 22,1-39.
Croft, W. (1993). The role of domains in the interpretation of metaphors and metonymies. Cognitive Linguistics 4(4):335-370.
Daniel, J-P. \& Tversky, B. (2012). How to put things together. Cognitive Processing, 13, 303-319.
Engel, U. (2002). Kurze Grammatik der deutschen Sprache. München: Iudicium.
Ericsson, K.A. \& Simon, H.A. (1984). Protocol Analysis. Verbal Reports as Data. Cambridge: MIT Press.
Gentner, D. (1983). Structure-Mapping: A Theoretical Framework for Analogy. Cognitive Science, 7, 155-170.
Gralla, L. (in prep.). Linguistic representation of problem solving processes in unaided object assembly Application of Cognitive Discourse Analysis. Unfinished doctoral dissertation, Department of Linguistics and Literary Studies, Bremen University.
Halliday, M.A.K. (1985). An Introduction to Functional Grammar. London: Edward Arnold.
Horton, W.S. \& Gerrig, R.J. (2002). Speakers' experience and audience design: knowing when and knowing how to adjust utterances to the addressee. Journal of Memory and Language, 47,589-606.
Lakoff, G. \& Johnson, M. (1980). Metaphors we live by. Chicago: University of Chicago Press.
Malt, B., Sloman, S.A., Gennari, S., Shi, M., \& Wang, Y. (1999). Knowing versus Naming: Similarity and the Linguistic Categorization of Artifacts. Journal of Memory and Language, 40, 230-262.
Rieser, H. (1996). Repräsentations-Metonymie, Perspektive und Koordination in aufgabenorientierten Dialogen. Projektreport SFB 360 Situierte Künstliche Kommunikation, Universität Bielefeld.
Rieser, H. (1997). Repräsentations-Metonymie, Perspektive und Koordination in aufgabenorientierten Dialogen. In C. Umbach, M. Garski, \& R. Hörnig (Eds.), Perspektive in Sprache und Raum. Aspekte von Repräsentation und Perspektivität. Wiesbaden: DUV.
Tenbrink, T., \& Seifert, I. (2011). Conceptual layers and strategies in tour planning. Cognitive Processing, 12, 109125.
von Stutterheim, C., Mangold-Allwinn, R., Baratelli, S., Kohlmann, U., \& Kölbing, H-G., (1993). Reference to objects in text production. In J. Nuyts \& E. Pederson (Eds.), Perspectives on language and conceptualization. Amsterdam: John Benjamins.

# Is Feedforward Learning more Efficient than Feedback Learning in Smart Meters of Electricity Consumption? 

Mona Guath (mona.guath@psyk.uu.se)<br>Department of Psychology, Uppsala University<br>SE-751 42, Uppsala, Sweden<br>Peter Juslin (peter.juslin@psyk.uu.se)<br>Department of Psychology, Uppsala University<br>SE-751 42, Uppsala, Sweden

Philip Millroth (philip.millroth@psyk.uu.se)<br>Department of Psychology, Uppsala University<br>SE-751 42, Uppsala, Sweden


#### Abstract

The most popular way to improve consumers' control over their electricity cost is by providing frequent and detailed feedback with "in-home displays" (IHD). In this study, we examined alternative ways to train experimental participants to control and optimize their use of electricity by "feedforward" training to map energy consuming behaviors to costs. The participants were trained in one of four experimental conditions, one feedback ("IHD") and three feedforward conditions before they had to control the electricity consumption in a simulated household. Results showed that one of the feedforward conditions produced somewhat higher utility and as good or better satisfaction of a monthly budget than the feedback training condition, despite never receiving any feedback about the monthly cost, but the generalization to a new budget constraint proved to be slightly poorer.


## Introduction

The use of so-called "smart electricity meters" is rapidly becoming common. It has been estimated that within the European Union alone some 51 billion euro is being invested in smart meters (Faruqui, Harris, \& Hledik, 2009). In many countries, household energy consumption is still billed once a month, but smart meters can offer feedback that is detailed and more frequent with so called In Home Displays (IHDs). Intuitively, the latter kind of feedback system seems more beneficial, and, indeed, many early studies suggested energy reductions up to $15 \%$. However, more recent studies point at consumption reductions at $2-4 \%$, few of them being significant (Klopfert \& Wallenborn, 2011). In the present study, we focus at in-home displays (IHDs), which only display the electrical consumption at different time intervals, and, unlike smart meters, they do not have a two-way communication with the central system. In a previous laboratory experiment (Guath, Millroth, Elwin, \& Juslin, 2012), we investigated how feedback about electricity consumption is best presented to electricity consumers in order to control and optimize their use of electricity. To measure a participant's energy efficiency
in an experimentally controlled environment, the participants took on the role of an inhabitant in a simulated household, performing different types of energy consuming behaviors within a given budget (Figure 1). The goal of decreasing electricity consumption is often emphasized, but the participant's task is actually an optimization problem that requires an appropriate balance between the cost of the electricity consumed and the benefit or utility obtained. The problem is illustrated in Figure 2, where the utility of electricity consumption is plotted against cost. The maximum utility obtainable at a given cost, assumed to be a decelerating function of the cost, is illustrated by the curve in Figure 2. The hypothetical utility obtained at a cost by a consumer is illustrated with a dot. The task is to move closer to the line for "maximal utility", however, this is associated with two constraints: achieving sufficient utility to make life bearable and not surpassing a constrained budget.

Guath et al. (2012) showed that in a deterministic system, frequent and detailed feedback was advantageous, but in probabilistic system, feedback aggregated over time was better, because it filtered out random noise.

## The Present Study

In the present study, we wanted to evaluate if the same improvement could be obtained by feedforward training, rather than feedback training (as in most IHDs), hence, minimizing the negative effects from feedback interventions as conceptualized in Kluger and DeNisi's (1996) study. Specifically, we wanted to avoid the decrease of effectiveness when attention is moved away from the task to the self, thus, making the effects of the training short-term. Another motive was to make the mapping task more flexible, not being dependent on the simulated household (Figure 1). Detailed and frequent feedback (an IHD) was compared to three feedforward conditions. Feedforward is defined as a process where knowledge is used to act directly to control the system, thus anticipating the changes that will occur (Basso \& Olivetti Belardinelli, 2006). In the present task, partici-
pants had to control the monthly cost of electrical consumption. Feedback training involved feedback about this criterion variable of daily and monthly cost of electrical consumption from experience with the task (running the simulated household in Figure 1). Feedforward training involved no feedback about the criterion variable (monthly electricity cost), but three different training schemes in various ways teaching the participants to directly map energy consuming behaviors to their costs ("map" refers to the mathematical concept of associating each element in a set with an element of another set, here the electrical cost to a certain electrical post).


Figure 1. The computer display in the simulated household in the experiment.


Figure 2. Illustration of the two ways of obtaining maximum utility at a specific cost (the solid nonlinear function), either by saving or by optimizing the use.

The choice of feedforward conditions was, in part, inspired by Pachur and Olsson's (2012) study of how learning tasks affect performance and strategy selection. They investigated two learning tasks, direct criterion learning and learning by comparison, and how these affected performance depending on the type of test (paired-comparison, classification, estimation) and decision environment (linear vs. non-linear). Pachur and Olsson (2012) concluded that direct criterion learning invites exemplar memory processes (Nosofsky, 1986), while learning by comparison invites processes of cue abstraction. Because our task is non-linear, if anything, exemplar memory should be a more efficient process than the abstract processes involved in cue abstraction,
which are constrained to mainly capture linear and additive tasks (Juslin, Karlsson, \& Olsson, 2008).

Our first feedforward condition, metric mapping (corresponding to direct criterion learning), informed about the function that relates the consumption to its cost, as studied in research on function learning (e.g., Kalish, Lewandowsky \& Kruschke, 2004). When training function concepts, a continuous stimulus variable is associated with a continuous response variable, in this case an electricity post (e.g., inner temperature) with its monthly cost. The metric mapping consisted of learning to map a certain electricity consuming activity (e.g., using hot water $15 \mathrm{~min} /$ day) to its cost (i.e., $262 \mathrm{SEK} / \mathrm{month}$ ).

The second condition was rank-order mapping (corresponding to learning by comparison) as conceptualized in decision by sampling (DbS) (Stewart et al., 2006). In DbS , it is assumed that people do not store metric knowledge in memory but only perform ordinal comparisons. Instead frequency accumulation in pair-wise comparisons are used for evaluating a target attribute against a decision sample. Indeed, the results in Pachur and Olsson (2012) suggested that at least in linear tasks learning by pairwise comparisons was more efficient than training with metric mapping, despite that the pairwise comparisons provide no explicit metric information about the criterion. On the other hand, if people also need to store metric knowledge in our task, then metric mapping should be more efficient. The rankorder evaluations are elicited by questions concerning the relation between two electricity consuming device (e.g., Which of the following has the highest monthly cost: A: Having the lights on for 60 minutes per day or B: Having the computer on for 10 minutes per day?).

The third feedforward condition was causal mapping training, in which the participant is encouraged to experiment with the individual and total monthly cost of the electrical posts in a minimalistic computer program. The causal mapping condition is inspired by the theory of causal nets (Holyoak \& Cheng, 2011) that accounts for how people learn about strength and structure as well as direction of causal relations. In view of this literature, we expected that invitation to manipulate the system in real time and experiment by changing individual variables during training should produce a more accurate (causal) model of relationships in the system.

Given that our decision task is non-linear, where the linear and additive integration afforded by cue abstraction is less appropriate, and the results suggesting that metric mapping invites exemplar memory (Pachur \& Olsson, 2012), performance in the metric mapping condition is expected to be better than in the rank-order and the causal mapping conditions. We also looked at the ability to generalize knowledge to a new budget (from 2000 SEK á month to 1500 SEK á month).

## Method

## Participants

One-hundred-and-twenty-nine students at Uppsala University volunteered to participate and were compensated with a cinema ticket (worth approximately $\$ 10$ ) or by
course credit. The sample consisted of 89 females and 40 males, with mean age 24.5 years ( $S D=4.66$ ).

## Material and Procedure

The experiment consisted of four parts presented in the following order: pre-test, systematic learning, post-test for effects of the systematic learning and post-test for ability to generalize to a new budget. Participants were given written and verbal instructions for each part.

The participant was presented with a sketch of a home on the computer screen indicating various energy consuming appliances, in all 18 posts (Figure 1). The task was to regulate the electricity consumption in the house so that its fictive inhabitant (an avatar called "Peter") received as much utility as possible from the consumption. The avatar was a way to incorporate the two parameters of utility and cost in the task in a comprehensible way. In Pretest and Posttest 1 of all conditions, the participant regulated the energy consumption each day for a period of 30 days with the aim to maximize the utility from the energy consumption within a budget of 2000 Swedish Crowns (SEK), approximately \$300, per month (changed to 1500 SEK in Posttest 2). On each new day, the participant could adjust the indoor temperature, the number of times of use per week for the dishwasher, washing machine, and tumble drier etc. When the settings had been made, they could not be changed for that day. On each day, the previous day's settings were presented as default, but they could be changed by the participant.

The utility of consumption for each appliance was presented by a bar on the right side of the screen. A separate bar for each post increased with the utility of consumption associated with this post, and a global sum in the upper right corner increased with the overall utility of the energy consumption. The utility $u_{\mathrm{i}}\left(t_{\mathrm{ij}}\right)$ obtained by consumption $t_{\mathrm{ij}}$ of post $i(i=1 \ldots 18)$ at level $j$ was,

$$
\begin{equation*}
u_{i}\left(t_{i j}\right)=\sum_{i=1}^{18} w_{i} \cdot\left(t_{i j}^{\alpha_{i}} / r_{i}^{\alpha_{i}}\right) \tag{1}
\end{equation*}
$$

where $w_{\mathrm{i}}$ is the linear weight in the overall summed utility $\left(\Sigma w_{\mathrm{i}}=1\right), r_{\mathrm{i}}$ is a ceiling on the allowable consumption, and $\alpha_{\mathrm{i}}$ is a parameter for the curvature of the utility function for post $i$. Eq. 1 defines utility functions with diminishing marginal return, where the posts differ both in the rate of the diminishing return $\left(\alpha_{\mathrm{i}}\right)$ and in their weight in the total utility $\left(w_{\mathrm{i}}\right)$. The parameters were selected to approximate realistic utility functions.
The total utility $U$ was the sum of the utility of each of the 18 posts,

$$
\begin{equation*}
U=\sum_{i=1}^{18} u_{i}\left(t_{i j}\right) \tag{2}
\end{equation*}
$$

In pretests and posttests, feedback was presented only for utility. Participants were given no feedback regarding the cost of their settings. After the pretest, the participants were assigned to one of four learning conditions.

## Detailed and Frequent Feedback ("IHD")

In the feedback condition, participants continued with 120 days in the simulated household. Furthermore, and
most importantly, they also received feedback about the cost of their consumption. After each day the participants received a bill containing feedback on the cost of energy consumption where the feedback was presented in terms of used kWh and the cost in SEK, as based on a fixed price of 1.40 SEK per kWh. The bill showed the cost for each appliance as well as the total sum for all appliances. This detailed and immediate feedback resembles that of an IHD. If the budget was exceeded, the total cost was red-lighted; if not, it was green-lighted. A normally and independently distributed random error, with standard deviation equal to $5 \%$ of the cost, was added to the cost of each post to simulate probabilism ${ }^{1}$. The total cost $C$ was the sum of the consumption cost $c\left(t_{\mathrm{ij}}\right)$ of the individual posts:

$$
\begin{equation*}
C=\sum_{i=1}^{18} c\left(t_{i j}\right) \tag{3}
\end{equation*}
$$

## Metric mapping

For participants in the metric mapping condition, the task was to learn to map consumption of each electricity post directly to its cost. They were presented with questions such as: "What is the monthly cost for having the computer on for 15 minutes a day?" Intervals of 5, 10, 15,30 and 60 minutes were used to give the participants a broad spectrum of the cost for each appliance. Participants reported their responses for one question at a time and were then given feedback on whether the response was correct or not and, if not, what the correct answer was. The program coded answers within $\pm 20$ percent of the correct answer as correct. A stop criterion was set for three correct responses in one block of five questions (one block involved the cost for $5,10,15,30$ and 60 minutes of use). When the participants achieved the stop criterion for one appliance they continued with the next appliance, until they had gone through all appliances in the house. Electric posts for lighting and hot water (shower and tap water) was lumped together, creating 12 different appliances from the 18 electricity posts in the simulated household. For appliances that were run on a weekly basis, such as the dish washing machine, participants were asked about the cost for number of runs per week.

## Rank-order Mapping

For participants in the rank-order mapping condition the task was to learn which of pairs of electricity consuming activities that is most costly. They answered questions such as "What has the highest monthly cost? A: Having the lights on for 60 minutes per day or B: Having the computer on for 10 minutes per day?" After each guess they were provided with feedback on whether the response was correct or not. The items were sampled from a pool of questions created by crossing the 12 applianc-

[^74]es (as in the direct-mapping condition) by the five time intervals (5, 10, 15, 30, 60 minutes, except for appliances run on a weekly basis). The items were sampled randomly with one constraint: each appliance had to appear at least once during the training session. Each participant received a unique sample. The participants started with 200 items; thereafter, they continued until they answered 19 of the latest 20 questions correctly.

## Causal Mapping

For participants in the causal mapping condition the task was to learn the relationship between consumption and electricity cost by interacting with the appliances on a real time basis, in order to obtain a sense of the cause and effect relationships. Again, the participants trained on 12 appliances with the same time intervals as in the other mapping conditions ( $5,10,15,30,60$ minutes, except for appliances run on a weekly basis). Participants were presented with a program where they could manipulate the usage of appliances on slide bars. Each slide bar had five levels for the time intervals of 5,10 , 15,30 and 60 minutes. Next to each slide bar, the cost of the appliance was indicated. The cost changed simultaneously with the manipulation of the slide bar. Participants could observe the monthly cost for their current settings on the top of the screen, also changing simultaneously with the manipulation of a slide bar. All appliances were presented on the same screen. The order in which the appliances were presented on the screen was randomized for each participant. A time limit of 15 minutes was set, and the participants were told to experiment and learn as much as possible in this time, with the goal to optimize their behavior in the household.

## Posttests

After participants had finished their respective training session, they all continued with another 30 days in the simulated household, similar to the 30 days in the pretest. After that first post-test, the participants continued with another round of 30 days in the simulated household, but this time the budget constraint was set to 1500 SEK instead of 2000 SEK, with the goal of investigating participants' ability to generalize the knowledge they had acquired in the systematic learning conditions.

## Design

The experiment involved a $4 \times 2$ mixed factorial design, with learning condition (detailed and frequent feedback, metric mapping, rank-order mapping, and causal-model mapping) as between-subjects independent variable, and budget constraint ( 2000 and 1500 SEK) as the withinsubjects independent variable. The participants were randomized to one of the between-subjects conditions, resulting in app. 30 participants in each condition. Dependent measures were the cost and utility of the use of electricity, with a particular eye to the maximization of utility within the indicated budget constraints of 2000 SEK and 1500 SEK per month.

## Results

In pretest, there were no significant differences between the conditions and all conditions exceeded the budget of

67 SEK/day. All conditions reduced their median cost from the Pretest to Posttest 1 (Wilcoxon Test, $T=1135$, $Z=7.187, p<.001$ across all four conditions; the same holds separately within each condition, all $p \mathrm{~s}<.005$ ).
In Figure 3, the median utility is plotted against the median cost for Posttest 1. The rank order condition was unable to satisfy the budget. The three conditions satisfying budget performed similarly, although metric mapping produced somewhat more utility than the other two conditions, which both fell below budget, as observed previously with feedback training (Guath et al, 2012).


Figure 3. The median utility obtained across the participants plotted as a function of the median cost in Pretest and Posttest 1 in each of the four training conditions ( $N \approx 30$ ). The vertical line represents the budget, while the curve is the maximum utility obtainable as a function of the cost.

There was a significant difference between the conditions in the median cost at Posttest 1 (Kruskal-Wallis test: $H(3, N=129)=18.607 p=.003)$, where pairwise multiple-comparisons indicate a significant difference between rank-order mapping and feedback training ( $p<.001$ ) and between rank-order and causal mapping ( $p$ $=.005)^{2}$. In both cases, rank order mapping has a significantly higher cost. There was also a significant difference between the conditions in median utility at Posttest 1 (Kruskal-Wallis test: $H(3, N=129)=12.557 p=.006)$, where pairwise multiple-comparisons indicate a significant difference between rank-order mapping and feedback training ( $p=.019$ ) and between rank-order mapping and causal mapping ( $p=.009$ ). In both cases, rank-order mapping has a significantly higher utility. The main differences is between the rank-order condition and the other conditions, with small differences between the other conditions, albeit with a slight hedge for metric mapping that comes closest to optimal performance (i.e., where the lines in Figure 3 intersect).

Figure 4 reports a more strict analysis only including those participants that roughly satisfied the budget (i.e.,

[^75]fell within $\pm 5$ units of 67 SEK/day). While app. 50 \% of the participants were able to satisfy the budget with feedback training and metric mapping only a minority of participants were able to satisfy the budget with causal mapping and rank order mapping ( $27 \%$ and $12 \%$, respectively). Among the only two conditions with many participants satisfying the budget, metric mapping produced significantly more utility than feedback training (Mann Whitney: $U=75, Z=2.179, p=.029$ ). Thus, if anything, metric mapping appears to produce somewhat better performance than feedback training in Posttest 1.


Figure 4. The median utility across the participants plotted as a function of the median cost in Posttest 1 in each of the four training conditions for the participants able to (app.) satisfy the budget (cost between 62 and 72 SEK/day). The vertical line represents the budget, while the curve is the maximum utility obtainable as a function of the cost. The percentages in parenthesis refer to the proportion of participants in each condition that satisfied the budget constraint.

All training conditions reduced their cost from Posttest 1 to Posttest 2 (new budget) (Wilcoxon test: all $p$ s $<$ .001), and reduced their utility (Wilcoxon test: all $p \mathrm{~s}<$ .001). As shown in Figure 5, in all conditions the median cost exceeded the budget, especially in the rank order condition. There is again modest difference between the other three conditions, although metric mapping exceeds the budget more than the other conditions.

There was a significant difference between the conditions in median cost at Posttest 2 (Kruskal-Wallis test: $H(3, N=129)=17.606 p<.001)$, where pairwise multi-ple-comparisons indicate significant differences between rank-order mapping and feedback training ( $p<$ .001), between rank-order and causal mapping ( $p=$ .033), and between feedback training and metric mapping ( $p=.045$ ). There was also a significant difference between the conditions in median utility at Posttest 2 (Kruskal-Wallis test: $H(3, N=129)=11.800 p=.008)$, where pairwise multiple-comparisons indicate a significant difference only between rank-order mapping and feedback training ( $p=.011$ ). The differences between rank order and metric mapping ( $p=.061$ ) and between rank order and causal mapping ( $p=.050$ ), however, ap-
proach significance. Rank order mapping produced a higher utility (but exceeds the budget). The rank-order condition thus provides the poorest performance, while, among the other three conditions, metric mapping seems to suffer most going from Posttest 1 to Posttest 2.


Figure 5. The median utility across participants plotted as a function of the median cost in Posttest 2 (new budget) in each of the four training conditions $(N \approx 30)$. The vertical line represents the budget, while the curve is the maximum utility obtainable as a function of the cost.


Figure 6. The median utility across the participants plotted as a function of the median cost in Posttest 2 (new budget), in each of the four training conditions for participants that were app. able to satisfy the budget (cost between 45 and 55 SEK a day). The vertical line represents the budget, while the curve is the maximum utility obtainable as a function of the cost. The percentages in parenthesis refer to the proportion of participants in each condition that satisfied the budget constraint.

Figure 6 shows the results of a stricter analysis only including participants with a cost falling within $\pm 5$ units of the budget ( $50 \mathrm{SEK} /$ day). This figure also illustrates that metric mapping training suffered more in the generalization test, with only $22 \%$ of the participants satisfy-
ing the new budget, in contrast to $41 \%$ with feedback and $45 \%$ with causal mapping training. In sum: the rank order condition again produced poor performance, while feedback training and causal mapping appear to allow better generalization of the knowledge obtained to satisfy also a new budget, as compared to metric mapping.

## Discussion

The results indicate that at an immediate test of performance direct mapping training is as good as or better than the detailed and frequent feedback in a typical "smart-meter". It should be noted that the performance of the participants in the direct mapping condition is quite impressive, with a cost virtually exactly on the budget and higher utility, despite never receiving any feedback about the total monthly cost in the house. An objection, of course, could be that these participants did not really learn anything from the metric mapping training, they just happened to be right because their prior conceptions about electricity consumption happened to be correct in regard to the simulated household (which is intended to be "realistic"). That all training groups changed their behavior significantly from the pretest to the posttest to accommodate the budget speaks against this explanation. Participants clearly learned to satisfy the monthly budget from metric mapping training, despite that they never received any feedback about it.

The good performance with metric mapping is in line with the results in Pachur and Olsson (2012), where the participants with direct criterion learning performed better than learning by comparison in a non-linear context due to the exemplar strategy. In the context of their interpretation, this suggests that our participants relied on exemplar memory rather than cue abstraction. An obvious question, in that case, is how people generate exemplar representations of the complex stimuli used in our experiment. On plausible possibility, perhaps, is that they rely on "exemplars" in the form of partial configurations of electricity consumption that were associated with very successful (or unsuccessful) performance.

When generalizing to another budget, the feedback group performs somewhat better than the metric mapping group. This result was unexpected and we can only speculate as to what explains this difference. One possibility is that participants with metric mapping relied on a more exemplar-based strategy, which is known to offer less flexible generalization than the more analytic knowledge of cue-criterion relations. Another possible explanation could be the testing effect (Roediger \& Karpicke, 2006). Tests enhance retention more than additional study of the material, even when tests are given without feedback. In that perspective, the feedback training could be seen as a first test, and the post tests as yet another tests. Another interpretation could be that feedback training and causal mapping provided the participants with better knowledge of the underlying causes, that in turn, facilitated the generalization task.

In this task we found no evidence that people are especially apt at learning rank orders from pairwise comparison (Stewart et al., 2006), considering that the rank
order condition allowed few participants to satisfy the budget. In Posttest 1, only 27\% of the participants satisfied the budget criterion (and only $9 \%$ in Posttest 2). It might be the case that metric information is more crucial in the cost-benefit optimization task in our experiment.

To further investigate feedforward training, future experiments will explore the flexibility with which metric mapping and feedback training can adapt to new budgets and compare mapping involving different metrics (e.g., metrics instead referring to negative environmental effects). Also, it would be interesting to investigate the testing effect, and how it pertains to this context.

The results reported in this study opens for the possibility that shorter and cheaper feedforward training, for example, involving a 15 -minute session with a computer program, can be a cost effective alternative to the large scale implementation of complex information technology to monitor consumption and cost in real time.

## Acknowledgement

This work was supported by the Swedish strategic researchprogram StandUp for Energy (see http://www.standupforenergy.se).

## References

Basso, D. \& Olivetti Belardinelli, M. (2006). The role of the feedforward paradigm in cognitive psychology. Cognitive Processes, 7, 73-88.
Faruqui, A., Sergici, S., \& Sharif, A. (2010). The impact of informational feedback on energy consumption - survey of the experimental evidence. Energy, 35, 1598-1608.
Guath; M., Millroth, P., Elwin, E., \& Juslin, P. (2012). ReEvaluating Smart Meters with Detailed and Frequent Feedback of Electricity Consumption. Manuscript submitted for publication. Department of Psychology, Uppsala University, Sweden
Holyoak, K. J., \& Cheng, P. W. (2011). Causal learning and inference as a rational process: The new synthesis. Annual Review of Psychology, 62, 135-163.
Juslin, P., Karlsson, L., \& Olsson, H. (2008). Information integration in multiple-cue judgment: A division-of-labor hypothesis. Cognition, 106, 259-298.
Kalish, M. L., Lewandowsky, S., \& Kruschke, J. K. (2004). Population of linear experts: Knowledge partitioning and function learning. Psychological Review, 111, 1072-1099.
Klopfert, F. \& Wallenborn, G. (2011). Empowering Consumers Through Smart Metering. Report for BEUC, The European Consumer Organization.
Kluger, A. \& DeNisi, A. (1996). The Effects of Feedback Interventions on Performance: A Historical Review, a MetaAnalysis, and a Preliminary Feedback Intervention Theory. Psychological Bulletin, 2, 254-284.
Nosofsky, R. M. (1986). Attention, Similarity, and the Identi-fication-Categorization Relationship. Journal of Experimental Psychology: General, 115, 39-57.
Pachur, T. \& Olsson, H. (2012). Type of learning task impacts performance and strategy selection in decision making, Cognitive Psychology, 65, 207-240.
Roediger, H.L., III and J.D. Karpicke (2006). The power of testing memory: basic research and implications for educational practice. Perspectives on Psychological Science 1(3): 181-210.
Stewart, N., Chater, N., \& Brown, G. D. A. (2006). Decision by sampling. Cognitive Psychology, 53, 1-26.

# Fairness overrides reputation: The importance of fairness considerations in altruistic cooperation 

Şule Güney (s.guney@unsw.edu.au)<br>School of Psychology, University of New South Wales<br>Sydney, Australia<br>Ben R. Newell (ben.newell@unsw.edu.au)<br>School of Psychology, University of New South Wales<br>Sydney, Australia


#### Abstract

Behavioural findings in several strategic games indicate that people punish others if they think they are being treated 'unfairly' even at the cost of minimizing their own material payoff. We investigated the primary driving force behind such altruistic cooperation. In Experiment 1, we replicated previous findings indicating that the key mechanism contributing to the emergence of altruistic cooperation is fairness considerations. In Experiment 2, we investigated the effect of the opportunity for reputation building and future interaction on altruistic cooperation and found that these factors become effective only when fairness considerations are removed.


Keywords: altruistic cooperation; mini ultimatum game, fairness, reputation building, future interaction.

## Introduction

Human altruistic cooperation presents a puzzle from the perspectives of both the standard economic models of the 'self-interested actor' and the evolutionary models of the 'self-regarding individual' because it involves some characteristics that are difficult to reconcile with the predictions of standard game theoretical and evolutionary analyses. In particular, these characteristics are rewarding the cooperators (i.e., altruistic rewarding) and punishing the norm violators (i.e., altruistic punishment), at a personal cost, even though the probability that this cost will be repaid (either by third parties or by that specific agent in the future) is very low (Gintis et al., 2003).

Evidence for the existence of altruistic cooperation largely comes from laboratory experiments in which the respective behavioral pattern has been observed through economic games. One of the best-known economic games used to demonstrate altruistic cooperation (especially, altruistic punishment) is the Ultimatum Game (UG) (Güth, Schmittberger, \& Schwarze, 1982), in which two players are presented with a sum of money, and one of them is assigned to the role of Proposer while the other one to the Responder. The Proposer is asked to offer any portion of the given money to the Responder. If the Responder accepts the amount offered by the Proposer, the money is distributed in accordance with the proposal. If the Responder rejects the offer, both get nothing. According to standard economic theory of self-interest, a rational Proposer offers the minimum possible amount, and a rational Responder never rejects any amount unless it is zero (Binmore, 2007). The
underlying assumption in this prediction is that both parties care only about how much money they get. However, the vast majority of experimental studies has shown that the modal offers by the Proposers lie between $40 \%-50 \%$ of the total amount and the Responders frequently reject offers below 25\% (Güth et al., 1982; Roth, 1995; Henrich et al., 2005). This pattern of results has been replicated crossculturally (Henrich et al., 2005) and shown to be robust with large stakes (Cameron, 1999). The experiments reported here aimed to investigate the role of several factors (i.e., fairness considerations vs. perceived opportunity of reputation building and future interaction) that might contribute to the emergence of altruistic cooperation in experimental contexts.

## Altruistic cooperation as a function of fairness considerations

Some researchers argue that the underlying mechanism of such non self-regarding behaviors (i.e., high offers by the Proposers and frequent rejections by the Responders) in the UG is not to get as much money as possible, but to maintain fairness norms among players (Gintis et al., 2003; Fehr \& Gachter, 2002). In fact, the motivation behind the Proposers' high offers can be explained with or without the involvement of fairness considerations: They simply may not want to offer an amount that can be easily turned down by the Responder, so they are willing to distribute the money in a relatively fair way. Thus the Proposers' main concern still might be getting as much as possible in the end, rather than treating the Responders fairly (Declerck et al., 2009). However, for Responders, the role of fairness concerns is more apparent and must be stronger because they seem to accept ending up with nothing rather than being treated unfairly. Even though the Responders could have been better off by accepting any amount offered, they prefer to punish the Proposer' unfairness, at a cost to themselves. This pattern of response indicates that the Responders engage in altruistic punishment in response to the unfairness of the Proposer.
A special version of UG has been used to demonstrate how much the Responders care about unfair acts of the Proposers. The structure of the so-called Mini UG (see Table 1) is the same as the standard UG, with an exception: The Proposer is again asked to distribute an amount of money but unlike the standard UG, only in one of two ways.

Table 1. General structure of Mini Ultimatum Games.

| Mini Ultimatum Games* |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | (5/5) Game |  | (2/8) Game |  | (10/0) Game |  | (8/2) Game |  |
| Possible distributions | (8/2) | (5/5) | (8/2) | (2/8) | (8/2) | (10/0) | (8/2) | (8/2) |
| Perceived fairness of the (8/2) distribution |  | Unfair | Reason | nfair ** |  |  |  |  |

* The numbers in the parentheses denote how much the Proposer could get/how much the Responder could get.
** The Proposer seems to have an excuse for offering the more inequitable distribution (8/2), because otherwise he would be unfair to himself [i.e., by offering the (2/8) distribution, he would give 8 to the Responder, and take 2 himself].

Both players participate in four consecutive Mini UGs, and throughout all these games one way of distribution is always fixed while the alternative distribution is always different across games. The fixed distribution is a relatively inequitable one (i.e., the Proposer can take $\$ 8$ for himself, and offer $\$ 2$ to the Responder, see table 1 ).
However, the available alternative distribution varies in terms of the outcome fairness, sometimes yielding a more equitable outcome (i.e., the Proposer can take $\$ 5$ for himself, and offer $\$ 5$ to the Responder, see table 1), and sometimes yielding an even more unequal outcome (i.e., the Proposer can take $\$ 10$ for himself, and offer $\$ 0$ to the Responder, see table 1). Under the standard assumptions, rejection rates for the fixed distribution (8/2) were expected to be the same regardless of its alternatives, as its monetary value stays unchanged across games (Falk et al., 2003). However, this particular distribution was rejected much more frequently when the Proposer intentionally ignored the more equitable alternative distribution [i.e., the (5/5) distribution] than when he ignored the more unequal alternative distribution [i.e., the (10/0) distribution] (Falk et al., 2003; Sutter, 2007). Thus the rejection decisions made by the Responders seem not to be determined by the absolute amount of the offer (i.e., $\$ 2$ ), but by whether the offer is seen as relatively unfair [i.e., in comparison to (5/5) split] or fair [i.e., in comparison to ( $10 / 0$ ) split]. See table 1 for the perceived fairness of the fixed distribution (8/2) across four games.

These findings indicate that the Responders punish the unfairness of the Proposers by rejecting an amount of money in one case and appreciate the fairness of the Proposer by accepting the very same amount in another case. It has been argued therefore that fairness considerations must be the underlying motive behind altruistic cooperation (Gintis et al., 2003; Fehr \& Gachter, 2002).

## Altruistic cooperation as a function of misperceived opportunity of reputation building and future interaction

Although the importance of fairness considerations in such bargaining games has been widely accepted, the real reasons for altruistic cooperation (i.e., the Responders'
rejection/acceptance behaviors in the UG) have been a source of much debate (Declerck et al., 2009). As mentioned earlier, by rejecting a non-zero offer, the Responders seem to engage in actions that are opposite to their self-interest, in order to maintain the fairness norms between parties. Thus fairness considerations seem to override the self-regarding/rational motives. Confidence in such a conclusion mainly comes from the two critical features of the above-mentioned experiments: The identities of both players are kept hidden (i.e., anonymous) and they will never meet again in another round (i.e., one-shot encounter). These specific features, therefore, eliminate the possibility of reputation building (henceforth, RB) and future interaction (henceforth, FI) as potential sources of this seemingly fairness-driven behavior (Fehr \& Fischbacher, 2003). Any involvement of the possibility of RB and FI would be especially critical in this context because the altruistic behavior obtained in these experiments then could be explained within the boundaries of self-regarding motives: It is rational and adaptive to reject unfair offers if the possibility of re-encountering the same game partner in the future is high enough or if the possibility of building a reputation among other players is at stake. The underlying reason for this claim is that rejecting unfair offers protects the player from being offered with unequal distributions by the same game partner in the future or by third parties, and thus this behavior serves the player's self interest (Burnham \& Johnson, 2005; Hagen \& Hammerstein, 2006).

This argument goes further in the direction that people engage in altruistic cooperation in one-shot and anonymous encounters simply because they confuse the experimental settings with the more familiar environments where interactions are normally repeated and non-anonymous (Burnham \& Johnson, 2005). In fact, the participants might still be responding to implicit cues suggesting that future interaction is possible or that their reputation is at stake. One finding that supports this interpretation is that the presence of eyespots on the computer desktop, which triggers the sense that participants are being watched, leads to increased generosity in another money allocation game (Haley \& Fessler, 2005). Some other studies suggest that even the
perception of being involved in a situation where future interaction and reputation building is possible triggers altruistic cooperation in one-shot, and anonymously played economic games (Kiyanori et al., 2000). Thus behaving in an altruistically cooperative manner in the UGs might not solely result from the concern for the maintenance of fairness norms, but from the misperceived opportunity of reputation building and future interaction (Haley \& Fessler, 2005; Bateson, Nettle, \& Roberts, 2006).

## Present Experiments

Previous studies have already established that the (8/2) distribution is rejected at different levels depending on whether the alternative distributions are perceived as fair or not (i.e., highest rejections observed when the alternative was more equitable). However their findings diverge in terms of rejection rates of the $(8 / 2)$ distribution when the alternative distribution was more inequitable. More specifically, $9 \%$ of the Responders rejected the (8/2) distribution in the (10/0) game in Falk et al.'s (2003) study whereas almost $28 \%$ rejected in Sutter's (2007) study. Considering these differences in previous findings, we found it necessary to re-establish the basic phenomenon observed in the Mini UG (presented in Table 1) in our own subject pool, in Experiment 1.

In Experiment 2, we aimed to understand the combined effect of the real possibility of RB and FI in the Mini $\mathrm{UG}^{1}$. If the real reason behind the rejections in one-shot and anonymously played games is the misperceived possibility of RB and FI (and thus for maximizing the material pay-off, for the maintenance of fairness norms), then an increase in the level of altruistic cooperation should be expected when the actual possibility is added to the context. Although such an additional effect of possibility of RB and FI has not been investigated in the Mini UG, there are two main reasons for expecting such an increase. First, the importance given to equality is expected to be elevated (Rottemberg, 2008) because the fairness norm (i.e., distributing the allocated money evenly) is strengthened in presence of the possibility of RB and FI (Hertel et al., 2002). Second and more importantly, the sanctions inflicted upon the unfairness of a game partner through altruistic cooperation might be considered as an effective tool for maximizing future gains (Kiyanori et al., 2000). In addition, there were two main reasons for using the Mini UG, instead of the standard UG: First, its structure would allow us to see how the possibility of RB and FI, along with the fairness concerns, would contribute to the Responders' rejections especially when altruistic punishment (i.e., when the alternative offer yielded a more equitable distribution) is

[^76]expected to take place. Second, in the Mini UG, there is one special game [the ( $8 / 2$ ) game, see table 1] in which the Proposer has no choice, but to offer the fixed amount. This particular case would enable us to detect the sole effect of the possibility of RB and FI on the Responders' decisions when an unequal distribution was offered without any (un)fair intentions of the Proposer involved.

## Experiment 1

We expected the rejection rate of the $(8 / 2)$ distribution to be different across different Mini UGs. More specifically, the highest rejection rate expected to be in the $(5 / 5)$ game. In addition we expected to find statistically significant differences between the rejection rates of the $(8 / 2)$ distribution in the $(5 / 5)$ and the $(10 / 0)$ games.

## Method

Participants: Fifty first year psychology students ( M age $=$ 19.5, 36 female) at UNSW participated in the experiment as a part of their course requirement, and were informed that they would be paid, contingent on the outcome of their choices.
Procedure: There were 10 experimental sessions in total, and 5 participants were tested at a time in each experimental session. Participants were seated in separate rooms and their identities were kept hidden throughout the whole experiment. All participants played the Mini UG as the Responders since our main interest was to see whether we would be able to replicate the choice pattern of the Responders obtained in previous studies (i.e., Falk et al., 2003). However, each participant was told that only one participant in each group of 5 would be assigned to the Responder role and that the rest would be playing as Proposers. This procedure made them believe that the offer in each game would come from an actual but different participant (Proposer) rather than from the computer. The offers made by the computer mimicked the actual rate of proposals offered by real Proposers in the study of Falk et al. (2003). For instance, in that study, the (8/2) distribution was offered by $31 \%$ of the Proposers in the $(5 / 5)$ game, and $73 \%$ in the $(2 / 8)$ game. Thus the Responders in Experiment 1 were offered (8/2) distribution with the probability of .31 in the $(5 / 5)$ game, and that of .73 in the $(2 / 8)$ game. The participants played the games for real money, but currency was defined as Monetary Unit (MU), where 1 MU was equal to 0.5 AUD. The experiment was conducted and run with the Runtime Revolution Software.
Design: The Responders participated in all four Mini UGs presented in Table 1. They were asked to indicate their acceptance/rejection decisions for each of the two possible distributions in each game before hearing the actual distribution offered [i.e., the strategy method was used, see Falk et al., (2003) for further information regarding this method]. For example, in the (10/0) game, the Responders were asked whether they would accept or reject if the Proposer offered them the (10/0) distribution

Table 2. Rejection rates of (8/2) distribution across games in Experiment 1 and 2.

|  | Rejection rates of $(8 / 2)$ distribution |  |  |  |
| :--- | :---: | :---: | :---: | :---: |
|  | $(5 / 5)$ Game | $(2 / 8)$ Game | $(10 / 0)$ Game | $(8 / 2)$ Game |
| Experiment 1 | $60 \%$ | $42 \%$ | $18 \%$ | $14 \%$ |
| Experiment 2* | $52 \%$ | $41 \%$ | $18 \%$ | $50 \%$ |

* Rejection rates reported for Experiment 2 were averaged across rounds.
instead of $(8 / 2)$; and they were subsequently asked whether they would accept or reject if the Proposer offered the $(8 / 2)$ distribution instead of $(10 / 0)$. If the game was ( $8 / 2$ ), they were simply asked what they would do if the Proposer had no choice but to offer the $(8 / 2)$ distribution. Once the Responders indicated their rejection/acceptance decision for each possible distribution, they simply moved on to the next game. After the completion of all four games, the Responders were informed about the overall outcomes and debriefed about the real set-up of the experiment (i.e., the offers were not made by actual proposers). The presentation order of the Mini UGs and that of the possible distributions in each game were randomized.


## Results

Table 2 (top row) shows the rejection rates of (8/2) distribution in different games. The main pattern observed in the previous studies (i.e., Falk et al., 2003; Sutter, 2007) was replicated. To test the overall rejection rate differences across four games, we ran Cochran's Q test. The test confirmed that the rejection rates of the $(8 / 2)$ distribution were significantly different across four games ( $\mathrm{p}<.0001$ ). The rejection rate of the $(8 / 2)$ distribution in the $(5 / 5)$ game was the highest among four games. McNemar change tests were performed for the pairwise comparisons and it showed that the rejection rate in the $(5 / 5)$ game was significantly higher than that of the (10/0) ( $\mathrm{p}<.0001)^{2}$. These results confirmed the previous findings that the rejections to an (unfair) offer were indeed not determined by the absolute amount of money, but by how fair or unfair that offer was perceived in comparison to the other available offers.

## Experiment 2

In order to test the effect of the possibility of RB and FI we changed the structure of the Mini UG from being oneshot and anonymously played to being iterated and nonanonymously played. We predicted that the rejection rates of the $(8 / 2)$ distribution in the Mini UG should be (i) even higher when its alternative was the $(5 / 5)$ distribution

[^77]because it is adaptive to build the reputation that one is a tough bargainer who rejects unfair offers, and (ii) even lower when its alternative was the $(10 / 0)$ distribution because it is adaptive to give the message for future interactions that one is capable to discern and will reward fair intentions.

## Method

Participants: Ninety-six first year psychology students (M age $=19.63$, 62 female) at UNSW participated in the experiment as a part of their course requirement and were informed that they would be paid depending on the outcome of their choices. Four participants were tested in each experimental session and there were 24 sessions in total.

Instructions phase: First, the participants were randomly allocated to their roles, (with 2 being Proposers, and the other 2 being Responders) and warned against revealing their allocated roles to the others. Individual players were then given detailed verbal instructions (along with a written instructions document) regarding the general structure of the game play, what their roles required them to do, and what the consequences of their accept/reject decisions would be. They were specifically informed that they would play the game for more than one round with the same partner, and that their decision would be announced to other players before they switched their partners. However, the players were not given any information about how many rounds they would play in total (i.e., in order to make the 'shadow of the future' long enough), when exactly they would switch partners (i.e., in order to make the possibility of RB stronger). In order to eliminate a potential wealth effect, the participants were told that the overall amount that they would receive would be determined by a coin flip at the end of the experiment. If the coin toss came up heads, then they would get paid the amount that they earned in the first half of the experiment, and if tails, the amount earned in the second half. Afterwards, the instructions documents were collected, and the players were taken to the separate rooms to complete a short quiz measuring whether all the instructions were understood clearly.

Design: Each experimental session consisted of 4 consecutive rounds and in each round the participants played a different Mini UG game [i.e., the (5/5) game in Round 1, the (8/2) game in Round 2 and so on. Note that the allocation of the games into particular rounds was randomized]. Each player was matched with his/her first game partner (i.e., Proposer 1 with Responder 1) before

Round 1 and played two consecutive rounds (e.g., Round 1 and Round 2) with the same partner. After the completion of Round 2, they switched their partners (i.e., Proposer 1 started playing with Responder 2) and played the following 2 rounds (Round 3 and Round 4) with their new partners. At the end of each round, decisions of both players (and the resulting outcomes) were announced to the players. These announcements were done privately (i.e., only between the pairs) after Round 1 and after Round 3; but publicly (i.e., to all players) after Round 2 and Round 4. For example, the decisions of Responder 1 and Proposer 1 were announced only to these two players after they completed Round 1, but their overall decisions in Round 1 and Round 2 were announced to all players just before they switched their partners.

Game play: In all Mini UGs, the Proposer was asked to choose one of the two available distributions (see Table 1). Simultaneously the Responder, without knowing what the Proposer actually had chosen to offer, was asked to indicate his/her acceptance/rejection decisions for each of the two possible distributions. (If the Responder had accepted the offer that the Proposer had actually chosen, the amount was distributed in accordance with the proposal. Otherwise, both got nothing). Both players were informed about the outcome right after the game was over, and then they moved on to the next game. The currency in the experiment was defined in Monetary Units (MU), where 1 MU equals . 5 AUD. The experiment conducted and run with z-Tree (Fischbacher, 2007). After the game play was over, both players received a questionnaire. The Proposers were asked to indicate why they offered the amount they offered and the Responders were asked why they rejected/accepted the (8/2) distribution.

## Results

All participants passed the quiz distributed before the game play, thus all responses were included in the analysis. Table 2 (the bottom row) presents the overall rejection rates of the $(8 / 2)$ distribution in different games. The highest rejection rate was obtained in the $(5 / 5)$ game and the lowest in the (10/0) game. These rejection rates of the (8/2) distribution were significantly different across four groups ( $\mathrm{p}=.0011$, Cochran's Q test). Interestingly, half of the participants rejected the $(8 / 2)$ distribution in the $(8 / 2)$ game. McNemar change tests indicated that the rejection rate in the $(5 / 5)$ game was significantly higher than that in the $(10 / 0)$ game, $p=.0006$ but not than those in the $(2 / 8)$ and the $(8 / 2)$ games, $p=.30$, and $\mathrm{p}=.83$, respectively ${ }^{3}$.

A cross-experimental comparison demonstrated that the rejection rates of the $(8 / 2)$ distribution between Experiment 1 and Experiment 2 did not significantly differ in the $(5 / 5)$ games $\left[\chi^{2}(1, \mathrm{~N}=98)=.62, \mathrm{p}=.43\right]$, the $(2 / 8)$ games $\left[\chi^{2}(1, \mathrm{~N}=98)=.00, \mathrm{p}=.97\right]$, and the $(10 / 0)$ games $\left[\chi^{2}(1, \mathrm{~N}=98)=.01, \mathrm{p}=.92\right]$. Contrary to our expectations, the rejection rates of the (8/2) distribution did not increase when the alternative distribution was
(5/5), and did not decrease when the alternative distribution was ( $10 / 0$ ). However, the ( $8 / 2$ ) distribution was rejected in the $(8 / 2)$ game much more frequently in Experiment 2 than Experiment 1, $\chi^{2}(1, \mathrm{~N}=98)=13.12, \mathrm{p}=$ .0003 . To explore the pattern of results obtained in the (8/2) game in detail, we examined round by round rejection rates. The rejection rates of the (8/2) distribution were especially high in Round 1 and Round 3 in which the first encounters with the game partners took place (see Figure 1). Possible reasons for this special pattern are addressed in the General Discussion section.


Figure 1. Rejection rates of (8/2) distribution in the $(8 / 2)$ game across rounds.

## General Discussion

In Experiment 1, we (re)established the phenomenon that people (negatively) respond to intentional unfairness in a Mini UG at a cost to their own material payoff. Contrary to our predictions, results of Experiment 2 indicated that the additional effect of possibility of RB and FI did not lead to an increase in altruistic cooperation [i.e., rejection rates of the $(8 / 2)$ distribution did not change especially when the Responders were expected to punish unfair offers (i.e., the $5 / 5$ game) or to reward fair offers (i.e., the 10/0 game)]. Two potential but competing explanations could be made. One possibility is that the possibility of RB and FI is indeed (mis)perceived in one-shot and anonymously played games, and thus did not lead to any differences in the pattern of responses when it was explicitly incorporated into the context (Haley \& Fessler, 2005; Bateson, Nettle, \& Roberts, 2006). The other possibility is that the explicit incorporation of the possibility of RB and FI did not have any additional effect on the responses in the presence of the influence of fairness considerations (that are already effective enough to determine the rates of rejection).

Unexpectedly high rejection rates observed in the (8/2) game in Experiment 2 provide supporting evidence for the latter explanation. The possibility of RB and FI changed the responses only in a particular game where the intention of the Proposer was not assessable (the $8 / 2$ game), but not in the other games in which the intentions were assessable (i.e., the $5 / 5$, the $10 / 0$, and the $2 / 8$ games). This pattern of results supports the governing role of fairness considerations in two ways. First, rejection rates of the ( $8 / 2$ ) distribution may have already reached a maximum level in the $(5 / 5)$ game or a minimum in the (10/0) game even in one-shot and anonymously played

Mini UG just as a result of the perceived fairness of the distribution. Thus there was no room for an additional effect induced by the possibility of RB and FI. This is an indication of how dominant the fairness concerns are in determining the level of altruistic cooperation.

Second, finding no evidence for pronounced levels of altruistic cooperation in respective games [i.e., the (5/5), $(2 / 8)$ and $(10 / 0)$ games] might be an indication of the effect of RB and FI being too weak to overcome the effect of fairness considerations. The Responders might only be taking the perceived intentions of the Proposers into consideration as a determinant of their accept/reject decisions for an unequal offer, and thus might not need to have additional reasons/concerns to change those decisions even when RB and FI are possible. However, once the fairness consideration is weakened as a result of the removal of the possible intentions behind an offer in the (8/2) game, the effect of RB and FI becomes effective in changing their responses/concerns: It makes the Responders (negatively) react against the unfairness of the outcome of the (8/2) distribution, most likely, in order to increase the possibility of being treated fairly in the future (Hertel et al., 2002; Kiyonari et al., 2000). The round-wise analysis of the (8/2) game (see Figure 1) confirmed that the increase in rejections (in response to unfair distribution) was indeed resulting from the effect of the possibility of RB and FI. Round 1 and Round 3, in which the highest rejections were observed, were particularly important for the Responders to convey their message for the future encounters. The implicit message given under such condition could be that they don't like to be offered an unequal distribution. The Responders' selfreports collected after the game play also indicate that the main purpose of the rejections in this game was indeed to tell the Proposers that 'I will reject again if you ever propose such an unequal distribution'.

The current set of studies explicitly reveals the importance of fairness considerations in determining the level of altruistic cooperation, especially in the presence of other dominant factors such as the possibility of RB and FI. Demonstrating that these other factors may become effective only in the absence of an important aspect of the fairness concerns [i.e., (un)fairness of intentions] provides a new avenue for the investigation of economic behavior in interactive environments.

## Acknowledgments

This research was supported by a UNSW Overseas Postgraduate Research Scholarship awarded to the first author and Australian Research Council grants (DP110100797; FT110100151) to the second author.

## References

Bateson, M., Nettle, D., \& Roberts, G. (2006). Cues of being watched enhance cooperation in a real world setting. Biology Letters, 2, 412-414.
Binmore, K. (2007). Playing for real: A text on game theory. New York: Oxford University Press.

Burnham, T. C., \& Johnson, D. D. P. (2005). The biological and evolutionary logic of human cooperation. Analyse \& Kritik, 27, 113-135.
Cameron, L. A. (1999). Raising the stakes in the ultimatum game: Experimental evidence from Indonesia. Economic Inquiry, 37, 47-59.
Declerck, C. H., Kiyonari, T., \& Boone, C. (2009). Why do responders reject unequal offers in the Ultimatum Game? An experimental study on the role of perceiving interdependence. Journal of Economic Psychology, 30, 335-342.
Falk, A., Fehr, E., \& Fischbacher, U. (2003). On the nature of fair behavior. Economic Inquiry, 41, 20-26.
Fehr, E., \& Fischbacher, U. (2003). The nature of human altruism. Nature, 425, 785-791.
Fehr, E., \& Gachter, S. (2002). Altruistic punishment in humans. Nature, 415, 137-140.
Fischbacher, U. (2007). z-Tree: Zurich toolbox for readymade economic experiments. Experimental Economics, 10, 171-178.
Gintis, H., Bowles, S., Boyd, R., \& Fehr, E. (2003). Explaining altruistic behavior in humans. Evolution and Human Behavior, 24, 153-172.
Güth, W., Schmittberger, B., \& Schwarze, J. (1982). An experimental analysis of ultimatum bargaining. Journal of Economic Behavior and Organization, 3, 367-388.
Hagen, E. H., \& Hammerstein, P. (2006). Game theory and human evolution: A critique of some recent interpretations of experimental games. Theoretical Population Biology, 69, 339-348.
Haley, K. J., \& Fessler, D. M. T. (2005). Nobody's watching? Subtle cues affect generosity in an anonymous economic game. Evolution and Human Behavior, 26, 245-256.
Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H. et al. (2005). "Economic man" in crosscultural perspective: Behavioral experiments in 15 small-scale societies. Behavioral and Brain Sciences, 28, 795-855.
Hertel, G., Aarts, H., \& Zeelenberg, M. (2002). What do you think is "fair?": Effects of ingroup norms and outcome control on fairness judgments. European Journal of Social Psychology, 32, 372-431.
Kiyonari, T., Tanida, S., \& Yamagishi, T. (2000). Social exchange and reciprocity: Confusion or a heuristic? Evolution and Human Behavior, 21, 411-427.
Roth, A. E. (1995). Bargaining experiments. In J. H. Kagel \& A. E. Roth (Eds.), Handbook of experimental economics (pp. 253-348). Princeton, NJ: Princeton University Press.
Rottemberg, J. J. (2008). Minimally accepted altruism and the ultimatum game. Journal of Economic Behavior and Organization, 66, 457-476.
Sutter, M. (2007). Outcomes versus intentions: On the nature of fair behavior and its development with age. Journal of Economic Psychology, 28, 69-78.

# A Probabilistic Constraint Satisfaction Model of Information Distortion in Diagnostic Reasoning 

York Hagmayer (york.hagmayer@kcl.ac.uk)<br>Olga Kostopoulou (olga.kostopoulou@kcl.ac.uk)<br>Department Primary Care and Public Health Sciences, King’s College London<br>42 Weston Street, London, SE1 3QD


#### Abstract

Information distortion is a cognitive bias in sequential diagnostic reasoning. Assumptions about the diagnostic validity of later evidence are distorted in favor of the leading hypothesis. Therefore the bias contributes to a primacy effect. Current parallel constraint satisfaction models account for order effects and coherence shifts, but do not explain information distortion. As an alternative a new, probabilistic constraint satisfaction model is proposed, which considers uncertainty about diagnostic validity by defining probability distributions over coherence relations. Simulations based on the new model show that by shifting distributions in favor of the leading hypothesis an increase in coherence can be achieved. Thus the model is able to explain information distortion by assuming a need for coherence. It also accounts for a number of other recent findings on clinical diagnostic reasoning. Alternative models and necessary future research are discussed.


Keywords: Diagnostic reasoning; information distortion; parallel constraint satisfaction model.

## Information Distortion in Diagnostic Reasoning

Diagnostic reasoning is an important cognitive activity in many areas. Based on available evidence decision makers infer hidden properties or diagnoses that account for the observations made. Diagnostic reasoning is maybe most important in the clinical domain. Making accurate diagnoses is essential for physicians. Unfortunately clinical diagnostic reasoning is affected by many biases, which may result in medical error (Croskerry, 2003; Kostopoulou et al., 2008). One of these biases is information distortion. When deriving a diagnosis clinicians have been shown to bias their interpretation of newly arriving evidence to support their preferred hypothesis (Kostopoulou, Russo, Keenan, Delaney \& Douri, 2012). More precisely, clinicians alter their assumptions about the diagnostic validity of observed signs and symptoms (i.e., the likelihood of the diagnosis given the sign) so that they lend greater support to the favored diagnostic hypothesis. Similar findings on pre-decisional distortion of evidence have been reported for other professions like sales (Russo et al., 2006). Information distortion has been explained by a need for coherence (Russo, Medvec, \& Meloy, 1996). By interpreting new evidence as supportive of the leading hypothesis decision makers increase the coherence among the favored diagnostic hypothesis and the evidence. Consistency theories in turn account for the need for coherence (cf. Simon et al., 2004).

Parallel constraint satisfaction models, especially Thagard's (1989) ECHO model, have been used to implement coherence-based accounts of diagnostic reasoning (e.g., Gloeckner, Betsch \& Schindler, 2009). These models were extended to sequentially arriving evidence, which affords frequent updating of diagnostic hypotheses (Mehlhorn, Taatgen, Lebiere, \& Krems, 2011; Wang, Johnson, \& Zhang, 2006). Although these models can account for biased decision making, they cannot fully explain information distortion. Constraint satisfaction models in general assume that coherence relations among evidence and hypotheses, which represent assumptions about diagnostic validity, are stable. Research on information distortion, however, shows that decision makers are uncertain about these relations and may change respective beliefs during decision making (Kostopoulou et al., 2012; Russo et al., 1996; 2006). To account for these findings we will put forward a new, probabilistic constraint satisfaction model.

In the paper, we will first briefly describe a recent study on information distortion to exemplify methods and findings. Then we outline a standard constraint satisfaction model of sequential diagnostic reasoning and discuss its shortcomings. Next we propose a new, probabilistic constraint satisfaction model. Results from a simulation study will show that the model predicts information distortion and other findings from the literature. Finally, alternative models will be discussed and necessary future research will be pointed out.

## Exemplary Empirical Findings

Kostopoulou and colleagues (2012) recently published a study on information distortion in the clinical domain. Physicians were confronted with case vignettes presenting diagnostic evidence and asked to evaluate two competing diagnostic hypotheses A and B. Evidence was presented sequentially in a particular order. The first set of cues strongly favored Hypothesis A over B, the next set of cues equally supported both hypotheses, while the third set strongly favored Hypothesis B over A. Overall the evidence was ambiguous. Participating clinicians were asked to make two judgments after each new item: (i) to rate how much this particular item favors either hypothesis (i.e., the item's differential diagnostic validity), and (ii) to rate the likelihood of the diagnoses given all information received so far. Both ratings were made on a scale ranging from one hypothesis to the other. In addition, a control group of
physicians rated each item individually. Information distortion was calculated by computing the difference between individual cue ratings and mean control ratings. From a normative perspective, no information distortion should be expected as the diagnostic validity of individual cues should be constant. Hence, any changes in assumptions about diagnostic validity, which create additional support for the favored hypothesis, constitute a bias.

Three findings are important for the purpose of this paper (see Kostopoulou et al., 2012, for complete results). First, there was a substantial variation between clinicians with respect to the assumed diagnostic validity of cues, which indicates that clinicians were uncertain about how much each piece of evidence supported the hypotheses. Second, participants exaggerated or reduced the diagnostic validity of individual items to support the initially preferred hypothesis. This was especially true for the neutral cues. Third, a majority (56\%) kept the initially preferred hypothesis, while $38 \%$ switched to the hypothesis favored by the evidence coming in last. Only 6\% correctly judged the hypotheses as equally likely. A good model should be able to account these findings.

## Constraint Satisfaction Models of Diagnostic Reasoning

There are many cognitive models to describe sequential hypothesis testing, including Bayesian and logical accounts. We focus on parallel constraint satisfaction models here as they have been very successful in modeling sequential diagnostic reasoning. They also account directly for the frequently found primacy, recency and coherence effects (Mehlhorn et al., 2011, Wang et al., 2006). Thirdly, they are supported by consistency theories, which provide a psychological plausible explanation for why people strive for coherence (Simon et al., 2004).

Many constraint satisfaction models are based on ECHO, a connectionist model of the theory of explanatory coherence (Thagard, 1989). The theory assumes that the acceptance of a belief depends on its relations to other beliefs. By accepting and rejecting beliefs, the overall coherence of the belief set can be maximized. Roughly speaking, a set of beliefs is coherent, if (i) beliefs connected by a positive link (i.e., mutual support, consistency or entailment) are both accepted or rejected, and (ii) only one of the beliefs connected by a negative link is accepted (see next paragraph for formal details).

ECHO has been implemented as a connectionist network (see Figure 1). Hypotheses and items of evidence are represented by nodes, while coherence relations are represented as symmetrical links. Hypotheses and evidence are connected by links with positive weights if they are coherent with each other (e.g., if the evidence is a diagnostically valid indicator), by negative links if they are incoherent (e.g., if the evidence indicates the absence of the diagnosis), or they are not related if they irrelevant for each other. Evidence nodes are assumed to have a special status as their acceptance not only depends on coherence with
other beliefs but on observations. Therefore they receive external activation from a special activation unit (not shown in Figure 1). Evidence nodes in turn activate potential diagnoses. Hypotheses coherent with the evidence get positive activation, while contradicted hypotheses are negatively activated. Different hypotheses are assumed to compete in explaining the observations. Therefore they are negatively related. Activations are passed through the network and added to each other until a stable state is reached. More precisely, the activation of each unit $j$ is updated by combining its current activation $\mathrm{a}_{\mathrm{j}}(\mathrm{t})$ with the net effect (net ${ }_{j}$ ) of all the units i connected to it according to the following formalism (Thagard, 1989; see also McClelland \& Rumelhart, 1981):

$$
\begin{aligned}
& \mathrm{a}_{\mathrm{j}}(\mathrm{t}+1)=\mathrm{a}_{\mathrm{j}}(\mathrm{t})^{*}(1-\Theta)+\left(\text { net }_{\mathrm{j}} *\left[\text { max }-\mathrm{a}_{\mathrm{j}}(\mathrm{t})\right] \text {, if net } \mathrm{t}_{\mathrm{j}}>0\right. \\
& =\mathrm{a}_{\mathrm{j}}(\mathrm{t}) *(1-\Theta)+\left(\text { net }_{\mathrm{j}} *\left[\mathrm{a}_{\mathrm{j}}(\mathrm{t})-\mathrm{min}\right] \text {, if net }{ }_{j} \leq 0\right. \\
& \text { with net }{ }_{j}=\Sigma_{\mathrm{i}} \mathrm{rel}_{\mathrm{ij}} * \mathrm{a}_{\mathrm{i}}
\end{aligned}
$$

The parameter $\Theta$ represents a decay and min and max the maximum and minimum activation (usually 1 and -1 ). Final activations represent acceptance. Hence, the hypothesis, which receives the highest positive activation in the end, is the preferred diagnosis.

The coherence of a belief set can be calculated by summing up the products of final activations and relations. This measure has been called harmony (Thagard, 1989).

$$
\text { Harmony }=\Sigma_{\mathrm{i}} \Sigma_{\mathrm{j}} \operatorname{rel}_{\mathrm{ij}} * \mathrm{a}_{\mathrm{i}} * \mathrm{a}_{\mathrm{j}}
$$

To account for sequentially arriving evidence, the external activation of evidence nodes is assumed to shift towards the new arriving evidence (Wang et al., 2006). In line with findings on the limited capacity of attention, the received activation is kept constant and is decayed exponentially across items. The activation received by an item of evidence is calculated according to the following equation:

$$
\mathrm{a}_{\mathrm{ev}}=\mathrm{a}_{\mathrm{ev}}^{*} \exp (-1 * \Lambda[\text { Number of subsequent items seen }])
$$

The parameter $\Lambda$ represents how strongly the activation of an item is decayed due to later items. Little or no decay results in primacy effects, i.e., the first evidence biases decisions in favor of the initially preferred hypothesis. Strong decay leads to recency effects (Wang et al., 2006).


Figure 1. Parallel constraint satisfaction model of sequential diagnostic reasoning. Nodes represent hypotheses (HypA/B) and pieces of evidence (e.g., ProA). Solid lines represent coherent, dashed lines incoherent relations. Pieces of evidence arrive sequentially along the time line.

Figure 1 shows the structure of a constraint satisfaction network with two competing hypotheses (Hyp A, Hyp B) and nine pieces of evidence. The first three observations (ProA 1-3) support Hypothesis A (indicated by the solid lines) and contradict Hypothesis B (indicated by the dashed lines). The next observations (Ambig 1-3) support both hypotheses, while the final set favors Hypothesis B over A (ProB 1-3). This is the order of evidence clinicians received in the study by Kostopoulou et al. (2012).

This model predicts that Hypothesis A will be favored over Hypothesis B unless there is a very strong decay of the initial evidence (see simulations for respective evidence). But it cannot explain information distortion. As outlined above, information distortion means that assumptions about the diagnostic validity, i.e., the relations between evidence and hypotheses are distorted. The model presented here keeps these relations constant assuming that decision makers have stable beliefs about coherence relations. Hence the model cannot account for the findings by Kostopoulou and colleagues (2012) that participating clinicians distorted their assumptions about diagnostic validity for a particular case depending on their currently favored hypothesis.

## A probabilistic constraint satisfaction model

Decision makers may be uncertain about the coherence relations among evidence and hypotheses. Consider the medical domain. Although a particular diagnostic cue may have a positive predictive value for Diagnosis A, there will be cases in which another diagnosis will prove to be true. Standard constraint satisfaction models of diagnostic reasoning do not allow us to represent this uncertainty. This uncertainty can be captured by conceptualizing the relations connecting evidence and hypotheses as beliefs and defining probability distributions over these beliefs. Probability distributions are used to represent the uncertainty in many cognitive models, e.g., Bayesian models (Chater \& Oaksford, 2008), but they have not been used in constraint satisfaction models so far. Nevertheless, their application seems straightforward. There are three types of coherence relations: positive links, negative links, and no links (representing irrelevance). The probability distribution defines the likelihood that evidence and hypothesis are connected by a positive, a negative or no relation.

For example, to represent the assumption that a piece of evidence X almost always supports a Hypothesis A the probability of a positive link between $X$ and $A$ is set to a high value (i.e., $\mathrm{P}\left(+_{A x}\right) \approx 1$ ) while the probabilities of a negative or no link are assumed to be very small (i.e., $\mathrm{P}(-\mathrm{ax})$ $\left.\approx 0, \mathrm{P}\left(0_{\mathrm{AX}}\right) \approx 0\right)$. To derive predictions for a particular probability distribution, a set of constraint satisfactions networks is instantiated and run. Based on the resulting activations of the networks the likelihood that each hypothesis will receive the highest final activation is estimated. In addition, the mean resulting harmony is calculated to estimate the expected overall coherence.

Like standard parallel constraint satisfaction models the probabilistic models can account for primacy and recency effects by assuming differential decay of sequentially arriving information. Moreover, they may also account for information distortion. By shifting the probability distribution over coherence relations the overall coherence (i.e., harmony) may be increased. Thus a need for coherence may cause a change in beliefs about coherence relations resulting in information distortion. There is a limit however. To preserve the belief that a certain piece of evidence is coherent with a hypothesis in general, the probability distribution can only be shifted to a certain degree. To be more precise, the sign of the sum of weights of the relations multiplied with their respective probabilities has to remain the same. For example, if Hypothesis A and Evidence X are assumed to be coherent $\Sigma \mathrm{P}\left(\right.$ relation $\left._{\mathrm{AX}}\right){ }^{*}$ relation $_{\mathrm{AX}}>0$. Thus probabilistic constraint satisfaction models may predict information distortion without assuming an outright change in beliefs about the diagnostic validity of cues.

## Simulations

To explore the predictions of probabilistic constraint satisfaction models, we implemented the model shown in Figure 1 with various probability distributions over coherence relations (see Table 1). The overall relation between each piece of evidence and the two hypotheses was kept the same across all distributions. The first three pieces of evidence were generally coherent with Hypothesis A and incoherent with $B$, the ambiguous evidence supported both hypotheses, and the final set contradicted A and favored B.

Table 1: Probability distributions over coherence relations of the model depicted in Figure 1.

| Relation | $\begin{aligned} & \text { HypA - } \\ & \text { Pro A1-A3 } \end{aligned}$ |  |  | $\begin{aligned} & \text { HypB - } \\ & \text { Pro A1-A3 } \end{aligned}$ |  |  | $\begin{aligned} & \text { HypA - } \\ & \text { Amb1-Amb3 } \end{aligned}$ |  |  | $\begin{aligned} & \text { HypB - } \\ & \text { Amb1-Amb3 } \end{aligned}$ |  |  | $\begin{aligned} & \text { HypA - } \\ & \text { Pro B1-B3 } \end{aligned}$ |  |  | $\begin{aligned} & \text { HypB - } \\ & \text { Pro B1-B3 } \end{aligned}$ |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Distrib. | P(+) | P(-) | $\mathrm{P}(\mathrm{o})$ | $\mathrm{P}(+)$ | P(-) | $\mathrm{P}(\mathrm{o})$ | $\mathrm{P}(+)$ | P(-) | $\mathrm{P}(\mathrm{o})$ | $\mathrm{P}(+)$ | P(-) | $\mathrm{P}(\mathrm{o})$ | $\mathrm{P}(+)$ | P(-) | $\mathrm{P}(\mathrm{o})$ | P(+) | P(-) | $\mathrm{P}(\mathrm{o})$ |
| M1 fixed | 1 | - | - | - | 1 | - | 1 | - | - | 1 | - | - | - | 1 | - | 1 | - | - |
| M2 | . 9 | - | . 1 | - | . 9 | . 1 | . 9 | . 05 | . 05 | . 9 | . 05 | . 05 | . 05 | . 9 | . 05 | . 9 | . 05 | . 05 |
| M3 ProA | . 9 | - | . 1 | - | . 9 | . 1 | . 9 | . 05 | . 05 | . 5 | . 25 | . 25 | . 3 | . 4 | . 3 | . 4 | . 3 | . 3 |
| M4 ProB | . 9 | - | . 1 | - | . 9 | . 1 | . 5 | . 25 | . 25 | . 9 | . 05 | . 05 | . 05 | . 9 | . 05 | . 9 | . 05 | . 05 |
| M5 | . 5 | - | . 5 | - | . 5 | . 5 | . 5 | . 25 | . 25 | . 5 | . 25 | . 25 | - | . 5 | . 5 | . 5 | - | . 5 |
| M6 ProA | . 5 | - | . 5 | - | . 5 | . 5 | . 9 | . 05 | . 05 | . 5 | . 25 | . 25 | . 3 | . 4 | . 3 | . 4 | . 3 | . 3 |
| M7 ProB | . 5 | - | . 5 | - | . 5 | . 5 | . 5 | . 25 | . 25 | . 9 | . 05 | . 05 | . 05 | . 9 | . 05 | . 9 | . 05 | . 05 |

Table 2: Results of simulations. Harmony (i.e., degree of coherence of beliefs) after each new piece of evidence and distribution of finally preferred hypotheses
$\left.\begin{array}{lcccccccccccc}\text { Preferred Hypothesis }\end{array}\right)$

The first distribution (M1) was identical to standard models and assumed no uncertainty about the coherence between evidence and hypotheses. The second (M2) closely resembled the standard model and assumed the same relations with a high probability of .9. The third distribution (M3ProA) represents shift of assumptions in favor of Hypothesis A after the first three pieces of evidence. The ambiguous evidence (Amb1-Amb3) is considered less supportive of Hypothesis B, and the evidence clearly favoring Hypothesis B (ProB1-ProB3) as less contradictory for $A$ and less supportive of $B$. The forth distribution (M4ProB) represents a shift in favor of Hypothesis B. Now the ambiguous evidence is considered less supportive for Hypothesis A. If the model adequately captures the predictions of consistency theories, we should see an increase in coherence for M3 over M2 and M4.

The fifth probability distribution over coherence relations (M5) represents another set of basic assumptions. It assumes that all pieces of evidence are considered moderately supportive of the respective hypotheses. Distribution M6ProA again represents a shift of distribution M5 in favor of Hypothesis A while M7ProB represents a shift of M5 in favor of Hypothesis B. A comparison of the results for these distributions will show whether any of these shifts would increase coherence.

Model parameters were set to random values or kept constant for all simulations. Links of coherence had a weight of +.05 , incoherence links of -.05 . The incoherence link between hypotheses was set to -.2. Initial activations of hypotheses were set to random values between -.2 and +.2 . Evidence nodes were added sequentially to the network after activations settled. Resulting activations were transferred to the next step. External activations received by evidence nodes were decayed when new evidence arrived. The decay parameter $\Lambda$ was randomly set to values between 1 (strong exponential decay) and .1 (almost not decay). The activation added through the evidence nodes was kept constant at .5 for all steps. In line with previous studies we found that the qualitative pattern of activations hardly depended on the specific parameters (Thagard, 1989). Therefore only one set of results is reported here.

## Results

For each probability distribution 10.000 constraint satisfaction models were instantiated and run. The results of the simulations are depicted in Table 2. Harmony, i.e., the resulting overall coherence of the belief network, is shown for each new piece of evidence. In addition, the percentage of cases in which Hypothesis A was preferred over B is given. For six out of seven distributions, Hypothesis A was preferred over Hypothesis B. Thus a primacy effect resulted, which is in accordance to the results of Kostopoulou et al., (2012). As expected, decay had a strong impact on results. When the decay was strong ( $\Lambda=1$ ), that is, the last piece of evidence was strongly activated while previous evidence hardly received any activation, a recency effect sometimes occurred and Hypothesis B was favored. When decay was weak ( $\Lambda=.1$ ), that is, initial evidence was activated only slightly less than the latest evidence, a primacy effect resulted even when the distributions were shifted in favor of Hypothesis B. Note that recent research indicates that weak or no decay fits best with people's actual decisions (Mehlhorn et al., 2011).

A comparison of distributions M1 and M2 shows that a probabilistic network with high probabilities basically results in the same overall preferences as a deterministic network which is identical to standard constraint satisfaction models. Overall coherence was only slightly reduced when relations became uncertain. A comparison of distributions M2, M3ProA and M4ProB indicates that the coherence of beliefs increased substantially when the probability distribution over coherence relations was shifted in favor of Hypothesis A, but not when it was distorted in favor of B. Note that an increase in coherence for M3ProA already resulted for the ambiguous items of evidence, after which it stayed at an elevated level. Thus the model predicts information distortion especially for the ambiguous items of evidence. This is what has been found empirically.

A comparison of distribution M5 to distributions M6ProA and M7ProB shows a different picture. Starting from less assertive assumptions about the diagnostic validity of the evidence, more coherence could be gained by shifting
assumptions in favor of Hypothesis B. A closer analysis shows that coherence increased for the ambiguous pieces of evidence by shifting assumptions towards Hypothesis A, but that this gain evaporated when the evidence favoring Hypothesis B arrived. Interestingly, a shift towards B only yielded substantially more coherence for the last few items. Thus the model predicts that people being uncertain should be less likely to distort but more likely to end up choosing Hypothesis B. This is what Kostopoulou and colleagues (2012) found.

## General Discussion

A probabilistic constraint satisfaction model has been proposed to explain information distortion in sequential diagnostic reasoning. The model takes into account that diagnosticians may be uncertain whether a certain piece of evidence supports a diagnostic hypothesis for a particular case. To be more precise, it takes into account that people are aware of the fact that a piece of evidence may not always be present when a diagnosis is given and vice versa. Note that the model like constraint satisfaction models in general does not differentiate between the sensitivity of a diagnostic sign (i.e., the probability of the sign given the diagnosis) and the positive predictive value of the sign (i.e., the probability of the diagnosis given the sign). The model does, however, differentiate between believing a certain piece of evidence and believing that the information has diagnostic implications for a hypothesis.

The model has been implemented by using the standard formalism of ECHO (Thagard, 1989) and an exponential activation decay function to account for the sequentially arriving evidence. Uncertainty about diagnostic relations is represented by probability distributions over coherence relations among evidence and hypotheses. Belief in the observed evidence and hypotheses is represented by activations of the respective nodes.

Simulations were run to investigate the properties of the model and to find out whether it is able to predict findings reported in the literature. An analysis of the predictions of different probability distributions yielded several interesting results. First, the model shows a primacy effect which is reported frequently in the literature when people first receive several pieces of evidence favoring one hypothesis over others (Brownstein, 2003; Hogarth \& Einhorn, 1992). However, many other models predict order effects, so this prediction is not unique.

Second, the model predicts information distortion. The results show that by distorting assumptions about the diagnostic validity of the observed cues, i.e., by shifting probability distributions over coherence relations, more coherent beliefs can be achieved. Importantly, coherence can be increased without giving up general assumptions about the coherence between cues and hypotheses. Thus, the model explains how the need for coherence can drive changes in beliefs about diagnostic validity and why information distortion may result.

Crucially, the simulations also showed that not all changes in beliefs about diagnostic relations may result in higher coherence. They also indicated that a shift in beliefs has an impact on coherence at a particular point during the diagnostic process. Thus the model allows for very specific predictions once initial beliefs about diagnostic relations are known.

## Alternative Models

A number of parallel constraint satisfaction models has been proposed in the literature to account for diagnostic reasoning (e.g., Mehlhorn \& Jahn, 2009; Mehlhorn et al., 2011; Wang et al., 2006; Gloeckner \& Betsch, 2008; Gloeckner, Betsch, \& Schindler, 2009).

The parallel constraint satisfaction model of Gloeckner and colleagues (2009) was designed to account for distortions in validity in multiple-cue judgment. In their research they found that participants changed their assessments of diagnostic validity depending on the favored option for a particular case. Although the model was devised for concurrent, non-sequential decision making it may be extended to cover sequential decision making. The structure of the model is highly similar to the model depicted in Figure 1 with cues being related to two alternative options, which compete with each other. Cues are assumed to be related to an activation unit. But, relations and activations are given an interpretation that is very different from our proposed model. The relation to the activation unit is assumed to represent the general validity of the cue, while the activation of each cue is considered to represent the validity of this cue for the particular case. This model is able to account for many findings in the judgment and decision making literature (cf. Gloeckner et al., 2009). Nevertheless, the model has difficulty to account for information distortion, because it does not differentiate between the validity of a cue and the diagnostic validity of the cue for a particular hypothesis. The results on information distortion (Kostopoupou et al., 2012) show that participants may increase the diagnostic validity with respect to Hypothesis A while decreasing the diagnostic validity with respect to Hypothesis B. The activation of a node, however, cannot increase and decrease at the same time. Thus the activation may represent whether a piece of evidence is considered valid or invalid, but not whether it is considered valid with respect to a diagnosis. The probabilistic constraint satisfaction model allows for this differentiation. Assumptions with respect to diagnostic validity are represented by probability distributions over coherence links. Therefore assumptions about the diagnostic validity with respect to several hypotheses may change independently from each other. Such a probabilistic constraint satisfaction model may account for the findings of Gloeckner et al. (2009). It also predicts that participants would lower their belief in the validity of cues contradicting the preferred option, as the resulting activation for these nodes would be negative.

The constraint satisfaction model proposed by Wang and colleagues UECHO (2006) was specifically developed to capture sequential belief updating and learning with a parallel constraint satisfaction network. The structure of the model is the same as the model shown in Figure 1. As outlined above, the model accounts for sequentially arriving evidence by a decay function over the activation distributed by the special activation unit. We adopted this idea for our model. The second important novel idea of UECHO is that the strength of the coherence links may change due to learning from feedback. We did not consider this idea for two reasons. First, clinicians are very unlikely to change their generic diagnostic knowledge in experimental studies on diagnostic reasoning and information distortion. Second, assigning specific weights to coherence links violates the fundamentally qualitative notion of coherence stressed by Thagard (1989). Either some evidence is coherent or incoherent with a hypothesis, or it is irrelevant. Our probabilistic model keeps this notion by assuming that links are either positive, negative or zero, while at the same time defining a probability distribution over these links representing the idea that evidence may be found even when the coherent hypothesis turns out to be false. Learning from feedback could be added to our model by adding a Bayesian learning algorithm that updates the probability distribution over coherence links. This seems to be a viable and elegant alternative to the proposal of Wang and colleagues (2006). In principle, UECHO may be extended to account for information distortion by assuming that the weight of individual coherence links may change for a particular case (i.e., change without learning). Like we envisioned for our probabilistic model, this shift may be driven by increased overall coherence (i.e., harmony). Such a model, however, would not be able to represent the uncertainty about coherence relations like the probabilistic model does.

## Future Directions

The proposed probabilistic constraint satisfaction model shares important features with other parallel constraint models. In contrast to the other models it explains information distortion. As outlined above, the model allows for a number of specific predictions including conditions under which information distortion should not be found. However, to test these predictions, assumptions about diagnostic relations have to be assessed on an individual level and compared to later measurements of information distortion. Respective research still needs to be done. Hence, only future studies will show whether probabilistic constraint satisfaction models are able to successfully predict preferred hypotheses, information distortion and validity judgments of individual diagnosticians.

## References

Brownstein, A.L. (2003). Biased predecision processing. Psychological Bulletin, 129(4), 545-68.
Chater, N. \& Oaksford, M. (2008). The probabilistic mind. Prospects for Bayesian cognitive science. Oxford: Oxford University Press.
Croskerry, P. (2003). The importance of cognitive errors is medicince and strategies to minimize them. Academic Medicine, 78, 775-780.
Glöckner, A., \& Betsch, T. (2008). Multiple-reason decision making based on automatic processing. Journal of Experimental Psychology: Learning, Memory, and Cognition, 34, 1055-1075.
Gloeckner, A., Betsch, T., \& Schlindler, N. (2009). Coherence shifts in probabilistic inference tasks. Journal of Behavioural Decision Making, 23, 439-462.
Hogarth, R. \& Einhorn, H. (1992). Order effects in belief updating: the belief-adjustment model. Cognitive Psychology, 24, 1-55.
Kostopoulou, O., Delaney, B.C., \& Munro, C.W. (2008). Diagnostic difficulty and error in primary care - a systematic review. Family Practice, 25, 400-413.
Kostopoulou, O., Russo, J.E., Keenan, G., Delaney, B.C. \& Douiri, A. (2012). Information distortion in physicians’ diagnostic judgments. Medical Decision Making. Published online. DOI: 10.1177/0272989X12447241
McClelland, J. L., \& Rumelhart, D. E. (1981). An interactive model of context effects in letter perception. Part 1. An account of basic findings. Psychological Review, 88, 375-407.
Mehlhorn, K., Taatgen, N.A., Lebiere, C., \& Krems, J. (2011). Memory activation and the availability of explanations in sequential diagnostic reasoning. Journal of Experimental Psychology: Learning, Memory \& Cognition, 37, 1391-1411.
Russo, J.E., Meloy, M.G., Wilks, T.J. (2000). Predecisional distortion of information by auditors and salespersons. Management Science, 46(1), 13-27.
Russo, J.E., Medvec, V.H., \& Meloy, M.G. (1996). The distortion of information during decisions. Organizational Behavior and Human Decision Processes, 66, 102-110.
Simon, D., Snow, C. J., \& Read, S. J. (2004). The redux of cognitive consistency theories: Evidence judgments by constraint satisfaction. Journal of Personality and Social Psychology, 86, 814-837.
Thagard, P. (1989). Explanatory coherence. Behavioral and Brain Sciences, 12, 435-502.
Wang, H., Johnson, T.R., \& Zhang, J. (2006). The order effect in human abductive reasoning: an empirical and computational study. Journal of Experimental and Theoretical Artificial Intelligence, 18, 215-247.

# Spatial cognition: the return path 

# Kai Hamburger (kai.hamburger@ psychol.uni-giessen.de) <br> Lena E. Dienelt (lena-eowyn.dienelt@psychol.uni-giessen.de) <br> Marianne Strickrodt (marianne.strickrodt @ psychol.uni-giessen.de) Florian Röser (florian.roeser@psychol.uni-giessen.de) 

Justus Liebig University Giessen, Department of Psychology, Experimental Psychology and Cognitive Science<br>Otto-Behaghel-Strasse 10 F<br>35394 Giessen, Germany


#### Abstract

The cognitive representation of a return path is a rather unexplored topic including different issues, e.g., perception, mental imagery, mental spatial processing, and language. We here investigated the return path with landmarks located on different positions (optimal, suboptimal). Participants learned a total of 24 routes and had to produce the return paths ( $N=20$ ). In a second experiment the different positions plus map learning versus verbal directions were investigated ( $N=20$ ). Both experiments reveal that the position of a landmark at an intersection (structural salience) has an influence on wayfinding performance. However, the results are somehow ambiguous. Therefore, we also present first approaches for predicting behavior (e.g., optimal route descriptions) and for modeling the perceptual and cognitive processes involved in finding the return path, including visibility, structural salience, mental representation/ transformation, and language.


Keywords: return path; structural salience; landmarks; mental transformation

## Introduction

Imagine that you are on a vacation in an unknown foreign city. After your arrival at the hotel you want to explore the surroundings and maybe visit a place of interest or a touristic feature (e.g., a famous building such as the Eiffel Tower in Paris). You may base your search on different means for successfully reaching your goal. You may want to use a verbal description that you received at the reception desk of your hotel, maybe you want to make use of a city map in your tourist guide, or, if you do not have these means at hand, you may want to ask a pedestrian on the street for giving you directions to your goal location. There is also the possibility of using a mobile navigation system. This latter example is part of the debate on "extended cognition" (e.g., Clark \& Chalmers, 1998), which is beyond the scope of this project. Here, the focus is rather on the "innate" navigation system, perceptual and cognitive processes that enable humans to navigate without getting lost (most of the times). In general, wayfinders use so-called landmarks, objects or buildings that stand out of their environment, to aid navigation (e.g., Lynch, 1960; Presson \& Montello, 1988; Caduff \& Timpf, 2008). Let us return to our initial example. One important question is whether the verbal description is on its own sufficient for reaching the
goal without being distracted or being led into a wrong direction? Or, would it be better to supplement the verbal description with a map, or maybe make only use of the map instead? This is not only a question of not getting lost (e.g., Dudchenko, 2010), but also a question about cognitive economy, namely, reaching the goal with the least cognitive or physical effort. Let us assume that we successfully reached the goal. We are now faced with a new, maybe more difficult, problem. We need to return to our hotel!
Finding a return path is an everyday problem but has rarely been investigated empirically (retrace the same route; e.g., Golledge, 1997, Büchner, Hölscher, \& Strube, 2007; Papinski, Scott, \& Doherty, 2009). We are able to manage this task, but we do not yet know the underlying cognitive and neural processes enabling us to find the return path.
One of the most important aspects for the return path is probably the structure of the environment (e.g., structural landmark salience; Sorrows \& Hirtle, 1999; Klippel \& Winter, 2005). Since we assume visual salience (or better perceptual salience) -that is how much an object stands out from its environment (e.g., Caduff \& Timpf, 2008)- and semantic salience of landmarks -that is for example its name, meaning, or function (e.g., Hamburger \& Knauff, 2011)- to be less important, we here try to control for these aspects and rather focus on the structural aspects as we have done in several previous experiments on structural salience (e.g., Röser, Hamburger, Krumnack, \& Knauff, 2012a; Röser, Krumnack, Hamburger, \& Knauff, 2012b).
There are two optimal positions for landmarks to be located on a regular/initial path: before the intersection (Klippel \& Winter, 2005) in direction of the turn and behind the intersection in direction of the turn (Röser et al., 2012a). Most important is that the landmark is located somewhere in direction of the turn (Röser et al., 2012a). But, for the return path, two different positions might be the optimal ones: the positions before the intersection in direction of the turn and behind the intersection opposite to the direction of the turn. These positions are invariant for the return path (they remain unchanged). The other two positions are variant, since they have to be mentally and verbally transformed for the return path (e.g., "before the intersection opposite to the direction of turn" becomes "behind the intersection and in direction of the turn" on the way back). Further details on
this theoretical assumption are provided in the section "Theoretical assumptions, modeling, future research".

## Experiment 1

## Method

## Participants

A total of 20 Psychology students from the University of Giessen participated (16 females). They had a mean age of 23.5 years ( $S D=4.08$ ). All participants were naive with respect to this study, provided informed written consent, and received course credits for participation. They had normal or corrected-to-normal visual acuity and were free of any preexisting psychiatric or neurologic illness (e.g., epilepsy).

## Materials

The equipment included a custom $19^{\prime \prime}$ monitor (Dell), a Personal Computer (HP Compaq 6000 Pro), and a Response Pad (RB-530 Cedrus Corporation©). For presentation and data recording SuperLab 4.0 Stimulus Presentation Software (Cedrus Corporation©) was employed.
The virtual environment (maze) was set up with Google® SketchUp 8 (compare to SQUARELAND; Hamburger \& Knauff, 2011), which in its original version is made of $10 \times 10$ cuboids, representing regular orthogonal intersections, and proofed very flexible in terms of experimental manipulations. Here, 24 routes, each with eight intersections in an egocentric perspective, were created. The directions left or right were used. Every intersection $(24 \times 8=192)$ contained one distinct landmark one of 192 different words on a white sign (Figure 1). These distinct landmark words were used to prevent interferences of previously learned landmark and direction combinations (e.g., in Route 1 you have to turn right when you see the word "horse"; in a later route you might have to turn left when you see the word "horse"). Hence, a landmark which was shown once to the participant does not appear again later in another route. We controlled for all landmarks being comparably imaginable by using familiar, everyday words. A landmark was placed on both sides of the corresponding facades of a corner, so it was visible from both directions of travel.


Figure 1: Screenshot of an intersection in the virtual maze (decision point). The landmark (word; Apfel = apple) is presented on both facades at one corner (position).

To control for direction or landmark position effects, the number of right/left turns and the position of landmarks (before or after the intersection, in or against moving direction) were balanced for single routes. This balancing applies to both the regular travel direction (forward) in the learning phase and to the return path in the wayfinding phase.

## Procedure

Participants learned a route of eight intersections via successively presented pictures of each of the intersections (Figure 1). Every intersection was shown for duration of eight seconds (learning phase). Subsequently, participants were instructed to find the same path again either in the normal (forward from origin to destination) or the reverse travel direction (backwards from destination to origin; wayfinding phase). Every intersection was presented via pictures and served as a decision point (right or left) for which direction decisions had to be made.
After one route was navigated (eight direction decisions), the learning phase of the next route started. No feedback about the decisions was given. The total of 24 routes had to be learned by each participant. Overall, half of the routes had to be found in the forward run direction, while for the other half the return path was required. Therefore, two experimental versions were used where navigation direction in the wayfinding phase was interchanged (e.g., Route 1 had to be found again in forward direction in version 1, but in the backwards direction in version 2). The order of the routes was randomized for every participant. Correct decisions and response times served as dependent variables. At the end participants were asked to indicate any strategies they had used during the experiment.

## Results

The mean correct route decisions on the return path were about $87 \%$ in this experiment (chance level 50\%).
An analysis of variance with repeated measures for the wayfinding phase was performed. Within-subject factors were navigation direction (forward/backward) and landmark position (all four possible positions). Both for correct decisions and response times a significant main effect for navigation direction (correct decisions: $F(1,19)=19.865$, $p<.001$; response times: $F(1,19)=21.571, p<.001)$, but not for landmark position (correct decisions: $F(3,57)=1.020$, $p=.391$; response times: $F(3,57)=.871, p=.461)$ could be found. Participants were better and faster in navigating the original route direction (forward) compared to the reverse direction (backwards), but the position of a landmark did not lead to any performance differences.
A wide range and variability of learning strategies was reported by the participants and different levels of selfconfidence in performance were expressed. Thus, we were interested in possible group differences. Therefore, we divided our sample in participants with an overall better ( $N=14$ ) and an overall worse performance ( $N=6$ ) with respect to mean overall performance. This mean-
performance-grouping now functioned as a between-subject factor in a re-analysis of variance with repeated measures. For response times again only a main effect for direction ( $F(1,18)=16.196, p=.001$ ), but not for landmark position $(F(3,54)=.508, p<.678)$ occurred. However, for correct decisions again a significant main effect for direction $(F(1,18)=22.322, p<.001)$ and an additional interaction of navigation direction $\times$ landmark position $\times$ group $(F(3,54)=3.895, p=.025)$ emerged. This means that the wayfinding performance with landmarks on varying positions differs with navigation direction and depends on the participant being a high or a low performer.

## Discussion

An overall effect for the wayfinding direction could be found. People were faster and better when travelling the route in the originally learned direction (forward) compared to navigating the return path, which is not very surprising but has not been investigated systematically before. No landmark position effect was found. Only the mean performance of low performers indicates that some people (maybe depending on spatial ability and learning strategy) might be affected by structural differences (positions), and that the helpfulness of a landmark might differ depending on the direction of travel (forward, backwards). Such data need to be further analyzed in future research with the focus on individual strategies (wayfinding performance vs. sense-of-direction; e.g., Kato \& Takeuchi, 2003).
In Experiment 2 we used a more realistic setup: video sequence from an egocentric perspective with approximated true physical sizes on a projection screen; only one route but with the option of going straight; two learning conditions; more than just eight intersections, etc.

## Experiment 2

## Method

## Participants

A total of 20 Psychology students from the University of Giessen participated ( 13 females). They had a mean age of 26.1 years ( $S D=9.03$ ). All participants were naive with respect to this study, provided informed written consent, and received course credits for participation. They had normal or corrected-to-normal visual acuity and were free of any preexisting psychiatric or neurologic illness (e.g., epilepsy).

## Materials

The same setup was used as in Experiment 1 but this time the routes were presented on a customary projection screen (171x238 cm) with a projector (Panasonic PT-F100NT).
For the experiment two different routes through the maze, with 20 intersections each, were created. Therefore, a total of 20 different words served as landmark objects (Figure 1). The words were derived from a catalog of pictograms which made them visually similar, realistic, and easy to imagine. In the maze the landmarks were again placed on both
facades of a corner (position), so they were visible from both directions of travel.
Videos of the two routes were generated from an egocentric perspective, with an eye height of 1.70 m and a constant walking speed of about $2 \mathrm{~m} / \mathrm{s}$. For presentation and data recording SuperLab 4.0 Stimulus Presentation Software (Cedrus Corporation©) was employed.

## Procedure

Participants were assigned to two groups: One of them learned a path with 20 intersections via a map, the other one through verbal description (allocentric vs. egocentric learning condition). After a five minute break, the learned path was shown as video in reverse order through the virtual maze, which was stopped at every intersection (decision point) for participants to indicate the path directions right, left, or straight (wayfinding phase). Learning condition (map/description) and landmark position (optimal/ suboptimal) served as independent variables while correct route decisions and response times served as dependent variables.

## Results

With landmarks being located in (assumed) optimal positions correct decisions on the return path were made in about $67.5 \%$ (chance level 33.3\%) if the initial path was learned via a route description. When the path was encoded via a map about $65 \%$ correct route decisions were made. With landmark objects being in suboptimal positions on the return path, $59 \%$ correct decisions were made for the description condition and $65 \%$ for the map condition.
For the optimal positions the response times were lower ( 3900 ms ) in the description condition, compared to the map condition ( 4960 ms ). Responses for intersections with landmarks on suboptimal positions revealed again a shorter response time for the description condition (4175ms), in comparison to the map condition ( 4825 ms ).

An analysis of variance with the within-subject factor landmark position (optimal/suboptimal) and the betweensubject factor learning condition (map/verbal description) was performed. It revealed a significantly higher performance for landmarks on optimal positions ( $F(1,18)=4.99, p=.038)$. But, the position did not reveal significant differences according to the response times ( $F(1,18)=.033, p=.858$ ). The learning conditions did neither differ significantly in the wayfinding phase with respect to correct decisions $(F(1,18)=.066, p=.800)$, nor with respect to response times $(F(1,18)=.621, p=.441)$. The three possible options of choice on the intersections (left, right, straight on) did not lead to significant differences according to correct decisions ( $F(2,38)=.818, p=.449$ ). No interactions were obtained.

## Discussion

The landmark position led to significant differences in performance (correct decisions), though this was not the case for the decision times. Consistent with the expectations
better decisions were made if landmarks were located on optimal positions. Since no decision time differences could be obtained, this effect cannot be due to longer viewing times for the landmarks. We may therefore conclude that the quality of a landmark as a point of reference for finding the return path very much depends on its position, as has previously been assumed for the "initial path" (forward run; Klippel \& Winter, 2005; Röser et al., 2012a,b).
The different learning conditions map and description (allocentric/egocentric) did not lead to a significant difference in the wayfinding phase, neither for correct decisions nor for the response times. This absence of an effect may be explained by the "dual coding theory of human wayfinding knowledge" (Meilinger, Knauff, \& Bülthoff, 2008). It assumes that environmental information is (sometimes) encoded in a spatial format alone but sometimes additionally in a verbal format. Information learned through maps (allocentric) is encoded verbally as well as information learned through descriptions (egocentric mental imagery). The similar performances after studying a map or a verbal description may be attributed to verbal representations existing for both encoding conditions (Meilinger \& Knauff, 2008).
In Experiment 2 position effects were found in comparison to Experiment 1. It is possible that Experiment 1 only tested the direction memory (memory task), while Experiment 2 represents a realistic wayfinding task. Since these results are not conclusive, more theoretically driven assumptions and empirical research are required.

## Theoretical assumptions, modeling, future research

In the following we present current ideas on how landmarks, places, and directions might be cognitively processed for the return path. As we have seen so far from our first two experiments on the return path and which role landmarks and landmark positions play in this context, more systematic empirical work is required.
As can be seen in Figure 2 we need to differentiate between an allocentric and an egocentric perspective. In the allocentric perspective (forward run) the assumed optimal position (Klippel \& Winter, 2005; Röser et al., 2012a) is position D , before the intersection and in direction of the turn. For the forward run optimal positions have been suggested theoretically/mathematically (Klippel \& Winter, 2005) and have been evaluated empirically (Röser et al., 2012a,b). For the return path the optimal positions are not yet known. We assume that position D should still be optimal, since it is before the intersection in direction of the turn (identical to the initial path), and this position is invariant independent of direction (no right/left encoding necessary). According to the findings by Röser et al. (2012a) position $C$ could be optimal as well in the egocentric perspective and A could be optimal in both perspectives, since $A$ is also invariant (opposite to the direction of the turn) as is the case for D .


Figure 2: Possible optimal (dark gray) and suboptimal (light gray) landmark positions for the forward run and the return path in the allocentric and egocentric perspective. See text for details.

Another important issue in the egocentric perspective is the so-called "visibility" (Winter, 2003; Röser et al., 2012b). This means that different locations have different visibilities depending on the observers own position (Figure 3). Visual attention is generally paid to the direction of turn. It seems that in an egocentric perspective it is important that a landmark is at least located in direction of the turn and that before and behind become less important.


Figure 3: Visibility from two different positions: initial path (left) and return path (right). $\mathrm{X}=$ position of individual; $\boldsymbol{\rightarrow}=$ walking direction. In the allocentric perspective each position is equally visible for both directions, not so for the return path. The small images on the bottom visualize the sight in the egocentric perspective. See text for details.

For the return path it is important to take this into account. This means that for the return path the optimal position in the allocentric perspective remains the same (D), since this location is still before the intersection and in direction of the turn (invariant; see central section of Figure 2). According to the above findings and the previous logic, in the egocentric perspective the optimal positions should now be C and D. However, position C was a suboptimal one on the forward run and therefore it may now be doubted that it becomes optimal on the return path, since it is a variant
position (forward run: before the intersection and opposite to the direction of the turn; return path: behind the intersection and in direction of the turn). This would require some additional mental transformation for the observer in order to correctly find the return path.
Now it is interesting to see that positions D and A are invariant for the initial and the return path, while B and C are variant locations (see right section of Figure 2). But, this is only the case if the spatial information is unspecific; that is right has to be transformed into left on the return path (direction specific), while turn into direction of $D$ or turn in the opposite direction of $A$ remain the same for the return path (direction unspecific).

According to the concept of "advanced visibility" (Winter, 2003) it is furthermore important in the egocentric perspective, whether both facades at one location at the intersection are visually identical/similar (e.g., same color, texture) or totally different (e.g., one facade is brown and the other white). This may change the recognizability on the return path in a dramatic way (Figure 3). For instance, if both facades are similar, then this information can be used for the return path, but if they differ significantly, then position D becomes useless on the return path, since it cannot be recognized anymore (only if the observer turns the head on the initial path at the intersection). From a perceptual point of view the object must be recognizable. If this condition is not fulfilled, the former optimal position D might become totally worthless (see Tables 1 and 2 for theoretical predictions; please note the lower right value, which has the most dramatic effects depending on visibility and equal appearance).

Not only the visibility represents an important issue but also language and how it is used when giving instructions, learning new pathways, and transforming them mentally (for the return path). As mentioned above, there are at least two ways of spatial directions: direction specific and direction unspecific information (Figure 4).

Table 1: Visibilities for the different landmark positions (A-D) in Figures 2 and 3 for the initial path, the theoretical return path, and for the real return path; 0 indicates that no facade is visible, 0.5 indicates that one facade in visible, and 1 means that both possible facades of a building at an intersection are visible. Here, both facades of a single building have the same characteristics/appearance. Thus, position D has a visibility of 0.5 on the return path, since the visible facade is similar to the one seen on the initial path.

| Path | Initial <br> Path | Return Path <br> (hypothetical) | Return Path <br> (real) |
| :--- | :---: | :---: | :---: |
| Position | 1 | 1 | 1 |
| A | 1 | 0.5 | 0.5 |
| B | 1 | 1 | 1 |
| C | 0.5 | 0.5 | $\underline{0.5}$ |
| D | 0.5 |  |  |

Table 2: Visibilities for the different landmark positions. In comparison to Table 1 we now assume that the two facades of each building are different in their appearance. This leads to a visibility of 0 for position D , since here the new facade on the return path does not contain any information about this position compared to the initial path.

| Path | Initial <br> Path | Return Path <br> (hypothetical) | Return Path <br> (real) |
| :--- | :---: | :---: | :---: |
| Position | 1 | 1 | 1 |
| A | 1 | 0.5 | 0.5 |
| B | 0.5 | 1 | 0.5 |
| C | 0.5 | 0.5 | $\underline{\mathbf{0 . 0}}$ |
| D |  |  |  |



Figure 4: Examples for verbal directions in the forward run and the return path. Note that the descriptions for positions D and A vary only slightly (if at all), while larger changes occur for positions C and B.

Direction specific here means that a precise direction with a single spatial word is provided, e.g., left or right. At first glance this information is easy to understand and simple to use. But, it becomes complicated if the return path has to be constructed, since then a left turn needs to become a right turn and vice versa (note that straight remains straight on the return path). Thus, an additional mental transformation is required. Additional in this sense means that it is also possible to encode directions in an unspecific way (without directions but rather based on landmark locations). In other words, the verbal direction turn in the direction of the gas station does not need to be verbally or mentally transformed if it is located on position D (the same is true for position A with the instruction turn in opposite direction of $A$ ). On the return path, both locations and unspecific directions would remain the same: in the mental representation the gas station would still either be in direction of the turn (D) or opposite to the direction of the turn (A). This would require one mental processing step less, since no transformation would
be required (left $\rightarrow$ right) resulting in less cognitive load. But, is this how wayfinders encode spatial information and spatial directions? Theoretically, direction unspecific information would be less effortful and therefore preferable over a direction specific strategy that results in higher cognitive load.
Therefore, it is important in a first step to systematically investigate how wayfinders encode given (unfamiliar) routes and how they transform them into a return path; and in a second step it is necessary to model the optimal strategies (also with respect to individual abilities) to make predictions about spatial performance.

## General Discussion and Conclusion

In this position paper we presented first empirical data on return path research and how this information is processed to aid wayfinding (unfortunately, we could not present all empirical research within this study). As can be seen from our theoretical assumptions, much more research is required within this context. We offered a few interesting issues, e.g., structural importance, visibility, language, mental transformation, which should be investigated further. So far we did not focus on brain imaging and neural correlates of wayfinding. But, investigating the cognitive processes of how we learn and encode initial pathways and how we later transform them into new routes (especially return paths) is also of relevance for the neuroscientific branch of this research. Thus, our findings and assumptions about the return path make up an interesting project for interdisciplinary future cognitive research.

## Acknowledgement

This study was supported by the German Research Foundation (DFG HA5954/1-1). We thank Markus Knauff for valuable comments on the manuscript.

## References

Büchner, S. J., Hölscher, C., \& Strube, G. (2007). Path choice heuristics for navigation related to mental representations of a building. In S. Vosniadou, D. Kayser, A. Protopapas (Eds.), Proceedings of the $2^{\text {nd }}$ European Cognitive Science Conference (pp. 504-509). Hillsdale, NJ: Lawrence Erlbaum Associates.
Caduff, D., \& Timpf, S. (2008). On the assessment of landmark salience for human wayfinding. Cognitive Processing, 9(4), 249-267.
Clark, A., \& Chalmers, D. (1998). The extended mind. Analysis, 58(1), 7-19.
Dudchenko, P. A. (2010). Why people get lost - The psychology and neuroscience of spatial cognition. Oxford: University Press.
Golledge, R. G. (1997). Defining criteria in path selection. In D. F. Ettema \& H. J. P. Timmermans (Eds.), Activitybased approaches to travel analysis (pp. 151-169). New York: Elsevier.

Hamburger, K., \& Knauff, M. (2011). Squareland: A virtual environment for investigating cognitive processes in human wayfinding. PsychNology Journal, 9(2), 137163.

Hamburger, K., \& Röser, F. (2011). The meaning of Gestalt for human wayfinding - How much does it cost to switch modalities? Gestalt Theory, 33(3/4), 363-382.
Kato, Y., \& Takeuchi, Y. (2003). Individual differences in wayfinding strategies. Journal of Environmental Psychology, 23(2), 171-188.
Klippel, A., \& Winter, S. (2005). Structural salience of landmarks for route discrimination. In A. G. Cohn \& D. Mark (Eds.), Spatial Information Theory. International Conference COSIT (pp. 347-362). Berlin: Springer.
Lynch, K. (1960). The image of the city. Cambridge, MA: MIT Press.
Meilinger, T., Knauff, M., \& Bülthoff, H. H. (2008). Working memory in wayfinding- a dual task experiment in a virtual city. Cognitive Science, 32(4), 755-770.
Meilinger, T., \& Knauff, M. (2008). Ask for your way or use a map: A field experiment on spatial orientation and wayfinding in an urban environment. Journal of Spatial Science, 53(2), 13-23.
Papinski, D., Scott, D. M., \& Doherty, S. T. (2009). Exploring the route choice decision-making process: A comparison of planned and observed routes obtained using person-based GPS. Transportation Research Part F, 12(4), 347-358.
Presson, C. C., \& Montello, D. R. (1988). Points of reference in spatial cognition: Stalking the elusive landmark. British Journal of Developmental Psychology, 6(4), 378-381.
Röser, F., Hamburger, K., \& Knauff, M. (2011). The Giessen virtual environment laboratory: Human wayfinding and landmark salience. Cognitive Processing, 12(1), 209-214.
Röser, F., Hamburger, K., Krumnack, A., \& Knauff, M. (2012a). The structural salience of landmarks: Results from an on-line study and a virtual environment experiment. Journal of Spatial Science, 57(1), 37-50.
Röser, F., Krumnack, A., Hamburger, K., \& Knauff, M. (2012b). A four factor model of landmark salience - A new approach. In N. Rußwinkel, U. Drewitz \& H van Rijn (Eds.), Proceedings of the $11^{\text {th }}$ International Conference on Cognitive Modeling (ICCM) (pp. 82-87). Berlin.
Sorrows, M. E., \& Hirtle, S. C. (1999). The nature of landmarks for real and electronic spaces. In C. Freksa \& D. M. Mark (Eds). Spatial Information Theory: Cognitive and Computational Foundations of Geographic Information Science, International Conference COSIT 1999 (pp. 37-50). Stade: Springer.
Winter, S. (2003). Route adaptive selection of salient features. In W. Kuhn, M.F. Worboys \& S. Timpf (Eds.), Spatial Information Theory: Cognitive and Computational Foundations of Geographic Information Science, International Conference COSIT (pp. 37-50). Berlin: Springer.

# Attentional Processing in Bistable Perception is Influenced by Genetic Effects Associated with Sinistrality. 

Roeland Hancock (rhancock@email.arizona.edu)<br>Department of Psychology, 1503 E. University Blvd.<br>Tucson, AZ 85721 USA


#### Abstract

Binocular rivalry displays and ambiguous figures such as the Necker cube elicit a perceptual reversal effect mediated by attentional and perceptual processes. Perceptual dominance times are highly variable between individuals and may be partially influenced by genetic factors. This study examined the role of putative genetic effects associated with familial sinistrality, derived from a novel pedigree-based genetic model. In a continuous Necker viewing task, dominance times were significantly correlated ( $R^{2}=.36$ ) with a multifactorial estimate of genetic effects associated with non right-handedness. No association with genetic estimates was found in an intermittent viewing condition. These results suggest that genetic factors associated with functional asymmetries may also affect noisebased perceptual alternation, but not short term visual memory.


Keywords: Bistable Perception; Attention; Individual Differences; Familial Sinistrality

## Introduction

Functional cerebral lateralization has been proposed to arise from a combination of constraints on interhemispheric connectivity (Ringo, Doty, Demeter, \& Simard, 1994) and the properties, such as excitability, of neural assemblies within each hemisphere (Levitan \& Reggia, 2000). Cerebral lateralization, particularly for language, is often associated with individual hand preference as well as the handedness of close relatives (Tzourio-Mazoyer et al., 2010), suggesting a genetic contribution to functional asymmetry. Genetic effects associated with familial handedness may in turn be linked to subtle neurobiological differences, including the distribution of neurotransmitters involved in regulating neural excitability, noise or interhemispheric connectivity. We adopt a simple behavioral paradigm, the Necker cube viewing task, that may be sensitive to these factors at multiple time scales to identify possible associations between neural function and familial handedness.

A variety of ambiguous visual displays can induce bistable (or multistable) perception in observers, a state in which the subjective interpretation of the display alternates between perspectives. This can be induced through ambiguous figures (e.g. the Necker cube and Rubin's face/vase illusion), ambiguous motion or binocular rivalry displays in which different images are shown to each eye. Binocular rivalry displays allow more control over stimulus presentation and are more widely studied than ambiguous displays, but similar processes are assumed to drive alternation in both types of displays. Dominance times-the duration for which a particular interpretation is perceived-are highly variable between individuals (Aafjes, Hueting, \& Visser, 1966) and may be partially influenced by genetic factors in both binocular rivalry
and Necker cube displays (Shannon, Patrick, Jiang, Bernat, \& $\mathrm{He}, 2011$ ). Several factors, including the visual properties of the stimuli (e.g. intensity (Levelt, 1967)), eye movements and attention, have been proposed to affect perceptual reversals.

Eye movements are related to perceptual reversals, but it is not clear that eye movements directly cause reversals. Fixating on one face of the Necker cube biases perception, but does not eliminate perceptual alternation, suggesting that reversals are not merely the result of scanning different areas of the visual field (Toppino, 2003). Scotto, Oliva, and Tuccio (1990) showed Necker cube images that were partially corrected for eye-movements which resulted in an increased reversal rate over normal viewing conditions, further suggesting that bistable perception is not merely a product of scanning eye movements across the visual field. van Dam and van Ee (2005) examined the relation between perceptual reversals and eye movements and found no correlation between reported reversals and involuntary micro-saccades, however saccades and eye blinks were suppressed prior to button presses. This suppression occurred both when subjects reported perceptual flips and experimentally controlled stimulus flips. The number of blinks and saccades did not decrease prior to random button presses that were unrelated to the stimulus. Thus, bistable perception appears to be related to a cognitive process that then directs visual attention, rather than the direct result of visual field changes.

The present study investigates individual variability in Necker cube dominance times as a function of putative genetic effects associated with hand preference. Handedness and familial sinistrality are known to affect lateralization of other brain functions, notably language (Tzourio-Mazoyer et al., 2010). Familial sinistrality has also been found to influence spatial processing (e.g. Mckeever (1986)) and resting brain states (Hancock, 2012), but has not been widely studied with respect to the cognitive processes involved in bistable perception. Although Necker cube viewing is typically not a lateralized task, fMRI and EEG studies suggest that the reversal phenomenon is right-lateralized, at least initially (Pitts, Martínez, Stalmaster, Nerger, \& Hillyard, 2009; Britz \& Pitts, 2011) and dominance time is associated with grey matter volume in the right, but not left, superior parietal lobe (Kleinschmidt, Sterzer, \& Rees, 2012). Unilateral transcranial magnetic stimulation and caloric stimulation can disrupt binocular rivalry and Necker cube reversals (Miller et al., 2000), supporting bilateral contributions to bistable perception. While interhemispheric transfer across the corpus
may play some part in normal switching, it is unlikely to have a critical effect on dominance cycles (Miller et al., 2000; Pettigrew, 2001). Instead, Pettigrew suggests that a subcortical oscillatory mechanism, perhaps in the ventral tegmental area, may drive rivalry. A model of interhemispheric switching between two percepts is also difficult to reconcile with the similar dynamics found in multistable perception, including tristability (Wallis \& Ringelhan, 2013).

Factors known to influence asymmetric brain functionsuch as familial sinistrality-may have effects on bistable perception, but these likely reflect neural differences within hemispheres that could contribute to lateralization during development (Levitan \& Reggia, 2000). Specifically, we hypothesize that familial sinistrality may partially account for individual differences in neural excitability or noise, two critical parameters that determine dominance times in many models of bistable perception (e.g. Moreno-Bote, Rinzel, \& Rubin, 2007). In addition, bistable perception offers an interesting test case for identifying such effects at multiple levels of processing. The two Necker cube viewing tasks used here-a continuous viewing and intermittent viewing condition-are likely to differentially recruit perceptual processes. Bistable perception recruits a fronto-parietal network of brain regions, presumed to reflect contributions of both frontal, higherlevel attention processes and lower-level perceptual processes (Sterzer, Kleinschmidt, \& Rees, 2009). Intermittent viewing conditions extend dominance times (Leopold, Wilke, Maier, \& Logothetis, 2002), likely reflecting in increased role for visual memory processes in posterior brain regions. Functional studies suggest that brain asymmetries are multifactorial with partially independent factors controlling lateralization in brain regions associated with language, attention and vision (Liu, Stufflebeam, Sepulcre, Hedden, \& Buckner, 2009). Genetic effects associated with familial sinistrality are proposed to influence one or more of these functions, which may be observed as individual variability in the continuous Necker cube task (reflecting contributions of attention and visual processing, modulated by short-term dynamics ) or in the intermittent viewing task (reflecting primarily visual working memory or slow neural processes).

## Genetic Models of Handedness

Most studies of familial handedness effects treat familial handedness as a categorical variable, broadly classifying subjects as those reporting only right-handed relatives (FS-) or reporting at least one non right-handed relative (FS+). Considering the low heritability of handedness (20-30\%, Medland et al. (2009)), problems introduced by considering variable family sizes and family envelopes of interest (Bishop, 1990), and placing subjects who may have vastly different susceptibilities into the same group, categorical classification is likely to be an underpowered approach to studying genetic effects associated with handedness. Power can be increased substantially with the use of non-categorical measures of familial sinistrality, such as the proportion of left-
handed relatives (Corey \& Foundas, 2005), but these are not widely used. Even when more genetically informed familial handedness measures are used, these are sometimes based on a particular theory of genetic transmission and expression, thus confounding familial handedness effects with a specific, and likely incorrect, genetic model (e.g. McManus (1995)). Applying these theories to calculate the likelihood of individual genotypes (which might then be used a covariate) also requires assumptions regarding allele frequencies to be made. Recently, Karev (2011) proposed a quantitive measure of familial handedness based on weighting left-handed relatives by relatedness, finding that this measure correlates with hand preference measures.

In an effort to increase power and more robustly associate familial sinistrality effects with putative genetic effects, a standard multifactorial threshold model was used to estimate genetic load for left handedness in individual subjects. This model treats the phenotype as function of multiple genetic loci, each probably associated with small genetic effect, and multiple environmental effects. Unlike single gene models (Annett, 1985; McManus, 1991; see McManus, Davison and Armour (in press) for a multi-locus variant of the dextral chance model), this approach does not assume that the genetic component of the phenotype is controlled largely or entirely by posited genes with a given transmission probability and penetrance, however it can account for such effects when alleles make additive contributions to the phenotype (as is this case in McManus's model). This approach extends Karev (2011) by considering not only the relatedness of left-handed relatives but also their expected genetic effects. This is similar to McManus (1995), but relies only a standard multifactorial model, as opposed to a particular major gene model. The basic assumption of the multifactorial genetic model is that multiple loci additively contribute to a continuous phenotype and interact with environmental effects. For example, a multifactorial model of height would assume a set of loci with alleles at each loci adding or subtracting a small amount from the phenotype. The summed contribution of these alleles is the additive genetic effect or genetic load. The genetic effect interacts with environment, for example hormones or nutrient levels, to produce the expressed phenotype. Formally, we can consider the contribution of genes to a phenotype as being described by the linear model

$$
\begin{equation*}
y=X \beta+A a+e \tag{1}
\end{equation*}
$$

where $y$ is the measured phenotype, $X$ is a matrix of observations for fixed effects (sex and founder status, in this case), $\beta$ is the vector of fixed effects, $a$ is the vector of additive genetic effects and $e$ is a vector of environmental effects. $A$ is an $n \times n$ matrix describing how $n$ individuals are related in terms of their expected allele sharing. For example, $A_{i, j}=.5$ indicates that individuals $i$ and $j$ share (theoretically) $50 \%$ of their genes, i.e. $i$ and $j$ are siblings or a parent/offspring pair. For simplicity, we consider only autosomal effects.

We employ the Bayesian formulation of (1)

$$
\begin{equation*}
y \mid \beta, a, \sigma_{e}^{2} \sim N\left(X \beta+A a, I \sigma_{e}^{2}\right) \tag{2}
\end{equation*}
$$

with

$$
\begin{equation*}
a \mid A, \sigma_{a}^{2} \sim N\left(0, A \sigma_{a}^{2}\right) \tag{3}
\end{equation*}
$$

and an uninformative prior for fixed effects $\beta$, e.g.

$$
\begin{equation*}
\beta \sim N\left(0,10^{6}\right) \tag{4}
\end{equation*}
$$

For binary phenotypes, we assume that the observed phenotype is obtained by thresholding a latent continuous variable (the liability) (Falconer, 1965). If the liability, $\ell$, exceeds a threshold value $t_{0}$ (say $t_{0}=0$, since the liability distribution can be shifted arbitrarily), then the binary phenotype (e.g. disease or left-handedness) is expressed. The liability is distributed as in (2)

$$
\begin{equation*}
\ell \mid \beta, a, \sigma_{e}^{2} \sim N\left(X \beta+A a, I \sigma_{e}^{2}\right) \tag{5}
\end{equation*}
$$

where $\ell$ is mapped to a categorical phenotype $c=1,2, \ldots, k$ by the partition

$$
\begin{equation*}
t_{0}<t_{1}<t_{2}<\ldots<t_{k-1}<t_{k} \tag{6}
\end{equation*}
$$

so that

$$
\operatorname{Pr}\left(y_{i}=c \mid l_{i}, t_{c-1}, t_{c}\right)= \begin{cases}1 & t_{c-1}<l_{i} \leq t_{c}  \tag{7}\\ 0 & \text { otherwise }\end{cases}
$$

Fitting this model to family pedigrees provides a posterior distribution of additive effects for each individual in the pedigree that can be then be used as a covariate.

## Pedigrees

Data for fitting the model were obtained from a survey of hand preferences for blood relatives distributed to 4561 students in introductory psychology courses at the University of Arizona. Respondents were asked to report hand preferences (left, right, ambidextrous or unknown) for each parent and grandparent. For brothers, sisters, aunts and uncles, respondents were asked to provide the number of relatives in each relation with left, right, ambidextrous or unknown hand preference from the options $0,1,2$ and " 3 or more" (coded as 3 relatives). Respondents also reported their sex, hand preferences for writing, drawing and throwing a ball (on a five point scale from 'strong left' to 'strong right') and foot preference for kicking a ball. Respondents were considered righthanded if they reported a 'moderate right' or 'strong right' preference for writing hand, ambidextrous if they reported 'no preference' and left-handed if they reported a 'strong left' or 'moderate left' hand preference. Respondents were not given any criteria for determining relatives' handedness. A similar questionnaire was administered to each experimental participant.

Since the observed liability for right-handed individuals is always below threshold, purely right-handed families are uninformative with respect to random effect estimates, i.e.
the posterior distribution of additive genetic effects is fixed by the prior distribution. To improve the mixing properties of the model, effects were estimated only for individuals in FS+ families and individuals with missing data who did not form an ancestral link in a given pedigree were removed from the pedigree. For example, avuncular relatives and grandparents of unknown handedness were removed, but a parent of unknown handedness who had multiple children of known handedness would be retained. Families with fewer than three phenotype individuals were also removed. Pedigrees were pruned using the prunePed function in the $R$ MCMCglmm package (Hadfield, 2010). Phenotypes of the remaining unknown individuals were replaced at random, conditional on known family members. Convergent logit models were estimated with MCMCglmm using a slice sampling Monte Carlo Markov chain (MCMC) algorithm with block updating (Sorensen \& Gianola, 2002; Hadfield, 2010) for $2,500,000$ iterations with a burn in period of 500,000 iterations and thinning interval of 100 .

## Methods

## Participants

Nineteen participants ( 10 female) with no reported non righthanded relatives and 26 ( 14 female) participants with at least one reported left-handed relative, all native English speakers, were recruited from introductory psychology courses at the University of Arizona, Tucson, AZ, USA and received course credit for participation. Participants were screened by self report for history of brain injury, neurological and psychiatric disorders, medication, and normal vision. Research was approved by the University of Arizona Institutional Review Board.

## Procedure

A white-on-black Necker cube (Figure 1A) was presented, centered on a 15 inch LCD display at a viewing distance of 15 inches. Stimuli subtended a visual angle of approximately 4 degrees. Stimuli were presented using the MATLAB-based (Mathworks, Natick, MA) Psychtoolbox (Brainard, 1997) on Windows XP. Participants were seated at eye level with the center of the display using a chin rest to minimize movement. Participants were instructed to focus on a small cross in the center of the Necker cube, to avoid deliberately switching perspectives and to report perspective changes by pressing a key corresponding to the new perspective. The left and right arrow keys of a standard keyboard were labeled with images of the "left" (Figure 1B) and "right" (Figure 1C) perspectives, respectively. Participants were instructed to respond with the first and second fingers of their right hand and were familiarized with the two perspectives prior to the experiment. The experiment consisted of a continuous viewing condition, designed to elicit spontaneous perceptual reversals followed by an intermittent viewing condition designed to extend dominance times (Leopold et al., 2002) and recruit memory processes (Pastukhov \& Braun, 2008) or slow neural dynamics.


Figure 1: A Necker cube (A) and the perceived left (B) and right (C) perspectives.

In the continuous viewing condition, the Necker cube was continuously presented during four blocks, with self-paced breaks between each block. In each block, the complete Necker cube was shown for 200 seconds. As a check for random responses, one perspective, selected at random, of the complete Necker cube was reduced in intensity over 30 seconds to gradually produce a partially unambiguous cube and a corresponding increase in dominance time. This process began 0-30 seconds (uniformly distributed) after the end of each block, followed by a 30 second presentation of the complete Necker cube. In the intermittent viewing condition, participants saw an alternating sequence of a Necker cube followed by a fixation screen without the cube. On each intermittent trial, the Necker cube was shown for 750,1225 or 2000 ms followed by a fixation screen for 1000, 2236 or 5000 ms . All combinations of stimulus duration and inter stimulus interval (ISI) were repeated 24 times. Participants reported their perspective at the beginning of each trial and again if a reversal was experienced.

## Results

Log-transformed dominance times, defined as the time between alternate-side button presses, were analyzed using linear mixed effects models with random intercepts per subject and sex, block, perspective and genetic load as fixed effects. Likelihood ratio tests were used to select models. Reported p-values were estimated using MCMC as implemented in the languageR package (Baayen, 2007).

## Continuous Viewing

Dominance time increased for all participants during the partially disambiguated display periods $(M=11.88 \mathrm{sec}$, $S D=6.94$ ), indicating their responses were not stimulusindependent. No significant effect of sex or block was found. There was a significant positive association between additive genetic effects and dominance time ( $\beta=.68, p=.005$ ).

Dominance times were significantly shorter for the "right" perspective ( $\beta=.47, p<.001$ ).

Mean Dominance Times


Figure 2: Correlation $\left(R^{2}=.36\right)$ between continuous viewing dominance times and estimated additive genetic effects (a). FS- individuals are shown for comparison only.

## Intermittent Viewing

Intermittent viewing times were uncorrelated with continuous viewing times $(r=-.05, p=.74)$ or estimated genetic effects ( $r=.13, p=.53$; Figure 3) and were significantly longer ( $M=17.84 \mathrm{sec}, S D=17.65$ ) than continuous dominance times $(t(43)=6.70, p<.001)$.

## Discussion

Increasing estimated genetic load for non right-handedness in right-handed subjects was associated with increased dominance times in the continuous viewing condition. Although no control for differences in overall motor response speed as a function of familial sinistrality was included, the increase in dominance time during disambiguated display periods and intermittent viewing suggests that participants' responses were directly related to the perceived orientation of the Necker cube. Current models of bistable perception propose that reversals are the result of an interaction between parietal/occipital perceptual and visual working memory processes and noisy frontal attentional processes (Sterzer et al., 2009). In the context of these models (Gigante, Mattia, Braun, \& Del Giudice, 2009; Moreno-Bote et al., 2007), the dissociation between familial sinistrality effects in the continuous and intermittent viewing conditions suggests that the genetic effects of interest selectively act on frontal processes or the interaction between frontal and parietal regions.

No group differences were found when familial sinistrality was considered as a dichotomous variable. Although the FS-


Figure 3: Correlation between intermittent viewing dominance times and estimated additive genetic effects (a). FSindividuals are shown for comparison only.
group might be expected to have fewer genetic effects found in the FS+ group, and therefore show an overall group difference in dominance times, the obtained result is consistent with a gene-gene interaction model for handedness. Under this model, the estimated genetic effects, $a$, associated with phenotypic handedness may be decomposable into alleles that influence bistable perception and possibly other aspects of cognition and brain organization $\left(a_{b}\right)$ and one or more alleles $\left(a_{h}\right)$ that contribute to hand preference in the presence of $a_{b}$ effects. Thus the FS- and FS+ groups may have similar distributions of $a_{b}$ effects and dominance times while the FS- group lack the $a_{h}$ alleles needed to express non righthandedness in conjunction with $a_{b}$ alleles. With handedness as a phenotype, the $a_{b}$ effects of interest can only be estimated in conjunction with $a_{h}$ effects and not at all in the FS- group.

These findings validate the utility of applying a multifactorial genetic model of handedness to account for individual differences in cognitive behavior and support views of lateralization and related brain function as a complex trait, both in terms of phenotype and genetics. The neural mechanisms of the familial sinistrality-related increase in dominance times are unknown, but likely reflect differences in attentional processing or the coupling of attentional and visual levels of processing. The increases in dominance times could reflect changes at the local circuit or neurotransmitter level leading to decreased levels of neural noise that trigger perceptual reversals (e.g. Moreno-Bote et al., 2007) .

## Acknowledgments

This research was supported by a fellowship to R.H. from the Arizona Center for the Biology of Complex Diseases.

Ian Rios and Ankit Takyar assisted with data collection. Dr. Thomas G. Bever provided comments and laboratory resources.

## References

Aafjes, M., Hueting, J. E., \& Visser, P. (1966). Individual and interindividual differences in binocular retinal rivalry in man. Psychophysiology, 3(1), 18-22.
Annett, M. (1985). Left, right, hand and brain. Psych Press UK.
Baayen, R. H. (2007). Analyzing Linguistic Data: A practical introduction to statistics using R. Cambridge Univ Press.
Bishop, D. V. M. (1990). On the futility of using familial sinistrality to subclassify handedness groups. Cortex, 26(1), 153-155.
Brainard, D. H. (1997). The Psychophysics Toolbox. Spatial Vision, 10(4), 433-436.
Britz, J., \& Pitts, M. A. (2011). Perceptual reversals during binocular rivalry: ERP components and their concomitant source differences. Psychophysiology, 48(11), 1490-1499.
Corey, D. M., \& Foundas, A. L. (2005). Measuring familial sinistrality: problems with dichotomous classification. Laterality, 10(4), 321-335.
Falconer, D. (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives. Annals of Human Genetics, 29(1), 51-76.
Gigante, G., Mattia, M., Braun, J., \& Del Giudice, P. (2009). Bistable perception modeled as competing stochastic integrations at two levels. PLoS Computational Biology, 5(7), e1000430.
Hadfield, J. (2010). MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. Journal of Statistical Software.
Hancock, R. (2012, May). Bayesian estimates of genetic handedness predict oscillatory brain activity. Presented at the 14th Annual Meeting of the International Behavioral and Neural Genetics Society.
Karev, G. B. (2011). Quantitatively assessed familial sinistrality in right-, mixed-, and left-handers. Laterality, 16(1), 93-106.
Kleinschmidt, A., Sterzer, P., \& Rees, G. (2012). Variability of perceptual multistability: from brain state to individual trait. Philosophical Transactions Of The Royal Society BBiological Sciences, 367(1591), 988-1000.
Leopold, D. A., Wilke, M., Maier, A., \& Logothetis, N. K. (2002). Stable perception of visually ambiguous patterns. Nature Neuroscience, 5(6), 605-609.
Levelt, W. J. (1967). Note on the distribution of dominance times in binocular rivalry. British Journal of Psychology, 58(1), 143-145.
Levitan, S., \& Reggia, J. A. (2000). A computational model of lateralization and asymmetries in cortical maps. Neural computation, 12(9), 2037-2062.
Liu, H., Stufflebeam, S. M., Sepulcre, J., Hedden, T., \& Buckner, R. L. (2009). Evidence from intrinsic activity that
asymmetry of the human brain is controlled by multiple factors. Proceedings of the National Academy of Sciences, 106(48), 20499-20503.
Mckeever, W. F. (1986). The influences of handedness, sex, familial sinistrality and androgyny on language laterality, verbal ability, and spatial ability. Cortex, 22(4), 521-537.
McManus, I. C. (1991). The inheritance of left-handedness. Ciba Foundation symposium, 162, 251-67; discussion 267-81.
McManus, I. C. (1995). Familial sinistrality: the utility of calculating exact genotype probabilities for individuals. Cortex, 31(1), 3-24.
McManus, I. C., Davison, A., \& Armour, J. (2013). Multi-locus genetic models of handedness closely resemble single-locus models in explaining family data and are compatible with genome-wide association studies. Annals of the New York Academy of Sciences.
Medland, S. E., Duffy, D. L., Wright, M. J., Geffen, G. M., Hay, D. A., Levy, F., ... Boomsma, D. I. (2009). Genetic influences on handedness: data from 25,732 Australian and Dutch twin families. Neuropsychologia, 47(2), 330-337.
Miller, S. M., Liu, G. B., Ngo, T. T., Hooper, G., Riek, S., Carson, R. G., \& Pettigrew, J. D. (2000). Interhemispheric switching mediates perceptual rivalry. Current biology : CB, 10(7), 383-392.
Moreno-Bote, R., Rinzel, J., \& Rubin, N. (2007). NoiseInduced Alternations in an Attractor Network Model of Perceptual Bistability. Journal of Neurophysiology, 98(3), 1125-1139.
Pastukhov, A., \& Braun, J. (2008). A short-term memory of multi-stable perception. Journal of Vision, 8(13), 7-7.
Pettigrew, J. D. (2001). Searching for the switch: Neural bases for perceptual rivalry alternations. Brain and Mind, 2(1), 85-118.
Pitts, M. A., Martínez, A., Stalmaster, C., Nerger, J. L., \& Hillyard, S. A. (2009). Neural generators of ERPs linked with Necker cube reversals. Psychophysiology, 46(4), 694702.

Ringo, J. L., Doty, R. W., Demeter, S., \& Simard, P. Y. (1994). Time is of the essence: a conjecture that hemispheric specialization arises from interhemispheric conduction delay. Cerebral Cortex, 4(4), 331-343.
Scotto, M. A., Oliva, G. A., \& Tuccio, M. T. (1990). Eye movements and reversal rates of ambiguous patterns. Perceptual and motor skills, 70(3 Pt 2), 1059-1073.
Shannon, R. W., Patrick, C. J., Jiang, Y., Bernat, E., \& He, S. (2011). Genes contribute to the switching dynamics of bistable perception. Journal of Vision, 11(3).
Sorensen, D., \& Gianola, D. (2002). Likelihood, Bayesian and MCMC methods in quantitative genetics. Springer.
Sterzer, P., Kleinschmidt, A., \& Rees, G. (2009). The neural bases of multistable perception. TRENDS in Cognitive Sciences, 13(7), 310-318.
Toppino, T. C. (2003). Reversible-figure perception: mechanisms of intentional control. Perception and Psy-
chophysics, 65(8), 1285-1295.
Tzourio-Mazoyer, N., Petit, L., Razafimandimby, A., Crivello, F., Zago, L., Jobard, G., ... Mazoyer, B. (2010). Left hemisphere lateralization for language in right-handers is controlled in part by familial sinistrality, manual preference strength, and head size. Journal of Neuroscience, 30(40), 13314-13318.
van Dam, L. C. J., \& van Ee, R. (2005). The role of (micro)saccades and blinks in perceptual bi-stability from slant rivalry. Vision Research, 45(18), 2417-2435.
Wallis, G., \& Ringelhan, S. (2013). The dynamics of perceptual rivalry in bistable and tristable perception. Journal of Vision, 13(2).

# Intrusions and the Decision to Terminate Memory Search 

J. Isaiah Harbison (isaiah.harbison@gmail.com)<br>Department of Psychology, University of Maryland at College Park College Park, MD 20742

Eddy J. Davelaar (eddy.davelaar@gmail.com)<br>Departiment of Psychological Sciences, Birkbeck University of London Malet Street, WC13 7HX, London, UK

Erica C. Yu (erica.c.yu@gmail.com)<br>Erika K. Hussey (erikahussey@gmail.com)<br>Michael R. Dougherty (mdougher@umd.edu)<br>Department of Psychology, University of Maryland at College Park College Park, MD 20742


#### Abstract

Little is known about the how the decision is made to terminate memory search, though there have been several recent attempts to uncover this process. In one recent study, Miller et al. (2012), re-analyzed data from a large number of free-recall experiments and identified intrusions as a factor that influenced search termination decisions. One potential problem with this re-analysis is that all the data were drawn from experiments in which it was impossible to determine if or when search was terminated. Using data from experiments in which search termination decisions were directly measured, we confirmed Miller at al.'s (2012) original findings but also demonstrated that intrusions influence the time taken to generate the final retrieval and the time between the final retrieval and search termination. The pattern of data is consistent with a simple, sample-withreplacement model in which intrusions are less active than items from the target list.


Keywords: recall; memory search termination; stopping rules

Every search of memory is eventually terminated. When an individual decides to terminate their own memory search (e.g., when they are not interrupted or given a fixed time limit for search), what factors influence this decision? The long history of memory research is relatively silent regarding this question as most memory recall experiments give participants a pre-determined amount of time to search memory (a closed-interval) and have no method of determining when or, even if, participants terminate their search before the retrieval interval expires. When participants are allowed to terminate their own search, as is the case in the open-interval design discussed below, a number of dependent variables emerge that allow the measurement of memory search termination decisions. These variables include the total time spent in search ( total time or $\mathrm{T}_{T}$ ), which is controlled by the participant in this design. The total time can be divided into the time from the beginning of search to the time of the final retrieval (time-to-last retrieval or $\mathrm{T}_{L}$ ) and the time between the final retrieval and search
termination (exit latency or $\mathrm{E}_{L}$; Dougherty \& Harbison, 2007). These variables allow for the testing of different memory search stopping rules previously proposed in the literature, with much of the available data uniquely supporting the cumulative-failures stopping rule (Harbison, Dougherty, Davelaar, \& Fayyad, 2009) proposed within the search of associative memory model (SAM; Raaijmakers \& Shiffrin, 1981). According to this rule, every retrieval attempt that does not produce a new retrieval is counted as a retrieval failure and search is terminated when the number of these failures reaches a threshold.

Recent research, however, has suggests that search termination might also be influenced by the presence of memory intrusions (Miller, Weidmann, \& Kahana, 2012; Unsworth, Brewer, \& Spiller, 2011). Miller et al. (2012) showed that memory search was more likely to be terminated after an intrusion from a previous list (prior list intrusion or PLI), an extra list intrusion (ELI), or after outputting a list word that had previously been retrieved (repetition). They suggested that the increase in the probability of stopping may be due to the effect such retrieval errors have on subsequent recall. Each retrieved word is thought to influence subsequent retrievals either by the use of the retrieved word as a cue for subsequent retrieval attempts (Kimball, Smith, \& Kahana, 2007; Sirotin, Kimball, \& Kahana, 2005) or by the retrieved item shifting the contextual retrieval cues closer to the retrieved items own context (Howard \& Kahana, 2002). Intrusions and repetitions then would decrease the probability of retrieving a new target list word. A PLI would increase the relative probability of another word from the prior list being retrieved; an ELI would increase the probability of sampling related extra-list words, and repetitions would increase the probability of retrieving other words that have already been retrieved.

One potential problem with the Miller et al. analysis is that they had to infer when participants terminated search since the experiments they used in their analysis used a closed-interval design. To determine
when participants might have terminated search, they used data from an open-interval experiment (Dougherty \& Harbison, 2007) to set an inter-retrieval time longer than participants were found to search memory before terminating search. However, the factors they identified as increasing the probability of search termination would also slow down retrieval. That is, if intrusions increase the probability of sampling non-target words, this decreases the probability of sampling target list words. When the probability of sampling a word decreases, the expected time to sample that word increases. Therefore, it might be that intrusions did not increase the probability of search termination but simply slowed retrieval sufficiently for no additional words to be output in the retrieval interval. Participants might have continued searching but to no avail.

In the present study, we used data from experiments that used an open-interval paradigm to examine whether intrusions increase the probability of search termination when participants were required to indicate when they terminated search. Second, we tested the hypothesis that the retrieval of an intrusion changes the probability of subsequent retrieval types. Third, we tested if terminating search after an intrusion changes the temporal variables of search termination $\left(\mathrm{T}_{T}, \mathrm{~T}_{L}\right.$, and $\left.\mathrm{E}_{L}\right)$. Fourth, we evaluated whether the results could be modeled using a cumulative-failures stopping rule, and if so what assumptions were needed.

## Open- vs Closed-Interval Retrieval

The difference between the standard free recall paradigm, or the closed-interval design, and the openinterval design is depicted in Figure 1. In the closedinterval design participants are given a predetermined length of time for retrieval (e.g., 60 seconds). In this design, the decision to terminate search is obscured. During the retrieval interval participants might continue to search memory throughout the entire interval, terminate search immediately after the final retrieval, or even retrieve items after terminating search (e.g., a participant might terminate search then stumble upon another list word while letting their mind wander). In contrast, the open-interval design allows participants to continue search until they decide to terminate their own search (depicted in Figure 1 with an " X "). As mentioned above, three temporal variables emerge as measures of memory search termination and it has been found that the $\mathrm{T}_{T}$ and $\mathrm{T}_{L}$ increases while the $\mathrm{E}_{L}$ decreases with the number of items retrieved (Dougherty \& Harbison, 2007; Harbison et al., 2009; Unsworth, Brewer, \& Spillers, 2011).

The open-interval design is therefore particularly useful for examining search termination decisions and provides not only a method of replicating the Miller et al. results, but also a method of extending them. Using this design it is possible to determine if search is terminated


Figure 1: Comparison of A) Closed-Interval and B) Open-Interval Experimental Designs.
more quickly if an intrusion is the final outputted word relative to when a word from the target list is the final retrieval.

We re-analyzed data from two previous experiments using the open-interval design to test for these results. Importantly, the data from these two experiments utilized what we call a multi-target cued recall paradigm. Within this paradigm, participants studied multiple separate lists of words successively, with all words from each list paired with a single cue word. After studying all the lists, participants were presented with a cue and asked to recall as many words as possible that had appeared with the cue. This aspect of the experiments is particularly useful for examining the role of intrusions on search termination because it should increase the probability of PLIs. The multi-target cue task can be contrasted with the standard procedure for list recall experiments, which is to have participants study a single list at a time and then, possibly after a filler task to clear short-term memory (e.g., solving simple math problems), prompt participants to retrieve words from the studied list.

## Analysis of Previous Experiments

As mentioned above, both experiments used a multitarget cued recall paradigm. For the first experiment (Dougherty \& Harbison, 2007), three lists were shown in each block for four total blocks of lists providing data from twelve lists per participant. For the second (Experiment 1, Harbison et al., 2009), four lists were shown in each of the four blocks providing sixteen lists per participant. The two experiments also differed in list length. There were ten words per list in the first experiment and eight words per list in the second. Both experiments used lists of high (KF i 50) and low (KF ; 10) frequency words drawn from the MRC Psycholinguistic Database (Wilson, 1988; available from http://www.psy.uwa.edu.au/mrcdatabase.uwa_mrc.htm). For the second experiment, in addition to high and low
frequency lists, two of the lists per block were a mix of high and low frequency words (four of each). One mixed-frequency list had a high frequency cue word per block and one had a low frequency cue word. In both experiments, word frequency did not have a significant effect on the memory search termination variables. Therefore, the analyses collapsed across lists of high and low frequency words.

Table 1: Mean Number Retrieved

| Source | List | PLI | ELI |
| :--- | :---: | :---: | :---: |
| Dougherty \& Harbison, 2007 | 3.40 | .25 | .18 |
| Exp 1. Harbison et al., 2009 | 2.12 | .53 | .42 |
| New Experiment | 1.84 | .32 | .06 |

Table 1 shows the mean number of list words, PLIs, and ELIs per participant per list. Note that the number of PLIs though still small made up $6 \%$ and $17 \%$ of the total number of items retrieved, respectively. Following the procedure of Miller et al. (2012), we examined the probability of terminating search as a function of both the previous retrieval type (list word, PLI, ELI) and output position. Note that for all results reported in this study statistical significance was determined by biasedcorrected and accelerated (BCa) bootstrap estimates of the $95 \%$ confidence intervals. Here, we test the differences between list words and the two types of intrusions. The results from both previous experiments were consistent with the PLIs and ELIs increasing the probability of search termination, as shown in Figure 2A, but none of these differences were significant for the Dougherty and Harbison (2007) experiment. For the Harbison et al. (2009) experiment, shown in Figure 2B, output positions two and three were significant for both ELIs and PLIs, also ELIs were significant for the first position. Therefore, these results replicated those of Miller et al. (2012).

The hypothesis for why intrusions increase the probability of search termination was also tested. As mentioned above, the explanation is that intrusions increase the probability of non-list items being sampled after an initial intrusion. These intrusions could be words from a previous list, as should be the case for PLIs, or words from outside the experiment, as should be the case for ELIs. Figure 3A and B shows the probability of list words, PLIs, ELIs, and search termination immediately after each type of retrieval. For both experiments, after a PLI participants were more likely to terminate search, less likely to output a list word, and, importantly, more likely to output another PLI relative to after they had generated a list word. In contrast, after ELIs there were only significant differences in the probability of generating list words and terminating search and this was only found to be significant in the Harbison et al. Experiment


Figure 2: Probability of terminating search as a function of the previous retrieval and output position for A) Dougherty \& Harbison, 2007; B) Harbison et al., 2009, Exp 1; and C) the new experiment.

1 data.
These results are consistent with the explanation proposed by Miller et al.(2012). Not only are PLIs associated with an increase in the probability of termination, but they are also associated with an increase in the probability of generating other words from previous lists. The support for ELIs is mixed, but still consistent with this explanation.

Overall the pattern of results using the open-interval design were consistent with those reported using the closed-interval design. Search termination was more likely after an intrusion, especially a PLI. Furthermore, the retrieval of a PLI does appear to be correlated with an increase in the probability of retrieving words from previous lists and a decrease in the probability of retrieving subsequent list words. However, the replication of difference in the probability of terminating search by type and output position was only significant for one of the two re-analyzed experiments. Therefore, to further test Miller et al.'s results a new experiment using the open-interval and multi-target cued recall was conducted.


Figure 3: Probability of retrieval type or termination as a function of the previous retrieval type.

## Experiment

103 Participants were randomly assigned to one of two conditions. Participants in one condition were given two blocks of four high frequency word lists followed by two blocks of low frequency word lists while participants in the other condition were given four blocks of low frequency word lists. This manipulation was designed to test a hypothesis about the role of individual differences in motivation on the influence of previous retrieval experience on stopping decisions (Dougherty \& Harbison, 2007) which is outside of the scope of the present study. As was found in the previous experiments, word frequency did not influence stopping decisions outside of the impact on number of words retrieved. Therefore, consistent with the previous experiments, all lists were combined for the purpose of examining stopping decisions.

## Stimuli

Both high and low frequency words were drawn from MRC linguistics database (Wilson, 1988), the same source for lists in the previous experiments, and the same criteria for high and low frequency words were used. Lists of eight to be recalled words and one cue word were random generated for each participant.

## Procedure

The same open-interval, multi-target cued recall procedure was used as in the previous experiments. During learning, participants were presented with the cue word and each list word from the first, second, third, and fourth list. The learning phase was then repeated. Thus, participants saw each list word with that list's cue word twice. Participants were then asked to retrieve the words from the first, second, third, and fourth lists, in that order. There were a total of four blocks of lists. As before, participants were provided an open-interval for retrieval. They indicated when they were finished retrieving from each list by saying "Stop" and pressing the space bar.

## Results and Discussion

The exit latency and total time results replicated the findings from previous open-interval experiments (Dougherty \& Harbison, 2007; Harbison et al., 2009; Unsworth et al., 2012). $\mathrm{E}_{L} \mathrm{~s}$ were negatively correlated with number retrieved and $\mathrm{T}_{L}$ and $\mathrm{T}_{T}$ were positively correlated with number retrieved. The mean within participant gammas of $-0.293,0.224$, and 0.472 , respectively, were each significant.
Intrusions and Search Termination As shown in Figure 2C, the results were again consistent with Miller et al. (2012). PLIs and ELIs were consistently more likely to be the final word retrieved before termination across output positions relative to target list words. However, only the output positions with the most participants contributing to them, the first three positions, were significantly different between the intrusion types (both PLI and ELI) and list words.

The shift in the probability of retrieving subsequent list items and intrusions after retrieving an intrusion was found in the present experiment, matching the results of the re-analyzed experiments. Participants were more likely to have a PLI after a PLI and more likely to terminate search while also being less likely to retrieve a target item. Likewise, after an ELI, participants were more likely to terminate and less likely to generate a target word. One difference from the results reported above is that participants were also more likely to generate an ELI after an ELI in the present experiment. This pattern is shown in Figure 3C.
Intrusions and the Time Course of Termination The open-interval design also allows for the testing of temporal effects of intrusions. Specifically, does the generation of an intrusion have a different profile in terms of exit latency $\left(\mathrm{E}_{L}\right)$, time-to-last $\left(\mathrm{T}_{L}\right)$, and total time $\left(\mathrm{T}_{T}\right)$ ? To test this, we compared temporal variables when search was terminated after both types of intrusions with instances where retrieval was terminated after a target list word was generated. Note that rate transformations were used for the purposes of the analyses and
that we again used BCa to estimate the $95 \%$ confidence intervals.


Figure 4: A) Exit Latency, B) Time to Last, and C) Total Time as a function of final retrieval type and total number retrieved.

We combined the data from the present experiment and the experiments included in the above reanalysis. The combined $\mathrm{E}_{L}, \mathrm{~T}_{L}$, and $\mathrm{T}_{T}$ results are shown in Figure 4 as a function of final retrieval type and total number retrieved. The $\mathrm{E}_{L}$ after a PLI was significantly shorter than when the final word was from the target list for four of the five cases. Furthermore, the $\mathrm{T}_{L}$ was longer after a final PLI for all five total number retrieved. The pattern of results was less clear for the differences between target list words and ELIs. The $\mathrm{E}_{L}$ was shorter for three of the five cases and the $\mathrm{T}_{L}$ was longer for two of them. The results were even less consistent for $\mathrm{T}_{T}$. The difference between list words and PLIs were significant for two output positions and the difference between list words and ELIs was only significant for one of the positions.

The temporal data add another portion of the picture of the role of intrusions in search termination. Combined, the present results suggest that participants are more likely to terminate search following an intrusion, more likely to generate a subsequent intrusion of the same type (particularly for PLIs), that the time to generate the final item is longer if it is an intrusion (par-
ticularly for PLIs), and that the time between the final retrieval and termination is shorter after an intrusion (particularly for PLIs). This pattern of data provides a new challenge for models of recall and particularly stopping rules to account for. The next section examines how well as simple sample-with-replacement model is able to account for these results.

## Stopping Rules and Intrusions

For the present simulation, we tested the explanation that intrusions have lower activations relative to list words. Like intrusions, items with lower activation are retrieved later in recall. Here we tested if items with lower activation shared the additional characteristic of intrusions. The present focus is on the sampling process itself. Therefore, for the sake of simplicity, we assumed a set of activations instead of modeling encoding and the activation process (Harbison, Hussey, Dougherty, \& Davelaar, 2012; Rohrer, 1996).

A sample-with-replacement model equipped with the cumulative failures stopping rule was used to predict search. The pattern of activations used to test the model was $(.5, .5, .5, .5,1,1,1.5,1.5,2,2)$. We ran the sample-with-replacement procedure where the probability of sampling item $i$ was determined by the act ${ }_{i}$ divided by the sum over all item activations. A retrieval attempt was successful if the sampled item had not previously been retrieved and if the activation of the sampled item was greater than the recovery threshold of 0.5 . Four unrecoverable items were included in the activation pattern. The total number of retrieval failures was tracked and once this number exceeded the stopping threshold of 30 search retrieval was terminated as prescribed by the cumulative failure stopping rule (Harbison et al., 2009). The stopping threshold was within the range tested in previous applications of the model (Harbison et al., 2012) and the activation pattern used, specifically the inclusion of four items that were not recoverable, was chosen to increase the variability in the number of items retrieved. Also to this end, for each simulation run a subset of six items was chosen at random (and with equal probability) from the complete pattern of activations. Without this, the number of items retrieved was too consistent to be at all comparable to participant data.

The results from 10,000 independent runs of list recall are presented in Figure 5. Items that were relatively less active show the same profile as intrusions. The probability of terminating search after an item with less activation was greater than for an item of greater activation, as shown in Figure 5A. Also, if items with lower activation were the final retrieved, it took longer to generate them ( $\mathrm{T}_{L}$ is greater) and search was terminated more quickly after the final retrieval ( $\mathrm{E}_{L}$ is smaller), as shown in Figure 5B. Therefore, the present results relating intrusions to search termination can be accounted for by


Figure 5: A) Probability of search termination and B) $\mathrm{E}_{L}, \mathrm{~T}_{T}$, and $\mathrm{T}_{L}$ by item activation.
a simple sample-with-replacement model equipped with the cumulative failures stopping rule as long as it can be assumed that intrusions are relatively less active than words from the target list. This seems a reasonable assumption as models taking into account multiple sources of association (context, experimental word associations, and semantic word associations) often a assume a multiplicative use of search cues (Sirotin, Kimball, \& Kahana, 2005; Kimball, Smith, \& Kahana, 2007) and intrusions should at least have a lower association strength to the target context than list words.

In contrast, the model does not account for the observed impact of intrusions on subsequent retrieval. That is, it cannot account for the greater probability of retrieving a PLI when the immediately preceding item is a PLI. Therefore, it might be necessary to assume that the retrieval of an intrusion changes the contextual landscape which determined the probability of retrieving subsequent PLIs and target list items. Provided this extra assumption, the cumulative failure stopping rule should have no trouble accounted for the present results.

## Conclusion

The results from three experiments using the openinterval design replicate the results from a large-scale re-analysis of closed-interval experiments (Miller et al., 2012) that participants are more likely to terminate search after an intrusion (either from previous list or from outside the experiment) than after a correct recall. Furthermore, search was terminated more quickly after an intrusion (smaller exit latency) and the time taken to generate an intrusion as the final output was greater (greater time-to-last retrieval). These results were found to be consistent with the predictions of a sample-withreplacement model equipped with a cumulative failure stopping rule. As such, the results suggest that the role
of intrusions in search termination is an indirect one. Intrusions could lead to a greater probability of retrieval failures, with the total number of retrieval failures still being the direct cause of search termination.

## Acknowledgements

We gratefully acknowledge NSF (grant BCS-1030831) for support of this research.

## References

Dougherty, M. R., \& Harbison, J. I. (2007). Motivatied to retrieve: How often are you willing to fo back to the well when the well is dry? Journal of Experimental Psychology: Learning, Memory, and Cognition, 33, 1108-1117.
Harbison, J. I., Dougherty, M. R., Davelaar, E. J., \& Fayyad, B. (2009). On the lawfulness of the decision to terminate memory search. Cognition, 111, 416-421.
Harbison, J. I., Hussey, E. K., Dougherty, M. R., \& Davelaar, E. J. (2012). Self-terminated vs. experimenterterminated memory search. In N. Miyake, D. Peebles, \& R. P. Cooper (Eds.), Proceedings of the 34th annual conference of the cognitive science society (pp. 426431). Austin, TX: Cognitive Science Society.

Howard, M. W., \& Kahana, M. J. (2002). A distributed representation of temporal context. Journal of Mathematical Psychology, 46, 269-299.
Kimball, D. R., Smith, T. A., \& Kahana, M. J. (2007). The fSAM model of false recall. Psychological Review, 114, 954-993.
Miller, J. F., Weidemann, C. T., \& Kahana, M. J. (2012). Recall termination in free recall. Memory $\&$ Cognition, 40, 540-550.
Raaijmakers, J. G., \& Shiffrin, R. M. (1981). Search of associative memory. Psychological Review, 88, 93134.

Rohrer, D. (1996). On the relative and absolute strength of a memory trace. Memory $\&$ Cognition, 24, 188201.

Sirotin, Y. B., Kimball, D. R., \& Kahana, M. J. (2005). Going beyond a single list: Modeling the effects of prior experience on episodic free recall. Psychonomic Bulletin \& Review, 12, 787-805.
Unsworth, N., Brewer, G. A., \& Spillers, G. J. (2011). Factors that influence search termination decisions in free recall: An examination of response type and confidence. Acta Psychologica, 138, 19-29.
Wilson, M. D. (1988). The MRC psycholinguistic database: Machine readable dictionary, version 2. Behavioral Research Methods, Instruments and Computers, 20, 6-11.

# Autism, optimism and positive events: Evidence against a general optimistic bias 

Adam J. L. Harris (adam.harris@ucl.ac.uk)<br>Department of Cognitive, Perceptual \& Brain Sciences, University College London, Gower Street, London, UK<br>Punit Shah (ps00113@surrey.ac.uk) and Caroline Catmur (c.catmur@surrey.ac.uk)<br>Department of Psychology, University of Surrey, Guildford, UK

Geoffrey Bird ${ }^{1}$ (geoff.bird@kcl.ac.uk) and Ulrike Hahn (u.hahn@bbk.ac.uk)
Department of Psychological Sciences, Birkbeck College, University of London, Malet Street, London, UK


#### Abstract

The conclusion that people are optimistic concerning personal risk does not have a sound evidential basis. Following Harris and Hahn's (2011) critique of unrealistic optimism research, we consider the evidence from a recent series of high profile neuroscience papers. We demonstrate that the methods used are fundamentally flawed. A simulation and an empirical comparison of autism spectrum condition participants with typical adults confirm that we have learnt nothing about optimism from these studies.


Keywords: Optimism; human rationality; belief updating; statistical artifact

## Introduction

It is a long-established 'fact' that human judgment is biased when estimating personal risk. It has been argued that people underestimate their risk of encountering negative events (compared to their estimate of the average person's risk), and overestimate their chances of encountering positive events (e.g., Weinstein, 1980). This pattern of optimistic self-estimates has been termed 'unrealistic optimism'.

The majority of evidence for unrealistic optimism originated from research employing what we will term the 'comparison method'. Participants are asked to directly (Weinstein, 1980) or indirectly (e.g., van der Velde, van der Pligt, \& Hooykas, 1994; for a review, see Helweg-Larsen \& Shepperd, 2001) compare their chance of experiencing a negative life event with that of the average individual. It is argued that the existence of unrealistic optimism is supported by a pattern of results where the mean of participants' estimates of self-risk is significantly lower than the mean of their estimates of the average person's risk. However, a recent analysis of studies using the comparison method casts doubt over the existence of unrealistic optimism. Harris and Hahn (2011) showed that perfectly rational (non-biased) agents generated personal risk estimates that would be classified as unrealistically optimistic on the comparison method for purely statistical reasons. Unfortunately, the comparison method has been used in the vast majority of unrealistic optimism studies to date, meaning that rather than being a characteristic of
healthy human thought (e.g., Sharot, 2012; Taylor \& Brown, 1988), 'unrealistic optimism' may instead be a statistical artifact resulting from a flawed methodology.

## Identifying Optimism: A New Development

Unrealistic optimism is widely regarded as a central cognitive bias, yet the fact that almost all past research is based on the comparison method means that there is very little empirical evidence for its existence. The frailties associated with the comparison method have, however, only recently been identified, and common belief amongst both researchers and the general public is that risk estimates are characterised by an optimism bias (e.g., Sharot, 2012).

With this position as their starting point, Sharot and colleagues (Sharot, Guitart-Masip et al., 2012; Sharot, Kanai, et al., 2012; Sharot, Korn, \& Dolan, 2011) have begun an investigation into the neural underpinnings of the phenomenon. Sharot et al. (2011) claimed that people maintain an optimistic view of the future through selectively incorporating new desirable information into their beliefs more than new undesirable information. If this result is robust, it seems to shift the body of evidence back in favor of an optimistic bias in human likelihood judgments.

The Update Method We shall refer to the method introduced in Sharot et al. (2011; see also, Sharot, GuitartMasip, et al., 2012; Sharot, Kanai, et al., 2012) as the update method (Figure 1). Participants estimate their chance of experiencing each of a series of negative events. They are subsequently given the probabilities with which these negative events are experienced by the average individual (base rates). Subsequently, participants re-estimate their chances of experiencing the same negative life events. The critical variable is the degree to which participants update their personal risk estimates in light of the given information about the average person's risk. Updates are compared across two types of trial: trials where the base rate of the negative event was lower than the participant's initial estimate ('desirable information'), and trials where it was higher ('undesirable information'). Participants are found to update their estimates more in response to desirable

[^78]information, thus enabling the preservation of an unrealistically optimistic view of the future.


Figure 1: The update method.

## Overview

The present paper proceeds in two stages. In the first stage, we present a formal critique of the update method. We subsequently demonstrate (Study 1) that, as a result of the identified flaws, the update method gives rise to a seemingly optimistic pattern of belief updating in a population of rational Bayesian agents. Finally, in anticipation of a counter-argument that the demonstration of neural moderators of the effect (e.g., Sharot, Kanai, et al., 2012) provides further evidence for its existence, we present results from a study comparing autism spectrum condition (ASC) participants with typical adult (TA) controls. With a suitable addition to the update method, we demonstrate that we can infer nothing about the neural correlates of optimism from the results currently present in the literature.

## What's Wrong With the Update Method?

There are two distinct ways in which we might receive new information relevant to our individual risk. We may receive new information about the base rate, or we may receive diagnostic information that pertains to us personally (e.g., genetic vulnerability). Both types of information are relevant, both can be desirable or undesirable, and both should be combined via Bayes' Theorem to provide our best estimate of risk (e.g., Kahneman \& Tversky, 1973), as is illustrated in the following example.

Normative Risk Updating 55-year-old Tim estimates that the average 55 -year-old's risk of contracting heart disease (HD) (the base rate) is 20\%. In the absence of any other information, Tim's best estimate of his own likelihood of contracting HD is $20 \%$.

If Tim possesses any diagnostic information that differentiates his risk from the average person's, he should normatively combine the base rate with this diagnostic information. For example, if he does not have a family history of HD, his risk is lower than the average person's. Bayes' Theorem prescribes how this information should be combined (e.g., Kahneman \& Tversky, 1973):

$$
\begin{equation*}
P(h \mid e)=\frac{P(h) P(e \mid h)}{P(h) P(e \mid h)+P(\neg h) P(e \mid \neg h)} \tag{1}
\end{equation*}
$$

Bayes' Theorem prescribes the probability, $P(h \mid e)$, of experiencing an event $h$ (e.g., HD) in light of evidence $e$ (no family history of HD). The best estimate of experiencing that event is a function of the base rate of the event, $P(h)$, and the diagnosticity of the evidence - the likelihood ratio of true positives, $P(e \mid h)$, relative to false positives, $P(e \mid \neg h)$. In Tim's case, $P(e \mid h)$ reflects how likely an HD patient is to have no family history of HD, whereas $P(e \mid \neg h)$ reflects the probability of no family history of HD in those who do not contract HD. From a longitudinal study (Hawe, Talmud, Miller, \& Humphries, 2003), we can calculate $P(e \mid h)=.52$ and $P(e \mid \neg h)=.66$. Tim's estimate of $P(h)$ is $20 \%$ and therefore his best estimate of his chance of contracting HD is:

$$
\begin{equation*}
\frac{.2 \times .52}{.2 \times .52+.8 \times .66}=16 \% \tag{2}
\end{equation*}
$$

The base rate of HD is actually $33 \%$ for 55 year-old males (Bleumink et al., 2004). If Tim receives this information, he should recalculate his personal risk once more, using Bayes' Theorem, replacing his previous base rate estimate (20\%) with $33 \%$, which will result in an increased 'best estimate':

$$
\begin{equation*}
\frac{.33 \times .52}{.33 \times .52+.67 \times .66}=28 \% \tag{3}
\end{equation*}
$$

Given the two basic components to normative probability judgments - base rates and evidence - there are thus two ways to receive undesirable (desirable) new information: One can receive new, diagnostic information which suggests that one is more (less) at risk, or one may discover that the base rate is higher (lower) than previously thought.

In Equation 2, Tim knows the accurate base rate, calculates his personal risk rationally, and yet his personal risk is different from the base rate. In contrast to the central assumption of the update method, individuals should not necessarily change their estimate of personal risk simply because it lies above or below the base rate. Researchers can only prescribe what effect the new base rate information should have on a participant's risk estimate if they know the participant's previous estimate of the base rate. Without this knowledge, it is impossible to classify a particular trial as 'desirable' or 'undesirable' and therefore to say in which direction (and how much) the participant's estimate should change. Study 1 highlights the flaws in the update method by simulating optimistic data from non-biased agents!

## Study 1: A Simulation

Take a hypothetical sample of 100 Bayesian agents, 25 of whom assume base rates of $.1, .2, .3$, and .4 respectively (mean $=.25$ ) for Disease $X$, which has a true base rate of .25. Before the study, these agents receive evidence
reflecting their vulnerability to Disease X with the following characteristics: $P(e \mid h)=.5 ; P(\neg e \mid \neg h)=.9$.
$P(h) P(e \mid h)+P(\neg h) P(e \mid \neg h)$ defines the proportion who received evidence suggesting increased risk, here: . $25 \times .5+$ $.75 \times .1=.2$. Thus $20 \%$ of agents have evidence suggesting they will get the disease ('positive evidence') and $80 \%$ have evidence suggesting they will not ('negative evidence'). At each base rate, 5 agents will receive positive evidence, and 20 will receive negative evidence.

In the study (Table 1), agents calculate their initial risk estimates normatively via Bayes’ Theorem, using their subjective base rates; their second estimate recalculates Bayes' Theorem using the experimenter-provided true base rate (see Equations 2 and 3).

Agents whose subjective base rate estimates were below the true base rate (.25) truly receive undesirable information: Disease X is more prevalent than they thought. Agents whose subjective base rate estimates were above the true base rate truly receive desirable information: Disease X is less prevalent than they thought. However, the update method classifies information as 'desirable' or 'undesirable' based on the relationship between initial estimate and true base rate, thus misclassifying $30 \%$ of the sample (grey columns).

The final experimental 'result' is obtained by averaging across those agents receiving 'desirable' and 'undesirable' information (as defined by the experimenter). As each positive evidence group represents 5 agents, and each negative group 20 agents, the resulting absolute means are: Desirable Group $=0.04$; Undesirable Group $=0.03$. Thus, these rational agents show 'greater updating' in response to 'desirable' than 'undesirable' information and would be labelled optimistic - though rational by definition - simply due to incorrect classification. Although this is a somewhat small effect, it becomes much more pronounced when base rate estimates are regressive toward the midpoint of the scale, as is typical of people's probability estimates in many
contexts (see e.g., Harris \& Hahn, 2011; Moore \& Small, 2008, and references therein). If the true base rate were .21 , i.e., below the agents' mean estimate, then the seeming difference in updating rises to $8 \%$ ('desirable’ = . 084 ; 'undesirable' $=.004$ ) - easily able to account for extant experimental data. Figure 2 demonstrates that this pattern of results is not dependent on the precise parameters used in this illustrative example. The preponderance of positive differences in updating (where people update more in response to desirable than undesirable information) is clear from both Figure 2a (where mean estimates of the underlying base rate are correct) and Figure 2b (where base rate estimates are regressive, \& consequently represents more realistic simulations).

## Implications of the Update Method's Flaws

The key implication of the simulation results presented above is that a difference in participants' amount of updating in response to 'desirable' and 'undesirable' information, as defined in Sharot et al. (2011; see also, Sharot, Guitart-Masip et al., 2012; Sharot, Kanai et al., 2012), cannot be taken as evidence of optimism. As with the statistical confounds in the comparison method, however, there is a simple test that disassociates the predictions of an optimism account from a statistical artifact account: the inclusion of positive events (Harris \& Hahn, 2011). Our simulation result is blind to the valence of the events being judged. Consequently, for positive events the statistical account again predicts that updating will be greater in response to information that the base rate is lower than you originally believed your own risk to be. For positive events, which we want to experience, these would be classified as 'undesirable' trials, and hence this pattern of results would suggest pessimism. Thus, updating is undoubtedly optimistic (despite the flaws in the method) if participants update more in response to desirable than undesirable information for both negative and positive events.

Table 1: Artifactual Unrealistic Optimism in Study 1

| Those with positive evidence |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Subjective base rate | $\begin{gathered} 0.1 \\ (n=5) \\ \hline \end{gathered}$ | $\begin{gathered} 0.2 \\ (n=5) \\ \hline \end{gathered}$ | $\begin{gathered} 0.3 \\ (n=5) \\ \hline \end{gathered}$ | $\begin{gathered} 0.4 \\ (n=5) \\ \hline \end{gathered}$ | 0.1 <br> $(n=20)$ <br> 0.058 | 0.2 <br> $(n=20)$ | 0.3 <br> $(n=20)$ <br> 0.192 | $\begin{gathered} 0.4 \\ (n=20) \\ \hline \end{gathered}$ |
|  |  |  |  |  |  |  |  |  |
| Initial Estimate | 0.357 | 0.556 | 0.682 | 0.769 | 0.058 | 0.122 | 0.192 | 0.270 |
| True Base Rate | 0.25 | 0.25 | 0.25 | 0.25 | 0.25 | 0.25 | 0.25 | 0.25 |
| Experimenter-defined Desirability ${ }^{\text {a }}$ | Des | Des | Des | Des | Undes | Undes | Undes | Des |
| True Desirability ${ }^{\text {a }}$ | Undes | Undes | Des | Des | Undes | Undes | Des | Des |
| Correctly Classified? | NO | NO | YES | YES | YES | YES | NO | YES |
| Final Estimate | 0.625 | 0.625 | 0.625 | 0.625 | 0.156 | 0.156 | 0.156 | 0.156 |
| Amount of update (IE - FE) | -0.268 | -0.069 | 0.057 | 0.144 | -0.098 | -0.034 | 0.036 | 0.114 |

[^79]Furthermore, Figure 2 demonstrates how a non-selective change to a probabilistically relevant quantity (e.g., changing the perceived diagnosticity of individuals' evidence; or altering the regressiveness of their initial base rate estimates) that affects all agents equally, can lead to a seemingly selective effect: a sharp increase in the difference between updating for 'desirable' vs. 'undesirable' information. Thus, the selective effects of L-DOPA (Sharot, Guitart-Masip et al., 2012) and transcranial magnetic stimulation (TMS) (Sharot, Kanai, et al., 2012) on belief updating might be entirely unrelated to optimism, and simply reflect (for example) better learning (i.e., less conservative updating - formally equivalent to an increase in the diagnosticity of information) following receipt of L DOPA.


Figure 2: Seemingly 'optimistic' updating. Positive values in the z-axes demonstrate 'optimistic' updating for Bayesian agents receiving evidence with the properties shown in the x-axes, $P(e \mid h)$, and y-axes, $P(\neg e \mid \neg h)$. Panel a), agents estimate the base rate as either: $.1, .2, .3, .4$, and the true base rate $=.25$ (as in Table 1). Panel b), true base rate $=.21$. The mass of data points in both plots spuriously suggests optimistic updating.

## Study 2: An Experiment

The results reported above led us to run an experiment in which we followed Sharot et al.'s (e.g., 2011) method, but crucially included a selection of positive events in addition to the negative events. We were further interested in a comparison of ASC participants, with typical adult (TA) controls. Sharot et al. (2011) reported that optimistic updating was largely mediated by the medial prefrontal cortex. Given that ASC is associated with hypoactivity in this region across a range of tasks (Gilbert et al., 2008), and ASC participants have been suggested to be less susceptible to emotional biases in decision making (de Martino et al., 2008), one might predict these individuals to show less of an optimistic bias, similar to the attenuation of the effect in participants who have received TMS to the left inferior frontal gyrus (Sharot, Kanai, et al., 2012). Should this effect be observed, we are crucially interested in the results from the positive events in this study. If TAs do show a seeming pessimism bias (as predicted by the statistical account), will the individuals with ASC show more of a 'pessimism bias' (as they appear to be less optimistic than TAs with negative events), or will this effect also be attenuated - suggesting that ASC has not had a selective effect on 'optimism'?

## Method

Participants 20 male participants with a mean age of 34 years participated in return for $£ 10$. The 10 ASC participants all received diagnoses from an independent clinician. Of the 10, using the Autism Diagnostic Observation Schedule (Lord et al., 2000), 5 met the criteria for autism, and 5 for autistic spectrum disorders. The ASC and NT participants were matched for IQ (Weschler, 2011) and age, $t \mathrm{~s}<1$.

Design The experiment was a $2 x 2 x 2$ (group $x$ event type [positive/negative] $x$ information desirability) design, with event type and information desirability manipulated withinparticipants. In a departure from Sharot et al.'s (e.g., 2011) method, the base rates presented to participants were derived from participants' initial estimates of their own risk (IE1), rather than being externally sourced. This enabled the experimental manipulation of information desirability, as well as controlling for the potential differential accuracy of base rate information pertaining to negative events (for which there is plenty of information) versus positive events (for which there is a scarcity of information). Base rates were computed according to the following formula: A random percentage between $17 \%$ and $40 \%$ (uniform distribution) of IE was either added to, or subtracted from, the IE, according to trial type, and rounded to the nearest integer. Thus, for example, on positive desirable trials a random percentage of the IE was added to the IE resulting in a derived probability indicating that the positive event was more likely to occur than had previously been estimated, and on negative desirable trials a random percentage of the IE was subtracted from the IE, indicating that the negative event was less likely than had been estimated. All
probabilities were capped between $3 \%$ and $80 \%$ as participants were informed that this was the range of possible probabilities (see also, e.g., Sharot et al., 2011).

Stimuli and Procedure Eighty short descriptions of life events, many of which had previously been used in the study of unrealistic optimism (e.g., Sharot et al., 2011; Weinstein, 1980), were presented in a random order. Half of the events were positive and half negative. Very rare or very common events were avoided and all known probabilities lay between $10 \%$ and $70 \%$ ( $M=32.6, S D=18.8$; Office for National Statistics and PubMed). The procedure followed that of Sharot et al. (2011) (see Figure 1), with one addition. Following their IE, participants were asked to estimate the chance of the event occurring to the average person ${ }^{2}$. Following the main task, participants' memory for the base rates presented to them was tested, and they provided four salience ratings for each event ${ }^{3}$. Finally, a funneled debrief procedure (Bargh \& Chartrand, 2000) showed that no participants suspected that the derived base rates were inaccurate.

Scoring For each event the amount of update was calculated first by computing the absolute difference between the updated estimate in the second session and the IE, and second, coding the difference as positive when the update was in the direction of the base rate and negative when the update was away from the base rate. Mean updates for each participant in each condition were then calculated after removal of outliers ( $\pm 3 \times$ the interquartile range) and trials for which a derived probability could not be applied (e.g., when a participant's IE was already at the lowest extreme of the probability range, but the trial-type required that a lower base rate be supplied).

## Results

TA Controls For negative events, we replicated the central result from Sharot et al. (2011). Participants updated more in response to desirable information than to undesirable information, $t(9)=5.02, p=.001$ (see Figure 3). This result demonstrates that the minor changes to the update method's procedure were not consequential for the general paradigm. For positive events, however, the reverse pattern was observed, with greater updating in response to undesirable information than desirable information, $t(9)=4.71, p=$ .001. Across both event types, a significant event type x desirability interaction was observed, $F(1,9)=35.45, p<$ .001, with no main effect of event type or, crucially, information desirability ( $F \mathrm{~s}<1$ ).

TA Controls and ASC Participants Figure 3 shows that ASC participants showed significantly less asymmetry in their information updating than did TAs, with a significant group x event type x desirability interaction, $F(1,18)=$

[^80]

Figure 3: Mean updates in Study 2. Error bars are +/- 1 SE.
4.83, $p<.041$. None of the effects observed in the TAs were observed in the autistic participants ( $p \mathrm{~s}>.17$ ). The significant 3-way interaction demonstrated that updating in the ASC group was different from in the TA group, but there was no evidence that they were either more or less optimistic.

## Discussion

Sharot, Kanai et al. (2012) administered TMS to the left IFG in one group of their participants. Participants subsequently completed an update task comprised entirely of negative events. Participants who had received TMS to the left IFG were shown to exhibit less of an asymmetry in updating to desirable versus undesirable information, and it was therefore concluded that the left IFG is involved in the suppression of updating beliefs about vulnerability in response to undesirable information. Study 2 can be seen as analogous to this study, with the presence or absence of ASC the quasi-experimental variable in place of the location of TMS. Study 2 included positive events in addition to negative events. For positive events, TAs did not show an optimistic bias in belief updating, but rather a seeming pessimistic bias. This resulted in a $2 x 2$ interaction, such that TAs updated more in response to desirable information for negative events, but more in response to undesirable information for positive events. It was this interaction, rather than a specific optimism bias that was attenuated in the ASC participants. One can only speculate whether the same effect would occur with TMS of the left IFG.

## Summary

Optimism is currently a research area of much interest. Studying optimism with real-world events is complicated by our inability to know the objective probability of an event in a given individual's future (see also, Weinstein \& Klein, 1996). The very prevalent comparative method has been shown to be insufficient (Harris \& Hahn, 2011), and here
we have demonstrated that a recent ('update') method also fails as a test of human rationality, through 'demonstrating' that rational agents are irrational (Study 1)! Despite this weakness, the addition of positive events can determine whether there is evidence for an optimistic bias. With this addition, no evidence is found (Study 2). The inclusion of ASC participants in Study 2 also demonstrated the difficulties inherent in making conclusions about the effects of individual differences or experimental manipulations on optimistic belief updating, without the inclusion of both positive and negative events.

Our resulting conclusion, that there is no evidence for a general optimistic bias in human likelihood judgments, might seem surprising in the light of historical common belief. It is, however, consonant with recent results where participants overestimate the likelihood of negative events relative to neutral events in the laboratory (e.g., Bilgin, 2012; Harris, Corner, \& Hahn, 2009; Risen \& Gilovich, 2007; Vosgerau, 2010). On the basis of the extant evidence, we cannot conclude that an optimistic view of the future characterises healthy human thought (Taylor \& Brown, 1988).

## References

Bargh, J.A., \& Chartrand, T.L. (2000). The mind in the middle: A practical guide to priming and automaticity research. In H.T. Reis \& C.M. Judd (Eds.), Handbook of research methods in social and personality psychology. Cambridge: Cambridge University Press
Bilgin, B. (2012). Losses loom more likely than gains: Propensity to imagine losses increases their subjective probability. Organizational Behavior and Human Decision Processes, 118, 203-215.
Bleumink, G. S., Knetsch, A. M., Sturkenboom, M. C. J. M., Straus, S. M. J. M., Hofman, A., et al. (2004). Quantifying the heart failure epidemic: Prevalence, incidence rate, lifetime risk and prognosis of heart failure. European Heart Journal, 25, 1614-1619.
De Martino, B., Harrison, N., Knafo, S., Bird, G., and Dolan, R. (2008). Explaining enhanced logical consistency during decision making in autism. Journal of Neuroscience, 28, 10746-10750.
Gilbert, S.J., Bird, G., Brindley, R., Frith, C.D., and Burgess, P.W. (2008). Atypical recruitment of medial prefrontal cortex in autism spectrum disorders: An fMRI study of two executive function tasks. Neuropsychologia, 46, 2281-2291.
Harris, A. J. L., Corner, A., \& Hahn, U. (2009). Estimating the probability of negative events. Cognition, 110, 51-64.
Harris, A. J. L., \& Hahn, U. (2011). Unrealistic optimism about future life events: A cautionary note. Psychological Review, 118, 135-154.
Hawe, E., Talmud, P. J., Miller, G. J., \& Humphries, S. E. (2003). Family history is a coronary heart disease risk factor in the second Northwick Park heart study. Annals of Human Genetics, 67, 97-106.

Helweg-Larsen, M., \& Shepperd, J. A. (2001). Do moderators of the optimistic bias affect personal or target risk estimates? A review of the literature. Personality and Social Psychology Review, 5, 74-95.
Kahneman, D., \& Tversky, A. (1973). On the psychology of prediction. Psychological Review, 80, 237-251.
Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Jr., Leventhal, B. L., DiLavore, P. C., et al. (2000). The autism diagnostic observation schedule-generic: A standard measure of social and communication deficits associated with the spectrum of autism. Journal of Autism and Developmental Disorders, 30, 205-223.
Moore, D. A., \& Small, D. A. (2008). When it is rational for the majority to believe that they are better than average. In J. I. Krueger (Ed.), Rationality and Social Responsibility. Essays in Honor of Robyn Mason Dawes. New York: Psychology Press.
Risen, J. L., \& Gilovich, T. (2007). Another look at why people are reluctant to exchange lottery tickets. Journal of Personality and Social Psychology, 93, 12-22.
Shah, P., Harris, A. J. L., Hahn, U., Catmur, C., \& Bird, G. (2013). A pessimistic view of optimistic belief updating. Manuscript submitted for publication.
Sharot, T. (2012). The Optimism Bias: Why we're wired to look on the bright side. London: Constable \& Robinson.
Sharot, T., Guitart-Masip, M., Korn, C. W., Chowdhury, R., \& Dolan, R. J. (2012). How dopamine enhances an optimism bias in humans. Current Biology, 22, 14771481.

Sharot, T., Kanai, R., Marston, D., Korn, C. W., Rees, G., \& Dolan, R. J. (2012). Selectively altering belief formation in the human brain. Proceedings of the National Academy of Sciences of the United States of America, 109, 1705817062.

Sharot, T., Korn, C. W., \& Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. Nature Neuroscience, 14, 1475-1479.
Taylor, S. E., \& Brown, J. D. (1988). Illusion and wellbeing: A social psychological perspective on mental health. Psychological Bulletin, 103, 193-210.
Van der Velde, F. W., van der Pligt, J., \& Hooykas, C. (1994). Perceiving AIDS-related risk: Accuracy as a function of differences in actual risk. Health Psychology, 13, 25-33.
Vosgerau, J. (2010). How prevalent is wishful thinking? Misattribution of arousal causes optimism and pessimism in subjective probabilities. Journal of Experimental Psychology: General, 139, 32-48.
Weinstein, N. D. (1980). Unrealistic optimism about future life events. Journal of Personality and Social Psychology, 39, 806-820.
Weinstein, N. D., \& Klein, W. M. (1996) Unrealistic optimism: Present and future. Journal of Social and Clinical Psychology, 15, 1-8.
Weschler, D. (2011). Weschler Abbreviated Scale of Intelligence (2 ${ }^{\text {nd }}$ Edition). San Antonio, TX: Pearson.

# The Neural Computation of Scalar Implicature 

Joshua K. Hartshorne (jkhartshorne@gmail.com)<br>Department of Brain and Cognitive Sciences, MIT, 77 Massachusetts Avenue Cambridge, MA 02139 USA

Jesse Snedeker (snedeker@wjh.harvard.edu)<br>Department of Psychology, Harvard University, 33 Kirkland St. Cambridge, MA 02138 USA

Albert Kim (albert.kim@colorado.edu)<br>Institute of Cognitive Science, University of Colorado-Boulder, 594 UCB<br>Boulder, CO 80309 USA


#### Abstract

What psychological and linguistic processes allow one to go beyond the literal meaning of a sentence and infer what was meant but not said ("reading between the lines")? Theorists have differed as to whether these phenomena are driven by complex, online inference processes or by relatively rote rules. The present study uses ERP to explore the cognitive and neural mechanisms involved in scalar implicature (the inference that, e.g., "some" indicates "some but not all"), a test case that has been subject to considerable behavioral research but limited neuropsychological research. Our results challenge both rote-processing and rich-inference accounts. We provide the first ERP results showing that scalar implicature processing depends on context, challenging roteprocessing theories of implicature. However, we also fail to find evidence of a processing cost associated with implicature processing, as predicted by many rich-inference accounts. These results point to a novel conceptualization of pragmatic processing in scalar implicature.


Keywords: language; pragmatics; ERP; scalar implicature

## Rich vs. Rote Pragmatics

Understanding language appears to involve two broad but distinct kinds of processes: derivation of the semantic meaning (those things entailed to be true) and pragmatic inferences that go beyond this "literal" meaning (Bach, 1999; Grice, 1989; Morris, 1938). For example, given sentence (1), the fact that Gabe is the agent of the drinking event would typically be attributed to semantic decoding, while the inference that he is an inconsiderate lout who has annoyed the speaker would generally be construed as pragmatic.
(1) Gabe drank all of the milk and put the carton back in the fridge.

There is, however, considerable controversy about where semantics ends and pragmatics begins and about how to distinguish the representations and processes underlying each, as well as their interaction. One particularly contentious point is whether pragmatic inferences result from complex, rich reasoning processes (Grice, 1989; Sperber \& Wilson, 1986) or from relatively rote, automatic, almost grammatical rules (Chierchia, Fox, \& Spector, 2012; Levinson, 2000).

Perhaps the most-researched test case is scalar implicature, illustrated in (2):
(2) a. John ate some of the cookies.
b. John did not eat all of the cookies.

When we hear a sentence like (2a), we typically assume that (2b) is true as well. Although this inference is robust, it can be cancelled (3a), distinguishing it from semantic entailments, which cannot be cancelled (3b) (Hirschberg, 1991; Horn, 1972).
(3) a. John ate some of the cookies. In fact, he ate all of them.
b. *John ate some of the cookies. In fact, he ate none of them.

On the classical view (Horn, 1972), scalar implicature requires rich online counter-factual reasoning: Listeners only infer (2b) from (2a) if they believe i) the speaker knows whether John ate all the cookies, ii) it is relevant whether John ate all the cookies, and iii) assuming (a-b) hold, the speaker would tell them that John ate all the cookies. This view has been questioned, originally by Levinson (2000), who argued that scalar implicatures are triggered automatically, prior even to compositional processing (i.e., processing language at the level of phrases or sentences).

Much of the work addressing this theoretical dispute has been indirect, testing whether scalar implicatures are slow and computationally costly as a proxy for being rich and complex (Bott \& Noveck, 2004; De Neys \& Schaeken, 2007; Grodner, Klein, Carbary, \& Tanenhaus, 2010; Huang \& Snedeker, 2009). Results have been inconsistent and controversial. More problematically, the link between "slow and costly" and "rich and complex" can be disputed: Grodner and colleagues (2010) argue that scalar implicature is rich, complex, and fast; similarly nothing in principle forbids an automatic process from being slow.

A more direct route is as follows: If scalar implicature is the result of a complex inference process, it should be possible to create contexts in which scalar implicatures are more or less likely to be calculated. If, on the other hand, scalar implicature is an automatic process, it should be relatively impervious to context. A handful of behavioral
studies have reported contextual manipulations that affect scalar implicature calculation (Bergen \& Grodner, 2012; Bonnefon, Feeney, \& Villejoubert, 2009; Hartshorne \& Snedeker, submitted; Noveck, Chierchia, Chevaux, Guelminger, \& Sylvestre, 2002); we return to these in the Discussion. Nothing is known on a neuropsychological level about scalar implicature's context sensitivity, as no such studies have been reported. ${ }^{1}$ Thus, we conducted the present study in order to confirm the behavioral results and extend them to the neuropsychological level.

## Grammatical Context

We compare the ERPs elicited by carefully matched sentences that do or do not evoke scalar implicatures. Our method derives from previous findings that scalar implicatures are more likely to be calculated in declarative sentences (4a) than in the antecedent of a condition (4b).
(4) a. Addison ate some of the cookies before breakfast this morning, and...
b. If Addison ate some of the cookies before breakfast this morning, then...

This has been explained by Chierchia and colleagues (2012) as an effect of entailment context. Scalar implicature usually operates to deny the truth of a logically stronger statement. Since Addison ate all of the cookies entails that Addison ate some of the cookies, stating the latter implicates that the former is not true. In contrast, If Addison ate all of the cookies, then $Q$ does not entail but rather is entailed by If Addison ate some of the cookies, then $Q$; thus stating the latter does not implicate that the former is not true.

While this entailment manipulation is linguistic in nature, it is nonetheless difficult to account for on a strict automatic processing account like Levinson's (2000), on which scalar implicature is triggered lexically prior to any compositional processes - that is, before sentential context, which is by definition compositional, can play a role.

While intuitions that conditional sentences suppress implicature seem robust, experimental evidence consists of a single published judgment study (Noveck et al., 2002). ${ }^{2}$ Thus at best we do not know whether the entailment manipulation in (4) affects scalar implicature online.

The present study addresses this gap as follows:
(5) a. Addison ate some of the cookies before breakfast this morning, and the rest are on the counter.
b. If Addison ate some of the cookies before breakfast this morning, then the rest are on the counter.

[^81]Note that the rest is only felicitous if Addison has not eaten all of the cookies, which is exactly what the scalar implicature implies; thus, by hypothesis the rest should be more felicitous in (5a) than (5b). Thus, by testing whether the ERPs to the rest are different in (5a) and (5b), we test whether entailment context rapidly modulates scalar implicature, affecting interpretation of content later in the sentence. In addition, by comparing ERPs at some - the word that triggers the scalar implicature - we will gain valuable information about the neural processes supporting scalar implicature calculation.

One methodological concern remains: Declarative and conditional sentences differ in numerous ways, not just in how they affect scalar implicature. Thus, any differences observed may be due to implicature-irrelevant processes. Thus, we included a control version of the experiment, where some was everywhere replaced with only some, a phrase that semantically forces the subset ("not all") reading. Thus, the crucial analyses are interactions differences seen between the critical declarative and conditional sentences not seen between the control declarative and conditional sentences.

## Method

## Subjects

49 monolingual native English-speaking right-handed individuals participated. Two were excluded for equipment failure and ten for excessive artifact, leaving 19 in the experimental condition and 16 in the control condition.

## Materials and Procedure

Each participant saw 30 critical declarative sentences and 30 critical conditional sentences. Filler sentences consisted of 60 matched in structure - but not content - to the critical sentences but with continuations that did not mention "the rest" and 35 which additionally swapped the word some for all. These fillers prevented subjects from inferring that all sentences would refer to "the rest" of a previouslymentioned collection. An additional 42 filler sentences involved relative clauses and no quantifiers. Four lists were created by converting the critical declarative sentences into conditional sentences and vice versa and by reversing the order of all stimuli (except the first four stimuli, which were always the same fillers). The four experimental and four control lists were identical except that in the latter, the word some was always preceded by only.

Sentences were presented in eight blocks, with breaks in between. 61 of the sentences were followed by comprehension questions, which were not analyzed. Sentences were presented roughly one word at a time. Wherever two short words appeared consecutively, we presented them together (e.g., Sally/saw/a cat/on the/table). This allowed us to present the critical phrase the rest as a single unit, rather than in two parts which would potentially add noise to the ERP. Some was always presented singly. Stimuli were presented in the center of the screen for 350
ms with a 250 ms blank interval between words. The intertrial interval ranged from 1600 to 2000 ms , not counting any time spent on questions.

## Acquisition and Analysis

Ongoing EEG was recorded from 128 scalp locations using a geodesic sensor net (Electrical Geodesics, Eugene, OR) as subjects read the sentences silently. EEG was recorded relative to a vertex channel and later re-referenced to the average of the mastoid channels. Impedances were maintained below $100 \Omega$. Signals were recorded at 250 Hz and down-sampled to 200 Hz post-acquisition. A $0.1-30 \mathrm{~Hz}$ bandpass filter was applied. Epochs of 1500 ms were selected following the critical phrase (some or the rest) and were corrected with a 200 ms pre-stimulus baseline. Bad channels were replaced and epochs containing artifact (eye blink, eye movement, etc.) removed, both by computer algorithm. Only participants with at least 19 epochs per cell were included in analyses.

## The Bootstrap Cluster Algorithm

The previous literature has focused on the role of the N400 in processing scalar implicature violations. Because no previous study has looked for components indexing scalar implicature generation, we needed a mechanism for selecting and analyzing exactly those electrodes in those time periods with the greatest differences between conditions without allowing multiple comparisons to inflate our Type I error rate (cf Vul, Harris, Winkielman, \& Pashler, 2009). We adapted the recently developed bootstrap cluster analysis of Maris and Oostenveld (2007).

We calculated the context by condition interaction using a mixed effects model with maximal random effects for each electrode at each time point (to speed processing, we further down-sampled the data to 50 Hz ) and recorded the t-value. We then identified all clusters of data points with $t$-values greater than 1.96 or less than $-1.96 .{ }^{3}$ Clustering crossed both time (consecutive super-threshold data points on the same electrode were placed in the same cluster) and space (superthreshold data points from the same time point and belonging to neighboring electrodes were placed in the same cluster). Although data points with positive effects (positive t -values) may represent the same underlying dipole as data points with negative effects (negative $t$-values), we adopted the conservative approach of placing $t$-values of different polarities in different clusters. Clusters are assigned scores, which are the sum of their t -values; thus, clusters with larger statistical effects and/or which are extended in time and space are assigned larger scores.

Statistical significance was assessed through bootstrapping. The condition labels for the subjects (experimental/control) were shuffled, as were the context

[^82]codes (declarative/conditional) for each subject's average ERPs. The clustering algorithm was re-run, and the scores for the largest positive and negative clusters were recorded. This process was repeated 100 times. P-values for a given cluster in the actual data are estimated as the number of larger clusters from the bootstrapped data (calculated separately for positive and negative clusters).

## Results

As can be seen in Figure 1, the interaction between context and condition in the ERPs evoked by some were weak, and none of the resulting clusters were significant ( $p \mathrm{~s}>.2$ ).

In contrast, at the rest an interaction was observed, frontally distributed and lasting from approximately 400 to 1300 ms post-stimulus ( $p=.04$; see Figures $2 \& 3$ ). Inspection of the four waveforms for the four conditions revealed at this interaction was due to a positive deflection for the conditional/experimental sentences relative to the other three conditions. That the conditional/experimental sentences should be the odd one out is expected: only in that condition should the rest be difficult to process, and in fact in our norming studies, the conditional/experimental sentences were judged to be less felicitous than the other three types; this effect disappeared if the sentences were truncated prior to the rest.

Thus, we interpret the interaction at the rest to be due to a positive deflection for the conditional/experimental sentences, reflecting the infelicity of the rest, perhaps due to the difficulty assigning its reference.

## Discussion

A previous judgment study (Noveck et al., 2002) found that scalar implicatures were more likely in declarative than conditional sentences. If this is the case, and if this contextual manipulation affects processing rapidly, then the rest should be more difficult to process in (5b) than (5a). Indeed, we found that the contextual manipulation affected the ERPs to the rest. Interestingly, we did not find an effect of the manipulation on the ERPs to the scalar trigger some.

We address theoretical implications of these findings momentarily. First, we consider their robustness. Given recent concern about replicability in the cognitive sciences (Hartshorne \& Schachner, 2012), we conducted a replication closely matched to the above experiment with the following differences: EEG was recorded using $\mathrm{Ag} / \mathrm{AgCl}$ electrodes attached to an elastic cap following the extended 10-20 system (Newer et al., 1998), and blink artifact was corrected through linear regression. We coded the stimuli so that the ERPs to some in the filler sentences - which up through some are indistinguishable from the critical sentences could be included in analysis, doubling the number of trials for that analysis. Analyses were conducted in identical fashion and with the same result, demonstrating their robustness: no significant clusters were found at some ( $p \mathrm{~s}>.15$ ), but an extended, frontally-distributed cluster was found after the rest ( $p<.01$ ).


Figure 1: Bootstrap cluster analyses at some. In each panel, electrodes grouped into left-hemisphere, midline, and righthemisphere, with more anterior electrodes placed higher. Panel A: $t$-values. Panel B: clusters (distinct color for each cluster).


Figure 2: Bootstrap cluster analyses at the rest. In each panel, electrodes grouped into left-hemisphere, midline, and righthemisphere, with more anterior electrodes placed higher. Panel A: $t$-values. Panel B: clusters (distinct color for each cluster).


Figure 3: Difference waves (declarative - conditional) at the rest. Topographical plots are shown at 600 ms poststimulus. Four representative electrodes are depicted for the entire epoch. The relative negativity for the difference waves in the experimental sentences is driven by a positive deflection for the conditional sentences (see main text).

We consider first the positive results. As predicted, scalar implicatures are less likely to be calculated in the antecedents of conditionals (as evidenced by results at the rest), confirming a strong prediction of Chierchia and colleagues' Grammatical Theory. Moreover, context's effect was sufficiently rapid to affect processing of content (the rest) later in the sentence. This is difficult to reconcile with a strong lexicalist position like Levinson's (2000), on which scalar implicatures are always triggered by words like some, though they may be explicitly cancelled as in (3a). Note that not only was the implicature not cancelled in our conditional sentences, not calculating the implicature renders the sentences infelicitous.

Perhaps the most intriguing finding was the lack of any effect at the scalar implicature trigger some. This finding matches those of five self-paced reading experiments involving similar stimuli, for which Hartshorne and Snedeker (submitted) similarly report no effect: some was read no faster or slower whether a scalar implicature was calculated or not. These finding are in apparent conflict with a single experiment by Bergen and Grodner (2012), which showed slower reading times for some in implicaturepromoting conditions, which they interpreted as indexing the computational cost of scalar implicature calculation. However, Bergen and Grodner used a different
manipulation, an issue we return to shortly. ${ }^{4}$
There are at least three logically possible explanations of the result. The first is that the scalar implicature processing's effect on ERP (and self-paced reading) is quite small and thus we had insufficient statistical power to detect it. This would raise an interesting question: Why is the effect so small relative to typical language ERP effects (such as the effect we observed on the rest) which are observable with a study this size?

A second possibility is that scalar implicature is always triggered by some, and thus the ERPs were identical across conditions (Levinson, 2000). As already noted, this runs afoul of our results at the rest; we would have to stipulate that the entailment context acts to cancel the implicature in the conditional sentences. What the mechanism would be is unclear. Moreover, the effect of the cancellation should be measurable, and though we explored ERPs to several words subsequent to some, we saw no evidence of it.

A third possibility is that not computing the scalar implicature is also a complex and lengthy inference, sufficiently similar to actually computing the scalar implicature that the two could not be distinguished in our study. On the Grammatical Theory, the parser attempts to insert scalar implicatures at any appropriate insertion site, and they are retained if certain criteria are met, such as its resulting in a more informative (i.e., logically stronger) interpretation of the utterance. Presumably, the only way the grammar can know that these criteria have been met is to actually carry out the operations; thus, similar work is done whether the scalar implicature is ultimately endorsed or not. Similarly, on the Gricean account, scalar implicatures are calculated only when certain conditions are met (e.g., the speaker would make the stronger statement if it were true and the speaker knows whether the stronger statement is true). Whether these conditions are met affects whether the implicature is endorsed, not whether the complex set of conditions must be checked. In short, even if calculating a scalar implicature is costly, that does not necessarily mean that manipulations which affect whether the implicature is ultimately endorsed affect the computational cost.

Why then did Bergen and Grodner find an effect on some? While we manipulated whether the scalar implicature was appropriate, they manipulated the salience of the stronger alternative (e.g., all). Since scalar implicature processing requires a stronger alternative to get off the ground, their manipulation may have more directly affected whether processing happened at all.

## Conclusion

We find the grammatical entailment context modulates scalar implicature processing rapidly enough to affect

[^83]processing of subsequent words in the sentence. At the same time, this manipulation did not affect the EEG evoked by the scalar implicature trigger (some). These findings present a first step in uncovering the neural processes underlying the factors driving scalar implicature and also present challenges to existing theories.

## Acknowledgments

The authors thank Miki Uruwashi and especially Stephanie Liem for help with data acquisition, and attendees at AMLaP 2012 for comments and suggestions. This work was supported by NDSEG and NSF GRFP fellowships to JKH and NSF-BCS 0921012 to JS.

## References

Bach, Kent. (1999). The semantics-pragmatics distinction: What it is and why it matters. In K. Turner (Ed.), The Semantics-Pragmatics Interface from Different Points of View (pp. 65-84). Oxford, UK: Elsevier.
Bergen, Leon, \& Grodner, Daniel J. (2012). Speaker knowledge influences the comprehension of pragmatic inferences. Journal of Experimental Psychology: Learning, Memory, and Cognition, 38(5), 1450-1460.
Bonnefon, Jean-Francois, Feeney, Aidan, \& Villejoubert, Gaelle. (2009). When some is actually all: Scalar inferences in face-threatening contexts. Cognition, 112, 249-258.
Bott, Lewis, \& Noveck, Ira A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. Journal of Memory and Language, 51, 437-457.
Breheny, Richard, Katsos, Napoleon, \& Williams, John. (2006). Are generalised scalar implicatures generated by default? An on-line investigation into the role of context in generating pragmatic inferences. Cognition, 100, 434-463. doi: 10.1016/j.cognition.2005.07.003

Chierchia, Gennaro, Fox, Danny, \& Spector, Benjamin. (2012). The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. In C. Maienborn, K. v. Heusinger \& P. Portner (Eds.), Semantics: An International Handbook of Natural Language Meaning (Vol. 2, pp. 2297-2332). Boston, MA: De Gruyter Mouton.
De Neys, Wim, \& Schaeken, Walter. (2007). When people are more logical under cognitive load: dual task impact on scalar implicature. Experimental Psychology, 54, 128-133.
Grice, P. (1989). Studies in the way of words. Cambridge, MA: Harvard University Press.
Grodner, Daniel J., Klein, Natalie M., Carbary, Kathleen M., \& Tanenhaus, Michael K. (2010). "Some," and possible all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. Cognition, 116, 42-55.

Hartshorne, Joshua K., \& Schachner, Adena. (2012). Tracking replicability as a method of postpublication open evaluation. Frontiers in Computational Neuroscience, 6(8).
Hartshorne, Joshua K., \& Snedeker, Jesse. (submitted). The speed of inference: Evidence against rapid use of context in calculation of scalar implicature.
Hirschberg, Julia. (1991). A Theory of Scalar Implicature. book, 1-220.
Horn, Laurence R. (1972). On the semantic properties of logical operators in English. book, 1-309.
Huang, Yi Ting, \& Snedeker, Jesse. (2009). Online interpretation of scalar quantifiers: Insight into the semantics--pragmatics interface. Cognitive Psychology, 58, 376-415. doi: 10.1016/j.cogpsych.2008.09.001

Levinson, Stephen C. (2000). Presumptive Meanings: The Theory of Generalized Conversational Implicature, Speech, and Communication. book, 1-480.
Maris, Eric, \& Oostenveld, Robert. (2007). Nonparametric statistical testing of EEG- and MEG-data. Journal of Neuroscience Methods, 164, 177-190. doi: 10.1016/j.jneumeth.2007.03.024

Morris, Charles. (1938). Foundations of the theory of signs. International Encyclopdia of Unified Science, 1(2).
Newer, M. R., Comi, G., Emerson, R., FuglsangFrederiksen, A., Guerit, J. M., Hinrichs, H., . . . Rappelsburger, Peter. (1998). IFCN standards for digital recording of clinical EEG. Electroencephalography and Clinical Neurophysiology, 106, 259-261.
Nieuwland, Mante S., Ditman, Tali, \& Kuperberg, Gina R. (2010). On the incrementality of pragmatic processing: An ERP investigation of informativeness and pragmatic abilities. Journal of Memory and Langauge, 63, 324-346.
Noveck, Ira A, Chierchia, Gennaro, Chevaux, Florelle, Guelminger, Raphaele, \& Sylvestre, Emmanuel. (2002). Linguistic-pragmatic factors in interpreting disjunctions. Thinking and Reasoning, 8, 297-326.
Noveck, Ira A, \& Posada, Andres. (2003). Characterizing the time course of an implicature: an evoked potentials study. Brain and language, 85, 203-210. doi: 10.1016/S0093-934X(03)00053-1
Panizza, Daniele, Chierchia, Gennaro, \& Clifton Jr., Charles. (2009). On the role of entailment patterns and scalar implicatures in the processing of numerals. Journal of Memory and Language, 61, 503-518.
Sperber, D., \& Wilson, D. (1986). Relevance: Communication and cognition. Oxford: Basil Blackwell.
Vul, Edward, Harris, Christine, Winkielman, Piotr, \& Pashler, Harold. (2009). Puzzlingly High Correlations in fMRI Studies of Emotion , Personality, and Social Cognition. Perspectives on Psychological Science, 4, 274-290.

# Causal model and sampling approaches to reducing base rate neglect 

Brett K. Hayes (B.Hayes@unsw.edu.au)<br>Ben R. Newell (Ben.Newell@unsw.edu.au)<br>Guy E. Hawkins (G.Hawkins@unsw.edu.au)<br>School of Psychology, University of New South Wales<br>Sydney, 2052, Australia


#### Abstract

Two studies examined how sampling of base rate information and causal explanation of false positives facilitate intuitive probability judgments. Experiment 1a varied these two manipulations factorially. Each had an additive effect on reducing base rate neglect and increasing choice of the normatively correct solution. Experiment 1 b showed that description of relevant distributional information produced similar facilitation to sequential sampling. These results indicate that causal and sampling approaches impact on different components of probability judgment.


Keywords: Causal reasoning, Sequential sampling, Base rate neglect, Bayesian judgment, Belief updating

## Introduction

One of the most commonly observed biases in human judgment is neglect of relevant base rate information (Eddy, 1982; Gigerenzer \& Hoffrage, 1995; Tversky \& Kahneman, 1974). For example, when people attempt to solve intuitive probability problems like that in Figure 1 (standard version), they typically ignore the low base rate ( $p$ (Cancer) $=.01$ ), generating probability estimates that are much higher than the normative Bayesian solution ( $p$ (Cancer $\mid$ Mammogram + ) $\approx 0.051$, see Appendix for a derivation) .

Previous work has suggested a number of solutions to the problem of base rate neglect. These include the use of frequency rather than probability formats for relevant statistics (Gigerenzer \& Hoffrage, 1995), and instructions that clarify set relations between the relevant samples (Barbey \& Sloman, 2007; Evans, Handley, Perham, Over, \& Thompson, 2000).

Two novel approaches to explaining and reducing base rate neglect have recently been proposed. The first involves consideration of the intuitive causal models that people construct when solving probability problems. Krynski and Tenenbaum (2007) outline a "causal-Bayes" account of probability judgments which assumes that errors arise when the statistics in a given problem do not readily map onto an intuitive causal model. In the standard mammogram problem for example, no causal explanation for the false positive rate (the probability of a positive mammogram in the absence of cancer) is given. According to Krynski and Tenenbaum (2007) this makes it difficult to integrate the false positive rate into Bayesian calculations, leading to inflated probability estimates. The problem can be overcome by providing a causal explanation for the relevant statistics. Krynski and Tenenbaum (2007) found that when
such an explanation was supplied (see the causal version in Figure 1) there was a marked increase in the accuracy of probability estimates.

Alternately Hogarth and Soyer (2011) suggest that people are less likely to neglect relevant statistics when they have had "experience" with the relevant sample. Specifically, they suggest that trial-by-trial sampling of the frequency of an event from the relevant probability distribution can lead to more accurate estimates in problems involving low base rates (cf. Lejarraga, 2010; Sedlmeier, 1999). Hence, Hogarth and Soyer (2011) allowed some participants to draw sequential samples of women with a positive mammogram from a distribution with a low base rate of cancer. Sampling led to more accurate probability estimates than when only a description of the base rate was provided.

## Mammogram problem

Doctors often encourage women at age 50 to participate in a routine mammography screening for breast cancer.

From past statistics, the following is known:
$1 \%$ of women had breast cancer at the time of the screening Of those with breast cancer, $80 \%$ received a positive result on the mammogram
[Standard version] Of those without breast cancer, 15\% received a positive result on the mammogram
[Causal version] 30\% of the women had a benign cyst at the time of screening. Of those with a benign cyst, $50 \%$ received a positive test on the mammogram

All others received a negative result
Suppose a woman gets a positive result during a routine
mammogram screening. Without knowing any other symptoms, what are the chances she has breast cancer? __ \%
Figure 1. The mammogram problem

## Combining causal model and sampling approaches

A key motivation for the current work was that the causal model and sampling approaches appear to address different components of intuitive probability problems. Krynski and Tenenbaum (2007), focused on incorporating information about false positive rates into a causal model of the problem. In contrast, Hogarth and Soyer's (2011) sequential sampling approach aimed at improving sensitivity to the low base
rate. The major goal of the current research was to combine these two approaches to overcoming base rate neglect. If our analysis is correct, then the causal model and sampling approaches should have additive effects on performance in intuitive probability problems.

A secondary goal was to address a number of methodological limitations of previous work using sequential sampling to overcome base rate neglect. First, Hogarth and Soyer (2011) asked participants to answer the same probability problem on three occasions; after a summary description of the base rate, after sampling experience, and a final estimate. For mammogram problems like that in Figure 1 this led to a complex pattern of results with accuracy increasing when probability problems were solved after sampling, but a marked decrease in accuracy when participants subsequently solved the same problem after reading a description of the base rate. To allow for a more straightforward assessment of the effects of description and experience, Experiment 1a used a betweensubjects manipulation in which half the participants provided an answer to the intuitive probability problem after reading a description and having relevant sampling experience, whereas the remainder answered on the basis of the description alone.

Second, Hogarth and Soyer (2011) assessed intuitive probability accuracy using a relatively liberal performance measure (participants had to choose the correct estimate from four options). This is likely to yield higher levels of accuracy than the more conventional method of requesting point estimates of probability. To facilitate comparison of the sampling and causal model approaches we therefore assessed performance using both open-ended estimates (as used by Krynski \& Tenenbaum, 2007) and forced choice questions.

Third and most importantly, we aimed to clarify the nature of the information that gives rise to improved base rate representations. Hence, in Experiment 1b participants were provided with a yoked description of sampling outcomes (e.g., out of 4 people observed, 1 person had cancer) to examine whether improved performance was a result of sequential sampling per se or simply the distributional information provided by the sample (cf. Rakow, Demes, \& Newell, 2008).

## Experiment 1a

This study examined the respective contribution of causal explanation of false positives and sampling experience to performance on the mammogram problem (Figure 1). Each factor was varied factorially and performance was assessed using both point estimate and forced choice methods. Based on the previous work of Krynski and Tenenbaum (2007) and Hogarth and Soyer (2011), we expected that providing causal information and relevant sampling experience would each lead to improved probability judgments. Based on our argument that each of these approaches addresses a different component of the task, we further predicted that these effects would be additive.

## Method

Participants. One hundred undergraduate students $\left(M_{\text {AGE }}=\right.$ 20.1 years) participated for course credit. Equal numbers were randomly allocated to the four experimental conditions.

Design and Procedure. The experiment followed a 2 (False positive information: standard vs. causal) x 2 (Base rate presentation: description only vs. description + sampling) design with both factors manipulated between subjects.

All participants were presented with the mammogram problem shown in Figure 1 (cf. Krynski \& Tenenbaum, 2007, Experiment 2). The problem was presented in either the standard or causal version, with each version administered to an equal number of participants. In both versions the Bayesian solution to the question about the likelihood of cancer given a positive mammogram was (approximately) 5\%.

In all conditions the problem description (the text in normal font in Figure 1) was first presented on a computer screen. In the Description only condition, an open-ended question asking for an estimate of the likelihood of cancer in a woman with a positive mammogram appeared after 15 s . As per Krynski and Tenenbaum (2007), the format of this estimate was a \% chance of cancer between 0 and 100. Participants were invited to use an on-screen calculator to assist in solving the problem. After a likelihood estimate was entered, the cancer estimation question was repeated together with four alternative "answers that people commonly give to this question" ( $1 \%, 5 \%, 65 \%, 80 \%$ ). Participants used a mouse to click on the option they thought was "closest to the correct answer".

Those in the sampling condition received an additional sampling phase between the problem description and the request for a likelihood estimate. In this phase they were told that to assist task completion they would be able to draw samples of women who had received a positive mammogram. Each time a participant clicked a "simulate" button they were told whether or not a sampled woman had cancer. In the standard condition the feedback for cancerabsent cases was "this woman does not have cancer". In the causal condition it was "this woman has a benign cyst". Samples were drawn randomly from a uniform distribution of 10000 cases ${ }^{1}$. There was no limit on the number of samples that could be drawn. At any time during the sampling process participants could also click an on-screen button to view a running tally of a) samples with cancer; b) samples without cancer; and c) total samples viewed. To familiarize participants with the sampling tool, prior to commencing the main experiment they were shown the outcomes of 10 samples of tossing an unbiased coin. After the sampling phase those in the sampling condition received the same open-ended and multiple choice questions as the

[^84]description only condition, but were not provided with an on-screen calculator. There was no time limit on any part of the procedure.

An on-screen version of the 4-item Berlin Numeracy Test (Cokely, Galesic, Schulz, Ghazal \& Garcia-Retamero, 2012) was also administered. Numerical ability $\left(M_{\text {CORRECT }}=2.44\right)$ did not differ across experimental conditions ( $p>.35$ ).

## Results and Discussion

As a preliminary step we examined behavior in the sampling condition. The number of samples drawn ranged from 3 to $50\left(M_{\text {SAMPLES }}=17.26, \quad \mathrm{SD}=12.22\right)$. A majority of participants experienced no positive cases of cancer (42\%) or only one positive case (34\%). The mean number of samples did not differ between the causal or standard versions of the sampling condition ( $p$ 's $>0.5$ ).

Intuitive probability - Open-ended estimates. Estimates of the likelihood of cancer were analyzed by computing the simple deviation of the estimate from the normative solution ( $5.1 \%$, see Figure 2). To examine group differences in estimate accuracy, deviation scores were entered into a 2(description vs. sampling) x 2(standard vs. causal version) analysis of variance (ANOVA). Estimates in the sampling condition ( $M_{\text {DEVIATION }}=25.69$ ) were closer to the normative solution than those in the description only condition ( $M_{\text {DEVIATION }}=39.91$ ), $F(1,96)=4.78, p=.03$, Cohen's $d=$ 0.43. There was a non-significant trend for estimates in the causal condition $\left(M_{\text {DEVIATION }}=26.87\right)$ to be closer to the normative solution than those in the standard condition $\left(M_{\text {DEVIATION }}=38.74\right), F(1,96)=3.33, p=.07, d=0.36$. There was no interaction between base rate presentation and causal factors, $p=.45 .^{2}$

As per Krynski and Tenenbaum (2007), we also tallied the frequency of estimates that could be classified as correct (estimates in the range $4 \%-6 \%$ ) or as base rate neglect (estimates $>=65 \%$ ). Binary logistic regression showed that "neglect" estimates were less common in sampling than description only ( $24 \%$ vs. $44 \%$ of responses in the respective conditions), Wald (1) $=4.36, p=.04$, and less common in causal than the standard condition ( $24 \%$ vs. $44 \%$ ), Wald $(1)=4.36, p=.04$. However, the frequency of estimates classified as normatively correct ( $M=12 \%$ ) did not differ across conditions. No interactions between the sampling and causal factors were found ( $p$ 's $>.4$ ).

Intuitive probability - Forced choice. These responses were classified as correct (a choice of 5\%), base rate overuse (1\%), or base rate neglect (a choice of 65 or $80 \%$ ). The proportion of responses in each category within each condition is given in Figure 3. Logistic regression was again used to examine changes in the proportion of each

[^85]type of response across conditions. Figure 3 shows that selection of the correct response was more common in the sampling than the description only condition, Wald (1) = $7.24, p=.007$, and in the causal than the standard condition, Wald (1) $=8.96, p=.003$. The interaction between these factors was not significant ( $p>.35$ ). Choice of the base rate neglect options was less common in the causal than the standard condition, Wald (1) $=4.87, p=.03$. These choices were unaffected by the sampling manipulation and there was no interaction with causal version ( $p$ 's > .15). Neither manipulation affected selection of the base rate overuse option, ( $p$ 's > .15).


Figure 2. Deviation scores for probability estimates (with standard error bars).

Additional analyses. In the description only condition, accessing the on-screen calculator during testing was positively correlated with the likelihood of giving the correct estimate on the open-ended test, $r(49)=0.28, p=$ .04 , and with selection of the correct alternative in forced choice, $r(49)=0.35, p=.01$. In the sampling condition, no sampling statistics (number of samples drawn, number of cancer positive cases observed, proportion of cancer positive cases observed) were correlated with any performance measures (all $p$ 's > 0.1). However, the frequency with which the summary tally was accessed was positively correlated with the likelihood of providing a correct estimate, $r(49)=0.32, p=.02$.

Summary. The accuracy of judgments of cancer probability was facilitated by an opportunity to sample instances with a positive mammogram and by causal explanation of false positives. Although correct probability estimates were rare, both causal and sampling manipulations led to a downward shift in estimates in the direction of the normative solution. Both manipulations increased choice of the correct estimate and decreased choice of the neglect option. Notably these effects were additive, supporting the view that the causal and sampling manipulations affect different components of intuitive probability judgment.



Figure 3. Experiment 1a. Proportion of forced choices.

## Experiment 1b

The beneficial effects of sampling found in Experiment 1a and in past work (Hogarth \& Soyer, 2011) could arise from a range of mechanisms. Hogarth and Soyer (2011) suggest that "across time, a person observes sequences of outcomes that can be used to infer the characteristics of the data generating process" (p. 435). However it is unclear whether sampling experience per se is critical here. Sequential sampling may be just one of many methods of obtaining information about the distribution of positive and negative cases. Other methods such as description of a frequency distribution (cf. Gigerenzer \& Hoffrage, 1995) could convey the same information, and hence may also reduce base rate neglect. Some support for this view comes from the Experiment 1 finding that use of a summary tally was correlated with estimate accuracy.

Experiment 1 b examined this possibility by presenting all participants with a summary tally of positive and negative cases of cancer from a sample of women with a positive mammogram. This 'enhanced description' presents the same base rate information that was present in the sampling condition of Experiment 1a, but without trial-by-trial sampling. To allow for close matching of the statistical information presented to participants, the sampling tallies used in this study were yoked to the outcomes of sequential sampling in Experiment 1a (see Rakow et al., 2008, for a
related manipulation). If sampling experience is crucial for gaining a more accurate representation of the problem, then the Experiment 1a sampling condition should yield superior probability estimates to enhanced description. If the critical issue is the generation of a representative distribution of positive and negative cases, then enhanced description should do as well as sequential sampling. As in Experiment 1a, descriptions of the problem included a standard or causal explanation of the false positives.

## Method

Participants
Fifty undergraduate students $\left(M_{\mathrm{AGE}}=19.3\right.$ years $)$ participated for course credit. Equal numbers were randomly allocated to causal and standard conditions.

## Procedure

The general procedure was similar to the causal and standard description only conditions in Experiment 1a, with the important exception that all participants were given an on-screen tally of positive and negative cases of cancer from samples of women with a positive mammogram. Fifty tallies were generated based on sampling outcomes in the sampling condition of Experiment 1a. For example, if a participant in the earlier study drew 20 samples containing 1 cancer positive and 19 negative instances, then a tally containing the same information was constructed for an enhanced description participant. An on-screen calculator was available to assist in answering the problem.

## Results and Discussion

Intuitive probability - Open-ended estimates. Estimation performance was again examined by calculating the deviation of estimates from the normative solution (see Figure 2). Accuracy as measured by deviation scores was not affected by causal explanation, $F(1,49)=0.2, p=.66$.

The more important issue was how estimation performance compared with the sampling and description conditions in Experiment 1a. Inspection of Figure 2 suggests that the pattern of deviation scores in enhanced description was more similar to the sampling than the description condition from the earlier study. These trends were examined using a cross-experimental task (description only, sequential sampling, enhanced description) x causal framing ANOVA. Planned comparisons compared performance in the enhanced description condition with the description only and sampling conditions respectively. The analysis confirmed that estimates in the enhanced description condition $\left(M_{\text {DEVIATION }}=16.56\right)$ were more accurate than in the description only condition, $F(1,144)=$ 14.04, $p<.001, d=0.74$, but did not differ from estimates in the sampling condition, $F(1,144)=2.24, p=.14$. No significant influence of causal framing was found. ${ }^{3}$

[^86]Intuitive probability - Forced choice. Forced choice responses in the enhanced description group are given in Figure 4. Binary logistic regression found no significant differences between the enhanced description and sampling groups for any type of response, and no interactions with causal framing, $p$ 's > .06. Correct responses were more common in enhanced description than in the description only conditions, Wald (1) $=9.66, p=.002$, and neglect responses were less common, Wald (1) $=7.93, p=.005$. Across the enhanced and description only conditions, causal framing led to a higher rate of correct responding than standard framing, Wald (1) $=4.06, p=.04$, but this effect was stronger in the description only condition, Wald (1) = 4.87, $p=.03$. The enhanced description and description only groups did not differ in base rate overuse.


Figure 4. Experiment 1b. Proportion of forced choices.
Summary. This study examined whether sampling experience is necessary to reduce base rate neglect in intuitive probability. When people were given a description of the relevant sampling information they performed as well as those in the sampling condition of Experiment 1a, and better than those in description only. It appears that what is crucial is having relevant information about the distribution of positive and negative cases; this can be obtained through sampling or a description of the frequency distribution.

A puzzling finding was that causal framing, which had a positive effect on probability estimates in Experiment 1a, had little impact on enhanced description estimates. This may have been due to the accuracy of intuitive probability estimates in the standard version of enhanced description being higher than the standard conditions in the earlier study (see Figure 1). In other words, estimates may have been approaching ceiling in the enhanced description standard group, reducing the likelihood of finding further facilitation due to causal explanation.

## General Discussion

These studies examined how providing sampling information about base rates and a causal explanation of false positives can improve intuitive probability judgments. Experiment 1a found that these manipulations led to a shift in probability judgments toward the normative response,
and away from inflated estimates that would usually be classified as base rate neglect. Moreover, each manipulation increased choice of the normative solution.

The results replicate and extend previous findings of a positive effect of causal framing (Krynski \& Tenenbaum, 2007) and sampling experience (Hogarth \& Soyer, 2011) on intuitive probability judgment. Experiment 1a, however, was the first study to combine these manipulations. An important result was that effects of sampling and causal explanation were additive. This is consistent with the view that these manipulations address different aspects of probability judgment. The sampling and enhanced description manipulations helped establish greater sensitivity to the base rate. The causal manipulation facilitated the incorporation of false positives into the problem solution.

Experiment 1b clarified the role of sampling experience in improving probability judgment. Contrary to the views of Hogarth and Soyer (2011), we found that sequential sampling was not essential for reducing base rate neglect. A similar level of facilitation was obtained when the relevant statistical information was conveyed by a description of sampling outcomes. This is consistent with other findings in the judgment and decision-making literature which show that detailed descriptions of statistical information can produce equivalent effects to sequential sampling (e.g., Rakow et al., 2008).
The causal facilitation effects in these studies are consistent with the broader perspective on probability judgments outlined by Krynski and Tenenbaum (2007). This "causal Bayesian" view suggests that encoding the relevant statistics in an intuitive probability problem will not lead to accurate judgments, unless the statistics can be incorporated into a causal model of the problem. In the current studies both standard and causal groups were given equivalent statistical information about false positives but only the latter were supplied with a cause. According to Krynski and Tenenbaum (2007) this allows those in the causal condition to construct an intuitive model with two generative nodes that provide alternative explanations for positive mammograms. More broadly, these findings are consistent with the idea that people often fail to spontaneously consider alternative causes for probabilistic outcomes but can do so when prompted (e.g., Fernbach, Darlow, \& Sloman, 2011).

It is notable that although both causal explanation and sampling shifted open-ended probability estimates in the right direction, neither manipulation increased the rate of normatively correct estimation. Similar results have been reported in previous work on base rate neglect. Krynski and Tenenbaum (2007) found that although causal explanation of false positives reduced base rate neglect, most participants in the causal condition still failed to produce a normative probability estimate. Likewise, although Gigerenzer and Hoffrage (1995) found that frequency formats for relevant statistics improved the accuracy of probability estimates, the majority of participants still gave
normatively incorrect answers to the mammogram problem. This raises the question of what additional barriers need to be overcome to produce normative probability estimates.

The causal Bayesian perspective suggests one answer. According to this view the solution of probability problems proceeds in three stages. The first involves constructing an intuitive causal model of the problem. The second involves encoding the relevant statistics and mapping these onto the various nodes of the causal model. The third stage uses Bayesian inference to update beliefs in the light of the observed statistics. Arguably, the causal and sampling manipulations in the current studies impacted on the first two stages. The finding that a majority still do not produce normative estimates suggests that people may need further assistance with the final stage of implementing Bayes' rule (cf. Sedlmeier \& Gigerenzer, 2001).

A final caveat is that although an opportunity to draw samples and description of sampling outcomes facilitated performance, sampling should not be regarded as a panacea for the problem of representing base rates. It is important to note that in the current studies and in Hogarth and Soyer (2011), samples were conditionalized on a woman having a positive mammogram. This ensured that with sufficient draws, a representative base rate was observed. However, samples outside the laboratory are not always constrained in this way. Sampling based on incorrect conditionalization (e.g., drawing samples of women with cancer and seeing whether they have a positive mammogram) can actually lead to more biased intuitive probability estimates (e.g., Fiedler, Brinkmann, Betsch, \& Wild, 2000).

These studies suggest that the causal Bayesian approach represents a useful framework for analyzing the subcomponents of intuitive probability problems, and intervening on these components to improve judgment accuracy. Our findings show that using experienced or described samples can reduce base rate neglect, and that supplying a cause for false positives increases the likelihood that these will be considered in probability judgments.

## Acknowledgments

This work was supported by Australian Research Council Discovery Grant DP120100266 to the first two authors and a Future Fellowship FT110100151 to the second author. We thank Chris Moore and Ann Martin for help with programming and running the experiments.

## References

Barbey, A. K., \& Sloman, S. A. (2007). Base rate respect: From ecological rationality to dual processes. Behavioral and Brain Sciences, 30, 241-297.
Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., \& GarciaRetamero, R. (2012). Measuring risk literacy: The Berlin numeracy test. Judgment and Decision Making, 7, 25-47.
Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine. In D. Kahneman, P. Slovic, \& A. Tversky (Eds.), Judgment under uncertainty: Heuristics and biases (pp. 249-267). Cambridge, UK: CUP.

Evans, J. S. B. T., Handley, S. J., Perham, N., Over, D. E., \& Thompson, V. A. (2000). Frequency versus probability formats in statistical word problems. Cognition, 77, 197213.

Fernbach, P. M., Darlow, A., \& Sloman, S. A. (2011). Asymmetries in predictive and diagnostic reasoning. Journal of Experimental Psychology: General, 140, 168185.

Fiedler, K., Brinkmann, B., Betsch, T., \& Wild, B. (2000). A sampling approach to biases in conditional probability judgments: Beyond base rate neglect and statistical format. Journal of Experimental Psychology: General, 129, 399-418.
Gigerenzer, G., \& Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. Psychological Review, 102, 684-704.
Hogarth, R. M., \& Soyer, E. (2011). Sequentially simulated outcomes: Kind experience versus nontransparent description. Journal of Experimental Psychology: General, 140, 434-463.
Krynski, T. R., \& Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. Journal of Experimental Psychology: General, 136, 430-450.
Lejarraga, T. (2010). When experience is better than description: Time delays and complexity. Journal of Behavioral Decision Making, 23,100-116.
Rakow, T. R., Demes, K., \& Newell, B. R. (2008). Biased samples not mode of presentation: Re-examining the apparent underweighting of rare events in experiencebased choice. Organizational Behavior and Human Decision Processes, 106, 168-179.
Sedlmeier, P. (1999). Improving statistical reasoning: Theoretical models and practical applications. Mahwah, NJ: Erlbaum.
Sedlmeier, P., \& Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. Journal of Experimental Psychology: General, 130, 380-400.
Tversky, A., \& Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185, 11241131.

## Appendix

The normative probability of cancer (C) given a positive mammogram ( M ) is given by:

$$
\begin{aligned}
p(\mathrm{C} \mid \mathrm{M}) & =\frac{p(\mathrm{C})^{*} p(\mathrm{M} \mid \mathrm{C})}{p(\mathrm{C})^{*} p(\mathrm{M} \mid \mathrm{C})+p(\neg \mathrm{C})^{*} p(\mathrm{M} \mid \neg \mathrm{C})} \\
& =\frac{0.01 * 0.80}{0.01 * 0.80+0.99 * 0.15} \\
& \approx \quad 0.051
\end{aligned}
$$

# The Role of Familiarity, Priming and Perception in Similarity Judgments 

Laura M. Hiatt and J. Gregory Trafton<br>U.S. Naval Research Laboratory<br>Washington, DC 20375 USA<br>\{laura.hiatt, greg.trafton\} @ nrl.navy.mil


#### Abstract

We present a novel way of accounting for similarity judgments. Our approach posits that similarity ratings stem from three main sources: familiarity, priming, and inherent perceptual similarity. We present a process model of our approach in the cognitive architecture ACT-R, and match our model's predictions to data collected from a human subject experiment which involved simple perceptual stimuli. Familiarity accounts for rising ratings over time; priming accounts for asymmetric effects that arise when the stimuli are shown with different frequencies. Pure perceptual similarity also predicts trends in the results. Overall, our model matched the data with $R^{2}$ of 0.99 .


## Introduction

Similarity is a critical and pervasive part of human cognition (Medin, Goldstone, \& Gentner, 1993). Similarity measures are integral to object categorization and classification (Nosofsky, 1992). Similarity is also pervasive in problem solving (Novick, 1990), decision-making (Medin, Goldstone, \& Markman, 1995), and memory (Roediger, 1990). As with many aspects of human cognition, however, the mechanisms that determine similarity are not yet fully understood. Various theories abound, with none yet able to capture enough different types of situations to be called the winner (Rorissa, 2005).

One interesting result in this field is asymmetries that have been shown to arise when making similarity judgments, even of very simple perceptual stimuli (Tversky, 1977; Rosch, 1975). Rosch (1975) argued that such similarity is based on mapping stimuli onto one another and, intuitively, nonprototypical stimuli map more easily onto prototypical stimuli than vice versa. Tversky (1977) argued that is due to weighted feature matching, where the salience of features in the current context determines their weight; others agree with this thought in general (Medin et al., 1993; Glucksberg \& Keysar, 1990).

These two explanations, however, assume that either there is a clear prototype inherently present in the experiment (such as the more perceptually complex stimulus), or that stimuli have various features which have a clear inherent order of cognitive preference and saliency (such as symmetry). They do not, however, provide any concrete explanations for why complexity or symmetry may lead to prototypicality or saliency.

Polk, Behensky, Gonzalez, and Smith (2002) shed light on the situation by presenting an experiment that avoids the question entirely by using perceptual stimuli where the only feature was color (so there are no features to comparatively weigh), and where the color hues are fairly similar (so there is
no clear prototype). The experiment showed a striking asymmetry in similarity judgments between the different colors when they were presented with different frequencies during an irrelevant training task: colors which had been trained on less often were considered more similar to colors which had been trained on more often than the other way around. To account for these low-level results, Polk et al. (2002) implemented a neural network which simulated the asymmetry by measuring the ease with which the network switches between different activation patterns; those that are more stable (e.g., high-frequency patterns) were easier to assimilate to.

In our approach, we match the human subject data from Polk et al. (2002) while attempting to address three additional points. First, there was a second significant effect, that the ratings in general increased over time, that the above models do not address. Second, we wanted our approach to provide explicit cognitive processes for similarity ratings. Third, we believe that inherent perceptual similarity also plays a part in these types of similarity judgments (e.g., purple is inherently more similar to blue than to orange) (Smith \& Heise, 1992).

To this end, we begin our approach with the cognitive architecture ACT-R (Adaptive Characterization of Thought Rational) (Anderson, 2007). Using ACT-R, we account for similarity judgments by considering three values provided $a$ priori by the architecture. The first, familiarity, is represented as a base-level activation value of a concept, which represents its frequency and recency of use. The second, priming, is based on spreading activation, which disperses activation between different, associated concepts in declarative memory (Anderson, 1983; Harrison \& Trafton, 2010). In addition, we utilize an extension to ACT-R which provides it with a calculation for measuring color similarity (Breslow, Ratwani, \& Trafton, 2009; Breslow, Trafton, \& Ratwani, 2009).

Our model starts without any pre-existing declarative knowledge or network structures; all knowledge is created during the experiment. Over time, our cognitive model builds a network of concepts (e.g., color blocks) by learning associations between them in the form of subsymbolic connections between their representations in declarative knowledge. During each similarity judgment, the model combines its measure of perceptual similarity with base-level and spreading activation to determine its response. On the first trial of an experiment, there is no spreading activation since there is no declarative knowledge and so the judgment is based purely on base-level activation and perceptual similarity; however, over time the model builds up declarative memories that may contribute to spreading activation in later trials. This explains the two main effects found in Polk et al.'s experiment. Dur-
ing judgments made on stimuli which have been previously viewed with different frequencies, less activation is spread from high- to low-frequency color patches than from lowto high-frequency color patches because that is the direction which priming favors. In addition, base-level activation is higher at the end of the experiment than in the beginning due to increased familiarity with the colors, leading to increased ratings over time.

The primary contribution of this work is a general account for similarity which provides implemented, explicit, processlevel mechanisms for calculating similarity values. Other work has used activation for similarity in more abstract terms: the neural network written by Polk et al. (2002) relies on activation patterns; and other accounts also exist (Kozima \& Furugori, 1993; Ulhaque \& Bahn, 1992). Tversky (1977)'s discussion of salience, and Rosch (1975)'s on prototypicality can also be seen as broadly touching upon activation in similarity. Our work solidifies these accounts in a cognitive setting by positing that the abstract notion of salience (or prototypicality) translates to familiarity and priming in a cognitive model. In addition, our work is distinguished because we also account for inherent perceptual similarity. We show that our model is an excellent fit for empirical human subject data on similarity judgments on simple perceptual stimuli.

## Experiment

There were three phases to the experiment: a pre-test phase, a training phase, and a post-test phase (Polk et al., 2002). In the pre-test phase, participants viewed two patches of different colors that were the same size and were asked to rate their similarity on a scale of 0 to 9 ( 0 as highly dissimilar, 9 as highly similar). Five different hues of green and five different hues of blue were used, designated as bluel...blue5, and green $1 .$. green 5 . Greens and blues were never compared to each other; only hues of the same color were shown concurrently. During a trial, the stimuli were presented as part of a text question to emphasize the directionality of the judgment: "How similar is (color patch 1) to (color patch 2)?" The color blocks were labeled "Blue1" and "Blue2," or "Green1" and "Green 2 " as appropriate, with the label displaying below the color patches; note that this is irrespective of whether the color itself was blue1, blue5, etc. The blocks were 140X140 pixels. Once a user entered a rating, the screen was cleared for 500 ms before the next comparison appeared. Each pair of colors was presented twice in each direction for a total of four times each. The order in which the pairs were presented was randomized, except that the same hue was not present in consecutive trials. The sentence was centered both horizontally and vertically.

In the training phase, participants saw two patches of the same color but different sizes (125X125, 131X131, 138X138 and 144X144 pixels, appearing with equal probability) and were asked to specify which was larger. The key part of the experiment is that, during this phase, two of the five greens and two of the five blues were presented ten times more fre-
quently as others, 110 times instead of 11 (half the participants saw blue1, blue2, green1 and green 2 with a higher frequency, called the "1-2 group", the other saw blue4, blue5, green4 and green 5 presented more often, the " $4-5$ group").

The third phase was a second testing phase that was an exact repeat of the first phase. Forty-five participants took part in the experiment, with ten being excluded due to inaccuracy on their size judgments or self-reporting of a lack of concentration. For more details, see (Polk et al., 2002).

The experiment produced two interesting results. First, there was no significant difference between "forward" (less frequent color on the left, more frequent color on the right) and "backward" (more frequent color on the left, less frequent on the right) comparisons in the pre-test phase; however, these comparisons showed a striking asymmetry effect during the post-test phase. Specifically, forward comparisons were ranked as significantly more similar than backward comparisons during this testing phase. A second find was that similarity ratings were significantly higher in the post-test than in the pre-test. No other effects were reported as significant by the authors.

## ACT-R

At the core of our approach is the cognitive architecture ACTR (Anderson, 2007). At a high level, ACT-R is a hybrid symbolic/subsymbolic production-based system. In other words, given declarative knowledge (fact-based memories, or "chunks") and procedural knowledge (rule-based memories, or "productions"), as well as input from the world (e.g., visual), it decides what productions to fire next; these productions can either change its internal state (e.g., by creating new knowledge) or its physical one (e.g., by pressing a key on a keyboard). Knowledge has both a symbolic component, such as who was where at what time, and subsymbolic one, such as how relevant a fact is to the current situation.

ACT-R is made of up several major components. First, it has several limited-capacity buffers. Each buffer is backed by one (or more) theoretically motivated modules (e.g., declarative, visual, aural, etc.); in addition, there is the procedural module, which does not have an associated buffer. Each module represents a specific cognitive faculty and has been shown to have anatomical correspondences in the brain (Anderson, Albert, \& Fincham, 2005; Anderson, 2007).

Chunks consist of a set of slots, whose values determine the concept that the chunk represents. At any point in time, there may be at most one chunk in any individual buffer; a module's job is to decide when to put chunks into its corresponding buffer. Then, a central pattern matcher uses the contents of the buffers, if any, to match specific productions which, when fired, can modify the current buffer contents.

The relevant modules of ACT-R to this paper are the declarative, intentional, imaginal, visual and motor modules, which are associated with the retrieval, goal, imaginal, visual and visual-location, and motor buffers, respectively. The declarative module manages the creation and storage of the model's
factual memory; in addition, when requested, chunks can be accessed via the retrieval buffer. It has been shown to be an astonishingly good predictor of human declarative memory (Anderson, Bothell, Lebiere, \& Matessa, 1998; Anderson, 1983; Schneider \& Anderson, 2011). The intentional and imaginal modules provide support for task-oriented cognition. The goal buffer (associated with the intentional module) typically contains chunks that identify and placekeep the model's current goal; the imaginal module usually contains intermediate problem state representations. Finally, the visual and motor modules interface ACT-R with the world, allowing ACT-R to see objects on a computer screen (including their height and color) and press buttons on a keyboard.

To quantify the perceptual difference between the color blocks, we consider a measure of color similarity proposed by Breslow, Ratwani, and Trafton (2009). They introduced a component to ACT-R which supports high-level color processing that can detect both color similarity and brightness difference between colors. It is based on the CIEDE2000 algorithm (CIE, 2001), and has been shown to match well with human subject data. Given two color values, this color similarity component returns a numeric measure of how similar they are perceived to be by the cognitive model.

## Subsymbolic Information

A key aspect of declarative memory in ACT-R is priming, or the subsymbolic activation of chunks. Activation consists of three primary components: base-level activation, spreading activation, and noise. Base-level activation is a measure of familiarity that is learned over time and is a function of the frequency and recency of references to the chunk, where a single reference is defined as (for purposes of this paper) being added to and then removed from a buffer. It is designed to represent the activation of a chunk over longer periods of time. Spreading activation, on the other hand, is temporary and based on the current context, allowing chunks that are the focus of attention to prime related memories for short periods of time. Noise is a random component that models the noise of the human brain; since its presence would not affect our results, we ignore noise in the rest of this paper.

Spreading activation is spread along subsymbolic associative links between chunks. Links are created from a chunk $j$ to a chunk $i$ when: (1) chunk $i$ contains chunk $j$, or has chunk $j$ as one of its slot values; or (2) chunks $i$ and $j$ are both matched by the same production (called co-occurrence) (Anderson, 1983; Harrison \& Trafton, 2010). There are other ways of creating links, as well, but we do not utilize them in this model and so forgo their discussion. Once established, links have an associated strength value which affects how much activation is passed along the link from chunk $j$ to chunk $i$. Link strengths, intuitively, reflect the probability that chunk $i$ will be needed when chunk $j$ is being referred to by a production. They are a function of how many times chunks $j$ and $i$ have been referred to by a production at the same time, vs. how many times chunk $j$ was referred to by a production without chunk $i$. Note, then, that while links
stemming from co-occurrence are always present in both directions (i.e., chunk $j$ activates chunk $i$ and vice versa), the links may be of different strengths if the chunks have not always been referred to by productions at the same time, or with the same frequency.

Spreading activation sources from the goal buffer. When a chunk $i$ is in the goal buffer, the buffer's source activation is divided equally among all chunks $j$ which have an outgoing link to chunk $i$ (such as a slot value of chunk $i$, or a chunk that has co-occurred with chunk $i$ in the past). The $j$ chunks then use their source activation as the basis of spreading activation along all of their outgoing links. Note that in ACT-R, this is a one-step process; the chunks that receive spreading activation from the $j$ chunks do not, in turn, spread activation along their outgoing links (Anderson, 1983).

## Model and Fit

The model itself is fairly simple. It starts out with no declarative knowledge, but with the productions necessary to complete the experimental task. For each trial during the two testing phases, the model starts by looking at the color block on the left. Once the model has the block chunk in its visual buffer, the model requests a retrieval of the chunk of the color associated with it, and looks for the object on the right. Once the left color chunk has been retrieved, the model places it in the imaginal buffer and removes it from the retrieval buffer. Then, when it sees the object on the right, it can retrieve the color of the second color block. When that chunk has been retrieved, the model has each of the color chunks in a buffer and it can then proceed to making the similarity judgment.

The model draws similarity from three sources: how perceptually similar the two colors are to each other, how familiar the right color is, and to what degree the right color is primed. Recall that mathematically, perceptual similarity is calculated from the RGB values of two colors; familiarity is represented as base-level activation, which numerically represents the recency and frequency of use of a color; and, finally, priming is measured via the amount of spreading activation that a color receives in the current context. Therefore, the model calculates the perceptual similarity of the two colors, and looks at the total activation (both base-level and spreading) that the right color has at the time of the judgement. Note that while the right color receives spreading activation from sources other than the left color, the amount of "other" spreading activation is constant across trials, making the spreading activation from the left color the cause of asymmetry effects. After the model has these two numbers available, it presses the button corresponding to its rating, the trial is finished and the model waits for the next one to begin.

During a training trial, the model first looks at the color block on the left. It then stores the block's color as part of the goal representation in case it is later needed (such as if the subsequent visual search fails), and looks for another block of the same color. Once it sees the second color block, it retrieves the first one. Then, while thinking about the first
block and looking at the second one, it compares their heights and responds accordingly.

Subsymbolically, during each testing trial, co-occurrence links are created (or strengthened) between: each of the block chunks and the current goal chunk; each of the color chunks and the current goal chunk; and the two color chunks. Block chunks also have an incoming, containment link from their associated color chunk. During a training trial, co-occurrence links are created (or strengthened) between: the two block chunks; and each of the block chunks and the current goal chunk. The goal chunk and the block chunks also each have an incoming link from the color chunk because they contain it as a slot value. Figure 1 shows this in diagrammatic form, showing co-occurrence links as bi-directional for simplicity.


Figure 1: Subsymbolic connections between chunks at various phases of model execution. Here, the model performed two pre-test trials (with colors blue5/blue1, and bluel/blue2, respectively), and one training trial (where bluel is the color). In order to maintain clarity, this diagram is slightly simplified from the model's actual subsymbolic network (e.g., it does not contain containment links for the visual location slots of the left and right blocks, which have no bearing in the spreading activation process here).

In terms of parameters, the associative learning rate, which affects the rate at which links are strengthened, was set to 6.5 , which represents a fairly brisk rate of learning. There is no standard value for this parameter. The base level learning decay parameter was set to 0.4 instead of its default of 0.5 . All other parameters were set to their default values.

## Model Predictions

First, the model predicts that later comparisons will, overall, be more similar than earlier comparisons. Before the exper-
iment begins, the colors are not familiar to the model, and so do not have very high base-level activations. During the pre-test, those values increase as the color chunks are referenced many times. Throughout the training phase the chunks' base-level activations decay, since the color chunks are not referenced in those productions. Base-level activation values then increase again during the post-test, leading to higher familiarity with the colors in the post-test than in the pre-test. Since all colors are shown equally during the pre-test and post-test phases, base-level activation does not predict any sort of asymmetry effect. Additionally, within the pre-test and post-test conditions, our model predicts that there will be ordering effects, with later stimuli being rated as more similar than earlier stimuli; these effects, however, should average out given randomization of stimuli across participants.

As we have mentioned before, the different strengths of the subsymbolic links between two colors can cause asymmetries to arise in the degree to which they prime each other. Consider, on an intuitive level, Figure 1(b). Now, imagine that a model with this subsymbolic structure has the goal to judge how similar blue5 and bluel are. First, both color chunks receive equal source activation from the goal chunk due to their equivalent co-occurrence links with it. They then provide each other with spreading activation according to the appropriate link strength. Here, as its greater number of links implies, bluel has been needed more times than blue5; this means that the link blue $1 \rightarrow$ blue 5 is weaker than the link blue $5 \rightarrow$ bluel. Therefore, bluel will receive more spreading activation than blue5, leading to an asymmetry in their similarity rankings.

As a result of this asymmetry, priming in the model predicts different effects for the pre- and post- tests. For the pretest, the model predicts differences in similarity of forward and backward comparisons based solely on ordering effects of the stimuli. Given enough participants, these ordering effects average out over time to result in equal pre-test forward and backward comparisons. For the post-test, the model predicts that less frequently shown colors will spread more activation to more frequently shown colors than vice versa; i.e., it predicts that colors in forward comparisons will be ranked as more similar than those in backward comparisons.

Finally, our model predicts that the green ratings will be higher overall than the blue ratings, as well as that the $4-5$ group's ratings will be higher overall than the 1-2 group's ratings. This is because of the specific hues chosen and is a purely perceptual point. The color similarity values do not differ depending on the direction of the comparison, or on whether the test is a pre- or post-test.

## Model Fit

In addition to the experimental results published in the original article (Polk et al., 2002), we also examined more detailed aggregate data provided to us by the authors. The data included the averages, for each subject, of ratings for trials of each condition (e.g., the average rating for each subject of all pre-test forward trials of blue hue, etc.). Since our model
is sensitive to the order in which stimuli are presented, we would have preferred to replicate the experiment exactly, including presenting the stimuli in the same order as in the original experiment. Because this information was not available, we instead used our model to simulate data from 1000 participants, in order to allow effects to better converge on the model's true predictions.

Our measurement of the model fit is done in two steps. First, the model needs to transform the similarity measures into an overall similarity rating. We do this post-hoc by fitting a linear regression model to the data, with the perceptual similarity value and total activation as the explanatory variables and the human participants' ratings as the dependent variable. We use total activation to maintain cognitive fidelity; it is unclear whether human minds can separately consider baselevel and spreading activation values during cognitive tasks. Individual data points were the different conditions (e.g., the average rating across all participants in the 1-2 group of pretest forward trials of blue hue, etc.). The model only considers the main effects of the two variables; this is because our goal is to show that the two similarity measures are the primary components of similarity ratings in this task, not to make any claims about how they are combined by the brain into a numerical rating. We take this approach because there are very few theories or accepted practices of how to convert continuous, numerical data to a rating scale.

Second, with this step completed, we compared the model's predicted ratings with the human participants' ratings. The model does indeed produce the two main significant effects of the human subject data, showing both a directional asymmetry in post-test comparisons as well as an increase in similarity ratings overall in the post-test. Figure 2 shows graphs of the numerical results for both the model and the human subject data. Note that error bars for the human data are not available. $R^{2}$ (multiple) for these graphs across all data was 0.99 ; for blue only, 0.96 ; for green only, 0.96 ; for the 1-2 group (which saw blue1, blue2, green1 and green 2 more frequently), 0.98, and for the $4-5$ group (which saw blue 4 , blue5, green 4 and green 5 more frequently), 0.91.

For differences in color, the model's results, where blue pre-test ratings are slightly lower than green pre-test ratings, do not match well with the data's trends. The model does have overall higher ratings for the 4-5 group pre-test than the 1-2 group pre-test, but not to the extent of the data. The effects do not present a difficulty for the model, however, in large part because the experiment did not find these effects to be significant, presumably due to its small sample size of 35 .

In terms of their individual contribution to explaining the data, color alone yields an $R^{2}$ of 0.09 ; color and base-level activation produce an $R^{2}$ of 0.79 ; and color, base-level activation and spreading activation an $R^{2}$ of 0.99 . This is intuitive and consistent with our account of the data: base-level activation is responsible for the larger main effect between the pre- and post-tests, while spreading activation correlates to the more modest interaction effect of forward versus back-


Figure 2: Main Results.
ward comparisons.

## Discussion

In this paper, we have introduced an account for similarity ratings that combines familiarity, priming, and perceptual similarity into a single judgment. We match our account to human subject data involving simple perceptual stimuli from Polk et al. (2002). Our approach explains that ratings rise over time because participants become more familiar with the stimuli, in general. Priming explains the asymmetry effect found because, inherently, low frequency concepts prime high frequency concepts more than the opposite. Finally, although the experiment did not reveal any significant differences between colors, our approach predicts that different pairs of colors will be slightly more similar than others due to pure perceptual similarity. Using these mechanisms, we show an excellent match to the human data.

Our approach is significant for at least two main reasons. First, we provide explicit, process-level mechanisms for determining similarity that explain, in a sense, how others' work (Rosch, 1975; Tversky, 1977) is realized by the human mind. Second, and perhaps more importantly, the mechanisms we offer as the basis of similarity have been shown to be preexisting mechanisms in cognition that are also used for other
cognitive processes such as retrieval of memories, categorization, and problem solving (Anderson, 2007; Altmann \& Trafton, 2002). This strengthens our approach since it also explains the pervasiveness of similarity in human cognition that has been found by a plethora of other research.

Finally, it is worth noting that familiarity, priming and perceptual similarity are not intended to be characterized as the ultimate, and only, way to determine similarity ratings. While they work well with simple perceptual stimuli, and we expect that their success will also extend to more complicated perceptual stimuli and simple concepts, we recognize that more complicated mechanisms are likely at work in, for example, the similarity of complex perceptual scenes, or the similarity of two short stories. We believe that such judgments likely involve some sort of structure alignment process as others have hypothesized for similarity judgments of higher-level stimuli or concepts (Markman, 1999; Goldstone, 1994). Instead, this paper is intended to introduce familiarity, priming and perceptual similarity as the foundation for similarity which other mechanisms can augment.

## Acknowledgments

This work was supported by the Office of Naval Research, grant numbers N0001412WX30002 and N0001411WX20516 to GT. The views and conclusions contained in this document do not represent the official policies of the U.S. Navy.

Many thanks to Thad Polk for making his data available to us. We would also like to acknowledge Wallace E. Lawson and Anthony M. Harrison for their helpful discussions during this work.

## References

Altmann, E. M., \& Trafton, J. G. (2002). Memory for goals: An activation-based model. Cognitive Science, 26, 39-83.
Anderson, J. R. (1983). A spreading activation theory of memory. Journal of Verbal Learning and Verbal Behavior, 22(3), 261-295.
Anderson, J. R. (2007). How can the human mind occur in the physical universe? Oxford University Press.
Anderson, J. R., Albert, M. V., \& Fincham, J. M. (2005, August). Tracing problem solving in real time. Journal of Cognitive Neuroscience, 17, 1261-1274.
Anderson, J. R., Bothell, D., Lebiere, C., \& Matessa, M. (1998). An integrated theory of list memory. Journal of Memory and Language, 38(4), 341-380.
Breslow, L. A., Ratwani, R. M., \& Trafton, J. G. (2009). Cognitive models of the influence of color scale on data visualization tasks. Human Factors, 51(3), 321-338.
Breslow, L. A., Trafton, J. G., \& Ratwani, R. M. (2009). A perceptual process approach to selecting color scales for complex visualizations. Journal of Experimental Psychology: Applied, 15(1), 25-34.
CIE. (2001). Inprovements to industrial colour-difference evaluation (Tech. Rep. No. CIE 142-2001). Vienna: Commission Internationale de l'Eclairage.

Glucksberg, S., \& Keysar, B. (1990). Understanding metaphorical comparisons: Beyond similarity. Psychological Review, 97, 3-18.
Goldstone, R. L. (1994). Similarity, interactive activation, and mapping. Journal of Experimental Psychology: Learning, Memory and Cognition, 20(3-28).
Harrison, A. M., \& Trafton, J. G. (2010). Cognition for action: an architectural account for "grounded interaction". In Proceedings of the Annual Conference on Cognitive Science.
Kozima, H., \& Furugori, T. (1993). Similarity between words computed by spreading activation on an English dictionary. In Proceedings of the Sixth Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg, PA, USA.
Markman, A. B. (1999). Knowledge representation. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
Medin, D. L., Goldstone, R. L., \& Gentner, D. (1993). Respects for similarity. Psychological Review, 100(2), 254278.

Medin, D. L., Goldstone, R. L., \& Markman, A. B. (1995). Comparison and choice: Relations between similarity processes and decision processes. Psychonomic Bulletin and Review, 2, 1-19.
Nosofsky, R. M. (1992). Exemplar-based approach to relating categorization, idenfitication, and recognition. In F. G. Ashby (Ed.), Multidimensional models of perception and cognition (p. 363-393). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.
Novick, L. R. (1990). Representational transfer in problem solving. Psychological Science, 128-132.
Polk, T. A., Behensky, C., Gonzalez, R., \& Smith, E. E. (2002). Rating the similarity of simple perceptual stimuli: Asymmetries induced by manipulating exposure frequency. Cognition, 82(3), B75-B88.
Roediger, H. L. (1990). Implicit memory: Retention without remembering. American Psychologist, 45, 1043-1056.
Rorissa, A. (2005). Perceived features and similarity of images: An investigation into their relationships and a test of tversky's contrast model. Unpublished doctoral dissertation, University of North Texas.
Rosch, E. (1975). Cognitive reference points. Cognitive Psychology, 7, 532-547.
Schneider, D. W., \& Anderson, J. R. (2011). A memorybased model of hick's law. Cognitive Psychology, 62(3), 193-222.
Smith, L. B., \& Heise, D. (1992). Perceptual similarity and conceptual structure. Advances in Psychology - Percepts, Concepts and Categories: The Representation and Processing of Information, 93, 233-272.
Tversky, A. (1977). Features of similarity. Psychological Review, 84, 327-352.
Ulhaque, E., \& Bahn, K. D. (1992). A spreading activation model of consumers' asymmetric similarity judgment. Advances in Consumer Research, 19, 782-786.

# Large-Scale Empirical Analyses of the Abstract/Concrete Distinction 

Felix Hill (fh295@cam.ac.uk) ${ }^{1}$, Anna Korhonen (alk23@cam.ac.uk) ${ }^{1}$, Christian Bentz (cb696@cam.ac.uk) ${ }^{2}$<br>${ }^{1}$ Computer Laboratory, University of Cambridge<br>${ }^{2}$ Department of Theoretical and Applied Linguistics, University of Cambridge


#### Abstract

We present original evidence that abstract and concrete concepts are organized and represented differently, based on statistical analyses of thousands of concepts in publicly available datasets. First, we show that abstract and concrete concepts have differing patterns of association with other concepts. Second, we test recent hypotheses that abstract concepts are organized according to association, whereas concrete concepts are organized according to (semantic) similarity. Third, we present evidence suggesting that concrete representations are more strongly feature-based than abstract representations. We argue that degree of featurebased structure may fundamentally determine concreteness, and discuss implications for cognitive and computational models of meaning.


Keywords: Concreteness; concepts; similarity; association.

## Introduction

A large body of empirical evidence indicates important cognitive differences between abstract concepts, such as guilt or obesity, and concrete concepts, such as chocolate or cheeseburger. It has been shown that concrete concepts are more easily learned and remembered than abstract concepts, and that language referring to concrete concepts is more easily processed (Schwanenflugel, 1991). Moreover, there are cases of brain damage in which either abstract or concrete concepts appear to be specifically impaired (Warrington, 1975). In addition, functional magnetic resonance imaging (fMRI) studies implicate overlapping but partly distinct neural systems in the processing of the two concept types (Binder et al., 2005). Despite these widely known findings, however, there is little consensus on the cognitive basis of the observed differences (Schwanenflugel, 1991). Indeed, while many studies of conceptual representation and organization focus on concrete domains, comparatively little has been established empirically about abstract concepts. ${ }^{1}$

In this paper we test various theoretical claims concerning the abstract/concrete distinction by exploiting large publicly-available experimental datasets and computational resources. By analyzing thousands of abstract and concrete concepts, our approach marginalizes potential confounds more robustly than in smaller-scale behavioral studies. In Analysis 1 we show that abstract concepts are associated in the mind to a wider range of other concepts, although the degree of this association is typically weaker than for concrete concepts. In Analysis 2 we explore the basis of these associations by testing the hypothesis that similarity

[^87]predicts association for concrete concepts to a greater extent than for abstract concepts. In Analysis 3, we show that freeassociation is a more symmetric relation for abstract concepts than for concrete concepts. The findings together suggest contrasts in both the organization and representation of abstract and concrete concepts. We conclude by discussing the implications of the findings for existing theories and models of conceptual representation.

## Data

Our analyses exploit three publicly available resources compiled to assist psychological modeling and analysis.

USF Norms All three experimental analyses use the University of South Florida (USF) Free-association Norms (Nelson \& McEvoy, 2012). The USF data consists of over 5,000 words and their associates. In compiling the data, more than 6,000 participants were presented with cue words and asked to "write the first word that comes to mind that is meaningfully related or strongly associated to the presented word". For a cue word $c$ and an associate $a$, the Forward Association Probability (FAP) from $c$ to $a$ is the proportion of participants who produced $a$ when presented with $c$. FAP is thus a measure of the strength of an associate relative to other associates of that cue.

Many of the cues and associates in the USF data have a concreteness score, derived from either the norms of Paivio, Yuille and Madigan (1968) or Toglia and Battig (1978). In both cases contributors were asked to rate words based on a scale of 1 (very abstract) to 7 (very concrete). ${ }^{2}$

WordNet WordNet is a tree-based lexical ontology containing over 155,000 words produced manually by researchers at Princeton University (Felbaum, 1998). The present work used WordNet version 3.0.

Brown Corpus Word frequencies were extracted from the one million-word Brown Corpus (Kucera \& Francis, 1967), chosen because it is an American corpus compiled at a similar time to the USF data. Word tokens in the Brown Corpus are tagged for their part of speech (POS). For a word type it is then possible to extract the majority POS (the POS with which the type is most frequently tagged).

[^88]
## Analyses

Each of our analyses is motivated by characteristics of the abstract/concrete distinction proposed in theoretical and behavioral studies.

## Analysis 1: Patterns of Association

Motivation Schwanenflugel's Context Availability Model (1991) offers a theoretical basis for the aforementioned empirical abstract/concrete differences. Her exposition of the model relies on the following hypothesis: ${ }^{3}$
(H1) Abstract concepts have more (but weaker) connections (to other concepts) than concrete concepts.

Schwanenflugel presents only small-scale behavioral experiments (64 words, 40 participants) in support of H1. In Analysis 1 we test H 1 on a far larger data set.

Method We extracted those 3,255 pairs in the USF data for which the concreteness of the cue-word was known. Since cue words are connected to a finite set of associates by FAP values, we can isolate a probability distribution over associates for each cue. Since our measure of association strength (FAP) is relative, it is not possible to compare these strengths directly across cue words. Nonetheless, we can make inferences about absolute cue associate strength from properties of the FAP distributions. If a cue has many associates with little variance in the FAP distribution, each FAP value must necessarily be low (and absolute association strength intuitively weak). In contrast, for a given number of associates, higher variance implies that some FAP values are notably higher than the mean, and thus likely to be strong absolutely. Therefore, to address H1 we considered both the dimension (number of associates) and the variance of the FAP distribution for each cue word.

In an initial analysis of the data, we noted a moderate but significant negative correlation between concreteness and frequency, $r(3255)=-.16, p<.001$. Therefore, a multiple regression analysis was conducted with $\log$ (Frequency), Number of Associates and Variance of FAP as predictors, and Concreteness as dependent variable. Because the Concreteness/Frequency multicolinearity was exacerbated by high frequency abstract prepositions and verbs, a second analysis was conducted solely over cue words with majority POS 'noun' $(n=2,320)$.

Results and Discussion In both cases the regression model explained $17 \%$ of the variance of Concreteness and was statistically significant. The beta coefficients in Table 1 indicate that concreteness correlates negatively with both \#Associates and FAP Variance. Both are highly significant

[^89]predictors even when controlling for frequency as an independent predictor.

We have shown that abstract words have more associates than concrete words and lower variance in FAP distributions. This is consistent with the idea that the strength of their associates is on average weaker than for concrete words. Fig. 1 represents the strength of this effect visually. Whilst this confirmation of H 1 is consistent with Schwanenflugel's Context Availability model, it is also consistent with other theoretical characterizations of the abstract/concrete distinction (Paivio, 1986; Markman and Stilwell, 2001). We thus investigate the distinction in more detail in Analyses 2 and 3.

|  | All words |  | Nouns only |  |
| :---: | :---: | :---: | :---: | :---: |
|  | Coeff. ( $\beta$ ) | $t$ | Coeff. ( $\beta$ ) | $t$ |
| \# Assocs | -0.04*** | -16.70 | -0.04*** | -15.97 |
| Variance | -18.01*** | -5.85 | -15.64*** | -4.41 |
| $\log$ (Freq) | -0.18*** | -14.21 | -0.12*** | -7.87 |
| $R^{2}=.17$ |  |  | $R^{2}=.17$ | $57.51^{* * *}$ |

Table 1: Multiple regression analysis of Concreteness


Figure 1: Average FAP mass at each associate rank over the 500 most abstract and concrete cue words in the USF data. Note the stronger initial associates in the concrete case and the longer tail of weak associates in the abstract case.

## Analysis 2: Distinct Conceptual Organization?

Motivation Based on recent behavioral studies of healthy and brain-damaged subjects, (see e.g. Crutch et al., 2009), Crutch and colleagues argue that abstract and concrete concepts differ "qualitatively" in how they relate to other concepts. More specifically, they propose the following:
(H2) Concrete concepts are organized in the mind according to similarity whilst abstract concepts are organized according to association.

The terms association and similarity refer to the ways the concept pairs [car, bike] and [car, petrol] are related: Car is said to be semantically similar to bike, and associated with (but not similar to) petrol. Intuitively, car and bike may be understood as similar because of their common physical features (wheels), their common function (transport), or because they fall within a clearly definable category (modes of transport). By contrast, car and petrol may be associated because they often occur together or because of the functional relationship between them. The two relations are neither mutually exclusive nor independent; bike and car are related to some degree by both association and similarity.

In support of H2, Crutch et al. (2009) asked 20 participants to select the odd-one-out from lists of five words appearing on a screen. The lists comprised either concrete or abstract words (based on ratings of six informants) connected either by similarity (e.g. dog, wolf, fox etc.; theft, robbery, stealing etc.) or association (dog, bone, collar etc.; theft, law, victim etc.), with an unrelated odd-one-out item in each list. Controlling for frequency and position, subjects were both significantly faster and more accurate if the related words were either abstract and associated or concrete and similar. These results support H2 on the basis that decision times are faster when the related items form a more coherent group, rendering the odd-oneout more salient.

Despite the consistency in these findings, each of Crutch et al.'s experiments tested a small sample of subjects ( $<20$ ) with a small ( $<20$ ) number of concepts. It is therefore possible that the observed differences resulted from semantic factors particular to the subjects and items but independent of concreteness. Analysis 2 exploits the USF data and WordNet to investigate H2 more thoroughly.

Method Because similarity and association are not mutually exclusive, H 2 can be interpreted in terms of differing interactions between these two relation types. If concrete concepts are organized in the mind to a greater extent than abstract concepts according to similarity, then the associates of a given concrete concept should be more similar to that concept than the associates of a given abstract concept. In other words, there should be greater correlation between similarity and association over concrete concepts than abstract. We test for this effect with a multiple regression over cue-associate pairs, with FAP as dependent variable and Concreteness, Similarity and their interaction as predictors. Relevant to H 2 is the presence or absence of a positive interaction between concreteness and similarity.

Following other studies of conceptual structure (Markman \& Wisniewski, 1997), we model similarity as proximity in a conceptual taxonomy, in this case, WordNet. Various measures of similarity have been developed for WordNet (see e.g. Resnik, 1995). PathSim, based on the shortest path between two senses, is perhaps the simplest,
and mirrors the manual approach taken by Markman \& Wisniewski (1997). For this experiment, SIM, a measure of the similarity of two words $w_{l}$ and $w_{2}$ on the range [0, 1], was defined as the maximum PathSim between all senses of $w_{1}$ and all senses of $w_{2}$. Since verbs, adjectives and nouns occupy separate taxonomic structures in WordNet, PathSim does not effectively measure similarity across these categories. We thus restrict our analysis to those 18,672 pairs in the FAP data for which cue concreteness and FAP are known and the majority POS for both words is 'noun'.

As a pre-test, SIM was evaluated on Rubinstein and Goodenough's (1965) similarity data for 65 word pairs, ${ }^{4}$ previously used as a benchmark for automatic similarity measures. The correlation between these judgments and SIM, $r(63)=.77, p<.05$, was comparable to other more complex WordNet metrics such as Resnik's (1995) Information Content, $r(63)=.79, p<.05$, and approaching the human replication baseline, $r(63)=.90$ (Resnik, 1995).

Results and Discussion As detailed in Table 2, the regression model was significant, $F(2,3252)=194.53$, and, as expected, SIM was a significant predictor of FAP. The interaction term SIM:Concreteness was positive, as predicted by H2, and a significant predictor of FAP.

Table 2: Multiple regression analysis of FAP over cue (noun) - associate (noun) pairs

|  | Coeff. ( $\beta$ ) | $t$-value |
| :---: | :---: | :---: |
| SIM | 0.048 | 3.66*** |
| Concreteness | 0.003 | 1.64 |
| SIM:Conc | 0.005 | 2.07* |
| $R^{2}=.03, F(3,18665)=194.53$ |  |  |

The positive interaction between similarity and concreteness in our model lends some support to H2. However, the size of this effect is small: the model explains less than .1 of a percentage point more variance in FAP than a model with no interaction term. While statistically significant, this difference is not consistent with a "qualitative difference" in conceptual organization between abstract and concrete concepts, as Crutch and Warrington (2005) propose. Rather, our analysis supports a gradual contrast in patterns of organization along a continuum from concrete to abstract. Of course, qualitative or categorical differences may exist that are too subtle to be detected by the current method. We intend to examine this possibility in future work, using the USF data and WordNet to generate appropriate items for larger-scale behavioral experiments.

## Analysis 3: Distinct Conceptual Representation?

Motivation Hypothesis H2 (Analysis 2) characterizes the abstract/concrete distinction in terms of conceptual

[^90]organization. With respect to the differences in representation that cause the H 2 effect, Crutch and Warrington offer only speculative hypotheses. For instance, they suggest that that "abstract concepts are represented in associative neural networks", whilst "concrete concepts have a categorical organization" (Crutch \& Warrington, 2005, p. 624). Weimer-Hastings and Xu (2005) address this question empirically, and find that people tend to generate fewer "intrinsic" and proportionally more "relational" properties for abstract concepts. Nevertheless, given the untimed, conscious nature of their feature-generation task, and the fact they test only 31 subjects with 36 concepts, the strength of their findings is limited in a similar way to those of Crutch et al. In Analysis 3 we test for evidence of specific representational differences that could explain H2 and the other concreteness effects detailed in the Introduction.

Although the limitations of classical theories of representation with strict binary property specifications are well known, many recent theories characterize representations as feature-based in a more dynamic sense (see e.g. Plaut \& Shallice, 1993). Indeed, the idea of concepts as complexes of conceptually basic features underlines explanations of various empirical observations, including typicality effects, category learning and categoryspecific semantic impairments (Tyler et al. 1995).

Feature-based models are not ubiquitous. Competing approaches such as spatial models (See e.g. Shepard, 1957) or associative networks (Steyvers \& Tennembaum, 2005) have also captured various established cognitive phenomena. One criticism of such models, however, is they naturally model relatedness with a symmetric operation: for all concepts $x$ and $y$, relatedness $(x, y)=\operatorname{relatedness}(y, x)$. As often observed, (Griffiths, et al., 2007; Tversky, 1977) empirical measures of conceptual promixity are in general asymmetric. For instance, it is common to find concept pairs $X$ and $Y$ for which subjects judge the statement ' $X$ is like $Y$ ' to be more acceptable than ' $Y$ is like $X$ '. This effect can be particularly evident when one concept is more salient or prototypical than the other ('Justin Bieber is like Elvis' vs. 'Elvis is like Justin Bieber'?). Asymmetries are also observed in priming effects and free-association, for instance with category name/member or whole/part pairs (Alsatian primes dog more than dog primes Alsatian).

A noted strength of feature-based models is that they naturally capture the asymmetry of semantic relations. In the Contrast Model, Tversky (1977) proposes that the similarity of conceptual representations is computed as some continuous function of their common and distinctive features. Such operations are generally asymmetric, particularly given a disparity in the number of features. For instance, suppose the concept jackal is represented with the features $\{4 L E G S, F U R, H O W L S\}$ and the concept $d o g$ with the features $\{4 L E G S, F U R, T A I L, C O L L A R, ~ L O Y A L$, DOMESTIC\}. Tversky argues that it is more natural to say that jackals are like dogs than vice versa because two thirds of jackal features are shared by dog, whereas only one third of $d o g$ features are shared by jackal. As with other theories
of representation mentioned previously, Tversky's demonstrations are typically confined to concrete words. Nevertheless, his conclusions could be aligned with H2 (Analysis 2) if the following hypothesis held:
(H3) Concrete representations have a high degree of feature-based structure. Abstract representations do not.

Indeed, the soundness of H 3 could point to a causal explanation of the H 2 effect. By H 3 , computing the similarity of abstract concepts by mapping features would be relatively hard. Alternative types of relation would thus be required to group sets of abstract concepts in the mind.

Proposals similar to H3 have been made by several researchers. Plaut and Shallice (1993) showed that integrating differential degrees of feature-based structure into connectionist simulations of dyslexia leads to more accurate replication of established concrete word advantages. Additionally, Markman and Stilwell's (2001) analysis of conceptual category subtypes is entirely consistent with H3. On this view, feature-based categories include those noun concepts typically considered very concrete, whereas abstract noun, prepositions and verbs are all relational categories. Feature-based categories are represented by some configuration of (featural) information 'subordinate to' or 'contained within' that representation (p. 330), whereas relational categories are defined by external information, such as the position of the representation in a relational structure. Finally, H3 is also compatible with the feature-generation data of Weimer-Hastings and Xu (2005).

In Analysis 3 we exploit the USF data to test a prediction of H3. If Tversky's demonstration that asymmetry derives from features is sound, there should be greater asymmetry between concrete concepts than between abstract concepts.

Method Although Tversky's reasoning pertains to a similarity relation, we use the USF data to explore asymmetries in association. Similarity is an important factor in association in general, as evidenced by the high SIM/FAP correlation (Analysis 2). We thus expect asymmetries deriving from similarity to be reflected in FAP values, noting that asymmetry of free-association has been observed previously (Michelbaker et al., 2011).

For each of the 18,668 ordered cue-associate pairs [ $c, a$ ] for which the concreteness of $c$ and $a$ is known, we calculate the (additive) asymmetry $|F A P[c, a]-F A P[a, c]|$. We define the total cue asymmetry, CueAsymm $(c)$, as the sum of the additive asymmetries over all associates of that cue. For a given cue item in our analysis, we experiment with three different measures of concreteness. The first is the cue concreteness Conc(c). Since Tversky's explanation of asymmetry relies on both concepts having a feature-based representation, for each pair $[c, a]$ we also calculate both the sum and the product of concreteness scores. We then define ConcSum(c) as the sum of the sums over all associates, $\operatorname{ConcSum}(c)=\sum_{a} \operatorname{Conc}(\mathrm{c})+\operatorname{Conc}(\mathrm{a})$, and ConcProd(c) as the sum of products $\operatorname{ConcProd}(c)=\sum_{a} \operatorname{Conc}(c) \operatorname{Conc}(\mathrm{a})$. To control for the possibility that FAP asymmetries are caused exclusively by a disparity in frequency between cue and
associate, we also define the measure FreqDisp(c); the sum of the absolute differences between the frequency of a cue word and that of each of its associates, FreqDisp $(c)=$ $\sum_{a} \mid$ Freq(a) - Freq(c)|. We analyse the relationship between CueAsymm (dependent variable) and the three measures of concreteness (predictors) in separate multiple regression models, with FreqDisp as an independent predictor in each.

Results and Discussion The results in Table 3 show a significant positive correlation between the concreteness measure and CueAsymm in all three models, confirming the prediction of H3. Moreover, the model with ConcProd ( $R^{2}$ $=13.73$ ) accounts for more of the CueAsymm variance than with ConcSum ( $R^{2}=.12$ ), which in turn accounts for more than with Conc ( $R^{2}=.08$ ). These two comparisons show that information about the concreteness of both cue and associate is important for predicting asymmetry, consistent with Tversky's explanation of the link between features and asymmetry. It is also notable that FreqDisp is a (marginally) significant predictor in only one of the three models. Therefore the predictive relationship between concreteness and asymmetry (illustrated in Fig. 2) does not derive from discrepancies in frequency between words.

Table 3: Multiple regression analyses of CueAsymm

|  | Coeff. ( $\beta$ ) | $t$ |
| :---: | :---: | :---: |
| Conc | 0.001*** | 16.28 |
| FreqDisp | -0.000 | -1.44 |
| $R^{2}=.08, F(2,3252)=135.60^{* * *}$ |  |  |
| ConcSum | 0.003*** | 21.33 |
| FreqDisp | -0.000* | -2.43 |
| $R^{2}=.12, F(2,3252)=230.92^{* * *}$ |  |  |
| ConcProd | 0.001*** | 22.60 |
| FreqDisp | -0.000 | -0.39 |
| $R^{2}=.14, F(2,3252)=258.81 * * *$ |  |  |

Table 4: USF pairs with highest and lowest asymmetry

| Cue (conc) | Associate | FAP | Backward AP | Additive asymmetry |
| :---: | :---: | :---: | :---: | :---: |
| Keg (6.87) | Beer (5.83) | 0.885 | 0 | 0.885 |
| Text (5.80) | Book (6.09) | 0.881 | 0 | 0.881 |
| Fish (5.84) | Trout (5.93) | 0.036 | 0.913 | 0.877 |
| How (1.57) | Method (2.2) | 0.014 | 0.014 | 0 |
| Honor (1.75) | Courage (2.51) | 0.014 | 0.014 | 0 |
| Immoral (1.81) | Dishonest (2.63) | 0.014 | 0.014 | 0 |

In a separate analysis, we observed that the ConcProd model over pairs in which the cue word is a noun ( $R^{2}=$ 0.1325 ) fits better than the model over pairs in which the cue is a non-noun $\left(R^{2}=0.0987\right)$ or specifically a verb $\left(R^{2}=\right.$ 0.114 ). Indeed, across all 18,668 pairs, the mean additive asymmetry when both cue and associate are nouns (.071) is
significantly greater than when both are not (.066), $t$ (9351.3) $=2.78, p<.01$. Together with Tversky's analysis, these observations are consistent with Markman and Stilwell's proposal that many noun representations are feature-based whereas representations of verbs and prepositions rely on features to a lesser extent.


Figure 2: Scatterplot of CueAsymm vs. ConcProd.

## Conclusion

In this study we have reported the following effects of increasing conceptual concreteness:

## 1. Fewer, but stronger associates (Analysis 1).

2. A stronger correlation between the similarity of concepts and the strength of their association (Analysis 2).
3. Greater asymmetry of association (Analysis 3).

These findings derive from analyses of thousands of concepts and data from thousands of subjects, an approach that significantly increases their robustness in comparison with previous behavioral experiments.
Finding 3 is consistent with, and arguably suggestive of, the view that concrete representations are more strongly feature-based than abstract concepts. Instead of a strongly feature-based structure, abstract representations encode patterns of relations with other concepts (both abstract and concrete). We hypothesize that the degree of feature-based structure is the fundamental cognitive correlate of what is intuitively understood as concreteness.
On this account, computing the similarity of two concrete concepts would involve a (asymmetric) feature comparison of the sort described by Tversky. In contrast, computing the similarity of abstract concepts would require a (more symmetric) comparison of relational predicates such as analogy or structure-mapping (Markman \& Gentner, 1993). Because of their representational structure, the feature-based operation would be simple and intuitive for concrete concepts, so that similar objects (of close taxonomic categories) come to be associated. On the other hand, for abstract concepts, perhaps because structure mapping is more complex or demanding, the items that come to be associated are instead those that fill neighboring positions in
the relational structure specified by that concept (such as arguments of verbs or prepositions). Intuitively this would result in a larger set of associates than for concrete concepts, as confirmed by Finding 1. Moreover, such associates would not in general be similar, as supported by Finding 2.

If this is correct, it is likely that computational models of meaning could be improved by integrating a dimension of concreteness. For instance, models that connect words via syntagmatic co-occurrence would be particularly appropriate for modeling human association in abstract domains, whereas approaches based on taxonomies, or those measuring paradigmactic co-occurrence, would better reflect similarity and be more apt for concrete domains. In future work we plan to test these hypotheses by analyzing how concreteness is reflected in running text corpora.

## Acknowledgments

Thank you to Barry Devereux for helpful comments. The work was supported by St John's College, the Royal Society and the Arts and Humanities Research Council.

## References

Binder, J., Westbury, C., McKiernan, K., Possing, E., \& Medler, D. (2005). Distinct brain systems for processing concrete and abstract concepts. Journal of Cognitive Neuroscience 17(6), 905-917.
Crutch, S., Connell, S., \& Warrington, E. (2009). The different representational frameworks underpinning abstract and concrete knowledge. Quarterly Journal of Experimental Psychology, 62(7), 1377-1388.
Crutch, S., \& Warrington, E. (2005). Abstract and concrete concepts have structurally different representational frameworks. Brain, 128(3), 615-627.
Felbaum, C. (1998). WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.
Griffiths, T., Steyvers, M., \& Tenembaum, J. (2007). Topics in semantic representation. Psychological Review, 114 (2), 211-244.

Kucera, H., \& Francis, W. (1967). Computational Analysis of Present-day American English. Providence, RI: Brown University Press.
Markman, A, \& Gentner, D. (1993). Structural alignment during similarity comparisons. Cognitive Psychology,25, 431-467.
Markman, A, \& Stilwell, C. (2001). Role-governed categories. Journal of Theoretical and Experimental Artificial Intelligence, 13, 329-358.
Markman \& Wisniewski, E. (1997). Similar and different: The differentiation of basic level categories. Journal of Experimental Psychology: Learning, Memory and Cognition. 23(1), 54-70
Michelbaker, L., Evert, S., \& Schütze, H. (2011). Asymmetry in corpus-derived and human word associations. Corpus Linguistics and Linguistic Theory, 7(2), 245.

Nelson, D., \& McEvoy, C. (2012). The University of South Florida Word Association Norms. Retrieved online from: http://web.usf.edu/FreeAssociation/Intro.html.
Paivio, A., Yuille, J., \& Madigan, S. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. Journal of Experimental Psychology Monograph Supplement, 76(1, Pt. 2).
Plaut, D., \& Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. Cognitive Neuropsychology, 10, 377-500.
Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. Proceedings of IJCAI-95.
Rubenstein, H., \& Goodenough, J. (1965). Contextual correlates of synonymy. Communications of the ACM 8(10), 627-633.
Schwanenflugel, P. (1991). Why are abstract concepts hard to understand? In P. Schwanenflugel. The psychology of word meanings (pp. 223-250). Hillsdale, NJ: Erlbaum.
Shepard, R. (1957). Stimulus and response generalization: a stochastic model relating generalization to distance in psychological space. Psychometrika, 22, 325-345.
Steyvers, M., \& Tennembaum, J. (2005). The large-scale structure of semantic networks. Cognitive Science 29(1), 41-78.
Toglia, M., \& Battig, W. (1978). Handbook of semantic word norms. Hillsdale, N.J: Erlbaum.
Tversky, A. (1977). Features of similarity. Psychological Review, 84, 327-352.
Tyler, L., Moss, H., \& Jennings, F. (1995). Abstract word deficits in aphasia: Evidence from semantic priming. Neuropsychology, Vol 9(3), 354-363.
Warrington, E. (1975). The selective impairment of semantic memory. Quarterly Journal of Experimental Psychology 27(4), 635-657.
Wiemer-Hastings, K., \& Xu, X. (2005). Content differences for abstract and concrete concepts. Cognitive Science 29(5), 719-736.

# How Episodic and Working Memory Affect Rule- and Memory-Based Judgments 

Janina A. Hoffmann (janina.hoffmann@unibas.ch)<br>Bettina von Helversen (bettina.vonhelversen@unibas.ch)<br>Jörg Rieskamp (joerg.rieskamp@unibas.ch)<br>Department of Psychology, University of Basel, Missionsstrasse 62a<br>4055 Basel, Switzerland


#### Abstract

Making accurate judgments is an essential skill in everyday life. However, little is known about the basic cognitive skills required for accurate judgments. Research on judgment and categorization processes suggests that people rely on various strategies when making judgments. These strategies differ in the cognitive abilities they require. Specifically high working memory capacity may benefit rule-based judgments, whereas good long-term memory may be crucial for memory-based judgments. We investigated this hypothesis following an individual differences approach. 177 participants performed two judgment tasks that were either best solved by a rulebased or a memory-based strategy. Additionally, we measured working memory capacity and episodic memory with three tests. Consistent with our hypothesis structural equation modeling showed that working memory capacity predicted judgment accuracy in the rule-based task whereas episodic memory predicted judgment accuracy in the memory-based task. Apparently, different memory abilities are essential for successfully adopting different judgment strategies.


Keywords: Judgment and decision making; working memory; episodic memory

Long-term memory and working memory are to a varying degree engaged in many daily activities. On a shopping trip, for example, people need episodic long-term memory to remember items on the shopping list. Trying to quickly sum up the prices of a shopping basket, however, draws upon working memory. Similarly, everyday judgments, such as judging the skills of a job candidate or the suitability of an apartment, may require both working memory and episodic memory. In this paper, we investigate how memory skills relate to people's success in solving judgment tasks.

## Multiple Cue Judgments

In multiple-cue judgment tasks, people are asked to repeatedly estimate a continuous criterion such as the price of a shopping basket based on a number of cues, for instance the products in the shopping basket. To make such judgments, recent research suggests that people rely on two kinds of judgment strategies: rule-based and memory-based strategies (Juslin, Karlsson, \& Olsson, 2008; von Helversen \& Rieskamp, 2008).

Rule-based strategies assume that people try to explicitly abstract the relationship between the cues and the criterion and integrate this information in a linear additive way. To
estimate the price of a shopping basket, for instance, the shopper may try to estimate the price of each product and add up all prices. Mathematically, this integration process can be described with a linear additive model. The criterion estimate $\hat{c}_{p}$ of an object $p$ is the weighted sum of the cue values $x_{p i}$ :
$\hat{c}_{p}=k+\sum_{i=1}^{I} w_{i} \cdot x_{p i}$
where $w_{\mathrm{i}}$ are the cue weights for each cue $i$ and $k$ is a constant intercept.

In contrast, memory-based strategies assume that people judge a new object (the probe) by retrieving previously encountered objects (exemplars) from memory. For example, when estimating the price of a shopping basket people may recall how much they spent the last time they went shopping. The more similar a retrieved exemplar (previous shopping baskets) is to the probe (current shopping basket), the more this exemplar influences the probe's criterion estimate. If a shopper bought the same items last time, for instance, he may just recall this price from memory to estimate the new prize.

This judgment strategy is mathematically described with an exemplar model (Juslin, Olsson, \& Olsson, 2003). To determine the similarity, first the distance $d_{p j}$ between the probe $p$ and exemplar $j$ is calculated. This distance is the summed absolute difference of their cue values $x_{p i}$ and $x_{j i}$ on each cue $i$, weighted by a sensitivity parameter $h$.

$$
\begin{equation*}
d_{p j}=h\left(\sum_{i=1}^{I}\left|x_{p i}-x_{j i}\right|\right) \tag{2}
\end{equation*}
$$

These distances are then transformed into similarities $S(p, j)$ with an exponential decay function (Nosofsky \& Zaki, 1998):

$$
\begin{equation*}
S(p, j)=e^{-d_{p j}} \tag{3}
\end{equation*}
$$

To estimate the criterion value $\hat{c}_{p}$, the similarities are weighted with their corresponding criterion values $c_{j}$ and averaged (Juslin et al., 2003).

$$
\begin{equation*}
\hat{c}_{p}=\sum_{j=1}^{J} S(p, j) \cdot c_{j} / \sum_{j=1}^{J} S(p, j) \tag{4}
\end{equation*}
$$

Past research suggests that people shift between rulebased and memory-based judgment strategies depending on
task structure (Juslin et al., 2008; von Helversen \& Rieskamp, 2008). In linear additive judgment tasks, that is in tasks where the criterion can be approximated by a linear additive function of the cues, people generally rely on rulebased strategies. In contrast, in multiplicative judgment tasks, where the criterion can be approximated by a multiplicative function of the cues, memory-based strategies are more frequently used (Hoffmann, von Helversen, \& Rieskamp, 2013; Juslin et al., 2008). However, little attention has been paid to the cognitive abilities these strategies draw upon and how individual differences in cognitive abilities influence strategy selection and performance.

## Memory Processes in Multiple-Cue Judgments

Theories in judgment and categorization propose that rulebased and memory-based judgment strategies build on different memory abilities. For instance, Ashby and O'Brien (2005) suggested that executing simple rule-based categorization strategies requires working memory capacity, whereas exemplar retrieval involves episodic memory. In a similar vein, Juslin et al. (2008) argued that cue abstraction could be conceived as a capacity-constrained sequential process, whereas memory-based judgment strategies rely on a controlled retrieval process.

Previous research has often studied how working memory influences judgment and categorization performance. In line with a capacity-constrained abstraction process, cognitive load impairs performance in rule-based categorization tasks more than performance in implicit information-integration tasks (Zeithamova \& Maddox, 2006). Indeed, cognitive load may even induce people to shift from a rule-based to a memory-based strategy suggesting that memory-based strategies require less cognitive control (Hoffmann et al., 2013). Yet, some research also suggests that working memory may play a crucial role in learning in all judgment tasks. Indeed, performance in a range of judgment tasks can be predicted by measures of working memory and intelligence (Weaver \& Stewart, 2012). Similarly, Lewandowsky (2011) found that high working memory capacity benefitted learning in rule-based as well as memory-based categorization tasks. Thus, it is unclear whether high working memory capacity only benefits rule abstraction processes or whether it benefits performance in all kinds of judgment tasks.

Research relating episodic memory to judgment performance is scarce. Exemplar models predict a relationship between recognition and categorization and, indeed, have successfully modeled both recognition and categorization performance (Nosofsky, 1988). Consistent with a controlled retrieval process, the instruction to learn all exemplars by heart improves performance in a difficult memory-based judgment task (Olsson, Enkvist, \& Juslin, 2006). Also, memorization of single exemplars enhances recognition of these exemplars in a later recognition test (Palmeri \& Nosofsky, 1995). The importance of episodic memory for category learning, however, has been severely
disputed (Knowlton, 1999), leading to a call for more experimental studies (Ashby \& O’Brien, 2005). Taken together, although some evidence suggests that people engage in a controlled retrieval process when solving memory-based judgment tasks, the role of episodic memory in categorization and even more for judgments is still unclear.

## The Present Study

Our study investigates how episodic and working memory skills affect judgment performance in rule-based and memory-based judgments. We test the hypothesis that judgment accuracy is related to working memory capacity when people rely on a rule-based judgment strategy. Likewise judgment accuracy should be related to episodic memory when people adopt a memory-based judgment strategy. To test these hypotheses, the participants solved a linear as well as a multiplicative judgment task. In addition, we measured participants' working memory and episodic memory skills using three different tests.

## Participants

177 participants (113 female, $M_{\text {Age }}=24.1, S D_{\text {Age }}=6.2$ ) were recruited at the University of Basel. Participants received a participation fee of 20 CHF per hour (approx. 22 US-\$) and an additional bonus in the judgment tasks ( $M=$ $10.3, S D=2.4$ ). One subject was excluded from the analysis because he guessed in the judgment tasks.

## Automated Working Memory Span Tasks

Working memory span tasks were designed to measure both storage and processing of information in working memory (Redick et al., 2012). In working memory span tasks, participants process one set of stimuli while remembering another set of stimuli. For instance, in each trial of the operation span task, participants first see a simple equation. After they have solved the equation and given the answer, they see the first letter that has to be remembered. Subsequently, another equation is presented and another letter has to be remembered, until the set size (the number of presented letters) is reached. Finally, participants are asked to recall the letters in the order of their appearance. Trials with different set sizes are randomly interspersed, with each set size repeated three times. All span tasks were taken from Unsworth et al. (2009) and translated to German.

Reading Span In the reading span participants judged the plausibility of a sentence while remembering letters. Set size varied from 3 to 7 .

Operation Span Participants were asked to solve mathematical equations while remembering letters. Set size varied from 3 to 7 .

Symmetry Span Participants judged the symmetry of a chessboard picture while remembering the position of squares in $4 \times 4$ matrix. Set size varied from 2 to 5 .

## Episodic Memory Tasks

We measured episodic memory with three different tasks: a free recall task with pictures, a cued recall task with numbers, and a recognition test of verbs.

Picture Free Recall We selected 20 pictures from a picture database (Rossion \& Pourtois, 2004) that had high ratings on imagery and concreteness. Each picture was presented for 3 s on the screen and participants were asked to remember them. After a retention interval of 2 minutes participants recalled the pictures.

Cued Number Recall We assessed cued number recall with a computerized version of the Cued Number Recall task from the BIS 4 (Jäger, Süß, \& Beauducel, 1997). 15 pairs of a two- and a three-digit number were presented for 10 s each on the screen. After a retention interval of 2 minutes participants saw the cued number pair as well as four threedigit distractors and had to indicate which three-digit number was initially presented together with the two-digit number.

Verb Recognition We selected 40 verbs with 5 to 7 letters from the Hager and Hasselhorn database (1994) rated high on imagery and concreteness. Participants learned half of the verbs for 3 s each. After a retention interval of 2 minutes participants indicated whether they recognized the 40 verbs from the learning phase by classifying them as old or new.

## Judgment Tasks

Participants solved both a linear and a multiplicative judgment task. In the linear judgment task, we expected participants to use a rule-based strategy; that is, their judgments should be well described by a linear regression model. In contrast, in the multiplicative judgment task, participants should rely on a memory-based strategy (Juslin et al., 2008).

In the linear judgment task, the criterion $y$ was a linear, additive function of the cues and could thus be perfectly predicted by a rule-based strategy:

$$
\begin{equation*}
y=4 c_{1}+3 c_{2}+2 c_{3}+c_{4} \tag{5}
\end{equation*}
$$

where $c_{1}$ reflects the most important cue according to its cue weight. Each cue value varied between 0 and 5 .

In the multiplicative judgment task the function generating the criterion $y$ included a multiplicative combination of the cues:

$$
\begin{equation*}
y=\left(4 c_{1}+3 c_{2}+2 c_{3}+c_{4}+2 c_{1} c_{2} c_{3}+c_{2} c_{3} c_{4}\right) / 8.5 \tag{6}
\end{equation*}
$$

Because of the interacting cues, abstracting linear additive rules does not help solve the task. Therefore, people should switch to exemplar-based strategies and store the objects and the associated criterion values in exemplar memory (Juslin et al., 2008).

We used two different cover stories for the linear and the multiplicative multiple-cue judgment task. In the linear judgment task, participants judged how well a comic figure performed in a game on a scale from 0 to 50 . In the multiplicative judgment task, participants estimated how
toxic a bug was on a scale from 0 to 50 . The stimuli for the two cover stories consisted of pictures of either bugs or comic figures. These bugs and comic figures varied on four different continuous cues. The bugs varied on the length of their legs, their antennae, and their wings and the number of points on their back. The comic figures had different sizes of their ears and their nose, a different number of hairs and stripes on their shirt. These visual features were randomly assigned to the cues.

Both tasks consisted of a training phase and a test phase. During the training phase, participants learned to estimate the criterion values for 25 exemplars. In each trial, participants first saw a picture of a bug or a comic figure and were asked to estimate its criterion value. Afterwards they received feedback about the correct value, their own estimate and the points they earned. The training phase ended after 10 blocks. In the subsequent test phase, participants judged 15 new probes four times, but did not receive any performance feedback.

To motivate participants to reach a high performance, participants could earn points in every trial. The number of points they earned was a truncated quadratic function of the deviation of their judgment $j$ from the criterion $y$ :

$$
\begin{equation*}
\text { Points }=20-(j-y)^{2} / 7.625 \tag{7}
\end{equation*}
$$

At the end of the judgment tasks, the points earned were converted to a monetary bonus ( 1500 points $=1 \mathrm{CHF}$ ). In addition, participants earned a bonus of 3 CHF if they reached $80 \%$ of the points in the last training block.

## Procedure

Participants solved all tasks on one day with half an hour break between the two sessions. The tasks were presented in the same order to each participant. In the first session, participants began with the linear judgment task, moved on to the operation span, solved the verb recognition and the picture free recall task, and finally completed the symmetry span. The second session started with the multiplicative judgment task. Afterwards, participants completed the reading span and finally the cued number recall task.

## Results

## Task Performance

We first analyzed participants' average performance in the memory and the judgment tasks (see Table 1 for descriptive statistics). In the working memory tasks, we used the partial credit score, the sum of items recalled in the correct position, as the dependent variable (Conway et al., 2005). If a participant recalled all items correctly, he achieved a score of 75 in the operation span and the reading span and a score of 42 in the symmetry span. Overall, participants recalled more items in the operation and the reading span than in the symmetry span, replicating normative data (Redick et al., 2012). In the episodic memory tasks, we used the percentage of correctly recalled items as the dependent variable. On average, participants remembered a higher

Table 1: Descriptive statistics for the memory and the judgment tasks.

| Task | $M$ | $S D$ | Skew | Kurt |
| :--- | :---: | :---: | :---: | :---: |
| Operation Span | 57.7 | 12.3 | -1.2 | 1.7 |
| Reading Span | 57.1 | 12.2 | -1.8 | 2.3 |
| Symmetry Span | 29.6 | 7.4 | -0.6 | 0.1 |
| Recognition (\% recalled) | .87 | .09 | -0.7 | 0.5 |
| Cued Recall (\% recalled) | .42 | .19 | 0.2 | -0.2 |
| Free Recall (\% recalled) | .46 | .17 | 0.1 | -.01 |
| Linear Judgment |  |  |  |  |
| $\quad$ Last training block | 6.0 | 2.2 | 0.9 | 1.9 |
| $\quad$ Test (Mean) | 5.4 | 1.8 | 0.7 | 0.8 |
| Multiplicative Judgment |  |  |  |  |
| $\quad$ Last training block | 5.2 | 1.8 | 0.7 | 0.6 |
| Test (Mean) | 5.0 | 1.8 | 1.0 | 0.8 |

Note: Skew $=$ Skewness; Kurt $=$ Kurtosis
percentage of items correctly in the recognition task than in the cued recall or the free recall task.

Learning performance in the judgment tasks was measured with the root mean squared deviation (RMSD) between participants' judgments and the correct criterion in the last training block. The learning performance showed that on average participants learned the judgment tasks quite well. However, the multiplicative judgment task was learned more easily than the linear judgment task. Could participants generalize this good performance to judgments for new items in the test phase? We measured judgment performance in the test phase as the RMSD between the correct criterion and participants' mean judgments; that is, the judgment for each probe averaged over the four presentations in the test phase. Performance for new items in the test phase was comparable to performance in the training phase indicating that participants successfully generalized their performance to new items.

To determine which judgment strategy described participants' judgments best, we fitted a linear regression model (see equation 1) and an exemplar model (see equations 2-4) to participants' judgments in the last three training blocks and predicted participants' mean judgments in the test phase (von Helversen \& Rieskamp, 2008). We compared those models to a baseline model that simply estimated participants' mean judgment. Participants were classified as following the strategy that led to the smallest RMSD between model predictions and participants' mean judgments in the test phase. As shown in Figure 1 the judgment process of the participants was highly task sensitive: In the linear judgment task most participants were best described by a linear model, whereas in the multiplicative judgment task, most participants were best described by an exemplar model, $\chi^{2}(2)=95.3, p<.001$.

## Measurement Models

To understand which memory abilities underlie human judgment processes we followed a structural equation


Figure 1. Strategy classification of participants in the linear and the multiplicative judgment task.
modeling approach. Structural equation modeling allows detecting relationships between latent constructs while correcting for the distinct variance of the measures (for a review see Tomarken \& Waller, 2005).

We first estimated two separate measurement models for memory and judgment abilities. These models were later combined into one structural model. All models were estimated using a maximum likelihood estimator with robust standard errors (MLR) because descriptive data indicated some deviations from multivariate normality. The reported $\chi^{2}$ difference tests were performed using the Satorra-Bentler scaled $\chi^{2}$ values (Satorra \& Bentler, 2001).

Measurement Models for Memory Abilities To measure memory abilities, we hypothesized that episodic memory and working memory capacity can be conceived of as two separate latent constructs that may be correlated (Brewer \& Unsworth, 2012). We first fitted a two-factor latent variable model to the memory data assuming no correlation between working memory and episodic memory. All working memory span tasks loaded on one latent factor, while all episodic memory tasks loaded on a second latent factor. Because the residual variance of the manifest variable recognition was estimated to be negative, we fixed it to 0 . This model fitted reasonably well, $\chi^{2}(10)=16.11, p=.10$, $\mathrm{CFI}=.95, \mathrm{RMSEA}=.06, \mathrm{SRMR}=.08$. Allowing working memory capacity and episodic memory to correlate did not significantly improve model fit, $\chi^{2}(9)=14.85, p=.10$, CFI $=.95$, RMSEA $=.06$, SRMR $=.06$. Finally, a one-factor model assuming a correlation of 1 between episodic memory and working memory capacity fitted worse than the two-factor model, $\chi^{2}(10)=128.2, p<.001, \mathrm{CFI}=.01$, RMSEA $=.26$, SRMR $=.16$. In sum, memory abilities in our study were best described by assuming two separate, uncorrelated latent constructs for working memory and episodic memory.

Measurement Models for Judgment Abilities To find out whether performance depends on the judgment task, we fitted three different measurement models for judgment abilities to judgment performance in the four test blocks of the linear and the multiplicative judgment task. We first estimated a two-factor latent variable model assuming no


Figure 2: Structural equation model relating working memory capacity and episodic memory to judgment performance in the test phase. All loadings and correlations are standardized.
correlation between the factors. One factor predicted judgment performance in the linear judgment task, the second factor predicted judgment performance in the multiplicative judgment task. This model did not describe the judgment data well, $\chi^{2}(20)=38.54, p<.01, \mathrm{CFI}=.975$, RMSEA $=.07$, $\mathrm{SRMR}=.11$. Allowing a correlation between the judgment factors improved model fit, $\chi^{2}(19)=$ $28.24, p=.08, \mathrm{CFI}=.99$, $\mathrm{RMSEA}=.05, \mathrm{SRMR}=.03$. Finally, we estimated a one-factor model assuming a correlation of 1 between judgment performance in the linear and the multiplicative task. This one-factor model could not account for the judgment data, $\chi^{2}(20)=362.14, p<.001$, $\mathrm{CFI}=.54, \mathrm{RMSEA}=.31, \mathrm{SRMR}=.22$. In sum, a twofactor model with correlated factors captured performance variations within the judgment tasks best. This suggests that although performance in rule-based and memory-based judgment tasks is correlated, distinct processes may account for performance differences between the tasks.

## Linking Memory Skills to Judgment Performance

Next, we investigated the link between memory abilities and judgment performance. Based on our prediction, we estimated a structural model (depicted in Figure 2) relating working memory capacity to judgment performance in the linear task and episodic memory to judgment performance in the multiplicative task. This model provided a good fit to the data, $\chi^{2}(75)=89.93, p=.12, \mathrm{CFI}=.98, \mathrm{RMSEA}=.03$, $\mathrm{SRMR}=.08$. Allowing a correlation between working memory capacity and judgment performance in multiplicative tasks and a correlation between episodic memory and judgment performance in linear tasks did not significantly improve the fit of the structural model, $\chi^{2}$ (73) $=85.27, p=.15, \mathrm{CFI}=.99, \mathrm{RMSEA}=.03, \mathrm{SRMR}=.06$. Also, a structural model assuming that memory abilities do not predict judgment abilities could not account for the data,
$\chi^{2}(77)=107.48, p=.01, \mathrm{CFI}=.97, \mathrm{RMSEA}=.05, \mathrm{SRMR}$ $=.10$. Indeed, setting the weight from working memory to linear task performance to 0 decreased model fit, $\Delta \chi^{2}(1)=$ $4.10, p=.04$. Likewise, setting the weight from episodic memory to multiplicative task performance to 0 decreased model fit, $\Delta \chi^{2}(1)=12.67, p<.001$. Thus, while judgment accuracy in rule-based tasks was predicted by working memory capacity, judgment accuracy in memory-based tasks was predicted by episodic memory.

## Discussion

Our study sheds light on which memory abilities people rely when making judgments, a topic that has received little attention in the literature. As the first study linking memory abilities to performance in judgment tasks, we found that working memory capacity predicted judgment accuracy in a linear task, whereas episodic memory predicted judgment accuracy in a multiplicative task. Furthermore, participants relied on a rule-based strategy in the linear task and a memory-based strategy in the multiplicative task. In line with theories of judgment and categorization (Ashby \& O'Brien, 2005; Juslin et al., 2008) this suggests that the two strategies draw upon different memory abilities.

Our results suggest that working memory capacity only predicted judgment performance in rule-based judgment tasks. This result seems to contradict research linking working memory capacity to performance in rule-based and memory-based categorization tasks (Lewandowsky, 2011). One reason for these diverging results may be that our study focused on the differences between judgment tasks, namely the covariance that was not explained by a common judgment factor. Yet, Lewandowsky concentrated on the similarities among categorization tasks. Another reason for these diverging results may be that our study focused on the generalization to new items instead of the learning process.

Indeed, in Lewandowsky's study a learning parameter was strongly related to working memory capacity. Thus, while learning to apply a rule-based or a memory-based judgment strategy may require working memory capacity, only the correct execution of a rule-based judgment strategy may draw upon working memory capacity. Executing a memorybased judgment strategy may instead involve episodic memory skills.

Few studies have examined the link between episodic memory and judgment abilities. Our study clearly shows that episodic memory is related to performance in memorybased judgments. This result highlights the importance of episodic memory for judgments and resonates well with theories suggesting that exemplars are stored and deliberatively retrieved from long-term memory (Juslin et al., 2008). It is also in line with research arguing for exemplar processes in categorization (Nosofsky \& Zaki, 1998). Beyond that, our results highlight that a multitude of cognitive skills, not only working memory, is involved when people make judgments. Shifting the focus to longterm memory may open up new research questions and applications. For instance, memory-based judgment strategies may be more vulnerable to forgetting and interference. Knowledge about storage and retrieval processes in judgment may thus help improving judgments ranging from simple daily judgments such as estimating the price of a shopping basket to professional judgments such as judging the quality of a job candidate.

## Acknowledgments

This research was supported by a grant from the Swiss National Science Foundation (100014_130192/1).

## References

Ashby, F. G., \& O'Brien, J. B. (2005). Category learning and multiple memory systems. Trends in Cognitive Sciences, 9, 83-89.
Brewer, G. A., \& Unsworth, N. (2012). Individual differences in the effects of retrieval from long-term memory. Journal of Memory and Language, 66, 407-415.
Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., \& Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. Psychonomic Bulletin \& Review, 12, 769-786.
Hager, W., \& Hasselhorn, M. (Eds.). (1994). Handbuch deutschsprachiger Wortnormen. Göttingen: Hogrefe.
Hoffmann, J. A., von Helversen, B., \& Rieskamp, J. (2013). Deliberation's blindsight: How cognitive load can improve judgments. Psychological Science.
Jäger, A. O., Süß, H.-M., \& Beauducel, A. (1997). Berliner Intelligenzstruktur-Test. Form 4. Göttingen: Hogrefe.
Juslin, P., Karlsson, L., \& Olsson, H. (2008). Information integration in multiple cue judgment: A division of labor hypothesis. Cognition, 106, 259-298.
Juslin, P., Olsson, H., \& Olsson, A.-C. (2003). Exemplar effects in categorization and multiple-cue judgment.

Journal of Experimental Psychology: General, 132, 133156.

Knowlton, B. J. (1999). What can neuropsychology tell us about category learning? Trends in Cognitive Sciences, 3, 123-124.
Lewandowsky, S. (2011). Working memory capacity and categorization: Individual differences and modeling. Journal of Experimental Psychology: Learning, Memory, and Cognition, 37, 720-738.
Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. Journal of Experimental Psychology: Learning, Memory, and Cognition, 14, 700-708.
Nosofsky, R. M., \& Zaki, S. R. (1998). Dissociations between categorization and recognition in amnesic and normal individuals: An exemplar-based interpretation. Psychological Science, 9, 247-255.
Olsson, A.-C., Enkvist, T., \& Juslin, P. (2006). Go with the flow: How to master a nonlinear multiple-cue judgment task. Journal of Experimental Psychology: Learning, Memory, and Cognition, 32, 1371-84.
Palmeri, T. J., \& Nosofsky, R. M. (1995). Recognition memory for exceptions to the rule. Journal of Experimental Psychology: Learning, Memory, and Cognition, 21, 548-568.
Redick, T. S., Broadway, J. M., Meier, M. E., Kuriakose, P. S., Unsworth, N., Kane, M. J., \& Engle, R. W. (2012). Measuring working memory capacity with automated complex span tasks. European Journal of Psychological Assessment, 28, 164-171.
Rossion, B., \& Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's object set: The role of surface detail in basic-level object recognition. Perception, 33, 217-236.
Satorra, A., \& Bentler, P. M. (2001). A scaled difference Chi-square test statistic for moment structure analysis. Psychometrika, 66, 507-514.
Tomarken, A. J., \& Waller, N. G. (2005). Structural equation modeling: Strengths, limitations, and misconceptions. Annual Review of Clinical Psychology, 1, 31-65.
Unsworth, N., Redick, T. S., Heitz, R. P., Broadway, J. M., \& Engle, R. W. (2009). Complex working memory span tasks and higher-order cognition: A latent-variable analysis of the relationship between processing and storage. Memory, 17, 635-654.
von Helversen, B., \& Rieskamp, J. (2008). The mapping model: A cognitive theory of quantitative estimation. Journal of Experimental Psychology: General, 137, 7396.

Weaver, E. A., \& Stewart, T. R. (2012). Dimensions of judgment: Factor analysis of individual differences. Journal of Behavioral Decision Making, 25, 402-413.
Zeithamova, D., \& Maddox, W. T. (2006). Dual-task interference in perceptual category learning. Memory \& Cognition, 34, 387-398.

# Preschoolers' Trust is Sensitive to Variable Intentions 

Elena Hoicka (elena.hoicka@stir.ac.uk)<br>Psychology Department, University of Sheffield, Western Bank, Sheffield, S10 2PT, UK

Jessica Butcher, Felicity Malla<br>Psychology, School of Natural Sciences, University of Stirling, Stirling, FK9 4LA, UK

Paul L. Harris (paul_harris@gse.harvard.edu)
Harvard Graduate School, Harvard University, Appian Way, Cambridge, MA, 02138, USA


#### Abstract

This research demonstrates that preschoolers flexibly trust and mistrust the same individuals, as preschoolers recognize that their intentions may vary. In Study $1(N=101) 3$ - and 4-yearolds trusted speakers based on their current, rather than previous, intentions to give in/correct information. Thus preschoolers infer the meanings behind different intentions and recognize that intentions change within individuals over time. In Study $2(N=80) 3$ - to 5 -year-olds trusted speakers who were currently sincere, but previously intentionally inaccurate, rather than currently sincere, but previously ignorant, showing that preschoolers infer current knowledge from prior intentions. Preschoolers also trusted speakers who were currently knowledgeable, although previously ignorant, showing that they recognize knowledge is variable within individuals.


Keywords: Trust, Intention, Knowledge, Frame Problem, Preschoolers, Humor

## Introduction

A growing body of research suggests that children do not blindly trust just anyone; children consider whom to trust (e.g., Clément, Koenig \& Harris, 2004; Corriveau, Meints \& Harris, 2009; Koenig \& Harris, 2005). However, in a world where people have variable knowledge and intentions, an important question is whether children are prepared to trust individuals on some occasions but not others (e.g., Nurmsoo \& Robinson, 2009; Scofield \& Behrend, 2008; Shafto, et al., 2012). Thus, the question becomes not just whom to trust but also when to trust a given person. This is important not only for how we acquire information, but also for how we dismiss uninformative or irrelevant information. This is the Frame Problem (e.g., Dennett, 1984), which is still proving difficult to solve in Artificial Intelligence (AI, e.g., Ekbia \& Maguitman, 2001; Scherl \& Levesque, 2003), but which may be a relatively easy problem for preschoolers to solve.
Preschoolers consider past behaviors when deciding whom to trust. For example, when learning new words, both 3 - and 4 -year-olds trust a speaker who previously labeled familiar objects correctly over a speaker who labeled them incorrectly. Thus, children trust accurate over inaccurate speakers (e.g., Clément, et al., 2004; Koenig \& Harris, 2005). Children also trust knowledgeable over ignorant actors (Einav \& Robinson, 2011).

Our first goal was to discover whether preschoolers trust an informant's claims depending on their current intentions and not just on their previous accuracy or apparent knowledge. More specifically, we asked: (1) whether preschoolers are flexible when they trust others; (2) whether mental states, specifically intentions, play a role in trust; and (3) whether preschoolers understand that people's intentions can change over time. Most of the research to date suggests that people who were previously accurate or knowledgeable can be trusted in future, whereas people who were previously inaccurate or ignorant cannot. However people are not statically trustworthy or untrustworthy (e.g., Nurmsoo \& Robinson, 2009; Scofield \& Behrend, 2008; Shafto et al., 2012). Rather, a person can be trustworthy at times, but not at others.

Joking is a clear example of intentionally saying or doing the wrong thing (e.g., Hoicka, Jutsum, \& Gattis, 2008; Leekam, 1991), and so it is an occasion when the audience should not trust the information provided. Indeed, jokers want their audience to know about their falsehood, and they do not expect the audience to believe any part of it (e.g., Leekam, 1991). Thus, people provide cues when they are joking (e.g., Hoicka \& Gattis, 2012; Mireault, et al., 2012). In particular, parents express greater disbelief when joking as compared to acting literally (Hoicka, et al., 2008). Additionally, everyone jokes. For example, all 3-year-olds in a survey were reported by their parents to have produced novel jokes (Hoicka \& Akhtar, 2012). Therefore, joking is an ideal way to examine whether preschoolers use intent to say or do the wrong thing as a cue not to learn.

Research has started to consider whether young children understand the various contexts in which others intend to say or do the wrong thing. From 25 months, toddlers copy wrong actions marked as jokes (laughter), but correct the same wrong actions marked as mistakes ("Whoops!"; Hoicka \& Gattis, 2008). From 30 months, toddlers copy jokers who mislabel familiar objects, but they do not copy people who sincerely mislabel (Hoicka \& Akhtar, 2011). In the case of trust, children as young as 3 years understand that pretending is not a reliable cue for acquiring correct information compared to, for example, having direct experience with the relevant information (Koenig, 2012).

A critical aspect of intention is that it is not a stable mental state. People's intentions change over time (e.g.,

Cohen \& Levesque, 1990; Roy, 2009; van der Hoek, Jamroga, \& Wooldridge, 2007). According to Cohen and Levesque (1990, p. 214), people "keep (or commit to) intentions, but not forever; [they] discharge those intentions believed to have been satisfied". Thus, people can revise or complete their intentions, moving onto new intentions. When joking, the goal is to get a laugh at a moment in time, but not at everything forever.

Our first goal - examined in Experiment 1 - was to find out if children trust someone who currently intends to give correct information over someone who intends to give incorrect information, regardless of their past accuracy or intentions. A major goal of Experiment 2 was to discover whether children can infer a speaker's knowledge from his or her intention. When people intend to do or say the wrong thing, through joking, lying, or pretending (e.g., Hoicka \& Gattis, 2008), the speaker knows the correct information, but chooses not to say it (e.g., Leekam, 1991). Thus, if a joker previously said the wrong thing, he or she likely knew what the right answer was. By contrast, if an ignorant person said the wrong thing, this suggests that the person did not know the right answer. If children understand that joking is more likely to involve intentionally saying the wrong thing compared to being ignorant, then they should later trust a previous joker over a previously ignorant person when learning new information. Thus they may infer that the joker was more likely to know the information than the ignorant speaker, but withheld prior information.

## Experiment 1

Experiment 1 examined whether preschoolers base their trust on speakers' current intentions rather than their past accuracy or intentions. In the final test trials, one actor named novel objects while giving sincere cues whereas the other named them while giving joking cues. Depending on the condition, children had previously seen the actors display the same intentions (i.e., sincere versus joking), a switch in intentions, or they had had no prior exposure to the actors' intentions. The experimental question was whether children would be swayed by the actors' current intentions - as expressed in the test trials - or by their past intentions.

## Method

Participants Fifty-three 3-year-olds ( 33 females, $\mathrm{M}=3$ years, 5 months, range $=3$ years, 0 months -3 years, 11 months) and 484 -year-olds ( 25 females, $\mathrm{M}=4$ years, 4 months, range $=4$ years, 0 months -4 years, 11 months) were randomly assigned to one of three conditions: Consistent Intentions, Inconsistent Intentions, and No Prior Intentions. Children were of similar ages across conditions.

Materials The objects in the familiarization trials included a spoon, a bottle, a doll, and a brush. The objects in the action videos included a cookie, a cup, a scarf, and a hat. The objects in the test trials included a brown feathery cat toy, a
red, black and silver DIY object, a blue and white dog toy, and a red and white kitchen utensil. PowerPoint slideshows were made with each slide showing an object and/or actors, or a video (see procedure). Children's responses were videorecorded directly onto the laptop computer.

Design This study was a between-subjects design in which there were three conditions. In the Consistent Intentions condition, one actor was consistently joking and one actor was consistently sincere throughout the familiarization trials, action videos, and test trials. In the Inconsistent Intentions condition, one actor joked in the familiarization trials but was sincere in the action videos and test trials. The other actor was sincere in the familiarization trials but joked in the action videos and test trials. By showing that actors had switched intentions in the action videos, we anticipated that children would be prepared to recognize their new intentions in the test trials. In the No Prior Intentions condition, children saw the objects but not the actors in the training trials, and then participated in the full test trials. The dependent variable was whether children trusted the joker or sincere actor at test trials when learning new labels.

## Procedure

Familiarization Trials: At the start of the familiarization trials in the Consistent Intentions and Inconsistent Intentions conditions, children were shown a video of the two actors being asked to name an object (e.g., spoon). The joker laughed, named it incorrectly (e.g., duck) using a humorous intonation pattern (Hoicka \& Gattis, 2012) and said, "I'm being silly, only joking." The sincere actor labeled it correctly using a sincere intonation pattern (Hoicka \& Gattis, 2012) while smiling. After watching the video children were shown a slide with pictures of the object and the two actors who had named it and were asked, "She called it a [e.g., duck] and she called it a [e.g., spoon]. Can you tell me what it's called?" This repeated for the remaining three trials with different familiar objects. In the No Prior Intentions Condition, children were instead shown a slide with a picture of the familiar object, given two names for the object and then asked to name it. For example for the spoon, they were asked, "Is this a duck or a spoon?"

Action Videos: In the Consistent Intentions condition, the actor who had joked in the familiarization trials also joked in the action videos. Similarly, the sincere actor stayed sincere. In the Inconsistent Intentions condition, the actor who had joked in the familiarization trials became sincere during the action videos. Similarly, the actor who had been sincere in the familiarization trials became humorous.

For each action, each actor said the same line before performing an action such as, "I'm going to put this hat on". The sincere actor then did the correct action (e.g., putting the hat on her head) and the joker did the action incorrectly, (e.g., putting the hat under her arm) and saying, "I'm being silly, I'm only joking" and laughing. This continued for the other three actions. The No Prior Intentions condition did not include action videos.

Test Trials: Children watched four videos where a third actor asked the two actors what a novel object was called. The sincere actor smiled and said, e.g., "That's a mogo" with a sincere intonation pattern. The joker said, e.g., "That's a sepa" in a humorous intonation pattern, then, "I'm being silly, only joking" and laughed. Following the video the child was shown a slide with a photo of the object and the two actors and told, "She called it a sepa and she called it a mogo. Can you tell me what it's called?" This continued for the remaining three trials. In the Consistent Intentions condition, the actors played the same roles as they had for the familiarization trials and action videos. In the Inconsistent Intentions condition, the actor who had joked during the familiarization trials was now sincere (just as she had been during the action trials). By contrast, the actor who had been sincere during the familiarization trials was now joking (just as she had been during the action trials). In the No Prior Intentions condition, the test trials were the first time the children had seen the actors.

## Results

Data were analyzed with logit mixed effects models. Only significant effects and interactions are reported.
See Figure 1 for the percentage of trials on which children chose the sincere actor's label over the joker's, by Condition (Consistent Intentions, Inconsistent Intentions, No Prior Intentions) and Age. The base model was improved by Condition, $X^{2}(2)=13.91, p=.0010$, and Age, $X^{2}(1)=3.84$, $p=.0501$, as fixed effects. The resulting model (loglikelihood $=-149.17, N=340$ ) found children were significantly more likely to trust the sincere actor over the joker at test trials in the Consistent Intentions versus Inconsistent Intentions and No Prior Intentions conditions (both Odds-Ratio, $O R>2.43, p<.0256$ ). Four-year-olds were marginally more likely to trust the sincere actor than 3-year-olds ( $O R=1.78, p=.0526$ ).
When each condition was tested individually, children in the Consistent Intentions and No Prior Intentions conditions (both log-likelihood $>-59.22, \quad N=127 / 101$ ) were significantly more likely to trust the sincere actor than the joker at test trials (both $O R>2.66, p<.0012$ ). Age improved the model for the Inconsistent Intentions condition, $X^{2}(1)=6.03, p=.0141$. The overall model (loglikelihood $=-49.84, N=112$ ) found 4 -year-olds were more likely to trust the sincere actor than the joker at test trials than 3 -year-olds ( $O R=3.61, p=.0180$ ). Follow-up tests found that both 3- and 4 -year-olds (both log-likelihood $>$ $31.41, N=54 / 58$ ) were more likely to trust the sincere actor than the joker at test trials (both $O R>2.85, p<.0160$ ).

## Discussion

Across all three conditions, both 3- and 4 -year-olds were more likely to trust the actor who was sincere at test trials over the actor who was joking. This suggests children take into account the current intention of a speaker when deciding whether to learn from him or her. If the speaker's intention is sincere, children will learn. If the speaker's
intention is to joke, children will not learn. Thus, selective trust is not purely based on past experience with a speaker it is also based on a speaker's current intentions.

## Experiment 2

During the familiarization trials in both conditions in Experiment 2, one actor mislabeled familiar objects due to ignorance, while the other actor mislabeled familiar objects because she was joking. During test trials, the previously ignorant actor gave cues suggesting that she was now knowledgeable, whereas the joker continued to give joking cues (Knowledge Inconsistent condition). Alternatively, both actors gave cues that they were sincere (Knowledge Inferred condition).

The first experimental question concerned the Knowledge Inconsistent condition. Would children recognize that someone who was once ignorant could become knowledgeable, and would be better to trust than a previous joker who intended to continue saying the wrong thing? The second experimental question concerned the Knowledge Inferred condition. Would children distinguish the two types of inaccuracy during familiarization trials; more specifically, would they recognize that a previous joker was more likely to know the correct labels compared to an ignorant speaker, but chose not to say them. If so, during test trials children should trust a previous joker who becomes sincere, and intends to say the right thing, over a previously ignorant actor, who is also sincere, but is less likely to be knowledgeable.

## Method

Participants Thirty 3-year-olds ( 14 females, $\mathrm{M}=3$ years, 5 months, range $=3$ years, 1 month -3 years, 11 months), 28 4 -year-olds ( 13 females, $\mathrm{M}=4$ years, 5 months, range $=4$ years, 0 months -4 years, 10 months), and 225 -year olds ( 13 females, $M=5$ years, 6 months, range $=5$ years, 1 month -5 years, 11 months) were randomly assigned to one of two conditions: the Knowledge Inconsistent condition and the Knowledge Inferred condition. Children were of similar ages across conditions.

Materials Same as Study 1, except that there were an additional two familiarization objects (car, pig).

Design This study was a between-subjects design in which there were two conditions. In the Knowledge Inconsistent condition, one actor was ignorant during training, but knowledgeable during testing, whereas the other actor consistently joked. In the Knowledge Inferred condition, again, one actor was ignorant during training whereas the other joked. During action and test trials, both actors were sincere. Action trials were included to show a change of intentions in the joker, as in Experiment 1.

There were six training trials in both conditions. The joker joked for four trials, and was knowledgeable for two trials. The ignorant actor was ignorant for four trials, and
knowledgeable for two trials. By showing that the ignorant actor could switch knowledge states during training, we anticipated that children would be prepared to recognize the ignorant actor's new knowledge in the test trials in the Knowledge Inconsistent condition. Jokers also had two knowledgeable trials in order to keep (in)accuracy consistent between actors. Training was the same in the Knowledge Inferred condition in order to ensure conditions were comparable. The dependent variable was who children trusted when learning new labels - the previous joker, or previously ignorant actor.

## Procedure

Familiarization Trials: The task proceeded in the same way as in Experiment 1. In both conditions, the joker gave incorrect labels paired with joking cues for four out of six familiarization trials and correct labels paired with knowledge cues for two familiarization trials. The ignorant actor gave incorrect labels paired with ignorance cues for four out of six familiarization trials and correct labels paired with knowledge cues for two familiarization trials. Humorous cues were the same as in Experiment 1. For ignorance cues, the actor shrugged her shoulders and labeled the object incorrectly saying, e.g., "I don't know, that's a train?" Knowledgeable cues involved displaying their knowledge and labeling an object correctly, e.g., "I know this one. That's a spoon."

Action Videos: In the Knowledge Inferred condition only, children were shown action videos which were the same as those used in Experiment 1. The actor who had joked during the familiarization trials carried out the four sincere actions. The actor who had been ignorant during familiarization also carried out the four sincere actions.
Test Trials: The test trials were the same as in Experiment 1. In the Knowledge Inconsistent condition, the previous joker continued to joke during test trials saying, e.g., "That's a sepa, I'm being silly, only joking" and laughing, whilst the previously ignorant actor was now knowledgeable saying, e.g., "I know this one. It's a mogo". In the Knowledge Inferred condition, both actors labeled the novel object giving sincere cues, where they would smile and say, "That's a mogo" or "That's a sepa".

## Results

We built logit mixed effects models as in Experiment 1. No gender or age (over, under 4.5 years) differences were found. See Figure 2 for the percentage of trials on which children chose the previous joker over the previously ignorant actor, by Condition (Knowledge Inconsistent, Knowledge Inferred). The base model was improved by Condition, $X^{2}(1)=50.05, p<.0001$ as a fixed effect. The resulting model (log-likelihood $=-177.49, N=307$ ) found an effect of Condition ( $O R=6.21, p<.0001$ ). When each condition was tested individually, children in the Knowledge Inconsistent condition (log-likelihood $=-73.42$, $N=157$ ) were significantly more likely to trust the previously ignorant actor (now knowledgeable) versus the
previous joker (still joking) at test trials ( $O R=4.77, p<$ .0001). In the Knowledge Inferred condition (log-likelihood $=-93.02, N=137$ ) children were significantly more likely to trust the previous joker over the previously ignorant actor when both were sincere at test trials $(O R=1.40, p=.0505)$.

## Discussion

Experiment 2 shows that children consider intentions in combination with knowledge when deciding whom to trust for information. When both actors were sincere during test trials in the Knowledge Inferred condition, children were more likely to trust the previous joker than the previously ignorant actor. Because both actors were equally inaccurate during the training trials, accuracy could not be used as a cue. Moreover, children in Experiment 1 did not trust the joker, even when no previous training was given, suggesting that children do not simply prefer jokers. Thus, children inferred that despite the joker previously being inaccurate, she likely actually knew the correct information, at least compared to the ignorant actor, and would thus express the correct information when being sincere.

Another possible way to explain the results is that children avoided learning from someone who was previously ignorant. However, this cannot be the case because children chose to learn from the previously ignorant actor rather than the previous joker when she showed signs of knowledge in the test trials in the Knowledge Inconsistent condition. This demonstrates that children are flexible in their trust, and understand that people's knowledge can vary. They acknowledge that sometimes people know words, and sometimes they do not.

## General Discussion

Experiment 1 showed that children trust speakers based on the speakers' current intentions, rather than their previous intentions or accuracy. Children trusted the actor who was currently being sincere versus joking when learning new labels, even when the actors' previous intentions were different or had not been made available to the preschoolers. Thus, preschoolers recognized that when speakers joke, they intend to say the wrong thing, and so should not be trusted to provide accurate information.

Experiment 2 showed that preschoolers combine intention and knowledge states to determine whether information is trustworthy. Specifically, children inferred that when both actors were sincere at test trials, the actor who previously said the wrong thing in the context of a joke was more likely to know the correct labels compared to the actor who previously said the wrong thing due to ignorance. Thus, children recognized that a joker is more likely to know the truth compared to an ignorant speaker, but chooses not to say it. However, when the previous joker continued to joke, and the previously ignorant actor showed that she was knowledgeable at test trials, children were flexible and preferred the previously ignorant actor, recognizing a change in the ignorant actor's knowledge.

## Stable Traits

Much of the research to date has portrayed children's trust as involving the attribution of a stable trait concerning previous accuracy or knowledge (e.g., Clément, et al., 2004; Corriveau, et al., 2009; Einav \& Robinson, 2011; Koenig \& Harris, 2005). This is the first empirical research to show that children consider not only previous accuracy or knowledge, but also the speaker's past and current intentions, when deciding whom to trust. This is an important skill to have because speakers shift rapidly in their intentions, joking at one moment and being sincere the next. This adds to a body of research showing that children are flexible in their trust (e.g., Nurmsoo \& Robinson, 2009; Scofield \& Behrend, 2008; Shafto, et al., 2012).
In the case of intentionally saying the wrong thing, such as joking, it is highly unlikely that someone would always joke, even if most people do joke at certain times (e.g., Hoicka \& Akhtar, 2012) to trust them. The current findings converge with evidence from computational models, which suggest children consider intention and knowledge, and not just accuracy, when deciding whom to trust (Shafto, et al., 2012).

## Intention

Although much research has considered toddlers' understanding that people intend to do the right thing (e.g., Carpenter, Akhtar, \& Tomasello, 1998), understanding complex intentions, such as intentions to do the wrong thing, may be a more refined test of intention understanding as it involves considering why, and not just whether, someone would do something intentionally. Our experiments show that preschoolers respond appropriately to complex intentions from 3 years. Specifically, they recognize that people can intend to do different things for different reasons. They can intend to say the right thing to teach others, or they can intend to say the wrong thing to joke.

A growing body of research suggests that preschoolers understand that people can intend to do the wrong thing (Hoicka \& Akhtar, 2011; Hoicka \& Gattis, 2008; Rakoczy, Tomasello, \& Striano, 2004). The current experiments extend this prior research by showing that preschoolers can make use of this insight when learning new information. Children are thus flexible learners, accepting new information only when appropriate. Preschoolers are therefore able to solve the Frame Problem (e.g., Dennett, 1984) to a relatively sophisticated degree, tracking speakers’ prior and current accuracy, knowledge, and intentions to decide when to accept versus reject information. An important question that follows is how they solve the Frame Problem. In the current studies, social cues clearly helped. Indeed, in the case of humor and humorous intentions, parents scaffold infants' and toddlers' understanding through cues and explicit expressions of disbelief (Hoicka \& Gattis, 2012; Hoicka, et al., 2008; Mireault, et al., 2012). Thus, the Frame Problem may be solved to some extent through social cues and parental scaffolding.

## Knowledge

Analysis based in philosophy and AI shows that intention is not a stand-alone mental state. Rather, to have an intention, one must also have other mental states such as beliefs and knowledge (Cohen \& Levesque, 1990; van der Hoek, et al., 2007). Thus, for children to truly understand others' intentions, they must also understand others' beliefs or knowledge. Experiment 2 provides the first experimental evidence that children as young as 3 years infer knowledge from intentions, and use inferred knowledge to learn from a previously inaccurate person later on.

Although past research demonstrates that children can infer intentions to do the wrong thing when joking (e.g., Hoicka \& Akhtar, 2011; Hoicka \& Gattis, 2008), it was not clear from this research whether children understood that the actor actually knew the correct information. The current research shows that they consider intention alongside other mental states, specifically knowledge.
Experiment 2 also suggests that preschoolers can attribute knowledge to people who were previously ignorant if they later demonstrate cues showing knowledge. This is consistent with previous findings on perceptual access. Children did not trust an informant who could not perceive the information that they needed, but later trusted the same informant when he or she could perceive that information (Nurmsoo \& Robinson, 2009; Robinson, et al., 2011). This flexibility makes sense. For example, sometimes people forget information, but not always. Sometimes people have some knowledge in a domain, but not all knowledge, for example vocabulary, for which there is variation amongst parents (e.g., Huttenlocher, 1998). Thus, a speaker may know some labels, but not all labels. Being sensitive to cues which suggest when someone has knowledge and when they do not, even within the same domain, would thus be a useful tool in selectively trusting, and acquiring information, from others.

## Figures



Figure 1: Percentage of test trials for which children trusted the currently sincere actor over the current joker in Study 1. * $p<.05$. Lines indicate where differences were examined.

Parentheses indicate results summed across groups.


Figure 2. Total percentage of test trials during which children trusted the previously ignorant actor over the previous joker, by condition. ${ }^{*} \mathrm{p} \leq .05$

## Acknowledgments

We thank parents and toddlers for participating in our research. This research was supported by a British Academy Small Research Grant SG110479 awarded to Elena Hoicka.

## References

Carpenter, M., Akhtar, N., \& Tomasello, M. (1998). Fourteen- through 18 -month-old infants differentially imitate intentional and accidental actions. Infant Behavior \& Development, 21, 315-330.
Clément, F., Koenig, M., \& Harris, P. (2004). The ontogenesis of trust. Mind \& Language, 19, 360-379.
Cohen, P. R., \& Levesque, H. J. (1990). Intention is choice with commitment. Artificial Intelligence, 42, 213-261.
Corriveau, K. H., Meints, K., \& Harris, P. L. (2009). Early tracking of informant accuracy and inaccuracy. British Journal of Developmental Psychology, 27, 331-342.
Dennett, D. C. (1984). Cognitive wheels: The frame problem of AI. In C. Hookway (Ed.), Minds, machines and evolution. Cambridge, United Kingdom: Cambridge University Press.
Einav, S., \& Robinson, E. J. (2011). When being right is not enough: Four-year-olds distinguish knowledgeable informants from merely accurate informants. Psychological Science, 22, 1250-1253.
Ekbia, H. R., \& Maguitman, A. M. (2001). Context and relevance: A pragmatic approach. In V. Akman (Ed.), Modeling and using context. Berlin: Springer Verlag.
Hoicka, E., \& Akhtar, N. (2012). Early humor production. British Journal of Developmental Psychology, 30, 586603.

Hoicka, E., \& Akhtar, N. (2011). Preschoolers joke with jokers, but correct foreigners. Developmental Science, 14, 848-858.
Hoicka, E., \& Gattis, M. (2012). Acoustic differences between humorous and sincere communicative intentions. British Journal of Developmental Psychology, 30, 531549.

Hoicka, E., \& Gattis, M. (2008). Do the wrong thing: How toddlers tell a joke from a mistake. Cognitive Development, 23, 180-190.
Hoicka, E., Jutsum, S., \& Gattis, M. (2008). Humor, abstraction, and disbelief. Cognitive Science: A Multidisciplinary Journal, 32, 985-1002.
Huttenlocher, J. (1998). Language input and language growth. Preventive Medicine: An International Journal Devoted to Practice and Theory, 27, 195-199.
Koenig, M. A. (2012). Beyond semantic accuracy: Preschoolers evaluate a speaker's reasons. Child Development, 83, 1051-1063.
Koenig, M. A., \& Harris, P. L. (2005). Preschoolers mistrust ignorant and inaccurate speakers. Child Development, 76, 1261-1277.
Leekam, S. R. (1991). Jokes and lies: Children's understanding of intentional falsehood. In A. Whiten, \& A. Whiten (Eds.), Natural theories of mind: Evolution, development and simulation of everyday mindreading. Cambridge, MA US: Basil Blackwell.
Mireault, G., Poutre, M., Sargent-Hier, M., Dias, C., Perdue, B., \& Myrick, A. (2012). Humour perception and creation between parents and 3- to 6-month-old infants. Infant and Child Development, 21, 338-347.
Nurmsoo, E., \& Robinson, E. J. (2009). Children's trust in previously inaccurate informants who were well or poorly informed: When past errors can be excused. Child Development, 80, 23-27.
Rakoczy, H., Tomasello, M., \& Striano, T. (2004). Young children know that trying is not pretending: A test of the 'behaving-as-if' construal of children's early concept of pretense. Developmental Psychology, 40, 388-399.
Robinson, E. J., Butterfill, S. A., \& Nurmsoo, E. (2011). Gaining knowledge via other minds: Children's flexible trust in others as sources of information. British Journal of Developmental Psychology, 29, 961-980.
Roy, O. (2009). A dynamic-epistemic hybrid logic for intentions and information changes in strategic games. Synthese, 171, 291-320.
Scherl, R. B., \& Levesque, H. J. (2003). Knowledge, action, and the frame problem. Artificial Intelligence, 144, 1-39.
Scofield, J., \& Behrend, D. A. (2008). Learning words from reliable and unreliable speakers. Cognitive Development, 23, 278-290.
Shafto, P., Eaves, B., Navarro, D. J., \& Perfors, A. (2012). Epistemic trust: Modeling children's reasoning about others' knowledge and intent. Developmental Science, 15, 436-447.
van der Hoek, W., Jamroga, W., \& Wooldridge, M. (2007). Towards a theory of intention revision. Synthese, 155, 265-290.

# When Is Language a Window into the Mind? Looking Beyond Words to Infer Conceptual Categories 

Kevin J. Holmes (kjholmes@berkeley.edu)<br>Department of Linguistics, University of California, Berkeley, 1203 Dwinelle Hall, Berkeley, CA 94720 USA<br>Phillip Wolff (pwolff@emory.edu)<br>Department of Psychology, Emory University, 36 Eagle Row, Atlanta, GA 30322 USA


#### Abstract

Language is often regarded as a rich source of evidence about the mind. However, a number of findings challenge this position, at least at the level of words: Where languages differ in their lexical distinctions, conceptual differences are not always observed. We ask here how language might serve as a window into the mind despite an apparently loose connection between words and concepts. We propose that prominent conceptual distinctions, though not necessarily captured by individual words, may be revealed by elements of meaning shared by multiple words. Testing this hypothesis in the domain of space, we show that clusters of spatial terms, identified through dimensionality reduction analyses of semantic similarity data, align with conceptual categories spontaneously accessed during the perceptual discrimination of spatial relations. These findings suggest that aspects of semantic structure beyond the level of words may provide considerable insight into the conceptual system. Implications for research on linguistic relativity are discussed.


Keywords: language and thought; word meaning; concepts; semantic structure; space; categorical perception.

## Introduction

Many cognitive scientists regard language as a window into the mind (Chomsky, 1975; Lakoff, 1987; Pinker, 2007). Complicating this view, however, is the observation that languages differ dramatically in how they partition the world by name (Malt \& Wolff, 2010). Critically, this semantic diversity is not necessarily mirrored by corresponding conceptual diversity: Where languages differ in their lexical distinctions, conceptual differences are not always observed (e.g., Malt et al., 1999; Munnich, Landau, \& Dosher, 2001; Papafragou, Hulbert, \& Trueswell, 2008). Such findings suggest that, at least at the level of words, language may not be a particularly good window into the mind. In this research, we look beyond individual words to identify other aspects of semantic structure that might prove more tightly connected to the conceptual system. In particular, we propose that prominent conceptual distinctions may be revealed by elements of meaning shared by multiple words. Investigating this hypothesis in the domain of space, we identify clusters of spatial prepositions with similar meanings and assess the extent to which those meanings are spontaneously accessed during the nonlinguistic processing of spatial relations. Our ultimate conclusion will be that language can provide an illuminating window into the mind-if you know where to look.

## Dissociations between words and concepts

A large literature documents the pervasiveness of semantic diversity, with cross-linguistic variation in word meaning observed in such disparate domains as artifacts (Malt et al., 1999), spatial relations (Levinson et al., 2003), and number (Frank et al., 2008), among many others. If language is a window into the mind at the level of words, such diversity should also be observed at the conceptual level. That is, speakers of different languages should perform differently on relevant nonlinguistic tasks, in a manner that aligns with the lexical distinctions of their respective languages.

Although this prediction has been supported by a number of studies investigating the Whorfian hypothesis (see Wolff \& Holmes, 2011), other studies have shown striking asymmetries in performance on linguistic and nonlinguistic tasks. Malt et al. (1999) found that speakers of English, Spanish, and Chinese differed markedly in how they named a set of common household containers (e.g., bottles, jars, etc.), yet showed remarkable agreement when sorting the objects based on overall similarity. Munnich et al. (2001) observed that English, Japanese, and Korean speakers differed in their naming, but not their memory, of spatial locations. Papafragou et al. (2008) found that English and Greek speakers described motion events differently despite showing similar attentional patterns when viewing the events. Together, these findings suggest that the distinctions picked out by words are not invariably salient at the conceptual level, implying some degree of dissociation between words and conceptual representations.

Several factors might account for this word-concept mismatch. Word meanings are shaped, to a much greater degree than conceptual knowledge, by historical forces such as language contact and past speakers' concerns (Malt, Gennari, \& Imai, 2010), and by communicative pressures, such as the need to maximize informativeness and minimize cognitive load (Kemp \& Regier, 2012). As a consequence, the words of a language will tend to reflect the language's history and support efficient communication, but may often fail to capture salient conceptual distinctions - despite the long-standing intuition that they should (cf. Rosch, Mervis, Gray, Johnson, \& Boyes-Braem, 1976).

## The semantic clusters hypothesis

Although the factors outlined above render individual words an unreliable guide to conceptual representations, there may be other ways in which language can provide insight into
the conceptual system. Several recent approaches combine semantic data from multiple languages, motivated by the idea that cross-linguistically frequent semantic distinctions may be linked to prominent, perhaps even universal, conceptual ones (Malt et al., 2011; Regier, Khetarpal, \& Majid, in press). However, a similar idea can be applied to a single language: Elements of meaning that are shared by many words - and hence apply across a wide range of communicative contexts - may be particularly likely to capture key conceptual distinctions. Words that share the same element of meaning can be likened to snapshots of the same underlying concept: No single word will capture the concept on its own, but by examining multiple words with closely related meanings, the concept may emerge (cf. Regier et al., in press). Accordingly, conceptually salient distinctions may be revealed by clusters of related words. We call this proposal the semantic clusters hypothesis.

Testing this hypothesis requires (1) identifying clusters of words within a given domain, and (2) assessing their conceptual salience. The first step may be achieved by obtaining a measure of the similarities among all of the words in a domain. A common method for collecting semantic similarity data is to have people divide words into groups based on their meanings (e.g., Wolff \& Song, 2003). Words with similar meanings will tend to be grouped together often, while words with dissimilar meanings will rarely be grouped together. These co-occurrences may be combined across participants to construct a similarity matrix, which in turn may be analyzed using dimensionality reduction techniques, such as multidimensional scaling (MDS). Any clusters of words, or latent categories, within the semantic similarity space for the domain are likely to be revealed by such techniques.

The second step requires examining the extent to which the latent categories factor into cognitive processes unrelated to language. One way to establish the role of categories in nonlinguistic processing is to show that the category membership of a set of items influences how the items are perceived. Items from different categories are often easier to tell apart than items from the same category, even after controlling for the physical distance between the items - a phenomenon known as categorical perception (CP; Goldstone \& Hendrickson, 2010). In the case of latent categories, CP could be tested by having people discriminate among items from the domain of interest, with the items coming from either different latent categories or the same latent category. CP would be indicated by superior discrimination on between- compared to within-category trials. Such an effect, if found, would indicate that the latent categories are spontaneously accessed in a nonlinguistic context, providing evidence for their conceptual salience.

We adopted the approach outlined above to test the semantic clusters hypothesis in the domain of space, a perennial battleground in research on the language-thought interface (Li \& Gleitman, 2002; Majid et al., 2004). In Experiment 1, participants sorted a large inventory of spatial prepositions into groups, and MDS was used to identify
latent categories. Experiment 2 examined the conceptual salience of these categories, using CP as a diagnostic. Recent evidence suggests that CP is stronger in the left hemisphere than the right (Gilbert et al., 2006), even for unnamed categories (Holmes \& Wolff, 2012) - consistent with specialization of the left hemisphere for categorical processing independent of language (Kosslyn et al., 1989). Thus, even though the items within a latent category might share no common name, we expected CP for such categories to be left-lateralized.

## Experiment 1

The goal of Experiment 1 was to obtain a measure of the semantic structure of the spatial domain, from which clusters of prepositions could be identified.

## Method

Participants Sixty-three Emory University undergraduates, all native English speakers, participated for course credit or payment. One participant was excluded for not following instructions.
Materials An inventory of English spatial prepositions was assembled by adapting a comprehensive list from Landau and Jackendoff (1993). Forty-two prepositions were selected from the original list, omitting archaic (e.g., betwixt, without), intransitive (e.g., apart, downstairs), nonspatial, (e.g., ago, despite) and predominantly metaphorical (e.g., in line with) prepositions, and those requiring a phrasal verb construction (e.g., through, as in "pierce through"). The resulting inventory is shown in Table 1.

Table 1: Spatial prepositions used in Experiment 1.

| about | atop | in | past |
| :--- | :--- | :--- | :--- |
| above | before | in back of | to the left of |
| across | behind | in front of | to the right of |
| after | below | inside | to the side of |
| against* | beneath | near | toward |
| along | beside | off | under |
| alongside | between | on | underneath |
| amid | beyond | on top of | up |
| among | by | opposite | within |
| around | down | outside |  |
| at | far from | over |  |
| *excluded from analyses |  |  |  |

Each of the prepositions was printed in bold at the top of a $4 " \times 6$ " index card. Below each term were two example sentences reflecting prototypical spatial usages of the term.
Procedure The experiment consisted of two phases. In the first phase, participants were presented with the randomly ordered stack of index cards and were asked to write a definition for each preposition based on the two example sentences. The purpose of this task was to encourage participants to think relatively deeply about the meanings of the prepositions. For a subset of participants, the term
against was inadvertently omitted from the stack of cards; as a result, this term was excluded from analyses.

In the second phase, participants were asked to divide the index cards into as many groups as they felt were appropriate. They were told that the prepositions in each group should have "essentially the same meaning." Participants were given as much time as they needed to complete both phases of the experiment.

## Results and discussion

The number of groups of prepositions ranged from 5 to 29 ( $M=14.1, S D=5.8$ ). The raw sorting data were converted into a pairwise similarity matrix, with the similarity between each pair of prepositions taken to be the proportion of participants who grouped them together. For example, if all 62 participants grouped above and below together, the similarity between them would be $62 \div 62=1$; if 31 participants grouped above and below together, the similarity between them would be $31 \div 62=.5$, and so on.

The similarity matrix was submitted as input to a MDS algorithm, ALSCAL (ordinal model), and solutions of increasing dimensionality were generated. Because the largest decline in stress (a measure of the degree of fit between the actual and estimated inter-item distances) occurred between 1 and 2 dimensions, the 2-dimensional solution $($ stress $=.26)$ is shown in Figure 1. ${ }^{1}$


Figure 1: Multidimensional scaling solution of sorting data from Experiment $1 . K$-means clusters are marked on the solution and labeled for descriptive purposes.

To help identify clusters within the solution, the estimated inter-item distances were combined into a new pairwise similarity matrix. This matrix was then submitted as input to a series of $K$-means clustering analyses, using increasing values

[^91]of $K$ (i.e., number of clusters). In these analyses, substantial reduction in within-cluster variance occurred up to $K=4$, with only minimal further reduction thereafter. These results suggest that the MDS similarity space is most optimally partitioned into four clusters. These clusters -labeled abovebelow, front-back, left-right, and in - are marked on the solution in Figure 1. ${ }^{2}$ Notably, three of the clusters contain words that are essentially opposite in meaning. This suggests that the clusters cannot be reduced to individual word meanings, but instead may be viewed as latent categories. ${ }^{3}$ The next experiment investigated the conceptual salience of these categories; that is, the extent to which they play a role in the nonlinguistic processing of spatial relations.

## Experiment 2

In Experiment 2, participants were presented with multiple pictures showing spatial relations from the above-below, left-right, and front-back categories. ${ }^{4}$ Their task was to decide whether the pictures were perceptually identical or one of the pictures (the target) was different from the others (the distractors). On "different" trials, the target was from either the same category as the distractors (within-category; e.g., above vs. below) or a different category (betweencategory; e.g., above vs. left). CP would be revealed by faster or more accurate performance on between- than within-category trials. Note that because the target and distractors had different names on both within- and between-category trials, CP - if observed - would reflect the influence of categorical rather than linguistic representations. Given evidence that CP for both named and unnamed categories is left-lateralized (Holmes \& Wolff, 2012), we expected that latent categories would likewise yield left-lateralized CP. To examine this possibility, the location of the target was varied, with left-lateralized CP indicated by stronger CP effects when the target is presented in the right visual field (RVF; i.e., left hemisphere) than the left visual field (LVF).

[^92]
## Method

Participants Twenty-two Emory University undergraduates, all right-handed native English speakers, participated for course credit or payment. Four participants were excluded, 3 for low accuracy ( $<65 \%$ correct on "different" test trials) and 1 for a mean reaction time (RT) greater than 2.5 standard deviations above the mean for all participants.
Materials The materials were 12 pictures of a bird and an airplane (see Figure 2). Each picture displayed the objects from one of 3 perspectives (front, side, or top view). There were 4 pictures from each perspective, each showing the bird in a different location. The distance from the bird to the airplane, as determined by their closest edges, was the same across locations.


Figure 2: Stimuli used in Experiment 2. (a) front view: above, below, left, and right. (b) side view: above, below, front, and back. (c) top view: left, right, front, and back.

In the discrimination task, each display consisted of a fixation marker surrounded by 4 pictures, all from the same perspective (see Figure 3). In each picture, the center of the airplane subtended $11.5^{\circ}(\mathrm{h}) \times 12.8^{\circ}(\mathrm{v})$ visual angle.


Figure 3: Examples of displays used in Experiment 2.
(a) within-category trial (below target, above distractors).
(b) between-category trial (front target, above distractors).

Design and procedure There were 3 blocks of trials, each consisting of 16 practice trials and 192 test trials. All displays in each block were from a single perspective (front, side, or top). The order of the blocks was counterbalanced.

On half of the test trials in each block, the 4 pictures in the display were identical ("same" trials). This resulted in 4 unique "same" displays in each block, with each display presented 24 times. On the other half, 3 pictures (distractors) were identical and the fourth (target) was different ("different" trials). There were 2 kinds of "different" trials: (a) within-category, in which target and distractors were from the same category (above-below, leftright, or front-back); and (b) between-category, in which target and distractors were from different categories (see Figure 3). Across "different" trials, each picture served as the target at all 4 positions in the display ( $2 \mathrm{LVF}, 2 \mathrm{RVF}$ ), resulting in 48 unique "different" displays per block (16 within-category, 32 between-category), each presented twice. Trials were presented in random order.

On each trial, a fixation marker appeared centrally for 500 ms , followed by one of the displays for 200 ms (to discourage eye movements). Participants indicated whether there was an "odd one out" (i.e., target) by pressing the "S" key for same (i.e., no odd one out) or the "D" key for different, using their left and right index fingers, respectively. The next trial began after participants logged a response. Feedback was provided after practice, but not test, trials.

Following the discrimination task, participants wrote a brief description of the relative locations of the bird and airplane in each of the 12 pictures, presented on index cards. This served as a manipulation check to verify that participants interpreted the pictures as showing the spatial relations they were intended to depict.


Figure 4: Results of Experiment 2.
Error bars are $95 \%$ within-subjects confidence intervals. $\mathrm{LVF}=$ left visual field; $\mathrm{RVF}=$ right visual field.

## Results and discussion

As shown in Figure 4, the categories above-below, leftright, and front-back elicited CP: Participants were faster to discriminate spatial relations from different categories than from the same category. Notably, this effect was found only in the RVF, indicating that CP was left-lateralized, consistent with a left hemisphere specialization for categorical processing (Kosslyn et al., 1989) and previous left-lateralized CP studies (e.g., Holmes \& Wolff, 2012).

On the discrimination task, mean accuracy was $89.1 \%$ ( $S D=7.5$ ), with no difference between "same" and
"different" trials $(p>.1)$. Subsequent analyses focused on the "different" trials, for which CP could be assessed. Trials in which participants responded incorrectly (12.2\%) or RT was greater than 2.5 SD from individual means ( $2.8 \%$ ) were excluded. A 2 (visual field: LVF vs. RVF) $\times 2$ (category relation: within- vs. between-category) repeated-measures analysis of variance (ANOVA) on RT for the remaining trials yielded main effects of visual field, $F(1,17)=8.83, p$ $=.009$, and category relation, $F(1,17)=8.53, p=.01$, and an interaction, $F(1,17)=11.60, p=.003$. Participants responded faster on between- than within-category trials with RVF targets, $t(17)=5.17, p<.0001$, but not LVF targets, $t(17)=.34, p>.7$ (see Figure 4), indicating leftlateralized CP. An analogous ANOVA on the accuracy data yielded no significant effects (all $p \mathrm{~s}>.05$ ), suggesting that there was no speed-accuracy tradeoff.


Figure 5: Results of Experiment 2 by category.
Planned comparisons revealed left-lateralized CP for each of the 3 categories assessed. Within-category trials were divided according to the category membership of target and distractors: above-below (i.e., above target and below distractors, or below target and above distractors), left-right, and front-back. For RVF targets, discrimination was faster on between-category trials than on each of the 3 kinds of within-category trials [above-below: $t(17)=2.17, p=.04$; left-right: $t(17)=2.66, p=.02$; front-back: $t(17)=2.13, p=$ .05 ; see Figure 5]. For LVF targets, none of these differences reached significance ( $p \mathrm{~s}>.2$ ).

The results of the picture description task were as expected. Across the 12 pictures, $87.5 \%$ of the descriptions included prepositions from the intended category (e.g., "the bird is above the plane" or "the plane is below the bird" for above pictures); for 6 of the pictures, there was $100 \%$ agreement. The descriptions of the 4 front-back pictures were the most variable. Six participants consistently described these pictures using horizontal or vertical terms (e.g., "to the right of" for the front picture in Figure 2b), implying that they viewed them as 2-dimensional. Importantly, however, any ambiguity in the stimuli could not itself account for left-lateralized CP, as the existence of multiple interpretations for a given picture would presumably lead to slower responses on both within- and between-category trials (if not more so for the latter, given that individual participants occasionally used the same
preposition to describe pictures from different categories, but never the same category). In addition, none of the 216 descriptions included superordinate terms (e.g., "horizontal" for left-right pictures), suggesting that left-lateralized CP was not driven by linguistic representations.

In sum, the results of Experiment 2 provide evidence for the conceptual salience of the above-below, left-right, and front-back categories identified in Experiment 1. The findings support the semantic clusters hypothesis in showing that clusters of related words align with conceptual categories that are spontaneously accessed during nonlinguistic processing, with consequences for simple perceptual decisions.

## General Discussion

Recent research on the language-thought interface has led to a paradox. Although language has long been viewed as a window into the mind, a number of studies have suggested that word meanings may often be dissociated from conceptual representations. The present research offers a potential resolution to this discrepancy: Language may be a better reflection of the conceptual system at the level of clusters of words than at the level of individual words. According to the semantic clusters hypothesis, clusters of words capture salient conceptual distinctions. Consistent with this hypothesis, groups of spatial prepositions that clustered together in a semantic similarity space in Experiment 1 aligned with conceptual categories that yielded CP in Experiment 2. These findings suggest that language and the conceptual system may share a common underlying structure that is obscured when focusing solely on individual word meanings and their conceptual analogues. In related work, we have shown that clusters of words elicit stronger CP effects than individual words (Holmes, 2012), further supporting our conclusions.

We used CP as a tool for assessing conceptual salience, but our findings also inform the nature of CP itself. In particular, our findings show that CP can occur in the perception of relations, not just in the perception of objects or object properties - the focus of the vast majority of previous CP research (Goldstone \& Hendrickson, 2010). Our findings also lend support to the generality of left-lateralized CP effects, recently contested in the domain of color (e.g., Witzel \& Gegenfurtner, 2011), though they challenge the dominant linguistic interpretation of such effects (cf. Gilbert et al., 2006). Given that left-lateralized CP occurred for categories whose members share no common label, the phenomenon appears to be driven by categories rather than their names (Holmes \& Wolff, 2012), despite the propensity to link left hemisphere processing to language.

Although our findings demonstrate a clear connection between language and the conceptual system, they do not address the origins of this connection, including the possibility that language is the causal agent. The spatial clusters identified here, though shown to be conceptually salient in the minds of English speakers, are not necessarily universal. In principle, the clusters might vary cross-
linguistically, and those differences could lead to differences in nonlinguistic spatial processing. On the one hand, such differences seem unlikely because the elements of meaning associated with the clusters presumably reflect, and are constrained by, structure in the world to a much greater extent than are individual word meanings (see Malt et al., 2010). On the other hand, striking cross-linguistic differences in the meanings of spatial terms have been documented (e.g., Levinson et al., 2003), suggesting the possibility of further differences at deeper levels of semantic structure.

If such differences exist, an investigation of analogous conceptual differences would provide a strong test of the semantic clusters hypothesis. Though agnostic with respect to the universality of semantic structure, the hypothesis would predict that clusters of words within a particular language should be conceptually salient for speakers of that language. Thus, speakers of languages with different sets of clusters should show correspondingly different patterns of CP for those clusters. Notably, this kind of Whorfian effect would be unlike any previously reported, in that it would be driven by categories not explicitly encoded in the semantic system and of which many language users likely have no conscious awareness. Probing the existence of such effects may represent the next frontier in research on linguistic relativity.

## Acknowledgments

We are grateful to Larry Barsalou, Harold Gouzoules, Stella Lourenco, Laura Namy, Marjorie Pak, and Terry Regier for helpful feedback, Ted Guio for graphic design, and Tyler Brown and Lee Kugelmann for help with transcribing and coding descriptions. This research was supported by an APA Dissertation Research Award to KJH.

## References

Chomsky, N. (1975). Reflections on language. New York: Pantheon.
Clark, H. H. (1973). Space, time, semantics, and the child. In T. E. Moore (Ed.), Cognitive development and the acquisition of language. New York: Academic Press.
Frank, M. C., Everett, D. L., Fedorenko, E., \& Gibson, E. (2008). Number as a cognitive technology: Evidence from Pirahã language and cognition. Cognition, 108, 819-824.
Gilbert, A. L., Regier, T., Kay, P., \& Ivry, R. B. (2006). Whorf hypothesis is supported in the right visual field but not the left. PNAS, 103, 489-494.
Goldstone, R. L., \& Hendrickson, A. T. (2010). Categorical perception. WIREs Cognitive Science, 1, 69-78.
Holmes, K. J. (2012). Language as a window into the mind: The case of space. Doctoral dissertation, Department of Psychology, Emory University.
Holmes, K. J., \& Wolff, P. (2012). Does categorical perception in the left hemisphere depend on language? Journal of Experimental Psychology: General, 141, 439-443.
Kemp, C., \& Regier, T. (2012). Kinship categories across languages reflect general communicative principles. Science, 336, 1049-1054.

Kosslyn, S. M., Koenig, O., Barrett, A., Cave, C. B., Tang, J., \& Gabrieli, J. D. E. (1989). Evidence for two types of spatial representations: Hemispheric specialization for categorical and coordinate relations. Journal of Experimental Psychology: Human Perception and Performance, 15, 723-735.
Lakoff, G. (1987). Women, fire, and dangerous things. Chicago: University of Chicago Press.
Landau, B., \& Jackendoff, R. (1993). "What" and "where" in spatial language and spatial cognition. Behavioral and Brain Sciences, 16, 217-265.
Levinson, S. C., Meira, S., \& The Language and Cognition Group (2003). 'Natural concepts' in the spatial topological domain-Adpositional meanings in crosslinguistic perspective: An exercise in semantic typology. Language, 79, 485-516.
Li, P., \& Gleitman, L. (2002). Turning the tables: Language and spatial reasoning. Cognition, 83, 265-294.
Majid, A., Bowerman, M., Kita, S., Haun, D. B. M., \& Levinson, S. C. (2004). Can language restructure cognition? The case for space. Trends in Cognitive Sciences, 8, 108-114.
Malt, B. C., Ameel, E., Gennari, S., Imai, M., Saji, N., \& Majid, A. (2011). Do words reveal concepts? Proceedings of the 33rd Annual Conference of the Cognitive Science Society (pp. 519-524). Austin, TX: Cognitive Science Society.
Malt, B. C., Gennari, S., \& Imai, M. (2010). Lexicalization patterns and the world-to-words mapping. In B. C. Malt \& P. Wolff (Eds.), Words and the mind: How words capture human experience. New York: Oxford University Press.
Malt, B. C., Sloman, S. A., Gennari, S., Shi, M., \& Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. Journal of Memory and Language, 40, 230-262.
Malt, B. C., \& Wolff, P. (Eds.) (2010). Words and the mind: How words capture human experience. New York: Oxford University Press.
Munnich, E., Landau, B., \& Dosher, B. A. (2001). Spatial language and spatial representation: A cross-linguistic comparison. Cognition, 81, 171-207.
Papafragou, A., Hulbert, J., \& Trueswell, J. (2008). Does language guide event perception? Evidence from eye movements. Cognition, 108, 155-184.
Pinker, S. (2007). The stuff of thought: Language as a window into human nature. New York: Penguin Books.
Regier, T., Khetarpal, N., \& Majid, A. (in press). Inferring semantic maps. Linguistic Typology.
Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., \& Boyes-Braem, P. (1976). Basic objects in natural categories. Cognitive Psychology, 8, 382-439.
Witzel, C., \& Gegenfurtner, K. R. (2011). Is there a lateralized category effect for color? Journal of Vision, 11.
Wolff, P., \& Holmes, K. J. (2011). Linguistic relativity. WIREs Cognitive Science, 2, 253-256.
Wolff, P., \& Song, G. (2003). Models of causation and the semantics of causal verbs. Cognitive Psychology, 47, 276332.

# Preschoolers' Understanding of Preferences is Modulated by Linguistic Framing 

Taylor F. Holubar (tholubar@stanford.edu)<br>Department of Psychology<br>Stanford University

Ellen M. Markman (markman@stanford.edu)<br>Department of Psychology<br>Stanford University


#### Abstract

Reasoning about others' preferences is an important aspect of understanding the social world. Although there is some evidence that young children reason appropriately about others' discrepant preferences, there are reasons to suspect this ability remains fragile through the preschool years. In particular, we argue that the way preferences are expressed may tap into humans' lifelong tendency toward naïve realism, the belief that my way of seeing the world is the normative, correct one. We present data demonstrating that tolerance for unconventional opinions increases during the preschool years but remains susceptible to influence by linguistic framing.


Keywords: theory of mind; preferences; naïve realism; linguistic framing effects.

## Introduction

Statements of preference are a profoundly strange phenomenon. In principle, preferences are subjective: although finding ice cream delicious may be normative and all but universal, "ice cream is delicious" is not true or false in the way that "ice cream is sugary" is true or "ice cream is hot" is false. Yet statements of this form--describing a subjective valuation as if it were an objective fact--are widespread and remarkably unremarkable. Could talking about preferences as if they were facts influence the way people reason about preferences? We suggest that preferences are a difficult concept to reason about (and particularly susceptible to effects of framing) because they hold a fundamentally different epistemic status from facts. They require acknowledging that one's own (often strongly held) beliefs are not verifiably correct, and that even totally opposite beliefs should be respected as valid. These recognitions require sophisticated perspective taking skills, which preschoolers notoriously lack. However, adults also show similar biases, as the literature on naive realism demonstrates (Robinson, Keltner, Ward, \& Ross, 1995). These observations lead us to ask: how do children reason about preferences that differ from their own? Does the way we talk about preferences impact their reasoning?

It is well documented that preschoolers begin passing classic tests of false-belief understanding around age four (see Wellman, Cross, \& Watson, 2001, for a meta-analysis). A prominent explanation of this shift is that children develop the insight that the mind operates on representations of the world, rather than veridical copies of the world. In other words, for a pre-representational child, mental states are taken as exactly reflecting the state of the world. One
consequence of a non-representational theory of mind could be that a child's evaluative opinions (such as whether she finds ice cream delicious or not) are also taken as direct reflections of objective properties of the world. Such a belief should lead pre-representational children to expect everyone else also to find ice cream delicious, and the subjective, idiosyncratic nature of preferences to be lost.

Some evidence does exist that very young children in fact treat subjective properties like "deliciousness" as features of objects rather than as mental states tied to individuals. Gergely, Egyed, \& Király (2007) report that 14-month-olds expect adults to treat an object in accordance with the total amount of liking or disliking for the object the infant has witnessed, irrespective of whether a particular adult has previously demonstrated liking or disliking for the object. In other words, infants do not reason about an object's likability in terms of a particular person's opinion about that object. Instead, they seem to aggregate information across individuals and associate that totality with the object, not with the individuals who produce the information. This finding is consistent with a relatively impoverished understanding of mental states, whereby children conceptualize beliefs (including those about an item's likability or desirability) as reflecting objective states in the world.

At the same time, the strongest version of this hypothesis is not borne out in the data: at least in some cases, slightly older children respond appropriately to desires different from their own. Children as young as 18 months recognize that adults may desire different foods than they themselves do: when an adult who has previously shown positive affect towards broccoli and negative affect toward goldfish crackers requests more food, 18-month-olds (but not 14-month-olds) respond appropriately to requests for food (Repacholi \& Gopnik, 1997). In some contexts, two-yearolds also show similar understanding of others' desires ( Ma \& $\mathrm{Xu}, 2011$ ). For example, children predict that a character will be happy when she satisfies her desire to play with a given toy, even when the participant has previously said she would choose to play with a different toy over that toy (Wellman \& Woolley, 1991).

Given these discrepant findings, important work remains in charting out how children develop an adult-like understanding of preferences. In this paper, we present two studies in which we specifically probe how preschoolers reason about unconventional preferences--preferences that, while likely not espoused by the child, nonetheless are not
simply false or mistaken in the way that facts can be. We argue that achieving this understanding is an on-going process through the preschool years. Our evidence suggests that preschoolers' understanding of the subjective nature of preferences is fragile and easily disrupted.

## Study 1

## Participants

68 preschoolers from Bing Nursery School at Stanford University participated. Twenty-three 3-to-5-year-old children (12 boys) participated in a marked framing condition. The mean age was 4 years 6 months. Forty-five 3-to-5-year-old children (21 boys) participated in an unmarked framing condition. The mean age was 4 years 5 months.

## Procedure

To evaluate how children understand unconventional preferences, we developed a paradigm in which participants were shown pictures of other children who were said to have answered questions about familiar foods. Questions either concerned factual properties, such as a food's color or texture, or subjective qualities, such as how appealing the food is. All foods used for opinion questions were chosen to be conventionally popular, desirable foods. Characters' responses to opinion questions were either conventional or unconventional (i.e., expressed a positive or negative attitude towards a popular food). Responses to factual questions were either true or false. This permitted us to analyze how children treated conventional and unconventional responses to opinion questions and to compare their treatment of opinion questions with their treatment of factual questions. Each child heard responses to two examples of each of the four statement types (true and false facts, conventional and unconventional opinions) for a total of eight items.

To examine how framing might influence performance, we manipulated how characters responded to opinion questions. In an unmarked framing condition, responses took the form "Ice cream is delicious" or "Ice cream is yucky." In a marked framing condition, responses took the form "I like ice cream" or "I don't like ice cream." We expected children to be more likely to reject an unconventional opinion when it was unmarked than when it was marked as an opinion with "I like."

To assess children's reactions to characters' statements, we directed four questions to participants after each question and statement pair. First, children were asked to explain why the character had made that statement (e.g., "Why did Martin say ice cream is delicious?"). Next, children were asked if the character made a mistake and if the character was "just being silly." (These two questions were designed to cover two of the most frequent categories of explanations that adults gave in pilot work.) Finally, children were shown a picture of another character who was said to hold the opposite belief of the first character (i.e., to hold a negative attitude when the first character held a

| Statement Type | EXAMPLE STATEMENTS |  |
| :---: | :---: | :---: |
|  | Marked <br> Condition | Unmarked <br> Condition |
| True Fact | Milk is white. | Milk is white. |
| False Fact | Milk is green. | Milk is green. |
| Conventional <br> Opinion | I like ice cream. <br> Ice cream is <br> delicious. |  |
| Unconventional <br> Opinion | I don't like ice <br> cream. | Ice cream is <br> yucky. |

Figure 1: Summary of experimental design.
positive attitude or to hold a false belief when the first character expressed a true belief, and vice versa). They were then asked if that contrasting statement "could be right." We thus had three converging yes/no measures of whether children deemed a statement acceptable or mistaken, as well as an open-ended explanation of the characters' beliefs. Although we are primarily interested in children's judgments about opinion statements, including factual questions allowed us a baseline estimate of how often children would label as mistaken statements that adults would also describe as mistakes, against which we could compare children's treatments of opinion statements.

## Results: yes/no questions

For the three yes/no questions, we scored each response as correct or incorrect as follows. True facts should be judged as not a mistake and not silly; when the second character expressed the opposite belief (i.e., a false fact), his response should be judged as not right. False facts should be judged as a mistake and as silly; when the second character expressed the opposite belief (i.e., a true fact), his response should be judged as right. Conventional and unconventional opinions, on the other hand, should be judged the same way as each other: as neither a mistake nor silly. When the second character expressed the opposite opinion (whether conventional or unconventional), their response should be judged as right. To test for possible developmental changes, we divided children into older and younger groups based on the median age for each condition.

As a first analysis, we conducted a logit mixed model using statement type, framing, and age (older or younger) to predict correct responding. We found a significant three-way interaction between statement type, framing, and median age, $\mathrm{b}=-3.4$, S.E. $=1.5, \mathrm{z}=-2.1, \mathrm{p}=.04$. To unpack this threeway interaction, we next analyzed each statement type separately, examining effects of framing and age on children's performance.

True factual statements We first asked how children treated true factual statements (e.g., "Milk is white"). A logit mixed model using framing condition and age group to predict correct responding to questions about true factual statements showed a significant developmental difference $(b=2.4, z=2.4, p=.01)$. The effect of framing was nonsignificant, as expected, since the factual statements were the same across framing conditions. This confirms that children did not differ in their baseline responses to identical statements of true facts across conditions.

To examine performance in a more fine-grained fashion, we created composite scores for each participant by summing the number of correct responses to all yes/no questions for each statement type (e.g., true fact, conventional opinion). The maximum possible score was 6 (three questions per two items for each statement type). In the marked framing condition, both older and younger children appropriately answered questions about true facts at well above chance rates (older: $M=5.5$ out of $6, \mathrm{SD}=.52$, $\mathrm{t}(10)=16.2, \mathrm{p}<.001$; younger: $M=4.3, \mathrm{SD}=1.8, \mathrm{t}(11)=2.5, \mathrm{p}=$. 03). However, as the logit model indicated, older children performed significantly better, $\mathrm{t}(12.9)=2.2, \mathrm{p}=.05$. In the unmarked framing condition, similar patterns emerged, with both older ( $M=5.4, \mathrm{SD}=1.4$ ) and younger children ( $M=4.3$, $\mathrm{SD}=1.8$ ) performing above chance rates, $\mathrm{t}(21)=8.1, \mathrm{p}<.001$ and $\mathrm{t}(22)=3.4, \mathrm{p}=.003$ respectively. As in the marked framing condition, older children somewhat out-performed younger children, $\mathrm{t}(41.1)=2.3, \mathrm{p}=.03$. Thus preschoolers seem to appropriately answer questions about true facts, judging that they are not mistakes at well above chance rates.

False factual statements We then asked how children treated false factual statements (e.g., "Milk is green"). A logit mixed model using framing condition and age group to predict correct responding to these questions showed neither the main effect of age or of framing, nor their interaction, reached significance. The non-significance of the effect of framing was, again, expected, since factual statements were identical across conditions.

On the composite measures, performance was strong for all groups. Questions about false facts were answered appropriately by older children in the marked framing condition at a rate significantly above chance ( $M=5.2$ out of $6, \mathrm{SD}=1.08$ ), $\mathrm{t}(10)=6.7, \mathrm{p}=<.001$. Likewise, younger children also answered appropriately at a rate that exceeded chance $(M=4.8$ out of $6, \mathrm{SD}=1.14), \mathrm{t}(11)=5.32, \mathrm{p}<.001$. Their overall performance was comparable to that seen for the older children, $\mathrm{t}(21.0)=-.93$, n.s.

In the unmarked framing condition, younger children answered an average of 5.2 questions correctly ( $\mathrm{SD}=1.0$ ), which exceeded the number expected by chance, $\mathrm{t}(22)=10.2$, $\mathrm{p}<.001$. Similarly, older children answered an average of 5.3 questions correctly ( $\mathrm{SD}=.98$ ), again more than expected by chance, $\mathrm{t}(21)=10.8, \mathrm{p}<.001$. Older children's performance was not significantly better than younger children's, $\mathrm{t}(43.0)=.18$, n.s. It is not clear why older children would out-perform younger children when questions were asked about true facts but not about false facts. Anecdotally, we have observed that some children are perplexed about the
pragmatics of asking whether a manifestly true fact (like milk being white) is a mistake; it is plausible that younger children answered "yes" to those questions due to some uncertainty about the experimenter's intentions. In any case, these data provide robust evidence that children understand the notion of mistaken facts, leading us to ask how they apply these concepts to opinions.

Conventional opinion statements Preschoolers' treatment of conventional opinions ("I like ice cream" or "Ice cream is delicious") was similar to their treatment of true facts. A logit mixed model using framing condition and age group to predict correct responding to questions about conventional opinion statements showed neither the main effect of age or of framing, nor their interaction, reached significance. When asked about conventional opinions, preschoolers showed a high degree of acceptance without major developmental differences or strong influences of framing. In the marked framing condition, younger children correctly answered these questions at above-chance rates $(\mathrm{M}=4.7, \mathrm{SD}=1.78)$, $\mathrm{t}(11)=3.25, \mathrm{p}=.008$, as did older children $(\mathrm{M}=4.8$ out of 6 , $\mathrm{SD}=1.33), \mathrm{t}(10)=4.54, \mathrm{p}=.001$. The mean number of questions answered correctly did not differ by age group, $\mathrm{t}(20.2)=-.23$, n.s.

In the unmarked framing condition, younger children answered an average of 4.5 questions correctly ( $\mathrm{SD}=1.5$ ), more than expected by chance, $\mathrm{t}(22)=5.0, \mathrm{p}<.001$. Likewise, older children correctly answered an average of 5.0 questions correctly, also more than expected by chance, $\mathrm{t}(21)=6.5, \mathrm{p}<.001$. Performance did not differ by age groups on questions about conventional opinions.

Unconventional opinion statements A different picture emerged when we analyzed preschoolers responses to unconventional opinions ("I don't like ice cream" or "Ice cream is yucky"). The logit mixed model predicting correct responses with framing and age showed a marginal main effect of framing but a significant interaction between framing and age. For younger children, framing had little effect: performance was weak in both framing conditions. In the marked framing condition, younger children correctly answered only 1.6 of the 6 questions about unconventional opinions ( $\mathrm{SD}=2.2$ ), a rate significantly below chance, $\mathrm{t}(11)=-2.2, \mathrm{p}=.05$. In the unmarked framing condition, younger children answered an average of just .70 questions correctly ( $\mathrm{SD}=.93$ ), again well fewer than expected by chance, $\mathrm{t}(22)=-11.0, \mathrm{p}<.001$.

Older children showed a relatively strong understanding of unconventional opinions in the marked framing condition (at least compared to younger children), answering an average of 3.7 correctly ( $\mathrm{SD}=1.5$ ). However, unlike for all other statement types, this success rate did not differ from chance, $\mathrm{t}(10)=1.6$, n.s. Older children's performance was, however, significantly better than younger children's, $\mathrm{t}(19.4)=-2.8, \mathrm{p}=.01$. In the unmarked framing condition, performance dropped dramatically for older children. Older children answered an average of only .59 questions correctly ( $\mathrm{SD}=.80$ ), again fewer than expected by chance, $\mathrm{t}(21)=-14.2, \mathrm{p}<.001$. The difference between conditions was highly significant for older children, $\mathrm{t}(12.9)=6.5, \mathrm{p}<.001$.


Figure 2: Summary of performance by statement type, condition, and age.

Performance in the unmarked framing condition was equally poor for both age groups, $\mathrm{t}(42.5)=.41$, n.s.

Taken together, these data indicate that younger children did not answer our questions about unconventional opinions correctly in either framing condition. Older children did better than younger children in the marked framing condition, although not as well as they had done with questions about facts or conventional opinions. However, framing had a pronounced impact on how older children reasoned about unconventional opinions, reducing their performance to the same level as was seen for younger children.

## Results: open-ended explanations

While the yes/no questions allow us one way of probing children's reasoning about unconventional opinions, there are concerns in the literature that such explicit measures might underestimate children's ability to reason in a sophisticated, adult-like fashion about others' opinions (Banerjee et al., 2007). Exploring children's open-ended explanations for why characters might respond with an unconventional opinion provides a less constrained window into children's spontaneous reactions to unconventional opinions.

It is worth noting that children's "explanations" frequently did not constitute what an adult would call an explanation. Frequently, they were simply comments on or responses to what the character had said. Whatever we call them, though, these comments provide a useful probe. To analyze these data, we coded children's responses as to
whether they indicated agreement or disagreement with the character's statement. Below, we report the mean number of times children disagreed with each statement type. There were two trials featuring each statement type, so the maximum number of disagreements is two. Unlike with the yes/no questions, no main effects of framing or age were found, but for consistency's sake we present means broken down by those variables.

True factual statements Disagreements were rare in response to true factual statements. Younger children in the unmarked framing condition expressed a mean of .25 disagreements, while no disagreements were seen for older children in the unmarked framing condition or in either age group in the marked framing condition.

False factual statements Disagreements were much more frequent in response to false factual statements. In the marked framing condition, younger children disagreed an average of 1.2 times, while older children disagreed an average of 1.1 times. In the unmarked framing condition, younger children disagreed an average of 1.0 times, while older children disagreed an average of 1.2 times Thus across age groups and framing, children expressed disagreement with false facts on more than half of trials. This may not seem terrifically high, but it is worth reiterating that children were asked to explain why the character had made their statement. Disagreements then weren't particularly good answers for the question that had been posed.

Conventional opinions In the marked framing condition, only one younger child ever disagreed, and did so on only one trial. Similarly, in the unmarked framing condition, a total of one disagreement was recorded among younger children and one among older children.

Unconventional opinions. In the marked framing condition, younger children disagreed on an average of .75 trials, while older children disagreed on an average of only . 25 trials. In contrast, in the unmarked framing condition, younger children disagreed on an average of .85 trials, while older children disagreed on an average of .68 trials.

To test whether preschoolers' tendency to disagree differed across statement types, we conducted a logit mixed model using statement type to predict whether the child disagreed on a trial. Contrasts were specified to treat the unconventional opinion as the baseline. Children were significantly less likely to disagree with true factual statements $(b=-3.2, \quad \mathrm{z}=-5.5, \mathrm{p}<.001)$ and conventional opinions ( $\mathrm{b}=-4.1, \mathrm{z}=-5.2, \mathrm{p}<.001$ ) than they were to disagree with unconventional opinions. However, children were more likely to disagree with false factual statements than with unconventional opinions ( $\mathrm{b}=1.5, \mathrm{z}=4.7, \mathrm{p}<.001$ ).

These data present a somewhat different picture from that observed with the yes/no questions. To begin with, age and and framing did not exert significant influences on whether children disagreed with statements. Moreover, although children disagreed with unconventional opinions relatively often, they did not disagree with unconventional opinions as often as they did with false facts. Thus, whereas the yes/no data might lead us to suggest that preschoolers--especially younger ones--robustly fail to understand that opinions differ from facts in that even unconventional opinions are not wrong, our open-ended explanations suggest that preschoolers' understanding is somewhat more nuanced and that they do not entirely conflate facts and opinions.

## Study 2

An important feature of both the framings that we have used above is that they articulate properties of long duration. If ice cream is delicious, it is always delicious, not merely delicious right now. Likewise, if I like ice cream, the suggestion is that I like ice cream in general, not just right now. This permanence (or at least longevity) is very characteristic of facts. In contrast, desires are frequently temporary; they disappear when they have been satisfied. Importantly, even if I like ice cream or agree that ice cream is delicious, I may not always want ice cream, right now. In the literature reviewed above, preferences have always been conceptualized or demonstrated in terms of a situationspecific want. For example, Wellman and Woolley (1991) probe 2-year-olds' understanding of discrepant desires by asking participants which of two equally desirable activities they would want to engage in and saying the character wanted to do the other. This situation of choosing between two attractive options and temporarily prioritizing one is likely to be familiar even to 2 -year-olds, and likely poses weaker demands on children's incipient theory of mind abilities.

One possibility, then, is children's relatively poor understanding of unconventional opinions in study one is that our framings made characters' preferences sound immutable and permanent. If children associate these features with facts but not with desires, children might show themselves to be more tolerant of unconventional opinions if the framing emphasized the transient nature of the unconventional opinion. Below, we present work in progress that tests this possibility by implementing a "want" framing.

## Procedure

The procedure was identical to that of Study 1, except opinion statements were presented using a "want" framing instead of marking them with "I like" or not marking them. Thus, characters who expressed a conventional opinion said, for example, "I want some ice cream" while characters who expressed an unconventional opinion said, for example, "I don't want some ice cream."

## Results

Data collection is on-going. However, in the condition of interest, unconventional opinions, it is clear that this "want" manipulation has not improved performance. The modal number of correct responses is zero; the maximum number of correct responses thus far is three out of six. This pattern is virtually indistinguishable from that seen for the unmarked condition reported above.

We do not wish to over-interpret this preliminary data. Nonetheless, the performance of children in this pilot indicates that using a want framing is no magic bullet. Using language that emphasizes that an unconventional desire might be temporary does not seem to radically improve children's performance.

## General Discussion

Much has been made of quite young children's ability to reason about preferences in several different circumstances. There is no denying that children's appreciation of others' mental lives is far richer than was once thought, but the data we present here underscore that developing a theory of mind is a complex and protracted process. In our samples, even older four-year-olds-who, the literature suggests, would pass traditional tests of false-belief understanding-rejected unconventional opinions as mistaken or silly, at least when those opinions were expressed using factual-sounding language.

Given that toddlers seem able to respond appropriately to desires they do not share, why would our substantially older preschoolers persist in rejecting unconventional preferences? The possibility remains that reasoning about enduring, temporally unbounded likes or dislikes might be more demanding than reasoning about a desire in a specific situation. Our data collected thus far, however, suggests that a very simple change to the framing of characters' desiresusing the stem "I want" instead of "I like"-- is not enough to render preschoolers accepting of unusual wants. It is premature to jump to conclusions, of course. A more explicit
change to the framing to emphasize the temporariness of the desire might have a greater impact.

It may also be more difficult for preschoolers to reason about discrepant preferences when the preferences are highly surprising. In the existing literature, children were asked to reason about comparably attractive options (Wellman \& Woolley, 1991). In these cases, children in fact probably like both alternatives but picked one over the other on that occasion. It might not be difficult to recognize that someone might prefer to go to the park and someone else to the beach on a given occasion. Children also succeed on relatively unfamiliar items ( Ma and $\mathrm{Xu}, 2011$ ), where they may have only a weak opinion of their own.

Moreover, children are highly attuned to statistical information, as work in a broad array of domains demonstrates (Gergely et al., 2007; Kushnir, Xu, \& Wellman, 2010; Ma \& Xu, 2011; Saffran, Aslin, \& Newport, 1996). Liking for ice cream is likely highly overlearned, perhaps so much so that children cannot conceive of someone disliking it. Such a sensitivity to statistical regularities could interact with a non-representational theory of mind, reinforcing the notion that some mental states are veridical copies of the world itself.

Importantly, adults show analogous difficulties with reasoning about preferences and opinions in some circumstances. Ross, Greene, \& House (1977) coined the term "false consensus effect" to describe the robust bias in adults to assume that more people will endorse a belief when the participant also endorses that belief. This is of course a more subtle phenomenon than outright rejecting opinions that do not concord with one's own. Nonetheless, many of the explanations that have been offered to explain the false consensus effect in adults may shed light on children's behavior. For example, adults using anchoring and adjustment heuristics, whereby a person makes an initial prediction about others' behavior based on one's own behavior and then adjusts for others' idiosyncrasies, tend to systematically under-adjust (Epley \& Gilovich, 2006). If children are prone to egocentrism, the failure to adjust sufficiently for differences among people might be especially pronounced.

Likewise, motivational effects have been posited to explain adults' false consensus effects: adults simply want to be in the mainstream, and assuming others agree with them makes them feel good (Marks \& Miller, 1987). The idea of social norms is becoming especially salient in the preschool years (Nucci \& Turiel, 1978), and children may well view liking the right foods as one such norm. Indeed, liking the right foods is an important social signal for adults (e.g., eating caviar or eating french-fries). Not wanting the right foods may thus be a mistake in preschoolers' eyes in a social sense that differs from the way they apply that term to false facts.

These considerations illustrate the range of cognitive factors that underpin reasoning about preferences. Even as children's understanding of mental lives increases through the preschool years, many of the factors that lead to naïve realism in adults render children's understanding of unconventional preferences susceptible to the influence of linguistic framing. The robustness with which children
reject unconventional preferences provides a compelling demonstration of the challenges children face in learning to reason about the social world in an adult-like fashion.

## References

Banerjee, R., Yuill, N., Larson, C., Easton, K., Robinson, E., \& Rowley, M. (2007). Children's differentiation between beliefs about matters of fact and matters of opinion. Developmental Psychology, 43(5), 1084-96.
Epley, N., \& Gilovich, T. (2006). The anchoring-andadjustment heuristic: Why the adjustments are insufficient. Psychological Science, 17(4), 311-318.
Gergely, G., Egyed, K., \& Király, I. (2007). On pedagogy. Developmental Science, 10(1), 139-46.
Kushnir, T., Xu, F., \& Wellman, H. M. (2010). Young children use statistical sampling to infer the preferences of other people. Psychological science, 21(8), 1134-40.
Ma, L., \& Xu, F. (2011). Young children's use of statistical sampling evidence to infer the subjectivity of preferences. Cognition, 120(3), 403-11.
Marks, G., \& Miller, N. (1987). Ten years of research on the false-consensus effect: An empirical and theoretical review. Psychological Bulletin, 102(1), 72-90.
Nucci, L. P., \& Turiel, E. (1978). Social interactions and the development of social concepts in preschool children. Child Development, 49(2), 400-407.
Repacholi, B. M., \& Gopnik, A. (1997). Early reasoning about desires: Evidence from 14- and 18 -month-olds. Developmental Psychology, 33(1), 12-21. doi: 10.1037/0012-1649.33.1.12

Robinson, R. J., Keltner, D., Ward, A., \& Ross, L. (1995). Actual versus assumed differences in construal: "Naive realism" in intergroup perception and conflict. Journal of Personality and Social Psychology, 68(3), 404-417.
Ross, L., Greene, D., \& House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. Journal of Experimental Social Psychology, 13(3), 279-301.
Saffran, J. R., Aslin, R. N., \& Newport, E. L. (1996). Statistical learning by 8 -month-old infants. Science, 274(5294), 1926-1928.
Wellman, H. M., Cross, D., \& Watson, J. (2001). Metaanalysis of theory-of-mind development: The truth about false belief. Child Development, 72(3), 655-84.
Wellman, H. M., \& Woolley, J. D. (1990). From simple desires to ordinary beliefs: The early development of everyday psychology. Cognition, 35, 245-275.

# Eye movement optimization in visual search 

Ryan M. Hope and Wayne D. Gray<br>Cognitive Science Department<br>Rensselaer Polytechnic Institute


#### Abstract

In the present study we investigated whether eye movements in visual search are optimized to reduce time on task. Subjects task was to find a target object in a large field of objects that differed based on shape, color, size and numeric label. The target specification was manipulated, directly influencing the average number of fixations it took subjects to find the target object. Although a microstrategy that allowed for parallel saccade programming and information processing was found to be more efficient in terms of time, a serial microstrategy where saccade programming always follows information processing was found to be the more prevalent microstrategy.


Keywords: visual search, eye movements, microstrategy, optimization, return saccades

## Introduction

Visual search may well be our most ubiquitous cognitive task. Many (dozens? hundreds?) times a day we scan our desk for books or memos, our fridge for eggs or beer, the streets for oncoming traffic, cable television for shows we want to watch, and crowded rooms for faces we recognize. Although our natural scan environments are seldom completely novel, the objects in them and their places in these environments are seldom constant.

How do we do such searches? Most of them have the flavor of being a "one off" on at least one of several dimensions. Even the clutter on top of my dresser varies, if only slightly, from day to day. Do we develop optimal search strategies for each environment? - My fridge? Driving on I-87? Looking for friends to sit next to during Cognitive Science talks in Sage 4101? This strikes us as a likely possibility, meaning that there may be no general high level strategy for scanning my dresser top, the fridge, and I-87. But what about lower level strategies? What about the process of moving our eyes to a location, fixating that location, processing the perceptual and semantic information at that location, and deciding whether that location contains the target of our search, or whether we need to saccade to the next location? Can this strategy be optimized? Can it be optimized for all or many search environments? What would such an optimization strategy optimize?

Although in such a small paper it is obvious that we cannot address most of these issues, we believe we have a good start on addressing the later; namely, what would an optimized fixation and saccade strategy look like - this is the subject of our paper.

## Background

The key characteristic of visual search for an active vision task (Findlay \& Gilchrist, 2003), is one of moving the eyes from one location to another until we find our search target. But what does this really entail and can we bring to bear strategies in this task that cannot be applied to the simpler case of being tachistoscopically presented with a single item and asked whether that is our search target or not?

The case of a serial sequence of single items, each of which demands a "yes" or "no" answer before we are shown the next item, seems to define a procedure in which we perceptually process the visual object, semantically process the results of that perception, and decide whether the current object is a member of the target set defined for us by the experimenter. This strategy or procedure seems well suited for the given task environment and maybe defines the optimal strategy in this environment. Indeed, this procedure could be applied when I am searching my dresser top or searching for the large-red-star in Figure 1. Indeed, vision researchers often describe searches that entail the visual scanning of a busy screen in exactly these terms (e.g., Deubel \& Schneider, 1996; Henderson, 1992). From this point forward we will refer to this strategy as the serial microstrategy.

An alternative strategy differs from the serial microstrategy only in that the programming and subsequent execution of the next eye movement does not wait for the semantic processing and decision processing to complete. Programming of the next eye movement could start as soon as the previous saccade is completed, in parallel with information processing and decision making processes. From this point forward we will refer to this strategy as the parallel microstrategy.

Evidence for the parallel strategy comes from a number of visual search studies (Engel, 1977; Gould, 1973, Hooge \& Erkelens, 1996) where it was reported that subjects often fixated the target, made an eye movement away from the target, and then, on the next fixation, returned to the target (return saccades). Hooge and Erkelens (1996) also reported a number of missed targets. This evidence implies that the fixation durations can be too short to recognize the target. The occurrence of return saccades and missed targets suggest that saccade preparation may start before foveal processing is complete and that complete foveal processing is not necessarily the trigger for the subsequent saccade. For further details of both serial and parallel models of eye movements in visual search see Hornof and Halverson (2003), Hornof and Kieras (1997).

The parallel and serial microstrategies can also be differentiated based on their temporal costs. Since the parallel strategy allows for the concurrent processing of information and saccade preparation, fixation durations could be shorter when the parallel strategy is used than when the serial strategy is used because target identification can continue during the saccade. However, Becker and Jürgens (1979) showed that it is possible for saccades to be aborted during the preparation phase. This could lead to someone using the parallel strategy with apparent fixation durations closer to those observed in the serial strategy. On the other hand, since saccades in the parallel strategy do not necessarily wait for a decision of target presence, the parallel strategy could involve two extra saccades and fixations as a result of return saccades. Therefore in the extreme cases, the parallel strategy has a fixed cost of 2 saccades and 2 fixations whereas the serial strategy has a cost that grows with each fixation. Therefore there exists some threshold, in terms of number of fixations, where the parallel strategy will eventually become more efficient than the serial strategy.

The task used in the present study is a visual search task first used by Williams (1966). In the Williams search task, subjects have to find a target object in a very large field of objects that differ by size, shape, color and a numbered label (e.g., " 11 ", " 25 ", etc). Williams found that when he manipulated which target features were known (e.g., "large blue circle" versus "small yellow" versus "triangle"), a high proportion of fixations were on objects of the specified color and only a moderate proportion of fixations were on objects of the specified size or shape. When two or more characteristics were specified, fixations were generally based on a single characteristic. Additionally, the average number of fixations required to find the target differed based on the target specifications. This aspect of the task makes it perfect for eliciting the use of different microstrategies as the optimal strategy should depend on target specifications.

It seems likely that at least some subjects on at least some trials used a mix of serial and parallel microstrategies. However, the parallel and the serial strategies predict the same pattern of saccades for all but the last two saccades to and from target object. Hence, for this initial report, we make the simplifying assumption that if we find a return saccade, that we can classify the entire trial as having been accomplished using the parallel strategy. Likewise, the absence of return saccades were used to classify trials as having been accomplished using a serial strategy. Based on this classification scheme we had the following hypotheses:

1. Due to the temporal costs of the serial microstrategy with increasing number of fixations, the proportion of trials with return saccades (indicating the use of the parallel microstrategy) should be higher on trials where search is inefficient than it is on trials where search is efficient. Here efficiency corresponds to the number of fixations need to find a target, a function of the target specification or number of cue features.
2. The average fixation duration (all fixations on a given
trial) should be shorter on trials that exhibit return saccades (indicating the use of the parallel microstrategy) than on trials that do not exhibit return saccades.
3. For trials that do not contain return saccades, fixation durations should increase as a function of number of cue features due to increasing processing requirements.
4. Subjects will satisfice by using the microstrategy that results in the most time savings across all trial types.

## Method

## Subjects

Subjects were 15 undergraduate students at Rensselaer Polytechnic Institute ( 10 men, 5 women) who were given course credit for their participation. Subjects were prescreened for their dependence on eyeglasses or contact lenses; only subjects that reported needing neither were allowed to participate in the experiment.

## Apparatus

The experiment was displayed on a 22 " Dell widescreen LCD with a resolution of $1680 \times 1050$ (pixels) and physical dimensions of $473.76 \times 296.1(\mathrm{~mm})$. Eyetracking was performed with a SensoMotoric Instruments RED500 eyetracker running at a sample rate of 500 Hz . On average, subjects were positioned $700 \pm 100 \mathrm{~mm}$ away from the LCD.

## Task

Subjects task on each trial was to find a target in a field of 48 randomly dispersed objects. Each object had a unique combination of shape (4 levels), color (4 levels) and size (3 levels). On each shape was a randomly assigned numeric id which ranged from 01 to 48 . Each trial starts with the search objects masked and a cue at the center of the display. The cue was a text description of the target object. The probe always contained the numeric id and up to three other features (shape, size and color). The particular features shown in the cue, in addition to the numeric id, was systematically manipulated throughout the experiment such that each subject experienced one trial of each object and cue combination. The non-id cue features were ordered randomly with the id always showing last (see Figure 1 for an example of the cue). The 48 objects and 8 cue combinations results in 384 unique trials. Because of the random dispersion of objects on each trial, no subjects experienced identical trials. Subjects were instructed to study the cue until they felt they had memorized it, at which point, they were instructed to press spacebar on the keyboard to reveal the search field and begin searching for the target. The cue remained on screen during the search phase. Once subjects find the target, they end the trial by using the mouse to click on the target. Search time was measured as the time between the spacebar press and the first correct click on the target object. Subjects were given no explicit instruction to emphasize speed, their task was simply to find the target.


Figure 1. An example of a trial search display that contains an end of trial return saccade. In this trial, the cue is "large 18 ". The target is the large yellow oval in the bottom right corner. Eye gaze data is overlaid; red dots correspond to samples that belong to saccades, black dots correspond to samples that belong to fixations. Circles with a 1 degree visual angle radius have been drawn around the center of mass for each fixation. As indicated by the grid formed by the x and y axes, the scanpath for this Ss is: e4, f4, h5, h5, f7, e8, c7, b9, i8, h2, f2, i2. Notice how the last 3 fixations include 2 fixations on the target object separated in time by a fixation on a different object. Neither the grid nor the light gray circles around each object were visible during the trial.

## Stimuli

The four shapes used in the task were star, oval, crescent and cross. The four colors used in the task were red, yellow, green and blue and had hue values (in HSV space) of 0,72 , 144 and 216 respectively. The saturation and value of all four colors was set to 50 and 100 respectively. The search field was a $1050 \times 1050$ (pixel) square centered on the screen and had $65 \%$ gray background color. See Figure 1 for an example of the shapes and colors used in this task. The three object sizes were small ( 48 pixels), medium (119 pixels), and large (191 pixels) and correspond to visual angles of 1.1, 2.7 and 4.4 degrees with a potential error up to $15 \%$ depending on head position.

## Gaze Data Classification

Raw gaze data were classified into events (saccades and fixations) by an algorithm that uses both velocity and acceleration thresholds. The algorithm followed the following general procedure:

1. Convert the $x, y$ screen coordinates from pixels to degrees of visual angle relative to the center of the screen.
2. Compute smoothed first and second order derivatives of the $\mathrm{x}, \mathrm{y}$ visual angle components.
3. Label gaze samples with corresponding velocity and acceleration components that both exceed their respective thresholds as saccades, the rest are labeled as fixations.
4. Identify saccades that last for less than 20 ms , reclassify them as fixations.

Velocity and acceleration (first and second order derivites) were calculated using a Savitzky-Golay filter (as recommended by Nyström and Holmqvist (2010)) for its ability to preserve local minima and maxima. The Savitzky-Golay filter used was a second order filter with a window length of 11 samples which allowed for accurate detection of saccades down to 20 ms in duration (Nyström \& Holmqvist, 2010). The velocity and acceleration thresholds used in the present study were $30^{\circ} / \mathrm{s}$ and $8000^{\circ} / \mathrm{s}^{2}$ respectively and were based on the "cognitive configuration" of the EyeLink software (SR Research Ltd. 2007).

## Results

## Number of Fixations

Our first hypothesis stated that the proportion of return saccades should increase with number of fixations and that number of fixations was directly related to the cue specification (trial type) and number of cues. Therefore, in order to evaluate our first hypothesis we first need to show that trial type has an impact on number of fixations. In order to accomplish this we performed one-way repeated measures analysis of variance. The analysis of variance revealed a significant effect of trial type on proportion of return saccades, $F(7,98)=76.56$, $p<.001, \eta_{g}{ }^{2}=0.73$. The means and standard errors can be seen in Figure 2. The general trend revealed from this analysis is that the number of fixations decreased as the number of cue features increased. We performed a second analysis of variance to test the effect of number of cues on number of fixations. As expected by hypothesis 3 , the greater the number of cues, the fewer fixations, $F(3,42)=96.62, p<.001$, $\eta_{g}{ }^{2}=0.76$. The effect size of number of cue was stronger than the effect of trial type on number of fixations.

## Return Saccades

Return saccades were identified through scanpath analysis. First, the search field was divided into a $9 \times 9$ array which resulted in 81 total cells. The width and height of each cell was approximately 2.6 degrees of visual angle. Second, fixations were recoded as belonging to one of the 81 cells based on their center of mass. After recoding, consecutive fixations occurring in the same cell were combined. This effectively removes any microsaccades and other small amplitude corrective saccades. Finally, for scanpaths of length 3 or greater, the last fixation was compared to the fixation 2 back. In order for a trial to be classified as containing a return saccade two


Figure 2. The average number of fixation for each trial type. Error bars represent standard error. The dashed line represents the threshold where the parallel strategy becomes more efficient than the serial strategy, determined empirically using Equation 5


Figure 3. The average proportion of return saccades for each trial type. Error bars represent standard error.
criterion had to be met. First, the fixation 2 back from the final fixation had to be in the same cell, or any of the surrounding cells, as the last fixation. Second, the fixation 1 back from the last fixation could not be on the center cell where the cue was located (see center of Figure 11). Of the 5939 total trials, 2273 trials contained a return saccade.

In order to test the effect of trial type on the proportion
of return saccades a one-way repeated measures analysis of variance was performed. The analysis of variance revealed a significant effect of trial type on proportion of return saccades, $F(7,98)=12.62, p<.001, \eta_{g}{ }^{2}=0.33$. The means and standard errors can be seen in Figure 3 . The general trend revealed from this analysis is that the proportion of return saccades decreased as the number of cue features increased. In order to confirm this trend we performed a second oneway repeated measures analysis of variance to test the effect of number of cue features on the proportion of return saccades. The analysis of variance revealed a significant effect of number of cue features on proportion of return saccades, $F(3,42)=25.57, p<.001, \eta_{g}{ }^{2}=0.46$. Similar to the analysis of variance involving number of fixations, the effect size of number of cues was stronger than the effect of trial type for proportion of return saccades. In addition, the correlation between proportion of return saccades and number of fixations with respect to trial type, which can be seen by comparing Figure 2 and Figure 3 , is strong; $r=.90, n=8, p=.002$.

## Fixation Durations

In order to test the effect of return saccades and number of cue features on average fixation duration (all fixations within a trial) we performed a $2 \times 4$ (return saccade by number of cue features) repeated measures analysis of variance. The analysis of variance revealed a significant main effect of return saccade, $F(1,14)=27.44, p=.001, \eta_{g}{ }^{2}=0.22$; a significant main effect of number of cues, $F(3,42)=6.68$, $p<.001, \eta_{g}{ }^{2}=0.11$; and a marginally significant interaction effect $F(3,42)=2.62, p=.063, \eta_{g}{ }^{2}=0.02$. The means and standard errors can be seen in Figure 4


Figure 4. The average fixation duration (all fixations within a trial) for each level of number of cue features. Error bars represent standard error.

## Comparison of Search Strategies

The serial strategy search time $T_{\text {serial }}$ can be approximated by the following equation:

$$
\begin{equation*}
T_{\text {serial }}=(N-2) *\left(F_{\text {serial }}+S\right) \tag{1}
\end{equation*}
$$

where N is the average number of fixations on a given trial type, and $S$ is the average saccade duration. Similarly, the parallel strategy search time $T_{\text {parallel }}$ can be approximated by the following formula:

$$
\begin{equation*}
T_{\text {parallel }}=N *\left(F_{\text {parallel }}+S\right) \tag{2}
\end{equation*}
$$

In order to approximate the values for $F_{\text {serial }}$ and $F_{\text {parallel }}$, we first performed mixed effects regression on mean fixation duration with return saccade and number of cues as fixed factors and with subjects as a random factor. The regression yielded the following equation (rounded to the nearest millisecond):

$$
\begin{equation*}
F=213-E * 24+C * 5 \tag{3}
\end{equation*}
$$

where F is fixation duration, E is 1 for return saccade trials ( 0 for non-return saccade trials) and C is the number of cues. The intercept value of 213 ms is consistent with previous research on average fixation durations in visual search (Rayner, 2009, Salthouse \& Ellis, 1980). The 24 ms difference between trials with and with out return saccades (assumed to be associated with the time required to make a decision of target presence) is psychologically plausible (Neisser, 1963; van Diepen, De Graef, \& D'Ydewalle, 1995). In addition, the 5 ms per cue feature seems psychologically plausible. Using Equation 3 we can compute the average fixation duration, $F_{\text {serial }}$ and $F_{\text {parallel }}$, for a given trial type and search strategy by using the average number of fixations, N, as shown in Figure 2 The value of S was set to 45 ms in all computations based on the average of saccade durations in the empirical data.

The time savings (or loss) from using the parallel search strategy can then be computed as follows:

$$
\begin{equation*}
T_{\text {diff }}=T_{\text {serial }}-T_{\text {parallel }} \tag{4}
\end{equation*}
$$

The results of Equation 1 and Equation 2 applied to each of the 8 probe combinations applied to Equation 4 is shown in Figure 5. The parallel search strategy saved more time than the serial search strategy for all trials where color was not an available cue feature. The sum of all $T_{d i f f}$ values was 972.02 ms indicating that overall, the parallel strategy could be more efficient.

This analysis can be taken one step further by setting $F_{\text {serial }}$ equal to $F_{\text {parallel }}$ and solve for N , resulting in Equation 5 , to find the threshold in terms of number of fixations where the parallel strategy becomes more efficient than the serial strategy.

$$
\begin{equation*}
N=\frac{2 * F_{\text {serial }}+2 * S}{F_{\text {serial }}+F_{\text {parallel }}} \tag{5}
\end{equation*}
$$

By using a value of 213 for $F_{\text {serial }}$ and 189 for $F_{\text {parallel }}$ (computed from Equation 3) and a value of 45 for $S$, the threshold turns out to be 21.5 fixations. This threshold is depicted as the dashed horizontal line in Figure 2 Interestingly, for all trials in which color is available as a probe feature, the average number of fixations is less than this threshold. In other words, when color is not available, the parallel search strategy will be more efficient in terms of time.

## Discussion

The goal of this study was to determine if people optimize their eye movements during visual search. In order to accomplish this goal we had subjects perform a difficult visual search task where they had to find a target object in a field of objects that differed in size, shape color and numeric label. We hypothesized that subjects would optimized their eye movements by using the more efficient of two microstrategies: the serial and parallel microstrategies. We used the presence of return saccades as a marker of the parallel microstrategy. Consistent with our first hypothesis, the proportion of trials that contained return saccades were higher on trials that required lots of fixations to find the target as well as on trials where there were few target cues compared to trials that required only a few fixations to find the target or trials that had many target cues. One interpretation of this result is that subjects are indeed sensitive, consciously or unconsciously, to the temporal costs of serial strategy which ensures a target presence decision before the eye is moved to the next location.

Our second hypothesis predicted that fixation durations would be shorter on trials that contained return saccades that on trials that did not contain return saccades. This indeed was the case. We also predicted in our third hypothesis that on trials that did not include a return saccade that fixation durations should increase with respect to the number of known target features. This prediction was mildly supported by the marginally significant interaction we found in our analysis of variance involving return saccades and number of cue features on fixation duration.

Our conservative cost analysis of the serial and parallel microstrategies does show that the parallel microstrategy can more cost effective in terms of time on task. However, return saccades were only observed on $38 \%$ of the trials in our experiment. This could be interpreted as evidence that subjects did not or could not optimize their eye movements. However, we are not positive that is the case. It's possible that return saccades are just not a good enough measure of the parallel search strategy since it is possible for early eye movement programming (that would have cause the eye to move before target analysis) to be aborted in the parallel strategy. This could not only potentially reduce the number of observed return saccades but it could also affect the average fixation durations in trials with and with out return saccades. It is possible that someone could be using the parallel strategy throughout most of a trial but then abort an eye movement that would


Figure 5. The estimated difference ( $T_{d i f f}$ ) in time from using the parallel or serial microstrategy, computed using Equation 4 Positive values favor the parallel strategy, negative values favor the serial strategy.
have resulted in a return saccade. This would result in the lowering of the average fixation duration for trials with out a return saccade due to improper strategy classification. Additionally, a trial that shows a return saccade could still have contained many aborted early saccades inflating the average fixation duration of return saccade trials.

## Conclusion

Our study has provided evidence that eye movements in visual search can be sensitive to millisecond level cost-benefit trade-offs. Whether or not people can actually optimize their eye movements to take advantage of these cost-benefit tradeoffs is still not clear. In addition, the fact that the parallel strategy could under some circumstances appear as if it were actually a serial strategy allows for the possibility that the serial strategy does not even exist. This idea is consistent with the findings of Hornof and Kieras (1997) and Hornof and Halverson (2003). Further research on this topic will need to find better ways to quantify the prevalence of parallel processing in eye movement microstrategies.

## Acknowledgments

The work was supported, in part, by grants N000140910402 and N000141310252 to Wayne Gray
from the Office of Naval Research, Dr. Ray Perez, Project Officer.

## References

Becker, W. \& Jürgens, R. (1979). An analysis of the saccadic system by means of double step stimuli. Vision research.
Deubel, H. \& Schneider, W. X. (1996). Saccade target selection and object recognition: evidence for a common attentional mechanism. Vision research, 36(12), 1827-37.
Engel, F. (1977). Visual conspicuity, visual search and fixation tendencies of the eye. Vision research, (June).
Findlay, J. M. \& Gilchrist, I. D. (2003). Active vision. Oxford Univ. Press.
Gould, J. D. (1973). Eye movements during visual search and memory search. Journal of experimental psychology, 98(1), 184-95.
Henderson, J. M. (1992). Visual attention and eye movement control during reading and picture viewing. In K. Rayner (Ed.), Eye movements and visual cognition: scene perception and reading (pp. 260-283). New York: Springer.
Hooge, I. T. C. \& Erkelens, C. J. (1996). Control of fixation duration in a simple search task. Attention, Perception, $\mathcal{E}$ Psychophysics, 58(7), 969-976.
Hornof, A. J. \& Halverson, T. (2003). Cognitive strategies and eye movements for searching hierarchical computer displays. In Proceedings of ACM CHI 2003: Conference on Human Factors in Computing Systems. New York, NY: ACM.
Hornof, A. J. \& Kieras, D. E. (1997). Cognitive modeling reveals menu search is both random and systematic. In Acm chi'97 conference on human factors in computing systems (pp. 107-114). ACM.
Neisser, U. (1963). Decision-Time without Reaction-Time: Experiments in Visual Scanning. The American Journal of Psychology, 76(3), 376-385.
Nyström, M. \& Holmqvist, K. (2010). An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data. Behavior Research Methods.
Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. The Quarterly Journal of Experimental Psychology, 62(8), 1457-1506.
Salthouse, T. \& Ellis, C. (1980). Determinants of eye-fixation duration. The American journal of psychology.
SR Research Ltd. (2007). EyeLink 1000 User Manual (9/13/2007).
van Diepen, P. M., De Graef, P., \& D'Ydewalle, G. (1995). Chronometry of foveal and peripheral information encoding during scene perception. In J. M. Findlay, R. Walker, \& R. W. Kentridge (Eds.), Eye movement research: mechanisms, processes and applications (pp. 349-362). Amsterdam: North Holland.
Williams, L. G. (1966). The effect of target specification on objects fixated during visual search. Perception and Psychophysics, 1(5), 315-318.

# More than a feeling: When emotional reactions don't predict moral judgments 

Zachary Horne (horne2@illinois.edu)<br>Department of Philosophy, 810 S. Wright Street<br>Urbana, IL 61801 USA<br>Derek Powell (derekpowell@ucla.edu)<br>Department of Psychology, 1285 Franz Hall, Los Angeles, CA 90095 USA


#### Abstract

Many moral psychologists have proposed that the difference between people's moral judgments about the Trolley and Footbridge dilemmas can be explained by their differing emotional responses to the dilemmas. In two experiments, we tested this explanation by presenting the dilemmas and measuring participants' reactions using a self-report emotion measure (PANAS-X). As might be expected, participants experienced more intense emotions after reading moral dilemmas when compared to a non-moral dilemma. However, participants' emotional reactions to the Trolley and Footbridge dilemmas did not differ. Our findings call the oft cited emotion explanation into question.


Keywords: moral psychology; emotion; PANAS-X; decision making.

Recently, research on moral dilemmas has crossed disciplines, branching out from ethics into psychology and cognitive neuroscience (e.g., Ciaramelli, Muccioli, Làdavas, \& Di Pellegrino, 2007; Greene, Sommerville, Nystrom, Darley, \& Cohen, 2001; Greene, Nystrom, Engell, Darley, \& Cohen, 2004; Koenigs, Young, Adolphs, Tranel, Cushman, Hauser, \& Damasio, 2007). Within these lines of inquiry, there has been widespread interest in participants' judgments about two famous moral dilemmas: the Trolley dilemma and the Footbridge dilemma. In the Trolley dilemma, participants are told they are at the wheel of a runaway train that will cause the deaths of five workmen if it proceeds on its present course. The only way to avoid the deaths of these workmen is to hit a switch, redirecting the train to a side track where it will kill a single workman instead. This dilemma is often contrasted with another famous moral dilemma, the Footbridge dilemma, which asks participants to consider a situation in which they are on a footbridge, in between a runaway trolley and five workmen who will be killed if nothing is done. Participants are then told that the only way to save the lives of the five workmen is to push a stranger off the bridge and onto the tracks below where his large body will stop the trolley. Surprisingly, in the Trolley dilemma, approximately $80 \%$ of people judge that it is appropriate to take action (hit the switch, killing one person), yet in the Footbridge dilemma, about $80 \%$ judge that is it inappropriate to take action (push the man off the bridge, killing him). Given the apparent similarity of these moral situations, researchers have sought
to understand why people give such drastically different judgments about these cases.

Many researchers have suggested that these two dilemmas can be distinguished by the emotional reactions they elicit. Specifically, the Footbridge dilemma is thought to elicit strong negative emotional reactions, whereas the Trolley dilemma is thought to elicit weak negative emotional reactions. We will refer to this claim as the Emotion Explanation.

The Emotion Explanation has been interpreted as having wide-ranging implications for theories in psychology, cognitive neuroscience, and ethics. For instance, some psychologists have suggested that the Emotion Explanation is revealing of the psychological mechanisms recruited in moral judgment in general (e.g., Greene et al., 2001; Koenigs et al., 2007). Meanwhile, philosophers have employed the Emotion Explanation to diverse ends, using it to advance or undercut normative ethical theories, and to make arguments about the epistemic status of moral intuitions (e.g., Singer, 2005).

There are several reasons to take this explanation seriously. First, there is good evidence that emotions do importantly influence moral judgments (Haidt, 2001). Most directly, cognitive neuroscientists using functional Magnetic Resonance Imaging (fMRI) have reported finding activity in areas of the brain associated with emotional processing while participants made 'deontological' moral judgments, such as judging that it is inappropriate to push the man in the Footbridge dilemma (Greene et al., 2001, 2004). Researchers have also examined the moral judgments of participants with damage to the ventromedial prefrontal cortex (vmPFC), an area of the brain associated with affect. When given a battery of moral dilemmas, these participants gave 'utilitarian' moral judgments in the Footbridge dilemma, whereas healthy control participants tended to give deontological judgments (Koenigs et al., 2007).

Despite this evidence, there are a number of reasons to examine the Emotion Explanation more closely. First, it is not clear how the Emotion Explanation coheres with existing theories about morally relevant emotions, such as guilt. For example, social functionalist theories of emotion suggest that emotions like guilt serve both intrapersonal and interpersonal functions (Keltner, Haidt, \& Shiota, 2006). On this view, guilt is typically elicited as a result of one's perceived transgressions against another, which can damage
relationships (Keltner, 1995). Upon contemplating some action, the experience of guilt might serve an intrapersonal function, deterring the potentially damaging action. Alternatively, if the action is ultimately taken, guilt can serve an interpersonal function, helping to repair whatever damage was done (Baumeister, 1994).

While it is clear that moral dilemmas can differ in the intensity of the emotions they evoke (e.g., dilemmas that ask participants to consider killing their child versus lying), the emotionally salient aspects of both the Trolley and Footbridge dilemmas are held constant. In both scenarios, participants are asked if they should kill a stranger in order to save five other strangers. Causing the death of another person seems likely to place a strain on interpersonal relationships, whatever one's motivation for doing so. Cast in this light, social functionalist theories of emotion seem to predict that people will experience feelings of guilt even as they approve of taking action in the Trolley dilemma.

Second, extant research in moral psychology provides only indirect support for the Emotion Explanation. The most direct support comes from fMRI investigations by Greene and colleagues (2001, 2004). In these studies, researchers presented participants with two large groups of moral dilemmas. One group was composed of the Trolley dilemma and 18 other dilemmas, and the other group was composed of the Footbridge dilemma and 24 other dilemmas. They called the dilemmas they considered Trolley-like 'impersonal', and the dilemmas they considered Footbridgelike 'personal', due to the 'closeness' of the action being performed in the dilemma. They found that the 'impersonal' dilemmas activated areas of the brain associated with deliberative cognition, whereas the 'personal' dilemmas activated areas of the brain associated with emotion. They regarded these results, in part, as evidence for the Emotion Explanation.

However, in order for the data from the studies by Greene and colleagues $(2001,2004)$ to provide compelling evidence for the Emotion Explanation per se, each of their 'impersonal' and 'personal' dilemmas would have to elicit emotions with valence and intensity similar to the Trolley and Footbridge dilemmas, respectively. It is doubtful that such parity was achieved. Dilemmas deemed 'impersonal' frequently did not involve physically harming someone (even though the Trolley dilemma clearly does), whereas those deemed 'personal' almost invariably did. In fact, only 10 of the 19 'impersonal' dilemmas involved harm, whereas 24 of the 25 'personal' dilemmas involved physically harming someone (Moore, Clark, \& Kane, 2008). This difference is a problematic confound because moral situations concerning the bodily harm of another have been shown to elicit stronger negative emotional responses than those not involving any bodily harm (Heekeren, 2005). Misgivings about the representativeness of the dilemmas in these sets are further supported by an item analysis performed by McGuire and colleagues (2009) on Greene and colleagues' data. This analysis demonstrated that only a subset of the moral dilemmas was responsible for the
significantly different patterns of neural activation Greene and colleagues observed, again impugning evidence for the Emotion Explanation.

Given the confound in these materials, and the findings of the item analysis (McGuire, Langdon, Coltheart, \& Mackenzie, 2009), we suspect that people experienced more intense emotions when considering 'personal' dilemmas because the dilemmas much more frequently involved physical harm, not because the Footbridge dilemma itself is more emotionally engaging than the Trolley dilemma. ${ }^{1}$ In other words, Greene and colleagues' (2001) findings do not provide clear evidence for the Emotion Explanation, and thus corroboration of the Emotion Explanation requires a more direct test. These criticisms apply equally to a number of other studies by researchers who have used these same materials and have claimed to have found evidence for the Emotion Explanation (e.g., Ciaramelli et al., 2011; Koenigs et al., 2007). To our knowledge, there is no empirical data taken as evidence for the Emotion Explanation that avoids this criticism.

We sought to test the Emotion Explanation by examining people's emotional responses to the Trolley dilemma and Footbridge dilemma, individually and specifically. We measured participants' emotional responses to moral dilemmas using the PANAS-X, a comprehensive emotional state, trait, and mood self-report measure (Watson, Clark, \& Tellegen, 1988) that has been shown to correlate with neural activation in the amygdala (Irwin, Davidson, Kalin, Sorenson, \& Turski, 1998), as well as in the vmPFC (Zald, Mattson, \& Pardo, 2002). Importantly, use of this self-report measure allowed us to investigate people's emotional responses to individual moral dilemmas, rather than to a battery of different dilemmas. More advanced methodologies such as GSR or fMRI, while in many ways superior to self-report, are ill-suited for investigating responses to individual stimuli.

In line with previous research, we hypothesized that considering the Footbridge dilemma would elicit negative emotions. In contrast to the Emotion Explanation, we predicted that considering the Trolley Dilemma would also elicit increased guilt, as well as other negative emotions. Moreover, because the Trolley and Footbridge dilemmas contain very similar emotionally-relevant content (they both call on participants to imagine killing another person), we expected to find very little difference in people's emotional reactions to these dilemmas. These last two predictions stand in contrast to the way moral psychologists have conceived of the Footbridge and Trolley dilemmas (e.g., Ciarmelli et al., 2007; Greene et al., 2001; Greene et al.,

[^93]2004; Koenigs et al., 2007); hence support for our predictions would cast doubt on the Emotion Explanation.

## Experiment 1

## Method

Participants The participants in Experiment 1 were 442 students enrolled in various undergraduate courses at Arizona State University. Approximately 54\% of the participants were males. The mean age of participants was 20.1 years old.

Materials Four different scenarios were presented to participants in this study, with each participant reading one scenario. Three of these scenarios were moral vignettes, the 'Trolley,' the 'Footbridge,' and the 'Crying Baby' vignette, all taken verbatim from Greene et al. (2001).

One potential concern is that the PANAS-X subscales might not be sensitive enough to detect differences between moral dilemmas of different emotional intensity. To evaluate the sensitivity of our chosen PANAS-X subscales, we presented some participants with a highly emotional dilemma, the Crying Baby dilemma. In this dilemma, participants must consider smothering their own infant child to save the lives of their townspeople.

The Crying Baby dilemma also afforded us the opportunity to explore whether there was a relationship between emotions and moral judgments within an emotionally engaging case. Prior research shows that people tend to be evenly divided over whether or not it is appropriate to take action in the Crying Baby dilemma, permitting meaningful comparisons between those who approve and those who disapprove. The fourth dilemma was a non-moral control dilemma (Coupon) adapted from Greene et al. (2001).

In Experiment 1, we used selected sub-scales from the PANAS-X to measure the extent to which participants experienced several relevant emotions. Participants rated how strongly they felt certain feelings on a 1 to 5 Likert scale, where 1 corresponded to 'Very slightly or not at all' and 5 corresponded to 'Extremely.' Importantly, participants were asked to report on their current emotional state rather than to assess the vignette or to speculate as to how they might feel if placed in the situation they read. These individual ratings were then averaged to obtain a score for each emotion subscale. Data were collected for the Positive Affect, Negative Affect, Hostility, Guilt, Joviality and Self-Assurance sub-scales from the PANASX. The Guilt sub-scale included words like 'Guilty', 'Ashamed', and 'Disgusted With Self', and the Hostility subscale included words like 'Angry', 'Disgusted', and 'Hostile'. The Joviality subscale included words like 'Happy', 'Joyful', and 'Cheerful', and the Self-Assurance subscale included words like 'Proud', 'Strong', and 'Confident'. As the Positive and Negative Affect scales contain many of the same emotion words found in the Guilt, Joviality, Hostility and

Self-Assurance subscales, we do not discuss analyses of the Positive and Negative Affect scales.

Procedure Experiment 1 was administered and data were collected via pen-and-paper questionnaires. The questionnaires consisted of text describing a moral dilemma, followed by an emotion measure (i.e., the PANAS-X). Participants were asked to respond to the emotion measure after considering a statement of the form, 'You are thinking about (action) in order to (outcome).' For example, when considering the Trolley dilemma, participants would read the statement, 'You are thinking about hitting the switch in order to avoid the deaths of the five workmen.' The text indicating the outcome was italicized for all conditions. On the opposite side of the questionnaire participants were asked to make a judgment of appropriateness. In making their judgments of appropriateness, participants were asked, 'Is it appropriate for you to (action) in order to (outcome).' They were instructed to circle 'yes' or 'no.' Participants were also asked to describe the emotions they experienced while reading the vignettes, and were given several lines to write in their response. These descriptions were used to identify problematic responses or inconsistencies between participants' reported emotions on the PANAS-X. Finally, participants were asked to provide demographic information.

Experiment 1 was conducted during various lecture classes at Arizona State University. The questionnaires were passed out at the beginning of the class, and students were given approximately 10 minutes to complete them. Instructions were given both verbally and in writing. Participants were instructed not to talk to each other while completing the survey, and were observed for compliance.

## Results

Of the 442 participants originally involved in the experiment, 12 gave written descriptions of their emotions that were primarily non sequitur or conflicted with the emotion ratings they had given on the PANAS-X. These 12 participants were removed from further analyses. Only participants who had complete data sets for all the emotion subscales were included in the final analysis. This resulted in 86 participants responding to the control vignette, 147 to the Crying Baby, 71 to the Footbridge, and 77 to the

Table 1: Participants' judgments across conditions (percent approval).

| Control | Trolley | Footbridge | Crying Baby |
| :---: | :---: | :---: | :---: |
| $92 \%$ | $91 \%$ | $25 \%$ | $46 \%$ |



Figure 1: Mean emotion ratings on each PANAS-X sub-scale for the four dilemmas in Experiment 1.

Trolley. Additional participants were included in the Crying Baby in order to achieve adequate statistical power for comparisons to be made within that condition. Mean subscale ratings for each vignette are shown in Figure 1.

Participants' judgments are summarized in table 1. Their responses accord with those obtained by previous investigations of these scenarios.

PANAS-X scores were examined with four ANOVA tests with vignette type as a between-subjects variable. Significance was obtained on the $\operatorname{Guilt}(F(3,377)=62.57$, MSE $=1.334, p<.001)$, Hostility $(F(3,377)=65.66$, MSE $=.953, p<.001)$, Joviality $(F(3,377)=21.19$, MSE $=.453$, $p<.001)$ and Self-Assurance $(F(3,377)=2.74, \mathrm{MSE}=$ $.737, p=.04)$ subscales.

For each of these ANOVA tests, planned comparisons examined differences between the three moral vignettes and the control vignette in the emotion ratings on the Guilt, Hostility and Joviality subscales. The moral vignettes elicited significantly more Guilt $(t(377)=8.69, p<.001)$ and Hostility $(t(377)=7.90, p<.001$, and significantly less Joviality $(t(377)=-6.83, p<.001)$ than the control vignette. Though significance was obtained on the respective ANOVA analyses, no significant difference was found between the moral vignettes and control vignette for SelfAssurance ratings $(t(377)=-.475, p=.635)$, suggesting that the Self-Assurance subscale was not relevant to moral judgments. No further analyses were conducted for SelfAssurance ratings.

As predicted, participants' ratings of Guilt, Hostility and Joviality were compared between the Trolley and Footbridge vignettes with three separate t-tests. These tests found no significant differences between the two groups for Guilt $(t(146)=-.09, p=.93)$, Hostility $(t(148)=-.33, p=$ $.74)$ or Joviality $(t(148)=1.24, p=.22)$. This finding is consistent with our prediction that such highly similar moral

Table 2: Summary of Logistic Regression Models Predicting Moral Judgments in the Crying Baby Dilemma in Experiment 1.

|  | Model 1 |  |  |  | Model 2 |  |  |  |  |
| ---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $\beta$ | SE | $e^{\beta}$ | $p$ | $\beta$ | SE | $e^{\beta}$ | $p$ |  |
| Hostility | -.21 | .23 | .81 | .36 | -.22 | .24 | .81 | .37 |  |
| Guilt | -.06 | .19 | .94 | .76 | -.02 | .20 | .98 | .91 |  |
| Joviality | .21 | .40 | 1.24 | .59 | .41 | .41 | 1.20 | .66 |  |
| Gender |  |  |  |  | 1.19 | .35 | 3.27 | .001 |  |
| $\chi^{2}$ |  | 4.08 |  |  | 15.81 |  |  |  |  |
| df | 3 |  |  | 4 |  |  |  |  |  |
| $p$ |  | .253 |  |  | .003 |  |  |  |  |

scenarios were unlikely to elicit meaningfully different emotional reactions.

Participants who read the Trolley vignette reported greater levels of Guilt $(t(377)=4.66, p<.001)$ and Hostility $(t(377)=4.22, p<.001)$, and lower levels of Joviality $(t(377)=-5.39, p<.001)$ than those who read the control vignette. Further planned comparisons tested whether the Crying Baby vignette was more emotionally salient than the Footbridge and Trolley vignettes. The Crying Baby vignette elicited more Guilt $(t(377)=9.11, p<.001)$ and Hostility $(t(377)=10.23, p<.001)$, and less Joviality $(t(377)=-2.79, p<.001)$ than the Footbridge and Trolley vignettes. These differences indicate that the PANAS-X is sufficiently sensitive as an emotion measure to detect differences between moral dilemmas of different emotional intensity.

Finally, we used logistic regression analyses to assess the relationship between emotions and moral judgments in the Crying Baby dilemma. We evaluated this relationship using two different regression models. The resulting equations are summarized in Table 2. The first model uses the Guilt, Hostility, and Joviality subscale scores as predictors of people's moral judgments. No significant relationship was found between participants' emotion ratings and their moral judgments for the Crying Baby problem $\left(\chi^{2}(3)=4.082\right.$, $p=$ .25). The second model adds gender to the three regressors from the original model. This model significantly predicted moral judgments $\left(\chi^{2}(4)=15.813, p<.01\right)$. Within this model, the gender coefficient was statistically significant ( $\beta$ $=1.185, e^{\beta}=3.271, p<.01$ ), indicating that women were less likely to judge it appropriate to smother the child. Importantly, this relationship between gender and moral judgments was significant over and above the emotion ratings.

## Experiment 2

As the results of Experiment 1 stand in stark contrast to the Emotion Explanation, we sought to replicate the findings of Experiment 1 in a different population.

## Method

The participants in Experiment 2 were 221 workers recruited using Amazon Mechanical Turk (mTurk). Approximately $53 \%$ of the participants were males. The mean age of participants was approximately 31 years old.

Materials The materials and measurements used in Experiment 2 were identical to those of Experiment 1, with two exceptions. The results of Experiment 1 indicated that participants' emotion ratings on the Self-Assurance subscale in moral conditions did not differ from participants in the control condition. Consequently, in Experiment 2, we only used the Guilt, Joviality and Hostility subscales to measure participants emotional responses. In addition, rather than asking participants to give moral judgments as a binary 'yes/no' response, we asked them to rate the moral rightness or wrongness of the proposed act on a 1-6 Likert scale. The end points of the Likert scale were labeled 'Completely Inappropriate' and 'Completely Appropriate.' This change was made in order to increase our statistical power for detecting any potential relationship between emotions and moral judgments in the Crying Baby dilemma.

Collection Experiment 2 was administered and data was collected through the mTurk work-distribution website. Eligible workers were redirected to Qualtrics, where they completed the study. Afterwards, workers were directed back to mTurk, where they were compensated with $\$ .20$.

Procedure The survey presented to participants in Experiment 2 was nearly identical to the survey in Experiment 1, save for being computerized rather than pen and paper based. It consisted of text describing a moral dilemma, followed by an emotion measure (PANAS-X). After rating their emotions, participants made a moral judgment about the dilemma presented to them.

## Results

Experiment 2 replicated the findings from Experiment 1. In Experiment 2, 61 participants read the Coupon vignette, 51 read the Crying Baby, 57 read the Footbridge, and 52 read the Trolley. On average, participants judged it appropriate to use the coupon (mean $=5.51$ ) and to switch the track in the Trolley dilemma (mean $=4.69$ ). Participants were divided over whether to smother the baby (tending to disapprove somewhat, mean $=2.88$ ), and in general disapproved of pushing the man $($ mean $=2.32)$.

An ANOVA revealed significant differences between the moral and control vignettes in terms of the emotions they elicited, with significance obtaining for the Guilt $(F(3,217)$ $=31.86, \mathrm{MSE}=1.389, p<.001)$, Hostility $(F(3,217)=$ 52.57, MSE $=.929, p<.001)$, and Joviality $(F(3,217)=$ 15.339, MSE $=.824, p<.001$ ) subscales. Planned comparisons again revealed the sensitivity of the PANAS-X for measuring the different emotions elicited by different moral cases. The Crying Baby vignette elicited more Guilt

Table 3: Summary of Linear Regression Model Predicting Moral Judgments in the Crying Baby Dilemma in Experiment 2.

|  | $b$ | SE | t | $p$ |
| ---: | :---: | :---: | :---: | :---: |
| Guilt | -.020 | .345 | -.059 | .954 |
| Hostility | .298 | .392 | .758 | .452 |
| Joviality | .175 | .336 | -.521 | .605 |
| Intercept | 2.216 | 1.038 |  |  |
| F |  | .764 |  |  |
| df |  | 47 |  |  |
| $p$ |  | .52 |  |  |
| $\mathrm{R}^{2}$ |  | .047 |  |  |

$(t(217)=4.08, p<.001)$ and Hostility $(t(217)=5.72, p<$ .001) than did the Footbridge and Trolley vignettes. The same contrast for the joviality subscale approached significance $(t(217)=1.81, p=.07)$.

As was observed in Experiment 1, participants who read the Trolley vignette reported greater levels of Guilt ( $t(111)$ $=6.05, p<.001)$ and Hostility $(t(111)=7.01, p<.001)$, and lower levels of Joviality $(t(111)=-3.67, p<.001)$ than those who read the non-moral control vignette. To provide a second test of the Emotion Explanation, participants' ratings of Guilt, Hostility and Joviality were compared between the Trolley and Footbridge vignettes with three separate $t$-tests. These tests again found no significant differences between the two groups for Guilt $(t(107)=-1.42, p=.16)$, Hostility $(t(107)=-.34, p=.74)$ or Joviality $(t(107)=1.27, p=.21)$.

The use of Likert scale ratings of moral judgments in Experiment 2 enabled us to assess the relationship between emotions and moral judgments in the Crying Baby dilemma using a simpler linear regression model. The model used the three PANAS-X sub-scale scores as predictors of people's moral judgments. The resulting equation is summarized in Table 3. Replicating our findings in Experiment 1, no significant relationship was found between participants' emotion ratings and their moral judgments for the Crying Baby dilemma $(F(3,45)=.90, p=.45)$. This model accounted for less than 5 percent of the variance in participants' Crying Baby judgments, indicating an extremely weak (and non-significant) relationship between emotions and moral judgments about the case.

## Discussion

We hypothesized that the Trolley dilemma would be more emotionally engaging than many moral psychologists have claimed. Additionally, we suspected that because of the high degree of similarity between the Trolley and Footbridge dilemmas, any differences between the emotions elicited by the two dilemmas would not be sufficient to explain the large difference in participants' judgments about the dilemmas. The results of Experiments 1 and 2 confirm these two predictions. We found that participants reported significantly more guilt and hostility in response to the Trolley dilemma than to the control dilemma, and
significantly less joviality. Additionally, we found no difference in emotion ratings from participants who responded to the Footbridge dilemma and those who responded to the Trolley dilemma, despite the rather large sample size of our experiment.

To our knowledge, this is the first use of the PANAS-X to compare the emotions elicited by different moral dilemmas. Additionally, one of our primary hypotheses was that there would be no differences between the emotions elicited by the Footbridge and Trolley judgments. This made it important to confirm that the PANAS-X is sufficiently sensitive to detect emotional changes, not just between moral dilemmas and a non-moral dilemma, but also between different moral dilemmas. We found that participants who responded to the Crying Baby dilemma reported significantly stronger negative emotions and weaker positive emotions than did participants who responded to the Trolley and Footbridge dilemmas, demonstrating the sensitivity of the PANAS-X.

The Emotion Explanation, as stated, concerns the Trolley and Footbridge dilemmas specifically. However, this explanation has been employed to a number of different ends, not all of which necessarily hinge on facts about these specific dilemmas. In this way, some proponents of the Emotion Explanation might argue for a more general claim, viz., in general, deontological moral judgments recruit gut, emotional processes. In an emotionally salient case like the Crying Baby dilemma, proponents of this more general Emotion Explanation ought to predict a relationship between emotions moral judgments, such that the stronger the negative emotional reaction a participant experiences, the more likely they are to give a deontological judgment. Across two experiments, we observed no such relationship between emotions and moral judgments in the Crying Baby dilemma.

In sum, our findings indicate that the Trolley, Footbridge and Crying Baby dilemmas fail to conform to the more general Emotion Explanation. It is a question for future research whether these dilemmas are simply exceptions to the rule, or whether the more general Emotion Explanation should be rejected. Our findings cast doubt on the Emotion Explanation and may call for revision of the psychological, ethical, and epistemological theories in which it has been employed.

## Acknowledgments

Preparation of this paper was supported by grant N000140810186 from the Office of Naval Research.

## References

Baumeister, R. F., Stillwell, A. M., \& Heatherton, T. F. (1994). Guilt: an interpersonal approach. Psychological Bulletin, 115(2), 243-267.
Ciaramelli, E., Muccioli, M., Làdavas, E., \& Di Pellegrino, G. (2007). Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex. Social cognitive and affective neuroscience, 2(2), 84-92.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., \& Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. Neuron, 44(2), 389-400.
Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., \& Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. Science, 293(5537), 2105-2108.
Haidt, J. (2001). The Emotional Dog and is Rational Tail: A Social Intuitionist Approach to Moral Judgment. Psychological Review, 108(4), 814-834.
Heekeren, H. R., Wartenburger, I., Schmidt, H., Prehn, K., Schwintowski, H. P. et al. (2005). Influence of bodily harm on neural correlates of semantic and moral decisionmaking. NeuroImage, 24(3), 887-897
Irwin, W., Davidson, R. J., Kalin, N. H., Sorenson, J. A. \& Turski, P. A. 1998 Relations between human amygdala activation and self-reported dispositional affect. J. Cogn. Neurosci. (Suppl. S), 109.
Keltner, D. (1995). Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. Journal of Personality and Social Psychology, 68(3), 441-454.
Keltner, D., Haidt, J., \& Shiota, M. N. (2006). Social Functionalism and the Evolution of Emotions. In Schaller, M., Simpson, J. A., \& Kenrick, D. T. (Eds.) Evolution and Social Psychology (115-142). New York, NY: Psychology Press.
Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., \& Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgments. Nature, 446(7138), 908-911.
McGuire, J., Langdon, R., Coltheart, M., \& Mackenzie, C. (2009). A reanalysis of the personal/impersonal distinction in moral psychology research. Journal of Experimental Social Psychology, 45(3), 577-580.
Singer, P. (2005). Ethics and Intuitions. The Journal of Ethics, 9.
Watson, D., Clark, L. A., \& Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect schedule. Journal of Personality and Social Psychology,54(6), 1063-1070.
Zald, D. H., Mattson, D. L., \& Pardo, J. V. (2002). Brain activity in ventromedial prefrontal cortex correlates with individual differences in negative affect. Proceedings of the National Academy of Sciences of the United States of America, 99(4), 2450-4.

# Young children's developing sensitivity to discourse continuity as a cue to reference 

Alexandra C. Horowitz<br>ahorowitestanford.edu<br>Department of Psychology<br>Stanford University

Michael C. Frank<br>mcfrank@stanford.edu<br>Department of Psychology<br>Stanford University


#### Abstract

Children can learn new words in pedagogical contexts, but they may also infer reference using a variety of other information sources. Here we investigate children's sensitivity to the placement of novel labels within discourse structure as a possible mechanism for word learning. In Experiment 1, children ages 2-6 years participated in word learning trials featuring two novel items and one novel label. In critical trials, the labels were embedded between two sentences about the same item, whereas in a control condition, the label was introduced after two sentences about the item. Children of all ages were more likely to attribute the label to the toy whose descriptions bracketed the embedded label, and response strength increased with age. Children across all ages responded at chance in the control condition. In Experiment 2, adults showed the same patterns of responses as children in both critical and control conditions. Together, these results suggest that discourse continuity is a reliable cue to reference for both children and adults.


Keywords: word learning; discourse; social cues; language development.

## Introduction

Children use a variety of strategies for learning new words. In overtly pedagogical situations, children can use cues such as pointing, joint attention, and labeling to establish a direct mapping between an object and name. However, many situations do not feature ostensive labeling events. In these cases, children must rely on other strategies to infer the referent of a novel word. One source of information may come from discourse structure (the order of utterances and how they relate to each other). Recognizing how speakers relate topic information may help children resolve reference that would otherwise be ambiguous in the absence of the broader context.

For example, a child may not have an idea of what chinchilla means from an utterance such as, "I love chinchillas!", but she may apply her knowledge of discourse structure to infer its meaning when the same utterance is related to topical information, such as "I got a new pet. I love chinchillas! They're so soft." Children are exposed to information about discourse structure whenever they hear speech, and their accumulation of experience may help them to update and refine their expectations about topic relationships. This developing expertise may allow children to infer meaning that is locally ambiguous, yet resolvable in the context of broader discourse.

Little work has explored whether children use discourse structure to scaffold reference disambiguation. Our aim is to address this question by investigating children's and adults' recognition of communicative structure when dissociated from other social and ostensive cues. Understanding the contribution of discourse knowledge in children's reference disambiguation may help identify opportunities for children to infer meaning from topic coherence.

A large body of research has been devoted to children's ability to map names to inferred referents through disambiguation of a single new item and label. In the presence of a known item and an unknown item, children map a novel label to the novel item rather than the already-named item (Markman \& Wachtel, 1988; Merriman, Bowman, \& MacWhinney, 1989; Mervis \& Bertrand, 1994; Clark, 1990). Though the mechanisms at play are debated, this finding establishes that children can make inferences about a speaker's likely intended meaning in constrained contexts. While employing their repertoire of word-object mappings may allow children to disambiguate some novel referents, there may be other situations in which multiple items are unknown.

Children can also use social-pragmatic cues to infer novel word reference. By age 2, children map novel labels to novel objects that the speaker attends to rather than what they themselves may be attending to (Baldwin, 1991), and even after a time delay (Baldwin, 1993). Young children apply new terms to the target but not non-target items of a speaker's search (Tomasello \& Barton, 1994), show evidence of considering not only their own novelty perspective, but also what is novel to a speaker in a discourse context (Akhtar, Carpenter, \& Tomasello, 1996), and can recognize that speaker naming events may convey social-pragmatic implications about information that is expected to be shared (Diesendruck \& Shemer, 2006).

Despite this body of evidence on general pragmatic cues, few studies have investigated children's sensitivity to discourse structure. Nevertheless, some work suggests that discourse continuity may provide important opportunities for learning. Frank, Fernald, and Tenenbaum (2013) examined a video corpus of caregivers interacting with their 6-18 monthold children. In these natural settings, they found discourse continuity (that utterances in close succession are likely to relate to the same topic) was as reliably linked to reference as were social cues such as pointing and gaze (Frank, Tenenbaum, \& Fernald, 2013). Although this result suggests that discourse continuity may be an available cue to disambiguate reference in the presence of competitors, their work did not provide evidence that learners actually make use of this information. Therefore, our present studies provide the first test of whether discourse position can be used to determine reference in word learning.

We discuss two experiments indicating that both children and adults can use discourse position to resolve reference ambiguity. Additionally, our control condition shows that children and adults rely on the informativeness of discourse structure rather than simpler heuristics such as temporal proxim-


Figure 1: Schematic order of events for "A" trials across conditions. In the During condition, the experimenter makes eye contact without other gaze cues and introduces the naming event between two descriptions of a single toy. In the After control, events are identical except that the experimenter introduces the naming event after two descriptions of a toy.
ity. Our findings suggest that language users are able to apply their knowledge about how utterances relate within a discourse to make inferences about speakers' intended referents. Overall, this work suggests that children are not constrained to social and contextual cues from individual utterances, but can evaluate how information may relate across broader discourses.

## Experiment 1

We designed a novel task to investigate children's recognition of discourse structure as a cue to reference. In a large sample of 2- 6 year-olds, we introduced children to two novel toys accompanied by only one novel label, and we manipulated where in discourse the name was introduced. We were interested in which toy children selected as the referent of the label. The experimenter made eye contact with the child but gave no gaze or gesture cues to the referent of the label during the naming event, so the only cue to reference was the location of the naming event within the broader discourse. In the critical During condition, children were introduced to a novel label within descriptions about either the first ("A") or second ("B") toy (see Figure 1). Because the label was embedded within descriptions of the same toy in critical trials, we were interested in whether children would infer topic continuity and link the label to that toy.

An alternative explanation for why children might choose the toy whose descriptions bracketed the naming events is that children are making a temporal association between the label and the toy descriptions rather than considering discourse structure per se. That is, children may be selecting the toy that is described closest to the naming event, which would always correspond with the toy surrounding the introduction of the label in During trials. Therefore, we also ran a control condition to dissociate temporal proximity from discourse coherence. In the control After condition, the naming event was introduced after descriptions of either Toy A or Toy B rather than between descriptions of that toy. Thus the naming event could occur next to descriptions of both toys in After $A$ trials, or next to only Toy B in After B trials. Sample scripts for each trial type are listed in Table 1.

Our design allows us to make the following predictions: If children recognize discourse continuity as a cue to reference, they should infer that new information contained within a single topic is likely to also refer to that topic. Therefore, children should select Toy A in During A trials and Toy B in During $B$ trials, but not have a clear strategy in After trials. If children rely on temporal association rather than topic coherence, they should select a referent according to what comes closest to the naming event, i.e. Toy A in During $A$ and Toy B in During $B$ trials, as well as Toy B in After $B$ trials because
it is the only description proximal to the naming event. These predictions are outlined in Table 2.

Table 1: Sample scripts for each condition (During or After) and trial type (Toy A or Toy B). The green sentences are descriptions of Toy A, and the blue sentences are descriptions of Toy B.

| During A | After A |
| :--- | :--- |
| The top of this one is wobbly. | The top of this one is wobbly. |
| Have you seen a toma before? | Look how to move this switch. |
| Look how to move this switch. | Have you seen a toma before? |
| The sides of this one are bumpy. | The sides of this one are bumpy. |
| Look how to squish the top down. Look how to squish the top down. |  |
| During B | After B |
| The top of this one is wobbly. | The top of this one is wobbly. |
| Look how to move this switch. | Look how to move this switch. |
| The sides of this one are bumpy. | The sides of this one are bumpy. |
| Have you seen a toma before? | Look how to squish the top down. |
| Look how to squish the top down. Have you seen a toma before? |  |

Table 2: Predictions for reference selections (Toy A or Toy B) across each trial type if participants rely on discourse continuity or temporal association.

|  | Discourse Continuity | Temporal Association |
| :--- | :---: | :---: |
| During A | A | A |
| During B | B | B |
| After A | either | either |
| After B | either | B |

## Methods

Participants One hundred sixty-six children were recruited from the San Jose Children's Discovery Museum to complete a planned sample of 128 children. Children were given a sticker and certificate as compensation their participation. Parents were asked to fill out a short demographic form about their children's language background, and only children who were reported to hear English at least $75 \%$ of the time were included in the study. Twelve children were excluded due to insufficient English exposure, 12 children whose language information was not reported were excluded, and 14 children were excluded for not completing all four trials of the study.

Children were recruited in four age groups: 2-year-olds ( $\mathrm{n}=32,18$ girls, mean age 2 years 6 months), 3 -year-olds years ( $\mathrm{n}=32$, 10 girls, mean age 3 years 6 months), 4-yearolds ( $n=32$, 14 girls, mean age 4 years 6 months), and 5-yearolds ( $n=32,18$ girls, mean age 5 years 4 months).
Stimuli Four pairs of unusual items (e.g. a faucet aerator and a spaghetti measure) served as the novel toys. An additional item was used for training.

Procedures Participants were seated across from the experimenter in a quiet room at the museum. Children participated
in a training trial featuring ostensive labeling of a single toy ("This toy is called a blicket. Can you point to the blicket?") before seeing four discourse disambiguation trials. For the discourse disambiguation trials, the experimenter placed two toys on the table and described each in turn (see Figure 1). The toy pair remained in view of the child throughout the duration of the trial. All children heard the same scripts used to describe the toys; the only difference was the discourse location where the label was introduced. In a between-subjects manipulation, half the participants in each age group ( $\mathrm{n}=16$ per age) participated in four During trials; the other half participated in four After control trials. Order and toy pairs were counterbalanced across participants. At the end of each trial, the experimenter prompted the child to identify the named item by pointing. If children did not respond immediately, they were prompted again to make their best guess. The sessions were videotaped and coded offline. The entire task took about 5 minutes to complete.

Sample scripts for each condition and trial type are shown in Table 1. In During trials, the experimenter introduced the naming event between two sentences about the same toy (e.g. "You can push this button. Hey [child's name]! Have you seen a toma before? Tomas are so neat! What cool handles."). For two trials the label introduction was embedded during descriptions of the first toy (Toy A) and for two trial it was introduced during descriptions of the second toy (Toy B). When describing the toys, the experimenter directed her gaze to the toy and demonstrated a feature of the toy. There was a brief pause between each sentence. For the naming event, she disengaged from the toy and maintained a neutral position while drawing the children's attention using their names and establishing eye contact. The experimenter did not give any gaze cues or other indicators to the referent of the novel name. Thus, the naming event in itself carried no information to guide disambiguation; the only cue available was its location within discourse. In During trials, the naming event was always embedded between descriptions of a single toy. The After trials were identical except that the naming event appeared after the two descriptions about a toy (e.g. "You can push this button. What cool handles. Hey [child's name]! Have you seen a toma before? Tomas are so neat!").

## Results and Discussion

Figure 2 illustrates the proportion of children selecting the second toy (Toy B) as the referent of the label across conditions (During and After) and trial types (whether the label was introduced with Toy A or Toy B). The figure also includes adult performance across conditions and trial types from Experiment 2.

Children showed increased sensitivity to discourse coherence over development. Overall, children in the During condition were more likely to select Toy B when the label was embedded during descriptions about Toy B , and were less likely to select Toy B (thus more likely to select Toy A) when the naming event was bracketed by descriptions about Toy A. This pattern became more pronounced as children got older.


Figure 2: Combined data from Experiments 1 and 2. Mean proportion of selection of the second toy across condition (During or After) and trial type (label given with first toy: A, or second toy: B). Chance performance is at 0.5 , error bars represent $95 \%$ confidence intervals.

This finding suggests that children's sensitivity to discourse coherence as a cue to disambiguate reference increases across ages 2-6 years.

In After trials, children were at chance in selecting either toy for both After $A$ and After $B$ trials for all age groups. This result shows that children did not develop a consistent strategy to disambiguate reference when cues to topic coherence from discourse structure were not available.

To test the reliability of these patterns, we ran a generalized linear mixed model predicting toy selection as an interaction between condition (During or After), trial type (trial location A or B), and age with random effects of participant. There was a significant three-way interaction between condition, trial type, and age ( $\beta=0.89, p=0.03$ ), indicating that, with increasing age, participants were more likely to select Toy A in A trials and Toy B in B trials only for the During condition. No other factors were significant.

We also ran a series of paired $t$-tests to examine response differences between trial types (naming event with Toy A or Toy B) within condition (During or After) for each age group $(2-3 \mathrm{~s}, 3-4 \mathrm{~s}, 4-5 \mathrm{~s}$, and $5-6 \mathrm{~s}$ ) (Table 3). Significant differences in reference selection were found across trial types in the During condition for children ages 3-6 years; at these ages, children were more likely to select Toy A in During A trials and Toy B in During $B$ trials. By age 3, children were able to consider the broader discourse structure to help disambiguate the target referent.

In the naming events, children shared eye contact with the experimenter and were introduced to a novel word, but there were no indicators of the referent of the label other than its location within the broader discourse. Children's systematic responding to selecting the toy whose descriptions bracketed the naming event thus suggests that they can recognize and
refer to discourse coherence to infer reference in the absence of other social cues.

Could participants have assumed that the novel labels referred to both toys at once? The uniqueness of the toys makes this situation unlikely. Toys were distinct artifacts with different colors, shapes, and functions, and items in a pair were presented at opposite ends of the table, giving no visual signal that the toys were grouped. Additionally, if participants believed that both toys in a pair were examples of a novel category, we would have observed responding at chance across both During and After conditions. Thus we do not believe that a superordinate interpretation of the novel terms would explain the pattern of data we observed.

Finally, children did not appear to use temporal proximity to disambiguate reference in After trials. We found no significant differences between reference selections across After A and After $B$ trials for any age group, suggesting that children did not have consistent strategies for disambiguating the intended meaning of a novel term when topic coherence information is not available. While the After trials lacked definitive cues to establish reference, we were unsure whether adult users might have strategies for interpreting these trials.

## Experiment 2

In Experiment 2, we extended our design to adult participants. We wanted to confirm adults' sensitivity to discourse continuity as a cue to word learning in During trials, and assess strategies for referent disambiguation in After trials to compare with our developmental results.

## Methods

Participants Twenty-five adult participants were recruited from the San Jose Children's Discovery Museum, and were

Table 3: Results from paired t-tests examining response differences across trial types (naming location with Toy A or Toy B) within condition (During or After) for each age group (2$3 \mathrm{~s}, 3-4 \mathrm{~s}, 4-5 \mathrm{~s}$, and $5-6 \mathrm{~s}$ ). Children ages 3-6 years show significant differences in their response selections between During $A$ and During $B$ trials. No other significant differences across trial types are found.

|  | During condition |  |  | After condition |  |  |
| :---: | ---: | ---: | ---: | ---: | ---: | ---: |
| Age | $t$-value | df | $p$-value | $t$-value | df | $p$-value |
| $2-3$ | -1.23 | 15 | 0.24 | -1.58 | 15 | 0.14 |
| $3-4$ | -2.52 | 15 | $\mathbf{0 . 0 2}$ | -1.23 | 15 | 0.24 |
| $4-5$ | -5.48 | 15 | $<\mathbf{0 . 0 1}$ | -1.95 | 15 | 0.07 |
| $5-6$ | -8.22 | 15 | $<\mathbf{0 . 0 1}$ | -1.58 | 15 | 0.14 |

offered a sticker and certificate for their participation. They were informed that the task was designed for children. Only participants who reported using English at least $75 \%$ of the time were included in the study. One participant was excluded for reporting English use under this threshold.

Stimuli and procedure The stimuli and procedure were identical to Experiment 1, with the exception that adults did not undergo a training trial to practice pointing. Otherwise, adults were randomly assigned to either the During or After condition, and trial order and toy pairs were counterbalanced across participants.

## Results and Discussion

Results were coded for whether participants selected the second toy (Toy B) as the referent of the novel label (see Figure 2). Participants almost never selected Toy B in During A trials, but were near ceiling at selecting Toy B for During $B$ trials. Responses to each trial type were significantly different from chance in exact binomial tests ( $p<0.01$ for both trial types), and significantly different from each other ( $\beta=-2.06, p=0.01$ ) in a generalized linear mixed model predicting toy selection by condition (During or After) and trial type (trial location A or B) with random effects of participant. These results illustrate that adults were sensitive to naming events within discourse structure as informative cues to referent disambiguation; like the results in Experiment 1, adults demonstrate referent selections that correspond with discourse coherence, selecting the toy whose descriptions bracketed the naming event.

For After trials, adult participants were at chance in selecting Toy B in both After A and After B trials ( $p>0.3$ in exact binomial test for each). Performance was not significantly different between After $A$ and After $B$ trials ( $\beta=0.85$, $p=0.15$ ). These findings indicate that adults did not exhibit a strategy for disambiguating reference in After trials; they were at chance in determining a referent when the naming event followed the descriptions of either toy. This pattern of results also parallels the developmental results. Adults, like children, did not show systematic response patterns when discourse information was not available.

Together, these results suggest that language users are sensitive to how information relates within discourse structure. Adults and children systematically disambiguated reference when they could infer topic coherence in During trials. In contrast, we found that listeners did not develop consistent response strategies for information that is isolated from social and discourse context in After trials. Neither children nor adults followed heuristics such as resolving reference by temporal association.

## General Discussion

We investigated whether adults and children could use position in discourse as a cue to resolve reference. In our experiments, adults made use of discourse position effectively and children showed increasing sensitivity to discourse position across childhood. All groups except the youngest in our study successfully used discourse position to infer the mapping of a label. Taken together, our findings suggest that language users learn to make inferences about reference not only from pragmatic or social cues, but also from information about the general discourse in which a novel label is embedded.

Our experimental design ruled out two alternative explanations. The first is that children were simply selecting the referent most proximal to the naming event. In the During condition, this temporal proximity account would make the same predictions as a discourse-based account. Our After condition allowed us to rule out this possibility. While temporal proximity remains ambiguous in After $A$ trials, it is unambiguous in After B trials because Toy B is the only toy described proximate to the naming event. However, children at all ages responded around chance for both After $A$ and After B trials with no difference between naming location, suggesting that children did not use temporal proximity alone to make their judgments.

Our After condition rules out a second possible interpretation as well: that children's mappings are driven purely by novelty. By age 2, children state novel rather than given information in their productions (Baker \& Greenfield, 1988) and apply similar expectations to other speakers by mapping new labels to items novel to the speaker's perspective (Akhtar et al., 1996). Children's behavior in the During condition is consistent with a novelty account: they chose the toy that was most recently attended to by the experimenter (the one newest to the discourse). But on this account, children should assign Toy B as the referent in After A trials, because the naming event directly precedes the introduction of this toy. Instead, responses from the After A condition were at chance between the two toys, suggesting that novelty alone also did not account for our findings.

Children can learn from ostensive naming events when they are available, but many situations they encounter are not overtly pedagogical, and the ability to extract information that is embedded in discourse may help children deduce the meanings of words in these cases. As our initial chinchilla example illustrated, discourse position is a powerful informa-
tion source for understanding language and for learning new words. Children who can infer how new information relates to the current topic may be able to accumulate knowledge more accurately and more efficiently. Because of its accessibility to young children in our study (and the possibility that even younger children might use discourse position in a simpler task), the use of discourse structure to help disambiguate reference might be one of the array of learning mechanisms that helps explain children's rapid vocabulary growth. Using topic coherence to make inferences about novel terms and information may allow children to access learning opportunities that would otherwise be unavailable.

## Acknowledgments

Special thanks to the staff and families at the San Jose Children's Discovery Museum.

## References

Akhtar, N., Carpenter, M., \& Tomasello, M. (1996). The role of discourse novelty in early word learning. Child Development, 67(2), 635-645.
Baker, N., \& Greenfield, P. (1988). The development of new and old information in young children's early language. Language Sciences, 10(1), 3-34.
Baldwin, D. (1991). Infants' contribution to the achievement of joint reference. Child Development, 62(5), 874-890.
Baldwin, D. (1993). Early referential understanding: Infants’ ability to recognize referential acts for what they are. $D e$ velopmental Psychology, 29(5), 832.
Clark, E. (1990). On the pragmatics of contrast. Journal of Child Language, 17(02), 417-431.
Diesendruck, G., \& Shemer, G. (2006). Young children's expectation of competence in word learning. Journal of Child Language, 33(2), 321.
Frank, M., Tenenbaum, J., \& Fernald, A. (2013). Social and discourse contributions to the determination of reference in cross-situational word learning. Language, Learning, and Development.
Markman, E., \& Wachtel, G. (1988). Children's use of mutual exclusivity to constrain the meanings of words. Cognitive psychology, 20(2), 121-157.
Merriman, W., Bowman, L., \& MacWhinney, B. (1989). The mutual exclusivity bias in children's word learning. Monographs of the society for research in child development.
Mervis, C., \& Bertrand, J. (1994). Acquisition of the novel name-nameless category (N3C) principle. Child development, 65(6), 1646-1662.
Tomasello, M., \& Barton, M. (1994). Learning words in nonostensive contexts. Developmental Psychology, 30(5), 639.

# Anxiety Can Influence Analogy-Making Both Positively and Negatively Depending on the Complexity of the Mapping Task 

Penka Hristova (phristova@cogs.nbu.bg)<br>Department of Cognitive Science and Psychology, New Bulgarian University, 21 Montevideo Street Sofia 1618, Bulgaria<br>Yoana Petkova (yoana.l.petkova@gmail.com)<br>Department of Cognitive Science and Psychology, New Bulgarian University, 21 Montevideo Street Sofia 1618, Bulgaria<br>Boicho Kokinov (bkokinov@nbu.bg)<br>Department of Cognitive Science and Psychology, New Bulgarian University, 21 Montevideo Street Sofia 1618, Bulgaria


#### Abstract

In attempt to resolve the controversial issue of the influence of the anxiety state on analogy-making this paper presents a replication of the original Tohill and Holyoak study extending it with a new factor - the complexity of the mapping. It turns out that the anxiety influence interacts with the complexity of the mapping task. This has certain implications for the models of analogy and for the further study of the role anxiety plays in analogy-making.


Keywords: analogy; emotion, state anxiety, complexity.

## How Anxiety Changes Analogical Mapping

The chronologically first studies that focus on analogies under anxiety, suggest that analogical performance drops significantly when the state anxiety is heightened. Leon and Revelle (1985) manipulated anxiety via time pressure. They found that people are more accurate under low timepressure, but unfortunately, as Tohill and Holyoak (2000) have argued, these results could also be interpreted as a speed/accuracy trade-off. The negative correlation between anxiety and accuracy reported for the low time-pressure group is also inconclusive with respect to the possible causal link between anxiety and analogies.

Tohill and Holyoak (2000), however, also found decline in identification of the relational mappings under an induced state anxiety in a well-controlled study (Exp2). Prior to the analogy-making session, they manipulated the state anxiety via a serial subtraction task with negative feedback. The procedure turned out to differentiate successfully the anxiety of the two groups, measured with the State-Trait Anxiety Inventory (STAI) (Spielberger, Gorsuch, \& Lushene, 1970). Analogy performance was measured by a cross-mapping visual task. Participants saw two pictures on the screen for 15 seconds, then one object in the first picture was pointed at and they were instructed to indicate either which object from the second picture "goes with" the pointed one (Exp1), or to indicate the object from the second picture that "relationally matches" the pointed one (Exp2). The task is difficult because the target object has always two plausible matches - one at the level of object similarity (i.e., the same
or an extremely similar object present in the second picture) and another at the level of relational similarity (i.e., a superficially distinct object that plays an analogous role in the second picture). Less relational responses (Exp1) and less accuracy of identifying the relational match (Exp2) were found for participants from the high state-anxiety group. The decline in relational reasoning was explained by the Processing Efficiency Theory of Eysenck and Calvo (1992). According to that theory and its successor - the Attentional Control Theory (Eysenck, Deraksan, Santos \& Calvo, 2007), attention of highly anxious individuals is distracted by the anxiety related thoughts. Thus, instead of focusing on the main task, anxious individuals ruminate about the threatening situation, the stimuli or the potential failure. Hence, as Holyoak (2012) has pointed out, anxiety seems to place individuals in a dual task situation that definitely reduces the available working memory resources and in turn, changes analogical mapping from relational (based on common relations) to superficial mode (based on common features). Besides, it was argued that both the capability to integrate multiple relations and to inhibit distracting information during analogy-making depend crucially on the available resources (Cho, Holyoak, \& Cannon, 2007; Krawczyk, Morrison, Viskontas, Holyoak, Chow, Mendez, Miller, \& Knowlton, 2008; Sweis, Bharani, \& Morrison, 2012; Viskontas, Morrison, Holyoak, Hummel, \& Knowlton, 2004). To sum up, it seems reasonable to consider that state anxiety reduces the capability of an individual to integrate the relevant relations and to inhibit the irrelevant ones, which are directly connected to analogical mapping.

However, analogy-making involves not only choosing among various potential relational matches, but is a much more complicated process that integrates perception and encoding of the existing relations in the environment (among the many possible relations some are relevant for the analogy, others are not), building hypotheses about possible pairs of corresponding relations, and choosing among these hypotheses which are forming the most consistent mapping.

Anxiety could have an influence on all these processes, including the perception of relations. However, it was shown that threatening stimuli prompt participants to search faster for both threatening and non-threatening stimuli (Becker, 2009). Likewise, Phelps, Ling, and Carrasco (2006) reported that contrast sensitivity increases after a presentation of a fearful face. In addition, Pacheco-Unguetti, Acosta, Callejas, and Lupianez (2010) reported neuroimaging data suggesting that state and trait anxiety modulate differentially the work of the three attentional networks presumed by Posner and Petersen (1990) and Posner, Rueda, and Kanske (2007). State anxiety was associated with overfunctioning of the alerting and the orienting attentional networks, while trait anxiety with insufficiency of the executive control network. In other words, state anxiety reinforces bottom-up processing (i.e., perception and selection of relevant information), and traitanxiety hampers the top-down control (i.e., voluntary action control and conflict resolution). That is why Hristova and Kokinov (2011) studied the influence of anxiety on the perception of relations. They reported that people in heightened state anxiety are superior in encoding of relations between superficially dissimilar geometric figures: they were both faster and significantly more accurate than participants in the control group in recognizing identical relations between two sets of figures. Therefore, it was argued that the superior encoding of relations under heightened state anxiety may improve, instead of impeding analogical mapping. Indeed, some research supports this hypothesis.

Richert, Whitehouse, and Stewart (2005) showed that religious rituals, performed in high anxiety states, lead to a higher percentage of generated reflections (including ones, based on analogies) and that percentage increased significantly over a one-month time period. They argued that rarely performed religious rituals, accompanied by high physiological arousal become the focus of conscious rumination and in that way advantage the drawing of more and deeper analogies between the current anxious situation and the individual personal memory. This leads to a better memory for a given ritual and binds the religious ideas to personal experience. The reported results, however, were inconclusive, since arousal was manipulated through the ritual itself (high arousal and low arousal rituals) and analogies were only part of the interpretations that were measured.

Later, Feldman, Hristova, and Kokinov (2010) demonstrated that participants in high state anxiety were more prone to relational matches between superficially dissimilar stimuli instead of superficial matches between structurally dissimilar stimuli in a match-to-sample task. Participants were shown a sample set that consists of three geometric figures and were asked to indicate which of the two target sets of figures is more similar to the sample one. High state anxiety participants were significantly more likely to indicate the relational target instead of the superficially similar one, for a comparable amount of time
(i.e. the difference between RTs in the anxiety and the control group was insignificant). Thus, surprisingly, it was demonstrated that induced state anxiety can support analogies in a situation quite similar to the one used in the first experiment of Tohill and Holyoak (2000). Both experiments require a choice between superficially similar and relationally similar options. The authors (Feldman et al., 2010) discussed the controversy between the data and suggested that it can be potentially explained by the difference in the experimental procedures that possibly eliminated the benefits of the superior encoding (if any) for the anxiety group in the Tohill and Holyoak's case: maybe the 15 sec stimulus presentation used in the Tohill and Holyoak's study (Tohill \& Holyoak, 2000), but not in theirs ${ }^{1}$, had restricted the effect of the superior encoding of relations that was discussed above.

Alternatively, however, as Feldman et al. (2010) discussed, the same empirical inconsistency may be due to the difference in either the intensity of the induced state anxiety or the difficulty of the analogical tasks used in either or both studies. That particular explanation corresponds to the well-known Yerkes-Dodson Law (Yerkes \& Dodson, 1908) that postulates a nonlinear inverted U-shaped relationship between stimulus strength and the rapidity of habit formation for tasks varying in discrimination difficulties. Taking into account the fact that later, it was largely assumed that the same U-shaped relationship is also valid when describing the relation between (emotional) arousal and performance (Broadhurst, 1957), and that arousal is largely recognized as the physiological component of any emotional state, including anxiety, it is reasonable to assume that the studies of Tohill and Holyoak (2000) and Feldman et al. (2010) could have been run with anxiety states that differ in intensity. Unfortunately, the two studies differ in both the procedure used for anxiety induction $^{2}$ and the manipulation-check instrument ${ }^{3}$, so it was not possible to reject that explanation at that point. Moreover, those studies used different stimuli ${ }^{4}$ and tasks ${ }^{5}$. Hence, if one of the tasks was more difficult than the other,

[^94]the respective optimal levels of arousal will also differ, since the more difficult the task is, the lower the respective level of optimal arousal is.

Thus the goal of the present study is to replicate Experiment 2 of Tohill and Holyoak (2000), while explicitly varying the complexity of the task. And we expect interaction between the anxiety state (low and high anxiety) and the complexity of the analogy task. The complexity here is operationalized by the number of potential hypotheses that could participate in the competition for the best match. This is manipulated in the task by the number of alternatives offered among which the answer is to be chosen.

## Experiment: Many vs. Small Number of Alternatives

This experiment varies within-subject the number of the suggested alternatives for the answer. This seems an easiest way to manipulate the complexity of the task without changing the stimuli themselves: choosing between two alternatives is easier than choosing between four.

## Method

## Design

We used a $2 x 2$ mixed factorial design with two levels of state anxiety (anxiety and control) and two options for complexity of the task (choosing between 2 answers vs. choosing between 4 answers). The level of state anxiety was varied between-subject by serial subtraction task, described thoroughly in the procedure section below. The complexity of the task was operationalized as the number of possible alternatives for the answer and it was varied within-subject. Half of the trials for each participant required a choice of answer between 2 alternatives, the other half between 4 alternatives. The dependent variables of this study were the accuracy, defined as correct identification of the relational match, and the response time.

Between-subject counterbalancing: Stimuli were betweensubject counterbalanced with the number of alternatives (i.e. 2 or 4 alternatives for answer). In addition, the letters of the available alternatives for answer (i.e. A, B, C or D) were also balanced across stimuli and participants.

## Stimuli

The stimuli were 14 analogical picture-pairs with crossmapping (i.e., they allow both an attribute mapping of a key object and a relational one). Nine of those pairs were redrawn from the original stimuli used in the study of Tohill and Holyoak (2000) ${ }^{6}$. Taken as a whole, the pictures look

[^95]quite different from their originals, but they consist of the same relations and preserve the cross-mapping structure of their originals. We add 5 new picture-pairs that also have a cross-mapping structure. So, overall we used 14 black and white picture-pairs in our study. For all of them, the key object in the top picture (i.e. the circled one) corresponds to two objects in the bottom picture in the same picture-pair: in picture-pair 1 in Figure 1 the fisherman in the picture corresponds to both the fisherman and the seagull in the bottom picture. The former correspondence is based on shared physical characteristics and that is why, it can be considered as a result from an attribute mapping, while the latter is based on shared relations (i.e. they both catch the fish) and hence it can be considered to be based on a relational mapping. Similarly, the circled girl in the top picture of pair 2 (Figure 1), maps both the girl and the teddy-bear in the bottom picture. The former mapping is an attribution one (i.e. the two girls possess similar physical characteristics), the latter is an relational one (i.e. they both receive an injection).


Figure 1: Examples of picture-pairs that contain crossmapping. Picture-pair1 is a redrawing of one of the original drawings used in the Tohill and Holyoak study (2000).
Picture-pair2 is an example of one of the new stimuli with cross-mapping created purposely for the current study. Picture-pair3 is the training picture-pair, used to explain the meaning of relational correspondence.

## Procedure

Participants were tested individually in a soundproof cubicle. They were randomly assigned to the control or anxiety condition. All of them were informed that they would take part in a study on representation of numbers and they might be asked either to count backward or forward from a given number. They were told that the specific direction for each of them will be chosen randomly. However, since the counting procedure should be accomplished twice they were notified that we would asked them to take part in two short unrelated investigations inbetween: the first one, connected to relational reasoning and the second one, connected to an on-going standardization of a questionnaire for a Bulgarian population.

Participants in the anxiety condition were asked to count aloud backward beginning at 1000, 970 or 950 in increment of 13 . The starting point for each participant was randomly
assigned. Two experimenters took part in the experiment: one of them corrected the mistakes, while the other one measured the time and urged participants to count faster and faster, since the predetermined time of 45 seconds was about to expire. When the counting task was finished, one of the experimenters left the cubicle. Participants in the control condition were asked to count forward aloud beginning at 1 for 45 sec . They were instructed to count at a pace that relaxes them.

Both groups started the analogy-making task immediately after the counting. The procedure for stimuli presentation and the instruction given to the participants were analogous to those used by Tohill and Holyoak in Experiment 2 of their study (2000). Participants were told that they would be shoun picture-pairs one by one on the screen. Some of the objects in the top picture would be numbered and some of the objects in the bottom picture would be lettered. The two pictures will stay on the screen for 15 seconds before one of the numbered objects in the top picture will be circled in red. Their task is to indicate by pressing the respective button on the BBOX, which of the lettered objects from the bottom picture corresponds to the circled one from the top picture. They were trained to focus on the relations between objects with the robot example used by Tohill and Holyoak (2000) (Figure 1). Frist, the participants were asked to think and say, whether robot "A", "B" or "C" from the bottom picture is related analogously to the robot 1 from the top picture. Independent of the answer all participants received the same explanation: "Robot " 1 " corresponds to robot "A" because they took part in similar relations, i.e. they both are using weapons".

Then participants were instructed that some trials would require a choice between two alternatives, others between four alternatives. Finally, they were asked to indicate their answer as accurately and as fast as possible. If participants confirmed that they have understood the task and they have no questions concerning the experiment they move on to the target trails.

After the analogy-making task the participants filled out the Bulgarian adapted version of Spielberger's STAI (Щетински, Паспаланов, 1989). Then they count again forward or backward, depending on the condition and were not urged so frantically this time to count faster. In fact, the last counting was exclusively made to restore participant's mood after anxiety manipulation. Anxious participants were told that they have counted backward exceptionally well during their second turn. In addition, although we did not revealed the real hypothesis of the experiment, we told them that the counting task sometimes leads people to feel a little bit disturbed or tense, so if they feel in that way it may well be because of the task.

The whole experiment lasts between 15 and 20 minutes. The timing of events during the whole experiment is presented in Figure 3.

The experiment was double-blind: the experimenters knew that they should induce anxiety via the serial subtraction task and that we are looking for some differences in performance due to that anxiety. However
they didn't know what kind of differences were expected between conditions.


Figure 3: Stimulus displays and the timing of events.

## Participants

90 ( 44 female and 46 male) volunteers took part in this experiment. All of them were students at New Bulgarian University from different university specialties. The mean age was 23.5 years ranging from 19 to 39 years. The control group consisted of 44 participants, the anxious group - of 46. The groups were balanced on gender.

## Results and Discussion

The anxiety manipulation was successful; the mean state anxiety scores of STAI significantly differed between groups: the mean state anxiety for the anxious group was 40.59 ( $\mathrm{SD}=10.701$ ) and for the control group was 36.05 ( $\mathrm{SD}=9.977$ ). That difference in state anxiety turned to be significant tested with ANOVA: F $(1,89)=4.327$, $\mathrm{p}=0.040$. The mean trait anxiety scores, however, did not differed significantly between groups: $\mathrm{F}(1,89)=0.293, \mathrm{p}=0.590$ (means of 41.57 ( $\mathrm{SD}=10.100$ ) for the control condition and of 42.70 ( $\mathrm{SD}=9.661$ ) for the anxious condition). Hence, any difference between the groups should be due to that change in state anxiety, instead of some individual differences in trait anxiety. The difference in the state anxiety scores, however, is only 4.54 , while the same difference in the Tohill and Holyoak Experiment 2 was $9.2^{7}$. Therefore, any direct comparisons between the two experiments seem not well-grounded.

The mean percentage of relational responses for all conditions is presented in Table 1.
A Repeated Measures ANOVA with one within-subject variable (complexity of the task: task with 2 options and task with 4 options) and one between-subject variable (the level of state anxiety: anxiety and control) was carried out on the accuracy data. The main effect of state anxiety on accuracy was not significant $\left(F(1,88)=0.001, p=0.970, \eta_{p^{2}}\right.$ $=0.000$ ) but the main effect of complexity on accuracy was significant $\left(F(1,88)=8.996, p=0.004, \eta_{p^{2}}=0.093\right)$ such that

[^96]accuracy was higher when two alternatives were considered (86\%), compared to four (78\%).

Table 1: Mean percentage of relational mapping per condition.

| Condition | \%relational mappings |
| :--- | :---: |
| Control_2 alternatives | $89 \%$ |
| Control_4 alternatives | $75 \%$ |
| Anxiety_2 alternatives | $83 \%$ |
| Anxiety_4 alternatives | $81 \%$ |

There was a significant interaction between complexity and anxiety, $\mathrm{F}(1,88)=5.724, \mathrm{p}=0.019, \eta_{\mathrm{p}^{2}}=0.019$. The impact of complexity on accuracy was only significant for the control condition (Figure 4).
Interestingly, the $2 x 2$ Repeated Measures ANOVA on response time revealed a main effect of complexity ( F $(1,88)=6.606, p=0.012, \eta_{p^{2}}=0.070$ : response times were less for 2 alternatives than for 4 alternatives condition, means of 4255.093 msec and 4692.856 msec , respectively) but not of anxiety $\left(F(1,88)=0.027, p=0.871, \eta p^{2}=0.000\right)$. Besides, the interaction between complexity and anxiety was not significant: $F(1,88)=0.274, p=0.602, \eta_{p^{2}}=0.003$. Thus, the accuracy data are not explainable in terms of performance time, they are rather due to real differences in processing.


Figure 4: Mean number of relational mappings per condition.

To sum up, the obtained difference in accuracy but not in response time data only for the control group speaks in favor of processing that compensate the complexity of the task in high state anxiety. An increase in the number of alternatives slows the answers down significantly for both the control and anxiety groups. The accuracy, however, differs. If 4 alternatives require a higher level of inhibition to come up with a relational mapping as the results in the control condition seem to suggest, then the logical result would be that anxiety would diminish the accuracy for more complex tasks. Moreover, as Holyoak (2012) has pointed out, if anxiety reduces the available cognitive resources, crucial for the inhibition of irrelevant information, then it
would hamper analogical mapping even more than in the control condition. That seems not to be the case in our data: participants in the anxiety condition do not differ in terms of accuracy when searching for a relational match among 2 or 4 alternatives.

However, the fact that state anxiety fosters encoding of objects as well as relations (Hristova \& Kokinov, 2011) may suggest a plausible explanation of the obtained results. When the relational match is chosen between 2 and 4 alternatives, it is not only that the mapping becomes harder with the number of the available alternatives, the number of required relations that should be considered also increases. Hence, the state anxiety group would have an appreciable encoding superiority over the control group in the case of four alternatives: they will encode the necessary relations faster and may use the time for resolving the harder mapping. In other words, the superior encoding of the anxiety group may compensate for the difficulties in the subsequent mapping, associated with the more complex task. Of course, it is possible, actually quite probable, that the improvement of relational encoding and the suggested impoverished inhibition due to the state anxiety, depend on the level of anxiety. On one hand that may explain the inconsistency between the reported results and Tohill and Holyoak's data (2000, Experiment 2): the difference between the anxiety levels in the control and anxiety groups in their experiment was almost twice the difference in the current study. That might indicate quite different outcomes with regard to the Yerkes-Dodson Law (1908).

The experiment described here, however, points to an interesting interplay between the subprocesses of analogy making under anxiety.

## Conclusion

The data from this experiment suggest that anxiety influences differentially the encoding of relations and the inhibition of alternative hypotheses, which are crucial for the final analogical mapping. Complexity does change relational mapping under low but not under high state anxiety.

In a recent paper Vendetti, Knowlton, and Holyoak (2012) varied the semantic distance between analogical domains and showed that anxiety does not decrease the number of correct relational responses, but increases the number of false alarms in verbal $\mathrm{A}: \mathrm{B}:: \mathrm{C}: \mathrm{D}$ analogies. This was interpreted as switching to a non-analogy strategy which looks for the superficial overlap of the domains, rather than their structure. That explanation, however, does not seem applicable to our data, since the superficial overlap between the two structures represented in the two pictures of each stimulus pair are identical, but differ only in the number of relations that should be considered in the case of 2 and 4 alternatives.

Finally, the current conflict between the data in the field can be explained both via differences in the complexity of the tasks and in the obtained level of state anxiety, which in turn highlights the importance of experiments that
manipulate anxiety on at least three levels (in order to capture a particular nonlinear relationship between anxiety and performance on analogy tasks), while controlling for the complexity of the given task.

## Acknowledgments

We express our gratitude to Art Markman and Keith Holyoak, who gave us the opportunity to work with the original stimuli used in the studies, respectively of Markman and Gentner (1993) and Tohill and Holyoak (2000). We also thank the artist Kristian Sarbinov who redrew all the pictures and drew the new ones under our guidance. This study, would be impossible without the work of our experimenters Dilyana Nikolova and Pavlina Petrakieva, who did their best to apply the anxiety induction procedure without knowing its purport.

Finally, we would like to thank the anonymous reviewers for their helpful comments and suggestions.

This work has been supported by the European Office for Aerospace Research and Development under grant FA8655-10-1-3061.

## References

Becker, M. (2009). Panic Search: Fear produces efficient visual search for nonthreatening objects. Psychological Science, 20(4), 435-437.
Broadhurst, P.L. (1957). Emotionality and the YerkesDodson law. Journal of Experimental Psychology, 54, 345-352.
Cho, S., Holyoak, K. J., \& Cannon, T. D. (2007). Analogical reasoning in working memory: Resources shared among relational integration, interference resolution, and maintenance. Memory and Cognition, 35(6), 1445-1455.
Eysenck, M., \& Calvo M. (1992). Anxiety and performance: The processing efficiency theory. Cognition and Emotion, 6, 409-434.
Eysenck, M. W., Deraksan, N., Santos, R., \& Calvo, M. G. (2007). Anxiety and cognitive performance: Attentional control theory. Emotion, 7, 336-353.
Feldman, V., Hristova, P., \& Kokinov, B. (2010). How does anxiety influence analogical mapping? Proceedings of the 32nd Annual Conference of the Cognitive Science Society. Hillsdale, NJ: Erlbaum.
Holyoak, K. J. (2012). Analogy and relational reasoning. In K. J. Holyoak \& R. G. Morrison (Eds.), The Oxford handbook of thinking and reasoning. New York: Oxford University Press.
Hristova, P., \& Kokinov, B. (2011). Anxiety fosters relational encoding. In: B. Kokinov, A. Karmiloff-Smith, \& N. Nersessian (Eds.), European Perspectives on Cognitive Science. Sofia: NBU Press.
Krawczyk, D. C., Morrison, R. G., Viskontas, I., Holyoak, J., Chow, T. W., Mendez, M. F., Miller, B. L., \& Knowlton, B. J. (2008). Distraction during relational reasoning: The role of prefrontal cortex in interference control. Neuropsychologia, 46, 2020-2032.

Leon, M., \& Revelle, W. (1985). Effects of anxiety on analogical reasoning: A test of three theoretical models. Journal of Personality and Social Psychology. 49(5), 1302-1315.
Markman, A. B., \& Gentner, D. (1993). Splitting the differences: A structural alignment view of similarity. Journal of Memory and Language, 32, 517-535.
Pacheco-Unguetti, A., Acosta, A., Callejas, A., \& Lupianez, J. (2010). Attention and anxiety: Different attentional functioning under state and trait anxiety. Psychological Science, 21(2): 298-304.
Phelps, E.A., Ling, S., \& Carrasco, M. (2006). Emotion facilitates perception and potentates the perceptual benefits of attention. Psychological Science, 17, 292-299.
Posner, M.I., \& Petersen, S.E. (1990). The attention system of the human brain. Annual Review of Neuroscience, 13, 25-42.
Posner, M.I., Rueda, M.R., \& Kanske, P. (2007). Probing the mechanisms of attention. In J.T. Cacioppo, J.G. Tassinary, \& G.G. Berntson (Eds.), Handbook of psychophysiology. Cambridge, England: Cambridge University Press.
Richert, R., Whitehouse, H., \& Stewart, E. (2005). Memory and analogical thinking in high arousal rituals. In: H. Whitehouse \& R.H. McCauley (Eds.), Mind and religion: psychological and cognitive foundations of religiosity. Walnut Creek, CA: AltaMira Press.
Spielberger, C.D., Gorsuch, R.L., \& Lushene, R. (1970). STAI manual for the State-Trait Anxiety Inventory. Palo Alto, CA: Consulting Psychologists Press.
Sweis, B.M., Bharani, K.L. \& Morrison, R.G. (2012). The time course of inhibition in analogical reasoning: An event-related potential approach. In N. Miyake, D. Peebles, \& R. P. Cooper (Eds.), Proceedings of the 34th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society, 1030-1035.
Tohill, J., \& Holyoak, K. (2000). The impact of anxiety on analogical reasoning, Thinking \& Reasoning, 6:1, 27 - 40.
Vendetti, M., Knowlton, B., \& Holyoak, K. (2012). The impact of semantic distance and induced stress on analogical reasoning: A neurocomputational account. Cognitive, Affective \& Behavioral Neuroscience, 12, 804812.

Viskontas, I. V., Morrison, R. G., Holyoak, K. J., Hummel. J. E., \& Knowlton, B. J. (2004). Relational integration, inhibition, and analogical reasoning in older adults. Psychology and Aging, 19(4), 581-591.
Yerkes, R. M., \& Dodson, J. D. (1908) The relation of strength of stimulus to rapidity of habit-formation. Journal of Comparative Neurology and Psychology, 18, 459-482.
Щетински, Д., Паспаланов, И. (1989). Методическо пособие за работа с българската форма на въпросника за оценка на тревожността на Ч.Спилбъргър. БАН - Институт по психология. (Bulgarian standardized version of Spielberger's STAI).

# A Neural Model of Human Image Categorization 

Eric Hunsberger (ehunsberger@uwaterloo.ca)<br>Peter Blouw (pblouw@uwaterloo.ca)<br>James Bergstra (jbergstra@uwaterloo.ca)<br>Chris Eliasmith (celiasmith@uwaterloo.ca)<br>Centre for Theoretical Neuroscience<br>University of Waterloo, Waterloo, ON, Canada N2L 3G1


#### Abstract

Although studies of categorization have been a staple of psychological research for decades, there continues to be substantial disagreement about how unique classes of objects are represented in the brain. We present a neural architecture for categorizing visual stimuli based on the Neural Engineering Framework and the manipulation of semantic pointers. The model accounts for how the visual system computes semantic representations from raw images, and how those representations are then manipulated to produce category judgments. All computations of the model are carried out in simulated spiking neurons. We demonstrate that the model matches human performance on two seminal behavioural studies of image-based concept acquisition: Posner and Keele (1968) and Regehr and Brooks (1993).


Keywords: category representation; image categorization; neural engineering framework; vector symbolic architecture

## Introduction

Although studies of categorization have been a staple of psychological research for decades, there continues to be substantial disagreement about how the mind represents information about unique classes of objects. Theories involving prototypes, exemplars, and explanatory schemas have all been shown to account for only a subset of known categorization phenomena, and progress toward a unified theory of category representation has been limited (for reviews, see Murphy, 2002; Machery, 2009; Smith \& Medin, 1981). Historically, the difficulty in modelling category representation has been to balance generality and accuracy.

On one hand, many of the models developed from these theories have a fairly narrow scope of application. Consider, for instance, similarity-based accounts of concept reference; these models produce impressive results at matching human behaviour in tasks that involve feature comparisons (see Smith \& Medin, 1981), but they do not generalize well to other tasks that require the use of deeper category knowledge or explanatory inferences (see Murphy, 2002).

On the other hand, approaches with greater scope tend to pay a price in terms of predictive accuracy or viability. For example, Barsalou's (1999) theory of perceptual symbol systems is a more or less unified account of category representation, but it lacks a corresponding computational model (Dennett \& Viger, 1999). Rogers and McClelland's (2004) account of semantic cognition provides a powerful model that performs well across a range of categorization tasks, but employs both an idealized neural architecture and an idealized set of inputs (i.e. it is an abstract connectionist network that
does not use raw percepts as input). Many researchers now recognize that object perception and conceptual cognition are not distinct (Palmeri \& Gauthier, 2004), making it important that models integrate both perception and cognition.

In this paper, we argue that advances in our understanding of the visual system and new principles for the design of neural architectures can be used to overcome many of the difficulties in providing a viable, neurally grounded, computational model of image categorization. We use the techniques of the Neural Engineering Framework (NEF) (Eliasmith \& Anderson, 2003) to develop a model of category representation that connects retinal activity to high level cognitive judgments using a class of vector-symbolic representations called semantic pointers (Eliasmith et al., 2012). The model receives natural images as input, produces category judgments as output, and carries out intermediate processing in simulated neurons. The proposed model replicates human performance on two independent studies of human judgment in prototype-based and exemplar-based image categorization, with no changes to the model. Semantic pointer architectures have been shown to support several important cognitive capacities (e.g. Stewart \& Eliasmith, 2011; Eliasmith et al., 2012). Our study extends this line of research, showing that semantic pointers computed by a plausible visual system model can be used to replicate human category judgments.

## Model Description

We developed a model of human image categorization that consists of a feed-forward visual perception model (similar to Hinton \& Salakhutdinov, 2006) driving a vector-symbolic associative memory (see Gayler, 2003; Plate, 2003). The model was first constructed using a rate approximation of the spiking leaky integrate-and-fire (LIF) neuron model for the visual system and explicit vector computations for the associative memory. The model was then implemented fully in spiking neurons using the principles of the NEF.

The visual system component of the model is a sequence of feed-forward neural populations that compresses high dimensional input images into comparatively low dimensional vectors, which we refer to as semantic pointers. The first population, analogous to the retina and lateral geniculate nucleus (LGN), is a rasterized image, as would be captured by a conventional digital camera. Like the retina, a camera adapts to global lighting conditions and provides an image with standard intensity levels. Our (small) LGN population corre-
sponds to a square $30 \times 30$ image region that is best compared to a small portion from the centre of the visual field. The second population, analogous to visual area V1, comprises 2500 neurons with local connectivity: each neuron responds to a randomly chosen $9 \times 9$ patch in LGN. Neurons in the third (V2), fourth (V4), and fifth (inferotemporal (IT) cortex) populations (with size 900,400 , and 225 respectively) are connected to all neurons in the previous population. The activation pattern in the fifth population (with latency similar to visual area IT) is the semantic pointer representing the image stimulus. Representations generated in this manner are stored in an associative memory as category exemplars (during training), and used to probe the associative memory to yield a category judgment during testing (see Figure 2).

## Adaptation to Natural Images

A large fraction of neuron cells in visual area V1 are well modelled as luminance edge detectors (Hubel \& Wiesel, 1968; DiCarlo, Zoccolan, \& Rust, 2012). There is mounting evidence that visual system neurons behave as they do because they continuously adapt to statistical patterns in visual stimuli (Olshausen \& Field, 1996; Hyvärinen, 2009). Computer vision systems inspired by principles of adaptation to unlabelled visual images are among the best-performing computer vision systems, and reproduce several phenomena discovered by electrode recordings (Lee, Ekanadham, \& Ng, 2008; Le et al., 2012). One strategy for adaptation to unlabelled images is the autoencoder (Ackley, Hinton, \& Sejnowski, 1985; Rumelhart, Hinton, \& Williams, 1985), which was first applied to images by Cottrell, Munro, and Zipser (1987).

The connection weights of our visual system model were trained as a deep autoencoder, with an additional $\ell_{1}$ penalty on the hidden node activation rates to model the energy cost of spiking and encourage sparser activation patterns. The objective function for one layer is given by

$$
\begin{equation*}
O=\frac{1}{K} \sum_{i, k}\left(x_{i}^{(k)}-y_{i}^{(k)}\right)^{2}+\lambda \sum_{j}\left|q_{j}-\rho\right| \tag{1}
\end{equation*}
$$

where $x_{i}^{(k)}$ is the value of visual node $i$ for example $k, y_{i}^{(k)}$ is the autoencoder's reconstruction of node $i$ example $k, q_{j}$ is a running average of the activation of hidden node $j$, and $\lambda$ and $\rho$ control the importance of sparsity and the desired sparsity level, respectively. Uniquely, our autoencoder used an LIF response function as the feature activation function.

The autoencoder was trained on random $30 \times 30$ natural image patches chosen from the first 10 images of the van Hateren Natural Image Database (van Hateren \& van der Schaaf, 1998). with each patch normalized to zero mean and unit variance. We trained only on un-whitened images, which contain the full spectrum of spatial frequencies. We found that whitening was not required to extract Gabor-like filters from the statistics of the natural images (Figure 1), and was in fact undesirable since it removed some low-frequency features important for classification.


Figure 1: Filters from the first layer of the visual system, after autoencoder training on natural images. Like neurons in area V1, our model neurons detect luminance edges at a variety of frequencies and orientations.

Like Hinton and Salakhutdinov (2006), each layer of the autoencoder was pretrained individually; however, layers were pretrained as autoencoders, not restricted Boltzmann machines, allowing us to use an LIF response function for the neuron nonlinearity. The layers were then combined into a deep autoencoder and trained together using backpropagation.

## Semantic Pointers: Memory and Retrieval

We refer to the vectors processed by the model as semantic pointers because they function as compressed symbol-like representations that encode the statistically relevant aspects of the information they represent (Eliasmith, in press). In the non-visual component of the architecture, semantic pointers representing the compressed images are bound with category labels using the mathematical operation of circular convolution (see e.g. Plate, 2003). Subsequently, the bound representations are added to the memory via superposition. This process is captured formally by Equation 2:

$$
\begin{equation*}
M=\sum_{i=1}^{N}\left(P_{i} * L_{i}\right) \tag{2}
\end{equation*}
$$

where $P_{i}$ is a semantic pointer produced by the visual system from the $i^{\text {th }}$ raw image, $L_{i}$ is a vector representing the category label associated with the image, $M$ is the memory pointer, and $*$ is the circular convolution operator.

Once the memory is built up with a number of learned category exemplars, it can be used to produce categorization judgments in response to novel input images via the use of an inverse of the convolution operation. This inverse operation probes the memory for the category label that is most likely to fit the input image on the basis of prior learning. As a whole, the categorization process conforms to the following mathematical description:

$$
\begin{equation*}
c=\operatorname{argmax}_{c}\left[\left(P^{-1} * M\right) \cdot L_{c}\right] \tag{3}
\end{equation*}
$$

where $c$ refers to the resulting category judgment, $P^{-1}$ refers to the pseudoinverse of the semantic pointer corresponding to the test image, $L_{c}$ refers to the label pointer of category $c$, and $a \cdot b$ refers to the dot product of $a$ and $b$. For the rate model, these operations were implemented directly in vectors; for the spiking model, the operations were implemented in spiking LIF neurons using the NEF.


Figure 2: The schematic of our visual categorization model has three components. Left: The visual system comprises four populations of leaky integrate-and-fire neurons corresponding to the LGN, visual areas V1, V2, V4, and IT, which we take to represent a semantic pointer. The connections between these populations are adapted to natural scene statistics by unsupervised learning. Upper Right: The memory of our model is encoded as a single semantic pointer, which is the sum of several labelled training patterns (three are shown here). Labels have been bound to their corresponding image representations through the mathematical operation of circular convolution. Lower Right: At test time, our model labels visual stimuli by deconvolving the activity patterns of IT with the memory vector, and matching the result against several possible label decisions.

In short, the model builds category representations by storing compressed and labelled percepts, and produces categorization judgments by evaluating the similarity between an input percept and the exemplars stored in memory. However, since all labelled percepts are compressed into the same vector, there is significant interaction between stored percepts; this can be likened to creating a prototype based on the percepts. The model categorization system thus falls part way in between pure exemplar-based categorization and pure prototype-based categorization; it has elements of both.

## Experiment 1: Prototype-based Categorization

To account for the sort of phenomena that have traditionally motivated prototype theories of category representation, we tested the model on a task from Posner \& Keele's (1968) classic study of pattern classification. We chose to model Experiment 3 of the study, which was designed to test whether human subjects are learning about class prototypes when they only ever see distorted examples. In the study, subjects are trained to classify classify random dot patterns into three mutually exclusive categories. Each pattern consists of nine dots dispersed over a $30 \times 30$ grid, with each dot occupying one cell in the grid. The patterns used for training are generated from three prototypes; each training pattern is created by choosing a prototype pattern, and moving each dot according to a random distortion rule (see Figure 3.) Thirty (30) subjects were trained by corrective feedback to classify twelve 'high distortion' patterns (four from each category). After training, the subjects were asked to classify twenty-four pat-
terns without feedback: patterns from the training phase (2 per prototype, 6 total), the prototype patterns (3), prototype patterns with a smaller degree of distortion (6), new highly distorted prototype patterns (6), and entirely random new patterns (3). Subjects were tested on these patterns on two consecutive days, in terms of both accuracy and reaction time.

The protocol for evaluating our categorization model was nearly identical. We presented the model with the twelve training images, and it stored the semantic pointers associated with the labels and the images into the model's memory (see Figure 2). Then we presented our model with each of the twenty-four testing patterns. Figure 4 compares the accuracy of our model to the classification accuracy of the human subjects. Since our model lacks motor output, we did not evaluate it on reaction time. Figure 4 shows the results of our model; in sum, the model performs much like the human subjects.

## Experiment 2: Exemplar-based Categorization

To account for effects more commonly aligned with exemplar theories of category representation, we tested the model on a task from Regehr \& Brook's (1993) study of the comparative influence of analytic and non-analytic processes on categorization behaviour. The study involves a number of experiments in which subjects are asked to classify simple drawings of imaginary animals into one of two categories. The animals all possess an analytic structure that varies along five binary dimensions (e.g. a round vs. angular body), but the exact perceptual manifestation of a particular dimension value (i.e.


Figure 3: Sample stimuli for Experiment 1, modelling a classic study by Posner \& Keele (1968). The dot patterns are created by distorting three randomly drawn prototype images (left) with low (centre) and high (right) levels of noise. Subjects are trained to classify a set of twelve high-distortion patterns and tested without feedback on the same prototypes at different distortion levels.
feature) can vary across animals. For example, two animals might have round bodies, and thus be analytically equivalent to some extent, but the actual roundness of their respective bodies might be quite distinct (see Figure 5). This allows for the construction of stimuli sets that possess drawings that are analytically equivalent but perceptually dissimilar, along with drawings that are analytically distinct but perceptually similar. By training subjects through corrective feedback to classify these images into categories defined by an analytic rule, Regehr \& Brooks were able to test hypotheses regarding the relative importance of perceptual similarity and analytic structure during categorization.

In the experiment 1C of Regehr \& Brooks' study, 32 subjects are placed into one of two conditions and then trained to classify a set of eight images into two categories. For subjects in the first condition, the perceptual manifestations of a given dimension are constant across the images (See Figure 5, left). For subjects in the second condition, the perceptual manifestations of a given dimension vary across images (See Figure 5 , right). Every subject was trained to learn one of four labelling rules based on analytic structure. The rules had the form: An image is a 'builder' if it has at least two of X, Y, and spots, otherwise it is a 'digger'. The criteria X and Y referred to things like "long neck", "angular body", "short legs" and so on (see Regehr \& Brooks, 1993, for details). Training occurs through corrective feedback and is considered complete after five runs through the image set.

During the transfer phase of the experiment, subjects are asked to classify a set of sixteen images, eight of which are from the training set and eight of which are qualitatively similar, but new. The new images have been designed to pair up with a specific training image, and only differ on the dimension of "spots on body." Half of the new images belong in the same category as their twin from the training set, while the other new images have a different category from their twin. The idea motivating this experimental design is that if sub-


Figure 4: Comparison of human and model performance for Experiment 1. The model is able to account for human results when presented with the schema, low distortion (5), and high distortion (7) patterns. Occasional random errors by human subjects may explain the discrepancy on training examples. Error bars indicate $95 \%$ confidence intervals. Human data from Posner \& Keele (1968).
jects attend primarily to the analytic structure of the images during testing, then they should make relatively few errors on the new bad transfer items (because both analytic structure and perceptual similarity favour the correct judgment). Alternatively, if subjects attend primarily to similarity to past exemplars, then they should make relatively more errors on the bad transfer items (because perceptual similarity and analytic structure favour opposing judgments). The study is designed to test the effect of appearance on subjects' use of structural vs. perceptual mental representations.

We model experiment 1C of Regehr \& Brooks' study with the same model that we used in Experiment 1. We presented our model with the same eight training images used in the original experiment (though downsampled to fit in a $30 \times 30$ patch), drawn either in the composite style or in the individuated style. The semantic pointers created by the visual system, together with semantic pointers for the corresponding image labels, were stored into the model's memory, as described by Equation 2 shown in Figure 2. We tested the representations of our visual system by classifying the goodtransfer and bad-transfer test images, as well as the original training images. The accuracy of our model in each case is presented in Figure 6. Our model provides a good match to human performance, and replicates the effect that perceptually individuated stimuli foster substantially different error profiles than perceptually un-individuated stimuli.

Individuated Parts




Figure 5: Sample stimuli for Experiment 2, modelling experiment 1C of Regehr and Brooks (1993). (Left) Images are composed of interchangeable (composite) feature manifestations. (Right) Images expressing the same attributes are drawn in a more coherent (individuated) style. Regehr \& Brooks (1993) drew a distinction between good transfer and bad transfer test stimuli. A test stimulus is a good transfer case when the addition or removal of spots matches a training case with the same label, and a bad transfer case if adding or removing spots matches a training case with the opposite label. (Adapted from Regehr \& Brooks (1993) Figure 2A).


Figure 6: Comparison of human and model performance for Experiment 2. Our model accounts for the key difference in human performance on the good transfer (GT) versus bad transfer (BT) pairs for the individuated stimuli. Error bars indicate $95 \%$ confidence intervals. Human data from Regehr \& Brooks (1993).

## Discussion

Posner \& Keele's (1968) study is considered to be seminal in the development of prototype theory, and the result that subjects categorize the training patterns and prototype patterns equally well is taken to indicate that the subjects are abstracting information about the prototypes during the training phase. Our model's replication of this performance provides good evidence that our approach is capable accounting for the sort of prototype effects that the study uncovered. Interestingly, the spiking version of the model performs slightly worse than humans on the prototypes, indicating that it might be performing a more exemplar-based classification. However, we hypothesize that adding more neurons to the associative memory will attenuate this effect.

Regehr \& Brooks' (1993) study is more easily located in the tradition of exemplar theories of category representation. The fact that the model replicates the effects of interference from exemplar memories on more analytic categorization approaches suggests that it is well-equipped to deal exemplarbased phenomena. Moreover, the architecture of the model almost trivially assures that this is true-the contents of the associative memory essentially are exemplars produced from visual experience. It is thus reasonable to expect that phenomena found in studies using different kinds exemplars will be reproducible with the model.

Overall, the results of the simulations indicate that our model is able to account for an important set of phenomena closely associated with exemplar and prototype theories of category representation. The fact that the simulation employs a neural architecture for all stages of processing, and that it begins with raw image input, provides an important contribution to the current literature.

However, the model has several limitations as it stands. Nevertheless, we believe it is reasonable to expect that the architecture is capable of capturing an even wider range of phenomena. We identify two requirements of scaling that an architecture utilizing semantic pointers can likely achieve.

For one, it is possible to incorporate a more realistic account of category learning into the model. In the actual experiments, subjects learn the relevant categories through corrective feedback, and the feedback process continues either for a set number of trials, or until the subjects can accurately classify all of the items without error. By comparison, our model learns by memorizing a set of training images labelled with the correct category. In the model, the label/image relationships are forgotten when the model is shown another set of stimuli. However, the recent development of methods for introducing biologically plausible learning rules into the neural framework we employ indicates that this simplification could be removed in the future. Other cognitive models that make use of semantic pointers have already incorporated a form of reinforcement learning using such rules (e.g. Eliasmith et al., 2012).

Second, we believe it is possible to account for a broader range of categorization phenomena. The architectural prin-
ciples used in our model have also been used to construct what is currently the world's largest functional brain model, able to account for tasks involving serial-order recall, syntactic induction, and the manipulation of numerical concepts (Eliasmith et al., 2012). The fact that other large-scale cognitive models make use of the same representations and processes as this model provide good reason to think that a similar scale of functionality can achieved with models specifically focused on category representation. These two extensions will be the focus of future work.

## Conclusion

This paper has presented a neural architecture for categorizing visual stimuli using a semantic pointer architecture. Our model replicates human behaviour on two important studies of visual categorization: Posner \& Keele's (1968) and Regehr \& Brooks’ (1993). Modelling efforts have traditionally had to face the dilemma of choosing between plausibility and scope. The end-to-end neural model described here takes a suggestive first step in addressing this dilemma. Overall, this promise of scalability adds further theoretical significance to the empirical results we describe. The combination of a hierarchical visual model and a neurally implemented vector-symbolic architecture yields a new, effective approach to building models of category representation that are scalable, biologically plausible, and comprehensive, in that they capture the stages of processing from perception to judgment.

## Acknowledgments

We gratefully acknowledge the financial support of NSERC, SSHRC, Canada Research Chairs, the Canadian Foundation for Innovation, the Ontario Innovation Trust, and the Banting Postdoctoral Fellowships Program.

## References

Ackley, D. H., Hinton, G. E., \& Sejnowski, T. J. (1985). A Learning Algorithm for Boltzmann Machines. Cognitive Science, 9(1), 147-169.
Barsalou, L. (1999). Perceptual symbols systems. Behavioral and Brain Sciences, 22, 577-660.
Cottrell, G., Munro, P., \& Zipser, D. (1987). Learning internal representations from gray-scale images: An example of extensional programming. In Proc. COGSCI 9 (pp. 462473).

Dennett, D., \& Viger, C. (1999). Sort-of symbols? Behavioral and Brain Sciences, 22(4), 613.
DiCarlo, J. J., Zoccolan, D., \& Rust, N. C. (2012). How does the brain solve visual object recognition? Neuron, 73, 415-34.
Eliasmith, C. (in press). How to build a brain: A neural architecture for biological cognition. Oxford, UK: Oxford University Press.
Eliasmith, C., \& Anderson, C. (2003). Neural engineering: Computation, representation, and dynamics in neurobiological systems. Cambridge, MA: MIT Press.

Eliasmith, C., Stewart, T., Choo, F.-X., Bekolay, T., DeWolf, T., Tang, Y., et al. (2012). A large-scale model of the functioning brain. Science, 338(6111), 1202-1205.
Gayler, R. (2003). Vector symbolic architectures answer Jackendoffs challenges for cognitive neuroscience. In P. Slezak (Ed.), ICCS/ASCS (p. 133-138).

Hinton, G. E., \& Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. Science (New York, N.Y.), 313(5786), 504-7.
Hubel, D., \& Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. Physiology, 195(1), 215-43.
Hyvärinen, A. (2009). Statistical models of natural images and cortical visual representation. Topics in Cognitive Science, 2, 251-264.
Le, Q. V., Ranzato, M. A., Monga, R., Devin, M., Chen, K., Corrado, G. S., et al. (2012). Building high-level features using large scale unsupervised learning. In Proc. ICML 29 (pp. 81-88).
Lee, H., Ekanadham, C., \& Ng, A. Y. (2008). Sparse deep belief net model for visual area V2. In Proc. NIPS 20 (pp. 873-880).
Machery, E. (2009). Doing without concepts. Oxford, UK: Oxford University Press.
Murphy, G. (2002). The big book of concepts. Cambridge, MA: MIT Press.
Olshausen, B. A., \& Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature, 381, 607-691.
Palmeri, T. J., \& Gauthier, I. (2004). Visual object understanding. Nature reviews. Neuroscience, 5(4), 291-303.
Plate, T. (2003). Holographic reduced representation: Distributed representation for cognitive structure. Stanford, CA: CSLI Publications.
Posner, M., \& Keele, S. (1968). On the genesis of abstract ideas. Experimental Psychology, 77, 653-663.
Regehr, G., \& Brooks, J. (1993). Perceptual manifestations of analytic structure - the priority of holistic individuation. Experimental Psychology: General, 122(1), 92-114.
Rogers, T., \& McClelland, J. (2004). Semantic cognition: A parallel distributed processing approach. Cambridge, MA: MIT Press.
Rumelhart, D., Hinton, G., \& Williams, R. (1985). Learning internal representations by error propagation (Tech. Rep. No. ICS-8506). UCSD Institute for Cognitive Science.
Smith, E., \& Medin, D. (1981). Categories and concepts. Cambridge, MA: Harvard University Press.
Stewart, T., \& Eliasmith, C. (2011). Neural cognitive modeling: A biologically constrained spiking neuron model of the tower of hanoi task. In Proc. COGSCI 33 (p. 656-661).
van Hateren, J. H., \& van der Schaaf, A. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex. Proc. Biological Sciences, 265(1394), 359-366.

# How to improve on Quinian bootstrapping - a response to nativist objections 

Zoltan Jakab (zoltan.jakab@barczi.elte.hu)<br>Institute for Psychology of Special Needs, Loránd Eötvös University, Ecseri út 3 1097 Budapest, Hungary


#### Abstract

Quinian bootstrapping is Susan Carey's solution to Fodor's paradox of concept learning. Carey claims that contrary to Fodor's view, not all learning amounts to hypothesis testing, and that there are ways in which even primitive concepts can be learned. Recently Georges Rey has argued that Carey's attempt to refute radical concept nativism is unsuccessful. First it cannot explain how the expressive power of mental representational systems could increase due to learning. Second, both Fodorian circularity charges and Goodmanian problems of indeterminacy apply to Carey's examples of Quinian bootstrapping. I argue that Carey's examples of bootstrapping can be amended to escape Fodorian and Goodmanian objections. I suggest some ways to improve on our models of concept learning to this end. I also argue that skill learning is the way for mental representational systems to increase their own expressive power, that is, to enrich their conceptual repertoire beyond what compositionality alone affords.


Keywords: Quinian bootstrapping; nativism; concept learning; expressive power

## Introduction: Fodor's paradox

Jerry Fodor's famous argument for concept innateness has taken different forms. According to its early version (Fodor, 1975, 1981), all learning is hypothesis testing, and to formulate a hypothesis one needs to possess all the concepts that the hypothesis involves. Therefore, in the course of learning one can seek for evidence supporting or undermining a hypothesis, but formulating hypotheses will never result in the acquisition of new concepts. In order to work, this argument needs an immediate qualification. Some concepts are structured, whereas others are not; they are primitive. At any rate, this distinction holds on Fodor's own representational theory of mind which endorses compositionality. Complex concepts arise as combinations of primitives (or simpler complexes). For example, WHITE RAVEN ${ }^{1}$ is a complex concept which has two constituents: WHITE and RAVEN; the two constituents are related by a conjunction. Now if one wishes to test the hypothesis that White ravens are quite rare, and it is conceded that forming WHITE RAVEN out of WHITE and RAVEN counts as learning a new concept, then evidently hypothesis formation makes room for concept learning. For more complex cases the idea that compositionality affords concept learning does

[^97]often sound intuitively plausible. ${ }^{2}$ Hence Fodor's early view on concept acquisition, according to which primitive concepts cannot be learned, and so they must be innate. In Fodor $(1990,1998)$ this idea is supplemented by argument that most of our concepts are primitive, giving rise to radical concept nativism.

More recently, Fodor (2008) has found this conclusion much too weak, and formulated a stronger version according to which the obstacle to concept learning is not that most concepts are primitive, but rather that compositionality cannot increase the expressive power of cognitive systems. Very roughly, expressive power is the range of concepts and hypotheses (theories, conceptions) that a given representational system could formulate, or express, given its primitive symbols (concepts) and rules of combination. As we currently understand cognition, any case of learning seems to be underlain by some mental process that exploits compositionality: forming complex mental representations out of simpler ones governed by rules, plus adjusting certain parameters of the primitives. ${ }^{3}$ If this is how cognition operates, then all that learning can achieve is the manifestation of what's born with us: we actually come to express what we are innately capable of expressing. In formal logic, building new complexes out of primitives (symbols and rules of combination) does not count as increase in expressive power - only adding certain new primitives ${ }^{4}$ does. The same restriction seems to apply in the realm of mental representation, if indeed compositionality is the only game in town for cognitive theorizing.

In sum, since no cognitive mechanism that we can currently think of transcends compositionality, and compositionality cannot increase expressive power beyond one's innate endowment, no cognitive mechanism can increase the expressive power of mental representational

[^98]systems. Now if genuine concept learning requires increase in expressive power, then no cognitive mechanism that we can currently think of amounts to genuine concept learning - only triggering of some sort. Thus the challenge is: supply a theory of cognition which can somehow pass beyond the limits of ordinary compositional systems, and is a candidate for expressive power increase. This seems like a challenge that is next to impossible to meet. This is how I understand Fodor's claim (2008) that it is true and a priori that the whole idea of concept learning is per se confused.

In the rest of this paper I first look at Susan Carey's (2009) response to Fodor's challenge. Then I briefly present Georges Rey's (2012) critique of Carey's views. Finally, I suggest some ways to address Rey's and Fodor's critique and reinstate concept learning roughly along the lines Carey proposed.

## Quinian bootstrapping

Susan Carey (2009) makes the following claims about concept acquisition. First, some of our concepts are innate and modular in origin, whereas the majority of our concepts are constructed out of the innate resources and originate from the explicit knowledge systems (roughly, domain general central systems). Second, the newly constructed concepts have more expressive power than the innate system. Third, Carey suggests a mechanism of acquisition for the second set of concepts that is allegedly more powerful than ordinary compositionality based on formal logic. This mechanism is called Quinian bootstrapping (hereafter QBS), and a number of examples of it are supplied throughout Carey's book. The general scheme of QBS is that in trying to understand some new ideas (e.g., fractional numbers, or density as mass per volume), children first form an ordered set of empty placeholders ("mental symbols") which are then gradually filled up with relevant content as a result of analogical reasoning. Empirical evidence for transition from an initial to a more powerful set of concepts involves within-child consistency of performance over a wide range of relevant tasks, and observations that the acquisition of the new set of concepts is difficult. That is, initially children try to assimilate the new terms to their earlier concepts - for instance, they take numerals to be quantifiers of some sort.

Carey's answer to Fodor's argument in particular is that QBS constitutes concept learning, because it does not consist in construction from antecedently available concepts using the machinery of compositional semantics alone. The newly bootstrapped concepts are definitional primitives in Carey's view, therefore they cannot arise from logical construction. Carey takes this line of argument to undermine Fodor's claims that (i) all learning mechanisms reduce to hypothesis formation and testing, and that (ii) the relevant hypotheses must be formed in terms of available concepts via compositional semantics. In addition to the details she herself provides, Carey endorses other proposals that primitive concepts can be learned (Margolis, 1999; Margolis \& Laurence, 2002).

## Nativist objections redrawn

Rey (2012) argues that even though Carey supplies valuable data on how children in fact acquire concepts, she fails to meet Fodor's nativist challenge. This is so because she conflates certain semantic issues, namely how expressive power might increase, with epistemological ones, that is, cognitive accounts of mental representation and its development. In logic, expressive power increases only if new primitives (e.g., operators, predicates) are added to a logical system - primitives which cannot be expressed by combinations of the antecedently available ones. The development of mental representation, on the other hand, includes episodes in which new complexes are formed out of certain innately available primitives. So how could QBS increase the expressive power of mental representational systems?

In addressing this issue, Rey points out an interesting parallel between Quinian bootstrapping and Ramsey sentences. Ramsey sentences are logical formulas that serve to define certain new theoretical terms by specifying their relations to already given old terms. A Ramsey sentence is a huge formula that is a conjunction of all the claims made by some theory. This conjunction involves all the relevant concepts: old ones, and new ones which are to be theoretically defined by the entire formula. The Ramsey sentence contains a unique existential quantification of each of the new terms. For example, on certain varieties of functionalism about the mind, types of mental states are characterized by the relations that they exhibit to other states (mental states, stimuli, and behaviors). Here the old terms are those of stimuli and behaviors; the new ones are those of mental states. For instance, the mental state of hunger is elicited by certain bodily conditions, it tends to evoke thoughts about food and food-seeking, and food-seeking behavior (unless some higher-order motivational states like the desire to lose weight cancel that behavioral effect out). Ramsey sentences could also be formed about theories in other domains of science.

According to Rey, this way of introducing new terms is reminiscent of Quinian bootstrapping. As we have seen, QBS consists in introducing some new set of empty placeholder symbols together with the relations in which they stand to one another, then gradually filling up the placeholders with content. The analogy between Ramsey formalism and QBS may not be perfect, as only the former assumes that the meaning of the newly introduced theoretical terms is exhausted by their relations to other terms. Rey then suggests that even if Ramsey sentences cannot capture the format of mental representation, they can at least capture its content. Mental representation probably does not have the format of Ramsey sentences, but some Ramsey sentences might have content equivalent to conceptual representations in the mind.

Ramsey-sentences can even capture the general type of content-enrichment that arises from locking to new environmental kinds. Simply adding a rigidifying operator (Kaplan, 1978) to the logical devices utilized by Ramsey
sentences would capture the reference-fixing aspect of cognitive architecture. Kaplan's dthat operator has the following interpretation: dthat $\mathrm{F}:=$ the unique thing that is F in the actual world, but not in all possible worlds, where F is some non-rigid role concept. For example, 'watery stuff' picks out $\mathrm{H}_{2} \mathrm{O}$ in the actual world (but not in all possible worlds). 'dthat watery stuff' designates $\mathrm{H}_{2} \mathrm{O}$ - the actual filler of the watery-stuff role - in every possible world. This is also the semantic function of Fodorian conceptual primitives like WATER: according to Fodor, the concept WATER refers to the very natural kind that happens to fill the watery stuff role on Earth in our world. Therefore these dthat expressions can be equivalent in content (if not equivalent in format) to referring conceptual primitives. The question now is, may dthat expressions as reference fixers increase the expressive power of a representational system into which they are introduced? Rey notes that this is a purely logico-semantic issue which is crucial for resolving the nativism debate, still empirical evidence concerning concept acquisition appears irrelevant to it.

Rey still acknowledges that some form of QBS, or the introduction of new referring primitives, might increase the host system's expressive power. However, currently we cannot see how this could happen, because Carey's solution cannot adequately handle Fodorian problems of circularity and Goodmanian worries of indeterminacy. ${ }^{5}$ Rey's argument for this claim is based on one of Carey's key examples, the acquisition of NATURAL NUMBER. On Carey's account, the crucial step of induction in this process is when the child notices that the numerals name successive sets each of which has one more member than the previous one (Carey, 2009; Rey, 2012). This noticing, however, could not happen without tokening the concept ONE MORE THAN, aka SUCCESSOR, which is the very essence of the concept NATURAL NUMBER. Without tokening SUCCESSOR the child could generalize in a number of other ways when seeing the sets in question, never arriving at the proper meanings of the numerals. Thus Carey's bestelaborated example of Quinian bootstrapping is still subject to both Fodorian and Goodmanian worries.

Rey also supplies some examples to argue that known cases of analogical reasoning (a crucial process in QBS) are subject to the same objections. He says in his footnote 13: "Or consider the bewildering analogy often provided in introductions to General Relativity, in terms of a rubber sheet, whose shape is of course ordinarily deformed by rolling onto it a heavy metal ball. 'The curvature of fourdimensional space-time is just the same,' we're told, 'except that there's no rubber sheet, no gravity and the deformation

[^99]occurs in four dimensions'! Analogies may help in causing the manifestation of a concept, but it's hard to see how they'd be sufficient for the bringing out the possession of one." Thus it seems that nothing short of innately possessing concepts can result in their manifestation. We still cannot see a non-circular account of concept construction that can also handle the inherent indeterminacy in induction.

## Getting out of the circle

In this section I propose a slightly amended, non-circular account of bootstrapping SUCCESSOR, trying to save Carey's original account. I shall also point out some ways to handle Goodmanian worries associated with the same concept. Then I suggest one example of analogical reasoning in order to provide an existence proof that analogies play a role in concept learning, and not just triggering. Following these examples, I shall argue that some forms of procedural learning, namely the acquisition of certain skills is necessary for learning new referring concepts, and this kind of learning is essential for increasing expressive power in human minds. Finally, I briefly discuss Fodor's new nativist argument outlined in the Introduction.

## Bootstrapping SUCCESSOR

Step 1. Imagine that a child is playing with toy horses and riders; she has a bunch of both, and she is trying to mount exactly one rider on the back of each horse. Two general outcomes are possible: (i) each horse has a rider, and each rider is sitting on a horse (ii) there remain horseless riders, or riderless horses. Suppose that when (i) happens it makes the child happy giving her a feeling like "it's all nice and complete". So (i) is a Good Case. On the other hand, (ii) leaves the child with a mild frustration, as the pairing activity cannot be finished - this is a Bad Case. Thus a Good Case is one where the pairing activity can be completed, and there remain no unpaired objects in either set.

Step 2. Suppose that the child, on carefully observing a Good Case, notices two possible groupings. One is that all members in each group (e.g., horses, riders) look alike; the other is that two different objects (a horse and a rider) constitute a pair. Looking alike is judged on the basis of perceptual similarity; but the immediate question is, does the child need to be born with PAIR, in order to proceed with the second recognition? The answer is no: let us make a little detour to see how PAIR might be bootstrapped.

Here we take for granted something like Margolis and Laurence's account of primitive concept acquisition (Margolis, 1999; Laurence \& Margolis, 2002). The first step in bootstrapping PAIR is to take two generic kind-concept frames, and fill into them the two perceptual prototypes (horses and riders, in our case). I take it that such generic concept frames have two placeholders: one for perceptual information (I call this placeholder the P-slot), the other for abstract information, represented by other concepts (A-slot). For any particular concept, these slots may have a varying amount of information in them, from minimal to very rich.

Suppose that in the horses and riders case, the A-slot of both concepts initially contains only the core cognition concept OBJECT. What distinguishes the two concepts is the perceptual information in their P-slots. In order to represent a particular pair of objects, the two concepts need to be linked by means of association. In the present case this is based on the fact that the two were manipulated together (one in the child's left hand, the other in the right, etc.). To obtain the generic concept PAIR from context-dependent representations of particular object pairs, we need abstraction which in our case takes the form of feature elimination. While forming representations of different pairs of objects on different occasions, the child notices that there are varying as well as constant features in these representations. That there are two concept frames (i.e., two objects connected in experience) is the constant part; P-slot contents, and the experiences that set up the associative connections may vary. Thus PAIR will be the abstract schema TWO CONNECTED OBJECTS (i.e., two generic kind concept frames linked by association).

We can now return to SUCCESSOR. In Step 3 the child notices that the Good Cases consist of pairs only, whereas the Bad Cases contain leftover single entities for which pairing cannot be finished. Once again a move of feature elimination is needed: for being a Good Case, the quantity of pairs does not matter. Quantity of pairs may be represented by the analog magnitude system, so there is no threat of circularity: we do not need NATURAL NUMBER to account for GOOD CASE, for instance.

Step 4. Suppose that the child does further experiments with some Good Cases, trying to find out how to turn a Good Case into a bad one. She discovers that adding pairless objects of one of the two kinds will do. She then notes the minimum effort to spoil a Good Case: adding exactly one object (horse or rider, as in our example). The concept ONE is available from set-based quantification, so we have still not closed Fodor's circle. ${ }^{6}$ To summarize, the recipe for obtaining SUCCESSOR is this: take a Good Case, spoil it with the minimum effort. Disassemble the pairs; form two groups of the two kinds of objects. Now the kind group that was added an extra item a moment ago is the successor of the other (redo the pairing if you forget which one is which). At the end, MINIMALLY SPOILED GOOD CASE will serve as SUCCESSOR.

Let us now turn to Goodmanian worries. The question is, what cognitive factor is shepherding mental construction in this particular direction, given that there are so many other possible ways to assemble structures from representational primitives? Perhaps the directing forces are the innate concepts PAIR and SUCCESSOR, lurking in the background? That need not be the case. Children's and

[^100]adults' constructions do in fact proceed in many different directions, forming a lot of representational complexes. There is a lot of search going on in the vast logical space of compositional representation, but many of the constructs are soon dropped as useless - either as a result of social influence, or in an effort to understand what is going on in the environment. Some constructs, however get promoted.

For an example, imagine that Daddy and little Victor are playing a board game. At some point, little Victor wants to know who has more game pieces on the board. As a matter of fact, Daddy has six blue ones, whereas little Victor has eight red ones. Little Victor stubbornly thinks that the proper way to count is $1,2,3,4,2,2,2,2$, (repeating 2 ad infinitum). He notices that Daddy's set of pieces maps onto 2, and so do his ones, therefore he tentatively concludes that both of them have the same number of pieces. But he also has the impression that the blue and red pieces are different in number. He manages to prove this by pairing them up, producing a Bad Case. Meanwhile, Daddy is vehemently arguing that he is doing his counting in the wrong way, and offers a different system. Now unless little Victor is no smarter than the present prime minister of Hungary, he will quickly realize that something has gone wrong with his system of counting. So he switches to the one proposed by Daddy, and resolves the inconsistency. Very similar stories could be told about learning to play chess, and a number of other cases; note that this solution is quite close to Goodman's original one, namely entrenchment (Goodman \& Putnam, 1983).

## Analogical reasoning

I declare up front that at present $I$ do not have a bootstrapping story for CURVATURE OF FOURDIMENSIONAL SPACE. I can only show, using a much simpler example, that analogical reasoning may indeed be an important means of actually learning, and not just triggering, a new concept. Here is my story. The carburetor was invented by two Hungarians, Donát Bánki, and János Csonka. According to an anecdote, a key step in the process happened when Csonka was walking by a florist in a busy boulevard of Budapest. The lady was using her spray bottle to water her flowers. Csonka saw the event, and immediately realized that that was the way to manage fuel introduction into the cylinders of internal combustion engines (instead of evaporating gasoline by engine heat, as was done in some early motors). This is a fairly simple, yet powerful analogy: the recipe is, replace water by gasoline, and flowers by steel cylinders, and you pretty much got the carburetor. So we can meaningfully claim that feeding the engine with gas is like using a spray bottle - only there is no spray bottle.

This solution, however, does not generalize to the concept of four-dimensional space. An important difference between the carburetor and the 4D-space cases is that all by itself the image of a steel ball rolling on a rubber sheet may not give a physics student the crucial insight into the target concept; this image may simply be an initial intuition pump, or even
just a funny attention-grabber. I presume that a convincing account of how to learn FOUR-DIMENSIONAL SPACE will be much more complex than that of learning CARBURETOR; still we seem to have an existence proof that analogies may contribute to genuine learning.

For another example, why suppose that LIGHT BULB must be innate, given that learners can construct this concept somewhat similarly to how Edison constructed the light bulb itself? It is a matter of accumulating experience to realize that white hot metals may serve as light sources especially if they don't melt, don't burn, and don't set your house on fire. Satisfying these constraints was the result of a long cultural development, and an obvious example of skill acquisition. The crown was put on by the inventor, but he was far from being the only one to contribute. And while the relevant skills are possessed only by some expert workers, the basic principle can be understood by virtually anyone.

## Reference and ramsification

Recall Fodor's new argument for nativism, according to which (P1) compositionality cannot increase expressive power, and (P2) expressive power increase is the hallmark of concept learning, therefore (C) from the principle of compositionality no account of concept learning can be derived. What supports this argument - in particular, what supports (P1)? Fodor and Rey suggest the following analogy. Just like in formal logic, where compositionality cannot increase expressive power, in mental representation systems (MRS) combination cannot increase expressive power either. To this analogy the immediate reply is, minds are not exactly like systems of formal logic - not even on Fodor's own view. Next, however, Rey contends that no matter how the format of MRS differs from formal logic, contentwise it can be captured by logical formalism. Now the argument becomes, IF the content of MRS can be captured by logical formalism, AND compositionality in logical formalism cannot increase expressive power, THEN the kind of compositionality that obtains in MRS cannot increase expressive power either, never mind the differences in format. This is a giant leap. It needs support, which Rey provides by (i) reiterating the circularity objection and Goodmanian worries, trying to show that they undermine Carey's examples of QBS, and (ii) arguing that not even newly acquired primitive concepts can increase the expressive power of MRS. We have seen that via (i), the general argument leads us back to particular learning accounts: if learning theories can take care of circularity and Goodmanian indeterminacy objections, that will undermine Fodor's and Rey's nativist arguments. In defending (ii), Rey makes two points. First, conceptual role is essential for concept possession; a primitive concept that causally covaries with some environmental kind, but is causally or inferentially detached from other concepts does not increase the expressive power of its host MRS. Second, consider Twin-Earth thought experiments: any denizen of Earth or Twin Earth has the power to represent either XYZ or $\mathrm{H}_{2} \mathrm{O}$, even before one of these substances appears in their
environment. In this sense, there is no increase in their mind's expressive power at all.

Here is my response to this argument. Referring is an ability; it is not a logical construct. If you have the ability to refer to Fs, then some mental symbol of yours carries information about Fness. Information carried about Fness is a semantic issue, whereas the mechanism or ability that secures the locking and thereby endows the symbol with content belongs to metasemantics. As such, it is not captured by logical formalism. The distinction between skills or procedural knowledge on the one hand, and factual knowledge on the other, is a familiar idea in philosophy. Putnam (1981) argued that logical formalism can never unambiguously determine its own interpretation. However, the referential grounding of symbols in human minds reduces interpretational indeterminacy. It has also been argued that procedural knowledge does not reduce to factual knowledge captured by logical formalism or statements in natural language (Carroll, 1895; Winch, 1990; see also Lewis, 1990).

Keeping semantics and metasemantics separate is quite important. Metasemantic factors are the sources of mental (representational) content, thus candidates for a source of increasing expressive power. Semantics alone cannot account for the origin of mental content. Metasemantics does that, and a metasemantic account of how mental content arises may well involve psychological mechanisms - for example, skill leaning.

Some examples of perceptuo-motor skills that I have in mind are the following.
(1) The cultural development of artifacts. The example of light bulbs above is a case in point: the creation of new artificial kinds comes with the invention of new conceptual primitives.
(2) Deference. Learn to communicate with experts, ask for information. Is this ring made of gold or brass? Is that animal an insect or a crustacean? By gaining relevant information, you can ground new concepts.
(3) Actions. Learn or invent new types of action, and call them dances, martial arts, singing, etc.
(4) Theoretical concepts. This is admittedly the most difficult case, since on current views in philosophy of science theoretical concepts are not introduced by skill learning in the first place. Rather, they are formulated on the basis of earlier theories, and scientists' creativity. For example, scientists first inferred that electrons must exist (an exercise of already possessed inference skills), then they developed methods to detect them (detection of electrons was a newly constructed skill routinely applied later), and not the other way around. I agree that a lot more needs to be said about theoretical concepts to make a skill-based account of concept learning more plausible. One way to address this problem is to develop further the ideas of concept construction supplied above, and applying them to the acquisition of theoretical concepts. The kind of constructivism that I have proposed so far is Piagetian in
spirit, but a possible way to develop it further is to consider certain neoconstructivist approaches (e.g., Johnson, 2009).

In his paper, Rey distinguishes between functioning psychological expressive power, and semantic expressive power; he argues that learning can increase the former, but not the latter. Here is how this distinction is to be understood. Rey thinks that learning and innateness are not incompatible. Hypothesis testing and experience may select from innately specified concepts and hypotheses, similarly Chomsky's principles and parameters view of the acquisition of grammar. Experience may tell us which of the innately available concepts of ours are useful in understanding our environment at many different levels of description, ranging from social to microphysical.

I am defending a view stronger than this one: as I have argued, there exist ways in which even the semantic expressive power of human minds may increase. I agree with Carey that many of our concepts arise in ontogenesis as a result of bootstrapping from experience and a smaller set of innate concepts. Therefore they are not innate in the sense Rey thinks they are. New logical constructions out of antecedently given concepts do not increase expressive power, but new logical constructions which also contain new referring primitives inexpressible in antecedent vocabulary do so, at least according to the standards of formal logic. Moreover, as I have argued, skill learning paves the way to learning new referring primitives.

Let me add one more note on the skill-based account presented here. Fodor would say that even if you learn the skill to refer to Fs, this does not entail that you learn the concept F - you may simply trigger an innate concept by learning a skill.

What motivates this distinction between learning the relevant skill whereas only triggering the concept? In Fodor's view, what keeps this distinction compelling is that no particular skill is necessary for possessing the concept F. The criterion of concept possession is locking a mental symbol to its extension, and locking is multiply realizable (Fodor, 1998). Here is my reply. The idea that referring is an ability can easily accommodate multiple realizability: there are many different skills that can ground a given concept. My concept ELM is a deferential one at present, but I could become an expert at recognizing trees in the future. One needs to learn some of a range of relevant skills in order to come to possess a particular referring concept.

Finally, note that Fodor's new, expressive-power-based argument for nativism is not nearly an a priori one. As we have seen, this argument takes us straight back to particular theories of concept learning when we check the support for its premises. I think the really serious problems of concept learning remain Fodor's circularity objection, Goodmanian indeterminacy, and Fodor's other question of how concept learning can be anything other than hypothesis formation. Solving these problems takes a lot of work: we need to devise detailed accounts of the acquisition of particular concepts, or types of concepts. But that is just what Susan Carey started doing in her book.

## Acknowledgments

An earlier version of this paper was presented at the Budapest Workshop in Philosophy of Mind and Cognitive Science at Central European University, May 14, 2012, as a commentary on Georges Rey's talk. I wish to thank the organizers Katalin Balog and Barry Loewer for the invitation, and Georges Rey for critical discussion. I am also grateful for the valuable commentaries by the four anonymous reviewers of this submission.

## References

Carey, S. (2009). The origin of concepts. New York NY: Oxford University Press.
Carroll, L. (1895). What the Tortoise Said to Achilles. Mind, 4, 278-80.
Fodor, J. A. (1975). The Language of Thought. Cambridge MA: Harvard University Press.
Fodor, J. A. (1981). The present status of the innateness controversy. In Representations: philosophical essays on the foundations of cognitive science. Cambridge MA: MIT Press.
Fodor, J. A. (1990). A theory of content and other essays. Cambridge MA: MIT Press.
Fodor, J. A. (1998). Concepts: Where cognitive science went wrong. New York NY: Oxford University Press.
Fodor, J. A. (2008). LOT 2: The Language of Thought Revisited. New York NY: Oxford University Press.
Goodman, N., \& Putnam, H. (1983). Fact, fiction, and forecast. Cambridge MA: Harvard University Press.
Johnson, S. P. (2009). Neoconstructivism. New York, NY: Oxford University Press
Kaplan. D. (1989). Demonstratives. In J. Almog, J. Perry, \& H. Wettstein (Eds.), Themes from Kaplan. New York NY: Oxford University Press.
Laurence, S., \& Margolis, E. (2002). Radical concept nativism. Cognition, 86, 25-55.
Lewis, D. (1990). What experience teaches. In G. W. Lycan (Ed.), Mind and Cognition. Hoboken, NJ: WileyBlackwell
Margolis, E. (1999). How to acquire a concept. In E. Margolis \& S. Laurence (Eds.), Concepts: Core readings. Cambridge MA: MIT Press.
Putnam, H. (1981). Reason, truth, and history. Cambridge UK: Cambridge University Press.
Rey, G. (2012). Learning, expressive power and mad dog nativism: The poverty of stimuli (and analogies) yet again. Unpublished manuscript.
Winch, P. (1990). The idea of a social science and its relation to philosophy. London, UK: Routledge

# Gestures for Thinking 

Azadeh Jamalian (aj2334@columbia.edu)<br>Teachers College, Columbia University, 525 W. $120^{\text {th }}$ Street

New York, NY 10027 USA

Valeria Giardino (Valeria.Giardino@ens.fr)<br>Institut Jean Nicod, Ecole Normale Supérieure, 29, rue d'Ulm, F-75005 Paris

Barbara Tversky (btversky@stanford.edu)<br>Teachers College, Columbia University, $525 \mathrm{~W} .120^{\text {th }}$ Street<br>New York, NY 10027 USA


#### Abstract

Can our gestures help us think, and, if so, how? Previous work suggests that they can. Here, students, alone in a room, studied descriptions of environments for later tests of knowledge. The majority of participants spontaneously gestured while reading the descriptions, and most of those also gestured while answering true-false questions. They did not gesture proportionately more time for environments with many landmarks than for environments with few. Their gestures laid out the environments, primarily using points to places and lines for paths. Descriptions and questions accompanied by gestures were remembered more accurately. Participants rarely looked at their hands. Gestures seem to promote learning by establishing embodied representations of the environments.


Keywords: Gesture; embodiment; spatial representation; spatial memory; route/survey perspectives; navigation.

## Introduction

Gestures serve many ends and have many forms. People gesture in communications to others, but also for themselves, that is, they gesture to think (Goldin-Meadow, 2003; McNeill, 1992). Gestures for thinking help thinking in different ways. They help people find words (Krauss \& Hadar, 2001). They offload memory (Cook, Yip, \& GoldinMeadow, 2012; Goldin-Meadow, Nusbaum, Kelly, \& Wagner, 2001). They help people perform mental rotation (Chu \& Kita, 2008; Wexler, Kosslyn, \& Berthoz, 1998; Wohlschlager \& Wohlschlager 1998). They help people count (Carlson, Avraamides, Cary, \& Strasberg, 2007).

Gestures are actions in space, and as such, can readily represent spatial structures and spatial actions. In fact, gestures help people solve spatial problems (Kessell \& Tversky, 2006; Schwartz \& Black, 1996). Interestingly, in solving spatial problems, gestures can serve much like diagrams. When given paper and pencil during problem solving, one group diagrammed the same spatial problems that another group gestured to solve (Kessell \& Tversky, 2006). Diagrams also offload memory, but they serve cognition in many other ways. Creating a good diagram entails extracting the crucial information and structuring it to represent a problem to be solved or information to be
comprehended and learned felicitously, yielding an integrated external model of the information that can be inspected and mentally manipulated (e. g., Tversky, 2011). Gestures are crude, and as such almost necessarily abstract. They can also create integrated external models. In explaining complex environments or scientific systems, people produced a coordinated and integrated series of gestures that modeled the spaces of environment (Emmorey, Tversky, and Taylor, 2000), family trees (Enfield, 2003), and scientific processes (Kang, Tversky, and Black, 2013) to be learned.

People gesture to explain spatial environments to others, creating external models with their hands. Will they do so for themselves, as aids to comprehension and memory? Here, we investigate whether people, alone in a room studying descriptions of complex environments will gesture for themselves. If so, what is the nature of their gestures? And does gesturing help them learn and remember the environments?

Gesturing could help learning and memory indirectly by off-loading memory to another modality. Gestures have been shown to be effective in off-loading memory during explanations (Goldin-Meadow, et al, 2001). But gestures could also help learning and memory in direct ways, by constructing an external model of the environment to be learned. Half the environments participants studied had 4 landmarks and half had 8 landmarks; the latter should put greater stress on working memory (e. g., Jonides, Lewis, Nee, Lustig, Berman, and Moore, 2008). If the primary role of gestures is to offload working memory, participants should gesture more when studying descriptions with more landmarks. If the primary role of gestures is to construct a model of the environment, much like a diagram, then there is little reason to expect more gesturing for the environments with more landmarks. Gestures can reflect mental representations (e. g., Alibali, Bassok, Olseth, Syc, and Goldin-Meadow, 1999). Description perspective was manipulated because route and survey descriptions yield different mental representations early (but not late) in learning (Lee and Tversky, 2005).

## Method

Participants. 48 ( 28 female, 20 male), primarily graduate students from Columbia University, were paid to participate in the study. Participants were native English speakers or have graduated from an English speaking high school.
Descriptions. The environments had 4 or 8 landmarks. There were three outdoor environments, Etna City, Chinatown, and the Financial district, and three indoor environments, a spa, an electronics show, and a grocery store. There were 8 landmarks and 4 landmarks versions of each of these.

There were also versions of each environment from route $(\mathrm{R})$ or survey (S) perspectives. A route perspective takes an imaginary traveler, you, through an environment describing the turns and landmarks with respect to "you" in terms of your left, right, front, and back. A survey perspective takes an overview of an environment and describes landmarks with respect to each other in terms of north-south-east-west. The route descriptions always began with cardinal directions so that participants could answer questions from a survey perspective. The descriptions and the environments were based on earlier work (Taylor \& Tversky, 1992).

The average length of the 8R descriptions was 141 words, of the 8 S descriptions, 127 words, of the 4 R descriptions, 69 words, and of the 4 S descriptions, 72 words. Table 1 shows an example of a description of an outdoor environment with 4 landmarks from a survey perspective, and of an indoor environment with 8 landmarks from a route perspective.

## Table 1: Examples of descriptions

## Example 1: 4S outdoor environment

Etna is a charming town nestled in an attractive valley, entered on River Highway. River Highway runs eastwest at the southern edge of the town of Etna. Toward the eastern border, River Highway intersects with Mountain Rd, which runs north of it. At the northwest corner of the intersection is a gas station. North of the gas station, Mountain Road will intersect with Maple Ave, which runs west.

## Example 2: 8R indoor environment

Rock Creek Center is a showcase for new electronic devices. Enter Rock Creek Center from the east side of the building near the southeast corner. As you enter, you see, on the left wall, a Bulletin Board. Past the Bulletin Board, on your right is the Video Camera room and on your left is the Office stretching to the corner of the building. Past the office you are forced to turn right and you will find the Cafeteria on your left stretching to the corner of the building. After the Cafeteria, you are forced to turn right and you will find a large room with Mobile Phones on your left. On your right you will see the Televisions room. At the end of the hallway, turn right and you will find the Laptop Center on your left. Past the Laptop Center, you will return to the entrance on your left.

Design. Each participant read four descriptions, one with 4 landmarks and one with 8 landmarks from each perspective. The specific environment for each condition was chosen from the set of three outdoor environments and three indoor environments. All variables, size, perspective, environment, order were counter-balanced and appeared equally often across participants.
True-false Questions. Verbatim and inference statements were designed for each description, 10 for the 8 landmark environments and 6 for the 4 landmark environments. For the 8 landmark environments, there were 2 statements taken verbatim from the text with the same perspective, 2 statements taken verbatim from the text with the other perspective, and 6 inference statements, 3 route, and 3 survey. For the 4 landmark environments, there were a total of 6 statements: 1 verbatim from the route perspective, 1 verbatim from the survey perspective, 2 inference from a route perspective, and 2 inference from a survey perspective. Inference statements could be verified from information provided in the descriptions. Half of the statements were true and the other half was false. The statements were presented in a random order for each participant. Table 2 shows examples of true/false statements for Etna.

Table 2: Examples of true/false statements

|  | Verbatim | Inference |
| :---: | :--- | :--- |
| Route | Going east on River <br> Highway, at the <br> intersection with <br> Mountain Rd, you will <br> find a gas station on <br> your left. | From Mountain Rd, <br> turn right on River <br> Highway and you <br> will have the Gas <br> Station on your <br> right. |
| Survey | North of the gas <br> station, Mountain <br> Road will intersect <br> with Maple Ave, <br> which runs east. | South of Maple Ave <br> to the west of <br> Mountain Rd is the <br> Gas Station. |

Procedure. Participants first signed a consent form, assenting to participating in the experiment and to being videotaped. They were additionally asked for permission to show their videos in presentations of the research. They then completed a paper version of the Mental Rotation Task (Vandenberg \& Kuse, 1978), a common test of spatial ability.

Participants were seated in front of a Mac OS X 10.7, as shown in Figure 1. Video records of the computer screen and front views of participants were captured with Silverback© software, and participants' side views with a videocam. The experimenter explained the procedure to each participant: "In this study you will be asked to read 4 text descriptions of environments. After reading each description, your memory for the information in the text will be tested. You will start with a practice text description. Throughout the study, you will not have access to a
keyboard and will send commands to the computer with your voice." The participants responded verbally, saying "next", "yes", or "no" when appropriate, to advance from screen to screen. Their responses were analyzed by the Mac speech recognition program and used to advance screens and record responses. This left participants' hands free to gesture, on or off the table.

Participants first had a practice trial. The first screen explained the task: "You will be asked to read the description of an environment as practice. Once you are done reading the description say aloud "Next". After the description you will be asked to judge the truth of some statements about the environment. You may take as much as time you need." Then participants read a description of an amusement park. The complete description was on the screen. Participants were free to read the practice and experimental descriptions as long as they liked. Immediately after reading the description, participants were presented with 4 true/false questions, one on each screen. They said "yes" for true and "no" for false. After the practice trial, the experimenter answered any questions the participant had, and then left the room.

Participants then proceeded through the experiment, reading each of the four descriptions and answering the corresponding true/false questions after each.


Figure 1: Experimental Setup. Participant gesturing while studying description.

## Results

Coding. Two trained coders, coded 10 of 48 videos for gesturing while studying, Kappa $=0.76(\mathrm{p}<0.001)$, for length of time spent gesturing while studying, $\mathrm{t}(39)=0.244$, $\mathrm{p}=0.809$, for looking at their hands while gesturing while studying, Kappa $=0.56(p<0.001)$, for studying time, $\mathrm{t}(39)=1.402, \mathrm{p}=0.169$, for gesturing while verifying statements, Kappa $=0.90(p<0.001)$, for looking at their hands while gesturing in verifying statements, Kappa $=0.44$ ( $\mathrm{p}<0.001$ ), and for length of time to verify statements, $t(359)=0.120, p=0.90$. Any movement of hands or fingers, excluding beat gestures, was coded as gesturing. Any glance at hands while gesturing was coded as looking. The coded duration of the gesture included active movements and periods when individuals left their hands still on the table or in mid-air in a certain position and form. Times were coded from the Silverback© videos of the screen and by using ELAN software. In cases of disagreement coders consulted
a third coder. One coder coded the remaining videos, discussing uncertain cases with the second coder. Qualitative coding of the gestures is ongoing, but it is clear that gestures indicating places, primarily points, and indicating connections between places, drawing lines or placing the edge of a hand, predominate. Most gestures were performed on the table, but some were in the air (see Figures 1 and 2).
Gesture at study. Seventy-three percent of participants (35 out of 48) gestured at least once for at least one description during study. Twelve participants ( $25 \%$ ) gestured for all four descriptions, 7 gestured for three, 10 gestured for two, and 6 for only one. Notably, number of landmarks in the environments ( 4 vs. 8) did not influence whether participants gestured at study, $\chi^{2}(1, \mathrm{~N}=48)=1.132, \mathrm{p}=$ 0.289 . Similarly, neither perspective (route vs. survey), $\chi^{2}(1, \mathrm{~N}=48)=.023, \mathrm{p}=0.879$, order $\left(1^{\text {st }}, 2^{\text {nd }}, 3^{\text {rd }}\right.$, or $\left.4^{\text {th }}\right)$, $\chi^{2}(3, \mathrm{~N}=48)=1.687, \mathrm{p}=0.171$, or MRT score, $\mathrm{F}(1,185)=$ $0.089, \mathrm{p}=0.765$, influenced gesturing at study.

For each participant, the percentage of time gesturing while studying was computed. Neither spatial ability F(1, $45)=0.357, \mathrm{p}=0.553$ ) nor gender $(\mathrm{F}(1,45)=0.505, \mathrm{p}=$ $0.481)$ affected the percent of time gesturing
Gesture at test. Sixty-five percent of participants (31 out of 48) gestured at least once when verifying the true/false statements. Table 3 shows number of statements for which participants gestured both when studying and answering, only when studying, only when answering, or not at all, out of the total of 1526 statements (excluding 10 cases in which participants' answers were missing).


Figure 2. Participant gesturing while answering question.
Table 3: Number of questions and gesture behavior

| Gesturing | Frequency | Percentage |
| :--- | :---: | :---: |
| Both at study and when <br> verifying | 547 | $35.8 \%$ |
| Only at study | 220 | $14.4 \%$ |
| Only when verifying | 21 | $1.4 \%$ |
| None | 738 | $48.4 \%$ |

As evident from Table 3, participants were far more likely to gesture to verify statements for the descriptions they gestured at study. Only $1.4 \%$ of the questions received
gestures at verification that had not received gestures at study.

Moreover, for $85 \%$ of the descriptions accompanied by gesture, at least one question was also accompanied by gesture. Participants, who did not gesture at all while studying the descriptions did not gesture when answering questions. Specifically, $27 \%$ of participants ( 13 out of 48) did not gesture either at study or at verification.

Overall, neither the environment's perspective (survey vs. route), $\chi^{2}(1, \mathrm{~N}=48)=.743, \mathrm{p}=0.389$, nor question perspective, $\chi^{2}(1, \mathrm{~N}=48)=.264, \mathrm{p}=0.608$, nor number of landmarks ( 4 vs. 8 ), $\chi^{2}(1, \mathrm{~N}=48)=.028, \mathrm{p}=0.868$, nor type of statement (verbatim vs. inference), $\chi^{2}(1, \mathrm{~N}=48)=.439$, $\mathrm{p}=0.508$, nor MRT scores, $\mathrm{F}(1,1520)=0.899, \mathrm{p}=0.343$ influenced whether participants gestured at verification.

In short, most participants gestured while studying and verifying and most who gestured at verification had also gestured at study. Neither spatial ability nor length nor perspective of the descriptions or questions affected whether participants gestured.
Accuracy. As evident in Figure 3, when participants had gestured at study, they were more likely to be accurate at testing ( $\mathrm{M}=0.821, \mathrm{SD}=0.29$ ) than when they had not gestured at study $(\mathrm{M}=0.743, \mathrm{SD}=0.30), \mathrm{F}(1,1517)=$ $8.249, \mathrm{p}=0.004<0.01$. Not surprisingly, accuracy was higher for the 4 landmark environments ( $M=0.810$, $\mathrm{SD}=$ 0.24 ) than for the 8 landmark environments ( $\mathrm{M}=0.760, \mathrm{SD}=$ $0.28), \mathrm{F}(1,1517)=6.561, \mathrm{p}=0.011<0.05$. Accuracy improved with spatial ability, $F(1,1517)=10.210, p=0.001$ $<0.01$ but the correlation between accuracy and spatial ability was low and not significant. Accuracy varied with kind of statement, $F(1,1517)=7.182, \mathrm{p}<0.001$. Replicating Taylor and Tversky (1992), post-hoc analyses showed that verbatim statements ( $\mathrm{M}=0.838, \mathrm{SD}=0.21$ ) were more accurate than inference statements ( $M=0.720, \mathrm{SD}=0.31$ ), $\mathrm{t}(1513)=3.809, \mathrm{p}<0.01$, and that for inference statements, there was no advantage for statements in the perspective of reading (same perspective ( $\mathrm{M}=0.727, \mathrm{SD}=0.30$ ); other perspective $(M=0.718, S D=0.31), t(1513)=0.311, p=$ 0.756 ), indicating that memory representations were perspective-free.


Figure 3. Accuracy by gesturing at study. Error bars represent standard error

The effects of gesturing at verification were analyzed separately. Participants were more likely to be accurate verifying statements when they gestured ( $M=0.814, S D=$ 0.23 ), than when they $\operatorname{did}$ not ( $\mathrm{M}=0.757, \mathrm{SD}=0.29$ ), $\mathrm{F}(1$, $1515)=5.325, \mathrm{p}=0.038<0.05$. As before, accuracy increased with spatial ability, $F(1,1515)=10.191, p=0.001$ $<0.01$, and was affected by statement category in the same ways as the previous analysis, $\mathrm{F}(1,1515)=17.084, \mathrm{p}<$ 0.001 .


Figure 4: Accuracy by gesturing at verification.
To examine the effects of gesture at study and gesture at response, participants were divided into 4 groups: gesture at both, gesture only at study, gesture only at response, no gesture. Gesture behavior had an effect on accuracy, $\mathrm{F}(3$, 1494) $=3.593, \mathrm{p}=0.013<0.05$. Post-hoc analyses showed that participants were more accurate at testing when they had gestured both at study and verification ( $\mathrm{M}=0.780, \mathrm{SD}=$ 0.27 ), than when they did not gesture at all $(\mathrm{M}=0.705, \mathrm{SD}=$ 0.32 ), $\mathrm{t}(1494)=2.491, \mathrm{p}=0.013<0.05$. Similarly, they were more accurate when they only gestured at study ( $\mathrm{M}=0.816$, $\mathrm{SD}=0.23$ ), than when they did not gesture at all, $\mathrm{t}(1494)=$ 2.655, $p=0.008<0.01$. However, there was not a significant improvement for gesture only at response ( $M=0.811$, $\mathrm{SD}=0.25$ ) than for no gesture, $\mathrm{t}(1494)=0.333, \mathrm{p}=0.739$; this could be due to the severely limited number of cases in which participants only gestured at response (See Table 3).
To make sure that the advantage of gesturing was not because the better learners gestured, comparisons were done within participants who gestured when studying two or three descriptions, but not all descriptions. For those who gestured sometimes, accuracy was higher when they gestured at study ( $\mathrm{M}=0.762, \mathrm{SD}=0.29$ ) than when they did not $(\mathrm{M}=0.677, \mathrm{SD}=0.35), \mathrm{F}(1,513)=3.938, \mathrm{p}=0.048<$ 0.05 . Similarly, they were more accurate verifying statements when they gestured ( $\mathrm{M}=.764, \mathrm{SD}=0.29$ ) than when they did not gesture ( $\mathrm{M}=0.628, \mathrm{SD}=0.35$ ), $\mathrm{F}(1$, $513)=3.910, p=0.049<0.05$. So, gesturing itself helps - it is not just that those who tend to gesture also remember better.
Studying Times. As expected, participants took longer to study the longer descriptions with 8 landmarks ( $\mathrm{M}=$ $112.14 \mathrm{sec}, \mathrm{SD}=28.43$ ) than the shorter ones with 4 landmarks ( $\mathrm{M}=56.57 \mathrm{sec}, \mathrm{SD}=28.43$ ), $\mathrm{F}(1,187)=94.104$, p
$<0.001$. Gesturing did not influence study time, $\mathrm{F}(1,187)=$ $1.212, p=0.272$. Similarly, neither spatial ability, $F(1,187)=$ $2.198, p=0.140$, nor text perspective, $F(1,187)=0.101, p=$ 0.752 , affected study times.

Verification Times. Figure 5 shows that gesture behavior influenced verification time, $\mathrm{F}(3,1441)=3.431, \mathrm{p}=0.016<$ 0.05 . Post-hoc Bonferroni comparisons showed that participants were faster to verify statements when they had only gestured at study $(M=8.95 \mathrm{sec}, \mathrm{SD}=2.61)$ than when they had not gestured at all ( $\mathrm{M}=10.35 \mathrm{sec}, \mathrm{SD}=4.16$ ), $\mathrm{p}<$ 0.001 . By contrast, answering took longer when participants only gestured at verification $(M=15.65 \mathrm{sec}, \mathrm{SD}=6.19)$ than when they only gestured at study, $\mathrm{p}=0.004<0.01$. There was no difference on verification time when participants gestured both at study and at verification ( $\mathrm{M}=11.47 \mathrm{sec}$, $\mathrm{SD}=3.88$ ), compared to when they did not gesture at all. Spatial ability, perspective, and size of environment did not effect verification times. Thus gesture at study decreased verification time while gesture at responding increased verification time, and in cases when they gestured both at study and at verification, the two effects cancelled each other.


Figure 5: Verification time by gesture behavior
Did participants look at their hands while gesturing? For the most part, participants did not look at their hands while gesturing; they looked at their hands for $35.8 \%$ of the texts during reading but they were typically brief glances. Out of the 35 participants who gestured at least once when reading texts, 15 never looked at their hands. At verification, participants looked at their hands for less than $10 \%$ of the statements they gestured while verifying. Out of the 31 participants who gestured for at least one of the statements, 16 never looked at their hands.

## Discussion

Participants, alone in a room, read descriptions of a variety of complex environments that they were to learn for later questions. While they were studying, most of them gestured at least once, and the majority gestured for most of the descriptions, in the absence of any communication. The
descriptions accompanied by gestures were remembered better than those that were not, and the questions that were accompanied by gestures were answered more accurately than those that were not. The advantage of gesturing on memory cannot be explained as the better participants both gestured and remembered better. Even within those participants who frequently gestured, gesturing at study and at responding improved memory. Gestures modeled the structures of the environments, pointing to places and outlining paths between places. Except on rare occasions, participants did not look at their hands as they gestured, suggesting that it is the actions per se that serve comprehension and learning, rather than the visual accompaniments. Overall, spontaneous gesturing at learning and spontaneous gesturing at memory retrieval promoted learning. Gestures appeared to improve learning by establishing embodied representations of the structures of the environments and appear to improve memory by redintegrating the queried parts of the environments.

In addition to providing embodied representations of the environments, gestures might also have served to offload memory, as in previous research (e. g., Cook, et al., 2012; Goldin-Meadow, et al., 2001), just as diagrams offload memory. However, the proportion of study time gesturing did not increase as memory load increased from light to heavy. Thus, the role of gesture in lightening memory load appears to be less important for comprehending and learning complex environments than other features of gestures, notably, creating embodied representations.

Gestures are actions, and thereby provide an additional code beyond the verbal code participants read. Multiple codes in multiple modalities are known to promote memory (e. g., Paivio, 1986). Motor codes in particular augment memory (e. g. Engelkamp \& Zimmer, 1994; Hommel, Musseler, Aschersleben, \& Prinz, 2001) but the cases that have been studied have primarily been cases where the memory was for the action per se. In the present case, the actions served memory not for the actions but rather for what the actions represented.

Actions, like diagrams and words, can represent, that is, they can stand for something other than themselves. Certainly for the case of words but also for the case of diagrams, representation seems to be their primary function. Not so for actions. Actions can represent, but they are primarily used for the ordinary (and extraordinary) tasks of life, manipulating objects and navigating environments. Gestures are a special class of actions that serve to represent rather than to act on or in the world. Similar to diagrams, gestures can represent more directly than purely symbolic words; they bear some resemblance to what they represent (e. g., Tversky, 2011).

Like diagrams, gestures can use space to represent ideas that are spatial or metaphorically spatial (e. g., Enfield, 2003; Emmorey, et al., 2000; Tversky, 2011; Tversky, Heiser, Lee, \& Daniel, 2009). Like diagrams, gestures are spatial and visual. However, it seems that the spatial and action components of representational gestures serve
comprehension and memory rather than the visual. Participants rarely looked at their hands. Researchers in art, sketching, and design refer to drawing as gesture. Blindfolded architects gesture copiously as they design, and they cannot see either their gestures or their designs. Nevertheless, their designs equal those they create without blindfolds (Bilda and Gero, 2006). Together, these findings suggest that some of the benefits of gesturing to those who gesture may be the embodiment of thought into action.

## Acknowledgments

The authors are indebted to NSF IIS-0725223, NSF IIS0855995 and NSF IIS-0905417 for supporting parts of the research and manuscript preparation. Valeria Giardino is grateful to The Italian Academy for Advanced Studies in America at Columbia University for supporting her visit.

## References

Alibali, M. W., Bassok, M., Olseth, K. L., Syc, S. E., \& Goldin-Meadow, S. (1999). Illuminating mental representations through speech and gesture. Psychological Science, 10, 327-333.
Bilda, Z and Gero, J.S. (2006) Reasoning with internal and external representations: A case study with expert architects. In R Sun (ed), CogSci/ICCS 2006, Lawrence Erlbaum, pp. 1020-1026.
Carlson, R. A., Avraamides, M. N., Cary, M., \& Strasberg, S. (2007). What do the hands externalize in simple arithmetic? Journal of Experimental Psychology: Learning, Memory, and Cognition, 33(4), 747-756.
Chu, M., \& Kita, S. (2008). Spontaneous gestures during mental rotation tasks: Insights into the microdevelopment of the motor strategy. Journal of Experimental Psychology: General, 137, 706-723.
Cook, S. W., Yip, T. \& Goldin-Meadow, S. (2012). Gestures, but not meaningless movements, lighten working memory load when explaining math. Language and Cognitive Processes, 27, 594-610.
Emmorey, K., Tversky, B., \& Taylor, H. (2000). Using space to describe space: Perspective in speech, sign, and gesture. Spatial Cognition and Computation, 2, 157-180.
Enfield, N. (2003). Producing and editing diagrams using co-speech gesture: Spatializing non-spatial relations in explanations of kinship in Laos. Journal of Linguistic Anthropology, 13, 17-50.
Engelkamp, J. and Zimmer, H. D. (1994). Human memory: A multimodal approach. Seattle: Hogref and Huber.
Goldin-Meadow, S. (2003). Hearing gesture: How our hands help us think, Cambridge, Mass.: The Belknap Press.
Goldin-Meadow, S., Nusbaum, H., Kelly, S. D., \& Wagner, S. (2001). Explaining math: Gesturing lightens the load. Psychological Science, 12, 516-522.
Hommel, B., Musseler, J., Aschersleben, G., \& Prinz, W. (2001). The theory of event coding (TEC): A framework for perception and action planning. Behavioral \& Brain Sciences, 24, 849-937.

Jonides, J., Lewis, R. L., Nee, D. E., Lustig, C. A., Berman, M. G., and Moore, K. S. (2008). The mind and brain of short-term memory. Annual Review of Psychology, 19, 193-224.
Kang, S., Tversky, B. \& Black, J. B. (2013) Gesture and speech in explanations to experts and novices. Submitted.
Kessell, A. M. \& Tversky, B. (2006). Using gestures and diagrams to think and talk about insight problems. In R. Sun and N. Miyake (Eds) Proceedings of the 28th Annual Conference of the Cognitive Science Society. P. 2528.
Krauss R. M. \& Hadar U. (2001). The role of speech-related arm/hand gestures in word retrieval. In R. Campbell \& L. Messing, (Eds.) Gesture, speech, and sign. Pp. 93-116. Oxford: Oxford University Press.
Lee, P. U. and Tversky, B. (2005). Interplay between visual and spatial: the effect of landmark descriptions on comprehension of route/survey spatial descriptions. Spatial cognition and computation, 5, 163-185.
McNeill, D. (1992). Hand and mind: What gestures reveal about thought. Chicago: University of Chicago Press.
Paivio, A (1986). Mental representations: a dual coding approach. Oxford: Oxford University Press.
Schwartz, D. L., \& Black, J. B. (1996). Shuttling between depictive models and abstract rules: Induction and fallback. Cognitive Science, 20, 457-497.
Taylor, H. A., \& Tversky, B. (1992). Spatial mental models derived from survey and route descriptions. Journal of Memory and Language, 31, 261-282.
Tversky, B. (2011). Visualizing thought. Topics in Cognitive Science. 3, 499-535.
Tversky, B., Heiser, J., Lee, P. and Daniel, M.P. (2009). Explanations in gesture, diagram, and word. In K. R. Coventry, T. Tenbrink, \& J. A. Bateman (Editors), Spatial language and dialogue. Pp. 119-131. Oxford: Oxford University Press.
Vandenberg, S. G. and Kuse, A. R. (1978). Mental rotations: A group test of three-dimensional spatial visualization. Perceptual Motor Skills, 47, 599-604.
Wexler, M., Kosslyn, S. M., \& Berthoz, A. (1998). Motor processes in mental rotation. Cognition, 68, 77-94.
Wohlschläger, A., \& Wohlschläger, A. (1998). Mental and manual rotatation. Journal of Experimental Psychology: Human Perception and Performance, 24, 397-412.

# Relational labels can improve relational retrieval 

Anja Jamrozik (a.jamrozik@u.northwestern.edu)<br>Dedre Gentner (gentner@northwestern.edu)<br>Department of Psychology, Northwestern University<br>2029 Sheridan Road, Evanston, IL 60208 USA


#### Abstract

Retrieval that is based on common relational structure, such as an underlying principle or pattern, is useful but typically rare. Based on evidence that comparison-derived schema abstraction can improve relational retrieval, we asked whether the use of relational labels can also promote abstraction and improve relational retrieval. Using a cued-recall paradigm, we varied the presence of relational labels at encoding and test. As compared to a no-label baseline condition, relational retrieval improved when relational labels were given at encoding and at test and also when relational labels were given only at encoding. The findings demonstrate that one way to improve relational retrieval is through the use of labels that name relational structure.


Keywords: relational retrieval; relational language; inert knowledge

## Introduction

When encountering a new example or problem, we sometimes retrieve examples from memory that share relational structure with the current example. This can be very useful as it allows us to transfer existing knowledge to the new example. For instance, if a social psychology student learns about the classic findings that a person's attitude can become resistant against very persuasive arguments after the person has argued against weak versions of such arguments (e.g., McGuire, 1961; McGuire \& Papageorgis, 1961), then this might remind the student of how someone can become immune to a disease after being exposed to a weakened form of that disease. Based on this connection, they may be able to draw some conclusions about the new situation, such as why the attitude becomes resistant to change, or why the initial arguments against the attitude have to be weak.

As useful as it can be, relational retrieval-retrieval based only on common relational structure-is typically rare (e.g., Gentner, Rattermann, \& Forbus, 1993; Gick \& Holyoak, 1980, 1983; Holyoak \& Koh, 1987; Ross, 1987, 1989). Instead, memory retrieval is likely to be based either on overall similarity or on surface commonalities, such as matching entities (e.g., Brooks, Norman, \& Allen, 1991; Gentner et al., 1993; Holyoak \& Koh, 1987; Ross, 1987, 1989). This is an instance of the inert knowledge problem (Whitehead, 1929) - that people are often unable to retrieve knowledge and apply it to new situations even when that information has been stored in memory (e.g., Barnett \& Ceci, 2002; Bransford, 1979).

Even though relational retrieval is typically rare, it is more likely for experts in a domain. For example, when solving challenging science problems, experts often retrieve problems that share common relational structure (e.g., Clement, 1988). Likewise, the likelihood of relational retrieval is better for students with greater mathematical expertise than for novices (Novick, 1988), and mathematical expertise can predict the likelihood of transferring a solution strategy to analogous math problems (Novick \& Holyoak, 1991).

What contributes to experts' improvement in relational retrieval? Two factors that might be involved are having experienced many opportunities to compare examples and acquiring a technical vocabulary. There is abundant evidence that comparison of examples can improve the likelihood of relational retrieval. When learners compare two instances of the same relational structure, the process of alignment renders their common structure more salient. This process of schema abstraction increases the likelihood of retaining this common relational structure and transferring it to other instances (e.g., Gentner Loewenstein, Thompson, \& Forbus, 2009; Gentner \& Markman, 1997; Gick \& Holyoak, 1983; Markman \& Gentner, 1993; Reeves \& Weisberg, 1994; Ross \& Kennedy, 1990).

In the current research, we focus on the second factor and ask how acquiring a vocabulary may impact relational retrieval ${ }^{1}$. As someone develops expertise in a domain, they may acquire terms that name common relations or relational patterns in that domain. We ask whether the use of such relational language can improve relational retrieval.

## Relational Language

In the current research, we ask whether an abstraction process like the one that operates during comparison also applies when relational language is used. Specifically, we ask whether using known relational terms like reciprocity or inoculation to label examples promotes the abstraction of their relational structure and leads to improved relational retrieval. The idea is that using a relational term to label a situation can promote the abstraction of relational structure

[^101]and change the construal of the situation by shifting the focus to its relational structure (e.g., Gentner, 2003, 2010; Genter, Angorro, \& Klibanoff, 2011; Gentner \& Loewenstein, 2002). This may result in improved relational retrieval in response to another example that shares the same relational structure.

Applying relational labels shares some similarities with comparison of examples, in that both involve the extraction and retention of relational structure. The difference is that in the case of relational labels, the alignment is between a labeled example and the relational structure conveyed by the label (which has already been abstracted). For instance, if the label inoculation is applied to a situation in which someone becomes immune to the stronger form of smallpox after being exposed to a milder form of smallpox, this invites a construal of the situation with a heightened focus on the common relational structure, such as 'exposure to a weakened form of something protects against the stronger form'. If so, then this might increase transfer to a further, more distant example of the relational structure. For instance, if someone encodes the example from medicine labeled inoculation and later receives an example about attitude change labeled inoculation, then both of the examples should be construed with a focus on the relational structure that each shares with inoculation. These similar relational construals should then improve the likelihood of relational retrieval of one example given the other.

The logic of the current experiment is to vary the presence of relational labels and to test the likelihood of subsequent relational retrieval. In the experiment, participants read a series of stories at encoding either with or without relational labels. Later, they were given test stories that shared relational structure with the encoding stories and asked to write out of the encoding stories they were reminded of. In one condition, the same relational labels were used at encoding and at test. The prediction is that this will lead to heightened relational retrieval, because the labels will invite the same construal across examples. Of course, if the same label is used at encoding and at test, then this leaves open the possibility that the label may also be acting as a common surface feature between examples that share relational structure.

We will also test a more interesting possibility: that labels may improve relational retrieval even if they are only present at encoding. If a relational label is applied to the initial example, then this can result in a construal in which there is a greater focus on the invited relational structure. Later, if the relational structure of the test example is apparent (or even partially apparent), then it should be more likely to match the initial (stored) example than it would have been had the initial example not been labeled. For instance, if the initial medical example is labeled inoculation, then this may increase its likelihood of being retrieved given a new passage involving the analogous phenomenon in attitude change.

Finally, it is possible that relational labels present at test only may likewise improve relational retrieval. There is some evidence that deriving a relational abstraction from comparison can improve relational retrieval of past examples that share the same relational structure (Gentner et al., 2009). If, as we have suggested, labels also promote abstraction of relational structure in much the same way as comparison, then they may also improve relational retrieval when they are given at test.

In the current experiment, we tested whether relational labels can improve relational retrieval in one of the ways described above. We varied the presence of labels at encoding and at test. As outlined above, the pattern of results can inform us about the underlying processes through which relational labels might be having their effect.

## Logic of the Experiment

We used a cued recall paradigm similar to that used in previous studies of relational retrieval (e.g., Gentner et al, 1993). Participants studied one set of stories during an encoding phase. After a delay, in the test phase they received a new set of stories. For each story, their task was to write down any stories that they were reminded of from the encoding phase.

Each story in the test phase described the same relational pattern (e.g., positive feedback, reciprocity) as one story in the encoding phase and came from the same domain (e.g., medicine, political science) as another story in the encoding phase. Thus for each test story there were two likely retrieval candidates, one sharing the same relational pattern as the test story (the relational match), and one coming from the same domain as the test story (the domain match).

We expected that domain matches would be quite likely to be retrieved because stories from the same domain often involve both surface commonalities and associative connections. As reviewed earlier, retrieval is likely to be based on surface commonalities (e.g., Gentner et al., 1993; Holyoak \& Koh, 1987; Ross, 1987) and there is also evidence that the degree of association between examples affects retrieval (e.g., Howard \& Kahana, 2002; Pollio, Richards, \& Lucan, 1969; Wolfe, 2005). Assuming that domain retrieval would be dominant and relational retrieval relatively rare at baseline, we could assess improvement in relational retrieval when relational labels were added.

We varied whether participants received relational labels during the encoding and test phases in a $2 \times 2$ betweensubjects design (no relational labels at encoding or at test; relational labels at encoding only, relational labels at test only, or relational labels at both encoding and at test). The condition in which no relational labels were given during either the encoding or test phases provides a baseline of performance against which we could measure gains in relational retrieval. We tested whether (a) the use of relational labels can improve the likelihood of relational retrieval, and (b) whether this benefit occurs only when the
same relational labels are used both at encoding and at test, or whether there is also a benefit for the use of relational labels only during encoding or only during test.

The relational labels we used were schema noun labels that named the relational patterns described by the stories (e.g., positive feedback, reciprocity, inoculation). Schema nouns are a subtype of relational nouns, which name categories whose members share relational structure (e.g., Gentner \& Kurtz, 2005; Goldwater, Markman, \& Stilwell, 2011; Markman \& Stilwell, 2001). Since schema nouns name entire relational structures, they seemed particularly well-suited for promoting abstraction and a relational construal of the stories.

The most obvious prediction is that giving relational labels at encoding and at test should improve relational retrieval. This effect would be consistent with the interpretation that the label invited a similar relational construal for the two examples. Unfortunately, this result by itself is subject to another interpretation. Perhaps relational labels also act as a common surface feature between examples that share relational structure. If labels invite relational construal and/or act as a common feature, there should only be a benefit of relational labels on retrieval if labels are present at both encoding and test.

The two conditions of most interest are whether participants' relational retrieval improves when relational labels are given only during the encoding phase or during the test phase. If relational labels promote a construal in which there is a greater focus on the invited relational structure, then this may enable it to be retrieved more easily if the same relational structure (or a part of this relational structure) is encountered in future situations. If this is the case, then providing relational labels at encoding should also improve retrieval. It is possible that providing relational labels at test might also improve relational retrieval. If relational labels help people abstract the named relational structure and retrieve prior examples that share the same structure, then receiving labels during the test phase may also lead to improvement in relational retrieval. In sum, the two single labeling conditions allow us to examine the effects of selective abstraction on relational retrieval. The pattern of results will inform us about how relational labels might be having their effect.

## Method

## Participants

Participants $(N=60$, 39 female, mean age $=22.48)$ were recruited from the Northwestern University community and were paid or received course credit for their participation in the experiment. All participants were native English speakers.

Participants were assigned to one of the four conditions as follows: 15 participants received no relational labels at encoding or at test, 15 participants received labels only at
encoding, 16 participants received labels only at test, and 14 participants received labels at encoding and at test. An additional four participants were tested but excluded from further analyses, either for failing to follow the test instructions (one participant), or for failing to respond to at least half of the test items (three participants).

## Materials and Design

The materials consisted of two sets (A and B) of fourteen stories each that served as the encoding and test story sets (with A/B assignment counterbalanced). Each of these two story sets was made up of ten key stories and four filler stories. The ten key stories described relational patterns. The relational patterns were chosen to be applicable in different domains. For example, as described earlier, the relational pattern inoculation can appear in domain of medicine and also in the domain of psychology. Half of the key stories described causal systems (e.g., a positive feedback system), and the other half described other relational schemas (e.g., reciprocity). Some of the causal systems stories were adapted from an earlier study by Rottman, Gentner, and Goldwater (2012).

Within each of the two story sets (i.e., the encoding and test sets), each of the test stories was set in a different domain (e.g., mechanical engineering, political science, psychology). Across the two story sets, each test story matched one encoding story in terms of its relational structure, and matched a different encoding story in terms of its domain. For example, in one set, the story describing reciprocity was set in the domain of political science. In the other set, the story describing reciprocity was set in the domain of psychology, and the story set in the domain of political science described a different relational pattern.

As a check on our manipulation of domain-relatedness, we used Latent Semantic Analysis (LSA) to measure the degree of relatedness between the stories (Landauer \& Dumais, 1997). For each of the ten test stories in story set A, we calculated the LSA relatedness scores (Landauer \& Kintsch, 1998) to the story in set B that came from the same domain and to the story from set B that described the same relational pattern. Stories that came from the same domain had higher relatedness scores $(M=0.35, S D=0.20)$ than did stories that shared relational structure $(M=0.14, S D=$ $0.07), t(9)=3.38, p=.008, d=1.40$. This confirmed that, as intended, stories that came from the same domain were more semantically related than stories that described the same relational pattern.

Each of the two story sets also contained four filler stories. Across the two story sets, the filler stories were matched both in their relational pattern and in their domain setting.

The presence or absence of relational labels during the encoding and test phases was varied in a 2 (label present vs. absent at encoding) x 2 (label present vs. absent at test) between-subjects design.

## Procedure

There were two phases: an encoding phase and a test phase. Before the encoding phase, participants were informed that the experiment was composed of two parts. They were told that they would read a number of stories and that they would use the information they read about during the second part of experiment.

In the encoding phase, participants read one set of stories, either story set A or B. The order of the stories was randomized. If participants were in either of two labeling conditions (labels at encoding, labels at encoding and test), an additional sentence was added to the end of each story that described the relational pattern (e.g., "This is an example of reciprocity"). Participants were allowed as much time as they needed to read the stories.

After the encoding phase, participants completed a 15 minute filler task and then began the test phase of the experiment. In the test phase, participants received the set of stories that they did not receive during the encoding phase. For each story, they were asked to write down any of the original stories of which they were reminded. They were told that they could write down multiple original stories if they were reminded of them by one test story. Likewise, they could write down an original story multiple times if they were reminded of it by multiple test stories.

The test stories were presented one at a time on the screen, with a large text box below each story for participants to write down their responses. Participants were allowed as much time as needed to make their responses. As in the encoding phase, in the two labeling conditions (labels at test, labels at encoding and test), each of the stories ended with a sentence describing the relational pattern.

For each participant, we calculated two measures: the number of relational matches retrieved and the number of domain matches retrieved. A trained research assistant, who was blind to condition, coded the responses as relational matches, domain matches, or other responses. Other responses included extraneous retrievals that were neither domain matches or relational matches, and responses that were ambiguous or that did not provide enough information to be classified.

## Results

As predicted, relational labels led to more relational retrievals and fewer domain retrievals overall. Label condition had an effect on the number of relational matches, $F(3,56)=15.51 p<.001, \eta^{2}=0.454$, and domain matches, $F(3,56)=3.51, p=.021, \eta^{2}=0.158$, retrieved. As expected, Tukey HSD tests ${ }^{2}$ revealed that participants who received labels at encoding and at test retrieved more relational matches $(M=6.86, S D=2.45)$ than participants who received no labels $(M=1.87, S D=1.60), p<.001, d=$

[^102]2.52, participants who received labels at encoding ( $M=$ 4.13, $S D=2.23$ ), $p=.005, d=1.21$, and participants who received labels at test $(M=2.81, S D=2.04), p<.001, d$ $=1.87$ (see Figure 1). More interestingly, participants who received labels only at encoding also retrieved more relational matches than participants who received no labels, $p=.023, d=1.21$.

Domain matches showed a somewhat complementary pattern. Participants who received no labels retrieved more domain matches $(M=3.60, S D=1.96)$ than participants who received labels in both the encoding and test phases ( $M$ $=1.43, S D=1.55), p=.015, d=1.27$. There were no other differences between conditions in the number of domain matches retrieved. There were no differences between conditions in the total number of items (domain, relational, and other) retrieved, $F(3,56)=1.76, p=.166, \eta^{2}=0.086$.


Figure 1: The average number of relational matches retrieved by condition.

## Discussion

As predicted, we found that relational labels are able to improve relational retrieval. Relational retrieval improved over the baseline level when participants received relational labels at encoding and at test. More interestingly, relational retrieval also improved significantly when participants received relational labels only at encoding. There was evidence that relational labels can also promote a relational focus: participants who received relational labels at encoding and at test retrieved fewer domain matches than did participants who did not receive labels.

Earlier, we suggested that the use of known relational labels might improve relational retrieval by promoting a consistent relational construal of examples that share the same label. Further, we proposed that if an example is construed with a focus on the relational structure invited by the label, then this construal might be more likely to be
retrieved in response to another example that shares the same relational structure. Finally, we suggested that if a label promotes abstraction of relational structure, that this might improve the likelihood of retrieving prior examples that share this relational structure. We found evidence for the first and the second predictions, but not for the third.

The likelihood of relational retrieval was greatest when relational labels were present at both encoding and test, consistent with our hypothesis that labels promote a consistent relational construal (though, as noted earlier, it could also have resulted simply through labels acting as a common feature). Turning to the second prediction, we found that relational labels at encoding improved relational retrieval. This suggests that the labels promoted abstraction and storage of the named relational structure so that it could be retrieved more easily when a test example shared the same relational structure.

We did not find an effect of labels at test. One reason for this asymmetry between encoding and test may be that relational labels at encoding could have primed people to adopt a relational focus that carried over into the test.

At the start of the paper, we raised the question of why relational retrieval becomes more likely with domain expertise. Our findings suggest that one factor may be learning a technical vocabulary to name relational patterns in the domain.

The present findings have implications for learning and education. In order to promote relational retrieval and transfer, it could be useful to provide learners with labels for important relational structures. Our findings suggest that relational labels may highlight relational structure and make it more likely that it will be accessed again in the future. Additionally, the finding that relational retrieval is best when the same labels are used at encoding and retrieval suggests that labels for relational structures should be consistent. This fits with Forbus et al.'s (1995) claim that uniform relational encoding promotes relational retrieval.

Our findings are compatible with prior work demonstrating that other kinds of relational language can improve relational retrieval and transfer. For example, Clement, Mawby, and Giles (1994) found that using the same or synonymous verbs to describe relational structure in analogous situations improved the likelihood of relational retrieval. This suggests that using relational terms that invite a similar construal of situations, even if the labels are not identical, can increase the likelihood of noticing their similarity. Another finding related to this work is that receiving relational terms when learning about a new domain can improve the likelihood of relational transfer (Son, Doumas, \& Goldstone, 2010). In Son et al.'s (2010) studies, participants completed a tutorial about a domain they did not know about (Signal Detection Theory) and later solved transfer problems that involved the same principles. Half of the participants received relational terms (e.g., target, distracter, false alarm) in the tutorial. Since the
domain was new to participants, these particular uses of the terms were novel to them. The studies manipulated whether the semantics of the relational terms matched the tutorial scenario and whether the tutorial and transfer scenarios were easily alignable. The use of relational terms improved transfer performance the most dramatically when the semantics of the relational terms matched the tutorial and when the tutorial and transfer situations were easily alignable. These findings are consistent with the framework we presented earlier. If a relational term already had some initial stored meaning, then if it were applied in a new context (the tutorial scenario), this should lead to an alignment that resulted in the abstraction of their common relational structure. If participants then received the easily alignable transfer scenario and attended to its identical relational structure, then they should be able to retrieve the past case that shared this relational structure and transfer solution strategies from it to the transfer scenario.

In sum, the current research suggests that one way to improve relational retrieval is through the use of relational labels. This raises a number of interesting questions about how relational labels bring about this improvement. In this research, we investigated the effects of known relational labels. What effects would unknown or partially understood labels have? What would be the most effective way to introduce new vocabulary to improve relational retrieval and transfer? Would the benefits of relational labels on retrieval remain with greater delays? This line of research promises to shed light on the way in which symbolic learning and analogical processes combine in the acquisition of expertise.

## Acknowledgments

This research was supported by ONR Grant N00014-08-10040. We thank Laura Willig for her help in coding the data and members of the Cognition and Language Lab for their helpful suggestions.

## References

Barnett, S. M., \& Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. Psychological Bulletin, 128, 612-637.
Bransford, J. D. (1979). The role of prior knowledge. In J. D. Bransford (Ed.), Human cognition: Learning, understanding, and remembering. Belmont, CA: Wadsworth.
Brooks, L. R., Norman, G. R., \& Allen, S. W. (1991). Role of specific similarity in a medical diagnostic task. Journal of Experimental Psychology: General, 120, 278-287.
Clement, J. (1988). Observed methods for generating analogies in scientific problem solving. Cognitive Science, 12, 563-586.
Clement, C.A., Mawby, R., \& Giles, D.E. (1994). The effects of manifest relational similarity on analog retrieval. Journal of Memory and Language, 33, 396-420.

Forbus, K. D., Gentner, D., \& Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. Cognitive Science, 19, 141-205.
Gentner, D. (2003). Why we're so smart. In D. Gentner and S. Goldin-Meadow (Eds.), Language in mind: Advances in the study of language and thought. Cambridge, MA: MIT Press.
Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. Cognitive Science, 34(5). 752-775.
Gentner, D., Anggoro, F. K., \& Klibanoff, R. S. (2011). Structure-mapping and relational language support children's learning of relational categories. Child Development, 82(4). 1173-1188.
Gentner, D., \& Kurtz, K. J. (2005). Learning and using relational categories. In W. K. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, \& P. W. Wolff (Eds.), Categorization inside and outside the laboratory. Washington, DC: APA.
Gentner, D., \& Loewenstein, J. (2002). Relational language and relational thought. In J. Byrnes \& E. Amsel (Eds.), Language, Literacy, and Cognitive Development. Mahwah, NJ: LEA.
Gentner, D., Loewenstein, J., Thompson, L., \& Forbus, K. (2009) Reviving inert knowledge: Analogical abstraction supports relational retrieval of past events. Cognitive Science, 3, 1343-1382.
Gentner, D., \& Markman, A. B. (1997). Structure mapping in analogy and similarity. American Psychologist, 52, 4556.

Gentner, D., \& Medina, J. (1998). Similarity and the development of rules. Cognition, 65, 263-297.
Gentner, D., \& Namy, L. (1999). Comparison in the development of categories. Cognitive Development, 14, 487-513.
Gentner, D., Rattermann, M. J., \& Forbus, K. D. (1993). The roles of similarity in transfer: Separating retrievability and inferential soundness. Cognitive Psychology, 25, 524-575.
Gick, M. L., \& Holyoak, K. J. (1980). Analogical problem solving. Cognitive Psychology, 12, 306-355.
Gick, M . L., \& Holyoak, K J. (1983). Schema induction and analogical transfer. Cognitive Psychology, 15, 1-38.
Goldwater, M. B., Markman, A. B, Stilwell, C. H. (2011). The empirical case for role-governed categories. Cognition, 118, 359-376.
Holyoak, K. J., \& Koh, K. (1987). Surface and structural similarity in analogical transfer. Memory \& Cognition, 15, 332-340.
Howard, M. W., \& Kahana, M. J. (2002). When does semantic similarity help episodic retrieval? Journal of Memory \& Language, 46, 85-98.
Landauer, T. K., \& Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge.

Psychological Review, 104, 211-420.
Landauer, T., \& Kintsch, W. (1998). Latent Semantic Analysis. [Online computer software]. Available from http://lsa.colorado.edu/.
Markman, A. B., \& Gentner, D. (1993). Structural alignment during similarity comparisons. Cognitive Psychology, 25, 431-467.
Markman, A. B., \& Stilwell, C. H. (2001). Role-governed categories. Journal of Experimental \& Theoretical Intelligence, 13, 329-358.
McGuire, W. J. (1961). The effectiveness of supportive and refutational defenses in immunizing and restoring beliefs against persuasion. Sociometry, 24, 184-197.
McGuire, W. J., \& Papageorgis, D. (1961). The relative efficacy of various types of prior belief defense in producing immunity against persuasion. Journal of Abnormal Social Psychology, 62, 327-337.
Novick, L. (1988). Analogical transfer, problem similarity, and expertise. Journal of Experimental Psychology: Learning, Memory, and Cognition, 14, 510-520.
Novick, L. R., \& Holyoak, K. J. (1991). Mathematical problem solving by analogy. Journal of Experimental Psychology: Learning, Memory, and Cognition, 17, 398415.

Pollio, H. R., Richards, S., \& Lucas, R. (1969). Temporal properties of category recall. Journal of Verbal Learning \& Verbal Behavior, 8, 529-536.
Reeves, L. M., \& Weisberg, R. W. (1994). The role of content and abstract information in analogical transfer. Psychological Bulletin, 115, 381-400.
Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. Journal of Experimental Psychology: Learning, Memory, and Cognition, 13, 629-639.
Ross, B. H. (1989). Distinguishing types of superficial similarities: Different effects on the access and use of earlier problems. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15, 456-468.
Ross, B. H., \& Kennedy, P. T. (1990). Generalizing from the use of earlier examples in problem solving. Journal of Experimental Psychology: Learning, Memory, and Cognition, 16, 42-55.
Rottman, B. M., Gentner, D. \& Goldwater, M. B. (2012). Causal systems categories: Differences in novice and expert categorization of causal phenomena. Cognitive Science, 36(5), 919-932.
Son, J.Y., Doumas, L.A.A., \& Goldstone, R.L. (2010). When do words promote analogical transfer? Journal of Problem Solving, 3, 52-92.
Whitehead, A. N. (1929). The Aims of Education and Other Essays. New York: Macmillan Company.
Wolfe, M. B. (2005). Memory for narrative and expository text: Independent influences of semantic associations and text organization. Journal of Experimental Psychology: Learning, Memory, and Cognition, 31(2), 359-364.

# Verb bias and structural priming in non-linguistic grammar acquisition task 

Marius Janciauskas (m.janciauskas@liv.ac.uk)<br>Franklin Chang (franklin.chang@liverpool.ac.uk)<br>University of Liverpool, School of Psychology, Bedford Street South, Liverpool, L69 7ZA UK


#### Abstract

Domain-general statistical learning (SL) is thought to support language phenomena like verb bias and structural priming. We explored this idea by inducing these phenomena within a non-linguistic serial reaction time (SRT) task where participants learned an English-like artificial language using SL. In a series of two experiments we found error rates to be sensitive to verbs' structural preferences and abstract structural priming. The similarities between the behaviour in this task and previous linguistic research suggests that this method may be useful for studying the nature of SL in language learning and processing.


Keywords: statistical learning; verb bias; structural priming.
An important question in the study of language is the degree to which language acquisition depends on language-specific mechanisms or general-purpose statistical learning (SL) mechanisms (e.g., Kidd, 2012). Research has found that SL takes place in real and artificial language learning tasks (Fine \& Jaeger, 2013; Qi, 2012; Saffran, 2003, Wells, Christiansen, Race, Acheson, \& MacDonald, 2009). However, the use of language or auditory stimuli could cause language-specific systems to become activated in these tasks (Gervain, Nespor, Mazuka, Horie, \& Mehler, 2008). Non-linguistic artificial grammar learning (AGL) or serial reaction time (SRT) tasks provide a paradigm for studying grammar learning that is independent of linguistic knowledge. But the grammars used in the existing studies (e.g. Hunt \& Aslin, 2010) are quite different from real language and it is hard to link findings in these studies to human syntactic phenomena. Thus, it is still not known if domain-general SL can account for the acquisition and processing of human syntactic knowledge.

The present study set out to develop a method to study SL processes within a non-linguistic task designed to approximate the contexts in which certain linguistic phenomena occur. We developed an SRT task where participants had to implicitly learn statistical regularities in symbol sequences generated from an English-like grammar in a symbol-matching task. If participants learn this language as they process it (linguistic adaptation; Chang, Janciauskas, \& Fitz, 2012), then their accuracy and reaction times should reveal how linguistic phenomena arise out of general-purpose SL.

We applied our paradigm to explain two language phenomena: verb bias and structural priming. Verb bias is the tendency for individual verbs to prefer particular structures. For example, if a verb occurs more often in the double object (DO) structure as in "the man gave the woman the dress" rather than the prepositional dative (PD) structure "the man gave the dress to the woman", the verb is
said to have double object dative bias. This phenomenon is thought to occur as a result of learning distributional relationships between verbs and structures (Juliano \& Tanenhaus, 1994). In this example, the DO bias arises from stronger probabilistic association of the verb 'give' with the DO structure. A verb's occurrence in its preferred structure (verb-structure match hereafter) is known to influence structural choices and reduce processing time at the choice point where alternating structures diverge (Ferreira, 1996; Garnsey, Pearlmutter, Myers, \& Lotocky, 1997; Stallings, MacDonald, \& O'Seaghdha, 1998).

Another phenomenon of interest is structural priming, which is the tendency for participants to repeat previously produced sentence structures (Bock, 1986). For example, if participants heard the DO sentence like "the boy threw the dog the ball" and are then given a picture which can be described using a DO (e.g. "the man gave the woman the dress") or a PD (e.g. "the man gave the dress to the woman") structure sentence, they were more likely to use the same DO structure. Structural priming has been found to persist over time, suggesting that it is supported by learning (Bock \& Griffin, 2000). Chang, Dell and Bock (2006) used a connectionist model to show that priming could be explained as SL over abstract structural representations. Like verb bias, structural priming influences structural choices in sentence production and comprehension times at the post-verbal position (Corley \& Scheepers, 2002; Weber \& Indefrey, 2009).

In sum, verb bias and structural priming are thought to depend on SL processes involving linguistic units. If these processes are not specific to language then it should be possible to find verb bias- and structural priming-like effects in a non-linguistic SRT task. The present studies are a step towards such a paradigm.

## Study 1: Dative Alternation SRT Task

The first study used a variant of Hunt and Aslin's (2001) SRT study. In the centre of a computer screen participants saw sequences of letters appearing one at a time, which required them to find that letter on a circle of 21 letters surrounding the centre by moving a mouse cursor over it. The sequences were structured based on a grammar that included English dative alternation-like structures. For example, the symbol string "H J Z C M" approximated a PD sentence without articles like "man gave dress to woman". The corresponding DO symbol string was "H J M Z" ("man gave woman dress"). Verb bias was created by varying the frequency of the symbols (verbs hereafter) appearing in the verb's position with particular structures. For example, J and B occurred more often with PD structure, while D and

N occurred more often with DO structures. Structural priming was tested by manipulating the structures of adjacent strings (prime and target) so that half of them had matching structures (e.g. DO-DO) and half of them had mismatching structures (e.g. DO-PD). The input was created in sections of 24 items consisting of eight prime-target pairs separated by structurally unrelated fillers ( 120 items total). Twelve of the 16 PD and DO strings in each section contained a verb which matched its preferred structure and four strings contained a verb that mismatched it. These sections created temporal points at which the behaviours associated with learning structural constraints could be assessed.

Like in language studies, verb bias and structural priming effects were measured at the first 'post-verbal' symbol (e.g. after ' $G$ '), hypothesizing that verb-structure match and structural match between adjacent strings would reduce error rate and the reaction times taken to process that element. If these effects are the result of learning, we predicted that verb bias and structural match between adjacent strings would show a growing influence over the different sections of the study.

## Method

## Participants and materials

An opportunity sample of 79 participants was recruited from the University of Liverpool student population. The visual display consisted of letter symbols forming a circle (Fig. 1) and a space in the centre where stimulus strings were presented one symbol at a time. The language from which the strings were created consisted of 17 letters randomly allocated to 7 categories that resembled syntactic categories found in English language (Table 1). The categories were combined following English grammar rules to create grammatical letter strings (Table 2).

To test structural priming, presentation of the strings was structured so that PD and DO occurred in all combinations in pairs (prime and target) followed by one filler sentence of either intransitive (IN) or transitive (TR) structure. To create 'verb bias', the DVERBP (PD bias) and DVERBD (DO bias) categories occurred in PD and DO structures respectively $75 \%$ of the time.

The letter strings were generated by randomly selecting symbols from the appropriate categories with no overlap in symbols between adjacent strings with the priority given to the lower frequency members to ensure equal distribution. A total of 120 letter strings were used in the experiment. The development of verb bias and structural priming effects was tested every 24 items, which created 5 temporally different sections containing 8 instances of prime-target pairs.

After the letter-matching task people were given a grammaticality judgment test. Twenty-four randomly generated whole grammatical strings were presented side-by-side with another string that was identical to the target string but with two members belonging to the different
categories swapped to create ungrammatical transitions (e.g. MBHF and MHBF).


Figure 1: Visual display for Experiment 1 (left) and Experiment 2 (right, production trial)

Table 1: Category type, names, and symbols in Exp 1.

| Category Type | Category | Symbols |
| :--- | :--- | :--- |
| Animate Noun | ANOUN | X,M,Y,H |
| Inanimate Noun | INOUN | F,Z,Q,P |
| Intransitive Verb | IVERB | W,L |
| Transitive Verb | TVERB | S,G |
| Dative verb with PD bias | DVERBP | J, B |
| Dative verb with DO bias | DVERBD | D,N |
| Preposition | PREP | C |

Table 2: Rules used to create letter strings.

| Type | Category | Letter string <br> (English-equivalent) |
| :---: | :--- | :--- |
| IN | ANOUN IVERB | X W (Boys sleep) |
| TR | ANOUN TVERB | Y G Z (Girls like |
|  | INOUN | books) |
| DO | ANOUN | M B H F (Woman |
|  | DVERBD/DVERBP | showed boys car) |
|  | ANOUN INOUN |  |
| PD | ANOUN | H J Z C M (Man gave |
|  | DVERBD/DVERBP | dress to woman) |
|  | INOUN PREP ANOUN |  |

## Procedure

Participants were tested in a quiet room, with up to six people on individual computers per session. They were not told that the letter strings followed certain rules. They processed the strings by matching the letters appearing in the centre to those on the circle using a mouse on a letter-by-letter basis. Each response reset the position of the mouse cursor to the centre and triggered the next symbol. Letter strings were separated by a blank screen. After 120 items participants received a grammaticality judgment task (described above). Participants were told that the strings they saw earlier followed certain rules and that their task was to decide which of the two strings was grammatical. The experiment took approximately 20 minutes to complete.

## Data collection and analysis

Error rates and reaction times in milliseconds were recorded for the time taken to move the mouse cursor to the correct symbol on the circle. Verb bias and priming were tested after the DVERBD/DVERBP symbol ( $3^{\text {rd }}$ position). For the reaction time data, only correct items were used and responses were log-transformed. Reponses that were two standard deviations above or below the mean were removed.

## Results

The grammaticality judgment task was assessed using a one-sample $t$-test against chance ( $50 \%$, two-tailed). Participants successfully recognized $56 \%$ of grammatical strings $(t(76)=4.4 ; p<.001)$ showing that they had learned some structural aspects of the language.

The task produced a total of 36,340 responses with an error rate of $5.7 \%$. To assess the influence of verb bias on error rate, accuracy data (correct or incorrect response) were submitted to a binomial mixed model with verb-structure match (match vs. mismatch, effect coded) crossed with section (centered) as predictor variables. Participants were included as a random factor with maximal random structure (Barr, Levy, Scheepers, \& Tily, 2013). We found a significant two-way interaction ( $b=-0.22, S E=0.11, z=2.09$, $p=.04$ ), showing that verb-structure match reduced the likelihood of making an error and that this knowledge grew as the participant learned the language (Figure 2).


Figure 2: Mean proportions of errors (top) and reaction times (bottom) in each section when verb and structure matched (solid line) or mismatched (dashed line).

Similar mixed model assumptions were used for the reaction time analysis unless specifically mentioned. Reaction times were submitted to a mixed model. We found faster reaction times as the experiment progressed, showing a general learning effect ( $b=-0.03, S E=0.005, X^{2}=26.13$, $p<.001$ ). Verb-structure match increased processing times (Fig. 2; $\quad b=0.05, \quad S E=0.01, \quad X^{2}=19.02, \quad p<.001$ ). This
suggested speed-accuracy tradeoff but the exact nature of this effect remains to be established. No two-way interaction was observed.

For the structural priming error analysis, a binomial mixed model was used with structural prime-target match (match vs. mismatch, effect coded) crossed with section (centered) as predictor variables. Since verb bias was varied in these items, verb was included as an additional random factor and maximal models were fitted. We found no significant main effects or interactions. For reaction times, a main effect of section was observed ( $b=-0.03, S E=0.005$, $X^{2}=7.67, p=.006$ ) indicating that reaction times decreased as the experiment progressed, showing a general learning effect. In sum, participants implicitly acquired knowledge of symbol strings such that they were better than chance at judging their grammaticality. We found a growing verb bias effect in the error rates but no structural priming effects were observed, suggesting that people may not have learnt to distinguish the required structures well enough.

## Study 2: Semantic and Task Constraints

We postulated that the lack of a priming effect was due in part to the difficulty in distinguishing the PD/DO structures. In natural language, non-linguistic animacy provides a cue that enhances the distinctiveness of these structures (e.g. gave the dress/woman). In addition, the random position of the letters on the circle made anticipation more difficult. Therefore, we conducted a second experiment where symbols were grouped together (Figure 1b). To add animacy cues, we replaced the animate noun letters with the stick figures and the inanimate noun symbols with object-like symbols.

Since abstract priming is not always found in reaction times in comprehension (Tooley \& Traxler, 2010), we added a production-like string generation task (production hereafter), where participants occasionally saw the whole string in the centre and were required to produce it from memory by selecting the appropriate symbols in the circle (Figure 1b). Studies of human sentence production often use sentence recall to test verb bias or priming (e.g., Potter \& Lombardi, 1998).

Like before, we predicted that verb-structure match and structural prime-target match would influence processing times and error rates in both comprehension and production tasks. If these effects are learned over the study, they would increase over section.

## Participants and materials

39 participants were recruited from the pool of university students participating for course credits. The task was identical to Experiment 1 with the following changes. To aid category learning, letters belonging to the same category were grouped together on the circle (Figure 1b). ANOUN and INOUN letters were replaced with symbols providing semantic cues to those categories. To implement the production task, participants were shown the whole string, which disappeared once the mouse was moved. They were
then required to produce the string from memory by selecting the appropriate symbols as quickly as possible. Twenty-four production trials were added by replacing four comprehension trials (target strings) in each section. An additional section of 24 items was added for a total of 144 items.

## Results

Participants' grammatical knowledge at the end of the experiment was assessed using a one-sample $t$-test comparison against chance ( $50 \%$, two-tailed). They successfully recognized $64 \% \quad(t(41)=6.63, p<.001)$ of grammatical strings showing that they learned the language as in other artificial language learning studies.


Figure 3: Mean proportions of errors and reaction times in each section in production and comprehension tasks when verb and structure matched (solid line) or mismatched (dashed line).

The task produced a total of 22,464 responses, $5.5 \%$ of which were incorrect. Error data and reaction times were analyzed as in Experiment 1 with the addition of task type (production or comprehension, effect coded) fully crossed with the other variables. We found that error rates went down over section ( $b=-0.2, S E=0.06, z=-3.15, p=.002$ ) showing general learning effect. Participants made more errors in production than in comprehension ( $b=1.47$, $S E=0.23, z=6.34, p<.001$ ), but also improved more in production over sections $(b=-0.26, S E=0.12, z=-2.22$, $p=.03$ ). Finally we found that error rates were higher when verb and structure matched $(b=0.71, S E=0.21, z=3.36$, $p<.001$ ), which contradicts our prediction. However, this is due to the fact that the majority of errors belonged to the
target category ( $67 \%$ ), indicating that in most cases people anticipated the correct category but chose the wrong symbol. This was likely to be due to the grouping and visual similarity of the symbols.

The reaction time analysis revealed a general learning effect in which participants reacted faster across trials ( $b=-0.03, S E=0.003, X^{2}=122.38, p<.001$ ). Participants were also faster in production than in comprehension ( $b=-0.32$, $\left.S E=0.902, X^{2}=83.04, p<.001\right)$ due to task differences. Verbstructure match produced a significant main effect where reaction times decreased when verb matched its structure ( $b=-0.07, S E=0.01, X^{2}=21.42, p<0.001$ ). The mismatch with error rate resulted from speed-accuracy tradeoff where faster reaction times in verb-structure match condition resulted in more errors ( $b=0.001, S E=0.0004, z=2.95, p=.003$ ).


Figure 4: Mean proportions of errors in each section in production and comprehension tasks when prime structure was the same (solid line) or different (dashed line).

To examine structural priming, error rates and reaction times were submitted to similar mixed models as in Experiment 1 with the addition of task type fully crossed with other variables. A general learning effect was indicated by decreasing error rates over sections ( $b=-0.18, S E=0.07$, $z=-2.68, p=.007$ ). Participants produced more errors in production than in comprehension ( $b=1.14, S E=0.23$, $z=5.23, p<.001$ ), reflecting task demands. Finally, there was a three-way interaction between structural match, section and task type ( $b=-0.98, S E=0.24, z=-4.0, p<.001$ ), indicating that the reduction in error rates due to prime structure was greater in production relative to comprehension as section increased (Figure 4).

The reaction time analysis found a general improvement over section ( $b=-0.04, S E=0.004, X^{2}=84.13, p<.001$ ) and
faster responding in production $(b=-0.32, S E=0.02$, $X^{2}=11.12, p<.001$ ), reflecting a general learning effect and the nature of the task respectively. No priming effect was found in reaction times.

In sum, the production task and semantic grouping gave rise to structural priming in participants' errors. The fact that this priming is only evident at the end of the study suggests that participants had to learn structures before generalizing across the different strings (structural priming as language learning, Chang, Dell, \& Bock, 2006). Verb bias seemed to be present early in the experiment suggesting that grouping of the letters on the circle made verb bias acquisition relatively easy and did not allow capturing the growth of the effects over time.

## Discussion

This is the first study to provide evidence that verb bias and structural priming in a non-linguistic artificial language task arise from learning distributional constraints. Verb bias was found in Experiment 1, where participants were less likely to make errors when the structure matched the verb's preference. Structural priming was found in Experiment 2, where participants were less likely to make an error in producing a target sentence from memory if the previous sentence was of the same structure. Importantly, the prime and target shared no common symbols, so this effect cannot be due simply to the recall of particular symbol combinations. Both verb bias and priming effects increased over the experiment as participants learned the language, showing that these effects resulted from learning some language-related knowledge and not from some methodspecific features. This supports the prediction that such linguistic effects would also manifest in non-linguistic tasks, pointing to the commonalities in the underlying SL mechanisms (Chang, Janciauskas, \& Fitz, 2012).

One may note, however, that reaction times and errors showed conflicting results for verb bias, where in Experiment 1, verb-structure match created fewer errors, but slower reaction times, while the opposite pattern was observed in Experiment 2. The main difference between the two studies was the semantic similarity and grouping of the stimuli (verb bias did not interact with task type in Exp. 2). In Experiment 2, the semantic grouping meant that anticipation of the category that resulted from verb's bias (left for ANOUNs, right for INOUNs) resulted in faster reaction times, but also triggered more errors, particularly for the same category members. However, the exact cause for the patterns observed in Experiment 1 remains to be established but it is likely to be due to the differences in the way the letters were distributed on the screen in the two experiments. Interestingly, speech errors in natural language also exhibit speed-accuracy tradeoffs with speech rate (MacKay, 1982) and within-category effects (Dell, 1986) warranting further investigation of these effects in such nonlinguistic tasks.

Although our task is an artificial grammar-learning task, there are intriguing similarities with dissociations in human
verb bias and priming tasks. Errors in this study are related to structural choice in production tasks, because an "error" at the choice point can become a grammatical utterance depending on how the participant completes the sentence. Reaction time is related to graded measures like sentence initiation time or comprehension reading time. Effects of verb preferences on structural choice are well documented (Ferreira, 1996; Stallings et al., 1998), but the results for reaction times are mixed, with some studies finding facilitation (Garnsey et. al, 1997) and other studies finding no effect (e.g. Kjelgaard \& Speer, 1999). Likewise, structural priming results are robust in production (Pickering \& Ferreira, 2008), but abstract priming across verbs is less robust in comprehension studies (Arai, Van Gompel, \& Scheepers, 2007; Branigan, Pickering, \& McLean, 2005). Since the computational properties of the linguistic and non-linguistic mechanisms show these similarities, the differences observed in the way people express their knowledge in the comprehension-like task and the production-like task suggests that it may be beneficial to study these mechanisms within such an SRT paradigm, where both tasks are closely matched and the input tightly controlled.

To conclude, linguistic theories have long claimed that language involves specialized linguistic systems (Chomsky, 1965). These systems help to explain why verbs govern the structures that they appear in and how children acquire abstract syntactic representations from experience with word sequences. For these reasons, it should be difficult to use a domain-general visual-motor task to model the acquisition of a new language and find behaviours that mirror linguistic phenomena like verb bias and abstract structural priming. The fact that we observed these effects is particularly intriguing considering the short time taken to learn our language (our study took 20 minutes compared to 360 minutes in Hunt and Aslin, 2001). These difficulties were overcome in part due to the integrated learningprocessing approach taken here. The approach that is often used in SL studies involves separating testing from learning in order to test novel combinations that provide a strong test of abstract grammatical knowledge. Instead, we used linguistic adaptation of the existing representation in response to the input as evidence for abstraction and learning (e.g., structural priming). Since these items can be tested multiple times, it is possible to factor out individual variation and see changes as learning unfolds. The addition of semantic cues made it easier for participants to exhibit structural knowledge and allowed linguistic adaptation to take place at a higher level with these categories as lower level elements. Since our goal was to look at how representations change over time, rather than how they emerge from scratch, building semantics into the task is justified as children have an animacy distinction before they fully acquire structures like the dative (Gropen et al., 1989). Finally, an addition of a production task showed that the effects of learning manifest differently depending on the task that draws upon the acquired knowledge. In sum,
although there are still methodological issues to address, our results so far suggest that this task could be a way to examine the processes that take place in language production and comprehension.

## Acknowledgments

This research was funded by an ESRC Ph.D. studentship (grant number ES/J500094/1).

## References

Arai, M., Van Gompel, R. P. G., \& Scheepers, C. (2007). Priming ditransitive structures in comprehension. Cognitive Psychology, 54(3), 218-250.
Barr, D. J., Levy, R., Scheepers, C., \& Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language, 68(3), 255-278.
Bock, J. K. (1986). Syntactic persistence in language production. Cognitive Psychology, 18(3), 355-387.
Bock, K., \& Griffin, Z. M. (2000). The persistence of structural priming: Transient activation or implicit learning? Journal of Experimental Psychology: General, 129(2), 177-192.
Branigan, H. P., Pickering, M. J., \& McLean, J. F. (2005). Priming prepositional-phrase attachment during comprehension. Journal of Experimental Psychology: Learning, Memory, and Cognition, 31(3), 468.
Chang, F., Dell. S. G., \& Bock, K. (2006). Becoming syntactic. Psychological Review, 113(2), 234-272.
Chang, F., Janciauskas, M., \& Fitz, H. (2012). Language adaptation and learning: Getting explicit about implicit learning. Language and Linguistics Compass, 6(5), 259278.

Chomsky, N. (1965). Aspects of the theory of syntax. Cambridge, MA: MIT Press.
Corley, M., \& Scheepers, C. (2002). Syntactic priming in English sentence production: Categorical and latency evidence from an internet-based study. Psychonomic Bulletin \& Review, 9(1), 126-131.
Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. Psychological Review, 93(3), 283321.

Ferreira, V. S. (1996). Is it better to give than to donate? Syntactic flexibility in language production. Journal of Memory and Language, 35(5), 724-755.
Fine, A. B., \& Jaeger, T. F. (2013). Evidence for Implicit Learning in Syntactic Comprehension. Cognitive Science, 37(3), 578-591.
Garnsey, S. M., Pearlmutter, N. J., Myers, E., \& Lotocky, M. A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. Journal of Memory and Language, 37(1), 58-93.
Gervain, J., Nespor, M., Mazuka, R., Horie, R., \& Mehler, J. (2008). Bootstrapping word order in prelexical infants: A Japanese-Italian cross-linguistic study. Cognitive Psychology, 57(1), 56-74.

Gropen. J., Pinker. S., Hollander. M., Goldberg. R., \& Wilson, R. (1989). The learnability and acquisition of the dative alternation in English. Language, 65(2), 203-257.
Howard, D., Nickels, L., Coltheart, M., \& Cole-Virtue, J. (2006). Cumulative semantic inhibition in picture naming: Experimental and computational studies. Cognition, 100(3), 464-482.
Hunt, R. H., \& Aslin, R. N. (2001). Statistical learning in a serial reaction time task: Access to separable statistical cues by individual learners. Journal of Experimental Psychology: General, 130(4), 658-680.
Hunt, R. H., \& Aslin, R. N. (2010). Category induction via distributional analysis: Evidence from a serial reaction time task. Journal of Memory and Language, 62(2), 98112.

Juliano, C., Tanenhaus, M. K. (1994). A constraint-based lexicalist account of the subject/object attachment preference. Journal of Psycholinguistic Research, 23(6), 459-471.
Kidd, E. (2012). Implicit statistical learning is directly associated with the acquisition of syntax. Developmental Psychology, 48(1), 171-184.
Kjelgaard, M. M., \& Speer, S. R. (1999). Prosodic facilitation and interference in the resolution of temporary syntactic closure ambiguity. Journal of Memory and Language, 40(2), 153-194.
MacKay, D. G. (1982). The problems of flexibility, fluency, and speed-accuracy trade-off in skilled behavior. Psychological Review, 89(5), 483-506.
Qi, Z. (2012). Neurocognitive plasticity in verb bias learning in children and adults. Dissertation. University of Illinois.
Pickering, J. M., \& Ferreira, S. V. (2008). Structural priming: A critical review. Psychological Bulletin, 134(3), 427-459.
Potter, M. C., \& Lombardi, L. (1990). Regeneration in the short-term recall of sentences. Journal of Memory and Language, 29(6), 633-654.
Saffran, J. R. (2003). Statistical language learning mechanisms and constraints. Current Directions in Psychological Science, 12(4), 110-114.
Stallings, L. M., MacDonald, M. C., \& O'Seaghdha, P. G. (1998). Phrasal ordering constraints in sentence production: Phrase length and verb disposition in heavyNP shift. Journal of Memory and Language, 39(3), 392417.

Tooley, K. M., \& Traxler, M. J. (2010). Syntactic priming in comprehension: A critical review. Language and Linguistic Compass, 4(10), 925-937.
Weber, K., \& Indefrey, P. (2009). Syntactic priming in German-English bilinguals during sentence comprehension. NeuroImage, 46(4), 1164-1172.
Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J., \& MacDonald, M. C. (2009). Experience and sentence processing: Statistical learning and relative clause comprehension. Cognitive Psychology, 58(2), 250-271.

# Not so innocent: Reasoning about costs, competence, and culpability in very early childhood 

Julian Jara-Ettinger (jjara@mit.edu)<br>Joshua B. Tenenbaum (jbt@mit.edu)<br>Laura E. Schulz (lschulz@mit.edu)<br>Department of Brain and Cognitive Sciences, MIT<br>Cambridge, MA 02139 USA


#### Abstract

Social evaluations depend on our ability to interpret other people's behavior. In adults, these evaluations are influenced by our perception of the competence and motivation of the agent: helping when it is difficult to help is praiseworthy; not helping when it is easy to help is reprehensible. Here we look at young children's capacity to make competence attributions and its relation to their social evaluations. We find that as early as 18-months, infants can use the time and effort associated with achieving a goal-directed action to distinguish agents, and that infants prefer more competent agents. When asked to choose between two agents who act as moral bystanders and refuse to engage in a helpful action, we find a sustained preference for the more competent agent until the age of three, when the preference is reversed. We argue that the ability to calculate the cost and benefits of goal-directed action originates in early childhood and plays a fundamental role in moral reasoning.


Keywords: Action Understanding; Morality; Social Cognition; Theory of Mind.

## Introduction

The past decade has seen a revolution in our understanding of psychosocial reasoning in early childhood. Recent findings suggest that infants infer the false beliefs of others (Onishi \& Baillargeon, 2005; Southgate, Senju, \& Csibra, 2007; Kovács, Téglás, \& Endress, 2010), distinguish helpers, hinderers, and moral bystanders (Kuhlmeier, Wynn, \& Bloom, 2003; Hamlin, Wynn, \& Bloom, 2007); draw different inferences about actions directed towards member of in-groups and out-groups (Baillargeon et al., in press); predict actions based on social dominance (Thomsen, Frankenhuis, Ingold-Smith, \& Carey, 2011), judge third party agents transitively, based on how they interact with moral transgressors (Hamlin, Wynn, Bloom, \& Mahajan, 2011; Sloane, Baillargeon, \& Premack, 2012); and consider agents' knowledge about a target agent's preferences in making moral judgments (Hamlin, Ullman, Tenenbaum, Goodman, \& Baker, 2013). The discovery of infants' sophisticated social intelligence is among the most exciting recent developments in the field of cognitive science. However, to date relatively little is understood about the types of computations that underlie these social judgments.

## Rational Planning, Social Evaluations, and a Naïve Utility Calculus

Here we propose a new approach to thinking about social reasoning in infancy, drawing on the insight that the ability to reason about goal-directed action is at the core of our cognition about agents. (See Carey, 2009; Gergely \& Csibra,

2003 for review). Consistent with a large body of prior work, we assume that inferences about agents' goal-directed actions are governed by a principle of rational expectation: the idea that agents act efficiently to achieve their goals (e.g., Scott \& Baillargeon, 2013; Gergely \& Csibra, 2003). Computational work on the principle of rational expectation as probabilistic inference over rational planning has been used to successfully model adults' reasoning about agents' goals (Baker, Saxe, \& Tenenbaum, 2009, 2011; Ullman et al., 2010; Jara-Ettinger, Baker, \& Tenenbaum, 2012).

The principle of rational expectation is predicated on the understanding that agents act in ways that will minimize costs and maximize rewards. We propose that the ability to compute the costs and benefits of actions forms the heart of a naïve utility calculus that supports inference at the earliest stages of children's theories of agency. (See Jara-Ettinger, Gweon, Tenenbaum, \& Schulz in prep, for a detailed version of this argument and experimental studies in childhood). Here we provide an informal description of this approach and test one of its qualitative predictions: that an analysis of the cost functions associated with agent actions is central to the moral judgments even of very young children.

Intuitively, adult social evaluations are influenced by our perception of how much an action will cost the agent who performs it. Imagine for instance, that your neighbor, Sally, watches a child struggle to reach a package on the top shelf of a grocery store. Sally stands by and does nothing at all. Although there is no intrinsic relationship between height and moral worth, you may well judge Sally less harshly if she is $4^{\prime} 11^{\prime \prime}$ than if she is an NCAA Division 1 basketball player.

What analysis underlies this inference? We suggest that in evaluating and predicting agents' actions, observers automatically compute the cost of actions. The perceived cost of an action (controlling for constraints imposed by the environment) reflects inferences about the agents' level of competence; the perceived benefits of the action to the agent reflect inferences about the agents' level of motivation. Motivation and competence jointly affect the probability of the agents' actions so the two attributions trade-off with each other. If we know that an agent is highly motivated and she fails to act, we may infer that she is incompetent; conversely, if we know the agent is highly competent and she fails to act, we may infer that she is unmotivated. Morally, lack of competence to help is an exonerating factor; lack of motivation is not.

More generally social evaluations depend heavily on the agent's motivation (Cushman, 2008; Knobe, 2005; Young, Cushman, Hauser, \& Saxe, 2007). This may not be known and must then be inferred. If the agent performs a morally worthy action, then the higher our estimate of the cost of the action, the higher our estimate of the agents' motivation to act morally. Similarly, if the agent fails to act, then the lower our estimate of the cost of action, the lower our estimate of the agent's motivation. Ambiguity arises when the agent acts but the cost of action is very low, or the agent fails to act but the cost of action is very high; in such cases, we may be unsure of the agent's level of motivation. In our example, if Sally is $4^{\prime} 11^{\prime \prime}$, there is a high cost to reaching the shelf. This renders her failure to act ambiguous. Did she not want to help or was it simply too hard for her to do so? By contrast, if Sally is an NCAA basketball player, we can infer that the cost of reaching a shelf is low; thus we are more confident that her failure to act derives from a morally suspect lack of motivation.

We propose that these kinds of considerations are part of a general calculation of a cost function that, even early in development, is used to reason about goal-directed behavior and interpret agents' actions. However, to date no empirical work has looked at how differences in the cost function of agent actions affect children's evaluative and moral judgments. Similarly, no previous computational work has looked at how learners might compute the cost function of agent actions; work on goal inference has implicitly assumed that the cost function of actions is known (e.g., Baker et al., 2009; Ullman et al., 2010).

Here we test the prediction that very young children can estimate the cost functions associated with agents' actions and that this analysis affects children's moral judgments. In Experiment 1, we test the basic premise that children can use the perceived cost of actions to estimate agents' competence. We predict that at baseline children will prefer more (versus less) competent agents. In Experiment 2, we look at whether children can use differences in the cost of actions to infer differences in agents' motivations. We predict that when agents are moral bystanders, children may overcome their baseline preference for competent agents and be more likely to consider the merits of incompetent (but potentially more wellintended) agents.

## Experiment 1: Early Competence Attribution

In Experiment 1 we look at whether toddlers can use the time and effort associated with achieving a goal-directed action to estimate the cost of the action to the agents. We also look at whether toddlers have an early preference for competent agents.

## Participants

Twenty-four toddlers (mean age (SD): 21.19 months (97 days), range 16.8-28 months, 16 males) were tested at an ur-
(a) Experiment 1


Figure 1: Procedure for Experiments 1 and 2. Both experiments begin by introducing two puppets and a toy. One puppet (the Competent agent) was able to make the toy play music on the first attempt; the other puppet (the Incompetent agent) succeeded only after many attempts. In Experiment 1 (blue arrow), children were then asked to choose one of the puppets to play with. In Experiment 2 (green arrows), after the child saw both puppets activate the toy, the parent turned around and asked each puppet for help with the toy. Both puppets refused. As in Experiment 1, children were then asked to choose one of the puppets to play with.
ban children's museum ${ }^{1}$. Five children were excluded from analysis: four by decision of a blind coder and one for parental interference (See Results). All subjects were tested at an urban children's museum.

## Stimuli

Participants were shown two puppets and a yellow cylindrical toy with a black button at the top. The toy played music when the button was pressed.

## Procedure

Participants were tested in a quiet room at the museum. The child's parent was seated on a chair facing away from the testing table and the parent was asked to hold the toddler over his or her shoulder. Thus the child could see the stimuli but the parent could not.

Once the parent and toddler were positioned the experimenter presented the yellow toy to the child and introduced the two puppets. See Figure 1. He said, "Here are my two friends! They are going to show you how the toy works." Both puppets were continuously present throughout the experiment and each puppet approached the toy (order counterbalanced between participants) one at a time. The puppet said, "It's my turn!" and then pressed the button. When the toy activated, the toy played a song for approximately $10 \mathrm{sec}-$

[^103]

Figure 2: Results from Experiment 1: Number of children choosing each agent. $* *=p<.001$ by binomial test.
onds and then the puppet released the button. During this time, both puppets moved rythmically to the sound of the song. After releasing the button, each puppet said "Yay!" to celebrate the success.

The puppets differed in how many attempts it took them to activate the toy. The competent agent was always able to make the toy play music on the first attempt. The incompetent puppet tried several times to activate the toy (flattening his hand over the button but not depressing it fully). After the third or fourth failed attempt, the incompetent puppet backed away to look at the button, and then tried again. The incompetent puppet made a few more failed attempts and then successfully activated the toy. (The number of total attempts ranged from 6-8 trials across participants, allowing some flexibility in maintaining the child's attention to the task.) After the show, the parent was asked to turn around and to place their child at a marker on the middle of the edge of a lower table. The experimenter placed both puppets on opposite sides of the table equidistant from the child and asked the child which one she wanted to play with.

## Results and Discussion

All videotapes were coded by a coder blind to condition. Four children were excluded from analysis due to the coders' judgment that the puppets were not placed equidistant from the child. One additional child was excluded from analysis due to parental interference. The coder recorded the toddlers' first contact with a puppet following the prompt. If the child did not make a choice within a 30 -second window following the prompt, the experiment was ended. Three children did not make a choice. Of the 16 children who did make a choice, 15
preferred the competent agent ( $p<0.001$ by binomial test). See Figure 2.

In our design, the incompetent agent both made more attempts to activate the toy and took longer to activate the toy. Additionally, after some initial failures, the incompetent puppet studied the toy before trying again. Thus there were redundant cues to the agent's incompetence and we do not know whether toddlers' preferences were driven by the overall effort to achieve the goal, the time to achieve the goal (and thus perhaps the relative novelty of the puppet who achieved the goal more quickly), or a more abstract judgment about these factors as indices of competence per se. Future research might look at the range of factors that affect toddlers' inferences about the cost of agent actions. However, the result of Experiment 1 give strong evidence that by 18 months, children distinguish agents from differential cues to competence and prefer agents who appear to incur fewer costs to achieve a goal.

## Experiment 2: Competence and Social Evaluations

In Experiment 2, we look at how children's judgment of agent competence affects their social evaluation. Because pilot work suggested that the task in Experiment 2 was more demanding than the one in Experiment 1, we tested slightly older children: two and three-year-olds.

## Participants

Seventeen two-year-olds (mean age (SD): 30.8 months (83 days), range 26.6-34.9 months, 9 males); one was dropped from analysis for failure to make a choice. Thirty three-yearolds (mean age (SD): 42 months (104 days), range 36-50.09 months, 17 males) were recruited in the test condition; 7 were dropped from analysis, 4 by decision of a blind coder and 3 for failures to make a choice. An additional 9 three-year-olds (mean age (SD): 35.4 months (131.67 days), range 29.1-42.03 months, 4 males) were recruited for a control condition, 1 was dropped from analysis due to failure to make a choice. All subjects were tested at an urban children's museum ${ }^{2}$.

## Stimuli

The stimuli used in Experiment 2 were identical to stimuli used in Experiment 1.

## Procedure

The protocol in Experiment 1 was identical to the protocol in Experiment 2 with the following exceptions (See Figure 1). Because the children were older, they were given a choice of sitting in a small chair or standing in front of the testing table, behind the parent's chair. Additionally, before the experiment began, the parents were given a script to read telling them that when prompted to do so, they should turn around and pick up

[^104]Experiment 2: Children's preferences as a function of age


Figure 3: Children's choice of puppet in the test condition of Experiment 2 as a function of their age. The logistic regression with $95 \%$ confidence interval is shown on top. At early ages we find a preference for the competent agent, which disappears in older subjects.
the toy from the table. The experimenter would then place one puppet at a time in front of them. Parents in the test condition were instructed to ask each puppet: "Can you help me make the toy go?" Parents in the control condition were instructed to ask each puppet "Do you have a toy like this at home?"

Otherwise, the first part of the protocol proceeded as in Experiment 1. After both the competent and incompetent puppets successfully made the toy play music, but before the child was asked to make a choice, both puppets were removed and the toy was placed in the middle of the table. At this point the parent was asked to turn around. The parent picked up the toy and the experimenter returned a puppet to the middle of the table (order of puppets counterbalanced). Only one puppet was visible at a time. After the parent asked the puppet the target question, the puppet looked at the toy, then at the parent and said "No!" The puppet then turned around and hid under the table. This was repeated with the next puppet. To ensure that the child understood, in the test condition the experimenter said, "No one seems to want to help!" In the control condition he said, "No one seems to have this toy!" The questions and answers were then repeated with each puppet a second time.

After each puppet had said "no" twice, the experimenter took the toy from the parent and asked the child to stand on a marker in the center of a table edge. As in Experiment 1, the experimenter then set each puppet on opposite sides of the table, equidistant from the child and asked the child which puppet she would rather play with.

## Results and Discussion

Results were coded from videotape by a coder blind to conditions, as in Experiment 1. Children were excluded from analysis if, in the coder's judgment, the puppets were not placed equidistant from the child or if children did not make a choice within the 30 -second window, resulting in 162 -year-olds and 23 3-year-olds in the test condition and 83 -year-olds in the control condition (See Participants).

In the test condition, a logistic regression showed an effect


Figure 4: Number of subjects choosing to play with the competent (red) and incompetent agent (blue) in each age and condition of Experiment 2. $*=p<.05$.
of age on children's preferences: older children were more likely than younger children to choose the incompetent puppet ( $p<0.02$ ). See Figure 3 .

We followed-up with planned comparisons of the two and three-year-olds separately. See Figure 4. The two-year-olds showed a robust preference for the competent puppet. Of the 16 two-year-olds who made a choice, 12 chose the competent puppet ( $p<.05$ by binomial test). By contrast, the three-year-olds in the test condition chose between the puppets at chance; 13 of the 26 three-year-olds chose the competent puppet ( $p=n s$ by binomial test).

These results are consistent with the possibility that three-year-olds can use differences in agents' competence to attribute differences in agents' motivation, and can overcome their baseline preference for competent agents if agents fail to act helpfully. Arguably however, the three-year-olds chose at chance because they simply forgot which puppet was more competent (perhaps because the three-year-olds were more engaged than the two-year-olds by the puppets' refusals).

To see whether three-year-olds retained the competence information we looked at three-year-olds' performance in the control conditions. Failure to recall the more competent puppet seems unlikely to explain the results; preliminary results from the control condition suggest that the three-year-olds have no difficulty remembering which puppet was more competent when moral culpability is not at issue: 6 of the 8 three-year-olds showed a preference for the more competent puppet.

These results suggest that by the age of three, children can override a preference for competent agents if those agents act as moral bystanders. Given that it is morally objectionable to refrain from helping when a helpful action is relatively low cost, three-year-olds seem to be able to look more favorably on agents who have the excuse of incompetence to exonerate them.

The current findings are consistent with previous work sug-
gesting that both toddlers (Tomasello et al., 2005) and chimpanzees (Call, Hare, Carpenter, \& Tomasello, 2004) differentiate between agents who are unwilling to act helpfully from those who are unable to act helpfully. Here we also find that children distinguish competence to help from motivation to help. Critically however, participants in the earlier studies could assess the agents' motivation directly, from overt behavioral cues: serious, if failed, attempts to help indicated a motivated agent; "teasing" indicated an unmotivated one. Additionally, the "unable" agent in the earlier studies was genuinely unable: every attempt the agent made failed.

By contrast, in the current study, both agents were unwilling to help (both puppet said "no" and turned away from the parent) and neither agent was unable to help (both puppets were in fact able to activate the toy). Children could only evaluate the agents on the basis of graded differences in the agents' competence; however, this is precisely the kind of ability children should have if, as we have proposed, social reasoning in early childhood is informed by a naïve utility calculus, supporting computations of the costs and benefits of actions.

## General Discussion

Consistent with the idea that a naïve utility calculus is integral to children's understanding of agents, we found that inferences about the relative cost of agents' actions affect children's social evaluations from very early in development. Toddlers seem to be sensitive to cues associated with the relative competence of agents and prefer agents who achieve goals quickly and easily to agents who achieve the same goals at apparently higher costs. By the age of three, children seem to be able to use differences in agent competence as grounds for evaluating agents differently, even when the agents act identically in refusing to act at all.

As noted, we provided redundant cues to the competence and incompetence of these agents, including the time it took for each agent to make the toy play music, and the number of times each puppet pressed the button. We do not know to what extent toddlers' preferences were driven by each individual cue, or if their choice was guided by a more abstract representation of competence. Future research can shed light on the full range of cues we use to infer an agent's competence both in the physical domain and the epistemological domain, where some form of competence preference has also been found (Koenig, Clément, \& Harris, 2004; Pasquini, Corriveau, Koenig, Harris, et al., 2007).

There are several hypotheses consistent with the developmental change we observed between the age of two and three. One possibility is that toddlers distinguish competent and incompetent agents, but they do not infer that relative competence implies an obligation to act helpfully, or that relative incompetence exonerates an agent from such actions. A related possibility is that toddlers might make categorical distinctions between classes of behavior (e.g., "helping", "not helping", and "hindering") but make no distinctions within each cate-
gory; because both puppets in our paradigm refused to help, toddlers might find them equally blameworthy. A final intriguing possibility is that both two and three-year-olds can integrate judgments of agents' competence with moral judgments, but the children find themselves in a moral dilemma (and resolve the dilemma differently at different ages): they believe the incompetent agent is less culpable; however, they also believe it is a good idea to affiliate with competent agents. Future work is necessary to disambiguate these possibilities.

Additionally, because the children were given a forced choice between two agents, we do not know whether the one and two-year-olds' choices were based on a preference for the competent agent, an aversion to the incompetent one, or both. Similarly, we do not know whether the three-year-olds' choices reflect a relatively greater preference for the less competent (and therefore morally exonerated) agent, a relative devaluing of the more competent (and therefore morally culpable) agent, or both. Further research might disambiguate the specific attributions underlying children's preferences.

What this study does show is that human beings are sensitive to the cost of actions very early in development and form an early preference for competent agents. As children progress through early childhood, they become increasingly able to use inferences about an agent's competence to draw inferences about the agent's moral status. At an age when children themselves are still largely both incompetent and innocent, their ability to understand how the one characteristic might bear upon the other suggests remarkably sophisticated inferential abilities and highlights the importance of building a new theoretical synthesis for understanding the development social reasoning.

## Acknowledgments

We thank the Boston Children's Museum and the families who volunteered to participate. Thanks to Salvador Esparza, Eric Garr, and Aviana Polsky for help with data collection. We thank Rachel Magid for help with coding the data. Special thanks to Hyowon Gweon and Kim Scott for helpful comments and discussions. This research was funded by the National Science Foundation and the Simons Center for the Social Brain.

## References

Baillargeon, R., Scott, R. M., He, Z., Sloane, S., Setoh, P., Jin, K., et al. (in press). Psychological and sociomoral reasoning in infancy. APA Handbook of Personality and Social Psychology: Vol. 1. Atittudes and Social Cognition..
Baker, C. L., Saxe, R., \& Tenenbaum, J. B. (2009). Action understanding as inverse planning. Cognition, 113, 329349.

Baker, C. L., Saxe, R., \& Tenenbaum, J. B. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In Proceedings of the Thirtieth Third Annual Conference of the Cognitive Science Society (p. 2469-2474).

Call, J., Hare, B., Carpenter, M., \& Tomasello, M. (2004). unwillingversus unable: chimpanzees understanding of human intentional action. Developmental science, 7(4), 488498.

Carey, S. (2009). The origin of concepts. Oxford University Press, USA.
Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. Cognition, 108(2), 353-380.
Gergely, G., \& Csibra, G. (2003). Teleological reasoning in infancy: The naïve theory of rational action. Trends in Cognitive Sciences, 7(7), 287-292.
Hamlin, J. K., Ullman, T., Tenenbaum, J., Goodman, N., \& Baker, C. (2013). The mentalistic basis of core social cognition: experiments in preverbal infants and a computational model. Developmental science, 16(2), 209-226.
Hamlin, J. K., Wynn, K., \& Bloom, P. (2007). Social evaluation by preverbal infants. Nature, 450, 557-560.
Hamlin, J. K., Wynn, K., Bloom, P., \& Mahajan, N. (2011, December). How infants and toddlers react to antisocial others. Proceedings of the National Academy of Sciences of the United States of America, 108(50), 19931-6.
Jara-Ettinger, J., Baker, C. L., \& Tenenbaum, J. B. (2012). Learning what is where from social observations. In Proceedings of the Thirty-Fourth Annual Conference of the Cognitive Science Society (pp. 515-520).
Jara-Ettinger, J., Gweon, H., Tenenbaum, J. B., \& Schulz, L. E. (in prep). You can't always get what you want: Inferring competence from preference in early childhood.
Knobe, J. (2005). Theory of mind and moral cognition: exploring the connections. Trends in Cognitive Sciences, 9(8), 357-359.
Koenig, M. A., Clément, F., \& Harris, P. L. (2004). Trust in testimony: Children's use of true and false statements. Psychological Science, 15(10), 694-698.
Kovács, A. M., Téglás, E., \& Endress, A. D. (2010, December). The social sense: susceptibility to others' beliefs in human infants and adults. Science (New York, N. Y.), 330(6012), 1830-4.
Kuhlmeier, V., Wynn, K., \& Bloom, P. (2003, September). Attribution of Dispositional States by 12-Month-Olds. Psychological Science, 14(5), 402-408.
Onishi, K. H., \& Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? Science, 308(5719), 255258.

Pasquini, E. S., Corriveau, K. H., Koenig, M., Harris, P. L., et al. (2007). Preschoolers monitor the relative accuracy of informants. Developmental psychology, 43(5), 12161226.

Scott, R. M., \& Baillargeon, R. (2013). Do infants really expect agents to act efficiently?: A critical test of the rationality principle. Psychological Science.
Sloane, S., Baillargeon, R., \& Premack, D. (2012). Do infants have a sense of fairness? Psychological science, 23(2), 196-204.

Southgate, V., Senju, A., \& Csibra, G. (2007). Action Anticipation through Attribution of False Belief By Two-YearOlds. Psychological Science, 18, 587-592.
Thomsen, L., Frankenhuis, W. E., Ingold-Smith, M., \& Carey, S. (2011, January). Big and mighty: preverbal infants mentally represent social dominance. Science (New York, N.Y.), 331(6016), 477-80.
Tomasello, M., Carpenter, M., Call, J., Behne, T., Moll, H., et al. (2005). Understanding and sharing intentions: The origins of cultural cognition. Behavioral and brain sciences, 28(5), 675-690.
Ullman, T. D., Baker, C. L., Macindoe, O., Evans, O., Goodman, N. D., \& Tenenbaum, J. B. (2010). Help or hinder: Bayesian models of social goal inference. In Advances in Neural Information Processing Systems 22 (pp. 18741882).

Young, L., Cushman, F., Hauser, M., \& Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. Proceedings of the National Academy of Sciences, 104(20), 8235-8240.

# Developmental See-Saws: Ordered visual input in the first two years of life 

Swapnaa Jayaraman, Caitlin M. Fausey, \& Linda B. Smith<br>(\{swapnaa, cfausey, smith4\} @indiana.edu)<br>Department of Psychological and Brain Sciences, Indiana University<br>1101 East Tenth Street, Bloomington, IN 47405 USA


#### Abstract

The first two years of life are characterized by considerable change in all domains - perception, cognition, action, and social interactions. Here, we consider the statistical structure of visual input during these two years. Infants spanning the ages from 1 to 24 months wore body-mounted video cameras for 6 hours at home as they engaged in their daily activities. Our data strongly suggest that the statistical structure of the learning environment is dynamic and ordered. The available visual statistics are not stationary, but rather they are gated by young children's developmental level. We find a rolling wave of "See-Saw" patterns over developmental time in two classes of important social stimuli: First faces, then hands; and within hands, first other-then-self-then-touching-then-holding. These ordered environments may help learning systems "start small," find the optimal path to the optimal solution, and determine the architecture of the system that does the learning.


Keywords: natural statistics; first-person camera; faces; hands

## Introduction

Growing evidence across many domains indicates that human learners, including infants, are highly sensitive to the statistical regularities in the learning environment (Saffran, Aslin \& Newport, 1996) and that in many domains the regularities in the learning environment contain sufficient information to yield deep conceptual representations (Chater, Tenenbaum \& Yuille, 2006) of the kind that appear responsible for syntax (Griffiths, Steyvers \& Tenenbaum, 2007), semantics (Griffiths, et al., 2007), categories (Madole \& Oakes, 1999), and human visual object recognition (Logothetis \& Sheinberg, 1996; Quinn, Eimas \& Tarr, 2001). As methods for understanding structure in large data sets advance, it seems likely that we will discover rich insights in even broader domains about how the statistical structure of learning environments shapes human learning and knowledge. The research presented in this paper makes two contributions to this endeavor: by extending the study of the statistical structure of the learning environment to social stimuli - faces and hands; and by showing that the statistical structure of the learning environment is not stationary. Rumelhart and McClelland's (1986) model of the learning of the past tense was once famously criticized (Pinker \& Prince, 1988) as a cheat because the network was presented with learning examples in an ordered way rather than as batch statistics. However, the statistics of the learning environment can change substantially with development itself. The present findings provide one clear and dramatic example of this reality. We return to the more
general issue of what constitutes a developmentally appropriate conceptualization of statistical learning in the General Discussion.

## Faces and Hands as Important Social Stimuli

The first two years of life are characterized by considerable change in all domains - perception, cognition, action, and social interactions. Findings in all of these domains indicate the important role of other people, in scaffolding and supporting developmental process (Tomasello, 1988). Research into the social behaviors of mature partners that support infant learning have centered on two body regions face and hands. Research into the adult actions that infants attend to as guides to learning and understanding the world also focus on two body regions - faces and hands. And, indeed, a very large literature suggests that infants are highly sensitive to what are very small movements in these body regions - a shift in eye gaze (Butterworth \& Jarrett, 1991), a mouth opening (Moll \& Tomasello, 2012), a point (Leung \& Rheingold,1981), and a grasp (Woodward, 1998). It seems likely that faces and hands are everywhere in early infant experience. From the statistical learning perspective, this would mean that face and hand experiences present a very large data set for mining the structure and meaning of social gestures.

However, contemporary research is also consistent with the idea that this statistical learning might be modulated by internal (and innate, Meltzoff \& Moore, 1977; Slater, 1999) biases that privilege faces early in development. Research on infant face perception begins from the perspective that faces are a special class of stimuli. A large set of findings using diverse tasks indicate that very young infants are differentially sensitive to face-like visual stimuli relative to other stimulus categories (Goren, Sarty \& Wu, 1975; Johnson, et al., 1991) and can discriminate familiar faces from unfamiliar ones (Field et al, 1984) shortly after birth. Moreover, the earliest social interactions consist of face-toface play (Stern, 1971) and these have been characterized as "proto-conversations" that teach critical components of turn-taking and seem a likely context for learning about the facial cues that modulate social interactions and infant learning. Other evidence suggests that this early sensitivity to faces plays a critical role in tuning face perception processes: By 6 months infants recognize faces that are similar to those that have dominated their visual experiences (same race) better than faces that are dissimilar (different
race) (Kelly et al., 2007). Early visual deprivation appears to disrupt indices of face expertise such as configural processing of faces (Maurer, Le Grand \& Mondloch, 2002), and identification of faces with altered orientations and expressions (Geldart et al, 2002).

Systematic attention to hand actions as social indicators as pointers to objects to which infants should attend -- has been shown in 4 month olds (Rohlfing, Longo \& Bertenthal, 2012) but as far as we know this is the youngest demonstration of an understanding of a hand action. Most of the evidence indicating infant attention to and understanding of the meaning of hand actions - both in the context of language learning (Bates et al, 1989) and in the context of understanding the causal structure of events (Baldwin, 1991; Woodward, 1998) - focuses on older infants. For example, 10 and 12 months olds have been shown to use hand actions to predict causal sequences (Sommerville \& Woodward, 2005), 11 month olds have been shown to use the structure of a hand action to predict where an event will occur (Canon \& Woodward, 2012), in a large number of experiments 9 to 14 month olds have been shown to use points and other hand gestures to determine the intended referent of a heard word (Rader \& Zukow-Goldring, 2012), and 18 month olds may even understand hand gestures that mimic actions as pointers to objects and events more readily than words (Namy \& Waxman, 1998), as may 2-4 year olds (Hahn \& Gershkoff-Stowe, 2010). Several recent studies on 12 to 18 month old infants' attention in naturalistic contexts indicate that these older infants differentially - and perhaps systematically - look to the hand actions of mature partners when engaged in joint play with the parent (de Barbaro, Chiba \& Deák, 2011; Franchak, et al., 2010; Yoshida \& Smith, 2008), a result that suggests that these older infants know that hand actions contain important social information.

These findings indicate that infants may know about faces as sources of social information before they know about hands and suggest the following hypothesis: Although faces and hands are equally ubiquitous in the learning environments of infants, learning about these two classes of social cues is gated by infants’ early differential sensitivity to faces.

## Ordered input

Traditional approaches to statistical learning have concentrated on non-incremental learning tasks, tasks in which the entire training set is fixed at the start of learning and then is either presented in its entirety or randomly sampled. From this perspective, if learning needs to be constrained in some way or directed to some portion of the input, it must be accomplished by internal constraints on the learning system (Markman \& Hutchinson, 1984; Pinker, 1989), such as an innate sensitivity or interest in faces. But infants do not encounter the world as a single set of fixed statistics; they encounter it one learning instance at a time.

Because of this, they may encounter and experience selected regularities in just one small region of the total batchstatistics learning environment (Smith \& Gasser, 2005). Given the dramatic changes in the skills of human infants over the first two years of life, this seems highly likely. Given the ordered nature of these changes - first rolling over, then reaching, then sitting stably, then crawling, and then walking -- these selected statistics will also be ordered. West and King (1987) proposed the concept of ontogenetic niche: the idea that developmental level orders experiences in ways that constrains and canalizes developmental process. For example, in humans (and most mammal and some bird species) the young require constant caretaking and this constant caretaking limits as well as structures the input, and thus the regularities that can be learned one at a time. Humans' changing sensory motor abilities seem likely to constrain and expand visual experiences in different ways at different times. Human infants spend their first 6 months where others place them - on the floor, in infant seats, in a crib, in arms - and see what is in those places and what their mature caretakers care to show them. By 12 months, infants are much more masters of their own visual environments placing themselves in different locations and actively selecting what they will show themselves (Adolph et al., 2012).

These considerations raise an alternative hypothesis about faces and hands: Although faces and hands are equally ubiquitous in human environments, they are not equally ubiquitous in the visual environments of infants of different ages; instead, experiences of faces and hands are ordered, with dense experiences of faces characterizing the early ontogenetic niche and dense experiences of hands characterizing the later ontogenetic niche.

## Rationale for the present approach

The findings reported here are part of a larger program of research examining the statistical structure of natural visual environments as it relates to social cues and language learning. We build on the approach of a growing number of researchers using ego-centric cameras (Fathi, Hodgkins \& Rehg, 2012; Kanade, 2009) to capture first person visual environments. Studies of infants’ first person perspectives (mostly small laboratory studies, Aslin, 2009; Franchak et al., 2011; Smith, Yu, \& Pereira, 2011) have shown that these first person environments are characterized by properties of early visual experience that are not evident from third-person observer perspectives (Yoshida \& Smith, 2008) and have also documented the impact of infant body movements on infant visual experience (Kretch et al., 2012). Intriguingly, all the head-camera studies conducted with toddlers to date have noted that faces are rarely in the head camera images whereas hands - the child's and social partner's - are often in view (Franchak et al, 2011; Frank, 2012; Smith et al, 2011; Yoshida \& Smith, 2008). These studies, however, did not broadly sample the natural or
representative experiences of participants. The present study was designed to do just this. Infants spanning the ages from 1 to 24 months wore body-mounted video cameras for 6 hours at home as they engaged in their daily activities. The first questions we asked of these data, the results we report here, are these: How prevalent are faces and hands in the visual environment? Do the frequencies of faces and hands change systematically with development?

## Method: Capturing early visual environments

## Participants

23 infants and toddlers provided up to 6 hours of video each. This visual corpus consists of four subsets grouped by age. All videos within an age range are treated as a set.

Table 1: Infant and toddler visual corpus

| Age (months) | n | Hours of video | Frames coded |
| :---: | :---: | :---: | :---: |
| $1-3$ | 7 | 19.26 | 13,865 |
| $7-9$ | 5 | 21.88 | 15,754 |
| 18 | 6 | 22.64 | 16,303 |
| 24 | 5 | 14.59 | 10,505 |
| Total | 23 | 78.37 | 56,427 |

## Materials and Procedure

A small, lightweight camera was used to record the visual environments of infants and toddlers (Looxcie 2, Looxcie, Inc.). The diagonal FOV is 62 degrees with a 2 ":infinity depth of focus. The camera was secured to a wearable hat or harness. Parents were given a camera, hat and/or harness, and instructions about camera operations. Parents recorded up to 6 hours of video when their child was awake.

## Video pre-processing

Recorded videos were screened for private content and blank screens (e.g., camera was left turned on while not on child). Remaining videos were converted to images sampled at one frame for every five seconds of video. This first-of-akind corpus has approximately 78 hours of video and 56,000 frames of the natural visual environments of infants and toddlers in the first two years of life.

## Video coding: A reliable crowd-sourced approach

Frames were presented to coders on Amazon's Mechanical Turk (mturk.com) and analyses consider only those frames for which at least $75 \%$ of coders agreed (across all coding passes, $93.7 \%$ reliable judgments). Coding proceeded in six separate passes through the data. For each pass, coders viewed an instructions page with example images.

## Faces and Hands

The first broad passes coded for the presence of Faces and Hands. The infant (1-3, 7-9 months) and toddler (18, 24 months) data were coded with slightly different protocols. For infants, each coder saw up to eight frames and answered several questions about each frame. The two relevant questions were: (1) Do you see a human face or face part? and (2) Do you see other body parts or skin? If yes, which do you see? (a) bare hands/fingers, (b) bare feet/toes, (c) other body parts (neck, shoulder, knee, etc.), (d) body parts covered in clothes, (e) two or more of the above. In these analyses, only responses that indicated the presence of bare hands were further analyzed. Four unique coders judged each frame. For toddlers, each coder saw up to 100 frames and answered the same yes-or-no question for all frames. In separate passes, coders answered either (1) Do you see a human face in this picture? or (2) Do you see a human hand in this picture? Five unique coders judged each frame.

## Free, touching and holding hands

The next coding passes focused specifically on hands. First, we identified whether hands in the visual input belonged to the child or to someone else. Then, we identified whether the child's own hand was free, touching something, or holding a small object. In four distinct passes, coders answered one of these questions: Does any hand you see belong to the child wearing the camera?, Is the child's own hand touching something?, Is the child's own hand holding onto something?, Is the child's own hand holding something that can be carried?

## Results: Ordered visual input

## Body parts in the visual environment

How prevalent are faces and hands in the visual environments of infants and toddlers? The relative frequency of these two key body parts depends on the developmental stage of the child (Figure 1). The visual environments of the infants had more faces than hands (1-3 months: . 29 Faces, .01 Hands; 7-9 months: . 15 Faces, .06 Hands). For toddlers, hands were more prevalent than faces (18 months: . 11 Faces, .28 Hands; 24 months: 07 Faces, .32 Hands); $\chi^{2}(3, N=17962)=6936.84, p<.001$.

Faces and hands appear to trade-off, suggesting ordered visual input: Faces first, then hands. The developmental trend is not just increased variability of body parts in the visual input: The total proportion of faces and hands together is more stable across the first two years of life (.30, $.21, .34, .34$, for each age range respectively). The key finding is a "See-Saw" pattern: What is first available to infants (here, Faces; "See") fades to developmental history ("Saw") as infants creates new tasks for themselves with advancing motor, language and social skills.


Figure 1: Faces and hands in early visual environments

## Hands in the visual environment

What kinds of hands are potentially in view for infants and toddlers? The answer to this question changes over the first two years of life. Here, we focus on three kinds of developmentally relevant input: hand identity, contact between hands and the world, and holding small objects. For each, we find patterns of ordered visual input: What is visually available early is replaced by something else later.

Hands Identity: Other-to-Own Whose hands are available in the visual input to infants and toddlers? Over the first two years of life, the child's own hands are increasingly available (Figure 2a). Early, the kinds of hands in the visual environment are overwhelmingly other people's hands, but by toddlerhood the child's own hands are nearly half the available hand visual input (1-3 months: . 05 Own; 7-9 months: . 35 Own; 18 months: . 44 Own; 24 months: . 46 Own), $\chi 2(3, \mathrm{~N}=9096)=152.64, \mathrm{p}<.001$. The pattern is: First someone else's hands, then your own hands.

Hands Contact: Free-to-Touch When your visual environment includes your own hand, what else is potentially in view? Over the first two years of life, infants increasingly make manual contact with the world (Figure 2b). Early, hands are free -- flailing and reaching. The visual environment of 1-3 month-old infants does not include their own hands contacting the world. But, by 24 months, over two-thirds of the views of their own hands also include something they touch (1-3 months: 0 Touching; 7-9 months: . 68 Touching; 18 months: . 68 Touching; 24 months: . 77 Touching), $\chi^{2}(3, \mathrm{~N}=3936)=61.76, p<.001$. The pattern is: First hands free, then hands touching the world.

Hands on Objects: Touch-to-Hold How often do early visual environments include one's own hand holding an


Figure 2: Ordered visual input of Hands
object? Over the first two years of life, an increasing proportion of visual instances of touching the world are instances of holding objects (Figure 2c). Before 18 months, the visual environment includes few instances of hands together with objects (7-9 months: . 19 of all touching instances). Toddlers' visual input, however, includes many of these instances (18 months: .54; 24 months: .61). That is, infants and toddlers find themselves in very different visual circumstances with respect to hands-on-objects, $\chi^{2}(2, \mathrm{~N}=$ $2814)=141.61, p<.001$. The pattern is: First hands touching, then hands holding.

## General Discussion

Our corpus of visual environments is unprecedented in scope: We are capturing visual regularities throughout the first two years of human life. Importantly, we capture environments throughout these two years, rather than zooming in to focus on one unique time, or zooming out to collapse across many different times. Everyday acting and thinking happens within nested timescales and complete theories of how environmental regularities matter for human cognition demand evidence from each scale: from realtime measures of in-the-moment attention through summaries of long-term experience. Our project provides critical insight into a scale currently missing from theories of statistical learning: developmental time.

Our data strongly suggest that input is dynamic and ordered. Visual regularities in developmental time may be a rolling wave of "See-Saw" patterns. Here, we see this across two classes of important social stimuli - faces, then hands. We also see this within hands, going from Other-to-Self-to-Touching-to-Holding. If our investigation into early visual statistics had zoomed into 3-month-old infants, we would have missed important regularities about hands; if we had zoomed out to batch statistics over the first two years of life,
we would have concluded that the environments of infants and toddlers include roughly one-third body parts. Instead, we find a key pattern in environmental regularities: Developmental statistics are dynamic and ordered.

The available visual statistics are gated by young children's developmental level. A 3-month-old infant who is placed and carried finds herself in different visual environments than a walking, talking 24-month-old. Like other species, our data suggest that humans experience distinct ontogenetic niches as they progress toward adultlike motor, social, and language abilities. These visual niches may do a lot of important filtering for young learners: rather than sophisticated internal attentional control, "starting small" in structured input may be accomplished by other developmental constraints. Of course, the fact that developmental changes in many skills constrain the visual environment does not rule out the possibility of additional attentional gating. It may, however, reduce the challenges that attentional gating must resolve.

Does this temporal ordering of statistical regularities matter? It could be that outcomes at 2 years are best predicted by the total set of regularities and not by the order of those visual environments. Alternatively, some paths through the search space may be optimal, and mother-nature may optimize social learning by guiding the learner along optimal paths. More radically, the order of these experiences may not just enhance the optimal solution, but may determine the class of outcomes. Developmental process consists not just in the sampling of information but also in the change in the very internal structure of the learner. Considerable evidence from a psycho-biological perspective shows that the ordering and timing of sensory information play a critical role in brain development (Held \& Hein, 1963; Lord, 2012; Turkewitz \& Kenny, 2004). Reordering the usual sensory experiences within a developmental individual changes the architecture of the brain, not just what is known but what is knowable (Knudsen, 2006). A related idea, from cognitive theorists is the "starting small" hypothesis: limits that arise from the immaturity of the neural system constrain the input and, rather than holding back development, play a role in fostering development (Dominguez \& Jacobs, 2003; Elman, 1993; Fox, Levitt \& Nelson, 2010; Newport, 1990; Westermann, 2000). Between birth and 2 years, human infants travel through a set of highly distinct developmental environments determined first by their early immaturity and then by their growing emotional, motor, and cognitive competence. These ordered environments may help learning systems "start small," find the optimal path to the optimal solution, and determine the architecture of the system that does the learning.

## Acknowledgments

This work was supported by NICHHD 28675, NIH NRSA HD007475-18, and NIH R21HD068475. The authors wish to thank Char Wozniak and Ariel La for data collection.

## References

Adolph, K. E., Cole, W. G., Komati, M., Garciaguirre, J. S., Badaly, D., Lingeman, J. M., \& Sotsky, R. B. (2012) How do you learn to walk? Thousands of steps and dozens of falls per day. Psychological Science.
Aslin, R. N. (2009). How infants view natural scenes gathered from a head-mounted camera. Optometry and vision science: official publication of the American Academy of Optometry, 86(6), 561.
Baldwin, D. A. (1991). Infants' contribution to the achievement of joint reference. Child development, 62(5), 874-890.
Bates, E., Thal, D., Whitesell, K., Fenson, L., \& Oakes, L. (1989). Integrating language and gesture in infancy. Developmental Psychology, 25(6), 1004.
Butterworth, G., \& Jarrett, N. (1991). What minds have in common is space: Spatial mechanisms serving joint visual attention in infancy. British journal of developmental psychology, 9(1), 55-72.
Cannon, E. N., \& Woodward, A. L. (2012). Infants generate goal-based action predictions. Developmental Science.
Chater, N., Tenenbaum, J. B., \& Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. Trends in Cognitive Sciences, 10(7),287-291.
de Barbaro, K., Chiba, A., \& Deák, G. O. (2011). Micro-analysis of infant looking in a naturalistic social setting: insights from biologically based models of attention. Developmental Science, 14(5), 1150-1160.
Dominguez, M., \& Jacobs, R.A. (2003). Developmental constraints aid the acquisition of binocular disparity sensitivities. Neural Computation, 15(1), 161-182.
Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. Cognition, 48(1), 71-99.
Fathi, A., Hodgins, J. K., \& Rehg, J. M. (2012, June). Social interactions: A first-person perspective. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on (pp. 1226-1233). IEEE.
Field, T. M., Cohen, D., Garcia, R., \& Greenberg, R. (1984). Mother-stranger face discrimination by the newborn. Infant Behavior and Development, 7(1), 19-25.
Fox, S.E., Levitt, P., \& Neslon III, C.A. (2010). How the timing and quality of early experiences influence the development of brain architecture. Child development, 81(1), 28-40.
Franchak, J. M., Kretch, K. S., Soska, K. C., Babcock, J. S., \& Adolph, K. E. (2010, March). Head-mounted eyetracking of infants' natural interactions: a new method. In Proceedings of the 2010 Symposium on Eye-Tracking Research \& Applications (pp. 21-27). ACM.
Frank, M. C. (2012). Measuring children's visual access to social information using face detection. Proceedings of the 34th Annual Meeting of the Cognitive Science Society.
Geldart, S., Mondloch, C. J., Maurer, D., De Schonen, S., \& Brent, H. P. (2002). The effect of early visual deprivation on the development of face processing. Developmental Science, 5(4), 490-501.

Goren, C. C., Sarty, M., \& Wu, P. Y. (1975). Visual following and pattern discrimination of face-like stimuli by newborn infants. Pediatrics, 56(4), 544-549.
Griffiths, T. L., Steyvers, M., \& Tenenbaum, J. B. (2007). Topics in semantic representation. Psychological review, 114(2), 211.
Hahn, E. R., \& Gershkoff-Stowe, L. (2010). Children and adults learn actions for objects more readily than labels. Language Learning and Development, 6(4), 283308.

Held, R., \& Hein, A. (1963). Movement-produced stimulation in the development of visually guided behavior. Journal of Comparative and Physiological Psychology, 56(5), 872.
Johnson, M. H., Dziurawiec, S., Ellis, H., \& Morton, J. (1991). Newborns' preferential tracking of face-like stimuli and its subsequent decline. Cognition, 40(1), 1-19.
Kanade, T. (2009). First-person, inside-out vision. In IEEE Workshop on Egocentric Vision, CVPR (Vol. 1).
Kelly, D. J., Quinn, P. C., Slater, A. M., Lee, K., Ge, L., \& Pascalis, O. (2007). The other-race effect develops during infancy evidence of perceptual narrowing. Psychological Science, 18(12), 1084-1089.
Knudsen, E. I. (2006). Neural Derivation of Sound Source Location in the Barn Owl. Annals of the New York Academy of Sciences, 510(1), 33-38.
Kretch, K., Franchak, J., Brothers, J., \& Adolph, K. (2012). What infants see depends on locomotor posture. Journal of Vision, 12(9), 182-182.
Leung, E. H., \& Rheingold, H. L. (1981). Development of pointing as a social gesture. Developmental Psychology, 17(2), 215
Logothetis, N. K., \& Sheinberg, D. L. (1996). Visual object recognition. Annual review of neuroscience, 19(1), 577621.

Lord, K. (2012). A Comparison of the Sensory Development of Wolves (Canis lupus lupus) and Dogs (Canis lupus familiaris). Ethology.
Madole, K. L., \& Oakes, L. M. (1999). Making sense of infant categorization: Stable processes and changing representations. Developmental Review,19(2), 263-296.
Markman, E. M., \& Hutchinson, J. E. (1984). Children's sensitivity to constraints on word meaning: Taxonomic versus thematic relations. Cognitive psychology, 16, 1-27.
Maurer, D., Grand, R. L., \& Mondloch, C. J. (2002). The many faces of configural processing. Trends in cognitive sciences, 6(6), 255-260.
Meltzoff, A. N., \& Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. Science, 198(4312), 75-78.
Moll, H., \& Tomasello, M. (2010). Infant cognition. Current Biology, 20(20), R872-R875.
Namy, L. L., \& Waxman, S. R. (1998). Words and gestures: Infants' interpretations of different forms of symbolic reference. Child development,69, 295-308.
Newport, E.L. (1990). Maturational constraints on language learning. Cognitive Science, 14(1), 11-28.

Pinker, S. (1989). Learnability and cognition: The acquisition of argument structure. The MIT Press.
Pinker, S., \& Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. Cognition,28(1), 73-193.
Quinn, P. C., Eimas, P. D., \& Tarr, M. J. (2001). Perceptual categorization of cat and dog silhouettes by 3-to 4-monthold infants. Journal of experimental child psychology, 79(1), 78-94.
Rader, N. D. V., \& Zukow-Goldring, P. (2012). Caregivers’ gestures direct infant attention during early word learning: the importance of dynamic synchrony.Language Sciences.
Rohlfing, K. J., Longo, M. R., \& Bertenthal, B. I. (2012). Dynamic pointing triggers shifts of visual attention in young infants. Developmental Science.
Rumelhart, D. E., McClelland, J. L., \& CORPORATE PDP Research Group. (1986). Parallel distributed processing: explorations in the microstructure of cognition, vol. 2: psychological and biological models.
Saffran, J. R., Aslin, R. N., \& Newport, E. L. (1996). Statistical learning by 8-month-old infants.
Slater, A. M. (1999). Perceptual development: Visual, auditory and speech perception in infancy. Psychology Press.
Smith, L. B., Yu, C., \& Pereira, A. F. (2010). Not your mother's view: The dynamics of toddler visual experience. Developmental science, 14(1), 9-17.
Smith, L., \& Gasser, M. (2005). The development of embodied cognition: Six lessons from babies. Artificial Life, 11(1-2), 13-29.
Sommerville, J. A., \& Woodward, A. L. (2005). Pulling out the intentional structure of action: the relation between action processing and action production in infancy. Cognition, 95(1), 1-30.
Stern, D. N. (1971). A Micro-Analysis of Mother-Infant Interaction. Behavior Regulating Social Contact Between a Mother and her 3 1/2 Month-Old Twins. Journal of the American Academy of Child Psychiatry, 10(3), 501-517.
Tomasello, M. (1988). The role of joint attentional processes in early language development. Language Sciences, 10(1), 69-88.
Turkewitz, G., \& Kenny, P.A. (2004). Limitations on input as a basis for neural organization and perceptual development: A preliminary theoretical statement. Developmental Psychobiology, 15(4), 357-368.
West, M. J., \& King, A. P. (1987). Settling nature and nurture into an ontogenetic niche. Developmental psychobiology, 20(5), 549-562.
Westermann, G. (2000). Constructivist neural network models of cognitive development (Doctoral dissertation, University of Edinburgh).
Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. Cognition, 69(1), 1-34.
Yoshida, H., \& Smith, L. B. (2008). What's in view for toddlers? Using a head camera to study visual experience. Infancy, 13(3), 229-248.

# Chinese-speaking adults' understanding of argument structure 

Lu Jiang (kyouro2003@yahoo.co.jp)<br>Graduate School of Education, University of Tokyo<br>7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

Etsuko Haryu (haryu@ p.u-tokyo.ac.jp)<br>Graduate School of Education, University of Tokyo<br>7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan


#### Abstract

Syntactic constructions roughly correspond to sentence meanings. Previous research has shown that Chinese children can associate an SVO construction with a causative event at age 2, but do not always map an SV construction to a noncausative event even after reaching 5 years of age. The latter results may be attributed to the fact that Chinese allows argument-dropping (Jiang \& Haryu, 2010). This paper investigated Chinese adults' syntax-semantics knowledge and found that even adults do not always map an intransitive construction to a non-causative event, although they are likely to use an intransitive construction to describe a non-causative event. The results suggest that although Chinese adults understand that causative and non-causative events should typically be described using transitive and intransitive constructions, respectively, the use of this knowledge in inferring novel verb meanings seems to be regulated by the actual usage of SV sentences in Chinese.


Keywords: argument structure; Mandarin Chinese; adults; verb meanings; intransitive; transitive.

## Introduction

It is said that for children, learning verbs is difficult (e.g., Gentner, 1978, 1982; Imai et al., 2008). This is due to the fact that when a novel verb is introduced for a particular scene, there are an infinite number of possibilities concerning which aspect of the scene the verb refers to. For example, if we hear the novel verb "gorping" while watching a scene in which a girl is walking with a dog, the verb "gorping" may refer to "walking," "taking a dog for a walk," or "moving from one place to another." Even for adults, it is difficult to infer the meaning of a given verb if it is presented without any syntactic information (Gillette, Gleitman, Gleitman, \& Lederer, 1999). When they are told in what syntactic construction the verb appears, however, adults find it much easier to infer its meaning. Thus, the syntactic constructions in which verbs appear provide us with a very important cue to verb meanings, since the types of verb meanings roughly correspond to the syntactic structures in which those verbs appear (Gleitman, 1990). For example, a verb that appears in an intransitive construction with a single argument (e.g., "The boy goes") is likely to describe a non-causative event, while a verb that appears in a transitive construction with two arguments (e.g., "The girl pushed the boy") typically refers to a causative event.

Developmental psycholinguists have investigated whether and when children are able to use syntactic constructions to infer verb meanings. In recent studies, a forced-choice pointing task has been used to investigate this problem. In this task, children are typically presented with two videos side-by-side, one showing a causative event and the other a non-causative event, and asked to select a scene that matches a presented sentence involving a novel verb. Two types of test sentences are used: In one type, the novel verb is presented in a transitive construction, and in the other type, the novel verb is embedded in an intransitive construction. For example, while watching two events, a causative one in which a duck is pushing a bunny into a squat position and a non-causative one in which a duck and a bunny are moving one of their arms in a circle, children hear the novel verb "blick" in a transitive sentence such as "The duck is blicking the bunny," or in an intransitive construction with a conjoined noun such as "The duck and the bunny are blicking." The children are then asked to point to the event that matches the presented sentence. Thus, these studies have focused on whether children would select a causative event for a transitive sentence, and a noncausative event for an intransitive one.

These previous studies have found that English-learning 2-year-olds associate a transitive construction with a causative event. However, children of the same age do not always map an intransitive construction to a non-causative event (Arunachalam \& Waxman, 2010; Noble, Rowland, \& Pine, 2011). They become able to map an intransitive construction to a non-causative event by 3 years of age (Noble et al., 2011). Research that examined Chineselearning children using the same method also found that Chinese 2-year-olds were able to map a transitive construction to a causative event. However, Chinese children did not always associate an intransitive construction with a non-causative event even after reaching 5 years of age (Jiang \& Haryu, 2010). That is, in both English and Chinese, children seem to have some difficulty in acquiring knowledge of intransitive constructions, and their acquisition of intransitive constructions is later than that of transitive constructions.

The fact that it takes longer for children to become able to use intransitive constructions to infer verb meanings may be partly attributed to the fact that there are some verbs that have a general meaning and can be used in an intransitive construction but can refer to a causative event, not only in

English but also in Chinese. For example, the verb "play" can be used in an intransitive construction to refer to a causative event in which a girl makes a boy perform an action, by saying "The girl and the boy are playing." Thus, the existence of such intransitive verbs may contribute to the fact that both English- and Chinese-learning children need more time to acquire knowledge of intransitive constructions, compared to the time they take to acquire knowledge of transitive constructions.

Furthermore, a certain characteristic of Chinese might make it even more difficult for Chinese-speaking children to learn the correspondence between an intransitive construction and a non-causative event. Unlike English, Chinese allows pervasive ellipsis of noun arguments. Either or both the subject and the object can be dropped from the sentence. Therefore, in Chinese, an SV sentence could be either an intransitive sentence or a transitive sentence with the object omitted. As a result, SV sentences Chineselearning children hear in their daily life do not always refer to a non-causative event. This may also contribute to the difficulty that Chinese children have in learning the correspondence between an intransitive construction and a non-causative event. Given this characteristic of the Chinese language, it may also be the case that Chinese adults do not associate an intransitive construction with a non-causative event.

In the present research, two experiments were carried out to investigate whether Chinese-speaking adults associate a sentence with a single argument with a non-causative event in the same way that they associate a sentence with two arguments with a causative event. In Experiment 1, by presenting Chinese adults with two videos, one showing a non-causative event and the other a causative event, we examined whether they would map an SV sentence to a noncausative event, and an SVO sentence to a causative event, respectively. In Experiment 2, we presented Chinese adults with a video showing either a causative or a non-causative event, and asked them to select an appropriate sentence to describe the scene out of two types of test sentences, an SV sentence and an SVO sentence.

## Experiment 1

In Experiment 1, we investigated Chinese adults' syntaxsemantics knowledge using a forced-choice pointing task, which has been used in recent studies on children. In the current experiment, participants were presented with a novel verb placed in a transitive construction or in an intransitive construction with a conjoined noun ("the woman and the man") as the subject while they watched two events, one causative and the other non-causative. The participants were then asked to point to the event that matched the presented sentence. If, as argued by Jiang \& Haryu (2010), the pervasive ellipsis of noun arguments in Chinese makes it difficult not only for Chinese-learning children but also for Chinese-speaking adults to map an SV sentence to a noncausative event, then the adults would not map an
intransitive construction to a non-causative event, even when they assign a transitive construction to a causative event.

## Method

Participants The participants were 40 undergraduate students ( 20 males and 20 females, mean age 21 years, range 20 to 24 years). The participants were randomly assigned to two conditions: the intransitive condition and the transitive condition. In each condition, there were the same number of males and females. All the participants were native speakers of Mandarin Chinese.

Materials Six sets of videos were used (see Table 1 for details). They were the same videos that were used in the experiment with Chinese children in the previous research conducted by Jiang and Haryu (2010). Each set consisted of two videos, one showing a non-causative event and the other showing a causative one. In the non-causative event, a young woman and a young man performed the same repetitive action separately, side by side. In half of the causative videos, the young woman made the young man perform an action, while in the other half the man made the woman perform an action (see Figure 1 for an example).

As novel verbs, six monosyllabic nonsense words, the same ones in Jiang \& Haryu (2010), "xia3," "kao2," "pa3," "de4," "mu1," and "tie2," were used. Ten college students who were native speakers of Mandarin Chinese agreed that these words are senseless in that language. However, in the present experiment, these novel words were all low-pass filtered so that the sounds did not cause the participants to remember similar-sounding verbs that already exist in Chinese. The auditory stimuli were created by embedding these low-pass filtered words in the verb position of SV or SVO sentences, which were recorded clearly by a female adult native speaker of Mandarin Chinese.

Procedure Participants were tested individually using a forced-choice pointing task. The video stimuli were presented on a note PC using PowerPoint. All videos lasted about 10 seconds. In each test trial, a sentence was presented twice using Windows Media Player while the participant was watching two videos side-by-side, one showing a causative event and the other a non-causative event. The participant was then asked to point to the matching video. Participants in the intransitive condition heard a novel verb in an intransitive construction, such as "A1yi2 he2 shu1shu zai4 X (The woman and the man are Xing)," while those in the transitive condition were presented with a novel verb embedded in a transitive construction, such as "A1yi2 zai4 X shu1shu (The woman is X -ing the man)." Each participant received six test trials.

## Results and discussion

The selection of a causative event was scored as a causative response. The mean proportions of causative responses were calculated for each condition (see Figure 2).

Table 1: Stimulus materials used in Experiment 1

| Set | Non-causative events | Causative events | Novel <br> verb |
| :---: | :--- | :--- | :--- |
| 1 | A woman and a man sway side by side. | A woman tugs at a man's hand. | xia3 |
| 2 | A man and a woman move up and down <br> by bending their knees. | A man shakes a woman by the shoulders. | kao2 |
| 3 | A man and a woman twist their torsos <br> from left to right. | A man makes a woman bend down by <br> pressing on her shoulders. | pa3 |
| 4 | A woman and a man bow repeatedly. | A woman pats a man on his shoulder. | de4 |
| 5 | A woman and a man swing both of their <br> arms up and down together. | A woman turns a man's body in a circle. | mu1 |
| 6 | A man and a woman stamp their feet. | A man holds a woman's hand and waves <br> it. | tie2 |



Figure 1: A sample set of video events used in Experiment 1 (Set 1)


Figure 2: Mean proportions of causative responses in Experiment 1

The participants in the transitive condition selected causative events .88 of the time, which was significantly above chance level $(t(19)=10.48, p<.001, d=2.34)$. However, the participants in the intransitive condition
selected causative events .48 of the time, which was not different from chance $(t(19)=.64, p=.53, d=.14)$. That is, participants in the transitive condition matched a transitive sentence with two arguments to a causative event, while those in the intransitive condition did not always select a non-causative event for an intransitive sentence. Furthermore, an unpaired $t$-test revealed a significant difference between the two conditions, $t(38)=7.64, p$ $<.001, d=2.42$, indicating that participants in the transitive condition chose causative events much more frequently than those in the intransitive condition. The above results suggest that while Chinese adults assign a transitive construction to a causative event, they do not always map an intransitive construction to a non-causative event.

To summarize, when shown a non-causative and a causative event and asked to select which of the two events the given sentence described, Chinese adults were likely to select a causative event in response to a transitive construction, while they did not show a clear tendency to choose a non-causative event over a causative event in response to an SV construction. Their behavior was consistent with that of the young Chinese-speaking children in Jiang \& Haryu (2010).

## Experiment 2

The results of Experiment 1 suggest that Chinese adults think an SVO construction describes a causative event while at the same time they think that an SV construction can be used to describe not only a non-causative event but also a causative event. This belief seems to be consistent with the usage of SV sentences in Chinese. However, does this mean that there is no typical scene that should be described by using an SV construction? In Experiment 2, we investigated this problem by presenting the participants with a noncausative or a causative event and asking them to choose which of two given sentences, i.e., an SV sentence and an SVO sentence both containing the same novel verb, matched the event.

## Method

Participants Twenty undergraduate students, who were not tested in Experiment 1, took part in this experiment. The participants consisted of 8 males and 12 females (mean age 20 years, range 18 to 21 years), who were all native speakers of Mandarin Chinese.

Materials and procedure The visual materials were the same videos (six non-causative and six causative) as used in Experiment 1. In addition to the six nonsense words ("xia3," "kao2," "pa3," "de4," "mu1," and "tie2") used in Experiment 1, another six ones ("pei3," "ne1," "mai1," "diu4," "ka2," and "hua3") were also used as novel verbs. These 12 words were all confirmed as nonsense in Mandarin Chinese by 10 college students whose native language was Mandarin Chinese. Unlike Experiment 1, these words were not low-pass filtered, because participants in this experiment were asked to select one out of two sentences involving the same novel word, and thus the sounds of the novel words would not affect their performance.

While a video was shown, two sentences (i.e., an SV sentence and an SVO sentence, both involving the same novel verb) were presented. The participants were asked to select which sentence better matched the video. For example, when presented with a causative event in which a woman was tugging at a man's hand, the participants heard the intransitive sentence "A1yi2 he2 shu1shu zai4 Xia3 (The woman and the man are Xia3-ing)" together with the transitive sentence "A1yi2 zai4 Xia3 shu1shu (The woman is Xia3-ing the man)" and were asked to choose the one that matched the event. This procedure was repeated for 12 videos. That is, each participant received six causative trials and six non-causative trials.

## Results and discussion

We counted the number of responses in which the participants chose an intransitive sentence in response to a non-causative event, and a transitive sentence in response to a causative event, respectively. The mean scores each for the non-causative and the causative events were 5.7 ( $\mathrm{SD}=$ 0.57 ) and $5.7(\mathrm{SD}=0.57)$ out of 6 , respectively. Two $t$-tests were conducted to see whether these scores were
significantly above chance. The analyses revealed that participants were more likely to choose intransitive sentences to describe a non-causative event than expected by chance, $t(19)=21.14, p<.001, d=9.11$, and that they described a causative event by using transitive sentences more frequently than chance, $t(19)=21.14, p<.001, d=$ 9.11.

These results suggest that Chinese adults prefer SV sentences to SVO sentences when describing non-causative events, and use SVO sentences more often when referring to causative events.

## General Discussion

The present research examined Chinese-speaking adults' understanding of argument structure through two experiments that tested whether Chinese adults associate an SV and an SVO construction with a non-causative and a causative event, respectively. In Experiment 1, we found that when shown two events (a non-causative and a causative one) and asked to choose which one matched the presented sentence, Chinese adults were willing to map an SVO sentence to a causative event, while at the same time they did not always associate a given SV sentence with a non-causative event, which was also the case with Chinese young children (see Jiang \& Haryu, 2010). In contrast, when given two sentences (an intransitive and a transitive one) and asked to select which matched the given event in Experiment 2, Chinese adults were likely to assign a transitive sentence and an intransitive sentence to a causative event and a non-causative event, respectively.

The results of Experiment 1, together with those of Jiang and Haryu (2010), indicate that Chinese speakers, whether young children or adults, do not assume that SV sentences refer to non-causative events. At the same time, they think that SVO sentences describe causative events. This attitude in Chinese speakers is in contrast with what was found in English speakers who match SV constructions to noncausative events as well as matching SVO constructions to causative events (Arunachalam \& Waxman, 2010; Noble et al., 2011). This difference between Chinese- and Englishspeakers appears to come from the fact that English does not allow argument-dropping whereas Chinese allows pervasive ellipsis of noun arguments. Due to this particular property of Chinese, SV sentences could be either an intransitive construction or a transitive construction with the object dropped. Therefore, it is not appropriate for Chinese speakers to assume that a given SV sentence always refers to a non-causative event. In this respect, the behavior of Chinese speakers that do not automatically associate an SV sentence with a non-causative event seems to be in accordance to the actual usage of SV constructions in the language, indicating the possibility that the knowledge of argument structure is learned from the language input.

However, at the same time, as shown in Experiment 2, when Chinese adults are asked which of two constructions, an SV or an SVO construction, should be used to describe a
causative and a non-causative event, respectively, they answer that an SV construction rather than an SVO construction should be used to describe a non-causative event. In addition, they prefer to use an SVO construction rather than an SV construction to describe a causative event. This belief appears inconsistent not only with the Chinese input they have received, but also with the fact that Chinese speakers do not always map SV constructions to noncausative events.

Two possibilities may be considered as the origin of such asymmetrical behavior in Chinese speakers. The first possibility is that the knowledge of argument structure may be universal and innate, but one of the characteristics of Chinese (the fact that it allows the pervasive ellipsis of noun arguments) may guide people to regulate the use of this knowledge in inferring the meaning of a given sentence. Lidz, Gleitman, \& Gleitman (2003) argued for the view that the knowledge of argument structure is universal, based on their findings: Children learning Kannada make use of the number of noun arguments rather than morphological inflections as a cue to determine whether the given sentence refers to a causative or a non-causative event, although in Kannada not number of arguments but morphological inflections are definitive cues to the causativity of described events.

The other possibility is that owing to the pragmatic demands of communication, people independently of the properties of their native language prefer to use an SVO construction to describe a causative event and an SV construction to describe a non-causative event. When an SV construction such as "The woman and the man are playing" is used to describe a causative event in which the woman is making the man perform a certain action, what action the woman is in fact making the man perform is not known. In order to precisely convey what is happening, the different roles played by different agents should be described separately, using SVO constructions. On the other hand, using SV constructions may convey that all the agents play the same role in the event. Such pragmatic needs may guide people to prefer to use SVO constructions to describe causative events, and SV constructions for non-causative events, even though SV constructions do not always correspond to causative events in the Chinese input.

In sum, the present research has shown that although Chinese adults prefer to use SVO and SV constructions to describe causative and non-causative events, respectively, they do not always use this knowledge of syntax-semantics correspondences in deciding whether a given sentence refers to a causative or a non-causative event. The latter result suggests that the particular property of Chinese that allows argument-dropping might guide Chinese speakers not to automatically map an SV sentence to a non-causative event. However, despite this property of the Chinese language, why do Chinese speakers prefer to use SV and SVO constructions to describe non-causative and causative events, respectively? Is this because the knowledge of syntaxsemantics correspondences is universal, as suggested by

Lidz et al. (2003)? Or does it relate to the pragmatic demands of communication? Further research is required to investigate this question.

## References

Arunachalam, S., \& Waxman, S. R. (2010). Meaning from syntax: Evidence from 2-year-olds. Cognition, 114, 442446.

Gentner, D. (1978). On relational meaning: The acquisition of verb meaning. Child Development, 49, 988-998.
Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In S.A. Kuczaj (Ed.), Language Development: Vol. 2. Language, thought, and culture (pp. 301-334). Hillsdale, NJ: Erlbaum.
Gillette, J., Gleitman, H., Gleitman, L., \& Lederer, A. (1999). Human simulations of vocabulary learning. Cognition, 73, 135-176.
Gleitman, L. (1990). The structural source of verb meanings. Language Acquisition, 1, 3-55.
Imai, M., Li, L., Haryu, E., Okada, H., Hirsh-Pasek, K., Golinkoff, R., \& Shigematsu, J. (2008). Novel noun and verb learning in Chinese-, English-, and Japanesespeaking children. Child Development, 79, 979-1000.
Jiang, L., \& Haryu, E. (2010, November). Young Chinesespeaking children's understanding of the correspondence between verb meaning and argument structure. Paper presented at the Boston University Conference on Language Development, Boston, MA.
Lidz, J., Gleitman, H., \& Gleitman, L. (2003). Understanding how input matters: Verb learning and the footprint of universal grammar. Cognition, 87, 151-178.
Noble, C. H., Rowland, C. F., \& Pine, J. M. (2011). Comprehension of argument structure and semantic roles: Evidence from English-learning children and the forcedchoice pointing paradigm. Cognitive Science, 35, 963-982.

# Efficient codes for multi-modal pose regression 

Leif Johnson Joseph Cooper Dana Ballard<br>Department of Computer Science, The University of Texas at Austin, USA<br>\{leif, jcooper, dana\}@cs.utexas.edu


#### Abstract

Redundancy reduction, or sparsity, appears to be an important information-theoretic principle for encoding natural sensory data. While sparse codes have been the subject of much recent research, they have primarily been evaluated using readily available datasets of natural images and sounds. In comparison, relatively little work has investigated the use of sparse codes for representing information about human movements and poses. This paper proposes a basic architecture for evaluating the impact of sparsity when coding human poses, and tests the performance of several coding methods within this framework for the task of mapping from a kinematic (joint angle) modality to a dynamic (joint torque) one. We show that sparse codes are indeed useful for effective mappings between modalities and examine in detail the sources of error for each stage in the model.


## Overview

Recent work from machine learning (Ranzato, Boureau, \& LeCun, 20077; Lee, Battle, Kaina, \& Ng, [2007) and neuroscience (Olshausen \& Field, [996; Smith \& Lewicki, [2006) has emphasized the role that sparsity, or redundancy reduction (Barlow, 1961), appears to play when coding sensory data. Sparse codes are well suited to represent natural sensory data because the space of all possible inputs (e.g., all possible $1000 \times 1000$ images) is not uniformly covered by the samples (e.g., photos or retinal inputs) that we tend to encounter in the natural world. In fact, many types of natural data are theorized to lie along a low-dimensional manifold embedded in the larger space (Olshausen \& Field, [2004). Sparse codes are effective for representing data along such lowdimensional manifolds because the basis vectors in the code can be used efficiently (i.e., using just a few nonzero coefficients) to indicate, for a particular data point, its location along the manifold rather than its coordinates in the higher-dimensional space.

Many of the results in this area of research have focused on codes for sensory information like images and sounds. Concurrently, research in control theory has suggested that human movements might also lie along a relatively low-dimensional manifold embedded in the space of all possible movements (Scholz \& Schöner, [1999; Latash, Scholz, \& Schöner, [002). Sparse codes might be useful, then, for representing information about movement and pose in humans. To our knowledge, however, such codes have not been evaluated extensively on natural movement or pose data.

This paper proposes a basic architecture for testing the effectiveness of a broad class of coding techniques when mapping from kinematic (joint angle) to dynamic
(torque) data in human poses. While computationally straightforward, the model allows us to compare and evaluate several possible approaches to this coding and regression task. We show that, for the class of techniques captured by our model, sparsity is indeed useful for representing and manipulating pose data. In fact, even though the absolute decoding error associated with sparse codes can be larger than the corresponding absolute error for dense codes, the information captured by each coefficient in a sparse code is larger than for dense codes. In addition, sparse codes appear to facilitate the task of mapping or regressing from one information modality to another, making these codes particularly interesting from the perspective of a whole organism, which must integrate information from many different sources of information to make effective survival decisions.

## Problem setting

The human body is marvelously complex, with over 630 muscles and, by some estimates, more than 240 degrees of freedom (Winter, [200); Zatsiorsky \& Prilutsky, [2012). Despite this complexity, humans are skilled at controlling their bodies to make movements that accomplish a wide variety of tasks in the world. Humans also seem to be skilled at transforming information about movement between different modalities. For example, a person can normally mimic the posture of another person without much conscious effort, even though this task requires some sort of conversion from the visual configuration of the conspecific's body (possibly expressed in world or visual coordinates) to the kinematic configuration of their own. Along these lines, it is conceivable that the tasks of computing potential movements, evaluating proposed movements, and selecting and executing a particular movement all require separate ways of looking at the movements.

Studying human movements is difficult: the parameters describing movements are high-dimensional and time-varying, and, in addition, most of the quantities that are relevant for describing the control of movement are invisible to an outside observer. Although we do not have a practical way to observe the control signals or even accurately measure all of the joint torques or angles during a complex, multijoint human movement, we can use technologies like motion capture (Figure 『) to measure the external aspects of movement with high accuracy. Given motion capture data, Cooper and Ballard (2012) proposed a technique to compute the angles


Figure 1: Information processing architecture for multimodal coding and regression. Information from a frame of one modality of pose data, such as angles (orange), is to be mapped onto information from another modality, such as torques (blue). This mapping is accomplished by coding a frame of angle data, augmented with its derivative $\Delta$, using parameters $\mathbf{P}$; likewise, torques augmented with derivatives $\Delta$ are encoded using parameters $\mathbf{Q}$. Finally, a parametric regression $\mathbf{R}$ is computed between the codes (yellow and green).
and forces that would have been required for a simplified model of the human skeleton to effect the same movements. These computed angles and forces, while still a coarse proxy for some of the information that might be used by the central nervous system, constitute the data for this paper.

Theoretically, one could transform information from one modality into another by amassing a large quantity of corresponding data from these two modalities and computing regression coefficients directly. However, this is inefficient for at least two reasons. First, computing a regression between two datasets becomes increasingly problematic as the dimensionality of the data increases; this difficulty is compounded when there is noise in the data. Second, if the manifold hypothesis is accurate, then each modality of the raw data will have statistical redundancies that would need to be captured by the regression process. Rather than working in the space of raw measurements, then, we hypothesize that manipulating or combining movement information is more efficient in a space defined by codes that somehow represent the raw signals (Srivastava \& Salakhutdinov, [ШII2). The question addressed by this paper is, which types of codes are most efficient for processing information about movement?

## Pose coding and regression

We assume that we have a set of data that represents kinematic and dynamic views of human motion, modeled using an articulated body with $n$ degrees of freedom, and measured over a consecutive sequence of $m$ discrete time
steps. Formally, we represent a sequence of raw joint angles as a matrix $\mathbf{B} \in \mathbb{R}^{n \times m}$, where each column $\mathbf{b}^{(t)}$ represents a single frame of angle data. Similarly, we represent a sequence of raw joint torques as a matrix $\mathbf{U} \in \mathbb{R}^{n \times m}$ whose columns $\mathbf{u}^{(t)}$ each contain a frame of torque data. We define these matrices as complementary views of a single motion trajectory, so that for any frame $t$, the joint angles $\mathbf{b}^{(t)}$ correspond to the torques $\mathbf{u}^{(t)}$.

As mentioned above, movement is complex to model because it is high-dimensional ( $n$ is often large) and varies over time ( $m$ is often large). Rather than attempt to tackle both of these challenges at once, we simplify the modeling task here by considering the task of mapping between these two modalities for single poses (frames). Such a simplification makes the modeling task obviously difficult, since a single frame of kinematic pose data, for instance, does not indicate the direction in which the joint angles will be changing in subsequent frames. To address this issue, we make use of a common technique from speech recognition (Picone, [1993) and augment each of the raw data frames in our system with its first derivative. This provides information about the rates at which the angles and torques are changing, which could be useful when trying to compute torques on the basis of angles. The augmented data matrices $\mathbf{A}, \mathbf{V} \in \mathbb{R}^{2 n \times m}$ are then defined as

$$
\mathbf{A}=\left[\begin{array}{c}
\mathbf{B} \\
\Delta \mathbf{B}
\end{array}\right] \quad \text { and } \quad \mathbf{V}=\left[\begin{array}{c}
\mathbf{U} \\
\Delta \mathbf{U}
\end{array}\right],
$$

where $\Delta \mathbf{X}^{(t)}=\left(\mathbf{x}^{(t+1)}-\mathbf{x}^{(t-1)}\right) / 2$ represents a secant approximation to the derivative at each frame.

Having created a set of kinematic and matched dynamic data describing sequences of human poses, we propose an information processing architecture for computing regressions from angles to torques. In this framework (see Figure [) , a single frame of $n$ input angles, augmented with its derivative, is encoded first into $k$ angle-code coefficients using a coding model characterized by parameters $\mathbf{P} \in \mathbb{R}^{k \times 2 n}$. Then a regression model characterized by parameters $\mathbf{R} \in \mathbb{R}^{k \times k}$ transforms the $k$ angle-code coefficients into a $k$ torque-code coefficients. Finally, this torque encoding is converted back into a frame of $n$ torques, augmented with its first derivative, by inverting the torque coding model characterized by parameters $\mathbf{Q} \in \mathbb{R}^{2 n \times k}$. If the manifold hypothesis holds for human pose data, then code parameters $\mathbf{P}$ and $\mathbf{Q}$ can be learned independently, because these parameters will describe the structure of the manifold for each modality of pose data; codes for each manifold should then be useful for a wide variety of other information processing tasks (Vincent, Larochelle, Lajoie, Bengio, \& Manzagol, (20II). Regression parameters $\mathbf{R}$ can then be learned using the encoded data from each modality.

Given our parametric framework for coding and regression, the coding approaches considered here all as-
sume a finite "codebook" $\mathbf{D} \in \mathbb{R}^{2 n \times k}$ whose columns $\mathbf{d}_{i}$ each represent a "basis vector" that is used in some way to encode data. This paper does not focus specifically on learning the codebook, though it does mention a few approaches to codebook learning below.

Separating the model into coding and regression stages brings three advantages to the problem at hand. First, it allows us to manipulate the number of parameters in the model in a controlled way. The multistage model contains $k^{2}+4 k n$ parameters, while direct regression requires $4 n^{2}$ parameters. When $k<2 n$, the multistage model has fewer parameters than direct regression, but when $k$ exceeds the dimensionality of the data, the multistage model has more parameters. Models with more parameters tend to be more accurate, but they might overfit the data and capture more noise than desired. Second, separate modules for coding in each modality, and regression between codes, allows for in-depth analysis of the performance of each module: codes for one modality that provide for low decoding error could also be ones that do not permit easy regression, for example. Finally, defining distinct coding modules permit an analysis of the degree to which coding, in isolation, provides an efficient representation of the data.

## Coding algorithms

Mathematically, this paper treats coding as a general term for transforming a vector of raw data $\mathbf{x} \in \mathbb{R}^{2 n}$ into another vector of coefficients $\mathbf{z} \in \mathbb{R}^{k}$, such that $\mathbf{z}$ contains sufficient information to recover $\mathbf{x}$ with some tolerated level of error. More formally, coding is often defined in terms of minimizing a cost function

$$
\|g(\mathbf{D}, \mathbf{z})-\mathbf{x}\|_{2}^{2}+\lambda R(\mathbf{z})
$$

where $g(\mathbf{D}, \mathbf{z})$ refers to a decoding operation that converts coefficients $\mathbf{z}$ into an estimate of the raw data $\hat{\mathbf{x}}$, and $R$ is a regularizer that can be chosen to prevent overfitting, promote sparsity in the code, etc. For this paper, we evaluate several approaches to coding, each described briefly below.

PCA Principal component analysis (PCA) is widely used as a data preprocessing technique, but here we use PCA to refer to an encoding that simply computes the inner product of a data point $\mathbf{x}$ with each of the codebook vectors $\mathbf{d}_{i}$ : that is, $\mathbf{z}=\mathbf{D}^{T} \mathbf{x}$.

The PCA codebook consists of the eigenvectors of the covariance matrix of the data, sorted in decreasing order of magnitude of the corresponding eigenvalue. By retaining only the first $k$ eigenvectors in the codebook, PCA retains the maximally varying components of the data first, and progressively refines the error by retaining smaller components.

Used in this way, PCA implicitly models the underlying data as a multivariate normal distribution. Because the codebook for PCA is composed of eigenvec-
tors (which are orthogonal to each other), there can be at most as many codebook vectors as dimensions in the input data.

K-means K-means (MacQueen et al., 1967) can be seen as an extremely sparse coding technique that represents a data point $\mathbf{x}$ using only the closest basis vector in the codebook: for this approach, $\mathbf{z}=\left[\xi_{1} \ldots \xi_{k}\right]^{T}$ such that

$$
\xi_{i}= \begin{cases}1 & \text { if }\left\|\mathbf{x}-\mathbf{d}_{i}\right\|_{2}<\left\|\mathbf{x}-\mathbf{d}_{j}\right\|_{2} \text { for } j \neq i \\ 0 & \text { otherwise }\end{cases}
$$

The codebook corresponding to this coding approach is learned from the data by setting the $\mathbf{d}_{i}$ to random elements from the training data, and then iteratively adjusting the columns of $\mathbf{D}$ so that the sum of the Euclidean distances from each data point to its closest codebook vector is minimized.

Sparse coding Sparse coding (Tibshirani, 1996), also called lasso regression, computes the coefficients $\mathbf{z}$ for data point $\mathbf{x}$ by minimizing a least-squares cost function with a regularization penalty on the magnitude of the code coefficients:

$$
\mathbf{z}=\underset{\zeta}{\arg \min }\|\mathbf{D} \zeta-\mathbf{x}\|_{2}^{2}+\lambda\|\zeta\|_{1}
$$

$\lambda$ is a parameter that controls the tradeoff between accurate representation and sparsity; following results from the literature, we set $\lambda \propto 1 / \sqrt{n}$.

Coates and Ng (2011) reported that sparse coding worked well across many types of codebooks for their tasks (image classification). For this coding algorithm, then, we tested several different codebooks: random, sampled, and learned. The random codebook consisted of IID vectors drawn from the standard normal distribution, normalized to unit length. The sampled codebook contained samples drawn uniformly from the training data, also normalized to unit length. The learned codebook used a fast, online algorithm developed by Mairal, Bach, Ponce, and Sapiro ( 20019 ) to tune the codebook to the data. Briefly, the general algorithm is a variation of coordinate descent, where sparse encoding computations are alternated with codebook updates. The codebook learning process attempts to minimize the lasso cost function above, both with respect to the codes $\mathbf{z}$ and also with respect to the codebook $\mathbf{D}$.

## Regression

Once codes have been computed for the source and target datasets, the next task is to compute a regression matrix $\mathbf{R}$ that will convert coefficients from one modality into coefficients from another. We used ridge regression (Hoerl \& Kennard, 1970$)$ to compute the best parameters for inferring coefficients across coded modalities. We can express the regression task between codes $\mathbf{z}_{\alpha}$ and $\mathbf{z}_{\beta}$


Figure 2: The motion capture environment consists of a full-body motion capture suit (black, with red LEDs), and a treadmill centered in the motion capture space.
as optimizing the cost function

$$
\left\|\mathbf{R} \mathbf{z}_{\alpha}-\mathbf{z}_{\beta}\right\|_{2}^{2}+\lambda\|\mathbf{R}\|_{F}^{2}
$$

where $\lambda$ captures the degree to which the modeler is willing to tolerate large values in $\mathbf{R}$ when explaining the observed data. Essentially, ridge regression is the same as linear regression, but it adds a penalty on large values of the coefficients that are used to describe the data. In our experiments, the value of $\lambda$ was set empirically by cross-validation on the training set.

## Experiments

To measure and collect human movement data, we used a 16-camera Phasespace ${ }^{\text {m }}$ motion capture system in conjunction with a standard treadmill (Figure [7]). Human subjects in the motion tracking area wore a full-body suit equipped with active-pulse LED motion tracking markers and were recorded as they walked and ran on the treadmill at a variety of speeds.

For the results reported here, we recorded the positions of $L=48$ markers from one subject as he walked at speeds ranging from 0.22 to $2.68 \mathrm{~m} / \mathrm{s}$. The recording lasted twenty minutes. The Phasespace system produces frames of motion capture data at a rate of 120 Hz , so this recording resulted in more than 120,000 frames of raw motion-capture data. These frames were processed using the articulated forward model proposed by Cooper and Ballard (2012), resulting in three sequences of measurements for the observed motion: the sequence of interpolated marker positions $\mathbf{X}=\left[\mathbf{x}^{(1)} \ldots \mathbf{x}^{(N)}\right]$

[^105]representing the positions of the segments of the articulated model over time; the sequence of observed angles $\mathbf{A}=\left[\mathbf{a}^{(1)} \ldots \mathbf{a}^{(N)}\right]$ for each of the 54 degrees of freedom in the model; and the corresponding torques $\mathbf{V}=\left[\mathbf{v}^{(1)} \ldots \mathbf{v}^{(N)}\right]$ that were necessary to cause those angles to move through the observed dynamic trajectory of the model.

## Preprocessing

For this paper, we were concerned with mapping angles to torques, so we discarded the marker data $\mathbf{X}$. To obtain datasets for training and testing the coding and regression models, we needed to perform some preprocessing to obtain matched sets of frames that would permit a fair comparison.

First, the sequences obtained from the model were smoothed by convolving each channel in each modality with a 5 -sample ( 42 millisecond) rectangular window over time. After smoothing, each channel of the data was normalized by subtracting out the mean value and dividing by the standard deviation. These steps ensured that the data did not contain residual noise due to marker dropouts, and also that the data values were all approximately the same scale.

Each frame of data was then augmented with an approximation of its first derivative by calculating the secant approximation of these quantities using the neighboring two frames. ${ }^{\text {[] }}$

Next, the smoothed, normalized, derivativeaugmented frames were segmented into three distinct regions, each containing 24000 frames ( 200 seconds) of data: the first (segment A) consisted of slow walks, the second (segment B) consisted of fast walks, and the third (segment C ) consisted of running movements. To evaluate the coding and regression models, each segment was further partitioned into disjoint training, validation, and test sets such that $10 \%$ of frames from each segment were used for validation, $10 \%$ were used only for testing, and the remainder were available for training.

## Coding efficiency

We first analyzed the performance of the different coding techniques discussed above when reconstructing the raw torque data using the torque codes. Formally, after training the dictionaries as needed, we computed $\mathbf{z}$ for each frame of augmented torque data $\mathbf{v}$ in the test set, and then computed the decoding operation to obtain an estimate $\hat{\mathbf{v}}$. The decoding error $e_{\mathbf{v}}$ was then defined as the RMS value of the residual:

$$
e_{\mathbf{v}}=\sqrt{\frac{1}{n}\|\hat{\mathbf{v}}-\mathbf{v}\|_{2}^{2}}
$$

[^106]

Figure 3: Mean RMS decoding error for joint torques, measured with respect to the size of the codebook. Larger codebooks result in codes that capture more of the variance in the data, even when the codebook is created using IID standard normal samples. A log scale has been used on both axes to reveal trends more clearly.

Figure 3 shows the mean RMSE for each coding approach, measured with various sizes of codebooks, and applied solely to the torque data. (Results for the angle data were similar.) Unsurprisingly, larger codebooks were able to capture more of the variance in the data than smaller codebooks, regardless of the coding method. Perhaps more interesting, however, was the finding shown in Figure 74: when measured by the number of nonzero coefficients used in the code, sparse codes produced more accurate reconstructions than dense codes. This was somewhat vacuously true of K-means, since it only uses 1 coefficient for each $\mathbf{z}$; in comparison, however, this was not true for sparse coding combined with the random codebook.

## Predicting torques from angles

In addition to comparing the effectiveness of different coding schemes for torque data, we also used our framework to compare the encoding methods in a larger context, namely predicting torque values on the basis of angle values. This task could be seen as a coarse approximation for a control task: given a target kinematic pose, what might be the torques that would be associated with that pose?

Because the analysis framework proposed in this paper breaks down this task into three separate stagesencoding, regression, and decoding-we can analyze the regression component of the task separately from the other components. In general, RMS error for the regression task alone (Figure ${ }^{\text {a }}$ ) followed the same pattern as errors for the encoding and decoding components: larger codebooks tended to yield lower errors. However, sparsity played a critical role in this task, since K-means yielded the lowest regression errors, while PCA yielded


Figure 4: Mean RMS decoding error for joint torques, measured per nonzero coefficient in the encoding. Sparse codes like lasso regression were more effective, per coefficient, than dense codes like PCA, but only when the codebook was tuned to the dataset.


Figure 5: RMS regression error, measured with respect to encoded torque values.
relatively large errors.
Finally, we compared the overall torque regression error for all components of the model together, as measured by comparing the outputs from our processing model with the true torques measured during the experiment (Figure []). As a baseline, we computed a direct regression from angle to torque data: this resulted in an RMS error of 0.65 on the test set. PCA performed at baseline for complete codebooks, which is unsurprising since PCA simply rescales the data. However, some of the sparse coding approaches did outperform PCA by a large margin (up to $30 \%$ reduction in error). In particular, using lasso coding combined with a large, learned dictionary produced lower RMS errors than any of the other approaches examined here.


Figure 6: Mean RMS reconstruction error, measured with respect to true torque values for each frame. Note that the RMS reconstruction error for direct regression between true angles and true torques is 0.652 , which is approximately at the asymptote shown for PCA.

## Discussion

This paper presented an efficient coding and regression model for human pose information, and used this model to examine the performance of several coding algorithms on human pose information. The model allowed us to examine separately the errors in coding information about poses and in regressing from one modality to another. We learned that even though some approaches produce extremely low coding and decoding errors, and other approaches were conducive to learning regressions between codes, in order to perform well on the task of predicting information across information modalities, a coding approach must have extremely low error on both tasks.

In several ways this paper is just a first look at this sort of modeling on human pose information. In particular, we limited our examination of human pose information to snapshots of single moments in time. Movement, however, is fundamentally dynamic, so we plan to expand the techniques presented here to temporal sequences of poses, by learning codes for entire movements.

## References

Barlow, H. (1961). Possible principles underlying the transformation of sensory messages. Sensory Communication, 217-234.
Coates, A., \& Ng, A. (2011). The importance of encoding versus training with sparse coding and vector quantization. In Proc. 28th Intl. Conf. on Machine Learning (Vol. 8, p. 10).
Cooper, J., \& Ballard, D. (2012). Realtime, physicsbased marker following. In Proc. Motion in Games (pp. 350-361). Springer.
Hoerl, A., \& Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Tech-
nometrics, 12(1), 55-67.
Latash, M., Scholz, J., \& Schöner, G. (2002). Motor control strategies revealed in the structure of motor variability. Exercise and Sport Sciences Reviews, $30(1), 26-31$.
Lee, H., Battle, A., Raina, R., \& Ng, A. (2007). Efficient sparse coding algorithms. Advances in neural information processing systems, 19, 801.
MacQueen, J., et al. (1967). Some methods for classification and analysis of multivariate observations. In Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability (Vol. 1, p. 14).
Mairal, J., Bach, F., Ponce, J., \& Sapiro, G. (2009). Online dictionary learning for sparse coding. In Proc. 26th Intl. Conf. on Machine Learning (pp. 689-696).
Olshausen, B., \& Field, D. (1996). Emergence of simplecell receptive field properties by learning a sparse code for natural images. Nature, 381 (6583), 607609.

Olshausen, B., \& Field, D. (2004). Sparse coding of sensory inputs. Current Opinion in Neurobiology, 14(4), 481-487.
Picone, J. (1993). Signal modeling techniques in speech recognition. Proc. IEEE, 81 (9), 1215-1247.
Ranzato, M., Boureau, Y., \& LeCun, Y. (2007). Sparse feature learning for deep belief networks. Advances in neural information processing systems, 20, 1185-1192.
Scholz, J., \& Schöner, G. (1999). The uncontrolled manifold concept: identifying control variables for a functional task. Experimental Brain Research, 126(3), 289-306.
Smith, E., \& Lewicki, M. (2006). Efficient auditory coding. Nature, 439(7079), 978-982.
Srivastava, N., \& Salakhutdinov, R. (2012). Multimodal learning with deep boltzmann machines. In $A d-$ vances in neural information processing systems 25 (pp. 2231-2239).
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 267-288.
Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., \& Manzagol, P. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. The Journal of Machine Learning Research, 11, 3371-3408.
Winter, D. (2009). Biomechanics and Motor Control of Human Movement. Wiley.
Zatsiorsky, V., \& Prilutsky, B. (2012). Biomechanics of Skeletal Muscles. Human Kinetics.

# A soft barrier model for predicting human visuomotor behavior in a driving task 

Leif Johnson (leif@cs.utexas.edu)<br>Department of Computer Science, University of Texas at Austin, USA<br>Brian Sullivan (brians@ski.org)<br>Smith-Kettlewell Eye Research Institute, San Francisco, CA, USA<br>Mary Hayhoe (mary@mail.cps.utexas.edu)<br>Department of Psychology, University of Texas at Austin, USA<br>Dana Ballard (dana@cs.utexas.edu)<br>Department of Computer Science, University of Texas at Austin, USA


#### Abstract

We present a task-based model of human gaze allocation in a driving environment. When engaged in natural tasks, gaze is predominantly directed towards task relevant objects. In particular in a multi-task scenario such as driving, human drivers must access multiple perceptual cues that can be used for effective control. Our model uses visual task modules that require multiple independent sources of information for control, analogous to human foveation on different task-relevant objects. Building on the framework described by Sprague and Ballard (2003), we use a modular structure to feed information to a set of PID controllers that drive a simulated car and introduce a barrier model for gaze selection. The softmax barrier model uses performance thresholds to represent task importance across modules and allows noise to be added to any module to represent task uncertainty. Results from the model compare favorably with human gaze data gathered from subjects driving in a virtual environment.


Keywords: Visual attention; eye movements.

## Introduction

Humans routinely interact with complex, noisy, dynamic environments to accomplish tasks in the world. For example, while driving a car, a person navigates to a desired destination (e.g., grocery store) while paying attention to different types of objects in the environment (pedestrians, vehicles, etc.) and obeying traffic laws (speed limit, stop signs, etc.). Humans are able to balance competing task demands while simultaneously gathering information from the world through a foveated visual system, which must be actively moved to different targets to obtain high-resolution imagery.

During the deployment of attention, in particular overt eye movements towards an object, humans are sensitive to bottom-up salience (color, motion, etc.) as well as top-down task priority and the rewards associated with a task (Knudsen, 2007; Wolfe, Butcher, Lee, \& Hyle, 2003). In particular when engaged in "natural" tasks, eye movements are largely directed towards task relevant objects (Hayhoe \& Ballard, 2005; Land \& Hayhoe, 2001). Typically in natural environments, there are multiple task relevant objects spread over space and time that require active visual strategies to properly gather information. While human vision research has often focused on models of visual saliency, i.e., a stimulus based controller of attention (Bruce \& Tsotsos, 2009; Itti \& Koch, 2001; Zhang, Tong, Marks, Shan, \& Cottrell, 2008),
such models are inappropriate to address task-based behavior because they do not incorporate information about the state of the agent whose vision is being modeled. An alternative approach is to consider vision as part of a control process where information from the senses is used to guide motor behavior (Butko \& Movellan, 2010; Nunez-Varela, Ravindran, \& Wyatt, 2012; Senders, 1980; Sprague \& Ballard, 2003; Sullivan, Johnson, Ballard, \& Hayhoe, 2011). Both stimulus and taskbased approaches have led to a variety of formulations concerning how eye movements should be selected, e.g., using energy models, information theoretic measures, or measures of reward and uncertainty. In the present work, we focus on how selection of eye movement targets may be controlled in part by task related uncertainty and reward.

We present a model of visual processing and control that simultaneously takes into account the reward and uncertainty in multiple tasks associated with a dynamic, noisy driving environment. The model successfully accounts for variations in gaze deployment seen in humans driving in a virtual reality driving environment. Additionally, we discuss future research allowed by inversion of the soft barrier model. Inversion allows human data to be mapped into parameters in the model space so that it can be understood and compared quantitatively within the model framework.

## Model

The model proposed in this paper follows the modular architecture of Sprague and Ballard (2003) by factoring complex behaviors like driving into a set of simple control modules that each focus on a well-defined task-for example, a module to follow the road and another to avoid oncoming cars. Intuitively, a module is an abstract black-box controller that can be used alone to guide an agent through a single task. More interestingly, modules can be used together dynamically to engage in multiple ongoing behaviors. While the human visual system is highly parallel, processing and attentional focus are largely biased towards the fovea, meaning humans typically get information in a serial fashion by foveating different objects over time. In our model, multiple task modules run concurrently; however, to incorporate the foveation constraint, only one module at a time actively gains new percep-
tual information.
At a high level, modules are responsible for gathering and updating information about specific aspects of the state of the world, and for using that information to generate control signals for the agent. A central component of the model is that it requires a usable control policy even in the absence of updating its state information. Human short term memory decays with time, so to simulate this we allow the state information upon which the actions are computed to be corrupted by noise. We incorporate these into our model using simple scalar values for each module. Formally, we define a module as a tuple $M=\left(S, \mathcal{A}, \pi, \mathbf{s}^{*}, \rho, \varepsilon\right)$, where:

- $\mathcal{S}=\left\{s_{1}, \ldots, s_{n}\right\}$ is a set of the $n$ state variables that are relevant to the module,
- $\mathcal{A}=\left\{a_{1}, \ldots, a_{k}\right\}$ is a set of the $k$ action variables that are relevant to the module,
- $\pi: \mathbb{R}^{n} \rightarrow \mathbb{R}^{k}$ is a control policy that maps state values onto actions,
- $\mathbf{s}^{*} \in \mathbb{R}^{n}$ is a vector of target state values,
- $\rho$ is a scalar uncertainty threshold value for the module, and
- $\varepsilon$ is a scalar noise value for the module.

The first three elements of $M$ are common to typical Markov decision process (MDP) scenarios. The state space, spanned by elements of $\mathcal{S}$, represents all possible combinations of world state that are relevant for the task. The action space, spanned by elements of $\mathcal{A}$, describes all possible actions that the agent can take. The control policy maps states to actions; an optimal policy maps states to the best action for each state. The fourth element of the tuple, $\mathbf{s}^{*}$, is a vector of target values for each state variable. These target values are used in place of the more traditional formulation of scalar reward; this is discussed in further detail below. Finally, each module incorporates explicit values for both task priority $1 / \rho$ and task uncertainty $\varepsilon$, which are also explained below.

A learning agent is equipped with $N$ individual modules $M^{(1)}, \ldots, M^{(N)}$ that each specialize in one task and can be used in conjunction to control behavior in the world. To simplify the control problem, in our model all modules share a common set of action variables. In the driving environment described in this paper, there are two action variables: one represents changes in the vehicle's speed and another represents changes in the vehicle's heading.

## State estimates

Each module depends on a set of world-state variables that are relevant to the module's specific task. When driving, relevant state variables for a car-following task, for example, could include the agent-centric distance and angle to the leader car, the speed and heading relative to the leader car, etc. Relevant state variables for a target-speed task might be as simple as monitoring the absolute speed of the agent.


Figure 1: Evolution of state variable and uncertainty information for two single-variable modules. On the left, the solid blue lines represent the observed values of the state variable $A$ over time, while the shaded blue regions represent the region in which the true value of the state variable is likely to occur. On the right, the observed values and uncertainty regions are shown in green for a different state variable, $B$. Vertical dashes in each plot indicate times where the state estimates are updated with a new observation of the true value; these updates also reduce the magnitudes of the uncertainties in each estimate towards zero. The $\varepsilon$ parameter in this scenario is greater for the module tracking variable $A$ than for the module tracking variable $B$.

In MDP scenarios, agents are assumed to have constant access to accurate state variable information. Humans, on the other hand, have a foveated visual system that often requires active serial collection of updated state information. We assume that this serial process requires that when one visual task is accessing new information all other tasks must rely on noisy memory estimates.

To incorporate this state uncertainty into the model, each module $M^{(i)}$ maintains an explicit estimate of the current value of each of its state variables, $\hat{\mathbf{s}}^{(i)}(t)$. (We will henceforth omit the module superscript except to resolve ambiguities.) This estimate could be designed to incorporate many sorts of prior information about the evolution of the world, but the model in its current state simply treats state estimates as samples drawn IID from a spherical normal distribution

$$
\hat{\mathbf{s}}(t) \sim \mathcal{N}\left(\mu(t), \sigma^{2}(t) I\right)
$$

where $\mu(t)=\left[\mu_{1}(t) \ldots \mu_{n}(t)\right]^{T}$ is a vector of the most recently observed state values, and $\sigma(t)$ is the standard deviation for the state variable estimates in the module. Figure 1 shows the state updates over time for a simple, hypothetical system containing two modules, each tracking one state variable.
Uncertainty propagation Uncertainty propagates over time within each module by maintaining a small set of $J$ "uncertainty particles" $\mathcal{E}=\left\{\beta_{1}(t), \ldots, \beta_{J}(t)\right\}$. Each particle represents one potential path of deviation that the true state value might have taken from the last-observed state value. At every time step in the simulation, all uncertainty particles are displaced randomly by a step drawn from $\mathcal{N}(0, \varepsilon)$, thus defining a random walk for each particle. The root-mean-square value of the uncertainty particles is then used to define the standard
deviation

$$
\sigma(t)=\sqrt{\frac{1}{J} \sum_{j=1}^{J} \beta_{j}^{2}(t)}
$$

of state estimates for this module. Periodically, a module will be updated with accurate state information from the world (described below); when this happens, the magnitude of each uncertainty particle for the module is reduced according to $\beta_{j}(t+1)=(1-\alpha) \beta_{j}(t)$. After an informal parameter search, we set $\alpha=0.7$ for all modules; with $\alpha=1$, the model tends to produce many short updates because uncertainty is instantly reduced to 0 , but with $\alpha<1$ the uncertainty increase due to noise competes with the uncertainty reduction from the updated state information. Figure 2 shows the uncertainties over time for the hypothetical two-module system shown in Figure 1.

The state estimation approach described here can be seen as a sort of particle filter (e.g., Arulampalam et al., 2002), using an uninformed proposal distribution and equal weights for all particles. Interestingly, the behavior of the simulation was largely unaffected by the choice of $J$; for our simulation, we used $J=10$.

## Control policy

Each module relies on a policy to determine which action to take when the world is in a particular state. There are multiple ways an MDP may be solved for a control policy, e.g. in reinforcement learning a $Q$-table can be learned, which explicitly represents the expected future reward for each possible state and action combination; the policy is then given by a simple maximum over available actions for each state.

For a task like driving, however, continuous variables are the most natural representation of state (distance to another car, current speed, etc.) and action (change in speed, change in steering) variables. Although MDP algorithms can converge on policies for tasks in continuous spaces, for many real-world tasks the resulting policies can be more easily modeled using a simple parametric function. In addition, many algorithms for solving MDPs require significant training time to arrive at these regularly-shaped policies. The model described here instead uses a continuous proportional-integral-derivative (PID) control strategy.
PID controllers A PID controller $C(e)$ is a feedback control functional that maps state errors $e(t)$ onto control signals $u(t)$. Formally,

$$
C(e)=K_{P} e(t)+K_{I} \int_{0}^{t} e(v) d v+K_{D} \dot{e}(t)
$$

where $K_{P}, K_{I}$, and $K_{D}$ are parameters that affect the convergence speed and stability of the PID controller output when encountering a step change in error. In our model, these parameters are tuned manually for each module in isolation (O’Dwyer, 2006) to produce qualitatively appropriate driving behavior.

Each module in the model uses one PID controller for each state variable. Given estimates $\hat{\mathbf{s}}(t)$ of the current values of each variable and a vector of target state values $\mathbf{s}^{*}$, the control policy becomes

$$
\pi(\hat{\mathbf{s}}(t))=U\left[C_{1}\left(\hat{s}_{1}(t)-s_{1}^{*}\right) \ldots C_{n}\left(\hat{s}_{n}(t)-s_{n}^{*}\right)\right]^{T}
$$

where $U$ is a $k \times n$ mixing matrix that combines control policy recommendations from each PID controller into a final control output for each action variable. Note that, in this model, the control policy $\pi$ does not have access to the true state values $\mathbf{s}(t)$, but rather to the module's estimates of those state values $\widehat{\mathbf{s}}(t)$.

The composition of $U$ is determined by the needs of the modeling task. For the driving task, for example, each module generally has one state variable corresponding to a desired distance, and another corresponding to a desired heading. For this case, $U$ is set to the $2 \times 2$ identity matrix, since the PID controller that is monitoring a distance variable provides a natural control signal for vehicle speed, and the PID controller that monitors an angle variable provides a control signal for the vehicle heading. The exception to this is the module focusing on maintaining a target speed; this module only monitors current speed in the world, so it always provides a zero-output control value for the change-of-heading action variable.

## Priority

Modules can be prioritized by increasing their importance relative to one another, to allow modular agents to perform one task (for example, following a leader car) in preference to another (like achieving a target speed). In a traditional MDP scenario, this is modeled by controlling the ratio of reward values between two subsets of world states. In the present model, module priority is manipulated through the $\rho$ parameter: as $\rho$ increases, the module's relative priority decreases.

This relative priority value is incorporated into the model as a soft bound on the diffusion of uncertainty for each module. The specifics of this integration of uncertainty and priority are described in more detail next, as part of the perceptual arbitration process.

## Simulation

In simulation, an agent is placed in a two-dimensional virtual driving world. The world contains a single road with multiple lanes. Several non-agent cars are placed on the lanes at random locations, and one of the non-agent cars is designated as the leader car.

The basic simulation loop updates the state of the world at a fixed frequency $f_{s}$ (set to 60 Hz to match experimental conditions from (Sullivan, Johnson, Rothkopf, Ballard, \& Hayhoe, 2012)) according to an elementary physics simulation. At each time step, each car in the world moves ahead proportionally to its speed, in a direction given by its heading. For the non-agent cars, the simulator constrains these speed and heading values so that the cars always follow the lanes in the world.


Figure 2: Example random walks for uncertainty particles in the two modules shown in Figure 1. The individual particles are shown as small dots in each plot; their RMS uncertainty value $\sigma(t)$ is shown with a solid blue (left) and green (right) line. Vertical dotted lines indicate time steps when each module received an update; these updates reduce the magnitudes of the uncertainty particles towards 0 . The uncertainty threshold $\rho$ for each module is indicated by the shaded gray region in the center of the plot; in this example, the module tracking state variable $B$ (right) has higher priority (lower uncertainty threshold $\rho$ ) than the module tracking variable $A$. However, because the module on the left has a higher noise parameter $\varepsilon$, it receives more updates than the module on the right in the same duration of time.

After moving all vehicles in the world, the simulation additionally requests a control update from the learning agent, which changes the heading and speed of the agent before the next frame begins. Every time the simulator requests a control update for the learning agent, the modules are also updated by displacing their uncertainty particles according to each module's noise parameter $\varepsilon$.

## Perceptual Arbitration

If the simulation only performed the steps above, the agent's performance would become increasingly erratic over time, because the uncertainty particles would drift further away from zero. The resulting erroneous state value estimates would produce poor PID controller outputs, and the resulting actions chosen by the agent would further compound the uncertainty in the state estimates. In a human driver, this behavior would be analogous to taking one look at the world when getting into the car, and then driving with a blindfold thereafter.

Clearly this is not what humans tend to do when driving. Instead, people continually and regularly reposition their gaze toward objects in the environment as the driving task progresses. The final step in our model is to incorpate a scheduler to arbitrate between task modules, such that updated sensory information is delivered to the PID controllers dependent on task uncertainty and priority. Like Sprague and Ballard (2003), we hypothesize that this repositioning serves to reduce uncertainty about the state of relevant variables in the world-distance to a leader car, current speed, etc. To capture this behavior, the simulator periodically selects a module for receiving updated state information through a perceptual arbitration mechanism. This selection process happens at a constant frequency $f_{p}$ (set to 3 Hz for the results reported here
to approximate the frequency of human gaze behavior).
The perceptual arbitration process incorporates priority and uncertainty in the following way. We first define, for each module $M^{(i)}$, a weighted uncertainty at time $t$ that incorporates both the RMS uncertainty and the scalar priority of the module:

$$
\zeta^{(i)}(t)=\sigma^{(i)}(t)-\rho^{(i)}
$$

We also define a global variable $\phi(t)$ to represent the index of the module that gets updated at time $t$. Then the soft barrier model defines the probability that module $M^{(i)}$ is selected for update at time $t$ as a Boltzmann distribution over each of the priority-weighted module uncertainties:

$$
p\left(\phi(t)=i \mid \zeta^{(1)}(t), \ldots, \zeta^{(N)}(t)\right)=\frac{\exp \left(\zeta^{(i)}(t)\right)}{\sum_{j=1}^{N} \exp \left(\zeta^{(j)}(t)\right)}
$$

Intuitively, if the uncertainty in $M^{(i)}$ is currently above the threshold for that module-that is, if $\sigma^{(i)}(t)>\rho^{(i)}$-then $M^{(i)}$ is much more likely to be selected for update than another module, especially if none of the other modules have uncertainties exceeding their thresholds. However, the softmax selection process allows for nondeterminism: even if $\zeta^{(i)}(t)>\zeta^{(j)}(t)$ for $j \neq i$, there is some nonzero probability that $i$ will not be selected for update at time $t$. Finally, because module updates are always selected at frequency $f_{p}$ by sampling from the above distribution at the appropriate time, a module might be selected for update even if none of the agent's task modules have exceeded their uncertainty boundary (i.e., if $\zeta^{(i)}<0$ for all $i$ ).

Although inspired by diffusion models of decision making, this model contrasts somewhat with traditional models. Many diffusion models with "hard" bounds were developed for forced-choice, two-alternative tasks (e.g., Carpenter \& Williams, 1995); our model, in comparison, is designed to incorporate a wider variety of tasks. The "soft" barrier, driven at a fixed frequency, can incorporate more than two choices into the model simultaneously, while accounting for biologically realistic delays in planning and executing saccades.

As described above, and illustrated in Figures 1 and 2, when a module is selected for update, it is provided with the true state of each world state variable in $S^{(i)}$, and each of its uncertainty particles $\beta_{j}$ is reduced towards zero for every simulation frame until another module is selected for update.

## Simulation results

We implemented the model described above ${ }^{1}$ and ran several simulations to assess its qualitative behavior. Our simulated driving environment was identical in layout to the virtual environment used by Sullivan et al. (2012), so that we could directly compare our results to human performance. Our implementation consisted of three modules: a "speed" module

[^107]$M^{(s)}$ that attempted to drive at a particular target speed; a "follow" module $M^{(f)}$ that attempted to follow a lead car, and a "lane" module $M^{(l)}$ that attempted to steer so as to follow the nearest lane on the road. All cars in the simulation drove in a simulated 2-dimensional world, described above. Each time gaze was allocated to a new module, we recorded the module that received the gaze, as well as several behavioral measurements (e.g., distance to leader car, current speed, etc.) to verify that the agent was driving appropriately.

## Categorizing looks

The gaze selection process in our model is Markovian, meaning that each selected module is independent of the previously-selected modules; more formally, $p(\phi(t) \mid \phi(t-$ $n), \cdot)=p(\phi(t) \mid \cdot)$ for all $n>0$. Thus, it is possible that multiple consecutive module updates are directed at the same module, or $\phi(t)=\phi(t-n)$. Similar refixation behavior exists in human gaze during complex tasks; presumably observers use the visual information across multiple fixations for a continuous control signal for a single task. To make analysis simpler and more consistent between simulation and human results, we grouped multiple consecutive updates for a given module into a single "look." For instance, in the example shown in Figures 1 and 2, updates are provided first to module A, then B, then A twice, then B twice. In this example, each module receives two "looks," with the second look for each module being twice as long as the first.

## Comparison with human results

Sullivan et al. (2012) instructed subjects driving in a virtual environment to follow a leader car and maintain a certain speed, but the priority of which of the two tasks was most important was varied so that one was high and the other was low. Additionally, subjects drove in some conditions where noise was added to the speed of the car, with the intent of disrupting the maintenance of a constant speed. These manipulations resulted in four conditions where either following a leader or maintaining a constant speed was most important, and velocity noise was either present or absent. They found that task priority increased fixation behavior on task-related objects. Additionally, an interaction between priority and uncertainty was found, whereby uncertainty alone did not guarantee increased fixation behavior. Instead, only if a task related object had sufficiently high priority did the addition of uncertainty further increase fixation behavior. Look duration histograms for this experiment are replicated in the top row of Figure 3.

We ran a set of simulations with our model attempting to replicate this behavior using parameters set to mimic the orginal human driver conditions. We used a simple grid search to locate these parameters. Because all of the parameters taken together can present a scaling ambiguity (e.g., if all $\varepsilon^{(i)}$ and $\rho^{(i)}$ are multiplied by 2 , then the same qualitative behavior will result) we fixed $\rho^{(f)}=1$ and explored only settings for the other parameters.

Once we identified the parameter settings corresponding to the experimental conditions, we evaluated our model by running it in each of these conditions 10 times, with each simulation run for approximately 4000 steps. The sequence of module updates for each simulation run was stored and labeled as looks as described above, then normalized to form a probability distribution. These results were compared the distributions of look durations from the human data. The model was able to capture several important aspects of the human data, including a sensitivity to both noise and priority, but also a gating effect whereby noise in low-priority tasks had a smaller effect than noise in high-priority tasks. Our results, shown under the human data in Figure 3, are qualitatively similar to the human performance in a virtual driving environment.

In addition to our scheduling model, a baseline fixation scheduler was run in the simulation. This scheduler incorporated only the priority of each task in selecting modules for update, but uncertainty was not incorporated. The results from this baseline scheduler are shown in the bottom row of Figure 3. The probability distributions from our scheduler and the baseline compared against the human data via the Kullback-Leibler (KL) divergence. Over all the conditions, our model had an average KL divergence of 2.20, versus 4.43 for the baseline scheduler (lower numbers are better).

## Discussion

This paper described a modular, "soft" barrier approach for modeling eye movements in human drivers. The model includes explicit measurements of an agent's estimates of external world state, and uses a random walk to model the uncertainty in these estimates over time. Uncertainty, modulated by the priority of a task, is then used to arbitrate gaze allocations among competing modules. Our priority-plusuncertainty model provides a better fit of a set of human driving data than a priority-only baseline fixation scheduling model. We are currently working on comparing this model to predictions from a standard salience model (Itti \& Koch, 2001), a central bias model (Tatler \& Vincent, 2009) and the original scheduling model that inspired our work (Sprague \& Ballard, 2003). In addition, the softmax approach to selecting modules for update permits a clean inversion of the model; that is, given human eye fixation behavior, the model can be inverted to provide the most likely set of parameter settings to explain those data. We plan to develop this inversion more fully so as to replace the grid search described in this paper. Once the inversion process is in place, we can use this model to recover the task priorities and uncertainty levels that human drivers appear to be experiencing in these conditions.

## References

Arulampalam, M., Maskell, S., Gordon, N., Clapp, T., Sci, D., Organ, T., \& Adelaide, S. (2002). A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. IEEE Transactions on Signal Processing, 50(2), 174-188.


Figure 3: Distribution of look durations for human subjects (a-d; from Sullivan et al., 2012), model predictions (e-h), and baseline predictions (i-1). In all plots, look duration (in seconds) is shown along the abscissa, with the proportion of looks indicated on the ordinate. Looks to the speedometer are plotted with green squares; looks to the leader car are plotted with blue circles. ( $a, b$ ) In conditions where driving at a target speed was emphasized, human looks at the speedometer were approximately matched in duration to looks at the leader car. (c, d) In conditions where following a leader car was emphasized, human looks at the speedometer were brief. Noise added to the car's speed (b, d) affected human looks in the speed task more than looks in the following task. Similar results hold for our model (e-h), but not for a baseline model that incorporates task priority but ignores the effects of uncertainty (i-l).

Bruce, N., \& Tsotsos, J. (2009). Saliency, attention, and visual search: An information theoretic approach. Journal of Vision, 9(3).
Butko, N., \& Movellan, J. (2010). Detecting contingencies: An infomax approach. Neural Networks, 23(8), 973-984.
Carpenter, R., \& Williams, M. (1995). Neural computation of log likelihood in control of saccadic eye movements. Nature, 377(6544), 59-62.
Hayhoe, M., \& Ballard, D. (2005). Eye movements in natural behavior. Trends in cognitive sciences, 9(4), 188-194.
Itti, L., \& Koch, C. (2001). Computational modeling of visual attention. Nature reviews neuroscience, 2(3), 194-203.
Knudsen, E. (2007). Fundamental components of attention. Annu. Rev. Neurosci., 30, 57-78.
Land, M., \& Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? Vision research, 41(25), 3559-3565.
Nunez-Varela, J., Ravindran, B., \& Wyatt, J. (2012). Gaze allocation analysis for a visually guided manipulation task. Proc. From Animals to Animats, 44-53.
O'Dwyer, A. (2006). Handbook of PI and PID controller tuning rules (Vol. 2). World Scientific.
Senders, J. (1980). Visual scanning processes. Soest, Nether-
lands: Drukkerij Neo Print.
Sprague, N., \& Ballard, D. (2003). Eye movements for reward maximization. Advances in neural information processing systems, 16.
Sullivan, B., Johnson, L., Ballard, D., \& Hayhoe, M. (2011). A modular reinforcement learning model for human visuomotor behavior in a driving task. In Proc. AISB Symposium (pp. 33-40).
Sullivan, B., Johnson, L., Rothkopf, C., Ballard, D., \& Hayhoe, M. (2012). The role of uncertainty and reward on eye movements in a virtual driving task. Journal of Vision, 12(13).
Tatler, B., \& Vincent, B. (2009). The prominence of behavioural biases in eye guidance. Visual Cognition, 17(67), 1029-1054.

Wolfe, J., Butcher, S., Lee, C., \& Hyle, M. (2003). Changing your mind: on the contributions of top-down and bottomup guidance in visual search for feature singletons. Journal of Experimental Psychology: Human Perception and Performance, 29(2), 483.
Zhang, L., Tong, M., Marks, T., Shan, H., \& Cottrell, G. (2008). SUN: A bayesian framework for saliency using natural statistics. Journal of Vision, 8(7).

# Meaning Overrides Frequency in Idiomatic and Compositional Multiword Chunks 

Hajnal Jolsvai (jolsvai@cornell.edu)<br>Stewart M. McCauley (smm424@cornell.edu)<br>Morten H. Christiansen (christiansen@cornell.edu)<br>Department of Psychology, Cornell University, Ithaca, NY 14853 USA


#### Abstract

In line with usage-based accounts, recent psycholinguistic studies have confirmed that frequency of occurrence impacts processing latencies for multiword strings (e.g., Arnon \& Snider, 2010). However, these studies have not been concerned with the meaning of the multiword chunks in question, which is central to accounts of formulaic language rooted in cognitive linguistics (e.g., Wray, 2002). Here, we address this issue by comparing processing latencies for three types of multiword chunks: idiomatic expressions, meaningful compositional phrases, and less meaningful fragments. All three chunk types were matched for wholeand sub-string frequency. Our results show that frequency facilitates processing for all three chunk types, but to a lesser extent than their "meaningfulness" (as assessed in a separate norming study), indicating that the meanings of multiword expressions may have implications for models of language processing which extend beyond those of frequency of occurrence.


Keywords: idiomatic phrases; chunks; distributional statistics; Construction Grammar (CG); usage-based approach; cognitive linguistics

## Introduction

It has been recognized through a large number of corpus studies that everyday language involves a wide array of fixed co-occurring multiword sequences (such as let it go), or formulaic sequences (e.g., Sinclair, 1991). Usage-based theories suggest that strings of words can become fixed linguistic patterns through simple repetition of use (e.g., strong tea or powerful computer, as opposed to powerful tea or strong computer; Bybee, 2006; Bybee \& Hopper, 2001; Tomasello, 2003). Indeed, recent psycholinguistic work has shown that the processing of a multiword sequence is affected by the frequency of the sequence as a whole (e.g., Arnon \& Snider, 2010; Tremblay, Derwing, Libben, \& Westbury, 2011).

But how much insight does frequency alone give us into formulaic language processing? Is it possible that some multiword chunks become conventionalized because they convey a highly specific meaning, in much the same way as do individual words (e.g., let it be)? In addition to their frequency of use, the "meaningfulness" of multiword chunks may be another factor affecting stored linguistic units, in line with what would be expected from a cognitivegrammar perspective (e.g., Fillmore, Kay, \& O’Connor, 1988; Langacker, 1987).

Idioms are perhaps the most canonical form of formulaic language, involving a particular string of words which is linked to a specific meaning that does not follow compositionally from the meanings of the constituent words (e.g., spill the beans). Importantly, it has traditionally been assumed that idiomatic phrases are represented and retrieved from memory as single units and thus processed faster than compositional phrases (e.g., Swinney \& Cutler, 1979). Whereas idioms are often treated as special cases, Construction Grammar approaches suggest that both idiomatic and high-frequency compositional multiword expressions are stored as conventionalized form-meaning pairs (Goldberg, 2003).

Supporting the lexicalized idioms assumption (e.g., Swinney \& Cutler, 1979), an eye-tracking study by Underwood, Schmidt and Galpin (2004) found fewer and shorter fixations for the last words of idiomatic expressions (e.g., met the deadline by the skin of his teeth) as compared to the same words in non-idiomatic contexts (e.g., the dentist looked at his teeth). In a self-paced reading study, Conklin and Schmidt (2008) found that idiomatic expressions (e.g., hit the nail on the head) were read faster when compared to similar phrases (e.g., hit his head on the nail). Moreover, a recent eye-tracking study demonstrated a processing advantage for idiomatic expressions (e.g., at the end of the day) over similar phrases (e.g., at the end of the war) (Siyanova-Chanturia, Conklin, Schmitt, 2011). However, a shortcoming of these studies, which reveal faster processing for formulaic expressions, is that wholeand sub-string frequencies were not controlled for in a systematic way.

Nevertheless, recent experimental evidence has shown that language users are sensitive to the frequency of compositional multiword phrases. Typically, in these studies, pairs of high- and low-frequency phrases are compared, using corpus data to control for substring frequencies. Bannard and Matthews (2008) found that 3 - and 4 -years-olds repeat more frequent variants of compositional four-word phrases (such as a drink of milk) more easily than similar but less frequent four-word strings (such as a drink of tea). Similarly, high-frequency compositional multiword phrases are processed faster by adults than comparable sequences of lesser frequency (e.g., Arnon \& Snider, 2010; Tremblay \& Baayen, 2010).

In addition to findings of frequency effects for compositional multiword phrases, recent studies have investigated whether similar frequency effects can be
revealed for multiword sequences crossing syntactic boundaries. For instance, higher $n$-gram frequency chunks crossing syntactic boundaries (such as in the middle of the) are read faster compared to less frequent non-constituents (e.g., in the front of the) when embedded in a sentence (Tremblay, Derwing, Libben, \& Westbury, 2011). Similarly, Arnon and Cohen-Priva (in press) found no effect of constituency when comparing production latencies for highand low-frequency constituent and non-constituent pairs. The presence of frequency effects for not only phrases, but for sentence fragments as well, suggests that constituency may be a less important factor for theoretical accounts seeking to incorporate multiword chunks as linguistic units in their own right. Alternatively, it is possible that meaning may also need to be considered in order to reveal constituency effects for multiword sequences.

Although these findings make it clear that more frequent multiword phrases are processed faster, there is work to suggest that other factors may need to be taken in to consideration. For instance, Siyanova-Chanturia, Conklin and van Heuven (2011) found that readers are also sensitive to the fixedness of multiword chunks, above and beyond their frequency of occurrence. In this eye-tracking study, more frequent formulaic binominal phrases (such as bride and groom) were more easily read by participants as compared to their less preferred reversed forms, which nevertheless shared the same syntactic and semantic attributes (groom and bride). More importantly, even lower frequency formulaic chunks with fixed configurations (like sweet and sour), were read faster than higher frequency reversed phrases (such as west and east).

The above-mentioned studies are limited in that the "meaningfulness" of multiword chunks is not taken into account. Language users may also be sensitive to the relative meaningfulness of a multiword sequence as a whole, in line with predictions made by cognitive linguists (e.g., Langacker, 1987; Wray, 2002).

In the current study, we attempt to fill a gap left by previous psycholinguistic studies concerned with the processing of multiword sequences, which have tended to focus primarily on distributional properties, such as frequency of occurrence. We investigate the processing of multiword chunks that vary in the degree to which people find them meaningful as a whole unit, while controlling for frequency. Specifically, our goal is to determine whether the relative meaningfulness of multiword chunks may impact their processing over and above the well-known effects of frequency. Moreover, we aim to test the idea that highly meaningful chunks show processing advantages independently of their compositional status, in accordance with Construction Grammar approaches that treat both idiomatic and compositional sequences as pairings between form and meaning (e.g., Goldberg, 2003).

We address these issues by examining readers' processing of three different types of 3-word sequences; idiomatic expressions, highly meaningful compositional phrases. and less meaningful fragments which cross syntactic boundaries.

All three types of trigrams were selected from the same corpora and matched for phrase (trigram) and substring (bigram, unigram) frequencies. Both idiomatic expressions and compositional phrases were rated as being equally meaningful in an initial norming study. Furthermore, participants in this norming study rated sentence fragments as being significantly less meaningful than their frequencymatched idiomatic and meaningful compositional phrases. A different set of participants found the three types of tokens to be equally plausible as part of an English sentence in a separate norming study (thus, our items differed only in their meaningfulness, and not plausibility as possible sequences of words in English sentences). Instead of comparing pairs of high and low $n$-gram frequency sequences differing by only one word (don't have to worry vs. don't have to wait), as in previous studies, we investigated sequences which varied in the whole-string frequency ( $\log _{2}$ transformed) with which they appeared in a large corpus, along a continuum ranging from 1.0 to 10.4.

We predicted that idiomatic expressions and compositional phrases would be processed faster than less meaningful fragments because they vary in the extent to which people find them meaningful as units. We also predicted that when controlled for phrase and substring frequencies, idiomatic expressions would not be processed faster than compositional meaningful sequences. This prediction is at odds with the traditional view that contrasts faster access to stored idiomatic expressions with slower processing of computed compositional phrases (e.g., Swinney and Cutler, 1979). However, our prediction is in line with a Construction Grammar perspective (e.g., Goldberg, 2003).

## Methods

Three different types of trigrams were extracted from the American National Corpus (ANC; Reppen, Ide, \& Suderman, 2005) and the Fisher corpus (Cieri, Graff, Kimball, Miller, \& Walker, 2004, 2005). The two corpora were combined into a single corpus containing a total of 39 million words of American English. The Fisher corpus comprises spoken language (telephone conversations), while the ANC consists of spoken as well as written texts. From this combined corpus, we selected all 3-word idiomatic expressions using the following collections: McGraw-Hill's Essential American Idioms Dictionary (Spears, 2008); Handbook of Commonly Used American Idioms (Makai, Boatner, Gates 1991); and the IdiomQuest (http://www.idiomquest.com) and American Idioms (http://www.americanidioms.net) online idiom dictionaries. Eighty-two 3-word idiomatic expressions from these collections appeared in the combined corpus. For each idiomatic expression (e.g., over the hill) we selected frequency-matched compositional phrases (e.g., had a dream) and frequency-matched fragments (e.g., by the postal). $\log _{2}$ transformation was applied to all raw phrase and substring frequencies prior to the selection process (described below). Table 1 presents examples of the three
groups of items alongside their frequencies in the corpus. Each token belongs to a "triad" of whole- and sub-string frequency-matched items, with one from each of the three conditions (idiomatic expression, compositional phrase and fragment).

Table 1: Example triads of idiomatic expressions, compositional phrases, and fragments with their $\log _{2}$ transformed phrase frequencies

| Idiomatic <br> Expressions |  | Compositional <br> Phrases |  | Fragments |
| :--- | :--- | :--- | :--- | :--- | :--- |

Both the compositional phrases and the fragments were frequency-matched to a corresponding idiomatic expression such that trigram frequency, first bigram, second bigram, first unigram, second unigram, and third unigram frequencies were within $\pm 10 \%$ of the corresponding idiomatic phrase's frequencies, respectively. Table 2 shows the results of the individual ANOVA tests of the phrase and substring frequencies across idiomatic expressions, compositional phrases, and fragments.

Table 2: Individual ANOVA tests showing no differences between the averages of the six frequency measures across the three experimental conditions

|  | Df | F-score | p-value |
| :--- | :---: | ---: | ---: |
| Phrase | 2 | 0.0338 | 0.9668 |
| $1^{\text {st }}$ bigram | 2 | 0.092 | 0.9121 |
| $2^{\text {nd }}$ bigram | 2 | 0.0259 | 0.9745 |
| $1^{\text {st }}$ unigram | 2 | 0.8341 | 0.4368 |
| $2^{\text {nd }}$ unigram | 2 | 0.6037 | 0.5485 |
| $3^{\text {rd }}$ unigram | 2 | 0.05 | 0.9513 |

To ensure that the items differed only in the extent to which they would be judged as meaningful (i.e., that the fragments were less meaningful than the idiomatic expressions and compositional phrases, which in turn should not differ from one another) as opposed to plausibility of occurrence, we conducted two norming studies. In the first norming study, participants judged the plausibility of each trigram. In the second norming study, a different set of
participants rated the tokens according to how meaningful they were as units.

## Norming Study 1: Plausibility

The purpose of the first norming study was to collect judgments regarding the plausibility of the idiomatic expressions, compositional phrases, and fragments. This was necessary to ensure that stimuli from each condition were equally plausible as strings in American English, despite any differences in the extent to which items from each condition conveyed coherent meanings as units.

Participants Thirty-three native speakers of American English from the Cornell undergraduate population participated in the study for extra credit (mean age = 19.69; $\mathrm{SD}=1.74$ ). Data from 2 participants were omitted because their overall performance fell below $80 \%$ in a random memory recall task (see below).

Materials The materials consisted of 3-word sequences of 82 idiomatic phrases, 236 compositional phrases, and 218 fragments. The 3 -word sequences were presented to participants on a computer screen (one 3 -word sequence at a time). As a control, 90 impossible 3-word combinations were also included. The impossible tokens were created by scrambling matching compositional phrases and fragment tokens that were not introduced in the experimental material.

Procedure Participants' task was to rate each trigram according to how plausible the sequence of words was as part of an English sentence. Participants rated each token on a 1-7 scale by pressing a key. To ensure that participants read each sequence, a random memory recall test was included. In $10 \%$ of the trials for each condition (idiomatic expression, compositional phrase, and fragment), participants were asked to type an English sentence that included the 3-word sequence they had just seen.

Results Table 3 shows the mean scores for each condition in the plausibility norming study. As the trigrams were rated as equally plausible, we submitted them to a second norming study evaluating their meaningfulness.

Table 3: Mean ratings and standard deviations of plausibility (Norming Study 1) scores for idiomatic expressions, compositional phrases, and fragments.

|  | Mean | SD |
| :--- | ---: | ---: |
| Idiomatic Expressions | 6.84 | 0.32 |
| Compositional Phrases | 6.91 | 0.22 |
| Fragments | 6.87 | 0.22 |

## Norming Study 2: Meaningfulness

The purpose of the second norming study was to collect judgments regarding the meaningfulness of the idiomatic expressions, compositional phrases, and fragments as units.

Participants A different set of 33 native speakers of American English from the Cornell undergraduate population (mean age $=19.84$; $\mathrm{SD}=1.15$ ) participated in the study for extra credit. Data from two subjects were removed because their overall performance on the random memory recall test (see below) was less than $80 \%$.

Materials The materials consisted of the same 3-word sequences used in the first norming study (82 idiomatic phrases, 236 compositional phrases, and 218 fragments). The 3-word sequences as a whole were presented to participants one-by-one on a computer screen.

Procedure Participants' task was to rate each trigram according to how meaningful they found each sequence as a unit. Participants rated each token on a scale of 1-7 by pressing a key. To ensure that participants read each sequence, a memory recall test was included. In $10 \%$ of the trials for each condition (idiomatic expression, compositional phrase, and fragment), participants were asked to type an English sentence that included the last 3word sequence they had seen.

Results Table 4 shows the mean scores for each condition in the meaningfulness norming study.

Table 4: Mean ratings and standard deviations of meaningfulness (Norming Study 2) scores for idiomatic expressions, compositional phrases, and fragments.

|  | Mean | SD |
| :--- | ---: | ---: |
| Idiomatic Expressions | 5.90 | 1.02 |
| Compositional Phrases | 5.89 | 0.86 |
| Fragments | 1.98 | 0.48 |

In order to arrive at the final set of experimental items, we submitted the following to a selection algorithm (more details below): the arcsine-transformed proportion of subjects rating each item 6 or 7 ( 1 or 2 for fragments) in Norming Study 2, the arcsine transformed proportion of subjects rating each item as a 6 or 7 in Norming Study 1, and the whole- and sub-string frequencies of each item. The algorithm selected the set of 40 triads (each comprising an idiomatic expression, a compositional phrase, and a fragment) which differed least according to plausibility norming scores as well as phrase and substring frequencies across the three trigram types, while also differing maximally in Norming 2 scores between fragments and the other two conditions.

## Reaction Time Study

Our prediction was that the overall meaningfulness of a sequence would facilitate processing over and above mere frequency of use, independently whether the sequence was an idiomatic phrase. Thus, we predicted that processing latencies for idiomatic phrases and compositional meaningful phrases would not differ from one another. However, both idiomatic phrases and compositional phrases should show processing advantages over fragments.

Participants An additional 40 native speakers of American English from the Cornell undergraduate population were recruited, none of which participated in either of the two norming studies (mean age $=20.5 ; \mathrm{SD}=1.58$ ).

Materials A final set of frequency-matched tokens from the trigrams rated in the two norming studies was selected for the reaction time study (using a selection algorithm which sought to minimize differences along the frequency dimensions as well as the norming scores between the conditions). This set consisted of 40 triads, each comprising an idiomatic phrase, a compositional phrase and a fragment. The resulting set of 40 idiomatic expressions, 40 compositional phrases and 40 fragments did not differ significantly along the 6 frequency dimensions (trigram, first bigram, second bigram, first unigram, second unigram, third unigram) or according to the percentage of subjects rating items as 6 or 7 in the first plausibility norming study. Additionally, the items were constrained such that the idiomatic and compositional phrases did not differ in terms of their meaningfulness ratings from second norming study, whereas the fragments were chosen to have the lowest meaningfulness scores possible. All comparisons: p $>0.4$.

The $\log _{2}$ phrase frequencies of the final set of 40 triads introduced in the behavioral study ranged between 1 and 10.4. Besides the 40 experimental triads, 120 impossible sequences (such as hear I isn't) were used as fillers.

Procedure We based our reaction time study on Arnon and Snider's (2010) phrasal decision task (which in turn is based on the classic lexical decision task). Participants were presented with the three-word sequences (120 experimental and 120 impossible filler tokens) separately, in random order, on a computer screen, and asked to judge (by quickly hitting one of two keys) whether they formed possible word combinations in the context of English sentences.

Data Analysis Data points corresponding to reaction times of less than 200 ms were removed, along with extreme outliers (defined as those reaction times exceeding the upper quartile by more than three times the inter-quartile range), resulting in a $1.5 \%$ data loss. The data were then submitted to a linear mixed-effects (LME) analysis, with Item and Subject as random effects, and the scores from the second norming study (hereafter referred to as the Meaningfulness Scores), Trigram Type (using Idiomatic Expressions as the base case), Frequency (whole-string), substring frequencies
(including frequency predictors for First Bigram, Second Bigram, First Unigram, Second Unigram, and Third Unigram), Length in Characters, and the Trigram Type x Frequency interaction term as fixed effects. Because the Meaningfulness scores had a fixed range of 1 to 7 , they were converted to proportions ( $\mathrm{n} / 7$ ), which were then logittransformed prior to entry in the model (cf. Armitage \& Berry, 1984). As the reaction times were not normally distributed, they were log-transformed prior to the analysis.

The LME resulted in no significant interaction between Frequency and Trigram Type and was therefore simplified to involve Item and Subject as random effects, and Meaningfulness scores, Trigram Type (using Idiomatic Expressions as the base case), Length in Characters, Frequency (whole-string), and the substring frequency predictors as fixed effects. There did not appear to be substantial multicollinearity between the fixed effects: The condition number for the matrix of predictors (cf. Belsley, Kuh, \& Welsch, 1980) was only 6.6.

The final model was compared to a version without the fixed effects ( $\chi^{2}=140.63, \mathrm{p}<0.0001$ ) as well as a version of the model without the variables of interest (Meaningfulness Score, Frequency, Trigram Type; $\chi^{2}=111.07, \mathrm{p}<0.0001$ ), indicating that the full model captured more of the variance in both cases.

## Results

As predicted, participants showed sensitivity to the meaningfulness of the trigrams. Decision times were faster for more meaningful tokens, as revealed by a highly significant main effect of Meaningfulness Score ( $\beta=-0.037$, $\mathrm{p}<0.0001$ ). Also in line with our predictions, reaction times were affected by trigram type: It took longer for participants to decide whether fragments were possible strings in English ( $\beta=0.089, \mathrm{p}<0.05$ ). However, decision times for compositional meaningful phrases were no slower than for idiomatic expressions ( $\beta=0.003, \mathrm{p}>0.3$ ). Table 5 presents the mean RTs and standard deviations for each condition.

Table 5: Mean RTs and standard deviations for idiomatic expressions, compositional phrases, and fragments.

|  | Mean | SD |
| :--- | ---: | ---: |
| Idiomatic Expressions | 766.4 | 247.9 |
| Compositional Phrases | 789.4 | 260.3 |
| Fragments | 949.9 | 343.7 |

Frequency (whole-string) also reached significance ( $\beta=-0.014, \mathrm{p}<0.05$ ) indicating that subjects responded faster to trigrams with greater whole-string frequency. Importantly, the effect of Frequency was less than that of Meaningfulness Score.

Of the fixed effects included to control for substring frequency and character length, only Length in Characters ( $\beta=0.08, \mathrm{p}<0.05$ ) and Third Unigram $(\beta=0.014$, $\mathrm{p}<0.05$ ) reached significance.

## General Discussion

In this study, we compared the processing latencies for triads consisting of idiomatic expressions, compositional phrases and less meaningful fragments. In each triad, both the meaningful compositional phrases and the fragments were frequency-matched to the idiomatic phrases. The aim of the study was to investigate (i) whether the relative meaningfulness of multiword chunks affects processing latencies in addition to their frequency, and (ii) whether idiomatic and meaningful compositional phrases are processed similarly, once frequency is adequately controlled for.

Our results indicate that participants were sensitive to the meaningfulness of the chunks. The meaningfulness of a given trigram-as indicated by the second norming studysuccessfully predicted reaction times in the final experiment. Participants' decision times for more meaningful trigrams were faster than for less meaningful ones. Furthermore, our findings show that the whole-string frequency of the tokens predicted the processing latencies, but to a lesser extent than the meaningfulness of the different chunks. These results thus suggest that in addition to frequency of occurrence, as emphasized by usage-based theories, the relative meaningfulness of multiword chunks should also be considered in accounts of language processing.

Our findings also showed that processing latencies for idiomatic phrases did not differ from frequency-matched compositional meaningful phrases, while processing latencies for less meaningful fragments were significantly greater. This suggests that meaningful compositional sequences may be represented and processed similarly to idiomatic phrases, despite their compositional nature. This is at odds with traditional distinctions between stored idioms and compositional phrases (Swinney \& Cutler, 1979). However, these results are in line with Construction Grammar approaches, suggesting that there are no fundamental differences between the representation and processing of idiomatic constructions and compositional phrases; they are both instances of conventionalized formmeaning mappings (Goldberg, 2003).

Additionally, higher meaningfulness scores were associated with reduced processing latencies. These results, when viewed alongside the weaker frequency effect, provide a step forward for studies of formulaic language, suggesting that that the meaningfulness of multiword chunks may be as important to their processing as their distributional properties.

It is possible that the processing latencies for compositional phrases in the current study were affected by their status as constituents, whereas fragments crossed syntactic boundaries. There exists only one study that has investigated whether constituency affects multi-word sequence processing while controlling for frequency. In a recent production study, Arnon and Cohen-Priva (in press) found that constituency did not affect processing. In their study, similar frequency effects were found when
comparing high- and low-frequency variants of constituents (a lot of work vs. a lot of years) as well as non-constituents crossing syntactic boundaries (as far as $I$ vs. as far as you). Similar frequency effects for both phrases and fragments suggest that constituency may be a less important feature of multiword chunk processing. On the other hand, it is possible that the lack of a constituency effect in the Arnon and Cohen-Priva study stems from not taking chunk meanings into account. Further studies are needed to examine the exact nature of the relationship between constituency and chunk meaningfulness.

To conclude, our results provide new insights into the representation and processing of formulaic expressions; they suggest that multiword compositional phrases that people find highly meaningful are likely to be processed similarly to idiomatic phrases, as a linguistic unit in its own right. Our findings are thus relevant for usage-based approaches to language, indicating that meaning provides an additional dimension that such approaches must take into account, in line with a number of expectations derived from cognitive linguistics (e.g., Langacker, 1987).

## Acknowledgments

We would like to thank Jaclyn Jeffrey-Wilensky and Julia Krasnow for their assistance with the reaction time study. Thanks also go to Jaclyn Jeffrey-Wilensky and László Kálmán for their help during material selection. Finally, we are grateful to Haim Bar for his valuable comments on our statistical analyses.

## References

Armitage, P. \& Berry, G. (1994). Statistical methods in medical research (3rd ed.). Oxford: Blackwell.
Arnon, I. \& Cohen-Priva, U. (in press). More than words: The effect of multi-word frequency and constituency on phonetic duration. Language and Speech.
Arnon, I. \& Snider, N. (2010). More than words: Frequency effects for multi-word phrases. Journal of Memory and Language, 62, 67-82.
Bannard, C. \& Matthews, D. (2008). Stored word sequences in language learning. Psychological Science, 19, 241-248.
Belsley, D. A., Kuh, E., \& Welsch, R. E. (1980). Regression diagnostics: Identifying influential data and sources of collinearity. NY: Wiley.
Bybee, J. (2006). From usage to grammar: the mind's response to repetition. Language, 82, 711-733.
Bybee, J., \& Hopper, P. (2001). Frequency and the emergence of linguistic structure. Amsterdam: John Benjamins.
Cieri, C., Graff, D., Kimball, O., Miller, D., \& Walker, K. (2004). Fisher English training speech (part 1) transcripts. Philadelphia: Linguistic Data Consortium.
Cieri, C., Graff, D., Kimball, O., Miller, D., \& Walker, K. (2005). Fisher English training speech (part 2) transcripts. Philadelphia: Linguistic Data Consortium.
Conklin, K., \& Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic
language by native and non-native speakers? Applied Linguistics, 29, 72-89.
Fillmore, C., Kay, P., \& O’Connor, M. K. (1988). Regularity and idiomaticity in grammatical constructions: The case of let alone. Language, 64, 501-38.
Goldberg, A. E. (2003). Constructions: A new theoretical approach to language. Trends in Cognitive Sciences, 7, 219-224.
Langacker, R.W. (1987). Foundations of cognitive grammar, Vol. 1: Theoretical prerequisites. Palo Alto: Stanford University Press.
Makai, A., Boatner, M. T., Gates, J. E. (1991). Handbook of commonly used American idioms. NY: Barron's Educational Series Inc.
Reppen, R., Ide, N., \& Suderman, K. (2005). American national corpus (ANC): Second release. Philadelphia: Linguistic Data Consortium.
Spears, R..A. (2008). Essential American idioms dictionary. Columbus, OH: McGraw-Hill.
Siyanova-Chanturia, A., Conklin, K., Schmitt, N. (2011). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. Second Language Research, 27, 251-272.
Siyanova-Chanturia, A., Conklin, K., \& Van Heuven, W. J. B. (2011). Seeing a phrase "time and again" matters: The role of phrasal frequency in the processing of multiword sequences. Journal of Experimental Psychology: Learning, Memory, and Cognition. 37, 776-784.
Sinclair, J. (1991). Corpus, concordance, collocation. Oxford: Oxford University Press.
Swinney, D., \& Cutler, A. (1979). The access and processing of idiomatic expressions. Journal of Verbal Learning and Verbal Behaviour, 18, 523-34.
Tomasello, M. (2003). Constructing a language: A usagebased theory of language acquisition. Cambridge: Harvard University Press.
Tremblay, A. \& Baayen, H. (2010). Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall (pp. 151-167). In D. Wood (Ed.), Perspectives on formulaic language, NY: Continuum International Publishing.
Tremblay, A., Derwing, B., Libben, G., \& Westbury, C. (2011). Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. Language Learning, 61, 569-613.
Underwood, G., Schmitt, N., \& Galpin, A. (2004). The eyes have it: An eye-movement study into the processing of formulaic sequences. In N. Schmitt (Ed.), Formulaic sequences: Acquisition, processing, and use (pp. 153172). Amsterdam: John Benjamins.

Wray, A. (2002). Formulaic Language and the Lexicon. Cambridge, UK: Cambridge University Press.

# The effects of dual verbal and visual tasks on featural vs. relational category learning 

Wookyoung Jung (jung43@illinois.edu)<br>Department of Psychology, 603 E. Daniel Street<br>Champaign, IL 61820 USA

John E. Hummel (jehummel@illinois.edu)<br>Department of Psychology, 603 E. Daniel Street<br>Champaign, IL 61820 USA


#### Abstract

Many studies have examined the distinction between featureand relation-based categories (Gentner, 2005; Genter \& Kurtz, 2005; Jung \& Hummel, 2009; Tomlinson \& Love, 2011). Those findings suggest that featural and relationl categories have fundamentally different learning algorithms, where relational categories rely on explicit representations and thus require working memory and attention, as opposed to featural categories which may be learned more implicitly. In this study, we investigated further the distinction between feature-and relation-based category learning using a dual task methodology. Our results revealed an interaction: featural category learning was more impaired by a visuospatial dual task than by a verbal dual task, whereas relational category learning was more impaired by the verbal dual task. Our results suggest that in contrast to featural category learning, which may involve mainly non-verbal mechanisms, relational category learning appears to place greater demands on more explicit and attention-demanding verbal or verbally-related learning mechanisms.


Key words: featural category learning; relational category learning; dual task; verbal dual task; visuospatial dual task; category learning algorithms

The ability to categorize plays a central role in human mental life. We use categories to makes sense of the world. They allow us to generalize knowledge form one situation to another, to decide which objects in the world are fundamentally the same, and to infer the unseen properties of novel category members. Research on categorization has mainly focused on feature-based categories-that is, categories defined by their exemplars' features, as when "bugs" in one category tend to have a particular kind of head, body and tail and "bugs" in the opposite category tend to have a different kind of head, body and tail (e.g., Taylor and Ross, 2009)—and comparatively little on relationbased categories-i.e., categories by the relations between exemplars' parts, or by relations between category exemplars and other objects in the world (for reviews, see Gentner, 2005; Goldwater, Markman, \& Stilwell, 2011; Jung \& Hummel, 2009; Kittur, Hummel \& Holyoak, 2004).

The distinction between featural and relational categories matters because features and relations are very different things-so different that we can have little or no
confidence that anything learned about category learning using feature-based categories will generalize at all to the case of relational categories. For example, the kinds of learning algorithms that work well with feature-based categories (i.e., various kinds of statistical learning) are completely incapable of learning relational categories (Doumas, Hummel \& Sandhofer, 2008; Hummel \& Holyoak, 2003; Jung \& Hummel, 2009; Kittur et al., 2004, 2006).

One of the clearest examples of this difference comes in the form of peoples' ability to learn probabilistic (aka family resemblance) category structures. It has been known since the 1970s that people have no difficulty learning categories with probabilistic structures, in which any given feature is likely to belong to a given category (e.g., "bugs" in category A are likely to have one kind of head whereas "bugs" in category B are likely to have another), but no feature is deterministically associated with any given category (e.g., sometimes, bugs from category B will have heads typical of bugs from category A and vice-versa; see Murphy, 2002, for a review). However, as noted by Kittur et al. (2004), such prototype effects have always been observed with feature-based categories. With categories defined by the relations between their exemplars' features, such prototype effects have proven difficult or impossible to observe (Jung \& Hummel, 2009, 2011; Kittur et al., 2004, 2006).

These differences between peoples' ability to learn featural and relational categories are consistent with the claim that fundamentally different learning algorithms may be at work in the two cases. For example, whereas associative learning may work in the case of featural categories, relational category learning may require a more sophisticated algorithm based, for example, on structured intersection discovery, in which learners compare examples to one another, retaining what the examples have in common and discarding or discounting the details on which they differ (Gick \& Holyoak, 1983; Hummel \& Holyoak, 2003; Jung \& Hummel, 2009, 2011; Kittur et al, 2004, 2006).

A fundamental assumption underlying this intersection discovery hypothesis is that people's mental representations of relational categories are explicitly relational (see Hummel \& Holyoak, 2003; Jung \& Hummel, 2009, 2011). That is, we assume that people notice and explicitly represent the relations between objects (and object parts) and use these relations as the basis for making their categorization responses. This assumption also leads to another critical contrast with feature-based approaches to mental representation and categorization. In contrast to featurebased representations, which come to us effortlessly, relational representations require attention and working memory (see, e.g., Hummel \& Holyoak, 1997, 2003; Logan, 1994; Maybery et al., 1986).

In this study, we examined what kinds of working memory might be involved in feature- or relation-based category learning. In particular, our interest was in how featural and relational category learning tasks respond to verbal and visuospatial dual tasks. If featural and relational category learning are based on different learning algorithms, then they might be differentially sensitive to different kinds of dual tasks.

Other researchers have also argued for multiple systems of category learning (Ashby et al., 1998). Miles and Minda (2011) showed that verbal dual tasks, which impose an executive functioning load, impaired rule-defined category learning, whereas a visual dual task impaired non-rule-defined learning regardless of executive functioning demand. Their findings provided evidence that verbal working memory and executive functioning are engaged in the rule-defined system, and visual processing is more engaged in the non-rule-defined system.

Our experiment will test the prediction that relational category learning will be more subject to verbal dual-task interference than feature-based category learning. By contrast, feature-based learning will be more subject to visuospatial dual-task interference than relational learning.

We used deterministic category structures; i.e., there was always one relation or feature that was deterministically predictive of category membership. The reason for using deterministic categories is that the categories must be learnable, even in the relational case, so that we can observe the effects of our manipulation on trials to criterion (i.e., how long it takes subjects to learn the categories).

We orthogonally crossed relational vs. feature-based categories with verbal dual task vs. visual dual task vs. no dual task. In the verbal dual task conditions, subjects had to perform a task known to interfere with relational processing (memorizing digits) while they simultaneously performed the category learning task. In the visual dual task condition, subjects had to memorize the locations of filled squares in 3 X 3 grids while simultaneously learning the categorization. In the no dual task condition, subjects simply performed the category learning task by itself.

## Method

Participants. A total of 75 subjects participated in the study for course credit. Each participant was randomly assigned to one of the six conditions.

Materials. Each exemplar consisted of a grey ellipse and a grey rectangle. Each exemplar had both relational properties (e.g., ellipse bigger than rectangle) and featural properties (e.g., ellipse of size 4). Each subject was tasked with deciding whether the objects they saw belonged to one of two featural or one of two relational categories.

Each exemplar was defined by three category-relevant properties: size (absolute in the featural condition or relative in the relational condition), darkness (absolute or relative) and orientation (absolute or relative). In the featural condition, the orientation of the ellipse was deterministically associated with category membership (i.e., horizontal orientation for category A, vertical for category L), whereas in the relational category condition, the relative orientation of the ellipse and rectangle (i.e., either same or different) was deterministically associated with category membership (with same for category A and different for category L). The other properties were probabilistically associated with category membership.


Figure 1. Three relevant properties in the featural condition: category A (above) and L (below)

For the featural category condition, the prototypes of the categories were defined as $[1,1,1]$ for category A and [ $0,0,0$ ] for $L$, where $[1,1,1]$ represents an rectangle size 3 [out of 9] for category A, 7 for category L, the color 3 [out of 9] for category A, 7 for category L, and horizontal orientation for category A, vertical for category L (Figure 1). Similarly, for the relational category condition, the prototypes were defined as $[1,1,1]$ for category $A$ and
[ $0,0,0$ ] for L , where [1,1,1] represents an ellipse larger, darker, and same orientation and $[0,0,0]$ represents a rectangle larger, darker, and different orientation (Figure 2). Exemplars of each category were made by switching the value of one dimension in the prototype (e.g., relational category A exemplar [1,0,1] would have the ellipse larger, lighter, and same orientation as the rectangle). Four copies of each exemplar type were presented on each block, two paired with a "Yes" responses on the dual task and two with a "No" responses, resulting in 32 trials per category per block.


Figure 2. Three relevant properties in the relational condition: category A (above) and L (below)

Design. The experiment used a 3 (dual task: none vs. verbal vs. visuospatial) X 2 (relevant property: features vs. relations) between-subjects design.

Procedure. Participants were assigned randomly to one of the six groups. For the dual task conditions, on each trial, a memory task was provided first and followed by a categorization task and by a recall task. For the control conditions, only the categorization task was provided (Figure 3). Both categorization and dual task responses were followed by accuracy feedback.

Participants in the verbal dual-task condition were first given a verbal working memory task, in which 5 random digits were displayed for two seconds with spaces between them (so that they appeared to be individual numbers rather than digits of a single number). Participants were asked to memorize the digits while they performed the categorization task. In the categorization task, an exemplar consisting of a rectangle and an ellipse was shown. Participants were instructed to press the A key if the stimulus belonged to category A and the L key if it belonged to L. Each exemplar
remained on the screen until the participant responded. Responses were followed by accuracy feedback. Participants then saw one random digit and were asked to decide whether it was in the set they saw previously.


Figure 3. Experimental design by each condition
In the visuospatial dual-task condition, a 3 by 3 grid was displayed in the middle of a screen for two seconds with two randomly-chosen cells filled. Participants were asked to memorize the locations of the filled cells until they completed the categorization task. In the recall task, one filled cell was displayed in the grid and participants were asked whether the cell had been filled in the original display. The experiment consisted of 30 blocks ( 960 trials) and continued until the participant responded correctly on at least twenty nine of thirty two trials ( $90.6 \%$ correct) for two consecutive blocks or until all 30 blocks had transpired, whichever came first.

## Results

Dual task accuracy. We discarded the data from participants whose accuracy was below $70 \%$ correct on the dual task ( 2 subjects in the verbal/featural condition). Mean accuracy on the verbal dual task was $M=.94(S D=.03)$ for the featural category learning condition, and $M=0.91$ ( $S D=$ 0.06 ) for the relational learning condition. Mean accuracy on the visual dual task was $M=0.91(S D=0.06)$ for the featural condition, and $M=0.89$ ( $S D=0.04$ ) for the relational condition. There was no reliable difference between the verbal and visuospatial tasks $[t(51)=1.61, p$ $=$.114], suggesting that these tasks occupied cognitive resources to roughly the same extent.

## Category learning task accuracy: trials to criterion.

 Since our primary interest is the rate at which participants learn the categories as a function of the dual tasks, we reportour data first in terms of trials to criterion. These analyses are conservative in the sense that participants who never learned to criterion were treated as though they reached criterion on the last block. Figure 4 shows the mean trials to criterion by condition. A 3 (dual task) $\times 2$ (category learning task) between-subjects ANOVA revealed a main effect of dual task $[F(2,69)=5.058, M S E=579014.858, p<0.01]$. Since our main interest is in how different dual tasks affect the different kinds of category learning, one-way ANOVAs were conducted for the featural and relational learning conditions. The results revealed reliable differences between dual tasks in the featural category learning condition $[F(2,35)=4.981, M S E=617725.846, p<0.05]$. Planned comparisons in the featural category learning showed that there was a reliable difference between the verbal ( $M=386$, $S D=387)$ and visuospatial dual task $(M=697, S D=411)$ [ $t(35)=-2.288, p<0.05]$. There was also a reliable difference between the visuospatial and the control condition ( $M=262, S D=191$ ) $[t(35)=3.014, p<0.01]$. The difference between the verbal and the control condition was not reliable [ $t(35)=0.877, p<0.386$ ]. The ANOVA results from the relational condition revealed reliable differences between the dual tasks $[F(2,34)=7.641, M S E=$ 799483.887, $p<0.01]$. Planned comparisons revealed that there was a reliable difference between the verbal ( $M=739$, $S D=352$ ) and visuospatial dual task $(M=330, S D=362)$ $[t(34)=3.221, p<0.01]$. There was also a reliable difference between the verbal and control conditions ( $M=$ 276, $S D=222$ ) $[t(34)=3.014, p<0.01]$. The difference between the visuospatial and control conditions was not reliable [ $t(34)=0.404, p<0.689$ ]. No other main effects were statistically reliable. Most interestingly, there was a reliable interaction between dual task and category learning, indicating that relational category learning was disrupted more by the verbal dual task, whereas featural category learning was disrupted more by the visuospatial dual task $[F(2,69)=2.475, M S E=855659.946, p<0.01]$.

Response times. Since the category learning accuracy results yielded a reliable interaction between the dual and category learning tasks, we also analyzed these tasks in terms of participants' mean response times on individual trials in order to gain insight about the strategies participants in each condition may have adopted. A 3 (dual task) $\times 2$ (category learning task) between-subjects ANOVA revealed a main effect of dual task $[F(2,69)=3.202, M S E=0.961$, $p<0.05]$. One-way ANOVAs were also conducted in each category learning condition. The main effect of dual task was not reliable $[F(2,35)=2.137, M S E=0.612, p=0.133]$ in the featual learning condition. But since the accuracy data showed that participants in visuospatial feature-learning required many more trials than to reach to the criterion than participants in verbal featural learning, we expected a reliable difference between two conditions in a planned comparison analysis. Our prediction was confirmed. There was a reliable difference between the verbal ( $M=0.99$, $S D$ $=0.31$ ) and visuospatial dual task ( $M=1.41, S D=0.78$ )
$[t(35)=-2.037, p<0.05]$, indicating that response times in visuospatial feature-learning condition were longer than those in verbal feature-learning. No other differences were statistically reliable. There were no reliable differences in the relational learning condition. Also, ANOVA showed a reliable main effect of category learning $[F(1,69)=3.883$, $M S E=1.166, p=0.053$ ], indicating that feature learning ( $M=1.17, S D=0.55$ ) was marginally faster than relational learning ( $M=1.42, S D=0.56$ ) (Figure 5).


Figure 4. Accuracy by category learning condition


Figure 5. Response times by dual condition

## Discussion

To the extent that relational concepts are qualitatively similar to feature-based concepts, our understanding of concepts can be expected to generalize from the (extensively investigated) case of feature-based categories to the (largely neglected) case of relational categories. However, there is reason to believe they are not, casting doubt on our ability to generalize our conclusions from studies using feature-based categories to the case of relational concepts.

Most notably, people have no difficulty learning feature-based categories in which no single feature remains invariant across all members of a category (see Murphy, 2002). By contrast, Kittur and colleagues showed that relational categories are extremely difficult to learn when there is no such relational invariant (i.e., property that holds over all members of a category; Kittur et al., 2004, 2006). Jung and Hummel (2009, 2011) provided additional evidence that relational learning requires some kind of invariant in order to succeed. These findings suggest that featural and relational learning rely not only on qualitatively different forms of mental representation (namely, features vs. relations; see, e.g., Hummel, 2010; Hummel \& Holyoak, 1997, for a discussion of the difference) but also that they rely on qualitatively different kinds of learning algorithms (e.g., associative learning in the featural case and something more akin to structured intersection discovery in the relational case; Jung \& Hummel, 2009, 2011).

The current experiment provides additional evidence for this sharp distinction between featural and relational category learning. In the current experiment, featural learning was impeded by a visual dual task (i.e., one that might be expected to interfere with visual feature processing as required for featural learning) but not by a verbal dual task. Relational category learning, in sharp contrast, was interfered with by a verbal dual task (which has been shown to interfere with relational processing; Waltz et al., 2000), but not by a visual dual task. This double dissociation between visual vs. verbal dual task interference on the one hand and featural vs. relational category learning on the other adds to the growing evidence that these two kinds of category learning rely on qualitatively different and dissociable learning systems.

## Acknowledgements

This research was supported by a grant from the University of Illinois Research Board.

## References

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., \& Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. Psychological Review, 105, 442-481.

Doumas. L. A. A., Hummel, J. E., \& Sandhofer, C. M.(2008).A theory of the discovery and predication of relational concepts. Psychological Review, 115, 1-43.
Gentner, D. (2005). The development of relational category knowledge. In L. Gershkoff-Stowe \& D. H. Rakison, (Eds.), Building object categories in developmental time. (pp. 245-275). Hillsdale, NJ: Erlbaum
Gentner, D., \& Kurtz, K. (2005). Relational categories. In W. K. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman \& P. W. Wolff (Eds.), Categorization inside and outside the lab. (pp. 151-175). Washington, DC: APA.
Gick, M. L., \& Holyoak, K. J. (1983). Schema induction and analogical transfer. Cognitive Psychology, 15, 1-38.
Goldwater, M. B., Markman, A. B., \& Stilwell, C. H. (2011). The empirical case for role-governed categories. Cognition, 118, 359-376.
Hummel, J. E. (2010). Symbolic vs. associative learning. Cognitive Science, 34, 958-965.
Hummel, J. E., \& Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. Psychological Review, 104, 427-466.
Hummel, J. E., \& Holyoak, K. J. (2003). A symbolicconnectionist theory of relational inference and generalization. Psychological Review, 110, 220-264.
Kittur, A., Holyoak, K. J., \& Hummel, J. E. (2006). Using ideal observers in higher-order human category learning. Proceedings of the Twenty Eight Annual Conference of the Cognitive Science Society (pp. 435-440). Hillsdale, NJ: Lawrence Erlbaum Associates.
Kittur, A., Hummel, J. E., \& Holyoak, K. J. (2004). Feature- vs. relation-defined categories: Probab(alistic)ly Not the Same. Proceedings of the Twenty Six Annual Conference of the Cognitive Science Society (pp. 696701).Hillsdale, NJ: Lawrence Erlbaum Associates.

Jung, W., \& Hummel, J. E. (2009) Probabilistic relational categories are learnable as long as you don't know you're learning probabilistic relational categories. Proceedings of The 31st Annual Conference of the Cognitive Science Society, Society (pp. 1042-1047). Hillsdale, NJ: Lawrence Erlbaum Associates.
Jung, W., \& Hummel, J. E. (2011). Progressive alignment facilitates learning of deterministic but not probabilistic relational categories. Proceedings of the 33rd Annual Conference of the Cognitive Science Society.
Logan, G. D. (1992). Shapes of reaction time distributions and shapes of learning curves: A test of the instance theory of automaticity. Journal of Experimental Psychology: Learning, Memory and Cognition, 18, 883914.

Maybery, M. T., Bain, J. D., \& Halford, G. S. (1986). Information processing demands of transitive inference. Journal of Experimental Psychology: Learning, Memory, and Cognition, 12, 600-613.
Miles, S. J. \& Minda, J. P. (2011). The effects of concurrent verbal and visual tasks on category learning. Journal of Experimental Psychology: Learning Memory \& Cognition, 37, 588-607.

Murphy, G. L. (2002). The big book of concepts, Cambridge, MA: MIT Press.
Taylor, E. G. \& Ross, B. H. (2009). Classifying partial exemplars: Seeing less and learning more. Journal of Experimental Psychology: Learning, Memory, \& Cognition, 35(5), 1374-1380.
Tomlinson, M.T., \& Love, B.C. (2010). When learning to classify by relations is easier than by features. Thinking \& Reasoning,16, 372-401.
Waltz, J. A., Lau, A., Grewal, S. K., \& Holyoak, K. J. (2000). The role of working memory in analogical mapping. Memory \& Cognition, 28, 1205-1212.

# The Effects of Repeated Sequential Context on Recognition Memory 

George Kachergis, Gregory E. Cox, and Richard M. Shiffrin \{gkacherg, grcox, shiffrin\}@indiana.edu<br>Department of Psychological \& Brain Science / Cognitive Science Program<br>Bloomington, IN 47405 USA


#### Abstract

Many people have had the experience of knowing what song will play next on an album (even one heard only a few times). Conversely, many people fail to recognize an acquaintance encountered in an unfamiliar context. Associations can likely form simply because items appear nearby in time, and not only due to semantic similarity. Using surprise recognition testing, we examine the automatic storage of associations between successively encountered words on a list of incidentally studied words. Many modern memory models assume storage of such associations, but with little evidence as yet (e.g., Cox \& Shiffrin, 2012; REM-II Mueller \& Shiffrin, 2006). We find evidence for sequential associations, which are further improved by shared semantics or study context. We also find improved accuracy and response time for old words preceded by old words, and for new words preceded by new words - regardless of the previous response.


Keywords: recognition; episodic memory; temporal context; sequential association; priming

## Introduction

We have all had the experience of knowing what song will play next on an album that we have listened to several times, even without having looked at the list of songs. Conversely, we have also had the experience of seeing an acquaintance in a new context and not immediately recognizing them. Without realizing it, we often form associations between co-occurring events in a context, and memory is strengthened if the context reoccurs. In general, associates stored together help us remember if they are present at test. When any event is experienced there are a host of potential associations that make up the contextwhen and where and with what other things did the event happen to me? The memory and its context, and the retrieval of both, are termed episodic memory. The current study investigates the formation and retrieval of one type of context: the other words in a presented sequence of words. A critical factor in this research is the existence of source confusion. For example, given a recognition test of a word, test word familiarity is partly governed by the familiarity of the previous test word. At both study and test we confuse features of nearby events. For example, Jacoby and Whitehouse (1989) found in a recognition experiment when unstudied words were preceded by a subliminal prime ( 50 or 35 ms ) of the same word, people were more likely to incorrectly endorse the word as a studied one (i.e., false alarm). When the prime was a different unstudied word than the target, false alarms decreased for the 50 ms primes, but oddly not for 35 ms primes. In contrast, for liminal primes ( 200 or 600 ms ), a studied prime decreases hits for a
matching target, and an unstudied prime reduces false alarms to a matching target.

The ROUSE-Responding Optimally with Unknown Sources of Evidence - model of short-term priming (Huber, Shiffrin, Lyle, and Ruys, 2001) incorporates feature leakage from the prime to the target, leading to biased responses. However, ROUSE's decision rule has a discounting mechanism that attempts to correct for leakage: underdiscounting explains why primed words are chosen after passive priming, and over-discounting accounts for foil preference after active priming. Although ROUSE was applied mainly to identity and orthographically similar primes, semantic priming and leakage of semantic features also occur, and all these features should (with some probability) be incorporated in the storage of the next few events, and in the test probe of the next few tests. This might suggest episodic-recognition context effects would match those in perceptual recognition, but Malmberg and Annis (2012) investigated sequential dependencies in recognition and found patterns that did not seem to match those found in perceptual experiments. We will investigate this issue in some detail in this research.

We examine the storage of associations between adjacent words in a studied list, and how memory for a studied word is affected at test by the presence of its study-list neighbor. When people expect a memory test, they will form explicit associations between nearby items using a variety of coding schemes. Since our main interest is in automatic and nonstrategic storage and retrieval, we limit explicit associative strategies by using an incidental study task: participants make alternating pleasantness/animacy judgments at study.

Evidence for temporal associations have been found in recall following explicit attempts to remember. Participants are serially shown individual, unrelated words (e.g. 'crow', 'bottle', 'house', ...) and then asked to recall words from the list in any order. Given that a participant recalls a word (e.g. 'bottle'), the next word they recall is very likely to be the next word that was presented (e.g. 'house'; Kahana, 1996). In recognition tests, participants are shown words one at a time, some from the studied list, and some new, and asked to indicate those studied. A positive recognition response is thought to occur when the test word seems sufficiently familiar, via a fast and automatic parallel search of memory, or when its study event is recalled explicitly, typically via a slow and strategic process (Malmberg, Holden \& Shiffrin 2004). Models such as REM and TCM explicitly have a role for word context. We seek to understand such effects when study is incidental.

## Experiment 1

This study explores the automatic formation and retrieval of associations in recognition memory between temporally proximal events. Specifically, we varied the relation of two successive words at study for incidental judgments, and explored the effect when words related to these were tested successively, each for separate judgments of presence during study. For example, if "banana" is followed by "chair" at study, is "chair" recognized better or differently at test when preceded by a test of "banana"? The words in this example are semantically unrelated, but some of the adjacent words were made to be semantic associates.

The conditions we used included identical repeats, i.e. the same successive words at study and test, the case probably most likely to produce recognition benefits. In another condition the context word itself does not repeat, but its meaning does: The forward migration of matching semantic features at both study and test could produce improved recognition. In addition, meaning could be altered by the meaning of a recent word. For example, bank might be encoded as an earthen side if preceded by river, but encoded as a monetary institution if preceded by money. Table 1 shows examples of each condition, as well as the possible features that the preceding word (cue) may contribute to the target word at test: Familiarity (F) if the cue was a studied word; Semantics (S) if the cue is semantically related to the target; and Context (C), if the cue was also the target's study neighbor.

| Cue Type | Study | Test | F | S | C |
| :--- | :---: | :---: | :---: | :---: | :---: |
| Same, Related | cash bank | cash bank | 1 | 1 | 1 |
| Same Sense | cash bank | robber bank | 1 | 1 | 0 |
| Different | cash bank | river bank | 1 | $1^{*}$ | 0 |
| Sense |  |  |  |  |  |
| Different, | cash bank | sloth bank | 1 | 0 | 0 |
| Unrelated |  |  |  |  |  |
| Same, | sloth bank | sloth bank | 1 | 0 | 1 |
| Unrelated | sloth bank | glass bank | 1 | 0 | 0 |
| Unrelated | $\ldots$ bank | lamp bank | 0 | 0 | 0 |

Table 1: Features (Familiarity, Semantics, Context) that the cue may contribute to the target at test in each condition. *=related in the lexicon.

## Subjects

Participants were 57 undergraduates at Indiana University who received course credit for participating.

## Stimuli \& Procedure

We selected 40 common polysemous words (e.g., diamond) and their two strongest forward associates for each meaning (e.g., ruby/emerald and spade/ace) from the free association norms (Nelson, McEvoy, and Schreiber, 1998). For each participant, the 40 polysemes are assigned randomly to one of five conditions. In the Same, Related (SR) condition, the strongest associate of the dominant
meaning is presented just prior to the polyseme at both study and test (e.g. ruby-diamond). In the Same, Unrelated (SU) condition, an unrelated word is presented just prior to the polyseme at both study and test (e.g. lawn-diamond). The remaining conditions all have the strongest associate of the dominant meaning immediately prior to the polyseme at study, and a word that was studied elsewhere presented prior to the polyseme at test. In the Same Sense (SS) condition, a different associate from the same meaning was presented prior to the polyseme at test (e.g. emeralddiamond). In the Different Sense (DS) condition, the strongest associate from the other meaning was presented prior to the polyseme at test (e.g. spade-diamond). Finally, in the Different, Unrelated (DU) condition, an unrelated word was presented prior to the polyseme at test (fruitdiamond), for comparison to SR.

These 40 pairs of words are shuffled among 80 common filler words to compose a study list of 160 words. At study, each word was rated for either animacy or pleasantness, in an alternating fashion, in order to induce belief that this was the primary task and to reduce explicit encoding of successive words in identical ways. Each word was presented for 900 ms , followed by $2,000 \mathrm{~ms}$ of prompting for a response (which was not recorded), followed by 800 ms of blank screen before the next word was presented. After the study list was completed, participants were instructed that they would now perform a recognition test for the words they had just studied. The 160 studied words were randomly shuffled among 160 new words for surprise yes/no recognition testing. In order to reduce the use of strategic and explicit recollection we required participants to respond to the old-new test task within 700 ms . Slow responses elicited a "Too slow!" feedback message. Feedback on correctness was given on each test trial in Experiment 1.

## Results \& Discussion

Of the 60 subjects, 12 were removed for having a mean accuracy not significantly above chance (.522). Of the remaining responses, $2.8 \%$ were removed for being faster than 150 ms . The remaining 13,397 responses were analyzed using mixed-effects logistic regression, which is more appropriate than ANOVAs for analyzing accuracy (Jaeger, 2008). As regressors, we used the features that the cue may contribute to the target (see Table 1): Familiarity, Semantics, and Context. The logistic regression (see Table 2) shows that each of the three factors increase the odds of recognizing the target, with Semantics being the strongest cue ( $O R=2.16$ ), followed by Context $(O R=1.60)$, and finally Familiarity (OR=1.09).

| Factor | Coefficient | $Z$ | odds | p-value |
| :--- | :---: | :---: | :---: | :---: |
| (Intercept) | 0.40 | 9.67 | 1.49 | $<.001$ |
| Familiarity | 0.09 | 2.52 | 1.09 | $=.01$ |
| Context | 0.47 | 4.95 | 1.60 | $<.001$ |
| Semantics | 0.77 | 9.57 | 2.16 | $<.001$ |

Table 2: Logistic regression coefficients for Experiment 1.
Shown by condition in Figure 1, participants were most likely to respond old to old items in the SR condition, followed by the SS and DS conditions, then the SU condition, and finally the DU condition. The SR and SU findings imply that automatically encoded temporal context affects recognition, although we cannot say how much of the effect is due to a bias shift vs. a performance shift (because the design did not have equivalent conditions of cuing preceding new trials). We note in particular that the presence of a semantic relationship between the polyseme and the previous word at study (DU) or at study and test (SS, DS) increases the probability of giving an old response. From these results, it is clear that automatic associations are formed between both related and unrelated temporally proximal items. We also infer that familiarity accruing to the preceding test item tends to make the next test word seem familiar.


Figure 1: "Hits', p(old|old), for polysemous conditions in Exp. 1 (with feedback).

## Experiment 2

In Experiment 1, in contrast to traditional recognition memory experiments, we provided corrective feedback after each response at test. It may be that participants used the feedback signal from the previous trial to classify their feeling of familiarity and strategically used it on the next trial in any of several ways. Thus, in Experiment 2 we did not provide accuracy feedback at test.

## Subjects

Participants were 57 undergraduates at Indiana University who received course credit for participating.

## Stimuli \& Procedure

The same stimuli and procedure were used in Experiment 2, except at test there was no accuracy feedback given.

## Results

Of the 57 subjects, 4 were removed for having a mean accuracy not significantly above chance (.522). Of the
remaining responses, $1.8 \%$ were removed for being faster than 150 ms . The remaining 15,275 responses were analyzed using multilevel logistic regression. As in Experiment 1, we found positive effects of Semantics ( $\mathrm{OR}=1.84$ ), Context ( $\mathrm{OR}=1.51$ ), and Familiarity ( $\mathrm{OR}=1.25$; see Table 4). Thus, we have evidence for all of these three cues influencing the proximal trial, with and without feedback, when responses are limited to within 700 ms . In both experiments, semantics had the strongest effect, followed by context, and then familiarity.

| Factor | Coefficient | $Z$ | odds | p-value |
| :--- | :---: | :---: | :---: | :---: |
| (Intercept) | 0.40 | 7.03 | 1.49 | $<.001$ |
| Familiarity | 0.22 | 5.77 | 1.25 | $<.001$ |
| Context | 0.41 | 4.00 | 1.51 | $<.001$ |
| Semantics | 0.61 | 7.17 | 1.84 | $<.001$ |

Table 4: Logistic regression coefficients for Experiment 2.
Figure 2 shows the probability of a "hit" (old to an old test item) by condition for the polysemous manipulations in Exp. 2, which look much like those in Exp. 1. The only qualitative difference is that Same Sense was higher than Different Sense in Exp. 2, whereas a trend in the opposite direction was found in Exp. 1. Even given this difference, the experiments-with and without feedback-had much the same results.


Figure 2: "Hits" for polysemous conditions in Exp. 2 (without feedback).

The previous results, from both studies, were those for the carefully balanced conditions. However there also many filler items that were studied and tested, and many new items tested. Analyses of these items and their sequential effects are taken up in the next section.

## Further Sequential Analysis

Analyses of the filler words and new words showed a more general sequential context effect. We analyzed all of the data in both experiments using mixed-effects logistic
regression, trying to predict correct responses (old for old, new for new) as a function of the current item's oldness, the previous item's oldness, the correctness of the response to the previous item, and feedback (i.e., Experiment).

| Factor | Coeff | $Z$ | p-value |
| :--- | :---: | :---: | :---: |
| (Intercept) | 0.29 | 4.31 | $<.001$ |
| Prev. Correct | 0.37 | 5.39 | $<.001$ |
| Previous Old | 0.13 | 1.53 | $=.13$ |
| Current Old | 0.37 | 4.60 | $<.001$ |
| Feedback | 0.16 | 1.73 | $=.08$ |
| PrevCorr*PrevOld | -0.58 | -0.86 | $=.39$ |
| PrevCorr*CurOld | -0.66 | -4.19 | $<.001$ |
| PrevOld*CurOld | 0.08 | 0.67 | $=.51$ |
| PrevCorr*Feedback | -0.12 | -1.37 | $=.17$ |
| PrevOld*Feedback | -0.52 | -4.49 | $<.001$ |
| CurrOld*Feedback | -0.47 | -4.30 | $<.001$ |
| PrevCorr*PrevOld* | 1.06 | 6.88 | $<.001$ |
| CurrOld |  | $<.001$ |  |
| PrevCorr*PrevOld* | 0.50 | 3.43 | $=.07$ |
| Feedback |  |  |  |
| PrevCorr*CurrOld* | 0.26 | 1.85 | $<.001$ |
| Feedback |  |  |  |
| PrevOld*CurrOld* | 1.00 | 6.01 | $<.001$ |
| Feedback |  |  |  |
| PrevCorr*PrevOld* | -0.89 | -4.22 |  |
| CurrOld*Feedback |  |  |  |

Table 3. Coefficients for accuracy in both experiments.
Being correct on the previous trial increases the odds of being correct on the current trial (previous: $\mathrm{M}_{\text {corr }}=.64$, $\mathrm{M}_{\text {incorr }}=.61$ ). The odds of being correct on the current trial also increase if the previous trial was an old (i.e. studied) word rather than a new (i.e. unstudied) word (prev old $\mathrm{M}=$ .64 , prev new $M=.61$ ). There is no significant effect of the current item's familiarity. There was a significant interaction of previous correctness and the current item type, showing that if a new cue was misidentified as old, subjects were much worse at the current trial (. 59 vs. .64). Most strikingly, there was a significant interaction of the cue's and target's familiarity: old targets were more likely to be identified after an old cue (Old|Old $=.73$, Old|New $=.56$ )regardless of the response to the cue-and new items were similarly more likely to be correctly identified as new after a new cue (New $\mid$ New $=.64$, New $\mid$ Old $=.54$ ).

Figure 3 displays correct rejection of unstudied (New) items and recognition of studied (Old) items as a function of the previous trial's familiarity and response correctness for Exp. 1 (with feedback). Figure 4 displays the same information for Exp. 2.


Figure 3: Proportion of correct responses for unstudied (New) items and studied (Old) items by panel, broken down according to the studied/unstudied status of the item on the previous trial, as well as the correctness of the response on the previous trial. Note that New|(Previous New) items are more likely to be correctly rejected than New|(Previous Old), regardless of the correctness of the response on the previous trial. Similarly, Old|Old accuracy is greater than Old|New.

Previous Incorrect ○


Figure 4: Without feedback, almost the same pattern is evident: Old|Old responses are more accurate than Old|New responses, regardless of correctness on the previous trial. New|(New,Correct) responses are better than New|(Old,Correct), but New|(New,Incorrect) trends lower than $\mathrm{New} \mid$ (Old, Incorrect), breaking the pattern.

We also investigated the 18,023 correct response times using log-linear mixed-effects regression. Shown in Table 4, there was a significant main effect of the previous item's oldness (Previous Old), and a significant interaction of previous oldness with current oldness (PrevOld*CurOld). The mean correct RT when the previous item was old was 506 ms vs. 504 ms when the previous item was new. When
the current item is new, Ss were faster after new items ( 504 ms ) than old items ( 526 ms ). When the current item is old, Ss were faster after old items $(490 \mathrm{~ms})$ than new items ( 507 ms ). This corroborates the accuracy fluency finding, showing an advantage when the current item is the same oldness as the previous item. There was also a marginally significant interaction of previous response correctness, previous oldness, and feedback.

| Factor | Coeff | $t$ | p-value |
| :--- | :---: | :---: | :---: |
| (Intercept) | 497.26 | 65.24 | $<.001$ |
| Prev. Correct | 6.05 | 1.52 | $=.13$ |
| Previous Old | 21.44 | 4.20 | $<.001$ |
| Current Old | -3.20 | -0.69 | $=.49$ |
| Feedback | 0.07 | 0.01 | $=.99$ |
| PrevCorr*PrevOld | -10.05 | -1.58 | $=.11$ |
| PrevCorr*CurOld | 4.31 | 0.73 | $=.47$ |
| PrevOld*CurOld | -27.43 | -3.97 | $<.001$ |
| PrevCorr*Feedback | -4.04 | -0.75 | $=.45$ |
| PrevOld*Feedback | -3.77 | -0.53 | $=.59$ |
| CurrOld*Feedback | 5.12 | 0.80 | $=.42$ |
| PrevCorr*PrevOld* | 3.28 | 0.38 | $=.70$ |
| CurrOld |  |  |  |
| PrevCorr*PrevOld* | 15.97 | 1.83 | $=.07$ |
| Feedback |  | $=.63$ |  |
| PrevCorr*CurrOld* | 3.88 | 0.48 | $=.48$ |
| Feedback |  |  |  |
| PrevOld*CurrOld* | -6.73 | -0.71 | $=.27$ |
| Feedback |  |  |  |
| PrevCorr*PrevOld* | -12.98 | -1.09 | $=.27$ |
| CurrOld*Feedback |  |  |  |

Table 4. Coefficients for correct RTs in both experiments.

In summary, in an incidental-study recognition memory task with fast responding, we found that the oldness of the prior tested word affects the response time and accuracy on this word. When the current test word is studied, having seen a studied word on the previous trial makes you, on average, faster and more accurate on the current trial - regardless of your response on the previous trial. The accuracy effect happened with and without feedback, so the responses cannot merely be driven by feedback. Seeing a studied word reinstates context features from the study list, and those features contribute to the correct recognition on this trial. For unstudied items preceded by other unstudied items, there is no reinstated context from the previous trial to discount. The need for discounting may explain why correct responses for unstudied items preceded by studied items were drastically slower than for unstudied items preceded by unstudied items.

## Discussion

In two recognition memory experiments with time-limited responses-limiting the role of recollection-we found evidence that associations form between incidentallystudied items. Although oldness and semantics can also serve to increase the likelihood of correct recognition, enhanced recognition due purely to sequential context was also observed.

## Context Effects

Roughly additive priming effects were found for oldness (familiarity), semantics, and sequential context. Although many models could account for one or even two of these effects straightforwardly, additional assumptions would be required to account for all three. We begin by making a common assumption in memory modeling that study events are encoded as a set of features and that recognition decisions are made on the basis of a comparison of a probe-also consisting of a set of features-to each stored trace in memory with "old" responses given if this comparison is strong enough (e.g., Hintzman, 1988; Murdock, 1992; Shiffrin \& Steyvers, 1997; Nelson \& Shiffrin, in press). We make the further assumption that some features sampled during the preceding trial are able to "leak" into the probe features present on the current trial. The same leakage is assumed to occur at study, with features of recent items being present in short term memory during the encoding of a subsequent item, and hence joining that item's stored trace (implemented by Nelson \& Shiffrin, in press).

Thus, because the memory trace contains some features from the preceding study item, preceding it by the same item at test leads to a stronger match between the test probe and the memory trace. The same account explains the positive but smaller priming when the preceding test item is semantically related-some of the semantic features overlap, but not the many physical features that also overlap when identity priming is used.

Priming due to oldness or semantics independent of sequential context requires yet more modeling assumptions, for which we turn to the dynamic model of recognition of Cox \& Shiffrin (2012). This model was able to account for the Jacoby-Whitehouse illusion by assuming, as we have thus far, that primes (in this case, previous test items) contribute some features to the current test probe, at least initially (see Cox, Lewis \& Shiffrin, under review, for more details). As more features are sampled and added to the probe, its match to memory evolves over time. If the probe begins with no features at all, the match to memory tends to go down slightly with the first few features sampled, regardless of whether the test item was studied or not. This is because, even if the test item is a target, it will tend not to match most of the other studied items and these mismatches outweigh the single target match until a sufficient number of features are sampled. Thus, after a few features have been sampled, the match for a target test will tend to increase while the match for a foil test will tend to continue to
decrease. If, however, a few features are present at the start of the trial, these initial negative steps are avoided for both targets and foils, leading to a bias to say old. This bias is proportional to the similarity between the prime and the test item. Thus, an old unrelated item will lead to a slight bias, and a semantically related item will lead to a larger bias, as observed in the present studies.

## Old/New Effects

This mechanism is qualitatively consistent with the observed effects of oldness and correctness of the previous trial in the no-feedback condition. If the preceding item is new, it will tend to contribute features that do not match anything on the list, minimizing the similarity not just with the current test item, but with all the traces in recent memory, leading to a lower tendency to respond "old" than if the preceding item had been old. All else being equal, if we assume that the decision on the preceding trial reflects the quality of evidence provided by the test item on the preceding trial, the effect of the oldness of the previous test item should interact with correctness. For example, if the previous trial was a false alarm, then although the previous item was new, it had to contain enough old features to lead the participant to judge it as old. This account then predicts that the effect of the oldness of the previous test item on the current trial is mainly a function of whether the participant thought the previous item was old, manifesting as a crossover interaction between oldness and correctness on the previous trial.

This is the pattern observed in the no-feedback condition, and is consistent with the idea that there is little or no discounting (a la ROUSE; Huber, et al., 2001) of previous item features in that condition. This interaction is absent from the feedback condition, however: one is still more likely to make an "old" response when the previous item was old, but correctness does not have a large effect on responses to old items; rather, correctness only seems to affect responses to new items, with incorrect responses on the previous trial leading to an overall bias to respond "old" on the current trial. In terms of ROUSE's discounting mechanisms, these data suggest that participants might engage in discounting when the previous trial was incorrect, but they only discount new features. One problem with this account, of course, is that it is unclear whether "old" and "new" features can be identified and differentially discounted. Another problem is that there is no clear reason why participants would only discount new features since doing so only leads to more errors.

An alternative explanation in terms of response criteriae.g., requiring more evidence to respond after an errordoes not hold up either, since that would predict increased accuracy after an error, the opposite of what is observed here. In short, although current models of memory might account for most of the results reported here, the old/new effects in the feedback condition seem to require additional mechanisms that will require further research to elucidate.

Further questions include: What is the effect of using lures that were not studied, but are semantically related to
the polyseme, as cues? If an associate of the nondominant meaning is the cue at study, does it still provide an advantage? If the associate is presented after the polyseme at study, is the association still formed?

In the world, things that occur nearby in time (or space) are often related, and if these relations can be remembered they may prove important. Having shown that automatic associations are formed-even between unrelated items-in recognition memory, much work remains to be done to determine how these associations are represented in memory, and what other forms of context they capture.

## Acknowledgments

Thanks to Brendan Johns for helpful discussions.

## References

Cox, G. E., \& Shiffrin, R. M. (2012). Criterion setting and the dynamics of recognition memory. Topics in Cognitive Science, 4(1), 135-150.
Huber, D. E., Shiffrin, R. M., Lyle, K.B., and Ruys, K.I. (2001). Perception and preference in short-term word priming. Psychological Review 108, 149-182.
Jacoby, L. L. and Whitehouse, K. (1989). An illusion of memory: False recognition influenced by unconscious perception. Journal of Experimental Psychology 118, 126-135.
Kahana, M. J. (1996). Associative retrieval processes in free recall. Memory \& Cognition 24, 103-109.
Malmberg, K. J., \& Annis, J. (2012). On the relationship between memory and perception: Sequential dependencies in recognition testing. Journal of Exp. Psychology: General, 141(2), 233-259.
Malmberg, K. J., Holden, J. E., and Shiffrin, R. M. (2004). Modeling the effects of repetitions, similarity, and normative word frequency on old-new recognition and judgments of frequency. Journal of Experimental Psych.: Learning, Memory, and Cognition, 30, 319-331.
Mueller, S. T. \& Shiffrin, R. M. (2006). REM II: A model of the developmental co-evolution of episodic memory and semantic knowledge. Proceedings of the International Conference on Learning and Development.
Nelson, D. L., McEvoy, C. L., and Schreiber, T. A. (1998). University of South Florida word association, rhyme, and word fragment norms. www.usf.edu/FreeAssociation/
Nobel, P. A. and Shiffrin, R. M., (2001). Retrieval processes in recognition and cued recall. Journal of Experimental Psych.: Learning, Memory, \& Cognition, 27, 384-413.
Ratcliff, R. (1978). A Theory of mental retrieval. Psychological Review, 95, 59-108.
Schwartz, G., Howard, M. W., Jing, B., and Kahana, M. J. (2005). Shadows of the past: Temporal retrieval effects in recognition memory. Psych. Science, 16, 898-904.
Shiffrin, R. M. and Steyvers, M. (1997). A model for recognition memory: REM: Retrieving effectively from memory. Psychological Bulletin \& Review. 4(2), 145-166.
Turvey, M. (1973). On peripheral and central processes in vision: Inferences from an information processing analysis of masking with patterned stimuli. Psychological Review, 80, 1-52.

# More Naturalistic Cross-situational Word Learning 

George Kachergis ${ }^{1}$ and Chen $\mathbf{Y u}^{2}$<br>${ }^{1}$ george.kachergis@gmail.com, ${ }^{2}$ chenyu@indiana.edu<br>${ }^{1}$ Psychology Department, Leiden University, the Netherlands<br>${ }^{2}$ Department of Psychological \& Brain Science / Cognitive Science Program, Indiana University, USA


#### Abstract

Previous research has found that people can use word-object co-occurrences from ambiguous situations to learn word meanings (e.g., Yu \& Smith, 2007). However, most studies of cross-situational learning present an equal number of words and objects, which may simplify the problem by, for example, encouraging learners to use assumptions such as each word going with one object. This paper presents several conditions in which the number of words and objects do not match: either additional objects appear at random, or objects appear sometimes without their intended words. These manipulations do generally hurt learning in comparison to balanced conditions, but people still learn a significant proportion of word-object pairings. The results are explored in terms of statistics of the training trials - including contextual diversity and context familiarity-and with the uncertainty- and familiarity-biased associative model. Parametric differences between conditions hint at hidden cognitive constructs.


Keywords: statistical learning; cross-situational learning; language acquisition

## Introduction

Human infants learn words quite quickly despite many challenges facing them, including uncertainty and ambiguity in the language environment. Recent research has studied how learners may acquire word meanings from regularities in the co-occurrence of words and referents (e.g., objects). Such cross-situational statistical word learning relies on two assumptions: 1) that spoken words are often relevant to the visible environment, and 2) that learners can to some extent remember the co-occurrence of multiple words and objects in a scene. Thus, as words and their intended referents are observed in different situations over time, learners can apprehend the correct word-object mappings. Relying only on the regularity of the linguistic environment and basic memory and attention processes, this may be an important method of learning nouns for infants, and even adult travelers in a foreign country.

In adult cross-situational learning studies (e.g., Yu \& Smith 2007), participants are asked to learn the meaning of alien words by watching a series of training trials. On each trial learners see an array of unfamiliar objects (e.g., four sculptures) and hear pseudowords (e.g., stigson, bosa). The meaning of each pseudoword is ambiguous on a given trial, because although each word refers to a single onscreen object, the intended referent is not indicated. In a typical learning scenario, participants attempt to learn 18 wordobject pairings from 27 trials, with four words and four objects given per trial. In this design, each word-referent pair is presented six times over the five-minute training period. Learning a correct word-object pairing requires
some form of accumulation of word-object co-occurrences. When tested on each word and given four trained objects to choose from, participants chose the correct object for half of the 18 words, on average (Yu \& Smith, 2007).

However, learning environments in the real world are likely not as simple: there may be objects present that go unnamed, some spoken words (e.g. articles) do not refer to particular objects, and object names may be spoken when the intended object is not visible. These forms of noise likely interfere with learning to some extent. When a word is heard without the object it previously co-occurred with several times, is a learner to map it to a new object? What if that object already has a name? Conversely, when an object is seen, but the word it previously occurred with is not heard, will learners lose certainty about the old mapping, and even associate a new word with it?

In this study, we take baseline conditions from $\mathrm{Yu} \&$ Smith (2007) that present an equal number of words and objects on each trial and either add or remove words or objects in a systematical way in order to change various cooccurring statistics that learning may rely on. We investigate several critical factors that matter to learning, such as conditional probability of words given objects during learning, final test probability, and contextual diversity-the number of other pairs each pair appears with (Kachergis, Yu, \& Shiffrin, 2009b). Following Fazly, Alishahi, and Stevenson (2010), we also investigate additional two factors - age of exposure (i.e., when a pair first appears) and context familiarity (the mean frequency of the objects a given pair appears with). Not only are these factors likely to influence how well people learn, but likely so will the fact that the trials contain an unequal number of words and objects. Previous studies have also typically presented an equal number of words and objects on each trial, which may induce participants to only consider 1-to-1 mappings (although see Vouloumanos, 2008 as well as mutual exclusivity investigations: Kachergis, 2012; Ichinco, Frank, \& Saxe, 2009; Yurovsky \& Yu, 2008).

Finally, we use a recent associative model of crosssituational learning (Kachergis, Yu, \& Shiffrin, 2012) to shed light on differences between the conditions. The model assumes that learners have access to both their familiarity and their uncertainty about the word-object pairings present on a given trial, and that attention competes for uncertain stimuli and for already-strong pairings. This model matches adult behavior in a number of previous cross-situational experiments (Kachergis, 2012; Kachergis, Yu, \& Shiffrin, 2013).

## Experiment

Participants were asked to learn 18 word-referent pairs from a series of individually ambiguous training trials using the cross-situational word learning paradigm (Yu \& Smith, 2007). Each training trial was comprised of a display of two or more novel objects and two or more spoken pseudowords, depending on condition. With no indication of which word refers to which object, on a single trial, learners can only guess at the correct word-referent mappings. However, since words always appear on trials with their intended referents, the correct pairings may be learned over the series of trials.

In this study, many conditions were created by manipulating training conditions from Yu and Smith (2007)-the $2 \times 2$ (i.e., 2 word-object pairs per trial), $3 \times 3$, and $4 \times 4$ conditions- to introduce different types of noise which is arguably more in line with real-world learning, such as a non-referential word, an unnamed object, or both. In every condition, participants experienced a series of training trials with a total of 18 intended word-object pairs. The same pair was never allowed to appear in neighboring trials in conditions, as previous studies have shown such temporal contiguity improves learning (Kachergis, Yu, \& Shiffrin, 2009a; Kachergis, Yu, \& Shiffrin, 2013). In the baseline $\mathbf{2 \times 2}$ ( 54 trials), $\mathbf{3 \times 3}$ (36 trials), and $\mathbf{4 \times 4}$ ( 27 trials) conditions, each word and object appear 6 times. Every time a given object is present, the intended word is presented ( $p(\mathrm{w} \mid \mathrm{o})=1)$, and every time a given word is presented, the intended object is present $(p(\mathrm{o} \mid \mathrm{w})=1)$. Most conditions in Table 1 were built from these three baseline conditions. We manipulate the number of words and objects per trial, thus changing their frequency. This also changes the probability of hearing the word for a given object on a trial (in Table 1,

Trial $p(\mathrm{w} \mid \mathrm{o}))$. The probability of seeing an object given that its label was heard was always 1 (Trial $p(\mathrm{o} \mid \mathrm{w})$ ). Test $p(\mathrm{o} \mid \mathrm{w})$ in Table 1 shows the probability of guessing the intended object for a given word after experiencing all of the training.

In the $2 \times 4$ condition, words appeared 6 times and objects 12 times, so on each trial the probability of hearing the intended word for a given object is $p(\mathrm{w} \mid \mathrm{o})=.5$. In the $2 \times 3$ condition, objects appear 9 times, making $\mathrm{p}(\mathrm{w} \mid \mathrm{o})=.67$. In the $\mathbf{2 x 4}$ condition, each word appears 6 times and each object appears 12 times. In the $\mathbf{3 x 3}+\mathbf{1 w} / \mathbf{o}$ condition, an additional random word and object were shown on each trial. In the $\mathbf{4 x 4}+\mathbf{2 w} / \mathbf{o}$ condition, two additional random words and objects were shown per trial. In the $\mathbf{3 x 4}$ condition, each word appears 6 times, each object 8 times $(p(w \mid 0)=.75)$. In the $3 \times 41 / .5$ condition, words appear 6 times, and 12 objects appear only with their words $(p(0, w)=1)$, while 6 objects appear 12 times $(p(\mathrm{w} \mid \mathrm{o})=.5)$. In the $\mathbf{3 \times 4} \mathbf{1 / . 6 6}$ condition, words appear with their objects 6 times $(p(\mathrm{w} \mid \mathrm{o})=1)$, but 12 objects appear 3 additional times $(p(\mathrm{w} \mid \mathrm{o})=.66)$ without their words. In the $3 \times 4+60$ condition, 18 word-object pairs cooccur 6 times, and 6 additional objects occur as noise.

The $1 \times 3$ condition divided each trial of the $3 \times 3$ condition into 3 trials with one word and 3 objects, and shuffled the trials so no objects (or words) repeated trial-to-trial. Thus, words appeared 6 times, and objects 24 times $(p(\mathrm{w} \mid \mathrm{o})=.33)$. The $1 \times 4$ condition divided the $4 \times 4$ trials as $1 \times 3$ did for $3 \times 3$, meaning that objects appeared 24 times $(p(\mathrm{w} \mid \mathrm{o})=.25)$.

Calculated for each item per condition, Table 1 also shows the average "Age" of Exposure (trial the pair first appears), Context Familiarity (defined by Fazly, Alishahi, and Stevenson (2010) as the mean co-occurrence with all other pairs across exposures), and Context Diversity (the number of unique pairs a pair co-occurs with over training).

| Condition | Word <br> Freq. | Object <br> Freq. | Trial <br> $p(\mathrm{w} \mid \mathrm{o})$ | Test <br> $p(\mathrm{o} \mid \mathrm{w})$ | Context <br> Familiarity | "Age" of <br> Exposure | Context <br> Diversity | Num. <br> Subjs. | Correct |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\mathbf{2 x 2}$ | 6 | 6 | 1 | 0.5 | 3.5 | 5.6 | 5.1 | 19 | 0.79 |
| $2 \times 3$ | 6 | 9 | 0.67 | 0.33 | 3.3 | 6.0 | 9.1 | 55 | 0.56 |
| $2 \times 4$ | 6 | 12 | 0.5 | 0.25 | 3.3 | 5.6 | 11.8 | 33 | 0.30 |
| $3 \times 3+1 \mathrm{w} / \mathrm{o}$ | 9 | 9 | 1 | 0.22 | 5.1 | 4.3 | 12.9 | 39 | 0.17 |
| $4 \times 4+2 \mathrm{w} / \mathrm{o}$ | 12 | 12 | 1 | 0.12 | 6.6 | 3.8 | 16.2 | 39 | 0.10 |
| $\mathbf{3 x 3}$ | 6 | 6 | 1 | 0.33 | 3.5 | 3.7 | 8.8 | 36 | 0.43 |
| $1 \times 3$ | 6 | 18 | 0.33 | 0.33 | 3.2 | 17.5 | 8.7 | 63 | 0.52 |
| $3 \times 4$ | 6 | 8 | 0.75 | 0.25 | 3.4 | 3.7 | 12.3 | 25 | 0.19 |
| $3 \times 4+60$ | 6 | 6 | 1 | 0.25 | 3.5 | 3.7 | 13.6 | 20 | 0.27 |
| $3 \times 41 / .5$ | 6 | 8 | $1 / .5$ | 0.25 | 4.3 | 3.7 | 11.3 | 25 | 0.22 |
| $3 \times 41 / .66$ | 6 | 8 | $1 / .66$ | 0.25 | 3.6 | 3.7 | 12.1 | 25 | 0.21 |
| $\mathbf{4 x 4}$ | 6 | 6 | 1 | 0.25 | 3.5 | 2.8 | 12.2 | 77 | 0.31 |
| $1 \times 4$ | 6 | 24 | 0.25 | 0.25 | 3.1 | 19.9 | 12.0 | 40 | 0.19 |

Table 1. Summary of conditions in the Experiment.

## Subjects

Participants were undergraduates at Indiana University who received course credit for participating. The number of participants in each condition are shown in Table 1 (Num. Subjs. column). None had participated in previous crosssituational experiments.

## Stimuli

Each training trial consisted of an array of 2-4 uncommon objects (e.g., sculptures) and 2-4 spoken pseudowords, depending on condition (see Table 1). The computergenerated pseudowords are phonotactically-probable in English (e.g., "bosa"), and were spoken by a monotone, synthetic female voice. The words and objects used for each condition were drawn from a set of 72 words and 72 objects.

Training for each condition consisted of between 27 and 108 trials. Each training trial began with the appearance of two to four objects (differing by condition) which remained visible for the entire trial. After 2 s of initial silence, each word was heard (1s per word, with 2 s of silence after each).

## Procedure

Participants were told they would see a series of trials with some objects and alien words, but that the words would be presented in random order. They were also told that their knowledge of which words belong with which objects would be tested at the end.

After each training block, participants' knowledge of word-object mappings was assessed using 18-alternative forced choice (18AFC) testing: on each test trial a single word was played, and the participant was instructed to choose the appropriate object from a display of all 18 trained objects. Each of the 18 words was tested once in a random order.

## Results \& Discussion

As shown in Fig. 1, all of the conditions had mean performance significantly above chance (18AFC chance $=$ .056). The $2 \times 2$ baseline condition had by far the highest performance ( $M=.79$ ). Adding another object to each trialwithout it's intended word-harmed learning ( $2 \times 3$ : $M=.56$ ). $2 \times 4$ adds yet another object, further decreasing both Trial $p(\mathrm{w} \mid \mathrm{o})$ and Test $p(\mathrm{o} \mid \mathrm{w})$, resulting in even lower performance $(M=.30)$. Adding an extra pair ( $3 \times 3+1 \mathrm{w} / \mathrm{o}$ ) or two ( $3 \times 3$ $+2 \mathrm{w} / \mathrm{o}$ ) is even more harmful ( $M=.17, M=.10$, resp.); it both lowers Test $p(\mathrm{o} \mid \mathrm{w})$ and creates more possible pairings to consider on each trial. For another example, the $1 \times 3$ and 1 x 4 conditions are identical in all of the other factors except that there were 1 word and 3 objects in the $1 \times 3$ condition ( 0.33 ) but 1 word and 4 objects in the $1 x 4$ condition ( 0.25 ). This one change caused a dramatic performance difference from $M=.53$ to $M=.19$. Meanwhile, it may not seem like there is a dramatic difference between the $1 \times 3$ and $2 \times 3$ conditions. All this suggest that given multiple factors that can be used to characterize statistical information in training data, and the flexibly of human statistical learning systems, it is difficult to pull apart all of the effects in terms of conditionsespecially in the 3 -word and 4 -word conditions-as a
change in one factor is often correlated with changes in several other factors (e.g., contextual diversity and context familiarity).


Figure 1: Mean accuracy by training condition. Performance was variable, but all conditions were above chance (18AFC=.056). Error bars show +/-SE.

To better understand the effects of the various factors, we fit a logistic mixed-effects regression model to the trial-level accuracy data using the lme4 package in R (Bates and Maechler, 2010; R Development Core Team, 2010). Mixed logit models are more appropriate for forced-choice data than ANOVAs, especially when different conditions yield different amounts of data, as in the present experiment (Jaeger, 2008). The model included subject as a random factor, and Trials/Condition, Word Frequency, Object Frequency, Trial $p(\mathrm{w} \mid \mathrm{o})$, Test $p(\mathrm{o} \mid \mathrm{w})$, Contextual Diversity, Age of Exposure, and Context Familiarity as fixed factors. All of these factors were scaled to remove collinearity. Shown in Table 2, there was a significant negative intercept, showing that participants were less likely to choose the correct answer than the incorrect answer. Trials/Condition and Test $p(\mathrm{o} \mid \mathrm{w})$ both had significant, large positive coefficients (. 75 and .78 ), showing that longer training periods were better, as well as stronger correct associations-both of which occur more in the conditions with fewer pairs per trial (i.e., $2 \times 2$ rather than $4 \times 4$ ).

| Factor | Coefficient | $Z$ | p-value |
| :--- | :---: | :---: | :---: |
| (Intercept) | -0.75 | -9.20 | $<.001$ |
| Trials/Cond | 0.75 | 4.57 | $<.001$ |
| Word Freq | -0.10 | -0.92 | $=.36$ |
| Obj Freq | -0.58 | -2.75 | $<.01$ |
| Trial $p$ (w\|o) | -0.14 | -0.88 | $=.38$ |
| Test $p(\mathrm{o} \mid \mathrm{w})$ | 0.78 | 5.67 | $<.001$ |
| Cont. Fam. | 0.20 | 2.82 | $<.01$ |
| Age of Exp | -0.08 | -1.93 | $=.05$ |
| Cont. Div. | 0.17 | 2.24 | $<.05$ |

Table 2. Summary of logistic regression results.

Word frequency did not contribute significantly to correctness, but object frequency had a negative coefficient, showing that additional repetitions of an object on trials without the intended word indeed inhibited learning of that object. Trial $p(\mathrm{w} \mid \mathrm{o})$ was not a significant predictor of accuracy; it seems the other (partially-correlated) factors better capture the variance. Context Familiarity and Contextual Diversity both have significant positive coefficients (. 20 and .17). Though they are correlated ( $r=.56$ ), these two factors both help people learn words. Age of Exposure had a marginally significant negative coefficient (-.08), showing that earlier-appearing pairs are indeed a bit more likely to be learned.

In total, these results offer a look at the factors that influence cross-situational word learning, and an estimate of their relative magnitudes. We now apply a recent associative model of cross-situational word learning to see whether it can account for word-learning in these noisy environments, and to see whether the recovered parameters yield any additional insight.

## Model

To better understand how the condition demands differ, we introduce an associative model of cross-situational word learning proposed by Kachergis, Yu, and Shiffrin (2012a).

The model assumes that learners do not equally attend to all word-object pairings on a trial (i.e., store all cooccurrences). Rather, selective attention on a trial is drawn to strengthen associations between words and objects that have co-occurred previously. This bias for familiar pairings competes with a bias to attend to stimuli that have no strong associates (e.g., as a novel stimulus). The competing familiarity and uncertainty biases allow the model to exhibit fast mapping, since a novel word-novel object combination will demand more attention, and mutual exclusivity: a novel word will only become weakly associated with an alreadyknown referent (Kachergis, Yu, \& Shiffrin, 2012). For example, suppose word $w_{l}$ and object $o_{l}$ have appeared together and are thus somewhat associated, while $w_{7}$ and $o_{7}$ are novel. Given a trial with both pairs: $\left\{w_{1}, o_{1}, w_{7}, o_{7}\right\}, w_{1}-o_{1}$ demands more attention than $w_{7}-o_{1}, w_{1}-o_{7}$, or $w_{7}-o_{7}$, since $w_{l}-o_{l}$ is stronger than baseline. However, attention is also pulled individually to $w_{7}$ and to $o_{7}$, since both of these novel stimuli have no strong associates. Uncertainty is measured by the entropy of each stimulus' association strengths. Because of the high joint uncertainty of $w_{7}$ and $o_{7}$, more attention is given to the association $w_{7}-O_{7}$. Thus, attention is mostly divided between $w_{I}-o_{I}$ and $w_{7}-O_{7}$, although the other pairings will be strengthened a bit.

Formally, let $M$ be an $n$ word $\times n$ object association matrix that is incrementally built during training. Cell $M_{w, o}$ will be the strength of association between word $w$ and object $o$. Strengths are subject to forgetting (i.e., general decay) but are augmented by viewing the particular stimuli. Before the first trial, $M$ is empty. On each training trial $t$, a subset $S$ of $m$ word-object pairings appears. If new words and objects are seen, new rows and columns are first added.

The initial values for these new rows and columns are $k$, a small constant (here, 0.01).

Trial-to-trial, association strengths decay and then a fixed amount of associative weight, $\chi$, is distributed among the presented word-object associations and added to the strengths. The rule used to distribute $\chi$ (i.e., attention) balances a bias for attending to unknown stimuli with a bias for strengthening already-strong associations. When a word and referent are repeated, extra attention (i.e., $\chi$ ) is given to this pair: a prior knowledge bias. Stimuli with no strong associates also attract attention, whereas pairings between uncertain objects and known words, or vice-versa, draw little attention. Stimulus uncertainty is measured by entropy $(H)$, which is 0 when the outcome of a variable is certain (e.g., a word appears with one object, and has never appeared with any other object), and maximal $\left(\log _{2} n\right)$ when all of the $n$ possible object (or word) associations are equally likely (e.g., when a stimulus has not been observed before, or if a stimulus were to appear with every other stimulus equally). In the model, on each trial the entropy of each word (and object) is calculated from the normalized row (column) vector of associations for that word (object), $p\left(M_{w}\right)$, like so:

$$
H\left(M_{w, \cdot}\right)=-\sum_{i=1}^{n} p\left(M_{w, i}\right) \cdot \log \left(p\left(M_{w, i}\right)\right)
$$

The update rule for allocating attention and adjusting strengths for the stimuli presented on a trial is:

$$
M_{w, o}=\alpha M_{w, o}+\frac{\chi \cdot e^{\lambda \cdot(H(w)+H(o))} \cdot M_{w, o}}{\sum_{w \in S} \sum_{o \in S} e^{\lambda \cdot(H(w)+H(o))} \cdot M_{w, o}}
$$

In this equation, $\alpha$ is a parameter governing forgetting, $\chi$ is the weight being distributed, and $\lambda$ is a scaling parameter governing differential weighting of uncertainty and prior knowledge (familiarity). As $\lambda$ increases, the weight of uncertainty (i.e., the exponentiated entropy term, which includes both the word's and object's association entropies) increases relative to familiarity. The denominator normalizes the numerator so that exactly $\chi$ associative weight is distributed among the potential associations on the trial. Only decay operates for stimuli not on the current trial. After training, a learner is tested with each word and chooses an object from $n$ alternatives in proportion to the association strengths of each alternative to that word.

In sum, this associative model learns trial-by-trial by distributing attention in a way that corresponds with both our intuitions about word-learning-using competing biases for familiar pairings and uncertain stimuli-and a number of empirical findings. Three parameters ( $\chi, \alpha$, and $\lambda$ ) were fit using $\log$ likelihood to each individual's choices in each training condition. The model achieved quite a good fit to the data, with $R^{2}=.98$. Figure 2 shows mean model performance for individuals' model fits by condition. Figure 3 shows individuals' mean performance in each condition versus the model's performance. Next, we investigate the
parameter values for each condition to see what they tell us about the cognitive effects of different types of noise.


Figure 2. Model performance closely matches human performance (Fig. 1) and variability in the Experiment.


Figure 3. Individual performance versus model fit: the model was capable of closely matching the behavior of most participants.

We first looked at correlations between each parameter and performance. $\chi$ was positively correlated with learning (Pearson's $r=.72, t(494)=22.74, p<.001$ ), which is consistent with our interpretation of $\chi$ as a learning rate; how much associative weight can be distributed per trial. $\lambda$ was negatively correlated with performance ( $r=-.22, t(494)=-$ $5.04, p<.001$ ): greater focus on uncertain stimuli seems to harm learning, at least in the conditions of this Experiment. $\lambda$ and $\chi$ were also negatively correlated ( $r=-.20, t(294)=-$ 4.64, $\mathrm{p}<.001$ ), meaning that uncertainty-focused learners tended to have slower learning rates. All other correlations were $<|.03|$, and not significant.

We also investigated whether there were differences in parameters by condition. Ideally, the parameters of a cognitive model should be cognitively interpretable. For
example, in our model, $\chi$ is for now a learning rate per trial, but should likely depend on how many possible associations there are on a trial and how much time there is to consider them. If systematic differences in particular parameters were required to fit performance in some of the conditions, then we may be able to pinpoint which factors learning rate and memory decay depend on, and redefine them in more meaningful units. An ANOVA by condition for each parameter showed significant differences for all three parameters $(\chi: \mathrm{F}(12,482)=11.63, p<.001 ; \lambda: \mathrm{F}(12,482)=2.13$, $p=.01 ; \alpha: \mathrm{F}(12,482)=2.70, p<.01)$. Table 3 shows the mean parameters found for each condition. We emboldened the highest mean values for each parameter and italicized the lowest in order to highlight the conditions with unusual mean parameter values.

For $\chi$, the $2 \times 2$ condition has the highest value (19.47), and this condition also yields the highest performance in humans. $2 \times 2$ also has the lowest $\lambda$ (i.e., more focus on familiarity) and $\alpha$ (i.e., faster decay), the latter of which may mitigate the high learning rate a bit. Conditions with the next-highest learning rates- $2 \times 3$ (9.87) and $1 \times 3$ (10.32)-had the next-highest performance (.56 and .52). $1 \times 3$, along with $1 \times 4$ also had the highest mean $\alpha=.94$ (memory fidelity). These two conditions have the shortest trials (5s), along with the fewest possible associations: only one word and three or four objects. The conditions with the lowest learning rates, $3 \mathrm{x} 4(\chi=.35)$ and $1 \mathrm{x} 4(\chi=.47)$, have fairly low performance (. 19 and .19). The short trial time for the 1 x 4 condition may not give subjects enough time to pick out the single correct pairing.

| Condition | Correct | $\chi$ | $\lambda$ | $\alpha$ |
| :--- | :---: | :---: | :---: | :---: |
| $2 \times 2$ | 0.79 | $\mathbf{1 9 . 4 7}$ | 5.0 | 0.85 |
| $2 \times 3$ | 0.56 | 9.87 | 6.9 | 0.92 |
| $2 \times 4$ | 0.30 | 1.73 | 9.3 | 0.88 |
| $3 \times 3+1 \mathrm{w} / \mathrm{o}$ | 0.17 | 0.91 | 8.6 | 0.89 |
| $4 \times 4+2 \mathrm{w} / \mathrm{o}$ | 0.10 | 3.01 | 8.7 | 0.87 |
| $3 \times 3$ | 0.43 | 6.30 | 6.1 | 0.90 |
| $1 \times 3$ | 0.52 | 10.32 | 7.3 | $\mathbf{0 . 9 4}$ |
| $3 \times 4$ | 0.19 | 0.35 | 7.5 | 0.89 |
| $3 \times 4+6 \mathrm{o}$ | 0.27 | 5.07 | $\mathbf{9 . 2}$ | 0.89 |
| $3 \times 41 / .5$ | 0.22 | 0.99 | 8.0 | 0.92 |
| $3 \times 41 / .66$ | 0.21 | 1.58 | $\mathbf{9 . 1}$ | 0.88 |
| $4 \times 4$ | 0.31 | 2.80 | 7.9 | 0.87 |
| $1 \times 4$ | 0.19 | 0.47 | $\mathbf{9 . 1}$ | $\mathbf{0 . 9 4}$ |

Table 3. Mean of best-fitting parameters for each condition. The largest and smallest mean values of each parameter are emboldened and italicized, respectively.
In the $3 \times 4$ condition, there are again more objects than words, and many possible associations. The other $3 \times 4$ conditions also had low performance and low learning rates,
except for $3 x 4+60$, in which participants may have had little difficulty ignoring the extraneous objects (which are less confusing since they are occur infrequently, and never with a consistent name). It is hard to see a pattern in $\lambda$, the relative focus on uncertainty vs. familiar pairings (roughly, explore vs. exploit). We do not yet have any reason to believe $\lambda$ should remain fixed; learners may well change it-implicitly or strategically-depending on task demands. Moreover, previous investigations found that $\lambda$ has little effect on the shape of learning curves (Kachergis, $\mathrm{Yu}, \&$ Shiffrin, 2012b).

## Discussion

In the language environment, many objects in a scene may go unlabeled, whether they are novel or familiar. For the sake of simplicity, previous studies of cross-situational learning presented an equal number of words and objects on each trial, and a word's intended referent was always present (and vice-versa; e.g. Yu \& Smith, 2007; Kachergis, Yu, \& Shiffrin, 2009a, 2009b). In this study, we presented learners with a variety of conditions with different kinds and degrees of statistical noise (e.g., extra objects, mismatched words and objects). Although performance varied widely in different conditions, learners performed significantly above chance in all conditions.

To better understand what factors influence learning, we measured various statistics about items in each condition, and tried to predict learning from these statistics. Greater contextual diversity-how many pairs a pair appears with during training, context familiarity-the average frequency of pairs a pair appears with, trials per condition, and overall strength of the correct pairing all significantly improved the odds of learning a pair. Being exposed to a pair earlier in training improved learning of that pair, but being exposed to an object more often inhibited learning, because in this study extra occurrences of an object were likely to be noise (e.g., appearing on a trial where it goes unnamed). These conditions and measures provide important constraints for word-learning models, as well as demonstrating that crosssituational learning is robust under a variety of types of noise.

We applied a recent associative word-learning model to these data, and found that it could account for individuals' behavior in each of the conditions. We investigated the average parameter values for individuals in each condition, and found that they differed. The learning rate parameter was strongly linked to overall performance, and was high when there were few pairings to consider on each trial (e.g., $2 \times 2,1 \times 3$ )-unless most of them were noise, and presented quickly (e.g., $1 \times 4$ ). There less memory decay in conditions with very one word per trial, and thus few associations ( $1 \times 3$, $1 \times 4$ ), although the most decay occurred in the $2 \times 2$ condition, but that was balanced by the fast learning rate. Overall, we have a somewhat clearer idea of what the model's parameters do under different noise conditions, but we do not yet have a wholly satisfactory psychological interpretation of them.

In summary, cross-situational learning is robust under a many noise conditions that more closely resemble situations learners may encounter in the real world than in previous studies. Moreover, we have presented a large dataset that we hope will inspire new experiments to test the limits of crosssituational learning, and will constrain and inform modeling efforts.

## References

Bates, D. \& Maechler, M. (2010). lme4: Linear mixedeffects models using S 4 classes. R package version 0.999375-37. http://CRAN.R-project.org/package=lme4

Fazly, A., Alishahi, A., \& Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. Cognitive Science, 34, 1017-1063.
Gleitman, L. (1990). The structural sources of verb meanings. Language Acquisition, 1(1), 1-55.
Ichinco, D., Frank, M.C., \& Saxe, R. (2009). Crosssituational word learning respects mutual exclusivity. In N. Taatgen, H. van Rijn, J. Nerbonne, \& L. Schomaker (Eds.) Proceedings of CogSci 31 (pp. 749-754).
Jaeger, T. F. (2008) Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. Journal of Memory and Language, 59, 434-446.
Kachergis, G., Yu, C., \& Shiffrin, R. M. (2009a). Temporal contiguity in cross-situational statistical learning. In N . Taatgen, H. van Rijn, J. Nerbonne, \& L. Schomaker (Eds.) Proceedings of CogSci 31. Austin, TX: Cognitive Science Society.
Kachergis, G., Yu, C., \& Shiffrin, R. M. (2009b). Frequency and contextual diversity effects in cross-situational word learning. In N. Taatgen, H. van Rijn, J. Nerbonne, \& L. Schomaker (Eds.) Proceedings of CogSci 31 (pp. 755760).

Kachergis, G., Yu, C., \& Shiffrin, R.M. (2012a). An associative model of adaptive inference for learning word-referent mappings. Psychonomic Bulletin \& Review, 19(2), 317-324.
Kachergis, G., Yu, C., \& Shiffrin, R. M. (2012b). Crosssituational word learning is better modeled by associations than hypotheses. IEEE Conference on Development and Learning / EpiRob 2012.
Kachergis, G., Yu, C., \& Shiffrin, R. M. (2013). Actively learning object names across ambiguous situations. Topics in Cognitive Science, 5(1), 200-213.
Markman, E.M. \& Wachtel, G.F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. Cognitive Psychology, 20, 121-157.
Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. Cognition, 107(2), 729-742.
Yu, C. \& Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. Psychological Science, 18, 414-420.
Yurovsky, D. \& Yu, C. (2008). Mutual exclusivity in crosssituational statistical learning. Proceedings of CogSci 30 (pp. 715-720). Austin, TX: Cognitive Science Society.

# The role of thinking-aloud instructions and prior domain knowledge in information processing and source evaluation during Web search 

Yvonne Kammerer (y.kammerer@iwm-kmrc.de)<br>Knowledge Media Research Center, Schleichstr. 6, 72076 Tuebingen, Germany<br>Peter Gerjets (p.gerjets@iwm-kmrc.de)<br>Knowledge Media Research Center, Schleichstr. 6, 72076 Tuebingen, Germany


#### Abstract

This paper examines the impact of thinking-aloud (TA) instructions as well as of individuals' prior domain knowledge on information processing and source evaluation during Web search on a health-related topic. With regard to TA instructions, prompted instructions that entailed evaluation prompts (as used in some previous Web search studies) were compared to neutral instructions (in line with the standards defined by Ericsson \& Simon, 1993) and to a silent condition. To measure participants' $(\mathrm{N}=44)$ information processing and source evaluation we used a rich multi-method approach including eye-tracking methodologies, log file data, and verbal protocols. Results indicate that prompted TA instructions as compared to neutral instructions significantly increased participants' verbal reflections on information quality and on structural aspects of Web pages, given that participants possessed at least a moderate level of prior domain knowledge. In addition, prompted instructions resulted in less linear viewing sequences on the search engine results pages than the silent condition. Finally, the higher participants' prior domain knowledge the more intensely they scrutinized the search results presented by the search engine and the smaller were their average pupil sizes, which indicated lower cognitive load. The significance of the results is considered in light of methodological as well as educational implications.


Keywords: Web search; source evaluation; prior domain knowledge; thinking-aloud instructions; eye tracking

## Introduction

In recent years, the World Wide Web (WWW) has evolved into a major information resource offering easy access to billions of Web pages on almost any topic. However, as anyone can publish virtually any information on the WWW, the quality of Web pages can vary widely. That is, misleading and low-quality Web pages, for example, in the field of medicine and health care, are as common as those providing neutral, high-quality information. Hence, to avoid the selection and use of doubtful or even false information, it is important that Web users themselves critically evaluate the quality of information they retrieve from the Web.

Previous empirical findings about Web users' source evaluation as indicated by verbal reports, however, are inconclusive. Whereas some studies indicate that Web users mainly evaluate search results and Web pages only based on the topical relevance or their ranking in the search engine results page (e.g., Hargittai, Fullerton, Menchen-Trevino, \& Thomas, 2010; Savolainen \& Kari, 2006), others suggest that Web users are also concerned to a substantial extent
about information quality (i.e., the accuracy, authority, objectivity, or currency of information; e.g. Rieh, 2002; Tombros, Ruthven, \& Jose, 2005). According to Tombros et al. (2005) also the structure of Web pages (i.e., the clarity of the Web page or the organization of the information therein) is evaluated extensively.

The aim of the present paper was to identify potential reasons for the divergent findings. Specifically, we examined the impact of the instructions used in the studies as well as of participants' prior domain knowledge on information processing and source evaluation during Web search, using a rich multi-method approach including eye tracking methodologies, log file data, and verbal protocols.

## The Role of Prompted Thinking-Aloud Instructions

One reason for the divergent findings might be a methodological one, namely that in the studies by Rieh (2002) and Tombros et al. (2005) participants were instructed beforehand to explain what criteria they used to evaluate Web information (Tombros et al., 2005) or to select good or credible information during Web search and to explain their evaluation processes in the form of postsearch interviews including specific evaluation-related questions (Rieh, 2002).

According to the seminal work by Ericsson and Simon (1993) and the meta-analysis by Fox, Ericsson, and Best (2011), however, procedures for verbal reporting are only nonreactive (i.e., do not alter thought processes and task performance) when instructions to think aloud are given in a neutral way, by instructing participants to verbalize their thoughts per se. In contrast, procedures that entail describing or explaining thoughts and actions - as it was the case in the studies by Tombros et al. (2005) and Rieh (2002) - are significantly reactive, altering participants' course of cognitive processing and leading to higher task performance than silent conditions. Hence, the instructions used by Tombros et al. (2005) and Rieh (2002) might have increased participants' awareness of the necessity of critically evaluating the information retrieved during Web search. Indirect evidence for this assumption comes from the studies by Hargittai et al. (2010) and Savolainen and Kari (2006) that used neutral thinking-aloud (TA) instructions and only found few utterances related to information quality or structural aspects of Web pages. A first aim of the present study, thus, was to further prove this assumption by directly comparing participants' verbal utterances when given TA
instructions entailing explicit prompts to explain what criteria they used to evaluate search results and Web pages (hereinafter referred to as prompted thinking-aloud) with neutral TA instructions (in line with Ericsson \& Simon, 1993).

Furthermore, previous research that directly compared participants' information processing while thinking-aloud (with neutral or prompted instructions) to a silent condition, showed that with prompted TA instructions that entailed explanation prompts participants who had to solve a series of information tasks on a Website explored the information more extensively than when they worked in silence (Hertzum, Hansen, \& Andersen, 2009). That is, in the former case participants showed a more distributed visual exploration of the screen (i.e., they searched or scanned more across the screen), scrolled more frequently within Web pages, and navigated to more Web pages than in the silent condition. Between the neutral TA instructions and the silent condition, in contrast, only few marginal differences were found. With the present study we aimed at expanding these findings to a Web search scenario including multiple Websites being accessed via a search engine results page (SERP), with participants either receiving prompted TA instructions (including evaluation prompts), neutral TA instructions, or working in silence.

## The Role of Prior Domain Knowledge

Referring to the Elaboration Likelihood Model (ELM; Petty \& Cacioppo, 1986) from persuasion research, Metzger (2007) postulates that the extent to which Web users engage in source evaluations is dependent both on their motivation (which might be increased by evaluation prompts), but also on their ability (e.g. their prior domain knowledge). In this regard it should be noted that participants in Rieh's (2002) study reported to use their prior domain knowledge to evaluate information quality, indicating that they possessed a certain amount of prior knowledge on the search topics. Importantly, previous research has not only shown differences in source evaluations between domain experts and novices, but also between groups of novices varying in their level of prior domain knowledge. For example, a casestudy by MaKinster, Beghetto, and Plucker (2002) that investigated undergraduate students' Web search on a complex science topic through postsearch interviews that entailed evaluation-related questions indicated that students with moderate domain knowledge scrutinized search results more thoroughly, by examining the titles, the page excerpts, and the URLs of the search results, than low-knowledge students. In addition, Bråten, Strømsø, and Salmerón (2011) found that when reading multiple documents dealing with a science-related topic undergraduates with low knowledge on the subject matter trusted the different documents to the same extent irrespective of the type of source, whereas students with higher knowledge judged an article from a company with vested interests about the addressed issue as less trustworthy than the other documents. According to Bråten et al. (2011) a possible explanation for the
undifferentiated or lacking trustworthiness evaluations of low-knowledge readers is that they have to invest more cognitive effort in comprehending the content of the documents than readers with higher domain knowledge. As a consequence, low-knowledge readers might have less cognitive resources available to engage in evaluations that go beyond content. Thus, another central aim of our study was to further examine the effects of prior domain knowledge on information processing and source evaluation during Web search as well as potential interactions between prior domain knowledge and TA instructions.

## Hypotheses of the Present Study

Given the theoretical considerations and prior empirical findings our hypotheses were as follows:

First, with respect to the evaluation of information quality and structural aspects of Web pages as indicated by verbal utterances, we hypothesized that prompted TA instructions would increase the number of respective verbal utterances during Web search as compared to neutral instructions, but only when participants' possessed a certain level of prior knowledge (cf. Metzger, 2007) (H1). Low-knowledge participants' quality-related and structure-related verbal utterances should not be increased due to participants' lack of cognitive resources (cf. Bråten et al., 2011).

Second, based on the findings by Hertzum et al. (2009) we hypothesized that prompted TA instructions as compared to a silent condition would result in a more distributed scanning behavior on SERPs (H2) and a significant increase in Web pages selected from the SERPs during Web search (H3). In contrast, neutral instructions should only result in slight changes in regard to these measures as compared to a silent condition.

Third, according to the case-study results of MaKinster et al. (2002) we hypothesized that the higher participants' domain knowledge the more thoroughly they would scrutinize the search results, as indicated by longer total fixation times (H4). Finally, based on the theoretical assumption provided by Bråten et al. (2011) we hypothesized that the lower participants' domain knowledge, the more cognitively demanding the Web search would be for them, which should result in larger mean pupil sizes (H5) which is known to be an indicator of increased cognitive load (for review e.g. see Hyönä, Tommola, \& Alaja, 1995; Wang, 2010).

## Method

## Participants and Design

Participants ${ }^{1}$ were 44 undergraduates ( 18 male, $M=25.02$ years, $S D=3.68$ ) from different majors at a large German

[^108]university, who were rewarded with either course credit or payment. Participants had normal or corrected to normal vision. They reported to use Google as their primary search engine and judged their computer- and Web search experience and skills between intermediate and high ( $M=$ $3.50, S D=0.67,4$ items, on a scale from $1=$ very low to $5=$ very high, Cronbach's $\alpha=.78$ ).

Thinking-aloud condition served as a between-subjects factor, with participants being randomly assigned to a neutral TA condition, a prompted TA condition, or a silent condition. Neutral instructions to think aloud were worded in line with the standards described by Ericsson and Simon (1993), that is "Please think aloud during your Web search, that is, verbalize everything that comes to your mind." In contrast, the prompted TA instructions were similar to the instructions used, for instance, by Tombros et al. (2005) or Rieh (2002) including evaluation prompts, that is "Please think aloud during your Web search, that is, mention the evaluation criteria you apply to select search results and to assess Web pages." In the silent condition participants performed the task silently.

There were no differences between the three conditions regarding participants' age $\left(\chi^{2}(2, N=44)=2.24, p=.33\right)$, gender $(F<1)$, or computer- and Web search experience and skills $(F<1)$.

As a second factor participants' self-reported prior knowledge on the subject matter of the task (i.e., diets and nutrition) was assessed (see 'Measures' for details) and used as a continuous between-subjects factor.

## Tasks and Web Materials

Participants were given the task of seeking information on the WWW about two competing weight loss methods, namely low carb(ohydrate) diets and low fat diets, in order to give informed advice to a fictitious overweight friend who wants to lose weight by changing her diet.

To complete the experimental task, which was limited to 20 minutes, participants were provided with three prefabricated Google-like SERPs (with ten search results each), retrieved for the search terms "low fat", "low carb", and "low carb versus low fat". Participants were instructed to access all three SERPs during their Web search. They could access all 30 Websites corresponding to the list of search results presented on the SERPs, but were not allowed to generate new search results by changing the search terms. All search results and Websites were relevant to the search topic in regard to the content of information provided, but differed with regard to the type of sources including Websites provided by scientific institutions, journalists, industry and companies, and laypeople.

## Measures

Self-reported prior domain knowledge To assess participants' prior knowledge on diets and nutrition, participants were administered a questionnaire with eight statements that had to be rated on five-point scales ( $1=$ totally disagree; $5=$ totally agree); example items are "I
know more about diets and nutrition than my family and friends" or "I can describe the concept of low carb diets". Cronbach's alpha was .87 for the eight items. The range of participants' prior domain knowledge was 1 to 4.38 ( $M=$ $2.48, S D=0.70$ ). There were no differences between the three experimental conditions regarding participants' prior knowledge on diets and nutrition $(F<1)$.

## Dependent variables

Participants' thinking-aloud protocols (in the prompted and neutral TA conditions) were segmented at a small grain size: Each sentence or utterance preceded and followed by a pause was considered a separate segment. According to the studies by Rieh (2002) and Tombros et al. (2005) utterances were coded as information quality when participants on SERPs or Websites reflected on whether (or not) the information is good, valid, credible, or current, or the source provides trustworthy, reliable, or official information, and as structure of Web pages when participants addressed the clarity of a Web page or the organization of the information therein (cf. Tombros et al., 2005). Besides, utterances addressing the topical relevance (i.e., utterances on SERPs or Websites about whether or not the information matches with the search topic) were coded (e.g., Savolainen \& Kari, 2006). Two raters familiar with the search task and the Web materials scored $30 \%$ of the protocols. Interrater reliability computed on this subsample of protocols yielded a Cohen's kappa of .76. Disagreements were resolved through discussion between the raters. One rater scored the remaining protocols.

In addition, for participants of all three conditions their viewing and selection behavior on the SERPs were analyzed. Eye movements and mouse clicks were recorded by a 50 Hz Tobii 1750 remote eye-tracking system supported by the software ClearView 2.7.1.

To analyze participants' eye-tracking data, for each of the ten search results on a SERP a polygonal area of interest (AOI) was defined around the search result (covering the title, the excerpt, and the URL), in order to determine for how long and in which order a participant was looking at a search result. The minimum fixation duration was set to 100 milliseconds with a fixation radius of 30 pixels. On this basis, the total fixation time of each search result, that is, the overall amount of time participants scrutinized a search result to decide whether to click on it or not, was calculated. Second, as a measure of distributed viewing behavior across the screen (cf. Hertzum et al., 2009) participants' viewing sequences on the SERPs, that is, the order in which the search results on a SERP were inspected, were computed, by comparing them with a perfectly linear sequence from top to bottom (search result 1, search result 2, search result 3 , search result 4 , etc.) by means of the Levenshtein distance (cf. Josephson \& Holmes 2002). Levenshtein distance values were transformed into a similarity percentage: A higher similarity percentage indicated more linear top-tobottom viewing sequences. The similarity percentages were averaged across the three SERPs. Third, participants' mean
pupil size (in mm ) during task processing was computed (averaged across eyes).

Furthermore, based on the mouse click recordings, the number of Websites participants' accessed during their Web search was analyzed (cf. Hertzum et al., 2009).

## Procedure

Participants were tested in individual sessions of approximately one hour. The lighting conditions were kept constant during all examinations. Before participants started on the search task, control variables and self-reported prior domain knowledge were assessed. Furthermore, participants received some general instructions about the Web search experiment as well as the TA instructions according to their experimental condition. Next, they performed a practice task (structured in the same way as the subsequent main task) for approximately five minutes to get acquainted with the Web search environment and with the TA method. After the practice task, participants received the instruction for the main task (i.e., the request of the fictitious friend). Then, they were calibrated on the eye-tracking system using a nine-point calibration and started their Web search. In the two TA conditions, whenever participants stopped verbalizing their thoughts, the experimenter reminded them (after 5 seconds) to keep thinking aloud or to mention the evaluation criteria, respectively. After 20 minutes participants were asked to stop the task. Subsequent to their Web search, they had to decide which of the two weight loss methods they would recommend. However, in the present study only process measures during Web search were analyzed.

## Results

ANCOVAs with TA conditions (neutral, prompted, and silent) and prior domain knowledge (z-scored) as well as an interaction term between TA conditions and prior domain knowledge were conducted. Significant interaction effects between TA condition and prior domain knowledge were probed according to the procedure outlined by Aiken and West (1991). Table 1 shows means and standard errors for the dependent variables.

## Verbal Utterances

With respect to verbal utterances about topical relevance neither TA condition nor prior domain knowledge had a significant effect and there was no significant interaction between the two factors (all $F \mathrm{~s}<1.90, p \mathrm{~s}>.18$ ).

With respect to verbal utterances about information quality the ANCOVA showed a significant effect of TA condition $\left(F(1,25)=10.95, p=.003, \eta_{\mathrm{p}}^{2}=.30\right)$ and of prior knowledge $\left(F(1,25)=6.22, p=.02, \eta_{\mathrm{p}}{ }^{2}=.20\right)$. These effects were qualified by a significant interaction between the two factors $\left(F(1,25)=4.62, p=.04, \eta_{\mathrm{p}}{ }^{2}=.16\right)$. Simple comparisons according to the procedure outlined by Aiken and West (1991) revealed that, consistent with the predictions of $H 1$, only high-knowledge participants (i.e., 1
$S D$ above the sample mean) and moderate-knowledge participants (i.e., at the sample mean) expressed more information-quality utterances during prompted TA than during neutral TA $(\beta=.80, t(25)=3.85, p=.001$ and $\beta=$ $.48, t(25)=3.31, p=.003$, respectively). In contrast, lowknowledge participants (i.e., $1 S D$ below the sample mean) did not significantly differ in the two TA conditions ( $\beta=$ $.16, t(25)=0.79, p=.44)$. In addition, in the prompted TA condition prior knowledge was significantly positively related to the number of information-quality utterances ( $\beta=$ $.67, t(25)=3.30, p=.001$ ), whereas in the neutral TA condition it wasn't $(\beta=.05, t(25)=0.24, p=.81)$.
With respect to verbal utterances regarding the structure of Web pages the ANCOVA showed similar effects: a significant effect of TA condition $(F(1,25)=13.69, p=$ $.001, \eta_{\mathrm{p}}^{2}=.35$ ), a marginal effect of prior knowledge ( $F(1$, 25) $=3.49, p=.07, \eta_{\mathrm{p}}^{2}=.12$ ), and a significant interaction between the two factors $\left(F(1,25)=6.01, p=.02, \eta_{\mathrm{p}}^{2}=.19\right)$. Simple comparisons revealed that only high-knowledge participants and moderate-knowledge participants expressed more structure-related utterances during prompted TA than during neutral $\mathrm{TA}(\beta=.89, t(25)=4.34, p<.001$ and $\beta=$ $.53, t(25)=3.70, p=.001$, respectively). In contrast, lowknowledge participants did not differ in the two TA conditions ( $\beta=.17, t(25)=0.85, p=.41$ ). In addition, as for the information-quality utterances, in the prompted TA condition prior knowledge was significantly positively related to the number of structure-related utterances ( $\beta=$ $.62, t(25)=3.07, p=.01)$, whereas in the neutral condition it wasn't $(\beta=-.08, t(25)=-0.41, p=.69)$.

## Eye-Tracking and Mouse Clicks

With regard to the linearity of participants' viewing sequences on the SERPs, as measured by the similarity percentages of participants' string to a linear top-to-bottom string, the ANCOVA showed a significant main effect of TA condition $\left(F(2,38)=6.19, p=.01, \eta_{p}{ }^{2}=.25\right)$. Bonferroni post-hoc tests revealed that in line with the predictions of H 2 , in the prompted TA condition participants' viewing sequences were significantly less linear than in the silent condition ( $p=.004$ ). The differences between the neutral TA condition and the silent condition were only marginally significant ( $p=.08$ ). The two TA conditions did not differ significantly ( $p=.89$ ). Prior knowledge had no effect on participants' viewing sequences and there was no interaction with TA conditions (both $F \mathrm{~s}<$ 1).

Regarding the total fixation time of each search result the ANCOVA showed a significant main effect of prior knowledge $\left(F(1,38)=4.00, p=.05, \eta_{\mathrm{p}}{ }^{2}=.10\right)$. In line with the predictions by H 4 , the higher participants' domain knowledge, the more intensely they scrutinized the search results to decide whether or not to click on them. TA conditions had no effect $(F<1)$ and there was no interaction with prior knowledge $(F(1,38)=1.52, p=.23)$.

Table 1: Means (and standard errors) of the dependent variables as a function of thinking-aloud (TA) conditions.

|  | Thinking-aloud condition |  |  |
| :--- | :---: | :---: | :---: |
| Dependent variables | Prompted instructions | Neutral instructions | Silent condition |
| \# utterances about topical relevance | $6.92(1.21)$ | $4.51(1.26)$ | NA |
| \# utterances about information quality | $12.11(1.55)$ | $4.74(1.60)$ | NA |
| \# utterances about structure of Web pages | $5.52(0.84)$ | $1.05(0.87)$ | NA |
| \% linearity of viewing sequences on SERPs | $62.11(3.90)$ | $68.10(4.10)$ | $81.40(4.00)$ |
| total fixation time (in s) of each search result | $3.02(0.35)$ | $3.03(0.36)$ | $2.62(0.36)$ |
| mean pupil size (in mm) | $3.71(0.10)$ | $3.86(0.11)$ | $3.58(0.10)$ |
| \# Websites accessed (in \%) | $47.14(4.09)$ | $52.91(4.30)$ | $42.49(4.20)$ |

With regard to participants' mean pupil size during task processing, after controlling for participant's average eye-to-screen-distance (which had a strong positive effect on pupil size, $\left.F(1,37)=7.89, p=.01, \eta_{\mathrm{p}}^{2}=.18\right)$, the ANCOVA showed a significant effect of prior knowledge $(F(1,37)=$ $5.87, p=.02, \eta_{\mathrm{p}}^{2}=.14$ ). In line with the predictions by $H 5$, less prior domain knowledge was associated with larger pupil sizes. Besides, there was no significant effect of TA conditions $(F(2,37)=1.71, p=.20)$, nor a significant interaction with prior knowledge $(F<1)$.

With regard to the number of Websites participants accessed during Web search, the ANCOVA showed no significant main effect of TA conditions $(F(1,38)=1.51, p$ $=.24)$, but of prior knowledge $\left(F(1,38)=5.40, p=.03, \eta_{\mathrm{p}}{ }^{2}\right.$ $=.12$ ). This effect was qualified by a significant interaction with TA conditions $\left(F(2,38)=7.11, p=.002, \eta_{p}{ }^{2}=.27\right)$. Simple comparisons revealed that participants with high prior knowledge accessed significantly more Websites in the prompted TA condition than in the silent or neutral TA condition $(\beta=.45, t(38)=2.26, p=.03$ and $\beta=.53, t(38)=$ 2.32, $p=.03$, respectively), which did not differ significantly $(\beta=.09, t(38)=0.42, p=.68)$. Thus, the assumptions of $H 3$ are at least partly confirmed. However, other than expected, low-knowledge participants accessed significantly more Websites in the neutral TA condition than in the silent or prompted TA condition $(\beta=.51, t(38)=$ $2.41, p=.02$ and $\beta=.72, t(38)=3.65, p=.001$, respectively), which did not differ significantly ( $\beta=.21$, $t(38)=0.97, p=.34)$.

In addition, in the prompted TA condition prior knowledge was significantly positively related to the number of selected search results $(\beta=.93, t(38)=4.13, p<$ .001), whereas in the silent and neutral TA condition no significant relationship with prior knowledge were found ( $\beta$ $=-.38, t(38)=-1.66, p=.11$ and $\beta=.25, t(38)=1.17, p=$ .25 , respectively).

## Discussion and Conclusion

With the present study we sought to examine the impact of TA instructions - especially of "prompted" TA instructions that entail evaluation prompts as used in some previous studies (e.g. Rieh, 2002; Tombros et al., 2005) - as well as of prior domain knowledge on participants' information processing and source evaluation during Web search.

First, with regard to verbal utterances about information quality and the structure of Web pages, in line with $H 1$, our study showed that prompted TA instructions (i.e., to mention the evaluation criteria applied to select search results and to assess Web pages) as compared to neutral TA instructions (i.e., to verbalize everything that comes to their mind) significantly increased participants' verbal reflections on information quality and on structural aspects of Web pages, given that participants possessed at least a moderate level of prior knowledge on diets and nutrition. In contrast, irrespective of the instructions given, students with no or little prior domain knowledge only rarely expressed such utterances. Thus, to conclude, the findings by Rieh (2002) and Tombros et al. (2005) seem to have resulted from a combination of prompted TA instructions and a certain level of prior domain knowledge of the participants. Utterances about topical relevance, on the contrary were expressed by all participants to a similar extent, irrespective of prior knowledge or instructions. That is, the evaluation of whether or not a search result or a Web page addresses the search topic seems to be a default process that guides every Web search task.

Second, with regard to the effects of prompted or neutral TA instructions on participants' information processing during Web search as compared to a silent condition, in line with $H 2$ and the findings by Hertzum et al. (2009) about Web page inspection, we found that the prompted instructions resulted in less linear, that is, more distributed viewing sequences on the SERPs. That is, instead of simply following the list order (a typical effect on SERPs) participants examined the search results in a rather free way. This can be seen as an indication of own evaluations, instead of simply relying on the order presented by the search engine. Between the neutral TA instructions and the silent condition, in contrast, only marginal respective differences were found (cf. Hertzum et al., 2009). With regard to the number of Websites accessed during Web search the results showed a more complex pattern than expected. For participants with rather high prior knowledge similar effects were found as in the study by Hertzum et al. (2009), namely that participants in the prompted TA condition accessed more Websites than in the other conditions (in line with H3). However, for participants with a moderate level of prior knowledge no differences were
found between conditions, and participants with low prior knowledge accessed more Websites with neutral TA instructions than with prompted instructions or when working in silence. A post-hoc explanation for the latter finding might be that low-knowledge participants in the neutral TA condition were overwhelmed by the situation and did not really know what to say and to do and, thus, simply selected (almost) all Websites.

Third, with regard to further effects of prior knowledge, in line with H 4 we found that the higher students' selfreported prior knowledge on diets and nutrition the more intensively they scrutinized the search results on the SERPs to decide whether to select them (cf. MaKinster et al., 2002). Furthermore, in line with $H 5$ participants with lower prior knowledge across conditions had increased pupil sizes, which is an indicator of increased cognitive load. This effect provides further evidence for the assumption proposed by Bråten et al. (2011) that low-knowledge users have to employ more cognitive effort in order to comprehend the content of the Web pages than Web users with higher domain knowledge. As a consequence, they seem to lack cognitive resources to engage in evaluations that go beyond content. This is also indicated by the fact that in both TA conditions low-knowledge users hardly ever verbally reflected on the information quality or on the structure of Web pages. Future research should examine these effects in greater detail, analyzing more detailed moment-to-moment changes in participants' pupil sizes and using withinsubjects task variations.

In sum, from a methodological perspective, the present findings indicate that prompted TA instructions increase participants' verbal reflections on information quality and on structural aspects of Web pages, and, therefore, might overestimate, users' spontaneous reflections about these issues. Yet, from an educational point of view the findings of the present study suggest that simple evaluation instructions that prompt Web users to evaluate search results and Web pages have the potential to increase their awareness about the necessity to critically evaluate the information found on the Web - given that they possess a certain amount of prior knowledge on the subject matter of the search task at hand. In order to improve source evaluations of searchers with no or little prior domain knowledge, however, more comprehensive instructional support seems to be required that also includes measures to reduce individuals' cognitive burden during task processing. Furthermore, when the WWW is used as a research tool in formal learning settings (e.g., in school), it seems advisable that the teacher first provides some general information on the topic before students start with their Web search.

## References

Aiken, L. S., \& West, S. G. (1991). Multiple regression: Testing and interpreting interactions. Newbury Park, CA: Sage.
Bråten, I., Strømsø, H. I., \& Salmerón, L. (2011). Trust and mistrust when students read multiple information sources
about climate change. Learning and Instruction, 21, 180192.

Ericsson, K. A., \& Simon, H. A. (1993). Protocol analysis: Verbal reports as data. Cambridge, MA: MIT Press.
Fox, M. C., Ericsson, K. A., \& Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. Psychological Bulletin, 137, 316-344.
Gerjets, P., Kammerer, Y., \& Werner, B. (2011). Measuring spontaneous and instructed evaluation processes during web search: Integrating concurrent thinking-aloud protocols and eye-tracking data. Learning and Instruction, 21, 220-231.
Hargittai, E., Fullerton, L., Menchen-Trevino, E., \& Thomas, K. Y. (2010). Trust online: young adults' evaluation of Web content. International Journal of Communications, 4, 468-494.
Hertzum, M., Hansen, K. D., \& Andersen, H. H. K. (2009). Scrutinising usability evaluation: Does thinking aloud affect behaviour and mental workload? Behaviour and Information Technology, 28, 165-181.
Hyönä, J., Tommola, J., \& Alaja, A.-M. (1995). Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. Quarterly Journal of Experimental Psychology, 48A, 598-612.
Josephson, S. \& Holmes, M. E. (2002). Visual attention to repeated internet images: testing the scanpath theory on the world wide web. In A.T. Duchowski (Ed), Proceedings of the ETRA 2002 Symposium. ACM Press.
MaKinster, J. G., Beghetto, R. A., \& Plucker, J. A. (2002). Why can't I find Newton's Third Law?: Case studies of students using of the Web as a science resource. Journal of Science Education and Technology, 11, 155-172.
Metzger, M. J. (2007). Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. Journal of the American Society for Information Science and Technology, 58, 2078-2091.
Petty, R. E., \& Cacioppo, J. T. (1986). Communication and persuasion: Central and peripheral routes to attitude change. New York: Springer.
Rieh, S. Y. (2002). Judgment of information quality and cognitive authority in the Web. Journal of the American Society for Information Science and Technology, 53, 145161.

Savolainen, R., \& Kari, J. (2006). User-defined relevance criteria in Web searching. Journal of Documentation, 62, 685-707.
Tombros, A., Ruthven, I., \& Jose, J. M. (2005). How users assess Web pages for information-seeking. Journal of the American Society for Information Science and Technology, 56, 327-344.
Wang, J. T. (2010). Pupil dilation and eye-tracking. In M. Schulte-Mecklenbeck, A. Kuhberger, \& R. Ranyard (Eds.), A handbook of process tracing methods for decision research: a critical review and user's guide. New York: Psychology Press.

# Effects of Numeric Magnitude on the Cortical Valuation Network 

Frank Kanayet (kanayet.1@osu.edu)<br>Department of Psychology, 255 Psychology Building<br>Columbus, OH 43206 USA

John Opfer (opfer.7@osu.edu)<br>Department of Psychology, 245 Psychology Building<br>Columbus, OH 43206 USA

William Cunningham (cunningham@psych.utoronto.ca)<br>Department of Psychology, University of Toronto, 100 St George St Toronto, ON M5S 3G3 Canada


#### Abstract

Previous work has identified a distributed, network of neural systems involved in appraising the value of rewards, such as when winning $\$ 100$. We hypothesized that involvement of intraparietal sulcus (IPS) in this network is specialized for processing numeric rather than monetary value. To test our hypothesis, we manipulated numeric magnitude and units to construct a range of economic rewards (e.g., $+\$ 1,+100 \phi$ ) in response to simple decisions. Consistent with our hypothesis, BOLD activity in IPS was related to changes in numeric magnitude, independent of monetary value, whereas activity in OFC was associated with monetary value, independent of numeric magnitude. Finally, by using representation similarity analysis, we found that the information represented in IPS and OFC was more consistent with the patterns expected if representations of numeric magnitudes or monetary values, respectively, were in a compressive scale. Together, these findings show the importance of numerical cognition for understanding how the brain processes monetary rewards.


Keywords: monetary rewards; IPS; OFC; fMRI, representation similarity analysis, numerical cognition

## Introduction

Humans are continuously faced with choices, often involving incommensurable options. When choosing among options for a romantic date, for example, it is not clear how to compare the esthetic pleasure of a movie with the gustatory pleasure of a nice dinner. Making decisions apparently requires computing the value of each option in some way that would register the difference in values (Montague, King-Casas \& Cohen, 2006).

Economic models have long assumed that humans behave 'as if' they compute value for each option in a common mental currency (i.e., subjective value), and experiments on the distributed neural correlates of valuation (i.e., valuation network) suggest that this assumption may be correct (Kable \& Glimcher, 2009; Padoa-Schioppa \& Asad, 2006, 2008; but see Vlaev, et al., 2011; Tremblay \& Schultz, 1999; Seymour et al., 2007; Nieuwenhuis et al., 2005).

One of the sources of information that comprise this common mental currency is the magnitude of the reward. Being rewarded two cookies feels better than being
rewarded just one. A special case of using magnitude for valuation is the use of monetary rewards. By using money, dissimilar goods can be compared on the same scale (e.g., dollars or cents) and can be described with just one value, its numeric magnitude (e.g., +300 or -300 ).

Although translation of value into a numeric scale has many benefits, it may also come with a price. Numeric magnitude, like luminance and loudness, has a compressive psychophysical scale (Fechner, 1860/1966; Weber, 1846/1948; Dehaene, 1997). Thus, the difference between 10 and 15 appears larger than the difference between 120 and 125.

The compressive nature of numerical judgments is important because it may play a large role in how the brain tracks monetary value and makes economic decisions (Furlong \& Opfer, 2009; Peters et al., 2008). Indeed, "unit effects" on decision-making have been known for many decades. For example, Kahneman and Tversky (1981) observed that participants were willing to trade 20 minutes of their time to save $\$ 5$ on a $\$ 15$ calculator, but not on a $\$ 125$ jacket, even though in both cases they are trading 20 minutes of their time for the same amount of money (i.e. $\$ 5)$. Although, these effects can be explained by assuming that subjects pay more attention to the proportional gains, the compressive function of numeric representations might provide another explanation. Because larger numerals have smaller psychological distances between them, the difference between a $\$ 125$ and $\$ 120$ jacket is subjectively less than the difference between a $\$ 15$ and $\$ 10$ calculator.

In this paper, we were interested in why neural activation also appears to devalue marginal monetary gains. Specifically, we addressed whether the neural response to increasing quantities of money are caused by increases in objective monetary value (the value hypothesis), by increases in the numeric magnitude used to represent the magnitude of the monetary reward (the number hypothesis), both, or neither. This issue is important because neuroeconomists typically assume that the brain areas responsible for processing monetary rewards are not affected by the magnitudes of the numerals that represent the rewards, but this assumption has never been tested.

## The Valuation Network: Neural Correlates of Monetary Value

Research in the field of neuroeconomics has suggested the existence of a neural valuation network. This network computes the subjective value of options under consideration and uses that valuation to make choices. The most critical brain areas associated with economic value are the orbitofrontal cortex (OFC)/ventromedial prefrontal cortex (VMPFC), striatum, anterior cingulate cortex (ACC), and posterior parietal cortex (PPC) (Glimcher, 2009; Kable \& Glimcher, 2009). In theory, the function of this valuation network is to integrate the multiple value dimensions of an option to provide a one-dimensional scale of subjective value according to which choices can be ranked for future decisions.

Of critical importance for this paper are the roles of OFC and PPC. Both studies in monkeys and humans have consistently shown the importance of OFC in the valuation process. There are, however, different ways in which that value can be represented. Although some studies have found neurons in OFC that are associated with absolute value (Paddoa-Schioppa \& Assad, 2006, 2008; Tom et al., 2007), other studies have found that other neurons in OFC are also associated with relative value with adaptive scaling (Kennerley, Behrens \& Wallis, 2011; Tremblay \& Schultz, 1999).

Parietal cortex activity related to valuation processes has been located in the lateral inferior parietal cortex (LIP) of monkeys and intraparietal sulcus (IPS) - its human homologue (Clithero, Carter, \& Huettel, 2009; Kable \& Glimcher, 2009; Platt \& Glimcher, 1999). For instance, using pattern classification techniques, activity in IPS was related to the value of options, and it was even able to distinguish between intertemporal and probabilistic valuations (Clithero et al., 2009). Also, their data suggest that IPS is critical for the initial stages of valuation by representing and integrating the information necessary for computation of economic value in OFC and the striatum. Also, activation of IPS has recently been related to the outcome of monetary rewards, but not to the outcome of social rewards (Lin, Adolphs \& Rangel, 2011). This result is important since it shows that the presence of (numeric) magnitude information in the reward presented may be a critical component of the value representation in IPS.

However, the meaning of magnitude in the studies reviewed is ambiguous. Because numeric magnitude and value magnitude typically go hand in hand, it is not possible to know if the increases in activation in IPS (or OFC) are related to an increase in the value of the reward or in the numbers used to represent that value. Moreover, there is strong evidence that suggests that IPS is a central area in the processing of numeric information (Arsalidou \& Taylor, 2011 for a meta analysis). Therefore, we suggest - as an alternative hypothesis - that while OFC does process reward value, the role of IPS in these studies is to process the numeric magnitudes of the rewards being considered. If true, the activity of the valuation network would be
susceptible to manipulations of numeric magnitude even when these manipulations do not change the objective monetary value of the rewards.

## Present Study

Recently, Furlong and Opfer (2009) provided a method to discern between these two possibilities at the behavioral level. Although economic theories assume that the magnitude of the numbers should not affect economic behavior, Furlong and Opfer showed that in fact numeric magnitude and not economic value explained the degree of cooperation of participants in a prisoner's dilemma task. The device used to prove this point was exceedingly simple. By manipulating the unit of the rewards between dollars and cents, it was possible to achieve rewards with the same objective economic value while drastically changing the numeric magnitude associated with the same reward (e.g. \$1 $=100 \not \subset$ ). This simple manipulation makes it possible to provide participants with a variety of rewards in such a way that allows to parametrically vary numeric magnitude and economic value independently.

To test whether IPS processes numeric magnitude or economic value, we conducted a functional magnetic resonance imaging (fMRI) study that was designed to introduce linear transformations to the magnitudes of rewards. In order to properly disambiguate the effects of numeric magnitude from those of monetary value on the valuation network, we developed a scratch-off lottery game in which we could manipulate the units (between dollars and cents) of the monetary rewards given to participants.

## Method

## Participants

Seventeen adults participated (mean age 22.2, range 18-41; 10 female). All were right-handed, had normal or corrected-to-normal vision and reported no neurological problems. One participant was excluded for failing to complete the task and complaints of headaches during scanning.

## Design and Procedure

Participants were recruited to play a lottery game; $\$ 15$ was guaranteed for playing plus the chance to earn $\$ 0$ to $\$ 20$ more depending on the value of tickets uncovered during the experiment. To uncover extra money, participants had to choose between two covered tickets (represented as two gray rectangles on a computer screen) by pressing one of two buttons on a button box. After choosing a ticket, the amount of extra money earned (or lost) would be revealed. Participants had only one second to choose a ticket lest the choice be made for them; were 25 tickets missed during the session, all extra money would be forfeit.

Unbeknownst to participants, the lottery was rigged in several ways to optimize data for our experiment. First, the sequence of rewards and the jittered intertrial interval were presented in a pseudo-random order, determined by a custom MATLAB script (Poldrack, 2011) that optimizes
contrast efficiencies of fMRI event-related designs. Jittered intertrial intervals varied from 2 s to 8 s and were derived from a pseudoexponential distribution (mean ITI $=4 \mathrm{~s}$ ). The optimization routine was created for each of the 5 individual runs, and order of runs varied randomly between subjects. Thus, participants had no actual control of the amount of money they received.

Critically, values of tickets came from all possible combinations of 5 numbers (i.e., $0,1,3,100,300$ ), 2 units (i.e., dollars and cents) and 2 valences (i.e., win and loss). Combined, these components yielded a range of 17 possible tickets: $(-1 \phi,-3 \phi,-100 \phi,-300 \phi,-\$ 1,-\$ 3,-\$ 100,-\$ 300,0$, $+1 \phi,+3 \phi,+100 \phi,+300 \phi,+\$ 1,+\$ 3,+\$ 100,+\$ 300)$. To control for number of digits and position of units, rewards were presented such that valence signs always appeared in the leftmost position, units rightmost, and numbers between with three digits and one dot (e.g., "- $1.00 \not \subset "$ ".

The experiment consisted of 5 fMRI runs of 8 minutes each. Each run contained 57 trials, 51 trials corresponding to 3 repetitions of each of the 17 different tickets, and 6 extra tickets. Extra trials were added because equal repetitions of all tickets would yield no net gain thereby earning participants no extra money. Instead, the lottery was rigged so all participants earned an extra $\$ 10.50$ from the 30 extra tickets distributed randomly over the 5 runs.

## fMRI Scanning Parameters

Imaging data was collected on a Siemens Tim MAGNETOM Trio 3T MRI scanner. For registration of images, we used a T1-weighted MPRAGE sequence (TR = $1900 \mathrm{~ms} ; \mathrm{TE}=4.68 \mathrm{~ms}$ ). In each run, we acquired 237 whole-brain T2* weighted echo planar images (TR = $2100 \mathrm{~ms} ; \mathrm{TE}=25 \mathrm{~ms}$; flip angle $90^{\circ}$ ). The first 4 volumes of images were discarded to allow for stabilization of the scanner. Parameters of functional scans were selected to minimize susceptibility problems associated with imaging of prefrontal cortex (PFC).

## Data Analysis

fMRI data were analyzed using FEAT 5.98 (FMRI Expert Analysis Tool) from FSL toolbox (www.fmrib.ox.ac.uk/fsl). Preprocessing of data consisted of brain extraction, motion correction, spatial smoothing with a 5 mm (FWHM) Gaussian kernel, and registration to standard MNI space.

Statistical analyses were conducted with a whole-brain GLM parametric analysis in which parametric regressors were created to model wins and losses separately to account for the different subjective value functions predicted by prospect theory (Kahneman \& Tversky, 1979). Specifically, activity for each trial was modeled using units (i.e. dollars = 1 , cents $=-1$ ) and numbers (i.e. $1,3,100,300$ ) as regressors:

BOLD(wins) $=$ Units(wins) + Number(wins) + (Number(wins) * Units(wins))
BOLD $($ losses $)=$ Units $($ losses $)+$ Number(losses $)+$ (Number(losses) * Units(losses))
Where the Number*Units interaction corresponds to the objective monetary value of each ticket. In addition to these regressors of interest, motion correction parameters from

MCFLIRT motion correction procedure were also included in the models as regressors of no interest. Whole brain statistical analyses were performed using a multi-stage approach to implement a mixed-effects model treating participants as random-effects. Regressors of interest were constructed by convolving a boxcar function representing the onset time of the stimulus, the magnitude of the parametric regressor and its duration with a canonical double-gamma (HRF). All reported results in the following section were assessed for cluster-wise significance ( $\mathrm{P}<$ 0.05 , FWE-corrected) using a defining threshold of $Z>2.3$.

## Results

## Behavioral Results

To ensure that participants were paying attention to the task, they were instructed to choose a lottery ticket within 1 s of their onset on the screen. The typical participant was very attentive and only missed 3.88 tickets (range $0-11$ ).

## Imaging Results

The experimental design of this study allowed examining the effects of manipulating numeric magnitude, units and valence on neural activity. Wins and losses were modeled separately in agreement with prospect theory (Kahneman \& Tversky, 1979). For space reasons, we will only describe results concerning winning trials.

## Win Trials

As predicted by the number hypothesis, we found that bilateral activation of IPS was related to increases in numeric magnitude, but not to increases in monetary value. The activation clusters associated with the number parametric regressor extended to adjacent areas in lateral occipital cortex, inferior temporal gyrus, superior parietal lobule and angular gyrus. Also, we found significant clusters in middle frontal gyrus. (Fig. 1). Further, no clusters showed a negative relation between number and neural activation. These patterns are consistent with the literature on number processing (Arsalidou \& Taylor, 2011) and show the importance of numeric information in the processing of monetary rewards. Further, these findings contradict the idea that the role of IPS in the valuation network is to compute economic value (Glimcher, 2009).

Conversely, activity in bilateral OFC, insula, inferior frontal gyrus, ACC, VMPFC, angular gyrus, and lateral occipital gyrus, (Fig. 2) was associated with increases in monetary value. These results are consistent with what is known about the neural correlates of absolute value (Kable \& Glimcher, 2009; Padoa-Schioppa \& Assad, 2008).

Activity associated with the units regressor (i.e. greater activity for dollars than for cents) was found in areas of bilateral OFC, insula, VMPFC, paracingulate cortex, ACC, and left striatum (Fig. 1). Here, areas that show significant relations with receiving rewards in dollars overlap with the areas associated with increases in monetary rewards. This overlap makes sense because - all else being equal -


Figure 1: Regions for which activation was significantly modulated by numeric magnitude (red), monetary value (blue), dollars (yellow) and cents (green) of winning tickets.
changing the unit of the received lottery ticket from cents to dollars entailed a 100 -fold increase in monetary value. Conversely, areas of left postcentral gyrus, anterior IPS, and right lateral occipital gyrus showed greater activity for cents than for dollars. Following a similar logic, a change from dollars to cents - holding the amount of money constant entailed a 100 -fold increase in numeric magnitude. However, as can be seen in Fig. 1 and unlike in the previous case, the areas in posterior parietal cortex that showed significant BOLD activity related to cents do not overlap with the areas associated with increases in numeric magnitude.

Evidence that PPC activation is associated with numeric magnitude adds to the current literature of the neural correlates of valuation by pointing out an important confound present in all studies of valuation that have used monetary rewards. Several of these studies have reported IPS activity and as a result have suggested that PPC is directly implicated in the network that computes economic value (Ballard \& Knutson, 2009; Clithero et al., 2009; Hare, et al., 2011; Lin, et al., 2011; Louie, et al., 2011; Nieuwenhuis, et al., 2005; Platt \& Glimcher, 1999). However, the present results suggest that the involvement of IPS in the valuation network is related to the processing of numeric magnitude information and not to economic value.

Conversely, signatures of absolute monetary values were obtained in OFC, VMPFC, striatum and ACC. In these areas, the magnitudes of the numbers used to represent the economic values did not affect the representation of economic value. These results are consistent with the previous literature (Kable \& Glimcher, 2009) since these are all major areas of the suggested neural network charged with processing economic value. Combined, the findings from winning tickets are in agreement with the idea that there are in fact multiple valuation networks that may have different properties.

## Representation Similarity Analysis

An important question underlying this study is whether brain regions like IPS or OFC treat $100 ¢$ more like $\$ 1$ (same economic value) or like $\$ 100$ (same numeric value). One way to answer this question is to examine how the


Figure 2: Regions of OFC, VMPFC and insula for which activation was significantly modulated by increases in monetary value for winning tickets.
information of interest is represented in a particular brain area. Here, the main goal is not just to detect activation, but to characterize the information present in the particular area (Kriegeskorte \& Bandettini, 2007). Representation similarity analysis (RSA; Kriegeskorte, Mur \& Bandettini, 2008) is one kind of multivariate approach to fMRI data analysis that tries to accomplish this. RSA aims to find correspondences between the relations among stimuli, and the relations between the patterns of brain activation in a particular brain area in response to the same stimuli. Therefore, RSA can be applied to the present problem, since it can provide an answer to the question of whether a given brain area treats the full matrix of monetary rewards more like the numerical representation of those rewards or more like a sequence of distinct monetary values.

Furthermore, we were interested in comparing multiple theories of how the brain patterns of activity elicited by the full set of stimuli might be related. For the purposes of this study, RSA allowed us to compare the patterns of brain activity from anatomical regions of interest (ROI's) to the patterns expected if the given brain area processes numeric magnitude (both in linear and logarithmic scales), or monetary value (both in linear and logarithmic scales). Additionally, by using RSA we were able to check if positive and negative rewards were treated equally or not.

To conduct RSA, we computed dissimilarity matrices (DSMs) among all presented stimuli for the patterns of activity in each ROI, as well as for each theoretical model. Once these DSMs were obtained, Spearman correlations were computed between the ROIs and the model DSMs. This analysis allowed us to rank order the ROI-model similarities (Fig. 3).

As can be seen in Fig. 4, the DSM obtained from IPS activation was more similar to the DSM expected if the information represented were numeric magnitude in a log scale. Conversely, the DSM obtained from OFC activation was more similar to the DSM expected if the information represented were monetary value in a log scale (Fig. 4). These results not only are consistent with the GLM results presented in the previous section, but also provide important additional information regarding the details of the neural


Figure 3: Middle: Single subject IPS dissimilarity matrix (DSM); Top Left: Absolute-magnitude Log Number DSM;
Top Right: Log Number DSM; Bottom Left: Absolutemagnitude Log Monetary Value DSM; Bottom Right: Log Monetary Value DSM. In RSA the brain DSM is correlated with Spearman correlations to the model DSMs to provide a rank order of which model matches better the brain data.
representations. In particular, the results from RSA show that both numeric magnitude in IPS, and monetary value in OFC are represented in compressive scales.

## Discussion

We proposed that the neural response to monetary rewards could be accurately predicted by the cognitive components of valuation. One such component, the processing of numeric magnitude, seemed likely to be especially important, though previous studies had not controlled it systematically. We thought this an important oversight: because the function relating objective numeric magnitudes to subjective magnitudes is compressive, a similar relation might exist in the neural response to monetary rewards.

The results presented in this paper generally confirm this hypothesis. In particular, when winning money of varying amounts, IPS activity was strongly associated with the numeric - and not monetary - value of the rewards. In contrast, activity of OFC, insula, ACC and VMPFC was strongly associated with the monetary - and not numeric value of rewards. Further, RSA showed that numeric information in IPS and monetary value in OFC are represented in a compressive scale.

The fact the IPS was associated with increases in numeric magnitude and not to monetary value provides a new way to understand previous studies about the valuation network (Ballard \& Knutson, 2009; Clithero et al., 2009; Hare, et al., 2011; Louie, et al., 2011; Nieuwenhuis, et al., 2005). In these studies IPS activity was interpreted as processing monetary value, but our results suggest that it is better understood as processing numeric information. Moreover, this conclusion is strengthened by the fact that IPS has been continuously associated with processing of numeric and mathematical information (Arsalidou \& Taylor, 2011).

On the other hand, the finding that activity in OFC,


Figure 4: Spearman correlations between model DSMs and group-averaged DSM for IPS (left) and OFC (right).

ACC, VMPFC, and insula was related to monetary value, is consistent with a wealth of studies that have established strong relations between these areas and the process of monetary value (Cunningham et al., 2009, Glimcher, 2009, O’Doherty, et al., 2003; Padoa-Schioppa \& Assad, 2008). Thus, it seems that though the value hypothesis applies to OFC, the number hypothesis applies to the IPS.

Our results from RSA suggest that both numeric magnitude on IPS and monetary value in OFC are represented in compressive scales. An interesting question that should be explored further is whether the information is first compressed in one area (e.g., OFC uses the already compressed numerical information from IPS when processing monetary rewards) or whether information is compressed in both areas independently. Thus, performing effective connectivity analyses such as dynamic causal modeling (Friston et al., 2003) could provide useful information about the interactions between these brain areas.

Combined with the effects that numeric information have on economic behavior (Furlong \& Opfer, 2009; Peters et al., 2008), the implications of these findings can be far reaching. The fact that simply changing the numerical magnitude of a reward (without altering at all the monetary value) can create these stark effects on the neural valuation network and in particular in IPS - implies that individual differences in IPS should predict differences in decisions that involve monetary information. Therefore, people who suffer dyscalculia or neurological disorders that affect the functionality of parietal cortex (such as Williams or Turner Syndrome) may be at risk for deficits in economic decisionmaking.

For example, Peters and collaborators (2008) found that individual differences in both numeracy and number sense had an impact on the use of numeric information on economic decisions. Activity in IPS in response to monetary value can provide a neural link to this line of research.

Finally, the fact that both numeric magnitude and monetary value are represented in a compressive scale suggests that more attention has to be paid at the way we present monetary information when important decisions have to be made. For example, recent political discussions about deficit reduction deal with extremely large numeric
magnitudes. The compressive scales that we use to represent money imply that it is very likely that decisions made with very large values would not be consistent with the decisions made in an equivalent situation with smaller numeric magnitudes. If monetary values are treated differently when only the numbers used to represent them are different, people might be easily deceived in supporting proposals that go against their own preferences.

## References

Arsalidou, M., \& Taylor, M. J. (2011). Is $2+2=4$ ? Metaanalyses of brain areas needed for numbers and calculations. NeuroImage, 54(3), 2382-3293.
Ballard, K., Knutson, B. (2009). Dissociable neural representations of future reward magnitude and delay during temporal discounting. NeuroImage, 45, 143-150.
Clithero, J.A., Carter, R.M., \& Huettel, S.A. (2009). Local pattern classification differentiates processes of economic valuation. NeuroImage, 45, 1329-1338.
Cunningham, W. A., Kesek, A., Mowrer, S.M. (2009). Distinct orbitofrontal regions encode stimulus and choice valuation. Journal of Cognitive Neuroscience, 21, 19561966.

Dehaene, S. (1997). The number sense: How the mind creates mathematics. New York: Oxford University Press.
Fechner, G. T. (1966). Elements of psychophysics. (H. E. Adler, Trans.) New York: Holt, Rinehart and Winston. (Original work published 1860).
Friston, K. J., Harrison, L. \& Penny, W. (2003). Dynamical causal modeling. NeuroImage, 19(4), 1273-1302.
Furlong, E. E., \& Opfer, J. E. (2009). Cognitive constraints on how economic rewards affect cooperation. Psychological Science, 20, 11-16.
Glimcher, P. W. (2009). Neuroeconomics: Decision making and the brain. P. W. Glimcher, C. F. Camerer, E. Fehr \& R. A. Poldrack (Eds.). New York: Academic Press.

Hare, T. A., Schultz, W., Camerer, C., O'Doherty, J., \& Rangel, A. (2011). Transformation of stimulus value signals into motor commands during simple choice. Proceedings of the National Academy of Sciences of United States of America, 108, 18120-18125.
Kable, J.W. \& Glimcher, P.W. (2009). The Neurobiology of decision: Consensus and controversy. Neuron, 63(6), 733745.

Kahneman, D., Tversky, A., 1979. Prospect theory: An analysis of decision under risk. Econometrica, 47, 263 291.

Kennerley, S.W., Behrens, T.E. \& Wallis, J.D. (2011). Double dissociation of value computations in orbitofrontal and anterior cingulate neurons. Nature Neuroscience, 14, 1581-1589.
Kriegeskorte, N. \& Bandettini, P. A. (2007). Analyzing for information, not activation, to exploit high-resolution fMRI. NeuroImage, 38(4), 649-662.
Kriegeskorte ,N., Mur, M. \& Bandettini, P. A. (2008). Representational similarity analysis - connecting the
branches of systems neuroscience. \#Frontiers in Systems Neuroscience. doi:10.3389/neuro.06.004.2008.
Lin, A., Adolphs, A., \& Rangel, A. (2011). Social and monetary reward learning engage overlapping neural substrates. Social Cognitive and Affective Neuroscience, 7(3), 274-281.
Louie K., Grattan L.E., \& Glimcher, P.W. (2011). Reward value-based gain control: Divisive normalization in parietal cortex. Journal of Neuroscience, 31(29), 1062710639.

Montague, PR, King-Casas, B, Cohen, JD. (2006). Imaging valuation models in human choice. Annual Review of Neuroscience, 29, 417-448.
Moyer, R. S. \& Landauer, T. K. (1967). Time required for judgments of numerical inequality. Nature, 215, 15191520.

Nieuwenhuis, S., Heslenfeld, D.J., Alting von Geusau, N.J., Mars, R.B., Holroyd, C.B., \& Yeung, N. (2005). Activity in human reward-sensitive brain areas is strongly context dependent. NeuroImage, 25, 1302-1309.
O’Doherty, J., Critchley, H., Deichmann, R., \& Dolan, R. J. (2003). Dissociating valence of outcome from behavioral control in human orbital and ventral prefrontal cortices. Journal of Neuroscience, 23(21), 7931-7939.
Padoa-Schioppa, C. \& Assad, J. A. (2006). Neurons in the orbitofrontal cortex encode economic value. Nature, 441(7090), 223-6.
Padoa-Schioppa, C. \& Assad, J. A. (2008). The representation of economic value in the orbitofrontal cortex is invariant for changes of menu. Nature Neuroscience, 11(1), 95-102.
Peters, E., Slovic, P., Västfjäll, D., \& Mertz, C. K. (2008). Intuitive numbers guide decisions. Judgment and Decision Making, 3(8), 619-635.
Platt, M. L. \& Glimcher, P. W. (1999). Neural correlates of decision variables in parietal cortex. Nature, 400, 233238.

Seymour, B., Daw, N., Dayan, P., Singer, T. \& Dolan, R. (2007). Differential encoding of losses and gains in the human striatum. Journal of Neuroscience, 27(18), 48264831.

Tom, S. M., Fox, C. R., Trepel, C. \& Poldrack, R. A. (2007). The neural basis of loss aversion in decision making under risk. Science, 315, 515-518.
Tremblay, L. \& Schultz, W. (1999). Relative reward preference in primate orbitofrontal cortex. Nature, 398, 704-708.
Tversky, A., \& Kahneman, D. (1981). The framing of decisions and the psychology of choice. Science, 211, 453-458.
Vlaev, I, Chater, N., Stewart, N. \& Brown, G. D. (2011). Does the brain calculate value? Trends in Cognitive Sciences. 15(546-554).
Weber, E. H. (1948). Concerning Touch. In D. Wayne (Ed.) Readings in the History of Psychology. New York: Appleton-Century-Croft

# The Funny Thing About Incongruity: A Computational Model of Humor in Puns 

Justine T. Kao ${ }^{1}$ (justinek @ stanford.edu), Roger Levy ${ }^{2}$ (rlevy @ucsd.edu), Noah D. Goodman ${ }^{1}$ (ngoodman@stanford.edu)<br>${ }^{1}$ Department of Psychology, Stanford University. ${ }^{2}$ Department of Linguistics, UC San Diego.

Abstract<br>Researchers showed the robot ten puns, hoping that one of them would make it laugh. Unfortunately, no pun in ten did.


#### Abstract

What makes something funny? Humor theorists posit that incongruity-perceiving a situation from different viewpoints and finding the resulting interpretations to be incompatiblecontributes to sensations of mirth. In this paper, we use a computational model of sentence comprehension to formalize incongruity and test its relationship to humor in puns. By combining a noisy channel model of language comprehension and standard information theoretic measures, we derive two dimensions of incongruity-ambiguity of meaning and distinctiveness of viewpoints-and use them to predict humans' judgments of funniness. Results showed that both ambiguity and distinctiveness are significant predictors of humor. Additionally, our model automatically identifies specific features of a pun that make it amusing. We thus show how a probabilistic model of sentence comprehension can help explain essential features of the complex phenomenon of linguistic humor.


Keywords: Humor; language understanding; probabilistic models

## Introduction

Humor plays an essential role in human interactions: it has important positive effects on children's development (Frank \& McGhee, 1989), success in the work place (Duncan et al. 1990), coping with illness and traumatic events (Gelkopf \& Kreitler, 1996), and marital satisfaction (Ziv \& Gadish, 1989). Indeed, in a study on gender differences in desired characteristics of relationship partners, both men and women rated sense of humor as more important than physical attractiveness and earning potential (Stewart et al., 2000). In this paper, we are interested in understanding how this fundamental and ubiquitous phenomenon works from the perspective of cognitive science. What makes something funny? How might defining characteristics of humor shed light on the ways in which the mind processes and evaluates information?

A leading theory of humor posits that incongruityperceiving a situation from different viewpoints and finding the resulting interpretations to be incompatible-contributes to sensations of mirth (Veale, 2004, Forabosco, 1992, Martin, 2007; Hurley et al., 2011; an idea that dates to Kant's theories about laughter and the sublime (Veatch, 1998). Although there is disagreement about whether incongruity alone is sufficient, most theorists accept that incongruity is necessary for producing humor: as Veale (2004) states, "Of the few sweeping generalizations one can make about humor that are neither controversial or trivially false, one is surely that humor is a phenomenon that relies on incongruity." However, definitions of incongruity are often ambiguous and difficult to operationalize in empirical research. In this paper, we use a computational model of language understanding to formalize a notion of incongruity and test its relationship to humor.

Language understanding in general, and particularly humor, relies on rich commonsense knowledge and discourse understanding. To somewhat limit the scope of our task, we focus on applying formalizations of incongruity to a subset of linguistic humor: puns. Writer and philosopher Henri Bergson defined a pun as "a sentence or utterance in which two ideas are expressed, and we are confronted with only one series of words." This highlights the fact that one sentence must evoke two different interpretations in order to be a pun, which aligns with the concept of incongruity as a requisite of humor.

We develop our model on homophone puns-puns containing words that sound identical to other words in the English language-because the space of possible interpretations of a homophone pun is relatively constrained and well-defined. An example helps to illustrate:

## "The magician got so mad he pulled his hare out."

This sentence allows for two interpretations:
(a) The magician got so mad he performed the trick of pulling a rabbit out of his hat.
(b) The magician got so mad he (idiomatically) pulled out the hair on his head.

If the comprehender interprets the word "hare" as itself, he will arrive at interpretation (a); if he interprets the word as its homophone "hair," he will arrive at interpretation (b). The sentence-level differences between interpretations (a) and (b) can thus be approximated by the two interpretations of the observed word "hare." In general, distinct interpretations of a homophone pun hinges on one phonetically ambiguous word, allowing the two lexical forms of the homophone word to stand in for competing interpretations of the entire sentence.

Critically, even though the example we gave was a written pun and the reader sees the word "hare" explicitly on the page, the "hair" interpretation is still present and even salient in the context of the sentence. Here we explore the idea that puns such as these arise and are funny when they are due to noisy-channel processing. Noisy channel models of sentence processing posit that language comprehension is a rational process that incorporates uncertainty about surface input to arrive at sentence-level interpretations that are globally coherent (Levy, 2008; Levy et al., 2009). Comprehenders can thus consider multiple word-level interpretations ("viewpoints") to arrive at more than one interpretation of a sentence, each coherent but potentially incongruous with each other. The notion of incongruity thus fits naturally into a noisy channel model of sentence comprehension.

Our purposes for developing a formal model of linguistic humor are two-fold. First, we wish to formalize the concept
of incongruity and test assumptions adopted by leading theories in humor research. Secondly, we aim to show that a noisy channel of language processing allows for flexible context selection and sentence comprehension that gives rise to sophisticated linguistic and social meaning such as humor.

## Model

Incongruity is a property of the interpretations derived from a sentence. In order to formalize incongruity, we first describe a probabilistic model of sentence comprehension. Our model aims to infer the topic of a sentence (a coarse representation of its meaning) from the observed words. Unlike previous such models, however, we take a noisy channel approach, assuming that the comprehender maintains uncertainty over which words reflect the sentence topic and which are noise. From this model we derive two quantities that may contribute to humor: ambiguity and distinctiveness. Intuitively, if the resulting interpretation is unambiguous, then no incongruity exists and the sentence is unlikely to be funny. However, since many ambiguous sentences are not funny (e.g. "I went to the bank"), ambiguity alone is insufficient. This is because the interpretations of such sentences are not supported by distinct topical subsets of the sentence (or "viewpoints"). In other words, there must be a set of words in the sentence that support one interpretation and a set that supports the other, and these two sets must be different or "distinct" from each other in order to evoke a sense of incongruity.

Assume our sentence is composed of a vector of content words $\vec{w}=\left\{w_{1}, \ldots, w_{i}, h, w_{i+1}, \ldots, w_{n}\right\}$, including a phonetically ambiguous word $h$. We will use a simple generative model for $\vec{w}$ (see Figure 1): given the latent sentence topic $m$, each word is generated independently by first deciding if it reflects the topic (the indicator variable $f_{i}$ ). If so it is sampled based on semantic relevance to $m$; if not it is sampled from a fixed unigram prior over words. We thus view the sentence as a mixture of topical and non-topical words. Similar approaches have been used in generative models of language to account for words that provide non-semantic information, such as topic models that incorporate syntax (Griffiths et al. 2005). Our model is motivated by the important role that semantic priming plays in lexical disambiguation during sentence processing (Seidenberg et al. 1982; Burke \& Yee, 1984); while ignoring the other non-semantic factors of interpretation (which may also be important).

We make the simplifying assumption that the plausible candidate topics $m$ of the sentence correspond to the potential interpretations of the homophone word $h$, which are constrained by phonetic similarity to two alternatives, $m_{1}$ and $m_{2}$. For example, in the magician pun described above, $h$ is the phonetically ambiguous target word "hare," and $m_{1}$ and $m_{2}$ are the candidate interpretations hare and hair. The two potential topics of the sentence can be identified by the two interpretations hare and hair. This assumption reduces the ill-defined space of sentence meanings to the simple proxy of alternate spellings for phonetically ambiguous words.


Figure 1: Generative model of a sentence. Each word $w_{i}$ is generated based on the sentence topic $m$ if the indicator variable $f_{i}$ puts it in semantic focus; otherwise it is generated as noise (from a unigram distribution).

Using the above generative model, we can infer the joint probability distribution $P(m, \vec{f} \mid \vec{w})$ of the sentence topic $m$ and the indicator variables $\vec{f}$ that determine whether each word is in semantic focus. This distribution can be factorized into:

$$
\begin{equation*}
P(m, \vec{f} \mid \vec{w})=P(m \mid \vec{w}) P(\vec{f} \mid m, \vec{w}) \tag{1}
\end{equation*}
$$

The two terms on the right-hand side are the basis for our derivations of measures for ambiguity and distinctiveness respectively. Ambiguity means the presence of two similarly likely interpretations and can be quantified as a summary of the distribution $P(m \mid \vec{w})$. Distinctiveness measures the degree to which two interpretations are supported by "distinct" viewpoints of the sentence, which we represent as the divergence between sets of words that are in semantic focus given the two values of $m$; it can be quantified as a summary of the distribution $P(\vec{f} \mid m, \vec{w})$. Together, these two measures constitute our formalization of incongruity.

Ambiguity Let $M$ denote the distribution $P(m \mid \vec{w})$, a binomial distribution over the two meaning values $m_{1}$ and $m_{2}$ given the observed words. If the entropy of this distribution is low, this means that the probability mass is concentrated on only one meaning, and the alternative meaning is unlikely given the observed words. If entropy is high, this means that the probability mass is more evenly distributed among $m_{1}$ and $m_{1}$, and the two interpretations are similarly likely given the contexts. The entropy of $P(m \mid \vec{w})$ is thus a natural measure of the degree of ambiguity present in a sentence. We compute $P(m \mid \vec{w})$ as follows:

$$
\begin{align*}
P(m \mid \vec{w}) & =\sum_{\vec{f}} P(m, \vec{f} \mid \vec{w})  \tag{2}\\
& \propto \sum_{\vec{f}} P(\vec{w} \mid m, \vec{f}) P(m) P(\vec{f})  \tag{3}\\
& =\sum_{\vec{f}}\left(P(m) P(\vec{f}) \prod_{i} P\left(w_{i} \mid m, f_{i}\right)\right) \tag{4}
\end{align*}
$$

We approximate $P(m)$ as the unigram frequency of the words that represent $m$. For example, $P(m=$ hare $)$ is approximated
as $P(m=$ "hare"). We also assume a uniform probability that each words is in focus-hence $P(\vec{f})$ is a constant. As for $P(m \mid \vec{w})$, note that it is driven in part by the semantic relationship between $m$ and $\vec{w}$ and in part by the prior probability of $m$, which we approximate using the unigram probability of the words $m_{1}$ and $m_{2}$. From the generative model,

$$
P\left(w_{i} \mid m, f_{i}\right)= \begin{cases}P\left(w_{i}\right), & \text { if } f=0 \\ P\left(w_{i} \mid m\right), & \text { if } f=1\end{cases}
$$

Once we derive $P(m \mid \vec{w})$, we then compute its entropy as a measure of ambiguity.

Distinctiveness We next turn to the distribution over focus sets, given sentence topic. This may be computed as follows:

$$
\begin{equation*}
P(\vec{f} \mid m, \vec{w}) \propto P(\vec{w} \mid m, \vec{f}) P(\vec{f} \mid m) \tag{5}
\end{equation*}
$$

Since $\vec{f}$ and $m$ are independent, $P(\vec{f} \mid m)=P(\vec{f})$.
Let $F_{1}$ denote the distribution $P\left(f \mid m_{1}, \vec{w}\right)$ and $F_{2}$ denote the distribution $P\left(f \mid m_{2}, \vec{w}\right) . F_{1}$ and $F_{2}$ represent the distributions over semantic focus sets assuming the sentence topic $m_{1}$ and $m_{2}$, respectively. We use a symmetrized Kullback-Leibler divergence score $D_{K L}\left(F_{1} \| F_{2}\right)+D_{K L}\left(F_{2} \| F_{1}\right)$ to measure the distance between $F_{1}$ and $F_{2}$. This score measures how "distinct" the semantic focus sets are given $m_{1}$ and $m_{2}$. A low KL score would indicate that meanings $m_{1}$ and $m_{2}$ are supported by similar subsets of the sentence; a high KL score would indicate that $m_{1}$ and $m_{2}$ are supported by distinct subsets of the sentence.

## Evaluation

By generating a large corpus of sentences involving the same words and measuring subjective funniness of each sentence we can evaluate the contribution of each of our quantitative measures, ambiguity and distinctiveness, to humor. We evaluate our model and measures on a set of 235 sentences, consisting of 65 puns, 40 "de-punned" control sentences that are matched with a subset of the puns, and 130 non-pun control sentences that match the puns in containing the same phonetically ambiguous words.

## Materials

We selected 40 pun sentences from a large collection of puns on a website called Pun of the Day, which contains over one thousand puns. Puns were selected such that the ambiguous item is a single phonetically ambiguous word, and no two puns in the collection have the same ambiguous item. To obtain more homophone pun items, a research assistant generated an additional 25 pun sentences based on a separate list of homophone words.

We constructed 40 sentences to be minimally different from the pun sentences that we collected from "Pun of the Day," which we will call de-punned sentences. A second research assistant who was blind to the hypothesis was asked to replace one word in each of the pun sentences (without
changing the homophone word itself) so that the sentence is still grammatical but is no longer a pun. This resulted in sentences that differed from the pun sentences by one word each.

The 130 non-pun sentences were chosen to match each pun sentence on its ambiguous word as well as the alternative homophone. The sentences were taken from an online version of Heinle's Newbury House Dictionary of American English (http://nhd.heinle.com/). We selected sample sentences included in the definition of the homophone word. This design ensured that puns, de-punned, and non-pun sentences all contain the same set of phonetically ambiguous words. Table 1 shows example sentences from each category.

| Type | Example |
| :--- | :--- |
| Pun | The magician got so mad he pulled his hare out. |
| De-pun | The professor got so mad he pulled his hare out. |
| Non-pun | The hare ran rapidly across the field. |
| Non-pun | Some people have lots of hair on their heads. |

Table 1: Example sentences from each category

## Human ratings of semantic relatedness

As described in the model section, computing our measures requires the prior probabilities of meanings $P(m)$ (approximated as the unigram probabilities of the words that denote the meanings), the prior probabilities of words $P(w)$, and the conditional probabilities of each word in the sentence given a meaning $P(w \mid m)$. While we computed $P(w)$ and $P(m)$ directly from the Google Web unigram corpus, $P(w \mid m)$ is difficult to obtain through traditional topic models trained on corpora due to data sparsity. However, since each meaning we consider has a single word as proxy, we may approximate $P(w \mid m)$ using an empirical measure of the semantic relatedness between $w$ and $m$, denoted $R(c, m)$. We use $R(c, m)$ as a proxy for point wise mutual information between $c$ and $m$, defined as follows:
$R(w, m)=\log \frac{P(w, m)}{P(w) P(m)}=\log \frac{P(w \mid m)}{P(w)}=\log P(w \mid m)-\log P(w)$
We assume that human ratings of relatedness between two words $R^{\prime}(w, m)$ approximate true relatedness up to an additive constant $z$. With the proper substitutions and transformations,

$$
\begin{equation*}
P(w \mid m)=e^{R^{\prime}(w, m)+z} P(w) \tag{6}
\end{equation*}
$$

To obtain $R^{\prime}(w, m)$ for each of the words $w$ in the stimuli sentences, we recruited 200 subjects on Amazon's Mechanical Turk to rate distinct word pairs on their semantic relatedness. Since it is difficult to obtain the relatedness rating of a word with itself, we used a free parameter $r$ and fit it to data. Function words were removed from each of the sentences in our dataset, and the remaining words were paired with each of the interpretations of the homophone sequence (e.g., for the pun in Figure 1, "magician" and "hare" is a legitimate word pair, as well as "magician" and "hair"). This resulted in 1460


Figure 2: (a) In the example pun (top), two candidate meanings of $h$ are each more related to a subset of the content words. In the non-pun, only one candidate meaning is more related. (b) Content words are similarly related to both candidate meanings in puns; more related to alternative meanings in de-puns; more related to observed meanings in non-pun. (c) Funniness varies across the sentence types in a pattern that reflects the balance of relatedness to candidate meanings shown in (b).
distinct word pairs. Each subject saw 146 pairs of words in random order and were asked to rate how related each word pair is using a scale from 1 to 10 . The average split-half correlation of the relatedness ratings was 0.916 , indicating that semantic relatedness was a reliable measure.

Figure 2 a) shows the relatedness of each content word with the two homophone interpretations for two example sentences. In the top sentence, which is a pun, the word "magician" is rated as significantly more related to "hare" than it is to "hair", while the word "pulled" is rated as significantly more related to "hair" than it is to "hare." In the bottom sentence, which is a non-pun, all words except the neutral word "through" are more related to the word "hare" than to "hair."

Figure 2b) shows the average relatedness ratings of words and the two homophone interpretations across the three types of sentences. In pun sentences, the average relatedness of words to the two homophone interpretations are not significantly different. In the de-punned sentences, the average relatedness of words to the alternative meaning is significantly higher than to the observed meaning. In the non-pun sentences, the average relatedness of words to the observed meaning is significantly higher than to the alternative meaning. These analyses suggest that relatedness ratings for the two candidate meanings capture the presence or absence of multiple interpretations in a sentence. It further supports our model's prediction that ambiguity of meaning and the distinctiveness of supporting context words can help distinguish among the three types of sentences.

## Human Ratings of Funniness

We obtained funniness ratings of the 235 sentences from 100 subjects on Amazon's Mechanical Turk. Each subject read roughly 60 sentences in random order, counterbalanced

|  | Estimate | Std. Error | p value |
| :--- | ---: | ---: | :--- |
| Intercept | -0.699 | 0.180 | $<0.0001$ |
| Ambiguity | 1.338 | 0.245 | $<0.0001$ |
| Distinctiveness | 0.183 | 0.053 | $<0.0001$ |

Table 2: Regression coefficients using ambiguity and distinctiveness to predict funniness ratings
for the sentence types, and rated each sentence on funniness and correctness. The average split-half correlation of funniness ratings was 0.83 . Figure 2 (c) shows the average funniness ratings of puns, de-punned, and non-pun sentences. Pun sentences are rated as significantly funnier than de-punned sentences, and de-punned sentences are rated as significantly funnier than non-pun sentences $(F(2,232)=$ 415.3, $p<0.0001$ ). Figure 2 (b) and Figure 2 (c) together suggest that the balance of relatedness between the two interpretations is a predictor of funniness.

## Results

Following the derivations described in the model section and using the relatedness measures described above, we computed an ambiguity and distinctiveness value for each of the 235 sentences. Our model has two free parameters-the additive constant $z$ in equation (6) and the constant $r$ that indicates the relatedness of a word with itself -which we optimized using $R^{2}$ in the linear regression summarized in Table 2 As predicted, ambiguity differs significantly across sentence types $(F(2,232)=25.42, p<0.0001)$ and correlates significantly with human ratings of funniness across the 235 sentences ( $r=0.33, p<0.0001$ ). Furthermore, distinctiveness scores differ significantly across sentence types as well $(F(2,232)=5.76, p<0.005)$ and correlates signifi-

| $m 1$ | $m 2$ | Type | Sentence and Semantic Focus Sets | Amb. | Disj. | Funniness |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| hare | hair | Pun | The magician got so mad he pulled his hare | 0.570 | 3.405 | 1.714 |
|  |  | De-pun | The professor got so mad he pulled his hare out. | 0.575 | 2.698 | 0.328 |
|  |  | Non-pun | The hare ran rapidly through the fields. | 0.055 | 2.791 | -0.400 |
|  |  | Non-pun | Most people have lots of hair on their heads. | $2.76 E^{-5}$ | 3.920 | -0.343 |
| tiers | tears | Pun | It was an emotional wedding. Even the cake was in tiers. | 0.333 | 3.424 | 1.541 |
|  |  | De-pun | It was an emotional wedding. Even the mother-in-law was in tiers. | 0.693 | 2.916 | 0.057 |
|  |  | Non-pun | Boxes are stacked in tiers in the warehouse. | 0.018 | 3.203 | -0.560 |
|  |  | Non-pun | Tears ran down her cheeks as she watched a sad movie. | $1.73 E^{-5}$ | 4.397 | -0.569 |

Table 3: Semantic focus sets, ambiguity/disjointedness scores, and funniness ratings for two groups of sentences. Words in red are in semantic focus with $m_{1}$; green with $m_{2}$; blue with both. Semantic focus sets for all sentences can be found at http://www.stanford.edu/~justinek/Pun/focusSets.html


Figure 3: Standard error ellipses of ambiguity and distinctiveness across sentence types. Puns score higher on ambiguity and distinctiveness; de-puns are less supported by distinct focus sets; non-puns have low ambiguity.
cantly with human ratings of funniness ( $r=0.21, p<0.005$ ).
A linear regression showed that both ambiguity and distinctiveness are significant predictors of funniness. Together, the two predictors capture a modest but significant amount of the variance in funniness ratings $\left(F(2,232)=20.86, R^{2}=\right.$ $0.145, p<0.001$; see Table 27. Using both ambiguity and distinctiveness as dimensions that formalize incongruity, we can distinguish among puns, non-puns, and de-punned sentences, as shown in Figure 3 Figure 3 shows the standard error ellipses for each of the three sentence types in the two-dimensional space of ambiguity and distinctiveness. Although there is a fair amount of noise in the predictors (likely due to simplifying assumptions, the need to use empirical measures of relatedness, and the inherent complexity of humor) we see that pun sentences tend to cluster at a space with higher ambiguity and distinctiveness. While de-punned sentences are also high on ambiguity (e.g. it is ambiguous whether the word "hare" in "The professor got so mad he pulled his hare out" should be interpreted as hair), they tend to have lower distinctiveness measures. Non-puns score the


Figure 4: Funniness contours smoothed using a 2-D Loess regression with ambiguity and disjointedness measures as predictors. Sentences become funnier as they move to high ambiguity and distinctiveness space.
lowest on ambiguity with moderate distinctiveness measures.
Figure 4 shows the funniness contours in the twodimensional ambiguity-distinctiveness space smoothed using a 2-D Loess regression. Not only do the three types of sentences differ along the two dimensions, but sentences become funnier as they increase in ambiguity and distinctiveness. These results suggest that our measures of incongruity capture an important aspect of humor in pun sentences.

Beyond predicting the funniness of a sentence, our model can also tell us which particular features of a pun make it amusing. By finding the most likely semantic focus sets $\vec{f}$ given each latent meaning variable $m$ and the observed words, we can identify words in a funny sentence that are critical to producing incongruity and humor. Table 3 shows the most likely semantic focus sets given each meaning for two groups of sentences. Sentences in each group contain the same pair of candidate meanings for the target word $h$. However, they differ in measures of ambiguity, distinctiveness, and funniness. Words in the most likely focus sets given $m_{1}$ are in red; words in the most likely focus sets given $m_{2}$ are in green; and
words in the most likely focus sets of both meanings are in dark blue. We observe that visually, the two pun sentences (which are significantly funnier) have more distinctive and balanced sets of focus words for each meaning than other sentences in their groups. De-punned sentences tend to have fewer words in support of $m_{1}$, and non-pun sentences tend to have no words in support of the interpretation that was not observed. Moreover, imagine if you were asked to explain why the two pun sentences are funny. The colorful words in each pun sentence-for example, the fact that magicians tend to perform magic tricks with hares, and people tend to be described as pulling out their hair when angry-are what one might intuitively use to explain why the sentence is a pun. Our model thus provides a natural way of not only formalizing incongruity and using it to predict when a sentence is a pun, but also to explain what aspects of a pun make it funny.

## Discussion

Researchers in artificial intelligence have argued that given the importance of humor in human communication, computers need to generate and detect humor in order to interact with humans more effectively (Mihalcea \& Strapparava, 2006). However, most work in computational humor has focused either on joke-specific templates and schemata (Binsted, 1996; Kiddon \& Brun, 2011) or surface linguistic features that predict humorous intent (Mihalcea \& Strapparava, 2006, Reyes et al., 2010). Our work moves beyond these approaches and directly utilizes a model of sentence comprehension to derive theory-driven measures of humor.

While the measures we developed account for a significant amount of variance in funniness ratings, there are several ways to improve our model of language in order to more accurately capture the subtleties of linguistic humor. By making the simplifying assumption that semantic association drives sentence comprehension, we disregarded the sequential structure of language that is often important for understanding a pun. For example, "The actors had one great movie after another. They were on a role." scores high on funniness but low on our measures because it leverages the idiomatic expression "on a roll" to boost the interpretation roll. Since our bag-ofwords model does not account for word sequences, the measures we derive fail to fully capture the incongruity of many pun sentences that contain idiomatic expressions. In future work, we aim to incorporate information about the sequential structure of a sentence to further improve our language model and measures of incongruity.

In this paper, we showed how a basic model of sentence comprehension can illuminate incongruous sentence interpretations with rich social and linguistic meaning. Although our task in this paper is limited in scope, we believe that it represents a step towards developing models of language that can explain complex phenomena such as humor. From the perspective of language understanding, such phenomena can serve as probes for developing models of language that account for the subtleties of linguistic behavior. We hope that
our work contributes to research in humor theory, computational humor, and language understanding, with the aim to one day understand what makes us laugh and build robots that appreciate the wonders of word play.

## Acknowledgments

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship to JTK and a John S. McDonnell Foundation Scholar Award to NDG. We thank Stu Melton and Mia Polansky for helping prepare part of the materials used in this paper.

## References

Binsted, K. (1996). Machine humour: An implemented model of puns.
Burke, D., \& Yee, P. (1984). Semantic priming during sentence processing by young and older adults. Developmental Psychology, 20(5), 903.
Duncan, W., Smeltzer, L., \& Leap, T. (1990). Humor and work: Applications of joking behavior to management. Journal of Management, 16(2), 255-278.
Forabosco, G. (1992). Cognitive aspects of the humor process: The concept of incongruity. Humor: International Journal of Humor Research; Humor: International Journal of Humor Research.
Frank, M., \& McGhee, P. (1989). Humor and children's development: A guide to practical applications. Routledge.
Gelkopf, M., \& Kreitler, S. (1996). Is humor only fun, an alternative cure or magic? the cognitive therapeutic potential of humor. Journal of Cognitive Psychotherapy, 10(4), 235-254.
Griffiths, T., Steyvers, M., Blei, D., \& Tenenbaum, J. (2005). Integrating topics and syntax. Advances in neural information processing systems, 17, 537-544.
Hurley, M., Dennett, D., \& Adams, R. (2011). Inside jokes: Using humor to reverse-engineer the mind. Mit Pr.
Kiddon, C., \& Brun, Y. (2011). Thats what she said: double entendre identification. In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies (pp. 89-94).
Levy, R. (2008). A noisy-channel model of rational human sentence comprehension under uncertain input. In Proceedings of the conference on empirical methods in natural language processing (pp. 234-243).
Levy, R., Bicknell, K., Slattery, T., \& Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. Proceedings of the National Academy of Sciences, 106(50), 21086-21090.
Martin, R. (2007). The psychology of humor: An integrative approach. Academic Press.
Mihalcea, R., \& Strapparava, C. (2006). Learning to laugh (automatically): Computational models for humor recognition. Computational Intelligence, 22(2), 126-142.
Reyes, A., Buscaldi, D., \& Rosso, P. (2010). The impact of semantic and morphosyntactic ambiguity on automatic humour recognition. In H. Horacek, E. Mtais, R. Muoz, \& M. Wolska (Eds.), Natural language processing and information systems (Vol. 5723, p. 130-141). Springer Berlin / Heidelberg.

Seidenberg, M., Tanenhaus, M., Leiman, J., \& Bienkowski, M. (1982). Automatic access of the meanings of ambiguous words in context: Some limitations of knowledge-based processing. Cognitive psychology, 14(4), 489-537.
Stewart, S., Stinnett, H., \& Rosenfeld, L. (2000). Sex differences in desired characteristics of short-term and long-term relationship partners. Journal of Social and Personal Relationships, 17(6), 843-853.
Veale, T. (2004). Incongruity in humor: Root cause or epiphenomenon? Humor-International Journal of Humor Research, 17(4), 419-428.
Veatch, T. (1998). A theory of humor. Humor, 11, 161-215.
Ziv, A., \& Gadish, O. (1989). Humor and marital satisfaction. The journal of social psychology, 129(6), 759-768.

# Nuanced Social Inferences about Trustworthiness from Observation of Mimicry 

Liam C. Kavanagh (lkavanag@ucsd.edu)<br>Department of Psychology, University of California, San Diego<br>9500 Gilman Drive La Jolla, CA

Giti Bakhtiari (giti.bakhtiari@uni-wuerzburg.de)
Department of Psychology, University of Würzburg
Röntgenring 10, Würzburg, DE
Christopher L. Suhler (clsuhler@ucsd.edu)
Department of Philosophy, University of California, San Diego
9500 Gilman Drive La Jolla, CA
Patricia S. Churchland (pschurchland@ucsd.edu)
Department of Philosophy, University of California, San Diego
9500 Gilman Drive La Jolla, CA

Rob Holland (r.holland@psych.ru.ni)<br>Department of Social and Cultural Psychology, Radboud University Nijmegen 6500 HE Nijmegen, Netherlands

Piotr Winkielman (pwinkielman@ucsd.edu)
Department of Psychology, University of California, San Diego
9500 Gilman Drive La Jolla, CA


#### Abstract

Mimicry and imitation are crucial mechanisms of social learning and rapport. Further, mimicry informs essential social judgments formed not only by the interacting party but also by third-party observers. How sophisticated are observer's inferences from mimicry? We examined this in the context of observers' use of mimicry to judge trustworthiness. Participants observed a dyadic interaction in which a target mimicked or did not mimic a model. Prior to observation, the model's honesty was earlier defamed, or praised, in front of some, but not other, targets. Observers always knew the model's reputation. Observers also knew which targets were aware of the model's reputation. Results suggest that observers' use of mimicry in trust judgments is very sophisticated It reflects not just the presence of mimicry, but also the model's moral reputation and, critically, knowledge of the target's awareness of the model's reputation. This sophistication leads observers to rate targets as trustworthy when they mimic untrustworthy models, but only when the observers know that the model reputation is unknown to the target.


Keywords: Mimicry; Imitation; Inference, Social Judgment; Trust

## Introduction

Mimicry is an essential part of the human social repertoire that is inexorably bound up to basic social processes of empathy, bonding, and in-group formation (Churchland, 2011; Kashima, 2008; Lakin \& Chartrand, 2003; Lakin, Jefferis, Cheng, \& Chartrand 2003; Preston \& De Waal, 2002). We have greater rapport with those who mimic us
(Bernieri 1998), and are more prosocial after being mimicked (Lakin, Jefferis, Cheng, \& Chartrand, 2003). Interestingly, many of these effects occur without participants' consciousness of mimicry (Chartrand \& Bargh, 1999).

To date, the mimicry literature has focused on the interacting dyad (for a review, see Chartrand \& van Baaren, 2009). However, the social context of mimicry often includes many interacting parties. Here, human observers can use information about who mimics whom to make social judgments. This was shown in a recent series of experiments by Kavanagh, Suhler, Churchland, and Winkielman (2011). Participants (observers) viewed videos of one-on-one interviews, and evaluated the interviewee's competence. In some videos the interviewee mimicked the interviewer's gestures (leg crossing, chin-rubbing) and in the other videos the interviewee did not mimic. Additionally, the attitude of the interviewer towards the interviewee was manipulated: in some videos, the interviewers were cordial to the interviewee and in others they were condescending to the interviewee.

The results showed that the impact of mimicry on the observers' (participants') judgments of interviewees' competence depended on whom the interviewee mimicked. When the interviewer was rude to the interviewee, mimicking interviewees were rated as significantly less competent than non-mimicking interviewees. When the interviewer was cordial to the interviewee, mimicking
interviewees were rated as non-significantly more competent. Finally, when the interviewer was cropped out of the videos, thus preventing participants from noticing synchronous movements, all of the above effects disappeared. Despite these effects, participants showed a lack of conscious awareness of mimicry in debriefings.

Thus mimicry, when done in the wrong context, can negatively affect observers' judgments of our competence. But this research leaves open two important questions, which will addressed presently.

## How Complex are Inferences From Mimicry?

The above-discussed judgments by observers can be seen as relatively "sophisticated", as they took the attitude of the model (interviewer) towards the mimic (interviewee) into account. Observers clearly did not simply equate mimicry with competence. This seems to show that information gleaned from mimicry is integrated with broader social information in a subtle manner.
However, an alternative "non-intelligent" interpretation of these findings is that gestural mimicry simply enhances the perceived similarity between the interacting parties. As a result, negative attributes of the model (e.g., rudeness or cordiality) "rubbed off" on the mimic but not on the nonmimic. This can be explained as a relatively simple associationist inference that a target person who behaves like the model probably shares further traits with the model (Andersen, Moskowitz, Blair \& Nosek, 2007), or as a reflection of observers' belief that mimicry functions as a means of enhancing perceived similarity (Over \& Carpenter, 2012).

On the other hand, much research argues that mimicry itself is a complex, and even intelligent, process. Mimicry generation (despite its unconscious origins) depends on the context and social relationship between the mimicker and the model. People reduce their mimicry or even engage in anti-mimicry when interacting with a partner who is disliked, represents an out-group, or has different goals (Bourgeois \& Hess, 2008; Likowski, Muehlberger, Seibt, Pauli, \& Weyers, 2008; McIntosh, 2006; Stel et al., 2010). It would thus seem maladaptive for perceivers to interpret mimicry in a context-free manner.

## Mimicry, competence, and trust

It is also interesting and important to understand whether perception of traits other than competence can be influenced by perceived mimicry. Competence is one of the two main dimensions of social judgment (Judd et al., 1995). The other dimension is trust, which is critical to group cohesion, relationships, and most social transactions. Indeed, mimicry is posited to be part of the process of developing empathy and interpersonal trust (Bavelas, Black, Lemery, \& Mullett, 1987), and been called "social glue" (Lakin \& Chartrand, 2003). Thus, the connection between mimicry and trust is of obvious interest.

The link between mimicry and trust is particularly important in situations where an observer watches an interaction involving a person with a persuasive agenda. It has been shown, for example, that children imitate others more when they are attempting to persuade them to do unpleasant things, such as eat unappetizing foodstuffs, rather than enjoyable things (Thelen, Miller, Fehrenbach, Frautschi, \& Fishbein, 1980). Mimicry has also been shown to be an effective technique in adult negotiations (Maddux, Mullen, \& Galinsky, 2008). In short, it is important, and also novel, to explore how mimicry influences third-party inferences about trust.

## Present Study

The goals of the present study were twofold. First we examined whether third-party judgments about mimicry result from a simple assumption of trait similarity between similarly behaving individuals (i.e., the transference or "rub off" effect described above), or whether they instead reflect more nuanced social inferences. In particular, we tested whether observers' inferences about the target's trustworthiness reflected not only the presence/absence of mimicry between the target and the model, and the model's past trust-relevant behavior, but also, critically, the observer's knowledge of the target's epistemic state with regard to the model's past behavior. If mimicry inferences are indeed complex, observers should be sensitive to whether the mimicker "knows" about the model's reputation-related behavior. This is not unlike sensitivity that observers, even relatively young children, show to the epistemic state of an actor in "theory-of-mind" tests (Premack \& Premack, 1995).
Secondly, we explicitly attempt to show that mimicry can influence third-party observers' impressions of the trustworthiness of dyad members. We do this by directly influencing participants' impressions of the trustworthiness of one of the dyad members (the interviewer), and then testing whether mimicry (and the mimic's epistemic state) moderates the extent to which this reputation carries over to the other member (the interviewee).
The current paradigm was based on the previous work on third-party observation (Kavanagh et al. 2011), with some important changes. Participants again observed interviews and made social judgments about interviewees. Interviewees either mimicked or did not mimic their interviewer (model). Additionally, in the current study, participants also had to play an economic "trust" game (Berg, Dickhaut, \& McCabe, 1995) with interviewees.

Critically, in the current study, the interviewer exhibited the same neutral behavior in all videos - that is, she was not directly cordial or directly rude to the target, as in Kavanagh et al. (2011). Rather, subjects' perceptions of the interviewer were manipulated by an experimental confederate (also posing as a subject) who relayed a story. The interviewer was depicted as trustworthy to half of subjects, and as untrustworthy to the other half.

Participants' understanding of whether they and the interviewees had common knowledge of the interviewer was also manipulated. We did this by having some of the interviewees be present as "subjects" in the waiting room, along with the actual participant. There they also heard either the praising or defamatory story about the interviewer. Other interviewees were not among the subjects in the waiting room. Thus, the observers knew that some interviewees, having heard the story, were aware of the interviewer's high/low trustworthiness, but that other interviewees, having not heard the story, were not aware of the interviewer's high/low trustworthiness.
This step was motivated by the moral psychology literature, as well as philosophical and legal perspectives on responsibility more broadly, which suggest that observers take into account an agent's mental states (e.g., intent, deliberation, knowledge) when determining culpability for right/wrong actions (Suhler \& Churchland, 2009). Importantly, it seems that such considerations are relatively automatic (Young \& Saxe, 2009).
The current paradigm allows for a test of the sophisticated inference hypothesis of mimicry. If mimicry (similar movements) simply leads dyad members to be seen as similar, then observers should judge interviewees mimicking the trustworthy interviewer more favorably than interviewees mimicking the untrustworthy interviewer, regardless of the interviewee's state of knowledge about the interviewer's trustworthiness/untrustworthiness. However, the sophisticated inference account generates a more nuanced set of predictions in the context of trust-related situations.
The most straightforward prediction is that mimicking a trustworthy interviewer should benefit interviewees who are personally knowledgeable about his or her trustworthiness. This prediction should be offered with the caveat that work within the dyad has shown that some level of mimicry is expected in a normal face-to-face interaction (Dalton, Chartrand, \& Finkel, 2010) and so it may be that very strong mimicry would be required in order to "make a positive impression" on the viewer.

Figure 1: Experimental Conditions


The second and perhaps more interesting prediction is that mimicking an untrustworthy interviewer may benefit interviewees who are not knowledgeable about his or her misdeeds. After all, for the observer, when an innocent mimics an undesirable individual (i.e., shows affiliative behavior toward an undeserving party), the mimic should be seen as a particularly trusting (or naïve) and prosocial individual. All this should result a three-way interaction between mimicry (present/absent), trustworthiness (positive/negative), and knowledge (present/absent).

## Method

Participants and Procedure. 123 UCSD undergraduates participated for class credit. Upon arrival in the lab, subjects were greeted by the experimenter, who consulted a list and then told them that they would be in the rater condition and placed in a waiting room marked "rater condition." Another waiting room, clearly visible, was marked "interviewer condition" (this language was chosen to minimize ingroup effects by emphasizing situational assignment to the interviewee role). See Figure 1 for timeline (panel A) and spatial schematic (panel B). Several minutes before the supposed experimental start time, the confederate (henceforth "the gossip") who would be used to manipulate opinions about the trustworthiness of the interviewer was brought to the waiting room posing as another subject. Two other confederate "subjects" (confederate interviewees), who would eventually be transferred to the interviewee condition, entered the room one at a time right around the supposed experimental start time. This made their entrance (i) noticeable by subjects and (ii) minimized their chances of being engaged in conversation by subjects. After the last confederate arrived, the experimenter then fetched another confederate ("the interviewer"), introduced her to the subjects, and while the interviewer stood on, told the subjects (accurately) that they would see four interviews and then would play with the interviewees an economic game in which the trustworthiness of one's partner would be crucial to success. Subjects were also (deceptively) told that interviews would take place live in the interview room and that the video would be broadcast to computer screens in the rating rooms via the local intranet. After all these instructions, the experimenter left to take the interviewer to the interview room.

While the interviewer was gone, "the gossip" loudly told a story meant to either erode or build the participants' trust in the interviewer. In the praise condition the gossip recounted that the interviewer had driven to his home to return a lost wallet intact. In the defamation condition, participants were told that the interviewer was a friends' roommate who avoided paying all bills and shunned any communication.

After sufficient time, the interviewer returned and said that the experiment was ready to start but that two "subjects" currently in the rater condition would need to be transferred to the interviewee condition. The confederate interviewees were chosen for transfer and taken to the
interview room, while subjects and the gossip were escorted to small rooms equipped with a computer.

The computer portion of the experiment (see Figure 1, right panel) consisted of viewing and responding to 4 videos of interviews with 4 interviewees, all of whom were paired with the confederate interviewer. At the start of the computer portion, participants also viewed a detailed explanation of the Trust Game (see below). To back up the experimenter's cover story that the videos were in fact live, a screen of variable duration saying simply "waiting for live feed" in slowly flashing text was added to each video. Each subject now saw two interviewees (knowledgeable interviewees) who had heard the same story that they did, and two that did not (non-knowledgeable interviewees). Within each of these knowledge conditions, videos were arranged so that one member mimicked the interviewer and one did not. Mimicry was implemented as is standard in the literature, with neutral gestures (e.g., chin rubbing and leg-crossing) and no differences between verbal or meaningful nonverbal content.

Participants provided Likert scale (7-point) ratings for each interviewee on trustworthiness. As a behavioral measure, subjects also indicated how much money (from \$0 to 10) they would entrust with the target in the investor game (Berg et al., 1995). In this game, the participant is endowed with $\$ 10$ and can transfer some amount of this endowment to the target. This money is then tripled, and a partner could (if scrupulous) then return some money back to the participant. Participants also rated targets on control dimensions of competence, likability, and friendliness. Finally, as a manipulation check, participants also rated the interviewer herself on the same measures. Postexperimental funneled debriefing was also performed to check for conscious awareness of mimicry, belief that videos were live, and skepticism towards the waiting room manipulation (gossip).

Manipulation-Checks Analysis of debriefing questionnaires revealed that 20 participants were suspicious of the waiting room manipulation. A further 6 participants were excluded because they personally knew one of the confederates, and 1 participant was excluded because they noticed mimicry. The final sample thus consisted of 96 participants (59 females and 37 males) who were, on average, 20.23 years old ( $S D=1.58$ ).
We tested how participants perceived the interviewer as function of reputational manipulation (praise/defamation) and as a function of interacting with a mimicking or nonmimicking interviewee. Across all dependent measures, a 2 x 2 MANOVA revealed only a main effect of reputation, with no ME or interactions with mimicry on any measures. Specifically, in the Trust Game probe, participants were willing to give the interviewer more money in the praise condition $[M=4.8$ vs. $3.17, F(1,94)=10.30, p=.002]$. They also rated the interviewer as more trustworthy $[M=5.05 \mathrm{vs}$. 3.67, $F(1,94)=21.5, p<.001]$. In the praise condition, the interviewer was also rated as significantly more likeable
( $F=15$ ), competent ( $F=5.1$ ), and marginally more friendly ( $F=3.8$ ), all $p s<.05$. However, of the 4 ratings, the trust rating was particularly strongly influenced, as reflected in the 2-way interaction of reputation by rating type, $[F(3,94)$ $=5.9, p=.01]$. In short, our manipulation was very successful in changing the perception of the interviewer (model), with the effects particularly pronounced on the trust-related dimensions.

## Results

As described above, the "sophisticated inferences from mimicry" hypothesis predicts that evaluations made by third-party observers should vary as a function of the reputation of the model, presence or absence of mimicry by the target (interviewee), and the observer's knowledge about whether the target is aware of the reputation of the model (interviewer). Our central prediction was the observers would take the target's epistemic state into account. This should lead to the inferences that are more than simply a function of the goodness/badness of the model, as predicted by an associationist/rub-off account.

Because our reputational manipulation (praising vs. defamatory story) targeted the model's trustworthiness, we chose trustworthiness ratings, and monetary investment as our critical DVs. We analyzed these variables as a function of reputation (praise/defamation), knowledge (informed/ignorant), and mimicry (presence/absence), using a 3-way, mixed-models MANOVA.

We first focused on the trustworthiness rating (Figure 2) There were no significant main effects or interactions in the knowledgeable condition (all Fs < 1). Critically, in the ignorant condition, we found a 2-way interaction of reputation and mimicry $[F(1,93)=5.87, p=.02$. Simple effects (two-tailed) showed that ignorant participants who mimicked bad (untrustworthy) models were perceived as mo

re trustworthy than participants who did not mimic bad models ( $p<.06$ ), and as more trustworthy than ignorant participants who mimicked good (trustworthy) models ( $p<$ .10). Overall, this pattern resulted in a significant 3-way interaction $[F(1,93)=4.58, p=.03]$. We also found a main effect of knowledge, such that targets who witnessed the reputational manipulation were rated as more trustworthy
than targets who did not $[F(1,93)=11, p<.01]$. This effect (also seen on other ratings, as discussed below) presumably reflects that participants were personally familiar with targets in the knowledge, but not the ignorant, condition.

On the investment measure of trust, we also found a 2way interaction of reputation and mimicry $[F(1,94)=5.41$, $p=.02$ ]. Simple effects (two-tailed) showed that when ignorant participants mimicked bad models, they were perceived as marginally more trustworthy than participants who did not mimic bad models ( $p<.12$ ), and as more trustworthy than ignorant participants who mimicked good models ( $p<.01$ ). There were no significant effects in the "knowledgeable" condition (all Fs < 1). Overall, this pattern resulted in a 3-way interaction $[F(1,94)=3.42$, $p=.07]$. No other main effects or interactions were reliable.


Other ratings. Similar analyses were conducted on other ratings that were not directly related to trust. Analyses revealed a main effect of knowledge, such that "knowledgeable" targets were rated as overall friendlier ( $p<$ .01 ) and more competent than "ignorant" targets ( $p<.05$ ). As mentioned above, this effect may be due to participants' personal familiarity with targets in the knowledge condition (due to having spent time with them in the waiting room). Critically, none of the other ratings showed the 3-way interaction involving mimicry, reputation and knowledge (Fs $<1$ ). This suggests that observers’ inferences were restricted to the relevant trust-related dimensions that were relevant to our praising/defaming story manipulation.

## Discussion

Our central question was whether third-party observers' inferences from mimicry are simple or sophisticated. Thus, we tested whether such inferences take into account not only the presence or absence of mimicry, but also the reputation of the model and the target's knowledge about the model. We examined these third-party inferences of mimicry in the context of morality-related judgments of trust - an important social dimension.
Overall, our results support the idea that inferences made about third-party dyad members on the basis of observed mimicry are nuanced. Specifically, participants' judgments of trustworthiness reflected (i) whether the target mimicked,
(ii) the reputation of the person they mimicked, and also (iii) whether the target was aware of their model's reputation. As such, the results speak against the hypothesis that third party judgments of mimics reflect simple "rub-off", where the mimic is merely assigned traits of the model. More generally, our results suggest that people can integrate their perception of rapport with higher level social information. Critically, this higher level information includes knowledge about the target's epistemic state.
The observed 3-way interaction was largely driven by the fact that ignorant interviewees benefited from mimicking the untrustworthy interviewer. As we suggested earlier, this could reflect observers' perception that mimicry is an affiliative, courteous social behavior, so extending it, unknowingly, to an undeserving party is demonstrating a particularly trusting, perhaps gullible personality disposition. However, we feel this explanation should be tested explicitly in future extensions.
Another question for future research is why knowledgeable targets were not seen as more trustworthy after mimicking a trustworthy model. Targets were also not "blamed" for mimicking a model they knew to be bad. One possibility, consistent with the trust literature is that people do not give moral credit or blame for social courtesy (mimicry) when such courteous behavior is expected (Wojciszke, 2005). As mentioned, past work shows that some level of mimicry is expected in a normal social interaction (Dalton et al., 2003).
Another possibility is that mimicry effects disappeared in the knowledgeable condition because the targets had previously been personally encountered in the "waiting room" portion of the experiment. As mentioned, this brief acquaintance non-specifically enhanced several judgments, including trust, friendliness, and competence. Familiarity may also have caused these targets actions' to be judged in a more "charitable" manner. It may also be related to the fact that people make more situational inferences about acquaintances and more dispositional ones about strangers.
Also noteworthy is that, as expected, there was no decrement in competence ratings when targets mimicked a "defamed" interviewer. We believe this is because, unlike previous work (Kavanagh et al, 2011) interviewers were never directly condescending to the interviewee in this study. This contextualizes previous work, in which interviewees who mimicked to interviewers who condescended directly to them were seen as less competent. This provides further evidence that inferences from mimicry are subtle and situated. Though the full extent and nature of such inferences is not entirely clear, the results are the first to point towards the integration of unconsciously processed embodied signals with epistemic (e.g. Theory of Mind) knowledge.

## References

Andersen, S.M., Moskowitz, G.B., Blair, I.V, Nosek, B.A. 2007. Automatic thought. In E.T. Higgins \& A.W. Kruglanski (Eds.), Social Psychology: Handbook of Basic Principles (2 ${ }^{\text {nd }}$ Ed.) (pp. 138-75). New York, NY: Guilford Press.

Baird, J. A., \& Astington, J. W. (2004). The role of mental state understanding in the development of moral cognition and moral action. New Directions for Child and Adolescent Development, 103, 37-49.
Bavelas, J. B., Black, A., Lemery, C. R., \& Mullett, J. (1987). Motor mimicry as primitive empathy. In N. Eisenberg \& J. Strayer (Eds.), Empathy and its development (pp. 317-338). Cambridge, UK: Cambridge University Press.
Berg, J., Dickhaut, J. \& McCabe, K. (1995) Trust, reciprocity, and social history. Games and Economic Behavior, 10, 122-42.
Bernieri, F. (1988). Coordinated movement and rapport in teacher-student interactions. Journal of Nonverbal Behavior, 12, 120-138.
Bourgeois, P., \& Hess, U. (2008). The impact of social context on mimicry. Biological Psychology, 77, 343352.

Chartrand, T. L., \& Bargh, J. A. (1999). The chameleon effect: The perception-behavior
link and social interaction. Journal of Personality and Social Psychology, 76, 893-910.
Chartrand, T. L., \& van Baaren, R. B. (2009). Human mimicry. Advances in Experimental Social Psychology, 41, 219-274.
Churchland, P. (2011). Braintrust: What neuroscience tells us about morality. Princeton, NJ: Princeton University Press.
Heyes, C. M. (2011) Automatic imitation. Psychological Bulletin, 137, 463-483.
Judd, C. M., James-Hawkins, L., Yzerbyt, V., \& Kashima, Y. (2005). Fundamental dimensions of social judgment: Understanding the relations between judgments of competence and warmth. Journal of Personality and Social Psychology, 89, 899-913.
Kashima, Y. (2008). A social psychology of cultural dynamics: How cultures are formed, maintained, and transformed. Social and Personality Psychology Compass, 2, 107-120.
Kavanagh, L., Suhler, C., Churchland, P., \& Winkielman, P. (2011). When it's an error to mirror: The surprising reputational costs of mimicry. Psychological Science, 22, 1274-1276.
Kitayama, S., \& Cohen, D. (2007). Handbook of cultural psychology. New York, NY: Guilford Press.
Lakin, J. L., \& Chartrand, T. L. (2003). Using nonconscious behavioral mimicry to create affiliation and rapport. Psychological Science, 14, 334-339.
Lakin, J. L., Jefferis, V. E., Cheng, C. M., \& Chartrand, T. L. (2003). The Chameleon effect as social glue: Evidence for the evolutionary significance of
nonconscious mimicry. Journal of Nonverbal Behavior, 27, 145-162.
Likowski, K. U., Muehlberger, A., Seibt, B., Pauli, P., \& Weyers, P. (2008). Modulation of facial mimicry by attitudes. Journal of Experimental Social Psychology, 44, 1065-1072.
Maddux, W. W., Mullen, E., Galinsky, A. D. (2008). Chameleons bake bigger pies and take bigger pieces: Strategic behavioral mimicry facilitates negotiation outcomes. Journal of Experimental Social Psychology, 44, 461-468
McIntosh, D. N. (2006). Spontaneous facial mimicry, liking and emotional contagion. Polish Psychological Bulletin, 37, 31-42.
Niedenthal, P. M., Barsalou, L., Winkielman, P., KrauthGruber, S., \& Ric, F. (2005). Embodiment in Attitudes, Social Perception, and Emotion. Personality and Social Psychology Review, 9, 184-211
Over, H., \& Carpenter, M. (2012). Putting the social into social learning: Explaining both selectivity and fidelity in children's copying behavior. Journal of Comparative Psychology, 126, 182-192
Premack, D. and Premack, A.J. (1995) Origins of human social competence. In The cognitive neurosciences (Gazzaniga, M.S., ed.), pp. 205-218, MIT Press
Preston, S. D., \& de Waal, F. B. M. (2002). Empathy: Its ultimate and proximate bases. Behavioral and Brain Sciences, 25, 1-71.
Stel, M., Blascovich, J., McCall, C., Mastop, J., van Baaren, R. B., \& Vonk, R. (2010). Mimicking disliked others: Effects of a priori liking on the mimicry-liking link. European Journal of Social Psychology, 40, 867880.

Suhler, C.L. and Churchland, P.S. (2009). Control: conscious and otherwise. Trends in Cognitive Sciences, 13, 341-347
Thelen, M. H., Miller, D. J., Fehrenbach, P. A., Frautschi, N. M., \& Fishbein, M. D. (1980). Imitation during play as a means of social influence. Child Development, 51, 918-920.
Tomasello, M. (1999). The human adaptation for culture. Annual Review of Anthropology, 28, 509-529.
Wojciszke, B. (2005). Morality and competence in person and self-perception. European Review of Social Psychology, 16, 155-188.
Young, L., Saxe, R. (2009). An fMRI investigation of spontaneous mental state inference fo r moral judgment. Journal of Cognitive Neuroscience, 21, 1396-1405.

# Paradoxical Cultural Categories and Paradoxical Social Relationships: The Case of Cooperation and Competition 

Josh Keller (jwkeller@ntu.edu.sg)<br>Division of Strategy, Management \& Organisation, Nanyang Technological University<br>50 Nanyang Ave. S3-B2C-97, Singapore 639798<br>Jeffrey Loewenstein (jloew@illinois.edu)<br>College of Business, University of Illinois, Urbana-Champaign<br>1206 South Sixth Street MC-707 Champaign, IL 61820 USA<br>Martin Kilduff (m.kilduff@ucl.ac.uk)<br>Department of Management Science and Innovation, University College London<br>Gower Street London WC1E 6BT United Kingdom<br>Jin Yan (yanjin@zju.edu.cn)<br>School of Management, Zhejiang University<br>Gudun Road, Hangzhou, Zhejiang 310058 China


#### Abstract

How individuals think about opposing or paradoxical categories influences their social relationships. We found that Chinese managers were more likely than US managers to categorize attempts to outperform others as an instance of both competition and cooperation. Further, the Chinese managers were more likely than the US managers to perceive a given working relationship as being both cooperative and competitive. The two findings were linked: culturally-guided beliefs about whether the cooperation-competition paradox should be integrated or kept separate influenced how individuals understood their social relationships. More broadly, the implication is that category membership and relations between categories are guided by cultural influences distinct from the particulars of the categories themselves that normally enter into cognitive science research on categories. In addition, those categorization choices are consequential for the network of social relationships individuals form.


Keywords: Categories; paradox; cooperation; competition; culture; relationships; China.

## Introduction

"[O]ur two countries gain far more when we cooperate with one another than when we descend into an unhealthy competition."

Hillary Clinton, US Secretary of State, Beijing, September 5, 2012, at a joint press conference with Chinese Foreign Minister Yang Jiechi.

Choices to engage in cooperation and competition are fundamental to a wide range of social life, ranging from diplomacy between nations down to working relationships between individuals. Actors form competitive relationships as they seek to maximize their own outcomes and form cooperative relationships as they seek to achieve group goals. Further, most actors, most of the time, have mixed motives-they are concerned with both their individual
outcomes and their group's outcomes. Yet it is not clear whether and why actors might choose to engage in both cooperation and competition.

We will suggest that categories play a key role in the choice to engage in both cooperation and competition. As a result, we raise new issues in the study of culture, categories, and complex social relationships. The specific account that we develop centers on what we term paradoxical categorization, or the classification of a single situation as a member of both of two opposing categories. In our case, the paradoxical categorization of interest is the classification of a situation as both an instance of cooperation and an instance of competition. We show that culture influences whether individuals engage in paradoxical categorization. Then we show that paradoxical categorization predicts whether managers have working relationships that are both cooperative and competitive.

## Paradoxical Cultural Categories

Multiple streams of work are now challenging longstanding assumptions about the relation between cooperation and competition, and they are converging to make the joint use of cooperation and competition an important question. One such longstanding assumption in research on cooperation and competition, also implicit in the quote from Secretary Clinton, is that cooperation and competition are separate. Cooperation and competition have long been defined as mutually exclusive types of relationship (Deutsch, 1949), mutually exclusive types of behavior (Komorita \& Parks, 1996), and mutually exclusive types of motivation (McClintock \& Allison, 1989). However, there are now multiple proposals about why cooperation and competition could be integrated (e.g., Brandenberger \& Nalebuff, 1996; Van de Vliert, 1999), suggesting that cooperation and competition can co-occur.

Another longstanding assumption in research on cooperation and competition (Fulop, 2004), also implicit in
the quote from Secretary Clinton, is that cooperation and competition are the same for everyone. However, it is now clear that, for example, individuals in the United States and China view cooperation differently (Keller \& Loewenstein, 2011), and that individuals in Hungary and Japan view competition differently (Fulop, 2004). There are also strong theoretical arguments suggesting that the relation between cooperation and competition likely differs across cultures (Chen, 2008): Western cultural philosophies (e.g., US, UK, Australia) seem to emphasize the separation of cooperation and competition and East Asian cultural philosophies (e.g., China, Japan, Korea) seem to emphasize the integration of cooperation and competition. Accordingly, culturallyguided beliefs may affect when and why individuals choose to engage in both cooperation and competition.

A third longstanding assumption in research in cooperation and competition (Stanne, Johnson \& Johnson, 1999), but that Secretary Clinton's quote rejects, is that there is only one kind of competition. Instead, there appear to be distinct consequences to healthy or appropriate competition, such as the attempt to outperform others, and unhealthy or zero-sum competition, such as the attempt to sabotage others (Stanne et al., 1999). Different kinds of competition may be differently compatible with cooperation. That is, the overall semantic relation between cooperation and competition may be antonymic (Herrmann, Conti, Peters, Robbins, \& Chaffin, 1979), as noted in both American (Merriam-Webster, 2006) and Chinese (He, 2009) thesauruses. However, even if the categories as a whole are antonyms, it is an open question as to whether the two categories may still overlap and share members.

We generate a new account of the relation between cooperation and competition consistent with the three new possibilities just discussed. Our starting point is to conceptualize cooperation and competition as cultural categories (Atran, Medin \& Ross, 2005; Douglas, 1986; Keller \& Loewenstein, 2011). Through social interactions, people learn the conventions in their culture (Millikan 2005) for categorizing interpersonal situations and relationships as cooperative and as competitive. The question then is why an individual might categorize an item as being both cooperative and competitive. Two influences seem key: beliefs about paradoxes and contradictions, and the type of interpersonal situation.

There is ample evidence that individuals who are members of Chinese culture are more likely than members of American culture to hold dialectical beliefs (SpencerRodgers et al, 2010), meaning they tend to tolerate contradictions, expect change, and seek to integrate paradoxes. One consequence is that Chinese individuals tend to be more likely than American individuals to engage in paradoxical categorization. For example, they are more likely to categorize themselves as both shy and outgoing (Spencer-Rodgers, Boucher, Mori, Wang \& Peng, 2009) and as both happy and sad (Bagozzi, Wong \& Yi, 1999).

These general tendencies should apply to cooperation and competition. To be clear, we are not claiming that a general
tendency towards dialecticism, derived from one's culture, predicts a willingness to believe that any competitive situation is also a cooperative situation, or even more starkly, that Chinese individuals always engage in both cooperation and competition and American individuals never do. Rather, we are suggesting that dialecticism licenses individuals to grant that cooperation and competition could co-occur. Specifically, a general tendency towards dialecticism, derived from one's culture, should predict an individual's willingness to categorize a seemingly contradictory situation with features of both cooperation and competition as being both cooperative and competitive, rather than being forced to pick one or the other.

The key situations with features of both cooperation and competition are acts of healthy competition, such as attempts to outperform another person. If attempts to outperform others are interpreted as efforts to gain higher relative standing (a key feature of competitive behavior, Johnson \& Johnson, 1989), these efforts could be classified as competitive. If attempts to outperform others are also seen as efforts to advance group gains (a key feature of cooperative behavior, Tyler \& Blader, 2000) then they have the potential to be classified not only as competitive but also as cooperative. By contrast, acts of unhealthy competition, such as attempts to sabotage another person, are unlikely to be seen as incorporating any feature of cooperation (they lower group gains; Stanne et al, 1999), but are likely to be seen as competitive (they are efforts to gain higher relative standing). Thus, paradoxical categorization could occur for attempts to outperform others but is unlikely for attempts to sabotage others.

Taken together with the prior point about culture, the full prediction is that because individuals who are members of East Asian cultures are more likely than individuals from Western cultures to hold dialectical beliefs, they should be more likely to generate the paradoxical categorization that attempts to outperform others are acts of both cooperation and competition.

## Paradoxical Social Relationships

Most research on social relationships has described a stark choice between cooperative colleagues giving each other advice versus rivals battling to get ahead (e.g., Burt, 1987). Yet just as researchers examining the same data can radically disagree concerning whether cooperation or competition represents the best explanation of observed patterns (Kilduff \& Oh, 2006), individuals also sometimes struggle to comprehend the meaning of their colleagues' actions. We see people inventing terms like "coopetition" and "frenemies" to account for such complex social relationships.

Individuals are embedded in networks of cooperative working relationships as they collaborate with others. But people are also embedded in networks of competitive relationships as they vie for status and resources (Burt, 1992; Lazega \& Patterson, 1999). Because social relationships are complex (Ingram \& Zou, 2008), an
individual could have relationships that are both cooperative and competitive. When are relationships likely to be recognized as both cooperative and competitive? Two concerns seem to be key: the frequency of interaction and paradoxical categorization.

We focus on managers' working relationships, as a subset of social relationships. In managerial work contexts, frequency of interaction typically implies that individuals engage in reciprocal patterns of sharing knowledge (McAllister, 1995). In addition, managers who work together frequently are also more likely to have their performance compared (Brown et al, 1998) and to contend for resources (Burt, 1992). Thus, working together frequently is likely to provide the opportunity for individuals to experience and to reciprocate acts of cooperation and acts of competition.

The cycles of reciprocated behaviors that individuals experience in their social relationships should guide how they interpret those relationships (Gouldner, 1960; Koster \& Sanders, 2006). So, for example, if individuals experience others sharing knowledge, they may interpret those acts as cooperation and reciprocate with cooperative behaviors of their own, leading them to characterize their relationship as cooperative. Accordingly, if managers' working relationships involve frequent contact, then this should provide the potential for developing relationships that are both cooperative and competitive.

Frequent interaction only provides the potential for forming working relationships that are both cooperative and competitive because individuals might tend to reciprocate mainly one as opposed to both kinds of behavior. Consistent with our earlier arguments, we suggest that individuals from different cultural groups and with differing cultural categories should differ in how they resolve the paradoxical tension (Miron-Spektor, Gino, \& Argote, 2011; Smith \& Lewis, 2011) of encountering opportunities for, or behaviors indicating, both cooperation and competition.

If Chinese individuals are more likely than American individuals to categorize attempts to outperform others as instances of both cooperation and competition, then this may indicate a more general willingness to integrate and reciprocate both cooperation and competition. That is, Chinese individuals may be more likely than American individuals to experience someone attempting to outperform them, perceive it as cooperative and competitive, and reciprocate with acts of cooperation as well as acts of (presumably healthy) competition. In contrast, American individuals may be more likely than Chinese individuals to experience someone attempting to outperform them, perceive it as competitive and not cooperative, and reciprocate with acts of competition and non-cooperation. The end result is a difference in the frequency of experiencing both cooperation and competition within the same working relationship. Thus, paradoxical categorization should predict whether, for those working relationships with frequent interaction that allow for developed chains of
reciprocity, individuals are likely to characterize those working relationships as both cooperative and competitive.

In the study that follows, we examined Chinese and American managers for their beliefs about the paradoxical categorization of cooperation and competition. A week later, we gathered their evaluations of their working relationships. We expected that Chinese managers would be more likely than American managers to endorse paradoxical categorization and to characterize their frequent working relationships as both cooperative and competitive.

## Methods

## Participants

A total of 111 managers in the United States and 139 managers in China participated in the study. The American managers were, on average, 29 years old and the Chinese managers were about 31 years old. The American managers ( $76 \%$ ) and the Chinese managers ( $63 \%$ ) tended to be male. All participants had earned college degrees and had at least three years of full-time work experience. Within each sample, each major industry, including technology, services, and manufacturing, was represented. Participation in the study was voluntary.

## Procedure and Materials

Time 1 Survey Participants listed up to 24 people within their organization with whom they had an ongoing working relationship (as in, for example, Chua, Ingram, \& Morris, 2008). Participants then completed a categorization task, as described below. Finally, participants provided demographic information about themselves and information about their organization.

Categorization task The categorization task followed protocols developed within cognitive anthropology (see Weller, 2007 for a review). Using a separate sample from the main study, we asked 40 participants from China and 40 participants from the United States to describe situations that indicated competition. We used existing data on cooperation from Keller and Loewenstein (2011). We created 25 items describing situations that were mentioned by members of both cultures as either cooperative or competitive. All items were in Chinese in China and in English in the US. To ensure language equivalence, we engaged in a coding, translating and back-translating process by coders not informed about the purposes of the study (Brislin, 1970).

The key items concerned outperforming (5 items), sabotaging ( 4 items) and knowledge sharing ( 2 items). We included an additional 14 filler items to reduce demand effects. The 25 situations were presented to participants three separate times; once each for whether the situation could be categorized as cooperation, as competition and as commitment (to provide a filler between the cooperation and competition categorization tasks). Half the participants rated situations for cooperation first and competition third,

Table 1: Categorization of situations as cooperation and competition

|  |  | Knowledge Sharing | Sabotaging | Outperforming |
| :--- | :--- | :---: | :---: | :---: |
| USA | Cooperation | $3.83(1.40)^{*}$ | $1.53(0.55)^{*}$ | $3.10(0.73)$ |
|  | Competition | $2.25(1.32)^{*}$ | $4.22(0.70) *$ | $4.16(0.42) *$ |
| China | Cooperation | $3.98(1.18)^{*}$ | $1.61(0.51)^{*}$ | $3.62(0.55) *$ |
|  | Competition | $2.01(1.01)^{*}$ | $4.12(0.64) *$ | $4.26(0.46) *$ |

Standard deviations are in parentheses.

* $\mathrm{p}<.01$ from one-way t -tests ( $\min \mathrm{df}=110, \min \mathrm{t}=13$ ) for differences from 3, with above 3 indicating cooperation or competition, and below 3 indicating non-cooperation or non-competition.
and half rated situations in the reverse order. We found no effects of order of presentation.

Time 2 Survey One week later, participants evaluated each working relationship they had listed on the Time 1 survey. They rated the level of competition, cooperation, and the frequency with which they worked together, as well as other information beyond the scope of the current paper. The order of presentation of the questions about cooperation and competition was counterbalanced, and we found no effects of the order of presentation.

## Measures

Categorization Participants rated knowledge sharing, sabotaging, and outperforming situations twice on scales from $1=$ non-cooperative/ non-competitive and $5=$ cooperative/ competitive.

Paradoxical categorization We used participants' ratings of how cooperative outperforming situations were as a measure of paradoxical categorization. We found similar patterns if we use measures based on their ratings of both cooperation and competition.

Frequent interaction Working relationships with "at least daily" interaction were coded as a working relationship with frequent interaction.

Paradoxical working relationships Participants rated each working relationship on a 5-point scale for cooperation ( $1=$ very non-cooperative, $2=$ slightly non-cooperative, $3=$ neither cooperative nor non-cooperative, $4=$ slightly cooperative, and $5=$ very cooperative) and a similar scale for competition. A working relationship that was rated a 4 or 5 on both the "cooperative" and "competitive" scales was coded as a working relationship that had both cooperation and competition.

Number of working relationships Participants could have reported up to 24 working relationships and we included the number they listed as a control variable.

Demographic Variables Age and gender were included as control variables because they commonly influence interactions within organizations.

## Results

## Categorization data

Table 1 shows the means and standard deviations for the categorization of each situation type (knowledge sharing, sabotaging and outperforming) for respondents from the US and China. As expected, Chinese managers and American managers categorized knowledge sharing as cooperative and non-competitive and sabotaging situations as competitive and non-cooperative. Both Chinese and American managers categorized outperforming situations as competitive. Finally and most critically, Chinese managers categorized outperforming situations as cooperative $(M=3.62, S D=$ 0.55 ) whereas American managers did not $(M=3.10, S D=$ $.73), t(249)=6.35, p<.01$. Thus, Chinese managers showed greater willingness than American managers to engage in paradoxical categorization.

## Working relationship data

Table 2 reports hierarchical non-linear logistic regression models predicting paradoxical working relationships. We found no effects of gender, age, and number of ties (second level control variables) or cultural group (a second-level variable). As expected, frequent interaction predicted paradoxical working relationships (a first-level variable; $\mathrm{B}=$ $.77, \mathrm{SE}=.15$ ). Also as expected, there was an interaction between cultural group and frequent interaction ( $\mathrm{B}=.73$, SE $=.30)$, as Chinese managers $(M=.20)$ reported that more of their frequent interaction relationships were paradoxical working relationships than did American managers ( $M=$ .14), $t(2342)=3.38, p<.01$.

Paradoxical categorization helped to explain the effect of cultural group. Paradoxical categorization predicted paradoxical working relationships (a second-level variable; $\mathrm{B}=.05$, $\mathrm{SE}=.01$ ). When including paradoxical categorization with cultural group, frequent interaction and the interaction of cultural group and frequent interaction, the effect of the interaction was still significant yet reduced (B $=.51, \mathrm{SE}=.30$ ). A bootstrapped test of an indirect effect of the interaction of cultural group and frequent interaction on

Table 2: Predictors of Paradoxical Working Relationships

|  | Controls |  | Cultural Group |  | Frequent Contact |  | Interaction |  | Paradoxical categorization |  | Full Model |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | B | SE | B | SE | B | SE | B | SE | B | SE | B | SE |
| Intercept | -2.18 | 0.61 | -2.03 | 0.60 | -2.21 | 0.06 | -2.25 | 0.65 | -0.04 | 0.09 | -0.33 | 0.18 |
| Gender (F) | 0.16 | 0.17 | 0.15 | 0.17 | 0.02 | 0.02 | 0.15 | 0.18 | 0.02 | 0.02 | 0.16 | 0.19 |
| Age | 0.02 | 0.02 | 0.01 | 0.02 | 0.01 | 0.02 | 0.01 | 0.02 | 0.00 | 0.00 | 0.01 | 0.02 |
| Number of Ties | 0.00 | 0.01 | -0.01 | 0.01 | 0.00 | 0.01 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.01 |
| Cultural Group |  |  | 0.16 | 0.16 |  |  | -0.37 | 0.26 |  |  | -0.52 | 0.31 |
| Frequent Interaction |  |  |  |  | 0.77 | 0.02 | 0.24 | 0.18 |  |  | 0.28 | 0.19 |
| Cultural Group <br> *Frequent Interaction |  |  |  |  |  |  | 0.73 | 0.30 |  |  | 0.51 | 0.30 |
| Paradoxical Categorization |  |  |  |  |  |  |  |  | 0.05 | 0.01 | 0.32 | 0.16 |

Note: bold if $\mathrm{p}<.05$.
paradoxical working relationships through paradoxical categorization found support, estimating an effect of 0.04
( $95 \%$ CI: 0.01-0.07). Therefore, the influence of Chinese culture on paradoxical categorization is linked to the particular likelihood of Chinese managers' having paradoxical working relationships among their frequent interaction partners.

## Discussion

Chinese managers, relative to American managers, were more likely to categorize outperforming situations as both cooperative and competitive, and in turn were more likely to describe working relationships with frequent interaction as both cooperative and competitive. Simultaneously cooperative and competitive working relationships were not randomly distributed rare occurrences. For those who categorized outperforming situations as both cooperative and competitive (a set of mostly Chinese and some American managers), the median manager worked every day with two people with whom they both cooperated and competed. Yet for those who did not categorize outperforming situations as both cooperative and competitive (a set of mostly American and some Chinese managers), the median manager did not work with anyone with whom they both cooperated and competed. Thus, cultural support for paradoxical categorization, combined with enabling social situations, shape the social experience of cooperation and competition and, more broadly, opportunities for paradoxical working relationships.

Part of the account is about the influence of culture on categorization. Our account, drawing on prior literature, was that cultural philosophies can provide a basis for dialecticism. Dialectical beliefs then enable paradoxical categorization. In related research, we have documented the mediating role of dialectical beliefs in the link between cultural group membership and paradoxical categorization.

The robustness of the link between culture and paradoxical categorization is suggestive of the importance of studying culture to studying categories. The nature of the relation people perceive between categories, and category membership itself, is not just a function of the features or properties of the categories, their members, or the categories to which they are associated (e.g., Goldstone, 1996). Whether attempts to outperform others are instances of cooperation is ambiguous and appears to be resolved by principled social convention.

The more general implication is that attempts at studying category membership and relations among categories without considering cultural influences has the potential to be misleading. Part of individuals' understandings of categories-which tends not to be the focus of cognitive science research-is shaped by cultural use of the specific category and the culturally normative views about categories more generally. Studying artificial categories is wonderfully useful, as is studying concrete object categories that are fairly consistent across cultural communities. Cooperation and competition are not typical of the categories cognitive science researchers tend to study (they are relational categories; Gentner \& Kurtz, 2005). Yet cooperation and competition are arguably among the most frequently used categories in social life, and relational categories more generally account for much of our expert knowledge. The culturally-guided aspects of these categories' meanings are, therefore, highly consequential and so worthy topics of study.

Cultural factors shape categories because category use is so often social. In the current case, beliefs about categories are linked to perceptions of relationships. In other work, we also show that these beliefs about cooperation and competition predict behavior in a workgroup context. These perceptions and behaviors are consequential. People's choices are guided by how they perceive others, and those
perceptions can be self-reinforcing because of reciprocity. For example, a direct implication of the findings in this paper are that general cultural beliefs about paradoxes could, by shaping categories and relations between categories, shape the networks of social relationships that comprise our lives.

## References

Atran, S., Medin, D. L., \& Ross, N. O. 2005. The Cultural Mind: Environmental Decision Making and Cultural Modeling Within and Across Populations. Psychological Review, 112(4): 744-776.
Bagozzi, R. P., Wong, N., \& Yi, Y. 1999. The Role of Culture and Gender in the Relationship between Positive and Negative Affect. Cognition \& Emotion, 13(6): 641672.

Brandenburger, A., \& Nalebuff, B. 1996. Co-Opetition. New York: Doubleday.
Burt, R. S. 1987. Social Contagion and Innovation: Cohesion Versus Structural Equivalence. American Journal of Sociology, 92(6): 1287.
Burt, R. S. 1992. Structural holes: The social structure of competition. Cambridge, MA: Harvard University Press
Chen, M.-j. 2008. Reconceptualizing the CompetitionCooperation Relationship: A Transparadox Perspective. Journal of Management Inquiry, 17: 288.
Chua, R., Ingram, P., \& Morris, M. W. 2008. From the head and the heart: Locating cognition- and affect-based trust in managers' professional networks. Academy of Management Journal, 51: 436-452.
Deutsch, M. 1949. A theory of cooperation and competition. Human Relations, 2: 199-231.
Douglas, M. 1986. How institutions think. Syracuse, NY: Syracuse University Press.
Fulop, M. 2004. Competition as a culturally constructed concept. In C. Baillie, E. Dunn \& Y. Zheng (Eds.), Travelling facts. New York: Campus Verlag.
Gentner, D., \& Kurtz, K. (2005). Relational categories. In W. K. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman \& P. W. Wolff (Eds.), Categorization inside and outside the lab (pp. 151-175). Washington, DC: APA.
Goldstone, R. L. (1996). Isolated and Interrelated Concepts. Memory \& Cognition, 24, 608-628
Gouldner, A. W. 1960. The Norm of Reciprocity: A Preliminary Statement. American Sociological Review, 25(2): 161-178.
He, G. 2009. Xiandai Hanyu Fanyi Cidian: Shanghai Cishu Chubanshe.
Herrmann, D., G. Conti, et al. 1979. "Comprehension of antonymy and the generality of categorization models." Journal of Experimental Psychology: Human Learning and Memory, 5(6), 585-597.
Ingram, P., \& Zou, X. 2008. Business friendships. Research in Organizational Behavior, 28: 167-184.

Johnson, D. W., \& Johnson, R. 1989. Cooperation and competition: Theory and research. Edina, MN: Interaction book.
Keller, J., \& Loewenstein, J. 2011. The cultural category of cooperation: A cultural consensus model analysis for China and the United States. Organization Science, 22(2), 299-319.
Kilduff, M., \& Oh, H. 2006. Deconstructing diffusion: An ethnostatistical examination of medical innovation network data reanalyses Organizational Research Methods, 9: 432-455.
Komorita, S. S., \& Parks, C. D. 1996. Social Dilemmas. Boulder, CO: Westview.
Koster, F., \& Sanders, K. 2006. Organizational citizens or reciprocal relationships? An empirical comparison. Personnel Review, 35(5): 519-537.
Lazega, E., \& Pattison, P. E. 1999. Multiplexity, generalized exchange and cooperation in organizations: a case study. Social Networks, 21(1): 67-90.
McClintock, C. G., \& Allison, S. T. 1989. Social Value Orientation and Helping Behavior<sup>1</sup>. Journal of Applied Social Psychology, 19(4): 353-362.
Merriam-Webster. 2006. The Merriam-Webster Thesaurus: Merriam-Webster.
Millikan, R. G. 2005. Language: a biological model OUP.
Miron-Spektor, E., Gino, F., \& Argote, L. 2011. Paradoxical frames and creative sparks: Enhancing individual creativity through conflict and integration. Organizational Behavior and Human Decision Processes, 116(2): 229240.

Smith, W. K., \& Lewis, M. 2011. Toward a Theory of Paradox: A Dynamic Equilibrium Model of Organizing. Academy of Management Review, 36(2): 381-403.
Spencer-Rodgers, J., Boucher, H., Mori, S., Wang, L., \& Peng, K. 2009. The Dialectical self-concept: Contradiction, Change, and Holism in East Asian cultures. Personality and Social Psychology Bulletin, 35: 29-44.
Spencer-Rodgers, J., Williams, M. J., \& Peng, K. 2010. Cultural Differences in Expectations of Change and Tolerance for Contradiction: A Decade of Empirical Research. Personality \& Social Psychology Review, 14(3): 296-312.
Stanne, M. B., Johnson, D. W., \& Johnson, R. T. 1999. Does competition enhance or inhibit motor performance: A meta-analysis. Psychological Bulletin, 125(1): 133-154.
Tyler, T. R., \& Blader, S. L. 2000. Cooperation in groups : procedural justice, social identity, and behavioral engagement. Philadelphia, PA: Psychology Press.
Van de Vliert, E. 1999. Cooperation and competition as partners. European Review of Social Psychology, 10, 231-257.
Weller, S. C. 2007. Cultural consensus theory: Applications and frequently asked questions. Field Methods, 19(4) 339-368.

# Minimally Supervised Learning for Unconstrained Conceptual Property Extraction 

Colin Kelly (colin.kelly@cl.cam.ac.uk), Anna Korhonen (anna.korhonen@cl.cam.ac.uk)<br>Computer Laboratory, University of Cambridge<br>15 JJ Thomson Avenue, Cambridge, CB3 0FD, UK<br>Barry Devereux (barry@csl.psychol.cam.ac.uk)<br>Centre for Speech, Language, and the Brain, University of Cambridge Downing Street, Cambridge, CB2 3EB, UK


#### Abstract

We present a highly performant, minimally supervised system for the challenging task of unconstrained conceptual property extraction (e.g., banana is fruit, spoon used for eating). Our technique employs lightly supervised support vector machines to acquire promising features from our corpora (Wikipedia and UKWAC) and uses those features to anchor the search for plausible unconstrained relations in our corpus. We introduce a novel backing-off method to find the most likely relation for each concept/feature pair and produce a number of metrics which act as potential indicators of true relations, training our system using a stochastic search algorithm to find the optimal reweighting of these metrics. We also introduce a human semantic-similarity dataset; our output shows a strong correlation with human similarity judgements. Both our gold standard comparison and direct human evaluation results improve on those of previous approaches, with our human judgements evaluation showing a significant 20 percentage point performance increase.


## Introduction

Recent theories in cognitive psychology attest a propertybased, distributed and componential model of conceptual representation for concrete concepts (e.g., elephant, screwdriver) in the brain (Farah \& McClelland, 1991; Tyler, Moss, Durrant-Peatfield, \& Levy, 2000; Randall, Moss, Rodd, Greer, \& Tyler, 2004). To explore the validity of these theories, researchers employ real-world knowledge taken from property norming studies where human volunteers are asked to list properties for concepts. McRae, Cree, Seidenberg, and McNorgan (2005) performed the largest such study to date, collecting properties for over 500 concrete nouns (we call these the 'McRae norms'). Some example properties from these norms can be found in Table 1.

However, as has been widely discussed (Murphy, 2002; McRae et al., 2005), these studies suffer from a number of weaknesses. For example, human participants often underreport certain properties, even when they are facts presumably known by the volunteers: though all participants are likely to have known that animals have hearts, has heart is not reported as a property for any animal concept. Similarly, is animal is listed as a property of all animals in the norms while breathes is only cited as a property for whale. A related issue is inconsistency across similar concepts: has legs is listed as a property of leopard but is absent for tiger.

Our task is to automatically extract such conceptual representations from large text corpora using NLP techniques. We

Table 1: Top ten properties from McRae norms with production frequencies for knife and pig.

| knife | pig |  |  |
| :--- | :--- | :--- | :--- |
| is sharp | 29 | an animal | 21 |
| used for cutting | 25 | lives on farms | 20 |
| is dangerous | 14 | is pink | 20 |
| has a handle | 14 | has a tail | 17 |
| has a blade | 11 | has a curly tail | 15 |
| a weapon | 11 | has a snout | 12 |
| a utensil | 9 | eaten as bacon | 11 |
| made of steel | 8 | oinks | 9 |
| is serrated | 8 | is fat | 8 |
| found in kitchens | 8 | is dirty | 8 |

hope to extract features for a given concept as well as those features' relationship with that concept; specifically, we aim to extract properties in the form of concept relation feature triples (e.g., knife used for cutting, pig lives on farm), where both the relation and the feature are unconstrained. Our task is particularly challenging because while we seek a very specific 'type' of information (namely, conceptual properties), there is an enormous amount of variation across the features and relations of properties which exhibit such characteristics.

Previous approaches to our specific conceptual property extraction task (Baroni, Murphy, Barbu, \& Poesio, 2009; Devereux, Pilkington, Poibeau, \& Korhonen, 2009; Kelly, Devereux, \& Korhonen, 2010, 2012) have been successful to varying degrees, however each has suffered from limitations. Baroni et al., for example, did not explicitly offer relations between their extracted concepts and features. The relations extracted by the Devereux et al., system were rather unsophisticated, with the relation corresponding to the verb found along the grammatical relation path linking concept to feature. The Kelly et al. (2010) system had reasonable performance but was founded on manually constructed rules and relied heavily on WordNet for its feature selection.

The system of Kelly et al. (2012) approached this task as one of relation classification. The relations generated were derived directly from its training set; it was therefore unable to posit new or unseen relationships between its extracted concepts and features. We believe their feature output, however, was promising and we extend and enhance their feature extraction method in the first component of our own system.

Our system works by first employing a wealth of lexical, syntactic and semantic machine-learning attributes to train a support vector machine for feature-extraction. Unlike other
approaches, we make heavy use of unlabelled training data, rendering our system only very lightly supervised. Next, we return to our unlabelled corpus to find relations for the extracted features, using a novel, probabilistically motivated backing-off technique. In doing so, we are not constrained by relations found in the McRae norms: our method allows for the extraction of any relation.

## Data

## Recoded norms

We used the same set of recoded norms employed by Kelly et al. (2012) to train our system. This set, containing 510 concepts in total, is a coding of an anglicised version ${ }^{1}$ of the McRae norms into a uniform concept relation feature format, where each feature and concept contain one word; the relation slot can contain one or more words.

## Corpora

We used Wikipedia and the more general UKWAC corpus (Ferraresi, Zanchetta, Baroni, \& Bernardini, 2008), containing English-language webpages, as corpora. Together these offered a suitable balance of general and encyclopaedic text. We used the C\&C-parser (Clark \& Curran, 2007) to extract grammatical relations (GRs) and part of speech (POS) information from sentences, allowing us to construct a GR-POS graph for each. We trained our system on the corpora individually and in combination.

## Chunking

We also used chunked versions of our two corpora. Chunking is a technique which identifies the constituent blocks of a sentence (verb phrase, noun phrase, prepositional phrase, etc.). To chunk our corpora, we used the Apache OpenNLP 1.5 suite (Baldridge, 2005), using the Tokenizer, POS Tagger and Chunker tools. The various components of the suite were trained using models supplied with the OpenNLP package.

## Method

We trained our system with 466 of the 510 concepts in the anglicised McRae set to fix our training parameters and evaluated with the remaining 44 concepts, those in the ESSLLI expansion set (Baroni, Evert, \& Lenci, 2008) (discussed later).

## Feature derivation

In the first stage we focussed on extracting terms relevant to our concepts in order to generate a promising set of features, similar to those found in our norms.

Machine learning attributes Support vector machines (SVMs) are non-probabilistic binary linear classifiers which take a set of input data and predict, for each given input, which of two possible classes it corresponds to. This works by plotting training data points in a high-dimensional space and separating them with a hyperplane which has the largest

[^109]distance (or margin) to the nearest training data points of each class. This plane is subsequently used to classify unseen data points. SVMs can also be extended to the multi-class case.

We trained an SVM by constructing paths through each sentence's GR-POS graph from the concept to prospective features and used the GR path labels, POS tags, relation verb instances and path-length as machine learning attributes. We augmented this (mostly syntactic) set of machine learning attributes to incorporate additional semantic and lexical information: bigrams and concept/feature clusters. ${ }^{2}$ The intuition behind this was that similar types of concepts/features (as exhibited by cluster membership) might also exhibit similar types of relationships (e.g., 'tool' concepts and used for relations); the aim was to enable the SVM to detect the regularities that exist in the relationships between different semantic classes of concepts and features.

Every possible attribute across the training set corresponded to a distinct dimension of the vector space. The majority of the co-ordinates of the training data points took binary values depending on whether the dimension's corresponding attribute appeared in the path (except the clustering and path-length attributes which took integer values). Each training data point was labelled with its relation (or 'class').
Learning instances We applied the SVM Light software ${ }^{3}$ (Joachims, 1999) to our learning attributes to extract an SVM score (the sum of absolute values of the decision function values, which can be interpreted as a measure of confidence of the SVM in its classification) for each concept-feature pair. We also calculated log-likelihood (LL) (Dunning, 1993) and pointwise mutual information (PMI) (Church \& Hanks, 1990) statistics across the top 200 returned concept-feature pairs for each concept.

Previous work has ignored a large amount of potentially instructive training data by only examining sentences which link entities explicitly found in the training set. However, the use of 'negative' information could prove informative and therefore we trained on all GR-POS paths linking one of our concepts to any potential feature term ${ }^{4}$ in each sentence. The size of our training set was 5.52 million instances for the Wikipedia corpus and 20.07 million instances for the UKWAC corpus. ${ }^{5}$ As we were unaware of the nature of the relationship between these concept/feature terms, we labelled these unknown training paths as unknownrel.

Our system was therefore only very lightly supervised: only $6.8 \%$ of the UKWAC input and $8.7 \%$ of the Wikipedia input to the system was labelled with relations drawn from the McRae norms. Consequently, our SVM classified every

[^110]concept/feature pair into the unknownrel relation class. We therefore ignored the relation output from this stage of the system, instead using the top 200 returned concept/feature pairs ranked by their SVM scores as input to the next stage. In this way, we were interpreting a higher-rated SVM score as a proxy for the likelihood that a feature would have some kind of relationship with the concept at hand.

## Relation extraction

The underlying hypothesis of our relation extraction stage was that if we found sequences of chunks in our corpus sentences which were anchored at each end by a known concept and feature (from the previous stage), and those chunks' labels matched the labels of our chunked property norms, then we could use the surface text of the chunk(s) between the anchors as the relation in our concept relation feature format.
Chunk pattern selection To decide which patterns of chunks were likely to be indicative of property norm relations, we turned to our training set. We passed the full text of the non-ESSLLI McRae norms through the chunker, and manually examined the output for chunk label patterns likely to indicate relations.

Using this output, we created a ruleset for selecting sentence fragments (chunk sequences) which were similar in structure to our property norms. We called a sequence of three labelled chunks a three-chunk, a sequence of four chunks a four-chunk, etc. We employed the first four most frequent label combinations (NP VP NP; NP VP PP NP; NP VP ADJP; and, NP VP ADVP) to form our ruleset; together these covered $95.6 \%$ of the three- and four-chunk label patterns generated from our training set. By using the NP VP PP NP-labelled four-chunks we were able to extract multi-word, prepositional verbs (e.g., worn on, used for) as potential relations: previous approaches to our task have not attempted this.

Chunk pre-selection We needed to select those chunks most relevant to our relation extraction task. To do this we passed through our chunked corpus, generating sets of 3 and 4 sequential chunks and pre-selecting those which were relevant to our concepts. Our criterion for relevancy at this stage was that the final term contained within the first chunk, when lemmatised, corresponded to a training concept.

Chunk to triple conversion Having pre-selected our chunks we generated triples from the chunk text. For threechunks we did this by simply taking the final term in the first, second and third chunks and lemmatising each to give our concept, relation and feature terms respectively. For fourchunks we followed the same process for the first and fourth chunks to yield our concept and feature. To extract the relation we took the final term of the second (VP) chunk and compounded it with the final term of the third (PP) chunk; the only exception to this was if the POS of the final term of the second chunk was VBG, in which case we lemmatised that term and compounded it with the third chunk's final term. For example:

- [NP Mirrors_NNS] [VP are_VBP found_VBN] [PP in_IN] [NP the_DT bedroom_NN] became mirror found in bedroom
- [NP Most_JJS cats_NNS] [VP have_VBP] [NP furry_NN tails_NNS] became cat have tail
- [NP The_DT microwave_NN] [VP was_VBD running_VBG] [PP on_IN] [NP electricity_NN] became microwave run on electricity


## Relation selection

The third stage of our system worked by taking each con-cept-feature pair from both the SVM and chunking output, and finding the best relation for that pair from the chunking output to generate a triple. It also assigned to that triple a number of metrics relating to its constituent parts, their relative frequency and association scores.

We assumed that each concept-feature pair had one corresponding relation. We called the set of extracted triples generated by Stage 2, $T$ (with triples $(c, r, f) \in T$ ) and the set of all extracted relations from Stage $2, R$. For each concept, we also generated a final potential feature set, $F_{c}$, which, for a given concept, was the union of the top 200 features from Stage 1 (ranked by their SVM score) and the top 200 features from Stage 2 (ranked by their frequency in the extracted relations, but excluding features which appeared only once).

We defined Concept Feature Frequency (CFF) to be the number of times a concept, $c$, and feature, $f$, co-occurred across our extracted relations:

$$
\begin{equation*}
\operatorname{CFF}(c, f)=\sum_{r \in R} \operatorname{freq}(c, r, f) \tag{1}
\end{equation*}
$$

We also calculated a Distinct Relation Score which measured the number of distinct relations linking $c$ to $f$ :

$$
\begin{equation*}
\operatorname{DRS}(c, f)=\left|D_{c, f}\right| \text { for } D_{c, f}=\{r:(c, r, f) \in T\} \tag{2}
\end{equation*}
$$

We next wanted to choose relations for our various con-cept-feature pairs, $(c, f) \in C \times F_{c}$. We did this in three steps.

Step 1 For each concept, $c$, and feature, $f$, we iterated through all relations relating to that pair and calculated an Exact Match Score:

$$
\begin{equation*}
\operatorname{EMS}(c, f)=\max \{\operatorname{freq}(c, r, f): r \in R\} \tag{3}
\end{equation*}
$$

If $\operatorname{EMS}(c, f)>0$ then we selected as our best relation, $\hat{r}$, the relation corresponding to that score. If there was more than one relation with the same score, then we chose the least common (i.e., that which had the lowest frequency across all our relations). If $\operatorname{EMS}(c, f)=0$ then we left $\hat{r}$ undefined.

Step 2 Our first step only retrieved a relation if there was an exact match amongst our relation extraction output.

If there wasn't, we took a split approach; given a particular concept, $c$, and feature, $f$, we calculated separate probabilities across all our relations of $c$ occurring with each relation, and of $f$ occurring with each relation. We then calculated for each relation, $r$, a combined score for the combination of $c$, $r$ and $f$ by multiplying the constituent probabilities together. Our pairwise combination score was defined:

$$
\begin{equation*}
p(c, r)=\sum_{f \in F} \frac{\operatorname{freq}(c, r, f)}{\operatorname{freq}(c) \cdot \operatorname{freq}(r)} \tag{4a}
\end{equation*}
$$

$$
\begin{gather*}
p(r, f)=\sum_{c \in C} \frac{\operatorname{freq}(c, r, f)}{\operatorname{freq}(r) \cdot \operatorname{freq}(f)}  \tag{4b}\\
\operatorname{PCS}(c, f)=\left\{\begin{array}{l}
p(c, \hat{r}) \cdot p(\hat{r}, f) \text { if } \hat{r} \text { defined } \\
\max \{p(c, r) \cdot p(r, f): r \in R\}
\end{array}\right. \tag{4c}
\end{gather*}
$$

If we had not already selected a best relation, $\hat{r}$, then we defined it as the relation, $r$, which corresponded to this pairwise combination score. Again, if there was more than one relation with the same score, then we chose the least common.
Step 3 Our final step assigned relations to concept/feature pairs which lacked an exact mutually linking relation. This occurred around $17 \%$ of the time and was usually due to both the concept and feature terms being relatively low frequency. ${ }^{6}$

To achieve this, we backed-off to semantic feature clusters: we defined $f_{\star}$ as the cluster for feature $f$, and $F_{\star}$ as the set of all feature clusters, and defined our Feature Cluster Score, $\operatorname{FCS}\left(c, f_{\star}\right)$, analogously to our Pairwise Combination Score, merely substituting all instances of $f$ for $f_{\star}$. Our best relation, $\hat{r}$, was defined as the relation corresponding to this FCS.

## Reweighting

In our system's fourth and final stage we used the metrics derived above to assign an overall score for each triple using a weighting of parameters; we used our training set to derive the most optimal values for these parameters. We normalised our various metrics so that they all lay between 0 and 1 .

Our relation selection stage had already fixed a relation, $\hat{r}$, for each concept and feature. Hence we calculated for each of our triples $t=(c, \hat{r}, f)$ an overall score:

$$
\begin{align*}
\operatorname{score}(t)= & \beta_{\mathrm{PMI}} \cdot \operatorname{PMI}(t)+\beta_{\mathrm{LL}} \cdot \mathrm{LL}(t)+\beta_{\mathrm{SVM}} \cdot \operatorname{SVM}(t) \\
& +\beta_{\mathrm{CFF}} \cdot \mathrm{CFF}(t)+\beta_{\mathrm{DRS}} \cdot \operatorname{DRS}(t)+\beta_{\mathrm{EMS}} \cdot \operatorname{EMS}(t)  \tag{5}\\
& +\beta_{\mathrm{PCS}} \cdot \operatorname{PCS}(t)+\beta_{\mathrm{FCS}} \cdot \operatorname{FCS}(t)
\end{align*}
$$

We wished to optimise our parameters for superior feature F-score performance against our training set. We employed a stochastic process to find best-possible values for our training parameters, using a random-restart hill-climbing algorithm, repeated 1000 times and selecting the output (and $\beta$ values) offering the best F-score across these iterations.

This process offered a reasonable approximation of the best possible F-scores our system could produce and their corresponding $\beta$ values; following this process, our best F-scores were $0.2739,0.2803$ and 0.2996 for our Wikipedia, UKWAC and combined corpora respectively.

## Evaluation

We evaluated our system using gold standard, human semantic-similarity and direct human evaluations.

## Gold standard evaluation

We began by comparing our top twenty output using the ESSLLI gold standard set. This 'expansion' set comprises the top

[^111]Table 2: Our best precision, recall and F-scores against the synonym-expanded ESSLLI norms across our corpora, found using the training $\beta$ parameters.

| Relation | Corpus | Prec. | Recall | F |
| ---: | :--- | :--- | :--- | :--- |
| With | Wikipedia | 0.1131 | 0.2265 | 0.1509 |
|  | UKWAC | 0.1000 | 0.2005 | 0.1335 |
|  | Combined | 0.1214 | 0.2431 | 0.1620 |
|  | Kelly et al. | 0.1238 | 0.2493 | 0.1654 |
|  | Wikipedia | 0.1214 | 0.2431 | 0.1620 |
|  | UKWAC | 0.1048 | 0.2101 | 0.1398 |
|  | Combined | 0.1298 | 0.2598 | 0.1731 |
| Without | Wikipedia | 0.2798 | 0.5603 | 0.3732 |
|  | UKWAC | 0.2560 | 0.5132 | 0.3416 |
|  | Combined | 0.2798 | 0.5606 | 0.3733 |
|  | Kelly et al. | 0.2417 | 0.4847 | 0.3225 |

ten lemmatised properties for each of 44 concepts from the recoded McRae norms, together with a feature expansion set generated for each concept relation feature triple. One of the reasons for using this set is that McRae et al. normalised their features by channelling synonymous properties into a single representation. The ESSLLI set undoes some of these normalizations, expanding the feature terms to a set of synonyms. In this way, loud, noise and noisy (for example) can all be counted as matches against the property is loud. The relations were not expanded.

Our results can be found in Table 2. We also assessed our system using the full text of the relations found in the original McRae norms as additional 'relation synonyms'; these augmented results can be found under the 'With (aug.)' relation heading. We have exceeded the performance of Kelly et al. (2012) (best F-score of 0.1654 ) with a best overall F-score of 0.1731 for the combined corpus.

We also note that performing these evaluations on the top ten properties returned further improved the situation (perhaps unsurprising since the ESSLLI set contains only ten properties per concept); for example, evaluating our top ten triples against the relation synonyms set returned a precision of 0.2215 for the combined corpus. Furthermore, the precision on the combined corpus for the top ten evaluation of features-only was 0.4409 , surpassing Baroni et al. (2009) who offer a best score of 0.239 on the same evaluation.

## Human-generated semantic similarity

Comparison with the ESSLLI gold standard is still an incomplete evaluation: not all conceptual properties for a given concept are contained therein, and lexical variation can mark valid relations as wrong. Furthermore, one of the primary advantages of our computational approach is its ability to extract a large number of properties for a given concept. Hence, we introduced an alternative approach to calculate how semantically meaningful our output was by evaluating the triples' capacity to predict human-rated similarity between words.

We asked five native English speakers to rate the similarity of 90 concept pairs, where concepts in the pairs were all drawn from the ESSLLI set. The raters were given instructions explaining the task and then presented with each concept pair, one by one, a scale of 1 to 7 and asked to rate how

Table 3: Pearson correlation ( $r$ ) results with confidence intervals between our $V_{\text {Human }}$ vector and our similarity vectors $V$ (with dimensionality $D$ and derived from the top $n$ properties) from our system.

| Relation | V | $n$ | D | $r$ | Conf. Int. |
| :---: | :---: | :---: | :---: | :---: | :---: |
| With | Wikipedia | 20 | 654 | 0.598 | [0.446, 0.716] |
|  | UKWAC |  | 712 | 0.629 | [0.486, 0.740] |
|  | Combined |  | 692 | 0.671 | [0.539, 0.771] |
|  | Wikipedia | 300 | 3585 | 0.693 | [0.568, 0.787$]$ |
|  | UKWAC |  | 3442 | 0.683 | [0.555, 0.780$]$ |
|  | Combined |  | 3380 | 0.723 | [0.606, 0.809] |
| Without | Wikipedia | 20 | 478 | 0.720 | [0.603, 0.807] |
|  | UKWAC |  | 456 | 0.754 | [0.649, 0.832] |
|  | Combined |  | 475 | 0.742 | [0.632, 0.822 ] |
|  | Wikipedia | 300 | 7324 | 0.782 | [0.685, 0.851] |
|  | UKWAC |  | 8698 | 0.806 | [0.719, 0.868] |
|  | Combined |  | 8727 | 0.807 | [0.721, 0.869] |
| WithWithout | McRae |  | 410 | 0.785 | [0.691, 0.854] |
|  |  |  | 355 | 0.787 | [0.693, 0.855$]$ |
|  | LSA |  | 300 | 0.708 | [0.586, 0.798] |

similar the two concepts were.
To compare our system with these ratings we constructed a vector space of dimension $D$, where $D$ was the number of distinct properties across our triples. For each of our 44 concepts, we generated a concept-score vector with non-zero entries by inserting the triple scores, $\operatorname{score}(t)$, into their correct entries in the concept-score vector. We then constructed a $44 \times 44$ symmetric pairwise similarity matrix across our concepts by calculating the cosine similarity between their concept-score vectors. From this we extracted a similarity vector, $V$, for our 90 pairwise comparisons.

We calculated twelve such matrices (using the top twenty and top 300 extracted triples, across three corpora and excluding and including the relation term). We also generated two such matrices using both the feature-heads and the full text of the McRae property norms, using the norm production frequencies as entries in each concept's vector, as well as comparing our ratings with LSA-predicted (Landauer, Foltz, \& Laham, 1998) similarities. ${ }^{7}$ Our results are in Table 3. ${ }^{8}$

Our systems' performance, evaluating with and without relation and when using the top twenty triples, was comparable to LSA (correlation 0.708) with average correlations across our corpora of 0.754 and 0.671 respectively. Including the top 300 extracted triples brought our correlations up to 0.807 and 0.754 respectively, an extremely strong result given that the average Pearson coefficient of correlation across the five judges (considering all pairwise combinations) was 0.820 .

## Human evaluation

In our final evaluation, we asked two native English speaking human judges to assess the accuracy of our triples. Following the methodology of Devereux et al. (2009), we asked them to classify output triples for 15 concepts into four categories: 'correct' (c), 'plausible' (p), 'related' (r) and 'wrong'

[^112]Table 4: Inter-annotator agreement and judgements for our extraction system applied to our three corpora.

| Corpus |  | Audge | A | B | Avg | c/p |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | \(\left.\begin{array}{c}Kappa <br>

(Agree)\end{array}\right]\)
(w). Our judges were unaware of the aims of the evaluation. We concatenated their ratings using the methodology of Devereux et al. ${ }^{9}$ however our instructions reflected the fact that, unlike previous systems, our output contained prepositional relations and we therefore did not wish our volunteers to allow for absent prepositions. This evaluation offers an important insight into the viability of our method as a property extraction system. Our results are in Table 4, and Table 5 shows a sample of our output and the corresponding judgements.

It is clear that our best results were again in the combined corpus, where an impressive $71.3 \%$ of our returned triples were marked as either plausible or correct with a Kappa (Fleiss, 1971) score of 0.7229 indicating substantial agreement between annotators. This constitutes a major improvement over Kelly et al. (2012) who evaluated on the same set of concepts and whose corresponding score was just $51.1 \%$.

## Discussion

As the first system to offer viable unconstrained property norm-like extraction, this paper brings research into conceptual property extraction to the next level. Our system employs both full parsing and chunking to extract features and relations respectively and introduces a novel multi-step backingoff method for relation selection. Our gold standard performance exceeded that of previous approaches, and our human evaluation indicated that we have outperformed the system of Kelly et al. (2012) by a significant margin. We also introduced a semantic similarity evaluation for this task, showing a strong Pearson correlation of 0.754 with human ratings when employing just 20 extracted properties per concept, with the correlation rising to 0.807 when using 300 properties. In this latter case, the predicted similarities were almost as correlated with human judgements as the human judgements are with each other.

Potential criticisms of our system include the fact that our chunk to triple conversion process won't necessarily always yield a true reflection of the sentence's original meaning. It is, for example, possible for the final chunk to contain adjectives which modify the final noun. These could have importance from a conceptual representation perspective (e.g., features such as long neck for giraffe has long neck). Also, the modifying portion of a chunk may be semantically significant, altering the final term's meaning (e.g., a tea bag is quite different from a bag). It should be possible to have more gen-

[^113]Table 5: Judges' assessments of the top twenty extracted relation/feature pairs (combined corpus) for two concepts.

| knife |  | Judge |  | Judge |  |
| :--- | :---: | :--- | :--- | :--- | :--- |
|  | A | B | pig | A | B |
| sharpened by hand | c | c | eat piglet | c | p |
| based on design | c | c | get fat | c | c |
| made of steel | c | c | produce pork | r | c |
| be small | c | p | breed farm | r | r |
| pick on fork | r | r | put into sausage | c | c |
| be make | p | r | be large | p | p |
| crafted from metal | c | c | have baby | c | c |
| scaled for use | p | p | be different | p | p |
| make cut | c | c | stunned through use | r | w |
| be sharp | c | c | be bacon | c | r |
| be weapon | c | c | be welfare | r | r |
| have edge | c | c | discover sheep | c | c |
| have handle | c | c | killed for meat | c | c |
| be serrated | c | c | used for food | c | c |
| made of stainless | w | r | label cattle | w | w |
| is for cutting | c | c | be animal | c | c |
| have blade | c | c | shackled by ham | r | r |
| be useful | p | c | chew tail | c | c |
| be tool | c | c | have disease | c | c |
| be dangerous | c | c | found in guinea | c | c |

eral chunk to triple extraction (e.g., by using a larger corpus to mitigate the sparsity associated with multi-word terms).

Finally, a major issue is our lack of comprehensive training/testing data; our norms are incomplete insofar as there were a large number of 'correct' properties absent from our gold standard. In future work we hope to implement largescale evaluation of our system's output (e.g., using Amazon Turk) which would allow us to rapidly obtain large amounts of human-generated feedback. We could then use activelearning to introduce a feedback loop of human-annotation to better pinpoint inaccurate features or relations. Feedback which strongly indicated that certain properties were uninteresting could prove invaluable in getting even closer to a conceptual structure-like representation of concepts.

## Acknowledgments

This research was supported by EPSRC grant EP/F030061/1 and the Royal Society University Research Fellowship, UK. We are grateful to McRae and colleagues for making their norms publicly available, and to the anonymous reviewers for their comments and feedback.

## References

Baldridge, J. (2005). The Apache OpenNLP project.
Baroni, M., Evert, S., \& Lenci, A. (Eds.). (2008). ESSLLI 2008 Workshop on Distributional Lexical Semantics.
Baroni, M., Murphy, B., Barbu, E., \& Poesio, M. (2009). Strudel: A corpus-based semantic model based on properties and types. Cognitive Science, 1-33.
Church, K., \& Hanks, P. (1990). Word association norms, mutual information, and lexicography. Computational Linguistics, 16(1), 22-29.
Clark, S., \& Curran, J. (2007, 1). Wide-coverage efficient statistical parsing with CCG and log-linear models. Computational Linguistics, 33(4), 493-552.

Devereux, B., Pilkington, N., Poibeau, T., \& Korhonen, A. (2009). Towards unrestricted, large-scale acquisition of feature-based conceptual representations from corpus data. Research on Language \& Computation, 1-34.
Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19(1), 61-74.
Farah, M., \& McClelland, J. (1991). A computational model of semantic memory impairment: Modality specificity and emergent category specificity. Journal of Experimental Psychology: General, 120(4), 339-357.
Ferraresi, A., Zanchetta, E., Baroni, M., \& Bernardini, S. (2008). Introducing and evaluating UKWAC, a very large web-derived corpus of English. In Proceedings of the 4th Web as Corpus Workshop (wac-4) - Can we beat Google? (pp. 47-54).
Fisher, R. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika, 10(4), 507-521.
Fleiss, J. (1971). Measuring nominal scale agreement among many raters. Psychological bulletin, 76(5), 378.
Joachims, T. (1999). Making large scale SVM learning practical.
Kelly, C., Devereux, B., \& Korhonen, A. (2010). Acquiring human-like feature-based conceptual representations from corpora. In First Workshop on Computational Neurolinguistics (p. 61). Association for Computational Linguistics.
Kelly, C., Devereux, B., \& Korhonen, A. (2012). Semisupervised learning for automatic conceptual property extraction. In Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics (pp. 11-20). Association for Computational Linguistics.
Landauer, T., Foltz, P., \& Laham, D. (1998). An introduction to latent semantic analysis. Discourse Processes, 25, 259284.

McRae, K., Cree, G., Seidenberg, M., \& McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. Behavioral Research Methods, Instruments, and Computers, 37, 547-559.
Murphy, G. (2002). The Big Book of Concepts. The MIT Press.
Randall, B., Moss, H., Rodd, J., Greer, M., \& Tyler, L. (2004). Distinctiveness and correlation in conceptual structure: Behavioral and computational studies. Journal of Experimental Psychology Learning Memory and Cognition, 30(2), 393-406.
Taylor, K., Devereux, B., Acres, K., Randall, B., \& Tyler, L. (2011). Contrasting effects of feature-based statistics on the categorisation and basic-level identification of visual objects. Cognition, 122(3), 363-74.
Tyler, L., Moss, H., Durrant-Peatfield, M., \& Levy, J. (2000). Conceptual structure and the structure of concepts: A distributed account of category-specific deficits. Brain and Language, 75(2), 195-231.

# Musicality and non-native speech sound processing are linked through temporal, pitch and spectral acuity. 

Vera Kempe (vkempe@abertay.ac.uk)<br>Division of Psychology, School of Social and Health Sciences, Bell Street<br>Dundee, DD1 1HG, United Kingdom<br>Dennis Bublitz (dennis.bublitz@gmail.com)<br>Department of Psychology, College of Staten Island, City University of New York Staten Island, NY 10314 USA

Patricia J. Brooks (patricia.brooks@csi.cuny.edu)
Department of Psychology, College of Staten Island, City University of New York Staten Island, NY 10314 USA


#### Abstract

Are observed links between musicality and non-native speech sound processing due to superior sensory processing of temporal, pitch, and spectral information, which benefits both musical and linguistic processing? Native English speakers discriminated Norwegian tonal contrasts, non-linguistic puretone analogues, Norwegian vowels, and short tones differing in temporal, pitch and spectral characteristics. Musicality was measured using Gordon's (1989) Advanced Measures of Musical Audiation (AMMA). After controlling for effects of sex, non-verbal IQ and previous language learning experience, the link between AMMA scores and tonal contour discrimination was partially mediated by pitch acuity. In addition, tonal contrast, pitch contour and vowel discrimination were predicted by temporal and spectral acuity. No independent effects of musical training were found. Thus, links between musicality and non-native speech sound processing appear to be mainly due to superior temporal, pitch or spectral acuity, which, in turn, may play somewhat different roles in processing different speech sounds.


Keywords: Non-native phoneme processing; temporal acuity, pitch acuity; spectral acuity; musicality; tonal contrast; vowel contrast.

## Introduction

A number of studies have documented links between musicality and the ability to discriminate non-native speech sounds (Delogu, Lampis, Belardinelli, 2006, 2010; Marie, Delogu, Lampis, Belardinelli \& Besson, 2011; Slevc \& Miyake 2006; Wong, Skoe, Russo, Dees \& Kraus, 2007). These studies typically use complex psychometric measures of musical aptitude, which leaves open the question as to which specific sub-components of musical aptitude are associated with non-native speech sound processing. Tests like the Wing test (Wing, 1968) or Gordon's Advanced Measures of Musical Audiation (AMMA) (Gordon, 1989) rely on working memory for musical and rhythmic phrases as well as on the ability to discriminate subtle differences in pitch, timbre, intensity, and rhythm. It is not clear which of these sub-components of musical aptitude are linked to non-
native sound processing, especially because different types of acoustic information may be important for distinguishing different types of speech sounds. Specifically, the perception of vowels, which differ in spectral information associated with the first and second formants, should be most strongly predicted by sensitivity to timbre. In contrast, the perception of consonants, which are often distinguished by temporal information such as Voice Onset Time or formant transitions, should benefit from sensitivity to rapid temporal changes. Finally, lexical tones require sensitivity to pitch and, to the extent that they encompass differences in pitch contour, also sensitivity to temporal information. Thus, different aspects of auditory sensory acuity may be important for the processing of different types of non-native speech sounds. So far, the relationship between musicality and non-native speech sounds processing has been quite consistently established for tonal contrasts (Marie, Delogu, Lampis, Belardinelli \& Besson, 2011; Slevc \& Miyake, 2006; Wong, Skoe, Russo, Dees \& Kraus, 2007), but the findings are less clear for phonological contrasts (Delogu, Lampis \& Belardinelli, 2006, 2010).

A primary aim of this study was therefore to examine the specific contributions of temporal, pitch, and spectral acuity to the processing of different non-native speech sounds, and to determine whether general measures of musical aptitude can explain additional variance in non-native speech sound processing above what is explained by these basic sensory processes. To this end, we examined both a tonal and a vowel contrast that exist in Norwegian, a language unfamiliar to our participants. Many dialects of Norwegian have lexical tone such that rising and falling-rising pitch accents distinguish minimal pairs of segmentally identical bi-syllabic words. For example, 'Hammer', spoken with the rising tone, is a Norwegian proper noun while 'hammer', spoken with the falling-rising tone, denotes the tool. These contrasts encompass temporal changes in fundamental frequency in the range of several hundreds of milliseconds. Norwegian also contains a vowel contrast not present in English, the /i/ - /y/ contrast. The existence of these Norwegian contrasts offers the possibility to use linguistic
stimuli rather than isolated synthesized segments, which may be processed in ways that differ from processing of natural linguistic materials. Moreover, to control for differences between linguistic and non-linguistic stimuli, we also used the extracted pure-tone analogues of the Norwegian tonal contrasts as stimuli.

Musicality is a complex construct encompassing musical aptitude as well as musical expertise (Nardo \& Reiterer, 2009). Studies comparing non-native speech sound processing between musicians and non-musicians (Marie et al., 2011; Wong et al., 2007) suggest that musical practice hones abilities such as sensory acuity, working memory, or attentional control, which may transfer to the linguistic domain. However, it is also possible that certain sensory or cognitive abilities benefit both musical and linguistic processing. To see whether musical expertise incurs benefits for non-native speech sound processing in addition to benefits associated with superior auditory sensory acuity, we also examined whether the duration of musical training would explain variance in non-native speech sound processing over and above measures of musical aptitude.

## Method

Native speakers of English completed AX discrimination tasks for non-native tonal and vowel contrasts, as well as for synthesized sounds differing in temporal, pitch, and spectral characteristics. Musical aptitude was tested using Gordon's AMMA (Gordon, 1989). To control for non-verbal intelligence we administered Cattell's Culture Fair Intelligence Test (Cattell \& Cattell, 1973). Music and language background questionnaires inquired about length of musical training and number of languages learned at home or at school, and elicited self-ratings of proficiency in each language (L2 and L3).

## Participants

One hundred and three native speakers of American English ( 58 women, 45 men ) aged 19-22 years participated in the study. Three participants failed to provide L3 proficiency self-ratings, and one participant failed to provide pitch discrimination data. These participants are missing from analyses including these variables.

## Materials and Tasks

Advanced Measures of Musical Audiation (AMMA): Gordon's (1989) AMMA consists of 30 items, each of which comprises a short musical 'statement' followed after four seconds by a short 'answer' of the same length. These items contain either one or more tonal changes, or one or more rhythmic changes, but not both. Participants have to decide whether the phrases are the same or different. For 'different' items, participants are asked to decide whether the difference is a tonal or rhythm change. The test yields a tonal and rhythm score, as well as a composite score.
AX ('same-different') - tasks: All AX tasks comprised 32 'same' and 32 'different' trials. For the Norwegian tonal and vowel contrasts, 'same' trials comprised different
within-category instantiations obtained from repeated recordings of the same word.
Temporal Acuity. We synthesized eight 250 Hz sinusoidal carrier waves with an overall duration of 600 ms differing in amplitude envelope onset rise times, and otherwise devoid of segmental, spectral and pitch information. The onset of the amplitude envelope was faded in with rise times to reach maximum amplitude at $0 \mathrm{~ms}, 10 \mathrm{~ms}, 20 \mathrm{~ms}, 30 \mathrm{~ms}, 60 \mathrm{~ms}$, $70 \mathrm{~ms}, 80 \mathrm{~ms}$ and 90 ms . 'Different' trials comprised pairs of sounds differing in rise times by 60 ms (e.g. 0 ms vs. 60 ms or 10 ms vs. 70 ms etc.), centered around 45 ms , a value which has been reported as the category boundary between 'bowed' and 'plucked' sounds (Cutting \& Rosner, 1974).
Pitch Acuity. We created eight 500 ms pure tone sinusoidal carrier waves ranging from 100 to 3000 Hz in steps increasing by 100 Hz resulting in tones of $100 \mathrm{~Hz}, 200$ $\mathrm{Hz}, 400 \mathrm{~Hz}, 700 \mathrm{~Hz}, 1100 \mathrm{~Hz}, 1600 \mathrm{~Hz}, 2200 \mathrm{~Hz}$, and 3000 Hz , as well as contrasts with a frequency increased by $2 \%$ resulting in tones of $102 \mathrm{~Hz}, 204 \mathrm{~Hz}, 408 \mathrm{~Hz}, 714 \mathrm{~Hz}, 1122$ $\mathrm{Hz}, 2244 \mathrm{~Hz}$, and 3060 Hz . The cumulative increase was designed to create sound pairs that subjectively sampled the pitch range at roughly similar intervals, taking into account the non-linearity of pitch perception. For the 'different' trials, each sound was paired with its corresponding contrast sound resulting in pairs of 100 Hz vs. $102 \mathrm{~Hz}, 200 \mathrm{~Hz}$ vs. $204 \mathrm{~Hz}, 400 \mathrm{~Hz}$ vs. 408 Hz etc.)
Spectral Acuity. To test spectral acuity, we incorporated the pure tones created for the pitch acuity test into complex tones comprising low (e.g. 100 Hz or 200 Hz ), middle (e.g. 700 Hz or 1100 Hz ) and high (e.g. 2200 Hz or 3000 Hz ) frequencies. These frequencies were chosen to broadly mimic the fundamental frequency and the first two formants of speech, which are crucial for vowel perception. For 'different' pairs, one of the component tones was increased by $2 \%$, and this change affected either the middle or the high frequency. For example, a 'different' pair might include a complex tone consisting of frequencies of 100 Hz , 1100 Hz and 3000 Hz and a complex tone consisting of frequencies of $100 \mathrm{~Hz}, 1122 \mathrm{~Hz}$ and 3000 Hz .
Norwegian tonal contrast. Recordings by a male native speaker of Norwegian of eight minimal pairs of Norwegian words containing a contrast between rising and falling-rising tonal contours were taken from Kempe, Thoresen, Kirk, Schaeffler \& Brooks (2012). Four pairs contained short vowels in the first (stressed) syllable (mean length 64 ms ); the remaining four pairs contained long vowels (mean length 187 ms ). Crucially, words with rising and with falling-rising tones did not differ in length of the first vowel ( 118 vs. $133 \mathrm{~ms}, \mathrm{p}=.5$ ), overall word length ( 396 vs .417 $\mathrm{ms}, \mathrm{p}=.2$ ), and metric stress; thus, duration and stress could not be used as additional cues. Corresponding short and long vowel pairs were matched for their initial phoneme.

To ensure that a male advantage in the discrimination these tonal contrasts, as found in previous research (Kempe et al., 2012), was not an artifact of the male voice presenting the stimuli, we also created a female voice version of the stimuli. To control for indexical features, we submitted the
male voice stimuli to the voice gender change algorithm in PRAAT (Boersma \& Weenink, 2011) using a fundamental frequency of 220 Hz and scaling the first formant up by 20 $\%$. All results below are averaged over the male-voiced and the female-voiced version of the AX-task.
Pitch contour (non-speech analogue of tonal contrast). The non-speech equivalents of the Norwegian tonal contrast comprised sine waves with pitch contours extracted from the fundamental frequency modulation of both the malevoiced and the female-voice Norwegian tonal contrasts. These stimuli contained no information other than the pitch contour of the Norwegian tones. Again, results below are averaged over the male-voiced and the female-voiced version of the AX-task.
Norwegian vowel contrast. We used eight minimal pairs of Norwegian mono-syllabic words containing the vowel /i:/ or /I/ vs. /y:/ or /Y/, a contrast between high front unrounded and rounded vowels which does not exist in English. Recordings of a male native speaker of Norwegian were taken from Kempe et al. (2012). Four word pairs contained the short vowels $/ \mathrm{I} /$ and $/ \mathrm{Y} /$ (mean length 67 ms ), and the remaining four word pairs contained the long vowels /i:/ and $/ \mathrm{y}: /$ (mean length 150 ms ). On average both members of a minimal pair did not differ in vowel length ( 108 vs .108 ms , $\mathrm{p}=.9$ ); thus, duration could not serve as additional cue.

Other measures: Participants also completed Cattell's Culture-Fair Test of Nonverbal Intelligence, Scale 3, Form A (Cattell \& Cattell, 1973), a music background questionnaire on which they provided information about duration of musical training, and a language background questionnaire on which they indicated the number of languages learned, and rated their reading, writing, speaking and comprehension abilities in all languages on a scale from 1 (very poor) to 6 (native-like).

## Procedure

AX discrimination tasks were presented in three blocks, with the first block containing the temporal, pitch and spectral AX tasks, the second block containing the malevoiced and female-voiced tonal AX tasks as well as the vowel AX task, and the third block containing the two AX tasks presenting the extracted pitch contours of the malevoiced and female-voiced Norwegian tonal stimuli. The fixed block sequence ensured that variance due to order effects was not confounded with participant variance, although task order was counterbalanced within blocks. AMMA and Culture Fair Intelligence Test were interspersed between blocks with their order counterbalanced as well. Informed consent was obtained and background questionnaires were completed prior to any of the tasks.

In each of the AX-tasks, participants received eight practice trials with feedback, followed by 64 test trials without feedback. Within a trial, sound stimuli were separated by an inter-stimulus interval of 200 ms ; the intertrial interval was 500 ms . Participants were asked to press the ' $s$ ' key if they perceived the sounds to be the same and the ' $d$ ' key if they perceived them to be different.

## Results

Participants' performance on the AX-tasks was converted into A', a sensitivity measure that corrects for differences in response bias, and ranges from 0 to 1 , with 0.5 corresponding to chance. Table 1 shows performance with tonal contrasts, non-speech contour analogues, and vowels. As previous research had shown a male advantage for the processing of some non-native speech sounds (Kempe et al., 2012; Bowles, Silbert, Jackson \& Doughy, 2011), results are given for male and female participants separately. A 3 (Condition) x 2 (Sex) ANOVA yielded a main effect of Condition, $\mathrm{F}(2,202)=8.2, \mathrm{p}<.001$. Bonferroni-corrected post-hoc tests indicated that performance was superior for the vowels compared to both tonal conditions, all t's $>3.4$, all p 's $<.01$, which did not differ from each other, $\mathrm{p}=.7$. The main effect of Sex, $\mathrm{F}(1,101)=2.3, \mathrm{p}=.1$, and the interaction, $\mathrm{F}(2,202)=2.8$, p $=.06$, fell short of significance. This trend towards a male advantage in processing tonal contours, but not vowels confirms the previous findings (Kempe et al., 2012).

Table 1: Mean A' and standard deviations (in parentheses) for discrimination of Norwegian tonal contrasts, extracted pitch contours and Norwegian vowels.

|  | tonal contrast | pitch contour | vowel contrast |
| :--- | :--- | :--- | :--- |
| males | $0.781(0.103)$ | $0.790(0.087)$ | $0.800(0.101)$ |
| females | $0.748(0.097)$ | $0.747(0.104)$ | $0.802(0.102)$ |

Zero-order correlations between the predictors of nonnative speech-sound processing are provided in Table 2. Noteworthy findings involve a positive correlation between non-verbal intelligence and both AMMA scores, indicating that comparison of musical phrases relies to some extent on mechanisms shared with psychometric intelligence, such as working memory and cognitive control (Duncan, Emslie, Williams, Johnson \& Freer, 1996). Also, as expected, pitch and spectral acuity were positively correlated with both AMMA scores. In contrast, temporal acuity was not correlated with the AMMA scores, which may reflect the fact that the temporal processing relevant for music involves a longer time scale than the rapid temporal changes in the order of tens of milliseconds presented in our temporal acuity test. Instead, temporal acuity was positively correlated with non-verbal intelligence, confirming the documented link between rapid temporal auditory processing and psychometric intelligence (Rammsayer \& Brandler, 2007).
The results of multiple regression analyses of non-verbal intelligence, language background, sex (dummy-coded) and the musicality measures on performance with tonal contrasts, pitch contours and vowels are presented in Table 3 (upper panel). We found that non-verbal intelligence showed a trend towards a positive association with discrimination of tonal contours and non-linguistic pitch contours. As indicated above, there was also a statistically marginal male advantage for these stimuli. Crucially, the

AMMA rhythm score predicted performance with tonal contrasts, even if they were extracted and presented without linguistic information. For non-native vowels, only selfrated proficiency in L3 predicted performance. Thus, while tonal performance was related to musicality, vowel performance was not, confirming observations by Delogu et al. (2006; 2010). Note that there was no effect of musical training.

Next, we added temporal, pitch and spectral acuity to the model (lower panel of Table 3) to see whether auditory acuity explains the link between musicality and non-native tonal processing. All Variance Inflation Factors were below 3.8, suggesting that multi-collinearity was not a problem in this data set. For the tonal contrast, we found a significant effect of pitch acuity; the effects of temporal and spectral acuity fell short of significance. For the extracted pitch contour, we found a significant effect of spectral acuity; the effect of temporal acuity fell short of significance. For the vowel contrast, we found significant effects of temporal and spectral acuity. In other words, the data showed a tendency for temporal and spectral acuity to predict discrimination of all non-native speech contrasts while the predictive effect of pitch acuity was confined to the tonal contrasts.

Table 2: Zero-order Pearson correlations between predictor variables. CFI - Culture Fair Intelligence test, \# of Ls number of learned languages, N 's range from 99 and 103 depending on missing values, ${ }^{* *} \mathrm{p}<.01$, ${ }^{*} \mathrm{p}<.05$.

|  | 2 | 3 | 4 | 5 |  |
| :--- | :--- | :--- | :--- | :--- | :--- |
|  | $.25^{*}$ | .03 | $.23^{*}$ | .09 |  |
| 1. CFI |  | $.42^{* *}$ | $.79^{* *}$ | .04 |  |
| 2. \# of Ls |  |  | $.56^{* *}$ | .17 |  |
| 3. L2 (rating) |  |  |  | .12 |  |
| 4. L3 (rating) |  |  |  |  |  |
|  | 6 | 7 | 8 | 9 | 10 |
|  | $.22^{*}$ | $.30^{* *}$ | $.24^{*}$ | .14. | .07 |
| 1. CFI | .11 | .19 | $.24^{*}$ | .11 | .12 |
| 2. \# of Ls | .06 | .03 | .13 | .06 | .07 |
| 3. L2 (rating) | .19 | .14 | $.21^{*}$ | .11 | .11 |
| 4. L3 (rating) | $.8^{* *}$ | $.25^{*}$ | .13 | .09 | $.28^{* *}$ |
| 5. music (years) | $.21^{*}$ | $.71^{* *}$ | .19 | $.23^{*}$ | $.25^{*}$ |
| 6. tonal score |  |  | .14 | $.34^{* *}$ | $.32^{* *}$ |
| 7. rhythm score |  |  |  | .14 | .12 |
| 8. temporal |  |  |  |  | $.50^{* *}$ |
| 9. pitch |  |  |  |  |  |
| 10. spectral |  |  |  |  |  |

To test explicitly whether the association of the AMMA rhythm score with the tonal and the extracted pitch contour contrasts was mediated by auditory acuity, we performed mediation analyses employing bootstrapping to estimate the $95 \%$ confidence intervals of the indirect effect using a procedure introduced by Hayes and Preacher (2013) for multiple predictor variables (SPSS-macro MEDIATE, http://www.afhayes.com/spss-sas-and-mplus-macros-andcode.html). A relative indirect effect is deemed to be statistically significant at $p=.05$ if these confidence
intervals do not include zero. This analysis revealed that the effect of the AMMA rhythm score on the tonal contrast was partially mediated by pitch acuity (the obtained lower and upper boundaries of the confidence interval were .0001 and .0048 , respectively). For the extracted pitch contour, there were no indirect effects.

Table 3: Standardized regression coefficients and proportion of variance accounted for in regression analyses with performance on tonal contrasts, pitch contours and vowel contrasts as criterion variables and sex, non-verbal intelligence, language background and musical ability measures as predictors at the first step (upper panel) and temporal, pitch and spectral acuity added at the next step (lower panel), ${ }^{* * *} \mathrm{p}<.001,{ }^{* *} \mathrm{p}<.01,{ }^{*} \mathrm{p}<.05,{ }^{\dagger} \mathrm{p}<.1$.

|  | tonal <br> contrast | pitch <br> contour | vowel <br> contrast |
| :--- | :--- | :--- | :--- |
| sex | $-.16^{\dagger}$ | $-.19^{\dagger}$ | .02 |
| CFI | $.16^{\dagger}$ | $21^{*}$ | .06 |
| \# of Ls | -.04 | -.06 | -.17 |
| L2 (rating) | .15 | -.07 | .03 |
| L3 (rating) | .03 | .06 | $.36^{*}$ |
| music (years) | .09 | .00 | -.01 |
| tonal score | -.03 | .01 | .16 |
| rhythm score | $.40^{* *}$ | $.30^{*}$ | .11 |
| adj. R | .21 | .15 | .09 |
| F(9,97) | $4.29^{* * *}$ | $3.10^{* *}$ | $2.23^{*}$ |
|  |  |  |  |
|  | tonal | pitch | vowel |
|  | contrast | contour | contrast |
| sex | -.10 | $-.18^{\dagger}$ | .02 |
| CFI | $.16^{\dagger}$ | $.19^{\dagger}$ | .04 |
| \# of Ls | -.06 | -.10 | -.23 |
| L2 (rating) | .16 | -.06 | .04 |
| L3 (rating) | .00 | .05 | $.36^{*}$ |
| music (years) | .02 | -.07 | -.10 |
| tonal score | -.09 | -.03 | .09 |
| rhythm score | $.32^{*}$ | $.26^{\dagger}$ | .10 |
| temporal | $.17^{\dagger}$ | $.18^{\dagger}$ | $.29^{* *}$ |
| pitch | $.22^{*}$ | .03 | -.05 |
| spectral | $.19^{\dagger}$ | $.22^{*}$ | $.26^{*}$ |
| adj. R | .34 | .20 | .20 |
| F(9,97) | $5.54^{* * *}$ | $3.23^{* *}$ | $3.23^{* *}$ |

## Discussion

When different measures of auditory sensory acuity relevant to non-native speech-sound processing were added into a multiple regression model, they had an independent effect beyond effects of musical aptitude. For Norwegian tonal contrasts, the effect of musical aptitude was partially mediated by pitch acuity. For non-linguistic pitch contours and for vowels, performance was mainly predicted by temporal and spectral acuity rather than musical aptitude. This suggests that associations between musical aptitude and non-native speech-sound processing predominantly
arise from superior sensory processing encompassing the ability to make subtle distinctions in temporal, pitch, and spectral properties of the sounds. This finding is incompatible with claims that musical and linguistic processing exploits different cues-with language mainly relying on rapid temporal processing and music relying on processing of pitch and spectral information (Zatorre, Belin \& Penhune, 2002). By showing that musical aptitude contributes little to non-native speech sound processing beyond effects of temporal, pitch and spectral acuity, our findings underscore the importance of these basic sensory processes for both music and speech sound processing, and thereby support the idea of partially shared mechanisms (Patel, 2003; Strait, Hornickel \& Kraus, 2011). This is not to say that musical aptitude does not encompass other components that may or may not be shared with language; rather, we suggest that the links between musical aptitude tests and non-native speech sound processing reported in the literature may be due to basic temporal, pitch and spectral acuity to a significant degree.
Before sensory acuity measures were added to the regression model, effects of musical ability were found only for tonal contrasts and their non-linguistic analogues, but not for the phonemic vowel contrast. This confirms reports that the link between musicality and non-native speech sound processing is strongest for tonal contrasts (Delogu et al., 2006, 2010). Adding temporal, pitch and spectral acuity to the model qualified this picture: We had hypothesized that discrimination of tonal contrasts would be predicted by temporal as well as pitch and spectral acuity, whereas discrimination of vowel contrasts would be predicted mainly by spectral acuity. Indeed, for tonal contrasts and pitch contours this prediction bore out, although some effects fell short of significance. This confirms the previously observed role of temporal information, in addition to pitch and spectral information, in the processing of pitch contours, whether embedded in linguistic material or presented on their own (Kempe et al., 2012)-an important finding as researchers often conceive of lexical tones as predominantly involving pitch processing, even though lexical tones often entail changes of pitch and spectral information over time.
Counter to our prediction, we found that, along with spectral acuity, temporal acuity was also a significant predictor for vowel processing. This is surprising because we had carefully controlled for vowel length and metrical stress to exclude temporal information as an additional cue. Still, it is possible that participants were sensitive to subtle durational differences in other segments when trying to discriminate between the Norwegian words. Interestingly, performance with the vowel contrast was also significantly predicted by self-rating in an L3, which suggests that some of the languages participants studied later in life might have provided prior exposure to the $/ \mathrm{i} /-/ \mathrm{y} /$ vowel contrast, and this experience may have transferred to our stimuli. The effects of prior language experience notwithstanding, our findings suggest that both temporal and spectral acuity were
important predictors for the particular phonological contrast used in this study; the lack of a significant correlation of either AMMA scores with temporal acuity might explain why musical aptitude did not predict performance for this particular contrast.
More generally, our findings suggest that temporal, spectral and, to some extent, pitch acuity, underlie processing of a variety of non-native speech sounds. This adds an important facet to our understanding of speechsound processing in light of approaches that focus on rapid temporal auditory processing as the main sensory component underlying language (Goswami et al., 2002; Tallal, 1980). Moreover, subtle differences in the role of spectral, pitch, and temporal acuity in the processing of different phonemes might account for why links between musicality and non-native phoneme processing sometimes remain elusive (Delogu et al., 2006, 2010). For the tonal contrasts, on the other hand, the residual effects of musical aptitude may reflect shared working memory components (Williamson, Baddeley \& Hitch, 2010), in addition to sensory components, due to the somewhat longer duration of these stimuli. In addition, the finding that sensory acuity played a similar role in the processing of tonal contrasts and their non-linguistic analogues challenges the view that linguistic stimuli enjoy a special status with respect to basic sensory processing (Gandour et al., 2000)-a conclusion that needs to be verified through neuro-imaging studies.
The finding that the AMMA rhythm score was a better predictor of non-native tonal and pitch contour processing than the tonal score confirms similar findings with respect to non-native speech sound processing (Nardo \& Reiterer, 2009), as well as auditory working memory and reading performance in children (Strait et al., 2011). It appears that information related to repetitiveness, predictability and the sequential nature of sound sequences as measured by the rhythm score has greater diagnostic validity for detecting subtle changes in linguistic stimuli than information related to pitch differences.

A number of studies have conceptualized musicality as musical expertise and suggested that exposure to, and regular practice of music may hone sensory and cognitive abilities, which can subsequently benefit language processing (Marie et al., 2011; Wong et al., 2007). If this were correct one would expect experience with music, measured as number of years of musical training, to exert an independent effect on non-native speech sound processing. Musical training in our sample ranged from 0 to 20 years. Still, no independent effects of length of musical training on non-native speech-sound processing were found. Thus, our findings do not support the notion that musical expertise has independent effects over and above musical aptitude; rather, they suggest that auditory sensory acuity benefits performance both in the musical and the linguistic domain.

Finally, it is worth mentioning that our findings partially replicated the sex difference in non-native speech sound processing observed in previous studies (Bowles et al., 2011; Kempe et al., 2012). This sex difference was assumed
to be due to a general male advantage in rapid temporal auditory processing (Wittman \& Szelag, 2003) and temporal discrimination tasks (Rammsayer \& Troche, 2010), and, thus, should be found for processing of speech sounds that require rapid temporal auditory processing. Indeed, the difference in sensitivity for tonal contrasts and extracted pitch contours pointed towards a male advantage, although the effect fell short of statistical significance in the multiple regression analyses. In further support, a separate analysis for just the tonal and pitch contour stimuli yielded a main effect of sex, $F(1,101)=4.7, p<.05$. Surprisingly, however, the predicted male advantage was not observed for temporal acuity, $\mathrm{p}=.9$; instead, we found a male advantage in pitch acuity, $\mathrm{t}(101)=2.06, \mathrm{p}<.05$. This suggests that while a male advantage in processing of non-native tonal contrasts seems to be a robust phenomenon, future research may have to explore alternative explanations with respect to the underlying mechanisms.

In sum, our findings further illuminate the link between musical aptitude and non-native speech-sound processing by demonstrating that this link is largely explained by sensory acuity. Our results also suggest that temporal, pitch, and spectral acuity all contribute to the processing of a range of non-native speech sounds.

## References

Boersma, P. \& Weenink, D. (2011). Praat: doing phonetics by computer [Computer program] Version 5316, Available: http://wwwfonhumuvanl/praat/. Accessed 2012, May 30.
Bowles, A. R., Silbert, N. H., Jackson, S. R. \& Doughy, C. J. (2011). Individual differences in working memory predict second language learning success. Poster presented at the 52nd Annual Meeting of The Psychonomic Society, Seattle, WA.
Cattell, R. B. \& Cattell, H. E. P. (1973). Measuring Intelligence with the Culture-Fair Tests. Champaign, IL: Institute for Personality and Ability Testing.
Cutting, J. E. \& Rosner, B. S. (1974). Categories and boundaries in speech and music. Perception and Psychophysics, 16, 564-570.
Delogu, F., Lampis, G., Belardinelli, M. O. (2006). Music-to-language transfer effect: may melodic ability improve learning of tonal languages by native nontonal speakers? Cognitive Processes, 7, 203-207.
Delogu, F., Lampis, G., Belardinelli, M. O. (2010). From melody to lexical tone: Musical ability enhances specific aspects of foreign language perception. European Journal of Cognitive Psychology, 22, 46-61.
Duncan, J., Emslie, H., Williams, P., Johnson, R. \& Freer, C. (1996). Intelligence and the frontal lobe: The organization of goal-directed behavior. Cognitive Psychology, 30, 257-303.
Gandour, J., Wong, D., Hsieh, L., Weinzapfel, B., Van Lancker, D., et al. (2000) A crosslinguistic PET study of tone perception. Journal of Cognitive Neuroscience, 12, 207-222.

Gordon, E. E. (1989). Advanced Measures of Music Audiation. Chicago: GIA.
Goswami, U., Thompson, J., Richardson, U., Stainthorp, R., Highes, D., Rosen, S., \& Scott, S. K. (2002). Amplitude envelope onsets and developmental dyslexia: a new hypothesis. Proceedings of the National Academy of Sciences USA, 99, 10911-10916.
Hayes, A. F. \& Preacher, K. J. (2013). Statistical mediation analysis with a multicategorical independent variable [White paper]. Retrieved from http://www.afhayes.com
Marie, C., Delogu, F., Lampis, G., Belardinelli, M. O. \& Besson, M. (2011). Influence of musical expertise on segmental and tonal processing in Mandarin Chinese. Journal of Cognitive Neuroscience, 23, 2701-2715.
Nardo, D., Reiterer, S. M. (2009). Musicality and phonetic language aptitude. In: Dogil, G., Reiterer, S. M. (Eds.). Language talent and brain activity. Trends in applied linguistics Berlin: Mouton De Gruyter.
Patel, A. (2003). Language, music, syntax and the brain. Nature Neuroscience, 6, 674-681.
Rammsayer, T. H. \& Brandler, S. (2007). Performance on temporal information processing as an index of general intelligence. Intelligence, 35, 123-139.
Rammsayer, T., \& Troche, S. (2010). Sex differences in the processing of temporal information in the sub-second range. Personality and Individual Differences, 49, 923927.

Slevc, L. R., Miyake, A. (2006). Individual differences in second-language proficiency: Does musical ability matter? Psychological Science, 17, 675-681.
Strait, D. L., Hornickel, J. \& Kraus, N. (2011) Subcortical processing of speech regularities underlies reading and music aptitude in children. Behavioral and Brain Functions, 7.
Tallal, P. (1980). Auditory temporal perception, phonics, and reading disabilities in children. Brain and Language, 9, 182-198.
Williamson, V. J., Baddeley, A. D. \& Hitch, G. J. (2010). Musicians' and nonmusicians' short-term memory for verbal and musical sequences: Comparing phonological similarity and pitch proximity. Memory \& Cognition, 38, 163-175.
Wing, H. D. (1968). Tests of musical ability and appreciation: An investigation into the measurement, distribution, and development of musical capacity (2nd ed.). London: Cambridge University Press.
Wittmann, M., \& Szelag, E. (2003). Sex differences in perception of temporal order. Perceptual and Motor Skills, 96, 105-112.
Wong, P. C., Skoe, E., Russo, N. M., Dees, T., Kraus, N. (2007). Musical experience shapes human brainstem encoding of linguistic pitch patterns. Nature Neuroscience, 10, 420-422.
Zatorre, R. J., Belin, P. \& Penhune, V. B. (2002). Structure and function of auditory cortex: music and speech. Trends in Cognitive Sciences, 6, 37-46.

# Maximizing Students' Retention Via Spaced Review: Practical Guidance From Computational Models Of Memory 

Mohammad Khajah (mohammad.khajah@colorado.edu) Robert Lindsey (robert.lindsey@colorado.edu)<br>Michael Mozer (mozer@colorado.edu)<br>Department of Computer Science<br>University of Colorado<br>Boulder, CO 80309-0430


#### Abstract

During the school semester, students face an onslaught of new material. Students work hard to achieve initial mastery of the material, but soon their skill degrades or they forget. Although students and educators both appreciate that review can help stabilize learning, time constraints result in a trade off between acquiring new knowledge and preserving old knowledge. To use time efficiently, when should review take place? Experimental studies have shown benefits to long-term retention with spaced study, but little practical advice is available to students and educators about the optimal spacing of study. The dearth of advice is due to the challenge of conducting experimental studies of learning in educational settings where material is introduced in blocks over the time frame of a semester. In this paper, we turn to two established models of memory-ACT-R and MCM-to conduct simulation studies exploring the impact of study schedule on long-term retention. Based on the premise of fixed time each week to review, converging evidence from the two models suggests that an optimal review schedule obtains significant benefits over haphazard (suboptimal) review schedules. Further, we identify two scheduling heuristics that obtain near optimal review performance: (1) review the material from $\mu$-weeks back, and (2) review material whose predicted memory strength is closest to $\theta$. The former has implications for classroom instruction and the latter for the design of electronic tutors.


Keywords: spacing effect; memory model; ACT-R, MCM, optimization, learning, review

## Introduction

At every level of the educational system, from grade school through college through professional school, instructors and textbooks typically introduce students to new material in blocks. These blocks-sometimes called sections or units-represent conceptually coherent chunks of knowledge. For example, in a foreign language class, students may learn conversational skills concerning foods and restaurants one week, traveling the next week, and vacation activities the following week. In medical school, students may study vascular, pulmonary, and renal systems in consecutive months.

At the end of each block, teachers typically administer a quiz or assign a problem set to encourage students to master the material in the block. Because the students are rewarded for focusing on this task, they have little incentive at that moment to rehearse and practice material they have learned previously. As a result, forgetting is inevitable. Although anyone who has taught a
class appreciates the need for review, the time demands of review of old material must be balanced against the need to introduce new material, explain concepts, and encourage students toward initial mastery.

Achieving this balance requires an understanding of when students will most benefit from review. Reviewing material when it is fresh provides minimal benefit; however, waiting until material has been forgotten is also costly because the earlier study provides little benefit. A long history of research in experimental psychology has shown that the temporal distribution or spacing of study has a substantive impact on long-term retention. Selecting the ideal spacing of study can lead to nearly doubling retention of material on an educationally relevant time scale of a year (Cepeda, Vul, Rohrer, Wixted, \& Pashler, 2008). Evidence for the benefit of spaced study is found not only in the domain of declarative learning, but in conceptual understanding and cognitive skill acquisition (Carpenter et al., in press), and spacing manipulations have been shown to be effective in the classroom (e.g., Sobel, Cepeda, \& Kapler, 2011).

The goal of this paper is to leverage computer simulations to offer educators practical guidance about the optimal spacing of review in the context of a semesteror quarter-long class. In such a context, we assume that the class is divided into blocks, new material is introduced in each block, and some time during each block is allotted for review of old material. The issue at hand is what material should be reviewed and when. To state the issue formally, suppose that a semester consists of $B$ blocks, and in block $b, b=1 \ldots B$, the opportunity exists to review material from $N$ previous blocks, denoted $R_{b, n}, 1 \leq R_{b, n}<b$ and $n=1 \ldots N$. What review schedule, $\mathbf{R} \equiv\left\{R_{b, n}\right\}$, will maximize the students' memory for material following some retention interval $R I$ weeks after the end of the semester?

Conducting controlled experimental studies to answer this question is not feasible. Even if the opportunity is afforded for the review of only one $(N=1)$ block, the number of review schedules is $1 \times 2 \times \ldots \times(B-1)=$ $(B-1)$ !, and the combinatorics get worse for larger $N$. A typical high-school semester or a typical college quarter may have $B=10$ weeks of new material, for which $9!=362880$ possible review schedules exist.

Although the number of candidate schedules could be greatly pruned, it would be a significant undertaking to conduct an experimental study comparing even two alternative schedules over a time window spanning ten study blocks and a subsequent final evaluation.

Because of the difficulty in conducting multi-session studies over extended time periods, nearly all prior research on spacing has either focused on the case of two study sessions or spanned such a compressed time scale that its educational relevance is questionable. (Kang, Lindsey, Mozer, \& Pashler, submitted, offer a contrasting example.) Without recourse to controlled laboratory studies, one might conclude that cognitive science has little to offer educators beyond the qualitative advice to review material occasionally.

However, a trustworthy computational model can be used to optimize study, i.e., to search for the study schedule that will maximize student retention at some specified point or time window in the future. The cost of predicting performance with a computational model under a given study schedule is negligible relative to the cost of conducting a behavioral experiment. In past work, we've shown the potential benefits of optimizing study via a cognitive model (Lindsey, Mozer, Cepeda, \& Pashler, 2009). In the present work, we use models to explore a range of scheduling algorithms in order to identify both optimal schedulers and heuristic schedulers that well approximate the optimum in an extended classroom setting.

## Spaced Study And Memory Models

The spacing effect has been investigated for over a hundred years (Ebbinghaus, 1885/1964), and in additional to qualitative theories, many mathematical and computational models have been proposed to explain the phenomenon (e.g., Benjamin \& Tullis, 2010; Raaijmakers, 2003). However, two recent efforts have been fairly comprehensive in obtaining quantitative fits to data and both have shown promise in predicting the outcome of experimental studies: an extension of the ACT-R model of memory (Pavlik \& Anderson, 2005, 2008), and a model we developed called the Multiscale Context Model or MCM (Mozer, Pashler, Cepeda, Lindsey, \& Vul, 2009). We summarize the two models and then turn to using the models as a proxy for human performance to predict the optimal spacing of study. Lindsey et al. (2009) compared qualitative predictions of ACT-R and MCM in hypothetical situations, and the models gave some contrasting results. However, these earlier simulation studies did not explore the predictions of the models in a practical educationally relevant setting.

## ACT-R

ACT-R (Anderson et al., 2004) is an influential cognitive architecture whose declarative memory module is often
used to account for explicit recall following study. ACT$R$ assumes that a separate trace is laid down each time an item is studied, and the trace decays according to a power law, $t^{-d}$, where $t$ is the age of the memory and $d$ is the power law decay for that trace. Following $n$ study episodes, the activation for an item, $m_{n}$, combines the trace strengths of individual study episodes:

$$
m_{n}=\ln \left(\sum_{k=1}^{n} b_{k} t_{k}^{-d_{k}}\right)+\beta
$$

where $t_{k}$ and $d_{k}$ refer to the age and decay associated with trace $k$, and $\beta$ is a student- and/or item-specific parameter that influences memory strength. The variable $b_{k}$ reflects the salience of the $k$ th study session (Pavlik, 2007): larger values of $b_{k}$ correspond to cases where, for example, the participant self-tested and therefore exerted more effort.

To explain spacing effects, Pavlik and Anderson (2005; 2008) made an additional assumption: the decay for the trace formed on study trial $k$ depends on the item's activation at the point when study occurs:

$$
d_{k}\left(m_{k-1}\right)=c e^{m_{k-1}}+\alpha
$$

where $c$ and $\alpha$ are constants. If study trial $k$ occurs shortly after the previous trial, the item's activation, $m_{k-1}$, is large, which will cause trace $k$ to decay rapidly. Increasing spacing therefore benefits memory by slowing decay of trace $k$. However, this benefit is traded off against a cost incurred due to the aging of traces $1 \ldots k-1$ that causes them to decay further. The probability of recall is monotonically related to activation:

$$
p(m)=1 /\left(1+e^{\frac{\tau-m}{s}}\right)
$$

where $\tau$ and $s$ are additional parameters. In total, the variant of the model described here has six free parameters.

Pavlik and Anderson (2008) use ACT-R activation predictions in a heuristic algorithm for within-session scheduling of trial order and trial type (i.e., whether an item is merely studied, or whether it is first tested and then studied). They assume a fixed spacing between initial study and subsequent review. Thus, their algorithm reduces to determining how to best allocate a finite amount of time within a session. Although they show an effect of the algorithm used for within-session scheduling, we focus on the complementary issue of between-session scheduling. The between-session manipulation has a far greater impact on long-term retention (Cepeda, Pashler, Vul, Wixted, \& Rohrer, 2006).

## MCM

ACT-R is posited on the assumption that memory decay follows a power function. We developed an alternative model, the Multiscale Context Model or MCM (Mozer et al., 2009), which provides a mechanistic basis
for the power function. Adopting key ideas from previous models of the spacing effect (Kording, Tenenbaum, \& Shadmehr, 2007; Raaijmakers, 2003; Staddon, Chelaru, \& Higa, 2002) MCM proposes that each time an item is studied, it is stored in multiple item-specific memory traces that decay at different rates. Although each trace has an exponential decay, the sum of the traces decays approximately as a power function of time. Specifically, trace $i$, denoted $x_{i}$, decays over time according to:

$$
x_{i}(t+\Delta t)=x_{i}(t) \exp \left(-\Delta t / \tau_{i}\right)
$$

where $\tau_{i}$ is the decay time constant, ordered such that successive traces have slower decays, i.e., $\tau_{i}<\tau_{i+1}$. Traces $1-k$ are combined to form a net trace strength, $s_{k}$, via a weighted average:

$$
s_{k}=\frac{1}{\Gamma_{k}} \sum_{i=1}^{k} \gamma_{i} x_{i}, \text { where } \Gamma_{k}=\sum_{i=1}^{k} \gamma_{i}
$$

and $\gamma_{i}$ is a factor representing the contribution of trace $i$. In a cascade of $K$ traces, recall probability is simply the thresholded strength: $P($ recall $)=\min \left(1, s_{K}\right)$.

Spacing effects arise from the trace update rule, which is based on Staddon et al. (2002). A trace is updated only to the degree that it and faster decaying traces fail to encode the item at the time of study. This rule has the effect of storing information on a time scale that is appropriate given its frequency of occurrence in the environment. Formally, when an item is studied, the increment to trace $i$ is negatively correlated with the net strength of the first $i$ traces, i.e.,

$$
\Delta x_{i}=\epsilon\left(1-s_{i}\right)
$$

where $\epsilon$ is a step size. We adopt the retrieval-dependent update assumption of Raaijmakers (2003): $\epsilon=1$ for an item that is not recalled at the time of study, and $\epsilon=\epsilon_{r}$ $\left(\epsilon_{r}>1\right)$ for an item that is recalled.

The model has only 5 free parameters $\left(\epsilon_{r}\right.$, and 4 parameters that determine the contributions $\left\{\gamma_{i}\right\}$ and the time constants, $\left\{\tau_{i}\right\}$ ). MCM was designed such that its parameters could be fully constrained by data that are easy to collect-the function characterizing forgetting folllowing a single study session-which then allows the model to make predictions for data that are difficult to collect-the function characterizing forgetting following a study schedule consisting of two or more study sessions. MCM has been used to obtain parameter-free predictions for various results in the spacing literature.

## Methodology

## Model Parameterization

Different parameterizations of ACT-R and MCM are critical to accounting for a range of learning scenariosscenarios that encode the ability and background knowledge of students, the difficulty of material, the manner of


Figure 1: The 105 power-function forgetting curves used to represent a diversity of learning scenarios (i.e., learning tasks varying in material difficulty, student ability, manner of study, and potential interference.
study, and the degree to which previously learned material interferes with or facilitates the learning of new material. Because our goal is to obtain results that have some generality across scenarios, we simulate a wide range of scenarios and base our results on the average over scenarios. We summarize the many factors that comprise a scenario in terms of a forgetting curve, which specifies the probability that material learned in a single study session will be available at some later point in time. Figure 1 shows a family of 105 forgetting curves, all of which decay according to a power function of time. This family expresses a diverse range of naturally occurring degrees of forgetting.

For MCM, we search for model parameters that well approximate each forgetting curve. MCM has five free parameters, one of which $\left(\epsilon_{r}\right)$ was set based on previous simulations, and the other four of which directly determine and are fully constrained by the shape of the forgetting curve. For ACT-R, we fixed $b_{k}=1$, but because its remaining free parameters are not fully constrained by the forgetting curve, we used the parameterized MCM to generate data which was then used to fit ACT-R parameters, ensuring that matched parameter sets had a loose correspondence. The generated data consisted of two study sessions with intersession intervals ranging from minutes to weeks, and a subsequent final test days to months later. This procedure yielded 105 matched instantiations of MCM and ACT-R, reflecting a wide range of scenarios.

## Simulated Learning Experiment

We conducted separate simulations of MCM and ACT$R$ to model the performance of a student learning new material in each of $B=10$ weekly blocks. We assumed homogeneity of material in a block, allowing the block's material to be distilled into a single item for the purpose of the simulation. Initial study was simulated as a single training trial to the model, though this training trial - and the corresponding memory trace - is intended to correspond to the net effect of concentrated study over multiple trials by a student learner.

Review was included in the curriculum starting after a



Figure 2: (left, middle panels) Activation trace from MCM for 10 blocks of material for good and poor review schedules. (right panel) Predicted performance on cumulative exam as a function of week in semester for alternative review schedules.
$D$-week delay. Review consists of selecting one previous block's material and presenting it as a training trial to the model. We simulated the $(B-1)!/(D-1)$ ! distinct review schedules. We allowed $D$ to vary because when review begins earlier in the semester, the number of sensible review schedules significantly shrinks. For example, with $D=1$, the only option for week 2 review is week 1 ; this selection has consequences the next week because in week 3 , review of week 1 again adds little benefit, so a sensible option is to review week 2 ; and so forth.

To evaluate the effectiveness of a review schedule, mean recall accuracy over the $B$ blocks was assessed by querying the model with a final recall test following a retention interval of $R I$ weeks past the end of the semester.

## Alternative Review Schedulers

To summarize, we consider two models of human learning (ACT-R and MCM), 105 scenarios (model parameterizations), 3 retention intervals ( $R I=1,4,26$ weeks), and 3 review delays ( $D=1,2,3$ weeks), for a total of 1890 distinct combinations. For each combination, we conducted an exhaustive search through the set of distinct review schedules to determine the optimal schedule - the schedule that yields the highest average accuracy on the final test according to the model.

In addition, we considered various heuristic schedulers. Our goal is to identify heuristics that produce a close-to-optimal schedule. The two best heuristic schedulers were as follows. A $\mu$-back scheduler follows a simple rule: in week $i$, review material from week $\max (1, i-\mu)$. A $\theta$-threshold scheduler is motivated by Bjork's (1994) notion of desireable difficulty - that material should be restudied as it is on the verge of being forgotten. Using a memory model to determine the strength of each week's material, this scheduler selects the material whose recall probability closest to $\theta$. Because we use the same model for scheduling as we use for modeling the student, this scheduler offers a best-case use of the $\theta$-threshold. (We also explored several variants of the threshold scheduler which yielded poorer performance. One variant uses a scaled threshold rule whereby the threshold value is relative to the range of performance over all weeks' material. Another uses an asymmetric threshold where the selection is for material whose recall probability is close to the threshold on one side - either above or below.)

## Results

Figure 2 provides an intuition about the operation of our model-based scheduling. The left panel of the Figure shows ten curves, each representing the memory strength predicted by MCM for one block of material as a function of weeks into the semester. The color coding from red to blue indicates blocks $1-10$, respectively. In this example, block $i$ is introduced in week $i$ and is then reviewed in week $i+1$. As a result, the block gets a bump in strength in weeks $i$ and $i+1$, and then decays from that point on. The curves in the Figure represent the average over the 105 learning scenarios, and the ordinate of the graph shows the expected recall probability over these scenarios. The absolute probability is immaterial and is a consequence of the specific scenarios we simulate. However, relative probabilities matter. To emphasize this point, the middle panel of the Figure shows an activation trace for an arbitrary and somewhat bad review schedule. The right panel shows the same time history of activation, but averaged over the individual blocks to obtain a prediction of cumulative-exam performance (weighting all blocks equally) at a given time. The superiority of the one-back schedule (left panel) over the arbitrary (middle panel) is reflected in a higher average recall probability. Four weeks following the end of the 10 -week semester, the better review schedule achieves a $89.7 \%$ improvement in retention over no review, and a $16.1 \%$ improvement in retention over the poorer quality review schedule.

## Exhaustive Search Of Alternative Schedules

Figure 3 shows a set of curves that reflect the expected perfomance of all possible review schedules for a given simulation, sorted from worst to best. The average is taken over learning scenarios. Each graph shows three simulations, one per retention interval $(R I=1,4,26$ weeks). The top and bottom rows are simulations of MCM and ACT-R, respectively. The columns from left-to-right correspond to simulations in which review begins following weeks 1,2 , and $3(D=1,2,3)$. The colored squares on the left of each graph indicate the performance of a 'no review' condition for the retention interval of the corresponding color. Not surprisingly, all review schedules are superior to no review, and well-timed review is as much as $33 \%$ better than poorly-timed review.


Figure 3: Each curve is the sorted performance of all possible review schedules (The x-axis is an index over distinct review schedules). Each curve corresponds to one retention interval (RI). Top and bottom rows are simulations of MCM and ACT-R, respectively. The columns correspond to simulations in which review onset begins following weeks 1,2 , and 3 , respectively. Colored squares indicate the performance of a 'no review' schedule. Digits indicate the performance of the heuristic $\mu$-back scheduler, for $\mu=1,2,3$. The green disk (.) indicates the performance of the $\theta$ threshold scheduler for optimal $\theta$.


Figure 4: (a) Relative performance predicted by MCM and ACT-R for the $\theta$-threshold heuristic as a function of $\theta$ (for $D=1, R I=1$ ). (b) Relative performance of the 2-back schedule over all learning scenarios, sorted from best to worst (for $D=1, R I=1$ ). In both graphs, performance is relative to the exhaustive space of schedules.

## $\theta$-Threshold Heuristic Scheduler

The key result in Figure 3 concerns the performance of heuristic schedulers relative to the optimum schedule discovered by exhaustive search. For each curve, the location of the green disk indicates the relative ranking of the $\theta$-threshold schedule, for the best setting of $\theta$. The further to the right along the x -axis, the higher the ranking. The two models are consistent in predicting that the $\theta$-threshold scheduler is as good or nearly as good as the best schedule found by exhaustive search. Figure 4a shows how the predicted performance varies as a function of $\theta$ for the two models, for a delay of $D=1$ week and a retention interval of $R I=1$ week. The ordinate indicates the relative performance in the range defined by the complete space of schedules, where $100 \%$ and $0 \%$ correspond to the best and worst sched-
ules found by exhaustive search, respectively. Notably, the two models yield very similar curves, and although the $\theta$-threshold scheduler does not produce the very best schedule, it comes reasonably close. Notably, both MCM and ACT-R are consistent in indicating that a threshold in the neighborhood of $\theta=.4$ is best. We have shown the curve for $D=1$ and $R I=1$, but curves for the other values of $D$ and $R I$ are quite similar, and all have the same optimum for $\theta$.

The limitation of a threshold scheduler is that it requires an accurate model to predict memory strength as a function of time given some history of study. In our simulation, we've assumed that the model we use for determining memory strength - either MCM or ACT-R - is a veridical model of our (simulated) student. An important question for future research concerns how the accuracy of the model used for scheduling affects the performance of the $\theta$-threshold scheduler. However, it is clear that whatever model is used must take into account the history and spacing of past study, because the effect of distributed practice - as embodied in both MCM and ACT-R - is central to the difference in performance across review schedules.

## $\mu$-Back Heuristic Scheduler

Figure 3 also depicts the performance of the 1-, 2-, and 3 -back schedules, all of which do reasonably well across models, delays, and retention intervals. However, because ACT-R predicts the 1-back schedule to be inferior for $D=2,3$, and because MCM predicts the 3 -back schedule to be slightly worse for $D=1,2$, we suggest that the $\mu=2$, or the 2 -back schedule, might be adopted as a robust solution across conditions.

All results we've presented to this point are the average over the 105 learning scenarios. It's possible that
the $\mu$-back schedules work well on average but not for specific scenarios. To examine the performance of the 2-back schedule across scenarios, Figure 4b shows the performance in each scenario, sorted from best to worst. The curves for MCM and ACT-R are remarkably similar, and indicate that the 2-back schedule performs well for the majority ( $60-80 \%$ ) of scenarios we considered, further supporting our claim of its robustness.

## Discussion

In a metaanalysis of the experimental literature, the optimal spacing of study was found to grow monotonically with the retention interval (Cepeda et al., 2006). Although in past work we've shown that MCM and ACT-R both predict this characteristic, neither model strongly predicts that the best $\mu$ in the $\mu$-back scheduler should increase with the retention interval (Figure 3). Most likely, this inconsistency is due to the fact that as $\mu$ increases, the initial $\mu+1$ weeks of study become focused on the first week's material, and there are diminishing returns of this focus. Consequently, the benefits of increased spacing must be outweighed by the cost of illspent review time. This result suggests to us the importance of moving beyond laboratory studies of spacingtypically with two study sessions and a single block of material to be learned-to situations more reflective of real-world educational constraints, i.e., semesters in which multiple blocks of material are presented staggered in time and initial study must be interlaced with review.

Our results provide practical guidance to educators: To preserve learning beyond the end of a semester, a 2-back review schedule should generally be appropriate. Although classroom teachers do not have access to mathematical models of human memory, and therefore cannot exploit the $\theta$-threshold scheduler, we see great potential of incorporating model-based scheduling into electronic tutors used in synchronization with classroom instruction (Lindsey et al., in preparation). Indeed, such an approach opens the possibility to personalized review appropriate for a specific student rather than a one-size-fits-all approach. Our caveat in suggesting this approach is that it requires accurate psychological models of memory. Models based on intuition-as embodied in existing web-based flashcard apps-are unlikely to be adequate.

## Acknowledgments

This research was supported by the UCSD Temporal Dynamics of Learning Center (NSF award SBE-0542013, Garrison Cottrell, PI), and by NSF award BCS-0720375.

## References

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., \& Qin, Y. (2004). An integrated theory of mind. Psych. Rev., 111, 1036-1060.
Benjamin, A. S., \& Tullis, J. (2010). What makes distributed practice effective? Cog. Psych., 61, 228-247.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe \& A. Shimamura (Eds.), Metacognition: Knowing about knowing (pp. 185-205). Cambridge, MA: MIT Press.
Carpenter, S. K., Cepeda, N. J., Rohrer, D., Kang, S. H., \& Pashler, H. (in press). Using spacing to enhance diverse forms of learning: Review of recent research and implications for instruction. Ed. Res.
Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., \& Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. Psych. Bull., 132, 354-380.
Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., \& Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. Psych. Sci., 19, 1095-1102.
Ebbinghaus, H. (1885/1964). Memory: A contribution to experimental psychology. New York: Dover.
Kang, S. H. K., Lindsey, R. V., Mozer, M. C., \& Pashler, H. (Submitted). Retrieval practice over the long term: Expanding or equal-interval spacing?
Kording, K. P., Tenenbaum, J. B., \& Shadmehr, R. (2007). The dynamics of memory are a consequence of optimal adaptation to a changing body. Nature Neuroscience, 10, 779-786.
Lindsey, R., Mozer, M. C., Cepeda, N. J., \& Pashler, H. (2009). Optimizing memory retention with cognitive models. In A. Howes, D. Peebles, \& R. Cooper (Eds.), Proc. 9th intl. conf. on cog. modeling (ICCM). Manchester, UK.
Mozer, M. C., Pashler, H., Cepeda, N., Lindsey, R., \& Vul, E. (2009). Predicting the optimal spacing of study: A multiscale context model of memory. In Advances in neural information processing systems 22 (pp. 1321-1329). La Jolla, CA: NIPS Foundation.
Pavlik, P. I. (2007). Understanding and applying the dynamics of test practice and study practice. Instruc. Sci., 35, 407-441.
Pavlik, P. I., \& Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: An activationbased model of the spacing effect. Cognitive Science, 29(4), 559-586.
Pavlik, P. I., \& Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. J. Exptl. Psych.: Applied, 14, 101-117.
Raaijmakers, J. G. W. (2003). Spacing and repetition effects in human memory: application of the SAM model. Cognitive Science, 27, 431-452.
Sobel, H. S., Cepeda, N. J., \& Kapler, I. V. (2011). Spacing effects in real-world classroom vocabulary learning. Appl. Cog. Psych., 25, 763-767.
Staddon, J. E. R., Chelaru, I. M., \& Higa, J. J. (2002). Habituation, memory and the brain: The dynamics of interval timing. Behav. Proc., 57, 71-88.

# Spatial terms across languages support near-optimal communication: Evidence from Peruvian Amazonia, and computational analyses 

Naveen Khetarpal (khetarpal@uchicago.edu) ${ }^{\text {a }}$ Grace Neveu (gkneveu@berkeley.edu) ${ }^{\text {b }}$<br>Asifa Majid (a.majid@let.ru.nl) ${ }^{\text {c }}$<br>Lev Michael (levmichael@berkeley.edu) ${ }^{\text {b }}$<br>Terry Regier (terry.regier@berkeley.edu) ${ }^{\text {b,d }}$<br>${ }^{\text {a Department of Psychology, University of Chicago, Chicago, IL } 60637 \text { USA }}$<br>${ }^{\mathrm{b}}$ Department of Linguistics, University of California, Berkeley, Berkeley, CA 94720 USA<br>${ }^{\text {c }}$ Center for Language Studies, Radboud University, Nijmegen, Netherlands<br>${ }^{\mathrm{d}}$ Cognitive Science Program, University of California, Berkeley, Berkeley, CA 94720 USA


#### Abstract

Why do languages have the categories they do? It has been argued that spatial terms in the world's languages reflect categories that support highly informative communication, and that this accounts for the spatial categories found across languages. However, this proposal has been tested against only nine languages, and in a limited fashion. Here, we consider two new languages: Maijiki, an under-documented language of Peruvian Amazonia, and English. We analyze spatial data from these two new languages and the original nine, using thorough and theoretically targeted computational tests. The results support the hypothesis that spatial terms across dissimilar languages enable near-optimally informative communication, over an influential competing hypothesis.


Keywords: Spatial terms; semantic universals; informative communication; language and thought; semantic maps.

## Spatial categories across languages

Spatial terms across languages often pick out different categories, as illustrated in Figure 1. Yet at the same time similar or comparable categories often recur across unrelated languages.


Figure 1: 10 spatial scenes, as categorized in 2 languages: Tiriyó and Yélî-Dnye. Source: Levinson et al. (2003).

A central question in cognitive science is why languages have the categories they do - in this case, why spatial categories exhibit the constrained cross-language variation they do (Bowerman \& Pederson, 1992; Bowerman, 1996; Talmy, 2000; Levinson et al., 2003).

## Informative communication

Recently, an answer to this question has been proposed that is grounded in general communicative principles. Khetarpal, Majid, \& Regier (2009) argued that across languages, spatial categories are shaped by the need to support informative communication. On this view, the many different spatial systems observed across languages represent different means to this same end. This argument mirrors analogous arguments that have recently been advanced for the semantic domains of color (Regier, Kay, \& Khetarpal, 2007) and kinship (Kemp \& Regier, 2012), and also reflects a more general recent focus on informative communication as a central force that explains why languages take the forms they do (e.g. Fedzechkina, Jaeger, \& Newport, 2012; Piantadosi, Tily, \& Gibson, 2011).

Khetarpal et al. (2009) considered the 71 spatial scenes of the topological relations picture series or TRPS (Bowerman \& Pederson, 1992), illustrated in part in Figure 1, as named by speakers of 9 unrelated languages: Basque, Dutch, Ewe, Lao, Lavukaleve, Tiriyó, Trumai, Yélî-Dnye, and Yukatek (Levinson et al., 2003). Each of these languages groups TRPS scenes together into languagespecific spatial categories, and Khetarpal et al. (2009) asked whether these attested groupings support near-optimally informative communication. In a series of computational simulations, they asked whether each of these linguistic spatial systems supports informative communication better than a comparison class of hypothetical systems. They found that this is indeed the case. They concluded that spatial terms across languages reflect near-optimally informative spatial categories, and that this functional force may help to explain which spatial categories appear in the world's languages.

However, this earlier work is limited in three important respects. First, it considered data from only nine languages.

Such data are difficult and time-consuming to collect, and we are grateful to our colleagues at the MPI Nijmegen for sharing their data with us. Still, this is a very small sample, so it is possible that other languages would break the generalization made on the basis of these nine. Second, the earlier work tested the near-optimality claim against these nine languages in a narrow and limited way. Each language was compared to only 69 hypothetical systems that were intended to be comparable to it. Thus it is possible that many other, unexamined hypothetical systems may exist that are more informative than the attested system - again potentially breaking the generalization and undercutting the central theoretical claim. Third, the earlier work did not test the informativeness proposal against alternative explanations for constrained semantic variation.

Here we bring new data and analyses to bear on the claim that spatial categories across languages support informative communication, and that this force may account for the observed variation in spatial systems. The new data are from Maijiki, an under-documented language of Peruvian Amazonia, and English. The new analyses compare eleven languages (Maijiki, English, and the nine languages from Levinson et al., 2003) to much larger and more theoretically targeted sets of hypothetical systems. Critically, unlike the earlier analyses, the new analyses explicitly pit the claim of near-optimal informativeness against the competing and influential theoretical claim that semantic categories tend to pick out connected regions of conceptual or perceptual space (e.g. Croft, 2003; Haspelmath, 2003; Roberson, Davies, \& Davidoff, 2000; Roberson, 2005).

In what follows we first describe Maijiki and its spatial system, comparing it with that of English. We then lay out the hypotheses to be tested, our analyses of the eleven languages under consideration, and the results of these analyses. We conclude from these results that spatial systems across languages do indeed reflect near-optimally informative categories, and that this proposal is supported over the competing claim that categories pick out connected regions of conceptual or perceptual space. We suggest that the functional goal of informative communication may account for the wide but constrained variation found in spatial systems across languages.

## Maijiki

Maijiki is an under-documented Western Tukanoan language of Peruvian Amazonia, spoken in the departmento of Loreto, near the Colombian-Peruvian border. The language is spoken by approximately 100 individuals, of whom some 25 are Maijiki-dominant, although there are no monolingual speakers. The language is currently being documented as part of the Maijiki Project, a multi-year effort to produce a grammar, text collection, and dictionary of the language (Michael, Beier, \& Farmer, 2012). Maijiki is unrelated to the other languages that we consider in this paper.

The spatial system of Maijiki has only recently been investigated, and is described in detail by Neveu and

Michael (in preparation). Spatial meanings are conveyed in Maijiki by several means, including spatial adpositions and spatial verbs. For simplicity we focus on the major spatial adpositions, listed in Table 1 (tone marks are suppressed here and elsewhere in this paper).

Table 1: Spatial adpositions in Maijiki.

| Adposition | Approximate meaning |
| :---: | :--- |
| guibi | under |
| gunu | near an edge |
| imijai | on top or above |
| jeteruru | behind |
| sanu | inside at bottom |

The extensions of these Maijiki spatial adpositions are illustrated in Figure 2 below, as subsets of the full set of 71 scenes in the TRPS. Also shown for comparison are spatial categories in English. In each of the 71 scenes, the figure object is shown in orange, the ground object in black, and the corresponding spatial meaning is the spatial relation between the figure and the ground. As can be seen, the spatial categories of Maijiki differ from those of English. We seek general principles that help to determine which logically possible groupings of scenes constitute categories that are attested in the world's languages.

## Hypotheses

We consider two hypotheses, which our analyses pit against each other, using data from Maijiki, English, and the nine languages of Levinson et al. (2003).

## Near-optimally informative communication

The first hypothesis is the one sketched above: that spatial categories across languages appear as they do because these categories maximize or near-maximize the informativeness of communication. We take a communicative system to be informative to the extent that it supports accurate mental reconstruction by a listener of a speaker's intended meaning (cf. communication accuracy: Lantz \& Stefflre, 1964). This general idea, which also applies to other semantic domains, can be made concrete through the following communicative scenario.

A speaker has a particular spatial relation in mind, and wishes to communicate it to a listener. To that end, the speaker produces a spatial term that describes this spatial relation. The listener must then mentally reconstruct the original spatial relation that the speaker intended, from the term used. Because the listener knows only that the intended spatial relation falls in the general category named by the spatial term, the listener's mental reconstruction is the set of all spatial relations that are named by the term. We define the reconstruction accuracy to be the similarity of this mental reconstruction to the original intended spatial relation. In general, we hold that informative categories, and informative systems of categories, are those that support high reconstruction accuracy.


Figure 2: A semantic map showing spatial categories from Maijiki (red) and English (blue). Categories that appear in both languages are shown in black. Links connect scenes that are presumed to be universally related across languages. All displayed categories in both Maijiki and English pick out connected regions of the map.

We formalize these ideas as follows. ${ }^{1}$ Let $S$ be the set of all possible spatial relations (here approximated by the spatial scenes of the TRPS, or the subset of those scenes that are assigned names by a given language). Let $\operatorname{sim}(x, y)$ be the similarity between two spatial relations $x$ and $y$ (here, similarity is gauged empirically as described below, and ranges from $0=$ completely dissimilar to $1=$ maximally

[^114]similar). Let $s$ be the specific spatial relation the speaker intends to convey, let $t$ be the spatial term used to describe that spatial relation, and let $\operatorname{cat}(t)$ be the category or set of all spatial relations described by $t$, including $s$. Finally, let $\operatorname{era}(s)$ be the expected reconstruction accuracy of scene $s$, i.e. the similarity between the target spatial relation $s$ and the listener's reconstruction of that spatial relation, based on the speaker's spatial term $t$. This is the average, over all spatial relations $r$ in the same named category $\operatorname{cat}(t)$ as $s$, of the similarity between $r$ and $s$ :
\[

$$
\begin{equation*}
\operatorname{era}(s)=\frac{1}{|\operatorname{cat}(t)|} \sum_{r \in \operatorname{cat}(t)} \operatorname{sim}(r, s) \tag{1}
\end{equation*}
$$

\]

The overall expected accuracy of reconstruction, over all possible stimuli, is then given by:

$$
\begin{equation*}
R=\frac{1}{|S|} \sum_{s \in S} e r a(s) \tag{2}
\end{equation*}
$$

$R$ is a measure of how well a given communicative system supports informative communication. The first hypothesis we consider is that attested linguistic spatial systems will tend to exhibit high $R$, compared with hypothetical systems.

## The semantic map connectivity hypothesis

The second hypothesis we consider holds instead that attested categories pick out connected regions of a universal network of meanings called a semantic map (e.g. Croft, 2003; Haspelmath, 2003). Figure 2, in which we saw the spatial systems of Maijiki and English, is an example of a semantic map. Here the meanings are spatial meanings, represented by the spatial scenes of the TRPS. These spatial meanings are assumed to be universally available, and the links in the network represent presumed universally available connections between closely related spatial meanings. As we have seen, different languages often group these meanings into categories differently, and these language-specific groupings are also displayed in the map. Thus a semantic map represents both presumed universal semantic structure and language-specific parcelings of that structure.

The core idea behind a semantic map is that across languages, semantic categories will always pick out connected regions of the network. In other words, a category should correspond to a group of meanings (here, scenes) that are connected in the sense that one may travel from any meaning in the category to any other by repeatedly traversing links in the network. The semantic map in Figure 2 was inferred automatically (Regier, Khetarpal, \& Majid, in press) to accommodate, as connected regions, the spatial categories of the nine languages of Levinson et al. (2003). As can be seen, this network generalizes well to Maijiki and English: all the displayed Maijiki and English spatial categories also pick out connected regions of this map, although Maijiki and English were not considered in its construction. ${ }^{2}$ This fact suggests that the inferred universal structure of this semantic map, and the criterion of connectedness implicit in it, may in fact be an important constraint on semantic categories across languages. Similar ideas emphasizing the importance of connectedness as a determinant of what makes a good or natural category may also be found elsewhere (e.g. Levinson et al., 2003; Roberson, Davies, \& Davidoff, 2000; Roberson, 2005).

[^115]
## Goal of our analyses

It has been previously suggested (e.g. Croft, 2003: 138; Cysouw, 2001: 609; Regier et al., in press) that connectedness in a semantic map may be too loose a constraint on category shape, in part because it allows elongated categories with no clear central region; thus, semantic categories in actuality may tend to be more compact and coherent than is suggested by this constraint alone. However it has not yet been determined whether informativeness provides a better account of cross-language variation in semantic systems. The analyses we present below seek to answer this open question, by deliberately pitting informativeness and connectedness against each other.

## Analyses

We reasoned with the following predictions. The informativeness hypothesis predicts that attested linguistic spatial systems will support informative communication more effectively than almost all hypothetical systems - even if those hypothetical systems all pick out connected regions of a semantic map. The connectedness hypothesis in contrast does not make this prediction. Instead, on that hypothesis, it is connectedness rather than informativeness that plays a privileged role in determining which possible systems are actually attested - and so the informativeness of an attested linguistic spatial system should not tend to be any greater than the informativeness of other connected hypothetical systems.

For this reason, in our analyses we compared the informativeness of an actual linguistic spatial system with that of hypothetical variants, all of which correspond to connected regions of the semantic map of Figure 2. If informativeness is a major determinant of attested category systems, we expect the actual linguistic spatial system to support informative communication better than the connected hypothetical variants.

## Crawling a semantic map

We generated hypothetical connected variants of existing systems by randomly "crawling" a semantic map, by analogy with web-crawling - that is, through random graph traversal of a semantic map. We began with the semantic map in Figure 2, but with no labels assigned to the scenes. Then, for a given target language (e.g. English), we construct a hypothetical connected variant of that language as follows. Start by randomly selecting one the spatial terms in the language-call this term $t$ and the number of scenes associated with it $k$. Now randomly select one of the scenes in the graph and label it $t$. Then select another scene at random from the set of as-yet-unlabeled scenes directly connected to some scene already labeled $t$, and label that new scene $t$ as well; if there are no such scenes from which to select, the procedure terminates and begins again with no labels on any nodes. This step of extending the label $t$ to neighboring scenes is repeated until there are $k$ scenes
associated with $t$. The process as a whole is repeated for all terms in the language.

## Methods

We conducted semantic-map-crawling analyses separately for each of the eleven languages under consideration: Maijiki, English, Basque, Dutch, Ewe, Lao, Lavukaleve, Tiriyó, Trumai, Yélî-Dnye, and Yukatek. For each language, 2000 hypothetical connected variants were generated as described above, each with the same number of categories, and the same number of scenes per category, as the original. For each real or hypothetical spatial naming system, we calculated $R$, our measure of reconstruction accuracy, using equations 1 and 2 above. The categories $\operatorname{cat}(t)$ used to label specific scenes were determined by the naming system under consideration. The similarity of each pair of scenes $x$ and $y, \operatorname{sim}(x, y)$, was determined empirically by pile-sorting. Khetarpal et al. (2009) had asked speakers of English and Dutch to sort the TRPS scenes into piles on the basis of the similarity of the spatial relation portrayed, and they took the similarity of any two scenes to be the proportion of all their participants who sorted those two scenes into the same pile. ${ }^{3}$ We used the pile-sort-derived similarity judgments from that earlier study. For each language, we then compared the reconstruction accuracy $R$ for the language itself to the distribution of $R$ obtained for hypothetical connected variants of that system.

## Results

Figure 3 below presents the results of our analysis of Maijiki. The red line shows the informativeness $(R)$ of the Maijiki spatial adpositional system, and the blue histogram shows the frequency with which various values of $R$ were exhibited by hypothetical connected variants of Maijiki, obtained by randomly crawling the semantic map of Figure 2.


Figure 3: Informativeness of communication supported by the Maijiki spatial adpositional system (red line), compared with that supported by 2000 hypothetical variants derived by randomly crawling a semantic map (blue histogram).

The actual Maijiki system supports informative communication more effectively than any of the sampled

[^116]hypothetical connected variants. These results are consistent with the claim that languages tend to have highly informative spatial systems, and that informativeness is more relevant to the shape of such systems than is connectedness. Similar results from other languages would strengthen this conclusion.

Figure 4 below presents analogous results for English. Again, the actual English system supports informative communication more effectively than any of the sampled hypothetical connected variants.


Figure 4: Informativeness of communication supported by the English spatial system (red line), compared with that supported by 2000 hypothetical variants derived by randomly crawling a semantic map (blue histogram).

Finally, Table 2 below presents summary results of semantic map crawling analyses for all eleven languages we consider. In this case, the results are given numerically, as the proportion of hypothetical variants that the actual linguistic system scores higher than in $R$ (reconstruction accuracy). The results shown here for Maijiki and English summarize the results from the histograms displayed above; for the remaining nine languages, we present results in summary form only, to conserve space. In all cases, the actual linguistic system outperforms most of the sampled hypothetical connected variants, and in several cases it outperforms all of them.

Table 2: Summary results of semantic map crawling analyses for all languages considered in this study.

| Language | Result |
| :--- | :--- |
| Basque | $>99.95 \%$ |
| Dutch | $>100.00 \%$ |
| English | $>100.00 \%$ |
| Ewe | $>99.95 \%$ |
| Lao | $>96.20 \%$ |
| Lavukaleve | $>99.75 \%$ |
| Maijijki | $>100.00 \%$ |
| Tiriyó | $>100.00 \%$ |
| Trumai | $>100.00 \%$ |
| Yélî-Dnye | $>97.35 \%$ |
| Yukatek | $>99.95 \%$ |

In sum, each of the 11 languages considered supports informative communication more effectively than most sampled hypothetical variants of those systems - even when
the variants are connected regions of a semantic map. These results are consistent with the hypothesis that informativeness shapes category systems across languages, and that it does so more than connectedness in a semantic map.

## Conclusions

Our findings support the claim that spatial systems across languages reflect the need for informative communication. They do so based on new evidence, including evidence from an under-documented language, and on new large-scale analyses that directly pit informativeness against the competing claim that natural categories pick out connected regions of a semantic map.

These findings also leave a number of issues unresolved, suggesting directions for future investigation. Theoretically, our analyses have focused on the informativeness of a given system, by comparing that system to competitors of comparable complexity - thus deliberately controlling for, and not investigating, the complexity of these systems. A more complete account would investigate both informativeness and complexity, and the tradeoff between these two general forces (e.g. Kemp \& Regier, 2012). Empirically, eleven languages is still a small sample when considered relative to all existing languages. We feel that every new language considered adds important evidence, particularly under-documented languages such as Maijiki but consideration of more languages will allow more definitive conclusions.

Nonetheless, the present results lend substantial new support to the hypothesis that informativeness plays an important role in shaping spatial semantic systems across languages. In so doing, these results add to the current literature that suggests that the need for informative communication may be a key functional force that explains why languages have the forms that they do.

## Acknowledgments

This work was supported by NSF under grants SBE0541957 and SBE-1041707, the Spatial Intelligence and Learning Center (SILC). We thank Alex Carstensen for her help with the Maijiki data.

## References

Bowerman, M. (1996). Learning how to structure space for language: A cross-linguistic perspective. In P. Bloom, M. Peterson, M. Garrett, \& L. Nadel (Eds.) Language and space (pp. 385-436). Cambridge, MA: MIT Press.
Bowerman, M. \& Pederson, E. (1992). Cross-linguistic studies of spatial semantic organization. In Annual Report of the Max Planck Institute for Psycholinguistics 1992 (pp. 53-56).
Croft, W. (2003). Typology and universals: Second edition. Cambridge, UK: Cambridge University Press.

Cysouw, M. (2001). Review of Martin Haspelmath, Indefinite Pronouns (1997). Journal of Linguistics 37, 607-612.
Fedzechkina, M., Jaeger, T. F., \& Newport, E. L. (2012). Language learners restructure their input to facilitate efficient communication. PNAS early edition.
Garner, W. R. (1974). The processing of information and structure. Potomac, MD: L. Erlbaum Associates
Haspelmath, M. (2003). The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In M. Tomasello (Ed.), The new psychology of language, vol. 2 (pp. 211-242). Mahwah, NJ: Erlbaum.
Kemp, C. \& Regier, T. (2012). Kinship categories across languages reflect general communicative principles. Science, 336, 1049-1054.
Khetarpal, N., Majid, A., Malt, B., Sloman, S., \& Regier, T. (2010). Similarity judgments reflect both language and cross-language tendencies: Evidence from two semantic domains. In S. Ohlsson and R. Catrambone (Eds.), Proceedings of the 32nd Annual Meeting of the Cognitive Science Society.
Khetarpal, N., Majid, A., \& Regier, T. (2009). Spatial terms reflect near-optimal spatial categories. In N. Taatgen et al. (Eds.), Proceedings of the 31st Annual Meeting of the Cognitive Science Society.
Lantz, D., \& Stefflre, V. (1964). Language and cognition revisited. Journal of Abnormal and Social Psychology, 69, 472-481.
Levinson, S., Meira, S., \& the Language and Cognition group (2003). 'Natural concepts' in the spatial topological domain-adpositional meanings in crosslinguistic perspective: An exercise in semantic typology. Language, 79, 485-516.
Michael, L., Beier C., \& Farmer, S. (2012). Diccionario Bilingüe Maijiki-Castellano y Castellano-Maijiki. Maijiki Project.
Neveu, G., \& Michael, L. (in preparation). The semantics and pragmatics of topological spatial relations in Maijiki. Ms.
Piantadosi, S. T., Tily, H., \& Gibson, E. (2011). Word lengths are optimized for efficient communication. PNAS, 108, 3526-3529.
Regier, T., Kay, P., \& Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. PNAS, 104, 1436-1441.
Regier, T., Khetarpal, N., \& Majid, A. (in press). Inferring semantic maps. Linguistic Typology.
Roberson, D., Davies I. \& Davidoff, J. (2000). Color categories are not universal: Replications and new evidence from a stone-age culture. Journal of Experimental Psychology: General, 129, 369-398.
Roberson, D., Davidoff, J., Davies, I.R.L., \& Shapiro, L.R. (2005). Color categories: Evidence for the cultural relativity hypothesis. Cognitive Psychology, 50, 378-411.
Talmy, L. (2000). How language structures space. In L. Talmy (Ed.) Toward a cognitive semantics, Volume 1 (pp. 177-254). Cambridge, MA: MIT Press.

# Negotiation Strategies with Incongruent Facial Expressions of Emotion Cause Cardiovascular Threat 

Peter Khooshabeh (khooshabeh@ict.usc.edu) ${ }^{1,3}$<br>Celso de Melo (demelo@usc.edu) ${ }^{2}$<br>Brooks Volkman (volkman@psych.ucsb.edu) ${ }^{1}$<br>Jonathan Gratch (gratch@ict.usc.edu) ${ }^{3}$<br>Jim Blascovich (blascovi@psych.ucsb.edu) ${ }^{1}$<br>Peter J. Carnevale (carnevale@usc.edu) ${ }^{2}$<br>${ }^{1}$ Department of Psychological and Brain Sciences, University of California, Santa Barbara, CA 93106 USA<br>${ }^{2}$ Marshall School of Business, University of Southern California, Los Angeles, CA 90089<br>${ }^{3}$ Institute for Creative Technologies, University of Southern California, Los Angeles, CA 90094


#### Abstract

Affect is important in motivated performance situations such as negotiation. Longstanding theories of emotion suggest that facial expressions provide enough information to perceive another person's internal affective state. Alternatively, the contextual emotion hypothesis posits that situational factors bias the perception of emotion in others' facial displays. This hypothesis predicts that individuals will have different perceptions of the same facial expression depending upon the context in which the expression is displayed. In this study, cardiovascular indexes of motivational states (i.e., challenge vs. threat) were recorded while players engaged in a multiissue negotiation where the opposing negotiator (confederate) displayed emotional facial expressions (angry vs. happy); the confederate's negotiation strategy (cooperative vs. competitive) was factorially crossed with his facial expression. During the game, participants' eye fixations and cardiovascular responses, indexing task engagement and challenge/threat motivation, were recorded. Results indicated that participants playing confederates with incongruent facial expressions (e.g., cooperative strategy, angry face) exhibited a greater threat response, which arises due to increased uncertainty. Eye fixations also suggest that participants look at the face more in order to acquire information to reconcile their uncertainty in the incongruent condition. Taken together, these results suggest that context matters in the perception of emotion.


Keywords: facial expressions, negotiation, context in emotion

## Introduction

Negotiation is relatively common in personal and professional settings. A child might ask a parent whether she can leave the dinner table. The parent might sternly command the child to finish her vegetables and the child could make a counter offer to finish the peas but not the broccoli. This could ensue into a strategic and emotionally charged social interaction.

Emotion is an important human factor in motivated performance situations (i.e., those that are self-relevant and therefore task engaging and require instrumental cognitive responses; Blascovich, 2008). Such interactions are rarely
affectively neutral; that is, they are associated with interactants' positive or negative emotional states. Clearly, negotiations represent motivated performance situations to interested partners. And, experimental negotiation tasks are no exception, including those involving real human players (Van Kleef, De Dreu, \& Manstead, 2004) and digital agents (i.e., player representations driven by computer algorithms, de Melo, Carnevale, \& Gratch, 2011).

The current work examines individuals' motivational responses, using physiological indexes, to emotionally expressive virtual characters in a multi-issue negotiation task. Specifically, we focus on the question of how situational context affects emotion perception from facial expressions. In person-to-agent negotiation tasks, experimenters often insert communicative cues such as agent facial expressions intended to strategically manipulate user's emotions. Agents that show emotion have now been used in several domains such as education, entertainment, training, therapy and commerce (for a review see Beale \& Creed, 2009). In a multi-issue negotiation task, de Melo and colleagues (2011) reported that participants made more concessions to a virtual human that displayed an angry facial expression compared to a happy facial expression.

Most research on the effects of virtual characters' emotional facial expressions has relied on subjective responses from participants (e.g., Beale \& Creed, 2009). However, given the evidence that emotion is processed via non-conscious pathways, perhaps more so than conscious pathways (Tamietto \& De Gelder, 2010), validated physiological measures related to affect should provide confirmation of the operation of non-conscious emotional processes involved in motivated performance tasks such as negotiation (Blascovich \& Mendes, 2010).

## Psychophysiological Measurement of Motivational States

Psychophysiological research is now a well-established technique to infer peoples' affective reactions to various situations (Blascovich, Vanman, Mendes, \& Dickerson,


Figure 1: The angry (left) and happy (right) facial expressions displayed by the virtual confederate.
2011). However, a lot of research involving peripheral physiological markers has been based on unitary physiological responses such as heart rate variability (Rienerman-Jones, Cosenzo, \& Nicholson, 2010) or electrodermal activity (Meehan, Insko, Whitton, \& Brooks, 2002) mostly as indexes for workload and stress.

Motivational research suggests that relying on unitary indexes can mask important processes. For example, the physiological indexes specified by the bio-psychosocial model (BPS; Blascovich, 2008) of challenge and threat provide a much more informative index of task motivation. Briefly, the BPS model is based on the neuroendocrine underpinnings (i.e., Dienstbier, 1989) of cardiovascular responses involving the sympathetic-adrenal-medullary (SAM) axis as well as the hypothalamic pituitary-adrenalcortical (HPA) axis.

Psychologically, challenge motivation occurs when an individual's consciously and unconsciously evaluated resources outweigh evaluated task demands. Threat occurs when resources are evaluated as not meeting task demands. Both states involve the activation of the SAM axis, while only the threat state involves both the axes.

Accordingly, activation of common SAM axis sympathetic neural and adrenal medullary endocrine processes affect cardiovascular responses underlying both challenge and threat including increased heart rate (HR) and increased ventricular contractility (VC; i.e., decreased preejection period or "PEP"), and task engagement. However, cardiac output (CO) and total peripheral resistance (TPR) differ depending on motivational state. A challenge state results in decreased TPR and an increase in CO, whereas a threat state leads little or no change or a decrease in CO and little or no change or an increase in TPR (Blascovich \& Mendes, 2010).

There is evidence that individuals' explicit responses in experimental tasks are not always congruent with underlying physiological markers (e.g., Blascovich, Mendes, Hunter, Lickel, \& Kowai-Bell, 2001). By utilizing the physiological markers specified by the BPS model of challenge and threat, one can identify motivational responses to a stimulus that are not typically accessible to a participant during a motivated performance situation.

## Theoretical Motivation and Research Questions

Previous work on cognition and emotion perception proposes that context matters when people decode others' emotions (Barrett, Mesquita, \& Gendron, 2011; Lanzetta \& Englis, 1989; Singer et al., 2006). In their review of the literature, Barrett and colleagues point out how the visual scene in a stimulus can give rise to different interpretations of an emotional state conveyed by a facial expression. A scowling face can convey anger or disgust depending on the body posture with which it is paired. Individuals' patterns of behavior can also serve as context cues that affect emotional processing.

Two similar individuals can give rise to different emotional responses in their interaction partners based on their behaviors and actions. For example, Singer and colleagues (2006) led participants to believe that confederate players in a Prisoner's Dilemma game were fair or not based on the confederates' game investment strategy. Experimenters then randomly cued participants that either the (fair/unfair) confederate or the participant herself would receive a painful shock. Participants exhibited more empathic neural activity (fronto-insular and anterior cingular cortex) when they observed a fair player receive a shock compared to the unfair player (Singer et al., 2006). This is compelling because the only difference between the individuals was the contextual information of their game playing strategy.

On the basis of this research (i.e., Barrett et al., 2011; Singer et al., 2006), we can infer that different contexts can shape the perception of emotion as well as give rise to different neurophysiological responses to facial expressions. In particular, it is possible that an experimental confederate that employs a fair strategy will be perceived differently as a function of whether the individual smiles or scowls. Similarly, a fair individual that smiles might be perceived differently compared to an unfair smiling individual. In this study, we utilized virtual humans as research confederates in order to manipulate their facial expressions and negotiation strategies while keeping other aspects of the interaction under experimental control.

The research question driving this work was: Do differences in virtual humans' emotional facial expressions coupled with their behavioral strategies in a negotiation task affect neuropsychological processes related to motivation and affect?

The contextual emotion hypothesis suggests that if the confederate's negotiation strategy affects perceptual process


Figure 2: The multi issue bargaining negotiation task interface with areas of interest.
of facial expressions, then individuals will show different responses to the same facial expression depending on context. Specifically, individuals will have a different cardiovascular response to angry faces when paired with a tough strategy compared to angry faces paired with soft strategies.

Van Kleef and colleagues have argued that if partners in a social interaction lack information about the other's needs, desires, and goals, then emotional displays help people make sense of situations (Van Kleef, De Dreu, \& Manstead, 2010). It follows then that people will tend to look more at emotionally significant facial features when there is uncertainty in social interactions. Therefore, with respect to the eye tracking measure, the contextual emotion hypothesis predicts that individuals will fixate more on diagnostic facial features when the confederate's negotiation strategy is incongruent with his facial expression.

An alternative hypothesis suggests that emotion perception is driven purely by the structural features of a face alone. This hypothesis predicts that individuals will show heightened threat responses to angry faces compared to happy faces-regardless of the confederate's strategy with which they are coupled-and there should be more eye fixations on threatening faces (Mogg, Garner, \& Bradley, 2007; Tracy \& Robins, 2008).

## Method

## Participants, Design, Materials, Apparatus

Eighty participants were recruited from university undergraduate psychology courses. Participants played a multi-issue bargaining task (Van Kleef, De Dreu, \& Manstead, 2004). The task involves a scenario in which
participants act as mobile phone sellers and have to negotiate over three issues: a price, length of service contract, and warranty duration with the virtual human (see Figure 2, Payoffs). Each issue had a level that denoted its worth to the participants. Given that the participant was the seller, she would get the most points by selling the mobile phones for the highest price ( $\$ 150$, level 9 ) in order to gain 400 points; the shortest warranty period ( 1 month, level 9) corresponding to 120 points; and the shortest service contract (1 month, level 9) corresponding to 240 points. Participant's maximum score was therefore 760 points.

The confederate was an intelligent agent that displayed emotional facial expressions to convey anger or happiness. The study employed a 2 X 2 fully-crossed factorial betweensubjects experimental design. The two factors were the virtual human's emotional facial expression (happy or angry) crossed with his negotiation strategy (tough or soft). When the virtual human used a tough strategy (competitive), he made small concessions from his initial offer compared to the larger concessions he made using a soft (i.e., cooperative) strategy.

Both the soft and tough negotiating confederates made the initial offer to the participant, which was level 1 of price ( $\$ 110$, zero points to the participant), level 2 of warranty period ( 8 months, 15 points to the participant), and level 1 of length of service contract ( 9 months, zero points to the participant). From this 1-2-1 initial offer by the confederate, the soft and tough agent followed different counter offer policies. In both cases, the confederate's offer was not contingent on the participant's counter offers (see Table 1).

Table 1: Progression of Soft and Tough Negotiation offers through the six round task

| Round | Soft | Tough |
| :--- | :--- | :--- |
| 2 | $1-2-3$ | $1-2-2$ |
| 3 | $1-4-3$ | $1-3-2$ |
| 4 | $3-4-3$ | $2-3-2$ |
| 5 | $3-4-5$ | $2-3-3$ |
| 6 | $5-4-5$ | $3-3-3$ |

While participants interacted with the virtual character during the negotiation game, various measures related to their cardiovascular states were recorded. Electrocardiographic (EKG) and impedance cardiographic (ZKG) signals were recorded continuously with a Biopac MP150 system, using a standard lead II electrode configuration (for EKG) and a tetrapolar aluminum-mylar tape electrode system (for ZKG); blood pressure was continuously recorded using an automated blood pressure device. The automated blood pressure recorded readings via a cuff placed around the participants' wrists and fingers of their non-dominant left hand. The EKG and ZKG signals were scored using an interactive software program that produces ensemble-averaged values for heart rate (HR), pre-
ejection period (PEP). Additionally, cardiac output (CO) was calculated from stroke volume (SV) recordings via impedance and heart rate; and total peripheral resistance (TPR) was calculated using impedance and blood pressure readings as a measure of vascular activity.

An SMI RED eye tracker was used with 60 Hz sampling rate and a 17 " flat screen monitor. The eye tracking camera was positioned to the monitor and as such was unobtrusive to the participants during the task.

## Procedure

Participants completed a health screening questionnaire and informed consent was obtained prior to their participation. No one refused to participate. Female research assistants proceeded to apply the necessary sensors for physiological recording including impedance tape electrodes, EKG spot electrodes and blood pressure sensors. A five point calibration was used to ensure proper eye tracking measurement.

Next, the participant sat comfortably at rest for five minutes prior to receiving any task instructions. Finally, the participants were instructed to play 1 practice round to learn the user interface, during which the virtual human was not visible. Next, the criterion negotiation game commenced for 6 rounds. Afterwards, participants completed surveys that recorded their subjective and open-ended responses to the virtual human.

## Results

## Negotiation Task

Performance on the negotiation task was calculated based on how much the participants conceded to the virtual confederate over the six rounds. Each issue in the negotiation was summed at each round to compute a demand score. The best outcome for the participant would have been a demand score of 760 points. The final performance measure was the difference between the first and last round demand scores. If participants conceded more over the six rounds, this difference score would be higher. A univariate ANOVA with two factors (1. emotion: happy or angry; 2. strategy: tough or soft) showed no main effects or interactions (all $p$ 's $>0.5$, see Table 2 for means).

Table 2: Mean demand score difference from the first and last round (SD). A score of zero indicates no concession.

|  | Angry | Happy |  |
| :--- | :--- | :--- | :--- |
| Soft | $150.2(223)$ | $129.4(217)$ | $140.6(217)$ |
| Tough | $113.3(155)$ | $142.5(193)$ | $127.6(173)$ |
|  | $131.8(190)$ | $136.3(202)$ | 133 |

## Cardiovascular Physiological Indexes

We predicted that individuals interacting with the virtual confederate would exhibit task engagement, and that those interacting with an incongruent virtual confederate (e.g., soft strategy but angry face) would experience threat.

## Task Engagement

According to the Biopsychosocial Model, task engagement is indexed by increases from baseline in sympathetically driven cardiovascular responses. As is common in this research, we calculated changes from baseline in preejection period (PEP), a purely sympathetically driven cardiovascular measure (Tomaka, Blascovich, Kelsey, \& Leitten, 1993).

We established average baseline values of HR and PEP by averaging baseline minutes 4 and 5 for each of these measures. PEP decreased during the task ( $M=133.3 \mathrm{~ms}, S D$ $=15.76$ ) compared to the baseline $(M=135.8 \mathrm{~ms}, S D=$ 16.06), as predicted, two-tailed paired samples $t$-test, $t(78)=$ $3.31, p=.001$.

## Challenge and Threat

Total Peripheral Resistance (TPR) scores were computed by subtracting TPR during baseline from TPR during the negotiation task. A univariate ANOVA did not show a main effect of either strategy or emotion (both $F$ 's $<1$ ). There was also no interaction, $F(1,62)=1.47, p=.23$.

Cardiac output (CO) reactivity scores were computed by subtracting CO during baseline from CO during the negotiation task. A univariate ANOVA, controlling for baseline CO, with two factors showed no main effects of emotion or strategy ( $F$ 's $<1$ ). There was an interaction between emotion and strategy, $F(1,76)=8.34, p=0.005$, $\eta_{p}{ }^{2}=.098$.

Using a Bonferroni adjustment, simple effects analyses revealed that participants in the soft strategy condition significantly differed from each other, $F(1,77)=5.34, p=$ $0.024, \eta_{p}{ }^{2}=.065$. Participants who interacted with the virtual confederate that displayed an angry facial expression while using a soft (more conceding) strategy had further reduced cardiac output compared to those participants in the soft-happy condition (see Figure 3).

## Eye Tracking

BeGaze eye tracking analysis (SMI) software was used to construct areas of interest (AOI) on different components of the task interface as well as the virtual confederate's facial regions.


Figure 3: Cardiac output reactivity scores in the two virtual human strategy and facial expression conditions.

A multivariate ANOVA was conducted with the eight AOI (see Figure 2). The MANOVA showed a significant difference among the different AOI, $F(7,504)=77.06, p<$ $0.001, \eta_{p}{ }^{2}=.517$. As Figure 4 shows, participants fixated on the offer section of the game interface the longest percentage of time throughout the task. However, the results suggest people also spend considerable time looking at the face. In fact, the percentage of time fixated on the total face-aggregate of eyes, mouth, and remainder of the face-did not differ from the time fixated on the offer, $t(72)$ $=.54, p=.6$.

## Mouth AOI

The main differentiating facial feature for the angry and happy expressions was the mouth area. Past work indicates


Figure 4: Percentage of time during the whole negotiation fixated on different areas of interest.
that individuals from samples similar to ours tend to fixate more on the mouth region (Blais, Jack, Scheepers, Fiset, \& Caldara, 2008; Jack, Blais, Scheepers, Schyns, \& Caldara, 2009). Thus, we conducted an ANOVA with the two factors


Figure 5: Percentage of time on the mouth as a function of the confederate's emotion and strategy.
of strategy and emotion on the percentage of time fixated on the mouth AOI.

There was no main effect of strategy or emotion, $F$ 's $<1$. There was a marginal interaction of strategy and emotion, $F(1,68)=3.281, p=.074, \eta_{p}{ }^{2}=.046$. As Figure 5 shows, participants in the soft-angry condition tended to fixate the mouth for a longer time compared to participants in the softhappy condition.

## Discussion

Virtual human confederates in a negotiation game caused a threat motivational response (reduced cardiac output) when their facial expressions were not congruent with their strategies. Specifically, participants had lower cardiac output when the virtual human negotiated using a soft strategy but displayed an angry facial expression. Additionally, despite not reaching significance, similar effects occurred when participants engaged with a tough agent that showed happy facial expressions. This incongruence could cause more uncertainty, which is related to increases in task demands (Tomaka et al., 1993).

Eye tracking results provide converging evidence. Participants in the incongruent strategy and emotion condition (e.g., soft-angry) tended to fixate on the most diagnostic facial region longer compared to participants in the congruent condition (e.g., soft-happy). This result suggests that participants tended to fixate longer at the mouth in order to try to gain potential cues to reconcile their uncertainty from the conflicting strategy and emotion coupling.

These results are compatible with the suggestion that people look at others' facial expressions in an attempt to reduce inherent uncertainty that occurs in social decision making situations with counterparts that might have different priorities and objectives (Van Kleef et al., 2010). Our results show specific psychophysiological evidence for this process, especially when there is an incongruence between the counterpart's strategy and the facial displays.

These results are also in line with other research which suggests that context matters when people decode others' emotions (Barrett et al., 2011; Lanzetta \& Englis, 1989; Singer et al., 2006; Szczurek, Monin, \& Gross, 2012). An
identical angry facial expression gave rise to different motivational states depending on the strategic context in which it was displayed during the negotiation task.

Finally, the results can have practical implications for the design of human-computer interaction systems. This work suggests that cardiovascular measures are sensitive at detecting incongruence and uncertainty in human users and suggests that affecting the context in which emotions are shown (for instance, in virtual humans) can lead to concomitant changes to the user's challenge/threat motivational state.

## Acknowledgments

Peter Khooshabeh performed this research while on appointment as a U.S. Army Research Lab Postdoctoral Fellow. We thank Kyle O’Donnell for help with analysis.

## References

Barrett, L. F., Mesquita, B., \& Gendron, M. (2011). Context in Emotion Perception. Current Directions in Psychological Science, 20, 286-290. doi:10.1177/0963721411422522
Beale, R., \& Creed, C. (2009). Affective interaction: How emotional agents affect users. International Journal of Human-Computer Studies, 67(9), 755776. doi:10.1016/j.ijhcs.2009.05.001

Blais, C., Jack, R. E., Scheepers, C., Fiset, D., \& Caldara, R. (2008). Culture shapes how we look at faces. PLoS ONE, 3(8).
Blascovich, J. (2008). Challenge and threat appraisal. In Handbook of Approach and Avoidance Motivation (pp. 432-444). CRC Press.
Blascovich, J., \& Mendes, W. B. (2010). Social psychophysiology and embodiment. In Handbook of Social Psychology (5th Edition., pp. 194-227). New York: Wiley.
Blascovich, J., Mendes, W. B., Hunter, S. B., Lickel, B., \& Kowai-Bell, N. (2001). Perceiver threat in social interactions with stigmatized others. Journal of Personality and Social Psychology, 80(2), 253267.

Blascovich, J., Vanman, E., Mendes, W., \& Dickerson, S. S. (2011). Social psychophysiology for social and personality psychology. Los Angeles: SAGE.
De Melo, C. M., Carnevale, P., \& Gratch, J. (2011). The effect of expression of anger and happiness in computer agents on negotiations with humans. In The 10th International Conference on Autonomous Agents and Multiagent Systems - Volume 3 (pp. 937-944). Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems. Retrieved from http://dl.acm.org/citation.cfm?id=2034396.203440 2
Dienstbier, R. A. (1989). Arousal and physiological toughness: Implications for physical and mental health. Psychological Review, 96(1), 84-100.

Jack, R. E., Blais, C., Scheepers, C., Schyns, P. G., \& Caldara, R. (2009). Cultural confusions show that facial expressions are not universal. Current Biology, 19, 1543-1548.
Lanzetta, J. T., \& Englis, B. G. (1989). Expectations of cooperation and competition and their effect on observers' vicarious emotional responses. Journal of Personality and Social Psychology, 56(4), 543554.

Meehan, M., Insko, B., Whitton, M., \& Brooks, F. P. (2002). Physiological measures of presence in stressful virtual environments (p. 645). ACM Press. doi:10.1145/566570.566630
Mogg, K., Garner, M., \& Bradley, B. P. (2007). Anxiety and orienting of gaze to angry and fearful faces. Biological Psychology, 76(3), 163-169. doi:10.1016/j.biopsycho.2007.07.005
Rienerman-Jones, L., Cosenzo, K., \& Nicholson, D. (2010). Subjective and objective measures of operator state in automated systems. Presented at the Applied Human Factors and Ergonomics, Miami, FL.
Singer, T., Seymour, B., O'Doherty, J. P., Stephan, K. E., Dolan, R. J., \& Frith, C. D. (2006). Empathic neural responses are modulated by the perceived fairness of others. Nature, 439(7075), 466-469. doi:10.1038/nature04271
Szczurek, L., Monin, B., \& Gross, J. J. (2012). The Stranger Effect: The Rejection of Affective Deviants. Psychological Science, 23(10), 1105-1111. doi:10.1177/0956797612445314
Tamietto, M., \& De Gelder, B. (2010). Neural bases of the non-conscious perception of emotional signals. Nature Reviews Neuroscience, 11(10), 697-709. doi:10.1038/nrn2889
Tomaka, J., Blascovich, J., Kelsey, R. M., \& Leitten, C. L. (1993). Subjective, physiological, and behavioral effects of threat and challenge appraisal. Journal of Personality and Social Psychology, 65(2), 248260. doi:10.1037/0022-3514.65.2.248

Tracy, J. L., \& Robins, R. W. (2008). The automaticity of emotion recognition. Emotion, 8(1), 81-95. doi:10.1037/1528-3542.8.1.81
Van Kleef, G. A., De Dreu, C. K. W., \& Manstead, A. S. R. (2004). The interpersonal effects of anger and happiness in negotiations. Journal of Personality and Social Psychology, 86(1), 57-76.
Van Kleef, G. A., De Dreu, C. K. W., \& Manstead, A. S. R. (2010). An Interpersonal Approach to Emotion in Social Decision Making. In Advances in Experimental Social Psychology (Vol. 42, pp. 4596). Elsevier. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S006526 011042002X

# A model of population dynamics applied to phonetic change 

James Kirby (j.kirby@ed.ac.uk)<br>School of Philosophy, Psychology, and Language Sciences, University of Edinburgh, Edinburgh EH8 9AD UK<br>Morgan Sonderegger (morgan.sonderegger@mcgill.ca)<br>Department of Linguistics, McGill University, Montreal, Quebec H3A 1A7 Canada


#### Abstract

We consider the problem of language evolution in a population setting, focusing on the case of continuous parameter learning. While theories of phonetic change tend to emphasize the types of transmission errors that could give rise to a shift in pronunciation norms, it is challenging to develop a model that allows for both stability as well as change. We model the acquisition of vowel-to-vowel coarticulation in both single- and multiple-teacher settings, considering progressively more restrictive prior learning biases. We demonstrate that both stability and change are possible at the population level, but only under fairly strong assumptions about the nature of learning and production biases.


Keywords: Language evolution; sound change; computational modeling; phonetics; coarticulation

## Introduction

The problem of language evolution and change has received increased attention from a computational perspective in recent years (e.g. Niyogi \& Berwick, 1995; Wedel, 2006; Kirby, Dowman, \& Griffiths, 2007). Most of this work has focused on modeling either lexical or syntactic change, where the task is usually cast as deciding between competing discrete representation, e.g. different grammars (Baker, 2008). A similar approach is often taken in models of the evolution of sound patterns, where the learning problem is cast as one of deciding between discrete pronunciation variants (e.g. Niyogi, 2006).

However, learning a sound pattern of a language also consists of learning continuous phonetic cue distributions that describe how the sounds of that language are realized. Understanding the dynamics of these distributions is important for understanding sound change, because the seeds of categorylevel change are often claimed to be based in continuous phonetic variation (Ohala, 1993). In this paper, we address the evolution of sound patterns by considering the acquisition of continuous parameter distributions in a population setting. While we consider the particular example of a phonetic parameter, the basic results are applicable to the learning of continuous parameters more generally.

## Stability and change in phonetic realization

In all languages, when a sound is produced in a connected stream of speech, its phonetic realization is influenced by the preceding and/or following context. This contextual variability, termed coarticulation, has often been argued to underlie a wide variety of sound changes in the world's languages. One example is a historical process known as primary umlaut, attested in Old High German beginning c. 750 AD, in which short low /a/ was fronted and raised to /e/ when a high front vowel or glide occurred in the following syllable,
e.g. *[gasti:] >/gesti/ 'guests' (modern German Gäste). It has been proposed that the roots of umlaut may be traced to vowel-to-vowel coarticulation (Iverson \& Salmons, 2003); however, vowel-to-vowel coarticulation did not invariably result in umlaut. For example, even while primary umlaut was spreading throughout West Germanic, it is clear that it did not affect Gothic (Campbell, 1998:75). The umlaut example illustrates a more general point: the mere presence of a potential trigger does not imply that phonetic change is inevitable. Thus, any empirically adequate model of how the sound pattern of a language evolves must account for instances of stability as well as change (Weinreich, Labov, \& Herzog, 1968).

## Learning bias in phonetic change

An important body of research on phonetic change has focused on establishing the preconditions for change to occur in a single speaker-hearer (Ohala, 1993). Similarly, computational models of phonetic change have mostly considered individuals, focusing on how biases in learning or in speech production/perception impact whether or not change occurs (Pierrehumbert, 2001). However, even if a change were to obtain at the level of a single speaker, its spread in the speech community is far from inevitable: social and cultural factors may conspire to inhibit or enhance a change in the population at large. In addition, the dynamics of linguistic populations are complex: how assumptions about individual speakers play out in population dynamics can be surprisingly nontrivial and dependent on assumptions about population structure (Niyogi, 2006). For both reasons, the general plausibility of accounts of contextually-driven phonetic change, and what role channel and learning biases play, cannot be properly assessed until their dynamics at the population level are better understood.

This paper explores the effects of different assumptions about bias and population structure on the evolution of phonetic categories in a population, as applied to a simplified version of primary umlaut in Germanic. In particular, we consider six models of learning how /a/ is pronounced before a high vowel. Our aim is a model which satifies three goals:

1. Stability of limited coarticulation in the population, as in pre-Old High German
2. Stability of full coarticulation in the population (e.g. umlaut), as in Old High German
3. Change from stable limited coarticulation to stable full coarticulation.


Figure 1: (a) Parallel diffusion chains (classic iterated learning). (b) Single-teacher scenario. (c) Multiple-teacher scenario.

## Properties of populations

Fig. 1 illustrates three types of population structure. Fig. 1a shows a classic iterated learning (IL) scenario, also known as a diffusion chain (Smith, Kirby, \& Brighton, 2003). In IL, each learner of generation $t+1$ receives input from a randomly chosen member of generation $t$. Thus, every member of a generation functions both as learner and as teacher. Fig. 1b illustrates single-teacher learning with replacement. This scenario differs crucially from classic IL in that while the input for a learner comes from exactly one teacher, some teachers may provide input to more than one learner, while others may not provide any. Finally, Fig. 1c illustrates multipleteacher learning with replacement. Here, input may come from more than one teacher, although some teachers may not provide data to any leaners in the following generation.

The IL scenario has frequently been assumed in work on language evolution and change (e.g. Smith et al., 2003; Kirby et al., 2007). ${ }^{1}$ While these models have well-understood dynamics and may be appropriate in some situations, in general different dynamics emerge in population learning scenarios (1b-c) (Dediu, 2009; Smith, 2009). In this work we focus on scenarios (1b-c), aiming to determine (1) what type of biases are necessary for stability and change to obtain in a population of learners, and (2) if and how such biases interact with differences in the number of teachers. We begin with a naive learning model (no prior) and then turn to consider progressively more restrictive priors.

## Framework

We assume that (1) speech sounds have been organized into discrete segments; (2) the phonetic realisation of segments is subject to coarticulation; and (3) the learner has access to the complete segmental inventory. We consider here a simple language with the lexicon $\Sigma=\left\{\mathrm{V}_{1}, \mathrm{~V}_{2}, \mathrm{~V}_{12}\right\}$, where $\mathrm{V}_{12}$ represents $\mathrm{V}_{1}$ in the potentially coarticulation-inducing context of $V_{2}$. For the primary umlaut example, $V_{1}$ and $V_{2}$ can be thought of as the vowels $/ \mathrm{a} /$ and $/ \mathrm{i} /$ in isolation, and $\mathrm{V}_{12}$ as the vowel /a/ in a context where it is coarticulated towards /i/.

For simplicity, vowel tokens are represented by their first formant (F1) value, an acoustic measure of vowel height. We assume that the F 1 distributions of $\mathrm{V}_{1}$ and $\mathrm{V}_{2}$ are known to all learners, are the same for all learners, and do not change over time. The distribution of $\mathrm{V}_{12}$ differs from that of $\mathrm{V}_{1}$ only by an offset to the mean $p$, indicating how much $\mathrm{V}_{1}$ is affected by coarticulation (i.e. raised) in the $\mathrm{V}_{2}$ context. In all

[^117]derivations below, we assume in particular that for a learner with parameter $p$, the three categories $\mathrm{V}_{1}, \mathrm{~V}_{2}$, and $\mathrm{V}_{12}$ follow normal distributions in a single dimension (F1):
\[

$$
\begin{equation*}
\mathrm{V}_{1} \sim N\left(\mu_{a}, \sigma_{a}^{2}\right), \mathrm{V}_{2} \sim N\left(\mu_{i}, \sigma_{i}^{2}\right), \mathrm{V}_{12} \sim N\left(\mu_{a}-p, \sigma_{a}^{2}\right) \tag{1}
\end{equation*}
$$

\]

We assume that learners are divided into discrete generations $\mathcal{G}_{t}$ of size $M$. Each learner in generation $\mathcal{G}_{t}$ receives $n_{\mathrm{V}_{12}}$ examples of $\mathrm{V}_{12}$ from the members of generation $\mathcal{G}_{t-1}$. The learner's task is then simply to infer $p$. The state of the population $\mathcal{G}_{t}$ can thus be characterized by the distribution $p \sim \pi_{t}(p)$. For simplicity, we assume that $M$ is infinite, so the evolution of the population is not a stochastic process.

In $\mathcal{G}_{t+1}$, each learner is presented with $n$ examples of $\mathrm{V}_{12}$ drawn from a sequence of teachers in $\mathcal{G}_{t}$ chosen by some sampling procedure $\mathcal{S} .{ }^{2}$ Given these examples, the learner applies some learning algorithm $\mathcal{A}$. Assuming $\mathcal{S}$ and $\mathcal{A}$ are the same for all agents in $\mathcal{G}_{t+1}$, this implies the following evolution equation for $\pi_{t}$ :

$$
\begin{equation*}
\left(\pi_{t+1}\right)=f_{S, \mathcal{A}}\left(\pi_{t} \mid \text { constants }\right) \tag{2}
\end{equation*}
$$

For a given $\mathcal{A}$ and $\mathcal{S}$, our goal is to determine $f$, and characterize its behavior, in particular which (if any) of our modeling goals it satisfies.

## Models

This section describes the evolutionary dynamics of a population of learners who estimate the degree of coarticulation from training data based on the assumption that these examples are independently and identically (i.i.d.) generated by a single source with a fixed $p$. We consider learners with three types of prior bias in estimating $p$, corresponding to three choices of $\mathcal{A}$ : no prior $\left(\mathcal{A}_{\text {naive }}\right)$, a simple prior $\left(\mathcal{A}_{\text {simple }}\right)$, and a more complex prior ( $\left.\mathcal{A}_{\text {complex }}\right)$. These learners are embedded in two of the types of populations shown in Fig. 1, corresponding to two choices of $\mathcal{S}$ : (1b), in which a learner's input is provided by a single teacher ( $\mathcal{S}_{\text {single }}$ ), and (1c), in which her input may be drawn from multiple teachers ( $\mathcal{S}_{\text {multiple }}$ ).

## $\mathcal{A}_{\text {naive }}$ : Naive learning models

We first consider maximum-likelihood (ML) learners who are "naive" in the sense of having no prior over estimates of $p .{ }^{3}$

[^118]Model 1.1: Naive learning, single teacher First we consider a situation in which a learner in generation $\mathcal{G}_{t+1}$ receives examples from a single member of generation $\mathcal{G}_{t}$. Each learner is associated with a value $p_{\text {parent }}$ (one draw from the $\pi_{t}$ distribution, representing the single teacher's degree of coarticulation), which is used to generate $n$ training examples $\vec{y}=\left(y_{1}, \ldots, y_{n}\right)$. Let $\bar{y}$ be the mean of this sample. Each example is normally distributed, following (1): $P\left(y_{i}\right)=N\left(\mu_{a}-p_{\text {parent }}, \sigma_{a}^{2}\right)$. The sample's mean is also normally distributed, with the same mean and reduced variance:

$$
\begin{equation*}
P\left(\left(y_{1}+\cdots+y_{n}\right) / n \mid p_{\text {parent }}\right)=N\left(\mu_{a}-p_{\text {parent }}, \sigma_{a}^{2} / n\right) \tag{3}
\end{equation*}
$$

Given $\bar{y}$, the learner's maximum-likelihood estimate of $p$ is $\hat{p}=\mu_{a}-\bar{y}$. Thus, using Eq. 3, the distribution over values of $\hat{p}$ the learner could acquire given $p_{\text {parent }}$ is:

$$
\begin{equation*}
P\left(\hat{p} \mid p_{\text {parent }}\right)=N\left(p_{\text {parent }}, \sigma_{a}^{2} / n\right) \tag{4}
\end{equation*}
$$

We are interested in the evolution of the distribution $\pi_{t}$ : that is, the marginal distribution of $\hat{p}$ as a function of the distribution of $p_{\text {parent }}$. Abbreviating $p_{\text {parent }}$ as $p$, this is:

$$
\begin{align*}
\pi_{t+1}(\hat{p}) & =\int \underbrace{P((\hat{p} \mid p)}_{\text {Eq } \cdot 4} \pi_{t}(p) d p \\
& =\int \pi_{t}(p) \cdot N_{\hat{p}}\left(p, \sigma_{a}^{2} / n\right) d p \tag{5}
\end{align*}
$$

To get a sense of the evolution of $\pi_{t}$, we can compute how its mean and variance change over time. Let $p$ be the random variable distributed according to $\pi_{t}$, and $\hat{p}$ the same for $\pi_{t+1}$. The expected value of $\hat{p}$ is then:

$$
\begin{align*}
E[\hat{p}]=\int \pi_{t+1}(\hat{p}) \hat{p} d \hat{p} & =\int\left[\int \pi_{t}(p) \cdot N_{\hat{p}}\left(p, \sigma_{a}^{2} / n\right) d p\right] \hat{p} d \hat{p} \\
& =\int \pi_{t}(p) \underbrace{\left[\int N_{\hat{p}}\left(p, \sigma_{a}^{2} / n\right) \hat{p} d \hat{p}\right]}_{=E[\hat{p} \mid p]=p} d p \\
& =\int \pi_{t}(p) p d p=E[p] \tag{6}
\end{align*}
$$

By a similar derivation for $E\left[\hat{p}^{2}\right]$, the variance of $\hat{p}$ can be shown to be:

$$
\begin{align*}
\operatorname{Var}(\hat{p}) & =E\left[(\hat{p}-E[\hat{p}])^{2}\right]=E\left[\hat{p}^{2}\right]-E[\hat{p}]^{2} \\
& =\sigma_{a}^{2} / n+\operatorname{Var}(p) \tag{7}
\end{align*}
$$

Thus, the distribution of $p$ in the $n+1^{\text {th }}$ generation has the same mean as in the $n^{\text {th }}$ generation, but larger variance; i.e., the distribution becomes more diffuse with each generation. ${ }^{4}$

[^119]Model 1.2: Naive learning, multiple teachers We now consider the case where a learner in generation $\mathcal{G}_{t+1}$ receives each training example from a randomly-chosen teacher in generation $\mathcal{G}_{t+1}$. This is equivalent to drawing $n$ values of $p$ from $\pi_{t}, \vec{p}=\left(p_{1}, \ldots, p_{n}\right)$, and for each $p_{i}$ generating one training example $y_{i}$ :

$$
\begin{equation*}
P\left(y_{i} \mid p_{i}\right)=N\left(\mu_{a}-p_{i}, \sigma_{a}\right) \quad i=1, \ldots, n \tag{8}
\end{equation*}
$$

As in the single-teacher case, we assume that the learner chooses the ML estimate for $p, \hat{p}=\mu_{a}-\bar{y}$. Using (8) and the fact that the $y_{i}$ are independent and normally distributed:

$$
\begin{align*}
P(\bar{y} \mid \vec{p}) & =N\left(\mu_{a}-\left(p_{1}+\cdots+p_{n}\right) / n, \sigma_{a}^{2} / n\right) \\
\Longrightarrow P(\hat{p} \mid \vec{p}) & =N\left(\bar{p}, \sigma_{a}^{2} / n\right) \tag{9}
\end{align*}
$$

where $\bar{p}=\left(p_{1}+\cdots+p_{n}\right) / n$. Thus, the learner's estimate $\hat{p}$ is the mean of the $p$ values which generated the training data, plus some noise.

To obtain $\pi_{t+1}(\hat{p})$, the marginal distribution of $\hat{p}$, we integrate out $p_{1}, \ldots, p_{n}$ from (9):

$$
\begin{equation*}
\pi_{t+1}(\hat{p})=\int N_{\hat{p}}\left(\bar{p}, \sigma_{a}^{2} / n\right) \prod_{i=1}^{n} \pi_{t}\left(p_{i}\right) d p_{i} \tag{10}
\end{equation*}
$$

As in the single-teacher case, we can get a sense of how the distribution of $p$ evolves by computing the mean and variance of $\pi_{t+1}$. Let $p_{t}$ be the random variable with distribution $\pi_{t}$. The expected value and variance of $\hat{p}$ can be shown to be:

$$
\begin{equation*}
E(\hat{p})=E\left(p_{t}\right), \quad \operatorname{Var}(\hat{p})=\sigma_{a}^{2} / n+\operatorname{Var}\left(p_{t}\right) / n \tag{11}
\end{equation*}
$$

Some algebra shows that

$$
\left(\operatorname{Var}(\hat{p})-\alpha_{*}\right)=\left(\operatorname{Var}(p)-\alpha_{*}\right) / n
$$

where $\alpha_{*}=\sigma_{a}^{2} /(n-1)$. The variance of the distribution of $p$ moves over time towards $\alpha_{*}$; if already at $\alpha_{*}$, it stays there forever. Thus, the mean of the distribution of $p$ stays the same over time, but its variance moves towards a single value.

Summary Whether the single-teacher (1.1) or multipleteacher (1.2) scenario is assumed, the naive learning models predict that the average degree of coarticulation in the population will not change over time. It follows that, under these assumptions, change from little coarticulation to full coarticulation (Goal 3) is not possible.

The single-teacher model has an additional problem. The variability of the degree of coarticulation in the population is predicted to increase with each generation., i.e. speakers come to coarticulate increasingly differently. Intuitively, because each production is noisy, the learner's estimate of the degree of coarticulation is inherently noisy (Eq. 4): it is impossible to exactly acquire the target value of the parent from a finite sample. Increasing population-level variation in the degree of coarticulation over time is clearly empirically inadequate, because the effects of umlaut are generally either present or absent in a given population. Thus, Model 1.1 does not allow for stability of a population with little coarticulation (Goal 1) or full coarticulation (Goal 2).

## $\mathcal{A}_{\text {simple }}$ : Simple prior models

Intuitively, the reason that the single-teacher naive learning model fails to allow for stability around a particular value of $p$ is that there is no force counteracting the noise in each learner's estimate (4), which causes the distribution of $p$ values to spread out over time. In this section, we consider the effect of a simple prior learning bias on the evolution of $p$.

As above, we assume the learner estimates $p$ based on the assumption that data is generated i.i.d. from a source with a fixed $p$. The distribution of the data under this assumption is

$$
\begin{align*}
P(\vec{y} \mid p) & =P\left(y_{1} \mid p\right) \cdots P\left(y_{n} \mid p\right)  \tag{12}\\
& =\frac{\exp \left[-\sum_{i=1}^{n}\left(y_{i}-\left(\mu_{a}+p\right)\right)^{2} /\left(2 \sigma_{a}^{2}\right)\right]}{\left(2 \pi \sigma_{a}^{2}\right)^{n / 2}} \tag{13}
\end{align*}
$$

However, we now assume that learners' knowledge about $p$ is probabilistic: they begin with a prior distribution $(P(p)=$ $\alpha(p)$ ) on how likely different values of $p$ are a priori, which is updated to a posterior distribution based on the data $(P(p \mid \vec{y}))$.

Recall that the population of naive learners from a single teacher did not show the simplest possible empirically adequate dynamics: stability of the distribution of $p$ over time near $p=0$; i.e., most people have a minimal (but fixed) degree of coarticulation. As a first pass to see if this behavior is possible with learners who reason probabilistically about $p$, we assume learners have a prior biasing them towards values of $p$ near 0 , with values away from 0 becoming increasingly less likely. Intuitively, this prior "should" sharpen the distribution of $p$ towards $p=0$ over time, counteracting the effect of transmission noise which tends to make the distribution of $p$ spread out more in each generation (as in Model 1.1).

For simplicity, we assume a Gaussian prior $\alpha \sim N\left(0, \tau^{2}\right)$. The posterior is then simply:

$$
\begin{equation*}
P(p \mid \vec{y})=P(\vec{y} \mid p) P(p) / P(\vec{y}) \tag{14}
\end{equation*}
$$

The learner must pick a point estimate of $p$, denoted $\hat{p}$, using $P(p \mid \vec{y})$. The two familiar strategies, choosing the maximum or expected value of the posterior (abbreviated MAP, EV), turn out to be equivalent:

$$
\begin{equation*}
\hat{p}_{\mathrm{MAP}}=\hat{p}_{\mathrm{EV}}=\frac{\left(\mu_{a}-\bar{y}\right)}{1+\sigma_{a}^{2} / n \tau^{2}} \tag{15}
\end{equation*}
$$

Abbreviating the denominator of (15) as $K=1+\sigma_{a}^{2} /\left(n \tau^{2}\right)$, these estimates of $p$ may then be equivalently written as $\hat{p}=$ $\left(\mu_{a}-\bar{y}\right) / K$.

As above, we can now consider the consequences of this learning strategy under different population scenarios. ${ }^{5}$

Model 2.1: Simple prior, single teacher We again first assume a scenario in which each learner in generation $\mathcal{G}_{t+1}$ receives $n$ training examples from a single member of generation $\mathcal{G}_{t}$, who has coarticulatory parameter $p_{\text {parent }}$, abbreviated as $p$. The distribution of $p$ is $P(p)=\pi_{t}(p)$.

[^120]We first determine the distribution of a learner's estimate $\hat{p}$, given fixed $p$ and data $\vec{y} . \bar{y}$ is normally distributed, as described by (3), as in the no-prior case. Because $\bar{y}$ is normally distributed and $\hat{p}=\left(\mu_{a}-\bar{y}\right) / K$, the distribution of $\hat{p}$ is

$$
\begin{equation*}
P\left(\hat{p} \mid p_{\text {parent }}\right)=N\left(p_{\text {parent }} / K, \sigma_{a}^{2} / n K^{2}\right) \tag{16}
\end{equation*}
$$

Thus, on average, the learner's estimate of $p$ is closer to 0 than the parent's value.

We can now compute the marginal distribution of $\hat{p}$ :

$$
\begin{align*}
\pi_{t+1}(\hat{p})=P(\hat{p}) & =\int \underbrace{P(\hat{p} \mid p)}_{(16)} P(p) d p  \tag{17}\\
& =\int N_{\hat{p}}\left(p / K, \sigma_{a}^{2} / n K^{2}\right) \pi_{t}(p) d p \tag{18}
\end{align*}
$$

As in the no-prior case, we can gain some understanding of the evolution equation (18) by examining how the expectation and variance of $p$ evolve. By a similar derivation to (6), it can be shown that the expectation of $\hat{p}$ is:

$$
\begin{equation*}
E(\hat{p})=E\left(p_{\text {parent }}\right) / K \tag{19}
\end{equation*}
$$

Because $K>1$ (for any values of $\sigma_{a}, n$, and $\tau$ ), the expected value of the coarticulation parameter decreases with each generation. By a similar derivation to the no-prior case, the variance of $\hat{p}$ can be shown to be

$$
\begin{equation*}
\operatorname{Var}(\hat{p})=\left[\sigma_{a}^{2} / n+\operatorname{Var}(p)\right] / K^{2} \tag{20}
\end{equation*}
$$

and some algebra shows that

$$
\left(\operatorname{Var}(\hat{p})-\alpha_{*}\right)=\left(\operatorname{Var}(p)-\alpha_{*}\right) / K^{2}
$$

where $\alpha_{*}=\frac{\sigma_{a}^{2}}{n\left(K^{2}-1\right)}$. Because $K>1$, the variance of $p$ moves over time towards the fixed point $\alpha_{*}$, as in Model 1.2. Thus (as noted by Smith, 2009 in other settings), the distribution of coarticulation in the population does not converge to the prior, unlike the well-known result of Griffiths and Kalish (2007).
Model 2.2: Simple prior, multiple teachers The situation is similar under $\mathcal{S}_{\text {multiple }}$. The mean of $p$ can be shown to evolve exactly as in the single-teacher case (19), towards 0 . Similarly, the variance looks very similar to the evolution in the single-teacher case (20), except for an extra factor of $n$ in the denominator. The variance again evolves towards a fixed point, now $\alpha_{*}=\sigma_{a}^{2} /\left(n K^{2}-1\right)$, but in this case more quickly than in Model 2.1. Intuitively, this means that because learners have a strong prior against coarticulation, evidence for coarticulation at the level of the individual is mitigated and is unlikely to spread throughout the population.
Summary For both single- and multiple-teacher scenarios, a simple Gaussian prior drives the value of $p$ to 0 , predicting phonologization of coarticulation to be impossible. Thus, both Model 2.1 and 2.2 meet modeling Goal 1 (stability of little coarticulation), but neither of Goals 2 or 3 .

## $\mathcal{A}_{\text {complex }}$ : Complex prior models

The preceding section has shown that the distribution of $p$ in populations of learners with a Gaussian prior always converges to 0 . This simple prior model is empirically inadequate, because it fails to predict the possibility of stable coarticulation in a population. We therefore considered several more complex priors. Here we discuss one such prior, a quadratic polynomial with a minimum at $\left(\mu_{a}-\mu_{i}\right) / 2$ which is concave up between 0 and $\mu_{a}-\mu_{i}$ :

$$
\begin{equation*}
P(p) \propto\left[a\left(\mu_{a}-\mu_{i}\right)^{2}+\left(p-\left(\mu_{a}-\mu_{i}\right) / 2\right)^{2}\right] \tag{21}
\end{equation*}
$$

Here, $a$ is a scale parameter controlling the "strength" of the prior: as $a \rightarrow 0$, values of $p$ near the endpoints are maximally weighted relative to values near $\left(\mu_{a}-\mu_{i}\right) / 2$; as $a$ is increased, the prior is progressively flatter.
We assume the learner takes the MAP estimate of $p$ for values of $p$ in $\left[0, \mu_{a}-\mu_{i}\right]$, which does not have a closed form solution, but can be found numerically. We thus proceeded by simulation to determine the evolution of the distribution of $p$ over time in this case. The results reported here assume $\mu_{a}-\mu_{i}=200$ and a strong prior $(a=0.01)$ in a multipleteacher setting. We used large generation sizes ( $M=10000$ ) to approximate the deterministic behavior of infinite populations. Due to space constraints we discuss only the results for multiple-teacher models; the results for analogous singleteacher models are similar, in terms of our modeling goals.
Model 3.1: Complex prior First, we considered cases where there is little coarticulation in the population ( $p_{0} \sim$ $N(10,10)$ ) and where primary umlaut is effectively complete ( $p_{0} \sim N(190,10)$ ). The evolution of density estimates for $p$ over 1000 generations can be seen in the first panel of Fig. 2. While there is a slight shift in the mean and variance, they reach relatively stable values by around 1000 generations. However, the strength of the prior, in terms of the value of $a$, is important: as seen in the second panel of Fig. 2, for a weak prior ( $a=0.99$ ), the variance of $p$ in the population increases quickly over time, with results similar to the no-prior case discussed above.


Figure 2: Evolution of density of $p$ over time (indicated by color) with (left) a strong polynomial prior ( $a=0.01$ ) or (right) a weak polynomial prior ( $a=0.99$ ).

The simulation results suggest that a strong polynomial prior can result in a stable distribution for $p$ in the population over time, with most learners having values near 0 (little coarticulation) or 200 (full coarticulation). Model 3.1 thus satisfies our Goals 1 and 2.
Model 3.2: Complex prior, production bias More extensive simulation with Model 3.1, however, suggests that it does not satisfy Goal 3: it is never possible for a population to transition from a stable state of little articulation to a stable state of full coarticulation. The reason is intuitively clear: the prior is strong enough to bias learners towards either $p=0$ or $p=\mu_{a}-\mu_{i}$, but there is no countervailing force which could bias learners towards full coarticulation.

One plausible type of bias is an external force that increases the likelihood of coarticulated variants. Here, we implement a systematic production bias by assuming that some percentage of the learner's data have been moved towards $\mu_{i}$ by a quantity $\ell \sim N(\lambda, \lambda / 2)$; that is, they are coarticulated more than expected from the teacher's value for the coarticulatory parameter. This kind of bias, corresponding to a general tendency in speech production to over- or undershoot articulatory targets, is commonly considered in computational models of phonetic change (e.g. Pierrehumbert, 2001).

Assuming a strong polynomial prior $a=0.01$, change from no to full coarticulation turns out to indeed be possible, but only for a sufficiently large bias. Fig. 3 illustrates this with bias factors $\lambda=2$ and $\lambda=10$, starting in a state with little coarticulation ( $p_{0}=10$ ), in which $10 \%$ of tokens in each generation were subject to a lenition bias. As in Model 3.1, for a strong enough prior (low $a$ ) with no bias, the littlecoarticulation state is stable. As the amount of bias $(\lambda)$ is increased past a critical value, there is a rapid shift of the population to a stable state where most learners have full coarticulation. That is, there is a bifurcation where the amount of bias has overcome the stabilizing affect of the prior, and the little-coarticulation state becomes unstable. These results also illustrate a more general tradeoff between the strength of the prior and the amount of bias observed in further simulations (not shown here): for a stronger prior, the critical value of $\lambda$ increases: more bias is needed to overcome the prior.

Thus, Model 3.2 meets all three of our modeling goals: (1) a stable population with little coarticulation, (2) a stable population with full coarticulation, and (3) a rapid transition from little to full coarticulation are all possible, for particular initial conditions and values of the system parameters $(a, \lambda)$.

## Discussion

Our main goal in this paper was to evaluate how assumptions about bias and population structure for a population of learners of a continuous parameter translated into population-level models capable of modeling three empirically-observed scenarios of stability and change.

One interesting result was that population structure did not necessarily have much effect on the dynamics. For the naive learning scenario, the single-teacher and multiple-


Figure 3: Evolution of density of $p$ over time (indicated by color) with a strong polynomial prior ( $a=0.01$ ), with $10 \%$ of tokens subject to bias factor $\lambda=2$ (left) or $\lambda=10$ (right).
teacher models had qualitatively different dynamics: the variance in the degree of coarticulation stabilized over time in the multiple-teacher model, but increased over time in the singleteacher model. But for the simple prior and complex prior models, whether a single-teacher or multiple-teacher scenario was assumed largely impacted the rate at which a stable state was reached, rather than changing the qualitative outcome. Given that social structure plays a key role in the actuation and spread of language change (Labov, 2001), future work should further explore the role of different population structures with more complex teacher-learner relations. ${ }^{6}$

On the other hand, assumptions about bias mattered a great deal. When no or weak learning bias $\left(\mathcal{A}_{\text {naive }}\right.$, or $\mathcal{A}_{\text {simple }}$ with high $a$ ) was assumed, stability of the distribution of the coarticulatory parameter $p$ in the population was impossible. When a strong bias towards non-coarticulation was assumed ( $\mathcal{A}_{\text {naive }}$ with low $a$ ), stability of minimal coarticulation was possible (Goal 1), but stability of full coarticulation and change between the two (Goals 2,3 ) were not. It was only after assuming learners have a strong prior biasing them towards either little or full coarticulation, along with introducing an explicit unidirectional pressure to coarticulate, that it was possible to have primary umlaut: change (Goal 3) from stability of little coarticulation (Goal 1) to stability of full coarticulation (Goal 2).

Model 3.2, which met all three goals, shows a bifurcation: change from one stable state (little coarticulation in the population) to another (full coarticulation in the population) occurred suddenly as a system parameter (the amount of production bias) was varied past a critical value. Bifurcations in linguistic populations have been suggested as a potential mechanism underlying the actuation of linguistic change, but to our knowledge have previously only been shown to occur in models of change in discrete parameters (e.g. Niyogi, 2006). Our demonstration that bifurcations are possible in a population of learners of a continuous parameter supports the

[^121]hypothesis that bifurcations play a key role in the actuation of language change more generally. Future work should explore whether such bifurcations emerge in models that more accurately reflect the social structure of speech communities, and where the outcome of learning is a distribution over multiple phonetic cues, rather than a single cue.

## References

Baker, A. (2008). Computational approaches to the study of language change. Language and Linguistics Compass, 2(2), 289-307.
Boyd, R., \& Richerson, P. J. (1985). Culture and the evolutionary process. Chicago: University of Chicago Press.
Campbell, L. (1998). Historical linguistics: An introduction. Cambridge, MA: MIT Press.
Dediu, D. (2009). Genetic biasing through cultural transmission: Do simple Bayesian models of language evolution generalise? Journal of Theoretical Biology, 259, 552-561.
Griffiths, T., \& Kalish, M. (2007). Language evolution by iterated learning with Bayesian agents. Cognitive Science, 31, 441-480.
Iverson, G. K., \& Salmons, J. C. (2003). The ingenerate motivation of sound change. In R. Hickey (Ed.), Motives for language change. Cambridge: Cambridge University Press.
Kirby, S., Dowman, M., \& Griffiths, T. (2007). Innateness and culture in the evolution of language. Proceedings of the National Academy of Sciences, 104, 5241-5245.
Labov, W. (2001). Principles of linguistic change vol. 2: Social factors. Oxford: Oxford University Press.
Niyogi, P. (2006). The computational nature of language learning and evolution. Cambridge, MA: MIT Press.
Niyogi, P., \& Berwick, R. (1995). The logical problem of language change (AI Memo No. 1516). MIT.
Ohala, J. J. (1993). The phonetics of sound change. In C. Jones (Ed.), Historical Linguistics: Problems and Perspectives. London: Longman.
Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In J. Bybee \& P. Hopper (Eds.), Frequency and the emergence of linguistic structure. Amsterdam: John Benjamins.
Smith, K. (2009). Iterated learning in populations of Bayesian agents. In N. Taatgen \& H. van Rijn (Eds.), Proceedings of the 31th Annual Conference of the Cognitive Science Society (pp. 697-702). Austin: Cognitive Science Society.
Smith, K., Kirby, S., \& Brighton, H. (2003). Iterated learning: A framework for the emergence of language. Artificial Life, 9, 371-386.
Wedel, A. (2006). Exemplar models, evolution and language change. The Linguistic Review, 23, 247-274.
Weinreich, U., Labov, W., \& Herzog, M. (1968). Empirical foundations for a theory of language change. In W. Lehmann \& Y. Malkiel (Eds.), Directions for historical linguistics. Austin: University of Texas.

# Do older Gaelic-English bilinguals show an advantage in inhibitory control? 

Neil W. Kirk (n.kirk@abertay.ac.uk)<br>Department of Psychology, University of Abertay Dundee, Bell Street, Dundee DD1 1HG UK

Kenneth C. Scott-Brown (k.scott-brown@abertay.ac.uk)<br>Department of Psychology, University of Abertay Dundee, Bell Street, Dundee DD1 1HG UK

Vera Kempe (v.kempe@abertay.ac.uk)<br>Department of Psychology, University of Abertay Dundee, Bell Street, Dundee DD1 1HG UK


#### Abstract

We examined whether a bilingual advantage can be found in older bilinguals that share the same cultural background with monolinguals. Sixteen Gaelic-English bilinguals over the age of 60 years were compared with three monolingual control groups in performance on the Simon task, as well as in general intelligence and socio-economic status. Some of the monolinguals were bidialectal allowing us to also test whether switching between dialects can incur similar cognitive benefits as bilingualism. Results showed no group differences in overall reaction times as well as in the Simon effect suggesting that individuals that share a cultural background may not exhibit differences in inhibitory control even if they routinely use another dialect or another language. This opens up the possibility that other factors associated with bilingualism, like immigrant status, may be responsible for the bilingual advantage found in some but not in other studies. Keywords: Bilingualism; bildialectism; inhibitory control; Simon test.


## Introduction

A considerable number of studies have demonstrated a bilingual advantage in executive processing (for a review see Bialystok, Craik, Green \& Gollan, 2009), which seems to be most pronounced in young children and older adults (Bialystok, Martin \& Viswanathan, 2005). It has been suggested that knowing and using two or more languages on a regular basis requires individuals to inhibit one language while using the other, both at the level of selecting the appropriate linguistic setting as well as on the level of selecting individual words (Hilchey \& Klein, 2011). Thus, bilingualism has been causally linked to improved executive processing which transfers to non-linguistic domains. In older individuals, such improved executive processing may be beneficial for maintaining cognitive flexibility later in life (Bialystok et al., 2004) so much so as to even delay the onset of dementia (Bialystok, Craik \& Freedman, 2007).

However, because random assignment is not possible in quasi-experimental studies with bilingual participants, there is always the possibility that bilingualism is confounded with differences in a variety of hidden factors (Hilchey \& Klein, 2011), most notably socio-economic status (SES), but also educational and cultural background (Hakuta, Ferdman
\& Diaz, 1986), variables that can affect cognitive functioning (Mezzacappa, 2004). While recent studies reporting a bilingual advantage try to match bilinguals and monolinguals on SES, it is often difficult to match participants in cultural background, and immigrant status especially for older participants. For example, in Bialystok et al. (2004), the older monolinguals resided in North America while the majority of older bilinguals resided in India. Similarly, in Bialystok et al. (2008), 20 out of 24 older bilinguals were immigrants who had arrived in North America as children or adolescents suggesting that they belonged to an immigrant community likely to differ culturally from monolingual North American controls. Finally, while Schroeder and Marian (2012) do not explicitly report immigrant status or age of arrival in North America for their bilinguals, the range of languages spoken by their participants suggests that they were predominantly first or second generation immigrants from different cultural backgrounds than the monolinguals.

There is evidence that differences in cultural background are associated with differences in executive processing (Sabbagh, Xu, Carlson, Moses \& Lee, 2006). This can be attributed to culture-specific parenting attitudes or educational and leisure practices which influence exposure to activities that require and promote executive processing, such as musical training (Bialystok, 2011), playing of video games (Green \& Bavelier, 2003) and a host of other, as of yet, unknown factors.

There is also the possibility that genetic effects may be responsible for cultural differences in executive processing: For example, population-genetic studies have shown that the prevalence of the 7-repeat allele of the dopamine receptor gene (DRD4), is markedly lower in East and South East Asia compared to North America (Chang, Kidd, Kivak, Pakstis, \& Kidd, 1996). This allele has been associated with attention-deficit hyperactivity disorder (ADHD; Faraone, Doyle, Mick, \& Biederman, 2001), which, in turn, often manifests itself in poor executive processing (Schachar, Tannock, Marriott, \& Logan, 1995); although the relationship between DRD4 and ADHD itself seems to be subject to cross-cultural variation as culture may affect the
phenotypic realisation of this genotype (Nikolaidis \& Gray, 2010). Somewhat contradictory, Chen, Burton, Greenberger \& Dmitrieva (1999) have shown a link between the long alleles of DRD4 and population migration patterns, indicative of migration selecting for traits like noveltyseeking and openness. The personality trait of openness, in turn, has been associated with better performance in some aspects of executive functioning (Williams, Suchy, Rau, 2009). This may imply the possibility of a reverse causal relationship between bilingualism and executive processing: individuals with superior executive abilities might be more likely to be bilingual because of a potentially greater propensity to make life choices leading to migration or, when placed in a multi-lingual environment, greater success in maintaining use of multiple languages. Although reconciling these different findings is beyond the scope of this paper, they point to the intriguing possibility of a genetic origin of group differences in executive processing, which may co-vary with cultural background, immigrant status and bilingualism. Indeed, Morton and Harper (2007) failed to observe a superior inhibitory control when comparing non-immigrant bilingual with monolingual children matched for SES and cultural background. However, a similar study controlling for cultural background and immigrant status has not yet been conducted with older bilinguals. The present study therefore aims to test the bilingual advantage in executive processing in older bilinguals that share cultural background with the monolingual controls.

Studying Gaelic-English bilinguals allowed us to address this issue because Gaelic, a Celtic minority language, is spoken by a non-immigrant community of about 58,000 individuals residing mainly in the West of Scotland. Since Gaelic language schooling was abolished in 1872 and has been reintroduced only in 2006 there are no Gaelic monolinguals. Rather, older Gaelic-English bilinguals acquired Gaelic in early childhood before being introduced to English in school, and tend to use Gaelic in the home and in the local bilingual community. However, in terms of cultural attitudes and values, educational practices, leisure activities, media exposure and immigrant status, these bilinguals do not differ from English monolinguals.

In this study, we used the Simon test, closely modeled after Experiment 1 in Bialystok et al. (2004) to test whether Gaelic-English bilinguals exhibit benefits in inhibitory control compared to monolinguals recruited from the same cultural background. We restricted our exploration to the testing of inhibitory control, one component of executive processing, because this component had been examined in older bilinguals before (Bialystok et al., 2004; Bialystok et al., 2008; Schroeder \& Marian, 2012). In the Simon Task, participants have to inhibit a pre-potent spatially cued response when responding to the colour of a stimulus. This requires inhibitory control (Lu \& Proctor, 1995), which has been shown to be superior in bilinguals (Bialystok et al., 2004).

One issue that arises in a Scottish context is related to choosing appropriate monolingual controls: Britain is a country with extraordinary dialectal diversity and speakers of British English are often exposed to various local varieties of English. Specifically, in Scotland 85\% of the population report using one of the local varieties of the Scots dialect to varying degrees (Scottish Government Social Research, 2010), in addition to Standard Scottish English (SSE). Even though dialects of the same language are traditionally considered to be mutually intelligible there is considerable variability rendering the boundaries between languages and dialects fluid. Consequently, the linguistic classification of Scots, a Germanic language variety, is subject to much debate with some considering it a separate language, while others classifying it as a dialect of English or as a register used in specific social contexts (see Aitken, 1985). Indeed, local varieties of Scots differ considerably from SSE in their phonetic, lexical and even some syntactic features (Smith \& Durham, 2012). Thus, bidialectal speakers must monitor continuously who can or cannot be addressed in Scots, choose appropriate articulatory settings, and inhibit phonetic and lexical variants pertaining to the variety not currently used. It is therefore important to carefully control dialect use in the monolinguals. Moreover, the question as to whether use of multiple dialects can incur executive processing benefits similar to those observed in bilinguals is an interesting question in its own right, and will also be explored in this study. We tested three monolingual control groups: (1) bidialectal speakers who reported switching continuously between SSE and Dundonian, a local variety of Scots spoken in Eastern Scotland, (2) monodialectal speakers of SSE residing in the same locale as the bidialectals but who reported never or rarely using Dundonian, and (3) monolingual speakers of Anglo-English, a variety spoken in the South of England, for whom Scots was for the most part unintelligible. Note that the label monodialectal is used to refer to those monolingual participants who share a geographical and cultural background with the bidialectal participants. If inhibitory control advantages arise for different languages only then one would expect to find faster reaction times and a smaller Simon effect only in the Gaelic-English bilinguals. If regular switching to dialect also results in an inhibitory control advantage one would expect bidialectals to also exhibit shorter reaction times and a smaller Simon effect compared to monodialectals and monolinguals.

## Method

## Participants:

Sixty-four older adults ( $M=70.3$ years, SD $=7.6$ years, range $=60.2-88.7$ years) participated in the experiment. The 16 bilingual participants ( 6 men ) were speakers of Gaelic and SSE, the 16 bidialectal participants ( 7 men ) were speakers and regular users of both SSE and Dundonian Scots, the 16 monodialectal participants ( 5 men ) were monolinguals speakers of SSE who did not use Dundonian Scots, and the 16 monolingual participants ( 6 men ) were
speakers of Anglo-English. The monodialectal and bidialectal participants were recruited from the Dundee area, the Gaelic-SSE bilinguals were recruited from the Western Isles and the West coast of Scotland, and the English monolinguals were recruited from different parts of England and Scotland (all but one had not lived in Scotland for any considerable length of time and were either visitors or had recently retired to the area).

The Background Questionnaires (described below) revealed that the bilinguals' daily use of Gaelic and the bidialectals' use of Dundonian Scots ranged between 30\% and $70 \%$ of times. The monodialectals reported less than 25 $\%$ use of Dundonian Scots. Three other participants reported predominantly using Dundonian Scots. As it proved impossible to recruit further monodialectal speakers of this type, these monodialectals were excluded from the study. One bilingual participant was excluded due to $90 \%$ SSE and only $10 \%$ Gaelic usage, and one participant failed to perform the Simon Task correctly and was also excluded.

## Materials:

Background Questionnaire: A background questionnaire was used to gather relevant background information about the participants' educational background (including the age they left school, whether they continued to further or higher education and which qualifications they gained) as well as the occupations they had held throughout their working lives. It also inquired about their dialect usage and any second languages they had learned. The Gaelic-SSE bilinguals additionally received a modified version of the LEAP-Q (Marian. Blumenfeld \& Kauschanskaya, 2007), a questionnaire designed to determine bilingual language status that has been validated using behavioural measures of language proficiency.

Wechsler Abbreviated Scale of Intelligence (WASI): Two subscales of the WASI were used to determine participants' verbal and non-verbal IQ. The Vocabulary subscale tested the participants' verbal reasoning ability and required them to give definitions of words with increasing difficulty. The Matrix Reasoning subscale consisted of patterns designed to measure abstract non-verbal reasoning ability. Participants' raw scores were converted to $t$-scores which are normalised for each age range and combined to give an overall score from which a final IQ score was determined.

Simon Task: The Simon Task was modelled after Experiment 1 in Bialystok et al. (2004). Participants were presented with red and blue squares, half of which appeared on the left side of the screen, and the other half on the right. Participants were asked to press a key on the left (the ' 1 ' key) or the right (the ' 0 ' key) of the keyboard depending on the colour of the square. Assignment of colours to keys was counterbalanced across participants.

In congruent trials, the response associated with the colour of the square corresponded to the presentation location; in incongruent trials, the square was presented on the opposite side of the location of the response key. Thus,
in these trials participants had to inhibit the pre-potent response of selecting the spatially congruent key, and instead had to select the key associated with the colour of the square. The reaction time difference between incongruent and congruent trials is considered to be a measure of inhibitory control. Participants were given 4 congruent and 4 incongruent practice trials with feedback before moving on to the 28 critical trials (7 each of congruent red, congruent blue, incongruent red, incongruent blue) presented without feedback.

## Procedure:

Participants were first given the Background Questionnaire, which inquired about their knowledge and use of the various languages and varieties of English. The monolingual speakers were asked about their daily usage of different varieties of English and other foreign languages; for the Scottish participants these questions pertained to their use of Dundonian Scots. The responses indicated to what extent participants were fluent in one or two varieties and were using them on a daily basis. For the bilingual speakers, these questions pertained to their use of Gaelic and SSE. The bilinguals also received the LEAP-Q after the Background Questionnaire, to obtain information about their self-rated proficiency in each language, the age at which they starting learning each language, the age at which they became fluent and the proportion of time they currently use each language.

Participants were then given the Vocabulary and Matrix Reasoning subscale of the WASI. In the Vocabulary subscale, participants have to provide definitions of words. In the Matrix Reasoning subscale, participants were shown series of shapes instantiating a rule and were asked to identify which shape fits in the missing slot.

Finally, the Simon task was presented on a Toshiba laptop, with presentation controlled by Eprime. Participants first saw a fixation cross in the middle of the screen for 800 ms , followed by an interval of 250 ms . Half of the participants were randomly assigned to press the ' 1 ' key for 'red' and the ' 0 ' key for 'blue'; the assignment was reversed for the other half of participants. The keys were marked with white stickers on the keyboard. Then, a red or blue square appeared either to the left or the right of the screen, subtending five degrees of visual angle. The squares were visible for 1000 ms if there was no response. Timing began with the onset of stimulus, and was terminated with the response. The next item started after a 500 ms blank interval. The experiment began with 8 practice trials for which participants received feedback. Practice was followed by the 28 critical trials presented without feedback. Order of the 14 congruent and 14 incongruent trials was randomised.

## Results

We first compared the four groups on linguistic, demographic and cognitive measures which are presented in Table 1.

Table 1: Means and standard deviations (in parentheses) for linguistic, demographic and cognitive measures (voc: Vocabulary subscale of WASI, matrix: Matrix Reasoning subscale of WASI, skill: skill level as measure of SES,
\%use: \% daily use of Anglo-English or SSE). F denotes Fvalue in one-way ANOVAs with $\mathrm{df}=3,60$ for all conditions except the WASI subscales, where $\mathrm{df}=3,59$ ).

|  | monolinguals |  |  | biling | $\begin{aligned} & \hline F \\ & (3,60) \end{aligned}$ |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  | monoling (AngloE) | monodial (SSE) | bidial |  |  |
| age | $\begin{aligned} & 70.3 \\ & (7.6) \end{aligned}$ | $\begin{aligned} & 69.7 \\ & (7.6) \end{aligned}$ | $\begin{aligned} & \hline 72.4 \\ & (8.6) \end{aligned}$ | $\begin{aligned} & \hline 69.8 \\ & (5.5) \end{aligned}$ | 0.5 n.s. |
| voc | $\begin{aligned} & 60.3 \\ & (9.6) \end{aligned}$ | $\begin{aligned} & 57.1 \\ & (8.4) \end{aligned}$ | $\begin{aligned} & 55.5 \\ & (6.7) \end{aligned}$ | $\begin{aligned} & 57.9 \\ & (9.0) \end{aligned}$ | 0.5 n.s. |
| mat | $\begin{aligned} & 61.0 \\ & (10.5) \end{aligned}$ | $\begin{aligned} & 59.5 \\ & (10.3) \end{aligned}$ | $\begin{aligned} & 59.1 \\ & (7.7) \end{aligned}$ | $\begin{aligned} & 59.5 \\ & (5.7) \end{aligned}$ | 0.1 n.s. |
| skill | $\begin{aligned} & 3.13 \\ & (1.20) \end{aligned}$ | $\begin{aligned} & 2.88 \\ & (0.89) \end{aligned}$ | $\begin{aligned} & 2.37 \\ & (0.89) \end{aligned}$ | $\begin{aligned} & 3.37 \\ & (0.81) \end{aligned}$ | $\begin{aligned} & 3.2 \\ & \mathrm{p}<.05 \end{aligned}$ |
| \%use | $\begin{aligned} & 100.0 \\ & (0.0) \end{aligned}$ | $\begin{aligned} & 94.6 \\ & (7.3) \end{aligned}$ | $\begin{array}{r} 52.6 \\ (9.7) \\ \hline \end{array}$ | $\begin{aligned} & 44.3 \\ & (15.3) \end{aligned}$ | $\begin{aligned} & 135.7 \\ & \mathrm{p}<.001 \end{aligned}$ |

Percent language use: Participants' self-reported percentages of daily use of either Anglo-English or SSE were submitted to a one-way ANOVA to compare the four groups. This analysis yielded a significant effect of Group (see Table 1). Post-hoc tests using Tamhane's T2 for unequal variances indicated that bilinguals and bidialectals reported significantly less use of English (i.e. only an average of $48 \%$ of time) than monolinguals and monodialectals, all p's $<.001$.

Socio-economic status (SES): To determine SES, we used the 2010 Standard Occupation Classification (UK Office of National Statistics) to categorise participants' occupations into one of four skill levels based in the amount of formal qualifications or work-based training estimated to be necessary to perform the occupational tasks. These skill levels ranged from 1 (occupations requiring general education) to 4 (professional/managerial occupations requiring degree-level education). We disregarded participant income as another measure of SES as $75 \%$ of participants were retired.
A one-way ANOVA for skill levels yielded a significant effect of Group (see Table 1). Post-hoc comparisons using Tamhane's T2 for unequal variances indicated that the bilingual group had a significantly higher skill level than the bi-dialectal group, $\mathrm{p}<.05$. No other significant differences were found.

WASI: WASI scores were missing for one monodialectal participant who was unable to complete the session. One-way ANOVAs comparing performance of the groups on each of the subscales separately yielded no significant effects (see Table 1).

Simon Task: Participants committed a total of $3.4 \%$ of errors. Error rates were submitted to a 4 (Group: bidialectal, monodialectal, bilingual, monolingual) x 2 (Trial Type:
congruent, incongruent) ANOVA, which yielded no significant effects.
For correct trials, reaction times greater than 2.5 standard deviations above the mean were excluded from the analysis (Ratcliffe, 1993), which affected an additional 56 (2.9\%) of trials. For the reaction times, a 4 (Group: bidialectal, monodialectal, bilingual, monolingual) x 2 (Trial Type: congruent, incongruent) ANOVA yielded a main effect of Trial Type, $F(1,60)=80.3, \mathrm{p}<.001$ (see Figure 1). There was no main effect of Group nor was there a significant interaction between Group and Trial Type. Thus, as expected, performance on incongruent items was slower indicating that inhibiting the incongruent spatial location of the stimulus required additional effort. However, overall reaction time and Simon effect did not differ between the groups.


Figure 1: Reaction times for congruent and incongruent trials in the Simon task in bidialectal, monodialectal, bilingual and monolingual speakers. Error bars show 1 S.E.

One possible explanation for the discrepant findings between this and the Bialystok et al. (2004) study may be related to differential treatment of reaction time outliers. Bialystok et al. (2004) do not report any exclusion of outliers. To achieve comparability with that study, we repeated the ANOVA with all reaction times from the correct responses included. This analysis yielded a main effect of Trial Type, $\mathrm{F}(1,60)=9.29, \mathrm{p}<.01$, but no effect of Group and no interaction between the two factors.

Because SES is associated with executive processing (Morton and Harper, 2007), we included skill level, our measure of SES, as a covariate in the ANOVA for the reaction times (outliers excluded), which did not change the outcome of the analysis. Moreover, an analysis ignoring language group and including only skill level as the between-subjects variable did not yield any significant effects either.

In sum, while all 4 language groups showed significantly slower reaction times for incongruent trials in the Simon Task, there were no significant differences between any of
the groups in global reaction time and levels of inhibitory control.

## Discussion

Our findings did not show a bilingual advantage in nonlinguistic inhibitory control for older Gaelic-English bilinguals, nor did we find such an advantage for bidialectal speakers who routinely switch between Dundonian Scots and SSE. Moreover, we also did not find a global reaction time advantage for bilinguals and bidialectals, which has been interpreted as an indicator of improved general executive processing. This is in contrast to a substantial body of evidence demonstrating a bilingual advantage in executive processing in general, and in inhibitory control specifically. We therefore carefully compared our findings to the three other studies that had tested older bilinguals and monolinguals to determine whether differences in administration of the Simon task may have resulted in these discrepant results.

Our experiment was closely modeled after Experiment 1 in Bialystok et al. (2004). For the monolinguals, that experiment showed mean reaction times of 1437 ms for the congruent trials, and 3150 ms for the incongruent trials. For the bilinguals, the reaction times were somewhat faster (congruent: 911 ms , incongruent: 1959 ms ). These are unusually slow reaction times, in stark contrast to the much faster reaction times in our study (see Figure 1), which contained the same timing, the same number of trials, and a comparable sample size. Moreover, in Experiment 2 of the Bialystok et al. (2004) study, participants received a centered control condition and a 4 -colour condition in addition to the standard 2-colour condition, as well as an increased number of trials. Still, reaction times in the comparable 2 -colour condition were of a similarly large magnitude (older monolinguals: 1012 ms vs. 1595 ms , older bilinguals: 889 ms vs. 1101 ms , for congruent and incongruent trials, respectively). Again, these overall reaction times and the Simon effect are far beyond what is considered to be the standard Simon effect in older adults (Hilchey \& Klien, 2011; Van der Lubbe \& Verleger, 2002). This leaves open the possibility that group differences between older bilinguals and monolinguals emerge only for unusually long reaction times which may be indicative of a substantial slowing of cognitive performance in some older populations, perhaps due to diminished experience with computerised testing or due to sub-clinical effects of dementia. However, the fact that Bialystok et al. (2008) and Schroeder and Marian (2012) found a bilingual advantage in the Simon effect for older bilinguals with overall reaction times similar to the ones reported here suggests that the bilingual advantage is not an artifact of long reaction times but emerges when bilinguals and monolinguals differ in cultural background and immigrant status.

Although an analysis of other age groups is beyond the scope if this paper, it is worth mentioning that a similarly inconsistent picture emerges for studies of inhibitory control
in children, and in younger and middle-aged adults. While a considerable number of studies report a smaller Simon effect for bilinguals (Bialystok et al., 2004, 2005; Bialystok, 2006), others failed to find such a difference (Humphrey \& Valian, 2012; Kosaie \& Phillips, 2012a,b; Paap \& Greenberg, 2013). Findings of a bilingual advantage also tend to be inconsistent for other tests of executive processing (e.g. Stroop task, Flanker task, anti-saccade task) and for different aspects of executive processing (e.g. response suppression, switching, monitoring, updating - for overviews see Hilchey \& Klein, 2011, and Paap \& Greenberg, 2013). We would like to suggest that differences in cultural background and immigrant status are likely candidates for explaining the differences between monolinguals and bilinguals.

It should be mentioned that our failure to find an executive processing advantage in older Gaelic-English bilinguals contrasts with the advantage of Gaelic-English bilingual children in various measures of verbal and nonverbal IQ such as the Block design, Vocabulary and Arithmetic sub-tests of the Wechsler Intelligence Scale for Children reported by Lachlan, Parisi and Fadda (2012). As the bilingual children were all schooled in Gaelic, the authors conclude that schooling in the minority language may have consolidated their bilingualism, in contrast to a group of Sardinian-Italian bilingual children who were not schooled in the minority language and did not differ from Italian monolinguals in performance on the same tests. We agree that Gaelic schooling may indeed have been a beneficial factor for children's intellectual development as considerable resources have been expended by the Scottish government on re-introduction of Gaelic-medium education, perhaps making it more compelling for more aspirational parents to enroll their children into the better funded Gaelic tracks. For these reasons, and because psychometric intelligence does not constitute a direct measure of executive processing, we are not convinced that this finding constitutes support for superior executive processing in Gaelic-English bilinguals.

In sum, our failure to replicate an inhibitory control advantage in older Gaelic-English bilinguals and Scots-SSE bidialectals points to the importance of controlling factors like cultural background and immigrant status when studying the link between bilingualism and executive processing.

## Acknowledgements:

Part of this research was sponsored by a Carnegie Travel Grant to Neil Kirk and by Leverhulme Trust Research Project Grant RPG-375 to Vera Kempe.

## References:

Aitken, A. H. (1985). Is Scots a language? English Today, 1, 41-45.
Bialystok, E. (2006). Effect of bilingualism and computer video game experience on the Simon Task Canadian Journal of Experimental Psychology, 60, 68-79.

Bialystok, E. (2011). How does experience change cognition? Evaluating the evidence. British Journal of Psychology, 102, 303-305.
Bialystok, E., Craik, F. I. M., \& Freedman, M. (2007). Bilingualism as a protection against the onset of symptoms of dementia. Neuropsychologia, 45, 459-464.
Bialystok, E., Craik, F. I. M., Green, D. W., \& Gollan, T. H. (2009). Bilingual minds. Psychological Science in the Public Interest, 10, 89-129.
Bialystok, E., Craik, F. I. M., Klein, R., \& Viswanathan, M. (2004). Bilingualism, aging, and cognitive control: Evidence from the Simon task. Psychology \& Aging, 19, 290-303.
Bialystok, E., Craik, F., \& Luk, G. (2008). Cognitive control and lexical access in younger and older bilinguals. Journal of Experimental Psychology: Learning, Memory, and Cognition, 34, 859-873.
Bialystok, E., Martin, M. M., \& Viswanathan, M. (2005). Bilingualism across the lifespan: The rise and fall of inhibitory control. International Journal of Bilingualism, 9, 103-119.
Chang, F.-M., Kidd, J. R., Kivak, K. J., Pakstis, A. J., \& Kidd, K. K. (1996). The world-wide distribution of allele frequencies at the human dopamine D4 receptor locus. Human Genetics, 98, 91-101.
Chen, C., Burton, M., Greenberger, E., \& Dmitrieva, J. (1999) Population migration and the variation of dopamine D4 receptor (DRD4) allele frequencies around the globe. Evolution and Human Behavior, 20, 309-324.
Faraone, S. V., Doyle, A. E., Mick, E., \& Biederman, J. (2001). Meta- analysis of the association between the 7repeat allele of the dopamine $D(4)$ receptor gene and attention deficit hyperactivity disorder. American Journal of Psychiatry, 158, 1052-1057.
Green, C. S., \& Bavelier, D. (2003). Action video game modifies visual selective attention. Nature, 423, 534-537.
Hakuta, K., Ferdman, B. M., \& Diaz, R. M. (1986). Bilingualism and cognitive development: Three perspectives and methodological implications. Los Angeles: Center for Language Education and Research, University of California.
Hilchey, M. D., \& Klein, R. M. (2011). Are there bilingual advantages on nonlinguistic interference tasks? Implications for plasticity of executive control processes. Psychonomic Bulletin \& Review, 18, 625-658.
Humphrey, A. D. \& Valian, V. V. (2012) Multilingualism and cognitive control: Simon and Flanker task performance in monolingual and multilingual young adults. Paper presented at the $53^{\text {rd }}$ Annual Meeting of the Psychonomic Society, Minneapolis, MN.
Kousaie, S. \& Phillips, N. A. (2012a). Conflict monitoring and resolution: Are two languages better than one? Evidence from reaction time and event-related brain potentials. Brain Research, 1446, 71-90.
Kousaie, S. \& Phillips, N. A. (2012b). Aging and bilingualism: absence of a "bilingual advantage" in Stroop interference in a nonimmigrant sample. The

Quarterly Journal of Experimental Psychology, 65, 356369.

Lauchlan, F., Parisi, M., \& Fadda, R. (2012). Bilingualism in Sardinia and Scotland: Exploring the cognitive benefits of speaking a 'minority' language. International Journal of Bilingualism, DOI: 10.1177/1367006911429622.
Lu, C-H., \& Proctor, R. W. (1995). The influence of irrelevant location information on performance: A review of the Simon and spatial Stroop effects. Psychonomic Bulletin and Review, 2, 174-207.
Marian, V., Blumenfeld, H. K., \& Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. Journal of Speech, Language and Hearing Research, 50, 940-967.
Mezzacappa, E. (2004). Alerting, orienting, and executive attention: Developmental properties and sociodemographic correlates in an epidemiological sample of young, urban children. Child Development, 75, 1373-1386.
Morton, J. B., \& Harper, S. N. (2007). What did Simon say? Revisiting the bilingual advantage. Developmental Science, 10, 719-726.
Nikolaidis, A., \& Gray, J. R. (2010). ADHD and the DRD4 exon III 7 -repeat polymorphism: an international metaanalysis. SCAN, 5, 188-193.
Paap, K. R. \& Greenberg, Z. I. (2013). There is no coherent evidence for a bilingual advantage in executive processing. Cognitive Psychology, 66, 232-258.
Ratcliff, R. (1993). Methods of dealing with reaction time outliers. Psychological Bulletin, 114, 510-532.
Sabbagh, M.A., Xu, F., Carlson, S.M., Moses, L.J., \& Lee, K. (2006). The development of executive functioning and theory of mind: a comparison of Chinese and US preschoolers. Psychological Science, 17, 74-81.
Schachar, R., Tannock, R., Marriott, M., \& Logan, G. (1995). Deficient inhibitory control in attention deficit hyperactivity disorder. Journal of Abnormal Child Psychology, 23, 411-437.
Schroeder, S. R., \& Marian, V. (2012). A bilingual advantage for episodic memory in older adults. Journal of Cognitive Psychology, DOI: 10.1080/ 20445911.2012.669367.

Scottish Government Social Research (2010) Public attitudes towards the Scots language. Retrieved from: http://www.scotland.gov.uk/Publications/2010/01/061051 23/0. Accessed 1/2/13.
Smith, J., \& Durham, M. (2012). Bidialectalism or dialect death? Explaining generational change in the Shetland Isles, Scotland. American Speech, 87, 57-88.
Van der Lubbe, R. H. J., \& Verleger, R. (2002). Aging and the Simon task. Psychophysiology, 39, 100-110.
Williams, P. G., Suchy, Y., \& Rau, H. K. (2009). Individual differences in executive functioning: implications for stress regulation. Annals of Behavioral Medicine 37, 126140.

# Bayesian Speech Production: Evidence from Latency and Hyperarticulation 

Christo Kirov(kirov@cogsci.jhu.edu)<br>Department of Cognitive Science, 3400 N. Charles Street Baltimore, MD 21218 USA<br>Colin Wilson (colin@cogsci.jhu.edu)<br>Department of Cognitive Science, 3400 N. Charles Street<br>Baltimore, MD 21218 USA


#### Abstract

Both response latency and phonetic variation reflect competition among alternatives during the speech production process. A review of the literature finds an apparent contradiction in the latency results. In some tasks where latency is measured, similarity between targets and competitors results in slower reaction times. In other tasks, similar competitors appear to facilitate production times relative to non-similar competitors (though a lack of any competition at all results in the shortest response latencies). With respect to phonetic realization, experiments suggest that high levels of competition induced by sufficiently similar competitors result in hyperarticulation of target utterances. We present a Bayesian model of speech production that formalizes the selection and planning of spoken forms as noisy-channel communication among different levels of processing. The model resolves the apparent contradiction found in the latency results, and establishes a novel connection between those results and observed patterns of hyperarticulation.


Keywords: Speech production; competition; Bayesian modeling

## Introduction

Competition among alternatives, and the need to resolve competition efficiently and correctly, are pervasive in speech perception and speech production (e.g., Luce \& Pisoni, 1998; Marslen-Wilson \& Zwitserlood, 1989; Dell \& Gordon, 2003). A number of studies have examined how such competitive processes are reflected in the time it takes to plan speech, and in the fine-grained phonetic realization of speech sounds. The goal of this paper is to develop a unified explanation of these potentially conflicting results, which have typically been treated independently.

In various speech production tasks, response latency is affected by the relationship between the target response and any primes, distractors, or competitors in the experimental speech environment (e.g., masked priming (Ferrand et al., 1996), plan switching (Meyer \& Gordon, 1985; Yaniv et al., 1990), cue distractor tasks (Gordon \& Meyer, 1984; Galantucci et al., 2009; Roon, 2012). A review of these results reveals an apparent contradiction with respect to how similarity between targets and competitors affects response latency.

In some production tasks, similarity between target utterances and competitors results in delayed (longer) response latencies (Meyer \& Gordon, 1985; Yaniv, Meyer, Gordon, Huff, \& Sevald, 1990; Roelofs, 1999). One example is the plan-switching task (Meyer \& Gordon, 1985), in which participants are prompted to plan to say one form (e.g., the syllable UP), but are sometimes cued to say an alternative (e.g.,
the syllable UB). The findings from this task are summarized in Table 1. When the alternative response is highly similar to the originally planned target response, the time to initiate the alternative is lengthened. This effect drops off rapidly with increased phonological/phonetic distance. Only alternative responses that are about one feature away from the target seem to induce a significant delay.

Table 1: Plan Switching Task: Similarity = Higher Latency

| Planned | Alternative | Difference | Latency |
| :---: | :---: | :---: | :---: |
| UP | UB | voicing | high |
| UP | UT | place | high |
| UP | UD | voicing + place | low |

In cue-distractor tasks, on the other hand, similarity seems to play the opposite role (Gordon \& Meyer, 1984; Galantucci et al., 2009; Roon, 2012). In a cue-distractor task, participants are taught to associate a visual cue with a particular verbal response (e.g., the syllable KA or GA). Upon receiving the cue, the participant attempts to produce the associated response as quickly as possible. However, before the subject is able to initiate speech (e.g., at 200 ms after the cue), an auditory or visual distractor is presented (e.g., the syllable PA).

In spite of the fact that the subject has been given instructions to ignore the distractor, it has an effect on response latency as summarized in Table 2. It seems that when the distractor is sufficiently similar to the target response, production is facilitated relative to the case when the distractor is at a greater distance. However, it is always the case that the presentation of a distractor, no matter how it is related to the target, results in some production delay relative to the nodistractor case.

Table 2: Cue-Distractor Task: Similarity = Lower Latency

| Response | Distractor | Difference | Latency |
| :---: | :---: | :---: | :---: |
| KA | none | NA | minimal |
| KA | GA | voicing | low |
| KA | TA | place | low |
| KA | DA | voicing+place | high |

Finally, high levels of competition have been shown to influence phonetic realization: salient competitors in the speech environment give rise to hyperarticulation of spoken forms.


Figure 1: Experimental paradigm.

For example, Buz \& Jaeger (2012) found that word duration in a corpus of running speech is negatively correlated with distance to the nearest previously mentioned neighbor: neighbors mentioned in the recent past, against which the current word must plausibly compete, condition longer phonetic realizations. Baese-Berke \& Goldrick (2009), using the same paradigm as our own experiments reviewed below, found VOT lengthening for voiceless-initial target words in the context of voiced-initial neighbors (e.g., the word CAP in the context of the word GAP). ${ }^{1}$

The Baese-Berke \& Goldrick (2009) paradigm is designed to simulate a situation in which a speaker must accurately communicate a target word to a listener in the presence of contextually-salient competitor words that could delay recognition or cause miscommunication. The paradigm involves two participants, one playing the role of speaker and the other the role of listener. Each participant sits at a separate computer terminal (which is not visible to the other participant). In each trial of the experiment, two or more words appear on both screens: a target word along with competitor words that are sometimes phonological neighbors of the target. After approximately 1 s , the target word is highlighted on the speaker's screen, who then produces it aloud. At this point, the listener clicks the word that was heard - the same word produced by the speaker, if communication is successful. The speaker's pronunciation of the target is recorded and analyzed acoustically after the experiment. The experimental setup is illustrated in Figure 1. This paradigm has the advantage of being able to precisely control a target word's "context" (the neighbors that appear on-screen with it) and including motivation for the speakers to communicate clearly, as they receive feedback indicating whether the listener has selected the intended word.

Using the same paradigm, we performed a battery of experiments (see Kirov \& Wilson (2012) for details) to determine in what ways competitors could differ from the target utterance and still induce VOT hyperarticulation. The results, summarized in Table 3, point to the following generalization. Competition induces hyperarticulation only when competitors are sufficiently similar to the target word (a difference

[^122]of roughly one phonological feature). The effect drops off quickly as phonological and/or phonetic distance increases.

This nonlinear relationship between feature distance and effect size mirrors the pattern found in the plan-switching task described earlier, suggesting that both effects might be linked through a common mechanism. Although, to our knowledge, no published experiment has directly attempted to correlate response latency with VOT hyperarticulation, there is some additional evidence that latency and hyperarticulation are linked. Bell et al. (2009) suggest that lexical access latency and utterance duration are correlated. Munson (personal communication) has also found that latency in a picture naming task is a good predictor of overall vowel-space expansion: longer latencies are associated with greater vowel expansion, which is a well-known type of hyperarticulation that can be conditioned by lexical competition (e.g., Wright, 2003; Munson \& Solomon, 2004).

Table 3: Summary of Hyperarticulation Results

| Target | Competitor | Difference | Effect |
| :---: | :---: | :---: | :---: |
| CAP | DOLL | unrelated | X |
| CAP | CAD | coda | X |
| CAP | CUP | vowel | X |
| TAP | NAP | onset voicing + nasality | X |
| CAP | TAP | onset place | $\checkmark$ |
| CAP | GAP | onset voicing | $\checkmark$ |

In this paper, we present a Bayesian model of speech production that resolves the apparent contradiction present in the latency data, and links the latency results to the hyperarticulation results, explaining why these effects share the same rapid drop-off as feature distance increases between competitors in speech production. We are not aware of previous work that has attempted to unify this body of results. Indeed, Roon (2012) has recently suggested that since the planswitching task and cue-distractor tasks show different effects of similarity they must engage different levels of representation/processing. However, ascribing the effects to different processing levels would not make it clear why both tasks are sensitive to distance in the same phonetic/phonological space, and most importantly would not explain why some effects of similarity are inhibitory and others facilitative.

The proposed model posits that selection and planning of spoken forms involves optimal communication over noisy channels that link levels of mental processing/representation. Like well-known models of perception and recognition, our model takes Bayesian belief updating to be a fundamental component of psychological algorithms. This is in the spirit of other recent attempts to explore the mechanistic, as opposed to computational, utility of the Bayesian formalism (e.g., Sanborn et al., 2010).

## Bayesian Word Production Introduction

Bayesian models have been productively applied to many aspects of perception (e.g., Knill \& Richards, 1996; Girshick et al., 2011), including speech perception (Feldman, Morgan, \& Griffiths, 2009; Norris \& McQueen, 2008) and written word recognition (Norris \& McQueen, 2008). In perceptionoriented modeling, the mental system interprets noisy signals gained by the senses, updating internal beliefs about the external state as more and more evidence accumulates.

In the Bayesian word production model developed here, shown schematically in Figure 2, the signals of interest originate and are processed wholly within the mental system. Instead of interpreting noisy signals from the external world, the levels of processing/representation studied here interpret noisy messages from other levels. Each level maintains a probability distribution over representational states, receives noisy messages from one or more other levels indicating which state it should take adopt, and in turn sends noisy messages to other levels.

As is standard in Bayesian models of perception, we take noise to be an ineluctable feature of any communication system: noise is present in a signal regardless of whether that signal originates externally (from the environment, or the senses) or internally (from another mental level). One of the simplest approaches to successful transmission over a noisy channel is to use a repetition code. Repeated sampling in perception can result in a more accurate representation of the external world. For the same reason, repeated transmission of the same message to a level of processing can lead it to adopt a more functionally-appropriate representational state.


Figure 2: Bayesian Word Production Schematic

To further clarify the model, we will explain how the message passing process works, using the link between the lemma and phonology levels as an example. The construction of a message is shown schematically in Figure 3. Each
possible lemma can send a characteristic message consisting of a phonological feature vector. The simulations reported here used phonologically realistic feature representations, but for reasons of space we show only part of each vector in the figure. In the construction of a message, first one lemma is sampled from the lemma distribution. The characteristic message of that lemma is then corrupted by noise and passed to the phonology level.


Figure 3: Message Construction

The receipt of the message by the phonology level, and the way in which the message is used to update the distribution over word forms at that level, is shown in Figure 4. Each phonological form represented by this level expects a particular message. The difference between this expected message and the message received is passed through a likelihood function to determine the probability that the message received corresponds to that particular form ( $p$ (message|form)). The form of the likelihood function is determined by the type of noise that corrupts the message; in the simulations reported here we assume that noise in the word production system has a Gaussian distribution, but we have found that other types of noise are equally compatible with the experimental results (e.g., random flipping of binary feature values). Using the likelihood value, and the prior probability of each form, the level's probability distribution is updated according to Bayes' Rule:

$$
p(\text { form } \mid \text { message }) \propto p(\text { message } \mid \text { form }) p(\text { form })
$$

When simulating word production using the model, a phonological form is chosen for production when it passes a high threshold probability. In the simulations reported here, the threshold is set to 0.95 , which means that a form can be chosen only when it has $95 \%$ (or more) of the total probability after an instance of the Bayesian belief update. In most situations, a single message will not provide sufficient evidence for any form to reach this threshold after a single update. The necessary level of evidence is accumulated through multiple messages over time. This temporal repetition code for communication among mental levels will lead to accurate word


Figure 4: Bayesian Belief Updating
form selection with high probability, and lends itself well to accounting for latency and other effects observed in production.

## Resolving Latency Contradictions

We will demonstrate how the model can resolve the apparent contradiction in the latency data with the help of characteristic examples. We begin with the plan switching task, where similarity between target and competitor appears to have an inhibitory effect. There are two relevant conditions. In the first case, shown in Table 4, the participant must plan to say a target utterance (the syllable UP), but is given a cue to say a different but similar alternative (the syllable UB) instead. Initially, the distribution of forms at the phonology level favors the target utterance. After the cue, this level begins to receive messages favoring the alternative. Since the target and the alternative are very similar, the likelihood function favors both of them, and the posterior distribution after each message is received is only slightly different from the prior distribution. Thus, it takes many messages (i.e., higher latency) for the alternative to reach the threshold probability required for production.

Table 4: Similar Alternative - Plan UP with potential alternative UB: Each message causes small posterior change.

|  | UP | UB |
| :--- | :---: | :---: |
| 1) Initial state | 0.75 | 0.25 |
| 2) UB message likelihoods | 0.85 | 0.95 |
| 3) Updated state | 0.73 | .27 |

Table 5 shows the case when alternative response (UD) is
substantially different from the target (UP). Once again, the initial distribution at the phonology level favors the target. This time, however, the likelihood function responds differently to the messages received after the response cue. Since the target and alternative are substantially different, the likelihood favors the alternative but not the target. As a result, the posterior distribution after each message is received is more significantly shifted. Since the posterior distribution experiences a larger change with each incoming message, it takes many fewer messages - hence less time - for the alternative response to reach threshold probability.

Table 5: Non-similar Alternative - Plan UP with potential alternative UD: Each message causes large posterior change.

|  | UP | UD |
| :--- | :---: | :---: |
| 1) Initial state | 0.75 | 0.25 |
| 2) UD message likelihoods | 0.25 | 0.95 |
| 3) Updated state | 0.44 | .56 |

Overall then, latency is higher when the alternative response is more similar to the target, since both the alternative and the target are favored by the likelihood (i.e., there is evidence to produce both forms).

Next, we consider the cue-distractor task, where it seems a similar distractor has a facilitatory effect, relative to a different distractor. Again, there are two relevant conditions. In both cases, we will follow the setup in Roon (2012): depending on a response cue, the participant must say either KA or GA. We will assume that the KA cue is given, and some time has passed so that the distribution at the phonology level has shifted in favor of KA. In the first case, shown in Table 6, some time after the response cue the participant is presented with a distractor $(\mathbf{P A})$ similar to the target, and a few messages corresponding to the distractor are sent to the phonology level. Since the distractor is similar to the target and different from its competitors, the likelihood function provides high evidence for the target and low evidence for any competitors, resulting in a favorable shift in posterior distribution. ${ }^{2}$ Note that if the message received corresponded to the target exactly and not just a similar distractor, the target likelihood would be even higher, and the distribution would shift more favorably. Hence, latency is lowest when there is no distractor.

In the second case, shown in Table 7, the distractor presented after the cue (BA) is substantially different from the target, but similar to the alternative response. The distractor messages now provide low evidence for the target and high evidence for its competitors, causing the posterior distribution to shift in the wrong direction. Correcting this shift requires collecting more evidence for the target, resulting in greater latency.

[^123]Table 6: Similar distractor - PA: Distractor message provides more evidence for target than competitors.

|  | KA | GA |
| :--- | :---: | :---: |
| 1) Initial state | 0.75 | 0.25 |
| 2) PA message likelihoods | 0.85 | 0.25 |
| 3) Updated state | 0.91 | 0.09 |

Table 7: Non-similar distractor - BA: Distractor message provides more evidence for competitors than target.

|  | KA | GA |
| :--- | :---: | :---: |
| 1) Initial state | 0.75 | 0.25 |
| 2) BA message likelihoods | 0.25 | 0.85 |
| 3) Updated state | 0.47 | 0.53 |

In sum, a non-similar distractor causes a larger delay than a similar distractor because it provides strong evidence for the target's competitors and creates a shift in posterior probability towards them which must be overcome. There are certain situations where this generalization will not hold. In particular, if the non-similar distractor is also very different from all competitors (e.g., if the target is very similar to all possible alternatives), then it may create a smaller posterior shift towards the competitors than a similar distractor. Such situations have not arisen in the cue-distractor experiments to date, and so remain a novel prediction of the model.

Overall, we see that if speech production is a Bayesian process as proposed in this paper, the apparent contradiction found in the latency literature is resolved. In the planswitching task, similarity is inhibitory because messages for the correct form also support the originally planned competitor form. In the cue-distractor task, similarity facilitates responses because messages from the distractor are transient and favor the correct form more than any competitors.

## Linking Latency and Hyperarticulation

As shown in Figure 2, the phonology level in the model can be linked to a phonetics level that maintains a distribution over prototypical phonetic realizations. Formally, the channel between phonology and phonetics works identically to the channel between lemmas and phonology, or any other pair of connected levels. The phonology level sends messages to the phonetics level indicating which phonetic realization is preferred, and the phonetic level updates its distribution according to Bayes' rule.

The message passing between phonology and phonetics stops when a decision about which form to produce is made at the phonology level (i.e., some form achieves threshold probability). At this point, the phonetic realization of that form can be extracted as a deterministic function of the posterior distribution in the phonetic level.

Figure 5 shows the results of a series of simulations that
varied the distance between the target utterance and its closest competitor in the salient-competitor paradigm of BaeseBerke \& Goldrick (2009). As feature distance increases, there is a rapid drop-off in both the time it takes for the phonology level to settle on the target form and the value of the phonetic parameter associated with the form. This pattern arises with a variety of model parameterizations with respect to noise and likelihood functions.


Figure 5: Simulation Results: Selection time and VOT hyperarticulation as distance between target and competitor varied from 1 to 5 features.

As previously shown for the plan-switching experiment, decisions at the phonology level take longer when competitors are very similar to the target. These longer planning times allow more messages to be sent from the phonology level to the phonetics level, so the latter will ultimately be presented with a greater amount of evidence for the max VOT prototype. This results in a more skewed posterior distribution and ultimately a longer VOT value.

The crucial result is that the modeling approach presented here predicts that hyperarticulation is a mechanical consequence of planning latency, and the two are closely correlated. This would explain why both types of effects show a similarly rapid drop as feature distance between competitors increases.

## General Discussion and Future Research

We have presented a Bayesian model of word production that resolves an apparent contradiction found in latency-centered word production studies, and links latency results with results describing hyperarticulation. The fundamental claim of the model is that the selection and preparation of spoken forms should be formalized as Bayesian communication among levels of the speech production system. The model occupies a unique place in the overall space of production models, having distinct advantages and avenues for further development.

Most modeling based on interactive activation (e.g., Dell \& Gordon, 2003) has not attempted to explain latency data, focusing instead on what errors the model makes after running for a predetermined amount of time. While the Bayesian model presented in this paper can be pushed to make errors by increasing the level of noise, it is left to future research to determine if the error distribution predicted conflicts with the available empirical data.

Some models, including Roon's dynamic field theory (DFT)-based production model (2012) and Roelofs'

WEAVER++ (1997), have addressed latency results, but have not been simultaneously used to explain hyperarticulation results. In addition, the extant models do not appear to address the full range of latency effects, focusing only on those cases, such as the cue-distractor task, in which similarity between targets and competitors appears to facilitate production. It remains to be seen whether cases in which similarity has an inhibitory effect, including the results from plan-switching tasks, can be accounted for by the models in their present form.

Up to this point, we have focused on modeling empirical data where the speaker's potential utterances were limited to a small closed set. Many studies deal with situations where any word in the lexicon is a potential output utterance. These studies typically examine the effects of global lexical factors such as frequency and neighborhood density on word production. Words with low lexical frequency and high neighborhood density tend to be produced with an expanded vowel space and longer VOT (Munson \& Solomon, 2004; Wright, 2003; Goldinger \& Summers, 1989).

An important question to pursue with respect to our model (or any model) is whether or not it can scale up to explain these results. One convenient feature of the Bayesian model is that the likelihood calculation performed when updating the distribution in a level quickly rules out competitors that differ significantly from the target utterance (i.e., their likelihood is close to 0 ). This means that selection among a large open set of potential outputs quickly begins to resemble selection between a small closed set of the type used in our experiments and simulations.

## References

Baese-Berk, M., \& Goldrick, M. (2009). Mechanisms of interaction in speech production. Language and Cognitive Processes, 24, 527-554.
Bell, A., Brenier, J. M., Gregory, M., Girand, C., \& Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational english. Journal of Memory and Language, 60, 92-111.
Buz, E., \& Jaeger, T. F. (2012). Effects of phonological confusability on speech duration (poster). In The 23th CUNY sentence processing conference.
Dell, G. S., \& Gordon, J. K. (2003). Neighbors in the lexicon: Friends or foes? In N. O. Schiller \& A. S. Meyer (Eds.), Phonetics and phonology in language comprehension and production: Differences and similarities (p. 9-39). Mouton, New York.
Feldman, N. H., Morgan, J. L., \& Griffiths, T. L. (2009). The influence of categories on perception: Explainng the perceptual magnet effect as optimal statistical inference. Psychological Review, 116(4), 752-782.
Ferrand, L., Segui, J., \& Grainger, J. (1996). Masked priming of word and picture naming: The role of syllabic units. Journal of Memory and Language, 35(708-723).
Galantucci, B., Fowler, C. A., \& Goldstein, L. M. (2009). Perceptuomotor compatibility effects in speech. Attention,

Perception, \& Psychophysics, 71(1), 1138-1149.
Girshick, A. R., Landy, M. S., \& Simoncelli, E. P. (2011). Cardinal rules: Visual orientation perception reflects knowledge of environmental statistics. Nature Neuroscience, 14(7), 926-934.
Goldinger, S., \& Summers, W. V. (1989). Lexical neighborhoods in speech production: A first report. In Research on Speech Percetion Progress Report (p. 331-342). Bloomington.
Gordon, P. C., \& Meyer, D. E. (1984). Perceptual-motor processing of phonetic features in speech. Journal of Experimental Psychology: Human Perception and Performance, 10(2), 153-178.
Kirov, C., \& Wilson, C. (2012). The specificity of online variation in speech production. In Proceedings of the 34th annual meeting of the cognitive science society. Sapporo, Japan.
Knill, D., \& Richards, W. (1996). Perception as bayesian inference. Cambridge, UK: Cambridge University Press.
Luce, P. A., \& Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activativation model. Ear and Hearing, 19, 1-36.
Marslen-Wilson, W., \& Zwitserlood, P. (1989). Accessing spoken words: The importance of word onsets. Journal of Experimental Psychology: Human Perception and Performance, 15, 576-585.
Meyer, D. E., \& Gordon, P. C. (1985). Speech production: Motor programming of phonetic features. Journal of Memory and Language, 24, 3-26.
Munson, B., \& Solomon, N. P. (2004). The effect of phonological neighborhood density on vowel articulation. Journal of Speech, Language, and Hearing Research, 47, 10481058.

Norris, D., \& McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. Psychological Review, 115(2), 357-395.
Roelofs, A. (1997). The WEAVER model of word-form encoding in speech production. Cognition, 64, 249-284.
Roelofs, A. (1999). Phonological segments and features as planning units in speech production. Language and Cognitive Processes, 14, 173-200.
Roon, K. (2012). The dynamics of phonological planning. Unpublished doctoral dissertation, New York University.
Sanborn, A. N., Griffiths, T. L., \& Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. Psychological Review, 117(4), 1144-1167.
Wright, R. (2003). Factors of lexical competition in vowel articulation. In J. Local, R. Ogden, \& R. Temple (Eds.), Phonetic Interpretation: Papers in Laboratory Phonology VI (p. 75-87). Cambridge University Press.
Yaniv, I., Meyer, D. E., Gordon, P. C., Huff, C. A., \& Sevald, C. A. (1990). Vowel similarity, connectionist models, and syllable structure in motor programming of speech. Journal of Memory and Language, 29(1), 1-26.

# Transitive and periphrastic sentences affect memory for simple causal scenes 

Melissa Kline ${ }^{1}$ (mekline@mit.edu)<br>Paul Muentener ${ }^{1}$ (pmuenten@mit.edu)<br>Laura Schulz (lschulz@mit.edu)<br>Department of Bran and Cognitive Sciences, MIT<br>77 Massachusetts Avenue<br>Cambridge MA 02139


#### Abstract

Can linguistic structures influence how people perceive and remember causal events? Using a change-detection method, we presented participants with direct causal scenes paired with either transitive (He stretched the toy) or periphrastic sentences (He made the toy stretch.) Participants then viewed movies with changes to the manner of action (stretching the toy with palms up vs. down), the result (stretching it a shorter vs. longer distance), or no change. Participants judged whether the two movies were identical. Reading periphrastic sentences made people more likely to notice a change in manner than a change in result. Reading transitive sentences had the reverse effect - participants were more likely to notice changes in result. This work provides an important advance in our understanding of how rich conceptual representations map into the grammatical structures of language. We discuss how this novel method can provide insight into the nonlinguistic representations recruited by particular sentence structures.


Keywords: Causal language; event structure, change blindness, memory and language

## Introduction

How do speakers map between richly structured event representations and structured linguistic descriptions? For many kinds of events, speakers have a wide range of options. A speaker who sees a boy breaking a window with a baseball can choose to say: "The window broke"; "The boy broke the window"; "The boy broke the window with a baseball"; "The boy broke the window with a baseball during a Little League game" or, if the event was unintentional, "The boy accidentally broke the window." Each of these choices selectively highlights some aspects of the event (the result, the cause, the manner, the context, the intent, etc.) perhaps at the cost of neglecting others ("The boy broke the South Rose window of Notre-Dame"). Many theories of verb argument structure and event representation have been proposed to explain the conceptual primitives that might underlie these descriptions (Gleitman 1990; Jackendoff 1990; Levin \& Rappaport-Hovav 2005; Pinker 1989; Talmy 1985).

Although all of these sentences describe the same actual occurrence, how the speaker represents the event will
influence the type of description chosen. Conversely of course, the way an event is described influences how people represent it. Fausey and Boroditsky (2010) showed for instance that listeners were more likely to attribute blame and financial responsibility to the perpetrator of a causal event following agentive descriptions ("He broke the vase") than non-agentive descriptions ("The vase broke"). Other studies of event representations involved in language have focused on generalizations above the level of individual event-description pairings (Fausey and Boroditsky 2011; Lakusta \& Landau 2012). For instance, Fausey and Boroditsky show that English speakers are more likely than Spanish speakers to remember the perpetrator of an accidental causal event, even when the events are presented non-linguistically. They suggest that this may be because typical descriptions of accidental causal events in English ("He broke the vase") focus on the agent whereas typical descriptions of accidental causal events in Spanish do not ("Se rompió el florero", roughly "the vase broke itself"). However, because this was a purely nonlinguistic task, we cannot conclude whether the memory differences in this study were primarily an effect of these particular sentences, as opposed to other effects of language or culture.

Nonetheless, when different sentences include different components of the event (i.e., by including or omitting reference to the causal agent) it seems evident that linguistic descriptions might influence event representation (and vice versa). However, in some cases, more than one sentence is available even to describe the same components of the event (e.g., "The boy broke the window"/"The boy made the window break").

What nonlinguistic event representations might underlie linguistic distinctions like these? One factor known to influence event descriptions is the directness of the causal event. In direct causal events, the causal agent immediately impacts the causal patient. By contrast, in mediated causal events, the causal agent's action on the causal patient is less direct; for example, acting through an intermediary (e.g., a tool used to bring about the effect).

Work comparing direct and mediated causal events has predominantly examined two types of linguistic structures: lexical causatives and periphrastic causatives. While lexical causatives encode the result in the main verb of a transitive sentence ("The boy broke the window"), periphrastic

[^124]causatives ("The boy made the window break") are multiclausal and encode the result in the embedded verb. The exact syntactic differences between these sentence types are subject to the particular linguistic framework used, but the situations under which speakers tend to use each type have been studied extensively. Research has shown that adults both prefer and produce transitive sentences more often to describe direct causal events than mediated events (Wolff, 2003; Song \& Wolff 2005).

Here, we look at the impact of specific syntactic structures on adults' memory for events. We examine transitive and periphrastic descriptions because unlike agentive and non-agentive sentences, these two sentence types both encode the causal agent and the result. When both transitive and periphrastic sentences are acceptable, does sentence choice affect participants' visual memory for causal scenes?

We used a change blindness paradigm (Pashler 1988; Simons \& Levin 1998; Simons \& Chabris 1999) in which we asked participants to report whether a movie changed between the first and second viewing. Our hypotheses concerned the effect that reading different sentences would have on change detection. Wolff (2003) suggests that causal transitive sentences should lead listeners to expect direct
causal scenes. Motivated by this hypothesis, we predicted that when viewing intentional, direct causal scenes, participants who read transitive sentences (e.g., "The boy stretched the accordion") would be relatively better at detecting result changes (e.g. stretching an accordion toy a little vs. a lot) and relatively worse at noticing changes in manner (e.g. stretching an accordion toy with hands facing up vs. down). In contrast, since periphrastic (but not transitive) causal sentences can be used to describe mediated causal events, we expected that participants who read these sentences would be better at detecting manner changes.

## Experiment

## Method

Participants 329 adult participants took part in the experiment, which was conducted on Amazon's Mechanical Turk platform. Participants were screened to be located in the United States and self-reporting as native English speakers. Testing was conducted over several days, and care was taken (by monitoring user ID numbers assigned by


Figure 1. Sample stimuli images from 2 events. Each of 12 events had a base movie and 2 change movies (manner change, result change). In the "Roll" event, the manner change involved the woman switching from using one hand to using two hands. In the result change, the toy truck rolled across the table with its wheels up, rather than its wheels down. In the "Rattle" event, the manner change involved the woman changing the direction she shook the toy ring. The result change involved changing the sound the rattle made. The 10 additional event type triads were Bend, Bounce, Close, Drop, Tip Over, Ring, Rotate, Spill, Spin and Stretch. In addition to the critical movies, all participants received 6 control 'base-movie/no-change' trials (not depicted).


Figure 2. A visual depiction of the procedure for a stimulus (ROLL) in the Periphrastic x Manner-Change condition. Notice how the woman's hands are positioned in each movie. For each experimental trial $(\mathrm{n}=12)$, participants viewed the target sentence followed by a base movie. After a 5-s delay they saw the altered movie followed by the detection question.

Amazon) that participants did not take the survey more than once.

Materials We constructed 12 stimulus movie sets, each based around a simple, intentional causal action that could be described in a simple transitive sentence. Two example stimulus sets are shown in Figure 1, and videos of all stimuli used the experiment are available online at http://mit.edu/~mekline/www/KMS_cogsci13.html. In addition to the base movie, each set included a mannerchange version and a result-change version. In addition to the twelve target stimuli, six movies used in a previous study of direct, intentional causal actions (Muentener \& Lakusta, 2011) were included as control 'no-change' stimuli.

Stimuli were presented online using the Python package EconWillow (http://econwillow.sourceforge.net).

Procedure Each participant was randomly assigned to one of six conditions, crossing sentence type (Transitive, Periphrastic, Baseline/no sentence) and change type (Manner, Result) in a between-subjects design. To ensure that participants were able to view and hear the movies presented over the Internet, and to check language skills, all participants first watched a movie similar to the experimental stimuli and provided a short description. Participants were informed that they would view pairs of movies and be asked to report whether they were the same or different.

A schematic of a sample trial is shown in Figure 2. On each trial, participants were first instructed to get ready for the next movie, with the target sentence (or no sentence, in the Baseline condition) printed below. Then they saw the base movie for that stimulus - playback controls were disabled so that participants could not watch movies more
than once. After reading the target sentence again, participants performed math problems during a 5 second delay. Finally, they viewed a second movie. In the nochange trials, they simply saw the initial movie a second time. In the change trials, they saw the altered version of the movie that was appropriate for their condition (Manner or Result.) Participants were asked whether they thought the second movie was the same or different from the first, and feedback was given after every trial. In total, participants saw 12 change trials and 6 no-change trials.

## Results

To ensure that participants were not simply reporting that all movies contained changes, performance on the nochange trials was used as criteria for inclusion in the analysis. 206 participants (mean 34.3 per condition) answered at least 5 of 6 no-change trials correctly and were included in all analyses below.

Figure 3 plots participants' accuracy on change trials. There was a significant Change x Sentence interaction ( $\mathrm{F}(2$, $200)=4.54, \mathrm{p}<0.02$ ) as well as a significant main effect of Change type $(\mathrm{F}(1,200)=8.22, \mathrm{p}<0.01$.) In the Transitive condition, participants were significantly better at noticing Result changes than Manner changes $(\mathrm{t}(75)=3.53$, $\mathrm{p}<$ 0.01 ); this difference was marginal in the Baseline condition $t(63)=1.74, p=0.086$.) For Periphrastic sentences, there was no difference in accuracy between Manner and Result conditions $(\mathrm{t}(62)=0.61, \mathrm{p}=0.55)$. As predicted, a planned comparison showed that result changes were detected more often following transitive sentences and manner changes were detected more often following periphrastic sentences $(\mathrm{t}(200)=3.22, \mathrm{p}<0.01)$.


Figure 3. Participants' accuracy on the 12 change trials

Because there was an unanticipated difference in detection rates for the Baseline (no sentence) conditions, we examined detection rates on each of the 12 individual 'change' items. This difference was almost entirely due to just three items where the salience of the manner and result changes were not well matched. For these three items, the result change was easier to detect at baseline than the manner change. (Close $-\mathrm{X}^{2}=7.79, \mathrm{p}<0.01$; Drop $-\mathrm{X}^{2}=$ 7.61, $\mathrm{p}<0.01$; Tip-Over $-\mathrm{X}^{2}=8.90, \mathrm{p}<0.01$ ). To clarify the nature of the differences observed for Transitive and Periphrastic sentence conditions, we removed these three items from subsequent analyses.

Following this removal of salience-mismatched items, the only significant omnibus result was a significant Change x Sentence interaction $(F(1,200)=4.14 \mathrm{p}<0.02)$. Again, the planned comparison was significant: result changes were detected with greater success after reading transitive sentences, while manner changes were detected more often following periphrastic sentences $(\mathrm{t}(200)=2.99, \mathrm{p}<0.01)$. These results are depicted in Figure 4.

## Discussion

As predicted, the choice of transitive or periphrastic descriptions had a marked impact on participants' memory


Figure 4. Accuracy on the 9 trials which did not show a significant baseline difference in detection rate between Manner and Result changes.
for scenes. Participants who heard transitive sentences were more likely to detect changes in the result of a direct causal event than changes in the manner. Participants who heard periphrastic causal sentences showed the reverse pattern, showing better performance when detecting manner changes. This pattern persisted when three items which were not matched on baseline manner/result salience were removed.

One concern with these results is that the periphrastic is a less frequent and more complex linguistic description than the transitive. As a result, participants may have simply been more attentive to the events after they were described with periphrastic sentences. Arguing against this interpretation however, participants in the periphrastic conditions were not more attentive to event changes across the board: indeed, they were more likely to neglect result changes. However, to further address this alternative explanation, we are currently investigating participants' memory for manner and result changes when they read other complex or infrequent sentences.

Note that the effect of sentence structure on scene representation in this experiment consists of relative inattention to particular change categories. While the manner and result detection rates are different for transitive and periphrastic sentences, the more frequently detected change in each case is statistically identical to the baseline detection rate $(\mathrm{t}(63)=0.96, \mathrm{p}=0.34 ; \mathrm{t}(79)=0.23, \mathrm{p}=0.82$.) This finding is consistent with the within/between category effect found for color words (Winawer et al 2007.) Russian speakers, who have separate basic color words for light and dark blue (goluboy and siniy), showed a between-category advantage for color perception. When they were asked to distinguish between color chips that were both siniy or both goluboy, they showed decreased performance compared to color chips which were equally similar but crossed the naming boundary. English speakers showed no such advantage for dark blue vs. light blue colors.

Together with Winawer et al's study, the current results suggest that event perception helps us identify the conceptual categories that are mapped to particular linguistic structures. When no sentence is presented, both manner and result changes are considered potentially relevant. However, when people read a sentence description, a particular perspective is imposed on their event representation which seems to make some categories important and some less important. For transitive sentences, manner changes which preserve the result (e.g. bending a toy with right vs. left hand, but reaching the same final position) seem to constitute a relatively unimportant difference, and changes are neglected. In contrast, the result of the action is central to the event representation, and participants continue to notice these changes. For periphrastic causal descriptions, the reverse is true: minor changes in the result are seen as relatively unimportant whereas minor changes in the manner are seen as central to the event.

Note that Wolff's theory of causal descriptions suggests that lexical (transitive) causatives are used only for direct causal events, while periphrastics are also used to describe mediated events. Wolff focuses on how people choose between the two sentence types for different scenes but does not make specific reference to how these descriptions affect attention to the manner in which a causal event is brought about. In this study, we are able to extend this work by showing that these descriptions have a specific effect on event perception. Even when considering events that can be described with either type of sentence, participants pay more attention to how an event took place after reading periphrastic causal descriptions than after reading transitive descriptions. With this change-detection method, it will also be possible to test other event aspects that have to do with the types of events Wolff has studied, such as changes in instrument or type of agent-patient contact.

Moving beyond causal descriptions, this method can also be used to test other hypothesized correspondences between syntactic structures and particular event features or semantic concepts. After viewing sentence-event pairings, the prediction is that participants will be more sensitive to changes that have to do with the event feature representations that map to the sentence. In contrast, when changes of the same salience are made to event aspects that are not central to the sentence-event mapping, participants will fail to notice these changes. Thus, patterns of memory and attention can allow us to discover the specific semantic content of particular sentence types.

This work provides an important advance in our understanding of how rich conceptual representations map onto the grammatical structures of language, a key problem in the study of language and thought. The mapping between language and thought goes in both directions - language provides the tools to describe a wide range of event construals, and in turn, the specific descriptions we use can influence event perception, altering which components of event representations are seen as most important. By testing how memory for events changes when people encounter different types of sentences, we can experimentally discover the underlying event features which structure our cognitive and linguistic representations, and begin to understand how these representations are used in the moment to understand and describe events in the world.

## Acknowledgments

This project was supported by funding from National Science Foundation and the John Merck Scholars Fund. Thanks also to Daniel Friel and Tara Venkatesan for their assistance with stimuli creation, and to the entire Early Childhood Cognition Lab at MIT for helpful discussion and feedback.

## References

Fausey, C. M., \& Boroditsky, L. (2010). Subtle linguistic cues influence perceived blame and financial liability. Psychonomic Bulletin \& Review, 17(5), 644-650. doi:10.3758/PBR.17.5.644
Fausey, C. M., \& Boroditsky, L. (2011). Who dunnit? Crosslinguistic differences in eye-witness memory. Psychonomic Bulletin \& Review, 18(1), 150-157. doi:10.3758/s13423-010-0021-5
Gleitman, L. R. (1990). The structural sources of verb meanings. Language Acquisition, 1, 3-55.
Jackendoff, R. (1990). Semantic structures. Cambridge, MA: The MIT Press.
Lakusta, L., \& Landau, B. (2012). Language and Memory for Motion Events: Origins of the Asymmetry Between Source and Goal Paths. Cognitive Science, 36(3), 517-544. doi:10.1111/j.1551-6709.2011.01220.x
Levin, B., \& Rappaport Hovav, M. (2005). Argument realization. Cambridge, UK: Cambridge University Press.
Muentener, P., \& Lakusta, L. (2011). The intention-to-CAUSE bias: Evidence from children's causal language. Cognition, 119(3), 341-355. doi:10.1016/j.cognition.2011.01.017
Pashler, H. (1988). Familiarity and visual change detection. Perception \& Psychophysics, 44(4), 369-378. doi:10.3758/BF03210419
Pinker, S. (1989). Learnability and Cognition: The acquisition of Argument Structure. Cambridge, Mass.: MIT Press.
Simons, D. J., \& Chabris, C. F. (1999). Gorillas in our midst: sustained inattentional blindness for dynamic events. Perception, 28(9), 1059-1074. doi:10.1068/p2952
Simons, D. J., \& Levin, D. T. (1998). Failure to detect changes to people during a real-world interaction. Psychonomic Bulletin \& Review, 5(4), 644-649. doi:10.3758/BF03208840
Song, G., \& Wolff, P. (2005). Linking perceptual properties to the linguistic expression of causation. In M. Achard \& S. Kemmer (Eds.), Language, Culture, and Mind (pp. 237-250). Stanford, CA: CSLI Publications.
Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical forms. Language typology and syntactic description, 3, 57-149.
Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., \& Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. Proceedings of the National Academy of Sciences, 104(19), 7780-7785.
Wolff, P. (2003). Direct causation in the linguistic coding and individuation of causal events. Cognition, 88(1), 1-48. doi:10.1016/S0010-0277(03)00004-0

# Semantic Distance Modulates the N400 Event-Related Potential in Verbal Analogical Reasoning 

Matthew J. Kmiecik (mkmiecik@luc.edu)<br>Department of Psychology<br>Loyola University Chicago<br>1032 W. Sheridan Rd.<br>Chicago, IL 60626 USA<br>Robert G. Morrison (rmorrison@luc.edu)<br>Department of Psychology, Neuroscience Institute<br>Loyola University Chicago<br>1032 W. Sheridan Rd.<br>Chicago, IL 60626 USA


#### Abstract

Computational accounts have traditionally focused on mapping between structured representations as fundamental to analogical processing. However, a recent connectionist model has been used to argue that structured representations may not be necessary to solve verbal analogies. Green and colleagues (2010) have shown that brain areas associated with analogical mapping become more engaged as semantic distance increases between verbal analogy source and targets. Herein, we had participants verify verbal analogies characterized for semantic distance while we monitored their brain waves using EEG. Our results suggest that the semantic distance between the source and target of a verbal analogy does influence early semantic processing as reflected in the N400 Event-Related Potential. However, successfully differentiating valid and invalid verbal analogies engages areas of prefrontal cortex widely associated with inhibitory processing and the integration of abstract relations in working memory. Thus, it appears that traditional semantic priming alone is likely insufficient to explain the full extent of analogical processing.


Keywords: analogy, semantic distance, EEG, N400

## Introduction

Analogical reasoning is fundamental to the way that humans learn and reason in day-to-day life.. Likewise, analogies have long been considered to be a core component of analytic intelligence (Spearman, 1927) and of great importance in learning and discovery (Holyoak \& Thagard, 1995). For nearly a century, researchers in cognitive science have developed theories and computational models to offer potential mechanisms for analogical processing (French, 2002). More recently patient-based (Morrison et al., 2004; Krawczyk et al., 2008) and functional neuroimaging studies (e.g., Bunge et al. 2005; Bunge et al. 2009; Green et al. 2010; Krawczyk et al., 2010; Volle et al., 2010; Watson \& Chatterjee, 2012) have begun to identify a network of brain areas, particularly the prefrontal cortex (PFC), essential for analogical reasoning.

Four-term verbal analogies have long been used as both a standard measure of intelligence and vocabulary knowledge. According to traditional accounts of analogical processing, to solve this type of problem the reasoner needs to (1) retrieve word meanings from semantic memory, (2) bind words into explicit abstract relations in working memory, and (3) perform a mapping in working memory between corresponding sets of words in the source and target. For instance, to verify the analogy:

> animal : zoo :: person : house
participants may (1) retrieve the meanings of the words animal, zoo, person, and house, (2) bind housed (animal), lived-in (zoo), housed (person), and lived-in (house) (3) and then map lives-in (animal, zoo) to lives-in (person, house) specifically discovering that animal analogically maps to person and zoo maps to house. Several researchers have used this type of approach as embodied in the LISA model (Hummel \& Holyoak, 1997; 2003) to account for patterns of verbal analogy performance (Morrison et al., 2004; Michael Vendetti \& Knowlton \& Holyoak, 2012).

In contrast, recent connectionist models of analogy (Leech, Mareschal \& Cooper, 2008) have proposed that four-term verbal analogies may be solved without the use of structured relations via a mechanism utilizing guided pattern completion in semantic memory. Contrary to previous accounts of analogical priming (Spellman, Holyoak, \& Morrison, 2001), Leech and colleagues argue that this mechanism of analogy could occur automatically without the use of explicitly represented relations and analogical mapping.

In addition to many experimental studies (see Holyoak \& Hummel, 2008) the former traditional relationally explicit approach is supported by findings showing that solving verbal analogies engages anterior regions of the PFC (Bunge et al., 2005; Green et al., 2010) frequently associated with processing abstract information (e.g., Christoff et al., 2009; Nikitin \& Morrison, 2011; Volle et


Figure 1: (a) Trial timeline. Participants were instructed to think of how the A:B were related and then decide whether the C:D pair was related in the same way. Calculated ERPs were time-locked to presentation of the C:D pair. (b) C:D pairs were used for all four conditions across four blocks of trials. Valid and invalid problems were matched for semantic distance using LSA for both near and far conditions.
al., 2010) and explicit relational integration (e.g., Bunge et al., 2009; Nikitin \& Morrison, 2011; Volle et al., 2010; Watson \& Chatterjee, 2012). However, a recent set of functional magnetic resonance imaging (fMRI) studies suggest that verbal analogical reasoning may exist on a continuum between the two approaches depending on the nature of the analogies. Green et al. (2010) developed a problem set of four-term verbal analogies that varied in the semantic distance between the source and target as measured using Latent Semantic Analysis (LSA, Landauer, Foltz, \& Laham, 1998). Green and colleagues found that the anterior regions of PFC frequently associated with relational integration and/or abstract information processing were engaged to a greater extent when the source and target domains of the analogy were more distant ("far" analogies). This result suggests that "near" analogies may employ processing less dependent on structured representations.

To further explore this distinction we employed Green and colleagues' method of differentiating near and far analogies to develop a large set of verbal analogy problems for use with scalp electroencephalography (EEG). Researchers interested in the use of semantic memory during language processing have frequently used EEG analyzed with event-related potentials (ERPs) to investigate the time-course of semantic processing. Specifically, the

N400 is a negative ERP component that typically peaks around 400 ms after presentation of the stimulus. The N400 increases in negativity as a stimulus (usually a single word) becomes more incongruous from its context (Kutas \& Federmeier, 2011). The N400 effect was first documented in sentence processing. For example, the italicized word in the sentence, "The cat will bake the food" will elicit a more negative N400 relative to, "The cat will eat the food" (Kutas \& Hillyard, 1980). Many studies of language processing and semantic processing have shown the N400 to be sensitive to contextual semantic meaning (Kutas \& Federmeier, 2011). Subsequent work has shown that the N400 effect is elicited in response to conceptual incongruities in other domains. For example, incorrect answers to simple symbolic (e.g., " 4 x $4=2 l "$ ) and verbal (e.g., "Twelve plus three equals sixteen.") arithmetic problems elicit an N400 effect (e.g., Niedeggen \& Rosler, 1999).

Importantly, this type of automatic semantic congruity processing as measured by ERP methodology occurs earlier in the time course of processing than structured comparisons such as syntactic processing or analogical mapping. For instance, a positive ERP component typically peaking around 600 ms after stimulus presentation (the P600) is sensitive to violations of syntax within a sentence (e.g., "The student will studying the lecture the professor gave on tuesday."; Osterhout \& Mobley, 1995). Likewise the P600 is also sensitive to violations in structure of music (Patel et al., 1998). Likewise, Nikitin and Morrison (2011) found that an ERP component linked to the comparison of relational structures during visual analogical reasoning began approximately 500 to 600 ms post stimulus presentation, once again after the N400.

To further explore the influence of semantic distance on analogical reasoning, we recorded EEG while participants solved sequentially presented four-term verbal analogy problems (e.g. $A: B:: C D$; see Figure 1) varying in the semantic distance between the source ( $\mathrm{A}: \mathrm{B}$ ) and target (C:D) word pairs. Semantic distances between the first and second word pairs were split into near (semantically similar) and far (semantically less similar) analogies. We hypothesized that near analogies would be more likely than far analogies to be solved via automatic semantic priming and thus the N400 ERP would be less negative for near than far analogies.

## Method

## Participants

Seventeen Loyola University Chicago undergraduate students ( $\mathrm{M}=21.4$ years old) participated in the experiment. Participants gave informed consent to take part in the study and were paid according to procedures approved by the Loyola University Chicago Institutional Review Board. One participant was excluded from the analysis due to poor EEG recording quality.

## Materials

Four-term verbal analogy problems were constructed from pairs of words representing one of five possible relations: kept in (e.g. animal:zoo), kind of (e.g. aluminum:metal), made of (e.g. candle:wax), used to (e.g. train:travel), and works for (e.g. curator:museum). To ensure that word pairs were representative of the claimed relation we had an earlier group of 10 participants perform a relation naming task with a candidate list of word pairs. In the present study we only included word pairs in which participants could quickly name the stated relation from the five possibilities. Word pairs with identical relations were paired to form valid analogies and pairs representing different relations were paired to form invalid analogies.

Following the methods of Green and colleagues (2010) we further divided valid and invalid analogies based on 4term semantic distance using Latent Semantic Analysis (LSA) into either near or far analogies. LSA performs complex algorithms on large corpora of text (semantic spaces) to produce semantic similarity ratings for pairs of words (Landauer, Foltz \& Laham, 1998). The Matrix Comparison feature in LSA allows users to enter a list of $n$ terms or word pairs and produce similarity ratings between all terms or pairs of words ( $n \times n$ ) entered in the list. Similarity ratings within the source ( $\mathrm{A}: \mathrm{B}$ ) and target ( $\mathrm{C}: \mathrm{D}$ ) of each analogy problem, as well as similarity ratings between the source and target, were acquired. The source word pairs and overall analogies were characterized for both near and far semantic distance using the obtained similarity ratings from LSA.

Two counterbalanced versions of 360 unique problems were created with 90 of each type of trial. For each version every C:D word pair was used in all four conditions. To minimize the chance of confounding the N400 due to repetition effects we divided each version into quarters so that CD word pairs could be separated in time, one in each quarter. Both valid and invalid problems differed with respect to semantic distance in the same way, and this was consistent across the quarters of the experiment. Importantly, problems did not differ with respect to mean word length or word frequency as measured using HAL (Burgess \& Lund, 1997; Balota et al., 2007) across problem types.

## EEG Recording Procedure

EEG was recorded from each participant using a 72-channel Biosemi Active2 EEG system. 64 electrodes were located at equidistant locations in a nylon cap. To expand the coverage of EEG monitoring we placed two electrodes on the inferior edge of the orbit of each eye. Raw EEG was re-referenced to an average of the two mastoid electrodes and then highpass filtered at 0.01 Hz . The signal was then band-stop filtered from 59 to 61 Hz to remove any AC electrical contamination. EEG signal was corrected for ocular artifacts using a spatial PCA filter, a method available in EMSE (Source Signal Imaging, San Diego CA). Signal was further cleaned via a $\pm 100 \mu \mathrm{~V}$ rejection criterion. Included
participants had fewer than $15 \%$ of trials rejected due to EEG artifacts. A 20 Hz low-pass filter was applied to ERPs for visualization only.

## Procedure

After being prepared for the EEG recording, participants sat in a quiet room equipped with a 21 -inch CRT monitor and an electronic response box. Participants sat 100 cm from the monitor. Stimulus width was adjusted to 4 degrees of visual angle. The task was run and data were collected using e-Prime 2.0 (Psychology Software Tools, Sharpsburg, PA).

After task instructions the participant performed 20 practice trials with feedback. Each trial began with a randomly jittered fixation screen that lasted 500 to 1000 ms (See Figure 1a). The first word pair was presented at the center of the screen for 3.3 s . Participants were instructed to think of how the pair of words was related. Following an equal sign presented for 750 ms , a second pair of words was presented for 3.5 s during which participants were decide whether the two pairs of words were related in the same way (i.e., formed a valid analogy). Participants indicated their choice by pressing one of two buttons with two fingers from their right hand. The entire experiment consisted of 360 trials divided into twelve blocks separated by 20 s rests.

## Results

## Behavioral Results

Participants were more accurate in judging near than far analogies (see Figure 2a; $F(1,15)=28.6, p<.001, \eta_{\mathrm{p}}{ }^{2}=$ .66). There was no difference in accuracy with respect to validity $\left(F(1,15)=.2, n s, \eta_{\mathrm{p}}^{2}=.01\right)$; however, there was a significant interaction between semantic distance and validity $\left(F(1,15)=8.5, p=.01, \eta_{\mathrm{p}}^{2}=.36\right)$. Further contrasts suggested that this interaction was driven by participants being more accurate for valid near than valid far problems $\left(F(1,15)=24.9, p<.001, \eta_{\mathrm{p}}^{2}=.62\right)$.

Participants were also faster in judging near than far analogies (see Figure 2b; $F(1,15)=12.1, p=.003, \eta_{\mathrm{p}}{ }^{2}=$ .45), faster in judging valid compared to invalid problems $\left(F(1,15)=9.7, p=.007, \eta_{\mathrm{p}}^{2}=.39\right)$, and the two factors also interacted $\left(F(1,15)=15.9, p=.001, \eta_{\mathrm{p}}^{2}=.51\right)$. The interaction was driven by participants being faster for valid near than valid far problems $\left(F(1,15)=22.6, p<.001, \eta_{\mathrm{p}}{ }^{2}=\right.$ .60) and faster for valid than invalid near problems $(F(1,15)$ $\left.=18.5, p=.001, \eta_{\mathrm{p}}^{2}=.55\right)$.


Figure 2: Verbal analogy (a) accuracy and (b) RT.

## EEG Results

We calculated grandaverage ERPs for correct trials for each of the four conditions (see Figure 3a). Initially we divided the first 1400 ms of processing into seven 200 ms epochs and performed a 2 (near vs. far) x 2 (valid vs. invalid) x 7 (Time) repeated measures ANOVA on mean amplitudes from a central electrode (i.e., Cz ) frequently used in N400 analyses. There were reliable main effects of semantic distance $\left(F(1,15)=20 ., p<.001, \eta_{\mathrm{p}}^{2}=.57\right)$, and time $\left(F(6,90)=11.8 ., p<.001, \eta_{\mathrm{p}}^{2}=.44\right)$, and a trend towards a difference based on the validity of the analogy $(F(1,15)=$ 3.5., $p=.08, \eta_{\mathrm{p}}^{2}=.19$ ). Importantly there was a three way interaction of type, validity and time $(F(6,90)=5.1 ., p<$ $.001, \eta_{\mathrm{p}}{ }^{2}=.26$ ). As can be seen in Figure 3a far problems have more negative ERPs beginning around the N400; however, near invalid problems later join far valid and invalid problems as being more negative than near valid problems. While topographies based on valid/invalid subtractions (see Figure 3b) are broad they tend to suggest that differences in the near valid vs. invalid conditions move from more central to more right frontal distributions.

Second, we focused on the N400 and calculated mean amplitude for an early $300-500 \mathrm{~ms}$ time window typically used for analyzing the N400 in studies of semantic priming. We ran a 2 (near vs. far) x 2 (valid vs. invalid) repeated measures ANOVA. The N400 as measured in this time window was more negative for far than near problems regardless of problem validity $\left(F(1,15)=19.5 ., p<.001, \eta_{\mathrm{p}}{ }^{2}\right.$ $=.57)$, with no interaction $\left(F(1,15)=1.5, p=.24, \eta_{\mathrm{p}}{ }^{2}=.09\right)$.

To compare our results to those of Green and colleagues (2010) and to attempt to understand the time course of neural activity with respect to semantic distance and topography, we conducted an additional analysis on just correct near and far valid problems (see Figure 4). In this analysis we focused on the early $300-500 \mathrm{~ms}$ time window previously mentioned and a later $900-1100 \mathrm{~ms}$ time window closer to the response. Nikitin and Morrison (2011) have previously shown this later time window to be associated with analogical mapping in a visual analogy task. Adapting
the methods of McCarthy and Woods (1985) we normalized the subtraction of near and far mean amplitudes for each time window. A 2 (near vs. far) x 2 (early vs. late) repeated measures ANOVA demonstrated that the normalized near/far subtraction reversed from initially being greater in central areas to later being greater in frontal areas yielding a reliable interaction (see Figure $4 \mathrm{~b} ; F(1,15)=9.6, p=.007$, $\eta_{\mathrm{p}}{ }^{2}=.39$ ). Normalized subtractions showed an increase in frontal channels over time $\left(F(1,15)=5.9, p=.02, \eta_{\mathrm{p}}^{2}=.28\right)$, while central channels showed a trend towards a decrease $\left(F(1,15)=3.6, p=.08, \eta_{\mathrm{p}}^{2}=.19\right)$. Thus, we believe that Green and colleagues (2010) result that frontopolar areas were more active for far than near analogies may be driven by later processing likely reflective of the greater reliance on analogical mapping while solving far analogies.

## Discussion

As hypothesized, we found that the semantic distance between source and target word pairs in verbal analogy problems modulated the mean amplitude of the N400 ERP with near analogies eliciting less negative N400s compared to far analogies. The N400 ERP is sensitive to word repetition, semantic integration, and semantic expectancy effects (Kutas \& Federmeier, 2011). Controlling word repetition effects by utilizing identical second word pairs across all four conditions ensured semantic integration processes were isolated when analyzing N400 modulation. A more negative N 400 for far analogies can be explained by the 'knowledge integration effort' view, which suggests negativity in N400 amplitude is directly proportional to the integration effort required to extract lexical representations for each target (Holcomb, 1993). Increases in semantic integration effort in far analogies were reflected in more negative N400 mean amplitude as semantic distance between source and target analogs increased.

The knowledge integration effort view also explains the less negative N400 mean amplitude observed in near analogies. As semantic distance between source and target analogs decreased, less semantic integration effort was


Figure 3: a) Grandaverage stimulus-locked ERPs (electrode Cz ) for correct Valid and Invalid, near and far analogies. b) Topographic maps of valid/invalid subtractions for near and far analogies across the time course.

b) Early ( $\mathbf{3 0 0}-500 \mathrm{~ms}$ )

Late ( $900-1100 \mathrm{~ms}$ )

$-2.5 \mu \mathrm{~V}$
$+2.5 \mu \mathrm{~V}$
Figure 4: a) Stimulus-locked ERPs for correct valid near and far conditions. Grandaverage ERPs for each condition were calculated from clusters of frontal and central elecctrodes indicated above. b) Topographic maps of near/far subtractions for the early and late time windows.
required to derive lexical representations. In other words, deriving the semantic information of the first word pair and determining the source analog relation facilitated access to the second word pair, resulting in a less negative N400 in near analogies. A previous study demonstrated that analogous source pairs facilitated lexical access to target words when participants were instructed to attend to and use relational information (Spellman, Holyoak \& Morrison, 2001). Since participants were solving analogies and attending to the relations between words, analogical priming
may have facilitated semantic integration particularly in the near analogy condition.
However, while semantic priming may be sufficient to explain near valid analogies it may not be sufficient to reject false analogies or perform far analogies. Beginning at the N400, valid and invalid ERPs for near analogies diverge. Closer inspection of the topographies (see Figure 4) suggests engagement of areas of the brain traditionally associated with inhibitory processing during analogy (Cho et al. 2010; Watson \& Chatterjee, 2012). In fact, Morrison and colleagues (2004) have previously demonstrated that frontal patients have great difficulty rejecting lures in twochoice verbal analogy problems where semantic congruity for the false item is greater than for the true item.
Likewise, while far analogies do show a more negative N400 than near analogies, suggesting that automatic semantic processing is indexing semantic distance in analogy, there is no difference in the N400 between far valid and invalid analogies demonstrating that semantics alone are insufficient for complete analogical processing. In fact, like invalid near condition analogies, far analogies engage prefrontal cortex to a greater extent, consistent with findings by Green and colleagues (2008).
Thus, our findings suggest approaches relying solely on tranditional semantic priming, such as recent connectionist approaches (e.g., Leech, Mareschal \& Cooper, 2008),have limited applicability when the distance between the source and targets of analogies increases, or when the reasoner must choose between alternative analogues where semantics alone do not indicate the more relationally similar match.

## Acknowledgements

The authors would like to thank Ryan Brisson and Krishna Bharani for technical support. The Provost and the Mulchay Scholars Program of Loyola University Chicago (MJK, RGM), the American Federation of Aging Research/Rosalinde and Arthur Gilbert Foundation (RGM), and the Illinois Department of Public Health (RGM) provided support.

## References

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., . . . Treiman, R. (2007). The english lexicon project. Behavior Research Methods, 39(3), 445459.

Bunge, S. A., Helskog, E. H., \& Wendelken, C. (2009). Left, but not right, rostrolateral prefrontal cortex meets a stringent test of the relational integration hypothesis. NeuroImage, 46(1), 338-342.
Bunge, S. A., Wendelken, C., Badre, D., \& Wagner, A. D. (2005). Analogical reasoning and prefrontal cortex: Evidence for separable retrieval and integration mechanisms. Cerebral Cortex (New York, N.Y.: 1991), 15(3), 239-249.
Burgess, C., \& Lund, K. (1997). Modeling parsing constraints with high-dimensional context space. Language and Cognitive Processes, 12, 177-210.

Cho, S., Moody, T. D., Fernandino, L., Mumford, J. A., Poldrack, R. A., Cannon, T. D., Knowlton, B. J., \& Holyoak, K. J. (2010). Common and dissociable prefrontal loci associated with component mechanisms of analogical reasoning. Cerebral Cortex, 20, 524-533.
Christoff, K., Keramatian, K., Gordon, A. M., Smith, R., \& Madler, B. (2009). Prefrontal organization of cognitive control according to levels of abstraction. Brain Research, 1286, 94-105.
French, R. M. (2002). The computational modeling of analogy-making. Trends in Cognitive Sciences, 6(5), 200205.

Green, A. E., Kraemer, D. J. M., Fugelsang, J. A., Gray, J. R., \& Dunbar, K. N. (2010). Connecting long distance: Semantic distance in analogical reasoning modulates frontopolar cortex activity. Cerebral Cortex, 20(1), 70-76.
Holcomb, P. J., \& Anderson, J. E. (1993). Cross-modal semantic priming: A time-course analysis using eventrelated brain potentials. Language and Cognitive Processes, 8(4), 379-411.
Holyoak, K. J., \& Thagard, P. (1995). Mental leaps: Analogy in creative thought. Cambridge, MA, US: The MIT Press.
Holyoak, K. J.,\& Hummel, J. E. (2008). No way to start a space program. Commentary on R. Leech, D. Mareschal and R. P. Cooper, Analogy as relational priming: A developmental and computational perspective on the origins of a complex skill, Behavioral and Brain Sciences, 31, 388-389.
Hummel, J. E., \& Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. Psychological Review, 104(3), 427-466.
Hummel, J. E., \& Holyoak, K. J. (2003). A symbolicconnectionist theory of relational inference and generalization. Psychological Review, 110(2), 220-264.
Krawczyk, D. C., Morrison, R. G., Viskontas, I., Holyoak, K. J., Chow, T. W., Mendez, M. F., . . . Knowlton, B. J. (2008). Distraction during relational reasoning: The role of prefrontal cortex in interference control. Neuropsychologia, 46(7), 2020-2032.
Kutas, M. \& Federmeier, K.D. (2011) Thirty years and counting: Finding meaning in the N400 component of the Event-Related Brain potential (ERP), Annual Review of Psychology Vol. 62: 621-647.
Kutas, M., \& Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. Science, 207(4427), 203-205.
Landauer, T. K., Foltz, P. W., \& Laham, D. (1998). An introduction to latent semantic analysis. Discourse Processes, 25(2-3), 259-284.
Leech, R., Mareschal, D., \& Cooper, R. P. (2008). Analogy as relational priming: A developmental and computational perspective on the origins of a complex cognitive skill. Behavioral and Brain Sciences, 31(4), 375-378.
McCarthy, G., \& Wood, C. C. ~1985!. Scalp distributions of event-related potentials: An ambiguity associated with
analysis of variance models. Electroencephalography and Clinical Neurophysiology, 62, 203-208.
Morrison, R. G., Krawczyk, D. C., Holyoak, K. J., Hummel, J. E., Chow, T. W., Miller, B. L., \& Knowlton, B. J. (2004). A neurocomputational model of analogical reasoning and its breakdown in frontotemporal lobar degeneration. Journal of Cognitive Neuroscience, 16(2), 260-271.
Morrison, R.G., Holyoak, K.J., \& Truong, B. (2001). Working memory modularity in analogical reasoning. Proceedings of the Twenty-fourth Annual Conference of the Cognitive Science Society (pp. 663-668). Mahwah, NJ: Erlbaum.
Niedeggen, M., Rosler, F., \& Jost, K. (1999). Processing of incongruous mental calculation problems: Evidence for an arithmetic N400 effect. Psychophysiology, 36(3), 307324.

Nikitin, S., \& Morrison, R. G. (2011). Analogical reasoning in human prefrontal cortex: an event-related potential approach. Cognitive Neuroscience Annual Meeting, San Francisco, CA.
Osterhout, L., \& Mobley, L. A. (1995). Event-related brain potentials elicited by failure to agree. Journal of Memory and Language, 34(6), 739-773.
Patel, A. D., Gibson, E., Ratner, J., Besson, M., \& Holcomb, P. J. (1998). Processing syntactic relations in language and music: An event-related potential study. Journal of Cognitive Neuroscience, 10(6), 717-733.
Spearman, C. (1927). The abilities of man; their nature and measurement. New York: Macmillan Co.
Spellman, B. A., Holyoak, K. J., \& Morrison, R. G. (2001). Analogical priming via semantic relations. Memory \& Cognition, 29(3), 383-393.
Vendetti, M., Knowlton, B. J., \& Holyoak, K. J. (2012). The impact of semantic distance and induced stress on analogical reasoning. Cognitive, Affective, and Behavioral Neuroscience, 12, 804-812.
Volle, E., Gilbert, S. J., Benoit, R. G., \& Burgess, P. W. (2010). Specialization of the rostral prefrontal cortex for distinct analogy processes. Cerebral Cortex, 20(11), 2647-2659.
Watson, C.E. \& Chatterjee, A. (2012). A bilateral frontoparietal network underlies visuospatial analogical reasoning. NeuroImage, 59: 2831-2838

# Multiple Proposal Memory in Observational Word Learning 

Judith Koehne (judith.koehne@uni-bamberg.de)<br>Otto-Friedrich-Universität Bamberg<br>Markusplatz 3, 96045 Bamberg

John C. Trueswell (trueswel@psych.upenn.edu)<br>Lila R. Gleitman (gleitman@psych.upenn.edu)<br>Institute for Research in Cognitive Science, University of Pennsylvania, 3401 Walnut Street, Suite 400A, Philadelphia, PA, USA


#### Abstract

The temporal co-occurrence of a novel word and a visual referent undoubtedly facilitates establishing the meaning of a word. It is less understood, however, how precisely learners can keep track of the frequencies of these co-occurrences across situations. Observational learning may rely on one or few highly informative exposures (propose-but-verify) or it may be driven by the collection of evidence in a more gradual and parallel manner (multiple-hypotheses tracking). We evaluated both hypotheses within two experiments and found that learners were able to keep track of more than one hypothesis for a novel word. However, this memory was strongly dependent on each learner's individual learning path (i.e., which meanings they had considered before) and influenced by the order of presentation of potential referents. We argue for an account of a multiple-proposal memory rather than a multiple co-occurrence memory.


Keywords: observational word learning; memory; crosssituational analysis; multiple hypotheses tracking; propose-but-verify; individual learning paths

## Observational Word Learning

While observing the world can be a very direct path to the meaning of a novel word (fast mapping, Carey, 1978), the relationship between both sources of input is often too ambiguous to make a promising immediate guess. The learner could solve this problem in various ways: On each learning instance, she could store multiple possible solutions and then identify the best solution across several learning instances through an intersective process, an assumption that is commonly understood to underlie the idea of crosssituational word learning (Quine, 1960; Yu \& Smith, 2007). Alternatively, she could make an immediate guess about the word's meaning and wait for confirmation or rejection. In this case, the learner would have no memory for the alternatives that were not guessed, but maximally a memory for the different guesses tried along the way until the correct one is identified.

While experiments reported in Medina, Snedeker, Trueswell, \& Gleitman (2011) and Trueswell, Medina, Hafri, \& Gleitman (2013) support this latter idea (propose-butverify account), other studies indicate that learners are able to extract multiple hypotheses on each learning instance (Vouloumanos, 2008; Vouloumanos \& Werker, 2009; Koe-
hne \& Crocker, 2011). An important aspect that is ignored in these studies, however, is the role that each learner's individual learning path plays. It is therefore unclear whether one and the same person in fact stores multiple possibilities for a word and if this is the case, under which circumstances. One factor that has been shown to be relevant to this question is the order in which the language novice has encountered and re-encountered potential referents (Medina et al., 2011).

We evaluate the way learners exploit observational word learning situations within two experiments, employing the standard paradigm of psychologically investigating crosssituational word learning. Importantly, we consider both the learner's individual learning path and the order of exposures and re-exposures of potential referents. We moreover address the possibility that the different outcomes in different studies may be due to the implemented experimental procedure. In particular, we compare a procedure, in which participants make a choice on each learning trial (Exp. 1) to a passive look-and-listen learning phase (Exp. 2).

## Learning based on Co-occurrence Frequencies

Trueswell et al. (2013) examined learners' memory in observational learning situations in a series of experiments. During the learning phase, participants were presented two or five visual referents and a spoken sentence containing one novel noun per trial. The task was to choose that referent (by mouse click) in each trial that the learner believed to match with the novel noun. Trueswell et al. found that even if the learning situations were greatly simplified but still ambiguous (just two possible referents), participants later showed no sign of memory for any referent other than the one they had selected. Specifically, when a learner reencountered a noun (e.g., mipen), he was at chance at selecting the correct referent (e.g., bear) if he had made the wrong choice the previous time he had encountered mipen (e.g., if he had chosen door rather than the correct bear). Had he remembered that the unselected (but correct) referent (bear) had co-occurred with mipen, he could have unambiguously identified it as correct in the current situation.

Interestingly, other studies indicate that learners are able to precisely differentiate the co-occurrence frequencies of different alternatives for one noun. Vouloumanos (2008)
employed a passive look-and-listen learning phase with one referent and one noun per trial. Over the course of the experiment, each noun co-occurred with several referents with varying frequencies. In a final forced-choice vocabulary test, learners could differentiate between these alternatives based on small differences in their co-occurrence statistics. However, since there was only one referent per learning trial, this study does not answer the question whether multiple possibilities are memorized from one situation. Addressing this issue, Koehne \& Crocker (2011) integrated a learning procedure with four objects depicted for each novel noun. As in Vouloumanos (2008), nouns co-occurred with objects with different frequencies ( $83 \%, 50 \%$, and $17 \%$ ). Interestingly, when the $83 \%$ referent was not available in a final forced-choice test, learners preferred the $50 \%$ referent over $17 \%$ alternatives. This result suggests sensitivity to differences in co-occurrence statistics even when learning trials are ambiguous.

Differences between Trueswell et al. and Koehne \& Crocker could be due to the experimental procedure (forced choice vs. look-and-listen during learning). However, individual learning paths were not considered in Koehne \& Crocker: It is unclear whether selecting the $50 \%$ referent depended on the choices, or proposals, the learner had made before and whether one and the same learner had stored multiple alternatives for one noun.

Indeed, as noted by Trueswell et al. (2013), the strictest version of a propose-but-verify procedure, in which only a single meaning is ever maintained, is inadequate because it fails to explain the learning of ambiguous words. They therefore propose that "when a confirmed (and even reconfirmed) hypothesis for a word is then not supported by a later context, the learner would actively search memory for past rejected hypotheses, and may ... establish a second meaning for the word." Here we call this multiple-proposal memory, in which only previously proposed meanings are available in memory rather than entire referential sets from past learning instances (i.e., the context) as stipulated by the most common cross-situational accounts.

The method used in the two experiments presented here allows us to differentiate between the predictions from the propose-but-verify versus the multiple-hypothesis-tracking account. In particular, it addresses the question whether one and the same learner keeps track of more than one hypothesis for a novel word.

## Experiment 1

Experiment 1 addresses the questions of whether and how learners track multiple meanings for a novel word and what role both the learning path and the order of (re-) exposures of potential referents play in this process.

## Methods

Participants 36 participants were tested, four of which had to be excluded due to technical and eye-tracking problems.

Data of 32 participants (11 Male, average age 22) was analyzed.

Design, Materials, \& Procedure The overall task of Experiment 1 was to learn the meanings of 16 novel nouns. Learning trials consisted of one spoken English sentence containing one of the novel words (e.g., I see a moke!) and four objects that were depicted on the screen. During training, each noun had six learning trials, intermixed with the other learning trials. Crucially each of the 16 nouns was assigned two meanings with different co-occurrence frequencies: One referent was present whenever the noun was present (six times, $100 \%$ referent, e.g., television), the other referent was present in only half of the cases the noun was (three times, $50 \%$ referent, e.g., dog). All other objects cooccurred only once with a noun (17\%). We manipulated the order in which trials including and excluding the $50 \%$ referent were presented within four levels (within participants): Firstly, the $50 \%$-present (P) and $50 \%$-absent (A) trials could be either blocked (AAAPPP and PPPAAA) or not blocked (APAPAP and PAPAPA); secondly, the first encounter of a noun could be either an A trial (AAAPPP and APAPAP) or a P trial (PPPAAA and PAPAPA).

On each learning trial, participants selected by mouse click the referent they thought belonged to the novel noun. After each response, they gave a confidence rating for their selection (on a scale from 1 to 9 ). No feedback was given and participants were not informed that nouns may have multiple meanings.

After all six learning trials had been encountered for a word, a final test was given for each word, in which eight objects and one spoken word were presented and learners were asked to again select the matching referent and indicate their confidence. The $100 \%$ referent, however, was not available which means that the $50 \%$ object was the one with the highest co-occurrence rate - all other objects were $0 \%$ and $17 \%$ referents.

The experiment consisted of two parts: Eight novel nouns were taught and tested (Block 1) before the other eight noun were taught and tested (Block 2). Order of presentation of learning and test trials was pseudo-randomized: Between two exposures of the same noun, there was always at least one but not more than 8 trials with other nouns. Participants were run individually and the experiment lasted approximately 30 minutes.

Predictions Standard cross-situational accounts (such as Yu \& Smith, 2007, Vouloumanos, 2008, and Koehne \& Crocker, 2001) predict that learners precisely keep track of the co-occurrence frequencies between nouns and referents. The $50 \%$ referent should therefore be chosen at final test above chance in all conditions, independent of both the learning path and the order of (re-)exposures of $50 \%$-present trials.

According to a strict propose-but-verify account selection of the $50 \%$ alternative at final test would occur if and only if it is the current working hypothesized meaning - that is, if
the $50 \%$ referent had been selected on the preceding learning instance. This is impossible when the $50 \%$ referent is Absent on the last learning trial, predicting chance performance in conditions PPPAAA and PAPAPA. During the final test in the other two conditions, the $50 \%$ alternative would be selected above chance on those rare occasions when the learner had selected the $50 \%$ referent on the last learning instance (i.e., when they failed to learn the $100 \%$ target by Instance 6).

According to the weaker propose-but-verify account, a final test with the $100 \%$ referent absent will trigger consideration of all past proposed meanings. This means that abovechance performance on the $50 \%$ referent is expected if and only if the learner had previously selected (clicked on) a $50 \%$ referent during the learning phase. One might expect such a memory to have a recency component: More recently proposed meanings will be easier to remember. Moreover, early encounters of a string of $50 \%$ referents (i.e., PPPAAA) will increase the probability that this referent will be selected during the learning phase and thus more likely to be recalled at test. Conversely, it is very unlikely during learning that the $50 \%$ referent will be selected on any trial when these occurrences are grouped late in the sequence (AAAPPP): Most learners will have already locked onto the $100 \%$ item as the referent by Instance 4, and thus rarely select the $50 \%$ referent during learning. Therefore, they will not select it at test either.

## Data Analysis, Results, \& Discussion

The results are most consistent with the weaker propose-but-verify account. Across conditions, participants selected the $50 \%$ referent in the final test significantly more often than chance $(25.4 \%$ vs. $12.5 \%$ chance; $t(31)=7.77, p<$ $.001)^{1}$. Both confidence ratings and reaction times support that this difference is meaningful: Ratings were significantly higher $(\chi 2(1)=17.87, p<.001)$ and reaction times were significantly lower $(\chi 2(1)=9.36, p<.01)$ when the $50 \%$ referent was chosen than when it was not. Moreover, it was chosen significantly more often than any other of the seven ( $0 \%$ and $17 \%$ ) objects.

While this trend holds for all four conditions, differences to chance were significant only in Conditions PPPAAA $(34.4 \% ; t(31)=4.91, p<.001)$, PAPAPA $(27.3 \% ; t(31)=$ $4.32, p<.001)$, and APAPAP ( $23.4 \% ; t(31)=2.80, p<.05)$ but not in AAAPPP $(16.4 \% ; t(31)=1.02, p=.32$; Figure 1). This finding is inconsistent with a standard cross-situational account because all conditions should have been above chance independent of presentation order. It is also inconsistent with the strict propose-but-verify account because PPPAAA and PAPAPA ought not be above chance, but they are. Consistent with the weaker propose-but-verify, PPPAAA offers the best performance overall whereas AAPPP offers the worst.

To get insight into the roles of ordering and learning paths on the final test, we analyzed the effects of Condition and

[^125]Previous Selection of the 50\% Referent, that is, whether the $50 \%$ referent had been chosen in the previous encounter when it had been present. Note that this trial was in different positions depending on condition: It was the last trial in Conditions AAAPPP and APAPAP in Condition PAPAPA, and the fourth to last trial in Condition PPPAAA.


Figure 1: Selections in test, Exp. 1
Consistent with the weaker version of propose-but-verify, we found that participants were only above chance at selection of the $50 \%$ referent at test if they had selected the referent on its last encounter during learning (Figure 2). Note that the number of observations contributing to each proportion differs in the way expected if learners were using the weaker propose-but-verify procedure during the learning phase; the $50 \%$ referent was selected on the previous encounter during learning only 24 times (out of 128) in AAAPPP, but 56 times in PPPAAA. In APAPAP it was chosen 27 times and in PAPAPA 32 times. If selected during learning however, it was recalled at final test at similar rates regardless of condition (i.e., Figure 2).

To confirm the reliability of the effects in Figure 2, we conducted a multi-level logistic regression using Condition and Previous 50\% Referent Accuracy as predictors of selecting the $50 \%$ referent at test, entering both as fixed effects (using the lme4 package in R, Bates, 2005). Random intercepts and slopes of Subjects and Items were integrated. If a model did not converge, random effects were reduced until convergence was reached (always discarding the random effect with the smallest effect). Main effects were tested using model comparison (Chi-Square values are reported; Baayen, Davidson, \& Bates, 2008). We found a significant effect of Previous 50\% Referent Accuracy only $(\chi 2(1)=$ 80.21, $p<.001$ ) but no effect of Condition $(\chi 2(3)=3.67, p$ $=.30)$ and no interaction $(\chi 2(3)=4.52, p=.21)$. T-tests confirm that for that subset of trials for which it was not the case that the $50 \%$ referent had been chosen in the previous learning trial in which it had been present, selecting the $50 \%$ referent was not above chance $(t(31)=-.68, p=.50)$. This reveals that, independent of condition, the $50 \%$ referent was only chosen reliably if it had also been chosen in the previous encounter for which it had been present.

Interestingly, 50\% selection was still above chance when it additionally was the case that the $100 \%$ referent had been chosen two to five times during learning $(t(25)=6.43, \mathrm{p}<$ .001). This means that one and the same learner could consider the $100 \%$ referent as the correct referent and still be sensitive to the fact that the $50 \%$ referent was a better candidate than the $17 \%$ objects as long as the $50 \%$ referent, as well, had been considered.

This pattern of results supports the weaker version of the propose-but-verify account: While in fact a referent is only stored as the potential meaning if it has been actively considered before, this consideration does not need to happen in the absolutely previous encounter of the noun but only in the last common encounter of the noun and that referent. This means that learners do not only memorize the last guess they made for a noun but also less recent guesses. Our results are clearly not in line with the hypothesis that learners are equipped with a general multiple co-occurrence memory.


Figure 2: 50\% referent selections in test, Exp. 1
It is possible that the results from Experiment affected by the employed learning procedure: Forcing a selection on each trial may enforce the influence of the learning path (i.e., previous accuracy). We address this possibility in Experiment 2.

## Experiment 2

Experiment 2 investigates whether learning path and conditions have the same effect on memorizing potential referents if learners are not forced to make a choice on learning trials.

## Methods

Participants 39 participants were tested, seven of which had to be excluded due to technical and eye-tracking problems. Data of 32 participants ( 16 Male, average age 23) was analyzed.

Design, Materials, \& Procedure The learning paradigm, design, materials, and procedure were exactly the same as in Experiment 1 except that participants were asked to simply look and listen during learning trials while trying to figure out what the novel nouns mean. As in Experiment 2, however, trial change was self-paced (elicited by button press). Moreover, participants' eyes were tracked using a Tobii 1750 eye-tracker (sampling rate 50 Hz ).

Predictions Hypothesizing that clicking does not influence the learner's behavior predicts that one will find the same results as in Experiment 1. Hypothesizing that clicking enforces previous accuracy to be crucial on the other hand predicts a weaker effect of the learning path on the memory for the $50 \%$ referent.

## Data Analysis, Results, \& Discussion

Selecting the $50 \%$ referent in the test again was significantly more frequent than would be expected by chance ( $22.7 \%$ vs. $12.5 \% ; t(31)=6.07, \mathrm{p}<.001$ ) and than selecting any of the other candidates. As in Experiment 1, confidence ratings were higher $(\chi 2(1)=13.12, p<.001)$ and reaction times were lower $(\chi 2(1)=5.12, p<.05)$ when the $50 \%$ referent was selected than when another object was chosen.

Selection rates were (at least marginally) significantly above chance in all four conditions (PPPAAA: 29.9\%, $t(31)$ $=4.53, \mathrm{p}<.001$; PAPAPA: $18.8 \%, t(31)=1.76, \mathrm{p}=.09$; APAPAP: $22.7 \%, t(31)=3.13, \mathrm{p}<.01$; AAAPPP: $19.5 \%$, $t(31)=1.83, \mathrm{p}=.08$, Figure 3$).$


Figure 3: Selections in test, Exp. 2
To evaluate the effect of the learning path although no choices were made during learning, we used learners' eye movements as a predictor: Specifically, we coded test trials for the frequency of $50 \%$-present learning trials in which the $50 \%$ referent had been fixated more often than any of the three other candidates after the novel noun was presented (i.e., from onset of the noun until the self-paced end of the trial). The rationale of this coding was that looking at a referent most reveals that participants had paid attention to it, indicating that it was selected as the potential referent.

We then included this measurement of Previous 50\% Referent Accuracy as a predictor, together with Condition (Fig-
ure 4). Similar to Experiment 1, we found that choosing the $50 \%$ referent at test was not predicted by Condition $(\chi 2(3)=$ $.96, p=.31$ ) but by Previous 50\% Referent Accuracy ( $\chi 2(1)$ $=7.49, p<.01)$. Again, there was no interaction $(\chi 2(3)=$ $0.89, p=.83$ ). And again the number observations across conditions patterned like in Experiment 1 in terms of how often the $50 \%$ referent was 'selected' (by eye) on its last occurrence during learning ( $\mathrm{N}=22$ for AAAPPP; $\mathrm{N}=44$ for PPPAAA; $\mathrm{N}=24$ for APAPAP; and $\mathrm{N}=23$ for PAPAPA).


Figure 4: 50\% referent selections in test, Exp. 2
Interestingly, however, the $50 \%$ referent was still chosen significantly more often than chance at test if it was not looked at most often in the previous encounter $(t(31)=2.43$, $\mathrm{p}<.05$ ). We therefore also coded test trials for whether the $50 \%$ referent had been looked at most in any (i.e., at least one) learning trial (Any Accuracy). We found that if this was not the case, selecting the $50 \%$ referent was not more frequent than chance $(t(21)=-.21, p=.83)$. Any Accuracy was a marginally significant predictor $(\chi 2(1)=3.44, p=.06)$ whereas Condition was not $(\chi 2(3)=3.45, p=.33)$ and both did not interact $(\chi 2(3)=1.27, p=.74)$.

Similar to Experiment 1, having looked at the $100 \%$ referent most often in two to five learning trials did not change this pattern: The $50 \%$ referent was still chosen significantly more often than chance as long as it was also looked at most at least once $(t(31)=3.80, p<.001)$.

These results suggest that learners' behavior when choices were not forced during learning was similar to their behavior when they were forced to respond (i.e., as in Experiment 1). While it may be less crucial that the $50 \%$ referent was paid particular attention to exactly the last time it was encountered, the data indicates that it is was necessary that it at some point in learning it had been attended to. While this difference could suggest that memory in Experiment 2 was better than in Experiment 1 (i.e., that learners stored all proposals rather than only the last one), the different measurements of Previous 50\% Referent Accuracy cannot be perfectly compared with one another.

Most important, however, is that even if the learner is not forced to make decisions during learning, it is still crucial for a potential referent to be paid particular attention to at some point. We interpret this as a confirmation of our findings from Experiment 1: Learners show no sign of a general multiple co-occurrence memory but they are able to memorize more than one proposal they have made.

## Analyses Experiments 1 \& 2

In order to evaluate a potential difference between Experiments 1 and 2 regarding the influence of Condition, we entered data from both into one analysis. Experiment (Experiment 1: click vs. Experiment 2: no click) and Condition were used as fixed factors. We found a marginal effect of Condition ( $\chi 2(3)=7.61, p=.06$ ), no effect of Experiment $(\chi 2(1)=1.59, p=.21)$, and no interaction $(\chi 2(3)=3.22, p=$ .36 ; Figure 5). We then grouped the four conditions into two: $50 \%$ present in first trial (PPPAAA \& PAPAPA) versus $50 \%$ absent in first trial (AAAPPP \& APAPAP) and repeated the analysis. While selecting the $50 \%$ object was significantly more frequent in the first-trial present than the first-trial absent conditions ( $\chi 2(1)=6.63, p<.05$ ), still neither an effect of Experiment $(\chi 2(1)=1.04, p=.31)$ nor an interaction was found $(\chi 2(1)=2.31, p=.13)$. Within experiments, however, both condition groups differed significantly only for Experiment $1(\chi 2(1)=8.04, p<.01)$ but not for Experiment $2(\chi 2(1)=0.51, p=.47)$. It is therefore not quite clear whether the order of exposure and re-exposure was equally meaningful to both Experiments. Possibly, it was slightly more important in Experiment 1 than Experiment 2 that a referent's first encounter happened early, as also indicated by the missing significance of selecting the 50\% referent in Condition AAAPPP in Experiment 1 (Figure 1). Either way, for both experiments, the effect of Previous Accuracy was a much clearer predictor than Condition.


Figure 5: 50\% referent selections in test, Exp. $1 \& 2$

## Conclusions \& General Discussion

Results from Experiments 1 and 2 reveal that learners successfully learned to differentiate between co-occurrence
frequencies of $50 \%$ versus $17 \%$ and $0 \%$ even though another referent co-occurred perfectly ( $100 \%$ ). However, this was only the case if the $50 \%$ referent was in the learner's attention at least once before (or if it even was actively selected). Importantly, the $50 \%$ referent was also stored even if it was not the only referent that the learner had considered (i.e., when both the $100 \%$ referent and the $50 \%$ referent were in the learner's focus of attention at some point during learning). These findings clearly reveal that while cooccurrences were not generally all stored, 'multiple proposal' memory is possible in observational word learning. This is not in line with the standard cross-situational account whereas it generally supports a propose-but-verify account. Interestingly though, selecting the $50 \%$ referent was above chance in conditions PPPAAA and PAPAPA; unlike a strict propose-but-verify theory would predict, learners can memorize more than the most recent choice they have made.

Asking participants to select a referent during learning trials did not generally suppress memorizing multipleproposals. While it may be the case that it is more important for a forced-choice learning procedure than the non-forced choice one that the $50 \%$ referent is considered exactly in the previous encounter of it, a clear comparison between choosing and looking is impossible. If the difference is real, it would indicate that forcing a choice enhances the role of previous consideration, possibly because a stronger memory trace is built by actively (and physically) making a selection than by mental consideration.

Our results moreover at least indicate that there is a possible influence of the order in which referents are firstly encountered and re-encountered: Early on, when the hypotheses space is still completely open, learners are more willing to memorize co-occurring objects as potential meanings than later, when other hypotheses (or considerations) have already been made for a novel noun. This may be more strongly the case when selections are forced even early on in learning (in Experiment 1).

## Summary

We investigated learners' memory for co-occurrence frequencies in referentially ambiguous observational-word learning situations within two experiments. Our data reveals that while participants were able to recall more than one potential meaning for a noun, this memory was dependent on the person's single considerations during learning: Only if a potential meaning had been proposed before (i.e., selected or paid particular attention to), it was stored. However, learners memorized more than the most recent proposal they had made for a novel word. Moreover, a meaning was more likely to be proposed if it co-occurred with a noun early on the learning path. While this whole pattern was very similar independent of the learning procedure (choice made during learning, Experiment 1, vs. no choices made, Experiment 2), the influence of being proposed early may be enhanced when choices are made. In line with a moderate version of the propose-but-verify account (Medina et al.,

2011; Trueswell et al., 2013), our results can be accounted for by a multiple-proposal memory rather than a multiple-co-occurrence memory. Indeed, such a procedure is logically necessary to explain the learning of words with more than one meaning (i.e., homophones). Future research is necessary to explore the conditions under which ambiguous words are successfully learned, taking into account the mutually exclusive occurrence of appropriate referents (Meaning 1 vs. Meaning 2), which was not modeled experimentally here (i.e., the $100 \%$ referent was simultaneously present alongside the $50 \%$ referent on each " P " learning trial). Moreover, other distinguishing contextual features likely support the differentiation of two meanings for the same word. Finally, future work must examine how well these observations hold for naturally occurring word-learning environments in which referential ambiguity is greater and the contexts of word use are more variable. Artificial stimuli like those used here offer better experimental control and thus allow for closer examination of the learning mechanism but do not address how this mechanism responds to more typical input (Medina et al., 2011; Trueswell et al., 2013).

## References

Baayen, R., Davidson, D., \& Bates, D. (2008). Mixedeffects modeling with crossed random effects for subjects and items. Journal of Memory and Language, 59(4), 390412.

Bates, D. (2005). Fitting linear mixed models in r. R News, (pp. 27-30).
Carey, S. (1978). The child as a word learner. In J. Bresnan \& G. Miller (Eds.), Linguistic theory and psychological reality (pp. 264-293). Cambridge, MA: MIT Press.
Koehne, J. \& Crocker, M.W. (2011). The interplay of multiple mechanisms in word learning. In L. Carlson, C. Hoelscher, \& T. F. Shipley (Eds.), Proceedings of the 33rd Annual meeting of the Cognitive Science Society (pp. 1930-1936). Austin, TX: Cognitive Science Society.
Medina, T., Snedeker, J., Trueswell, J., \& Gleitman, L. (2011). How words can and cannot be learned by observation. Proceedings of the National Academy of Sciences of the United States of America, 108, 914-919.
Quine, W. (1960). Word and Object. Cambridge, MA: Cambridge University Press.
Trueswell, J., Medina, T., Hafri, A., \& Gleitman, L. (2013). Propose but verify: Fast mapping meets cross-situational word learning. Cognitive Psychology, 66, 126-156.
Vouloumanos, A. \& Werker, J. F. (2009). Infants' learning of novel words in a stochastic environment. Developmental Psychology, 45(6), 1611-7.
Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. Cognition, 107(2), 729-42.
Yu, C. \& Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. Psychological Science: A journal of the American Psychological Society, 18(5), 414-20.

# Awareness in Decision-Making: Implicit Influences or Measurement Biases? 

Emmanouil Konstantinidis (emmanouil.konstantinidis.09@ucl.ac.uk) \& David R. Shanks (d.shanks@ucl.ac.uk)<br>Department of Cognitive, Perceptual and Brain Sciences<br>University College London<br>26 Bedford Way, WC1H 0AP, London,UK


#### Abstract

Can our decisions be guided by unconscious or implicit influences? According to the somatic marker hypothesis, emotionbased signals guide our decisions in uncertain environments outside awareness. However, evidence for this claim can be questioned on the grounds of inadequate assessments of conscious knowledge. Post-decision wagering, in which participants make wagers on the correctness of their decisions, has been recently proposed as an objective and sensitive measure of conscious content. In the experiments reported here, we employed variations of a classic decision-making paradigm, the Iowa Gambling Task, in combination with wagering in order to investigate the role played by unconscious influences. We also examined biases that affect wagering strategies such as the definition of the optimal strategy and loss aversion. Our results demonstrate the inadequacy of post-decision wagering as a direct measure of conscious knowledge and also question the claim that implicit processes influence decision-making.


Keywords: Iowa Gambling Task; decision-making; postdecision wagering; awareness; implicit learning; loss aversion.

## Introduction

One of the most influential paradigms in the study of decision-making under uncertainty is the Iowa Gambling Task (IGT), a four-armed bandit task which requires participants to sample from decks of cards (labeled A-D) with different monetary payoffs (Bechara et al., 1994). Each card selection yields either a win or a combination of a win and loss. Overall, two of the decks are bad, leading to higher immediate wins but long-term losses (a net loss of $-\$ 25$ per card) whereas the remaining two are good, leading to lower immediate rewards but long-term wins (a net win of $+\$ 25$ per card). Another feature of the task is that the decks differ in the probability of loss with two having infrequent $(p=.1)$ and two decks having frequent $(p=.5)$ losses.

The IGT was initially designed to test empirically the somatic marker hypothesis (SMH), according to which bodily affective signals or markers can assist decision-making processes by marking response outcomes with an emotional signal, thus facilitating the selection of the most rewarding options in situations of uncertainty (see Damasio, 1996). A major assumption regarding somatic markers is that they can inform advantageous decision-making even when participants are not aware of the quality of their decisions.

## Measures of Conscious Knowledge in the IGT

In a highly influential study, Bechara et al. (1997) assessed participants' knowledge by pausing the task every 10 trials and asking them to report whatever they knew and felt about the task. In addition, participants' electrodermal skin conductance responses (SCRs) were measured as an index of emotional arousal. The crucial finding was that participants se-
lected cards from the good decks before they had conscious knowledge that those decks were the best. Importantly, the SCRs were higher prior to selections from the bad decks, suggesting that non-conscious biases (or markers) helped participants to avoid disadvantageous selections.

The assertion that our decisions are guided by implicit influences has been extensively criticized on the basis of weaknesses in the method that Bechara et al. (1997) used to probe awareness. Broad questions often fail to capture the rich spectrum of conscious knowledge, and moreover participants may not fully report their knowledge due to a conservative response criterion (see Newell \& Shanks, in press). Following these methodological considerations, Maia and McClelland (2004) developed a more sensitive test of awareness in the form of a quantitative questionnaire. They asked their participants to rate each deck on a numerical scale, to report the expected wins and losses from each deck, and to indicate which deck they would pick if they could only choose from one deck for the rest of the task. Using this assessment, Maia and McClelland found that optimal performance on the IGT was accompanied by accurate reports of the decks' payoffs, indicating that there is little evidence that implicit or emotional biases are essential for successful learning of the task structure.

## Post-Decision Wagering

Despite the apparent evidence in favor of conscious processing in the IGT, Bechara et al. (2005) suggested that the method of probing awareness used by Maia and McClelland (2004) was intrusive and so may render participants more rapidly aware of the task structure and thus undermine the role of somatic markers. In order to avoid the methodological inconsistencies created by verbal reports and numerical confidence ratings, Persaud, McLeod, and Cowey (2007) developed a novel non-verbal method of assessing awareness in the IGT, in which participants are required to make wagers after their card selections (post-decision wagering; PDW). The rationale is that if a participant maximizes her winnings through advantageous wagering (high wager after a good deck choice or low after a bad deck), then this is taken to indicate conscious knowledge.

Persaud et al. (2007) examined the influence of different types of questioning in combination with wagering in three different groups. In their version of the IGT, participants were instructed to wager $£ 10$ or $£ 20$ after each deck selection. Selections from the bad decks yielded a win of two times the wager whereas selections from the good decks returned the amount of the wager. The net outcome was ei-
ther a win of $£ 75$ (good decks) or a loss of the same amount (bad decks) per 10 cards if participants randomly allocated their wagers. Persaud et al. measured on which trial good deck selection and advantageous wagering began and conjectured that if a significant difference between these measures emerged, with performance revealing a preference for good decks before advantageous wagering emerged, this would indicate an unconscious influence on decision-making. In fact, the groups who were asked no questions or the questions of Bechara et al. (1997) demonstrated a lag between good deck selection and advantageous wagering, indicating a dissociation between performance and awareness. On the other hand, when awareness was probed by the questionnaire of Maia and McClelland (2004), performance and advantageous wagering developed simultaneously. These findings led Persaud et al. to conclude that there are implicit influences on decisionmaking which are masked if the measure of awareness makes participants aware of the nature of the task.

## Overview of the Experiments

Although post-decision wagering holds much promise as a method of measuring awareness, it has received many criticisms (e.g., Clifford, Arabzadeh, \& Harris, 2008; Fleming \& Dolan, 2010; Schurger \& Sher, 2008). In the present study, we will focus on two problematic and rather contradictory aspects of the method: the definition of the optimal strategy and loss aversion. Paradoxically, the optimal strategy is always to wager high as this strategy will give the same outcome if deck discrimination is at chance but will increase winnings if it is greater than chance. In this sense wagering high can be said to be a weakly dominant strategy with Persaud et al.'s (2007) payoff matrix, as it is either no worse than wagering low, or better. Consequently, a rational participant would always wager high, regardless of her knowledge about the task.

A second important issue is the influence of loss aversion in wagering strategies. According to prospect theory (Kahneman \& Tversky, 1979), people have an asymmetric conceptualisation of wins and losses; for example, the prospect of losing $£ 5$ is considered to be of greater psychological magnitude than that of winning the same amount. Subjective measures of awareness require participants to set a criterion about, for example, whether to wager high or low. Hence, any criterion may be modulated by cognitive biases such as loss aversion.

The present study examines the validity and sensitivity of post-decision wagering by looking closely at the two aforementioned problematic aspects and by measuring participants' awareness in combination with Maia and McClelland's (2004) quantitative questions.

## Experiment 1: The Weakly Dominant Strategy

In order to overcome the problem that wagering high is the rational strategy irrespective of the acquired knowledge, Clifford et al. (2008) proposed a solution by modifying the original payoff matrix used by Persaud et al. (2007) (see Table 1). In this modified matrix participants are discouraged

Table 1: Payoff matrices for the different combinations of deck selection and wager.

| Persaud et al. |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Clifford et al. |  |  |  |  |  |  |
| Schurger \& Sher |  |  |  |  |  |  |
| Wager | Good | Bad | Good | Bad | Good | Bad |
| Low | +1 | -1 | +2 | -1 | +1 | -2 |
| High | +2 | -2 | +5 | -5 | +10 | -10 |

from wagering high all the time. This can be shown by computing the expected payoff from wagering low which is $+1 / 2$ $[=(+2-1) / 2]$ when deck selection is random compared to $0[=(+5-5) / 2]$ from wagering high. However, when deck discrimination is better than chance, it is more rewarding to wager high due to a higher payoff with a good-high combination $(+5)$ than a good-low one $(+2)$. Based on this matrix a rational participant (i.e., a participant who seeks to maximize gains) would start to wager high only when her discrimination or level of awareness is $4 / 7$ or .57 . The latter can be computed from the differential loss of wagering on a bad decision $(5-1=4)$ divided by the sum of the differential loss and the differential gain of wagering on a good decision $(5-2=3)$.

We also addressed the question of how sensitive wagering is compared to another measure of awareness by using the quantitative questions of Maia and McClelland (2004).

## Method

Participants Twenty-one volunteers participated (13 females, age $M=23.45, S D=3.56$ ), all of whom were recruited via the departmental subject pool. They were paid $£ 2$ for their participation and an additional amount between $£ 0$ and $£ 3$, dependent on their performance in the task.
Task A probabilistic variation of the original IGT was employed. There were four decks of cards with identical physical appearance, labelled A-D. The payoff structure of each deck was different to the original IGT; the matrix of Clifford et al. (2008) was used as the basis of the payoffs received by participants on each trial, in such a way that the amount won or lost was dependent on card selection and wagering. Based on the contingencies in Table 1, on the majority of trials a payoff of +2 was associated with a good deck selection and a low wager. Similarly, the combination of a bad deck and high wager produced a payoff of -5 . On a minority of trials the signs of the payoffs were reversed (e.g., the payoff was -2 for good deck selection/low wager). For decks A and C, the majority outcome occurred on a randomly-chosen $75 \%$ of trials while for decks B and D, the majority outcome occurred on $60 \%$ of trials, resulting in different overall expected payoffs for each deck. In contrast to the original IGT, the outcome on each trial was either a net win or a loss and participants could win or lose points, which translated into money at the end of the experiment.

The task comprised 100 card selections. After each card selection, participants could place a wager, either High or Low.

They were told that if they were confident that their choice would give them some net winnings, then they should wager high, otherwise, they should make a low wager. Based on the combination of deck selection and wagering, participants were presented with a single amount, either a win or a loss. Along with wagering, participants' conscious knowledge was assessed using a modified version of Maia and McClelland's (2004) questionnaire, which was administered every 20 trials.

## Results

Choice and Wagering In order to examine whether decision-making strategies in the IGT are dissociable from awareness we computed the average proportion of good deck selections (decks C and D) and advantageous wagering (either a high wager after a good deck or a low wager after a bad deck) over successive blocks of 10 trials. If learning of the good decks emerged earlier than awareness (as indexed by wagering) then that would indicate an unconscious influence on decision-making strategies in the IGT. However, performance exceeded the chance level on block 1 for both measures (Choice: $t(20)=2.83, p=.01,95 \%$ CI [0.52, 0.65], Wagering: $t(20)=3.80, p<.001,95 \%$ CI [0.57, 0.73]) (see Figure 1A). It is important to note that even though they use the same scale, the two measures cannot be compared directly because advantageous wagering is dependent on the first-order decision (e.g., deck selection) and this creates the possibility of functional differences between the measurement scales. For example, if a participant always chooses a good deck (with the proportion of good deck selections therefore being 1.0), but decides to make both high and low wagers because she is more confident on some trials than others, then advantageous wagering cannot attain a value of 1.0.

A repeated-measures ANOVA on good deck selections with polynomial contrasts revealed significant linear, $F(1,20)=72.38, M S E=3.12, p<.001, \eta_{g}^{2}=0.41$, and quadratic trends, $F(1,20)=7.34, M S E=1.83, p=$ $.013, \eta_{g}^{2}=0.04$. Wagering performance closely followed the optimal decision-making strategy as demonstrated by a linear effect of block, $F(1,20)=19.57, M S E=4.26, p<$ $.001, \eta_{g}^{2}=0.19$. These results indicate that participants favored the good decks and became gradually capable of maximizing their winnings by placing appropriate wagers. Since choice and wagering displayed similar patterns there is no evidence of a dissociation between learning and awareness of the optimal strategy, a pattern which is at odds with the main claim of Persaud et al. (2007) about learning without awareness in the IGT.

An important feature of our task variant is that each deck yielded different overall expected outcomes. If participants had explicit insight about their choices, they could discriminate not only between good and bad decks but also within each pair of decks (A vs B and C vs D). More selections from the most rewarding of the good decks (deck C) would suggest that participants possess substantial knowledge about the payoff structure of the task. Figure 1B shows that participants indeed selected more cards from the deck with the highest win


Figure 1: (A) Proportion of good deck selections and advantageous wagering across block of trials (lines) and proportion of participants who favored the good decks in the quantitative questions (markers). Error bars are $\pm 1$ SEM. (B) Overall proportion of deck selections. The win probability associated with each deck is depicted on the top of each bar.
probability (.75) and there was a significant preference over the second best deck, $t(20)=3.98, p<.001, d=0.87$.
Questionnaire Two of the questionnaire measures, in which participants provided a rating for each deck and indicated which deck they would select cards from for the remainder of the task (Ratings and Deck Selected in Figure 1), reflect knowledge about the general quality of the decks. The remaining two measures, in which participants reported the average net outcome from each deck (Reported Net), refer to the actual payoffs. Participants also reported the average win and loss and the frequency of losses, based on which a net amount for each deck is obtained (Calculated Net).

Figure 1 (markers) shows the proportion of participants whose answers favored the good decks in each of the questionnaire measures. Participants whose verbal responses did not discriminate between good and bad decks (i.e., they gave the same ratings or same reported net for all decks) do not count towards this proportion. Inspection of the figure shows that participants exhibited substantial knowledge about the quality of each deck even in the first assessment period (trial 20). Not only did they rate the good decks higher than the bad decks, but also they had a firm basis for such an attribution as revealed by their reported and calculated net payoffs.

Another way of examining the two measures is to look at participants' deck selections and wagering in the trials following the administration of the questionnaire (trials 21, 41, 61,81 ) in order to examine whether participants who behaved advantageously had knowledge of the advantageous strategy. Specifically, we are interested in the verbal reports and wagers of those participants who behaved advantageously (i.e., selected good decks) in these trials. Figure 2 shows that the majority of participants who made good deck selections also demonstrated knowledge of the advantageous strategy in all the questionnaire items. However, wagering underestimates the acquired knowledge in all trials following the question-
naire compared to the verbal reports (the figure also shows wagers on trial 100, immediately prior to the final set of questions). Thus, it is evident that the detailed and structured questions revealed higher levels of awareness compared to wagering.


Figure 2: Percentage of participants who showed knowledge of the advantageous strategy in the questionnaire items versus in their wagers. Wagering indicates the percentage of participants who made an advantageous wager (high on a good deck choice) on the trial immediately following the administration of the questionnaire.

## Discussion

The key point of Experiment 1 is that wagering did not lag behind the selection of good decks, with both measures becoming reliably better than chance very early in the task. Under the conditions tested here, awareness as measured by wagering tracked deck selections quite closely. In addition, the results of the quantitative questions revealed that there was actually conscious knowledge that was left undetected by wagering as participants' wagers were less sensitive than their verbal reports. This places a question mark over the validity of post-decision wagering as a valid and sensitive method of assessing awareness.

An interesting finding is the early onset of learning and awareness (Block 1) which can be primarily explained by the probabilistic allocation of wins and losses on each trial. Fellows and Farah (2005) found that in their shuffled IGT version (the order of the decks was changed so that losses from the bad decks occurred immediately at the start of the task) normal controls selected more cards from the good decks even in the first 20 trials and they kept on choosing the good decks throughout the task. This setup of the payoff schedule removes the reversal learning component of the IGT which can delay the learning of the optimal decisions.

## Experiment 2: Dealing with Loss Aversion

Depending on the payoff matrix, participants may employ different response criteria to place high or low wagers which can make detection of acquired knowledge very difficult. This leads to the possibility that the expression of awareness via wagering may be constrained by factors other than knowledge itself. For instance, several studies have shown that loss
aversion affects awareness as indexed by wagering (e.g., Dienes \& Seth, 2010; Fleming \& Dolan, 2010).

Schurger and Sher (2008) proposed that the design of the payoff matrix should take into account participants' tendency to evaluate losses worse than equivalent wins. Unlike Clifford et al.'s matrix, which encourages low wagers when certainty is low, "subjects seem to need precisely quite the opposite sort of encouragement" (Schurger \& Sher, 2008, p. 209). Table 1 shows the matrix devised by Schurger and Sher as a means to counter loss aversion. Specifically when discrimination between good and bad decks is at chance it is more advantageous to wager high due to a negative expected payoff from wagering low $[(+1-2) / 2=-1 / 2]$ compared to a neutral payoff from wagering high $[(+10-10) / 2=0]$. Following this, it can be shown that a rational participant would switch to high wagers even when her discrimination is below chance ( $50 \%$ ), at $8 / 17$ or $47 \%$. Specifically, the differential loss of wagering on a bad decision is $8(=10-2)$ divided by the sum of the differential loss and the differential gain of wagering on a good decision $(10-1=9)$.

Despite the fact that this matrix discourages participants from wagering low under uncertainty, its weights regarding high wagers are two times bigger compared to the matrix of Clifford et al. On the one hand, the larger loss following a low wager after an incorrect decision discourages participants from wagering low, thus overcoming the problem of loss aversion. On the other hand, the bigger weights for high wagers could discourage participants from wagering high, even when knowledge about the quality of the decks exists. Thus, employment of this matrix might reveal that the remedy proposed to counter loss aversion cannot be achieved due to the increased weights associated with high wagering.

## Method

Participants We tested a total of 30 participants ( $24 \mathrm{fe}-$ males, age $M=25.08, S D=4.02$ ), recruited from UCL's psychology subject pool. Participants were rewarded between $£ 1$ and $£ 5$, proportional to their performance on the task.

Task The payoffs of each deck were different to the original IGT, but their overall expected payoffs reflected the ratio of losses to wins of the original task. There were four decks of cards each having 100 associated wins and losses, one for each trial. A randomly drawn (win or loss) value was then computed for each trial, which constituted the payoff on that deck for that trial. Decks A and B were bad decks, with an overall net outcome of -500 points (a net loss of -5 per card). These decks had high rewards (from $15-25$ points), but large losses (from $25-75$ ). Decks C and D were good decks, with an overall net outcome of +500 points (a net win of +5 per card). They had lower rewards (from $5-15$ ), but their losses were smaller too. Decks A, B, and C had a loss on $50 \%$ of trials, whereas Deck D had a loss on $10 \%$ of trials. The characteristics of each deck matched the original IGT, including the probabilities and relative magnitudes of losses, except for deck B. The losses on deck B were dis-
tributed over 50 trials (as against originally 10 trials only). We did this to avoid a major loss if participants were unlucky enough to encounter the deck B loss with a high wager. The post-decision wagers comprised multipliers, with the payoff schedule as proposed by Schurger and Sher (2008). Accordingly, a given IGT trial payoff was multiplied by a factor of 2 when wagering low on decks A and B , and by 1 when wagering low on decks C and D . When wagering high, all deck payoffs were multiplied by a factor of 10 .

## Results

Performance exceeded chance on block 1 for both measures (Choice: $t(29)=2.39, p=.023,95 \%$ CI [0.51, 0.64], Wagering: $t(29)=2.52, p=.018,95 \%$ CI $[0.51,0.61]$ ) (see Figure 3 ). This result indicates that participants' optimal decisionmaking and learning occurred very early in the task, a pattern that is not observed in previous studies which have utilised a payoff schedule similar or identical to the original IGT.


Figure 3: (A) Proportion of good deck selections and advantageous wagering across block of trials. Error bars are $\pm 1$ SEM. (B) Overall proportion of deck selections.

Two separate repeated-measures ANOVAs were conducted to investigate the progression of good deck selections and advantageous wagering across blocks. Polynomial contrasts revealed a significant linear effect of block on the proportion of good deck selections, $F(1,29)=143.40, M S E=2.34, p<$ $.001, \eta_{g}^{2}=0.37$. However, the same trend was not observed on the proportion of advantageous wagers as the linear effect was not significant, $F(1,29)=2.06, M S E=8.44, p=.16$. Even though wagering was above chance even from block 1, it never exceeded 0.7. In a situation where high wagers have much greater stakes than low wagers, participants may wager conservatively throughout the task, independent of learning and awareness, due to an aversion to large losses. Additionally, advantageous wagering was above chance in all blocks of trials.

Figure 3B shows the proportion of deck selections throughout the task. Deck B was not selected as often as in previous studies using the IGT, a fact which reflects the change of the loss probability. When the occurrence of losses is more frequent (.5), the prominent deck B phenomenon is not ob-
served. On the other hand, deck D (loss probability .1) was selected more often than deck C (loss probability .5) even though both decks have the same expected value.

## Discussion

This experiment confirms the hypothesis that loss aversion modulates wagering strategies by making participants more sensitive to losses. While the payoff matrix we used encourages high wagering under uncertainty, the probabilistic IGT variant we employed was found to be easier to learn than the classic IGT and thus participants were able to grasp the payoff schedule in the first 10 trials, indicating that they did not go through a phase of exploration or uncertainty. Having learned the probabilistic structure of wins and losses early in the task, it might be expected that wagering would simultaneously follow the optimal choices. This was the case in the first 2 blocks where participants had learned about the good strategy and made high wagers. Yet a random loss which may occur from the selection of a good deck with a high wager $(\times$ 10) would result in a large amount of points being deducted from the total sum. Hence, a "lose-less" strategy seems to overtake the tendency to maximize winnings, and in this particular case leads to suboptimal wagering. In other words, loss aversion constrains participants from wagering high on their good deck selections. This is indicative of a bias regulating wagering strategies and not lack of awareness. While good deck selections gradually increased to reach the maximum point by the end of the task, it would be unreasonable to argue that this was the effect of an unconscious mechanism.

The present experiment also highlights the inadequacy of post-decision wagering to measure awareness objectively and directly. Small changes in the payoff matrix can dramatically change the expression of awareness as cognitive or response biases overtly influence the reasoning behind participants' wagering strategies.

## General Discussion

The purpose of the present article was twofold: first, to evaluate how sensitive and direct post-decision wagering is as a measure of awareness, and secondly to investigate whether there are implicit influences on decision-making under uncertainty or whether past suggestive results have been the byproduct of use of insensitive measures.

We examined two main response biases, dominance and loss aversion, which arise from the design of the payoff matrix. In both cases, there was a direct effect of the design of the payoff matrix on the wagering strategies that participants employed. In Experiment 1, no delay was observed in the onset of awareness relative to deck selection, as advantageous wagering closely followed learning of the good decks. The early onset of awareness can be accounted for by either of two factors: first, the values in the payoff matrix which may have encouraged low wagers, and secondly, the quantitative questions which might have influenced the development of awareness of the deck values. However, participants' wagering performance was better than chance even
before the first administration of the questionnaire, indicating that the explicit nature of the questions did not make participants more aware of the decks' payoffs.

In Experiment 2, we tried to control for the effects of loss aversion on wagering strategies, mindful of the possibility that the high values in the wagering matrix could make participants reluctant to place high wagers. The matrix proposed by Schurger and Sher (2008) attempts to eliminate loss aversion in situations of uncertainty, that is when knowledge about a response option is weak. Participants were able to discriminate between the decks after a few trials. Although wagering performance was better than chance from the beginning of the task, it did not lead participants to maximize their earnings. One explanation lies in the design of the task: with random losses occurring even on selections from the good decks and wagers treated as multipliers of the actual payoffs, the prospect of losing a significant amount could inhibit the placement of high wagers.

Even though the decision-making paradigms employed in our experiments are not identical to the original IGT, they maintain its key characteristics such as sequential choices, rewards, punishments, and uncertainty, while removing some of its problematic features such as participants' tendency to focus on decks with infrequent losses, the lack of explorationexploitation, and the dual presentation of the outcomes. These latter issues make the interpretation of IGT data quite difficult (see Steingroever et al., 2013). Nevertheless, future work could seek to determine whether advantageous deck selection and wagering develop in parallel when our probabilistic version includes the reversal learning element of the original IGT task.

The claim that unconscious biases are essential for successful performance in the IGT has not been confirmed in either of the experiments reported here. In fact, participants' advantageous selections developed in parallel with their conscious knowledge, as revealed by verbal reports and wagering. Learning of the task contingencies through emotional markers cannot be ruled out, but does not seem to precede conscious evaluation of the experienced outcomes. In other words, the activation of an unconscious emotional system which provides critical information for decision-making processes is not essential or sufficient to explain learning in the IGT. Dunn, Dalgleish, and Lawrence (2006) suggested that there is little evidence to support the view that deck contingencies are consciously impenetrable, and that what needs to be tested is whether participants have an explicit understanding of the payoff schedule or whether they can simply discriminate the quality of the decks. Our questionnaire results demonstrate that participants not only were able to show a general preference for the good decks but that in addition they could justify their preferences by accurately reporting the average wins and losses associated with each deck. This knowledge develops very early in the task, leaving little room for unconscious influences.

## Acknowledgments

We thank Ojas Mehta for his help in collecting the data of Experiment 2 and for his useful comments. This work was supported by a UCL Graduate School fellowship (EK).

## References

Bechara, A., Damasio, A. R., Damasio, H., \& Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. Cognition, 50, 7-15.
Bechara, A., Damasio, H., Tranel, D., \& Damasio, A. R. (1997). Deciding advantageously before knowing the advantageous strategy. Science, 275, 1293-1295.
Bechara, A., Damasio, H., Tranel, D., \& Damasio, A. R. (2005). The Iowa gambling task and the somatic marker hypothesis: Some questions and answers. Trends in Cognitive Sciences, 9, 159-162.
Clifford, C. W. G., Arabzadeh, E., \& Harris, J. A. (2008). Getting technical about awareness. Trends in Cognitive Sciences, 12, 54-58.
Damasio, A. R. (1996). The somatic marker hypothesis and the possible functions of the prefrontal cortex. Philosophical Transactions of the Royal Society B: Biological Sciences, 351, 1413-1420.
Dienes, Z., \& Seth, A. K. (2010). Gambling on the unconscious: A comparison of wagering and confidence ratings as measures of awareness in an artificial grammar task. Consciousness and Cognition, 19, 674-681.
Dunn, B. D., Dalgleish, T., \& Lawrence, A. D. (2006). The somatic marker hypothesis: A critical evaluation. Neuroscience \& Biobehavioral Reviews, 30, 239-271.
Fellows, L. K., \& Farah, M. J. (2005). Different underlying impairments in decision-making following ventromedial and dorsolateral frontal lobe damage in humans. Cerebral Cortex, 15, 58-63.
Fleming, S. M., \& Dolan, R. J. (2010). Effects of loss aversion on post-decision wagering: Implications for measures of awareness. Consciousness and Cognition, 19, 352-363.
Kahneman, D., \& Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, 47, 263291.

Maia, T. V., \& McClelland, J. L. (2004). A reexamination of the evidence for the somatic marker hypothesis: What participants really know in the Iowa gambling task. Proceedings of the National Academy of Sciences, 101, 1607516080.

Newell, B. R., \& Shanks, D. R. (in press). Unconscious influences on decision making: A critical review. Behavioral and Brain Sciences.
Persaud, N., McLeod, P., \& Cowey, A. (2007). Post-decision wagering objectively measures awareness. Nature Neuroscience, 10, 257-261.
Schurger, A., \& Sher, S. (2008). Awareness, loss aversion, and post-decision wagering. Trends in Cognitive Sciences, 12, 209-210.
Steingroever, H., Wetzels, R., Horstmann, A., Neumann, J., \& Wagenmakers, E.-J. (2013). Performance of healthy participants on the Iowa gambling task. Psychological Assessment, 25, 180-193.

# The impact of bottom-up and top-down saliency cues on reference production 

Ruud Koolen (R.M.F.Koolen@uvt.nl)<br>Emiel Krahmer (E.J.Krahmer@uvt.nl)<br>Marc Swerts (M.G.J.Swerts@uvt.nl)<br>Tilburg Center for Cognition and Communication (TiCC), School of Humanities, Tilburg University<br>PO Box 90153, 5000 LE, Tilburg, The Netherlands


#### Abstract

This study investigates to what extent visual saliency cues in realistic visual scenes cause speakers to include a redundant color attribute in their definite descriptions of objects, and in particular how such cues guide speakers in determining which objects in the scene are relevant distractors, and which not. First, regarding bottom-up cues, the results revealed that the presence of clutter positively affected the redundant use of color, but that the distance between a target and a distractor did not have an effect in this respect. Second, an effect of topdown saliency (i.e., whether a target's type was mentioned in the instructions) was only partially borne out by the data. We argue that these findings are problematic for algorithms that aim to generate psychologically realistic object descriptions, since these generally select properties that help to distinguish a target from all distractors that are present in a scene.


Keywords: Definite reference; Overspecification; Bottom-up and top-down saliency; Computational models.

## Introduction

When producing definite object descriptions (such as "the green chair"), speakers must decide on the information that they include in order to make a target object identifiable for the addressee. Many referential tasks require distinguishing a target from one or more distractor objects. The properties that speakers include to make the target identifiable seem to be largely determined by the properties of the distractor objects. For example, consider the two scenes depicted in Fig. 1.


Figure 1: Two simple visual scenes.
Although the target is the same in both scenes (a large brown chair, as indicated by the arrows), the distractor will probably cause a speaker to describe it in different ways. In the left scene, where the distractor is a small brown chair, there is a high chance that a speaker produces a description like "the large chair". However, in the right scene, where the distractor is a large green chair, a description like "the brown chair" is more likely to be uttered.

While the distractor object(s) seem to play a large role in the production of target descriptions in simple visual scenes such as the ones depicted in Fig. 1 (involving comparisons of structurally different minimal pairs of objects), it is the question whether a similar process is at play when speakers refer to target objects in realistic, more complex scenes. For example, imagine a speaker asking her listener to hand her a plate that is lying on a table full of objects. Do speakers then regard all these objects as relevant distractors? Or may there be reasons why certain objects are excluded from the set of distractors? And, most importantly, how does this influence the production of reference? These are the questions that we address in the current paper.

## Background

In recent years, the production of referring expressions has received considerable attention, both from a computational and from a psycholinguistic perspective (van Deemter et al., 2012). In computational linguistics, for instance, researchers have developed several Referring Expression Generation (REG) algorithms, most notably the Incremental Algorithm (IA) introduced by Dale and Reiter (1995). The IA is a computational model that focuses on content planning: the algorithm iteratively selects attributes (e.g., type, color, size) in order to distinguish a target from one or more distractor objects in the distractor set. In order to do this, the IA uses a preference order that contains all attributes that occur in the given domain, where it considers frequently used attributes for inclusion before less frequent attributes. In this paper, we assume that type is at the head of the preference order (before color), since it is needed to generate a proper noun phrase (Levelt, 1989).
So how does the IA define the distractor set? Dale and Reiter (1995) write: "We define the context set to be the set of entities that the hearer is currently assumed to be attending to" (p. 236), where the distractor set consists of all elements that are present in a visual scene except the target. Thus, Dale and Reiter do not explain explicitly how the set of distractors should be determined for a scene, and whether it should be restricted in a certain communicative situation or not. This means that the IA generally selects the content of its object descriptions by searching for properties that help to distinguish the target from all distractors that are present in a visual scene. This might be problematic from a psychological perspective: for example, regarding discourse
structure, Krahmer and Theune (2002) argue that the set of distractor objects may change during a discourse (e.g., when speakers repeatedly refer to the same object), while Kelleher and Kruijff (2006) and Van der Sluis (2005) add that visual salience may play a role in this as well. However, to the best of our knowledge, earlier work that systematically tests how human speakers are driven by visual salience to dynamically restrict the distractor set for a given scene is lacking.
In the current paper, we test three visual saliency cues that may guide speakers in determining the set of distractors. We base our manipulations on Itti and Koch's (2000) model of visual attention, stating that an object pops out of a scene if it is sufficiently salient (Itti and Koch call this bottom-up, perceptual saliency), or if the viewer's attention is guided to it (also referred to as top-down, conceptual saliency).

With regard to bottom-up scene processing, we expect two kinds of cues to guide speakers in determining the distractor set. First, we expect the presence of visual clutter to play a role. We define visual clutter here as a collection of objects that are thematically related to the target object, and assume that the amount of visual clutter is positively correlated to the amount of objects in a scene (Bravo \& Farid, 2008). In previous research, clutter has been shown to affect speakers' response times when describing naturalistic scenes (Coco \& Keller, 2009), with slower reactions for cluttered scenes. In line with this, we expect that since a cluttered scene contains more objects (and may thus be more difficult to process), it is unlikely that speakers 'calculate' for every distractor how it can be distinguished from the target object.

Secondly, again regarding bottom-up scene processing, we expect distractor distance (that is, the distance between the target and a distractor) to guide speakers in determining the distractor set. For reference in dialogue, Beun and Cremers (1998) suggest that a speaker's focus of attention limits the number of relevant distractors: in their experiment, they find that speakers generally consider only visually close objects when referring to targets. In an eye-tracking study, BrownSchmidt and Tanenhaus (2008) have similarly shown that distractors that are visually close to the last mentioned target are most likely to be in the speaker's focus of attention. In this paper, we study if the same goes for reference where no preceding discourse is involved.
Thirdly, related to top-down scene processing, we expect the specificity of the referential task to affect speakers when determining the set of distractors, where we hypothesize that a general task (such as "describe this object") will leave the speaker with a bigger, less restricted set of distractors than a more specific task (such as "describe this plate", where the target's type is mentioned). In the latter case, speakers might leave objects other than plates unattended, while any object that is present in the scene might be regarded as a relevant distractor in the former case.

## The current study

Our experiment was a reference production task, in which participants were presented with realistic scenes on a screen. The scenes contained one target and several distractors, and
the participants were asked to describe the target in such a way that an addressee could uniquely identify it. Crucially, the trials were set up in such a way that color was never needed to do this, enabling us to take the proportional use of redundant color attributes (i.e., color attributes that were not necessary for identification) as our dependent variable. In doing this, we follow Koolen et al. (to appear), who used the redundant use of color to study how speakers differ in their perception of low-variation and high-variation scenes. Our stimuli were designed in such a way that the Incremental Algorithm would not select color to distinguish the target: it would always select type and size, irrespective of any experimental condition.

What do we predict regarding the redundant use of color? Firstly, we expect to find an effect of clutter (with speakers using more color when clutter objects are present), because a cluttered scene contains more distractors and hence might be more difficult to process. Secondly, we expect distractor distance to affect speakers' redundant color use: redundant color attributes might get used more often when a potential distractor is placed close to the target object as compared to when it is distant. Thirdly, we expect to find effects of the specificity of the referential task: we hypothesize that when speakers are instructed in a general way (e.g., "Describe this object"), they will be more likely to redundantly use color as compared to when the instruction includes the target's type (e.g., "Describe this plate").

## Experiment

## Method

Participants. 43 undergraduate students from Tilburg University ( 30 female, 13 male) took part in the experiment. All (mean age 21 years and 1 month, range 18-34 years) were native speakers of Dutch (the language of the study) and participated for course credits.

Materials. The stimulus materials consisted of 80 trials, all of which were photo-realistic pictures of objects on either a kitchen table or an office desk. In the 40 critical trials, there were always at least three objects present on the table: one target object and two distractor objects. Crucially, one of the distractors had the same type and color as the target object (meaning that it could only be ruled out by means of its size), and was always positioned next to the target object (either left or right). The second distractor always had a different type and color as compared to the target, and its positioning varied across conditions. Besides that, two other principal factors were manipulated: one related to the presence of clutter in the scene, and one related to the specificity of the task that was given to the speakers.
The first manipulation was related to whether or not there was clutter present in the visual scene. We define clutter as a collection of all kinds of objects that are thematically related to the target and its two main distractors. The clutter objects were not systematically varied, and were unique in the sense that they all had different types. The color of these


Fig. 2: Examples of critical trials in Experiment 1. The left scenes have a close distractor, whereas this distractor is distant in the right scenes. The upper scenes do not contain clutter, whereas the lower scenes do. Note that both the small and the large plate could be the target in the experiment.
clutter objects was kept as neutral as possible; it was at least made sure that the clutter objects did not have the same color as compared to the target and its two distractors. In the cluttered pictures in Fig. 2, five objects are added that one would expect to see on a breakfast table, where most do not have a salient color: a bag of bread, a newspaper, a piece of cheese, a cheese slicer, and a pack of chocolate sprinkles. Clutter was added in half of the critical trials, and the same clutter objects were used for the scenes with a close and a distant distractor.
The second manipulation (distractor distance) was related to the distance between the target object and the second distractor. This distance was manipulated as follows: in half of the trials, the distractor was positioned close to the target (with the two distractors placed in the same corner of the table), whereas this distance was maximized in the other half of the trials (with the target and the first distractor in one corner of the table, and the second one in the opposite corner). In Fig. 2, the left pictures had a close distractor, and the right pictures had a distant distractor. When a scene had a distant distractor, this object was always positioned in the corner opposite the target. Note that mentioning the target's type and size was sufficient to identify the target in both the close and distant conditions, implying that the use of color would inevitably result in overspecification.

The experiment had eighty trials: forty critical trials and forty fillers. Regarding the critical trials, we used ten scenes: five scenes with objects on an office desk and five scenes with objects on a kitchen table. These ten scenes were all manipulated in a 2 (distance) x 2 (clutter) design, resulting
in four within-conditions as described above: one picture with two close distractors but without clutter, one with a close and a distant second distractor without clutter, one with two close distractors and with clutter, and one with a close and a distant distractor and with clutter. Note that the target object could be positioned in all four corners of the table (and not necessarily in the left bottom corner, as is the case in Fig. 2). Since there were always two similar objects in a scene (one of which being the target object), we marked the small object as the target in half of the scenes, and the large object in the other half of the scenes.
Besides distractor distance and the presence of clutter presence (both manipulated as within participants factors), the experiment also had one between participants variable (hence called specificity of the referential task), which was related to the instruction that was given to the participants. As mentioned earlier, it was the participants' task to describe each target object in such a way that it could be distinguished from the other objects in the visual scene. All participants were presented with the same stimuli, but two kinds of instructions were used. Half of the participants had the task to "describe this object" (which means that they took part in the low specificity condition), whereas the other half of the participants (in the high specificity condition) had a more specific task. In this condition, the target's type was mentioned in the instruction. For example, in Fig. 2, these speakers were asked to "describe this plate".

Table 1: Overview of the experimental design and the number of descriptions within each cell.

|  | No clutter |  | Clutter |  |
| :--- | :---: | :---: | :---: | :---: |
|  | Close | Distant | Close | Distant |
| Low specificity | 210 | 210 | 210 | 210 |
| High specificity | 220 | 220 | 220 | 220 |

The experiment had forty fillers: twenty from the kitchen table domain and twenty from the office desk domain. These fillers were set up in the same way as the critical trials, in the sense that there were scenes containing few objects that were positioned in the same way as those in the critical trials, and scenes containing many different objects (in line with the clutter scenes that served as critical trials). Again, one of the objects was marked as the target and was described by the participants, with the crucial difference that the objects in the filler pictures did not differ in terms of their color. In this way, speakers were discouraged from using color when describing the fillers.

Procedure. The experiment was performed in a lab, and had an average running time of 10 minutes. After participants had entered the lab, they were randomly assigned to one of the two conditions: 21 participants took part in the low specificity condition, and 22 in the high specificity condition. Thereafter, they were seated opposite the listener (who was a confederate of the experimenter), and were instructed so as to describe a target in such a way that their listener could uniquely identify it. For each new trial, the participants were instructed by means of a pre-recorded task (for example: "describe this object" in the low specificity condition and "describe this plate" in the high specificity condition). Speakers could take as much time as needed to describe a target, and their descriptions were recorded with a voice recorder.

The trials were presented to participants on a computer screen. We made one block of eighty trials in a fixed random order (which was presented to one half of the participants), and a second block containing the same trials in the reverse order (which was presented to the other half of the speakers). There were two practice trials. The listener had a paper booklet in front of her, containing - for each trial - separate pictures of all the objects that occurred in that given scene. These pictures were taken from the pictures the speaker was presented with. Based on the speaker's descriptions, the listener marked the object that she thought was referred to on an answering form. In order to prevent speakers from including location information in their target descriptions (e.g., 'The plate in the left bottom corner'), the instructions emphasized that the listener was presented with the same objects ranked in a different order. The listener always acted as though she understood the descriptions, and never asked clarification questions. This was done to enable a focus on content planning of initial descriptions ('first mentions'). Once the listener had identified a target, this was communicated to the speaker, who then went on to describe the next target.

Design and statistical analysis. The experiment had a $2 \times 2$ x 2 design (see table 1) with two within participants factors: distractor distance (levels: close, distant) and clutter presence (levels: no clutter, clutter), and one between participants factor: specificity of the referential task (levels: low, high). The experiment had one dependent variable: the proportion of descriptions containing a color attribute. As described above, we made sure that speakers never needed color in order to distinguish the target object from its distractors: mentioning the target's type and size was always sufficient. Thus, when speakers mentioned color, this always resulted in an overspecified description.
Our statistical procedure consisted of Repeated Measures ANOVAs: one on the participant means $\left(F_{1}\right)$ and one on the item means $\left(F_{2}\right)$. We only report on interactions where these are significant. In order to compensate for departures from normality, we applied a standard arcsin transformation to the proportions before running the ANOVAs. For the sake of readability, we report the untransformed proportions in the results section.

## Results

In total, 1720 target descriptions were produced in this experiment. All of these contained a type attribute, and most ( $85.8 \%$ ) contained a size attribute. In the rest of the cases, other additional attributes were mentioned to distinguish the target object (such as its orientation). All descriptions were fully distinguishing. Speakers redundantly mentioned color in $39 \%$ of the descriptions.

Results for clutter. The first factor that we expected to affect speakers' redundant use of color was the presence of visual clutter in the scene. Fig. 3 displays the proportional use of color as a function of clutter presence.


Fig. 3: The proportional use of color (plus standard deviations) as a function of clutter presence.

As can be seen in Fig. 4, the presence of clutter positively affected the redundant use of the attribute color $\left(F 1_{(1,41)}=\right.$ 13.38, $\left.p=.001 ; F 2_{(1,36)}=3.91, p=.06\right)$. In other words, speakers were more likely to include color when presented with visual scenes containing clutter $(M=.43, S D=.05)$ as compared to when the scene did not contain clutter ( $M=$ $.35, S D=.06$ ).

Results for distractor distance. We also studied the effect of distractor distance on the redundant use of color: whether a distractor was placed close to or far from the target object. The results are shown in Fig. 4.


Fig. 4: The proportional use of color (plus standard deviations) as a function of distractor distance.

As can be seen in Fig. 3, distractor distance did not affect the proportional use of the redundant attribute color $\left(F 1_{(1,41)}\right.$ $\left.=.068, p=.80 ; F 2_{(1,36)}=.00, p=.99\right)$. More specifically, color was mentioned color exactly as many times when the distractor was close $(M=.39, S D=.05)$ as compared to when it was distant ( $M=.39, S D=.05$ ).

Results for specificity of the referential task. The third factor that we manipulated was related to the specificity level of the task that was given to the speakers: for half of the speakers this specificity level was low, while it was high for the other half of the speakers.


Fig. 5: The proportional use of color (plus standard deviations) as a function of the specificity of the instruction.

As reflected in Fig. 5, the specificity of the instructions to some extent affected the use of the redundant attribute color,
but this effect was only significant by items $\left(F 1_{(1,41)}=.355\right.$, $\left.p=.55 ; F 2_{(1,36)}=15.81, p<.001\right)$. More specifically, this means that speakers that took part in the low specificity condition ( $M=.42, S D=.07$ ) did use color more frequently as compared to those taking part in the high specificity condition ( $M=.36, S D=.07$ ), but that we did not find a convincing effect of instruction specificity.

## Discussion

In this paper, we have investigated how bottom-up and topdown saliency cues - as defined by Itti and Koch (2000) guide speakers in determining which objects in a scene belong to the set of relevant distractor objects. In doing this, we have studied how these cues affect speakers' production of object descriptions, and in particular, to what extent they cause speakers to mention a redundant color attribute. On average, $39 \%$ of the object descriptions in our experiment redundantly contained a color attribute, which is more than the proportions reported by, among others, Belke and Meyer (2002), Koolen et al. (to appear), and Pechmann (1989).

Regarding top-down scene processing, we hypothesized that participants in the low specificity condition (who were asked to "describe this object") would be more likely to use a redundant color attribute than participants that took part in high specificity condition (who were asked to "describe this X", e.g., "this plate"), since in the latter case (where the target's type was used in the instruction) only the distractor with the same type would remain to rule out (which could always be done by mentioning size). We indeed found a numerical difference between the conditions in the predicted direction, but this was only significant in the $F_{2}$ analysis. We plan to further study this effect in future research.
Secondly, regarding bottom-up scene processing, we have found that - at least for the visual scenes used here - the distance between the target and a distractor does not affect the redundant use of color. This might be due to an artefact of the experimental setup: given that our speakers knew that the addressee was presented with - for every given scene separate pictures of all objects that were depicted in that scene, this might have caused them to ignore the distance between the target and the distractors in the scene. In future research, we aim to improve our manipulation of distractor distance.
Thirdly, again regarding bottom-up scene processing, our results showed that speakers are more likely to mention a redundant color attribute when there is visual clutter present in the scene as compared to when this is not the case. One explanation for this might be that a scene with clutter simply contains more distractor objects than a scene without clutter. As suggested earlier, the latter might lower the chance that speakers exactly 'calculate' for each distractor how it can be distinguished from the target in the most efficient way. Our results suggest that speakers tend to process cluttered scenes in a 'faster' way: they might rely on heuristics (Tversky \& Kahnemann, 1982) when they have to uniquely describe an object. With regard to reference production, heuristics can be defined as general rules that say, for example, that color
should always be included in the case of a cluttered scene. In recent years, such heuristics have indeed been claimed to influence reference production (e.g., Dale \& Viethen, 2009), since speakers' limited processing capacity might prevent them from calculating the shortest possible description in a given referential task (van Deemter et al., 2012).
As we have explained in the introduction of this paper, for the current REG algorithms (most notably Dale and Reiter's IA introduced in 1995) it is not explained explicitly how the distractor set should be defined for a given scene (Krahmer \& Theune, 2002): such algorithms select the content of their descriptions by searching for those properties that help to distinguish the target from all distractors that are present in the scene. Given that our findings (at least partly) suggest that perceptual and conceptual cues affect speakers' object descriptions, and that there are situations in which speakers do not take all distractors into account, or do not 'calculate' the most efficient way to describe a target because too many objects are present, the question is what the implications of our findings are for REG algorithms such as the IA.
For one thing, our results show that speakers often use a color attribute when algorithms such as the IA would not do this: in our stimuli, the algorithm would select type and size instead of color (assuming that, as explained earlier, type is placed at the head of the preference order, followed by color and other, less preferred attributes such as size). So how can the IA account for this frequent color use? For the specific case of clutter (which delivered us with the most convincing results), we propose that one solution might be to make the algorithm redundantly include color more often when it has to describe a target object in a cluttered scene as compared to a scene without clutter. For example, this could be done by dynamically adapting the preference order to the amount of clutter that is present in a particular scene: when clutter is present, color could be placed at the head of the preference order (causing it to be selected if there is any color variation in the scene), whereas the preference order can remain as we assume it is now (with type before color) for visual scenes that do not contain clutter.

## Conclusion

Bottom-up cues (i.e., the presence of visual clutter, but not distractor distance), and to a lesser extent top-down saliency cues (i.e., specificity of the referential task) influence the redundant use of color in definite object descriptions. This is problematic for Referring Expression Generation algorithms that aim to automatically generate psychologically realistic object descriptions.

## Acknowledgments

The research reported in this paper is part of the NWO VICI project "Bridging the gap between psycholinguistics and computational linguistics: the case of referring expressions" (Grant 277-70-007). We thank Anouk van Helvoort for help in creating the stimuli, Kristel Bartels and Elsa Jonkers for running the experiment and annotating the data, and Jette Viethen for comments on an earlier draft of this paper.

## References

Beun, R., \& Cremers, A. (1998). Object reference in a shared domain of conversations. Pragmatics and Cognition, 6 (1/2), 121-152.
Belke, E., \& Meyer, A. (2002). Tracking the time course of multidimensional stimulus discrimination: Analysis of viewing patterns and processing times during "same" "different" decisions. European Journal of Cognitive Psychology, 14, 237-266.
Bravo, M. \& Farid, H. (2008). A scale invariant measure of clutter. Journal of Vision, 8 (1), 1-9.
Brown-Schmidt, S., \& Tanenhaus, M. (2008). Real-time investigation of referential domains in unscripted conversation: a targeted language game approach. Cognitive Science, 32 (4), 643-684.
Coco, M., \& Keller, F. (2009). The impact of visual information on reference assignment in sentence production. In Proceedings of the $31^{\text {st }}$ annual conference of the Cognitive Science society, 274-279. Amsterdam, The Netherlands.
Dale, R., \& Reiter, E. (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. Cognitive Science, 18, 233-263.
Dale, R., and Viethen, J. (2009). Referring Expression Generation through attribute-based heuristics. Proceedings of the $12^{\text {th }}$ European workshop on $N L G$ (ENLG), 58-65.
Itti, L., \& Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. Vision research, 40, 1489-1506.
Kelleher, J. \& Kruijff, G.J. (2006). Incremental generation of spatial referring expressions in situated dialogue. In Proceedings of COLING/ACL '06. Sydney, Australia.
Koolen, R., Goudbeek, M., and Krahmer, E. (to appear). The effect of scene variation on the redundant use of color in definite object descriptions. Cognitive Science.
Krahmer, E., \& Theune, M. (2002). Efficient contextsensitive generation of referring expressions. In: K. van Deemter \& R. Kibble (Eds.). Information sharing: Givenness and newness in language processing (pp. 223264). CSLI publications, Stanford.

Levelt, W. (1989). Speaking: from intention to articulation. MIT Press, Cambridge/London.
Pechmann, T. (1989). Incremental speech production and referential overspecification. Linguistics 27, 89-110.
Tversky, A. \& Kahneman, D. (1982). Judgement under uncertainty: heuristics and biases. In D. Kahneman, P. Slovic \& A. Tversky (Eds.), Judgement under uncertainty, heuristics and biases. Cambridge: Cambridge University Press.
Van Deemter, K., Gatt, A., Van Gompel, R., \& Krahmer, E. (2012). Toward a computational psycholinguistics of reference production. Topics in Cognitive Science, 4 (2), 166-183.
Van der Sluis, I. (2005). Multimodal reference. PhD thesis, Tilburg University, The Netherlands.

# A spreading-activation model of the semantic coordination of speech and gesture 

Stefan Kopp (skopp@TechFak.Uni-Bielefeld.DE)<br>Kirsten Bergmann (kirsten.bergmann@Uni-Bielefeld.DE)<br>Sebastian Kahl (skahl@TechFak.Uni-Bielefeld.DE)<br>Faculty of Technology, Center of Excellence "Cognitive Interaction Technology" (CITEC),<br>Collaborative Research Center "Alignment in Communication" (SFB 673)<br>Bielefeld University, P.O. Box 100 131, D-33501 Bielefeld, Germany


#### Abstract

In naturally occurring speech and gesture, meaning occurs organized and distributed across the modalities in different ways. The underlying cognitive processes are largely unexplored. We propose a model based on activation spreading within dynamically shaped multimodal memories, in which coordination arises from the interplay of visuo-spatial and linguistically shaped representations under given communicative and cognitive resources. An implementation of this model is presented and first simulation results are reported.


Keywords: Speech, gesture, conceptualization, semantic coordination, activation spreading

## Introduction

Gestures are an integral part of human communication and they are inseparably intertwined with speech (McNeill \& Duncan, 2000). The detailed nature of this connection, however, is still a matter of considerable debate. The data that underlie this debate have for the most part come from studies on the coordination of overt speech and gestures showing that the two modalities are coordinated in their temporal arrangement and in meaning, but with considerable variations. When occurring in temporal proximity, the two modalities express the same underlying idea, however, not necessarily identical aspects of it: Iconic gestures can be found to be redundant with the information encoded verbally (e.g., 'round cake' + gesture depicting a round shape), to supplement it (e.g., 'cake' + gesture depicting a round shape), or even to complement it (e.g., 'looks like this' + gesture depicting a round shape). These variations in meaning coordination-in combination with temporal synchrony-led to different hypotheses about how the two modalities encode aspects of meaning and what mutual influences between the two modalities could underlie this. However, a concrete picture of this and in particular of the underlying cognitive processes is still missing.

In previous work (Bergmann \& Kopp, 2009) we explored how the surface form of speech and gesture is determined and how this formulation process can be simulated in a computational model. In this paper we turn to the preceding stage, namely, conceptualization by which meaning is structured, portioned and distributed across the two modalities, yielding different kinds of semantic coordination one can see in reallife natural behavior. We thereby focus on speech along with shape-depicting (iconic) gestures. We start with reviewing the
empirical findings on semantic coordination of speech and gesture, and we discuss mechanisms and models that have been put forward to explain it. We argue that building computational models helps to elucidate the mechanisms and to bridge the gap between descriptive models and observable behavior. We propose the first model to present a detailed cognitive account of how meaning can be organized and coordinated in speech and gesture. It is based on tenets of activation spreading in multimodal memory representations and it entails a number of, now explorable, assumptions about conceptualization of speech and gesture. We describe an implementation of this model and present first results on how it can simulate and explain different cases of semantic coordination reported in the literature.

## Background

## Semantic coordination of speech and gesture

A number of studies have shown that concomittant speech and gesture are coordinated in meaning. One line of evidence coming from cross-linguistic studies suggest that packaging of content for co-speech gestures is influenced by the information packaging for the accompanying speech. For example, Kita and Özyürek (2003) showed that speakers of English who are able to combine manner and path of a movement in a single clause (e.g. 'he rolled down' or 'he swings') accompanied this by a single gesture encoding both semantic features. In contrast, Turkish and Japanese speakers encoded manner and path separately in two clauses (e.g. 'he descended as he rolled') and are more likely to use two separate gestures for these two features. Along the same line, when native speakers of Turkish (L1) speak English as their second language (L2) at different levels of proficiency, their gestures were shown to follow the information packaging strategy they adopt (Özyürek, 2002): Advanced L2 speakers typically encoded manner and path information in one clause and their gestures followed, where as speakers at lower proficiency levels typically used two-clause constructions in speech thus following the structure of Turkish, accompanied by separate gestures for manner and path. A subsequent study (Kita et al., 2007) showed that this effect also occurrs when native speakers of English are forced to produce one- or two-clause descriptions of manner and path.

Other studies have investigated the cognitive factors that influence frequency and nature of gesturing, including its coordination with speech. Bavelas, Kenwood, Johnson, and Philips (2002) found that speakers are more likely to produce non-redundant gestures when their addressees could see them, as opposed to when their gestures are not visible and hence less essential for their partners. Bergmann and Kopp (2006) report results from an analysis of natural gesturing in direction-giving, indicating that supplementary iconic gestures are more likely in cases of problems of speech production (e.g. disfluencies) or when the information conveyed is introduced into the dialogue (and thus conceptualized for the first time). In line with this, recent work has suggested that speakers indeed produce more gestures at moments of relatively high load on the conceptualization process for speaking (Kita \& Davies, 2009), in particular on the linearization and the focusing components of conceptualization (Melinger \& Kita, 2007). Hostetter and Alibali (2007) report findings suggesting that speakers who have stronger visual-spatial skills than verbal skills produce higher rates of depictive gestures than other speakers. In a later study, Hostetter and Alibali (2011) found that the speakers with high spatial skills also produced a higher proportion of non-redundant gesturespeech combinations than other speakers, whereas verbaldominant speakers tended to produce such gestures more in case of speech disfluencies. The authors hypothesize that "non-redundant gesture-speech combinations occur because mental images are more active in speakers minds at the moment of speaking than are verbal codes" [р.45]. Taken together, this suggests that non-redundant gesture-speech combinations are the result of speakers having both strong spatial knowledge and weak verbal knowledge simultaneously, and avoiding the effort of transforming the one into the other.

## Models of speech and gesture production

Different models of speech and gesture production have been proposed. One distinguishing feature is the point where cross-modal coordination can take place. The Growth Point Theory (McNeill \& Duncan, 2000) assumes that gestures arise from idea units combining imagery and categorial content. This combination is unstable and initiates dynamic cognitive events through which speech and gesture unfold. Speech and gesture, in this view, are inseparable and interact throughout the production process.

Assuming that gestures are generated "pre-linguistically", Krauss, Chen, and Gottesman (2000) hold that gesture are generated from a mental representation of a source concept comprising a set of semantic features (size, color, shape etc.) that are encoded in propositional and/or spatial format. While there is no influence of language production onto gesture in this model, the readily planned and executed gesture facilitates lexical retrieval through cross-modal priming.

De Ruiter (2000) proposed that speech-gesture coordination arises from a multimodal conceptualization process that selects the information to be expressed in each modality and assigns a perspective for the expression. A propositional rep-
resentation is transformed into a preverbal message, and an imagistic representation is transformed into a so-called sketch and sent to a gesture planner. Kita and Özyürek (2003) agree that gesture and speech are two separate systems interacting during the conceptualization stage. Based on cross-linguistic evidence, their account holds that language shapes iconic gestures such that the content of a gesture is determined by three factors: (1) a communicative intention, (2) action schemata selected on the basis of features of imagined or real space, (3) bidirectional interactions between speech and gesture production processes at the level of conceptualization, i.e. the organization of meaning. An additional link between the speech formulator and the preverbal message generator allows for feedback from grammatical or phonological encoding to the conceptualizer and thus to gesture.

Hostetter and Alibali (2008) proposed the Gestures as Simulated Action framework that emphasizes how gestures may arise from an interplay of mental imagery, embodied simulations, and language production. According to this view, language production evokes enactive mental representations which give rise to motor activation. Whether a gesture is produced or not depends on the amount of motor activation, the speaker's variable gesture threshold, and the simultaneous engagement of the motor system for speaking.

Inspite of a consistent theoretical picture starting to emerge, many questions about the detailed mechanisms remain open. A promising approach to explicate and test hypotheses are cognitive models that allow for computational simulation. However, such modeling attempts for the production of speech and gestures are almost inexistent. Only Breslow, Harrison, and Trafton (2010) proposed an integrated production model based on the cognitive architecture ACT-R (Anderson, Bothell, Byrne, Lebiere, \& Qin, 2004). This account draws on two major assumptions: (1) on Jackendoff's claim that language representations include some irreducibly spatial components; (2) on Goldberg's approach according to which language processing is based on constructions which consist of both semantic and syntactic components. The authors assume such constructions to prescribe spatial representations for what they call linguistic spatial gestures and which they assume to provide "little information not included in the accompanying language" [p.14]. In this view, constructions are selected first and then words and gestures are determined so as to realize the construction. Accordingly, semantic coordination is predetermined and does not result from a coordination process based on problems with lexicalization or high activation of particular visuo-spatial features. This model hence has difficulties, e.g., to explain gestures that clearly complement or supplement verbally encoded meaning.

## A spreading-activation model

We investigate to what extent semantic coordination of speech and gesture can be explained by cognitive principles of activation-based processing on multimodal memory. This account is embedded in a larger production model (Kopp,

Bergmann, \& Wachsmuth, 2008) that comprises three stages: conceptualization, where a message generator and an image generator work together to select and organize information to be encoded in speech and gesture, respectively; formulation, where a speech formulator and a gesture formulator determine appropriate verbal and gestural forms for this; motor control and articulation to finally execute the behaviors. Motor control, articulation, and formulation have been subject of earlier work (Bergmann \& Kopp, 2009). What is missing is a model for multimodal conceptualization that accounts for the range of semantic coordination we see in real-life speechgesture combinations.

## Basic assumptions

We posit that the semantic coordination of speech and gesture emerges from (1) the communicative goal, (2) the need to activate, retrieve and organize multimodal information to achieve this goal, and (3) the expressive as well as cognitive resources available to the speaker at the moment. To model this process, we make a number of assumptions, partly in line with previous models. First, language production requires a preverbal message to be formulated in a symbolic-propositional representation that is linguistically shaped (Slobin, 1996; Levelt, 1989) (SPR, henceforth). During conceptualization the SPR, e.g. a function-argument structure denoting a spatial property of an object, often needs to be extracted from visuo-spatial representations (VSR), e.g. the mental image of this object. We assume this process to involve the invokation and instantiation of memorized supramodal concepts (SMC, henceforth), e.g. the concept 'round' which links the corresponding visuo-spatial properties to a corresponding propositional denotation. Co-verbal iconic gestures are then shaped by (1) the imagistic content in VSR, (2) the invoked SMCs, and (3) the organization of SPR for linguistic processing. We assume that units or entries of these memory structures can be selectively activated and that activation spreads along links between them. Fig. 1 illustrates the overall relation between the three memory structures.


Figure 1: Multimodal memory structures involed in speechgesture production (activations indicated by bold lines).

## Overall production process

Fig. 2 shows an outline of the overall production architecture. Conceptualization consists of cognitive processes that operate upon the abovementioned memory structures to create a,
more or less coherent, multimodal message. These processes are constrained by principles of memory retrieval, which we assume can be modeled by principles of activation spreading (Collins \& Loftus, 1975). As in cognitive architectures like ACT-R (Anderson et al., 2004), activations float dynamically, spread across linked entities (in particular via SMCs), and decay over time. Activation of more complex SMCs are assumed to decay more slowly than activation in lower VSR or SPR.


Figure 2: Overall production architecture.

Production starts with the message generator and image generator inducing local activations of modal entries, evoked by a communicative goal. VSRs that are sufficiently activated invoke matching SMCs, leading to an instantiation of SPRs representing the corresponding visuo-spatial knowledge in linguistically shaped ways. The generators independently select modal entries and pass them on to the formulators. As in ACT-R, highly activated features or concepts are more likely to be retrieved and thus to be encoded. Note that, as activation is dynamic, feature selection depends on the time of retrieval and thus available resources. The message generator has to map activated concepts in SPR onto grammatically determined categorical structures, anticipating what the speech formulator is able to process (cf. (Levelt, 1989)). Importantly, interaction between generators and formulators in each modality can run top-down and bottom-up. For example, a proposition being encoded by the speech formulator results in reinforced activation of the concept in SPR, and thus increased activation of associated concepts in VSR.

In result, semantic coordination emerges from the local choices generators and formulators take, based on the activation dynamics in multimodally linked memory representations. Redundant speech and gesture result from focused activation of supramodally linked mental representations, whereas non-redundant speech and gesture arise when activations scatter over entries not connected via SMCs.

## Computational simulation

We have implemented the activation-based model of semantic coordination within our larger speech and gesture production architecture (Bergmann \& Kopp, 2009). Newly implemented parts are the VSR, SPR and SMC memory structures, the activation dynamics upon these structures, and the generator modules operating on them.

## Representations

To realize the VSR and part of the SMC, we employ a model of visuo-spatial imagery called Imagistic Description Trees (IDT) (Sowa \& Kopp, 2003). The IDT model was designed, based on empirical data, to cover the meaningful visuo-spatial features in shape-depicting iconic gestures. Important aspects include (1) a tree structure for shape decomposition with abstract object schemas as nodes, (2) extents in different dimensions as an approximation of shape, and (3) the possibility of dimensional information to be underspecified. The latter occurs, e.g., when the axes of an object schema cover less than the three dimensions of space or when an exact dimensional extent is left open but only a coarse relation between axes like "dominates" is given. This allows to represent the visuo-spatial properties of SMCs such as 'round', 'left-of' or 'longish'. Applying SMC to VSR is realized through graph unification and similarity matching between object schemas, yielding similarity values that assess how well a certain SMC applies to a particular visuo-spatially represented entity (cf. Fig. 1). SPR are implemented straight forward as predicate-argument sentences.

## Activation dynamics

Each memory entry in VSR, SPR and SMC has a timedependent activation value $a_{t}$. Activation dynamics results from simple update and spreading rules applied to these values in each iteration of a stepwise cognitive simulation process. At each step all of the following updates are performed:

- Activation update for memory entries: $a_{t+1}=a_{t}-d+r$, with decay $d$, random noise $r$ (order of magnitude $10^{-1}$ )
- Activation spreading within VSR: $a_{t+1}=\frac{a_{t}}{c \cdot l}$, where $c$ is the number of outgoing links (fan-out effect) and $l$ is the depth in the hierarchical IDT structure (fade-out effect)
- Activation spreading from SPR towards VSR via SMC: $a_{t+1}^{v s r}=\frac{a_{t}^{s s r}+t_{t}^{s p r}}{2}+\alpha \cdot\left(a_{t}^{s m c}-a_{t}^{v s r}\right)+r-d$, where $\alpha$ controls the rate of convergence towards the SMC activation.
- Activation spreading from VSR towards SPR via SMC:
$a_{t+1}^{s p r}=\frac{a_{t}^{v s r}+a_{t}^{s p r}}{2}+\alpha \cdot\left(a_{t}^{s m c}-a_{t}^{s p r}\right)+r-d$
The first formula models the decay and random noise of each entry's activation, the second realizes local spreading of activation within VSR, the latter two at a global level between VSR and SPR. Especially the global multimodal activation spreading is important as it ensures that linked visuo-spatial and propositional codes align and mutually stabilize. Fig. 3
(left) shows the activations of two linked entries. At point $t=200$ one entry gets temporarily activated and the activation of the linked entry follows. The second important property of this rule is that activation of the more global SMC $a_{t}^{\text {smc }}$ spreads to both linked entries, such that both are "pulled" towards this value. This can be seen in Fig. 3 (right) where the SMC's activation is increased by 2.0 at point $t=100$. Note that activation of SMCs decays more slowly than activation of VSR and SPR entries. Activations of linked entries thus stabilizes at a higher level, such that stable multimodal information packages emerge for a limited period of time. The duration of this time period depends on the decay rate of SMC activations. Finally, memory retrieval depends on the activation of the entries being retrieved. We adopt the ACT-R approach to map activation onto retrieval probability: $p=1 /\left(1+e^{\frac{-\left(a_{t}-s\right)}{r}}\right)$, where $s$ is a threshold and $r$ the noise in the activation levels.


Figure 3: Activations of two memory entry linked via an SMC: temporary activation of one entry (left); activation of the linking SMC (right).

## Generators

The message generator has to package activated SPR information in a way that the speech formulator can produce an appropriate construction. We employ an LTAG-based (Lexicalized Tree Adjoining Grammar) sentence planner for speech formulation (cf. (Bergmann \& Kopp, 2009)). To make sure that all facts necessary to generate a specific construction are available, the message generator applies networks that reflect the encoding options provided by the speech formulator's LTAG grammar (this conforms the view that the conceptualizer learns to anticipate the formulator's abilities (Levelt, 1989)). These message networks consist of type nodes for entities, properties of entities and relations between them. These are connected via weighted links reflecting the combination of particular linguistic types in a language. For instance, relation nodes are strongly linked to two (or more) entity nodes, while links between entity and property nodes are weaker.The message generator matches the activated propositions in SPR against nodes of possible message networks and determines their initial activations. Activation, again, spreads via the weighted links and finally results in an overall activation pattern of a pre-verbal message. This has been implemented for a limited part of our domain of investigation (corresponding to NPs about buildings and their properties).

The image generator retrieves visuo-spatial information from activated VSR and SMC entries in memory. It is in charge of unifying this information into an imagistic representation, from which the gesture formulator can derive a gesture form specification (based on Bayesian decision networks learned from empirical data (Bergmann \& Kopp, 2009)). For instance, information about shape is combined with information about the object's size or position. Depending on the knowledge encoded here, the gesture formulator is able to plan a shape-depicting gesture or rather a localizing deictic or placing gesture.

## Simulation results

The implemented model offers-and simulates-detailed explanations of how semantic coordination between speech and gesture arises (see next Section). In particular, it allows us to manipulate the interaction between modality-specific production processes. As a first exploration, we report results on how processing time as a cognitive resource affects the observable meaning coordination.

The production process is initiated by setting the communicative intention "introduce churchwindow-1". Upon receiving this goal, the image generator activates visuo-spatial imagery of the church window in VSR, and the message generator activates symbolic representations of non-spatial semantic concepts in SPR. These activations spread through memory and lead to invokation of SMCs for, e.g., 'round' (bound to churchwindow-1) and 'at-top-of' (the church-tower), as well as instantiation of the corresponding SPR entries. SMCs along with their linked entries in VSR and SPR attain highest and most slowly decaying activation values.

After a preset number of processing cycles, both generators retrieve modality-specific information from memory with a probability depending on current activation values, leading to 'round' and 'at-top-of' concepts being encoded in speech and gesture in a less coordinated way: the message generator may retrieve only information about the salient shape of the window, but not about its position relative to other entities. Accordingly, a sentence like "The church has a round window" gets formulated. The image generator, on the other hand, may receive information about the entity's position as well. This can result in shape depicting gestures, like drawing the shape of the window in the air, or a static posturing gesture where the hands becoming a model of the circular shape. As the position of the entity is also available, the gesture would be performed in that part of gesture space. So, the gesture would be non-redundant to speech, supplementing it with the position of the entity.

If more time is available, however, the contents expressed either verbally or gesturally tend to converge. The message generator will start to (re-)activate those entries being retrieved and selected by the speech formulator. This results in multimodal representations being better coordinated when the modality-specific formulators start with their generation work, as it is more likely that both generators receive the same
information about shape and position of the entity. Accordingly, the speech formulator is now enabled to plan a sentence like "The church has a round window at the top" which-like the gesture(s) described previously-encodes both, shape information and the entity's relative position.

To quantify these observations, we ran a simulation experiment in which we manipulated the available time (in terms of memory update cycles) before the model had to come up with a sentence and a gesture. We analyzed the resulting sentences and gestures for semantic redundancy/non-redundancy. We defined two conditions: A time-constrained condition with a certain number $N$ of cycles and a condition with twice as many cycles. We ran the model 100 times in each condition. Fig. 4 shows that non-redundant (supplementary) gestures dominate in those runs with stricter temporal limitations, while redundant ones become more likely when time available is increased. Notably, the information conveyed by gesture was similar in both conditions. So, the higher redundancy in the less time-constrained condition is mostly due to the fact that the verbal utterances were richer in content.


Figure 4: Number of semantic gesture features encoded redundantly vs. non-redundantly with speech in 100 simulation runs in more (left) or less (right) time-constrained conditions (note that a gesture may carry more than one feature).

## Discussion and conclusions

We have presented the first model to explain semantic coordination between speech and gesture in terms of (1) how visuospatial and symbolic-propositional memory entries are dynamically linked, (2) how activation spreads in these concept structures, and (3) how this interacts with modality-specific processes of conceptualization and formulation. We believe that this model offer mechanisms and thus possible explanations for many empirical findings and hypotheses put forth in literature: The hypothesis that gestures are more likely if activation in visuo-spatial memory is higher, is directly explained by the activation-based retrieval probabilities when the image generator accesses memory; the hypothesis that non-redundant gestures are more likely when spatial codes are not transformed into verbal codes is accounted for by entries in VSR and SPR not being linked via SMC, leading to less coordinated conceptual structures and activations. Fi-
nally, the shaping of gesture by speech is accounted for, first, through SPR and SMC schematizing VSR in linguistically shaped ways and, second, through choices in linguistic formulation reinforcing activations in SPR and thus VSR.

Our simulation study showed that the model also offers a natural account for the finding that non-redundant gesture are more likely when conceptualization load is high, based on the assumption that memory-based cross-modal coordination consumes resources (memory, time) and is reduced or compromised when, e.g., time is limited. This examplifies how a model like ours can help to make hypothesis testable by giving rise to predictions that can be explored in computational simulations as well as in appropriately set up empirical experiments. While the model presented here mainly accounts for information distribution, work is underway to extend the model to account also for different ways to package information over multiple clauses, e.g., depending on available linguistic or gestural resources. This will enable to simulate cross-linguistic differences in co-speech gesturing. Another issue for future work will be to go beyond objectrelated gestures accompanying NP constructions, and to address descriptions of action events with a more complex internal structure and thus a more demanding semantic coordination to be achieved by the cognitive processes involved in multimodal conceptualization.

## Acknowledgements

This research is supported by the Deutsche Forschungsgemeinschaft (DFG) in the Collaborative Research Center 673 "Alignment in Communication" and the Center of Excellence 277 "Cognitive Interaction Technology" (CITEC).

## References

Anderson, J., Bothell, D., Byrne, M., Lebiere, C., \& Qin, Y. (2004). An integrated theory of the mind. Psychological Review, 111(4), 1036-1060.
Bavelas, J., Kenwood, C., Johnson, T., \& Philips, B. (2002). An experimental study of when and how speakers use gestures to communicate. Gesture, 2(1), 1-17.
Bergmann, K., \& Kopp, S. (2006). Verbal or visual: How information is distributed across speech and gesture in spatial dialog. In D. Schlangen \& R. Fernandez (Eds.), Proceedings of the 10th workshop on the semantics and pragmatics of dialogue (pp. 90-97).
Bergmann, K., \& Kopp, S. (2009). Increasing expressiveness for virtual agents - autonomous generation of speech and gesture for spatial description tasks. In Proc. of AAMAS 2009 (p. 361-368).
Breslow, L., Harrison, A., \& Trafton, J. (2010). Linguistic spatial gestures. In D. Salvucci \& G. Gunzelmann (Eds.), Proceedings of the 10th international conference on cognitive modeling (pp. 13-18).
Collins, A. M., \& Loftus, E. F. (1975). A spreading-activation theory of semantic processing. Psychological Review, 82(6), 407-428.

Hostetter, A., \& Alibali, M. (2007). Raise your hand if you're spatial-relations between verbal and spatial skills and gesture production. Gesture, 7, 73-95.
Hostetter, A., \& Alibali, M. (2008). Visible embodiment: Gestures as simulated action. Psychonomic Bulletin and Review, 15/3, 495-514.
Hostetter, A., \& Alibali, M. (2011). Cognitive skills and gesture-speech redundancy. Gesture, 11(1), 40-60.
Kita, S., \& Davies, T. S. (2009). Competing conceptual representations trigger co-speech representational gestures. Language and Cognitive Processes, 24(5), 761-775.
Kita, S., \& Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. Journal of Memory and Language, 48, 16-32.
Kita, S., Özyürek, A., Allen, S., Brown, A., Furman, R., \& Ishizuka, T. (2007). Relations between syntactic encoding and co-speech gestures: Implications for a model of speech and gesture production. Language and Cognitive Processes, 22, 1212-1236.
Kopp, S., Bergmann, K., \& Wachsmuth, I. (2008). Multimodal communication from multimodal thinking - towards an integrated model of speech and gesture production. Semantic Computing, 2(1), 115-136.
Krauss, R., Chen, Y., \& Gottesman, R. (2000). Lexical gestures and lexical access: A process model. In D. McNeill (Ed.), Language and gesture (pp. 261283). Cambridge, UK: Cambridge University Press.

Levelt, W. J. M. (1989). Speaking: From intention to articulation. MIT Press.
McNeill, D., \& Duncan, S. (2000). Growth points in thinking-for-speaking. In D. McNeill (Ed.), Language and gesture (pp. 141-161). Cambridge, UK: Cambridge University Press.
Melinger, A., \& Kita, S. (2007). Conceptualisation load triggers gesture production. Language and Cognitive Processes, 22(4), 473-500.
Özyürek, A. (2002). Speech-gesture relationship across languages and in second language learners: Implications for spatial thinking and speaking. In Proceedings of the 26th annual boston university conference on language development (pp. 500-509).
Ruiter, J. de. (2000). The production of gesture and speech. In D. McNeill (Ed.), Language and gesture (pp. 284311). Cambridge, UK: Cambridge University Press.

Slobin, D. I. (1996). From "thought and language" to "thinking for speaking". In J. J. Gumperz \& S. C. Levison (Eds.), Rethinking linguistic relativity (p. 70-96). Cambridge Univ. Press.
Sowa, T., \& Kopp, S. (2003). A cognitive model for the representation and processing of shape-related gestures. In Proc. European Cognitive Science Conference.

# Space and Time are Mutually Contagious in Sound 

Alexander Kranjec (kranjeca@duq.edu)<br>Psychology Department, Duquesne University<br>Center for the Neural Basis of Cognition, Carnegie Mellon University<br>Pittsburgh, PA, USA<br>Matthew Lehet (mil@andrew.cmu.edu)<br>Psychology Department, Carnegie Mellon University<br>Pittsburgh, PA, USA<br>Anjan Chatterjee (anjan@mail.med.upenn.edu)<br>Neurology Department and Center for Cognitive Neuroscience, University of Pennsylvania<br>Philadelphia, PA, USA


#### Abstract

Time is talked about in terms of space more frequently than the opposite is true. Past experimental evidence suggests that this asymmetry runs deep, with results suggesting that temporal concepts and percepts find structure in spatial representations but not vice versa. However, these studies frequently involve verbal and/or visual stimuli. Because vision makes a privileged contribution to spatial processing it is unclear whether these results speak to a deep asymmetry between time and space, or a modality specific one. The present study was motivated by this ambiguity and a complementary correspondence between audition and temporal processing. In an auditory perceptual task, duration and spatial displacement judgments were shown to be mutually contagious. Irrelevant temporal information influenced spatial judgments and vice versa with a larger effect of time on space. The results suggest that the perceptual asymmetry between domains does not generalize across modalities and that time is not fundamentally more abstract than space.


Keywords: space and time; language and thought; metaphor; embodiment

## Introduction

Time is frequently talked about using the language of space (Clark, 1973; Hasplemath, 1997; Tenbrink, 2007). Events can be long or short, and can occupy a place that is either behind or in front of us in time. Space is used to talk about time not only frequently but also meaningfully. We talk about temporal extent or duration in terms of distance, and the past and future in egocentric locational terms. These ways of talking and thinking about space and time are thought to reflect something about how we experience these domains together. We may talk about duration in terms of length because it takes more time to visually scan or travel through more extended space, and the past as behind
because as we walk forward, objects we pass begin to occupy the unseen space behind our bodies becoming accessible only to memory as part of a temporal past. Experimental studies support the idea that the ways in which we experience space play a role in structuring the semantics of time (Boroditsky, 2000, 2001; Boroditsky \& Ramscar, 2002; Casasanto \& Boroditsky, 2008; Kranjec, Cardillo, Schmidt, \& Chatterjee, 2010; Kranjec \& McDonough, 2011; Matlock, Ramscar, \& Boroditsky, 2005; Miles, Nind, \& Macrae, 2010; Nunez, Motz, \& Teuschner, 2006; Nunez \& Sweetser, 2006; Torralbo, Santiago, \& Lupianez, 2006).

In language, time-space relations are relatively asymmetrical. Not only is time talked about in spatial terms much more frequently than space is talked about in terms of time, but in many ways time must be talked about using the language of space, whereas the opposite is not true. These linguistic patterns have been interpreted to suggest a deeper conceptual organization. According to conceptual metaphor theory (Lakoff \& Johnson, 1999) we think about relatively abstract target domains (like time) in terms of more concrete source domains (like space). This basic organizational principle is purported to serve the functional role of making more abstract concepts easier to talk and think about. It is argued that we depend on such a hierarchy because, for example, we can directly see and touch things "in space" in a way that we cannot "in time." This suggests that thinking about time in terms of space runs cognitively deep, and reflects a mental organization more fundamental than that observed at the relatively superficial level of words.

In a widely cited paper, Casasanto and Boroditsky (2008) sought strong experimental evidence for this theoretical organizational principle. Specifically, they wanted to know if the asymmetry of space-time metaphors in language predicted a similar asymmetry in perception. They reasoned
that low-level perceptual biases demonstrating concordant asymmetry with patterns found in language would provide strong evidence that temporal representations are grounded on more concrete spatial representations.

In their study, participants viewed growing or static lines one at a time on a computer screen. Lines could be of nine durations crossed with nine displacement sizes to produce 81 unique stimuli. After the presentation of each line, participants were randomly prompted to either reproduce a line's spatial extent (by dragging a mouse) or a line's duration (by clicking a mouse). Each line was presented twice: once in each kind of reproduction trial (i.e., displacement or duration estimation).

They found that the remembered size of a line in space concordantly modulated recall for its duration, but not vice versa. That is, (spatially) longer lines were remembered as being presented for longer times, but lines of greater durations were not remembered as having greater spatial extent. The results were consistent with the idea that asymmetrical patterns of space-time mappings in language are preserved further down at the level of perception. They concluded, "these findings provide evidence that the metaphorical relationship between space and time observed in language also exists in our more basic representations of distance and duration" (p. 592).

That we use space to think about time is now widely acknowledged. The idea that time is fundamentally (i.e., ontologically) more abstract than space is regarded as a prerequisite for this relation. However, there are still reasons to question this general view. First, neural data supporting the idea that our temporal concepts are grounded in embodied spatial representations is scarce, partly because it is not entirely clear what an embodied spatial representation is in the first place (Kranjec \& Chatterjee, 2010). Furthermore, recent fMRI evidence suggests that temporal and spatial concepts do not necessarily have privileged relations in the brain (Kranjec, Cardillo, Lehet, \& Chatterjee, 2012). By focusing on space, embodied theories have neglected to investigate temporal conceptual grounding in neural systems that instantiate time perception in the body. Importantly, studies in this area of research tend to rely on visual tasks. This in particular makes it unclear whether observed behavioral asymmetries between time and space reflect (1) general ontological (or even metaphysical) relations dependent on each domain's relative level of "abstractness" or (2) a less general, modalityspecific contribution of visual representations in humans.

To distinguish between these two alternatives, the present study directly probes time-space relations in the auditory domain. Audition was selected because there are intuitive reasons to think that those time-space asymmetries observed in vision might actually be reversed in sound. This is because time, more than space, seems to be an intimate
part of our auditory experience. [But see (Shamma, 2001) for a dissenting view.] For example, whereas spatial relations and visual objects tend to be persistent, sound, like time, is relatively transient (Galton, 2011). While the retina preserves analog spatial relations in early representations, the cochlea does not (Moore, 1977; Ratliff \& Hartline, 1974). Sound localization is less precise than object localization in vision (Kubovy, 1988). And generally, temporal information is more critical and/or salient in common forms of experience grounded in sound perception (e.g., music and speech). In the context of music, "when" a sound occurs matters much more than "where" it does. In speech, the ability to perceive differences in voice onset time is critical for discriminating between phonological categories (Blumstein, Cooper, Zurif, \& Caramazza, 1977). Thus, one might argue that in many critical contexts, relations between sound and time are relatively more concrete than relations between sound and space. The present research directly addresses these issues with a task closely following Casasanto and Boroditsky (2008) but using auditory instead of visual stimuli.

## Methods

Twenty members of the University of Pennsylvania community participated for payment. All participants were right-handed, native English speakers, and between 18-26 years of age. The participants were equiped with headphones and seated at a computer for a self-paced experiment. Participants initiated the beginning of each new trial and the start of each within-trial component. Each trial consisted of two sounds, a target sound followed by a playback sound. In the first part of each trial, the target sound was presented, and participants were instructed to attend to both spatial and temporal aspects of the stimulus. Target sounds consisted of bursts of white noise that changed in location relative to a participant's head position across time. White noise bursts were of nine durations (lasting between 1000 and 5000 ms with 500 ms increments) and 9 distances (moving between .5 and 4.5 m in increments of .5 m ). All durations and distances were crossed to create 81 distinct target sounds. The initial location of the target sound was an average of 2.75 m to the left or right of the listener with a jitter of between .1 and .5 m . Starting locations on the right indicated leftward moving trials and starting locations on the left indicated rightward moving trials. Starting locations were randomly assigned to stimuli with an even number of right and leftward moving trials. The plane of movement was 1 meter in front of the listener. Stimuli were created using Matlab and played using the OpenAL library provided with Psychophysics Toolbox extensions (Brainard, 1997). The OpenAL library is designed to model sounds moving in virtual metric space.

After attending to the target sound, participants were prompted to reproduce either the sound's duration or distance and then instructed to press the spacebar to begin
the playback sound. In this second part of each trial, the playback sound provided the medium for the participant's response. The playback sound began in the final location of the preceding target sound and moved in the reverse direction. So, if a target sound moved rightward, the playback sound moved leftward, and vice versa. On distance trials, participants were instructed to respond when the playback sound reached the start location of the target sound, thereby reproducing the distance from head to start point. In this manner, the participant's head provided a fixed reference point for judging distance. On duration trials, participants were instructed to respond when the playback sound duration was equal to the target sound duration. The playback sound lasted for a fixed 8500 ms and moved 3.5 m past the starting location of the target sound or until the participant responded. The playback sounds were designed in such a manner as to allow participants the possibility to both overshoot and undershoot their estimates. Participants heard each target sound in both duration and distance conditions for a total of 162 trials.

## Results

The results (Fig. 1) demonstrate that actual durations affected estimates of spatial displacement (Fig. 1A: y $=$ $0.0002 \mathrm{x}+1.4208, \mathrm{r}^{2}=.98, \mathrm{df}=7, \mathrm{p}<.01$ ) and that actual spatial displacement affected estimates of duration (Fig. 1B: $\mathrm{y}=128.97 \mathrm{x}+2532.8, \mathrm{r}^{2}=.88, \mathrm{df}=7, \mathrm{p}<.01$ ). On distance trials, for stimuli of the same average duration ( 3000 ms ), sounds shorter in length were judged to be of shorter duration, and sounds longer in length were judged to be of longer duration. On duration trials, for stimuli of the same average displacement ( 2.5 m ) sounds of shorter durations were judged to be shorter in length, and sounds of longer durations were judged to be longer in length. Time and space were mutually contagious in that irrelevant information in the task-irrelevant domain affected participants' estimates of both duration and spatial displacement. Compatible effects were found using multiple regression analyses. Duration was significantly correlated with distance judgments even when variance associated with each trial's actual distance was removed $[\rho r(81)=.81 ; \mathrm{p}<.01]$. Distance was significantly correlated with duration judgments when variance associated with actual duration was removed $[\rho r(81)=.64$; $\mathrm{p}<.01]$. There was no effect of direction (left-moving vs. right moving trials).

Figure 1A-D: (Opposite column) Averaged duration and spatial displacement estimates. The top scatterplots depict between domain effects. The dotted lines represent the line predicted by perfect performance. All space and time intervals were fully crossed. The average of all 9 displacement intervals is 2.5 m at each duration (1A) and the average of the all duration intervals is 3000 ms at each displacement length (1B). The bottom scatterplots (1C and 1D) depict within domain effects. Error bars refer to standard error of the mean.


Participants' overall estimates of duration and displacement were very accurate. The effects of actual duration on estimated duration (Fig. 1C: $\mathrm{y}=0.6805 \mathrm{x}+813.64, \mathrm{r}^{2}=.99$, $\mathrm{df}=7, \mathrm{p}<.001$ ) and actual displacement on estimated displacement (Fig. 1D: $y=0.6374 x+0.4115, r^{2}=.99, \mathrm{df}=$ $7, \mathrm{p}<.001$ ) were also very similar to each other and to analogous analyses of accuracy in Casasanto \& Boroditsky (2008). This suggests that participants were approximately equal in accuracy when making duration and distance judgments within the present experiment and between comparable experiments using auditory and visual stimuli.

The effect of duration on displacement was significantly greater than the effect of displacement on duration (Fig. 2) ( 2 A vs. 2 B : difference of correlations $=0.10 ; z=1.7$ onetailed, $p<.05$ ). However, some caution should be taken when interpreting this result. It is unclear to us whether differences in perceptual judgments between domains can be directly compared at such a fine grain when arbitrarily defined scales, intervals, and ranges (e.g., in seconds and meters) are used to define temporal and spatial aspects of the stimuli. This is a concern even though spatial and temporal judgments focused on identical stimuli. It is possible that other scaled relations could yield different patterns of results.


Figure 2: Comparing differences between effects. Effects of: Duration on Displacement (Dur|Dis); Displacement on Duration (Dis|Dur); Duration on Duration (Dur|Dur); Displacement on Displacement (Dis|Dis). A-D refer to corresponding scatterplots in Figure 1.

## Discussion

While strong claims about asymmetrical ontological relations between space and time in the auditory domain are premature, we can report a significant pattern of time-space asymmetry in the auditory domain. This asymmetry is predicted by the temporal nature of auditory processing and runs in the opposite direction of the asymmetry found in the visual domain as predicted by patterns of language use (Casasanto \& Boroditsky, 2008) and the relatively spatial nature of vision.

A prior study using visual stimuli found strong evidence for an asymmetrical relationship between space and time, such
that the remembered size of a stimulus in space modulated recall for its duration, but not vice versa (Casasanto \& Boroditsky, 2008). In contrast, the present study having an analogous design but using auditory stimuli found that space and time are mutually contagious. Furthermore, as predicted by the privileged relation between auditory and temporal processing, the perceived duration of a stimulus had a larger effect on perceived displacement than the reverse. While this is suggestive of a perceptual asymmetry running opposite to that observed in the visual domain, broader claims regarding a deep ontological asymmetry between time and space in the auditory domain are currently unwarranted. Although "in sound," time appears to influence judgments of spatial displacement more than vice versa, these results may not generalize to other scales, intervals, and ranges of time-space relations. And importantly, the effect of spatial displacement on duration estimates was still strong in the auditory domain ( $\mathrm{r}^{2}=.88$ ). In Casasanto and Boroditsky's 2008 study, actual duration had no effect on spatial displacement judgments.

These results demonstrate that time is not necessarily or fundamentally more abstract than space, and suggest that previously observed verbal and mental asymmetries of representing time in terms of space may at least be partially dependent on the human disposition to think visually. The general idea that visuospatial representations are central to how we think and reason in a more general sense about the world is well established (Johnson-Laird, 1986; Tversky, 2005) as is the more specific idea that spatial, visual, and verbal representations are deeply intertwined in giving rise to abstract semantics (Chatterjee, 2001, 2008; Jackendoff, 1996; Talmy, 2000). In the context of previous research demonstrating a strong asymmetry for time-space relations, the results of the present study suggest something very important about the nature of those "embodied spatial representations" that appear to structure patterns in language and thought. That is, such representations are likely visuospatial in nature.

It should be noted that the present results do not refute those reported in Casasanto and Boroditsky's analogous (2008) study. Rather, our results suggest that the hypothesis that time is more abstract than space at the level of a deep ontology and/or basic cognitive architecture may need to be revised. This should not come as a total surprise because "space" is itself a very abstract concept and, like "time," cannot be directly seen, touched, or heard. The present data suggest that what makes certain spatial or temporal relations more or less abstract is the sensory modality in which those relations are processed or experienced. As such, the present results support a refined but intuitive view of embodied cognition that takes into account contributions of a particular sensory modality in processing the abstract qualities of a stimulus. While space and time may be equally abstract, relations between objects immersed in either substrate (whether seen or heard) may be more or less
so depending on a range of species-specific and contextual variables.

## References

Blumstein, S. E., Cooper, W. E., Zurif, E. B., \& Caramazza, A. (1977). The perception and production of voiceonset time in aphasia. Neuropsychologia, 15(3), 371-372.
Boroditsky, L. (2000). Metaphoric structuring: understanding time through spatial metaphors. Cognition, 75(1), 1-28.
Boroditsky, L. (2001). Does language shape thought?: Mandarin and English speakers' conceptions of time. Cognitive Psychology, 43(1), 1-22.
Boroditsky, L., \& Ramscar, M. (2002). The roles of body and mind in abstract thought. Psychological Science, 13(2), 185-189.
Brainard, D. H. (1997). The psychophysics toolbox. Spatial vision, 10(4), 433-436.
Casasanto, D., \& Boroditsky, L. (2008). Time in the mind: Using space to think about time. Cognition, 106(2), 579-593.
Chatterjee, A. (2001). Language and space: some interactions. Trends in Cognitive Science, 5, 55-61.
Chatterjee, A. (2008). The neural organization of spatial thought and language. Seminars in Speech and Language, 29, 226-238.
Clark, H. H. (1973). Space, time, semantics and the child. In T. Moore (Ed.), Cognitive development and the acquisition of language. New York: Academic Press.
Galton, A. (2011). Time flies but space does not: Limits to the spatialisation of time. Journal of Pragmatics, 43(3), 695-703.
Hasplemath, M. (1997). From space to time: Temporal adverbials in the world's languages. Munich \& Newcastle: Lincom Europa.
Jackendoff, R. (1996). The architecture of the linguisticspatial interface. In P. Bloom, M. A. Peterson, L. Nadel \& M. F. Garrett (Eds.), Language and Space (pp. 1-30). Cambridge, MA.
Johnson-Laird, P. N. (1986). Mental models: Towards a cognitive science of language, inference, and consciousness: Harvard Univ Pr.
Kranjec, A., Cardillo, E. R., Lehet, M., \& Chatterjee, A. (2012). Deconstructing events: The neural bases for space, time, and causality. Journal of Cognitive Neuroscience, 24(1), 1-16.
Kranjec, A., Cardillo, E. R., Schmidt, G. L., \& Chatterjee, A. (2010). Prescribed spatial prepositions influence how we think about time. Cognition, 114(1), 111116.

Kranjec, A., \& Chatterjee, A. (2010). Are temporal concepts embodied? A challenge for cognitive neuroscience. Frontiers in Psychology, 1(240), doi: 10.3389/fpsyg.2010.00240.

Kranjec, A., \& McDonough, L. (2011). The implicit and explicit embodiment of time Journal of Pragmatics, 43(3), 735-748.
Kubovy, M. (1988). Should we resist the seductiveness of the space: time:: vision: audition analogy?
Lakoff, G., \& Johnson, M. (1999). Philosophy in the Flesh. New York, NY: Basic Books.
Matlock, T., Ramscar, M., \& Boroditsky, L. (2005). On the experiential link between spatial and temporal language. Cognitive Science, 29(4), 655-664.
Miles, L. K., Nind, L. K., \& Macrae, C. N. (2010). Moving through time. Psychological Science, 21(2), 222223.

Moore, B. C. (1977). Introduction to the psychology of hearing. Baltimore: University Park Press.
Nunez, R., Motz, B., \& Teuschner, U. (2006). Time after Time: The psychological reality of the Ego-and Time-Reference-Point distinction in metaphorical construals of time. . Metaphor and Symbol., 21, 133-146.
Nunez, R., \& Sweetser, E. (2006). With the Future Behind Them : Convergent Evidence From Aymara Language and Gesture in the Crosslinguistic Comparison of Spatial Construals of Time. Cognitive Science, 30(3), 133-146.
Ratliff, F., \& Hartline, H. K. (1974). Studies on Excitation and Inhibition in the Retina: A collection of Papers from the Laboratories of $H$. Keffer Hartline: Rockefeller Univ. Press.
Shamma, S. (2001). On the role of space and time in auditory processing. Trends in Cognitive Sciences, 5(8), 340-348.
Talmy, L. (2000). Towards a cognitive semantics: Concept structuring systems. Cambridge, MA: The MIT Press.
Tenbrink, T. (2007). Space, time, and the use of language: An investigation of relationships. Berlin: Mouton de Gruyter.
Torralbo, A., Santiago, J., \& Lupianez, J. (2006). Flexible conceptual projection of time onto spatial frames of reference. Cognitive Science, 30(4), 745-757.
Tversky, B. (2005). Visuospatial reasoning. The Cambridge handbook of thinking and reasoning, 209-249.

# Competition, event comprehension, and dynamic location information in sentence processing 

Anuenue Kukona (a.b.bakerkukona@dundee.ac.uk)<br>School of Psychology, University of Dundee<br>DD1 4HN UK

Gerry T. M. Altmann (g.altmann@psych.york.ac.uk)<br>Department of Psychology, University of York YO10 5DD UK

Yuki Kamide (y.kamide@dundee.ac.uk)
School of Psychology, University of Dundee
DD1 4HN UK


#### Abstract

Two experiments in the visual world paradigm investigated competition in sentence processing from dynamic eventrelated information about location. In Experiment 1, listeners viewed visual arrays with container objects like a bowl, jar, pan, and jug, while they heard sentences like "The boy will pour the sweetcorn from the bowl into the jar, and he will pour the gravy from the pan into the jug. But first/And then, he will taste the sweetcorn." While "But first" contexts referred to the "source" location of the discourse-final noun (e.g., "sweetcorn"), "And then" contexts referred to its "goal" location. In Experiment 2, listeners always heard "And then" contexts. We found that listeners rapidly fixated contextrelevant locations. Crucially, they also fixated locations that were context-irrelevant, but related to the discourse-final noun, suggesting object competition, or consistent with abstract location information implied by "But first" (source) or "And then" (goal), suggesting location competition.


Keywords: Competition; Event comprehension; Location; Visual world paradigm.

## Introduction

Everyday, we use language in dynamic real world settings that change along any number of dimensions. One such dimension is location: for example, objects like car keys and TV remotes are routinely involved in actions and events that result in (often frustrating!) changes of location. Findings from the Visual world paradigm (Tanenhaus, SpiveyKnowlton, Eberhard, \& Sedivy, 1995), which presents listeners with spoken language about a visual context, have revealed that listeners' eye movements are rapidly guided by dynamic location information (i.e., information about the multiple instantiations of an object at different locations over event time) during sentence processing.

For example, Chambers and San Juan (2008) instructed listeners to move objects around a visual array, and then they presented listeners with sentences like "Now return the..." They found that listeners anticipatorily fixated previously moved objects, consistent with "return," compared to previously unmoved objects. Similarly, Altmann and Kamide (2009) presented listeners with visual
scenes with objects like a glass (on the floor), table, and bookshelf, followed by a blank screen, and sentences like "The woman will put the glass onto the table. Then, she will pick up the bottle, and pour the wine carefully into the glass." At the discourse-final "the glass," they found that listeners were more likely to fixate the glass's new location (the prior location onscreen of the table), consistent with the sentence context, compared to its initial location on the floor.

## Location-based competition?

Thus, when language comprehenders have dynamic eventrelated information about an object's location (i.e., information about where an object is and/or where it was and/or where it will be), they must resolve which locations are relevant to a sentence context, and which are not. The findings of Chambers and San Juan (2008) and Altmann and Kamide (2009) suggest that language comprehenders rapidly integrate location information with sentence context information, and rapidly retrieve context-relevant locations. Here, we ask a closely related question: do language comprehenders also retrieve context-irrelevant location information? In other words, do irrelevant locations compete with relevant locations?

Hoover and Richardson (2008) addressed this question in a study that used a memory recall task. Their listeners heard spoken facts from a burrowing creature at different locations in a visual display, followed by a question about one of the facts. During the question, they found that listeners were more likely to fixate the location where the queried fact had been presented (compared to distractor locations that the creature had not visited). However, listeners also fixated the location where the non-queried fact had been presented.

Hoover and Richardson (2008)'s findings suggest that location information drives competition effects in eye movements: fixations to locations that are related to target information, but context-irrelevant. This complements competition observed elsewhere: for example, Huettig and Altmann (2005) presented listeners with visual arrays with objects like a trumpet and goat, while they heard words like
"piano." They found that listeners were more likely to fixate trumpet, which was of the same category as "piano" (e.g., musical instrument), compared to distractors (e.g., goat). Hoover and Richardson's findings suggest that in addition to effects based on long-term, semantic knowledge (Huettig \& Altmann, 2005), competition in language also stems from short-term, situated location information.

However, Hoover and Richardson (2008)'s findings also raise a number of new questions. First, evidence for location-based competition is mixed. For example, Altmann and Kamide (2009) observed no competition during "the glass" in their study: listeners were no more likely to fixate the glass's (context-irrelevant) initial location on the floor compared to distractors (e.g., bookshelf). Thus, it is unclear how location-based competition impacts online sentence processing.

Second, and more importantly, Hoover and Richardson (2008)'s findings seem to depend more precisely on object, rather than location, competition. In a second condition in which two different creatures presented their facts, they found no competition effects (i.e., listeners did not fixate the location where the non-queried fact had been presented). Thus, perhaps a more precise way of thinking about their results is that the creature was competing with itself, insomuch as it had to be represented at two locations, rather than that the associated locations were competing. Indeed, this claim is compatible with recent work by Hindy, Altmann, Kalenik, \& Thomspon-Schill (2012): they found that conflict-associated brain regions were activated during sentences that described a state change (e.g., "The squirrel will crack the acorn"). They suggest that event-related changes activate multiple instantiations of an object, and that the representation of the object before the change competes with (and engenders conflict with) the representation of the object after the change.

## Current experiments

In the current study, we tested for location-based and objectbased competition effects in sentence processing. In two experiments, we addressed a critical difference between Altmann and Kamide (2009) and Hoover and Richardson (2008): the predictability of the context-relevant location information. In Altmann and Kamide, the discourse contexts were highly predictable (e.g., the discourse-final "the glass" could be anticipated based on the verb selectional restrictions of "pour"), and consequently they observed strong anticipatory effects (e.g., listeners fixated contextrelevant locations prior to the discourse-final noun). By contrast, Hoover and Richardson (2008) queried facts (and their associated locations) at random. Here, we used discourse contexts that were closely related to Altmann and Kamide (2009), but that did not allow for anticipation. Listeners viewed visual arrays with container objects like a bowl, jar, pan, and jug (Figure 1), while they heard sentence pairs like ( $1 \mathrm{a}, \mathrm{b}$ ).
(1a) The boy will pour the sweetcorn from the bowl into the jar, and he will pour the gravy from the pan into the jug.
(1b) But first/And then, he will taste the sweetcorn.
Two critical referents (e.g., "sweetcorn" and "gravy") were described in (1a), so that listeners could not anticipate the discourse-final noun in (1b) (half of trials re-referred to the first critical referent ["sweetcorn"], and half to the second critical referent ["gravy"]). In Experiment 1, listeners heard both "But first" contexts, which referred to the "source" location (e.g., bowl) of the discourse-final noun (e.g., "sweetcorn"), and "And then" contexts, which referred to its "goal" location (e.g., jar). In Experiment 2, listeners only heard "And then" contexts, which always referred to the "goal" location.

Crucially, our design allowed us to disentangle locationbased and object-based competition. In both experiments, we tested for object-based competition (e.g., competition between sweetcorn and itself) by examining fixations to container objects that were related to the discourse-final noun but inconsistent with "But first/And then" (e.g., jar, the goal location of "sweetcorn," following "But first, he will taste the sweetcorn"). On the other hand, we tested for location-based competition (e.g., competition between source/goal locations, which were not linked via an object) by examining fixations to container objects that were consistent with "But first" (source location) or "And then" (goal location) but unrelated to the discourse-final noun (e.g., pan, the source location of "gravy," following "But first, he will taste the sweetcorn").

## Experiment 1



Figure 1: Example visual array from Experiments 1 and 2.

## Methods

Participants Forty-eight individuals from the University of Dundee community participated for course credit or $£ 4$.

Materials We constructed 48 sentence pairs like (1a,b). The first sentence described the critical contents of two containers moving either from their initial locations into two new locations (1a), or into two new locations from their initial locations. The second sentence referred to either the


Figure 2: Average ( $95 \%$ CI) proportions of fixations to source and goal locations of the sweetcorn (target referent) and gravy (competitor referent) during "But first..." (A) and "And then..." (B) "he will taste the sweetcorn" in Experiment 1.
first (e.g., "sweetcorn") or second (e.g., "gravy") of the critical contents, and either its initial ("But first") or new ("And then") location (1b). Each item had eight forms, reflecting the crossing of movement description (from-into and into-from), conjunction ("But first" and "And then"), and discourse-final noun ("sweetcorn" and "gravy"), which were rotated across participants (see Appendix A). Visual arrays (Figure 1) depicted the container objects in the four corners of the display, but not their contents.

For each critical referent, the "source" location was the location of the object before the described movement (e.g., sweetcorn: bowl; gravy: pan), and the "goal" location was
the location of the object after the described movement (e.g., sweetcorn: jar; gravy: jug).

Procedure We used an SR Research EyeLink II headmounted eye tracker, sampling at 500 Hz from one eye (viewing was binocular). The experiments involved a look-and-listen task: participants were instructed to look carefully at the visual arrays, and to listen carefully to the sentences. The onset of the visual stimulus preceded the onset of the spoken stimulus by $1,000 \mathrm{~ms}$. A trial ended $3,000 \mathrm{~ms}$ after the offset of the sentence.

The eye tracker was recalibrated after every eighth trial. The experiment began with four practice trials, and included

12 filler trials (which described a single critical object, rather than two). The experiment was approximately 40 minutes in length.

## Results

Average proportions of fixations to source and goal locations of the sweetcorn (target referent) and gravy (competitor referent) in the visual array are plotted during "But first, he will taste the sweetcorn" in Figure 2A and during "And then, he will taste the sweetcorn" in Figure 2B. Eye movements were resynchronized at the onset of each of the plotted windows ("But first/And then," "he will taste the," "sweetcorn").

We analyzed eye movements during three time windows: during "But first/And then, he will taste the," at the offset of "sweetcorn," and between sentence offset and 500 ms following sentence offset. These windows directly precede, and follow, the critical discourse-final noun. Average proportions of fixations to each region of interest (ROI) are reported within each time window in Table 1 ("But first...") and Table 2 ("And then..."). We submitted proportions of fixations to planned pairwise comparisons (using paired $t$ tests).

Table 1: Average (SD) proportions of fixations to source and goal locations of the sweetcorn (target referent) and gravy (competitor referent) during "But first, he will taste the," at the offset of "sweetcorn," and between sentence offset and 500 ms past sentence offset in Experiment 1.

| ROI | "But" | "sweetcorn" | $\mathbf{+ 5 0 0}$ |
| :--- | :---: | :---: | :---: |
| sweetcorn (source) | $.19(.07)$ | $.35(.13)$ | $.41(.16)$ |
| sweetcorn (goal) | $.25(.08)$ | $.26(.09)$ | $.26(.10)$ |
| gravy (source) | $.20(.06)$ | $.16(.08)$ | $.14(.07)$ |
| gravy (goal) | $.26(.08)$ | $.14(.10)$ | $.11(.07)$ |

Table 2: Average (SD) proportions of fixations to source and goal locations of the sweetcorn (target referent) and gravy (competitor referent) during "And then, he will taste the," at the offset of "sweetcorn," and between sentence offset and 500 ms past sentence offset in Experiment 1.

| ROI | "And" | "sweetcorn" | $\mathbf{+ 5 0 0}$ |
| :--- | :---: | :---: | :---: |
| sweetcorn (source) | $.18(.08)$ | $.23(.10)$ | $.24(.10)$ |
| sweetcorn (goal) | $.26(.08)$ | $.36(.14)$ | $.42(.14)$ |
| gravy (source) | $.19(.08)$ | $.14(.08)$ | $.11(.08)$ |
| gravy (goal) | $.26(.08)$ | $.18(.11)$ | $.14(.09)$ |

"But first." During "But first, he will taste the," fixations to goal locations were reliably greater than to source locations ( $p$ 's $<.001$ ). However, fixations to sweetcorn and gravy source locations did not differ reliably

[^126]( $p=.68$ ), and fixations to sweetcorn and gravy goal locations did not differ reliably ( $p=.60$ ).

At the offset of "sweetcorn," fixations to the (contextrelevant) sweetcorn source location were reliably greater than to the sweetcorn goal location ( $p<.01$ ), gravy source location ( $p<.001$ ), and gravy goal location ( $p<.001$ ); fixations to the sweetcorn goal location were reliably greater than to the gravy source location ( $p<.001$ ) and gravy goal location ( $p<.001$ ), capturing object competition; and fixations to the gravy source location were not reliably different from the gravy goal location $(p=.49)$.

Finally, during the 500 ms time window following sentence offset, fixations to the sweetcorn source location were reliably greater than to the sweetcorn goal location ( $p$ $<.001$ ), gravy source location ( $p<.001$ ), and gravy goal location ( $p<.001$ ); fixations to the sweetcorn goal location were reliably greater than to the gravy source location ( $p<$ .001 ) and gravy goal location ( $p<.001$ ), capturing object competition; and fixations to the gravy source location were reliably greater than to the gravy goal location ( $p<.05$ ), capturing location competition.
"And then." During "And then, he will taste the," fixations to goal locations were reliably greater than to source locations ( $p$ 's < .001). However, fixations to sweetcorn and gravy source locations did not differ reliably ( $p=.66$ ), and fixations to sweetcorn and gravy goal locations did not differ reliably $(p=.81)$.

At the offset of "sweetcorn," fixations to the (contextrelevant) sweetcorn goal location were reliably greater than to the sweetcorn source location ( $p<.001$ ), gravy goal location ( $p<.001$ ), and gravy source location ( $p<.001$ ); fixations to the sweetcorn source location were reliably greater than to the gravy goal location $(p=.05)$ and gravy source location ( $p<.001$ ), capturing object competition; and fixations to the gravy goal location were marginally greater than to the gravy source location ( $p=.08$ ), capturing location competition.

Finally, during the 500 ms time window following sentence offset, fixations to the sweetcorn goal location were reliably greater than to the sweetcorn source location ( $p<.001$ ), gravy goal location ( $p<.001$ ), and gravy source location ( $p<.001$ ); fixations to the sweetcorn source location were reliably greater than to the gravy goal location ( $p<.001$ ) and gravy source location ( $p<.001$ ), capturing object competition; and fixations to the gravy goal location were marginally greater than to the gravy source location ( $p$ $=.08$ ), capturing location competition.

## Discussion

During the discourse-final noun, we found that listeners were more likely to fixate context-relevant locations compared to all other locations (e.g., see fixations to the sweetcorn source location in Figure 2A and sweetcorn goal location in Figure 2B). Consistent with Chambers and San Juan (2008) and Altmann and Kamide (2009), these results suggest that listeners rapidly integrate location and sentence


Figure 3: Average ( $95 \% \mathrm{CI}$ ) proportions of fixations to source and goal locations of the sweetcorn (target referent) and gravy (competitor referent) during "And then, he will taste the sweetcorn" in Experiment 2.
context information. We also found that listeners had a strong bias to fixate goal locations prior to the discoursefinal noun for both "And then" (in which goal information was relevant) and "But first" (in which source information was relevant) contexts. This result suggests that listeners may be biased to track "current" location information: indeed, by the end of the second sentence, the "current" location corresponds to the goal location if one assumes that the description of the events in the language, and the events themselves, are closely time locked.

Crucially, we also found evidence for object-based competition: listeners were more likely to fixate locations that were related to the discourse-final noun but inconsistent with "But first/And then" compared to completely unrelated locations (e.g., see fixations to the sweetcorn goal location vs. gravy goal location in Figure 2A and sweetcorn source location vs. gravy source location in Figure 2B). This result suggests that representations of the sweetcorn at contextirrelevant locations were competing with representations of the sweetcorn at context relevant locations.

Finally, we also found evidence for location-based competition that was independent of object-based competition. Following "sweetcorn," listeners were more likely to fixate the gravy source location compared to the gravy goal location with "But first," which was consistent with source locations, although both gravy locations were unrelated to "sweetcorn." The opposite pattern was also observed with "And then," although the effect was marginal. These results suggest that source locations were competing, based on abstract information about whether a location was a source or a goal of an event, even though no object was present at two source locations.

Next, we asked: can location-based competition be reduced in a setting in which there is less uncertainty about
which referents are relevant? Thus, in Experiment 2 listeners always heard "And then" contexts, in which goal locations were relevant.

## Experiment 2

## Methods

Participants Twenty-four individuals from the University of Dundee community participated for course credit or $£ 4$.

Materials Materials were identical to Experiment 1, except that the second sentence always referred to the goal location ("And then...").

Procedure The procedure was identical to Experiment 1.

## Results

Average proportions of fixations to source and goal locations of the sweetcorn (target referent) and gravy (competitor referent) in the visual array are plotted during "And then, he will taste the sweetcorn" in Figure 3. We performed identical analyses to Experiment 1. Average proportions of fixations by ROI are reported in Table 3.

During "And then, he will taste the," fixations to goal locations were reliably greater than to source locations ( $p$ 's $<.001$ ). However, fixations to sweetcorn and gravy source locations did not differ reliably ( $p=.19$ ), and fixations to sweetcorn and gravy goal locations did not differ reliably ( $p$ $=.50$ ).

At the offset of "sweetcorn," fixations to the (contextrelevant) sweetcorn goal location were reliably greater than to the sweetcorn source location ( $p<.001$ ), gravy goal location ( $p<.001$ ), and gravy source location ( $p<.001$ );
fixations to the sweetcorn source location were reliably greater than to the gravy source location ( $p<.001$ ), capturing object competition, but not the gravy goal location ( $p=.21$ ); and fixations to the gravy goal location were reliably greater than to the gravy source location ( $p<.05$ ), capturing location competition.

Finally, during the 500 ms time window following sentence offset, fixations to the sweetcorn goal location were reliably greater than to the sweetcorn source location ( $p<.001$ ), gravy goal location ( $p<.001$ ), and gravy source location ( $p<.001$ ); fixations to the sweetcorn source location were reliably greater than to the gravy goal location ( $p<.001$ ) and gravy source location ( $p<.001$ ), capturing object competition; and fixations to the gravy goal location were reliably greater than to the gravy source location ( $p<$ .05 ), capturing location competition.

Table 3: Average (SD) proportions of fixations to source and goal locations of the sweetcorn (target referent) and gravy (competitor referent) during "And then, he will taste the," at the offset of "sweetcorn," and between sentence offset and 500 ms past sentence offset in Experiment 2.

| ROI | "And" | "sweetcorn" | $\mathbf{+ 5 0 0}$ |
| :--- | :---: | :---: | :---: |
| sweetcorn (source) | $.16(.06)$ | $.19(.07)$ | $.22(.08)$ |
| sweetcorn (goal) | $.29(.06)$ | $.44(.13)$ | $.48(.14)$ |
| gravy (source) | $.17(.04)$ | $.12(.05)$ | $.09(.05)$ |
| gravy (goal) | $.28(.05)$ | $.17(.07)$ | $.13(.06)$ |

## Discussion

The pattern of results in Experiment 2 was similar to Experiment 1. We found that listeners were more likely to fixate the context-relevant sweetcorn goal location compared to all other locations. Similarly, listeners also fixated the sweetcorn source location based on object competition, and the gravy goal location based on location competition. Further, these results suggest that competition does not depend on mentioning both source/goal locations.

## General Discussion

In two experiments, we found evidence for both locationbased and objects-based competition in sentence processing. While our sentence contexts modulated fixations to contextually-relevant locations, they did not fully inhibit fixations to contextually-irrelevant "competitor" locations. Consistent with Hoover and Richardson (2008) and Hindy et al. (2012), our results suggest that representations of an object (e.g., sweetcorn) before an event-related change compete with representations of the object after the change. In our case, the crucial event-related change was one of location, and the impact of this competition was reflected in eye movements to context-irrelevant locations. Critically, we also found evidence for object-independent location competition (e.g., between goal locations, although they corresponded to different objects [i.e., sweetcorn vs. gravy]). Taken together, these findings suggest that these
two sources of competition - objects and abstract location information - are separable, and have differential effects on sentence processing.

But our findings also diverged from the results of Altmann and Kamide (2009), who did not observe competition during "the glass" (see the Introduction). As we have suggested, a critical difference between the current experiments and their study was the predictability of the context-relevant location. Indeed, they did observe anticipatory competition effects: just prior to "the glass" (i.e., during "the wine carefully into"), listeners were reliably more likely to fixate competitor locations than distractors. Taken together, these findings suggest that location information does compete, and that this competition can precede the mention of the critical referent.

## Acknowledgments

We thank Karl Smith Byrne, Aneta Lisowska, and Robert Petrie for their help with data collection. We gratefully acknowledge support from ESRC Grants RES-062-23-2749 to GTMA and YK and RES-062-23-2842 to YK.

## Appendix A

1a. The boy will pour the sweetcorn from the bowl into the jar, and he will pour the gravy from the pan into the jug.

1b. The boy will pour the sweetcorn into the jar from the bowl, and he will pour the gravy into the jug from the pan.

2a. But first, he will taste the sweetcorn.
2 b . And then, he will taste the sweetcorn.
2c. But first, he will taste the gravy.
2d. And then, he will taste the gravy.

## References

Altmann, G. T. M., \& Kamide, Y. (2009). Discoursemediation of the mapping between language and the visual world: eye movements and mental representation. Cognition, 111, 55-71.
Chambers, C.G., \& San Juan, V. (2008). Perception and presupposition in real-time language comprehension: Insights from anticipatory processing. Cognition, 108, 2650.

Hoover, M. A., \& Richardson, D. C. (2008). When facts go down the rabbit hole: contrasting features and objecthood as indexes to memory. Cognition, 108, 533-42.
Huettig, F. \& Altmann, G.T.M. (2005). Word meaning and the control of eye fixation: semantic competitor effects and the visual world paradigm. Cognition, 96, 23-32.
Tanenhaus, M.K., Spivey-Knowlton, M., Eberhard, K., \& Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. Science, 268, 1632-1634.
Hindy, N.C., Altmann, G.T.M., Kalenik, E., \& ThomsponSchill, S.L. (2012). The effect of object-state changes on event processing: Do objects compete with themselves? The journal of neuroscience, 32, 5795-5803.

# Evolution and Modularity: The limits of mechanistic explanation 

Jaakko Kuorikoski (jaakko.kuorikoski@helsinki.fi)<br>Social and Moral Philosophy, P.O. Box 24<br>University of Helsinki, 00014 Finland

Samuli Pöyhönen (samuli.poyhonen@helsinki.fi)<br>Social and Moral Philosophy, P.O. Box 24<br>University of Helsinki, 00014 Finland


#### Abstract

${ }^{1}$ Accounts of mechanistic explanation require that complex cognitive phenomena can be decomposed into simpler subtasks. We provide a theory of explanation that rationalizes this requirement, and then we use a simple genetic algorithm exercise to demonstrate that evolution can produce designs that violate this functional modularity requirement.


Keywords: mechanism; explanation; evolution; modularity; genetic algorithm

## Introduction

Connectionism, dynamical systems theory, and new robotics have questioned whether the search for informationprocessing mechanisms provides a feasible approach to the study of biologically evolved cognitive systems such as the human mind. Whereas approaches that have their origins in classical AI tend to conceive of cognition as a set of computational operations to be mapped onto physiological parts according to functional decompositions inspired directly by the programmer's intuitions about possible efficient subroutines, the alternative research programs emphasize that biological evolution is likely to produce unintuitive designs of such complexity that renders heuristics based on decomposability and programming intuitions unusable.

In this paper we analyze the problems that evolved solutions raise to the mechanistic understanding of cognitive phenomena. The problem of understanding non-intuitive designs produced by natural selection is well-known in philosophy of psychology (e.g., Clark 1997, Ch. 5), philosophy of biology (Wimsatt 2007), and now even in popular psychology (Marcus 2008), but it has proved to be difficult to articulate without a clear idea of what exactly it is that evolutionary tinkering is supposed to hinder. The main challenge for scientific understanding is often framed and explained by pointing to the path-dependent nature and the resulting unfamiliarity of the evolved design (Jacob 1977). We argue that this is not the whole story. The aim of this paper is to provide an explicit theory of mechanistic explanation and understanding that will move us beyond intuitions towards a more systematic analysis of the nature of these challenges. We also combine our theory of explanation with a computational application of

[^127]evolutionary design: problem-solutions generated by genetic algorithms. By analyzing the nature of solutions that genetic algorithms offer to computational problems, we suggest that evolutionary designs are often hard to understand because they can exhibit non-modular functionality, and that this creates problems for strategies of mechanistic explanation.

## Mechanistic Explanation in the Cognitive Sciences

According to the proponents of the mechanistic approach to explanation (Bechtel 2008; Craver 2007; Piccinini \& Craver 2011), a central goal of the cognitive sciences is to provide understanding of system-level properties of the cognitive system in terms of the properties of its physical component parts and their organization. The most developed philosophical account of strategies for reaching such mechanistic understanding is Bechtel and Richardson's (2010) study of the heuristics of decomposition and localization (DL). The DL procedure goes roughly as follows. First, the different phenomena that the system of interest exhibits are differentiated. Then the phenomenon of interest is functionally decomposed, i.e., analyzed into a set of possible component operations that would be sufficient to produce it. One can think of this step as the formulation of a preliminary set of simpler functions that, taken together, would constitute the more complex input-output relation (the system-level phenomenon). The system is also structurally decomposed into a set of component parts. The final step is to try to localize the component operations by mapping them onto appropriate structural component parts. If this cannot be done, the fault may lie with the functional and structural decompositions or with the very identification of the phenomenon, and these may then have to be rethought. The identification and decomposition procedures will in the beginning be guided by earlier theories and common sense, but empirical evidence can always suggest that a thorough reworking of the basic ontology and the form of the possible explananda may be in order.

What the schema of Bechtel and Richardson lacks is an explicit theory of explanation providing an account of what makes such decomposition and localization exercises explanatory. Whereas cognitive theories of explanation (Churchland 1989; Thagard 2012; Waskan 2006) focus on the internal models and processes of the individuals engaged in explanation-related tasks, such conceptualization is
misleading when thinking about the goal of research: understanding. We use the term 'understanding' in order to shift our focus from single explanations to a broader collective epistemic goal. Scientific understanding proper is not what happens inside individual heads, but is constituted by the collective abilities of the scientific community to reason about and to manipulate the objects of investigation. To conceptualize scientific understanding directly based on models of individual explanatory cognition is to commit a fallacy of composition.

We therefore approach understanding as a public, behavioral concept. Understanding is a regulative label, which is attributed with regard to manifest abilities in action and correctness of reasoning. Suitable cognitive processes (comprehension), and possibly the possession of right mental models, taking place in the privacy of individual minds, are a causal prerequisite for possible fulfillment of these criteria, but these processes themselves are not the facts in virtue of which something is understood or not. They are not the criteria of understanding in the sense that we would have to know them in order to say whether somebody really understands something. The correctness of internal mental models is judged according to manifest cognitive performance, not the other way round (Ylikoski \& Kuorikoski 2010).

We take the principal criterion of understanding to be inferential performance: whether someone understands a phenomenon is assessed based on whether he or she can make correct inferences related to it. Thus our view of understanding can be linked to Woodward's (2003) widely accepted account of scientific explanation, which tells us more specifically what kinds of inferences are constitutive of specifically explanatory understanding (see also Craver 2007). Explanation consists in exhibiting functional dependency relations between variables. This is the connection between explanation and understanding: knowledge of explanatory relationships grounds understanding by implying answers to what-if-things-had-been-different questions concerning the consequences of counterfactual or hypothetical changes in the values of the explanans variable. This is the important difference between explanatory information and purely descriptive information. Whether someone understands a phenomenon is evaluated according to whether he or she can make inferences not only about its actual state, but also about possible states of the phenomenon or system in question.

## Modularity and Understanding

According to Bechtel and Richardson, decomposability is a regulative ideal in mechanistic model construction because complex systems are psychologically unmanageable for humans. Decomposition allows the explanatory task to be divided into parts that are manageable for cognitively limited beings, thereby rendering the system intelligible (Bechtel \& Richardson 2010). The idea comes originally from Simon (1962), who claimed that complex systems have to be nearly-decomposable in order to be
understandable for finite cognitive agents. Neardecomposability means that the system can be decomposed into parts in such a way that the intrinsic causal properties of the parts are more important for the behavior of the system than their relational causal properties, which are constituted also by their environment and interaction. Neardecomposable systems are thus hierarchical in the sense that the complex whole can be seen as made from a limited set of simpler parts and interactions. Hierarchical systems are more manageable for cognitively limited beings because their 'complete description' includes recurring or irrelevant elements describing similar recurring parts and nonimportant interactions. The removal of such descriptions does not hamper our understanding of the system and thus eases cognitive load.

Although there are a number of arguments that conclusively show that such informational economy by itself is not constitutive of understanding, ${ }^{2}$ we agree with Simon (and Bechtel and Richardson) in that a property closely related to near-decomposability, namely modularity, is a necessary condition for mechanistic explanations. In the case of causal-mechanistic explanations, the explanatory dependencies track the consequences of interventions (Woodward 2003; see fig. 1) and causal knowledge thus enables the goal-directed manipulation of the object of explanation. These answers are the basis of the inferential performance constitutive of causal understanding.


$$
\begin{aligned}
& X=f(U) \\
& Z=g(X) \\
& Y=h(Z, U)
\end{aligned}
$$

$$
\begin{aligned}
& X=f(U) \\
& Z=Z \\
& Y=h(z, U)
\end{aligned}
$$

Figure 1. Invariance under exogenous interventions distinguishes "deep", causal, dependencies from mere correlations. $\mathrm{P}(\mathrm{Y} \mid \mathrm{Z}=\mathrm{z})$ is not the same as $\mathrm{P}(\mathrm{Y} \mid \operatorname{set}(\mathrm{Z}=\mathrm{z}))$.

Such answers to what-if questions are derived from internal or external representations of the object of understanding. In order for these answers to be well defined, the dependencies grounding the answers have to possess some degree of independence such that a local change in an aspect of the phenomenon under study cannot ramify uncontrollably or intractably. If local modifications in a part of a system disrupt other parts (dependencies) in a way that is not explicitly specified (endogenized) in the (internal or external) representation of the system according to which the what-if inferences are made, the consequences of these changes are impossible to predict and counterfactual assertions impossible to evaluate (Woodward 2003, 333).

[^128]Therefore, a necessary condition for a representation to provide explanations, and thus understanding, of a phenomenon is that the modularity in the representation matches the modularity in the phenomenon.

If we intervene on a causal input corresponding to variable $X_{i}$ in a model of the studied system, and the intervention, no matter how surgical, also changes the dependencies within the system, or values of other variables themselves affecting variables causally downstream of $X_{i}$, the model does not give correct predictions about the consequences of the intervention. Hence, the model does not provide correct causal understanding of the system and the causal role of the variable in it. If the system cannot be correctly modeled on any level of description or decomposition so that it is modular in the way described above - if the system itself is not causally modular - no what-if-things-had-been-different questions concerning interventions in the system can be answered. This would mean that every local change brings about intractable changes elsewhere in the system to such an extent that there can be no representation that would enable a cognitively finite being to track these changes and make correct inferences about their consequences.

The problem of understanding causally non-modular systems has received some attention in the philosophy of science literature (e.g., Bechtel and Richardson 2010, Ch. 9). However, according to the DL schema, before we can even start thinking about searching for the causalmechanistic implementation of the complex system behavior, we need to formulate hypotheses about the possible functional decompositions of the behavior (see also Cummins 1983). For example, what kind of simpler subtasks could possibly produce complex cognitive capacities such as language production and comprehension, long-term memory, and visual object-recognition? Importantly, this task is separate, though not independent, from hypotheses concerning the implementation of the capacity. Although the understanding offered by the functional decomposition is not, strictly speaking, causal component operations do not cause the whole behavior because they are constitutive parts of it - the modularity constraint on understandability still applies in the following way. We can only understand the complex behavior by having knowledge of its component operations, if we can make reliable what-if inferences concerning the possible consequences of changes in the component operations for the properties of the more complex explanandum capacity. For example, we provisionally understand working memory if we can infer from possible changes in its hypothesized component operations (such as differences in the properties of the postulated phonological loop or episodic buffer) to changes in the properties of the capacity. These inferences are only possible if the functional decomposition itself is suitably modular, i.e., the consequences of "local" changes in component operations do not ramify in an intractable way, making the behavior of the whole completely holistic. We now argue that genetic algorithms demonstrate that
design-by-selection can lead to such non-modular complex behavior.

## Genetic Algorithms

From the point of view of AI, genetic algorithms (henceforth GAs) are a form of non-exhaustive but massively parallel search in the search space of a problem (Holland 1975; Mitchell 1996). Although GAs are not the only strand of evolutionary programming, they serve our purpose well because their basic principles are easy to understand and they are the most well-known kind of evolutionary programming outside computer science (Clark 1997, 2001; Mitchell 2009). GAs are useful for a number of different purposes, but here we use a simple example originally from Mitchell (2009, Ch. 9), where a GA is used to evolve a behavioral strategy for a simulated agent.

Mitchell's model shows how an algorithm mimicking biological evolution can be used to develop a controlling program for a robot picking up soda cans on a 10x10 grid. Robby the robot can only see the squares adjacent to its location (center, North, South, East, West), and each turn it can either move one step to a particular direction, move at random, try to pick up a can, or do nothing. Each simulation run lasts for a predetermined amount of time steps (originally 200), and Robby's task is to pick up as many randomly situated cans as possible.


Genome G:
$[1,0,3,2,0,2,0,0,3,6,6,6,6,2,2,6,6,6,0,2,5,7,3,8,5,10,8$, $3,0,3,3,2,10,3,0,3,6,6,6,2,2,2,6,6,6,2,8,7,6,4,9,9,6,6$, $2,1,1,2,2,0,7,0,1,6,0,6,6,5,8,4,0,6,7,9,8,9,6,6,2,1,0,1$, $4,1,2,0,2,6,2,7,6,0,6,6,6,6,1,1,3,1,4,8,8,9,4,8,6,9,3,0$, $1,3,2,4,3,2,9,6,6,6,2,2,4,6,1,4,0,7,4,6,3,7,2,2,10,1,4$, $1,5,7,8,9,0,5,6,0,6,2,7,2,7,6,2,10,8,1,1,5,10,0,2,6,2,0$, $3,2,0,7,0,0,9,6,6,4,6,6,1,6,6,8,8,6,7,8,5,5,8,9,4,3,4,9$, $3,3,6,0,10,10,6,0,9,6,10,4,10,3,9,3,10,4,5,5,7,8,7,1,2$, $6,5,8,5,10,7,6,10,6,0,6,2,2,5,9,2,8,2,6,3,1,5,6,8,6,2]$

Figure 2. Each "locus" in the genome G corresponds to one of the possible immediate environmental states of Robby, and each digit (the allele) to a move in that situation (e.g., ' 0 ' $\rightarrow$ 'move north', '5' $\rightarrow$ 'pick up') (see Mitchell 2009, 137).

Initially a random population of software individuals is generated, each with a "genome" consisting of 243 random numbers. Each locus in the genome guides Robby's behavior in a particular situation (Fig 2). The fitness score of each candidate in the population is calculated by running several simulation trials: crudely, the more cans the robot is able to pick up on average, the higher its fitness. Programs with the highest fitness are then used to form the next generation: they are paired randomly, and the genomes of the two parents are crossed over at a randomly chosen point to create the genomes of new individuals. Finally, for each locus of a descendant's genome, there is a small probability (.005) that a mutation occurs in it. As a result, the new generation is based on the most successful variants among the previous generation, and the process loops back to the fitness-calculation phase. Thus the GA continues searching
for efficient solutions by charting new regions of the search space.

After a few hundred generations, the evolved strategies start to achieve impressive results. As we replicated Mitchell's simulation, we observed that after the $800^{\text {th }}$ generation, the best strategies among evolved Robbys started to have higher fitness scores than a Robby programmed by a human designer (ultimately 480 vs. 440 points). ${ }^{3}$ However, although solutions found with GAs are efficient, their behavior is often hard to understand. The ingenious behavioral strategies that the programs employ cannot be deciphered by simply looking at individual genes or sets of genes. Instead, it is necessary to look holistically at the broad phenotypic behavior of the robot. A nice illustration of this impenetrability of such evolved solutions is the fact that in some cases when a high-fitness Robby is in the same square with a can, it decides not to pick it up, but rather moves away from the square. While this behavior seems prima facie irrational, looking at the total behavioral profile of the robot uncovers a clever strategy: Robby uses cans as markers to remember that there are other cans on its side, and it explores the adjacent squares for extra cans before picking up the marker can. Thus by not treating cans only as targets but also as navigational tools, Robby uses its environment to extend its severely limited visual capacities and to compensate for its total lack of memory.

Moreover, by examining the behavior of a highly efficient $1500^{\text {th }}$ generation Robby, it can be seen that this marker strategy manifests in slightly different ways in different environmental situations. It is not a discrete adaptation, but rather a collection of independently evolved sub-strategies. Furthermore, the marker strategy is tightly intertwined with another environment-employing "hack" that the sophisticated Robby uses: when there is already a lot of empty space on the grid, Robby employs a "vacuumcleaner" movement strategy. It follows the walls of the board, departing toward the center when it detects a can, employs the marker strategy if possible, and immediately after cleaning up its local environment, returns directly to the south wall to continue its round around the board. This strategy also includes an ingenious "bounce" feature: when Robby arrives to the corner preceding the wall that is parallel to its default navigation direction in an empty field, it bounces off the wall to increase the range of this search pattern.

Such "kluges" are common to designs created by GAs. Like biological evolution, GAs can come up with solutions that a human designer would not think of. These solutions often offload parts of problem solving to the environment, and thus rely on a tight coupling between the system and its environment. And, as pointed out by Clark (1997, 2001), recurrent circuitry and complex feedback loops between different levels of processing often feature in systems designed by GAs. Such designs are often difficult to understand.

[^129]We suggest that these difficulties in understanding are often created by the lack of modularity in the functional decomposition of the behavior. The high-fitness Robby (genome G in Fig. 1) mentioned in the paragraph above only leaves cans as markers in some specific situations, and only the totality of this selective marking strategy - together with navigational strategies utilizing cans and walls - constitutes the effectiveness of the can-search procedure. Looking at isolated genes in Robby's genome only reveals trivially modular elements corresponding to elementary subtasks in its behavior: one gene corresponds to an elementary move in a specific environmental situation. But we cannot make inferences from local hypothetical changes in these elemental behaviors to consequent effects on fitness. The connection between any single elementary behavioral rule and the strategy is simply too complex and context dependent. A change in a single rule (in situation B; a can present; whether to pick up or not to pick up the can) has consequences for the effectiveness of the other elementary behavioral rules. Explanatorily relevant inferences would require an extra "level" of modular sub-operations between the individual movements and the strategy as a whole.

The marker and vacuum-cleaner strategies mentioned above appear to be examples of such middle-level suboperations, but by themselves they are insufficient to yield understanding of the whole behavior of our most successful Robby. This is because the effectiveness of leaving a can is a result of the evolved coupling between the specific situations in which Robby leaves a can and the rest of the navigation behavior. Therefore, there is no way of independently altering these middle-level strategies. Also "the bounce" is intertwined with the rest of the vacuuming navigation and cannot be independently altered. In general, genetic algorithms do not often produce easily discernible designs. Rather, the interesting heuristics in the system's behavior can only be revealed by simultaneously looking at constellations of different genes, and eventually, the whole genome.

To recapitulate, our example exhibits several distinct (yet related) challenges to understanding:
(1) The discernible middle-level strategies (marker, vacuum-cleaner) do not have dedicated structural bases. Instead, the nature of the design process leaves all atomic structural elements (the 243 DNA elements) open for exploitation by all capacities serving the main goal. Consequentially, the system is neither structurally nor behaviorally nearly-decomposable, but instead has a "flat hierarchy," and strategies are implemented in highly distributed structures.
(2) Challenge 1 above means that the interactions between subtasks tend to be strong: a change in one subtask constituting a part of the marker-behavior also affects the functioning of the vacuum-cleaner navigation. In general, the middle-level strategies can only be discerned and defined in a very abstract way, and the interaction-effect on their contribution to the overall fitness is so large as to make
any inferences about the consequences of partial changes in one strategy next to impossible.
(3) The way in which the strategies contribute to the fitness of the individual is highly context-dependent and depends on the properties of the environment as well as the DNA of the agent. Even small modifications to the environment can lead to drastic changes in the performance of a strategy. For instance, we observed that adding only a few randomly placed extra walls on the grid radically collapses the average score of the successful Robby described above.

Extrapolating from this very simple case, we contend that GAs may design behavior that cannot be tidily decomposed into hierarchical and modular subtasks, whose individual contributions would be easy to understand (i.e., we could infer how a change in a sub-routine would affect the behavior of the mother-task). Instead, feedback, many tasks using the same subtasks as resources, and tight systemenvironment coupling lead to holistic design where almost "everything is relevant for everything." The evolved functional architecture is flat in that there are few discernible levels of order between the elementary operations and the complex behavior. The counterintuitiveness of such flat architectures is apparent in the deep mistrust faced by connectionist suggestions for nonhierarchical design of cognitive capacities (see e.g., Rumelhart and McClelland 1986 vs. Pinker and Prince 1988).

Furthermore, GAs underscore the path dependence of evolutionary problem solving. For sufficiently complex computational problems there are often several local maxima in the fitness landscape of the problem, and the population can converge to different maxima in different runs of the simulation. The functional decomposition that a human designer comes up with is just one possible solution among several others. Perhaps our biological evolution actually ended up with a radically different one.

## Lessons for the Study of Mind

Genetic algorithms demonstrate that evolution can, in principle, lead to non-modular functionality. This imposes a limit on our ability to understand such behavior: if we cannot trace the consequences of changes in the suboperations, we cannot answer what-if questions concerning the complex behavior. Such behavior constitutes a thorny problem for mechanistic understanding of the implementation of the said behavioral capacities, since the DL heuristic cannot get off the ground: we do not even know what we are supposed to localize. We can now ask two questions: should we expect to find such non-modular functionality in nature, especially in human cognition, and if so, what attitude should we adopt with respect to this problem. Should the aim of causal-mechanistic understanding of the brain be given up, and be replaced, for example, with non-mechanistic dynamical models often
employing a limited set of instrumentally interpreted macrovariables?
There are important disanalogies between GAs and biological evolution. As is the case with Robby, there is often no genotype-phenotype distinction. In biological evolution, however, genes do not directly cause properties of the phenotype, but rather participate in guiding ontogenesis. It has been suggested that ontogenesis itself favors modular design. GAs may also seem a problematic platform for exploring the possibilities of DL heuristics, since the lowest level of functional organization and the level of implementation are identical (i.e., the genome). However, we see no reasons why this would affect our argument. Moreover, the argument developed here is not only about genetic selection, but about selection in general, and failures of functional modularity may in principle also arise in the course of development - at least if the idea of neuronal group selection or "neural Darwinism" is taken seriously.
The recent research on biological control networks (metabolic and gene regulatory networks) suggests that evolved modular organization is in fact the rule rather than the exception: control networks exhibit network modularity and the recurring modules (motifs) have easily discernible modular functions. Therefore, the question in the recent years has rather been to formulate an evolutionary explanation for this modular design. Genetic algorithms have been used to argue that modularity is not selected for, but that it is instead a byproduct of specialization of gene activity (Espinasa-Soto \& Wagner 2010) or of selection against densely connected networks and long connections (Clune, Mouret \& Lipson 2013).

Most interesting for our case, Kashtan and Alon (2005; see also Kashtan et al. 2007) have demonstrated that when the goals themselves are composed of modularly varying sub-goals, evolution tends to produce modular functionality. It seems easy to see why this is the case. If the tasks to which the system has to adapt remain the same, the selection environment does not change, and the peaks in the fitness landscape are stable, then selection favors strategies that offload problem solving to that particular environment as much as possible. But if the task itself is composed of changing subtasks, it makes sense to design the adaptive response in such a way that a particular sub-operation can locally adapt to a local change in a subtask without altering the totality of the otherwise well-functioning behavior.
In their research, Kashtan and Alon evolved several network models to compute complex Boolean functions, with fitness calculated according to how close the network output was to the target. They found that by modularly varying goals, it is often possible to considerably speed up the evolution. In our Robby simulation, we studied the effects of changing environment for the evolution of modularity by allowing the environment to change discretely from an initial no-walls (torus) condition to one with walls, and eventually to one with also random obstacles. Our results suggest that although "modularity in
tasks" does speed up learning, it can often prematurely weed out diversity in the population in such a way that, in the end, the global maximum for the main target task cannot be reached.

It seems likely that our cognition has evolved in at least partly modularly changing selection environment, but the extent to which we should expect to find modular functionality in human cognition is hard to estimate. We suspect that the usefulness of many of the existing computational models investigating the evolution of biological modularity is constrained by the fact that the tasks (e.g. simple categorization, logic circuits) solved by the algorithms are straightforwardly computational and do not really involve any interesting behavioral aspects. This is why the Robby platform has certain advantages for exploring evolved functionality: The dynamic nature of the simulation allows the "emergence" of novel and irreducibly top-level strategies in a way that is lacking in the more static contexts.

Because of the uncertainty related to the actual extent of non-modularity in human cognition, we stress the conditional nature of our argument. Our study of genetic algorithms and our analysis of the properties of the resulting designs only demonstrates that evolution can create designs, which are in principle beyond the understanding of unaided cognitive beings such as us.

Yet there is nothing mysterious in such designs. Simon pondered whether the apparent abundance of hierarchical, nearly decomposable complexity was due to our selective attention to precisely such systems, but we believe this to be a somewhat hasty conjecture. We have no trouble finding and delineating systems, such as Robby, or possibly ourselves, that manifest functionally non-decomposable behaviors sustained by a flat architecture. However, there certainly might be a psychological bias that makes us see hierarchical design also where there is none. One way of coping with this obstacle to understanding is to realize that there are no fundamental reasons to limit the relevant epistemic agent to be an unaided human. Although only a human agent can experience a sense of understanding, this feeling should not be confused with understanding itself. Therefore brute computational approaches can produce understanding as long as the epistemic subject, the cognitive unit whose inferential abilities are to be evaluated, is conceived as the human-computer pair.

## References

Bechtel, W., \& Richardson, R. (2010). Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research. Cambridge MA: The MIT Press.
Churchland, P. M. (1989). A Neurocomputational Perspective: The Nature of Mind and the Structure of Science. Cambridge, MA: MIT Press.

Clune, J., Mouret, J-B., \& Lipson, H. (2013). The evolutionary origins of modularity. Proceedings of the Royal Society B, 280.
Clark, A. (1997). Being There: Putting Brain, Body and World Together Again. Cambridge MA: The MIT Press.
Clark, A. (2001). Mindware: An Introduction to the Philosophy of Cognitive Science. Oxford: Oxford UP.
Cummins, R. (1983). The Nature of Psychological Explanation. Cambridge MA: MIT Press.
Espinosa-Soto C, Wagner A. (2010). Specialization Can Drive the Evolution of Modularity. PLoS Computational Biology, 6(3).
Holland, J. (1975). Adaptation in Natural and Artificial Systems. Ann Arbor: University of Michigan Press.
Jacob, F. (1977). Evolution and Tinkering. Science 196 (4295): 1161-1166.

Kashtan, N., \& Alon, U. (2005). Spontaneous evolution of modularity and network motifs. PNAS 102 (39): 1377313778.

Kashtan, N., Noor, E. \& Alon, U. (2007). Varying environments can speed up evolution. PNAS 104 (34): 13711-13716.
Levins, R. (1973). The Limits of Complexity. in Pattee, H. (Ed.), Hierarchy Theory: The Challenge of Complex Systems. London: Braziller.
Marcus, G. (2008). Kluge: The Haphazard Construction of the Human Mind. Boston and New York: Houghton Mifflin.
Mitchell, M. (1996). An Introduction to Genetic Algorithms. Cambridge MA: MIT Press.
Mitchell, M. (2009). Complexity. A guided tour. Oxford: Oxford University Press.
Piccinini, G., \& Craver, C. (2011). Integrating Psychology and Neuroscience: Functional Analyses as Mechanism Sketches. Synthese 183 (3): 283-311.
Pinker, S., \& Prince, A. (1988). On Language and Connectionism: Analysis of a Parallel Distributed Processing Model of Language Acquisition. Cognition 23: 73-193.
Rumelhart, D., \& McClelland, J. (1986). On Learning the Past Tenses of English Verbs. In McClelland \& Rumelhart et al. (Eds.), Parallel Distributed Processing vol. I. Cambridge MA: MIT Press.
Simon, H. (1962). The Architecture of Complexity. Proceedings of the American Philosophical Society 106: 476-482.
Thagard, P. (2012). The Cognitive Science of Science. Cambridge MA: MIT Press.
Waskan, J. (2006). Models and cognition: Prediction and explanation in everyday life and in science. Cambridge, MA: The MIT Press.
Wimsatt, W. (2007). Re-Engineering Philosophy for Limited Beings. Cambridge MA: Harvard UP.
Woodward, J. 2003. Making Things Happen. Oxford UP.
Ylikoski, P., \& Kuorikoski, J. (2010). Dissecting Explanatory Power. Philosophical Studies 148, 201-219.

# Incremental processing in the pragmatic interpretation of contrastive prosody 

Chigusa Kurumada<br>kurumada@stanford.edu<br>Dpt. of Linguistics<br>Stanford University

Meredith Brown<br>mbrown@bcs.rochester.edu<br>Dpt. of Brain and Cognitive Sciences<br>University of Rochester

Sarah Bibyk<br>sbibyk@bcs.rochester.edu<br>Dpt. of Brain and Cognitive Sciences<br>University of Rochester

Daniel F. Pontillo<br>dpontillo@bcs.rochester.edu<br>Dpt. of Brain and Cognitive Sciences<br>University of Rochester

Michael K. Tanenhaus<br>mtan@bcs.rochester.edu<br>Dpt. of Brain and Cognitive Sciences<br>University of Rochester


#### Abstract

We present an eye-tracking experiment investigating the time course with which listeners derive pragmatic inferences from contextual information. We used as a test case the construction "It looks like an X" pronounced either with (a) a nuclear pitch accent on the final noun, or (b) a contrastive $\mathrm{L}+\mathrm{H}^{*}$ pitch accent and a rising boundary tone, a contour that can support a complex contrastive inference (e.g., It LOOKS like a zebra...(but it is not)). The contrastive intonational contour elicited higher proportions of fixations to non-prototypical target pictures (e.g., a zebra-like animal) during the earliest moments of processing the target noun. Further, when the display only contained a single related pair of pictures, effects of the contrastive accent on "looks" emerged prior to the target noun, indicating that efficient referential resolution is supported by rapidly generated inferences based on visual and prosodic context.


Keywords: Prosody, contrastive accent, pragmatic inferences, eye-tracking.

## Introduction

Few, if any, would question the claim that addressees must make use of context to infer the intentions of a speaker (speaker meaning). Herb Clark (1992) gives a lovely example to illustrate the richness of context-based inferences. Clark describes a situation in which he addressed the utterance, "I'm hot", to his school-age son, Damon. After going through the plausible pre-compiled senses, Clark notes that none captures his intended (and immediately understood) meaning of his utterance, which could only be inferred from the specific context. Herb and Damon were playing poker and Damon was about to make a large bet. Herb was warning Damon that he should think twice about it.

Despite countless everyday examples of this sort, there is also a widely held view that pragmatic inference is external to the core mechanism of language comprehension. For example, this assumption underlies Levinson's (2000) influential proposal that common inferences might be pre-compiled as automatically generated defaults, by-passing the need for making a slow and resource intensive inferences (e.g., Neely, 1977; Posner \& Snyder, 1975; Shiffrin \& Schneider, 1977). This idea receives support from the hypothesis that the remarkable speed and ease of real-time language processing is possible, in part, because of its modularity in the processing system. A syntactic module, for example, performs computations on restricted inputs without appealing to slow


Figure 1: A sample visual display used in Sedivy et al. (1999) for an instruction "Pick up the tall glass"
and resource-demanding processes, such as inference (e.g., Fodor, 1983).

This modularity hypothesis, however, lacks an explanation for cases in which expectations based on context can effectively constrain parsing decisions. In fact, there is now a large body of research demonstrating that listeners rapidly use information from the linguistic and visual context to resolve ambiguity (e.g., Altmann 1998; Chambers, Tanenhaus \& Magnuson, 2004; Snedeker \& Trueswell, 2003). In this constraint-based approach, the context of language use is integral to effective and incremental language processing in guiding expectations (e.g., Levy, 2008). Furthermore, Piantadosi, Tily and Gibson (2012) propose that inherent ambiguity in the linguistic signal is in fact a design feature of an efficient encoding system, given the assumption that listeners can integrate context information to inferentially resolve ambiguity.

Consistent with these accounts, a number of studies using online measures have shown that listeners can, and do, incorporate visual information to process linguistic input incrementally. For example, Sedivy et al. (1999) examined listeners' processing of prenominal adjectives during incremental language processing. They asked participants to manipulate objects based on spoken instructions such as "Pick up the tall glass". In Figure 1, the pitcher on the lower left is the tallest object, but the glass on the upper left is both tall by comparison to glasses in general, and taller than the other glass in the upper right-hand corner. Sedivy et al. found that the partial instruction "Pick up the tall --" elicited fixations to the tall member of the contrast pair (e.g., the tall glass) rather than
the other tall object (e.g., the pitcher) in the display. This suggests that listeners rapidly integrate context-specific contrast information to begin resolving referential ambiguity prior to the head noun.

Nonetheless, it remains to be understood how readily listeners can derive more complex inferences such as conversational implicature. For example, some experimental studies on the English quantifier some (but not all) have concluded that even the basic scalar implicature is indeed slow and costly, compared with computing its logical meanings (i.e., at least one, possibly all) (e.g., Bott \& Novek, 2004; Huang \& Snedeker, 2009). On the other hand, there is a recent body of work (e.g., Grodner et al., 2010, Degen \& Tanenhaus, under review) suggesting that delays arise only when use of some in the particular context is less natural than another rapidly available alternative.

In this current study, we approach this problem by examining the time course of English speakers' comprehension of contrastive prosody. In English, the pitch accent $\mathrm{L}+\mathrm{H}^{*}$ is known to evoke an alternative set of referents and invites a contrastive inference (e.g., Katie did not win a $T R U C K_{L+H *}$ (but won a motorcycle), Ito \& Speer, 2008). Previous work has found that the use of $\mathrm{L}+\mathrm{H}^{*}$ in an appropriate discourse context restricts the domain of reference during incremental language comprehension. For instance, the $\mathrm{L}+\mathrm{H}^{*}$ in "Give me the red ball. Now give me the $\operatorname{GREEN}_{L+H *}$-" triggers anticipatory eye-movements to a green object of the same type as the preceding referent (i.e., a green ball).

While this contrast-evoking function of $\mathrm{L}+\mathrm{H}^{*}$ is known to be robust (Weber et al., 2006), previous experimental work has almost exclusively focused on prenominal adjectives highlighting color or size contrast. Moreover, studies so far have found incremental processing of contrastive prosody only when a member of the relevant contrast set was linguistically mentioned in prior discourse. These limitations make it difficult to scale up previous findings to cases where contrastive accent triggers complex, and hence allegedly costly, conversational implicatures.

To address this, we used a different linguistic construction, "It looks like an X", which can support two opposing pragmatic interpretations depending on its prosodic realization. A canonical declarative prosodic contour, with a nuclear pitch accent on the final noun (as illustrated in Figure 2, left panel, henceforth Noun-focus prosody), typically elicits an affirmative interpretation (e.g., It looks like a zebra and I think it is one). When the verb "looks" is lengthened and emphasized with a contrastive $\mathrm{L}+\mathrm{H}^{*}$ accent and the utterance ends with a rising L-H\% boundary tone (Figure 2, right, Verb-focus prosody), it can trigger a negative interpretation (e.g., It LOOKS like a zebra but its actually not one (Kurumada, Brown, \& Tanenhaus, 2012).

In the current study, we tested if and how the listeners develop the two different interpretations as they receive prosodic information. Specifically, we asked the following questions:


Figure 2: Examples of Noun-focus prosody (left) and Verbfocus prosody (right).

1. Can listeners integrate visually represented contrasts with prosodic information to guide pragmatic interpretation?
2. Do listeners process intonational contours and develop pragmatic expectations incrementally?

## Experiment overview

We examined the time course of pragmatic intonation interpretation using the visual world paradigm (Tanenhaus et al., 1995). Participants listened to the construction "It looks like an X" produced with either Noun-focus or Verb-focus prosody, and they were asked to click on the corresponding referent in a four picture display. In each display, there was at least one pair of visually similar items (e.g., a zebra and an okapi; Figure 3-a, bottom row). We assumed based on previous work that Noun-focus prosody would bias responses toward the more prototypical member of each pair (e.g., a zebra for "It looks like a ZEBRA"), while Verb-focus prosody would bias responses toward the less prototypical member (e.g., an okapi for "It LOOKS like a zebra"). Thus, our first hypothesis is that listeners should integrate the contrasting relation between the prototypical and non-prototypical target pictures in their interpretation of the utterance intonation.

A previous study has shown that listeners can develop a similar contrastive inference in a visual search task (Dennison \& Schafer, 2010). Their study used the construction "X HAD Y" (e.g., "Lisa HAD a bell..." (but she no longer has one)), but they found no evidence of incremental processing. They proposed that the contrastive inference requires both the pitch accent and the boundary tone, and hence occurs only after the sentence offset.

In the current study, we hypothesize that listeners can compute an implicature incrementally, based on the prosodic and visual context. We tested this hypothesis by comparing the time course of eye movements in a display with a single pair of contrasting items, to those in a display with two pairs. In the one-contrast display, we predicted that participants would be able to use the contrastive pitch accent to infer that the likely referent is a member of the contrast pair (more specifically, the less prototypical member) prior to the processing of the target word. In the two-contrast display, the target referent cannot be determined until it has been explicitly mentioned, which should result in effects of prosody emerging later, i.e., during the processing of the target word.


Figure 3: Sample visual displays

## Methods

## Participants

Twenty-four students from University of Rochester were paid (\$10) to take part in the experiment. They were native speakers of American English with normal or corrected-to-normal vision and normal hearing.

## Stimuli

We selected 16 imageable high-frequency nouns and embedded them in the sentence frame "It looks like an X". A native speaker of American English recorded two tokens of each item with the Noun-focus and the Verb-focus prosodic patterns. The same speaker also recorded 44 filler items in which a target referent was described (e.g., "Can you find the one with white fur?").

We selected 16 more items to form pairs with the 16 target nouns. In each pair, the items were visually similar to each other (e.g., a zebra and an okapi) and one item (e.g., a zebra) was always more common. Hereafter, the picture from each pair that is more common (e.g., a zebra, Figure 3, bottom left) is referred to as the prototypical target picture, and the other (e.g., an okapi, Figure 3, bottom left) is referred to as the nonprototypical target picture.
Prototypicality and nameability norming To create visual stimuli, we ran two types of norming studies using Amazon Mechanical Turk, an online crowd-sourcing platform. In the first study, 40 subjects provided names and nameability ratings (on a seven-point rating scale) for each of the $240 \mathrm{im}-$ ages. In a second norming study, we presented 40 subjects with the images along with a label and collected ratings of referential fit for both adult-directed speech and, as a separate response, child-directed speech. The non-prototypical pictures (e.g., okapi) were always presented with the names of their respective prototypical items (e.g., zebra) in order to establish an empirical measure of prototypicality.

Based on this information we constructed 60 visual scenes ( 16 critical trials and 44 filler trials). Each of the scenes consisted of four pictures including one pair of target pictures described in an auditory stimulus. We created two types of visual scenes: a) 1 target pair +2 singletons ( 1 -contrast condition), and b) 1 target pair and 1 distractor pair ( 2 -contrast
condition) (Figure 3). Singletons in 1-contrast trials consisted of one easily nameable picture and one less-nameable picture to equate the complexity of the visual display across trials.

## Procedure

Participants were first presented with a cover story in which a mother and a child are looking at a picture book. The mother is helping the child to identify various objects and animals by verbally commenting on them. Each trial began with the presentation of a visual display containing four pictures. After 1 second of display preview, participants heard a spoken sentence over Sennheiser HD 570 headphones and clicked on a picture described by the sentence. Their mouseclickng responses were collected while their eye movements were tracked using a head-mounted SR Research EyeLink II system sampling at 250 Hz , with drift correction procedures performed after every fifth trial.

Eight lists were constructed by crossing the 1 ) item presentation order, 2) the location of the prototypical and the nonprototypical items on the display, and 3) the prosodic contour (Noun-focus vs. Verb-focus). All lists started with three example items to familiarize participants with the task. The mean duration of the experiment was 12 minutes.

## Results and discussion

We analyzed three dependent measures to obtain converging evidence about the role of prosody and visual display characteristics in the processing of critical items: response choices in the picture selection task, response times, and proportions of fixations to different alternatives within the display. Each variable was assessed with multi-level generalized linear regression models implemented using the lmer function within the lme4 package in R (R Development Core Team, 2010; Bates et al. 2008) ${ }^{1}$.

We first confirmed that participants selected a correct target picture in $96 \%$ of filler trials, indicating that participants did not have difficulty completing or attending to the picture selection task. We then analyzed their responses in the 16 critical trials to ask if they encoded the visual contrasts of items on the screen and associated them with the two prosodic contours. Participants selected the prototypical target picture $65.6 \%$ of the time in the Noun-focus prosody condition, but only $25.5 \%$ of the time in the Verb-focus prosody condition. A multilevel logistic regression model of responses confirmed that depending on the prosodic contour, participants reliably chose either a prototypical or a non-prototypical item

[^130]

Figure 4: Proportions of fixation to pictures in response to the Noun-focus (solid line) and to the Verb-focus (dashed line). The lines are aligned at the onset of the final noun.
( $\beta=6.37, z=4.394, p<.0001$ ). Thus, without any explicit mention of a contrasted item, participants encoded a relevant contrast set in the visual field and developed a contrastive inference based on the Verb-focus prosody.

Response times indicated that Verb-focus prosody elicited slower responses (mean RT=2204 ms) than Noun-focus prosody (mean $\mathrm{RT}=1969 \mathrm{~ms}, \beta=-.242, t=-2.09, p<.05$ ). However, the effect of prosody was dependent on whether the prototypical or non-prototypical target picture was selected ( $\beta=.509, t=2.94, p<.005$ ). On trials with Nounfocus prosody, RTs were significantly faster when a prototypical picture was selected (mean $\mathrm{RT}=1762 \mathrm{~ms}$ ) than when a non-prototypical picture was chosen (mean $\mathrm{RT}=2364 \mathrm{~ms}$, $\beta=-.272, t=-3.20, p<.005$ ). On trials with Verbfocus prosody, however, there was a numerical trend in the opposite direction (mean RT=2540 ms for prototypical target responses vs. 2089 ms for non-prototypical target responses, $(\beta=.201, t=1.10, p>.10)$. This finding suggests that responses deviating from the expected association between prosody and picture type were associated with greater processing difficulty, further supporting the hypothesis that participants were interpreting the prosodic contour based on
the visual contrasts.
Next we analysed the eye-tracking data to examine the time course of processing the contrastive prosody. Our analysis focused on two regions, which were both defined with respect to the point at which segmental information from the target word would be expected to influence processing. The first region, which we termed the pre-target region, was defined as the region beginning at 200 ms after the offset of "looks" and ending at 200 ms after the onset of the target word. This roughly corresponds to the region indicated with the caption "like a" in Figure 4, shifted to the right by 200 ms . Because it takes approximately 200 ms to program and execute an eye movement, fixations within this window should not be influenced by segmental information from the target word. The only information that would be expected to influence eye movements within this window, then, is information from preceding prosody (e.g. the contrastive accent on "looks").

The second region, the early target-word region, began at 200 ms after target word onset and ended at 200 ms after the offset of the first syllable of the target word. This roughly corresponds to the region indicated with the caption "ze-" in Figure 4 , shifted to the right by 200 ms . Fixations within this window were expected to reflect the integration of expectations based on preceding prosody and initial effects of incrementally presented target word information. Within each window, mean proportions of fixations to each picture were calculated and then transformed using the empirical logit function (Cox, 1970) for the purposes of linear regression analysis.

Pre-target fixations We analyzed logit-transformed proportions of fixations averaged across the pre-target region in two linear mixed-effects regression models. The first model examined effects of prosody contour, display type (i.e., onevs. two contrast sets), and trial number on logit-transformed mean proportions of fixations to the distractor pictures vs. either member of the target contrast set (e.g. the zebra and okapi). The goal of this analysis was to assess prosody- and display-wise differences in anticipating the target contrast set. We predicted that Verb-focus prosody would bias participants to fixate members of the target contrast set in one-contrast trials but not in two-contrast trials.

Results from the regression analysis revealed that the predicted three-way interaction between prosody condition, picture type, and display type was significant ( $\beta=.754, t=1.98$, $p<.05$ ). Analyzing proportions of fixations by display type revealed that effects of prosody condition were dependent on picture type in one-contrast trials $(\beta=-.322, t=-2.61$, $p<.01$ ). Participants were no more likely in the Nounfocus condition to fixate the target picture (mean untransformed proportion of fixations=.209) and the distractor pictures (mean=.186, $\beta=.007, t=.068, p>.1$ ), but they exhibited a significant bias toward the target contrast set in response to Verb-focus prosody (mean=. 245 vs. . $167, \beta=$ $-.315, t=-3.34, p<.001$ ). This interaction was not sig-
nificant in two-contrast trials $(\beta=-.058, t=-.529, p>.1)$.
The second linear mixed-effects regression model examined effects of prosody condition, display type, and trial number on logit-transformed mean fixation proportions to prototypical vs. non-prototypical target pictures. We predicted that the difference between fixations to non-prototypical pictures and fixations to prototypical pictures would be greater in response to Verb-focus prosody than Noun-focus prosody. Indeed, the regression model revealed a marginal two-way interaction between prosody condition and picture type ( $\beta=$ $.138, t=1.71, p=.087$ ). In the Noun-focus prosody condition, fixations to bad target pictures (mean=.238) did not differ significantly from fixations to good target pictures (mean=.181, $\beta=-.124, t=-1.20, p>.1$ ). In the Verb-focus condition, however, participants were significantly more likely to fixate the bad target picture (mean=. 300 vs. .190, $\beta=-.243, t=-2.69, p<.01)$.

Taken together, these findings suggest that participants rapidly encode the visual attributes of and relations between potential referents, and rapidly integrate this visual information with the incoming prosodic input. When a single contrast set is present in the display, contrastive Verb-focus prosody biases listeners to fixate members of that set. This trend is illustrated in Figure 4. In the 1-contrast condition (Figure 4-a), the fixation proportions to the non-prototypical target based on the Verb-focus prosody begins to diverge in the pre-target region. On the other hand, such divergence is delayed in the 2-contrast condition (Figure 4-b).
Early target-word fixations For the early target-word region, we again analyzed logit-transformed mean proportions of fixations in two linear mixed-effects regression models, to compare effects of prosody condition, display type, and trial number on (a) target-picture fixations to distractor fixations, and (b) prototypical target-picture fixations to nonprototypical target-picture fixations.

For the analysis comparing logit-transformed mean proportions of fixations to target pictures vs. distractor pictures, we predicted that the three-way interaction between prosody condition, picture type, and display type would no longer be significant. Instead, the main prediction was that fixations to both target pictures would be significantly higher than fixations to distractors across trial types.

The results of the analysis indicated that neither display type nor prosody condition accounted for a significant proportion of variance in target vs. distractor fixations in the early target-word region. Instead, the main finding was that participants were significantly more likely to fixate target pictures (mean $=.280$ ) than distractor pictures (mean $=.127, \beta=-.240$, $t=-4.12, p<.0001$ ), reflecting their early use of incoming segmental information to restrict the referential domain to the two target pictures.

The analysis of fixations to the two target pictures was predicted to show that the difference between non-prototypical target picture fixations and prototypical target picture fixations would continue to be greater in the Verb-focus condi-


Figure 5: Mean fixation proportions to the non-prototypical item in response to the Verb-focus prosody. Error bars represent standard errors.
tion. In addition, display type was no longer predicted to significantly influence patterns of fixations, since the segmental information from the initial sounds of the target word should restrict the domain of reference to the target contrast set.

The second linear mixed-effects regression model indeed revealed a two-way interaction between prosody condition and picture type ( $\beta=-.472, t=-4.02, p<.0001$ ). In the Verb-focus condition, participants were significantly more likely to fixate non-prototypical target pictures (mean=.374) than prototypical target pictures (mean $=.222, \beta=-.335$, $t=-3.97, p<.0001$ ). In the Noun-focus condition, however, there was a non-significant trend in the opposite direction, with more fixations to prototypical target pictures (mean $=.293$ ) than non-prototypical target pictures (mean $=.231, \beta=.138, t=1.49, p>.1$ ). This interaction between prosody condition and picture type suggests that listeners rapidly integrated incoming segmental information from the target word with their pragmatic expectations for a prototypical vs. non-prototypical referent based on preceding prosodic information.

Figure 5 summarizes mean fixation proportions to the nonprotptypical item based on the Verb-focus prosody. Within the pre-target region, participants were looking at the nonprototypical item more when they were in the 1-contrast condition than in the 2 -contrast condition. This trend was even more magnified when the segmentatal information of the target noun becomes available. This demonstrates that the contrastive pitch accent was processed incrementally under the constraints of the visual context.

## Conclusion

The results show that participants generated complex pragmatic interpretations in an incremental manner. In a context with only one contrast pair, listeners began to launch eye movements to a less prototypical target picture even before segmental cues to the final noun become available. This is of particular interest because, unlike in previous studies, the contrastive accent in the current study was used with the verb.

The contrast was not simply based on individual visual features of objects (e.g., color or size); rather, it was mediated by the implicature based on different predicates. Namely, "It LOOKS like an X" is contrasted with "It IS an X", and therefore interpreted as "It is not an X". Our results demonstrate that such complex pragmatic reasoning can develop online.

The results also highlighted the facilitative roles of visual context in intonation interpretation. The early timing of the prosody effect in the 1-contrast condition suggests that listeners made use of visually represented contrast to guide their inference. This enabled us to demonstrate that inferences based on contrastive prosody do not require explicit previous mention of a contrasting item and can be made incrementally on the basis of partial prosodic contour as well as visual information. These findings together advance our knowledge about the remarkably rapid and robust inferential mechanisms supporting online language comprehension and pragmatic communication.

## Acknowledgements

Thanks to Eve V. Clark, Christine Gunlogson and T. Florian Jaeger for valuable discussion, and to Dana Subik for support with participant testing. This research was supported by the NICHD grant HD27206 (MKT).

## References

Altmann, G.T.M. (1998). Ambiguity in Sentence Processing. Trends in Cognitive Sciences, 2, 146-152.
Baayen, R., Davidson, D., \& Bates. D., (2008). Mixedeffects modeling with crossed random effects for subjects and items. JML, 59, 390-412.
Bates, D.M., Maechler, M., \& Dai, B. (2008). lme4: Linear mixed-effects models using S 4 classes.
Bott, L., \& Noveck, I. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. JML, 51, 437-457.
Chambers, C. G., Tanenhaus, M. K., \& Magnuson, J. S. (2004). Actions and affordances in syntactic ambiguity resolution. JEP:LMC, 30, 687-696.
Clark, H. H. (1992). Arenas of language use. Chicago: University of Chicago Press.
Cox, D.R. (1970). The analysis of binary data. London: Methuen.
Degen, J., \& Tanenhaus, M. K. (under review). Processing scalar implicature: A constraint-based approach.
Dennison, H. Y., \& Schafer, A. (2010). Online construction of implicature through contrastive prosody. Speech prosody 2010 conference.
Fodor, J. A. (1983).The modularity of mind: An essay on faculty psychology. MIT Press.
Grodner, D. J., Klein, N. M., Carbary, K. M., \& Tanenhaus, M. K. (2010). Some, and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. Cognition, 116, 42-55.

Huang, Y. T., \& Snedeker, J. (2009). Online interpretation of scalar quantifiers: insight into the semantics-pragmatics interface. Cognitive Psychology, 58, 376-415.
Ito, K., \& Speer, S. R. (2008). Anticipatory effects of intonation: Eye movements during instructed visual search. JML, 58, 541-573.
Kurumada, C., Brown, M., \& Tanenhaus, M. K. (2012). Prosody and pragmatic inference: It looks like speech adaptation. Proceedings of the 34th Annual Conference of the Cognitive Science Society.
Levinson, S. C. (2000). Presumptive meanings: The theory of generalized conversational implicature. MIT Press.
Levy, R. (2008). Expectation-based syntactic comprehension. Cognition, 106, 1126-1177.
Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. JEP: G., 106, 226-254.
Piantadosi, S. T., Tily, H., \& Gibson, E. (2012). The communicative function of ambiguity in language. Cognition, 122, 280-91.
Pierrehumbert, J. \& Hirschberg, J. (1990) The meaning of intonational contours in the interpretation of discourse. In P. Cohen, J. Morgan, and M. Pollack (eds), Intentions and plans in communication and ciscourse (pp. 271-311). MIT Press.
Posner, M. I., \& Snyder, C. R. (1975). Facilitation and inhibition in the processing of signals. In P. M. A. Rabbitt \& S. Dornic (Eds.), Attention and performance (pp. 669-682). New York: Academic Press.
R Development Core Team. (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
Tanenhaus, M. K., Spivey-Knowlton, M., Eberhard, K., \& Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. Science, 268, 1632-1634.
Tanenhaus, M. K., Magnuson, J. S., Dahan, D., \& Chambers, C. G. (2000). Eye movements and lexical access in spoken language comprehension: Evaluating a linking hypothesis between fixations and linguistic processing. Journal of Psycholinguistic Research, 29, 557-580.
Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., \& Carlson, G. N. Achieving incremental semantic interpretation through contextual representation. Cognition, 71, 109147.

Shiffrin, R. M., \& Schneider, W. (1977). Controlled and automatic human information processing. Psychological Review, 84, 127-190.
Snedeker, J., \& Trueswell, J. (2003). Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. JML, 48, 103-130.
Weber, A., Braun, B., \& Crocker, M. W. (2006). Finding referents in time: Eye-tracking evidence for the role of contrastive accents. Language and Speech, 49, 367-392.

# Contextual inferences over speakers' pragmatic intentions: Preschoolers' comprehension of contrastive prosody 

Chigusa Kurumada<br>kurumada@stanford.edu<br>Department of Linguistics<br>Stanford University


#### Abstract

We investigate pre-schoolers' ability in drawing pragmatic inferences based on prosodic information. Previous work has found that young children are generally oblivious to intonational meaning of utterances. In particular, the ability to comprehend contrastive prosody develops late during language acquisition (after the age of 6). In three experiments, we show that preschoolers can engage in prosody-based pragmatic inferences if the context provides supports for them. Furthermore, we find that preschoolers' interpretation of prosody involves complex counter-factual reasoning ('what the speaker would have said if she had intended another meaning'). The picture emerging from our studies contrasts with previous work: Through rich contextual inferences, four-year olds are able to bootstrap their interpretation of prosodic information, and achieve adult like performance in intonation interpretation.


Keywords: Prosody, language acquisition, contrastive accent, Principle of Contrast, rational inference

## Introduction

In learning new words, young children can make use of pragmatic inference to bootstrap their knowledge about new word-object mapping. For example, in a situation where an adult utters "Give me the TOMA (nonce word)" ${ }^{1}$ when a familiar object (e.g., a spoon) and an unfamiliar object (e.g., a whisk) are present in a visual field, a child as young as two years of age is likely to reach for the whisk. This is considered to be based on a cognitive bias for a unique objectlabel mapping (e.g., Markman and Wachtel (1988)) or the inference that the mother should have used the familiar word (spoon) if she had intended to refer to it (e.g., Clark (1990)).

Such pragmatic dispositions provide immense leverage in word learning because there is inherent uncertainty associated with mappings between speakers' intentions, linguistic signals, and their referents (Frank, Goodman, \& Tenenbaum, 2009). One way to systematically solve this puzzle is to estimate the probability assigned to a possible intention-signal mapping relative to other possible mappings warranted by the same context. For instance, the probability of the signal "Give me the TOMA" expressing the speaker's intention of picking out the non-spoon object is estimated in proportion to probabilities of (1) the signal being generated by the intention of picking out a spoon; and (2) other signals (including [a spoon]) being generated by the intention of picking out the target (non-spoon) object.

[^131]

Figure 1: A schematic representation of a pragmatic model of intention-signal mapping

In the model illustrated in Figure 1, a speaker's intention (to express meaning m 1 ) generates a particular linguistic signal (signal s1); If the speaker had meant to express meaning m 2 , she would have generated a different signal (signal s2). A word learner's job is to work backwards from the observed signal to infer the speaker's intention (as indicated with arrows) while updating her belief about the signal-intention mappings, including any associations newly introduced.

The current study extends this idea to a new domain: Young children's interpretation of contrastive prosody. Just as words are associated with speaker intentions, different prosodic representations are probabilistically mapped onto intentions that had generated them (Figure 2). Compared to words, however, prosodic signals are continuous and variable, which can make the mapping puzzle much harder to solve. Furthermore, the intentions that prosodic representations encode are often very abstract (e.g., contrastiveness) and are not always disambiguated in an observable context. In other words, signal-intention mappings for prosody include much more uncertainty (indicated by the thinner arrows in Figure 2 ) for listeners to overcome. The current study suggests that, despite this additional complexity, the rational inferences attested in word-learning provide leverage in young children's discovery of pragmatic functions of contrastive prosody as well.

Prosodic information is known to encode structural boundaries and phrasing but also speakers' intentions (Pierrehumbert \& Hirschberg, 1990; Ladd, 2008; Büring, 2003). Much attention has been paid to how listeners interpret context-relevant contrast based on a low-high-low (an L+H*) pitch accent. (e.g., KATIE $\left(\mathrm{L}+\mathrm{H}^{*}\right)$ did not win a truck (but LAURA did); Ito \& Speer, 2008). Previous work has generally agreed that inferences based on an $\mathrm{L}+\mathrm{H}^{*}$ accent present great difficulty to preschoolers, and even young school children fail to achieve adult-like performance in experimental settings (e.g., Solan, 1980; Cruttenden, 1985; Wells, Peppe,


Figure 2: A schematic representation of a model for intention-signal mapping in prosodic interpretation


Figure 3: Waveforms (top) and pitch contours (bottom) of the utterance "It looks like a zebra". The affirmative interpretation It is a zebra is typically conveyed by the pattern on the left, while the negative interpretation It is not a zebra is conveyed by the pattern on the right.
\& Goulandris, 2004; Sekerina \& Trueswell, 2012). Since even young infants can use prosodic information for finding word boundaries or affective communication (Cutler \& Swinney, 1987), the difficulty is not usually attributed to their sensitivity to prosodic information per se, but to limited cognitive resources and memory span (Speer \& Ito, 2009). The current study shows a supportive discourse context allows children to comprehend contrastive prosody earlier than previously reported.

Furthermore, we demonstrate that acquisition of contrastive prosody is supported by the same pragmatic inference underlying the word-intention mappings described above. That is, upon receiving a signal, the listener works backward to inferentially identify the intention most likely to generate the particular signal observed. Crucially, a learner can bootstrap her knowledge about a new signal-intention mapping based on other mappings warranted in the same context.

In the three experiments reported below, 4-year-olds and adults were asked to interpret an English construction "It looks like an X", which can evoke different pragmatic meanings depending on its prosodic realization. A canonical accent placement (as illustrated in Figure 3, left panel, henceforth noun-focus prosody) typically elicits an affirmative interpretation (e.g. It looks like a zebra and I think it is one). When the verb "looks" is lengthened and emphasized with a contrastive accent $\left(\mathrm{L}+\mathrm{H}^{*}\right)$ and the utterance ends with a L H\% boundary tone (Figure 3, right, verb-focus prosody), it can trigger a negative interpretation (e.g. It LOOKS like a zebra but its actually not one; see also Dennison \& Schafer, 2010).

The results show that, replicating the previous studies, preschoolers do not show adult-like understanding of con-
trastive prosody provided in an experimental setting (Experiment 1). However, the difficulty is alleviated when the prosodic input is preceded by a question that highlights what the alternative would be (Experiment 2). This suggests that preschoolers have prosodic representations while exhibiting difficulty in evoking a context-relevant alternative. Furthermore, children can use their knowledge about a construction (It's an X) that is more familiar to them to bootstrap their knowledge about contrastive inference (It LOOKS like a zebra...). These data together suggest that preschoolers use contextual information and pragmatic inferences to achieve adult-like performance in understanding contrastive prosody.

## Experiment 1

In Experiment 1, a game-like task was used to elicit preschoolers' interpretations of the two prosodic contours: Noun-focus and Verb-focus prosody. This was done to replicate the past studies' finding that young children fail to derive contrastive inference based on the prosodic information.

## Methods

Participants 12 children acquiring English as their first language ( 6 girls, 6 boys; mean age $4 ; 1$, age range $3 ; 2-4 ; 8$ ) were recruited and tested at a local nursery school in Stanford, California. For a comparison, 20 adults were also tested in the same paradigm using Amazon Mechanical Turk. The adult participants were all self-reported native speakers of American English residing in the United States.

Stimuli Sixteen high-frequency animal names were embedded in the sentence frame [It looks like an X] (e.g., It looks like a zebra). Half of the items were produced with Noun-focus prosody (e.g., "It looks like a ZEBRA!") and the other half were produced with Verb-focus prosody (e.g., "It LOOKS like a zebra"). The pronunciation patterns were counter-balanced between two lists: items pronounced with Noun-focus prosody in List 1 were produced with Verb-focus prosody in List 2 (and vice versa).

Sixteen more animal terms were chosen to form pairs in which the animals resembled each other in visual features (e.g., a zebra and an okapi, Figure 4). In each pair, the target animal named in the input sentence ("it looks like a zebra") was the more common of the two and was more familiar to the children being tested. Hereafter, the target named in a sentence (e.g., a zebra) is referred to as the "mentioned" animal and the paired animal (e.g., an okapi) is referred to as the "unmentioned" animal. The two animals in each pair served as likely referents for one or the other of the two prosodic contours used in the task (e.g., a zebra for "It looks like a ZEBRA!" and an okapi for "It LOOKS like a zebra")

Procedure The experimenter began by introducing a puppet and telling the child that they would play a guessing game together. The game had two parts. The first part was a picturenaming phase, in which the child saw seven pictures on a computer screen and labeled them one by one. This was done to ensure that the names of the objects were familiar to the


Figure 4: Experimental setup: Participants heard the puppet's clue while looking at a picture like (a) and were subsequently asked to guess which of the two pictures (as in (b)) was hidden behind the barrier.
child. In the second part, the test phase, the child and the puppet took part in seven trials (two practice and five critical) of a two-alternative forced-choice task. In each trial, the child and puppet were presented with a picture partially occluded by a gray barrier (Figure 4a). The puppet was then allowed to peek behind the barrier and give the child a clue about what he saw. The puppet's clue took the form of the "it looks like X" construction, pronounced with either Noun-focus prosody or Verb-focus prosody. All the puppet's speech was vocalized during the task by an experimenter who was a native speaker of American English. Following the puppet's clue, the child was presented with two pictures - the mentioned and unmentioned (Figure 4-b) - and asked to point which animal was hidden by the barrier. When the child pointed to a picture, she got feedback about which animal was the target referent. After completing seven trials, the child named five more animals and participated in five more guessing game.

## Results and Discussion

Figure 5 shows the proportion of mentioned animals chosen by the children and adult participants (e.g., choosing a zebra when the input sentence was [it looks like a zebra]). Adults responded to the stimuli in the expected ways: they reliably picked a mentioned animal based on the Noun-focus prosody (e.g., "It looks like a ZEBRA") and an unmentioned animal based on the Verb-focus prosody (e.g., "It LOOKS like a zebra...").

However, the four-year-olds did not differentiate the two patterns $\left(\chi^{2}(1)=1.36, p>.24\right)$. Overall, they showed a weak bias towards the picking an unmentioned animal ( $60 \%$ of all the responses), which might be due to their preference for a new, and often funnier looking, animal. Thus, replicating the previous findings, four-year-olds did not seem to make the contrastive inference based on the prosodic contour.

What makes the comprehension of contrastive prosody difficult for young children? Recall the model presented in Figure 2 . In order to correctly interpret the prosodic contours, a listener needs to be aware that they are mapped onto two distinct speaker intentions. In particular, it is critical to understand that the two prosodic patterns signal two intentions (answers) relevant to a question at hand (e.g., Is the animal a


Figure 5: Proportions of the mentioned animal chosen in Experiment 1 (children and adults) and Experiment 2 (children). Error bars represent standard error of the mean.
zebra or not?) While this reasoning comes naturally to adult listeners, young children might need more contextual support for establishing these assumptions that are plugged into the interpretation of contrastive prosody. To test this prediction, in Experiment 2, additional discourse-contextual information was provided to ensure that the children have better access to the contextual alternatives.

## Experiment 2

In Experiment 2, effects of an explicit guess-question are examined. The puppet asks the child to make a guess about the hidden animal, which establishes a question that needs to be answered (e.g., Is the hidden animal a zebra or not?) If this manipulation has a positive effect on the child's understanding of contrastive prosody, it would mean that the comprehension difficulty observed in Experiment 1 is at least partially attributed to their difficulty in detecting contextual alternatives.

## Methods

Participants 12 children ( 7 girls, 5 boys; mean age $4 ; 2$, age range $3 ; 6-4 ; 6$ ) were recruited and tested at the same nursery school as in Experiment 1.
Stimuli The stimuli were identical to those of Experiment 1.

Procedure The procedure was almost identical to that of Experiment 1 except that the puppet first pointed to the partially occluded picture and provided a guess-question: "What do you think is hiding behind the wall?" When the child did not give an answer, or the child's answer was unrelated to the trial item, the puppet followed up by saying, "I'm gonna guess it's an X (e.g., a zebra). But let me take a peek and give you a clue." This was done to provide an additional cue to ground the current "question under discussion" (Roberts, 2004), namely, whether the identity of the hidden animal was an X or not an X .

## Results and Discussion

A mixed logit regression analysis with the full two (Nounfocus vs Verb-focus prosody) by two (Experiment 1 vs. Experiment 2) design was employed to predict children's likelihood of choosing a mentioned animal for each stimulus sentence (e.g., choosing a zebra when the input sentence was "it looks like a zebra"). The model reported here has a full factorial random effect structure justified by the data, which contains random by-subject and -item slopes. All the predictors were sum-coded and there was no sign of collinearity. The original model contained the item order as a fixed effect, but it was removed from the analysis that follows based on a null effect in model comparison.

Overall, children showed a marginal but non-significant preference for choosing the mentioned animal when they heard the Noun-focus prosody ( $\beta=.62, p<.07$ ). Importantly, children were overall more likely to choose the mentioned animal in Experiment 2, where an explicit guessquestion was present ( $\beta=1.43, p<.003$ ) (Figure 5). There was also a significant interaction term between the prosodic input and the conditions ( $\beta=2.41, p<.001$ ), such that children were more likely to choose a mentioned animal based on the Noun-focus prosody in Experiment 2. That is, the explicit guess-question about a target animal facilitated their comprehension of the pragmatic interpretations of the two prosodic patterns. An additional analysis revealed that there was also an effect of age: older children tended to choose mentioned animals across conditions and input patterns significantly more often ( $\beta=.08, p<.04$ ).

How did the contextual support lead to adult-like judgement patterns? Two follow-up analyses were conducted to test if the presence and the types of guesses could predict children's choice behaviours. In Experiment 2, children were willing to make guesses $79 \%$ of the time, $53 \%$ of which included an animal mentioned in the input sentences. Consequently, the target animal was introduced to the discourse by the child $42 \%$ of the time, and by the puppet $58 \%$ of the time (e.g., "I'm gonna guess it's an X"). The children's responses to the puppet's guess-question were coded as binary predictors: (1) whether the child offered an animal name or not; and (2) whether the child made a correct guess. These predictors were included in two different models of children's choice of a mentioned animal in the Experiment 2. However, neither of these predictors was significant ( $p>.8$ and $p>.7$ respectively). This suggests that the facilitative effect observed in Experiment 2 cannot be reduced to children's expectation about a particular animal. Whether or not the child guessed correctly at the outset, the explicit introduction of the animal name provided support for their prosodic interpretations.

## Experiment 3

Experiment 3 tests a hypothesis based on the model of backward inferencing in prosodic comprehension. A structurally simpler, semantically less ambiguous, sentence "It's an X" was added to test if the presence of a familiar signal-intention
mapping supports children's inferences about a less-familiar signal and its pragmatic meaning. A few novel features were added to Experiment 3. First, pre-recorded speech was used for the input to rigorously control the input children received. Also, physically manipulable props (a picture card, a cardboard box) replaced the computer screen for the stimuli presentation.

## Methods

Participants 36 children acquiring English as their first language ( 24 girls, 12 boys; mean age $4 ; 6$, age range $3 ; 8-5 ; 2$ ) were recruited and tested. They were randomly assigned to one of the three conditions described in Table 1. 60 adults were also tested on-line, using Amazon Mechanical Turk. 3 adult participants were excluded from the data because their participation time was two standard deviation below the mean.

Stimuli 16 high-frequency animal names were embedded in two sentence frames: [It looks like an X] and [It is an X]. Tokens of [It looks like an X] with Noun-focus and Verbfocus prosody, as well as tokens of [It is an X], were recorded by a female native speaker of American English for use in the presentation of each trial.
Procedure Participants took part in a two-alternative forced choice task similar to Experiments 1 and 2. It consisted of a total of 16 trials (two practice trials and 14 critical trials). As in Experiment 1, the child participant was first introduced to a puppet. A mini portable speaker was attached to the puppet in order to play the audio stimuli. The child first participated in a picture-naming task, in which they labeled eight animals one-by-one. Then child and puppet took part in a guessing game where first they were presented with a box and told that it contained many different pictures of animals. Then the puppet was allowed to peek inside the box and give the child a clue. Next, each child was presented with two pictures (e.g., a zebra and an okapi) and prompted to indicate which of the two pictures was hidden in the box. Then the experimenter took a picture card from the box and showed the child which animal the puppet had actually "seen". After the first eight trials, the child was given eight more pictures of animals to name, and participated in eight more test trials.

Manipulation Children are put into one of the three conditions: Prosory-only, Form-only, and Combined conditions. In the Prosody-Only condition, as in Experiment 1, the puppet used either a Noun-focus or Verb-focus contour with [It looks like an X] to give a hint and a warning respectively. Hence the Prosody-only condition is expected to replicate the results from Experiment 1. In the Form-Only condition, the puppet said [It's an $\mathrm{X}!$ ] as a hint when the target animal was an X (e.g., "It's a ZEBRA" when the target picture depicted a zebra), compared to the puppet saying [It looks like an X] also with a focus on the final noun, as a warning when the picture was not an X (e.g., "It looks like a ZEBRA" when the target

Table 1: The between-subject manipulation of Experiment 3. The shaded cells indicate sentence patterns used for the warning function (identifying the hidden animal as not being the mentioned animal)

|  | Prosody-only X | Form-only | Combined |
| :--- | :--- | :--- | :--- |
| It's an X |  | "It's a ZEBRA" | "It's a ZEBRA" |
| Noun-focus prosody | "It looks like a ZEBRA" | "It looks like a ZEBRA" |  |
| Verb-focus prosody | "It LOOKS like a zebra..." |  | "It LOOKS like a zebra..." |

picture depicted an okapi). Notice that this manipulation is done based on the assumption that the Noun-focus prosody is in principle semantically ambiguous and it can be interpreted as it is an $X$ or It is not an $X$ depending on a speaker's preference and a context (Kurumada, Brown, \& Tanenhaus, 2012). It is hypothesized that children can better distinguish the pragmatic intentions based on these formal cues due to their reliance on lexically encoded information over prosodically encoded information in online processing (Snedeker \& Trueswell, 2003).

Finally, in the Combined condition, the puppet used [It's an X] for a hint with Noun-focus prosody, and Verb-focus prosody (e.g., "It LOOKS like an X") for a warning. Recall the word-learning situation with a spoon and a whisk. Presence of a familiar association (the word "spoon" and an intention to pick out a familiar object) results in higher confidence in a novel mapping compared to a situation in which there is no such familiar association. Experiment 3 tests if children can discover pragmatic meaning of contrastive prosody via a similar pragmatic inference: In the Combined condition, compared to the Prosody-Only condition, "It LOOKS like an x" becomes a more likely candidate for conveying the [it is not an X ] meaning because speaker should otherwise have said "It's an X" if she had meant that.

## Results and Discussion

All responses from the 14 critical trials were included in the analysis, as shown in Figure 6 for the children's response patterns. In the Prosody-only condition, which replicated Experiment 1, children's responses did not deviate from chance, and their judgments for each of the two prosodic patterns ("It looks like an X" and "It LOOKS like an X") did not differ significantly. In the Form-only condition, children showed more sensitivity to the contrast intended by the speaker ( $p .<.3$ ): They showed more diverged responses for the two types of sentences intended for a hint and a warning presumably due to their confidence in the interpretation of "It's an X".

In the Combined condition, the children showed nearly categorical and opposing responses for the two types of prompts, reliably choosing the mentioned animal when they heard "It's an X" and the unmentioned one when they heard "It LOOKS like an X". The most interesting comparison can be made between the responses for the Verb-focus prosody in the Prosody-only and the Combined conditions. They are acoustically identical and yet interpreted differently depending on the other sentence type used for a hint function. The difference between the Form-only and the Combined condi-


Figure 6: Proportions of an mentioned animal chosen by 4-year-olds in Experiment 3. Error bars represent standard error of the mean.


Figure 7: Proportions of an mentioned animal chosen by adults in Experiment 3
tion suggests that children were indeed aware of the pragmatic function of the contrastive prosody used in the Verbfocus prosody.

Figure 7 illustrates adults' responses in the three conditions. The most significant difference between the children's and adults' responses can be found in the Prosody-only condition. While children's responses for both of the prosodic patterns are at chance, adult listeners were almost categorically choosing a mentioned and an mentioned animals based on the Noun-focus and the Verb-focus contours respectively. The judgments were less categorical in the Form-only condition, which suggests that the children's and adults' response patterns were similar to each other in the Form-only condition as well as in the Combined condition.

## General Discussion

The results of the three experiments indicate that discourse contexts provide strong support for preschoolers' comprehension of contrastive prosody. In particular, Experiment 2 confirmed that an explicit question in a preceding context helped children to be tuned into the prosodic differences. This was considered to be because the question made it easier for them to derive two distinct speaker intentions (i.e., It is an $X$ and It is not an X ). This is in line with previous findings in which preschoolers' difficulty in computing scalar implicature was alleviated by an explicit depiction of contextual alternatives (Barner, Brooks, \& Bale, 2011).

Experiment 3 provided evidence that children engage in a rather complex probabilistic inferences when interpreting contrastive prosody: they interpret prosodic contours conditionally, depending on what other speech signals are used by the same speaker. When the speaker uses a more familiar, and semantically less ambiguous, sentence (i.e., It's an X), they can effectively infer that the prosodic prominence on LOOKS signals a distinct speaker intention (i.e., It is not an X).

Previous studies have viewed the interpretation of contrastive prosody as part of children's domain-specific knowledge about mappings between specific patterns of acoustic signals (i.e., $\mathrm{L}+\mathrm{H}^{*}$ ) and pragmatic meanings. However, that approach cannot explain how hearing other forms (e.g., "It's an $\mathrm{X}^{\prime \prime}$ ) in the same context affects children's understanding of contrastive prosody. The current results suggest that children are trying to solve a bigger inference problem, where they cope with uncertainty regarding different speech signals, and estimate the likelihood with which prosodic signals are mapped onto different meanings.

These results highlight the possibility that such contextual inferences allow children to process pragmatic interpretations of prosody even before they acquire fully-fledged understanding of the prosody-pragmatics interface. As we saw in Experiment 1 and 3, four-year-olds do not yet exhibit adult-like knowledge about contrastive prosody. In other words, they do not reliably call to mind a contrast set based solely on prosodic minimal pairs (e.g., It looks like an X vs. It LOOKS like an $x$ ) and make an inference about a speaker's intentions. Nevertheless, with more contextual and linguistic information, they engage in inferences that result in adult-like responses. The role of such contextually-supported inferences has been discussed almost exclusively in word-learning while its full implications for language acquisition remain to be understood. The current results suggest that a similar model can explain pre-schoolers' intonational interpretation. A contrastive meaning of prosody LOOKS unattainable for those young children, but it is not in a context.

## Acknowledgments

Thanks to Sarah Bibyk, Meredith Brown, Marisa Casillas, Eve V. Clark, Christine Gunlogson, T. Florian Jaeger, Molly Lewis, and Michael K. Tanenhaus for valuable discussion.

## References

Barner, D., Brooks, N., \& Bale, A. (2011). Accessing the unsaid: the role of scalar alternatives in childrens pragmatic inference. Cognition, 118(1), 84-93.
Büring, D. (2003). On d-trees, beans, and b-accents. Linguistics \& Philosophy, 26(5), 511-545.
Clark, E. V. (1990). On the pragmatics of contrast. Journal of Child Language, 17(02), 417-431.
Cruttenden, A. (1985). Intonation comprehension in ten-year-olds. Journal of Child Language, 12, 643-661.
Cutler, A., \& Swinney, D. A. (1987). Prosody and the development of comprehension. Journal of Child Language, 14(01), 145-167.
Dennison, H. Y., \& Schafer, A. J. (2010). Online construction of implicature through contrastive prosody. In Speech prosody 2010 conference.
Frank, M. C., Goodman, N. D., \& Tenenbaum, J. B. (2009). Using speakers referential intentions to model early cross-situational word learning. Psychological Science, 20(5), 578-585.
Ito, K., \& Speer, S. R. (2008). Anticipatory effects of intonation: Eye movements during instructed visual search. Journal of Memory and Language, 58(2), 541-573.
Kurumada, C., Brown, M., \& Tanenhaus, M. K. (2012). Prosody and pragmatic inference: It looks like speech adaptation. In Proceedings of the 34th annual conference of the cognitive science society.
Ladd, D. R. (2008). Intonational phonology (2nd ed.). Cambridge University Press.
Markman, E. M., \& Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. Cognitive Psychology, 20(2), 121-157.
Pierrehumbert, J., \& Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. Intentions in Communication, 271-311.
Roberts, C. (2004). Discourse context in dynamic interpretation. In L. Horn \& G. Ward (Eds.), Handbook of contemporary pragmatic theory (pp. 197-220). Blackwell.
Sekerina, I. A., \& Trueswell, J. C. (2012). Interactive processing of contrastive expressions by russian children. First Language, 32(1-2), 63-87.
Snedeker, J., \& Trueswell, J. (2003). Memory and language using prosody to avoid ambiguity : Effects of speaker awareness and referential context. Journal of Memory and Language, 48(1), 103-130.
Solan, L. (1980). Contrastive stress and children's interpretation of pronouns. J Speech Hear Res, 23(3), 688-698.
Speer, S. R., \& Ito, K. (2009). Prosody in first language acquisition acquiring intonation as a tool to organize information in conversation. Language and Linguistics Compass, 3(1), 90-110.
Wells, B., Peppe, S., \& Goulandris, N. (2004). Intonation development from five to thirteen. Journal of Child Language, 31(04), 749-778.

# Communicatively efficient language production and case-marker omission in Japanese 

Chigusa Kurumada<br>kurumada@stanford.edu<br>Department of Linguistics<br>Stanford University

T. Florian Jaeger<br>fjaeger@bcs.rochester.edu<br>Department of Brain and Cognitive Sciences<br>University of Rochester


#### Abstract

Recent studies hypothesize that language production is governed by the principle of efficient information transmission: Speakers tend to omit elements whose information content is contextually predictable, while providing more linguistic signal to convey otherwise less predictable information. However, previous findings in support of this hypothesis are also compatible with alternative accounts based on production difficulty. To distinguish between these competing accounts, we conducted experiments on speaker's preference in optional casemarking in Japanese. The results suggest that Japanese speakers are more likely to omit the object case-marker when an associated noun has properties (e.g., animacy) that are prototypical to a grammatical object. Moreover, case-marker omission was facilitated when other elements in a sentence made the grammatical function assignment more predictable. The results were obtained with all the factors related to production difficulty held constant, and thus provide support for the models of communicatively efficient language production.


Keywords: Case-marker omission, Japanese, language production, efficiency, information transfer

## Introduction

Language often offers multiple options for expressing semantically equivalent or near-equivalent messages. For example, speakers have a choice between different word orders and voice choices (e.g. the ditransitive alternation in English; active vs. passive; scrambling in languages like German, Japanese, and Hindi). Another common type of alternation involves the choice between more or less linguistic form (e.g., phoneme duration, morphemes, phonological alternation) to encode the same meaning. For example, in English sentence like (1), speakers can, but do not have to, produce the complementizer "that". The past psycholinguistic studies have used these alternations as windows into the cognitive processes underlying language production.
(1) My boss said (that) we were absolutely crazy.

Recent psycholinguistics work has proposed that speakers' preferences in such alternations provide evidence that the computational system underlying language production is organized to facilitate efficient information transfer (e.g., Aylett \& Turk, 2004; Levy \& Jaeger, 2007; Jaeger, 2010). This line of work has focused on the observation that it seems to be predictable linguistic material, and hence material that is low in Shannon information, that tends to be reduced or even completely omitted (e.g., Bell, Brenier, Gregory, Girand, \& Jurafsky, 2009; Frank \& Jaeger, 2008; Resnik, 1996). For example, whether English speakers produce the complementizer that depends on how predictable it is for the continuation
of words so far (e.g., My boss said) to take a complement clause, as in (1): when the complement clause is much expected, that is likely to be omitted; when it is less so, that is likely to be inserted (Jaeger, 2010).

Observations like these can be accounted for in terms of communicative efficiency since the reduction/omission of predictable material and lengthening/insertion of less predictable material results in more uniform distribution of information density across the speech signal, which is proven to facilitate efficient -i.e. fast and robust- information transfer (Genzel \& Charniak, 2002) and to minimize processing difficulty (Levy \& Jaeger, 2007). A trade-off between the amount of information and the amount of linguistic signal expended is also expected under models of boundedly rational communication: If speakers are rational and aim to balance production effort with communicative success, they should provide a more perceptual signal when a word or structure is less contextually inferable (the Ideal Speaker Model in Jaeger, 2011, see also Piantadosi, Tily, \& Gibson, 2011).

However, for much of the evidence that has been cited in favor of communicative efficiency accounts, it is an open question to what extent it could be accommodated in competing accounts. In particular, it is well-established that difficulty with the retrieval, processing, or articulation of upcoming material can affect the degree of reduction/omission of (typically immediately preceding) material. This includes effects on phonetic reduction (Fox Tree \& Clark, 1997) and the optional production of "that" (e.g., Ferreira \& Dell, 2000) described above. For example, speakers are more likely to produce the optional "that", if the onset of the complement clause (its subject) is frequent, short and has previously been mentioned (Ferreira \& Dell, 2000; Roland, Elman, \& Ferreira, 2005). We follow previous work and refer to these effects as availability-based effects.

Here we investigate a morpho-syntactic alternation that provides a less ambiguous test case for the communicativeefficiency accounts. We present three production experiments on optional case-marking in Japanese. Japanese is a verb-final language with relatively flexible word order: in a transitive sentence like (2), both SOV and OSV are possible word orders, although the latter is considerably less frequent. Case relations are marked with post-nominal particles (case-markers) as shown in (2), where "-ga" is the nominative marker and "-o" is the accusative marker. Here we focus on the optionality of the accusative case-marker, indicated in () by parentheses.
(2) Taro-ga sushi-(o) tabe-ta.

Taro-NOM sushi-(ACC) eat-PAST.
Taro ate sushi.
Compared to languages like English or Mandarin Chinese, the flexibility in Japanese word order implies higher uncertainty about the grammatical function assignment (henceforth GF-assignment). In other words, it is not immediately obvious what is the subject and what is the object especially when a case-marker is not present. While case-markers are obligatory in written discourse, however, they are frequently omitted in conversational speech - often, like in (2), without change in meaning.

Such optional case-marking has received considerable attention in linguistic work (e.g., Aissen, 2003), but has remained comparatively understudied in psycholinguistics. Here it is of interest because it allows us to distinguish between the predictions of communicative efficiency accounts, specifically the ideal speaker model proposed in Jaeger (2011), and those of availability-based accounts.

The ideal speaker model predicts that speakers should be more likely to produce case-marking when the intended GFassignment is unexpected given the other information provided in the sentence (for details, see Jaeger, 2011). This is the prediction we test here. In the remainder of this paper, we will use the more succinct statement that "speakers should be more likely to mark the unexpected" to refer to this prediction. In Experiment 1, we investigate whether object typicality, manipulated via changes in animacy, affects speakers' preference during production of optional case-marking in Japanese. We compare the production of sentences that only differ in the animacy of the object, which is either a human referent (e.g., the student) or an inanimate referent (e.g., the fire-engine)(Figure 1-a). If speakers are more likely to mark the unexpected, they should be more likely to produce object case markers if the object is human [atypical] compared to if it is inanimate [typical] (Aissen, 2003).

Preliminary evidence for this prediction comes from qualitative work (Lee, 2007), which examined optional casemarking in conversational Korean. She tested the hypothesis suggested by typological work on case-marking systems. Lee found that both definiteness and animacy are significant predictors of Korean speakers' use of subject- as well as object case-markers in the direction predicted here. However, in a similar corpus study in Japanese, Fry (2003) found no effect of animacy in the object-case marking. While inanimate and indefinite (i.e. atypical) subjects were significantly more likely to be case-marked than animate and definite subjects ( $69 \%$ vs. $64 \%$ ), he found no effect of animacy on object case-marking. It is unclear, however, whether this result indicates an actual absence of an effect of animacy, or whether this was simply due to the sparseness of data or confounding factors, not controlled in the corpus study. Experiment 1 addresses this question by manipulating only object animacy. Further, items were constructed such that all materials following the direct object were held constant, so that no differ-
ences in case-marking preferences between the conditions are predicted by availability-based production (Ferreira \& Dell, 2000).

Experiment 2 and 3 put the ideal speaker model's prediction to a stronger test. We investigate if Japanese speakers' case-marking preference can be directly affected by the inferability of GF-assignment, beyond categorical factors like animacy, givenness, or definiteness. To this end we manipulate plausibility of GF-assignments and investigate their effect on the production of case markers.

## Experiment 1

We employ a spoken recall paradigm to test if Japanese speakers' use of the accusative case-marker -o is sensitive to animacy of the direct object. We manipulated the animacy of the direct object and whether the original stimulus contained a direct object case-marker or not. If optional case-marking is affected by a preference for communicative efficiency, speakers should be more likely to produce responses with a casemarker for animate (atypical) objects compared to inanimate (typical) objects.

## Methods

Participants 20 native speakers of Japanese in Stanford area participated in this study. They received $\$ 7$ for their participation.

Materials In a Latin-square design, each list contained 24 items and 48 fillers. As described above, items were transitive sentences with either animate or inanimate direct objects and with or without a case-marker. Sentence patterns for recall stimuli are illustrated in Figure 1. The nominative casemarker was always present, avoiding ambiguity about the intended GF-assignment and hence about the meaning. Additionally, all items were presented in the subject-before-object order, which is hugely more frequent in Japanese.

Fillers were length-matched sentences with intransitive verbs and longer adverbial phrases. There was no lexical overlap between any of the stimuli. Stimuli were grouped into pairs so that there were 24 item-filler pairs and 12 filler-filler pairs, totalling 36 trials. The order of items and fillers within a pair and the order of pairs were held constant across participants. All stimuli were recorded by a female native speaker of Japanese, using the same prosody for all conditions. In addition, the same speaker recorded 72 recall prompts, one for each sentence (always the verb).

Procedure Each trial consisted of an encoding phase and a recall phase. During encoding, participants listened to pairs of sentences and were instructed to remember them. During recall, participants heard the verb of one of the two sentences (the prompt). They then recalled and produced the full sentence corresponding to that verb. Subsequently, the second prompt was played and participants produced the second sentence. In half of the trials, the sentence encoded first was also recalled first. In the other half, the order was reversed. Following standard procedure (Ferreira \& Dell, 2000), target


Figure 1: Sentence patterns in a) Experiment 1 (left) and b) Experiment 2 and 3 (right).

Table 1: Percentage of recall error for each of the four conditions in Experiment 1.

|  | Animate | Inanimate |
| :--- | :--- | :--- |
| Present | $8.9 \%$ | $8.9 \%$ |
| Absent | $15.6 \%$ | $8.9 \%$ |

items were never recalled directly after encoding. That is, if an item was encoded first, it was recalled either first or second. If it was encoded second, it was always recalled second. Across participants the recall order for each pair of stimuli was held constant.

Scoring All $20 \times 24=480$ recorded items were transcribed and coded by a native Japanese undergraduate research assistant who was unaware of the purpose of this experiment. Sentences with recall errors anywhere in the sentence were excluded ( $11.6 \%$ ), leaving 424 responses for our analysis. The error rate was doubled for sentences with animate objects without case-marker compared to the other three conditions (Table 2, $\chi^{2}(3)=.7, p<.05$ ). Although these sentences are technically not ambiguous (recall that all sentences contained subject case-marking), sentences with animate objects that lack case-marking are the ones we hypothesized to be most confusing. The high error rate indicates that the recall production was in fact difficult with this type of sentences.

We also coded the presence of disfluency ( 88 out of the 424 analyzable sentences contained at least one disfluency). There was no significant difference across sentence types $\left(\chi^{2}(1)=.03, p>.8\right)$.

## Results and discussion

A mixed logit regression analysis with the full 2 (animacy of object) x 2 (presence of case-marker in stimulus) design was employed to predict the presence of a case-marker. We report the results for the model with maximum random effect structure justified by the data based on model comparison (Jaeger, 2008), which contained random by-subject and item intercepts as well as by-item slopes for the presence of case-marker in stimulus. The effect of interest reported below was robustly significant even in the full random effect model (by-subject and by-item slopes for all factors and their interaction). All predictors were contrast coded and there were no signs of collinearity (fixed effect $r s<.07$ ).


Figure 2: Proportion of case-marker use in Experiment 1. The error bars represent $+/-1$ standard error.

Unsurprisingly, participants were more likely to produce a object case-marker if the original stimulus contained a object case-marker ( $\beta=2.3, p<.0001$ ). Crucially, participants were also more likely to produce a case-marker if the object was animate ( $\beta=.45, p<.03$ ). The two factors did not interact ( $p>9$ ). Figure 2 summarizes the two main effects.

In this recall production experiment, we thus found that animacy of direct objects affected Japanese speakers' casemarking preferences. This suggest that the failure to find animacy-based effects on object case-marking in casual speech in previous work is indeed due to a lack of power and confounding effects of other variables that optional casemarking is sensitive to (Fry, 2003). Our results hence replicate for Japanese what has been found in corpus-based studied on Korean speech (Lee, 2007). Our findings are also compatible with the qualitative description of differential casemarking across languages of the world (Aissen, 2003).

Most relevant to the current purpose, the animacy-based effect on case-marking support the ideal speaker model. The effects we observe are expected under the ideal speaker model and cannot be reduced to availability-based production since the material following the direct object was held constant within items. As discussed above, it is also not clear how production-oriented ('accessibility-based') accounts of the type that have been proposed for phonetic reduction (Arnold, in press; Bell et al., 2009) would account for the observed effects. More generally, it seems unlikely that the effects
observed here are reducible to production difficulty. While we found an increase proportion of recall errors in sentence with animate (atypical) objects without case marker, the proportion of recall errors for animate objects with case markers was identical to that of inanimate objects. Furthermore, the interaction observed for recall errors is not observed in our analysis of speakers' case-marker preferences. Finally, as mentioned above, we did not observe any differences in the distribution of disfluencies across the animacy conditions.

## Experiment 2

If the animacy effect we saw in Experiment 1 was indeed due to effect of communicative efficiency on the encoding of GFassignment, the effect should remain when the animacy of the arguments are controlled. Using animate-animate noun pairs, we examined if the plausibility of GF-assignment affects Japanese speakers' use of the accusative case-marker.

In Figure 1-b (right), the GF-assignment is more plausible when the doctor treats the patient rather than the other way around. Hence the ideal speaker model predicts that speakers should be more likely to produce the object case-marker when the doctor is the object of the sentence, compared to when the patient is the object.

## Methods

Participants 32 native speakers of Japanese in Stanford area participated in this study. They received $\$ 7$ for their participation.
Stimuli There were 24 items, consisting of subject, object, adverb, and verb. Each item consisted of eight conditions ( $=192$ stimuli), resulting from crossing 1) the plausibility of GF-assignment based on the order of two noun phrases and the verb, 2) presence/absence of the case-marker, and 3) the identity of the verb. This is illustrated in Figure 1.

In the example in Figure 1-b, plausibility was coded as high in "The doctor treated the patient in a hospital room" and "The patient waited for the doctor in a hospital room" and as low in "The patient treated the doctor in a hospital room" and "The doctor waited for the patient in a hospital room". Below we collapse over the 2-way within-item verb contrast and treat the design as a $2 \times 2$ (plausibility by case-marker presence in the input), since the identity of the verb is of no theoretical interest here (the effects of verbs on plausibility are already capture by our coding of plausibility). Furthermore, the verb identity was balanced within item and within the two plausibility conditions.

As in Experiment 1, the 24 items were combined with 48 length-matched fillers in a Latin-square design that held order of stimuli constant across lists. There was no lexical overlap between any of the stimuli. The grouping of stimuli into pairs was the same as in Experiment 1. All stimuli were recordings of a female native speaker of Japanese. In addition, the same speaker recorded 72 prompts, one for each sentence (always the verb).

Table 2: Percentage of recall error for each of the four conditions in Experiment 2.

|  | Plausible object | Implausible object |
| :--- | :--- | :--- |
| Present | $8.6 \%$ | $16.7 \%$ |
| Absent | $8.6 \%$ | $12.3 \%$ |

Norming studies The stimuli described below were created using an online norming study (40 native speakers of Japanese). Participants were presented the 24 animateanimate noun pairs used in our items (e.g., doctor-patient). Participants first rated the relative naturalness of the two potential patterns of grammatical function assignment on a 10 point rating scale (e.g., The doctor (did something to) the patient vs. The patient (did something to) the doctor). Second, we asked them to provide verbs that would make the most natural continuation of a given pair of nouns for each pattern (e.g, doctor-NOM patient-ACC / patient-NOM doctor-ACC). Among the set of verbs given by the informants, we selected two verbs for each noun pair such that one verb maximizes the object probability of one noun, and the other verb maximizes the object probability of the other noun. For example, for the "doctor-patient" pair, we chose "(to) treat", which makes "patient" as the plausible object, and "(to) wait for", which makes "the doctor" more plausible.

A second norming study (40 native speakers of Japanese; no overlap in participants with first norming study or either of the experiments) asked participants to rate the relative naturalness of the two patterns of GF-assignment when the verb is present (e.g., "doctor-NOM patient-ACC hospital-roomLOC treat" vs. "patient-NOM doctor-ACC hospital-roomLOC treat").

Procedure The procedure was identical to that of Experiment 1. Since some of the stimuli (e.g., The patient treated the doctor.) were meant to be less plausible, participants were instructed to listen to sentences carefully and produce them faithfully to the input even when they were "somewhat surprising".

Scoring All $32 \times 24=744$ responses were transcribed and coded by a native speaker of Japanese. Sentences with recall errors were excluded ( $11.6 \%$ [sic]), leaving 658 responses for our analysis. Numerically, the error rate was higher for sentences with implausible GF-assignments ( $14.5 \%$ vs. 8.6\%) but the difference was not statistically significant $\left(\chi^{2}(1)=\right.$ $.24, p>.62$ ). $13 \%$ of the error free sentences contained at least one instance of disfluency, but the occurrence rate did not differ across conditions ( $\chi^{2}(3)=.80, p>.3$ ). As in Experiment 1, the distribution of disfluencies did not differ across conditions ( $\chi^{2}(3)=.80, p>.3$ ).

## Results and Discussion

Using the same statistical approach as in Experiment 1, we analyzed the remaining $2 \times 2$ design defined by the plausibility of GF-assignment (high vs. low) and case-marker presence


Figure 3: Proportion of case-marker use in Experiment 3 by plausibility of GF-assignment (high vs. low) and presence of the object case marker in the recall stimulus ( $-o$ present vs. absent). The error bars indicate $+/-1$ standard error.
in the originally presented stimuli (presence vs. absence) to predict the presence of a case-marker in the responses. Unsurprisingly, participants were more likely to produce an object case-marker if the original stimulus contained an object case-marker ( $\beta=2.6, p<.0001$ ). More importantly, they were also more likely to produce a case-marker for implausible GF-assignment ( $\beta=0.8, p<.003$ ). The two factors did not interact ( $p<.7$ ). Thus, this model confirmed the predicted effect of object plausibility in the form of categorical factor. The effects are illustrated in Figure 3.

Next, we analyzed the gradient effects of values for the plausibility of GF-assignments as rated in the norming studies described above. The effect of two different plausibility ratings was examined: plausibility of GF-assignment given only the two nouns (Norming Study 1, part 1) or given the full sentence (Norming Study 2). Plausibility ratings for full sentences returned the expected significant effect ( $\beta=-0.43$, $p<.03$ ), namely that the speakers were more likely to use the case-marker when the GF-assignment had been normed to be less plausible. However, the ratings based on only the two noun phrases returned no such effect ( $\beta=-.05, p>.9$ ). This suggests that what speakers assess is the plausibility of an intended message (including the verb information) rather than just the properties of nouns.

## Experiment 3

We replicated Experiment 2 by using the subject noun, instead of the verb, as a recall cue. This allows us to rule out the possibility that the plausibility effect observed in Experiment 2 is an artifact of the recall cue being the verb of a sentence. As Japanese is a SOV language, it is possible that the use of the (sentence final) verb as the recall cue in Experiment 1 and 2 forced participants to adopt task-specific production strategies that do not reflect normal language production.


Figure 4: Proportion of case-marker use in Experiment 3 by plausibility of GF-assignment (high vs. low) and presence of the object case marker in the recall stimulus ( $-o$ present vs. absent). The error bars indicate $+/-1$ standard error.

## Methods

Participants 26 native speakers of Japanese in Stanford area participated in this study. Data from one subject were excluded because of a problem with the recording device.

Stimuli The stimuli and their presentation order were mostly identical to those of Experiment 2. New recall cues (the subject nouns of the target sentences) were recorded by the same speaker who recorded the stimuli sentences for Experiment 2.

## Results and Discussion

Using the same statistical approach as in Experiment 1 and 2 , we analysed the binary predictors of the plausibility of GF-assignment (high vs. low) and case-marker presence in the originally presented stimuli (presence vs. absence of $-o$ ) to predict the occurrence of a case-marker in the responses. We replicated the finding in Experiment 2: the expected effect of the object case-marker presence in the original stimulus ( $\beta=1.83, p<.001$ ), and the low inferrability GFassignments ( $\beta=-1.12, p<.007$ ). The two effects did not interact ( $p>.2$ ). Thus, this model confirmed that the effect of object plausibility (in the form of categorical factor) remains when the recall cue was the subject of the sentence. The effects are illustrated in Figure 4.

Also, we replicated the effects of two different plausibility ratings. Plausibility ratings for full sentences returned the expected significant effect $(\beta=-0.40, p<.05)$ whereas the ratings based on only the two noun phrases returned no such effect ( $\beta=-.03, p>.6$ ). There was a negative interaction term between the plausibility value given the full sentence and the presence of the case-marker: Highly plausible GFassignments were even less likely to be case-marked when the original sentences lacked the object case-marking ( $\beta=-.74$, $p<.01$ ).

## General Discussion

Our results suggest that Japanese speakers prefer to produce an object NP without case marking when grammatical function of a noun is made more predictable given the semantics of the noun (e.g., animacy) and the other linguistic elements in the sentence (e.g., plausibility of GF-assignment given the subject, object, and verb). The plausibility effect we saw in Experiment 2 and 3 strongly suggests that speakers have fine-grained probabilistic knowledge about the plausibility beyond categorical factors like animacy and definiteness. More generally, this is one of the first studies showing a systematic effect of plausibility in the morpho-syntactic encoding of speech.

Recent studies have found that speakers' use and nonuse of case-markers are interacting with the word order as well. Speakers tend to use case-makers when the word order does not conform to an expected pattern (see Lee and Kim (2012) in Korean, Fedzechkina, Jaeger, and Newport (2012) for experiments using an artificial language). This is also predictable by the assumption of efficient communication. An explicit case marking becomes more likely when the non-canonical word order biases against an intended GFassignment pattern.

This body of research including the current study constitutes strong support for the view that language production is optimized to maximize the efficiency of information transmission (Jaeger, 2010; Levy \& Jaeger, 2007), and, in particular, the ideal speaker model (Jaeger, 2011). Unlike most previous work, the current results cannot be accounted for in terms of availability, even if availability-based production is extended to include the availability of upcoming syntactic structures.

Furthermore, our results provide broader implications for studies investigating the effects of communicative pressure on cross-linguistically attested phenomena such as differential case-marking (Aissen, 2003). Our results and Lee's (2007) study in Korean support the hypothesis that the languages with an optional case-marking system are sensitive to the same factors that are known to affect more categorical caseencodings in various languages. This may suggest that the functional pressure for efficient communication underlies at least some of the universal features found across languages.

## References

Aissen, J. (2003). Differential object marking: iconicity vs. economy. Natural Language and Linguistic Theory, 21, 435-483.
Arnold, J. E. (in press). Reference production: Productioninternal and addressee-oriented processes. Language and Cognitive Processes.
Aylett, M. P., \& Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. Language and Speech, 47(1), 31-56.

Bell, A., Brenier, J., Gregory, M., Girand, C., \& Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. JML, 60(1), 92-111.
Fedzechkina, M., Jaeger, T. F., \& Newport, E. L. (2012). Language learners restructure their input to facilitate efficient communication. Proceedings of the National Academy of Sciences.
Ferreira, V., \& Dell. (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. Cognitive Psychology, 40, 296-340.
Fox Tree, J. E., \& Clark, H. H. (1997). Pronouncing "the" as "thee" to signal problems in speaking. Cognition, 62, 151-167.
Frank, A., \& Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In Proceedings of the 30th annual meeting of the Cognitive Science Society.
Fry, J. (2003). Ellipsis and wa-marking in japanese conversation. New York: Routledge.
Genzel, D., \& Charniak, E. (2002). Entropy rate constancy in text. In Proceedings of the Association of Computational Linguistics (pp. 199-206). Philadelphia, PA.
Jaeger, T. F. (2008). Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models. JML, 59, 434-446.
Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. Cognitive Psychology, 61, 23-62.
Jaeger, T. F. (2011). Communicative efficiency and adaptiveness in the ideal speaker (ms). University of Rochester.
Lee, H. (2007). Case ellipsis at the grammar/pragmatics interface: A formal analysis from a typological perspective. Journal of Pragmatics, 39(9), 1465-1481.
Lee, H., \& Kim, N. (2012). Non-canonical word order and subject-object asymmetry in korean case ellipsis. In Proceedings of the 19th international conference on head-driven phrase structure grammar (pp. 427-442). Stanford, CA: CSLI Publications.
Levy, R., \& Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In B. Schlökopf, J. Platt, \& T. Hoffman (Eds.), Advances in Neural Information Processing Systems (NIPS) (Vol. 19, p. 849-856). Cambridge, MA: MIT Press.
Piantadosi, S., Tily, H., \& Gibson, E. (2011). Word lengths are optimized for efficient communication. Proceedings of the National Academy of Sciences, 108(9), 3526.

Resnik, P. (1996). Selectional constraints: An informationtheoretic model and its computational realization. Cognition, 61, 127-159.
Roland, D., Elman, J. L., \& Ferreira, V. S. (2005). Why is that? Structural prediction and ambiguity resolution in a very large corpus of English sentences. Cognition, 1-28.

# Snapshots of Working Memory: Using Early Eye Movements to Capture Temporal Dynamics 

Nicholas D. Lange (ndlange@gmail.com) $\dagger$.<br>Rick P. Thomas (Rickey.P.Thomas-1@ou.edu) $\ddagger$ Daniel R. Buttaccio (buttacciodr@ou.edu) $\ddagger$ Eddy J. Davelaar (e.davelaar@bbk.ac.uk) †<br>$\dagger$ Birkbeck, University of London, Department of Psychological Sciences, Malet Street, London, UK, WC1E 7HX<br>$\ddagger$ University of Oklahoma, Department of Psychology<br>Lindsey Street, Norman, OK, USA 73019


#### Abstract

Research investigating top-down attentional capture has demonstrated a tight coupling of working memory content with attention and eye movements. By capitalizing on this relationship, we have developed a novel methodology called the Memory Activation Capture (MAC) procedure for measuring the dynamics of working memory content supporting complex cognitive tasks (e.g., decision making, problem solving). By observing which items are preferentially fixated in task irrelevant arrays containing task relevant information, we gain a measure of working memory content as the task evolves through time. The efficacy of the MAC procedure is demonstrated in a hypothesis generation task. Results suggest a two-stage process following hypothesis retrieval whereby it undergoes a brief period of heightened activation before entering a lower activation state while being maintained for output. The present effects are of additional general interest as they represent the first demonstrations of top-down attentional capture driven by participant-established WM content retrieved from long-term memory.


Keywords: attention, memory, decision making, eye tracking, process tracing, hypothesis generation

## Introduction

The considerable interest in understanding the cognitive dynamics of information use over time is underscored by the proliferation of process-tracing methodologies within several domains. Think-aloud procedures, in which a participant provides concurrent verbalization of their cognitive states while performing a task, were among the first of these techniques to be developed (Ericcson \& Simon, 1993; Ford, Schmitt, Schechtman, Hults, \& Doherty, 1989; Montgomery \& Svenson, 1976; Svenson, 1979) and still enjoy widespread use today (Schulte-Mecklenbeck, Kühberger \& Ranyard, 2011). The usage of eye movements as a window to dynamic cognitive processing has flourished as of late with application in several fields including decision making (Franco-Watkins \& Johnson, 2011; Glaholt, Wu, \& Reingold, 2009, Glaholt \& Reingold, 2011, Sutterlin, Brunner, \& Opwis, 2008), problem solving (Ellis, Glaholt, \& Reingold, 2011), categorization (Rehder \& Hoffman, 2005a, 2005b), language comprehension (Cooper, 1974; Tanenhaus, 1995), and diagnostic reasoning (Renkewitz \& Jahn, 2012).

The methodology forwarded here shares the same goal as process-tracing techniques, namely to gain better understanding of cognitive dynamics by measuring information use as the task naturally unfolds. Our method, however, takes a novel approach towards revealing the content of working memory by relying on the recent literature addressing attentional "top-down capture" effects demonstrating a tight coupling between the content of WM and the deployment of attention and eye movements. For instance, Soto, Heinke, Humphreys, and Blanco (2005) and Soto \& Humphreys (2007) provide evidence that attention is automatically captured by the contents of WM (although for contrasting evidence see Woodman \& Luck, 2007). As the presently forwarded methodology relies on eye movement data it is of particular importance that Soto et al. (2005) found eye movements to be sensitive to the spatial congruency of target and WM-matching items. Moreover, Moores, Laiti, and \& Chelazzi (2003) found that first saccades were biased towards WM-matching items as well as semantic associates of items maintained in WM.

As eye movements are biased by the content of WM, it may be possible to capitalize on this bias to develop a measure of WM content deployable in complex cognitive tasks. Specifically, by presenting brief visual arrays containing task related information at various points in time, differences in the oculomotor guidance towards the items contained in such "WM probe arrays" could be taken as evidence regarding the active content of WM at the time of the array presentations. In this way our methodology can be thought of as an effort to capture snapshots of WM across time. We refer to our methodology as the Memory Activation Capture (MAC) procedure. Although the logic of this procedure (as well as its specific advantages) have been treated elsewhere (Lange, Thomas, Buttaccio, \& Davelaar, 2012), the present experiment represents the first deployment of this procedure in a complex cognitive task.

In the present paper, we deploy this procedure in the context of a memory retrieval task to investigate the temporal dynamics of hypothesis generation. We define hypothesis generation as a general case of cued recall in which the observation of one or multiple cues can lead to the retrieval of one or multiple hypotheses (Dougherty, Thomas, \& Lange, 2010; Thomas, Dougherty, Sprenger, \& Harbison, 2008). In our day-to-day lives, we utilize this
process to better understand the occurrences we witness in our environment. For instance, if a friend is acting differently than usual you may generate various explanations for their behavior. A professional example comes to us through medical diagnosis in which a physician observes various symptoms from a patient and retrieves associated diagnoses from long term memory (LTM).

Recently, we have investigated the influences of time and sequence on hypothesis generation by formulating a model addressing the influence of WM dynamics during information acquisition on the retrieval (i.e., generation) of hypotheses and confirmed model predictions (Lange, Thomas, and Davelaar, 2012 Lange, Thomas, Buttaccio, Illingworth, \& Davelaar, 2012, Lange, Davelaar, \& Thomas, In Press). This model assumes that the memory activation associated with each piece of acquired information (i.e., data) undergoes a dynamic rise and fall over time in accordance with 1) competition from other acquired items, 2) bottom-up activation, and 3) its self-recurrent excitation (see Davelaar, Goshen-Gottstein, Ashkenazi, Haarmann, \& Usher (2005) for a fuller treatment and computational details). We hypothesize that the memory activations of acquired data and retrieved hypotheses are subject to the same competitive WM dynamics. We now provide a hypothetical example of how the memory activations of data and hypotheses may trade off in a simplified medical diagnosis task and use this example to illustrate a hypothetical deployment of the MAC procedure.

Figure 1 provides a hypothetical example of the deployment of our procedure in the context of a simplified medical diagnosis task (e.g., hypothesis generation, diagnostic reasoning). The task is initiated with the presentation of a patient symptom (e.g., fever). As this information is acquired, its associated memory representation becomes activated and rises. Shortly after this data has been acquired, the memory activation associated with an associated diagnosis begins to rise and is generated when its memory activation crosses a threshold distinguishing the content of WM. The memory activation of the diagnosis continues to rise while at the same time the activation associated with the symptom decreases due to competitive WM processes. The points labeled T1-T4 represent points at which the WM Probe Array could be presented. In this example, we assume that the probe array (represented visually) contains four items: the diagnosis and three distractor items. At T1, the diagnosis would not be fixated more than the distractors in the probe array. However, at T2, the diagnosis has been retrieved and resides in WM. At this point, we would expect to see the diagnosis being fixated more often than the distractor items, indicating that it is active in WM. Moreover, at time points T3 and T4 we might see a rise and fall in the fixation rate of the diagnosis due to the rise and fall of its associated memory activation. An important aspect regarding our use of topdown oculomotor capture as a measure of WM content is that, unlike any visual search task, the WM Probe Arrays used in the present experiments are completely task
irrelevant. That is, the participant does not have a task to perform on the array and is not instructed for any response to the arrays.


Figure 1: Hypothetical deployment of the MAC procedure in the context of a simplified medical diagnosis task. Time points T1-T4 represent the presentations of the WM Probe Arrays where eye movements are measured.

We now present an experiment deploying the MAC procedure to investigate the time course of memory retrieval in the context of a hypothesis generation task. The task is explained to the participants as a "Cause and Effect learning task" in which they are to learn associations between colors, some representing Causes and some representing Effects emanating from those Causes. Thus, the present task contains the essential structure for a hypothesis generation task in which one reasons from events (Effects) to explanations (Causes).

## Deploying the MAC Procedure

In this experiment we test the efficacy of the MAC procedure to detect the retrieval of a likely hypothesis into working memory and its sensitivity to retrieval timing.

Participants Twenty-three participants from the University of Oklahoma participated for course credit.

Apparatus, Stimuli, and Procedure Eye movements were recorded monocularly (dominant eye) via an Eye Link 1000 (SR Research) at a sampling rate of 1000 Hz and a distance of 60 cm from a 17" monitor. Stimulus presentation and data recording were controlled via Experiment Builder. A ResponsePixx box was used to collect manual responses during the experiment. Eight colors were used during the experiment (blue, green, orange, purple, red, salmon, white, and yellow). Gray was used as the background color throughout the experiment and the fixations were black. Prior to the start of the experiment, the colors were randomly assigned as causes, effects, or distractors.

The experiment consisted of two main phases, a training phase in which the participants learned probabilistic relationships between the Causes and Effects followed by an elicitation phase in which the MAC procedure was deployed. Training consisted of two parts, passive exemplar training and active exemplar training in what could be considered as a "probabilistic paired-associates category
learning task". Participants first went through the passive training portion which was followed by active training and the entirety of the training phase constituted four repeated pairings of passive and active blocks. In passive training, the participant was presented with many individual exemplars in which a single "Cause" and "Effect" pairing with an arrow going from the Cause towards the Effect. Each exemplar appeared for $1,500 \mathrm{~ms}$ after which point the participant pressed the response box to view a new Cause and Effect exemplar. There were four screen configurations in which the pairing could appear and these were randomly selected on each trial to ensure that the Causes and Effects were balanced across spatial locations.

During active training, the participant was presented with an exemplar in which the Cause was absent and the participant had to select the likely Cause with a manual button press. The participant then received feedback (correct/incorrect) for each trial and was shown the correct Cause on incorrect trials. For the first block of active training, participants had $3,000 \mathrm{~ms}$ to respond with the related Cause and this decreased to $1,500 \mathrm{~ms}$ for the second, third, and fourth blocks.

The statistical associations between the Causes and Effects were dictated by the values in Table 1. Note, Effect 1 was highly diagnostic of Cause 1 and Effect 2 was highly diagnostic of Cause 2 (while Effects 3 and 4 were nondiagnostic). For example, there is a $90 \%$ chance that Effect 1 will be present given Cause 1 as described in Table 1, therefore when Effect 1 is observed it is highly likely that Cause 1 is responsible. Additionally, it is important to note that Effect 1 and Effect 2 were complementary with one another as were Effects 3 \& 4. For instance, in medical diagnosis context Effect 1 could represent "fever" and Effect 2 would represent "no fever".

Table 1: The Cause-Effect contingency table governing the associations between the Causes and Effects.

|  | Effect 1 | Effect 2 | Effect 3 | Effect 4 |
| :--- | :---: | :---: | :---: | :---: |
| Cause 1 | 0.9 | 0.1 | 0.9 | 0.1 |
| Cause 2 | 0.1 | 0.9 | 0.9 | 0.1 |

The elicitation phase, in which we deployed the MAC procedure (and recorded eye movements), commenced following the fourth round of active training. In this phase, participants were instructed that on each trial they would be presented with an Effect and would have to respond (manually with left/right button press) to select the most likely Cause given the effect. On $2 / 3$ of trials a WM Probe Array was briefly presented for 396 ms containing four colored disks (top to center $=15 \mathrm{~mm}$ and right to center $=$ 14). Two of the colors were those of the Causes and the two other colors were those assigned as distracter colors at the beginning of the experiment (which had not appeared at any point prior in the experiment). These four colored disks were positioned around a circle (unseen) with a radius of 86 mm . Relative to a clock face one disk appeared at 1 or 2 o'clock, another at 4 or 5 o'clock, another at 7 or 8 o'clock, and the last at 10 or 11 o'clock. Each of the four items
(Causes \& Distractors) were randomly assigned to these four positions in the WM probe array.


Figure 2: Trial schematic demonstrating the sequence of events for trials on which the WM Probe Array appeared.

The focal independent variable was the timing of the WM Probe Array on the trials in which an array appeared. The WM Probe Arrays were manipulated to appear at a variable SOA following the onset of the Effect. For the Short SOA condition, the ISI with a fixation cross was presented for a brief duration ( 48 ms ) and for the Long SOA condition the fixation was presented for a longer duration ( 600 ms ). The relative duration of the second fixation cross was inverted from the duration of the first fixation ( 600 ms for Short SOA trials and 48 ms for Long SOA trials). On the remaining third of trials, no WM Probe Array appeared. These trials were included to limit the expectation of the WM Probe Array's appearance throughout elicitation. On these "no-probe trials", the Effect was followed by a fixation for 1092 ms prior to the selection screen. Thus, total trial time was equal across all trials. Participants completed 36 trials (12 Short SOA, 12 Long SOA, and 12 No-Probe trials). Within each of these conditions, three trials occurred with each Effect. Participants were not informed of WM probe array's appearance nor were they provided instruction for any response when it appeared.

Although Effects 3 and 4 were presented in the elicitation phase, we were not concerned with these trials as these Effects were non-diagnostic. Effects 1 and 2, on the other hand, were highly diagnostic and, accordingly, it is on these trials where our interest and predictions fall. On these trials, the likely hypothesis should be retrieved into WM and bias eye movements towards its matching representation in the WM probe array through top-down capture. Thus, we hypothesized that participants would fixate the likely hypothesis first more often than the unlikely hypothesis and distractors. Additionally, we hypothesized that a difference in the time course of the generation process might emerge between the two SOA conditions as a result of the time pressure applied in the active training.

## Results

For eye movement analyses, regions of interest (ROIs) were centered on each colored disk appearing in the WM Probe Arrays measuring 34.5 mm top to center and 32 mm right to center. A disk was considered fixated when a fixation landed in its corresponding ROI. For analysis we took our primary DV as the first WM Probe Array disk fixated. Only trials in which participants were presented with a diagnostic

Effect (i.e., Effect 1 or 2) were analyzed. Trials on which the participant selected the less likely Cause at the end of the trial were considered as incorrect trials and discarded prior to analysis ( $24 \%)^{1}$. Two additional criteria were applied to each trial for inclusion in the analysis 1) the participant must have been fixating within an ROI at the center of the screen at the beginning of the trial ( 32 mm tall x 34 mm wide) and 2) an item in the array must have been fixated. An additional $8 \%$ of the total trials were discarded for central fixation criterion and an additional $37 \%$ of the total trials were discarded for the array item fixation criterion.

As displayed in Figure 3 the likely Cause was more often fixated first than the unlikely Cause and distractors in the Short SOA condition, $\mathrm{z}=4.3, p<0.001$, and $\mathrm{z}=4.96, p<$ 0.001 , as well as in the Long SOA condition, $\mathrm{z}=1.91, p<$ 0.056 , and $\mathrm{z}=3.45, p<0.001$ (although this difference was marginal between the likely and unlikely Cause). More importantly, logistic regression revealed that the likely Cause was more often fixated first in the Short SOA condition than in the Long SOA condition, $\chi^{2}(1)=5.92, p<$ .05. No such differences were found for the unlikely Cause, $\chi^{2}(1)=1.36, p=.24$, or distractors, $\chi^{2}(1)=2.9, p=.08$.


Figure 3: Proportion of trials on which each item type was the first array item fixated. Results demonstrate increased fixation of the Likely Cause at the Short SOA relative to the
Long SOA and greater fixation rates for the Likely Cause relative to the Unlikely Cause and Distractors.

## Discussion

We have developed a novel methodology to non-intrusively measure the content of WM in complex cognitive tasks as they unfold over time. Here we deployed the MAC procedure in the context of a hypothesis generation task in which participants retrieved a hypothesis from LTM based on the presentation of an associated cue. Our procedure shares the aims of the multitude of process tracing approaches that have been developed over the last thirty years - to assess moment by moment cognitive dynamics

[^132]and changes in the representations utilized en route to final task output. By capitalizing on the tight connection between WM content and attentional allocation via top-down capture (Soto, Heinke, Humphreys, \& Blanco, 2005; Soto \& Humphreys, 2007), we have developed a new method of such assessment. Moreover, by designing our procedure to assess WM content briefly and on task-irrelevant arrays, we have aimed to develop a procedure that will be less interfering to the processes under investigation than traditional processing measures which essentially constrain the participant with a dual-task (see Russo, 1978; Russo, Johnson, \& Stevens, 1989).

Two important effects manifest in the present experiment: 1) The Likely Cause was most often fixated first relative to the other items in the WM probe arrays, and 2) There was an effect of SOA such that the likely Cause was more likely to be fixated at the shorter SOA. It has previously been suggested (Makovski \& Jiang, 2008) that biases towards WM matching content, as revealed through RTs, are sensitive to the representational strength of the WM content. Additionally, Lange, Thomas, Buttaccio, \& Davelaar (2012) provide preliminary evidence that eye movements are sensitive to WM activation. We interpret the present effect of SOA for the likely hypothesis as demonstrating differences in the memory activation (i.e., representational strength) possessed by the likely hypothesis between the short and long SOAs. The present results suggest that shortly after a hypothesis is retrieved into WM, it undergoes a brief period of heightened activation before moving into a decreased state of activation as it is maintained for output.

We refer to this initial heighted stage as a "retrieval input" stage as it is the act of retrieval from LTM that endows the hypothesis with this heightened activation state. We refer to the following stage of decreased activation as a "maintenance" stage as the hypothesis is being maintained in WM for eventual overt output. This account is readily captured by the context-activation model (Davelaar et al., 2005) which we have recently incorporated into a temporally dynamic model of hypothesis generation (Lange, Thomas, \& Davelaar, 2012). In the context-activation model, the memory activation of an item at each time step is determined by the item's activation on the previous time step, self-recurrent excitation that it recycles onto itself, inhibition from the other active items, external input, and noise ${ }^{2}$. Besides external input, the model can also be excited by information retrieved from LTM and the model readily produces the trend we see in the fixation data at the short and long SOAs. As demonstrated in Figure 4, when the model is provided "retrieval input" for 500 iterations, which is then removed for the final 500 iterations, the trend evidenced in the data is produced. Although the focus of this paper is not in modeling the empirical data, it is encouraging to see that a crucial component of our theoretical framework accounts for the data with such ease.

[^133]

Time (model iterations)
Figure 4: Context-Activation Model account of the SOA difference observed in Experiment 2 for the likely
hypothesis. Shortly after retrieval, the hypothesis enjoys a brief period of heightened activation in WM (measured at the short SOA) before moving into a less activated maintenance state (measured at the long SOA) prior to output.

Two related and recently developed methodological approaches deserve further consideration. Mehlhorn, Taatgen, Lebiere, and Krems (2011) used a lexical decision task to measure memory activation of candidate hypotheses in a diagnostic reasoning task. By interspersing the lexical decision task (yes/no response to indicate "hypothesis or not") at different time points in a diagnostic reasoning task they were able to draw conclusions regarding memory activation by assessing the relative speed with which the lexical decision was made for the various hypotheses of interest. This procedure and the MAC procedure share an emphasis on quickly assessing the content of memory with a brief "probe" presented to the participant. However, as with traditional process tracing, this modified lexical decision procedure requires a secondary (albeit not entirely concurrent) task in addition to the primary task of interest. Despite this difference, we believe the procedure of Mehlhorn et al. (2011) to be highly complementary to ours.

Also of note is the "memory indexing" technique of Renkewitz and Jahn (2012) capitalizing on the phenomenon of looking-at-nothing (Ferreira, Apel, \& Henderson, 2008). By holding the spatial locations of the task relevant information constant throughout the experiment, they were able to use eye movements relating to the spatial locations of this information in the testing phase as an index of what was actively being considered across time in the task (despite the fact that the screen was mostly blank as this data was collected and the participants were looking at nothing). This procedure has been successfully utilized to investigate multi-attribute choice (Renkewitz \& Jahn, 2012) as well as diagnostic reasoning (Jahn \& Braatz, 2012). Each of these three procedures (memory activation capture, modified lexical decision, and memory indexing) has their own strengths and weaknesses. By coordinating their utilization within the domain of hypothesis generation and diagnostic reasoning we may be well poised to gain a much deeper understanding of the dynamic memorial underpinnings of these tasks.

Lastly, although we have focused primarily on the domain of hypothesis generation and diagnostic reasoning here, it is important to note that the MAC procedure itself is entirely domain general. Although specific procedural details would need to be tailored for deployment in additional tasks (e.g., specific array stimuli), there is nothing in the logic or mechanics of the procedure that exclude it from use in other domains. We are hopeful that the application of the MAC procedure in domains such as problem solving, multiattribute choice, probability judgment, and hypothesis testing will foster additional insights concerning the cognitive dynamics operating in these domains as well.

## References

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. Cognitive Psychology, 6, 84-107.
Davelaar, E. J., Goshen-Gottstein, Y, Ashkenazi, A., Haarmann, H. J., \& Usher, M. (2005). The demise of short term memory revisited: Empirical and computational investigations of recency effects. Psychological Review, 112(1), 3-42. doi: 10.1037/0033295X.112.1.3
de Groot, A. D. (1965). Thought and choice in chess (1 ${ }^{\text {st }}$ ed.). The Hague: Mouton Publishers.
Dougherty, M.R.P., \& Hunter, J.E. (2003a). Probability judgment and subadditivity: The role of WMC and constraining retrieval. Memory \& Cognition, 31(6), 968982.

Dougherty, M.R.P., \& Hunter, J.E. (2003b). Hypothesis generation, probability judgment, and working memory capacity. Acta Psychologica, 113(3), 263-282.
Dougherty, M.R.P. \& Sprenger, A. (2006). The influence of improper sets of information on judgment: How irrelevant information can bias judged probability. Journal of Experimental Psychology: General, 135(2), 262-281.
Dougherty, M., Thomas, R., \& Lange, N. (2010). Toward an integrative theory of hypothesis generation, probability judgment, and hypothesis testing. In B. H. Ross, B. H. Ross (Eds.), The psychology of learning and motivation: Advances in research and theory (Vol 52) (pp. 299-342). San Diego, CA US: Elsevier Academic Press.
Ellis, J. J., Glaholt, M. G., \& Reingold, E. M. (2011). Eye movements reveal solution knowledge prior to insight. Consciousness and Cognition, 20, 768-776.
Ericsson, K.A. \& Simon, H.A. (1980). Verbal reports as data. Psychological Review, 87, 215- 251.
Ericsson K. A. \& Simon H. A. (1993). Protocol Analysis: Verbal Reports as Data. MIT Press, Cambridge, MA.
Ferreira, F., Apel, J., \& Henderson, J. M. (2008). Taking a new look at looking at nothing. Trends in Cognitive Sciences, 12, 405-410. doi: 10.1016/j.tics.2008.07.007
Franco-Watkins, A.M. \& Johnson, J.G. (2011) Applying the decision moving window to risky choice: Comparison of eye-tracking and mouse-tracing methods. Judgment and Decision Making, 6, 740-749.
Ford, J. K., Schmitt, N., Schechtman, S. L., Hults, B. M., \&

Doherty, M. L. (1989). Process tracing methods:
Contributions, problems, and neglected research questions. Organizational Behavior and Human Decision Processes, 43, 75-117.
Glaholt, M. G. \& Reingold, E. M. (2011). Eye Movement Monitoring as a Process Tracing Methodology in Decision Making Research. Journal of Neuroscience, Psychology, and Economics,4, 125-146.
Glaholt, M. G., Wu, M., \& Reingold, E. M. (2009).
Predicting preference from fixations. PsychNology Journal, 7, 141-158.
Jahn, G. \& Braatz, J. (2012) Memory Indexing of Sequential Symptom Processing in Diagnostic Reasoning. In N. Miyake, D. Peebles, \& R. P. Cooper (Eds.), Proceedings of the 34th Annual Conference of the Cognitive Science Society (pp. 497-502). Austin, TX: Cognitive Science Society.
Lange, N.D., Davelaar, E.J., \& Thomas, R.P. (In Press).
Data Acquisition Dynamics and Hypothesis Generation. Cognitive Systems Research. Invited Submission for Special Issue: Best of International Conference on Cognitive Modeling. doi: 10.1016/j.cogsys.2012.12.006
Lange, N.D, Thomas, R.P., \& Davelaar, E.J. (2012)
Temporal dynamics of hypothesis generation: Influences of data serial order, data consistency, and elicitation timing. Frontiers in Psychology: Cognitive Science 3:215.
Lange, N.D., Thomas, R.P., Buttaccio, D.R. \& Davelaar,
E.J. (2012) Catching a glimpse of working memory: Topdown capture as a tool for measuring the content of the mind. Attention, Perception, \& Psychophysics, 74(8):1562-7. doi: 10.3758/s13414-012-0378-9
Makovski, T. \& Jiang, Y. V. (2008). Indirect assessment of visual working memory for simple and complex objects. Memory \& Cognition, (36)6, 1132-1143. doi: 10.3758/МС.36.6.1132

Mehlhorn, K., Taatgen, N. A., Lebiere, C., \& Krems, J. F. (2011). Memory activation and the availability of explanations in sequential diagnostic reasoning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 37(6), 1391-1411.
Montgomery, H., \& Svenson, O. (1976). On decision rules and information processing strategies in multiattribute decision making. Scandinavian Journal of Psychology, 17, 283-291.
Moores, E., Laiti, L., \& Chelazzi, L. (2003). Associative knowledge controls deployment of visual selective attention. Nature Neuroscience, 6(2), 182-189. doi: 10.1038/nn996

Rehder, B., \& Hoffman, A. B. (2005a). Eyetracking and selective attention in category learning. Cognitive Psychology, 51, 1-41.
Rehder, B., \& Hoffman, A. B. (2005b). Thirty something categorization results explained: Selective attention, eyetracking, and models of category learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 31, 811-829.

Renkewitz, F., \& Jahn, G. (2012, April 30). Memory Indexing: A Novel Method for Tracing Memory Processes in Complex Cognitive Tasks. Journal of Experimental Psychology:Learning, Memory, and Cognition. Advance online publication. doi: 10.1037/a0028073

Russo, J.E., Johnson, E.J., \& Stephens, D.L. (1989). The validity of verbal protocols. Memory \& Cognition, 17, 759-769.
Russo, J.E. (1978) Eye Fixations Can Save the World: A Critical Evaluation and A Comparison Between Eye Fixations and Other Information Processing Methodologies. Advances in Consumer Research, 5, 561570.

Schulte-Mecklenbeck, M., Kühberger, A. \& Ranyard, R. (Eds.). (2011). A Handbook of Process Tracing Methods for Decision Research: A Critical Review and User's Guide. New York: Taylor \& Francis.
Soto, D., Heinke, D., Humphreys, G.W., \& Blanco, M. (2005). Early, involuntary top-down guidance of attention from working memory. Journal of Experimental Psychology: Human Perception and Performance, 31(2), 248-261. doi: 10.1037/0096-1523.31.2.248
Soto, D. \& D., Humphreys (2007). Automatic guidance of attention from verbal working memory. Journal of Experimental Psychology: Human Perception and Performance,33(3),730-757. doi: 10.1037/00961523.33.3.730

Sprenger, A.M., Dougherty, M.R., Atkins, S.M., FrancoWatkins, A.M., Thomas, R.P., Lange, N., \& Abbs, B. (2011) Implications of cognitive load for hypothesis generation and probability judgment. Frontiers in Psychology: Cognitive Science, 2:129. doi: 10.3389/fpsyg.2011.00129

Sutterlin, B., Brunner, T. A., \& Opwis, K. (2008).Eyetracking the cancellation and focus model for preference judgments. Journal of Experimental Social Psychology, 44, 904-911.
Svenson, O. (1979). Process descriptions of decision making. Organizational Behavior and Human Performance, 23, 86-112.
Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., \& Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. Science, 268, 1632-1634.
Thomas,R.P, Dougherty, M.R., Sprenger, A.M., and Harbison, J.I. (2008). Diagnostic hypothesis generation and human judgment. Psychological Review, 115, 155185.

Woodman, G.F. \& Luck, S.J. (2007). Do the contents of working memory automatically influence attentional selection during visual search? Journal of Experimental Psychology: Human Perception and Performance, 3(22), 363-377. doi: 10.1037/0096-1523.33.2.363

# Learning Representations of Animated Motion Sequences - A Neural Model 

Georg Layher (georg.layher@uni-ulm.de)<br>Institute of Neural Information Processing<br>Ulm University<br>James-Franck Ring, 89069 Ulm, Germany<br>Martin Giese (martin.giese@uni-tuebingen.de)<br>Section for Computational Sensomotorics<br>University Clinic Tübingen<br>Frondsbergstraße 23, 72074 Tübingen, Germany

Heiko Neumann (heiko.neumann@uni-ulm.de)<br>Institute of Neural Information Processing<br>Ulm University<br>James-Franck Ring, 89069 Ulm, Germany


#### Abstract

The detection and categorization of animate motions is a crucial task underlying social interaction and perceptual decisionmaking. Neural representations of perceived animate objects are built in the primate cortical region STS which is a region of convergent input from intermediate level form and motion representations. Populations of STS cells exist which are selectively responsive to specific animated motion sequences, such as walkers. It is still unclear how and to which extent form and motion information contribute to the generation of such representations and what kind of mechanisms are involved in the learning processes. The paper develops a cortical model architecture for the unsupervised learning of animated motion sequence representations. We demonstrate how the model automatically selects significant motion patterns as well as meaningful static form prototypes characterized by a high degree of articulation. Such key poses are selectively reinforced during learning through a cross-talk between the motion and form processing streams. Next, we show how sequence selective representations are learned in STS by fusing static form and motion input from the segregated bottom-up driving input streams. Cells in STS, in turn, feed their activities recurrently to their input sites along top-down signal pathways. We show how such learned feedback connections enable making predictions about future input as anticipation generated by sequence-selective STS cells. Network simulations demonstrate the computational capacity of the proposed model by reproducing several experimental findings from neurosciences and by accounting for recent behavioral data. Keywords: animated motion representation; implied motion; neural model; unsupervised learning; feedback.


## Introduction

Animated movements in actions, like walking, turning, etc., can be robustly detected from video sequence input and predictions about future occurrences can be derived from such spatio-temporal patterns. Giese \& Poggio (Giese \& Poggio, 2003) proposed a hierarchical feedforward network architecture that aims at explaining the computational mechanisms underlying the perception of biological motion, mainly from impoverished stimuli such as point-light walkers. In this paper, we propose a new learning-based hierarchical model for analyzing animated motion sequences. Prototypes in the form and motion pathways are established using a modified Hebbian learning scheme. We suggest how snapshot prototypes are automatically selected from continuous input video streams utilizing features from the motion pathway which are indicative for the occurrence of specific snapshots with strongly articulated configurations, serving as key
poses. Sequence-selective representations of articulated motions in cortical STS are driven jointly by input activations from both motion and form prototypes. In addition, feedback connections are learned to enable STS neurons predicting expected input from form selective IT and motion sensitive MST. We argue that for inputs presenting articulated postures without continuing motion, STS representations are fed by the corresponding snapshot prototype activations (Jellema \& Perrett, 2003). In turn, STS will send feedback to stages in the segregated pathways for form as well as motion processing. Stationary images which depict articulated postures, consequently generate effects of implied motion, which have been shown in functional magnetic resonance imaging (fMRI) studies (Kourtzi \& Kanwisher, 2000). We will argue here, that this can be accomplished by the proposed model through the action of fusing bottom-up input, driven by snapshot representation only, and the activated sequence representations sending feedback to both form and motion representations, thus amplifying motion representations even if no direct motion input is present.
Several computer vision approaches have been proposed for performing action recognition using different processing strategies of combining form and motion information. These approaches build upon the hierarchical architecture proposed by Poggio and coworkers which aims at defining a framework for form processing in the cortical ventral pathway (Riesenhuber \& Poggio, 1999). Extensions of the form processing model to analyze motion information responses in a separate pathway, like the Giese-Poggio model, have been suggested in e.g. (Schindler \& Van Gool, 2008). Here, the relative contributions of form and motion features to the classification of actions have been investigated. Details of the motion processing cascade alone have been studied in more detail in (Escobar \& Kornprobst, 2012). Here the authors contributed further evidence that detecting motion contrasts in sequences of animated motion is useful to distinguish action classes. In all these proposed models, the mechanisms for hierarchical motion (and form) processing are predefined and learning only occurs at the level of a final classifier to distinguish given categories. It still remains unclear to a large extent, how the motion and form prototypes (e.g., in cortical areas MST and IT, respectively) and the sequence-selective pattern representations in STS interact and which features are


Figure 1: Overview of the model architecture. The model consists of two separate processing streams, the motion and the form pathway, both converging into model area STS. Static form prototypes in area IT, as well as optic flow patterns in area MST are learned using an unsupervised Hebbian mechanism. A motion driven reinforcement signal between the two pathways is used to steer the learning of the IT prototypes. After the suppression of cells with low activities, IT and MST cells propagate into area STS, where sequenceselective cells learn corresponding spatio-temporal activity patterns using a similar Hebbian learning rule. In addition, the sequence-selective cells learn the output weights back to the segregated form and motion prototypes, that stabilizes the input processing and activity fusion.
used for learning. How can feature representations be learned automatically from given input streams at different levels of the distributed action sequence representations? Also no topdown influences have been considered so far and how such connectivity patterns may transfer different information between pathways to generate proper predictions concerning future input configurations.

## Model Architecture

The hierarchical model proposed here consists of two separate visual pathways for segregated form and motion processing as inspired by the work of (Giese \& Poggio, 2003) and extends it by combining it with models for the hierarchical feedforward and feedback processing of motion and form along the dorsal and the ventral pathway (Bayerl \& Neumann, 2004; Weidenbacher \& Neumann, 2009). Intermediate level form representations (in model IT) and prototypical optical flow patterns (in model MST) are established using a modified competitive Hebbian learning scheme with convergent weight dynamics. The two separate hierarchical learning approaches are influenced partly by the work of Rolls and
collaborators (Rolls \& Milward, 2000), in which the authors have suggested that layered neuronal structures arranged in a hierarchy with increasingly larger connectivity kernels can learn invariant representations of objects and specific motion patterns. Here, we propose how such learning in a hierarchy can be utilized for learning sequence-selective representations of animated movement prototypes from convergent form and motion input. In addition, we suggest how a motiondriven reinforcement mechanism automatically selects relevant snapshots in the form path from video input streams. The activities of the prototypical form and motion cells converge in the model complex STS, where correlated temporal activations for specific sequences are learned. Sequence-selective representations are established by combined bottom-up and top-down learning, both based on modified Hebbian mechanisms. An overview of the model is shown in Fig. 1. The details are outlined below.

## Form and Motion Processing

Processing the raw input data utilizes an initial stage of orientation and direction selective filtering (in model area V1). These responses are fed into separated pathways which are selective to static form representations (areas V2 and IT) and characteristic optical flow patterns (areas MT and MST). We use single compartment model neurons with gradual activation dynamics. The membrane potential of individual model neurons is calculated by conductance-based mechanisms of feed-forward integration of excitatory and inhibitory feeding input and a passive leakage. The potential can be enhanced by a gating mechanism to amplify the efficacy of the current potential by a matching top-down feedback signal. The membrane potential is finally regulated by a gain control mechanism that leads to activity normalization for a pool of neurons through mutual divisive inhibition. These mechanisms are summarized in a three-stage hierarchy of processing that includes input filtering, modulatory feedback, and pool normalization. The output of a cell is defined by a signal function which converts the membrane potential into a firing rate, or activity. Such model cells are grouped into layers which form abstract models of cortical areas.

## Learning of Form and Motion Prototypes

First, we investigated how intermediate level feature representations can be learned in a biologically plausible fashion by exposing the network architecture with realistic input sequences. In order to generate feature representations of complex form and motion patterns we employ an unsupervised learning mechanism based on a modified Hebbian learning scheme. The modification stabilizes the learning such that the growing of weight efficacies is constrained to approach (bounded) activity levels of the input or the output activation. Motivated by the invariance properties observed by (Wallis \& Rolls, 1997) we combined the modified Hebbian learning mechanism with a short-term memory trace of prolonged activity of the pre- or the post-synaptic cells (trace rule). The adaptation of weightings is controlled by post-synaptic cells
which, in turn, mutually compete for their ability to adjust their incoming connection weights. The particular details as well as the particular variations of the core architecture are explained below.

Hebbian learning in the form and motion pathways. In order to select the image regions that are fed to the learning of prototype representations a region of interest (ROI) is defined which represents a bounding box around the target object. Features within the target region are selected for learning feedforward connection weights in the form and the motion pathway, respectively. We employ the modified Hebbian learning rule

$$
\begin{equation*}
\Delta w_{j i}^{F F, s}=\eta_{s} \cdot \bar{v}_{i}^{p o s t} \cdot\left(u_{j}^{p r e}-\bar{v}_{i}^{p o s t} \cdot w_{j i}^{F F, s}\right) \tag{1}
\end{equation*}
$$

where $\Delta w_{j i}^{F F, s}$ represents the discretized rate of change in the efficacy of the weighted connections with the learning rate $\eta_{s} ; s \in\{$ form,motion $\}$ indicates that the same core mechanisms are devoted to learning in the form and motion pathway, respectively. The variables $u_{j}^{p r e}=f\left(x_{j}\right)$ and $v_{i}^{\text {post }}=f\left(y_{i}\right)$ are the firing rates driven by the membrane potential of pre- and post-synaptic cells, henceforth denoted as activity. The activity $\bar{v}_{i}$ of the post-synaptic cell is calculated by the temporal trace rule $\bar{v}_{i}^{t}=(1-\lambda) \bar{v}_{i}^{t-1}+\lambda v_{i}^{t}, 0<\lambda<1$ (Földiák, 1991). The trace rule (see also (Wallis \& Rolls, 1997; Rolls \& Milward, 2000)) has been proposed to incorporate a short-term memory function for the cells to keep their activation over a short temporal window while adapting their weights. The term in brackets on the r.h.s. of learning equation 1 serves as a biologically plausible mechanism to bound the growth of the cells' input synaptic weights (Oja, 1982). The post-synaptic cells (with activity $\bar{v}_{i}^{\text {post }}$ ) which gate the learning of their respective input weights are arranged in a competitive layer of neurons competing for the best matching response and their subsequent ability to adapt their kernel of spatial input weights. In a nutshell, the layer of post-synaptic neurons competes to select a winning node for a given input presentation which, in turn, is allowed to automatically adapt their incoming (instar) synaptic weights. The temporal trace (or short-term memory) establishes that categories learn their average input over a short temporal interval thus allowing small pertubations for the changing input signals.

Reinforcing snapshot learning. The Giese-Poggio model (Giese \& Poggio, 2003) suggests that sequence selectivity for biological motion recognition is driven by sequences of static snapshots. While the original model relies on snapshots that were regularly sampled temporally, we suggest a mechanism of how snapshots corresponding to strongly articulated poses can be selected automatically. Such snapshot representations are learned in the form channel by utilizing a gating reinforcement signal which is driven by the complementary representation of motion in the dorsal stage MT/MST. Formally, the weighted integration of motion energy over a given neighborhood is calculated by

$$
\begin{equation*}
m_{e}=\int_{\Omega} u_{\phi}(\mathbf{x}) \cdot \Lambda(\mathbf{x}) d \mathbf{x} d \phi \tag{2}
\end{equation*}
$$



Figure 2: IT prototypes trained using disabled and enabled reinforcement signal. Minima and maxima in motion energy correspond to articulated and non-articulated postures (bottom left). Continuous learning of IT prototypes leads to activation profiles with low selectivity (top right). Motion driven reinforcement leads to IT prototypes which signal snapshot poses in synchrony with the gait (bottom right; for details, see text).
with $\Lambda(\bullet)$ denoting a spatial kernel for weighting the relative contribution of motion responses $u_{\phi}(\bullet)$ at spatial locations $\mathbf{x}$ in the 2-D image plane and with direction selectivity $\phi .{ }^{1}$ The motion energy signal itself is a function of time which is used to steer the instar learning in the form pathway. We suggest that different subpopulations of static form, or snapshot, representations can be learned that correspond to either weakly or strongly articulated postures. Here, we focus on snapshot poses corresponding to highly articulated postures with signatures of maximum limb spreading. Motion energy at limbs drops during phases of high articulation when their apparent direction of motion reverses. We incorporate the function $g(\bullet)$ to control a vigilance in snapshot learning to favor form inputs which co-occur with local motion energy minima, i.e. when $\partial_{t} m_{e}=0$, given that $\partial_{t t} m_{e}>0$. In the weight adaptation, $\Delta w_{j i}^{F F, f o r m}$ in Eqn.1, the learning rate is now gated by the motion dependent reinforcement, $\eta_{\text {form }} \cdot g\left(m_{e}\right)$ which leads to the revised learning rule

$$
\begin{equation*}
\Delta w_{j i}^{F F, f o r m}=\eta_{\text {form }} \cdot g\left(m_{e}\right) \cdot \bar{v}_{i}^{p o s t} \cdot\left(u_{j}^{p r e}-\bar{v}_{i}^{\text {post }} \cdot w_{j i}^{F F, \text { form }}\right) . \tag{3}
\end{equation*}
$$

## Learning of Sequence-Selective Representations

Categorial representations in the form and motion pathway, namely in IT and MST, which were learned at the previous stage, feed forward their activations to the stage of STS. In order to stabilize the representations and activity distributions, even in the case of partial loss of input signals, the STS sequence-selective representations send top-down signals to their respective input stages.

[^134]

Figure 3: Response behavior of IT snapshot neurons, MST motion pattern neurons, and sequence-selective STS cells trained by video input for a walker moving from left to right. Activations in the model areas are shown for different input conditions for recall of the training sequence (top), opposite walker movement (middle), and walker displayed in reverse motion (bottom). Line styles and colors encode the input test cases on the right. For details and brief discussion, see text.

Learning of feedforward connections. Prototypical representations with spatio-temporal sequence selectivity are suggested to exist in the cortical STS complex where both form and motion pathways converge. The selectivities of model STS neurons are learned by again using a modified Hebbian instar learning mechanism similar to the separate learning of form and motion prototypes (Eqn.1),

$$
\begin{equation*}
\Delta w_{j i}^{i n, F F}=\eta_{s e q F F} \cdot \bar{v}_{i}^{p o s t} \cdot\left(u_{j}^{p r e}-\bar{v}_{i}^{p o s t} \cdot w_{j i}^{i n, F F}\right) . \tag{4}
\end{equation*}
$$

The weighting kernel $w_{j i}^{i n, F F}$ represents convergent IT $\rightarrow$ STS and MST $\rightarrow$ STS bottom-up input to a post-synaptic STS cell (instar). $\eta_{\text {seqFF }}$ denotes the learning rate and $u_{j}$ and $v_{i}$ are the firing rates of the pre- and post-synaptic neurons, respectively (the post-synaptic activity is again calculated via a temporal trace mechanism). The pre-synaptic activity for the receiving model STS cells are generated by concatenating form and motion output activations, namely $\mathbf{u}=\mathbf{u}^{I T} \cup \mathbf{u}^{M S T}$.

Learning feedback connections. An important component is that sequence-selective prototypes in STS in turn learn the output weights back to the segregated form and motion prototype representations, namely STS $\rightarrow$ IT + MST. Unlike the FF learning mechanisms, the learning here is gated by the pre-synaptic cell (in STS) for their top-down weights, which reads

$$
\begin{equation*}
\Delta w_{j i}^{o u t, F B}=\eta_{s e q F B} \cdot \bar{v}_{i}^{p r e} \cdot\left(u_{j}^{\text {post }}-w_{j i}^{\text {out }, F B}\right) \tag{5}
\end{equation*}
$$

with the same components as in the bottom-up learning formalism in Eqn.4. Bottom-up and top-down learning schemes slightly differ in the definition of the competitive terms (in brackets). In the feedback learning we employ a difference term between post-synaptic activity and the weighting, $u_{j}^{\text {post }}-w_{j i}^{\text {out }, F B}$, omitting the additional weighting of the connectivity strength via the pre-synaptic activity as in the Oja
rule. In steady-state each of the connection strengths emanating from STS cells assumes a value corresponding to the post-synaptic activity distribution, which defines the current input activation. Given an STS cell with attraction $\bar{v}_{i}^{p r e}$, the top-down weight vector approaches $\mathbf{u}^{\text {post }}=\mathbf{w}_{i}^{\text {out }, F B}$, thus learning the expected average input. Combined with the temporal trace, this establishes a representation in which each STS sequence-selective prototype encodes and memorizes in its weight pattern the expected driving input activity pattern configuration from the form and motion pathway. Such a topdown weighting pattern can then be used to generate predictions concerning the expected future input given the current maximally activated prototype at the STS level.

## Results

The model has been tested in various computational experiments, not all of which we can present here. In a first experiment, we probed the properties of snapshot selection from the input streams and their signature concerning static articulations. The latter property has been motivated by the fact that extremal articulation indicates configurations of implied motion, in turn, predictive for future motions. Results shown in Fig. 2 demonstrate that input activations (in V2) with strongly articulated shapes cohere with local motion minima. Such minima drive the reinforcement signal for learning whole body form prototypes. Temporal response signatures for IT prototypes are shown for disabled reinforcement $\left(g\left(m_{e}\right)=1\right.$, and when it is enabled ( $g\left(m_{e}\right)$ monotonically decreasing function of $m_{e}$ ).
We studied the response properties of STS representations and their motion sequence selectivity. There, a prototypical sequence-selective representation is learned for a walker that is traversing from left to right. After training of form, motion and sequence representations, the network is probed by


Figure 4: Response tunings of models cells in area IT (snapshots), MST (motion patterns), and STS (sequence-selective patterns) after training. Category representations have been learned for a walker moving along horizontal direction for $\phi=0^{\circ}$. Activities of prototypical cells are shown (bottom) which were probed by different inputs with varying movement directions, i.e. walkers approaching or receding at different angles with respect to the horizontal reference axis (top). Data has been summarized into box plots showing the response variabilities of models cells as well as the monotonic decline in response for deviations from the target tuning. The tuning width at half maximum response is around $\pm 40^{\circ}$. The variance of the MST / IT prototypes decreases towards larger deviations, depicting the loss of response selectivity of prototypes to different parts of a walkers gait.
three different movement scenarios: a forward moving walker with same profile and movement direction as in the training phase (recall), a forward moving walker traversing from right to left (opposite), and a backward moving walker (reverse). Form/motion prototypes and the sequence representation are triggered maximally in the recall case while in the opposite case form and motion prototypes only respond minimally, and so do the sequence-selective cells. In the reverse case the form prototypes selectively match the input at high articulation configurations, while the motion responses remain minimal. As a consequence, the sequence-selective representations respond at an intermediate level (Fig. 3). This evidence is in line with the experimental findings by (Oram \& Perrett, 1996) and recent observations by (Singer \& Sheinberg, 2010).

We further investigated the direction tuning of the sequence-selective prototypes. Here, we configured different walkers with varying movement directions and speeds with reference to a previously learned representation of a rightward moving walker at a speed of $1 \mathrm{~m} / \mathrm{s}$. Walking directions in the test cases were rotated by $\pm\left\{5^{\circ}, 10^{\circ}, 20^{\circ}, 40^{\circ}\right\}$. Model simulations result in a direction tuning of STS cells with half amplitude of approximately $\pm 40 \mathrm{deg}$ (Fig. 4). IT and MST cells, on the other hand, also show a drop in response but have


Figure 5: Selective removal of interconnections (lesioning). The model was trained using the same walking sequence as in the second experiment (see Fig. $3 /$ forward recall). The model was left untouched to provide a reference (top). Bottom-up (feedforward) connections between area MST and STS were removed, preventing any motion-related signal being propagated to STS (bottom). The amplitude of the IT prototype activities remains almost the same, whereas the sequence-selective STS cell responds only at about halfmagnitude (because of the missing support from the motion pathway). Note the feedback activities propagated from STS to MST optical flow pattern prototypes. We argue that this reflects the induction of increased fMRI BOLD response in human MT+ following the presentation of static implied motion stimuli.
a much larger variability.
In an additional experiment we selectively lesioned of the model architecture, particularly investigating the effects of extinguishing connections between model areas and the activity flow between learned representations (Fig.5). The fully connected model with learned IT / MST and STS feedforward and feedback connections was used as reference. When bottom-up connections from motion input (MST) were cut off the sequence-selective neuron responses in STS drop to approximately half their response amplitude. Feedback from STS invokes an amplification of activities in IT and MST representations. We observe that FF activation from IT alone can drive sequence neurons. Snapshot representations in IT drive the STS sequence neurons which, in turn, send feedback signals to the stages of IT and MST prototype representations. In the motion pathway such feedback elicits an increase in presynaptic activation. We argue that this reflects the induction of increased fMRI BOLD response in human MT+ following the presentation of static implied motion stimuli (Kourtzi \& Kanwisher, 2000).

## Discussion and Conclusion

We propose a biologically plausible model for the learning of animated motion sequences. The model builds upon neurophysiological evidence about the cortical sites and specific neuronal representations which contribute to articulated motion and implied motion perception. The main contributions of the paper are several-fold: First, we suggest how prototype representations in the form and motion pathways, namely in model cortical areas IT and MST, can be established on the basis of probing the model architecture by sequences containing animated motions. Learning mechanisms are based on modified Hebbian schemes which are stabilized through a trace mechanisms and the incorportion of an objective function taking the weight kernel saturation into account. Second, we suggest that sequence-selective cells in model area STS are learned by using the same learning mechanisms but now by combining the responses of intermediate level representations in the form and motion pathways. Third, the learning of articulated poses (snapshots) is controlled by a reinforcement mechanism that enables Hebbian learning in the form pathway through cross-pathway motion-form interaction. Given an animated motion sequence, snapshots are automatically selected as key poses corresponding to strong body pose articulations. Finally, the sequence-selective cells in model STS project to their respective input representations in the form and motion pathways. These feedback connections are again learned by a Hebbian mechanism. Together, the feedforward and the feedback interactions establish a loop of recurrent processing to stabilize the patterns of form, motion, and sequence representation. Via feedback, model STS cells generate a predictive signal through the backward connections' weights to encode the expected matching input that is suitable to match the currently activated sequence pattern. Together with the newly proposed feedback mechanism the model is able to account for various experimental findings, in particular, the ability to infer and predict future motion sequence development from articulated postures (implied motion). Importantly, cells in STS are responsive to both motion as well as static form (Oram \& Perrett, 1996). The model predicts that the presentation of static key poses from previously learned sequences alone leads to enhanced activation in STS sequence selective neurons as observed in (Jellema \& Perrett, 2003). The model also hypothesizes how the presentation of static articulated poses leads to the emergence of predictive motion perception and enhanced neural activations in the motion pathway (Kourtzi \& Kanwisher, 2000). Furthermore, learned sequence-selective prototype representations have direction tunings in response to walkers in the range of $\pm 40$, similar to those reported in (Perrett et al., 1989)). Once again, the model makes a testable prediction that articulated poses represent the snapshot frames that have been suggested by (Giese \& Poggio, 2003) and that have recently been tested experimentally by (Singer \& Sheinberg, 2010).

## Acknowledgements

GL and HN have been supported by the SFB Transregio 62 funded by the German Research Foundation (DFG).

## References

Bayerl, P., \& Neumann, H. (2004). Disambiguating visual motion through contextual feedback modulation. Neural Comput, 16(10), 2041-2066.
Escobar, M., \& Kornprobst, P. (2012). Action recognition via bio-inspired features: The richness of center-surround interaction. Comput Vis Image Und, 116(5), 593-605.
Földiák, P. (1991). Learning invariance from transformation sequences. Neural Comput, 3(2), 194-200.
Giese, M., \& Poggio, T. (2003). Neural mechanisms for the recognition of biological movements. Nat Rev Neurosci, 4(3), 179-192.
Jellema, T., \& Perrett, D. (2003). Cells in monkey STS responsive to articulated body motions and consequent static posture: a case of implied motion? Neuropsychologia, 41(13), 1728-1737.
Kourtzi, Z., \& Kanwisher, N. (2000). Activation in human MT/MST by static images with implied motion. J Cognitive Neurosci, 12(1), 48-55.
Oja, E. (1982). Simplified neuron model as a principal component analyzer. J Math Biol, 15(3), 267-273.
Oram, M., \& Perrett, D. (1996). Integration of form and motion in the anterior superior temporal polysensory area (STPa) of the macaque monkey. J Neurophysiol, 76(1), 109-129.
Perrett, D., Harries, M., Bevan, R., Thomas, S., Benson, P., Mistlin, A., ... Ortega, J. (1989). Frameworks of analysis for the neural representation of animate objects and actions. J Exp Biol, 146(1), 87-113.
Riesenhuber, M., \& Poggio, T. (1999). Hierarchical models of object recognition in cortex. Nat Neurosci, 2, 10191025.

Rolls, E. T., \& Milward, T. T. (2000, November). A model of invariant object recognition in the visual system: Learning rules, activation functions, lateral inhibition, and information-based performance measures. Neural Comput, 12(11), 2547-2572.
Schindler, K., \& Van Gool, L. (2008). Action snippets: How many frames does human action recognition require? In CVPR 2008 (pp. 1-8).
Singer, J., \& Sheinberg, D. (2010). Temporal cortex neurons encode articulated actions as slow sequences of integrated poses. J Neurosci, 30(8), 3133-3145.
Wallis, G., \& Rolls, E. (1997). Invariant face and object recognition in the visual system. Prog Neurobiol, 51(2), 167-194.
Weidenbacher, U., \& Neumann, H. (2009). Extraction of surface-related features in a recurrent model of V1-V2 interactions. PloS ONE, 4(6), e5909.

# Modeling disambiguation in word learning via multiple probabilistic constraints 

Molly Lewis<br>mll@stanford.edu<br>Department of Psychology<br>Stanford University

Michael C. Frank<br>mcfrank@stanford.edu<br>Department of Psychology<br>Stanford University


#### Abstract

Young children tend to map novel words to novel objects even in the presence of familiar competitors, a finding that has been dubbed the "disambiguation" effect. Theoretical accounts of this effect have debated whether it is due to initial constraints on children's lexicons (e.g. a principle of mutual exclusivity) or situation-specific pragmatic inferences. We suggest that both could be true. We present a hierarchical Bayesian model that implements both situation-level and hierarchical inference, and show that both can in principle contribute to disambiguation inferences with different levels of strength depending on differences in the situation and language experience of the learner. We additionally present data testing a novel prediction of this probabilistic view of disambiguation.


Keywords: Word learning; mutual exclusivity; Bayesian models.

## Introduction

A central property of language is that each word in the lexicon maps to a unique concept, and each concept maps to a unique word (Clark, 1987). Like other important regularities in language (e.g. grammatical categories), children cannot directly observe this general property. Instead, they must learn to use language in a way that is consistent with this generalization on the basis of evidence about only specific word-object pairs.

Even very young children behave in a way that is consistent with the one-to-one mapping between words and concepts in language. Evidence for this claim comes from what is known as the "disambiguation" effect. In a typical demonstration of this effect (e.g. Markman \& Wachtel, 1988), children are presented with a novel and familiar object (e.g. a whisk and a ball), and are asked to identify the referent of a novel word ("show me the dax"). Children in this task tend to choose the novel object as the referent, behaving in a way that is consistent with the one-to-one word-concept regularity in language, across a wide range of ages and experimental paradigms (Mervis, Golinkoff, \& Bertrand, 1994; Golinkoff, Mervis, Hirsh-Pasek, et al., 1994; Markman, Wasow, \& Hansen, 2003; Halberda, 2003; Bion, Borovsky, \& Fernald, 2013).

This effect has received much attention in the word learning literature because the ability to identify the meaning of a word in ambiguous contexts is, in essence, the core problem of word learning. That is, given any referential context, the meaning of a word is underdetermined (Quine, 1960), and the challenge for the world learner is to identify the referent of the word within this ambiguous context. Critically, the ability to infer that a novel word maps to a novel object makes the problem much easier to solve. For example, suppose a child hears the novel word "kumquat" while in the produce aisle
of the grocery store. There are an infinite number of possible meanings of this word given this referential context, but the child's ability to correctly disambiguate would lead her to rule out all meanings for which she already had a name. With this restricted hypothesis space, the child is more likely to identify the correct referent than if all objects in the context were considered as possible referents.

What are the cognitive processes underlying this effect? There are broadly two proposals in the literature. Under one proposal, Markman and colleagues (1988; 2003) suggest that children have a constraint on the types of lexicons considered when learning the meaning of a new word - a "mutual exclusivity constraint." With this constraint, children are biased to consider only those lexicons that have a one-to-one mapping between words and objects. Importantly, this constraint can be overcome in cases where it is incorrect (e.g. adjectives), but it nonetheless serves to restrict the set of lexicons initially entertained when learning the meaning of a novel word. Under this view, then, the disambiguation effect emerges from a constraint on the structure of lexicons.

Under a second proposal, the disambiguation effect is argued to result from online inferences made within the referential context (Clark, 1987; Diesendruck \& Markson, 2001). Clark suggests that the disambiguation effect is due to two pragmatic assumptions held by speakers. The first assumption is that speakers within the same speech community use the same words to refer to the same objects ("Principle of Conventionality"). The second assumption is that different linguistic forms refer to different meanings ("Principle of Contrast"). In the disambiguation task described above, then, children might reason (implicitly) as follows: You used a word I've never heard before. Since, presumably we both call a ball "ball" and if you'd meant the ball you would have said "ball," this new word must refer to the new object. Thus, under this account, disambiguation emerges not from a higherorder constraint on the structure of lexicons, but instead from in-the-moment inferences using general pragmatic principles.

These two proposals have traditionally been viewed as competing explanations of the disambiguation effect. Research in this area has consequently focused on identifying empirical tests that can distinguish between these two theories. For example, Diesendruck and Markson (2001) compare performance on a disambiguation task when children are told a novel fact about an object relative to a novel referential label. They found that children disambiguated in both conditions and argued on grounds of parsimony that the same pragmatic mechanism was likely to be responsible for both inferences. More recent evidence contradicts this
view: tests of children with autism, who are known to have impairments in pragmatic reasoning, find comparable performance on the disambiguation task between typically developing children and children with autism (Preissler \& Carey, 2005; de Marchena, Eigsti, Worek, Ono, \& Snedeker, 2011). This result provides some evidence for the view that disambiguation is due to a domain-specific lexical constraint.

We suggest that this competing-alternatives approach to the disambiguation effect should be reconsidered. In a disambiguation task, learners may be making use of both higherorder knowledge about how the lexicon is structured as well as information about the pragmatic or inferential structure of the task. Both of these constraints would then support children's inferences. In other words, these two classes of theories may be describing distinct, but complementary mechanisms that each contribute to a single empirical phenomenon, with their weights in any given task determined by children's age and language experience, the nature of the pragmatic situation, and other task-specific factors.

The model described here explores this proposal computationally. We constructed a Bayesian model that captures effects of both inferences within individual situations and hierarchical inferences about the structure of lexicons. Inferences about individual situations are modeled using an intentional/pragmatic model of word learning (Frank, Goodman, \& Tenenbaum, 2009), while generalizations about the nature of word-concept mappings are modeled as constraints on the set of lexicons that the model considers. We present a set of simulations and a developmental experiment showing that linguistic experience can influence the strength of disambiguation inferences at both levels.

The goal of our model is not to provide an algorithmic description of children's word learning, which we assume depends on psychological factors such as memory and cognitive control. Instead, we aim to provide an ideal observer analysis: to derive normative predictions given a well-articulated set of assumptions (Geisler, 2003). Human behavior can then be compared to this analysis, and deviations can be attributed to differences between the assumptions of the model and the realities of human psychology. Critically, neither our model (nor comparisons between it and human behavior) constitute claims of human optimality: Though our model employs optimal Bayesian inference, there is no implicit claim that human learners also do so (Frank, in press).

## Model

We model a word learner as performing Bayesian inference to infer the structure of a lexicon $l$, which we represent as a (sparse) bipartite graph connecting words $W=w_{1} \ldots w_{n}$ to objects $O=o_{1} \ldots o_{m}$. We write the full possible set of lexicons as $L$. An example enumeration of such lexicons for the case of $n=m=2$ is given in Fig. 2.

We assume a generative structure identical to the the model developed by Frank et al. (2009), with the added complexity of constraints placed on lexicons (described below; see Fig.


Figure 1: The generative process for our model.
1). The critical feature of this existing model is that words are assumed to be generated by intentions. This feature allows the model to jointly solve the problems of mapping a word to an object in ambiguous contexts and learning a long term mapping between a word and concept.

The learner infers a distribution over lexicons, given a corpus $S$ of situations (each consisting of sets of words $\bar{w}_{s}$ and objects $\bar{o}_{s}$ ). From Bayes' rule, the posterior probability of a lexicon is given by

$$
\begin{equation*}
P(l \mid S)=\frac{P(S \mid l) P(l)}{\sum_{l^{\prime} \in L} P\left(S \mid l^{\prime}\right) P\left(l^{\prime}\right)} \tag{1}
\end{equation*}
$$

We first define the likelihood term $P(S \mid L)$ and then return to the prior $P(L)$, which implements hierarchical constraints $C$ on lexicons.

Using the generative process in Fig. 1, we can write the likelihood of a particular situation in terms of the relationship between the objects that were observed in the situation $s$, the speaker's referential intention $i_{s}$ (a choice to speak about one of the objects), and the referring word $w_{s} .{ }^{1}$ As in our prior work, we assume that referential intentions are unobserved and sum across all possible intentions uniformly ${ }^{2}$ :

$$
\begin{equation*}
P(s \mid l)=\sum_{i_{s} \in \bar{o}_{s}} P\left(w_{s}, o_{s}, i_{s} \mid l\right) \tag{2}
\end{equation*}
$$

By the conditional independence of words and objects, we can expand to:

$$
\begin{equation*}
P(s \mid l)=\sum_{i_{s} \in \bar{o}_{s}} P\left(w_{s} \mid i_{s}, l\right) P\left(i_{s} \mid o_{s}\right) \tag{3}
\end{equation*}
$$

Finally, we aggregate across situations by taking the product of each independent situation:

[^135]

Figure 2: The posterior probability distribution over lexicons for our models for Simulation 1. Models were trained with situations establishing the mapping between $w_{1}$ and $o_{1}$ (the familiar word/object pair) and a disambiguation situation including $w_{2}$ and objects $o_{1}$ and $o_{2}$. The four different constraint models are distinguished by color in the main plot, while the 16 possible lexicons are shown on the horizontal axis. Lexicons are marked as links between words and objects, with the correct ( $w_{2}$ and $o_{2}$ ) mapping marked in green and the incorrect ( $w_{2}$ and $o_{1}$ ) mapping marked in red. The noise parameter $\alpha$ was chosen arbitrarily for display purposes and serves only to scale the results.

$$
\begin{equation*}
P(S \mid l)=\prod_{s \in S} \sum_{i \in \bar{o}_{s}} P\left(w_{s} \mid i_{s}, l\right) P\left(i_{s} \mid o_{s}\right) \tag{4}
\end{equation*}
$$

We assume that there is some level of noise in both the choice of word given intention $P\left(w_{s} \mid i_{s}, l\right)$ and the choice of intention given object $P\left(i_{s} \mid o_{s}\right)$, such that the speaker could in principle have been mistaken about their referent or misspoken their word. We implement this decision by assuming a constant probability of random noise for each of these, which we notate $\alpha$. For simplicity, $\alpha$ is assumed to be the same for both terms. The value of $\alpha$ serves only to scale the results we report below, but-as in nearly all probabilistic modelssome level of uncertainty about the individual observations is necessary to make graded predictions.

We now consider the prior distribution over lexicons. We define this prior hierarchically as being the product of a constraint over lexicons $c \in C$ :

$$
\begin{equation*}
P(l)=P(l \mid c) P(c) \tag{5}
\end{equation*}
$$

We consider a hypothesis space of four different constraints placed on the mappings between words and objects within lexicons: one word to one object ( $1-1$ constraint), one word to many objects ( 1 -many constraint), many words to one object (many- 1 constraint), and a null constraint. The 1-many constraint applies a restriction that each object maps to at most
one word in a lexicon. The many- 1 constraint applies a restriction that each word maps to at most one object in a lexicon. The 1-1 constraint applies both of these restrictions, and the null constraint applies neither of these restrictions. ${ }^{3}$ In practice, these hypotheses were implemented such that each lexicon consistent with a constraint was equiprobable, and all inconsistent lexicons had probability 0 . For simplicity, we assumed that $P(c) \propto 1$, although this assumption could easily be modified in future work.

For the simulations below, we were able to infer exact posterior distributions by enumerating all possible lexicons and normalizing (Equation 1).

## Simulation 1: Disambiguation at multiple levels

As a first test of our model on the disambiguation task, we trained the model on a corpus containing two situations. The first was an unambiguous situation in which word $w_{1}$ was associated with object $o_{1}$. This piece of evidence corresponded to the known word in the disambiguation task ("ball" in the example described above). We also included a disambiguation experimental situation, where the previously learned object $o_{1}$ (ball), a new object $o_{2}$ (whisk), and a novel word $w_{2}$

[^136]

Figure 3: Results from Simulation 2. Each panel shows the probability of inferring a 1-1 constraint on the lexicon given a different input corpus, described in text. Horizontal axes varies the overall number of distinct objects presented in the exposure corpus, while the colored lines denote different numbers of exposures to the corpus.
("dax").
Figure 2 shows the posterior distribution over lexicons inferred on the basis of this corpus. Each of the 16 possible lexicons (assuming a world with only two words and two objects) are represented along the x -axis, where lexicons are represented by object and word nodes connected by links.

In this maximally simple simulation of the disambiguation task, all four prior constraints give the highest posterior probability to the lexicon that links the novel word and the novel object. This result emerges from the structure of the inference problem: Given that the learner has already observed an association between $w_{1}$ and $o_{1}$, lexicons that posit a link between $w_{1}$ and $o_{2}$ are less probable than those that posit a link between $w_{2}$ and $o_{2}$. This result comes about because an object with two names ( $w_{1}$ and $w_{2}$ ) can be talked about in two different ways, and each of them is individually less probable than the one way of talking about an unambiguously-named object. (This result echoes the finding of mutual exclusivity in Frank et al., 2009).

These results suggest that disambiguation behavior in children could emerge without a 1-1 constraint on lexicons. On the other hand, prior constraints affected the strength of the disambiguation inference. Constraints barring 1-many and many-1 mappings increased the posterior probability of the correct lexicon; when both were in place, the correct lexicon had by far the highest probability. Thus, probabilistic inference and hierarchical constraints both support disambiguation behavior in the model.

## Simulation 2: Learning constraints on lexicons

Simulation 1 suggested that a learner could behave consistent with a 1-1 constraint on lexicons without assuming a hard constraint on the structure of lexicons. Nevertheless, imposing such a hard constraint raised the probability of a correct answer on the disambiguation task. In this simulation, we show that learners may induce a higher-order constraint on lexicons given the right kind of evidence.

To explore the model's ability to learn a hierarchical 1-1 constraint on lexicons, we trained our model on three corpora. Each corpus consisted of a set of situations with a single word and a single object, but we varied whether these mappings were consistent. The first, the "unambiguous exposure" corpus, showed unambiguous (1-1) mappings between words and objects. The second, the "occasionally ambiguous" corpus, showed the same body of data but with two contradictory mappings appended to the end. The final, "fully ambiguous," corpus consisted of one word that mapped to many objects. We varied both the number of exposures to the corpus (1-4) and the number of objects in the corpus (1-4). We then examined the posterior probability of the 1-1 constraint given these exposure corpora.

The results of this simulation are shown in Figure 3. Given 1-1 evidence, the model induces a 1-1 constraint on lexicons, and this bias becomes stronger as the number of observations increases. The posterior probability of the 1-1 constraint is decreased only slightly by a few ambiguous observations, reflecting the general robustness of this inference. In contrast, in the fully ambiguous condition, the model learns with relatively little data that a 1-1 constraint does not hold.

## Simulation 3: Stronger mappings result in stronger disambiguation

In Simulation 3, we explore whether providing more evidence for a link between the known word and object will in turn strengthen the probabilistic disambiguation effect between words and objects. Recall that the disambiguation effect in Simulation 1 emerged as a result of prior evidence for an association between the known word and object ("ball" and ball). Thus, this model predicts that if the learner receives more evidence for an association between the known word and known object, the disambiguation bias should become stronger.

We trained the model with either 1,2 , or 3 situations in which $w_{1}$ was unambiguously associated with $o_{1}$. We then tested the model in the disambiguation task with a known and


Figure 4: Model predictions under each of the four lexical constraints (left) and experimental results (right) for success in the disambiguation task as a function of the number of labels observed in training. Lower legend shows noise conditions for the model simulations.
unknown object and a novel word, as in Simulation 1, but using a Luce choice rule to compute the probability of a correct choice (Luce, 1963). If more observed associations between the known word and object lead to a stronger bias toward correct lexicons, we should expect the disambiguation bias to increase with the number of training situations.

Assuming a 1-1 constraint, the magnitude of the bias toward correct lexicons increases with number of training situations with the known word-known object association (Fig. 4, upper-left panel). In addition, the magnitude of this increase is sensitive to the noise parameter $\alpha$ that determines the probability that the wrong word was spoken to refer to an object.

## Experiment

We tested the prediction that confidence in the known word mapping leads to a stronger disambiguation inference in preschool children.

## Methods

We recruited 110 children ages $2 ; 1-4 ; 11$ from the floor of the Boston Children's Museum. In each one-year age group, we collected data from 35-38 children.

Each child completed four trials. Each trial consisted of a training and a test phase in a "novel-novel" disambiguation task (de Marchena et al., 2011). In the training phase, the experimenter presented the child with a novel object, and explicitly labeled the object with a novel label 1,2 , or 3 times ("Look at the dax"), and contrasted it with a second novel ob-
ject ("And this one is cool too") to ensure equal familiarity. In the test phase, the child was asked to point to the object referred to by a second novel label ("Can you show me the zot?"). Number of labels used in the training phase was manipulated between subjects. There were eight different novel words and objects. Object presentation side, object, and word were counterbalanced across children.

## Results

Responses were coded as correct if participants selected the novel object at test. As predicted, children showed a stronger disambiguation effect as the number of training labels increased, and as noise decreased with age (Fig. 4, right panel).

We analyzed the results using a logit mixed model to predict correct responses with age and number of labels as fixed effects, and participant as a random effect. There was a significant effect of age $(\beta=.044, p<.001)$ such that older children showed a stronger disambiguation bias. There was also a significant effect of number of labels, such that more training labels led to stronger disambiguation ( $\beta=.454, p<.001$ ). The interaction between age and number of labels was not significant ( $\beta=.019, p=.16$ ). Children's increased confidence in the disambiguation inference, as a function of number of training labels, is consistent with model predictions.

## General Discussion

The disambiguation effect suggests the presence of underlying cognitive mechanisms that help children solve the difficult mapping problem inherent of early word learning (Quine,
1960). Two classes of mechanisms have been proposed: a constraint on the structure of permitted lexicons, and in-themoment pragmatic inferences about the most likely referent given the context. We used a hierarchical Bayesian model to explore the independent contributions of these two effects and find that neither mechanism is necessary to create a bias, but either is sufficient. Disambiguation is strongest when both mechanisms jointly contribute.

This result has important consequences for attempts to experimentally differentiate between the proposed accounts of disambiguation. Given that both mechanisms can in principle lead to disambiguation behavior, experimental tests of disambiguation cannot distinguish between these two theories (as they are instantiated here). That is, evidence for disambiguation behavior is consistent with both a pragmatic account and a mutual exclusivity constraint account. Furthermore, there may be variability in the weights of these constraints across populations. For example, higher-order lexical constraints may play a larger role in disambiguation for individuals with impaired social-cognitive skills (e.g. autism), relative to typically developing children. Our results suggest that future research in this area should reconsider the assumption that a single mechanism must completely and independently give rise to the disambiguation effect.

Our model may provide useful insight into disambiguation in bilingualism. For bilingual learners, the structure of associations between words and objects in the environment differs from that of monolinguals. Bilingual learners typically observe two basic-level words associated with each object rather than one. To make sense of these associations, they might ultimately form an overhypothesis that there is a $1-1$ constraint on lexicons within each language, but they might nevertheless initially entertain a 1-many constraint as a hypothesis. Indeed, there is evidence that disambiguation behavior is weaker in bilingual and trilingual children (ByersHeinlein \& Werker, 2009).

Finally, it is important to consider the limits of an ideal observer analysis. While our results suggest that both mechanisms could contribute to disambiguation behavior, this finding does not entail that both mechanisms do in fact contribute. It remains possible that disambiguation behavior is the result of a single mechanism. Nonetheless, given evidence from other domains that the mind may simultaneously integrate basic probabilistic inferences with higher-order constraints (Tenenbaum, Kemp, Griffiths, \& Goodman, 2011), it seems likely that disambiguation behavior emerges from multiple underlying cognitive mechanisms.

## Acknowledgements

We gratefully acknowledge Jesse Snedeker for the suggestion that led to the Experiment, as well as for providing materials from de Marchena et al. (2011). We thank Joseph Austerweil for valuable feedback, and Melina Flores for assistance in data collection.

## References

Bion, R., Borovsky, A., \& Fernald, A. (2013). Fast mapping, slow learning: Disambiguation of novel word-object mappings in relation to vocabulary learning at 18,24 , and 30 months. Cognition, 126, 39-53.
Byers-Heinlein, K., \& Werker, J. (2009). Monolingual, bilingual, trilingual: infants' language experience influences the development of a word-learning heuristic. Developmental Science, 12, 815-823.
Clark, E. (1987). The principle of contrast: A constraint on language acquisition. Mechanisms of Language Acquisition. Hillsdale, NJ: Erlbaum.
de Marchena, A., Eigsti, I., Worek, A., Ono, K., \& Snedeker, J. (2011). Mutual exclusivity in autism spectrum disorders: Testing the pragmatic hypothesis. Cognition, 119, 96-113.
Diesendruck, G., \& Markson, L. (2001). Children's avoidance of lexical overlap: A pragmatic account. Developmental Psychology, 37, 630.
Frank, M. C. (in press). Throwing out the Bayesian baby with the optimal bathwater: Response to Endress (2013). Cognition.
Frank, M. C., \& Goodman, N. (2012). Predicting pragmatic reasoning in language games. Science, 336, 998-998.
Frank, M. C., Goodman, N., \& Tenenbaum, J. (2009). Using speakers' referential intentions to model early crosssituational word learning. Psychological Science, 20, 578.
Geisler, W. (2003). Ideal observer analysis. The Visual Neurosciences, 825-837.
Golinkoff, R., Mervis, C., Hirsh-Pasek, K., et al. (1994). Early object labels: The case for a developmental lexical principles framework. Journal of Child Language, 21, 125-125.
Goodman, N., Tenenbaum, J., Feldman, J., \& Griffiths, T. (2010). A rational analysis of rule-based concept learning. Cognitive Science, 32, 108-154.
Halberda, J. (2003). The development of a word-learning strategy. Cognition, 87, B23-B34.
Luce, R. (1963). Detection and recognition. Handbook of Mathematical Psychology, 1, 103-189.
Markman, E., \& Wachtel, G. (1988). Children's use of mutual exclusivity to constrain the meanings of words. Cognitive Psychology, 20, 121-157.
Markman, E., Wasow, J., \& Hansen, M. (2003). Use of the mutual exclusivity assumption by young word learners. Cognitive Psychology, 47, 241-275.
Mervis, C., Golinkoff, R., \& Bertrand, J. (1994). Two-yearolds readily learn multiple labels for the same basic-level category. Child Development, 65, 1163-1177.
Preissler, M., \& Carey, S. (2005). The role of inferences about referential intent in word learning: Evidence from autism. Cognition, 97, B13-B23.
Quine, W. (1960). Word and object (Vol. 4). The MIT Press.
Tenenbaum, J., Kemp, C., Griffiths, T., \& Goodman, N. (2011). How to grow a mind: Statistics, structure, and abstraction. Science, 331, 1279-1285.

# An integrated model of concept learning and word-concept mapping 

Molly Lewis<br>mll@stanford.edu<br>Department of Psychology<br>Stanford University

Michael C. Frank<br>mcfrank@stanford.edu<br>Department of Psychology<br>Stanford University


#### Abstract

To learn the meaning of a new word, children must solve two distinct problems: identify the referent under ambiguity and determine how to generalize that word's meaning to new objects. Traditionally, these two problems have been addressed separately in the literature, despite their tight relationship with one another. We present a hierarchical Bayesian model that jointly infers both the referent of a word in ambiguous contexts and the concept associated with a word. As a first step in testing this model, we provide evidence that our model fits human data in a simple cross-situational concept learning task.


Keywords: cross-situational word learning; Bayesian models

## Introduction

Learning a new word requires drawing a link in your mental lexicon between a word and a concept. But, children do not observe associations between words and abstract concepts; they observe associations between words and exemplars of those concepts. Furthermore, the associations between words and objects are ambiguous: a single word uttered in any particular context is consistent with an infinite number of possible interpretations (Quine, 1960). There are thus two problems a child must solve in order to learn the meaning of a new word: Determine which object is referred to by a word in context (the Mapping Problem) and determine the relevant concept of the object (the Generalization Problem; see Figure 1).

To understand these two problems more clearly, suppose you lived in an (impoverished) world with two words, "apple" and "cherry," and three objects, a green apple, a red apple, and a cherry. You hear the word "apple" in the context of a single red apple on the table. You somehow infer that "apple" refers to the red object on the table, and thus correctly solve the Mapping Problem. But you have not yet succeeded in solving the Generalization Problem. To correctly solve the Generalization Problem, you must decide whether "apple" also refers to the green apple, which is similar in shape to your observed apple exemplar, or whether it also refers to the cherry, which is similar in color to your observed apple exemplar. Or, alternatively, whether "apple" refers to neither of these other objects (i.e. a proper name). Thus, to learn the word "apple" in this world, you must infer both that "apple" refers to the red object on the table, and that "apple" should be generalized to other apple-shaped objects.

Separate learning mechanisms and constraints have been proposed to account for each of these problems. In the case of the Mapping Problem, one proposed constraint is crosssituational statistics (Pinker, 1984; Smith \& Yu, 2008; Yu \& Smith, 2007). Under this account, learners are hypothesized to aggregate the statistics of associations between words and


Figure 1: Schema of the two problems associated with learning the meaning of a word. Learning a new word requires that the child both identify which object the word refers to in the referential context (the Mapping Problem) and how to generalize that word to objects of the same kind (the Generalization Problem).
objects across situations. When considered in an isolated situation, the referent of a word may be ambiguous, but when situations are aggregated across, the learner is able to constrain the hypothesis space of likely meanings. There is evidence that children as young as 12-months-old can learn word meanings in this way (Smith \& Yu, 2008).

A second class of constraints on the Mapping Problem are accounts of the disambiguation effect. The disambiguation effect refers to children's tendency to select a novel, as opposed to familiar, object as a referent for a novel word. One account of this phenomenon is the principle of mutual exclusivity (Markman \& Wachtel, 1988; Markman, Wasow, \& Hansen, 2003). Under this proposal, there is a constraint on the types of lexicons considered when learning the meaning of a new word. With this constraint, children are biased to consider only those lexicons that have a one-to-one mapping between words and objects. Thus, when faced with an ambiguous referential context, the child solves the mapping problem by assuming that the novel word refers to the object for which she does not yet have a word in her lexicon. This is the inferred mapping because it is the only referent that allows the learner to maintain a one-to-one structure between words and concepts in the lexicon. Others have proposed that general pragmatic assumptions can also account for this effect (Clark, 1987; Diesendruck \& Markson, 2001).

There are also a range of proposals about how children might solve the Generalization Problem. One proposal is that children have a bias to generalize by shape (Smith, Jones, Landau, Gershkoff-Stowe, \& Samuelson, 2002). With this
bias, a child who has learned that "apple" maps to apple, for example, can generalize "apple" to all apple-shaped things. This bias allows learners to rule out alternative, less probable generalizations strategies, such as generalization along the dimension of color. A second proposal is that children have a bias to generalize to another object of similar kind, rather than to one that is thematically related ("Taxonomic Assumption"; Markman, 1990). For example, upon hearing the word "cherry," a child with this bias would be more likely to generalize the word to another fruit, as opposed to ice cream, despite the fact the ice cream and cherries often go together (see Xu \& Tenenbaum, 2007, for a probabilistic view).

Though theoretically distinct, and investigated separately, these the two problems are intimately related. If a child has solved the Generalization Problem for a particular category, the Mapping Problem becomes much easier. For example, suppose a child is faced with a never-before-seen apple and a novel object, and hears the word "dax". If the child has solved the Generalization Problem, the child can identify the apple as an exemplar of the APPLE ${ }^{1}$ concept, and determine the correct referent by mutual exclusivity. Conversely, if a learner can easily solve the Mapping Problem, the learner will accumulate more correct exemplars of a category, and thus be more likely to infer the correct concept. Thus, existing proposals about how each of these problems is solved takes the other problem for granted. But, importantly, a child acquiring language begins with neither of these problems solved; both must be solved in parallel. That is, a learner must determine both what object a word refers to, and how to generalize that meaning beyond the particular context. And, critically, she must do both at the same time.

There is limited work exploring how children might solve these two problems in parallel. A study by Akhtar and Montague (1999) begins to address this question by asking whether children might use cross-situational statistics to learn the relevant features for generalization. In their task, 2-4 year old children were presented with three novel objects that all shared a common feature (e.g. color), but varied along two other features (e.g. texture and shape). Children were able to correctly infer that the novel word referred to the shared feature. This result provides important evidence that children can infer word concepts cross-situationally. However, it is unclear whether this type of learning generalizes to the real world because the actual learning environment is not structured in a way that perfectly disambiguates word meanings cross-situationally.

Apart from word learning, the Generalization Problem has been well-studied in adults (Laurence \& Margolis, 1999; Rosch \& Mervis, 1975; Rosch, Mervis, Gray, Johnson, \& Boyes-Braem, 1976; Medin \& Ortony, 1989). However, limited research has attempted to extend this body of literature to work with children. One exception is work by Sloutsky and colleagues which adopts models of similarity to explain how

[^137]

Figure 2: The generative process for our model. Shading indicates observed variables
children generalize novel words (e.g. Sloutsky, Lo, \& Fisher, 2001).

We present a hierarchical Bayesian model that solves the Mapping and Generalization Problems in parallel. In modeling the Generalization Problem, we draw on the Boolean concept learning framework in which objects are defined by a set of features with a range of values (Shepard, Hovland, \& Jenkins, 1961). The goal for the word learner is conceptualized as the task of mapping a word to a set of features that define the relevant concept. In modeling the Mapping Problem, we focus on the role of cross-situational statistics. In particular, we build on the model developed by Frank, Goodman, and Tenenbaum (2009) that takes into account the intentions of the speaker in order to identify the referent in ambiguous contexts.

The plan for the paper is as follows. We first describe the design of this extended model, and then describe the results of an experiment that explores adult performance in a crosssituational Boolean concept learning task.

## Design of the Model

The goal of our model is to understand how children arrive at an understanding about the meanings of words, on the basis of limited evidence about the associations between words and objects. That is, the goal is to infer a lexicon - a set of word-concept mappings - on the basis of basis of observations of words and objects. To model this, we consider a set of variables relevant to this learning problem, and assume that they are related probabilistically. We assume an identical dependency structure as the model developed by Frank et al. (2009), with the addition of a concept layer to the generative process (see Figure 2). This model is the same underlying
model as presented in Lewis and Frank (2013) but with the addition of a theory of Boolean concepts. For completeness, we present the full model here, but details are identical except where noted.

We model a word learner as performing Bayesian inference to infer a lexicon $l$, which we represent as a (sparse) bipartite graph connecting words $W=w_{1} \ldots w_{n}$ to concepts $C=c_{1} \ldots c_{m}$. Concepts are written as a vector of features with values 1,2 or $*$. The $*$ notation denotes a feature that is irrelevant to the definition of a concept. For example, $[1 * *]$ represents a concept that is defined only by the value of the first feature. This hierarchical formulation of concepts is substantially similar to the concept learning model proposed by Goodman, Tenenbaum, Feldman, and Griffiths (2010). The full possible set of lexicons is denoted as $L$.

The learner infers a distribution over lexicons, given a corpus $S$ of situations (each consisting of sets of words $\bar{w}_{s}$ and objects $\bar{o}_{s}$ ). From Bayes' rule, the posterior probability of a lexicon is given by

$$
\begin{equation*}
P(l \mid S)=\frac{P(S \mid l) P(l)}{\sum_{l^{\prime} \in L} P\left(S \mid l^{\prime}\right) P\left(l^{\prime}\right)} . \tag{1}
\end{equation*}
$$

The prior $P(L)$ is assumed to be uniform over lexicons that map a concept to at most one word (one word to many concepts). We now define the likelihood term $P(S \mid L)$.

Using the generative process in Figure 2, we can write the likelihood of a particular situation in terms of the relationship between the objects that were observed in the situation $s$, the speaker's referential intention $i_{s}$ (a choice to speak about one of the objects), the concept $c_{s}$ selected by the speaker to represent the intention, and the referring word $w_{s}$. As in our prior work, we assume that referential intentions are unobserved and sum across all possible intentions uniformly:

$$
\begin{equation*}
P(s \mid l)=\sum_{i_{s} \in \bar{\sigma}_{s}} p\left(w_{s}, c_{s}, i_{s}, o_{s}, \mid l\right) \tag{2}
\end{equation*}
$$

By the conditional independence of words and objects, we use the chain rule to expand to:

$$
\begin{equation*}
P(s \mid l)=\sum_{i_{s} \in \bar{o}_{s}} P\left(w_{s} \mid c_{s}, l\right) P\left(c_{s} \mid i_{s}\right) P\left(i_{s} \mid o_{s}\right) \tag{3}
\end{equation*}
$$

Finally, we aggregate across situations by taking the product of each independent situation:

$$
\begin{equation*}
P(S \mid l)=\prod_{s \in S} \sum_{i \in \bar{o}_{s}} P\left(w_{s} \mid c_{s}, l\right) P\left(c_{s} \mid i_{s}\right) P\left(i_{s} \mid o_{s}\right) \tag{4}
\end{equation*}
$$

To find the key term in our concept model, $p\left(c_{s} \mid i_{s}\right)$, we use a noisy Naive Bayes classifier:

$$
P\left(c_{s} \mid i_{s}\right)=\prod_{j=1 \ldots f} \begin{cases}1-\alpha & \text { if }\left(c_{s}^{j}=i_{s}^{j}\right) \vee\left(i_{s}^{j}=*\right)  \tag{5}\\ \alpha & \text { otherwise }\end{cases}
$$

This formulation quantifies the probability of a concept given an intended object in terms of the match between the three features.


Figure 3: Experimental stimuli. Each object is defined by a binary value for each of three features: shape, appendage, and color.

We assume that there is some level of noise in both the choice of word given intention $P\left(w_{s} \mid i_{s}, l\right)$ and the choice of intention given object $P\left(c_{s} \mid i_{s}\right)$, such that the speaker could in principle have been mistaken about their referent or misspoken. We implement this decision by assuming a constant probability of random noise for each of these, which we notate $\alpha$; for simplicity, $\alpha$ is assumed to be the same for both decisions. The particular choice of $\alpha$ values only serves to scale the predictions, and does not influence the relative predictions of the test item types. However, as in nearly all probabilistic models, some level of uncertainty about the individual observations is necessary to be able to make graded predictions.

In the simulations reported here, we did inference by exact inference via full combinatoric enumeration of the space of possible lexicons.

## Experiment

Our model jointly solves the two problems associated with learning the meaning of a new word, the Mapping and Generalization Problems. As a first step in evaluating the model, we compared human and model performance in a crosssituational Boolean concept learning task. Participants were given a situation in which a word is seen in the context of two objects, but in a way that is ambiguous as to which of these objects (either or both) the label refers to. The learner is then presented with a second such situation. While each of these situations is individually ambiguous, the learner could aggregate information across situations to infer the concept associated with the word. As predicted by the model, we found that participants generalized the meaning of the label in a graded manner: the more features the training objects shared with the test object, the more likely participants were to generalize the label to the test object.

## Method and Materials

Participants Two hundred and and sixty-six adults were recruited from Amazon's Mechanical Turk. Twenty-two were


Figure 4: Bets on the probability of "dax bren nes" generalizing to each of the relevant test item types in each condition. Error bars represent $95 \%$ confidence intervals as computed via non-parametric bootstrap. Example items seen in the training situations are given in the top-right box of each plot. Given the particular training items shown here, an item in each of the relevant test item types (defined by the number of features shared with the training items) is shown along the $x$-axis. Actual training items (and thus test items) were counter-balanced across participants.
excluded for either not completing the task appropriately (e.g. by responding with values greater than 100) or failing to provide responses for all 8 test objects. All reported that they were native speakers of English.

Stimuli Each object in our stimulus set was constructed to have three binary features. The features of interest were shape, appendage and color. When fully permuted, this defines a space of eight possible objects (see Figure 3).

Procedure Participants viewed a webpage that showed two situations with two objects each. In the first situation, they were instructed: "Suppose you saw these two objects and heard 'dax bren nes.'" A multi-word novel label was used to avoid biases towards meanings consistent with the grammatical class of the word. In other words, we wanted to avoid participants inferring that because the word was an adjective, it was more likely to refer to a property (e.g. color) than a particular object (i.e. a proper noun), for example. Two more objects were presented below and participants were asked to "Now suppose you saw these two new objects and heard 'dax bren nes' again." They were then asked to "bet whether or not you think each of the objects below could also be called 'dax bren nes." Images of all eight objects (including the training items) were then presented, and participants were asked to provide a bet $0-100$ indicating their judgement.

Across participants, we manipulated the number of features shared within and across situations. ${ }^{2}$ We tested an unambiguous baseline condition in which the same object was paired with a different object in each situation and 5 ambiguous conditions in which the features of the objects were confounded either within or across situations. For the ambiguous conditions, we tested cases in which 1 or 2 features were shared within situations ("confounded within" conditions), 1 or 2 features shared across situations ("confounded across" conditions), and a case in which 3 features were shared both within and across situations (Figure 4).

[^138]

Figure 5: Mean bets in each experiment condition, plotted by predictions for each of the models (model predictions are scaled on the horizontal axis). Error bars represent $95 \%$ confidence intervals, as computed via non-parametric bootstrap. The line of best fit is plotted in red.

## Results and Model Fits

Participants showed a consistent gradient of generalization such that greater number of distinct features resulted in lower bets (weaker generalizations), consistent with previous experiments (Figure 4).

Model fits are shown in Figure 5. Our model fits the data with a correlation of $r=.95$. We compared this fit against a null model in which we calculated the target's total feature distance from objects in the situations. This was calculated by counting the number of features for which the target differed from each situation object (e.g. the feature distance [111] and [122] is 2) and summing across all four objects in the situation. This standard exemplar style model fits the data relatively well ( $r=.89$ ). Nevertheless, our model provides a substantial gain in fit. Using non-parametric bootstrap, the cross-situational concept-learning model fits the data significantly better than the feature distance model $(p<.05)$.

## General Discussion

In this cross-situational Boolean concept learning task, our model performed competitively with a simple feature distance model. Critically, however, our model has the machinery to solve not only this simple concept-mapping problem under minimal ambiguity, but can also deal with more complex worlds in which multiple words are present. Given that no existing model is able to jointly account for both the Mapping and Generalization Problems, this model provides a fruitful theoretical tool for future work to explore how children might solve these problems.

For example, this experiment could be straight-forwardly extended to introduce a more complex Mapping Problem component to the task. This could be done by adding additional words to the cross-situational learning context. In a
minimal version of this experiment, the learner could observe $w_{1}$ with [11] and [12] and $w_{2}$ with [22] and [12]. A learner who assumes that the speaker refers to both objects within each situation, might infer a mapping between $w_{1}$ with $[1 *]$ and a mapping between $w_{2}$ with $[* 2]$, given this referential context. Using situations such as these, this paradigm can be extended to directly explore joint inference of both the Mapping and Generalization problems.

An important underlying assumption of this model is that features are given a priori. This seems like an extreme position given that it is implausible that children acquiring language have an innate "appendage" feature, for example. It is, in a sense, the very goal of this model to explain how children acquire such abstract concepts as APPENDAGE. That is, features are themselves concepts that can be considered as primitives in the construction of more complex concepts. This problem, however, is not specific to the word learning problem, but rather is a challenge more generally to the Boolean concept learning framework. Nonetheless, a complete theory of how children acquire word concepts will need to provide an account for the origin of features.

Given this theoretical point, our model should be understood as a computational level description of the problem of acquiring word-concepts, given some set of concepts (i.e. features). Our model remains agnostic about the origins and nature of these initial concepts but, given some primitive set of concepts, our model describes how a learner might bootstrap from these primitives to infer more and more complex concepts. While it seems unlikely that children have an innate APPENDAGE feature, there is evidence that children may have certain perceptual categories, such as color, very early in development (Bornstein, Kessen, \& Weiskopf, 1976). Primitive perceptual features like color categories may provide the
initial building blocks for the construction of more complex concepts, given experience with the environment.

In sum, our model provides a rich framework for studying the word learning problem at the computational level. Previous research has explored how children might solve the two subproblems associated with word learning - the Generalization and Mapping Problems - separately. Our model contributes to this area by providing a unifed account for both of these problems. The experiment reported here suggests that our model is able to account for participants' behavior in solving one of these problems - the Generalization Problem in a simple cross-situational task. Importantly, our model's contribution to theories of the Generalization Problem is to provide an account of the generalization inferences, given an initial set of primitive concepts. This account, coupled with the ability to explore the Mapping Problem, lays the groundwork for a more cohesive understanding of how children learn the meanings of words.

## Acknowledgements

We thank Mia Kirkendoll for her assistance in data collection.

## References

Akhtar, N., \& Montague, L. (1999). Early lexical acquisition: The role of cross-situational learning. First Language, 19(57), 347-358.
Bornstein, M., Kessen, W., \& Weiskopf, S. (1976). Color vision and hue categorization in young human infants. Journal of Experimental Psychology: Human Perception and Performance, 2(1), 115.
Clark, E. (1987). The principle of contrast: A constraint on language acquisition. Mechanisms of language acquisition. Hillsdale, NJ: Erlbaum.
Diesendruck, G., \& Markson, L. (2001). Children's avoidance of lexical overlap: A pragmatic account. Developmental Psychology, 37(5), 630.
Frank, M. C., Goodman, N., \& Tenenbaum, J. (2009). Using speakers' referential intentions to model early crosssituational word learning. Psychological Science, 20(5), 578.

Goodman, N., Tenenbaum, J., Feldman, J., \& Griffiths, T. (2010). A rational analysis of rule-based concept learning. Cognitive Science, 32(1), 108-154.
Laurence, S., \& Margolis, E. (1999). Concepts: Core readings. In E. Margolis \& S. Laurence (Eds.), (chap. 1). MIT Press.
Lewis, M., \& Frank, M. C. (2013). Modeling disambiguation in word learning via multiple probabilistic constraints. In Proceedings of the 35th Annual Meeting of the Cognitive Science Society.
Markman, E. (1990). Constraints children place on word meanings. Cognitive Science, 14(1), 57-77.
Markman, E., \& Wachtel, G. (1988). Children's use of mutual exclusivity to constrain the meanings of words. Cognitive Psychology, 20(2), 121-157.

Markman, E., Wasow, J., \& Hansen, M. (2003). Use of the mutual exclusivity assumption by young word learners. Cognitive Psychology, 47(3), 241-275.
Medin, D., \& Ortony, A. (1989). Psychological essentialism. Similarity and Analogical Reasoning, 179-195.
Pinker, S. (1984). Language learnability and language development, with new commentary by the author (Vol. 7). Harvard University Press.
Quine, W. (1960). Word and object (Vol. 4). The MIT Press.
Rosch, E., \& Mervis, C. (1975). Family resemblances: Studies in the internal structure of categories. Cognitive Psychology, 7(4), 573-605.
Rosch, E., Mervis, C., Gray, W., Johnson, D., \& BoyesBraem, P. (1976). Basic objects in natural categories. Cognitive Psychology, 8(3), 382-439.
Shepard, R., Hovland, C., \& Jenkins, H. (1961). Learning and memorization of classifications. Psychological Monographs: General and Applied, 75(13), 1-42.
Sloutsky, Lo, Y.-F., \& Fisher, A. V. (2001). How much does a shared name make things similar? Linguistic labels, similarity, and the development of inductive inference. Child Development, 72(6), 1695-1709.
Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., \& Samuelson, L. (2002). Object name learning provides on-the-job training for attention. Psychological Science, 13(1), 13-19.
Smith, L. B., \& Yu, C. (2008). Infants rapidly learn wordreferent mappings via cross-situational statistics. Cognition, 106(3), 1558-1568.
Xu, F., \& Tenenbaum, J. (2007). Word learning as Bayesian inference. Psychological Review, 114(2), 245.
Yu, C., \& Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. Psychological Science, 18(5), 414-420.

# Proper Sequencing and Level-Bridging Scaffolding in Learning a Chemical System with Graphical Simulations 

Na Li (nl2284@tc.columbia.edu)<br>Teachers College, Columbia University, $525 \mathrm{~W} .120{ }^{\text {th }}$ street<br>New York, NY 10025 USA<br>John Black (black@exchange.tc.columbia.edu)<br>Teachers College, Columbia University, 525 W. $120^{\text {th }}$ street<br>New York, NY 10025 USA


#### Abstract

Secondary-level students encounter many difficulties in learning complex systems with hierarchical levels. Scaffolding is very critical in teaching complex systems. We have two complementary research questions on scaffolding: 1 . How can we chunk and sequence the learning activities in teaching complex systems? 2. How can we help students make connections across system levels? A simulation-based environment teaching a chemical system was used as the research instrument, and the study was conducted at a middle school setting. The results showed that the sequencing method following the "from concrete to abstract" principle produced better recall and comprehension of the system concepts (knowledge integration), while the sequencing method aligned with the casual structure of the system facilitated the construction of a better causal model for transfer. The results also demonstrated that explicit level-bridging scaffolding had positive effects on both knowledge integration and learning the deep causal structure.


Keywords: Complex systems; sequencing methods; levelbridging scaffolding; secondary-level science

## Research Background

Complex systems have become an important topic in today's science education. It is usually difficult for students to learn complex systems with hierarchical levels and abstract system dynamics (Jacobson \& Wilensky, 2006). Complex systems can be difficult from different perspectives. 1). Spatialtemporal extension of a system, e.g., there are many system levels with complex formation and interactivity 2). Abstract system levels and causal structures in a system, e.g., higherlevel patterns emerging from lower-level dynamics (Baryam, 1997). Although two types of difficulties always coexist in various complex systems, one may define the complexity more than the other in a specific context or at a certain learning stage. Biological and natural systems often have many system levels, diversified local behaviors and interactivity (Hmelo-silver \& Azevedo, 2006). For example, the human circulatory system has a "downward tree" structure with a large number of elements, and varied local element interactivity. Effective knowledge integration is a learning difficulty students have to conquer before learning the emergent processes involved in this type of system. Abstract levels and causal structures are often found in
chemical and physics systems (Stieff, 2011). For example, it is difficult to visualize "voltage in an electric circuit" emerging from electrons' behaviors and "gas pressure" from gas molecular activities.

Agent-based modeling and visualizing tools can create visual acuity of system levels and demonstrate cross-level dynamics (Levy \& Wilensky, 2009). However, given the complex nature of the learning content, mere perceptual grounding is not sufficient for effective learning. Scaffolding is a critical factor in learning complex systems (Jacobson et al., 2011). The first research question of this study: How can we chunk and sequence the learning activities in teaching complex systems? There have been contradictory findings to this question. However, analyzing the learning difficulties from different perspectives may address the debate over the sequencing methods.

Knowledge integration refers to students connecting scientific concepts and normative ideas, and providing coherent explanations to scientific phenomena (Linn, 2006). From the perspective of knowledge integration, the "topdown" approach starting from the concrete macro-level function of a system is effective. In Liu \& Hmelo-Silver (2009)'s study, participants learned the respiratory system with either the "top-down" or the "bottom-up" sequencing method, and the results showed that starting from the systemwide function ("how do we breathe") was better than starting from the lower-level substructures and entities. As can be seen, in this type of biological system, a higher-level function is concrete and easy to understand. And a function is often realized by the interactivity of a large number of diverse lower-level substructures. "Making science accessible" as a knowledge integration guideline informs us that concrete levels of a topic should come before abstract ones (Linn, 2006). Additionally, a top-down function-oriented sequencing method provides a good conceptual structure for knowledge integration (Liu \& Hmelo-silver, 2009).

Many studies demonstrate that the "bottom-up" approach is effective in teaching the implicit and abstract causal structure (e.g., emergence) of a complex system (Wilenky \& Stroup, 2002). This sequencing method allows students to experience the causal process of how the small effects of the micro-level elements can lead to the macro-level patterns. For example, in the Connected Chemistry Curriculum (Levy \& Wilensky, 2009), the approach is to let students manipulate and articulate the micro-level entity behaviors (e.g., how a
single gas molecule collide with the walls), and then gradually expand to the emergent processes and phenomena. It is claimed that the "bottom-up" approach help students conceptually understand the implicit linkages between the micro and macro level of the gas phenomena (Levy \& Wilensky, 2009).

While we are chunking and sequencing the tasks, we need to provide extra scaffolding for students to make connections across learning activities. Inter-level experience is critical in learning complex systems (Levy \& Wilensky, 2009). Thus the second research question of this study is: How can we help students make connections across system levels? Scaffolding that elicits self-explanation could significantly improve learning (Chi et al, 1994). Explicit level-bridging scaffolding such as inter-level questions facilitates selfexplanation, and is an effective strategy in teaching complex systems (Stieff, 2011). In this study, the effect of explicit level-bridging scaffolding was tested.

## Learning Materials and Instrument

Ideal gas law is a complex chemical system. A concrete phenomenon such as "an aerosol can explodes when it is thrown into the fire" can be defined as the "system-wide" or "pattern-level" function. This level is concrete, observable, and without complex dynamics. Temperature-pressurevolume relationship is an abstract macro level, which is analogous to and explains the observable pattern-level function, thus we define this level as the "mechanism level." This level depicts the mechanism of the "can explosion phenomenon"; meanwhile, this level emerges from the lower-level molecular activity defined as the "entity level".

A simulation-based environment with two simulations was used as the research instrument. The first simulation visualized the pattern-level function. Students could drag the fire icon towards the can and observe the can explodes (see Figure 1). The second simulation (see Figure 2) visualized the mechanism level (Temperature-pressure-volume relationship) and the entity level (molecular activity). The two simulations could be displayed separately on two pages. Students could switch to either simulation by clicking an arrow button, or they could be displayed on the same page and dynamically linked (see Figure 3). The dynamic link technique can facilitate information integration from multiple representations (van der Meji \& de Jong, 2006), and in this study, it was a part of the manipulation of the explicit levelbridging scaffolding condition.


Figure 1: Aerosol can simulation


Figure 2: Gas container simulation


Figure 3: Two simulations dynamically linked

## Variables \& Hypotheses

Two variables were tested in this study: 1. Sequencing methods; 2. Level-bridging scaffolding.

## Sequencing Methods

Three sequencing methods were compared in this study. The sequencing methods variable was manipulated by changing the delivery order of the three levels of the chemical system.

F-M-E sequencing method Starting from the function level to the mechanism level, and then to the entity level (F-M-E). This sequencing method followed the "from concrete to abstract" principle. It was function-oriented thus provided a good conceptual framework for knowledge integration.

E-M-F sequencing method Starting from the entity level to the mechanism level, and then to the function level (E-M-F). This sequencing method followed the "from cause to effect" principle because it was aligned with the causal structure of the ideal gas law system

M-E-F sequencing method Starting from the mechanism level to the entity level, and then to the function level (M-EF). This sequencing method did not follow the "from concrete to abstract" or the "from cause to effect" principle, thus it was hypothesized to be less effective than the other two methods.

## Hypotheses on Sequencing Methods

Hypothesis 1 the F-M-E sequencing method produces better knowledge integration when compared to the E-M-F and the M-E-F sequencing method.

Hypothesis 2 the E-M-F sequencing method produces better understanding of the deep causal structure when compared to the F-M-E and the M-E-F sequencing method.

## Hypotheses on Level-Bridging Scaffolding

Explicit level-bridging scaffolding and implicit levelbridging scaffolding were compared in this study.

Hypothesis 3 Explicit level-bridging scaffolding produces better knowledge integration when compared to implicit level-bridging scaffolding

Hypothesis 4 Explicit level-bridging scaffolding produces better understanding of the deep causal structure when compared to implicit level-bridging scaffolding.

## Method

## Participants

129 seventh graders from two inner city public middle schools participated in this study. Six cases were dropped from the sample, as these participants were absent from the second session of the study. The final sample included 123 participants. $78.9 \%$ were Hispanic, $13.8 \%$ Black, $4.1 \%$ white and $3.3 \%$ other. The mean age of this sample was 12.4 ( $\mathrm{SD}=0.53$ ). $48.8 \%$ were male and $51.2 \%$ female.

This study employed a $3 \times 2$ design. See Table 1 for the 6 treatment groups.

Table 1: $3 \times 2$ experimental design

|  | F-M-E | E-M-F | M-E-F |
| :--- | :--- | :--- | :--- |
| Explicit <br> level <br> bridging | Group 1 | Group 3 | Group 5 |
| Implicit <br> level <br> bridging | Group 2 | Group 4 | Group 6 |

## Procedure

Within each classroom, participants were randomly assigned. The data collection within each class was operated on two consecutive days. The total length of the two sessions was around 100 minutes.

Day 1 All participants took a pretest. Within the same classroom, participants were randomly paired up and randomly assigned to a condition. Each pair was assigned a laptop with the simulations; and each participant was assigned a booklet with 6 learning activities. Participants were asked to read the guidance and questions on the worksheets and write down their answers without any group discussion (for better control of extraneous factors). Three research assistants and the science teacher were present to monitor the learning progress, help change the simulation interfaces and solve technical problems. Participants completed 3-4 learning activities on Day 1.

Day 2 Participants were assigned to the same group as Day 1. They were asked to spend around 5 minutes reviewing their work from Day 1. Participants continued learning and completed the rest of the learning activities. Participants completed a posttest after the learning session.

## Manipulation

Sequencing Methods Sequencing methods were manipulated by changing the delivery order of these three system levels. The same learning activities on three system levels were arranged in different orders.

Level-Bridging Scaffolding This variable was manipulated on two aspects. For the explicit level-bridging condition: 1). Inserting inter-level questions among the learning activities 2). Two simulations were dynamically linked for the final learning activity (See Figure 3.). For the implicit levelbridging condition: 1). Inserting intra-level questions among the learning activities 2 ). Two simulations were not dynamically linked for the final learning activity.
The inter-level questions and intra-level questions were manipulated in a way that the same amount of information was delivered. Please see Table 2 for the two sets of questions. Where each question was inserted also depended on the sequencing method condition.

Table 2: Inter-level questions vs. Intra-level questions

| Inter-level questions |  |
| :--- | :---: |
| 1. What is the relationship <br> between temperature and <br> pressure? Use what you learned <br> about temperature and pressure <br> from the gas container <br> presentation, explain why the <br> aerosol can explodes? |  |
| 2. How do gas molecules <br> behave? Use what you learned <br> about gas molecules; explain <br> why as temperature rises, <br> pressure inside the container <br> also rises? |  |

3. Use the knowledge of gas molecules; explain what happens to the gas pressure inside the aerosol can as you drag the fire closer? Explain why the aerosol can explodes?

## Intra-level questions <br> 1.Explain why the aerosol can explodes?

2. Use what you learned about temperature and pressure from the gas container presentation, explain what is the relationship between temperature and pressure?
3. Use the knowledge of gas molecules; explain how do gas molecules behave?
4.Explain what happens to the aerosol can as you drag the fire closer?
5.What did you learn from the aerosol can presentation?
6.As temperature rises, pressure also rises, is this correct?
7.What did you learn about gas molecules?

## Measures

Pretest The pretest included two open-ended questions asking the participants to explain two ideal gas law problems: "using an ice pack to reduce tooth pain" and "car tires are more likely to explode in the summer than in the winter". No extra system information about ideal gas law was provided, as priming the participants with any level of the system might disrupt the manipulation of the sequencing methods.

Posttest. The posttest included four parts. Part I. short answer questions and labeling questions measuring recall of system knowledge. Part II. Two snapshots of the virtual experiment simulation were provided, participants were asked to describe what happened from Time A to time B. This open-ended question measured recall of simulation events; Part III. Four open-ended questions measured comprehension of the system knowledge, e.g., participants were asked to explain their understanding of "gas pressure", and "why the aerosol can explodes". Part IV. The same two ideal gas law problems as in the pretest were used as transfer questions. This part measured understanding of the deep causal structure of the system.

Most of the questions in the pre and posttest were openended questions. Participants' answers were coded on the absence or presence of important system knowledge units. All possible knowledge units were included in the coding scheme (possible maximum scores were high), but participants' actual scores were relatively low.

Two raters blind to the conditions coded the answers independently, and the inter-rater reliability was above $95 \%$ for all parts of the pre and posttest. Disagreement was resolved via discussion between the two raters.

## Results

## Pretest Scores

The possible maximum score of the pretest was 10. Pretest scores did not significantly differ across sequencing methods, $\mathrm{F}(2, \quad 117)=0.674, \quad \mathrm{p}=0,512 ; \quad$ or across level-bridging scaffolding conditions, $\mathrm{F}(1,117)=0.238, \mathrm{p}=0.789$. The pretest scores were used to establish equivalency and used as a covariate in further analysis.

Table 3. Pretest scores

|  | $\mathrm{F}-\mathrm{M}-\mathrm{E}$ | $\mathrm{E}-\mathrm{M}-\mathrm{F}$ | M-E-F | Marginal |
| :--- | :--- | :--- | :--- | :--- |
| Explicit | 1.33 | $\mathrm{M}=1.45$ | $\mathrm{M}=1.16$ | $\mathrm{M}=1.30$ |
| Level- | $\mathrm{SD}=0.65$ | $\mathrm{SD}=1.08$ | $\mathrm{SD}=0.90$ | $\mathrm{SD}=0.89$ |
| bridging | $\mathrm{N}=20$ | $\mathrm{~N}=19$ | $\mathrm{~N}=22$ | $\mathrm{~N}=30$ |
| Implicit | $\mathrm{M}=1.20$ | $\mathrm{M}=1.46$ | $\mathrm{M}=1.32$ | $\mathrm{M}=1.33$ |
| level- | $\mathrm{SD}=0.88$ | $\mathrm{SD}=1.05$ | $\mathrm{SD}=0.85$ | $\mathrm{SD}=0.93$ |
| bridging | $\mathrm{N}=20$ | $\mathrm{~N}=23$ | $\mathrm{~N}=19$ | $\mathrm{~N}=62$ |
| Marginal | $\mathrm{M}=1.26$ | $\mathrm{M}=1.45$ | $\mathrm{M}=1.23$ | Total |
|  | $\mathrm{SD}=0.77$ | $\mathrm{SD}=1.05$ | $\mathrm{SD}=0.87$ | $\mathrm{M}=1.10$ |
|  | $\mathrm{~N}=40$ | $\mathrm{~N}=42$ | $\mathrm{~N}=41$ | $\mathrm{SD}=0.65$ |
|  |  |  |  | $\mathrm{~N}=123$ |

## Posttest Scores

Knowledge integration was measured through recall and comprehension tasks (Part I, II, III), and understanding of the deep causal structure was measured through transfer tasks (Part IV). ANCOVA and helmert contrasts were conducted as inferential tests.

Part I. Recall of system knowledge (possible maximum score=8) Two statistical outliers were converted to the 98percentile value of the sample distribution. Descriptive data of this part please see Table 4 and Figure 4. Pretest scores as a covariate was marginally significant, $\mathrm{F}(1,116)=3.36$, $\mathrm{p}=0.069$. No interaction between sequencing methods and level-bridging scaffolding was found, $F(2,116)=0.212$, $\mathrm{p}=0.847$, indicating the $\mathrm{F}-\mathrm{M}-\mathrm{E}$ sequencing method and explicit level-bridging scaffolding had additive effects on the recall of system knowledge. The Helmert contrasts results showed that the F-M-E sequencing method produced significantly better recall when compared to the average of the other two sequencing methods, $\mathrm{t}(116)=2.56, \mathrm{p}=0.012$; no significant difference was found between the E-M-F and the M-E-F sequencing method, $t(116)=0.13, p=0.894$. This demonstrated that the "top-down" function-oriented sequencing method following the "concrete to abstract" principle (F-M-E) provided a desirable conceptual framework for knowledge integration. Explicit level-bridging scaffolding had significant positive effects on the recall of system knowledge, $\mathrm{F}(1,116)=7.24, \mathrm{p}=0.008$. The results supported Hypothesis 1 and 3.

Table 4. Recall of system knowledge

|  | $\mathrm{F}-\mathrm{M}-\mathrm{E}$ | E-M-F | M-E-F | Marginal |
| :--- | :--- | :--- | :--- | :--- |
| Explicit | $\mathrm{M}=4.30$ | $\mathrm{M}=3.45$ | $\mathrm{M}=3.50$ | $\mathrm{M}=3.77$ |
| Level- | $\mathrm{SD}=1.08$ | $\mathrm{SD}=1.19$ | $\mathrm{SD}=1.53$ | $\mathrm{SD}=1.26$ |
| bridging | $\mathrm{N}=20$ | $\mathrm{~N}=19$ | $\mathrm{~N}=22$ | $\mathrm{~N}=61$ |
| Implicit | $\mathrm{M}=3.45$ | $\mathrm{M}=3.00$ | $\mathrm{M}=3.00$ | $\mathrm{M}=3.14$ |
| level- | $\mathrm{SD}=1.19$ | $\mathrm{SD}=1.65$ | $\mathrm{SD}=1.20$ | $\mathrm{SD}=1.37$ |
| bridging | $\mathrm{N}=20$ | $\mathrm{~N}=23$ | $\mathrm{~N}=19$ | $\mathrm{~N}=62$ |
| Marginal | $\mathrm{M}=3.88$ | $\mathrm{M}=3.24$ | $\mathrm{M}=3.27$ | Total |
|  | $\mathrm{SD}=1.20$ | $\mathrm{SD}=1.38$ | $\mathrm{SD}=1.40$ | $\mathrm{M}=3.46$ |
|  | $\mathrm{~N}=40$ | $\mathrm{~N}=42$ | $\mathrm{~N}=41$ | $\mathrm{SD}=1.35$ |
|  |  |  |  | $\mathrm{~N}=123$ |



Figure 4. Recall of system knowledge

Part II. Recall of simulation events (possible maximum score=6) One statistical outlier was converted to the 99percentile value of this sample. Descriptive data of this part please see Table 5 and Figure 5. Pretest scores as a covariate was not significant, $\mathrm{F}(1,116)=1.11, \mathrm{p}=0.29$. The $\mathrm{F}-\mathrm{M}-\mathrm{E}$ \& Implicit level-bridging group recalled more simulation events when compared to the other treatment groups. As statistical evidence for that, the interaction between the sequencing methods contrast ( $\mathrm{F}-\mathrm{M}-\mathrm{E}$ vs. other) and the level-bridging scaffolding variable was significant, $\mathrm{t}(116)=2.03, \mathrm{p}=0.045$, meaning that the F-M-E was effective on the recall of simulation events only in the implicit level-bridging condition. The results from Part I and II indicated that the F-M-E sequencing method led to better recall in general. Given explicit level-bridging scaffolding, students were more likely to integrate important system concepts; while in the implicit level-bridging scaffolding condition, participants focused more on superficial simulation events.

Table 5. Recall of simulation events

|  | $\mathrm{F}-\mathrm{M}-\mathrm{E}$ | $\mathrm{E}-\mathrm{M}-\mathrm{F}$ | M-E-F | Marginal |
| :--- | :--- | :--- | :--- | :--- |
| Explicit | $\mathrm{M}=1.75$ | $\mathrm{M}=1.79$ | $\mathrm{M}=1.86$ | $\mathrm{M}=1.80$ |
| Level- | $\mathrm{SD}=0.55$ | $\mathrm{SD}=0.71$ | $\mathrm{SD}=0.83$ | $\mathrm{SD}=0.70$ |
| bridging | $\mathrm{N}=20$ | $\mathrm{~N}=19$ | $\mathrm{~N}=22$ | $\mathrm{~N}=61$ |
| Implicit | $\mathrm{M}=2.30$ | $\mathrm{M}=1.83$ | $\mathrm{M}=1.84$ | $\mathrm{M}=1.98$ |
| level- | $\mathrm{SD}=0.66$ | $\mathrm{SD}=0.83$ | $\mathrm{SD}=0.60$ | $\mathrm{SD}=0.74$ |
| bridging | $\mathrm{N}=20$ | $\mathrm{~N}=23$ | $\mathrm{~N}=19$ | $\mathrm{~N}=62$ |
| Marginal | $\mathrm{M}=2.02$ | $\mathrm{M}=1.81$ | $\mathrm{M}=1.85$ | Total |
|  | $\mathrm{SD}=0.66$ | $\mathrm{SD}=0.77$ | $\mathrm{SD}=0.73$ | $\mathrm{M}=1.89$ |
|  | $\mathrm{~N}=40$ | $\mathrm{~N}=42$ | $\mathrm{~N}=41$ | $\mathrm{SD}=0.72$ |
|  |  |  |  | $\mathrm{~N}=123$ |



Figure 5: Recall of simulation events
Part III. Comprehension (possible maximum score=15) Descriptive data of this part please see Table 6 and Figure 6. Pretest scores were significantly associated with the comprehension scores, $\mathrm{F}(1,116)=8.51, \mathrm{p}=.004$. The interaction between sequencing methods and level-bridging scaffolding was not significant, $\mathrm{F}(2,116)=0.049, \mathrm{p}=.952$, indicating the effects of sequencing methods and levelbridging scaffolding were additive. Although this part showed a similar pattern as Part I, the positive effects of F-M-E sequencing method over the average of the other two
was not significant, $t(116)=1.46, p=0.146$. Significant main effects of the explicit level-bridging scaffolding was found, $\mathrm{F}(1,116)=4.45, \mathrm{p}=0.037<0.05$. When comparing Part I and Part III, we may find that the effects of explicit level-bridging scaffolding on knowledge integration was more sustainable than the $\mathrm{F}-\mathrm{M}-\mathrm{E}$ sequencing method.

Table 6. Comprehension of system knowledge

|  | $\mathrm{F}-\mathrm{M}-\mathrm{E}$ | $\mathrm{E}-\mathrm{M}-\mathrm{F}$ | M-E-F | Marginal |
| :--- | :--- | :--- | :--- | :--- |
| Explicit | $\mathrm{M}=3.97$ | $\mathrm{M}=3.53$ | $\mathrm{M}=3.34$ | $\mathrm{M}=3.60$ |
| Level- | $\mathrm{SD}=1.41$ | $\mathrm{SD}=1.57$ | $\mathrm{SD}=2.01$ | $\mathrm{SD}=1.70$ |
| bridging | $\mathrm{N}=20$ | $\mathrm{~N}=19$ | $\mathrm{~N}=22$ | $\mathrm{~N}=61$ |
| Implicit | $\mathrm{M}=3.20$ | $\mathrm{M}=2.89$ | $\mathrm{M}=2.92$ | $\mathrm{M}=3.30$ |
| level- | $\mathrm{SD}=1.64$ | $\mathrm{SD}=1.27$ | $\mathrm{SD}=1.98$ | $\mathrm{SD}=1.61$ |
| bridging | $\mathrm{N}=20$ | $\mathrm{~N}=23$ | $\mathrm{~N}=19$ | $\mathrm{~N}=62$ |
| Marginal | $\mathrm{M}=3.58$ | $\mathrm{M}=3.18$ | $\mathrm{M}=3.15$ | Total |
|  | $\mathrm{SD}=1.56$ | $\mathrm{SD}=1.43$ | $\mathrm{SD}=3.15$ | $\mathrm{M}=3.30$ |
|  | $\mathrm{~N}=40$ | $\mathrm{~N}=42$ | $\mathrm{~N}=41$ | $\mathrm{SD}=1.68$ |
|  |  |  |  | $\mathrm{~N}=123$ |



Figure 6: Comprehension of system knowledge
Part IV. Transfer tasks (possible maximum score=10) Different from the comprehension questions in Part III, these two transfer questions required participants to recognize the problems as ideal gas law phenomena, and transfer the causal structure of the system to explain the problems. Two statistical outliers were converted to the 98 -percentile value of the sample distribution. Descriptive data of this part please see Table 7 and Figure 7. These two transfer questions were the same as the pretest questions. The mean pre-post gain was $0.68, \mathrm{SD}=1.33$, which was significantly different from 0 , $\mathrm{t}(122)=5.66, \mathrm{p}<.001$. However, the low pre-post gain indicated that transfer was inherently difficult.

The E-M-F sequencing method with explicit level-bridging scaffolding was the most effective treatment in teaching the deep causal structure of the system. As statistical evidence for the claim, the interaction of the sequencing methods contrast (E-M-F vs. other) and the explicit level-bridging scaffolding variable was significant, $\mathrm{t}(116)=2.04, \mathrm{p}=0.044$. This indicated that a "bottom-up" approach aligned with the causal structure was effective only when explicit level-
bridging scaffolding was provided. The results provided evidence to Hypothesis 2 and Hypothesis 4.

Table 7. Transfer_Understanding of the deep causal structure

|  | $\mathrm{F}-\mathrm{M}-\mathrm{E}$ | $\mathrm{E}-\mathrm{M}-\mathrm{F}$ | M-E-F | Marginal |
| :--- | :--- | :--- | :--- | :--- |
| Explicit | $\mathrm{M}=1.95$ | $\mathrm{M}=2.97$ | $\mathrm{M}=1.80$ | $\mathrm{M}=2.21$ |
| Level- | $\mathrm{SD}=1.15$ | $\mathrm{SD}=1.72$ | $\mathrm{SD}=1.46$ | $\mathrm{SD}=1.34$ |
| bridging | $\mathrm{N}=20$ | $\mathrm{~N}=19$ | $\mathrm{~N}=22$ | $\mathrm{~N}=61$ |
| Implicit | $\mathrm{M}=1.85$ | $\mathrm{M}=1.82$ | $\mathrm{M}=1.63$ | $\mathrm{M}=1.78$ |
| level- | $\mathrm{SD}=1.55$ | $\mathrm{SD}=1.22$ | $\mathrm{SD}=1.30$ | $\mathrm{SD}=1.34$ |
| bridging | $\mathrm{N}=20$ | $\mathrm{~N}=23$ | $\mathrm{~N}=19$ | $\mathrm{~N}=62$ |
| Marginal | $\mathrm{M}=1.90$ | $\mathrm{M}=2.35$ | $\mathrm{M}=1.71$ | Total |
|  | $\mathrm{SD}=1.35$ | $\mathrm{SD}=1.56$ | $\mathrm{SD}=1.34$ | $\mathrm{M}=2.00$ |
|  | $\mathrm{~N}=40$ | $\mathrm{~N}=42$ | $\mathrm{~N}=41$ | $\mathrm{SD}=1.45$ |
|  |  |  |  | $\mathrm{~N}=123$ |



Figure 7: Transfer_Understanding of the deep causal structure

## Conclusion

Scaffolding is critical in teaching complex systems. Different sequencing methods as procedural scaffolding were compared in this study. The F-M-E sequencing method which followed the "concrete to abstract" sequencing principle produced better knowledge integration. The E-M-F sequencing method which followed the "cause to effect" principle produced better understanding of the deep causal structure only when explicit level-bridging scaffolding was provided. The M-E-F sequencing which did not follow either principle was not very effective for either knowledge integration or understanding of the deep causal structure. These findings are valuable as they address the "top-down" vs. "bottom-up" debate in teaching complex systems. When teaching systems with many levels and detailed system dynamics, effective knowledge integration is very essential at an early stage, thus the "top-down" approach starting from concrete macro-level functions may produce better performance. While in teaching complex systems with abstract and implicit causal structures, a sequencing method aligned with the causal structure of the system may help learners construct better mental models for transfer. Different sequencing methods can be used in different contexts or at different learning stages.

The results also showed that explicit level-bridging scaffolding had positive effects on both knowledge integration and understanding of the causal structure. From the perspective of knowledge integration, level-bridging scaffolding and the F-M-E sequencing method had additive effects. In learning the deep causal structure, merely delivering the system knowledge in a "bottom-up" approach was not sufficient, and explicit level-bridging scaffolding was necessary in this process. The positive effects of the explicit level-bridging scaffolding are worth emphasizing. We need to explicitly encourage learners to make connections across system levels via inter-level questions and technology-enhanced techniques (e.g. dynamic link of two simulations). Future research is needed to study the separate effects of different level-bridging scaffolding strategies.

## References

Bar-Yam, Y. (1997). Dynamics of complex systems. Reading, MA: Addison-Wesley.
Chi, M. T. H., Deleeuw, N., Chiu, M. H., \& La Vancer, C. (1994) Eliciting self-explanations improves understanding. Cognitive Science, 18, 439-477.
Hmelo-Silver, C. E., \& Azevedo, R. (2006). Understanding complex systems: some core challenges. The Journal of the Learning Sciences, 15(1), 53-61.
Jacobson, M. J., Kapur, M., So, H., \& Lee, J. (2011). The ontologies of complexity and learning about complex systems. Instructional Science, 39, 763-783.
Jacobson, M. J. \& Wilensky, U. (2006). Complex systems in education: Scientific and educational importance and implications for learning sciences. The Journal of the Learning Sciences, 15(1), 11-34.
Levy, S. T. \& Wilensky, U. (2009). Crossing levels and representations: The connected chemistry (CC1) curriculum. Journal of Science Education \& Technology, 18(3), 224-242.
Liu, L., \& Hmelo-Silver, C. E. (2009). Promoting complex systems learning through the use of conceputal representations in hypermedia. Journal of Research in Science Teaching, 46(9), 1023-1040.
Linn, M. C. (2006). The knowledge integration perspective on learning and instruction. In K. Sawyer (Ed.). The Cambridge handbook of the learning sciences (pp. 243264). Cambridge University Press: Cambridge, UK.

Stieff, M. (2011). Improving representational competence using molecular simulations embedded in inquiry activities. Journal of Research in Science Teaching, 48(10), 1137-1158.
van der Meij, J., \& de Jong, T. (2006). Supporting students’ learning with multiple representations in a dynamic simulation-based learning environment. Learning and Instruction, 16, 199-212.
Wilensky, U., \& Stroup, W. (2002 April). Participatory Simulations: Envisioning the networked classroom as a way to support systems learning for all. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

# General and Efficient Cognitive Model Discovery Using a Simulated Student 

Nan Li (nli1@cs.cmu.edu)<br>Eliane Stampfer (estampfe@cs.cmu.edu)<br>William W. Cohen (wcohen@cs.cmu.edu)<br>Kenneth R. Koedinger (koedinger@cs.cmu.edu)<br>School of Computer Science, Carnegie Mellon University<br>5000 Forbes Ave., Pittsburgh PA 15213 USA


#### Abstract

In order to better understand how humans acquire knowledge, one of the essential goals in cognitive science is to build a cognitive model of human learning. Moreover, a cognitive model that better matches student behavior will often yield better instruction in intelligent tutoring systems. However, manual construction of such cognitive models is time consuming, and requires domain expertise. Further, manually-constructed models may still miss distinctions in learning which are important for instruction. Our prior work proposed an approach that finds cognitive models using a state-of-the-art learning agent, SimStudent, and we demonstrated that, for algebra learning, the agent can find a better cognitive model than human experts. To ensure the generality of that proposed approach, we further apply it to three domains: algebra, stoichiometry, and fraction addition. To evaluate the quality of the cognitive models discovered, we measured how well the cognitive models fit to student learning curve data. In two of those domains, SimStudent directly discovers a cognitive model that predicts human student behavior better than the human-generated model. In fraction addition, SimStudent supported discovery of a better cognitive model in combination with another automated cognitive model discovery method.


Keywords: cognitive model, machine learning, simulated student

## Introduction

One of the fundamental goals in cognitive science is to understand human knowledge acquisition. A cognitive model of human learning that fits data would be a significant achievement. This goal also complements with another goal in education, which is to provide individualized instruction based on students' abilities, learning styes, etc. Cognitive models provide intelligent tutoring systems with useful information on the learning task difficulties and transfer of learning among similar problems. A better cognitive model often leads to more effective tutoring. A cognitive model is a system that can solve problems in the various ways human students can. One common way of representing a cognitive model is a set of knowledge components ( $K C$ ) (Koedinger \& McLaughlin, 2010). The set of KCs includes the component skills, concepts, or percepts that a student must learn to be successful on the target tasks. For example, a KC "divide" in algebra encodes how to proceed given problems of the form $N v=N$ (e.g., $-3 x=6$ ), where $N$ stands for a number, and $v$ stands for a variable.

Nevertheless, manual construction of cognitive models remains time consuming and error prone. Traditional ways to construct cognitive models include structured interviews, think-aloud protocols, and rational analysis. Manual construction of cognitive models requires domain expertise, and
important instructional details may still be overlooked. Automated search methods such as Learning Factor Analysis (LFA) (Cen, Koedinger, \& Junker, 2006) are more objective: the algorithm searches through the space of human-provided factors to find a cognitive model that best matches with human data. Although automated search methods have found better models than manual construction, the quality of the discovered model depends on the quality of the human-provided factors. If there is a better model that can not be expressed by known factors, LFA will not be able to uncover it.

In Li, Matsuda, Cohen, and Koedinger (2011), we have proposed to use the state-of-the-art learning agent, SimStudent (Matsuda, Lee, Cohen, \& Koedinger, 2009), to automatically discover cognitive models without depending on human-provided factors. SimStudent learns skill knowledge from demonstration and problem solving experience. Each skill SimStudent acquires corresponds to a KC in the cognitive model. To demonstrate the generality of this approach, we present evaluations of the SimStudent-generated models in three domains: algebra, stoichiometry, and fraction addition. We validate the quality of the cognitive models using human student data as in Koedinger and MacLaren (1997). Instead of matching with performance data, we use the discovered cognitive model to predict human learning curve data. Experimental results show that for algebra and stoichiometry, SimStudent directly finds a better cognitive model than humans. For fraction addition, SimStudent results assist LFA in finding a better cognitive model than a domain expert. We have also carried out an in-depth study using Focused Benefits Investigation (FBI) (Koedinger, McLaughlin, \& Stamper, 2012) to better understand this machine learning approach, and discussed possible ways of further improvements.

## A Brief Review of SimStudent

SimStudent is an intelligent agent that inductively learns skills to solve problems from demonstrated solutions and from problem solving experience. It is a realization of programming by demonstration (Lau \& Weld, 1998) using a variation of the version space algorithm (Mitchell, 1982), inductive logic programming (Muggleton \& Raedt, 1994), and iterative-deepening depth-first search as underlying learning techniques. For more details, please refer to Matsuda et al. (2009). Recently, in order to build a more human-like intelligent agent, we have developed a model of representation learning, and integrated it into SimStudent's skill acquisition mechanism.


Figure 1: Original and extended production rules for divide in a readable format.

## Tutoring Strategy

To learn, SimStudent interacts with a tutor (human or automated). Given a problem, if SimStudent does not know how to solve it, it will ask the tutor to demonstrate a next step. If SimStudent knows how to proceed, it will propose the next step, and ask for feedback from the tutor. For each demonstrated step, the tutor specifies a tuple of $\langle$ selection, action, input $\rangle$ (SAI tuple) for a skill along with a skill label (e.g., divide). For instance, a demonstrated step for skill "divide" is $\langle(-3 x, 6)$, input text, (divide -3$)\rangle$.

SimStudent learns skills as production rules. The left side of Figure 1 shows an example production rule for skill divide in a readable format. A production rule shows "where" to look for useful information (i.e., perceptual information), "when" to apply the skill (i.e., precondition), and "how" to proceed (i.e., operator sequence). To illustrate, consider the rule on the left side of Figure 1. It states that given an equation (e.g., $-3 x=6$ ), if the left side does not have a constant term, then first get the coefficient of the left side (-3), and divide both sides by that coefficient.

Each skill corresponds to a KC. During training, the tutor can provide an initial set of KCs to SimStudent by labeling each demonstrated step with a skill name. The label given to SimStudent in the example production rule (i.e., the left side of Figure 1) is "divide". If no such initial KC is known, the tutor can simply label all of the steps with the same skill name. SimStudent's learning algorithm will automatically create the cognitive model as needed.

## Skill Learning

SimStudent has three learning components - each acquires one part of the production rules. The first component is a perceptual information (i.e., "where") learner that acquires the path to identify the useful information from its environment. In our case, the environment is a graphical user interface, but it could also be a physical world or an educational game. The elements in the environment are organized in a tree structure. SimStudent learns perception by moving from specific to general. That is, SimStudent tries to find the least general path in the perceptual hierarchy that covers all of the selections in the demonstrated steps. In the example skill "divide," the left and right sides of the equation can be found in
the last row that SimStudent entered input.
The second part of the learning mechanism is a precondition (i.e., "when") learner, which acquires the descriptions of desired situations in applying the skill. The learner is given a set of feature predicates to get a basic understanding of the problem. Each predicate is a boolean function that describes relations among objects in the domain (e.g. (has-coefficient $-3 x)$ ). The precondition learner utilizes FOIL (Quinlan, 1990), an inductive logic programming system that learns Horn clauses from both positive and negative examples. The learning process is general to specific, where the precondition learner starts by considering all situations applicable, and then gradually narrows down the condition based on negative examples. The precondition acquired for the example skill divide is (not (has-constant ?var-left)), which returns true if the left side does not have a constant.

The last component is the operator sequence (i.e., "how") learner. The learner is given a set of operator functions as prior knowledge. Operator functions specify (ideally) basic manipulations (e.g. (add 12 ), (get-coefficient $-3 x$ )) that SimStudent can apply to the problem. Given all of the demonstrated steps, the learning mechanism searches for the shortest operator sequence that could explain all of the records, using iterative-deepening depth-first search. For example, given a demonstrated step $\langle(-3 x, 6)$, (divide -3$)\rangle$, the shortest explanation sequence is (bind? coef (get-coefficient ?left-var)) (bind ?output (divide ?coef)).

There are two groups of operator functions, domainindependent operator functions and domain-specific operator functions. Domain-independent operator functions (e.g. (add 12 )) are basic skills applicable across multiple domains. Human students often have knowledge of these simple skills prior to class. Domain-specific operator functions (e.g. (add-term $5 x-25$ ), (get-coefficient $-3 x$ )) are more complicated skills that human students may not know before class. Thus providing such operators to SimStudent may produce learning behavior that is distinctly different from human students (Matsuda et al., 2009). As we will explain in the next subsection, by integrating representation learning with skill learning, we can reduce or remove SimStudent's dependency on domain-specific operator functions.

Finally, let's talk about how the KCs are discovered. SimStudent starts with a given set of skill labels associated with demonstrated steps. SimStudent tries to learn one rule for each label. It will fail when the perceptual information learner cannot find one path that covers all demonstrated steps, or the operator sequence learner cannot find one operator function sequence that explains all records. In that case, SimStudent learns a disjunctive rule just for the last record. This effectively splits the examples into two clusters. Later, for each new record, SimStudent tries to acquire a rule for each of the clusters with the new record, and stops whenever it successfully learns a rule with one of the clusters. If the new record cannot be added to any of the existing clusters, SimStudent creates another new cluster. By the end of learning, the set of


Figure 2: Different parse trees for -3 x and -x .
clusters defines a new cognitive model.

## Integrating Representation Learning with Skill Learning

As we can see, the prior knowledge given to SimStudent (e.g., the perceptual hierarchy, the operator functions) largely affects the cognitive model it discovers. The more knowledge engineering needed, the less human-like SimStudent is. Therefore, to get a better cognitive model, we need to reduce the amount of knowledge engineering required in constructing SimStudent. Previous studies (Chase \& Simon, 1973) have shown that one of the key differences between experts and novices is their different representations of the world. Recently, we have extended SimStudent to support representation learning, and integrated it into skill learning. It has been shown that by integrating representation learning and skill learning, we can automatically learn the tree-structured representation of the problem, and reduce or remove the need of domain-specific operator functions.

The representation learner extends a grammar induction technique to acquire a probabilistic context-free grammar (pCFG) for the problems based on a set of observations (e.g., $-3 x, 2 x+5)$. To integrate representation learning with skill learning, we extend the perceptual hierarchy to further include the most probable parse trees from the learned pCFG in the contents of the leaf nodes. For example, the left side of Figure 2 is a subtree for parsing " $-3 x$ " and is connected to the node associated with $-3 x$ in the perceptual hierarchy. This subtree ensures that the coefficient -3 is explicitly represented in the perceptual hierarchy. Then, the perceptual information learner and the operator function sequence learner determine how to extract the coefficient from the perceptual hierarchy. This path for identifying the coefficient is added to the perceptual information part of the production rules (See Figure 1, right side). Then, the operator function sequence part no longer needs the domain-specific operator function "get-coefficient". For more details, please refer to Li, Cohen, and Koedinger (2012).

## Cognitive Model Discovery Study

In Li et al. (2011), we demonstrated the effectiveness of using SimStudent to discover cognitive models in an algebra domain. In order to evaluate the generality of the proposed
approach, in this paper, we tested SimStudent in three domains, algebra, stoichiometry, and fraction addition.

## Method

In each domain, we compared the SimStudent model with the best human-generated model available, made by domain experts. To generate the SimStudent model, SimStudent was trained by interacting with automated tutors that simulate the automated tutors used by human students in the studies. The video demonstration in the original study was not used in training SimStudent. SimStudent was trained on problems used by humans students. Then, for each step a human student performed, we assigned the applicable production rule as the KC associated with that step. In cases where there was no applicable production rule, we coded the step using the human-generated cognitive model. Each time a student encounters a step using some KC is considered as an opportunity for that student to show mastery of that KC.

To evaluate the quality of the cognitive model, we measured how well the cognitive model fits with human student data using the Additive Factor Model (AFM) (Cen et al., 2006) to validate the coded steps. AFM is an instance of logistic regression that predicts the probability of a student making an error on the next step given each student, each KC , and the KC by opportunity interaction as independent variables.

$$
\ln \frac{p_{i j}}{1-p_{i j}}=\theta_{i}+\sum_{k} \beta_{k} Q_{k j}+\sum_{k} Q_{k j}\left(\gamma_{k} N_{i k}\right) .
$$

Where:
i represents a student i .
j represents a step j .
$\mathbf{k}$ represents a skill or KC k .
$p_{i j}$ is the probability that student i would be correct on step j.
$\theta_{i}$ is the coefficient for proficiency of student i.
$\beta_{k}$ is coefficient for difficulty of the skill or KC k
$Q_{k j}$ is the Q -matrix cell for step j using skill k .
$\gamma_{k}$ is the coefficient for the learning rate of skill k ;
$N_{i k}$ is the number of practice opportunities student i has had on the skill k;

## Domains

We carried out our study in three domains: algebra, stoichiometry, and fraction addition. The domains as well as the setup in each domain vary from one to another. This ensures that the experiment tests the generality of the proposed approach.

In algebra, we analyzed data from 71 students who used an Carnegie Learning Algebra I Tutor unit on equation solving. The students were typical students at a vocational-technical school in a rural/suburban area outside of Pittsburgh, PA. A

Table 1: AIC on SimStudent-Generated models and Human-Generated Models.

|  | Human-Generated <br> Model | SimStudent-Discovered <br> Model |
| :--- | :--- | :--- |
| Algebra | 6534.07 | $\mathbf{6 4 4 8 . 1}$ |
| Stoichiometry | 17380.9 | $\mathbf{1 7 2 1 8 . 5}$ |
| Fraction Addition | $\mathbf{2 1 1 2 . 8 2}$ | 2202.02 |

total of 19,683 transactions between the students and the Algebra Tutor were recorded, where each transaction represents an attempt or inquiry made by the student, and the feedback given by the tutor. We selected 40 problems that were used to teach students as the training set for SimStudent.

The stoichiometry dataset contains data from 3 studies. 510 high school and college students participated in the studies, and generated 172,060 transactions. Instructional videos on stoichiometry were intermingled with the problems. Instructional materials were provided via the Internet. It took students from 1.5 hours to 6.5 hours to complete the study. 8 problems in this study were used in training SimStudent.

In fraction addition, we analyzed data from 24 students who used an intelligent tutoring system as part of a larger study. Approximately half of the students were recruited from local schools. Students were given immediate correctness feedback on each step, and were offered on-demand text hints. Each interaction was logged through Datashop, and the 24 students yielded 4558 transactions. SimStudent was tutored with 20 problems from this study.

## Measurements

We used Akaike Information Criterion (AIC) and a 10 -fold cross validation (CV) to test how well the generated model predicts the correctness of human student behavior. AIC measures the fit to student data and penalizes over-fitting. We did not use BIC (Bayesian Information Criterion) as the fit metric, because based on past analysis across multiple DataShop datasets, it has been shown that AIC is a better predictor of cross validation than BIC is. The cross validation was performed over ten folds with the constraint that each of the training sets must have data points for each student and KC. We reported the root mean-squared error (RMSE) averaged over ten test sets.

## Results

As shown in Table 1 and Table 2, in algebra and stoichiometry, the SimStudent-discovered models that have lower AICs and RMSEs ( $p<0.001$ ) than the human-generated models. This means the SimStudent models better match the data (without over-fitting). However, in fraction addition, the human-generated model performs better than the SimStudent-discovered ones.

A closer look at the models reveals that in algebra, the SimStudent-discovered model splits some of the KCs in the human-generated model into finer grain sizes. For example, SimStudent creates two KCs for division, one for problems of the form $N v=N$, and one for problems of the form
$-v=N$. This is caused by the different parse trees for $N v$ and $-v$ as shown in Figure 2. Due to this split, the SimStudentgenerated model predicts a higher error rate on problems of the form $-v=N$ than problems of the form $N v=N$. It matches with human student error rates better than the human-generated model, which does not differentiate problems of these two forms.

In stoichiometry, instead of finding splits of existing KCs, SimStudent discovers new KCs that overlap with the original KCs. There are three basic sets of skills in this domain. Within each set, the human-generated KCs are assigned based on the location of the input, while the SimStudent-discovered KCs are associated with the goals of the input. Hence, suppose in two different problems, there are two inputs at the same location in the interface. If they are associated with different goals, the human-generated model will not differentiate them, while the SimStudent-discovered model will put them into two KCs. This indicates that SimStudent not only splits existing KCs, but also discovers totally different KCs.

The fraction addition problem set consists of three types of problems in increasing difficulty: 1) addends have equal denominators; 2) the denominator of one addend is a multiple of the other; 3) addends have unrelated denominators. The human-generated model differentiates these three types of problems in calculating the common denominators and the scaled numerators, and ends up having six KCs. SimStudent, however, associates all of the numerator scaling steps with one KC and associates the common denominator calculations with two KCs. In other words, in this domain, SimStudent partially recovered three out of six KCs, but did not further split them into six KCs. SimStudent did discover the other three KCs, but eventually removed them when they were superseded by more generalized rules. This bias towards more general production rules over specific ones regardless of computational cost appears to be a limitation of SimStudent as a cognitive model. Perhaps if we had let SimStudent keep a utility function for each production rule and retrieve them based on the computational cost, last retrieval time, and correctness, SimStudent may have arrived at all six KCs in the human-generated model.

## FBI Analysis and LFA on Fraction Addition

The differences of AIC and RMSE between the models are small. This is partially because the difference between the models is small. FBI, a recently developed technique, is designed to analyze which of these differences improves the model, and by how much. We applied FBI to the SimStudent and human-generated models in each domain to determine

Table 2: CV RMSE on SimStudent-Generated models and Human-Generated Models.

|  | Human-Generated <br> Model | SimStudent-Discovered <br> Model |
| :--- | :--- | :--- |
| Algebra | 0.4024 | $\mathbf{0 . 3 9 9 9}$ |
| Stoichiometry | 0.3501 | $\mathbf{0 . 3 4 8 8}$ |
| Fraction Addition | $\mathbf{0 . 3 2 3 2}$ | 0.3343 |

why the SimStudent models are better in two of the three cases. In the analysis, we set the human-generated models as the base.

FBI shows that in algebra, splitting "divide" reduces the RMSE of those steps by $1.02 \%$. Further, splitting subtraction and addition decreases the RMSE of those steps by $3.78 \%$ and $3.10 \%$, respectively. This also indicates that SimStudent is able to discover KCs of finer grain sizes that match with human data well.

The stoichiometry results are different. SimStudent discovered new KCs that were not part of any existing KCs. Given the 40 KCs in the human-generated model, SimStudent improved 26 of them. The biggest improvement is on skill molecular weight $(4.60 \%)$, since there are sometimes more than one skill applicable to the same step. The humangenerated model misses the additional skill, while the SimStudent model successfully captures both skills.

As described previously, SimStudent did not differentiate the numerator-scaling and common-denominator steps by problem type. This hurts the RMSE of the associated KCs in the SimStudent-generated model. Nevertheless, SimStudent considers finding the common denominator to be a different KC than copying it to the second converted addend. This split decreases by $7.43 \%$ for problems with unrelated denominators, and and by $0.12 \%$ for denominator steps of problems where one addend denominator was a multiple of the other.

Given the above results, we carried out a third study on fraction addition to test that whether the new KCs created by SimStudent can be used to discover better cognitive models. We used LFA to discover cognitive models given two sets of factors. The baseline LFA model was generated based on the factors (KCs) in the human-generated model. The other LFA model was discovered using both the factors (KCs) in the human-generated model and those in the SimStudentgenerated model. Both LFA models were better than the original human-generated model in terms of AIC and RMSE. Moreover, the LFA model using both human-generated and SimStudent-generated factors had better AIC (2061.4) and RMSE ( 0.3189 ) than the baseline LFA model (AIC: 2111.96, RMSE 0.3226). In other words, with the help of SimStudent, LFA discovered better models of human students.

## Related Work

The objective of this paper is to evaluate the generality of the proposed cognitive model discovery approach. Conati and VanLehn (1999) also applied machine learning techniques to generate cognitive models that fit with human data, but
they focused on assessing self-explanation instead of student learning. Additionally, there has been considerable work on comparing the quality of alternative cognitive models. LFA automatically discovers cognitive models, but is limited to the space of the human-provided factors. Other works such as Pavlik, Cen, and Koedinger (2009); Villano (1992) are less dependent on human labeling, but the models generated may be hard to interpret. In contrast, the SimStudent approach has the benefit that the acquired production rules have a precise and usually straightforward interpretation.

Other systems (e.g., Tatsuoka, 1983; Barnes, 2005) use a Q-matrix to find knowledge structure from student response data. Baffes and Mooney (1996) apply theory refinement to the problem of modeling incorrect student behavior. In addition, some research (e.g., Langley \& Ohlsson, 1984) uses artificial intelligent techniques to construct models that explain student's behavior in math domains. Besides SimStudent, there has also been considerable research on models of high-level learning (e.g., Laird, Rosenbloom, \& Newell, 1986; Anderson, 1993; Taatgen \& Lee, 2003; Sun, 2007; Tenenbaum \& Griffiths, 2001; Schmid \& Kitzelmann, 2011). Other research on creating simulated students (e.g., Chan \& Chou, 1997) is also closely related to our work. Nevertheless, none of the above approaches focused on modeling how representation learning affects skill learning. Moreover, none of them compared the system with human learning curve data. To the best of our knowledge, our work is the first combination of the two whereby we use cognitive model evaluation techniques to assess the quality of a simulated learner, and demonstrate it across multiple domains.

## Conclusion

In this paper, we evaluated the generality of an automatic cognitive model discovery method, and carried out an in-depth analysis to better understand the proposed approach. To avoid over-generalization of KCs, we would like to further extend the skill learning component to maintain utilities associated with each production rule. Further, we plan to investigate discovery of cognitive models for individual students, to provide more personalized learning. Finally, we plan to further integrate the perceptual learning component into skill learning, so that the representation acquired by the learner is refined during the process learning.

In the study, we show that the integration of the representation learning component into skill learning is key to the success of SimStudent in discovering cognitive models. Results indicate that in two out of three domains,

SimStudent-generated models are better predictors of human students' learning performance than human-coded models. For the third domain, when given the SimStudent- and human-generated KCs, LFA finds a better model than the human-generated one. A closer analysis shows that SimStudent is able to split existing KCs into finer grain sizes, discover new KCs, and uncover expert blind spots.

## Acknowledgments

The authors would like to thank Hui Cheng for running the experiments. This work was supported by the Pittsburgh Science of Learning Center, NSF Grant \#SBE-0836012.

## References

Anderson, J. R. (1993). Rules of the Mind. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
Baffes, P., \& Mooney, R. (1996). Refinement-based student modeling and automated bug library construction. Journal of Artificial Intelligence in Education, 7(1), 75-116.
Barnes, T. (2005). The Q-matrix method: Mining student response data for knowledge. In Proceedings aaai workshop educational data mining (pp. 1-8). Pittsburgh, PA.
Cen, H., Koedinger, K., \& Junker, B. (2006). Learning factors analysis - a general method for cognitive model evaluation and improvement. In Proceedings of the 8th international conference on intelligent tutoring systems (pp. 164-175).
Chan, T.-W., \& Chou, C.-Y. (1997). Exploring the design of computer supports for reciprocal tutoring. International Journal of Artificial Intelligence in Education, 8, 1-29.
Chase, W. G., \& Simon, H. A. (1973, January). Perception in chess. Cognitive Psychology, 4(1), 55-81.
Conati, C., \& VanLehn, K. (1999). A student model to assess self-explanation while learning from examples. In Proceedings of the seventh international conference on user modeling (pp. 303-305). Secaucus, NJ: SpringerVerlag New York, Inc.
Koedinger, K. R., \& MacLaren, B. A. (1997). Implicit strategies and errors in an improved model of early algebra problem solving. In Proceedings of the nineteenth annual conference of the cognitive science society (p. 382-387). Hillsdale, NJ: Erlbaum.
Koedinger, K. R., \& McLaughlin, E. A. (2010). Seeing language learning inside the math: Cognitive analysis yields transfer. In Proceedings of the 32nd annual conference of the cognitive science society (pp. 471-476). Austin, TX.
Koedinger, K. R., McLaughlin, E. A., \& Stamper, J. C. (2012). Automated student model improvement. In Proceedings of the 5th international conference on educational data mining (p. 17-24). Chania, Greece.
Laird, J. E., Rosenbloom, P. S., \& Newell, A. (1986). Chunk-
ing in soar: The anatomy of a general learning mechanism. Machine Learning, 1, 11-46.
Langley, P., \& Ohlsson, S. (1984). Automated cognitive modeling. In Proceedings of the fourth national conference on artificial intelligence (p. 193-197). Austin, TX: Morgan Kaufmann.
Lau, T., \& Weld, D. S. (1998). Programming by demonstration: An inductive learning formulation. In Proceedings of the 1999 international conference on intelligence user interfaces (pp. 145-152).
Li, N., Cohen, W. W., \& Koedinger, K. R. (2012). Efficient cross-domain learning of complex skills. In Proceedings of the eleventh international conference on intelligent tutoring systems (pp. 493-498). Berlin: SpringerVerlag.
Li, N., Matsuda, N., Cohen, W. W., \& Koedinger, K. R. (2011). A machine learning approach for automatic student model discovery. In Proceedings of the 4th international conference on educational data mining, eindhoven (p. 31-40).
Matsuda, N., Lee, A., Cohen, W. W., \& Koedinger, K. R. (2009). A computational model of how learner errors arise from weak prior knowledge. In Proceedings of conference of the cognitive science society.
Mitchell, T. (1982). Generalization as search. Artificial Intelligence, 18(2), 203-226.
Muggleton, S., \& Raedt, L. de. (1994). Inductive logic programming: Theory and methods. Journal of Logic Programming, 19, 629-679.
Pavlik, P. I., Cen, H., \& Koedinger, K. R. (2009). Learning Factors Transfer Analysis: Using Learning Curve Analysis to Automatically Generate Domain Models. In Proceedings of 2nd international conference on educational data mining (pp. 121-130).
Quinlan, J. R. (1990). Learning logical definitions from relations. Mach. Learn., 5(3), 239-266.
Schmid, U., \& Kitzelmann, E. (2011, September). Inductive rule learning on the knowledge level. Cognitive System Research, 12(3-4), 237-248.
Sun, R. (2007, September). Cognitive social simulation incorporating cognitive architectures. IEEE Intelligent Systems, 22(5), 33-39.
Taatgen, N. A., \& Lee, F. J. (2003). Production compilation: A simple mechanism to model complex skill acquisition. Human Factors, 45(1), 61-75.
Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. Journal of Educational Measurement, 345-354.
Tenenbaum, J. B., \& Griffiths, T. L. (2001). Generalization, similarity, and bayesian inference. Behavioral and Brain Sciences, 24, 629-640.
Villano, M. (1992). Probabilistic student models: Bayesian belief networks and knowledge space theory. In Proceedings of the 2 nd international conference on intelligent tutoring systems (p. 491-498). Heidelberg.

# Learned helplessness and generalization 

Falk Lieder (lieder@biomed.ee.ethz.ch)<br>Translational Neuromodeling Unit; University of Zurich, \& ETH Zurich, Zurich, Switzerland<br>Noah D Goodman (ngoodman@stanford.edu)<br>Department of Psychology; Stanford University, Stanford, CA, USA<br>Quentin JM Huys (qhuys@biomed.ee.ethz.ch)<br>Translational Neuromodeling Unit; University of Zurich, \& ETH Zurich and<br>Psychiatric University Hospital Zurich, Zurich, Switzerland


#### Abstract

In learned helplessness experiments, subjects first experience a lack of control in one situation, and then show learning deficits when performing or learning another task in another situation. Generalization, thus, is at the core of the learned helplessness phenomenon. Substantial experimental and theoretical effort has been invested into establishing that a state- and task-independent belief about controllability is necessary. However, to what extent generalization is also sufficient to explain the transfer has not been examined. Here, we show qualitatively and quantitatively that Bayesian learning of action-outcome contingencies at three levels of abstraction is sufficient to account for the key features of learned helplessness, including escape deficits and impairment of appetitive learning after inescapable shocks.


## Introduction

Helplessness is a failure to avoid punishment or obtain rewards even though they are under the agent's control. The aetiology and consequences of helplessness have been studied extensively in the animal learning literature using the learned helplessness paradigm [1]. In this paradigm, helplessness is induced in healthy animals by exposure to inescapable electric shocks. Helplessness is then measured by the subsequent failure to escape escapable shocks in a novel environment. The phenomenon was first demonstrated in dogs in the context of testing the two-factor learning theory using the shuttle-box escape task [2]. In the now classical version of the task, three animals are compared. A master rat experiences electrical shocks. These come on unpredictably, but can be terminated by some action, for instance turning a wheel. A yoked rat is exposed to the exact same sequence of shocks that are delivered to the master rat, but has no action available to terminate the shocks. A third rat, the control, is not exposed to shocks. Compared to the controls, the yoked rats are impaired at acquiring new instrumental responses, but the master rats are either unimpaired or may even show an improvement $[1,3]$. Effects reminiscent of those in animal learning have been demonstrated in humans. For instance, people who have been exposed to uncontrollable loud noise or insoluble problems are more likely give up on solving anagrams in a subsequent task [4]. Hopelessness theory is a translation of the helplessness concepts to the human and attributional realm [5].

Extensive animal and human experimentation and theoretical work have clarified that the crucial compo-
nent is a perceived lack of control rather than more specific explanations. First, exposure to one type of uncontrollable reinforcer in one situation can profoundly impair the acquisition of many other types of behaviours (including escape, jumping, immobility, lever pressing, and complex sequential behaviours) in a wide range of different situations [1]. Second, the effect of inescapable shocks is not due to shock-induced analgesia, because it also impairs reward seeking in the absence of negative reinforcers [6]; and as uncontrollable rewards can also induce helplessness [7-9]. Third, it is not reducible to an interference between learned motor inactivity as it also affects the ability to acquire behavioural suppression as an escape response [10]. Furthermore, in accordance with the finding that the probability of generalization increases with the variability of the examples (see [11]), learned helplessness lasts longer when it is induced by experiencing multiple mild stressors (chronic mild stress; [12]) than when it is induced by a single severe stressor. In conclusion, there is strong evidence suggesting that helplessness is learned by generalizing from one uncontrollable situation to believing that situations are uncontrollable in general.

Maier and Seligman formalized controllability in terms of the conditional probability of the reinforcer RF (reward or punishment) given whether or not action $A$ is taken $(A$ or $\bar{A})[1]$. According to their definition, the agent has control if and only if $P(\mathrm{RF} \mid A) \neq P(\mathrm{RF} \mid \bar{A})$. Huys and Dayan formalized the essence of this definition, i.e. that helplessness exists when altering behaviour does not alter outcomes, for multiple actions, outcomes, and degrees of desirability [13]. Here we continue [13]'s argument that generalization is central to learned helplessness by investigating two crucial interactions between perceived control and generalization:

1. Learning about the controllability of one situation transfers to novel situations via generalization.
2. Abstract knowledge about control determines how strongly the observation that a particular action had a particular effect will be generalized.

Using a hierarchical Bayesian model of action-outcomecontingencies we show that these interactions are sufficient to explain various deficits observed in the learned helplessness paradigm.


Figure 1: Hierarchical Bayesian model of statetransitions and controllability. $\boldsymbol{\theta}_{s, a}$ are the transition probabilities for taking action $a$ in state $s$ (level I). The second level, abstracts away from particular actions and represents the general outcome tendency $\boldsymbol{\beta}_{s}$ of situation $s$ and its controllability $\alpha_{s}$. The third level abstracts away from any particular state and represents how controllable the world is in general $(c)$ and how much states differ with respect to controllability $\left(\sigma_{c}^{2}\right)$.

## Methods

In a novel situation $s$, a rational agent may have to learn how likely each of the available actions $a_{1}, \cdots, a_{m}$ is to lead into each of the potential successor states $s_{1}, \cdots, s_{n}$. In the absence of knowledge about the particular situation $s$, the agent can bring experience in other situations $s^{\prime}$ to bear on the problem, i.e. it can use its knowledge about one part of the transition matrix to inform its belief about others. Hierarchical Bayesian formulations provide a normative framework for such generalizations $[11,14]$. In this section, we present such a model of state-transition probabilities with three levels of hierarchy (see Figures 1 and 2). At the lowest level are the probabilities that taking action $a$ in state $s$ will lead to state $s^{\prime}$. At the second level, the model represents the typical outcome probabilities of actions in any one particular situation $s$ and how different the outcomes of different actions tend to be. The more actions are believed to have similar outcomes, the less control there is. At the third and most abstract level, the model represents knowledge about how controllable situations are in general. In this model beliefs about the world's controllability acts as an over-hypothesis that shapes how the agent learns state-transition probabilities (cf. [14]). Concretely, the agent's belief about the state $S_{t+1}$ resulting from taking action $a$ in state $s$ is a multinomial distribution (Equation 1). The agent assumes that the transition probabilities $\boldsymbol{\theta}_{a, s}$ of the actions a available in state $s$ are all drawn from the same distribution: a Dirichlet distribution with the state-specific mean vector $\boldsymbol{\beta}_{s}$ and

$$
\begin{align*}
\forall a, s: S_{t+1} \mid S_{t}=s, A_{t}=a & \sim \operatorname{Multinomial}\left(\boldsymbol{\theta}_{a, s}\right)  \tag{1}\\
\forall a, s: \theta_{\mathbf{a}, \mathbf{s}} & \sim \operatorname{Dirichlet}\left(\alpha_{s} \cdot \boldsymbol{\beta}_{s}\right)  \tag{2}\\
\forall s:-\log \left(\alpha_{s}\right) & \sim \mathcal{N}\left(c, \sigma_{c}^{2}\right)  \tag{3}\\
\forall s: \beta_{s} & \sim \operatorname{Dirichlet}(\mathbf{1})  \tag{4}\\
c & \sim \mathcal{N}\left(\mu, \sigma_{\mu}^{2}\right)  \tag{5}\\
\sigma_{c}^{2} & \sim \operatorname{InvGamma}\left(\alpha_{\sigma}, \beta_{\sigma}\right) \tag{6}
\end{align*}
$$

Figure 2: The functional dependencies of the graphical model in Figure 1.
a second parameter $\alpha_{s}$ that determines the controllability of situation $s$ (Equation 2). If $\alpha_{s}$ goes to $\infty$, then the agent becomes sure that the transition probabilities $\boldsymbol{\theta}_{a_{1}, s}, \cdots, \boldsymbol{\theta}_{a_{N}, s}$ are independent of the agent's action $a$. This means that the situation is uncontrollable (corresponding to the second notion of controllability in [13]). Values of $\alpha_{s}$ close to zero corresponds to the belief that the transition probabilities $\boldsymbol{\theta}_{a_{1}, s}, \cdots, \boldsymbol{\theta}_{a_{N}, s}$ for different actions are uninformative about each other and hence can differ. To allow for the transfer of knowledge between states, a further level is needed: in addition to its belief about the controllability $\alpha_{s}$ of individual situations $s$, the agent also has a belief about how controllable situations are in general. This belief is described by a normal distribution on $-\log \left(\alpha_{s}\right)$ (Equation 3). The parameter $c=\mathbb{E}[-\log (\alpha)]$ expresses how controllable situations are on average, and $\sigma_{c}^{2}$ expresses how much controllability varies from situation to situation.

The average controllability $c$ and the variability of control $\sigma_{c}^{2}$ are unknown properties of the world that have to be learned from experience. We describe the agent's prior beliefs about these quantities by a normal distribution on $c$ (Equation 5) and an Inverse-Gamma distribution on $\sigma_{c}^{2}$ (Equation 6). In this model helplessness results from a probabilistic belief that one's control over the world is low on average (low $c$ ) and varies very little across situations (low $\sigma_{c}^{2}$ ).

Assuming that this hierarchical Bayesian model captures the subjects' internal representation of transition probabilities and control, we can examine how they infer the controllability $c$ of the world in general from the observations $\mathbf{o}=\left\{\left(s_{1}, a_{1}, s_{2}\right), \cdots,\left(s_{t-1}, a_{t-1}, s_{t}\right)\right\}$ of the state transitions $\left(s_{1}, s_{2}\right), \cdots,\left(s_{t-1}, s_{t}\right)$ and their actions $a_{1}, \cdots, a_{t}$. In addition, we can simulate the weaker generalization on the situation-specific controllability $\alpha_{s}$ and transition-tendency $\boldsymbol{\beta}_{s}$ by computing $P\left(\alpha_{s}, \boldsymbol{\beta}_{s} \mid \mathbf{o}\right)$. Finally, we can investigate how this generalization is shaped by abstract beliefs about control $\left(P\left(c, \sigma_{c}^{2}\right)\right)$.
Simulations The effects of controllable and uncontrollable shocks were modelled by simulating the learning process taking place during the shocks as Bayesian inference on $c$ and $\sigma_{c}^{2}$. The naive subjects' belief about controllability was modelled by a probability distribution with $\mathbb{E}[\alpha]=10$ and $\operatorname{Var}[\alpha]=100$; the variance of the prior beliefs was 100 for $c$ and 0.1 for $\sigma_{c}^{2}$. To model the observations resulting from controllable $\left(\mathbf{o}_{\mathrm{c}}\right)$ versus uncontrollable shocks ( $\mathbf{o}_{\neg \subset}$ ), we assumed one observation
per second during 64 shocks lasting 60 seconds each. For controllable shocks there was one action $\left(a_{1}\right)$ that would always terminate the shock $\left(s_{1} \rightarrow s_{2}\right)$ and four actions that did not $\left(s_{1} \rightarrow s_{1}\right)$, whereas there was no such action for uncontrollable shocks.

Learning after exposure to shocks was modelled as Bayesian inference on $\boldsymbol{\alpha}, \boldsymbol{\beta}$ given $c=\mathbb{E}\left[c \mid o_{\neg c}\right]$ and $\sigma_{c}^{2}=$ $\mathbb{E}\left[\sigma_{c}^{2} \mid o_{\neg c}\right]$ for uncontrollable shocks versus $c=\mathbb{E}\left[c \mid o_{c}\right]$ and $\sigma_{c}^{2}=\mathbb{E}\left[\sigma_{c}^{2} \mid o_{c}\right]$ for controllable shocks.

Inference was performed using Markov chain Monte-Carlo (MCMC) methods. To sample from $P\left(\boldsymbol{\alpha}, \boldsymbol{\beta}, c, \sigma_{c}^{2} \mid \mathbf{o}\right)$, we used a Metropolis-Hastings algorithm with Gaussian random-walk proposals on $c$ and $-\log (\alpha)$. The proposal for $\boldsymbol{\beta}_{t+1}$ was Dirichlet $\left(10 n \cdot \boldsymbol{\beta}_{t}\right)$ where $n$ is the number of states, and the proposal for $\sigma_{c, t+1}^{2}$ was an Inverse-Gamma distribution with mean $\sigma_{c, t}^{2}$ and variance 1. 50 Markov chains were run for 51000 iterations with a burn-in period of 1000 iterations. $P\left(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid c, \sigma_{c}^{2}, \mathbf{o}\right)$ was computed in the same way. The posterior expectation of $\boldsymbol{\theta}$ was computed using Monte-Carlo integration: $\mathbb{E}\left[\boldsymbol{\theta} \mid \mathbf{o}, c, \sigma_{c}^{2}\right]=\int \mathbb{E}[\boldsymbol{\theta} \mid \alpha, \beta, \mathbf{o}] \cdot p\left(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid \mathbf{o}, c, \sigma_{c}^{2}\right) d \boldsymbol{\alpha} d \boldsymbol{\beta} \approx$ $\frac{1}{m} \sum_{i=1}^{m} \mathbb{E}\left[\boldsymbol{\theta} \mid \boldsymbol{\alpha}_{i}, \boldsymbol{\beta}_{i}, \mathbf{o}\right]$ with $\boldsymbol{\alpha}_{i}, \boldsymbol{\beta}_{i} \sim P\left(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid \mathbf{o}, c, \sigma_{c}^{2}\right)$.

## Results

As a first step in assessing whether generalization can account for the differential effects of controllable versus uncontrollable stress, we simulated Bayesian learning from these experiences according to the model shown in Figure 1. Figure 3A shows the simulated changes in perceived controllability induced by the escapable and inescapable shocks administered in the learned helplessness paradigm [15]. After inescapable shocks, the subjects' perceived control $c$ was reduced, whereas controllable shocks increased it. Furthermore, controllable shocks increased the estimated variability of control across situations, whereas no such change was observed after inescapable shocks (Figure 3B). Thus, the two kinds of shocks have opposite effects on the subjects' high-level beliefs about controllability.

We next asked whether the beliefs induced by uncontrollable shocks were sufficient to impair escape learning in a different task, and whether the beliefs induced by controllable shocks have the opposite (mastery) effect. We modelled these beliefs by the inferred mean and variance of $c$ for escapable and inescapable shocks and simulated learning in the shuttle-box escape task. As a first step, we simulated learning from given observations with one action that did $\left(a_{1}\right)$ and four actions that did not cancel the shock $\left(a_{2}, \cdots, a_{5}\right)$. Concretely, we simulated how strongly naive subjects, subjects who had experienced inescapable shocks (yoked), and subjects who had experienced escapable shocks (masters) would believe that action $a_{1}$ cancels the shock after having taken action $a_{1}$ for $0,1,2,4,8,16$ or 32 times and each of the four other actions 8 times. Figure 4 shows that yoked subjects (red) acquired the escape response more slowly than naive subjects (blue): more evidence was required before they believed that the action was efficient in terminating the shock. Furthermore, the model


Figure 3: A: Expected controllability learned from controllable or uncontrollable electric shocks. The values on the x-axis are the change $\Delta c$ relative to the controllability expected by naive subjects and height of the bars shows how strongly the simulated agent beliefs in the corresponding value of $c$. B: Variance of controllability learned from controllable and uncontrollable electric shocks. The values on the x -axis are the change $\Delta c$ relative to the variability expected by naive subjects and height of the bars shows how strongly the simulated agent beliefs in the corresponding value of $\sigma_{c}^{2}$.


Figure 4: Simulated effects of controllable and uncontrollable shocks on the speed of learning that action 1 terminates the shock.
also captured mastery effects, whereby prior exposure to controllable shocks leads to faster learning (green; [16]).

To more quantitatively relate the learning dynamics shown in Figure 4 to empirical data, we simulated learning and decision making in the fixed-ratio operant conditioning task of [17]. In this task, rats have to learn to press a lever, but only every third lever press terminates the shock. This task was modelled as sequentialdecision making. To do so, we partitioned the 60 seconds of each trial in [17]'s experiment into 30 bins, each 2 seconds long, and simulated one decision, one observation, and one belief update for every bin. The simulated choices were to stay still $\left(a_{0}\right)$, to push the lever $\left(a_{1}\right)$, or to perform a different action $\left(a_{2}\right)$. The reward for staying still and receiving the shock was modelled as -1 $\left(r\left(s_{1}, a_{0}, s_{1}\right)=-1\right)$. Moving and receiving a shock was assumed to incur a small additional cost $\left(r\left(s_{1}, a_{2}, s_{1}\right)=\right.$ $-1.2)$. If the action stopped the shock, it was assumed to incur only the cost of movement $\left(r\left(s_{1}, a_{0}, s_{2}\right)=0\right.$ and $r\left(s_{1}, a_{i}, s_{2}\right)=-0.2$ for $\left.i \in\{1,2\}\right)$. We assumed that rats


Figure 5: The left panel shows empirical data acquired by [17] as shown in [1]. The plots on right show our simulations of the experiment.
learn the probability $P\left(S_{t+1}=s_{2} \mid S_{t}=s_{1}, A_{t}=a_{i}\right)$ that an action $a_{i}$ terminates the shock (an alternative model might consider treating different numbers of lever presses as separate actions). The subjects' internal representation of transition probabilities was modelled as the hierarchical Bayesian model shown in Figure 1. The rat's decision making was simulated by a sampling algorithm to produce behaviour akin to probability-matching [18]. Specifically, we assumed that the rat simulates five outcomes $u_{i, j}=r\left(s, a_{i}, s_{j}^{\prime}\right), s_{j}^{\prime} \sim P\left(S_{t+1} \mid S_{t}=s, A_{t}=a_{i}\right)$ of each action $a_{i}$ and chooses the action $a_{i}$ for which the average utility $\frac{1}{5} \sum_{j=1}^{5} u_{i, j}$ was largest, and ties were broken at random. Under these assumptions, the learning dynamics shown in Figure 4 capture the qualitative effects of uncontrollable shocks on the probability to escape shock and the time required to do so [17]: yoked subjects failed to escape more often than naive subjects (Figure 5, left panel), and when they succeeded to escape it took them longer (Figure 5, right panel). Furthermore, our model accounts for the mastery effect that rats who had been exposed to controllable shocks prior to the task, escaped faster than rats with no prior exposure to shock.

As outlined in the introduction, learned helplessness impairs not only the ability to learn from punishments but also from rewards. To assess whether our model captures this effect, we simulated the experiment by [6]. In the experiment's appetitive choice task, rats were rewarded with food for going into one of two chambers after they had been trained to prefer the other chamber. We modelled this task as a sequence of decisions, observations, and belief updates as described above. As Figure 6 shows our model captures that uncontrollable shocks reduced the probability that a rat would first seek out the chamber in which a reward would be delivered. Thus this apparently anhedonic behaviour can be explained purely in terms of impaired associative learning due the generalization that the world is uncontrollable.

Next, we asked whether our model can account for the finding that the effect of learned helplessness is most pronounced in tasks that are complex and require persistence. To answer this question, we simulated decision


Figure 6: Simulation of the appetitive choice distinction task by [6]. Our simulation captures that yoked rats performed worse than naive rats across all 10 blocks of the experiment.


Figure 7: Simulation of the experiment by [19]. Dashed lines are model predictions; diamonds are data points. The three columns correspond to the experimental conditions requiring 1,2 , or 3 lever presses.
making and learning in the experiment by [19]. In this experiment, yoked rats did learn to escape response when one or two, but not when three lever presses were required. In Figure 7, we show that the model can quantitatively capture the increasing penetrance of inescapable shock exposure with increasing escape response requirements.

## Discussion

Our results indicate that a normative account of generalization of action-outcome contingencies is sufficient to produce a wide range of phenomena observed in learned helplessness experiments. The account captures (i) how helplessness is induced by uncontrollable stressors and why it transfers to novel situations, (ii) why controllable stress fails to induce helplessness, (iii) that helplessness results from impaired learning that different actions have different effects, (iv) mastery effects, (v) impaired reward seeking, and (vi) the interaction between helplessness and task requirements. This suggests that the generalization of experienced control may be sufficient to account for many learned helplessness effects.

Note that our model explains helplessness as the consequence of rational learning and generalization (cf. $[11,14])$ from uncontrollable stress. Mirroring the fact that learned helplessness induces depression-like states in healthy animals and affects healthy humans, this sug-
gests one pathway by which learned helplessness may arise as a rational adaptation to an uncontrollable environment rather than from negative biases or dysfunctional information processing. In the present account, the generalization that leads to helplessness is that different actions tend to have the same effect. This generalization can occur at two levels of abstraction: (i) from the outcomes of some actions in a given situation to the outcomes of all actions available in that situation, and (ii) from the controllability of one situation to the controllability of situations in general. Our model predicts that after uncontrollable stress generalizations of the first type will be unusually strong. This may capture overgeneralization - a frequent feature of depressive thought in humans [20] - and the model explains how this learning style increases the risk for helplessness by fostering the belief that all actions are equal.

According to the classical notion [1], control requires that taking an action or not alters the probability of outcomes. Our model generalizes this notion by allowing for multiple actions, multiple outcomes, and two additional levels of abstraction, and it expands it from a binary distinction to a graded, quantitative measure of controllability. As a result, our model instantiates [13]'s second type of controllability which captured on the achievability of different outcomes. While the notion of control presented here does not take into account the outcomes' desirability (type 3 in [13]), it refines [13]'s proposal of how helplessness might be learned by generalization in two regards. First, [13] juxtaposed two extremes of generalization: the controllabilities of different environments are (i) independent (no generalization) or (ii) identical (full generalization). Arguably, both extremes correspond to pathologically inaccurate models of the world. By contrast, the hierarchical generative model proposed here formalizes the more realistic intermediate assumption that although some situations are more controllable than others, knowing about the controllability of one situation is informative about the controllability of other situations. Second, inference in this model captures an important aspect of attribution: Was the outcome due to the action I took or due to the situation I was in? In the model, a perceived lack of control induces misattributions that impair learning: the over-hypothesis that the world is uncontrollable renders implausible any interpretation according to which different actions have different effects. Therefore the perceived lack of control biases people to attribute the outcome of taking action $a$ in state $s$ to the state $s$ rather than to the action that they have taken. For extreme helplessness the situation's action-independent outcome tendency $\boldsymbol{\beta}_{s}$ will be updated just as much as the outcome probabilities $\boldsymbol{\theta}_{s, a}$ of taking action $a$ in this situation. Conversely, the actionspecific outcome probabilities $\boldsymbol{\theta}_{s, a}$ will be updated no more than $\boldsymbol{\beta}_{s}$, and the outcomes of actions $b, c, d, \cdots$ will influence the belief about $\boldsymbol{\theta}_{s, a}$ almost as much as the outcomes of action $a$ itself. This increases the amount of evidence required to discover that there is an action that achieves the goal while most other actions do not, and this is how the perceived lack of control impairs learn-
ing. Thus, according to our model, helpless behaviour in simple tasks results from slowed learning of transition probabilities. This complements a recent model of how the perceived lack of control impairs planning in complex, sequential decision problems [21].

Despite the encouraging results reported here, more research is needed to establish the validity of this modelling approach. Our simulation of the experiment reported in [19] did not fully capture the rats' learning dynamics and overestimated their performance in the simplest condition. These discrepancies could be due to the simplistic assumption that rats do not associate shock termination with the number of lever presses but only with their most recent action. As a result, the fit achieved by the reported simulation is not substantially better than the fit obtained by reducing the subjective intensity of the shock (data not shown). Therefore the results of our simulations (Figure 7) could be mimicked by analgesia [19, 22]. Furthermore, our model failed to capture the precise pattern of the data in [6]. However, the two hypotheses might be discernible using data that reveal the dynamics of learning, or by directly probing subjects' beliefs about outcome probabilities in the absence of rewards.

There are important aspects of helplessness that our model does not capture yet. It needs to assume surprisingly weak priors to explain why helplessness can be learned so rapidly -an assumption that is difficult to extend to mature animals with a lifetime of experience. Conversely, it would be challenging to explain with our model why even severe shock-induced helplessness tends to fade within 48 hours. To capture these two suggestions of temporal (in)stability of helplessness, our model could be extended by replacing the constant $c$ by a Gaussian random walk $C(t)$. This would take into account that how much control a person has, changes throughout his or her life. It could be used to explain why a very brief period of uncontrollable stress can induce helplessness despite a lifetime of experience in a controllable environment. This extension may also explain why the speed of learning about controllability increases with the environment's perceived volatility (cf. [23, 24]). Beyond these cognitive signatures of helplessness, our model has yet to engage with the affective and subjective aspects of helplessness.

Our model suggests a number of avenues for future work. As normative inference, exposure to sufficient controllability will lead to the correct inference, suggesting that experience of control should heal helplessness, and thus that the escape deficit should be temporary. Why and how the experience of shocks that an animal did not attempt to escape may nurture helpless beliefs may necessitate adaptations in the inference. Nevertheless, the model does capture that controllable stress can immunize subjects against the effects of uncontrollable stress, and suggests two computational mechanisms: first a higher expected controllability, and second a higher estimate of the variability of control across situations. If the initial expected controllability is higher, then a given reduction in perceived control may no longer be sufficient
to induce learned helplessness. Alternatively, a higher estimate of the variability of control across situations would reduce the strength of the generalization from uncontrollable stress in one situation to the controllability of the world in general. This theoretical work does, however, also echo the likely limited benefits of exposure because a subjects' helpless choices in some situations, paired with strong tendencies to generalise, can produce sufficient evidence of lack of control to drown any islands of controllability.

Our model may be able to illuminate why and how the effects of chronic-mild stress [12] differ from the effects of severe-acute stress in their scope, severity, and duration. The results suggests that the underlying mechanisms can be understood in terms of well-studied general generalization phenomena [11]. For instance, since the strength of a generalization increases with the variability of the examples, the experience of multiple stressors should render the effects of chronic mild stress more general than the equivalent amount of stress experienced in a single situation.

Learned helplessness is a behavioural paradigm with parallels in humans and animal models, and with established validity in research on depression [25]. Using methods from reinforcement and machine learning, this work has shown that abstract learning and generalization about controllability explain many of the key features of learned helplessness in animals.

## References

[1] S. F. Maier and M. E. Seligman, "Learned helplessness: Theory and evidence.," J EXP PSYCHOL GEN, vol. 105, pp. 3-46, 1976.
[2] B. Overmier and M. Seligman, "Effects of inescapable shock upon subsequent escape and avoidance responding.," J COMP PHYSIOL PSYCH, vol. 63, pp. 28-33, 1967.
[3] S. F. Maier and L. R. Watkins, "Stressor controllability and learned helplessness: the roles of the dorsal raphe nucleus, serotonin, and corticotropin-releasing factor.," NEUROSCI BIOBEHAV R, vol. 29, pp. 829-841, 2005.
[4] C. Peterson, S. F. Maier, and M. E. P. Seligman, Learned Helplessness: A theory for the age of personal control. OUP, 1993.
[5] L. Y. Abramson, G. I. Metalsky, and L. B. Alloy, "Hopelessness depression: A theory-based subtype of depression," Psychol. Rev., vol. 96, pp. 358-372, 1989.
[6] R. A. Rosellini, J. P. DeCola, and N. R. Shapiro, "Crossmotivational effects of inescapable shock are associative in nature.," J EXP PSYCHOL ANIM B, vol. 8, pp. 376388, 1982.
[7] F. Goodkin, "Rats learn the relationship between responding and environmental events: An expansion of the learned helplessness hypothesis," LEARN MOTIV, vol. 7, 1976.
[8] R. L. Welker, "Acquisition of a free-operant-appetitive response in pigeons as a function of prior experience with response-independent food," LEARN MOTIV, vol. 7, pp. 394-405, 1976.
[9] Q. J. M. Huys, J. Vogelstein, and P. Dayan, "Psychiatry: Insights into depression through normative decisionmaking models," in Adv. Neural Inf. Process. Syst. 21 (D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, eds.), pp. 729-736, MIT Press, 2009.
[10] R. L. Jackson, S. F. Maier, and P. M. Rapaport, "Exposure to inescapable shock produces both activity and associative deficits in the rat," Learn. Motiv., vol. 9, pp. 69-98, 1978.
[11] J. B. Tenenbaum and T. L. Griffiths, "Generalization, similarity, and bayesian inference.," Behav Brain Sci, vol. 24, 2001.
[12] P. Willner, "Validity, reliability and utility of the chronic mild stress model of depression: a 10-year review and evaluation," Psychopharm, vol. 134, pp. 319-29, 1997.
[13] Q. J. M. Huys and P. Dayan, "A bayesian formulation of behavioral control.," Cognition, vol. 113, pp. 314-328, 2009.
[14] C. Kemp, A. Perfors, and J. B. Tenenbaum, "Learning overhypotheses with hierarchical bayesian models," Dev. Sci., vol. 10, pp. 307-321, 2007.
[15] M. Seligman and S. Maier, "Failure to escape traumatic shock.," J EXP PSYCHOL, vol. 74, pp. 1-9, 1967.
[16] J. R. Volpicelli, R. R. Ulm, A. Altenor, and M. E. P. Seligman, "Learned mastery in the rat," LEARN MOTIV, vol. 14, pp. 204-222, 1983.
[17] R. D. Hannum, R. A. Rosellini, and M. E. Seligman, "Learned helplessness in the rat: Retention and immunization.," DEV PSYCHOL, vol. 12, pp. 449-454, 1976.
[18] R. J. Herrnstein and D. H. Loveland, "Maximizing and matching on concurrent ratio schedules.," J Exp Anal Behav, vol. 24, pp. 107-116, 1975.
[19] M. Seligman and G. Beagley, "Learned helplessness in the rat.," J COMP PHYSIOL PSYCH, vol. 88, pp. 534541, 1975.
[20] A. T. Beck, A. J. Rush, B. F. Shaw, and G. Emery, Cognitive therapy of depression. Guilford Press, 1979.
[21] F. Lieder, N. D. Goodman, and Q. J. M. Huys, "Controllability and resource-rational planning.," Cosyne $A b$ stracts 2013, 2013.
[22] R. L. Jackson, S. F. Maier, and D. J. Coon, "Long-term analgesic effects of inescapable shock and learned helplessness," Science, vol. 206, no. 4414, pp. 91-93, 1979.
[23] T. E. J. Behrens, M. W. Woolrich, M. E. Walton, and M. F. S. Rushworth, "Learning the value of information in an uncertain world.," Nat Neurosci, vol. 10, pp. 12141221, 2007.
[24] C. Mathys, J. Daunizeau, K. J. Friston, and K. E. Stephan, "A bayesian foundation for individual learning under uncertainty.," Front Hum Neurosci, vol. 5, 2011.
[25] L. Y. Abramson, L. B. Alloy, M. E. Hogan, W. G. Whitehouse, M. Cornette, S. Akhavan, and A. Chiara, "Suicidality and cognitive vulnerability to depression among college students: a prospective study.," J Adolesc, vol. 21, pp. 473-487, 1998.

# Is it Time Bayes went Fishing? Bayesian Probabilistic Reasoning in a Category Learning Task 

Marcus Lindskog (marcus.lindskog@psyk.uu.se), Anders Winman (anders.winman@psyk.uu.se), Peter Juslin (peter.juslin@psyk.uu.se)<br>Department of Psychology, Uppsala University, P. O Box 1225, Uppsala, 75142 Sweden


#### Abstract

People have generally been considered poor at probabilistic reasoning, producing subjective probability estimates that far from accord to normative rules. Features of the typical probabilistic reasoning task, however, make strong conclusions difficult. The present study, therefore, combines research on probabilistic reasoning with research on category learning where participants learn base rates and likelihoods in a category-learning task. Later they produce estimates of posterior probability based on the learnt probabilities. The results show that our participants can produce subjective probability estimates that are well calibrated against the normative Bayesian probability and are sensitive to base rates. Further, they have accurate knowledge of both base rate and means of the categories encountered during learning. This indicates that under some conditions people might be better at probabilistic reasoning than what could be expected from previous research.


Keywords: Probabilistic reasoning, category learning, Bayes’ theorem, base rate

## Introduction

Research concerned with human probability judgment has been dominated by the general conclusion that people are poor at reasoning with probabilities because they substitute hard facts about probabilities with subjective variables that are conveniently available (see e.g., Gilovich, Griffin, \& Kahneman, 2002). In fact, with respect to tasks requiring people to integrate probabilities according to Bayes’ theorem the verdict is even harder, as summarized by a quote from Kahneman and Tversky (1972, p. 450): "In his evaluation of evidence, man is apparently not a conservative Bayesian: he is not a Bayesian at all." In the present study, we present results indicating that, at least under some conditions, the claim by Kahneman and Tversky might have been somewhat premature.

To appreciate the kind of task our participants are faced with, imagine going to catch fish in a lake where the fishing authorities have farmed two kinds of bass: copper bass, and silver bass. The two kinds of bass look identical and the only feature that distinguishes them is that the copper bass weighs, on average, a little less than the silver bass. While looking identical they, however, taste very differently. If you want a delicious dinner, you should go for the silver bass while if you want to feel sick you should choose a
copper bass. To make sure that the lake is not over fished the authorities have also decided that, at all time, the ratio of copper to silver bass should be 8:2, a piece of information not made publically available.

The fish scenario illustrates a type of situation that people engage in frequently in their everyday lives. The fishermen estimate the probability of a new object belonging to a category based on previous experience. That is, each time a fish is taken out of the lake the fisherman needs to estimate the probability of a given fish being a copper or a silver bass. The estimate is informed by experience with fish previously taken up out of the lake and cooked for dinner, thus effectively categorized as copper or silver bass. More specifically, this illustrates a situation where an observer needs to learn base rates and likelihoods from experience and later integrate this information to reach an estimate of a posterior probability. In such, the fish scenario incorporates two areas of cognitive psychology: probabilistic reasoning and category learning, that have been extensively investigated separately, but seldom together (but see, Nilsson, Olsson, \& Juslin, 2005).

## Probabilistic Reasoning

Research on human probabilistic reasoning has mainly been concerned with the evaluation of subjective probability estimates against normative rules of probability. In the typical experiment, the subjective estimates are informed by a set of probabilities explicitly stated in the task. Consider, for example, the cab problem (Tversky \& Kahneman, 1980) where participants are asked to estimate the probability of a cab involved in an accident being blue rather than green based on the base rates of blue (.15) and green (.85) cabs and the hit-rate (.8) of an eyewitness with both the base rate and hit-rate being explicitly stated in the task. The normative answer (.41) can be found by integrating the information in the problem using Bayes' theorem.

When presented with the cab problem, and similar problems, people tend to give probability estimates that are much higher than what is implied by Bayes' theorem. Often the modal response is closer to the hit-rate of the eyewitness (.8). This pattern of results is commonly interpreted as a captivation in participants by the hit-rate along with neglect of the base rate (.15). The dominating explanation to this apparent neglect of base rates has been that people are prone
to use judgmental heuristics (e.g. the representativeness heuristic) that ignore base rates (e.g., Kahneman \& Tversky, 1972; but see, Koehler, 1996). More recent accounts of probabilistic reasoning, suggesting that people are prone to linear additive information integration, argue instead that the non-normative answers are the result of how probabilities are integrated rather than the use of heuristics per se (Juslin, Nilsson, \& Winman, 2009; Juslin, Nilsson, Winman, \& Lindskog, 2011).

Regardless of the underlying mechanisms explaining the results, the use of complex normative rules, such as Bayes’ theorem, to integrate probabilities seems to be beyond the ability of most people (e.g., Eddy, 1982; Gigerenzer \& Hoffrage, 1995). In fact, even explicit instructions regarding how to use Bayes' theorem to integrate the information is insufficient to improve people's judgments (Juslin et al., 2011). It should be noted, however, that the despite the somewhat discouraging picture painted by previous research, recent accounts of human cognition (e.g., Oaksford \& Chater, 2009; Tenenbaum, Kemp, Griffiths, \& Goodman, 2011) have indicated that people are rational Bayesian agents with a remarkable ability to integrate information in accordance with the laws of Bayesian probability theory.

The extent to which people's probability estimates in Bayesian reasoning tasks coincide with the normative answer has largely been tested using tasks similar to the cab problem. Three features of these types of tasks are noteworthy, features that might influence the conclusions that can be drawn about human probabilistic reasoning. First, the information to be integrated (base rates, likelihoods, etc.) is explicitly given to participants in the form of probabilities (e.g., Kahneman \& Tversky, 1972) or, sometimes, frequencies (e.g., Gigerenzer \& Hoffrage, 1995). Second, the tasks are commonly set up to give a posterior probability that is low, often .40 or smaller. Finally, the outcome for which the posterior probability is estimated is often binary (blue or green cab, disease or no disease, engineer or lawyer, etc.). All of these task features make it difficult to draw strong conclusions about the ability of people to integrate probabilistic information. In everyday life, people are unlikely to come across situations where probabilities are explicitly stated. They rather encounter situations, like the fishing example above, where probabilities are learned from experience. Many real life situations also include an outcome, for which the posterior probability is estimated, that is continuous rather than binary. Furthermore, the restriction of the range of posterior probabilities makes conclusions about the extent to which people are calibrated against the Bayesian probability difficult due to regression effects. In order to address these three issues it is necessary to find a task where participants learn probabilities from experience and where it is possible to elicit probability estimates on the entire 0 to 1 range for a
continuous outcome variable. One promising candidate is found in category learning.

## Probabilistic Reasoning and Category learning

In the typical categorization task participants are presented with a number of stimuli from two or more categories and are asked to assign an appropriate category to each based on a set of features. During learning, the categorization is often followed by feedback regarding the correct category.

The literature contains several different models of how categorization is made, including prototype, exemplar, and decision-bound models (Ashby \& Maddox, 2005). The purpose of this study is not to distinguish between the different kinds of models. Rather, we draw upon the notion that most models of human categorization make assumptions about: a) how and what information is accessed from the categories and what computations are performed on this information and, b) how a response is selected after computations are made (Ashby \& Alfonso-Reese, 1995). For most models that assume a probabilistic, in contrast to a deterministic, response selection process, the decision rule subjects are assumed to use could be described as; respond category A to stimulus $x$ with probability $M(\mathrm{x})$ where:

$$
\begin{equation*}
M(x)=\frac{\beta_{A} S_{x A}}{\beta_{A} S_{x A}+\beta_{B} S_{x B}} \tag{1}
\end{equation*}
$$

In this expression $\beta_{i}$ is the response bias towards category $i$ and $S_{x i}$ is a measure of the similarity between stimulus $x$ and category $j$. At least under some conditions Eq. 1 can be reduced to

$$
\begin{equation*}
M(x)=\frac{\hat{P}(A) \hat{f}_{A}(x)}{\hat{P}(A) \hat{f}_{A}(x)+\hat{P}(B) \hat{f}_{B}(x)} \tag{2}
\end{equation*}
$$

where $\hat{P}(i)$ and $\hat{f}_{i}$ are estimators of the base rate and probability density function of category $i$ respectively (Anderson, 1991; Ashby \& Alfonso-Reese, 1995). Ashby and Alfonso-Reese (1995) argued that these properties of the categorization task transform it into a density estimation task where participants are faced with estimating base rates and probability density functions of each category. Indeed, several investigations of models of categorization have shown that they are mathematically equivalent to density estimation (e.g., Anderson, 1991; Ashby \& Alfonso-Reese, 1995; Griffiths, Sanborn, Canini, \& Navarro, 2008)
The similarities between Bayes' theorem and Eq. 2 suggest that categorization tasks are similar to probabilistic reasoning tasks with the difference that while probabilities are explicitly stated in the reasoning task they need to be learned from trial-by-trial feedback in the categorization task. Further, while the literature on probabilistic reasoning is somewhat pessimistic about people's ability to integrate
probabilities the categorization literature suggests that people are quite apt at categorization (Ashby \& Maddox, 2005). However, while research on categorization has been extensively concerned with how categories are represented and the processes leading up to a categorization (Ashby \& Maddox, 2005) it has put much less focus on the extent to which base rates and likelihoods are learned. Further, the typical categorization task requires participants to assign a stimulus to a category leaving the question of whether $M(\mathrm{x})$ in Eq. 2 is close to the normative posterior probability unanswered.

It should be noted that categorization research indicates that people are able to learn base rate information from experience (Medin \& Edelson, 1988), at least under some conditions, and that models of categorization can be seen as the cognitive substrate of subjective probability estimates (Nilsson et al., 2005).

## The Present Study

The present study investigates the accuracy of subjective probability estimates in a Bayesian probability reasoning task. Instead of being presented with base rates and likelihoods explicitly, however, participants learn them through experience in a categorization task.

Further, we elicit probabilities from the entire range of possible posterior probabilities for a continuous outcome variable in order to have a task that is as ecologically valid as possible.

To investigate factors that might influence the learning of base rates and likelihoods as well as the process used to elicit probability estimates, we manipulate both base rate and the distance between categories (i.e., the likelihood ratio).

## Method

## Participants

Participants were 40 ( 24 female and 16 male) undergraduate students from Uppsala University with a mean age of 25.1 years ( $S D=4.3$ years). They received a movie ticket or course credits for their participation.

## Design

The experiment used a $2 x 2$ between-subjects design with base-rate-ratios (8:2/6:4) and category-distance (short / far) as independent variables.

## Materials and Procedure

The computerized task was carried out on a PC and consisted of a learning phase and a test phase. On each of the 200 trials in the learning phase, participants categorized an exemplar to one of two categories ( A and B ) along a single dimension. The number of exemplars from each category was determined by the base-rate-ratio. In the 8:2condition the ratio of the number of exemplars in the two
categories was 8:2 (i.e., 160 A-exemplars and 40 Bexemplars) and in the other condition it was $6: 4$. The 200 items were presented in an individually randomized order.

A unique training set was created for each participant by randomly sampling stimuli from two Gaussian distributions with equal standard deviation $(\sigma=6)$. In the short categorydistance condition, the mean of the two Gaussians were 40 and 49 respectively while in the far condition they were 40 and 52. Whether category A or B had the highest mean was counterbalanced over participants.

The experiment used two cover stories. Either the categories where two types of projectors (Braun / Kodak) categorized on their brightness (lumens) or two types of disease (Buragamo / Terrigitis) categorized on the fictitious PKS-value. Participants were told that the values they would experiences were created specifically for this study and that they could not use any prior knowledge to solve the categorization task. The two cover stories, and which category was A or B, was counterbalanced over participants.

On each of the 52 trials in the test phase participants were presented with a value (lumens or PKS) not seen in training and were asked to state the probability (in percent) that the item belonged to category A (i.e., the category with the highest base rate). To create the 52 items for the test phase the range of the training set was divided into eleven intervals based on the posterior probability $p_{A x}$ that a test item $x$ belonged to category $A\left(p_{A x}=0,0<p_{A x}<.1, .1 \leq\right.$ $p_{A x}<.2, \ldots, .9 \leq p_{A x}<1.0, p_{A x}=1.0$ ). For each of the nine middle intervals ( $0<p_{A X}<.1, \ldots 9 \leq p_{A x}<1.0$ ) four items were randomly drawn uniformly from that interval. Six items each were randomly drawn from the two extreme intervals, where the posterior probability is 0 and 1 . Finally, four critical items with an equal distance to the category means were included in the test set. After completing the test phase participants gave explicit estimates of the base rates and means of the two categories.

## Results

## Learning Performance

To investigate learning performance, the learning phase was divided into 10 blocks of 20 trials each. For each block, we calculated the proportion of correct categorizations. Figure 1 illustrates that participants quickly learn to categorize the training stimuli to the appropriate category with proportion correct reaching .8 at the end of the training phase.

We investigated the extent to which the base-rate-ratio and category-distance manipulations influenced the rate of learning by entering proportion correct as dependent variable into a $2 \times 2 \times 10$ split plot ANOVA with base-rateratio ( $8: 2$ / 6:4) and category-distance (short / far) as between-subjects independent variable and training block as within-subjects independent variable. The analysis revealed a significant main effect of training block $(F(9,324)=4.95$, $M S E=0.012, p<.001$ ) with a significant difference between the first and last block. Further, there was a significant main effect of category-distance $(F(1,36)=$
5.09, $M S E=0.068, p=.03$ ) where participants in the far condition performed better ( $M=.78, S E M=.018$ ) than participants in the short condition ( $M=.72$, $S E M=.018$ ). Notably this difference was significant also in the last training block $(t(38)=2.6, p=.01)$.


Figure 1: Proportion correct as a function of training block. Vertical bars denote 95 \% - confidence intervals.

Neither the main effect of base-rate-ratio $(F(1,36)=2.83$, $M S E=0.068, p=.10$ ) nor any of the interactions (all $p: s>$ .13) reached significance. Notice that while the main effect of category-condition indicates that it was easier for participants to learn the categories with means far apart as opposed to close together, the lack of interactions suggest an equal learning rate in all conditions.

## Subjective Probability Estimates

In the test phase participant gave explicit estimates of the posterior probability that an item $x$ belongs to category A (i.e., the category with the highest base rate). Figure 2 shows the mean estimated probability plotted against the normative Bayesian probability. In the figure, estimates are grouped into the eleven intervals described above.

As is evident from the figure participants are on average fairly well calibrated in their subjective probability estimates. To investigate the effect of base-rate-ratio and category-distance on the subjective estimates of posterior probability we calculated the mean absolute difference between the estimated and normative probability. The difference was entered as dependent variable into a 2 x 2 factorial ANOVA with base-rate-ratio (8:2 / 6:4) and category-distance (short / far) as between-subjects independent variables. There were no significant effects (all $p: s>.18$ ). Thus, probability estimates were on average not influenced by base-rate-ratio or category-distance.

To investigate a possible bias in the probability estimates the signed difference (rather than absolute difference) was entered into the corresponding ANOVA. Once again there were no significant effects (all $p:$ s > .26) and a single
sample t-test on the signed difference revealed that it did not differ significantly from $0(t(39)=.96, p=.35)$.


Figure 2: Subjective probability plotted against the normative Bayesian probability. Dotted line indicates perfect calibration.

The results illustrated in Figure 2 indicate that the accuracy of subjective probability estimates might vary as a function of the Bayesian posterior probability. To investigate this probability we conducted a more fine grained analysis where Bayesian probability interval was added as a within-subjects factor in the analysis of absolute error. This $2 \times 2 \times 11$ split-plot ANOVA revealed two significant effects. First, the main effect of Bayesian probability interval was significant $(F(10,360)=3.07$, MSE $=0.018, p<.001$ ). The effect is due to absolute errors for the larger probability intervals being smaller than those for the lower intervals. Second, the significant probability interval by base-rate-ratio $(F(10,360)=2.79, M S E=0.018$, $p<.001$ ) is illustrated in Figure 3 by means of a calibration curve. As can be seen in the figure, the interaction is due to estimates in the low probability intervals being slightly better for participants in the 6:4-condition than for participants in the 8:4-condition while it is the opposite in the high probability intervals.

The analysis above suggests that the base-rate-ratio manipulation might influence the extent to which participants use base rates to inform their subjective probability estimates. To investigate this possibility we analyzed participants’ probability estimates of the critical items included in the test set. Remember that the critical items are positioned with the same distance to both category means. If participants disregard the base rate information and instead use the ratio of the distance from a test item to each of the two means as a proxy for the posterior probability, or some similar strategy, they should estimate the posterior probability of all critical items to be .5.


Figure 3: Subjective probability plotted against the normative Bayesian probability for the two base-rate-ratio conditions separately. Dotted line indicates perfect calibration.

Figure 4 displays the distribution of responses to the critical items. As is evident from the figure a majority of responses are larger than .5 , indicating that participants take the base rate of the two categories into account when giving subjective probability estimates. To further investigate the use of base rates the subjective probability estimates of critical items were entered as dependent variable into a 2 x 2 factorial ANOVA with base-rate-ratio and category distance as between-subjects factors. One participant, considered an outlier ( $|z|>2.5$ ), was excluded from the analysis. The ANOVA revealed a significant main effect of base-rate-ratio $(F(1,35)=4.63, M S E=0.037, p=.038)$ with higher probability estimates in the $8: 2$-condition $(M=.76, S D=$ .14 ) then in the $6: 4$-condition ( $M=.62, S D=.24$ ). None of the other effects reached significance (both $p: s>.20$ ). More importantly in all conditions, participants gave estimates larger than .5 , even though not significantly larger in the short-6:4-condition, indicating sensitivity to base rates.


Figure 4: Distribution of subjective probability estimates of critical items in the test phase.

A further indication of sensitivity to base rates is given by the explicit estimates of base rates illustrated in Figure 5.


Figure 5: Means of explicit estimates of base rate for the four different true base rates separately. Vertical bars denote $95 \%$ - confidence intervals.

As can be seen in the figure the explicit estimates are sensitive to the experienced base rates. In addition there is little difference in the accuracy of estimates in the different conditions indicating that the differences in use of base rates seen above is not an effect of differences in learning.

## Discussion

Research on probabilistic reasoning has long been dominated by the general conclusion that people are very poor at integrating information according to the laws of probability (e.g., Bayes’ theorem). At the same time research concerned with category learning, indicates that people are quite apt at solving categorization tasks that, at least mathematically, are similar to probabilistic reasoning tasks. In the present study, we therefore combined these two research traditions by eliciting subjective posterior probabilities from base rates and likelihoods learned in a categorization tasks.

Performance in the learning phase indicated that our participants quickly learned to categorize the stimuli correctly. Performance was somewhat better when category means were far apart as opposed to close together. This was expected because the closer the two category means get the more two their probability density functions overlap, which in turn makes it more difficult to distinguish the two categories.

The subjective probability estimates given by participants in the test phase were, as is illustrated in Figure 2, well calibrated against the normative Bayesian probability. There was no systematic bias in the estimates and the pattern of results seen in Figure 2 suggests that the deviations from the normative Bayesian probability could be attributed to
regression effects. Notably, even though there was a difference in learning between the two category-distance conditions, this did not affect the correspondence of the subjective estimates.

The explicit estimates of base rates and category means indicated that participants learned these category properties. Arguably, however, they might not have used them to reach a subjective probability estimate. However, the analysis of the critical items included in the test phase showed that participants in all conditions were sensitive to the base rate and, at least to some extent, integrated this knowledge in their probability estimates.

Similar to previous research demonstrating that people can be good at reasoning under some conditions (e.g., Baron, 2000), the results of the present study show that when people are allowed to learn base rates and likelihoods in a category learning task they are at least under some conditions able to produce subjective probability estimates that are well calibrated and sensitive to base rates. This suggests that the conclusion by Kahneman and Tversky (1972, p. 450) may have been somewhat premature. An interesting question for future research is to investigate the processes leading up to what is apparently a normative answer.

## Acknowledgments

This research was sponsored by the Swedish Research Council.

## References

Anderson, J. R. (1991). The adaptive nature of human categorization. Psychological Review, 98, 409-429.
Ashby, F. G., \& Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. Journal of Mathematical Psychology, 39, 216-233.
Ashby, F. G., \& Maddox, W. T. (2005). Human category learning. Annual Review of Psychology, 56, 149-178.
Baron, J. (2000). Can we use human judgments to determine the discount rate? Risk Analysis, 20, 861-868.
Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, \& A. Tversky (Eds.), Judgment Under Uncertainty: Herustics and Biases (pp. 249-267). Cambridge, UK: Cambridge University Press.
Gigerenzer, G., \& Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. Psychological Review, 102, 684-704.
Gilovich, T., Griffin, D. W., \& Kahneman, D. (Eds.). (2002). Heuristics and Biases: The psychology of intuitive judgment. New York: Krieger Publishing Company.
Griffiths, T. L., Sanborn, A. N., Canini, K. R., \& Navarro, D. J. (2008). Categorization as nonparametric Bayesian density estimation. In N. Chater \& M. Oaksford (Eds.), The Probabilistic Mind: Prospects for Bayesian Cognitive

Science (pp. 303-328). Oxford, UK: Oxford University Press.
Juslin, P., Nilsson, H., \& Winman, A. (2009). Probability theory, not the very guide of life. Psychological Review, 116, 856-874.
Juslin, P., Nilsson, H., Winman, A., \& Lindskog, M. (2011). Reducing cognitive biases in probabilistic reasoning by the use of logarithm formats. Cognition, 120(2), 248-267.
Kahneman, D., \& Tversky, A. (1972). Subjective probability: A judgment of representativeness. Cognitive Psychology, 3, 430-454.
Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. Behavioral and Brain Sciences, 19, 1-53.
Medin, D. L., \& Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. Journal of Experimental Psychology: General, 117, 68.
Nilsson, H., Olsson, H., \& Juslin, P. (2005). The cognitive substrate of subjective probability. Journal of Experimental Psychology: Learning, Memory, and Cognition, 31, 600-620.
Oaksford, M., \& Chater, N. (2009). Précis of Bayesian rationality: The probabilistic approach to human reasoning. Behavioral and Brain Sciences, 32, 69-120.
Tenenbaum, J. B., Kemp, C., Griffiths, T. L., \& Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. Science, 331, 1279-1285.

# Extreme Expertise: Exploring Expert Behavior in Tetris 

John K. Lindstedt (lindsj3@rpi.edu) \& Wayne D. Gray (grayw@rpi.edu)<br>Cognitive Science Department<br>Rensselaer Polytechnic Institute


#### Abstract

Expertise is easy to identify in retrospect. It is the most expert player who wins the meet and the most proficient team that wins the playoffs. However, sometimes during play we see a masterful move that clearly separates one player from the competition. Our goal, in this work, is to identify the masterful moves or elements of expertise that predict the continuum of performance in the game of Tetris. As a first step we have collected data from a wide variety of Tetris Tournament players and used it to derive metrics of global, local, and immediate interactions. Here we present statistical models of these data and report the initial success of these models at predicting level of expertise.


Keywords: expertise, skill acquisition, exploratory analysis, videogames, regression modeling, thin-slicing

## Introduction

It seems easy to identify which baseball players are experts. We can look at their outputs: batting average, fouls, or total runs scored. The trouble is, we can only really make assessments on these outputs after the fact, once all the numbers are in, and the point is somewhat moot. But there must be something different about these experts at a more fundamental level, something identifiable in the way they are playing the game that forms the basis for their continued excellent performance.

What are the hallmarks of the exceptional player's expertise? Is it something about the way they grip the bat, or their stance? Is it in their ability to hit a certain kind of pitch over others? Are they slightly faster to respond, or more deliberate with their actions? Is it because they know when to bunt? Moreover, how much of the player's performance do we need to see in order to make an informed assessment of his or her expertise?

These questions lay the groundwork for asking the question: can we identify elements of expertise, behaviors made from instant to instant during performance which will allow us to rank a person on a scale ranging from novice to expert by observing just a thin slice of their behavior? We investigate this question using the video game Tetris.

## Background

The history of the scientific study of human expertise is nearly as long as the history of scientific psychology, with publications dating back to the discovery of the plateau in skill gain in telegraph operators in 1897 (Bryan and Harter), to an overthrowing of that notion in favor of continuous,
if subtle, skill gains throughout the acquisition of expertise (Keller, 1958), and ultimately to a reconciliation of the two findings as valid depending on the measurement device (e.g., Robertson \& Glines, 1985).

Our reading of the historical literature is that the discrepancy of major claims about the nature of expertise highlights the importance of metrics and of the available theoretical constructs. Although Bryan and Harter collected some data with millisecond accuracy, their general methodology lacked a few important controls and their main theoretical construct was stated in intuitive terms. Fifty years later, Keller (one of the foremost behaviorists of his day) had much higher standards for experimental design as well as a theoretical framework, behaviorism, that had no room for unobservable hierarchical structures. Just 30 years after Keller, Robertson and Glines had available to them the hierarchical theories of the information processing theorists as well as an understanding of the ways in which adopting different strategies could lead to differences in performance. As a consequence, unlike Keller when they looked, they found abundant evidence for individual differences in plateaus that seemed to reflect differences in strategies available or discoverable by students with different intellectual backgrounds (i.e., primarily engineers versus humanities students).

Our longterm goal is to provide a theoretical account of extreme expertise in dynamic tasks; that is, those which require an integration of real-time decision-making with a (figurative) tight loop among cognition, perception, and action. Examples of such skills include laproscopic surgery (Keehner et al., 2004), piloting jet aircraft and helicopters (Proctor, Bauer, \& Lucario, 2007; Hays, Jacobs, Prince, \& Salas, 1992), and detection of enemy submarines hiding in deep waters (Ehret, Gray, \& Kirschenbaum, 2000). Of course, we lack access to surgeons, helicopter pilots, and submarine commanders. However, we do have people who have spent thousands of hours acquiring extreme expertise in videogames. These people are the subject of our study and our first attempt at thinslicing the expertise in Tetris is the subject of this paper.

## Why Tetris?

Tetris is a videogame that is both easy to comprehend and difficult to achieve mastery over. The game is simple in that it has relatively simple rules (introduced in the next section) and players make decisions based on a limited set of potential actions (arranging and placing game pieces). However, there is much for a novice player to learn. The game space changes as a result of decisions made by the player. Errors accumulate
and one error tends to lead to another error until catastrophic failure (i.e. the end of the game) occurs. As the player succeeds, time pressure increases so that decisions have to be made within decreasing time windows. Furthermore, achieving the highest rewards requires performing maneuvers that risk error and reaching levels of the game where time pressure is highest.

To become highly proficient in the task, players must learn to effectively negotiate the error cost and the increasing time pressure by employing cognitive abilities such as: use of working memory, mental rotations, perceptual comparisons, strategic planning, and prediction, as well as the dexterous and rapid execution of chains of motor commands. Mastering Tetris requires the novice to coordinate the effective and efficient use all of these cognitive resources, abilities, and strategies. For these reasons, we see Tetris as an excellent platform for investigating the acquisition of expertise in a dynamic, real-time task.

In addition, Tetris has been used to document a variety of cognitive phenomena. A short list includes: epistemic versus pragmatic action (Destefano, Lindstedt, \& Gray, 2011; Kirsh \& Maglio, 1994), gains in cortical mass and BOLD response (Haier, Karama, Leyba, \& Jung, 2009), and near and far transfer (Sims \& Mayer, 2002).

## The Game of Tetris

(For readers already familiar with the game of Tetris, this section is optional review.)

Tetris is a game of increasingly fast-paced, generative puzzle-solving. When playing Tetris, a player manipulates a series of falling shapes, zoids, into an arrangement called the accumulation at the bottom of the game space. To score points, the player must clear rows. This is accomplished by filling at least one row in the accumulation. The immediate result is that points are scored and the row vanishes from the screen (thereby lowing the height of the accumulation). Since not all zoids fit perfectly together, the accumulation gradually rises as rows begin to go unfilled. When the accumulation reaches the top of the game space, the game ends. As the player clears lines, the game-level increases, speeding up the drop-rate of the zoid, and thus the difficulty, but also offering increased score payoffs for successfully cleared lines. Figure 1 illustrates the game screen as a player would see it.

Each zoid is one of seven unique shapes, all consisting of four contiguous block segments. Once a zoid is released into the game board, it begins automatically dropping, traversing the game space top to bottom in 12 seconds initially, down to 2 seconds at the highest difficulty level.

Scoring is nonlinear with respect to the number of lines cleared simultaneously. Initially, clearing 1 line awards 40 points, 2 lines awards 100 points, 3 lines awards 400 points, and clearing 4 lines simultaneously awards an extreme 1200 points. Clearing four lines simultaneously scores a Tetris, and
is notable because of both its high payoff and difficulty. Points awarded for $a$ Tetris are also modified multiplicatively by the current difficulty level.

Our version of Tetris, written in Flash, incorporates a robust logging system which captures all game events and states as they occur in real time. These events are detailed in the next section.


Figure 1. Example of the game environment.

## Events and Metrics

## Events in Tetris

Our basic unit of measurement is the episode, the time from when a zoid is released until it collides with and locks into the accumulation. It is in this time frame that all measurements of behavior and game state occur.

The player has available three kinds of actions: rotating clockwise and counterclockwise, moving a zoid to the left or right (i.e., translating between columns), and dropping the zoid (increasing the gravity intentionally). System events are any actions performed by the game environment, these include: automatically dropping the zoid due to gravity, clearing filled rows and awarding points, and releasing new zoids. Many of these actions occur within milliseconds of one another, a fact which is captured by time stamping in our continuous logging system.

Though the accumulation changes over time as zoids are placed and lines are cleared, during an episode the player interacts with one unique accumulation. Features of the accumulation are critical for understanding the player's current task status: its height determines how close the player is to failure, it may contain unreachable holes or pits which, for the game's continued success, must be uncovered (by clearing the rows which cover the pits) and filled, or overhangs (which can be thought of as a little cave that must somehow be filled by moving a zoid into it from its left or right side $a$ very difficult maneuver, especially for novices).

## Measure of Expertise

To assess the behavioral differences of expertise, we must define it quantitatively. Due to the difficulty of achieving high scores in Tetris, and the unlikelihood that a player will score highly simply "by accident," we consider a player's long-term ability to achieve high scores a basic measure of their expertise; that is, a player's expertise is equated to the maximum score the player was able to achieve during any of their games played during data collection. Because scores tend to increase nonlinearly (later levels award disproportionately more points) and seem to follow a somewhat exponential pattern, our metric of a player's expertise is the base-10 logarithm of their maximum game score.

## Predictive Measures

Because the task environment in Tetris is sufficiently simple, we are able to extract many details of task performance which may reflect differences in novice and expert behavior. It is important to point out that we are not searching only for those metrics which are the root cause of more expert performance, but also any metrics which reliably co-occur with expert ability. This investigation remains agnostic to this distinction between components and markers of expertise.

Our various metrics can be categorized at three successive time scales of human action (Newell, 1990, p. 122): global $\left(10^{2}\right)$, local $\left(10^{1}\right)$, or immediate $\left(10^{0}\right)$.
Global metrics. These assess the player's overall game status as reflected in the built accumulation. These metrics are associated most closely with survivability in the game, such as the overall height of the accumulation, or the number of unworkable holes, or pits, which the player has accrued during play. These metrics, averaged across sections of gameplay, indicate broad patterns of performance which may differentiate between novices and experts, particularly in terms of long-term strategies.
Average height: The average of all column heights in the accumulation.
Pits: The total number of unworkable pits (covered empty spaces) present in the accumulation.
Overhangs: The number of covered spaces into which a player may still dextrously maneuver a zoid.
Roughness: A measure of the "randomness" of the accumulation.
Levelness: Measures the relative flatness of the top of the board.
Spire: The difference between the highest point in the accumulation and the average height.
Tetris progress: The number of nearly-filled rows presently lined up in the accumulation, ready to produce highscoring line-clears.
Zoid-positions: The amount of "good" positions available for any kind of zoid. This is a rough measure of the
functional "goodness" of the accumulation the player has built.

Local metrics. These assess the kinds of zoid-placements the player selects in relation to possible positions on the accumulation. This includes features such as the number of perimeter segments matched during a placement (i.e., does that zoid fit flush with its surroundings, or does it stick out precariously?), or whether the placement creates pits or overhanging segments in the accumulation which complicate later gameplay decisions. Zoid placements are also compared across all potential placement locations and orientations for the current zoid, giving a ratio of assumed "goodness" for a placement. These local metrics account for the kinds of decisions made at each step of the game.

Matched edges: The number of segments of the placed zoid which are touching the surrounding accumulation.
Match ratio: Ratio of the number of matched edges to the maximum possible for all positions the zoid could have been placed this episode.
New pits: The number of new pits created by this move.
Uncovered pits: The number of pits uncovered by this move.
Filled overhangs: The number of overhang cavities filled by the current move.
Current zoid-positions: The number of "good" positions available for the current zoid, which may indicate a player's planning for the next zoid in the previous episode.

Immediate interaction metrics. These account for how a zoid placement is executed, what can be thought of as the sensory-motor aspects of the gameplay. These include measurements of reaction times for various actions, such as the first keypress in an episode, and the first commission of a zoid drop to indicate that a decision has been made. These measures account for the rapid interactive skills a player employs to perform the basic decisions in the local metrics.

Total translations: The number of times a zoid was moved left or right in the episode.
Total rotations: The number of rotation actions performed on the zoid this episode.
Grouped actions: The number of clusters of similar actions performed in sequence (i.e., 3-translations, 2-rotations, 16 -drops). This measure reduces the sequences of actions to more conceptually coherent segments, with lower numbers implying less scattered executions.
Drop ratio: The proportion of the zoid's downward movements (in this episode) that can be attributed to the player's intentional dropping versus the system's automatic dropping.
Initial latency: The time (in milliseconds) between the start of the episode and the first action taken by the player.

Average latency: The average time between actions taken by the player.
Drop latency: The time from the start of the episode until the player decides to drop the zoid.

Each of these metrics is tallied and recorded once per episode. By examining elements from these three categories of performance, we hope to capture a broad, detailed picture of each player's gameplay as it occurs in real time.

## Methods

## Data collection

To acquire data from a cross section of players with different levels of expertise, we sponsored a Tetris tournament at Rensselaer Polytechnic Institute's Genericon - a convention for gaming, comics, Japanese anime, and all things "nerd culture." Participants in the tournament were volunteers from the pool of all those attending the convention, comprised primarily of RPI undergraduates.

Before the tournament, participants played two rounds of Tetris to determine their eligibility for competing. Once entered, participants competed in pairwise elimination matches wherein the highest score wins. The top three players of each tournament were offered a cash prize, provided they came to the laboratory and played an additional hour of Tetris.

We collected data using this procedure at two successive Genericons in 2006 and 2007. At the end of data collection, we had data from 57 unique players, with game scores spanning six orders of magnitude (less than 100 points to over $1,000,000$ ).

## Data filtering

Games wherein a player did not clear any lines were omitted from analysis, as these represent sessions which were either aborted or wherein the player clearly did not understand the game rules. Additionally, we sometimes observed players self-aborting games by rapidly dropping zoids until a gameover was achieved. These episodes were omitted from analysis, as they reflected gameplay behavior with maligned goals.

## Observation window

An important consideration for our data set is that it is naturalistic: no experimental controls were put in place, and no manipulations were made to the basic game. As such, there is a great deal of unevenness in the data set. The task environment is influenced greatly by the randomness of the zoid selection and player strategy, as is the number of episodes it takes a player to advance to the next difficulty level (where game speed is increased), or even the number of episodes played before the game ends. To control these elements would be to interfere with the basic structure of the game and deviate from the way players would naturally approach the
game, hindering our ability to find natural expert players in the wild as such. Thus, we leave these vital game elements uncontrolled, and instead institute a moving window through which to examine the gameplay data.

A key element of this exploration is whether we can thinslice by predicting expertise from a relatively small amount of data. Across all subjects and games, the mean number of episodes per game was 264.74 [Min. $=41$, Max. $=1388$, S.D. $=210.97]$. For our thin-slicing, in all cases the observation window begins with the $1^{\text {st }}$ episode of each game, wherein all players have a completely empty accumulation with which to work. For each player, we then averaged the data for all games for episodes 1-2 (an extremely thin slice of behavior), $1-10,1-100$, and all (using all available data for the analysis). Averaging behavioral measures across this window results in aggregate measures of performance which are representative of a player's behavior for the chosen observation window. Our question is whether measures made on these different slices of performance are predictive of overall performance.

## Results

## Multiple linear regression models

Prior to modeling, the dataset was sampled using a simple random assignment, using $80 \%$ of the data for training and leaving $20 \%$ for testing model predictions. The samples were verified as having similar distributions for the dependent measure of expertise [Training set: Mean = 4.43, S.D. $=0.61$; Test set: Mean $=4.51$, S.D. $=0.73$ ].

For each of the four selected observation window sizes (2, 10,100 , and all episodes), we conducted a multiple regression on each training data set using all predictors detailed in the Predictive Measures section. To account for any suppressor effects, a backward step-wise selection process was used in the regression. Table 1 shows the results of these models, and Table 2 illuminates the significance of each model's predictors. Note that the number of predictors ultimately used in each model varies due to the stepwise selection process. Figure 2 shows the fit of each model to the training data.

## Prediction

To assess each model's ability to predict unseen data, we performed predictions on the test data set ( 20 percent of observations). The Predictions section of Table 1 shows the relative success of each model as determined by the fit of a Pearson's product-moment correlation. Figure 3 shows the fit of the test set data to the model predictions.

## Discussion

From these results, we see significant fits for models created using all sizes of observation windows, from data spanning just two episodes to the use of the entire data set. The two models sampling from just 2 and 10 episodes each are

Table 1
Results of linear regression model for all window sizes.

|  | Observation window size |  |  |  |
| :--- | ---: | ---: | ---: | ---: |
|  | eps | 10 eps | 100 eps | all eps |
| Multiple $\mathrm{R}^{2}$ | .4607 | .3913 | .5882 | .8185 |
| Adjusted $\mathrm{R}^{2}$ | .3686 | .2509 | .5058 | .7767 |
| DF | $(7,41)$ | $(9,39)$ | $(8,40)$ | $(9,39)$ |
| F-value | 5.003 | 2.786 | 7.141 | 19.55 |
| p-value | $<0.001$ | 0.01 | $<0.0001$ | $<0.0001$ |
|  |  |  |  |  |
|  |  |  |  |  |
| Prediction |  |  |  |  |
| Correlation | 0.344 | -0.235 | 0.697 | 0.757 |
| p-value | 0.27 | 0.46 | $<0.02$ | $<0.01$ |

Table 2
List of significant predictors across models of differing observation window sizes. Significance codes are: '*' $-p<0.05$; '**' $p<0.01$; '***' $p<0.001$; '.' = present but not significant.

|  | Window Size (episodes) |  |  |  |
| :---: | :---: | :---: | :---: | :---: |
|  | 2 | 10 | 100 | All |
| Intercept |  |  | ** | . |
| Global metrics: <br> Average Height <br> Pits <br> Overhangs <br> Roughness <br> Levelness <br> Spire <br> Tetris progress <br> Zoid-positions | $\begin{gathered} * \\ * \\ * \\ * \end{gathered}$ | * |  | * |
| Local metrics: <br> Matched edges <br> Match ratio <br> New pits <br> Uncovered pits <br> Filled overhangs <br> Current zoid-positions | *** <br> ** |  | ** | ** <br> *** |
| Immediate metrics: <br> Total translations <br> Total rotations <br> Grouped actions <br> Drop ratio <br> Initial latency <br> Average latency <br> Drop latency | * | $*$ $*$ | * | $\begin{gathered} * \\ * \\ * * * \\ \cdot \\ * * * \end{gathered}$ |

notable for their good fits, but both ultimately fail to predict unseen data. Nonetheless, their fits are encouraging in that they achieve a measure of success even when based on such a small proportion of the player's observable performance data.


Figure 2. Fit of multiple regression model to training data. Different plots for models sampling from A) 2 episodes, B) 10 episodes, C) 100 episodes, and D) all observed episodes per game.

Models sampling from more data are naturally able to account for more of the variance in the data, as seen by the increasing adjusted $\mathrm{R}^{2}$ values for those models with larger windows, with the model sampling all data presumably demonstrating a maximum of success. Interestingly, we see that the model sampling only the first 100 episodes (less than a quarter of all observed data), maintains a fit to the training data and ability to predict the test data comparable to that of the model sampling all data. This, too, is encouraging in our pursuit of using small proportions of data to predict long-term performance.

It is tempting to draw conclusions from the lists of significant predictors presented in Table 2, but there is, regrettably, a non-trivial sampling effect; depending on how the data set is partitioned into training and test sets, these significant variables tend to shift, vanish, and reappear on subsequent samplings. This is likely due to two underlying effects: a strong effect of individual differences, as suggested by Robertson and Glines (1985); and a high level of correlation between these variables, because many of them necessarily depend on one another (e.g., average height being necessary for Tetris progress). We cannot yet account for these covert effects and are not prepared to draw strong conclusions about the individual predictors' viability in predicting longterm Tetris performance. We can, however, offer two points of speculative commentary based on observation of these effects: first, some predictors seem to emerge as significant more frequently than others, and second, predictors representing all three categories (global, local, and immediate) tend to emerge as significant across samplings, indicating that there


Figure 3. Fit of predictions from models to the test data set. Different plots for models sampling from (A) 2 episodes, (B) 10 episodes, (C) 100 episodes, and (D) all observed episodes per game.
may exist latent factors within each of these categories which contribute independently to skilled performance.

## Conclusions

Our goal is to identify the elements of expertise that predict the continuum of performance in the game of Tetris. As a first step, we collected data from a wide variety of Tetris Tournament players and used it to derive metrics of global, local, and immediate interactions. Here we reported our first statistical models of these data and our initial success at predicting level of expertise from thin-slices of behavior.

Although our results are tentative, we are pleased with our initial success in applying a general cognitive task approach to extreme expertise. Our categories of global, local, and immediate interaction are based on three successive levels of the time scale of human action (Newell, 1990). At least some of our initial items for each scale shows some success as a predictor of expertise. Thin-slicing seems to produce valid predictions as, to our surprise, even the regression model based on the first two episodes of each game had some predictive validity. We are embolden by these initial successes and have made plans to collect an order of magnitude more data from an order of magnitude more players at all levels of expertise.

Our predictive modeling has thus far been limited to the statistical technique of multiple regression. Other techniques have been suggested and we are openly soliciting suggestions from the cognitive community. Further work will also seek to address the individual differences across players at the same skill level and will attempt to extract a more refined set of metrics of behavior with fewer co-dependencies.

## Acknowledgements

We thank Michael Kalsher for his advice on some of the statistical models presented in this paper. The work was supported, in part, by grants N000140910402 and N000141310252 to Wayne Gray from the Office of Naval Research, Dr. Ray Perez, Project Officer.

## References

Bryan, W. L. \& Harter, N. (1897). Studies in the physiology and psychology of the telegraphic language. Psychological Review, 4(1), 27-53.
Destefano, M., Lindstedt, J. K., \& Gray, W. D. (2011). Use of complementary actions decreases with expertise. In L. Carlson, C. Hölscher, \& T. Shipley (Eds.), Proceedings of the 33rd annual conference of the cognitive science society (pp. 2709-2014). Austin, TX: Cognitive Science Society.
Ehret, B. D., Gray, W. D., \& Kirschenbaum, S. S. (2000). Contending with complexity: developing and using a scaled world in applied cognitive research. Human Factors, 42(1), 8-23.
Haier, R., Karama, S., Leyba, L., \& Jung, R. (2009). Mri assessment of cortical thickness and functional activity changes in adolescent girls following three months of practice on a visual-spatial task. BMC Research Notes, 2, 1-7. Retrieved from http://dx. doi.org/10.1186/1756-0500-2-174
Hays, R. T., Jacobs, J. W., Prince, C., \& Salas, E. (1992). Flight simulator training effectiveness: a meta-analysis. Military Psychology, 4(2), 63-74. Retrieved from http: / / www . tandfonline . com / doi / abs / 10 . 1207 / s15327876mp0402_1
Keehner, M., Tendick, F., Meng, M., Anwar, H., Hegarty, M., Stoller, M., \& Duh, Q. (2004). Spatial ability, experience, and skill in laparoscopic surgery. American Journal of Surgery, 188(1), 71-75.
Keller, F. S. (1958). The phantom plateau. Journal of the Experimental Analysis of Behavior, 1(1), 1-13.
Kirsh, D. \& Maglio, P. (1994). On distinguishing epistemic from pragmatic action. Cognitive Science, 18, 513-549.
Newell, A. (1990). Unified theories of cognition. Cambridge, MA: Harvard University Press.
Proctor, M. D., Bauer, M., \& Lucario, T. (2007). Helicopter flight training through serious aviation gaming. The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology, 4(3), 277-294. Retrieved from http://dms.sagepub.com/content/4/3/277.abstract
Robertson, R. J. \& Glines, L. A. (1985). The phantom plateau returns. Perceptual and Motor Skills, 61(1), 55-64.
Sims, V. K. \& Mayer, R. E. (2002). Domain specificity of spatial expertise: the case of video game players. Applied Cognitive Psychology, 16, 97-115.

# Working memory capacity and fluid abilities: The more difficult the item, the more more is better 

Daniel R. Little (daniel.little@unimelb.edu.au)

School of Psychological Sciences, The University of Melbourne Parkville, VIC 3010 Australia

Stephan Lewandowsky (stephan.lewandowsky@uwa.edu.au)<br>School of Experimental Psychology, University of Bristol Bristol, UK BS81TU<br>School of Psychology, The University of Western Australia Crawley, WA 6009 Australia

Stewart Craig (craig03@student.uwa.edu.au)
School of Psychology, The University of Western Australia Crawley, WA 6009 Australia


#### Abstract

Recent evidence has suggested that the relationship between a test of fluid intelligence, Raven's Progressive Matrices, and working memory capacity (WMC) may be invariant across difficulty levels of the Raven's items. We show that this invariance can only be observed if the overall correlation between Raven's and WMC is low. We demonstrate that by using a composite measure of WMC, which yields a higher correlation between WMC and Raven's than reported in previous studies, that there was a significant positive relationship between Raven's item difficulty and the extent of the itemwise correlation with WMC. This result puts strong constraints on theories of reasoning and challenges some existing views. Keywords: Raven's Progressive Matrices; Working Memory Capacity.


## Introduction

Working memory (WM), the architecture responsible for the retention and manipulation of information over short periods of time, is a core component of human cognition. People's working-memory capacity (WMC) shares around $50 \%$ of the variance with general fluid intelligence (Kane, Hambrick, \& Conway, 2005) and is predictive of performance in a number of reasoning tasks and other measures of higher cognitive ability. However, there is some dispute about the exact nature of the relationship between WMC and one important assay of fluid intelligence, Raven's Progressive Matrices (e.g., Raven, Raven, \& Court, 1998).

Raven's test is designed such that items differ considerably in difficulty, with easy items-presented early in the test-solvable by more than $90 \%$ of participants and the hardest items-presented last-being solvable by fewer than $10 \%$ of participants. Carpenter, Just, and Shell (1990) presented a taxonomy of rule types that were used to create each of the Raven's items. To illustrate, Figure 1 presents two sample Raven's-like problems created using different rules. The matrix in panel A contains an incremental rule (i.e., the dots increase across items) and a distribution of 3 , permutation


Figure 1: Two examples of matrices like those in the Raven's test. A: Example of an item containing a pairwise incremental rule and a distribution of 3 permutation rule. B: Example of an item containing a constant rule and a distribution of $2(\mathrm{XOR})$ rule.
rule (i.e,. objects with 1,2 and 3 triangles are permuted across rows and columns). The matrix in panel B contains a constant rule (i.e., the center dot appears in all items) and a distribution of 2 (or logical XOR) rule (i.e., features which appear in the first two objects do not appear in the third object and features which appear only in one of the first two objects also appear in the third object). Carpenter et al.'s rule taxonomy also included feature decrements between objects, logical disjunction rules (OR) and logical conjunction rules (AND). Participants must infer these rules from the objects in the matrix and then predict and select the missing lower right object in the matrix from the set of possible response options.

Carpenter et al. (1990) compared two production system models that demonstrated the importance of the number and type of rules and WMC. Both of the models operated by finding correspondences between the


Figure 2: Left Panel: Observed accuracy from Unsworth \& Engle (2005). Right Panel: Observed itemwise point bi-serial correlations in Unsworth \& Engle (2005).
symbolically-coded features of the items, transferring these correspondences to a working memory buffer where any rule satisfied by the extracted correspondences was invoked, using the instantiated rules to generate the missing item, and finally, searching through the response options to find the best match. One model (called FAIRAVEN) had no strategic memory organization and did not have access to distribution of 2 (XOR) rules; the other model (called BETTERAVEN) was endowed with better control processes and contained access to all of the rules types. The assumptions of the models were consonant with observed accuracy, response time, and eye fixation data and the models were able to explain the performance of median Raven's performers and the very best Raven's performers, respectively.

If we assume that increased WMC allows for an improved ability to maintain goals and retain intermediate results and rules necessary to successfully solve the most difficult Raven's items, the implication of the modeling is that performance on more difficult items should be more highly correlated with WMC. In subsequent tests of that idea, Unsworth and Engle (2005) and Wiley, Jarosz, Cushen, and Colflesh (2011) examined the correlation between WMC and Raven's performance across ordinal item position, which is a proxy for item difficulty. Contrary to expectation, those studies found that the role of WMC remained invariant across item position. The left and right panels of Figure 2 show the accuracy and itemwise correlations observed by Unsworth and Engle (2005). Although the itemwise pattern is quite noisy, there appears to be no systematic relationship between ordinal item difficulty (on the abscissa) and the correlation between performance on those items and WMC (as measured by OSPAN). This impression of an invariant relationship was statistically supported by the failure to find an increasing correlation between OSPAN and the proportion correct within each quartile of the Raven's test.

Those reports of invariant itemwise correlations have been used to reject the model of Carpenter et al. (1990), or indeed any other proposal that cites the ability to hold
rules and goals in working memory as underlying Raven's performance. The failure to find a selective involvement of WMC has motivated alternative theorizing about the relationship between the Raven's test and WMC. For example, Unsworth and Engle (2005) concluded that the variance shared by WMC and Raven's reflected attentional control mechanisms, presumed to be implicated in both tasks, which were thought to be uniformly important across all of the Raven's items. Thus, irrespective of item difficulty, a person with larger WMC benefits from an enhanced ability to selectively focus on those features of an item that are relevant to the item-appropriate rule and to filter out distracting non-relevant goals and features. Although this account has not been quantitatively formalized, there is empirical support from other domains that working memory underwrites an ability to filter out distracting information (Conway, Cowan, \& Bunting, 2001).

The current state of affairs thus presents a conceptual puzzle: On the one hand, intuition and at least one theory (Carpenter et al., 1990) suggest that the importance of WMC should be accentuated for the more difficult Raven's items, for the simple reason that the easiest items are - by design - solvable by most participants and hence ought not to correlate much with WMC. On the other hand, there are now several reports that the role of WMC is invariant across item difficulty (Unsworth \& Engle, 2005; Wiley et al., 2011). Those results appear consonant with an attentional view of working memory but run counter to the model of Carpenter et al. (1990). However, there are several reasons to examine those reports further: First, the counter-intuitive nature of those results deserves to be underscored - after all, how can the correlation between WMC and performance be identical for items that are solved by $90 \%$ and $10 \%$, respectively, of participants?

There are other reasons to expect that the acceptance of an invariant relationship between Raven's performance and WMC may have been premature. By definition, those results rely on a failure to reject the null hypothesis, and the "noisiness" of the data is considerable (see Figure 2, right panel). Moreover, studies showing an invariant itemwise correlation were marred by the fact that only a single task (OSPAN) was used to measure WMC-consequently, measurement error or "method variance" from that single task might have masked a relationship between WMC and the more difficult Raven's items in the studies of Unsworth and Engle (2005) and of Wiley et al. (2011). In support of this claim, the correlations reported in those papers $(r=.335$ and $r=.33$, respectively) fall on the lower end of the range of correlations between WMC and Raven's identified in a recent meta-analysis (i.e., . 312 to .641; Ackerman, Beier, \& Boyle, 2005). Further, Unsworth and Engle (2005), participants were allocated 30 minutes to complete the

Raven's test rather than the standard 40 minutes, which likely resulted in decreased overall accuracy, that may have further obscured an increasing effect of WMC.

We suggest that there are strong and well-supported reasons to expect the involvement of WMC in performance to increase across item difficulty in the Raven's test. Reports to the contrary have relied on acceptance of the null hypothesis and have involved limited measures of WMC. The issue of how working memory relates to Raven's performance may therefore be worthy of further exploration. We revisit this issue and resolve it by presenting a behavioral study using a composite measure of WMC that correlates more strongly with Raven's and results in an increasing itemwise correlation-as predicted by Carpenter et al. (1990) and contrary to the null results reported to date.

## Behavioral Study

In this study, we sought to maximize the likelihood of finding an increasing itemwise corrrelation function by using multiple tasks and deriving a composite latent measure of WMC, thus reducing the task-specific variance and measurement error that beset a single-task measure such as OSPAN. We therefore expected the correlation between WMC and RAPM performance to be greater than in relevant previous studies. Why should we expect the overall correlation between WMC and Raven's to affect the itemwise correlation? The answer lies in the constraints imposed by the decreasing accuracy function across Raven's items: Because nearly everyone gets the early items correct, the corresponding point-biserial correlations for those items must be near zero. It follows that the overall correlation between WMC and Raven's can only express itself in the pointbiserial correlations for the later items where performance is more variable across individuals. Consequently, a greater overall correlation is preferentially observed in the later items, which necessarily translates into an increasing itemwise slope across the entire test.

This increasing slope fails to be observed only if performance on the final test items falls sufficiently close to the floor to constrain their variance, thereby curtailing the itemwise correlations for the last items. The shorter test duration used by Unsworth and Engle (2005) led to near-floor performance on the later test items, thereby preventing the detection of an increasing itemwise slope. This is likely to have been the case even if the overall correlation had been greater. For the increasing slope to be observed, performance on the later items ought to be off the floor and the overall correlation must be large. The standard 40 minute allocation in our study should act to increase accuracy for the later items, and the use of a battery of WM tests should serve to increase the overall correlation between WMC and Raven's performance.

## Method

Participants The participants were 130 volunteers (95 female; mean age 21.12) from the University of Western Australia campus community. Participants received either partial course credit for an undergraduate psychology course or $\$ 20$ for two 1-hour sessions.

Procedure In the first session of the study, participants completed a battery of four WMC tasks (see Lewandowsky, Oberauer, Yang, \& Ecker, 2010).

Memory updating (MU). The MU task required participants to (a) store a series of numbers in memory, (b) mentally update these numbers based on a series of arithmetic operations, and (c) recall the updated numbers. On each trial, three to five frames containing random digits were presented on the screen. Following memorization, successive arithmetic operations, (e.g., ' +4 ' or '-3') were presented in the frames, one at a time for a random number of steps before final recall was cued. The key dependent variable is the proportion of updated digits recalled correctly.

Operation span (OSPAN) and Sentence span (SS). On each OSPAN trial, a series of arithmetic equations were presented (e.g., $4+3=7$ ), each of which was followed by a consonant for memorization. Participants judged the equation for correctness and recalled the consonants immediately after list presentation in the original order. The SS task was identical to the OSPAN, except that instead of judging correctness of an equation, participants judged the meaningfulness of sentences (cf. Daneman \& Carpenter, 1980). For OSPAN and SS, the key dependent variable is the proportion of consonants recalled correctly.

Spatial short-term memory (SSTM). The SSTM task was adapted from Oberauer (1993) and involved memorization of the spatial location of circles presented, one-by-one, in various locations in a $10 \times 10$ grid. Participants used the mouse to indicate the memorized location of the dots in any order by clicking in the corresponding grid cells. For this task, participants are given a score based on how similar their recalled pattern was to the to-be-memorized pattern (see Lewandowsky et al., 2010).

Fluid intelligence tests (RAPM) In the second session, participants completed Sets I and II of the 1962 Raven's Advanced Progressive Matrices. As recommended by Raven et al. (1998), RAPM Set I was included to familiarize participants with the matrices. Participants had 5 minutes to complete the 12 items in Set I before being given the standard 40 minutes to complete the 36 items in Set II. We only report the results for the last 36 items (Set II).

## Results

Data from two participants who failed to complete all tasks were removed from analysis, and data from two

Table 1: Means $M$, standard deviations $S D$, skewness, and kurtosis for the operation span task (OSPAN), sentence span task (SS), spatial short-term memory task (SSTM), memory updating task (MU), and Raven's Advanced Progressive Matrices (RAPM).

| Measure | $M$ | $S D$ | Skewness | Kurtosis |
| :--- | :---: | :---: | :---: | :---: |
| OSPAN | 0.71 | 0.14 | -0.99 | 4.07 |
| SS | 0.70 | 0.15 | -0.70 | 3.30 |
| SSTM | 0.84 | 0.06 | -0.14 | 2.37 |
| MU | 0.66 | 0.18 | -0.34 | 2.48 |
| RAPM | 24.47 | 5.37 | -0.34 | 2.90 |

further participants were discarded for having WMC and Raven's scores less than three standards deviations below the mean, respectively. The final analyses thus used a sample size of $N=126$. Descriptive statistics for the four WMC tasks and RAPM are shown in Table 1. The top left panel of Figure 3 shows average performance on the RAPM items from Set II. The pattern conformed to expectation in that accuracy decreased with ordinal item position.
WMC and item difficulty For the correlational analyses, we computed a composite measure of WMC by first converting each participant's score on each WM task into a $z$-score, and then computing that person's average $z$-score across the four tasks ( $z \mathrm{WMC}$ ). As anticipated, the overall correlation between $z \mathrm{WMC}$ and the total RAPM score was moderately large, $r=.56, p<.001$, and larger than was found in previous studies using only a single measure of WMC.

The top right panel of Figure 3 shows the pointbiserial correlations between WMC and performance broken down across Raven's items, together with the best-fitting regression line. In contrast to Unsworth and Engle (2005), accuracy was high enough to permit inclusion of all of the Raven's items. The slope of the regression line (.004) was significantly greater than zero, $t(34)=2.87, p<.01, r^{2}=.20 .{ }^{1}$ The data confirm that when there is at least a moderate correlation between WMC and Raven's performance, the itemwise correlations increase with item difficulty.

Further, to analyse the relationship between $z \mathrm{WMC}$, item difficulty, and performance on Raven's, we conducted a multilevel logistic regression, which circumvents problems due to items with very high or very low accuracy by relying on the logistic (or inverse-logit) function to model the accuracy proportions for each item. We examined a model which includes WMC, ordinal item position (as a proxy for difficulty), and the inter-

[^139]

Figure 3: Left: Performance on Raven's Advanced Progressive Matrices items. Right: Observed correlations between working memory capacity ( $z \mathrm{WMC}$; based on a battery of four tasks) and performance on each Raven's item. The solid line represents the best-fitting regression line (intercept .13 , slope .004). Bottom panels: Results from a bootstrapping analysis resulting in correlations of $.52, .23$, and .09 , respectively, between WMC and overall Raven's performance. All bootstrap results are based on 10,000 replications and the shaded areas represent the $95 \%$ confidence regions for the bootstrapped means. The framed bottom-left panel matches the overall correlation and itemwise results in the top right panel.
action between these variables. ${ }^{2}$ We also systematically tested alternative random-effect models (i.e., letting one or more of intercept or ordinal item position vary randomly across participants. ${ }^{3}$ ) and determined the preferred model using BIC.

The logistic regression assumes that the predictors are linearly related to the logit transformation of the dependent variable; consequently, we examined the relationship between each variable and accuracy using a White test for nonlinearity (Lee, White, \& Granger, 1993). Ordinal item position showed a demonstrable nonlinear relationship with accuracy $\chi^{2}(2)=61.12, p<.001$. A Box-Tidwell analysis indicated that the nonlinearity could be removed by raising ordinal item position to a power of $1.704, \chi^{2}(2)=4.29, p=.12$ (see Box \& Tidwell, 1962). Because item position is only a proxy for difficulty, the transformation of that variable is acceptable because it retains the ordinal association with the unknown scale of actual difficulty. None of the other variables showed any nonlinear relationship with the largest

[^140]Table 2: Estimated parameters (and standard errors) of mixed effects modeling of the RAPM behavioral study. All significant coefficients are bolded.

| Parameters | Model 1 | Model 2 |
| :--- | ---: | ---: |
| Fixed |  |  |
| Intercept $\left(\beta_{0}\right)$ | $\mathbf{2 . 9 2 ( \mathbf { 0 . 1 1 ) }}$ | $\mathbf{2 . 9 8 ( \mathbf { 0 . 1 1 ) }}$ |
| zWMC $\left(\beta_{z}\right)$ | $\mathbf{0 . 5 3 ( \mathbf { 0 . 1 4 ) }}$ | $\mathbf{0 . 5 2 ( \mathbf { 0 . 1 4 ) }}$ |
| Item $\left(\beta_{\psi}\right)$ | $\mathbf{- 0 . 0 1}(\mathbf{0 . 0 0 0 3 )}$ | $\mathbf{- 0 . 0 1} \mathbf{( 0 . 0 0 0 4 )}$ |
| zWMC $\times$ Item | $\mathbf{0 . 0 0 1}(\mathbf{0 . 0 0 0 5 )}$ | $\mathbf{0 . 0 0 1} \mathbf{( 0 . 0 0 0 5 )}$ |
| Random |  |  |
| Intercept $s_{0}$ | $0.67(0.59)$ | $-0.05(0.06)$ |
| Item $s_{\psi}$ |  | $0.0001(0.0001)$ |
| Evaluation |  |  |
| df | 5 | 7 |
| BIC | 4089 | 4097 |

$\chi^{2}$ being for $z \mathrm{WMC}\left(\chi^{2}(2)=2.86, p=.24\right)$.
Exponentiating ordinal item position to correct for nonlinearity, our first model is given by the following equation:

$$
\begin{equation*}
y_{i j}=\beta_{0}+\beta_{z} z_{i}+\beta_{\psi} \psi_{j}^{\lambda}+\beta_{(z \times \psi)} z_{i} \times \psi_{j}^{\lambda}+\left(S_{i}+e_{i j}\right) \tag{1}
\end{equation*}
$$

where $y_{i j}$ is a binary response variable indicating whether participant $i$ made a correct (1) or incorrect (0) response on item $j, z_{i}$ is the $z \mathrm{WMC}$ score for participant $i, \psi_{j}$ is the ordinal item position of item $j, \lambda$ equals 1.704 (as indicated by the above Box-Tidwell analysis), $S_{i}$ is the set of subject random effects and $e_{i j}$ is an error term.

We tested this model using only the intercept as a random effect (e.g., Model 1, see Table 2) or the intercept plus $\psi^{\lambda}$ as random (Model 2). Comparison of the BICs pointed to the model in which only the intercept varied randomly as being preferable (i.e., Model 1). This model revealed significant effects of $z \mathrm{WMC}(p<.001)$, ordinal item position $\left(\psi^{\lambda}, p<.001\right)$, and the critical $z \mathrm{WMC} \times$ ordinal item position interaction $(p<.01)$. The latter interaction confirms that WMC played an increasingly important role as item difficulty increased, precisely paralleling our initial correlation-slope analysis.

Bootstrapping analysis We next conducted a bootstrapping analysis in which we simulate the effect of decreasing the overall correlation. In other words, to confirm that the magnitude of the overall correlation was responsible for the emergence of an item-difficulty effect in our study, we conducted bootstrapping analyses based on the observed subject $\times$ item $(126 \times 36)$ response matrix, with rows ordered according to the observed $z \mathrm{WMC}$. The overall correlation between $z \mathrm{WMC}$ and Raven's was manipulated by generating new $z \mathrm{WMC}$ scores for each participant and examining the effect of that manipulation on the itemwise correlations.

We created three conditions, each involving 10,000 bootstrapping runs. For each run, $n$, a new vector of $z \mathrm{WMC}$ scores was randomly derived from the observed values according to: $z \mathrm{WMC}^{(n)}=\nu \times z \mathrm{WMC}+\epsilon$ where $\epsilon \sim N\left(0, \sqrt{\left(1-\nu^{2}\right)}\right)$ and $\nu$ varied across conditions. This new vector contained $z \mathrm{WMC}$ scores which were derived from the observed $z \mathrm{WMC}$ scores but had a reduced correlation with the observed Raven's scores. The rows of the observed binary response matrix were then resorted according to the new vector $z \mathrm{WMC}^{(n)}$ yielding another bootstrapped replication with a specified correlation between $z \mathrm{WMC}$ and RAPM that maintained the overall itemwise error rate and overall Raven's correct for each participant observed in the study. Item-wise correlations were then computed between the bootstrapped replication and the $z \mathrm{WMC}$ scores.

The three bootstrapping analyses used $\nu=.95, .50$, and .20 , respectively, which yielded actual correlations $z \mathrm{WMC} \times \mathrm{RAPM}$ of $.53, .23$ and .09 (left, center, and right panel in bottom row of Figure 3, respectively). These actual correlations span a large range of possible overall correlations between WMC and Raven's. The bottom left panel provides an idea of the variability expected when the population correlation is approximately equal to that observed in our study. The remaining two panels show that as the population correlation decreases, so does the slope of the itemwise correlations. The center panel roughly corresponds to the correlation observed by Unsworth and Engle (2005) and confirms that the effect of item-difficulty is sufficiently small under those circumstances to escape detection when statistical power is insufficient.

Operation span and RAPM To provide further empirical confirmation that a reduction in the overall correlation between WMC and RAPM attenuates the itemwise effect, we examined the correlation between the OSPAN subtask and RAPM. For this task, the overall correlation with Raven's was much lower, $r=.36, p<.001$. The slope of the regression line for the point-biserial itemwise correlations was not significantly greater than zero, $t(34)=1.39, p=.17, r^{2}=.05$. Likewise, a multilevel logistic regression (see Equation 1) in which $z \mathrm{WMC}$ was replaced by OSPAN failed to find a significant interaction between OSPAN and exponentiated ordinal item position $(p=.09)$. This result replicated the invariant relationship found by Unsworth and Engle (2005), supporting our claim the previously published results were obscured by method-specific variance; that is, with a single task, the correlation includes taskspecific variance that hides the true magnitude of the underlying correlation between constructs.

## Discussion

There were two principal differences between our methodology and previous research. First, we used a
composite measure of WMC which resulted in a higher overall correlation between WMC and Raven's performance. Second, we extended the test time for RAPM to the recommended duration, which resulted in increased overall accuracy. Our results converge on the conclusion that when there is a moderate to strong overall correlation between WMC and performance on the Raven's test of fluid abilities, then the role of WMC becomes increasingly more important as item difficulty increases. Our results suggest that other studies failed to find an effect of item difficulty because in their cases the overall correlation involving WMC was small in magnitude (e.g., Unsworth \& Engle, 2005; Wiley et al., 2011). Moreover, the study by Unsworth and Engle (2005) was subtly biased against finding an itemwise effect because of their use of a shorter, non-standard time period for completion of the RAPM test ( 30 instead of 40 minutes). This nonstandard timing made it more likely that performance on the most difficult items would be near the floor (because most people ran out of time before solving those items), thereby necessitating their removal for lack of variance with an ensuant reduction in the power of the analysis.

On the surface, the present work may appear to be merely a statistical issue, but given the intense theoretical attention and interpretation this issue has received, its resolution has considerable psychological implications. In particular, our research cautions against reliance on a null result which has been a substantial barrier to theorizing in this domain. Previously, any model hoping to account for the relationship between WMC and Raven's also had to explain the invariant relationship across item difficulty. The present result shows that this invariance is of questionable generality. By contrast, although not presented here due to space limitations, we have replicated our result using the Raven's Standard Progressive Matrices in another study. Our results therefore open the door for quantitative models of WMC and Raven's that do not predict this invariance. We now know that any model attempting to explain the relationship between the two tasks has to predict that high WMC will allow you to do particularly well on hard items.

Our results are compatible with theoretical analyses of Raven's performance that appeal to working memory as a repository for rules and intermediate results (e.g., Carpenter et al., 1990). Although those theoretical views have fallen out of favor, largely due to the apparent absence of a modulating effect of item difficulty on the relation between WMC and Raven's performance, our results suggest that abandoning those approaches may have been premature.

## Acknowledgments

This work was supported by ARC Discovery Grant DP120103888 to the first and second author, an Aus-
tralian Professorial Fellowship and DORA to the second author and a Jean Rogerson postgraduate scholarship to the third author. We thank Charles Hanich for assistance with data collection and Klaus Oberauer for comments on an earlier version of the manuscript. Address correspondence to the first author at daniel.little@unimelb.edu.au.

## References

Ackerman, P. L., Beier, M. E., \& Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs? Psychological Bulletin, 131, 30-60.
Box, G. E. P., \& Tidwell, P. W. (1962). Transformation of independent variables. Technometrics, 4, 531-550.
Carpenter, P. A., Just, M. A., \& Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the raven progressive matrices test. Psychological Review, 97, 404-431.
Conway, A. R. A., Cowan, N., \& Bunting, M. F. (2001). The cocktail party phenomenon revisited: The importance of working memory capacity. Psychonomic Bulletin $\mathcal{E}$ Review, 8, 331-335.
Daneman, M., \& Carpenter, P. A. (1980). Individual differences in working memory and reading. Journal of Verbal Learning \& Verbal Behavior, 19, 450-466.
Kane, M. J., Hambrick, D. Z., \& Conway, A. R. A. (2005). Working memory capacity and fluid intelligence are strongly related constructs: Comment on ackerman, beier, and boyle. Psychological Bulletin, 131, 66-71.
Lee, T.-H., White, H., \& Granger, C. W. J. (1993). Testing for neglected nonlinearity in time series models. Journal of Economics, 56, 269-290.
Lewandowsky, S., Oberauer, K., Yang, L.-X., \& Ecker, U. K. H. (2010). A working memory test battery for matlab. Behavior Research Methods, 42, 571-581.
Oberauer, K. (1993). Die koordination kognitiver operationen: Eine studie ber die beziehung zwischen intelligenz und working memory (the coordination of cognitive operations: A study on the relation of intelligence and working memory). Zeitschrift fr Psychologie, 201, 57-84.
Raven, J., Raven, J. C., \& Court, J. H. (1998). Manual for raven's progressive matrices and vocabulary scales. section 4: The advanced progressive matrices. Oxford, UK: Oxford University Press.
Unsworth, N., \& Engle, R. W. (2005). Working memory capacity and fluid abilities: Examining the correlation between operation span and raven. Intelligence, 33, 67-81.
Wiley, J., Jarosz, A. F., Cushen, P. J., \& Colflesh, G. J. H. (2011). New rule use drives the relation between working memory capacity and raven's advanced progressive matrices. Journal of Experimental Psychology: Learning, Memory, and Cognition, 37, 256263.

# When Tuesday comes before Threesday: Cross-linguistic differences in numerical transparency of time words predicts temporal reasoning strategy and performance 

Nian Liu (nian.liu@ou.edu)<br>Department of Modern Languages, Literatures, and Linguistics, University of Oklahoma<br>780 Van Vleet Oval, Norman, OK 73072, USA<br>Benjamin Bergen (bkbergen@cogsci.ucsd.edu)<br>Department of Cognitive Science, UC San Diego<br>9500 Gilman Drive, La Jolla, CA 92093, USA


#### Abstract

Time concepts are named differently across the world's languages. In English, the names for days of the week and months of the year are opaque - to people learning and using English, there's no obvious reason why Friday or September have the names they do. But in other languages, like Chinese, time concepts have numerically transparent names-the days of the week and months of the year are named using sequential numbers. We investigated whether having opaque versus mathematically transparent time concepts affects how people reason about time. Results show that Chinese speakers are more likely to spontaneously employ arithmetic when doing temporal calculations, which in turn improves the speed and accuracy of some time calculations. English speakers appear to use other strategies, such as sequential recitation.


Keywords: time concepts; temporal reasoning, mathematical ability; linguistic relativity.

## Introduction

The world's languages encode time terms in different ways. In English, names of days of the week (DOW) or months of the year (MOY) are derived from planetary or mythical terms (Zerubavel, 1985), and as a result are largely opaque to contemporary users of the language-why for instance do Wednesday or April have the names they have? By contrast, many other languages exhibit more numerically transparent naming systems. Chinese Mandarin is exemplary of this pattern. It uses numbers in the names of months and days. For DOW, this system begins with Monday as the first day of the week. Thus, xingqi yi, "weekday one", is Monday, xingqi er, "weekday two", is Tuesday, etc. The only exception is Sunday, the last day of the week, xingqi ri, which translates to "weekday of sun". Similarly, months in Chinese are numbered from one to 12 , with "month one" denoting January, "month two" February, and so on.

Do differences in how languages name time concepts cause differences in how speakers of those languages reason about time? For instance, does the numerical transparency of DOW and MOY terms affect the kinds of cognitive mechanisms people use to make temporal calculations? There has been little work on this issue. The most relevant line of work has focused on cross-linguistic effects of transparency of number systems themselves, showing that differences in number naming systems can affect cognitive development and non-linguistic performance. Most
relevantly, acquisition studies (Miura et al., 1993, 1994; Miller et al., 1995; Paik \& Mix, 2003) have found that preschool-aged children whose native languages employ more systematic naming systems for their numbers outperform their counterparts who speak languages that use less transparent number naming systems, on both number matching and number identification tasks. When asked to demonstrate numbers using combinations of individual unit cubes representing the quantity one and long blocks representing ten, Asian children whose languages use numerical names that are congruent with base 10 numeration systems (Fuson, 1990) were much more likely to use the blocks of ten in constructing multi-digit numbers than their counterparts, whose native naming systems were not similarly transparent. This led the authors of that study to argue that "numerical language characteristics may have a significant effect on cognitive representation of numbers" (Miura et al., 1994, p. 410), which in turn may enhance the performance of Asian-language-speaking children on tasks involving the concept of place value.

The types of names given to various symbolic systems, such as numbers, have also been shown to affect the problem solving abilities of competent symbol users. Seron and Fayol (1994) noticed that the number naming system in French-speaking Belgium is simpler than the one used in France (in Belgium, 98 is roughly "ninety-eight" but in France, it's "four-twenty-eighteen"). They reported that second-graders in France made more errors in number production than their Belgian counterparts. The effects of naming system also extend into adulthood and mathematical performance. For instance, adult English speakers have more difficulty reversing two-digit numbers ending in 1 (e.g., saying " 14 " when shown " 41 ") than Chinese speakers do, presumably a result of English's idiosyncratic rules for naming numbers between 11 and 19 (Miller \& Zhu, 1991).

In sum, differences in number naming systems affect the acquisition and use of number concepts. The current study investigated whether the same is true for the naming of time concepts (in this case, DOW and MOY). In particular, we asked whether the mathematically transparent naming of time concepts confers advantages on acquisition of time concepts and reasoning about time. There has been limited work on this question. Kelly et al. (1999) is the only systematic experimental investigation. They asked college
students in China and the United States to name the day or month that occurs a specified length of time before or after another given day or month. Chinese college students performed these calculations faster than American college students. Kelly and colleagues argued that the difference resulted from the use of different strategies as a consequence of the naming systems used in Chinese and English. However, this argument about the mechanisms used was based on participants' self-reports. Moreover, the calculation distance in Kelly et al.'s (1999) experiment was held constant-4 for the DOW task and 7 for the MOY task. The use of the constant distance caused some participants to predict the answers; one of the strategies participants reported was memorizing specific pairs of items. The experiment described below, in which distances vary, indirectly assesses the mechanisms adults use to perform temporal reasoning tasks, and whether these mechanisms vary with native language.

## Testing mechanisms indirectly

Kelly et al., (1999) reported that adult Chinese speakers outperform their English-speaking counterparts in time calculation tasks. This could be due to the differences in the numerical transparency of time words in the languages they speak. That is, the transparent numerical structure of Chinese time words might facilitate time calculation, by allowing Chinese speakers to employ arithmetic strategies made possible by the use of numerical names. For example, "Four days after Monday is what day?" translates to "Four days after Weekday one is what day?" To the Chinese speaker, this directly evokes arithmetic, $4+1$, and as a result, they might be able to use arithmetic quickly to produce the answer: "Weekday five" (Friday). The same should be true of making calculations about months. By contrast, English speakers do not have the arithmetic laid out for them in tasks like this, so they might rely on alternate strategies, like reciting a sequence of days or months.

But Kelly et al.'s (1999) results could alternatively be the result of other cultural differences, for instance, in the depth, length, and nature of the math education each group receives. It might be that the Chinese population that was sampled simply was better at performing mental calculations than their English-speaking counterparts.

Our experiment teases apart these two possibilities with a nuanced design, based on three predictions. First, if the use of different arithmetic strategies by Chinese and English speakers is responsible for the performance difference, then the Chinese advantage should disappear in cross-week or cross-year calculations. Chinese speakers may encounter difficulties in calculating distances across boundaries since they have to convert the answers into modulo-7 or modulo12 systems; 3 days after "weekday 5 " is not "weekday 8 ", but rather 8 modulo 7 , thus "weekday 1 "; 3 months after "month 11 " is not "month 14 ", but 14 modulo 12 , thus "month 2". If English speakers use a non-arithmetic strategy by default, then they should exhibit less increase in difficulty when calculating across boundaries.

A second prediction of the hypothesis that Chinese speakers use arithmetic more than English speakers is that the calculation of distances from Sunday (which is called "weekday sun" in Chinese) should be more difficult for Chinese speakers than from any other day in a week, since number is not used in naming this day. This irregularity may cause trouble in applying the arithmetic strategy, thus slowing down Chinese speakers' calculations, compared to calculation involving other days of the week. Again, by contrast, English speakers should show no increased difficulty when making calculations relative to Sunday.

A third prediction is that the Chinese speakers' speed of calendar calculating should not vary much with longer temporal distance if they primarily use an arithmetic method in the calculation, while the English speakers' performance may be negatively affected by increases in temporal distance to be calculated if they are reciting sequences.

## Method

## Participants

Thirty-two (22 female and 10 male) native Chinese speakers, college students from Beijing United College, ranging in age from 18 to $31(M=21, S D=4.9)$, and 40 native English speakers (19 female and 21 male), undergraduate students at the University of Hawai'i, aged 18 to $29(M=21, S D=2.5)$, participated in the experiment either for extra credit in an introductory linguistics class or for five dollars or the equivalent.

## Materials and design

Each participant completed two temporal reasoning tasks, pertaining to DOW and MOY respectively. For each, we manipulated two factors-Boundary Type (Within/Across boundary) and Direction (Forward/Backward), which produced four question types, as below (showing only DOW). Both factors were manipulated within participants. Language was a between-participants factor. An additional factor, Sunday, applied only to the DOW blocks. Half of the DOW questions involved Sunday, to reveal eventual effects of this non-numerically-named day in Chinese.
(1a) If today is Monday, three days from now is what day?
(Within/Forward)
(1b) If today is Thursday, two days ago was what day?
(Within/Backward)
(2a) If today is Saturday, three days from now is what day?
(Across/Forward)
(2b) If today is Tuesday, five days ago was what day?
(Across/Backward)
Calculation distance differed across conditions; distance ranged from 1 to 7 for DOW questions (1-4 for Within and 2-7 for Across) and 2 to 12 for MOY questions (2-10 for Within and $2-12$ for Across) in order to match the cyclical nature of the weeks and months. There were 32 questions in the DOW block and 48 in the MOY block.

## Procedure

Participants were seated in front of a computer. Each trial began with a fixation cross in the center of the screen. Participants initiated the trial by pressing the SPACE bar when they were ready. They heard a recorded voice read a question to them over headphones. They were instructed to speak the answer as quickly and accurately as possible after hearing the question into a microphone, which was used as a voice key, connected to an E-Prime SR-BOX. A digital recorder was also used to record the answers.

Four practice items preceded the experiment, which was divided into two blocks, one for DOW and one for MOY. Block order was counterbalanced across participants. The questions were randomized within each block. There was a short break between blocks.

## Measures

Reaction times in milliseconds were measured from the offset of the question to the onset of the participant's answer. Accuracy (individual proportions of correct answers were arcsine transformed to comply with the normal distribution premise) were calculated as another dependent measure.

## Results

All filler syllables, tongue clicks, partial responses, and repeated responses (due to failure to trigger the voice key) were manually excluded. No participants were excluded because of outlying RTs or low accuracy.

## Error analysis

Figure 1 shows mean accuracy for each Boundary Type and Direction by language in the MOY and DOW tasks. We performed separate repeated-measures ANOVAs with participants $\left(\mathrm{F}_{1}\right)$ and items $\left(\mathrm{F}_{2}\right)$ as random factors. In the participants analysis, Language (Chinese/English) was between participants but Direction (Forward/Backward), and Boundary Type (Within/Across) were within subjects. For items, Language was a within-items factor and the other factors (Boundary Type and Direction) were between-items.

As predicted, Chinese speakers made fewer errors than English speakers on the MOY task-a main effect of Language: $F_{1}(1,68)=26.846, p<0.001, F_{2}(1,44)=54.265$, $p<0.001$. However, as Figure 1 shows, this advantage was carried by the Within Boundary questions, confirmed by an interaction between Language and Boundary Type, $F_{1}$ (1, $68)=34.842, p<0.001 ; F_{2}(1,44)=44.876, p<0.001$.

As the second part of Figure 1 shows, Language was a less robust factor in the DOW task. The main effect of Language did not reach significance by participants analysis, but it did by items, $F_{2}(1,28)=5.177, p=0.031$. This indicates that Chinese speakers did not have a global accuracy advantage over English speakers. The lack of Language effect was probably driven by the Chinese speakers' relatively poor performance in the AcrossBoundary condition. As with MOY, an interaction effect
was found between Language and Boundary Type, $F_{1}(1,68)$ $=18.537, p<0.001 ; F_{2}(1,28)=24.596, p<0.001$. Once again, Chinese speakers produced more errors in cross-week calculations compared to their English-speaking counterparts.


Figure 1: Accuracy rate on the MOY (upper) and DOW (lower) task for Chinese and English speakers.

Questions involving Sunday As discussed in the Method section, half of the DOW questions were Sunday-related questions. These questions could have been the cause of the Chinese speakers' low DOW accuracy if the irregular Sunday term ("weekday sun" instead of "weekday seven") affected the use of an calculation strategy. Statistical analysis confirms this. Chinese speakers were significantly less accurate on Sunday questions than non-Sunday questions, $F_{1}(1,29)=2.38, p=0.024 ; F_{2}(1,30)=2.83, p=$ 0.031 , while English speakers were not sensitive to Sunday and showed no difference in accuracy on Sunday questions (see Figure 2).


Figure 2: Chinese- and English-speaking participants' accuracy in different Sunday conditions.

## Reaction time analysis

All trials with incorrect responses were removed (16.7\% of the data). We also removed all responses that were greater
than 2.5 standard deviations from the mean of all responses in each of the four conditions (Boundary Type x Direction) for each language group. This excluded another $2.72 \%$ of the data. No participants or items were removed for reasons of accuracy or outlying $S D$. The reaction time data approximate normal distribution after this cleaning. Reaction times for correct responses were analyzed using a repeated-measures ANOVA with Boundary Type (Across/Within) and Direction (Forward/Backward) as within-subject factors and Language (Chinese/English) as a between-subject factor. An Item analysis took Language as a within-items factor and Boundary Type and Direction as between-items factors. The results in the MOY and DOW tasks, seen in Figure 3, show mean reaction time for each Boundary Type and Direction by language.

The results from MOY block mirrored the error analysis. A large main effect of Language, $F_{1}(1,68)=36.155, p<$ $0.001, F_{2}(1,44)=134.986, p<0.001$, confirmed that Chinese speakers were faster than English speakers. There were also significant interactions between Language and Boundary Type, $F_{1}(1,68)=35.809, p<0.001, F_{2}(1,44)=$ 9.613, $p<0.001$, and between Language and Direction, $F_{1}(1,68)=13.423, p<0.001, F_{2}(1,44)=8.407, p=0.006$. These effects showed that Chinese speakers gained more of a speed advantage from Within month questions than English speakers did, but that English speakers gained more of an advantage from Forward calculations.


Figure 3: Reaction times on the MOY (upper) and DOW (lower) task for Chinese and English speakers.

Reaction time results from the DOW task are more complicated because of the presence of Sunday. The Chinese speakers were consistently faster than their Englishspeaking counterparts, as shown by a main effect of language, $F_{1}(1,68)=20.617, p<0.001, F_{2}(1,28)=63.776, p$ $<0.001$. Both language groups answered Within questions faster than the Across ones, $F_{1}(1,68)=161.850, p<0.001$,
$F_{2}(1,28)=27.446, p<0.001$, and the Forward questions faster than the Backward ones, $F_{1}(1,68)=42.243, p<0.001$, $F_{2}(1,28)=11.496, p=0.002$. There was also an interaction between Language and Direction, $F_{1}(1,68)=15.407, p<$ $0.001, F_{2}(1,28)=9.244, p=0.005$. English speakers were much faster with Forward questions, $F_{1}(1,39)=86.977, p<$ $0.001, F_{2}(1,28)=13.090, p=0.001$. Chinese speakers were also faster in answering Forward questions than Backward questions, but the difference was much smaller, $F_{1}(1,29)=$ $13.460, p=0.001, F_{2}(1,28)=4.933, p=0.035$.

Questions involving Sunday The analysis of the accuracy data above showed that Chinese speakers made more errors on questions involving Sunday. But their reaction times were unaffected, as Figure 4 shows. A two-way ANOVA showed a main effect of Language, $F_{1}(1,68)=15.681, p<$ $0.001, F_{2}(1,30)=34.942, p<0.001$; Chinese speakers were faster than English speakers. But there was no interaction.


Figure 4: Reaction times in Sunday and Non-Sunday conditions for Chinese and English speakers.

## Calculation Distance analysis

Because the mean distance varied across conditions, we also did exploratory analyses by calculation distance, to see if different calculation difficulty revealed strategy differences. Reaction times by distance are presented in Figure 5.

As shown in Figure 5, the two language groups showed different level of sensitivity to the distance calculated. The Chinese speakers' reaction times were far less sensitive to either distance or direction, especially in the within-week condition. This is consistent with the hypothesis that they were using addition and subtraction of small numbers. The across-week condition is more complicated, as an extra step of modulo calculation could have been involved for the Chinese speakers.

The English speakers' reaction times were more strongly affected by the length of temporal distances. As can be seen on the right side of Figure 5, questions that required backwards calculations took much longer than Forward questions, presumably because counting backwards is less familiar. Moreover, reaction times increase steadily as the distances increase. The rise and drop of reaction times with longer distances suggests that the English speakers were applying different approaches when encountering calendar questions with different distances. When the distances grew longer, which makes verbal list counting a less efficient strategy, they may have flexibly and strategically switched to methods such as using numerical equivalents as a shortcut to reciting the list, such as "counting forward by the

12's or 7's complement to solve backward problems" (Kelly et al., 1999). The partial use of an arithmetic strategy was self-reported by some of the English participants in Kelly et al.'s work, and is consistent with the reaction time evidence we find here.


Figure 5: Reaction times of Chinese (left) and English (right) in Within (upper) and Across (lower) Week calculations (The $x$-axis shows calculation distances.)


Figure 6: Reaction times by Chinese (left) and English (right) speakers in Within (upper) and Across (lower) Year calculations (The $x$-axis shows calculation distances.)

The year cycle has longer distances and so can provide more information about strategies used in response to possible calculation difficulties. The two language groups' reaction time data were categorized by distance for the Within task and for the Across task, as shown in Figure 6.

Again, the Chinese speakers' reaction times were not affected nearly as much as the English speakers' were by calculation distance or direction, especially in the Within Year tasks, again consistent with the possibility that they are calculating arithmetically. On the other hand, the English speakers present a more complicated picture of responding to different distances. They were generally faster with Forward than Backward questions, and they spent more time when the distances got longer, except for the longest distances (10 through 12), with which they might be using strategies other than reciting due to the difficulty of counting all the way through the month list.

## Discussion

Languages differ in how numerically transparent their time words are, and we hypothesized that these differences would affect the temporal cognition of adult speakers of languages with distinct systems. Our results provide several types of experimental evidence that Chinese and English speakers use different strategies in temporal calculation tasks. As a result, Chinese speakers determine temporal distance calculations faster and more accurately. These findings are not consistent with the hypothesis that other factors, such as general mathematics ability, cause differences in overall accuracy or speed.

There are three pieces of relevant evidence. First, Chinese speakers were consistently more accurate in the Within Week and Within Year tasks, but their accuracy dropped significantly and their advantage disappeared with calculations across week or year boundaries. In contrast, the English speakers' answers were not particularly sensitive to boundary crossings. For the reaction time results, although the Chinese speakers were still faster than the English speakers in the Across Boundary calculations, they were slower compared to their own Within Week calculations. So although arithmetic strategies for time may be advantageous for some local calculations, the Chinese speakers' advantage in accuracy and speed diminishes when answering questions that involve temporal boundary crossings.

Second, the results confirmed that the irregularity of Sunday's Chinese name causes trouble in applying the arithmetic strategy, resulting in more mistakes by Chinese speakers in answering questions involving Sunday compared to calculations involving other days of the week. By contrast, English speakers showed no increased difficulty when making calculations relative to Sunday, and their accuracy rates were not affected by questions involving Sunday.

Finally, Chinese speakers' reactions seem not to systematically relate to distance or directions of calculation, whereas the English speakers' reaction times for Forward calculations increase steadily and substantially, implying an
increase in time spent reciting day or month lists, and their pattern on Backward questions were more complicated. We hypothesized that versatile/mixed strategies might be used due to the difficulty of reciting lists backwards.

In sum, these results suggest that differences in calendar terms between the two languages lead to dominant arithmetic or list-reciting strategies for speakers of those languages. The verbal list strategy is substantially slower than the number-transferring one, yet the latter results in more possible errors across boundaries because of required modulo 7 and modulo 12 calculations.

Previous cross-linguistic studies have shown that there is at least some causal influence from language to non-verbal cognition and unconscious habitual thought (Kay \& Kempton, 1984; Lucy, 1992; Gumperz \& Levinson, 1996, etc.). More specifically, the way time is described in a language can affect its speakers' conceptualization of time (Boroditsky, 2001) and can even shape low-level mental processes in psychophysical tasks (Casasanto et al., in revision). The current study adds to this line of study by showing that transparent numerical structure of the calendar might facilitate calendar calculation, causing Chinesespeaking adults to outperform their English-speaking counterparts in time calculation tasks, by exhibiting shorter reaction times and making fewer errors. We conclude that adults' temporal reasoning abilities differ, depending on the transparency of the naming systems that their languages employ for time sequences. In general, such a finding supports the hypothesis that linguistic differences can produce non-linguistic consequences, in this case in affecting people's reasoning about time (Boroditsky, 2000, 2003; Boroditsky \& Ramscar, 2002, Matlock et al., 2005; Núñez and Sweetser, 2006; Casasanto \& Boroditsky, 2008).

## Acknowledgments

This study was funded by National Science Foundation Doctoral Dissertation Grant \#1124006.

## References

Boroditsky, L. (2000). Metaphoric structuring: Understanding time through spatial metaphors. Cognition, 75(1), 1-28.
Boroditsky, L. (2001). Does language shape thought? English and Mandarin speakers' conceptions of time. Cognitive Psychology, 43(1), 1-22.
Boroditsky, L. (2003). Linguistic relativity. Encyclopedia of Cognitive Science. MacMillan Press.
Boroditsky, L. \& Ramscar, M. (2002). The roles of body and mind in abstract thought. Psych. Science, 13(2), 185188.

Casasanto, D., \& Boroditsky, L. (2008).Time in the mind: Using space to think about time. Cognition 106, 579-593.
Casasanto, D., Fotakopoulu, O., Pita, R., \& Boroditsky, L. (In Revision). How deep are effects of language on thought? Time estimation in speakers of English and Greek. Cognition.

Friedman, W. J. (1990). About time: Inventing the fourth dimension. Cambridge, MA: MIT Press.
Fuson, K.C. (1990). Conceptual structures for multiunit numbers: Implications for learning and teaching multidigit addition, subtraction, and place value. Cognition and Instruction, 7, 343-403.
Gumperz, J. \& Levinson, S. (1996). Rethinking linguistic relativity. Cambridge: Cambridge University Press.
Kay, P., \& Kempton, W. (1984). What is the Sapir-Whorf hypothesis? American Anthropologist, 86, 65-79.
Kelly, M., Miller, K., Fang, G., \& Feng, G. (1999). When days are numbered: calendar structure and the development of calendar processing in English and Chinese. J. of Experimental Child Psych. 73, 289-314.
Lucy, J. (1992). Language diversity and thought. A reformulation of the linguistic relativity hypothesis. Cambridge University Press.
Matlock, T., Ramscar, M., \& Boroditsky, L. (2005). On the experiential link between spatial and temporal language. Cognitive Science, 29, 655-664.
Miller, K. F., \& Zhu, J. (1991). The trouble with teens: Accessing the structure of number names. Journal of Memory and Language, 30, 48-68.
Miller, K. F., Smith, C. M., Zhu, J. J., \& Zhang, H. C. (1995). Preschool origins of cross-national differences in mathematical competence-the role of number-naming systems. Psychological Science, 6, 56-60.
Miura, I., Okamoto, Y., Kim, C., Steere, M., \& Fayol, M. (1993). First graders' cognitive representation of number and understanding of place value: Cross-national comparisons-France, Japan, Korea, Sweden, and the United States. J. of Educational Psychology, 85, 24-30.
Miura, I., Okamoto, Y., Kim, C., Chang, C-M., Steere, M., \& Fayol, M. (1994). Comparisons of children's cognitive representation of number: China, France, Korea, Sweden, and the United States. International Journal of Behavioral Development, 17, 401-411.
Núñez, R., \& Sweetser, E. (2006). With the future behind them: Convergent evidence from Aymara language and gesture in the crosslinguistic comparison of spatial construals of time. Cognitive Science, 30(3), 401-450.
Paik, Jae H., \& Mix, Kelly S. (2003) U.S. and Korean children's comprehension of fraction names: A reexamination of cross-national Differences. Child Development, 74(1), 144-154.
Seron, X., \& Fayol, M. (1994). Number transcoding in children: A functional analysis. British Journal of Developmental Psychology, 12, 281-300.
Zerubeval, E. (1985). The seven day circle: The history and meaning of the week. Chicago: Univ. of Chicago Press.

# Modeling Spatial Ability in Mental Rotation and Paper-Folding 

Andrew Lovett (Andrew@cs.northwestern.edu)<br>Kenneth Forbus (Forbus@northwestern.edu)<br>Qualitative Reasoning Group, Northwestern University, 2133 Sheridan Road<br>Evanston, IL 60208 USA


#### Abstract

Spatial ability tests like mental rotation and paper-folding provide strong predictions of an individual's achievement in science and engineering. What cognitive skills are involved in them? We use a computational model to analyze these tasks, asking how much information must be processed to perform them. The models demonstrate that in some cases stimuli can be vastly simplified, resulting in consistent performance regardless of stimulus complexity. The ability to produce a scaled-down representation of a complex stimulus may be a key skill underlying high spatial ability.


Keywords: spatial ability; mental rotation; paper-folding; cognitive modeling.

## Introduction

There is strong evidence linking spatial ability to academic achievement. Children who perform well on spatial ability tests are more likely to study STEM disciplines (Science, Technology, Engineering, and Mathematics) and to go into a STEM profession (Shea, Lubinski, \& Benbow, 2001; Wai, Lubinski, \& Benbow, 2009). This effect holds even when controlling for verbal and mathematical ability, suggesting that spatial ability is an independent component of intelligence. If we are to improve STEM achievement, it is critical that we better understand the skills that compose spatial ability and how they can be taught.

Traditionally, spatial ability has been evaluated using tasks such as mental rotation and paper-folding. In mental rotation (Figure 1A, 1B), individuals are shown two shapes and asked whether a rotation of one shape could produce the other. In paper-folding, they are shown a line-drawing of paper and asked to imagine the results of unfolding (Figure 1C) or folding up (Figure 1D) the paper. Both tasks appear to measure spatial visualization, the ability to manipulate mental representations of images (McGee, 1979). There is evidence that the tasks are linked, with training on one improving performance on the other (Wright et al., 2008). However, many questions remain about what skills enable people to perform them quickly and accurately.

Here, we study the mental rotation and paper-folding tasks using a computational model. The model operates directly on 2D line drawings (sketches), automatically generating representations, transforming them, and evaluating the results of the transformation. We use the model to analyze the tasks, asking how much information must be encoded and carried through the transformations to
perform each task consistently. This analysis allows us to address a longstanding debate about the effects of shape complexity on mental rotation. It also provides hypotheses about the skills supporting fast, efficient mental rotation, and thus the skills underlying spatial ability.

We begin with background on mental rotation and the question of shape complexity. We show how paper-folding appears to violate many researchers' conclusions, as it involves simple shapes but requires great deliberation and effort. We next present our computational model, which builds on previous cognitive models of perception, comparison, and visual problem-solving (Falkenhainer, Forbus, \& Gentner, 1989; Lovett \& Forbus, 2011). We apply the model to the two tasks, determining the amount of information that must be carried through the transformations, and showing why paper-folding is a more difficult task. Finally, we discuss the results and consider the ramifications for spatial ability in general.

## Background

## Mental Rotation

Mental rotation is frequently used to evaluate spatial ability (Vandenberg \& Kuse, 1978). Typically the distractors-the shapes that aren't a valid rotation-are mirror reflections. When they are presented sequentially, there is often a cue indicating what the orientation of the second shape will be (e.g., Cooper \& Podogny, 1976; Figure 1B). A common finding across task variations is that the response time is proportional to the angle of rotation between the shapes. That is, response times increase linearly with angular distance. This finding has led to the claim that people use a mental space, analogous to the physical space, and that they rotate their representation through this space just as an object might rotate physically (Shepard \& Cooper, 1982).

One common question concerns how shapes are rotated through mental space. Are they rotated piecemeal, with one part rotated at a time, or are they rotated holistically, with every part rotated together (Bethell-Fox \& Shepard, 1988)? These two possibilities produce different predictions about how shape complexity interacts with rotation speed. If shapes are rotated piecemeal, then people should rotate complex shapes more slowly, because there are more parts to rotate. If shapes are rotated holistically, then shape complexity may not affect rotation speed.


Figure 1. Mental rotation (A, B) and paper-folding (C, D) tasks (A: Shepard \& Metzler, 1971; B: Cooper \& Podogny, 1976; C: Ekstrom et al., 1976; D: Shepard \& Feng, 1972).

The results on shape complexity provide evidence for both piecemeal and holistic rotation. Overall, it appears that rotation speed depends more on other factors, such as the familiarity of the objects (Bethell-Fox \& Shepard, 1985; Yuille \& Steiger, 1982), the similarity of the distractors (Folk \& Luce, 1987), and the strategy and overall ability of the participant (Yuille \& Steiger, 1982; Heil \& JansenOsmann, 2008). These findings suggest that dealing with shape complexity may itself be a spatial skill. Skilled participants may apply heuristics to simplify shapes for rapid rotation. However, when these heuristics fail, they are reduced to rotating one piece at a time, which is slower.

One straightforward heuristic for simplifying a shape is to ignore parts of it. When Yuille and Steiger (1982) told participants they could complete a mental rotation task using only the top halves of the shapes, participants rotated the shapes more quickly. Alternatively, participants might utilize scalable representations (Schultheis \& Barkowsky, 2011) that support dynamic variation of detail based on task demands. Both the degree and the type of detail may vary. For example, while we can imagine both the locations and orientations of objects in space, a task might require considering only one of these. In this paper, we use the term spatial smoothing for any process that removes spatial detail, producing a simpler representation.

Participants may smooth out the details in complex shapes, producing representations with equal complexity to those of simpler shapes. However, when the distractors are particularly similar to the base shapes, participants may require additional detail, and so they may use more complex representations that are more difficult to rotate.

This hypothesis leads immediately to two predictions: 1) When similarity of distractors is kept constant and relatively low, people should rotate shapes at the same rate regardless of shape complexity. 2) As distractors become more similar,
people should rotate shapes more slowly, particularly when the shapes are complex. There is evidence supporting both predictions (1: Cooper \& Podgorny, 1976; 2: Folk \& Luce, 1987).

## Paper-Folding

In contrast with mental rotation, paper-folding has seen relatively little study. This is surprising, given that it is also often used to evaluate spatial ability (Ekstrom et al., 1976). Here, we focus on a version of paper-folding that emerged at about the same time as mental rotation (Shepard \& Feng, 1972). While this version is used less frequently in spatial ability evaluations, there is direct evidence linking it to mental rotation (Wright et al., 2008).

Figure 1D shows an example. The letters have been added for illustrative purposes and are not part of the stimulus. In this task, participants are shown six connected squares, representing the surfaces of a cube that has been unfolded. Two edges are highlighted by arrows, and one square is grayed out, indicating it is the base of the cube. Participants are asked whether the highlighted edges would align if the squares were folded back into a cube.

Unlike mental rotation, this task requires a sequence of rotations. For example, Figure 1D requires three rotations. One solution (Figure 2) would be: 1) Rotate squares A, B, and C up, so that they stick out from the plane. 2) Rotate squares $B$ and $C$ down to make the top surface of the cube. 3) Rotate square $C$ farther down, making the front surface of the cube. At this point, the two arrows align perfectly.

Surprisingly, even though each of these three rotations seems simple, they appear to be piecemeal rotations. Participants' response times are not a function of the number of rotations performed, but of the number of times every square is rotated. In this case, three squares are rotated (Figure 2B), then two squares (2C), then one square (2D),


Figure 2. Possible solution for Figure 1D.
so the overall number of squares rotated is $3+2+1=6$. The response times reflect the six rotations, suggesting participants rotate a single square at a time.

Why should participants require piecemeal rotation for such apparently simple shapes? We propose that, unlike many mental rotation tasks, little spatial smoothing can be performed. The precise location and orientation of every surface rotated is critical to performance. In Figure 2A, the location and orientation of square $A$ determines where the second rotation occurs, and the location and orientation of square B determines where the third rotation occurs.

If this proposal is true, it may shed light on how and when spatial smoothing can be applied, and what happens when it cannot be used. Understanding this requires determining how much spatial information must be rotated in each task. To better answer this question, we developed a computational model of the tasks.

## Model

The spatial ability model is built within CogSketch, a sketch understanding system. Below, we present CogSketch and its framework for cognitive modeling. We then describe how the model performs mental rotation and paper-folding.

## CogSketch

CogSketch is an open-domain sketch understanding system (Forbus et al., 2011). Users sketch a scene by drawing one or more objects. It is the user's responsibility to manually segment a sketch into objects, indicating when they have finished drawing one object and begun on the next.

Given a set of objects, CogSketch automatically generates a representation of the scene. While CogSketch does not model the process of visual perception, its representations are a model of those produced by human perception. The representations are based on two psychological claims: 1) Spatial representations include a qualitative or categorical component and a quantitative or metric component (Kosslyn et al., 1989; Forbus, et al 1991). When possible, people use the qualitative component during reasoning. CogSketch computes qualitative spatial relations between objects, e.g., indicating that one object is right of another or that two objects intersect. 2) Spatial representations are hierarchical, meaning they can represent a scene at different levels of abstraction (Palmer, 1977; Marr \& Nishihara, 1978). CogSketch can represent objects and the relations between them, or it can represent the edges of an individual object and the relations between those edges. To produce an edge-level representation, CogSketch segments an object's contour into edges at points of maximum curvature (e.g., Figure 3A; Lovett et al., 2012). Once edges have been
A)

B)

C)


Figure 3. A: Shape segmented into edges. B: Result of Gaussian smoothing. C: Result of selecting 4 longest edges.
computed, it generates qualitative spatial relations between the edges, e.g., indicating that two edges are parallel or that a corner between edges is convex.

CogSketch models visual comparison using the StructureMapping Engine (SME) (Falkenhainer et al 1989), a domain-general cognitive model based on Gentner's (1983) structure-mapping theory. SME compares two qualitative representations by aligning their common relational structure, highlighting commonalities and differences. For example, suppose SME is comparing two shapes like the one in Figure 3A. Each representation will contain entities (symbols representing each edge), attributes (features of the edges, such as straight vs. curved), first-order relations between edges (e.g., indicating that a corner between edges is convex), and higher-order relations between other relations (e.g., indicating that two corners are adjacent along the shape). By aligning the common relations, SME can determine the corresponding edges in the two shapes.

Modeling in CogSketch CogSketch possesses two key features that support modeling psychological experiments. First, in addition to sketching by hand, users can import shapes from another program such as PowerPoint. Perfectly straight line drawings from an experiment can be replicated in PowerPoint and imported into CogSketch, providing it with the same stimuli as those shown to human participants.

Second, CogSketch includes a Spatial Routines language for writing cognitive models. Spatial Routines, which builds on Ullman's (1984) concept of visual routines, provides modelers with a set of cognitive operations. These include visual perception, visual comparison, and spatial transformation operations. Modelers can parameterize these operations and combine them in different ways to produce a spatial routine. Each routine is both a theoretical model of how people perform a task and a fully automated computational model. The computational model can be run on visual stimuli, and its performance can be compared to human responses to evaluate the theoretical model.

We have previously modeled visual problem-solving tasks such as geometric analogy (Lovett \& Forbus, 2012), Raven's Progressive Matrices (Lovett, Forbus, \& Usher, 2010), and an oddity task (Lovett \& Forbus, 2011).

## Mental Rotation

We modeled a classic sequential mental rotation task because it presents the clearest evidence for efficient, holistic rotation. In this task (Figure 1B; Cooper \& Podogny, 1976), participants are presented with three stimuli in sequence: 1) They see the base shape. 2) They see an arrow indicating what the target shape's orientation will be. They are encouraged to mentally rotate the base shape to that orientation and press a button when they are done. 3) They see the target shape, and they indicate whether it is the same as the rotated base shape. The amount of time participants spend on step 2) indicates the rotation time. The key finding of the experiment was that rotation time did not increase as the shape complexity increased from a 6-sided
polygon to a 24 -sided polygon. Our model is designed to evaluate whether spatial smoothing can explain this finding.

Input In CogSketch, sequences of images can be input using a sketch lattice, a grid which divides the sketch space. For this model, we used a three-cell lattice to represent the three phases of each experimental trial. Stimuli were reproduced in PowerPoint. The experimenters traced over images of the original stimuli, ensuring that the number of sides in the new polygons was the same as in the original.

Representations CogSketch uses edge-level representations to perform two-dimensional shape transformations and comparisons. The qualitative edge-level representations describe spatial relations between edges, as summarized above. The quantitative representation includes for each edge: 1) The location of its center. 2) Its two-dimensional orientation. 3) Its length. 4) Its curvature.

When a shape is scaled, rotated, or reflected, each individual edge is transformed. This has little effect on the qualitative representation, but it can change each of the four features in the quantitative representation.

Shapes are compared in a two-step process. 1) Qualitative representations are compared using the Structure-Mapping Engine. This identifies the corresponding edges in the two shapes. 2) Each corresponding edge pair's four quantitative values are compared. If every pair is quantitatively the same, the shapes are identical.

Strategy Given a stimulus such as Figure 1B, the model automatically constructs an edge-level representation of each shape. It detects the orientation of the arrow and rotates the base (leftmost) shape accordingly. It then compares the rotated base shape to the target shape to determine whether they are identical.

Spatial Smoothing Recall that scalable representations allow two forms of spatial smoothing: spatial detail may be smoothed out, or certain types of spatial information may be removed entirely. There are many possible ways to smooth data, e.g. apply a Gaussian filter to the entire shape (Figure 3B). Such an approach would lose critical information about the nature of the edges making up the object. Here, we use a simple sampling strategy: we remove all but the four longest edges (Figure 3C). This operation produces representations of equal size for all shapes, regardless of their initial complexity, which is what we desire.

The quantitative representations contain four types of spatial information. We propose that spatial smoothing might remove three, leaving only a single type. In our evaluation, we test whether the task can be performed using only edge locations or using only edge orientations.

## Paper Folding

We modeled the paper-folding task shown in Figure 1D.
Input Each paper-folding stimulus was recreated in PowerPoint. The square representing the base of the cube
was given a solid gray fill (CogSketch can recognize elements by their color). CogSketch was given three objects: the unfolded cube and the two arrows pointing to the critical edges.

Representations This model required the development of a new representational level: surfaces. Surfaces are closed shapes making up the sides of three-dimensional objects. Each square of the unfolded cube is a separate surface. Surfaces can be computed easily using CogSketch's existing ability to find closed cycles of edges.

This model does not need to find corresponding elements, so it requires only quantitative representations. Each surface is represented with: 1) The location of its center. Locations are now in three-dimensional space, unlike with the previous model. 2) Its orientation, i.e., the orientation of a vector orthogonal to the surface. Three-dimensional orientations are unit vectors containing ( $x, y, z$ ) components, unlike the single value required for two dimensions. 3) A list of edges going around the surface. Each edge has its own individual location and orientation.

Three-dimensional rotations are performed about an axis in three-dimensional space. For example, in Figure 2B, surfaces $\mathrm{A}, \mathrm{B}$, and C are rotated about the edge connecting surface A to the base of the cube.

Strategy Given a stimulus, the model segments the object into edges and surfaces. Using the arrows, it identifies the two critical edges and their associated surfaces. In Figure 1D, the critical surfaces are the base of the cube and surface C. It folds each critical surface back into the cube shape via two spatial operations: 1) Trace along adjacent surfaces from the critical surface to the base surface. For surface C, this would produce the following trace: C->B->A->base. 2) Rotate $90^{\circ}$ about the edge between surfaces in the reverse trace order. First rotate surface A about the edge between the base and surface $A$. Because surfaces $B$ and $C$ are connected to A, they will also be rotated (Figure 2B). Next, rotate surface $B$ about the edge between surfaces $A$ and $B$. Because C is connected, it will also be rotated. And so on.

The model performs these two operations on each critical surface. In Figure 1D, the second critical surface is not rotated because it is already the base. The model takes the resulting shape and evaluates whether the two critical edges are aligned, i.e., have the same location and orientation.

Spatial Smoothing In this task, the location and orientation of most edges is irrelevant to the task; only the two critical edges matter. If an edge lies along only one surface, it can be ignored. If an edge lies between two surfaces (e.g., the edge between surfaces $A$ and $B$ ), it is important for determining the axis of rotation. However, because this task involves perfectly regular square shapes, a heuristic can be used: when rotating between two surfaces, place the axis of rotation halfway between the surfaces' centers, and orient it within the plane of those surfaces, perpendicular to the line connecting their centers. Due to this heuristic, all edges can be ignored except the two critical edges.

This means the following is being considered: the location and orientation of each rotated surface (and of the base), and the location and orientation of each critical edge. It may be possible, again, to consider only the location or orientation of the critical edges. However, both a surface's location and its orientation must be used in computing axes of rotation.

## Simulation

## Mental Rotation

The original experiment (Cooper \& Podogny, 1976) used five base shapes which varied in complexity from a 6 -sided figure to a 24 -sided figure. On each trial, participants were cued to rotate shapes some multiple of $60^{\circ}$ using the rotated arrow. They were then presented with a target shape at the new orientation. This could be: a) the same shape; b) a mirror-reflected shape; or c) a shape with some of the points permuted from the base shape. While distractors of type b) are commonly used, the distractors of type c) were added to test how carefully participants were rotating the shapes.

In this simulation, we ran the model on all six base shapes. However, we used only a single rotation value $\left(60^{\circ}\right)$, and only the mirror-reflected distractors. The single rotation was used because other rotations are mathematically equivalent and should not place additional demands on the model. The mirror-reflected distractors were used because they are the most common distractors found across different mental rotation tasks. In Future Work, we consider the challenge of recognizing permuted distractors.

Results Recall that the model spatially smoothed each shape, removing all but the four longest edges. This proved sufficient for recognizing that same shapes were the same and mirror-reflected shapes were different. Furthermore, when only edge orientations or only edge locations were used, either was sufficient for performing the task. We can conclude that when the distractors are sufficiently different, very little information must be rotated to perform mental rotation, and the complexity of the shapes is irrelevant.

## Paper-Folding

Shepard and Feng (1972) identify ten different classes of paper-folding problems, based on the number of folds and the number of squares per fold. For example, Figure 1D is a class I problem, in which $3+2+1$ squares must be rotated for one critical edge and no squares must be rotated for the other. Their paper provides one example of each class.

In this simulation, we ran the model on the single example of each class. Other instances of a class are mathematically equivalent. As in the original study, there was one nonmatch problem (where folding did not cause the critical edges to align) for each match problem. Nonmatch problems were created by randomly rotating an arrow so that it pointed to an adjacent edge in the same square.

Results Recall that the model rotated each surface's center and orientation. The two critical edges were rotated also, but
all other edges were ignored. This proved sufficient for solving all problems-the model correctly distinguished between matches and nonmatches. Furthermore, when only edge orientations or only edge locations were used to compare the critical edges, either was sufficient.

## Discussion

Having successfully modeled both tasks, we can now consider how much data must be transformed to perform mental rotation and paper-folding. The mental rotation model required only four values: the orientations or locations of the four longest edges.

Now, suppose the model were rotating two surfaces during paper-folding. This would require five values: the location and orientation of the two surfaces, and the location or orientation of the critical edge. Furthermore, these values are three-dimensional, whereas the mental rotation values were two-dimensional. In the computational model, threedimensional values are far more complex-for example, an orientation is a vector with ( $\mathrm{x}, \mathrm{y}, \mathrm{z}$ ) components. Spatial data is likely implemented differently in the brain, but there still may be increased processing demands for three dimensions (see Jolicoeur et al., 1985 for a discussion; but see Shepard \& Metzler, 1988).

These results support our initial hypothesis. Rotation rate appears to depend less on shape complexity than on the degree and type of detail required by the task. Paper-folding requires that more information be transformed, even when only two surfaces are being rotated and the surfaces are perfect squares. We propose that paper-folding overwhelms people's spatial working memory, forcing them to rotate one surface at a time in a piecemeal fashion.

## Conclusions and Future Work

This paper demonstrates how visual stimuli can be encoded, transformed, and compared. The computational model builds on existing cognitive models of visual representation and comparison. While we wish to avoid strong conclusions about how spatial information is represented and transformed in the brain, the model provides valuable information on the constraints of the modeled tasks.

In particular, the model suggests much of the spatial detail in a shape can be ignored during transformation. The detail needed depends on the task. In a task like mental rotation, where a single transformation is applied to all edges, very little detail is required. In a task like paper-folding, where the results of one rotation determine the axis for the next rotation, more detail must be carried through each transformation. Of course, even in mental rotation more detail will be required as the distractors become more similar to the shapes being rotated (Folk \& Luce, 1987; Yuille \& Steiger, 1982).

These findings are important for understanding spatial ability. Fast, holistic spatial transformations require spatial smoothing. Thus, a key component of spatial ability must be spatial abstraction: the ability to identify and encode critical spatial details while ignoring irrelevant features.

Questions remain about how skilled rotators perform spatial abstraction. The present approach of selecting the four longest edges, while effective, is only one heuristic. Others might include studying the distractors to determine which parts of a shape are most diagnostic (Yuille \& Steiger, 1982) and segmenting shapes into larger-scale parts (Hoffman \& Richards, 1984). In the future, we would like to study a larger stimulus set with distractors that vary in their similarity to the base shapes (Cooper \& Podogny, 1976; Folk \& Luce, 1987). By evaluating different spatial smoothing heuristics in the model, we can better understand the skills supporting spatial abstraction and spatial ability.

## Acknowledgments

This work was supported by NSF SLC Grant SBE-0541957, the Spatial Intelligence and Learning Center (SILC).

## References

Bethell-Fox, C. E., \& Shepard, R. N. (1988). Mental rotation: Effects of stimulus complexity and familiarity. Journal of Experimental Psychology: Human Perception \& Performance, 14(1), 12-23.
Cooper, L., \& Podgorny, P. (1976). Mental transformations and visual comparison processes: Effects of complexity and similarity. Journal of Experimental Psychology: Human Perception \& Performance, 2(4), 503-514.
Ekstrom, R. B., French, J. W., Harman, H. H., \& Derman, D. (1976). Manual for Kit of Factor-Referenced Cognitive Tests. Princeton, NJ: Educational Testing Services.
Falkenhainer, B., Forbus, K., \& Gentner, D. (1989). The structure mapping engine: Algorithm and examples. Artificial Intelligence, 41, 1-63.
Folk, M. D., \& Luce, R. D. (1987). Effects of stimulus complexity on mental rotation rate of polygons. Journal of Experimental Psychology: Human Perception \& Performance, 13(3), 395-404.
Forbus, K., Nielsen, P., \& Faltings, B. (1991). Qualitative spatial reasoning: The CLOCK project. Artificial Intelligence, 51(1-3), 417-471.
Forbus, K., Usher, J., Lovett, A., Lockwood, K., \& Wetzel, J. (2011). CogSketch: Sketch understanding for cognitive science research and for education. TopiCS, 3(4), 648-666.
Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. Cognitive Science, 7, 155-170.
Heil, M., \& Jansen-Osmann, P. (2008). Sex differences in mental rotation with polygons of different complexity: Do men utilize holistic processes whereas women prefer piecemeal ones? The Quarterly Journal of Experimental Psychology, 61(5), 683-689.
Hoffman, D. D., \& Richards, W. A. (1984). Parts of recognition. Cognition, 8(1-3), 65-96.
Jolicoeur, P., Regehr, S., Smith, L., \& Smith G. N. (1985). Mental rotation of representations of two-dimensional and three-dimensional objects. Canadian Journal of Psychology, 39(1), 100-129.
Kosslyn, S. M., Koenig, O., Barrett, A., Cave, C. B., Tang, J., \& Gabrieli, J. D. E. (1989). Evidence for two types of
spatial representations: Hemispheric specialization for categorical and coordinate relations. Journal of Experimental Psychology: $H$ P \& P, 15, 723-735.
Lovett, A., \& Forbus, K. (2012). Modeling multiple strategies for solving geometric analogy problems. Proceedings of CogSci '12.
Lovett, A., \& Forbus, K. (2011). Cultural commonalities and differences in spatial problem-solving: A computational analysis. Cognition, 121 (2), 281-287.
Lovett, A., Forbus, K., \& Usher, J. (2010). A structuremapping model of Raven's Progressive Matrices. Proceedings of CogSci '10.
Lovett, A., Kandaswamy, S., McLure, M., \& Forbus, K. (2012). Evaluating qualitative models of shape representation. Proceedings of $Q R$ '12.
Marr, D., \& Nishihara, H. K. (1978). Representation and recognition of the spatial organization of threedimensional shapes. Philosophical Transactions of the Royal Society of London, Series B, Biological Science, 200(1140), 269-294.
McGee, M. G. (1979). Human spatial abilities: Psychometric studies and environmental, genetic, hormonal, and neurological influences. Psychological Bulletin, 86(5), 889-918.
Palmer, S. E. (1977). Hierarchical structure in perceptual representation. Cognitive Psychology, 9, 441-474.
Schultheis, H., \& Barkowsky, T. (2011). Casimir: An architecture for mental spatial knowledge processing. Topics in Cognitive Science, 3(4), 778-795.
Shea, D.L., Lubinski, D., \& Benbow, C.P. (2001). Importance of assessing spatial ability in intellectually talented young adolescents: A 20-year longitudinal study. Journal of Educational Psychology, 93, 604-614.
Shepard, R. S., \& Cooper, L. A. (1982). Mental Images and their Transformations. Cambridge, MA: MIT Press.
Shepard, R. N., \& Feng, C. (1972). A chronometric study of mental paper folding. Cognitive Psychology, 3, 228-243.
Shepard, S., \& Metzler, J. (1988). Mental rotation: Effects of dimensionality of objects and type of task. Journal of Experimental Psychology: Human Perception \& Performance, 14(1): 3-11.
Shepard, R. N., \& Metzler, J. (1971). Mental rotation of three-dimensional objects. Science, 171(3972), 701-703.
Ullman, S. (1984). Visual routines. Cognition, 18, 97-159.
Vandenberg, S. G., \& Kuse, A. R. (1978). Mental rotations: A group test of three-dimensional spatial visualization. Perceptual and Motor Skills, 47, 599-601.
Wai, J., Lubinski, D., \& Benbow, C. (2009). Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. Journal of Educational Psychology, 101(4), 817-835.
Wright, R., Thompson, W. L., Ganis, G., Newcombe, N. S., \& Kosslyn, S. M. (2008). Training generalized spatial skills. Psychonomic Bulletin \& Review, 15(4), 763-771.
Yuille, J. C., \& Steiger, J. H. (1982). Nonholistic processing in mental rotation: Some suggestive evidence. Perception \& Psychophysics, 31(3), 201-209.

# Mnemonic Diffusion: An Agent-Based Modeling Investigation of Collective Memory 

Christian C. Luhmann (christian.luhmann@stonybrook.edu)<br>Department of Psychology, Stony Brook University<br>Stony Brook, NY 11790 USA

Suparna Rajaram (suparna.rajaram@stonybrook.edu)
Department of Psychology, Stony Brook University
Stony Brook, NY 11790 USA


#### Abstract

Past research has uncovered evidence of social influences on a wide variety of behaviors. Everything from our choice of clothing to smoking appears to be shaped by the people we know. However, little is known about the mechanisms that underlie these influences. Here, we report a series of agentbased simulations demonstrating that information diffuses across social networks in much the same way that behavior diffuses. These findings lead us to conclude that many previously observed social influences on behavior likely rely on a substrate of information transmission and representation.


Keywords: learning, memory, collaboration, social network, agent-based modeling

## Diffusion of Behavior

The idea of social influence has long been a topic of fascination for both scientists and the general public (Bartlett, 1932; Cialdini, 2001; Gladwell, 2000; Schelling, 2006). The general concept of social influence is an intuitive one. For instance, peer pressure is a factor in adolescents' tendency to drink, smoke, and engage in sexual behavior and some individuals slavishly follow the latest fashion trends, mimicking the styles seen on the runway or worn by celebrities. However, the intuition about social influence is far too narrow. The examples cited above are seen as exceptions; perhaps the susceptibility is restricted to a particularly impressionable population (e.g. adolescents) or perhaps influence is seen for relatively trivial behaviors (e.g., the clothes you wear). Such mindless imitation would not be seen, intuition suggests, concerning behaviors that are both deeply personal and of great consequence (e.g., suicide, how many children to have). Surprisingly, research has challenged this intuition, finding that social ties strongly influence a wide range of behaviors, including transformative life decisions (Christakis \& Fowler, 2009; Watts, 2003). For example, a program of research by Christakis and Fowler (2007) has revealed the surprising "contagion" of health-related attributes such as obesity and smoking.

In psychology, there is a long history of work exploring the influence of social context on behavior. Early work focused on the potentially deleterious behavioral effects of social influence. For example, Milgram (Milgram, Bickman, \& Berkowitz, 1969) examined social influences on arbitrary
behavior as a function of group size. In this study, groups of between one and 15 confederates stood on a New York street corner looking upward at a window overhead. Passing pedestrians were likely to mimic some aspect of the observed behavior (e.g., looking up) and this tendency increased with group size. The classic studies of Asch (e.g., Asch, 1951) demonstrate the power of social influence even more starkly because his participants were asked about matters of objective fact (e.g., the length of lines). Despite being accurate when making judgments individually, participants placed with confederates tended to conform, producing substantial errors.

Though past work has revealed the presence of social influences on a variety of behaviors, we know little about the mechanisms that underlie these influences. For example, it has been suggested that, "Social networks function...by giving us access to what flows within them" (Christakis \& Fowler, 2009, p. 91). But what does flow within our social networks that allows for these powerful influences on our behavior?

## Existing Models of Social Influence

Several mathematical models of social influence have been proposed (Easley \& Kleinberg, 2010; Jackson, 2008; Lopez-Pintado \& Watts, 2008). Among the most influential are linear threshold models (Granovetter, 1978). Such models assume that each individual has two mutually exclusive and exhaustive behavioral options available. For example, in Granovetter's classic example, each individual chooses whether or not to join a riot. In addition, each individual is assumed to observe the behavior of all other individuals. The decision of the individual is then a function of their own idiosyncratic threshold and the behavior observed in the group. If the number of other people observed to be rioting does not exceed the individual's threshold, she remains a bystander. If this number exceeds the individual's threshold, she begins to riot.

Several key details of the current crop of mathematical models should be noted. First, this work typically assumes "zero-intelligence agents" (Gode \& Sunder, 1993) that can do nothing but copy the behavior of their neighbors with some probability (e.g., Granovetter, 1978). This is undoubtedly convenient, but represents a substantial simplification, at least when attempting to model human behavior. Though some behaviors may be the result of
innate imitative mechanisms, many more behaviors are deliberative, relying on individuals' beliefs, goals, and desires. Second, many of the computational models of behavioral diffusion assume that individuals can occupy one of a small number of behavioral "states". For example, in the rioting example discussed above, individuals are assumed to either be rioting or not rioting, another vast over-simplification. Finally, because of these simplifications, previous models have largely avoided questions about the mechanisms by which behaviors are transferred between individuals. Indeed, these models expressly omit such mechanisms by assuming that mimicry is the critical basis of diffusion. To address the mechanisms themselves, the current study takes an agent-based modeling (ABM) approach (Smith \& Conrey, 2007; Stasser, 1988; Carley, Martin, \& Hirshman, 2009; Parunak, Belding, Hilscher, \& Brueckner, 2009; Coman, Kolling, Lewis, \& Hirst, 2012) in which agents are information processing units capable of representing information and learning.

The current study investigates the representation and transmission of information within social networks as fundamental mechanisms underlying these potent influences on our behavior. Specifically, we investigate how information is represented by individuals within a larger network and how the nature of social interactions shape the information as it flows through the network.

## Collaborative Memory in Small Groups

Recent behavioral work on the social transmission of memory in small groups has identified several key mechanisms that facilitate or inhibit information transmission in small groups, and how the interaction among these mechanisms shapes convergence amongst group members (what is referred to as collective memory). The collaborative memory paradigm provides a robust method for measuring the transmission of information in small groups of two or three participants (Rajaram, 2011). In this paradigm, each participant is first exposed to experimenter-provided stimuli (words, pictures, narratives). Participants then form groups and recall items collaboratively. Finally, participants recall items individually to assess the post-collaborative representations retained by each participant.

The consequence of collaboration on group memory is counterintuitive. Though a collaborating group recalls more than a given individual, the group recalls significantly less than its potential, a phenomenon called collaborative inhibition (Weldon \& Bellinger, 1997). To estimate the group's potential, performance is compared to that of a nominal group: the total, nonredundant recall of an equal number of participants who recalled individually (Blumen \& Rajaram, 2008; Weldon \& Bellinger, 1997). Although it seems reasonable to assume participants perform suboptimally because they feel less accountable while working in groups (social loafing; Latane, Williams, \& Harkins, 1979), experimental findings shows this is not the case (Weldon, Blair, \& Huebsch, 2000).

## Mechanisms Involved During Collaboration

The suboptimal performance of collaborative groups has been attributed to the retrieval disruption process where the output of one participant's recall disrupts other participants' attempts at recall, and as a result lowers the latter participants' output (B.H. Basden, Basden, Bryner, \& Thomas, 1997). Because each individual recalls less than her potential during collaboration, researchers have asked whether their post-collaborative representations would continue to exhibit this deficit. Though some forgetting does occur (Cuc, Koppel, \& Hirst, 2007) two mechanisms usually enhance the quantity and accuracy of postcollaborative representations; one, items not recalled during collaboration bounce back post-collaboratively (rebound) and, two, collaboration acts to expose each participant to items she might not have remembered otherwise (reexposure, Blumen \& Rajaram, 2008, 2009; Congleton \& Rajaram, 2011).
Several individual- and interaction-based properties influence these collaborative effects. One such change of note relates to increase in memory errors. As one example, social contagion errors arise when the stimuli activate plausible items for recall that were never presented (B.H. Basden et al., 2002; French, Gary,\&Mori, 2008; Reysen, 2007; Roediger, Meade, \& Bergman, 2001). Such contagion has been demonstrated in collaborative studies (B.H. Basden et al., 2002) using DRM stimuli (Roediger \& McDermott, 1995) in which a list of associatively-related words such as bed, rest, awake, tired, dream, wake, snooze, blanket, etc. leads participants to recall the never-presented lure (sleep) with great confidence. Propagation of memory errors in the real-world has been an enduring concern of cognitive scientists (e.g., Bartlett, 1932) but empirical investigations have remained elusive due to feasibility.
Memory representations in small groups are also characterized by the frequency with which information is processed before and during collaboration. For instance, the individual who dominates the collaborative discussion benefits most from rehearsal (Rajaram \& Pereira-Pasarin, 2010) and has the largest influence on the post-collaborative representations of other group members (Cuc, Ozuru, Manier, \& Hirst, 2006). Conversely, post-collaborative memory deficits occur for information not discussed during collaboration, either through omission (Cuc et al., 2007), rejection of correct responses (Merckelbach, van Roermund, \& Candel, 2007), or group conformity to incorrect responses (Reysen, 2005). We have further shown that frequency of discussion prior to or during collaboration changes both collaborative group recall and post-collaborative memory; when information is repeatedly processed prior to collaboration, it can reduce or even eliminate collaborative inhibition in group recall (Congleton \& Rajaram, 2011; Pereira-Pasarin \& Rajaram, 2011), and improve postcollaborative memory (Congleton \& Rajaram, 2011). Just as interestingly, when groups are given the opportunity to discuss more frequently this too reduces collaborative inhibition in group recall (B.H. Basden et al., 2000; Blumen
\& Rajaram, 2008). These behavioral outcomes raise intriguing questions about how frequency of discussion influences group-level representations in large social networks.

Yet another intriguing finding concerns the effects of group size. Even within small groups, research shows that as group size increases (from 2 to 3 or 4 members) collaborative inhibition increases with group size (B.H. Basden et al., 2000; Thorley \& Dewhurst, 2007). This raises the question about whether larger social networks would display an exaggerated version of this decline or whether the complex interplay of mechanisms would completely change how the group-level representation evolves.

## Current Approach

In the current study we investigate the processes that shape the transmission of information during both the collaborative remembering in the laboratory paradigm and more realistic social contexts. We take an agent-based modeling approach in which individuals are represented by computational agents and allowed to interact much as human subjects interact in the collaborative memory paradigm. The agents are endowed with simplified memory models capable of storing a set of $N$ items (e.g., words). The memory model consists of two separate representations. First, agents represent a set of inter-item associations that exist prior to any social interaction (a matrix denoted $S$ ). These associations represent pre-experimental knowledge such as the semantic associations between words. For the sake of simplicity, these inter-item associations were assigned random values (within the range $[0,1]$ ) in the current simulations. More systematic prior knowledge could obviously be constructed, particularly if such factors were important for specific research questions. For example, lists of categorized words can be simulated by constructing high within-category associations and low between-category associations, a strategy we have successfully used in recent modeling (Luhmann, Congleton, Zhou, \& Rajaram, 2013). The second representation is a set of $N$ activations (a vector denoted $A$, with elements bound to the range [ 0,1$]$ ), which allow for learning to occur during the experimental experience itself. For example, these activations capture recent experience studying experimenter-provided stimuli (e.g., word lists), items generated by collaborative partners, and even items generated by the agent itself.

Agents have two behaviors. First, they may encode a presented item by increasing the activation associated with the presented item (i.e., $\Delta A_{i}=\alpha\left[1-A_{i}\right]$, where $A_{i}$ is the activation of the item and $\alpha$ is a learning rate). This encoding occurs when items are presented by the experimenter (i.e., during the collaborative memory paradigm's initial study phase) and when agents are exposed to the items retrieved by other agents (e.g., during collaborative recall). Second, agents can retrieve an item. This is done by randomly selecting an item in proportion to the activation levels in $A$ (i.e., more active items are more likely to be generated). If the activity of the candidate item
is above the agent's recall threshold ( $\gamma$ ) and has not yet been generated by the group, then this item is successfully retrieved and generated (e.g., spoken out loud). Finally, associates of the retrieved item (from $S$ ) have their activations decreased (i.e., $\Delta A_{j}=-\beta S_{i j} A_{i}$ where $A_{i}$ is the retrieved item, $S_{i j}$ is the strength of the association between items $i$ and $j$, and $\beta$ is a forgetting rate).


Figure 1 - Simulations replicate the collaborative inhibition effect

## Simulation 1: Collaborative Inhibition

Our first investigation is of the most surprising finding to come out of the collaborative memory paradigm: collaborative inhibition. This was done for two reasons. First, the finding is elicited from a fairly simple experimental paradigm making these initial simulations relatively straightforward to construct. Second, to the degree we are capable of successfully replicating the least intuitive aspect of the empirical data, we can proceed with somewhat more confidence that our formalism has not overly simplified the cognitive processes involved.

To simulate the collaborative memory paradigm, we first presented the entire list of $N$ items to each agent in a random order. Three agents were then allowed to interact with one another. The interaction was structured such that each agent was given an opportunity to retrieve an item on each round. If an agent successfully retrieved an item, the retrieved item was encoded by the other two agents. Figure 1 illustrates the results of 1000 simulations of a collaborative condition and 1000 simulations of a nominal condition (i.e., total, nonredundant recall of three agents recalling individually) evaluated exactly as in the behavioral studies described above. As can be seen, the simulation results reproduce the collaborative inhibition findings describe above. This result is likely due to the fact that each agent is endowed with an idiosyncratic set of activations during the initial, individual study phase but then learns the contents of their peer's activations during the collaborative phase. Thus, the interaction between agents tends to increase the similarity of the agents' representations and minimize the idiosyncrasies that make the nominal groups more successful in generating greater quantity. Furthermore, exploratory simulations suggest that having individual agents repeatedly engage in isolated retrieval does not diminish the performance of the nominal group, suggesting that the collaborative inhibition is due to the social interaction (e.g., retrieval disruption, B.H. Basden et al., 1997) rather than repeated retrieval.


Figure 2 - Influence of group size on collaborative inhibition

## Simulation 2: Group Size

As mentioned above, prior work with the collaborative memory paradigm has manipulated a variety of different factors. One factor that has received surprisingly little attention is the size of the collaborating group despite its obvious real-world relevance. Only two studies have manipulated group size (B.H. Basden, Basden, and Henry, 2000; Thorley and Dewhurst, 2007), and both concluded that increasing group sizes produce more detrimental collaborative effects. Details of these studies limit interpretation, however. For example, Basden et al. (2000) tested 1-, 2-, and 4 -person groups. Though the 1-person groups recalled more than the 4-person groups, the 2-person groups were different from neither. Thorley and Dewhurst (2007) used DRM stimuli and were specifically interested in groups' tendency to falsely recall the lure items (rather than recall per se). Further, the group size tested was small in these studies (2-4 person groups) again limiting interpretation.

Here, we systematically manipulate group size and investigate the influence of this factor on collaborative inhibition. We simulated collaborative groups ranging in size from two to seven as well as nominal groups consisting of agents retrieving in isolation. As before, the entire list of $N$ items was first presented to each agent. The agents within a group were then allowed to interact with one another with the interaction structured as described above.

Figure 2 illustrates the results of 1000 simulations for each group size. The standard collaborative inhibition effect (measured here as Nominal - Collaborative) was found for all group sizes. However, the relationship between group size and collaborative inhibition was not entirely straightforward. As groups grew from two to four, collaborative inhibition increased (replicating Basden et al., 2000). However, as group size increased further, collaborative inhibition decreased. This non-monotonic relationship appears to be driven by the relative balance between the facilitative effects offered by collaboration (i.e., more agents increase the probability that the group will retrieve a given item) and the detrimental effects of retrieval disruption (i.e., more collaborators means more opportunities to be disrupted).

## Simulation 3: Diffusion of Collective Memory

The collaborative memory paradigms represent a realistic, real-world social network that is amenable to experimental study. However, the size of groups involved in this paradigm places obvious restrictions on the research questions that may be asked. The current simulation seeks to achieve substantially greater realism than the more traditional laboratory paradigms allow. Specifically, we wish to explore how the information represented by and shared between individuals makes its way through larger social networks.

To explore true social networks, we employed a larger population (60 computational agents of the kind described above), each of which was placed into a larger network structure. Though there are many potentially interesting network structures, we are most interested in those related to real world social networks. For this reason, the current simulation employs a so-called small world network (Watts \& Strogatz, 1998). Within such a network, the shortest distance between two nodes is short on average despite the network itself being relatively sparse (most nodes are not neighbors). These features give rise to the well-known "six degrees of separation" phenomenon. We further chose to set the average degree to 2 (meaning that agents were, on average, connected to 2 other agents).

As in the simulations reported above, each simulation began by presenting the entire list of 40 items to each agent individually. All subsequent interaction between agents occurred over this network. On each epoch of the simulation, a random agent was selected along with one of that agent's randomly selected neighbors. This pair was then allowed to interact just as the collaborative groups simulated above (each taking a turn to retrieve, etc.). To assess the diffusion of information across the network, we computed the similarities between pairs of agents' representations (i.e., the correlation between activation vectors, A) after the simulations were completed. This measure of similarity goes up with the overlap (both in what they represent strongly and what they have forgotten, i.e., collective memory) and goes down when one agent has forgotten an item that the other agent still remembers (i.e., is strongly active in $A$ ). We computed the similarity between all pairs of agents (i.e., both neighboring pairs and nonneighboring pairs) sorting these similarities on the basis of how close the two agents in each pair were to each other within the network (i.e., minimum distance). Neighboring agents would have a distance of 1 . Two non-neighboring agents that shared a common neighbor would have a distance of 2 and so on.

Figure 3 illustrates the results of a 60 -node small world network that was allowed to run for 1000 epochs. As can be seen, neighboring agents acquired very similar representations. This is not particularly surprising since neighboring agents will have interacted with each other and learned the contents of their neighbors' representations. What is surprising is that agents at a distance of two are highly similar as well. These agents never interacted with
one another, so direct communication cannot explain this similarity. Instead, the two agents' common neighbor presumably acted as a conduit through which information diffused, indirectly connecting the non-neighbors. Even more surprising then, is the fact that agents at distance three, separated by two intermediate agents, are also somewhat similar. After this point, the similarity between agents levels off, reflecting the boundaries of collective memory in large networks.

This similarity between non-neighboring agents is a phenomenon that has been observed in a variety of realworld social networks and is known as hyperdyadic spread (Christakis \& Fowler, 2009). For example, previous work has shown that people are 57\% more likely to become obese if a peer (e.g., friend) becomes obese and $20 \%$ more likely to become obese if a peer of a peer (e.g., a friend of a friend) becomes obese. Furthermore, in many of the behaviors studied within social networks, hyperdyadic spread from a given node in the social network has been found to extend to three "hops" from that node (e.g., to the friend of a friend of a friend) but not beyond - what Christakis and Fowler (2009) have termed the three degrees of influence rule. The fact that our simulations comply with this rule is interesting because the standard finding of hyperdyadic spread concerns the spread of behavior whereas the current results reflect the spread of information. Exploratory simulations employing other network structures (e.g., chains, trees) have either failed to uncover strong hyperdyadic spreading or failed to conform to the three degrees of influence rule. This suggests that this class of phenomena may be jointly driven by both the details of the social networks in which we live (e.g., small-world networks) and the constraints of human learning and memory.


Figure 3 - Hyperdyadic spread of collective memory in our simulation results

## Discussion

Broadly speaking, the goal of the current study has been to investigate social influence in real world social networks. Prior research has developed formal models to capture how behaviors diffuse amongst large groups, but these formalisms have been relatively agnostic about the underlying psychological mechanisms, instead modeling such behaviors as being literally contagious. Here, we have argued that the cognitive processes governing learning and
memory are likely candidates for such mechanisms as they are in prime position to influence the representations individuals hold and transmission of information between f

Past work using the collaborative memory paradigm has provided useful empirical data with which we can begin to explore our proposal. In this paradigm, groups of individuals first study items in isolation before collaborating in groups to recall these same items. Despite the practical limitations posed by this paradigm (e.g., small group sizes), the literature has provided a wealth of insights into the social influences on learning and memory. These insights include the role of retrieval disruption, re-exposure, and error correction, the influence of group size, and phenomena such as collaborative inhibition and error propagation.

In the current study, we have taken an agent-based modeling approach, simulating individuals as relatively simple information processing units capable of representing information, learning from experience, and interacting with other agents. In order to evaluate our proposal, we selected three different phenomena to explore. We first investigated the robust collaborative inhibition effect. Our simulations replicate the standard pattern of results, with collaborative groups under-performing relative their controls. We next simulated the somewhat less thoroughly studied role of group size on the collaborative inhibition effect. Here, we found that our simulations were capable of replicating the effects observed in the literature (increasing collaborative inhibition with increasing group size) but also made predictions about the boundary conditions of these effects. Finally, we extended our findings beyond the collaborative memory paradigm to investigate agents in a larger social network. Here, we found that our simulations exhibited hyperdyadic spread, a standard empirical finding in the diffusion of behavior across social networks.

We take the success of the current simulations as evidence in favor of our proposal. Our simulations demonstrate that the spread of information across connections in a social network mirrors the way in which behavior spreads across those same connections. Thus, it seems likely that social influence, and particularly the diffusion of behavior, relies on a substrate of information transmission and representation.

## References

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. In H. Guetzkow (Ed.), Groups, Leadership, and Men (pp. 177-190). Pittsburgh: Carnegie Press.
Bartlett, Sir F. C. (1932). Remembering: A Study in Experimental and Social Psychology. London: Cambridge University Press.
Basden, B. H., Basden, D. R., \& Henry, S. (2000). Costs and benefits of collaborative remembering. Applied Cognitive Psychology, 14(6), 497-507.
Basden, B. H., Reysen, M. B., \& Basden, D. R. (2002). Transmitting false memories in social groups. American Journal of Psychology, 115, 211-231.

Basden, B. H., Basden, D. R., Bryner, S., \& Thomas, R. L., III (1997). A comparison of group and individual remembering: Does collaboration disrupt retrieval strategies? Journal of Experimental Psychology: Learning, Memory, and Cognition, 23, 1176-1189.
Blumen, H. M., \& Rajaram, S. (2008). Influence of reexposure and retrieval disruption during group collaboration on later individual recall. Memory, 16, 231244.

Blumen, H. M., \& Rajaram, S. (2009). Effects of repeated collaborative retrieval on individual memory vary as a function of recall versus recognition tasks. Memory, 17, 840-846.
Carley, K. M., Martin, M. K., \& Hirshman, B. R. (2009). The etiology of social change. Topics in Cognitive Science, 1(4), 621-650.
Christakis, N. A., \& Fowler, J. H. (2007). The spread of obesity in a large social network over 32 years. New England Journal of Medicine, 357, 370-379.
Christakis, N. A., \& Fowler, J. H. (2009). Connected: The Surprising Power of our Social Networks and How they Shape our Lives. New York: Little, Brown and Company.
Congleton, A. R. \& Rajaram, S. (2011). The influence of learning methods on collaboration: Prior repeated retrieval enhances retrieval organization, abolishes collaborative inhibition, and promotes post-collaborative memory. Journal of Experimental Psychology: General, 140, 535-551.
Coman, A., Kolling, A., Lewis, A., \& Hirst, W. (2012). Mnemonic convergence: From empirical data to largescale Dynamics, Social Computing, Behavioral-Cultural Modeling and Prediction, 256-265.
Cuc, A., Koppel, J., \& Hirst, W. (2007). Silence is not golden: A case for socially shared retrieval-induced forgetting. Psychological Science, 18, 727-733.
Easley, D., \& Kleinberg, J. (2010). Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge: Cambridge University Press.
French, L., Gary, M., \& Mori, K. (2008). You say tomato? Collaborative remembering leads to more false memories for intimate couples than for strangers. Memory, 16, 262273.

Gladwell, M. (2000). The Tipping Point: How Little Things Can Make a Big Difference. New York: Little, Brown \& Company.
Granovetter, M. (1978). Threshold models of collective behavior. American Journal of Sociology, 83, 1420-1443.
Gode, D. K., \& Sunder, S. (1993). Allocative efficiency of markets with zero-intelligence traders: Market as a partial substitute for individual rationality. Journal of Political Economy, 101, 119-137.
Jackson, M. O. (2008). Social and Economic Networks. Princeton: Princeton University Press.
Latane, B., Williams, K., \& Harkins, S. (1979). Many hands make light the work: The causes and consequences of social loafing. Journal of Personality and Social Psychology, 37, 822-832.

Luhmann, C.C., Congleton, A.C., Zhou, X., \& Rajaram, S. (2013). When less is more: Retrieval-refined representations produce the testing effect. Manuscript submitted for publication.
Merckelbach, H., van Roermund, H., \& Candel, I. (2007). Effects of collaborative recall: Denying true information is as powerful as suggesting misinformation. Psychology, Crime, \& Law, 13, 573-581.
Milgram, S., Bickman, L., \& Berkowitz, L. (1969). Note on the drawing power of crowds of different size. Journal of Personality and Social Psychology, 13, 79-82.
Parunak, H. V., Belding, T. C., Hilscher, R., \& Brueckner, S. (2009). Understanding collective cognitive convergence Multi-Agent-Based Simulation IX (pp. 4659): Springer.

Pereira-Pasarin, L., \& Rajaram, S. (2011). Study repetition and divided attention: Effects of encoding manipulations on collaborative inhibition in group recall. Memory \& Cognition, 39, 968-976.
Rajaram, S. (2011). Collaboration both hurts and helps memory: A cognitive perspective. Current Directions in Psychological Science, 20, 76-81.
Rajaram, S., \& Pereira-Pasarin, L. (2010). Collaborative memory: Cognitive research and theory. Perspectives on Psychological Science, 5, 649-663.
Reysen, M. B. (2005). The effects of social pressure on false memories. Dissertation Abstracts International: Section B: The Sciences and Engineering, 65, 6066.
Roediger, H. L., \& McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. Journal of Experimental Psychology: Learning, Memory, and Cognition, 21, 803-814.
Roediger, H.L., III, Meade, M.L., \& Bergman, E.T. (2001). Social contagion of memory. Psychonomic Bulletin \& Review, 8, 365-371.
Smith, E. R., \& Conrey, F. R. (2007). Agent-based modeling: A new approach for theory building in social psychology. Personality and Social Psychology Review, 11, 87-104.
Stasser, G. (1988). Computer simulation as a research tool: The DISCUSS model of group decision making. Journal of Experimental Social Psychology, 24, 393-422.
Thorley, C. \& Dewhurst, S. A. (2007). Collaborative false recall in the DRM procedure: Effects of group size and group pressure. European Journal of Cognitive Psychology, 19, 867-881.
Watts, D. J., \& Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. Nature, 393, 440-442.
Weldon, M. S., \& Bellinger, K. D. (1997). Collective memory: Collaborative and individual processes in remembering. Journal of Experimental Psychology: Learning, Memory, and Cognition, 23, 1160-1175.
Weldon, M. S., Blair, C., \& Huebsch, D. (2000). Group remembering: Does social loafing underlie collaborative inhibition? Journal of Experimental Psychology: Learning, Memory, and Cognition, 26, 1568-1577.

# The effect of social roles on gaze cue utilisation in a real-world collaboration 

Ross G. Macdonald (rgmacdonald@dundee.ac.uk)<br>School of Psychology, University of Dundee, Nethergate, Dundee, DD1 4HN, UK

Benjamin W. Tatler (b.w.tatler@activevisionlab.org)<br>School of Psychology, University of Dundee, Nethergate, Dundee, DD1 4HN, UK


#### Abstract

During collaboration, people communicate using verbal and non-verbal cues, including gaze cues. Social factors can affect gaze allocation, however most research on gaze cueing has not considered these factors. The presence of social roles was manipulated in a collaborative task whilst eye movements were measured. In pairs, participants worked together to make a cake. Half of the pairs were given roles ("Chef" or "Gatherer") and the other half were not. Across all participants we found, contrary to the results of static image experiments, that participants spent very little time looking at each other, challenging the generalisability of the conclusions from lab-based paradigms. When given spoken instructions, listeners in the roles condition looked at the speaker significantly more than listeners in the no roles condition. We conclude that our tendency to seek the gaze cues of collaborators is affected either by our social perceptions of the collaborator or their perceived reliability.


Keywords: eye movements; joint attention; real world; gaze cues; social interaction.

## Introduction

When collaborating with another on a task, we need to communicate. As well as using spoken language, there are a number of non-verbal cues we can use, with the directional gaze cues given by the eyes being the most well-researched of these. Gaze cues are first used very early in life and continue to be given and followed throughout adulthood. People have a tendency to orient to and follow the gaze cues of others and can to do this with ease. However, there is evidence that the language accompanying a gaze cue and the social context of the cue can affect how people orient to and follow gaze cues. In the real world, gaze cues will always occur within a social context, yet this context is removed in most studies. The aim of the present study is to measure eye movements in a real-world setting to observe how the utilisation of gaze cues can be affected by social context in a natural collaboration.

When viewing images of faces, people have a tendency to look at the eyes (Yarbus, 1967) and when viewing images of social scenes people will seek out faces and eyes (Birmingham, Bischoff \& Kingstone, 2007; 2009) even when the person being fixated is not visually prominent and has no role for understanding the scene (Zwickel \& Võ, 2010). As well as orienting to these cues, people show a tendency to follow them. Friesen and Kingstone (1998) showed that incongruent gaze cues presented at fixation
could slow down responses in a Posner (1980) task, suggesting that the artificial gaze cue stimuli automatically shifted attention away from the target. Variants of this study looking at eye movements have found that participants will also look in the direction of the distracting gaze cue, even though they know there is no reason to do so (Ricciardelli, Bricolo, Aglioti, \& Chelazzi, 2002; Galfano et al, 2012). These findings have been used to suggest that humans are "hard-wired" to automatically follow the gaze cues of others (Emery, 2000).

The above research shows that people look at eyes and follow gaze cues when viewing isolated static images of others. However, in the real world, gaze cues usually occur alongside spoken language. There appears to be an intimate link between gaze allocation and spoken language, with people making anticipatory eye movements to objects that relate to what they hear (Altmann \& Kamide, 1999). Gaze cue utilisation in particular has been shown to be affected by spoken language; changing the syntactic structure of a sentence, whilst maintaining meaning changes the timing of gaze following (Knoeferle \& Kreysa, 2012). Reciprocally, Stuadte and Crocker (2011) showed the gaze cues can affect the understanding of spoken language; participants were shown videos of a robot describing the spatial and featural relations between a series of visible objects, whilst providing gaze cues. The robot made mistakes in his descriptions that could have been corrected in two different ways. The experimenters found that participants would correct in the way that was congruent with the gaze cue, suggesting that they were inferring meaning from the robot's gaze and assuming that the robot meant to refer to the object that it was gazing at. Given the effect gaze cues and language have on each other, it is important to use language in a paradigm investigating how gaze cues are used naturally in collaboration.

As well as mostly occurring alongside language, all gaze cues in the real world are provided in a social context. When interacting with another, where we look can be affected by our proximity to this other person (Argyle \& Dean, 1965). Social effects specifically on gaze seeking were investigated by Laidlaw, Foulsham, Kuhn and Kingstone (2011), who found that participants sitting in a waiting room were significantly more likely to look at a person on a monitor than the same person present in the room. Gallup et al (2012) found similar results for gaze following rather than seeking. They observed people walking past an attractive
item in a hallway and found that people were more likely to look in the same direction as somebody walking in front of them than somebody walking towards them. The results of these studies were explained by their respective authors as being due to participants trying to avoid potential interactions with strangers, which might be triggered by any gaze seeking or following behaviour detected by the oncoming person. These findings indicate that social factors can affect the way we utilise the gaze cues of strangers, which suggests that social context may have an effect on gaze utilisation in one-to-one interactions and collaborations.

Macdonald and Tatler (2013) considered gaze seeking and following behaviour in a real world communicative task, involving one-to-one interaction between an instructor (the experimenter) and a participant. The instructor manipulated his use of gaze as well as the specificity of his instructions in a simple block-building task. Participants were found to only seek and follow gaze cues when the language was ambiguous (it did not specify which single block the participant was meant to pick up), suggesting that gaze cues are used flexibly, depending on other information that is available. It was also noted that even when gaze cues supplied the only unambiguous information about which block to pick up (because the spoken instructions were ambiguous) participants did not seek and follow these all of the time. It was speculated (Macdonald \& Tatler, 2013) that social factors may have played a part in these results. More specifically, the social cost of looking at the instructor frequently in each trial may have deterred participants from seeking and following these gaze cues. Although this is speculation, these results make a case for manipulating social factors in a real-world gaze-cueing experiment.

One way to manipulate social factors in a gaze cueing task is to manipulate what the participant knows about the entity with which they are interacting. Participants carrying out a Posner (1980) task in Italy were shown distracter gaze cueing stimuli made from the faces of Italian political figures, including Silvio Berlusconi (Liuzza et al, 2011). The gaze of Berlusconi was found to cause significantly more interference in the task for right-wing voters (ingroup) than left wing voters (out-group). These results suggest that people may be more prone to following the gaze cues of others with shared beliefs. Crosby, Monin and Richardson (2008) showed that participants were more likely to look at an individual on a monitor if they thought the individual could hear comments that were potentially offensive to that individual. These results show that social factors such as beliefs about another individual can affect how others look at them as well as how others look at external objects whilst communicating with them. Although, these results show effects of prior beliefs about others on gaze behaviour, it is still unclear how beliefs about the role or knowledge of another affect the use of gaze cues in natural collaboration.

The present study manipulates participants' perception of their collaborator by assigning them roles in a task.

Participants, in pairs, were given a recipe to follow in order to make the batter for a cake. During this collaboration their eye movements were recorded using portable eye-trackers. When coding the data we were particularly interested in the time participants spent looking at each other (interpersonal gaze) or at the same object simultaneously (mutual gaze). Half of the pairs were given roles (chef or gatherer) to fulfil and the other half were not. By manipulating this we are able to investigate whether the perception of another's role in collaboration has any significant effect on the extent to which we seek and follow their gaze cues in a real-world interaction.

## Methods

## Participants

Twenty-four students from the University of Dundee participated in this experiment. They were split into twelve pairs to carry out the task. Six pairs were allocated to the roles condition and six were allocated to the no roles condition (see design).

## Materials

The experiment took place in a kitchen area on the University of Dundee campus. The kitchen was fully equipped with standard kitchen appliances, but only the oven and microwave were used. All items and foodstuffs that could be removed were removed before testing and the experimental materials were arranged carefully around the kitchen. This included the items and foodstuffs that were to be used for the procedure as well as a selection of distractor items. All of these items were placed in the same location for each pair of participants. A Recipe Procedure sheet was provided for each pair. This sheet explained, step-by-step, how to make the batter for a Victoria Sponge. There was also a Chef Guidelines sheet and a Gatherer Guidelines sheet for those in the roles condition. These sheets explained the responsibilities and duties for participants in the chef and gatherer roles.

## Design

This experiment had a between subjects design. The two independent variables for the analysis of mutual fixations and time participants spent looking at each other (interpersonal gaze) were the use of roles (roles or no roles) and the allocation of roles within the roles condition (chef or gatherer). For the analysis of the instruction statements the independent variables were the use of roles (roles or no roles) and the identity of participant (speaker or listener).

## Procedure

This experiment required two participants. The experimenter began by fitting a portable eye tracker to the first participant. At this point in the roles condition the first participant was given the Chef Guidelines and the second participant was given the Gatherer Guidelines. They were both instructed to read over their sheet and make sure they understood their roles. The Chef Guidelines informed the chef that they were in charge of preparing the recipe and that the gatherer was there to assist them. The sheet explained that the chef was expected to mix and prepare ingredients, following a recipe which they could not show to the gatherer. The chef would not be expected to collect any items or foodstuffs, but to delegate those duties to the gatherer. The chef would also be able to ask the gatherer to assist them with any aspect of the preparation they wished. The Gatherer Guidelines explained that the gatherer would not be expected to make any decisions concerning the preparation, but should instead do as instructed by the chef. Once the participants declared they understood their roles the gatherer was asked to remain outside whilst the experimenter and the chef entered the kitchen. The experimenter then gave the chef the Recipe Procedure sheet and told the chef where all of the necessary items and foodstuffs were located. The chef was then told they would have approximately three minutes to familiarise themselves with the kitchen and the locations of the items. During these three minutes the experimenter fitted another portable eyetracker to the gatherer. In the no roles condition the second eye-tracker was fitted straight after the first. At this point, in both conditions, both participants were brought into the kitchen and the eye-trackers were switched on.

The cameras were synchronised and the eye-trackers calibrated. Once calibration was complete, those in the no roles condition were directed to the Recipe Procedure sheet and informed that all of the items they would require were located around the kitchen. All participants were informed that the experimenter would be standing outside the kitchen, out of sight and that the participants must make no attempt to interact with him. The experimenter then told the participants that they may begin as soon as he was out of the room. The experimenter left and the procedure began. The procedure ended when the participants put the batter mixture in the oven.

## Eye movement and sound recording

Participants' eye movements were tracked using two Positive Science LLC mobile eye trackers, which allowed free head movement. Each eye tracker has two cameras mounted on the frame of a pair of spectacles: one records the scene from the participant's point of view and the other records the right eye. Data from these cameras were captured on digital camcorders. For one of the eye-trackers these camcorders were stored, alongside a power supply for the eye-tracker, in a lumbar pack worn by the participant. The camcorders connected to the second tracker were again
stored alongside a power supply, but were stored in a light backpack worn by the participant. This eye tracker also has a small microphone attached to the frame. This microphone recorded sound throughout the experiment and was able to pick-up the voices of both participants. Gaze direction was estimated off-line using Yarbus software provided by Positive Science, LLC, which tracks the pupil and corneal reflection. Calibration was carried out in two stages, one looking down at a counter and the other looking across the room. These two stages were used because by tracking one eye we are not able to directly measure the vergence of the eyes that occurs as participants focus on objects at different distances. Instead we fit the model to fixations on both proximal and distal points. If the tracker estimates in the scene video fell on the correct calibration positions the calibration was deemed adequate. Eye movement data were recorded at 30 Hz with a spatial accuracy of about 1 degree. Once videos for both participants were rendered with the eye movement information, Quicktime Pro was used to synchronise both videos in to one movie file, ready for analysis.

## Analysis

Eye tracking data were coded manually offline using Quicktime Media player and audio information was extracted using Audacity sound editing software. The first two dependent variables considered were (1) the proportion of time both participants fixated the same object (mutual fixations) and (2) the proportion of time a participant spent looking at their partner (interpersonal gaze). For these analyses, in each pair, one participant was labelled person A and the other was labelled person B . In the roles condition person A was the chef and B was the gatherer. Since there were not any defined roles in the no roles condition, participants in this condition were arbitrarily allocated as person A or B. The frame-by-frame coding of these data was split between the lead experimenter and three undergraduate volunteers from the School of Psychology. To begin, all four coders coded the same movie file and these were all compared by the lead experimenter to ensure a consistent and high quality of coding. Mutual fixations were compared across conditions by a t-test and the proportion of time spent on interpersonal gaze was analysed using a 2 (roles or no roles) by 2 (person A or person B) independent measures ANOVA.

The individual instructions were also coded and analysed. These were coded by the lead experimenter alone, using audacity sound editing software and the Quicktime movie files. For each pair, each instruction statement was numbered and transcribed, noting the speaker. The time that the speaker first looked (if at all) at the listener and vice versa was coded for each instruction statement. In the roles condition, the speaker was always the chef and the listener always the gatherer. In the no roles condition the participant who gave the instruction was considered to be the speaker. Therefore the identity of the speaker and listener would
switch throughout each movie in the no roles condition. From coding these data we considered the percentage of instructions in which the participant looked at the other participant. This was analysed using a 2 (role or no role) by 2 (speaker or listener) ANOVA.

## Results

## Overview of eye movements in collaboration

The first set of results is focused on the general eye movement behaviour of participants in the roles and no roles conditions. To investigate this behaviour we measured the proportion of time participants spent mutually fixating objects and the proportion of time spent fixating on the coparticipant (interpersonal gaze).

The mean proportion of time in which both participants fixated on the same item (mutual fixation) is shown in Figure 1.


Figure 1: The mean percentage of time in which mutual fixation occurred for participant pairs in the roles and no roles condition (with standard error bars).

A larger mean percentage of time was spent on mutual fixation in the roles condition (27.23\%) than the no roles condition (20.69\%). However this difference was not found to be significant $(\mathrm{t}(10)=1.37, \mathrm{p}=0.200)$

The mean percentage of time spent engaged in interpersonal gaze is shown in Figure 2. This plot shows the percentage of time that A spends looking at $B$ and vice versa for the roles and no roles conditions. The amount of time when participants A and B simultaneously looked at each other is also shown in Figure 2, for the roles and no roles conditions.


Figure 2: The mean percentage of time participants spent on interpersonal gaze for Person A, Person B and A and B simultaneously for both the roles and no roles conditions
(with standard error bars)
It can be seen from Figure 2 that on average person B spent more time looking at person A (3.77\%) than viceversa (1.92\%) in the roles condition, whilst participants spent only $0.43 \%$ of the total time on simultaneous interpersonal gaze. In the no roles condition, Person A was found to spend slightly more time looking at person B ( $2.62 \%$ ) than vice versa ( $2.31 \%$ ) and only $0.27 \%$ of the time was spent simultaneously looking at one another. A two (roles, no roles) by two (person A, person B) ANOVA was carried out on these results. No main effects of role condition $(\mathrm{F}(1,20)=0.171, \mathrm{p}=0.683)$ or participant $(\mathrm{F}(1,20)=0.701, \mathrm{p}=0.412)$ were found, nor was there any significant interaction $(F(1,20)=1.381, p=0.254)$.

## Analysis of eye movements during instructions

These results consider the eye movement behaviour during the periods when one of the participants was giving spoken instructions to the other. For the roles conditions the spoken instructions were always provided by the chef. For the no roles conditions, any instructions could have been provided by either participant. We investigated the mean percentage of (spoken) instructions in which interpersonal gaze occurred. For each of the roles and no roles conditions, we considered cases when the speaker looked at the listener, the listener looked at the speaker or both speaker and listener looked at each other at the same time (Figure 3).


Figure 3. The mean percentage of instructions in which interpersonal gaze occurred for speakers, listeners and both speakers and listeners in the roles and no roles conditions (with standard error bars).

A two (roles, no roles) by two (speaker, listener) ANOVA showed a main effect of identity of participant (speaker or listener) $(\mathrm{F}(1,20)=12.00, \mathrm{p}=0.002)$. The main effect of role condition was not significant $(\mathrm{F}(1,20)=3.21, \mathrm{p}=$ 0.089 ), however, there was a significant interaction ( $\mathrm{F}(1,20$ ) $=4.92, \mathrm{p}=0.038$ ). Post-hoc t-tests showed that listeners looked at speakers during significantly more instructions in the roles condition (50.20\%) than the no roles condition ( $20.28 \%, p=0.010$ ), but there was no significant difference found between speakers' looks to listeners across the roles ( $7.78 \%$ ) and no roles ( $10.97 \%$ ) conditions ( $p=0.766$ ).

## Discussion

The aim of this study was to investigate the effect of manipulating social context on the utilisation of gaze cues in a real world collaborative social interaction. Using portable eye trackers we were able to measure the eye movements of both collaborators for the duration of the task. The time participants spent looking at each other in this real world paradigm was much less than expected, given the results of experiments using static social scenes (Birmingham et al, 2007; 2009). Social context was actively manipulated in this paradigm by the presence or absence of roles as there is evidence from lab based studies (Crosby et al, 2008; Liuzza et al, 2011) that beliefs about a collaborator can affect gaze behaviour. The amount that listeners looked at speakers during instructions was affected by our manipulation of the roles of the two participants, providing evidence that the tendency to look at another individual during a real world interaction may be influenced by the social context provided by the roles of the individuals. This result is consistent with previous suggestions that gaze seeking and following may depend on the social context of the gaze cues (Gallup et al, 2012; Laidlaw et al, 2011; Macdonald \& Tatler, 2013).

There was no significant difference found between the percentage of time in which mutual fixations occurred in the roles and no roles conditions, with collaborators spending approximately one-quarter of task time mutually fixating on the same objects. There was also no significant difference
between the percentage of time that interpersonal gaze occurred across roles conditions. However, participants spent far less time (between $2-4 \%$ ) looking at each other than they spent mutually fixating other objects. This is notable as it appears to be at odds with the results of some previous lab-based studies. People have been shown to have a preference for looking at eyes when viewing pictures of people (Yarbus, 1967) or social scenes (Birmingham et al, 2009; Zwickel \& Võ, 2010), however in this task participants spent very little time looking at their partners. Given the potential informativeness of the eyes (Tomasello et al, 2007) and the ease with which people can interpret gaze direction (Anderson, Risko \& Kingstone 2011) this finding may seem surprising. However, studies using real people as stimuli may offer an explanation. Laidlaw et al (2011) showed that people were less likely to look at a present confederate than the same confederate on a video monitor and Gallup et al (2012) found that people were less likely to follow the gaze of strangers that could see them than strangers who could not. They concluded that this was due to there being potential consequences (social interaction) to looking at the present confederate or the oncoming stranger. A collaborator in the present study could potentially react to the looks of a participant, whereas the static and video images in lab based paradigms could not. Therefore, these lab-based studies may have over-estimated the tendency of people to look at eyes and faces in social settings.

These results present an obvious question; if people rarely look at each other in an interaction, can they still utilise gaze cues? Although our results cannot lead us to a definite answer, there are three main arguments for the ability to utilise gaze cues in these circumstances. Firstly, it has been shown that gaze cues can be followed and affect language comprehension, even when they are not directly fixated (Knoeferle \& Kreysa, 2012). Secondly, when gaze cues are fixated, the fixations do not necessarily involve long periods of time viewing the eyes. Looks to gaze cues may be very brief, but very informative. Thirdly, it may be the case that eyes are generally not sought out during a task, but are used effectively when required, for example, during instructions.

From our findings it is possible to speculate about the third possibility. Listeners were found to look at the speaker during significantly more instructions in the roles condition than the no roles condition. This finding shows that our preference for looking at others can be affected by social context. In the roles condition the listener was always the gatherer, following instructions given by an informed chef, who was in charge. In the no roles condition the identity of the listener would switch between the two equal partners, depending on who was giving the instruction. Macdonald and Tatler (2013) found that the degree of informativeness of gaze cues affected the extent to which the cues were sought out, with highly informative cues being sought most often. One possible interpretation of the present findings could be that our manipulation of the roles of the participants effectively manipulated the perceived
informativeness of the cues provided by the chef: listeners in the roles condition may consider the gaze cues of the chef to be highly informative, whereas the gaze cues of the speaker in the no roles condition may be considered less informative.

Alternatively, our pattern of results could arise from a social effect of authority. Liuzza et al (2011) found that right-wing voters were more heavily influenced by the gaze cues of their political leader than the gaze cues of the opposition leader. In the roles condition, the chef is in charge of the procedure and is therefore the leader of the gatherer. It is possible that, as well as being more inclined to follow the gaze cues of a leader, people are also more inclined to orient to the leader's gaze cues. Although the results do not allow us to favour one explanation over the other, these findings provide good evidence that the social context of collaboration can affect the extent to which collaborators look at each other during communication. A more controlled future experiment may be able to distinguish between the effects of the perceived reliability of a person and the perceived social role of a person.

This experiment investigated the effect of social roles on eye movement behaviour in a natural collaboration by using dual portable eye-trackers. We manipulated the roles of the participants to investigate the effect on gaze behaviour. Listeners were found to look more at a speaker providing verbal instructions if the speaker was playing the role of a chef. This suggests that our tendency to look at others is either affected by our social perceptions of a person or by our perception of their reliability. Additionally, we found that in this real social collaborative setting, people spent very little time looking at each other, challenging the generalisability of the conclusions from lab-based paradigms (Birmingham et al, 2007; 2009; Zwickel and Võ, 2010). Our results provide a strong case for investigating gaze cueing behaviour in highly naturalistic environments as well as providing evidence for the effect of social context on the utilisation of gaze cues.

## Acknowledgments

This paper was supported by an EPSRC Studentship awarded to Ross Macdonald. The authors would like to thank Anne-Joanna MacGregor, Julia McVean and Nicole Spittle for assistance with video coding.

## References

Altmann, G.T.M., \& Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. Cognition, 73, 247-264.
Anderson, N.C., Risko, E.F., \& Kingstone, A. (2011). Exploiting human sensitivity to gaze for tracking the eyes. Behavioural Research,43(3), 843-52.
Argyle, M., \& Dean, J. (1965). Eye-contact, distance and affiliation. Sociometry, 28(3), 289-304.

Birmingham, E., Bischof, W. F., \& Kingstone, A. (2007). Why do we look at people's eyes? Journal of Eye Movement Research, 1(1), 1-6.
Birmingham, E., Bischof, W. F., \& Kingstone, A. (2009). Get Real! Resolving the debate about equivalent social stimuli. Visual Cognition, 17 (6), 904-924.
Crosby J. R., Ronin B., \& Richardson D. R. (2008). Where do we look during potentially offensive behaviour? Psychological Science. 19(3): 226-228.
Emery, N. J. (2000). The eyes have it: the neuroethology, function and evolution of social gaze. Neuroscience and Biobehavioural Reviews, 24, 581-604.
Friesen, C.K., \& Kingstone, A. (1998). The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. Psychonomic Bulletin and Review, 5, 490-495.
Galfano, G., Dalmaso, M., Marzoli, D., Pavan, G., Coricelli, C., \& Castelli, L. (2012). Eye gaze cannot be ignored (but neither can arrows). Quarterly Journal of Experimental Psychology, 65(10), 1895-1910.
Gallup, A. C., Chong, A., \& Couzin, I. D. (2012). The directional flow of visual information transfer between pedestrians. Biology Letters, 8(4), 520-522.
Knoeferle, P., \& Kreysa, H. (2012). Can speaker gaze modulate syntactic structuring and thematic role assignment during spoken sentence comprehension? Frontiers in Psychology, 3, 538.
Laidlaw, K. E. W., Foulsham, T., Kuhn, G., \& Kingstone, A. (2011). Potential social interactions are important to social attention. Proceedings of the National Academy of Sciences, 108, 5548-5553.
Liuzza, M.T., Cazzato, V., Vecchione, M., Crostella, F., Caprara, G. V., \& Aglioti, S.M. (2011). Follow my eyes: The gaze of politicians reflexively captures the gaze of ingroup voters. PloS ONE 6(9), e25117.
Macdonald, R.G., \& Tatler, B.W. (2013). Do as eye say: Gaze-cueing and language in a real-world social interaction. Journal of Vision, 13(4):6, 1-12.
Posner, M.I. (1980). Orienting of attention. The Quarterly Journal of Experimental Psychology, 32, 3-25.
Ricciardelli, P., Bricolo, E., Aglioti, S. M., \& Chelazzi, L. (2002). My eyes want to look where your eyes are looking: Exploring the tendency to imitate another individual's gaze. Neuroreport, 13 (17), 2259-2264.
Staudte, M., \& Crocker, M. W. (2011). Investigating joint attention mechanisms through spoken human-robot interaction, Cognition, 120, 268-291.
Tomasello, M., Hare, B., Lehmann, H. \& Call, J. (2007). Reliance on head versus eyes in the gaze following of great apes and human infants: The cooperative eye hypothesis. Journal of Human Evolution, 52, 314-320.
Yarbus, A.L. (1967). Eye Movements and Vision. New York: Plenum Press.
Zwickel, J., \& Võ, M. L. H. (2010). How the presence of persons biases eye movements. Psychonomic Bulletin \& Review, 17, 257-262.

# Social pragmatic factors in reasoning from disjunctions of numerical estimations 

Robert Mackiewicz (rmackiew@swps.edu.pl)<br>University of Social Sciences and Humanities<br>Chodakowska 19/31, 03-815 Warsaw, Poland

Pawel Koniak (pkoniak@swps.edu.pl)<br>University of Social Sciences and Humanities<br>Chodakowska 19/31, 03-815 Warsaw, Poland


#### Abstract

The paper presents two studies that investigated how individuals reason from disjunctive statements that use numerical estimations. In the experiments two types of such statements were used. In the first type both constituents of a disjunction could be a logically correct answer. That is, if "The average life time of a fruit fly is either 9 or 27 days", any of those numbers is logically possible. In the second type that truth of one constituent excluded the truth of the other, e.g. "The average time of holidays in the EU is either higher than 9 days or else higher than 27 days". A simple repetition of any of those figures is an illusory inference as it renders both constituents true. The results of Experiment 1 proved that although the participants showed a tendency to repeat one of the disjuncts as their answer, this tendency was smaller when the content of the statements referred to politics and social life in comparison with the general knowledge questions. The results of Experiment 2 showed that individuals reveal the tendency to repeat opinions coming from speakers who are more likeable, even if such opinions are incorrect illusory inferences. The results of both studies show that illusory inferences appear also in the domain of numerical cognition but they may be reduced by pragmatic factors such as the content of the message and the knowledge about its source.


Keywords: reasoning, mental models, persuasion, illusory inferences, social pragmatics

## Introduction

Imagine that you heard the following statements from two different politicians:
Politician A: The average number of holiday days in The European Union is lower than 19 days.
Politician B: The average number of holiday days in The European Union is lower than 32 days.
How would you answer a question about the average number of holiday days in the EU, if you were informed that one of the above statements is definitely true and the other is definitely false? If both politicians are equally likely to be speaking the truth, it seems quite reasonable to expect one of their statements to be the correct answer. However, assuming that both options are equally possible would be a logical mistake as the truth of one of them excludes the truth of the other. Therefore, the correct estimate of the number of holiday days in the EU must lie between 19 and 32 days, which comes from the fact that the statement of politician A is false and the statement of politician B is true. This is the only possible answer because an assumption that the
statement of politician B is false leads to a contradiction, i.e. the average holiday time is both lower than 19 days and higher than 32 days.

## Mental models and the principle of truth

When two statements are presented in the form of a disjunction, one has to represent the fact that if one of them is true, then the other must be false. But naïve individuals seldom do this, as they typically represent only what is true at the expense of what is false. Forgetting about false possibilities is one of the principles of the theory of mental models (the model theory for short, Johnson-Laird, 2007).

The basic assumption of the model theory is that mental representations are iconic and they represent different possibilities as different mental models. Specifically, models represent what is common in all "possible worlds" when a certain type of relation holds. Therefore, the structure of models corresponds to the structure of what they represent (Johnson-Laird, 2006). The same objects may be represented by different types of mental models depending on the relation that was made salient in a given context. For example, if one is informed that the average life of a common fruit fly is three times longer than 9 days, he or she will understand this expression as a multiplication problem and will know that the correct answer is 27 , even though an experimenter has not yet started to ask any questions. On the other hand, if a participant of a psychological study was informed that:

The average life of a fruit fly is either shorter than 9 days or else it is shorter than 27 days.
he or she might see this expression as a disjunction of two possibilities, and represent them as two separate mental models:
shorter than 9 days

## shorter than 27 days

As none of those mental models represent false cases (i.e. not shorter than 27 days in the first model and not shorter than 9 days in the second model), the model theory predicts that individuals without training in logic should see both possibilities as equally probable. But choosing any of them as the correct answer is a so-called illusory inference as shorter than 9 means also shorter than 27 and shorter than 27 does not exclude it being shorter than 9 .

The model theory predicts such illusory inferences in all those reasoning tasks when models fail to represent what is false (e.g. Johnson-Laird \& Savary, 1999). Such illusory inferences have been proved to exist in different domains of the study of reasoning, e.g. conditionals (Barrouillet \& Lecas, 2000), probabilistic reasoning (Johnson-Laird \& Savary, 1996), quantified reasoning (Yang \& JohnsonLaird, 2000), and relations (Mackiewicz \& Johnson-Laird, 2012). All those studies prove that illusory inferences are quite compelling and in some of them all the participants succumb to drawing erroneous conclusions.

## Pragmatic modulations of reasoning process

Some researchers have tried to find antidotes to illusory inferences. The typical experimental manipulation in such studies provides some of the participants with an opportunity to learn how to falsify certain premises. For example, in the study of Newsome and Johnson-Laird (2006) the participants were explicitly asked to find conditions that make different statements false. Another method is to make the distinctive character of two premises more visible. Santamaria and Johnson-Laird (2000) used this type of manipulation by informing the participants that they should treat different pieces of information as different physical objects; in this case it was different advertisements cut from a newspaper. All such manipulations are semantic in nature as they aid the process of constructing fully explicit models.

The process of reasoning can be also modulated in a pragmatic way (Johnson-Laird \& Byrne, 2002). That is, general knowledge in the long-term memory or some information available from the context may help in forming the expanded representation of a problem (Sperber \& Wilson, 1995). In such a case reasoners may go beyond the logical form of the premises presented to them. For example, if you are informed that

The average time required to fulfill legal requirements necessary to open a new company is less than 9 days or else it is less than 27 days.
you may invoke from your memory a program recently presented on TV that urged for shortening the period of establishing a new company. Although you may not remember the details of this program, you might think that the whole process is definitely longer than a week and shorter than a month, and you could estimate its length at, say, 14 days. Such an answer is logically correct as it renders the first disjunct false and the second true. However, the process of arriving at this conclusion would be not a result of the analysis of what is true and what is false about the possibilities. Indeed, based on the same kind of recollections one would probably give the same estimation of 14 days, even if the disjunction was presented in the following way:

The average time required to fulfill legal requirements necessary to open a new company is less than 9 days or else it is more than 27 days.

In this case 14 days makes both disjuncts false and therefore it is logically incorrect. There is still a possibility that one may know that different types of companies require different legal formalities. Such a person would probably withdraw from giving any estimate of the time necessary to start a new business, claiming that the information given in the premises is incomplete.

As the study of Gigerenzer, Hoffrage, and Kleinbölting (1991) shows, people may use different types of background knowledge in order to arrive at numerical estimations of different facts. For example, they may think that a city that has a soccer team in the first league should be larger than one which team plays in the lower league. We believe that many such cues are available when numerical estimations refer to the domain of social life (e.g. percent of households connected to internet, average time of holiday in the EU) or governmental fiscal policy (e.g. public debt per capita, unemployment rate, percent of the EU funding used in the recent year). We used statements from those domains in our first experiment. As they are frequently mentioned in media coverage of politics, we refer to such statements as political in the rest of the paper. On the other hand, when no pragmatic clues are available, naïve reasoners should more often repeat one of the figures available from the premises as their own answer. We refer to such statements as general knowledge questions and in our studies we included in this group estimations of different facts of natural life (e.g. number of wolves hunting alone, average height of trees) and everyday human activities (e.g. number of words in the average email, number of letters in the average sentence).

We predict that naïve individuals rarely go beyond the principle of truth and in most cases they should err when the correct answer requires envisaging the situations in which some statements are false. However, it should happen less often in the case of political than general knowledge questions because pragmatic knowledge may suffice in recalling correct information from memory in the first type of questions. We verified this prediction in our first experiment.

Pragmatic modulations may also help individuals to see a certain set of premises not as a logical inference but rather as an attempt at persuasion. Recently, Mercier and Sperber (2011) have put forward a hypothesis that the main goal of reasoning is argumentation. Therefore, the main purpose of reasoning is to provide a set of logically related arguments that would support certain thesis. Or, if one is the target of a persuasion attempt, reasoning might be used in order to falsify the statements that other person uses to convince someone to his or her beliefs. Such social pragmatic factors may also include inferences about the intention or credibility of the source of a persuasive message (Bohner, Ruder, \& Erb, 2002) or its truthfulness (Eisend, 2006). In the case of reasoning in the political domain, some individuals might even rely on peripheral clues such as personal attractiveness of a politician (Bohner, Moskowitz, \& Chaiken, 1995) or potential gains and losses that he might attain (Priester \& Petty, 1995). In our second experiment we
wanted to check if reasoners are more likely to choose as their own answer an opinion presented by a person who is more likeable. So in this study we dissociated the content of the problem (politics vs. general knowledge questions) and the source of the message (likeable vs. non - likeable politician).

## Experiment 1

The first experiment compared how naïve individuals reason through disjunctions of general knowledge and political statements. We used two versions of both types of such disjunctions. In the first version, repetition of one of the disjuncts led to an illusory inference. We shall call them illusory problems throughout the rest of the paper. Two examples of such problems were given in the introductory section. Apart from the statements of the form
$A$ is lower than $X$ or else $A$ is lower than $Y$.
we also used statements:
A is bigger than X or else A is bigger than Y .
In all such statements the correct answer is any number that lies between X and Y as the principle of truth excludes all other possibilities.

The second set of disjunctions used pairs of statements that could not be both true at the same time but the falsity of one of them did not exclude the other as a possibly correct answer. We used two types of such statements:
$A$ is lower than $X$ or else $A$ is higher than $Y$.
A equals X or else A equals Y .
An example of the second type would be:
Either the average life time of the fruit fly equals 9 days or else it equals 27 days.
Given that one of those statements is true and the other is false, one cannot give any precise estimate of the average life time of the fruit fly. If the first disjunct is true than the other is false and vice versa. Such statements are typically tagged as control problems in the model theory research and so we will use this label throughout the rest of the paper.

## Method

Participants. 27 undergraduate psychology students from the Warsaw University of Social Sciences and Humanities took part in the study in exchange for a course credit. Although none of them had participated in a higher - level course in logic, it must be noted that an elementary course in this subject is obligatory at all higher education institutions in Poland. Participating in such a course should not, in fact, influence the results of our studies as it provides only basic level knowledge.

Design and materials. Participants acted as their own controls and were asked to give their own numerical estimations for twelve problems presented in the form of disjunctions "Either $X$ or else $Y$ ". Half of them were illusory, as in the examples described in the introduction to this experiment, and the other six were the control
inferences. Three of the illusory inferences used "higher" relations in both disjuncts and the other three used the predicate "lower". Three of the control inferences used "higher" - "lower" relations and the other three used exact numbers. Our main independent variable: general knowledge or political content was manipulated as a between group factor. Therefore, half of the participants were presented with general knowledge problems and half with political ones.

We used different contents for each of 12 general knowledge problems and for each of the political statements. Examples of both types of content are provided in the introductory section. The pairs of numbers presented in each of those problems were different. Nevertheless, each pair from the political domain matched one pair of numbers from the list of those presented with the general knowledge questions.

All problems were presented in the form of a small booklet. The instructions informed that the experiment was not a test of either intelligence or the personality of the participants. The instructions informed the participants that they would see different pairs of statements and that in each pair one of the statements is true and the other is false, but it is not possible to say which is true and which is false. The participants were also informed that they should give their own estimations only on the basis of the content that was used in each pair of statements. The two key instruction sentences were phrased in the following way: "Try to give your own estimate in each of the situations described below only on the basis of the information provided here. (...) Please write a number that reflects your estimation or, if you are not able to give your own answer, write " $X$ " in the answer line".

## Results and discussion

Our main concern was to investigate how often the participants repeat the figures provided in the disjunctive statements. In all control problems such repetitions should be considered as possibly correct answers, while in illusory problems a simple repetition was considered as an incorrect answer. Table 1 presents percentages of different types of answers for illusory and control problems with the general knowledge and the political content.

Similarly to all studies investigating illusory inferences, we obtained a strong effect of the type of disjunction. In general, the participants were more often correct on control than on illusory problems: $73 \%$ and $14 \%$, respectively (Wilcoxon test $z=4.02, p<.001$ ). As we predicted, the participants were more likely to repeat numerical estimates from the general knowledge statements than from those pertaining to the world of politics. As repetitions were correct in the control problem, this resulted in a bigger number of correct answers in the control problems with the general knowledge content in comparison with the political ones (Wilcoxon test $z=2.43, p=.008$, one tailed). However, in the case of illusory inferences repetitions were logically incorrect and so there were more such answers
with the general knowledge questions than with the political content (Wilcoxon test $z=2.16, p=.015$, one tailed). The participants did not show any regular tendency to choose the first or the second disjunct as their answer. However, the first one was chosen slightly more often in abstract (47\% choices of the first vs. $39 \%$ choices of the second disjunct) than in political problems ( $24 \%$ for the first one and $34 \%$ for the second one).

Table 1. Percentages of answers repeating and not repeating the numbers provided in disjunctive statements in
Experiment 1. (The column for the correct non repetitions in control problems is empty, as only repetitions could be
considered as correct answers; the percentages of logically correct answers are marked by italics.)

|  | Control problems |  |  |
| :---: | :---: | :---: | :---: |
| Content | Correct <br> repetitions | Correct non- <br> repetitions | Incorrect <br> non- <br> repetitions |
| Political <br> General <br> knowledge | $49 \%$ |  | $51 \%$ |
|  | $85 \%$ |  | $15 \%$ |
|  | Incorrect <br> repetitions | Correct non- <br> repetitions | Incorrect <br> non- <br> repetitions |
| Political <br> General <br> knowledge | $61 \%$ | $23 \%$ | $16 \%$ |

As all illusory problems were of the form: "A is higher (lower) than X or else A is higher (lower) than Y , the correct answers to all of them were estimates that fell between those two numbers provided in the statements. We called all such answers "correct non - repetitions" in Table 1. Very rarely did the participants gave such estimates in the general knowledge problems but they did it more often in the political ones (Wilcoxon test $z=1.72, p=.043$, one tailed).

In sum, the results of Experiment 1 proved our hypotheses. The participants were less likely to repeat numbers from the statements referring to broadly defined political issues than in the general knowledge problems. Although, generally there were more correct answers in control than in illusory problems, the caution evoked by the political content resulted in a smaller number of incorrect repetitions in illusory problems and more logically correct answers in those problems.

## Experiment 2

The previous study examined the influence of social pragmatic modulation on the tendency to repeat numbers provided in the form of a disjunction. Our Experiment 2 investigated whether it was possible to convince the participants to use the statements provided by some
speakers as participants' own estimations. As many of the studies from persuasion research show, one of the key factors that makes a message more persuasive is the attractiveness of its source (Petty \& Wegener, 1998). Following that, we hypothesized that participants would more often repeat the statements provided by more than less likeable speakers, even though endorsing the conclusions of such speakers may be logically incorrect.

## Method

Participants. We recruited a group of 27 participants from the same population as those described in Experiment 1. They were tested in small groups in exchange for a course credit.
Design and materials. We used the same two sets of 12 general knowledge and 12 political problems as those in Experiment 1. The design of this study was exactly the same as that of Experiment 1 with one exception. Both in political and general knowledge settings we wrote in the instructions that all the statements came from two different politicians. We manipulated their likeability by informing our student participants that one of them said in a television interview that "students are more mature than it is commonly believed", while the other said that "students are spoiled and do not know the true life". We assigned those two descriptions at random to both politicians, so the statements within each pair were assigned once to a likeable and once to a non - likeable politician. We also measured the sympathy for those two politicians by asking the participants to rate their likeness on a five-point scale with higher numbers meaning higher liking. The key instruction sentences were the same as in Experiment 1, so the participants were asked to give their own estimates only on the basis of the information provided in the experiment, and write "X", only if they were not able to figure out their own answers.

## Results and discussion

Again, we proved that there is a strong tendency to repeat numerical estimations provided in the form of disjunctions. As in Experiment 1, this led to correct answers in the control problems (57\%) and incorrect in the illusory problems (31\%). Only in $14 \%$ of their answers to illusory problems did the participants gave correct estimations that fell between two figures provided by the politicians depicted in the instructions (Wilcoxon test for comparison with correct answers in control problems yielded a significant result: $z=3.4, p=.001$ ). In comparison with the previous experiment the difference between the repetitions in general knowledge and political problems was not significant, though it was in the predicted direction: 69\% versus $52 \%$ (Wilcoxon test $z=1,49, p=.07$, one tailed). And also there was no difference between numbers of correct answers in illusory problems concerning general knowledge (18\%) and those connected with the world of politics (10\%).

As our main objective was to see if the participants tended to repeat the statements from a likeable politician more
often than from a non - likeable one, we first checked the effectiveness of our manipulation. Indeed, a politician presented as a person who liked students scored on average 3.96 on a five-points sympathy scales. This number was significantly higher than the average for a person who did not seem to be fond of students: 2.04 (Wilcoxon test $z=$ 3.97, $p<.001$ ). To check our main prediction for this experiment we calculated the Spearman correlations between the attractiveness scores of a likeable politician and the tendency to repeat his statement as the correct answer. The correlation coefficients are given in table 2.

Table 2. Correlations between the ratings of likability of a politician and the frequency of choosing his answer by the participants in Experiment 2. Correlations marked with an asterisk are significant at $p<.05$.

|  | Type of relation between <br> disjunctions |  |
| :--- | :---: | :---: |
| Inference <br> content | Control <br> General | Illusory <br> problems |
| knowledge | $.62^{*}$ | $.63^{*}$ |
| Political |  |  |

As Table 2 reveals, we observed significant positive correlations between the sympathy ratings of a politician who was presented as likeable and the frequency of choosing his statements as answers in both control and illusory problems that required some general knowledge. However, the participants were more skeptical about the expertise of a politician in illusory problems that used the content for the political domain. Only in this case the results did not follow our prediction but it was mainly due to a bigger frequency of participants' refraining from giving any estimates. From the logical point of view, this is also an incorrect answer but we counted the correlations only for answers that repeated the opinion of one of the politicians described in the instructions to this experiment.

The results of Experiment 2 repeated our findings from Experiment 1: the participants were more likely to repeat one of the statements as their own answers. This led to illusory inferences in those problems where the truth of one statement excluded the truth of the other. As our correlations show, there was quite a strong tendency to choose more often the estimates given by a more likeable politician. Our manipulation could seem somehow suspicious to the participants of the study: it is hard to believe that a politician knows anything about the life span of a fruit fly. We observed a smaller tendency of repetitions in illusory problems in this experiment than in the previous one. But possible skepticism did not lead to a bigger number of correct answers in illusory problems in comparison with Experiment 1. It seems as the participants in Experiment 2 more often refrained from giving any numerical estimates in comparison with those taking part in Experiment 1.

## General Discussion

The main purpose of the current study was to investigate how naïve individuals solved problems that used a pair of numerical estimations presented in the form of a logical disjunction. We used two versions of such problems. In the first version, one example being $A$ is lower than $X$ or else $A$ is higher than $Y$, the verbatim repetition of any of the statements had to be considered as a logically correct answer. However, in the second set of trials, including for example $A$ is lower than $X$ or else $A$ is lower than $Y$, the correct answer fell between X and Y and simple repetition of one of the disjuncts did not take into account that in such a case both disjuncts are true at the same time. Following the tradition in the psychological study of reasoning (Johnson-Laird \& Savary, 1996) we referred to such problems as illusory inferences.

A number of researches have tried to find an antidote to illusory inferences. Most of the previous research used direct exemplifications of situations in which logical statements are false. We used a different approach and manipulated the pragmatic factors that led our participants to look for alternative possibilities. In Experiment 1 we compared frequencies of illusory answers for disjunctions in which the statements referred to the general knowledge questions (e.g. the average life of a fruit fly) with the problems referring broadly to social life and politics (e.g. the average time necessary to open a new company). We assumed that social and political issues trigger more cues from the long term memory that participants could use to arrive at their own answers. Our predictions were confirmed. In Experiment 1 the participants were significantly more often correct in illusory problems from the political domain than the general knowledge problems that did not induce pragmatic cues. The "illusory effect" was also present in our Experiment 2 in which we wanted to cue the participants to repeat an opinion presented by a more likeable politician. In this experiment the difference between the proportions if illusory answers in political and the general knowledge questions was lost but indeed we observed a significant tendency to repeat an opinion of a more likeable politician. Our manipulation of the likeability that we used in Experiment 2 might seem suspicious to our student participants, so they more often avoided direct repetitions of opinions coming from a politician who may know nothing about the subject matter of a given question. We guess, however, that it is possible to design a manipulation that would more directly cue the participants to choose one of the provided opinions. Therefore the current results of Experiment 2 provide a useful stepping stone for future research.

Both experiments showed that illusory inferences can be elicited in the domain of numerical cognition. Although there are many studies in the cognitive psychology of reasoning that explicitly use numbers in experimental materials (cf. Oaksford, Chater, \& Larkin, 2000), they mainly use them as labels of probabilities or frequencies. Our study differs from such accounts as we investigated
how individuals reason from relational predicates such as "higher than X" or "lower than X". As it seems they do not only treat numbers as the representations of abstract numerical quantities, but are also able to treat numerical statements as entries into logical arguments.

Our study also differed from those typical for the area of persuasion research. Such studies normally measure the change of attitudes between and after the presentation of a persuasive message (e.g. Bohner, Einwiller, Erb, \& Siebler, 2003). In our research we asked the participants to draw logically valid conclusions taking into account only the information given in the premises. As it turned out, the reasoners were more likely to go beyond the logical form of the problem when its content provided pragmatic cues that triggered the search for available facts in the long term memory (Exp. 1). Although they tended to repeat more often an opinion from a person who was presented as more likeable, it did not help them when it came to finding the logically correct answers (Exp. 2).

As asserted by Gilbert (1991), understanding the message entails believing that it is true at least until it is falsified by some other clues or statements. The results of our experiments show that indeed people believe that two statements that are not overtly contradictory can be both true at the same time. Individuals do not assume that one of them can be false and, what is more, they do not follow the consequences of such assertions. Such an attitude may have an important influence on the world of politics and the role of democratic institutions. It is quite likely that during election campaigns voters do not analyze the relations between statements of different politicians but they rather choose those they would like to believe on the basis on pragmatic factors, such as the likeability of the source of the message.

## Acknowledgments

This research was made possible by The Polish Ministry of Science and Education grant no .2836/01/E/560/S/2012 to both authors. We are grateful to Wojciech Cwalina and Phil Johnson-Laird for discussions and advice.

## References

Barrouillet, P., \& Lecas, J-F. (2000). Illusory inferences from a disjunction of conditionals: A new mental models account. Cognition, 76, 167-173.
Bohner, G., Einwiller, S., Erb, H.-P., \& Siebler, F. (2003). When small means comfortable: Relations between product attributes in two-sided advertising. Journal of Consumer Psychology, 13, 454-463.
Bohner, G., Moskowitz, G.B., \& Chaiken, S. (1995). The interplay of heuristic and systematic processing of social information. In W. Stroebe, \& M. Hewstone (Eds.), European Review of Social Psychology (Vol. 6, pp 3368). Hoboken: Wiley.

Bohner, G., Ruder, M., \& Erb, H.P. (2002). When expertise backfires: Contrast and assimilation effects in persuasion. British Journal of Social Psychology, 41, 495-519.

Eisend, M. (2006). Source credibility dimensions in marketing communication - a generalized solution. Journal of Empirical Generalisations in Marketing Science, 10, 1-33.
Gigerenzer, G., Hoffrage, U., \& Kleinbölting, H. (1991). Probabilistic mental models: a Brunswikian theory of confidence. Psychological Review, 98, 506-528.
Gilbert, D.T. (1991). How mental systems believe. American Psychologist, 46, 107-119.
Johnson-Laird, P.N. (2006). Mental Models, sentential reasoning, and illusory inferences. In C. Held, M. Knauff, \& Q. Vosgerau (Eds.) Mental Models in Cognitive Psychology, Neuroscience, and Philosophy of Mind (pp. 1-25). Rotterdam: Elsevier.
Johnson-Laird, P.N. (2007). How we reason. Oxford: Oxford University Press.
Johnson-Laird, P. N., \& Byrne, R.M. (2002). Conditionals: A theory of meaning, pragmatics, and inference. Psychological Review, 109, 646-678
Johnson-Laird, P.N, \& Savary, F. (1996). Illusory inferences about probabilities. Acta Psychologica, 93, 6990.

Johnson -Laird, P.N., \& Savary, F. (1999). Illusory inferences: a novel class of erroneous deductions. Cognition, 71, 191-229.
Mackiewicz, R., \& Johnson-Laird, P.N. (2012). Reasoning from connectives and relations between entities. Memory \& Cognition, 40, 266-279.
Mercier, H., \& Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory of reasoning. Behavioral and Brain Sciences, 34, 57-74.
Newsome, M.R., \& Johnson-Laird, P.N. (2006). How falsity dispels fallacies. Thinking \& Reasoning, 12, 214-234.
Oaksford, M., Chater, N., \& Larkin, J. (2000). Probabilities and polarity biases in conditional inference. Journal of Experimental Psychology: Learning, Memory \& Cognition, 26, 883-899.
Petty, R.E., \& Wegener, D.T. (1998). Attitude change: Multiple roles for persuasion variables. In D. Gilbert, S. Fiske, \& G. Lindzey (Eds.), The handbook of social psychology (pp. 323-390). New York: McGraw-Hill.
Priester, J.R., \& Petty, R.E. (1995). Source attributions and persuasion: Perceived honesty as a determinant of message scrutiny. Personality and Social Psychology Bulletin, 21, 637-654.
Santamaria, C., \& Johnson-Laird, P.N. (2000). An antidote to illusory inferences. Thinking \& Reasoning, 6, 313-333.
Sperber, D., \& Wilson, D. (1995). Relevance. Communication and Cognition. Oxford: Blackwell.
Yang, Y., \& Johnson-Laird, J.P. (2000). Illusory inferences with quantified assertions: how to make the impossible seem possible, and vice versa. Memory \& Cognition, 28, 452-465.

# The Role of Cognitive Niches in Mediating Knowledge, Entropy \& Violence 

Lorenzo Magnani (lmagnani@unipv.it)<br>Department of Humanities, Philosophy Section<br>University of Pavia, ITALY

Tommaso Bertolotti (bertolotti@unipv.it)<br>Department of Humanities, Philosophy Section<br>University of Pavia, ITALY


#### Abstract

The aim of this paper is to briefly illustrate how the theoretical framework of cognitive niches can prove useful to frame not only the cultural development of human beings, but the naturalization of morality as well.


Keywords: Cognitive niches; Coalition enforcement hypothesis; Violence; Morality; Naturalization.

## Human Beings as Eco-Cognitive Engineers

Human beings usually make decisions and solve problems relying on incomplete information (Simon, 1955). Having incomplete information means that 1) our deliberations and decisions are never the best possible answer, but they are at least satisfying; 2) our conclusions are always withdrawable (i.e. questionable, or never final). That is, once we get more information about a certain situation we can always revise our previous decisions and think of alternative pathways that we could not "see" before; 3) a great part of our job is devoted to elaborating conjectures or hypotheses in order to obtain more adequate information. Making conjectures is essentially an act that in most cases consists in manipulating our problem, and the representation we have of it, so that we may eventually acquire/create more "valuable" knowledge. It is obvious that a great part of human conjectural activity is devoted to guessing hypotheses (that can be moral as well) about situations and events able to help subsequent decisions and actions. Conjectures (and thus "moral" conjectures) can be either the fruit of an abductive selection in a set of pre-stored hypotheses or the creation of new ones: in this sense, abduction - a term from the Peircean tradition - must be understood in an eco-cognitive perspective, which has been fruitfully applied in studies concerning Distributed and Embodied Cognition (Hutchins, 1995; Magnani, 2009). In order to make conjectures, human beings often need more evidence/data: in many cases this further cognitive action is the only way to simply make possible (or at least enhance) a way of reasoning that relies on "hypotheses" that are often hard to produce successfully.

Consider, for instance, diagnostic settings: often the information available does not allow a physician to make a precise diagnosis. Therefore, she has to perform additional tests, or even try some different treatments to uncover otherwise hidden symptoms. In doing so she simply aims at increasing her chances of making the appropriate decision. There are plenty of situations of that kind: for example, scientists are continuously engaged in a process of manipulating their research
settings in order to get more valuable information (Magnani, 2009). Most of this work is completely tacit and embodied in practice. The role of various laboratory artifacts is a clear example, but also in everyday life people face complex situations which require knowledge and manipulative expertise of various kinds - no matter who they are, whether teachers, policy makers, politicians, judges, workers, students, or simply wives, husbands, friends, sons, daughters, and so on. In this sense, human beings can be considered chance seekers, because they are continuously engaged in a process of building up and then extracting latent possibilities to uncover new valuable information and knowledge (Magnani \& Bardone, 2008).

Furthermore, as chance seekers, humans are also ecological engineers. Not only technologies and other artifacts are part of this ecology but also morality and, of course, violent modes of problem-solving. That is to say, humans (like other creatures) do not simply live in their environment, but they actively shape and change it while looking for suitable chances. In doing so, they construct cognitive niches through which the offerings provided by the environment in terms of cognitive possibilities are appropriately selected and/or manufactured to enhance their fitness as chance seekers (Tooby \& DeVore, 1987; Pinker, 1997, 2003). Hence, this ecological approach aims at understanding cognitive systems in terms of their environmental situatedness (Clancey, 1997; Magnani, 2005). Within this framework, "chances" are that kind of "information" which is not internally stored in memory or already available in an external resource, but that has to be "extracted" and then picked up upon occasion.

It is well-known that one of the main forces that shapes the process of adaptation is natural selection. That is, the evolution of organisms can be viewed as the result of a selective pressure that renders them well-suited to their environments. Adaptation is therefore considered as a sort of top-down process that goes from the environment to the living creature (Godfrey-Smith, 1998). In contrast to that, a small fraction of evolutionary biologists have recently tried to provide an alternative theoretical framework by emphasizing the role of niche construction (Laland, Odling-Smee, \& Feldman, 2000, 2001; Odling-Smee, Laland, \& Feldman, 2003).

According to this view, the environment is a sort of "global market" that provides living creatures with unlimited possibilities. Actually, not all the possibilities offered by the environment can be exploited by the human and non-human an-
imals populating a peculiar environment. For instance, the environment provides organisms with water to swim in, air to fly in, flat surfaces to walk on, and so on. However, there are no creatures able to take full advantage of all of those possibilities. ${ }^{1}$ Moreover, all organisms try to modify their surroundings in order to better exploit those elements that suit them and eliminate or mitigate the effect of the negative ones.
This process of environmental selection (Odling-Smee, 1988) allows living creatures to rebuild and shape "ecological niches". An ecological niche can be defined, following Gibson, as a "setting of environmental features that are suitable for an animal" (Gibson, 1979). It differs from the notion of habitat in the sense that the niche describes how an organism lives its environment, whereas the habitat simply describes where an organism lives.

In any ecological niche, the selective pressure of the local environment is drastically modified by organisms in order to lessen the negative impacts of all those elements toward which they are not suited. Indeed, this does not mean that natural selection is somehow halted, rather, this means that an adaptation cannot be considered only by referring to the agency of the environment, but also to that of the organism acting on it. In this sense, animals are ecological engineers, because they do not simply live their environment, but they actively shape and change it (Day, Laland, \& Odling-Smee, 2003).

It is well-known that, from the point of view of physics, organisms are far-from-equilibrium systems relative to their physical or abiotic surroundings. ${ }^{2}$ Apparently they violate the second law of thermodynamics because they stay alive, the law stating that net entropy always increases and that complex and concentrated stores of energy necessarily break down. It is said that they are open, dissipative systems (Prigogine \& Stengers, 1984), which maintain their status far from equilibrium by constantly exchanging energy and matter with their local environments. Odling-Smee, Laland and Feldman quote Schrödinger, contending that an organism has to "feed upon negative entropy [...] continually sucking orderliness from its environment" (Schrödinger, 1992, p. 73). To create cognitive niches is a way that an organism (which is always smartly and plastically "active", looking for profitable resources, and aiming at enhancing fitness) has to stay alive without violating the second law: indeed it "cannot" violate it. In this sense

[^141]cognitive niches can be considered obligatory: "To gain the resources they need and to dispose their detritus, organisms cannot just respond to their environments [...] to convert energy in dissipated energy" (p. 168).

Evolution is strictly intertwined with this process and so it has consequences not only for organisms but also for environments. Sometimes the thermodynamic costs are negligible (like in the heat loss caused by photosynthesis that is returned to the universe, "which is in effect infinite"- p. 169), sometimes they are not, in this case abiota of the environment have no capacity to contrast the niche-constructing activities of organisms (like for example, the atmosphere, which is in a new physical state of extreme disequilibrium in relation to exploitation of the Earth's limited resources). The only no-costs exception is when organisms die - and lose their far-fromequilibrium status). In this case the dead bodies are returned to the local environment in the form of dead organic matter (DOM), still a kind of niche construction, so to say, also called "ghost niche construction" (Odling-Smee et al., 2003, p. 170). Of course biota can resist most thermodynamic costs imposed on them by other niche-constructing organisms, often performing counteractive niche-constructing activities.

## Cognitive Niche Construction and the Mediation of Aggressivity

It is important to clarify the concept of cognitive niche that is at the basis of the possibility to grasp human moral and axiological systems in a naturalistic way, and the intertwined violence, which in this perspective still appears in all of its "banality". A recent book by Odling-Smee, Laland and Feldman (Odling-Smee et al., 2003) offers a full analysis of the concept of cognitive niche from a biological and evolutionary perspective. "Niche construction should be regarded, after natural selection, as a second major participant in evolution. [...] Niche construction is a potent evolutionary agent because it introduces feedback into the evolutionary dynamics" (Odling-Smee et al., 2003, p. 2). ${ }^{3}$ By modifying their environment and by their affecting, and partly controlling, some of the energy and matter flows in their ecosystems, organisms (not only humans) are able to modify some of the natural selection pressure present in their local selective environments, as well as in the selective environments of other organisms. This happens particularly when the same environmental changes are sufficiently recurrent throughout generations and selective change: "Even though spiders' webs are transitory objects [...] the spiders' genes 'instruct' the spider to make a new one" (Odling-Smee et al., 2003, p. 9). The fact that spiders on a web are exposed to avian predators suggests that webs can be a source of selection that produces further phenotype changes in some species, such as the marking of their webs to enhance crypsis or the creation of dummy

[^142]spiders probably to divert the attention of the birds that prey on them. Hence, also spiders adopt what humans call cheating and cognitively alter their cognitive niches to this aim. Cheating is part and parcel of aggressive predatory behavior (Bertolotti, Magnani, \& Bardone, 2013). It is of course not appropriate and clearly anthropomorphic to call these kinds of non human animal behavior "violent", but it remains clear that both in human and non human - especially gregarious - animals the construction of cognitive niches is related to the importance of triggering cooperation and of attacking, more or less violently, other living beings. So the cognitive niches also play, constitutively, the role of carriers of aggressiveness, and in humans, who intentionally build them, they can be legitimately called "moral" and "violent". This general description of cognitive niches is extremely interesting if matched with Gibson's definition of a niche as a "set of affordances" (Gibson, 1977). Relying on his concept of affordance, Gibson stresses how the niches characterizes how a peculiar individual acts within the niche itself and can be summed up in that individual's possibilities for action: one's cognitive niche is indeed made up of a series of possibilities extending between the agent and her environment.
While general inheritance (natural selection among organisms influences which individuals will survive to pass their genes on to the next generation) is usually regarded as the only inheritance system to play a fundamental role in biological evolution, niche construction may play a role over various generations, thus introducing a second general inheritance system (also called ecological inheritance by Odling-Smee). In the life of organisms, the first system occurs as a one-time, unique endowment through the process of reproduction (sexual for example); on the contrary, the second system can in principle be performed by any organism towards any other organism ("ecological" but not necessarily "genetic" relatives), at any stage of their lifetime. Organisms adapt to their environments but also adapt to environments as reconstructed by themselves or other organisms. ${ }^{4}$ From this perspective, acquired characteristics can play a role in the evolutionary process, even if in a non-Lamarckian way, through their influence on selective environments via cognitive niche construction. Phenotypes construct niches, which then can become new sources of natural selection, possibly responsible for modifying their own genes through ecological inheritance feedback (in this sense phenotypes are not merely the "vehicles" of their genes). Of course we have to remember that humans are not unique in their capacity to modify their environment, as we have already seen when referring to the case of the spiders that build "dummy spiders" (Wilcox \& Jackson, 2002): other species are informed by a kind of proto-

[^143]cultural and learning process that is very often intrinsically social, even if we have to say that animals seem to lack the ability to accumulate information as seen in the human cultural/technological case: Andy Clark ranks human language as one of the most powerful cognitive niches ever developed (Clark, 2006).

Indeed, it has to be noted that cultural niche construction alters selection not only at the genetic level, but also at the ontogenetic and cultural levels as well. For example the construction of various artifacts challenges the health of human beings:

Humans may respond to this novel selection pressure either through cultural evolution, for instance, by constructing hospitals, medicine, and vaccines, or at the ontogenetic level, by developing antibodies that confer some immunity, or through biological evolution, with the selection of resistant genotypes. As cultural niche construction typically offers a more immediate solution to new challenges, we anticipate that cultural niche construction will usually favor further counteractive cultural niche construction, rather than genetic change (OdlingSmee et al., 2003, p. 261).

With a broader explanatory reach than sociobiology and evolutionary psychology, the theory of niche construction simultaneously explains the role of cultural (and so moral) aspects (transmitted ideas), behavior (and so moral behavior, which directly orients the construction of niche construction itself), and ecological inheritance (artifacts, to be intended also as moral/violent mediators). Of course niche construction may also depend on learning. It is interesting to note that several species, many vertebrates for example, have evolved a capacity to learn from other individuals and to transmit this knowledge, thereby activating a kind of proto-cultural process which also affects niche construction skills: it seems that in hominids this kind of cultural transmission of acquired niche-constructing traits was ubiquitous, and this explains their success in building, maintaining, and transmitting the various cognitive niches in terms of moral systems of coalition enforcement. "This demonstrates how cultural processes are not just a product of human genetic evolution, but also a cause of human genetic evolution" (Odling-Smee et al., 2003, p. 27). From this viewpoint the notion of docility (Simon, 1993) acquires an explanatory role in describing the way human beings manage ecological and social resources to make their own decisions.
(Lahti \& Weinstein, 2005) and (Magnani, 2011, chap. 6) refer to the concept of viscosity to provide an explanation of the gap between the absolutism of morality and the empirical evidence that moral regulations are often infringed with no major consequences either for the whole moral system, or for the very individual who performs the infraction - alas, generating conflicts and violence. Viscosity is certainly constrained by docility, which favors the formation of "the state of being thick, sticky" but also of the state of being "semifluid in consistency, due to internal friction". We said that
the fact that morality is viscous hints at its thickness and being glue-like, thus meaning its capability to be deformed, stressed, pulled apart and reassembled without showing decisive harm to its own stability and reproducibility: this aspect also relates to docility. Viscosity and docility explain how our objectified moral cognitive niches are stable, and at the same time also vulnerable and modifiable. Thus it is easy to see in, a human individual, the stability of moral convictions depending on his stable moral niches, together with the spontaneous attitude to "disengage" them - for example resorting to a "re-engagement" in other moral conducts which are not dominant in his present moral cognitive niche, but still present as vestigial traces of previous - no longer dominant moral cognitive niches (Bandura, 1999; Magnani, 2011).
(Woods, 2013) touches a similar problem, related to docility, when, analyzing fallacious reasoning, he stresses the fact that "Whether full or partial, belief states are not chosen. They befall us like measles", in other words, "say so" induces belief (doxastic irresistibility). Similarly moral cognitive niches too "befall us like measles". The problem is related to the effect of what Gabbay and Woods call ad ignorantiam rule: "Human agents tend to accept without challenge the utterances and arguments of others except where they know or think they know or suspect that something is amiss" (Gabbay \& Woods, 2005, p. 27). The individual agent also economizes by unreflective acceptance of anything an interlocutor says or argues for, short of particular reasons to do otherwise, by applying the ad verecundiam fallacy. Accordingly, the reasoner accepts her sources' assurances because she is justified in thinking that the source has good reasons for them (the fallacy would be the failure to note that the source does not have good reasons for his assurances). Peirce contended, in a similar way, that it is not true that thoughts are in us because we are in them; "beings like us have a drive to accept the say so of others" (Woods, 2013).

It is noteworthy that all these information resources do not only come from other human beings. This would clearly be an oversimplification. Indeed, the information and resources that we continuously exploit are - so to speak - humanreadable. Both information production and transfer are dependent on various mediating structures, which are the result of more or less powerful cognitive delegations, namely, niche construction activities. Of course, it is hard to develop and articulate a rich culture as humans did, and still do, without effective mediating systems (writing, artifacts, material culture, etc.). Hence, we can say that, first of all, docility is more generally concerned with the tendency to lean on various ecological resources, which are released through cognitive niches. Secondly, social/moral learning cannot be seriously considered without referring to the agency of those mediating structures, whose efficiency in storing and transmitting information far exceeds, from many perspectives, that - direct and non-mediated method - of human beings.

## Cognitive Niches as Moral Niches

In the previous section we have tried to show that the concept of cognitive niche is an extremely appropriate intellectual instrument to grasp human cultural and moral systems, and their violent punishment counterparts, in a naturalistic way. It is important to present the moral and potentially violent dimension of cognitive niches. We have said that the activity of niche construction may enter evolution insofar as it modifies the selective pressures humans and other animals have to cope with. From this we can draw two major consequences.
First of all, the activity of cognitive niche construction potentially affects all those who participate and live in the same local environment in terms of cognitive chances made available (or not). That is, eco-cognitive modifications - brought about collectively (like herd-like behaviors) or by certain groups - may affect our shared cognitive repertoire amplifying it but also constraining or even impoverishing it. On certain occasions, eco-cognitive modifications may be considered by some individual (or particular groups of individuals) as threatening, impoverishing, or detrimental for their possibility to solve problems. Basically, they can perceive their cognitive system as if it is externally hacked so that they have to partly re-engineer their relationship with the environment, for instance, by modifying their previous habits or simply forcing them to cope with habits perceived as maladaptive or threatening for them or their group.

The second point deals with the role of the coalition enforcement hypothesis in cognitive niche maintenance. In fact, the construction of cognitive niches and the preservation and their maintenance through coalition enforcement has indeed a moral (and thus violent) dimension: that is, punishment, control and persecution of in-group free riders, and regulation of out-of-group conflicts (Magnani, 2011). The coalition enforcement hypothesis, put forward by Bingham (Bingham, 1999), aims at providing an explanation of the "human uniqueness" that is at the origin of human communication and language, in a strict relationship with the spectacular ecological dominance achieved by H. Sapiens, and of the role of cultural heritage. From this perspective, and due to the related constant moral and policing dimension of Homo's coalition enforcement history (which has an approximately two-million-year evolutionary history), human beings can be fundamentally seen as self-domesticated animals. In hominids, cooperation in groups (which, contrary to the case of non-human animals, is largely independent from kinship) fundamentally derived from the need to detect, control, and punish social parasites, who for example did not share the meat they hunted or partook of the food without joining the hunting party (also variously referred to as free riders, defectors, and cheaters) (Boehm, 1999). These social parasites were variously dealt with by killing or injuring them (and also by killing cooperators who refused to punish them) from a distance using projectile and clubbing weapons. In this case injuring and killing are cooperative and remote (and at
the same time they are "cognitive" activities). According to the coalition enforcement hypothesis, the avoidance of proximal conflict reduces risks for the individuals (hence the importance of remote killing). Of course, cooperative morality that generates "violence" against unusually "violent" and aggressive free riders and parasites can be performed in other weaker ways, such as denying a future access to the resource, injuring a juvenile relative, gossiping to persecute dishonest communication and manipulative in-group behaviors or waging war against less cooperative groups, etc.

From this perspective, the role played by morality (and, thus, violent punishment) is manifold: any activity that involves and signals a commitment toward cognitive niche construction and maintenance is potentially perceivable as violent against concurrent niches. To develop and to maintain some eco-cognitive modifications typical of a certain community implies that those modifications are indeed worth being preserved because they are perceived as good and useful, which immediately clashes with other possible ways of organizing an homologue cognitive niche. If a cognitive niche displays a univocal relationship with the group who developed it and cares for its maintenance, participating in the niche also involves a more or less public endorsement of the group that supports it. Of course one can partake of several niches (and hence of several groups) as long as they do not compete (or are perceived as not competing) in the same area, since no matter how polite the context may be, any conflict is ultimately about violence.

Morality can be considered as part of the niche's distributed knowledge, and it precisely concerns violence insofar as it regulates (also violent) relationships between individuals in the niche and with those that are confronted with it without actually partaking of it. Such a regulating activity is permitted by the dimension of violence embodied in rules and regulations and related punishments but also tacitly conveyed by the cognitive as we just observed: the most patent case of such in-niche morals are deontological codes typical of highly specific cognitive niches, but to different degrees they are traceable in every cognitive niche. Of course, the explicit dimension of normativity is concerned in this characterization of the cognitive niche as moral knowledge expressed in the different registers of rules and regulations is one of the pillars of niche maintenance. Even if a niche is not primarily involved in prescribing certain behaviors to its members, a contextual decency is required in order to obtain a state of homeostasis in intersubjective relationships. Should a niche seem to be totally devoid of general normativity, it would thrive insofar as it was laid upon a wider cognitive niche that is in turn heavily concerned with morals and norms, namely, religions, political and legal institutions and so on.

## Concluding Remarks

It is easy to see that the violent potential constitutively embedded in any cognitive niche actually displays the underlying dimension of structural and symbolic violence (Magnani,

2011, chap. 1). Structural violence is seen as morally legitimate insofar as it plays a crucial role in the activities of niche maintenance. Immediately we have to note that when parents, policemen, teachers and other agents inflict physical or invisible violence for legal and/or moral reasons, those reasons do not cancel the violence perpetrated and violence does not have to be condoned in so far as it is not always perceived as such. On the other hand it must be analyzed how in the case of structural violence those perpetrating agents do not seem to act only on their own behalf but on that of larger institutions that can be political, industrial, economic or religious. Such institutions populate structural violence not with actors but rather with what we call "violent mediators" (or in the extreme case of human beings that have turned themselves into violence mediating socio-cultural "artifacts", as in the role of the policeman in the framework of structural violence). ${ }^{5}$

Structural violence may acquire its most subtle and omnipresent form as the symbolic violence perpetrated by language. As a device of social mediation language is necessarily a cognitive niche mediator (and hence distributor) of violence as well. The violent nature of language is a fact too easily admitted to allow serious reflection, as if every speaker were aware of this horrible truth and wanted to get rid of it as soon as possible, even by simply acknowledging it and leaving it at that. As we have already pointed out, a gentle cluster of speech forms innocently distributes harmful, abusive, destructive, and damaging roles, commitments, inclinations and habits. Language, which is the very moral medium of cooperation and non-violence, also involves unconditional violence even against the speaker herself, insofar as by language one acquires and imposes dominion not only over fellow human beings but also over one's conscious and less conscious self, framing thoughts and emotions in the rigid crystallization brought about by words. The importance of symbolic violence should not be disregarded for one very simple reason: the only requirement to become a perpetrator is easy to meet as it consists in a basic knowledge of the niche language, and the very fact of speaking a language makes the speaker both potentially and actually violent in the symbolic dimension. Culture, knowledge and more highly developed speech abilities may not necessarily help, but conversely they positively turn an agent into an even more subtle perpetrator of violence: we already mentioned Gibson's definition of a niche as a set of affordances: such definition could perfectly fit the case of a moral, violent niche. Knowledge externalizations may constitute moral affordances, becoming possibilities for an individual's moral acting. Yet, should an individual not develop, or acquire, the correct moral affordances, she might be perceived as violent by the rest of the obliging community populating and maintaining the niche - and hence be violently

[^144]punished: what should be further studied is whether defining a moral niche as a set of moral affordances leads us to label it, as the other side of the coin, a set of violent affordances, of chances for violence.

## References

Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. Personality and Social Psychology Review, 3, 193-209.
Baumeister, R. F., Zhang, L., \& Vohs, K. D. (2004). Gossip as cultural learning. Review of General Psychology, 8(2), 111-121.
Ben Jacob, E., Shapira, Y., \& Tauber, A. I. (2006). Seeking the foundation of cognition in bacteria. From Schrödinger's negative entropy to latent information. Physica A, 359, 495-524.
Bertolotti, T., Magnani, L., \& Bardone, E. (2013). Camouflaging truth: A biological, argumentative and epistemological outlook from biological to linguistic camouflage. Journal of Cognition and Culture. (Forthcoming)
Bingham, P. M. (1999). Human uniqueness: a general theory. The Quarterly Review of Biology, 74(2), 133-169.
Boehm, C. (1999). Hierarchy in the Forest. Cambridge, MA: Harvard University Press.
Clancey, W. J. (1997). Situated Cognition: On Human Knowledge and Computer Representations. Cambridge: Cambridge University Press.
Clark, A. (2006). Language, embodiment, and the cognitive niche. Trends in Cognitive Science, 10(8), 370-374.
Dawkins, R. (2004). Extended phenotype - but not extended. A reply to Laland, Turner and Jablonka. Biology and Philosophy, 19, 377-397.
Day, R. L., Laland, K., \& Odling-Smee, F. J. (2003). Rethinking adaptation. The niche-construction perspective. Perspectives in Biology and Medicine, 46(1), 80-95.
Gabbay, D. M., \& Woods, J. (2005). The Reach of Abduction. Amsterdam: North-Holland. (Volume 2 of A Practical Logic of Cognitive Systems)
Gibson, J. J. (1977). The theory of affordances. In R. E. Shaw \& J. Bransford (Eds.), Perceiving, acting and knowing. Hillsdale, JN: Lawrence Erlbaum Associates.
Gibson, J. J. (1979). The Ecological Approach to Visual Perception. Boston, MA: Houghton Mifflin.
Godfrey-Smith, P. (1998). Complexity and the Function of Mind in Nature. Cambridge: Cambridge University Press.
Hutchins, E. (1995). Cognition in the Wild. Cambridge, MA: The MIT Press.
Lahti, D. C., \& Weinstein, B. S. (2005). The better angels of our nature: group stability and the evolution of moral tension. Evolution and Human Behavior, 2, 47-63.
Laland, K. N., Odling-Smee, F. J., \& Feldman, M. W. (2000). Niche construction, biological evolution and cultural change. Behavioral and Brain Sciences, 23(1), 131175.

Laland, K. N., Odling-Smee, F. J., \& Feldman, M. W. (2001).

Cultural niche construction and human evolution. Journal of Evolutionary Biology, 14, 22-33.
Laland, K. N., Odling-Smee, F. J., \& Feldman, M. W. (2005). On the breath and significance of niche construction: a reply to Grittiths, Okasha and Sterelny. Biology and Philosophy, 20, 37-55.
Magnani, L. (2005). Creativity and the disembodiment of mind. In P. Gervás, A. Pease, \& T. Veale (Eds.), Proceedings of CC05, Computational Creativity Workshop, IJCAI2005, Edinburgh (pp. 60-67).
Magnani, L. (2009). Abductive Cognition. The Eco-Cognitive Dimensions of Hypothetical Reasoning. Heidelberg/Berlin: Springer.
Magnani, L. (2011). Understanding Violence. Morality, Religion, and Violence Intertwined: a Philosophical Stance. Berlin/Heidelberg: Springer.
Magnani, L., \& Bardone, E. (2008). Sharing representations and creating chances through cognitive niche construction. The role of affordances and abduction. In S. Iwata, Y. Oshawa, S. Tsumoto, N. Zhong, Y. Shi, \& L. Magnani (Eds.), Communications and Discoveries from Multidisciplinary Data (pp. 3-40). Berlin: Springer.
Odling-Smee, F. J. (1988). The Role of Behavior in Evolution. Cambridge: Cambridge University Press.
Odling-Smee, F. J., Laland, K. N., \& Feldman, M. W. (2003). Niche Construction. The Neglected Process in Evolution. Princeton, NJ: Princeton University Press.
Pinker, S. (1997). How the Mind Works. New York: W. W. Norton.
Pinker, S. (2003). Language as an adaptation to the cognitive niche. In M. H. Christiansen \& S. Kirby (Eds.), Language Evolution (pp. 16-37). Oxford: Oxford University Press.
Prigogine, I., \& Stengers, I. (1984). Order out of Chaos. Man's New Dialogue with Nature. London: Bantam.
Schrödinger, E. (1992). What is life? with mind and matter and autobiographical sketches. Cambridge University Press: Cambridge. (Originally published in 1944)
Simon, H. A. (1955). A behavioral model of rational choice. The Quarterly Journal of Economics, 69, 99-118.
Simon, H. A. (1993). Altruism and economics. American Economic Review, 83(2), 157-161.
Sterelny, K. (2005). Made by each other: organisms and their environment. Biology and Philosophy, 20, 21-36.
Tooby, J., \& DeVore, I. (1987). The reconstruction of hominid behavioral evolution through strategic modeling. In W. G. Kinzey (Ed.), Primate Models of Hominid Behavior (pp. 183-237). Albany: Suny Press.
Wilcox, S., \& Jackson, R. (2002). Jumping spider tricksters: deceit, predation, and cognition. In M. Bekoff, C. Allen, \& M. Burghardt (Eds.), The Cognitive Animal. Empirical and Theoretical Perspectives on Animal Cognition (pp. 27-34). Cambridge, MA: The MIT Press.
Woods, J. (2013). Errors of Reasoning. Naturalizing the Logic of Inference. London: College Publications. (Forthcoming.)

# A Computational Theory of Subjective Probability [Featuring a Proof that the Conjunction Effect is not a Fallacy] 

Phil Maguire (pmaguire@cs.nuim.ie)<br>Department of Computer Science<br>NUI Maynooth, Ireland<br>Philippe Moser (pmoser@cs.nuim.ie)<br>Department of Computer Science<br>NUI Maynooth, Ireland

Rebecca Maguire (rebecca.maguire@ ncirl.ie)
School of Business, National College of Ireland
Mayor Street, IFSC, Dublin 1, Ireland
Mark Keane (mark.keane@ucd.ie)
School of Computer Science and Informatics
UCD, Dublin 4, Ireland


#### Abstract

In this article we demonstrate how algorithmic probability theory is applied to situations that involve uncertainty. When people are unsure of their model of reality, then the outcome they observe will cause them to update their beliefs. We argue that classical probability cannot be applied in such cases, and that subjective probability must instead be used. In Experiment 1 we show that, when judging the probability of lottery number sequences, people apply subjective rather than classical probability. In Experiment 2 we examine the conjunction fallacy and demonstrate that the materials used by Tverksy and Kahneman (1983) involve model uncertainty. We then provide a formal mathematical proof that, for every uncertain model, there exists a conjunction of outcomes which is more subjectively probable than either of its constituents in isolation.


Keywords: Conjunction fallacy; algorithmic statistics; likelihood judgments; surprise; subjective probability.

## Introduction

Breaking news: Pandemonium erupted today at the National Lottery headquarters as the numbers 1, 2, 3, 4, 5 and 6 were drawn for the third week in a row. Lottery officials, stunned by a sense of déjà vu, scrambled to release a statement insisting that the lottery drum selection mechanism meets the highest standards for randomness. Meanwhile thousands are celebrating after ignoring the opinions of mathematicians who had viewed the two previous draws as a statistical fluke. Commentators in the media are demanding an immediate investigation, describing the incident as a fiasco.

The mathematical concept of probability, originally formulated to describe the highly constrained environment of games of chance, has now found its way into everyday parlance, with people using it to quantify the likelihood of everything from the possibility of economic recession to the risk of global warming. Such has been the unquestioned adoption of the probability concept into mainstream culture that it has become the default assumption that probability theory provides the only logical way for people to think about likelihood.

For instance, Tverksy and Kahneman (1983) applied probability theory to real-world situations involving personality decisions, medical judgments, criminal motives and political forecasts. On observing consistent deviations from the mathematical theory, they interpreted their findings as evidence of a serious flaw in human reasoning (see Costello, 2009, for a review of the associated debate). In this article we adopt the alternative stance that consistent deviations between human reasoning and a simplified, artificial mathematical theory are far more likely to reflect deficiencies in the theory than they are to reflect sub-optimality in how people think about likelihood.

## Classical Probability

Probability theory was formalised by Kolmogorov in the 1930s through the notion of probability space, whereby a set of possible outcomes is mapped to a number that represents its likelihood by a probability measure function. For example, a perfect dice outputs the numbers from 1 to 6 with equal frequency. However, in the real world it is rarely feasible to identify the theoretical probability measure function which underlies the events we observe. Because we have to work backwards, using the events to deduce the original function, we can never be sure if the model we are using is correct.

For example, according to classical probability theory, no conceivable sequence of numbers produced by rolling a dice will ever lead us to revise our beliefs about the nature of the dice. Even if we rolled $1,1,1,1 \ldots$ the hypothesis of the sequence being a statistical fluke would always remain infinitely more likely than the possibility that the dice is biased.

In reality, nobody has beliefs which are strong enough to stand up to the requirements of classical probability theory. We strongly believe that the numbers drawn from the lottery are random, yet there are certain sequences which, as in the introductory lottery example, would cause us to question our assumptions and consider other possibilities. If the sequence
$1,2,3,4,5,6$ was drawn three weeks in succession, it might suggest that the balls were not equally weighted, the drum mechanism was defective, or that one of the lottery officials was playing a practical joke. The point where we start to ask questions reveals how strongly we hold our beliefs. But no matter how confident we are about a particular model of reality, there will always be some sequence of events which will cause us to change our mind.

This poses a crucial problem for probability theory. Let's consider the probability of $1,2,3,4,5,6$ being drawn in a lottery for three weeks in a row. If the draw is unbiased then this sequence of events is just as likely as any other. In a lottery with 45 numbers, the exact probability is $C(45,6)^{3}$. But if this sequence of events actually unfolded, it would lead us to believe that the draw mechanism is biased. Given the new updated belief, then the probability of getting $1,2,3,4$, 5,6 is actually far higher. So what is the true probability of this sequence of events?

To apply classical probability theory a single model of reality must be selected. We must assume either that the lottery draw is biased or that it isn't. But doing so would be a mistake because we don't actually know which world is the case. The situation involves model-outcome dependence, insofar as the outcome affects our beliefs about the system that generated it. Stating that the probability of drawing 1, 2, 3, 4, 5, 6 is $C(45,6)^{3}$ is misleading because, if this sequence of events actually occurred, we would no longer trust the assumptions involved in computing that probability.

## Uncertainty in the Real World

The issue here is that classical probability theory only applies to cases involving a definitive probability measure function, while models of reality always involve uncertainty. Though useful for reasoning about games deliberately engineered to generate pseudo-randomness, classical probability has less applicability to everyday life, where reducing uncertainty and optimising models of reality are the principal goals. In our previous work examining the difference between surprise and probability judgments (Maguire, Maguire, \& Keane, 2011) we presented a cognitive theory of uncertainty modeling which views the maintenance of an up-to-date representation of reality as the principal motivation guiding information seeking behaviour. People rely on observational data to continually refine their model of the environment, thus maintaining the optimality of their decision making. In particular, the signal that they rely on to diagnose discrepancies between their model and the real world is randomness deficiency.

The best model of a set of observational data is the one which describes it most concisely, so that the description of the data relative to the model is 'incompressible' or random (see Rissanen, 1978; Gács, Tromp, \& Vityányi, 2001). In the case that one's model of reality is optimal, then new sensory data should still be random with respect to it. The experience of randomness deficiency (i.e. a pattern which could be described more concisely using an alternative model) causes
alarm bells to go off, because it indicates that one's model is likely to be suboptimal. This is known as surprise.

When surprise occurs there are two potential resolutions. First, more observation data can be gathered, which might mitigate the randomness deficiency by revealing it to be a statistical fluke. If this does not resolve the discrepancy then the remaining alternative is to update one's model to fit the data. Either way, the resolution process necessitates urgent sampling of information from the environment. During the surprise response, eye widening, opening of the mouth and enlargement of the nasal cavity serve to facilitate the intake of sensory information (see Maguire et al., 2011).

Consider for example looking at the floor and seeing some crumbs which spell out the words "YOU ARE BEING WATCHED". When crumbs fall on the floor it is just as probable that they will arrange themselves into this pattern as any other. If we were certain that the crumbs had fallen randomly then it would not be interesting. However, where knowledge is uncertain then people respond to randomness deficiency. The pattern of crumbs is randomness deficient because there is another model which can explain it more concisely: Somebody might have deliberately arranged the crumbs in this way. The first strategy is to look at the rest of the floor. If the rest of the floor is covered in many crumbs which have no other patterns then the overall randomness deficiency is mitigated. If these are the only crumbs on the floor then finding a satisfactory explanation becomes critical.

People are motivated to seek out randomness deficiency in the world (Dessalles, 2006). The experience of randomness deficiency with subsequent resolution through representational updating is what makes subjects interesting, films entertaining and jokes funny (Schmidhuber, 2009). Accordingly, when people speak intuitively about likelihood and probability, it is the concept of representational updating which is relevant to them.

## Subjective Probability

Because it assumes a definitive probability measure function, classical probability theory cannot be applied to the concept of representational updating. This limitation means that the theory is, for the large part, irrelevant to everyday life and thus inappropriate for evaluating the nature of human reasoning.

Developments in algorithmic statistics have allowed probability theory to be extended to situations involving an uncertain probability measure function (see, e.g., Vityányi \& Li, 2000; Gács et al., 2001). The optimal model which can be derived from a set of observations is the one which maximizes the compression of that dataset, yielding the Minimum Description Length (MDL), a concept which formalizes Occam's razor.

Whenever an observation is no longer typical with respect to an MDL model it should be adjusted to lower the randomness deficiency of the data (see Li \& Vityányi, 2008, for details on how the updating process is carried out). We can quantify the extent of this representational adjustment in
terms of the amount of information that, given the original model, would be required to obtain the updated model. The more the information required, the more significant (and less likely) the update.

The model that people hold of reality represents the very best that they can do in representing their environment and provides the very best that they can achieve in terms of predictions. If we assume that our representation is a reliable predictor of events then the larger a potential update to that representation, the rarer it should be. Accordingly, we can apply probability theory to speak about the likelihood of an outcome requiring an update of a particular size. The uncertainty which precludes probability theory from being applied to real-world scenarios is circumvented by shifting the focus from an underdetermined probability measure function to the immutable mechanism of representational updating.

## Preliminaries

A computable probability density function $p$ can be interpreted as a model for a string generating device. Given such a device, described by $p$, there are some "type of strings" we expect to be output, whereas some others are surprising. String $x$ is said $p$-typical if it is a random string relative to the model described by $p$, i.e. the model already describes all the regularities in $x$.

Formally, let $\alpha>0$ be a constant, called the surprise threshold, which represents the level of randomness deficiency that necessitates representational updating. String $x$ is $p$-typical with surprise threshold $\alpha$ (or $(p, \alpha)$-typical) if the length of its shortest description given $p$ is at least the number of bits a Shannon-Fano code based on $p$ would require (an encoding where the more $p$-likely a string is, the shorter its encoding will be) after subtracting the surprise level $\alpha$, i.e.,

$$
K\left(x \mid p^{*}\right) \geq-\log p(x)-\alpha
$$

The idea behind the minimal description length (MDL) of a string $x$ (Gács et al., 2001) is to take the shortest (in description length) among all models for which $x$ is typical. To avoid overfitting (i.e. the model is specifically built for $x$ instead of for all "strings of type $x$ ") the description length of both the model and the string given the model, should be equal to the description of the string on its own. Formally, probability density function $p$ is optimal for string $x$ if the shortest description of $x$ has the same length (up to an additive constant) as the shortest description of $p$ plus the number of bits required for a Shannon-Fano encoding of $x$ based on $p$, i.e.,

$$
K(x)=K(p)-\log p(x) \pm O(1)
$$

where $O(1)$ means the equality holds up to an additive constant. The MDL of string $x$ is the shortest (description length) among all optimal probability density functions for $x$ for which $x$ is typical.

## Subjective information and probability

Suppose an observer experiences observations $d_{1}, d_{2}, \ldots$ generated by some source with computable probability density
$p_{\text {source }}$. The observer tries to learn the probability density $p_{\text {source }}$ by finding the shortest optimal model based on the observations made so far. Formally, after having observed strings $d_{1}, d_{2}, \ldots, d_{n}$, the observer seeks to construct a hypothetical model $p_{n}$ where

$$
\begin{aligned}
& p_{n}=\arg \min \left\{\left|p^{*}\right|: p \text { is optimal for } d_{1}, d_{2}, \ldots, d_{n}\right. \text { and } \\
& \left.\quad d_{1}, d_{2}, \ldots, d_{n} \text { are }(p, \alpha) \text {-typical }\right\}
\end{aligned}
$$

If the next observation $d_{n+1}$ is surprising, action may be required. Formally, observation $d_{n+1}$ is $\alpha$-surprising if the length of its shortest description given $p$ is less than the number of bits a Shannon-Fano code based on $p$ would require after subtracting the surprise level $\alpha$, i.e.,

$$
K\left(d_{n+1} \mid p_{n}^{*}\right)<-\log p_{n}\left(d_{n+1}\right)-\alpha
$$

If an update is performed, then the subjective information of $d_{n+1}$ (the "cost" of the update) is the amount of information needed to update the model to the latest, that is the length of the shortest description of the new model, given the old model, i.e.,

$$
\operatorname{subjective~information~}\left(d_{n+1}\right)=K\left(p_{n+1}^{*} \mid p_{n}^{*}\right)
$$

Subjective probability (the probability of the update) can then be quantified based on the amount of information it contains, i.e.,

$$
\operatorname{subjective} \operatorname{probability}\left(d_{n+1}\right)=2^{-K\left(p_{n+1}^{*} \mid p_{n}^{*}\right)}
$$

## Experiment 1

In the following experiment we investigated the hypothesis that people use subjective probability rather than classical probability to judge the likelihood for real-world events. We used an example for which the use of classical probability theory seems particularly compelling, namely lottery sequences (see Dessalles, 2006). A naive application of classical probability suggests that all lottery sequences are just as likely.

## Method

In a lottery system where 6 numbers are drawn from 45, each ordered sequence has a classical probability of $C(45,6)$. According to the theory outlined in the previous section, the subjective probability of an outcome is related to its randomness deficiency. People expect the lottery numbers to be Kolmogorov-random. The more they deviate from a typical random string, the lower the subjective probability that they reflect the output of a random source. The randomness deficiency of a string is quantified precisely by its MDL. However, since this theoretical construct is not computable in practice, we are obliged to create a heuristic compressor which approximates it.

We considered the patterns to which people are sensitive in discriminating predictable sequences from random ones. Overtly non-typical random patterns include ones in which
the numbers are consecutive (e.g. $3,4,5,6,7,8$ ) or where they increase in a constant step size. To compress these patterns we created a simple compressor which takes in an ordered sequence of six numbers, and computes the six step sizes between them (with the first number counting as the first step). A Huffman encoding scheme is then applied, which relates bit size to step size. A breakdown of the structure of the associated Huffman tree is provided in Table 1.

Using this system the sequence $10,32,33,35,39,45$ is transformed to step sizes of $+10,+22,+1,+2,+4,+6$ which is then encoded using $8+8+2+3+4+6=31$ bits. Analysing six years of bi-weekly Irish National Lottery draws revealed a mean compressed length of 30.9 bits, with a mode of 31 bits. The most randomness deficient of the 624 sequences was 2 , $4,32,34,36,37$ (description length of 20 bits), while the most random was $9,20,26,27,34,45$ (description length of 39 bits). The theoretical minimum description length of our system was 12 (e.g. $1,2,3,4,5,6$ ), while the theoretical maximum was 43 (e.g. $7,13,20,29,36,45$ ). The number of bits needed to perfectly encode an ordered random sequence of six numbers between 1 and 45 is 23.0 bits. Although our compressor cannot compute MDL, it delivers compression for randomness deficient outputs (i.e. it compresses below 23.0 bits for certain non-typical random sequences) and can therefore be used to evaluate the hypothesis that people use subjective rather than classical probability.

Table 1: Structure of Huffman encoding scheme.

| Level Depth | Leaves | \#Branches |
| :--- | :--- | :--- |
| 1 | - | 2 |
| 2 | +1, repeat | 2 |
| 3 | $+2,+3$ | 2 |
| 4 | +4 | 3 |
| 5 | +5 | 5 |
| 6 | +6 | 9 |
| 7 | $+7,+8$ | 16 |
| 8 | +9 up to +40 | - |

Participants 130 undergraduate students from NUI Maynooth participated voluntarily in this study.

Procedure As an initial step we purchased two quickpick (i.e. randomly selected) lottery tickets for the next week's Irish National Lottery, with six ordered numbers ranging from 1 to 45 . Participants were informed that we had purchased these tickets and that, for each of the two quickpick sequences, their goal was to identify it from among a group of five candidate sequences. No mention was made of how the other four sequences had been generated.

Each quickpick sequence was presented on a screen along with four other sequences randomly generated using our compressor algorithm. The four distractor sequences met the constraints of having compressed bit-sizes of between 15 and 18 bits, 19 and 22 bits, 23 and 26 bits, and 27 and 29 bits respec-
tively. As it happened, the first lottery ticket sequence had a compressed description length of 31 bits, and the second had a length of 30 bits. The ordering of the five sequences on the screen was randomized.

Participants ranked each set of five sequences in order of likelihood of being the quickpick sequence, from highest probability to lowest probability. After the process was complete participants were shown the actual lottery tickets so that, as promised, they could see if they had made the correct judgment or not.

Unfortunately for the experimenters, the lottery tickets did not turn out to be winning ones.

## Results and Discussion

An individual applying classical probability would view all sequences as equally likely and would thus only have a $20 \%$ chance of correctly identifying one quickpick sequence mixed with four others. However, $64 \%$ of participants correctly identified the numbers on the first ticket, and $66 \%$ on the second ticket (i.e. ranked these sequences in first place). When participants were shown the lottery tickets at the end of the experiment they were surprised that their intuition had, in the majority of cases, led them to make the correct choice.

Figure 1 shows the mean compressed bit size for sequences ranked from first to fifth place across the two presentations. The overall correlation between ranking and compressed description length was $0.965, p<.001$.


Figure 1: Mean compressed bit size according to rankings of likelihood.

These results demonstrate that, not only do people use subjective probability, they also enhance the accuracy of their judgments by using it. While the naive mathematician assumes all lottery sequences are equally likely, the savvy layperson realises there is an element of uncertainty involved in how those sequences were generated. The greater the randomness deficiency of a sequence, the greater the subjective probability that it was generated by a non-random generative mechanism.

Our central argument in this article is that, because models of reality always involve uncertainty, people apply subjective probability rather than classical probability in everyday life. In the following experiment we investigated whether the application of subjective probability can explain experimental observations which have previously been interpreted as examples of fallacious reasoning.

## Experiment 2

The conjunction effect is a situation in which people assert that a conjunction of two outcomes is more probable than either of those outcomes in isolation. According to classical probability theory this is a fallacy because requiring two outcomes to be validated is always a stricter criterion than requiring a single one to be validated (i.e. $P(x \wedge y) \leq P(y)$ ). The most celebrated example of the fallacy involves one of the materials used by Tverksy and Kahneman (1983), involving an individual named Linda.

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Which is more probable?
a) Linda is a bank teller
b) Linda is a bank teller and is active in the feminist movement.

Tverksy and Kahneman (1983) report that, when the two possible outcomes are listed together as above, $85 \%$ of people violate the conjunction rule by identifying $b$ ) as more probable. Tverksy and Kahneman's explanation is that people get confused by what they call 'representativeness'. They found that participants' responses reflect the extent to which the descriptions match a stereotype, with a correlation of 0.98 between mean ranks of probability and representativeness.

It is interesting to note that this correlation closely matches the observed correlation of 0.97 between mean ranks of probability and compressed description length in Experiment 1. This suggests the possibility that representativeness and randomness deficiency are closely related concepts.

In Experiment 1 we found that, when there is uncertainty as to the generative mechanism which produced an outcome, people rely on randomness deficiency to make judgments. The uncertainty in Experiment 1 concerned the fact that participants were given no information as to how four of the five lottery sequences were generated. Rather than assuming that all the sequences were generated randomly, they correctly used randomness deficiency to make inferences that resolved the uncertainty.

In the Linda example, some information about Linda is provided, but there is much about her that remains unknown (e.g. has she settled down since her student days?) In the case of uncertainty regarding the underlying probability measure function, then classical probability cannot be applied. For example, if we find out that Linda is a bank teller, then we might infer that she has settled down. In contrast, hearing that
she is still active in the feminist movement suggests that she has not changed much since her student days. Because these two models of Linda are quite different, there is no definitive probability measure function relative to which classical probability can be expressed.

## Method

In the following experiment we investigated whether the outcomes for the Linda scenario cause participants to adjust their model of Linda.

Materials For this experiment we altered the Linda scenario by including the outcomes as part of the description. We removed the information that she is single, outspoken and very bright and included at the end of the description either that "Linda is a bank teller" (Version 1) or "Linda is a bank teller and is active in the feminist movement" (Version 2). Participants were then asked to rate the probability of Linda having the attributes of being single, outspoken and very bright (from 0 to $100 \%$ ). In order for classical probability to be applicable, then the probabilities provided for Versions 1 and 2 should not differ significantly. Linda should be just as independent, outspoken and bright regardless of whether she is active in the feminist movement or not.

Participants 106 undergraduate students from NUI Maynooth participated voluntarily in this study.
Procedure Participants were randomly assigned either Version 1 or Version 2 of Linda's description and wrote down their probabilities for the three characteristics, which were randomly ordered along with three other filler characteristics (Linda plays golf, Linda is dyslexic, Linda suffers from anxiety).

## Results and Discussion

The mean probabilities for the three characteristics are shown in Table 2. When Linda was described as a bank teller and active in the feminist movement she was rated as significantly more likely to be single, demonstrating that the outcomes used in the Linda scenario cause participants to adjust their model of Linda.

Table 2: Mean probability ratings, t-test scores and significance for the two descriptions of Linda.

|  | Ver. 1 | Ver. 2 | t -test |
| :--- | :--- | :--- | :--- |
| Single | $47 \%$ | $64 \%$ | $t(104)=4.11, p<.001$ |
| Outspoken | $77 \%$ | $80 \%$ | $p>.05$ |
| Very Bright | $59 \%$ | $63 \%$ | $p>.05$ |

The numbers generated by a perfect dice never lead us to update our beliefs about the nature of the dice, yet finding out about Linda's current activities does lead people to update their beliefs about her. Because the model of Linda is uncertain, subjective probability must be applied. What people
are quantifying when they identify the conjunction as more probable is that the conjunction contains more subjective information, and that, relative to the process of representational updating, the likelihood of an outcome diminishes with the amount of subjective information it carries. Basing decisions on subjective probability is mathematically the correct approach when dealing with uncertainty regarding the underlying probability measure function.

In the following section we build on this result by proving that for every situation involving uncertainty (i.e. all real world scenarios) there is a conjunction of events which is more subjectively probable than either of its constituents in isolation.

## Proof that Conjunction Effect is not a Fallacy

In this section we prove that given any hypothetical model $p$, there are always two strings of events $x, y$ such that $x$ is a substring of $y$ but $y$ has higher subjective probability. The idea of the proof is that any long enough typical string of events can always be decomposed into a substring of events that carries greater subjective information.
Theorem 1. Let $E_{1}, E_{2}, \ldots E_{m}$ be $m$ independent events and let $p$ be the associated computable probability measure function. Let $\alpha>0$ be a surprise threshold. There exists a conjunction of events $A=A_{1} \wedge A_{2} \wedge \ldots \wedge A_{n}$ with a constituent $B$ (i.e. $p(A)<p(B))$ such that $B$ is $(p, \alpha)$-surprising (i.e. carries subjective information) and $A$ is $(p, \alpha)$-typical (i.e. has a subjective probability of 1).

Proof. Let $E_{1}, E_{2}, \ldots E_{m}, p$ and $\alpha>0$ be as above. Without loss of generality $m=2^{k}$ and $p$ can be seen as a probability on strings of length $k$ (each coding one event $E_{i}$ ) extended multiplicatively i.e., $p: 2^{k} \rightarrow[0,1]$ is extended multiplicatively by $p(x y):=p(x) p(y)$.

Let $n$ be a large integer. Let $y \in 2^{k n}$ be a $(p, \alpha)$-typical string. $y$ can be viewed as the concatenation of $n$ strings of length $k$ (i.e. the conjunction of $n$ events). By the pigeon hole principle, there must be such a string that occurs at least $n / 2^{k}$ times. Denote this string by $s$, and let $l$ be the number of occurences of $s$ in $y$, i.e. $l \geq n / 2^{k}$. Because $y$ is $(p, \alpha)$-typical we have $p(s)>0$. Thus $p(s)=2^{-c}$ for some $c>0$. Let $x$ be $l$ concatenations of $s$. Because $p$ is extended multiplicatively we have $p(x)>p(y)$.

Let us show that $x$ is $(p, \alpha)$-surprising. To describe $x$ it suffices to describe $l$ plus a few extra bits that say "print $s$ $l$ times". Since $l$ can be described in less than $2 \log l$ bits (by a prefix free program) we have $K(x)<3 \log l$ for $n$ large enough. We have

$$
\begin{aligned}
-\log p(x)-\alpha & =-\log p\left(s^{l}\right)-\alpha=-\log p(s)^{l}-\alpha \\
& =-l \log 2^{-c}-\alpha=c l-\alpha>3 \log l>K(x) \\
& \geq K\left(x \mid p^{*}\right)
\end{aligned}
$$

for $n$ large enough. Thus $x$ is $(p, \alpha)$-surprising, but $y$ is not.

## Conclusion

Although Tverksy and Kahneman (1983) identified an association between representativeness and the conjunction effect, they never provided an explanation for why such an association might exist, instead being satisfied to pass it off as an arbitrary reasoning fallacy. Had they questioned participants regarding their judgments, rather than dismissing them as fallacious, then the resultant findings may have facilitated the extension of classical probability theory. In sum, perhaps the most salient fallacy on display in Tverksy and Kahneman's (1983) study is the misplaced belief that mathematical theories which have been developed for precision models in the exact sciences retain their validity when used to describe complex cognition in the real world.

Tverksy and Kahneman (1983) posed the following question: "Why do intelligent and reasonably well-educated people fail to recognize the applicability of the conjunction rule in transparent problems?" Here, we have presented the answer: Because often it's not applicable.

## References

Costello, F. J. (2009). How probability theory explains the conjunction fallacy. Journal of Behavioral Decision Making, 22, 213-234.
Dessalles, J. L. (2006). A structural model of intuitive probability. In Proceedings of the seventh international conference on cognitive modeling (pp. 86-91). Trieste: Edizioni Goliardiche.
Gács, P., Tromp, J. T., \& Vityányi, P. M. B. (2001). Algorithmic statistics. IEEE Transactions in Information Theory, 47, 2443-2463.
Li, M., \& Vityányi, P. M. B. (2008). An introduction to Kolmogorov complexity and its applications. New York: Springer Verlag.
Maguire, R., Maguire, P., \& Keane, M. T. (2011). Making sense of surprise: An investigation of the factors influencing surprise judgments. Journal of Experimental Psychology: Learning,Memory and Cognition, 37(1), 176-186.
Rissanen, J. J. (1978). Modeling by the shortest data description. Automatica (Journal of IFAC), 14, 465-471.
Schmidhuber, J. (2009). Simple algorithmic theory of subjective beauty, novelty, suprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. Journal of the Society of Instrument and Control Engineers, 48(1), 2132.

Tverksy, A., \& Kahneman, D. (1983). Extentional vs. intuitive reasoning: The conjunction fallacy in probability judgment. Psychological Review, 90, 293-315.
Vityányi, P. M. B., \& Li, M. (2000). Minimum description length induction, Bayesianism, and Kolmogorov complexity. IEEE Transactions on Information Theory, 46, 446464.

# Dynamic Structure of Joint-Action Stimulus-Response Activity 

MaryLauren Malone (malonemo@mail.uc.edu)<br>Center for Cognition, Action \& Perception, University of Cincinnati, Cincinnati, OH, USA;

Ramon D Castillo (castilrn@mail.uc.edu)<br>Center for Cognition, Action \& Perception, University of Cincinnati, Cincinnati, OH, USA; Universidad de Talca, Chile

John G Holden (holdenjn@mail.uc.edu)
Center for Cognition, Action \& Perception, University of Cincinnati, Cincinnati, OH, USA;
Heidi Kloos (heidi.kloos@uc.edu)
Center for Cognition, Action \& Perception, University of Cincinnati, Cincinnati, OH, USA;
Michael J. Richardson (michael.richardson@uc.edu).
Center for Cognition, Action \& Perception, University of Cincinnati, Cincinnati, OH, USA;


#### Abstract

The mere presence of a co-actor can influence an individual's response behavior. For instance, a social Simon effect has been observed when two individuals perform a Go/No-Go response to one of two stimuli in the presence of each other, but not when they perform the same task alone. Such effects are argued to provide evidence that individuals co-represent the task goals and the to-be-performed actions of a co-actor. Motivated by the complex-systems approach, the present study was designed to investigate an alternative hypothesisthat such joint-action effects are due to dynamical (timeevolving) entrainment processes that perturb and couple the behavior of socially situated actors. To investigate this possibility, participants performed a standard Go/No-Go Simon task in joint and individual conditions. The dynamic structure of recorded response times (RTs) was examined using fractal statistics and instantaneous cross-correlation. Consistent with our hypothesis that participants responding in a shared space would become behaviorally coupled, the analyses revealed that RTs in the joint condition displayed decreased fractal structure (indicative of an interpersonal coupling perturbing and constraining participant behavior) compared to the individual condition, and were more correlated across a range of time-scales compared to the RTs of pseudo-pair controls. Collectively, the findings imply that self-organizing dynamic processes might underlie social stimulus-response compatibility effects and shape joint cognitive processes in general.


Keywords: joint action; stimulus-response compatibility; interpersonal coordination; pink noise; dynamical systems

## Introduction

Social interaction is a hallmark of everyday activity. Examples include a parent helping a child get dressed, a couple washing dishes together, people playing a team sport, or two workers carrying a heavy item up a flight of stairs. In each of these cases, a form of cooperation emerges such that the activity is coordinated across all participating
actors. Interestingly, coordination emerges even when no explicit coordination is required, for example when people are completing separate parts of the task. The present study aims to further investigate this latter form of coordination.

## Joint Stimulus-Response Compatibility (JSRC)

Over the past decade, a growing amount of research has been conducted investigating joint-action via so-called 'go/no-go tasks' (e.g., Sebanz, Knoblich, \& Prinz, 2005; Sebanz, Bekkering, \& Knoblich, 2006; Tsai, Kuo, Jing, Hung \& Tzeng, 2006). In such tasks, participants are instructed to 'go' when given a certain stimulus context (e.g., when they are presented with a red stimulus), and to 'not go' when given the alternative (e.g. a blue stimulus image). The compatibility aspect of these experiments lies in the spatial orientation of the stimulus relative to the location of the responding individual. For instance, if a stimulus is presented on the same side of a display with respect to where a participant is seated, the response is deemed "compatible". Alternatively, if a stimulus is presented on the opposite side of a display screen with respect to where a participant is seated, the response is deemed "incompatible".

To examine the effects of such stimulus-response mappings in a joint-action setting, the reaction times (RTs) are compared between two conditions: one in which the participant sits on one side of the display screen and responds alone to one stimulus type (the individual condition), and another where the task requirements are exactly the same except that another participant, seated on the opposite side of the display screen, responds to the alternative stimulus (the joint condition). The general finding is that even though participants in the joint condition are performing the exact same task as in the individual condition, a greater SRC effect exists when two people are completing the task in one another's presence compared to
when they complete the task alone. In other words, incompatible responses are significantly slower than compatible responses in joint conditions, but only marginally different (or not significantly different at all) for individual conditions.

These findings are generally taken as evidence for the co-representation of action goals during a joint-action setting, whereby actors form a shared representation of the collective task goal. That is, individuals mentally represent the actions of their co-actor and integrate them into their own action planning. This co-representation or action integration therefore results in slower RTs for incompatible stimulus situations compared to compatible stimulus situations. When completing the task alone, however, no such integration or co-representation occurs, and thus the spatial compatibility of the stimulus has little or no effect.

The JSRC effect has been observed across a wide range of stimulus and response manipulations, including hand posture (Cho, Proctor, \& Yamaguchi, 2007), non-biological response mechanisms (Buhlmann, Umilta, \& Wascher, 2007), orthogonality of stimulus location (Bae, Cho, \& Proctor, 2009; Figliozzi, Silvetti, Rubichi, \& Doricchi, 2010), and auditory stimuli (Buetti \& Kerzel, 2008). It is also known to be influenced by various social psychological variables, such as the facial features of a co-actor, and tasksharing paradigms (Philipp \& Prinz, 2010; Jung, Holländer, Müller, \& Prinz, 2011).

JSRC effects also appear to suggest that knowing what another person's task is during joint-action is the means by which an individual can understand others' action intentions and points to shared representations as the casual basis of this integration or modulation process. A consequence of this co- or shared-representation and action integration emphasis, however, is that no research has attempted to examine the time-evolution or behavioral dynamics of actors' responses during JSRC tasks, nor the degree to which JSRC effects are a result of the dynamical entrainment or coordination processes that are known to exist during co-present joint-action situations (Richardson et al., 2012; Schmidt et al., 2011). The aims of the present study were therefore to (i) examine the dynamical structure of JSRC task behavior and (ii) investigate whether the standard (visual) JSRC effect might be a result of dynamic coordination or entrainment processes coupling the response behavior of co-acting individuals.

## Examining the Dynamics of JSRC

At the crux of the traditional statistical analyses for JSRC experiments is a comparison of means, wherein each participant's time series of responses is represented as a single, unchanging number. The average RT response for each condition is understood as capturing the core and most meaningful aspect of the recorded RT behavior. The variability or time-evolution that occurs from trial-to-trial is simply discarded as error or mentioned only briefly in terms of how localized the mean is (for an exception see Vesper et
al., 2011). The temporal structure of RT variability (i.e., deviations from the mean over time), however, often provides additional and meaningful information about how behavior emerges over time (Gilden, 2001). For instance, there is evidence that the seemingly error-induced variation in responses may actually be reflective of how people execute discrete motor responses in a certain spatiotemporal context (Wing \& Kristofferson, 1973). Furthermore, even if the mean value and standard deviation are the same, the structure of RT time series that result in those means and standard deviations could in fact be quite different.

In order to examine the dynamic structure and unfolding variability of RTs over time, recent research has utilized fractal methods that provide deeper insight into the dynamics of an ongoing activity (Bassingthwaighte, Liebovitch, \& West, 1994; Gilden, Thornton, \& Mallon, 1995; Jensen, 1998; Van Orden, Holden, Turvey, 2003). Conceptually similar to geometric fractal patterns (Mandelbrot, 1982), fractal patterns in experimental timeseries data correspond to nested patterns of variability found across repeatedly measured behaviors. Instead of comparing the overall means, fractal analysis involves determining how the variability exhibited in a time-series changes with changes in time-scale. That is, fractal analysis involves determining if the structure of variability in an RT timeseries is statistically self-similar or scale invariant, such that small variations in the data have essentially the same structure as large variations (Brown \& Liebovitch, 2010; West \& Deering, 1995). As in geometrical fractal patterns, if one were to "zoom in" (i.e., examine a smaller scale) on the measurement time-series, one would discover essentially the same pattern of fluctuations evident at the larger scale (Holden, 2005). Accordingly, fractal statistical methods do not rely on partitioning the variability in measurement into different components, but rather assess the structure of the time-evolving variability observed.

A time-series containing random fluctuations (i.e., white noise) indicates that the observed variability is the result of unsystematic or unrelated changes from trial to trial (Van Orden, 2010). Alternatively, the variability in an RT time-series containing fractal or scale invariant structure contains trial-to-trial variability that is long-term correlated. In other words, the time-series contains nested patterns of variability wherein small variations in measurement have the same structure as large variations. Such structure in repeated measurements is often referred to as "pink noise" or $1 / f$ noise and are characteristic of a wide range of naturally occurring complex (interaction-dominant) systems and phenomena, from eye movement patterns (Aks, Zelinsky, \& Sprott, 2002) and heart rate variability (Eke et al., 2002), to self-reported mood change (Delignières, Fortes, \& Ninot, 2004).

There are numerous methods for determining the degree to which the variability in a behavioral or response time-series is scale invariant or pink (see Delignières et al., 2006 for a review). One of the most robust methods is detrended fluctuation analysis, commonly referred to as

DFA (Bassingthwaighte et al., 1994; Peng, Havlin, Stanley, \& Goldberger, 1995). DFA quantifies the long-term correlative properties of behavior by detrending the time series of adjacent bins, or collections of consecutive data points, at all time scales. The residual variance obtained from the least-square regression line subtraction of each bin is calculated for progressively larger bin sizes. Bin size is plotted against variance on a log-log plot, and the scaling exponent, $H$, is revealed by the slope of the best-fitting line. For DFA, $H \approx 1.0$ indicates that the response variability or "noise" is pink (i.e., fractal). White noise, however, corresponds to $H=0.5$.

Deviations away from 'perfect' pink noise (i.e., $H=$ 1.0) can result from changes in system flexibility (Kloos \& Van Orden, 2010). For instance, increasing task constraints or difficulty, such as coupling responses to external timers or events (i.e., metronomes), or increasing task speed, can whiten RT variability and result in $H \ll 1.0$ (Chen, Ding, \& Kelso, 2001; Delignières et al., 2009; Hausdorff et al., 1996). Changes in $H$ across conditions thus reveal how differing task manipulations result in processes that interact or constrain each other, as well as influence the overall organizational processes that underlie a series of behavioral responses (Van Orden, 2010). Accordingly, the question considered here was does the co-presence of an actor during a JSRC task change the fractal structure of an individual's RT behavior, and if so, how and why?

One possibility is that the behavior of individuals during joint-action conditions are subtly coupled or dynamically entrained and that this coupling or entrainment acts to constrain and/or perturb the behavioral responses of the individuals involved. There is a significant body of research demonstrating how the behaviors of co-present individuals often become dynamically coordinated or entrained (see Schmidt \& Richardson, 2008; Marsh, Richardson, \& Schmidt, 2009 for reviews) and that such processes can modulate and perturb individual behavior (Richardson et al., 2009, 2012; Riley, et al., 2011; Romero et al., 2012). If this is the case, then the fractal structure of the RT variability should be whiter in the joint condition compared to the individual condition.

To explore this possibility, we employed a standard SRC task, the Simon task (Craft \& Simon, 1970), and had participants complete the task under joint and individual go/no-go conditions. We performed a fractal analysis on the resting RT time-series using DFA, with the expectation that the joint condition would exhibit a whiter fractal structure ( $H$ closer to 0.5 ) compared to RT time-series in the individual condition. In addition to performing a fractal analysis, we also employed instantaneous cross-correlation (Barbosa, Yehia, \& Vatikiotis-Bateson, 2008) to index the degree to which the RTs of co-acting individuals were correlated (i.e., coordinated) with each other over time. If the behavioral responses of individuals are entrained during a joint-action situation, then the temporal correlation should be greater between the RT time series of individuals in the joint condition compared to RT time series of pseudo-pairs
created using RT time-series from participants who performed the task in the individual condition. We employed instantaneous cross-correlation because it allows one to determine how correlated two behavioral time-series are across multiple time-scales. The method is ideally suited for determining highly subtle non-synchronous coordination that occurs at variable time-lags. It essentially computes the correspondence between two signals recursively, generating a time-series of how past and future samples are correlated at all points in time. Setting a minimum $r$ value as a cut-off for what is considered to be correlated or not (i.e., $r=.25$ ) then allows one to calculate the percentage of points that resulted in correlation values greater than that cutoff. The resultant value is the proportion of correlated activity and can be understood as providing a measure of percent coupling.

## Methods

## Participants

Twenty-four undergraduate students from the University of Cincinnati ( 7 male, 17 female) participated in the study. They ranged in age from 18 to 22 years old and received class credit for participation in the experiment.

## Materials

A 19" Dell Flat Panel monitor was used to present stimuli. Stimuli included a blue " X " or red " X " ( 1 " high, $1 / 2$ " wide), displayed on the left or right of the screen (positioned $51 / 2$ " from the top and bottom of the screen, and 2 " from the left or right side of the screen, respectively). Stimulus presentation and data collection was controlled using Direct RT. An Apple keyboard, modified to be millisecond accurate, was used to collect reaction time data. The shift keys were used as response indicators on the keyboard. A red sticker was placed on the right shift key and a blue sticker was placed on the left shift key. The monitor and the keyboard were placed in the center of a desk, with the keyboard 7" from the front of the desk and 8 " from the monitor. Participants were seated in chairs that were placed next to each other in front of the keyboard. Each seated participant was positioned approximately 30 " from the display screen.

## Procedure

Participants completed a visual go/no-go Simon task in which they were instructed to respond with a key press to a specific stimulus color presented on the screen. Participants were assigned only one of the two stimulus colors (e.g. red) and were instructed to respond only to their designated color, regardless of location, while ignoring the alternative (e.g. blue). Participants completed the task in one of two experimental conditions: a joint condition or an individual condition. For the individual condition, participants performed the task alone. For the joint condition pairs of participants performed the task together. Similar to the procedure of Sebanz et al. (2003), subjects assigned to the
red key sat on the right, and subjects assigned the blue key sat on the left, regardless of condition (see Figure 1). A brief instruction screen was presented on the computer monitor prior to the start of the experiment. Clarifying instructions were administered verbally and an opportunity for questions or clarification was offered.

63).


Figure 2: Mean reaction time (RT) as a function of experimental condition and compatibility.

## Fractal Analysis

DFA was performed on the last 512 responses for each participant. Prior to analysis, the RTs were normalized by subtracting the relevant condition means for each participant in order to examine the variability of the residual fluctuations (see Gilden, 2001 for a detailed description of the rationale). Consistent with our hypothesis that participants responding in the joint condition would exhibit a whiter fractal structure of responses due to task constraints and coupling, a between samples one tailed $t$-test performed on $H$ values calculated using DFA revealed a significant effect of experimental condition, $t(22)=2.25, p<.05$, with the fractal structure of RTs in the joint condition being significantly lower $H(M=0.57, S D=0.06)$ than in the individual condition ( $M=0.63, S D=0.08$ ) (see Figure 2). One sample $t$-tests indicated that $H$ values were significantly different from a test value of 0.5 (hypothetical white noise) for both the individual, $t(11)=5.93, p<.01$, and the joint conditions, $t(11)=34.53, p<01$.


Figure 3: Mean Hurst $(H)$ as a function of experimental condition.

## Instantaneous Correlation

To determine the degree to which the RT time-series of participants in the joint condition were entrained or coupled to each other over time, we calculated the percentage of correlations within the time-series of instantaneous correlations for delays of -60 to 60 trials that had an $r>.25$. As mentioned above, the resultant value can be understood as a measure of percent coupling or the proportion of
correlated activity. We then used a between samples onetailed $t$-test to compare the percent coupling observed between participants in the joint condition to the percent coupling calculated between pseudo pairs of participants created by randomly pairing participants from the individual condition. Consistent with the hypothesis that the behavioral response of participants in the joint condition might be dynamically entrained or coupled, the analysis revealed that the percent coupling for the joint condition (\%30.9) was (marginally) significantly greater, $t(22)=1.65, p=.059$, compared to pseudo pairs (\%21.6) created from participants in the individual condition (see Figure 3).


Figure 4: The percent coupling calculated using instantaneous cross correlation as a function of experimental condition.

## Discussion

The experimental study present here was aimed at examining the behavioral dynamics of individuals during a joint-action stimulus-response compatibility task. We submitted recorded RT time-series during a JSRC task to both a standard comparison of means, and to various dynamical analysis methods in order to examine how RT variability evolved over time. We compared these patterns of variability between joint and individual conditions.

Consistent with previous research, we found a significant difference in the overall reaction times between the individual and joint conditions, as well as a significant compatibility effect in the joint condition. More importantly, by measuring the fractal structure of participants' RTs, we found that that the structure of variability in the joint condition was much whiter than in the individual condition, as predicted. The current results therefore extend previous research by demonstrating that the mere presence of another individual not only affects average RT, but also affects the dynamics of an individual's response behavior. This difference was theorized to be a consequence of the dynamical entrainment processes that mutually perturb and constrain the behavior of individuals in a shared environment (Schmidt \& Richardson, 2008).

To further examine whether the response behaviors of participants were dynamically coupled, an instantaneous correlation analysis was performed. We compared the degree to which the RT behavior of pairs in the joint condition was correlated to the degree of RT correlations that occurred for pseudo pairs created from participants who
completed the individual condition. The results of this analysis revealed that the response behavior of pairs in the joint condition exhibited greater temporal correlation compared to pseudo pairs, providing more evidence that the response behaviors of co-present individuals in the current go/no-go task were dynamically entrained. The magnitude of these temporal correlations was by no means large and occurred at non-synchronous time-lags. Thus, like most other forms of interpersonal entrainment or behavioral coupling (see e.g., Chartrand and Bargh, 1999; Schmidt \& Richardson, 2008; for reviews), the entrainment that occurred was most likely intermittent, rather than constant, and did not occur synchronously or at any fixed time lag. The weak and complex nature of the interpersonal influence should not be discounted, however, given the fact that the mean differences in RT are also relatively small (as is typically the case JSRC studies). Indeed, the relative change in mean RT, fractal dimension ( $H$ ), and \% coupling are all somewhat equivalent.

In conclusion, the current study provides the first evidence that the response behavior of co-actors during a JSRC task is dynamically entrained and that such dynamical entrainment processes operate to constrain and perturb the time-evolving response variability of co-acting individuals. Although not directly tested here, it is possible that these dynamic processes of constraints and coupling may underlie the JSRC effect, rather than some form of shared representation. In truth, the dynamical systems and representational accounts of such behavior are not mutually exclusive and may in fact provide complementary explanations for such joint-action phenomena. Future research should be directed towards investigating these issues in order to better understand how the dynamics of joint-action activity shape joint cognitive processes.

## Acknowledgements

We wish to acknowledge Guy Van Orden for his invaluable contribution to the field of complex systems.

## References

Aks, D.J., Zelinsky, G.J., \& Sprott, J.C. (2002). Memory across eye movements: $1 / f$ dynamic in visual search. Nonlinear Dynamics, Psychology, and Life Sciences, 6, 125.

Bae, G.Y., Cho, Y.S., \& Proctor, R.W. (2009). Transfer of orthogonal stimulus-response mappings to an orthogonal Simon task. The Quarterly Journal of Experimental Psychology, 62(4), 746-765.
Barbosa, A.V., Yehia, H.C., \& Vatikiotis-Bateson, E. (2008). Algorithm for Computing Spatiotemporal Coordination.
Bassingthwaighte, J. B., Liebovitch, L. S., \& West, B. J. (1994). Fractal physiology. New York: Oxford University Press.
Brown, C., \& Liebovitch, L. (2010). Fractal Analysis. Los Angeles: Sage.

Buetti, S., \& Kerzel, D. (2008). Time course of the Simon effect in pointing movements for horizontal, vertical, and acoustic stimuli: Evidence for a common mechanism. Acta Psychologica, 129, 420-428.
Buhlmann, I., Umilta, C., \& Wascher, E. (2007). Response coding and visuomotor transformation in the Simon task: The role of action goals. Journal of Experimental Psychology, 33(6), 1269-1282.
Chen Y., Ding M., \& Kelso J.A.S. (2001). Origins of timing errors in human sensorimotor coordination. Journal of Motor Behavior 33, 3-8.
Cho, Y.S., Proctor, R., \& Yamaguchi, M. (2007). Influences of response position and hand posture on the orthogonal Simon effect. The Quarterly Journal of Experimental Psychology, 61(7), 1020-1035.
Craft, J.L., \& Simon, J.R. (1970). Processing symbolic information from a visual display: Interference from an irrelevant directional cue. Journal of Experimental Psychology, 83(3), 415-420.
Delignières, D., Fortes, M., \& Ninot, G. (2004). The fractal dynamics of self-esteem and physical self. Nonlinear Dynamics in Psychology and Life Sciences, 8, 479-510.
Delignières D., Ramdani, S., Lemoine, L., Torre, K., Fortes, M., \& Ninot, G. (2006). Fractal analyses for 'short' time series: A re-assessment of classical methods. Journal of Mathematical Psychology, 50, 525-544.
Delignières D., Torre K., \& Lemoine L. (2009). Long-range correlation in synchronization and syncopation tapping: A linear phase correction model. PLoS ONE, 4(11): e7822.
Figliozzi, F., Silvetti, M., Rubichi, S., \& Doricchi, F. (2010). Determining priority between attentional and referential-coding sources of the Simon effect through optokinetic stimulation. Neuropsychologia, 48, 10111015.

Gilden, D.L., Thornton, T., \& Mallon, M.W. (1995). 1/f noise in human cognition. Science, 267, 1837-1839.
Gilden, D.L. (2001). Cognitive emissions of $1 / f$ noise. Psychological Review, 108(1), 33-56.
Hausdorff J.M., Purdon P.L., Peng C.K., Ladin Z., Wei J.Y., \& Goldberger A.L. (1996). Fractal dynamics of human gait: Stability of long-range correlations in stride interval fluctuations. Journal of Applied Physiology 80(5), 14481457.

Holden, J. G. (2005). Gauging the fractal dimension of response times from cognitive tasks. In M. A. Riley \& G. C. Van Orden (Eds.), Contemporary Nonlinear Methods for Behavioral Scientists: A webbook tutorial (pp. 267318).

Jung, C., Holländer, A., Müller, K., \& Prinz, W. (2011). Sharing a bimanual task between two: evidence of temporal alignment in interpersonal coordination. Experimental Brain Research, 211, 471-482.
Kloos, H., \& Van Orden, G. (2010). Voluntary Behavior in Cognitive and Motor Tasks. Mind \& Matter, 8(1), 19-43.
Marsh, K.L., Richardson, M.J., \& Schmidt, R.C. (2009). Social connection through joint action and interpersonal coordination. Topics in Cognitive Science, 1, 320-339.

Peng, C.K., Havlin, S., Stanley, H.E., \& Goldberger, A.L. (1995). Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. Chaos, 5(1), 82-87.
Philipp, A.M., \& Prinz, W. (2010). Evidence for a role of the responding agent in the joint compatibility effect. The Quarterly Journal of Experimental Psychology, 63(11), 2159-2171.
Richardson, M. J, Campbell, W.L, \& Schmidt, R. C. (2009). Movement interference during action observation as emergent coordination. Neuroscience letters, 449(2), 11722.

Richardson, M. J., Harrison, S. J., May, R., Kallen, R. W. \& Schmidt, R. C. (2012). Self-Organized Complementary Coordination: Dynamics of an Interpersonal CollisionAvoidance Task. BIO Web of Conferences 1, 00075.DOI: 10.1051/bioconf/20110100075.

Romero, V., Coey, C., Schmidt, R. C., \& Richardson, M. J. (2012). Movement Coordination or Movement Interference: Visual Tracking and Spontaneous Coordination Modulate Rhythmic Movement Interference. PLoS ONE 7(9): e44761.
Schmidt, R. C., \& Richardson, M. J. (2008). Dynamics of Interpersonal Coordination. In A. Fuchs \& V. Jirsa (Eds.). Coordination: Neural, Behavioral and Social Dynamics. (281-308)
Schmidt, R.C., Fitzpatrick, P., Caron, R., \& Mergeche, J. (2011). Understanding social motor coordination. Human Movement Science, 30, 834-845.
Sebanz, N., Knoblich, G., \& Prinz, W. (2003). Representing others' actions: Just like one's own? Cognition, 88, 11-21.
Sebanz, N., Knoblich, G., \& Prinz, W. (2005). How two share a task: Corepresenting stimulus-response mappings. Journal of Experimental Psychology, 31(6), 1234-1246.
Sebanz, N., Bekkering, H., \& Knoblich, G. (2006). Joint action: Bodies and minds moving together. TRENDS in Cognitive Science, 10(2), 70-76.
Tsai, C.C., Kuo, W.J., Jing, J.T., Hung, D.L., \& Tzeng, O.J.L. (2006). A common coding framework in self-other interaction: Evidence from joint action task. Experimental Brain Research, 175, 353-362.
Van Orden, G.C., Holden, J.G., \& Turvey, M.T. (2003). Self-organization of cognitive performance. Journal of Experimental Psychology: General, 132(3), 331-350.
Van Orden, G. (2010). Voluntary Performance. Medicina, 46(9), 581-594.
Vesper, C., van der Wel, R.P.R.D., Knoblich, G., \& Sebanz, N. (2011). Making oneself predictable: reduced temporal variability facilitates joint action coordination. Experimental Brain Research, 211, 517-530.
West, B. J., \& Deering, B. (1995). The lure of modern science. River Edge, NJ: World Scientific.
Wing, A.M., \& Kristofferson, A.B. (1973). Response delays and the timing of discrete motor responses. Perception and Psychophysics, 14(1), 5-12.

# Placing Numbers in Behavioral Space: Activity-Specific Interactions between Number and Space with a Single Response Button 

Tyler Marghetis* (tmarghet@ucsd.edu)<br>Jasmeen Kanwal* (jkanwal@ucsd.edu)<br>Benjamin K. Bergen (bkbergen@ucsd.edu)

University of California, San Diego
9500 Gilman Drive, La Jolla, CA 92093 USA


#### Abstract

How are we able to reason about abstract concepts that lie resolutely beyond the reach of perception? One strategy is to ground understanding in space. Numbers, for instance, are known to interact with egocentric space during rapid numerical judgments. A range of experimental results have demonstrated that, among literate Western people, this "mental number-line" goes from left to right, with smaller numbers associated with left space, and larger numbers with right space. But what is the nature of this "space"? Previous work has conflated multiple possible egocentric frames of reference-head-based, eye-based, action-based-leaving it unclear which space is interacting with number. In the present paper, two studies investigated whether a single centrallylocated button, stationary in hand- and eye-based coordinates, can nevertheless exhibit different spatial properties in virtue of task-specific activity. In a go/no-go paradigm, participants judged the magnitude (Exp. 1) and parity (Exp. 2) of singledigit numbers. Crucially, they responded only with the index or middle finger of a single hand. While judging magnitude (Exp. 1), participants were faster to respond to smaller numbers with the more leftward finger, and larger numbers with the more rightward finger, regardless of the hand being used. This effect disappeared when judging parity (Exp. 2), replaced by finger-specific associations on the left hand only. In sum, in a task-sensitive way, participants associated numbers with egocentric space-but a behavioral space defined relative to embodied interaction rather than head- or eye-based reference frames. We discuss implications for number representation and the nature of "space" in embodied activity.


Keywords: number; space; action; SNARC; embodiment; go/no-go; frames of reference; Merleau-Ponty

## Introduction

"[The body's] spatiality is not, like that of external objects or like that of 'spatial sensations', a spatiality of position, but a spatiality of situation."- Merleau-Ponty (1962, p. 114).

How are we able to reason about abstract concepts that lie resolutely beyond the reach of perception? One strategy is to ground understanding in space. Numbers, for instance, are tightly linked with space across human activity. We recycle the language of space to talk about numbers, counting $u p$ to higher numbers and down to lower numbers, and use space

[^145]to reason about numbers as abstract concepts (Lakoff \& Núñez, 2000; Núñez \& Marghetis, to appear). Mathematical diagrams often associate numbers with particular locations. And number interacts with space in less explicit ways during the online performance of mathematical activities. In a seminal study, Dehaene and colleagues (1993) asked participants to judge the magnitude (greater or less than 5?) or parity (even or odd?) of single digit numbers. Participants were reliably faster to respond to smaller numbers when responding with a button in left space, and to larger numbers when responding in right space-as if they were spontaneously thinking of numbers along a left-to-right "mental number line." This interaction between numerical magnitude and spatial location has been dubbed the "SNARC" effect.

In the two decades since, the literature on such numberspace associations has exploded (Hubbard et al., 2005; Wood et al., 2008). Similar effects have been found with bipedal responses (Schwarz \& Müller, 2006) and saccades (Fischer et al., 2003; Schwarz \& Keus, 2004), further reinforcing the genuinely spatial nature of this effect. While the particular direction of this "mental number-line" is quite flexible, shaped by such factors as habitual reading direction (Shaki, Fischer, \& Petrusic, 2009) and recent experience (Fischer, Mills, \& Shaki, 2010), there is a growing consensus that number and space are intimately related.

But doubts remain. A number of authors have suggested that the effects have less to do with a stable spatial representation of number, and more to do with flexible or non-spatial associations (e.g., Fischer et al., 2010; Gevers et al, 2010; Santens and Gevers, 2006). Others have pointed out that most studies force participants to respond spatially, using buttons that are distinguished by their spatial location, and thus implicitly inject space in virtue of the experimental setup (Núñez, Doan \& Nikoulina, 2011). Indeed, when other response modalities are used-e.g., responding with higher or lower pitches-participants exhibit interactions that are SNARC-like but non-spatial (Marghetis et al., 2011). Number-space associations, therefore, may be more flexible and context-sensitive than first assumed.

Beyond these concerns, one additional question has been largely unaddressed: the nature of the "space" that sometimes, undeniably, interacts with number. Previous work has conflated the multiple egocentric frames of reference that we use to encode space, which include head-,
eye-, object-, and action- based frames of reference (cf. Cohen and Anderson, 2002). For instance, in the classic SNARC paradigm, the response buttons are placed to the left and right of the participant's body, and thus are distinguished similarly in multiple frames of reference: they are lateralized relative to participants' heads, eyes, and motor responses. When space is forced on the subject in this way, through a forced-choice paradigm where responses are spatially distinguished, it remains unclear as to which "space" is interacting with number.

One possibility is that numbers interact with behavioral space, i.e. the space defined relative to task-specific embodied activity. If this is the case, then observation of the SNARC effect may not depend on the presence of spatial responses distinguished externally in head- or eye-based coordinates, but rather on some juxtaposition of taskrelevant actions within the specific context of the task. Testing this possibility requires an experimental design that can isolate effects of behavioral space from those due to head- or eye-based coordinates. We took up this challenge by investigating whether SNARC-like effects could be elicited using only a single, centrally-placed response button, which remained in the same location relative to head- and eye-based spatial coordinates. By requiring participants to press the button with different fingers of the same hand, the stationary button can move relative to taskrelevant embodied activity-that is, in behavioral space.

In two studies, participants judged the magnitude (Exp. 1) and parity (Exp. 2) of single-digit Arabic numerals. Crucially, and in contrast to previous studies, we used a go/no-go paradigm in both tasks: participants responded by pressing a single centrally-located button with only their index or middle finger. The finger used was manipulated between blocks, as was the "go" response criterion, while response hand was varied between subjects. By using a central button, kept stationary relative to head- and eyebased frames of reference, and by using a go/no-go paradigm in which different response fingers were never juxtaposed spatially within a block, but only temporally across blocks, we were able to focus on one particular spatial frame of reference: behavioral space, the space defined by the possible embodied actions within the task as a whole. Moreover, by manipulating the response finger,


Figure 1: Paradigm for Experiments 1 and 2. Here, the "Go" criterion would be to respond with the index finger of the left hand, if less than 5 (Exp. 1) or even (Exp. 2).
identified only by name (e.g. "index finger"), we were able to avoid any explicit spatial instructions.

If the SNARC effect is driven entirely by visual head- or eye-based frames of reference, then participants should not exhibit any SNARC-like effects here, since at no point are two response options spatially juxtaposed relative to head or eye. If, on the other hand, number interacts with space as enacted by task-specific interactions-that is, behavioral space-then we may find numbers systematically associated with space relative to activity-based frame of reference, with the more leftward finger faster for smaller numbers, and the more rightward finger faster for larger numbers.

## Experiment 1: Magnitude Task

## Participants

Undergraduate students ( $\mathrm{n}=32$, mean age $=20,19$ females) from a major research university participated in exchange for partial course credit.

## Procedure

In a go/no-go paradigm, participants judged the relative magnitude of visually-presented single-digit numerals, responding only if the number presented was greater than [/less than] 5. Participants responded by pressing a single, centrally-located button on a Serial Response Box placed at a comfortable distance in front of them, using either the index finger or middle finger. Response finger (index or middle) and response criterion (greater or less than 5) were fully crossed within participants, so each run had four blocks. Block order was counterbalanced, except that no two consecutive blocks used the same response finger (to avoid muscle fatigue). Response hand (left, right) varied between participants, so each participant maintained the same response hand throughout the experiment. Each participant, therefore, responded with their index and middle fingers for two blocks each.

Each trial began with a central fixation cross ( 500 ms .), followed by a centrally-presented single-digit number between 1 and 9 (excluding 5). If participants responded, the number would disappear; otherwise it would remain on the screen for 3 s., after which the trial would end automatically. See Figure 1. Each block began with 8 practice trials, followed by 80 experimental trials.

## Results

One participant was removed for failing to complete the experiment. Accuracy for the remaining 31 participants was high ( $\mathrm{M}>.99, \mathrm{SD}=.004$ ). Mean accuracy was analyzed with a $2 \times 2 \times 2 \times 2$ mixed ANOVA, with Magnitude (greater or less than 5), Parity (even, odd), and Response Finger (leftfinger, right-finger) as within-subjects factors, and Response Hand (left or right) as a between-subjects factor ${ }^{2}$. There were no significant effects on accuracy.

[^146]Before analyzing response time, incorrect trials were removed, followed by trials with reaction times that were slower than three standard deviations above each participant's mean response time, or faster than 200 ms ( $\mathrm{n}=83,1.7 \%$ of total trials).

Reaction times were analyzed with a $2 \times 2 \times 2 \times 2$ mixed ANOVA, with Magnitude (greater or less than 5), Parity (even, odd), and Response Finger (left-, right-finger) as within-subjects factors, and Response Hand (left, right) as a between-subjects factor. There was a main effect of Parity, with responses to odd numbers significantly faster than responses to even numbers ( $\mathrm{M}_{\mathrm{odd}}=382 \mathrm{~ms}, \mathrm{M}_{\text {even }}=398 \mathrm{~ms}$, $F(1,29)=54.8, p<0.001$; cf. Dehaene et al, 1993, who found no effect of Parity). There was also a marginally significant-but difficult to interpret-interaction between Parity and Response Hand $(F(1,29)=4.1, p=.053)$.

Crucially, the only other significant effect was an interaction between Magnitude and Response Finger ( $F(1$, 29) $=4.95, p=0.034, \eta_{\mathrm{p}}^{2}=.015$ ). Responses with the leftfinger (i.e. index finger of right hand, or middle finger of left hand) were faster for numbers less than 5 , while responses with the right-finger were faster for numbers greater than 5 (Fig. 2). There was no three-way interaction between Magnitude, Response Finger, and Response Hand $(F(1,29)=0.5, p=0.46)$, suggesting that the effect is due not to finger-specific associations, but to the location of the response fingers in each participant's behavioral space. The interaction between Magnitude and Response Hand, notably, was not significant $(F(1,29)=2.94, p=0.10)$.

## Discussion

When judging the magnitude of single-digit numbers, participants systematically associated smaller numbers with the left, and larger numbers with the right, even though the response button did not change location relative to head- or eye-based coordinates. Rather, magnitude was associated with locations in behavioral space, defined by the possible actions within the task: responding with one of two possible fingers. This was confirmed by the lack of a three-way interaction with Response Hand. In other words, numbers were not associated with particular fingers, as we might expect if participants were using a body-based frame of reference (e.g. DiLuca et al, 2006). Instead, they were associated with particular actions relative to task behavior: responses with the leftmost finger compared to responses with the rightmost finger. Moreover, this interaction arose despite the fact that left- and right-fingered responses were never juxtaposed within a single trial or block, but only manipulated between blocks and thus juxtaposed within the experiment as a whole.

Notably the interaction between Response Hand and Magnitude was not significant, contra the results of studies that have used bimanual responses (e.g. Dehaene et al, 1993; for review, see Wood et al, 2008). We attribute this to the fact that we manipulated response hand between subjects, not within. Responses with the left- and right-


Figure 2: Interaction between magnitude and finger-side in Experiment 1. (Error bars $=$ SE)
hands, therefore, were not contrasted within participants' task-internal embodied activity.

Does this effect arise automatically? The classic SNARC is often found even when magnitude is task-irrelevant, for instance when determining parity (even vs. odd) (Dehaene et al, 1993). This is taken to show that the interaction between magnitude and space is automatic and taskindependent, at least when responses involve lateralized buttons. Is this new action-based SNARC similarly automatic, or does it require explicit magnitude processing? Experiment 2 was designed to answer this question.

## Experiment 2: Parity Task

## Participants

Undergraduate students ( $\mathrm{n}=32$, mean age $=21,22$ females) from a major research university, who had not participated in the first experiment, participated in exchange for partial course credit.

## Procedure

The procedure was identical to Experiment 1, except participants had to respond based on the parity (even vs. odd) of each number, rather than the magnitude. For a given block, therefore, participants would only respond if the number was even [/odd] (Fig. 1). All nine numbers from 1 to 9 were used as stimuli, so each block began with 9 practice trials followed by 90 experimental trials.

## Results

Accuracy was high ( $\mathrm{M}=.99, \mathrm{SD}=.01$ ). Mean accuracy was analyzed with a $2 \times 2 \times 2 \times 2$ mixed ANOVA, with Magnitude (greater or less than 5), Parity (even, odd), and Response Finger (left-finger, right-finger) as within-subjects factors, and Response Hand (left or right) as a betweensubjects factor. The only significant effect was a main effect of Parity $(F(1,30)=10.1, p<0.01)$, with responses to odd numbers more accurate than those to even numbers $\left(\mathrm{M}_{\text {odd }}=98.6 \%, \mathrm{M}_{\text {even }}=97.6 \%\right)$.

Once again, before analyzing response times, incorrect trials were removed, followed by trials with reaction times that were slower than three standard deviations above each participant's mean response time ( $\mathrm{n}=85,1.7 \%$ of trials). Finally, trials where the target numeral was 5 were also removed, so we could include Magnitude (greater or less than 5) as a factor in our analysis.

In contrast with Experiment 1, numbers were not reliably associated with response side. Instead, while there was a marginal interaction between Magnitude and Finger Side $(F(1,30)=3.3, p=.08)$, this was complicated by a significant three-way interaction with Response Hand $(F(1,30)=9.3$, $p=0.005$ ). There was a similar effect for Parity: while there was no two-way interaction between Parity and Finger Side $(F(1,30)=2.1, p=0.16)$, there was a significant three-way interaction with Response Hand $(F(1,30)=4.9, p=.035)$. For both Magnitude and Parity, the effect was driven by the left hand, where the index finger was faster for odd or smaller numbers, and the middle finger was faster for even or larger numbers (Fig. 3). The only other significant effects were main effects of Parity and Magnitude $(F(1,30)=9.0$, $p=0.005$, and $F(1,30)=5.05, p=0.03$, respectively), and a hard-to-interpret three-way interaction between Magnitude, Parity, and Hand $(F(1,30)=4.2, p=0.05)$.

That the effect was driven by the left hand was confirmed by separate repeated-measures ANOVAs for each hand, with Response Finger (index, middle), Parity, and Magnitude as factors. For the right hand, Response Finger did not interact with Parity $(F(1,15)=0.2, p=0.90)$ or with Magnitude $(F(1,15)=0.02, p=0.88)$. But for the left hand there were significant interactions between Response Finger and Magnitude $(F(1,15)=5.07, p=0.039)$, and between Response Finger and Parity $(F(1,15)=7.1, p=.02)$.

## Discussion

When tasked with determining the parity of single-digit numbers, participants no longer exhibited the SNARC-like
 and magnitude (bottom), found on the left hand only, in Experiment 2. $($ Error bars $=\mathrm{SE})$.
effect found in Experiment 1. Instead, we found fingerspecific associations with both parity and magnitude, but only on the left hand. A change in task, therefore, induced new associations with numbers, tied to specific fingers rather than to a more general behavioral space. We return to the possible origins of these finger-specific associations in the General Discussion.

## General Discussion

What space do numbers inhabit? We conducted two studies to investigate the possibility that the form of "egocentric" space with which number interacts is behavioral space, the space of possible embodied interaction with the world. Indeed, contrary to what we would expect if number-space interactions are driven entirely by head- or eye-based coordinates, we found that numbers interacted with space even when participants responded with only a single, centrally-located button. This effect, moreover, was not driven entirely by body-based representations, since they were not specific to particular fingers or hands. Instead, responses to smaller numbers were faster with the more leftward finger of either hand, while responses to larger numbers were faster with the more rightward finger-a left-to-right "mental number line" defined entirely in terms of the embodied interaction between finger and apparatus.

This effect, however, was task-dependent, and disappeared when magnitude was not task-relevant. Instead, when participants were attending to parity, they associated specific fingers of the left hand with parity and magnitude: the index finger with odd or small numbers, and the middle finger with even or large numbers. The space with which numbers interacted, therefore, was flexibly tied to body and activity in a task-specific way.

## Task differences and finger-based representations

What might account for the different results of Experiments 1 and 2 ? One possibility is that parity and magnitude tasks require participants to attend to different information. Parity tasks, for instance, may activate linguistic and categorical representations, while magnitude tasks may activate analog visuospatial representations (van Dijk, Gevers \& Fias, 2009). It may be that the rather subtle spatial difference between fingers is sufficiently small that an interaction between magnitude and space requires the explicit activation of analog visuospatial representations of magnitude. Alternatively, since the classic bimanual Parity task explicitly distinguishes the response options by their positions on the left and the right, those linguistic labels may interact with categorical representations of numerical magnitude (cf. Proctor and Cho, 2006), explaining why the SNARC effect is seen in these types of Parity tasks but not in our deliberately modified setup.

Why, then, did magnitude and parity interact with specific fingers during the Parity task? One possibility is that the associations exhibited in Experiment 2 originate in culturally-specific gestures for numbers. In Quentin Tarantino's film Inglourious Basterds, an American spy
posing as a German is exposed when he orders two beers with the American rather than the German gesture: index and middle finger extended, instead of thumb and index finger. The participants at our American university, therefore, may have finger-specific associations as a result of the fingers they use to gesture for numbers: a single extended index finger for one, adding the middle finger for two. Within our task, these index-one and middle-two associations may have been extended to the rest of the numbers, with smaller or odd numbers associated with the index finger, and larger or even numbers associated with the middle finger-much like the relation between one (smaller and odd) and two (larger and even). This is only speculation, of course, although it does make a specific prediction: German participants should behave differently on our Parity task, responding faster with the thumb for smaller or odd numbers, and faster with the index finger for larger or even numbers, if we test these two fingers instead.

The results of Experiment 2 are illuminated further by recent research on finger-based representations of numbers. Fischer and colleagues, for instance, have suggested that stable finger-counting routines may explain cross-cultural variability in the direction of the SNARC; native English speakers, for instance, may count from left-to-right on their fingers, and also exhibit a left-to-right SNARC (Fischer, 2008; Lindemann, Alipour, and Fischer, 2011). Others, however, have found right-handed native English speakers to be ten times more likely to start counting on their right hand than on their left (Tschentscher et al., 2012), a pattern that we have also observed in pilot studies. Additionally, associations have been found (in the form of response-time facilitation effects) between specific numbers and the fingers used for those numbers in a habitual finger-counting routine (e.g. di Luca et al., 2006). The current study, by contrast, found categorical (i.e. magnitude and parity) rather than number-specific associations with finger. It may be that specific finger-number associations are only salient when multiple fingers are spatially juxtaposed at a single time.

## Which spaces?

The current results do not rule out the possibility that other spatial frames of reference also contribute to known interactions between number and space. Head- and eyebased coordinates may also play a role, and future studies should contrive to situate response-buttons in ways that tease apart the contributions of head- and eye-based coordinates, both from each other and from an action-based frame. Indeed, the classic SNARC effects may have been so pronounced exactly because they conflated multiple complementary frames of reference, which conspired to produce particularly strong effects.

Moreover, number may be associated with still other "spaces," including distinctions between peripersonal and distal space, although interactions between these spaces are still under-theorized and starkly under-explored. One study on the relation between peripersonal space and number bisection found an interaction between distance in
peripersonal space and number bisection, perhaps related to biases in lateral spatial attention (Longo and Lourenco, 2010), although the precise mechanism for this interaction is still unknown. In a study that contrasted finger-based and space-based representations of number, Riello \& Rusconi (2011) examined the possibility of a unimanual SNARC using a Two Alternative Forced Choice paradigm. Participants responded with the index and middle finger of the same hand, pressing buttons on either side of participants' midline. Response hand and orientation (face up or down) were also manipulated. They found co-existing hand-based and space-based representations of number, which were either complementary or incompatible depending on the hand and its orientation.

Two conditions in Riello \& Rusconi (2011) are of particular interest: responses with downward-facing left and right hands. In contrast with the results of the current studies, Riello and Rusconi only found a classic left-to-right SNARC effect on the right downward-facing hand. They explained this by positing an interaction between handbased (from thumb to little finger) and space-based (left to right) representations of number, which would be in conflict on the left hand when facing downward, but in accord on the right hand. The difference between our results and theirs may be due to a number of factors. For one, participants in Riello and Rusconi (2011) responded with two adjacent buttons-pressed by the index and middle finger of the same hand-that were placed on either side of the participants' midline, and thus were distinguished in multiple frames of reference (head-, eye-, hand-, and actionbased). Additionally, the simultaneous spatial juxtaposition of the two response options-unlike our design, in which different response options were only juxtaposed temporally between blocks-may have highlighted hand-based representations. Our results suggest that, when fingers are not spatially juxtaposed within a single block, behavioral space interacts spontaneously with number during magnitude judgments. Moreover, unlike previous studies (e.g. Dehaene et al, 1993), we did not find an interaction between hand side and magnitude. We attribute this to the fact that hand side was not contrasted within the task, but only manipulated between subjects, and thus this was not a salient distinction for the individual. Living organisms, after all, "enact a world as a domain of distinctions" (Varela, Thompson \& Rosch, 1993, p.140).

Previous attention to space, furthermore, has eclipsed attention to time. Existing studies have juxtaposed spatial responses within a single block, so that on any given trial there were always multiple spatial responses available. In our studies, which used a go/no-go paradigm, the different response options were contrasted across the experiment as a whole, rather than within a block, and so any particular trial involved only one possible spatial response (or lack of response). Said otherwise, the different possible spatial responses in our tasks were not juxtaposed at any particular slice of time in which a response was made. In spite of this, reliable associations between number and space emerged.

The relevant units of analysis, we conclude, are the behavioral contrasts within the temporally- and corporeallyextended cognitive ecology circumscribed by the task.

## Conclusion

In two experiments, we found that number interacts with space even when responses are not distinguished by their location in head- or eye-based spatial frames of reference. Numbers were associated with locations in behavioral space, enacted by the participant within the context of the task. This association between number and behavioral space, however, was task-specific, appearing only when numerical magnitude was directly task-relevant. While space is a ubiquitous and powerful cognitive resource (e.g., Kirsh, 1995; Tversky, 2011), it is neither fixed nor monolithic. The spaces of human activity are multiple, defined relative to varied frames of reference, and we deploy them flexibly during abstract though. As MerleauPonty (1962) argued a half century ago, the space that we inhabit is not pre-given, but constituted by the motility of the body. Our reasoning about abstract concepts is not grounded in some single, static, or stable representation of space, but in the space we enact through embodied activity.

## Acknowledgments

Thanks to Anthony Chan, Sarah Saturday, and Luke Eberle for data collection, and to Rafael Núñez for discussion.

## References

Cohen, Y.E., \& Anderson, R.A. (2002). A common reference frame for movement plans in the posterior parietal cortex. Nature Reviews Neuroscience, 3, 553-562.
Dehaene, S., Bossini, S., \& Giraux, P. (1993). The mental representation of parity and numerical magnitude. Journal of Experimental Psychology: General, 122, 371-396.
Di Luca, S., Grana, A., Semenza, C., Seron, X., and Pesenti, M. (2006). Finger-digit compatibility in arabic numeral processing. The Quarterly Journal of Experimental Psychology, 59:1648-1663.
Fischer, M.H. (2008). Finger counting habits modulate spatial-numerical associations. Cortex, 44, 386-392.
Fischer, M.H., Mills, R.A., \& Shaki, S. (2010). How to cook a SNARC: Number placement in text rapidly changes spatial-numerical associations. Brain and Cognition, 72, 333-336
Fischer, M.H., Castel, A.D., Dodd, M.D., Pratt, J. (2003). Perceiving numbers causes spatial shifts of attention. Nature Neuroscience, 6, 555-556.
Gevers, W, Santens, S, D'Hooge, E, Chen, Q, Van den Bossche, L, Fias, W, \& Verguts, T (2010). Verbal-spatial and visuo-spatial coding of number-space interactions. Journal of Experimental Psychology: General 139:180-90
Kirsh, D. (1995). The intelligent use of space. Artificial Intelligence 73, 31-68.
Lakoff, G., \& Núñez, R. (2000). Where mathematics comes from. New York: Basic Books.

Lindemann, O., Alipour, A., \& Fischer, M. H. (2011). Finger Counting Habits in Middle-Eastern and Western Individuals: An Online Survey. Journal of Cross-Cultural Psychology, 42, 566-578.
Longo, M. R., \& Lourenco, S. F. (2010). Bisecting the mental number line in near and far space. Brain and Cognition, 72, 362-367.
Merleau-Ponty, M. (1962). Phenomenology of Perception. Trans. by C. Smith. London: Routledge \& Kegan Paul.
Marghetis, T., Walker, E., Bergen, B., \& Núñez, R. (2011). Making SNAP judgments: Rethinking the spatial representation of number. In L. Carlson, C. Hölscher, \& T. Shipley (Eds.), Proceedings of the 33rd Annual Conference of the Cognitive Science Society (pp. 17811786). Austin, TX: Cognitive Science Society.

Núñez, R., Doan, D, Nikoulina, A. (2011). Squeezing, Striking, and Vocalizing: Is Number Representation Fundamentally Spatial? Cognition, 120, 225-235.
Núñez, R., \& Marghetis, T. (to appear). Cognitive Linguistics and the Concept(s) of Number. In R. CohenKadosh and K. Dowker (eds.), Oxford Handbook of Numerical Cognition. Oxford University Press.
Proctor, R.W., \& Cho, Y.S., (2006). Polaritycorrespondence: A general principle for performance of speeded binary classification tasks, Psychological Bulletin, 132, 416-442.
Riello, M., Rusconi, E. (2011). Unimanual SNARC effect: hand matters. Frontiers in Psychology, 2, 372.
Santens, S., \& Gevers, W., (2006). The SNARC effect does not imply a mental number line, Cognition, 108, 263-270.
Schwarz, W., \& Keus, I.M. (2004). Moving the eyes along the mental number line: comparing SNARC effects with saccadic and manual responses. Perception \& Psychophysics, 66, 651-664.
Schwarz, W., \& Müller, D. (2006). Spatial associations in number-related tasks: A comparison of manual and pedal responses. Experimental Psychology, 53, 4-15.
Shaki, S., Fischer, M, Petrusic W.M. (2009). Reading habits for both words and numbers contribute to the SNARC effect. Psychonomic Bulletin \& Review, 16, 328-31.
Tschentscher, N., Hauk, O., Fischer, M. H., \& Pulvermüller, F. (2012). You can count on the motor cortex: Finger counting habits modulate motor cortex activation evoked by numbers. NeuroImage, 59(4):3139-3148.
Tversky, B. (2011). Visualizations of thought. Topics in Cognitive Science, 3, 499-535.
van Dijck, J.P., Gevers, W., \& Fias, W., (2009). Numbers are associated with different types of spatial information depending on the task, Cognition, 113, 248-253.
Varela, F., Thompson, E., \& Rosch, E. (1993). The embodied mind: Cognitive science and human experience. Cambridge: MIT Press.
Wood, G., Willmes, K., Nuerk, H.-C., \& Fischer, M. (2008). On the cognitive link between space and number: a metaanalysis of the SNARC effect. Psychology Science Quarterly, 50, 489-52.

# Subgoal Labeled Worked Examples Improve K-12 Teacher Performance in Computer Programming Training 

Lauren E. Margulieux (l.marg @gatech.edu)<br>Georgia Institute of Technology, School of Psychology<br>Atlanta, GA 30332-0170 USA<br>Richard Catrambone (rc7@ prism.gatech.edu)<br>Georgia Institute of Technology, School of Psychology<br>Atlanta, GA 30332-0170 USA

Mark Guzdial (guzdial@cc.gatech.edu)<br>Georgia Institute of Technology, School of Interactive Computing<br>Atlanta, GA 30332-0760 USA


#### Abstract

Technology has become integrated into many facets of our lives. Due to the rapid onset of this integration, many current K-12 teachers do not have the skills required to supply the sudden demand for technical training. This deficit, in turn, has created a demand for professional development programs that allow working teachers to learn computer science so that they might become qualified to teach this increasingly important field. Subgoal labeled worked examples have been found to improve the performance of learners in highly procedural domains. The present study tested subgoal labeled worked examples in an online learning program for teachers. Teachers who received the subgoal labels solved novel problems more accurately than teachers who received the same worked examples without the subgoal labels. These findings have implications for the use of subgoal labels in professional development, other types of lifelong learning, and online learning.


Keywords: subgoal learning; worked examples; computer programming, K-12 teacher training.

## Introduction

As technology becomes ubiquitous, being technically trained is frequently necessary for individuals to be effective in their professional and personal lives. Technology has advanced at such a rapid pace, however, that many of our educators are not qualified to train students in technical fields. Thus, it is important to train teachers, who have full schedules and possibly no technical training, to become qualified to teach technical subjects. Fortunately, because technical subjects tend to be highly procedural, methods used for teaching other highly procedural subjects like mathematics can be used in technical education.

One of the methods that has been effective for teaching procedural domains (e.g., statistics and physics) is to manipulate the format of worked examples that students receive (e.g., Catrambone, 1996). Catrambone (1998) found that worked examples that included subgoal labels were effective for helping students learn to solve problems in a new domain. This intervention has also been found to be effective for teaching computer programming (Margulieux, Guzdial, \& Catrambone, 2012). Most of these subgoal
studies, however, have been conducted with undergraduate students in face-to-face learning environments. These are not the conditions that would be ideal for K-12 teacher professional development. The present study explores the effectiveness of the subgoal intervention for K-12 teachers interested in learning computer science in an online learning environment (i.e., with no face-to-face interaction).

Worked examples are an important instructional tool for learners in highly procedural domains like math or computer programming. Worked examples help learners because they provide specific information about how to apply domain principles to problem solving (Bassok, 1990). Furthermore, worked examples provide a step-by-step solution to a problem from which students can learn before they are able to solve problems independently (Atkinson, Derry, Renkl, \& Wortham, 2000). When learners are presented with all of the steps of an example solution at once, however, they often have difficulty determining what information is important for solving problems in that domain (i.e., structural information) and what information represents details relevant for solving only that problem (Catrambone, 1994).

Using subgoal labels to group steps of worked examples into meaningful units can help learners recognize structural information in the examples. Subgoals are functional components of complex problem solutions; each subgoal is a necessary part of the solution. How a subgoal is achieved might vary between and within problems, but the subgoals needed to complete a problem do not. Subgoals are specific to a domain, but not to a problem; a multitude of problems in a domain might have the same subgoal structure, so by learning the subgoals in a domain, students can learn to solve problems in that domain (Catrambone, 1994).

Learners who study materials that label the subgoals of a worked example are more likely to solve novel problems than learners who study the same examples without the subgoal labels (Catrambone, 1998). There are several possible theoretical explanations for this phenomenon. Subgoal labels can help learners chunk problem-solving steps which might reduce the cognitive load required to learn them (Catrambone, 1994). Furthermore, subgoal labels might help learners create mental models in a domain by
providing them with a framework (i.e., the set of subgoals) that they can use to organize information in a way that can guide transfer to future problems (Atkinson et al., 2000, Catrambone, 1996). Moreover, apprising learners of the structure of worked examples can help them recognize similarities among examples and promote self-explanation (Catrambone, 1998; Renkl \& Atkinson, 2002).

Expanding upon previous work (e.g., Catrambone, 1998), Margulieux et al. (2012) applied subgoal labeled worked examples to a previously untested domain, computer programming. They found that subgoal labels improved participants' performance on novel computer programming construction tasks (i.e., creating applications (apps) for Android devices). The present study expands upon this work by testing the intervention in a new environment and with a new population.

## Present Study

The present study manipulated the materials that K-12 teachers received to help them teach themselves how to program. Participants received either subgoal labeled worked examples or conventional worked examples (i.e., list of the steps of the solution with no labels). The conventional worked examples were adapted from material in the projects sections of the ICE Distance Education Portal (http://ice.cc.gatech.edu/dl/?q=node/641). The subgoals of the examples were determined using the TAPS procedure developed by Catrambone, Gane, Adams, Bujak, Kline, and Eiriksdottir (2013) and consultation with subject-matter experts (see Figure 1). The only difference between the materials that participants in the two conditions received was the added subgoal labels (see Figure 2).

```
Subgoal Labels
    1. Create components
    2. Set properties
    3. Handle events from My Blocks
    4. Set outputs from My Blocks
    5. Define variable from Built-In
    6. Set conditions from Built-In
    7. Emulate app
```

Figure 1. Subgoals Used In Instructional Material
The programming language that was used for the study is Android App Inventor, which is used to develop apps for Android devices. App Inventor is a drag-and-drop programming language; users are given pieces of code that they can drag from a menu and piece together in a programming area to make programs. Drag-and-drop programming languages can be useful for teaching novices because, instead of writing code, users select sections of code and piece them together like puzzle pieces. This type of code creation is easily understood by novices (Hundhausen, Farley, \& Brown, 2009).

## Subgoal labeled Materials

## Handle Events from My Blocks

1. Click on "My Blocks" to see the blocks for components you created.
2. Click on "clap"
3. Drag out a when clap.Touched block

## Set Output from My Blocks

4. Click on "clapSound" and
5. Drag out call clapSound.Play
6. Connect it after when clap.Touched

## Conventional Materials

1. Click on "My Blocks" to see the blocks for components you created.
Click on "clap"
Drag out a when clap.Touched block
Click on "clapSound"
Drag out call clapSound.Play
Connect it after when clap.Touched
Figure 2. Sample Materials from Two Groups
Over four sessions participants learned to make apps using App Inventor. In each session, participants received instruction for how to make one app and assessments asking them to modify or make new parts of an app (see Table 1).

Table 1: Sections of experimental sessions

| Session | $1^{\text {st }}$ section | $2^{\text {nd }}$ section | $3^{\text {rd }}$ section |
| :--- | :--- | :--- | :--- |
| 1 | Introduction | Instruction | Assessment |
| $2,3,4$ | Assessment | Instruction | Assessment |

In the first session, participants learned to make an app that played sounds when the user interacted with objects on the screen. In the second session, participants learned to make an app that selected and displayed text when a button was pressed. In the third session, participants learned to make an app that counted the number of times the user pressed a button in a time frame. In the fourth session, participants learned to make an app similar to the game Pong.

Instructional materials for each app included both a video demonstrating how to make an app and a text guide detailing how to make an app. Palmiter and Elkerton (1993) found that videos demonstrating how to complete tasks using a direct-manipulation interface can quickly and naturally teach users how to use the interface. They also concluded that only watching videos can lead to superficial processing while reading text instructions leads to deeper processing. Given that video demonstrations are a useful aid for learning to complete tasks using an unfamiliar interface and that text instructions lead to better transfer and retention for these tasks (Palmiter \& Elkerton, 1993), both types of instruction were used in the present study. Subgoal labels were presented in the videos as callouts to present the information succinctly without overshadowing any verbal instructions (see Figure 3, arrow added).


Figure 3. Sample of Subgoal Callout in Video
To assess participants' ability to solve problems using App Inventor, participants were asked to write the steps that they would take to program new features of an app. These assessment tasks were developed based on material that participants were exposed to during the sessions, but some assessment tasks required participants to use aspects of App Inventor that they had not used before to measure their ability to transfer their knowledge. Hints were given for tasks that required participants to use these unfamiliar features. The hints guided participants to the correct features but did not tell them how to use that feature (see Figure 4).
> "1.5 Write the steps you would take to make the screen change colors depending on the orientation of the phone; specifically, the screen turns blue when the pitch is greater than 2 (hint: you'll need to make an orientation sensor and use blocks from "Screen 1" in My Blocks)."
> "3.3 Write the steps you would take to create a list of colors and make the ball to change to a random color whenever it collided with something."

Figure 4. Sample of Assessment Tasks
Two types of assessments were given. One type was given at the end of each session and intended to measure participants' ability to solve novel problems, so it included near and far transfer tasks. The other type was given at the beginning of each session starting with the second session and intended to measure participants' retention of problem solving procedures, so it included only near transfer tasks.

Near transfer tasks required participants to follow an identical procedure that they had used in the instructional session but substituted blocks or components of the same type. For example, one task asked participants to program the clap sound to play when the phone was tilted up. To complete this task, participants could follow the same steps that they used in the instructional session to program the drum sound to play when the phone was tilted to the right,
but they had to replace the drum sound with the clap sound and the $x$-axis acceleration sensor with the $y$-axis acceleration sensor.

Far transfer tasks required participants to follow the same general scheme that they had used in the instructional session but substituted blocks or components of a different type. For example, one task asked participants to program an ImageSprite to move 5 pixels to the right when touched. The steps to do this task were different than the steps in the instructional session because the type of block was different, but the subgoals that needed to be completed were the same.

Participants were not permitted to use the video or text guides during the assessment period, but participants were encouraged to use the App Inventor interface to help them complete the assessment tasks. Participants were also allowed to access the apps that they had made during the session to serve as memory cues for the complex procedures they had learned in the session. Participants were instructed to not review instructional material between sessions, so their retention of problem solving procedures could be measured consistently.

## Method

## Participants

Participants were 18 K-12 teachers recruited through mailing lists for teachers interested in computer science education. Teachers with prior experience with Android App Inventor could not participate in the experiment, but they were not restricted by any other prior experience. The teachers had backgrounds that varied on a number of factors such as education, years as a teacher, years teaching computer science, level of computer science taught, and professional development completed. There were no correlations between participant performance and prior experience, so this issue will not be discussed further.

## Procedure

The experiment was conducted online with no face-to-face interaction. Instructions and media for the apps were emailed to participants, and the sessions were hosted on surveymonkey.com. Each SurveyMonkey survey gave participants instructions for completing the instructional session and assessment tasks (first session survey: http://www.surveymonkey.com/s/RVCWTBX, use "test" as participant number). Through the survey, participants were asked to record how long they spent on each instructional session and each assessment task. Participants were also asked how difficult they thought each instructional session and assessment task was on a Likert-type scale from "1Very Difficult" to "7-Very Easy."

The experiment comprised four sessions which were given one week apart. The timestamp on the surveys were checked to ensure participants completed the sessions at least six days apart. The sessions were similar to those in Margulieux et al. (2012) but adapted for online use. The major difference between the Margulieux et al. (2012) and
present administration of sessions is that the moderator instructions were given through text instead of speech. Each session taught participants how to make an app using a video and text guide. The video guide showed participants how to create the app, and the text guide gave step-by-step instructions for creating the app. After participants made the app for that session, they worked on the assessment tasks. Starting with the second session, participants also completed the retention assessment at the beginning of the session before they started making the app (see Table 1).

Completion rates for the sessions decreased during the study with a high level of participation for the demographic survey and low level for the last two sessions. Though the participants volunteered to be in the study, they did not receive any compensation for their time except instruction about App Inventor. Additionally, the assessment tasks were designed to be difficult in order to avoid a restriction of range problem caused by all participants performing well. Many participants commented that they were frustrated with the tasks. The teachers might have lost motivation to complete the sessions without more compensation. Few teachers experienced unforeseeable conflicts that ended their participation. There was not a recognizable pattern that distinguished participants who completed the study from those who did not. Data from only the first two sessions were analyzed due to low completion rates of the last two sessions.

These attrition rates are similar to those seen in other online learning environments such as Massive Open Online Courses (MOOCs). In an analysis of nearly 500,000 courses taken by over 40,000 students, Xu and Jaggars (2013) found that many of the factors that predict success in face-to-face learning environments also predict success in online learning environments (e.g., women were more successful, and students with higher GPAs were more successful). This finding suggests that attrition in online courses is similar to attrition in face-to-face courses but on a larger scale. However, the number of students that online courses can reach is much larger, so the number of students who complete an online course is generally greater than the number of students who complete an equivalent face-to-face course (Whiteman, 2013).

## Results and Discussion

Each solution of the assessment tasks was deconstructed into the components necessary to complete the solution; that is, the subgoals of the solution. As discussed earlier, the subgoals are inherent in the solutions, but the tasks did not provide any information about which subgoals were necessary to complete the solution. Because the solutions for the assessment tasks are complex, scoring the pieces of each solution instead of scoring the entire solution as correct or incorrect allowed for more sensitivity in the measurement.

Problem-solving performance is represented by two scores: a "correct" score and an "attempted" score.

Participants were given a point for each subgoal that they completed correctly and each subgoal that they attempted. Attempting a subgoal was operationally defined as listing at least one of the steps required to complete the subgoal, listing an incorrect step that would achieve a similar function, or describing the purpose of the subgoal in some way. Participant responses were scored by multiple raters, and interrater reliability was high with a one-way random model intraclass correlation coefficient of agreement (ICC(A)) of .87. There were 32 subgoals across the assessment task solutions, so participants could get a maximum score of 32 for both the attempted and correct problem-solving measurements.

## Correct Subgoals

Participants in the subgoal group ( $n=9$ ) completed $81 \%$ more subgoals correctly ( $M=26.6, S D=5.08$ ) than the conventional group ( $n=9, M=14.7, S D=6.63$ ), $F(1,16)=$ 18.23, $M S E=34.89, p=.001, \omega^{2}=.53, f=1.01$. These results mean that $53 \%$ of the variance for correct subgoals was accounted for by group. Furthermore, this is a very large effect size considering the amount of instruction that participants received (i.e., two, 30-45 minute instructional sessions). These findings suggest that the subgoal labeled worked examples, compared to conventional worked examples, can help people learn more efficiently to solve programming problems.

The difference between groups in this experiment is about twice as large as the difference between groups in Margulieux et al. (2012), $f=1.01$ vs. $f=.53$, respectively, even though the present study was conducted in a less controlled environment and its participants had more varied backgrounds. Participants in the present study also had as much time as they wanted to work on the assessments instead of being limited like in Margulieux et al. (2012).

One explanation for this larger effect could be that participants in this study were teachers who volunteered because they wanted to learn the material to further their career while participants in the Margulieux et al. (2012) studies were undergraduates who were less likely to be motivated to learn the material. Therefore, this difference could mean that the subgoal intervention is more effective for learners who are motivated to learn the material for the long-term than it is for lab participants who might only try to learn the material for the duration of the experiment.

Another possible explanation is that the participants in Margulieux et al. (2012) were students whose skills for learning new material were sharper than those of teachers who might have been out of school for decades. The difference between groups for the undergraduate sample might be smaller than for teachers because the students had better strategies for studying conventional worked examples than the teachers. Therefore, undergraduates who received the conventional worked examples would have performed better than teachers who received the conventional worked examples, thereby creating a smaller difference between groups in Margulieux et al. (2012) than the present study.

For both near and far transfer tasks, the subgoal group completed more subgoals successfully (Near: $M=10.6, S D$ = 1.94; Far: $M=7.1, S D=2.26$ ) than the conventional group (Near: $M=5.2, S D=3.70$; Far: $M=3.3, S D=2.35$ ), Near: $F(1,16)=14.65, M S E=8.74, p=.001, \omega^{2}=.48, f=$ .90, Far: $F(1,16)=12.11, M S E=5.31, p=.003, \omega^{2}=.43, f$ $=.82$. These results suggest that subgoal labels help performance on both near and far transfer tasks. Given the nature of the near and far transfer tasks, these findings could mean that the subgoal labels helped participants learn the material better (near transfer) and apply the material to novel problems (far transfer).

On the first end-of-session assessment tasks, participants in the subgoal group completed $223 \%$ more subgoals correctly ( $M=9.7, S D=1.41$ ) than the conventional group ( $M=3.0, S D=3.02$ ), $F(1,16)=27.04, M S E=5.56, p<$ $.001, \omega^{2}=.63, f=1.23$. These results mean that $63 \%$ of the variance for correct subgoals was accounted for by group. On the second end-of-session assessment tasks, participants in the subgoal group completed $70 \%$ more subgoals correctly ( $M=8.0, S D=2.83$ ) than the conventional group $(M=4.7, S D=3.57), F(1,16)=4.82, M S E=10.38, p=$ $.043, \omega^{2}=.23, f=.50$. These results mean that $23 \%$ of the variance for correct subgoals was accounted for by group.

The two series of assessments suggest the subgoal group was better at solving novel problems than the conventional group. Because the effect size of the second assessment was smaller than that of the first assessment $(f=.50$ vs. $f=1.23$, respectively), the difference between groups might decrease with repeated exposure to the same type of material. This decrease would be expected because as learners gain more knowledge, they are better able to identify important information and need less external guidance. This finding suggests that the subgoal labels are fulfilling the purpose for which they are intended: to highlight the information on which learners should focus so they can learn more effectively. Over time, both groups might achieve the same problem solving ability, but the learners who receive subgoal labels would reach a higher level faster than those who do not. This finding does not mean that subgoals are not valuable later, but it suggests that they are most effective when learners are first introduced to new material.

On the start-of-session assessment tasks (i.e., to measure retention of problem solving procedures), participants in the subgoal group completed $48 \%$ more subgoals correctly ( $M=$ 9.0, $S D=1.70$ ) than the conventional group ( $M=6.1, S D=$ 3.22), $F(1,16)=6.17, M S E=6.41, p=.024, \omega^{2}=.27, f=$ .57. These results mean that $27 \%$ of the variance for correct subgoals was accounted for by group. All of the tasks in this series were near transfer tasks, so to complete the tasks participants had to use procedures that they had learned in the previous session. This result suggests that the subgoal intervention promotes retention of the procedures.

## Attempted Subgoals

Participants in the subgoal group attempted $25 \%$ more subgoals ( $M=28.6, S D=3.50$ ) than the conventional group
( $M=22.8, S D=7.19$ ), $F(1,16)=4.70, M S E=31.70, p=$ $.046, \omega^{2}=.23, f=.51$. By attempting a subgoal, participants could be demonstrating that they know the solution needs a particular component. Therefore, this finding could mean that subgoal participants recognized more of the necessary components of the solutions than the conventional participants regardless of whether they were able to correctly complete the task.

## Time on Task and Difficulty

There were no statistically reliable differences between the groups on the time and difficulty measures (viz., time spent on instructional periods, difficulty rating of instructional periods, time spent on assessment periods, and difficulty rating of assessment periods; see Table 2). These results suggest that participants in the subgoal group performed better than the conventional group without taking longer to complete the instructions or tasks and without finding the instructions or tasks more difficult.

Table 2: Difference between groups for time and difficulty measures; time in minutes, difficulty on 7-pt. scale (1-Very Difficult and 7-Very Easy)

| Category | $M$ <br> subgoal | $M$ <br> conv | $\overline{S D}$ | $F$ | $p$ |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Time on <br> Instruction | 77.3 | 87.8 | 37.8 | .37 | .55 |
| Difficulty of <br> Instruction | 4.9 | 4.5 | 1.0 | .23 | .64 |
| Time on | 76.6 | 56.7 | 33.1 | 1.44 | .25 |
| Assessments <br> Difficulty of <br> Assessments | 4.3 | 3.8 | 1.1 | .66 | .43 |

This conclusion is supported by linear regression models. Group ( $\beta=.58, p=.005$ ) and time ( $\beta=.41 p=.031$ ) are both significant predictors of correct subgoal score suggesting that they account for different parts of the variance. When predicting attempted subgoal scores, group is no longer a significant predictor, and time ( $\beta=.54 p=$ .032) becomes the sole predictor. This model accounts for participants who spent relatively little time on the assessment tasks and did not write solutions (i.e., who did not attempt to solve the task). Furthermore, group ( $\beta=.62$, $p=.002$ ) and difficulty rating ( $\beta=.42 p=.024$ ) are both significant predictors of correct subgoal score suggesting that they also account for different parts of the variance in scores. When predicting attempted subgoal scores, however, group is no longer a significant predictor, and difficulty rating ( $\beta=.63 p=.009$ ) becomes the sole predictor. This model accounts for participants who did not attempt to solve the problems and rated the difficulty of the tasks as high. Due to a high correlation between time on task and difficulty rating ( $r=.60, p=.015$ ), these two predictors were analyzed in different models to avoid multicollinearity.

## Conclusion

Subgoal labeled worked examples have been effective for teaching students to solve problems in procedural domains such as statistics (Catrambone, 1998) and computer programming (Margulieux et al., 2012). Most of these studies have taken place in a laboratory with undergraduates. The present study extends prior work with results that suggest subgoal labeled worked examples are effective for K-12 teachers learning App Inventor in an online learning environment. These findings demonstrate that subgoal labels can be effective in a learning environment outside of the laboratory with a different population of learners.

It is encouraging that the subgoal intervention improved online learners' performance. The purpose of labeling subgoals in worked examples is to succinctly give the learner extra information to help them recognize the structure of the example. This type of extra information is what an instructor, who is an expert in the subject matter, might ideally provide to students in face-to-face instruction. Unfortunately, instructors are not always aware that they should provide this extra information, and even if they are aware, they do not necessarily know how to impart the information. In an online learning environment in which students rarely interact with an instructor, such as the one in this experiment, this extra information needs to be built into the instructions. Extra information could increase learning time. However, the present study demonstrates that, in the absence of an instructor, subgoal labeled worked examples provide enough extra information to help students learn more effectively without increasing the amount of time students take to learn.

The results of the experiments also imply that the subgoal intervention can be effective for populations other that undergraduates. The sample in the present experiment was heterogeneous in terms of age, education, and experience, so the amount of variance in the participants' performance scores that was accounted for by experimental group (over $50 \%$ in some cases) was surprisingly large. This finding can justify the use of resources to implement subgoal interventions in professional development, classrooms, and other instructional environments, including those online.

The present study demonstrates that subgoal labeled worked examples can be an effective intervention for teaching highly procedural domains outside of the laboratory. Additional experiments can examine the intervention in a variety of learning environments.

## Acknowledgments

Our thanks to Georgia Tech's GVU Center and IPaT for the grant that made this research possible. We also thank Barbara Ericson for her consultation and Andrea Crews for help scoring data.

## References

Atkinson, R. K., Derry, S. J., Renkl, A., \& Wortham, D. (2000). Learning from Examples: Instructional Principles from the Worked Examples Research. Review of the Educational Research, 70(2), 181-214. doi: 10.2307/ 1170661
Bassok, M. (1990). Transfer of Domain-specific Problemsolving Procedures. Journal of Experimental Psychology: Learning, Memory, and Cognition, 16(3), 522-533. doi: 10.1037/0278-7393.16.3.522

Catrambone, R. (1994). Improving examples to improve transfer to novel problems. Memory and Cognition, 22, 605-615. doi: 10.3758/BF03198399
Catrambone, R. (1996). Generalizing solution procedures learned from examples. Journal of Experimental Psychology: Learning, Memory, and Cognition, 22, 10201031. doi: 10.1037/0278-7393.22.4.1020

Catrambone, R. (1998). The subgoal learning model: Creating better examples so that students can solve novel problems. Journal of Experimental Psychology: General, 127, 355-376. doi: 10.1037/0096-3445.127.4.355
Catrambone, R., Gane, B., Adams, A., Bujak, K., Kline, K., \& Eiriksdottir, E. (2013). Task Analysis by Problem Solving (TAPS): A Method for Uncovering Expert Knowledge. Unpublished manuscript, Georgia Institute of Technology, Atlanta, GA.
Hundhausen, C. D., Farley, S. F., \& Brown, J. L. (2009). Can direct manipulation lower the barriers to computer programming and promote transfer of training?: An experimental study. ACM Transactions in CHI, 16(3). doi: 10.1145/1592440.1592442
Margulieux, L. E., Guzdial, M., \& Catrambone, R. (2012). Subgoal-labeled instructional material improves performance and transfer in learning to develop mobile applications. In Proceedings of the Ninth Annual International Conference on International Computing Education Research, 71-78. doi: 10.1145/2361276. 2361291
Palmiter, S., \& Elkerton, J. (1993). Animated demonstrations for learning procedural computer-based tasks. Human-Computer Interaction, 8(3), 193-216. doi:10.1207/s15327051hci0803_1
Renkl, A., \& Atkinson, R. K. (2002). Learning from examples: Fostering self-explanations in computer-based learning environments. Interactive Learning Environments, 10(2), 105-199. doi: 10.1076/ilee.10.2.105. 7441
Whiteman, W. (April, 2013). MOOCs in engineering education. Presented at the meeting of American Society of Engineering Education Georgia Tech Chapter, Atlanta, GA.
Xu, D., \& Jaggars, S. S. (2013). Adaptability to online learning: Differences across types of students and academic subject areas (CCRC Working Paper No. 54). Teachers College, Columbia University.

# The Ambivalence of Expert Categorizers 

Jessecae K. Marsh (jem311@lehigh.edu)<br>Department of Psychology, 17 Memorial Drive East<br>Bethlehem, PA 18015 USA

Naomi B. Rothman (nbr211@lehigh.edu)<br>Department of Management, 621 Taylor Street<br>Bethlehem, PA 18015 USA


#### Abstract

We explored people's reactions to expert categorizers who expressed difficulty in making a categorization decision. Specifically, we compared people's impressions of expert health professionals who either expressed certainty, uncertainty, or ambivalence about a categorization decision in the form of a diagnosis. We found that ambivalence resulted in the most negative impressions of these experts, including lower ratings of competence and decisiveness (Experiment 1). Impressions of ambivalence did not improve when the complexity of the decision was explicitly manipulated (Experiment 2). Implications for categorization are discussed.


Keywords: expert; categorization; decision-making; ambivalence.

## Introduction

People view the world as existing in clear, definable categories (Gelman, 2003). For example, when attempting to identify a bird sitting in our yard, we take as granted that there are clear delineations between different species of birds and that with enough knowledge a given bird can be neatly categorized into its appropriate category (Diesendruck \& Gelman, 1999; Estes, 2003). Believing the natural world is organized and divided in this way suggests that there are right answers to categorizing things that are accessible given enough knowledge. The people we turn to that possess this knowledge we call experts. Laypeople perceive that for different domains in the world, experts exist and possess knowledge specific to that domain (Wilson \& Keil, 1998). People defer to these experts when information or a decision is needed (Braisby, 2001, 2003; Danovitch \& Keil, 2004; Danovitch \& Keil, 2007; VanderBorght \& Jaswal, 2009). This deference has been described as a division of cognitive labor that allows a given person to be able to interact with elements of the world she does not understand because of the belief that there are experts that exist that do understand those elements (for a discussion see Wilson \& Keil, 1998).

Experts play an obviously important role in allowing people to survive in the modern world. Given the importance of experts, what happens when an expert expresses difficulty in making a categorization decision? For example, imagine a mechanic who can not decide if a car is malfunctioning because of a transmission problem or because of an exhaust problem, or a bird authority who cannot decide which of two species is the correct
categorization for a bird, or a doctor who is torn between two possible diagnoses for a patient. In short, how do we react to these experts who we have turned to for help when they express difficulty in making a categorization decision? Furthermore, are there differences in our reactions, depending on the type of difficulty experts are expressing? We delineate three different possible states a person making a categorization decision could experience: knowing the correct answer (certainty), not being clear at all as to what the correct answer is (uncertainty), and having narrowed the correct answer down but feeling tension and conflict as to which answer is the correct choice (ambivalence). In our study, we are specifically interested in this state of ambivalence. In the following we describe how people may react to ambivalent experts and then contrast this with possible reactions to uncertain experts.

How do we react to an expert who expresses being ambivalent about a categorization decision in her domain of expertise? One possibility is that ambivalence in experts is not perceived as problematic, but instead as a sign of effective decision-making. Seeing an expert express being torn over the correct categorization may verify our initial deference; this is a complex decision that is beyond our ability. Furthermore, expressions of ambivalence are often taken as a positive sign of more deliberative, and flexible thinking (Rothman, 2011). In addition, experiencing ambivalence is related to more creative (Fong, 2006) and more accurate (Rees, Rothman, Lehavy, \& Sanchez Burks, 2013) final decision-making. As such, an expert categorizer who is torn between placing something in one of two categories may be seen as a creative, deep thinker and valued for her expertise.

We believe it is much more likely that people react negatively to ambivalent experts. People largely act as if categories in the world exist with clear, delineable boundaries. That is, something is not partially a bird, or part of the bird and cat category. This belief in clear categorical boundaries has been linked to essentialism, or belief that a category has a causal essence that underlies the category, creates the features of that category, and must be possessed to be a member of that category (Gelman, 2003). Previous work has claimed that it is exactly this belief that categories possess essences that allows us to be willing to defer to experts in making a categorization decision; people are believed to defer to experts because they believe experts have the correct knowledge to recognize and identify the
causal essences that underlie category membership (Braisby, 2001; 2003). Now imagine this expert who is supposed to be able to recognize the essence underlying categorization being torn as to which of two categories something belongs. Because of the implications of essentialism, category membership should be all-or-none, with the item to be categorized only being in one of the two possible categories. ${ }^{1}$ Also, if anyone should be able to identify that causal essence, it should be this expert. In this way, an expert expressing ambivalence may elicit negative reactions because ambivalence does not mesh well with our beliefs about essences and the ability of experts to identify those essences. For these reasons, we predict that experts expressing ambivalence in making a categorization decision should be viewed negatively.

An interesting alternative to ambivalence in categorization decision-making is uncertainty. Previous literature on decision-making has distinguished uncertainty and ambivalence as two separate emotional and decisional states. Uncertainty is a state of not knowing the correct answer to a problem, whereas ambivalence reflects a state of being torn between two possible alternatives (Rothman, 2011). Said another way, a decision maker may be uncertain because not enough information is known to make a decision or because the person does not have enough expertise to know the correct answer. However, a decision maker who is ambivalent appears to have all of the information needed to make a decision but is torn between two possibilities. In our paradigm using expert decision makers, we predict that an uncertain expert may look like someone who just needs more information before a decision is possible. However, an ambivalent expert will seem to have all or at least more of the needed information since she is actively considering two possibilities. This should result in the ambivalent expert looking relatively more unable to make correct decisions. If this holds, we would expect that ambivalent experts could be viewed more negatively than uncertain experts.

In the following two experiments, we investigate how people view expert categorizers making a categorization decision. To ensure participants' familiarity with the type of expert categorizer and the decision domain, we presented participants with descriptions of a health professional making a difficult diagnostic decision. Interacting with this type of expert categorizer should be easy for participants to think about. We described the professional as deciding between two possible diagnoses and displaying one of three levels of categorization certainty: certain of the correct diagnosis, completely uncertain as to which of two

[^147]diagnoses is correct, or torn and conflicted as to which of two diagnoses was correct. We then measured participants' impressions of these expert categorizers.

## Experiment 1

In Experiment 1, we manipulated the levels of categorization certainty of an expert decision-maker and then measured people's perceptions of the quality of that decision-maker. If ambivalence is inherently unsettling for the reasons discussed above, participants should view ambivalent experts as lower quality and more indecisive than certain or even uncertain experts. However, if ambivalence is taken as a sign of the expert being thoughtful, then perceptions of these experts should be more favorable.

## Methods

Participants Sixty participants recruited through Amazon's Mechanical Turk participated for payment. Participation was restricted to Mechanical Turk workers in the United States.

Materials and Procedure Participants read a description of a hypothetical patient who was seeking advice from a health care professional for an ongoing health problem. Participants read that the patient was given a series of tests, the results of which suggested two potential diagnoses, labeled A and B . Participants were randomly assigned to read one of three statements that described the health professional as being certain the patient had diagnosis A and not B (Certain condition; $n=20$ ), uncertain as to whether diagnosis A or B was correct (Uncertain condition; $n=18$ ), or torn as to whether diagnosis A or B was correct (Ambivalent condition; $n=22$ ). We used the torn descriptor for the ambivalent condition because it conveys how the subjective state of ambivalence is likely to be expressed (see Rothman, 2011 for a more detailed discussion of this point). Participants were randomly assigned either to read that the person was seeking help from a physician and the diagnosis was one of two infections ( $n=27$ ) or was seeking help from a mental health clinician and the diagnosis was one of two mood disorders $(n=33)$. After reading the health interaction description, participants completed a series of different ratings that asked them to rate the provider on different dimensions or rate how a patient would react to the provider. We were specifically interested in three issues: how indecisive and how competent the provider from the previous exchange was seen to be, as well as how positively or negatively participants reacted to the provider. Indecisiveness was measured by asking participants to rate to what extent the physician possessed a series of personality traits, in which were embedded the following 8 traits related to indecision: Confused, Unsure, Uncertain, Indecisive, Hesitant, Not Definite, Faltering, and Wavering. Mean ratings across these 8 measures were used as a measure of indecisiveness. To measure competence, we asked participants to rate their agreement with a series of
statements assessing how likely they thought the physician would be to engage in a series of behaviors related to being well informed (mean of three statements: attend professional conferences, be aware of current research, be asked for an opinion by other professionals) and be seen as a quality practitioner (mean of three statements: be a high quality expert, make accurate diagnoses, create accurate treatment plans). We also measured participants' predictions of how a patient would react to the given expert through ratings of how likely the patient would be to recommend that others defer to this expert (mean of three statements: refer a friend to the health care professional, provide a strongly positive rating on a referral website, take his/her children to see this health care professional) and how likely the expert would be to be sued by a patient (assessed through a single question). These ratings were intermixed with other ratings of the professional that were unrelated to the measures we discuss here. All ratings were made on seven-point agreement scales with the exact anchor points of the scales varying by task (e.g., Not at all likely to Extremely likely).

Participants also made ratings related to behaviors of the patient, but we do not present those results here and these measures are not discussed further. The order of rating tasks was randomized for each participant.

## Results

There were no significant differences between the mental and medical health professionals on any of our measures, $p \mathrm{~s}$ $>.12$. As such, we collapsed across that manipulation. For all of the following analyses, we conducted one-way ANOVAs with categorization certainty (Ambivalent, Certain, Uncertain) as a between-subjects variable.

Decision Indecisiveness We first assessed perceptions of the expert's indecisiveness. We compared the mean ratings for participants in the ambivalent condition to ratings in the certain and uncertain conditions. Participants' perceptions of the expert's indecision differed significantly by condition, $F(2,57)=14.43, p<.001$. Planned contrasts demonstrated that the Ambivalent physician was judged as significantly more indecisive $(M=4.45, \mathrm{SD}=1.31)$ than the Certain physician $(M=2.47, \mathrm{SD}=0.94 ; t(57)=5.36, p<.001)$. Interestingly, the Ambivalent expert was also perceived as significantly more indecisive than the Uncertain physician ( $M=3.62$, $\mathrm{SD}=1.30 ; t(57)=2.19, p=.033)$. Not surprisingly, the Uncertain physician was perceived as more indecisive than the Certain physician, $t(57)=2.95, p=.004$.

Expert Competence Participants' perceptions of the expert's level of being informed differed significantly by condition, $F(2,57)=6.19, p=.004$. Planned contrasts demonstrated that the Uncertain and Certain experts were seen as equally informed, $p=.43$. However, the Ambivalent expert was judged as significantly less well informed ( $M=$ 4.14, $\mathrm{SD}=1.29$ ) than the Certain expert $(M=5.33, \mathrm{SD}=$ $0.97 ; t(57)=3.37, p=.001)$, or Uncertain expert $(M=5.04$, $\mathrm{SD}=1.15 ; t(57)=2.46, p=.017)$.

Similar results obtained with perceptions of the expert's quality. Expert's perceived quality differed significantly by condition, $F(2,57)=5.39, p=.007$. Planned contrasts found that Uncertain and Certain experts were seen as equal in quality, $p=.37$. As predicted, the Ambivalent expert was judged as significantly lower quality $(M=3.89, \mathrm{SD}=1.24)$ than the Certain physician ( $M=4.77, \mathrm{SD}=1.34 ; t(57)=$ $2.29, p=.026)$ or the Uncertain physician $(M=5.13, \mathrm{SD}=$ $1.10 ; t(57)=3.15, p=.003)$.

Reactions to the Expert Participants' perceptions of whether the patient would refer the expert differed significantly by condition, $F(2,57)=12.17, p<.001$. Planned contrasts demonstrated that the Ambivalent expert was judged as significantly less likely to be referred ( $M=$ 2.73, $\mathrm{SD}=1.11$ ) than the Certain expert $(M=4.47, \mathrm{SD}=$ 1.32; $t(57)=4.74, p<.001)$. The Ambivalent expert was also perceived less likely to be referred than the Uncertain expert $(M=4.02, \mathrm{SD}=1.13 ; t(57)=3.42, p=.001)$. Predicted referrals did not differ for the Uncertain and Certain experts, $p=.25$.

Participants' ratings of the likelihood the expert would be sued differed significantly by condition, $F(2,57)=7.54, p=$ .001. Planned contrasts demonstrated that the Ambivalent expert was judged as significantly more likely to be sued ( $M$ $=4.86, \mathrm{SD}=1.52)$ than the Certain expert $(M=3.15, \mathrm{SD}=$ $1.46 ; t(57)=3.77, p<.001)$. The Ambivalent expert was judged as more likely to be sued than the Uncertain expert ( $M=3.67, \mathrm{SD}=1.41 ; t(57)=2.56, p=.013)$. Likelihood to be sued did not differ for Certain and Uncertain experts, $p=$ . 28 .

## Discussion

Our results show that ambivalent experts are perceived uniformly more negatively than certain and uncertain experts. If this negative impression was just in reaction to the expert being anything but completely certain in his/her decision making, then we would expect to see uncertain experts being viewed more in line with ambivalent experts. Our findings instead support that uncertainty does not produce the same negative reactions as ambivalence. In fact uncertain experts were viewed as positively as certain experts, except for not surprisingly being viewed as more indecisive. One explanation for this finding is that uncertainty is interpreted as a case in which more information is simply needed before a decision can be made, rather than a sign of ineptitude. Conversely, ambivalence may be interpreted as a case in which all of the information is available, but the expert is simply inept. It is a question for future research as to whether these interpretations are what are driving our demonstrated results.

Given previous research on the benefits of ambivalence to the decision-making process (e.g., Fong, 2006; Rees et al., 2013), it may seem surprising that ambivalence is viewed so negatively. As we discussed earlier, laypeople may assume that categories exist and can be identified as long as the person has enough expertise. As such, experts should be
able to do this task easily. However, when experts express ambivalence (even more so than uncertainty), it appears from our results that such an expression conveys the expert's inability to make a decision. It seems possible, however, that such negative responses to ambivalent experts may be alleviated when the decision is described as complex. That is, because ambivalence is a typical reaction to complexity (e.g., Larsen, McGraw, \& Cacioppo, 2001; Tiedens \& Fong, 2002), expressed ambivalence may be more palatable to observers when observers are told that the decision/diagnosis is complex rather than simple. This may obtain because complexity provides a causal explanation for any ambivalence that is experienced (i.e., this is a complex case and it therefore makes sense it is hard to distinguish between two alternatives). This explanation may in turn make the decision-maker seem more justified for expressing the state of ambivalence (see Ahn, Novick, \& Kim, 2003). The same may hold true for expressions of uncertainty. However, this complexity could actually negatively influence impressions of certain experts because it may be difficult to understand how an expert is so certain even when the problem is complex. To investigate this possibility, we manipulated the stated complexity of the decision task in Experiment 2.

In Experiment 1 we ended the description of the health care interaction before a final diagnosis was provided. We did this because we were interested in how the expression of ambivalence is interpreted as it is first encountered, regardless of what decision the expert finally comes to. However, an alternative explanation for our results is that the certain provider was viewed more favorably than the ambivalent provider because the certain provider was perceived to have actually suggested a diagnosis whereas the ambivalent provider had not. ${ }^{2}$ In Experiment 2 we accounted for this issue by adding - across all conditions -the delivery of an actual diagnostic decision at the end of the health care interaction. It is possible that providing a diagnosis may make all decision states seem equally unproblematic; people may not care how an expert decisionmaker arrives at a decision once the decision is final. If this is true, perceptions of ambivalent experts may be equated to certain and uncertain experts. However, if people are focused on the process by which the decision maker arrives at a decision, then providing an actual diagnosis may not matter for the effects of certainty.

## Experiment 2

Experiment 2 expands from the basic design of Experiment 1 by equating all conditions on the delivery of a final diagnosis. We used the same manipulations as in Experiment 1 but added that all providers came to the same final diagnosis at the end of the interaction. We also manipulated the described complexity of the decision in Experiment 2, such that we would be able to assess if

[^148]describing a categorization decision as complex changes how ambivalence is perceived. As such, in the following analyses we will compare the effects of complexity within each certainty manipulation to see if it differentially influences impressions of each expressed decision state.

## Methods

Participants Ninety-two United States based participants recruited from Amazon's Mechanical Turk participated for payment.

Materials and Procedure The same basic materials and procedure was used as in Experiment 1 with the following exceptions. Because no differences were found between medical and mental health experts in Experiment 1, we used only medical experts. In addition, we manipulated the complexity of the decision: half of the participants received information that the diagnostic decision was complex in nature (Complex condition; $n=47$ ). The remaining participants did not receive this additional information (Control condition). Finally, all participants read that the physician made a decision of one diagnosis at the end.

After reading the description, participants went on to make the same ratings as in Experiment 1 related to their perceptions of the health care provider and the patient's follow up behaviors. To measure participants' conceptions of how long people spend in different certainty states during decision-making, we asked the following: "Think about the amount of time between learning about a problem and announcing a decision related to that problem. What percent of the time do people typically experience the following states in that time period?". Participants made ratings for three states: certain, uncertain, and torn and conflicted. For each state, participants dragged a slider bar to indicate the percentage time spent in that state, with percentages for all three states adding to 100 . Participants made these ratings once with the above prompt and then again while thinking of a complex decision. Finally, participants completed a series of post-test measures that asked them to indicate if they were a health care professional, how difficult they believed medical issue diagnosis to be, the level of expertise required to practice for several different types of medical professionals, an assessment of their desire to have final decisions (i.e., the Need for Closure scale), and their political leanings. For space purposes, we do not report the findings of the time spent deciding measure or the post-test measures.

## Results and Discussion

For the following analyses, we conducted 3 (Categorization Certainty: Ambivalent, Certain, Uncertain) x 2 (Complexity: complex vs. control) between-subjects ANOVAS with simple effects analyses within each certainty level comparing the complexity conditions. Bonferroni corrections were used in all of these analyses. In all ANOVAs, there was a significant main effect of Categorization Certainty, suggesting that adding the actual
diagnosis did not equate impressions of experts across the certainty manipulation. Importantly, in all ANOVAs we found a significant interaction between Categorization Certainty and Complexity. For simplicity sake, we focus on these interactions and present only the follow up simple effects analyses. Figure 1 depicts these analyses.

We first analyzed whether describing a decision process as complex altered perceptions of experts expressing uncertainty. Impressions of Uncertain experts were more positive when the decision was described as complex relative to the control condition: They were seen as marginally less indecisive $\left(M=4.07_{\text {control }}, \mathrm{SD}=1.54 \mathrm{vs} . M\right.$ $=3.33, \mathrm{SD}=1.10 ; p=.098)$, more informed $\left(M=3.31_{\text {control }}\right.$, $\mathrm{SD}=1.05$ vs. $M=4.73, \mathrm{SD}=1.50 ; p=.002$ ), and higher


Figure 1: Mean ratings across certainty conditions. * indicates $p<.05, * *$ indicates $p<.01, * * *$ indicates $p<.001, \dagger$ indicates $.05<p<.1$. Error bars represent standard error.
quality $\left(M=2.16_{\text {control }}, \mathrm{SD}=0.94\right.$ vs. $M=4.20, \mathrm{SD}=1.45 ; p$ $<.001$ ). Predicted patient reactions to Uncertain experts were more positive in the complex condition in that patients were seen as more likely to refer them to friends ( $M=$ $2.67_{\text {control, }} \mathrm{SD}=1.35$ vs. $M=4.16, \mathrm{SD}=1.23 ; p=.002$ ) and less likely to sue $\left(M=5.67_{\text {control }}, \mathrm{SD}=1.35\right.$ vs. $M=4.13$, $\mathrm{SD}=1.89 ; p=.008$ ).

This effect of complexity was reversed in Certain experts. Certain experts were seen in the Complex condition as more indecisive $\left(M=2.28_{\text {control }}\right.$, $\mathrm{SD}=0.99$ vs. $M=3.19$, $\mathrm{SD}=$ $1.35 ; p=.042$ ), marginally less informed ( $M=5.53_{\text {control }}$, SD $=0.96$ vs. $M=4.80, \mathrm{SD}=1.02 ; p=.095$ ), and marginally lower quality $\left(M=5.13_{\text {control }}, \mathrm{SD}=0.098\right.$ vs. $M=4.31$, SD $=0.96 ; p=.058$ ). Predicted patient reactions to Certain experts were less positive in the complex condition in that patients were seen as less likely to refer the expert to friends ( $M=4.93_{\text {control }}, \mathrm{SD}=1.18$ vs. $M=3.89, \mathrm{SD}=1.15 ; p=$ .028 ), and more likely to sue ( $M=3.27_{\text {control }}, \mathrm{SD}=1.33 \mathrm{vs}$. $M=4.53, \mathrm{SD}=1.77 ; p=.028$ ).

Interestingly, complexity of decisions did not alter impressions of Ambivalent experts, with no significant differences obtaining when the diagnosis was described as complex versus not. Ambivalent experts were seen as just as indecisive $\left(M=3.87_{\text {control }}, \mathrm{SD}=1.25\right.$ vs. $M=4.42, \mathrm{SD}=$ $1.00 ; p=.20)$, just as informed ( $M=4.16_{\text {control }}, \mathrm{SD}=1.36$ vs. $M=4.43, \mathrm{SD}=1.15 ; p=.51$ ), and of equal quality $(M=$ $3.27_{\text {control }}, \mathrm{SD}=1.44$ vs. $M=3.57, \mathrm{SD}=1.15 ; p=.46$ ). Predicted patient reactions did not differ across levels of complexity for likelihood to refer the expert ( $M=4.04_{\text {control }}$, $\mathrm{SD}=1.37$ vs. $M=3.57, \mathrm{SD}=1.35 ; p=.29$ ) or likelihood to sue $\left(M=4.80_{\text {control }}, \mathrm{SD}=1.66\right.$ vs. $M=5.29, \mathrm{SD}=1.16 ; p=$ .36).

## General Discussion

Relying on experts to aid in specialized decisions is a core feature of modern human reasoning. As such, it is vitally important to understand how people think about experts and their decision-making process. We have presented one of the first explorations of impressions of ambivalent experts by investigating how people perceive ambivalent versus certain and uncertain experts within the health domain. In two experiments, we present converging evidence that expressed ambivalence is particularly costly for experts (Experiment 1) and this cost holds regardless of the complexity of the task and the determination of a final categorization decision (Experiment 2). These results suggest that when an expert expresses ambivalence about a categorization decision in his or her area of expertise, observers react negatively to this expert regardless of the complexity of the task and whether a decision is eventually made.

Why do people react so negatively to ambivalence in experts? Thinking about our specific examples, the health care experts in our experiments were ambivalent as to how to classify patients into one of two categories. As discussed earlier, the idea that an expert may have difficulty categorizing something in their area of expertise may go
against our fundamental assumptions of how categories function in the world. If an expert expresses ambivalence in categorizing something within their domain, this may challenge the belief that categories are clearly defined and can be recognized with enough knowledge. In this sense, people may feel uncomfortable with ambivalent experts because they undermine assumptions about the nature of categories in the world that fall from essentialism.

If the negative reaction to ambivalent experts stems from implications of essentialism, then this would imply that ambivalent experts should be more acceptable for categories where essences are not inferred. Medical and mental health categories are seen as possessing causal essences that define the features of the category and are necessarily possessed by members of the category (Ahn, Flanagan, Marsh, \& Sanislow, 2006; Cooper \& Marsh, in preparation). If we use a domain that was inherently less essentialized, or not essentialized at all (e.g., artifacts, nominal kinds) we may see a shift in perceptions of experts' decision-making process. For example, if we interacted with an expert who could not categorize a man-made object, we may be more accepting of this expert's ambivalence precisely because there is not a defining causal feature by which to organize the object.

A tension has formed: ambivalence improves decisionmaking (Rees et al., 2013) but is perceived negatively by laypeople when expressed by an expert. This sets up the possible recommendation of telling experts to balance different ideas and be open to feeling ambivalent during decision-making, but under no circumstances express this ambivalence to others. This clearly seems like a less than ideal recommendation given that many experts are expressly charged with communicating their decision-making process to laypeople (e.g., the shared-decision making model of medicine). Further research is needed to understand how experts can convey ambivalence and not upset the laypeople they are tasked to help, as well as to understand for whom (e.g., what types of patients) ambivalence may be more or less palatable.

## Acknowledgments

This project was funded by Lehigh University research funds.

## References

Ahn, W., Flanagan, E. H., Marsh, J. K., \& Sanislow, C. (2006). Beliefs about essences and the reality of mental disorders. Psychological Science, 17, 759-766.
Ahn, W. K., Novick, L. R., \& Kim, N. S. (2003). Understanding behavior makes it more normal. Psychonomic Bulletin \& Review, 10, 746-752.
Braisby, N. (2001). Deference in categorisation: Evidence for essentialism? In J. D. Moore \& K. Stenning (Eds.), Proceedings of the $23^{\text {rd }}$ Annual Conference of the Cognitive Science Society. Mahwah, NJ: Lawrence Erlbaum Associates.

Braisby, N. (2003). Deference and essentialism in the categorization of chemical kinds. In R. Alterman \& D. Kirsch (Eds.), Proceedings of the 25th Annual Cognitive Science Society. Mahwah, NJ: Lawrence Erlbaum Associates.
Cooper, J. A., \& Marsh, J. K. (in preparation). Medical Experts' Essentialist Beliefs for Mental and Medical Disorder Categories.
Danovitch, J. H., \& Keil, F. C. (2004). Should you ask a fisherman or a biologist?: Developmental shifts in ways of clustering knowledge. Child Development, 75, 918931. doi: 10.1111/j.1467-8624.2004.00714.x

Danovitch, J. H., \& Keil, F. C. (2007). Choosing between hearts and minds: Children's understanding of moral advisors. Cognitive Development, 22, 110-123. doi: http://dx.doi.org/10.1016/j.cogdev.2006.07.001
Diesendruck, G., \& Gelman, S. A. (1999). Domain differences in absolute judgments of category membership: Evidence for an essentialist account of categorization. Psychonomic Bulletin and Review, 6, 338346.

Estes, Z. (2003). Domain differences in the structure of artifactual and natural categories. Memory and Cognition, 31, 199-214.
Fong, C. T. (2006). The effects of emotional ambivalence on creativity. Academy of Management Journal, 49, 10161030. doi: 10.5465/amj.2006.22798182

Gelman, S. A. (2003). The essential child: Origins of essentialism in everyday thought. New York, NY US: Oxford University Press.
Larsen, J., McGraw, A., \& Cacioppo, J. (2001). Can people feel happy and sad at the same time? Journal of Personality and Social Psychology, 81, 684-696.
Rees, L., Rothman, N. B., Lehavy, R., \& Sanchez-Burkes, J. (2013). The ambivalent mind can be a wise mind: Emotional ambivalence increases judgment accuracy. Journal of Experimental Social Psychology, 49, 360-367.
Rothman, N. B. (2011). Steering sheep: How expressed emotional ambivalence elicits dominance in interdependent decision-making contexts, Organizational Behavior and Human Decision Processes, 116, 66-82.
Fong, C. T., \& Tiedens, L. Z. (2002). Dueling experiences and dual ambivalences: Emotional and motivational ambivalence of women in high status positions. Motivation and Emotion, 26, 105-121. doi: 10.1023/a:1015198209285

VanderBorght, M., \& Jaswal, V. K. (2009). Who knows best? Preschoolers sometimes prefer child informants over adult informants. Infant and Child Development, 18, 61-71. doi: 10.1002/icd. 591
Wilson, R. A., \& Keil, F. C. (1998). The shadow and shallows of explanation. Minds \& Machines, 8, 137-159.

# Is it Really that Simple? The Complexity of Object Descriptions in Human-Computer Interaction 

Vivien Mast (viv@tzi.de)<br>SFB/TR8 Spatial Cognition, I5-[DiaSpace]<br>University of Bremen<br>Evelyn Bergmann (ebergman@uni-bremen.de)<br>SFB/TR8 Spatial Cognition, I6-[NavTalk]<br>University of Bremen


#### Abstract

In the literature on verbal human-computer interaction there is general consent that humans' preconceptualisations of the machine's capabilities lead to conceptual and syntactic simplifications of the language used. We present a Wizard of $\mathrm{Oz} /$ Confederate study where humans communicate with either a system or an expert in a localization task in a complex building, in a setting which encourages them to give as much information as possible. We analyzed the syntactic complexity of object descriptions. Although we did find differences concerning the complexity of object descriptions on the clausal level, there were no significant structural differences on the subclausal level.


Keywords: human-computer interaction; syntactic variation; complexity; dialogue systems; Wizard of Oz; object descriptions

## Introduction

Imagine you were lost in a building, but a dialogue system offered help if only it could locate you by a description of your surroundings. What kinds of information would you believe to be comprehensible to the system? And how would you shape your language in order to be understood?

According to prior research, humans communicating with artificial agents tend to use language that is both conceptually and syntactically simpler than when talking to other humans (Amalberti, Carbonell, \& Falzon, 1993; Tenbrink, 2005; Moratz, Fischer, \& Tenbrink, 2001). On the other hand, we know that humans adapt to the needs of their communication partners (Clark \& Wilkes-Gibbs, 1986; Clark \& Bangerter, 2004). In interaction with artificial agents, humans often do not know what level of knowledge and competence to expect from their interlocutor, and their expectations are influenced by different sources, such as preconceptualizations, domain, robot appearance, dialogue situation, and the course of the dialogue itself (Fischer, 2011).

When designing a system for user localization in complex buildings, it is central to determine what kinds of utterances should be expected, and whether findings from human-human interaction (HHI) research serve as a good basis for system design. In this paper, we will present a comparative study of human-computer interaction (HCI) vs. HHI in a user localization scenario designed to encourage the assumption of high cognitive and linguistic system capacities. Our analysis focuses on the number and syntactic complexity of object descriptions, as they are a central part of localization dialogues.

Based on the literature we expected participants' language to be more complex when talking to the expert than when talking to the system. Finally, we will discuss the consequences of our findings for research in human-computer interaction and system design.

## Human-Computer Interaction

Amalberti et al. (1993) summarize early Wizard of $O z$ research which found that in HCI participants tend to use fewer dialogue control acts, less structured dialogue, more "standard" forms, and simpler linguistic structures than in HHI. Linguistic simplifications include fewer referring expressions, less variation of syntactic structures, shorter verbal complements and a smaller vocabulary. For example, in a study comparing typed conversations, Kennedy, Wilkes, Elder, and Murray (1988) found that participants in HCI relied on a reduced lexicon, minimized usage of pronominal anaphor, and used shorter utterances, as compared to HHI.

A number of studies also report conceptual simplifications in HCI : when giving route instructions to a system in a mapbased task, speakers mainly rely on turn-by-turn instructions, as opposed to the more complex goal-oriented descriptions usually used by humans (Tenbrink, Ross, Thomas, Dethlefs, \& Andonova, 2010). In an experiment by Moratz et al. (2001), when instructing a robot to interact with objects, users tend to use fine-grained, path-based instructions, micromanaging the robot's movements; unlike known findings in HHI, they also consistenly use the robot's perspective.

## Influences on Expectations and Behaviour

The studies mentioned here seem to give a clear picture, indicating that humans use conceptually and linguistically simpler language when speaking to an artificial agent, as compared to humans. However, linguistic behaviour depends on a number of influencing factors, and the nature of the communication partner (human vs. machine) is only one of them.

When communicating with an artificial agent, humans do not know what degree of linguistic, cognitive, and sensorimotor capacities to expect from their interlocutor, be it a robot or an information-based computer system (Moratz et al., 2001; Fischer, 2011). Therefore, they are bound to form a hypothesis based on the information available. Fischer argues that both conceptual and linguistic behaviour of humans in HCI
depend on the user's conceptualization of the agent's affordances (Fischer, 2011). She shows that this conceptualization can be partially influenced by the physical appearance of the artificial agent, but more strongly so by users' preconceptions and the dialogue flow (Fischer, 2011, 2008).

It has also been widely demonstrated that speakers adapt to their partner during the course of a dialogue (Clark \& WilkesGibbs, 1986; Clark \& Bangerter, 2004; Garrod \& Pickering, 2007). This also holds for HCI. For example, speakers show linguistic adaptation to improve understandability (Oviatt, Bernard, \& Levow, 1998). Concerning the differences between HHI and HCI, while Kennedy et al. (1988) failed to manipulate the language of the user towards a more HHI-like style by more polite machine output, Amalberti et al. (1993) show that differences between HHI and HCI decrease over time, if the interlocutor's behaviour is identical in both conditions. Also, the mode of communication influences discourse behaviour. Generally, in oral communication speakers produce longer utterances than in written communication, and use a less normative style (Chafe, 1985).

In our opinion, crucial factors in influencing user's linguistic style are the domain and dialogue task. Early Wizard of $O z$ studies centered on problems that could be solved by exchanging relevant information in short question-answer pairs, like requiring information about which cells contain which geometrical shapes (Kennedy et al., 1988). Also, it was usually very clear which kind of information would be required in order to succeed in solving the task.

An extreme example of a different domain and dialogue task is $E L I Z A$, an early conversational agent that took the role of a Rogerian psychotherapist and was designed to draw "his patient out by reflecting the patient's statements back to him." (Weizenbaum, 1976). Though mainly intended as a demonstration gimmick, people who conversed with ELIZA became "deeply [...] involved with the computer and [...] unequivocally [...] anthropomorphized it." (Weizenbaum, 1976).

In the following, we present the setup of our study which was aimed at comparing HCI and HHI in a scenario designed to encourage participants to form high expectations of their interaction partner.

## Setup of the Study

In the study presented in this paper, we relied on two strategies to create a dialogue situation that would encourage participants to speak naturally to the system.

Firstly, the setting itself was chosen to be one where the precise nature of the information needed could not be easily guessed. Participants were brought to different positions in a complex building, and engaged in a remote spoken language dialogue with either the so-called "Infocenter expert" or "Infocenter system" whose supposed task it was to locate the participants in the building. No information was given to participants about the kind of information the sytem/expert had or could process, and it is evident that such a task does not provide a clear and easy solution. Any number of objects
and their features or relations to each other could be relevant, and there are numerous ways to describe these. Therefore the task and setting itself encouraged the participants to describe as much as possible so that they could be localised.

Secondly, participants were encouraged to give detailed descriptions by employing feedback methods (see section Dialogue Flow below). This is closer to natural discourse behaviour than just shaping questions more politely, as Kennedy et al. (1988) did.

## Procedure

We conducted the study in GW2, a complex building at the University of Bremen. The building has four floors with different layouts consisting of one or two main areas. Five positions in the building with different spatial layouts ( t intersections, open spaces, and an irregular intersection) were chosen for the experiment, making sure they were sufficiently far apart to make the dialogue situation plausible.

Before the task, participants filled in a questionnaire regarding basic demographic facts, prior knowledge of the building, and the Questionnaire on Spatial Strategies by Münzer and Hölscher (2011). They were then told that they would talk to either the "Infocenter system" (system condition) or the "Infocenter expert" (expert condition) that would try to locate them in the building and ask them questions. They were instructed to answer as well as they could. In order to enable inherently plausible dialogues about the physical environment, participants were told that the use of room numbers was not allowed.

Participants were brought to each point in ascending order. They were instructed to initiate the dialogue at each position with a predefined phrase, Ich bin bereit. (I am ready.). If participants asked the experimenter about the kind of expressions or information they should use, he/she repeated that they could say whatever they wanted, except for room numbers. No further information about the task was given.

## Participants

Overall, we tested 33 participants. One participant had to be excluded from evaluation due to technical problems. Of the remaining 32 participants, 17 interacted with the system, and 15 with the expert. All participants were students at the University of Bremen and reported native or near-native competence of German. There were 26 female ( 13 per condition) and 6 male participants (expert condition: 2, system condition: 4) aged $18-31$ years (mean: 22). Prior knowledge of the building was intermediary: On a 7-point Likert scale, scores ranged from 2 to 5 in both conditions, with means of 3.18 in the system condition $(s d=1.07)$ and 3.6 in the expert condition ( $s d=0.91$ ) There was no significant difference between conditions (two-sample t-test: $t=-1.194, d f=30$, $p=0.2418$ ).

## Technical Setup

Three experimenters took turns as wizard or confederate, each experimenter playing both roles. Great care was taken to


Figure 1: Overview of the dialogue flow for a given position.
provide a technical setup that ensured equivalent behaviour.
In the system condition we used a modified Wizard of Oz setup. The participant interacted with the system via spoken language, using a headset connected to a laptop. The participant's speech was sent via a one-way skype connection to the wizard laptop. The wizard classified the user utterance using an interface implemented for this purpose, whereupon the system response was automatically determined by the interface according to the dialogue flow described below. This ensured that the wizard and confederate behaved equivalently. The system response was sent as text via a socket connection to the user laptop where it was converted to speech using the MARY text-to-speech system (Trouvain \& Schröder, 2003).

In the expert condition a computer-aided confederate setup was created. The confederate used the same interface as the wizard in order to determine the next response. Instead of text-to-speech conversion, a two-way skype connection was used. The response text was shown on the wizard interface, and was then spoken by the confederate. Confederates were instructed to conform as closely to the wording of the response utterance as possible while still maintaining natural speech rather than reading out loud.

## Dialogue Flow

Care was taken to create a dialogue flow where participants felt an expectation to give as much information as possible without rendering the dialogue unnatural. Therefore, system/expert responses were designed to be as open as possible, and to give the general impression that the previous utterance(s) had been understood. In this respect, our dialogue flow was partly inspired by ELIZA. Although our wizard interface did not perform transformations on user utterances, it did use object names users uttered (which were typed in by the wizard/confederate) for generating replies, giving the impression that the user had been understood.

Additionally, in order to increase the naturalness of the dialogue and give the impression of high verbal capacity of the system, we produced a number of variant utterances for each system response type. The response variants were checked by two independent coders for equivalence. Responses that were semantically or pragmatically more constrained than the de-

| Code | Example utterance |
| :--- | :--- |
| WHERE | Bitte sage mir, wo du gerade stehst. <br> Please tell me where you are currently standing. |
| WHERE- | Wo genau in diesem/dieser <NAMED ELEMENT $>$ bist du gerade? <br> EXACTLY <br> Where exactly in this $<$ NAMED ELEMENT $>$ are you now? |
| WHAT- | Beschreibe mir bitte, was du von deiner momentanen Position aus <br> sehen kannst. <br> Please describe to me what you can see from your current position. |
| NO-INF | Ich habe leider keine Informationen über Einrichtungsdetails wie <br> $<$ NAMED OBJ $>$ oder <RANDOM IRRELEVANT OBJ $>$. <br> Unfortunately I don't have information about details of interior <br> furnishing such as <NAMED OBJ $>$ or $<$ RANDOM IRRELEVANT <br> OBJ $>$. |
| MORE | Was kannst du mir noch über diese(s/n) $<$ NAMED OBJ $>$ sagen? <br> What else can you tell me about this $<$ NAMED OBJ $>$ ? |
| ELSE- | Erzähle mir, was du dort noch sehen kannst. <br> Tell me what else you can see there. |
| ENE | Okay, ich habe herausgefunden wo du bist. Du kannst jetzt zum <br> nächsten Standpunkt gehen. <br> OkART |
| tion now! |  |

Table 1: Codes and example utterances for the different utterance types of the expert/system.
sired response were discarded.
Before the first position, the participant was greeted by the system/expert, and after the last position, the system/expert thanked them for their participation. As a general rule, the system/expert gave a success or failure message after the user had named at least 5 types of objects. Success and failure were predetermined and did not depend on the performance of the user. Position 2 led to failure, all others to success.

If less than 5 types of objects had been named, the system asked questions to elicit further information, depending on the course of the interaction. The detailed dialogue flow can be seen in Figure 1. Examples for each type of system/expert utterance are given in Table 1; the beginning of a dialogue between the system and a participant is shown in Table 2.

| Sp. | Code | Utterance |
| :--- | :--- | :--- |
| S | WHERE | Wo stehst du gerade? <br> Where are you standing now? |
| U |  | Ich stehe vor einem Gang, der durch eine Glastür geht und links <br> und rechts von mir ist jeweils auch ein Gang. <br> I am standing in front of a corridor that goes through a glass <br> door and left and to my left and right is also a corridor each. |
| S | MORE <br> -INF | Beschreibe mir bitte diesen Gang etwas genauer. <br> Please describe this corridor to me in a bit more detail. |
| U |  | Ahm, der Gang vor mir geht durch, äh, eine Glastür und ist noch <br> sehr lang. Ähm, von dem Gang gehen viele Türen ab. Der Gang <br> rechts von mir ist sehr kurz. Ähm, in diesem Gang befinden sich <br> zwei Säulen und der Gang links von mir ist auch relativ lang, <br> ähm. Geht aber ziemlich bald durch eine Glastür und auch in <br> diesem Gang befinden sich viele weiße Säulen. <br> Erm, the corridor in front of me goes through, erm, a glass door <br> and ist very long. Erm, from the corridor many doors go off. The <br> corridor to my right is very short. Erm, in this corridor there <br> are two pillars and the corridor to my left is also relatively long, <br> erm. But goes also quite soon through a glass door and also in <br> this corridor there are many white pillars. |
| S | ELSE- <br> SEE | Kannst du noch mehr sehen? <br> Can you see anything else? |
| U |  | Hinter mir ist ein Zeichen für einen Feuermelder, verschiedene <br> Informationsplakate. <br> Behind me is a sign for a fire-alarm, different information <br> posters. |

Table 2: Beginning of a dialogue between a user (U) and the system (S).

## Analysis

One major strategy for localization dialogues is a description of the current view, the spatial scene that surrounds the user. When asked where they were standing, participants answered with lists of concrete physical objects and optionally their relative position, but also with higher level descriptions of corridor constellations and region names that were inferred from signs or retrieved from knowledge. Another description strategy was to provide a route description to the current position. In our analysis, we focused on descriptions of objects. However, these could include descriptions of potential actions, as will be explained below.

The structure and complexity of object descriptions has been analyzed mainly from the point of view of how humans establish joint reference (Clark \& Wilkes-Gibbs, 1986; Clark \& Bangerter, 2004) and computational generation of referring expressions (Bohnet \& Dale, 2005). A referring expression has the structure of a more or less complex noun phrase, modified with adjectives of colour or form, or prepositional phrases that indicate parts or spatial location. However, in the scenario presented here, people give descriptions of objects to an interlocutor without knowing the amount and nature of information he/she has available. The goal is to provide information about the scene, therefore the syntactic structures are more complex than those of referring expressions.

On the other hand, people usually rely on two main strategies when describing complex multi-object scenes like a room, a city, or a desktop array. Either the discourse is organized sequentially as a mental tour or the objects are described in lines as a parralel structure (Linde \& Labov, 1975; Ullmer-Ehrich, 1982). Object configurations are referred to sequentially or in clusters, and usually the object's location is specified and not its orientation (Tenbrink, Coventry, \& Andonova, 2011). Although it could be expected in the current task that participants relied on the well-documented strategies for room descriptions, interestingly, they did not. For the purpose of localisation, it seems, the participants did not aim for completeness, hence no systematical discourse organisation.

In our analysis, we used object descriptions as the main unit of analysis, taking into account structures more complex than referring expressions, but below the level of full scene descriptions. Based on transcriptions of the original audio recordings, coders identified all object descriptions. The beginning of an object description was identified as follows: Any object that was introduced by a noun in a main clause in rheme position: Es gibt/ ich sehe/ ich stehe vor einer Treppe. (There is/ I'm looking at/ standing in front of a staircase) or in an elliptic main clause: [5] Und dann noch so ne Treppe. ${ }^{1}$ (and also a staircase), was regarded to be the beginning of an object description, unless the clause in question was the continuation of a prior object description.

Once a new description had been identified, all parts of the utterance that preserved anaphoric reference to the described

[^149]object were considered continuations of the given object description. Clauses containing repetitions or reformulations of the object name were considered continuations of the object descriptions only if they did not introduce a new object in rheme position.

## Categories of Elaborations

To address the complexity of object descriptions as explained above, we analyzed the number of attributes and elaborative features (henceforth elaborations) directly relating to the target object on the subclausal level, and the number and type of clauses of the object description. Elaborations and clauses were classified into 8 categories post-hoc on the basis of the data as described below. Annotations were carried out by 3 independent coders. Intercoder reliability was checked for by independent double coding for a subset of $10 \%$ of the data. Levels of agreement were either good or very good: Krippendorff's alpha computed on each of the 8 categories ranged between 0.785 (Adverbial Attributes) and 0.945 (Pronominal Clauses).

Compound Name: A compound noun was counted, if the initial noun describing the object was modified by a morpheme, but not if it was a simplex noun: [2510] KünstlerBüro (artist's office)

Adjective Attribute: indicates the number of dependent adjective attributes of the object: [5] so 'ne blaue Treppe (such a blue staircase)

Prepositional Attribute: number of dependent prepositional phrase attributes: [2320] eine Treppe mit Glaswänden (stairs with glasswalls)

Genitive Attribute: number of dependent genitive attributes: [2320] im Erdgeschoss, äh, des GW2 (on the first floor of the GW2)

Adverbial Attribute: number of adverbial attributes: [17] draußen im Flur (outside, on the corridor)

Pronominal Clause: number of dependent pronominal clauses: [33] Ähm, links von mir ist wieder so 'n Eingang, wo die Haupttreppe zum Kunstbereich kommt. (On my left is an entrance, where you can enter the art department)

Conjunction Clause: number of elaborating subordinate clauses introduced by conjunctions: [1152] Es ist ein Holzbrett davor, um die Tür aufzumachen. (There is a wooden piece in front of it to open the door)

Main Clause: indicates the number of main clauses which elaborate on the aforementioned object. This includes the primary introductory clause, and further clauses connected via 1) anaphoric pronouns „der, die, das, es, da": [9]Ähm, ich seh hier Zeitungen, Flyers- Die sind rechts von mir (I see magazines, Flyers- They are on my right); 2) they explicitly refer back to an object from the discourse history: Ähm, ich seh hier Zeitungen, Flyers. Die Flyers liegen rechts von mir
(I see magazines, flyers. The Flyers are on my right) 3) or they elaborate on parts of the aforementioned object which appear in theme position. Elaborations via main clauses can also be connected via ,also, und, oder" (thus, and, or). In this case, they are only counted as elaborations if they give further information about the described object, and do not introduce new objects. [1942] Und, ähm, man kann sich das vorstellen wie ein, wie ein Dreieck. Also ich stehe gerade, ich kann quasi geradeaus gehen, nach links oder nach rechts. (This is like a triangle. So, I can walk straight, to the left, or to the right)

## Results

In this section, we present our findings with respect to the number of object descriptions given and the usage of the different types of elaborations and clauses. The mean number of object descriptions produced by participants at each position, and the overall mean number of descriptions per position per speaker are shown in Table 3.

| Position | Expert | System |
| ---: | ---: | ---: |
| 1 | 8.87 | 7.59 |
| 2 | 11.07 | 6.65 |
| 3 | 11.13 | 6.82 |
| 4 | 10.33 | 7.18 |
| 5 | 10.13 | 7.59 |
| overall means | 10.31 | 7.16 |

Table 3: Mean number of object descriptions produced by speakers of both conditions at each position.

For the number of object descriptions given per position, we fitted a linear mixed model to the data, including fixed effects for Condition, Position and their interaction and a random intercept and random slope (with respect to Position) for each participant (Pinheiro \& Bates, 2000). There was a significant main effect of condition (HHI vs. HCI) $(F=5.37$ and $p=0.028$ ), but not for position ( $F=0.29$ and $p>0.05$ ). No interaction effect was found ( $F=1.16$ and $p>0.05$ ). In summary, this indicates that participants gave more object descriptions when talking to an expert than when talking to a system, and that there was no significant change in the number of object descriptions given during the course of the interaction.

## Complexity of Object Descriptions

Table 4 shows the frequency of the different types of elaborations and clauses in the corpus in each condition, and the mean frequency of each feature per object description. As can be seen from this table, all feature types are present in both system and expert condition, indicating that humans in both HHI and HCI use the full range of syntactic possibilities for describing objects. As Table 5 shows, while the relative frequency of subclausal elaborations per object description shows only a small difference between conditions, the relative number of clauses per description shows a larger difference.

For the number of clauses per object description, we fitted a linear mixed model with fixed effects for Condition, Position

| Elaborative Feature | Expert |  |  | System |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $\begin{aligned} & \hline \text { (in } \\ & 773 \\ & \text { OD) } \\ & \hline \end{aligned}$ | $\begin{aligned} & \text { mean } \\ & \text { per } \\ & \text { OD } \end{aligned}$ | $\begin{array}{ll} \hline \% & \text { of } \\ \mathrm{E} & \end{array}$ | $\begin{aligned} & \text { (in } \\ & 609 \\ & \text { OD) } \\ & \hline \end{aligned}$ | mean per OD | $\begin{array}{ll} \hline \% & \text { of } \\ \mathrm{E} & \end{array}$ |
| Compound Names | 276 | 0.36 | 10.57 | 219 | 0.36 | 12.01 |
| Adjective Attributes | 305 | 0.40 | 11.69 | 199 | 0.33 | 10.91 |
| PP-Attributes | 324 | 0.42 | 12.41 | 277 | 0.46 | 15.19 |
| Genitive Attributes | 7 | 0.01 | 0.27 | 3 | 0.00 | 0.16 |
| Adverbial Attributes | 251 | 0.32 | 9.62 | 173 | 0.28 | 9.48 |
| Main Clauses | 1147 | 1.48 | 43.95 | 816 | 1.34 | 44.74 |
| Promoninal Clauses | 209 | 0.27 | 8.01 | 87 | 0.14 | 4.77 |
| Conjunction Clauses | 91 | 0.12 | 3.49 | 50 | 0.08 | 2.74 |
| Total | 2609 |  | 100.00 | 1824 |  | 100.00 |

Table 4: Frequency of the different elaboration or clause types (E) when speaking to the expert vs. the system. The table shows total frequency in the corpus, and mean frequency per object description (OD).

| Feature Type | Expert |  |  | System |  |  |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: |
|  | total | mean | $\% \mathrm{E}$ | total | mean | $\% \mathrm{E}$ |
| Clausal | 1446 | 1.87 | 55.42 | 953 | 1.56 | 52.25 |
| Subclausal | 1163 | 1.50 | 44.58 | 871 | 1.43 | 47.75 |
| Total | 2609 |  | 100.00 | 1824 |  | 100.00 |

Table 5: Number of clauses and non-clausal elaborating features (E) in each condition. The table shows total frequency in the corpus, and mean frequency per object description (OD).
and their interaction and a random effect of Participant on the intercepts. We found a statistically significant effect of condition $(F=10.55$ and $p=0.003$ ), but no effect for position ( $F=2.17$ and $p>0.05$ ), and no interaction $(F=0.43$ and $p>0.05$ ), indicating that participants speaking to the expert used significantly more clauses per object description than those speaking to the system, regardless of the position. For the number of subclausal elaborations per object description, we fitted a linear mixed model with fixed effects for Condition, Position and their interaction and a random effect of Participant on the intercepts. No effect was found for condition ( $F=0.49$ and $p>0.05$ ), but a significant effect for position ( $F=6.41$ and $p<0.0001$ ), and no interaction ( $F=1.31$ and $p>0.05$ ), showing that on the subclausal level there was no systematic difference between HHI and HCI in our study. Using contrasts to break down the effect of position, a significant linear trend was found ( $t=3.32, p=0.001$ ), indicating a linear increase in subclausal elaborations in the course of the interactions.

## Discussion

In this paper, we have examined the differences between human-human und human-computer interaction in a user localization scenario designed to encourage participants to develop high expectations of the linguistic and cognitive capacities of an artificial communication partner. We have analyzed the number of object descriptions and their complexity as represented by 8 types of elaboration in HHI and HCI. Although the number of object descriptions given overall, and the number of clauses within these descriptions was higher for HHI than for HCI, participants in the HCI scenario showed the
full range of syntactic variability. They used all types of clauses and subclausal elaborations in sufficiently high frequency that they cannot be discarded as exceptions. Particularly, for subclausal elaborations, no significant difference in frequency between HHI and HCI could be found. These findings support our assumption that, given the appropriate scenario, HCI can be fairly natural and more similar to HHI than may be expected. In our opinion, this contradicts strong claims of computer talk as a separate register which is per se distinct from HHI (Zoeppritz, 1985). Rather, HCI shows parallels with intercultural communication where a number of individual and situational factors come together to shape (linguistic) behaviour, mediated by the interactant's conceptualizations (Fischer, 2011, 2007).

With regard to system design, the broad variability of the user's utterances shows that there is no way around developing systems with high verbal skills, which includes grammatical as well as conceptual competence. Future research could focus on further determining the influences and boundaries for shaping humans' linguistic behaviour towards artificial agents. Focusing on system design, this could help answer the question of how to frame human-computer interactions in a way that the users' expectations and the competence of the system are well matched.

## Acknowledgments

This research was supported by the $\mathrm{SFB} / \mathrm{TR} 8$ Spatial Cognition (Deutsche Forschungegemeinschaft, DFG). We would also like to thank the I5-[DiaSpace] and I6-[NavTalk] project groups, and especially Thora Tenbrink for support and insightful discussions, and Daniel Couto Vale and Mohammed Fazleh Elahi for their support with the design, technical setup, and data collection. We also thank our student assistants, Nadine Hagemann, Denise Rathjen, Gesa Schole, Christina Freihorst, Jonathan Burke and Verena Seidel for their reliability and speed with data collection and coding.

## References

Amalberti, R., Carbonell, N., \& Falzon, P. (1993). User representations of computer systems in human-computer speech interaction. International Journal of Man-Machine Studies, 38(4), 547-566.
Bohnet, B., \& Dale, R. (2005). Viewing referring expression generation as search. In Proceedings of the 19th international joint conference on artificial intelligence (pp. 1004-1009). San Francisco, CA: Morgan Kaufmann.
Chafe, W. L. (1985). Linguistic differences produced by differences betweeen spaking and writing. In D. Olson, N. Torrance, \& A. Hildyard (Eds.), Language, literacy, and learning (p. 105-123). Cambridge University Press.
Clark, H. H., \& Bangerter, A. (2004). Changing ideas about reference. In D. Sperber \& I. A. Noveck (Eds.), Experimental pragmatics. Hampshire, NY: Palgrave Macmillan.
Clark, H. H., \& Wilkes-Gibbs, D. (1986). Referring as a collaborative process. Cognition, 22(1), 1-39.
Fischer, K. (2007). Mensch-Computer-Interaktion als interkulturelle Kommunikation. In C. Sandten, K. Starck, \& M. Schrader-Kniffki (Eds.), Transkulturelle Begegnungen (pp. 35-50). Trier: Inputs.
Fischer, K. (2008). The role of user's concepts of the robot in human-robot spatial instruction. In T. Barkowsky,
M. Knauff, G. Ligozat, \& D. R. Montello (Eds.), Spatial Cognition V: Reasoning, Action, Interaction. International Conference Spatial Cognition 2006, Bremen, Germany, September 24-28, 2006, Revised Selected Papers. LNCS vol. 4387. (pp. 76-89). Springer.
Fischer, K. (2011). How people talk with robots: Designing dialogue to reduce user uncertainty. AI Magazine, 32(4), 31-38.
Garrod, S., \& Pickering, M. J. (2007). Alignment in dialogue. In G. Gaskell (Ed.), Oxford handbook of psycholinguistics. Oxford University Press.
Kennedy, A., Wilkes, A., Elder, L., \& Murray, W. S. (1988, oct). Dialogues with machines. Cognition, 30(1), 37-72.
Linde, C., \& Labov, W. (1975). Spatial networks as a site for the study of language and thought. Language, 50(4), 924-939.
Moratz, R., Fischer, K., \& Tenbrink, T. (2001). Cognitive modelling of spatial reference for human-robot interaction. International Journal On Artificial Intelligence Tools, 10(4).
Münzer, S., \& Hölscher, C. (2011). Entwicklung und Validierung eines Fragebogens zu räumlichen Strategien (Development and validation of a self-report measure of environmental spatial strategies). Diagnostica, 57(3), 111125.

Oviatt, S., Bernard, J., \& Levow, G.-A. (1998). Linguistic adaptations during spoken and multimodal error resolution. Language and Speech, 41(3-4), 419-442.
Pinheiro, J. C., \& Bates, D. M. (2000). Mixed-effects models in s and s-plus. New York, NY [u.a.]: Springer.
Tenbrink, T. (2005). Identifying objects on the basis of spatial contrast: An empirical study. In C. Freksa, M. Knauff, B. Krieg-Brückner, B. Nebel, \& T. Barkowsky (Eds.), Spatial cognition iv. reasoning, action, interaction (Vol. 3343, p. 124-146). Springer Berlin / Heidelberg.

Tenbrink, T., Coventry, K. R., \& Andonova, E. (2011). Spatial strategies in the description of complex configurations. Discourse Processes, 48(4), 237-266.
Tenbrink, T., Ross, R. J., Thomas, K. E., Dethlefs, N., \& Andonova, E. (2010). Route instructions in map-based hu-man-human and human-computer dialogue: A comparative analysis. Journal of Visual Languages and Computing, 21(5), 292-309.
Trouvain, J., \& Schröder, M. (2003). The German text-tospeech synthesis system MARY: A tool for research, development and teaching. International Journal of Speech Technology, 6, 365-377.
Ullmer-Ehrich, V. (1982). The structure of living space descriptions. In R. J. Jarvella \& W. Klein (Eds.), Speech, place, and action (p. 219-249). Chichester: Wiley.
Weizenbaum, J. (1976). Computer power and human reason: From judgment to calculation. San Francisco: Freeman.
Zoeppritz, M. (1985). Computer talk? (Tech. Rep. No. 85.05). IBM Scientific Center Heidelberg.

# Agency Intuitions in Physical Interactions 

Ralf Mayrhofer (rmayrho@uni-goettingen.de) Michael R. Waldmann (michael.waldmann@bio.uni-goettingen.de)<br>Institute of Psychology, University of Göttingen, Gosslerstraße 14, 37073 Göttingen, Germany


#### Abstract

The question how agent and patient roles are assigned to causal participants has largely been neglected in the psychological literature on force dynamics. Based on the linguistic theory of Dowty (1991), we propose that agency is a prototype concept. We adapted Dowty's theory to account for scenarios showing physical interactions. In the standard Michotte launching scenario the ball entering the scene is usually assigned the agent role, whereas the ball that is being launched is viewed as the patient. We showed in two experiments that agency intuitions were moderated by manipulations of the context prior to the launching event. Altering features such as relative movement, sequence of visibility, and self-propelled motion tended to increase agency attributions to the patient relative to the standard scenario. We suspect that shifts in figure-ground perceptions, and intuitions about characteristics of interventions may be the overarching reason for the efficacy of the tested criteria.


Keywords: force dynamics; causal reasoning; agency; Michotte task; physical causality

## Introduction

Currently there is a debate between two competing frameworks modeling causal reasoning. One prominent class is dependency theories, including covariation theories, counterfactual theories, and causal Bayes nets. The ontology expressed by these theories contains causal variables that either encode the presence or absence of events, facts, and properties, or different values of continuous quantities. These variables are interconnected by causal arrows that represent hidden mechanisms, and whose strength can be numerically expressed by causal strength parameters (see Waldmann \& Hagmayer, in press, for an overview).

A completely different view answers the question why an observed lawfulness holds by focusing on the participants involved in a causal relation, for example Ball A and Ball B in Michotte's task, or Aspirin and a person with headache in a medical scenario. One variant of this view, dispositional theories of causation, would say, for example, that the ingestion of Aspirin relieves headaches because Aspirin has an intrinsic property, a disposition (or capacity or power), to relieve headaches in suitable organisms (see, for example, Gnassounou \& Kistler, 2007; Mumford \& Anjum, 2011).

In psychology force dynamics, an example of a dispositional account, has become increasingly popular in recent years. Pinker (2007) has argued that force dynamics is a major competitor of Bayes net theories because it allows us to model intuitions about the generative processes underlying observed covariations. One attractive feature of dispositional theories and force dynamics in particular is that these
theories are capable of expressing abstract intuitions about mechanisms without requiring elaborate knowledge.

Force dynamics has been initially been developed in linguistics in the context of verb semantics (see Riemer, 2010; Talmy, 1988) but uses concepts that can be traced back to Aristotle (see Gnassounou \& Kistler, 2007). Aristotle explained efficient causation as a consequence of the interaction of two entities, an agent and a patient. An agent is, according to Aristotle, a substance operating on another substance, the patient, which is suffering the process of change. The acting agent who affects the patient therefore has the disposition, capacity or power to act; and the patient has the disposition, capacity or power to undergo the agent's action.

In linguistic theories of verb semantics and argument structure verbs place constraints on the possible participants mentioned in the noun phrases. For example, in "Peter pushes Mary", "push" has two arguments, one describing an agent (Peter), the other the patient (Mary). Typically, agents are assigned the syntactic subject position. Other participant roles (also called thematic or theta roles) have been postulated but there is no agreement in linguistics about the proper list (see Riemer, 2010, for an overview). Another important semantic theory for a theory of causation is Talmy's (1988) theory of force dynamics. He argues that intuitions about the interaction of forces are an important component of our general semantic intuitions.

Using a force dynamics framework, White (e.g., 2006, 2009) demonstrated the difference between intuitive causal representations and physics by studying Michotte type launching events. In Michotte's (1963) famous demonstrations of phenomenal causality, subjects observed moving objects. For example, in a launching scenario Object X, a ball, moves towards Object Y, another ball, and touches it. This stops Object X and sets Object Y into motion at the same or a slightly lesser speed. Observers typically describe this scenario as a case in which the movement of Object Y is caused by Object X (i.e., launching). Although according to Newtonian physics the force on body Y exerted by body X is equal in magnitude but opposite in direction to that on body X exerted by body Y, observers often see Object X as the cause and Object Y as the effect (causal asymmetry). Nobody would describe the scenario as a case of Object Y stopping Object X , although this would be a legitimate description.

The impression of causal asymmetry is also reflected in judgments of force. White (2009) presented participants with different launching events, and asked them to provide estimates of the relevant underlying forces. The results
showed that in such events more force is attributed to Object $X$ than to Object $Y$, and that Object $X$ is viewed as active and exerting a force on Object Y, whereas the initially stationary Object Y is viewed as inactive, exerting resistance to being moved. Thus, causal interactions are perceived as the result of the opposition between forces of agents (e.g., Object X ) and resistance of patients (e.g., Object Y).

A related theory that aims at elucidating our understanding of abstract causal concepts, such as "cause", "prevent", and "enable", is Wolff's (2007) theory of force dynamics (see also Wolff \& Song, 2003; Wolff et al., 2010). As in the theories of White (2009) and Talmy (1988), two entities are distinguished, which Wolff calls affectors and patients (i.e., the entity acted upon by the affector)(Talmy labels them antagonist and agonist). Force theory states that people evaluate configurations of forces attached to affectors and patients, which may vary in direction and degree, with respect to an endstate, that is, the possible result. Forces can be physical, psychological (e.g., intentions) or social (e.g., peer pressure). Causal relations are analyzed in terms of three components, (a) the tendency of a patient for an endstate, (b) the presence or absence of concordance between affector and patient, and (c) the degree to which the endstate is reached. For example, force theory would represent the singular causal fact "Winds caused the boat to heel" in terms of a patient (the boat) that had no tendency to heel (Tendency $=$ No), the affector (the wind) acted against the patient (Concordance $=$ No), and the result (heeling) occurred (Endstate approached = Yes).

Empirical support for the model was provided in a series of experiments in which participants made judgments about 3-D animations of realistically rendered objects (e.g., moving boats on a lake) with trajectories that were wholly determined by the force vectors entered into a physics simulator (see also Beller et al., 2009; Wolff et al., 2010; for further developments).

## The Empirical Basis of Agency Intuitions

In psychological research on force dynamics the main focus has been on how causal intuitions can be predicted on the basis of configurations of forces attached to agents and patients. The assignment of the roles of agent and patient to the causal participants has typically been treated as selfevident. In Wolff's (2007) example "Winds caused the boat to heel" there is no question that the winds should be assigned the agent role because obviously they play the role of actively overcoming the passive tendency of the boat. However, one deficit of current psychological versions of this theory is that no systematic set of empirical criteria has been laid out that unambiguously motivates the assignments of the agent and patient roles.

White's $(2006,2009)$ theory represents progress in this regard because he has pointed out one important criterion in the Michotte tasks, movement. For example, in a typical study White presented situations in which one ball stands still, and the other moves toward this ball launching it to the other side. In this scenario the moving ball is clearly as-
signed the role of the agent. However, when both balls were moving the assignment was less clear, and additional assumptions had to be made for agency assignments (see White, 2012, for an extended variant of White's, 2009, theory to predict pushing vs. pulling intuitions in cases in which both balls leave the scene attached to each other).

Another example of the ambiguity of agency assignments comes from a recent study by Mayrhofer and Waldmann (2013). In this study, the alien mind reader task was used that had been introduced in the literature by Steyvers et al. (2003; see also Mayrhofer, Hagmayer, \& Waldmann, 2010, for details). This scenario describes a set of aliens some of which have the capacity of picking up the thoughts of each other. In our studies the causal dependency relations were kept constant. In general a causal transmission process was described in which the thoughts of one alien (i.e., the cause) were transmitted into the heads of three other aliens (i.e., the effects). What we manipulated were subjects' assumption about the underlying dispositions responsible for the observed causal transmission. In one condition, the cause alien was assigned the role of the agent. Here the instructions stated that the cause alien has the capacity to send out his thoughts, and plant them into the heads of the effect aliens. In the contrasting condition, the effect aliens were described as the agents, being capable of reading the thoughts of the cause alien. We empirically ascertained that subjects shared our intuitions about the different assignments of the agent role. Interestingly, the results were clear-cut in the sender but ambiguous in the reader scenario. In the sender condition it was clear that the cause alien was the agent. However, in the reader scenario the intuitions did not uniquely attribute agency to the reader side, but equally divided agentive responsibility to the two sides, cause and effects. This may be analogous to the intuition that, although radios play an important part in picking up radio waves, the sending station also plays an active role.

What these results show is that it cannot always be a priori determined how the agency role is assigned, and sometimes the complementary participants may be both viewed as equally active. Hence our current goal is to empirically investigate empirical indicators of agency.

We are not the only ones who noticed that occasionally it is difficult to determine who should be assigned the agent role. For example, in the sentence "John hits Mary" John is clearly assigned the agent role. This example might suggest that grammatical subjects encode agent roles. However, in "John admires Mary", both participants play an active role so that a clear assignment is often impossible (similar with other psychological verbs, such as "mind reading")(see Dowty, 1991). Other ambiguous verbs include "buying" and "selling." They both require two active participants, and it is hard to uniquely assign the agent role.

Our experiments represent an initial attempt in a physical domain (launching events) to study factors moderating agency assignments. As a heuristic for criteria we will use Dowty's (1991) linguistic theory of the distinction between (proto-)agents and (proto-)patients. According to Dowty,
agency is a prototype concept that can be assigned on the basis of a number of empirical criteria. None of these criteria is necessary (hence prototype) but the confidence of the assignment should increase the more criteria are present. In Dowty's theory the agent features include (a) volitional involvement in the event or state, (b) sentience (and/or perception), (c) causing an event or change of state in another participant, (d) movement (relative to the position of the other participant), (e) exists (independently of the event named by the verb). The complementary patient features include (a) undergoes change of state, (b) incremental theme, (c) causally affected by another participant, (d) stationary relative to movement of another participant, (e) does not exist (independently of the event named by the verb). According to Dowty, when two participants are involved in a scenario, the relative number of properties from these lists decides about the assignment of roles. If there is an impasse, multiple assignments are possible. Dowty's criteria are developed to capture semantic implications of verbs. In scene perception other cues might additionally be used, including covariation.

How can this theory be applied to launching events? Obviously some of the criteria (e.g., sentience) do not apply. The remaining relevant criteria for agency include volitional involvement, causation, and relative movement. We believe that a unifying principle behind these three criteria is provided by the intervention concept popular in both the dependency account (Woodward, 2003) and force dynamic theories (White, 2006). In fact, White (2006) believes that force dynamic intuitions are grounded in sensomotoric experiences of actors sensing resistance from the objects they are attempting to manipulate.

The prototypic agent is a human confronted with a stationary scenario that either is constant or changes in a predictable way (i.e., the ground)(relative movement). The agent's act, which is considered independent of the target system, creates a change in a variable, which in turn affects other variables (i.e., the figure)(volition; causation). Following developmental evidence (Muentener \& Carey, 2010) we believe that people also make agency attributions when some of the features of the prototype are missing. For example, when no human agent is visible, the object behaving like it was manipulated by a hidden agent plays the role of the agent (e.g., one ball in the Michotte task). Apparently uncaused covariation against an invariant background has features of an intervention, which explains why the entity involved in the covariation will play the role of the agent. White's (2007) finding that the moving participant is assigned the agent role is also captured because a typical explanation of an apparently unmotivated movement is that there may be an invisible force causing it.

This theory, although derived from a force dynamic framework, is reminiscent of theories proposed within the dependency paradigm. The distinction between figure and ground is analogous to Cheng and Novick's (1991) criteria of the distinction between causes and enabling conditions. According to their theory, causal events that remain invari-
ant within a focal set are assigned the enabler role, whereas the event covarying (i.e., changing) with the effect within the focal set is the cause. This theory does not distinguish between agents and patients, however. Another closely related theory is the intervention account of causal Bayes net theory (Woodward, 2003). According to this theory a change of a variable by a free agent qualifies as an intervention. Thus, volitional involvement, movement, and causation are hallmarks of this concept. Note that this theory is not restricted to human agents. Every change of a variable that deterministically sets the variable and has characteristics of statistical independence with respect to the target system can play the formal role of an intervention.

Although there are analogies between force dynamic and dependency theories, there are also differences. The dependency theories mainly focus on covariation and event causation, whereas other criteria we will study, including relative or spontaneous movement, are neglected.

Our main empirical strategy will be to present scenarios involving an event with two participants (e.g., two balls) but manipulate across conditions properties of the participants possibly relevant for the assignment of participant roles (agent, patient). Our aim is to show that intuitions based on the proto-agency theory predict whether a participant in a fixed scenario is assigned the agent or patient role (or both).

## Experiment 1

In our experiments we employed variants of the Michotte task, which has been used as a classic demonstration for the usefulness of force theories. White $(2006,2009)$ has extensively studied this task, and has found a causal asymmetry effect: Agents are typically assigned greater force than patients. Another observation consistent with causal asymmetry is that the agent, for example a Ball X that is moving toward a stationary Ball Y, is typically described as causing the movement, but Ball Y is never described as stopping Ball X. Our goal is to manipulate the Michotte task in a way that either Ball X (the pushing ball) or Ball Y (the pushed ball) are more or less viewed as agents. In line with the proto-agency theory we predict that subjects differentiate between a stationary scenario (the ground) and an event that shares properties with hypothetical interventions. In Michotte's task a stationary scenario either consists of a set of balls at rest, or balls that are constantly moving in a predictable way. Given that no volitional agent (i.e., a human) is visible, other properties of causal agency apply.

As baseline condition (Condition A), we used the standard Michotte launching setup that was also used by White (2006, 2007, 2009): Ball Y is at rest in the middle of the display, Ball X is constantly moving and rolls from the left edge toward Ball Y. After contact, Ball Y moves and Ball X is at rest. In this condition Ball X should clearly be seen as the agent and Ball Y as the patient.

In three further conditions, we manipulated agency indicators for Ball Y while holding the properties of the physical interaction (i.e., the collision event) constant. Thus, in all conditions Ball Y is at rest in the middle of the screen im-
mediately prior to the collision. When Ball X hits Ball Y, then Ball X stops and Ball Y moves (with exactly the same speed and direction as Ball X had prior to the collision) towards and then beyond the edge of the screen.

What we manipulated were the conditions prior to the invariant launching event. In Conditions B and C, we manipulated relative movement. In both conditions Ball Y enters the screen from the bottom, moves towards its position in the middle of the screen, and stops there just before it is hit by Ball X. Thus, we added movement as an agency indicator expecting an increase of agency intuitions regarding Ball Y. We suspected that observing the movement of Ball X toward the middle position might lead some subjects to infer an intention to stop Ball X. Given that this possible intention is not successful (Ball Y will be launched by Ball X), this may not fully overcome the assignment of agency for Ball X, but a difference to the standard condition may be expected.

Furthermore, we manipulated which of the balls was seen first by restricting the section of the scene that is visible to the subject. In Condition B, we hid the left hand side margin of the scene; whereas in Condition C the lower margin was hidden. Thus, in Condition B the movement of Ball Y was seen first (i.e., Ball Y is the ground); in Condition C Ball X was already seen moving when Ball Y enters the scene (i.e., Ball Y is the figure). Our goal motivating the sequence of visibility was to test whether this subtle figure-ground manipulation affects agency attributions despite the fact that the underlying physical events are identical across the two conditions. We predicted that viewing Ball Y first as the figure would increase agency attributions regarding this ball.

In Condition D, Ball Y is at rest outside the trajectory of Ball X in the lower part of the screen. Suddenly, Ball Y starts to move so that it ends up in the same position as in the other conditions. Here the constant movement of X should be viewed as stationary (i.e., ground) with Ball Y behaving like an animate volitional agent. It is well known that spontaneous movement is seen as an indicator of animacy. In this condition the intuition that the self-propelled movement of Ball Y is a result of volition should be strongest, which should lead to the strongest agency intuitions within the set of conditions.

To sum up: From Condition A to Condition D we added more and more agency indicators for Ball Y (relative movement; relative visibility; volition). According to the proto-agency theory, we expect an increasing willingness of participants to judge Ball Y as the agent in the scenarios. Of course, given that Ball Y is always eventually launched by Ball X we did not expect a complete reversal of agency assignments.

## Method

Participants 39 students ( 27 women; mean age 23.4 years) from the University of Göttingen, Germany, participated in this experiment as part of a series of various unrelated com-
puter-based experiments in our computer lab. Participants received either course credit, or were paid $€ 8$ per hour.
Material For each condition, we constructed a flash movie of size $760 \times 760$ pixels that played for 3,000 milliseconds; the first and last 400 milliseconds presented a black screen resulting in an effective movie length of 2,200 milliseconds. Ball X and Ball Y were 120 pixels in diameter; one colored in red, the other in blue. In the standard condition (Condition A), Ball Y rests in the middle of the screen such that the left most point of $Y$ coincides with the center of the scene. After 20 milliseconds Ball X enters the scene from the left side on a horizontal trajectory with constant speed until it reaches the center of the screen (and, therefore, Ball Y) after 1,100 milliseconds. Then Ball X stops moving, and at the same time (no time lag) Ball Y starts moving with the same speed as Ball X towards the right hand side of the screen. Ball Y leaves the screen after 2,180 milliseconds. (Thus, the movie is symmetric in time and space.)

Keeping the movement shown in Condition A constant, we slightly altered the events prior to the launching event in the other conditions. We only manipulated the 800 milliseconds prior to the collision after which the movement pattern of the balls were identical across all conditions. In Conditions B and C, Ball Y entered the scene at the bottom after 20 milliseconds and moved vertically upwards until it reached its final position after 800 milliseconds. In Condition B, we covered 240 pixels of the scene's left hand side with a white panel; in Condition C, 240 pixels of the bottom were covered. Thus, in Condition B Ball X entered the scene after 700 milliseconds (whereas Ball Y was visibly moving the whole time); in Condition C Ball Y entered the scene after 840 milliseconds (whereas Ball X was visibly moving the whole time).

In Condition D, Ball Y was at rest in the lower half of the display (200 pixels above the bottom) and started moving upwards after 900 milliseconds (at the same speed Ball Y moves in Conditions B and C), and stopped at its final position after 800 milliseconds (i.e., movement time of 300 milliseconds). This sudden, apparently self-propelled movement was expected to suggest a volitional intervention into the trajectory of Ball X.

For counterbalancing purposes we additionally generated seven more movies per condition by rotating the scene by $90^{\circ}, 180^{\circ}$, and $270^{\circ}$, respectively, and switching colors of the balls yielding $4 \times 2=8$ movies per condition (i.e., in sum 32 movies).

Procedure We presented each subject with all 32 movies in random order. After seeing a movie (self-paced), we requested participants to select one of four sentences (presented in randomized order) as the best description of the scene:

1. The red ball launched the blue ball.
2. The blue ball stopped the red ball.
3. The blue ball launched the red ball.
4. The red ball stopped the blue ball.

Note that only two of the sentences actually described what was seen in the movie. If a subject selected one of the two nonsensical sentences we coded the answer as an error.

Design and Prediction We recoded subjects' responses according to the color coding as "X launched Y" vs. "Y stopped X " (plus error), and aggregated the eight color/rotation versions to align with a consistent $\mathrm{X} / \mathrm{Y}$ assignment. We expected an increasing selection rate for "Y stopped X " and a decreasing selection rate for " X launched Y", respectively, from Condition A to Condition D. ${ }^{1}$

## Results and Discussion

Fig. 1 shows the average selection rates for the two relevant scene descriptions across the four agency conditions. In line with previous research, Condition A revealed a strong preference in selecting Ball X as the agent ( $94.9 \%$ vs. $3.9 \%$ ). As predicted, selecting Ball X decreased from Condition A to Condition D, $F_{3,114}=24.0, p<.001, \eta^{2}=.39$. The preference for seeing Ball $Y$ as the agent increased analogously, $F_{3,114}=22.9, p<.001, \eta^{2}=.38$. The average error rate was 2.6\% and did not significantly differ across conditions, $F_{3,114}<1$.

Experiment 1 clearly demonstrates that agency intuitions are grounded in empirical indicators of agency, and confirmed the proposed proto-agency theory. However, it could be argued that the forced-choice format forced people to choose one description even when their intuition was in line with the symmetry assumptions of Newtonian mechanics. This argument does not explain why on average the choices did not even out, but we still wanted to replicate the results of Experiment 1 using a more unrestricted response format.

## Experiment 2

The goal of Experiment 2 was to replicate the findings of Experiment 1 with an unrestricted response format that allows subjects to express that they see both alternative sentences as valid descriptions of the scene. To accomplish this goal we presented subjects in Experiment 2 with rating scales that allowed them to judge the appropriateness of the scene descriptions independently.

## Method

Participants A new set of 34 students ( 23 women; mean age 23.4 years) from the University of Göttingen, Germany, participated in this experiment using the same design as in Experiment 1.

Material and Procedure We used the same set of 32 movies and the same procedure as in Experiment 1. Instead of a forced choice decision between scene descriptions, we presented subjects with the two sentences (adapted to the respective color version), and requested them to rate how well the sentences describe the scene using two separate rating

[^150]

Figure 1: Results of Experiment 1. Error bars indicate standard error of the means.
scales ranging from 0 ("not appropriate at all") to 10 ("highly appropriate"). Both sentences and rating scales were presented on a single screen; the order of the sentences was counterbalanced within subjects.

Design and Predictions We aggregated subject-wise across color/rotation conditions, which yielded a 4 (agency condition) x 2 (Ball X vs. Ball Y) within-subjects design with agency ratings as dependent measure. Since we expected decreasing ratings for Ball X and increasing ratings for Ball Y, we predicted an interaction between agency condition and the rated ball (X vs. Y).

## Results and Discussion

Fig. 2 shows the results for Experiment 2. As expected, the ratings for Ball X were higher as for Ball Y in Condition A with a decreasing trend for Ball X and an increasing trend for Ball Y from Condition A to Condition D. This pattern led to a significant interaction, $F_{3,99}=23.7, p<.001, \eta^{2}=.42$. Across conditions, Ball X received higher agency ratings than Ball Y, $F_{1,33}=34.3, p<.001, \eta^{2}=.51$, reflecting the fact


Figure 2: Results of Experiment 2. Error bars indicate standard error of the means.
that the salient end of the scene (Ball Y's leaving) overall dominates agency intuitions.

The results of Experiment 2 resemble the results of Experiment 1 closely. Although the difference between rating Ball X as the agent vs. rating Ball Y as the agent is much smaller in Condition A compared to Experiment 1, the overall pattern (decreasing ratings for Ball X and increasing ratings for Ball Y from Condition A to Condition D) was replicated, and showed that the findings were nor restricted to specific response formats.

## General Discussion

In contrast to dependency theories, force and dispositional theories of causal reasoning incorporate the distinction between agents and patients in causal interactions. The principal focus of research motivated by dispositional accounts was on how force configurations predict causal judgments, whereas the assignment of the agent and patient roles has largely been treated as self-evident. Various studies in both linguistics and psychology have shown, however, that role assignments are not always clear-cut. Occasionally it may even be necessary to assign the agent role to multiple causal participants.

Based on the linguistic theory of Dowty (1991), we proposed that agency is a prototype concept with multiple criteria, none of which necessary for the role assignment. We adapted this theory to account for physical interactions (e.g., Michotte type launching events). In the standard Michotte launching scenario the ball entering the scene (and launching the other ball) is typically assigned the agent role, whereas the ball that is being launched is viewed as the patient. We showed that agency intuitions are moderated by manipulations of the context prior to the launching event. Altering scene features, such as relative movement, sequence of visibility, and self-propelled motion tended to increase agency attributions to the patient relative to the standard scenario.

A unifying principle underlying these criteria may be that they all tend to lift the patient into the foreground (i.e., into the figure role), and appear to suggest some kind of volitional intervention. Intervention seems to be a central concept unifying dependency and force theories, although the criteria for determining agency are different in these two frameworks. We realize that our experiments just represent a first step. Future studies will have to go beyond launching scenarios to arrive at a more complete theory of agent/patient assignments.

## Acknowledgments

We wish to thank Sven Ritter for help in preparing the experiments. This research was supported by a research grant of the Deutsche Forschungsgemeinschaft (WA 621/22-1) that is part of the priority program "New frameworks of rationality" (SPP 1516).

## References

Beller, S., Bender, A., \& Song, J. (2009). Weighing up physical causes: Effects of culture, linguistic cues and content. Journal of Cognition and Culture, 9, 347-365.
Cheng, P. W., \& Novick, L. R. (1991). Causes versus enabling conditions. Cognition, 40, 83-120.
Dowty, D. (1991). Thematic proto roles and argument selection, Language, 67, 547-619.
Gnassounou, B. \& Kistler, M.(2007). Introduction. In M. Kistler \& B. Gnassounou (Eds.)(2007). Dispositions and causal powers. Aldershot: Ashgate.
Mayrhofer, R., Hagmayer, Y., \& Waldmann, M. R. (2010). Agents and causes: A Bayesian error attribution model of causal reasoning. In Proceedings of the Thirty-Second Annual Conference of the Cognitive Science Society (pp. 925-930).
Mayrhofer, R., \& Waldmann, M. R. (2013). Agents and causes: Dispositional intuitions as a guide to causal structure. Manuscript submitted for publication.
Michotte, A. E. (1963). The perception of causality. New York: Basic Books.
Muentener, P., \& Carey, S. (2010). Infants' causal representations of state change events. Cognitive Psychology, 61, 63-86.
Mumford, S., \& Anjum, R. L. (2011). Getting causes from powers. New York: Oxford University Press.
Pinker, S. (2007). The stuff of thought: Language as a window into human nature. New York: Viking.
Riemer, N. (2010). Introducing semantics. Cambridge: Cambridge University Press.
Steyvers, M., Tenenbaum, J. B., Wagenmakers, E-J., \& Blum, B. (2003). Inferring causal networks from observations and interventions. Cognitive Science, 27, 453-489.
Talmy, L. (1988). Force dynamics in language and cognition. Cognitive Science, 12, 49-100.
Waldmann, M. R., \& Hagmayer, Y. (in press). Causal reasoning. In D. Reisberg (Ed.), Oxford Handbook of Cognitive Psychology. New York: Oxford University Press.
White, P. A. (2006). The causal asymmetry. Psychological Review, 113, 132-147.
White, P. A. (2009). Perception of forces exerted by objects in collision events. Psychological Review, 116, 580-601.
White, P. A. (2012). Visual impressions of pushing and pulling: The object perceived as causal is not always the one that moves first. Perception, 41, 1193-1217.
Wolff, P. (2007). Representing causation. Journal of Experimental Psychology: General, 136, 82-111.
Wolff, P., \& Song, G. (2003). Models of causation and the semantics of causal verbs. Cognitive Psychology, 47, 276332.

Wolff, P., Barbey, A. K., \& Hausknecht, M. (2010). For want of a nail: How absences cause events. Journal of Experimental Psychology: General, 139, 191-221.
Woodward, J. (2003). Making things happen: A theory of causal explanation. Oxford: Oxford University Press.

# Reasoning with Inconsistent Causal Beliefs 

John V. McDonnell (john.mcdonnell@nyu.edu) ${ }^{1}$<br>Pedro Tsividis (tsividis@mit.edu) ${ }^{2}$ Bob Rehder (bob.rehder@nyu.edu) ${ }^{1}$


#### Abstract

Causal reasoning is a critical part of everyday cognition. We ask how people reason about causes when faced with inconsistent sources of knowledge. Causal models arise from multiple sources of information regarding their constituent parameters. Knowledge sources may be inconsistent both within parameters (when one source says a variable should appear often and another says it should appear rarely), and between parameters (when dependencies among parameters result in an internally inconsistent causal model). We provide a normative model for resolving both these sources of conflict. An experiment found that our model of belief integration predicted the qualitative pattern of adults causal inferences under uncertainty.


Keywords: Causal Learning; Causal Inference; Probabilistic Modeling

## Introduction

From deciding on investment strategies to predicting others' reactions in social situations, we are constantly making probabilistic judgments about causal systems. However, because we receive information from multiple sources, we are often faced with contradictory beliefs. Consider the problem faced by an epidemiologist trying to understand the causes of chronic hypertension in a particular population. She reads a review paper suggesting that smoking tobacco causes hypertension in $50 \%$ of patients, and that all other potential causes of hypertension can be ruled out. The epidemiologist knows from survey data that $25 \%$ of the population of interest are smokers, and (independently) that $25 \%$ have hypertension. If she assumes her maximum likelihood estimates are true, she is left with an incoherent causal model: If smoking is the only cause of hypertension, and is effective half the time, then there should be half as many people with hypertension as there are smokers. Arriving at coherent causal beliefs will require adjustment. Perhaps hypertension isn't really as prevalent as she thought, or perhaps the smoking always causes hypertension. In this paper, we propose a normative probabilistic model for reasoning with these inconsistencies and explore the implications of that model in an experiment in which participants receive conflicting sources of evidence.

## A Model

We assume that people's causal inferences are based on causal graphical models (CGMs), such as the one in Figure 1 in which $C 1$ and $C 2$ are believed to be independent causes of $E$. For simplicity, throughout the paper

[^151]we will assume that all variables are binary (present or absent) and that all causal links are generative, bringing about their effects through a noisy-or functional form. We allow for the possibility of additional causes of $E$ not shown in the graph by aggregating them into a single background cause that is always present (for a review of graphical models, see Koller \& Friedman, 2009). A fully parameterized CGM is sufficient to answer virtually any question one might want to ask about the variables involved, including questions of conditional or joint probability, counterfactual reasoning, and predicting the effect of interventions (Pearl, 2000). However, complications arise when one recognizes that CGMs are constructed from many individual beliefs held by the reasoner. Since these beliefs may come from multiple sources that vary in their reliability, it is inevitable that they will sometimes contradict one another. We ask: How should one draw causal inferences in light of such inconsistencies?

To answer this question, we first note that inconsistencies can be either between or across parameters, where parameters represent one's beliefs about each constituent of the model. For example, a belief about the probability of a cause being present is one constituent; a belief about the strength of a causal relation is another. In the first section we advance a new proposal for representing uncertainty in CGMs and show how it solves the problem of within-parameter conflicts. We then tackle the more challenging problem of between-parameter conflicts.

## Resolving conflicts within parameters

Consider the problem of representing the base rate of variable $C 1$, represented by parameter $c_{1}$ in Figure 1. We suppose that beliefs about base rates may come from first-hand observations (observing the prevalence of $C 1$ ), explicit, instruction (e.g., hearing that $C 1$ is uncommon), and prior beliefs (e.g., a tendency to believe that events of this type arise rarely). Because probability is bounded to the range $[0,1]$, the information from each of these sources can be encoded as a probability density function in that range. If knowledge is represented as a point estimate of the expected value of the variable combined with a confidence in that point estimate, this information can be encoded as a beta distribution. PDFs of beta distributions representing knowledge sources for $C 1$ are shown at the top of Figure 1. The shape of a beta distribution is controlled by two parameters, $\alpha$ and $\beta$, constrained to be positive (we will refer to such parameters as knowledge parameters, or $k$-parameters, to emphasize that they represent participants' knowledge and to avoid confusion with the constituent parameters).


Figure 1: A simple common-effect causal graphical model. Here $C 1$ and $C 2$ are potential causes of $E$. Variables of boxes indicate the parameters of the model. Our conflict-resolution model assumes participants encode their beliefs about each of these variables as beta distributions, depicted alongside each of the variables. Above, knowledge sources for $c_{1}$ and $m_{2}$ are depicted, also represented as beta distributions.

The expected value of the odds is the ratio of the two k -parameters, while their sum can be interpreted as the confidence in that estimate.

Reconciling different beliefs encoded as beta distributions is simple. Bayes' rule gives the posterior belief as the renormalized product of the prior and likelihood distributions. In the case of beta distributions, this is derived by summing the parameters. This means that for the knowledge sources in Figure 1, if we denote the k -parameters $\alpha_{\text {prior }}, \beta_{\text {prior }}, \alpha_{\text {instruct }}$, $\beta_{\text {instruct }}, \alpha_{\text {exp }}$, and $\beta_{\text {exp }}$, the posterior is simply $\operatorname{Beta}\left(\alpha_{\text {prior }}+\alpha_{\text {instruct }}+\alpha_{\text {prior }}, \beta_{\text {prior }}+\beta_{\text {instruct }}+\beta_{\text {prior }}\right)$. This process is depicted in Figure 1 for $c_{1}$.

A similar treatment can be applied to the strengths of the causal relations in Figure 1, shown in the figure for the case of $m_{2}$. Following Cheng (1997), we assume that each link is represented as a causal power: the propensity of the cause, when present, to bring about the effect. Because they are probabilities, beliefs about causal powers can also be stored as beta distributions.

As depicted in Figure 1, the model consists of six pa-
rameters: the base rates of $C 1$ and $C 2\left(c_{1}\right.$ and $\left.c_{2}\right)$, the strengths of causal relationships $C 1 \rightarrow E$ and $C 2 \rightarrow E$ ( $m_{1}$ and $m_{2}$ ), the strength of the background causes of $E(b)$, and the rate at which $E$ occurs (e). We suppose that belief in the value of each of them is represented as a posterior beta distribution.

## Resolving conflicts between parameters

Computing the posterior beta distribution for each model parameter does not eliminate all potential inconsistencies, however. As illustrated in the introduction, when all the causes of an effect are fully described, the effect variable's rate of occurrence is no longer free to vary. Because the value of $e$ is fixed, random draws from the individual beta posteriors will never return valid causal models (assuming infinite precision). This means we need a way to integrate information about effects into our beliefs over possible causal models without violating the constraints of the model.

Following Figure 1, let $\mathbf{V}$ represent the set of variables in the domain. For each $v \in \mathbf{V}$, belief in the probability of $v$ is characterized by the PDF of a beta distribution, denoted $\pi_{v}$. These correspond to $c_{1}, c_{2}$, and $e$ in the figure. Next, let $\mathbf{L}$ be the set of causal links in a model. For each $l \in \mathbf{L}$, the learner's belief in the causal power of $l$ is characterized by a beta-distributed PDF denoted $\pi_{l}$ ( $m_{1}$ and $m_{2}$ in the figure). Finally, let $\mathbf{E} \subset \mathbf{V}$ be the effects. For each $e \in \mathbf{E}$, the belief in the background causes of $e$ is characterized by a beta-distributed PDF denoted $\pi_{e}$ ( $b$ in the figure).

We now define the posterior over valid causal models. Let $\mathbf{r}, \mathbf{m}$, and $\mathbf{b}$ be vectors describing the base rate of every variable in $\mathbf{V}$, the strength of every link in $\mathbf{L}$, and the strength of the background causes for every effect in $\mathbf{E}$, respectively. Under a noisy-or causal model, all effects are explained by the likelihood of their causes and the strength of those causes. This means the rate of occurrence for an effect $e$ is constrained to be

$$
\begin{equation*}
r_{e}^{\prime}=1-\left(1-b_{e}\right) \prod_{(l \in \mathbf{L}) \wedge\left(l_{e}=e\right)}\left[1-r_{l_{c}} m_{l}\right] \tag{1}
\end{equation*}
$$

where $l_{c}$ and $l_{e}$ are the cause and effect variables associated with causal link $l$. Enforcing this consistency, we can define the joint probability of a fully parameterized model as
$P(\mathbf{r}, \mathbf{m}, \mathbf{b}) \propto\left\{\begin{array}{l}0 \quad \text { if } \exists e \in \mathbf{E}: r_{e} \neq r_{e}^{\prime} \\ \prod_{v \in \mathbf{V}} \pi_{v}\left(r_{v}\right) \prod_{l \in \mathbf{L}} \pi_{l}\left(m_{l}\right) \prod_{e \in \mathbf{E}} \pi_{e}\left(b_{e}\right) \text { otherwise. }\end{array}\right.$
This is equivalent to saying that the posterior over joint model values is defined as the result of sampling from the beta distributions characterizing each of the variables in the model, discarding the inconsistent models. Our central hypothesis is that, when inconsistencies among be-

|  | Rare Causes (25\%) | Common Causes (75\%) |
| :---: | :---: | :---: |
| One Cause (50\%) | Severely Underexplained <br> Implied base rate of effect: 12.5\% | Mildly Underexplained <br> Implied base rate of effect: 38\% |
| Three Causes (50\%) | Mildly Underexplained <br> Implied base rate of effect: 33\% | Severely Overexplained <br> Implied base rate of effect: 84\% |

Figure 2: Design of the experiment.
liefs exist, people draw inferences as if they are reasoning with the maximum a posteriori causal model.

## An Experiment

One important prediction of our model is that when faced with an invalid model, reasoners will make tradeoffs among the parameters to find a valid causal model (Equation 2 above). We tested this hypothesis by querying the adjustments individuals make to their instructed causal model given a variety of inconsistent beliefs.

Each participant learned about four binary variables in one of three domains: economics, meteorology, or sociology. Participants learned about four relevant variables and the beliefs of experts in their domain. In the domain of economics, for example, the four variables were interest rates (moderate/high), trade deficits (moderate/large), retirement savings (moderate/high), and job mobility (moderate/high). Depending on the condition, one or three of those variables (denoted $C_{1}, C_{2}$, and $C_{3}$ ) were been described as generative causes of a fourth $(E)$.

A $2 \times 2$ design (depicted in Figure 2) explored the effect of varying the number of causal links (one vs. three) and the base rates of those causes ("rare" vs. "common"). In all conditions, the effect was "somewhat common," (occurring $44 \%$ of the time), and the causes brought about the effect half the time.

Manipulation of these two parameters should of course result in changes to participants' estimates of the values of those parameters. However, because they also imply between-parameter conflicts, they should also result in compensating changes to other parameters in order to yield a consistent model. For example, consider the condition in which only one rare cause is instructed. Here the participants were told about a cause happening 4 times out of 16 and bringing about its effect $50 \%$ of the time. As indicated in the figure, these facts imply a base rate for the effect of .125 , which conflicts with its instructed base rate of .44. Thus, the effect is underdetermined by the causes in the model. There are many ways in which reasoners could compensate: They could
adjust the likelihood of the effect downward, increase the likelihood of the cause, increase the causal strength of the cause, or increase the likelihood of background causes. Participants must choose a combination of these adjustments to reason with a valid causal model.

Conversely, in the case where participants were instructed about three causes, each common, the implied base rate of the effect is .84 , that is, the effect is now over-determined, and the reverse adjustments are needed to help form a valid model. Note that because we did not explicitly control the confidence subjects should hold in their beliefs about individual parameters, we do not make predictions regarding which variables will be adjusted, but rather only that some subset will be adjusted to attain a consistent model.
To measure the adjustments made by participants, we followed this instruction with a series of questions designed to assess their beliefs about the parameters of the causal model.

## Method

Participants A total of 234 subjects were tested, consisting of 114 New York University undergraduates who received course credit and 120 online subjects who received a small cash incentive. Subjects were randomly assigned to the 1link/rare, 1 -link/common, 3 -link/rare, and 3 -link/common, conditions. Participants whose numerically-coded responses over the course of the experiment had a standard deviation of less than 2 were excluded, leaving $54,52,57$, and 47 participants in each of the conditions, respectively.
Materials Three knowledge domains were tested: Economics, meteorology, or sociology. In each domain, the same four variables were used, so the same variables always played the role of $C_{1}, C_{2}, C_{3}$, and $E$. Subjects in the 3 -link conditions learned three causal links: $C_{1} \rightarrow E, C_{2} \rightarrow E$, and $C_{3} \rightarrow E$. Those in the 1-link conditions learned only $C_{1} \rightarrow E$.
Base rate information was displayed as an instruction screen displaying a pictorial representation of the rates at which each of the variables was observed. Base rates of the causes were described as occurring $25 \%$ of the time ("rare"), or $75 \%$ ("common"). $E$ was always depicted as occurring $44 \%$ of the time ("somewhat common"). The base rates were illustrated using a pictorial graph showing them the values of each variable for 16 random systems from a "survey". For example, $\frac{7}{16}$ of systems were always shown to have the effect.

Causal information was conveyed in writing. For each causal relationship, participants were told that the cause brought about the effect $50 \%$ of the time. They were also given a short description of the mechanism underlying the relation. For example, if told that large trade deficits cause high job mobility, they were also told, "The flood of cheap imports means that many domestic manufacturing jobs are lost and workers must find new employment." Participants were also told that experts believed there were no other causes of the effect.
Procedure After being introduced to the domain, participants were presented with screens presenting experts' beliefs about the base rates of the variables and their causal relationships. After the instructions, online subjects were quizzed on their memory for the instructions. This repeated in a loop until they were able to correctly answer all questions.

Participants were asked four types of test questions presented in blocks. Block order was randomized. For all question types, participants responded with an estimated probability on a scale of $1-11$ by choosing one of 11 radio buttons, with the two extremes labeled "very unlikely" (on the


Figure 3: Causal model parameters derived from the empirical results from the experiment compared across all conditions (without modeling learning or prior belief). (A) The base rates of the causes (parameter $c_{1}$ ). (B) The strength of the causal links (parameter $m$ ). (C) The base rate of the effect (parameter $e$ ). (D) The strength of alternative causes (parameter b). Error bars are standard errors.
left) to "very likely" (on the right). The first questions were joint probability questions. In these questions, subjects were given the states of all four variables in their scenario and asked to rate the likelihood of those variable states being observed. All 16 possible questions formed by varying the state of the three binary variables were asked. A second question type consisted of conditional probability questions. For these questions, subjects were given the state of three of the four variables and asked to rate the likelihood of the fourth. All 8 possible conditional probability judgments formed asking for the probability of $E$ as a function of the presence or absence of the three causes were asked. In addition, eight questions asked for the probability of one of the causes with the effect either present or absent (with the other causes always stipulated to be absent). Finally, the third and fourth question types directly queried the base rates of the four variables and the strengths of the three potential causal links. To avoid the possibility that subjects would forget the initial information about the causal model, a "theory reminder" was presented on the right side of the screen, accompanying each question.

Finally, subjects were given the chance to learn more about their causal system by observing a sample of 32 instances from that domain. The sample was drawn from a model in which the causes had a base rate of .50 , two causal links $\left(C_{1} \rightarrow E\right.$ and $\left.C_{2} \rightarrow E\right)$ of strength .50 , with no alternative causes of $E$. Subjects were asked to "consider how these data might change your beliefs about the causal relationships in this system and the overall likelihoods of the variables involved." Participants were then asked to re-answer all the previous questions. Then they cycled through once more, again observing a sample of 32 instances, and again reanswering the questions. These responses were used in model fitting, but because learning effects appeared to be small the results of these test phases will not be reported here.

## Results

To characterize subjects' judgments of joint and conditional probability, we fit those judgments to a causal model with three causes and one effect, yielding eight parameters: $c_{1}, c_{2}$, and $c_{3}$ (the likelihoods of the causes),
$m_{1}, m_{2}$, and $m_{3}$ (the strengths of the putative causal links), $b$ (the background cause of $e$ ) and $e$, the likelihood of the effect (see the Appendix for details).

Figure 3 summarizes the effects of our two manipulations on the causal model parameters. For purposes of comparison, we only present those parameters that were involved in a causal relationship in all conditions $\left(c_{1}, m_{1}, e\right.$, and $\left.b\right)$. A $2 \times 2$ ANOVA with the causes' base rate and the number of causal links as factors was performed for each panel. A main effect of the base rate manipulation on estimates of base rates of $C_{1}$ (plotted in Figure 3A) confirmed the success of that manipulation $(F(1,206)=35.45, M S E=.034, p<.001)$. Importantly, the manipulation also resulted in an increase in the prevalence of the effect (parameter $e$ in Figure 3C, $F(1,206)=19.00, M S E=.074, p<.001)$ and a decrease in the strength of the background causes (parameter $b$ in Figure 3D $(F(1,206)=10.23, M S E=.058$, $p<.01$ ). That is, to accommodate the more prevalent causes, participants compensated by increasing the base rate of the effect and decreasing the effectiveness of alternative causes, as predicted by our model. There was no effect of the manipulation on causal strengths (parameter $m_{1}$ in Figure 3B).

The manipulation of the number of causal links also had two important effects. First, it reduced the base rate of $C_{1}$ (parameter $c_{1}$ in Figure $3 \mathrm{~A}, F(1,206)=35.45$, $M S E=.073, p<.001)$. Second, it reduced the strength of the $C_{1} \rightarrow E$ causal relationship (parameter $m_{1}$ in Figure $3 \mathrm{~B}, F(1,206)=20.67, M S E=.088, p<.001)$. Apparently, to accommodate two additional causal links, participants compensated by decreasing the effectiveness of the $C_{1} \rightarrow E$ link, reducing both $c_{1}$ and $m_{1}$ as predicted by our model. Changing the number of causal links did not have a significant effect on either the prevalence of the effect (parameter $e$ in Figure 3C, $F<1$ ) or the strength of alternative causes (Figure 3D, $F(1,206)=2.16, M S E=.058, p=.19)$. None of the 2-way interactions approached significance (all $F$ s $<1$ ).

## Discussion

The results of this experiment supported the claim that when people are given inconsistent information, they draw inferences as if they're reasoning with the most likely causal model. Increasing the base rates resulted in participants believing that the effect is more likely and alternative causes are weaker. Increasing the number of causes led participants to adjust their beliefs about the causes, weakening both their efficacy and their base rates. We now assess whether our theoretical model can provide not only a good qualitative account of these data, but an acceptable quantitative one as well.

## Theoretical Modeling

Recall that although our theoretical model of uncertainty and belief integration specifies that reasoners will adjust


Figure 4: Causal parameters derived from the fit of the theoretical model. (A) The base rates of the causes (parameter $c_{1}$ ). (B) The strength of the causal links (parameter $m$ ). (C) The base rate of the effect (parameter $e$ ). (D) The strength of alternative causes (parameter b).
parameters in order to reason with a consistent causal model, it does not specify which parameters will be adjusted in the absence of any information about the confidence with which beliefs about those parameters are held. To assess our model's potential for providing a quantitative account of the experiment, we fit it to those results treating confidence in each constituent parameter as a free parameter (a fifth control condition, not reported here, was also fit in which subjects were instructed on the same model that generated the learning data; see Procedure section). For any constituent parameter of the model $k$ (a variable or causal link), confidence is represented as a beta distribution with k -parameters $\alpha_{k}$ and $\beta_{k}$. We recast those k-parameters into the pair $v_{k}=\frac{\alpha_{k}}{\alpha_{k}+\beta_{k}}$, representing the expected value of the parameter, and $t_{k}=\alpha_{k}+\beta_{k}$, representing the overall confidence with which the expected value is held. We assume that reasoners may have different levels of confidence in the different types of parameters, represented as four free k-parameters: $t_{c}$ (the base rates of the cause), $t_{m}$ (the strength of the causal links), $t_{b}$ (the strength of the alternative causes), and $t_{e}$ (the base rate of the effect). Because we assume that subjects do not perfectly encode the initial numerical information provided about each parameter, those are free parameters as well. Let k-parameters $v_{m}, v_{b}$, and $v_{e}$ represent the instructed values of $m$ (causal link strengths, described to subjects as .5), $b$ (the strength of the alternative causes, described as 0 ), and $e$ (the base rate of the effect, described as .44), respectively. $v_{m 0}$ was the initial strength of the links on which subjects were not instructed. k-parameters $v_{c-r}$, $v_{c-m} v_{c-c}$ represent the initial base rates of the causes in the rare, moderate, and common conditions, respectively
(described as .25, .5, and .75).
This model was fit to the group level causal model parameters fit in the experiment. Eight parameters $\left(c_{1}, c_{2}\right.$, $c_{3}, m_{1}, m_{2}, m_{3}, b$, and $e$ ) were estimated per phase per condition. This included the learning phases as well as a fifth control condition not reported here. This involved fitting $8 \times 3 \times 5=120$ data points with 10 parameters. The parameters that minimized squared error were $t_{c}=106, t_{b}=299, t_{e}=3252, v_{c-r}=.162, v_{c-m}=.351$, $v_{c-c}=.485, v_{m}=.454, v_{m 0}=.069, v_{b}=.122$, and $v_{e}=.7$. The correlation between observed and predicted values was .964. The predictions are depicted in Figure 4, which is analogous to Figure 3. Figure 4 reveals that the model is able to capture the effects of the causal strength manipulation, namely the base rate of $C_{1}$ (parameter $c_{1}$, Figure 4 A ) and the strength of the $C_{1} \rightarrow E$ link (parameter $m_{1}$, Figure 4 B ) both decrease as the number of causal relations increases (compare with Figure 3A and B, respectively). It is also able to reproduce the effects of the base rate manipulation on the strength of the alternative causes (parameter $b$, Figure 4D), namely, that alternative causal strength decreases as the base rates of the causes increase (compare with Figure 3D). Less successfully, it predicts an increase in the base rate of the effect with larger base rates of the causes (parameter $e$, Figure 4C), although the magnitude of that change is much smaller than the one exhibited by subjects (compare with Figure 3C). Note the insensitivity of parameter $e$ to changes in the other causal model parameters is a manifestation of the large confidence the model places on its initial value $\left(t_{e}=3252\right.$, vs. all other $t \mathrm{~s}<300$ ).

## General Discussion

In ecologically valid settings, causal reasoning often takes places with multiple knowledge sources that are potentially inconsistent with one another. To specify how causal inferences should be drawn in such situations, we developed an account of how uncertainty about causal models might be represented and then showed how to derive the most likely causal model that is sensitive to each knowledge source yet resolves inconsistencies between them. Our central hypothesis was that people would draw causal inferences as if they were reasoning with the most likely consistent model.

The qualitative predictions of this model were confirmed in an experiment manipulating two instructed parameters: the base rates of the causes and the number of causal links. Making causes more prevalent resulted in alternative causes becoming weaker and the effect becoming more prevalent. Making causal relations more numerous resulted in the causes becoming rarer and other causal links becoming weaker. We know of no other model that is capable of predicting these sorts of effects.

Our model also yielded moderately good quantitative fits to the data. One result it was unable to reproduce was participants' tendency to adjust the likelihood of the effect ( $e$ ) when the causal base rates were adjusted but not when the number of causes was adjusted. Moreover, we acknowledge that these fits used a large number of parameters, necessitated by the fact that confidence in each instructed model parameter was not specified experimentally and so needed to be free parameters. We are conducting follow-up studies manipulating instructed confidence in the information provided to participants.

Although our representation of uncertainty was sufficient to account for our empirical results, its assumption that the distributions of the causal model parameters are independent is unrealistic in some situations. For example, Lu et al. (2008) have modeled the traditional causal learning experiment as one in which the prior distribution is a two-dimensional density function on the strength of the to-be-learned causal link and the strength of alternative causes. Multivariate representations of uncertainty like this may be common. In addition, one might imagine that reasoners not only have experiential knowledge about the base rates of variables (Figure 1), but also about configurations of variables. Finally, in addition to changing parameters to attain consistency, reasoners might also change the function relating an effect to its causes (e.g., by assuming that the causes combine interactively rather than independently) or even the structure of the model itself (e.g., deleting a causal link, as proposed by Griffiths \& Tenenbaum, 2005).

One facet of the data not discussed above was the distinction between our explicit queries (e.g., directly querying the causal efficacy of $C_{1}$ ) and our implicit ones, such as judgments of conditional and joint probability. Perhaps unsurprisingly, explicit queries more closely resembled the likelihood information participants were given, unadjusted for consistency, suggesting that such questions do not invoke inconsistency resolution processes we have specified here. We also observed (but did not report here) that participants' causal models changed little as a result of observing data. Whether this result reflected a kind of anchoring effect (initial judgments influenced later ones) or participants large confidence in the initial domain theories on which they were initially instructed remain questions for future research.

## Acknowledgments

This work was supported by the Air Force Office of Scientific Research, Grant No. FA9550-09-NL404.

## References

Cheng, P. (1997). From covariation to causation: a causal power theory. Psych. Rev. 104, 367.
Griffiths, T. L. \& Tenenbaum, J. B. (2005). Structure and strength in causal induction. Cog. Psych. 51, 334-384.

Koller, D. \& Friedman, N. (2009). Probabilistic graphical models: principles and techniques. Cambridge, MA: The MIT Press.
Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., \& Holyoak, K. J. (2008). Bayesian generic priors for causal learning. Psych. Rev. 115, 955-984.
Pearl, J. (2000). Causality: models, reasoning, and inference. Cambridge, UK: University Press.

## Appendix

Participants' causal models were simultaneously fit to participants' judgments of conditional and joint probability. Fitting assumed that participants formed a common-effects model with $C_{1}, C_{2}$, and $C_{3}$ as potential causes of $E$. The joint is defined as the probability that the four variables will take any particular combination of values. From the axioms of probability, we derive

$$
\begin{equation*}
p\left(E, C_{1}, C_{2}, C_{3}\right)=p\left(E \mid C_{1}, C_{2}, C_{3}\right) p\left(C_{1}, C_{2}, C_{3}\right) \tag{3}
\end{equation*}
$$

Because $C_{1}, C_{2}$, and $C_{3}$ are assumed to be independent,

$$
\begin{equation*}
p\left(E, C_{1}, C_{2}, C_{3}\right)=p\left(E \mid C_{1}, C_{2}, C_{3}\right) p\left(C_{1}\right) p\left(C_{2}\right) P\left(C_{3}\right) \tag{4}
\end{equation*}
$$

Assuming that the causes bring about their effects according to a noisy-or rule, the probability that $E$ is present given the status of the causes is given by

$$
\begin{equation*}
p\left(E=1 \mid C_{1}, C_{2}, C_{3}\right)=1-(1-b) \prod_{i \in\{1,2,3\}}\left(1-m_{i} C_{i}\right) \tag{5}
\end{equation*}
$$

where presence or absence is coded as 1 or 0 , respectively.
Equations 4 and 5 are sufficient to specify the probability of any combination of the variables as a function of the parameters $c_{1}, c_{2}, c_{3}, m_{1}, m_{2}, m_{3}$, and $b$.

Separate $c, m$, and $b$ parameters were estimated for each participant for each test phase. To transform responses on the 1-11 scale into probabilities, we applied a nonlinear (power) transformation. This necessitated fitting a power parameter, $\gamma$. Each subject's rankings were predicted as follows:

$$
\begin{align*}
& \text { rating }_{\text {cond }}\left(r_{b, i}\right)=10 p_{k}\left(r_{i} ; c_{b}, m_{b}, B_{b}\right)^{\gamma_{\text {cond }}}+1  \tag{6}\\
& \operatorname{rating}_{\text {joint }}\left(o_{b, i}\right)=10 p_{k}\left(r_{i} ; c_{b}, m_{b}, B_{b}\right)^{\gamma_{\text {joint }}}+1 \tag{7}
\end{align*}
$$

where $r_{b, i}$ and $o_{b, i}$ are the subject's conditional and joint judgments, respectively, on trial $i$ in phase $b . \gamma \mathrm{s}$ were fit within participant and question type, and constrained to the range $[0,5]$. This resulted in $7 \times 3=21$ causal model parameters and the two $\gamma$ parameters (23 in total) used to fit $32 \times 3=96$ responses. Parameters were fit using gradient descent.

The model fits included a control condition in which participants were instructed with a causal model that conformed to the data they observed. The results are not elaborated in this paper due to space constraints.

This model achieved a respectable correlation between its predictions and the empirical data points (.950; . 783 averaged over subjects).

# Naïve Physics - the wrong theory? 

I.P.L. McLaren (i.p.1.mclaren@exeter.ac.uk), K.Wood, \& R.P. McLaren<br>School of Psychology, University of Exeter, UK.


#### Abstract

In this paper we examine the idea of a "naïve physics" in humans solving physics problems. This invokes the idea that people have a theory of motion in their heads that is nonNewtonian, and hence leads to systematic errors on these problems. We are able to show that, by selecting our problems carefully, it is possible to obtain answers that are consistent with this naïve physics and inconsistent with it; suggesting that it is not used to solve these problems but sometimes offered as post-hoc justification for the answers given. We offer evidence that the answers given owe more to past experience than any theory, and that a theory that postulates extrapolation on the basis of associative memory can give a good account of our results.


Keywords: Associative, Memory, Naïve Physics, Theory.

## Introduction

McCloskey, Caramazza and Green (1980) and McCloskey, Washburn and Felch (1983) have proposed that people consistently make particular mistakes when asked to predict the path of an object, given certain initial conditions, because they are applying the wrong theory, an intuitive mechanics or "naïve physics", to this type of problem. Their evidence is that, when given a relatively simple physical situation and asked to extrapolate on the basis of the information supplied, people tend to make certain types of error rather than others, and justify this with verbal reports that indicate a non-Newtonian approach to the problem (even though the instructions encourage that type of approach). One example of such a problem would be for participants to be asked to imagine looking down on the curved tube (which is held horizontally) in Figure 1, while a ball bearing is inserted with some speed, v, as shown. Their task is to draw the path the ball bearing takes on exiting the tube, ignoring such factors as friction between ball bearing and tube, and any wind resistance. The plan view is intended to take gravity out of the picture for the purposes of this problem, and the correct, Newtonian solution, is to draw a straight line as shown in the figure (solid line) as the ball bearing leaves the tube. Instead, many participants draw something approximating the curved dotted line as their answer, and justify this by claiming that the ball bearing has acquired "curvy impetus" as a result of its journey through the tube and this continues to cause its path to curve on exiting the tube. A more sophisticated version of this account will claim that this impetus dissipates with time, and so the curved path will gradually straighten as the ball bearing gets further from the tube (see McCloskey et al, 1980).


Figure 1: The ball bearing and curved tube problem.
Another classic problem studied by McCloskey and colleagues concerns what they call the "straight down belief" (McCloskey, Washburn \& Felch 1983). The idea here is that people tend to predict a straight down trajectory for objects that are dropped whilst being carried, whereas they predict a parabolic path for objects that fall when moving independently. A classic example of this is a cannonball projected off a cliff. If fired horizontally from a cannon with initial velocity, $v$, an "out and down" approximation to the parabolic path is a typical response from participants asked to draw its subsequent path. But if carried (by some overhead conveyor belt arrangement) to the edge of the cliff with velocity, v , and then released, even though this is the identical problem in physical terms, participants are much more likely to describe the cannonball's motion as straight down. The reason they give for this is that, in the first case, the cannonball possesses its own impetus when it leaves the gun. This horizontal impetus takes it out past the cliff edge, but starts to dissipate. At the same time, gravity takes hold, and accelerates it downwards - hence the "out and down" parabolic trajectory. But, in the case of the cannonball being carried by a conveyor belt, participants think it has no impetus of its own, and so, when released, gravity takes hold immediately and it drops straight down.

The research reported in McCloskey et al's paper investigates the basis for this finding. On the one hand, it may be just as it seems and as participants in these experiments claim, i.e. that they have applied a naïve physics (or natural intuition) to the problem which is, in some sense the "wrong" theory as it is non-Newtonian, and it is this that leads to the consistent error in predicting the path of the object. On the other hand, it may be that the
account given by participants of why they drew that particular path is more by way of a post-hoc rationalization of what they did, rather than an account of what caused them to do it. Instead, the causal factor in producing this consistent error could be extrapolation on the basis of experience, by which we mean to imply associativelymediated retrieval of memories based on some surface similarity to the problem just posed. On this account, the reason why a carried object is often portrayed as taking a path straight downwards when released is because that is the perceived experience that we have, and call upon, of this situation in real life. Imagine you are cycling along and you drop a package. To a first approximation at least, the package appears to fall straight down. This is because both the package and the observer on the bicycle share a (moving) frame of reference, and in that frame of reference there is no horizontal motion with respect to one another. But a thrown package, which, in naïve physics terms, is one moving independently, does not have this property and so will follow a curved trajectory.

We realize that proponents of the naïve physics view will argue that it is exactly episodes of this kind that lead to induction of a naïve physics i.e. the incorrect theory is derived from these types of experiences, and it is only by running carefully controlled experiments that allow for friction, wind resistance, and frames of reference that the proper Newtonian theory can be arrived at. But we would differ from this view in arguing that the effect of experience is primary, and that its impact via retrieval from memory is what drives the response, not its impact via some naïve physics induced on the basis of these experiences. This stance makes the prediction that, if we are able to find scenarios where experience would predict that a carried object would follow a curved trajectory, or an object moving independently should take a path straight down, then the result of putting these problems to participants should be quite different to that predicted on the naïve physics account. If participants have a theory that drives their responses, then it should apply across different situations, as long as the particular scenario employed does not change the essential physics of the problem. Equally, if the outcome of experimental investigation of this proposition were to be that the responses made to a problem involving a carried object were to predict a path straight down (independent of considerable variation in the surface features of the scenario), then this would be inconsistent with an account in terms of associative memory (to the extent that different memories would be expected to lead to different predictions.

## Experiment

## Method

## Participants

27 University of Exeter students with ages ranging from 18-35 participated in this experiment. All were
undergraduates studying psychology, but were naïve to the hypotheses under test in this experiment.

## Design

Eight physics problems, featuring falling objects, were devised. These problems all had the same underlying structure, and therefore the same answer (in terms of Newtonian physics), but different surface and contextual features. The problems were of two types: Those in which the object was carried prior to being released, and those where the object had been moving freely (independently) throughout. These two problem types were further divided to give two subsets in which the expected answers were either congruent or incongruent with the predictions of a naïve physics theory.

Four different types of problems were, therefore, presented to the participants in a questionnaire:

Type 1: Carried - Congruent (CC) - Problems in which objects are carried prior to release and where our predicted answer "falls straight down" is in accordance with naïve physics theory.

Type 2: Carried - Incongruent (CI) - Problems in which objects are carried prior to release and where our predicted answer is not in accordance with naïve physics theory.

Type 3: Free - Congruent (FC) - Problems in which objects are moving freely/independently and where our predicted answer "curved forwards/parabolic trajectory" is in accordance with naïve physics theory.

Type 4: Free - Incongruent (FI) - Problems in which objects are moving freely/independently and where our predicted answer is not in accordance with naïve physics theory.

Two imaginary scenarios were devised for each type of problem. These were constructed using MS Powerpoint and Word software on a Macintosh computer as follows:

## Table 1

1. Carried Congruent condition (CC):

Problem 1. Bird in flight dropping ice cream
Problem 5. Plane dropping crate
2. Carried Incongruent condition (CI):

Problem 2. Swinging monkey drops banana
Problem 6. Cricket bowler drops ball at release
3. Free Congruent condition (FC):

Problem 3. Cannonball fired off cliff
Problem 7. Skier approaching a crevasse
4. Free Incongruent condition (FI):

Problem 4. Skateboarder dropping in to a half-pipe Problem 8. Water falling over a cliff


Figure 2. The figure shows one of the scenarios (Problem 1) presented for the CC condition. Participants were asked to select one of the 5 responses to indicate the trajectory of the dropped object (e.g. the ice-cream)

The problems in Table 1 were chosen so that the CC and FC examples, i.e. the congruent problems, were closely modeled on problems that have typically been used in previous experiments on naïve physics. As far as we could tell, there were no associations to events or situations that contradicted the predictions of a naïve physics theory for these examples. The incongruent problems (CI and FI) were chosen so that they conformed to the "Carried" or "Free" designation, but had associations that seemed to us to suggest that a response that was incongruent with a naïve physics theory would be given. Thus, a cricket bowler (e.g. Problem 6) is typically seen as projecting the ball forwards, not dropping it straight down. If you observe someone or something else (e.g. the monkey in Problem 2) in a state of motion carrying an object that they drop, then the typical perceived experience is for that object to continue to follow that state of motion. Because waterfalls (e.g. Problem 8) are typically seen from front-on and below, the modal experience is of them falling nearly straight down; and when


Figure 3. This shows examples of the other three types of scenario presented to our participants. The top panel is an example of a CI (Type 2 problem), the middle a FC (Type 3 problem) and the bottom a FI (Type 4 problem) example. In each case the same five response options were offered, illustrated by mini-drawings to show the path suggested (see Figure 2 for examples).
skateboarders drop in (e.g. Problem 4), they typically appear to take an initial path that is straight down. So, our hope was that these scenarios would predispose our participants to choose paths that were not expected on a naïve physics account.

## Procedure

The eight problems were collated into a nine page A4 sized booklet printed in black and white using size 12 Arial font. The front page consisted of instructions to the participants, a consent form and a question asking for information about their level of physics knowledge. Each of the following eight pages featured one of the physics problems, a pictorial multiple-choice selection showing the available answers and a text box with space for participants to write an explanation of the rationale behind their selection. The presentation of the problems was randomized, with each problem appearing once, and equally likely to appear on any page.

Participants were asked to look at each problem and attempt to predict the path of the object as it fell to the ground. They were then asked to select, by circling the relevant number from the multiple choice answers, which of the five responses: 1. Backwards, 2. Straight Down, 3. Out \& Down, 4. Curved forwards (Parabolic) and 5. Straight Forwards, most closely resembled the path they had thought of. Participants were also asked to write a short rationale for their choice of path to make clear the connection between their answer and their intuition. After completing all problems participants returned the booklet and were thanked for their contribution.

## Results

The data of interest are the frequencies of each response (1-5) provided for a given problem. Table 2 gives these frequencies for each problem. The responses that might be considered consistent with naïve physics (NP) for a given problem are shown in green, those that are not and are better explained by an appeal to associative memory (AM) are shown in red.

## Table 2

| Response |  |  |  |  |  |
| ---: | ---: | ---: | ---: | ---: | ---: |
| Question | 1 | 2 | 3 | 4 | 5 |
| 1 | 7 | 18 | 3 | 3 | 5 |
| 2 | 4 | 12 | 2 | 17 | 1 |
| 3 | 0 | 3 | 11 | 21 | 1 |
| 4 | 0 | 25 | 4 | 5 | 2 |
| 5 | 9 | 21 | 0 | 2 | 4 |
| 6 | 1 | 3 | 3 | 28 | 1 |
| 7 | 0 | 0 | 13 | 21 | 2 |
| 8 | 0 | 5 | 24 | 7 | 0 |

We took the view that for a "Carried" problem (Problems 1, 5, 2 and 6) the NP congruent response would be 2 ("Straight Down") and the AM (or NP Incongruent) response would be 3 ("Out \& Down") or 4 ("Curved Forwards"). For a "Free" problem (Problems 3, 7, 4 and 8) the NP responses we allowed as congruent were either 3 ("Out \& Down") or 4 ("Curved Forwards"). This was done
because, in practice, distinguishing between responses 3 and 4 was difficult. The AM (or Incongruent) equivalent response for these problems was 2 ("Straight Down"). Responses 1 and 5 were relatively rarely used, so when computing Chi-Square values we collapsed responses 1 and 2 together, and responses 3,4 and 5 together. This gave two basic classes of answer, which we can characterise as mostly straight down and certainly not forwards (1 and 2), and mostly curved forwards (3, 4 and 5).

We first of all collapsed over the two problems per condition, and then ran a $\chi^{2}$ as a $4 \times 5$ contingency table ( 4 conditions by 5 responses) to see if there was any effect of condition on responding. The resultant $\chi^{2}=157$, 12df, $\mathrm{p}<.005$ suggests that there is. We then collapsed further so that there were only two response classes as already detailed (to ensure that the expected values in each cell were sufficiently high), and carried out a series of $\chi^{2}$ analyses to investigate the basis of this effect. A $2 \times 2$ contingency table (hence, 1df) analysis of congruency by response gave a $\chi^{2}=0.95, \mathrm{p}=\mathrm{ns}$, indicating no main effect of this factor. Analysis of "Free" vs. "Carried" by response gave a $\chi^{2}=26.13, \mathrm{p}<.001$, showing that this factor exerted a strong influence over choice of response, with, as expected, "Free" problems tending to produce mostly curved forwards responses and "Carried" problems mostly straight down responses. If we break this down further, then the $\chi^{2}=3.06$, $\mathrm{p}=\mathrm{ns}$ for the Incongruent data suggested that there was no significant difference in the distribution of responses caused by this factor in these problems, but the $\chi^{2}=78.06, \mathrm{p}<.001$ for the Congruent data indicates that it is these problems that drove the strong tendency for the two types of problem to lead to different responses.

The data of main interest, however, are how the Congruency factor influences performance on the "Free" and "Carried" problems. Taking the "Free" problems first, a $2 \times 2$ contingency table analysis with congruency as a factor, and collapsed response as the other, produced a $\chi^{2}=28.66, \mathrm{p}<.001$, with Congruent problems favouring a curved forwards response over straight down answers, and Incongruent problems significantly reducing this tendency. The "Carried" data show an even clearer effect, with Congruent problems favouring straight down responses and Incongruent data reversing this effect to show a strong tendency to elicit curved forwards responses, $\chi^{2}=34.09$, $\mathrm{p}<.001$. It seems that we were successful in our attempt to select problems that either favoured the response expected on the basis of naïve physics (Congruent), or were not congruent with this prediction and instead owed more to associative memory (Incongruent). This effect was particularly marked for the "Carried" problems, where there is essentially a pure interaction, with the Congruent problems behaving exactly as naïve physics would predict and the Incongruent problems showing quite the opposite pattern. The results corresponding to these analyses are shown in an easily interpretable form in Figure 4.


Figure 4. This shows the response difference score for each of the four conditions in our experiment. This score is simply the difference between the aggregated responses, $(3+4+5)-(1+2)$, and measures preference for one response class over the other, with a positive score denoting a bias in favor of the curved forwards class over the straight down alternative.

In essence, Figure 4 demonstrates that whilst it is entirely possible to get the pattern of results predicted by a naïve physics model for how people solve this type of problem, it is also possible to choose problems such that the effect is eliminated.

## General Discussion

What are we to conclude from these findings? Perhaps the first, and most obvious conclusion, is that a simple naïve physics theory that predicts the "straight down effect" because carried objects do not have any impetus of their own is not going to be able to explain these results. Either the theory is wrong, or it is not being applied in these situations. And this last is a real possibility. By giving considerable context to the problems, we have definitely biased participants in the way that they approach them, perhaps they do not invoke a naïve physics in these circumstances because the problems are not abstract enough?

Our objection to this analysis would be that naïve physics is exactly that which should be able to deal with these "real world" situations. And furthermore, there is as good qualitative evidence for naïve physics being used in our experiment as there is in other studies that have used this evidence to argue for a naïve physics theory. If we take the CC class of problem first, we have the following quotes that are typical of the approach taken to these problems by our participants.
For Problem 1: "As soon as the seagull lets go there is no forward momentum, so therefore it will drop straight down."
And for Problem 5: "The box dropped straight out of the plane so would not have been affected by the movement of the plane."

Both these problems are Carried Congruent scenarios. The typical response selection was No. 2, "straight down", and the explanation offered is the classic "an object has no impetus of its own if carried" justification given on the basis of a naïve physics. But if we now consider the explanations given for the Carried Incongruent problems we have for Problem 6: "From my experience of ball games, they don't just drop downwards."
Which clearly indicates a reliance on experience that we take to be mediated by associative memory. It might be argued that the memory simply overrides the predictions of the theory in this instance, and of course this is a possible explanation of the forward path typically chosen for this problem. But when we come to Problem 4, one of the Free Incongruent scenarios, then the justification offered for choosing the straight down response (No. 2) is in Problem 4: "That's how skaters do it."
So we would have to argue that once again recall based on experience is overriding the predictions of the theory (which would predict that the path is curved forwards). At this point the reader will notice that in every case the answer chosen is one consistent with generic experience of the world, and this impression is confirmed by our final class of problem, Free Congruent, where the typical response is curved forwards as a naïve physics would predict, and the generic explanation for offering this response is in Problem 7:
"Due to moving at speed."
Which fits nicely with the idea that a freely moving object has impetus. Thus we have two conditions where we have the results and the rationale expected on a naïve physics view, and two conditions where we have the opposite. But in all cases, the responses and rationale seem grounded in experience, and an explanation based on extrapolation from experience is tenable. Surely in these circumstances it is more parsimonious to attribute the answers given to memory-based extrapolation from experience, rather than invoke some abstracted theory that has to be overridden much of the time?

But we do not think that just any memory-based extrapolation from experience will serve to explain our results. Instead, we believe that the memory involved is associative in nature, so that it has captured the basic statistical regularities embedded in experience and retrieves them on the basis of a surface similarity to the problem. As such, we believe that an error-correcting system (e.g. McClelland and Rumelhart, 1985, Rumelhart, Hinton and Williams, 1986, and see McLaren and Dickinson, 1990 for a discussion) is required, as this will be able to extract the necessary structure. A version of such a system that can then function as a model of associative memory would be ideal (e.g. see the model in McLaren, 2011, based on McLaren, 1993, and the most recent version of this in McLaren, Forrest and McLaren, 2012). Finally, the ability to capture structure over time will also be needed, and for this the SRN (Elman, 1990) and it's more sophisticated variant the Augmented SRN (Cleeremans and McClelland, 1991, see
also Yeates, Jones, Wills, McLaren and McLaren, in press, and the APECS variant in Jones and McLaren, 2001) fits the bill.

This more primitive, associative system is only part of the story, however, we also postulate another rule-based system that takes the output of associative memory and then constructs a story about the answer given around it. In doing this we are advocating a dual-process theory of cognition along the lines of that given in McLaren, Green and Mackintosh (1994), and illustrated in Spiegel and McLaren (2003, 2006) and Jones and McLaren (2009). It is this combination of extrapolation based on experience, followed by induction of some heuristic to explain why a particular answer has been given, that we believe has led to the notion of a naïve physics. It results in reliably incorrect answers to physics problems, and a narrative that accompanies these answers. The point of our research, however, is to show that if we frame what is essentially the same problem in a different way, so that we access a different type of experience, then the answer changes and so does the accompanying narrative. Clearly, if a deeper physical analysis of the problems were involved in accessing experience, the answer to all the problems studied here would be the same, a parabolic path forwards. Hence we have to postulate retrieval on the basis of surface similarity for this aspect of our theory to work. We would then argue that actually the inductive inference that suggests a naïve physics is more a matter of an attempt to "make sense" of our participants intuitive response to these scenarios.

## Conclusion

We have arrived at a position where the statements made by participants attempting to solve simple physics problems and taken to support the existence of a naïve physics are seen as post-hoc rationalization for the answer given rather than causally implicated in that answer. We believe that it is extrapolation based on experience, via retrieval from associative memory (that is itself the product of associative learning), which is responsible for the reliably incorrect answers given to the problems studied here. We would go further, and say that our position also applies to the more abstract problems often studied in naïve physics experiments, though clearly here the experiential memories involved must be of a more generic nature. Take as an example the problem shown in Figure 1. How are we to explain that result? What memory could be accessed for that problem? There are not many retrieval cues, just a curved tube and a ball bearing. But this is enough to retrieve memories of water emerging from a garden hose (as these are often curved) - and the path the water takes is typically curved as well. The analogy between hose and problem is, of course, incorrect - but the superficial similarity exists and this is what drives associative processing. The result is an extrapolation to an incorrect, curved path, because it feels right. And then we tell a story about why we gave that answer. The great advantage of this explanation is that it
generalizes to the results reported in this paper. And so we conclude that as a theory of why we seem to have the wrong idea about how objects move, it is to be preferred to the naïve physics point of view.

## References

Cleeremans, A. \& McClelland, J.L. (1991). Learning the structure of event sequences. Journal of Experimental Psychology: General, 120, 235-253.
Elman, J.L. (1990). Finding structure in time. Cognitive Science, 14, 179-211.
Jones, F.W., \& McLaren, I.P.L. (2001). Modelling the detailed pattern of SRT sequence learning. In J.D. Moore, \& K. Stenning (Eds.), Proceedings of the twenty-third annual conference of the cognitive science society (pp. 465-470). LEA: Mahwah, NJ.
Jones, F.W. and McLaren, I.P.L. (2009). Human Sequence Learning Under Incidental and Intentional Conditions. Journal of Experimental Psychology: Animal Behavior Processes, 35, 538553.

McCloskey, M., Caramazza, A., and Green, B. (1980). Curvilinear Motion in the Absence of External Forces: Naïve Beliefs about the Motion of Objects. Science, 210, 1139-41.
McCloskey M, Washburn A \& Felch L (1983) Intuitive Physics: The Straight Down Belief and Its Origin, Journal of Experimental Psychology: Learning, Memory and Cognition, Vol 9 (4), 636-649.
McClelland, J.L. and Rumelhart, D.E. 1985. Distributed memory and the representation of general and specific information. Journal of Experimental Psychology: General, 114, 159-88.
McLaren, I.P.L. and Dickinson, A. (1990). The conditioning connection. Phil. Trans. R. Soc. Lond. B, 329, 179-186.
McLaren, I.P.L. (1993). APECS: a solution to the sequential learning problem. Proceedings of the XVth Annual Convention of the Cognitive Science Society, 717-22.
McLaren IP, Green RE \& Mackintosh NJ (1994) Animal Learning and the Implicit/Explicit Distinction in Ellis N (ed) Implicit and Explicit Learning of Languages, London, Academic Press.
McLaren, I.P.L. (2011). APECS: An adaptively parameterised model of associative learning and memory. In Alonso, E. \& Mondragón, E. (Eds.). Computational Neuroscience for Advancing Artificial Intelligence: Models, Methods and Applications. Hershey, PA: IGI Global.
McLaren, I. P. L., Forrest, C.L., and McLaren, R.P. (2012). Elemental representation and configural mappings: Combining elemental and configural theories of associative learning. Learning and Behavior, 40 (3), 320-33.
Rumelhart, D.E., Hinton, G.E. and Williams, R.J. 1986. Learning internal representations by error propagation. In D.E. Rumelhart and J.L. McClelland (Eds.) Parallel Distributed Processing. Vol. I. Cambridge. Mass. Bradford Books. pp. 318-362.
Spiegel, R. \& McLaren, I.P.L. (2003). Abstract and associativelybased representations in human sequence learning. Phil Trans. Roy. Soc B., 358.
Spiegel, R. and McLaren, I.P.L. (2006). Associative Sequence Learning in Humans. Journal of Experimental Psychology: Animal Behavior Processes, 32, 2, 156-63.
Yeates, F., Jones, F.W., Wills, A.J., McLaren, R.P. and McLaren, I.P.L. (in press). Modelling human sequence learning under incidental conditions. Journal of Experimental Psychology: Animal Behavior Processes.

# Sequential Diagnostic Reasoning with Verbal Information 

Björn Meder (meder@mpib-berlin.mpg.de)<br>Max Planck Institute for Human Development, Center for Adaptive Behavior and Cognition (ABC)<br>Lentzeallee 94, 10495 Berlin, Germany

Ralf Mayrhofer (rmayrho@uni-goettingen.de)
Department of Psychology, University of Göttingen
Gosslerstrasse 14, 37073 Göttingen, Germany


#### Abstract

In sequential diagnostic reasoning, the goal is to infer the probability of a cause event from sequentially observed effects. Typically, studies investigating such tasks provide subjects with precise quantitative information regarding the strength of the relations between causes and effects. By contrast, we examined people's performance when this information is communicated through qualitative, rather vague verbal terms (e.g., " $X$ occasionally causes symptom $A$ "). We conducted an experiment in which we compared subjects’ judgments with a Bayesian model whose predictions were derived using numeric equivalents of various verbal terms from an unrelated study with different subjects. We found a remarkably close correspondence between subjects' diagnostic judgments based on verbal information and the model's predictions, as well as compared to a matched control condition in which information was presented numerically. Additionally, we observed interindividual differences regarding the temporal weighting of evidence.


Keywords: Sequential diagnostic reasoning; verbal reasoning; causal inference; Bayesian models; recency effects; linguistic probability terms; evidence accumulation

## Introduction

In diagnostic reasoning, the goal is to infer the probability of a cause event from observing its effects. The characteristic feature of sequential diagnostic reasoning is that multiple pieces of evidence are observed at different points in time. For instance, a doctor whose aim is to infer the cause of a patient's symptoms may take a blood sample and order different diagnostic tests. The test results may come in distributed over time, with each result potentially providing evidence for different diseases. Thus, sequential diagnostic reasoning requires keeping track of the evidence and the hypotheses under consideration.

We investigated different aspects of sequential diagnostic reasoning. Theoretically, we considered different ways in which such tasks can be modeled. For instance, standard probability calculus (e.g., Bayes's rule) is not sensitive to the temporal dynamics of evidence accumulation. Yet, there are ways to incorporate temporal weighting of evidence into probabilistic models of diagnostic reasoning and to model its potential influence on people's inferences.

Empirically, we were interested in investigating diagnostic reasoning with verbal information. In many real-world situations, everyday language is used to communicate probability or frequency information (Budescu \& Wallsten,

1995; Teigen \& Brun, 2003). For example, we might find it unusual if a doctor told us that the probability of a particular disease causing some symptom is $66 \%$. By contrast, a statement such as "disease $X$ frequently causes symptom $A$ " may feel more natural, despite the apparent lack of preciseness (Wallsten, Budescu, Zwick, \& Kemp, 1993).

Although using verbal probability terms is common in many real-world situations, they do not easily fit with computational accounts of cognition. As a consequence, in most behavioral studies subjects are provided with precise quantitative information (e.g., Meder, Mayrhofer, \& Waldmann, 2009). By contrast, we investigated reasoning with verbal information by using the numeric equivalents of linguistic terms (Bocklisch, Bocklisch, \& Krems, 2012) to derive model predictions. This allowed us to examine people's capacity to make diagnostic inferences in the absence of quantitative information and to compare their judgments to different accounts, including variants of Bayesian and linear models. To test for the temporal weighting of information, we varied the testing conditions by manipulating whether all evidence obtained so far was directly available when making a judgment or had to be partially retrieved from memory.

## Modeling Sequential Diagnostic Reasoning

The characteristic feature of sequential diagnostic reasoning is that different pieces of evidence are acquired step by step. Consider the causal model shown in Figure 1a. There are two (mutually exclusive) cause events, $X$ and $Y$; each can generate effects $A, B, C$, and $D$. In our experiment the cause variables were two chemical substances and the effects were different symptoms caused by these substances. The symptoms were observed sequentially and the goal was to infer whether $X$ or $Y$ caused them. How can such inferences be formally modeled?

## Standard Model: "Simple" Bayes

Let $S$ denote a set of symptoms $\left\{S_{1}, \ldots, S_{T}\right\}$, and let $X$ and $Y$ denote two mutually exclusive causes that can generate $S$. Since $X$ and $Y$ are mutually exclusive, $P(Y \mid S)=1-P(X \mid S)$. The posterior probability of cause $Y$ given the symptoms, $P(Y \mid S)$, can be computed using Bayes's rule:

$$
\begin{equation*}
P(Y \mid S)=\frac{P(S \mid Y) P(Y)}{P(S \mid Y) P(Y)+P(S \mid X) P(X)} \tag{1}
\end{equation*}
$$

where $P(S \mid Y)$ denotes the likelihood of the symptoms given cause $Y, P(Y)$ is the base rate of cause $Y$, and $P(S \mid X)$ and $P(X)$ denote the corresponding estimates for the alternative cause.

We considered only situations in which $X$ and $Y$ were equally likely a priori, thus $P(X)=\mathrm{P}(Y)=.5$. In this case, Eq. (1) simplifies to

$$
\begin{equation*}
P(Y \mid S)=\frac{P(S \mid Y)}{P(S \mid Y)+P(S \mid X)} \tag{2}
\end{equation*}
$$

Thus, the posterior probability of $Y$ is a function of the likelihood of the set of symptoms $S$ under each of the two hypotheses $X$ and $Y$.
a)

Causal model

b)

Relations and presentation formats

| Likelihoods | Numeric | Verbal |
| :--- | :--- | :--- |
| $P(A \mid X)$ | $88 \%$ | "almost always" |
| $P(A \mid Y)$ | $19 \%$ | "infrequently" |
| $P(B \mid X)$ | $66 \%$ | "frequently" |
| $P(B \mid Y$ | $29 \%$ | "occasionally" |
| $P(C \mid X)$ | $29 \%$ | "occasionally" |
| $P(C \mid Y)$ | $66 \%$ | "frequently" |
| $P(D \mid X)$ | $19 \%$ | "infrequently" |
| $P(D \mid Y)$ | $88 \%$ | "almost always" |

Figure 1: (a) Causal structure used in our diagnosis task, and (b) strength of the individual links (likelihoods) in the numeric and verbal formats used in the experiment.

## Temporal Weighting of Evidence: "Memory" Bayes

For the simple Bayes model, the temporal order in which observations are made does not matter: The resulting probabilities are the same regardless of whether beliefs are updated sequentially according to the individual symptoms or conditional on all symptoms at once.

However, we were also interested in modeling the sequential dynamics of evidence accumulation. For instance, diagnostic inferences can be influenced by memory limitations, such as the (partial) neglect of earlier obtained evidence. To model the influence of time, we applied the log odds form of Bayes's rule to the target inference:

$$
\begin{equation*}
\varphi=\log \frac{P(Y \mid S)}{P(X \mid S)}=\log \frac{P(S \mid Y)}{P(S \mid X)}+\log \frac{P(Y)}{P(X)} \tag{3}
\end{equation*}
$$

Assuming both hypotheses are equally likely a priori, we can omit the prior odds from the derivation and expand the likelihood odds by summing over the sequence of symptoms $S_{1}, \ldots, S_{T}$ given their conditional independence:

$$
\begin{equation*}
\varphi=\sum_{t=1}^{T} \log \frac{P\left(S_{t} \mid Y\right)}{P\left(S_{t} \mid X\right)} \tag{4}
\end{equation*}
$$

where $t$ is the current symptom and $T$ is the total number of symptoms observed so far.

The log posterior odds can then be transformed into a conditional probability by an inverse-logit transformation:

$$
\begin{equation*}
P(Y \mid S)=\frac{1}{1+e^{-\varphi}} \tag{5}
\end{equation*}
$$

This equation is mathematically equivalent to the standard form of Bayes's rule for the posterior probability of $Y$ given the set of symptoms $S$ as shown in Eq. (2).

Importantly, the log-odds form allows us to introduce an exponential decay parameter $\delta$ that controls the weighting of symptoms in the course of their presentation (Steyvers, Tenenbaum, Wagenmakers, \& Blum, 2003). Therefore, we replace Eq. (4) by

$$
\begin{equation*}
\varphi=\sum_{t=1}^{T}\left[\log \frac{P\left(S_{t} \mid Y\right)}{P\left(S_{t} \mid X\right)}\right] e^{-\frac{T-t}{\delta}} \tag{6}
\end{equation*}
$$

In the limit, if $\delta=\infty$, there is no decay and Eq. (6) reduces to Eq. (4). In this case, all symptoms are equally weighted and symptom order does not matter (as predicted by the simple Bayes model). The closer $\delta$ is to 0 , the more weight is given to the current symptom. If $\delta=0$, the posterior probability depends on only the most recent symptom, yielding an inference strategy that is completely ignorant of past information (e.g., an agent without memory). Thus, "memory" Bayes can be used to model recency effects (Hogarth \& Einhorn, 1992; Trueblood \& Busemeyer, 2011).

## Mapping Verbal Terms to Numbers

A key question of our research was how accurately people reason with verbal information, absolutely and relative to situations in which quantitative information is available. Answering this requires translating verbal expressions into a numeric representation that can be used to derive precise model predictions for the verbal reasoning task.

We used numeric equivalents of verbal expressions from a study by Bocklisch et al. (2012). They asked subjects to provide numeric estimates for different verbal expressions in a frequency format (e.g., It is frequently necessary to work at a rapid pace means "in $X$ of 100 work tasks/cases"). This mapping of words to numbers provided the basis for our empirical study, in which we used the four verbal expressions "infrequently", "occasionally", "frequently", and "almost always" to convey the strength of the cause-effect relations. ${ }^{1}$ The corresponding numeric mean estimates were $19 \%, 29 \%, 66 \%$, and $88 \%$ (Figure 1b).

These estimates were used to derive posterior probabilities of the causes given the symptoms via Bayes's rule [Eq. (1)], which served as normative benchmarks for evaluating subjects' diagnostic judgments. Note that the numeric equivalents were elicited from a different, unrelated sample from the one used in our study.

[^152]
## Experiment

The main goal of our study was to investigate sequential diagnostic reasoning with verbal information and compare different presentation formats with respect to the temporal weighting of evidence.

The first factor we manipulated was the way in which subjects were informed about the strength of the relations between causes and effects. In the verbal condition the strength of the individual relations was conveyed through four verbal terms ("infrequently", "occasionally", "frequently", "almost always"). In the numeric condition, causal strengths were presented in a percentage format. The two formats were matched using the estimates from Bocklisch et al. (2012). For instance, in the verbal condition subjects learned that $X$ "almost always" causes $A$, whereas in the numeric condition subjects learned that the probability of $X$ causing $A$ is $88 \%$ (see Figure 1b).

With the second manipulation we aimed to investigate possible influences of temporal weighting on diagnostic judgments (i.e., recency effects). In the single-symptom condition, only the current symptom was visible on the computer screen when participants made the diagnostic judgment. In the all-symptoms condition, the full set of symptoms reported so far was visible on the screen when they made a diagnosis. ${ }^{2}$ The rationale behind this manipulation was that there might be a tendency to overweight the currently presented symptom when previously obtained evidence needs to be recalled from memory.

## Method

Participants One hundred fifty-six students (103 women; $M_{\text {age }}=23.4$ years) from the University of Göttingen, Germany, participated in this experiment as part of a series of various unrelated computer-based experiments. Subjects either received course credit or were paid $€ 8$ per hour.

Materials and Procedure We used a hypothetical medical diagnosis scenario in which the subjects' task was to find out which of two fictitious chemical substances was the cause of certain symptoms in patients. The instructions asked subjects to take the role of a doctor being responsible for the workers at two chemical plants. At one plant workers may come in contact with the substance "Altexon"; at the other they may have contact with "Zyroxan". Each of these substances can cause four symptoms: dizziness, fever, headache, and vomiting. The assignment of labels to causes (substances) and effects (symptoms) was randomized.
Subjects were informed that their task would be to diagnose a series of workers who had had contact with either of the two substances. The instructions explicitly stated that accidents were equally likely to happen in each of the plants (i.e., the base rate of each cause was $50 \%$ ). Subjects were

[^153]also told that the patients would report their symptoms sequentially.
The experiment consisted of two phases: a learning phase, in which subjects learned the strengths of the individual causal relations, and a test phase, in which subjects were sequentially presented with symptoms of different patients and had to make a diagnostic judgment after each symptom.

Figure 1b illustrates the strengths of the relations between substances and symptoms according to the two presentation formats. In the learning phase, the subjects' task was to learn the strength of the individual relations in a trial-bytrial fashion. On each trial, subjects were shown a substance along with a symptom and had to estimate how often the substance causes the symptom. In the verbal condition, possible answers were "infrequently", "occasionally", "frequently", and "almost always". In the numeric condition, the corresponding answers were $19 \%, 33 \%, 66 \%$, and $88 \%$. After making an estimate, subjects received feedback regarding the actual relation. The eight relations were presented block-wise, with the order randomized within a block. To proceed to the test phase, subjects needed to answer two consecutive blocks correctly (or pass 12 blocks in total).
In the test phase, the subjects' task was to make diagnostic judgments for different sequences of symptoms, with each symptom sequence referring to a different patient who had come in contact with either $X$ or $Y$. Each test trial consisted of three sequentially presented symptoms (e.g., $A-D-C$ ), with a diagnostic judgment requested after each symptom. In the all-symptoms condition, all symptoms reported so far were present on the screen. In the single-symptom condition, only the current symptom was displayed. All judgments were given on an 11-point scale from 0 to 100 , with the endpoints labeled as "The patient definitely had contact with Altexon" and "The patient definitely had contact with Zyroxan".
Table 1 shows the six symptom sequences together with the posterior probabilities derived using the likelihoods shown in Figure 1b, assuming $P(X)=P(Y)=.5$. Additionally, we presented the six symptom sequences that entailed identical posterior probabilities for $X$ (e.g., $P(Y \mid A-D-$ $C)=P(X \mid D-A-B)$ such that diagnoses were counterbalanced. Thus, each subject saw 12 sequences in total. The corresponding pairs were later recoded and aggregated.
The test trials were administered in random order. After the test phase, we tested subjects again with respect to the strength of the individual substance-symptom relations (as learned in the learning phase) by presenting an additional

Table 1: Test trials with sequentially presented symptoms.

| Posterior | Symptom sequence |  |  |  |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $A-D-C$ | $D-A-C$ | $B-C-A$ | $C-B-A$ | $A-C-D$ | $C-A-D$ |
|  | .18 | .82 | .31 | .69 | .18 | .69 |
|  | .50 | .50 | .50 | .50 | .33 | .33 |
|  | .69 | .69 | .18 | .18 | .69 | .69 |

Note. Numbers refer to the posterior probability of cause $Y$ given a set of symptoms $S$ according to the simple Bayes model.
block of the learning phase (without feedback). This served as a manipulation check to ensure that subjects still remembered the relations between substances and symptoms.

Design Subjects were randomly assigned to one of the 2 (numeric vs. verbal) $\times 2$ (single vs. all symptoms) betweensubjects conditions. Within each subject, we aggregated over the (recoded) judgments within the counterbalanced pairs of trials, yielding 6 (trials) $\times 3$ (symptoms) withinsubject conditions with judged $P(Y \mid S)$ as dependent measure.

## Results and Discussion

Learning Criterion At the end of the experiment we tested subjects on the eight substance-symptom relations presented in the learning phase. Because the strength of the individual relations is the basis for the diagnostic judgments, we excluded all subjects who could not reproduce at least seven of the eight relations correctly. Accordingly, $28.2 \%$ of the subjects were excluded from the analyses, yielding between 27 and 30 valid subjects per condition (total $N=112$ ).

Overall Fit Figure 2 shows subjects' mean diagnostic judgments for the different symptom sequences along with the posterior probabilities derived from the simple Bayes model. A first inspection of the data indicates that subjects' judgments were remarkably accurate, with estimates being close to the true posteriors. This was the case regardless of whether information was provided in a verbal or numeric format (especially in the all-symptoms condition; see lefthand side of Figure 2). Thus, subjects were capable of making pretty accurate inferences when reasoning with verbal information. This close correspondence is particularly remarkable because the numeric equivalents of the verbal terms were taken from a different sample of subjects who participated in an unrelated study (Bocklisch et al., 2012).

Before conducting the model-based analysis, we ran a mixed analysis of variance with the 2 (numeric vs. verbal) $\times$ 2 (single vs. all symptoms) conditions as between-subjects factors and the 6 (trials) $\times 3$ (symptoms) conditions as with-in-subject factors. The key result of this analysis was that there was no main effect of presentation format, $F(1,108)<1$, a weak effect of the single- vs. all-symptom presentation manipulation, $F(1,108)=3.1, p=.08, \eta_{\mathrm{p}}=.03$, and no interaction $(F<1)$.

To evaluate subjects' overall accuracy we computed the correlation and mean squared error (MSE) between the empirical judgments and the posterior probabilities derived from the simple Bayes model (note that no fitting is involved here). To address if symptoms are weighted differently in sequential reasoning, we fitted the decay parameter $\delta$ of the "memory" Bayes model to the data (separately for each condition, using the MSE as fitting criterion). ${ }^{3}$ The relative size of the decay parameter $\delta$ in the single- vs. allsymptoms condition gives an idea of whether the testing

[^154]

Figure 2: Mean diagnostic judgments ( $\pm 95 \% \mathrm{CI}$ ) and predictions of the simple Bayes model. Rows represent the different trials (see Table 1), separately by testing procedure (all-symptoms vs. single-symptom presentation).
procedure influences people's judgments, in particular whether there is a tendency to neglect previous evidence when only the current symptom is shown when performing a diagnostic judgment. This should result in lower values of the decay parameter $\delta$ for the single-symptom condition relative to the all-symptoms condition. ${ }^{4}$

Table 2: Fits of the "simple" and "memory" Bayes models.

| Format | Symptoms | Simple Bayes |  | Memory Bayes |  |  |
| :--- | :--- | :---: | :---: | :---: | :---: | :---: |
|  |  | $r$ | $M S E$ | $r$ | $M S E$ | $\delta$ |
| Verbal | All | .991 | .0008 | .991 | .0009 | $>$ le +10 |
|  | Single | .952 | .0089 | .983 | .0036 | 2 |
|  | All | .996 | .0004 | .996 | .0004 | 40 |
|  | Single | .971 | .0028 | .988 | .0014 | 4.5 |

Table 2 shows the fits of the two models. Overall, both the (high) correlations and the (low) MSE indicate that the models' predictions fit well with subjects' judgments. In the all-symptoms conditions, the fit for simple Bayes was almost perfect ( $r=.991$ and $r=.996$, respectively), mirroring that for 35 of 36 ( 6 trials $\times 3$ symptoms $\times 2$ formats [verbal vs. numeric]) data points, the model's predictions fell inside the $95 \%$ confidence interval.

The results also indicate some neglect of previous evidence in both single-symptom conditions, in which only the current symptom was displayed on the screen when subjects made a diagnostic judgment (cf. Figure 2). Here, in both the verbal and the numeric condition, lower values for $\delta$ were obtained than in the all-symptoms conditions (see Table 2). Consistent with this finding, for these conditions the memory Bayes model achieved a higher fit than the simple Bayes model, in terms of both the correlation and the MSE. This result indicates that subjects were more likely to overweight the current evidence when previous symptoms had to be recalled from memory.

Model-Based Clustering Temporal weighting of cumulative evidence might not be due to the characteristics of the task or the reasoning context alone but might also result from interindividual differences or strategies. We therefore explored if it is possible to identify homogenous subgroups of subjects differing with respect to their temporal weighting of symptoms (i.e., that differ in the $\delta$ parameter).
To identify such clusters, we adapted the model-based clustering technique introduced by Steyvers et al. (2003), which was inspired by $K$-means clustering. The clustering problem requires solving two problems simultaneously: first, assigning subjects to clusters such that clusters are homogenous with respect to the model predictions, and second, estimating the best-fitting $\delta$ parameter for each cluster. This problem can be approximately solved by a recursive algorithm that starts with a random assignment of subjects to clusters and then iterates over two steps, namely, fitting and re-assignment, until no subject changes cluster. ${ }^{5}$

[^155]Table 3: Results of the model-based clustering.

| Format | Symptoms | Cluster 1 |  |  |  | Cluster 2 |  |  |  |
| :--- | :--- | :--- | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | $\delta$ | $n$ | $r$ | MSE | $\delta$ | $n$ | $r$ | $M S E$ |
| Verbal | All | $\infty$ | 28 | .991 | .0008 | - | - | - | - |
|  | Single | $\infty$ | 15 | .994 | .0014 | 0 | 12 | .993 | .0047 |
| Numeric | All | $\infty$ | 24 | .997 | .0003 | 1.5 | 3 | .969 | .0056 |
|  | Single | 85 | 21 | .993 | .0010 | 0.6 | 9 | .989 | .0017 |

Note. $\delta=\infty$ means that the estimate is greater than $1 \mathrm{e}+10$; in this case there is essentially no difference from the predictions of the simple Bayes model. $\delta=0$ means that the estimate is smaller than $1 \mathrm{e}-10$; in this case there is essentially no difference from the prediction of Bayes's rule taking into account only the currently presented symptom.

We applied this procedure to each of the four conditions; the results are shown in Table 3. The verbal all-symptoms condition yielded only one cluster as a solution, whereas the other three conditions yielded stable two-cluster solutions. Remarkably, in each condition the majority of subjects were assigned to a cluster that is best represented by a very high $\delta$ parameter of the memory Bayes model. Essentially, this means that these subjects are best described by a prediction profile that is almost identical to the predictions of the simple Bayes model. In the single-symptom conditions, however, a substantial proportion of people were best described by a quite low $\delta$ parameter, meaning that their diagnostic judgments were almost exclusively determined by the currently presented symptom. Taken together, the clustering results strengthen the findings we already obtained by the overall fitting of the data.

Linear Models of Diagnostic Judgment In our study, people's diagnostic judgments corresponded strongly to the predictions of Bayes's rule. Can alternative models approximate these predictions? We here consider one alternative class of models, namely, weighted-additive (WADD) approaches. From this view, the cause event is inferred using an average (i.e., linear) combination of symptom weights:

$$
\begin{equation*}
P(Y \mid S)=\frac{1}{T} \sum_{t=1}^{T} w_{S_{t}} \tag{7}
\end{equation*}
$$

where $t$ is the current symptom and $T$ is the total number of symptoms observed so far.

We tested three different linear models that make different assumptions regarding the decision weights. The simplest model, tallying, simply counts symptoms. In our scenario, symptoms $A$ and $B$ are more likely to be generated by $X$, whereas $C$ and $D$ are more likely to be generated by $Y$. Given a set of symptoms, one simply tallies the evidence. For instance, given the sequence $A-C-D$, two of the three symptoms provide evidence for $Y$; accordingly, the resulting

[^156]estimate would be $2 / 3$. Note that this result is very close to the true probability, which is .69 in this case.

The second linear model assumes that the decision weights reflect the strength of the cause-effect relations; we therefore call it likelihood WADD. This model simply sums over the likelihoods and normalizes the result by dividing it by the number of presented symptoms. Given the sequence $A-C-D$, this model would predict that the probability of $Y$ is $.58[(.19+.66+.88) / 3]$, which for this sequence is quite close to the true probability of 69 .

Finally, we examined the predictions of an "optimal" WADD model by fitting the weights to the data, using MSE minimization as a criterion. This model essentially serves as a benchmark, as it provides the best fit given the functional form of the model (linear combination) and the data.

Table 4: Fits of the linear models.

| Format | Symptoms | Tallying |  | $\begin{aligned} & \hline \text { Likelihood } \\ & \text { WADD } \\ & \hline \end{aligned}$ |  | Optimal <br> WADD |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | $r$ | MSE | $r$ | MSE | $r$ | MSE |
| Verbal | All | . 861 | . 0278 | . 901 | . 0093 | . 901 | . 0085 |
|  | Single | . 817 | . 0320 | . 856 | . 0242 | . 857 | . 0197 |
| Numeric | All | . 864 | . 0266 | . 896 | . 0106 | . 898 | . 0094 |
|  | Single | . 848 | . 0285 | . 880 | . 0108 | . 881 | . 0102 |

The results (Table 4) show that all linear models achieved a respectable fit, but none could match the Bayesian models. These results speak against the idea that our subjects used a linear-additive strategy to make judgments.

## General Discussion

Although verbal terms such as "infrequently", "occasionally", and "frequently" are rather vague and imprecise, they are commonly used in many real-world situations. In contrast, researchers interested in human probabilistic thinking and judgment under uncertainty usually provide their subjects with precise numeric information in order to compare their behavior and inferences to the predictions of computational models, which typically also require numeric input.

A key motivation underlying the present work was to investigate subjects' reasoning in situations that more closely resemble real-world situations, in which inferences must often be drawn in the absence of reliable quantitative information. Using a sequential diagnostic reasoning task, we observed that people's inferences were surprisingly accurate when information on cause-effect relations was conveyed merely through linguistic terms. In fact, performance was almost indiscernible from a control condition in which subjects were provided with numeric information. The fact that we took the numeric equivalents from a different study (Bocklisch et al., 2012) supports research showing that the interpretation of linguistic frequency terms is relatively stable across populations (Mosteller \& Youtz, 1990).

Generally, subjects' diagnostic judgments closely resembled the prediction of a simple Bayes model that operates on matched numeric values. This is a promising finding for applying computational models of cognition to verbal reasoning tasks. It is particularly interesting for Bayesian mod-
eling, as this approach is not restricted to numeric point estimates (e.g., mean of an elicited frequency term distributions) but can also operate on full distributions (e.g., fitted Beta distributions).

Furthermore, we investigated the temporal weighting of evidence. We found that symptoms were equally weighted when all relevant symptoms were available during judgment, but we also observed a neglect of previous evidence when only the current symptom was present. Model-based cluster analyses revealed that this was due to a subgroup of subjects who considered only the current symptom, whereas most people took into account all evidence in a normative fashion. Overall, our results contrast with views that consider human probabilistic reasoning as flawed and error prone.

## Acknowledgments

Both authors contributed equally to this research. We thank Franziska Bocklisch, Georg Jahn, Felix Rebitschek, and Michael Waldmann for helpful comments on this project. This research was supported by grants Wa 621/22 and Me 3717/2 from the Deutsche Forschungsgemeinschaft (DFG) as part of the priority program "New Frameworks of Rationality" (SPP 1516).

## References

Bocklisch, F., Bocklisch, S. F., \& Krems, J. F. (2012). Sometimes, often, and always: Exploring the vague meanings of frequency expressions. Behavior Research Methods, 44, 144-157.
Budescu, D. V., \& Wallsten, T. S. (1995). Processing linguistic probabilities: General principles and empirical evidence. Psychology of Learning and Motivation, 32, 275318.

Hogarth, R. M., \& Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. Cognitive Psychology, 24, 1-55.
Meder, B., Mayrhofer, R., \& Waldmann, M. R. (2009). A rational model of elemental diagnostic inference. In Proceedings of the 31st annual conference of the Cognitive Science Society (pp. 2176-2181). Austin, TX: Cognitive Science Society.
Mosteller, F., \& Youtz, C. (1990). Quantifying probabilistic expressions. Statistical Science, 5, 2-34.
Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J., \& Blum, B. (2003). Inferring causal networks from observations and interventions. Cognitive Science, 27, 453-489.
Teigen, K. H., \& Brun, W. (2003). Verbal expressions of uncertainty and probability. In D. Hardman (Ed.), Thinking: Psychological perspectives on reasoning, judgment and decision making. New York, NY: Wiley.
Trueblood, J., \& Busemeyer, J. R. (2011). A quantum probability account of order effects in inference. Cognitive Science, 35, 1518-1552.
Wallsten, T. S., Budescu, D. V., Zwick, R., \& Kemp, S. M. (1993). Preferences and reasons for communicating probabilistic information in numerical or verbal terms. Bulletin of the Psychonomic Society, 31, 135-138.

# Comparative Evidence For Associative Learning In Task Switching 

# Christina Meier (CM374@exeter.ac.uk) 

Charlotte L. D. Forrest (CLF206@exeter.ac.uk)
Stephen E. G. Lea (S.E.G.Lea@exeter.ac.uk)
Katharina Angerer (KA288@exeter.ac.uk)

Ian P. L. McLaren (I.P.L.McLaren@exeter.ac.uk)<br>Department of Psychology, Washington Singer Laboratories<br>Exeter, EX4 4QG, UK


#### Abstract

Humans can perform several different tasks on the same set of stimuli in rapid alternation. Each task, signaled by a distinct task cue, may require the classification of stimuli using a different stimulus attribute. However, such "task switching" performance comes at a cost, as expressed by weaker performance when switching rather than repeating tasks. This cost is often claimed to be the consequence of a mental reorientation away from the previous task and towards the new task, requiring executive control of behavior. Alternatively, task switching could simply be based on the retrieval of different cue-stimulus-response associations. In this experiment, pigeons learned go-left/go-right discriminations between grating patterns according to either their spatial frequency or their orientation, depending on the color of the pattern (the task cue). When humans solved the same tasks on the basis of verbalizable rules, they responded more slowly and made more errors on trials where they had to switch between tasks than when repeating the same task. Pigeons did not show this "switch cost"; but like humans, their performance was significantly worse when the response (left or right) to a given stimulus varied between tasks than when it stayed the same (the "congruency effect"). Larger effects of both switch costs and congruency were observed in humans learning the tasks by trial and error. We discuss the potential driving factors behind these very different patterns of performance for both humans and pigeons.


Keywords: executive control; associative learning; task switching; humans; pigeons; comparative cognition.

## Introduction

Humans are able to perform two or more different tasks on the same stimulus material when cued to do so (called "task switching"). Typically, each task requires the classification of a set of stimuli according to a different stimulus attribute. The task that is to be performed in a given trial is indicated by a specific task cue (for example, subjects might be asked to judge a grating pattern by its spatial frequency when the color yellow is presented, or to classify the same stimulus according to whether it is vertically or horizontally orientated when the color red appears).

However, it is still a matter of debate which cognitive mechanisms underlie human task-switching ability. Humans may classify the stimuli they see based on rules, and a common phenomenon of task switching, namely longer reaction times and higher error rates after switching tasks compared to repeating the same task ("switch costs"), has been assumed to reflect the executive control processes
associated with this rule use. For example, humans might sort a series of stimuli based on their orientation while ignoring other available stimulus dimensions such as spatial frequency. In this context, switching from one task to the next involves executive control when identifying the current task, retrieving its specific stimulus-response rules into one's working memory (while deleting the rules of the previous task) and adjusting one's response reaction to the new requirements: in short, a mental disengagement from the previous task and preparation for the currently relevant task, known as "task-set reconfiguration" (Vandierendonck, Liefooghe \& Verbruggen, 2010). Switch costs are thought to reflect the need for such reconfiguration in switch trials but not in repeat trials, for which the task-set is already available (Monsell, 2003).

But, if we believe that humans have multiple processes available that support learning (McLaren, Green \& Mackintosh, 1994), task-switching phenomena might be the result of associative learning mechanisms, i.e., the retrieval of cue-stimulus-response associations (Logan \& Bundesen, 2003). Learning to respond correctly in a task-switching paradigm could be accomplished by associating the overall appearance of a stimulus with a certain response (Lea \& Wills, 2008). Each stimulus could be categorized by using its combined dimensions and comparing its similarity to a stimulus to which the correct response is known.

A task-switching phenomenon often observed in addition to switch costs, the effects of stimulus congruency (Monsell, Yeung \& Azuma, 2000), might indeed be better explained by associative learning processes than task-set reconfiguration. As each task makes use of the same set of multidimensional stimuli, stimulus values on individual dimensions can be defined as either congruent or incongruent in relation to the correct response towards them. If a stimulus is congruent, it always requires the same response regardless of the current task; learning to discriminate between different congruent stimuli thus takes the form of a component discrimination, in which the correct response depends on a single element of a multidimensional stimulus. However, when an incongruent stimulus is shown, the correct response varies depending on the current task in the manner of a biconditional discrimination. Given that there is good evidence that such discriminations are difficult to learn (Harris \& Livesey, 2008), it is no surprise that on trials in which a congruent stimulus is shown, reaction time and error rate are distinctly
lower compared to trials with an incongruent stimulus, and humans can exhibit large congruency effects (Monsell, Yeung \& Azuma, 2000). Experiments intended to elicit an associative approach to task switching in humans, either by only providing cue-stimulus-response contingencies instead of full task instructions or by forcing participants to learn how to respond by trial and error, have yielded very large congruency effects and switch costs that were considerably smaller than the effects of congruency (Forrest, Elchlepp, Monsell \& McLaren, 2012).

Humans can communicate the extent to which they refer to certain rules when reacting to a stimulus. It is therefore potentially possible to identify a group of participants who learned the responses based on rules about the stimulus dimensions and those who did not, and compare their performance to that of animals, who might not have the same cognitive mechanisms available.

Stoet and Snyder (2003) were the first to explicitly investigate task-switching effects in nonhuman animals. Their two rhesus macaques behaved very similarly to Forrest et al.'s (2012) humans who were presumed to be learning associatively: while their performance produced a large congruency effect, switch costs were rather small, and in fact absent in one animal. Stoet and Snyder (2003) acknowledged that monkeys might lack at least one of the cognitive control mechanisms necessary to solve taskswitching paradigms in the typical human way, but they did not doubt that their subjects used some form of executive control. However, the possibility remains that both humans and nonhuman animals might solve a task-switching paradigm associatively.

If humans who claim to be unaware of any rules underlying a task-switching paradigm employ an approach similar to that in animals assumedly solving the task by purely associative processes, such as the pigeon (Mackintosh, 1988), their performance would be expected to resemble that of those animals. For this purpose, pigeons might make a more suitable comparison than primates; they can also be tested in larger numbers than monkeys, so more reliable results should be obtained.

To design a paradigm suitable for use with both humans and animals, instead of relying on language-based stimuli or cues, we used varying values of the visual dimensions of color, orientation and spatial frequency to indicate a correct response. Additionally, we used trial-and-error training on cues and stimuli that resembled the conditioning procedures usually employed in animal testing. To investigate whether human performance under these conditions can be compared to that based entirely on associative-learning processes, we trained humans and pigeons on the same paradigm.

## Method

## Subjects

Twenty-four Psychology undergraduate students, in exchange for course credit, and eight pigeons (Columbia livia) participated in this experiment. Pigeons were kept in
an indoor aviary ( $2 \times 1 \times 2.5 \mathrm{~m}$ ) that housed 15 pigeons at the time of the experiment. They were maintained at or above $80 \%$ of their free-feeding weight. Both humans and pigeons were naïve to the testing stimuli, though pigeons had previously been trained to peck at a white observing key presented in the center of a black touch-sensitive display, followed by a peck at a red, blue, green or yellow colored circle appearing in the same position, and finally, to peck at a white reward key randomly presented either to the left or to the right of the display.

## Apparatus

All experiments were carried out inside the Washington Singer Laboratories at the University of Exeter. Pigeons were tested in one of eight identical $71 \times 50.5 \times 43.5 \mathrm{~cm}$ operant chambers. Each pigeon was always tested in the same chamber. One of the long walls of the chamber was fitted with a $31 \times 23.5 \mathrm{~cm}$ (15") touch monitor (Model 1547L 1024x768pxl TFT monitor, CarrollTouch infrared detector, ELO Touchsystems Inc.) mounted 12cm above the grid floor of the chamber. Two 2.8 Watt white houselights were mounted to either side above the screen; below the screen, mounted 4 cm above the chamber floor and directly below each house light, two $6 \times 5 \mathrm{~cm}$ apertures gave access to grain hoppers when solenoids were activated. The food hoppers were illuminated by a 2.8 Watt light when activated and contained a $2: 1$ mixture of hemp seed and conditioner. Also mounted below the screen between the two food hoppers, a 50 Ohm loudspeaker played white noise into the box as well as indicating effective pecks to target areas with an immediate beep. The interior of the box was monitored by a video camera attached to the short wall of the chamber opposite the chamber door. Contingencies were controlled and data collected using a PC computer running the Whisker system (Cardinal \& Aitken, 2010) with client programs written in Visual Basic 6.0. Humans were tested in a small experimental room on an iMac. The program was written in MatLab R2008b ${ }^{\circledR}$ asing the Psychtoolbox (Kleiner, Brainard \& Pelli, 2007) add-on and run using MatLab2011b ${ }^{\circledR}$.

## Procedure

For pigeons, each trial began with the presentation of an observing key (100 pixels in diameter) presented in the center of a black display. Following two pecks at the observing key, it was replaced by one of four task cues, a color-filled circle of 200 pixels in diameter, in the display center. Each of the two tasks was associated with two distinct cues: these were blue or yellow for task A, and red or green for task B. For humans, a trial started immediately with the presentation of the cue; that is, no observing key was presented. Pigeons had to peck the task cue twice, after which the task stimulus appeared, superimposed on the cue, making both the cue and the stimulus visible simultaneously. Humans were asked to mouse-click once on the cue, upon which the stimulus appeared in the same way as for pigeons.

Stimuli, made up as circular Gaussian patches of 200 pixels in diameter, consisted of one of four sinusoidal grating patterns, differing from another in two dimensions: spatial frequency - either low ( 2 cycles per 100 pixels (c/100px) for pigeons and $3 \mathrm{c} / 100 \mathrm{px}$ for humans) or high ( $12 \mathrm{c} / 100 \mathrm{px}$ for pigeons and $10 \mathrm{c} / 100 \mathrm{px}$ for humans) - and line orientation - either horizontal or vertical. All combinations of cue color, spatial frequency and orientation were used, resulting in 16 visually distinct images.

The correct response towards a stimulus depended on the task. For example, for some participants, task A required responding to the spatial frequency of the grating pattern, e.g., if a stimulus, regardless of the orientation of the pattern, had a low spatial frequency, the correct response towards this stimulus was to choose the left reward location, while stimuli with a high spatial frequency afforded choosing the right reward location. Conversely, in task B, stimuli would have to be classified according to the orientation of the grating pattern, regardless of its spatial frequency. That is, if a stimulus showed a horizontal pattern, it required a response to the left reward location, while a vertical pattern indicated a response to the right location as the correct one. Although blue and yellow were always assigned to task A, the stimulus attributes (spatial frequency or orientation) that were important for classification in task A and the reward location that was associated with any cuestimulus combination were counterbalanced across pigeons and across humans. As each stimulus always contained both spatial frequency and orientation information, some stimuli always required the same response, e.g., a horizontal pattern of low spatial frequency might always require a left response regardless of the current task. In addition to these congruent stimuli, responses to incongruent ones depended on the task at hand, for example, a horizontal stimulus with a high spatial frequency pattern might require a response to the left reward location on the orientation task but a response to the right location if the spatial frequency was to be judged.

For pigeons, pecking twice at the composite stimulus in the display center resulted in it being deleted from the center and simultaneously reappearing 200 pixels to the left and to the right of the display center as response keys. Pigeons made a final response by choosing the correct reward location (left or right) by pecking at the stimulus presented on that side. The two response keys were effective between 3 and 6 seconds after the onset of their presentation, after which a single peck at the correct key resulted in the activation of the corresponding food magazine for 2.5 seconds. During training only, if a pigeon developed a position bias, i.e., showed a strong tendency to peck one of the two response keys, responses to the more attended side were made ineffective for one to two (or more if necessary) seconds longer than to the less attended side. The release of the food magazine ended a trial. The inter-trial-interval to the next presentation of the observing key lasted between 15 and 30 seconds. Human participants were asked to mouseclick on the stimulus in the center of the screen, which led to
the appearance of two square, white response keys to the left and right side of the stimulus; the stimulus also remained on display. Participants responded to the stimulus by clicking on the response key that was associated with the present cue-stimulus combination. If the correct response key was chosen, the stimulus and response keys disappeared from the screen and the word "Correct" appeared in white letters next to a golden star for two seconds before the next trial began. If the wrong response key was clicked, the entire display was replaced by the phrase "WRONG!" in white letters. Participants were asked to respond as quickly as possible while making as few mistakes as possible.

Training Both pigeons and human participants received training on each task separately before attempting the taskswitching paradigm. The order in which the tasks were learned was counterbalanced across individuals of each species.

Pigeons received daily training sessions of 3 blocks of 24 trials each, showing each possible combination of the two cues of the task to be trained and the different variations of spatial frequency and orientation three times per block. The first block included a 25th trial (a repeat of the first trial of the session), as that first trial was not included in analyses, resulting in 73 trials per day in total. The order of cuestimuli combinations was randomized within blocks. Discrimination of the stimuli was considered successful if the pigeon responded correctly on at least $80 \%$ of trials within a daily training session, in at least three consecutive sessions. Pigeons thus received a minimum of 3 sessions, or 219 trials, on a task before starting training on the other task. The number of sessions on each task was gradually reduced until pigeons were able to switch between tasks from one day to the next and still perform at or above $80 \%$ correct responses in each session. For humans, training on each separate task was carried out in four blocks of trials. A block was considered successful if subjects reached the criterion of $80 \%$ or more correct responses in the previous trials that included each stimulus at least twice. Thus, the criterion was based on at least eight consecutive trials, two for each of the four different combinations of spatial frequency and orientation. The first training block of a task contained at least 32 trials, then, the second task was trained in at least 32 trials. After this, the first task was repeated for a minimum of another 16 trials until criterion was reached; finally, the second task was repeated for at least 16 trials until the participant reached criterion in this fourth and final training block.

Test Once each task was trained separately to success criterion, subjects entered the task-switching part of the experiment, in which task $A$ and task $B$ trials were intermingled. The task sequence was partially randomized to produce a switch trial in one third of the trials; for nonswitch trials, the two task cues alternated so that the same cue was never shown for two trials in a row. Pigeons received 20 sessions of 73 trials each, or 1460 trials in total; in each block, the four combinations of spatial frequency
and orientation were presented three times per task. Humans completed 24 blocks of 25 trials, a total of 600 trials, in the same manner as described above, with each of the combinations occurring twice on a task-repeat trial and once on a task-switch trial per two blocks. After completion of the task-switching procedure, we determined, via a questionnaire, which approach they used to solve the experiment and assessed their ability to describe the rules that defined a correct response. If a participant was able to correctly identify the contingencies between a task cue and certain stimulus characteristics, he or she was considered to have understood and successfully applied the underlying rule. If participants could not explain any relationship between stimuli, cues and the correct response, this was taken as an indication that they had not used task rules.

## Results

The only basis for comparing the two species was accuracy (errors) when choosing a response key, as it was not possible to obtain an accurate estimation of response latencies for pigeons, although we did record each subject's latency to peck/click on a response key. Restricting pigeons' time to respond would have required differential reinforcement of short response latencies, which could potentially have impaired learning of the cue-stimulusresponse contingencies. Thus, all results reported are for error rates when making a response. We ran four of the birds on ten more sessions with a strongly reduced inter-trial interval after they had completed the main study to assess whether allowing for unrestricted response times potentially decreased any effects, and this yielded similar results to those reported below. Nevertheless, the possibility remains that the particular timing requirements of the task we used may play an important role in producing our results.

For the human data, we calculated participants' error rate when choosing a response key as a percentage for each pair of consecutive blocks, i.e., for 48 trials (the first trial of each block was excluded from analysis, since it was neither a switch nor a repeat trial), resulting in 12 block pairs per participant. Pigeons' performance was calculated for each of the 20 sessions, excluding the first trial of each session.

The different training methods we employed for humans and pigeons resulted in substantially lower error rates for
pigeons. However, while it was necessary to train pigeons to produce error rates below $20 \%$ throughout, we were reluctant to give more training to humans as it would have increased the chances of humans inferring the task rules. As it was, nine of the 24 human participants were able to verbalize the rules for both tasks at the end of the experiment. A further eight reported having discovered one of the two rules or having made up their own solving strategies. Because of the ambiguity as to what mechanisms these participants relied on to solve the tasks, we did not include their data in any further analyses. The remaining seven participants stated that they were not aware of any relationships between the stimuli and the correct response, and it is these participants that most naturally allow comparison with performance by the pigeons on this task.

To investigate the extent to which both humans and pigeons were influenced by switch costs or congruency effects, we conducted a 3-way repeated measure ANOVA using Switch/Repeat Trials, Stimulus Congruency and Block Pair Sequence/Session as within-subject factors. Analyses were carried out separately for humans, according to the number of rules humans could name (No Rule and 2 Rules) and for pigeons. F and p values for the effects mentioned below are reported in Table 1. All results were subjected to Huynh-Feldt correction.

Humans received an average of 137 trials before entering the task-switching stage; pigeons entered the test phase after an average of 109 training sessions. Since the pigeons had received substantially longer training, we conducted all analyses on the the first half of the sessions as well as on the full data set, to rule out potential floor effects. Results were the same for both data sets as the pigeons did not significantly improve their performance over time; accordingly, the results reported are from the full data set of 20 sessions.

Performance of humans was influenced by whether they were able to verbalize the discrimination rules or not; error rates were significantly lower for people who were able to verbalize both tasks ( 2 Rules; $\mathrm{M}=12.1 \%, \mathrm{SD}=2.0$ ) than if No Rule ( $\mathrm{M}=36.1 \%, \mathrm{SD}=2.3$ ) had been inferred $(\mathrm{F}(1,14)=$ $62.20, \mathrm{p}<.001$ ). Pigeons' error rates were low ( $\mathrm{M}=9.1 \%$, $\mathrm{SD}=3.7$ ), due to the amount of they received.

Table 1: Overall F and p values for Switch/Repeat Trials, Stimulus Congruency, Block Pair Sequence/Sessions, and significant interactions between factors for humans depending on the number of rules they named and for pigeons. P values below .05 are marked in bold.

|  | Both Rules (N=9) |  |  | No Rules (N=7) |  | Pigeons (N=8) |  |
| :--- | :--- | :---: | :--- | :---: | :---: | :---: | :---: |
| Factor | $\mathrm{F}(\mathrm{df})$ | p | $\mathrm{F}(\mathrm{df})$ | p | $\mathrm{F}(\mathrm{df})$ | p |  |
| Block Pairs/Sessions | $8.92(11,88)$ | $<.001$ | $0.59(11,66)$ | .720 | $1.04(19,133)$ | .413 |  |
| Switch/Repeat | $57.29(1,8)$ | $<.001$ | $18.83(1,6)$ | $\mathbf{. 0 0 5}$ | $0.13(1,7)$ | .731 |  |
| Stimulus Congruency | $68.06(1,8)$ | $<.001$ | $19.10(1,6)$ | $\mathbf{. 0 0 5}$ | $71.03(1,7)$ | $<. \mathbf{0 0 1}$ |  |
|  |  |  |  |  |  |  |  |
| Session * Switch | $2.98(11,88)$ | $\mathbf{0 0 2}$ | $0.38(11,66)$ | .574 | $0.62(19,133)$ | .885 |  |
| Session * Congruency | $3.68(11,88)$ | $\mathbf{0 0 2}$ | $1.42(11,66)$ | .222 | $1.37(19,133)$ | .155 |  |
| Switch * Congruency | $14.18(1,8)$ | $\mathbf{0 0 5}$ | $10.69(1,6)$ | $\mathbf{. 0 1 7}$ | $0.55(1,7)$ | .484 |  |



Figure 1: Stimulus congruency effects (difference in error rates between incongruent and congruent stimuli) in trials in which the task repeats and those in which it switches from the previous trial, and across all trials ('Total').

The factor of Stimulus Congruency strongly influenced performance for all groups; human participants and pigeons made more errors when faced with incongruent stimuli than when dealing with congruent ones (Figure 1, 'Total').

Similarly, there was a highly significant effect of the factor Switch/Repeat for humans regardless of the number of rules verbalized: they performed less well on switch trials, which required executing the opposite task to the one on the previous trial, than on repeat trials (Figure 2, 'Total'). However, while the effect was present in both human groups, pigeons demonstrated a noticeable lack of switch costs (Figure 2, 'Total').

All human participants showed significantly higher switch costs on trials with incongruent stimuli than on those trials in which the stimulus was congruent, i.e. there was a significant interaction between the two factors for all three human groups (Figure 2).

The sequence of Block Pairs (or Sessions for pigeons) was reliable for 2-Rules users only, implying that these participants learned to make fewer mistakes as the experiment carried on, while No-Rule users and pigeons maintained their initial level of performance throughout. For those participants who were able to verbalize the two rules, both the effects of Switch/Repeat trials and Stimulus Congruency declined over the course of the experiment, i.e., this group experienced interaction effects of Block Pairs with the two other main factors.
Although not relevant for the species comparisons, it can be noted that, in their reaction times, human showed a similar pattern to what has previously been observed in humans using different learning approaches (Forrest et al., 2012); that is, those who inferred both rules suffered from switch costs much more ( $\mathrm{M}=149.70 \mathrm{~ms}$ ) than from congruency effects ( $\mathrm{M}=94.78 \mathrm{~ms}$ ), whereas humans who did not use any rules were largely affected by congruency effects ( $\mathrm{M}=50.34 \mathrm{~ms}$ ) but barely showed any costs in switching between tasks ( $\mathrm{M}=0.58 \mathrm{~ms}$ ).


Figure 2: Switch cost (difference in error rates between task switch trials and task repeat trials) for congruent and incongruent stimuli, and overall ('Total').

## Discussion

Forrest et al. (2012) showed that humans in the cue-stimulus-response (no rule) condition of their task-switching experiment expressed reduced switch costs and larger congruency effects relative to a Tasks group that were told both of the applicable rules at the start of the experiment. They offer this as a "signature" of associatively-based performance on this type of task. We are not in a position to make a direct comparison with their study, as we did not run an equivalent of their Tasks group. Our 2-Rules participants are perhaps an approximation to this group, but had to induce the rules, and were not instructed to apply them. However, these participants demonstrated significant switch costs and exhibited a congruency effect, similar to the effects usually found when humans are informed of the task rules before engaging in a task-switching paradigm. This group was the only one that significantly decreased their error rates over the course of the experiment; it can be assumed that this was due to participants "figuring out the tasks": during the first few blocks, performance essentially matched that of No-Rule users, but then it dramatically improved to a level similar to the performance of the pigeons, i.e., at error rates of $10 \%$ or less. A third of participants were unable to report any task rules; these might instead be classified as employing an associative approach to task switching. In addition to a generally high error rate, solving the tasks without any knowledge of their underlying rules had considerable impact on the magnitude of typical task-switching phenomena: while a stimulus's congruency only moderately affected performance in those who used both rules, it heavily influenced humans' ability to solve the tasks if they had been unaware of the rules. This comes to no surprise, as it will always be easier, especially in regard to associative learning, to learn the correct response to a given stimulus when it is the same in both tasks - that is, when that stimulus is congruent - than when it varies between tasks, as it does for incongruent stimuli.

Clearly, there is a significant congruency effect in the pigeon data and no cost of switching between tasks. Similar levels - a much bigger congruency effect than switch costs are also observed in the No-Rule humans and in Forrest et
al.'s (2012) study. The typical signature for a task-based approach in humans instructed to use tasks is the reverse, a larger switch cost and a smaller congruency effect. Further research will establish if this true of the tasks used here.

Although we can draw parallels between the performance of pigeons and No-Rule-using humans, there are some very clear discrepancies between the pigeon data and that of either of the human groups. It is especially apparent that in incongruent trials, there is some switch cost in each of the human groups, which was also observed in Forrest et al.'s (2012) results. Even the human participants who were not using any rules exhibited some switch cost for the incongruent stimuli, yet the pigeons show no discernible trace of any such effect but are able to "task switch". Why is this?

The most interesting possibility is that pigeons simply do not suffer from a switch cost in this paradigm. That is, when given a combination of component and biconditional discriminations, they do not exhibit any difficulty in switching from one hypothetical task to another, even in the case of the biconditional discrimination (i.e., the discrimination involving the incongruent stimuli). This result would imply that there is no switch cost in associatively-mediated task switching, and lead to the conclusion that the switch costs in all our human groups were due, in some sense, to contamination by rule use. This would fit rather well with theories that explain switch costs in terms of task-set reconfiguration (Monsell \& Mizon, 2006) but less well with theories that attempt to explain switch costs in associative terms (e.g., Logan \& Bundesen, 2003). A potential way of testing this assumption would be to compare the groups' performance when faced with novel stimuli in a generalization task.

Another possibility would be that pigeons do exhibit switch costs in this paradigm, but that we either lack the power to detect it, or there is another factor masking it. The former possibility cannot be ruled out given that the predicted effect would be small, in any case, but there is simply no evidence of any switch cost to suggest that it would be worth running many more pigeons in an attempt to increase the power. The latter possibility is, however, worth investigating, especially if switch costs are potentially only present for incongruent stimuli while there will be little or no switch costs on congruent trials. A close examination of the paradigm reveals the possibility of an unwanted interaction between the difference between switch and repeat trials and a preference for novelty (e.g. in matching to sample, see Wright \& Delius, 2005). Pigeons might preferentially respond to trials in which there is some change in stimulation (either in the form of a different stimulus or different response) compared to the previous trial, and avoid those in which both the stimulus and the response location are the same as in the preceding trial. The latter, for incongruent stimuli, is only possible on repeat trials, so, other things being equal, performance on those trials should then on average be worse than on switch trials. A disadvantage for repeat trials over switch trials for
incongruent stimuli could potentially cancel out any switch costs in those trials, which by definition compose a disadvantage for switch trials over repeat trials. Whether this is the case or not is a matter for further empirical investigation.

## References

Cardinal, R.N. \& Aitken, M.R.F. (2010). Whisker: a clientserver high-performance multimedia research control system. Behavior Research Methods, 42, 1059-1071.
Forrest, C.L.D., Elchlepp, H., Monsell, S. \& McLaren, I.P.L. (2012). Task switching without knowledge of the tasks. In N. Miyake, D. Peebles, \& R.P. Cooper (Eds.), Proceedings of the 34th Annual Conference of the Cognitive Science Society (pp. 1548-1553). Austin, TX: Cognitive Science Society.
Harris, J. A. \& Livesey, E.J. (2008). Comparing patterning and biconditional discriminations in humans. Journal of Experimental Psychology: Animal Behavior Processes, 34, 144-54.
Kleiner,M., Brainard, D. \& Pelli, D. (2007). What's new in Psychtoolbox-3? Perception, 36, ECVP Abstract Supplement.
Lea, S.E.G. \& Wills, A.J. (2008). Use of multiple dimensions in learned discriminations. Comparative Cognition and Behaviour Reviews, 3, 115-133.
Logan, G.D., \& Bundesen, C. (2003). Clever homunculus: Is there an endogenous act of control in the explicit taskcuing procedure? Journal of Experimental Psychology: Human Perception and Performance, 29, 575-599.
Mackintosh, N.J. (1988). Approaches to the study of animal intelligence. British Journal of Psychology, 79, 509-526.
McLaren, I.P.L., Green, R.E.A. \& Mackintosh, N.J. (1994) Animal learning and the explicit/implicit distinction. In N.C. Ellis (Ed.), Implicit and Explicit Learning of Languages. London: Academic Press.
Monsell, S. (2003). Task switching. TRENDS in Cognitive Sciences, 7, 134-140.
Monsell, S., \& Mizon, G.A. (2006). Can the task-cuing paradigm measure an endogenous task-set reconfiguration process? Journal of Experimental Psychology: Human Perception and Performance, 32, 493-516.
Monsell, S., Yeung, N. \& Azuma, R. (2000). Reconfiguration of task-set: Is it easier to switch to the weaker task? Psychological Research, 63, 250-264.
Stoet, G. \& Snyder, L.H. (2003). Executive control and task-switching in monkeys. Neuropsychologia, 41, 13571364.

Vandierendonck, A., Liefooghe, B. \& Verbruggen, F. (2010). Task switching: Interplay of reconfiguration and interference. Psychological Bulletin, 136, 601-626.
Wright, A.W. and Delius, J.D. (2005). Learning Processes in Matching and Oddity: The Oddity Preference Effect and Sample Reinforcement. Journal of Experimental Psychology: Animal Behavior Processes, 31, 425-32.

# Relations between Body Motion and Emotion: Analysis based on Laban Movement Analysis 

Junya Morita, Yukari Nagai, Tomoyuki Moritsu<br>j-morita@jaist.ac.jp, ynagai@jaist.ac.jp, moritsu.tom@gmail.com<br>School of Knowledge Science, Japan Advanced Institute of Science and Technology<br>1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan


#### Abstract

This study examined the relationship between human body movements and emotion based on Laban Movement Analysis (LMA). Ten participants participated in the experiment in which they stayed at a small resting room while hearing pleasant or unpleasant sounds. After the stay at the room, the participants rated their subjective emotional states. Participants' body movement were also recorded with four video cameras. The movement analysis based on LMA reveled significant differences in movement features between experimental conditions. In addition, significant correlations between movement features and subjective mood ratings were observed. These results suggest a strong relationship between human body movements and emotion.


Keywords: Emotion; Body motion; LMA

## Introduction

The English word "emotion" comes from the Latin word "emovere," meaning to "move out." This derivation suggests a close relationship between emotion and body movements. In fact, this relationship, which we will refer to as the motion-emotion relationship in this paper, has been repeatedly discussed (Damasio, 1994; Darwin, 1890; James, 1892). Researchers have attempted to reveal causal relationships between emotion and body movements or the evolutionary advantages of emotion. This work has also described how emotion is expressed in body movements. However quantitative relationships between body movements and emotion categories have yet to be fully clarified.

Experimental studies on the motion-emotion relationship have so far, been conducted in the field of emotion perception (Atkinson, Dittrich, Germmell, \& Young, 2004; Dittrich, Troscianko, Lea, \& Morgan, 1996; Field, Hampson, \& Rose, 2005; Kaiser \& Keller, 2011; Pollick, Paterson, Bruderlin, \& Sanford, 2001). For example, in Dittrich et al. (1996) presented participants with stimuli that expressed human body movements, and participants rated their impressions of those movements on basic emotional category scales. The results showed that participants could judge the emotional states even from biological motion produced by pointlight displays.

Pollick et al. (2001) also conducted experiments that followed the paradigm proposed by Dittrich et al. (1996). Unlike the other previous studies, Pollick and colleagues quantified features of body movements as a way to reveal the motionemotion relationship. The results showed that motions with strong velocity tended to be perceived as anger or happiness, while motions with weak velocity tended to be perceived as sadness or tired.

The movement stimuli used in emotion perception studies are usually created from performances of expert dancers or actors. On the basis of this method, several body movement theories have been proposed in the field of drama and dance.

Laban Movement Analysis (LMA) is one of the most famous theories of body expressions in dance (Laban, 1980). This theory assumes two basic opposing forms of body movement: fighting form and indulging form. Fighting form involves active, prominent, brisk body movements, while indulging form is unsteady weak body movements.

LMA assumes that such forms of body movement reflect subjective inner attitudes, refered to as efforts. The theory classifies effort into the following three axes:

- Weight effort that stands for the vigorousness of body movement.
- Space effort that stands for the degree of directional deflection.
- Time effort stands for the hurriedness in the changes of movement.

The terms fighting and indulging refer to opposing or enhancing the characteristics of a type of effort defined in the above three axes. Fighting form has strong weight, space, and time efforts. Indulging form has weak weight, space and time efforts.

LMA does not directly explain the motion-emotion relationship. Instead, this theory was proposed to describe body movements in dance. Laban (1980) did not mention any correspondence between the efforts and emotion categories. However, several researchers have applied LMA to emotion expressions programmed in robots (Hachimura, Takashina, \& Yoshimura, 2005; Masuda \& Kato, 2009; Nakata \& Mori, 2002). For example, Nakata and Mori (2002) defined the three effort axes based on Euclidean vector operations. Here, participants rated impressions of robot motions that manipulated the three efforts based on LMA. The results indicated correlations between emotion ratings and the efforts based on LMA.

Studies using robots are useful in being able to produce precise manipulations of body movement. Researchers can freely create robot's movements, and these movements are easily quantified. However, studies using robots have the same limitations as studies assessing emotion perception. Both fields have only examined intentionally expressed emotion. Ekman and Friesen (1975) pointed out that there are uncontrolled, involuntary, true emotion expressions, as well as
qualified, modulated, or false expressions. Past studies have examined Ekman's second type of emotion expressions. Few studies have challenged the first type of emotion expressions. For a deeper understanding of the motion-emotion relationship, the first type of emotion expressions needed to be examined.

The present study aimed to apply past findings of the motion-emotion relationship from uncontrolled experimental situations (Ekman's first type). To achieve this goal, we quantified involuntarily body movements using LMA. The motion-emotion relationship was explored by calculating correlations between quantitative features of body movements and subjective ratings of emotional states. In addition, this study explored how the environment influences on the motion-emotion relationship. It can be reasonably assumed that the motion-emotion relation will be affected by environment. We prepared two experimental conditions that were expected to arouse different emotional states.

## Methods

## Design

This study utilized audio stimuli that were assumed to arouse participants' emotion. A series of studies have previously examined environmental sounds related to pleasant or unpleasant emotion (Shimai, Tanaka, \& Terasaki, 1990). From this work, we set up an experiment to examine the motionemotion relationship using different environmental sound settings.

A between subjects design was used in the experiment. One group heard a pleasant environmental sound (pleasant sound group), and the other group heard an unpleasant environmental sound (unpleasant sound group).

## Participants

Twenty graduate students from Japan Advanced Institute of Science and Technology participated in the experiment; they were divided into two groups of 10 participants each; all participants were in their 20s and naive to the purpose of the study.

## Apparatus

We did not prepare any specific tasks for the participants because we wanted to examine emotion elicitation and body movement in uncontrolled situations. The participants were just asked to relax in a resting room while hearing the environmental sounds. We prepared the following resting room and environmental sounds:

## - Resting room

The resting room was designed like a Japanese teahouse ( $2 \mathrm{~m} \times 2 \mathrm{~m} \times 2 \mathrm{~m}$ ). The room had two tatami on the floor, and one small window on the wall. The participants' body movements were recorded with four video cameras affixed to the ceiling (Panasonic BB-HCM515).

- Environmental sound

The environmental sounds were prepared using free sound libraries on the web. The pleasant sound was selected from "pdsounds" (http://www.pdsounds.org/). The unpleasant sound was created at "sound 101" (http://www.sound101.org/). Ten sound files were arbitrarily downloaded from these two web sites. Three raters, who were naive to the purpose of the study, assessed the pleasantness of these sounds. The sound that had the highest rating was chosen as the pleasant sound, and the sound that had the lowest rating was selected as the unpleasant sound. The selected pleasant sound was the sound of a brook ( 17 seconds), while the unpleasant sound was composed of several noises, such as sirens, microphone feedback, and scratching noises ( 20 seconds). Each environmental sound was looped during the experiment. The participants heard the sound through wireless headphones (Sony DRBT50). The maximum output sound level was kept below 70 db for both groups.

## Procedure

The procedure for the experiment was composed of the following three steps:

## 1. Instructions

The participants were told that this experiment was conducted to evaluate the resting room with environmental sound. They were asked to rest while hearing the environmental sound. They were not told that their movements would be recorded while in the room.

## 2. Resting room

The participants entered the resting room alone, and took off their shoes. They put on a set of wireless headphones and an orange jump suit. The jump suit was used to analyze body movements with more ease. The participants could not bring any personal items into the room, including their mobile phone. After 30 minutes, the experimenter announced the end of the experiment.
3. Subjective emotional state ratings

Immediately after being in the resting room, the participants rated their emotional states using the POMS (Profile of Mood States) brief test (Pollock, Cho, Reker, \& Volavka, 1979). This test is usually used to assess transient and distinct mood states, which includes 30 questions classified into six factors: tension (anxiety), depression (dejection), anger (hostility), vigor (activity), fatigue (inertia), and confusion (bewilderment). This test outputs standardized scores with an average of 50 points.

The Ethics Committee of Japan Advanced Institute of Science and Technology approved the study.

## Analysis

We analyzed the video data by applying LMA. Although the LMA effort axes were quantified by Nakata and Mori (2002), their method targeted pre-programed movements with parameters such as joint torque or angles of movement. Since our study obtained data in uncontrolled situations, the method from this previous study could not be directly applied. To quantify the efforts of Weight, Space and Time, we used optical flows estimated by image processing.

Weight effort is defined as the following equation.

$$
\begin{equation*}
\text { Weight }=\sum_{n=1}^{t} \sum_{i=1}^{x} \sum_{j=1}^{y} \frac{\left\|v_{n i j}\right\|}{t \times x \times y} \tag{1}
\end{equation*}
$$

where $t, x$, and $y$ indicates the number of frames ( 25 fps ), the width of the frame ( 240 px ), and the height of the frame (320px), respectively. $\left\|v_{n j k}\right\|$ indicates the strength of the optical flow at pixel $i j$ of frame $n$. Weight indicates the time space average of the vector strength. If the participants moved actively, this index would increase.

The following equation defines Space effort.

$$
\begin{equation*}
\text { Space }=\sum_{n=1}^{t} \sum_{i=1}^{x} \sum_{j=1}^{y} \frac{\mu_{n} \cdot v_{n i j}}{t \times x \times y} \tag{2}
\end{equation*}
$$

where $\mu_{n}$ is a mean vector (by-center) of optical flows in frame $n$, which indicates overall direction of body movements in the frame. Space is calculated as the time space average of dot products between a mean vector and individual optical flows. This value would increase when optical flows in the frame had consistent direction. Conversely, the value would decrease if direction of optical flows diverged.

Based on the above two feature quantities, Time (W) and Time ( S ) are defined as hurriedness in the changes of movement.

$$
\begin{gather*}
\operatorname{Time}(W)=\sum_{n=2}^{t} \frac{\left|W_{e i g h t}(n)-W \operatorname{Weight}_{( }(n-1)\right|}{t}  \tag{3}\\
\operatorname{Time}(S)=\sum_{n=2}^{t} \frac{\mu_{n} \cdot \mu_{n-1}}{t} \tag{4}
\end{gather*}
$$

Time (W) and Time (S) represent the difference of Weight and Space between two continuous frames respectively. Time (W) represents the degree of changes of movement strength. Time (S) indicates the degree of changes of movement direction.

These feature quantities were averaged over the four video cameras. Optical flows were determined by Lucas and Kanade (1981)'s gradient method. The frames were preprocessed by gray-scale processing and background differencing technique. For all steps of the above analysis, we used OpenCV that is an open-source library for image processing.


Figure 1: Subjective rating of emotional states after staying in the resting room.


Figure 2: Changes in the features of body movements while in the resting room.

## Results

## Subjective ratings on emotional states

Figure 1 shows the average rating scores on the POMS after staying in the resting room. A 2 (group: pleasant sound vs. unpleasant sound) by 6 (emotion categories: tension, depression, anger, vigor, fatigue, confusion) mixed design Analysis of Variance (ANOVA) revealed a significant interaction between the environmental sounds and the emotion categories $(F(5,90)=3.80, p<.01)$. Simple main effects of the environmental sounds were significant for tension, depression, fatigue, and confusion (tension: $F(1,18)=8.97, p<.01$, depression: $F(1,18)=7.89, p<.05$, anger: $F(1,18)=$ 2.64, n.s., vigor: $F(1,18)=0.01$, n.s., fatigue: $F(1,18)=$ $11.12, p<.01$, confusion: $F(1,18)=5.74, p<.05)$. These results indicate changes in emotional states influenced by the environmental sounds.

Table 1: Correlations between the subjective ratings of emotional states and features of body movements (Overall).

|  | Weight | Space | Time(W) | Time(S) |
| :--- | :---: | :---: | :---: | :---: |
| tension | $0.632^{*}$ | $0.543^{*}$ | $0.655^{*}$ | $0.691^{*}$ |
| depression | 0.436 | 0.275 | 0.405 | $0.560^{*}$ |
| anger | $0.589^{*}$ | 0.372 | $0.560^{*}$ | $0.610^{*}$ |
| vigor | -0.072 | 0.056 | -0.036 | -0.023 |
| fatigue | $0.601^{*}$ | $0.473^{*}$ | $0.601^{*}$ | $0.520^{*}$ |
| confusion | $0.473^{*}$ | 0.327 | 0.435 | $0.468^{*}$ |
| Note $* p<.05$ |  |  |  |  |

## Feature quantities of body movement

Figure 2 shows the feature quantities of body while the participants were in the resting room. Four separete 10 (time: $3-30$ ) by 2 (environmental sounds: pleasant vs. unpleasant) mixed design ANOVAs were conducted for the feature quantities. The analysis did not reveal significant interactions between time and the environmental sounds (Weight: $F(9,162)=0.38$,n.s., Space: $F(9,162)=0.59$, n.s., Time $(\mathrm{W}): F(9,162)=0.49$,n.s., Time(S): $F(9,162)=0.85$, n.s.). We obtained significant main effects of the environmental sounds (Weight: $F(1,18)=12.77, p<.01$. Space: $F(1,18)=11.64, p<.01$. Time(W): $F(1,18)=8.54, p<.01$. Time(S): $F(1,18)=$ 23.92, $p<.01$ ), and significant main effects of time for the four feature quantities (Weight: $F(9,162)=12.17, p<.01$. Space: $F(9,162)=6.22, p<.01$, Time(W): $F(9,162)=$ $12.89, p<.01$. Time(S): $F(9,162)=15.30, p<.01)$.

The main effects of the environmental sounds indicate that the body movements in the unpleasant sound condition were active, directed, and rapid. This suggests that the unpleasant sound changed body movements toward fighting form. On the other hand, the pleasant sound changed the body movement toward indulging form.

The main effects of time indicate that the participants made more movements immediately after entering the resting room. We confirmed qualitative differences in behaviors between the early and latter stages of the experiment. During the first few minutes, several participants observed interiors of the resting room. Next, they tended to lie down on the tatami mats until the end of the experiment.

## Correlations between emotional states and body movement

Table 1 shows correlation coefficients between the rated emotional states and the feature quantities of body movements. In this analysis, we used the feature quantities in the latter 12 minutes to examine body movements that strongly connected to the emotional states.

From the table, we can observe some significant correlations between body motion and the emotion ratings. The emotion categories other than vigor appear to be related to some of the body movements.

However, we cannot exclude the possibility that these correlations were caused by features of the environmental sound

Table 2: Correlations between the subjective ratings of emotional states and features of body movements (pleasant sound condition).

|  | Weight | Space | Time(W) | Time(S) |
| :--- | :---: | :---: | :---: | :---: |
| tension | -0.136 | 0.113 | -0.014 | -0.181 |
| depression | -0.028 | 0.129 | 0.267 | -0.101 |
| anger | -0.531 | -0.405 | -0.355 | $-0.657 *$ |
| vigor | 0.338 | 0.182 | 0.187 | -0.005 |
| fatigue | -0.610 | -0.493 | -0.436 | -0.591 |
| confusion | 0.324 | 0.413 | 0.577 | -0.259 |
| Note. ${ }^{*} p<.05$. |  |  |  |  |

Table 3: Correlations between the subjective ratings of emotional states and features of body movements (unpleasant sound condition).

|  | Weight | Space | Time(W) | Time(S) |
| :--- | :---: | :---: | :---: | :---: |
| tension | -0.616 | 0.389 | 0.580 | $0.798^{*}$ |
| depression | 0.291 | -0.050 | 0.140 | 0.540 |
| anger | $0.731^{*}$ | 0.422 | $0.661^{*}$ | $0.892^{*}$ |
| vigor | -0.232 | 0.039 | -0.123 | -0.020 |
| fatigue | $0.725^{*}$ | 0.451 | 0.621 | $0.719^{*}$ |
| confusion | 0.343 | -0.001 | 0.158 | 0.510 |
| Note. ${ }^{*} p<.05$. |  |  |  |  |

because the participants in the two groups heard different sounds. The body movements might have been influenced by acoustic features such as rhythm or tempo. To exclude this possibility and obtain stronger evidence of the motionemotion relationship, we calculated correlations disaggregated by the environmental sound.

Table 2 and 3 show correlation coefficients between the emotional states and the feature quantities of body movements in the pleasant and the unpleasant sound conditions, respectively. In the pleasant sound condition, anger was negatively correlated with time (S). In the unpleasant sound condition, anger was positively correlated with Weight, Time (W), and Time ( S ). There was a positive correlation between tension and Time ( S ) in the unpleasant sound condition. Fatigue in the unpleasant sound condition was also correlated with Weight and Time (S).

From these results, we confirmed a motion-emotion relationship in a situation where the same environmental sound was presented. However, there were large differences in the pattern of the motion-emotion relationship between the two groups. In the following section, we will try to explain each of the observed correlations.

## Discussion

This study explored the motion-emotion relationship in uncontrolled experimental situations. We prepared two conditions that differed in the environmental sounds presented. The results confirmed that (1) changes in emotional states as a function of environmental sound occured, (2) differences in body movements between different environmental sound
conditions emerged, and (3) correlations between emotional states and body movements in each environmental sound condition were observed. This following section provides an interpretation of these results and discusses implications and limitations of the current findings.

## Changes in emotional states by environmental sound

This study confirmed that environmental sounds changed emotional states as assessed by the POMS test. Although this study does not focus on the relationship between environmental sounds and emotion, we consider this result useful for future experimental studies on emotional arousal. Presentation of environmental sounds is non-invasive and an easy way to manipulate emotional states.

However, Figure 1 also shows limitations of our emotional manipulation. The influence of the environmental sound was limited to specific emotion categories. There were no differences in anger and vigor between the two groups. It is difficult to interpret why these differences occured. To provide a more accurate emotion manipulation, we need to conduct further experiments that explore the mechanisms underlying the relationships between emotion categories in this context.

## Changes in body movements by environmental sound

In addition to the emotional states, the environmental sound influenced body movements. The feature quantities of body movements in the unpleasant sound condition were higher than those in the pleasant sound condition. The participants in the unpleasant sound condition exhibited active, directed, and rapid movement. Such body movements are consistent with fighting form of LMA. In contrast, body movements observed in the pleasant condition are consistent with indulging form. Considering the results of the POMS test, we can assume emotion categories like tension, depression, fatigue and confusion were related to fighting form.

However, this conclusion is not sufficient given that the environmental sounds in the two groups had different physical features. The difference between the two groups might be explained by certain mechanisms, such as sensory-motor coordination (Repp, 2005). This limitation can be addressed by examining correlations disaggregated by the environmental sounds.

## Relationships between emotional states and body movement

Our study observed significant correlations between emotional states and body movements. This was especially the case when we consider that the correlations disaggregated by the environmental sounds are important. We provide an interpretation for the observed correlations in the following section.

Expressions of anger in response to the unpleasant sound In the unpleasant sound condition, anger was correlated with all of the features except Space. From this result, we can con-
sider that anger is expressed as fighting form in LMA. This interpretation is consistent with past studies of emotion perception. As noted in the Introduction, Pollick et al. (2001) observed a correlation between anger perception and velocity of body movements. Masuda and Kato (2009), who analyzed robot motion based on LMA, also observed a strong correlation between anger perception and Time effort. The present study succeeded in extending these findings of emotion perception to an uncontrolled situation.

Expressions of tension in response to the unpleasant sound In the unpleasant sound condition, tension was correlated with Time (S), which quantifies the degree of hurriedness in movement changes. This result is consistent with the classic finding indicating that tension is often accompanied by muscle overactivity (Sainsbury \& Gibson, 1954). Darwin (1890) also classified this action as an involuntary reflex movement. The present study described these relationships using the LMA framework.

Expressions of fatigue in response to the unpleasant sound Weight and Time ( S ) were correlated with fatigue in the unpleasant sound condition. POMS defines fatigue as decrease in motivation or energy. This result is inconsistent with our intuition that fatigue decreases body movements.

However, this result does not necessarily indicate a direct causal relationship between fatigue and these movement features. One possible explanation for this result might be that other emotion categories that covaried with fatigue caused an increase in the movement features.

In the unpleasant sound condition, fatigue was positively correlated with tension ( $r=0.738, p<.05$ ) and anger ( $r=$ $0.698, p<.05$ ), which were also correlated with the movement features (Table 3). From these correlations, we can speculate that fatigue in the unpleasant sound condition was caused by strong and rapid movements accompanying these emotional states. However, our results cannot directly confirm this hypothesis. This possibility needs further study using more sophisticated methods such as path analysis.
Expressions of anger in response to the pleasant sound Compared to the unpleasant sound condition, the pleasant sound condition did not show many correlations between body movements and emotional states. In the pleasant sound condition, anger was negatively correlated with Time (S). This result is incontrast to the positive correlation observed in the unpleasant sound condition.

As in the above case, this contrasting finding might be explained by examining emotion categories that were negatively correlated with anger. In Pollick et al. (2001)'s experiment, both anger and happiness were expressed as active motion. Considering their finding, our negative correlation between anger and Time ( S ) is not too suprising. The increase in happiness might accompany a decrease in anger. However, we cannot confirm with the present results. Future studies that employ emotional ratings including happiness, should be conducted.

## Summary

This study attempted to reveal motion-emotion relationships in uncontrolled situations. As a result, we observed that both motion and emotion were changed by environmental sounds. More importantly, we obtained correlations between emotion and body movements in each environmental sound condition. From these results, we confirmed the motion-emotion relationship from uncontrolled, involuntary emotional expressions.

Our study is characterized by a method of quantifying body movement features based on LMA. This quantitative method is useful for connecting motion and emotion in uncontrolled experimental situations. We consider that our approach can contribute not only to scientific studies of emotion but also to engineering applications. Automatic emotion assessment is applicable to various pragmatic situations, including dance, education and marketing. LMA is a useful tool for systematically describing body movements. Past studies using LMA have not included involuntary motion/emotion in uncontrolled situations. Our study extends the application fields of these past studies.

However, our results also revealed complexities in the motion-emotion relationship. As shown in Table 2 and 3, emotion is expressed differently in different situations. Additionally, there is no one-to-one correspondence between body movements and emotional states. A single emotion category can be connected to several body movements and vice versa. Furthermore, as speculated earlier, emotion categories are intricately intertwined. To reveal the deeper mechanisms of the motion-emotion relationship, studies with larger sample sizes will be needed.

## References

Atkinson, A., Dittrich, W., Germmell, A., \& Young, A. (2004). Emotion perception from dynamic and static body expressions in point-light and full-light displays. Perception, 33, 717-746.
Damasio, A. (1994). Descartes' error: Emotion, rationality and the human brain. New York: Putnam.
Darwin, C. (1890). The Expression of the Emotions in Man and Animals (London ed.). HarperCollins.
Dittrich, W. H., Troscianko, T., Lea, S. E. G., \& Morgan, D. (1996). Perception of emotion from dynamic point-light displays represented in dance. Perception, 25(6), 727-738.
Ekman, P., \& Friesen, W. V. (1975). Unmasking the face: A guide to emotions from facial cues. Prentice-Hall. New Jersey.
Field, D., Hampson, S., \& Rose, D. (2005). The perception of emotion from body movement in point-light displays of interpersonal dialogue. Perception, 34, 1171-1180.
Hachimura, K., Takashina, K., \& Yoshimura, M. (2005). Analysis and Evaluation of Dancing Movement Based on LMA. In Ieee international workshop on robot and human interactive communication, 2005. (pp. 294-299). IEEE.

James, W. (1892). Psyhology (Brief Course). University of Notre Dame Press.
Kaiser, R., \& Keller, P. E. (2011, July). Music's impact on the visual perception of emotional dyadic interactions. Musicae Scientiae, 15(2), 270-287.
Laban, R. (1980). The Mastery of Movement (4th ed.; L. Ullmann, Ed.). Hampshire, UK: Dance Books.
Lucas, B. D., \& Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In Proceedings of imaging understanding workshop (pp. 121130).

Masuda, M., \& Kato, S. (2009). Emotion detection from body motion of human form robot based on laban movement analysis. In Proceedings of the 12th international conference on principles of practice in multi-agent systems (pp. 322-334). Nagoya.
Nakata, T., \& Mori, T. (2002). Analysis of impression of robot bodily expression. Journal of Robotics and Mechatronics, 14(1), 27-36.
Pollick, F., Paterson, H., Bruderlin, A., \& Sanford, A. J. (2001). Perceiving affect from arm movement. Cognition, 82, 51-61.
Pollock, V., Cho, D. W., Reker, D., \& Volavka, J. (1979). Profile of Mood States: The factors and their physiological correlates. Journal of Nervous and Mental Disease, 167, 612-614.
Repp, B. (2005). Sensorimotor synchronization: A review of the tapping literature. Psychonomic Bulletin Review, 12(6), 969-992.
Sainsbury, P., \& Gibson, J. G. (1954). Symptoms of anxiety and tension and the accompanying physiological changes in the muscular system. Journal of Neurology, 17, 216224.

Shimai, S., Tanaka, M., \& Terasaki, M. (1990). Plesantness and unplesantness of environmental sounds. Perception and Motor Skills, 71, 960-962.

# The missing baselines in arguments for the optimal efficiency of languages 

## Fermín MOSCOSO DEL PRADO (fmoscoso@linguistics.ucsb.edu)

Department of Linguistics<br>University of California, Santa Barbara

Santa Barbara, CA 93106-3100 USA


#### Abstract

I argue that linear correlations between log word frequency, and lexical measures, cannot be taken as evidence for a "Principle of Minimum Effort". The Principle of Maximum Entropy indicates that such relations are in fact the ones most probable to be found. For such claims, one needs to compare the correlations with adequate baselines reflecting what would be expected in a purely random system. I then introduce a way of computing such baselines, and use it to show that the correlations found in a corpus are actually weaker than what one would expect to find by chance. Therefore, if an argument were to be made based on them, it would paradoxically be that language is worse for communication than what one would expect to find in a random system. More appropriately however, what these results reflect is that such correlations are not the best places to look for linguistic optimality.


Keywords: Corpus Study; Lexical Ambiguity; Principle of Maximum Entropy; Zipf's Law of Abbreviation

## Introduction

Arguments about language being optimal for communication have a long tradition within the cognitive sciences, dating at least as far back as Zipf (1935). Zipf observed that, across many texts, there is an inverse correlation between a word's frequency of occurrence and its length in characters, which is now referred to as Zipf's Law of Abbreviation (ZLA). This observation led him to his "Principle of Least Effort" (Zipf, 1949): Humans prefer shorter words to refer to frequent concepts, so that the overall length of utterances will be minimized, and so will the effort required to produce them. In this form, from the speaker's (or writer's) point of view, the optimality of human language would be measured by the amount of effort required by a speaker to produce an utterance. Zipf also realized that, from the comprehender's perspective, optimality would not be so much concerned with the length of an utterance as it would with the ease with which it can be unambiguously decoded. Jointly considering both the perspective of the speaker and that of the comprehender, the structure of language would be subject to a trade-off between utterance length and degree of ambiguity. Zipf was somewhat vague with respect to how such trade-off could be measured, but his general idea is considered valid ever since.

As compelling as Zipf's arguments seem, very early on, researchers in Information Theory and Psychology noticed that they may not be as informative as Zipf thought. Both Mandelbrot (1953) and Miller (1957) realized that the negative correlation between a word's frequency of occurrence and its length in characters (i.e., ZLA) would also arise in randomly generated texts that lack any linguistic structure or communicative value whatsoever; what Mandelbrot dubbed a "typing monkeys" process, and Miller -somewhat less
graphically- called an "intermittent silence" process. The validity and importance of Zipf's original observations on the distribution of word frequencies and word lengths is beyond doubt, as is evidenced for instance in a whole family of power-law distributions and phenomena across many unrelated fields of science being currently named in Zipf's honor (e.g., Zipf's Laws, Zipfian distributions, Zipf-Mandelbrot distribution). However, Zipf's interpretation that such properties reflect the optimization of human language structure is disconfirmed by the fact that those very same properties are also found in systems that are not the result of any optimization process. The properties are therefore not informative about the optimality of the process that generated them. This highlights a problem that is exhibitted by many claims on language properties that reflect some form of optimization: The lack of a non-optimal baseline against which to test whether such inferences are perhaps non sequiturs.

Let us consider a non-linguistic example. Suppose I put forward a theory on the processes governing the outcome obtained when throwing two particular dice. The dice themselves would be beyond my possible observations (e.g., inside a black box), but I would have access to the sum of their outcome. My theory could state that the dice are loaded so that they strongly favor a non-extreme (or optimal) outcome of three or four dots. In order to test my theory, I would collect data from many dice throws (with access only to their summed values). After obtaining a few thousand throws, if I found that the average value of the sum is seven (with some preset degree of precision), which is fully consistent with my theory. As it happens, however, seven would also be the most likely value of the sum, even if the dice were not loaded. Therefore, I could not take the evidence from the average value as support for my theory, as it would also be consistent with the a priori more likely theory that the dice are not loaded. As we will see below, one can objectively say that the unloaded theory is more probable a priori using mere probabilistic arguments (the Principle of Maximum Entropy).

In the example above, the prediction used to test the hypothesized property holds trivially for the most probable outcome. One can of course design situations in which the seven sum property does not hold (e.g., by loading the sixes on dice). Still, even if it is possible to artificially design such a scenario, it is still the case that the most probable outcome, whether or not any optimization is at work, is that the property will hold (i.e., the dice will sum up to seven). However, even if the property were less evident than the one used in this example, testing it also on a few non-optimal baselines would enable us to see that such property does not signal the
presence of optimization.
This discussion is motivated by the recent publication of several papers making claims on the optimality of human language which suffer from the same lack of baseline problem. In what follows, I begin by summarizing four of these recent arguments for language optimality. I then introduce the Principle of Maximum Entropy, and use it to show analytically that the findings presented as evidence for communicative optimality turn out to be trivial predictions that will also be observed in the most probable non-optimized baselines. In the data section, I analyze a corpus of English to assess the strength of the effects used to argue for optimality. The results show that these effects are in fact significantly weaker than what one would expect to find by mere chance. In other words, if one took such correlations as evidence for optimization, one would have to conclude that human languages are actually less optimized than one would expect by chance. I conclude with a discussion of how Information-Theory can be used to make predictions on the optimality of human languages that do indeed survive the non-optimal baselines tests.

## Some Information Theoretical Arguments for Communicative Efficiency

In a recent study, Piantadosi, Tily, and Gibson (2012) extend ZLA to the domain of lexical ambiguity. Following Zipf, they argue that short words require less effort to be produced than longer words would. Therefore, by a similar principle of economy, it would be beneficial to encode as many meanings as possible using the shorter words, and then use the redundancy present in the context to disambiguate them. They support this claim by showing that, in corpora of three languages, there is indeed a negative correlation between word ambiguity and word length when other factors (e.g., frequency) are considered.

A second prediction of Piantadosi et al. (2012) considers the fact that more frequent words have a more accessible lexical representation, as is evidenced by the fact that they elicit shorter reaction times and lower error rates in a broad range of lexical processing experiments (e.g., Oldfield \& Wingfield, 1965). That a word is easy to access makes it a desirable candidate to carry many meanings (or be associated to many uninflected word lemmas) in a system that is optimized to make the production and comprehension of words as effortless as possible. Therefore, they predict a positive correlation between a word's frequency and the number of distinct meanings (or lemmas) associated with it. Their analyses of several corpora indeed find this correlation.

As convincing as these arguments might seem, I will argue below that the findings are but trivial consequences of ZLA, and do not provide support the communicative hypothesis that is put forward. I will further discuss, ZLA is itself a trivial property of most symbolic sequences, irrespective of whether they are optimized.

Piantadosi et al. (2012) also argue that, if one computes the probability of a word according to a triphone (i.e., phoneme
trigram) model, those words with the highest probability correspond to those that provide a more prototypical example of the phonotactics of a language. Those words that conform better to the phonotactic constraints of the language will be easier to pronounce and recognize. Following the "least effort" argument, they predict that words with high phonotactic predictability should be associated with more meanings (or word lemmas) than words with lower phonotactic probability.

## The Principle of Maximum Entropy

Before making any claims that a particular distribution of linguistic probabilities (of words, words lengths, degrees of ambiguity, etc.) constitutes evidence for language being "optimized for human communication", one should check what kind of such distributions one would expect to observe by mere chance, irrespective of the presence any hypothetical optimization process. The relevant properties of the distribution (e.g., ambiguous words being more frequent, etc.) should be found to be significantly more (or less) marked in actual linguistic data than one would expect them to be.

This raises the problem of how to determine, among the infinite possible discrete probability distributions that words could have, which ones are the most probable a priori. The Principle of Maximum Entropy (PME; Jaynes, 1957a, 1957b) states that, among all probability distributions satisfying a set of constraints, the most probable one will be the one that has the highest entropy. The entropy (Shannon, 1948) of a probability distribution defined over a discrete set of words $W=\left\{w_{1}, w_{2}, w_{3}, \ldots\right\}$ is given by

$$
H(W)=-\sum_{w \in W} \mathrm{P}(W=w) \log \mathrm{P}(W=w)
$$

where $P(W=w)$ denotes the probability of encountering word $w$ in a corpus of text (i.e., its relative frequency of occurrence). In what follows, I will use the abbreviated notation $\mathrm{p}_{i}=\mathrm{P}\left(W=w_{i}\right)$. The most probable assignment of values for the $\mathrm{p}_{i}$ is the one leading to the highest value of $H(W)$, with the obvious constraint that the values of the $\mathrm{p}_{i}$ must all sum to one, so that they form an actual probability distribution. ${ }^{1}$

If no additional constraints were present (i.e., any assignment of probabilities could be considered), then, for a finite set $N$ probabilities, the maximum entropy would be the uniform distribution with $p_{i}=1 / N$. Of course, when one considers the probabilities of words, not all probability distributions are valid assignments. Rather, these distributions need to satisfy some basic constraints. These specific constraints can be requirements such as the existence of a mean word length or an average degree of ambiguity. Constraints of this type can be expressed as values of the means of some given functions. For instance, one such function can be the length of a word (measured in either characters, phonemes, or syllables), which maps words into natural numbers ( $\ell: W \mapsto \mathbb{N}$ ),

[^157]and another function can the degree to which words are ambiguous, which maps words into the non-negative real numbers $\left(\mathcal{A}: W \mapsto \mathbb{R}^{+}\right)$. The constraints would then be expressed as the existance of a mean word length $\left(\langle\ell(w)\rangle_{W}=L\right)$ and a mean degree of ambiguity $\left(\langle\mathcal{A}(w)\rangle_{W}=A\right)$.

The most probable distribution that satisfies the constraints would then be the solution to the maximization problem:

$$
\arg \max -\sum_{w_{i} \in W} \mathrm{p}_{i} \log \mathrm{p}_{i}
$$

subject to

$$
\left\{\begin{array}{cll} 
& \sum_{w_{i} \in W} \mathrm{p}_{i} & =1  \tag{1}\\
\langle\ell(w)\rangle_{W}= & \sum_{w \in W} \mathrm{p}_{i} \ell\left(w_{i}\right) & =L \\
\langle\mathcal{A}(w)\rangle_{W}= & \sum_{w \in W} \mathrm{p}_{i} \mathcal{A}\left(w_{i}\right) & =A
\end{array}\right.
$$

Notice that I have added one constraint to indicate that the resulting probability distribution must be normalized. The solution to such a problem is found analitically using the method of Laplace multipliers. It must have the form of a Boltzmann canonical distribution, ${ }^{2}$

$$
\begin{equation*}
\mathrm{p}_{i}=\mathrm{e}^{\lambda_{0}+\lambda_{1} \ell\left(w_{i}\right)+\lambda_{2} \mathcal{A}\left(w_{i}\right)} \tag{2}
\end{equation*}
$$

where the parameters $\lambda_{0}, \lambda_{1}$, and $\lambda_{2}$ are Laplace multipliers whose value is uniquely determined by the individual values of the word lengths $\ell\left(w_{i}\right)$, ambiguities $\mathcal{A}\left(w_{i}\right)$ as well as their average values $L$ and $A$.

## Implications of the PME

Taking logs on both sides of Eq. 2 reveals that a priori assuming the existence of a mean word length $(L)$ and an average degree of ambiguity $(A)-$ the most probable relation between our variables of interest is

$$
\begin{equation*}
\log \mathrm{p}_{i}=\lambda_{0}+\lambda_{1} \ell\left(w_{i}\right)+\lambda_{2} \mathcal{A}\left(w_{i}\right) \tag{3}
\end{equation*}
$$

This equation already makes important predictions. We should expect that -everything else being equal- the log probability of a word (i.e., its $\log$ frequency) should be linearly related to both its length and to its degree of ambiguity. No assumptions about language being optimized for communication are necessary to make this prediction, it just happens to be to most probable type of relation. The signs and values of the Laplace multipliers $\lambda_{1}$ and $\lambda_{2}$ will determine the strength and direction of the correlations. They therefore provide baselines for any effects whose presence is argued to reflect a form of efficiency or optimality. Without any need for efficiency, we should expect to find correlations with the strengths given by $\lambda_{1}$ and $\lambda_{2}$.

A negative value of $\lambda_{1}$ would indicate that ZLA is in fact the most likely relation that one should expect to find between word frequency and word length. Therefore, in order to claim that ZLA provides evidence for communicative efficiency, one should observe that the relation between

[^158]log word frequency and word length is more negative than $\lambda_{1}$. This finding complements the previous arguments of Mandelbrot (1953), Miller (1957), or Ferrer i Cancho and Moscoso del Prado (2011) that random processes also exhibit ZLA. It provides a baseline to assess whether the ZLA observed in a real corpus is stronger than what one would have expected by mere chance.

Similarly, $\lambda_{2}$ indexes the relation between log word frequency and degree of ambiguity. Piantadosi et al. (2012)'s finding of a positive correlation between a word's ambiguity and its frequency of occurrence (i.e., frequent words are more ambiguous) can only be interpreted as evidence for optimality if the regression coefficient found for the degree of ambiguity is more positive than $\lambda_{2}$.

A simple rewrite of Eq. 3 results in

$$
\begin{equation*}
\mathcal{A}\left(w_{i}\right)=\frac{\lambda_{0}-\log \mathrm{p}_{i}+\lambda_{1} \ell\left(w_{i}\right)}{-\lambda_{2}} . \tag{4}
\end{equation*}
$$

This indicates that, when word frequency is kept constant or controlled for, one should also expect to find a linear relationship between a word's length and its degree of ambiguity (with a regression coefficient $-\lambda_{1} / \lambda_{2}$ ), as was documented by Piantadosi et al. (2012). As before, in order to accept Piantadosi and colleagues' interpretation that their finding is indicative of some form of communicative efficiency, one needs to ensure that such relation is less marked than $-\lambda_{1} / \lambda_{2}$.

Although, for reasons of space I do not detail it here, it is easy to show that a word's frequency of occurrence in a corpus is expected to be directly proportional to that word's phonotactic probability as computed from an n-phone (e.g., diphone, triphone, ...) model whose parameters were computed on that same corpus. If we denote a word's triphonebased probability as $T_{i}$, we can therefore say that $k \mathrm{p}_{i} \approx T_{i}$ for some value $0<k \leq 1$ constant across all words. If we substitute on Eq. 3, we obtain

$$
\begin{equation*}
\log T_{i}-\log k \approx \lambda_{0}+\lambda_{1} \ell\left(w_{i}\right)+\lambda_{2} \mathcal{A}\left(w_{i}\right) . \tag{5}
\end{equation*}
$$

Dividing both sides of Eq. 5 by $\ell\left(w_{i}\right)$ with a simple rearrangement results in

$$
\begin{equation*}
\frac{\log T_{i}}{\ell\left(w_{i}\right)} \approx \lambda_{1}+\frac{\log k+\lambda_{0}+\lambda_{2} \mathcal{A}\left(w_{i}\right)}{\ell\left(w_{i}\right)} \tag{6}
\end{equation*}
$$

Therefore, when the degree of ambiguity is controlled for, a word's log triphone (or diphone, ...) probability normalized by its length, is expected to be non-linearly related to word length itself (i.e., linearly related to the reciprocal of word length). One would therefore expect to find the nonlinearities that Piantadosi et al. (2012) found. Hence, such non-linear relation -by itself- cannot be interpreted to be the product of an optimization process, contrary to what was argued by Piantadosi and his colleagues.

In summary, I have shown that the linear relationships between $\log$ word frequency, word length, and word ambiguity -by themselves- do not warrant an interpretation that language is optimized for communicative efficient. In the following section, I will estimate the values of the parameters
$\lambda_{0}, \lambda_{1}$, and $\lambda_{2}$ from corpus data, and I will assess whether or not they provide evidence for (or against) any sort of optimization.

## Corpus Study

In order to test whether the values of the Laplace multipliers $\left(\lambda_{1}, \lambda_{2}\right)$ provide support for the hypothesis that language is optimized for communicative efficiency, I selected the 29,025 most frequent English content words (adjectives, adverbs, nouns, and verbs) present in WordNet (Miller, 1995). ${ }^{3}$ The selection was done using their surface word spoken frequency according the CELEX lexical database (Baayen, Piepenbrock, \& Gulikers, 1995), from where the corresponding word length in phonemes was obtained. ${ }^{4}$ For each word I counted its number of distinct senses (i.e., 'synsets') listed in WordNet (Miller, 1995). The log of the number of senses was taken as the measure of a word's ambiguity $\left(\mathcal{A}\left(w_{i}\right)=\log N_{i}\right.$, where $N_{i}$ is the number of distinct senses of $\left.w_{i}\right) .{ }^{5}$

I normalized the word frequency counts into relative frequencies adding up to one, and estimated the mean word length as the weighted average

$$
L=\left\langle\ell\left(w_{i}\right)\right\rangle_{W}=\sum_{w_{i} \in W} \mathrm{p}_{i} \ell\left(w_{i}\right)
$$

with $\mathrm{p}_{i}$ being the corpus based relative frequency of $w_{i}$. Similarly, the average degree of ambiguity was estimated as

$$
A=\left\langle\mathcal{A}\left(w_{i}\right)\right\rangle_{W}=\sum_{w_{i} \in W} \mathrm{p}_{i} \mathcal{A}\left(w_{i}\right)=\sum_{w_{i} \in W} \mathrm{p}_{i} \log N_{i} .
$$

Using these estimates of $L$ and $A$, and the individual values of $\ell\left(w_{i}\right)$ and $\mathcal{A}\left(w_{i}\right)$, the values of the multipliers $\lambda_{0}, \lambda_{1}$, and $\lambda_{2}$ were estimated by nonlinear maximization (using a Newtontype algorithm) of

$$
H(W)=-\lambda_{0}-\lambda_{1} L-\lambda_{2} A
$$

subject to the constraints of Eq. 1. In order to keep the results comparable to those of Piantadosi and colleagues, additional parameters were added to separate the different grammatical categories.

The values of the Laplace multipliers were estimated to be $\lambda_{0}=-10.37, \lambda_{1}=-.14$, and $\lambda_{2}=1.07$. As I discussed in the previous section, that $\lambda_{1}<0$ indicates that ZLA (a negative correlation between word length and word frequency) is the most likely relationship between these two variables,

[^159]whether or not any optimization is at work. Similarly, $\lambda_{2}>0$ implies that we should expect a priori a positive correlation (all other factors equal) between a word's frequency and its degree of ambiguity. As suspected -by itself- the positive relation between frequency and ambiguity does not warrant the interpretation of optimization, it is rather what one should expect, contrary to Piantadosi et al. (2012).


Figure 1: Non-linear effects of word length (left panels) and degree of ambiguity (right panels) on the lengthnormalized $\log$ triphone probability (top panels) and the length-normalized $\log$ a priori frequency (bottom panels).

To see the actual values of these correlations in the corpus itself, I performed a linear regression predicting a word's log probability from its length and its degree of ambiguity (once more, I also included additional parameters to separate the grammatical categories). As Piantadosi et al. (2012), I found significant effects of both word length and degree of ambiguity (both with $\mathrm{p}<.0001$ ). Interestingly, the estimated coefficient for word length $(\beta=-.05 \pm .003)$ constitutes a much weaker effect than the one we would have expected by chance ( $\lambda_{1}=-.14$ ). This means that the supposed optimization from ZLA is actually weaker than what one should have expected, not supporting any optimization. In a similar vein, the coefficient estimated for the effect of ambiguity $(\beta=.84 \pm .01)$ is also a weaker effect than the chance level $\left(\lambda_{2}=1.07\right)$. Again, it seems that the negative relation between frequency and ambiguity that was claimed by Piantadosi and his colleagues to reflect optimization, is actually significantly weaker than the expected chance level. This illustrates the importance of having meaningful baselines before interpreting lexical statistics.

As discussed in the previous section the ratio $-\lambda_{1} / \lambda_{2}=.13$
indexes the strength of the correlation between word length and word ambiguity (after controlling for word frequency) that we should expect by chance. Indeed, we should therefore expect by chance a positive correlation between a word's length and its degree of ambiguity, irrespective of any optimization process. The strength of this relationship in the data is given by the ratio between the corresponding regression coefficients $\beta[$ length $] / \beta[$ ambiguity $]=.05$, which is once more weaker than what we would have expected by chance.

In order to assess the non-linear effects of a word's phonotactic probability, I trained a triphone model using the Brown corpus of English (Kucera \& Francis, 1967) after transcribing all the words into the phonemic forms using the CMU Pronouncing Dictionary. ${ }^{6}$ I used this triphone model to estimate the phonotactic probability of each word $\left(T_{i}\right)$. For each the length-normalized $\log$ trigram probabilities $\left(\log T_{i} / \ell\left(w_{i}\right)\right)$ and the length-normalized $\log$ a priori probabilities estimated using the $\lambda\left(\log \mathrm{p}_{i} / \ell\left(w_{i}\right)\right)$, I fitted a generalized additive model with a linear predictor for log word frequency (estimated for the corpus or a priori) and penalized spline smoothers terms for word length and degree of ambiguity. Fig. 1 plots the estimated curves. As predicted, the shape and strength of the non-linear relations is basically the shame for the actual triphone probabilities (top-panels), as it is for the word probabilities that would be predicted a priori (bottom panels). Once more, the shape of the relation does not warrant the interpretation of optimization.

The values of the Laplace multipliers can be used with Eq. 3 to compute what should be the a priori distribution of words, considering only their length and degrees of ambiguity. The log relative frequencies predicted by the method exhibit a remarkably strong correlation with the relative $\log$ frequencies actually observed ( $r=.45, t[29004]=86.41, \mathrm{p}<$ .0001 ). This suggest that the frequency distribution of words is not that different from the distribution one would expect to find by chance. In other words, it does not appear to reflect much specific optimization.

It could be argued that, by using the actual values of word length and word ambiguity as estimated from the corpus, I am covertly exploiting the possible correlations that are already present in their distributions, even before considering word frequency. To control for this possible confound, I used a Jackknife technique. New values of word lengths and ambiguity were randomly assigned to each word by randomly sampling (with replacement) from the original distributions. In this way, one obtains distributions of word length and ambiguity which are fully uncorrelated, but retain their original distributional shapes. I repeated this process two hundred times, re-estimating the values of the Laplace multipliers in each case. Fig. 2 compares the original $\lambda$ estimates (blue dots), the distribution of $\lambda$ values obtained in the resampling (box and whiskers plots), and the $\beta$ parameters of the regressions on the actual corpus (red crosses). As it can be seen, even after fully decoupling word length and ambiguity, the

[^160]expected effects are much stronger than those observed in the corpus. In short, the effects that allegedly reflect optimization of language for communicative efficiency are actually much weaker than we should have expected them to be.

## Conclusion

These results do not question either ZLA or Piantadosi et al. (2012)'s effects, rather the effects themselves are indeed replicated here. What is questioned is the interpretation of such effects as evidence for optimization. I have shown that by themselves- the linear relationships between log word frequency, word length, and degree of ambiguity, do not warrant the interpretation that language is optimized for communicative efficiency. The shape and direction of the effects reported in Piantadosi et al. (2012), as well as ZLA, are precisely what one would expect to obtain by chance (i.e., by a random assignment of probabilities). Furthermore, if anything, I find that the effects present in a corpus are actually weaker than what one would obtain by chance. Following the classical "Principle of Least Effort" interpretation is therefore not warranted by this type of correlations.

These findings complement previous studies showing that mere random typing processes (e.g., Mandelbrot, 1953; Miller, 1957; Ferrer i Cancho \& Moscoso del Prado, 2011) also exhibit ZLA. More generally, the most likely observation is that a word's log probability of occurrence is linearly related with any property of the word for which a mean value exists. Therefore, in order to claim that observations of this kind are indicative of any type of process (optimization or otherwise) giving rise to the word frequency measures are not warranted, unless the effects are explicitly found to be significantly the stronger than the effects one should find by chance. In other words, such effects are meaningless unless compared to random baselines.

Notice that the same conclusions would be reached if, instead of unigram word frequencies, I had considered the $a$ priori distribution of word bigrams or trigrams. It would merely be a question of applying the PME to the whole matrix of $n$-grams and we would expect to obtain the same type of linear relations between $\log n$-gram frequencies and lexical measures. Thus, similar arguments as those expressed in Piantadosi, Tily, and Gibson (2011), would suffer from exactly the same problems I discussed above.

By this I do not intend to claim that language is not optimized for human communication. Rather the opposite, I am strongly convinced that this is indeed the case. However, pure correlational values between lexical measures (or for that matter $n$-gram measures) are not sufficient evidence to support such claims. Of course, there is a certain commonsensical aspect to the claim that human language is optimal for communication: It would be difficult to find a cognitive scientist who disagrees with such a statement. However, claims on optimization of human language should rely on specific mechanisms by which the optimization takes place, together with explicit mathematical (e.g., variational) de-


Figure 2: Estimated relations between log word frequency, word length, and degree of ambiguity. The red crosses indicate the magnitude of the effects observed in the corpora. The blue dots plot the magnitude that we should expect to observe $a$ priori. The box and whisker plots plot the distribution of the a priori predictions once word length and ambiguity have been decoupled using Jackknife. The leftmost and middle panels respectively plot the effects of word length and ambiguity of log word frequency. The rightmost panel plots the direct relation between word length and word ambiguity. Notice the different vertical scales between the panels.
scriptions of how such an optimization proceeds, as exemplified by some recent studies (e.g., Ferrer i Cancho \& Solé, 2003; Ferrer i Cancho \& Díaz-Guilera, 2007).

## References

Baayen, R. H., Piepenbrock, R., \& Gulikers, L. (1995). The CELEX lexical database (CD-ROM). University of Pennsylvania, Philadelphia, PA: Linguistic Data Consortium.
Ferrer i Cancho, R., \& Díaz-Guilera, A. (2007). The global minima of the communicative energy of natural communication systems. Journal of Statistical Mechanics: Theory and Experiment, 2007, P06009.
Ferrer i Cancho, R., \& Moscoso del Prado, F. (2011). Information content versus word length in random typing. Journal of Statistical Mechanics: Theory and Experiment, 2011, L12002.
Ferrer i Cancho, R., \& Solé, R. V. (2003). Least effort and the origins of scaling in human language. Proceedings of the National Academy of Sciences of the United States of America, 100, 788-791.
Jaynes, E. T. (1957a). Information theory and statistical mechanics. Physical Review, 106, 620-630.
Jaynes, E. T. (1957b). Information theory and statistical mechanics II. Physical Review, 108, 171-190.
Jaynes, E. T. (2003). Probability Theory: The Logic of Science. Cambridge, UK: Cambridge University Press.
Kucera, H., \& Francis, W. N. (1967). Computational analysis of present-day American English. Providence, RI: Brown University Press.
Mandelbrot, B. B. (1953). An information theory of the statistical structure of language. In W. Jackson (Ed.), Com-
munication Theory (p. 503512). New York, NY: Academic Press.
Miller, G. A. (1957). Some effects of intermittent silence. The American Journal of Psychology, 70, 311-314.
Miller, G. A. (1995). WordNet: A lexical database for English. Communications of the ACM, 38, 39-41.
Moscoso del Prado, F. (2007). Co-occurrence and the effect of inflectional paradigms. Lingue e Linguaggio, 2, 247263.

Moscoso del Prado, F. (2011). Macroscopic thermodynamics of reaction times. Journal of Mathematical Psychology, 55, 302-319.
Oldfield, R. C., \& Wingfield, A. (1965). Response latencies in naming objects. Quarterly Journal of Experimental Psychology, 17, 273-281.
Piantadosi, S. T., Tily, H. J., \& Gibson, E. (2011). Word lengths are optimized for efficient communication. Proceedings of the National Academy of Sciences of the USA, 108.

Piantadosi, S. T., Tily, H. J., \& Gibson, E. (2012). The communicative function of ambiguity in language. Cognition, 122, 280-291.
Shannon, C. E. (1948). A mathematical theory of communication. The Bell System Technical Journal, 27, 379-423, 623-656.
Zipf, G. K. (1935). The Psychobiology of Language; an Introduction to Dynamic Philology. Boston, MA: HoughtonMifflin.
Zipf, G. K. (1949). Human Behavior and the Principle of Least Effort. Reading, MA: Addison-Wesley.

# Social Roles and Category Use: A Study of Creativity Assessment 

Jennifer S. Mueller (jmueller@sandiego.edu)<br>University of San Diego, 5998 Alcalá Park<br>San Diego, CA 92110 USA<br>Jeffrey Loewenstein (jloew@illinois.edu)<br>University of Illinois, Urbana-Champaign, 1206 South Sixth Street MC-707<br>Champaign, IL 61820 USA<br>Shimul Melwani (shimul_melwani@kenan-flagler.unc.edu)<br>University of North Carolina, Chapel Hill, 4725 McColl CB 3490<br>Chapel Hill, NC 27599 USA


#### Abstract

We show that social roles alter creativity assessments. Specifically, the two main roles in the innovation process generator roles for producing new ideas and implementer roles for selecting ideas to pursue - invoke different lay theories about what is creative. Study 1 showed that implementers rated a low novelty version of an idea as more creative than a high novelty version, but generators did the opposite. Study 2 showed that generators rated a low usefulness idea as more creative than a high usefulness idea, but implementers did the opposite. Thus, complementary roles prompted competing perspectives. These findings underscore a new challenge for the social distribution of knowledge-intensive work.


Keywords: Social Roles; creativity; categories; lay theories.

## Introduction

Editor's response to Sylvia Plath: There certainly isn't enough genuine talent for us to take notice.

Editor's response to Rudyard Kipling: I'm sorry Mr. Kipling, but you just don't know how to use the English language.

Many creative ideas are rejected and not necessarily more kindly than Plath's and Kipling's were. Thus, it is not enough to produce creative ideas. For cultural and scientific advancement, others need to recognize that the ideas are creative. One longstanding view is de gustibus non est disputandum-assessments are idiosyncratic. In contrast, current creativity theory and research claims that assessments are guided by domain knowledge. People within a community develop lay theories surrounding the category of creativity-causal and relational knowledge about what counts as creative and what is a more and less central member of the category. With expertise in the area (Amabile, 1982; Csikszentmihalyi, 1988), or even just moderate exposure to the area (Hennessey, Amabile, \& Mueller, 2010; Sawyer, 2012), individuals appear to converge with others in their assessments of creativity. However, given how complex the causal and relational
knowledge is that underpins judgments about creativity, we suggest that individuals' assessments of creativity are guided by more than just their domain knowledge. We suggest that they are also guided by their social roles.

We explore the possibility that the editors failed to recognize Plath and Kipling's creative ideas not because of lack of knowledge or idiosyncratic taste but because of their roles as editors. Specifically, we examine two key roles studied by organizational psychologists examining organizational innovation (Elsbach \& Kramer, 2003; Klein \& Knight, 2005; Klein \& Sorra, 1996): implementers, such as a book editor, and idea generators, such as a book author. We propose that adopting an implementer role leads to a different view of what is creative than adopting a generator role. The argument we make is parallel to one made about roles and person perception: different roles can produce different expectancies (Biddle, 1986), which then lead individuals in those roles to form different assessments of the same focal person (e.g., Winquist, Mohr, \& Kenny, 1998). Accordingly, it is not just that expectancies can highlight some aspects of the causal and relational structure underpinning a complex category and so alter judgments about that category, but also that those expectancies are systematically tied to particular social roles.

Social roles could shape assessments of creativity by shifting the lay or implicit theories (Paletz \& Peng, 2008; Sternberg, 1985) people use to evaluate ideas for creativity. This would resolve a puzzle in the creativity literature, which provides evidence of multiple, potentially conflicting lay theories about creativity. While there is widespread agreement that creative ideas combine novelty and usefulness (George, 2007), one lay theory is that novelty is the dominant characteristic in creativity assessments (Amabile, Barsade, Mueller, \& Staw, 2005), whereas another lay theory highlights that usefulness is the essential component of creativity assessments (Cooper, 2006). Which concern, novelty or usefulness, is deemed most causally central might be critical, as there is now evidence that people's assessments of novelty and usefulness are negatively related (Rietzschel, Nijstad, \& Stroebe, 2009). One reason why is a lay theory that highly novel ideas are not very useful, as they are likely to fail to solve problems
on time and within budget (Elsbach et al., 2003), and to fail in the marketplace (Fleming, 2001), rendering them less creative (Mueller, Melwani, \& Goncalo, 2012). Another reason why is a lay theory within scientific communities that high levels of usefulness can indicate that an idea is not very novel, as usefulness indicates taking on a smaller challenge and making a smaller change from current practice (Drazin, Glynn, \& Kazanjian, 1999). We suggest that people's social roles guide which lay theories they use, and so which causal factor underlying their model of creativity is central and shaping their creativity assessments.

We focus on two social roles that are fundamental to the social division of labor in the innovation process, generator and implementer roles (Elsbach et al., 2003). The innovation literature notes that these roles are complimentary, and their coordination is key to the process of innovation (Krishnan \& Ulrich, 2001): generators create new ideas, products and processes that implementers then assess, select and pursue. For example, scientists generate articles that editors vet, entrepreneurs generate business ideas that venture capitalists fund, and researchers generate product ideas that managers implement.

The two roles generate different expectancies. Generator roles include expectations around generating new ideas and overcoming challenges to solve problems in novel ways (Drazin et al., 1999). Hence, generator roles may activate lay theories about novelty being key to creativity and about highly useful ideas being less creative due to less opportunity for overcoming novel challenges. Implementer roles include expectations around maximizing efficiency by meeting timelines as well as budgetary and resource constraints (Drazin et al., 1999). Accordingly, implementer roles may activate lay theories about creative ideas being distinguished by usefulness and about highly novel ideas being less creative due to challenges of implementation.

If these predictions hold, then it will provide support for the importance of social roles and expectancies in assessing creativity. Managers may want creativity, but as implementers, they may adopt perspectives that lead them to reject the ideas that creators find most compelling. Thus, our account provides an explanation for the phenomenon of managers saying they want creativity but nonetheless end up rejecting creative ideas, a phenomenon that is widely noted in the popular press (e.g., Bussey, 2012; Hindo, 2007) and in the innovation research literature (DeFillippi, Grabher, \& Jones, 2007; Smith \& Lewis, 2011; Staw, 1995). The broader theoretical implication is that two complementary roles, such as implementer and generator roles in the innovation process, may bring with them complementary knowledge, but they may also bring with them competing causal models of the same categories that may thwart their ability to coordinate, communicate, and perform together.

## Experiment 1

This study examined how individuals in generator and implementer roles assessed a high and a low novelty idea.

We expected generators to assess the high novelty idea as more creative than the low novelty idea, because of their lay theory that novelty is the distinguishing characteristic of creative ideas. The key prediction though is that we expected implementers to rate the high novelty idea as less creative than the low novelty idea, because of a lay theory that highly novel ideas are less useful and so less creative. Critically, we examine these predictions about role effects by ensuring there were no systematic differences in domain knowledge (an effect that may otherwise accompany roles) by randomly assigning people to roles (as classically done in Anderson \& Pichert, 1978, among other work).

## Method

Participants and Design We recruited 176 people from Amazon's Mechanical Turk ( $62 \%$ male, $M=28.41$ years, $S D=9.43)$. Participants averaged $8.09(S D=8.94)$ years of work experience. Participants were randomly assigned to one of four conditions of a 2 (role: generator, implementer) X 2 (idea novelty: high, low) between-subjects design. Each cell contained more than 34 cases.

Procedure and Materials In Part one, participants were assigned to either a generator or an implementer role at a "large innovative product development firm" that "develops high performance outdoor gear." Generators were responsible for "generating new ideas, brainstorming technologies, and developing products and processes." Implementers were responsible for "cost savings, profitability, decreased time to market, meeting deadlines and product functionality." These descriptions were from prior research outlining generator and implementer role expectations (Drazin et al., 1999). Participants then wrote about the "important things that help you perform this role." In a pilot study $(\mathrm{N}=152)$, three coders (average coder-pair agreement was $94 \%$ ) rated whether participants described resources for novelty (e.g., "good team to help bounce ideas off of," "inspiring workspace") and for usefulness (e.g., "efficient staff," "computer with accounting programs"). Implementers $(96 \%)$ mentioned usefulness more than generators ( $36 \%, \chi^{2}(1)=61.46, p<.01$ ). Generators ( $80 \%$ ) mentioned novelty more than implementers ( $13 \%, \chi^{2}$ $(1)=70.01, p<.01)$. Thus, this manipulation led participants to adopt the intended roles and associated concerns.

In Part two, participants rated an idea for a "waterproof fleece," which "uses a soft, breathable and waterproof fabric using bio-mimicry to replicate the properties of a leaf in the Amazon rain forest that repels water yet is also very soft and pliable to the touch." The high novelty idea was described as a "completely new technology not currently available in the marketplace," while the low novelty idea was described as "an existing technology currently available in the marketplace." In a pilot study $(\mathrm{N}=54)$, participants (not put in any role) rated the high or low novelty idea as: new and original (assessing novelty, $\alpha=.86$ ), useful and valuable (assessing usefulness, $\alpha=.73$ ), and creative and
innovative (assessing creativity, $\alpha=.83$ ). Participants rated the low novelty idea $(\mathrm{M}=4.70)$ as less novel than the high novelty idea $(M=5.83, t(53)=3.53, p<.01)$, but comparably useful $(M=6.00)$ to the high novelty idea ( $\mathrm{M}=$ $5.89, t(53)=-.45, \mathrm{~ns})$. In addition, participants rated the low novelty idea ( $\mathrm{M}=5.22$ ) as less creative than the high novelty idea $(\mathrm{M}=5.70, t(53)=2.68, p<.05)$. Thus, the level of novelty was noticeable and produced a shift in perceived creativity. Participants in the main study rated the ideas using the same creativity, usefulness and novelty scales (all alphas above .70).

## Results

A multivariate ANOVA identified interactions between role (generator or implementer) and idea type (high or low novelty) when predicting creativity ratings $(F(1,172)=$ $\left.38.38, p<.01, \eta^{2}=.15\right)$, novelty ratings $(F(1,172)=9.00$, $\left.p<.05, \eta^{2}=.03\right)$ and usefulness ratings $(F(1,172)=$ 7.75, $p<.05, \eta^{2}=.05$ ). Planned comparisons revealed a crossover interaction such that generators rated the high novelty idea as significantly more creative $(M=6.04)$ than implementers $\left(M=5.43, t(86)=3.48, p<.01, \eta^{2}=.19\right.$; Figure 1). Generators rated the low novelty idea as significantly less creative $(M=4.62)$ than implementers ( $M$ $\left.=5.91, t(86)=-4.40, p<.01, \eta^{2}=.02\right)$. The critical finding was that implementers saw the low novelty idea as more creative than the high novelty idea $(t(71)=-2.06, p<$ $.05, \eta^{2}=.14$ ), whereas generators saw the low novelty idea as less creative than the high novelty idea $(t(101)=6.08, p$ $<.01, \eta^{2}=.04$ ).

Regarding novelty, generators saw the low novelty idea as less novel $(M=4.57)$ than implementers $(M=5.30, t(86)=$ $\left.-2.31, p<.05, \eta^{2}=.04\right)$, but generators $(M=5.47)$ and implementers $(M=5.29)$ had comparable ratings of the high novelty idea $\left(t(86)=.81, \mathrm{p}=\mathrm{ns}\right.$; see Figure $\left.2, \eta^{2}=.003\right)$. Generators rated the high novelty idea as more novel than the low novelty idea $\left(t(101)=3.48, p<.01, \eta^{2}=.07\right)$, whereas implementers did not show a detectable difference $\left(t(71)=-.06, n s, \eta^{2}=.00\right)$.

Regarding usefulness, implementers saw the high novelty idea as less useful $(M=5.36)$ than generators $(M=5.79$, $t(86)=1.99, p<.05$; see Figure 3, $\left.\eta^{2}=.03\right)$. Implementers also saw the low novelty idea as more useful $(M=5.95)$ than generators $(M=5.53, t(73)=-.2 .14, p<$ $.05, \eta^{2}=.02$ ). While generators saw no difference in the extent to which high $(M=5.79)$ and low $(M=5.53)$ novelty ideas were useful $\left(t(101)=1.49, \mathrm{p}=\mathrm{ns}, \eta^{2}=.01\right)$, implementers viewed the low novelty idea as more useful than the high novelty idea $\left(t(71)=-2.41, p<.05, \eta^{2}=\right.$ .04).

To further examine the relationship between roles, novelty and usefulness, and assessments of creativity, we conducted two parallel mediation analyses, one for each role, in which both novelty and usefulness were entered as possible mediators of creativity assessments (using the approach in Preacher, Rucker \& Hayes, 2007). We found that for generators, there was an indirect effect of high and
low novelty ideas on creativity assessments through novelty (mean effect estimate $=.57, \mathrm{SE}=.205 ; 95 \%$ CI 1.025 to .225 , i.e., does not include 0 ), but there was no indirect effect through usefulness (mean effect estimate $=.05, \mathrm{SE}=$ $.04 ; 95 \%$ CI -.0061 to .166 , i.e., includes 0). For implementers, we found the opposite pattern. There was evidence of an indirect effect of high and low novelty ideas on creativity assessments through usefulness (mean effect estimate $=.098, \mathrm{SE}=.064,95 \% \mathrm{CI} .0024$ to .2598 ), but not through novelty (mean effect estimate $=.0084, \mathrm{SE}=.145$; $95 \%$ CI -.287 to .289 ). Accordingly, generators' ratings of creativity seemed driven by novelty and implementers' ratings of creativity seemed driven by usefulness.


Figure 1. Means and $95 \%$ confidence intervals of creativity ratings by role (implementer, generator) and idea type (high novelty, low novelty), Experiment 1.

## Discussion

Adopting generator and implementer roles can lead people to form different assessments of creativity. Generators perceived a high novelty idea as more novel and creative but no more useful than a low novelty idea. This is consistent with the generator role evoking a lay theory that emphasizes novelty when assessing creativity. However, separating an effect of role-based expectancies from task demands driven by role instructions (we told generators to focus on novelty, and so they did) is challenging. We will present a better test of the generator role in Study 2.

The more striking pattern came from implementers, who assessed a high novelty idea as less useful and creative, but not more novel, than a low novelty idea. These data do not suffer from the same concern as the generator role data. Implementers were told to focus on usefulness, but they were not told to ignore novelty, and the ideas that they rated provided no direct information about usefulness. Thus, it is noteworthy that the implementers evaluated the high novelty idea as less useful than the low novelty idea, because it suggests that they were employing the lay theory that highly novel ideas are likely untested and risky and so lower in usefulness and lower in creativity.

## Experiment 2

Study 2 tested high and low usefulness ideas. The key prediction is that generators should assess a high usefulness idea as less novel and creative than a low usefulness idea, due to a lay theory that an idea high in usefulness indicates less novelty and so less creativity. Our account also predicts that implementers should assess the high usefulness idea as more useful and more creative than the low usefulness idea.

## Method

Participants and Design We recruited 161 participants from Amazon's Mechanical Turk ( $64 \%$ male, $M=30.42$ years, $S D=12.17$ ). Participants had an average of 9.75 (SD $=10.62$ ) years of work experience. Participants were randomly assigned to one of four conditions of a 2 (role: generator, implementer) X 2 (idea usefulness: high, low) between-subjects design. Each cell contained more than 29 cases.
Procedure and Materials The only difference with Study 1 was the idea being rated. We used the high novelty idea from Study 1, so everyone rated an idea that was explicitly marked as being highly novel. We added that the idea was "cheap and easy to make" (in the high usefulness condition) or "costly and difficult to make" (in the low usefulness condition).

## Results

A multivariate ANOVA identified interactions between role (generator or implementer) and idea type (high or low usefulness) when predicting creativity ratings $(F(1,157)=$ 20.06, $p<.01, \eta^{2}=.11$ ) and novelty ratings $(F(1,157)=$ $8.51, p<.01, \eta^{2}=.05$ ), but only a marginally significant trend for usefulness ratings $\left(F(1,157)=3.10, \mathrm{p}=.08, \eta^{2}=\right.$ .01 ). Planned comparisons revealed a crossover interaction for creativity ratings (Figure 2). Generators rated the high usefulness idea as less creative $(M=5.14)$ than implementers $\left(\mathrm{M}=6.07, t(76)=-3.95, p<.01, \eta^{2}=.10\right.$; see Figure 4). Generators also rated the low usefulness idea as more creative $(M=5.88)$ than implementers $(M=5.46$, $\left.t(81)=2.19, p<.05, \eta^{2}=.03\right)$. Implementers saw the high usefulness idea as more creative than the low usefulness idea $\left(t(88)=3.36, p<.05, \eta^{2}=.06\right)$, whereas generators recognized the low usefulness idea as more creative than the high novelty idea ( $t(69)=-2.96, p<.05, \eta^{2}=.06$ ).

Regarding novelty and usefulness, generators saw the high usefulness idea as less novel $(\mathrm{M}=4.97)$ than implementers $\left(M=5.81, t(76)=-3.11, p<.05, \eta^{2}=.06\right)$, but generators $(M=5.61)$ and implementers $(M=5.41)$ gave comparable novelty ratings to the low usefulness idea $\left(t(81)=.83, \mathrm{p}=\mathrm{ns}, \eta^{2}=.004\right.$; see Figure 5). Generators rated the low usefulness idea as more novel than the high usefulness idea $\left(t(69)=-2.33, p<.05, \eta^{2}=.04\right)$, whereas implementers saw no difference $\left(t(88)=1.72, n s, \eta^{2}=.02\right)$. Also as expected, implementers viewed the high usefulness idea $(M=6.04)$ as more useful than the low usefulness idea $\left(\mathrm{M}=5.41, t(88)=2.63, p<.05, \eta^{2}=.04\right.$; see Figure 6). Generators did not rate the high $(\mathrm{M}=5.52)$ and low $(\mathrm{M}=$
5.51) usefulness ideas reliably differently in usefulness $\left(t(69)=.01, \mathrm{p}=\mathrm{ns}, \eta^{2}=.00\right)$.

For generators, we found evidence of an indirect effect of high and low usefulness ideas on creativity assessments through novelty (mean effect estimate $=.35, \mathrm{SE}=.176$; $95 \%$ CI .06 to .76 ), but there was no indirect effect through usefulness (mean effect estimate $=-.002, \mathrm{SE}=.09 ; 95 \% \mathrm{CI}$ -.129 to .218 ). For implementers, there was evidence of an indirect effect of high and low usefulness ideas on creativity assessments through usefulness (mean effect estimate $=.12$, $\mathrm{SE}=.063,95 \% \mathrm{CI} .028$ to .282 ), but not through novelty (mean effect estimate $=-.193, \mathrm{SE}=.111 ; 95 \% \mathrm{CI}-.423$ to .009 ). Once again, generators' ratings of creativity seemed linked to novelty and implementers' ratings of creativity seemed linked to usefulness.


Figure 2. Means and 95\% confidence intervals of creativity ratings by role (implementer, generator) and idea type (high usefulness, low usefulness), Experiment 2.

## Discussion

Study 2 found the complementary pattern to Study 1. Unsurprisingly, adopting the implementer role led people to rate the high usefulness idea as more useful and more creative than the low usefulness idea, although separating out role-based expectancies from task demand based on the role instructions is challenging. What is more clearly compelling though is that adopting the generator role led people to rate the high usefulness idea as less novel and less creative than the low usefulness idea. This occurred despite both ideas being described as completely new, which was the primary concern of the generator role. That high usefulness meant lower novelty for generators is consistent with our proposal that generator roles evoke a lay theory that high usefulness indicates a lesser challenge and deviation in practice and so a lower degree of creativity.

## General Discussion

The social roles used to distribute the process of innovation appear to lead to different assessments of creativity, the very issue on which people in those roles need to coordinate. The
generator role seemed to invoke lay theories that novelty is key to an idea being creative, and that highly useful ideas lack novelty. The implementer role seemed to invoke lay theories that usefulness is key to an idea being creative, and that highly novel ideas lack creativity because they are less useful.

A strength of the approach in these studies was that we ensured that there were no systematic knowledge differences by randomly assigning roles, as in practice role differences are likely confounded with knowledge differences. Knowledge differences could also, in addition to role-driven expectancies or other sets of goals, guide perceptions of novelty and usefulness, and alter creativity assessments. In additional exploratory data analyses, we also examined the possibility that work experience could have influenced responses. However, we found no signs of effects of years of work experience, nor did we find any signs of effects of whether or not participants had prior work experience in generator roles (about $30 \%$ did) or implementer roles (about 20\% did). Thus, we have evidence that assigning individuals to roles led them to adopt the goals and perspective of those roles, and that individuals with and without actual work experience in the roles produced comparable assessments.

One theoretical possibility highlighted by these studies is that at least for complex categories, such as creativity, meanings may vary systematically in multiple ways. There are already good reasons to believe that category meanings are not fixed for speakers of the natural language but rather vary across cultural communities (Clark, 1996; Keller \& Loewenstein, 2011). If the studies here generalize, then there may be further social fractionation in category meanings driven by roles. Of interest, just as individuals can code switch from one cultural vocabulary to another, they can also change roles. Thus, individuals' understandings and use of categories might shift systematically as they adopt different roles and identify with different communities. There are not just context effects but social context effects that draw on histories of experience and social interaction. Just because individuals generate their own understandings of categories does not imply that their understandings are constant, internally consistent, or driven by one goal or causal logic.

The effects of social roles on creativity also emphasizes the need for a more comprehensive theoretical account of social context on creativity assessments. This would go alongside work on the effects of social context on creative production (Amabile, 1983; Kim, Vincent, \& Goncalo, 2012). Future work could examine whether situational factors apart from roles also guide creativity assessments. For example, situational uncertainty (Whitson \& Galinsky, 2008) might activate the lay theory that highly novel ideas are not creative because they have uncertain use (cf., Mueller et al., 2012), and so guide creativity assessments.

The generator-implementer difference in creativity assessments that we found suggests a practical problem, because effective coordination hinges on mutual
understanding (Okhuysen \& Bechky, 2009). If implementers and generators do not agree about which ideas are creative, this should lead to conflict, frustration and rejection. Apparently, the division of labor in the innovation process brings with it a division of cognitive labor that is not just about who knows what (Keil, Stein, Webb, Billings, \& Rozenblit, 2010), but is also about perspectives and perhaps attitudes. Indeed, editors may select articles they view as "creative" but that researchers view as "same old, same old," whereas editors view that researchers often pursue "pie-in-the-sky" ideas without grounding them in existing methods. Governments may fund research projects that scientists think perpetuate existing paradigms rather than testing new ones, while grant decision-makers might view many scientists pursuing ideas with little practical value to society. Managers may view that designers generating new products focus on extremely cutting edge ideas that are too costly and expensive to produce at a profit, while the designers view that managers implement old and tired ideas to "make a buck." The result may well be a stubborn coordination challenge on the core task of generating and implementing creative ideas. Or, perhaps at different points in the innovation process, generators and implementers might consider adopting the perspective of the alternative role. Because the true paradox of this paper is that, ironically, both roles may be right.

## Acknowledgments

We thank Matt Cronin for thoughtful discussion of these issues, and acknowledge support from the College of Business at the University of Illinois, Urbana-Champaign.

## References

Amabile, T. M. 1982. Social psychology of creativity: A consensual assessment technique. Journal of Personality and Social Psychology, 43(5): 997-1013.
Amabile, T. M. 1983. The social psychology of creativity: A componential conceptualization. Journal of Personality and Social Psychology, 45(2): 357-376.
Amabile, T. M., Barsade, S. G., Mueller, J. S., \& Staw, B. M. 2005. Affect and creativity at work. Administrative Science Quarterly, 50: 367-403.
Anderson, R. C., \& Pichert, J. W. 1978. Recall of previously unrecallable information following a shift in perspective. Journal of verbal learning and verbal behavior, 17(1): 112.

Biddle, B. J. 1986. Recent development in role theory. Annual review of sociology, 12: 67-92.
Bussey, J. 2012. The Innovators Enigma. Wall Street Journal.
Clark, H. H. 1996. Using language. Cambridge: Cambridge University Press.
Cooper, R. 2006. Managing Technology Development Projects. Research Technology Management, 49(6): 23.
Csikszentmihalyi, M. 1988. Society, culture, and person: A systems view of creativity. Sternberg, Robert J.

DeFillippi, R., Grabher, G., \& Jones, C. 2007. Introduction to paradoxes of creativity: managerial and organizational challenges in the cultural economy. Journal of Organizational Behavior, 28(5): 511-521.
Drazin, R., Glynn, M. A., \& Kazanjian, R. K. 1999. Multilevel theorizing about creativity in organizations: A sensemaking perspective. Academy of Management Review, 24(2): 286-307.
Elsbach, K. D., \& Kramer, R. M. 2003. Assessing Creativity in Hollywood Pitch Meetings: Evidence for a DualProcess Model of Creativity Judgments. Academy of Management Journal, 46(3): 283-301.
Fleming, L. 2001. Recombinant uncertainty in technological search. Management Science. Special Issue: Design and development, 47(1): 117-132.
George, J. M. 2007. Chapter 9: Creativity in Organizations. The Academy of Management Annals, 1(1): 439-477.
Hennessey, B. A., Amabile, T. M., \& Mueller, J. S. 2010. Chapter 46: Consensual Assessment. , Encyclopedia of Creativity, 4th edition.
Hindo. 2007. At 3M, A Struggle Between Efficiency And Creativity. Businessweek.
Keil, F. C., Stein, C., Webb, L., Billings, V. D., \& Rozenblit, L. 2010. Discerning the division of cognitive labor: An emerging understanding of how knowledge is clustered in other minds. Cognitive science, 32(2): 259300.

Keller, J., Loewenstein, J. 2011. The cultural category of cooperation: A Cultural Consensus Model analysis for China and the US. Organization Science, 22(2), 299-319.
Kim, S. H., Vincent, L. C., \& Goncalo, J. A. 2012. Outside Advantage: Can Social Rejection Fuel Creative Thought?
Klein, K. J., \& Knight, A. P. 2005. Innovation implementation: Overcoming the challenge. Current Directions in Psychological Science, 14(5): 243-246.
Klein, K. J., \& Sorra, J. S. 1996. The challenge of innovation implementation. Academy of Management Review, 21(4): 1055-1080.
Krishnan, V., \& Ulrich, K. T. 2001. Product development decisions: A review of the literature. Management Science. Special Issue: Design and development, 47(1): 121.

Mueller, J. S., Melwani, S., \& Goncalo, J. A. 2012. The bias against creativity: People people desire yet reject creative ideas. Psychological Science, 21(1): 13-17.
Okhuysen, G. A., \& Bechky, B. A. 2009. 10 Coordination in Organizations: An Integrative Perspective. The Academy of Management Annals, 3(1): 463-502.
Paletz, S. B. F., \& Peng, K. 2008. Implicit Theories of Creativity Across Cultures Novelty and Appropriateness in Two Product Domains. Journal of Cross-Cultural Psychology, 39(3): 286-302.
Preacher, K. J., Rucker, D. D., \& Hayes, A. F. 2007. Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. Multivariate Behavioral Research, 42(1), 185-227.

Rietzschel, E., Nijstad, B., \& Stroebe, W. 2009. The selection of creative ideas after individual idea generation: Choosing between creativity and impact. British Journal of Psychology, 0(1-23).
Sawyer, R. K. 2012. Explaining Creativity: The Science of Human Innovation: The Science of Human Innovation: Oxford University Press, USA.
Smith, W. K., \& Lewis, M. W. 2011. Toward a theory of paradox: A dynamic equilibrium model of organizing. Academy of Management Review, 36(2): 381-403.
Staw, B. M. 1995. Why no one really wants creativity. In C. Ford, \& D. A. Gioia (Eds.), Creative Action in Organizations: Ivory Tower Visions and Real World Voices. Thousand Oaks, CA: Sage Publications, Inc.
Sternberg, R. J. 1985. Implicit theories of intelligence, creativity, and wisdom. Journal of Personality and Social Psychology, 49(3): 607-627.
Whitson, J., \& Galinsky, A. 2008. Lacking control increases illusory pattern perception. Science, 322(5898): 115.
Winquist, L. A., Mohr, C. D., \& Kenny, D. A. 1998. The female positivity effect in the perception of others. Journal of Research in Personality, 32(3): 370-388.

# Simultaneous Contradictory Belief and the Two-System Hypothesis 

Joshua Mugg (joshuamugg@gmail.com)<br>Department of Philosophy, 4700 Keele St. Toronto, ON M3J 1P3 Canada


#### Abstract

The two-system hypothesis states that there are two kinds of reasoning systems, the first of which is evolutionarily old, heuristically (or associatively) based, automatic, fast, and is a collection of independent systems. The second is evolutionarily new, perhaps peculiar to humans, is rule-based, controlled, slow, and is a single token system. Advocates of the two-system hypothesis generally support their claim by an inference to the best explanation: two systems are needed to explain experimental data from the reasoning, heuristics, and biases literature. The best evidence for this claim comes from simultaneous contradictory belief (henceforth SCB) (Sloman 1996, 2002). I argue that Sloman has not provided us with cases of SCB. In each of his examples there is no evidence that the beliefs are held simultaneously. I then offer the outline for an experimental setup that would offer compelling evidence for the existence of SCB and thereby support the two-system hypothesis.


Keywords: Dual-process; two-system hypothesis; simultaneous contradictory belief.

## Introduction

The two-system hypothesis states there are two reasoning systems (or at least two kinds of reasoning systems), the first of which (System 1 or 'S1') is evolutionarily old, heuristically (or associatively) based, automatic, fast, and is a collection of independent systems. The second (System 2 or 'S2') is evolutionarily new, perhaps peculiar to humans, is rule-based, controlled, and slow. ${ }^{1}$ The advocate of the two-system hypothesis must demonstrate that the two systems are distinct (what I call the distinctness claim), that the two systems are of different kinds (what I call the kind claim), and that S2 is a single system. Advocates of the twosystem hypothesis (I have in mind Evans (2004); Evans \& Over (1996); Sloman (1993, 1996); Stanovich (1999, 2004); Carruthers (2009); and Frankish (2004, 2009)) generally support their claim by an inference to the best explanation: two systems are needed to explain experimental data from the reasoning, heuristics, and biases literature. The best evidence for this claim comes from simultaneous contradictory belief (henceforth SCB) (Sloman 1996, 2002).
However, their inference to the best explanation only supports, if successful, the distinctness claim. While

[^161]criticism of the two-system hypothesis has focused on the kind claim (Samuels 2009; Evans 2011), I argue that Sloman has not provided us with cases of SCB. I then offer an experimental setup that would strongly support the existence of SBC.

## Why SCB?

Before examining Sloman's cases of SCB we need to understand why SCB is good evidence for the two-system hypothesis, and we need to understand what counts as evidence for SCB. Sloman seems to have a rival explanation in mind to account for human reasoning: there is just one reasoning system (call this the one-system hypothesis). The one-system hypothesis can and should allow that this system operates differently under different circumstances. It should, for example, sometimes operate deductively and other times inductively. The two-system theorist allows for a single system to operate inductively and deductive on different occasions, since S1 and S2 engage in both forms of reasoning. What, we should ask, would be the empirical difference between there being one reasoning system that operates differently under different stimuli and there being multiple systems? One reasoning system cannot have contradictory outputs for one input (Here I am in agreement with Sloman $(1996,2002)$ ). So the one-system hypothesis is committed to the following claim: for any question demanding reasoning and for which a reasoning system will produce only one answer, subjects will only offer one answer at any given time.
Sloman understands 'belief' broadly to mean "a propensity, feeling, or conviction that a response is appropriate even if it is not strong enough to be acted on" (384). This definition is not uncontroversial, but (for the sake of argument) I will grant it for the purpose of this paper. From Sloman's definition it follows that there are at least two ways in which subjects might offer more than one response at any given time. The first is behavioral. While people might explicitly say that they believe that $p$, they may exhibit behavior demonstrating that they believe not-p. This would be evidence that there is more than one system involved in reasoning. They believe (explicitly) that p but believe (dispositionally, tacitly, or implicitly) not-p. Second, a subject might feel a tension between $p$ and not-p. This is phenomenological evidence for SCB.
As an example of a task where subjects give simultaneous contradictory responses which indicates the existence of distinct systems, consider the Muller-Lyer illusion. Subjects believe that the two lines are equal, but cannot help but see them as different lengths, even after they have measured the
two lines. Subjects perceive the lines as different lengths, but believe that they are the same length. So the MullerLyer illusion supports the claim that perception and reasoning are governed by distinct systems. Because there are two systems operating simultaneously the tension between believing that the lines are the same length and seeing them as them as different lengths persists.
While in his (1996) Sloman seems to indicate that the Muller-Lyer supports the two-system hypothesis about reasoning, in his (2002) he mere takes it to indicate that perception and knowledge are governed by distinct systems-a conclusion he (2002) recognizes is consistent with the one-system hypothesis. He explains that "the conclusion that two independent systems are at work depends critically on the fact that the perception and the knowledge are maintained simultaneously" (385, emphasis added). So the Muller-Lyer supports the existence of two distinct systems, but not two distinct reasoning systems.
The preceding two paragraphs help elucidate what counts as evidence for SCB. It would be too stringent to require that subjects verbally claim $p$ and not-p at the same time. Alternatively, a subject might quickly alter her reposes (they might say 'it is valid...Wait, no it's not. Wait, yes it is...'). Call this a response toggle. A response toggle is neither necessary nor sufficient for there being SCB, though it might count in its favor. The two contradicting answers might come from an uncertainty in a step of a single system process (e.g. being asked whether an argument is sound, and the argument is valid but the subject is unsure about the truth of a premise). The uncertainty in that case reflects the uncertainty of one or more premises rather than two competing beliefs. So response toggle might count in favour of SCB, but there are cases in which a subject will response toggle without there being any SCB. ${ }^{2}$

While the existence of SCB would be good evidence for the existence of two reasoning systems, it is possible that there are two systems and that no SCB ever occurs. Indeed, it is possible that two systems exist and that they make (almost) no causal difference. I have two lungs, but (when both function properly) they operate in essentially the same way that one large lung would. So it is not the case that if there are two systems then those two systems will differ. That is, the existence of SCB is not necessary for the establishment of the two-system hypothesis, but if there is no SCB the case for the two-system hypothesis will be much weaker. If we can explain all the data using one system, then there is no need to posit a second system. Furthermore, if both the one-system and two-systems hypotheses offer equally plausible explanations then Ockham's Razor favours the one-system hypothesis since it posits the least kinds (and number) of entities.
One might object that SCB does is not sufficient for the establishment of the two-system hypothesis. One might reason as follows: if the two-systems hypothesis is true and

[^162]the two systems can deliver different solutions to the same problem, then there may be cases in which subjects have SCB. But then the existence of SCB supports a necessary condition (rather than sufficient) for the two-system hypothesis. Actually, the argument from SCB to the twosystem hypothesis is different. The reasoning is as follows:

1. There are cases of SCB.
2. If the one-system hypothesis is true, then there will be no cases of SCB.
3. So the one-system hypothesis is false.
4. If the one-system hypothesis is false, then there must be more than one reasoning system.
5. So there is more than one reasoning system.

It is true that this does not establish that there are only two reasoning systems, but if we need to posit more than one reasoning system a good place to start is with two reasoning systems. I now turn to the Sloman's examples of SCB.

## Sloman's cases of SCB

The first example of SCB to consider is Sloman's (1998) experiment in which subjects "tended to project properties from a superordinate category to a subordinate category only to the extent that the categories were similar" (2002, 387). He supports this claim through argument strength. Subjects were asked, assuming the first statement was true, to determine the strength of the following argument:

## Argument 1

Fact: Every individual piece of electronic equipment exhibits magnetic picofluctuation.
Conclusion: Every individual piece of audio equipment exhibits magnetic picofluctuation.

The mean of the subjects who affirmed the conclusion was .89, but of course if all audio equipment is electronic, then (given that the 'fact' above is true) a rational subject would give the conclusion a probability of 1 . Sloman points out that when the category in the conclusion was atypical of the category in the premise, the judgments were even lower. For the following argument (Argument 2) the mean probability judgment was .76 (among those who claimed that all kitchen appliances where electronic).

## Argument 2

Fact: Every individual piece of electronic equipment exhibits magnetic picofluctuation.
Conclusion: Every individual kitchen appliance exhibits magnetic picofluctuation.

During debriefing interviews, subjects agreed that there was good reason to assign this argument the maximum probability because of the category inclusion. However, subjects also thought that their lower probability assessments were also sensible, though "they inevitably failed to express why" $(2002,387)$. Sloman concludes that
after being shown the correct answer they had an associative S1 answer (a probability less than 1 ) and a rule-based S2 answer (a probability of 1 ).
There is a good alternative explanation to Sloman's claim that subjects had two answers in mind. There is an enthymeme in both arguments crucial to the argument going through. In Argument 1 it is "every individual piece of audio equipment is a piece of electronic equipment" and in Argument 2 it is "every individual kitchen appliance is a piece of electronic equipment." While both of these propositions seem plausible we should consider just how certain they are. Would a microphone stand count as part of the sound equipment? Does my wine opener or manual egg beater count as a kitchen appliance? Even if we answer negatively, the questions give us pause. So the enthymeme in the two arguments are not certain, and the level of each enthymeme is gauged by association. I am more inclined to exclude the microphone stand from audio equipment than I am a manual egg beater from kitchen appliances, which would explain why Argument 2 was deemed less certain than Argument 1. I propose that the subjects' responses reflect this uncertainty about the truth of the enthymeme, since subjects were not told to assume the enthymeme.

Sloman might respond that in debriefing interviews subjects admitted that there was good reason to assign each argument a probability of 1 . When researchers pointed out to subjects that they ought to have assigned each argument maximal probability, I suggest that subjects then took the enthymeme to be true. They were then considering a more complete argument in the debriefing interview. However, subjects can claim that their answers were reasonable, and rightly so. Their answers simply reflected the probability that the enthymeme was true, which is identical to the probability of the conclusion given the fact. I conclude that Sloman's (1998) study does not give us a case of SCB.
Next, Sloman cites Osherson, Smith, Wilkie, Lopez, and Shafir (1990) as a case of SCB. In this study subjects were asked to compare the strength of two arguments:

## Argument 3

Robins have an ulnar artery.
Therefore, birds have an ulnar artery.

## Argument 4

Robins have an ulnar artery.
Therefore, ostriches have an ulnar artery.
Most subjects claimed that Argument 3 was stronger than Argument 4. Sloman thinks that subjects believe the correct answer-that Argument 4 is at least as strong as Argument 3. However they associate robins with birds more easily than robins and ostriches, and so they claim that Argument 3 is stronger. Even assuming that subjects' mistake is not due to the ambiguity in quantification (which I suspect it is), Sloman gives us no reason to think that subjects hold the beliefs simultaneously. All that Sloman offers in support of his claim that these beliefs are held simultaneously is the
following: "this is a striking example in which a compelling logical argument fails to erase an even more compelling intuition: How much evidence can a fact about robins provide for an animal as dissimilar as an ostrich?" (2002, 388). So it is the intuition that is supposed to persist even after subjects realize that Argument 4 is at least as strong as Argument 3, but this is not a case of contradictory belief. Our intuitions (inductive reasoning) tell us that Argument 4 cannot be very strong, but deductive reasoning might then recognize that this implies that Argument 3 cannot be very strong either. The competing answers might be held at different times. First one thinks (wrongly) that Argument 3 is stronger than Argument 4. Then (after it being pointed out that ostriches are a kind of bird) Argument 4 looks at least as strong as Argument 3. There is no need for the two beliefs to be held simultaneously.
Sloman offers an example from Revlin, Leirer, Yopp, and Yopp (1980) as a case of SCB in syllogistic reasoning. Subjects were asked which of the following five possible conclusions followed from the premises:

## Argument 5

No members of the ad-hoc committee are women.
Some U.S. senators are members of the ad-hoc committee.
Therefore:
a. All U.S. senators are women.
b. No U.S. senators are women.
c. Some U.S. senators are women.
d. Some U.S. senators are not women.
e. None of the above is proven.

## Argument 6

No U.S. governors are members of the Harem Club. Some Arabian sheiks are members of the Harem Club. Therefore:
a. All Arabian sheiks are U.S. governors.
b. No Arabian sheiks are U.S. governors.
c. Some Arabian sheiks are U.S. governors.
d. Some Arabian sheiks are not U.S. governors.
e. None of the above is proven.

83\% responded correctly for Argument 5 (d. Some U.S. senators are not women) while only $67 \%$ of participants responded correctly to Argument 6 (d. Some Arabian sheiks are not U.S. governors). In Argument 5 the right conclusion accords with our standing beliefs while in Argument 6 our standing beliefs tell us that the stronger answer (b No Arabian sheiks are U.S. governors) is true. Sloman concludes from this example that "empirical belief obtained fairly directly through associative memory can inhibit the response generated by psycho-logic" (2002, 389). Again, Sloman has given us no reason to think that subjects simultaneously believe that, in Argument 6, b and d both follow and that only d follows. Do we currently have behavioral evidence that subjects simultaneously believe b
and only d? I think not. We only have evidence that more subjects chose wrongly in Argument 6, and perhaps they do so because when subjects lacks training in logic they fall back on their standing knowledge.

A final (well-known) example is Linda the feminist bank teller, which was devised by Tversky and Kahneman (1983). Subjects were given the following information: "Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations" (297). Subjects were then asked which of the two following is more likely: "A: Linda is a bank-teller. B: Linda is a bankteller and is active in the feminist movement" (297). In some experimental trials more than $80 \%$ of subjects said statement B was more likely than the statement A, but of course a conjunction can never be more likely than one of its conjuncts.

Sloman claims that he "can trace through the probability argument and concede its validity, while sensing that a state of affairs that [he] can imagine much more easily has a greater chance of obtaining" $(1996,12)$. He offers the phenomenological experience of another psychologist as well: "I know that the [conjunction] is least probable, yet a little homunculus in my head continues to jump up and down, shouting at me-'but she can't just be a bank teller: read the description'" (Gould, 469).
Sloman pointed out the difference in temporal relation between the responses in the Muller-Lyer illusion and the Necker cube illusion to five of his department colleagues. Namely, in the Muller-Lyer case the illusion that the two lines are different lengths persists even after one knows they are the same length, while in the Necker cube illusion one is only able to recognize one square as the front face at any given time. Sloman then asked his colleagues whether their experience in the Linda case was analogous to the MullerLyer or Necker cube case. All five agreed that it was like the Muller-Lyer illusion. Sloman then asked his colleagues whether the 'Monty Hall' case ${ }^{3}$ was analogous to the Muller-Lyer or Necker cube illusion. All of them (as well as Sloman) thought it was analogous to the Necker cube illusion. Sloman concludes that in the Monty Hall case the contradictory beliefs are not held simultaneously, whereas for the Linda case they are.

We should not put much weight on the phenomenology of the theorists. They are not naïve subjects and their phenomenological reports are suspect. Behavioral evidence of SCB from the subjects would be better. Sloman might

[^163]object citing that his definition of belief is such that it is not necessary that subject act upon their beliefs, and a belief might not even make a behavioral difference. Although a subject might not act upon his or her beliefs, if there is a propensity, feeling, or conviction that a given proposition is true then surely it will make some difference to the cognitive processes or behavior of that subject. To be is to have causal powers (Alexander's Dictum). So if there is a belief that p then that mental state has causal powers. Furthermore our mental states do not constitute a totally separate causal web from our behavior. That is, our behavior and mental states are causally connected. Therefore, if a subject believes that $p$ then that subject's behavior will be different than if they did not believe that $p$, even if that belief is too weak to deliberately act upon.

## Experiment to test for SCB

According to the two-system hypothesis there are two systems operating and the reason that subjects answer incorrectly in the examples above is S 2 is not given a chance to make a computation or S2's response is overwhelmed by S 1 . Since the two systems operate in parallel ${ }^{4}$ a subject who offers an incorrect answer and then is told what the correct (S2) answer is will keep her S1 answer at the S1 level. ${ }^{5}$ Take the Linda case for example. If the two-system hypothesis is right then subjects continue to believe (in some sense) that it is more likely that Linda is a feminist bank-teller than a mere bank-teller even after they admit that it is at least as likely that she is merely a bankteller. A subject must suppress her S1 systems if she is to maintain the correct answer while believing a contradictory claim at the S1 level. ${ }^{6}$ This kind of suppression requires the use of cognitive resources, thus temporarily depleting the subject's cognitive resources. So subjects who come to believe the correct answer after the test expend more cognitive energy than those who continue to believe (wrongly) that Linda is more likely a feminist and bankteller than just a bank-teller. Let me offer a rough sketch of an experimental setup (based on experiments performed by Richeson \& Shelton (2007)) to determine if this is right. A similar test could be run for other cases that Sloman and others think are instances of SBC.
In step one subjects are given a reasoning task that might involve SCB (e.g the Linda case). Subjects who offer the correct response should be dismissed. Those who offer the incorrect response should then be divided into two groups. The individuals are interviewed concerning the test that they just underwent. The first group will be made aware of their error during this interview. The experimenters will explain the conjunction fallacy and apply it to the Linda case. The

[^164]second group will not be made aware of their error. In the interviews with members of group two experimenters will explain some unrelated fallacy that is supposed to be similar in phenomenology to the Necker cube (say, the Monty Hall example) (this is step two). Immediately following this interaction members of both groups will be asked to complete a Stroop Test, a typical measurement of executive functioning and cognitive depletion ${ }^{7}$ (step three). If the twosystem hypothesis is right then we should expect that the first group will have slower response times or less accurate responses for the Stroop Test than the second because they will have had to use their executive functioning to suppress their S1 belief that it is more likely that Linda is a bankteller and feminist than she is a mere bank-teller.

It is important that those in group two have a conversation that is cognitively depleting. Otherwise the cognitive depletion in group one but not group two could be attributed to subjects in group one's having a conversation concerning logic beforehand (which depleted their cognitive resources) while the subjects in group two did not. Also, it will be important that subjects in the second group do not come to realize their mistake in the Linda case on their own. Experimenters may want to check at the end of the interview that subjects' in group two have not changed their responses. Alternatively, experimenters could ask subjects if they have changed their response to the initial case after completion of the Stroop Test.

## Conclusion

Sloman has put forth the most explicit reasons for the twosystem hypothesis: at least some of the experiments in the reasoning literature involve SCB, the explanation of which requires at least two reasoning systems. I have argued that the evidence thus far presented does not demonstrate that the beliefs are held simultaneously. However this does not imply a stalemate for advocates and skeptics of the twosystem hypothesis. I have offered the outline for an experiment that would demonstrate the existence of SCB and so give good evidence that humans have two distinct reasoning systems.

## References

Evans, Jonathan St. B. T.(Ed); Frankish, Keith(Ed) (2009a). In Two Minds and Beyond. New York, NY, US: Oxford University Press.
Evans, Jonathan St. B. T. and Frankish, Keith (2009b). "The duality of mind: an historical perspective." In In Two

[^165]Minds and Beyond. Evans, Jonathan St. B. T.(Ed); Frankish, Keith(Ed). New York, NY, US: Oxford University Press.
Evans, Jonathan St. B. T (2003). In two minds: dual-process accounts of reasoning. Trends in Cognitive Science 7:10, 454-459.
Evans, Jonathan St. B. T. (2011). Thinking twice: two minds in one brain. Oxford and New York: Oxford University Press.
Frankish, K. (2004). Mind and Supermind. Cambridge University Press.
Kahneman, D. (2011). Thinking, Fast and Slow. New York: Farrar, Straus and Giroux.
Reber, A.S. (1993) Implicit Learning and Tacit Knowledge, Oxford University Press.
Revlin, R., Leirer, V., Yopp, H., \& Yopp, R. (1980). The belief-bias effect in formal reasoning: The influence of knowledge on logic. Memory \& Cognition, 8, 584-592.
Richeson, J. A., \& Shelton, J. N. (2007). Negotiating interracial interactions: Costs, consequences, and possibilities. Current Directions in Psychological Science, 16(6), 316-320.
Osherson, D. Smith, E.E., Wilkie, O., Lopez, A., and E. Shafir. (1990) Category-based induction. Psychological Review, 97: 185-200.
Samuels, Richard. The magical number two, plus or minus: Dual-process theory as a theory of cognitive kinds. In In Two Minds and Beyond. Evans, Jonathan St. B. T.(Ed); Frankish, Keith(Ed). New York, NY, US: Oxford University Press.
Sloman, S. A. (1993). Feature-based induction. Cognitive Psychology, 25,231-280.
Sloman, S.A. (1996) "The Empirical Case for Two Systems of Reasoning." Psychological Bulletin 119, 3-22.
Sloman, S.A. (1998) "Categorical inference is not a tree: The myth of inheritance hierarchies. Cognitive Psychology, 35, 1-33.
Sloman, S.A. (2002) "Two Systems Reasoning." In T Gilovich, D. Griffin and D. Kahneman (Eds.) "Heuristics and Biases: The Psychology of Intuitive Judgment." Cambridge University Press, New York.
Stanovich, K. (1999). Who is Rational? Studies of individual differences in reasoning. Lawrence Erlbaum.
Stanovich, K. (2004). The Robot's Rebellion: Finding Meaning in the Age of Darwin. University of Chicago Press: London.
Tversky, A., \& Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. Psychological Review, 90, 293-315.
Tverky, A. \& Kahneman, D . (1986). Rational choice and the framing of decisions. Journal of Business, 59 s251s278.

# Bounded Optimal State Estimation and Control in Visual Search: Explaining Distractor Ratio Effects 

Christopher W. Myers (christopher.myers.29@us.af.mil)<br>Air Force Research Laboratory<br>Wright-Patterson AFB, OH USA

Richard L. Lewis (rickl@umich.edu)
University of Michigan
Ann Arbor, Michigan USA
Andrew Howes (howesa@bham.ac.uk)
University of Birmingham
Birmingham, UK


#### Abstract

We demonstrate that an ideal observer model bounded by known limitations of the human visual system can explain empirical evidence concerning two effects of distractor ratios on visual search-effects that have previously been explained with salience-based models. The model makes optimal state estimations based on Bayesian estimates of stimuli localization and optimal control decisions of where to fixate in order to maximize task performance. Analysis of the model's behavior under different task strategies and different constraints on the visual system reveal which aspects of the model are responsible for the effects: the distractor-ratio effects on number of fixations is a signature of optimal state estimation in the face of noisy spatial information, and the saccadic-bias effect is a signature of both optimal control and estimation under these same bounds.


Keywords: optimal state estimation, optimal control, ideal observer models, visual salience, distractor ratios

## Introduction

Visual search is so ubiquitous that we probably hardly notice ourselves doing it. We search for our car keys on a cluttered desk, for our family at the market, or for a reference in text. The current paper addresses how one adapts during visual search by determining what information to visually inspect. We address each of these questions through the development of a model based on the optimal integration of perceived information given a set of known constraints on the human visual system.

In a review of the literature on eye movements, Kowler (2011) describes two general approaches to modeling visual search processes. First are map-based approaches, such as salience maps (Itti, 2006; Itti \& Koch, 2000) and activation maps (Pomplun, 2003; Wolfe, 2007), where information is accumulated and processed to produce a topogrpahical map. Peaks in the map represent areas/items that differ from their surround, that contain attributes of the target, or both. Map peaks are used to guide search through the display using some peak selection routine, such as a greedy heuristic (Pomplun, 2003) or winner-take-all algorithm (Itti \& Koch, 2000). In general, the map-based approach assumes that saccades are programmed to move the fovea to an area of a stimulus that
stand out from its surroundings or that is similar in some way to a search goal.

Alternatively, visibility models (Kowler, 2011) such as ideal observer/searcher models (Geisler, 2011; Myers, Gray, \& Sims, 2011; Najemnik \& Geisler, 2008; Baron \& Kleinman, 1969), assume that saccades are programmed to direct foveated vision to areas of impoverished acuity in order to maximize information gained in service of task performance. Najemnik and Geisler (2005) found that the number, and spatial distribution, of saccades to find a target could be predicted by a model in which each saccade was directed to the ideal location (i.e., the highest probability of finding the target). Their model was sensitive to known human constraints on vision (e.g., decreasing contrast sensitivity with increasing eccentricity). Hence, saccadic selectivity could be considered a process that maximizes search performance by considering the effect of the eyes' subsequent fixation location.

In the current paper we build on the ideal observer approach by deriving a boundedly optimal adaptive visibility model capable of capturing empirical phenomena that demonstrate adaptation to changes in the proportion of available environmental features. More specifically, we use the model to explain phenomena associated with the distractor ratio paradigm (Bacon \& Egeth, 1997; Shen, Reingold, \& Pomplun, 2000; Zohary \& Hochstein, 1989)—phenomena that have previously been given interpretations in terms of salience maps. Key to this explanation is the incorporation of constraints on the representation of spatial information in the periphery into an ideal observer analysis .

The adaptive visibility model has a simple structure that decomposes visual search into optimal state estimation (the integration of perceptual evidence into a task-relevant representation of the external stimulus) and optimal control (the choice of overt task responses and information gathering actions; Stengel, 1994). The model incorporates a small number of constraints intended to abstractly characterize important properties of a noisy, foveated vision system (Tanner, 1961). Bayesian state estimation is used to optimally integrate the noisy percepts across fixations in service of two control deci-
sions: 1) where to next fixate and 2) when to issue a task response. Both the estimation and control processes are adapted to the simultaneous constraints of the vision system and the task at hand.

The structure of this model affords the formulation and exploration of a number of interesting theoretical questions concerning visual search phenomena. In this paper, we use modeling to determine whether distractor ratio effects are signatures of optimal state estimation, optimal control, (or both), and to identify the constraints of the visual system that are necessary for the effect to arise. To foreshadow the two key results, the model demonstrates (1) that distractor ratio effects may be understood as adaptation to changes in proportions of task-relevant environment features, and that these effects are signatures of optimal state estimation (not control) in the face of spatial uncertainty in the parafovea; and (2) that saccadic bias may be understood as a signature of both optimal control and optimal state estimation in the face of spatial uncertainty.

In the following sections we first discuss efficient visual search in the distractor ratio paradigm and introduce the boundedly optimal adaptive visibility model. We next discuss model results and their implications.

## Distractor Ratio Paradigm

The distractor ratio is the ratio between distractor sets that share features with a target for a fixed number of items on a display. For example, the distractor ratio when searching for a conjunction of a color and a shape (e.g., red O ) in a display of 48 items is the number of distractors that are the same color relative to the number of distractors that are of the same shape-same-color:same-shape. Hence, the distractor ratio for Figure 1(A) is 3:45, (B) is $24: 24$, and (C) is $46: 2$. Subjects are typically instructed to respond appropriately if they determine that a target is present or absent for each trial.

The distractor ratio paradigm has been used to distinguish between endogenous and exogenous influences on saccadic selectivity processes (Bacon \& Egeth, 1997; Zohary \& Hochstein, 1989). Exogenous influences are hypothesized to arise from the statistical properties of the visual environment, such as salience (Itti, 2006; Itti \& Koch, 2000), whereas endogenous influences are those that stem from knowledge brought to bear on the task through instructions (Yarbus, 1967) or learned during performance (Myers et al., 2011). Regardless of the endogenous/exogenous process distinction, results from distractor ratio experiments demonstrate adaptation to the changing structure of the search environment. Specifically, subjects prefer to actively search through the minority set of distractors that share a common feature with the target. Using eye-tracking, Shen et al. (2000) showed that subjects searching for a target (e.g., red O) preferred the same-color distractors (red X's in Figure 1A), yet adaptively shift this preference to same-shape distractors (i.e., green O) when presented with a distractor ratio where shape was the minority feature (e.g., Figure 1C). Importantly, this adaptation reduced response times and the number of fixations to


Figure 1: Distractor ratio stimuli when searching for a red O, and results from Shen et al. (2000). Panel (A) is a stimulus containing three distractors that share the same color feature as the target. Panel (B) has equal number of like-color and like-shape distractors. Panel (C) has two like shape distractors. Panel (D) demonstrates a $\cap$-shaped curve associated with an increasing number of same-color distractors for target absent (open circles) and target present (filled circles) trials, and represents the distractor-ratio effect.
locate the target (see Figure 1D), improving search efficiency (Bacon \& Egeth, 1997; Zohary \& Hochstein, 1989).

Saccadic selectivity in the distractor ratio paradigm demonstrates rational adaptation from the standpoint that subjects minimize their time to locate a target (c.f. Gray, Sims, Fu, \& Schoelles, 2006). Hence, response times and the number of fixations are minimal when a search stimulus has a minority feature (e.g., color or shape; see Figure 1A) relative to when the distractor ratio is equal to one (see Figure 1B) for target-present and target-absent trials (see Figure 1D). Interestingly, Shen et al. (2000) report that saccadic selectivity favoring the minority feature occurred as early as the very first saccade in a trial.

One potential explanation for the distractor ratio effect is that saccadic selectivity is exogenously influenced through stimulus salience (Theeuwes, 1993). In Figure 1A, the red X distractors stand out from the surrounding green O distractors. The reverse is true for Figure 1C. Hence, the salience approach predicts saccadic selectivity favoring the red X's in Figure 1A and the green O's in Figure 1B. Importantly, the inclusion of some inhibition of return mechanism (IOR; Klein,
2000) is a required addition to salience-based models in order to eliminate endlessly fixating the most salient areas of the display, which are not guaranteed to contain the target. Importantly, the IOR and salience mechanisms would be capable of not only reproducing an important hallmark of adaptive search in the distractor ratio paradigm (the $\cap$-shaped curves depicted in Figure 1D), but also another hallmark: saccadic bias in favor of the minority distractor set.

While the salience + IOR approach provides a potential explanation of adaptive search in the distractor ratio paradigm, we sought an explanation where the observed effects are a consequence of ideal adaptation to noisy encoding processes in the fovea and parafovea. In the following section we describe a reduced complexity version of the distractor ratio paradigm for testing our model.

## Horizontal Array Distractor-Ratio Paradigm

We reduced the task environment complexity from a two dimensional array (see Figure 1A, B, \& C) to a one-dimensional array. This reduction in complexity facilitated the running of a large number of model trials while maintaining the critical components of the distractor ratio paradigm. The reduced complexity version used for testing the model was a set of seven objects arranged horizontally, with $8.3^{\circ}$ of visual angle between adjacent items. The model searched through both target-present and target-absent trials for the same target throughout. Distractors were a conjunction of the same color as the target and a different shape, or vice versa. The model was tested over seven different distractor ratios (6:0, 5:1, 4:2, 3:3, 2:4, 1:5, 0:6; see Figure 2).


Figure 2: Three trials from the horizontal distractor-ratio paradigm. The target is a red O . Trial $5: 1$ corresponds with Figure 1A, $3: 3$ corresponds with Figure 1B, and 1:5 with Figure 1 C .

This one-dimensional version of the distractor-ratio paradigm facilitated computationally tractable Monte-Carlo evaluation of the model, while retaining the relevant features of the paradigm. In the following section we describe the model in detail and present results from the model evaluated on the one-dimensional version of the task.

## Adaptive Visibility Model

The goal of this modeling endeavor is to explain the phenomena associated with strategic adaptation observed in the distractor ratio paradigm, (i.e., the $\cap$-shaped curve and saccadic bias in favor of minority features) as adaptation to perceptual
noise. To achieve this goal we differentiate between state estimation and control (Stengel, 1994). The model we present optimally estimates the state of the visual environment given noisy input, and controls responses based on the optimal state estimation.

The process of active, effortful visual search can be decomposed into two key control decisions: 1) determining if the target is in the stimulus (i.e., the stopping rule), and 2 ) determining where next in the stimulus to inspect (i.e., saccadic selectivity). All search models must contain a stopping rule and a saccade location selection process.

Toward this end we first identified physiological constraints on the visual search process. Next, we assumed that subjects in distractor ratio experiments intended to minimize the time to complete a trial. This assumption has been used in other models as a subjective utility function when an objective utility function is not provided to subjects (Gray et al., 2006; Myers et al., 2011). Third, we determined a set of strategies that could be performed in the task environment. Finally, we used Monte Carlo simulations of the model to determine if the bounded optimal model could explain the distractor-ratio hallmarks of adaptive search. Further, we investigate which model constraints were critical to adaptive visual search observed in humans when performing distractor ratio tasks. We cover each of these steps in more detail in the following sections, and provide a walkthrough of the model process before presenting the model results.

## Constraints on Visual Search

The model begins a trial with a representation of whether the shape and color feature at each of the seven stimulus locations contains the same feature as the target. The model adopts a simple feature-vector coding of the display in which each of the seven locations is represented by 2 real-valued features (one for color, one for shape), where the value 1 is arbitrarily chosen as the target value for each dimension, and 0 as the non-target value. Thus, the true state of the display can be represented as a 14 -element vector of 1 s and 0 s .

There are two constraints on the model, each of which limit the accuracy of the perceived information for each fixation. First, visual acuity decreases with increasing eccentricity from the fovea; we capture this constraint in the model with feature noise. Second, information located in the parafovea is subject to localization error (Levi, 2008; Neri \& Levi, 2006), such that objects encoded in the parafovea may erroneously combine features from different objects at different locations (illusory conjunctions; Põder \& Wagemans, 2007). Each percept obtained by the model is simply the true 14 -element vector representing the display, corrupted by these two noise sources: feature noise and location noise.

The feature noise added to each true percept is a 14 element vector of values sampled from independent normal distributions with mean zero and standard deviations that increase as a step-function based on distance from the fovea. Standard deviations for determining feature noise within the fovea were set to 0.1 and 10 for outside the fovea (the qual-
itative results presented below do not depend on the precise shape of this acuity function). Localization noise was added to the model's percept by allowing the feature value for each position to be sampled from nearby positions with some probability. This probability was set to a low value for the fovea (representing an assumption of good feature binding in the focus of attention) and higher values for parafoveal positions (again, the qualitative results presented below do not depend on the precise values). The result from introducing these constraints was a model with a foveated visual system susceptible to illusory conjunctions. For each location, we sample all objects and obtain noisy feature information from these objects. For the fixation position this will very often be the true object but features will intrude for other positions in the periphery.

## Optimal State Estimation \& Control

The model uses Bayes' rule to optimally estimate the state of the display by integrating noisy perceptual information derived from each fixation. For each given noisy perceptual sample, the model computes the likelihood that the sample was generated from each of the possible target-absent and target-present displays and features at locations within those displays. This is accomplished as follows. First, the likelihoods of observing the perceptual sample at the feature level are computed (using the feature noise model). Second, the likelihoods that a sampled object at a particular location in the display has a specific feature value for each of the possible displays is computed (using the spatial noise model). Third, the probability that the percept was sampled for each display type is computed. Finally, the posterior probability over all the display types is computed using Bayes' rule.

## Search Strategies

There are four potential strategies for locating the target in the distractor ratio paradigm. First, one could choose to make no eye movements at all during a trial, continuing to stare straight ahead. We rule out the use of this strategy as its utility in a search environment such as the distractor-ratio paradigm is very low. The remaining strategies were random search, sequential search and the maximum a posterior (MAP) searcher of Najemnik and Geisler (2008), which we label here the look-for-targets strategy, which simply looks at the location most likely to contain the target.

The random search strategy was implemented by uniformly sampling a location with replacement from all the possible locations in the reduced complexity paradigm until a response was made. Consequently, the model could choose to re-fixate a location it just acquired a percept from. The sequential search strategy was implemented by starting in the middle location and searching from right to left, and back around until a response was made.

The MAP strategy took advantage of the posterior probabilities after each fixation. The model chose the next fixation location based on the posterior likelihood of containing a target. In the next sections we provide a walkthrough of
the model's process for completing a trial followed by results from each of the three strategies just described.

## Model Walkthrough

The model begins each trial with all possible displays being equally likely; consequently, the initial values for the targetpresent and target-absent decision variables equal 0.5 . Once a trial is presented to the model, it begins by fixating a location, obtaining a noisy percept from the fixated location, optimally integrating the noisy percept with previously acquired information from the trial, and calculating decision variables (i.e., target-absent and target-present) based on the optimally integrated percept. If neither of the decision variables reaches a decision threshold (arbitrarily set to 0.85 in the simulations, but which could be optimized to maximize utility in the face of imposed speed-accuracy tradeoffs), then the model selects a new location to fixate. If one of the decision variables is greater-than the threshold, then the model responds appropriately. A maximum number of fixations was set to 30 to prevent the model from infinitely fixating locations in the trial. To be clear, the model is not learning across trials, but is adapting to each trial, independently.

## Model Results

The model was run for 20,000 trials for each of the random, sequential, and look-for-targets strategies. Each trial completed by the model was randomly selected with replacement from all possible trials. Surprisingly, all strategies produced the $\cap$-shaped curve for target-absent and target-present trials, indicating that the distract-ratio effect may arise from optimal state estimation in the face of noisy perception, independent of the saccadic control strategy. We investigate this finding in more detail below.

Less surprisingly, the random strategy required, on average, more fixations to respond $\left(M_{\text {Target-Present }}=\right.$ 4.54; $M_{\text {Target-Absent }}=5.13$ ) than the sequential strategy $\left(M_{\text {Target-Present }}=3.84 ; M_{\text {Target-Absent }}=4.53\right)$, which in turn took more fixations to respond than the look-for-targets strategy $\left(M_{\text {Target-Present }}=2.94 ; M_{\text {Target-Absent }}=3.77\right)$.

The frequency of saccades directed toward objects containing a minority feature in a trial was evaluated to determine if it differed from what would be expected by chance (i.e., saccadic bias in right column of Figure 3; Shen et al., 2000). The analysis revealed that the look-for-targets strategy produced saccadic bias for target-present and target-absent trials whereas the random and sequential strategies did not. The results from the search efficiency and saccadic bias analyses demonstrate that the look-for-targets strategy produces both hallmarks of adaptive search within the distractor ratio paradigm.

To determine which perceptual constraint was required to yield the effects, we ran another round of simulations without location noise (one of two constraints in our ideal observer model). To make this determination we ran two sets of simulations: 20,000 trials for no-location-noise/high-feature-noise and 20,000 trials for no-location-noise/low feature-noise. The


Figure 3: Hallmarks of adaptive search in the distractor ratio paradigm for the random (top), sequential (middle), and look-for-targets strategies (bottom). The left column demonstrate search efficiency in the paradigm and correspond to the human data in Figure 1D. The right column demonstrates saccadic bias in the look-for-targets strategy and the absence of the bias in the other strategies.
removal of location noise eliminated the presence of the $\cap$ shaped curve, whereas high feature noise only contributed to greater fixations to respond relative to low feature noise (see Figure 4). Consequently, we argue that the $\cap$-shaped curve observed in distractor ratio tasks results from the potential for illusory conjunctions in the parafovea.

## Discussion \& Conclusions

The preliminary analysis presented above contrasted a wellknown salience based theory and a novel ideal observer based theory of distractor ratio phenomena. Despite the fact that the salience theory is widely accepted and that there is no previous ideal observer analysis of distractor ratio phenomena, we found that it offered a comprehensive explanation of the available empirical findings. Importantly, the different behaviors seen in people as a consequence of varying the statistical structure of the task environment emerge from a model that computes optimal state estimation and makes optimal control decisions given the constraints imposed by the human visual


Figure 4: Search efficiency results without location noise when feature noise was low (left) and when feature noise was high (right) for the look-for-targets strategy.

## system.

These preliminary findings are promising because the ideal observer, by virtue of the combination of optimal state estimation and control, offers the potential of a deeper explanation than the mechanistic salience model. The ideal observer combines a theory of the information processing mechanisms with an analysis of optimal state estimation and control. Furthermore, the estimation and control decomposition permits the exploration of specific hypotheses concerning the locus of the explanation for a given search phenomenon. Here, we determined that distractor ratio effects are signatures of optimal state estimation in the face of spatial noise in the periphery, while the saccadic bias effects are signatures of both optimal estimation and control.

Although these findings are encouraging, the model requires a number of important revisions before we can be fully confident that it provides a rigorous demonstration of the implications of the hypothesized visual processing constraints for behavior. In particular, we did not explore the full strategy space for directing saccades; although the look-fortargets (MAP) strategy may be close to optimal in this task, we must derive the optimal strategy in the full space and confirm that its predictions are consistent with those of the simple look-for-targets strategy.

Furthermore, we must conduct new human experiments that systematically test predictions of the ideal observer that differ from those of the salience model. We envisage that the new data will be collected using a utility maximization paradigm similar to those used by Trommershäuser and colleagues (Stritzke, Trommershäuser, \& Gegenfurtner, 2009; Trommershäuser, Maloney, \& Landy, 2003) in the explorations of perceptual motor control, and Lewis, Shvartsman, and Singh (to appear) in the exploration of eye-movements in linguistic tasks. Bounded optimal control models naturally predict differences in performance that arise when the payoff is changed but the task and stimuli are otherwise identical,
while salience-based models do not naturally predict such differences. A key advantage of these explicit-payoff paradigms is that assumptions regarding what subjects are maximizing during the experiment are grounded in the external payoffs, which are then used as the subjective utility functions in the optimal control models.

## Acknowledgments

This research was supported by a grant awarded from the Air Force Office of Scientific Research to the three authors (AFOSR \#12RH05COR) and grants awarded by NASA and the FAA to the second and third authors (NNX12AB08A and E2020272).

## References

Bacon, W. J., \& Egeth, H. E. (1997). Goal-directed guidance of attention: evidence from conjunctive visual search. Journal of Experimental Psychology: Human Perception and Performance, 23(4), 948-961.
Baron, S., \& Kleinman, D. L. (1969). The Human as an Optimal Controller and Information Processor. IEEE Transactions on Man-Machine Systems, 10(1), 10-17.
Geisler, W. S. (2011). Contributions of ideal observer theory to vision research. Vision Research, 51(7), 771-781.
Gray, W. D., Sims, C. R., Fu, W.-T., \& Schoelles, M. J. (2006). The soft constraints hypothesis: a rational analysis approach to resource allocation for interactive behavior. Psychological Review, 113(3), 461-482.
Itti, L. (2006). Quantitative modelling of perceptual salience at human eye position. Visual Cognition, 14(4-8), 959984.

Itti, L., \& Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Research, 40(10-12), 1489-1506.
Klein, R. M. (2000). Inhibition of return. Trends in Cognitive Sciences, 4(4), 138-147.
Kowler, E. (2011). Eye movements: the past 25 years. Vision Research, 51(13), 1457-1483.
Levi, D. M. (2008). Crowding An essential bottleneck for object recognition : A mini-review. Vision Research, 48, 635-654.
Lewis, R. L., Shvartsman, M., \& Singh, S. (to appear). The adaptive nature of eye-movements in linguistic tasks: How payoff and architecture shape speed-accuracy tradeoffs. Topics in Cognitive Science.
Myers, C. W., Gray, W. D., \& Sims, C. R. (2011). The insistence of vision : Why do people look at a salient stimulus when it signals target absence ? Visual Cognition, 19(9), 1122-1157.
Najemnik, J., \& Geisler, W. S. (2005). Optimal eye movement strategies in visual search. Nature, 434(7031), 387391.

Najemnik, J., \& Geisler, W. S. (2008). Eye movement statistics in humans are consistent with an optimal search strategy. Journal of Vision, 8(3), 4.1-14.

Neri, P., \& Levi, D. M. (2006). Spatial resolution for feature binding is impaired in peripheral and amblyopic vision. Journal of Neurophysiology, 96(1), 142-153.
Põder, E., \& Wagemans, J. (2007). Crowding with conjunctions of simple features. Journal of Vision, 7(2), 1-12.
Pomplun, M. (2003, April). Area activation: a computational model of saccadic selectivity in visual search. Cognitive Science, 27(2), 299-312.
Shen, J., Reingold, E. M., \& Pomplun, M. (2000). Distractor ratio influences patterns of eye movements during visual search. Perception, 29(2), 241-250.
Stengel, R. (1994). Optimal Control and Estimation. New York: Dover.
Stritzke, M., Trommershäuser, J., \& Gegenfurtner, K. R. (2009). Effects of salience and reward information during saccadic decisions under risk. Journal of the Optical Society of America A, 26(11), B1-B13.
Tanner, W. (1961). Physiological implications of psychophysical data. Annals of the New York Academy of Sciences, 89, 752-765.
Theeuwes, J. (1993). Visual selective attention: a theoretical analysis. Acta Psychologica, 83(2), 93-154.
Trommershäuser, J., Maloney, L. T., \& Landy, M. S. (2003, July). Statistical decision theory and the selection of rapid, goal-directed movements. Journal of the Optical Society of America. A, Optics, image science, and vision, 20(7), 1419-33.
Wolfe, J. M. (2007). Guided Search 4.0: Current Progress with a model of visual search. In W. D. Gray (Ed.), Integrated models of cognitive systems (pp. 99-119). New York: Oxford University Press.
Yarbus, A. L. (1967). Eye Movements and Vision. New York: Plenum.
Zohary, E., \& Hochstein, S. (1989). How serial is serial processing in vision? Perception, 18(2), 191-200.

# Is action-guiding vision cognitively impenetrable? 

Bence Nanay (bence.nanay@ua.ac.be or bn206@cam.ac.uk)<br>Centre for Philosophical Psychology, D 413, Grote Kauwenberg 18, Antwerp, 2000 Belgium<br>and<br>Peterhouse, Cambridge University, Cambridge, CB2 1RD, UK


#### Abstract

The aim of this paper is to argue that action-guiding vision is not cognitively impenetrable and arguments that suggest otherwise rely on an unjustified identification between actionguiding vision and dorsal vision - a functional and an anatomical way of describing the mind. The examination of these arguments show the importance of making a distinction between the functional and the anatomical level when addressing the problem of cognitive penetrability.


Keywords: Cognitive impenetrability; dorsal vision; actionguiding vision; perception and action; multimodal perception; top-down influences on perception

## The impenetrability of conscious versus unconscious perception

There are notorious ambiguities in the formulation of the cognitive penetrability debate. Some of these has been acknowledged and (more or less) resolved, like the one about the various senses in which a mental process could be considered to be cognitively penetrable (Fodor 1983). A seemingly simpler but often unacknowledged ambiguity is not about the predicate of the claim that perception is cognitively impenetrable, but its subject.

It seems that when philosophers and psychologists talk about the cognitive penetrability or impenetrability of perception, they mean very different things by 'perception'. Some (mostly philosophers) mean conscious perceptual experience (Lyons 2011, Siegel 2011), some others (mostly psychologists) mean perceptual processes in general, conscious or unconscious. As perception can be conscious or unconscious, even if it is true that conscious perceptual experience is cognitively penetrable (or impenetrable), this should not be generalized to unconscious perception and, as a result, to perception in general.

This distinction is especially important if we consider the following possibility. Many philosophers and psychologists now assume some version of the claim that conscious experiences are cognitively penetrable. But this leaves open the question whether unconscious perceptual states are cognitively penetrable. More specifically, this leaves open whether the most compelling cases for unconscious perceptual processes, namely, the ones that guide our goaldirected actions, are cognitively impenetrable.

According to a more and more influential view, the unconscious perceptual, or, more precisely, visual, processes that guide our goal-directed actions, which we can call 'action-guiding vision' is cognitively impenetrable (Milner 2008, Goodale \& Wolf 2009, Raftopoulos 2001, 2005, Norman 2002, Goodale 2011, Jacob \& Jeannerod 2003, Jeannerod \& Jacob 2005, Jacob 2005, Milner \&

Goodale 1995, 2008, Goodale \& Milner 2004, Rizzolatti \& Matelli 2003). For many philosophers and psychologists, the last refuge of the cognitive impenetrability thesis is unconscious action-guiding vision.

My aim is to argue that action-guiding vision is not cognitively impenetrable and arguments that suggest otherwise rely on an unjustified identification between action-guiding vision and dorsal vision - a functional and an anatomical way of describing the mind. The examination of these arguments show the importance of making a distinction between the functional and the anatomical level when addressing the problem of cognitive penetrability.

## Action-guiding vision

I call 'action-guiding vision' those, mainly unconscious, processes in visual perception that help us to perform the goal-directed movements of our actions. But why should we assume that there is such processes? And why should we think that they are unconscious?

The main reason for thinking that there is such a thing as unconscious action-guiding vision is that in certain circumstances, our conscious visual experiences represent the world differently from the way the perceptual processes that help us to perform goal-directed actions do (see, e.g., Loach et al. 2008, Xu et al. 2012). The most important examples for this comes from the study of optical illusions.

A number of optical illusions mislead our perceptual experience but not (or much less) our action-guiding vision. One such example is the three dimensional Ebbinghaus illusion. The two dimensional Ebbinghaus illusion is a simple optical illusion. A circle that is surrounded by smaller circles looks larger than a circle of the same size that is surrounded by larger circles. The three dimensional Ebbinghaus illusion reproduces this illusion in space: a poker-chip surrounded by smaller poker-chips appears to be larger than a poker-chip of the same diameter surrounded by larger ones. The surprising finding is that although our perceptual experience is incorrect - we experience the first chip to be larger than the second one -, if we are asked to pick up one of the chips, our grip-size is hardly influenced by the illusion (Aglioti et al. 1995, see also Milner \& Goodale 1995, ch. 6 and Goodale \& Milner 2004). Similar results can be reproduced in the case of other optical illusions, like the Müller-Lyer illusion (Goodale \& Humphrey 1998, Gentilucci et al. 1996, Daprati \& Gentilucci 1997, Bruno 2001), the 'Kanizsa compression illusion' (Bruno \& Bernardis 2002), the dot-in-frame illusion (Bridgeman et al., 1997), the Ponzo illusion
(Jackson \& Shaw 2000, Gonzalez et al. 2008) and the 'hollow face illusion' (Króliczak et al. 2006). ${ }^{1}$

What makes it possible for us to reach for the chip with the (more or less) appropriate grip size is action-guiding vision - the visual processes that help us to perform goaldirected actions. Our conscious experience represents the size-property of the chip one way (incorrectly) and our action-guiding vision represents it a different way (more or less correctly). Hence, action-guiding vision is different from our conscious perceptual experience. We cannot explain this behavior without postulating action-guiding vision. And as action-guiding vision represents the sizeproperty of the chip differently from the way conscious perception does, it must do so unconsciously.

I call the representations of action-guiding perceptual processes 'action-oriented representations' (Nanay 2012, Clark 1997, Mandik 2005). In the 3D Ebbinghaus case, it is the action-oriented representation that guides our action: it attributes, unconsciously, the (more or less) correct sizeproperty to the poker-chip - in spite of our very misleading conscious experience.

This focus on action-oriented representations as the mental states that mediates between perception and action provide a more and more popular third alternative to both classic computationalism/propositionalism and antirepresentationalism/enactivism. The mind is to be understood in terms of representations, but these representations are not all propositional, linguistically structured or uniquely human. Some are better compared to the mental representations of the predator that make it possible for it to catch its prey. These representations are simple, supposedly non-propositional, and maybe perceptual, representations and they are also inherently action-oriented (Norman 2002, Hummel et al. 2001, Grush 2004, Gendler 2008, Jeannerod 1997, Millikan 1995, 2004, Pacherie 2011, Jeannerod \& Jacob 2004, Clark 1997, Mandik 2005, Nanay 2011, 2012, 2013, in press).

It is not universally agreed upon that action-oriented representations can be considered to be perceptual states although there are some arguments for this conclusion (Bach 1978, Nanay 2011, 2012, 2013). But those accounts of action-guiding vision this paper is about, the ones that consider action-guiding vision to be cognitively impenetrable, invariably do consider representations that mediate between perception and action to be perceptual

[^166]states (Jeannerod \& Jacob 2003, Jacob \& Jeannerod 2005, Jacob 2005, Jeannerod 1997, Milner \& Goodale 1995, 2008, Goodale \& Milner 2004, Norman 2002).

So far, I treated action-guiding vision as a genuinely philosophical or theoretical category. But this philosophical way of raising the question may puzzle neuroscientists. They have long been studying the link between perception and action and we have a lot of empirical evidence about the nature of the processes that mediate between perception and action. And this body of evidence points to the dorsal visual subsystem. The dorsal visual subsystem is a genuine part of the perceptual system of mammals and its function is widely acknowledged to be the guiding of goal-directed actions.

In the light of the similarities between the dorsal visual subsystem and action-guiding vision, a very tempting suggestion would be to say that action-oriented representations must be the representations of the dorsal visual subsystem (see, e.g., Jacob \& Jeannerod 2003, Jacob 2005, Norman 2002, see also Matthen 2005 for a more cautious claim). The dorsal system guides action and represents the world in such a way that would help us perform actions - this sounds exactly like what actionguiding vision is supposed to do.

I will argue that we should resist this temptation to equate 'action-guiding vision' with 'dorsal vision'. And, more specifically, we should not identify action-oriented representations, the representations, posited on the functional level, with the representations of the dorsal stream. The dorsal stream plays an important role in the implementation of action-guiding vision, but it is unlikely that it plays the only role. We should be careful not to conflate the functional and the neural level.

## Dorsal vision

Humans (and other mammals) have two visual subsystems that use different regions of the central nervous system, the ventral and dorsal streams. To put it simply, the ventral stream is responsible for identification and recognition, whereas the function of the dorsal stream is the visual control of our motor actions. In normal circumstances, these two systems work together, but if one of them is removed or malfunctions, the other can still function relatively well (see Milner \& Goodale 1995, Goodale \& Milner 2004).

If the dorsal stream is malfunctioning, the agent can recognize the objects in front of her, but she is incapable of manipulating them or even localizing them in her egocentric space (especially if the perceived object falls outside the agent's fovea). This is called optic ataxia. If the ventral stream is malfunctioning, a condition called visual agnosia, the agent can perform actions with objects in front of her relatively well, but she is incapable of even guessing what these objects are.

The three dimensional Ebbinghaus illusion I mentioned above is normally explained as a nice demonstration of the dissociation between the dorsal and ventral visual subsystems in healthy human adults: the ventral subsystem is fooled by the illusion, but the dorsal is not. The other
examples in which optical illusions deceive the eye, but not the hand (Ponzo, Müller-Lyer, Kanizsa-compression, hollow face, etc) are analyzed in the same way. Sometimes our ventral visual subsystem attributes a different property to an object from the one the dorsal subsystem does.

The most important characteristics of the dorsal stream from the point of view of this paper is that it is taken to be informationally encapsulated from the rest of the brain. The original picture (in Milner \& Goodale 1995) was that dorsal processing is quick, automatic and insensitive not only to higher order mental processes, but also to processing in the ventral stream. While it has been very much debated whether the dorsal stream is insensitive to processing in the ventral stream (see below), the claim about the insensitivity of dorsal processing to higher order mental processes remains more or less uncontroversial even in the works of those who argue for various interactions between the dorsal and the ventral streams (see, e.g., Jeannerod and Jacob 2005, Rizzolatti \& Matelli 2003, Kravitz et al. 2011, Rossetti \& Pisella 2002).

Back to action-guiding vision. A very tempting suggestion would be to say that action-guiding vision is just dorsal vision and action-guiding representations are the representations of the dorsal visual subsystem. The dorsal system guides action, just like action-guiding vision. The dorsal system represents the world in such a way that would help us perform actions and so do action-oriented representations. Shouldn't we then just say that it is the dorsal stream that mediates between perception and action? If we were to accept this claim, it would follow that actionguiding vision is cognitively impenetrable. In fact, the proponents of the idea of the cognitive impenetrability of action-guiding vision take the informational encapsulation of the dorsal stream to be the main evidence for their claim (see esp. Goodale \& Wolf 2009, see also Jeannerod \& Jacob 2003, Norman 2002, Milner \& Goodale 1995, 2008, Raftopoulos 2001, 2005).

My aim is to carefully detach claims about action-guiding vision from claims about the dorsal stream. Action-guiding vision is not the same as dorsal vision. Whatever the neural implementation of action-guiding vision is, it surely includes the dorsal visual subsystem. But it cannot be restricted to the dorsal visual subsystem, for the following three reasons.

## Interactions between the dorsal and the ventral subsystems

First, the anatomical distinction between the dorsal and the ventral visual subsystems is not as neat and clear-cut as it was originally thought. It seems that there are interactions between the two streams at various point of perceptual processing (see, e.g., Jeannerod 1997, Franz \& Gegenfurtner 2008, Franz et al. 2000, Schenk \& McIntosh 2010, Rosetti \& Pisella 2002).

Further, to make things even more complicated, it has been argued that instead of two visual subsystems, we need to talk about three: the ventral, the ventrodorsal and the
dorsodorsal. To simplify matters considerably, what has been taken to be one single dorsal subsystem should be divided into two: one responsible for manipulating objects (dorsodorsal) and one responsible for localizing in egocentric space (ventrodorsal) (Rizzolatti \& Matelli 2003).

Even more recently it has been suggested that what was originally taken to be the dorsal stream is in fact the ensemble of three different visual subsystems (Kravitz et al. 2011). To talk about the dorsal stream as an independent chunk of the brain and to talk about action-oriented representations as the representations of this unified and independent bit of mental processing would be misleading to say the least.

## Is dorsal vision necessarily unconscious?

Second, there is a major debate both in vision science and in philosophy of cognitive science about whether dorsal vision is necessarily unconscious. The original proposal was that ventral visual processing may be conscious or unconscious, but dorsal processing is always unconscious. (see esp. Milner \& Goodale 1995, Goodale \& Milner 2004). But this view has been criticized both on empirical and on conceptual grounds (see Dehaene et al, 1998, Jeannerod 1997, Jacob \& Jeannerod 2003).

This debate does not seem to go away (see Brogaard 2011, Briscoe 2008, Milner \& Goodale 2008, Jeannerod \& Jacob 2004, Goodale 2011, Clark 2009, Kravitz et al. 2011). As action-oriented representations can be conscious or unconscious (although they are typically unconscious), if we were to equate action-oriented representations with dorsal perceptual processing, we would have to take sides in this grand debate, which proponents of the action-oriented representation approach would be well advised to avoid.

## Multimodality

The third reason why action-oriented representation is not to be identified with the representation of dorsal perception is the multimodality of perception. There is a lot of recent evidence that multimodal perception is the norm and not the exception - our sense modalities interact in a variety of ways (see Spence \& Driver 2004, Vroomen et al. 2001, Bertelson \& de Gelder 2004, O’Callaghan 2008). Information in one sense modality can influence the information processing in another sense modality at a very early stage of perceptual processing (often in the primary visual cortex in the case of vision (e.g., Watkins et al. 2006).

A simple example is ventriloquism, which is an illusory auditory experience caused by something visible (Bertelson 1999, O’Callaghan 2008). The auditory sense modality identifies the ventriloquist as the source of the voices, while the visual sense modality identifies the dummy. And the visual sense modality modifies the way we auditorily experience the scene. But there are more surprising examples: if there is a flash in your visual scene and you hear two beeps while the flash lasts, you experience it as two flashes (Shams et al. 2000).

Now, action-oriented representations are not necessarily visual: they can occur in any sense modality. But the dissociation between the dorsal and ventral subsystem is a distinction in the visual sense modality. Some have suggested a similar dissociation for speech perception (see, e..g, Hickock \& Poeppel 2007) and for touch (Reed et al. 2005), but the evidence for dissociations similar to the one in the case of vision is far from clear in audition, olfaction and the other sense modalities. Tying action-oriented representations to the dorsal visual subsystem would make it difficult to talk about action-oriented representations in sense modalities other than vision.

Further, the literature on the multimodality of perception clearly shows that our perceptual states in one sense modality are influenced by the information we receive in other sense modalities. And there are some recent behavioral experiments supporting the multimodality of action-oriented representations (see esp. Stein et al. 2004, Gentilucci et al., 1995). How about the dorsal stream? Although it seems clear that the dorsal stream is also multimodal (see, e.g., Battaglia-Mayer \& Caminiti 2002), but the exact extent of the crossmodal influences on dorsal processing has been debated (see, e.g., Lewis \& Van Essen 2000, Rozzi 2008). Again, it seems that the neural correlate of action-oriented representation has a lot to do with the dorsal stream, but the current empirical evidence on multimodal perception does not quite support the claim that it is identical to, or fully exhausted by, the dorsal stream.

To sum up, these findings all point in the direction of a theoretical framework where it is clear that the dorsal stream plays an important role in the implementation of actionguiding vision, but it is unlikely that it plays the only role. Whatever the neural implementation of action-guiding vision is, it surely includes the dorsal visual subsystem. But we have strong reasons to doubt that it is restricted to the dorsal visual subsystem.

## Is action-guiding vision cognitively impenetrable?

We are finally in the position to assess the claims about the cognitive impenetrability of action-guiding vision. The proponents of the idea of the cognitive impenetrability of action-guiding vision take the informational encapsulation of the dorsal stream to be the main evidence for their claim (see esp. Goodale and Wolf 2009, see also Jeannerod \& Jacob 2003, Norman 2002, Milner \& Goodale 1995, 2008, Raftopoulos 2001, 2005). But as I argued, action-guiding vision and dorsal vision should not be conflated. As a result, the argument from the informational encapsulation of the dorsal stream will bear no direct implications for the cognitive impenetrability of action-guiding vision.

But we have even stronger reasons to mistrust the suggestion that action-guiding vision is cognitively impenetrable. We have some positive evidence that actionguiding vision is sensitive to various top-down factors, like the subject's affective life (Morgado et al. 2011), her language skills (Pulvermuller et al. 2005) and her
expectations or knowledge. The example I will be focusing on is the following: two very widely used brand of matches in the UK are 'Swan Vestas' and 'Scottish Bluebell'. The box of Swan Vestas is $25 \%$ larger than that of Scottish Bluebells. And it turns out that the brand of the match boxes influences our grip size when grasping them (McIntosh \& Lashleya 2008). When the subjects were grasping the 1.25scale replica of the Scottish Bluebell box, their grip size was smaller than it was when grasping the normal Swan Vestas of the same size. And when they were grasping the 0.8 -scale replica of Swan Vestas box, their grip size was larger than it was when grasping the normal Scottish Bluebell box. Hence, the brand of the match boxes (but at the very least, the recognition thereof) influences grip size: it influences our action-guiding vision.

Dorsal vision may or may not be informationally encapsulated. But action-guiding vision, as the McIntosh and Lashleya experiment shows, is cognitively penetrable. Then this finding can be used as an independent evidence for the claim that dorsal vision and action-guiding vision are different and they should not be confused. ${ }^{2}$

But there is a more general lesson to be learned from the controversy about the cognitive impenetrability of actionguiding vision. Dorsal vision is an anatomical concept - it is identified by means of anatomical criteria. Action-guiding vision is a functional concept - it is identified by means of functional criteria. To confuse the two is to confuse the functional and the anatomical ways of describing the mind.

And this confusion is especially dangerous when it comes to the cognitive impenetrability debate. The proponents and opponents of the cognitive impenetrability of perception agree that whatever is meant by perception in this debate, for example, whether it is conscious experience or some unconscious perceptual process, it is to be identified by means of functional criteria. But then we should be suspicious of using anatomical data in support of, or against, the cognitive impenetrability claim.

If we are to keep the functional and the anatomical levels of describing the mind separate, then using anatomical data to bear directly on the cognitive impenetrability debate is a methodological mistake. The example of using the anatomical data of the dorsal stream to argue for the cognitive impenetrability of action-guiding vision is a good case study of this.

## Acknowledgments

This work was supported by the EU FP7 CIG grant PCIG09-GA-2011-293818 and the FWO Odysseus grant G.0020.12N.

[^167]
## References

Aglioti, S., DeSouza, J.F.X., \& Goodale, M.A. (1995). Sizecontrast illusions deceive the eye but not the hand. Current Biology, 5, 679-685.
Bach, K. (1978). A representational theory of action, Philosophical Studies 34, 361-379.
Battaglia-Mayer, A., \& Caminiti, R. (2002). Optic ataxia as a result of the breakdown of the global tuning fields of parietal neurones. Brain, 125, 225-237.
Bertelson, P. \& de Gelder, B. (2004). The psychology of multimodal perception. In C. Spence \& J. Driver (Eds), Crossmodal Space and Crossmodal Attention, Oxford: Oxford University Press.
Brand, M. (1984). Intending and Acting. Cambridge, MA: MIT Press.
Bridgeman B, Peery S. \& Anand S. (1997). Interaction of cognitive and sensorimotor maps of visual space. Perception \& Psychophysics 59, 456-459.
Briscoe, R. (2008). Another look at the two visual systems hypothesis. Journal of Conscious Studies 15, 35-62.
Brogaard, B. (2011). Are there unconscious perceptual processes? Consciousness and Cognition 20, 449-63.
Bruno, N. (2001). When does action resist visual illusions? Trends in Cognitive Sciences 5, 385-388.
Bruno, N. \& Bernardis, P. (2002). Dissociating perception and action in Kanizsa's compression illusion. Psychonomic Bulletin \& Review 9, 723-730.
Clark, A. (1997). Being There. Cambridge, MA: MIT Press.
Clark, A. (2009). Perception, action, and experience. Neuropsychologia, 47, 1460-1468.
Daprati, E. \& Gentilucci, M. (1997). Grasping an illusion. Neuropsychologia 35, 1577-1582.
Dehaene, S., Naccache, L., Le Clec'H, G., Koechlin, E., Mueller, M., Dehaene-Lambertz, G., van de Moortele, P.F. \& Le Bihan, D. (1998). Imaging unconscious semantic priming. Nature, 395, 597-600.
Dijkerman, H. C., \& De Haan, E. H. F. (2007). Somatosensory processes subserving perception and action. Behavioral and Brain Sciences, 30, 189-201.
Fodor, J.A. (1983). The Modularity of Mind. Cambridge, MA: MIT Press.
Franz, V. (2001). Action does not resist visual illusions, Trends in Cognitive Sciences, 5, 457-59.
Franz, V. (2003). Manual size estimation: A neuropsychological measure of perception?, Experimental Brain Research, 151, 471-77.
Franz, V. \& Gegenfurtner, K. (2008). Grasping visual illusions: consistent data and no dissociation. Cognitive Neuropsychology 25, 920-50.
Franz, V. , Bülthoff, H. \& Fahle, M. (2003). Grasp effects of the Ebbinghaus illusion: obstacle avoidance is not the explanation. Experiental Brain Research 149, 470-477.
Franz, V., Gegenfurtner, K., Bülthoff, H. \& Fahle, M. (2000). Grasping visual illusions: No evidence for a dissociation between perception and action, Psychological Science, 11, 20-25.

Gendler, T. (2008). Alief in action (and reaction) Mind \& Language, 23, 552-585.
Gentilucci, M., Cheiffe, S., Daprati, E., Saetti, M. C. \& Toni, I. (1996). Visual illusion and action. Neuropsychologia 34, 369-376.
Gentilucci, M., Daprati, E., Toni, I., Chieffi, S., \& Saetti, M. C. (1995). Unconscious updating of grasp motor program. Experimental Brain Research, 105, 291-303.
Gillam, B. 1998 Illusions at Century's End. In J. Hochberg (Ed.), Perception and Cognition at Century's End. San Diego: Academic Press.
Gonzalez, C., Ganel, T., Whitwell, R., Morrissey, B. \& Goodale, M. (2008). Practice makes perfect, but only with the right hand, Neuropsychologia, 46, 624-631.
Goodale M.A. (2011). Transforming vision into action. Vision Research 51, 1567-87.
Goodale, M.A. \& Humphrey, G.K. (1998). The objects of action and perception. Cognition 67, 181-207.
Goodale, M.A. \& Milner, A.D. (2004). Sights Unseen. Oxford: Oxford University Press.
Goodale, M.A. \& Wolf, M. (2009). Vision for action. In D. Dedrick \& L. Trick (Eds.), Computation, Cognition, and Pylyshyn. Cambridge, MA: MIT Press.
Grush, R. (2004). The emulation theory of representation, Behavioral and Brain Sciences 27, 377-442.
Haffenden, A. \& Goodale, M.A., (1998). The effect of pictorial illusion on prehension and perception. Journal of Cognitive Neuroscience 10, 122-136.
Haffenden, A. M., Schiff, K. C., \& Goodale, M. A. (2001). The dissociation between perception and action in the Ebbinghaus illusion. Current Biology 11, 177-181.
Hickock, G. \& Poeppel, D. (2007). The cortical organization of speech processing. Nature Reviews Neuroscience,8, 393-401.
Hommel, B., Müsseler, J., Aschersleben, G., \& Prinz, W. (2001). The theory of event coding. Behavioral and Brain Sciences, 24, 849-878.
Jackson, S. \& Shaw, A. (2000). The Ponzo illusion affects grip-force but not grip-aperture scaling during prehension movements, Journal of Experimental Psychology HPP, 26, 418-23.
Jacob, P. (2005). Grasping and perceiving an object. In A. Brooks \& K. Akins (Eds.), Cognition and the Brain. Cambridge: Cambridge University Press.
Jacob, P. \& Jeannerod, M. (2003). Ways of Seeing. Oxford: Oxford University Press.
Jeannerod, M. (1994), The representing brain: Neural correlates of motor intention and imagery, Behavioral and Brain Sciences, 17, 187-245.
Jeannerod, M. (1997). The Cognitive Neuroscience of Action. Oxford: Blackwell.
Jeannerod, M. \& Jacob, P. (2005). Visual cognition: a new look at the two-visual systems model. Neuropsychologia, 43, 301-312.
Kaas, J. H., \& Hackett, T. A. (1999). 'What' and 'where' processing in auditory cortex. Nature Neuroscience, 2, 1045-1047.

Kravitz, D. J., Kadharbatcha S., Saleem, S., Baker, C. I. \& Mishkin, M. (2011). A new neural framework for visuospatial processing. Nature Neuroscience 12, 217-230 Króliczak, G., Heard, P., Goodale, M.A. \& Gregory, R.L. (2006). Dissociation of perception and action unmasked by the hollow-face illusion. Brain Research, 1080, 9-16.
Lewis, J. \& Van Essen, D. (2000). Corticocortical connections of visual, sensorimotor and multimodal processing areas in the parietal lobe of the macaque monkey. Journal of Comp. Neurology, 428, 112-37.
Loach, D., Frischen, A., Bruce, N. \& Tsotsos, J.K. (2008). An attentional mechanism for selecting appropriate actions afforded by graspable objects. Psychological Science, 19, 1253-57.
Lyons, J. (2011). Circularity, reliability and the cognitive penetrability of perception, Philosophical Issues 21, 289311.

Mandik, P. (2005). Action oriented representation. In A. Brook \& K. Akins (Eds.), Cognition and the Brain: The Philosophy and Neuroscience Movement. Cambridge: Cambridge University Press.
Matthen, M. (2005). Seeing, Doing, and Knowing: A Philosophical Theory of Sense-Perception. Oxford: Clarendon Press.
McIntosh, R.D., \& Lashleya, G. (2008). Matching boxes: Familiar size influences action programming. Neuropsychologia, 46, 2441-2444.
Millikan, R.G. (1995). Pushmi-pullyu representations. Philosophical Perspectives, 9, 185-200.
Millikan, R.G. (2004). Varieties of Meaning. Cambrdige, MA: The MIT Press.
Milner, A.D. (2008). Conscious and unconscious visual processing in the human brain. In L. Weiskrantz \& M. Davies (Eds.), Frontiers of Consciousness. Oxford: Oxford University Press.
Milner, A.D. \& Goodale, M.A. (1995). The Visual Brain in Action, Oxford: Oxford University Press.
Milner, A.D., \& Goodale, M.A. (2008). Two visual systems re-viewed. Neuropsychologia, 46, 774-785.
Morgado N., Muller D., Gentaz E., \& Palluel-Germain R. (2011). Close to me? The influence of affective closeness on space perception. Perception, 40, 877-879)
Nanay, B. (2011). Do we see apples as edible? Pacific Philosophical Quarterly, 92, 305-322.
Nanay, B. (2012). Action-oriented perception. European Journal of Philosophy, 20, 430-446.
Nanay, B. (2013). Between Perception and Action. Oxford: Oxford University Press.
Nanay, Bence (in press). Success semantics: The sequel. Philosophical Studies
Norman, J. (2002). Two Visual Systems and Two Theories of Perception. Behavioral and Brain Sciences, 25, 73-144.
O'Callaghan, C. (2008). Seeing what you hear: Crossmodal illusions and perception. Philosopical Issues, 18, 316-38.
Pacherie, E. (2011). Nonconceptual representations for action and the limits of intentional control. Social Psychology, 42, 67-73.

Pavani, F., Boscagli, I., Benvenuti, F., Rabuffetti, M., and Farnè, A. (1999). Are perception and action affected differently by the Titchener circles illusion? Experimental Brain Research, 127, 95-101.
Pulvermuller, F., \& Hauk, O. (2005). Functional links between motor and language systems. European Journal of Neuroscience , 21, 793-797
Raftopoulos, A. (2001). Is perception informationally encapsulated? The issue of the theory-ladenness of perception. Cognitive Science, 25, 423-451.
Raftopoulos, A. (2005). Perceptual systems and a viable form of realism. In A. Raftopoulos (Ed.), Cognitive Penetrability of Perception. Hauppauge: Nova Science.
Reed, C.L., Klatzky, R.L. \& Halgren, E. (2005). What vs. where in touch: an fMRI study. NeuroImage, 25, 718-26.
Rizzolatti, G., \& Matelli, M. (2003). Two different streams form the dorsal visual system: Anatomy and functions. Experimental Brain Research, 153, 146-157.
Rossetti, Y. and L. Pisella 2002 Several 'vision for action' systems: a guide to dissociating and integrating dorsal and ventral functions. Attention and Performance 19: 62-119.
Rozzi, S., Ferrari, P. F., Bonini, L., Rizzolatti, G. \& Fogassi, L. (2008). Functional organization of inferior parietal lobule convexity in the macaque monkey. European Journal of Neuroscience 28, 1569-1588.
Schenk, T. \& McIntosh, R. D. (2010). Do we have independent visual streams for perception and action? Cognitive Neuroscience 1, 52-78.
Shams, L., Kamitani, Y., \& Shimojo, S. (2000). What you see is what you hear. Nature, 408, 788.
Siegel, S. (2011). Cognitive penetrability and perceptual justification. Nous 46, 201-222.
Spence, C. \& Driver, J. (2004, Eds.). Crossmodal Space and Crossmodal Attention, Oxford: Oxford University Press.
Stein, B.E., Stanford, T.R., Wallace, M.T., Vaughan, J.W. \& Jiang, W. (2004). Crossmodal spatial interactions in subcortical and cortical circuits. In C. Spence \& J. Driver (Eds.), Crossmodal Space and Crossmodal Attention. Oxford: Oxford University Press.
Vishton, P. (2004). Human vision focuses on information relevant to a task, to the detriment of information that is not relevant. Behavioral and Brain Sciences, 27, 53-54.
Vishton, P. \& Farbe, E. (2003). Effects of the Ebbinghaus illusion on different behaviors, Spatial Vision, 16, 377-92.
Vroomen, J., Bertelson, P., \& de Gelder, B. (2001). Auditory-visual spatial interactions. In B. de Gelder, E. de Haan \& C. Heywood (Eds.), Out of Mind, Oxford: Oxford University Press.
Watkins, S., Shams, L., Tanaka, S., Haynes, J. D., \& Rees, G. (2006). Sound alters activity in human V1 in association with illusory visual perception. NeuroImage 31, 1247-1256.
Xu, Y., O’Keefe, S., Suzuki, S. \& Franconeri, S.L. (2012). Visual influence on haptic torque perception. Perception 41, 862 - 870.

# Finding hidden types: Inductive inference in long-tailed environments 

Daniel J. Navarro (daniel.navarro@adelaide.edu.au)<br>School of Psychology, University of Adelaide, SA 5005, Australia


#### Abstract

Making inference in everyday life often requires people to make inferences about low frequency events. In the most extreme case, some types of object or event may have never been previously observed. An experiment is presented in which participants needed to infer the existence and number of unobserved event types, based solely on the frequency distribution of a set of observed events. Results indicate people's inferences are sensitive to the shape of the distribution over the observed events, even when the number of observed events and event types is held constant, and that people are able to infer abstract rules that describe entire classes of event distributions. Human inferences are shown to be similar to those made by a hierarchical Bayesian model.


Keywords: inductive inference, Bayesian cognition, frequency effects, concept learning

Imagine you are walking through the bushlands in a foreign land. You are accompanied by a local guide, who comments on the plant life around you. So far she has described 20 plants as alba, 20 plants as glabra and another 20 as eburnia. On this basis it is tempting to think that albas, glabras and eburnias are the only types of plants around, or at least the only plant types that your guide is intending to label for you. You could not be certain that this is the correct inference of course, but it seems sensible.

Contrast this with a slightly different scenario, in which your guide refers to 58 of the plants as albas, points to one example of a glabra and one example of an eburnia. Again, it is impossible to be sure what to believe, but it seems much less reasonable to conclude that these are the only three plant labels that your guide is ever going to use. Both scenarios involve 60 plants and 3 category labels, yet they do not feel equivalent.

The logic behind this intuition is relatively straightforward. In the second example, you have evidence of the existence of low-frequency types, whereas in the first example you do not. The fact that some types are relatively rare suggests that there may be other rare types that you have not yet seen. In other words, the shape of the frequency distribution plays a powerful role in shaping our inductive inferences in this problem. This is illustrated in Figure 1.

In essence, this is a category learning problem: the learner has encountered a new kind of object (the foreign plants) and is attempting to learn the extension of the category with respect to a particular feature (the labels). Viewed as a category learning problem, the different inferences drawn in the two cases are an example of a frequency effect, though of a
rather different character than the usual exemplar frequency effects. The key difference is that the effect does not pertain to a specific exemplar, but instead is an effect that pertains to the overall frequency distribution. In the first case, the learner has evidence that the frequency distribution is homogeneous: the observed exemplars have equal frequency. In the second case, the evidence implies that the frequency distribution is long-tailed, meaning that there are a small number of items that are very common, but most observations are quite rare.

## Frequency effects in categorization and choice

Exemplar frequency effects are well-established in the categorization literature: for instance, high-frequency exemplars are classified more accurately, and are judged to be more typical of the category than are low-frequency items (Nosofsky, 1988). However, although the role of item frequency is wellstudied (Nosofsky, 1988; Barsalou, 1985; Barsalou, Huttenlocher, \& Lamberts, 1998), the inductive inference described earlier is rather different to exemplar frequency effects as they are traditionally conceived. In both examples the $o b$ served frequency of blue, purple, white or any other color flower is zero, yet they differ in terms of the expected subjective frequency. That is, changing the distribution of the same set of three types (albas, glabras, eburnias) alters the expectations about the probabilities associated with as-yetunobserved types.

Frequency effects of a different kind arise in the judgment and decision making literature. In this literature the focus is on how much weight people place on low-frequency outcomes when evaluating possible options, whereas the concept learning literature tends to focus on the role of high-frequency items. Although much of the early evidence (Kahneman \& Tversky, 1979; Tversky \& Fox, 1995) suggested that people tend to overweight low-frequency events, there is some evidence indicating that this applies primarily to described frequencies, and not to experienced ones (e.g. Barron \& Erev, 2003; Hertwig, Barron, Weber, \& Erev, 2004), though much of this difference can be attributed to the different information and feedback available to participants (e.g. Rakow, Demes, \& Newell, 2008; Camilleri \& Newell, 2011). As with the category learning literature, these studies have focused on events whose observed frequency is at least one, rather than looking at the inferences people make about never-observed events.


Figure 1: Illustration of why the shape of the type-token distribution matters. Suppose you had observed several observations of types A, B, and C, all of which are equally frequent (panel a). In order to believe that there are more hidden types $\mathrm{D}, \mathrm{E}$ and F , one is required to postulate that the true distribution looks like panel b . If, however, the empirical frequencies observed were asymmetric (panel c), then in order to believe in hidden types $\mathrm{D}, \mathrm{E}$ and F , one is required only to postulate a rank-frequency distribution like the one in panel d. To the extent that the distribution in panel $b$ feels less natural than those in panels $a$, c and d, people should be expected to draw different inferences about unobserved types when presented with uniform data than when they are presented with asymmetric data.

## Learning kinds of feature distribution

A recent topic of interest in the concept learning literature is how people learn abstract rules ${ }^{1}$ that guide inductive inference in new situations (e.g. Kemp, Goodman, \& Tenenbaum, 2010; Perfors \& Tenenbaum, 2009). Applied to the current context, the idea would be that people do not merely learn that a single category shows a skewed frequency distribution over object types. Instead, people can learn that "skewness" is a property that is possessed by multiple categories. For example, if we know that the distributions of flood and fire severity are long-tailed (two categories of natural disaster for which a reasonable of data are available to people), we might also guess that the distribution of asteroid strikes (a category of natural disaster largely unknown to people) has a similar shape. One goal of the current work is to see whether people are willing to draw abstract inferences about distributional shape, and use these inferences to alter their guesses about unobserved event types.

## Overview

The goal of this paper is to investigate how people infer the existence of unobserved event types, and whether people are

[^168]

Figure 2: Sample stimulus in the pencil-and-paper version of the task. This was the first trial in the skewed condition (lower left panel in Figure 3). There are 10 tokens of the $\Im$ type, 1 token of the $\sqrt{ }$ type, and 1 token of the $\varnothing$ type. The computerized task was the same, but types were differentiated by color as well as by symbol, and the assignment of symbols was randomized.
sensitive to distributional form when doing so. The structure of this paper is as follows. An experiment is described in which participants were asked to guess how many types of marbles exist in a bag that is only partially observed, where the distribution of observations is manipulated. Human responses in this task are compared to the predictions of a hierarchical Bayesian model that learns both the number of types and the shape of the distribution over types. The implications of the results for the black swan problem that motivated the experiment are discussed.

## Experiment

## Method

Participants 101 participants ( $68 \%$ female) were recruited from the University of Adelaide community: 33 were undergraduates participating for course credit, 57 were recruited through a paid participant list, and 11 were graduate students. The 57 paid participants did a computerized version of the task, while the other 44 participants completed a pencil and paper version.
Materials \& Procedure The task took the form of a guessing game involving 7 trials. On each trial participants were shown 6,12 or 18 marbles, and told these had been drawn from a bag containing 100 marbles in total. Each marble belonged to one of several types, indicated by a symbol displayed on the marbles surface, and the participant was asked to guess how many types were represented in the full set of 100 marbles. No feedback was given as to the true number of types. Figure 2 illustrates how a set of 12 marbles belonging to 3 types was displayed.

Participants were randomly assigned into one of two conditions, referred to as the "uniform data" condition and the "skewed data" condition. The number of marbles observed and the number of types they belonged to was identical across conditions. For example, the first trial always showed 12 marbles (tokens) belonging to 3 types, and the second trial always showed 18 tokens that represented 4 types of marble regardless of condition. The conditions differed only in the frequency distribution over types. In the uniform condition, the tokens were evenly divided among types: on trial 1 , for instance, there were 4 marbles of each of the 3 types (i.e., a 4-4-4 split). In the skewed condition, the split was highly uneven, with most marbles belonging to a single type: on trial 1 , the split was $10-1-1$. The complete set of frequency dis-


Figure 3: Experimental design. Each panel shows a rank-frequency plot of the marbles on a single trial. The top row shows the type-token distribution for all 7 bags in the uniform condition. The bottom row shows the corresponding distributions for the skewed condition.

Table 1: Descriptive statistics. On the left is a summary of the observations shown to participants on each trial. The middle columns show the $5 \%$ trimmed mean response broken down by bag number and condition. The right columns show the proportion of "extrapolative" responses, namely the proportion of responses that imply the existence of at least one unobserved type.

|  |  |  | Mean |  | Extrapolation |  |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| Bag | Tokens | Types | Unif. | Skew. | Unif. | Skew. |
| 1 | 12 | 3 | 4.35 | 6.38 | 0.34 | 0.47 |
| 2 | 18 | 4 | 4.40 | 7.64 | 0.18 | 0.47 |
| 3 | 18 | 5 | 5.75 | 10.54 | 0.34 | 0.63 |
| 4 | 6 | 3 | 4.58 | 8.64 | 0.45 | 0.70 |
| 5 | 18 | 2 | 2.42 | 2.79 | 0.14 | 0.30 |
| 6 | 6 | 2 | 3.02 | 4.13 | 0.43 | 0.49 |
| 7 | 6 | 1 | 1.32 | 1.77 | 0.25 | 0.47 |

tributions used in the experiment is shown in Figure 3. Note that the final trial was identical in both conditions.
Exclusions Data from 7 participants were excluded either because they gave impossibly large or impossibly small responses, indicating that they did not understand the task. An 8th participant was excluded for omitting responses. An additional 6 participants gave sensible but qualitatively different responses ${ }^{2}$ to the remaining 87 . As such, the data from these two groups should not be aggregated, but the minority group is too small to analyze separately.

## Results

Table 1 presents an overview of the data. For all seven trials, the average number of types estimated by participants was larger in the skewed distribution condition than in the uniform distribution condition. Moreover, if we classify re-

[^169]sponses into two categories - those "extrapolative" responses in which participants inferred the existence of at least one hidden type, and responses in which they did not - we observe the same pattern. Participants were more likely to infer the existence of hidden types when the observed frequency distribution was skewed. ${ }^{3}$

To determine if the tendency to estimate more types in the skewed condition represents a significant effect, it is convenient to code the responses in terms of the number of unobserved types the participant predicted, rather than the total number of types estimated for the bag. When coded in this fashion, a response of " 3 types" on the first trial is treated the same as a " 1 type" response on the last one, because in each case the participant has indicated that he or she does not believe there are any hidden types. This has the advantage that a "0 hidden types" response always represents "no extrapolation", and all other responses represent "the extent of the extrapolation" from the sample shown to the participant.

Once the data are coded in this fashion, they can be analyzed using linear mixed effects models, which are wellsuited for describing data with a repeated measures structure. In addition to including a fixed effect of condition, the model includes a random effect of bag for each participant in order to capture individual differences in responding. ${ }^{4}$ Moreover, because the responses are skewed due to the presence of a floor effect (i.e., "zero" hidden types is a natural lower bound on responses), a Poisson error distribution was used instead of assuming normality. ${ }^{5}$ The key result is that the Wald test for the

[^170]

Figure 4: Two different biases that the model can learn, for bags containing $k=6$ types. In panel (a), the type frequencies are highly uniform ( $\alpha=100$ ), and the expected rank-frequency plot is quite flat (it becomes perfectly flat as $\alpha \rightarrow \infty$ ). In panel (b), the type frequencies are highly variable ( $\alpha=.5$ ) and the expected rankfrequency plot is very skewed.
main effect of condition was significant ( $z=3.11, p=.002$ ): participants did in fact guess that more unobserved types existed in the skewed condition than in the uniform condition.

The previous analysis demonstrated that participants in the skewed condition tended to estimate more hidden types than participants in the uniform condition. In addition to showing that this effect exists across the whole experiment, it is particularly useful to focus on bag 7, as this represents the purest test of whether people were forming theories about bags in general. A two sample Wilcoxon test ${ }^{6}$ applied to the bag 7 data revealed a significant difference ( $Z=-2.09, p=.037$ ). Despite the fact that the final bag was identical in both conditions, participants estimated more unobserved types when the preceding bags had revealed a skewed distributional shape.

## A probabilistic model of the task

This section outlines a computational analysis of the induction problem used in the experiment. The analysis relies on a probabilistic model of how bags of marbles are generated and how observations are sampled from those bags. It is related to the Bayesian concept learning model used by Kemp, Perfors, and Tenenbaum (2007), but differs in a key respect. Kemp et al. (2007) assume the learner knows the true number of object types in advance, whereas the model used here treats the number of types as an unknown quantity that must be inferred. As with most computational analyses, the model does not describe the processes people use to arrive at estimates. Rather, it provides a sensible standard against which human judgments in this task can be assessed.

## Generative model for bags

Suppose that a bag contains $k$ types of marbles, and let $\theta_{i}$ denote the probability that a particular marble will be of the $i$-th type. We may characterize the bag itself using a vector of

[^171]

Figure 5: Structure of the model. Shaded circles denote variables that had been observed by participants on or before trial $b$ of the experiment. Unshaded circles denote variables whose values must be inferred. The question asked of participants on trial $b$ corresponds to the value of $k_{b}$.
type probabilities $\boldsymbol{\theta}=\left(\theta_{1}, \ldots, \theta_{k}\right)$. A set of $n$ observed marbles from the bag $\boldsymbol{x}$ can be treated as a multinomial sample of size $n$ generated with probabilities $\boldsymbol{\theta}$. The unobserved marbles can be viewed as a multinomial sample of size $100-n$ from the same distribution. ${ }^{7}$ This model implies that, in a sample of size $n$, the learner should expect to see $n \theta_{i}$ exemplars of type $i$. As such, if $n$ and $\theta_{i}$ are both small, it is quite possible that zero exemplars of type $i$ appear in the learner's observations; it therefore becomes an unobserved type.

This formalism can be extended to provide a generative model for bags, which comes in two parts. First, the number of types $k$ is sampled from some distribution. This paper uses a binomial distribution for this purpose, though this choice is somewhat arbitrary. Second, once $k$ is sampled, the vector of type probabilities $\boldsymbol{\theta}$ is generated. A convenient choice is a Dirichlet distribution with symmetry parameter $\alpha$. This distribution is widely used by Bayesian concept learning models (e.g. Anderson, 1991; Kemp et al., 2007), and allows the learner to have strong beliefs about the shape of the frequency distribution without knowing a priori which types are more common. If $\alpha$ is small, the learner has a strong expectation that some types of marble will be frequent (Figure 4b) while others will be rare. In contrast, if $\alpha$ is large, the learner possesses a strong expectation that all types of marble should occur with approximately equal frequency (Figure 4a).

An important characteristic of this model is that it satisfies the intuitive constraint illustrated in Figure 1. The uniform distribution in panel a is the expected pattern when $k=3$ and $\alpha$ is large. The skewed distributions in panels c and d are the expected patterns produced by small $\alpha$ values, with $k=3$ and $k=6$ respectively. In contrast, although the distribution shown in panel $b$ is possible within the model, it is not highly likely under any choice of $k$ and $\alpha$.

Formally, the model is written as follows: if bags are generated with symmetry parameter $\alpha$, then we obtain the following sampling model for the observations $\boldsymbol{x}$ :

$$
k \mid \lambda \sim \operatorname{Binomial}(\lambda, n)
$$

[^172]

Figure 6: Schematic illustration of the inferences drawn by the model when the observed data are uniform. Things marked "??" refer to values that are inferred by the model rather than observed.

$$
\begin{aligned}
\boldsymbol{\theta} \mid k, \alpha & \sim \operatorname{Dirichlet}\left(\alpha \mathbf{1}^{(k)}\right) \\
\boldsymbol{x} \mid \boldsymbol{\theta} & \sim \operatorname{Multinomial}(\boldsymbol{\theta}, n)
\end{aligned}
$$

where $\mathbf{1}^{(k)}$ denotes a vector of length $k$ that contains only 1 s , and $n$ is treated as a fixed property of the experiment and is not part of the generative process over observations. The prior over $\alpha$ is assumed to be a gamma distribution.

The structure of this model as a whole is illustrated graphically in Figure 5. On trial $b$ of the experiment, the learner has access to the samples $\boldsymbol{x}_{1}, \boldsymbol{x}_{2}, \ldots, \boldsymbol{x}_{b}$ from the first $b$ bags (shaded circles). The task as stated is to estimate the the number of types in the $b$-th bag, $k_{b}$, which is one of the several unobserved variables (white circles) whose value is inferred via Bayesian inference.

## Learning abstract rules about bags

One of the important patterns in the empirical data is the fact that participants give different responses to bag 7 in the two conditions. The model reproduces this pattern because the symmetry parameter $\alpha$ describes an abstract regularity that attaches to all bags. As such, the model is able to learn the value $\alpha$ across trials. If the model is shown several samples with uniform distributions over observed types, the model will gradually raise the value of $\alpha$. The value of $\alpha$ tends to decrease when the observed type frequencies are consistently non-uniform.

The consequences of this learning are illustrated in Figures 6 and 7. In the bottom row of Figure 6, the observed samples are evenly split across types, so the model infers a large value for $\alpha$ (top row). The most plausible way to have uniform distributions and remain consistent with the raw data is to have no unobserved types (middle row). Contrast this with the skewed-data scenario in Figure 7. Here the model infers a small value for $\alpha$ and assumes that all of the frequency distributions are also skewed (middle row). Skewed distributions over types imply that at least some types are low probability, so it is entirely plausible to believe that unobserved types exist. As a consequence, the model makes different predictions about the final bag in Figure 7 than it does for the exact same bag in Figure 6.


Figure 7: Schematic illustration of the inferences drawn by the model when the observed data are skewed. Things marked "??" refer to values that are inferred by the model rather than observed.

## Model implementation

Although the model specifies many latent variables, the quantity of interest for the $b$-th bag is $P\left(k_{b} \mid \boldsymbol{x}_{1}, \boldsymbol{x}_{2}, \ldots, \boldsymbol{x}_{b}\right)$, the posterior probability that bag $b$ contains $k_{b}$ types of marbles, given all of the samples observed so far. This posterior probability cannot be computed analytically: given this, the model was implemented in JAGS (version 3.1.0) and numerical estimates were obtained using Markov chain Monte Carlo. For each bag $b$, samples were drawn from the joint posterior distribution over all latent variables, and these samples were used to approximated the posterior probabilities of interest.

Because the data presented to participants is different on each trial, fitting the model to the data requires 14 separate model runs. Each of these 1750 model runs involved drawing 1000 samples from the posterior distribution over $k$ after a burn in of 1000 samples. Moreover, because the model predictions depend on the choice of priors, a grid search using 125 different parameter sets was tried. The value of $\lambda$ was varied from .05 to .25 , and the shape and scale parameters for the prior over $\alpha$ were both varied from 1 to 5 . The best performing parameter values correspond to a prior over $k$ that is $\operatorname{Binomial}(0.15,100)$ for all bags, and a prior over $\alpha$ that is $\operatorname{Gamma}(4,2)$.

## Modeling human data

The model predictions are generally in close agreement with human responses, but there are some differences. The main one is that the model never generates extremely large estimates: human participants occasionally guessed that a bag contained more than 5-6 hidden types, whereas guesses of this kind do not appear at all among the 1000 samples from the model posterior. In other words, although the model produces a distribution over responses that is qualitatively in agreement with human responses, it contains fewer very large values. This difference appears to be due to the fact that the model does not incorporate individual differences: it assumes that all participants have the same priors and rely on the same probabilistic assumptions about the task. ${ }^{8}$ Nevertheless, there

[^173]

Figure 8: Comparing the model fits against human responses. The Pearson correlation between model and human is $0.89(p<.001)$.
are individual differences in how people solve the task. There are a few people who consistently estimate large numbers of hidden types, but most do not. This makes it difficult to directly compare model estimate of $k$ against the raw human responses.

A simple solution to the problem is to compare the qualitatively important distinction in the task, namely whether or not a particular response implies the existence of at least one hidden type. That is, instead of fitting the model to the mean number of types estimated by participants (middle columns in Table 1), it is fit to the proportion of human responses in which the number of estimated types was larger than the number of types revealed in the same (right columns in Table 1). These responses are "extrapolative" in that they indicate that the participant has extrapolated beyond the observed data and guessed that there exists at least one hidden type.

Figure 8 plots model estimate of the probability that a bag contains at least one hidden type against the proportion of extrapolative responses in the empirical data. Circles denote bags in the uniform condition, and diamonds represent bags in the skewed condition, and the text denotes bag number. The correlation between model predictions and human responses is $0.89(p<.001)$ for the best fitting parameter values. However, the model fit is robust: the average correlation across all 125 parameter sets was 0.84 , never fell below 0.66 , and was significant at $p<.01$ in all cases.

## Discussion

The close agreement between model predictions and human responses implies that people are sensitive to the information contained in the shape of the distribution of events they have experiences when making inferences about types of events they have never seen. Moreover, the fact that systematic differences existed on the final trial of the experiment, and that these differences are captured via a hierarchical Bayesian
model, implies that people are able to use the information from one context (i.e., one bag) to inform the inferences they draw in another one.

One potential extension to this work is to is to consider the role of information search. In the current study, the number of observations sampled from each bag was fixed by the experimenter. However, in many real world decision making problems, people have some degree of control over how much information they collect before making choices. It seems plausible to think that, when the true event distribution is very uneven, people will adopt a very different search strategy than if the frequency distribution is uniform. As such, the constraint that the number of types and tokens observed needed to be matched across experimental conditions, although important from a methodological perspective, may obscure one of the key differences in how people make inferences and choices more generally. In preliminary work investigating this question, we have found some evidence that information search process is indeed influenced by distributional shape, but this is work in progress.

## Acknowledgements

This research was supported by ARC grant DP110104949, with salary support from ARC grant FT110100431. I thank Frances Nettle and Natalie May for their assistance with data collection, and Amy Perfors and Nancy Briggs for many helpful discussions.

## References

Anderson, J. R. (1991). The adaptive nature of human categorization. Psychological Review, 98(3), 409429.
Barron, G., \& Erev, I. (2003). Small feedback-based decisions and their limited correspondence to description-based decisions. Journal of Behavioral Decision Making, 16(3), 215233.
Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. Journal of Experimental Psychology: Learning, Memory, and Cognition, 11(4), 629.
Barsalou, L. W., Huttenlocher, J., \& Lamberts, K. (1998). Basing categorization on individuals and events. Cognitive Psychology, 36(3), 203272.
Camilleri, A. R., \& Newell, B. R. (2011). When and why rare events are underweighted: A direct comparison of the sampling, partial feedback, full feedback and description choice paradigms. Psychonomic bulletin \& review, $18(2), 377384$.
Hertwig, R., Barron, G., Weber, E. U., \& Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. Psychological Science, 15(8), 534539.
Kahneman, D., \& Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica: Journal of the Econometric Society, 263291.
Kemp, C., Goodman, N. D., \& Tenenbaum, J. B. (2010). Learning to learn causal models. Cognitive Science, 34(7), 11851243.
Kemp, C., Perfors, A., \& Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical bayesian models. Developmental science, 10(3), 307321.
Nosofsky, R. M. (1988). Similarity, frequency, and category representations. Journal of Experimental Psychology: Learning, Memory, and Cognition, 14(1), 54.
Perfors, A. F., \& Tenenbaum, J. B. (2009). Learning to learn categories. In Proceedings of the 31st annual conference of the cognitive science society (pp. 136-141). Austin, TX: Cognitive Science Society.
Rakow, T., Demes, K. A., \& Newell, B. R. (2008). Biased samples not mode of presentation: Re-examining the apparent underweighting of rare events in experience-based choice. Organizational Behavior and Human Decision Processes, 106(2), 168179.

Tversky, A., \& Fox, C. R. (1995). Weighing risk and uncertainty. Psychological review, 102(2), 269.

# Grounded Spatial Belief Revision 

# Jelica Nejasmic (jelica.nejasmic@ psychol.uni-giessen.de) 

## Leandra Bucher (leandra.bucher@psychol.uni-giessen.de)

Markus Knauff (markus.knauff@psychol.uni-giessen.de)<br>Justus-Liebig University, Experimental Psychology and Cognitive Science, Otto-Behaghel-Str. 10F, 35394 Giessen, Germany


#### Abstract

Beliefs frequently undergo revisions, especially when new pieces of information are true but inconsistent with current beliefs. In previous studies, we showed that spatial belief revision is often guided by the functional asymmetry between the reference object and the located objects of the spatial relation. Here we first draw a connection between spatial belief revision and grounded cognition. In two experiments, we explored whether imagined physical properties of objects influence which object is relocated and which remains at its initial position. Participants mentally revised beliefs about the arrangement of objects which could be envisaged as small and large (Experiment 1) or easy to move and difficult to move (Experiment 2). The results show that (1) small objects are more often relocated than larger objects and (2) easy to move objects are faster relocated than difficult to move objects. The findings are in line with the idea of grounded cognition.


Keywords: Spatial cognition, grounded cognition, mental models, belief revision, spatial reasoning

## Introduction

Imagine you have a date with a friend in a foreign city. He described to you how to come to the meeting point: "When you get off the train, you will see the kiosk to the left of you, and an ice cart to the right of you. To the left of the kiosk, I will wait for you." This description is compatible with the following mental model:
Kiosk - I - ice cart

Almost arriving you get a call from your friend who tells you: "I made a mistake. The kiosk is to the right of the ice cart." On which side is your friend waiting for you? In fact there are two possibilities:

$$
\begin{align*}
& \text { I - ice cart - kiosk }  \tag{1}\\
& \text { Ice cart - kiosk - I }
\end{align*}
$$

In everyday life, we are often confronted with such problems. People describe how to find certain objects and then realize that the description is wrong ("I left your key on the kitchen table, but it is actually on the table in the living room"); someone describes how to find a certain place in a foreign city and on your way, you realize that his
description was wrong; your partner describes where he parked your car, but it is parked somewhere different, and so on.

All this has to do with the field of "belief revision". Researchers in this field explore how people change their mind in the light of new contradicting information. The experimental studies mostly used conditional reasoning problems in which an inconsistency arises between a fact, contradicting a valid conclusion, and the conditional and categorical premises. Within this research, psychologists were able to show that belief revision is affected by many factors, including asymmetries between particular facts and general laws (Revlis, Lipkin, \& Hayes, 1971), conditional and categorical premises (Elio \& Pelletier, 1997; Dieussaert, Schaeken, De Neys, \& d’Ydewalle, 2000; Girotto, JohnsonLaird, Legrenzi, \& Sonino, 2000; Revlin, Cate, \& Rouss, 2001), major and minor premises (Politzer \& Carles, 2001), and reliable and unreliable information sources (Wolf, Rieger, \& Knauff, 2012).

The present work is part of our current attempt to extend the cognitive research on human belief revision to the area of spatial reasoning. Our main motivation is that (1) spatial inferences are ubiquitous in our daily life (Goodwin \& Johnson-Laird, 2005; Knauff, 2013), (2) reasoning with spatial relations is often easier than reasoning with conditionals (Johnson-Laird \& Byrne, 1991; Knauff, 2007), and that space is one of the most fundamental dimensions of our physical and psychological reality (Gattis, 2001; Knauff, 1999, 2013). In our previous work, we have identified three main principle of spatial belief revision (Knauff, Bucher, Krumnack, \& Nejasmic, 2013):

1. Spatial reasoning relies on mental models. A mental model is a unified representation of what is true if the premises are true. Reasoners use the meaning of assertions and general knowledge to construct single models of possibilities compatible with these assertions. Spatial relations are not represented explicitly in a propositional format but rather they are inherent in the model and thus can be (and must be) 'read off' from the model by mental inspection processes (Johnson-Laird \& Byrne, 1991; Polk \& Newell, 1995; Goodwin \& Johnson-Laird, 2005).
2. Spatial belief revision relies on the revision of mental models. If newly available information is inconsistent
with the current model (and new information must be taken for granted), a model revision is necessary to establish consistency. The revision process relies on local transformations in which tokens in the model are moved to new positions. If not all available information can be true at the same time, people "decide" which of the information to retain and which one to discard (Bucher, Krumnack, Nejasmic, \& Knauff, 2011; Krumnack, Bucher, Nejasmic, \& Knauff, 2011; Bucher \& Nejasmic, 2012).
3. The model revision process is sensitive to the functional asymmetry between the "reference object" (RO) and the "located object" (LO). For instance, in the statement " A is to the right of C ", the C is the RO and the A the object that is located in relation to the RO. To regain consistency the LO of the inconsistent statement seems to be relocated within the initial constructed mental model (Bucher et al., 2011; Krumnack et al., 2011; Bucher \& Nejasmic, 2012; Knauff et al., 2013). The distinction has been made by several psychologists and linguists (Miller \& Johnson-Laird, 1972; Talmy, 1983; Landau \& Jackendoff, 1993). The common idea of all these theories is that a spatial relation refers to the position of a particular object in focus relative to another object or area (Tenbrink, Andonova, \& Coventry, 2011).

In the present work we mainly focus on the third principle (LO-RO-asymmetry) and combine it with the idea that cognitive processes are not only abstract symbolic manipulations but grounded in perceptual, motoric, or emotional experience (for an overview, see de Vega, Glenberg, \& Graesser, 2008). Imagine, for instance, you are helping a friend to move into a new apartment. You have to carry many things (sofas, tables, books, porcelain, washing machine, hopefully no piano, etc.) from his old apartment to the furniture truck and then later from the furniture truck into the new apartment. If you do that, it is very likely that you try to avoid carrying heavy objects and prefer to move objects which are less heavy. But, does this heaviness also affect how you think about the location and relocation of objects? In principle, the weight of objects should not matter if we just mentally move objects in a mental model from one position to another. We do not have to carry the objects physically; so why should their weight matter? On the other hand, the theory of grounded cognition claims that such physical properties have an effect on how we think. The process of thought is a mental simulation of bodily experiences and therefore the weight of objects should affect how we mentally process the information from a spatial reasoning problem (Glenberg \& Robertson, 1999; Kaschak \& Glenberg, 2000; Kaschak \& Glenberg, 2002; Rinck \& Bower, 2004; Bergen \& Chan, 2005; Pulvermüller, 2005).

## Two experiments of spatial belief revision and grounded cognition

In the following, we present two experiments on grounded spatial belief revision. In the experiments, participants received spatial information about the location of small or large objects (Experiment 1) or easy to move or difficult to move objects (Experiment 2), and then have to revise their initial model in the light of new contradicting information. Is this revision process affected by the size or movability of the objects that can be relocated? Do people prefer to relocate the smaller and easy to move object as the theory of grounded cognition suggests? We present a pilot study, because we did not want to use the actual physical mass and size of objects but rather how they are psychologically perceived and represented (although that should highly correlate). Then we report two experiments: in experiment 1, the size of the objects was varied; in experiment 2 the movability was varied. These factors were combined with the role of the objects as being the RO or LO of the relation in the newly available information. In our previous experiments, we found a strong preference of relocation the LO because the RO is considered less flexible. Can the size or movability of the envisaged objects overwrite this general preference? In the last part of the paper we discuss our findings and draw some general conclusion about grounded cognition and spatial belief revision.

## Pilot study

Our first task was to define the set of objects to use in our experiments. To select the set of "large" and "small" objects, we developed a questionnaire with 64 objects. 46 participants rated the size of the objects on a five-point scale with the poles "very small" and "very large". Then they rated the same objects regarding movability on a five-pointscale from "easy to move" to "difficult to move". The order of objects was randomized. For the analysis, we computed the means over the group of participants and selected the objects with the lowest and highest mean ratings for size and movability. The objects are presented in Table 1. These objects were used for the following experiments.

Table 1: Objects used in the experiment according to their property

| small | large | easy to <br> move | difficult to <br> move |
| :---: | :---: | :---: | :---: |
| screen | power mast | wheelchair | pillar |
| vase | bridge | bicycle | counter |
| printer | railway station | carriage | gravestone |
| post | high rise | scooter | oven |
| lamp | spire | barrow | hydrant |

## Experiment 1: Small vs. large objects

## Method

Participants. 21 students from the University of Giessen (9 male; age: $M=22.86 ; S D=5.27$ ) were tested individually. They gave written informed consent and received course credits for their participation.

Materials, design, and procedure. Each participant solved 48 revision problems. Six practice trials (not analyzed) preceded the experimental trials. All stimuli were generated and presented using Superlab 4.0 (Cedrus Corporation, San Pedro, CA, 1999) with a RB-530, on a standard personal computer with a 19 " monitor. Number of correct decisions and corresponding revision times were recorded.

The structure of the problems was as follows: First, participants received sequentially two statements, also called premises (1, 2), which described the spatial relation between three objects, for example:

> "A is to the left of $B "$
> "B is to the left of C"

From these two premises the participants inferred that the three objects are in the arrangement $\mathrm{A}-\mathrm{B}-\mathrm{C}$. They did that by choosing one of two arrangements (correct arrangement/correct arrangement mirrored) that were presented on the screen.

In the next step, participants were confronted with an additional statement, e.g., ' A is to the right of C '. This is the critical point in time where participants in our experiments had to realize that something must be wrong with their initial model about the layout of the three objects. Not all three statements can be true at the same time because the third statement contradicts the logical inference from the first two premises. Participants were told that the third statement is irrefutably true so they could not ignore the third statement. The only option was to decide which one of the first two premises may be abandoned. If the first premise is discarded this results in the arrangement $\mathrm{B}-\mathrm{C}-\mathrm{A}$; if the second premise is discarded this results in the arrangement C - A - B. It is essential to see that the first revision strategy corresponds to the LO-Relocation, whereas the second revision strategy corresponds to the RO-Relocation. All statements used the relation "left of" and "right of" and were presented sequentially. Positions of the arrangements as well as the relations were counterbalanced across the experiment.

To study the effect of object size the terms, A, B, C were instantiated with the small and large objects from Table 1. To boost the possible effect of object properties we integrated a "you" into the problems. We expected that this would foster the perspective taking and that the participants are therefore even more sensitive to the object properties.
(1) premise: "The vase is to the left of you."
(2) premise: "You are to the left of the spire."

Initial model: Vase - you - spire
Inconsistent fact: "The spire is to the left of the vase."
Examples of revised orders:
Spire - vase - you ("relocation of the large object") vs. You - spire - vase ("relocation of the small object")

Participants received instructions on the computer screen. They were instructed to imagine an arrangement determined by the premises and subsequently to choose the respective arrangement (on the screen) by pressing the corresponding button. Afterwards participants had to decide whether or not the presented fact is consistent with this model. For the problems with a fact contradicting the initial model (which was the case in half of the problems), participants were asked to revise the arrangement and to define the revised arrangement by pressing the corresponding button.

## Results and discussion

Mean percentage rate of correctly constructed models was $98 \%(S D=2.15)$ and in $94 \%(S D=8.54)$ of the inconsistent problems participants correctly identified the inconsistency between the initial model and the contradictory fact. Erroneous trials were excluded from further analysis.

An ANOVA with the factors object size (small vs. large) $\times$ person's position (leftmost, middle, rightmost) was conducted for the revision choice and the revision times. Level of significance was 5\%.

This ANOVA revealed a significant main effect of object property for revision choices $\left[F(1,20)=8.21 ; p<.05 ; \eta^{2}=\right.$ .29]. The main effect of person's position and the interaction were non-significant ( $p>.87$ ).

T-tests revealed that choosing of revised arrangements were based significantly more often on relocations of small objects ( $M=56 \%, S D=8.95$ ) compared to large objects ( $M$ $=44 \%, S D=8.95 ; t(20)=3.09 ; p<.01$ ) (see Fig. 1).

Results of the ANOVA for revision times were nonsignificant (all $p \mathrm{~s}>.53$ ). Implicitly, the analyses also showed that the differences between LO and RO were less important than the differences between small and large objects.

Here is an example problem:


Figure 1: Relative frequency (in \%) of model selections based on the relocation of small and large objects. Error bars show standard errors.

Our results show that physical properties of objects have an effect on how people revise their existing belief about the arrangement of objects in space. People have a strong tendency to relocate those objects that would also be easier to move physically. This finding agrees with the grounded cognition approach and is more difficult to explain based on purely symbolic cognitive theories. The finding also agrees with the mental model theory of reasoning, in which people reason spatially by constructing, inspecting and varying spatial mental models that mirror the situation described in the premises (Knauff, Rauh, \& Schlieder, 1995; Ragni, Knauff, \& Nebel, 2005; Rauh, Hagen, Kuss, Knauff, Schlieder, \& Strube, 2005; Nejasmic, Krumnack, Bucher, \& Knauff, 2011). If such a model is then contradicted by a new fact, people try to revise the model by local transformations. In fact, they move the objects in the model around to obtain a model consistent with the newly available information. We could show these mental operations are affected by the imagined physical properties of the objects. With the next experiment, we tried to replicate this effect with objects which are easy or difficult to move. The question is again: does object property affect reasoning and belief revision? Is the physical challenge related to a difficult movable object somehow reflected when we manipulate it mentally?

## Experiment 2: Easy and difficult to move objects

## Method

Participants. 24 students from the University of Giessen (5 male; age: $M=22.71 ; S D=6.60$ ) were tested individually. They gave written informed consent and received course credit for their participation. Data from one participant were excluded from the analysis due to a technical problem.
Materials, procedure, and design. The instructions on the computer and the procedure were the same as in experiment

1. The only difference between experiment 1 and 2 was the object property. Again, we manipulated two factors: object mobility (easy and difficult to move objects) and person's position ("you") in the spatial arrangement (leftmost, middle, and rightmost) as the independent within subject factors. Again, the dependent variables are revision choice and revision times.

## Results and discussion

As in experiment 1, participants selected correct arrangements in more than $95 \%(M=97 \%, S D=3.70)$ of the cases. They detected inconsistencies between the initial constructed arrangements and the contradictory facts in 94\% ( $S D=4.97$ ), correctly. Erroneous trials were excluded from further analysis.

An ANOVA with the factors object property (easy and difficult to move objects) $\times$ person's position ("you") in the spatial arrangement (leftmost, middle, and rightmost) was conducted for the revision choice and the respective revision times. Level of significance was $5 \%$.

The ANOVA for revision times revealed a significant main effect of object property $\left[F(1,16)=4.76 ; p<.05 ; \eta^{2}=\right.$ .23]. The main effect of person's position and the interaction were non-significant ( $p s>.23$ ). One-tailed t-tests revealed that participants needed less time for a revision based on a relocation of an easy to move object ( $M=2956.04 \mathrm{~ms}, S D=$ 1257.80) compared to a difficult to move object ( $M=$ $4189.31 \mathrm{~ms}, S D=714.77 ; t(22)=1.93 ; p<.05)$ (see Fig. 2). In this experiment, the ANOVA for revision choices were non-significant ( $p>.41$ ). Although participants did not show a clear preference for a relocation of an easy to move object compared to a difficult to move object, they needed more time to establish consistency, when the revision was based on a relocation of a difficult to move object.


Figure 2: Revision times (in ms) based on the relocation of the respective object. Error bars show standard errors.

## General discussion

We reported two experiments on the connection of spatial belief revision and grounded cognition. We showed that (1) small objects are more often relocated than larger objects
and (2) easy to move objects are faster relocated than objects which are difficult movable. The findings are in line with the idea of grounded cognition. According to this approach, the mind is embodied, and thus cognitive processes must be grounded in perceptual, motoric, or emotional experience (for an overview, see de Vega et al.,2008). From this point of view, human thought is almost exclusively based on perceptual simulations and modalityspecific representations (Barsalou, 2008, 2010).

Our results suggest that object properties effect spatial belief revision. The (imagined) properties of objects can overwrite the general preference for LO-relocation (experiment 1) and the time it takes to revise a mental model can be modulated by different object properties (experiment 2). These findings indicate that the physical effort that would be necessary when an object is actually relocated is reflected in revision preferences and revision speed. Moreover, the effort for different object properties (like size and movability) was reflected by different psychological measures in our experiment. On the one hand, size affected the frequency with which objects were mentally relocated. Participants preferably changed the position of small objects. Large objects were preferably left in the same position. What happens when we transfer this result to the physical world? It basically means that a vase is relocated preferably to a house. That makes sense, given that we indeed relocate vases more often and more easily than houses. On the other hand, movability modulated the time individuals needed for the revision process. This also agrees with real actions in the physical world. We might not have preferences in relocating ovens compared to wheelchairs, but it is more time consuming to move an oven than a wheelchair. Consequently, participants needed more time relocating heavy ovens than mentally "pushing" the wheelchair.

Of course, often size and movability (and many more properties of objects) are confounded. That was also the case in our experiments. However, we made an attempt to disentangle size and movability. Our results suggest that we were successful in doing so. An important corollary from our studies is that different properties affect different dependent measurements. This might be of concern in further experiments in this field.

We are aware that our findings are too weak to make a strong case for the grounded cognition approach in reasoning and belief revision. This is in particular important with regard to findings suggesting that effects of grounded cognition are more based on experimental demand characteristics considering the work of During et al. (2009). Another critical point is that, for instance, research on wayfinding and navigation shows that some kind of spatial representations have an amodal representational format, which does not agree with the embodied cognition approach (Giudice, Klatzky, \& Loomis, 2009). In that sense, our experiments can only be a very first step to draw a new connection between two fields that still work in isolation. However, another corollary from our findings is that the
theory of grounded cognition and the theory of mental models fit well together. The theory of mental models also assumes that we reason by mentally simulating what might be the case. However, the model theory so far has focused on the simulation of spatial relations between objects. Today, the vast majority of researchers consider the model theory to be the empirically best supported theory of human spatial inference (Vandierendonck, Dierckx, \& de Vooght, 2004; Goodwin \& Johnson-Laird, 2005; Knauff, 2009, 2013; for an exception see: van der Henst, 2002). The present work shows that the idea of a mental simulation in mental models can also incorporate other physical features. We will continue to investigate this connection between mental models and grounded cognition in future experiments.

## Acknowledgments

This work was supported by grants of the Deutsche Forschungsgemeinschaft (DFG) to M. Knauff and B. Nebel KN 465/6-1 and KN 465/6-2. We thank Isabell Tapia-Leon for assistance with the experiments.

## References

Barsalou, L. W. (2008). Grounded cognition. Annual Review of Psychology, 59, 617-645.
Barsalou, L. W. (2010). Grounded Cognition: Past, Present, and Future. Topics in Cognitive Science, 2, 716-724.
Bergen, B., \& Chang, N. (2005). Embodied Construction Grammar in simulation-based language understanding. In Östman, J.-O. \& Fried, M. (eds.), Construction Grammar(s): Cognitive and Cross-Language Dimensions. Johns Benjamins.
Bucher, L., Krumnack, A., Nejasmic, J., \& Knauff, M. (2011). Cognitive processes underlying spatial belief revision. In L. Carlson, C. Hölscher, \& T. Shipley (Eds.), Proceedings of the 33 rd annual conference of the Cognitive Science Society (pp. 3477-3483). Austin, TX: Cognitive Science Society.
Bucher, L., \& Nejasmic, J. (2012). Relocating multiple objects during belief revision. In C. Stachniss, K. Schill, \& D. Uttal (Eds.), Lecture notes in artificial intelligence: Spatial cognition (pp. 476-491). Berlin, Germany: Springer.
De Vega, M., Graesser, A. C., \& Glenberg, A. M. (2008). Reflecting on the debate. In M. de Vega, A. M. Glenberg, \& A. C. Graesser (Eds.), Symbols and embodiment: Debates on meaning and cognition (pp. 397-440). Oxford, U.K.: Oxford University Press.

Dieussaert, K., Schaeken, W., De Neys, W., \& d'Ydewalle, G. (2000). Initial belief state as a predictor of belief revision. Current Psychology of Cognition, 19, 277-286.
Durgin, F. H., Baird, J., Greenburg, M., Russell, R., Shaughnessy, K., \& Waymouth, S. (2009). Who is Being Deceived? The Experimental Demands of Wearing a Backpack. Psychonomic Bulletin and Review 16 (5): 964968.

Elio, R., \& Pelletier, F. J. (1997). Belief change as propositional update. Cognitive Science, 21, 419-460.
Gattis, M. (Ed.). (2001). Spatial schema and abstract thought. Cambridge, MA: MIT Press.
Girotto, V., Johnson-Laird, P. N., Legrenzi, P., \& Sonino, M. (2000). Reasoning to consistency: How people resolve logical inconsistencies. In J. Garcia-Madruga, M. Carriedo, \& M. J. Gonzalez-Labra (Eds.), Mental models in reasoning (pp. 83-97). Madrid, Spain: Universidad Nacional de Educatión a Distancia.
Giudice, N. A., Klatzky, R. L., \& Loomis, J. M. (2009). Evidence for amodal representations after bimodal learning: Integration of haptic-visual layouts into a common spatial image. Spatial Cognition and Computation, 9, 287-304.
Glenberg, A. M., \& Robertson, D. A. (1999). Indexical Understanding of instructions. Discourse Processes, 28, 1-26.
Goodwin, G. P., \& Johnson-Laird, P. N. (2005). Reasoning with relations. Psychological Review, 112, 468-493.
Johnson-Laird, P. N., \& Byrne, R. (1991). Deduction. Hove, UK: Erlbaum.
Johnson-Laird, P. N., Legrenzi, P., \& Girotto, V. (2004). How we detect logical inconsistencies. Current Directions in Psychological Science, 13, 41-45.
Kaschak, M. P., \& Glenberg, A. M. (2000).Constructing meaning: The role of affordances in grammatical constructions in sentence comprehension. Journal of Memory and Language, 43, 508-529.
Kaschak, M. P., \& Glenberg, A. M. (2002). Grounding language in action. Psychonomic Bulletin \& Review, 9, 558-565.
Knauff, M. (1999). The cognitive adequacy of Allen's interval calculus for qualitative spatial representation and reasoning. Spatial Cognition and Computation, 1, 261290.

Knauff, M. (2007). How our brains reason logically. Topio, 26, 19-36.
Knauff, M. (2009). A neuro-cognitive theory of deductive relational reasoning with mental models and visual images. Spatial Cognition and Computation, 9, 109-137.
Knauff, M. (2013). Space to reason. Cambridge, MA: MIT Press.
Knauff, M., Bucher, L., Krumnack, A., \& Nejasmic, J. (2013). Spatial belief revision. Journal of Cognitive Psychology, 1, 1-10.
Knauff, M., Rauh, R., \& Schlieder, C. (1995). Preferred mental models in qualitative spatial reasoning: A cognitive assessment of Allen's calculus. In Proceedings of the $17^{\text {th }}$ Annual Conference of the Cognitive Science Society (pp. 200-205). Mahwah, NJ: Lawrence Erlbaum.
Krumnack, A., Bucher, L., Nejasmic, J., \& Knauff, M. (2011). Efficiency and minimal change in spatial belief revision. In L. Carlson, C. Hölscher, \& T. Shipley (Eds.), Proceedings of the 33rd annual conference of the Cognitive Science Society (pp. 2270-2275). Austin, TX: Cognitive Science Society.

Landau, B., \& Jackendoff, R. (1993). "What'" and "where" in spatial language and spatial cognition. Behavioral and Brain Sciences, 16, 217-265.
Miller, G. A., \& Johnson-Laird, P. N. (1976). Language and perception. Cambridge, MA: Harvard University Press.
Nejasmic, J., Krumnack, A., Bucher, L., \& Knauff, M. (2011). Cognitive processes underlying the continuity effect in spatial reasoning. In L. Carlson, C. Hölscher, \& T. Shipley (Eds.), Proceedings of the 33rd annual conference of the Cognitive Science Society (pp. 11271132). Austin, TX: Cognitive Science Society.

Politzer, G., \& Carles, L. (2001). Belief revision and uncertain reasoning. Thinking and Reasoning, 7, 217-234.
Polk, T. A., \& Newell, A. (1995). Deduction as verbal reasoning. Psychological Review, 102, 533-566.
Pulvermüller, F. (2005). Brain mechanisms linking language and action. Nature Reviews Neuroscience, 6, 576-582.
Ragni, M., Knauff, M., \& Nebel, B. (2005). A Computational Model for Spatial Reasoning with Mental Models. In Proceedings of the $27^{\text {th }}$ Annual Conference of the Cognitive Science Society, (pp. 1797-1802). Mahwah, NJ: Erlbaum.
Rauh, R., Hagen, C., Kuss, T., Knauff, M., Schlieder, C., \& Strube, G. (2005). Preferred and alternative mental models in spatial reasoning. Spatial Cognition and Computation, 5(2\&3), 239-269.
Revlin, R., Cate, C. L., \& Rouss, T. S. (2001). Reasoning counterfactually: Combining and rending. Memory and Cognition, 29, 1196-1208.
Revlis, R., Lipkin, S. G., \& Hayes, J. R. (1971). The importance of universal quantifiers in a hypothetical reasoning task. Journal of Verbal Learning and Verbal Behavior, 10, 86-91.
Rinck, M., \& Bower, G H. (2004). Goal-based accessibility of entities within situation models. In The Psychology of Learning and Motivation: Advances in Research and Theory, 44, ed. BH Ross, pp. 1-33. New York: Elsevier Sci.
Talmy, L. (1983). How language structures space. In H. L. Pick \& L. P. Acredolo (Eds.), Spatial orientation: Theory, research and application (pp. 225-282). London, UK: Plenum.
Tenbrink, T., Andonova, E., \& Coventry, K. R. (2011). Spatial strategies in the description of complex configurations. Discourse Processes, 48, 237-266.
Van der Henst, J.-B. (2002). Mental model theory versus the inference rule approach in relational reasoning, Thinking \& Reasoning, 8(3), 193 - 203
Vandierendonck, A., Dierckx, V., \& De Vooght, G. (2004). Mental model construction in linear reasoning: Evidence for the construction of initial annotated models. Quarterly Journal of Experimental Psychology, 57, 1369-1391.
Wolf, A. G., Rieger, S., \& Knauff, M. (2012). The effects of source trustworthiness and inference type on human belief revision. Thinking and Reasoning, 18, 417-440.

# Desirable Difficulty in Learning: A Computational Investigation 

Aida Nematzadeh, Afsaneh Fazly, and Suzanne Stevenson<br>Department of Computer Science<br>University of Toronto<br>\{aida,afsaneh,suzanne\} @cs.toronto.edu


#### Abstract

Certain difficulties of a word learning situation can promote long-term learning, and thus are referred to as "desirable difficulties". We use a computational modelling approach to examine the possible explanatory factors of the observed patterns in a cross-situational word learning experiment. Our results suggest that the within-trial ambiguity and the presentation duration of each trial in addition to other distributional characteristics of the input (experimental stimuli) may explain these results. Our findings also emphasize the role of computational modelling in understanding empirical results.


## Introduction

One of the important questions in language acquisition is how people learn the mappings between words and their meanings (Quine, 1960). A number of mechanisms and approaches have been proposed in an attempt to address this question (e.g., Tomasello, 1992; Pinker, 1989). A widelydiscussed mechanism is cross-situational learning, in which people learn word meanings by gathering evidence from various exposures of words in different situations. Recent word learning experiments also confirm that both adults and children keep track of cross-situational statistics across individually ambiguous learning trials, and infer the correct wordmeaning mappings even in highly ambiguous conditions (Yu \& Smith, 2007; Smith \& Yu, 2008). These experiments have gained popularity in recent years (e.g., Yurovsky \& Yu, 2008; Vlach, Sandhofer, \& Kornell, 2008), and provide opportunities for further investigating the observed patterns in naturalistic word learning.

One interesting aspect of word learning that can be studied in such experiments, is its interaction with other cognitive processes such as memory and attention. An example is the experiments of Vlach et al. (2008) on children, which examine the spacing effect, i.e., the observation that people generally learn better when the presentations of the items to be learned are distributed (spaced) over a period of time. This and other similar patterns in human learning are referred to as "desirable difficulties": Although a more difficult learning situation may hinder short-term recall of learned material, it may promote long-term retention.

In this work, we use a computational model to shed light on one such case of an observed "desirable difficulty" in crosssituational word learning, studied by Vlach and Sandhofer (2010). Notably, Vlach and Sandhofer (2010) attribute their findings to desirable difficulties in learning, but do not provide an explanation of why and how the sort of difficulty they focus on facilitates long-term retention of the learned words. Computational modelling enables us to investigate the precise
learning mechanisms, and the variations in the input conditions, that might explain these findings. We first introduce our computational model of cross-situational word learning, and then explain and analyze the experimental data and results of Vlach and Sandhofer (2010) in the context of our model. Finally, we describe the way we simulate these experiments using our model, and how this enables us to examine the role of several different factors in the observed pattern of word learning.

## The Computational Model

In this section, we present our computational model of word learning that was first published in Nematzadeh, Fazly, and Stevenson (2012a). Our model builds on the word learning model of Fazly, Alishahi, and Stevenson (2010), which takes an incremental approach in learning probabilistic associations between words and their meanings. In Nematzadeh et al., we integrated new functionality into this model to capture forgetting (i.e., an effect of memory) and attention to novelty. Our proposed model accounts for several observed patterns of the spacing effect in children and adults, in which experimental subjects learn presented items better when they are spaced apart in time, than when they are shown in immediate succession. We provide a brief overview of the model before turning to modelling of other kinds of "desirable difficulties."

## Learning from an Input Pair

Our model learns about the meaning of words by incrementally processing a corpus that contains a sequence of utterances paired with a semantic representation of a scene, which is the hypothetical perception of a learner upon hearing the utterance. Each input to the model pairs a set of words (the representation of the utterance) with a set of semantic features (the representation of the scene), as in:

```
Utterance: { she,drinks, milk }
Scene: { ANIMATE, PERSON, FEMALE, CONSUME,
    DRINK, SUBSTANCE, FOOD, DAIRY-PRODUCT }
```

We create corpora drawn from child-directed speech, in which lemmatized, transcribed utterances are paired with artificially generated semantics, based on WordNet or other semantic featural representations of the entities and actions corresponding to the words. In the experiments here on novel word learning, nonce words are paired with these naturalistic semantic representations, in which features corresponding to meaning properties are probabilistically associated with a word.

When processing an input pair, the model bootstraps its current knowledge of word meanings to hypothesize the
strength of association between the words in the current input and the meaning features in the current scene. These probabilistic alignments between the words and features of the current input are then used to update the model's knowledge of word meanings.

More formally, for each word, the model learns a meaning probability, which is a probability distribution over all possible semantic features. The model starts with uniform meaning probabilities for all words; i.e., before processing any input, all features are equally likely for every word. At each time step $t$, the model processes an input pair and calculates an alignment score, $a_{t}(w, f)$, between each word $w$ and semantic feature $f$ in the input pair. This alignment score reflects how strongly the $w-f$ pair are associated at time $t$, by considering two sources of information: (1) the meaning probabilities of all the words in the utterance $\mathrm{U}_{t}$ (representing the knowledge of the model of word meanings up to that point), and (2) the novelty of words, capturing the attention a learner might pay to the novel words compared to the familiar words (explained below). The alignment score is formulated as:

$$
\begin{equation*}
a_{t}(w, f)=\frac{p_{t}(f \mid w)}{\sum_{w^{\prime} \in \mathrm{U}_{t}} p_{t}\left(f \mid w^{\prime}\right)} * \operatorname{novelty}_{t}(w) \tag{1}
\end{equation*}
$$

where $p_{t}(f \mid w)$ is the probability of $f$ being part of the meaning of word $w$ at time $t$, right before processing the input pair, and novelty $_{t}(w)$ is a multiplicative attentional factor.

This factor, novelty ${ }_{t}(w)$, taps into empirical studies on attention showing that people attend to novel items in a learning scenario more than other items, leading to improved learning of those items (e.g., Snyder, Blank, \& Marsolek, 2008; MacPherson \& Moore, 2010; Horst, Samuelson, Kucker, \& McMurray, 2011). In the word learning scenario, this corresponds to a person focusing on determining the meaning of novel words. We model this observation by incorporating the multiplicative novelty $y_{t}(w)$ in the above formula, providing an increase in word-feature association for a more novel word. The novelty ${ }_{t}(w)$ measures the degree of novelty of a word as a simple inverse function of recency: The more recently a word $w$ has been observed by the model $\left(t_{\text {last }_{w}}\right)$, the less novel it appears to the model at the current time $t$ :

$$
\begin{equation*}
\operatorname{novelty}_{t}(w)=1-\operatorname{recency}\left(t, t_{\text {last }_{w}}\right) \tag{2}
\end{equation*}
$$

where recency $\left(t, t_{\text {last }_{w}}\right)$ is inversely proportional to the difference between $t$ and $t_{\text {last }_{w}}$. We set novelty $(w)$ to be 1 for the first exposure of the word.

## Accumulating Evidence over Time

The model keeps track of the accumulation of all the alignment scores of all word-feature pairs, and uses these scores to update the meaning probabilities of the words. These alignment scores reflect the model's knowledge of the associations between words and various potential meanings. To simulate the effect of forgetting in memory, these alignments undergo
a decay over time. At each time $t$, the strength of association of a word and a feature is formulated as:

$$
\begin{equation*}
\operatorname{assoc}_{t}(f, w)=\sum_{t^{\prime}} \frac{a_{t^{\prime}(w, f)}}{\left(t-t^{\prime}\right)^{d_{a_{t^{\prime}}}}} \tag{3}
\end{equation*}
$$

where $t^{\prime}$ is the time at which the alignment $a_{t^{\prime}}$ is calculated, and $d_{a_{t^{\prime}}}$ is the decay rate associated with this alignment. We note that our formulation of assoc is inspired by the ACT-R model of memory (Anderson \& Lebiere, 1998), in which the sum of individual memory strengthenings for an item determines the item's activation. We assume that stronger alignments should be more entrenched in memory and thus decay more slowly than weaker alignments. Thus, each alignment undergoes a decay which is dependent on the strength of the alignment:

$$
\begin{equation*}
d_{a_{t^{\prime}}}=\frac{d}{a_{t^{\prime}}(w, f)} \tag{4}
\end{equation*}
$$

where $d$ is a constant parameter. Note that the alignments between a word and different features may be forgotten at different rates.

This association score is then normalized using a smoothed version of the following to yield the meaning probability of that feature $f$ for that word $w$ at time $t$ :

$$
\begin{equation*}
p_{t}(f \mid w)=\frac{\operatorname{assoc}_{t}(f, w)}{\sum_{f^{\prime} \in \mathcal{M}} \operatorname{assoc}_{t}\left(f^{\prime}, w\right)} \tag{5}
\end{equation*}
$$

where $\mathcal{M}$ is the set of all observed meaning features.

## Desirable Difficulties in Word Learning

Vlach and Sandhofer (2010) - henceforth V\&S - explore the factors involved in "desirable difficulty" through a set of (now standard) cross-situational word learning experiments on adults, varying the presentation and testing conditions. In each $N \times N$ trial, subjects see some number $N$ of novel objects on a computer screen, while hearing $N$ novel words (in arbitrary order) that label the displayed objects; see Figure 1. In testing, subjects hear a single word, and are asked to select the corresponding object from a display of 4 objects. Across three presentation conditions, the total number of word-object pairs, and the number of times each is seen, are held constant, while there is increasing within-trial ambiguity - i.e., the number of possible pairings between the words and the objects within a single presentation: $2 \times 2$, $3 \times 3$, and $4 \times 4$. Furthermore, participants were tested at each of three times: immediately after training, 30 minutes after, and one week after.

V\&S find that in the immediate testing condition, as expected, the number of correctly learned pairs decreases as the within-trial ambiguity increases. That is, the participants performed the best in the $2 \times 2$ condition and the worst in $4 \times 4$ (Figure 2). However, when tested after 30 minutes of delay, there was no significant difference between the performance


Figure 1: Example stimuli from $2 \times 2$ condition taken from V\&S.
of the participants in the $2 \times 2$ and the $3 \times 3$ conditions, while $4 \times 4$ still had the worst performance. Interestingly, in testing after one week, the participants performed better in the $3 \times 3$ than the $2 \times 2$ condition. (Again, $4 \times 4$ still had the worst performance.) In summary, what should be the "easiest" condition $(2 \times 2)$ has the best performance in immediate testing, but a more difficult condition $(3 \times 3)$ has better performance one week later.


Figure 2: The results of V\&S's experiment.
V\&S relate their findings to "desirable difficulties" in learning: they argue that the difficulty of a learning situation might hinder immediate performance, but promote longer term performance. However, they do not discuss why the performance of the $4 \times 4$ condition is the worst compared to the other conditions for all testing intervals. That is, why is the level of difficulty in $3 \times 3$ desired, but is not so for $4 \times 4$. Moreover, they do not explain why and how difficulty can boost learning in the long term in this learning scenario.

We observe that, in the V\&S experiments, the $2 \times 2$ condition has more learning trials, each of which is seen for less time, than in the $3 \times 3$ condition (and similarly for $3 \times 3$ compared to $4 \times 4$ ). This occurs because the total number of word-object pairs, the number of times each is seen, and the total presentation time of the full set of items, are all held constant across the three presentation conditions. We can thus identify three factors that differ across the V\&S conditions, each of which may contribute to the observed pattern: (1) the within-trial ambiguity, (2) the presentation duration of each trial, and (3) the average spacing interval (where spacing is
the number of trials between the two presentations of a wordobject pair).

Computational modelling can be used as a tool to study the necessity and the interaction of these three factors (the within-trial ambiguity, the presentation time of each trial, and the average spacing interval) in a cross-situational learning scenario. In our model, the increase in within-trial ambiguity results in more competition among the possible alignments since there are more words and meanings to potentially align; this results in lower association scores and therefore decreased performance in word learning. We argue that the second factor, the presentation duration, is related to forgetting. In the following section (Methodology), we will explain how we incorporate differences in the presentation duration into our model. The third factor (the spacing interval) relates to the interaction of forgetting and attention to novelty in the model: As the spacing interval becomes larger, the amount of forgetting increases, resulting in lower association scores between words and features; however, the novelty of words and consequently their association scores increases as the spacing interval gets larger. Thus, varying the spacing interval affects the performance of the model (see Nematzadeh et al., 2012a for more details). We use our model to study the interaction of these three factors, with the goal of providing a more precise explanation for the desirable difficulty observed in the experiments of V\&S. Next, we explain our methodology, including our input generation, and the simulation of the V\&S experiments.

## Methodology

## Input Generation

To generate the input stimuli for our model, we need to pair words with a meaning representation that corresponds to the depiction of the corresponding object in the experimental situation of Figure 1. To do so, we draw on the input-generation lexicon of Nematzadeh, Fazly, and Stevenson (2012b), which was previously used to automatically annotate corpora of child-directed utterances with meaning features corresponding to the words in those utterances. Here, we use the lexicon to provide a source of naturalistic meaning representations ("novel object descriptions") for a set of "novel" words (i.e., the words in the input stimuli are unknown to the model, as in the experiments we are modeling).

The true meaning of each word in the lexicon, $\operatorname{tm}(w)$, is a vector of semantic features and their assigned scores or weights (Figure 3). ${ }^{1}$ When a word is used in an input trial, its meaning features are probabilistically sampled from $\mathrm{tm}(w)$ according to the weight of each feature in the lexical entry of the word. This probabilistic sampling captures our intuition that a participant, when faced with a trial in the crosssituational experiment of Figure 1, will grasp some features of the novel objects but not necessarily all. Each trial of the input is then composed of a set of $N$ words ( 2,3 , or 4 words,

[^174]depending on the condition), paired with a set of features which is the union of the $N$ sets of meaning features sampled for each of the words in that trial.

```
apple: { FOOD:1, SOLID:.72, PRODUCE:.63,
    EDIBLE-FRUIT:.32, PLANT-PART:.22,
    PHYSICAL-ENTITY:.17, WHOLE:.06, .. }
```

Figure 3: True meaning features \& probabilities for apple.

To produce a full set of experimental trials, we first convert the exact stimuli of V\&S to the format of our input. That is, in their stimuli, we replace each word with a specific word from our lexicon, and each object with the probabilisticallygenerated meaning representation for its corresponding word (as explained above). The precise combination of corresponding word/object pairs in each trial, and the order of the trials, are exactly the same as in the V\&S stimuli. We refer to this data as the input of V\&S.

The V\&S input includes 18 novel word-object pairs, each of which occurs 6 times, resulting in 54, 36, and 27 trials in the $2 \times 2,3 \times 3$, and $4 \times 4$ conditions, respectively. We note that the V\&S input, as a specific set of stimuli, might have particular spacing properties that contribute to their results. Thus we also randomly generate input stimuli in order to evaluate the effect of arbitrary variation in the precise presentation order of the word/object pairs. We randomly generate 20 sets of input stimuli for each condition, keeping the number of pairs, their frequency, and the number of trials the same as in the V\&S input. We use the same novel words that we used in generating V\&S data, and randomly generate their meaning representations as explained. The result is that we can experiment both with the precise data of V\&S, as well as 20 randomly generated sets of input stimuli with the same basic properties.

## Modeling of the Presentation Duration

One aspect of the V\&S experimental conditions that we cannot directly replicate in our model is the presentation duration of each trial in a stimulus set. Recall that because of the various properties of the stimuli, the individual trials in each of the three conditions ( $2 \times 2,3 \times 3$, and $4 \times 4$ ) have different presentation durations. Our model does not have a notion of "presentation duration" - it simply processes each input as it receives it. Thus to simulate these differences, different degrees of forgetting decays are used in the model (see Eqn. (4)). The intuition is that subjects forget faster in a condition with a shorter presentation duration, since they have less time to absorb the stimuli in each trial. The forgetting decay is thus set to a larger value in the $2 \times 2$ condition (where the presentation time is the smallest), and successively smaller in each of the $3 \times 3$ and $4 \times 4$ conditions.

## Simulation of the V\&S Experiments

We train our model by presenting the set of inputs for a given condition, where it learns incrementally in response to each trial. Similarly to V\&S, we evaluate our model at three points
of time after training: immediately after processing the last input (time $=t$ ), at $t+30$, and at $t+350$. These times were chosen to loosely reflect the three time intervals in V\&S's experiments. We will use the labels "no delay", "brief delay", and "lengthy delay", to refer to these timings in describing our results.

To evaluate the performance of the model at each testing point, we measure how well each word is acquired by comparing its learned meaning $\operatorname{lm}(w)$ - a vector holding the values of the meaning probability (Eqn. (5)) - to its true meaning $\operatorname{tm}(w)$ from the input-generation lexicon (see Figure 3):

$$
\begin{equation*}
\operatorname{acq}(w)=\operatorname{sim}(\operatorname{lm}(w), \operatorname{tm}(w)) \tag{6}
\end{equation*}
$$

where sim is the cosine similarity between the two meaning vectors, $\operatorname{tm}(w)$ and $\operatorname{lm}(w)$. The higher acq $(w)$ is, the more similar $\operatorname{lm}(w)$ and $\operatorname{tm}(w)$ are. We use the average acq score at time $t$ of all the words in the input to reflect the overall learning of the model at that time.

## Results

We first examine the behavior of our model when trained on the V\&S input, and then compare these with results on our randomly generated stimuli.

## The Input of V\&S

The results of training and evaluating our model on the V\&S input are presented in Figure 4. We see the same interesting pattern as found in V\&S (shown in Figure 2) for the $2 \times 2$ and the $3 \times 3$ conditions. That is, $2 \times 2$ is better with no delay, but similar with brief delay and worse with lengthy delay, even though $3 \times 3$ is "harder" due to its higher degree of withintrial ambiguity. Unlike the V\&S results, $3 \times 3$ and $4 \times 4$ are similar for all delays.


Figure 4: Average acq score of words (from the model) given the three conditions and three time intervals similar to the V\&S experiments.

We consider these findings in the context of the discussed factors of presentation duration, within-trial ambiguity, and average spacing of items, which we proposed might explain
the desirable difficulty in learning. The differences in presentation duration (shortest for $2 \times 2$ and longest for $4 \times 4$ ) entails that, generally, the learning in the $2 \times 2$ condition should decline most steeply over time, and learning in the $4 \times 4$ should decline least steeply: i.e., for each set of same-coloured bars in Figure 4, we expect learning to decrease over time, and more rapidly for lower values of $N$ in the $N \times N$ conditions. We see this predicted behaviour with our model, which results from our modeling of presentation duration with an inversely proportional decay rate (i.e., the shorter the presentation duration, the greater the degree of forgetting).

It is expected that in the absence of other factors, increasing within-trial ambiguity from the $2 \times 2$ to the $4 \times 4$ conditions results in a decline in average acq score, since greater ambiguity should lead to decreased learning. However, in our model, the presentation duration also plays a role. Similar to results of V\&S, we see the decline pattern in the "no delay" condition, and in the "brief delay" condition (albeit with less difference), due to the increased competition for word-meaning alignments that occurs with a higher number of items in a trial (see Figure 4). However, we do not see this pattern in the lengthy delay condition.

To summarize, our results are similar to those of V\&S, who found that while the $2 \times 2$ condition led to best learning when tested immediately, it led to poorer performance than the $3 \times 3$ condition given a lengthy delay before testing - a pattern V\&S attribute to the "desirable difficulty". It seems that these factors of presentation duration and withintrial ambiguity may interact, such that the steep decline in performance in subsequent testing in the $2 \times 2$ condition more than offsets the advantage it has from the lesser within-trial ambiguity.

In the experiments of V\&S, the performance in the $4 \times 4$ condition is always worse than the two other conditions. However, our model produces very similar results for the $3 \times 3$ and the $4 \times 4$ conditions. Also, the role of the spacing interval is not clear in these results. The problem is that by just considering one set of stimuli within each $N \times N$ condition (each of which has a set spacing of items), we do not have a variation of the average spacing interval that is independent of the presentation duration and the within-trial ambiguity. We turn to these issues in the next subsection.

## Randomly Generated Input

We observed that the performance of the model in the $3 \times 3$ and $4 \times 4$ conditions on the V\&S input is very similar. We also investigate a condition here with higher within-trial ambiguity to see if such a condition might be "hard" enough for the model (because of the higher within-trial ambiguity) so that it results in a similar patten to the $4 \times 4$ condition in V\&S. As with the others, we generate 20 sets of input stimuli for this $6 \times 6$ condition, using 18 word-object pairs, each of which occurs 6 times, producing 18 trials. Thus the generated input stimuli for the four conditions allows us to examine both the role of average spacing interval, and the impact of a more difficult condition with higher within-trial ambiguity.

We train our model on the randomly-generated inputs (with different average spacing intervals) for all four $N \times N$ conditions. To evaluate the performance of the model, the average acq score of words for all 20 sets of inputs within a single $N \times N$ condition are averaged (see Figure 5). We can see that when tested with "no delay", the $2 \times 2,3 \times 3$, and $4 \times 4$ conditions have similar scores. Moreover, we can see a pattern similar to V\&S's experiments: the $3 \times 3$ and $4 \times 4$ conditions have the best results after the "lengthy delay". We also observe that by increasing difficulty in the $6 \times 6$ condition (due to the high within-trial ambiguity), the model produces a pattern similar to the pattern observed in the $4 \times 4$ condition in V\&S's experiments. This confirms our hypothesis that for our model, the $4 \times 4$ condition is not "hard" enough to result in a steep decline over time intervals as in the V\&S's results.


Figure 5: Average acq score of words (from the model) given the four conditions and the three time intervals, averaged over 20 sets of stimuli.

However, we also see that, in contrast to V\&S's results (and our model's performance on the V\&S data), the $2 \times 2$ condition with no delay fails to show better learning than the other conditions.

To better understand this difference between the two sets of results, we look more closely at the scores of the individual randomly-generated stimuli sets. We find that there is a notable difference in the average $a c q$ score across the 20 input files for the $2 \times 2$ condition, which shows its maximum value of 0.76 for the V\&S's data, while the minimum is 0.50 . This suggests that the characteristics of the particular input (as a result of varying the average spacing interval) may be responsible for some of the observed patterns in the V\&S's results.

We were interested to understand why the V\&S data has the maximum score, especially since there was a sizable gap between the score of this input and the next best score among the randomly-generated inputs (of 0.64). In an attempt to identify the factor behind this variation, we measured various statistics for each input set, such as the following: (1) the average spacing interval of words, which has been shown to affect learning both in people (Vlach et al., 2008) and in our model (Nematzadeh et al., 2012a); (2) the average time since
the last occurrence of words, that impacts the amount of forgetting that occurs; and (3) the average context familiarity of words (that is, the familiarity of the words that occur with a word in an utterance), a factor that has been noted to affect word learning (see, e.g., Fazly, Ahmadi-Fakhr, Alishahi, \& Stevenson, 2010). However, we found that none of these measures explain the variation of the scores in all the inputs. Future research is needed to fully understand the impact of the properties these measures tap into, and whether they may (individually or in combination) contribute to explaining the pattern of the results.

## Summary

The "desirable difficulty" of a learning condition can promote the long term retention of the learned items. We have used a computational model to investigate the possible factors behind one such case of a "desirable difficulty" in a crosssituational word learning experiment (Vlach \& Sandhofer, 2010). Notably, the experimental results were not clearly pointing to the factors causing the patterns observed in the performance of the human participants. Using a computational model, we have suggested that an interaction between two factors (the within-trial ambiguity of the learning trials, and the presentation duration of each trial) might explain the observed patterns. In addition, our results point to other distributional characteristics of the input (experimental stimuli) that might have an impact on the performance of the learner. These findings illustrate the role of computational modelling, not only in explaining observed human behaviour, but also in fully understanding the factors involved in a phenomenon. There are several factors involved in a cross-situational word learning experiment, such as the contextual familiarity of words, and the average spacing interval of words. Our findings signify the importance of controlling for these factors in order to understand the reasons behind the observed patterns. But it is difficult do so in human experiments because the factors can interact in complex ways.

Our work is an initial attempt at shedding light on the interaction of memory, attention and word learning, and understanding "desirable difficulty" in learning. Other factors (e.g., working memory) might play a role in the performance of people as well. For example, because the number of items that people can store in their working memory is limited (Miller, 1956), the participants might store more trials in their working memory in the $2 \times 2$ condition, compared with the other conditions. The participants might use this information of the multiple trials (in their working memory) to make inferences about word-object mappings that repeat in successive trials. One future direction would be to incorporate a working memory module into our word learning model, and examine the impact of such inferences in a cross-situational learning scenario.

## Acknowledgements

We would like to thank Haley Vlach for valuable discussion, and for providing us with the stimuli of her experiments. We
are also grateful for the financial support from NSERC of Canada, and University of Toronto.

## References

Anderson, J. R., \& Lebiere, C. (1998). The atomic components of thought. Lawrence Erlbaum Associates.
Fazly, A., Ahmadi-Fakhr, F., Alishahi, A., \& Stevenson, S. (2010). Cross-situational learning of low frequency words: The role of context familiarity and age of exposure. Proc. of CogSci, 10 .
Fazly, A., Alishahi, A., \& Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. Cognitive Science, 34(6), 1017-1063.
Horst, J. S., Samuelson, L. K., Kucker, S. C., \& McMurray, B. (2011). Whats new? children prefer novelty in referent selection. Cognition, 118(2), 234-244.
MacPherson, A. C., \& Moore, C. (2010). Understanding interest in the second year of life. Infancy, 15(3), 324-335.
Miller, G. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological review, 63(2), 81.
Nematzadeh, A., Fazly, A., \& Stevenson, S. (2012a). A computational model of memory, attention, and word learning. In Proceedings of the 3rd workshop on cognitive modeling and computational linguistics. Association for Computational Linguistics.
Nematzadeh, A., Fazly, A., \& Stevenson, S. (2012b). Interaction of word learning and semantic category formation in late talking. In Proceedings of the 34th Annual Conference of the Cognitive Science Society.
Pinker, S. (1989). Learnability and cognition: The acquisition of argument structure. Cambridge, Mass.: MIT Press.
Quine, W. V. O. (1960). Word and object. MIT Press.
Smith, L., \& Yu, C. (2008). Infants rapidly learn wordreferent mappings via cross-situational statistics. Cognition, 106(3), 1558-1568.
Snyder, K. A., Blank, M. P., \& Marsolek, C. J. (2008). What form of memory underlies novelty preferences? Psychological Bulletin and Review, 15(2), 315-321.
Tomasello, M. (1992). The social bases of language acquisition. Social development, 1(1), 67-87.
Vlach, H. A., \& Sandhofer, C. M. (2010). Desirable difficulties in cross-situational word learning. In Proceedings of the 32nd annual conference of the cognitive science society.
Vlach, H. A., Sandhofer, C. M., \& Kornell, N. (2008, October 0). The Spacing Effect in Children's Memory and Category Induction. Cognition, 109(1), 163-167.
Yu, C., \& Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. Psychological Science, 18(5), 414-420.
Yurovsky, D., \& Yu, C. (2008). Mutual exclusivity in crosssituational statistical learning. In Proceedings of the 30th annual conference of the cognitive science society (pp. 715720).

# The Individuation and Recognition of Emotion 

Albert Newen (albert.newen@rub.de)<br>Institute for Philosophy II, Universitätsstraße 150<br>44780 Bochum, Germany

Anna Welpinghus (anna.welpinghus@rub.de)<br>Institute for Philosophy II, Universitätsstraße 150<br>44780 Bochum, Germany


#### Abstract

First, we argue for the metaphysical claim that emotions are individuated as patterns of characteristic features. Our second claim concerns the epistemology of emotion recognition: We demonstrate that emotion recognition is a process pattern recognition that relies on the same type of pattern recognition typical for object recognition. The analogy allows us to defend a variant of a direct perception account of emotion recognition. We distinguish two forms of directly perceiving emotions: 1. perceiving an emotion (almost) without any top-down-processes, 2. perceiving an emotion involving some significant top-down-processes (including expectations and background knowledge), and in addition 3. an inferencebased evaluation of an emotion.


Keywords: Emotion, Philosophical Theory, Nature of Emotion, Recognition of Emotion

## Introduction

What is the nature of emotions and how can we recognize them in other human beings? In this short paper, we cannot do justice to the complex discussions concerning both questions. Nevertheless, we develop our own view with the core claim that emotions are individuated as pattern and that emotion recognition is a process of pattern recognition that has much in common with the recognition of objects by perception. This paper's focus lies on an outline of our account and its advantages, while we only briefly mention the criticism of alternative views.

## The metaphysical debate: An overview

In the metaphysical debate we have two extreme positions: emotions are individuated as social constructs (Lutz, 1986; Harré, 1986), on the one hand, or they are individuated as evolutionary anchored affect programs (Ekman, 1972; Griffith, 1997), on the other. Both accounts have severe deficits (Welpinghus \& Newen, 2012). Let us mention only the two main deficits: psychoevolutionary accounts state that shared evolutionary history is the only criteria to identify types of emotions. They do not provide any classificatory schemes which do not refer to each category's evolutionary history but for many emotion categories referred to not only in everyday speech but also in psychological theories, it is far from clear whether their members share the same evolutionary history. Thus, the psychoevolutionary account has difficulties providing adequate classificatory schemes, for example for studying
emotions in a social context. In principle, psychoevolutionary accounts of emotions can easily account for basic emotions but have problems to account for the role of cognitive contents in so-called cognitive emotions (Zinck \& Newen, 2008). On the other hand, the social constructionist can easily account for the latter including the cultural variety of emotion phenomena but they underestimate the strong overlap of the emotion repertoire despite the cultural variation. Here Ekman (1972) has shown that basic emotions like joy, fear, anger, sadness etc. are accompanied with the same facial expression. There is an open debate which phenomena are basic emotions but a strong part of the community presupposes that there are basic emotions which are evolutionary old, shared with animals and develop early in ontogeny (Ekman, 1999; Griffiths, 1997; Zinck \& Newen, 2008). The evolutionary anchor of basic emotions constraints our emotion repertoire and undermines the social constructivist view that emotions are entirely created by cultural factors. What could be an alternative? We need to do justice to both features, the evolutionary anchor of basic emotions and the cultural dependence of some emotions. We suggest that the claim that emotions are individuated as pattern is the best alternative: 1. pattern can easily involve both, evolutionary anchored as well as culturally shaped features and thus account for both observations; 2. this view especially helps to distinguish emotion concepts in a society and their natural basis, i.e. some emotions concepts are categorizing only conventional constructs while others are actually anchored in natural kinds (which empirical science has to discover). 2. The account of emotion as pattern is nicely connecting with our folk psychological way of thinking about emotions (noticing the many faces of emotions), 3 . the best reductive scientific accounts of emotions have (at least so far) not succeeded in reducing emotions to a very few necessary features which are constituting a type of emotion, 4. furthermore, we observe a great variety of realizations of one type of emotion (e.g. types of fear; types of phobia which are classified according to the intentional object but also according to other features) which indicates that the analysis as a pattern consisting of characteristic features is the best we can gain because it makes it easy to account for the variety of emotion types without loosing the evolutionary anchor. Let us illustrate some of these points while enfolding the view that emotions are individuated as pattern. One important aspect of pattern is that in principle all
features are dispensable: there only has to remain a minimal package of features constituting the phenomenon but none of them is necessarily instantiated. Even this radical feature of pattern is true for several emotions but not for all. ${ }^{1}$

## Emotions as patterns

The idea that emotions are organized in patterns and that thus the recognition of emotion is basically a process of pattern recognition is not new (e.g. it is used in Izard, 1972 and Izard et al., 2000). But we aim to develop a detailed own account of it that enables us to analyze emotion recognition in more detail. We will now list and shortly characterize all types of features relevant for the individuation of emotions before focusing on the recognition part:
(1) Autonomic physiological responses: William James (1884) famously claimed that an emotion is the perception of bodily changes, especially changes in autonomic nervous system (ANS) activity. Physiological parameters controlled by ANS activity are crucial for Jamesian emotion theories. However, since not every emotion has a clearly distinct ANS pattern, especially when it comes to more fine-grained distinctions, even clearly Jamesian accounts of emotion such as the account of Prinz (2004) have to provide an account of emotion which grants that two different emotions need not have different patterns of ANS activity. They can be distinguished by other features. ${ }^{2}$
(2) Expressive actions and action tendencies: The psychologist Nico Frijda has coined the term 'action tendency' (which stems from Arnold, 1960) for a felt urge to perform a certain kind of action. This urge is also manifest in bodily changes which are suitable preparations for these actions. He largely equates emotions and felt action tendencies (Frijda, 1986, 71), a move we do not share. Specific actions or action tendencies as the only features are neither constitutive for every type of emotion nor can they alone constitute any emotion: Emotions like happiness are not accompanied by a specific action tendency. An action tendency may become manifest in actual expressive actions or not. Although actions out of emotions are rather flexible, there are typical behaviors for some emotions, such as freezing or fleeing in fearful situations. We routinely rely on expressive actions as well as signs for mere action tendencies for emotion recognition.
(3) Bodily expression: Bodily expression can be divided into (a) facial expression, (b) posture (c) gesture, (d) vocal

[^175]expression such as screams, roars or laughter and (e) tone of voice. We subsume these under the heading 'bodily expression' because all of these depend on muscle contractions and because these components are usually perceived together. We see a proud person talking: We see her smile, her erect posture; we hear her self-satisfied tone of voice. This leads to an overall impression of her as proud; under normal conditions, we do not pay attention to any of these components separately. Laboratory studies point towards an early integration of these visual and auditory cues (Campanella \& Belin, 2007). On the other hand, experiments show that face (Ekman, Friesen \& Ellsworth, 1972), but also posture and gesture alone is sufficient to recognize some coarse-grained distinctions (Atkinson et al., 2004).
(4) Phenomenal experience: Emotions are normally accompanied by a particular phenomenal quality or feeling. We do not consider phenomenal experience as necessary because there may be rare cases of emotions in the absence of such feelings, when typical physiological, expressive and cognitive aspects are present (the paradigmatic case is a person with a strong disposition to repress her fear [Sparks et al., 1999; Weinberger, 1990; Weinberger \& Davidson, 1994]).
(5) Cognitive features comprise (a) attitudes and (b) shifts of attention and perception. Emotions are often accompanied by cognitive attitudes. Belittling thoughts about the rival are characteristic for jealousy and a judgment that one has been treated unfairly is characteristic (though neither necessary nor sufficient) for anger, to give two examples. Furthermore, sometimes such attitudes are manifest in behavior or are verbally reported. Cognitive components that are less demanding than attitudes are e.g, a shift of attention, for example being alert to dangers in a state of fear. Through such shifts, emotions can make us perceive things we otherwise would not have perceived. They also make us perceive things in a certain light. Some theorists put characteristic ways of perceiving the world at the center of their theory of emotion (Döring, 2003). We will not equate emotion with ways of perceiving a world, but we do acknowledge the role of changes in perceiving the world due to having an emotion connected with a special phenomenology and its role in guiding behavior.
(6) The intentional object: 'Intentional object' is a technical term for the object the emotion is about. 'This can be a particular thing or person (that pudding, this man), an event or an action (the earthquake, your hitting me), or a state of affairs (my being in an aeroplane)' (Goldie, 2000, 16-17). It is not always real. The intentional object (or the type of object) is crucial for the more fine-grained classification of emotions, for example for distinguishing envy from jealousy. ${ }^{3}$

[^176]The features can be thought of as variables, which can take different values (the variable facial expression can take the values 'fearful expression', 'sad expression', etc.) Some values are likely to occur together, others are very unlikely to occur together. This is due to a range of causal mechanisms and constraints, but probably also constitutive and conceptual links. Illuminating these links further is a task for the empirical sciences as well as philosophy. Because some values are likely to occur together, we can learn to distinguish typical patterns of values and form a concept of the overall phenomenon. To illustrate, figure 1 shows a fear pattern.


Figure 1: Fear pattern.

## Our understanding of 'being constitutive of an emotion'

Are these features constitutive of an emotion? This depends on the underlying conception of 'constitution'. According to an essentialist conception of constitution, a constitutive part of something is also an essential part, a necessary part in all possible worlds. If the part was taken away, the leftovers could not form the same type of phenomenon as before anymore. According to our view of emotions as patterns, emotional features are not constitutive in the essentialist sense, because we do not rule out the possibility to have a token of the same type of an emotion lacking some characteristic features. We think that if there are only a few characteristic features they can be sufficient for an emotional episode to constitute it to be a token of the type anger. Those features of an emotional episode which contribute to it being a token of a specific emotion type are the constitutive parts of an emotion pattern ${ }^{4}$. The fact that a
entail the claim that the object is partly the bearer of the mental state in this case.
${ }^{4}$ We are presupposing a realism about patterns such that in principle for each combination of features it is possible to clarify whether this is an instantiation of the relevant pattern or not. We can of course grant that there are borderline cases and there exists a bundle of transfer cases.
person is sitting in a car during her angry episode may be part of a token of anger but it is not constitutive of her anger. We suggest a notion of being constitutive for X that accounts for X being a pattern. A feature f is constitutive for a pattern X if it is part of at least one set of features which is minimally sufficient for a token to belong to a type X. 'Minimally sufficient' means that these features are jointly sufficient for the episode to be of type X , but if one of them would be taken away the episode would not count as a instance of X anymore. According to our view, emotions usually include overt bodily features (expression or characteristic actions) - if characteristic expressive aspects occur during an emotion episode, they are constitutive and not just an effect of the proper emotion, or a cause for it but emotions do not always involve expressive components. ${ }^{5}$

One might wonder whether our list of features is exhaustive. ${ }^{6}$ Should we include neural correlates, for example? Features like ANS response, behavior, cognitive attitudes, perceptual shifts, and expression all have neural correlates. Neural correlates are insofar not an extra component in addition to the others - we could mention them as part of the individuation features of emotion but we do not have to. ${ }^{7}$ They may be used in a clinical or scientific context to infer whether a person is having emotions but are not used in ordinary contexts, since we cannot access them in an everyday context. Also, it is crucial to distinguish the factors relevant for individuating emotions from those which facilitate (or impair) the recognition of emotions. Contextual factors for the recognition of emotions include most prominently (a) the pragmatic context in which it occurs and (b) the personality, goals, beliefs, etc. of the person having the emotion. These features do not belong to the emotion pattern but are only relevant for the recognition

[^177]of other people's emotions. ${ }^{8}$ One may also wonder whether the list of features includes too many features and whether some can be reduced to others: This may be possible - and desirable - in principle but given the state of art in cognitive sciences we have no convincing candidate for a theory which reduces emotions to fewer features and would nonetheless allow us to describe the great variety of emotion phenomena.

## Emotion recognition as pattern recognition

Pattern recognition is a general method of classifying world phenomena. We aim to analyze the recognition of emotion as a process of pattern recognition on the basis of perceiving characteristic features and integrating them into a unity by recognizing an emotion and ascribing it to the other. ${ }^{9}$ The latter presupposes observers possessing normal conceptual ${ }^{10}$ competences. In providing a new version of explicating emotion recognition as pattern recognition we propose a new variant of a perceptual theory of recognizing emotions. We will characterize the process of pattern recognition as one of cue combination and cue integration that culminates in an activation of the most plausible representation of an emotion pattern. This pattern recognition involves a multifactorial weighting process organizing sensory cues from different sense modalities, on the one hand, and accounting for other cues, like social and personal background information, on the other. It is fruitful to distinguish between two types of emotion recognition: The process of pattern recognition by activating the most likely representation of an emotion pattern can take place either without almost any top-down processes or strongly constrained by those. We understand top-down processes as specific processes involving prefrontal activation of the brain; this presupposes as a minimal consensus that prefrontal activation of the brain is necessarily involved in the activation of complex cognitive processes which can be loosely understood as conceptual. We can rely on studies which show that such prefrontal activation is at least sometimes involved in standard perceptual processes. ${ }^{11}$ Thus, we distinguish 1. (a basic form of) perceiving an emotion (almost) without any top-down-processes, 2. perceiving an emotion by strongly involving top-downprocesses (a strongly concept-modified form of perception).

[^178]Both types of perceiving emotions can be distinguished from 3. an inference-based evaluation of an emotion pattern. The latter presupposes a stable evaluation of an emotion as being F , which then may be modified or reevaluated by reflecting on the information (Newen \& Welpinghus, under review).

## Cue combination and cue integration in recognition of emotions

We have argued that emotions are individuated on the basis of a number of features that constitute an emotion pattern. Our aim now is to characterize the relevant process of pattern recognition in the case of emotions in more detail. To reach that aim, we analyze the process of recognizing emotions in parallel to the process of perceiving objects. We take the relevant model of object perception from Ernst and Bülthoff (2004): It describes how a stable percept is developed by merging the senses in a process of cue combination and cue integration. This model of object recognition can be directly transferred into a model of emotion recognition including four aspects: (i) The bottomup processes that start with sensory cues of an emotion, which lead to a first estimate of an emotion. A detailed analysis of emotion recognition would have to distinguish at least the following cues: facial expression (F), the gestures (G), the whole body posture (B), the tone of voice (T), autonomic features like sweating (A), the movement/action of the person (M) and the event whose registration triggers the emotion (C). Registering these features also leads to an internal activation of my own, the perceiver's bodily state, e.g. mirror neuron activation, which has been shown at least in the case of seeing disgust in the face of the other person (Calder et al., 2000), and (at least) sometimes also of a visceral activation. (ii) The relevant sensory cues are used for cue combination (combining nonredundant features) and cue integration (selecting one dominant out of a group of redundant features); it is plausible that the process of registering an emotion starts with cue combination of complementary features like F and M producing an intermediate (normally unconscious) estimate E1 of an emotion state, as well as a combination of G, B, A and M, which may constitute a different (normally unconscious) estimate E2 of an emotion state. In a second step, we are in need of integration of redundant cues like E1, E2 and E3 while the latter may be triggered by the auditory cue (T) together with registering the cause (C). (iii) Cue combination and integration leads to an activation of the most likely emotion pattern (on the basis of Bayesian principles) which is connected with a (normally conscious) stable percept. Thus, there is a plausible sense in which we can say that we see an emotion while seeing a person having the features mentioned above. The percept we have is directly associated with the activation of an emotion pattern which we are able to distinguish from others. (iv) Furthermore, the development of an emotion percept may be influenced by top-down processes. Thus, the percept of an emotion is a product of sensory cue combination and
integration influenced by top-down processes: In the case of emotions, situational and person-specific background knowledge can modify the basic bottom-up processes that predicted the emotional state of the other.

## Concluding remarks

Describing emotions as patterns of characteristic features has a number of advantages: First, in the debate between evolutionary theorists and social constructionists, it offers an account that allows to integrate evolutionary anchored as well as culturally shaped features. This especially helps to distinguish emotion concepts in a society and their natural basis, i.e. some emotions concepts are categorizing only conventional constructs while others are actually anchored in natural kinds. The pattern theory accounts for a significant degree of conceptual flexibility in the realm of emotions - based on formal objects and on similarity relations among the various emotion features; given the nature of our affect programs we can actually form a number of coherent classificatory schemata for emotions. Social constructionists often concentrate on emotion categorization in different culture which may be different even if the behavioral dispositions of the people from different cultures would be very similar. We can account for this by characterizing emotions concepts by the significant pattern including behavioral as well as cognitive features. Evolutionary theorists claim that despite cultural variation that any real emotion is constituted by an underlying common core of a few evolutionary anchored features, e.g. fear involves the disposition to freeze or to initiate flight behavior, it has its typical phenomenology. These observations are accounted for by distinguishing rather inflexible and more flexible features of an emotion type. On the basis of the metaphysical view that emotions are individuated as pattern, we argued for a theory of emotion recognition such that it is a process of pattern recognition which is furthermore parallel to object recognition: the same processes of cue combination and cue integration are relevant to develop a percept of an emotion when observing the affective state of another human being. We indicated that we can characterize the perceptual access to emotions as direct, i.e. unmediated by any inferences but we could not develop this view in detail here: we only outlined that the direct perception thesis needs qualification by working out the role of top-down processes in emotion recognition (see Newen/Welpinghus under review). But despite the difference in the relevant features of recognition, both recognition of emotions as well as of objects are analyzed as relying on the same functional processes constituting perception of entities (of any kind). Thus, our model has the advantage of being very parsimonious.

## References

Arnold, M. B. (1960). Emotion and personality. New York: Columbia University Press.
Atkinson, A., Dittrich, W., Gemmell, A., \& Young, A. (2004). Emotion perception from dynamic and static body
expressions in point-light and full-light displays. Perception, 33, 717-46.
Döring, S. (2003). Explaining Action by Emotion. The Philosophical Quarterly, 53, 214-30.
Döring, S. (2009). Philosophie der Gefühle heute. In S. Döring (Ed.), Philosophie der Gefühle (pp. 12-65). Frankfurt am Main: Suhrkamp.
Durand, K., Gallay, M., Seigneuric, A., Robichon, F., \& Baudouin, J. Y. (2007). The development of facial emotion recognition: the role of configural information. $J$ Exp Child Psychol, 97(1), 14-27.
Ekman, P., Friesen, W. V., \& Ellsworth P. (1972). Emotion in the human face. Oxford: Pergamon Press.
Ekman, P. (1999). Basic emotions. In T. Dalgleish \& M. Power (Ed.), Handbook of cognition and emotion (pp. 45-60). Chichester: Wiley.
Ernst, M. O., \& Bülthoff, H. (2004). Merging the senses into a robust percept. Trends in Cognitive Sciences, 8(4), 162-168.
Frijda, N. H. (1986). The Emotions. Studies in Emotions and Social Interactions. New York, NY: Cambridge University Press.
Goldie, P. (2000). The Emotions. A Philosophical Exploration. Oxford: Clarendon Press.
Gopnik, A. and Meltzoff, A. N. (1997): Words, thoughts, and theories. Cambridge, MA.: MIT Press.
Griffiths, P. E. (1997). What emotions really are. The problem of psychological categories. Chicago: University of Chicago Press.
Harré, R. (1986). An outline of the social constructionist viewpoint. In R. Harré (Ed.), The social construction of the emotions (pp. 2-14). Oxford: Basil Blackwell.
Izard, C. E. (1972). Patterns of emotions. A new analysis of anxiety and depression. With chapters coauthored by Edmund S. Bartlett and Alan G. Marshall. New York, NY: Academic Press.
Izard, C. E., Ackerman, B. P., Schoff, K. M., \& Fine, S. E. (2000). Self-organization of discrete emotions, emotion patterns, and emotion-cognition relations. In S. E. Lewis \& I. Granic (Eds.), Emotion, development, and selforganization: Dynamic systems approaches to emotional development (pp. 15-36). New York, NY: Cambridge University Press.
James, W. (1884). What is an emotion? Mind, 9(34), 188205.

LeDoux, J. E. (1996). The emotional brain: The mysterious underpinnings of emotional life. New York, NY: Simon \& Schuster.
Lutz, C. (1986). Emotion words on Ifaluk. In R. Harré (Ed.), The social construction of the emotions (pp.267-288). Oxford: Basil Blackwell.
Lyons, W. (1980). Emotion. Cambridge, MA: Cambridge University Press.
Marks, J. (1982). A Theory of Emotion. Philosophical Studies, 42, 217-242.

Newen, A., \& Bartels, A. (2007). Animal Minds and the Possession of Concepts. Philosophical Psychology, 20, 283-308.
Newen, A., \& Welpinghus, A. (under review): Emotion recognition as pattern recognition: the relevance of perception.
Norton, D., McBain, R., Holt, D. J., Ongur, D., \& Chen, Y. (2009). Association of impaired facial affect recognition with basic facial and visual processing deficits in schizophrenia. Biol Psychiatry, 65, 1094-8.
Parkinson, B., Fischer, A., \& Manstead, A. (2005). Emotions in social relations. New York: Psychology Press.
Prinz, J. (2004). Gut Reactions. A Perceptual Theory of Emotion. New York, NY: Oxford University Press.
Robinson, J. (1983). Emotion, Judgment, and Desire. The Journal of Philosophy, 80, 731-41.
Scherer, K. R. (2009). The dynamic architecture of emotion: Evidence for the component process model. Cognition and Emotion, 23, 1307-51.
Sparks, G. G., Pellechia, M., \& Irvine, C. (1999). The repressive coping style and fright reactions to mass media. Communication Research, 26, 176-192.
Vuilleumier P., \& Pourtois, G. (2007). Distributed and interactive brain mechanisms during emotion face perception: evidence from functional neuroimaging. Neuropsychologia, 45, 174-94.
Weinberger, D. A. (1990). The construct validity of the repressive coping style. In J. L. Singer (Ed.), Repression and dissociation (pp. 337-85). Chicago, Ill: University of Chicago Press.
Weinberger, D. A., \& Davidson, M. N. (1994). Styles of inhibiting emotional expression: Distinguishing repressive coping from impression management. Journal of personality, 62, 587-613.
Welpinghus, A., \& Newen, A. (2012). Emotion und Kultur: Wie individuieren wir Emotionen und welche Rolle spielen kulturelle Faktoren dabei? Zeitschrift für philosophische Forschung, 66, 367-392.
Zinck, A., \& Newen, A. (2008). Classifying emotion: A developmental account. Synthese, 161, 1-25.

# Schema-Driven, Space-Supported Random Accessible Memory Systems for Manipulation of Symbolic Working Memory 

Nader Noori (nnoori@usc.edu), Laurent Itti (itti@usc.edu)<br>Department of Computer Science, University of Southern California Los Angeles, CA 90089 USA


#### Abstract

We present an execution model for manipulation of working memory content during intellectual symbolic working memory tasks, which allows random access of WM content through a schema-operated sensory-motor spatial working memory. The core concept of this framework is binding symbolic items to spatial locations which are accessible via selective mechanisms of attention in space. An operational schema implements basic WM management operations such as insertion, deletion and fetching through sequences of shifts in spatial attention towards registry locations. We apply the model to a serial recall task (both forward and backward orders). We show that the model provides a better fit to human data in backward recall compared to forward recall, which conforms with the evidence for leveraging spatial strategies for backward recall and phonological strategies for forward recall in normal subjects. We discuss additional possible implications of our model and its assumption of spatial organization of WM content and access through shifts of attention.


Keywords: Memory Manipulation; Operational Schema; Forward Recall; Reverse Recall; Computational Modelling; Intellectual Tasks; Working Memory.

## Introduction

Cognitive psychologists use the term working memory (WM) to emphasize on the use of temporarily stored information in connection with cognitive tasks that involve processing information (Baddeley, 1992). However, a review of the literature shows that the information processing aspect of cognitive tasks mostly applies to and have been explored using intellectual tasks with symbols. Tasks such as random digit generation, forward and backward digit or word span, mental arithmetic, $n$-back recall, double counting and sorting are prevalent in the context of cognitive psychology to explore the ability for manipulation and maintenance of information in working memory (see Repov and Baddely's review paper (Repovs \& Baddeley, 2006)).

Although the credit for popularizing the term working memory goes to cognitive psychologists, the concept of working memory as the ability of temporarily storing information for the use in the upcoming task has been applied in other domains and to tasks that lack symbolic or intellectual features. For example, working memory which is of the interest in the perception community is related to maintenance and manipulation of information for sensory tasks such as visual search (Oh \& Kim, 2004), or in action-perception domain for performing action routines (Arbib, 1987).

However what distinguishes WM in different domains is beyond differences in particular instances of information and indeed is mostly related to their execution models: the functional principles for management or manipulation of information. In particular what different models of working memory
in the domain of cognitive psychology ( CP ) share is a metaconcept for their execution models which can be referred to as the dichotomy of process-storage. Applying this metaconcept to management of information for human working memory was one the most important contributions of Alan Baddeley and Graham Hitch to cognitive psychology which was originally presented in their seminal work (Baddeley \& Hitch, 1974), and ever since has become the common denominator of all models of WM in CP. In this dichotomy which was inspired by Von Neumann's architectural design for modern digital computers (Von Neumann, 1982), the role of execution and processing is given to a central processing unit -namely the Central Executive (CE)—which controls the flow of information between and within storage slave units. However, a long debate over the nature of storage in CP community (Jonides et al., 2008) has restricted elaborations on functional mechanisms of CE.

The concept of CE in WM management did not prove as successful as its counterpart in Von Neumann's proposal in achieving a working memory management schema which helps information processing. What distinguishes the central processing unit (CPU) in Von Neumann's architecture from CE in Baddeley's proposal is that the CPU had all mechanisms for control of storage units built in, while Baddeley and Hitch use the central executive as a metaphor for a central and powerful executive unit with no specific detail as to how CE controls slave storage units (Baddeley, 1992). As Baddeley himself has stated in several occasions because of this lack of specificity, CE has become the rag bag of unanswered questions(Repovs \& Baddeley, 2006) or a homunculus (Repovs \& Baddeley, 2006; Baddeley, 1996). What is known about the executive role of the central executive, for the most part, is postulated by Baddeley and colleagues. Inspired by Norman and Schallice's idea of the Supervisiory Attentional System (SAS) (Norman \& Shallice, 1986), Baddeley has proposed that CE plays a role in controlling limited resources of executive attention (Repovs \& Baddeley, 2006; Baddeley, 1996). However, adding the function of controlling executive attentional resources has not been able to fill the void of a paradigm for an executive model for manipulation of information and to yield a model that explains how executive paradigm are encoded.

To give an example of an alternative meta-concept for manipulation of information we can refer to Arbib's work on information processing in perception-action loops. Arbib in his neuroethologic studies used the concept of information processing in a mechanistic fashion (Arbib, 1980) which was in-
formed by Norbert Wiener's theory of control and the concept of interplay between information and action in controlling biological organisms (Wiener, 1948). The term 'Schema' was the key concept in Arbib's terminology for describing how neural systems interplay to exchange information to achieve a biological goal (Arbib, 1992). In his later work on modeling visually-guided actions, he included the concept of working memory as a mechanism for sustaining information representations relevant to upcoming actions, as long as they remain relevant (Arbib, 1987).

While we share our target of study with contemporary cognitive psychologists, in devising an execution model for manipulation of information in the intellectual and symbolic domain, we are influenced by Wiener's system-theoretic and biologically-plausible concept of flow of information, and by Arbib's schema-theoretic approach (Arbib, 1992). We also have one specific additional assumption for our execution model for memory manipulation, which is the use of spacesupported sensory-motor systems in manipulation of information. We argue that, from an evolutionary standpoint, it is plausible that the capacity of performing intellectual symbolic tasks, which are very recent in our evolutionary history, might have emerged by re-using or co-opting rudementary systems for action and perception. Thus, we try to re-use sensory-motor systems as the building blocks of our approach to a model of working memory for intellectual working memory tasks.

We refer to this schema-driven and space-supported sensory-motor system which provides a random-accessible memory system for manipulation of information as the spatial registry system (SRS). The following section of this paper explains the general concepts of SRS. To demonstrate the power of this paradigm, we present a simple SRS model for immediate forward and backward recall. We show that the model explains the human patterns of errors in backward recall, which has been argued to utilize space for memory organization. Finally, we discuss what we learned from this simulation effort.

## The Spatial Registry System (SRS)

The focus of this section is description of a system for random access of symbolic content of working memory during intellectual mental tasks. Random-accessible working memories are the critical components for dynamic manipulation of information. Yet, they are not the only working memory systems in the context of intellectual working memory tasks. We later discuss a serially-accessible memory as an additional utility memory which collaborates with SRS systems for achieving a complete functioning working memory system.

We propose that symbolic items of the working memory can register with spatial locations in a grounded sensorymotor system which is supported by a spatial representation. Examples of such system -as we presented elsewhere (Noori \& Itti, 2011) - can be occulomotor system, or a kinesthetic system that helps proper configuration of body parts in
space using proprioception and muscle movements (think of a profoundly blind individual's ability for performing tasks in space without any visual reference).

This registry mechanism provides spatial addressing for random access to items of working memory. What is critical is how this addressing is used in the process of memory manipulation. The critical component is spatial selective attention (SSA) as a means to shift between items that are registered with space. An operational schema (OS) defines the sequence of shifts between registry locations.

For example, imagine the case of a concurrent mental headcounting of adults and children in a party. As your gaze shifts to a person in the living room, first your visual system becomes engaged in identifying whether the person at focus is an adult or a child. In the next step, one of two running counts that matches the identified category should be increased by one. The challenge is keeping track of two numbers and associating them to categories. A spatial registry strategy is associating the existing count of adults $n_{a}$ to location $l_{a}$ (e.g., left side in visual field or under pinky finger of the left hand) and the existing count of children $n_{c}$ to location $l_{c}$ (e.g., right side of visual field or under index finger of the left hand). Identifying the next child will trigger a shift of spatial attention to $l_{c}$, to fetch the current count of children. Once the increment operation is applied on the current count the result will replace (by first deletion and then insertion) the old count. Note how attention shifts might be used both for perception of the external world and for selection of WM items, which, under the SRS hypothesis, might give rise to conflicts in some situations, which in turn provides ways to test the hypothesis (see Noori and Itti's paper in this proceedings (Noori \& Itti, 2013) where they report the effect of congruency of shift of spatial attention for target detection and shift of selective attention in internal domain during triple-counting of visual targets).

The operational schema can be conceptualized as a list of mappings of the current state onto the next action. Here is a formal representation of an alternative OS for our head-counting scenario.
OS $1:\left\{\right.$ child $\Rightarrow$ shift to $l_{c} ;$ adult $\Rightarrow$ shift to $\left.l_{a}\right\}$
$O S_{2}:\left\{\right.$ at $l_{c} \Rightarrow$ fetch $n_{c} ;$ at $l_{a} \Rightarrow$ fetch $\left.n_{a}\right\}$
$O S_{3}:\left\{\right.$ at $l_{c} \& n_{c}$ is retrieved $\Rightarrow n_{c} \rightarrow n_{c}+1$
; at $l_{a} \& n_{a}$ is retrieved $\left.\Rightarrow n_{a} \rightarrow n_{a}+1\right\}$
$O S_{4}:\{$ shift the gaze to next person \& identify the category $\}$
Each of these schemas may include other sub-schemas. For example $n_{c} \rightarrow n_{c}+1$ in $\mathrm{OS}_{3}$ may include a sequence of operations over internal representation such as deletion and insertion (binding to space).

We have discussed the neural evidence for this hypothesis elsewhere (Noori \& Itti, 2011). Here we only mention one neuropsychological study which provides a critical evidence for our proposed model(Koenigs, Barbey, Postle, \& Grafman, 2009). Koenig et al. showed that patients who have sustained damage to their superior parietal lobule (SPL) generally lose their capacity for mental operations that need rearrangement
of information and thus they concluded that SPL is critical for manipulation of information in working memory. Interestingly, SPL is a part of the association cortex in the posterior parietal cortex (PPC) and sits at the junction of several sensory processing regions, with projections to motor area of the brain. SPL is shown to be critical for a wide range of routines that need sensory-motor integration, such as navigation, visual search, etc (Rizzolatti, Fogassi, \& Gallese, 1997).

We need to add that the spatial registry system is not a unitary system and several SRS systems might collaborate in running the executive machinery of working memory. However, what all SRS instances have in common is, first, their build-in internal space representation, second, a mechanism to shift the attention to those locations, and, third, a binding mechanism which can associate locations with symbolic representations.

In our view, spatial registry systems are complemented by other systems that mimic a serially-accessible memory, to provide a layer of working memory for intellectual symbolic tasks. An example of such system can be the sensory-motor system that supports speech perception-vocalization (Wilson, 2001) which is believed to be critical for spoken language acquisition (Baddeley, Gathercole, \& Papagno, 1998).

## A Spatial Registry System for Serial Recall

Here we present an SRS model for the immediate serial recall task for both forward and backward recall directions. As we will discuss later in our review of the literature, experimental evidence suggest that forward and backward recall draw on different brain systems. Forward recall seem to take impact from phonological characteristics of the list items (Bireta et al., 2010), suggesting that forward recall relies on phonological resources of the brain. On the other hand, backward recall is disrupted in individuals with deficit in spatial cognition (Rudel \& Denckla, 1974), suggesting that backward recall relies on spatial encoding. We used a spatial registry model applied to both directions of recall; however, as we will discuss, we learned that our SRS model provides a better fit for human behaviour, which conforms with what is assumed about involvement of different systems of working memory in two immediate recall directions.

## A brief review of the literature

Serial digit span tasks are common in both clinical assessment and neuropsychological studies (Rudel \& Denckla, 1974). However, forward recall, disproportionately, has received more attention in modeling attempts. This is mainly related to the importance of temporal serial order in everyday tasks (Glasspool, 2005). As the result, there are many neural models, behavioural models, and mathematical models dedicated to describing forward recall. In contrast, for backward recall, theoretical efforts mostly have focused on augmenting or reusing models of forward recall. In the face of abundant behavioural and neural evidence that serial recall in forward and backward directions draw on different brain mechanisms, it is not surprising that models of backward recall
have obtained remarkably less success in describing human behaviour compared to forward recall (Bireta et al., 2010). Only flexible mathematical models with enough degrees of freedom ,such as the Temporal Ratio Model (Brown, Neath, \& Chater, 2007), have been able to successfully model both recall tasks in one shot (Bireta et al., 2010).

In terms of similarity, recalling in both orders shows recency and primacy effect (Henson, 1996; Li \& Lewandowsky, 1995). Yet ,in forward recall a stronger primacy effect is observed (Henson, 1996), while in backward recall a stronger recency effect is observed ( Li \& Lewandowsky, 1995). Several studies have revealed the difference between recalling in two directions. Bireta et al. tested four benchmark effects that demonstrate the role of phonological resources in immediate forward recall tasks the word length effect, the irrelevant speech effect, the acoustic confusion effect and the concurrent articulation effectfor both directions of recall. They reported that the benchmark effect 'was either absent or greatly attenuated with backward recall despite being present with forward recall'. On the other hand, Li and Lewandowsky observed that altering visual-spatial characteristics of the recall list affected backward recall and not forward recall (Li \& Lewandowsky, 1995).

Neuropsychological evidence also supports that neurological damage to phonological resources of the brain impairs forward digit span while damage to spatial resources of the brain impairs backward digit span (Rudel \& Denckla, 1974). Consistent with these observations, neuroimaging studies also have revealed differences in cortical regions which are active during the two different recall orders (Sun et al., 2005; Hoshi et al., 2000). In particular, these studies have revealed significant activation of cortical areas with spatial processing characteristics in backward recall compared to forward recall.

In terms of modling efforts, Bietra et al. have briefly reviewed existing models. Their review indicated that those models that take the phonological aspect of forward serial recall are not successful in modeling backward recall, and only models that are agnostic to the difference in underlying mechanisms of serial recalls in two different directions are relatively successful in modeling both tasks.

In sum, available evidence suggests that a model that confers a special role to space may be necessary for a mechanistic model for the backward recall. In the following section we detail such a model, built based on the specifications of the SRS model for visuospatial working memory as the spatial registry.

## Simulation

A population coding of a one dimensional space in the form of an array of neurons was used as the registry space. Population coding of neurons has been extensively explored (Pouget, Dayan, \& Zemel, 2000) in the literature and is popular for neural modeling of visuospatial working memory (Constantinidis \& Wang, 2004). This array of neurons encoded a parametric space spanning the range of -1 to 1 . The
tuning curve for neurons in this array was characterized by $\sigma_{0}+x_{n} \times \kappa$ where $\sigma_{0}$ is the tuning band parameter of the neuron at the center of space, $x_{n}$ is neuron's peak response location, and $\kappa$ a constant which controls the variability of tuning band in the array of neurons.

Registering with a specific location would trigger noisy activation in the population around the target memory field. The share of a registry at $x_{r}$ in activation amplitude of a neuron at $x_{n}$ is determined by $A_{0} e^{-\frac{\left(x_{r}-x_{n}\right)^{2}}{2 \times \sigma_{n}{ }^{2}}}$. In case of registering several items in the activation of a neuron is defined as the sum of evoked signals of all registries as long as the sum of signals is less than a saturation value $\mathcal{S}$. So the base response amplitude of neuron $n$ is defined as follow:

$$
\begin{equation*}
\max \left(\mathcal{S}, \sum_{i=1}^{N} A_{i}(t) \times e^{-\frac{\left(x_{r_{i}}-x_{n}\right)^{2}}{2 \times \sigma_{n}^{2}}}\right) \tag{1}
\end{equation*}
$$

where $i$ is the index for registered items, $x_{r_{i}}$ is the registry location of the item $i$ and $A_{i}(t)$ denotes the effective amplitude of the ith registry at time $t$ which is defined by:

$$
\begin{equation*}
A_{0} \times e^{-\frac{t-t_{i}}{\tau_{d}}} \tag{2}
\end{equation*}
$$

where $t_{i}$ is the registry time of item $i$ and $\tau_{d}$, the damping factor, controls the decay rate of registry effects.

The schema for the immediate recall task includes two phases: binding and recall. During the binding phase, independent of the recall order, items of the list orderly register with locations from left to right so that each item in the list registers on the right side of previously registered item (except the first item). The exact times and locations of registries are perturbed by different random distributions. The distances between registry locations are determined by a Weibull distribution with two parameters (shape factor and scale factor). Duration of registry and recall processes are defined by two separate Gaussian distributions, which adds four more parameters to our model.

In the recall phase, a part of the schema is independent of recall direction, which is the condition for identifying the most active neuron, and for selecting the next item (until all items are removed from the registry space). Neurons in the array compete for gaining control of a registry recalling unit. The item at the closest registry location to the selected neuron will be recalled. Recalling memory items from registry involves inhibiting neurons in the array associated with registration of the recalled item.

Another part of the recall schema which is sensitive to the direction of recall is characterized by a bias. The bias is applied by a multiplicative exponential factor of the position which acts as a biased modulation of neural activities. For forward recall, this bias will enhance the activity of neurons on the left side of the space, and during the backward recall this bias enhances activities of neurons on the right side of the space. So, a part of the schema for recall is selecting the bias direction; however, once the bias direction is selected items will be selected only based the order of most active neurons.

| Par | Description | Par | Description |
| ---: | :--- | ---: | :--- |
| $\sigma_{0}$ | Spatial tuning at the center | $\kappa$ | Tuning band var factor |
| $\beta$ | Bias factor | $\nu$ | Noise factor |
| $\tau_{d}$ | Damping factor | K | Binding shape factor |
| $\lambda$ | Binding scale factor | $\mathcal{S}$ | Saturation factor |
| $\mu_{b}$ | Mean for binding duration | $\sigma_{b}$ | STD for binding duration |
| $\mu_{r}$ | Mean for fetching duration | $\sigma_{r}$ | STD for fetching duration |

Table 1
Parameters of the SRS model for serial recall


Figure 1: Positional error for the best two-way fit for both directions of recall compared to the human performance.

This implementation only accommodates positional or movement errors in which items are recalled in the wrong order. This type of error is the most prevalent error among adults (McCormack, Brown, Vousden, \& Henson, 2000) in recall tasks. However there are other types of errors such as omissions, intrusions and repetitions with less significant effect. Table 1 summarizes all parameters of this implementation.

## Results and Discussion

To explore tuning parameters we used serial position error for a list of five items from Li and Lewandowsky's study (Li \& Lewandowsky, 1995). An evolutionary algorithm was used to optimize the parameters based on the sum of absolute distance of predicted positional error over the ground truth data for both directions. So optimization of parameters was performed with regard to ground truth data for both directions simultaneously and forward and backward error data played equal roles in the evaluation function. However a closer inspection of the result revealed that the final parameters shifted in favour of the backward data. The best fitting parameters among 2857 independently generated solutions yielded a prediction for backward recall with $5.6 \%$ absolute distance to the human data (out of $500 \%$ maximum possible distance) while the same set of parameters yielded a prediction for the forward recall with $14.8 \%$ absolute distance to the human data (see Figure 1). Further analysis of best first 100 independent solutions of the optimiza-
tion process showed that the quality of predicted solution for backward recall was significantly better than forward recall $(t(198)=47.93, p<0.0001)$, where the difference between mean of fitness qualities was $6.6 \%$ in favour of the backward recall.

Moreover, a closer inspection of all generated parameter sets during during optimization process revealed two highly distinguishable modes for $\sigma_{0}$, the first order tuning curve parameter. A population of solutions with narrow tuning curve at the center peaked around $\sigma_{0}=0.04$ which included 847 solutions all with $\sigma_{0}<0.1$. Another population of wide tuning curve at the center peaked around $\sigma_{0}=0.57$ all with $\sigma_{0}>0.38$ included 2010 solutions. Later analysis of the fitness values of these solutions showed that the population of wide tuning curve (WTC) on average scored better fitness value than the population of narrow tuning curve (NTC). The wide tuning curve population (WTC) generally scored better in each of recall types compared with the narrow tuning curve population. Moreover, WTC and NTC populations were also highly separable with regard to other parameters including the bias factor, and temporal characteristics of binding and recall of item. In particular the for WTC the average duration of the task was correlated with the damping factor of neural activity while the duration of the task was independent of the damping factor of neural activities. In sum, WTC population provided both better solutions and more plausibility.

To test the predicting power of the model we used the parameter of the best solution discovered in the optimization of the previous phase to simulate the movement errors (the distance between order of an incorrectly recalled item, and its true order; e.g., if item 3 is recalled as item 2, the movement error is 1) in forward recall data for six items, from another study (McCormack et al., 2000). Note that number of items for training was different than for testing. Moreover, positional error data, which is used for optimization of parameters, is independent of movement errors (which we confirmed through simulation, not shown here).

Figure 2 shows the result of our simulation in the same graph with the data of two adult human subject groups, tested in two different experiments with different settings for a forward recall task (McCormack et al., 2000). Our simulation result sits in between the data points for two different results for adult human subject groups, which demonstrates that our prediction is in the range of the variability of the performance of human subjects, and clearly demonstrates the predictive power of the model.

In sum, the result of our simulation shows that SRS for immediate serial recall can account for human behaviour. However, as it was explained, the quality of our solution for backward recall is significantly better than the quality of our result for forward recall once both recall orders played the same role in optimization of parameters. This may be related to the fact that normal subjects leverage their phonological resources for forward recall (Bireta et al., 2010).

This does not mean that visual-spatial resources cannot be
used for forward recall. In fact previous studies have shown that articulatory suppression during working memory task with written verbal material can eliminate the effect of other signature effects such as word length effect or acoustic confusion effect without diminishing subjects' capacity for remembering the serial order (Wilson, 2001). These evidences suggest that once the speech recognition-vocalization system as the primary source of encoding serial recall is no longer accessible (by articulatory suppression) and working material are presented in visual format, another mechanism is utilized for encoding serial recall which does not rely on phonological resources. We argue that one could use a visual-spatial strategy for forward recall too. In this case, the prediction of our SRS model is that the overall performance would not be significantly better (see Figure 1a). However, using phonological resources for the forward recall has at least one advantage, which is freeing visual-spatial resources for other tasks. In contrast the ability to perform a backward recall task with impaired spatial resources is restricted (Rudel \& Denckla, 1974), in agreement with our finding that visualspatial resources are used for backward recall.


Figure 2: Prediction of SRS model along experimental data for movement errors during serial recall of six items.

## General Discussion

In this paper we presented the idea of a space-supported, schema-driven, random-accessible memory system in the domain of intellectual working memory tasks of cognitive psychology. Our proposal included a strong evolutionary assumption about what would constitute an executive model for a working memory system in the intellectual domain, which can be built atop sensory-motor systems that support perception-action routines. Perception-action routines, such as prey catching, evolutionarily, preceded the intellectual routines, such as mental subtraction, and thus we suggest that sensory-motor working memory systems for regulating the former routines might have been reused for maintaining and
manipulation of information needed by the evolutionary more recent latter routines.

The presented model provides a randomly-accessible working memory, yet to get a full-function working memory model that explains human behaviour across different domains, one may need to take a serially-accessible working memory subsystem into consideration too. A speech perception-vocalization subsystem -which resembles the phonological loop in Baddeley and Hitch's three-component model of working memory (Baddeley \& Hitch, 1974)- may be considered as an alternative serial component of working memory machinery in the domain of symbolic tasks. However, we argue that a serial system is not sufficient to explain humans' flexible memory manipulation of symbolic information in the intellectual domain, and one may need to include a faster and more flexible working memory system for random access to its content.

Finally one may argue that the visuospatial sketchpad in Baddeley and Hitch's model (Baddeley \& Hitch, 1974) can achieve the same function of our proposed SRS system. We can summarize the differences or our spatial registry system and the visuospatial sketchpad as follow.

First, while visuospatial sketchpad is merely a visualspatial system our SRS is a generic schema-driven system and as previously suggested several instances of sensory-motor working memory systems (e.g. occulomotor or kinesthetic system) may fulfil the characteristics of SRS.

Second, our SRS comes with a built-in executive system in the form of the operational schema (OS), while the visuospatial sketchpad outsources the execution functions to the CE, with no specifications of how this executive functions are exerted. In this sense, SRS provides a mechanistic model of manipulation of WM items while the sketchpad is a passive storage resource.

Third, in Baddley's model visuospatial sketchpad is a domain-specific storage slave unit which stores task-relevant visual-spatial information, while in our proposal an SRS system may play a general role in manipulation of symbolic information with no immediate visual or spatial features.

## Acknowledgments

This work was supported by the National Science Foundation (CRCNS Grant No. BCS-0827764), the Army Research Office (W911NF-11-1-0046), and the U.S. Army (W81XWH-10-2-0076).

## References

Arbib, M. (1980). Interacting schemas for motor control. Tutorials in motor behavior, 71-81.
Arbib, M. (1987). Levels of modeling of mechanisms of visually guided behavior. Behavioral and Brain Sciences, 10(03), 407436.

Arbib, M. (1992). Schema theory. Encyclopedia of artificial intelligence, 2, 1427-1443.
Baddeley, A. (1992). Working memory. Science, 255(5044), 556.
Baddeley, A. (1996). Exploring the Central Executive. The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology, 49(1), 5.

Baddeley, A., Gathercole, S., \& Papagno, C. (1998). The phonological loop as a language learning device. Psychological review, 105(1), 158.
Baddeley, A., \& Hitch, G. (1974). Working memory. The psychology of learning and motivation, 8, 47-89.
Bireta, T., Fry, S., Jalbert, A., Neath, I., Surprenant, A., Tehan, G., et al. (2010). Backward recall and benchmark effects of working memory. Memory \& cognition, 38(3), 279-291.
Brown, G., Neath, I., \& Chater, N. (2007). A temporal ratio model of memory. Psychological Review, 114(3), 539.
Constantinidis, C., \& Wang, X. (2004). A neural circuit basis for spatial working memory. The Neuroscientist, 10(6), 553-565.
Glasspool, D. (2005). Serial order in behaviour: Evidence from performance slips. Connectionist models in cognitive psychology, 241.

Henson, R. (1996). Unchained memory: Error patterns rule out chaining models of immediate serial recall. The Quarterly Journal of Experimental Psychology: Section A, 49(1), 80-115.
Hoshi, Y., Oda, I., Wada, Y., Ito, Y., Yamashita, Y., Oda, M., et al. (2000). Visuospatial imagery is a fruitful strategy for the digit span backward task: a study with near-infrared optical tomography. Cognitive Brain Research, 9(3), 339-342.
Jonides, J., Lewis, R. L., Nee, D. E., Lustig, C. A., Berman, M. G., \& Moore, K. S. (2008, January). The Mind and Brain of ShortTerm Memory. Annual Review of Psychology, 59(1), 193-224.
Koenigs, M., Barbey, A. K., Postle, B. R., \& Grafman, J. (2009). Superior Parietal Cortex Is Critical for the Manipulation of Information in Working Memory. J. Neurosci., 29(47), 14980-14986.
Li, S., \& Lewandowsky, S. (1995). Forward and backward recall: Different retrieval processes. Journal of Experimental Psychology: Learning, Memory, and Cognition, $21(4), 837$.
McCormack, T., Brown, G. D. A., Vousden, J. I., \& Henson, R. N. A. (2000, July). Children's Serial Recall Errors: Implications for Theories of Short-Term Memory Development. Journal of Experimental Child Psychology, 76(3), 222-252.
Noori, N., \& Itti, L. (2011). Spatial Registry Model: Towards a Grounded Account for Executive Attention. Proceedings of the 33rd Annual Conference of the Cognitive Science Society, 31873192.

Noori, N., \& Itti, L. (2013). Where what you count is what really counts. In Proceedings of the 35th annunal conference of the cognitive science society.
Norman, D. A., \& Shallice, T. (1986). Attention to Action: Willed and Automatic Control of Behavior. In Consciousness and selfregulation (Vol. 4, pp. 1-18). New York: Plenum.
Oh, S., \& Kim, M. (2004). The role of spatial working memory in visual search efficiency. Psychonomic Bulletin \& Review, 11(2), 275-281.
Pouget, A., Dayan, P., \& Zemel, R. (2000, November). Information processing with population codes. Nat Rev Neurosci, 1(2), 125132.

Repovs, G., \& Baddeley, A. (2006, April). The multi-component model of working memory: Explorations in experimental cognitive psychology. Neuroscience, 139(1), 5-21.
Rizzolatti, G., Fogassi, L., \& Gallese, V. (1997). Parietal cortex: from sight to action. Current opinion in neurobiology, 7(4), 562567.

Rudel, R., \& Denckla, M. (1974). Relation of forward and backward digit repetition to neurological impairment in children with learning disabilities. Neuropsychologia, 12(1), 109-118.
Sun, X., Zhang, X., Chen, X., Zhang, P., Bao, M., Zhang, D., et al. (2005). Age-dependent brain activation during forward and backward digit recall revealed by fMRI. Neuroimage, 26(1), 3647.

Von Neumann, J. (1982). First Draft of a Report on the EDVAC. In R. B. (Ed.), The Origins of Digital Computers (pp. 383-92). Springer-Verlag, Berlin.
Wiener, N. (1948). Cybernetics; or control and communication in the animal and the machine.
Wilson, M. (2001). The case for sensorimotor coding in working memory. Psychonomic Bulletin \& Review, 8(1), 44-57.

# Auditory Overshadowing in Preschoolers: A Preference for the Input, the System, or Both? 

Allison O’Leary (oleary.83@osu.edu)<br>Department of Psychology, 1835 Neil Avenue<br>Columbus, OH 43210, USA

Vladimir M. Sloutsky (sloutsky.1@osu.edu)<br>Department of Psychology, 1835 Neil Avenue<br>Columbus, OH 43210, USA


#### Abstract

Auditory overshadowing occurs when the presence of an auditory stimulus interferes with visual processing. The current study tested whether this occurs due to a privileged attentional status of auditory input or due to the dynamic characteristics of auditory input. To address these questions, preschoolers completed one of four discrimination tasks. In the sound, motion, and item baseline conditions, children discriminated these single information types by judging whether paired stimuli were the same or different. In the combined condition, children discriminated changing sounds, motions, or items in the face of competing input in the other two dimensions. Although children's discrimination of all information types attenuated in the combined condition relative to baseline, motion and item discrimination attenuated more than auditory discrimination. This provides evidence that early in development auditory information receives privileged processing in the face of competing input.


Keywords: Attention; Cross-modal processing; Cognitive development; Preschoolers

## Introduction

To successfully navigate and form an understanding of our environment, we must develop the ability to efficiently integrate information from multiple sensory modalities. There is evidence that some multisensory integration occurs even in neonates, as demonstrated by the structural convergence of input from different sensory modalities (Stein \& Meredith, 1993). However, it is clear that the ability to integrate multimodal information is not fully mature at birth and, in fact, exhibits a protracted developmental trajectory. For example, in contrast to normally functioning adults, infants and young children exhibit a phenomenon known as auditory overshadowing (Robinson \& Sloutsky, 2004b). This occurs when the presence of an auditory stimulus interferes with one's processing of a visual stimulus. That is, infants and young children more easily discriminate visual stimuli when presented in isolation than when paired with labels or sounds. In some cases, children completely fail to discriminate changes in a visual stimulus when presented simultaneously with a sound (Napolitano \& Sloutsky, 2004). Importantly, although visual discrimination is impaired in the presence of auditory information, auditory discrimination does not suffer in the presence of visual
information (Sloutsky \& Napolitano, 2003; Robinson \& Sloutsky, 2010b).

The importance of audio-visual integration is especially apparent for processes like word learning, where individuals must map verbal labels to objects in visual space. If visual processing is inhibited when an auditory stimulus is present, auditory overshadowing could be a major contributor to the difficulty children face mapping words to objects. Indeed, 10 -month-olds encoded only auditory information when presented with visual information and a verbally presented label. However, 16-month-olds demonstrated able processing of both the visual information and the label (as indicated by looking time preferences; Robinson \& Sloutsky, 2004a).

Although the ability to process auditory and visual information progresses through infanthood, even preschoolaged children show difficulties discriminating static visual stimuli in the presence of sounds and labels. Napolitano and Sloutsky (2004) demonstrated that 4-year-olds were susceptible to auditory overshadowing when asked to discriminate changes in the visual and auditory aspects of a target stimulus. Specifically, this effect was most profound when the visual and auditory stimuli were unfamiliar. Further, the authors demonstrated that this overshadowing was resistant to explicit instruction, in that it still occurred when children were asked to attend exclusively to visual information.

More recent research has aimed to detail the conditions in which auditory overshadowing occurs and to elucidate the basic cognitive mechanisms underlying these effects. Input from the visual and auditory modalities seem to "compete" for processing resources early in development, whereas adults are able to process multimodal information (Robinson \& Sloutsky, 2004a). Robinson and Sloutsky (2007a) also argued that overshadowing may occur during initial processing/encoding as well as during response selection. In addition, unfamiliar auditory stimuli slow visual processing, whereas familiar auditory stimuli do not (Robinson \& Sloutsky, 2007b). Despite our understanding of the subtleties of the effect, major questions remain regarding the privileged status of auditory input and a number of hypotheses explaining auditory overshadowing effects have been proposed.

One possibility is that auditory and visual information are processed serially, and that auditory stimuli are processed more thoroughly because they are faster to engage attention. Visual processing may be inhibited until attention is disengaged (Robinson \& Sloutsky, 2010a). However, there are at least two possible explanations as to why auditory input would more easily engage attention. First, adults process auditory information more quickly than visual information, which may reflect privileged processing of auditory input per se (Green \& von Gierke, 1984). Because the auditory system begins maturing before the visual system, it is possible the auditory system processes information more efficiently even during childhood. Alternatively, it is possible that the nature of the input plays a role. For example, auditory stimuli are typically dynamic (i.e., exhibit change over time) and have more abrupt onset than visual stimuli. Perhaps it is this dynamicity that quickly engages attention.

However, previous research could not distinguish between these possibilities because all previous studies of auditory overshadowing have utilized static visual stimuli. The current study aimed to test these explanations by increasing the dynamicity of visual stimuli. This was done by adding motion to the visual stimulus, which is a powerful bottom-up attentional cue (Egeth \& Yantis, 1997). If the dynamic nature of auditory input is responsible for auditory overshadowing, one would expect the increased dynamicity of the visual stimulus to attenuate auditory overshadowing, resulting in more thorough visual processing. In this situation, visual information may even overshadow auditory information. On the other hand, if the auditory system itself is privileged, one would not expect such attenuation, even if the visual cue is dynamic. Distinguishing between these possibilities would result in better understanding of the mechanisms underlying auditory overshadowing.

In previous studies where young children were presented with auditory and visual information, visual discrimination attenuated significantly whereas auditory discrimination showed no significant attenuation. The current study aimed to investigate the effect when presented stimuli had an auditory, visual, and motion component. In this "combined" condition, target and test stimuli included all of these components. Children were instructed to say "same" if the target and test stimuli were the same in all 3 aspects. Children were instructed to say "different," however, if the sound component changed, the motion component changed, or the item appearance component changed. To determine the extent of overshadowing, we assessed children's discrimination of changes in the sound, image, or motion of these more complex stimuli relative to their ability to discriminate these types of information when presented in isolation (i.e., sound, motion, and item discrimination baselines). If auditory overshadowing stems from the dynamic nature of sound, then the dynamic motion cue should attenuate auditory processing.

## Method

## Participants

Eighty-two four-year-olds (41 girls and 41 boys, $M=4.48$ years, $S D=.28$ years) participated in this experiment. Children completed one of four conditions: sound baseline ( $N=20$ ), motion baseline ( $N=20$ ), item baseline ( $N=20$ ), or a combined (sound-motion-item) condition ( $N=22$ ). Children were recruited through local daycares and preschools located in Columbus, Ohio. The majority of participants were Caucasian.

## Stimuli

In the combined condition, children saw two moving "toys" which were each paired with a sound, and were asked to discriminate changes in the sound, motion, or appearance of these toys. Each stimulus in this condition consisted of a sound, motion, and an object, referred to as "item". There were 8 sound, 8 motion, and 8 item appearance possibilities. All of these were intended to be novel to children, as overshadowing effects are sensitive to stimulus familiarity (Robinson \& Sloutsky, 2010b). For example, auditory stimuli consisted of dynamic sounds like camera clicks and notes from an organ. Each auditory stimulus was 1500 ms in duration. The 8 items all consisted of a central ' $X$ ' on which four colored shapes were placed. Each of these items was animated in 8 different motion patterns (e.g., $360^{\circ}$ rotation, looming) to produce 512 total sound-motion-item combinations. Each stimulus presentation proceeded in the following way: children viewed the static image in silence for 500 ms , after which the motion and sound began simultaneously, lasting for the remaining 1500 ms of stimulus presentation.

Each of the baseline conditions aimed to assess children's discrimination when a single type of information is presented. In the sound baseline, children were asked to discriminate 2 of the 8 sounds used in the combined condition. In this baseline, the only visual stimulus was a small fixation cross presented in the center of the screen. In the motion baseline, children were asked to discriminate 2 of the 8 motion types. Here, the same item was used throughout the task, and these visual stimuli were presented in silence. In the item baseline, children discriminated 2 of the 8 items, presented statically and in silence.

## Procedure

In all conditions, children observed pairs of stimuli presented sequentially and were asked to indicate whether the two stimuli were the same or different. Because of their novelty, the experimenter described the stimuli as "toys from outer space." Children were told they would see two of these toys and that their job was to tell the experimenter whether the two toys were different in any way (i.e., if they noticed a difference in the item, motion, or sound). During each trial, children were first presented with a fixation cross in the center of the screen. Once the child fixated on this


Figure 1. Target and test stimuli used in the sound, motion, and item baseline conditions. The top images depict example target stimuli. Bottom images depict potential test stimuli for that target.


Figure 2. Target and test stimuli used in the combined condition.
cross, the experimenter pressed the spacebar to present the target stimulus. The child then saw another fixation cross.

Upon the child's fixation, the experimenter pressed the spacebar to present the test stimulus. Each target and test stimulus was 4000 ms in duration. Following test stimulus presentation, the child was asked to indicate verbally whether the two toys they just observed were the same or different. Half of the trials were "same" trials, in that the target stimulus was identical to the test stimulus. The remaining half of trials were "different" trials, such that theitem, motion, or sound presented at test differed from that in the target. Each child completed 4 blocks which each consisted of 12 trials. Children saw a 12 -second cartoon between blocks.

In the combined condition, each target trial consisted of a moving toy paired with a sound. Each sound, motion, and item was selected randomly. During the "same" test trials, children were presented with a stimulus identical to the target stimulus. In one third of the "different" trials, only the sound changed at test, and the item and motion remained the same. In one third of these trials, only the motion of the item changed, and the item itself as well as the sound remained the same. In the remaining third of these "different" trials, the item changed, but the motion and sound remained the same. Of interest was whether children showed differential ability to discriminate these 3 types of information when presented simultaneously, relative to when presented in isolation (i.e., performance in the respective baselines).

The baseline conditions assessed children's ability to discriminate sound, motion, and item information individually. In the sound baseline, target trials consisted of an auditory stimulus (selected randomly from the 8 possibilities) presented with only a fixation cross in the center of the screen. During "same" trials, children heard the same sound. During "different" trials, children heard a different sound. The same structure applied to the motion and item baselines.

## Results

To ensure that children understood and were engaged with the tasks, only those children whose overall performance was above chance ( $50 \%$ accuracy) were included in subsequent analyses. This eliminated 2 children in the sound baseline condition, 2 children in the motion baseline condition, and 2 children in the item baseline condition.

## Baseline conditions

To evaluate the meaning of differential performance (discriminating sound, motion, and item) in the combined condition, it was necessary to first establish children's ability to discriminate each type of information presented independently. Children demonstrated high performance across the sound ( $M=.91, S D=.08$ ), motion ( $M=.87, S D$ $=.10)$, and item $(M=.92, S D=.07)$ baselines. A one-way ANOVA indicated no significant differences between children's performance in these three conditions, $F(2,51)=$ 2.03, $p=.142$.

A one-way ANOVA comparing children's reaction times to "different" trials in the 3 conditions (including only correct responses) revealed a main effect of information type, $F(2,51)=8.77, p<.005$. Post hoc multiple comparisons using Fisher's LSD revealed that individuals in the item baseline condition $(M=1486.81, S D=1066.36)$ discriminated more quickly than those in the motion ( $M=$ 2409.69, $S D=817.29$ ) and sound ( $M=2653.24, S D=$ 724.60) baselines. Average RTs in the motion and sound baselines did not differ significantly from one another. These results suggest that item discrimination was somewhat easier than discrimination of sounds or patterns of motion.

## Combined condition

In the combined condition, children indeed showed differential accuracy discriminating sound, motion, and item changes, $F(2,63)=7.11, p<.005$. Post-hoc multiple comparisons indicated children were significantly more accurate when discriminating sounds $(M=.70, S D=.25)$ than both motions ( $M=.48, S D=.30$ ), and items ( $M=.39$, $S D=.30$ ). Children's discrimination of motions and items did not differ significantly. Further, children's reaction times did not differ significantly when discriminating the different information types, $F(2,63)=1.45, p=.242$.

## Comparing baseline and combined conditions

Because the assessment of overshadowing requires the comparison of performance on the between-subjects baseline conditions and the within-subjects combined condition, we calculated difference scores to describe each individual's attenuated performance in the combined condition relative to baseline. For example, we subtracted the mean accuracy of all individuals in the sound baseline from each individual's average accuracy to "different sound" trials in the combined condition.

As indexed by these difference scores, children's performance in the combined condition attenuated significantly across all the information types (i.e., all difference scores differed from zero): sound, $t(21)=-3.83, p$ $<.005$; motion, $t(21)=-6.03, p<.001$; and item, $t(21)=-$ $8.29, p<.001$.

A one-way ANOVA comparing children's difference scores revealed a main effect of condition, $F(2,63)=7.472$, $p<.005$. Post-hoc multiple comparisons using Fisher's LSD revealed children's performance in the combined condition attenuated less in the auditory condition $(M=-.20, S D=$ $.25)$ than in the item condition $(M=-.53, S D=.30)$ as well as the motion condition ( $M=-.39, S D=.30$ ). Difference scores in the item and motion conditions did not differ from one another. Children's reaction times differed amongst the baseline and combined conditions only when it came to discriminating item information, $t(38)=-2.53, p<.05$.


Figure 3. Proportion of accuracy to "different" trials for sound, motion, and item information by condition.


Figure 4. Average reaction time to discriminate "different" trials for sound, motion, and item information by condition. Baseline reaction times include only correct responses.

In sum, children's discrimination of sound, motion, and item information attenuated when presented simultaneously rather than in isolation. Sound discrimination attenuated the least, followed by motion discrimination, and then item discrimination, which attenuated most. These results indicate that auditory information was more likely to overshadow visual motion than the other way around.

These findings, in turn, suggest that processing of auditory information may indeed be privileged compared to visual information, even if visual information is dynamic in nature.

## General Discussion

The current experiment compared children's ability to discriminate motion, sound, and item appearance information when presented simultaneously versus in isolation. When combined, children's discrimination of all 3 information types attenuated relative to baseline performance. This indicates that all 3 types of information were somewhat susceptible to interference from one another. It is well established that when auditory and static visual stimuli are presented simultaneously, processing of the visual stimulus suffers whereas auditory processing remains intact. Thus, the additional motion information in the current study contributes to the overshadowing of auditory processing. This provides support to the idea that the dynamicity of visual and auditory input plays a role in the allocation of attention to these two modalities.

Although children in the baseline conditions exhibited high accuracy across the board, the levels of attenuation in the combined condition differed significantly across information types. The most attenuation occurred for item discrimination, followed by motion discrimination, and the least attenuation to sound discrimination. The fact that auditory discrimination was overshadowed least suggests that the auditory system may have privileged status early in development.

Interestingly, children's reaction times in the item discrimination baseline were significantly faster than in the other baseline conditions. Note that accuracy in this condition was high and equivalent to the other conditions. Therefore, it is unlikely that differential attenuation occurred due to differences in baseline discrimination. If this had been the case, one would have expected the least overshadowing for item discrimination.

Overall, the current study demonstrates that there may be multiple influences at play in the manifestation of auditory overshadowing of visual information. Pulls on attention appear to be sensitive to increased stimulus dynamicity. In addition, auditory information receives privileged processing even in the face of competing dynamic visual input. In terms of the directionality of overshadowing effects, however, the current study is limited in its potential conclusions. Children were exposed to one information type in the baseline conditions and all three information types in the combined condition. It is impossible to speculate the direction of overshadowing effects without comparing the extent of children's overshadowing in the combined condition to conditions involving only two types of information. For example, this would allow us to compare overshadowing when sound, motion, and visual information are combined to the traditional overshadowing effect involving only auditory and static visual stimuli. Without this manipulation (which is currently in progress), we are unable to identify whether the additional motion
information contributes to the attenuation of item discrimination.

Further research will also need to elucidate whether increasing the dynamicity of visual input could benefit young children's word learning. Previous work indicates that, in some cases, novel sounds like the ones used in the current experiment overshadow visual input, whereas familiar sounds and words do not. Perhaps the use of more familiar auditory stimuli would result in less overshadowing overall. If this were the case, this increased attention to visual and motion information would provide more optimal conditions for learning object-word mappings. Future work could also investigate whether these mappings occur more easily when visual stimuli are presented dynamically.

In addition, it will be interesting to investigate the developmental trajectory of auditory overshadowing in the context of dynamic visual input. Perhaps there are times throughout development in which motion facilitates visual processing, and times when it hinders or distracts. Further, a developmental investigation can help us identify related processes that could help explain the typically observed reduction in overshadowing with age. For example, perhaps increases in working memory capacity are associated with the developing ability to bind auditory and visual information (Allen, Hitch, \& Baddeley, 2009). This research could illuminate ways of presenting particular stimuli to encourage processing between different modalities. Once we understand the developmental process contributing to audio-visual association, we can apply those principles to identify optimal learning conditions in order to teach children more efficiently.

## Conclusions

In sum, the current research points to several important findings. When auditory stimuli were presented with dynamic visual stimuli, children's processing of auditory information attenuated. This finding is the first evidence that visual information may interfere with auditory processing early in development. Although auditory discrimination attenuated, it attenuated less than the discrimination of motion patterns or item appearances. The combination of these findings indicates that auditory overshadowing may have multiple underlying causes. Children's attention to dynamicity in the visual modality seemed to pull attention ordinarily devoted to auditory processing. However, the auditory information was still processed better than motion or item appearance information in the face of this competing input. Thus, auditory overshadowing seems to be a function of privileged processing of auditory information.

## Acknowledgements

This research was supported by a National Science Foundation (NSF) grant (BCS-0720135) to Vladimir Sloutsky. We thank the children, parents, and preschools that participated in this research.

## References

Allen, R. J., Hitch, G. J., \& Baddeley, A. D. (2009). Crossmodal binding and working memory. Visual Cognition, 17(1-2), 83-102.
Egeth, H. E., \& Yantis, S. (1997). Visual attention: Control, representation, and time course. Annual review of psychology, 48(1), 269-297.
Green, D. M., \& Von Gierke, S. M. (1984). Visual and auditory choice reaction times. Acta psychologica, 55(3), 231-247.
Napolitano, A. C., \& Sloutsky, V. M. (2004). Is a picture worth a thousand words? The flexible nature of modality dominance in young children. Child Development, 75, 1850-1870.
Robinson, C. W., \& Sloutsky, V. M. (2004a). Auditory dominance and its change in the course of development. Child Development, 75, 1387-1401.
Robinson, C, W., \& Sloutsky, V. M. (2004b). The effect of stimulus familiarity on modality dominance. In K. Forbus, D. Gentner, \& T. Regier (Eds.), Proceedings of the XXVI Annual Conference of the Cognitive Science Society, 1167-1172 Mahwah, NJ: Erlbaum.
Robinson, C. W., \& Sloutsky, V. M. (2007a). Auditory dominance: Overshadowing or response competition? In D. McNamara \& G. Trafton (Eds.), Proceedings of the XXIX Annual Conference of the Cognitive Science Society (pp. 605-610) Mahwah, NJ: Erlbaum.
Robinson, C. W., \& Sloutsky, V. M. (2007b). Visual processing speed: Effects of auditory input on visual processing. Developmental Science, 10, 734-740.
Robinson, C. W., \& Sloutsky, V. M. (2010a). Development of cross-modal processing. Wiley Interdisciplinary Reviews: Cognitive Science, 1135-141.
Robinson, C. W., \& Sloutsky, V. M. (2010b). Effects of multimodal presentation and stimulus familiarity on auditory and visual processing. Journal of Experimental Child Psychology, 107, 351-358.
Sloutsky, V. M., \& Napolitano, A. (2003). Is a picture worth a thousand words? Preference for auditory modality in young children. Child Development, 74, 822-833.
Stein, B. E., \& Meredith, M. A. (1993). The merging of the senses. The MIT Press.

# The interaction of different working memory mechanisms and sentence processing: A study of the P600 

Polly O’Rourke (porourke@casl.umd.edu)<br>Center for the Advanced Study of Language<br>University of Maryland, $700552^{\text {nd }}$ Avenue<br>College Park, MD 20742 USA


#### Abstract

While previous research has shown that working memory capacity (WMC) predicts sentence processing ability, the understanding of the relationship is limited as almost all studies have used the reading span task as their sole measure of WMC. The current study examined how the effects of garden-path sentences and filler-gap dependencies (as indexed by the P600) related to four measures of working memory (reading span, operation span, anti-saccade and $n$-back). P600 effects for garden-path sentences correlated positively with operation span score while effects for object relatives correlated negatively with n-back accuracy. These results indicate that, though both sentence types are associated with increased working memory demands, the resolution of temporary syntactic ambiguity and filler-gap dependencies recruit distinct working memory mechanisms.


Keywords: Garden-Path; Object Relative, Reading Span, $N$ back, P600.

## Introduction

Two major sources of difficulty in sentence processing are temporary syntactic ambiguity and syntactic complexity (see sentences 1 and 2, respectively). Sentences containing temporary ambiguities (i.e. garden-path sentences) usually lead to an initial incorrect parse of the syntactic structure which must be reanalyzed when the temporary ambiguity is resolved in order for the sentence to be correctly interpreted. In sentence 1 below, the initial interpretation may be that the patient met the doctor and the nurse but upon arriving at the word "showed" (the disambiguating verb), it is apparent that that interpretation is incorrect and must be revised. A prime example of syntactic complexity is an object relative which represents a filler-gap dependency in which direct or indirect object is displaced from the verb from which it gets its thematic role. In order to resolve the filler-gap dependency, the object ("to whom" in sentence 2 below) must be maintained active until it can be mapped onto the thematic grid of the relevant verb ("showed").

1. The patient met the doctor and the nurse with the white dress showed the chart during the meeting.
2. The patient met the doctor to whom the nurse with the white dress showed the chart during the meeting. ${ }^{1}$
[^179]Behavioral research has shown longer reading times and reduced comprehension accuracy for sentences with gardenpath sentences (Frazier \& Rayner, 1982; Ferreira, Bailey \& Ferraro, 2002) and object relatives (Frazier, 1987; King \& Just, 1991) compared to simple controls. Research using the noninvasive event-related potential (ERP) technique of recording brain activity has also provided evidence for the increased difficulty of these sentence types. The key potential of interest is the P600, a positive shift which emerges 500 to 800 ms post-stimulus, typically largest over posterior sites. The P600 is generally considered to be an index of syntactic integration difficulty (Kaan et al., 2000). It is elicited by syntactic violations of all types but is also sensitive to syntactic ambiguity and syntactic complexity in well-formed sentences. Garden-path sentences elicit P600 effects relative to non-garden-path sentences (Osterhout, Holcomb \& Swinney, 1994; Kaan \& Swab, 2003; Gouvea, Phillips, Kazanina \& Poeppel, 2010). In addition, several studies have found P600 effects when comparing sentences containing object relative clauses to simple declarative sentences (Kaan, Harris, Gibson \& Holcomb, 2000; Gouvea et al. 2010). Gouvea et al. (2010) found that the P600 effect for garden-path sentences is more robust and of longer duration than that of unambiguous object relative structures.

One potential source of the increased difficulty is the increased demand these sentences place on working memory resources. Working memory (WM) is "a multicomponent system responsible for active maintenance of information in the face of ongoing processing and/or distraction" (Conway, Kane, Bunting, Hambrick, Wilhelm \& Engle, 2005, p. 770) which facilitates goal directed behavior. Individual differences in working memory capacity (WMC) impact sentence processing ability. A number of studies have found that individuals with high WMC have faster reading times and improved comprehension performance for garden-path sentences (Just \& Carpenter, 1992; Friederici, Steinhauer, Mecklinger \& Meyer, 1998) and object relative dependencies (Just \& Carpenter, 1992; King \& Just, 1991) than low WMC participants. The P600 is modulated by WMC as well. ERP studies have shown that individuals with high WMC have greater P600 effects at the disambiguation point of sentences containing temporary ambiguities (Friederici et al., 1998; Vos, Gunter, Schriefers \& Friederici et al., 2001; Vos \& Friederici, 2003; Bornkessel et al., 2004), indicating increased reanalysis in the high WMC participants. Neither

Bornkessel et al. (2004) nor Friederici et al. (1998) found group differences in the P600 effects for syntactically complex sentences which did not contain temporary ambiguities. This suggests that the resolution of syntactic ambiguity and of filler-gap dependencies differ in terms of working memory demands.

Understanding of the relationship between WMC and different types of syntactic processing is limited by the fact that the vast majority of studies that have examined the connection have used the reading span task (Danemann \& Carpenter, 1980) as the sole index of WMC, while many different assessments, which tap different working memory mechanisms, exist (Conway, Kane, Bunting, Hambrick, Wilhelm \& Engle, 2005). This makes it impossible to determine if the difference between the working memory demands of two syntactic processes are quantitative or qualitative. Parsing complex but unambiguous syntax may be less costly than reanalysis of garden-path structures, or it could recruit working memory mechanisms not indexed by the reading span task.

The goal of the current study was to determine if working memory mechanisms other than those indexed by reading span are relevant to the online sentence processing as indexed by P600 effect size. To this end, WMC was assessed using four different measures: reading span, operation span, $n$-back and anti-saccade. The reading span is a complex span task which assesses an individual's ability to maintain and process information (i.e. read sentences) under divided attention. Operation span, another complex span task, is very similar except that the processing component is the performance of mathematical operations. These two tasks correlate with each other (Conway et al. 2005) and both have been shown to predict sentence comprehension performance (Turner \& Engle, 1989). Also, as mentioned before, reading span has been found to predict P600 effect size for garden-paths (Friederici et al., 1998; Vos et al., 2001, Vos \& Friederici, 2003; Bornkessel et al., 2004). $N$-back performance reflects the ability to maintain, monitor and regularly update information. The relationship between $n$-back and sentence processing is unclear and largely untested. Sprouse, Wagers and Phillips (2012) found no evidence of a relationship between $n$-back performance and island effects (i.e. effects of syntactic complexity) on acceptability judgments. Novick and colleagues, however, found that individuals who improve their $n$-back performance via training show reduced garden-path effects in their comprehension accuracy relative to those who do not respond to training (Novick, Hussey, Teubner-Rhodes, Dougherty, Harbison \& Bunting, in press). The anti-saccade task tests the ability to suppress a prepotent response. Bilinguals are known to out-perform monolinguals in tasks tapping this skill (Bialystok, 2006, Bialystok, Craik, \& Luk, 2008), suggesting a possible connection with language processing. By including a wider range of working memory assessments, the current study aimed to enrich understanding of the cognitive underpinnings of sentence processing.

## Methods

## Participants

Data was collected from 65 right handed participants. Data from two participants was excluded because it was revealed that they didn't meet the participation criteria. Six participants were excluded due to technical issues with data collection. An additional 6 were excluded due to excessive EEG artifacts. Data from the remaining 51 participants (29 female) between the ages of 18 and 40 (mean age $=21.5$, S.D. $=2.33$ ) were included in the analysis. All participants were neurologically normal, native speakers of English. None had had started learning a second language before age 12.

## Sentence Stimuli

The sentence stimuli consisted of garden-path, object relative and control sentences (see 3-5 above, respectively. The critical word in each condition was a ditransitive verb ("showed" below). A total of 108 triplets were prepared using 108 different critical verbs such that each sentence in a triplet is identical except for the region at the beginning of the second clause (bolded below). Ninety of the 108 came from Gouvea et al. (2010)'s stimuli set. In the garden-path sentences, the critical verb indicated the need for reanalysis. In the object relative sentences, the critical verb indicated the thematic position of the wh-phrase ("to whom") and, thus, allowed the resolution of the filler-gap dependency. Each list contained 36 sentences in each condition. The presentation of sentences was counterbalanced such that each sentence appeared in one condition per list. In addition, 72 filler sentences (matched for length and complexity) were included. Fifty percent of all sentences were followed by a comprehension question. Questions came either from Gouvea et al. (2010)'s stimuli or were created for the new sentences. The comprehension questions did not specifically target the resolution of the garden-path structure. In total, six lists were created such that each sentence appeared in each condition, with or without a comprehension question, across lists.
3. The patient met the doctor and the nurse with the white dress showed the chart during the meeting. (Garden-Path)
4. The patient met the doctor to whom the nurse with the white dress showed the chart during the meeting. (Object Relative)
5. The patient met the doctor while the nurse with the white dress showed the chart during the meeting. (Control)

## Working Memory Tasks

Reading Span Automated Reading-Span (Unsworth, Heitz, Schrock \& Engle, 2005) was used in this experiment. Participants were presented with a series of sentences and asked to indicate, via button press, if the sentences make
sense. After each sentence they were then presented with a letter that they must remember. At the end of the sequence, they had to recall the letters in the order of presentation. Their score reflects the total number of letters recalled in the correct order.

Operation Span Automated Operation-Span (Unsworth et al., 2005) was used in this experiment. Operation span is identical to reading span as described above except instead of making sense judgments on sentences, participants had to solve math problems involving multiple operations.
$N$-Back In the $n$-back task, participants were presented with a sequence of single letters and asked to judge if the current letter is the same as the one that occurred $n$ places back in the sequence. For example, in a 4-back task, the third " X " in the following sequence would be a target: X U P X X U U. Lures, which appeared one space before a target ( $n-1$; the second "X") or one space after a target ( $n+1$; the third "U") were also included. Participants in the current experiment performed 2-back and 4-back. Accuracy for four item types (target, non-target, $n-1$ lure and $n+1$ lure) were averaged across $n$ level (2-back and 4-back).

Anti-Saccade In the anti-saccade task, participants performed a letter monitoring task. They were first presented with a flashing cue that appears on either the left or right side of the computer screen. The cue was followed by a letter. The letter was either on the same side of the screen as the cue (pro-saccade) or on the opposite side (antisaccade). Participants had to suppress the impulse to shift their gaze to the cue in order to maximize performance. Accuracy for anti-saccade trials was included in the analysis.

## EEG Recording

Electroencephalographic (EEG) data was acquired using the Electrical Geodesics Inc. (EGI) NetStation 128-channel system. The HyrdoCel Geodesic Sensor Net is an elastic structure containing $\mathrm{Ag} / \mathrm{AgCl}$ electrodes, individually housed underneath a sponge pedestal, which is soaked in a saline solution ( KCl ) and placed carefully over the participant's head. The signal was high-pass filtered online at 0.1 Hz , low-pass filtered at 100 Hz , and notch filtered at 60 Hz . The EEG signal was sampled at 250 Hz . Impedances were kept below $50 \mathrm{~K} \Omega$ where possible and otherwise under $100 \mathrm{~K} \Omega$. Prior to averaging, drift, eye blinks and other movement artifacts were corrected via either the EP Toolkit for MatLab (Dien, 2010). EEG were recorded using CZ as a reference and later re-referenced to the global mean.

## Procedure

After signing a consent form and background questionnaire, the experimenter applied the sensor net. Participants were seated in a sound attenuated booth using a chin rest in order to reduce movement artifacts. EEG data was collected during the sentence processing task. Sentences appeared
word-by-word in a rectangular box in the center of a high resolution computer screen. The rectangular box appeared continuously on the monitor. Each word was presented for 300 ms , followed by a blank of 200 ms . The final word of the sentence was presented with a period sign and was followed by a 5.5 second rest period. $50 \%$ of the test sentences were followed by comprehension questions. The questions were presented in their entirety above the rectangular frame for 2500 ms , followed by a rest period of 3500 ms . Key presses with the right and left index fingers (counterbalanced across subjects) were used to for yes and no responses to the questions. Within the session, the stimuli were broken into 6 runs consisting of 27 sentences and lasting approximately 8 minutes each. The EEG session, including electrode application and removal, lasted approximately 1.5 hours. After electrode removal, participants performed the four working memory assessments (also in a sound attenuated booth). The order of the four working memory tasks was counterbalanced across participants. Completion of the working memory tasks took no more than one hour and, thus, the entire session lasted approximately 2.5 hours.

## Data Analysis

Upon completion of pre-processing, ERPs were computed for each individual in each experimental condition for a 1500 ms interval time-locked to the presentation of the critical verb ("showed" above) relative to a 200 ms prestimulus baseline. The following time windows were considered in the analysis: 500-700 and 700-900. The analyses were performed on midline, dorsal and ventral electrodes. The midline electrodes were divided into anterior (FPZ, FZ, FCZ, CZ) and posterior (CPZ, PZ, POZ, $\mathrm{OZ})$ sections. The dorsal electrodes were grouped by anterior-posterior (AP) location and hemisphere: Left anterior (FP1, AF3, F3, 20, FC3, C3), right anterior (FP2, AF4, F4, 118, FC4, C3), left posterior (CP3, 53, P3, P1, 59, PO7) and right posterior (CP4, 86, P4, P2, 91, PO8). The ventral electrodes were similarly grouped: Left anterior (F7, FT7, FC5, T10, 40), right anterior (F8, FT8, FC6, T11, 109), left posterior (T3, TP7, CP5, 50, P5, T5, P9) and right posterior (T4, TP8, CP6, 101, P6, T6, P10).

The effects of the garden-path/object relatives compared to controls on brain activity were assessed in the dorsal and ventral regions with three way ANOVAs (sentence type $x$ AP $x$ hemisphere) and in the midline electrodes with a twoway ANOVA (sentence type x AP). In addition, the mean amplitude in posterior midline electrodes in the $700-900 \mathrm{~ms}$ time window was used in the correlational analyses of the working memory assessment and behavioral data.

Scores and accuracy data from the four working memory assessments were used in the correlational analysis. First, the correlations between the working memory measures were assessed. Second, the correlations between the WM measures and garden-path/object relative effects in the comprehension accuracy and P600 data were assessed. The sentence type effects were calculated for the comprehension
data by subtracting accuracy for control sentences from that of garden-path/object relatives. Likewise, for the ERP data, the mean amplitude at over posterior midline electrodes during the $700-900 \mathrm{~ms}$ time window for the control sentences was subtracted from that of garden-path/object relative sentences.

## Results

## Behavioral Data

Accuracy was lower for garden-path (79.2\%, S.D. 17.1) and object relative ( $80.6 \%$, S.D. 19.0) sentences than for controls $(84.3 \%, 15.3)$. There was an effect of sentence type for the garden-path/control comparison $(F(1,50)=4.03, p=$ .050 ) but not for the analysis of object relatives ( $p>.2$ ).

## ERP Data

In the $700-900 \mathrm{~ms}$ time window, garden-path sentences elicited increased positivity compared to controls. At midline regions, there was a significant interaction of type and $\mathrm{AP}(F(1,50)=6.08, p<.05)$ such that garden-paths are more positive than controls at posterior sites. Simple comparisons showed a marginal effect in posterior sites $(\mathrm{F}(1,50)=3.61, \mathrm{p}=.06)$. Over dorsal regions, there was a main effect of sentence type $(\mathrm{F}(1,50)=5.39, \mathrm{p}<.05)$ and an interaction of sentence type and AP $(F(1,50)=5.24, p<.05$. Simple comparisons showed a significant effect of type at posterior electrodes $(F(1,50)=10.4, p<.005)$ such that garden-paths elicited greater positivity. There was also an interaction of sentence type and AP over ventral sites $(F(1,50)=4.28, \mathrm{p}<.05)$. Simple comparisons revealed no significant effects. For the object relatives, there was a main effect of sentence type $(\mathrm{F}(1,50)=4.12, \mathrm{p}<.05)$ such that object relatives were more positive.

## Correlations

The correlational analysis of the working memory assessments showed significant correlations between operation span and reading span ( $r=.353, p<.05$ ), operation span and anti-saccade accuracy ( $r=.334 . p<.05$ ), and anti-saccade and $n$-back target accuracy $(r=.349, p<$ .05).

Analysis of the accuracy effects showed a correlation between garden-path effects and reading span ( $r=-.294, p<$ .05). There was also a significant correlation between object relative effects and accuracy for $n+1$ lures in the $n$-back task ( $r=.317, p<.05$ ). A similar pattern was seen in the P600 data. The P600 effect for garden-path sentences correlated positively with operation span score ( $\mathrm{r}=.381, \mathrm{p}<.01$ ) and marginally with reading span score ( $r=.266, p=.06$ ). The P600 effect for object relatives correlated negatively with $n$ back accuracy for $n-1$ lures ( $r=-.352, p<.05$ ) and $n+1$ lures ( $r=-.332, p<.05$ ).

## Discussion

The effects of sentence type (both P600 and accuracy) are consistent with previous findings. The effects for gardenpath sentences were significant and typical in terms of distribution and time course. With respect to object relatives, Gouvea et al. (2010) got no significant effects for object relatives versus controls while the current study found a main effect of sentence type (with no interactions with topographical factors). This difference is likely due to power as the current study had 50 participants while Gouvea et al. (2010) had twenty. Accuracy for garden-path sentences was significantly lower than controls, as in previous studies (Frazier \& Rayner, 1982; Ferreira, Bailey \& Ferraro 2002) but there was no effect for object relatives. Gouvea et al. (2010) with almost the same materials got no effects of sentence type whatsoever in the accuracy data. Though the garden-path effects differ from Gouvea et al. (2010) in this respect, they are consistent with previous studies. The cross-correlations between the working memory measures (specifically the lack of correlation between the complex span tasks and $n$-back) were also consistent with established findings (Kane, Conway, Miura \& Colflesh, 2007; Unsworth, Schrock \& Engle, 2004). The two complex span tasks (reading span and operation span) did correlate significantly with each other which was also expected based on previous findings (Conway et al. 2005). The correlational analyses between the WM and sentence processing effects were, therefore, run on data showing standard effects and not anomalous in any way.

The key finding of the correlational analysis is that, while performance on the complex span tasks (reading span and operation span) does predict garden-path effects, it is $n$-back accuracy that predicts the effect for object relatives. This is seen in both P600 data and accuracy. Though this is a novel finding, both assessments are intuitively related to their respective sentence types.

In garden-path sentences, individuals must maintain the linear sequence of words active in memory while reanalyzing the syntactic structure. The resolution of a garden-path does, therefore, require dividing attention between storage and sentence processing as does the reading span task. The correlation with operation span was positive for the P600 effect, indicating greater effects for individuals with higher operation span. The marginal correlation with reading span was also positive. These findings are consistent with the Early Commitment Model (Friederici et al., 1998) which argues that high span individuals, when faced with a syntactic ambiguity, commit early to one structure and proceed accordingly rather than entertaining multiple possibilities. The consequence, in the case of a garden-path, is that they then must execute the costly reanalysis process (Friederici et al., 1998). Reading span score correlated with the garden-path effect on comprehension accuracy, similarly predicting greater effects for high span participants. Taken together, these results affirm complex span performance as a predictor online and offline garden-path effects.

In the object relative sentences, the relative pronoun must be maintained active and eventually matched to the appropriate thematic position. This involves assessing each new word in the sentence sequence to determine if it is the relevant predicate. In this way, the process resembles $n$-back in which each new letter must be checked against the letter that is $n$ places back in the sequence. The correlation was negative for the P600 effect suggesting that increased $n$ back accuracy is associated with either (1) reduced sensitivity to structural relationships or (2) increased efficiency in that processing such that complex syntactic structure is less disruptive. The positive correlation between the comprehension accuracy effects accuracy for $n+1$ (indicating reduced effects of syntactic complexity among individuals with high $n+1$ lure accuracy) does not enable a distinction between the two accounts, as both would predict reduced differences between object relative and control sentences. While this is a question for future research, the finding remains that working memory mechanisms reflected in $n$-back are recruited during the resolution of filler-gap dependencies.

The lack of a relationship between n-back accuracy and garden-path effects is somewhat surprising given the findings of Novick et al. (in press). This could be due to methodological differences (for example, Novick et al. used self-paced reading). It could also indicate, in addition to increasing n-back accuracy as a result of training, participants gained some strategic skills that facilitated task performance.

In contrast to the n-back and complex span tasks, antisaccade also showed no relationship with the online and offline sentence type effects. Mendelsohn (2002) also failed to find a relationship between anti-saccade and garden-path effects but he did succeed in with a verbal sorting task that also measured the ability to inhibit automatic responses. It is possible that anti-saccade is a poor predictor of language performance.

The finding that separate working memory measures correlate with P600 effects for the two sentence types leads to the speculation that the late positive components elicited by these two processes may be categorically distinct. The notion of distinct late positive components is not new (see Kutas, van Petten, \& Kluender 2006 for review) but findings have been mixed. Friederici, Hahne and Saddy (2002) found differences in time course and topographical distribution in the P600s effects of grammaticality and syntactic complexity. Kaan and Swaab (2003), however, found no difference in P600 effects of grammaticality and dispreferred structure. The current results, that late positivities elicited by the resolution of temporary syntactic ambiguity and syntactic complexity are underpinned by distinct working memory mechanisms provides a fresh perspective on this question. In addition to the correlational effects, the two effects did show topographical differences such that the garden-path sentences elicited a late positivity with a posterior distribution while that of the object relatives was not limited to posterior sites. While this difference is
slight, due to the relatively coarse topographical analyses it is not possible "to rule out the possibility that the P600 is the consequence of a disparate set of processes that happen to elicit topographically similar responses" (Gouvea et al. 2010, p. 177). While the proposal is speculative at this time, consideration of the relationship of working memory mechanisms to these components in future research will likely provide valuable insights.

In conclusion, the results of the current study suggest a more complex relationship between working memory mechanisms and online sentence processing than has previously been considered, such that different working memory mechanisms support the resolution of different types of difficult structures. Furthermore, the findings provide evidence for functionally distinct late positive ERP components. Future research on the interaction of working memory and language must include a variety of working memory assessments in order to increase understanding of the cognitive underpinnings of sentence processing.

## References

Bialystok, E. (2006). Effect of bilingualism and computer video game experience on the Simon task. Canadian Journal of Experimental Psychology, 60, 68-79.
Bialystok, E., Craik, F. I. M. \& Luk, G. (2008). Cognitive control and lexical access in younger and older bilinguals. Journal of Experimental Psychology: Learning, Memory and Cognition, 34, 859-873.
Bornkessel, I.D., Fiebach, C.J. \& Friederici, A.D. (2004). On the cost of syntactic ambiguity in human language comprehension: an individual differences approach. Cognitive Brain Research, 21, 11-21.
Conway, A.R., Kane, M.J., Bunting, M.F., Hambrick, D.Z., Wilhelm, O. \& Engle, R.W. (2005). Working memory span tasks: A methodological review and user's guide. Psychonomic Bulletin \& Review, 12(5), 769-786.
Daneman, M., \& Carpenter, P.A. (1980). Individual differences in working memory and reading. Journal of Verbal Learning and Verbal Behavior, 19, 450-466.
Dien, J. (2010). The ERP PCA Toolkit: an open source program for advanced statistical analysis of event-related potential data. Journal of Neuroscience Methods, 187(1), 138-145.
Ferreira, F., Bailey, K.G.D., Ferraro, V., 2002. Goodenough representations in language comprehension. Current Directions in Psychological Science, 11, 11-15.
Frazier, L. (1987). Syntactic processing: Evidence from Dutch. Natural Language and Linguistic Theory, 5, 519559.

Frazier, L. \& Rayner, K. (1982). Making and correcting errors during sentence comprehension: eye movements in the analysis of structurally ambiguous sentences. Cognitive Psychology, 14, 178-210.
Friederici, A. D., Hahne, A., \& Saddy, D. (2002). Distinct neurophysiological patterns reflecting aspects of syntactic complexity and syntactic repair. Journal of Psycholinguistic Research, 31, 45-63.

Friederici, A. D., Steinhauer, K., Mecklinger, A., \& Meyer, M. (1998). Working memory constraints on syntactic ambiguity resolution as revealed by electrical brain responses. Biological Psychology, 47, 193-221.
Gouvea, A.C., Phillips, C., Kazanina, N. \& Poeppel, D. (2010). The linguistic processes underlying the P600. Language and Cognitive Processes, 25(2), 149-188.
Just, M.A., \& Carpenter, P.A. (1992). A capacity theory of comprehension: Individual differences in working memory. Psychological Review, 99, 122-149.
Kaan, E., Harris, T., Gibson, E., Holcomb, P.J., 2000. The P600 as an index of syntactic integration difficulty. Language and Cognitive Processes. 15, 159-201.
Kaan, E., Swaab, T.Y.Y., 2003. Repair, revision, and complexity in syntactic analysis: an electrophysiological differentiation. Journal Cognitive Neuroscience, 15, 98110.

Kane, M.J., Conway, A.R., Miura, T.K. \& Colflesh, G.J. (2007). Working memory, attention control, and the Nback task: A question of construct validity. Journal of Experimental Psychology: Learning, Memory and Cognition, 33(3), 615-622.
King, J.W., \& Just, M.A. (1991). Individual differences in syntactic processing: The role of working memory. Journal of Memory and Language, 30, 580-602.
Kutas, M., Van Petten, C., \& Kluender, R. (2006). Psycholinguistics electrified II: 1994-2005. In M. Traxler \& M.A. Gernsbacher (Eds.), Handbook of Psycholinguistics, $2^{\text {nd }}$ Edition. New York: Elsevier.
MacDonald, M. C., \& Christiansen, M. (2002). Reassessing working memory: Comment on Just and Carpenter (1992) and Waters and Caplan (1996). Psychological Review, 109, 35-54.
Mendelsohn, A. (2002). Individual differences in ambiguity resolution:Working memory and inhibition. Unpublished doctoral dissertation, Northeastern University.
Novick, J. M., Hussey, E., Teubner-Rhodes, S., Dougherty, M. R., Harbison, J. I., \& Bunting, M. F. (in press). Clearing the garden-path: Improving sentence processing through executive control training. Language and Cognitive Processes.
Novick, J.M., Trueswell, J.C. \& Thompson-Schill, S.L. (2005). Cognitive control and parsing: Reexamining the role of Broca's area in sentence comprehension. Cognitive, Affective \& Behavioral Neuroscience, 5(3), 263-281.
Osterhout, L., Holcomb, P.J., Swinney, D.A., 1994. Brain potentials elicited by garden-path sentences: evidence of the application of verb information during parsing. Journal of Experimental Psychology: Learning Memory and Cognition, 20, 786-803.
Sprouse, J., Wagers, M. \& Philips, C. (2012). A test of the relation between working-memory capacity and syntactic island effects. Language, 88(1), 82-123.
Turner, M.L. \& Engle, R.W. (1989). Is working memory capacity task independent? Journal of Memory and Language, 28, 127-154.

Unsworth, N., Schrock, J.C. \& Engle, R.W. (2004). Working memory capacity and the antisaccade task: Individual differences in voluntary saccade control. Journal of Experimental Psychology: Learning Memory and Cognition, 30, 1302-1321.
Unsworth, N., Heitz, R. P., Schrock, J. C., \& Engle, R. W. (2005). An automated version of the operation span task. Behavior Research Methods, 37, 498-505.
Vos, S.H., Gunter, T., Schriefers, H. y Friederici, A.D. (2001). Syntactic parsing and working memory: The effects of syntactic complexity, reading span and concurrent load. Language and Cognitive Processes, 16, 65-103.

## Acknowledgements

This research is based upon work supported, in whole or in part, with funding from the United States Government. Any opinions, findings and conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the University of Maryland, College Park and/or any agency or entity of the United States Government. Nothing in this article is intended to be and shall not be treated or construed as an endorsement or recommendation by the University of Maryland, United States Government, or the author of the product, process, or service that is the subject of this article. Correspondence concerning this article should be addressed to Polly O'Rourke, Center for Advanced Study of Language, University of Maryland, $700552^{\text {nd }}$ Avenue, College Park, MD 20742 (e-mail: porourke@casl.umd.edu). Special thanks to Colin Phillips, Mike Bunting and Joseph Dien.

# Can Help Seeking Behavior in Intelligent Tutoring Systems Be Used as Online Measure for Goal Orientation? 

Christine Otieno (otieno@psychologie.uni-freiburg.de)<br>Rolf Schwonke (schwonke@psychologie.uni-freiburg.de)<br>Department of Educational and Developmental Psychology, University of Freiburg<br>Engelbergerstr. 41, D-79085 Freiburg, Germany<br>Ron Salden (rsalden@m-iti.org)<br>Madeira Interactive Technologies Institute, University of Madeira<br>Caminho da Penteada, 9020-105 Funchal, Portugal<br>Alexander Renkl (renkl@psychologie.uni-freiburg.de)<br>Department of Educational and Developmental Psychology, University of Freiburg<br>Engelbergerstr. 41, D-79085 Freiburg, Germany


#### Abstract

Questionnaires to assess goal orientation are widely used. However, recent research indicates some shortcomings. Most significantly, questionnaire data are unable to capture developments and changes in students' goal orientation during the learning process. Therefore, it seems appropriate to supplement questionnaire data with online measures that directly tackle students' behavior. We analyzed data of 57 students who participated in a study with the Cognitive Tutor Geometry. Specifically, we analyzed relationships between questionnaire data on goal orientation, the use of hints and a glossary while working with the Tutor as potential online indicators for goal orientation, and learning outcomes. Results of our analyses show that our potential online indicators systematically differ from questionnaire data of goal orientation, yet have high predictive power for learning outcomes. Therefore, online indicators may be used to supplement questionnaire data of goal orientation and/or to further optimize adaptation in intelligent tutoring systems.


Keywords: Motivation, Goal Orientation, Self-regulated Learning, Intelligent Tutoring Systems

## Introduction

Motivation and self-regulated learning are inseparably intertwined. One specifically important and wellinvestigated area of motivation is that of achievement goal theory (Pintrich, 2000). Initially, the theory's basic distinction was between mastery goal and performance goal orientation (e.g., Dweck, 1986). Mastery goal orientation refers to the goal of reaching understanding and mastery in a field. Performance goal orientation refers to the goal to perform better in comparison to others (Elliot \& McGregor, 2001). Mastery goal oriented students have often been found to show more effort and persistence during learning and, as a result, better learning outcomes compared to performance goal oriented students (Urdan, 1997). Elliot and McGregor (2001) introduced "valence" as an additional dimension to describe goal orientation; that is, approaching success versus avoiding failure. This additional distinction leads to four aspects of goal orientations: mastery-
approach, mastery-avoidance, performance-approach, and performance-avoidance goal orientation.

Initially, goal orientation was regarded as a relatively stable personality trait (e.g., Dweck \& Leggett, 1988). Later studies, however, put an emphasis on the influence of situational variables and task characteristics on goal orientation (e.g., Butler, 1993). Some researchers pointed out as early as in the 90s, that students may not clearly belong to one or the other group of learners in classroom situations (i.e., performance versus mastery goal oriented or approach versus avoidance oriented). In contrast, it is highly plausible that students show both mastery and performance goal orientations at the same time, albeit at different levels. Also, there may be variations in the students' predominant goal orientation during learning phases depending on the task at hand and level of expertise (e.g., Meece \& Holt, 1993). In analogy to the state-trait concept of anxiety first introduced by Cattell and Scheier (1961), recent research points to the reciprocal influences of state and trait measures in the field of goal orientation (Chen, Gully, Whiteman, \& Kilcullen, 2000).

Typically, goal orientation is measured via self-report questionnaires. This approach is rooted in the traditional view of goal orientation as a personality trait and can be considered to measure habitual goal orientation. Despite their long tradition and proven utility in the field, interpretation of questionnaire data of achievement goal orientation can be problematic. More specifically, ambiguity between different questionnaires with respect to their conceptual overlap often makes it difficult to compare findings from different studies (Hulleman, Schrager, Bodmann, \& Harackiewicz, 2010). Also, data measured before or after a learning phase using self-report questionnaires lack the ability to capture decisions and states of the learners as they arise from circumstances in the learning environment and develop during the learning process (Richardson, 2004). Consequently, recent research calls for measurement of achievement goals not only by questionnaires but also by online measures to grasp
moment-to-moment actions and thereby the state aspect of goal orientation at a fine-grained level. One way to track goal orientation online is through traces in online learning environments (e.g., Zhou \& Winne, 2012).

In an attempt to investigate potential relationships between online and offline measures for goal orientation and their predictive power for learning outcomes, Zhou and Winne (2012) enriched an instructional text presented online with a set of hyperlinks and tags referring to the four different goal orientations (mastery-approach, masteryavoidance, performance-approach, performance-avoidance). Hyperlinks to be selected were presented next to the text (e.g., "take a practice test on this"; performance-approach). Within the text students could use highlighting to structure the text and label highlights (e.g., "I want to learn more about this"; mastery-approach; Zhou \& Winne, 2012, p.415). Selection of hyperlinks and tags were interpreted as online traces for the respective goal orientation. Zhou and Winne (2012) found that goal orientation as assessed by questionnaire data do not correlate with goal orientation as assessed by the traces captured online during the learning process. These findings are in line with earlier research indicating that self-reported measures of study tactics do hardly correspond to respective online measures collected during learning (Jamieson-Noel \& Winne, 2003). The differences between online measures and questionnaire data for goal orientation could be partly seen as an indication of state-trait differences in goal orientation. Additionally, Zhou and Winne found an advantage of online traces over questionnaire data to predict learning outcomes, which raises the following question: Are online measures "better" than questionnaire data to assess goal orientation for educational purposes?
Another potential advantage of measuring goal orientation online is that it is less obtrusive compared to explicitly asking questions upon which students have to reflect. For example, intelligent tutoring systems could use tracking data that are collected during the learning process to estimate students' goal orientation at any given point in the learning phase and adapt their responses to the students' motivation and attitudes (e.g., Arroyo, Cooper, Burleson, \& Woolf, 2010). Such adaptation could make intelligent tutoring systems even more effective (for an overview of recent advances in intelligent tutoring systems, see Graesser, Conley, \& Olney, 2012).

Cognitive Tutors and other intelligent tutoring systems have proven to be very effective in supporting individual students' learning in a variety of domains such as mathematics or genetics (for an overview, see Koedinger \& Corbett, 2006) and are widely used in schools across the United States as part of the regular mathematics curriculum. Based on an online assessment of students' learning, Cognitive Tutors provide individualized support for guided learning by doing. Specifically, the Tutor selects appropriate problems, gives just-in-time feedback, and provides hints. Additionally, students can use a glossary to look up definitions and explanations. Hints provide direct
instructions for the next step a student has to determine; they are context sensitive and therefore adaptive to the situation. The glossary offers definitions and explanations for principles to be understood and learned; it is context insensitive and therefore not adaptive to the specific situation (Koedinger \& Aleven, 2007). In light of achievement goal theory one could interpret the use of hints as performance-goal oriented and the use of a glossary as mastery-goal oriented behavior. Although hints can be used in a mastery-goal oriented way, specifically if students reflect upon them, they are often not used in this way. Their adaptive nature to the problem at hand suggests their use in order to immediately solve a problem rather than to deeply reflect and understand the underlying principle. Sometimes students even abuse hints in order to proceed quickly through the learning environment, a behavior referred to as "gaming-the-system" (Baker, Corbett, Koedinger, \& Wagner, 2004). The glossary, in contrast, is not directly related to a to-be-determined problem step at hand. Therefore, we assume that it is consulted whenever students are interested in information that goes beyond the immediate problem-solving step. We claim that this behavior may be related to mastery-goal orientation as, in contrast to hint use, it does not primarily improve immediate performance in the learning environment but understanding. Using online tracking data of hint and glossary use could therefore be an unobtrusive and more proximal, "state-like" indication of goal orientation compared to the more reflected and "trait-like" measures gained by questionnaires. In addition, the data are tracked automatically, not taking up additional resources on either the side of the program or the learner.

## The Present Study

Attempting to test if the findings of Zhou and Winne (2012) can be conceptually replicated in a different learning environment, and if hint and glossary use could be valid behavioral indicators for goal orientation, we reanalyzed a data set from an earlier study where students learned geometry principles using the Cognitive Tutor Geometry ${ }^{\circledR}$ (Salden, Aleven, Renkl, \& Schwonke, 2009). First, we tested if self-reported goal orientations as assessed by a questionnaire correspond to the respective online measures.

Second, we assumed, as in the study by Zhou and Winne (2012), a positive relationship of glossary use and learning outcomes (i.e., understanding) and a negative relationship of hint use and learning outcomes. In our study, the learning outcome tests (i.e., posttests) - presented immediately after the learning phase and one week later - measured not so much knowledge of routines but application and understanding of the principles learned in the Cognitive Tutor. Our expectations were also in line with earlier studies showing better learning outcomes for masteryoriented students than for performance-oriented students (for an overview, see Urdan, 1997). In addition, other studies on the Cognitive Tutor found negative relations between hint use and learning outcomes (e.g., Aleven \&

Koedinger, 2001).
Third, while Zhou and Winne (2012) did not find a significant relationship of questionnaire data with performance on posttest, theoretical considerations as well as earlier studies led to the expectation that such a relationship may exist (for an overview, see Urdan, 1997). We therefore addressed the ("two-sided") research question (as did Zhou and Winne) if online and questionnaire data alike relate to learning outcomes.
Fourth, we checked if behavioral indicators for mastery as well as performance goal orientation (i.e., glossary and hint use, respectively) are stronger predictors of learning outcomes than respective questionnaire data.
To even out potential influences of prior knowledge on posttest performance we controlled for math grade (the strongest predictor of learning outcomes in this study) in all calculations involving posttest performance. More specifically, we addressed the following research questions:
(RQ1) Do self-reported goal orientations from the questionnaire correlate with respective behavioral indicators (i.e., hint use with performance goal orientation and glossary use with mastery goal orientation)?
(RQ2) Is there a positive relationship between glossary use and learning outcomes and a negative relationship between hint use and learning outcomes?
(RQ3) What is the relationship between questionnaire data of goal orientation and learning outcomes?
(RQ4) Are behavioral indicators for goal orientation better predictors of learning outcomes than the respective questionnaire data (i.e., are glossary and hint use better predictors for learning outcomes than selfreport measures)?

## Method

## Sample and Design

Participants in our study were 57 students (19 in $9^{\text {th }}$ grade and 38 in $10^{\text {th }}$ grade; age: $M=15.63, S D=.84$ ) from a German "Realschule", which is equivalent to an American high school. The original study comprised three conditions to which participants were randomly assigned resulting in an equal distribution of 19 students per condition. In two conditions students were provided with worked examples to solve the mathematical problems. Worked examples were either faded out according to a fixed procedure (fixed fading condition) or according to the student's individual skill level (adaptive fading condition). The third condition served as a control and did not receive any worked examples (problem condition; Salden et al., 2009). For the purpose of the reanalysis of our data for this paper, that is to investigate potential relationships between online and questionnaire measures of goal orientation and learning outcome, we examined all 57 participants as one group. To preclude potential influences of conditions on the observed
relationships, however, we routinely calculated all analyses for the separate conditions and checked for potential significant differences. However, no such differences were found.

Learning Environment - The Cognitive Tutor


Figure 1. Screenshot of the Cognitive Tutor Geometry®
Cognitive Tutors provide adaptive feedback and model students' skill acquisition based on two algorithms: model tracing and knowledge tracing (Koedinger \& Corbett, 2006). Simulating the problem solving process enables the Tutor, for example, to provide specific hints for a problem situation. Also, all steps (i.e., all actions a student takes while working with the program) are tracked in a logfile. This data are used online for adaptation. For the purpose of this paper we analyzed part of this logfile data, specifically the amount of hint and glossary use (percentage in relation to all activities of the student in the learning environment), and correlated them with offline data of a goal orientation questionnaire and posttest scores.

Learning Materials During the learning phase with the Cognitive Tutor we asked students to work on fifteen problems in a Cognitive Tutor lesson on geometry, covering four geometry principles. The first eight problems required the application of only one geometry principle. The last seven problems combined different principles and were therefore more complex. Before the learning phase we provided students with instructions about the different tools in the Tutor. More specifically, after giving an overview of the learning environment, hints were introduced as an assistance tool to use when "having trouble solving a task or when reaching an impasse. The glossary was introduced as an assistance tool to use if "you are unsure when to use a certain mathematical principle or which is the corresponding formula". These instructions were routinely used in several of our studies involving the Cognitive Tutor Geometry (e.g., Salden et al., 2009; Schwonke et al., 2012).

## Instruments

Pretest The pretest was integrated in the Cognitive Tutor and consisted of four geometry problems related to the lessons taught later during the learning phase with the program. All Cognitive Tutor help facilities (e.g., hints) were disabled during the pretest. On average students needed 21 minutes to complete the pretest. Mathematics grade was a significantly stronger predictor of posttest performance than the pretest. Therefore, we included mathematics grade and not pretest scores in all analyses referring to posttest performance.
Goal Orientation Questionnaire Before solving the posttest, students were asked to answer 8 items concerning their learning goal orientation while working with the program on a scale from 1 to 6 . Items were adapted from Elliot and McGregor (2001) and reflected mastery-approach and performance-approach goal orientations only.

Posttest A posttest consisting of the same problems as the pretest was implemented in the learning environment. Additionally, all participants were asked to complete a paper-pencil test immediately after working with the Tutor and one week later (delayed posttest). Immediate and delayed posttests were identical. On average students needed 31 minutes to complete the posttest and 21 minutes to complete the delayed posttest.

## Procedure

The first experimental session lasted 90 minutes on average and was divided into three parts: pretest and introduction, learning phase in the Cognitive Tutor, and questionnaire on goal orientation as well as posttest. First, students' general prior knowledge was assessed by their mathematics grade together with additional demographic data such as age and gender. Then they received a brief introduction on how to use the Cognitive Tutor followed by a short pretest implemented in the Tutor measuring their prior knowledge. After completing this pretest, students read an instructional text providing information about the rules and principles that were later addressed in the Cognitive Tutor. In the tutoring part, students worked with their respective version of the Cognitive Tutor. This learning phase was followed by a questionnaire measuring goal orientation with self-report measures and a knowledge test. The students worked again on the knowledge test in a second session (one week later).

## Results and Discussion

To test if questionnaire data for goal orientation align with respective online measures (RQ1) we determined Pearson's correlations between assumed behavioral indicators for goal orientation (i.e., glossary use for mastery goal orientation and hint use for performance goal orientation) and self-report questionnaire data. There was no significant relationship between glossary use and mastery goal orientation ( $r=.13, \mathrm{p}=.339$ ) or hint use and performance goal orientation ( $r=-.14, \mathrm{p}=.298$ ). These
findings are in line with Zhou and Winne (2012). The missing relationship between behavioral data collected online and questionnaire data collected after the learning phase may indicate that the two measures capture different constructs. One theoretically plausible interpretation is, that both the online measures collected by Zhou and Winne and our behavioral data, that is, hint and glossary use may reflect state goal orientation while questionnaire data may capture the trait aspect of goal orientation. However, one could argue that state and trait measures of other psychological constructs are generally correlated which raises the question of construct validity of the online measures. Therefore, more data is needed to decide if online measures and specifically behavioral data as the ones used in our study can be validly used as indicators for (state) goal orientation, if they differ systematically from the assumed trait measures of questionnaire data, and how both state and trait mutually affect each other. However, our data provides some initial evidence for the validity of hint and glossary use as measures for goal orientation:

First, we determined the correlation between glossary and hint use and found a very strong negative correlation: $r=-$ $.84, \mathrm{p}<.001$. This indicates that students had a relatively clear preference for either hints or glossary which is in line with the assumption that the type of tool use indicates whether the students were primarily concerned about solving the problems (i.e., performance orientation) or understanding the principles (i.e., mastery orientation) while working on the Cognitive Tutor lessons.

Second, we tested if glossary use is positively related and hint use is negatively related to learning outcomes (RQ2) which should be the case if these online measures can be associated with goal orientation. We determined partial correlations between glossary and hint use and the immediate and delayed posttest performance, controlling for prior knowledge. Results indicate a significant positive relationship for glossary use and immediate ( $r=.37, \mathrm{p}=$ .008) posttest score. Correlation of glossary use and the delayed posttest score slightly failed to reach statistical significance ( $r=.28, \mathrm{p}=.050$ ). There was a significant negative relationship between hint use and performance on the immediate ( $r=-.48, \mathrm{p}<.001$ ) as well as the delayed ( $r=$ $-.36, \mathrm{p}=.009$ ) posttest score. These relations can be seen as evidence that glossary and hint use may indeed be valid indicators for goal orientation. This may specifically be true as our posttests measured deep understanding of the principles learned in the Cognitive Tutor and not so much knowledge of routines. In a test measuring the later, differences between primarily performance versus mastery goal oriented students may not be as pronounced. Additionally, interpreting hint use as a measure for performance goal orientation may provide one explanation for the repeatedly found negative relations between hint use in the Cognitive Tutor and performance on posttest.

We further tested the relationship between self-reported mastery and performance goal orientation (i.e., questionnaire data) and learning outcomes (RQ3). We
found a significant positive relation between mastery goal orientation and delayed posttest scores ( $r=.41, \mathrm{p}=.003$ ). The relationship between mastery goal orientation and immediate posttest scores ( $r=.25, \mathrm{p}=.076$ ) slightly failed to reach statistical significance. There was also no significant relationship between self-reported performance goal orientation and immediate posttest scores ( $r=-.21, \mathrm{p}=$ .144); the relationship between performance goal orientation and delayed posttest scores ( $r=-.24, \mathrm{p}=.086$ ) failed to reach statistical significance. These results are, at least partly, in contrast to Zhou and Winne (2012) who observed no statistically significant correlations between self-reported goal orientations and posttest performance. However, the results are in line with theoretical assumptions and earlier studies using questionnaire data on goal orientation and further corroborate the aforementioned relation between goal orientation and learning outcomes.
To test if online measures or their respective questionnaire data are better predictors for learning outcomes (RQ4) we calculated separate stepwise linear regression analyses, one for mastery goal orientation (glossary use and respective questionnaire data) and one for performance goal orientation (hint use and respective questionnaire data) as potential predictors for immediate and delayed posttest performance. Concerning the predictive power of mastery goal orientation (glossary use vs. questionnaire data) for posttest scores results are mixed: While glossary use was the sole best predictor for immediate posttest scores, questionnaire data was the best predictor for delayed posttest scores (Table 1). With regard to the predictive power of performance goal orientation (hint use vs. questionnaire data) for posttest scores, there was a clear advantage of the behavioral data: Hint use was the sole best predictor for both immediate and delayed posttest scores (Table 2). Taken together, our results indicate that specifically for mastery goal orientation questionnaire data might yield predictive power beyond behavioral online data, at least for long-term learning effects. These results are not fully in line with Zhou and Winne (2012) who consistently found online measures to be the stronger predictors of learning outcomes in regression models. There might be methodological explanations for the differences between the two studies: We used a different questionnaire as basis for our goal orientation items and measured only two and not four aspects of goal orientation. Also, the questionnaire used by Zhou and Winne did not relate significantly to learning outcome measures. In addition, utilizing hint and glossary use as indicators for goal orientation might be a little more "indirect" as compared to the online data collected by Zhou and Winne (2012). For example, hint use might also be elicited by errors made when trying to determine solution steps, that is, it may be related to rather poor performance in the learning environment. However, the very strong negative correlation of $r=-.84$ between hint and glossary use cannot be explained by these errors (partial correlation controlling for errors is still highly significant with $r=-.73$,
$p<.001$ ).
Table 1: Glossary Use and Mastery Goal Orientation as Predictors for Learning Outcomes

|  |  | $B$ | $S E$ | $\beta$ |
| :--- | :--- | :--- | :--- | :--- |
| Posttest | Step 1 <br> Glossary Use | .22 | .06 | $.42^{* *}$ |
| Delayed | Step 1 |  |  |  |
| Posttest | Items on <br> Mastery Goal <br> Orientation | .09 | .03 | $.38^{* *}$ |
|  | Step 2 <br> Items on <br> Mastery Goal <br> Orientation <br> Glossary Use | .08 | .03 | $.35^{* *}$ |
|  | (16 | .06 | $.30^{*}$ |  |

Note. Posttest: $R^{2}=.18$ for Step 1; Delayed Posttest:
$R^{2}=.15, \Delta R^{2}=.09$ for Step $2(p<.05)$.

* $\mathrm{p}<.05$ and ${ }^{* *} \mathrm{p}<.01$.

Table 2: Hint Use and Performance Goal Orientation as Predictors for Learning Outcomes
$\left.\begin{array}{lllll}\hline & & B & S E & \beta \\ B\end{array}\right]$

Note. Posttest: $R^{2}=.28$ for Step 1; Delayed Posttest:
$R^{2}=.18$.
${ }^{* *} \mathrm{p}<.01$ and ${ }^{* * *} \mathrm{p}<.001$.
Taken together, both behavioral online and offline questionnaire data provide us with important insights for understanding learners' goal orientation and can be used to supplement rather than to replace each other for the sake of scientific advancement. Given the high predictive value of behavioral online data, however, their use should be considered for educational purposes in classrooms and specifically in online learning environments, where an unobtrusive and efficient collection of goal orientation data could improve adaptation in intelligent tutoring systems and thereby foster the learning process. In addition, one should keep in mind that self-report measures of characteristics such as goal orientation are potentially subject to a social desirability bias which could be circumvented with (indirect) online measures.

Can help seeking behavior in intelligent tutoring systems be used as online measure for goal orientation? Even
though we cannot answer this question based on our data conclusively, our results provide a first and promising indication that online behavior in intelligent tutoring systems provides an unobtrusive and efficient additional or even alternative measure to questionnaire data to assess goal orientation in educational settings.

## Acknowledgments

This work was supported by the Pittsburgh Science of Learning Center which is funded by the National Science Foundation; award number SBE-0354420. We are grateful for Vincent Aleven's assistance in conducting the original study and would also like to thank him for his helpful comments on an earlier version of this paper.

## References

Aleven, V., \& Koedinger, K. R. (2001). Investigations into help seeking and learning with a Cognitive Tutor. In R. Luckin (Ed.), Papers of the AIED-2001 Workshop on Help Provision and Help Seeking in Interactive Learning Environments (pp. 47-58).
Arroyo, I., Cooper, D. G., Burleson, W., \& Woolf, B. P. (2010). Bayesian networks and linear regression models of students' goals, moods, and emotions. Presented at the Third International Conference on Educational Data Mining (EDM2010), Pittsburgh, PA, USA.
Baker, R. S., Corbett, A. T., Koedinger, K. R., \& Wagner, A. Z. (2004). Off-task behavior in the Cognitive Tutor classroom: When students "game the system". In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 383-390). New York, NY, USA: ACM. doi:10.1145/985692.985741
Butler, R. (1993). Effects of task- and ego-achievement goals on information seeking during task engagement. Journal of Personality and Social Psychology, 65(1), 1831. doi:10.1037/0022-3514.65.1.18

Cattell, R. B., \& Scheier, I. H. (1961). The meaning and measurement of neuroticism and anxiety. Oxford England: Ronald.
Chen, G., Gully, S. M., Whiteman, J.-A., \& Kilcullen, R. N. (2000). Examination of relationships among trait-like individual differences, state-like individual differences, and learning performance. Journal of Applied Psychology, 85(6), 835-847. doi:10.1037/0021-9010.85.6.835
Dweck, C. S. (1986). Motivational processes affecting learning. American Psychologist, 41(10), 1040-1048. doi:10.1037/0003-066X.41.10.1040
Dweck, C. S., \& Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. Psychological Review, $95(2)$, 256-273. doi:10.1037/0033295X.95.2.256
Elliot, A. J., \& McGregor, H. A. (2001). A 2 X 2 achievement goal framework. Journal of Personality and Social Psychology, 80(3), 501-519.
Graesser, A. C., Conley, M. W., \& Olney, A. (2012). Intelligent tutoring systems. In K. R. Harris, S. Graham, T. Urdan, A. G. Bus, S. Major, \& H. L. Swanson (Eds.),

APA educational psychology handbook, Vol 3: Application to teaching and learning. (pp. 451-473). Washington, DC, US: American Psychological Association.
Hulleman, C. S., Schrager, S. M., Bodmann, S. M., \& Harackiewicz, J. M. (2010). A meta-analytic review of achievement goal measures: Different labels for the same constructs or different constructs with similar labels? Psychological Bulletin, 136(3), 422-449. doi:10.1037/a0018947
Jamieson-Noel, D., \& Winne, P. H. (2003). Comparing selfreports to traces of studying behavior as representations of students' studying and achievement. Zeitschrift für Pädagogische Psychologiel German Journal of Educational Psychology, 17(3-4), 159-171. doi:10.1024//1010-0652.17.34.159
Koedinger, K. R., \& Aleven, V. (2007). Exploring the assistance dilemma in experiments with Cognitive Tutors. Educational Psychology Review, 19(3), 239-264.
Koedinger, K. R., \& Corbett, A. (2006). Cognitive Tutors: Technology bringing learning sciences to the classroom. In R. K. Sawyer (Ed.), The Cambridge handbook of the learning sciences. (pp. 61-77). New York, NY US: Cambridge University Press.
Meece, J. L., \& Holt, K. (1993). A pattern analysis of students' achievement goals. Journal of Educational Psychology, 85(4), 582-590. doi:10.1037/00220663.85.4.582

Pintrich, P. R. (2000). The role of goal orientation in selfregulated learning. In M. Boekaerts, P. R. Pintrich, \& M. Zeidner (Eds.), Handbook of self-regulation. (pp. 451502). San Diego, CA US: Academic Press.

Richardson, J. T. E. (2004). Methodological issues in questionnaire-based research on student learning in higher education. Educational Psychology Review, 16(4), 347358.

Salden, R., Aleven, V., Renkl, A., \& Schwonke, R. (2009). Worked examples and tutored problem solving: Redundant or synergistic forms of support? Topics in Cognitive Science, l(1), 203-213.
Schwonke, R., Ertelt, A., Otieno, C., Renkl, A., Aleven, V., \& Salden, R. (2012). Metacognitive support promotes an effective use ofs instructional resources in intelligent tutoring. Learning and Instruction. doi:10.1016/j.learninstruc.2012.08.003
Urdan, T. C. (1997). Achievement goal theory: Past results, future directions. In M. L. Maehr \& P. R. Pintrich (Eds.), Advances in motivation and achievement (Vol. 10, pp. 99-141). Greenwich, CT: JAI Press.
Zhou, M., \& Winne, P. H. (2012). Modeling academic achievement by self-reported versus traced goal orientation. Learning and Instruction, 22(6), 413-419. doi:10.1016/j.learninstruc.2012.03.004

# Social categories create (biased) semantic interference during face naming 

Dominic J. Packer (djp208@lehigh.edu)<br>Psychology and Cognitive Science, Lehigh University, 17 Memorial Dr. E. Bethlehem, PA 18015 USA

Pádraig G. O'Séaghdha (pat.oseaghdha@lehigh.edu)
Psychology and Cognitive Science, Lehigh University, 17 Memorial Dr. E. Bethlehem, PA 18015 USA

Almut Hupbach (hupbach@lehigh.edu)<br>Psychology and Cognitive Science, Lehigh University, 17 Memorial Dr. E. Bethlehem, PA 18015 USA

Joseph E. Bates (x5a512@lehigh.edu)
Psychology, Lehigh University, 17 Memorial Dr. E. Bethlehem, PA 18015 USA


#### Abstract

Semantic interference in word retrieval has been observed for both well-learned and ad hoc inter-item relations. We tested whether such semantic interference extends to the blocked cyclic naming of racially homogeneous vs. heterogeneous faces. No information except arbitrarily assigned names was provided for novel faces. Yet we observed interference in naming individuals in homogeneous groups. Moreover, consistent with other findings in the social domain, interference occurred for other-race but not for own-race faces. Because this interference effect does not require a rich knowledge base about individuals, it is consistent with the view that interference arises in adjustments to the strength of conceptual-lexical links rather than in knowledge structures themselves. Evidence of modulation by target race further suggests that interference effects may provide an effective tool for exploration of social categorization processes.


Keywords: Semantic interference; language production; blocked cyclic naming; social categorization; intergroup bias

## Introduction

This paper presents an initial investigation of how basic memory retrieval and language production processes are affected by social context. When people name objects, they often exhibit semantic interference in which retrieving a target word from memory disrupts retrieval of words that belong to the same semantic category (e.g., Damian, Viglocco \& Levelt, 2001). In the current research, we examined whether similar interference effects occur when naming faces of members from social categories, specifically, racial groups. We further investigated whether people exhibit a 'name retrieval bias', such that there is greater interference when naming other-race compared to own-race faces.

## Semantic Interference in Language Production

Semantic interference highlights the competitive nature of word selection in language production. In the blockedcyclic naming paradigm, for example, participants repeatedly name small sets of pictures (e.g., four pictures each named individually four times). The pictures are either presented in homogenous blocks in which they share a common semantic relation, or in heterogeneous blocks in which they do not have identifiable semantic links. Naming times are slower in homogeneous than heterogeneous blocks, and the level of interference often increases over cycles (e.g., Damian et al., 2001; Schnur et al., 2009). Semantic interference occurs in this and similar paradigms because retrieving a word co-activates semantically related words, which compete with and slow selection of the target word. Further, retrieving a word primes its subsequent retrieval, making it a stronger competitor when later naming related words (Howard, Nickels, Coltheart \& Cole-Virtue, 2006). The change in lexical accessibility is long lasting, suggesting that a learning mechanism that damps accessibility of competitors while strengthening the current item, rather than short-term modulation of activation, is at the core of semantic interference (Oppenheim, Dell \& Schwartz, 2010; see Navarrete, DelPrato \& Mahon, 2012). fMRI studies have localized lexical selection to left inferior frontal gyrus [LIFG] (believed to be involved in competition resolution) and linked areas of temporal cortex (Schnur et al., 2009).

Traditionally, semantic interference experiments have investigated how shared membership in fixed taxonomic categories (e.g., animals, vegetables, minerals) generates interference. In these cases, items share both category memberships and overlapping semantic features (e.g., legs, heads, locomotion). However, recent research has also shown interference for items linked by a semantic theme
(e.g., garden links slug, gardener and rake; see Abdel Rahman \& Melinger, 2007), as well as for items that are linked as members of an ad hoc category. Abdel Rahman and Melinger (2011) had participants complete a cyclicblocked naming task with pictures that had no obvious semantic relation to one another (e.g., stool, knife, bucket, river), but that could be combined as members of an ad hoc category (e.g., "things present on a fishing trip"). No interference was found when participants were unaware of the ad hoc category; however, interference (longer naming latencies) arose when participants were informed about the category. These findings demonstrate the highly dynamic nature of the semantic system, such that items without fixed shared semantic features nevertheless rapidly become associated (i.e., exhibit shared activation and competition) when linked by a thematic context.

## Effects of Social Categories

Social categories (e.g., racial group memberships) function in many ways like other categories. People tend, for example, to exaggerate within group similarities and between group differences for both social and non-social categories (e.g., Levin \& Angelone, 2002; Tajfel \& Wilkes, 1963). As such, by virtue of their category membership, individuals are assumed to possess common features, and are often stereotyped accordingly (e.g., Kunda \& Spencer, 2003). Further, the effects of social categories often emerge rapidly and automatically (e.g., Devine, 1989). For example, ERP studies have observed category-based differences in neural signals associated with early visual processing of different race faces (Ito \& Bartholow, 2009). In these studies, white participants show heightened P100 and N170 responses (which have been linked to early face processing) when viewing White versus Black faces (e.g., Ito \& Urland, 2003; Cunningham, Van Bavel, Arbuckle, Packer \& Waggoner, 2012). The robust influence of social categories extends to a wide range of cognitive and affective processes, ultimately shaping behaviors including affiliation, cooperation and conflict (Turner, Hogg, Oakes, Reicher \& Wetherell, 1987).

Prior research has also shown effects of social categories on language use, such that people strategically use language to enhance ingroups and derogate outgroups. In a phenomenon known as the 'linguistic intergroup bias', people tend to describe positive ingroup and negative outgroup behaviors more abstractly than they describe negative ingroup and positive outgroup behaviors (Maass, 1999; Maass, Salvi, Arcuri \& Semin, 1989). The use of relatively abstract words (e.g., adjectives - helpful, aggressive) to communicate "our" desirable and "their" undesirable actions implies that these are enduring and global characteristics. In contrast, using relatively concrete words (e.g., action verbs - help, hit) to communicate our undesirable and their desirable actions conveys that these behaviors are situationally-specific and transient.

In the current research, we investigated whether social categories affect basic language production processes under
controlled experimental conditions. We did this in the domain of face naming. Specifically, we asked whether shared social categories induce semantic interference effects during person (face) naming. We had participants learn the names of 16 novel faces belonging to four different racial groups (four faces in each). They then completed a blocked-cyclic naming task with these faces. In homogenous blocks, participants cycled through naming four faces that all belonged to the same racial group; in heterogeneous blocks, participants cycled through naming four faces that each belonged to a different racial group.

In contrast to common objects, faces are processed through partly specialized cortical networks, including fusiform gyrus. In addition, because person names are arbitrary, their retrieval from face configurations may be more difficult than object naming (e.g., Valentine, Brennen \& Brédart, 1996; see also Griffin, 2010). However, given that social categories exert robust effects on a wide range of psychological processes and function similarly to non-social categories, we expected to observe semantic interference in face naming. Due to their common category membership (and overlapping visual and possibly semantic features e.g., stereotypes), retrieving the name of one group member should increase co-activation of the names of other members, which will compete with and slow selection of the target name.

## Social Categorization Biases

Social categories are not entirely analogous to non-social categories. In particular, people often exhibit intergroup biases, such that members of ingroups and outgroups are processed differently (e.g., the linguistic intergroup bias described above). These biases come in different forms, but many are reducible to the observation that outgroup members tend to be processed more categorically than ingroup members. Whereas ingroup members are typically individuated and treated as distinct entities, outgroup members are often treated as relatively interchangeable exemplars of their group (Brewer, 1988; Fiske \& Neuberg, 1990).

One such bias emerges in facial recognition. The 'other race effect' (or 'own race bias') refers to the well-replicated finding that people are generally better at recognizing members of their own versus other racial groups (e.g., in incidental recognition paradigms). Although perceptual expertise is a contributor (i.e., people typically have more experience processing own than other race faces), recent research suggests that this bias is largely a categorization driven effect (Hugenberg, Young, Bernstein \& Sacco, 2010; Van Bavel, Packer \& Cunningham, 2011). According to Hugenberg et al.'s (2010) Categorization-Individuation Model, for example, classifying faces as exemplars of a category focuses attention on category-diagnostic (shared) features, which reduces subsequent ability to discriminate among category members. In contrast, when faces are not categorized but instead individuated, attention is focused on distinct features, enhancing the ability to discriminate them
later. Critically, intergroup biases in facial recognition emerge because outgroup faces tend to activate their categories more strongly than ingroup faces (Hugenberg et al., 2010; Levin, 1996, 2000; Stroessner, 1996).

We anticipated that a similar bias might also occur in name production. If outgroup faces invoke categorization more strongly than ingroup faces, semantic interference should be stronger for outgroup than ingroup faces. Retrieving an outgroup member's name should co-activate the names of other outgroup members. In contrast, because ingroup members tend to be individuated, retrieving the name of an ingroup member should not increase activation of other ingroup members' names, at least not to the same extent. If so, homogenous ingroup naming may not differ from heterogeneous condition naming.

To summarize, we predicted semantic interference during naming of faces when they are racially grouped. Based on ingroup/outgroup differences in social-cognitive processing, we further hypothesized that such interference would be stronger for other-race than own-race faces.

## Methods

## Participants

Eighteen introductory psychology students at Lehigh University participated for partial course credit. This sample size provides good power because the repeated measures design collects many observations from each participant. The average age was 19.22 years, and there were 9 males and 8 females (one did not report gender). All participants spoke English as a first language, and reported European ethnic origins during pre-testing. A racially homogeneous sample was important in this case for testing hypothesized ingroup/outgroup differences. During testing, one participant indicated a mixed ethnic background (European and Asian). Analyses including vs. excluding this participant yielded identical findings; we therefore report analyses including all participants.

## Design

We used a 2 (Context: heterogeneous, homogeneous) X 4 (Race: Asian, Black, Middle-Eastern, White) X 4 Replication (1, 2, 3, 4) within subjects design.

## Procedure

Learning Phase. After familiarization with the picture naming set-up, participants first learned arbitrarily assigned names for 16 male faces. Four faces belonged to each of four racial groups: Asian, Black, Middle Eastern and White. All names were single syllable, of European origin, and common in the North American context. Each name also started with a different letter (e.g., Bill, Chris, Dan). The names were assigned to faces in two different randomizations, which were counter-balanced across participants. Participants initially viewed the 16 face/name pairings in a randomized $4 \times 4$ matrix on the computer
screen for 2 minutes, and were instructed to try to memorize as many as they could. Each face/name was then presented twice for two seconds in random order, with participants instructed to read the names aloud.
Testing Phase. The testing phase consisted of four replications, each containing eight sets of 16 trials. Each set comprised four faces repeated semi-randomly (for each participant) across four repetition cycles. Four of the sets were racially homogenous, four were racially heterogeneous, and the order of sets within each replication was randomly determined for each participant. In total, participants completed 512 trials.

At the beginning of each set, participants were shown four faces along with their names and were asked to read each name aloud. Then, to confirm that they remembered the names, they were presented with each face individually (in random order) and were asked to provide the name (which appeared on the screen upon vocalization to confirm or correct participants' responses). This was repeated until participants named all four faces correctly. In most cases, no repetitions were required.

Each trial began with a fixation cue $\left(^{*}\right)$ displayed for 100 -milliseconds (ms), along with a warning sound, followed by a face. Naming latencies were measured with a voice key. The face remained on the screen until a name was produced, or for a maximum of 1500 ms . After naming, the face disappeared and was followed by a blank screen for 1500 ms . Participants were instructed to speak clearly, and to name the faces as quickly and accurately as possible.

## Results

Following standard practice for this type of design, the heterogeneous context responses in each replication were sorted to match the corresponding homogeneous groupings. We then conducted a 2 (Context: heterogeneous, homogeneous) X 4 (Race: Asian, Black, Middle-Eastern, White) X 4 Replication (1, 2, 3, 4) analysis on the speed with which participants named faces. Specifically, we implemented a multi-level model in which trials were nested within participants using the PROC MIXED procedure in SAS. Multi-level models allow for more accurate estimates of effects by accounting for interdependence among trials within participants. We removed error trials on which participants named a face incorrectly (1.7\%), the voice key was triggered by something other than a name (e.g., a cough or stutter, $0.9 \%$ ) or participants did not respond within the time window (1.0\%). We also removed trials with RTs < 200 ms .

Extending prior research on semantic interference effects in blocked cyclic naming paradigms, there was a significant main effect of Context, $\mathrm{F}(1,17)=20.05, \mathrm{p}<.001$. Overall, participants were slower to name faces in racially homogeneous ( $\mathrm{M}=701, \mathrm{SD}=179$ ) versus heterogeneous contexts $(M=671, S D=160)$. Critically, however, the effect of Context was modulated by Race [Context x Race interaction: $\mathrm{F}(3,51)=3.56, \mathrm{p}<.05$ ]. As shown in Figure 1 A , naming faces in homogenous (vs. heterogeneous)
contexts produced statistically significant interference effects for Asian ( $\mathrm{p}<.05$ ), Black and Middle-Eastern faces (ps <.01). However, there was no evidence of an effect for White faces ( $\mathrm{p}>.30$ ). The Context X Race interaction was not moderated by Replication ( $\mathrm{F}(9,107$ ) $=0.85, \mathrm{p}>.50$ ), indicating that the pattern was stable throughout the experiment.

Although they were relatively rare, examination of naming errors showed that they exhibited the same pattern as the reaction time data. Specifically, naming errors were more frequent in homogeneous than heterogeneous contexts for faces of all races except White (see Figure 1B)

## Discussion

Our study shows several novel findings. First, we observed semantic interference in proper name retrieval in blocked cyclic naming. To our knowledge, this has not been reported previously. Second, the basis of the interference was racial grouping of the faces, extending previous reports of semantic interference among taxonomic, thematic or ad hoc associates to social categories instantiated by facial features. In this domain, the basis for interference in name retrieval is quite slender, comprising modulation of the mapping from face to name by the mere knowledge that the named individuals belong to a distinct racial group. Third, the effect was present for three "other race" groups, but was absent for the "own race" of the white participants. This is interesting both as a new manifestation of the own race bias (i.e., a name retrieval bias), and as evidence that semantic interference does not arise under all conditions. The set of white faces could certainly be construed as a category in the context of this experiment, and yet we observed no evidence of interference. The error data even suggest that naming of homogeneous white faces may have been facilitated.

Previous researchers of semantic interference have taken pains to show that the effect is not an artifact of visual similarity (e.g., Damian et al., 2001). This concern also arises in the case of face naming, because racially homogeneous faces might be more difficult to discriminate. However, the data do not support this possibility. The visual similarity explanation would predict a context effect for all of the groups (not the case), and greater difficulty in naming outgroup than ingroup members in heterogeneous contexts (also not the case). Thus our findings are clearly driven by categorization of outgroup faces and not by a perceptual similarity confound.

One way to interpret our findings (and link them to previous findings with ad hoc categories) is that an autonomous face recognition process is followed by categorically constrained name retrieval. Against this, however, is the finding of Van Bavel et al. (2011) that social categories other than race affect the functioning of the fusiform face area and that nonracial group affiliations can even trump visually salient characteristics such as race. If context dynamically modulates the functioning of face processing areas, it may be more appropriate to conceive of

Figure 1: Response Times and Error Rates as a Function of Race and Heterogeneous or Homogeneous Context. (RTs and their standard errors (pooled) are estimated from the multi-level model).

## A. Response Times


B. Face Naming Errors

semantic interference in name retrieval as engaging relatively extensive neural networks.

To our knowledge, the present study is the first to demonstrate interference in name retrieval for faces from different racial categories. Research currently under way in our research group attempts to shed light on the basis of this effect by teasing apart whether common category membership, overlapping semantic features, or even shared visual features contribute to the effect. Another question we may examine in future is the cultural domain of the names. In this experiment, the names were all European in origin and they may have been perceived as more congruent with the White category. It is not clear, however, that this could account for the observed pattern of effects because tighter linkages among the names within the White category should presumably tend to increase rather than decrease interference and vice-versa for the other ethnic groups.

Additionally, we are investigating whether the observed name retrieval bias is specific to racial groups or extends to other social categories. Ongoing research is, for example,
using a minimal group paradigm in which participants are randomly assigned to novel and arbitrary groups (e.g., teams). Minimal groups trigger many of the same biases as other social categories (e.g., in face recognition; Van Bavel et al., 2011), and we anticipate that they may in this domain as well. Importantly, minimal groups do not differ systematically in visual features (e.g., members of all groups can belong to the same race), and participants do not possess semantic information (e.g., stereotypes) about the groups. To the extent that similar effects are observed with minimal groups, it will illuminate the role that categorization per se can play in interference effects. Our research contributes to the mounting evidence that influences of categorization on interference effects in word/name retrieval are dynamic, shifting as a function of currently available or salient categories (Abdel Rahman \& Melinger, 2011).

## Wider Implications

The own race bias in facial recognition has profound and disturbing social implications. For example, difficulties distinguishing between members of other races may be a significant cause of eyewitness misidentification and wrongful conviction in criminal cases (Scheck, Neufeld \& Dwyer, 2003). The current research suggests that a similar bias may occur in name retrieval, such that people have greater difficulty retrieving the names of other race individuals. Most of the time, the consequences of a name retrieval bias may be minor, but in certain contexts this bias could have pernicious effects. For example, teachers may be less likely to call on other race students, perhaps particularly if those students tend to be encountered proximally (e.g., seated together). The correlates and consequences of the name retrieval bias shown here merit further investigation.

## Acknowledgments

This research was supported by a Lehigh University CORE (Collaborative Research Opportunity) award. We thank Nevada Heft for contributions to the project.

## References

Abdel Rahman, R. \& Melinger, A. (2007). When bees hamper the production of honey: Lexical interference from associates in speech production. Journal of Experimental Psychology: Learning, Memory, and Cognition, 33, 604-614.
Abdel Rahman, R. \& Melinger, A. (2011). The dynamic
microstructure of speech production: Semantic interference built on the fly. Journal of Experimental Psychology: Learning, Memory, and Cognition, 37, 149161.

Brewer, M. B. (1988). A dual process model of impression formation. In T. Srull \& R. Wyer (Eds.), Advances in Social Cognition, Vol. 1 (pp. 1-36). Hillsdale, NJ: Erlbaum.

Cunningham, W. A., Van Bavel, J. J., Arbuckle, N. L., Packer, D. J., \& Waggoner, A. S. (2012). Rapid social perception is flexible: approach and avoidance motivational states shape P100 responses to other-race faces. Frontiers in Human Neuroscience, 6.
Damian, M. F., Vigliocco, G., \& Levelt, W. J. (2001). Effects of semantic context in the naming of pictures and words. Cognition, 81, B77-B86.
Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. Journal of Personality and Social Psychology, 56, 5-18.
Fiske, S. T., \& Neuberg, S. L. (1990). A continuum of impression-formation, from category-base to individuating processes: Influence of information and motivation on attention and interpretation. Advances in Experimental Social Psychology, 23, 1-74.
Griffin, Z. M. (2010). Retrieving personal names, referring expressions, and terms of address. Psychology of Learning and Motivation, 53, 345-387.
Howard, D., Nickels, L., Coltheart, M., \& Cole-Virtue, J. (2006). Cumulative semantic inhibition in picture naming: Experimental and computational studies. Cognition, 100, 464-482.
Hugenberg, K., Young, S. G., Bernstein, M. J., \& Sacco, D. F. (2010). The categorization-individuation model: An integrative account of the other-race recognition deficit. Psychological Review, 117, 1168-1187.
Ito, T. A., \& Bartholow, B. D. (2009). The neural correlates of race. Trends in cognitive sciences, 13(12), 524-531.
Ito, T. A., \& Urland, G. R. (2003). Race and gender on the brain: electrocortical measures of attention to the race and gender of multiply categorizable individuals. Journal of Personality and Social Psychology, 85, 616-626.
Kunda, Z., \& Spencer, S. J. (2003). When do stereotypes come to mind and when do they color judgment? A goalbased theoretical framework for stereotype activation and application. Psychological Bulletin, 129, 522-544.
Levin, D. T. (1996). Classifying faces by race: The structure of face categories. Journal of Experimental Psychology: Learning, Memory, and Cognition, 22, 1364-1382.
Levin, D. T. (2000). Race as a visual feature: Using visual search and perceptual discrimination tasks to understand face categories and the cross-race recognition deficit. Journal of Experimental Psychology: General, 129, 559574.

Levin, D. T., \& Angelone, B. L. (2002). Categorical perception of race. Perception, 31, 567-578.
Maass, A. (1999). Linguistic intergroup bias: Stereotype perpetuation through language. Advances in Experimental Social Psychology, 31, 79-121.
Maass, A., Salvi, D., Arcuri, L., \& Semin, G. R. (1989). Language use in intergroup contexts: The linguistic intergroup bias. Journal of Personality and Social Psychology, 57, 981-993.

Navarrete, E., Del Prato, P., and Mahon, B. (2012).Factors determining semantic facilitation and interference in the cyclic naming paradigm. Frontiers in Psychology, 3, Article 38, 1-15.
Oppenheim, G. M., Dell, G. S., \& Schwartz, M. F. (2010). The dark side of incremental learning: A model of cumulative semantic interference during lexical access in speech production. Cognition, 114, 227-252.
Scheck, B., Neufeld, P., \& Dwyer, J. (2003). Actual innocence: When justice goes wrong and how to make it right. New American Library.
Schnur, T., Schwartz, M., Kimberg, D., Hirshorn, E., Coslett, H., \& Thompson-Schill, S. (2009). Localizing interference during naming: Convergent neuroimaging and neuropsychological evidence for the function of Broca's area. PNAS, 106, 322-327.
Stroessner, S. J. (1996). Social categorization by race or sex: Effects of perceived non-normalcy on response times. Social Cognition, 14, 247-276.
Tajfel, H., \& Wilkes, A. L. (1963). Classification and quantitative judgement. British Journal of Psychology, 54, 101-114.
Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., \& Wetherell, M. S. (1987). Rediscovering the Social Group: A Self-Categorization Theory. Cambridge, MA: Blackwell.
Valentine, T., Brennen, T., \& Brédart, S. (1996). The cognitive psychology of proper names: On the importance of being Ernest. London: Routledge.
Van Bavel, J. J., Packer, D. J., \& Cunningham, W. A. (2011). Modulation of the fusiform face area following minimal exposure to motivationally relevant faces: evidence of in-group enhancement (not out-group disregard). Journal of Cognitive Neuroscience, 23, 33433354.

# Changing minds by tracking eyes: Dynamical systems, gaze and moral decisions 

Philip Pärnamets (philip.parnamets@lucs.lu.se)<br>Petter Johansson (petter.johansson@lucs.lu.se)<br>Christian Balkenius (christian.balkenius@lucs.lu.se)<br>Lars Hall (lars.hall@lucs.lu.se)<br>Lund University Cognitive Science, Lund University, Kungshuset, Lundagård, 222 22, Lund, Sweden<br>Michael J. Spivey (spivey@ucmerced.edu)<br>Department of Cognitive Science, University of California, Merced, Merced, CA 95344, USA<br>Daniel C. Richardson (dcr@eyethink.org)<br>Cognitive, Perceptual and Brain Sciences, University College London, Gower Street, London, WC1E 6BT, UK


#### Abstract

Decision making is a dynamic process. Alternatives compete over time, and this competition plays out in sensorimotor processes. This is true not just for perceptual decisions or simple categorisation tasks, but also for moral decisions, which are the outcome of a complex interplay of intuition, emotion and reasoning. In this experiment, we first establish a descriptive and causal link between gaze and moral judgement. We then use eye movements to track the time course of participants' moral decisions and show that by interrupting their decision process based on their gaze position, we are able to influence what they decide. We interpret this as evidence for a dynamical systems view of decision making and argue that our results provide new insights into how judgements are reached and constructed in our embodied minds.


Keywords: Decision making; morality; dynamic systems; eye tracking

Imagine a jury, evenly split over a verdict concerning a murder. One jury member is yet to make her decision, which will decide the fate of the accused. She weighs up her choice, looking between the faces of those who argued for and against conviction, glancing at the evidence and police reports on the table. The foreman clears her throat: the jury must take a vote now.

The jury member's decision could be analysed in terms of the evidence that is presented and how it is framed, and many experiments have investigated such factors. But here we are interested in one particular, often overlooked aspect: the precise moment of choice, in this case when the foreman cleared her throat. We claim that the precise timing of events like these may have a causal influence over a decision. To make our case, we adopt a perspective viewing decision-making as a fundamentally dynamic process. In the decision process, two, or more, options compete over time until one option reaches a threshold or the process is interrupted and the system is forced to reach a conclusion. Secondly, we present evidence that eye movements can reveal something of this process. Where the jury member looked, moment-by-moment showed what option she was considering. Our novel claim is that by manipulating when someone is forced to make a decision, and, hence, knowing
where they are in their decision trajectory, an influence can be exerted of what is decided

## Process of moral judgements

Recent research in moral psychology emphasizes how contextual factors influence processes underlying moral judgements, factors such as emotional state (Wheatley \& Haidt, 2005), political preferences (Graham, Haidt \& Nosek, 2009) and causal structure of the moral problem (Cushman, Young \& Hauser, 2006). In particular the interplay between intuitions, emotions and reasoning has been of central concern.
In two of the most influential models in this tradition, Haidt's Social Intuitionist Model (Haidt, 2001) and Greene's Dual Process Model (Greene et al. 2008), moral cognition is viewed as being comprised of a number of modules each dedicated towards processing specific forms of information. These modules then discretely combine their output to produce a moral judgement, however, the computational properties of the system are typically not spelled out nor how strict the modular metaphor is to be interpreted.

Taking a cue from dynamical systems modelling of cognition we propose an alternative to stage-based accounts. We view the processes of making moral judgements as a stochastic system of graded, probabilistic representations, in which a judgement can be understood as a temporary settling of the system around an attractor basin in a decision space (McKinstry, Dale \& Spivey, 2008; Spivey, 2007). In this study, we exploit a proposed coupling between cognition and gaze behavior to show the dynamic nature of moral judgements.

## Dynamic and embodied minds

Minds can be understood and modelled as complex dynamic systems. The discrete symbol and motor output that characterises language and action according to standard models can be generated by graded, probabilistic processes on a continuous timescale, extending beyond the discrete partitions that our everyday practices impose on our understanding of ourselves (Spivey, 2007; Van Orden,

Holden \& Turvey, 2005). There is evidence for the neural plausibility stemming from studies measuring and influencing saccades and saccadic programming in real time (Gold \& Shadlen, 2000) as well as the large scale probabilistic nature of neural populations (Pouget, Dayan \& Zemel, 2003).

A key element of this dynamic view of mind is the tight coupling between sensorimotor outputs and cognitive processing in general - an embodied view of cognition. This is evidenced, for example, during linguistic processing where persons glance towards phonological competitors while viewing an array of objects, by for example looking towards a candle when hearing 'candy' (Tanenhaus, SpiveyKnowlton, Eberhardt \& Sedivy, 1995) and when two people are engaged in a conversation with each other (Richardson, Dale \& Tomlinson, 2009). Eye movements have been shown to closely follow cognition during spatial indexing tasks in adults (Hoover \& Richardson, 2008) and infants as young as 6 months old (Richardson \& Kirkham, 2004).

Similarly, mouse movements will show curvature towards distracting alternatives indicating competition in categorisation tasks, for example when classifying whales as being fish or mammals, or, analogously with the eye movement result above, when processing linguistic inputs (Freeman, Dale \& Farmer, 2011).

Graded representations have also been found in more advanced reasoning tasks. In one study, participants were asked to judge the truth of a number of propositions that were selected to represent various steps of veridicality. For the propositions with intermediate truth values mouse movements would veer longer between answers thus tracking the more arduous cognitive task of assigning a truth value in these trials (McKinstry et al., 2008).

## Gaze preference and choice

The link between gaze and decision making has also been investigated by a number of studies investigating preference formation and decisions. In one study participants were asked to choose which of two faces they found more attractive. Their eye movements exhibited a bias towards the about to be chosen alternative, a finding dubbed the gaze cascade effect (Shimojo et al. 2003). The increasing likelihood to gaze towards a preferred alternative has also been demonstrated for participants considering difficult moral dilemmas (Pärnamets, 2008), indicating that gaze could contribute to moral judgements as well.

In addition, the experiment by Shimojo et al. (2003) demonstrated a possibility to bias preference judgements by actively directing gaze towards one face for longer periods of exposure than the alternative. Similar methods have been utilised to bias consumer decisions for candy bars (Armel, Beaumel \& Rangel, 2008). In all these experiments, however, different information is being presented to participants by artificially directing their gaze towards different alternatives.

By contrast, we propose in our experiments that choice can be influenced by manipulating only the timing of the decision and not the stimuli the participant attends to.

## Hypothesis

We investigated whether the coupling between eye movements and cognitive processes could be leveraged to influence the discrete end-state of the dynamic process.

We hypothesised, as suggested by pilot experiments from our lab (Richardson, Spivey \& Hoover, 2009), that the direction of participants' gaze could be an index of which attractor basin in their decision space they are gravitating towards. Using this information we would be able to collapse their decision function and bias their judgement to the currently favoured alternative, even if that alternative might not have been the option they would have preferred, had the decision process been allowed to take its noninterrupted course.

## Experiments

We devised a series of experiments that attempted to bias participants' decisions by monitoring their gaze. The first experiment, comprised of two studies (1a and 1b), was designed with a view of establishing an upper bound for our expected effect as well as exploring the link between gaze, as an index of thought, and judgement necessary for our paradigm to work. The second experiment replicates findings for face preference and consumer decisions by manipulating direction and duration of gaze and allows us to establish a causal link between gaze and choice for moral judgements.

The third experiment is our main study which demonstrates the hypothesised effect; influencing decisions solely on the basis of timing. Experiment 4 addresses a possible objection to our procedure and replicates our main finding.

## Equipment and materials

Eye tracking was performed using an SMI RED 250 eye tracker running at 250 Hz on a 19" screen with a resolution of $1680 * 1050$ pixels. Stimuli were presented using PsychoPhysics Toolbox (Kleiner, Brainard \& Pelli, 2007) running on MatLab 2010b (The MathWorks, Natick, MA.). Gaze was sampled by the MatLab script with 10 ms intervals. Calibration was performed on each subject at the start of the experiment using 5 points followed by 4 validation points. Calibrations with error exceeding $1^{\circ}$ visual angle in more than one case were rerun. Average error was less than $0.5^{\circ}$.

There were a total of 98 items that participants were asked to listen to and make judgements about. Of these, 63 were moral items and 35 were factual items. The factual items were used previously in our pilot studies (Richardson et al., 2009) and were all propositions that have an average $50 \%$ truth value, meaning that a large sample of persons were found equally likely to judge the propositions as being true
or false. An example proposition would be "Is coffee bad for your health?" with the alternatives "Yes" and "No".

The moral items were derived from Moral Foundations Theory (MFT, Graham et al., 2009) and propositions were designed to fit with each of the five categories found in MFT. In addition a few propositions were of a meta-ethical character. An example item is "Murder is sometimes justifiable" with the alternatives "Sometimes justifiable" and "Never justifiable". The alternatives were such that it would be informative to view both for the participant.

## General Procedure

Participants were asked to sit in front of a computer screen wearing a pair of headphones. They were instructed that during each trial they would hear a sentence stating either a moral or a factual proposition. Two alternatives would then be shown on the screen, one on the left-hand side of the screen and the other on the right-hand. Their task was to use their judgement to select the alternative that they thought was right in relation to the sentence they heard.

Participants indicated their selection by clicking the right or left mouse button, where the buttons corresponded to the alternatives presented on either the right or left side of the screen. The alternatives were visible for a maximum of 3000 ms , or until the experimental 'trigger' went off. This trigger, based on their eye movements, varied between experiments as explained below. The participants then saw a prompt asking them to "Choose now!" Participants were instructed to respond quickly once the prompt was shown on the screen. After each trial, a 1-7 continuous confidence scale was also presented.

Participants were told that the alternatives would be visible for a random and short amount of time each trial and were asked to view both alternatives. Unbeknownst to the participants the timing of each trial was dependent on their eye movements which were being concurrently recorded. The experimental trigger determining the length of each trial was based on the input from the eye tracker during each trial. It was sent to go off as soon one of the alternatives had accumulated at least 750 ms of dwell time and the other alternative had accumulated at least 250 ms of dwell time. These criteria ensured that the trigger would not go off until both alternatives had been seen by the participant. The exact conditions governing the trigger varied between the experiments and are detailed below.

If participants did not set off the trigger then the trial would time out after 3000 ms , and the participants would then be asked to make a choice. Trials timed out either because the participants failed to look at both alternatives, or because the eye-tracker momentarily lost track of the participants' gaze and so failed to capture when the participants shifted their focus between the two choice options. In both cases, there is no way for us using this paradigm to interrupt a decision process where participants are drawn between two competing alternatives. All such time-out trials were removed from further analysis.

Participants indicated during debrief that they occasionally would fail to understand an item and in those cases typically indicated very low confidence following that trial. Trials with confidence $<1.5$ of 7 were removed for that reason.

## Experiment 1a

Procedure In this first experiment we wanted to establish that the coupling between eye movements and cognition would be present even for our moral items. We did not attempt to bias participants' decisions at this point. We simply wanted to show that there is a relationship between the distribution of gaze across two alternatives, and which alternative is eventually chosen. This also allows us to establish an upper bound for the effect size of our later attempts at influencing decisions.

In each trial of this experiment the first 300 ms of viewing time were not counted towards the trigger, giving participants some extra time to orient themselves during each trial. From 300 ms onwards, we kept track of how long the participant looked at each alternative. As soon as one alternative was viewed for at least 750 ms and the other for 250 ms , the trigger was set off. We termed the alternative that had been looked at the longest the target. A success was counted if that alternative was chosen, otherwise a failure.

Fifteen persons (11 female) participated in this experiment with a mean age of $20.80 \quad(\mathrm{SD}=2.04)$. Participants were recruited through both the public and student-based subject pools at University College London.

Results and Discussion For the moral items 603 trials ( $67.67 \%$ ) were successful ( $p<0.0001$, Binomial test). For the factual items 280 trials ( $65.57 \%$ ) were successful ( $\mathrm{p}<0.0001$, Binomial test) (see figure 1).

This version demonstrates the plausibility of using gaze as an index of mental trajectories and using this information to collapse decision space towards an alternative under consideration. We used a 300 ms wait time in order to follow the procedure of our pilot work (Richardson et al. 2009). But these early eye movements could presumably also contribute to the decision vector. We therefore ran a second version of the experiment without the 300 ms wait time to investigate this.

## Experiment 1b

Procedure This Experiment was identical to 1a above, apart from we did not use a 300 ms wait time before eye movements to both alternatives were measured. Twenty persons ( 10 female) were recruited through both the public and student-based subject pools at University College London. Participants had a mean age of $27.20(\mathrm{SD}=8.07)$.

Results and discussion For the moral items, 716 trials ( $60.02 \%$ ) were successful ( $p<0.0001$, Binomial test). For the factual items 385 trials (68.14\%) were successful ( $\mathrm{p}<0.0001$, Binomial test) (see figure 1 ).

From Experiments 1 a and 1 b we conclude that eye movements are closely linked to the decision making process. If we actively interrupt participants during their deliberation, they are more likely to choose the alternative they have looked at for longer, even for complex moral judgements. We cannot yet claim the causal connexion between the timing of our interruption and the content of the judgement, of course. However, these experiments allow us to establish an upper bound for an expected effect size for the later experiments in which we attempt to bias their decisions in a predetermined manner.

## Experiment 2

In experiment 2 we wanted to establish the causal connexion between gaze and choice for the judgements which we were interested in. We adopted the methods used by, for example, Shimojo et al. (2003), where the combination of gaze and exposure, but not exposure alone, had been shown to influence choice. We constructed an experiment where we would be directing participants' gaze towards alternatives so that they would be more exposed to the target alternative compared to the non-target.

Procedure. The procedure in Experiment 2 differs significantly from the general procedure of the other experiments reported here.

Once participants had heard the item, they were presented with one alternative at a time with each presentation lasting 400 ms . One alternative was always shown on the right-hand side and the other on the left-hand side. The different alternatives appeared pseudo-randomly, such that the target alternative was given a $3: 1$ exposure weighting. Total combined viewing time for both the alternatives was 3200 ms . Target and non-target alternatives were presented in random order. Choice was indicated after the presentation sequence had completed, as in Experiment 1. Nineteen persons ( 13 female, mean age 22.36, $\mathrm{SD}=3.82$ ) participated in Experiment 2. Participants were recruited through both the public and student-based subject pools at University College London.

Results and discussion For the moral items 600 (53.29\%) trials were successful ( $\mathrm{p}<0.05$, Binomial test), while 287 (54.99\%) trials were successful for the factual items ( $\mathrm{p}<0.05$, Binomial test) (see figure 1 ).

We conclude that there is a causal connexion between gaze and choice. In addition, we demonstrate the possibility to bias moral and factual judgements with the help of directed gaze and exposure effects, and this on the relatively small time scales that our paradigm is operating on. Typical trials in the literature using this method usually last around twice as long as ours.

Additionally, to our knowledge, this is the first empirical demonstration of this effect for moral judgements, and as such, it is a remarkable finding in itself.

## Experiment 3

Our goal in this experiment was to exert an influence over the decisions that participants made by manipulating nothing but the timing of their decisions. On each trial, we randomly determined which alternative we would try to bias the participant towards. Unlike in previous experiments that have biased decisions by changing stimuli or directing gaze, including our Experiment 2, our participants looked freely at the alternatives in front of them. We simply tracked the time course of their eye movements during the decision process, and prompted the participants to decide when we judged that their gaze suggested they were at a particular point in their decision space that favoured the option we were trying to influence them to choose.

Procedure Experiment 3 was identical to experiment 1 b in all respects except that here the experiment program would, for each trial, randomly designate one alternative as the target. The trigger would only go off if that target alternative had accumulated at least 750 ms of dwell time and the other, non-target, alternative had accumulated at least 250 ms of dwell time.

Twenty persons (14 female, mean age 29.60, $\mathrm{SD}=13.14$ ) participated in experiment 3. Participants were recruited through both the public and student-based subject pools at University College London.

Results and discussion For the moral items 609 (58.22\%) trials were successful ( $\mathrm{p}<0.0001$, Binomial test). For the factual items 282 (56.51\%) of trials were successful ( $\mathrm{p}<0.005$, Binomial test) (see figure 1 ).

This finding demonstrates that we are able to influence participants' judgements in both moral and factual decisions by tracking their gaze alone. We merely asked them to respond at a given point in time when their eye movements reveal them being in a position in their decision space that indicates them gravitating towards a given alternative. We claim that this finding supports our view of the dynamic nature of judgements, where judgements can be understood as trajectories in decision space travelling between alternatives conceived of as attractor basins in that space.

We also wish to highlight the difference between this experiment, where participants are using their gaze actively and are unconstrained in the environment, and Experiment 2 where participants, while moving their eyes, are passive recipients of information. Given this difference and the fact that our paradigm allows varying degrees of relative exposure to the alternatives, in virtue of how the trigger is set up, we argue that our finding in Experiment 3 represents a novel connexion between gaze and choice.

## Experiment 4

One possible objection to our claims is that perhaps participants have already made their decisions well before the trigger is set off. It is conceivable that participants gaze towards the target for longer than 750 ms , settle on that alternative, and then simply avert their gaze towards the
other alternative out of boredom, which sets off the trigger. Participants then choose the designated target, and the trial is counted as a success. But, the objection goes, if they were allowed to indicate their choice as soon as they'd made it, they would have clicked much earlier in the trial. Experiment 4 was designed to meet this objection by removing the constraint that participants are unable to respond before the prompt screen.

Also, in addition to confidence ratings two extra followup questions were asked of participants after each trial. One concerned asked if they had been able to hear the items and see the alternatives properly. The second concerned asked how important the issue raised by the item was to them, which pertained primarily to the moral items. The two additions were made to meet to further objections to experiment 2 , namely that participants would only be biased towards our target on items they felt were unimportant or where they failed to fully understand the item or alternative. These were also set in terms of 1-7 continuous scale.

Procedure The trigger was set up to work as in experiment 2. In addition, time participants were instructed that they could also indicate their choice when they had made up their minds, clicking the mouse in the same way as they would during the 'Choose now!' screen.

21 persons ( 17 female, mean age 21.81, $\mathrm{SD}=5.38$ ) participated in experiment 3. Participants were recruited through both the public and student-based subject pools at University College London.

Results and discussion For the moral items 240 (21.27\%) trials were such that the participant responded before the experimental trigger was activated. The corresponding number for the factual trials was 264 ( $45.28 \%$ ). Since these are cases where the trigger has not been activated, these trials were not analysed further.

Of the remaining trials 496 (55.86\%) were successful ( $\mathrm{p}<0.001$, Binomial test) for the moral items, and for the factual items 189 (59.24\%) trials were successful ( $\mathrm{p}<0.005$, Binomial test) (see figure 1). Analysing the various ratings participants made after their decision, we found that there were no significant difference between trials that were


Figure 1: Results from the experiments for all items. $95 \%$ confidence intervals are shown.
successfully biased and those that were not. For moral items, there was no difference in comprehension between successful trials ( $\mathrm{M}=6.57$ ) and unsuccessful trials $(\mathrm{M}=6.50)$ $(\mathrm{t}(809.223)=1.67, \mathrm{p}=0.09)$; and no difference in perceived importance on successful trials ( $\mathrm{M}=4.98$ ) and unsuccessful trials $(\mathrm{M}=4.92)(\mathrm{t}(847.692)=0.57, \mathrm{p}=0.57)$. Similarly, for factual items there was no difference in comprehension between successful trials $(M=6.55)$ and unsuccessful trials $(\mathrm{M}=6.52)(\mathrm{t}(262.72)=0.261, \mathrm{p}=0.79)$; and no differences in importance ratings either ( $M=3.53$, successful, $M=3.35$, unsuccessful, $\mathrm{t}(268.412)=0.92, \mathrm{p}=0.36)$.

We find in this experiment that there are trials where participants make up their minds before our manipulation is triggered. But these cases do not explain our findings, since when they are excluded from the analysis, our biasing effect remains. Experiment 4 still demonstrates judgements malleable to influence depending solely on the timing of judgement based on measuring gaze indexed thought trajectories.

## General Discussion

We have argued that decision making is a dynamic system exhibiting a tight coupling between eye-movements and judgement. We have demonstrated a causal link between gaze and choice using a paradigm utilising exposure and directed gaze. We then, following the logic of dynamical systems, have shown that we are able to influence participants' moral and factual judgements using gaze only as an index of thought, and by manipulating nothing but the timing of the decisions. The results from Experiment 4 suggest that our effects are no mere artefacts of the experimental procedure.

A surprising aspect of our results is the demonstration of the malleability of moral judgments on very small timescales across the wide spectrum of moral domains which our stimuli encompass. This malleability is present even when we manipulate only the timing of decisions, rather than by adding information to the situation, as has typically been the case in the literature. We emphasise that while we interpret our findings in the light of a dynamical systems perspective on mind, the effects on moral judgements are of significant interest by themselves for understanding our moral mind.

One valid concern about our findings is the relatively small effects sizes. This is not too surprising, however, due to the fact that gaze and decision making processes, while linked, cannot be not rigidly yoked together. For one thing, eye movements have various biological constraints and are necessarily discrete, whereas thought processes could be continuous and graded; for another, one can chose between options while fixating a single point. Given the partial though pervasive (Spivey, Richardson \& Dale, 2009) - link between eye movements and cognition, it is not surprising that gaze is an imperfect indicator of decision processes, and our bias effects are the size that they are. Indeed, we would argue that it is remarkable that they exist at all.

A significantly higher effect size would be surprising for other reasons as well. One reason is that this would jar with our capacity to, at times, be moral agents. Recall, we are not making a claim about a possible lack of moral agency in our participants, only a claim about morality's dynamic nature. In a sense the size of our effect says something very real about the strength of our participants' moral systems. It opens up avenues for a more detailed exploration of individuals' and groups' moral landscapes, as well as understanding the complex interplay between cognition, sensorimotor systems and the environment.

In future work we hope to expand the range of moral decisions under examination, and develop computational models of the process. The class of judgements we have used here are first person judgements about one's personal moral values. These are judgements which are known to be open to manipulation (Hall, Johansson \& Strandberg, 2012). In further research we plan to extend our results to third person judgements and concrete moral action. Eventually, we hope to model the dynamics of moral decision marking. Drawing on our current findings, we aim to develop a model that exhibits the same drift towards attractor states when interrupted as we found in our participants and compare this to alternative accounts such as drift-diffusion and accumulator models. But from the results we have presented here, we hope the case has been made that that much can be learned about our moral selves by focusing on the deep integration between cognitive and perceptual functions and how this integration plays out in time.

## Acknowledgments

This research was made possible with the support of Stiftelsen Erik och Gurli Hultengrens fond för filosofi vid Lunds universitet and Makarna Ingeniör Lars Henrik Fornanders fond to P.P. We thank Abigail Hodder for lending her voice to recordings of the items used in these experiments.

## References

Armel, K.C., Beaumel, A. \& Rangel, A. (2008) Biasing simple choice by manipulating relative visual attention. Judgment and Decision Making. 3(5), 296-403.
Cushman, F., Young, L. \& Hauser, M. (2006) The role of conscious reasoning and intuition in moral judgment. Psychological Science, 17(12), 1082-1089.
Freeman, J.B., Dale, R. \& Farmer, T.A. (2011) Hand in motion reveals mind in motion. Frontiers in Psychology. 2, 59.
Gold, J.I. \& Shadlen, M.N. (2000) Representation of a perceptual decision in developing occulomotor commands. Nature. 404, 390-394.
Graham, J., Haidt, J. \& Nosek, B.A. (2009) Liberals and conservatives rely on different sets of moral foundations. Personality Processes and Individual Differences. 96(5), 1029-1046.
Greene, J.D., Morelli, S.A., Lowenberg, K., Nystrom, L.E. \& Cohen, J.D. (2008) Cognitive load selectively interferes
with utilitarian moral judgment. Cognition. 107, 11441154.

Haidt, J. (2001) The emotional dog and its rational tail. Psychological Review. 108(4), 814-834.
Hall, L., Johansson, P. \& Strandberg, T. (2012) Lifting the veil of morality: Choice blindness and attitude reversals on a self-transforming survey. PLOS One. 7(9), e45457
Hoover, M. A. \& Richardson, D. C. (2008). When Facts Go Down the Rabbit Hole: Contrasting Features and Objecthood as Indexes to Memory. Cognition, 108(2), 533-42
Kleiner, M., Brainard, D. \& Pelli, D. (2007) What is new in Psychtoolbox-3? Perception 36 ECVP Abstract Supplement.
McKinstry, C., Dale, R. \& Spivey, M.J. (2008) Action dynamics reveal parallel competition in decision making. Psychological Science. 19(1), 22-24.
Pouget, A., Dayan, P. \& Zemel, R.S. (2003) Inference and computation with population codes. Annual Review of Neuroscience. 26, 381-400.
Pärnamets, P. (2008) Perceptual components in moral decision making. Unpublished master's thesis. Lund University, Lund, Sweden.
Richardson, D.C., Dale, R. \& Tomlinson Jr, J.M. (2009) Conversation, gaze coordination and beliefs about visual context. Cognitive Science, 33, 1468-1482.
Richardson, D.C \& Kirkham, N.Z (2004) Multimodal events and moving locations: Eye movements of adults and 6-months-olds reveal dynamic spatial indexing. Journal of Experimental Psychology: General, 133(1), 4662.

Richardson, D.C, Spivey, M.J., \& Hoover, M.A. (2009) How to influence choice by monitoring gaze. In N.A. Taatgen \& H. van Rijn (Eds.), Proceedings of the 31th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.
Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhardt, K.M. \& Sedivy, J.C. (1995) Integration of visual and linguistic information in spoken language comprehension. Science. 268(5217), 1632-1634.
Shimojo, S., Simion, C., Shimojo, E. \& Scheier, C. (2003). Gaze bias both reflects and influences preferences. Nature Neuroscience, 6(12), 1317-1322.
Spivey, M.J. (2007). The continuity of mind. New York, Oxford University Press.
Spivey, M.J, Richardson, D.C, \& Dale, R. (2009). The movement of eye and hand as a window into language and cognition. In E. Morsella, J. A. Bargh, \& P. M. Gollwitzer (Eds.), Oxford Handbook of Human Action (pp. 225-249). New York: Oxford University Press.
Van Orden, G.C., Holden, J.G. \& Turvey, M.T. (2005) Human cognition and 1/f scaling. Journal of Experimental Psychology: General. 134(1), 117-123.
Wheatley, T. \& Haidt, J. (2005) Hypnotic disgust makes moral judgements more severe. Psychological Science, 16(10), 780-784.

# Multimodal Networks of Interpersonal Interaction and Conversational Contexts 

Alexandra Paxton (paxton.alexandra@gmail.com) Rick Dale (rdale@ucmerced.edu)<br>Cognitive and Information Sciences, University of California, Merced Merced, CA 95341 USA


#### Abstract

In interpersonal interaction, the terms synchrony or alignment refer to the way in which communication channels like speech or body movement become intertwined over time, both across interlocutors and within a single individual. A recent trend in alignment research has targeted multimodal alignment, exploring how various communication channels affect one another over time (e.g., Louwerse et al., 2012). While existing research has made significant progress in mapping multimodal alignment during task-based or positively valenced interactions, little is known about the dynamics of multimodal alignment during conflict. We visualize multimodal alignment during naturalistic affiliative and argumentative interactions as networks based on analyses of body movement and speech. Broadly, we find that conversational contexts strongly impact the ways in which interlocutors' movement and speech systems self-organize interpersonally and intrapersonally.


Keywords: alignment; conflict; conversation; interaction; movement; network; speech; synchrony

## Introduction

Interpersonal communication is a multimodal activity. Conversation incorporates multiple channels of communication that enrich the interaction, like hand gestures, facial expressions, posture, and speech. To effectively communicate with one another, interlocutors cue in to each of these channels simultaneously, often without realizing the importance placed on each of them.

Considerable work has surveyed multimodal qualities of interaction (e.g., Norris, 2004). However, in general, experimental study of interpersonal communication in cognitive science tends to target single behavioral channels. This has led to significant advances in our understanding of these specific channels, but there is still much work to be done in investigating the connections among them. Continued multimodal research will likely yield interesting-and possibly unexpected-relationships among channels that have been extensively explored unimodally. For this reason, the current research explores multimodal alignment situated within different conversational contexts.

Specifically, the present research examines interpersonal communication through alignment dynamics. Research on interpersonal alignment focuses on how affect, behavior, and cognition of interacting individuals affect one another over time. Over the past several decades, researchers have explored interpersonal alignment over a range of channels, from movement (e.g., Richardson et al., 2007) to speech (e.g., Garrod \& Pickering, 2004) to cognition (e.g., Brennan, Galati, \& Kuhlen, 2010). As aforementioned, the majority
of this work centers on one or two behavioral channels, but a growing body of literature has begun to investigate how multiple channels align during communication (e.g., Louwerse, Dale, Bard, \& Jeuniaux, 2012).

A distinction can be drawn between alignment in a general sense and synchrony. ${ }^{1}$ We use the term alignment to refer broadly to the concept that individuals, over time, change their affect, behavior, and cognition as a direct result of their interaction with another individual. This umbrella term encompasses everything from mimicry, in which individuals are performing highly similar behaviors to their interaction partner (e.g., Chartrand \& Bargh, 1999), to more complementary behavior patterns like synergy (e.g., Riley, Richardson, Shockley, \& Ramenzoni, 2011). Synchrony, on the other hand, can be considered a specific pattern of alignment and refers exclusively to the in-phase entrainment of behavior or communication channels.

In the spirit of intrapersonal alignment research spanning the last several decades (e.g., Haken, Kelso, \& Bunz, 1985), the current research explores the dynamics of interpersonal alignment as a self-organizing property of human interaction. The present research focuses specifically on two channels of communication-speech and body movementand the ways in which these channels are affected by each other, by conversation partners, and by the conversational context. If it is true that human interaction self-organizes around, for example, conversational goals, we should expect to see multimodal alignment patterns changing across these different contexts.

## Body Movement and Speech Alignment

Previous work on speech and body movement demonstrates rich interactivity between and within individuals. Some of the earliest work in alignment research presents evidence for interpersonal and intrapersonal multimodal alignment between movement and speech channels (Condon \& Ogston, 1966). Since then, research has continued to explore intrapersonal alignment, both in speech (Reitter, Moore, \& Keller, 2010) and movement (Beek, Peper, \& Daffertshofer, 2002). However, body movement and speech have both been studied extensively within the interpersonal alignment literature as well, generally within affect-neutral, positively valenced, or task-oriented settings.

Studies of body movement alignment have spanned a wide variety of behaviors, including gesture (Bernieri \&

[^180]Rosenthal, 1991), stepping (Miles, Griffiths, Richardson, \& Macrae, 2010), and overall levels of body movement (Paxton \& Dale, in press). Often, work on body movement alignment incorporates elements of social psychology, like investigating how interpersonal alignment affects liking (Chartrand \& Bargh, 1999). Broadly speaking, these findings generally cast the phenomenon as a pervasive and relatively automatic process that can be enhanced with liking or rapport. However, limited research suggests that higher-level social factors may inhibit bodily alignment (Miles et al., 2010; Paxton \& Dale, under revision).

Individuals also align over numerous measures of speech. Over time, interacting individuals have been shown to use more similar acoustic features (Kousidis \& Dorran, 2008), sentence structures (Cleland \& Pickering, 2003), and even respiratory patterns (McFarland, 2001). The tendency toward alignment during interaction is so powerful that individuals even align to simulated partners (Krämer, Kopp, Becker-Asano, \& Sommer, 2012). These and related findings of interpersonal alignment in speech have led some to suggest that this is an automatic tendency driven in part by shared cognitive representations (Brennan et al., 2010).

These past findings point to a distinct temporal structure of speech and movement during interaction. Recently, researchers have begun to emphasize the importance of investigating interpersonal multimodal alignment on a large scale (e.g., Delaherche \& Chetouani, 2010; Louwerse et al., 2012). These questions allow researchers to more fully understand the complex, interdependent structure of communication channels during interaction.

## Dynamics of Interpersonal Alignment

Mechanistic models of body mechanics (e.g., interlimb coordination; Haken et al., 1985) have influenced recent work on the dynamics in interpersonal interaction (e.g., Miles et al., 2010). Researchers have begun to explore the forms and functions of alignment, going beyond earlier studies simply investigating its existence. Such work is dedicated to exploring the time course of alignment with the belief that-like many other phenomena-alignment is neither static nor uniform across contexts.

Like the work discussed above, research on interaction dynamics has also focused on speech and movement channels. From gaze patterns and postural sway (Shockley, Richardson, \& Dale, 2009) to speech production (Tilsen, 2009), researchers have found support for dynamical interpersonal and intrapersonal alignment, both unimodal and multimodal. Individuals' patterns of alignment change with task demands (Sebanz, Bekkering, \& Knoblich, 2006), social context (Miles, Lumsden, Richardson, \& Macrae, 2011), and even physical environment (Richardson et al., 2007), providing further support for claims of context dependence in alignment (Riley et al., 2011).

## The Present Study

Previous work has pointed to distinct patterns of organization of speech and body movement between and
within individuals, and recent trends have begun to situate this alignment within the context of multimodal interaction. Yet, so far, the dynamics of multimodality in alignment have remained relatively unexplored. Moreover, research on alignment thus far has sampled only a small percentage of the total space of human communicative contexts, focusing primarily on task completion or friendly interactions.

Our primary goal in the present study is to add to the growing literature on the dynamics of multimodal alignment. We aim to extend various elements of previous findings on unimodal and multimodal alignment, both interand intrapersonally. The present study explores participants’ speech patterns and their relation to body movement with three initial hypotheses. First, we hypothesize that individuals will align to one another's speech patterns but will not demonstrate in-phase synchrony of speech, due to the natural constraints of turn-taking. Second, we anticipate that individuals will exhibit multimodal intrapersonal synchrony, tending to move and speak at the same time. Third, although we anticipate that individuals will be most likely to move while speaking (and the reverse), we expect that analyses will reveal some evidence for interpersonal multimodal alignment (e.g., due to nodding).

The current project extends previous work on multimodal alignment further by situating the research within different conversational contexts, namely affiliation and argument. As part of a larger line of research investigating alignment in various contexts, the present research brings a focus on asymmetric contexts-interactions in which individuals have conflicting, differing, or opposing goals-to bear on questions of multimodal alignment. Previous research has demonstrated that conflict significantly decreases levels of interpersonal bodily synchrony (Paxton \& Dale, under revision). We continue to explore alignment during conflict in the present project. Compared to non-asymmetric contexts, we anticipate that argument will affect alignment in several ways: first, that individuals will demonstrate a more rigid turn-taking structure (possibly, e.g., to satisfy implicit social demands for reciprocity); second, that levels of intrapersonal multimodal synchrony will remain consistent; and third, that levels of interpersonal multimodal alignment will decrease. This pattern of results-balancing stable conversational structures with sensitivity to contextual factors-would reinforce claims of contextdependent, emergent properties of human interactions.

In addition to analyzing these data, the present study also hopes to begin work towards descriptive models of interpersonal and intrapersonal multimodal and unimodal alignment in different conversational contexts. After presenting our analyses, we highlight our findings in visualization networks of multimodal alignment dynamics.

## Method

## Corpus

The data presented here were collected by the authors as part of a larger corpus comparing interpersonal alignment
during argument and affiliation. The corpus comprised over 35 naïve participant dyads engaged in different conversational settings, collected from the University of Memphis and the University of California, Merced. As a further exploratory analysis building on previous findings (Paxton \& Dale, in press; Paxton \& Dale, under revision), the present analyses were performed on a subset of the participants from the University of California, Merced, based on uniformity of experimental conditions. The audio data had not yet been analyzed, separately or in conjunction with body movement.

## Participants

24 undergraduate participants (mean age $=20.14$ years) were recruited as 12 dyads ( 6 female, 6 mixed-sex) through the school's online subject pool system. Participants signed up independently and were unable to see their partner's identity beforehand. Only one dyad reported having known one another prior to participation. One mixed-sex dyad was dropped from present analyses because their opinions were too similar to achieve any argument during the experiment.

## Procedure

Upon arrival, participants were separated and individually completed a number of questionnaires prior to interacting with one another. One of the questionnaires was an opinion survey that included a number of sociopolitical (e.g., abortion, death penalty, legalization of marijuana) and university-specific (e.g., a campus rule forbidding freshmen students from bringing cars to campus) issues. The opinion survey posed the issues as open-ended, opinion-neutral questions. For each item, participants were given several lines to write their opinion and were directed to indicate the strength of that opinion on a Likert-style scale from 1 (feel very weakly) to 4 (feel very strongly).

Participants were randomly assigned to one of two conditions based on the order of the two target conversations that they were prompted to have. All dyads held a brief introduction conversation without the experimenter present ( $\sim 3 \mathrm{~min}$ ) and two target conversations (10min each). Half of the dyads were given an affiliative prompt first and an argumentative prompt second; half of the dyads experienced the reverse order. After each target conversation, participants were separated to complete postconversation measures. Participants were not informed in advance of the conversation topics. After holding both target conversations, participants were thanked and debriefed.

For the affiliative conversation, dyads were instructed to discuss popular media that both participants enjoyed. ${ }^{2}$ Experimenters identified the argumentative prompt for each dyad based on participants' responses to the opinion survey. The topic for which participants expressed

[^181]opposing views (e.g., one pro-life, the other pro-choice) and for which both participants indicated strong feelings was chosen. Up to two additional argumentative prompts were chosen using the same criteria and were given to participants if they could not continue the argumentative conversation on the first topic for the entire time.

## Materials

Movement data were collected automatically using a framedifferencing method (FDM; Paxton \& Dale, in press). Participants sat facing one another during their conversations and were captured in profile in a single frame on a high-definition camcorder (Canon VIXIA HF M31). ${ }^{3}$ The videos were downsampled at 8 Hz to a series of still frames. The FDM tracked movement by registering changes in pixels across frames (see Figure 1 for toy visualization) and applying a filter to remove extraneous pixel changes (e.g., due to fluctuations in light sources). For additional detail on the FDM, see Paxton and Dale (in press). See also Grammer, Honda, Jüette, and Schmitt (1999) for related methods.


Figure 1: Sample FDM sequence of interacting dyad, aggregated over multiple frames for visualization purposes.

Speech data were collected using individual lapel microphones (Audio-Technica ATR 3350) and a mixer (Azden CAM-3) so that each participant's audio was captured on a separate channel. The present research used on/off speech states as the measure for speech. On/off
speech states were obtained for each participant using the sound finder function in Audacity. Decibel cutoffs were individually determined for each dyad in order to maximize sensitivity and specificity.

Samples of data obtained from each interaction type are graphed in Figure 2. Each sample includes 250 sec of interaction. Taken from the same dyad, the figures graph changes in body movement as red and blue lines, with speech events depicted as boxes of corresponding color behind the lines.

## Results

The present analyses tested for unimodal and multimodal interpersonal and intrapersonal alignment. Crosscorrelations were calculated for interpersonal unimodal (e.g., participant A's movement to participant B's movement), interpersonal multimodal (e.g., participant A's speech to participant B's movement), and intrapersonal

[^182]

Figure 2: Sample body movement time series of a single dyad during 250s of interaction during an affiliative (left) and an argumentative (right) conversation. Speech data are represented as shaded boxes of corresponding colors.
multimodal (e.g., participant A's movement to participant A's speech) channels within a $+/-3000 \mathrm{~ms}$ range, yielding a series of cross-correlation
coefficients ( $r$ ). Using cross-correlation coefficients permitted us to investigate both in-phase synchrony and longer-phase alignment trends within the data.

The data were primarily analyzed using a series of linear mixed-effects models (Baayen, 2008), using dyad and condition as random effects unless otherwise noted. All main and interaction terms were standardized prior to being entered into the models. As standardized values, the crosscorrelation coefficients can be interpreted as beta weights, a measure of effect size (Keith, 2005).

## Unimodal Alignment

Movement Previous analyses of the corpus found evidence for in-phase bodily synchrony (Paxton \& Dale, under revision). ${ }^{4}$ To ensure that the subset analyzed here exhibited similar patterns, our first model predicted interpersonal body movement alignment ( $r_{\text {mov }}$ ) with conversation type (affiliative or argumentative) and time lag (125ms increments). Results confirmed that the subset of dyads conformed to broader patterns within the whole corpus. Increases in time lag (i.e., comparing movement further removed in time) significantly predicted a drop in $r_{\text {mov }}$ ( $\beta=-$ .25; $p<.0001$ ), providing evidence for in-phase interpersonal synchrony. Changes in $r_{\text {mov }}$ were also significantly predicted by conversation type ( $\beta=-.19 ; p<.0001$ ), with lower levels of movement synchrony in argumentative conversations. Interestingly, while only trending toward significance in analyses of the entire corpus, the interaction term between conversation type and time lag reached significance in this subset of the data ( $\beta=.14 ; p<.01$ ): Interlocutors' body movements were more tightly synchronized during affiliative conversations, reaching higher peak $r_{\text {mov }}$ and falling more sharply as time lag increased.

[^183]Speech The second model tested interpersonal speech alignment during different conversation contexts, using conversation type and time lag (125ms) to predict interpersonal speech alignment $\left(r_{\text {speech }}\right)$. As anticipated, increases in time lag predicted increases in $r_{\text {speech }}$ ( $\beta=.15$; $p<.0001$ ), while argumentative conversation type significantly predicted a decrease in $r_{\text {speech }}$ ( $\beta=-.44$; $p<.0001$ ). The interaction term was also significant ( $\beta=-.11$; $p<.001$ ). Together, these results suggest that interlocutors generally respected the turn-taking structure during all conversations but were more likely to exhibit overlapping speech during affiliative conversations.

To better situate these results, we performed complementary analyses comparing participants' speech patterns during different conversation types, accounting for condition, conversation number, speaker, and dyad membership as random effects. In a model predicting turn length with conversation type, argumentative conversations predicted slightly but significantly longer turn lengths compared with affiliative conversations ( $\beta=.04 ; p<.005$ ). Another model predicted total number of speech events in a conversation by both participants using conversation type and found that argumentative conversations had significantly fewer speech events than affiliative conversations ( $\beta=-.31 ; p<.0001$ ).

## Multimodal Alignment

Interpersonal Next, we predicted interpersonal multimodal cross-correlation coefficients ( $r_{\text {multi }}$ ) with conversation type and time lag. The main effect for conversation type was again significant, with argumentative conversations predicting a significant drop in $r_{\text {multi }}$ ( $\beta=-.21 ; p<.0001$ ). Neither time lag ( $\beta=-.02 ; p=.39$ ) nor the interaction term ( $\beta=.01 ; p=.76$ ) reached significance.

Intrapersonal Our final model predicted intrapersonal multimodal cross-correlation coefficients ( $r_{\text {self }}$ ) with conversation type and time lag. As predicted, we found no significant effect of conversation type on $r_{\text {self }}$ ( $\beta=.01$; $p=.75$ ), suggesting that intrapersonal alignment may be less
sensitive to conversation context than interpersonal alignment. Increases in time lag again significantly predicted decreases in $r_{\text {self }}$ ( $\beta=-.34 ; p<.0001$ ), suggesting that interlocutors were more likely to be moving when talking and vice-versa. As with unimodal body movement alignment, the significance of time lag as a main effect provided evidence for the existence of in-phase synchrony. The interaction term also reached significance ( $\beta=.09$; $p<.05$ ): Although participants exhibited in-phase multimodal intrapersonal synchrony, individuals' own body movements and speech events were more tightly connected during affiliative conversations.

## Network Visualizations of Interaction

To create the network visualizations of interpersonal interaction, we used body movement (M) and speech (S) time series data rather than cross-correlation coefficients. The networks were intended to capture relationships as they occur in time. We created two independent networks, one for affiliative interactions and the other for argumentative interactions (Figure 3). Connection strengths were presented as beta weights obtained through a series of linear mixedeffects models. ${ }^{5}$ All models used condition, conversation number, and dyad as random effects; the intrapersonal models ( $\mathrm{M}_{1}: \mathrm{S}_{1}$ and $\mathrm{M}_{2}: \mathrm{S}_{2}$ ) included participant as an additional random effect. Models used the nodes as predictors of other nodes, according to their connections (e.g., predicting $\mathrm{M}_{1}$ with $\mathrm{M}_{2}$ ).

## Discussion

While we often intuitively acknowledge that conversational contexts affect the course of an interaction, our results suggest that there are fundamental differences in interpersonal dynamics during different contexts. During conflict, interpersonal body movement synchrony diminishes. Interlocutors have a more rigid turn-taking structure with fewer and longer turns. Dyads use fewer instances of any overlapping speech, including events like laughter and verbal tracking. Interpersonal multimodal alignment-when one interlocutor is talking and the other is simultaneously moving-drops. Furthermore, in many of these cases, the effect size of conversation type on these measures is quite large, suggesting a very strong impact of context on these aspects of interaction.

On the other hand, some types of behavior exhibit relatively more stable properties across context. Interlocutors multimodally synchronize their own speech and movement, tending to move and speak at the same time regardless of conversational contexts. However, intrapersonal multimodal synchrony can still be affected by context through interaction effects. We believe this reinforces a view of interpersonal interaction as inherently

[^184]

Figure 3: Network visualizations for affiliative (top) and argumentative (bottom) interactions. Colors correspond to those used in Figure 2. Connection strengths are shown as beta weights obtained from a series of linear mixed-effects models and can be interpreted as effect sizes. All connections are significant ( $p<.001$ ).
context-dependent, although the effects may be quite small for some elements or in some contexts.

Our findings paint conversation as a highly complex interpersonal communication structure. While there are some relatively stable elements within it (e.g., intrapersonal multimodal alignment), other elements are very sensitive to conversational contexts (e.g., interpersonal bodily synchrony). While complementary behaviors align across interactions, argument as a conversation context appears to exhibit additional constraints on alignment patterns. Based on these exploratory analyses, interpersonal communicative structures appear to be self-organizing within the interaction and with strong regard to the overall context.

The corpus analyzed here provides a rich source of interaction data in multiple conversational contexts. We intend to continue to mine these data in order to better understand the nature of multimodal communication and interaction and to collect additional corpora on other conversation contexts. In doing so, we hope to more fully develop the interaction network presented here. Future directions will pursue the creation of more predictive models of interpersonal (e.g., Mehler, Lücking, \& Weiß, 2010) and intrapersonal (e.g., Tilsen, 2009) multimodal alignment that can shape additional experimental work as
models like the HKB (Haken et al., 1985) have shaped intrapersonal unimodal alignment.

## Acknowledgments

The authors would like to thank undergraduates J.P. Gonzales and Stephanie Frewen for their help in data collection and preparation, respectively. This research was supported partially by NSF grants BCS-0826825 and BCS0926670.

## References

Baayen, R. H. (2008). Analyzing linguistic data: A practical introduction to statistics using $R$. Cambridge: Cambridge University Press.
Beek, P. J., Peper, C. E., \& Daffertshofer, A. (2002). Modeling rhythmic interlimb coordination: Beyond the Haken-Kelso-Bunz model. Brain and Cognition, 48(1), 149-165.
Bernieri, F., \& Rosenthal, R. (1991). Interpersonal coordination: Behavior matching and interactional synchrony. In R. Feldman \& B. Rime (Eds.), Fundamentals of nonverbal behavior: Studies in emotion and social interaction. New York: Cambridge University Press.
Brennan, S. E., Galati, A., \& Kuhlen, A. K. (2010). Two minds, one dialog: Coordinating speaking and understanding. In B. H. Ross (Ed.), The psychology of learning. Burlington: Academic Press.
Chartrand, T. L., \& Bargh, J. A. (1999). The chameleon effect: The perception-behavior link and social interaction. Journal of Personality and Social Psychology, 76(6), 893-910.
Cleland, A. A., \& Pickering, M. J. (2003). The use of lexical and syntactic information in language production: Evidence from the priming of noun-phrase structure. Journal of Memory and Language, 49(2), 214-230.
Condon, W. S., \& Ogston, W. D. (1966). Sound film analysis of normal and pathological behavior patterns. Journal of Nervous and Mental Disease, 143(3), 1-10.
Delaherche, E., \& Chetouani, M. (2010). Multimodal coordination: Exploring relevant features and measures. Proceedings of the Second International Workshop on Social Signal Processing (pp. 47-52). ACM.
Garrod, S., \& Pickering, M. J. (2004). Why is conversation so easy? Trends in Cognitive Sciences, 8(1), 8-11.
Grammer, K., Honda, M., Jüette, A., \& Schmitt, A. (1999). Fuzziness of nonverbal courtship communication unblurred by motion energy detection. Journal of Personality and Social Psychology, 77(3), 487-508.
Haken, H., Kelso, J. S., \& Bunz, H. (1985). A theoretical model of phase transitions in human hand movements. Biological Cybernetics, 51(5), 347-356.
Keith, T. Z. (2005). Multiple regression and beyond. Boston: Pearson Education.
Kopp, S. (2010). Social resonance and embodied coordination in face-to-face conversation with artificial interlocutors. Speech Communication, 52(6), 587-597.

Kousidis, S., \& Dorran, D. (2009). Monitoring convergence of temporal features in spontaneous dialogue speech. Proceedings of the First Young Researchers Workshop on Speech Technology. Dublin, Ireland.
Krämer, N., Kopp, S., Becker-Asano, C., \& Sommer, N. (2013). Smile and the world will smile with you - The effects of a virtual agent's smile on users' evaluation and behavior. International Journal of Human-Computer Studies, 71(3), 335-349.
Louwerse, M. M., Dale, R., Bard, E. G., \& Jeuniaux, P. (2012). Behavior matching in multimodal communication is synchronized. Cognitive Science, 36(8), 1404-1426.
McFarland, D. H. (2001). Respiratory markers of conversational interaction. Journal of Speech, Language and Hearing Research, 44(1), 128-143.
Miles, L. K., Griffiths, J. L., Richardson, M. J., \& Macrae, C. N. (2010). Too late to coordinate: Contextual influences on behavioral synchrony. European Journal of Social Psychology, 40(1), 52-60.
Miles, L. K., Lumsden, J., Richardson, M. J., \& Macrae, C. N. (2011). Do birds of a feather move together? Group membership and behavioral synchrony. Experimental Brain Research, 211(3-4), 495-503.
Mehler, A., Lücking, A., \& Weiß, P. (2010). A network model of interpersonal alignment in dialog. Entropy, 12(6), 1440-1483.
Norris, S. (2004). Analyzing multimodal interaction: A methodological framework. New York: Routledge.
Paxton, A., \& Dale, R. (in press). Frame-differencing methods for measuring bodily synchrony in conversation. Behavior Research Methods.
Paxton, A., \& Dale, R. (under revision). Argument disrupts interpersonal alignment.
Reitter, D., Moore, J. D., \& Keller, F. (2006). Priming of syntactic rules in task-oriented dialogue and spontaneous conversation. Proceedings of the Conference of the Cognitive Science Society (pp. 685-690). Vancouver, Canada.
Riley, M. A., Richardson, M. J., Shockley, K., \& Ramenzoni, V. C. (2011). Interpersonal synergies. Frontiers in Psychology, 2, 38.
Richardson, M. J., Marsh, K. L., Isenhower, R. W., Goodman, J. R. L., \& Schmidt, R. C. (2007). Rocking together: Dynamics of intentional and unintentional interpersonal coordination. Human Movement Science, 26(6), 867-891.
Sebanz, N., Bekkering, H., \& Knoblich, G. (2006). Joint action: bodies and minds moving together. Trends in Cognitive Sciences, 10(2), 70-76.
Shockley, K., Richardson, D. C., \& Dale, R. (2009). Conversation and coordinative structures. Topics in Cognitive Science, 1(2), 305-319.
Tilsen, S. (2009). Multitimescale dynamical interactions between speech rhythm and gesture. Cognitive Science, 33(5), 839-879.

# Getting to the Point: The Influence of Communicative Intent on the Kinematics of Pointing Gestures 

David Peeters (david.peeters@mpi.nl)<br>Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands<br>International Max Planck Research School for Language Sciences, Nijmegen, The Netherlands<br>Radboud University, Donders Institute for Brain, Cognition, and Behaviour, Nijmegen, The Netherlands<br>Mingyuan Chu (mingyuan.chu@mpi.nl)<br>Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

Judith Holler (judith.holler@mpi.nl)
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
School of Psychological Sciences, University of Manchester, UK
Aslı Özyürek (asli.ozyurek@mpi.nl)
Radboud University, Center for Language Studies, Nijmegen, The Netherlands
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
Peter Hagoort (peter.hagoort@mpi.nl)
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
Radboud University, Donders Institute for Brain, Cognition, and Behaviour, Nijmegen, The Netherlands


#### Abstract

In everyday communication, people not only use speech but also hand gestures to convey information. One intriguing question in gesture research has been why gestures take the specific form they do. Previous research has identified the speaker-gesturer's communicative intent as one factor shaping the form of iconic gestures. Here we investigate whether communicative intent also shapes the form of pointing gestures. In an experimental setting, twenty-four participants produced pointing gestures identifying a referent for an addressee. The communicative intent of the speakergesturer was manipulated by varying the informativeness of the pointing gesture. A second independent variable was the presence or absence of concurrent speech. As a function of their communicative intent and irrespective of the presence of speech, participants varied the durations of the stroke and the post-stroke hold-phase of their gesture. These findings add to our understanding of how the communicative context influences the form that a gesture takes.


Keywords: Pointing Gesture; Communicative Intent; Gesture Production; Action Planning; Deixis.

## Introduction

In everyday communication, people not only use speech but also meaningful hand gestures to convey information. One of the most intriguing questions in gesture research has been why such gestures take the physical form they do (Bavelas et al., 2008; Gerwing \& Bavelas, 2004; Krauss, Chen, \& Gottesman, 2000). The main focus so far in answering this question has been on gestures iconic in nature, i.e., gestures that in form and manner of execution visually resemble the simultaneously expressed meaning of the linguistic part of an utterance (McNeill, 1985), such as moving up and down
one's hand when talking about a basketball game. Typically, such studies have varied aspects of the communicative context, such as the visibility of gestures or the knowledge speaker and listener mutually share. Amongst other things, these studies have shown that speakers design their gestures for particular recipients and produce more (e.g., Alibali, Heath, \& Myers, 2001; Bavelas et al., 2008) as well as larger and more precise gestures when communicative intentions are enhanced (e.g., Gerwing \& Bavelas, 2004; Holler \& Stevens, 2007). Thus, iconic co-speech gestures seem closely linked to the speaker's specific communicative intent, and the particular form an iconic gesture takes depends on the context-bound communicative relation between speaker and addressee (see Holler \& Wilkin, 2011).

In contrast, it is unclear how the form of pointing gestures changes as a function of the gesturer's communicative intent. Pointing is a foundational building block of human communication (Kita, 2003) and allows us to directly connect our communication to the material world that surrounds us (Clark, 2003). Pointing has received much attention in the literature from an ontogenetic viewpoint because of its role in paving the way for the acquisition of language (Butterworth, 2003; Carpenter, Nagell, \& Tomasello, 1998; Tomasello, Carpenter, \& Liszkowski, 2007), as well as from a phylogenetic viewpoint with respect to declarative pointing being a uniquely human form of communication in a natural environment (Call \& Tomasello, 1994; Kita, 2003; Tomasello et al., 2007). In contrast, the exact form parameters that people vary in the production of pointing gestures in human adult communication remain largely unknown. Therefore, the present study aims at contributing further to our
understanding of why adults’ index-finger pointing gestures take the particular physical form they do in a communicative context.

There are some preliminary indications that suggest a relation between the form of a pointing gesture and the speaker's communicative intent. Cleret de Langavant et al. (2011) had participants repeatedly point to objects on a table in front of them. Two addressees were always sitting next to the table. At the onset of a communicative block, the participant was instructed to verbally address one of the addressees before the block started and was instructed before each trial of that block to point at a specific object for that addressee, who named the object after observing the participant's gesture. At the onset of a non-communicative block, the participant addressed nobody and was instructed before each trial to point at a specific object without having an addressee (and hence did not receive feedback from an addressee). Compared to the latter, non-communicative condition, the former condition yielded pointing gestures that had a trajectory and endpoint distribution that were tilted away from the addressee, arguably because the addressee's perspective on the target object was taken into account in the form of the gesture.

Everyday pointing gestures generally occur in a context in which two interlocutors share a joint attentional frame in which one person directs the attention of another person towards a location, event, or referent in the perceptual environment (Tomasello et al., 2007). An important prerequisite for a successful referential pointing gesture is that two interlocutors come to perceptually attend to the same entity or location and are mutually aware of the fact that they are both attending to the same thing (Clark, 1996). Therefore, instead of comparing a "communicative" situation (including addressing a listener and receiving verbal feedback) to a "non-communicative" situation (without addressing a listener and verbal feedback), as in Cleret de Langavant et al. (2011), we here compare two situations that are both communicative and differ only in the communicative intent of the speaker-gesturer. As a proxy of the communicative intent of the speaker-gesturer, we manipulate the degree of informativeness of the pointing gesture as a first factor in our design.

A second factor manipulated here is the presence of speech as a second modality. Pointing gestures often come with concurrent deictic speech such as spatial demonstratives (e.g., "this" and "that" in English). In the production of referring expressions, speech and gesture are tightly interconnected (Kendon, 2004; McNeill, 1992) and can be used independently or simultaneously to single out a referent (e.g., Bangerter, 2004), in contrast with iconic gestures that canonically come with speech. In the current study we manipulate the presence of such a second modality (speech) and explore the yet unaddressed question of whether the mere presence of speech as a second modality influences the form parameters people exploit in producing pointing gestures for their addressee, and whether the
presence of speech interacts with our manipulation of communicative intent.

The current study looks at different subcomponents (or: parameters) of the pointing gesture. We focus on the planning duration of the gesture, the duration of the stroke and the post-stroke hold-phase, as well as the point in time at which the apex is reached after the visual presentation of a referent (Levelt, Richardson, \& La Heij, 1985), and the amount of distance travelled by the pointing finger. Finally we also look at whether the synchronization of speech and gesture changes as a function of communicative intent.

## Method

## Participants

Twenty-four right-handed native speakers of Dutch (12 female; mean age 20.6), studying at Radboud University Nijmegen, participated in the experiment. They were compensated with $20 €$.

## Experimental Design and Set-up

Participants were seated at a distance of 100 cm from a computer screen that was placed back-to-back with another computer screen (henceforth: the back screen). Stimuli were four white circles in a horizontal line on the top of the screen, mirroring four circles on the back screen. The circles could light up either blue or yellow. A second participant (a confederate; henceforth: the addressee) looked at the back screen and the participant's pointing gesture via a camera. Figure 1 shows the view of the addressee via the camera (converted to a grayscale image). On all trials, participants referred to the circle that lit up. The addressee noted on a paper form to which of the four circles the participant


Figure 1: The addressee's view of the back screen and the pointing participant during a non-informative trial.
referred on each trial. In order to make the deictic act informative in one case but non-informative in the other, the following set-up was used. In both conditions, via a camera, the addressee observed the pointing gesture of the participant, as well as the circles at the back screen providing the corresponding view of the four circles the participant was seeing. This way, the addressee saw which of the four circles the participant pointed at, but without seeing which circle lit up on the participant's side of the screen, a crucial aspect in our manipulation (see below).

We manipulated the informativeness of the gesture (informative versus non-informative) as well as the modality of the deictic act (gesture-only versus gesture + speech) in a $2 x 2$ within participants design. In the informative condition, a circle turned blue or yellow only on the participants' screen but not on the back screen. Therefore the participant's pointing gesture was the only source of information on which the addressee had to base his decision in selecting the circle referred to by the participant. In the non-informative condition, the corresponding circles would light up on both the participant's and the addressee's screen. Thus, the participant's pointing gesture was non-informative, because the addressee saw one of the circles light up on the back screen at the same moment as the participant saw the corresponding circle light up (i.e., even before the onset of the participant's pointing gesture).

The modality factor was manipulated by having participants use either one or two modalities in referring to the circles. In gesture only blocks (G-only), participants pointed to a circle when it turned blue or yellow without producing speech. In gesture + speech blocks $(\mathrm{G}+\mathrm{S})$ participants pointed to the circle and said either die blauwe cirkel ("that blue circle") or die gele cirkel ("that yellow circle"), depending on the color of the circle. Note that, because on every trial only one circle turned blue or yellow, the speech was never informative (neither in the informative nor the non-informative blocks). The rationale for this was that we were interested in the possible effect of the mere presence of speech as a second modality, independently from the informativeness of the deictic act that was manipulated separately in the gesture.

Each trial started with a fixation cross, displayed for 500 ms , followed by the presentation of four white circles. After a jittered period of $500-1000 \mathrm{~ms}$, one of the circles turned yellow or blue. At this point, the participant was allowed to release her finger from a button, pointed to the blue or yellow circle, and named the circle (in the G+S blocks). The experiment consisted of 16 blocks of 20 trials each. Every condition in the experiment was represented by four blocks. The order of presentation of blocks was counterbalanced across participants. In half of the trials a circle lit up yellow, in the other half it lit up blue. The idea behind this was to create a slightly more complex and varied utterances to enhance the ecological validity in this very strictly controlled environment. Each block of 20 trials consisted of ten circles lighting up yellow and ten lighting up blue,
equally distributed over the four circles and the four conditions throughout the experiment, in a randomized way.

## Procedure

At the arrival of the participant, the experimenter explained that a second participant (i.e., the confederate addressee) would perform a behavioral task on the basis of the participant's gesture. The experimenter showed the participant the computer and form to be used by the addressee and demonstrated that the participant could be seen on the computer screen via a camera. Also, it was explained and shown to the participant that the addressee could only see the arm movement of the participant and the computer screen that was at the back of the computer screen that the participant saw. The addressee could not see the head of the participant, to avoid the participant from conveying information via the head and face. In order to keep participants motivated, it was emphasized that they were in a joint activity with the addressee and that the success of this joint activity depended on how well they worked together. The participant was then seated in a comfortable chair in the dimly-lit experiment room. The height of the screen was adjusted to the height of the eyes of the participant. The button used by the participant was placed at the height of the participant's elbow, 23 cm in front of the participant calculated from the vertical axis corresponding to the position of the participant's eyes. Participants were instructed to always rest their finger on this button, except when making the pointing gesture, which allowed calculating the duration and onset of the pointing gesture. A sensor was placed on the participant's right index finger nail to allow for motion tracking of the pointing movements. Participants’ electroencephalogram (EEG) was recorded continuously throughout the experiment. These results will be reported elsewhere.

After montage of the motion tracking sensor the experimenter picked up the addressee. The addressee was shown the room in which the participant performed the task, greeted the participant, and was seated in a chair in front of a computer in a room adjacent to the participant's room. In order to familiarize the participant with the different conditions and the task, thirty-two test-items (eight per condition) preceded the main experiment as a practice set. Participants received specific instructions to point with or without speech before each block. In addition, before each block, the participant was instructed whether the addressee could also see the same circles light up at the back screen or not during that block. Participants were asked to only move their hand and arm when pointing. During the experiment, participants were allowed to have a short break after every fourth block. Before and during the experiment, the communication between experimenter and addressee was minimal and fully scripted, in order to be constant across participants. After the experiment, the addressee was thanked for participation and left the room. Participants were debriefed, financially compensated, and thanked for participation.

## Kinematic recording and analysis

Behavioral and kinematic data were acquired throughout the experiment using experimental software (Presentation, Neurobehavioral Systems, Inc) and a 60 Hz motion tracking system and DTrack2 tracking software (both Advanced Realtime Tracking, Weilheim, Germany). For each trial, the Gesture Initiation Time (i.e. the moment the participant's finger left the button calculated from the moment a circle lit up) was calculated. This measure thus reflected the time it took to plan the pointing gesture. In addition, we collected for each trial the Apex Time (i.e. the moment of the endpoint of the gesture calculated from the moment a circle lit up). The endpoint of the gesture was defined as the point in time where the pointing index finger was at least 7 cm from the button and only moved forward less than 2 mm for two consecutive samples. The Stroke Duration was defined as the interval between the onset of the gesture (i.e., The Gesture Initiation Time) and the moment the apex was reached (i.e., the Apex Time). The Incremental Distance travelled by the pointing index finger was calculated for the complete stroke (similar to Levelt et al., 1985). Further, the Velocity of the hand movement was calculated for each trial on the basis of the Apex Time and the Incremental Distance. The Hold Duration of the pointing gesture was calculated by subtracting the Apex Time from the Retraction Time (i.e., the moment the index finger moved back in the direction of the button for at least 2 mm in two consecutive samples). In the G + S blocks, the Speech Onset Time was calculated from the moment one of the circles lit up. The Synchronization Time was defined as the difference between Apex Time and Speech Onset Time.

## Results

Trials on which the Gesture Initiation Time was below 100 ms or above 2000 ms were considered errors and excluded from all analyses ( $0.7 \%$ of total dataset). In addition, trials containing hesitations or errors in the participant's speech were removed from further analyses ( $0.2 \%$ of all data). Separate analyses of variance were performed for each
dependent variable with Informativeness (Informative versus Non-informative) and Modality (Gesture-only or Gesture+Speech) as within-subject factors.

A first analysis was performed on the Gesture Initiation Time. This analysis did not yield any significant main or interaction effect. Next, we analyzed the Stroke Duration. This analysis yielded a significant main effect of Informativeness, $F(1,23)=10.97, p=.003, \eta_{\mathrm{p}}{ }^{2}=.32$. This effect denoted that the duration of the stroke was significantly longer in the Informative condition ( $M=837$ ms ) than in the Non-informative condition ( $M=823 \mathrm{~ms}$ ). No significant main effect of Modality was found. There was no significant interaction between the two factors. Also an analysis on the Apex Time showed a significant main effect of Informativeness, $F(1,23)=8.15, p=.009, \eta_{\mathrm{p}}{ }^{2}=$ .26. This effect denoted that the apex was reached significantly later in the Informative condition ( $M=1379$ ms ) than in the Non-informative condition ( $M=1359 \mathrm{~ms}$ ). No significant main effect of Modality was found. There was no significant interaction between the two factors.

A further analysis was performed on the Incremental Distance. No significant main or interaction effect was found. Because the same amount of distance was travelled across conditions, but the apex was reached later in the Informative condition than in the Non-informative condition, the velocity of the pointing gesture must have been lower in the Informative condition compared to the Non-informative condition. Indeed, an analysis on the mean Velocity yielded a significant main effect of Informativeness, $F(1,23)=5.75, p=.025, \eta_{\mathrm{p}}^{2}=.20$. The velocity of the pointing gesture was significantly lower in the Informative condition ( $M=38.2 \mathrm{~cm} / \mathrm{s}$ ) than in the Noninformative condition ( $M=38.7 \mathrm{~cm} / \mathrm{s}$ ). Again, no significant main effect of Modality or interaction between the two factors was found. Another analysis, performed on the Hold Duration, yielded a significant main effect of Informativeness, $F(1,23)=10.17, p=.004, \eta_{\mathrm{p}}{ }^{2}=.31$. The Hold Duration was significantly longer in the Informative condition ( $M=1235 \mathrm{~ms}$ ) compared to the Non-informative condition ( $M=1143 \mathrm{~ms}$ ). No significant main effect of

Table 1: Overview of the results per condition in the experiment. Duration in ms is displayed for the Gesture Initiation Time (GIT), Stroke Duration (Stroke), Apex Time (Apex), Hold Duration (Hold), Speech Onset Time (SOT), and Synchronization Time (Sync). Further, the Incremental Distance in cm (Dist) and Velocity in $\mathrm{cm} / \mathrm{s}$ (Velocity) are provided. The standard error of the mean is indicated between parentheses.

| Condition | GIT | Stroke | Apex | Dist | Velocity | Hold | SOT | Sync |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
|  |  |  |  |  |  |  |  |  |
| Informative |  |  |  |  |  |  |  |  |
| Gesture-only | $534(21)$ | $834(30)$ | $1368(42)$ | $51(1)$ | $38.5(1)$ | $1252(135)$ |  |  |
| Gesture + Speech | $550(22)$ | $840(27)$ | $1389(39)$ | $51(1)$ | $37.8(1)$ | $1219(121)$ | $1385(65)$ | $4(54)$ |
|  |  |  |  |  |  |  |  |  |
| Non-informative | $532(22)$ | $819(29)$ | $1351(41)$ | $51(1)$ | $39.0(1)$ | $1138(116)$ |  |  |
| Gesture-only | $841(24)$ | $826(27)$ | $1367(40)$ | $51(1)$ | $38.5(1)$ | $1149(106)$ | $1351(66)$ | $16(54)$ |
| Gesture + Speech |  |  |  |  |  |  |  |  |

Modality was found. There was no significant interaction between the two factors.

In the $\mathrm{G}+\mathrm{S}$ conditions, participants referred linguistically to the circle on the screen while pointing. An analysis on the Speech Onset Time with Informativeness as the only withinsubject factor revealed a significant main effect, $F(1,23)=$ 6.79, $p=.016, \eta_{\mathrm{p}}^{2}=.23$. This effect reflected that the speech onset on average took place significantly later in the Informative condition ( $M=1385 \mathrm{~ms}$ ) than in the Noninformative condition ( $M=1351 \mathrm{~ms}$ ). An analysis on the Synchronization Time did not show a significant main effect of Informativeness ( $p=.16$ ), indicating that the onset of the speech and the apex of the gesture were aligned similarly and independently from the informativeness of the gesture. Table 1 summarizes all results.

## Discussion

Research investigating the production of iconic gestures has found that the form of such gestures changes on the basis of the communicative intent of the speaker-gesturer. Importantly, here we show that also in the case of pointing gestures speaker-gesturers exploit different form parameters as a function of their communicative intent. First, the duration of the stroke of pointing gestures was longer in the informative condition, which led to a gesture with a lower velocity and delayed the moment at which the apex was reached. Presumably participants did this in order to be as precise as possible in pointing to a target, which could be achieved by pointing more slowly. An additional benefit would then be that the addressee would have more time to identify towards which referent the gesture was heading. Second, the post-stroke hold-phase of the gesture was maintained longer, presumably in order to assure that the addressee had enough time to identify which referent the speaker pointed to. The form parameters under investigation here were not affected by the presence of deictic speech. Nevertheless, the onset of speech was synchronized with the moment at which the pointing gesture reached its apex.

A previous study compared a communicative to a noncommunicative situation and found that people may modify the trajectory and endpoint location of their pointing gesture to single out a referent for their addressee (Cleret de Langavant et al., 2011). The current study takes this research a step further by comparing two situations that are both communicative and identical except for the communicative intent of the gesturer. Cleret de Langavant et al. (2011) did not find a difference in the duration of the pointing gesture when comparing their communicative to their non-communicative condition. Here we did find an effect of communicative intent on the duration of the stroke and the post-stroke hold-phase. Thus, in addition to varying the endpoint location and trajectory of a pointing gesture (as in Cleret de Langavant et al., 2011), people may also use the duration of different sub-components of the pointing gesture in order to communicate effectively.

Participants temporally aligned the onset of the deictic linguistic expression with the moment the pointing gesture
reached its apex, regardless of whether the gesture was informative or not. This finding is in line with previous studies showing such temporal alignment of pointing and speech (e.g., Levelt et al., 1985; McNeill, 1992) and with models of speech and gesture production that underline the synchronization of speech and gesture (e.g., De Ruiter, 2000; Krauss et al., 2000). Here we show that this temporal synchrony between deictic speech and gesture is maintained irrespective of the speaker-gesturer's communicative intent.

We found a similar effect of communicative intent in situations where people only used gesture to communicate, compared to situations where speech and gesture were concurrently produced (Clark, 1996; Kendon, 2004). However, in our study, speech was purposefully never informative and very similar across trials, and there is indeed evidence that deictic speech can interact with the form of a simultaneously produced gesture (e.g., Gonseth, Vilain, \& Vilain, 2012). It is therefore possible that whenever speech itself is informative enough to single out a referent, speaker-gesturers no longer design their concurrent gesture to be maximally informative. Future research needs to shed more light on the influence of speech-gesture interaction on the form of deictic gesture and speech while manipulating the informativeness of the speech.

In general, the results of our study fit well with models of speech and gesture production that allow for a role of the speaker-gesturer's communicative intent in modulating the exact form of a gesture, such as the Sketch model (De Ruiter, 2000) and the Interface model (Kita \& Özyürek, 2003). Conversely, our data would argue against models of speech and gesture production that question whether the speaker’s communicative intent plays a role in determining the form of a gesture (e.g., Krauss et al., 2000). In our study, participants had the communicative intention of producing a pointing gesture towards a referent, either accompanied with referential speech or not. The Sketch model, which explicitly describes the production of pointing (in addition to other types of gesture), underlines that upon the intention to produce a pointing gesture, conventions such as which hand shape and finger to use can be retrieved from a knowledge store (called a "gestuary" by De Ruiter, 2000) in memory. This representation of the pointing gesture in the gestuary is only a template or abstract motor program, and there are a number of degrees of freedom that can be varied depending on the context in which the pointing gesture is performed. According to this model, in our study, participants retrieved a pointing gesture template from memory and subsequently exploited the duration of both the stroke (and as such the velocity and the moment the apex was reached) and the post-stroke hold-phase of the gesture as free parameters. Our study thus suggests that duration is a free parameter that people use to vary the execution of their pointing gesture, and further specifies in which specific components of the gesture duration is indeed varied.

The form a pointing gesture takes not only depends on the gesturer's communicative intent. Research has shown that it also depends on physical factors such as the spatial location
of a referent. For instance, people may raise their pointing arm and hand higher when a referent is more distant (Wilkins, 2003). Furthermore, the form of a gesture depends on cultural factors. In different cultures, different body parts are used for pointing (Kita, 2003; Wilkins, 2003). Finally, it may depend on socio-pragmatic factors. In a corpus study on Lao speakers, Enfield, Kita, and De Ruiter (2007) observed a distinction between relatively big points in which the whole arm is outstretched and relatively small points in which the hand is the main articulator. They argue that this difference in form is related to the pragmatic function of the utterance a gesture occurs in. Big points would do the primary work of an utterance, such as pointing out the location of an object, whereas small points would occur in utterances in which speech is central, adding a background modifier on the basis of social factors such as the common ground between interlocutors (p. 1738). Future studies could investigate interactions between such different physical, cultural, socio-pragmatic, and communicative factors.

To conclude, our study showed that people exploit the duration of the stroke (and as such its velocity and the moment the apex is reached) and the post-stroke hold-phase of their pointing gesture to communicate effectively. Thus, the form of a pointing gesture varies as a function of the speaker-gesturer's communicative intent. Similarly to iconic gestures, the form of pointing gestures is dependent, among other factors, on the context-bound communicative relation between speaker-gesturer and addressee.

## Acknowledgments

We thank Albert Russel for technical assistance. JH was supported through Marie Curie Fellowship \#255569 as well as through ERC Advanced Grant \#269484 INTERACT.

## References

Alibali, M. W., Heath, D. C., \& Myers, H. J. (2001). Effects of visibility between speaker and listener on gesture production: Some gestures are meant to be seen. Journal of Memory and Language, 44, 169-188.
Bangerter, A. (2004). Using pointing and describing to achieve joint focus of attention in dialogue. Psychological Science, 15, 415-419.
Bavelas, J., Gerwing, J., Sutton, C., \& Prevost, D. (2008). Gesturing on the telephone: Independent effects of dialogue and visibility. Journal of Memory and Language, 58, 495-520.
Butterworth, G. (2003). Pointing is the royal road to language for babies. In S. Kita (Ed.), Pointing. Where language, culture, and cognition meet. Mahwah, NJ: Lawrence Erlbaum.
Call, J., \& Tomasello, M. (1994). Production and comprehension of referential pointing by orangutans (Pongo pygmaeus). Journal of Comparative Psychology, 108, 307-317.
Carpenter, M., Nagell, K., \& Tomasello, M. (1998). Social cognition, joint attention, and communicative
competence from 9 to 15 months of age. Monographs of the Society for Research in Child Development, 255, Vol. 63, 1-174.
Clark, H. H. (1996). Using Language. Cambridge: Cambridge University Press.
Cleret de Langavant, L., Remy, P., Trinkler, I., McIntyre, J., Dupoux, E., Berthoz, A., \& Bachoud-Lévi, A.-C. (2011). Behavioral and neural correlates of communication via pointing. PLoS ONE, 6:3, e17719.
De Ruiter, J. P. (2000). The production of gesture and speech. In D. McNeill (Ed.), Language and gesture. Cambridge: Cambridge University Press.
Enfield, N. J., Kita, S., \& De Ruiter, J. P. (2007). Primary and secondary pragmatic functions of pointing gestures. Journal of Pragmatics, 39, 1722-1741.
Gerwing, J. \& Bavelas, J. (2004). Linguistic influences on gesture's form. Gesture, 4, 157-195.
Gonseth, C., Vilain, A., \& Vilain, C. (in press). An experimental study of speech/gesture interactions and distance encoding. Speech communication.
Holler, J., \& Stevens, R. (2007). An experimental investigation into the effect of common ground on how speakers use gesture and speech to represent size information in referential communication. Journal of Language and Social Psychology, 26, 4-27.
Holler, J., \& Wilkin, K. (2011). An experimental
investigation of how addressee feedback affects cospeech gestures accompanying speakers' responses. Journal of Pragmatics, 43, 3522-3536.
Kendon, A. (2004). Gesture: Visible action as utterance. Cambridge: Cambridge University Press.
Kita, S. (2003). Pointing. Where language, culture, and cognition meet. Mahwah, NJ: Lawrence Erlbaum.
Kita, S., \& Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. Journal of Memory and Language, 48, 16-32.
Krauss, R. M., Chen, Y., \& Gottesman, R. F. (2000). Lexical gestures and lexical access: A process model. In D. McNeill (Ed.), Language and gesture. Cambridge: Cambridge University Press.
Levelt, W. J. M., Richardson, G., \& La Heij, W. (1985). Pointing and voicing in deictic expressions. Journal of Memory and Language, 24, 133-164.
McNeill, D. (1985). So you think gestures are nonverbal? Psychological Review, 92, 350-371.
McNeill, D. (1992). Hand and Mind: What gestures reveal about thought. Chicago: University of Chicago Press.
Tomasello, M., Carpenter, M., \& Liszkowski, U. (2007). A new look at infant pointing. Child Development, 78, 705722.

Wilkins, D. (2003). Why pointing with the index finger is not a universal (in sociocultural and semiotic terms). In S. Kita (Ed.), Pointing. Where language, culture, and cognition meet. Mahwah, NJ: Lawrence Erlbaum.

# Comprehending with the body: Action compatibility in sign language? 

Pamela Perniss (p.perniss@ucl.ac.uk), David Vinson (d.vinson@ucl.ac.uk),<br>Neil Fox (neil.fox@ucl.ac.uk), Gabriella Vigliocco (g.vigliocco@ucl.ac.uk)<br>Department of Cognitive, Perceptual and Brain Sciences, 26 Bedford Way, WC1H 0AP London, UK


#### Abstract

Previous studies show that reading sentences about actions leads to specific motor activity associated with actually performing those actions. We investigate how sign language input may modulate motor activation, using British Sign Language (BSL) materials, some of which explicitly encode direction of motion, vs. written English, where motion is only implied. We find no evidence of action simulation in BSL comprehension, but replicate effects of action simulation in comprehension of written English. The results suggest that the perception of motor articulation in the language input interferes with mental simulation involving the motor system.


Keywords: embodiment; sign language; motor system; action-compatibility effect

## Introduction

There is now a body of evidence for an embodied view of language, according to which language comprehension is based in our bodily experience of the world and involves the same systems necessary for bodily experience. The grounding of language in perception and action has been evidenced in a wide range of behavioral and neuroscientific studies (e.g. Barsalou, Barbey, Simmons \& Wilson, 2003; Barsalou, 2008; Beauchamp \& Martin, 2007; Gallese \& Lakoff, 2005; Glenberg \& Kaschak, 2002; Glenberg \& Gallese, 2012; Zwaan \& Kaschak, 2008). For example, Stanfield and Zwaan (2001) have shown that sentence comprehension involves the activation of specific imagery related to the perceptual and action properties of an event. Participants were faster to verify that a pictured object (e.g. nail) appeared in a preceding sentence when the orientation of the object (e.g. horizontal vs. vertical) in the picture matched that implied in the sentence (e.g. John hammered the nail into the wall). Neuroscientific studies have likewise pointed to specific involvement of motor areas in understanding language related to action. For example, as found by Tettamanti, Buccino, Saccuman, Gallese et al. (2005) in an fMRI study, reading sentences describing actions using specific body parts (e.g. I bite the apple, I kick the ball) activates the area in the motor cortex related to physical use of that body part (e.g. mouth for bite, foot for kick) (see also Hauk, Johnsrude \& Pulvermüller, 2004).

The Action-Sentence Compatibility Effect (ACE), first demonstrated by Glenberg and Kaschak (2002), provides further compelling evidence that we involve our sensorimotor system in language comprehension by mentally simulating the actions and events encoded in language. In the Glenberg and Kaschak (2002) study, participants were presented visually with written sentences that implied either motion toward (Andy delivered the pizza to you) or away
from (You delivered the pizza to Andy) the body. Participants were asked to judge sentence sensibility by responding with a button press that required movement of the arm either toward or away from the body - i.e. in a direction congruent or incongruent with the direction of motion implied by the sentence. Participants were faster to respond to sentences when the implied motion was congruent with the response direction. This was true of sentences implying transfer in both the concrete and abstract domains (e.g. You communicated the message to Adam). In a related study, Borreggine and Kaschak (2006) provided evidence for the ACE when the same English sentences were presented to participants auditorily, unfolding in real time. The Action-Sentence Compatibility Effect suggests that sentence comprehension involves a dynamic mental simulation of the event, in this case, a motor simulation of performing the described action.

As the evidence supporting sensori-motor system involvement in language comprehension accumulates, we must also address the question of how this embodiment comes about. How does language come to be grounded in our bodily experience and what are the mechanisms by which language processing engages the sensori-motor system (cf. Perniss, Thompson \& Vigliocco, 2010)? Moreover, there is much debate about how embodiment effects may be modulated by context (cf. Willems \& Casasanto, 2011), and how effects may be constrained by different properties of language.

In this context, the strong role of action/motor simulation in sentence comprehension demonstrated by the ACE effect raises an interesting question with respect to the modality of language presentation. Neither the written visual nor the spoken audial presentation of sentences involves the physical use of the motor system. This situation is very different, however, in the case of sign languages. In sign languages, the natural languages of deaf people, meaning is encoded through movement of the hands and arms through the space on and in front of the body. The visual medium of sign language moreover affords a high degree of iconicity, or resemblance between linguistic form and meaning. This potential is exploited particularly for encoding sensorimotor information, such that meanings related to action and motion are expressed in highly iconic linguistic forms. Thus, many sign language verbs encoding transfer of the type studied by Glenberg and Kaschak (2002) explicitly realize directionality of motion in the event through a corresponding movement of the hands through space (i.e. toward or away from the body).

To date, embodiment effects have been studied looking at spoken/written language. Extending the investigation of
embodiment to language expressed in the visual modality, where the same motor articulators that perform nonlinguistic actions are used to encode actions linguistically, is an important step to understanding the nature of embodiment, and the conditions under which embodiment effects come about. The simulation effects observed in action sentence comprehension may well be modulated or constrained by inherent properties of language, particularly related to language modality and the potential for iconic representation. To address this question, we ask here how the iconic and motor properties of visual language may affect action simulation in sign language sentence comprehension?

One possible outcome is that we find an effect of action simulation in sentence comprehension consistent with what has been shown for English. The perception of motor action in language, where the linguistic (i.e. phonological) expression of action verbs is realized through directional motion, could boost the involvement of motor simulation in sentence comprehension. In this case, participants should be faster to respond in a direction corresponding to the event (and its realization in sign language), compared to when the response direction is incongruent with the event. It is also possible, however, that we observe an effect in the opposite direction, as the interpretation of the signed sentences may involve mentally taking the perspective of the signer producing them, as we discuss below.

An alternative outcome is that motor simulation of the encoded event in comprehension may be reduced or eliminated by the involvement of the motor system in the articulation of the action. Because viewing sign language means viewing physical movement of the same articulators necessary for the action simulation (here, arms and hands), the simulation itself may be blocked or even unnecessary. Thus, a contrasting prediction is that viewing sentences presented in a sign language would not yield an ACE effect.

To test these contrasting predictions, we investigate ACE effects in deaf users of British Sign Language (BSL).

## Experiment 1

We sought to replicate the Glenberg and Kaschak (2002) study using close translations of the original English sentences into BSL. All of the experimental sentences in this study implied directional motion, with equal numbers of sentences referring to motion toward the body and motion away from the body. Within these, half referred to concrete transfer events (as in You delivered the pizza to Andy) and half to abstract transfer events (as in You communicated the message to Adam).

The BSL sentences not only preserved the specific events described by the English sentences but also their $2^{\text {nd }} / 3^{\text {rd }}$ person reference structure. In BSL, as in other sign languages, person reference is achieved by directing signs at locations in space associated with the entities being talked about. Second person (you) is associated with a location directly opposite the signer, the canonical location of an
addressee. ${ }^{1}$ Third person (he/she/it) is associated with a location to the right or left of the signer. The body of the signer, specifically a location at the center of the signer's chest, is associated with first person ( $I$ ). Points to the appropriate locations indicate the arguments, e.g. subject and object, of a verb. In addition, verb signs themselves -so-called directional predicates - can indicate arguments by physically moving between the locations associated with the arguments. ${ }^{2}$

This is illustrated in the BSL example sentence shown in Figure 1, which corresponds to English James awarded the degree to you. In the example, $3^{\text {rd }}$ person reference to James is achieved in stills 2-3 of the figure, consisting of a sign for the letter ' J ' (for James), in still 2, followed by a pointing sign to a $3^{\text {rd }}$ person location to the right of the signer's body, in still 3. The predicate in the final still conveys the meaning he awards to you by moving from the $3^{\text {rd }}$ person location associated with James to the $2^{\text {nd }}$ person location - outward from and opposite the signer's body - associated with the participant/addressee viewing the sentence. Thus, in the BSL version of James awarded the degree to you, participants see the predicate move toward them, in the same way as the actual event would involve movement toward them. In the experiment, participants perceived directional verbs like award-to moving either toward or away from them, congruent with the direction of the event.


Figure 1: Glossed example of BSL sentence. (English translation: James awarded the degree to you.) The sentence includes a directional verb that moves from $3^{\text {rd }}$ to $2^{\text {nd }}$ person. The perceived motion in viewing the sentence is thus toward the participant's/addressee's body.

Not all predicates in BSL are directional, however, in this way. Translation of the set of verbs that appears in the original Glenberg and Kaschak (2002) study ignores a crucial contrast between directional vs. non-directional verb types in BSL. In non-directional predicates, the form of the verbs is the same regardless of the direction of the event (e.g. in the verb write, the hands represent writing at a location in front of the body regardless of who is writing to whom). Only directional predicates encode the direction of the event by moving between the locations associated with their arguments. (See the schematizations of verb types and direction used across experiments provided in Figure 2.)

[^185]A second set of sentences was designed to address this issue. This set encoded semantically similar events from those of Glenberg and Kaschak (2002), but using different verbs, in order to manipulate the number of directional vs. non-directional verbs that appeared in the sentences. This is important because the ACE effect hinges on the simulation of directional motion, and may be influenced here by the congruence of perceived phonological motion with the motion entailed by the event itself.


Figure 2. Schematization of verb types, person reference, and direction used across Experiments 1 and 2.

## Method

Participants 16 deaf adult signers of BSL were recruited from the greater London area. BSL age of acquisition ranged from $0-13$ years (mean 3.13; with 9 native signers who acquired BSL from birth). Participant age ranged between 19-59 years (mean age 34.69). All participants had normal or corrected to normal vision.

Materials For Experiment 1a, the original English sentences (Glenberg \& Kaschak, 2002) were translated into BSL by a native deaf BSL signer, and proficient BSLEnglish bilingual. BSL sentences were videotaped and edited into single sentence clips. All sentences depicted transfer from $2^{\text {nd }}$ to $3^{\text {rd }}$ or from $3^{\text {rd }}$ to $2^{\text {nd }}$ person, corresponding with direction of motion toward or away from the body, respectively (see Figure 1). 20 abstract and 20 concrete events were included, with two sentences depicting each event (one toward the body, one away). 40 nonsense sentences were also filmed, again closely resembling those used by Glenberg and Kaschak (2002). Different test lists were created so that each participant saw only one sentence referring to a given event, with equal numbers of abstract/concrete, toward/away sentences. Sentences were randomly ordered for each participant.

For Experiment 1b, we constructed BSL sentences around events involving 16 directional verbs and 13 non-directional verbs. ${ }^{3}$ For each verb we created four sentences, two sensible sentences (one toward the body, one away from the

[^186]body) and two nonsense sentences (one toward, one away). Each participant saw all four sentences involving a given verb, with materials divided into four blocks so that each verb occurred only once per block and so that conditions were approximately balanced within each block. Order of blocks and order of trials within a block were randomly ordered for each participant. In both experiments, we treated nonsense sentences as fillers, only analyzing the effects of implied directional motion in sentences depicting real events.

Procedure Participants sat directly opposite a computer screen with a response box oriented sagittally in front of them, and were told they would see BSL sentences addressed to them. Participants were prompted to press and hold the middle of three buttons on the response box upon the appearance of a fixation cross in the middle of the screen. Upon pressing the button, a video clip of a BSL sentence began to play, and continued to play as long as the middle button was held down. Participants judged the sensibility of the sentence by moving their finger to press a button either away from or toward their body from the middle button (i.e. to the near or far button on the response box). Participants were told to respond as quickly and accurately as possible. We measured the time it took for participants to release the central button, thus tapping into the motor planning necessary to make their responses (see Borreggine \& Kaschak, 2006). Participants came for two sessions, which differed only in the direction of the response for sensible sentences (toward vs. away from the body), with the order of response per session counterbalanced across participants. Experiments 1 a and 1 b were carried out separately in each session, again with the order counterbalanced across participants.

## Results

We analyzed only responses for sensible sentences, excluding errors and using button release latencies as the dependent measure. We analyzed Experiment 1a using $2 \times 2$ ANOVA (sentence direction $\times$ response congruence). The main effect of sentence direction was not significant; $F(1,15)=3.260, p=.091$ : a tendency for faster responses for transfer toward the body. Neither the main effect of response congruence and interaction were significant ( $F<1$ ).
For Experiment 1 b we conducted $2 \times 2 \times 2$ ANOVA also including the factor of verb type (directional vs. nondirectional). There was a main effect of sentence direction $\left(F(1,15)=6.772, \quad p=.020, \quad \eta_{\text {partial }}^{2}=0.311\right)$; the sentences moving toward the body were faster than those moving away from the body. There was also a main effect of verb type $\left(F(1,15)=124.933, p<.0001, \eta_{\text {partial }}^{2}=0.891\right)$; sentences with non-directional verbs were much faster than sentences with directional verbs (2091 vs. 2271 msec respectively), likely due to differences in sentence durations (e.g. directional verbs may take longer on average to execute than non-directional verbs). There was also an interaction between verb type and sentence direction $(F(1,15)=10.442$,
$p=.006, \eta_{\text {partial }}^{2}=0.410$ ), again presumably related to differences in verb durations. Crucially, the main effect of congruence was not significant $(F<1)$, nor were any of the interactions involving congruence (congruence $\times$ direction $F(1,15)=1.453, p=.247$; congruence $\times$ verb type, $F<1$; 3way interaction $F(1,15)=3.702, p=.074)$. See Figure 3.

We found no ACE effect: there was no main effect of response congruence, nor any interactions involving it. ${ }^{4}$


Figure 3: Results of Experiment 1a (left: BSL replication of Glenberg \& Kaschak, 2002) and 1 b (right: BSL $2^{\text {nd }} / 3^{\text {rd }}$ person comparing directional/non-directional verbs). We report correct button release times to sensible sentences as a function of sentence direction (Away (from) or Toward the body) and whether the response direction is congruent or incongruent with the directional event being described. Error bars reflect standard error of the mean (by subjects).

## Discussion

In Experiments 1a and 1 b we did not find an ACE effect; responses were not faster when the sentence implied an event moving in the same direction as the hand action required to make a sensibility decision. Furthermore, in Experiment 1 b there was no difference in response times between when the direction of motion entailed by the action was encoded phonologically (directional verbs) vs. when it was not (non-directional verb). This implies that the lack of action simulation - and thus lack of motor system involvement - may be related to perceiving sign language, which is produced by means of motor movement of the same articulators involved in the actual action event.

However, it may be the case that ACE requires that the direction of the button press converge with the direction of the event. In the $2{ }^{\text {nd }} / 3^{\text {rd }}$ person transfer used in Experiments 1 a and 1 b , the BSL directional verbs move diagonally, offset approximately $45^{\circ}$ from the center of the producer's body and offset approximately $45^{\circ}$ from the participant's direction of response (see Figure 2). If sentence compatibility effects require close directional convergence between the sentence judgment response and the simulated event, this discrepancy could reduce or eliminate ACE effects. Finally, if the lack of an effect in Experiments 1a and 1 b is due to the use of sign language in the task, we should be able to observe an ACE effect in the same

[^187]participants when they are reading English sentences. Obtaining an ACE effect using English is especially important in the face of the null effects we have reported here, showing that our procedure is sound and our participant number large enough to find evidence for ACE, if it were there.

## Experiment 2

Experiment 2a assesses whether the lack of convergence between the direction of motion encoded in the event and the response direction modulates simulation effects. We use BSL sentences that imply transfer between $1^{\text {st }} / 2^{\text {nd }}$ person, thereby maximizing the overlap in directionality. Sentences that encode transfer between $1^{\text {st }} / 2^{\text {nd }}$ person (e.g. I awarded the degree to you) involve phonological movement between the signer's body ( $1^{\text {st }}$ person) and a location opposite the signer's body ( $2^{\text {nd }}$ person). Thus, directional verbs move along the central axis, straight toward or away from the body. This modification of person reference in verbs creates complete directional convergence with the direction of button press response and with the direction of motion entailed by the actual action event. We otherwise used exactly the same materials and procedures as in Experiment 1b, where directional vs. non-directional verbs are compared.

Experiment 2 b assesses whether the lack of an effect in Experiment 1 is indeed specific to the use of sign language. The same (BSL-English bilingual) participants who took part in Experiment 2a, also carried out the original experiment by Glenberg and Kaschak (2002) with visual presentation of written English sentences.

## Method

Participants A new group of 16 deaf adult BSL-English bilinguals ${ }^{5}$ were recruited from the greater London area. BSL age of acquisition ranged from 0-11 years (mean 3.85; with 7 native signers who acquired BSL from birth). English age of acquisition ranged from 0-5 years (mean 2.19). Ages ranged between 18-59 years (mean age 30.75). All participants had normal or corrected to normal vision.

Materials BSL materials for Experiment 2a were the same as for Experiment 1 b , but all sentences depicted transfer from $1^{\text {st }}$ to $2^{\text {nd }}$ or from $2^{\text {nd }}$ to $1^{\text {st }}$ person. For the written English experiment, we used the original set of sentences from Glenberg and Kaschak (2002). List creation, task order etc. were the same as in Experiment 1.

Procedure The procedure was the same as in Experiment 1. For English sentences (Experiment 2b), participants saw the written English sentence appear in the middle of the screen.

[^188]
## Results

As in Experiment 1 we analyzed only the responses for sensible sentences, excluding error trials and using button release latencies as our dependent measures. We analyzed Experiment 2 a using $2 \times 2 \times 2$ ANOVA also including the factor of verb type (sentence direction $\times$ response congruence $\times$ verb type). None of the main effects reached significance (congruence $F<1$; direction $F(1,15)=2.299$, $p=.150$; verb type $F(1,15)=2.681, p=.122$ ), nor did any of the interactions (congruence $\times$ verb type, $F(1,15)=1.246$; $p=.282$; three-way interaction $F(1,15)=3.082, p=.1001$; all other $F<1$ ). As in Experiments 1a and 1b we found no ACE effect in BSL. ${ }^{6}$
We analyzed Experiment 2 b using $2 \times 2$ ANOVA. Here we found a main effect of congruence $(F(1,15)=10.888, p=.005$, $\left.\eta_{\text {partial }}^{2}=0.421\right)$ : an ACE effect in written English. There was a marginal main effect of sentence direction $(F(1,15)=4.353$, $p=.054$ ) and no interaction $(F<1)$. See Figure 4.


Figure 4: Results of Experiment 2a (left: BSL $1^{\text {st }} / 2^{\text {nd }}$ person comparing directional/non-directional verbs) and $2 b$ (right: English replication of Glenberg \& Kaschak, 2002 with BSL signers). We report correct button release times to sensible sentences as a function of sentence direction (Away (from) or Toward the body) and whether the response direction is congruent or incongruent with the directional event being described. Error bars reflect standard error of the mean (by subjects).

## Discussion

As in Experiment 1a and 1b, we find no ACE effect in Experiment 2a: responses to BSL sentences are no faster when the direction of response is congruent with the event depicted by a sentence. But we find reliable ACE effects in English (Experiment 2b), replicating the original Glenberg and Kaschak (2002) study with exactly the same participants as in Experiment 2a. Finding an ACE effect in English vs. no such effect in BSL, in the same population, suggests that action simulation is involved in language comprehension when visually perceiving written language, but not when perceiving signed language.

[^189]
## General Discussion

We assessed whether the same effects of action simulation observed during comprehension of English directional sentences can be observed in the comprehension of BSL directional sentences. The ACE effect has been argued to demonstrate that sentence comprehension relies on simulation of the actions encoded in the sentences. Specifically, the ACE effect shows that sentence comprehension is facilitated when the action implied by a sentence is directionally congruent with the action required to judge sentence sensibility.

Operating in the visual-spatial modality, sign language necessarily involves motor movement and utilizes the high potential for action iconicity that the medium affords. These properties of sign language led us to propose two possible outcomes regarding the role of action simulation in sign language comprehension: the action simulation is either boosted or blocked.

We found no evidence for an ACE effect in BSL sentence comprehension across three experiments ( $1 \mathrm{a}, 1 \mathrm{~b}, 2 \mathrm{a}$ ). The results thus suggest that viewing sign language does not engage the motor system in comprehension in the way that has been found for written and spoken presentation of English sentences. These results do not come about because of lack of power: we observed an ACE effect with the same participants when presented with English written sentences. This finding also indicates that it is not knowing a sign language per se that modulates the use of action simulation in sentence comprehension (i.e. in a second language).

The results further suggest that it is not the iconicity between the direction of motion of the action signs and the actual actions that blocks the involvement of action simulation in comprehension, as there was no difference found between directional vs. non-directional verbs.

Why then do we fail to see an ACE effect in BSL? A first possibility is that the lack of involvement of the motor system in comprehension may be related to perceiving the physical engagement of the motor articulators. This engagement would block the system from engaging in sentence simulation. It is also possible that the involvement of our sensori-motor systems in language comprehension depends on the format in which language is presented. The ACE effect has been found previously, and replicated here, only for language presented in a unichannel format - the written or the spoken word. These unichannel formats are not directly evocative of the events encoded in the sentences, they have no explicit visual correspondence to the events being described. They are impoverished in this sense compared to the rich, depictive event representations provided by the visual modality of signed language. Thus, it may be that an "impoverished" unichannel language representation relies on action simulation in comprehension, while a richer, multichannel language presentation - particularly involving depictive, iconic representation - does not. The action may not need to be "filled in" or simulated in the context of a rich, depictive representation of the event.

We cannot rule out, however, that the iconic properties of sign language action predicates play a role in affecting the involvement of the motor system. Even the non-directional verbs, which do not overtly encode the direction of motion of the action, were often iconic of the action in some way (e.g. BSL pour which resembles pouring a liquid, but does not vary in its direction depending on who is doing the pouring). Such iconic properties may engage the same effectors in simulation, perhaps with other aspects of the event such as hand configuration and orientation being more salient than generic aspects of directional motion. Our results also do not rule out the possibility that the perception of motor movement at all, and particularly non-iconic movement, may block the involvement of the motor system in comprehension.

Finally, there are a number of further modality-related differences between English and BSL that might have played a role in our study. First, the temporal unfolding of the event is different. Secondly, word order differences may play a role. While the English sentences follow a rigid Subject-Verb-Object order, the BSL sentences typically had the verb in final position (a word order common to many sign languages). Future research on ACE effects in spoken languages with verb-final word order, e.g. Japanese or Turkish, would be illuminating in this regard.

Another issue that bears further investigation relates to perspective-taking in sign language comprehension. Specifically, though participants were informed that they would see BSL sentences addressed to them, sentence comprehension may have involved mentally taking the signer's perspective. If participants mapped their own body onto that of the sign model producing the sentences, mentally imitating the sign model's motor production, this would create a conflict between the congruence between sentence direction and response direction. However, if this were the case, it is likely that we would have seen an ACE effect in the opposite direction. As no effect whatsoever was observed in the BSL versions of the experiments, we do not assume this to have played a role.

Thus, our research suggests that the involvement of action simulation in language comprehension is dependent on the format and modality of language presentation. This is important to our understanding of the conditions under which and the degree to which language comprehension involves simulation. The idea that the use of action simulation may be contextually dependent is in line with previous observations that contextual variables (e.g. abstract vs. concrete contexts) modulate effects of embodiment in terms of differential activation of sensori-motor representation in language processing (Mahon \& Carramazza, 2007; Willems \& Casasanto, 2011 for a review). Context dependency of the degree to which embodiment (i.e. the involvement of sensori-motor systems) is evident in language comprehension demonstrates a fundamental flexibility, rather than rigidity, of the architecture of language processing.

## Acknowledgments

Supported by UK Economic and Social Research Council grants RES-062-23-2012 to Gabriella Vigliocco and RES-620-28-6002 to the Deafness, Cognition and Language Research Centre (DCAL).

## References

Barsalou, L. W. (2008). Grounded cognition. Annual Review of Psychology, 59, 617-645.
Barsalou, L. W., Simmons, W. K., Barbey, A. K., \& Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. Trends in Cognitive Science, 7, 287-317.
Beauchamp, M. S., \& Martin, A. (2007). Grounding object concepts in perception and action: evidence from fMRI studies of tools. Cortex, 43, 461-468.
Borreggine, K. L., \& Kaschak, M. P. (2006). The ActionSentence Compatibility Effect: It's All in the Timing. Cognitive Science, 30, 1097-1112.
Gallese, V. \& Lakoff, G. (2005). The brain's concepts: The role of the sensory-motor system in conceptual knowledge. Cognitive Neuropsychology, 22, 455-479.
Glenberg, A.M., \& Kaschak, M.P. (2002). Grounding language in action. Psychon. Bulletin \& Review, 9, 558-565.
Glenberg A. M., \& Gallese, V. (2012). Action-based language: a theory of language acquisition, comprehension, and production. Cortex 48(7), 905-22.
Hauk, O., Johnsrude, I., \& Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. Neuron, 41, 301-307.
Mahon, B. Z. \& Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. Journal of Physiology Paris, 102, 59-70.
Perniss, P., Thompson, R. L., \& Vigliocco, G. (2010). Iconicity as a general property of language: evidence from spoken and signed languages. Frontiers in Psychology, 1, doi 10.3389/fpsyg.2010.00227.
Sandler, W. \& Lillo-Martin, D. (2006). Sign Language and Linguistic Universals. Cambridge: Cambridge Univ. Press
Stanfield, R. A., \& Zwaan, R. A. (2001). The effect of implied orientation derived from verbal context on picture recognition. Psychological Science, 12, 153-156.
Tettamanti, M., Buccino, G., Saccuman, M. C., Gallese, V., Danna, M., Scifo, P., et al. (2005). Listening to actionrelated sentences activates fronto-parietal motor circuits. Journal of Cognitive Neuroscience, 17, 273-281.
Willems, R. M. \& Casasanto, D. (2011). Flexibility in embodied language understanding. Frontiers in Psychology, 2, doi 10.3389/fpsyg.2011.00116.
Zwaan, R. A. \& Kaschak, M. P. (2008). Language in the brain, body, and world. In M. Robbins \& M. Aydede (Eds.), Cambridge Handbook of Situated Cognition. Cambridge: Cambridge University Press.

# Eye Gaze Affects Vocal Intonation Mimicry 

Marie Postma-Nilsenová (M.Nilsenova@tilburguniversity.edu)<br>Tilburg center for Cognition and Communication (TiCC), Warandelaan 2, 5037 AB Tilburg<br>Tilburg, The Netherlands

Niek Brunninkhuis (n.brunninkhuis@gmail.com)<br>Tilburg center for Cognition and Communication (TiCC), Warandelaan 2, 5037 AB Tilburg<br>Tilburg, The Netherlands

Eric Postma (E.O.Postma@tilburguniversity.edu)
Tilburg center for Cognition and Communication (TiCC), Warandelaan 2, 5037 AB Tilburg
Tilburg, The Netherlands


#### Abstract

Eye gaze and behavioral mimicry are important foundations of social interaction. Inspired by recent studies on eye-gaze mediated spontaneous behavioral mimicry of gestures, we studied the effect of eye gaze direction on vocal mimicry. Participants were instructed to repeat digits spoken by a virtual agent with a direct or averted eye gaze. As a measure of imitation, the vocal pitch was recorded and analyzed in order to determine if and to what extent vocal mimicry was modulated by eye gaze. The results showed that eye gaze direction affects vocal mimicry as measured by pitch slope. That is, when participants were exposed to an agent that gazed at them directly, they accommodated their intonation more to that of the agent, than when they were exposed to an agent that averted its gaze. These results suggest that in social interaction with a virtual agent, humans mimic vocal intonation and that the degree of mimicry depends on the eye-gaze direction of the agent. The implications for studies of social interaction are discussed.


Keywords: Eye Gaze; Vocal Mimicry; Virtual Agent

## 1. Introduction

In social interaction, eye gaze direction and behavioral mimicry are powerful nonverbal social signals (Stass \& Willis Jr, 1967; Kendon, 1967; Scherer, 1974; Cook \& Smith, 1975; Fukayama, Ohno, Mukawa, Sawaki, \& Hagita, 2002; Mason, Tatkow, \& Macrae, 2005; Baaren, Janssen, Chartrand, \& Dijksterhuis, 2009; Wang, 2012). Among different types of nonverbal cues, vocal pitch mimicry appears to play a fundamental role. The results of a range of experimental studies suggest that speakers effortlessly imitate and converge to the phonetic properties of recently heard speech (Delvaux \& Soquet, 2007; Gentilucci \& Bernardis, 2007; Natale, 1975; Pardo, 2006; Shockley, Sabadini, \& Fowler, 2004), including pitch (Babel \& Bulatov, 2012; Goldinger, 1998; Gorisch, Wells, \& Brown, 2012). Pitch - the perceptual correlate of fundamental frequency $\left(F_{0}\right)$ - is, arguably, the most important vocal source of information regarding emotions, stands and attitudes of the speaker. The $F_{0}$ region thus provides acoustic information for imitation exploited in promoting social convergence and status accommodation (Gregory, 1983; Gregory, Webster, \& Huang, 1993; Gregory \& Webster, 1996; Gregory, Dagan, \& Webster, 1997; Haas \& Gregory, 2005; Pardo, 2006) and expressing ingroup-outgroup bias (Babel, 2009; Pardo, Gibbons, Suppes, \& Krauss, 2012).

According to the Communication Accommodation Theory (CAT), in social interaction, people adjust their vocal characteristics to accommodate to each other (Giles, Coupland, \& Coupland, 1991). Support for CAT came from a study by (Gregory \& Webster, 1996), who analyzed Larry King Live television interviews. The results revealed that depending on the relative status of the interviewed guest, Larry King mimicked the vocal characteristics of his guests (in case of high status guests) or the guests mimicked the vocal characteristics of Larry King (in case of low status guests). In general, speakers who are perceived as attractive, likable and/or dominant influence listeners' pitch output, and pitch convergence can be seen as an indicator of cooperative behavior in communication dyads (Okada, Lachs, \& Boone, 2012). Pitch divergence, on the other hand, suggests that speakers may wish to be viewed as dissimilar and increase social distance between themselves (Babel, 2009).

Interestingly, empirical studies have shown that in social interactions, the direction of eye gaze influences the degree of behavioral mimicry (Kleinke \& Pohlen, 1971; Chartrand \& Bargh, 1999). A striking demonstration of the direct link between eye gaze direction and behavioral mimicry is due to Wang, Newport, and Hamilton (2011). In their study, participants were presented with a movie of an actress that either looked directly at the camera or averted her gaze from the camera. In both conditions there were movies of the actress opening her hand and movies in which she closed her hand. At the beginning of each trial, participants were instructed to either open or close their hand. The instructed hand movements could be congruent or incongruent with the displayed hand movements of the actress. After receiving the instruction, participants had to make the hand movement as quickly as possible. They were not instructed to mimic the hand movement of the actress. Not surprisingly, participants were significantly faster in making congruent hand movements than in making incongruent ones. Interestingly, though, the congruent hand movements in the direct gaze condition were considerably faster than those in the averted gaze condition, whereas for the incongruent hand movements gaze direction had no effect. These findings reveal that eye contact has a quick and specific effect on action mimicry.

Inspired by these results, we expect eye-gaze modulated mimicry in other response modalities, such as the vocal modality. Although vocal mimicry is a well-known phenomenon, the modulating effect of eye gaze has not yet been explored experimentally. The aim of this study is to investigate if vocal mimicry is modulated by eye gaze direction.

Instead of exploring the effect of eye gaze on reaction times, we determined its effect on the degree of vocal mimicry. We employed an experimental setting in which a virtual agent with either an averted or direct gaze utters single words with one of three pitch contours. The participants were instructed to repeat the words, but were not instructed to mimic the pitch contours (instead, they were distracted with another task).

## 2. Experiment

## Participants and Design

Forty-seven Dutch native speakers ( 24 male; mean age $21 ; 4$ ) were recruited from the Tilburg University student population. The experiment had a one-way within-subjects design with Eye Gaze Direction (direct, averted) as the independent variable and Vocal Mimicry as the dependent variable.

## Material

The stimulus material consisted of 8 visually presented words during which we measured speakers' baseline pitch, followed by 48 videos ( 16 experimental trials +32 distractors). Eight of the 16 experimental trials involved direct gaze of the agent (top figure 1), during the other 8, the agent either averted his gaze left or right (4 times each). In half of the distractor movies and only in these movies, the agent blinked his eyes; the ratio of gaze directions was 8:4:4 for direct:left:right in the blinking distractor group as well as in the no-blinking distractor group. In every stimulus, the agent expressed a single word (all Dutch monosyllabic digits between 0-10, in order to prevent possible emotional associations to the stimuli that might affect the speaker's pitch), followed by a blank screen.

The virtual agent was created with Poser (Smitch Micro Software inc, Aliso Viejo, California, U.S.), see Figure 1. The agent's lip movements were matched to the pre-recorded words and subtle head and eye movements were added to enhance the realism and to prevent the agent from being perceived as threatening (Ellsworth, 1975; Cook \& Smith, 1975; Argyle, Lefebvre, \& Cook, 1974). A film strip of a sample movie is shown in Figure 2.

The sound files used for the agent's voice were prerecorded in a sound attenuated booth by the second (male) author. His pitch values represented average vocal values for an adult male speaker in the Netherlands (i.e., in the 70-250 Hz range). Three different intonations were used in these recordings (a falling, a rising, and a late-rising tune, see Figure 3). Each movie lasted approximately 5 s , including a 0.5 s fade-from-black and fade-to-black to (i) smooth the transition between consecutive movies, (ii) mark the beginning and end of each stimulus, and (iii) avoid an unnatural and potentially
threatening gaze duration.

## Procedure

The experiment was set up in E-Prime (Psychological Software Tools Inc, Pittsburgh, Pennsylvania, USA) and presented with the help of a Dell Latitude E5510 laptop and a Trust HS-2100 headset. The distance between the participant's mouth and the headset's microphone was kept constant. For the baseline recordings, participants were presented with a random sequence of eight consecutive digits displayed


Figure 1: Impression of the virtual agent used in the experiment. From top to bottom: direct gaze, right averted gaze (while blinking in a distractor movie), and left averted gaze (while speaking).


Figure 2: Film strip of a sample movie as used in the experiment.
in white against a blue background. They were instructed to read each word out loud. Subsequently, they were shown a randomized sequence of videos in which the agent pronounced a digit and were instructed to repeat it. In order to ensure that the participants fixated the agent's eye region and did not focus on the imitation task, they were given the additional instruction to press the space bar whenever the agent blinked. The full length of the experiment was 10 minutes on average.


Figure 3: Graphical representations of the three different intonation patterns of the spoken digits as pronounced by the EIA (second experimenter's voice). From top to bottom: falling intonation, rising intonation, and late-rising intonation (in Hz ).

## Measurements

The experimental audio files collected during the experiment (1128 in total) were manually preprocessed to remove unvoiced speech and silent segments. After establishing the appropriate pitch threshold and ceiling for each individual voice, by auditory and visual inspection of the audio file and spectrogram, respectively, we extracted the pitch contour using the standard autocorrelation-based pitch detection function of Praat (Boersma, 2001). The frequency values were converted to semitones to allow for a comparison of male and female speakers (Borden \& Harris, 1980). For each audio file, we determined the values of two measures of pitch contour: pitch slope $\left(P_{\text {deriv }}\right)$ and pitch regression $\left(P_{\text {regline }}\right)$. The pitch slope is defined as the average difference of adjacent frequencies in the pitch contour. The pitch regression is defined as the slope of the linear regression line through the points making up the pitch contour.

Our measure of vocal mimicry is based on two variables. The first variable $\Delta_{P_{\text {baseline }}}$ is the absolute difference between the agent's pitch, $P_{\text {agent }}$ and the participant's baseline pitch, $P_{\text {baseline }}$, i.e., the pitch of the participant before repeating the agent.

$$
\begin{equation*}
\Delta_{P_{\text {baseline }}}=\left|P_{\text {agent }}-P_{\text {baseline }}\right| \tag{1}
\end{equation*}
$$

The second variable $\Delta_{P_{\text {rep }}}$ is the absolute difference between the agent's pitch and the participant's pitch while repeating the agent, $P_{\text {rep }}$.

$$
\begin{equation*}
\Delta_{P_{\text {rep }}}=\left|P_{\text {agent }}-P_{\text {rep }}\right| \tag{2}
\end{equation*}
$$

By subtracting the values of $\Delta_{P_{\text {baseline }}}$ and $\Delta_{P_{\text {rep }}}$, we obtain our measure of vocal mimicry $M_{V}$.

$$
\begin{equation*}
M_{V}=\Delta_{P_{\text {baseline }}}-\Delta_{P_{\text {rep }}} \tag{3}
\end{equation*}
$$

A positive value of the vocal mimicry $M_{V}$ indicates vocal mimicry, whereas a negative value indicates vocal complementarity. The value of vocal mimicry was calculated both for the pitch slope and the pitch regression separately.

## 3. Results

We start by reporting the vocal mimicry results regardless of eye gaze direction. Subsequently, we report the gazedependent vocal mimicry results. Non-parametric tests were used for data that were not normally distributed.

## Vocal Mimicry Results Independent of Gaze Direction

A Wilcoxon Signed-rank test was performed to establish if vocal mimicry occurred. Effect size estimates were computed using $r(=|Z / \sqrt{N}|)$, where $N$ equals the number of samples. The results for $P_{\text {deriv }}$ and $P_{\text {regline }}$ indicated a significant difference between the baseline and the experimental trial measurements, with a shift in the direction of the agent's pitch (see Table 1). Figures 4 and 5 illustrate the results: Both figures show the median values of the absolute differences between the agent's and the participant's pitch in the baseline and the repetition trial, with the results for males and females plotted separately.

Table 1: Results of Wilcoxon Signed-rank test for mimicry regardless of eye gaze direction (pitch measurements reported in semitones).

|  | $\Delta_{P_{\text {baseline }}}$ | $\Delta_{P_{\text {rep }}}$ | $Z$ | $r$ | $p$ |
| :--- | :--- | :--- | :--- | :--- | :--- |
| $P_{\text {deriv }}$ |  |  |  |  |  |
| Females | 41.473 | 37.926 | -2.829 | 0.417 | 0.005 |
| Males | 41.012 | 38.252 | -2.543 | 0.367 | 0.011 |
| Total | 41.813 | 37.926 | -3.852 | 0.397 | $<0.001$ |
| $P_{\text {regline }}$ |  |  |  |  |  |
| Females | 0.837 | 0.301 | -3.041 | 0.448 | 0.002 |
| Males | 0.847 | 0.255 | -3.114 | 0.449 | 0.002 |
| Total | 0.841 | 0.668 | -4.360 | 0.450 | $<0.001$ |



Figure 4: Plot of the results obtained for $P_{\text {deriv }}$ absolute differences with error bars ( $95 \% \mathrm{CI}$ ).


Figure 5: Plot of the results obtained for $P_{\text {regline }}$ absolute differences with error bars ( $95 \% \mathrm{CI}$ ).

These results indicate that participants accommodated the slope and regression line of their pitch contours to that of the agent. On both measures, when the dataset was split by gender, male and female participants showed similar effects.

## Vocal Mimicry Results Dependent on Gaze Direction

The results of the Wilcoxon Signed-rank test (Table 2) show a significant effect of gaze: compared to the participant's baseline, the slope of the pitch contour in the participant's repetition is more similar to that of the agent gazing towards the participant, than when the agent averted its gaze. A split-file analysis by gender showed a significant effect for male participants only. The medians of $P_{\text {deriv }}$ are visually presented in Figure 6, indicating cases of divergence in the condition with averted gaze in the male participant group.

Table 2: Results of Wilcoxon Signed-rank test for mimicry as measured by $P_{\text {deriv }}$ depending on eye gaze direction.

|  | Median | Median |  |  |  |
| :--- | :--- | :--- | :--- | :--- | :--- |
|  | $M_{V}$ | $M_{V}$ |  |  |  |
|  | Direct | Averted | $Z$ | $r$ | $p$ |
| $P_{\text {deriv }}$ |  |  |  |  |  |
| Females | 6.339 | 2.203 | -1.612 | 0.238 | 0.107 |
| Males | 5.013 | 2.394 | -1.971 | 0.285 | 0.049 |
| Total | 5.754 | 2.203 | -2.529 | 0.261 | 0.011 |

The results of a mixed within-between analysis of variants as measured by $P_{\text {regline }}$ are listed in Table 3. The results indicate no significant main effect of either gaze or gender, as well as no interaction effect of the two variables.

To explore possible individual variations in mimicry (taking pitch slope as the representative measure), we computed


Figure 6: The median values of $P_{\text {deriv }}$ absolute differences between the agent's and the participant's pitch for direct and averted gaze including error bars ( $95 \% \mathrm{CI}$ ).

Table 3: Results of ANOVA test for mimicry as measured by $P_{\text {regline }}$ depending on eye gaze direction.

|  | df | $F$ | $p$ | $\eta_{p}^{2}$ |
| :--- | :--- | :--- | :--- | :--- |
| $P_{\text {regline }}$ |  |  |  |  |
| Gender | 1 | 0.083 | 0.775 | 0.002 |
| Gaze | 1 | 1.772 | 0.190 | 0.038 |
| Gender $\times$ Gaze | 1 | 0.386 | 0.537 | 0.009 |

for each participant the average difference between $M_{V}$ in the averted gaze condition and $M_{V}$ in the direct gaze condition. The resulting scores reflect the individual effect of gaze on vocal mimicry; positive scores are associated with vocal mimicry, negative scores with vocal complementarity. Figure 7 is a graphical display of the scores, sorted from smaller to larger scores. Each bar represents a participant, and the height of the bar represents the magnitude of mimicry (positive) or complementarity (negative).


Figure 7: Individual differences in vocal mimicry
To sum up, we found that participants exhibit verbal mimicry as measured by $P_{\text {deriv }}$ and $P_{\text {regline }}$ and that at least in the case of $P_{\text {deriv }}$, the mimicry effect is stronger when the agent gazes at the participant. In addition, we observed individual varia-
tions in the degree of vocal mimicry. The implications of our findings are discussed in the next section.

## 4. General Discussion and Conclusion

Pitch is arguably the most important source of information regarding emotions, stances and attitudes of the speaker (Juslin \& Laukka, 2003) and pitch mimicry plays an important role in human interaction in that it reflects the closeness of the social bond between two individuals. Our findings indicate that pitch mimicry in social interactions may be modulated by eye gaze. The results of our experiment extend and generalize the findings obtained by Wang et al. (2011) for the visual (gesture) modality. The existence of eye-gaze modulated vocal mimicry underscores the importance of eye gaze as a social signal and lends further support to the close relation of eye contact and behavioral mimicry in social interaction.

The potential impact of eye gaze on the social bond between virtual agents and humans is of relevance to the development of future human-computer interfaces that display an interactive embodied agent and sense vocal and visual cues of the human interacting with the agent. Software controlling an interactive embodied agent, may confirm the establishment of a social bond with the human by instructing the agent to eye gaze and vocally address the human and subsequently sense the concomitant vocal mimicry.

In our experiment, participants were instructed to repeat the digit pronounced by the agent. It is not clear to what extent the mimicry observed depends on the type of instruction. Future work may experiment with alternative types of responses. For instance, participants could be instructed to complete a partial sentence uttered by the agent or to respond to a statement. In this way the dependency between an explicit instruction to repeat an utterance and behavioral mimicry can be determined.

Our use of a male human voice and a male virtual agent, may have caused gender effects that can be further explored in future studies. According to the CAT (Giles et al., 1991), talkers modify their speech to reinforce valued and socially meaningful differences between themselves and their interaction partners. Since male voices are lower pitched than female voices (Sachs, Lieberman, \& Erickson, 1973), females possibly reinforced the gender difference between themselves and the agent by deviating from the agents relatively low pitched voice and produce a higher pitched voice. Another issue to be explored in the future concerns the effect of joint attention. As well known, interlocutors are likely to follow each others gaze direction. It remains to be seen if contexts eliciting joint attention support vocal and other types of behavioral mimicry.

## References

Argyle, M., Lefebvre, L., \& Cook, M. (1974). The meaning of five patterns of gaze. European Journal of Social Psychology, 4(2), 125-136.
Baaren, R., van, Janssen, L., Chartrand, T., \& Dijksterhuis, A. (2009). Where is the love? The social aspects of
mimicry. Philosophical Transactions of The Royal Society B, 364(1528), 2381-2389.
Babel, M. (2009). Phonetic and social selectivity in speech accommodation. Unpublished doctoral dissertation, University of California.
Babel, M., \& Bulatov, D. (2012). The role of fundamental frequency in phonetic accommodation. Language and Speech, 55(2), 231-248.
Boersma, P. (2001). Praat, a system for doing phonetics by computer. Glot International, 5(9/10), 341-345.
Borden, G., \& Harris, K. (1980). Speech science primer: Physiology, acoustics and perception of speech. Baltimore, Maryland, USA: Williams \& Wilkins.
Chartrand, T., \& Bargh, J. (1999). The chameleon effect: The perception-behavior link and social interaction. Journal of Personality and Social Psychology, 76, 893-910.
Cook, M., \& Smith, M. (1975). The role of gaze in impression formation. British Journal of Social and Clinical Psychology, 14, 19-25.
Delvaux, V., \& Soquet, A. (2007). The influence of ambient speech on adult speech productions through unintentional imitation. Phonetica, 64, 145-173.
Ellsworth, P. (1975). Direct gaze as a social stimulus: The example of aggression. In L. Krames, T. Alloway, \& P. Pliner (Eds.), Nonverbal communication of aggression. New York, New York, USA: Plenum Press.
Fukayama, A., Ohno, T., Mukawa, N., Sawaki, M., \& Hagita, N. (2002). Messages embedded in gaze of interface agents - impression management with agent's gaze. In Proceedings of the sigchi conference on human factors in computing systems: Changing our world, changing ourselves (pp. 41-48). New York, New York, USA: ACM.
Gentilucci, M., \& Bernardis, P. (2007). Imitation during phoneme production. Neuropsychologia, 45, 608-615.
Giles, H., Coupland, J., \& Coupland, N. (1991). Accommodation theory: Communication, context, and consequence. In H. Giles, J. Coupland, \& N. Coupland (Eds.), Contexts of accommodation. developments in applied sociolinguistics (pp. 1-68). Cambridge University Press.
Goldinger, S. D. (1998). Echoes of echoes? an episodic theory of lexical access. Psychological Review, 105, 251279.

Gorisch, J., Wells, B., \& Brown, G. (2012). Pitch contour matching and interactional alignment across turns: an acoustic investigation. Language and Speech, 55, 57-76.
Gregory, S. (1983). A quantitative analysis of temporal symmetry in microsocial relations. American Sociological Review, 129-135.
Gregory, S., Dagan, K., \& Webster, S. (1997). Evaluating the relation of vocal accommodation in conversation partners' fundamental frequencies to perceptions of communication quality. Journal of Nonverbal Behavior, 21(1), 23-43.
Gregory, S., \& Webster, S. (1996). A nonverbal signal in voices of interview partners effectively predicts communication accommodation and social status perceptions. Jour-
nal of Personality and Social Psychology, 70, 1231-1240.
Gregory, S., Webster, S., \& Huang, C. (1993). Voice pitch and amplitude convergence as a metric of quality in dyadic interviews. Language \& Communication.
Haas, A., \& Gregory, S. (2005). The impact of physical attractiveness on womens social status and interactional power. Sociological Forum, 20, 449-471.
Juslin, P., \& Laukka, P. (2003). Communication of emotions in vocal expression and musical performance: different channels, same code? Journal of the Acoustical Society of America, 129, 770-814.
Kendon, A. (1967). Some Functions of Gaze-direction in Social Interaction. Acta Psychologica, 26, 22-63.
Kleinke, C., \& Pohlen, P. (1971). Affective and emotional responses as a function of other person's gaze and cooperativeness in a two-person game. Journal of Personality and Social Psychology, 17(3), 308-313.
Mason, M. F., Tatkow, E., \& Macrae, C. (2005). The look of love: gaze shifts and person perception. Psychological Science, 16, 236-239.
Natale, M. (1975). Social desirability as related to convergence of temporal speech patterns. Perceptual and Motor Skills, 40(3), 827-830.
Okada, B., Lachs, L., \& Boone, B. (2012). Interpreting tone of voice: Musical pitch relationships convey agreement in dyadic conversation. Journal of Acoustical Society of America, 132(3), EL208-EL214.
Pardo, J. (2006). On phonetic convergence during conversational interaction. Journal of the Acoustical Society of America, 119, 2382-2393.
Pardo, J., Gibbons, R., Suppes, A., \& Krauss, R. (2012). Phonetic convergence in college roommates. Journal of Phonetics, 40, 190-197.
Sachs, J., Lieberman, P., \& Erickson, D. (1973). Anatomical and cultural determinants of male and female speech. In R. Shuy \& R. Fasold (Eds.), Language attitudes: Current trends and prospects (pp. 74-84). Washington DC, USA: Georgetown University Press.
Scherer, S. (1974). Influence of proximity and eye contact on impression formation. Perceptualand Motor Skills, 38, 538.

Shockley, K., Sabadini, L., \& Fowler, C. (2004). Imitation in shadowing words. Perception \& Psychophysics, 66, 422429.

Stass, J., \& Willis Jr, F. (1967). Eye-contact, pupil dilation, and personal preference. Psychonomic Science, 7, 375-376.
Wang, Y. (2012). The control of mimicry by social signals. Unpublished doctoral dissertation, University of Nottingham.
Wang, Y., Newport, R., \& Hamilton, A. (2011). Eye contact enhances mimicry of intransitive hand movements. Biology Letters, 7(1), 7-10.

# The Power of Words in the Brain: Systematic Sound-Meaning Associations in Novel and Existing Words 

Marie Postma-Nilsenová (m.nilsenova@tilburguniversity.edu)<br>Department of Communication and Information Sciences, Tilburg University, P.O. Box 9013, 5000 LE Tilburg, The Netherlands

Constantijn Kaland (c.c.l.kaland@tilburguniversity.edu)<br>Department of Communication and Information Sciences, Tilburg University, P.O. Box 9013, 5000 LE Tilburg, The Netherlands

Leonoor Oversteegen (e.oversteegen@tilburguniversity.edu)<br>Department of Communication and Information Sciences, Tilburg University, P.O. Box 9013, 5000 LE Tilburg, The Netherlands


#### Abstract

In a series of experiments conducted with Dutch native speakers, we explored systematic size/power sound-symbolic associations in novel and existing words. In Experiment $1(N=64)$, participants associated vowel-intrinsic fundamental frequency with size/power, disregarding the modality of stimuli presentation (spoken, written), but depending on the lexical status of the stimulus (more strongly for novel then for existing words). In Experiment $2(N=56)$, we explored the idea that the order of vowels in a word affects soundsymbolic associations, as pitch contours emerge from a sequence of vowel-intrinsic fundamental frequencies. Participants perceived stimuli with 'rising' combinations of front-back vowels as less powerful than stimuli with 'falling' combinations. This finding indicates that even in non-tonal languages, sound symbolism is not bound to a single segment (phoneme). We compared the effect to the perception of tones in a tonal language, which we explored in Experiment 3 with Mandarin native speakers $(N=96)$ judging the perception of power in monosyllabic novel brand names with four different tones (rising, falling, level and fall-rise). In Experiment 4 ( $N=146$ ), we examined the effect of vowel-intrinsic intensity, which has previously remained un-noted. The results showed that like fundamental frequency, also intrinsic intensity influences size/power-symbolic associations.


Keywords: Sound Symbolism; Cross-modal Correspondence; Form-Meaning Arbitrariness; Frequency Code; Intrinsic Intensity; Intrinsic Fundamental Frequency.

## Introduction

A series of relatively recent experiments has shaken one of the most fundamental postulates about human language, namely the arbitrariness assumption regarding the relation between the sound structure of a word and its meaning. Originating with de Saussure, it has long been presumed that the number of pictorial, imitative, or onomatopoetic words in any language is very small (Hockett, 1958; de Saussure, 1916/1959). However, in violation of de Saussure's arbitrariness assumption, many studies confirm that both existing words and non-words in different languages evoke connotations based on their sound structure (Klink, 2000, 2001, 2003; Yorkston \& Menon, 2004; Lowrey \& Shrum, 2007; Shrum et al., 2012; Kovic et al.,

2010; Coulter \& Coulter, 2010; Aveyard, 2012; Parise \& Spence, 2012; Spence, 2011; Baxter \& Lowrey, 2011). The sound-meaning link appears to facilitate both L1 and L2 learning (Nygaard et al., 2009; Imai et al., 2008) by helping to solve the "cross-modal correspondence problem" (which stimuli to link to link together across the senses; Spence, 2011; Bremner et al., 2012) and has been claimed to hold universally (Ultan, 1978; Ramachandran \& Hubbard, 2005).

On the one hand, languages encode crossmodal relationships by mapping phonemes to specific shapes. For example, by virtue of the so-called "Bouba-Kiki" effect, round vowels are typically linked to rounded objects, whereas high-pitched un-round vowels bring to mind angularity (Köhler, 1929; Ramachandran \& Hubbard, 2001; Westbury, 2005; Bremner et al., 2012). The effect seems to hold for children as young as 2-2.5 years and concerns both objects and visual actions (Maurer, Pathman, \& Mondloch, 2006; Imai et al., 2008). On the other hand, sound-symbolic phenomena might give rise to a visual size/power association (Sapir, 1929). In various languages, words containing vowels with high intrinsic fundamental frequency, the acoustic correlate of pitch, (e.g., front vowels such as /ı, i, $\varepsilon /$; Lehiste \& Peterson, 1961; Whalen \& Levitt, 1995) are perceived as referring to lighter, thinner, smaller, less powerful and more feminine entities/individuals and, possibly, prickly, sour taste (Crisinel \& Spence, 2009) as opposed to words containing vowels with low intrinsic fundamental frequency (e.g., back vowels such as $/ \tau, o, \rho /$ ). A similar effect can be achieved by using different types of consonants that can influence fundamental frequency either directly, or by raising the frequency of the accompanying vowel (cmp. voiceless vs. voiced consonants, fricatives vs. plosives). The effect has primarily been observed for nonwords (e.g., invented brand names; Klink, 2000, 2001, 2003; Yorkston \& Menon, 2004; Lowrey \& Shrum, 2007; Shrum et al., 2012) but appears to hold for existing words as well, at least in limited domains: In a study in the area of consumer psychology, Coulter \& Coulter (2010) reported that vowel pitch in numerals can influence the perception of price discounts in English and Chinese.

The association between pitch and size/power can, for the most part, be explained by the mechanism of the Frequency Code (Ohala 1994; Gussenhoven, 2004), a psycholinguistic principle primarily used to account for interpretational effects of rising and falling utterance contours in different languages (Chen, 2004). According to the Frequency Code, high pitch is linked to smaller immature speakers (with thinner, smaller vocal folds and shorter vocal tracts producing higher fundamental frequency) and, hence, utterances carrying high and/or rising pitch are interpreted as less dominant, more uncertain, friendlier and questioning. Presumably, the effects of the Frequency Code can be translated from the suprasegmental domain (utterance prosody) to the micro-prosodic, segmental domain, and affect meaning with the help of motor-imitative mechanisms associated with speech perception.

## Current Study

In our current study, we attempted to explore further the effect of phonetic segment-intrinsic properties on interpretation of words and non-words. Surprisingly, past research of the sound-referent cross-modal mappings mostly involved written presentation of stimuli. Given that the Frequency Code is grounded in principles of speech production, the question arises what role the modality of presentation plays in the sound-size association. Are sound symbolic effects stronger if words are presented aurally rather than visually? Moreover, according to most models of reading (Coltheart, 2006; Coltheart, 2012), unlike existing words, pseudo-words and non-words can only be pronounced with the help of a system that associates visual segments (i.e., graphemes) with sounds. Are soundsymbolic effects stronger for pseudo-words and non-words than for existing words? We addressed these two questions in Experiment 1 conducted with Dutch native speakers.

Existing research in the area of phonetic symbolism has, so far, disregarded the possibility that not just individual phonemes, but also their sequences, might create meaningful associations between phoneme-intrinsic fundamental frequency and meaning, by virtue of producing a micro-prosodic contour. For example, in words consisting of two syllables, a sequence of high-low vowels (e.g., /CiCo/) produces a "falling contour", while the opposite sequence (/CoCi/) produces a "rising contour". In accordance with the predictions of the Frequency Code (Gussenhoven, 2004; Chen, 2004), these two contours should receive different interpretations (big/powerful and small/powerless, respectively). Do sound-symbolic associations vary with different sequencing of high- and low-pitched phonemes in a word? We explored the issue in Experiment 2 and compared the outcomes collected for a non-tonal language (Dutch) with perceptions of lexical tones in a tonal language (Mandarin Chinese) in Experiment 3.

The last issue raised in the current study concerns the effect of other segment-intrinsic prosodic features apart from fundamental frequency, in particular, segment-intrinsic intensity (Lehiste \& Peterson, 1959). As observed in
prosodic studies of interpersonal influence, intensity (loudness) is positively associated with judgments of dominance and potency (Scherer, 1974; Aronovitch, 1976; Buller \& Burgoon, 1986). If size/power-symbolic relations arise by transfer from the suprasegmental to the segmental domain, we might expect sounds with intrinsically higher intensity to have an effect comparable to that of low intrinsic fundamental frequency. To our knowledge, soundsymbolic associations of intrinsic intensity have not been explored in the past. Given the frequent co-occurrence of high intrinsic intensity and low intrinsic pitch, it might, in fact, be the case that at least some of the sound-meaning associations reported in the literature might be due to intrinsic intensity, rather than intrinsic fundamental frequency. Therefore, we also examined the interplay of the two acoustic features in Experiment 4.

To sum up, in our study, we addressed the following research questions:

1. Are sound symbolic effects stronger if words are presented aurally rather than visually? Are they stronger for pseudo-words than for existing lexemes? (Experiment 1)
2. Do sound-symbolic associations vary with different sequencing of high- and low-pitched phonemes in a word? (Experiment 2 and 3)
3. Are sound-symbolic associations affected by segmentintrinsic intensity? (Experiment 4)

## Experiment 1

In the first experiment, we addressed the question whether or not the modality in which stimuli are presented can affect sound-symbolic interpretation of vowel-intrinsic fundamental frequency. Given that most studies of crossmodal correspondences make use of non-words and pseudowords, we were also interested to find out if the effect is weaker (or perhaps absent) for items with denotative components.

## Participants and Design

Sixty-four native speakers of Dutch ( 32 male; $M_{\text {age }}=25.3$, $S D=10.5$ ) participated in the experiment on a voluntary basis. The experiment had a $2 \times 2 \times 2$ mixed withinbetween design. The between-participant variable was Modality in which participants perceived the stimuli (reading or listening). The participants were randomly assigned to the reading or listening condition (32 in each). The within-participant variables were Denotation (stimuli consisting either of existing words or pseudo-words) and Vowel-Intrinsic Fundamental Frequency (High or Low). The dependent variable was Perceived Power/Size.

## Material

The stimuli were either pseudowords consisting of phoneme and syllable combinations that are possible in Dutch but meaningless (e.g., Boloem, Wabboelan, Piripi and Kenep) or pseudowords composed of meaningful morphemes (e.g., Zondaar, Godenbad, Tinteling or Kietelkit). In total, 160 biand tri-syllabic stimuli were systematically divided into four
variants, resulting in lists of 40 items ( 20 existing words or compounds, 20 pseudo-words). For the listening condition, the stimuli were recorded in a soundproof booth by a native Dutch female speaker ( 50 y ) without any detectable dialect or speech habits.

## Procedure and Instrumentation

The experiment was cast as a marketing study of brand names for a generic type of shampoo. In the listening condition, the stimuli were presented on a Philips Go Gear MP4 with a headset. In the reading condition, the participants filled out a pen-and-paper form. Each experimental session lasted 6-10 minutes. Participants indicated their perception of the brand names on 4 VAS scales each 7.5 cm long (male/female, big/small, dominant/not dominant and unfriendly/friendly). The Cronbach's alpha coefficient of the four scales was .8 and the scales were reduced to a single Perceived Power/Size measure.

## Results

A mixed analysis of variance was used to explore the effect of the independent variables on the Perceived Power/Size. Stimuli with low intrinsic vowel frequency scored higher on the Power/Size scale than stimuli with high frequency vowels, $F(1,62)=326.77, p<.001, \eta_{p}{ }^{2}=.79$. The interaction effect of Denotation and Vowel-Intrinsic Fundamental Frequency was also significant, $F(1,62)=$ 41.29, $p<.001, \eta_{\mathrm{p}}{ }^{2}=.40$; the interpretation of pseudowords was more sensitive to the intrinsic vowel properties and their size/power-symbolic associations. Than the interpretation of existing words. There was no effect of modality of presentation.

## Discussion

The results of the first experiment show that the perception of pseudowords consisting of both meaningful and meaningless morphemes is influenced by vowel-intrinsic fundamental frequency: in accordance with the predictions of the Frequency Code, front vowels are associated with less power/smaller size than back vowels. The effect was found to be independent of the modality in which stimuli were presented. The effect was stronger for meaningless pseudowords than for those composed of existing morphemes, suggesting that the phoneme-to-grapheme route (Coltheart, 2006; Coltheart, 2012) might play a role.

## Experiment 2

In the second experiment, we addressed the question of the scope of vowel-intrinsic fundamental frequency effects: Are they bound strictly to the individual phonemes or can they combine into tones with different sound-symbolic interpretations, akin to lexical tones in tonal languages?

## Participants and Design

The experiment was conducted among 56 participants ( 25 male, $M_{\text {age }}=35.9, S D=13.0$ ), all native speakers of Dutch,
taking part on a voluntary basis. The independent variable was Vowel-Intrinsic Fundamental Frequency (High, Low, Falling and Rising and the dependent variable was perceived Power/Size of the stimulus (brand name).

## Materials

The stimulus list contained 32 meaningless pseudowords, 16 experimental and 16 fillers. The experimental items consisted of 8 bisyllabic and 8 monosyllabic meaningless pseudowords with either the sequence front vowel - back vowel or the opposite, or just a front or a back vowel, with a systematic variation of voiced and voiceless plosives and fricatives (see Table 1).

## Procedure and Instrumentation

The experiment was cast as a marketing study for a new product name and presented online via the LimeSurvey platform. The stimuli were presented in a visual modality only. Participants evaluated the stimuli on four 7-point scales (male/female, big/small, thick/thin and heavy/light). The mean Cronbach's alpha coefficient was .8 and for the purpose of the statistical analyses, the scales were thus reduced to a single Power/Size measure.

Table 1: Examples of experimental stimuli.

|  |  | High | Low | Falling | Rising |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  |  |  |  |
| C-Type | Voicing |  |  |  |  |
| Plosive | - | $T i$ | To | Tito | Toti |
|  | + | $B i$ | Bo | Bibo | Bobi |
| Fricative | - | Vie | Voe | Vievoe | Voevie |
|  | + | Soe | Sie | Siesoe | Soesie |

## Results

A repeated-measures ANOVA showed a significant effect of Vowel-Intrinsic Fundamental Frequency on the Perceived Power/Size of the novel product name, $F_{l}(3,165)=47.59$, $p<.001, \eta_{\mathrm{p}}{ }^{2}=.46, F_{2}(3,31)=4.23, p<.05, F_{\text {min }}^{\prime}(3,37)$ $=3.88, p<.05$. A post-hoc analysis indicated that the frontback (intrinsically falling) sequence was evaluated as more powerful ( $M=4.23, S D=.74, p=.017$ ) than the back-front (intrinsically rising) sequence ( $M=3.92, S D=.68, p<$. 001 ) and monosyllabic front-vowel stimuli (intrinsically high; $M=3.00, S D=.66, \mathrm{p}<.001$ ). The monosyllabic backvowel stimuli (intrinsically low) were perceived as the most powerful product names $(M=4.33, S D=.66)$ compared to the other three types $(p<.001)$.

## Discussion

In the second experiment, we tested the interpretation of vowel-intrinsic $F_{0}$ in combinations of two syllables, using the dimensions 'front-back' (/I-9/ and /i-u/). The vowel order had a large significant effect on the perception of Power/Size. Stimuli with 'rising' combinations of front-back vowels (e.g., /kuki/) were perceived as less powerful than stimuli with 'falling' combinations (e.g., /kiku/); they were
also judged as being less powerful than monosyllabic stimuli with a high (front) vowel. This finding indicates that sound symbolism that arises due to vowel-intrinsic fundamental frequency is not bound to a single phoneme but may be the result of combining the acoustic values for sequences of syllables.

## Experiment 3

In the third experiment, we explored sound-symbolic interpretations of tones in a tonal language (Mandarin Chinese). In spite of the long tradition of phonetic and phonological research into tone perception and use in tonal languages, no experimental studies have laid an explicit link between the Frequency Code and lexical tones (for a general overview, see Lapolla, 1994).

## Participants and Design

Ninety-six graduate and undergraduate Chinese students (39 male) from a university in Beijing participated in the study, 11 participants were excluded from the analysis due to incomplete data caused by server problems. The study was conducted in Mandarin Chinese and had a $2 \times 4$ design mixed within-between participant design. The withinparticipant variable was Tone (Tone 1 in high pitch height (H) - Level, Tone 2 in half-high pitch height (LH) - Rising, Tone 3 in half-low pitch height $(\mathrm{L}(\mathrm{H})$ ) - Low-Rising, and Tone 4 in low pitch-height (HL) - Falling; Gussenhoven, 2004; Wiedenhof, 2004). The between-participant variable was the writing system (Chinese characters or Pinyin); participants were randomly divided into one of the two conditions ( 45 in the character-written group and 40 in the Pinyin-written group).

## Material

Sixteen non-existent brand names (pseudowords) were written in matched Chinese characters and in Pinyin with four different tones (e.g., lū, lú, lǔ and lù). A pretest conducted with the Pinyin transcriptions showed that Chinese readers (same population, different group than the participant group) did not have any meaningful associations with the stimuli outside of the context of the experiment.

## Procedure and Instrumentation

Participants evaluated the visually presented stimuli on a 7point scale using the anchors "not powerful at all" (1) and "very powerful" (7). The study was conducted online with help of the Qualtrics platform.

## Results

There was a significant main effect of the type of tone on the average perceived power of brand names, $F(1.88$, 156.71) $=79.297, p<.001, \eta_{\mathrm{p}}{ }^{2}=.49$, with degrees of freedom corrected using Huynh-Feldt estimates of sphericity. Post-hoc comparisons revealed that the Tone 4 (Falling) was more powerful than the other three Tones. Tone 2 (Rising) and Tone 3 (Low-Rising) were more powerful than Tone 1 (Level/High); the perception of Tone 2
was not different from the perception of Tone 3, see Fig. 1. The writing system used for the presentation of stimuli did not have a significant effect.


Figure 1: Perception of Chinese lexical tones in novel brand names.

## Discussion

The results of the experiment show that the lexical tones in Mandarin Chinese are interpreted in a similar way as segmental and suprasegmental pitch (fundamental frequency) in non-tonal languages. Along the lines predicted by the Frequency Code, high (level) and rising tones are associated with less powerful brand names than falling tones.

## Experiment 4

In the fourth experiment, we identified another intrinsic acoustic property of vowels, namely intensity, as a feature that might play a role in cross-modal relationships. Given that the two acoustic values often correlate, it could be that at least some of the sound-symbolic effects discussed in the literature are caused by perceptions of intensity. Therefore, we were interested in a possible interplay of the two acoustic properties.

## Participants and Design

The participants were 146 native Dutch speakers ( 84 male, $\left.M_{\text {age }}=40.39, S D=13.5\right)$ who took part in the experiment on a voluntary basis. The experiment had a $2 \times 2$ withinparticipant design, with Fundamental Frequency (High, Low) and Intensity (High, Low) as independent variables and Perceived Power/Size as the dependent variable.

## Material

We used 40 fictitious brand names (pseudowords) as stimulus material. The brand names were constructed in accordance with the measurements of intrinsic fundamental frequency and intrinsic intensity of vowels (Lehiste \& Peterson, 1959; 1961) and within the limits of the Dutch
phonemic inventory in the following way: we selected four vowels as the most extreme representatives of the two acoustic dimensions (/u/ ~ [+ Fundamental Frequency, + Intensity], $/ \mathrm{i} / \sim$ [ Fundamental Frequency, - Intensity], /a/ $\sim$ [- Fundamental Frequency, + Intensity] and $/ \mathrm{I} / \sim$ [Fundamental Frequency, - Intensity]), see Fig. 2. The four vowels were systematically combined with both voiced and voiceless consonants in bi-syllabic non-words ("company names"), e.g. Manan, Doedoer, Tipit and Liediel.


Figure 2: Intrinsic fundamental frequency and intrinsic intensity of vowels (after Lehiste \& Peterson, 1959, 1961).

## Procedure and Instrumentation

The material was presented visually via the LimeSurvey platform as a part of a marketing study conducted on behalf of a foreign company in the Netherlands. Participants evaluated possible company names on four 5 -point scales: male/female, big/small, dominant/submissive and strong/weak. On the basis of the Cronbach's alpha (. 9 on average), the measures were reduced to a single scale.

## Results

The results of a repeated measures ANOVA show a significant effect of Fundamental Frequency, $F_{( }(1,145)=$ $223.44, p<.001, \eta_{\mathrm{p}}{ }^{2}=.61, F_{2}(1,36)=73.75, p<.001$, $F_{\text {min }}^{\prime}(1,62)=55.45, p<.001$, and a slightly weaker significant effect of Intensity, $F_{l}(1,145)=155.17, \mathrm{p}<.001$, $\eta_{\mathrm{p}}{ }^{2}=.52, F_{2}(1,36)=49.47, p<.001, F_{\min }^{\prime}(1,61)=37.51$, $p<.001$, as well as a trend of an interaction effect, $F_{l}(1$, $145)=26.66, p<.001, \eta_{p}{ }^{2}=.17, F_{2}(1,36)=3.76, p<$. $001, F_{\text {min }}^{\prime}(1,47)=3.30, p=.08$ on the perceived Power/Size, see Table 2.

Table 2: Mean Perceived Power/Size.

| Fundamental <br> Frequency | Intensity | $M$ | $S D$ |
| :---: | :---: | :---: | :---: |
| High | High | 2.34 | .75 |
|  | Low | 2.01 | .58 |
| Low | High | 3.00 | .65 |
|  | Low | 2.43 | .60 |

## Discussion

The results of Experiment 4 indicate that sound-size/power correspondence can arise not just thanks to vowel-intrinsic fundamental frequencies but also as a result of intrinsic intensity. Given that the two acoustic values appear to be in a negative correlation, it was not possible to pry apart their effects using written stimuli. In a follow-up study, their contribution could be explored in detail with the help of artificially synthesized stimuli. Another issue to explore in the future concerns individual differences in vowel pronunciation. As noted by Lehiste \& Peterson (1959, 1961), different speakers, in fact, vary in their use of the vowel space. The question thus arises if the sound-symbolic effects found in various studies are linked to speakers' own prototypical production targets or based on previous perceptual experience.

## General Discussion and Conclusion

In our study, we addressed a number of important questions regarding cross-modal correspondences in the auditory domain. We conducted a series of experiments under the assumption that the size/power sound-symbolic relations are due to intrinsic acoustic properties of phonemes captured by the psycholinguistic principle of the Frequency Code, rather than by idiosyncratic mechanisms akin to weak synaesthesia (see Parise \& Spence, 2012, for a discussion). The results of the experiments indicate that sound symbolic effects are not necessarily stronger when words are presented aurally rather than visually, however, the cross-modal sound-size correspondences are stronger for pseudo-words than for existing words. The intrinsic fundamental frequency effects are not bound to individual phonemes but can combine into tones with different sound-symbolic interpretations. In a sense, non-tonal languages thus resemble tonal languages where different lexical tones appear to be interpreted differently (at least on pseudo-words). Finally, other acoustic features apart from fundamental frequency may codetermine the cross-modal sound-based associations: in particular, intrinsic intensity seems to play a significant role.

## Acknowledgments

The authors would like to thank Jill Meijer, Milou Steenvoorden, Xiaohan Wang and Stijn Gimbrère for their help with collecting experimental data.

## References

Aronovitch, C.D. (1976). The voice of personality: Stereotyped judgments and their relation to voice quality and sex of speaker. Journal of Social Psychology, 99, 207-220.
Aveyard, M.E. (2012). Some consonants sound curvy: Effects of sound symbolism on object recognition. Memory and Cognition, 40, 83-92.

Baxter, S., \& Lowrey, T. (2011). Phonetic symbolism and children's brand name preferences. Journal of Consumer Marketing, 28, 516-523.
Bremner, A.J., Caparos, S., Davidoff, J., de Fockert, J., Linnell, K.J., \& Spence, C. (2013). "Bouba" and "Kiki" in Namibia? A remote culture makes similar shapesound matches, but different shape-taste matches to Westerners. Cognition, 126, 165-172.
Buller, D.B., \& Burgoon, J.K. (1986). The effects of vocalics and nonverbal sensitivity on compliance: A replication and extension. Human Communication Research, 13, 126-144.
Chen, A. (2004). Universal and language-specific perception of paralinguistic intonational meaning. Doctoral, dissertation, Radboud University, Nijmegen.
Coltheart, M. (2006). Dual route and connectionist models of reading: an overview. London Review of Education, 4, 5-17.
Coltheart, M. (2012). Dual-route theories of reading aloud. In J.S. Adelman (Ed.), Visual word recognition. New York: Psychology Press.
Coulter, K.S., \& Coulter, R.A. (2010). Small sounds, big deals: phonetic symbolism effects in pricing. Journal of Consumer Research, 37, 315-328.
Crisinel, A.-S., \& Spence, C. (2009). Implicit association between basic tastes and pitch. Neuroscience Letters, 464, 39-42.
Gussenhoven, C. (2004). The Phonology of Tone and Intonation. Cambridge University Press.
Imai, M., Kita, S., Nagumo, M., \& Okada, H. (2008). Sound symbolism facilitates early verb learning. Cognition, 109, 54-65.
Klink, R.R. (2000). Creating brand names with meaning: The use of sound symbolism. Marketing Letters, 11, 520.

Klink, R.R. (2001). Creating meaningful brand names: a study of semantics and sound symbolism. Journal of Marketing Theory and Practice, 9, 27-34.
Klink, R.R. (2003). Creating meaningful brands: The relationship between brand name and brand mark. Marketing Letters, 14, 143-157.
Kovic, V., Plunkett, K., \& Westermann, G. (2010). The shape of words in the brain. Cognition, 114, 19-28.
Köhler, W. (1929). Gestalt Psychology. Liveright: New York.
Lapolla, R.J. (1994). An experimental investigation into phonetic symbolism as it relates to Mandarin Chinese. In L. Hinton, J. Nichols \& J.J. Ohala (Eds.), Sound Symbolism. Cambridge: Cambridge University Press.
Lehiste, I., \& Peterson, G.E. (1959). Vowel amplitude and phonemic stress in American English. Journal of the Acoustical Society of America, 31, 428-435.
Lehiste, I., \& Peterson, G.E. (1961). Some basic considerations in the analysis of intonation. Journal of the Acoustical Society of America, 33, 419-425.

Lowrey, T.M., \& Shrum, L.J. (2007). Phonetic symbolism and brand name preference. Journal of Consumer Research, 34, 406-414.
Maurer, D., Pathman, T., \& Mondloch, C. (2006). The shape of boubas: Sound-shape correspondences in toddlers and adults. Developmental Science, 9, 316-322.
Nygaard, L.C., Cook, A.E., \& Namy, L.L. (2009). Sound to meaning correspondences facilitate word learning. Cognition, 112, 181-186.
Ohala, J.J. (1994). The Frequency Code underlies the sound-symbolic use of voice pitch. In J. Nichols \& J.J. Ohala (Eds.), Sound Symbolism. Cambridge: Cambridge University Press.
Parise, C.V., \& Spence, C. (2012). Audiovisual crossmodal correspondences and sound symbolism: a study using the implicit association test. Experimental Brain Research, 220, 319-333.
Ramachandran, V.S., \& Hubbard, E.M. (2001). Synaesthesia - A window into perception, thought and language. Journal of Consciousness Studies, 8, 3-34.
Ramachandran, V.S., \& Hubbard, E.M. (2005). The emergence of the human mind: Some clues from synesthesia. In L.C. Robertson \& N. Sagiv (Eds.), Synesthesia: Perspectives from Cognitive Neuroscience. Oxford: Oxford University Press.
de Saussure, F. (1916/1959). Course in General Linguistics. New York: Philosophical Library.
Scherer, K.R. (1974). Acoustic concomitants of emotional dimensions: Judging affect from synthesized tone sequences. In S. Weitz (Ed.), Nonverbal Communication. New York: Oxford University Press.
Shrum, L.J., Lowrey, T.M., Luna, D., Lerman, D., \& Liu, M. (2012). Sound symbolism effects across languages: Implications for global brand names. International Journal of Research in Marketing, forthcoming.
Spence, C. (2011). Crossmodal correspondences: A tutorial review. Attention, Perception, \& Psychophysics, 73, 971995.

Ultan, R. (1978). Size-sound symbolism. In J.H. Greenberg (Ed.), Universals of Human Language, Vol. 2: Phonology. Stanford: Stanford University Press.
Westbury, C. (2005). Implicit sound symbolism in lexical access: Evidence from an interference task. Brain and Language, 93, 10-19.
Whalen, D.H., \& Levitt, A.G. (1995). The universality of intrinsic F0 of vowels. Journal of Phonetics, 23, 349366.

Wiedenhof, J. (2004). Grammatica van het Mandarijn. Amsterdam: Uitgeverij Bulaaq.
Yorkston, E.A., \& Menon, G. (2004). A sound idea: phonetic effects of brand names on consumer judgments. Journal of Consumer Research, 31, 43-51.

# Justified True Belief Triggers False Recall of "Knowing" 

Derek Powell (derekpowell@ucla.edu)<br>Department of Psychology, University of California, Los Angeles<br>Los Angeles, CA 90095 USA<br>Zachary Horne (horne2@illinois.edu)<br>Department of Philosophy, University of Illinois, Urbana-Champaign<br>Urbana, IL 61802 USA<br>Angel Pinillos (pinillos@asu.edu)<br>Department of Philosophy, Arizona State University<br>Tempe, AZ 85281 USA

Keith J. Holyoak (holyoak@lifesci.ucla.edu)<br>Department of Psychology, University of California, Los Angeles<br>Los Angeles, CA 90034 USA


#### Abstract

Philosophers traditionally held that knowledge is justified true belief. Gettier (1963) challenged this view with thought experiments in which someone has a justified and true belief, but an element of luck is involved that disqualifies the belief from counting as knowledge. We examined laypeople's concept of knowledge using a semantic integration paradigm modeled after that of Gentner (1981). Participants read stories in which a character 'thought' something was true. On a subsequent recall task, readers sometimes falsely recalled the verb 'thought' as 'knew,' implicitly indicating that the reader had attributed knowledge to the character. False recall of 'knew' occurred more frequently when the story described a justified true belief than an unjustified belief. Justified true belief triggered these recall errors even in a so-called "Gettier case". The present findings suggest that semantic integration provides an empirical paradigm suitable for investigating lay notions about knowledge.


Keywords: belief; knowledge; semantic integration; false memory; experimental philosophy

People's beliefs are the primary drivers of their actions, yet these beliefs are often uncertain-the products of limited information about the world and interconnections between other (often uncertain) beliefs. For this reason, a capacity for evaluating the status of different beliefs is important for individuals in directing their own rational behavior, and for predicting the behavior of others. Understanding these processes requires an analysis of the concept of knowledge: the distinction between what is known versus what is merely believed, imagined, hoped for, or assumed. Research on metacognition has examined how people assess their confidence in their own beliefs (e.g., Klayman et al., 1999; Tsai, Klayman \& Hastie, 2008), and research on theory of mind (Premack \& Woodruff, 1978) has examined how beliefs are attributed to others (e.g., Birch, 2005) but there
has been very little psychological research examining the self-assessment and attribution of knowledge.

Making a decision about when to attribute knowledge, either to oneself or to another, hinges on one's conception of knowledge: it is a decision about whether or not the concept applies in a particular instance. Although other factors may play into this decision process, understanding the exact nature of the concept itself is essential to understanding the overall process of knowledge attribution. Philosophers have long contemplated the nature of knowledge, and have also developed a variety of methods for studying concepts. A common method involves using thought experiments.

One of the most influential of these thought experiments was proposed by Edmund Gettier (1963). Named for their progenitor, "Gettier cases" challenge the traditional conception of knowledge. Prior to the 1960s, most philosophers thought that knowledge should be analyzed as justified true belief. Today, many philosophers see Gettier cases as counterexamples to that analysis. Gettier cases are situations in which an agent has a true belief that is justified, but an element of luck is involved that disqualifies their cognitive state from being considered knowledge. To illustrate such a case, suppose that at $3: 34 \mathrm{pm}$ an agent comes to believe it is $3: 34 \mathrm{pm}$ by looking at her normally reliable watch. Suppose also that unbeknownst to the agent, her watch had been stopped for exactly 24 hours-she just happened to glance at her watch at the correct time. The agent's belief is not only true, but is also justified (since looking at one's normally reliable watch is a good way to form veridical beliefs about time of day). However, most philosophers judge that this agent does not know that it is $3: 34 \mathrm{pm}$, because her belief is true only by luck. If this judgment is correct, then this case is a counterexample to the traditional thesis that knowledge is justified true belief. It remains an open question whether philosophers' conceptions of knowledge are shared by laypeople. Recently,
experimental philosophers have sought to empirically investigate laypeople's intuitive judgments about philosophical thought experiments (for a review, see Knobe et al., 2012). In particular, researchers have turned their attention to examining whether philosophers' Gettier judgments are shared by laypeople (Weinberg, Nichols \& Stich, 2001; Turri, in press). Starmans and Friedman (2012) investigated laypeople's evaluations of Gettier cases by presenting participants with short vignettes that described agents forming beliefs under different circumstances. Three different versions of each scenario were created: the agents in the vignettes either formed a false belief, formed a justified true belief, or were "Gettiered"-the belief they formed was both justified and true, but was true only by luck. Starmans and Friedman then asked participants to judge whether the agents "knew" or "only believed" the proposition in question, and to rate how confident they were in their judgment. Participants attributed knowledge to agents in Gettier cases almost as readily as they did in cases of non-Gettiered justified true belief, suggesting that laypeople's concepts of knowledge may differ from those held by many philosophers. Their findings led Starmans and Friedman to conclude that laypeople view knowledge as justified true belief, in accord with the more traditional philosophical view.

However, the survey-based methodology used by Starmans and Friedman (2012) has limitations. Answers on such surveys may be influenced by demand characteristics (Orne, 1962). For example, if participants form some interpretation of the experimenter's hypothesis, they may attempt to confirm or disconfirm this hypothesis. In addition, participants commonly display apprehension about being evaluated (Weber \& Cook, 1972). This may lead them to give responses they perceive as either socially desirable, or likely to be considered "correct," irrespective of their actual attitudes or answers. For research on folk concepts, this type of evaluative apprehension might be manifested as a sort of amateur philosophizing, or attempts to avoid being "tricked" by the experimenter.

A further concern with survey-based methods is that asking participants to make knowledge attributions fails to isolate their knowledge concepts from other decision processes that could influence their judgments downstream. This is a general problem for survey-based methods, one that applies not only to investigations of concepts for knowledge, but also to investigations of lay concepts in general. Recently, social psychologists and experimental philosophers have investigated such concepts as intentional action (e.g., Knobe et al., 2012), causation (e.g., Livengood \& Machery, 2007), and explanation (Braverman et al., 2012), but thus far their methods have been primarily survey-based. In fact, almost all work in experimental philosophy utilizes these methods.

Here we propose and test a new method for examining people's concepts, based on psychological research related to semantic integration. Semantic integration is the cognitive process by which smaller units of semantic
information are combined to form larger meaningful structured representations, or "discourse meanings," during language processing. Many researchers in cognitive psychology and psycholinguistics have investigated how these structured representations are formed and how they are stored in memory. Early research by Sachs (1967) revealed that memory for the meaning of sentences is more robust than memory for their specific wordings. During language processing, the original form of presented material is stored only temporarily, just long enough to be comprehended. Once comprehended, the material's meaning is then encoded into long-term memory. Bransford and Franks (1971) reasoned that if semantic information is integrated during language processing, and it is the meaning of a passage that is actually encoded into memory, then human memory ought to exhibit productivity. That is, it should be possible for exposure to several basic, interrelated propositions to produce false verbatim memory for more complex propositions that express their combination, even when these propositions were never themselves experienced. These sorts of productive memory errors have been taken as evidence for semantic integration across a variety of language comprehension contexts (Flagg, 1976; Owen, Bower \& Black, 1979; Gentner, 1981; Sulin \& Dooling, 1974; Thorndyke, 1976).

To explain such findings, Gentner (1981) proposed a model of language processing in which linguistic propositions are considered both individually and in the broader context of the story in which they appear. Her model assumes that when a sentence is read within the context of a larger passage, the discourse meaning that a reader forms may incorporate information not contained in the original sentence.

In evaluating this model, Gentner (1981) focused her investigations on a relatively well-analyzed area of linguistics, the meanings of verbs. She was able to make specific predictions about how manipulations of the contextual information given in a passage of text would affect later recall for verbs within that passage. To illustrate, consider the relationship between the general verb 'give' and the more specific verb 'pay'. An informal analysis suggests that 'to give' some item is to take some action that transfers ownership of that item to a recipient. 'Paying' is a more specific form of giving, in which the giver owes the recipient. In her experiments, Gentner (1981) asked her participants to read paragraph-long stories that each featured a critical sentence containing some key verb of interest. For instance, one of these stories contained the critical sentence, "Max finally gave Sam the money." Two versions of this story were created, one that contained additional context explaining that Max owed Sam money, and a control story that lacked this information. After reading one version of the story, Gentner's participants performed a recall task, in which they were shown the critical sentence with the word 'gave' removed, and asked to fill in the word that had appeared in the story. In support of her predictions, Gentner found that participants who had been provided with the
additional contextual information were more likely to falsely recall the more specific verb 'paid' as having appeared in the critical sentence than participants who had read the control story.

We aimed to turn this methodology on its head: Whereas Gentner (1981) used a false recall paradigm to examine how known semantic structures are integrated during language processing, we used a similar paradigm to examine the semantic structure of the concept 'knowledge'. Following Gentner (1981), we constructed stories containing the generic cognition verb 'thought', and used false recall of the more specific verb 'knew' as a measure of the extent to which different contexts instantiate the semantic structure of 'knowledge'. By incorporating the relevant contextual information into our stories, we can examine whether the different scenarios imagined in various thought experiments differentially activate people's concept of knowledge.

Gentner's (1981) paradigm has several qualities that are desirable for our present purposes, relative to other semantic integration tasks. First, the use of free recall makes its results more compelling than tasks that rely on recognition judgments. Participants' responses to recognition tasks can be influenced by both true recollection as well as feelings of familiarity (Tulving, 1985). In contrast, explicit recall of the word 'knew' provides strong evidence for the semantic activation of the concept. Second, this paradigm focuses responses onto a single specific word of interest, whereas other semantic integration paradigms often ask participants to evaluate larger semantic units, such as phrases or sentences. This specificity may help reduce ambiguity in investigations of individual concepts.

Experiment 1 served as a proof of concept, demonstrating that semantic integration can be used to investigate laypeople's concept of knowledge. In Experiment 2 we examined the more controversial issue of how Gettier cases activate people's concept of knowledge.

## Experiment 1

## Method

In Experiment 1 we constructed two similar stories about a detective investigating a crime. In the first story, the detective forms the justified true belief (JTB condition) that his suspect is guilty: the omniscient narrator reveals that the detective's suspect committed the crime, and the detective uncovers evidence that his suspect is guilty. In the second story, the detective forms the unjustified belief (UB) that his suspect is guilty: he cannot find any useful evidence linking his suspect to the crime, and the narrator does not reveal whether the suspect is guilty. Following Gentner (1981), each story included a critical sentence with a critical word (italics added here for emphasis): "Whatever the DA's reservations, Dempsey thought Will was guilty." The critical word in both stories was 'thought', a generic cognition verb that could plausibly be recalled as 'knew', should the right conditions be met. This critical sentence was later presented with a blank in place of the critical word,
and participants were asked to recall the word that appeared in the story.

Based on prior research involving semantic integration and false recall paradigms (Gentner, 1981), we predicted that participants who read the story in which Dempsey's belief is justified and true would be more likely to falsely recall the word "knew" than participants who read the story in which Dempsey's belief is unjustified. Of course, it will not come as a surprise if a justified true belief more closely resembles people's concept of knowledge than an unjustified belief. Experiment 1 examined this simple case in order to demonstrate the potential of the semantic integration paradigm.

Participants This experiment was conducted online, with 147 participants ( 91 female) recruited from Amazon's Mechanical Turk (mTurk). The mean age of the participants was 34 years. They were all paid $\$ 0.50$ for their participation.

Procedure Participants were randomly assigned to the JTB and UB conditions, and were asked to read the corresponding story about the detective. Then, participants completed a distraction task, reading an approximately 1000 -word selection from a fictional article on gamma ray bursts (taken from Waskan et al., under review). Timing controls ensured that participants spent an adequate amount of time attending to each section of the experiment.

In the recall task, participants were shown five sentences from the detective story, each missing one word that was replaced with an underscored blank space. They were instructed to type in the word that originally appeared in the story. The critical sentence was always presented first.

After the recall task, participants were asked a direct question to assess their understanding of whether Dempsey had knowledge or not. Following Starmans and Friedman (2012), they were asked: "Would you say that Dempsey knew Will was guilty, or only thought Will was guilty?" They indicated their choices as 'knew' or 'thought,' and then rated their confidence on a 1-5 Likert scale.

## Results and Discussion

Recall task Participants' responses during the recall task were classified as either 'thought'-type responses or 'knew' responses. Words and phrases synonymous with 'thought' but neither stating nor implying knowledge were grouped together as 'thought'-type responses. In an effort to remain conservative, only the word 'knew' was counted toward the tally of 'knew' responses. Responses that were nonsense (i.e., were not verbs, were random letters typed in the blank, etc.) were excluded from analysis. After these exclusions, 64 participants remained in the JTB condition and 65 participants remained in the UB condition.

As predicted, participants recalled the word 'knew' significantly more often when they were assigned to the JTB condition than when they were assigned to the UB condition ( $39 \%$ vs. $18 \% ; X^{2}(1)=5.72, p=.016$ ). This
finding demonstrates that participants semantically integrated contextual information, specifying that Dempsey's belief was both true and justified, with the generic cognition verb 'thought,' leading to false recall of the more specific verb 'knew.'

Knowledge survey question Following Starmans and Friedman (2012), participants' 'knew' and 'thought' responses were assigned scores of 1 and -1 respectively, and these values were multiplied by the confidence ratings participants reported to produce a knowledge rating score. Knowledge ratings from participants in the JTB condition (-1.40) were significantly higher than in the UB condition $(-3.66), t(127)=6.94, p<.001$.

## Experiment 2

## Method

In Experiment 2 we used the semantic integration paradigm to examine a more substantive question about knowledge. Specifically, we investigated the extent to which Gettier cases activate people's concept of knowledge. Experiment 2 used three stories, adapted from the detective stories of Experiment 1.

In the first story, one character "Will" is guilty of a crime and "Dempsey," the detective in the story, finds authentic evidence of his guilt, forming the justified true belief that he is guilty (JTB condition). Meanwhile, another character "Beth", who is Will's girlfriend, observes the sequence of events that unfold and result in Dempsey thinking that Will is guilty. In the second story, Will is innocent of the crime, but is framed by his girlfriend Beth because she suspects that he is cheating on her. Dempsey finds evidence planted by Beth, and as a result forms the false belief that Will is guilty of the crime (FB condition). Finally, in the third story, Will is guilty of the crime, but he has eliminated all the authentic evidence of his crime. Beth, as part of a ploy to seek reprisals against Will, plants evidence that implicates him in the crime. Dempsey finds this evidence and forms the belief that Will is guilty. In this case, Dempsey's belief is both justified and true, but is only true by chance (Gettier condition). In each of these stories, the critical sentence read, "Whatever the ultimate verdict would be, Dempsey thought Will was guilty" (italics added here for clarity).

Importantly, the evidence Dempsey uncovered was the same in each version of the story. Moreover, the structure and wording of the stories was identical-save for the relevant manipulations-and the stories were closely matched in overall length.

Participants Experiment 2 was also conducted online, with 304 participants ( 164 female) recruited from mTurk. The mean age of participants was 31 years. All participants were paid $\$ 0.50$ for their participation.

Procedure Participants were randomly assigned to one of the three conditions, and were asked to read the
corresponding story about the detective. The distractor and recall tasks for Experiment 2 were the same as those in Experiment 1. After the recall task, participants were asked the same question about Dempsey's knowledge as in Experiment 1, and were also asked, "Should Dempsey have arrested Will?" Participants rated their confidence for both responses. After these questions, participants answered a pair of comprehension questions to ensure they had attended to central details of the story.

## Results and Discussion

Recall task Some participants were excluded from analysis after failing the reading comprehension check. Others gave ambiguous free recall responses that did not fit into either the 'knew' or 'thought' response categories. After these exclusions, 259 participants remained in the final analysis. Participants' recall responses were classified according to the same criteria as in Experiment 1. Figure 1 shows the proportion of 'knew' responses in the three conditions.

False recall of 'knew' was observed significantly more often in the JTB ( $42 \%$ ) and Gettier ( $47 \%$ ) conditions as compared with the FB ( $23 \%$ ) condition $\left(X^{2}(1)=5.94, p\right.$ $=.015$, and $X^{2}(1)=9.63, p<.01$, respectively). No significant difference was found between the frequencies of 'knew' recall in the JTB and Gettiered conditions ( $X^{2}(1)$ $=.20, p=.66)$. Participants thus seemed to believe that agents in Gettier cases possess knowledge, and apparently drew no distinction between Gettier cases and non-Gettier cases of justified true belief. This finding agrees with that reported by Starmans and Freidman (2012), and stands in contrast to how philosophers have understood the implications of Gettier cases.

Survey questions Participants' knowledge rating scores were calculated as in Experiment 1. This same procedure was also employed with the "arrest" question to calculate an action rating score, where a positive score indicated that participants endorsed Dempsey's arresting Will.


Figure 1: Percent false recall of 'knew' in critical sentence across conditions

Participants' knowledge ratings differed significantly between conditions, $F(2,256)=18.12, p<.001$. Post hoc tests using Bonferroni corrections indicated that participants' knowledge ratings in both the JTB (0.62) and Gettier conditions ( -0.11 ) were higher than those in the FB condition ( $-2.56 ; p<.001$ ), but that knowledge ratings did not differ significantly between the JTB and Gettier conditions ( $p=.65$ ). This finding is consistent with the results of the recall task, and replicates the pattern of results reported by Starmans and Friedman (2012). It is worth noting, however, that participants' actual knowledge rating scores are considerably lower in the present experiment than in that reported by Starmans and Friedman. This is likely due to the different experimental materials used, and in particular the quality of evidence depicted in the different stories: Starmans and Friedman's vignettes feature direct perceptual evidence, whereas our materials describe weaker physical and testimonial evidence.

Significant differences between conditions were also found for action ratings, $F(2,256)=52.81, p<.001$, where planned comparisons revealed the same pattern as for knowledge ratings: participants action ratings differed between the JTB (3.26) and FB ( -0.12 ) conditions ( $p<.001$ ), and between the Gettier (3.80) and FB conditions ( $p<.001$ ), but not between the JTB and Gettier conditions ( $p=.69$ ). Interestingly, participants endorsed Dempsey's action (arresting Will) even when Dempsey had been Gettiered. A number of philosophers hold that knowledge is intimately connected to action (Hawthorne \& Stanley, 2008). In particular, they hold that it is intuitive that if a person knows some proposition, then it is acceptable for them to use that proposition in reasoning and in action. Our result is consistent with this thesis: in conditions where participants say that Dempsey knows Will is guilty, they also tend to say that he should have arrested Will.

## General Discussion

## Summary

Experiment 1 provided a proof-of-concept that semantic integration tasks can be used to examine people's concept of knowledge. As expected, a story in which a character forms a justified true belief activated people's concept of knowledge more strongly than a story in which this same character forms an unjustified belief, as evidenced by an increase in false recall of the verb 'knew' in place of 'thought.'

The results of Experiment 2 corroborate Starmans and Friedman's (2012) findings on laypeople's reactions to Gettier cases, while avoiding alternative interpretations that might be raised with survey-based methods. As our participants believed they were completing a memory task, it is very unlikely that their responses were affected by unwanted demand characteristics, or are indicative of some sort of performance error (Kauppinen, 2007; Cullen, 2010). Rather, their responses presumably reflect their concept of knowledge, which apparently differs from the concept
developed by philosophers who have considered Gettier cases.

These results have important implications for philosophical research. Most philosophers have accepted that Gettier cases are not instances of knowledge, and thus that knowledge is not equivalent to justified true belief. At the same time, many of them also assume that when they analyze the concept of knowledge, they are investigating a concept that is shared by laypersons (or at least that lay judgments inform philosophical theories of knowledge). In light of our findings, as well as those of Starmans and Friedman (2012), it appears that philosophers may need to reconsider their assumptions (but see Nagel, San Juan \& Mar, 2013, for evidence that different variants of Gettier cases may yield divergent findings).

The present study demonstrates that semantic integration tasks provide a promising methodology for empirically investigating lay concepts, avoiding many of the pitfalls associated with survey-based methods. Semantic integration tasks minimize the likelihood that participants' responses are affected by demand characteristics, and make it possible to isolate the activation of concepts from downstream decision processes.

## Directions for Future Research

The fact that 'knowledge' has a verb form, as well as approximate near-synonyms such as 'thought' and 'believed', allowed us to model our investigations of knowledge directly on Gentner's (1981) research on the semantic integration of verb meanings. Of course, not all concepts of interest to psychologists and philosophers will necessarily exhibit these desirable traits. Where this is not the case, other semantic integration tasks may be more appropriate. For example, these constraints would not apply to semantic integration tasks measuring recognition for sentences or phrases (e.g., Bransford \& Franks, 1971; Owens, Bower \& Black, 1979). As described earlier, this type of task presents some disadvantages (e.g., increased ambiguity from the assessment of recognition memory over larger semantic units). However, these disadvantages are not insurmountable. In particular, employing a remember-know procedure (Tulving, 1985) could help distinguish between genuine recollection and familiarity. With sufficient care, it should be possible to craft phrases or sentences that unambiguously express whatever concept may be of interest to researchers.

The present investigation demonstrates the need for empirical research on knowledge attribution and knowledge concepts, and also illustrates a powerful method that may be applicable in future investigations. Further research is needed to explore both the implications of these early findings on Gettier cases, as well as other factors that are potentially relevant to knowledge, such as the salience of error (e.g., Schaffer \& Knobe, 2012) and agents' practical interests (e.g., Pinillos, 2012). Moreover, semantic integration tasks may prove useful for the study of other
philosophical concepts, including intention (Knobe et al., 2012) and causation (Livengood \& Machery, 2007).

## Acknowledgments

Preparation of this paper was supported by grant N000140810186 from the Office of Naval Research.

## References

Birch, S. A. J. (2005). When knowledge is a curse. Current Directions in Psychological Science, 14, 25-29.
Bransford, J. D., \& Franks, J. J. (1971). The abstraction of linguistic ideas. Cognitive Psychology, 2, 331-350.
Braverman, M., Clevenger, J. Harmon, I., Higgins, A., Horne, Z., Spino, J., \& Waskan, J. (2012). Intelligability is necessary for explanation but accuracy may not be. In N. Miyake, D. Peebles \& R. P. Cooper (Eds.), Proceedings of the Thirty-Fourth Annual Conference of the Cognitive Science Society (pp. 1368-1373). Austin, TX: Cognitive Science Society.
Cullen, S. (2010). Survey-driven romanticism. Review of Philosophy and Psychology, 1, 275-296.
Flagg, P. (1976). Semantic integration in sentence memory? Journal of Verbal Learning and Verbal Behavior, 15, 491-504.
Gentner, D. (1981). Integrating verb meanings into context. Discourse Processes, 4, 349-375.
Gettier, E. L. (1963). Is justified true belief knowledge? Analysis, 23, 121-123.
Hawthorne, J., \& Stanley, J. (2008). Knowledge and action. Journal of Philosophy, 105, 571.
Kauppinen, A. (2007). The rise and fall of experimental philosophy. Philosophical Explorations, 10, 95-118.
Klayman, J., Soll, J., González-Vallejo, C., \& Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. Organizational Behavior and Human Decision Processes, 79, 216-247.
Knobe, J., Buckwalter, W., Nichols, S., Robbins, P., Sarkissian, H., \& Sommers, T. (2012). Experimental philosophy. Annual Review of Psychology, 63, 81-99.
Livengood, J., \& Machery, E. (2007). The folk probably don't think what you think they think: Experiments on causation by absence. Midwest Studies in Philosophy, XXXI, 107-127.
Nagel, J., San Juan, V., \& Mar, R. A. (2013). Lay denial of knowledge for justified true beliefs. Cognition.
Orne, M. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. American Psychologist, 17, 776-783.
Owens, J., Bower, G., \& Black, J. (1979). The "soap opera" effect in story recall. Memory \& Cognition, 7, 185-191.
Pinillos, A. (2012). Knowledge, experiments, and practical interests. In J. Brown \& M. Gerkin (Eds.), Knowledge
ascriptions (pp. 192-219). New York: Oxford University Press.
Premack, D., \& Woodruff, G. (1978). Does the chimpanzee have a theory of mind? Behavioral and Brain Sciences, 1, 515-526.
Sachs, J. (1967). Recognition memory for syntactic and semantic aspects of connected discourse. Perception \& Psychophysics, 2, 437-442.
Schaffer, J., \& Knobe, J. (2012). Contrastive knowledge surveyed. Noûs, 46(4), 675-708.
Starmans, C., \& Friedman, O. (2012). The folk conception of knowledge. Cognition, 124, 272-283.
Sulin, R. A., \& Dooling, D. J. (1974). Intrusion of a thematic idea in retention of prose. Journal of Experimental Psychology, 103, 255-262.
Thorndyke, P. (1976). The role of inferences in discourse comprehension. Journal of Verbal Learning and Verbal Behavior, 15, 437-446.
Tsai, C., Klayman, J., \& Hastie, R. (2008). Effects of amount of information on judgment accuracy and confidence. Organizational Behavior and Human Decision Processes, 107, 97-105.
Tulving, E. (1985). Memory and consciousness. Canadian Psychology/Psychologie Canadienne, 26, 1-12.
Turri, J. (in press). A conspicuous art: Putting Gettier to the test. Philosopher's Imprint.
Waskan, J., Clevenger, J., Harmon, I., Horne, Z., \& Spino, J. (under review). Explanatory anti-psychologism overturned by lay and scientific case classifications.
Weber, S., \& Cook, T. (1972). Subject effects in laboratory research: An examination of subject roles, demand characteristics, and valid inference. Psychological Bulletin, 77, 273-295.
Weinberg, J. M., Nichols, S., \& Stich, S. (2008). Normativity and epistemic intuitions. Experimental Philosophy, 29, 17-45.

# Generic Priors Yield Competition Between Independently-Occurring Causes 

Derek Powell ${ }^{1}$ (derekpowell@ucla.edu)

M. Alice Merrick ${ }^{1}$ (m.a.merrick@gmail.com)<br>Hongjing Lu ${ }^{1,2}$ (hongjing@ucla.edu)<br>Keith J. Holyoak ${ }^{1}$ (holyoak@lifesci.ucla.edu)<br>Departments of Psychology ${ }^{1}$ and Statistics ${ }^{2}$, University of California, Los Angeles<br>Los Angeles, CA, USA


#### Abstract

Recent work on causal learning has investigated the possible role of generic priors in guiding human judgments of causal strength. One proposal has been that people have a preference for causes that are sparse and strong-i.e., few in number and individually strong (Lu et al., 2008). Evidence for the use of sparse-and-strong priors has been obtained using a maximally simple causal set-up (a single candidate cause plus unobserved background causes). Here we examine the possible impact of generic priors in more complex, multicausal set-ups. Sparse-and-strong priors predict that competition can be observed between candidate causes even if they occur independently (i.e., the estimated strength of cause A will be lower if the strength of uncorrelated cause B is high rather than low). Experiment 1 revealed such a cue competition effect in judgments of causal strength. Experiment 2 showed that, as predicted by a Bayesian learning model with sparse-and-strong priors, the impact of the prior diminishes as sample size increases. These findings support the importance of a preference for parsimony as a constraint on causal learning.


Keywords: causal learning; generic priors; causal strength; parsimony; Bayesian modeling

## Introduction

## Prior Beliefs in Causal Learning

Humans (and other intelligent organisms) are able to extract causal knowledge from patterns of covariation among cues and outcomes. Viewed from a Bayesian perspective, causal inferences are expected to be a joint function of likelihoods (the probability of observing the data given potential causal links of various possible strengths) and priors (expectations about causal links that the learner brings to the task). For relatively simple causal set-ups involving binary variables, human causal judgments can be described quite accurately by the power PC theory (Cheng, 1997), which uses a noisyOR likelihood function to integrate the influences of multiple generative causes (Griffiths \& Tenenbaum, 2005; Lu et al., 2008; see Holyoak \& Cheng, 2011, for a review).

Prior beliefs about causal relationships can also be formulated within a Bayesian framework for causal learning. Generic causal priors can be thought of as preferences for certain types of causal explanations, without
relying on domain-specific knowledge. Some Bayesian models have assumed uninformative priors (e.g., Griffiths \& Tenenbaum, 2005); however, other models have incorporated substantive generic priors about the nature of causes. In particular, Lu et al. (2008) proposed that people have a preference for causes that are sparse and strong: i.e. a preference for causal models that include a relatively small number of strong causes (rather than a larger number of weak causes). Such priors can be viewed as a special case of a more general pressure to encourage parsimony (Chater \& Vitanyi, 2003), which implies a combination of simplicity and explanatory power. The preference for parsimony has a number of expressions elsewhere in causal learning phenomena and theory. For instance, causal learners appear to make the default assumption that causes act independently in producing an effect, rather than interacting (Cheng, 1997; Novick \& Cheng, 2004). Moreover, people generally prefer simpler explanations to equally accurate but more complex explanations (Lombrozo, 2007).

## Generic Prior: Sparse-and-Strong (SS) Causes

Lu et al. (2008) formalized the "SS power" model with sparse-and-strong (SS) priors for simple causal models with a single candidate cue and a constantly-present background cause. When the candidate cause generates (rather than prevents) the effect, there is an expectation that the candidate cause is strong (strength close to 1) and the background is weak (strength close to 0 ), or vice versa. A single free parameter, $\alpha$, controls the strength of the prior (when $\alpha=0$, the distribution is uniform).

The possible role of generic priors in causal strength
Table 1. Contingency learning data for one experimental block (44 trials) by trial type

| Conditions |  | C | AC | BC | ABC |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Weak-B | E absent | 1 | 6 | 3 | 7 |
|  | E present | 10 | 5 | 8 | 4 |
| Strong-B | E absent | 1 | 6 | 9 | 10 |
|  | E present | 10 | 5 | 2 | 1 |



Figure 1: Sparse-and-strong prior distribution over causal strengths of three causes. Colors indicate the values of prior probability (red corresponds to highest probability).
judgments has so far only been examined for very simple causal graphs (e.g., one generative candidate cause and a constantly-present background cause). Lu et al. (2008) fit several causal learning models to parametric data for human strength judgments. They found the best fit was provided by a Bayesian implementation of the power PC theory that incorporated SS priors with an $\alpha$ value of 5 (not 0 ), implying a human preference for sparse-and-strong causes. When $\alpha$ value is set to zero, the prior distribution would be equivalent to a uniform distribution.

The generalization of the sparse-and-strong prior for more than two causes is straightforward. For the experiments reported here, the SS prior is constructed on the basis of three candidate causes, $A, B$, and $C$, which are all generative. The SS prior can be defined as,

$$
\begin{align*}
& P\left(w_{A}, w_{B}, w_{C}\right) \propto \\
& \quad e^{-\alpha\left(1-w_{A}\right)-\alpha w_{B}-\alpha w_{C}}+e^{-\alpha w_{A}-\alpha\left(1-w_{B}\right)-\alpha w_{C}}+e^{-\alpha w_{A}-\alpha w_{B}-\alpha\left(1-w_{C}\right)} \tag{1}
\end{align*}
$$

in which $w$ denotes causal strength for different causes.
Figure 1 illustrates the sparse-and-strong prior in the three-cause situation. A signature of SS priors is the preference for one strong cause coupled with other weak causes, i.e., a set of "ideal" causal strengths for the three causes might be $w_{\mathrm{A}}=1, w_{\mathrm{B}}=0$ and $w_{\mathrm{C}}=0$. This preference instantiated in SS priors implies a key empirical prediction: competition effects in judgments of causal strength when multiple causes co-occur. Strength competition implies that if a candidate cause appears along with another cause of greater strength (as defined by likelihoods), then the strength of the weaker candidate cause will be underestimated. This prediction goes beyond competition effects predicted by the likelihood function alone (i.e., a model assuming uninformative priors). The goal of the present paper is to test this key empirical prediction in a situation requiring inference based on multiple causes.

## Competition Between Causes

Various competitive dynamics are commonly observed in causal learning paradigms, including blocking (e.g., Shanks, 1985), overshadowing (e.g., Waldmann, 2001) and unovershadowing (De Houwer \& Becker, 2002). However, in all these paradigms the competition is between cues that cooccur in a systematic way. For example, blocking is typically obtained when cue A is first shown to produce the effect by itself, and then the compound cue $A B$ is introduced and also followed by the effect. From a Bayesian perspective, a lower causal strength judgment for the blocked cue, B , is entirely rational, as the learner has no opportunity to observe what happens when $B$ is presented without A (i.e., there will be greater uncertainty about the strength of B than of A). More generally, Bayesian models with uniform priors can readily account for a wide range of competition effects that involve cues occurring in a correlated fashion (Carroll, Cheng \& Lu, in press).

However, sparse-and-strong priors are unique in predicting competition between independently-occurring causes (e.g., the occurrence of cue A is uncorrelated with the occurrence of cue B). We will show simulation results confirming that when alternative causes A and B occur independently, a Bayesian model with uniform priors predicts that judgments of the strength of $A$ will not be influenced by the strength of $B$, or vice versa (also see Busemeyer, Myung \& McDaniel, 1993a). In contrast, an otherwise-identical model incorporating sparse-and-strong priors predicts that early in learning (when the impact of priors is maximal), independently-occurring causes will compete for strength (e.g., the strength of A will be judged to be lower if $B$ is strong rather than weak).

The present experiments include two conditions based on a set of contingency data, $D$, shown in Table 1. The occurrences of causes A and B are independent in both conditions. The causal power of A is held constant across the two conditions ( 0.5 ), but the causal power of B varies from one condition ( 0.2 , weak- $B$ condition) to the other ( 0.8 , strong-B condition).

For this set of contingency data the model computes the mean of estimated causal strength derived from the posterior distribution:

$$
\begin{equation*}
\bar{w}_{A}=\int_{0}^{1} w_{A} P\left(w_{A} \mid D\right) \tag{2}
\end{equation*}
$$

The posterior distribution $P\left(w_{A} \mid D\right)$ is obtained by applying Bayes rule to combine likelihood function and priors as

$$
\begin{equation*}
P\left(w_{A} \mid D\right)=\iint \frac{P\left(D \mid w_{A}, w_{B}, w_{C}\right) P\left(w_{A}, w_{B}, w_{C}\right)}{P(D)} d w_{B} d w_{C} . \tag{3}
\end{equation*}
$$

In our simulations, we employed the noisy-OR likelihood function (Cheng, 1997), since binary causes and effects were used in the experiments:

$$
\begin{aligned}
& P\left(E=1 \mid C_{A}, C_{B}, C_{C} ; w_{A}, w_{\boldsymbol{B}}, w_{C}\right) \\
& \quad=1-\left(1-w_{A} C_{A}\right)\left(1-w_{\boldsymbol{B}} C_{B}\right)\left(1-w_{C} C_{C}\right)(4)
\end{aligned}
$$

where $E$ and $C$ indicate the presence or absence of effect and causes, respectively.

Figure 2 shows the model predictions for causal strength of A in the two conditions. The Bayesian model with SS prior (center bards in Figure 3) predicts different estimates of $w_{\mathrm{A}}$ across conditions due to competition between causes A and B, even though the two cues occur independently. In contrast, a model with uniform prior (right bars in Figure 3) predicts that $w_{\mathrm{A}}$ will not vary across the two conditions. The latter simulation result confirms that a Bayesian model with uniform priors does not predict competition between independently-occurring causes when the likelihood function is a noisy-OR, extending the similar negative conclusion for the case in which the likelihood function is linear (Busemeyer et al., 1993a).

Testing these opposing predictions provides a means to discriminate between alternative possible priors for causal inference with multiple cues. The prediction of competition between independently-occurring causes has never been clearly tested. Busemeyer et al. (1993b) reported an experiment that obtained competition between uncorrelated cues, in a paradigm that may have drawn on causal learning mechanisms. However, this competition effect was observed only when participants were informed that the two cues would be of different strengths, one strong and one weak (see their Footnote 5, p. 194). It is possible that this instruction suggested to subjects that the cues were expected to be competitive. In general, relatively few studies of causal learning have used complex causal set-ups involving more than one or two candidate causes. The present experiments were designed to determine whether multiple candidate causes would compete for causal strength, and whether such effects can be modeled by assuming people have priors that causes will be sparse and strong.

## Experiment 1

## Method

Participants Participants were 90 undergraduate students at the University of California at Los Angeles (UCLA) who participated for class credit ( $80 \%$ female, mean age $=20$ years). Half were assigned to the strong- $B$ condition and half to the weak-B condition.

Procedure Participants read a cover story, as follows: "Imagine that you are assisting a doctor at a new island resort. Many of the guests at this new resort have become ill, and you are charged with helping to determine the cause
of the illnesses. Every day, at dinner, the resort provides a complimentary salad for its guests. The salads can be made with different exotic vegetables. The salads always have at least one exotic vegetable, and can be ordered with up to three different exotic vegetables. The resort's doctor thinks one or perhaps several of these exotic vegetables may be causing the illness <pictures of three vegetables are shown>. You will be reviewing a number of case files that describe what a guest ate and whether they became sick. After viewing these files you will be asked to give your assessment of which vegetable or vegetables are the culprits. Please pay attention to each case.... When you are done reviewing the cases you will be asked to estimate how many people each vegetable is likely to affect negatively."
These vegetables were labeled $\mathrm{A}, \mathrm{B}$, and C , and were shown as photographs of exotic vegetables (see Figure 2, top). These photographs depicted the actual vegetables radicchio, bitter melon, and black garlic. The assignment of vegetables to the labels $\mathrm{A}, \mathrm{B}$ and C was randomized across participants. During the learning phase, participants viewed "case files" for individual guests, showing which combination of vegetables they had eaten, and whether or not they had fallen ill.

There were four possible combinations of fruits: each guest had either eaten vegetable $C$ alone, vegetables $A$ and $C$, vegetables $B$ and $C$, or all three vegetables $A, B$, and $C$. These four combinations were presented in equal number, such that A and B both occurred 50 percent of the time, and the correlation between the occurrence of A and B was 0 . A total of 44 cases ( 11 of each type) was the minimum number required to reflect the underlying causal powers in the presented distribution of cause combinations and their associated outcomes.

The numbers of guests who became sick after eating each combination were determined by the causal powers assigned to each vegetable, calculated according to the noisy-OR likelihood function under the default assumption that each cause acts independently to produce the effect (Cheng, 1997). In both conditions, vegetable $A$ was assigned a causal strength of .50 , and vegetable $C$ was assigned a causal strength of .10. In the strong-B condition, vegetable B was assigned a causal strength of .80 , whereas in the weak-B condition, vegetable $B$ was assigned a causal strength of .20 . Cause A was the focus of the study, as we were interested in whether its judged strength would be influenced by the variation in the strength of cause B. Cause C served as an observable "background" cause, as it was shown to be present on every trial. The resulting contingency data is summarized in Table 1.

The 44 cases were presented sequentially in a different random order for each participant. After viewing all 44 learning trials, participants were asked to give a causal strength rating for all three vegetables. Participants were shown a picture of each vegetable along with text that read, "Imagine 100 healthy people ate this vegetable; how many do you estimate would get sick?" Participants then made


Figure 2: Example trial showing a guest who ate A, B, and C vegetables and became sick (top). Example response trial (bottom).
their rating using a slider, inputting a value between 0 and 100 (see Figure 3, bottom). The order of the three questions was randomized for each participant. After making all three ratings, participants were shown a summary of their responses and were asked to confirm that they had correctly entered their ratings. Participants were randomly assigned to one of two experimental conditions (weak-B or strong-B).

## Results and Discussion

Data from two participants were excluded due to technical issues. Data from another eight were excluded because they entered responses of zero to both cause A and cause B, suggesting errors or a lack of engagement with the task. Figure 4 shows the data for the critical A cue, along with the predictions derived from the SS power model and an otherwise-identical model with uniform priors. Participants in the strong-B condition underestimated the strength of cause A relative to participants in the Weak-B condition (mean of 34.05 versus 46.95 ), $t(79)=2.17, p<05$. The data

Table 2. Observed human strength ratings ( $0-100$ scale) and predictions based on sparse-and-strong (SS) priors for three different cues in Experiment 1.

|  | $\mathrm{A}(.50)$ |  |  | $\mathrm{B}(.20, .80)$ |  | $\mathrm{C}(.10)$ |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- |
|  | Pred. | Obs. | Pred. | Obs. | Pred. | Obs. |
| Weak-B | 52 | 47 | 16 | 34 | 13 | 18 |
| Strong-B | 35 | 34 | 78 | 63 | 16 | 18 |



Figure 3: Observed human strength ratings ( $0-100$ scale) and predictions based on sparse-and-strong (SS) versus uniform priors for cause A (0-100 scale) in Experiment 1.
were fit using Lu et al.'s (2008) "SS power" model, which provides a Bayesian formalization of sparse-and-strong priors. For modeling purposes we simply set $\alpha=5$ (the value estimated for the data sets reported by Lu et al., 2008), thus avoiding any need to fit a free parameter to the present data. The SS power model predicts the observed difference in the judged strength of $A$ in the weak- $B$ versus strong-B conditions, whereas the model with uniform priors does not.

Table 2 presents the mean ratings of causal strength obtained for three different cues, and Figure 3 plots the human data with predictions from the two models assuming different priors. Across all cues and conditions, the SS power model provides a good overall fit to the human data ( $R=.95$, root mean square deviation, $\mathrm{RMSD}=9.1$ ).

Although the overall fit of the SS power model is quite good, it bears noting that the predictions of the SS power model for cue $B$ were more extreme than the estimates given by participants. That is, when B was weak participants overestimated its strength relative to the model with SS priors; when B was strong participants underestimated it relative to SS priors. The estimates of the model using uniform priors deviate from the observed data in a similar (though marginally smaller) fashion. We speculate that these discrepancies may be due to memory limitations. Whereas the models assume perfect memory for contingency data, participants are likely to forget presented instances on some proportion of the trials, and therefore to have greater uncertainty in their strength estimates than predicted by the models. The models' estimates are computed from the mean of the posterior distribution, so increased uncertainty would lead to less extreme strength estimates for cue B (i.e., estimates closer to 50). Uncertainty would be expected to have less impact on estimates for cue A, for which the veridical strength in fact corresponds to a rating of 50 .

## Experiment 2

It is a natural feature of Bayesian models that the influence of priors diminishes as learners gather more data. Thus, the SS power model (Lu et al., 2008) predicts that competition between causes should be strongest when participants have made few observations, and will diminish as participants are exposed to more data.

Experiment 2 examined competition between causes after varying amounts of experience. The design was identical to that of Experiment 1, but added a second independent variable: sample size. Participants in both strong- and weakB conditions were asked to make judgments of causal strength three times, after viewing 44, 88 and finally 132 total cases. This resulted in a $2 \times 3$ factorial design, with one between-subjects factor (causal strength of cue B) and one within-subjects factor (number of cases observed).

The cover story was the same as it in Experiment 1, except for one sentence: "The resort's doctor thinks one or perhaps several of these exotic vegetables may be causing the illness" (Experiment 1) was revised to read, "The resort's doctor thinks these exotic vegetables may be causing the illness."

## Method

Participants Participants were 114 UCLA undergraduate students who participated for class credit ( $76 \%$ female, mean age $=20$ years).

Procedure Experimental materials were identical to those used in Experiment 1. Participants in Experiment 2 went through three blocks of learning trials, making causal strength estimates after 44, 88 and 132 learning trials. The distribution of types of cases (combinations of causes and outcome) were identical within each block. Order of presentation was randomized for each participant.

## Results and Discussion

One participant gave the same response on every trial, and six responded with extreme ratings of 0 or 100 for cause A, or ratings of 100 for Cause C. Data from these seven participants were excluded from analyses.

Figure 4 shows mean causal strength ratings for each vegetable at the end of each of the three learning blocks. A factorial repeated-measures ANOVA found no overall effect of increasing sample size, $F(2,210)=2.31, p=.10$, or of condition, $F(1,105)=1.16, p=.28)$. However, the analysis revealed a significant interaction between condition and learning block, $F(2,210)=5.61, p<.01$. As shown in Figure 4, Experiment 2 replicated the competition effect observed in Experiment 1 after 44 trials. After the first block, participants underestimated the strength of A when it was paired with a strong B cause, relative to when it was paired with a weak B cause (means of 56 versus $46 ; t(105)=$ 2.26, $p<.05$ ). As predicted, this difference disappeared with


Figure 4. Observed human strength ratings ( $0-100$ scale, top) and predictions of SS power model (bottom) for cause A across blocks in Experiment 2.
an increase in sample size, supporting the hypothesis that the observed competition effect is due to people's priors. The effect of the strength of $B$ on ratings of $A$ was not significant after 88 or 132 trials, $t(105)=0.75, p=.46$, and $t(105)=-0.26, p=.79$, respectively. Assuming $\alpha=5$ as before, the SS power model ( Lu et al., 2008) provides a good fit to the human data across all cues and conditions ( $R$ $=.96, \mathrm{RMSD}=12.87$ ). For cause A , the human data for the weak-B and strong-B conditions converge on the veridical value (50) more quickly than does the model's predictions (see Figure 4), perhaps reflecting the additional uncertainty participants experienced due to their fallible memory for the observations.

Causal strength estimates for all three vegetables were somewhat higher in Experiment 2 than Experiment 1. This difference may be due to a small change in instructions, which in Experiment 2 emphasized the doctor's belief that the vegetables were indeed causing the illness.

## General Discussion

The experiments reported here provide evidence for competition between independently-occurring causes in causal strength judgments, as predicted by a Bayesian
model of causal learning that assumes sparse-and-strong priors. After participants had made a relatively small number of observations, a cause of moderate strength was judged to be weaker when a competing (but uncorrelated) cause was strong than when the competing cause was weak. After additional cases were presented, the two conditions converged. This competition dynamic cannot be explained by naïve Bayesian models that assume uninformative priors (Busemeyer et al., 1993a), nor can such models explain why the competition effect diminishes as data is accumulated. The present results support the hypothesis that causal learners have generic prior expectations about causal relationships, and that a sparse-and-strong prior accurately characterizes these expectations.

The experiments presented here go beyond most previous investigations on causal learning by examining a more complex causal situation, one that included three observed generative causes. Examining a causal situation with multiple causes enabled a novel test of predictions that discriminated between alternative possible priors. Moreover, the relatively complex situation examined here may be more representative of the actual situations that causal learners encounter in the real world.

Using an iterative-learning method, Yeung and Griffiths (2011) empirically derived a different (but non-uniform) prior that was suggestive of a preference for strong causes, but that lacked the competitive pattern associated with the sparse prior. However, since the iterative method did not fully converge for the background cause, their results are open to multiple interpretations. Our task with multiple cues may provide a good way to further evaluate the generalization of empirical priors derived from the iterativelearning paradigm.

Lu et al. (2008) formalized sparse-and-strong priors for both generative and preventive causes. However, the preference for "sparseness" only applies across causes of the same polarity. In the generative case, sparseness is an influential factor even for simple causal set-ups, in which a single observed cause competes with an unobserved background cause (assumed by default to be generative). However, in the preventive case of the sparse-and-strong prior, competition dynamics are not evident when there is only a single preventive cause, as the observed cause is preventive whereas the background cause remains generative. The influence of sparseness, and hence the possibility of competition, is also predicted to arise in complex causal situations involving multiple preventers. As previous investigations have only examined the simplest cases, further research with more complex causal set-ups is needed to examine the possible impact of sparse-and-strong priors for preventive causes.

## Acknowledgments

Preparation of this paper was supported by grant N000140810186 from the Office of Naval Research.

## References

Busemeyer, J. R., Myung, I. J., \& McDaniel, M. A. (1993a). Cue competition effects: Theoretical implications for adaptive network learning models. Psychological Science, 4, 196-202.
Busemeyer, J. R., Myung, I. J., \& McDaniel, M. A. (1993b). Cue competition effects: Empirical tests of adaptive network learning models. Psychological Science, 4, 190195.

Carroll, C., Cheng, P. W., \& Lu, H. (in press). Inferential dependencies in causal inference: A comparison of beliefdistribution and associative approaches. Journal of Experimental Psychology: General.
Chater, N., \& Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? Trends in cognitive sciences, 7, 19-22.
Cheng, P. W. (1997). From covariation to causation: A causal power theory. Psychological Review, 104, 367405.

De Houwer, J., \& Beckers, T. (2002). A review of recent developments in research and theories on human contingency learning. Quarterly Journal of Experimental Psychology, 55B, 289-310.
Holyoak, K. J., \& Cheng, P. W. (2011). Causal learning and inference as a rational process: The new synthesis. Annual Review of Psychology, 62, 135-63.
Griffiths, T. L., \& Tenenbaum, J. B. (2005). Structure and strength in causal induction. Cognitive Psychology, 51, 334-384.
Lombrozo, T. (2007). Simplicity and probability in causal explanation. Cognitive psychology, 55, 232-57.
Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., \& Holyoak, K. J. (2008). Bayesian generic priors for causal learning. Psychological review, 115, 955-84.
Novick, L. R., \& Cheng, P. W. (2004). Assessing interactive causal influence. Psychological Review, 111, 455-85.
Shanks, D. (1985). Forward and backward blocking in human contingency judgment. Quarterly Journal of Experimental Psychology Section B: Comparative and Physiological Psychology, 37, 1-21.
Waldmann, M. R. (2001). Predictive versus diagnostic causal learning: Evidence from an overshadowing paradigm. Psychonomic Bulletin \& Review, 8, 600-8.
Yeung, S., \& Griffiths, T. L. (2011). Estimating human priors on causal strength. In L. Carlson, C. Hoelscher, \& T. F. Shipley (Eds.), Proceedings of the 33rd Annual Conference of the Cognitive Science Society (pp. 17091714). Austin, TX: Cognitive Science Society.

# Beyond backchannels: co-construction of dyadic stance by reciprocal reinforcement of smiles between virtual agents. 

Ken Prepin (ken.prepin@telecom-paristech.fr)<br>LTCI-CNRS/Telecom-ParisTech, 37-39 rue Dareau<br>75014 Paris, France

Magalie Ochs (magalie.ochs@ telecom-paristech.fr)<br>LTCI-CNRS/Telecom-ParisTech, 37-39 rue Dareau<br>75014 Paris, France

Catherine Pelachaud (catherine.pelachaud@telecom-paristech.fr)<br>LTCI-CNRS/Telecom-ParisTech, 37-39 rue Dareau<br>75014 Paris, France


#### Abstract

When two persons participate in a discussion, they not only exchange the concepts and ideas they are discussing, they also express attitudes, feelings and commitments regarding their partner: they express interpersonal stances. Endowed with backchannel model, several virtual agents are able to react to their partners' behaviour through their non-verbal behaviour. In this paper, we go beyond this approach, proposing and testing a model that enables agents to express a dyadic stance, marker of effective communication: agents will naturally coconstruct a shared dyadic stance if and only if their interpersonal stance is reciprocally positive. We focus on smile, which conveys interpersonal stance and is a particularly efficient signal for co-regulation of communication. With this model, a virtual agent, only capable to control its own individual parameters, can, in fact, modulate and control the dyadic stance appearing when it interacts with its partner. The evaluation of the model through a user perceptive study has enabled us to validate that the dyadic stance is significantly perceived as more positive (mutual understanding, attention, agreement, interest, pleasantness) when reinforcement of smile is reciprocal. Keywords: dyadic interaction; interactive behaviours; dynamical systems; dyadic stance; smile; virtual agent;


## Introduction

When we consider verbal communication, interlocutors not only exchange the concepts and ideas which constitute the subject of their discussion, they also express feelings, judgements or commitments regarding this subject. This "attitude which, for some time, is expressed and sustained interactively in communication, in a unimodal or multi-modal manner" corresponds to the stance: Chindamo, Allwood, and Ahlsén (2012) review the existing definitions and descriptions of stance; they show how these definitions have evolved from a focus on individual expression of stance to a more interactive and social description. Individual stance refers to two types of stance: epistemic and interpersonal stance (Kielsing, 2009). The epistemic stance is the expression of the relationship of a person to his/her own talk (for instance "certain"). The interpersonal stances convey the relationship of a person to the interlocutor (for example "warm" or "polite"). Moreover, during an interaction, "stances are constructed across turns rather than being the product of a single turn" (Chindamo et al., 2012). When interactants with individual epistemic and interpersonal stances are put in presence,
dyadic stances can be inferred (Prepin, Ochs, \& Pelachaud, 2012) from diachronic alignment between interactants. The effort of interlocutors to linguistically and non-verbally align through time is a marker of stance: it convey stance of mutual understanding, attention, agreement, interest and pleasantness (Louwerse, Dale, Bard, \& Jeuniaux, 2012).

The description of stance has not only evolved toward a distinction between individual and co-constructed stance. It has also evolved from a uniquely linguistic description (DuBois, 2007; Kielsing, 2009) to a description implying interactants’ Non-Verbal Behaviours (NVBs) (Scherer, 2005; Prepin et al., 2012). The non-verbal behaviours participate in maintaining contact between interactants and facilitate verbal exchange: they are an integral part of the communication process (Paradowski, 2011). NVBs actively convey stances through paralinguistic features (such as tone of voice, duration, loudness or prosody), facial expressions, and postures (Chindamo et al., 2012).

Models of interactive agents have mainly explored the automatic generation of virtual agent's behaviour aligned on the interlocutor's behaviour. Buschmeier, S., and Kopp (2010) combine a model of lexical alignment with a model generating behaviours based on linguistic information. Bailenson and Yee (2005) model the NVBs alignment of a speaking virtual agent to a listening human. They propose a Digital Chameleon (in reference to the Chameleon effect described by Chartrand and Bargh (1999)). Bevacqua, Hyniewska, and Pelachaud (2010) model the NVBs alignment of a listening agent to a speaking human: they propose a model of backchannels, i.e. NVBs aligned in time and nature, to facilitate human users to tell a story.

All these models focus on the adaptation of the virtual agent to its interlocutor, but do not take into account the reciprocal adaptation of this interlocutor: behaviours are computed in reaction to partner's behaviour, but not in interaction with partner's behaviour; the dynamical coupling associated to the mutual engagement of interactants is not modelled, and critical parameters of interaction such as synchrony and alignment which appear as side effects of this coupling (Paolo,

Rohde, \& Iizuka, 2008; Prepin \& Pelachaud, 2011, 2012a), are missed. In this paper, we aim at going further by proposing a model enabling virtual agents to co-construct their behaviours: agents will be enabled to adapt to each other behaviour on the fly (that is in the time scale of the coupling (Prepin \& Pelachaud, 2011)) and to perform a resulting behaviour which is a dynamically built mix of each other behaviour; agents will also be enabled to modulate how much their own behaviour is influenced by the behaviour of the other, and doing so, they can control the stance of the dyad.

In the present paper, we propose and test a model that enables virtual agents to co-construct a dyadic stance by taking advantage of the interactive loop existing between agents and the resulting conjugated effects of reciprocal alignments. Each virtual agent, only capable to control its own individual parameters, can, in fact, modulate and control the dyadic stance appearing when it interacts with its partner. We focus on smile behaviours for three reasons: (P1) a smile is one of the simplest and most easily recognized facial expressions (Ekman \& Friesen, 1982); (P2) recent works (Ochs, Niewiadmoski, \& Pelachaud, 2010) have shown that people are able to distinguish different types of smile when they are expressed by a virtual character; (P3) in multimodal communication, smile alignment appears in the form of synchronous smile expressions of interactants (Louwerse et al., 2012). These three properties of smile enable us to focus on the dynamical mechanisms of smiles alignment to model the co-construction of dyadic stances. For this purpose, based on the first property of smile (P1), we model the sensitivity to partner's smile as a motor resonance phenomenon. Considering the second property of smile (P2), we implement this model on a dyad of smiling virtual agents. Based on the third property of smile (P3), we enable the virtual agents' smiles occurring synchronously to reinforce each other depending on the two agents' individual stances.

## Model description

In order to create virtual agents able to co-construct a dyadic stance by taking advantage of the interactive loop they form with their partner, we focus on the agents capacity to mutually reinforce their smiles (see Introduction). The agents will be able to change the influence of their partner's smile on their own smile: the more their own actions are influenced by the perception of their partner's actions, the easier will be the coupling and the mutual reinforcement of the two agents smile; virtual agents will be able to control the dyadic stance they co-produce with their interlocutor.

## Smiles descriptions

In the proposed model, we focus on virtual agent's smiles. On one hand, smile is one of the simplest and most easily recognised facial expression (Ekman \& Friesen, 1982), and on the other hand it is one of the few behaviours often performed contingently by partners during interaction (Louwerse et al., 2012). The two muscles zygomatic major, on either side of

| Characteristics <br> of smile | Amused <br> smile | Polite smile | Embarrassed <br> smile |
| :--- | :---: | :---: | :---: |
| Cheek raising | + | - | - |
| Open mouth | + | - | - |
| Lips tension | - | - | + |
| Symmetry | + | + | - |
| Amplitude | + | - | - |

Table 1: Smiles characteristics depending on their type (table filled based on the results described in (Ochs et al., 2010)): + indicates significantly higher and - significantly lower values of the characteristic for a given type of smile than the others,
the face, have to be activated to create a smile, and are sufficient for an observer to recognize a facial expression as being a smile. However, subtle differences in dynamics and in muscular activations make smiles convey different messages (such as amusement and politeness). Ochs et al. (2010) have studied the characteristics of polite, amused, and embarrassed smiles of virtual agent's. Their results are summarized in Table 1 . The amused smiles are mainly characterized by large amplitude, open mouth, symmetry, and relaxed lips. Most of them also contain the activation of the cheek raising, and a long global duration. The polite smiles are mainly characterized by small amplitude, a closed mouth, symmetry, relaxed lips, and an absence of cheek raising. The embarrassed smiles often have small amplitude, a closed mouth, and tensed lips. They are also characterized by the absence of cheek raising and an asymmetry in the smile.

## Perception-Action mapping

In order to enable virtual agents to modify their facial expressions "on the fly" (that is dynamically and in real-time), as proposed in (Prepin \& Pelachaud, 2012b), facial expressions are updated frame by frame depending on both the speech expressed and the continuously incoming reactions of its partner. When an agent is performing an action (e.g. the display of a facial expression), it can have feedbacks concerning this action and can modify it "on the fly".
Several researches have shown that there is a natu$\mathrm{ral} /$ structural tendency to imitate the other and to better perceive the other when imitating back (Muir, 2005; Nadel, Prepin, \& Okanda, 2005). We model this property combining a mapping between the perceptive space and the motor space, and the self-activation of the motor space. Both the perceptive space and the motor space are defined by Action Units (AUs) in the Facial Action Coding System (Ekman \& Friesen, 1982) necessary to define smiles.

The self-activation of the motor space, with a weight $\alpha<$ 1 (see Figure.1), both simulates a short term memorisation of actions and facilitates the subsequent activation of similar actions (Schöner \& Thelen, 2006). The nearer $\alpha$ is to 1 , the longer the memorisation. We choose here $\alpha=0.95$ to ensure that this memorisation is "short term", i.e. that after 1 sec. (25 time steps), if there is no other stimulation, the activation of the AU is decreased by two thirds: $A U_{i}\left(t_{0}+25\right)<1 / 3$. $A U_{i}\left(t_{0}\right)$.

The mapping between perceptive and motor spaces corresponds to the links between perceived characteristics of smiles and generated characteristics of smiles. The mapping is based on the results on smiles reported in previous section. More precisely, the nodes in the perceptive and motor spaces correspond to the characteristics of the different types of smile ${ }^{1}$.

This mapping is represented in Figure 1 by links of different widths between the perceptive ( $A U_{p e r}$ ) and motor $\left(A U_{\text {prod }}\right)$ spaces. The dashed links ending with a circle represent inhibitory links.


Figure 1: Perceptive Space and Motor Space mapping.
The excitatory/inhibitory nature of links and their weight have been inferred from Table 1. We detail the modelled effects for each smile characteristic:

- Zygomatics: zygomatics appear in every smile and only their high amplitude indicates amused smile (Table 1); we assume that their perception will influence their production only if the perceived amplitude is over a threshold $t h_{\beta}$.
- Lips tension: amused and embarrassed smiles are incompatible (they have opposite characteristics, see Table 1); we assume that the specific AU of an embarrassed smile (i.e. lips tension) will inhibit and will be inhibited by the specific AUs of amused smile (i.e. cheeks raised and zygomatics over $t h_{\beta}$ ).
- Cheek raising: cheeks raising is an exclusive marker of amused smile (Table 1); we assume that its perception highly excites all the specific characteristics of amused smile (zygomatic above $t h_{\beta}$, cheeks raise and mouth opening).
- Mouth Opened: opening of mouth is not a specific characteristic of smiles. We assume that its perception only influence the opening of mouth production.
We stay at the level of a purely reactive model, only using muscular activations of produced and perceived signals. More cognitive modelling could infer emotions and intentions from these muscular activations.


## Interpersonal stance influence

Virtual agent's interpersonal stance (i.e. its stance regarding its interlocutor) influences the visuo-motor mapping (Fig.2).

[^190]For instance, a virtual agent with a cooperative attitude will be more sensitive to the interlocutor's perceived smile. Note that we do not model any cognitive model or strategy concerning the expression of stance, we just model how the interpersonal stance of the virtual agent modifies the way the agent is sensitive to its partner's behaviours: the agent will modify how much it is interactive, engaged and finally cooperative with its partner ${ }^{2}$

We assume here that interpersonal stance is represented as a single variable $\sigma$, in $[0,1]$, which multiplies all the influences between perceptive and motor spaces (see Fig.2). In the evaluation study, $\sigma$ only takes two values: $\sigma=0$ when the virtual agent is not cooperative, i.e. when its smiles are not reinforced by its partner's smiles; and $\sigma=0.45$ when the virtual agent is cooperative, i.e. when its smiles are reinforced by its partner's smiles. Note here that if $\sigma$ was higher than 0.45 , even without any communicative intention stimulating smiles, the reciprocal influence between agents would be too high to let smiles decrease.

## Virtual agents dyad

The last step in the design of our model is to put two virtual agents in presence, a speaker and a listener (Fig.2). For sake of simplicity and to focus on the dyadic effect of the smile expressions, the virtual listener has no access to the meaning of what the speaker says. The listener only perceives the speaker's non-verbal behaviour. On the other side, the speaker's speech directly influences its own actions in the motor space (see Fig.2).


Figure 2: Scheme of the interactive loop within the dyad.
We implement our model of virtual agents dyadic stance generation in the Leto/Prometheus Neural Network (NN) simulator (Gaussier \& Zrehen, 1994), interfaced with the virtual agent platform SEMAINE (Schröder, 2010). The NN simulator enables to design the architecture neuron by neuron and to control architecture dynamics in real-time (here frame by frame). The agent platform computes the communicative intention of the virtual character depending on its speech, and directly influences its actions in the motor space accordingly (see Fig.2). For instance, the utterance "I'm happy today" is automatically said with an amused smile.

[^191]In the context of face to face interaction, if both virtual agents have a cooperative interpersonal stance, they reciprocally reinforce their smiles (see Fig.3, (Prepin \& Pelachaud, 2012b)): a snowball effect on shared behaviours (when coupling occurs) and a decay/alignment of not-shared behaviours (when coupling is disrupted).


Figure 3: Dyadic dynamics of smiles. Solid and dotted lines are respectively for Agent1 and Agent2's intensity of smile.

The figure 4 shows the result of such an interaction on one agent: the virtual agent's smile is emphasized.


Figure 4: Snowball effect when smile reinforcement is reciprocal.

Finally, the proposed model enables one to simulate an interaction between two virtual agents with different smiling behaviour depending on the agents' interpersonal stance. The resulting interactions reflect different dyadic stances. In addition to cheeks raise and release of lips tension, the main side effect of mutual positive interpersonal stance is the snowball effect on smiles, i.e. the increase of smiles intensity and duration.

Indeed, considering that NVBs alignment and dynamical coupling are marker of the quality of the interaction (see Introduction), these side-effects (such as "snowball effect") are the cues that should give an impression of fruitful interaction. In order to validate that our model enables one to simulate interactions between virtual agents that convey different dyadic stances depending on the mutual reinforcement of their smiles, we have performed an evaluation presented in the next section.

## Evaluation of the model

To test that the proposed model enables one to simulate the co-construction of different stances, we have performed a user perceptive study. Our objective through this evaluation is to show that the smiles mutual reinforcement between two interacting virtual characters conveys specific stances. We have focused on the following dyadic stances: mutual understanding (the virtual characters seem to understand each other), mutual attention (the virtual characters seem to pay attention to each other), mutual agreement (the virtual characters seem
to be agreed with each other), mutual interest (the virtual characters seem to be interested to the discussion), mutual pleasantness (the virtual characters seem to spend a pleasant time to interact). These stances have been chosen since research (Louwerse et al., 2012) has shown that the mutual understanding, attention, agreement, interest and pleasantness are cues of the quality of an interaction between a speaker and a listener.
Hypothesis. The hypothesis we want to validate through the evaluation is the following:

The positive dyadic stance is significantly increased when reinforcement of smile is reciprocal.
More precisely, the evaluation aims to show that the mutual reinforcement of the smiles of the two interlocutors (i.e. the speaker and the listener) increases the impression of mutual understanding, attention, agreement, interest, pleasantness compared to an interaction in which only the listener's smiles are reinforced by the speaker's smiles (and not in the other way round).

A validation of this hypothesis will enable us to validate the proposed model which simulates virtual characters' dyadic stances through smiles mutual reinforcement and emerging snowball effect.
Procedure. In order to verify the hypothesis, we have performed the evaluation on the web. The evaluation was in French. Four video clips showing two virtual characters discussing were presented to participants. For each video clip, we asked the participants to answer 5 questions using a Likert scale of 5 points (from "strongly disagree" to "strongly agree"). The questions concerned their perception of the mutual understanding, attention, agreement, interest and pleasantness of the two virtual characters. An example of a question is "When you watch the two virtual characters discussing, according to you, do they understand each other?" (translated from French).
Video Clips. To evaluate the perception of the interaction between virtual characters in one way versus reciprocal conditions of smiles reinforcement, we have recorded the two conditions of interaction:

- reciprocal condition: both the speaker and the listener mutually reinforce their smiles depending on the smiles expressed by each other, "snowball effect" is enabled.
- control condition: only the listener reinforces its smiles according to the speaker's expressed smiles.
In the video clips, the virtual characters discuss using an unintelligible verbal language (corresponding to an acoustic deformation of French texts). By this way, we avoid an influence of what the virtual characters said on the user's perception. We have considered 6 different texts corresponding to the situation in which the virtual character tells a joke to its interlocutor. Given the text and the associated communicative intention, the virtual character expresses a polite smile at the beginning and an amused smile in the middle of the text. For each text, we have recorded video clips in the 2 conditions described above with a virtual character saying this text
with an acoustic deformation and another virtual character, in front, listening. In total, 12 video clips have been recorded. In order to visualize clearly the faces of the two virtual characters while keeping the impression that the virtual characters are face to face, we have used a film-making technique called split-screen (Fig.5). Before starting the evaluation on the web, to ensure that the instruction, the questions, and the video clips are understandable, the platform of test has been pre-tested with 7 participants.


Figure 5: Screen shot of a video clip of the two virtual characters interacting
Participants. Sixty-six individuals have participated in this evaluation on the web ( 34 females) with a mean age of $34(\mathrm{SD}=13)$. They were recruited via French mailing lists on line. The participants were predominantly from France ( $\mathrm{N}=63$ ). Each participant was shown and rated 4 video clips (two video clips selected randomly for each of the 2 conditions). The order of the presented video clips were counterbalanced to avoid any effect on the results.
Results (Fig.6). We have collected 264 video clips' ratings. Independent t-Test was conducted to compare the participants' ratings of the video clips in each condition. The analysis revealed statically significant effects of the condition on the participants' ratings of the mutual understanding ( $p<0.001$ ), the mutual attention ( $p<0.01$ ), the mutual agreement ( $p<0.001$ ), the mutual interest ( $p<0.001$ ), and the mutual pleasantness ( $p<0.001$ ).


Figure 6: Means and standard errors of the dyadic stances’ ratings for the two conditions. The significant differences between the condition are indicated by ${ }^{* *}$ for ( $p<0.001$ ), and * for $(p<0.01)$

Discussion of the results. The mutual understanding, attention, interest, agreement and pleasantness are perceived
significantly higher when the speaker and the listener mutually reinforce their smiles according to the other's smiles (reciprocal condition) than when only the listener reinforces its smiles depending on the speaker's expressed smiles (control condition). The impression of mutual understanding, attention, agreement, interest and pleasantness directly depends on the reciprocity of the interaction. These results are consistent with psychology studies which claim that the interaction effort must be shared and reciprocal to enable effective communication (Nadel et al., 2005; Paolo et al., 2008; Auvray, Lenay, \& Stewart, 2009; Fuchs \& DeJaegher, 2009). Finally, the results validate the hypothesis described above: The positive dyadic stance is significantly increased when reinforcement of smile is reciprocal and "snowball effect" is enabled.

## Conclusion

In the present paper, we have proposed a model enabling virtual agents to co-create different dyadic stances. We have described this model entwining each agent's ability to control its cooperation to the interaction and the dyadic effects emerging from the resulting agents coupling.
Agents are able to produce a continuum of smiling behaviours. They can modulate their own smiles depending directly on their perceptions of their partner's smiles. They can control the level of this modulation and doing so control their interpersonal stance: a highly cooperative agent reinforces its smiles when its interlocutor smiles. Finally when a speaking agent (which produces smiles in relation to its speech) and a listening agent are put together, their behaviours modulate each other reciprocally and dynamically form a new behaviour. Performing a user perceptive study, we have shown that this dyadic behaviour is the expression of the two agents dyadic stance: the specific dyadic dynamics which appear depending on each agent interpersonal stance convey information on agents' mutual understanding, attention, agreement, interest and pleasantness. The evaluation highlights that the virtual agent's backchannels (one way reactions) are less effective than reciprocal reactivity to convey some dyadic stances such as mutual understanding, attention, agreement, interest and pleasantness: The agents' reactions must be reciprocal, as proposed in our model, to enable side effects of dynamical coupling such as emphasise of smiles, increase in intensity and duration.
Future works. One of the aspect of the virtual agents modelling we have proposed is the fact that each agent of the dyad, has a different dynamic depending on the other agent stance: the agent's own smile dynamic (for instance the smile slope) changes according to whether or not the other agent has co-operative interpersonal stance. As a consequence, each agent, knowing its own interpersonal stance and detecting its own smile slope variation, could infer the other agent's interpersonal stance. Finally each agent can use this signal for modulating its own stance, its model of the other, or the way it interacts.

One of the next steps is to apply such a model to human-
virtual agent interaction. For this purpose, we are currently integrating in the SEMAINE platform a system to detect in real-time user's smiles ${ }^{3}$. In this condition of direct interaction between user and virtual agent, the user perception of the dyadic stances could be different since the user is directly engaged in the interaction (compared to the studied conditions in which users have a third person point of view when they watch virtual characters interacting).

## Acknowledgements

This work has been financed by the European Community Seventh Framework Program (FP7/2007-2013): the European Project VERVE and the European Project NoE SSPNet.

## References

Auvray, M., Lenay, C., \& Stewart, J. (2009). Perceptual interactions in a minimalist virtual environment. New ideas in psychology, 27, 32-47.
Bailenson, J., \& Yee, N. (2005). Digital chameleons: automatic assimilation of nonverbal gestures in immersive virtual environments. Psychological Science, 16(10), 814-9.
Bevacqua, E., Hyniewska, S., \& Pelachaud, C. (2010, October). Positive influence of smile backchannels in ecas. In International workshop on interacting with ecas as virtual characters (aamas 2010). Toronto, Canada
Buschmeier, H., S., K. B., \& Kopp. (2010). Adaptive expressiveness: virtual conversational agents that can align to their interaction partner. In Proceedings of the 9th international conference on autonomous agents and multiagent systems (pp. 91-98). Richland, SC: IFAAMAS.
Chartrand, T. L., \& Bargh, J. A. (1999). The chameleon effect: the perception-behavior link and social interaction. Journal of Personality and Social Psychology, 76(6), 893910.

Chindamo, M., Allwood, J., \& Ahlsén, E. (2012). Some suggestions for the study of stance in communication. In 2012 ase/ieee international conference on social computing (pp. 617-622). IEEE Computer Society.
DuBois, J. W. (2007). The stance triangle. In R. Engelbretson (Ed.), Stancetaking in discourse (pp. 139-182). Amsterdam/Philadelphia: John Benjamins Publishing Company.
Ekman, P., \& Friesen, W. V. (1982). Felt, False, And Miserable Smiles. Journal of Nonverbal Behavior, 6, 238-252.
Fuchs, T., \& DeJaegher, H. (2009). Enactive intersubjectivity: Participatory sense-making and mutual incorporation. Phenomenology and the Cognitive Sciences, 8(4), 465-486
Gaussier, P., \& Zrehen, S. (1994). Avoiding the world model trap: An acting robot does not need to be so smart! Journal of Robotics and Computer-Integrated Manufacturing, 11(4), 279-286.
Kielsing, S. F. (2009). Stance: Sociolinguistic perspectives. In A. Jaffe (Ed.), (pp. 171-194). Oxford: Oxford University Press.

[^192]Louwerse, M., Dale, R., Bard, E. G., \& Jeuniaux, P. (2012). Behavior Matching in Multimodal Communication is Synchronized. Cognitive Science, 36(8), 1404-26.
Muir, D. (2005). Emotional development. In J. Nadel \& D. Muir (Eds.), (pp. 207-233). Oxford, UK: Oxford University Press.
Nadel, J., Prepin, K., \& Okanda, M. (2005). Experiencing contingency and agency: first step toward selfunderstanding? In P. Hauf (Ed.), Making minds ii: Special issue of interaction studies 6:3 2005 (pp. 447-462). John Benjamins publishing company.
Ochs, M., Niewiadmoski, R., \& Pelachaud, C. (2010). How a virtual agent should smile? morphological and dynamic characteristics of virtual agent's smiles. In Intelligent virtual agent (iva).
Paolo, E. A. D., Rohde, M., \& Iizuka, H. (2008). Sensitivity to social contingency or stability of interaction? modelling the dynamics of perceptual crossing. New ideas in psychology, 26, 278-294.
Paradowski, M. B. (2011). The Embodied Language: Why Language Should Not Be Conceived of in Abstraction from the Brain and Body, and the Consequences for Robotics. SSRN eLibrary.
Prepin, K., Ochs, M., \& Pelachaud, C. (2012). Mutual stance building in dyad of virtual agents: Smile alignement and synchronisation. In Proceedings of international workshop on exploring stances in interactions, ase/ieee international conference on social computing (pp. 938-943).
Prepin, K., \& Pelachaud, C. (2011). Effect of time delays on agents' interaction dynamics. In Tumer, Yolum, Sonenberg, \& Stone (Eds.), The tenth international conference on autonomous agents and multi agents systems, aamas2011 (pp. 1-8). IFAAMAS.
Prepin, K., \& Pelachaud, C. (2012a). Basics of intersubjectivity dynamics: Model of synchrony emergence when dialog partners understand each other. In J. Filipe \& A. Fred (Eds.), Icaart 2011, revised selected papers, series: Communications in computer and information science (ccis) (Vol. 271, pp. 302-318). Springer-Verlag.
Prepin, K., \& Pelachaud, C. (2012b). Live generation of interactive non-verbal behaviours. In Conitzer, Winikoff, Padgham, \& van der Hoek (Eds.), 11th international conference on autonomous agents and multiagent systems, aamas 2012 (pp. 1-2). IFAAMAS.
Scherer, K. (2005, December). What are emotions? and how can they be measured? Social Science Information, 44(4), 695-729.
Schöner, G., \& Thelen, E. (2006). Using dynamic field theory to rethink infant habituation. Psychological Review, 113(2), 273-299.
Schröder, M. (2010). The semaine api: Towards a standardsbased framework for building emotion-oriented systems. Advances in Human-Computer Interaction, 2010, 21.

# Exploring word recognition with selected stimuli: The case for decorrelated parameters 

Athanassios Protopapas (aprotopapas@phs.uoa.gr)<br>Department of Philosophy \& History of Science (MITHE), Ano Ilissia University Campus<br>GR-15771 Zografos, Greece

Efthymia C. Kapnoula (ekapnoula@iowa.uiowa.edu)<br>Department of Psychology, E11 Seashore Hall<br>Iowa City, IA 52242-1409 USA


#### Abstract

We report a study of naming and lexical decision with 132 adult Greek speakers responding to 150 words and matched pseudowords with decorrelated frequency, length, neighborhood, syllable and bigram frequency, and transparency. This approach allowed us to individuate and accurately estimate the effects of each variable, and to assess their linearity and additivity. Significant effects of frequency, length, and syllable frequency were revealed, as well as several interactions. The results are informative for cognitive modeling of visual word recognition in more transparent orthographies.


Keywords: Visual word recognition; pseudowords; naming; lexical decision; mixed-effects models; Greek.

Models of visual word recognition posit distinct mechanisms and representations involved in the processing of orthographic stimuli. The implications of these hypotheses for response times (RT) to individual words and pseudowords are typically studied using naming and lexical decision tasks. A productive line of research concerns the effects of lexical and sublexical variables on RT distributions. In evaluating models and their properties, several variables have been examined in this light and have been found to be related to processing times, such as frequency, length, neighborhood, bigram and syllable frequency, and more (Balota, Yap, \& Cortese, 2006).

A common approach to such studies has been to form groups of stimuli differing in a parameter of interest, such as a low-frequency and a high-frequency group. Interactions are then examined by crossing levels of variables in subgroups. This has led to the identification of important effects but has recently been criticized for problems stemming from selection of atypical items and restricted parameter ranges (Balota, Yap, Hutchison, \& Cortese, 2012). More recent approaches have abandoned stimulus groupings in favor of regression approaches, in which wide value ranges of several variables are simultaneously entered in multivariate analyses (Yap \& Balota, 2009). Advances in statistical modeling have allowed more parameters to be examined (Baayen, 2008).

This approach has culminated in the "mega study" efforts, in which huge numbers of stimuli are presented to volunteer samples of unprecedented sizes. For example, the English Lexicon Project (Balota et al., 2007) includes about 40k words and pseudowords responded to by 1,260 participants. This allows examination of the full ranges of all parameters and their combinations. Special subsets may be selected for targeted analysis, so this method allows replication of previ-
ous studies and examination of potential biases or dependencies. Moreover, these data are available for future studies on as-yet unidentified parameters without further data collection. Thus mega-studies constitute a large step forward in our quest for understanding word processing.

The multiple regression approach is also vulnerable to criticism. Regression models typically include a linear effect for each variable, assuming that this effectively "partials out" all influence of the corresponding parameter, allowing the effects of other parameters to be accurately estimated. This is only the case when all effects are linear and independent. If not, the linear modeling removes only part of an effect, conflating the remainder with other correlated variables. The problems of linearity and additivity are compounded and cannot be addressed in stimulus sets with correlated effects because it cannot be known whether a variable has a true curvilinear effect or an interaction with a correlated variable. In both cases departures from linearity will be identified but the decision to "remove" one of the two will affect the other and may do so in a nonrepresentative or nonoptimal manner. Moreover, in the absence of curvilinearity and nonadditivity, correlated parameters necessarily lead to underestimated effects because shared variance is removed when one parameter is controlled.

Linearity is potentially informative as cognitive processing models can make specific predictions about the shape of the relationship between aspects of the stimuli and the time required to respond to them. Thus it is useful to establish the fitting functions on the actual RT distributions and incorporate them in modeling (cf. Balota \& Yap, 2011). Nonadditivity is also of great interest because it has implications for theoretical approaches insofar as the cause of each interaction must be understood within the context of relevant assumptions.

In the present study we begin to address these issues by examining RTs to a set of words and pseudowords in which variables typically examined in word recognition were uncorrelated. We work in Greek, a language with higher orthographic transparency and word length than English (Protopapas \& Vlahou, 2009), aiming to contribute to the cross-linguistic effort. We selected words from a corpus ensuring that there was no significant correlation among frequency, length, neighborhood, syllable and bigram frequency, and graphophonemic transparency. A set of pseudowords was then created, similarly uncorrelated, matching the word group in these vari-
ables. To the extent a reasonably wide range was sampled, this approach permits isolation of the effects of each variable and identification of individual interactions and nonlinearities. All items were presented in a naming and a lexical decision task. To maximize reliability and allow detection of small effects, we employed a relatively large sample, based on Rey, Courrieu, Schmidt-Weigand, and Jacobs (2009).

## Method

## Participants

The sample included 97 women and 35 men, native speakers of Greek, $18-36$ years old ( $M=23.3, S D=4.7$ ), mostly students (12-21 years of education, $M=15.4, S D=2.1$ ). Fourteen were left-handed.

## Stimuli

A set of 150 words were selected from the IPLR word list (Protopapas, Tzakosta, Chalamandaris, \& Tsiakoulis, 2012), $2-5$ syllables long (4-10 letters; 4-11 phonemes). In an iterative selection process, a set of properties were retrieved along with each word, including $\log$ frequency of occurrence; number of letters, phonemes, and syllables; orthographic and phonological neighborhood (Coltheart's N ); orthographic and phonological syllable frequency; letter and phoneme bigram frequency; and a nondirectional measure of graphophonemic transparency (log mean token "sonograph" probability; Spencer, 2009). A nonparametric index of association (Spearman's $\rho$ ) among all variables was calculated. The process terminated when groups of qualitatively distinct variables were not significantly correlated. The following variables were retained as most relevant for the analyses reported below: log frequency, number of letters, mean syllable and bigram frequency, and transparency. Figure 1 (top row) shows the distribution of each variable in the selected sets against the overall type and token distribution in the corpus.

A set of 150 pseudowords were constructed to resemble the words in basic phonological and orthographic structure and letter and phoneme distribution. The pseudowords were indistinguishable from the words in the target variables, as verified by the Kolmogorov-Smirnov test for equality of distributions. The results of these tests are listed in Table 1, along with the correlations among the variables for both sets.

A difficulty was encountered in pseudoword neighborhoods. Matching to the words required including items with many neighbors. Neighborhoods for long words are mainly due to grammatical inflection. Long words typically have few or no unrelated neighbors, but due to the rich inflectional system of Greek all content words have neighboring inflectional variants. Neighborhoods for pseudowords would therefore be limited to the inflectional families of word neighbors, so that some pseudowords would be strongly influenced by one lexical lemma. This was undesirable because pseudowords are known to activate lexical neighbors strongly (even assimilating their stress pattern; Protopapas, Gerakaki, \& Alexandri, 2007). Therefore, using pseudowords with neighbors would
result in a pseudoword set in which word activation might play a prominent role and suppress nonlexical effects. Thus we decided to minimize pseudoword neighborhoods and keep the pseudowords distinct from specific words, at the cost of matching and correlations involving neighborhoods.

## Procedure

A naming and a lexical decision task were implemented in DMDX (Forster \& Forster, 2003). In both tasks, each item was presented in Arial 36-pt white font at the center of a laptop $15.4^{\prime \prime}$ screen for 1.9 s . A few practice and warm-up items preceded the experimental stimuli. A short break was offered halfway through each block of 150 stimuli. For lexical decision, participants responded by pressing the left and right control keys. Words and pseudowords were intermixed randomly. The "word" response was set to the participant's preferred or nonpreferred hand, approximately counterbalanced. For naming, words and pseudowords were presented in separate blocks, in counterbalanced order between participants. Responses were recorded and onset times were subsequently verified using CheckVocal (Protopapas, 2007). The order of naming and lexical decision tasks was counterbalanced. In both tasks, items were presented in a different random order for each participant. A distractor task (digit span) was administered between the two tasks to minimize carryover effects.

## Results

Raw response times were logarithmically transformed and analyzed using mixed-effects models with crossed random effects for participants and items, separately for words and nonwords. Analyses were conducted using package lme4 (Bates, Maechler, \& Bolker, 2012) in R (R Core Team, 2012), mainly following Baayen (2008). All variables were centered. Models reported below included fixed effects and random slopes per participant for the linear effects of trial number and preceding RT, as well as random intercepts for participants and for items. These "baseline" effects are not discussed further.

For each task and stimulus type we examined the following: (a) A "full" model, with the complete set of variables, i.e., baseline effects plus all six experimental parameters (linear effects only, not interacting). This was used to estimate the "full" effects of each parameter. (b) For each parameter, a variant of the full model excluding that parameter. The item random intercepts of this model were used to estimate the "residual" effects of each parameter. (c) A baseline model with only the baseline effects. (d) For each parameter, a variant of the baseline model including only that parameter. This was used to estimate the "single" effects of each parameter. Comparison of this model to the baseline (via likelihood ratio) determined the significance of each parameter. (e) An "augmented" model, in which quadratic effects and interactions were added to the full model in a forward-backward procedure and retained when significant (determined via likelihood ratio at $p<.05$ ). The linear effects of all six parameters were retained whether or not they made significant contributions, and are listed in Table 1 (4 rightmost columns).

Table 1: Correlation coefficients (Spearman's $\rho$ ) among the 6 experimental parameters, for words (above the diagonal) and pseudowords (below the diagonal), and Kolmogorov-Smirnov tests for equality of distributions between words and pseudowords. The four rightmost columns show the estimated $\hat{\beta}$ for the corresponding linear effect in the augmented model (see text).

| Variable |  | Correlations |  |  |  |  | K-S test |  | Naming |  | Lexical decision |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | 2 | 3 | 4 | 5 | 6 | D | $p$ | Words | Pseudo | Words | Pseudo |
| 1 | frequency | -. 05 | . 00 | -. 02 | . 10 | . 00 |  |  | $-.0152^{*}$ |  | -.0311* |  |
| 2 | N letters |  | -. 01 | -. 05 | . 01 | -. 11 | . 06 | . 950 | +.0163* | +.0306* | +.0151* | +.0361* |
| 3 | neighborhood | $-.50^{\dagger}$ |  | -. 08 | . 07 | -. 02 | . 70 | . 000 | -. 0033 | $-.0668^{*}$ | -.0092* | +.0409* |
| 4 | syllable freq. | -. 08 | . 07 |  | . 09 | -. 09 | . 15 | . 079 | +.0032* | +.0022* | +.0044* | +. 0006 |
| 5 | bigram freq. | . 08 | -. 04 | . 06 |  | . 06 | . 09 | . 531 | -. 0039 | -.2131* | -. $2578 *$ | +. 1335 |
| 6 | transparency | . 00 | -. 03 | -. 12 | . 05 |  | . 05 | . 983 | +. 0000 | -. 0005 | +.0018* | -. 0010 |

Note: ${ }^{\dagger} p<.0005$; for all other correlations, $p>.1 ;{ }^{*}|t| \geq 2.0$

The variance of item random intercepts varied between tasks. In the baseline models, it was 4.98 for word naming, 16.03 for pseudoword naming, 9.99 for word lexical decision, and 7.90 for pseudoword lexical decision (all $\times 10^{-3}$ ). This sets an upper limit for the contributions of the experimental parameters, which are all item-related, indicating that there is more variance to be accounted for in pseudoword naming (the highest) than in word naming (the lowest). In comparison, the residual (error) variance of the baseline models were $13.1,19.3,37.9$, and 36.8 , respectively, suggesting that lexical decision tasks are "noisier" than naming tasks.

Linear individual effects The results for these analyses are summarized graphically in the four bottom rows of panels in Figure 1, in which the linear effect of each parameter is tested when all other parameters were in the model (residual vs. full model) and when no other parameters were in the model (single vs. baseline model). The two estimates were generally within one standard error of each other, indicating very close correspondence of the two types of analysis in estimating individual linear effects. The reduction in item variance (random intercept for items) by inclusion of each parameter was also very similar between the two approaches. There were some differences in significance but they concerned low variance proportions ( $1 \%$ ) or were associated with pseudoword orthographic neighborhood, which was not decorrelated.

Quadratic effects Examination of the raw and residual trends in Figure 1 indicated monotonic and largely linear effects of frequency, length, and syllabic frequency, especially for words. Some curvilinearity was apparent for other parameters, particularly for pseudowords. Nonlinearities were examined for every parameter in each case by testing quadratic terms added (centered) to the full model. The quadratic effect of bigram frequency was significant in pseudoword naming ( $\hat{\beta}=.1024, S E=.0435$ ), word lexical decision $(\hat{\beta}=.1720$, $S E=.0854$ ), and pseudoword lexical decision ( $\hat{\beta}=-.0778$, $S E=.0412$ ). The quadratic effect of orthographic neighborhood was significant in pseudoword lexical decision ( $\hat{\beta}=$ $-.0043, S E=.0021$ ) but seems spurious in light of the severely skewed distribution of this parameter. No other
quadratic terms were found to be significant. No higher-order terms or other nonlinear functions were tested.

Interactions Two-way interactions were examined by testing each pair of parameters. In word naming, there was an interaction between frequency and neighborhood ( $\hat{\beta}=.0054$, $S E=.0016$ ). In addition, there were interactions of transparency with frequency $(\hat{\beta}=.0007, S E=.0003)$, neighborhood ( $\hat{\beta}=-.0007, S E=.0003$ ), and syllabic frequency $(\hat{\beta}=.0002, S E=.0001)$. In pseudoword naming there was an interaction of length and neighborhood ( $\hat{\beta}=-.0201$, $S E=.0068$ ). In word lexical decision there were interactions of neighborhood with frequency $(\hat{\beta}=.0070, S E=.0023)$ and with transparency $(\hat{\beta}=-.0009, S E=.0005)$. Finally, in pseudoword lexical decision there was a marginal interaction of syllabic frequency with bigram frequency ( $\hat{\beta}=$ $.0050, S E=.0029$ ). There were no interactions involving the quadratic terms. No higher-order interactions were tested.

Following the addition of the aforementioned quadratic and interactive effects, the augmented models reduced the item variance (random intercepts) by half or more. Specifically, the proportion of item variance in the baseline model that was accounted for by the six experimental parameters was .48 for word naming, .71 for pseudoword naming, .48 for word lexical decision, and .50 for pseudoword lexical decision. Additional proportions in naming could be accounted for by modeling initial phoneme classes but there was no need for that in the present approach as the respective onset effects were adequately captured in the by-item random intercepts.

## Discussion

We employed a stimulus selection procedure to create a set of words and matched pseudowords with decorrelated parameters, aiming to examine curvilinear and interactive effects more accurately than with blind multiple regression. In this study we first confirmed that the linear effects are similar when estimated in full models vs. single-parameter models. Differences emerged for unmatched variables, as should be expected. Therefore this method achieves isolation of the effects of basic parameters, allowing further use in sit-

uations where complex modeling may be impractical or impossible, such as in fMRI. Use of decorrelated parameters to examine brain modulation in response to written stimuli has been previously reported for 465 monosyllabic English words (Graves, Desai, Humphries, Seidenberg, \& Binder, 2010). Our analysis supports this approach and extends it to multisyllabic words and a more transparent orthography.

Our results do not strongly challenge the common assumption of linearity, as most of the effects seem well approximated by a linear function. However, our analyses were based on $\log$ RT and not raw times. If linear fits on $\log$ RT can pass more stringent tests in comparison with a richer set of curvilinear alternatives, the implications for modeling are that models should predict logarithmic RT curves. The existing analysis techniques allow the field to progress from prediction of differences between conditions, or the mere existence of associations among parameters, toward more specific predictions of the relations between participants, items, and measures, as for example in the rate-amount (Faust, Balota, Spieler, \& Ferraro, 1999) and difference engine (Myerson, Hale, Zheng, Jenkins, \& Widaman, 2003) approaches. Simple correlation between predicted and measured times may become inadequate as more specific derivations and variance comparisons become increasingly feasible.

It remains to be established whether the effects uncovered in this analysis are properly accounted for in linear models of $\log$ RT. The effect of frequency, in particular, seems to level off somewhat towards lower frequencies. Although this may be an artifact of nonhomogeneous sampling affecting the lower end, it is consistent with a frequency effect less steep than logarithmic. Given that frequency has been log transformed, it may be fruitful to examine alternatives (e.g., power functions, ranks, or subjective estimates of familiarity) in accounting for the frequency effect (cf. Balota et al., 2012). It is reassuring that effect estimates and variance proportions in lexical decision were substantially greater (double) than in naming, consistent with the notion of frequency as a lexicalsemantic rather than surface variable.

The large effect of word length may seem surprising but it should be taken into account that multisyllabic words up to 10 letters long were involved. The Greek orthography is also rel-
atively transparent for reading (Protopapas \& Vlahou, 2009), conceivably supporting more serial approaches than English. This may explain why about half of the item variance in pseudowords was accounted for by length alone, almost as much in lexical decision as in naming. An interesting aspect of the data concerns the low-end shape of this relation, evident in pseudowords, although there may also be some flattening of the word curves. This may be related to the U-shape reported in other languages (see recent discussion in Yap \& Balota, 2009) and warrants further investigation.

No significant main effects of orthographic neighborhood were revealed in our analyses. This is surprising in light of consistent reports in the literature regarding neighborhood effects. However, there are issues with Greek word neighborhoods that warrant further scrutiny. Due to extensive inflection of nouns, verbs, and adjectives, many items counted as neighbors are inflectional variants, arguably linked to a single lexical lemma (contingent on one's theory of morphological representation in the lexicon). Moreover, the number of neighbors diminishes rapidly with word length, as there are fewer instances of words in longer letter-string space. This suggests that the emphasis on neighborhood effects may have resulted from an artifact of English being the most-studied language and allowing investigation restricted to short, single-syllable words. Alternatively, more flexible indices of orthographic distance may be required to express neighborhood density (e.g., Yarkoni, Balota, \& Yap, 2008).

In agreement with recent reports for other languages (Conrad, Tamm, Carreiras, \& Jacobs, 2010), inhibitory phonological syllabic frequency effects were found for words in both naming and lexical decision. A smaller but significant effect was found in pseudoword naming. No comparable effect was observed with orthographic syllable frequency (not reported above), consistent with the source of such effects lying within a phonological sublexical space. In contrast to syllables, orthographic bigram effects were minor, mainly restricted to pseudowords, and partly facilitatory, with a quadratic component resisting interpretation. There were no effects of phonological bigram frequency (i.e., phoneme pairs), consistent with an explanation for bigram effects related to orthographic familiarity with letter clusters.

Figure 1: (on previous page) The top row shows the distribution of parameter values for the stimulus set, separately for words (blue) and pseudowords (red), in comparison to all corpus types (light peach) and tokens (light blue). Bars display proportions of items, adding up to 1.0 . The other rows display the effects of each of the six experimental parameters on naming (Rows 2 and 3) and on lexical decision (Rows 4 and 5). Each box displays residual item effects (grey circles) in a model including baseline effects and all parameters except one. The red solid line plots a smoothed average (via function lowess) of these points. The dotted red line shows the effect estimate for this parameter when added to the predictor set, resulting in a full model. The teal solid line plots a smoothed average of the centered raw item means. The dotted green line shows the effect estimate of the parameter when included in a model with baseline and random effects only, absent all other parameters. The red and green numbers at the top of each panel are the corresponding effect estimates $(\hat{\beta})$ for the same-color dotted lines, whereas the numbers at the bottom of each panel are the proportions of item variance accounted for by this fixed effect; an asterisk denotes significant contribution (by likelihood ratio test). The vertical axis is scaled in $\log$ milliseconds (with respect to the grand mean intercept). Note different scaling of horizontal axes between parameters and also between distributions (top row) and effects panels.

There were no significant effects of consistency except in the augmented model for word lexical decision. This stands in contrast to large effects of regularity and consistency reported for English and may be attributed to the greater transparency of Greek. Nevertheless, there were consistent interactions involving transparency in word naming and lexical decision. Higher-probability grapheme-phoneme tokens were associated with increased RTs, an effect enhanced for higher-frequency words and diminished in larger neighborhoods. Instead of a pure consistency effect whereby more frequent mappings are decoded more rapidly, here we may have a situation in which systematic mappings permit greater confusion among lexical candidates. The fact that this only occurred for words-the (nonsignificant) trends for pseudowords being negative-suggests that it may be related to orthographic N being a poor index of neighborhood effects.

It should be kept in mind that our findings for pseudowords must be interpreted with caution as the stimulus set was not fully controlled and decorrelated due to the aforementioned neighborhood issue. This is not a major limitation of the study because most cognitive models typically focus on words and do not emphasize pseudoword processing.

Overall, the variance accounted for by our parameters was substantial but far from the $80 \%$ estimate Rey et al. (2009) gave for reproducible item variance in samples of this size. Although inclusion of initial phoneme class raised this proportion considerably, $80 \%$ was achieved only for pseudoword naming. Word naming lagged behind at $63 \%$, indicating that major sources of systematic item variance remain to be brought into the models (Adelman, Marquis, Sabatos-DeVito, \& Estes, in press). Morphological and semantic variables are obvious candidates to be examined in follow-up studies.

## Acknowledgments

Supported by the European Union (European Social Fund) and Greek national funds through the National Strategic Reference Framework - Research Funding Program THALIS-UOA-COGMEK: Cognitive mechanisms in the perception, representation, and organization of knowledge.

## References

Adelman, J. S., Marquis, S. J., Sabatos-DeVito, M. G., \& Estes, Z. (in press). The unexplained nature of reading. Journal of Experimental Psychology: Learning, Memory, and Cognition.
Baayen, R. (2008). Analyzing linguistic data: A practical introduction to statistics using R. Cambridge Univ. Press.
Balota, D., \& Yap, M. (2011). Moving beyond the mean in studies of mental chronometry: The power of response time distributional analyses. Current Directions in Psychological Science, 20(3), 160-166.
Balota, D., Yap, M., \& Cortese, M. (2006). Visual word recognition: The journey from features to meaning (a travel update). In M. J. Traxler \& M. A. Gernsbacher (Eds.), Handbook of psycholinguistics (2nd ed., pp. 285-375). Amsterdam, The Netherlands: Academic Press.

Balota, D., Yap, M., Hutchison, K., \& Cortese, M. (2012). Megastudies: What do millions (or so) of trials tell us about lexical processing? In J. S. Adelman (Ed.), Visual word recognition (Vol. 1, pp. 90-115). Psychology Press.
Balota, D., Yap, M., Hutchison, K., Cortese, M., Kessler, B., Loftis, B., et al. (2007). The English lexicon project. Behavior Research Methods, 39(3), 445-459.
Bates, D., Maechler, M., \& Bolker, B. (2012). lme4: Linear mixed-effects models using S4 classes. (R package version 0.999999-0)

Conrad, M., Tamm, S., Carreiras, M., \& Jacobs, A. (2010). Simulating syllable frequency effects within an interactive activation framework. European Journal of Cognitive Psychology, 22(5), 861-893.
Faust, M., Balota, D., Spieler, D., \& Ferraro, F. (1999). Individual differences in information-processing rate and amount: Implications for group differences in response latency. Psychological Bulletin, 125(6), 777-799.
Forster, K., \& Forster, J. (2003). DMDX: A Windows display program with millisecond accuracy. Behavior Research Methods, 35(1), 116-124.
Graves, W., Desai, R., Humphries, C., Seidenberg, M., \& Binder, J. (2010). Neural systems for reading aloud: A multiparametric approach. Cerebral Cortex, 20, 1799-15.
Myerson, J., Hale, S., Zheng, Y., Jenkins, L., \& Widaman, K. (2003). The difference engine: A model of diversity in cognition. Psychonomic Bulletin \& Review, 10, 262-288.
Protopapas, A. (2007). CheckVocal: A program to facilitate checking the accuracy and response time of vocal responses from DMDX. Behavior Research Methods, 39, 859-862.
Protopapas, A., Gerakaki, S., \& Alexandri, S. (2007). Sources of information for stress assignment in reading Greek. Applied Psycholinguistics, 28(4), 695.
Protopapas, A., Tzakosta, M., Chalamandaris, A., \& Tsiakoulis, P. (2012). IPLR: An online resource for Greek wordlevel and sublexical information. Language Resources and Evaluation, 46(3), 449-459.
Protopapas, A., \& Vlahou, E. (2009). A comparative quantitative analysis of Greek orthographic transparency. Behavior research methods, 41(4), 991-1008.
R Core Team. (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
Rey, A., Courrieu, P., Schmidt-Weigand, F., \& Jacobs, A. (2009). Item performance in visual word recognition. Psychonomic Bulletin \& Review, 16(3), 600-608.
Spencer, K. (2009). Feedforward, -backward, and neutral transparency measures for British English. Behavior Research Methods, 41(1), 220-227.
Yap, M., \& Balota, D. (2009). Visual word recognition of multisyllabic words. Journal of Memory and Language, 60(4), 502-529.
Yarkoni, T., Balota, D., \& Yap, M. (2008). Moving beyond Colthearts N: A new measure of orthographic similarity. Psychonomic Bulletin \& Review, 15(5), 971-979.

# Computational Models of Human Behavior in Wartime Negotiations 

David V. Pynadath (pynadath@usc.edu), USC Institute for Creative Technologies

Ning Wang (ningwang@curiouslab.com), Curious Lab LLC<br>Stacy C. Marsella (marsella@ict.usc.edu), USC Institute for Creative Technologies


#### Abstract

Political scientists are increasingly turning to game-theoretic models to understand and predict the behavior of national leaders in wartime scenarios, where two sides have the options of seeking resolution at either the bargaining table or on the battlefield. While the theoretical analyses of these models is suggestive of their ability to capture these scenarios, it is not clear to what degree human behavior conforms to such equilibrium-based expectations. We present the results of a study that placed people within two of these game models, playing against an intelligent agent. We consider several testable hypotheses drawn from the theoretical analyses and evaluate the degree to which the observed human decisionmaking conforms to those hypotheses.


Keywords: Decision making; game theory; intelligent agents

## Introduction

Political scientists are increasingly turning to computational models to understand and predict nations' wartime behavior (Fearon, 1995). Many such models combine both military and political processes, where battlefield decisions occur within the context of an overall negotiation over a contentious resource (Filson \& Werner, 2002; Smith \& Stam, 2004). Game-theoretic models of these processes seek to capture possible outcomes on the battlefield and at the bargaining table (Powell, 2001, 2004; Slantchev, 2003). These models hypothesize equilibrium strategies that correspond to the behaviors of nations in real-world scenarios.

While the theoretical analyses of such game-theoretic models is suggestive of their representational power, it remains an open question as to how well they capture actual human behavior in wartime negotiation. These models focus on equilibrium behavior, where both sides optimize their outcomes in response to the others' behaviors. The computational challenges of such optimization often require the equilibrium analyses to make simplifying assumptions (e.g., to reduce uncertainty about the opponent). However, people are not constrained to adopt these same assumptions when making their decisions, so it is possible that human behavior in the face of this uncertainty may greatly deviate from the predictions generated by such theoretical models.

On the other hand, these computational models easily lend themselves to an experimental setting, where we can pit a human player against an intelligent agent playing according to the model of interest. In other words, we can place human players within the game hypothesized by a model and have them negotiate with an agent. We can then observe human behavior and quantify the degree to which that behavior conforms to the expectations generated by the model.

This paper presents a human subject study where we implemented games corresponding to two wartime negotiation models from the literature (Powell, 2004; Slantchev, 2003). We present behavior hypotheses extracted from the theoretical analyses of these models. We analyze the observed human behavior to see that it generally satisfies these hypotheses. However, there are also interesting deviations from these theoretical expectations that suggest possible extensions to the models to better capture human decision-making.

## Wartime Negotiation Models

A number of formal models in the literature represent war as a costly process embedded within a negotiation game. In these models, two sides are in a dispute over a desirable resource, such as territory claimed by both sides. The game begins with some initial split of the territory. The game progresses round by round, with each round consisting of one side proposing a split of the territory, the other side responding to that proposal, and a possible battle. The game ends with a final split achieved by either an agreement on the proposed split or a decisive military victory by one side on the battlefield.

To facilitate a game-theoretic analysis, these models make simplifying assumptions regarding military outcomes. In particular, the probabilities associated with the battlefield are fixed, so that one side's probability of winning does not change during the course of the game, regardless of previous military outcomes. The costs of a single battle are also fixed throughout the course of the game. In our study, we present these costs to the human players in terms of troops lost.

A critical property of these models is uncertainty about the likelihood of battlefield outcomes. If both sides had complete information about their probability of winning battles, they could do an exact cost-benefit analysis and immediately agree upon an acceptable territorial split. The models we consider instead have incomplete information, where one side is ignorant of the probability of battlefield outcomes. As the war progresses, this side will gain information by observing battle and bargaining outcomes, re-evaluate its prospects, and make different decisions on offers and battles. This asymmetry lends itself to our human subject study, as we can give the agent complete information about the game probabilities, but hide that information from the human player. Our experiments will then allow us to study how people update their beliefs based on the information that is revealed in the game.

We chose two models (Powell, 2004; Slantchev, 2003) for this investigation, based on their impact on the field and their appropriateness for a human-agent game interaction.

## Powell Model

The Powell model proceeds as follows (Powell, 2001, 2004):

1. Player 1 makes an offer of $x \%$ of the territory.
2. Player 2 decides to accept, reject, or go to war.
(a) If acceptance, Player 2 gets $x \%$ of the territory, Player 1 gets $(100-x) \%$, and the game is over.
(b) If war, both sides incur the battle costs. Player 1 collapses with probability $p_{1}$ and Player 2 collapses with probability $p_{2}$.
i. If only Player 1 collapses, Player 2 gets all of the territory, and the game is over.
ii. If only Player 2 collapses, Player 1 gets all of the territory, and the game is over.
iii. Otherwise, return to Step 1.
(c) If rejection, Player 1 decides whether or not to go to war.
i. If war, a battle occurs exactly as in Step 2b.
ii. If not, return to Step 1.

The following properties distinguish this model from the Slantchev model (described in the next subsection):

Battle: Attacking is a choice (if an offer is rejected).
War state: In each battle, there is a fixed probability that you win (lose) the overall war and gain (lose) all of the territory.

Counteroffers: There are no offers made by Player 2.
Player 1, the offering side, does not know the probabilities of collapse ( $p_{1}$ and $p_{2}$ ), but Player 2 does know these probabilities. Thus, Player 1 is uncertain about the two sides' relative military strength and, consequently, the feasible agreements. The equilibrium behavior can be described as screening, where Player 1 will make a series of increasingly attractive offers, expecting weaker opponents to accept early in the process, thus screening them out before making the higher offers necessary to appease stronger opponents (Powell, 2004).

## Slantchev Model

The Slantchev model includes an additional variable, military position (in $\{0,1,2, \ldots, N\}$ ), that represents the relative gains made by the two sides in the war so far (Slantchev, 2003). The game under this model proceeds as follows:

1. The initiating player makes an offer of $x \%$ of the territory.
2. The responding player decides to accept or reject the offer.
(a) If acceptance, the responding player gets $x \%$ of the territory, the initiating player gets $(100-x) \%$, and the game is over.
(b) If rejection, continue to Step 3.
3. Battle occurs, and both sides incur the fixed costs. Player 1 wins the battle with probability $p$, Player 2 with probability $1-p$.
(a) If Player 1 wins, military position increases by 1 . If it reaches $N$, then Player 1 receives all the territory and the game is over.
(b) If Player 2 wins, military position decreases by 1 . If it reaches 0 , then Player 2 receives all the territory and the game is over.
4. Return to Step 1 with initiating and responding players reversed.

This model deviates from Powell's as follows:
Battle: There is a battle every round.

War state: A single battle does not directly end the war, but affects the military position variable. Collapse occurs only if military position hits its maximum or minimum value.

Counteroffers: Both sides alternate in making offers.
Like the Powell model, Player 1 does not know the battle probability ( $p$ ), but Player 2 does, so the equilibrium behavior again exhibits some screening. However, the Slantchev model provides Player 1 with the additional information source of Player 2's counteroffers. Furthermore, the military position provides another variable for the sides to consider, in that their offering behavior will change depending on which side is in a stronger position in the overall war.

## PsychSim Agents

We implemented both the Powell and Slantchev games within PsychSim, a multiagent framework for social simulation (Marsella, Pynadath, \& Read, 2004; Pynadath \& Marsella, 2005). PsychSim agents have their own goals, private beliefs, and mental models about other agents. They generate their beliefs and behaviors by solving partially observable Markov decision problems (POMDPs) (Kaelbling, Littman, \& Cassandra, 1998), whose quantitative transition probabilities and reward functions are a natural fit for the game-theoretic dynamics of our chosen models of wartime negotiation.

PsychSim agents have a theory of mind that allows them to recursively model other agents (e.g., their beliefs, rewards, etc.), form expectations about their behavior, and choose optimal actions as a best response. With sufficient computation, the PsychSim agent's optimal action corresponds to the equilibrium strategy. However, we can also limit the agent's horizon when computing expected values and the depth of recursion in modeling others. By doing so, the agent can quickly compute approximate best-response actions even when a human opponent deviates from the equilibrium.

The behavior of the PsychSim agents in both the Powell and Slantchev models roughly corresponds to the informed side's equilibrium strategy. For example, when starting with less territory in the Powell model, the PsychSim agent chooses war until its opponent makes an offer that exceeds its threshold of acceptability, computed as a function of the probability of military collapse. Under the Slantchev model, the agent also rejects any offer below a threshold, but that threshold changes based on the current military position. In particular, if the military position favors the agent's side, the threshold is higher than it would be otherwise. The counteroffers made by the agent are similarly lower when the military position is in its favor than they would be otherwise.

While both Powell and Slantchev focused on the case where the uninformed side also started as the satisfied side, we can also model the case where the uninformed side starts as the dissatisfied side. We change the initial distribution of territory from having the human player start with $72 \%$ of the territory as the satisfied side, to having the human player start with only $28 \%$. The PsychSim agent computes its policy of
behavior using the same algorithm in both cases, although the resulting strategy is slightly different. Under Powell, the agent playing the satisfied side (with $72 \%$ of the territory) will no longer attack when receiving an unacceptable offer. Instead, it will simply reject the offer, hoping to avoid a battle that will risk collapse and loss of all its territory. Under Slantchev, the agent's thresholds as the satisfied side will be universally higher than those as the dissatisfied side.

## Wartime Negotiation Study

We used these agents in a study of how people make decisions in wartime negotiation games. The subjects played each model twice: once as the satisfied side (starting with $72 \%$ of the territory) and once as the dissatisfied side (starting with $28 \%$ ), leading to four experimental conditions: Powell72, Powell28, Slantchev72, and Slantchev28. For each condition, the subject played as Player 1 against a PsychSim agent until the two sides agreed on a split or one side achieved a military victory. If neither occurred within 15 rounds, the game terminated with the sides staying at the initial division of territory. In the two Powell conditions, both sides have the same probability of collapse ( $p_{1}=p_{2}=0.1$ ). In the two Slantchev conditions, we use the same probability of winning for Player $1(p=0.3)$ and the same initial military position, with Player 1 slightly closer to losing the war ( 3 on a range from 0 to 10 ).

## Hypotheses

The Powell and Slantchev models yield hypotheses about behavior that we might see in our human subject data:
Screening Behavior The uninformed side tries to find the minimal offer that is acceptable to its opponent. It does so by progressively increasing its offer until its opponent accepts, gradually screening out weaker opponents who are willing to accept lower offers. We expect to see players make these increasing offers under both Powell and Slantchev models.

Principle of Convergence Warfare ceases to be useful when there is no information to gain, at which point the sides can both agree on a settlement. Given the static battle probability and the lack of signaling moves in the Powell model, the potential information gain should be exhausted sooner than in the Slantchev model. As a result, we would expect settlement to be reached sooner under the Powell model, where the only information gain is through rejected offers.

Information Asymmetry Because of the screening behavior, Slantchev claimed, "as war progresses, the outcome becomes less advantageous for the worse informed party." We thus expect settlements that take more rounds to be less favorable to the human players. Furthermore, we expect this trend to be more pronounced under Powell, where the players receive less information than they do under Slantchev.

Total victory Total military victory (i.e., one side winning all of the territory on the battlefield) is rare, as war typically reveals information quickly enough for both sides to reach
settlements instead. The possibility of collapse in a single battle under the Powell model should make total victory much more common. We would thus expect that negotiated settlements (as opposed to total victories) to be less common under the Powell model than under Slantchev.

War avoidance Both models provide incentives for sides to sacrifice territory to avoid a costly battle. We would expect players who expressed a more positive attitude toward war (ATW) to exhibit more of a willingness to engage in war and give up less territory in the final settlement.

Military Asymmetry If the uninformed side is also at a military disadvantage (as is the case for our human players), then we expect it to overestimate its probability of winning the war, thus making lower offers than it would make in the complete-information case. Therefore, we would expect to see lower offers when the players receive less information about their relative military strength (i.e., under Powell) than when they receive more (i.e., under Slantchev).

Battle outcomes A battle in the Slantchev model will make the victor more optimistic and more willing to delay agreement. We would thus expect players to make lower offers after winning a battle than they would after losing a battle.

Starting Territory Our four experimental conditions could engender different reference points (Neale \& Bazerman, 1991) when people play as the satisfied or dissatisfied sides (starting with $72 \%$ or $28 \%$ of the territory, respectively). We hypothesize that satisfied sides will make fewer concessions, as the endowment effect makes players less willing to give up territory already owned (Kahneman, Knetsch, \& Thaler, 1991). Thus, we expect a player to offer less when starting with more territory than when starting with less territory. Similarly, we expect the dissatisfied side will end up with less territory, because any territory gain, however small, would be more likely considered as satisfactory. Furthermore, because the satisfied side has more to lose through a military outcome, we expect that the difference between initial and final territorial splits will be more favorable for the dissatisfied sides.

## Study Population

We recruited 240 participants, of an average age of 35 , via Amazon Mechanical Turk. $51 \%$ of the participants were female, and $49 \%$ were male. $65 \%$ of the participants were from the United States, $29 \%$ from India, and $6 \%$ from other countries. Regarding the participants' highest level of education, $12 \%$ of the participants had some high school or high school diploma, $63 \%$ had some college or college degree, and $25 \%$ had some graduate school or graduate degree. $13 \%$ of the participants used a computer for 1-4 hours a day, $43 \%$ for 5-8 hours a day, and $44 \%$ for more than 8 hours a day.

## Procedure

After being assigned an anonymous ID, each participant read an information sheet about the study and then filled out a background survey. Next, the participant played the negotia-
tion game four times, each time against a different agent from one of the conditions. The order of the four agents the player negotiated with was randomized. During the negotiation, the participant filled out an in-game survey. Following each negotiation game, the participant filled out an opinion survey. We designed the study to be completed within an hour, and the average duration of the study was 32 minutes.

## Measures

Background Survey The background survey asked about the participant's age, gender, nationality, education, computer experience, Attitude Towards War (Dupuis \& Cohn, 2006), Social Orientation (Van Lange, De Bruin, Otten, \& Joireman, 1997) and attitude towards Inappropriate Negotiation (SINS, from (Robinson, Lewicki, \& Donahue, 2000)).
Opinion Survey The opinion survey contained questions regarding the participant's goals during the game, and questions from the Subjective Value Index (SVI) survey on opinions about oneself, the negotiation outcome, the process and the opponent (Curhan, Elfenbein, \& Xu, 2006).

In-Game Survey The in-game survey asked participants to predict the opponent's responses. For example, after making an offer in the Powell game, participants said whether they expected their opponent to accept the offer, reject it or attack.
Game Logs The game logs captured the participant's actions, the PsychSim agent's actions and the world states (e.g. number of troops, military position, and territory).

## Results

We had 240 participants in the study. Each participant played four different games, one under each condition. Data from incomplete games were discarded. In the end, we had 238 games in the Powell72 condition and 239 games in the Powell28, Slantchev72 and Slantchev28 conditions.

## Hypothesis: Screening Behavior

This hypothesis states that the participants' behavior is most likely to be screening, by making incrementally higher offers to find out the lowest offer that satisfies the opponent. We analyzed the dynamics of the participants' offers to see whether they increased, decreased, or stayed the same from one round to the next. The results of Figure 1 show that, by and large, the human players exhibit screening behavior (i.e., more increases than the alternatives).

## Hypothesis: Principle of Convergence

The hypothesis states that under the Powell model, settlement should be reached sooner, compared to the Slantchev model. We compared the number of rounds it took for both sides to reach an agreement under these two models, excluding games that ended with one side winning the war (instead of reaching an agreement). Results show that, contrary to the hypothesis, it took participants significantly more rounds to reach an agreement when interacting with the Powell model


Figure 1: Distribution over participants' offer dynamics.
than when interacting with the Slantchev model ( $p<.0001$, Mean $_{\text {Powell }}=3.13$, Mean $_{\text {Slantchev }}=1.95$ ).

## Hypothesis: Information Asymmetry

This hypothesis predicted that settlements taking more rounds to reach agreement would be less favorable toward the human player, with the effect being more pronounced in the Powell model. Excluding games with no agreement, we analyzed the correlation between the territory participants ended up with and the total number of rounds needed to reach the settlement. In general, there is a weak yet significant negative correlation between the territory participants got in the settlement and the number of rounds needed to reach that settlement ( $r=-.1965, p<.0001$ ), providing evidence in favor of this hypothesis. The correlation is of medium strength both in the Slantchev ( $r=-.2372, p<.0001$ ) and Powell games ( $r=-.2309, p=.0004$ ), failing to demonstrate the hypothesized difference between the two models.

## Hypothesis: Total Victory

Under this hypothesis, we expect that negotiated settlements to be less common under Powell than under Slantchev because of the possibility of immediate collapse in the former. The data bore out this hypothesis, as fewer games in the Powell model ended in a settlement than in the Slantchev model ( $p<.0001$, Mean $_{\text {Powell }}=48.0 \%$, Mean Slantchev $=60.8 \%$ ). It is also interesting to observe that settlements were much rarer when the player started with $28 \%$ territory than when starting with $72 \%\left(p<.0001\right.$, Mean $_{28}=42.77 \%$, Mean ${ }_{72}=66.25 \%$ ).

## Hypothesis: War Avoidance

We hypothesized that players who are more pro-war would be less willing to give up territory, and more willing to go to war. We measured the participants' attitudes towards war (ATW) in the background survey, where higher ATW scores indicate more of a pro-war attitude, and lower ones indicate an antiwar attitude. We did not find significant correlations between ATW and the average offers participants made ( $r=-.0172$, $p=.5950$ ). We also did not find a correlation between ATW and the number of rounds played in the game ( $r=.0220, p=$ .4979). Battles were not a choice in the Slantchev model. In the Powell72 condition, the human player never initiated an attack, because the agent would always do so first. Therefore, we analyzed the correlation between ATW and the number of player-initiated attacks in only the Powell28 condition and


Figure 2: Change of offers made after battle outcomes.
found a marginally significant weak correlation ( $r=.1107$, $p=.0877$ ). Thus, there was only the slightest of evidence in favor of his hypothesis.

## Hypothesis: Military Asymmetry

This hypothesis states that in the Powell model, where less information is revealed to the players, the players will make lower offers than they would under Slantchev. However, when interacting with the Powell model, participants made slightly higher offers than when they interacted with the Slantchev model ( $p<.0001$, Mean ${ }_{\text {Powell }}=36.26$, Mean $_{\text {Slantchev }}=33.29$ ), exactly the opposite of our hypothesis.

## Hypothesis: Battle Outcomes

We expect players to make lower offers after winning a battle than they would after losing. We ignore the Powell model, which gives players no information about battle outcomes (beyond game-ending collapses). Under Slantchev, players lost $68 \%$ of the battles and won only $32 \%$. When players won a battle, the offers they made next were significantly lower than when they lost $\left(p=.0441\right.$, Mean $_{\text {Won }}=31.70$, Mean $_{\text {Lost }}=35.88$ ), thus bearing out the hypothesis.

## Hypothesis: Starting Territory

Initial Offer We compared the offers that participants made at the beginning of the negotiation. Participants made significantly higher offers when starting with $28 \%$ territory compared to $72 \%\left(p<.0001\right.$, Mean $_{28}=34.20$, Mean $_{72}=27.95$ ).
Average Offer We compared the average offers the participants made during the negotiation. When starting with $28 \%$ territory, participants made significantly higher offers than with $72 \% ~\left(p<.0001\right.$, Mean $_{28}=38.84$, Mean $_{72}=30.71$ ).

End Territory We compared the percentage of territory the participant had when the game ended. When starting with $28 \%$ territory, participants ended up with signifi-


Figure 3: Participant offers in response to opponent actions.
cantly less territory than when starting with $72 \%$ ( $p<.0001$, Mean $_{28}=22.31$, Mean $_{72}=45.55$ ).

Net Territory Gain Beyond the impact on absolute territory, we also hypothesized that the starting territory would affect the relative gain/loss in territory from the beginning to the end of the game. When starting with $28 \%$ territory, participants lost significantly less territory than when they starting with $72 \%$ territory ( $p<.0001$, Mean $28=-5.69$, Mean ${ }_{72}=-$ 26.45). Thus, the observed behavior conformed to all of our expectations about the effect of the initial division of territory.

## Offers In Reaction to Opponent Actions

We also compared the participants' offers under both models in reaction to their opponent's actions. In the Powell model, when not accepting the participant's offer, the opponent chose to either simply reject the offer or to attack the participant. In the Slantchev model, attacking was not a choice, but the opponent did make counteroffers when not accepting the participant's offer. The differences were significant when interacting with the Powell model ( $p<.0001$ ) and Slantchev model ( $p<.0001$ ). The results of Figure 3 show that attacking the participants prompted lower offers, while less aggressive actions (e.g. rejecting without attacking, or making a counteroffer) typically resulted in higher offers.

## Discussion

As we can see from the previous section, much of the observed behavior conformed to the expectations generated by the theoretical analyses of the Powell and Slantchev models. The information asymmetry that is critical to both models had the expected impact on the human players, as they clearly exhibited the hypothesized screening behavior. Furthermore, we also observed direct evidence of the claim that "as war progresses, the outcome becomes less advantageous for the worse informed party" (Slantchev, 2003). Our agent-based experimental setup also allowed us to try a variation of the game on the starting territory, and the data provided strong evidence for our hypotheses regarding that variation.

However, there were also some interesting deviations from our hypothesized behavior. Slantchev's Principle of Convergence hypothesized that warfare ceases to be useful when there is no information to gain. Our derived hypothesis
viewed the Powell game as more information-poor than the Slantchev one, leading us to expect that settlements would be reached sooner in the former. However, our data showed the opposite trend. We can at least partly attribute this deviation to people assessing the information gains differently than was prescribed in the two models' equilibrium analysis. Under the Powell model, players may try to gather more observations of their opponents' behavior, because they do not realize that the agent's threshold of acceptance does not change. Conversely, under the Slantchev model, the players may feel overly certain in their beliefs upon observing only a few (or even only one) counteroffers from their opponent.

We also saw little evidence to support our War Avoidance hypothesis. We did not find any link between a participant's attitude towards war and the offers they made, the duration of the war, nor the number of attacks. It is likely that some participants did not carry their attitudes toward war over into this abstract game setting. However, we also need to further differentiate within the ATW scale about why people are antiwar and what types of war they are against.

The deviation from the military asymmetry hypothesis is harder to explain. From the very beginning of a Slantchev game, the players can observe that their military position puts them closer to losing the war than winning the war. As the game progresses, they can potentially observe that the battle probability is not in their favor. We would thus expect players to be more pessimistic about their chances under the Slantchev model and, thus, to make higher offers to quickly appease their opponents. However, players made higher offers in the Powell model, where there was no feedback about the war outcomes, nor the other side's valuations.

Therefore, we need a more fine-grained analysis of when players in Slantchev made the unexpectedly low offers. For example, the data supported our Battle Outcomes hypothesis, where players made lower offers after winning a battle. By isolating such cases, we may see that the general military asymmetry hypothesis holds, but we can also understand the in-game contingencies that would override the general trend.

Furthermore, despite the general conformity over all of the data, not every player's behavior conformed to our hypotheses. We must therefore analyze the data to identify the more fine-grained contingencies and the individual differences among our participants. Such an analysis would give us a better understanding of how the participants viewed, for example, the potential information received in the game, their possible military outcomes, etc. This analysis will also guide future studies by suggesting further instrumentation that is needed to gather the required in-game data.

With these additional analyses and data, we can build upon the field's game-theoretic models to develop higher-fidelity models of human behavior in such wartime negotiation scenarios. Ideally, these models will help bridge the gap between the theoretical computational frameworks and the decisionmaking we see in the real world. By doing so, they will provide an invaluable computational tool for political scientists
to explore a richer set of contingencies and individual differences and hopefully provide better predictions and explanations of behavior in wartime negotiation.

## Acknowledgments

This work was sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM). The content does not necessarily reflect the position or the policy of the Government. No official endorsement should be inferred.

## References

Curhan, J. R., Elfenbein, H. A., \& Xu, H. (2006). What do people value when they negotiate? Mapping the domain of subjective value in negotiation. J. of Personality and Social Psychology, 91, 493-512.
Dupuis, E. C., \& Cohn, E. S. (2006). A new scale to measure war attitudes: Construction and predictors. J. of Psychological Arts and Sciences, 3(1), 6-15.
Fearon, J. D. (1995). Rationalist explanations for war. International Organization, 49, 379-414.
Filson, D., \& Werner, S. (2002). A bargaining model of war and peace: Anticipating the onset, duration, and outcome of war. American J. of Political Science, 46, 819-838.
Kaelbling, L. P., Littman, M. L., \& Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. Artificial Intelligence, 101, 99-134.
Kahneman, D., Knetsch, J. L., \& Thaler, R. H. (1991). Anomalies: The endowment effect, loss aversion, and status quo bias. The J. of Economic Perspectives, 5(1), 193206.

Marsella, S. C., Pynadath, D. V., \& Read, S. J. (2004). PsychSim: Agent-based modeling of social interactions and influence. In ICCM (p. 243-248).
Neale, M. A., \& Bazerman, M. H. (1991). Cognition and rationality in negotiation. Free Press.
Powell, R. (2001). Bargaining while fighting. Doctoral dissertation, University of California, Berkeley.
Powell, R. (2004). Bargaining and learning while fighting. American J. of Political Science, 48(2), 344-361.
Pynadath, D. V., \& Marsella, S. C. (2005). PsychSim: Modeling theory of mind with decision-theoretic agents. In IJCAI (p. 1181-1186).

Robinson, R. J., Lewicki, R. J., \& Donahue, E. M. (2000). Extending and testing a five factor model of ethical and unethical bargaining tactics: Introducing the SINS scale. J. of Organizational Behavior, 21, 649-664.
Slantchev, B. L. (2003). The principle of convergence in wartime negotiations. American Political Science Review, 97, 621-632.
Smith, A., \& Stam, A. C. (2004). Bargaining and the nature of war. The J. of Conflict Resolution, 48(6), 783-813.
Van Lange, P. A. M., De Bruin, E. M. N., Otten, W., \& Joireman, J. A. (1997). Development of prosocial, individualistic, and competitive orientations: Theory and preliminary evidence. J. of Personality and Social Psychology, 74(4), 733-746.

# Using Property Induction to Evaluate Understanding of Mixing 

Connor Quinn ${ }^{1}$ (cq209@cam.ac.uk)<br>Michelle R. Ellefson ${ }^{1}$ (mre33@cam.ac.uk)<br>Anne Schlottmann ${ }^{2}$ (a.schlottmann@ucl.ac.uk)<br>Keith S. Taber ${ }^{1}$ (kst24@cam.ac.uk)<br>${ }^{1}$ Faculty of Education, University of Cambridge<br>${ }^{2}$ Psychology and Language Sciences, University College London


#### Abstract

Although reasoning skills have been investigated in a number of different domains, very little is known about how children and adults use them in chemistry. Here, participants from 4 years to adults saw various mixtures presented using a standard property induction paradigm. The category and appearance of everyday materials were varied to assess the extent that participants use these features to inform their judgments about what happens when these materials are mixed with water. In general, the results followed similar patterns seen when this paradigm has been applied to other domains, with both category and appearance informing inductive generalizations. The findings contrast with interview-based measures of children's understanding of chemistry and offer an important addition to the field.


Keywords: property induction; chemistry; reasoning; cognitive development

## Background

There is a growing consensus that children learn and reason about novel situations by basing their generalizations on their previous experiences (e.g., Wellman \& Gelman, 1998). Children have extensive experience of chemistry in their everyday world, e.g., baking or rusting. However, there are few studies in cognitive science exploring children's reasoning about the chemical world. Here, we present a novel application of a property induction paradigm to investigate how primary school children (ages 4 to 11) reason about one basic chemical phenomenon - the mixing of different materials. ${ }^{1}$

[^193]The focus of this study is mixing because it is one of the earliest chemical phenomena children are deemed capable of grasping (e.g., Au, Sidle, \& Rollins, 1993; Johnson, 2000; Rosen \& Rozin, 1993) and because very little work exists in this area (Çalýk, Ayas, \& Ebenezer, 2005). Most of the few existing studies have used interviews. The results of these interviews suggest that young children attend almost exclusively to what they can see, i.e., the macroscopic properties of the materials (Arnold, Moye, \& Winer, 1986; Ebenezer \& Erikson, 1996; Haider \& Abraham, 1991) and have little or no conception of the particulate nature of matter (Liu \& Lesniak, 2006; Nakhleh \& Samarapungavan, 1999; Renström, Andersson, \& Marton, 1990). Briefly, the particulate nature of matter refers to the idea that materials are made up of invisible, sub-microscopic particles, with molecules being the smallest particles of most materials. Some knowledge of the particulate nature of matter is necessary to understand materials and how they interact with each other; naïve (incomplete or incorrect) understanding of particles likely leads to misconceptions of chemical phenomena. The assumption is that because young children are not able to explain the particulate nature of matter or the microscopic properties of materials that they lack the ability to reason adequately about materials

One issue with these findings is that the interview method relies on children having the appropriate language of chemistry to be able to explain the phenomena. As a result, it may be the case that children's abilities in this area have been greatly underestimated. Extensive studies of naïve physics and naïve biology indicate that children's reasoning abilities about these science phenomena surpass their abilities to explain them verbally. Tasks that are not reliant on verbal ability indicate that even infants have some understanding of physics (Wellman \& Gelman, 1998). For example, infants know that solid objects cannot just appear/disappear or move through physical barriers (Spelke, Breinlinger, Macomber, \& Jacobson, 1992), are distinct from one another (Xu \& Carey, 1996), and once put into motion travel over distances related to the force of that motion (Kotovsky \& Baillargeon, 1998). These tasks indicate young children do have some appreciation of the properties of materials; 3-year-olds know that wooden pillows are hard (Kalish \& Gelman, 1992) and 4-year-olds know that material is conserved if the object is broken up
(e.g., a plastic toy taken apart is still plastic even if it no longer operates as a toy; Smith, Carey, \& Wiser, 1985).

The success of these language-sparse experimental paradigms in uncovering the foundations of young children's emerging understanding in naïve physics and biology might suggest that young children can make sense of chemical phenomena. Here, we use a language-sparse property induction paradigm to study early chemistry reasoning. Briefly, the property induction paradigm investigates how children use category and appearance information in their generalizations of natural kinds, typically biological kinds (Gelman \& Markman, 1986; Gelman \& Markman, 1987). For example, Farrar, Raney, \& Boyer (1992) showed 5- to 10-year-old children a familiar target object with its familiar name (e.g., egg) and taught them a novel property about that object (e.g., 'has mitochondria inside'). Next, children were asked whether the four test items below also had that novel property:

1. Same category, same appearance (e.g., plain egg);
2. Same category, different appearance (e.g., spotted egg);
3. Different category, same appearance (e.g., snow ball);
4. Different category, different appearance (e.g., leaf).

At all ages, generalizations depended both on category and appearance, but how children relied on these cues changed with age. Pre-school children generalized more to objects in the same category with the same appearance than to the other items; in other words, they thought the typical cue correlation was necessary. Second graders generalized more to objects that matched in category and appearance than to objects matching in only one cue than to objects matching in neither cue; that is, they realize that category and appearance are separable predictors. Only fourth graders generalized more to same category, different appearance items than to different category, same appearance items, realizing that category was a better predictor than appearance. This mature pattern appeared even for second graders in a second study varying knowledge of the categories/properties in question, but only when children reasoned about known categories/properties. Generalization about materials may include more features than category and appearance. For example, 8-year-olds seem to generalize more often to items with matching causal information compared to perceptual features and 5 -year-olds seem to be able to make use of causal information when it is not in competition with physical features (Hayes \& Thompson, 2007). This distinction may be relevant for chemistry where the causal factors that determine mixing outcomes may not correspond to perceptual features. The results from these property induction studies have indicated that children as young as 2 years are not limited to appearance-based reasoning when categories/properties are well known (Gelman \& Coley, 1990), but variations in knowledge continue to play a vital role at older ages.

Given the success of this paradigm in furthering the understanding of young children's reasoning, it seems well suited as an application for the chemical phenomena investigated here - mixing. More specifically, do children
generalize from one mixture outcome to another if the substances involved are of the same category or of the same appearance? How does this depend on age and on children's knowledge of the substances involved?

In contrast to studies of biological properties, generalization of mixture properties does not depend on category only, but on appearance as well. Mapping the category and appearance properties onto chemistry, it might be useful to think of categories in chemistry as relating to materials and appearances in chemistry as relating to forms like powder, granule or larger chunks. Whether different materials dissolve in water or not depends on a variety of factors related to molecular structure. For instance, water is polar and can break other polar or ionic materials like salt ( NaCl ) apart, but not non-polar or covalent materials like sand (SiO2); roughly, like dissolves like. How different forms of a material mix with water might depend on factors related to surface area. For instance, table salt ( NaCl , in granular form) usually dissolves more quickly in water than rock salt ( NaCl , in a large chunk) because it has a greater surface area. As such, in addition to examining the role of language in children's reasoning about basic chemical phenomena, the current design allows for an investigation of whether children's generalizations about chemical properties are similar to those in other domains. Specifically, will reasoning about chemistry follow both material (category) and form (appearance) cues in the same way as for biology, will there be a different pattern for chemistry, reflecting domain differences in cue efficacy, or will young children remain appearance-bound, as predicted by the findings from interview studies?

## Method

## Participants

A total of 142 participants $\left(N_{\text {female }}=81\right)$ took part in this experiment. There were 122 children recruited from schools in eastern England, including 24 children from reception ( $M$ $=4.87$ years, $S D=0.35, N_{\text {female }}=11$ ), 32 children from year two ( $M=6.62$ years, $S D=0.46, N_{\text {female }}=19$ ), 33 from year four ( $M=8.60$ years, $S D=0.41 N_{\text {female }}=15$ ) and 35 from year $\operatorname{six}$ ( $M=10.74$ years, $S D=0.28, N_{\text {female }}=21$ ). In addition, 20 adult participants were recruited from the university and local community ( $M=26.45, S D=6.70$, $N_{\text {female }}=15$ ). For simplicity, these different age groups are referred to here as 5 -year-olds, 7 -year-olds, 9-year-olds, 11 -year-olds, and adults. Adults were paid $£ 10$ for their participation and represented a range of chemistry experience. Children were invited to dress up as scientists for the duration of the study and were given stickers and their schools given a special science presentation by a local science outreach program. The participating schools were typical schools in terms of their range of student abilities and backgrounds according to publically available government data (www.ofsted.gov.uk).


Figure 1. An example of a target. The picture on the left shows the water and the target (e.g., granulated brown sugar) before mixing. The picture on the right shows the water and target after they were mixed.


Figure 2. An example of a set of probes: (1) same material, same form (e.g., granulated brown sugar); (2) same material, different form (e.g., brown sugar cube); (3) different material, same form (e.g., sand); and (4) different material, different form (e.g., a pebble).

## Materials

Everyday items (e.g., sugars, salts, sand, etc.) were selected as stimuli because children may reason better about familiar content and for safety reasons. Twelve sets of items were chosen, with each set including a target and four probes (see Table 1). The probes followed the conditions mentioned above: (1) same material, same form; (2) same material, different form; (3) different material, same form; and (4) different material, different form.

Several constraints were imposed on the selection of the targets and probes based on pragmatics and the experimental design: (1) the target and probes were safe and appropriate for use with young children; (2) the target-probe pairs had similar appearances for their matching forms (solid, granule, or powder); and (3) the targets and probes were balanced in terms of their outcomes when mixed with water. When controlling for mixing outcomes it was noticed that long names, (e.g., antacid) were often associated with exciting outcomes such as fizzing. To avoid this possible confound
some items were given alternative names ${ }^{2}$. Finally, the relative mass and volume of the targets and probes were as similar as possible so that these perceptual features would not act as additional cues to the outcomes. Transparent 400 mL plastic beakers (see Figure 1) were used to show the mixing of each target with water. The beakers were filled with 250 mL of water and had lids to allow mixing of the targets with the water without risk of spillage. The probes were presented in transparent 140 mL plastic containers (see Figure 2), sealed with clear plastic lids for safety.

## Procedures

Participants sat opposite the experimenter at a table in a quiet area of their primary school or university. A clear plastic beaker with water was placed on the table and identified as water. A transparent plastic tub containing the target was displayed and identified for the participant, using the phrase "See this tub? This tub has [target name]. I'm going to mix the [target name] with the water.'

All items were named for the participants. In order to ensure no cues about the type of material could be implied from the instructions, mass/count words were not used (e.g., "This is a vitamin." or "This is some sugar."). Instead, only general names were given (e.g., "This is vitamin." or "This is sugar."). The form of the target and probe were not mentioned.

The target was added to the water, the beaker was sealed and it was turned upside down once to facilitate mixing. Participants were asked to describe what happened both to the target and to the water. This step ensured they were attending to the mixing.

Table 1: List of Target and Probe Materials

| Target <br> Form | Target Material | Probe <br> Material |
| :---: | :---: | :---: |
| Solid | Chalk | Lolly |
|  | Chocolate | Almond |
|  | Vitamin | Sweet |
|  | Paint | Incense |
| Granule | Peppercorns | Candy |
|  | Bath Bomb | Wax |
|  | Sugar | Stone |
|  | Coffee | Stock cube |
| Powder | Coconut | Soap |
|  | Antacid | Washing Soap |
|  | Salt | Rice |
|  | Kool-Aid | Play-Doh |

After mixing the target with water, the experimenter displayed and identified each of the four probes, one at a

[^194]time in a pre-established randomized ordering. For each probe participants were asked if it would do the same as the target using the phrase 'See this tub? This tub has [probe name] in it. Do you think this would do the same as [target name] if I put it in water?'

Participants were instructed to give "Yes" or "No" replies. Simplifying the required responses in this way was important in order to make the task accessible for the youngest participants. For the younger groups two sheets of paper were also available, green and red, with 'yes' and 'no' written on them respectively. Children could point to these if they did not give a verbal response. Only one child made use of these sheets. If participants did not give a specific 'yes' or 'no' response, the question was repeated to prompt a 'yes' or 'no' answer. Sessions were video recorded so that replies could be verified and confirmed off-line.

The target remained in view on the table mixed with the water while the participants saw the probes. After each of the four probes was presented, the target and water were cleared out of view before the next set of items was presented.

There were 12 sets of items each containing a target to be mixed with water and four probes for a total of 48 trials. Both the order of the 12 sets and the order four probes within each set were presented pseudo-randomly. The youngest group always completed the study in two separate sessions.

## Results

The proportion of "Yes" responses given by the participants to the probes were analyzed using a $5 \times 2 \times 2$ repeated measures ANOVA with the between-subjects factor of age group (5-year-olds, 7 -year-olds, 9 -year-olds, 11 -year-olds, and adults) and the within-subjects factors of material (same vs. different from the target) and form (same vs. different from the target). This ANOVA was conducted using the restricted maximum likelihood technique ( $R E M L$; Bagiella, Sloan, \& Heitjan, 2000). There were no overall significant differences among the three forms (powder, granule, or solid), making it feasible to combine them together and focus the analyses on same vs. different form only.

There was a significant effect of both material, $F(1$, $139.1)=445.44, p<.0001$, and form, $F(1,139.2)=418.56$, $p<.0001$, as well as a significant interaction between material and form, $F(1,139.3$ ) $=119.42, p<.0001$ (See Figure 3). More specifically, participants responded 'Yes" most often when the probe was the same material, and same form as the target ( $M=.96, S D=.21$ ), followed by probes that were the same material and different form ( $M=.59, S D$ $=.49$ ), probes that were a different material and same form ( $M=.49, S D=.50$ ), and probes that were a different material and different form ( $M=.30, S D=.46$ ). Post-hoc tests using Tukey's HSD indicated that each probe type was different from the others.


Figure 3. The mean proportion of "Yes" responses made to the same and different materials and forms across all age groups.


Figure 4. The mean proportion of "Yes" responses made to the same and different materials and forms by each age group.

The main effect of age group was not significant, $F$ (4, $139.1)=0.89, p=.47$. Similarly, age group did not interact significantly with material, $F(4,139.1)=2.03, p=.09$, form, $F(4,139.1)=1.15, p=.34$, or material and form combined, $F(4,139.2)=1.53, p=.20$ (see Figure 4).

## Discussion

This paper presents a novel adaptation of the property induction paradigm to explore how children reason about the chemical process of mixing. The design was created in order to use language-sparse methods as a way of further examining children's reasoning in this domain by addressing whether: (1) young children display a better understanding of mixing processes when assessed using a language-sparse method compared to interviews; and (2) whether children differentially attend to the category (material) or appearance (form) of materials when generalizing about mixing.

In terms of the first question, the results confirm that young children's reasoning about these materials does not differ from older children and adults in terms of mixing in
this context. These findings are consistent with other property induction studies, but are contrary to the results of interview studies (Gelman \& Markman, 1986; Gelman \& Markman, 1987; Liu \& Lesniak, 2006; Au, Sidle, \& Rollins, 1993). As such, there is some indication that this type of language-sparse methodology might be useful in further exploring how young children reason about other chemical phenomena.

In relation to the second question, participants of all age groups attend to category and appearance when making generalizations about mixing. The presence of this finding for the youngest age group suggests that even young children bring their everyday reasoning skills to understanding chemistry despite not yet being able to articulate sophisticated explanations. The findings presented here replicate the overall pattern found with property induction studies in other domains (e.g., Gelman \& Markman, 1987). Specifically, the category seems to have more influence than appearance on the generalizations that were made.

However, these findings are distinctive to property induction studies in other areas. Specifically, this study did not replicate the common finding of age related differences in the use of category and appearance. In a chemistry context like that presented here, both features seem to be influencing generalizations, whereas in other studies from the domain of biology the categorical information becomes more important for generalizations in older children than younger children (e.g., Farrar et al., 1992).

One explanation for the discrepancy between the results found here and other property induction studies might be that this task used naturalistic materials and actual mixing events, whereas most of the previous studies used pictures, words or text (Farrar et al., 1992; Gelman \& Coley, 1990; Gelman \& Markman, 1986; Gelman \& Markman, 1987; Hayes \& Thompson, 2007). More specifically, these previous studies mostly frequently used artificially selected stimuli with a constrained set of properties that allowed for a limited number of inductions, whereas the materials used here are more ecologically valid but they do include a wider variety of properties and more possible inductions.

Using real materials might have inadvertently allowed participants to attend to properties other than the category and appearance properties explicitly examined here (e.g., the density of objects could have been assumed by participants to have played a role in the outcome). In contrast, when experimental stimuli are created to vary only on a limited set of properties, then participants may base their generalizations more on the specific properties for which these artificial stimuli were designed to control. Thus, the inherent complexity of real-world materials might have prevented well-controlled and systematic studies of reasoning about chemistry. This language-sparse design provides a platform from which additional studies might be developed that control for the wide variety of features that may play a role when natural stimuli are used in property induction studies, while still being more ecologically valid.

It may be the case that children exploit multiple redundant cues in their natural environment, so the pattern found here may be indicative of their reasoning in their everyday lives.

Another reason for this finding in chemistry, but not biology might be due to domain-specific differences in the way chemistry information is processed. It could be the case that reasoning skills are applied differently in the biological and chemical contexts because the features that help in terms of generalization have different predictive validity.

In chemistry, appearance might be both an unreliable predictor and necessary for making a prediction. Firstly, appearance alone is generally an unreliable predictor of category. For example, white powder can be any number of different materials with a wide range of possible chemical properties. Secondly, it is difficult to make a prediction about the outcome based on the knowledge of the category without information about appearance. For example, aluminum is inert as a solid block, but easily combusts in powdered form. In contrast, in biology, appearance might be a reliable predictor of behavior when it predicts category membership (e.g., wings might predict bird and flying).

Most property induction studies introduce unreliable correlations amongst features like appearance and category and assuming that biology naturally includes more reliable correlations amongst these features, then it could be the case that property induction studies introduce unnatural reasoning settings. As such, the age difference apparent in biology generalizations may reflect children's growing understanding of what to do when the correlations they experience in their everyday lives are broken by our experimental designs in property induction. On the other hand, in chemistry, the correlations are naturally unreliable, matching the usual property induction design. Thus, the property induction paradigm might be more representative of naturalistic reasoning in chemistry but not biology. If that were the case, then the more mature reasoning seen in this chemistry context might be related to the match between the experimental design and children's everyday experiences rather than the differences inherent to reasoning about biology and chemistry.

This aspect of chemistry raises an intriguing perspective on development of categorization and reasoning skills. Cognitive science includes a large body of studies investigating the basic building blocks of cognition with which children learn about the world. In physics, biology, and psychology evidence has suggested that children generalize existing knowledge to extend their ability to reason about the world. A debate remains about the origins of these basic reasoning skills. Most areas of reasoning struggle to separate the question of how much of reasoning is dependent upon domain-specific experiences and how much is due to domain-general strategies. Chemistry may offer a unique perspective for this debate. Like physics, biology, and psychology, children are exposed to chemical phenomena throughout childhood, but the differences between these domains in terms of predictive validity of features might provide a new direction for further study.

If it is indeed the case that children are sensitive to the idea that different cues are meaningful in biology versus chemistry, then it could be the case that children are bringing very sophisticated reasoning skills to their attempts to understand chemistry. However, this is the first study in this area and other relevant cues (e.g., density, naming, etc.) should to be explored before firm conclusions can be made.

In sum, this novel application of the property induction paradigm to chemistry raises important questions about the development of reasoning skills in chemistry and further offers directions of research to address key questions of how children learn to reason about the world. The question of how abstract reasoning skills develop is a core issue for education. Previous research into young children's understanding of chemistry has relied upon language-based measures. This study offers a more sensitive measure of chemistry reasoning that is not constrained by a child's language development. The findings presented here might be useful in re-evaluating the assumptions that educators make about the reasoning skills children bring to chemistry learning and could be applied to develop more effective ways of learning for chemistry students of all ages.

## Acknowledgements

This project was supported by The Leverhulme Trust (No. RPB-115). Special thanks to Polly Basak and Hyunji Kim for help with data collection and Elaine Wilson for advice and support. Further thanks to Katja Rieger, Gina Plant Cabrita, Sarah Laupper, and Matthew Cairnduff for help with selection and preparation of materials.

## References

Arnold, K. D., Moye, J., \& Winer, G. A. (1986). Illusion versus reality: Children's understanding of temperature adaptation. Journal of Experimental Child Psychology, 42(2), 256-272.
Au, T. K., Sidle, A. L., \& Rollins, K. B. (1993). Developing an intuitive understanding of conservation and contamination: Invisible particles as a plausible mechanism. Developmental Psychology, 29(2), 286-299.
Bagiella, E., Sloan, R. P., \& Heitjan, D. F. (2000). Mixedeffects models in psychophysiology. Psychophysiology, 37(1), 13-20.
Çalýk, M., Ayas, A., \& Ebenezer, J. V. (2005). A review of solution chemistry studies: insights into students’ conceptions. Journal of Science Education and Technology, 14(1), 29-50.
Ebenezer, J. V., \& Erickson, G. L. (1996). Chemistry students' conceptions of solubility: A phenomenography. Science Education, 80(2), 181-201.
Farrar, M. J., Raney, G. E., \& Boyer, M. E. (1992). Knowledge, concepts, and inferences in childhood. Child Development, 63(3), 673-691.
Gelman, S. A., \& Coley, J. D. (1990). The importance of knowing a dodo is a bird: Categories and inferences in 2-
year-old children. Developmental psychology, 26(5), 796-804.
Gelman, S. A., \& Markman, E. M. (1986). Categories and induction in young children. Cognition, 23(3), 183-209.
Gelman, S. A., \& Markman, E. M. (1987). Young children's inductions from natural kinds: The role of categories and appearances. Child Development, 58(6), 1532-1541.
Haidar, A. H., \& Abraham, M. R. (1991). A comparison of applied and theoretical knowledge of concepts based on the particulate nature of matter. Journal of Research in Science Teaching, 28(10), 919-938.
Hayes, B. K., \& Thompson, S. P. (2007). Causal relations and feature similarity in children's inductive reasoning. Journal of Experimental Psychology: General, 136(3), 470-484.
Johnson, P. (2000). Children's understanding of substances, part 1: recognizing chemical change. International Journal of Science Education, 22(7), 719-737.
Kalish, C. W., \& Gelman, S. A. (1992). On Wooden Pillows: Multiple Classification and Children's Category-based Inductions. Child Development, 63(6), 1536-1557.
Kotovsky, L., \& Baillargeon, R. (1998). The development of calibration-based reasoning about collision events in young infants. Cognition, 67(3), 311-351.
Liu, X., \& Lesniak, K. (2006). Progression in children’s understanding of the matter concept from elementary to high school. Journal of Research in Science Teaching, 43(3), 320-347.
Nakhleh, M. B., \& Samarapungavan, A. (1999). Elementary school children's beliefs about matter. Journal of Research in Science Teaching, 36(7), 777-805.
Renström, L., Andersson, B., \& Marton, F. (1990). Students' conceptions of matter. Journal of Educational Psychology, 82(3), 555-569.
Rosen, A. B., \& Rozin, P. (1993). Now you see it, now you don't: The preschool child's conception of invisible particles in the context of dissolving. Developmental Psychology, 29(2), 300-311.
Smith, C., Carey, S., \& Wiser, M. (1985). On differentiation: A case study of the development of the concepts of size, weight, and density. Cognition, 21(3), 177-237.
Spelke, E. S., Breinlinger, K., Macomber, J., \& Jacobson, K. (1992). Origins of knowledge. Psychological Review, 99(4), 605-632.
Wellman, H. M., \& Gelman, S. A. (1998). Knowledge acquisition in foundational domains. In W. Damon (Ed.), Handbook of child psychology: Volume 2: Cognition, perception, and language (pp. 523-573) Hoboken: Wiley.
Xu, F., \& Carey, S. (1996). Infants' metaphysics: The case of numerical identity. Cognitive Psychology, 30(2), 111153.

# Does the Semantic Integration of Emotion Words Depend on Emotional Empathy? N400, P600 and Localization Effects for Intentional and Proprioceptive Emotion Words in Sentence Contexts 

Natalia Rak ${ }^{\mathbf{1}}$, Jarmo Kontinen ${ }^{2}$, Lars Kuchinke ${ }^{1}$, and Markus Werning ${ }^{2}$<br>(\{natalia.rak, jarmo.kontinen, lars.kuchinke, markus.werning\}@rub.de)<br>${ }^{1}$ Department of Psychology, Ruhr University Bochum<br>${ }^{2}$ Department of Philosophy, Ruhr University Bochum


#### Abstract

Empathy with other persons' emotions has been suggested to root in a simulation process involving brain regions that play a crucial role in the production of one's own emotions. The current ERP study combines this approach with an embodiedsimulative view of semantics. This view implies that those very brain regions should also be involved in the semantic memory and linguistic comprehension of intentional and proprioceptive emotion words. The relation between cognitive empathy measured by the MET test and the size of the N400 effect occurring when semantic emotions words violate semantic expectations is investigated.


Keywords: empathy, semantic memory, N400, P600, emulative semantics, embodied cognition, simulation theory, mirror neuron system, emotion, proprioception, multifaceted empathy test (MET)

## Introduction

The current debate about the neural realization of linguistic meaning and semantic memory can be characterized by two opposing views: According to the abstract-symbolic view, semantic memory is a modular and amodal system. Semantic representations are considered as rather stable, decontextualized mental symbols that are processed in a largely informationally encapsulated way and do not essentially recruit mechanisms from perceptual, motoric or emotional brain processes. Combining various neurolinguistic findings, Friederici (2002) e.g. argues that "semantic processes are mainly subserved by the left temporal region and that the frontal cortex is recruited when strategic and/or memory aspects come into play".
The embodied-simulative view, in contrast, assumes that the processing of linguistic meaning essentially involves perceptual, motoric and emotional brain regions corresponding to the contents of the words to be comprehended. Based on a review of neurobiological data, Pulvermüller (1999) suggests that neural assemblies that pertain to the sensory-motor cortices and are bound by neural synchronization play an important role in understanding the meanings of words and sentences. These cortical sensory-motor action and perception circuits are interdependent in language comprehension. According to Barsalou (2005) semantic representations can be regarded as simulators of sensory-motor and emotional contents. Werning (2012) has coined the notion of Emulative Semantics and proposes a compositional, but non-symbolic
recurrent neural network model that generates simulations for semantic representations.
Support for the embodied-simulative view comes from a number of neuro-linguistic studies especially in the domain of action words. Neuroimaging investigations have shown that the linguistic comprehension of verbal stimuli involve motor circuits, i.e. specific motor activations can be found when subjects understand speech sounds, word meanings, semantic categories and sentence structures (Pulvermüller \& Fadiga, 2010) involving action words or words associated with actions. FMRI studies (Pulvermüller, 2005) regarding the understanding of action verbs, e.g., hint at a differential top-down activation of motor and pre-motor areas. Martin (2007) reports that the understanding of concrete nouns like hammer, for which not only features, but also affordances are salient, results in an activity distributed over the premotor and the visual cortex. Brain areas involved in motor control contribute to neural networks in which verb representations are grounded. Studies on motor deficits such as Parkinson disease, e.g., reveal impairment of patients' action naming (Rodríguez-Ferreiro et al., 2009).

Embodied-simulative accounts of the semantics of action words have been linked to mirror neuron systems. Mirror systems have been reported in humans not only for actions (Rizzolatti et al., 1996), but also for intentional emotions (Bastiaansen et al., 2009; disgust - Wicker et al., 2003; facial expressions - Carr et al., 2003), and proprioceptive emotions (pain - Avenanti et al., 2005; touch - Blakemore et al., 2005). Mirror neuron systems map the perceptions of actions and intentional as well as proprioceptive emotions of an observed person onto the perceiver's own somatosensory, viscero-motor, or motor representations of actions and emotions. Such a mapping is supposed to enable the observer of another person's actions and emotions to feel as if he were performing that action or experiencing that emotion himself. Since mirroring mechanisms may constitute sub-personal instantiations of embodied simulations, Gallese (2003) proposes mirror neuron systems as a neuronal basis of empathy.

These findings and theoretical considerations lay the ground for the current study. If the embodied-simulative view of linguistic meaning also applies to emotion words, the processes underlying empathy with other persons' emotions should be not entirely independent of processes underlying the comprehension of emotion words.

Table 1: Stimuli

|  | CON (congruent) | INCON (incongruent,) | UNREL (unrelated) |
| :---: | :---: | :---: | :---: |
| INT (intentional emotion) | Als Adrian von seinen hohen Gewinnen an der Börse erfährt, ist er darüber sehr erfreut. <br> When Adrian hears about his high gains on the stock market, he is very happy about it. | ... besorgt. <br> concerned | ...empört. <br> indignant |
| PROP (proprioceptive emotion | Nachdem Kerstin stundenlang ohne Wasser in der Hitze umherlief, ist sie nun sehr durstig. <br> After Kerstin has been walking around without water in the heat, she is now very thirsty. | ...hungrig. <br> hungry | ...hellwach. <br> awake |
| PHYS (physical control) | Da niemand die Türen jemals geölt hatte, begannen sie nach kurzer Zeit zu quietschen. <br> Since nobody had oiled the doors, they soon began to squeak. | ...bollern. <br> thud | ...tröpfeln. drip |

Furthermore, since the capacity and inclination to empathize with other persons' emotions varies across subjects, we consider it an interesting question whether good emotional empathizers "feel" semantic violations in the context of emotion words more strongly than poor emotional empathizers. One should thus predict that this results in stronger N400 effects.

The examination of the N400 effect in the event-related potentials (ERPs) is a common approach to investigate semantic integration in sentence processing. An N400 is a is a monophasic negativity between 200 and 600 ms after word onset, largest over centro-parietal sites (Kutas \& Federmeier, 2011). When comparing semantically expected and unexpected words, observed higher amplitudes of the N400 are discussed to reflect greater demands of semantic integration of an unexpected word at the sentence or the discourse level. Thus, N400 effects are particularly observed for critical words that do not fit into a sentence's context. Recent evidence from sentence processing has shown that the integration of contextual semantic information is dependent on emotional processing (Chwilla et al., 2011; Federmeier \& Kutas, 2011; Pinheiro et al., 2013).

## Methods

## Participants

25 female students from the University of Bochum ( $\mathrm{M}=24.19, \mathrm{SD}=2.58$ ) volunteered for the experiment. They were compensated with $10 €$ per hour for their time and effort. They were recruited through local advertisements on the university campus. Only healthy, right-handed women without a history of previous head injury, psychiatric and neurological disorders were included in the study. All were German native speakers, had normal or corrected-to-normal vision and were free of medication.

## EEG study

Sentence preparation A total set of $32 \times 3 \times 3 \times 3=864$ test sentences and $3 \times 32=96$ filler sentences was generated. Each sentence consisted of two clauses conjoined by a coordinating or subordinating conjunction. The target word was always the last word of the sentence and consisted in a medium frequent bisyllabic verb, adjective or participle. The logarithmic frequency of each target word was determined
from Wortschatz Leipzig (http://wortschatz.uni-leipzig.de) as the WL index. A WL index of $n$ means that the most frequent German word "der" is $2^{\text {n }}$ times more frequent than the target word. Following a $3 x 3$ design we introduced the three content categories INT "intentional emotion", PROP "proprioceptive emotion", and PHYS "physical control" and three congruency categories CON "congruent", INCON "incongruent", and UNREL "unrelated" for the target words.
The target words of category INT semantically denoted or lexically entailed an emotional relation between an experiencer and an intentional object. A further grammatical criterion was a verb/adjective valence of at least 2 . Words of category PROP semantically denoted or lexically entailed a proprioceptive feeling of an experiencer. Grammatically, these words had either a verb/adjective valence of less than 2 or were causatives resulting in a proprioceptive feeling. Category PHYS was designed as a control with non-mental target words. The target words for each of the content categories were grouped into triplets with one word for each of the three congruency categories (32 triplets for INT, PROP and PHYS each). For each triplet three different, but contentwise similar sentential contexts of the above mentioned two-clause structure were created such that the sentences completed by the word of condition CON would describe a semantically congruent and plausible scenario. The word of condition INCON was closely semantically related to that of condition CON (being typically an antonym or contrastive word), but would make each of the three sentential contexts semantically incongruent and implausible. The word of condition UNREL was not or only distantly semantically related to that of condition CON and would make each of the three sentential contexts grossly semantically incongruent and implausible. By combination altogether 9 sentences were created from each triplet and the three contexts. This allowed us to present the sentences to the subjects in random selection and order such that each subject saw all three target words of each triplet and all three of the corresponding sentential contexts without any repetition of either the target words or the contexts. Priming effects were thus avoided.
There were no significant differences in logarithmic frequency of the target words across the 9 conditions of the


Figure 1. Cluster-based permutation test for intentional emotion words. The test compares the congruency conditions UNREL to CON in the content condition INT for all channels and each 2 ms segment ( $\alpha=0.025$ ). The resulting significant clusters are marked red a) Negative cluster corresponding to an N400 effect with onset at 396 ms and offset at 498ms ( $\mathrm{p}=0.0096$ ). b) Positive cluster corresponding to a P600 effect with onset at 610 ms and offset at $694 \mathrm{~ms}(\mathrm{p}=0.0048)$.
$3 x 3$ design ( $\mathrm{M}=13.49, \mathrm{SD}=2.26$ ). The word classes were balanced between verbs, adjectives, and participles. The reason for the filler sentences was to balance semantically congruent (CON and Fillers) and incongruent (INCON and UNREL) scenarios. Each subject saw 288 test sentences and 96 filler sentences, i.e. altogether 384 sentences. See Tab. 1. Sentence Task Subjects viewed whole sentences, presented in Presentation software (Neurobehavioral Systems Inc., Albany, CA, USA) on the screen in front of them, in small chunks of words at a time for 500 ms each. The sentence started after the presentation of a centered fixation cross that stayed on the screen for 1000 ms . The words were presented in black letters on a grey background in the center of the screen. There was an inter-stimulus interval between the chunks of 50 ms , which showed a blank screen only. The last word was always the target word and determined the onset of the N400 measure epoch. After 33\% of the sentences, a question mark appeared 2000 ms after the offset of the target word which required the subject to press a button ("yes" or "no") for whether they considered the sentence to be sensible or not. The filler sentences mentioned above were necessary to enable an approximately equal number of button presses. The question mark was followed by a 2000ms blank screen until the next sentence started. The main purpose of the question served to keep participants engaged and alert during passive viewing.

## Background Measures

Multifaceted empathy test (MET) As a measure of empathy, the MET depicts 40 different photographs of various people in emotionally charged situations, with a varying degree of expression on their faces (Dziobek, 2008). In the computer task, each picture is presented three times with three different questions. Cognitive empathy is assessed by the question "How does the person feel?". By
pressing a number from one to four, the subject has to choose one of four possible emotional states, only one of which is defined as correct. The maximum score is 40 points. Two kinds of affective empathy are measured. Explicit empathy is assessed by the question "How much do you compassionate with this person?" and implicit empathy is compiled by the question "How strongly aroused are you by the picture". In both conditions, subjects are asked to rate their emotional engaging on a nine-point scale ranging from 1 (not at all) to 9 (very much). The implicit empathy measure reduces the subject's tendency to answer socially desirable. The maximum scores for both affective empathy measures are 9 points each.

## Procedure

Participants were seated comfortably in a chair, at a distance of 75 cm from the screen in a sound-proof and electrically shielded room with ambient lighting. Upon arrival, they signed informed consent, completed the Edinburgh handedness test and the eating disorders subtests of the DIPS (Diagnostisches Interview bei psychischen Störungen, reported elsewhere). Electrodes were applied to the scalp. After receiving task instructions, a short practice task of five training trials was introduced to ascertain comprehension of the task requirements. A total of 384 sentences, presented in a randomized order and split up into three equal presentation blocks, were presented on the screen. A keyboard with two response buttons was positioned in front of the subjects on the table. In between the blocks, participants had resting periods of at least 5 minutes in order to recover from fatigue and concentration loss. The total duration of the experimental task was 1 hour. By the end of the recording procedure, subjects completed the multifaceted empathy test (MET) offline.

## Electroencephalography Recording and Data Analysis

 The analysis is based on 30 active electrode channels that recorded an electroencephalogram (EEG) from the subjects' scalp surface with a BrainAmp acticap EEG recording system (BrainAmps amplifier, München) according to the international 10-20 system. Four additional electrodes measured participants' electrooculogramm (EOG) for both vertical and horizontal eye movements for later removal of eye movement artifacts. The reference electrode was placed in the position of the FCz and AFz served as ground

Figure 2. Grand average ERP waveforms for the electrode Cz. The two congruency conditions UNREL and CON are compared. The dotted lines mark the onsets and offsets of the N400 and, respectively, the P600 effects according to the cluster based permutation test. a) Intentional emotion words (INT). Cluster significance: $p(N 400)=0.0096$, $p(P 600)=0.0048$. b) Proprioceptive emotion words (PROP). Cluster significance: $p(N 400)=0.0004, p(P 600)=0.0325$. c) Physical controls (PHYS). Cluster significance: p(N400)= $0.0016, \mathrm{p}(\mathrm{P} 600)=0.0072$.
electrode. The sampling rate was 500 Hz and impedance was lowered to below $5 \mathrm{k} \Omega$. The EEG data were analyzed using BrainVision Analyzer 2.0 (BrainVision, München). After recording with a $0.5305-70 \mathrm{~Hz}$ online filter, the data were filtered off-line through a $0.5305-30 \mathrm{~Hz}$ bandpass zero phase Butterworth filter. Afterwards eye-blink artifacts were removed by an independent component analysis (ICA) which was performed for each subject. The reference electrode was offline re-referenced retrospectively to the linked mastoids comprising TP9 and TP10. An automatic artifact rejection removed all trials with amplitudes above $90 \mu \mathrm{~V}$ and below $-90 \mu \mathrm{~V}$. Segments from 200 ms pre-target onset until 1000 ms post-onset were separately extracted and averaged for every subject and for each of the $3 x 3$ conditions $\{I N T, ~ P R O P, ~ P H Y S\} \times\{C O N, ~ I N C O N$, UNREL\}.

## Results

## Onset and offset of N400 and P600

In order to determine the onset and offset of N400 and P600 effects, we compared the ERPs in the congruent condition against the ERPs in the unrelated condition for all three content conditions. The onset and the offset of the N400 and P600 were determined by a resampling procedure, the cluster-based permutation test: The averaged ERPs for each subject in the CON and UNREL condition were collected in a single set, which was then randomly partitioned into two equally sized subsets. The data-points (time x channel) were compared between the partitioned sets by a dependent $t$-test. The significantly different $-\alpha=0.025$ (The significance level of 0.05 was Bonferroni corrected by a factor 2 since both negative and positive clusters were of interest) - data points were then clustered according to temporal-spatial adjacency. The cluster-level statistics was calculated by taking the sum over the t-values for each cluster. This procedure was repeated 10,000 times. The p-values of the observed clusterlevel statistics were estimated as the proportion of partitions that resulted in a higher cluster-level statistics than the observed one (Maris \& Oostenveld, 2007).
The resampling procedure revealed the positive and negative clusters, i.e. collections of time-channel points where the measured amplitude in the UNREL condition was significantly higher (resp. lower) than in the CON condition. The onset of the cluster was taken to be the first time point contained in the cluster, whereas the offset was taken as the last time point in the cluster. The results of the cluster-based permutation test for condition INT is shown in Fig.1. Fig. 2 displays the grand averages on electrode Cz for all three content conditions with onsets, offsets and p-values for the N400 and P600 clusters determined by the clusterbased permutation test. Fig. 3 and Fig. 4 show the topographical distribution of the N400 and P600 effects.

## MET scores and N400

We were interested whether subjects' MET scores correlate with the effect-sizes of the N400 effects. To gain insight on


Figure 3. Topography of the $\mathbf{N 4 0 0}$ effects. The topographical mapping shows the difference between the congruency conditions UNREL and CON for the three content conditions averaged between onset and offset of the N400 effect according to the cluster-based permutation test. In all three content conditions the N400 effect has a centro-parietal extension.


Figure 4. Topography of the P600 effects. The topographical mapping shows the difference between the congruency conditions UNREL and CON for the three content conditions averaged between onset and offset of the P600 effect according to the cluster-based permutation test. a) For intentional emotion words (INT) the P600 effect has a medialfrontal focus. b) For proprioceptive emotion words (PROP) the P600 effect has a centro-parietal focus. c) In the physical control condition, the P600 effect has a centro-parietal focus with a right hemispheric dominance.


Figure 5. Comparing the N400 effect for intentional emotion words on the median split between HIGH and LOW cognitive empathizers. a) The difference between the congruency conditions UNREL and CON for intentional emotion words for the HIGH and LOW cognitive empathy group averaged over electrodes in the significant cluster. b) Topographical mapping of the differences between the HIGH and the LOW group (HIGH - LOW) for the N400 effect on the time interval determined by the cluster-based permutation test. The effect difference between HIGH and LOW cognitive empathizers peaks in the fronto-central region.
this matter, we did a median split of the tested group of subjects with respect to the three MET-scores: The subjects were split into HIGH ( $\mathrm{N}=12$ ) and LOW ( $\mathrm{N}=13$ ). We conducted a cluster-based permutation test to find out the significant differences between the HIGH cognitive empathy and the LOW cognitive empathy group in the time window of the main N400 effect for condition INT (396498ms).
The cluster-based permutation test confirmed that the HIGH cognitive empathy group has a significantly ( $\alpha=0.05$, $\mathrm{p}=0.0286$ ) stronger N 400 -effect than the LOW cognitive empathy group in the time interval $462-498 \mathrm{~ms}$ for condition INT, see Fig. 5 Due to space limitation the significant findings regarding proprioceptive emotion words are not discussed here.

## Discussion

As argued above, the embodied-simulative view of meaning predicts that subjects who have a higher capacity to empathize with the emotions of other persons should be more sensitive also to semantic violations that occur when an emotion word is embedded in an incongruent or even unrelated sentence context. This stronger sensitivity should correlate with a stronger N400 effect, which is a widely acknowledged measure for the violation of semantic expectations. As our study revealed, subjects with a high MET score for cognitive empathy with emotions, indeed, show a significantly stronger N400 effect when an intentional emotion word is embedded in a semantically unrelated sentence context than those with a low score. This difference is strongest in fronto-central regions of the brain.

Aside from these results interesting localization differences in the P600 effects were found between intentional and proprioceptive emotion words and the physical control condition while only marginal localization differences occurred for the N400 effects. Due to space limits a discussion of those will be deferred.

## References

Avenanti A., Bueti D., Galati G., Aglioti S. M. (2005). Transcranial magnetic stimulation highlights the sensorimotor side of empathy for pain. Nature Neuroscience;8(7):955-60.
Barsalou, L. W. Cohen, H. \& Lefebvre, C. (2005, Eds.) Situated conceptualization Handbook of categorization in cognitive science, Elsevier (pp. 619-50).
Bastiaansen J. A., Thioux M., Keysers C. (2009). Evidence for mirror systems in emotions. Philosophical transactions of the Royal Society of London. Series B, Biological sciences;364(1528):2391-404.
Blakemore S. J., Bristow D., Bird G., Frith C., Ward J. (2005). Somatosensory activations during the observation of touch and a case of vision-touch synaesthesia. Brain;128:1571-83.
Carr L., Iacoboni M., Dubeau M. C., Mazziotta J. C., Lenzi G. L. (2003). Neural mechanisms of empathy in humans: a relay from neural systems for imitation to limbic areas.

Proceedings of the National Academy of Sciences of the United States of America;100(9):5497-502.
Chwilla D. J., Virgillito D., Vissers C. T. (2011). The relationship of language and emotion: N400 support for an embodied view of language comprehension. Journal of Cognitive Neuroscience;23(9):2400-14.
Dziobek, I., Rogers, R., Fleck, S., Bahnemann, M., Heekeren, H.R., Wolf, O.T., Convit, A. (2008). Dissociation of Cognitive and Emotional Empathy in Adults with Asperger Syndrome Using the Multifaceted Empathy Test (MET). J Autism Dev Disord, 38:464-473.
Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. Trends in Cognitive Sciences, , 6, 78-84.
Gallese V. (2003). The roots of empathy: the shared manifold hypothesis and the neural basis of intersubjectivity. Psychopathology;36(4):171-80.
Kutas M., Federmeier K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). Annual Review of Psychology;62:621-47.
Maris E., Oostenveld R. (2007). Nonparametric statistical testing of EEG- and MEG-data. Journal of Neuroscience Methods;164(1):177-90.
Martin A. (2007). The representation of object concepts in the brain. Annual Review of Psychology;58:25-45.
Pinheiro A. P., Del Re E., Mezin J., Nestor P. G., Rauber A., McCarley R. W., Gonçalves O. F., Niznikiewicz M. A. (2013). Sensory-based and higher-order operations contribute to abnormal emotional prosody processing in schizophrenia: an electrophysiological investigation. Psychological Medicine;43(3):603-18.
Pulvermüller F. (2005). Brain mechanisms linking language and action. Nature Reviews Neuroscience;6(7):576-82.
Pulvermüller F., (1999). Words in the Brain's Language Behavioral and Brain Sciences, 1999, 22, 253-279.
Pulvermüller F., Fadiga L. (2010). Active perception: sensorimotor circuits as a cortical basis for language. Nature Reviews Neuroscience;11(5):351-60.
Rizzolatti G., Fadiga L., Gallese V., Fogassi L. (1996). Premotor cortex and the recognition of motor actions. Cognitive Brain Research;3(2):131-41.
Rodríguez-Ferreiro J., Menéndez M., Ribacoba R., Cuetos F. (2009). Action naming is impaired in Parkinson disease patients. Neuropsychologia;47(14):3271-4.
Werning, M. (2012) Non-symbolic Compositional Representation and Its Neuronal Foundation: Towards an Emulative Semantics. In Werning, M., Hinzen, W., \& Machery, M. (Eds.), The Oxford Handbook of Compositionality. Oxford University Press, Oxford (pp. 633-654).
Wicker B., Keysers C., Plailly J., Royet J. P., Gallese V., Rizzolatti G. (2003). Both of us disgusted in my insula: the common neural basis of seeing and feeling disgust. Neuron;40(3):655-64.

# The myth of cognitive decline 

Michael Ramscar, Peter Hendrix \& Harald Baayen<br>Seminar für Sprachwissenschaft, Wilhelmstraße 19<br>Universität Tübingen, Tübingen, Germany


#### Abstract

Across a range of psychometric tests, reaction times slow as adult age increases. These changes have been widely taken to show that cognitive-processing capacities decline across the lifespan. Contrary to this, we suggest that slower responses are not a sign of processing deficits, but instead reflect a growing search problem, which escalates as learning increases the amount of information in memory. A series of computational simulations show how age-related slowing emerges naturally in learning models, as a result of the statistical properties of human experience and the increased information-processing load that a lifetime of learning inevitably brings. Once the cost of processing this extra information is controlled for, findings taken to indicate declines in cognitive capacity support little more than the unsurprising idea that choosing between or recalling items becomes more difficult as their numbers increase. We review the implications of this for scientific and cultural understanding of aging.


Keywords: Learning; Language; Memory; Psychometric Testing;

## The Age of Tithonus

More and more people live longer and longer lives. Outside of 18 countries the UN describes as 'outliers' (Watkins et al, 2005), increased life expectancy and declining birth rates are raising median ages in populations across the globe. By 2030, 72 million Americans will be aged 65 or older, a twofold increase from 2000. The world's population is more aged than ever before in history, and its rate of aging is increasing.

People are clearly living longer; it is less clear that this is a blessing. In Greek mythology, Tithonus was the mortal lover of Eos, goddess of the dawn. While asking Zeus to make Tithonus immortal, Eos forgot to mention "eternal youth," dooming Tithonus to an eternity of decrepit babbling. The psychological and brain-sciences endorse the Tithonean view of aging, portraying adulthood as an extended period of mental decline: memories dim; thoughts slow; problem-solving abilities diminish (Naveh-Benjamin \& Old, 2008; Deary et al, 2009); and each year, the onset of cognitive decrepitude is set ever younger (Salthouse, 2009; Singh-Manoux et al., 2012). One crumb of comfort is that older adults are, on average, happier (Charles \& Carstensen, 2010), although in the circumstances, this might be taken as further evidence of their declining mental prowess.

In what follows, we show how the slowing response speeds that are taken as evidence of "cognitive decline" in adults emerge naturally in learning models (Baayen et al, 2011) as knowledge increases. These models, which are supported by a wealth of psychological (Ramscar et al, 2010) and neuroscientific (Schultz, 2006) evidence,
correctly identify greater variation in the cognitive performance of older adults, successfully predicting that older adults will show more sensitivity to fine-grained differences in test items than younger adults. The models run (and can be rerun) on computers, eliminating the possibility that aging hardware influences their performance, which instead reflects the informationprocessing costs incurred as knowledge increases. Once the demands of processing this extra information are taken into account, it becomes clear that much of the evidence for agerelated declines in cognitive capacity better supports the idea that information processing costs rise as the amount of information in a system increases.

## The problem with "processing speed"

Findings from a range of psychometric tests suggest that the rates at which the mind processes information increase from infancy to young adulthood, and decline steadily thereafter (Salthouse, 2011). Increasing reaction times are a primary marker for age-related cognitive decline (Deary et al, 2010), and are even considered its cause (Salthouse, 1996), yet they are puzzling. Practice improves speed and performance on individual cognitive tasks at all ages (Dew \& Giovanello, 2011). Since we get more practice using our cognitive capacities as we age, why does our performance on tests of them decline?

The answer lies in the way that psychometric tests neglect learning, and its relationship to the statistical patterns characteristic of human life. Learning is a discriminative process that serves to locally reduce the information processing costs associated with various aspects of knowledge and skill (Rescorla \& Wagner, 1972). However, age increases the range of knowledge and skills individuals possess, which increases the overall amount of information processed in their cognitive systems. This extra processing has a cost.

## Learning and the long tail of linguistic experience

Statistically, the distribution of human experience is highly skewed: Much of our day-to-day life is fairly repetitive, involving a small repertoire of common occurrences, such as reading the newspaper and going to work. At the same time, we experience a far more diverse repertoire of infrequent or even unique occurrences (we rarely read the exact same newspaper twice). When data is distributed like this, comparisons of means are often meaningless. Consider the problem of remembering birthdays: We are reminded of the birthdays of family members on an annual basis, and this usually makes us expert at remembering them. However, as we move through
life, we also learn about other birthdays. Sometimes we hear these dates only once, such as when we attend a party for someone we barely know. As each new birthday is learned, our mean exposure to all the birthdays we know will decline, and the task of recalling a particular birthday will become more complex. Recalling six hundred birthdays with $95 \%$ accuracy need not imply a worse memory than recalling six with $99 \%$ accuracy.

Standard psychometric tests do not take account of the statistical skew of human experience, or the way knowledge increases with experience. As a result, when used to compare age groups, they paint a misleading picture of mental development. This can be demonstrated most clearly in relation to language. Language is a central aspect of cognition, its statistics are more readily quantified than other aspects of human experience, and all psychometric tests involve some linguistic information processing.


Figure 1. The frequencies of the 1000 most common words in the Corpus of Contemporary American English plotted by rank.

Importantly, linguistic distributions are skewed at every level of description (Baayen, 2001). Consider the relationship between word types (e.g., dog) and tokens (how often "dog" occurs; Figure 1). In any large sample of English, a few words occur very frequently (the, and), such that half of a typical sample comprises tokens of only 100 or so high-frequency types. The relative frequency of these types decreases rapidly (the most-frequent word may be twice as frequent as the second-most), while frequency differences between types decrease as their relative frequency declines. This means that the other half of a typical sample is composed of ever-fewer tokens of a very large number of types, with ever-smaller frequency differences between them. Typically, around half of these types occur just once.

This is a very long-tailed distribution: $49 \%$ of the 425 million tokens in the Corpus of Contemporary American English (COCA; Davies, 2009) come from the 100 mostfrequent word types; the remaining $51 \%$ of tokens represent over 2.8 million word types. Although individual lowfrequency types are, by definition, rare, their distribution makes the chances of encountering a low-frequency token in any given sentence extremely high (Möbius, 2003). This distribution ensures that any English speaker learns only a
fraction of its total vocabulary, and that vocabularies grow steadily across the lifespan. However, the tests used to measure cognitive decline assume that vocabulary size is age-invariant in adults (Spearman, 1927; Carroll, 1993; Bowles \& Salthouse, 2008), an assumption seemingly confirmed by psychometric vocabulary measures, which suggest that vocabulary growth in adulthood is marginal (such that slight increases are only reliably detected in metaanalyses; Verhaeghen, 2003).

Psychometric vocabulary measures are virtually guaranteed to register these results, because they attempt to extrapolate vocabulary size from sets of test words. These tests, which are "normed" on the knowledge of schoolchildren, are heavily biased towards frequent wordtypes (Raven, 1965; Heim, 1970; Wechsler, 1997). Unfortunately, while extrapolation is feasible for frequent words, for the millions of low-frequency word-types, knowledge of one randomly sampled word does not predict knowledge of another. Since the distribution of types ensures that adult vocabularies overwhelmingly (and increasingly) comprise low-frequency words, it follows that reliably extrapolating their size or growth from a small test sample is mathematically impossible (Baayen, 2001).

## Simulating the effects of vocabulary learning on information processing

Most infants are sensitive to all the fine-grained phonetic discriminations made by the world's languages. As they learn a native vocabulary, this sensitivity to non-native phonetic distinctions diminishes (Werker \& Tees, 1984). Rather than indicating that cognitive decline begins in infancy, this loss in sensitivity can be seen as an inevitable result of learning. In discriminative learning models, the values of initially undifferentiated sets of cues are shaped by experience, which drives the discovery of cue values that best predict a learning environment (Rescorla, 1988). Because this process involves learning to ignore uninformative cues, it can explain why decreasing sensitivity to uninformative phonetic information goes hand in hand with increasing knowledge about informative phonetic distinctions (Ramscar et al, 2010).

The learning component of the model we use to simulate the effects of experience on reading works in precisely this way. It is an extension of the Naive Discriminative Reader (NDR; Baayen et al, 2011), a two-layer network in which letter unigrams and bigrams serve as input cues, and lexical items serve as outcomes. The values of these cues are initially undifferentiated, and are set competitively as the model learns to predict words from the letters it 'reads.' In the NDR, every predictive cue is linked to each lexical outcome to form a set of subnets. The cue-weights in these subnets are set by the equilibrium equations of the RescorlaWagner learning rule (Danks, 2003), and are completely determined by the distributional properties of the model's training corpus. Simulated latencies derived from these weights accurately capture a wide variety of empirical effects in reading (Baayen et al, 2011).

To model the influence of experience on different populations, the NDR was modified to make it sensitive to the physical and informational consequences of knowledge growth. Given that the amount of activation a given cue receives from the perceptual system remains constant over time (Attwell \& Laughlin, 2001), this modification keeps the total amount of activation spreading from cues to outcomes equal to the amount of activation arriving at them. Analogous to the principle of conservation of electric charge, this means that as vocabulary increases, and each cue becomes connected to an increasing number of outcomes, the amount of activation arriving at any given outcome decreases. Given a vocabulary of size $V$, the network support for any item $i$ is proportional to $a_{i} / V$ where $a_{i}$ is the activation an item receives from the cues in the input.

This modification also accounts for the effects that an increased number of outputs have on information processing in neural systems (Hentchel \& Barlow, 1991; supplementary materials). Shannon's source coding theorem shows that the smallest coding scheme for $V$ words requires, on average, $H(V)$ bits. Since $V$ determines the length of a message in a given code, the effective channel capacity $C$ of an ensemble of neurons decreases as code complexity increases and the amount of redundancy in signals across the network decreases (Hentchel \& Barlow, 1991). We denote these information costs by $f(V)$, where $f$ is an unknown nondecreasing function expressing the coding and signaling costs in a vocabulary of size $V$.

The response latency (RT) associated with reading (operationalized as reaction times to speeded judgments on written words) is modeled as a weighted sum of these components:

$$
R T_{i}=w_{1} V / a_{i}+w_{2} f(V)+c
$$

with $c$ a constant denoting the time required for response execution.

To simulate the effects of vocabulary-growth on adult reading, two NDR networks were trained on data drawn from the Google Trigrams Corpus (a large, naturalistic data set). The first network 'read' 500,000 word-trigram tokens, simulating reading to age 21, the typical age for "young adult" participants in studies; the second 'read' $3,000,000$ word-trigram tokens, ${ }^{3}$ simulating reading to age 70 (the typical age for "old adults"). Consistent with our analysis of the way linguistic distributions influence vocabulary growth, the old model acquired a much larger vocabulary: 32,536 word types, compared to the young model's 21,307 (Figure 2). These growth estimates are very conservative: the Trigram Corpus excludes trigrams with less than 40 occurrences, thereby omitting around $50 \%$ of the word types in the complete Google Corpus. Even with this constrained input, vocabulary expansion was far from asymptote, even after 5 million trigram tokens.

To examine the models' ability to simulate age-related reading differences, we compared their projected reading times for 2,904 English words to empirical latencies from older (mean age 73.6) and younger (21.1) readers for the
same items (Balota et al, 1999). The empirical data exhibit the expected effect of age: mean reaction times are 163 ms shorter for younger than older adults. Simulated reaction times mirror this difference, with an average difference of 167 ms .


Figure 2. Empirically observed vocabulary growth after sampling from the Google Trigrams Corpus.


Figure 3. Left panel: fit of a generalized additive model to the simulated response latencies from the old and young models. Right panel: fit of a generalized additive model to the empirical response latencies from young (mean age: 21.1) and old (73.6) adults (Balota et al, 1999).

The models also correctly predict an important qualitative difference in the empirical word-frequency effect. It is well established that lexical decision responses are slower for lower- (e.g., "whelp") than higher-frequency words ("where"). This overall effect of frequency is present for both young and old adults (Figure 3; right panel). However, while frequency effects asymptote at higher frequencies in both models, they only level off at the lowest frequencies in the younger model, a pattern that is also observed in the empirical data: older adults are far better attuned to frequency variations in the lower range of the test-set than younger adults.

These results can be explained by considering the way the models learn in more detail. In learning, weights on the links between the cues and outcomes are adjusted in two ways: They are strengthened whenever a cue and outcome co-occur; For example, the link between the bigram WH and the lexical target WHERE is strengthened when "where" is encountered in reading, and the link between WH and WHELP is strengthened when the "whelp" is encountered. Conversely, links are weakened when cues occur but outcomes do not.

| 1 | BLASH | 11 | CROME | 21 | TWERP | 31 | WHELP | 41 | BLEAT |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| 2 | SCHNOOK | 12 | GIBE | 22 | THWACK | 32 | SHUCK | 42 | CHIVE |
| 3 | LETCH | 13 | LISLE | 23 | DAUNT | 33 | MOOCH | 43 | WHIR |
| 4 | ZOUNDS | 14 | FLAYS | 24 | RETCH | 34 | JELL | 44 | CROON |
| 5 | JAPE | 15 | SPLOTCH | 25 | GYP | 35 | GROUCH | 45 | TAMP |
| 6 | SOUSE | 16 | VELDT | 26 | YAWL | 36 | AWN | 46 | BOSH |
| 7 | WHIG | 17 | SLOE | 27 | FLUB | 37 | MANSE | 47 | RILE |
| 8 | FILCH | 18 | CONK | 28 | STANCH | 38 | WRACK | 48 | BLANCH |
| 9 | RHEUM | 19 | FRAPPE | 29 | PAUNCH | 39 | HOOCH | 49 | LILT |
| 10 | PARCH | 20 | SKULK | 30 | JOWL | 40 | FLECK | 50 | JEER |

Table 1. The 50 lowest frequency items in the set used to test the models and the older and young adults; BLASH has the lowest frequency of these items, and JEER the highest. As can be seen, many of the letter bigrams in this set of words are comparatively rare in English.

Thus when "where" is encountered, WH occurs without WHELP, weakening the link between WH and WHELP. The distribution of high-frequency words results in their being encountered frequently, at a fairly constant rate over time. This will consistently reinforce the link between WH and WHERE, and consistently weaken the link between WH

In contrast, low-frequency words occur sporadically, so the link between WH and WHELP is reinforced far less (and the link between WH and WHERE weakened less). These imbalances result in "selection pressures" on word types that are reflected in the distribution of orthographic (and phonetic) cues across lexical items (see Zipf, 1949): high-frequency test items are both shorter $(\mathrm{t}(2901)=-10.58$, $\mathrm{p}<0.001$ ) and have higher mean bigram frequencies $(\mathrm{t}(2901)=8.98, \mathrm{p}<0.001)$ than low-frequency items. This means that, on average, low-frequency words contain both more cues, and more rare cues (Table 1). Although rare cues have relatively high values in small vocabularies, they are vulnerable to competition as vocabularies grow: newer vocabulary items also have low frequencies, and are more likely to contain the same rare cues.

All the predicted empirical effects were replicated in an analysis of a second, independent dataset (Figure 4).


Figure 4. Average percentile RT differences (old - young) for the naming latencies of 2,820 Single Syllable words (Yap et al, 2011) by young (mean age: 22.6) and old adults (73.6), plotted against the words' log frequency in the Google 1 -gram corpus, and a generalized additive model fit to the RT differences. As can be seen, the difference between older and younger readers' RTs increases as word frequency decreases.

## Modeling 'decline' in a non-lexical task

To examine whether the relationship between information load and response time also holds for "non-lexical" tests, we considered the letter classification task (Posner \& Mitchell, 1967), a standard non-lexical psychometric test in which two letters are presented in upper or lowercase (A, a, D, d, $\mathrm{E}, \mathrm{e}, \mathrm{R}, \mathrm{r}, \mathrm{H}, \mathrm{h}$ ) and participants judge whether they are alphabetically the same or different. Older subjects are typically slower than younger subjects in this task, a finding that is straightforwardly replicated in the NDR models once the coupling between letters and abbreviated meanings is accounted for (e.g, $H$ for entropy, $R$ for a statistical programming environment, $r$ for correlation, etc.). The network complexity function $f(V)$ in (1), which models response latencies as a function of the activation of the meanings of both letters in a letter pair, predicts longer latencies for older subjects as compared to younger subjects. In short, because the older model has a larger system of outcomes, it has more information to process, making "accessing" a letter harder, and reaction times concomitantly slower (see also Ramscar et al, 2010).

Psychometrically, letter classification is often described as an "information-processing" measure, and older adults' longer response times are taken as evidence of declining information-processing capacity. Yet information theorywhich defines the workings of information-processing system-is, at heart, a set of methods for formalizing the uncertainty in distributions (be they bits of code, or vocabulary items; Shannon, 1948). Information is a property of systems, and processing demands are measured in relation to them (MacKay, 2003). In letter classification, the system comprises the task, a participant, and, crucially, what that participant knows. Because psychometric tests neglect this knowledge, they are formally incapable of measuring information-processing in this task.

## Lexical knowledge and paired-associate learning

Wherever vocabulary size increases with experience, this increased knowledge predicts increasing processing costs and slower responses in psychometric tasks. As a consequence, slower latencies reflect learning, not "decline." Interestingly, this interaction between experience, vocabulary-size and response speed can also be seen in comparisons of monolingual and bilingual picture-
naming (Gollan et al, 2008): the response latencies of young-bilinguals more closely resemble older-monolinguals than younger-monolinguals or older-bilinguals. Notably, slower response times and increased tip-of-the-tongue rates are not taken as evidence of cognitive decline when observed in young-bilinguals (Gollan \& Acenas, 2004), but are instead seen to reflect the demands of processing the larger vocabularies that bilinguals inevitably learn.

The finding that bilinguals experience increased tip-of-the-tongue rates raises a question: could the same systemic effects of learning that account for increased lexical processing latencies explain age-related change in memory measures, such as Paired-Associate Learning (PAL; a psychometric measure of people's ability to learn and recall new information)? In PAL, e.g., the subtest of Wechsler's Memory Scale (WMS; desRosiers \& Ivison, 1988) participants have to learn more or less arbitrary pairings between word cues (e.g., baby; jury) and responses (cries; eagle). Although item-level performance is highly variable (Figure 5), older adults' overall PAL performance is slower and less accurate, and it has been suggested that aging causes encoding (Gilbert, 1941; MacKay \& Burke, 1990) and retrieval processing deficits (Burke \& Light, 1981).


Figure 5. Mean performance by item for 100 older (age 60-69) and 100 younger ( $20-29$ ) adults on forms 1 and 2 of the WMSPAL subtest (desRosiers \& Ivison, 1988). As in the lexical decision and naming data, the relationship between old and young PAL performance is nonlinear: again, older adults exhibit a more marked ability to discriminate between 'harder' (unrelated) and 'easier' (related) items than younger adults.

There is, however, no reason to think PAL performance should be age-invariant. Long-established principles of associative learning predict that well-known items should be harder to learn as Cues $\left(w_{I}\right)$ than newer items (Rescorla \& Wagner, 1972). Likewise, newer Response ( $w_{2}$ ) items should support better learning than well-known, predictable items (Kamin, 1969): $w_{1}-w_{2}$ pairs ought to become harder to learn when $w_{1}$ and $w_{2}$ occur independently at high rates (Rescorla, 1968; compare jury-eagle to baby-cries).

To examine whether age-related PAL differences simply reflect learning, we analyzed the relationship between the age-related variance in the performance of a large sample adults on the WMS-PAL subtest (desRosiers \& Ivison, 1988), and the factors that determine $w_{1}-w_{2}$ learnability. In a
regression analysis of item score differences (mean young mean old), $\underline{w}_{\underline{l}}$ predictability ( $\log$ frequency; $\mathrm{t}=-4.063$, $\mathrm{p}<0.001$ ), the relationship between $w_{2}$ and $w_{l}$ predictability ( $\log \left(w_{2}\right.$ frequency $) / \log \left(w_{l}\right.$ frequency $\left.) ; \mathfrak{t}=-2.935, \mathrm{p}<0.01\right)$ and actual $w_{l}-w_{2}$ co-occurrence rates (log Google frequency; $\mathrm{t}=6.773, \mathrm{p}<0.0001$ ) accounted for more than $75 \%$ of the variance in item performance between 20-29 and $60-69$ year-olds ( $\mathrm{F}(3)=16.432, \mathrm{r}=.87, \mathrm{p}<0.0001$ ).

All things being equal, the relative learnabilty of $w_{1}-w_{2}$ pairs can be estimated from $w_{l}-w_{2}$ co-occurrence and background rates. All things are not equal, however: Older adults have more experience, and learnability is a matter of experience. Accordingly, $w_{2}$ words will become more predictable the more they occur independently of $w_{1}$, and $w_{1}$ words will become less informative the more they occur independently of $w_{2}$; in each case, experience will make $w_{l^{-}}$ $w_{2}$ learning harder. A natural, predictable consequence of this is that PAL performance should increasingly reflect the distributional properties of the $w_{l}-w_{2}$ items as experience grows: if co-occurrence rates are low, a lifetime of learning that jury is uninformative about eagle should make learning jury-eagle harder; whereas high co-occurrence rates will reduce background rate effects, making baby-cries easier for older adults to learn relative to jury-eagle.


Figure 6. Mixed-effects slope estimates for the three learnability predictors and mean item performance of old (60-69) and young (20-29) adults in the WMS-PAL subtest (desRosiers \& Ivison, 1988). All predictor effects and interactions in the model are significant (see supplementary materials), and all slopes (except *) are significantly different from $0(t=>2)$. Older adults are more sensitive to background rate information (negative slopes) than young adults and, as the magnitude of the slopes shows, the overall performance of older adults reflects a far more systematic understanding of the English language.

A mixed-effects analysis of $w_{1}-w_{2}$ item scores by age confirmed the accuracy of this prediction (Figure 6). For each predictor, the magnitude of the slope for the older age group is greater than that for the younger age group, indicating that older subjects bring more lexical experience to the task. Consistent with our earlier findings, older adults' PAL performance reflects their greater knowledge of (and sensitivity to) the distributional properties of $w_{1}-w_{2}$ words, whereas younger adults' less varied performance reflects their more limited knowledge of them. As we noted above, the statistical properties of human experience makes comparing means invidious: in this case, it seems that high
mean PAL performance is a measure of ignorance, not "intelligence."

## Learning and Cognitive Maturation

These results suggest that older and younger adults' performance in psychometric testing largely reflects the same cognitive mechanisms, confronted with the task of processing different quantities of information. The performance of older adults on these tests is evidence of increased knowledge, not declining processing capacity.

When discussing these conclusions with colleagues, a question often arises: "Learning seems to predict linear patterns of change, but cognitive decline seems to kick in around age 60 or 70: how do you explain this?" To explain why, we first note that as people age, they encode less contextual information in memory (Naveh-Benjamin \& Old, 2008). Although this has been taken as evidence that the processes that "bind" contextual information in memory decline with age, learning theory predicts that experience will increasingly make people insensitive to context, because ignoring less informative cues is integral to learning.

Learning is also sensitive to the environment, and its predictions change with it: If a common environmental change-e.g., retirement-was to systematically reduce the variety of contexts people typically encounter in their lives, learning theory predicts that the amount of contextual information they learn will also drop, as the background rates of cues in remaining contexts rise. If these same people were to increasingly spend their time in environments where cues already have very high background rates (e.g., family homes), this effect will be exacerbated. In other words, because learning inevitably reduces sensitivity to everyday context, retirement is likely to make memories harder to individuate and more confusable, absent any change in cognitive processing, simply because it is likely to decrease contextual variety at exactly the time when, as a result of learning and experience, the organization of older adult's memories needs it most.

Learning can explain both the apparent changes in older adults "cognitive performance" around retirement-age, and the fact that these changes are not detected in testing. Similarly, the neglect of learning in the study of cognitive aging makes it highly likely that, like Tithonus, many of our beliefs about cognitive decline are myths. This does not mean that the diseases that can undermine cognition in old age are also mythical: However our understanding of these diseases can only be increased by a better understanding lifelong learning, and its sensitivity to the environment.

## References

Attwell D, \& S.B. Laughlin. (2001) An energy budget for signaling in the grey matter of the brain. J Cereb Blood Flow Metab 21: 1133-1145.
Baayen, R.H. et al. (2011) An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. Psychol. Rev. 118, 438-482.
Baayen. R.H. Word Frequency Distributions. (Kluwer Academic Publishers, 2001).

Balota, D.A. \& D.H. Spieler. (1998) The utility of item level analyses in model evaluation. Psychol Sci 9, 238-240.
Burke, D. M., \& Light, L. L. (1981) Memory and aging: The role of retrieval processes. Psychol. Bulletin, 90, 513-546.
Carroll, J.B. Human cognitive abilities: A survey of factor-analytic studies. (Cambridge University Press, 1993).
Danks, D. Equilibria of the Rescorla-Wagner model. (2003) J. Math. Psychol. 47(2), 109-121.
Davies, M. (2009) The 385+ Million Word Corpus of Contemporary American English (1990-present). Int. J. Corpus Ling. 14: 159-90.
Deary, I.J. et al. (2009) Age-associated cognitive decline. Brit. Med. Bull. 92, 135-152.
desRosiers G, Ivison D. (1986) Paired-associate learning: normative data for differences between high and low associate word pairs. J Clin Exp Neuropsychol; 8: 637-42.
Dew, I.T.Z. \& K.S. Giovanello. (2010) The status of rapid response learning in aging. Psychol. Aging 25(4), 898-910.
Gilbert, J.G. (1941) Memory loss in senescence. J. Abnormal \& Social Psychology, 36, 73-86.
Gollan, T.H. et al. (2008) Aging, bilingualism, and the weaker links hypothesis. J. Mem. Lang. 58, 787-814.
Hale, S., S.D. Lima \& J. Myerson. (1991) General Cognitive Slowing in the Nonlexical Domain: An Experimental Validation. Psychol. Aging 6(4), 512-521.
Hayes, D. P. \& M. Ahrens. (1988) Vocabulary Simplification for Children: A Special Case of 'Motherese'? J Child Lang 15, 395-410.
Heim, A.W. (1970) AH 4 group test of general intelligence. NFER-Nelson.
Hentchel, H.G.E. \& H.B. Barlow. (1991) Fininding minimum entropy codes with Hopfield Networks, Network, 2, 135-148.
Kamin L.J. (1969) Predictability, surprise, attention, and conditioning. In: Campbell B, Church R (eds). Punishment and Aversive Behaviour. Appleton-Century-Crofts: New York.
MacKay, D. G., \& Burke, D. M. (1990) Cognition and aging: A theory of new learning and the use of old connections. In T. Hess (Ed.), Aging and cognition: Knowledge organization and utilization (pp. 213-263). Amsterdam: North-Holland.
Naveh-Benjamin, M. \& S.R. Old. (2008) Aging and memory (humans). In J.H. Byrne, H. Eichenbaum, R. Menzel, H.L. Roediger, \& D. Sweatt (Eds.). Learning and memory: A comprehensive reference (2008).
Posner, M.J. \& R.F. Mitchell. (1967) Chronometric analsis of classification. Psychol. Rev. 74, 392-409.
Ramscar, M. et al. (2010) The Effects of Feature-Label-Order and their implications for symbolic learning. Cognitive Sci. 34(6), 909-957.
Raven, J.C. Guide to using the Mill Hill vocabulary test with progressive matrices. (HK Lewis, 1965).
Rescorla, R.A. \& A.R. Wagner. A theory of Pavlovian conditioning. In: Black \& Prokasy (Eds.), Classical Conditioning II: Current Research and Theory. (Appleton-Century-Crofts, 1972)
Salthouse, T.A. (1996) The processing-speed theory of adult age differences in cognition. Psychol. Rev. 103, 403-428.
Schultz, W. (2006). Behavioral theories and the neurophysiology of reward. Ann. Rev. Psychol. 57, 87-115.
Shannon, C.E. (1948) A mathematical theory of communication. Bell Systems Technical Journal 27(3), 379-423.
Singh-Manoux, A. et al. Timing of onset of cognitive decline: Results from Whitehall II prospective cohort study. BMJ 344, d7622 (2012).
Verhaeghen, P. Aging and vocabulary score: A meta-analysis. Psychol. Aging 18, 332-339 (2003).
Watkins, K. et al. UN Human Development Report: International Cooperation at a Crossroads-Aid, Trade and Security in an Unequal World. (United Nations Development Programme, 2005).
Wechsler, D. Wechsler Adult Intelligence Scale-Third Edition (WAIS-III). (Harcourt Assessment, 1997).
Yap, M.J., D.A. Balota, D.E. Sibley \& R. Ratcliff. (2011) Individual Differences in Visual Word Recognition: Insights From the English Lexicon Project. J Exp Psychol Hum Percept Perform 38(1), 53-79.
Zipf., G. K. Human behavior and the principle of least-effort. (AddisonWesley, 1949).

For a complete list of references and supplementary materials, please consult the article copy hosted at www.sfs.uni-tuebingen.de/

# The Evolution of Rhythmic Cognition: New Perspectives and Technologies in Comparative Research 

Andrea Ravignani** ${ }^{*}$ (andrea.ravignani@univie.ac.at), Bruno Gingras*, Rie Asano ${ }^{\dagger}$, Ruth Sonnweber*, Vicente Matellán ${ }^{\curvearrowright}$ \& W. Tecumseh Fitch*<br>*Department of Cognitive Biology, University of Vienna, Austria<br>\#Language Evolution \& Computation Research Unit, University of Edinburgh, UK<br>${ }^{\dagger}$ Department of Musicology, University of Cologne, Germany<br>${ }^{\diamond}$ Robotics Group, University of León, Spain


#### Abstract

Music is a pervasive phenomenon in human culture, and musical rhythm is virtually present in all musical traditions. Research on the evolution and cognitive underpinnings of rhythm can benefit from a number of approaches. We outline key concepts and definitions, allowing fine-grained analysis of rhythmic cognition in experimental studies. We advocate comparative animal research as a useful approach to answer questions about human music cognition and review experimental evidence from different species. Finally, we suggest future directions for research on the cognitive basis of rhythm. Apart from research in semi-natural setups, possibly allowed by "drum set for chimpanzees" prototypes presented here for the first time, mathematical modeling and systematic use of circular statistics may allow promising advances.


Keywords: The evolution of music; primate cognition; animal-machine interaction; chimpanzee drum set; vocal learning; rhythm; entrainment; beat; synchronization; social cognition; comparative cognition.

## Introduction

## Evolution of Music and Origins of Rhythm

Music as a cognitive system is one of the most prominent and distinctive human features. Since Darwin, the putative role of selection in the emergence of human music has been a topic of great debate. Numerous hypotheses, which attribute an adaptive value to music, have been proposed, all featuring some social component. While hypotheses on music origins are difficult to test directly, the comparative method in cognitive biology enables us to investigate the purported human uniqueness of particular musical abilities (Fitch, 2006). In this paper we focus on one aspect of music cognition, namely rhythm, and propose new perspectives and technologies for investigating its evolution.

## Rhythm and Cognition

Rhythm, characterized as a structured pattern of temporal change, plays a central role in music. Beats, defined as points in time occurring in a perceptually periodic way (Patel, 2008), are a basic element of musical rhythm. Grouping and meter are subsystems of musical rhythmic organization and are considered the basic structural components of rhythmic patterns (Lerdahl \& Jackendoff, 1983). Grouping refers to the organization of the musical stream into motives, phrases, and sections. Meter corresponds to a regular pattern of strong and weak beats. In metrical structures, beats are organized hierarchically according to their relative strength. Moreover, the impression of the speed of the performed pattern, the tempo,
influences the interpretation and perception of rhythmic structures. According to the tempo, humans may assign different organizations to grouping and metrical hierarchy. Hence, the cognition of musical rhythm should not be investigated solely holistically, but also in terms of beat, grouping, meter, and tempo. These, together, yield the flexibility of human rhythmic cognition: humans are able to extract structural properties from music and interpret them in multiple contexts. What are the basic capacities allowing this cognitive flexibility?

The metrical hierarchy mentioned above contains a particular hierarchical level called tactus, which listeners perceive as 'the (primary) beat' (Lerdahl \& Jackendoff, 1983), whose perception is robust to moderate tempo fluctuations (Patel, 2008). It seems that our internal processes underlying rhythm perception can be spontaneously synchronized, entrained, to external regular, periodic sensory cues (Grahn, 2012). In this entrainment model, the relative timing of events is processed by expecting their periods or phase and adjusting the expectations to actual occurrences (Grahn, 2012). This flexible beat processing mechanism is also the basis for synchronizing motor actions to musical stimuli, requiring (i) beat extraction, (ii) synchronization of an internal motor pulse to the inferred auditory beat (beat entrainment), and (iii) a motor pattern generation on the basis of the internal pulse (Fitch, 2012). A fundamental requirement of synchronization is hence the capacity to extract the beat, already present in newborns and infants, though not conclusively innate because of possible prenatal learning (Grahn, 2012). The capacity for beat perception and synchronization could be shared with other animals as an analogous or homologous evolutionary trait. In order to understand the nature and evolution of human cognitive capabilities for rhythm, different species must be tested on tasks requiring the three aforementioned skills.

## Rhythm and Beat Evidence in Non-human Animals

## Vocal Learning and Dissociation Hypotheses

Some non-human animal species have a particularly good control over their vocal tract. Among these, humans, elephants, many bird species and some marine mammals are capable of spontaneously imitating sounds which may or may not belong to their natural communication system. A promising hypothesis has been put forward connecting vocal learning and beat-based rhythmic abilities across species (Patel,


Figure 1: Phylogenetic tree of species showing: vocal learnering skills (underlined), ability to synchronize to a beat (bold) and spontaneous drumming behavior (italics). Notice how, while showing no evidence of vocal mimicry, California sea lions are capable of synchronization, and some apes exhibit natural percussive behavior.
2008): as both sorts of tasks are better performed with a tight connection between motor and auditory brain areas, which is found in some vocal learning species including humans, the skill of vocal mimicry would be a necessary prerequisite for beat perception and synchronization. Considerable experimental evidence supports this hypothesis.

As humans seem to be the only advanced vocal learners among primates, a key question is whether the ability to perceive, produce and entrain to musical rhythm is unique to humans among primates. Recent evidence (Honing et al., 2012) suggests that rhesus macaques (Macaca mulatta) can detect rhythmic grouping but not the downbeat in music. The authors formulate an "auditory timing dissociation hypothesis": Some cognitive skills allowing grouping are expected to exist in several primates due to common ancestry, while some others related to beat induction should be present in humans and other vocal learners due to convergent evolution. In fact, a generalized failure to produce beat-based rhythmic patterns in non-human primates would support the hypothesis of convergent evolution of vocal learning and beat perception and synchronization abilities.

Evidence from Vocal Learners Schachner et al. (2009) searched videos of putative animal entrainment to music using the global database YouTube. 1019 videos of non-human animals, half of which involved vocal mimicking species, were analyzed both for frequency and phase synchronization. Strikingly, all 33 videos showing convincing evidence of entrainment featured vocal learning species. Among species considered unable to learn new vocalizations, there was no evidence of synchronization ability. Within vocal mimics, all animals examined belonged to bird species, except for one Asian elephant (Elephas maximus). Schachner et al. (2009) also analyzed videos of sea lions (subfamily: Otariinae) which showed no evidence of entrainment (but see below for a recent study reporting evidence of entrainment in a sea lion).

This general result on synchronization abilities in vocal learning species is backed up by experimental evidence in three different avian species. Patel et al. (2009) analyzed the head bob movements of a sulphur-crested cockatoo ( Ca -
catua galerita eleonora) in response to a familiar song under unfamiliar tempo manipulations. In the absence of any training, the animal showed periods of entrainment matching phase and frequency of the musical beat. Schachner et al. (2009) provided additional evidence for entrainment in the same individual and a language-trained, African grey parrot (Psittacus erithacus).

Hasegawa et al. (2011) trained budgerigars (Melopsittacus undulates) to peck according to the beat of an audio-visual metronome. This study is particularly relevant as (i) it extends the sample size of the previous studies to 8 birds; (ii) it makes use of powerful analytical techniques from circular statistics and (iii) it compares actual performances to those of computer-simulated birds (simulating a range of neurophysiological constraints) in order to test the hypothesis that experimental subjects use "behavioral shortcuts" which could give the illusion of beat synchronization. Overall, Hasegawa et al. (2011) provide decisive evidence of trained synchronization ability in a vocal-mimicking species.

A Crucial Outlier In a recent study, Cook et al. (2013), investigated beat synchronization abilities in a pinniped, the California Sea Lion (Zalophus californianus). Crucially, sea lions, unlike seals and some other marine mammals, seem to have a low degree of vocal flexibility (Schusterman, 2008) and are usually grouped with non-vocal learners. Cook et al. (2013) trained the animal to bob its head in synchrony with different auditory stimuli at different tempi. This ability, trained first with metronome-like stimuli at different tempi, was easily transferred to novel tempi. Similarly, once trained with actual songs, the Sea Lion was able to transfer the synchronous head bobbing to new tempi and songs with no additional training.

This exciting finding opens new lines of research (see Figure 1). On the one hand, conclusive evidence on vocal mimicking abilities in sea lions is indispensable to contrast this finding with, and eventually update, the vocal learning hypothesis. On the other hand, Cook et al.'s (2013) discovery increases the likelihood of finding beat and (rhythmic) synchronization abilities in some vocal non-mimics. In particular apes and marine mammals, heterogeneous in vocal learn-
ing and advanced cognitive skills, offer a promising "testing field". Unfortunately, the evidence for apes and monkeys is either observational or not conclusive enough to prove or disprove beat entrainment.

## Contrasting Evidence from Macaques

Interval Timing Abilities Zarco et al. (2009) compared the ability of 20 human subjects and 3 rhesus macaques to synchronize to visual and auditory metronomes and to project this interval timing ability once the metric cue has been removed. They concluded that these monkeys were "not able to synchronize their tapping behavior to the sensory metronome as human subjects do" (Zarco et al., 2009). It is essential to notice that the authors based their conclusion on a linear test of "phase matching" (Patel et al., 2009). Zarco et al. calculated the average time difference between metronome cues and tap onset and compared this between species using a repeated measures ANOVA. As monkeys tapped, on average, 300 ms after the metronome and the ANOVA indicated a significant difference only between species, Zarco et al. (2009) interpreted this as evidence against synchronization. Further analyses suggested that the monkeys have, however, some form of timing prediction abilities, having shorter reaction times to stimuli with constant, rather than unpredictable, inter-onset intervals. Zarco et al. (2009) is a crucial contribution to the field, providing the first experimental paradigm for testing one component of rhythm in non-human primates. However it is unclear whether more specific tests from circular statistics would have led to the same conclusions in terms of phase or tempo synchronization.

Subsecond Beat Prediction Konoike, Mikami \& Miyachi (2012) conducted a similar experiment with two Japanese macaques (Macaca fuscata, closely related to rhesus macaques). The monkeys were reinforced for pushing a button in response to an audiovisual metronome. Crucially for our purposes, a synchronization threshold was set a priori: if a metronome beat was not matched with a tap within 350400 ms , the entire trial would be aborted. Reaction times were shorter with regularly-spaced beats when compared to an "unpredictable" inter-beat interval condition, as long as inter-beat intervals did not exceed one second. However, comparing the synchronization thresholds imposed by the authors to the reaction times, there could be a differential effect of the thresholds in shaping reaction times between subjects. As in the previous case, this study contributes to our understanding of what is unique about human rhythmic abilities. A suggestive hypothesis put forward by Konoike et al. (2012) is that their subjects' rhythmic control could depend on an automatic timing system rather than higher cognitive mechanisms. The a priori synchronization threshold and the lack of a statistical test on tempo matching prevent us from drawing conclusions about music-specific rhythmic abilities in these primates.

Synchronization of Arm Motion Nagasaka et al. (2013) reported mutual synchronization between pairs of Japanese
macaques in a laboratory setup. Interestingly, in each interaction, the ratios of BPM of the two subjects were small integers, suggesting periodical occurrence of synchronized taps. However, it seems that visual, rather than auditory, information had a decisive role in the macaques' synchronization accuracy when moving in response to a video of a conspecific.

## The Social Convergence Hypothesis

Recent findings (Large, Velasco \& Gray, 2008; Nagasaka et al., 2013) point towards the importance of social context in obtaining positive results when testing for rhythmic and music-related abilities. Children can already entrain to a pulse from 2.5 years of age onwards, being particularly accurate when drumming along with a human partner, rather than an artificial one (Kirschner \& Tomasello, 2009), suggesting that rhythmic abilities, coordination and cooperation could be partially connected within hominid evolutionary history. The recent "Social Convergence hypothesis" puts forward the importance of human social instincts in the development of rhythmic abilities: isochrony would be an easy way of achieving synchrony, which in turn is a form of coordinate, cooperative auditory signal generation (Fitch, 2012). Hence, evidence of entrainment in interactive contexts or from social species is required to support or refute this hypothesis.
Human-Bonobo Musical Interactions In the context of human-ape interaction, Large et al. (2008) reported an occurrence of entrainment. MIDI recordings from musical interactions between a human and three bonobos (Pan paniscus) were analyzed for evidence of synchronization. The authors claim that, after having identified " 37 episodes of rhythmic interaction, [...] in just under half of these episodes, statistical evidence of phase entrainment was found" (Large et al., 2008). The interactive nature of this study and the little published information leaves unclear the relative contribution of human and bonobo participants to rhythmic synchronization (Patel et al., 2009). Considering that bonobos are capable of synchronous hooting (de Waal, 1988), this result is, in principle, promising and worth further exploration.
Drumming by Wild Chimpanzees All three African Great Ape species engage in spontaneous drumming (Fitch, 2006). Chimpanzees (Pan troglodytes) can be observed hitting objects in order to produce loud sounds, especially during dominance displays. Arcadi, Robert \& Boesch (1998), analyzed chimpanzees' spontaneous drumming behavior on tree buttresses. Among other measures, Arcadi et al. (1998) reported an inter-beat interval distribution ranging up to 1.4 s , with a mean of 0.3 s and "most inter-beat intervals" less than 0.4 s . Transposing this into musical terms, the drumming behavior had a mean of 200 BPM (beats per minute) and was above 43 BPM, with most recorded patterns exceeding 150 BPM. ${ }^{1}$

[^195]Arcadi et al. (1998) found a number of individual differences in drumming behavior, notably in the inter-beat interval duration, the number of beats per "drumming session" and the length of sessions. Finally, the authors tested for statistical dependence between contiguous, non-adjacent beat patterns. One of the chimpanzees produced series of four beats, where a short interval between two beats statistically predicted another short interval between two following beats. This can be interpreted as showing a weak form of regularity in natural beat production and a sporadic, local steadiness in tempo. Percussive behaviors are hence naturally present in primates not capable of vocal mimicry. Together with Honing et al.'s (2012) findings on rhythmic grouping recognition, this suggests that rhythmic abilities across species might be graded, rather than dichotomous, suggesting that the evolution of musical rhythm be better investigated in a fine-grained manner.
Spontaneous Tapping in a Chimpanzee The languagetrained chimpanzee "Ai" has recently been shown capable of synchronizing her movements to an isochronous beat (Hattori, Tomonaga \& Matsuzawa, 2013). Three chimpanzees were trained to alternatively tap two keys of a keyboard at any preferred rate. The task was subsequently accompanied by steady auditory sequences of notes at three different tempos. One of the chimpanzees, Ai, spontaneously synchronized her tapping rate to one of the tempos. This is a remarkable result, presenting the first experimental evidence of behavioral synchronization in non-human primates. However, the authors point out that Ai’s lack of tempo flexibility and low phase accuracy would call for additional studies in order to clarify possible differences with normal human performance. Finally, Hattori et al. (2013) suggest that this study does not necessarily falsify the vocal learning hypothesis: the keyboard produced sounds, hence it is not conclusive proof of entrainment.

## Future Directions

If research on rhythmic cognition aims to advance and break new ground, there are some directions we propose it should take. First, a broader range of animal species should be tested: apes, marine mammals and non-avian vocal learners are key groups whose success or failure in beat and other rhythmic production tasks will arbitrate between a number of proposed hypotheses. We stress that such testing should happen as much as possible in an experimentally-controlled, though ecologically valid environment. Below we propose a viable approach for chimpanzees, using musical instruments explicitly built with those constraints in mind. Second, statistical techniques used to analyze entrainment data should be adequate to the purpose. If we think about statistics as a tool for getting closer to scientific facts, statistical techniques whose assumptions better fit the object under investigation will lead us closer to solid conclusions. Inference drawn from a statistical test resting on inadequate assumptions will lead to less robust conclusions. Third, mathematical modeling of the emergence of beat and rhythm is an important complement to
experiments. Analytical models and agent-based simulations can help sharpen hypotheses about which cultural, social and biological evolutionary processes endowed different species with different cognitive skills in terms of rhythm and music.

## Chimpanzee Drum-Set Prototypes

Towards Understanding Rhythmic Production in Chimpanzees Above we hinted at a viable methodological approach for testing beat and rhythm production abilities in higher primates. Chimpanzees already exhibit drumming behavior in the wild. A first step towards testing rhythm hypotheses in a semi-natural context could be to provide chimpanzees with a device they can use to produce sounds when manipulated. At the same time, such a "music-making device" should be particularly well adapted to the rigor of scientific experiments. No musical instrument or device, specifically designed for chimpanzees, sensing movements and feed-backing sounds, is currently available for purchase. Such a device should: (i) be resistant to chimpanzees' great strength, (ii) enable them to produce sound through object manipulation, (iii) systematically record data sensed from these movements, (iv) allow scientists to experimentally vary the sound properties of the object, without having to physically modify or replace it. We describe two prototypes specifically adapted to chimpanzees, which allow mapping sounds to physical movements and satisfy the requirements above. These prototypes constitute, to our knowledge, the first attempt at animal-computer rhythmic interaction. Here we outline their general features. For a thorough technical description and calibration data, see Ravignani et al. (in preparation).

Desiderata and General Features The prototypes were built with a main idea in mind: spurring the chimpanzee to spontaneous interaction and play. To maximize the chances of interaction, they were constructed and calibrated after scrutinizing videos of chimpanzees playing with objects, including the gum toy used in one of the prototypes. Each prototype consists of a sensing and a feedback unit. Sensing units feature acceleration and strain sensors embedded into manipulable objects. These units send acceleration or strain data to a computer, which converts them into sound and plays it in real time. The drum sets satisfy a number of logistic and technical desiderata. The sensing part is resistant, modular, low-voltage, inexpensive, interesting for the primate and easy to connect and configure. The software allows fast elaboration of data by performing few, simple operations, so as to limit the computational load. ${ }^{2}$
Prototype A: Wired The wired prototype is a parallelepiped containing piezoelectric sensors and connected to

[^196]a Mac computer via an Arduino ${ }^{3}$ board. A dedicated Python ${ }^{4}$ script is in charge of the auditory feedback. It can be mounted vertically on a wall or on the wire-mesh of chimpanzees' enclosures. This prototype has several advantages: (i) it is built with cheap and easy to find components, (ii) it entails no risk of electrocution and (iii) its ricochet property naturally suits the animal's tendency to hit and push objects.

Prototype B: Wireless The wireless prototype consists in a hollow dog toy enclosing a Wii Remote ${ }^{5}$. A computer receives data (via Bluetooth), which is processed and sonified using patches written in $\mathrm{Max}^{6}$. This device has several advantageous features: (i) chimpanzees generally enjoy manipulating objects, and chimpanzees have been both reported (Pruetz \& Bloomsmith, 1992) and observed by us to manipulate the model of toy used here (ii) its construction requires less work than the wired prototype and its components can be easily purchased, (iii) it has a wireless communication system, particularly advantageous in some applications.

## Circular Statistics

Most data coming from beat and rhythm experiments involve a periodic time component. Before applying a statistical test, it is essential to think about the nature and dimensionality of the data. As a parallel, if we wanted to compare the amount of rain falling on Britain over time, we should conceptualize rain as falling onto a 2-dimensional space, rather than on a real number line, $\mathbb{R}$. As the classical t-test for paired samples is defined on $\mathbb{R}$, it may be inappropriate to use it on geographical data. The fact that rhythmic data are originally associated with time makes time series analysis a possible approach to test a range of hypotheses.

The best option to use for periodic data, when possible, is circular statistics (Fisher, 1995). Its key feature consists in assuming that data is distributed on a circle, rather than on the usual real number line. This grants ideal analytical tools for data sets with a periodic time component, such as those deriving from beat and rhythm experiments. Several researchers in the field have successfully used these techniques on human (Kirschner \& Tomasello, 2009) and animal data (Hasegawa et al., 2011).

## The Importance of Modeling

Above we hinted at the importance of developing mathematical models of the emergence of rhythm. The last century has seen a radical increase in the quantitative approaches used in most areas of human knowledge. In particular, mathematical models and computer simulations have proven themselves particularly useful in testing the internal consistency of hypotheses, sharpening scientific assumptions and providing new viable directions for experimental testing. Scholars interested in the evolution and emergence of structure in

[^197]language, for instance, have provided quantitative accounts (Kirby, 2001), which have been later validated through cognitive experiments (Kirby, Cornish \& Smith, 2008).

Similarly, recent experiments (Honing et al., 2012) have shed some light on what can be accounted for by human culture or biology in rhythmic abilities. However, thorough explanations are still missing about the evolutionary forces, whether biological or cultural, that have shaped musical rhythm and the underlying human cognitive abilities. Quantitative evolutionary thinking can be used to study the emergence of music and rhythm, and models linking biology to culture could be an exciting second step.

While investigating what is unique about musical rhythm and which species possess the cognitive abilities to process it, human and animal experimental work should be complemented by models aimed at explaining the ultimate mechanisms of what is observed in everyday musical behavior. The lack of quantitative work trying to explain the emergence, cultural dynamics and biological evolution of music is surprising when we consider its pervasiveness in human lives.

## Conclusions

We suggested directions and methodologies for investigating the evolution of musical rhythm in a comparative, interdisciplinary perspective. Usage of a variety of statistical techniques on the same data set and replication are essential before conclusive claims of lack of synchronization can be made about a species or taxon. Moreover, experiments should be designed keeping in mind the critical theoretical distinctions introduced above.

Recent evidence provided by Honing et al. (2012) and Cook et al. (2013) may lead to newly redefined hypotheses, which in turn make the experimental testing of apes and marine mammals a fundamental prerequisite for a theory of human uniqueness of rhythmic abilities. The drum sets we presented are intended for apes to perform acoustic non-vocal rhytmhic production in a captive, though not restrained context. In general, as technological tools for human-machine interaction become available, new methodological paradigms for animal-machine interaction can be developed and used to test critical species in musical tasks. Mathematical modeling and agent-based simulations can be an important complement to empirical data, hopefully generating the same productive theory-experiments interplay seen in other disciplines.

Similarly to the broad variety of reaction time distributions across species and tasks, evolution has shaped animal brains and motor skills so that different species may require different statistical null hypotheses with respect to attempted synchronized motor behavior (for instance, due to perceptual or motor lower bounds on reaction times). Circular statistics, with its variety of theoretical distributions (von Mises, cardioid, wrapped normal, etc) and time-periodic tests, are ideal for testing hypotheses about rhythmic synchronization with different underlying assumptions.

The Vocal Learning and Social Convergence hypotheses make different predictions on which species should have
rhythmic abilities (Fitch, 2012). Both of them, however, are related to another uniquely human trait: language. Further development of experimental paradigms allowing social interactions under experimentally-controlled conditions will enable scientists to contrast these hypotheses and produce evidence relevant to the evolution and cognition of both music and language.

## Acknowledgments

A.R, B.G. and R.S. are supported by ERC Advanced Grant 230604 SOMACCA awarded to W.T.F. The development of the prototypes was additionally supported by an Exchange Grant ESF RNP "CompCog" (www.compcog.org) (06-RNP020) to A.R. We thank S. Kirby, U. Seifert, K. Smith, R. Hofer, C. Rodriguez, S. Pearson, A. Seed, K. Slocombe and A. Whiten. A.R. thanks the Gesellschaft für Kognitionswissenschaft for the travel support.

## References

Arcadi, A. C., Robert, D., \& Boesch, C. (1998). Buttress drumming by wild chimpanzees: Temporal patterning, phrase integration into loud calls, and preliminary evidence for individual distinctiveness. Primates, 39(4), 505-518.
Fisher, N. I. (1995). Statistical analysis of circular data. Cambridge University Press.
Cook, P., Rouse, A., Wilson, M., \& Reichmuth, C. (2013) A California Sea Lion (Zalophus californianus) Can Keep the Beat: Motor Entrainment to Rhythmic Auditory Stimuli in a Non Vocal Mimic. Journal of Comparative Psychology.
de Waal, F. B. (1988). The communicative repertoire of captive bonobos (Pan paniscus), compared to that of chimpanzees. Behaviour, 183-251.
Fitch, W. T. (2006). The biology and evolution of music: A comparative perspective. Cognition, 100(1), 173-215.
Fitch, W. T. (2012). The biology and evolution of rhythm: unraveling a paradox. In P. Rebuschat, M. Rohrmeier, J. A. Hawkins, \& I. Cross (Eds.), Language and Music as Cognitive Systems (73-95), New York: Oxford University Press.
Grahn, J. A. (2012). Neural Mechanisms of Rhythm Perception: Current Findings and Future Perspectives. Topics in Cognitive Science, 4(4), 508-606.
Hasegawa, A., Okanoya, K., Hasegawa, T., \& Seki, Y. (2011). Rhythmic synchronization tapping to an audio-visual metronome in budgerigars. Scientific reports, 1.
Hattori, Y., Tomonaga, M., \& Matsuzawa, T. (2013). Spontaneous synchronized tapping to an auditory rhythm in a chimpanzee. Scientific reports, 3.
Honing, H., Merchant, H., Háden, G. P., Prado, L., \& Bartolo, R. (2012). Rhesus Monkeys (Macaca mulatta) Detect Rhythmic Groups in Music, but Not the Beat. PloS one, 7(12).
Kirby, S. (2001). Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of
regularity and irregularity. IEEE Transactions on Evolutionary Computation, 5(2), 102-110.
Kirby, S., Cornish, H., \& Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. Proceedings of the National Academy of Sciences, 105(31), 10681-10686.
Kirschner, S., \& Tomasello, M. (2009). Joint drumming: Social context facilitates synchronization in preschool children. Journal of Experimental Child Psychology, 102(3), 299-314.
Konoike, N., Mikami, A., \& Miyachi, S. (2012). The influence of tempo upon the rhythmic motor control in macaque monkeys. Neuroscience Research.
Large, E. W., Velasco, M. J., \& Gray, P. M. (2008). Rhythmic analysis of musical interactions between bonobo and human. In Presentation given at the International Conference on Music Perception and Cognition. Sapporo, Japan.
Lerdahl, F., \& Jackendoff, R. (1983). A Generative Theory of Tonal Music. Cambridge: MIT Press.
McAuley, J. D. (2010). Tempo and Meter. In M. R. Jones, R. R. Fay, \& A. N. Popper (Eds.), Music Perception. Springer handbook of auditory research, Vol. 36, (165199). New York: Springer Science + Business Media.

Nagasaka, Y., Chao, Z. C., Hasegawa, N., Notoya T., \& Fujii N. (2013). Spontaneous synchronization of arm motion between Japanese macaques. Scientific reports, 3 .
Patel, A. D. (2008). Music, language, and the brain. Oxford University Press, USA.
Patel, A. D., Iversen, J. R., Bregman, M. R., \& Schulz, I. (2009). Studying synchronization to a musical beat in nonhuman animals. Annals of the New York Academy of Sciences, 1169(1), 459-469.
Pruetz J.D., Bloomsmith M.A., (1992). Comparing two manipulable objects as enrichment for captive chimpanzees. Animal Welfare 1:127-137.
Ravignani, A., Matellán, V. Gingras, B., Hofer, R., Rodriguez, C., \& Fitch, W.T. (in preparation) Sonification devices for acoustic pattern production in non-human primates using acceleration and strain sensors.
Schachner, A., Brady, T. F., Pepperberg, I. M., \& Hauser, M. D. (2009). Spontaneous motor entrainment to music in multiple vocal mimicking species. Current Biology, 19(10), 831-836.
Schusterman, R. J. (2008). Vocal learning in mammals with special emphasis on pinnipeds. The Evolution of Communicative Flexibility: Complexity, Creativity, and Adaptability in Human and Animal Communication, 41e70.
Zarco, W., Merchant, H., Prado, L., \& Mendez, J. C. (2009). Subsecond timing in primates: Comparison of interval production between human subjects and rhesus monkeys. Journal of neurophysiology, 102(6), 3191-3202.

# Training Principle-Based Self-Explanations: Transfer to New Learning Contents 

Alexander Renkl (renkl@psychologie.uni-freiburg.de), Judith Solymosi (judith.solymosi@gmail.com)<br>University of Freiburg, Germany, Department of Psychology, Engelbergerstaße 41, D-79085 Freiburg Germany

Michael Erdmann (michael.erdmann@web.de)<br>St. Ursula-Gymnasium Freiburg, Eisenbahnstraße 45<br>D-79098 Freiburg Germany

Vincent Aleven (aleven@cs.cmu.edu)<br>Human-Computer Interaction Institute, School of Computer Science, Carnegie Mellon University 5000 Forbes Ave, Pittsburgh, PA 15213 USA


#### Abstract

The present study tested the transfer effects of a short training intervention on principled-based self-explanations. The intervention used fables as well as mathematics examples and problems as "exemplifying" domains for training such selfexplanations. The effects were tested in a new learning environment about attribution theory and feedback messages. In this experiment, 58 German high-school students were randomly assigned to the self-explanation training condition or a control condition (i.e., training of mnemonic strategies). The learning outcomes from the learning environment about attribution theory and feedback did not significantly differ between groups. However, those students who also reported to have applied the strategies from the training intervention actually showed the best learning outcomes. Overall, the selfexplanation training intervention "convinced" just part of the learners to engage in principle-based self-explanations in a new environment. There seems to be two options to achieve more reliable effects by future training interventions: The learners have to be prompted more clearly that they should employ the learned strategies in the transfer learning environment or the short-term training intervention should be extended to have a stronger effect on spontaneous strategy application.


Keywords: Self-explanation, training intervention, transfer.

## Introduction

If students acquire cognitive skills, these skills should ideally be based on an understanding of the underlying domain principles (e.g., Chi \& VanLehn, 2010; Goldstone \& Day, 2012; Renkl, 2002). Such a conceptual underpinning facilitates the transfer of the acquired skills to new problems for which a modified solution procedure has to be found. In addition, deep conceptual understanding is considered to facilitate further procedural learning (e.g., Rittle-Johnson, Siegler, \& Alibali, 2001). In many learning situations, however, the learners acquire cognitive skills without understanding the corresponding domain principles. Thus, a major goal of instruction is to facilitate meaningful learning that strives for a principle-based understanding.
One way to induce a principle orientation for meaningful learning is to prompt learners for principle-based selfexplanations (Kalyuga, 2011). For example, Atkinson,

Renkl, and Merrill (2003) encouraged learners to determine the principle (here: probability rule) behind each step of a worked example. This prompting procedure fostered transfer to isomorphic and to novel problems, for which modified solution procedures had to be found. Principlebased prompting also worked in "verbal" domains without mathematical solution procedures. For example, Schworm and Renkl (2007) provided principle-based prompts to learners when they studied video examples of sound scientific argumentation. Such prompts help determine the argumentative structures and, thereby, the argumentation skill. Whereas the Atkinson et al. and the Schworm and Renkl studies analyzed example-based learning, Aleven and Koedinger (2002) showed that principle-based selfexplanation prompts also enhance learning by problemsolving (here: in the intelligent tutorial environment Cognitive Tutor). Further, there are numerous studies affirming the positive effects of prompting principle-based self-explanations (e.g., Berthold \& Renkl, 2009; Conati \& VanLehn, 2000; Renkl, 1997; Schworm \& Renkl, 2006).

The successful prompting procedures have, however, significant disadvantages. First, when prompts in the form of external guidance are provided, there is no guarantee that the learners do not fall back on rote learning when the prompts are not present anymore (cf. Wecker \& Fischer, 2011). Second, it is a substantial amount of work to enrich learning materials or environments with prompts; it may not be practical to do so for all materials a learner may need, or even to know what learning materials a learner may need in the future. It would be far preferable if the learners acquired self-explanation skills that they can use for further selfregulated learning in new learning environments.

There are several tried-and-tested self-explanation training interventions. However, they all have restrictions with respect to fostering principle-based self-explanations when learners study worked examples and solve problems in order to acquire cognitive skills. McNamara and colleagues focus on reading strategies in their selfexplanation training interventions SERT and iStart (McNamara, 2004; Levinstein, Boonthum, Pillarisetti, Bell, \& McNamara, 2007). These strategies are not tailored to
learning by examples and problem-solving. A restriction of other training interventions for self-explaining examples and problems that have been tested so far is that they employed the same type of materials in the training phase as in a subsequent learning phase (e.g., Bielaczyc, Pirolli, \& Brown, 1995). For example, Renkl, Stark, Gruber, and Mandl (1998) trained participants using examples of (compound) interest calculation in order to prepare them for a later learning phase dealing with the same domain. The self-explanation training of Wong, Lawson, and Keeves (2002) focuses on geometry learning in all phases.

The expectation that the self-explanation strategies addressed by these previous training will solely transfer to similar contents seems to be realistic because transfer to dissimilar contexts (e.g., different learning domain) is very hard to achieve (e.g., Detterman 1993; Goldstone \& Day, 2012; Perkins, 2009). Nevertheless, some researchers found some training effects that transfer over contents. For example, Chi and VanLehn (2010) had their learners work in an intelligent tutoring environment called "Pyrenees" (domain: probability) that demanded, among other things, a focus on domain principles. The learners were prompted to reason about the principles in order to determine sought values and they had to apply the principles to the problems at hand. It was found that this principle orientation transferred when working in another intelligent tutorial environment (i.e., "Andes"; domain: probability and physics); this was in particular true for learners with less prior knowledge. Note that there was not only a transfer across learning environments (Pyrenees to Andes) but also across learning domains (probability to physics).
Whereas Chi and VanLehn (2010) found transfer of a principle orientation acquired during physics learning, Busch, Renkl, and Schworm (2008) developed a training intervention with the "sole" purpose to foster selfexplanations. This short intervention (less than 30 min .) was conducted with the topic "fables." The learners were shown that in order to determine that a short story is a fable one has to self-explain whether some crucial principles were implemented in the story (e.g., animals as actors, hidden message). This intervention showed considerable transfer effects to a rather distant topic: example-based acquisition of scientific argumentation skills. Although this short-term training was surprisingly successful, it had a significant restriction. Although there was transfer from fables to scientific argumentation, it was "just" transfer between verbal domains. As the Busch et al. intervention did not refer to mathematical solution procedures, which are typical not only of mathematics but also of many science subdomains, we did not expect transfer to the latter domains. Hence, it is sensible to modify the Busch et al. training intervention by including mathematical contents.

## The Present Study

We trained high-school students providing selfexplanations in two domains. As in the study by Busch et al. (2008) we used fables as "verbal" exemplifying domain, and
mathematics as an algorithmic exemplifying domain. Afterwards the students learned from an example-based learning environment how to apply psychological attribution theory in order to provide feedback that has favorable motivational effects. This content domain was not taught or mentioned in the training intervention. Hence, we test the hypothesis that the self-explanation training using mathematics problems and fables as materials has positive effects on learning about the provision of productive feedback on the basis of attribution theory.

As control group, we did not use a non-treatment group, as these effects might be rather trivial. Instead, we compared the self-explanation intervention with a training intervention on mnemonic strategies. Although the latter strategies might be useful for remembering facts, we hypothesized that the self-explanation intervention is more favorable for highlevel learning goals (e.g., applying what has been learnt about feedback to evaluating new feedback messages).

When testing the effects of a modified version of the short training intervention by Busch et al. (2008), we tried to keep the training time short, that is, about half an hour (as in the original training intervention). Such a short training intervention is applicable within the usual class periods in schools. In the self-explanation intervention, we kept the basic example of a fable in order to demonstrate the value of principle-based self-explanations. In addition, we used mathematics examples in order to show how to self-explain while studying mathematics examples and while solving mathematics problems. We saved some training time in order to add mathematics contents by focusing on principlebased self-explanations and leaving out other types of selfexplanations (e.g., goal-operation elaborations) that were part of the original training intervention. Nevertheless, we had to shorten the treatment of self-explaining fables in order to keep the intervention time within the limits of about half an hour. A question that arose was whether the training intervention has still transfer effects to other verbal areas, even if the treatment of fables as verbal training examples was substantially reduced. The unique contribution of this study is the evaluation of a self-explanation training intervention that is designed to have across-domain transfer effects, that is, effects that are not bound to the "exemplifying" domains used during training.

## Method

## Participants and Design

We randomly assigned 58 female high-school students (age: $M=16.52, \mathrm{SD}=0.71$ ) to two conditions: training intervention on principle-based self-explanations ( $\mathrm{n}=31$ ) and training intervention on mnemonic strategies ( $\mathrm{n}=27$ ). The participants were members of elective courses in psychology from a "mono-educational" (i.e., just female students) Gymnasium (i.e., highest high-school track of the German three-track system). The main dependent variable was the learning outcomes in a learning environment that followed the different training interventions. This transfer
environment was about attribution theory and its application to providing productive feedback. The contents that the students learned in this experiment were not directly related to their currently treated topics in their psychology courses. However, they were (validly) informed that the topics fit the overall learning goals of these courses.

## Instruments and Materials

Short-term training environments. We compared the transfer effects of two training environments: Training of principle-based self-explanations versus training of mnemonic strategies. They lasted about half an hour. Both training interventions were parallel in a number of features. They both introduced the fictitious character Sarah who had learning difficulties (see Figure 1). In both cases, a friend helps out by suggesting some strategies (i.e., principle-based self-explanations or mnemonic strategies, respectively). Both training modules presented the contents within a dialogue between Sarah und her friend. During the program the learners in both conditions got work sheets in order to practice the respective strategies. Both modules ended with a short summary of the training contents.


Figure 1: Screenshot from the training intervention module on principle-based self-explanations (translated form German)

The training intervention on principle-based selfexplanation was divided into two main modules, which explained and practiced principle-based self-explanations when (a) studying an example and (b) solving a problem. The first example in the example-studying section was Aesop's fable "The fox and the crow." We showed that a fable is characterized by several principles or underlying features (e.g., animals as actors, principle of polarization, hidden message) and that the readers have to self-explain a story in terms of above-mentioned underlying features in order to identify the story as a fable (see Figure 1). Next the learners practiced principle-based self-explanations, supported by corresponding prompts, on a work sheet presenting a worked example applying the Pythagorean

Theorem. Hence, a first instance of inter-domain transfer was practiced. In the second part, we supported further transfer by presenting and practicing principle-based selfexplanations when solving diverse mathematics problems.

The training intervention on mnemonic strategies introduced and practiced three strategies: (a) Using mental images; (b) "Eselsbrücken," which is a German term for (in many cases funny) phrases that interconnect two items (e.g., word in a foreign language and translation). (c) "Mnemonic sentences" similar to "My very educated mother just served us nine pickles" for the planets and their distances to the sun (note, however, that we used other mnemonic sentences because this one does not work in German language).

Transfer environment. The transfer environment first introduced the concept of attribution and explained why attributions are important in learning contexts. Then it introduced the basics of Kelley's (1971) attribution theory, that is, the co-variation model. On this basis, it explained how feedback should be given to students so that functional attributions are fostered. Two small exercises were included in which the participants had to analyze feedback statements. Finally, a summary was provided. The learner worked on average $7.10 \mathrm{~min}(S D=2.02)$ in this module (no significant difference between the conditions).

Posttest. The posttest assessing the transfer effects of the self-explanation training consisted of 15 problems (average time: 23 min ). In addition, the posttest booklet asked three questions that were to be answered on 5-point rating scales at the very beginning (I found the first program useful; I found the second program useful; in the second program I applied the strategies that I have learned in the first program). After these questions, we presented the problems assessing the learning outcomes.

Three problems asked what should be emphasized in feedback in different circumstances. Six items asked for the attribution theory principles behind exemplary feedback messages (e.g., "In a dance class: A lot of people struggle with Tango" (the feedback message itself is printed in italics). Solution: Such feedback suggests attributions to task difficulty and it should "prevent" internal attributions when having difficulties). Four items required writing a short feedback statement for different circumstances. Finally, two items ask for identifying what is problematic with two suboptimal feedback statements. This scale had a good internal consistency (Cronbach's $\alpha$ of .86).

## Procedure

The students participated at experimental group sessions in a university computer laboratory (about 20 students per session). The students worked individually in front of a computer. The different computers were randomly assigned to one of the two experimental conditions. These sessions lasted about 100 min . At first glance, this duration is longer than to be expected from the average time of the single phases such as training intervention, transfer environment, and posttest. Note, however, that the faster students had to wait for the slower ones before going on to the next phase.

After some welcome words, we informed the students that they will learn about some learning strategies in a first computer-based learning environment and that they should apply these strategies in a second computer-based learning environment. Subsequently, students were asked to fill in a short paper-pencil questionnaire on demographic data (one page), previous school grade, and on learning goal orientation (Dweck \& Leggett, 1988). As the latter scale neither predicted learning outcomes nor interacted with the different treatments we did not consider the students' learning goal orientation in the following.
After completing the questionnaire the students worked on the training intervention modules. Subsequently the students learned about feedback and attribution theory in a second learning program. Finally, they took the posttest.

## Results

A significance level of .05 was used for all analyses. We used $d$ as an effect-size measure with values between .20 and .50 classified as small, values between .50 and .80 as medium, and values > . 80 as large (Cohen, 1988).
We did not find any significant differences between the groups with respect to the grade point average on the last report card or the experience with learning programs (both $F \mathrm{~s}<1$ ). Eight students said that they have never heard the term attribution. Fifty students said that they have already heard about this term but did not remember its meaning. No student was able to explain what attribution means. Overall, the student had hardly any prior knowledge.
The self-explanation condition scored descriptively higher on the posttest as compared to the mnemonics condition, $M$ $=.55, S D=.23$ vs. $M=.48, S D=.21(M \mathrm{~s}$ represent the percentage scores as compared to the theoretically possible maximum). However, this difference did not reach the level of statistical significance, $t(56)=1.08, p=.286, d=0.32$. This relatively weak and statistically not significant effect could be due to the following factors: (a) The effect of the self-explanation training intervention interacts with learning prerequisites (aptitude-treatment interaction explanation); (b) some of the learners superficially scanned or quickly read the training module (scan and skim explanation); (c) the training module was too difficult at least for some learners (difficulty explanation); (d) the learned selfexplanation strategies were not applied by some learners in the application environment on attribution and feedback (production deficiency explanation).
(a) Aptitude-treatment interaction explanation. The most important learning variable with respect to aptitudetreatment interaction is prior knowledge or achievement level (Kalyuga, 2007). The grade point average, as indicator of prior school achievement, was significantly related to the posttest ( $r=.37, p=.005$ ). However, there was no interaction between condition and grade point average, with respect to the posttest, $F<1$. Further exploratory analysis with other learning prerequisites (e.g., grades for mathematics or German; experience with computer-based learning program) did not indicate any aptitude-treatment
interaction. Hence, the aptitude-treatment interaction explanation is likely not true.
(b) Scan and skim explanation. If the weak and insignificant transfer effect was due to some learners' just scanning and skimming the training environment, there should be a correlation between learning time and training outcomes. However, the learning time in the training modules was not significantly related to learning outcomes, neither in the whole sample ( $r=.05, p=.699$ ) nor in the two sub-groups (self-explanation group: $r=.11, p=.551$; mnemonic strategies group: $r=-.15, p=.455)$. In this context, it should also be noted that the self-explanation group spend more time in the training module, $M=29.87$, $S D=6.70$, than the mnemonic group, $M=24.07, S D=6.89$, $t(56)=3.25, p<.002, d=0.85$. Overall, there is no indication that some learners in the self-explanation condition just quickly scanned the training module, which impeded their learning outcomes. Hence, the scan and skim explanation is likely not true.
(c) Difficulty explanation. If the self-explanation training intervention was too difficult for some learners, there should be a substantial number of errors in practice sheets that were included in the learning environment, and the number of errors in these practice sheets should predict lack of transfer. To test this explanation, we coded the quality of the students' responses to the four interspersed work sheets in the self-explanation training module from 1 (completely wrong) to 5 (correct, clear principle application). We found a mean of $4.35(S D=0.55)$, clearly indicating that the training was not too difficult for the learners. In addition, there was no significant correlation between the worksheet score and the posttest ( $r=.18, p=.180$ ). Overall, the difficulty explanation is likely not true.
(d) Production deficiency explanation. We asked the participants to rate on a five-point scale whether they applied the strategies learned in the first module (selfexplanation or mnemonics, respectively) in the second module on attribution, as suggested by the experimenter in the beginning of the session. When adding this rating in the prediction of learning outcomes (predictors: condition, rating, and condition by rating), we found a significant interaction effect between condition and reported strategy application with respect to the posttest, $F(1,54)=9.72, p=$ .003. To better understand this interaction, we determined the regression scores and their statistical significance in both conditions. In the self-explanation condition, the more the students reported that they applied the learned strategies, the better the posttest performance, $b=0.09, t(29)=2.59, p=$ .015. In the mnemonics condition, we did not find a significant relation between self-reported strategy application and posttest performance in the transfer environment, $\beta=-0.07, t(25)=-1.89, p=.071$. In accord with a production deficiency explanation, these findings indicate that only part of the students applied the learned strategies in the module on attribution and feedback and, thereby, profited with respect to learning outcomes.

In order to get an idea of how many non-applying students were "responsible" for the insignificant overall training effect, we conducted some post-hoc analyses. When we excluded the three students from the self-explanation condition who stated that they did not at all apply the strategies (i.e., choosing 1 on the 1 to 5 rating scale of strategy application), there was still no significant effect of condition on learning outcomes. However, when excluded an additional nine students, namely, all students who stated that they did not apply the strategies (i.e., choosing 2 on the 1 to 5 rating scale of strategy application), the condition effect gets to be statistically significant (self-explanation condition: $N=19 ; M=.63, S D=.22$; mnemonics condition (as already reported): $N=27 ; M=.48, S D=.21, t(44)=$ $2.23, p=.031, d=.70)$. Hence, only when we consider the (roughly) two thirds of the students that were convinced to apply the strategies, we get a significant effect of the selfexplanation training intervention.

The preceding post-hoc analysis might be criticized because we excluded only participants from the selfexplanation condition and we, therefore, had rather different group sizes. If we also exclude the ten participants from the mnemonic condition (i.e., roughly the lower third) that reported about low strategy application, we also got a significant group difference: self-explanation condition (as already reported): $N=19 ; M=.63, S D=.22$; mnemonics condition: $N=17 ; M=.42, S D=.23, t(34)=2.70, p=.011$, $d=.89$ ). This finding again underlines that the selfexplanation treatment was successful in about two thirds of the cases.

## Discussion

We tested whether we could successfully implement a short-term training intervention on principle-based selfexplanations that has positive effects on learning in a subsequent learning environment. Unfortunately, we got only a weak and statistically insignificant effect. According to our post-hoc analyses, it is unlikely that this weak effect was due to aptitude-treatment interactions with learning prerequisites, a scan-and-skim behavior of some learners, or the difficulty of the training intervention. Instead, only part of the students (about two thirds) was "convinced" by the training intervention to apply the learned strategies in a subsequent learning environment. Students who applied the strategies profited from the training intervention.
Why did some learners not apply the self-explanation strategies? There are at least three possible explanations: (a) These learners did not find the strategies in the selfexplanation training useful; (b) it was not salient enough, at least for some learners, that they were expected to apply the strategies that they have learned in the first environment in the second learning environment; (c) the training intervention was too short to fully change the students' habitual learning behavior.

The perceived usefulness argument can be evaluated by further post-hoc analyses. After completing the learning environment on attribution theory and before they took the
posttest, the learners rated how useful they found the strategy training module. The learners from the selfexplanation condition rated this module as rather useful ( $M$ $=4.03, S D=0.98$ (5-point scale from 1 , not at all, to 5 , fully agree). The perceived utility predicted to some extent whether the strategies applied ( $\mathrm{r}=.36, \mathrm{p}<0.05$ ). However, the perceived utility did not interact with the treatment, in contrast to the reported strategy application. Obviously, the (low) perceived usefulness was not a major cause for not applying the strategies and for reduced training effects.

How salient was it for the learners that they should apply the learned strategies in the second learning environment? In the beginning of the experimental sessions, the experimenter informed the students that they should apply the strategies to be learned in a first environment in the second computer-based learning environment. However, this prompt was not repeated (keep in mind that the question of to what extend the strategies learned in the first program were applied was posed after the transfer phase). Note also that in the beginning of the session, the students got a variety of information and were confronted with many new "impressions," that is, they came to a new building (i.e., Department of Psychology), they were introduced to the computer room and the experimenter, they were informed about various aspects of the study, etc. Thus, for some students, the instructions about strategy application might not have been very salient and they might not have been remembered when they began to work on the second learning environment. It seems plausible that - given the short duration of the training intervention so that no profound effect on habitual behavior can be expected - the students would need at least some form of "kick-off" prompt at the start of the transfer learning environment to apply the learned strategies to new contents.

As already argued, the short training duration makes it implausible that the students' habitual strategy use was changed. Against the background of the present intervention's short duration and the corresponding transfer literature (e.g., Detterman 1993; Goldstone \& Day, 2012), it can even be regarded as success that about two thirds of the learners transferred the newly learned self-explanation strategies across domains.

In this context, it should also be noted that we have replaced the verbal self-explanation training materials of Busch et al. (2008) to a large degree with mathematical examples. Given that Busch et al. found significant transfer effects across two verbal domains, it can be tentatively assumed that the "verbal part" was too much reduced. Hence, a sensible next step in improving the training intervention would be to extend the verbal part roughly to the length of the Busch et al. intervention. Thus, we extend the verbal part of our training intervention in a next step on our way to develop some type of "generic" self-explanation training. We also intend to test transfer effects on mathematical learning environments.

An alternative explanation for the positive training effect when looking at the two thirds of student reporting strategy
application is that the mnemonic intervention suppressed at least some students' tendency to self-explain spontaneously. Hence, further studies should also include a control condition allowing for spontaneous self-explanations.
Overall, the present study and Busch et al. (2008) have taken partly successful steps towards a self-explanation strategy training that has the potential to achieve acrossdomain transfer effects. Nevertheless, there is some further research to be done (e.g., extending the intervention; testing transfer to mathematical contents). However, the available findings justify some optimism that we can step by step come to a successful training approach.

## Acknowledgments

This work was supported by the Pittsburgh Science of Learning Center, which is funded by the National Science Foundation; award number SBE-0354420.

## References

Aleven, V., \& Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. Cognitive Science, 26, 147-179.
Atkinson, R. K., Renkl, A., \& Merrill, M. M. (2003). Transitioning from studying examples to solving problems: Combining fading with prompting fosters learning. Journal of Educational Psychology, 95, 774783.

Berthold, K., \& Renkl, A. (2009). Instructional aids to support a conceptual understanding of multiple representations. Journal of Educational Psychology, 101, 70-87.
Bielaczyc, K., Pirolli, P., \& Brown, A. L. (1995). Training in self-explanation and self-regulation strategies: Investigating the effects of knowledge acquisition activities on problem solving. Cognition \&Instruction, 13, 221-252.
Busch, C., Renkl, A., \& Schworm, S. (2008). Towards a generic self-explanation training intervention for example-based learning. In P. A. Kirschner, F. Prins, V. Jonker, \& G. Kanselaar (Eds.), Proceedings of the 8th International Conference of the Learning Sciences 2008 (CD version only). Utrecht, NL: ICLS.
Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.
Conati, C., \& VanLehn, K. (2000). Toward computer-based support of meta-cognitive skills: A computational framework to coach self-explanation. International Journal of Artificial Intelligence in Education, 11, 398415.

Chi, M. \& VanLehn, K. (2010). Meta-cognitive strategy instruction in intelligent tutoring systems: How, when, and why. Journal of Educational Technology and Society, 13, 25-39.
Detterman, D. L., (1993). The case for the prosecution: Transfer as epiphenomenon. In D. K. Detterman \& R. J. Sternberg (Eds.), Transfer on trial: Intelligence,
cognition, and instruction (pp. 1-24). Norwood, NJ: Ablex.
Dweck, C. S., \& Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. Psychological Review, 95, 256-273.
Goldstone, R. L., \& Day, S. B. (2012). Introduction to "New Conceptualizations of Transfer of Learning." Educational Psychologist, 47, 149-152.
Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. Educational Psychology Review, 19, 509-539.
Kalyuga, S. (2011). Cognitive load theory: How many types of load does it really need. Educational Psychology Review, 23, 1-19.
Kelley, H. H. (1971). Attribution in social interaction. New York: General Learning Press.
Levinstein, I. B., Boonthum, C., Pillarisetti, S. P., Bell, C., \& McNamara, D. S. (2007). iSTART 2: Improvements for Efficiency and Effectiveness. Behavior Research Methods Instruments and Computers, 39, 224-232.
McNamara, D. S. (2004). SERT: Self-explanation reading training. Discourse Processes, 38, 1-30.
Perkins, D. (2009). Making learning whole: How seven principles of teaching can transform education. San Francisco, CA : Jossey-Bass.
Renkl, A. (1997). Learning from worked-out examples: A study on individual differences. Cognitive Science, 21, 129.

Renkl, A. (2002). Learning from worked-out examples: Instructional explanations supplement self-explanations. Learning \& Instruction, 12, 529-556.
Renkl, A., Stark, R., Gruber, H., \& Mandl, H. (1998). Learning from worked-out examples: The effects of example variability and elicited self-explanations. Contemporary Educational Psychology, 23, 90-108.
Rittle-Johnson, B., Siegler, R. S., \& Alibali, M. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. Journal of Educational Psychology, 93, 346-362.
Schworm, S., \& Renkl, A. (2006). Computer-supported example-based learning: When instructional explanations reduce self-explanations. Computers \& Education, 46, 426-445.
Schworm, S., \& Renkl, A. (2007). Learning argumentation skills through the use of prompts for self-explaining examples. Journal of Educational Psychology, 99, 285296.

Wecker C., \& Fischer F. (2011). From guided to selfregulated performance of domain-general skills: The role of peer monitoring during the fading of instructional scripts. Learning \& Instruction, 21, 746-756.
Wong, R. M. F., Lawson, M. J., \& Keeves, J. (2002). The effects of self-explanation training on students' problem solving in high-school mathematics. Learning \& Instruction, 12, 233-262.

# Predicting similarity change as a result of categorization 

Irene Reppa (i.reppa@swansea.ac.uk)<br>Department of Psychology, Swansea University, Swansea, SA2 8PP<br>Emmanuel Pothos (e.m.pothos@gmail.com)<br>Department of Psychology, City University London, London, EC1V 0HB


#### Abstract

Learning a particular categorization leads to corresponding changes in the similarity structure of the categorized stimuli. The purpose of the current study was to examine whether different category structures may lead to greater or less similarity change. We created six category structures and examined changes in similarity as a result of categorization in between-participant conditions. The bestsupported hypothesis was that the ease of learning a categorization affects change in similarity, with the most change following learning of difficult category structures. There was also support for the hypothesis that similarity change is more likely to occur when the category boundary was not aligned with the physical dimension of variation. Finally, we discuss some methodological challenges in addressing this important research topic.


Keywords: similarity; categorization; learning difficulty; exemplar theory

There is widespread evidence that learning to categorize stimuli in a particular way leads to corresponding changes in the similarity structure of the stimuli (e.g., Gureckis \& Goldstone, 2008; Ozgen \& Davies, 2002; Schyns \& Oliva, 1999; Stevenage, 1998). For instance, stimuli categorized in the same category tend to be perceived as more similar to each other, compared to stimuli categorized in different categories (e.g., Goldstone, 1994; Schyns, Goldstone, \& Thibaut, 1997), and stimuli on either side of a category boundary tend to be more discriminable than stimuli on the same side of the boundary (e.g., Harnad, 1987). Similarly, differences have been reported on color perception across different linguistic communities (Roberson et al., 2005).

Research on the influence of categorization on perception has flourished for several reasons. It is theoretically important since it is at the heart of answering core issues regarding representation and the processing of sensory input. Do we perceive a faithful representation of sensory input? Or are our perceptual representations a compromise between constraints from sensory input and whatever categories are useful for the organism? Such research is also important for formal models of categorization, as most of them assume categorization models that assume representations which are stable across learning (e.g., Nosofsky, 1984).
The present research examined the effects of categorization on similarity. Changes in similarity might
correspond to perceptual changes or changes mediated through the addition of a category label (e.g., Goldstone, Lippa, Shiffrin, 2001; McMurray et al., in press; Roberson \& Davidoff, 2000; Sloutsky \& Fisher, 2004). Choosing to examine similarity is primarily a methodological simplification, since exploring directly changes relating to perception involves the technical challenge of eliminating (possible) effects from linguistic labeling. However, if across broadly matched category structures, for instance, in terms of learning difficulty, we find similarity changes following learning of some structures but not others, then one can make the additional step of inferring similarity changes over and above changes due to just the category label (see also Roberson et al., 2007).

Despite the numerous reports on the effects of categories on similarity, and perception in particular, there have been some reports of failures of such influence (e.g., Goldstone, 1994; Jiang et al., 2007; Freedman et al., 2003). The aim of present research was to examine possible factors might lead to changes in similarity.
We created six two-dimensional category structures, shown in Figure 1. Two category structures were designed so that the width dimension was diagnostic (Width easy and Width difficult), while the height dimension non-diagnostic, and two more category structures were defined so that height was the diagnostic dimension (Height easy and Height difficult) and width was non-diagnostic. Two versions of each category structure were created, one designed to be easy (e.g., Width easy), and one designed to be more difficult (e.g., Width difficult). Finally, two more category structures were created where both dimensions were relevant: the non-linearly separable (NLS) and the Diagonal structure, explained in more detail later.
Three different hypotheses regarding the effect of category learning on similarity changes were examined. One hypothesis was that category learning difficulty would affect the extent of similarity changes. A classification is easy (or more intuitive) if it is more readily obvious to naïve observers (Pothos \& Chater, 2002; Pothos et al., 2011). For example, when asked to freely classify a set of stimuli, participants will generate more intuitive classifications more frequently. These classifications will be typically easier to learn than non-intuitive ones. Category learning difficulty might influence similarity ratings in two ways. One possibility is that learning the easy category structures would lead to greater changes in similarity ratings. This is because, for easy category structures participants are able to
quickly learn the underlying categorization, perhaps with less emphasis on encoding the individual exemplars (cf. Ashby et al., 1999; Ashy \& Ell, 2002). Such inexact initial encoding of the exemplars may mean that exemplar representations end up being developed in a way that is more consistent with the underlying category structure (e.g., Edwards, Pothos, \& Perlman, in press). Support for this prediction comes from Folstein, Gauthier, and Palmeri (2010), who manipulated the complexity of the underlying stimulus space (not of category structure, as in the current study). Unlike previous related evidence showing that categorization does not influence similarity (e.g., Jiang et al., 2007; Freedman et al., 2003), they showed significant effects of categorization on perception, when the underlying stimulus space was simple.
The converse prediction, regarding the effect of category learning difficulty on similarity changes, is that learning a difficult category structure might result in more significant and enduring changes in the similarity structure of the stimuli. This possibility is motivated by evidence showing that supervised categorization processes can involve processes of selective attention or other changes in psychological space (e.g., through the sensitivity parameter; Nosofsky, 1984), though such research does not tell us whether such changes are enduring and on the actual stimulus representations.
A second hypothesis is that the linear separability of the learned categories might moderate changes in similarity. Overall, there is quite a lot of controversy regarding the role of linear separability in category learning and perhaps some of this controversy can be ultimately explained in terms of corresponding changes in the similarity structure of the categorized stimuli. Note that connectionist models require that NLS problems are transformed into linearly separable ones at their hidden layer, otherwise learning is not possible (indeed, the inability of perceptrons to learn NLS category structures has been at the heart of the critique of Minsky \& Papert, 1969; see also Rumelhart \& McClelland, 1986).

To examine possible influences of linear separability in relation to similarity changes, two categories were created to be broadly equal in terms of complexity but differ in whether they were LS or not. One was the NLS and the other was the Diagonal condition (Figure 1). In diagonally separated category structures, the members of one category can only be discriminated from their nearest neighbors in the other category with fine distinctions along both dimensions of variation. They have proved to be challenging for participants to learn (e.g., Ashby, Queller, \& Berretty, 1999; Ashby \& Ell, 2002).
If the cognitive system shares processing constraints with connectionist systems, maybe it would try to re-represent a NLS classification in an LS way, so that there would be more similarity change in learning an NLS classification, compared to an equally complex but LS one (the Diagonal one). Alternatively, it could be the case that more complex classifications are associated with less similarity change, if, for example, category learning of such classifications
involves rote memorization of the training exemplars (Blair \& Homa, 2003). In this case, NLS and Diagonal classification would lead to equivalent similarity changes.


Figure 1. The six category structures employed in the study.

Finally, a third hypothesis is that similarity change depends on whether the category boundary is aligned with a dimension of physical variation (e.g. height). According to the COVIS model of categorization (Ashby et al., 1999; Ashby \& Ell, 2002), whether the category boundary is along a physical dimension of variation can determine whether the executive (frontal) or the procedural system is engaged. To examine this hypothesis, the Diagonal category structure was compared against the condition best matched for difficulty with it (which turned out to be the Height Difficult category structure).

## Participants and Design

One hundred and eighty experimentally naïve participants, all Swansea University students, were tested. There were 20 participants in each of the six experimental groups, each learning one of the six different classifications shown in Figure 1. For all six classifications, successful learning is achieved when participants recognize that items in clusters A and B are in one category and items in clusters C and D are in another category. The dependent variable was changes in similarity as a result of category learning. The procedure for computing changes in similarity is described in the Results section.
Three independent variables were considered, to allow examination of the hypotheses examined. The first was
category structure learning difficulty, which was defined adhoc in terms of the number of trials to criterion. For this variable the Height Easy and Width Easy conditions were compared with their corresponding Height Difficult and Width Difficult conditions. The second independent variable was whether the category boundary was aligned with the dimension of physical variation or not. The third independent variable was linear separability with two levels, linearly versus non-linearly separable category structures.

For each of the six category structures, there was a corresponding control group providing similarity ratings for the stimuli, but without having gone through the categorization task first. For the Width Easy, Height Easy, and NLS conditions there was a common control group of 20 participants. For the Width Difficult and the Height Difficult conditions a different control of 20 participants, and for the Diagonal group yet another control group of 20 participants. The experiment lasted approximately 50 minutes for the experimental groups and 30 minutes for the control groups.

## Materials

We used yellow surface-rendered arrow-like shapes that varied in terms of two dimensions: the width of the arrowhead (horizontal dimension) and the length of the arrow (vertical). The smallest arrow's trunk measured 4.5 centimeters (cm) in height and its head measured 3.0 cm wide. Twenty-four more stimuli were created by incrementing trunk height and head width by $12 \%$. The stimuli employed in the experimental conditions were subsets of this original set of stimuli. The shortest arrow trunk in all six conditions was 4.5 cm high and the narrowest arrow head 3.0 cm wide. The tallest arrow trunk was 12.5 cm in the Width Easy, Height Easy, Diagonal, and NLS conditions and 7.1 cm in the Width Difficult and Height Difficult conditions. The widest arrow head was 8.3 cm in the Width Easy, Height Easy, and NLS conditions, 4.7 cm in the Width Difficult and Height Difficult conditions and 5.3 cm in the Diagonal condition.

## Procedure

A standard supervised categorization task was employed. A stimulus was presented at the center of a computer screen against a white background, until the participant decided whether it belonged to category A or B , at which point he/she received corrective feedback. Participants continued to categorize stimuli until no mistakes were made for 32 consecutive trials (i.e., all stimuli shown twice) or for a maximum of 256 trials. Five participants failed this criterion (three in the NLS condition and two in the diagonal condition) and these participants were not asked to complete the similarity part of the study. Participants, who completed the categorization task successfully subsequently received the similarity ratings task. In that task, each trial started with a 'Ready?' prompt at the center of the screen. Two stimuli appeared at the screen center for 500 ms each, one after the other, with an inter-stimulus interval of 500 ms . All possible
$16 \times 16=256$ stimulus pairs were presented and participants were asked to rate their similarity on a 1-9 scale, such that 1 corresponded to 'very dissimilar' and 9 to 'very similar'. Participants were encouraged to use the entire scale. Participants in the control groups went through the similarity ratings, without having done the categorization task first.

## Results

## Data cleaning

There were two simple checks that the participants were sufficiently attentive during the similarity ratings task. Participants who did not use the whole similarity rating scale (1-9) and those who did not rate two identical stimuli as identical (by giving them an average rating of seven or above) were excluded from the data. This procedure led to the elimination of 3 participants from the Width Easy group, 3 from the Height Easy group, 2 from the Width Difficult, 1 from the NLS group, and 3 participants from the control groups.

## Learning Results

Trials to criterion and errors correlated highly with each other ( $r=.84, p<.0005$ ). Both the trials to criterion and the errors varied across category structures $[F(5,105)=15.25$, $p=.0005$ and $F(5,105)=11.75, p=.0005$, respectively]. The 'easy' versions of category structures were easier to learn than the 'difficult' versions of the classifications. Also, participants found it easier to learn the 'width' classifications than the 'height' ones, a result showing that the perceptual salience of the two dimensions was not equivalent. Category structures defined along a single dimension (Width easy/difficult \& Height easy/difficult) were easier to learn than those defined along two dimensions (NLS and Diagonal), $t(109)=5.94, p=.0001$. There was no difference in ease of learning between LS category structures and the NLS one, $t(109)=.48, p>.05$. Finally, and as expected, the NLS and Diagonal classifications were the most difficult ones to learn, with no difference between them ( $\mathrm{p}>.05$ ).

## Similarity measures

Change in similarity as a result of learning could be quantified in various ways. The measures typically employed in studies of changes of perception, as a result of categorization, emphasize discriminability along diagnostic vs. non-diagnostic dimensions (e.g., Folstein, Gauthier, \& Palmeri, 2010; Goldstone, 1994). However, in the present study, any putative similarity changes as a result of categorization would relate to the categorization objective, that is, learning the different category structures. Therefore, it is more appropriate to consider a measure of similarity change, which is informed by the category structures in each case. Following theory on the determinants of category structure (Rosch and Mervis, 1975; Love, Medin, \& Gureckis, 2004; Pothos \& Chater, 2002; Pothos \& Bailey, 2009) and categorization work in general (e.g., Mathy et al.,
in press), we employed two dependent variables for how the similarity structure might change as a result of categorization: within category and between category similarity change. Note that similarity in these definitions is empirical similarity from participant ratings. Within and between category similarity change allow us to directly explore the circumstances when the similarity structure for a set of stimuli becomes more consistent with a learned categorization. Within category similarity was the average similarity from all possible pairs in the same category and between category similarity was the average similarity ratings for all pairs across different categories.

Table 1. Trials to criterion, errors, and change in within and between similarity values, as a result of learning, for the six category structures employed in this study. The category structures have been ordered in terms of difficulty of learning. Asterisks indicate that the difference between experimental and control groups for each condition, revealed by independent-samples t-tests, was significant. Positive values indicate that the mean similarity value was higher for the experimental group compared to the control group, while negative values indicate the opposite.

|  |  |  | Similarity change |  |
| :--- | :---: | :--- | :--- | :--- | :--- |
| Category structure | Trials to <br> criterion | Errors | Within | Between |
| Width Easy | 45.50 | 7.2 | $.28^{*}$ | -.13 |
| Width Difficult | 71.00 | 5.6 | $.45^{*}$ | .11 |
| Height Easy | 71.50 | 15.0 | .19 | -.07 |
| Height Difficult | 101.35 | 21.1 | $.59^{*}$ | .11 |
| NLS | 102.60 | 34.6 | $.57^{*}$ | .14 |
| Diagonal | 172.00 | 35.9 | .41 | $-.43^{*}$ |

To provide baseline similarity values, within and between category similarity was calculated for the control participants, following the calculation procedure for their respective experimental groups.
Once similarity values were computed for all groups (both experimental and control), similarity change values were computed for each experimental group. Clearly, any changes in similarity as a result of categorization are only meaningful compared to a pre-learning baseline. For example, suppose a participant provided similarity ratings for the stimuli after learning the Width Easy classification. We would then compute her, e.g., between similarity change value as the between similarity value from her similarity ratings minus the average between similarity value of all corresponding control participants. Henceforth, when we
refer to change in similarity values we imply similarity values computed in this way from the similarity ratings of the control participants, for each category structure. Adopting this analytical approach considerably simplifies comparisons of similarity changes across different category structures. Within and between similarity change can be understood as acquired equivalence and distinctiveness, respectively, but defined in terms of the learned categorizations, rather than stimulus dimensions.

Table 2. The $F$-tests examining the three hypotheses regarding similarity changes as a result of category learning.

| Hypothesis | Similarity change |  |
| :--- | :--- | :--- |
| Within | Between |  |
| Learning difficulty   <br> (Height vs. Width, $F(1,68)=5.40$, $F(1,68))=.92$, <br> easy and difficult) $p=.02$ $p=.89$ |  |  |
| Category boundary <br> aligned with physical | $F(1,36)=.48$, | $F(1,36)=5.66$, |
| variation <br> (Height Difficult vs. | $p=.49$ | $p=.02$ |
| Diagonal) <br> Linear separability <br>  | $F(1,55)=.13$, | $F(1,55)=.04$, |
| Diagonal vs. NLS) | $p=.72$ | $p=.84$ |

## Similarity Analyses

Similarity change for the six category structures we employed are shown in Table 1. For within similarity changes, positive values indicate changes in the similarity structure of the items more consistent with the learned classification. For between similarity changes, it is the other way round; between similarity is defined in terms of the similarity of items in different categories, so that if between similarity is negative this means that items in different categories become less similar (and therefore consistent with the learned classification).
The hypothesis that learning difficulty influences similarity change was examined in a 2 (Dimension: width vs. height) x 2 (Difficulty: easy vs. difficult) ANOVA. For within similarity changes, there was a significant main effect of Difficulty (shown in Table 2), with greater similarity change for difficult category structures than easy ones. There was no main effect of Dimension, $F(1,68)=.03$, $p>.05$, nor a significant interaction, $F(1,68)=.90, p>.05$. For between similarity change, there was no main effect of Difficulty, $F(1,68)=.92, \mathrm{p}>.05$, or Dimension, $F(1,68)=.01$, $p>.05$, and no significant interaction, $F(1,68)=.02, p>.05$.
The hypothesis that similarity changes would be determined by whether the category boundary is aligned with the dimension of physical variation, or not was examined in a one-way ANOVA (Height Difficult VS. Diagonal category structures). As shown in Table 2 this hypothesis was supported for between but not for within similarity changes. Finally, linear separability did not predict within or between similarity changes (Table 2).

## Discussion

There has been considerable interest in changes in similarity (and perception) induced as a result of categorization, though few researchers have attempted a systematic study of the factors that make such changes likely (for an exception see Folstein et al., 2010). The overarching question in this research was whether the nature of the category structure is a relevant factor in trying to understand changes in similarity as a result of categorization. Three main possibilities were considered. The first possibility was that category difficulty would influence similarity change. We suggested that in cases where there are well-separated categories, similarity change may correspond more to within similarity change (cf. ChinParker \& Ross, 2004), but for more poorly separated categories between similarity change may be more pronounced. In either case, more difficult category structures were expected to lead to greater similarity change. Our findings supported this hypothesis, but only partially. Difficulty of learning a category structure predicted changes in within category similarity, with stimuli in the same categories becoming more similar for more difficult, compared to the easier category structures.

The second possibility was that similarity changes are influenced by whether the category boundary was aligned with a dimension of physical variation. Indeed, this hypothesis was supported only for between category similarity change: when the category boundary was not aligned with a dimension of physical variation, then stimuli in different categories became less similar following categorization training. Although the influence of this factor could not be anticipated by prior work on similarity changes (e.g., work on categorical perception; Harnad, 1987), its role can be predicted within modern categorization theory (e.g., Ashby et al., 1998; Ashby \& Ell, 2002). For instance, when the category boundary is aligned with a dimension of physical variation, even when the categories are poorly separated, participants focus on within category information, rather than on between category contrasts. Work on the COVIS model of categorization shows that category boundaries aligned with a dimension of physical variation are simpler than ones which are not, even for poorly separated categories (Ashby et al., 1999). Therefore, the complexity of the category boundary instead of the actual difficulty of the category structures (as defined in this study), may be a factor driving between similarity change. This possibility needs further work to be fully supported.

Finally, linear separability predicted neither within nor between similarity changes, even though this factor was manipulated across conditions, which were broadly matched for learning difficulty.

One debate in the literature concerns the extent to which similarity changes reflect perceptual changes, changes in item representation, changes in the category's internal structure, the addition of a label as a feature in determining similarity, or simply task demands. This is an important issue that is beyond the scope of the current research. It is
important to note, however, that our finding that task difficulty influences the magnitude of similarity changes, is inconsistent with the view that similarity changes are due to the addition of category label to stimulus representations. That is, if a category label was added to stimulus representations in all cases, we should not have observed different degrees of similarity change for different category structures.

The current research revealed several methodological challenges in the study of changes in similarity as a result of categorization. First, several kinds of category structures are needed. Second, it is clearly of crucial importance to specify an appropriate index of similarity change, which takes into account possible differences between category structures. Indeed, in the present study, we did not observe equivalent results across the measures we introduced within and between similarity change. Researchers specifically interested in perception often consider acquired distinctiveness or equivalence, as a result of categorization (e.g., Goldstone, 1994; Harnad, 1987). Such measures are suitable when there are stimuli on either sides of a category boundary, but they are perhaps less suitable when the nearest neighbor stimuli on either side of a category boundary may be distant from each other in psychological space. This will often be the case for category structures that are meant to correspond to naturalistic ones (cf. Pothos et al., 2011).

A major methodological challenge was comparing effects of categorization on similarity for the different category structures. To do this we computed similarity values on the basis of similarity ratings, after they have been potentially modified by category learning (experimental participants) and without any categorization learning (control participants). Consequently, the dependent variables corresponded to the change of similarity ratings as a result of categorization. While we believe our solution to this problem to be adequate, it would be worthwhile to explore alternative approaches in future research.

Overall, the issue of whether some category structures are more or less likely to lead to corresponding changes in the similarity structure of categorized stimuli is a novel and exciting one. Here we presented a promising design to address it and a range of preliminary conclusions.

## References

Ashby, G. F. \& Ell, S. W. (2002). Single versus multiple systems of category learning: Reply to Nosofsky and Kruschke (2002). Psychonomic Bulletin \& Review, $\underline{9}$, 175-180.
Ashby, F. G., Queller, S., \& Berretty, P. M. (1999). On the dominance of unidimensional rules in unsupervised categorization. Perception \& Psychophysics, 61, 11781199.

Blair, M. \& Homa, D. (2001). Expanding the search for a linear separability constraint on category learning.
Memory \& Cognition, 29, 1153-1164.

Blair, M. \& Homa, D. (2003). As easy to memorize as they are to classify: The 5-4 categories and the category advantage. Memory \& Cognition, 31, 1293-1301.
Chin-Parker, S. \& Ross, B. H. (2004). Diagnosticity and prototypicality in category learning: A comparison of inference learning and classification learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 30, 216-226.
Edwards, D. J., Pothos, E. M., \& Perlman, A. (in press). Relational vs. absolute representation in categorization. American Journal of Psychology.
Folstein, J.R., Gauthier, I., \& Palmeri, T. J. (2010). Not all spaces stretch alike: How the structure of morphspaces constrains the effect of category learning on shape perception. Vision Sciences Society, 10th annual meeting.
Freedman, D. J., Riesenhuber, M., Poggio, T., and Miller, E. K. (2003). A comparison of primate prefrontal and inferior temporal cortices during visual categorization. Journal of Neuroscience, 23, 5235-5246.
Goldstone, R. L. (1994). Influences of Categorization on Perceptual Discrimination. Journal of Experimental Psychology: General, 123, 178-200.
Goldstone, R. L., Lippa, Y., Shiffrin, R. M. (2001). Altering object representations thought category learning. Cognition, 78, 27-43.
Gureckis, T. M. \& Goldstone, R. L. (2008). The effect of the internal structure of categories on perception. In B. C. Love, K. McRae, \& V. M. Sloutsky (Eds.), Proceedings of the 30th Annual Conference of the Cognitive Science Society (pp. 843). Austin, TX: Cognitive Science Society.
Harnad, S. (Ed.) (1987). Categorical Perception. Cambridge: Cambridge University Press.
Jiang, X., Bradley, E., Rini, R.A., Zeffiro, T., Vanmeter, J., and Riesenhuber, M. (2007). Categorization training results in shape- and category-selective human neural plasticity. Neuron 53, 891-903.
Love, B. C., Medin, D. L., \& Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. Psychological Review, 111, 309-332.
Mathy, F., Haladjian, H. H., Laurent, E., \& Goldstone, R. L. (under review). Similarity-dissimilarity competition in disjunctive classification tasks.
McMurray, B., Aslin, R. N., Tanenhaus, M.K., Spivey, M. J., \& Subik, D. (in press). Gradient sensitivity to withincategory variation in words and syllables. Journal of Experimental Psychology: Human Perception and Performance.
Medin, D. L. \& Schwanenflugel, P. J. (1981). Linear Separability in Classification Learning. Journal of Experimental Psychology: Human Learning and Memory, 7, 355-368.
Minsky, M. L. \& Papert, S. A. (1988). Perceptrons. $3^{\text {rd }}$ Edition. MIT Press: Cambridge, MA.
Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. Journal of Experimental Psychology: Learning, Memory, and Cognition, 10, 104114.

Ozgen, E. \& Davies, I. R. L. (2002). Acquisition of categorical color perception: A perceptual learning approach to the linguistic relativity hypothesis. Journal of Experimental Psychology: General, 131, 477-493
Pothos, E. M. \& Chater, N. (2002). A simplicity principle in unsupervised human categorization. Cognitive Science, 26, 303-343.
Pothos, E. M. \& Chater, N. (2005). Unsupervised categorization and category learning. Quarterly Journal of Experimental Psychology, 58A, 733-752.
Pothos, E. M. \& Bailey, T. M. (2009). Predicting category intuitiveness with the rational model, the simplicity model, and the Generalized Context Model. Journal of Experimental Psychology: Learning, Memory, and Cognition, 35, 1062-1080.
Pothos, E. M., Chater, N., \& Stewart, A. J. (2004). Information about the logical structure of a category affects generalization. British Journal of Psychology, 95, 371-386.
Pothos, E. M., Perlman, A., Bailey, T. M., Kurtz, K., Edwards, D. J., Hines, P., \& McDonnell, J. V. (2011). Measuring category intuitiveness in unconstrained categorization tasks. Cognition, 121, 83-100.
Roberson, D. \& Davidoff, J. (2000). The categorical perception of colors and facial expressions: The effect of verbal interference. Memory \& Cognition, 28, 977-986.
Roberson, D., Davidoff, J., Davies, I. R. L., \& Shapiro, L. R. (2005). Color categories: Evidence for the cultural relativity hypothesis. Cognitive Psychology, 50, 378-411.
Rosch, E. \& Mervis, C. B. (1975). Family Resemblances: Studies in the Internal Structure of Categories. Cognitive Psychology, 7, 573-605.
Rumelhart, D. E. \& McClelland, J. L. (1986). Parallel distributed processing: explorations in the microstructure of cognition. Cambridge: MIT Press.
Ruts, W., Storms, G., \& Hampton, J. (2004). Linear separability in superordinate natural language concepts. Memory \& Cognition, 32, 83-95.
Schyns, P. G. \& Oliva, A. (1999). Dr. Angry and Mr. Smile: When categorization flexibly modifies the perception of faces in rapid visual presentations. Cognition, 69, 243-265.
Schyns, P., Goldstone, R. L., \& Thibaut, J. (1997). The development of Features in object concepts. Behavioral and Brain Sciences, 21, 1-54.
Sloutsky, V. M. \& Fisher, A. V. (2004). Induction and categorization in young children: A similarity-based model. Journal of Experimental Psychology: General, 133, 166-188.
Smith, J. D., Murray, M. J. Jr.,\& Minda, J. P. (1997). Straight talk about linear separability. Journal of Experimental Psychology: Learning, Memory, and Cognition, 23, 659-680.
Stevenage, S. V. (1998). Which twin are you? A demonstration of induced categorical perception of identical twin faces. British Journal of Psychology, 89, 39-57.

# Inferring Sensory Experiences 

Kevin Reuter<br>kevin.reuter@rub.de<br>Department of Philosophy, Ruhr University Bochum


#### Abstract

Several scholars, e.g. Sellars (1956), Meltzoff \& Gopnik (1993), have construed the attribution of experiences as being governed by a folkpsychological theory in which experiences function as theoretical entities. However, so far this claim has not been convincingly supported by an account of how people infer the existence of experiences. In this paper I argue that the mechanisms that lead to the stipulation of experiences are fundamentally inferential and are applied in both self-attribution and third-person attribution of experiences. The two most common sources for going through such inferential processes are (i) disagreements between two people in how the world is presented to them, (ii) being aware of or suspecting differences between how the world is presented to a person and extraneous information the person has about the world. From situations like these, I show that 'experience' is a theoretically-acquired concept which refers to entities that play an explanatory role in virtue of fulfilling two conditions: a person entertains the concept experience if that person makes an appearance-reality distinction (C1) and considers the appearance to be subjective (C2).


Keywords: experiences, self-ascription, selfattribution, introspection, appearances, appearancereality distintion, theory-theory, Austin

## 1. The Appearance-Reality Distinction

It is widely held that self-ascribing experiences requires a person to conceive of the way things appear and not how they really are. Tye claims that "if you are attending to how things look to you, as opposed to how they are independently of how they look, you are bringing to bear your faculty of introspection" (2000, p. 46), and Dretske argues that when we self-ascribe experiences "we are conceiving of how things seem" (1994, p.266-7). What it means to conceive of how things seem, however, remains mostly unclear. More
specifically, it is hardly ever discussed, which appearance-statements count as attribution of experiences and which do not. This is especially curious as noone believes that every appearancestatement involves the attribution of an experience, e.g. "she looks chic" is an appearance-statement from which we cannot infer the attribution of an experience.

We often make appearance-statements when we know or suspect that we are in a situation in which it would be wrong to take what we seem to perceive at face value. E.g. we state that 'the Müller-Lyer lines only appear to be of different length' (illusion), 'my phone only appears to be ringing' (hallucination), 'the wall looks green to me but it is white' (unusual lighting conditions), 'the sponge looks like a rock' (different surface conditions) - see figure 1 below.


Figure 1: The Müller-Lyer illusion (left) and a sponge looking like a rock (right) are grasped by making an appearance-reality distinction.

It would be wrong, however, to infer without argument that in all of these cases attribution of experiences takes place. Austin makes an observation that deserves our attention. He states:
"It is perhaps even clearer that the way things look is, in general, just as much a fact about the world, just as open to public confirmation and challenge, as the way things are. I am not disclosing a fact about myself, but about petrol, when I say that petrol looks like water." (1962, p.43, my italics)

The case of a white wall looking green because it is illuminated by green light, and the example of a
sponge looking like a rock because it has a surface that has a rock-like structure, seem to support Austin's claim, that the way things look is as much a fact about the world, just as open to public confirmation and challenge, as the way things are. When it comes to visual examples, the photo test seems to be a possible means to track whether the way an object appears is largely a fact about the world: The illuminated wall will yield a greenish tone on the photo picture, and the sponge is indistinguishable from a rock on the photograph. In contrast, the way the Müller-Lyer lines look to a person does not seem to be subject to public confirmation and challenge. Making a photograph of the lines will not yield lines of differing length on the photograph (assuming of course a head-on photo). Hallucinations, dreams and afterimages are also not subject to public confirmation.

What is important for our discussion, is that Austin seems to have established a feature of appearancestatements which does not hold for all cases, and that this criterion supports our intuition that we selfascribe experiences just in cases in which the appearance is not open to public confirmation. However, it is of course possible for a person to ascribe a sensory state despite the reason for why something appears to be different from the way it really is, is dependent on external conditions of perception; and surely, people can self-ascribe their sensory states even though there is no reason to doubt the veridicality of their experiences. Hence, a simple classification of appearance-statements that are made because the senses are deceived as selfascriptive, and appearance-statements that are made because the conditions of perception are unusual as objective, seems false. What really seem to matter, so I will now argue, are the inferences people make when thinking about appearances. To illustrate this point, let us look at two examples: mirages and the moon illusion.

## 2. Inferring the Existence of Experiences

People are often aware or at least suspect that the world appears different from the way it really is. However, they also often lack an understanding of the reasons for why they make an appearancereality distinction. Mirages are most commonly associated with a thirsty and exhausted person traveling through the desert, suddenly seeming to see an oasis. People offer two distinct explanations
to account for this appearance: illusion and optical phenomenon. Whereas it could of course be the case that a traveler starts to hallucinate an oasis, mirages are properly explained (and most frequently happen) by the bending of light rays from distant objects and can be captured on camera - thus they are optical phenomena. The moon illusion, on the other hand, is a phenomenon that often occurs when people look at the moon which is just above the horizon. The moon appears to be larger on the horizon than it is high up in the sky (see figure 2 below). It was originally thought that due to light refraction in the atmosphere, the moon on the horizon occupies more space in the visual field than it would normally do. However, the moon illusion is not an optical phenomenon but is explained by the workings of our perceptual apparatus and is related to the Ponzo illusion, and cannot be captured on a photograph.


Figure 2: Example of the Moon illusion.
We can see from these two examples that people can be justifiably uncertain about whether the appearance of an object is different from its reality due to external physical conditions or internal psychological conditions. Sometimes it is simply very difficult to ascertain the reasons for a difference in appearance and reality. Misattributions of this difference go both ways: mirages are often misattributed to deranged sensory perception, the moon illusion is often misattributed to the physical properties of the atmosphere. But this also has the consequence that the meaning of an appearancestatement depends on which state - physical or psychological - is blamed for the conceived difference between appearance and reality. Although it is correct to say that the appearancestatement 'the moon appears larger on the horizon' cannot be subject to public confirmation or challenge, a person can say the same words but express a different appearance-statement, one that he thinks expresses openness to public
confirmation. Similarly, although the appearancestatement 'there seems to be an oasis' is usually taken to be subject to public confirmation, a person can utter the same sentence, but uses the sentence in a different sense which precludes public confirmation.

If people consider the appearance-statement to be open to public confirmation, they do not focus on their mental states as representing the world in a certain way, but rather talk about how the world is independent of their experiences. In contrast, they might use an appearance-statement to talk about their experiences themselves - statement which then of course should be classified as self-attributive. We are now in a position to formulate the two necessary and jointly sufficient conditions for self-attributing an experience.

Self-attribution of experiences: a subject $S$ selfattributes an experience iff,
(C1) S distinguishes the appearance from the reality of what $S$ experiences,
(C2) $\quad \mathrm{S}$ considers the appearance to be subjective.

Similarly, we can easily specify the conditions for attributing an experience to another person:

Attribution of experiences to others: a subject $S$ attributes an experience to agent O iff,
(O1) S distinguishes the appearance from the reality of what O experiences,
(O2) S considers the appearance to be subjective.

Being aware of or suspecting differences between the way the world presents itself and other beliefs we have about the world, is one of the two main sources for why people attribute experiences. We often suspect that the way the world presents itself is not how it really is with our less dominant senses, e.g. it may seem to us as if there is faint smell of burned toast in the air, but it might as well be just our imagination or a different scent that we misinterpret accordingly. In these and similar cases, suspicion often arises as to whether we experience the world differently from how it really is.

Disagreements between people about the way the world is perceived to be, are the second main source for attribution of experiences. Suppose you are at a ball, talking to a woman, and comment on how the
red-coloured dress suits her style. She responds by politely pointing out that the dress is of an orangecoloured shade. The disagreement about the colour might persist, and you cannot get yourself to see the orange colour. Once again, there seem to be two options: you can either blame the lighting conditions or some other external condition, or you can blame your way of perceiving the world for the disagreement, and thus attribute to yourself an experience.

Both the first-person and the third-person cases are very similar. People attribute experiences to other people for the same reasons that they self-attribute experiences: there is a difference between what seems to be the case for oneself and what seems to be the case for another person. This situation needs an explanation, and thus people postulate the existence of experiences.

We can generalise this case by following inferential theory-theoretic rule (TT):
(TT-A) If the perceptual statements of two honest people differ, and if these people
have sufficient knowledge and are in a position to make correct statements about the world, it is reasonable to draw a distinction between how the world appears and how the world really is.
(TT-B) If there is no obvious worldly cause for why people disagree with each other, it makes sense to consider the possibility that the way the world appears is dependent on the sensory system.
(TT-C) In 'blaming' the sensory system for the way the 'world' appears, attribution of experiences takes place.

## 3. The Explanatory Power of this Theory-Theory

This theory of self-ascription of experiences not only specifies two necessary and jointly sufficient conditions for the self-attribution of experiences, but also (a) analyses the concept of experience, and (b) captures the cognitive requirements that we need to be in possession of in order to be capable of metacognition.
(ad a) There is a need to specify the minimal but also sufficient conditions for the possession of the concept experience. (C1) and (C2) state that a
person entertains the concept of experience, if that person makes an appearance-reality distinction (C1) and considers the appearance to be subjective (C2). These two conditions have the advantage of specifying the conditions for entertaining the concept of experience in a non-circular way, i.e. we do not need to mention experiences in the description of the two conditions.
(ad b) Empirical results from the psychological literature on the appearance-reality task have a direct bearing on this discussion. According to this theory, children need to make two important advancements in their understanding of the nature of the world in order to have the ability to ascribe experiences. First, children need to understand that things can appear to be different from the way they are ( C 1 ). Second, they need to understand that the reason for why things can appear different is not always found in the environment itself (C2). In one of the well-known appearance-reality tasks, a sponge that looks like a rock, is presented to children. Several psychologists and philosophers argue that once children pass the appearance-reality task, they have gained the cognitive ability to selfascribe experiences (Taylor \& Flavell (1984), Nichols \& Stich (2003)). But this conclusion is not warranted. It is possible, and indeed highly plausible, to presume that children do not consider the reason for the appearance-reality distinction to be a matter of perceiving the sponge, but rather to blame the deceptive appearance on a visible property of the sponge. Thus, a child might pass the appearance-reality task because he associates the appearance of the sponge as a rock with the form of the sponge, e.g. the surface, shape, and size of the sponge is rock-like. The child does not need to think that the appearance of rock-like properties is explained by the way the sponge is perceived.

It is of course possible to self-ascribe a veridical experience without having any doubts whatsoever regarding the veridical nature of the experience. A person might entertain the concept of experience for other reasons (e.g. she just read about the brain-in-a-vat thought experiment) or for no obvious reason at all (e.g. sometimes thoughts enter our mind without us knowing where they 'came from'). The main conclusion of this paper is that self-awareness of experiences is often arrived at by an inferential process that is governed by a folk-psychological theory that we learn as children. The concept of experience is therefore a theoretically-acquired concept which refers to entities that play an explanatory role in that theory.

The theory I have presented belongs in certain respects to the group of theories which are commonly labelled 'theory-theories of selfawareness' (TTSA) and which originate in the theories of Ryle (1949) and Sellars (1956). Proponents of TTSA hold a theory according to which a person who self-attributes an experience, uses a theory to do so - hence the name 'theorytheory of self-awareness'. In a sense, I also claim that a person often uses a theory to self-ascribe her experiential states, but it must be clearly differentiated from other theory-theories that have been advocated in the literature. It is often argued that theory-theory accounts of self-awareness state that self-ascriptions of mental states are based on the same or similar processes as ascriptions of mental states to others (Kind (2005), Schwitzgebel (2010)). This claim is ambiguous: it can either mean that we determine the content of other people's mental states by the same or similar mechanisms that we use when determining the content of our own mental states; or it can mean that the mechanisms that lead from thoughts about the world to the attribution of experiences to others are the same or similar to the mechanisms when selfattributing mental states. Whereas I reject the former claim, I endorse a restricted form of the latter thesis. We do not infer what we experience but rather that we experience. What we experience, we know by perceptual attention and recognition. That we experience, we know by theorising about the circumstances we often find ourselves in

## 4. Objections

In this section $I$ evaluate two rather specific objections against construing the self-attribution of experiences in the way I have suggested above. First, Carruthers argues that the theory that children apply when they attribute experiences, is a nativistic theory. I show that Carruthers's claim is not warranted and that instead the inferential mechanisms that we apply when attributing experiences, are learned by children. Second, Papineau argues that experiences cannot be entities embedded in a folk-psychological theory because we can conceive of experiences as epiphenomenal states. I agree that Papineau's argument is valid, but he presupposes a conception of experiences as theoretical entities that my theory is not committed to.

Are self-attributive inferences nativistic or learned?

Carruthers claims:
"I favour such a nativistic theory-theory because if, firstly, young children are pictured as little scientists, constructing a mentalistic theory as the best explanation of the data, then it beggars belief that they should all hit upon the same theory, and at the same tender age too (at about the age of four, in fact). But if, secondly, the theory is supposed to be learned by the child from adult practitioners, then it is puzzling how this can take place without any explicit teaching or training;" (1996, p.23)

I agree with Carruthers that if we do conceive of young children as little scientists who need to construct their mentalistic theory all by themselves, then we should expect diverging theories to emerge from different children at different ages. But what is Carruthers's evidence that adult practitioners do not teach young people their theory of experiences? He gives none. So let me present what I consider to be a plausible story about how adults teach young children to self-attribute their experiences.

I have argued that the capability to make a distinction between appearance and reality is one of two necessary conditions required to grasp the concept experience. Adults not only provide children with the linguistic tools to make this distinction, they also actively educate children in making this distinction in appropriate circumstances and explain why appearance and reality can come apart. If a child asks what a rainbow is, the parent might tell the child that rainbows are not what they appear to be, but are only optical phenomena; or when the lights are switched off, then things do not become black, but only appear to be black. By then, children have also learned that bodily states like pains and itches are only felt by themselves, simply because they are the only ones who are 'connected' with their body. The distinction between public and private objects is, moreover, manifested in everyday conversations with adults. Adults ask children whether they are hungry, and ask them where the pain is located, but they tell them to watch out where they are going, tell them to eat food, and tell them to listen to what they say. Thus, children learn that (a) objects often do not change but rather their perceptual properties do, and that (b) they can draw a distinction between private objects and public objects of discourse. However, if it was not for situations in which children are required to combine
both ideas, the stipulation of experiences would seem unnecessary - at least from a folkpsychological standpoint. These situations exist though, and provide Kuhnian puzzles for children. Although illusions and hallucinations are the prime examples of non-veridical experiences in the philosophical literature, the most common and persistent non-veridical experiences are dream experiences. Children are not only regularly confronted with dreams they also want support from their parents especially after having had nightmares. But now the following curious situation occurs: parents tell their children that they need not worry because what they dreamed was not real. However, they also ask their children about what happened in their dreams. Thus, children are told that they had appearances that are private to them. They now only need to combine these two ideas, and thereby understand the concept of experience. I therefore reject Carruthers's claim that the concept of experience is nativistic.

## Do experiences necessarily have effects?

Papineau considers the possibility that selfattribution like 'this auditory experience', utilises a concept of experience, that is "some kind of theoretical concept, constituted by its role in some theory of experiences" (2007, p.121). He rejects this solution because he argues that if we really derive the concept of experience from some folkpsychological theory then we conceive of experiences as states with causes and effects. These causes and effects do not need to be specified by folk-psychology but they would be nonetheless an analytic part of our conception of experience. However, Papineau states, we can without contradiction think of experiences as epiphenomenal states without any subsequent effects like behaviour.

Papineau's argument is sound but only under a more specific reading of what a theoretical concept involves, which we do not need to accept. I argue that when children learn how to distinguish appearance from reality, they theorise that experiences have causes but not necessarily any effects. Possessing the concept of experience enables a person to conceptually distinguish veridical from illusory or hallucinatory experiences, and if there is a difference between appearance and reality, the person understands that the experience was not merely caused by worldly objects and events. However, I do not see any reason why we
should believe that this concept of experience is committed to the fact that experiences have cognitive or behavioural effects. Applying the appearance-reality distinction only implies that appearances are caused by the world, but not that appearances themselves have any effects. Thus, epiphenomenalism is consistent with phenomenal concepts being constituted by a theoreticallyacquired concept of experience.

## 5. Conclusion

In this paper I have argued that we understand experiences to be (a) appearances that are (b) subjective. In most circumstances in which a person attributes an experience, there are very good reasons for doing this. These reasons follow the twofold structure of the concept of experience. I have argued that there are two main 'sources' of experience attribution. First, a person might become aware of or suspect differences between how the world appears and how it really is. Second, disagreements between two people about how the world 'presents itself', can make people aware of such differences. The inferential nature of attribution of experiences was defended against two objections.

## References

1. Austin, J.L. (1962). Sense and sensibilia. Oxford: Oxford University Press.
2. Carruthers, P. (1996). Autism as mind-blindness: an elaboration and partial defence. in $P$. Carruthers, P. Smith (eds.) Theories of Theories of Mind, Cambridge University Press.
3. Dretske, F. (1994). Introspection. Proceedings of the Aristotelian Society. 94, 263-78.
4. Kind, A. (2012). Introspection. in The Internet Encyclopedia of Philosophy, ISSN 2161-0002, http://www.iep.utm.edu/introspe/.
5. Meltzoff, A., Gopnik, A. (1993). The role of imitation in understanding persons and developing theories of mind. in Baron-Cohen Understanding other minds: Perspectives from autism, Oxford: Oxford University Press.
6. Nichols, S, Stich, S. (2003). How to read your own mind: a cognitive theory of selfconsciousness. in Q. Smith and A. Jokic (eds.) Consciousness: New Philosophical Perspectives, Oxford: Oxford University Press.
7. Papineau, D. (2007). Phenomenal and perceptual concepts. in T. Alter and S. Walter (eds.)

Phenomenal Concepts and Phenomenal Knowledge, NY: Oxford University Press.
8. Ryle, G. (1949). The concept of mind. Chicago: Chicago University Press.
9. Schwitzgebel, E. (2010). Introspection. The Stanford Encyclopedia of Philosophy, E. Zalta (ed.), URL = http://plato.stanford.edu/entries/introspection/
10.Sellars, W. (1956/1997). Empiricism and the philosophy of mind. Cambridge, MA: Harvard University Press.
11.Taylor, M., Flavell, J. (1984). Seeing and believing: children's understanding of the distinction between appearance and reality. Child Development, 55, 1710-20.
12.Tye, M. (2000). Consciousness, Color and Content. Cambridge: MIT Press.

# Modeling the Emergence of Lexicons in Homesign Systems 

Russell Richie (russell.richie@uconn.edu)<br>Department of Psychology, 406 Babbidge Road, Unit 1020<br>Storrs, CT 06269-1020

Charles Yang (charles.yang@babel.ling.upenn.edu)<br>Departments of Linguistics, Computer Science, \& Psychology<br>619 Williams Hall<br>Philadelphia, PA 19104-6305

Marie Coppola (marie.coppola@uconn.edu)<br>Department of Psychology, 406 Babbidge Road<br>Department of Linguistics, 365 Fairfield Way<br>Storrs, CT 06269-1145


#### Abstract

Longitudinal data of conventionalization in emerging languages, combined with computational models explaining such data, are lacking in the literature on language emergence. In the present study we report on the emergence of gestural communication systems ("homesigns") invented by deaf individuals in Nicaragua. Analysis of longitudinal data from several families shows gradual convergence toward a gestural system with the essential characteristics of a shared lexicon. We propose a general computational framework to formalize the linguistic and social interactions among the individual signers such that a shared lexicon may arise. More specifically, a reinforcement learning process that adjusts the individual's probability of gesture use in response to others' actual gesture use provides a suitable account of the observed gestural convergence. Implications for language emergence are discussed.


Keywords: lexicon; homesign; conventionalization; language emergence; computational modeling; sign language; multiagent reinforcement learning model

## Introduction

How do languages emerge? What kinds of learners and environments, and particularly patterns of interaction among learners, give rise to language? The spontaneous emergence of gestural communication systems in deaf individuals not exposed to spoken or signed language (homesigners; Coppola \& Newport, 2005; Brentari \& Coppola, 2012) and of natural languages in deaf communities (Polich, 2005; Meir, Sandler, Padden \& Aronoff, 2010) offer unique opportunities to study the process of natural language emergence. Computational models, in contrast, allow formalization and testing of theories of language emergence. These two approaches clearly complement each other, yet there have been no integrations of the two in the literature on language emergence. To begin to rectify this, in this paper we compare empirical data from emerging sign systems to computational models to investigate emergence of a fundamental component of language: the lexicon. In particular, we investigate the process of conventionalization
of lexicons among small groups of individuals. We begin by reviewing extant literature on conventionalization.

Conventionalization of form-meaning mappings among interacting agents has been a major focus of language emergence research, mostly in experimental (see Galantucci, Garrod, \& Roberts, 2012 for review) and computational (Hutchins \& Hazlehurst, 1995; Barr, 2004; Steels \& Loetzsch, 2012) investigations. Human adults are brought into the lab to develop novel communication systems under various conditions (Selten \& Warglien, 2007), but in nearly all cases, conventionalization is observed among participants. In a related literature, researchers have investigated how language-learning biases shape communication systems as they are transmitted and learned across multiple generations (Kirby, Cornish, \& Smith, 2008). The basic finding is that human learners exposed to unsystematic form-meaning mappings will restructure these form-meaning mappings to be more compositional and learnable.

Conventionalization in natural language emergence is far less studied-the opportunities to observe the process are of course few and far between, and, when researchers become aware of a case, it is often well after a basic lexicon has conventionalized (R. Senghas, 1997). In fact, we are not aware of any studies observing conventionalization over time in emerging natural languages. We are only aware of studies of emerging systems that examine either inter-user consistency at a single point in time (e.g., Osugi, Supalla \& Webb, 1999), or intra-user consistency across a span of time (e.g., Goldin-Meadow, Butcher, Mylander \& Dodge, 1994). Showing images of objects and eliciting gestures for them, Osugi et al. (1999) investigated consistency in formmeaning mappings of lexical items among 21 deaf and hearing individuals in the geographically and genetically isolated Koniya region of Amami Island south of Japan. They show that individuals either Deaf or hearing were consistent with each other to the extent that they interacted. Goldin-Meadow et al. (1994) investigated the consistency over time of form-meaning mappings of gestures produced in a naturalistic context by a child homesigner called David
and his hearing mother. They found that David was more internally consistent than was his mother (and concluded that it was he who introduced into his system a noun-verb distinction, their primary object of interest).

In all, then, the two homesign studies, while shedding light on the outcome of conventionalization, reveal very little about the underlying process. The experimental research on conventionalization reviewed earlier, while suggestive, has not addressed conventionalization in natural linguistic settings. Computational modeling may provide explicit proposals of conventionalization mechanism, but it also suffers from the lack of connection with the empirical work. For instance, Barr (2004) investigated the effect of local vs. global information in conventionalization but the simulations are carried out on artificial data without making reference to experimental results or naturalistic case studies. The disconnect between experimental and computational approaches is a general concern for research on collective and cooperative behavior (see Goldstone \& Gureckis, 2009 for review).

In this paper, we take a step toward unifying empirical and computational work. We first, present new longitudinal data on conventionalization from naturally emerging homesign systems. We compare this data to preexisting non-longitudinal data on lexical consistency in Nicaraguan Sign Language (NSL), a natural sign language emerging in a vibrant Deaf community (Senghas \& Coppola, 2001; Senghas, 2003). We then present a general framework for studying conventionalization that incorporates elements of learning and social interactions. A specific implementation with reinforcement learning (Yang, 2002) appears to capture the observed trends of conventionalization. We conclude with a general discussion on the conditions for language emergence in a naturalistic setting.

## Homesign lexicons

In the present study, we examine conventionalization over a 9 -year period in form-meaning mappings for basic objects and concepts among deaf Nicaraguan homesigners and their family and friends.

## Method

Participants Participants were four deaf Nicaraguan homesigners [ 3 male; aged 11 to 33 years $(M=24)$ at various times of testing] and nine of their hearing family members and friends [4 male; aged 10 to $59(\mathrm{M}=30)$ at various times of testing; we henceforth refer to these family and friends as communication partners]. The homesigners have minimal or no interaction with other deaf individuals, including each other, and have minimal or no knowledge of Nicaraguan Sign Language or spoken or written Spanish, Instead, these homesigners have been using their respective invented gestural homesign systems all their lives. Despite their lack of linguistic input, they socialize with others, hold jobs, have families, and otherwise have typical lives. See Table 1
for relations between the homesigners and their family members.

Table 1: Homesigning groups

| Family 1 | Family 2 | Family 3 | Family 4 |
| :---: | :---: | :---: | :---: |
| Homesigner | Homesigner | Homesigner | Homesigner |
| Mother | Mother | Mother | Younger |
| brother |  |  |  |

Stimuli Stimuli were images of 22 basic objects and concepts. All items were familiar to participants. Nineteen of these objects and concepts were taken from Osugi et al. (1999), which itself was derived from Swadesh (1971). The stimulus items were: boy, cat, cold, cook, cow, dog, egg, fire, fish. flower, ice, girl, hot, moon, orange, palm tree, potato, rain, snake, stones, and sun.

Procedure In 2002, 2004, 2006, and 2011, M.C. showed participants images of the objects and concepts outlined above. Participants were tested individually. Using gesture and facial expressions, M.C. elicited participants' gestural responses to these images. Hearing participants were asked to use their hands to respond, and all were easily able to do the task. Participants responded to the camera, not to each other, and were not allowed to see each other's productions. All responses were videotaped for later analysis.

Coding Participants' responses were coded by a research assistant in consultation with R.R. A majority of responses contained more than one gesture (2 gestures: $40 \%$, 3 gestures: $15 \%, 4$ gestures: $4 \%$, and 5 gestures: $2 \%$ ), and so we coded every gesture individually for its Conceptual Component (CC), or aspect of the item's meaning that the gesture iconically represented. For example, a response to 'cow' might contain two gestures, one iconically representing horns (its CC is thus HORNS) and another iconically representing milking (its CC is thus MILKING) ${ }^{1}$.

## Results

Treating every CC as a dimension in a combinatorial space, every response can be represented as a binary-valued vector, with 1 representing the presence of a given CC and 0 the absence. The distance between two responses to the same object is thus a measure of conventionalization. We define distance here as the number of vector values by which two responses differ, and weight more heavily those vector values corresponding to CC's used more frequently (i.e. disagreement on the use of the CC ROUND will lead to a greater distance than disagreement on the infrequent CC

[^198]

Figure 1: Average distances, across objects tested, between a partner's lexicon and his/her homesigner's lexicon, per year. Partners converge with their respective homesigners.

MILKING) ${ }^{2}$. For a given object in a given year, we calculated this distance between each homesigner's response and that of each homesigner's communication partner's responses. For example, we calculate the distance between Homesigner 1's 2011 response to 'cow' and his mother's 2011 response to 'cow', as well as their 2006, 2004, and 2002 responses to 'cow'. For each homesignerpartner pair and year, we average these distances across all tested objects, yielding an overall measure of lexicon distance or conventionalization between a pair. Results are summarized in Figure 1 which shows decreases in lexicon distance across partners. To give a sense of the scale of weighted distance, consider a partner that with probability $P$ will agree with a homesigner in the usage of a CC. Simulations show that a partner agreeing with a homesigner $92.5 \%$ of the time gives a weighted distance of .069 , and agreeing $96 \%$ of the time gives a weighted distance of 0.036 $-\mathrm{a} \sim 50 \%$ reduction in error. This is roughly the change a typical communication partner (CP13) undergoes from 2002 to 2011.

We ran two tests to establish that (1) communication partners gradually converge with their respective homesigners, but that (2) even in 2011, convergence was not complete (where distance would be zero). To investigate our first question, we first extracted, for every partner, slopes of the linear regressions predicting homesigner-partner distance from year of testing. A one-tailed, one-sample Wilcoxon Signed Rank test on the nine slopes indicated that the median of this sample was significantly below 0 ( $\mathrm{W}=0$, $p<.01$ ), confirming the gradual convergence between homesigners and partners. To investigate our second question, we ran a series of one-tailed, one-sample

[^199]Wilcoxon Rank-Sum tests on the 2011 homesignercommunication partner distances. We found that these distances, despite decreasing over time, are still significantly greater than 0 ; all 9 of 9 such tests are highly significant (W's $\geq 91, p$ 's $\leq .001$ ).

## Discussion

We showed above that deaf homesigners slowly converge on form-meaning mappings with their hearing communication partners, but that convergence is not complete, even in 2011, the latest year in which we collected data. This contrasts sharply with the state of convergence in Nicaraguan Sign Language. The Deaf community in Managua, Nicaragua initially formed in 1978 (Polich, 2005), and by 1993 was holding 'standardization seminars' in smaller cities and towns outside the capital of Managua to spread the signs developed in Managua to the rest of the country (R. Senghas, 1997; López Gómez, Perez Castellón, Rivera Rostrán, \& Baltodano Baltodano, 1997). Thus, the NSL users in Managua must have converged on at least a basic lexicon in less than 15 years after coming together ${ }^{3}$. By 2011, all of the present homesigners had been using their respective systems for well more than 15 years, yet none of them had converged completely with any of their communication partners. What might explain this difference in rate of conventionalization between homesign and NSL? One possibility concerns the differences in patterns of interaction between users of homesign systems and users of NSL (and other Deaf community sign languages, Woll \& Ladd, 2003). While the deaf user of a homesign system uses the system for all interactions, the

[^200]hearing users only use the system to interact with that deaf user. In NSL and other deaf community sign languages, however, all users of the system interact with other users of the system using the system. In other words, the homesign interactive structure is one-to-many, while the NSL/deaf community structure is many-to-many. We now turn to our model, which replicates convergence, and allows us to test these predictions.

## Modeling Conventionalization

What are the conditions for conventionalization, whereby a shared lexicon emerges through strictly local linguistic interactions among linguistic individuals? At least two elements of process suggest themselves. First, the individuals must be "lexicon ready". In the simplest case, they must be able to maintain a list of form-meaning pairings. Similar to our study of homesigns, the individuals must be capable of making combinatorial use of constitutive units as in our case of Conceptual Components. Second, the individuals must be capable of learning, or modifying their lexicon as the result of linguistic and social interactions. In this section, we first describe a general framework to study lexical conventionalization. We then study its dynamics through the use of reinforcement learning (Bush \& Mosteller, 1951; Yang, 2002) as a model of learning and social interactions. Last, we use the model to test the hypothesis regarding the difference in conventionalization between homesign and NSL.

## The Framework

Consider a population of $N$ agents communicating a set of meanings through the combinatorial use of $C$ binary signs that are analogous to Conceptual Components in the homesign data. For a specific meaning, agent $i$ accesses a vector of probabilities $P_{\mathrm{c}}=\left\{p_{c}^{i}\right\}$, defined over these signs ( $j$ $=1,2, \ldots, C$ ) such that with probability $p_{c}^{i}$, the $c$ th sign is used by agent $i$ and with probability $\left(1-p_{c}^{i}\right)$, the $c$ th sign is not used. This representation can also be used to encode atomic use of signs, i.e., each meaning is expressed by one sign, in which case the vector $\sum_{c} p_{c}^{i}=1$ (i.e., agent $i$ has a probabilistic distribution of the signs and only one of them is chosen at each instance of use).

The central premise of the conventionalization model is that individuals adjust their choices of linguistic encoding in attunement with their communicative partners. To communicate a meaning, agent $i$ instantiates a vector $U_{i}$ of 0 's and 1's according to $P_{i}$. Agent $j$, the listener, generates a vector $U_{j}$ for that meaning according to its own $P_{j}$. (Note that the instantiations $U_{i / j}$ are not deterministic since the values are probabilistically chosen.) For each sign, agent $j$ compares $U_{j}$ against $U_{i}$ and makes adjustments to $P_{j}$ to agree with agent $i$ by the use of some learning algorithm. The changes in the distance between $P_{j}$ and $P_{i}$ over time represent the extent of convergence or conventionalization.

Linguistic communications among agents may also have a social component. Consider a matrix $S=\left[s_{i, j}\right]$, which defines the probabilities of communication between agents $i$ and $j$
such that $\forall i, \sum_{j} s_{i, j}=1$. The social matrix provides a general platform to encode patterns of interactions among agents. A matrix with positive probabilities only among the neighboring agents, for instance, is a straightforward implementation of Schelling (1971)'s classic model of segregation. The matrix may be fixed or it may change as the result of communication. For instance, it seems reasonable that agents would modify their partner preferences based on past successes or failures of communication, which can be modeled as $s_{i, j}$ increasing if a successful communication has occurred between agent $i$ and $j$ and decreasing upon failure.

As the result of the communicative interactions, the probability vectors for agents $\left\{P_{i}\right\}^{t}$ change over time, which characterizes the evolution of the lexicons in the population. In general, the dynamics of $\left\{P_{i}\right\}^{t}$ can be analyzed as a Markov Chain, first used by Berwick \& Niyogi (1997) to study language learning and change. Different choices of the learning algorithm ( $L$ ), which may be discrete or probabilistic (including Bayesian inference), the social matrix $S$ (and its own evolution), together with the current values in $\left\{P_{i}\right\}^{t}$ define the transition matrix $T^{t}$ at time $t$, which can be multiplied with $\left\{P_{i}\right\}^{t}$ to produce the next state of lexicon $\left\{P_{i}\right\}^{t+1}$. Similar models have been developed in the iterated learning framework (e.g., Kirby, Dowman \& Griffiths, 2007).

## Conventionalization through Reinforcement Learning

In what follows, we propose a specific learning model and consider several variant implementations relevant to the present study of sign convergence. The learning model is an instance of reinforcement learning (Bush \& Mosteller, 1951), a simple, efficient and domain general model of learning now with considerable behavioral and neurological support (see Niv, 2009 for review), and one which has been used in computational and empirical studies of language acquisition (Yang, 2002). Let agent $j$ 's current probability for sign $c$ be $p$. Upon each communication, the listener $j$ adjusts $p$ to match agent $i$ 's choices, following the Linear-Reward-Penalty ( $L_{R P}$ ) scheme of Bush \& Mosteller (1995) where the magnitude of change is a linear function of the current value of $p$ :

- Agent $i$ chooses 1: $p^{\prime}=p+\gamma(1-p)$
- Agent $i$ chooses $0: p^{\prime}=(1-\gamma) p$
where the learning rate $\gamma$ is typically a small real number. All probabilities are subsequently renormalized. Again, other models of learning can be studied in this fashion.

Social matrix: static vs dynamic We also consider the social communicative factors in conventionalization by manipulating the social matrix that defines the modes of individual interactions. As suggested above, we consider a case of adaptive social interactions where $\mathrm{s}_{\mathrm{i}, \mathrm{j}}$ increases if listener $j$ agrees with agent $i$ in all the choices of signs and decreases otherwise. The update rules for $S$ also follow the

LRP reinforcement learning scheme described above. Contrast this with static interactions where $\mathrm{s}_{\mathrm{i}, \mathrm{j}}$ 's remain constant.

Social matrix: homesign vs language An additional dimension of variation directly concerns the present study, for which we construct a homesign matrix in which one individual, the deaf signer (say agent 1 ), communicates with all other (hearing) individuals who do not use signs to communicate with each other. The matrix is initialized such that $\mathrm{s}_{\mathrm{i}, \mathrm{j}}=1 /(\mathrm{N}-1)$ where N is the total number of agents, $\mathrm{s}_{\mathrm{i}, 1}=1(\mathrm{i} \neq 1)$ and $\mathrm{s}_{\mathrm{i}, \mathrm{j}}=0(\mathrm{i}, \mathrm{j} \neq 1)$. We also consider what can be referred as the language matrix, where all agents are deaf and use signs to communicate with each other ( $\mathrm{s}_{\mathrm{i}, \mathrm{j}}=1$ / ( $\mathrm{N}-1$ ), $\mathrm{i} \neq \mathrm{j}$ ), which corresponds more closely to the sociolinguistic settings of typical sign language emergence (Woll \& Ladd, 2003). In all, we have four different modes of social interaction, that is, (home sign, sign language) $x$ (adaptive, static) and we explore their dynamical properties below.

Results In our simulations, we consider a population of $N=$ 5 agents. For each sign, we initialize the values in $P_{i}$ for each agent randomly between 0 and 1 ; they start out preferring either the use or the non-use of each sign with random probabilities. The learning rate $\gamma$ is set to 0.01 and is used for the adjustment of both $P_{i}$ 's and $S$, the social matrix that encodes the probabilities of inter-agent communications. For each simulation, we run the simulations over 2 million instances of communications; in the case of convergence, i.e., all $N$ agents in complete agreement with respect to sign usage (all $P_{i}$ 's at the value of 0 or 1 ), we record the number of iterations required for convergence. The main results are summarized in Table 2. Two things can be gleaned from these results: (1) there is no difference in convergence time between adaptive ( $p=0.412$ ) and static ( $p=0.435$ ) social structures and (2) there is a significant difference in convergence time between the homesign-type model and the language-type model ( $\mathrm{p}<10^{-12}$, for both social matrixes), indicating the importance of a mutually engaged community for the rapid emergence of a true linguistic system, and offering a potential explanation for the difference in rates of conventionalization between homesign and Nicaraguan Sign Language.

Table 2: Average number of iterations to convergence (percentage of simulations reaching convergence in 2 million iterations)

|  | Homesign | Language |
| :---: | :---: | :---: |
| Dynamic | $757 \mathrm{~K}(87 \%)$ | $281 \mathrm{~K}(100 \%)$ |
| Static | $698 \mathrm{~K}(80 \%)$ | $260 \mathrm{~K}(100 \%)$ |

## General Discussion

In the current work, we (1) presented longitudinal data showing conventionalization of lexicons among users of naturally emerging language-like systems (homesign gesture
systems); (2) showed that conventionalization in these homesign systems is slower than in Nicaraguan Sign Language (NSL), a recently emerging sign language used by a Deaf community; (3) formulated a general framework and causal model of conventionalization, in the form of a multiagent reinforcement learning model that obtains conventionalization; and (4) showed that an NSL-inspired model where all agents interact with each other converges significantly faster than a homesign-inspired model in which one agent (i.e. a deaf individual) interacts with every other agent (i.e. hearing individuals), but these other agents interact only with the first agent. We discuss implications our findings below, as well as open questions.

To the best of our knowledge, the present study is the first published observation of the lexicon, a fundamental component of language, emerging in natural human communication systems. Conventionalization has of course been obtained and studied numerous times in experimental settings (Galantucci et al., 2012), but our study is the first to connect the richness and complexity of real linguistic situations with well motivated models of learning. Surprisingly, variations in the dynamics of communications (the adaptive vs. static conditions in Table 1) led to little difference in the rate of convergence. The role of social/communicative factors in language emergence therefore deserves more careful consideration.

Our study is likewise, as far as we know, the first published paper to compare longitudinal or cross-sectional empirical data of naturally emerging languages to computational models of language emergence. As argued in the introduction, this synthesis is critical to a better understanding of language emergence. For example, many previous studies had established differences in linguistic complexity between homesign systems and natural sign languages (e.g., Coppola \& Senghas, 2010 regarding incorporation of deictic forms into syntax; Flaherty \& Senghas (2011) with respect to the existence of a count list), and had hypothesized about what differences between these systems' users affect language emergence (Senghas, 2005), but it has not been clear how exactly these differences influence language emergence. Our present data and model begin to answer this last question: more connected networks among users of the systems may accelerate conventionalization and language emergence.

Of course, alternative explanations of the different rates of conventionalization, and of complexity in general, in homesign systems and NSL do of course exist. For example, the hearing users of the homesign system have a spoken language to communicate with, and are thus under less pressure to use and conventionalize the homesign system. This contrasts with the situation faced by the deaf homesigner and users of NSL, who can only use their signed communication system and are thus behooved to conventionalize at a greater rate. Likewise, other learning models, e.g. Bayesian, can be studied in the general dynamic framework of language emergence. However, in the absence of more data to test the unique predictions of
different models, we opt here for one of the simpler possible models. We speculate that the general effects of network structure on conventionalization do not differ by class of model. These and other possibilities are not mutually exclusive and can be subject to future research. To identify a set of empirically motivated and verified conditions under which emergence takes place, or fails to do so (in a timely fashion), is an important first step toward to understanding the emergence of language.

## Acknowledgments

We would like to thank our deaf and hearing Nicaraguan participants and the members of the Coppola lab and the Sign Language Reading Group at the University of Connecticut. This research was supported by NSF IGERT grant \#1144399 to UConn and NIH grant P30 DC010751 to MC and Diane Lillo-Martin.

## References

Barr, D. J. (2004). Establishing conventional communication systems: Is common knowledge necessary? Cognitive Science, 28(6), 937-962.
Brentari, D., \& Coppola, M. (2012). What sign language creation teaches us about language. WIREs Cognitive Science. doi: 10.1002/wcs. 1212
Bush, R., \& Mosteller, F. (1951). A mathematical model for simple learning. Psychological Review, 68, 313-323.
Coppola, M. and Newport, E. L. (2005). Grammatical Subjects in home sign: Abstract linguistic structure in adult primary gesture systems without linguistic input. Proceedings of the National Academy of Sciences, 102(52), 19249-19253.
Coppola, M. and Senghas, A. (2010). The emergence of deixis in Nicaraguan signing. In D. Brentari (Ed.) Sign Languages: A Cambridge Language Survey. Cambridge, UK: Cambridge University Press.
Flaherty, M., \& Senghas, A. (2011). Numerosity and number signs in Deaf Nicaraguan adults. Cognition, 121(3), 427-436.
Galantucci, B., Garrod, S., and Roberts, G. (2012) Experimental Semiotics. Language and Linguistics Compass, 6(8), 477-493.
Goldin-Meadow, S., Butcher, C., Mylander, C. \& Dodge, M. (1994). Nouns and verbs in a self-styled gesture system: What's in a name? Cognitive Psychology, 27, 259-319.
Goldstone, R. L. \& Gureckis, T. M. (2009). Collective behavior. Topics in Cognitive Science, 1, 412-438. Hutchins, E. \& Hazlehurst, B. (1995) How to invent a lexicon: the development of shared symbols in interaction. In G. N. Gilbert and R. Conte, (Eds.), Artificial Societies: The computer simulation of social life. London: UCL Press.
Kirby, S., Dowman, M., \& Griffiths, T. L. (2007). Innateness and culture in the evolution of language. Proceedings of the National Academy of Sciences, 104(12), 5241-5245.

Kirby, S., Cornish, H., \& Smith, K. (2008). Cumulative cultural evolution in the laboratory: an experimental approach to the origins of structure in human language. Proceedings of the National Academy of Sciences, 105, 10681-10686.
López Gómez, J. J., Perez Castellón, A. M., Rivera Rostrán, J. M., \& Baltodano Baltodano, J. F. (Eds.) (1997). Diccionario del Idioma de Señas de Nicaragua. Managua, Nicaragua.
Meir, I., Sandler, W., Padden, C., \& Aronoff, M. (2010). Emerging sign languages. In M. Marschark and P. Spencer (Eds.) Oxford Handbook of Deaf Studies, Language, and Education, Vol. 2 (pp. 267-279). Oxford, UK: Oxford University Press.
Niv, Y. (2009). Reinforcement learning in the brain. The Journal of Mathematical Psychology, 53(3), 139-154.
Niyogi, P., \& Berwick, R. (1997). Evolutionary consequences of language learning. Linguistics and Philosophy, 20 (6), 697-719.
Osugi, Y., Supalla, T., \& Webb, R. (1999). The use of word elicitation to identify distinctive gestural systems on Amami Island. Sign Language \& Linguistics, (2)1, 87112.

Polich, L. (2005). The emergence of the deaf community in Nicaragua: With sign language you can learn so much. Washington, DC: Gallaudet University Press.
Schelling, T. C. (1971). Dynamic Models of Segregation. Journal of Mathematical Sociology, 1, 143-186.
Selten, R., \& Warglien, M. (2007). The emergence of simple languages in an experimental coordination game. Proceedings of the National Academy of Sciences, 104, 7361-7366.
Senghas, A. (2003). Intergenerational influence and ontogenetic development in the emergence of spatial grammar in Nicaraguan Sign Language. Cognitive Development, 18, 511-531.
Senghas, A. (2005). Language emergence: Clues from a new Bedouin sign language. Current Biology, 15(12), 463-465.
Senghas, R. J. (1997). An 'unspeakable, unwriteable' language: Deaf identity, language \& personhood among the first cohorts of Nicaraguan signers. (unpublished doctoral dissertation). University of Rochester, NY.
Senghas, A., and Coppola, M. (2001). Children creating language: How Nicaraguan Sign Language acquired a spatial grammar. Psychological Science, 12(4), 323-328.
Steels, L. \& Loetzsch, M. (2012). The grounded naming game. In L. Steels, (Ed.), Experiments in Cultural Language Evolution. Amsterdam: John Benjamins.
Swadesh, M. (1971). The origins and diversification of language. Sherzer, J. (Ed.), Chicago: Aldine.
Woll, B., \& Ladd, P. (2003). Deaf communities. In M. Marschark \& P. Spencer (Eds.), Oxford Handbook of Deaf Studies, Language, and Education, (pp. 151-162), Oxford: Oxford University Press.
Yang, C. (2002). Knowledge and learning in natural language. New York: Oxford University Press.

# A Bottom-up approach to the cultural evolution of bilingualism 

Seán G. Roberts (sean.roberts@mpi.nl)<br>Language and Cognition Department, Max Planck Institute for Psycholinguistics<br>Wundtlaan 1, Nijmegen 6525 XD The Netherlands


#### Abstract

The relationship between individual cognition and cultural phenomena at the society level can be transformed by cultural transmission (Kirby, Dowman, \& Griffiths, 2007). Top-down models of this process have typically assumed that individuals only adopt a single linguistic trait. Recent extensions include 'bilingual' agents, able to adopt multiple linguistic traits (Burkett \& Griffiths, 2010). However, bilingualism is more than variation within an individual: it involves the conditional use of variation with different interlocutors. That is, bilingualism is a property of a population that emerges from use. A bottom-up simulation is presented where learners are sensitive to the identity of other speakers. The simulation reveals that dynamic social structures are a key factor for the evolution of bilingualism in a population, a feature that was abstracted away in the top-down models. Top-down and bottom-up approaches may lead to different answers, but can work together to reveal and explore important features of the cultural transmission process. Keywords: Language; Cultural evolution; Bilingualism;


 Bottom-up.
## Introduction

Bilingualism is prevalent in many communities, but this is, intuitively, an unstable situation. What drives the emergence of bilingualism? Previous top-down computational models have emphasises the role of the learning biases of individuals such as a language's prestige (Abrams \& Strogatz, 2003), or expectations about variation in the input (Smith, 2009; Burkett \& Griffiths, 2010; Smith \& Thompson, 2012). However, these models assume that languages are monolithic, discrete and static. A bilingual is defined as an individual that adopts more than one linguistic variant from a set of discrete languages that has been defined a priori. This paper questions this assumption and explores how differences arise between the linguistic codes of communities in the first place and how they are subsequently adopted and maintained. In doing so, reveals that social factors, such as cultural identity and the dynamics of population structure, are crucial to the process of cultural evolution.

While low-level linguistic variants such as word order may plausibly have a discrete psychological reality (Diamond, 1991), 'languages' are more complex entities. Dividing continuous linguistic variation into categorical units or distinguishing 'languages' from 'dialects' is not straightforward and often involves complex notions such as politics, history, geography and identity (Sober, 1980; Haugen, 2009; Croft, in press; S. Roberts, 2012). The features of a language and the way it contrasts with others also changes over time. Therefore, this paper assumes that monolithic, static 'languages' (e.g. English, Welsh) have no psychological reality, but are emergent properties of populations and use. An abstract definition of bilingualism is presented which does not require
discrete categories: bilingualism is the amount of linguistic variation that is conditioned on social variables. That is, if I speak differently to Mary than to John, then I'm bilingual to some extent. This definition is compatible with the notion of audience design in sociolinguistics (Bell, 1984), and identifies bilingualism as a gradient property of interaction rather than a categorical feature that is identifiable in an isolated individual.

In order to explore the cultural evolution of bilingualism in this way, a bottom-up simulation is presented where learners are embedded in dynamic social structures and are sensitive to the identity of speakers. The top-down models assumed that social structures were static and focussed on cognitive explanations. In contrast, the bottom-up simulation will demonstrate that changes to social structures are an important factor. Rather than arguing that the bottom-up simulation is 'better' than the top-down model, this paper argues that different modelling tools tend to bias researchers towards making certain kinds of assumptions.

## Simulation definition

Bilingual cultural transmission is simulated as iterated stepwise linear regression. The representation of language is highly abstract, but allows the simulation of the emergence of bilingualism in a complex social structure. The linguistic space is a continuous one-dimensional space. A linguistic utterance is a point on this space (a real number). A meaning is represented as a point in a multi-dimensional meaning space. Each dimension of the meaning space represents a different semantic variable, such as properties of the environment that a linguistic utterance might be referring to (e.g. colour, number, size, tense etc.). A point in the meaning space, then, represents a particular combination of semantic elements (e.g. one big yellow thing). Each semantic variable has a set of hidden parameters which describe the distribution from which values are sampled. This systematic variation ensures that the linguistic signal has some structure to emulate. An important exception s a semantic variable that represents the identity of the speaker who described the event (speaker ID).

A 'learner' observes 'teachers' describing meanings with utterances and learns a mapping between the linguistic signal and the meaning space. The learning mechanism is a linear regression which results in an linguistic model that maps points in the meaning space to points in the linguistic space. The learner can then use this model to produce linguistic utterances itself.

Real languages exhibit flexibility with regards to which aspects of meaning condition linguistic variation. For example, in French, the form of a demonstrative (ce,cette,ces) depends


Figure 1: A diagram of how the bottom-up simulation works. Teachers produce linguistic and semantic data for a learner to observe. The learner uses stepwise regression to build a linguistic model, which is then used to describe some new semantic variables for the next generation.
on the grammatical gender of the referent and whether it is singular or plural, while the distance of the referent from the speaker is not important. In contrast, in Panjabi, the form of a demonstrative ( $i h, u h$ ) is conditioned on distance from the speaker, but not gender or number. In order to capture this flexibility, the linguistic model is selected by stepwise linear regression. A stepwise regression selects the minimum number of salient (semantic) variables that maximises the statistical fit, according to an information criterion. This allows the learner's linguistic model to prioritise or ignore different semantic variables in its linguistic utterances, including the identity of the speaker.

This learning process is iterated (Smith, Kirby, \& Brighton, 2003) in the following way. Learners observe a number of 'teachers' describing points in the meaning space with linguistic utterances, as above. The observations are affected by a small amount of noise. The probability of observing an utterance from a particular speaker depends on the structure of the community (see below). After the learners induce a linguistic model, the teachers are removed from the population. The learners become teachers for a new generation that is added to the population. This process repeats for many generations.

Learners have no explicit biases for particular 'languages', and the speaker ID variable is not privileged over other aspects of meaning. However, the learner is biased in a general way by the information criterion for the stepwise regression $I C$, which affects the number of variables an individual is willing to include in their linguistic model. The results presented below use the Akaike information criterion ( $I C=2$ ). In general, iterated stepwise linear regression has a bias towards shallow slopes and small intercepts. These do not affect the general results regarding bilingualism.

Population parameters Generations of individuals are separated by discrete timesteps $t_{1}, t_{2} \ldots t_{n}$. A population of $P_{t}$ learners in the current generation observe data produced by $P_{t-1}$ teachers in the previous generation. There are a number of communities in each generation and a set $C(t)$ of $P_{t}$ discrete labels represents which community each individual belongs to. A community interaction matrix $I(t)$ defines how much contact there is between each community. The probability $W(t)_{i, j}$ of learner $i$ receiving data from teacher $j$ is calculated as:

$$
\begin{equation*}
W(t)_{i, j}=\frac{I(t)_{C(t)_{i}, C(t-1)_{j}}}{S_{u} m_{W}} \tag{1}
\end{equation*}
$$

Where $\operatorname{Sum}_{W}$ is the sum of all weights between individuals. The community structure can therefore reflect situations from simple ones such as 'there are two communities' to a weighted, directed graph between individuals.

The community interaction matrix $I(t)$ can be simplified to a vector of single numbers by assuming that the probability of receiving data from any community that a learner does not belong to is equal.

$$
W(t)_{i, j}= \begin{cases}\frac{I(t)_{i}}{S_{u} m_{W}} & \text { if } C(t)_{i}=C(t-1)_{j}  \tag{2}\\ \frac{1-I(t)_{i}}{\text { Sum }_{W}} & \text { otherwise }\end{cases}
$$

This assumption will be adequate for the examples in this paper, and allows manipulation of the social structure through a single parameter for each community.

This framework allows different types of social structures. Given a situation where there are two teachers and two learners $P_{t-1}=P_{t}=2$ and two communities at each generation $C(t)=C(t-1)=\{A, B\}$, different settings of $I$ can then result in many social dynamics. Below I give some examples of matrices, with the learners (rows) labelled as $L_{1}$ and $L_{2}$ and the teachers (columns) labelled as $T_{1}$ and $T_{2}$. For example, a society with two communities that are completely integrated and balanced (effectively a single community):

$$
I(t)=\{0.5,0.5\} \longrightarrow \begin{array}{ccc} 
& T_{1} & T_{2}  \tag{3}\\
L_{1} & 0.5 & 0.5 \\
L_{2} & 0.5 & 0.5
\end{array}
$$

In the matrix above, for example, learner $1\left(L_{1}\right)$ has an equal probability of receiving data from either teacher. Alternatively, $L_{1}$ only receives data from $T_{1}$ and $L_{2}$ only receives data from $T_{2}$. This simulates two communities that are completely isolated:

$$
I(t)=\{1,1\} \longrightarrow \begin{array}{ccc} 
& T_{1} & T_{2}  \tag{4}\\
L_{1} & 1 & 0 \\
L_{2} & 0 & 1
\end{array}
$$

The prestige of a community can also be manipulated. Below is a matrix for a situation where one community only receives input from its own members (superstrate), while the other community receives input from both communities (substrate). It is predicted that this will lead to an analogue of a
minority language situation where everyone speaks one language (the majority language), and some speak a second language (the minority language).:

$$
I(t)=\{0.5,1\} \longrightarrow \begin{array}{ccc} 
& T_{1} & T_{2} \\
L_{1} & 0.5 & 0.5  \tag{5}\\
L_{2} & 0 & 1
\end{array}
$$

Measuring bilingualism Since 'languages' are not encoded in the simulation, the amount of bilingualism must be calculated from the bottom-up. This is defined as the amount of linguistic variation that is conditioned on social variables. In the simulation this is based on two measures of intelligibility, assuming that utterances are intelligible to speakers if their linguistic model would produce the same utterance given the same meaning (obverter assumption, similar to Oliphant \& Batali, 1997). The first is a measure of comprehensive intelligibility: the proportion of utterances that one speaker typically produces that another understands. For example, a monolingual speaker of English understands half of the utterances spoken by a balanced bilingual speaker of English and Welsh. In the simulation, this is a measure of the proportion of the variance in one learner's productions that is explained by another learner's linguistic model. If we're comparing individual $A$ and $B$, this is implemented in the following way (see figure 2):

1) Take identical samples of meanings $M_{A}$ and $M_{B}$
2) Sample speaker ID evenly in $M_{A}$ and $M_{B}$
3) Given $M_{A}$, produce utterances $U_{A}$ with $A$ 's linguistic model and given $M_{B}$, produce utterances $U_{B}$ with $B$ 's linguistic model
4) Calculate the correlation between $U_{A}$ and $U_{B}$

If two learners have the same linguistic model, then this correlation should be high. Individual A with a very different model from individual B will produce linguistic signals with a variation that is poorly explained by learner B's model, and so the correlation will be low.

We can also define a functional intelligibility score which is the proportion of utterances that interlocutors understand when they design their utterances for each other (figure 3). That is, a bilingual speaker of English and Welsh and a monolingual speaker of English could always make themselves understood by using English. In the simulation, this is calculated in a similar way to the comprehensive intelligibility score, except step (2) above is changed to

## 2) Set the speaker ID in $M_{A}$ to $B$ and in $M_{B}$ to $A$

In this case, an individual with a linguistic model that used speaker identity as a conditioning factor would adjust its variation to better suit its receiver (i.e. in the Welsh-English example, by speaking only English).

The two intelligibility measures can be combined to get a measure of bilingualism by subtracting the comprehensive intelligibility from the functional intelligibility. This can be
calculated for a population by taking the mean bilingualism score for all pairs of speakers.

$$
\begin{equation*}
B(t)=\frac{2}{n^{2}-n} \sum_{i=0}^{P_{t}-1} \sum_{j=i+1}^{P_{t}} \operatorname{Func}(i, j)-\operatorname{Comp}(i, j) \tag{6}
\end{equation*}
$$

Where $\operatorname{Func}(i, j)$ and $\operatorname{Comp}(i, j)$ calculate the functional and comprehensive intelligibility between two speakers.

If $B(t) \approx 0$, the comprehensive and functional intelligibility are, on average, similar. This means that everyone shares the same mapping between linguistic utterances and meanings what might be called the same linguistic code or 'medium' (Gafaranga, 2008). While this code might exhibit variation, this could be interpreted as a single 'language' (monolingualism).

If $B(t)<0$, the functional intelligibility between speakers is, on average, lower than the comprehensive intelligibility. For example, in the functional measure, speaker A would adapt their linguistic signal for speaker $B$ and $B$ would adapt their linguistic signal for speaker A . This yields a low functional similarity. However, their comprehensive similarity is high (their overall linguistic system is similar), so $B(t)$ is negative. This would be interpreted as bilingualism in the sense that each community is associated with a different mapping or 'code', and individuals can use each others' codes to some extent. A lower $B(t)$ means 'more' bilingualism in the lay sense. This is meant to represent the amount of linguistic variation that is conditioned on social variables, and so is analogous to an entropy-like measure where lower values indicate more order (the linguistic system is more conditioned on social factors $=$ bilingualism) and higher values indicate more disorder (the linguistic system is less conditioned by social factors $=$ monolingualism).
If $B(t)>0$, the comprehensive intelligibility score is lower than the functional intelligibility score. For example, A adapts their linguistic signal for B, but B does not adapt their linguistic signal for A . This leads to a high functional intelligibility, but a low comprehensive intelligibility. This means that both communities share one code, but one community has at least one other code. This might be interpreted as a minority situation in which one community speaks a minority language as well as the majority language. As we well see below, it's useful to be able to distinguish between 'balanced' communities $(B(t)<0)$ and 'minority' situations $(B(t)>0)$.

## Results

Since simulation is complex, basic findings are presented for simulations with 2 communities with 2 individuals each and 2 semantic variables, but the results scale up many semantic variables and hundreds of individuals. To summarise: unconditioned variation is unstable and bilingualism tracks social change. Figure 4 shows how $B$ changes in different social structures. When the two communities are completely integrated (integration parameter $I=0.5$ ), then they quickly converge to using the same linguistic code $(B=0)$. When the

Comprehensive intelligibility


Figure 2: The comprehensive intelligibility measure. Two individuals are given the same meanings and produce linguistic utterances with their linguistic models. The correlation between these utterances in the linguistic space is measured.


Figure 3: The functional intelligibility measure. Two individuals are given the same meanings, but the speaker ID is set to the other individual in the pair. They produce linguistic utterances with their linguistic models, and the correlation between the utterances in the linguistic space is measured.
two communities are partially isolated ( $I=0.8$ ), their varieties will take longer to converge and 'bilingualism' $(B<0)$ persists for some generations. The results are slightly different in a substrate/superstrate situation where learners from one community receive input equally from both communities (the minority, $I=0.5$ ), but the other community mainly receives input from speakers from its own community (the majority, $I=0.9999$ ). In this case, $B$ remains positive for many generations (a 'minority' language situation).

These results are for communities with static social structures. We can manipulate the social structure to demonstrate that linguistic diversity also tracks the change in social structures. Figure 5 shows the results of simulations with dynamic social structures. The communities go through cycles of being integrated, isolated, integrated and isolated again, with a few transition generations between each phase where the integration parameter is interpolated. As shown above, if two communities are integrated, they will come to speak effectively a single code ( $B \approx 0$, see figure 4 ). However, as the


Figure 4: The bilingualism score (means for 100 simulations) over generations for 2 communities of 2 individuals in three social structures: Integrated (circles), Isolated (triangles) and Minority (crosses).
communities become more isolated, $B$ increases. This is also in line with the results above. However, as the communities increase their interactions after this, $B$ decreases (everyone speaks a single code). Then we can split the communities apart and two codes will emerge again with some amount of bilingualism. That is, the distribution of linguistic variation tracks the changes in social structure.

More complex factors that affect $B$ were determined by analysing many runs of the simulation (analysis done using linear regression and stepwise linear regression). $B<0$ is inherently unstable in this simulation. As soon as individuals start mutually accommodating the linguistic signal of other communities, this neutralises the distinction over speaker ID. This is in line with the expectation that unconditioned linguistic variation is unstable (e.g. Smith \& Wonnacott, 2010).
$B<0$ is much more likely to emerge if speaker identity is the most important conditioning factor, while positive bilingualism scores can emerge if speaker identity is less important. Negative bilingualism is also more likely if individuals rank speaker identity in their models similarly. There are some more complex interactions. For example, $B<0$ tends to emerge when: the speaker ID is more important in the previous generation, except when communities are diverging, when it can be higher; when the community with the most complex linguistic model also considers speaker ID to be less important; when the mean and standard deviations of the speaker id rank are correlated; and when there is a stronger correlation between the difference in linguistic signal means and model fit ratio between communities.

Figure 5 shows that, after the first contact situation, only $B>0$ tends to emerge. This is partly due to the linguistic signal of two communities adapting to the same semantic distributions, and so becoming more alike. Situations where $B<0$ requires that there are large differences in the utterances of each community so that speaker ID conditions a large amount


Figure 5: How the integration parameter $I(t)$ (top, identical for both communities) affects the levels of bilingualism (bottom) in 2 communities with 2 learners each, which are alternatively integrated and semi-isolated (means over 100 runs with $95 \%$ confidence intervals).
of variation. When $B>0$ there is an imbalance in the extent to which different communities adapt to each other's linguistic signal.

It is possible to identify a 'superstrate' community as the one whose linguistic signal changes least between the generations of contact (as measured by the difference in a community's comprehensive intelligibility between generations). The difference in the linguistic utterance means between generations is the main determiner of the superstrate community. If community X's mean is absolutely greater than community Y's mean in the previous generation, then community X's linguistic models will change more than community Y. This affect arises due to the bias in the learning mechanism for small intercepts. However, this trend is only strong in the first generation of contact. During diverging generations, there is a $41 \%$ chance of a switch in superstrate community in the first two generations of divergence (from 100 simulations, significantly different from no switch: $\mathrm{t}=16.7, \mathrm{df}=399$, p -value $<0.001$, but also random switching: $\mathrm{t}=-3.55, \mathrm{df}=399, \mathrm{p}<$ 0.001 ). In contact situations, there is a $49 \%$ chance change of a switch in superstrate community in the first two generations (from 100 simulations, significantly different from no switch: $\mathrm{t}=13.8, \mathrm{df}=199, \mathrm{p}<0.001$; but not significantly different from random switching: $\mathrm{t}=-0.28, \mathrm{df}=199, \mathrm{p}=0.78$ ). In one generation a community might adapt to another, but this can cause the models in that community to better fit the data, leading to a pressure for the other community to adapt in the subsequent generation. Although a preliminary result, this may be compatible with phenomena such as 'mixed languages' where the emerging language in a contact situation uses the lexicon of one source language, but the grammar and
morphology of the other (e.g. Muysken, 1997). If lexical items and morphology take different amounts of time to learn (as suggested by Clahsen, Felser, Neubauer, Sato, \& Silva, 2010), then the 'mixing' might be partially due to this alternation in the community that adapts: the lexicon is taken first from one language, and later the morphology from another.

## Discussion

Dynamic social structures are a key aspect for explaining the emergence of bilingualism in this simulation. In the top-down models, social structures were static and so they could not form a part of the explanation. The bottom-up simulation can be more flexible because it doesn't require learners to be fully rational or optimal, as opposed to some Bayesian models.

The linguistic contrast between communities will diminish if there is no contrast in the social variables. However, it does not mean than bilingualism in the lay sense is unstable. Firstly, $B$ is not necessarily an index of an intuitive idea of bilingualism. Communities like those in Catalonia might actually have $B \approx 0$, because many people speak both Catalan and Spanish. Secondly, in the real world, linguistic variation might be dictated by social factors not simulated here, such as location, formality or stage of the conversation (e.g. Labov, 1963; Meyerhoff, 2008). Finally, this simulation includes no pressures to maintain a linguistic identity such as prestige, politics or resistance to freeriders (G. Roberts, 2010). Rather, it shows that bilingualism that can emerge just from the process of cultural transmission - a kind of baseline behaviour on top of which more complex factors may be applied.

The top-down models specified a prior bias over the amount of variation to expect in an agent's input, fitting the learning mechanism to the problem being addressed. In contrast, bilingualism emerges in the bottom-up simulation without individuals having a specific mechanism for dealing with bilingualism. All that is required is a general learning mechanism which conditions a linguistic signal on semantic variables. There are no expectations over the amount of variation to expect within or between speakers. Indeed, if social variables do not explain any of the variance, they do not play any role in an agent's linguistic internal representation.

Furthermore, the simulation maintains a division between population level phenomena and individual learning mechanisms: 'bilingualism' can emerge at the population level without discrete, static languages being encoded in the linguistic model of individuals. This suggests that that 'bilingualism' is a property of populations which is not necessarily related to specific individual learning biases. That is, whether humans have an expectation about the number of languages that will be in their input, or whether learning two languages is more difficult than learning one are not necessarily the most relevant questions. Rather, one should ask how contrasts in social variables support the maintenance of linguistic variation.

## Conclusion

The simulation works as a proof-of-concept for the abstract definition of bilingualism. Bilingualism is measurable in this simulation without encoding a discrete, monolithic, static concept of a 'language'. The measure behaves as we would expect in integrated, isolated and substrate/superstrate population structures. The results suggest that dynamic social structures are an important part of explaining the cultural evolution of bilingualism. This differs from the conclusions of top-down models, demonstrating that different kinds of models may be biased towards certain conclusions.

Top-down models considered the stability of bilingualism given assumptions about individual learning (Burkett \& Griffiths, 2010) and the most likely expectation for individuals to have about the number of languages in their input (Smith \& Thompson, 2012). These might suggest research directions such as estimating the expectations human learners have about the number of languages to expect in their input, the amount of noise in transmission or whether the social structure was one that caused bottlenecks on learning. However, in the bottom-up simulation, because bilingualism tracks social change, asking whether individuals should expect many languages in their input does not make sense without also thinking about dynamic social structures. This suggests that the questions asked by the top-down model are misleading. The bottom-up simulation suggests researching dynamic social structures, and how linguistic variation, social structures and learning biases coevolve.

Both the top-down and bottom-up models are very abstract, and it would be a difficult to determine which was more 'realistic' or fitted real data better. Instead, both approaches can be seen as converging on a common solution to the problem from different angles. The top-down model is better at yielding good analytic results, but the bottom-up model allows more flexibility in terms of social dynamics. The bottom-up simulation presented here has suggested that some of the assumptions of the top-down models require more scrutiny. In response, a top-down model could be built which addressed the most relevant points raised by the bottom-up simulation perhaps using techniques such as variational Bayesian analysis (e.g. Kurihara \& Sato, 2006). This process of exploring results and uncovering important assumptions using mutually supporting approaches can lead to more robust theories.

## Acknowledgments

Supported by the ESRC [ES/G010277/1]. Thanks to Simon Kirby, Kenny Smith and Antonella Sorace for comments.

## References

Abrams, D. M., \& Strogatz, S. H. (2003). Modelling the dynamics of language death. Nature, 424, 900.
Bell, A. (1984). Language style as audience design. Language in society, 13(2), 145-204.
Burkett, D., \& Griffiths, T. (2010). Iterated learning of multiple languages from multiple teachers. In A. Smith,
M. Schouwstra, B. de Boer, \& K. Smith (Eds.), The evolution of language: Proceedings of evolang 2010 (p. 58-65).
Clahsen, H., Felser, C., Neubauer, K., Sato, M., \& Silva, R. (2010). Morphological structure in native and nonnative language processing. Language Learning, 60(1), 21-43.
Croft, W. (in press). Language use and the evolution of languages. In K. Smith \& P. Binder (Eds.), The language phenomenon. Berlin: Springer.
Diamond, J. (1991). The rise and fall of the third chimpanzee. Radius, London.
Gafaranga, J. (2008). Talk in two languages. Palgrave Macmillan.
Haugen, E. (2009). Dialect, language, nation. American anthropologist, 68(4), 922-935.
Kirby, S., Dowman, M., \& Griffiths, T. L. (2007). Innateness and culture in the evolution of language. Proceedings of the National Academy of Sciences, 104(12), 5241-5245.
Kurihara, K., \& Sato, T. (2006). Variational bayesian grammar induction for natural language. Grammatical Inference: Algorithms and Applications, 84-96.
Labov, W. (1963). The social motivation of a sound change. Word, 19, 273-309.
Meyerhoff, M. (2008). Bequia sweet/bequia is sweet: syntactic variation in a lesser-known variety of caribbean english. English Today, 24(01), 33-40.
Muysken, P. (1997). Media lengua. Contact languages: a wider perspective, 17, 365.
Oliphant, M., \& Batali, J. (1997). Learning and the emergence of coordinated communication. Center for research on language newsletter, 11(1), 1-46.
Roberts, G. (2010). An experimental study of social selection and frequency of interaction in linguistic diversity. Interaction Studies, 11(1), 138-59.
Roberts, S. (2012). Evolutionary approaches to bilingualism. Unpublished doctoral dissertation, University of Edinburgh.
Smith, K. (2009). Iterated learning in populations of Bayesian agents. In N. Taatgen \& H. van Rijn (Eds.), Proceedings of the 31th Annual Conference of the Cognitive Science Society. Cognitive Science Society.
Smith, K., Kirby, S., \& Brighton, H. (2003). Iterated learning: a framework for the emergence of language. Artificial Life, 9(4), 371-386.
Smith, K., \& Thompson, B. (2012). Iterated learning in populations: Learning and evolving expectations about linguistic homogeneity. In T. C. Scott-Phillips, M. Tamariz, E. A. Cartmill, \& J. R. Hurford (Eds.), The Evolution of Language: Proceedings of the 9th International Conference (EVOLANG9) (p. 227-233). World Scientific.
Smith, K., \& Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. Cognition, 116(3), 444-449.
Sober, E. (1980). Language and psychological reality: some reflections on chomsky's rules and representations. Linguistics and Philosophy, 3(3), 395-405.

# Integration and Inference: Cross-situational Word Learning Involves More than Simple Co-occurrences 

Alexa R. Romberg and Chen Yu<br>\{aromberg, chenyu\}@indiana.edu<br>Department of Psychological and Brain Sciences, 1101 E. 10th Street<br>Bloomington, IN 47405 USA


#### Abstract

Statistical word learning involves forming and aggregating associations between words and objects that co-occur across contexts (e.g., Vouloumanos \& Werker, 2009; Smith \& Yu, 2008; Yu \& Smith, 2007). However, the mechanisms that support such learning are currently under debate, including the extent to which learners carry forward multiple ambiguous associations (e.g., Trueswell et al., 2013). The current study presented adults with a set of statistical word learning tasks designed to measure the statistical computations learners employ to build label-object mappings and to probe what information from past contexts is available to further process and integrate with new information. Results reveal that learners use the co-occurrence of label-object pairings to make inferences both about objects and labels currently present and those presented on previous trials. Further, the strength of learners' memory for past contexts moderated their inferences, suggesting a role for a rich information structure in cross-situational word learning.


Keywords: word learning; statistical learning; language acquisition; cross-situational learning

## Introduction

Imagine an infant on a walk with his father. The father, like many parents, comments on the things they see together: "There's a doggie and a kitty in the window!" and a few moments later: "Look, the man is walking the doggie!". How might the father's comments help the infant learn the meanings of words like doggie, kitty and man? Recent research has demonstrated that learners readily form label-object mappings by gathering co-occurrence statistics. Human infants (Smith \& Yu, 2008; Vouloumanos \& Werker, 2009), children (Scott \& Fisher, 2011) and adults (Kachergis, Yu \& Shiffrin, 2012; Suanda \& Namy, 2012; Yu \& Smith, 2007) are all capable of converting multiple individually ambiguous learning instances into specific knowledge as demonstrated by above-chance performance on a post-learning test or by an improvement in selection of the correct referent in a combined training and test procedure (Trueswell, Medina, Hafri \& Gleitman, 2013). However, the precise ways in which learners resolve the local ambiguities have been relatively unexplored.

Specifying how exactly learners use the information available is an important step to understanding the mechanisms contributing to success. When learners perform some computations but not others, this offers important constraints to any model of their learning and can inform discussion about the nature of the information stored. In the context of cross-situational word learning, two primary mechanisms, associative learning and hypothesis testing,
have been proposed for how learners accrue information over time. These mechanisms differ largely in the amount of information stored and, consequently, in how prior information influences later learning (Yu \& Smith, 2012). In particular, associative learning proposes that learners form multiple associations between the objects and labels present during each learning instance, storing a relatively rich information network. Hypothesis testing proposes that learners store only a single link between a label and possible referent, discarding other co-occurrence information. Distinguishing these possible mechanisms has been challenging thus far because of a lack of data regarding how learners process information on a trial-by-trial basis.

Details about what information learners store and how they use it during cross-situational word learning is vital for advancing theories of this process. The typical crosssituational word learning experiment uses a fairly large novel vocabulary (up to 18 to-be-learned label-object mappings) and consists of a series of trials that each present a subset of the labels and objects. Thus, the learner is faced with the difficult task of tracking these many labels and objects across trials (typically between 27 and 60 trials) and using what co-occurrences they can glean to generate as many correct mappings as possible. While this experimental design is daunting for the participant, it is also daunting for the experimenter, as there are inevitably many possible paths to success. One cannot know definitively how participants arrived at a particular mapping over the course of statistical learning or whether the same types of computations were used for all learned mappings.

The present study sought to alleviate these analytical ambiguities for the experimenter while maintaining the learning ambiguities for the participants. Rather than have participants view many trials across which to learn many mappings, learners were presented with a series of "miniature" cross-situational word learning tasks. These tasks consisted of only 2 or 3 trials and were constructed so that some, though not all, label-object mappings could (theoretically) be disambiguated, depending on which information learners stored and which inferences they made. The miniature tasks were constrained so that there was only one pathway to disambiguation, allowing us to infer the computations successful learners employed.

We focused on three fundamental processes that could serve as building blocks for sophisticated statistical learning. The first was the tracking of co-occurrence information - noticing that some labels and some objects appear together across multiple trials. The simplest of
statistical learning models, such as the "dumb associative model" outlined by Yu \& Smith (2012) do only this cooccurrence tracking to fill an association matrix. The second process was "forward integration", by which learners use information that they carry forward across trials about some objects and labels to make an inference about another object-label mapping. Mutual exclusivity (Markman, 1990) would be a strong form of forward integration, when one rules out objects with known labels as possible referents for a novel label. Recent evidence indicates that learners do employ mutual exclusivity during cross-situational word learning and that this type of inference could arise through basic attentional processes (Kachergis, Yu \& Shiffrin, 2012). The third process was "backward inference," by which learners use information on the current trial to infer something about an object/label experienced on a previous trial (but not on the present trial). This last process can be thought of as learning from negative evidence, as it entails noting the absence of particular objects and labels.

We compare performance on three different "miniature" cross-situational word learning tasks to assess learners' ability to use the available information in the three processes of co-occurrence tracking, forward integration and backward inference. The tasks were designed to look specifically at how trial-by-trial information is retained and processed. We also relate performance on the miniature tasks to a "full" cross-situational word learning task, to investigate whether these fundamental processes are also employed in larger-scale statistical word learning.

## Method

## Participants

Participants were 38 undergraduates (20 females) at Indiana University who earned course credit for their participation. The mean age was 20.9 years.

## Materials

The auditory stimuli consisted of 108 nonce words synthesized with the Ivona voice Jennifer using the TextSpeaker program. Nonce words consisted of one or two syllables ( 264 ms to 795 ms in duration) and followed English phonotactics. The visual stimuli were 123 color photographs of real objects or 3D models either of novel objects or objects that were not readily nameable. Images were displayed in the 4 corners of a 17 " monitor, on a white background at a size of approximately 3 " square.

## Experimental Design

There were 3 types of "mini-tasks" (see Figure 1), each made up of 2-3 training trials and then 3-4 test trials. Each mini-task was independent of the others and no stimuli were repeated across tasks. When objects and labels were repeated on multiple trials they were always presented in different spatial or temporal positions.

Across all parts of the experiment, each trial consisted of viewing 4 objects on a screen and listening to 4 nonce
words. Each trial began with the objects displayed in silence for 3 seconds. The onsets of the words were 3 seconds apart and the total trial length was 15 seconds. Every time an object was on screen the corresponding label was provided.

The training trial structure of each of the 3 types of minitasks is given in Figure 1. For all tasks, the R items refer to the object-label pairs $R_{1}, R_{2}$ and $R_{3}$, which were presented on Trials 1 and 2. T1 refers to the object-label pair presented on Trial 1 but not Trial 2. T2 refers to the object-label pair presented on Trial 2 but not Trial 1. The Base and Familiar Context tasks each consisted of a total of 5 word-object pairings: $\mathrm{R}_{1-3}$, T 1 and T2. The Novel Context task consisted of a total of 8 word-object pairings: $\mathrm{R}_{1-3}, \mathrm{~T} 1$ and T 2 and the 3 novel, label-object pairs presented only on Trial 3 ( $\mathrm{N}_{1-3}$ ).


Figure 1. Schematic representing the training trial structure for the 3 mini-tasks, with letters representing objects filling different roles in the design.

For each of the mini-tasks, the training trials were followed immediately by a series of test trials. On each test trial one word was presented auditorally and participants were instructed to click on the object the word most likely referred to out of all objects presented on the task plus a novel distracter object. For the Base and Familiar Context tasks, participants selected from 6 objects and for the Novel Context task participants selected from 9 objects. The tested words came from the different categories of items in the task (R, T1, T2 and, for Novel Context only, N). While the tested items aligned structurally across tasks, the information available to participants differed, enabling us to test hypotheses about what information participants track and what inferences they make.

All participants also completed a "full" cross-situational word learning task, based on Yu \& Smith, 2007, which consisted of 18 label-object pairings. These were presented 4 at a time across 27 training trials, so that each label cooccurred with its referent object 6 times. With the $4 x 4$ design, objects co-occurred with other labels, but such "spurious" correlations were limited to no more than 3 times across the 27 trials. Training was followed immediately by 18 test trials. On each test trial all 18 objects were displayed, one auditory label was presented and participants selected the best referent by mouse click.

## Procedure

Participants were given an overview of the experiment and informed consent was obtained. All participants first completed the Full CSL task. They were told that they there were 18 words and 18 objects, that they would see them 4 at a time and that the order of the labels on any trial did not correspond to the spatial location of the objects. They were instructed to learn as many label-object mappings as they could. Once participants completed the test for the Full task, they moved on immediately to the mini-tasks. There were a total of 15 mini-tasks, 5 of each task type. The tasks were grouped so that there was one of each type in each block of 3. The order of the 15 tasks was the same across all participants but the order of the test questions within each task were randomly determined for each participant. In the instructions for the mini-tasks, participants were told they would see a series of 15 tasks that were miniature versions of what they had just done and that they would be tested after just 2 or 3 training trials. They were told that no objects or words would be repeated across the mini-tasks. Participants were tested one at a time and listened to the auditory stimuli over headphones. The entire experiment took approximately 30 minutes.

## Predictions

The Base task provides a baseline measure of each of the three processes we are examining: The R items represent the co-occurrence tracking process. For each of the tasks, the precise object-label mappings within this group remain ambiguous. However, successfully tracking the repetition of this group of objects and labels theoretically enables learners to perform two types of inference to disambiguate the T1 and T2 mappings. The T2 items represent forward integration: whether participants can use the familiarity of the $R_{1}, R_{2}$ and $R_{3}$ pairs on Trial 2 to make a mapping between the relatively novel T2 label and object. Finally, the T1 items represent backward integration: whether participants use the absence of T 1 on Trial 2 to make a mapping between that label and object.

Backward integration relies on participants remembering the T1 pair across multiple ambiguous trials and was expected to be difficult. Thus, the other two mini-tasks were designed to test participants' memory for T1 by presenting it in either a novel or familiar context. This necessarily changes the interpretation of the T1 pair in the Novel and Familiar Context tasks, as participants no longer need rely solely on backward integration to learn the mapping.

In the Novel Context task, T1 is presented with 3 new objects and labels on the $3^{\text {rd }}$ trial. This task is the only task in which the association matrix distinguishes the T1 mapping, enabling a correct mapping if participants recognize T 1 from the first trial. It is also possible that learners could employ forward integration, mapping the familiar-looking object to the familiar-sounding label without any memory specifically linking the two. If, however, participants do not retain any memory of T 1 , they
should choose randomly from T1 and $\mathrm{N}_{1}, \mathrm{~N}_{2}$ and $\mathrm{N}_{3}$ on both the T1 and the N test trials.

Unlike the Novel Context task, the Familiar Context task does not provide any additional statistical certainty relative to the Base task. While participants get an additional T1 pairing, it occurs with the same items on both trials. However, participants could infer the T1 mapping by using forward integration in the same manner as the T2 item on Trial 2. Comparing performance between the Base and Familiar Context tasks provides further insight into how learners track information. In the most straightforward extrapolation from the Base task, accuracy on R and T 1 should improve due to the extra trial and accuracy on T 2 should decrease due to the extra trial between when T 2 is presented and tested. Further, within the Familiar Context task, accuracy on T1 is expected to be higher than T2, as participants can use the same process to infer them and T2 is presented on the last trial of the experiment.

## Results

All objects from the ambiguous groups ( $\mathrm{R}_{1-3}$ and $\mathrm{N}_{1-3}$ ) were scored as correct. The baseline for chance performance varied between test items and between tasks. For the Base task and Familiar Context task participants selected from 6 objects, so chance performance was $50 \%$ for R test trials and $16.7 \%$ for T1 and T2 test trials. For the Novel Context task, participants selected from 9 objects, so chance performance was $33.3 \%$ and $11.1 \%$, respectively. Statistical comparisons between trial types and tasks were performed with logistic mixed-effects models with random effects of subject (other random effect structures were tested but in no case improved model fit).

## Forward integration and backward inference

We first address performance on individual mini-tasks before turning to relationships between the mini-task and full task and comparisons between mini-tasks. Mean accuracy for each type of test item is shown in Figure 2. The results from the Base task reveal that learners do engage the three processes it was designed to test: co-occurrence tracking, forward integration and backward inference. Each of the three trial types has accuracy significantly above chance performance (see confidence intervals on figure). While forward integration accuracy was quite high, our prediction that backward inference would be relatively challenging was confirmed, with participants performing significantly better on T 2 items than T 1 items on the Base task ( $b=1.898, z=7.88, p<0.001$ ).

Results from the Novel Context and Familiar Context tasks point to the robustness of co-occurrence tracking and forward integration. In the Novel Context task, neither R items nor T 2 were presented in the final trial, so learners must maintain that information while concurrently learning about additional objects and labels. Despite this challenge, participants were significantly above chance on both R and T2 items for the Novel Context task (see Figure 2). In the Familiar Context task, T2 information must be maintained
while familiar objects and labels from Trial 1 are repeated in Trial 3. Again, learners were quite successful, performing significantly above chance. Surprisingly, there was no decrement in performance for T2 from the Base to the Familiar Context task (see further discussion below).

While accuracy was significantly above chance for backward integration T1 items on the Base task, it was not very high. Backward integration relies on memory for the T1 pair, as the inference must be made in the absence of the object and label. Remembering the T1 label and object may pose a particular challenge since the mapping between them is ambiguous when they are first presented; it is possible that rapid decay of this information is responsible for the relatively low performance on backwards inference.

However, results from the Novel Context task demonstrate that participants recognized the T1 pair as familiar on Trial 3 and distinguished it from $\mathrm{N}_{1-3}$. Indeed, accuracy on T1 is numerically much higher than the Base and Familiar Context tasks even though the chance baseline is lower. The error pattern also suggests that participants were not likely to confuse T 1 and $\mathrm{N}_{1-3}$. On N trials, participants selected T1 only $4.7 \%$ of the time, less than they selected T2 (10\%), which did not co-occur with the N group. On T1 trials, participants selected one of $\mathrm{N}_{1-3} 18.4 \%$ of the time, less than is expected for random guessing (33.3\%) and much less than they selected T1 (65.3\%).


Figure 2. Mean percentage correct for each of the three mini-tasks. The error bars are $95 \%$ confidence intervals of the mean. The horizontal lines within each bar represent chance performance for that test item.

Are forward integration and backward inference relevant for statistical word learning beyond the mini-tasks? While the mini-tasks used in our experiment are structured similarly to the design of the larger cross-situational word learning paradigm employed in previous research (e.g., Yu \& Smith, 2007; Yu, Zhong \& Fricker, 2012) the mini-tasks had much more trial-to-trial overlap than other crosssituational learning paradigms. Thus, it is possible that learners don't rely on these inferential computations in the larger task, but simply accumulate co-occurrence statistics.

In order to verify that forward integration and backward inference were relevant for cross-situational word learning in a larger set, we correlated participants' scores on the Base
task with their scores on the Full CSL task (see Figure 3). The proportion of correct object-label mappings was positively correlated for the Base mini-task and Full CSL task ( $r=0.485, p=0.002$ ), suggesting that these tasks tapped similar skills. We also tested correlations between participants' accuracy on the Full CSL and on each of the individual trial types in the Base task to investigate the role of the individual computations. Positive correlations were found for both backward inference (T1) items ( $r=0.407$, $p=0.011$ ) and forward integration (T2) items ( $r=0.387$, $p=0.016$ ). The relationship between accuracy on the cooccurrence tracking (R) items and Full CSL accuracy was marginally significant ( $r=0.303, p=0.065$ ). Accuracy on R items was in general quite high and this measure of cooccurrence tracking may not have been sensitive enough to detect a significant relationship. However, as described above, the tracking of co-occurrence information is necessary for the other two computations. Together, these results strongly suggest that forward integration and backward inference are processes integral to crosssituational learning.


Figure 3. Scatterplot depicting the correlation between percentage correct on the Base Mini-Task (horizontal axis) and the Full Task (vertical axis). Values have been jittered so that all data may be seen. Overall performance on the Base Task is shown in the top left panel, R test trials in top right, T 1 test trials in bottom left and T2 test trials in bottom right. Lines represent linear best fit.

## The role of tracking multiple co-occurrences

We now turn to comparisons between the mini-tasks to further explore the computations learners employed to infer label-object mappings and participants' memory for ambiguous prior information. The Familiar Context task repeated the information from Trial 1 on Trial 3. We predicted that this repetition would lead to higher accuracy
relative to the Base task for R and T 1 items and lower accuracy for T 2 items (because of the addition of Trial 3 between training on T 2 in Trial 2 and test). Of these predictions, only the improvement on T1 is confirmed by the data. Accuracy on R items was not significantly different between the two tasks ( $p>0.29$ ), suggesting that the additional co-occurrence information did not lead to better mappings.

Changes in performance on T1 and T2 items were tested with a logistic mixed-effect model with Task (Base or Familiar Context) and Trial Type (T1 or T2) as fixed effects and Subject as a random effect. This revealed a significant Task by Trial Type interaction ( $b=-1.00, z=3.13, p=0.002$ ). However, follow-up analyses confirmed that the pattern of effects was not as predicted. While accuracy on T1 was higher for the Familiar Context task than the Base task ( $b=0.631, z=2.94, p=0.003$ ), accuracy on T2 was not different across the two tasks ( $b=-0.347, z=1.48, p=0.14$ ). Accuracy was significantly higher for T 2 than T 1 for both tasks (Base: $b=1.898, z=7.88, p<0.001$; Familiar Context: $b=0.785, z=3.64, p<0.001$ ). Surprisingly, even though participants could use the same forward integration process to infer the T1 and T2 mappings and the T1 object and label were presented on the last training trial while T2 had to be maintained across this trial, participants were more accurate on T2 than T1.
Why might the results differ so much from our initial prediction? One possibility is that, for some participants, rather than increasing certainty, the repetition of the Trial 1 information on Trial 3 actually increased spurious correlations, and therefore the confusability, between $\mathrm{R}_{1-3}$ and T1. The pattern of errors across tasks supports this interpretation. On R trials participants were significantly more likely to incorrectly select T1 for the Familiar Context task than the Base task ( $b=1.153, z=2.92, p=0.004$ ). Selection of an R item on T 1 test trials was equivalent across the two tasks, despite the improved performance on T1 for the Familiar Context task ( $b=-0.282, z=1.06, p=0.29$ ).

The participants with the best memory of the first trial should be most likely to confuse $\mathrm{R}_{1-3}$ and T 1 . The nature of the backward inference requires memory for T 1 , as well as memory for the context in which T1 occurred (i.e., the other objects and labels), since it is the absence of T1 from this context on Trial 2 that allows the inference. If memory of Trial 1 increases confusability between $\mathrm{R}_{1-3}$ and T 1 , participants who were successful on backward inference in the Base task should improve less on T1 items for the Familiar Context task than participants who were not successful on backward inference.

Participants were split into two groups at the median for backward inference on the Base mini-task ( $\mathrm{N}=19$ in each group). Participants with $20 \%$ or less correct on T1 items for the Base task were labeled low-backward inference (low-BI, $M=0.116, S D=0.322$ ) and those with more than $20 \%$ correct were labeled high-backward inference (high-BI, $M=0.537$, $S D=0.501$ ). A logistic mixed-effect model predicting accuracy on T1 test items with Task (Base or Familiar

Context) and BI (high or low) as fixed factors and Subject as a random factor revealed a significant Task by BI interaction ( $b=-2.346, z=4.86, p<.001$ ). The low-BI group had significantly higher accuracy on the Familiar Context task ( $M=0.484, S D=0.502$ ) than the Base task ( $b=1.973$, $z=5.18, p<0.001$ ), while the high-BI group did not ( $M=0.453, S D=0.500 ; b=-0.349, z=1.19, p=0.23$ ). Thus, participants with a weak memory of T1 from the first trial benefitted from the repetition of that information on Trial 3, while participants with a strong memory did not. This is further confirmed by the pattern of accuracy for R items across the two tasks, which was also subject to a Task by BI-group interaction ( $b=-1.122, z=2.01, p=0.044$ ). The lowBI group improved slightly from the Base task to the Familiar Context (Base $M=0.768, S D=0.424$; FC $M=0.80$, $S D=0.402$ ) while the high-BI group actually declined (Base $M=0.905, \quad S D=0.294 ; \quad$ Familiar Context $\quad M=0.789$, $S D=0.410$ ), suggesting that the Trial 1 repetition disrupted their memory for the R items.
In contrast to the Familiar Context task, performance on the Novel Context task should only be aided by improved memory for T1 (and its first-trial context). Both high-BI and low-BI participants benefitted from the presentation of T1 in a novel context with high accuracy on T1 items (high-BI $M=0.726, S D=0.448$; low-BI $M=0.579, S D=0.496$ ).
While a better memory for context impeded performance on the Familiar Context mini-task, such memory should generally improve statistical word learning, as learners would have a more complete association matrix on which to build. We tested the role of contextual memory in crosssituational word learning by comparing performance in the Full CSL task for the high- and low-BI groups. As predicted, high-BI participants were significantly more accurate in the Full CSL task than low-BI participants (highBI $M=0.444, S D=0.192$; low=BI $M=0.254, S D=0.173$, $t(36)=3.2, p=0.003)$.

## Discussion

The present study investigated three fundamental processes that may contribute to cross-situational word learning. We found that learners readily tracked cooccurrence information trial by trial and used those cooccurrence statistics to infer label-object mappings in new learning situations, a process we termed forward integration. We also found that learners inferred label-object mappings when the disambiguating evidence was the absence of the label and object on trials on which they would otherwise be expected, a process we termed backward inference. Further, we found that participants retained multiple co-occurrences between objects and labels presented on previous trials. Importantly, participants who best remembered multiple object-label co-occurrences within a learning trial were most successful at cross-situational word learning.
Our results support the argument that cross-situational word learning involves learning a system of label-object mappings, in which learning about one set of items influences knowledge about other items. From a
straightforward co-occurrence information point of view, the T1 and T2 objects are not more strongly associated with the T1 or T2 labels than the $\mathrm{R}_{1}, \mathrm{R}_{2}$ and $\mathrm{R}_{3}$ labels for either the Base or Familiar Context task. In order to disambiguate these mappings participants must use the information available not just about the T1 and T2 pairs but also about the $R_{1}, R_{2}$ and $R_{3}$ pairs. In this way, participants draw on the entire association matrix to make inferences that are reasonable given their experience. Our results provide empirical evidence of these inferences, but do not tell us whether inferences were made by explicit reasoning or emerged from the dynamics of attention within and across trials (Kachergis, Yu \& Shiffrin, 2012; Yu \& Smith, 2012; Yu, Zhong \& Fricker, 2012). If replicated in young word learners, these results suggest an important role for the contexts in which word learning occurs.

There has been debate about the nature of information selection and information processing by cross-situational word learners. The presence of multiple objects and multiple labels on an individual learning instance means that learners could potentially associate all labels with all objects - the multiple association account (e.g., Yurovsky, Smith \& Yu, 2012). While equal attention may not be given to all possible mappings, this account predicts that learners will have a rich store of statistical information to draw on, so that if evidence for one particular mapping is contradicted (e.g., the label is given but the object is not present) there are other associations already in place that can inform the learners' inferences about the label's likely referent.

Alternatively, in the single-association account learners retain a single hypothesis for each object, discarding all other associations from a particular learning instance (Medina, Snedecker, Trueswell \& Gleitman, 2011; Trueswell, Medina, Hafri \& Gleitman, 2013). This account predicts that when a particular hypothesis is contradicted the learner must start from scratch, forming a new hypothesis at random based on the current learning instance.

These two accounts make disparate predictions for the present study, specifically within the Familiar Context task. The single-association account proposes that learners may form a hypothesis linking the T1 object and label during Trial 1 and that this hypothesis would be confirmed on Trial 3. However, because choice of hypotheses is random, there should not be systematic differences between which learners benefit from this extra information from one mini-task to the next. In direct contrast to this, our results suggest that for some learners, the repetition of information in Trial 3 was beneficial, improving accuracy on R and T1 items, and for some learners it was not. Crucially, what defined whether Trial 3 was beneficial was whether the participant had formed a strong memory for the first trial, both the potential T1 mapping AND the other objects present, as measured by their ability to perform backward inference. These findings raise important questions about how memory development may influence word learning in toddlers, as we found that better in the mini-tasks with high overlap, better memory led to potential interference, while in the larger task with
little trial-by-trial overlap, better memory (i.e., backward inference) led to better performance. Our data suggest that those learners who are successful in cross-situational learning tasks carry multiple possible associations forward. These associations are integrated in both the forward and backward directions to discover likely object-label pairs. Thus, statistical associative learning is a powerful mechanism that is within the repertoire of human cognitive systems.

## Acknowledgments

This work was supported by NIH R01 HD056029 to C. Yu and an NIH Ruth Kirschstein NRSA training grant to Indiana University: T32 HD007475-17.

## References

Kachergis, G., Yu, C., \& Shiffrin, R. M. (2012). An associative model of adaptive inference for learning word-referent mappings. Psychonomic Bulletin \& Review, 1-8.
Markman, E. M. (1990). Constraints children place on word meanings. Cognitive Science, 14, 57-77.
Medina, T. N., Snedeker, J., Trueswell, J. C., \& Gleitman, L. R. (2011). How words can and cannot be learned by observation. PNAS, 108(22), 9014-9019.
Scott, R. M., \& Fisher, C. (2011). 2.5-Year-olds use crosssituational consistency to learn verbs under referential uncertainty. Cognition, 122, 163-180.
Smith, L., \& Yu, C. (2008). Infants rapidly learn wordreferent mappings via cross-situational statistics. Cognition, 106, 1558-1568.
Suanda, S. H., \& Namy, L. L. (2012). Detailed Behavioral Analysis as a Window Into Cross- Situational Word Learning. Cognitive Science, 36, 545-559.
Trueswell, J. C., Medina, T. N., Hafri, A., \& Gleitman, L. R. (2013). Propose but verify: Fast mapping meets crosssituational word learning. Cognitive Psychology, 66, 126156.

Vouloumanos, A., \& Werker, J. F. (2009). Infants' learning of novel words in a stochastic environment. Developmental psychology, 45, 1611.
Yu, C., \& Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. Psychological Science, 18, 414-420.
Yu, C., \& Smith, L. B. (2012). Modeling cross-situational word-referent learning: Prior questions. Psychological Review, 119, 21.
Yu, C., Zhong, Y., \& Fricker, D. (2012). Selective attention in cross-situational statistical learning: evidence from eye tracking. Frontiers in Psychology, 3:148 doi: 10.3389/fpsyg.2012.00148

Yurovsky, D., Smith, L. B., \& Yu, C. (2012). Does Statistical Word Learning Scale? It's a Matter of Perspective. In N. Miyake, D. Peebles, \& R. Cooper (Eds.), Proceedings of the 34nd Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.

# A Dynamical Model of the Speech Perception-Production Link 

Kevin D. Roon (kdroon@nyu.edu)<br>Department of Linguistics, New York University, 10 Washington Place New York, NY 10003 USA

Adamantios I. Gafos (gafos@uni-potsdam.de)<br>Linguistics Department and Center of Cognitive Sciences, Universität Potsdam, Haus 14, Karl-Liebknecht-Straße 24-25 Potsdam 14476 Germany, and<br>Haskins Laboratories, 300 George Street<br>New Haven, CT 06511 USA


#### Abstract

How and to what extent the speech production and perception systems are linked is a question of longstanding debate (cf. Diehl, Lotto, \& Holt, 2004; Galantucci, Fowler, \& Turvey, 2006). Despite the long history of this debate and a vast number of studies providing experimental evidence indicating an intimate link between perception and production, formal proposals of this link have been conspicuously lacking in the literature. In this paper, we provide a computationally explicit, dynamical model of the process of phonological planning. In this model, the properties of a perceived utterance automatically serve as input to the ongoing planning of an intended utterance. Using tools from non-linear dynamics, we formalize how incoming inputs from perception influence the ongoing choice of phonological parameter values to be used in production. The use of a dynamical model enables establishing explicit bridges between phonological representations and response time data. Our model provides an account of response time modulations reported in independent experimental work, and makes additional concrete predictions that can be tested experimentally. In sum, our model provides a foundation for better understanding the cognition of speech perception, speech production, and the interaction between the two.


Keywords: Speech production; speech perception; dynamical modeling; perceptuo-motor effects; phonological planning.

## Perceptuo-Motor Effects

Many studies have provided evidence for the influence of the speech production system during the process of speech perception. Yuen, Brysbaert, Davis, \& Rastle (2010) showed that the articulations subjects produced could be modulated by stimuli they perceived immediately before producing a cued utterance. Specifically, subjects had increased alveolar closure in producing $s$ - or $k$-initial utterances when they heard a $t$-initial distractor, compared to baseline cases $(t$ is a sound produced with the tongue-tip making full contact at the alveolar ridge, but for fricatives like $s$ the tongue-tip contact is not complete, and for $k$ the contact is by a different articulator in a different location). D'Ausilio et al. (2009) administered transcranial magnetic stimulation to the areas of subjects' motor cortex that control lip or tongue movement and had subjects identify acoustic stimuli that were ambiguous as to place (labial vs. lingual). They found that subjects were more likely to mistakenly perceive stimuli as having the place whose corresponding motor cortex
area was being stimulated. Kerzel \& Bekkering (2000) and Galantucci, Fowler, \& Goldstein (2009) found that subjects response times (RTs) can be modulated systematically and involuntarily by various stimuli they perceive while speaking. We refer to these effects broadly as "perceptuo-motor effects" (Galantucci et al., 2009) because they are effects that indicate an influence of speech motor plans during the process of speech perception.

Much of the debate in the literature on the speech percep-tion-production link has centered on the claim of the Motor Theory of Speech Perception (Liberman \& Mattingly, 1985) that motor codes are the sole object of speech perception. However, as Lotto, Hickok, \& Holt (2009) point out, "there is no debate that speech production and perception interact in some manner [...] It is the 'nature' of the productionperception link that has not been established." The purpose of this study is not to disprove either side of the debate around that particular claim of the Motor Theory, but rather to address this latter point and provide a specific computationally explicit proposal regarding the nature of the percep-tion-production link.

In this paper, we propose a specific formalization of the perception-production link within a computational model of the dynamics of phonological planning. To illustrate our model, we focus on the response time data from the re-sponse-distractor experimental task used by Galantucci et al. (2009). In this task, subjects learned visual stimulus-spoken syllable pairings (e.g., if you see \&\& say ba, if you see \#\# say $d a$ ). While subjects were preparing the required responses (either $b a$ or $d a$ ), distractors were presented at varying delays (i.e., Stimulus Onset Asynchronies, "SOAs") relative to the presentation of the visual cue indicating the required response. The distractors were either a short tone, the same syllable the subject was preparing to say (e.g., $d a$ $d a$ ), or another syllable that differed in place of articulation from the response (e.g., $b a-g a$ ). In the Kerzel \& Bekkering (2000) study, video distractors were used instead of auditory distractors, with similar results.

Figure 1 summarizes the experimental results from Galantucci et al. (2009). First, the presence of any distractor resulted in longer RTs. Second, there was a monotonic effect of SOA on RTs. Both of these effects can be seen by looking at the Tone condition. RTs in the Tone condition (at both SOAs) were slower than on trials when there was no
distractor, and RTs in the Tone condition were longer at SOA 200 than at SOA 100 . From these two effects, it is evident that the mere presence of any distractor (linguistic or not) results in a slow-down in RTs. The Identity and Mismatch conditions introduce effects of linguistic (in)congruency between the responses and distractors in addition to whatever process generates the non-linguistic effects seen in the slow down due to a distractor presence and SOA. Crucially, RTs in the Mismatch condition were longer than in the Identity condition.


Figure 1: Result patterns from Galantucci et al. (2009).
These results motivate two broad computational principles, which in turn will inform the design of our model. These are the principles of excitation and inhibition. The fact that RTs were shorter in the Identity condition than in the Tone condition indicates the influence of excitation, since linguistic congruency offsets the slow-down introduced by the presence of a distractor. The longer RTs in the Mismatch condition compared to the Tone condition show the influence of inhibition due to linguistic incongruency, increasing the RTs beyond the effects of the mere presence of a distractor.

## Dynamical Model of Phonological Planning

We propose a formal, dynamical, computational model of the perception-production link, situating it in the planning process by which phonological parameters are set in speech production. The components of the model are shown in Figure 2. The model includes four dynamical planning fields (shaded rectangles), Inputs to these planning fields (ovals) that determine the actual parameter values to be produced, and a Monitor function that decides when all of the required values have been determined.

Figure 2 also shows an Implementation system that executes the motor plans for the intended utterance based on the production parameter values determined by the model. This Implementation system is not part of our model. The focus in our modeling work is on planning, that is, on the process of choosing values for phonological parameters. This process unfolds in time and, in the schematic shown in Figure 2, takes place before articulatory movement initiation and control, which are the business of the Implementation system. The Implementation system could be, e.g., either the Task Dynamics Model (Saltzman \& Munhall, 1989) or the DIVA model (Guenther, 1995).


Figure 2: Components of the model.
Planning Fields A key concept in the model is that of the planning field. Each phonological parameter of an intended utterance is assigned a planning field. Planning fields evolve over time and determine the specific parameter settings of the phonological parameters in an intended utterance. A planning field is defined by three axes: an axis representing the possible phonological parameter values, an axis representing the activation level associated with each possible phonological parameter value, and an axis representing time (see Figure 3).


Figure 3: Tongue Tip (TT) articulator planning field.
The phonological parameter relevant to the TT field is that of the constriction location for the tongue-tip articulator. Therefore, the phonological parameter axis in this field is represented by a continuum of constriction locations from dental (most anterior) to post-alveolar (most posterior). The use of a discrete planning field for each parameter is motivated by the desire to have our model be maximally compatible with extant models of phonological representation. The planning fields here correspond closely to the parameters used in Articulatory Phonology (Browman \& Goldstein, 1986, et seq.), with a field for each "tract variable", though our model could be applied to any appropriate system. As with the tongue-tip articulator, there are also planning fields for the other two primary oral articulators used in producing the syllables relevant to the experimental setting at handthe lower lip (LL) and tongue body (TB) -and one field for voicing. The parameter axis for the Voicing planning field is represented by the well-known continuum of Voice Onset Time ("VOT").

The planning fields evolve based on inputs to these fields. As we make explicit below, the dynamics of that evolution
are formalized within the computational framework of Dynamic Field Theory ("DFT", e.g., Erlhagen \& Schöner, 2002). Each field evolves such that after sufficient input(s), a peak of activation builds up and eventually stabilizes, with one parameter value having a maximum activation level. In Figure 3, following increasing values on the time axis, we can see the gradual evolution of a localized peak in activation at some value of constriction location intermediate between anterior and posterior.

Representing each articulator as its own field in the model with voicing as one separate field reflects the purpose of a planning field, which is to compute a single production value based on one or more potentially conflicting inputs. Our model assumes that these planning fields are the mechanism by which phonological planning of any utterance is achieved, that is, they are not specific to this experimental task. The design of the planning fields therefore reflects the general demands of speech production.

Input There are two sources of input to the evolving planning fields. One corresponds to the parameter values for the required response, and the other corresponds to the parameter values for the auditory distractor perceived during the planning of the utterance. Inputs are represented as twodimensional distributions of activation levels across the spectrum of possible values for a given parameter. Although not required by the framework, each given input in the present model is a normal distribution defined by the equation:

$$
\text { activation }_{\text {input }}=e^{-(x-\text { val }+ \text { noise })^{2}} / 2 \sigma^{2}
$$

val indicates the mean of the distribution, and was varied from trial to trial by adding a small noise term. Since constriction location does not vary in the examples used in the model of this task, the input values for constriction location did not materially change in the simulations. The standard deviation of the distribution $(\sigma)$ defines its width. Both responses and distractors in the task modeled here are voiced, so the input to the Voicing field was always the same.

Dynamics The purpose of the planning fields is to determine the phonological parameter values to be sent to implementation. Planning fields have two possible stable states. They either stay flat at their resting value, or they can have a single, sustained peak of activation. The value sent to implementation for a given field is the parameter value that has the maximum amount of activation when the field stabilizes in this second stable state. The fields in the model evolve based on the mechanisms of DFT. The dynamics of each of the three articulator planning fields (LL, TT, and TB ) are controlled by the equation:

$$
\begin{aligned}
& \tau d A(x, t)=-A(x, t)+h+r\left(\operatorname{input}_{\text {RESPONSE }}(x, t)\right) \\
& \quad+d\left(\text { input }_{\text {DISTRACTOR }}(x, t)\right)-\text { inhibition }_{\text {CROSS-FIELD }}(x, t) \\
& \quad+\text { interaction }(x, t)+\text { noise }^{\text {and }}
\end{aligned}
$$

$d A(x, t)$ is the change in activation level $A$ of $x$ at time step $t$. The rate of evolution of the field is controlled by $\tau$, with larger values of $\tau$ resulting in slower evolution of the field. $h$ is the resting level of the field. The inputs are added to the field, when appropriate, by the terms $\operatorname{input}_{\text {RESPONSE }}(x, t)$ and input $_{\text {DISTRACTOR }}(x, t) . r$ and $d$ encode the relative strengths of the inputs. The cross-field inhibition (indicated in Figure 2 by the bidirectional arrows between the articulator planning fields) introduced by any other articulator field(s) is added by the term $\operatorname{inhibition~}_{\text {Cross-FIELD }}(x, t)$ when the activation peak in another fields exceeds a threshold value $(\chi)$. The DFT dynamics (Erlhagen \& Schöner, 2002) capture the case of parameter setting for one effector. In our case, we have several articulators, and a single one must be chosen for any given speech segment. This is the motivation for the crossfield inhibition. In our model, cross-field inhibition follows a basic property of DFT in which inhibition comes into play when some threshold is crossed (we illustrate this with simulations below). Noise is added to introduce stochastic behavior into the model evolutions. The equation that defines the evolution of the Voicing field differs from the one that defines the evolution of the articulator fields only in that it does not contain a term for cross-field inhibition, because the Voicing field neither inhibits nor is inhibited by any other planning field. This design reflects the fact that voicing and articulator are cross-classifying parameters for English consonants (Chomsky \& Halle, 1968).

The interaction term, interaction $(x, t)$, the DFT "engine" that drives the evolution of the activation field through local excitation and global inhibition, is defined by:

$$
w(x)=w_{\text {excite }} e^{-\left(x^{2} / 2 \sigma_{w}^{2}\right)}-w_{\text {inhibit }}
$$

The interaction term induces changes in the field as some value(s) of $x$ approach a "soft" threshold ( $\theta$ ), which is determined by a sigmoid threshold function, defined by:

$$
f(u)=\frac{1}{1+\exp [-\beta(u-\theta)]}
$$

The use of a soft threshold means that some $x$ values below $\theta$ do engage the interaction term, but the contribution to the interaction of activation values less than $\theta$ diminishes with distance from $\theta$. The system is non-linear due to this soft threshold, in that incremental changes in activation levels have a non-uniform effect on the field's evolution. ${ }^{1}$

[^201]Since the required response and the perceived distractor both serve as input to the model, the evolution of the fields is driven by a combination of excitation and inhibition, depending on whether the two inputs have congruent parameter values. Congruent inputs to the model introduce excitation, while incongruent inputs inhibit each other.

Monitor The Monitor determines when activation has built up in required fields to a level that is sufficient to send to Implementation, based on a criterion value ( $\kappa$ ), which is the same across all four planning fields. The decision criteria for the Monitor are straightforward. The Monitor waits until the activation level for some $x$ value in both the Voicing field and one articulator field reach criterion. At that point it chooses the parameter values from those two fields with the highest activation level to be sent to Implementation. This has the effect that sometimes it is the Voicing field and sometimes an articulator field-whichever field evolves more slowly-that determines the RT on the trial.

## Simulations

Figure 4 illustrates the model dynamics by showing how the planning fields evolve during a single trial in three different conditions of the experimental study from Galantucci et al. (2009). The figures show the maximum activation level over time for each of the four planning fields. The dot-dashed red line shows the TT field evolution, the dashed blue line shows the LL field, the solid pink line shows the TB field, and the solid black line shows the Voicing field. Differences in the rate of rise of the maximum activation level of the fields predict differences in experimental RTs.

Figure 4A shows the evolution of the fields in the Tone condition. Since the tone distractor has no linguistic content, it serves as a baseline reference of how the planning fields evolve in the unperturbed case. The vertical dotted lines at time steps 100 and 500 indicate the duration of the required response input to the fields. Thus, the activation levels of the TT and Voicing fields start to rise at time step 100, the point at which the subject begins planning the required utterance based on the visual cue on that trial (here \# \# instructing the subject to say $d a$ ). The horizontal dotdashed black line drawn at activation level 0.7 indicates the soft threshold $(\theta)$ that determines the engagement of the within-field interaction term. The effects of the interaction term can be seen in that the rate of increase in the activation level of the TT and Voicing fields is not linear: as the activation level of each field approaches $\theta$, the steepness of the curve increases due to the local excitation being generated
activation fields: $\theta=0.7, w_{\text {excite }}=0.45, w_{\text {inhibit }}=0.1, \sigma=1$. For the sigmoid function, $\beta$ was always 1.5 . The constriction location input distribution for all articulator fields had a mean (val) of 0 and $\mathrm{SD}=$ 2, defined on an arbitrary scale of constriction locations that ranged from -10 to 10 . For the Voicing parameter, distributions for all voiced stimuli input had a mean of 5 ms VOT and $\mathrm{SD}=45 \mathrm{~ms}$. The criterion value ( $\kappa$ ) was 5 . The specific values of the variables in the above equations are not meaningful in and of themselves. Their values relative to each other are more informative.
in those fields by the interaction term. The TB and LL fields receive no input, and there is no change in their activation levels until around time step 200. At that point their activation levels start to drop due to the TT field reaching the cross-field inhibition threshold $(\chi)$, indicated by the horizontal dot-dashed teal line drawn at activation level 0 . The TT and Voicing activation levels continue to rise until they both have passed the criterion value ( $\kappa$ ), indicated by the solid line drawn at activation level 5. The time step at which the second field passes $\kappa$ (minus 100 , since that is the time step at which the cue is presented) is marked as the RT on that trial (the solid vertical line at about time step 425). The Monitor takes the maximum parameter values from the Voicing and TT fields and passes them to Implementation.


Figure 4: Evolution of planning fields in (A) the Tone, (B) Identity, and (C) Mismatch conditions. Dashed blue lines show the maximum activation level of the LL field, dot-dashed red of the TT field, solid pink of the TB field, solid black of the Voicing field. The horizontal black line at activation $=5$ shows the threshold $\kappa$ at which the Monitor chooses values for production, and the vertical black line perpendicular to it shows the simulated RT on the trial.

Figure 4 B shows the evolution of the fields in the Identity case on a trial with a $d a$ response and $d a$ distractor. From time step 0 to 200, all fields evolve in the same way as in the Tone condition. The vertical dashed lines at time steps 200 and 325 indicate the duration of the input from the distractor. Since the distractor inputs are qualitatively the same as those for the response, the activation level for the TT and Voicing fields rises at a much greater rate than in the Tone condition because both inputs add activation to the same range of parameter values, in addition to the local excitation being generated by the interaction term. The fields therefore cross $\kappa$ earlier than in the Tone condition, and the simulated RT is shorter.

Figure 4C shows the evolution of the fields in the Mismatch case on a trial with a $d a$ response and $b a$ distractor. Since the response and distractor share the same value of voicing, the evolution of the Voicing field in this condition is qualitatively the same as in the Identity case. The evolution of the TT field, however, is different. When the distractor input starts at time step 200, the activation level of the LL field begins to rise, and eventually crosses $\chi$, introducing cross-field inhibition to the TT field. The distractor input ends at time step 325, but by that time the LL field maximum is well above $\theta$, so it maintains a peak of activation for some time due to the interaction term, and the cross-field inhibition of the TT field by the LL field therefore persists. As a result, the rate of rise of the TT field activation level slows down compared to its rise in the Tone condition. The Monitor has to wait longer for the TT field to cross $\kappa$, and thus the RT on this trial is longer than in the Tone condition.


Figure 5: (A) Model simulations of the Galantucci et al. experiment. (B) Predictions for a scenario where distractor and response differ in Voicing rather than articulator.

Using our model, we simulated 150 trials for each of the three conditions (Identity, Tone, and Mismatch) at an SOA of 100 time steps for a total of 450 trials. On each trial, the time step at which the Monitor determined the RT was recorded. The activation level of each planning field was reset to its trial-initial state at the beginning of each trial. The simulated results are shown in Figure 5A. The model qualitatively replicates the experimental results from Galantucci et al. (2009). RTs in the Identity condition were shorter than the Tone condition, due to the reinforcing inputs and lack of any inhibition. On the other hand, RTs were longer in the Mismatch condition than in the Tone condition (and than the Identity condition). This is because the distractor and
response differed in articulator. As explained above, in this case cross-field inhibition slows down the evolution of the articulator field for the required response.

## Discussion

Our model fills a gap in the speech planning literature. Models of speech motor implementation (Saltzman \& Munhall, 1989; Guenther, 1995) explicitly capture how articulators move through space and over time to achieve their linguistic targets, but existing models of the sources of those target values either do not address the timecourse of the planning process (Chomsky \& Halle, 1968; Browman \& Goldstein, 1986), or assign little to no role to representations at the level of phonological features (Dell, 1986; Roelofs, 2000). Our results show the benefit of a model that addresses timecourse and phonological features explicitly.

Our model makes additional predictions for the cuedistractor task that can be tested experimentally. The model predicts that similar RT effects should be obtained when the Mismatch condition is such that the distractor and response differ in voicing rather than in articulator. The model predicts longer RTs in this Mismatch condition (e.g., da-ta) than in the Tone or Identity condition (e.g., da-da). In our model, the source of this difference in RTs is the withinfield inhibition that arises from the introduction of two incompatible inputs to the same field. This within-field inhibition is an inherent property of the DFT computational framework. Galantucci et al. (2009) did not test this condition, but results reported by Roon (2013) show perceptuomotor effects of voicing in the response-distractor task that are independent of articulator. This prediction is thus borne out. Figure 5B shows the model predictions for an experiment where the Identity condition is the same as the one reported in Figure 5A ( $d a-d a$ ), but where the Mismatch condition is da-ta. The model predicts slower RTs in the Mismatch condition than in the Tone condition. Future work will involve expanding the model to accommodate these new experimental results.

A second set of predictions concerns variation and phonetic detail in representations and processes. Since categories like voicing are defined as distributions on a phonetic continuum like VOT, compatible inputs need not be exactly the same in order to mutually excite each other: it is sufficient for the maximum activation peaks of two inputs to be near enough to each other. This excitation happens automatically, without any need to classify inputs categorically by defining category ranges. We plan to pursue this set of predictions in future work as well.

Most speech consists of utterances that are longer than monosyllables. Our present model does not address the planning field dynamics beyond CV syllables, which is what is required to account for reported perceptuo-motor effects. Future expansion of the model will address the dynamics involved in the planning of larger utterances.

Our model of the observed experimental effects bears directly on establishing the nature of the perceptionproduction link. In our model, speech perception is linked to
speech production as part of the process by which parameter values are set. The link between perception and production is the obligatory input of the perceived distractor to the motor planning field shown in Figure 2. Given the facilitation and inhibition based on (in)congruency between distractors and responses, there must be some intersection between the motor codes activated during motor planning of the required response and the codes activated during the perception of the distractor. The term "codes" refers to parameters such as voicing and articulator, and more precisely to the parameter values represented in our model. Our claim is not that the codes activated by perceiving the distractor must exclusively be motor codes. Rather, it is that the codes activated in the perception of the distractor must minimally be motor codes. Our study was not designed to address whether nonmotor codes are also activated. Our results are fully compatible with the Motor Theory (Liberman \& Mattingly, 1985). Our results are also consistent with theories that do or could propose a link between auditory-acoustic (or other) codes that are activated during the perception of the distractor, and motor codes corresponding to these auditory-acoustic codes (cf. Viviani, 2002: Figure 21.12), as long as a link between these other codes and the motor codes is assumed.

In sum, the perception-production link must be specified at the level of setting motor parameter values, including articulator and voicing, that need to be activated either directly or via associated codes. The effects of the perceptionproduction link are seen as the influence of a perceived distractor on the process of setting those parameters, i.e., on a production process, as seen in the reported RT modulations and their simulation by our model.

## Conclusions

During speech production, a speaker must retrieve the phonological representations of the required utterances by assembling a set of parameter values that specify the vocal tract actions corresponding to these utterances. We have presented a formal, dynamical, computational model of this process. In the model, assigning values to these parameters is a time-dependent process, captured as the evolution of a dynamical system over time. The model accounts for experimental results that have been proposed as evidence for an intimate link between perception and production. In our model, the perception-production link consists of the phonological parameter values of a perceived stimulus obligatorily contributing to the evolution of the activation levels of the fields engaged with the ongoing phonological planning of a required response. The present model can explain reported effects on response times, and makes new, experimentally testable predictions about similar response time modulations. The model therefore provides a foundation for a better understanding of speech production, perception, and the link between the two.

## Acknowledgments

KDR gratefully acknowledge research supported by NSF Grant 0951831 . Any opinions, findings, and conclusions or
recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. AIG gratefully acknowledges support from ERC AdG Grant 249440.

## References

Browman, C. P., \& Goldstein, L. M. (1986). Towards an articulatory phonology. Phonology Yearbook, 3, 219-252.
Chomsky, N., \& Halle, M. (1968). The Sound Pattern of English. New York: Harper \& Row.
D'Ausilio, A., Pulvermüller, F., Salmas, P., Bufalari, I., Begliomini, C., \& Fadiga, L. (2009). The motor somatotopy of speech perception. Current Biology, 19, 381-385.
Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. Psych. Review, 93, 283-321.
Diehl, R. L., Lotto, A. J., \& Holt, L. L. (2004). Speech perception. Annual Review of Psychology, 55, 149-179.
Erlhagen, W., \& Schöner, G. (2002). Dynamic field theory of movement preparation. Psych. Review, 109, 545-572.
Galantucci, B., Fowler, C. A., \& Goldstein, L. M. (2009). Perceptuomotor compatibility effects in speech. Attention, Perception, \& Psychophysics, 71(5), 1138-1149.
Galantucci, B., Fowler, C. A., \& Turvey, M. T. (2006). The motor theory of speech perception reviewed. Psychonomic Bulletin \& Review, 13(3), 361-377.
Guenther, F. H. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. Psych. Review, 102, 594-621.
Kerzel, D., \& Bekkering, H. (2000). Motor activation from visible speech: Evidence from stimulus response compatibility. Journal of Experimental Psychology: Human Perception and Performance, 26(2), 634-647.
Liberman, A. M., \& Mattingly, I. G. (1985). The motor theory of speech perception revised. Cognition, 21, 1-36.
Lotto, A. J., Hickok, G. S., \& Holt, L. L. (2009). Reflections on mirror neurons and speech perception. TRENDS in Cognitive Science, 13, 110-114.
Roelofs, A. (2000). WEAVER++ and other computational models of lemma retrievala and word-form encoding. In L. R. Wheeldon (Ed.), Aspects of Language Production (pp. 71-114). Philadelphia: Psychology Press.
Roon, K. D. (2013). The dynamics of phonological planning. Doctoral dissertation, Department of Linguistics, New York University, New York, NY.
Saltzman, E. L., \& Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. Ecological Psychology, 1(4), 333-382.
Viviani, P. (2002). Motor competence in the perception of dynamic events: a tutorial. In W. Prinz \& B. Hommel (Eds.), Common mechanisms in perception and action: Attention and performance XIX (pp. 406-442). Oxford/New York: Oxford University Press.
Yuen, I., Brysbaert, M., Davis, M. H., \& Rastle, K. (2010). Activation of articulatory information in speech perception. Proceedings of the National Academy of Sciences (Social Sciences), 107(2), 592-597.

# The spatial representation of grammatical number 

Timo B. Röttger (timo.roettger@uni-koeln.de)<br>IfL Phonetik, University of Cologne, Herbert-Levin-Str. 6, D-50931 Köln, Germany<br>Frank Domahs (domahs@uni-marburg.de)<br>Department of Germanic Linguistics, University of Marburg, Wilhelm-Röpke-Str. 6a, D-35032 Marburg, Germany


#### Abstract

Research on numerical cognition suggests a strong link between mental representations of space and quantity. The SNARC effect (Spatial-Numerical Association of Response Codes effect) is characterized by the association of small quantities with left space and large quantities with right space. While the majority of research on the spatial representation of number has been on number words or Arabic numerals, this study investigates quantity representations that are involved in the processing of grammatical number. We found that German words that were inflected for singular had a relative left hand advantage, and conversely, plurals had a relative right-hand advantage. However, this pattern was only found in relatively late responses. Moreover, it appeared to interfere with the opposite pattern caused by the MARC effect (Markedness Association of Response Codes effect) leading to a relative right-hand advantage for singulars. This interference appeared to depend mainly on response latency with MARC effects being more pronounced in early responses and SNARC-like effects being more pronounced in late responses. This work sheds light on the interaction of different stimulus-to-response mappings operating on the same stimulus dimension - grammatical number. Moreover, it suggests that spatial numerical associations go beyond explicit numerical information, as in number words or Arabic numerals.


Keywords: grammatical number, MARC effect; numerical representation; SNARC effect.

## Introduction

Many researchers have argued that the mental representation of quantity is intimately connected to space. This connection is often described using the metaphor of a mental number line, which (in Western cultures) is oriented from left to right. In line with this assumption, it has been shown that spatial response dimensions are associated to numerical magnitude: the SNARC effect is characterized by the association of small quantities to the left hand and large quantities to the right hand. In their seminal work, Dehaene, Bossini, and Giraux (1993) found that in a parity judgment task ("is the number even or odd?"), responses to larger numbers were consistently faster with the right hand than with the left hand, whereas responses to smaller numbers showed the opposite pattern. As the task was not explicitly focused on quantity information but on parity, the interaction between quantity and spatial orientation was taken to suggest automatic access to quantity representations which are organized horizontally. Several studies have found similar effects without hand movements, suggesting that the SNARC effect is not genuinely due to a
mapping to hands but to perceptual space (e.g., Fischer, Castel, Dodd, \& Praat, 2003; Loetscher, Schwarz, Schubiger, \& Brugger, 2008). The SNARC effect has been shown for both Arabic numbers and for spoken or written number words (cf., Landy, Jones, \& Hummel, 2008; Nuerk, Iverson, \& Willmes, 2004; Nuerk, Wood, \& Willmes, 2005).

In an alternative account, the SNARC effect could be attributed to polarity alignment (Landy et al., 2008; Proctor \& Cho, 2006; Santens \& Gevers, 2008). This account posits that in binary representations of dimensions, across both stimulus and response properties, one value of the dimension is "generally more available than the other" (Landy et al., 2008: 358). To account for the SNARC effect, e.g. the polarity correspondence principle (Proctor \& Cho, 2006) assumes that small numbers are coded as [-] polarity and large numbers as $[+]$ polarity. The response location is coded in a similar way: [-] polarity for a left response and $[+]$ polarity for a right response. Congruent polarities (small numbers/left space, large numbers/right space) cause faster response selection than incongruent polarities.
This model also accounts for the MARC effect (Markedness Association of Response Codes effect, cf., Nuerk et al., 2004; Reynvoet \& Brysbaert, 1999; Willmes \& Iversen, 1995). An example of the MARC effect are faster right hand responses to even numbers and faster left hand responses to odd numbers (see e.g., Nuerk et al., 2004). It is assumed that this effect is closely related to the concept of linguistic markedness (see Haspelmath, 2006, for an overview) which refers to the formal and conceptual asymmetry between linguistic categories: in a parity judgment task, in which the hand-to-response relation is manipulated within participants, the adjectives "right" and "even" are assumed to be linguistically unmarked (Zimmer, 1964). On the contrary, "left" and "odd" are assumed to be linguistically marked. Interference is observed if the markedness association between stimulus and response is incongruent, while facilitation is observed if the markedness association is congruent.
At least for numerals, SNARC and MARC effects may cooccur (e.g., Nuerk et al., 2004). However, they do not interfere with each other since they are linked to independent stimulus properties (SNARC is linked to relative magnitude, MARC is linked to parity).

## Grammatical number, quantity, and markedness

In addition to symbolic and lexical number representations, many languages encode quantity grammatically. In particular, languages such as English and German employ morphological markers that decode the distinction between one entity ("singular") and more than one entity ("plural"). Most commonly, nouns are grammatically marked for number by inflection, e.g., by adding an affix such as $-s$ to English nouns. The most frequent grammatical number systems restrict the number of available categories to singular (one entity) and plural (more than one entity) (Corbett, 2000). For a German example, compare (1), where the suffix $-n$ adds plural meaning to the noun lion.
(1) Löwe 'lion' vs. Löwen 'lions'

While, most research on mental quantity representation has focused on Arabic numerals or number words; much less is known about the semantic interpretation of grammatical number. Several developmental and behavioral studies demonstrated a tight connection between grammatical and conceptual number (Barner, Thalwitz, Wood, Yang, \& Carey, 2007; Berent, Pinker, Tzelgov, Bibi, \& Goldfarb, 2005; Sarnecka, Kamenskaya, Yamana, Ogura, \& Yudovina, 2007). For example, in a Stroop-like task, Berent et al. (2005) asked their participants to judge the quantity (one or two) of visually presented words while ignoring their contents. Letter strings consisted of both singular and plural nouns (Exp. 1), and of pseudowords with or without regular plural inflection (Exp. 3). Response latencies were higher when there was a mismatch between grammatical number and the quantity of words presented (e.g., dog dog vs. $d o g s d o g s$ ). The authors concluded that the extraction of semantic number from grammatical number is automatic and represented in a way that is comparable to the conceptual number that they extract from visual perception.
The present study follows up on those findings and links it to numerical cognition research. Grammatical number is an excellent testing ground for the interaction of contradicting stimulus-to-response mappings because it allows us to pit SNARC-based and MARC-based accounts against each other.

## The present study

The present study applies a binary classification task to German nouns inflected for singular or plural. Conceptual quantity is involved in the process of specifying the grammatical number of nouns because, typically, singular nouns refer to one entity and plural nouns refer to multiple entities. Although the plural does not represent a specific quantity, we assumed it to represent a quantity which is - on a (Western) mental number line - localized more towards the right relative to a singular quantity $(=1)$, thus leading to a SNARC-like effect. In other words, singular forms should
be responded to faster with the left hand whereas plurals should be responded to faster with the right hand.

This prediction goes against the predictions based on the MARC effect: In linguistic theory, singular is thought to be unmarked, and plural is thought to be marked (cf., Greenberg, 1966). For example, within a language, singulars are used more frequently than plurals. And, if a language has a morphological coding of number (such as an affix), then the plural is typically overtly coded, thus formally more complex, whereas singulars often lack an overt coding, as in the German example (1) above. The MARC effect predicts that if markedness of a stimulus (singular vs. plural) is congruent with the markedness of a response side (right vs. left), there should be facilitation. Hence, singular forms should be responded to faster with the right hand (unmarked) whereas plurals should be responded to faster with the left hand (marked).

Apparently, grammatical number poses a problem to polarity accounts. Two conflicting polarity alignments are potentially at work operating on the same stimulus dimension: one alignment coding singulars as [ + ] polarity due to its linguistically unmarked status, and one coding singulars as $[-]$ polarity due to the conceptual quantity representations. Typically, however, polarity alignment accounts do not deal with conflicting polarity associations and therefore they make no prediction about which polarity association should occur in a given setting. Moreover, if competing associations interfere with each other the model does not predict how interference affects behavior.

One level of dissociation of those effects might operate on processing depth: the SNARC effect may become stronger when magnitude processing is activated more intensively (Gevers, Verguts, Reynvoet, Caessens, \& Fias, 2006), i.e. the size of the SNARC effect depends on response latencies and the amount of semantic number processing required. In their meta-analysis, Wood, Willmes, Nuerk, \& Fischer (2008) found a positive correlation of the SNARC effect size and response latencies across studies. Moreover, they found SNARC effects to be more pronounced in studies in which the task required the active processing of numerical magnitude (see also De Brauwer \& Duyck, 2008; Fias, 2001). Because the SNARC effect requires a certain amount of semantic magnitude processing, we expect it to occur only in semantic tasks. The MARC effect on the other hand, could already occur in an asemantic task, since no semantic information is necessarily required to encode a plural inflection, which is a surface characteristic of a word. Thus, one might hypothesize that those two effects are potentially dissociated in respect to task requirements. To explore this possibility and to investigate a potential dissociation of SNARC and MARC, we introduced tasks requiring different processing depths.

## Method

We designed four different tasks corresponding to different stages of processing depth. In the first task, participants had
to decide whether the presented words were written in italics or not (surface processing, SURF). The second task was a lexical decision task: participants had to decide whether the presented letter strings were existing German words or not (lexical processing, LEX). In the third task, participants had to make animacy judgments, where they had to decide whether the nouns denoted creatures (animate) or objects (inanimate) (nonspecific semantic processing, SEM). In a fourth task, participants had to decide whether the nouns denoted one entity or more than one (specific semantic quantity processing, QUANT).
Because quantity information is assumed to be represented at a conceptual level of processing and the SURF and LEX conditions do not require conceptual access, we expected no SNARC effect to occur at SURF and LEX. On the other hand, both decisions in SEM and QUANT required access to conceptual representations, thus a SNARC effect is expected to occur at SEM and QUANT. A MARC effect, however, could already occur in asemantic tasks, thus we do not predict any task dependency of a potential MARC effect. In their interaction, with increasing processing depth the impact of the MARC effect should be increasingly attenuated by the impact of the SNARC effect.

## Participants

Fifty-two native speakers of German ( 33 female, 19 male), with an average age of 26.9 years ( $\mathrm{SD}=7.0$ ) volunteered to participate for payment. All of them had normal or corrected-to-normal vision.

## Stimuli

The stimuli consisted of four German nouns in both their singular and plural form, respectively (Kuh/Kühe 'cow(s)', Löwe/Löwen 'lion(s)', Münze/Münzen 'coin(s)', and Stuhl/Stühle 'chair(s)'). We applied the following selection criteria to fit the stimuli to the experimental design: two items were animate beings (Kuh, Löwe); the other two were inanimate entities (Stuhl, Münze). There were two grammatically masculine (Stuhl, Löwe) and two grammatically feminine nouns (Münze, Kuh). Plural forms of all nouns contained an umlaut. Because both singular and plural forms can have an $-e$ suffix and an umlaut, neither of these cues was valid for unambiguously detecting plural inflection. This was done to ensure that participants access lexical knowledge rather than focus their attention just to one particular orthographic cue.

## Procedure

All subjects participated in eight blocks of trials, i.e. two blocks per processing depth (SURF, LEX, SEM, and QUANT). After the first block of each processing depth, there was a short break, in which participants were instructed to reverse the assignment of response buttons. The order of response assignments to the right hand and the left hand, respectively, was counter-balanced across participants. Each block started with a training session in which all words were presented
once. In the test blocks, each word was presented ten times in randomized order.

The experiment was controlled using Superlab 2.04 software (Abboud, 1991) and a RB-830 response box (both Cedrus Corporation, San Pedro, CA, USA). Stimuli were displayed on a 16 "-monitor screen using black symbols against a white background. Stimuli were presented in Times New Roman, font size 90 , resulting in a maximum height of 15 mm and a maximum width of 50 mm . Responses were recorded by two response keys placed at a distance of 30 cm in front of the participants, centered in egocentric space and separated 10 cm from each other. At the beginning of each trial, a fixation stimulus consisting of five asterisks $(* * * * *)$ was presented in the center of the screen for 300 ms . Then, the target appeared and remained for 1300 ms , during which response time was measured. The inter-trial-interval was 1500 ms (blank screen). The instructions given to participants stressed both speed and accuracy.

## Analysis

Six participants were excluded from analyses because they showed difficulties in changing the response assignment in at least one task. In the remaining data set, $5.8 \%$ of the trials had to be excluded due to wrong responses (3.45\%), anticipations (RT faster than 200 ms ) ( $0.05 \%$ ), or RTs outside $\pm 3$ standard deviations from the individual mean of each task per hand association per speaker ( $2.31 \%$ ). There was no trade-off between mean RT and error rate ( $r=-.182$; $\mathrm{p}>.05$ ).

Reaction times were analyzed using a series of generalized linear mixed effects models implemented in the R software ( R Core Team, 2012) and the package lme 4 (Bates, Maechler, \& Bolker, 2012). We used a Gaussian error distribution and identity link function. Both subjects and items were used as crossed random intercept effects. Since we are interested in the interaction of stimulus and response, we included the factor Number (singular, plural) interacting with the factor responding Hand (right, left) as a fixed effect in the models.

In a first step, we tested if this interaction is dependent on task requirements, thus we included a three-way interaction of Hand $\times$ Number $\times$ Task (SURF, LEX, SEM, QUANT) as a fixed effect. In subsequent analyses we tested the Hand $\times$ Number interaction for each task separately.

We computed p -values comparing the models with the interactions in question to the models with only the noninteracting fixed effects via Likelihood ratio tests (LRTs).

## Results

Overall, responses to tasks differed substantially in terms of response latency, such that SURF was responded to fastest ( 513 ms ) followed by SEM ( 548 ms ), LEX ( 579 ms ), and QUANT ( 631 ms ). Crucially, the Hand $\times$ Number $\times$ Task interaction was significant $(\chi(9)=81.514, \mathrm{p}<0.0001)$, indicating that there was a stimulus-response interaction modulated by task specific effects (cf. Figure 1).


Figure 1: Estimated RT differences (dRT) and standard errors between right hand and left hand responses as a function of grammatical number ( $\mathrm{SG}=$ singular; $\mathrm{PL}=$ plural). Negative slopes indicate SNARC-like effects; positive slopes indicate MARC-like effects. Mean RT of each depth is given in brackets.

Subset analyses of each processing depth separately revealed that SURF showed a significant Hand $\times$ Number interaction $(\chi(1)=4.096, \mathrm{p}=0.043)$, such that the model estimated a greater right hand advantage for singular forms ( 9 ms ) than for plural forms ( 1 ms ) ( $\mathrm{SE}=3.88 \mathrm{~ms}$ ), as predicted by a MARC-based account. The two-way interaction Hand $\times$ Number did not reach significance in the LEX or SEM condition $(\chi(1)=1.226, \mathrm{p}=0.27$ and $\chi(1)=0.303$, $\mathrm{p}=0.58$, respectively). For the QUANT processing depth, there was a significant interaction of Hand $\times$ Number $(\chi(1)=35.11, \mathrm{p}<0.0001)$ such that the model estimated a left hand advantage for singular forms ( 6 ms ) and a right hand advantage for plural forms ( 26 ms ) ( $\mathrm{SE}=5.36 \mathrm{~ms}$ ), as predicted by a SNARC-based account.

Table 1: Overview of stimulus-to-response mappings as a function of task and RT bin.

| Task | Bin | dRT SG | dRT PL | Slope | SE |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  | 1 | -3,71 | -1,31 | -2,40 | 2,18 |
|  | 2 | -1,28 | 0,09 | -1,37 | 1,22 |
|  | 3 | -0,23 | -1,16 | 0,94 | 1,67 |
|  | 4 | -0,53 | -3,21 | 2,68 | 6,12 |
| 芹 | 1 | -2,04 | 2,46 | -4,50 | 2,59 |
|  | 2 | -0,33 | -1,59 | 1,26 | 1,34 |
|  | 3 | 1,69 | 1,21 | 0,47 | 1,80 |
|  | 4 | -4,52 | -7,16 | 2,64 | 6,04 |
| $\sum_{\sqrt[y]{n}}^{N}$ | 1 | -1,56 | 3,42 | -4,99 | 2,55 |
|  | 2 | -1,83 | 0,47 | -2,30 | 1,41 |
|  | 3 | 1,37 | 1,48 | -0,11 | 1,53 |
|  | 4 | 7,39 | -0,99 | 8,38 | 6,22 |
| $\frac{5}{2}$ | 1 | -3,33 | -3,40 | 0,07 | 3,19 |
|  | 2 | 2,09 | 1,73 | 0,36 | 1,97 |
|  | 3 | 1,96 | -3,14 | 5,10 | 2,56 |
|  | 4 | 3,65 | -16,95 | 20,60 | 8,57 |



Figure 2: Estimated RT differences (dRT) and standard errors between right hand and left hand responses as a function of grammatical number for RT bins (centered around zero). Negative slopes indicate SNARC-like effects; positive slopes indicate MARC-like effects. Mean RT of each bin is given in brackets.

Since the mean response latencies of the tasks differed substantially, the observed dissociation between SNARC and MARC might be due to overall processing time rather than required magnitude processing. To obtain a view of the time course, we rank ordered RTs for each subject and processing depth and divided them into four equal bins (Ratcliff, 1979). We tested if the Hand $\times$ Number interaction was dependent on the factor RT $\operatorname{bin}$ (bin 1-4) for each task separately. This was not the case for SURF, LEX or SEM $(\chi(9) \leq 12.75, p \geq 0.17)$. It was, however, for QUANT as indicated by a significant interaction of Hand $\times$ Number $\times$ $R T \operatorname{bin}(\chi(9)=20.77, \mathrm{p}=0.014)$. In this condition, there was a significant SNARC-like effect in late responses (bin 3 and 4) $(\chi(1) \geq 3.94, \mathrm{p} \leq 0.047)$, but not for early responses in bin 1 and $2(\chi(1) \leq 0.035, \mathrm{p} \geq 0.85)$ (cf. Table 1, Figure 2).

Numerical trends further indicate that SNARC-like effects are found in all tasks depending on overall processing time. This pattern of evidence suggests that these SNARC effects
could be accounted for by processing time only rather than processing depth. Visual inspection of the slopes over time yielded a similar pattern: Early responses exhibit positive or flat slopes indicating MARC-like patterns and/or the absence of SNARC-like effects, while late responses exhibit negative slopes indicating SNARC-like patterns. Moreover, the change of slope over the time course appears to be roughly linear.

## Discussion

The present study investigated stimulus-to-response mappings when processing grammatical number in binary tasks. We demonstrated that grammatical number markers elicit a SNARC-like effect, i.e. German words inflected for singular had a relative left hand advantage; plurals had a relative right hand advantage. At the same time, we demonstrated a MARC effect that showed the opposite pattern. There was a reliable MARC effect in a font classification task (SURF) and a reliable SNARC effect in a magnitude classification task (QUANT). In the light of our task dependent pattern of results, this evidence suggests that the SNARC effect is elicited in relatively late processing stages. A look at the overall RTs obtained reveals that QUANT indeed required the longest processing time. A significant interaction of reaction times and stimulus-toresponse mapping in the magnitude classification task as well as numerical trends in all tasks (cf. Table 1, Figure 2) underpin this interpretation. So, one may conclude that a simple explanation based on processing time is sufficient to account for the present pattern of results ("A MARC effect already appears in early responses while a SNARC effect only appears in late responses"). The appearance of SNARC in relatively late responses is in line with earlier findings on Arabic numerals and number words (e.g., Wood et al., 2008).

Polarity alignment accounts (Landy et al., 2008; Proctor \& Cho, 2006; Santens \& Gevers, 2008) explain both the SNARC and the MARC effect within the same framework. According to this account, congruent polarities lead to faster response selection than incongruent polarities. However, this account makes contradicting predictions regarding the response association for grammatical number: Based on the linguistic markedness dimension, singulars should be coded as $[+]$ polarity and plurals as $[-]$ polarity, thus leading to a facilation of right hand responses for singular forms. A quantity-based account makes the opposite prediction, which assumes that singulars are coded as [-] polarity and plurals as $[+]$ polarity (in analogy to numerals). Interestingly, the present study found both patterns, thus two conflicting polarity alignments have been shown to operate on the same stimulus dimension. Polarity alignment accounts in their present state, however, do not predict which polarity associations occur in a given setting and - if competing associations interfere with each other - how their interaction affects behavior. The present data indicate a temporal dissociation of these stimulus-to-response
associations with MARC effects being more dominant in early responses and SNARC-like effects being more dominant in late responses. Given the apparent linear change of slopes as a function of processing time, we might speculate that both effects co-occur, interfering with each other. Over time, the relative strength of one stimulus-toresponse mapping (MARC) decreases (or remains constant) while the alternative mapping (SNARC) increases. Due to the lack of statistically significant results for some conditions, this remains, however, speculative.

Generally, the presence of a SNARC effect in the quantity task demonstrates that a mental quantity representation may - in principle - be accessed from grammatical number in a similar way as during the processing of Arabic numbers and number words. One might argue, that the present data are ambivalent with respect to the question whether this quantity representation of grammatical number can be conceived as organised in a left-to-right oriented mental number line or not. One could, of course, doubt the relevance of the quantity-to-space nature of the response-tostimulus mapping and stick with a more neutral polarity account arguing that there is a coding of singular as [-] polarity and a coding of plural as [+] polarity. This interpretation does not require any reference to spatial quantity representation, and consequently our data would say nothing about the association between conceptual number and grammatical number. However, one would have to explain why singular is associated with [-] polarity and plural with a [+] polarity. To us, one possible interpretation is grounded in the spatial nature of the conceptual quantity representation.

Future research might shed light on these issues. An excellent testing ground are languages which have more complex morphological number systems: In addition to singular and plural, some number systems also have an additional grammatical category that is called "dual", which serves to refer to two distinct real-world entities (cf., Corbett, 2000). Other, more rarely occurring grammatical systems also contain a so-called "trial", in which nouns are marked for groups of exactly three distinct entities, or even a "paucal", in which a separate grammatical marker is used to refer to a small number of distinct entities. Grammatical systems in which more than two morphological categories are used to refer to quantity might further our understanding of the interaction of different stimulus-to-response associations in general and the interrelationship of linguistic and conceptual number in particular.

## Conclusion

To conclude, grammatical number elicits two contradicting stimulus-to-response mappings. A MARC effect based on the linguistic markedness of the grammatical categories singular and plural; and a SNARC-like effect based on its semantic reference to magnitudes. Similar to Arabic numbers and number words, this quantity representation seems to be organised in a rightward direction for increasing quantities. This SNARC-like effect,
however, only appears in relatively late responses, while the MARC effect appears to be restricted to relatively early responses. Linear trends in slope changes over time indicate that both effects interfere with each other.
In general, the use of linguistic categories beyond number words appears to be an interesting and promising avenue to investigate the relationship of different stimulus-to-response mappings.

## Acknowledgments

We would like to thank Bodo Winter, Martine Grice, and two anonymous reviewers for helpful comments and suggestions. Any remaining errors are our own.

## References

Abboud, H. (1991). SuperLab. Wheaton, MD: Cedrus.
Barner, D., Thalwitz, D., Wood, J., Yang, S., \& Carey, S. (2007). On the relation between the acquisition of singular-plural morpho-syntax and the conceptual distinction between one and more than one. Developmental Science, 10, 365-373.
Bates, D., Maechler, M., \& Bolker, B. (2012). Ime4: Linear mixed-effects models using S4 classes. R package version 0.999999-0. http://CRAN.R-project.org/package=lme4

Berent, I., Pinker, S., Tzelgov, J., Bibi, J., \& Goldfarb, L. (2005). Computation of semantic number from morphological information. Journal of Memory and Language, 53, 342-358.
Corbett, G. (2000). Number. New York: Cambridge University Press.
De Brauwer \& Duyck, W. (2008). The SNARC effect in the processing of second-language number words: Further evidence for strong lexico-semantic connections. The Quarterly Journal of Experimental Psychology, 61, 444458.

Dehaene, S., Bossini, S., \& Giraux, P. (1993). The mental representation of parity and number quantity. Journal of Experimental Psychology: General, 122, 371-396.
Fias, W. (2001). Two routes for the processing of verbal numbers: evidence from the SNARC effect. Psychological Research, 65(4), 250-259.
Fischer, M. H., Castel, A. D., Dodd, M. D., \& Pratt, J. (2003). Perceiving numbers causes spatial shifts of attention. Nature Neuroscience, 6(6), 555-556.
Gevers, W., Verguts, T., Reynvoet, B., Caessens, B., \& Fias, W. (2006). Numbers and space: a computational model of the SNARC effect. Journal of Experimental Psychology: Human Perception and Performance, 32(1), 32-44.
Greenberg, J. (1966). Language universals, with special reference to feature hierarchies. The Hague: Mouton.
Haspelmath, M. (2006). Against markedness (and what to replace it with). Journal of Linguistics, 42(1), 25-70.
Landy, D. H., Jones, E. L., \& Hummel, J. E. (2008). Why spatial-numeric associations aren't evidence for a mental
number line. Proc. 30th Ann. Conf. Cogn. Sci. Soc., 357362.

Loetscher, T., Schwarz, U., Schubiger, M., \& Brugger, P. (2008). Head turns bias the brain's internal random generator. Current Biology, 18(2), 60-62.
Nuerk, H. C., Iversen, W., \& Willmes, K. (2004). Notational modulation of the SNARC and the MARC (linguistic markedness of response codes) effect. Quarterly Journal of Experimental Psychology, 57A, 835-863.
Nuerk, H. C., Wood, G., \& Willmes, K. (2005). The universal SNARC effect - The association between number quantity and space is amodal. Experimental Psychology, 52, 187-194.
Proctor, R. W., \& Cho, Y. S. (2006). Polarity correspondence: A general principle for performance of speeded binary classification tasks. Psychological Bulletin, 132, 416-442.
R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.ISBN 3-900051-07-0, URL http://www.R-project.org/.
Reynvoet, B., \& Brysbaert, M. (1999). Single-digit and twodigit Arabic numbers address the same semantic number line. Cognition, 72, 191-201.
Santens, S., \& Gevers, W. (2008). The SNARC effect does not imply a mental number line. Cognition, 108, 263-270.
Willmes, K., \& Iversen, W. (1995 April). On the Internal Representation of Number Parity. Paper presented at the Spring Annual Meeting of the British Neuropsychological Society, London.
Wood, G., Willmes, K., Nuerk, H.-C., \& Fischer, M. (2008). On the cognitive link between space and number: a metaanalysis of the SNARC effect. Psychology Science Quarterly, 50(4), 489-525.
Zimmer, K. (1964). Affixed negation in English and other languages: An investigation of restricted productivity. Word, 20(2), Monograph No. 5.

# Cross-Linguistically Shared and Language-Specific Sound Symbolism for Motion: An Exploratory Data Mining Approach 

Noburo Saji (nons@sfc.keio.ac.jp)<br>Japan Society for the Promotion of Science/Keio University, 4-1-1 Hiyoshi, Kohoku<br>Yokohama, Kanagawa 223-8521 Japan<br>Kimi Akita (akitambo@lang.osaka-u.ac.jp)<br>Studies in Language and Culture, Osaka University, 1-8 Machikaneyama-cho<br>Toyonaka, Osaka 560-0043 Japan<br>Mutsumi Imai (imai@sfc.keio.ac.jp)<br>Faculty of Environment and Information Studies, Keio University, 5322 Endo<br>Fujisawa, Kanagawa 252-0882 Japan

Katerina Kantartzis (k.f.kantartzis@bham.ac.uk), Sotaro Kita (s.kita@bham.ac.uk)<br>School of Psychology, University of Birmingham, Edgbaston Birmingham B15 2TT UK


#### Abstract

This paper demonstrates a new quantitative approach to identify what is behind universally sensed sound symbolism and sound symbolism detected only by speakers of a particular language. We presented 70 locomotion videos to Japanese and English speakers and asked them to create a word that would sound-symbolically match each action, then to rate the action on five semantic dimensions. Multivariate analyses revealed that certain sound-meaning links (e.g., voicing and speed) were more consistent than others within and across languages. Language-specific sound symbolism was also found for some sound-meaning links (e.g., the affricate manner of articulation was associated with light motions in Japanese, but with heavy motions in English). This implies that cross-linguistically shared and language-specific parts of sound symbolism are intricately intertwined within each language. This research underscores the importance of a bottom-up approach which can exploratorily investigate the complex sound-symbolic systems as a whole.


Keywords: sound symbolism; mimetics; canonical correlation analysis

## Introduction

Traditional linguistics has long assumed that the relationship between the form and meaning of a word is arbitrary (de Saussure, 1916/1983). However, words whose forms are motivated by their meanings (i.e., sound-symbolic words) are widely found across languages. For example, bump and thump sound like what they mean: events with an abrupt end (Firth, 1935/1957). Some languages have a large lexical class of sound-symbolic words called "ideophones," "expressives," or "mimetics" (Voeltz \& Kilian-Hatz, 2001; Kita, 1997). For example, Japanese is rich in not only onomatopoeic (e.g., piyopiyo 'tweet-tweet') but also nononomatopoeic mimetic words (e.g., tobotobo 'plodding').
Sound symbolism is not limited to ideophones and mimetics. Sapir (1929) points out that English speakers associate novel words containing the vowel /i/ with
smallness more frequently than words containing /a/. Another celebrated example of sound symbolism is the association between sonorancy and roundness (Köhler, 1929/1947). It has been repeatedly observed that speakers of many languages prefer a round shape for maluma and an angular shape for takete (Brenner, Caparos, Davidoff, Fockert, Linnell, \& Spence, 2013; Davis, 1961; Holland \& Wertheimer, 1964).

Thus, there has been accumulating evidence that language does contain some non-arbitrary sound-meaning correspondences and people are sensitive to them. However, the exact nature of sound symbolism has not been fully clarified and one of the most important questions about sound symbolism is still open: what sound-meaning associations are shared by speakers of different languages, and why? In fact, researchers have recognized that not every case of sound symbolism may be detected as commonly as maluma/bouba vs. takete/kiki.

For example, Iwasaki, Vinson, \& Vigliocco (2007) examined whether English speakers can detect the meanings of some Japanese mimetics depicting motion events, by asking them to rate the mimetics on a set of semanticdifferential scales (e.g., energetic vs. non-energetic; fast vs. slow). Iwasaki et al. demonstrated that English and Japanese speakers' ratings agreed on some dimensions but not others. Specifically, Japanese speakers associated mimetics starting with a voiced consonant with the meaning component of "big person," and the mimetics with voiceless consonants with "feminine" and "formal" walking. English speakers agreed only with the former association (see also Haryu \& Zhao, 2007 for the language-specific nature of magnitudevoicing symbolism).

## Limitations of Comprehension Tasks

The question of universal and language-specific facets of sound symbolism has not been properly addressed or pursued in previous studies, mainly due to the nature of
their experimental method. Most experimental studies on sound symbolism have aimed at detecting the universality of sound symbolism and mainly employed comprehension tasks, such as forced-choice and semantic-differential rating tasks. These experiments were designed to examine whether subjects can detect "correct" sound-meaning correspondences, or how they rate each sound or word on a predetermined set of semantic scales, such as size and brightness. These tasks are effective in the examination of particular sound-meaning associations. However, no one knows how many such associations-how many sound patterns, how many meaning dimensions, and how many combinations of sounds and meanings-we have to examine before we reach the whole picture of the sound-symbolic system of a language, let alone its universality.

## The Present Study

The goal of the present research was to extract crosslinguistically shared and language-specific parts of sound symbolism and to give phonological or phonosemantic explanations to them. We approach this issue by examining intuitions for sound symbolism in Japanese speakers and English speakers. To rectify the above mentioned limitations in using comprehension tasks, we employed a production method in which participants were asked to make mimetic words that matched human locomotions in short video clips. This method would reveal an unlimited set of phonologically and phonotactically possible phoneme sequences available to the subjects. We then conducted a multivariate analysis which detects underling correlations between sounds and sounds, meanings and meaning, and sounds and meanings, and evaluates what sound-meaning correlations are more significant than others in Japanese and English. The comparison of the detected sound-meaning pairs in each language shows us the shared and languagespecific sound symbolism.

We will present the Japanese and English speakers' data separately in Experiments 1 and 2 , respectively.

## Experiment 1: Japanese

## Method

Materials We created 70 short video clips of various types of human locomotion ( $M=7.3 \mathrm{sec}, S D=2.7$ ). In each video, a person appeared from the left side of the monitor and moved to the right out of the frame in a certain manner of walking or running.

Participants and Procedure Ninety-three native Japanese speakers, all undergraduate students, participated in the experiment. They went through both an attribute rating task and a word creation task. The participants were first presented with the 70 video clips on a computer screen in random order and asked to evaluate them on five 11-point semantic-differential scales (from 1 to 11): "size" (bigsmall), "speed" (slow-fast), "weight" (heavy-light), "energeticity" (energetic-non-energetic), and "jerkiness"
(jerky-smooth). After the rating task, the videos were shown again to the participants in a random order. They were asked to generate sound-symbolic words and type them on a keyboard.

Data Preparation For analysis, we excluded soundsymbolic words that were obviously made on the analogy of existent nouns and verbs (e.g., robo-robo, cf. robotto 'robot'). We also excluded the data obtained for the videos whose most common semantic rating was " 6 " (neutral), which we assumed to blur the rest of the data. A total of 1,442 mimetics were submitted for analysis. They were phonetically coded and listed with the rating scores. For phonetic coding, we limited ourselves to the first moras $\left(C_{1} V_{1}\right)$ of the obtained mimetics, as they have been discussed to have particular sound-symbolic significance (Kawahara, Shinohara, \& Uchimoto, 2008; see also Hamano, 1998). The coding scheme for consonants, shown with the one for vowels in Table 1 (the coding for English will be used in Experiment 2), is based on Bailey \& Hahn (2005). The data is thus a $1,442 \times 13$ matrix, consisting of five semantic ratings and eight phonetic values for each mimetic.

Table 1: The coding scheme for phonetic features

|  | Japanese | English |
| :---: | :---: | :---: |
| $\mathrm{C}_{1}$ place of articulation | labial (Lab), velar (Vel), alveolar (Alv), glottal (Glot), palate (Pal), dental (Dent) | labial, <br> velar, <br> alveolar, <br> glottal, <br> palate, <br> dental |
| $\mathrm{C}_{1}$ sonorancy | sonorant (Son), <br> obstruent (Obs) | sonorant, obstruent |
| $\mathrm{C}_{1}$ manner of articulation | stop (Stop), <br> affricate (Aff), <br> fricative (Fric), <br> glide (Gld), <br> flap (Flap) | stop, affricate, fricative, glide, lateral (Lat), nasal (Nas), rhotic (Rhot) |
| $\mathrm{C}_{1}$ voicing | voiced, voiceless | voiced, voiceless |
| $\mathrm{C}_{1}$ palatalization | palatalized, not palatalized | n/a |
| $\mathrm{C}_{1}$ nasality | nasal, not nasal | n/a |
| $\mathrm{V}_{1}$ height | high, mid, low | high, mid-high, mid-low, low |
| $\mathrm{V}_{1}$ backness | front, central, back | front, central, back |

## Analysis and Results

Canonical Correlation Analysis We conducted a variant of canonical correlation analysis (CCA) designed for categorical variables (see Thompson, 2005 for its detailed algorism) developed by Van der Burg (1988). Generally speaking, CCA enables us to visualize an implicit structure underlying multiple datasets. In common with other multivariate analyses, such as principle component analysis, CCA attempts to explain all possible correlations in a lowdimensional space. While principle component analysis is applied to only one dataset, CCA investigates relationships
among two or more different variable sets and derives estimates by applying weights to the variables.

In the current context, CCA examines all possible correlations both within and across the two sets of variables (i.e., the sound and meaning datasets). This means that we can explore not only sound-meaning associations but also sound-sound or meaning-meaning correlations simultaneously, not limiting ourselves to a predetermined set of sound-meaning pairs. Notice that this analytical method is meaningful due to the very nature of sound symbolism, in which sound and meaning are intertwined with each other.

The Consistency of Sound-Meaning Associations The data matrix was fed into the program for canonical correlation analysis packaged in IBM SPSS Statistics 20 (IBM, 2012). We employed a two-dimensional solution, as the canonical correlation values of the first and second dimensions, which represent the latent correlations between the canonical variable of sounds and that of meanings, were significantly high ( $r \mathrm{~s}=.56$ (first dimension) and .25 (second dimension), $p \mathrm{~s}<.001$ ). These values guarantee consistent sound-meaning associations in the two dimensions, indicating systematic sound symbolism in the present free production experiment.

The Focal Sound-Meaning Associations in the SoundSymbolic System of Japanese To examine how sound and meaning are correlated in the present dataset, we considered the component loadings of each variable (see Table 2). As in principle component analysis, component loadings represent the correlation between the data and the extracted dimensions; each absolute value approximates the importance of the variables on each dimension.

Table 2: Component loadings of canonical correlation
analysis in Japanese

| Dataset | Variable (positive - negative ) | Dimension 1 | Dimension 2 |
| :---: | :---: | :---: | :---: |
| Meaning | Size (large - small) | . 40 | -. 36 |
|  | Speed (slow - fast) | . 56 | . 39 |
|  | Weight (heavy - light) | . 85 | . 07 |
|  | Energeticity (energetic - non-energetic) | -. 21 | -. 56 |
|  | Jerkiness (jerky - smooth) | . 31 | -. 35 |
| Sound | $\mathrm{C}_{1}$ _place | . 05 | . 17 |
|  | $\mathrm{C}_{1}$ _sonorancy | . 36 | . 26 |
|  | $\mathrm{C}_{1}$ _manner | . 05 | -. 42 |
|  | $\mathrm{C}_{1}$ _voicing | . 74 | -. 24 |
|  | $\mathrm{C}_{1}$ ppalatalization | -. 43 | -. 05 |
|  | $\mathrm{C}_{1}$ _nasality | . 38 | . 29 |
|  | $\mathrm{V}_{1}$ _height | . 05 | -. 40 |
|  | $\mathrm{V}_{1 \_}$backness | . 28 | . 33 |

Table 2 shows that the semantic attribute "weight" in the meaning group and the phonetic feature " $\mathrm{C}_{1}$ voicing" in the sound group obtained high positive loadings on Dimension 1 (. 85 and .74 , respectively). This suggests that the voicingweight association was critically important in Japanese
sound symbolism for motion. Noteworthy contributions were also observed for "speed" (.56), "size" (.40), and "jerkiness" (.31) among the meaning features and " $\mathrm{C}_{1}$ palatalization" (-.43), " $\mathrm{C}_{1}$ nasality" (.38), and " $\mathrm{C}_{1}$ sonorancy" (.36) among the sound features. The four semantic variables were positively correlated. Heavy, large, slow, and jerky (or light, small, fast, and smooth) manners of motion tended to appear together, corresponding to the four consonantal features. On the other hand, in Dimension 2, "speed" (.39) and " $\mathrm{V}_{1}$ backness" (.33) obtained high positive absolute values, while "size" ( -.36 ), "energeticity" (-.56), "jerkiness" (-.35), " $\mathrm{C}_{1}$ manner" ( -.42 ), and " $\mathrm{V}_{1}$ height" ( -.40 ) obtained high negative absolute values. This suggests that the correspondences between this set of consonantal and vocalic features and slow, small, nonenergetic, and smooth (or large, fast, energetic, and jerky) manners have the second most important status in Japanese sound symbolism for motion.

Details of the Sound-Meaning Associations The loading scores tell us which variables (e.g., manner of articulation) play a primary role in the discrimination of the dimensions, but it does not specify how much individual values in each variable (e.g., "affricate" in manner of articulation) contribute to those dimensions. We therefore computed the centroids of object scores for the semantic and phonetic values (see Van der Burg, 1988 for the details of this algorism). Specifically, each point in Figure 1 represents the weight of each value on the two dimensions. Note that the figure only shows sound variables for the sake of clarity; relevant meaning variables are indicated in the dimension labels, based on their loading scores above. The all abbreviations in Figure 1 is corresponds to those in Table1.

First, it is evident that the "voiced" and "voiceless" points are contrastively located in the positive and negative sides of Dimension 1, respectively. This is consistent with the large contribution of the voicing feature to this dimension in component loading. Moreover, Figure 1 reveals large positive contributions of the two phonetic values, "nasal" and "sonorant," to the same dimension, although the component loadings of the " $\mathrm{C}_{1}$ nasality" and " $\mathrm{C}_{1}$ sonorancy" variables were not as large as that of " $\mathrm{C}_{1}$ voicing." These coordinates indicate that voiced consonants that are nasal and sonorant (i.e., $[\mathrm{m}],[\mathrm{n}]$, as in moji and noro) have particular significance in Dimension 1. In contrast, the negative half of Dimension 1 features the voiceless obstruent that is "palatalized" and "affricate" (i.e., /ty/, realized as [tf]) as a sound that is strongly associated with a small, fast, light, smooth motion (e.g., tyoko).

Dimension 2 also shows clear contrasts for the variables which received high loading scores in Table 2: "fricative" and "affricate" ( $\mathrm{C}_{1}$ manner), "high" and "low" ( $\mathrm{V}_{1}$ height), and "back" and "central" ( $\mathrm{V}_{1}$ backness). Each of these contrasts is paired with a set of positive (slow, small, nonenergetic, smooth) or negative semantic values (fast, large, energetic, jerky) in Figure 1. The positive half of the same figure also contains "glottal," "palatal," "sonorant," and
"nasal." These results allow us to think of particular phones to be relevant to the present case of sound symbolism, such as [ $\mathrm{h}(\mathrm{j})]$ (fricative, glottal, (palatal)), [m] and [ n$]$ (nasal), and [u] (high, back) (e.g., heto, hura, moso). Similarly, in the negative half of Dimension 2, [tf] and [ts] (affricate) and [a] (central, low) are associated with large, fast, energetic, and jerky manners of motion (e.g., tyaki).


Figure 1: Category centroids for individual phonetic values in Japanese (See Table 1 for the explanations for the abbreviations)

## Experiment 2: English

## Method

Participants and Materials Twenty-seven English speakers at University of Birmingham, UK, participated in the experiment. The same 70 videos as we used in Experiment 1 were used as stimuli.

Procedure As in Experiment 1, English participants first saw randomly presented videos and were asked to rate them on five semantic dimensions. After the rating task, they watched the videos again and produced sound-symbolic words to depict the human motions shown in the video clips. Unlike Experiment 1, however, the participants were instructed to create $\mathrm{C}_{1} \mathrm{~V}_{1} \mathrm{C}_{2} \mathrm{~V}_{2}$ words that intuitively (or "sound-symbolically") matched the motions. This change was made because English speakers were not likely to be familiar with the notion of mimetics.

Data Preparation The data went through the same noise exclusion procedure as in Experiment 1. 1,227 "mimetic" words were retained for analysis. The $\mathrm{C}_{1} \mathrm{~V}_{1}$ of each mimetic was phonetically coded according to the scheme in Table 1. Thus, the resultant data matrix consisted of 1,227 rows of mimetics and 8 columns of phonetic/evaluative features.

## Analysis and Results

The Consistency of Sound-Meaning Associations Nonlinear canonical correlation analysis was conducted with the

English data matrix. We adopted a two-dimensional solution. The canonical correlation values for Dimensions 1 and 2 were .17 and .15 , respectively ( $p s<.01$ ). These values were substantially lower than their Japanese equivalents, indicating that the associations between the sound and meaning datasets in English are relatively weaker than those in Japanese. This may suggest that Japanese speakers have a better established sound-symbolic sense than English speakers due to the existence of the sound-symbolically systematized lexical class of mimetics in Japanese.

The Focal Sound-Meaning Associations in SoundSymbolic System of English The component loadings of each variable are listed in Table 3. It shows that "size" $(-.40)$, "speed" (.56), "energeticity" $(-.62)$, " $\mathrm{C}_{1}$ voicing" (.58), and " $\mathrm{V}_{1}$ height" ( -.39 ) obtained high absolute values in Dimension 1, while "weight" (.47), "energeticity" (-.32), "jerkiness" ( -.46 ), and " $\mathrm{C}_{1}$ place" ( -.70 ) were heavily weighted in Dimension 2. Thus, Dimension 1 is associated with small, slow, non-energetic motions, and Dimension 2 with heavy, non-energetic, smooth motions.

Table 3: Component loadings of canonical correlation analysis in English

| Dataset | Variable (positive - negative) | Dimension 1 | Dimension 2 |
| :---: | :---: | :---: | :---: |
| Meaning | Size (large - small) | -. 40 | . 12 |
|  | Speed (slow - fast) | . 56 | . 06 |
|  | Weight (heavy - light) | -. 11 | . 47 |
|  | Energeticity <br> (energetic - non-energetic) | -. 62 | -. 32 |
|  | Jerkiness (jerky - smooth) | -. 09 | -. 46 |
| Sound | $\mathrm{C}_{1}$ _place | . 04 | -. 70 |
|  | $\mathrm{C}_{1}$ _sonorancy | . 27 | . 07 |
|  | $\mathrm{C}_{1}$ _manner | . 18 | -. 19 |
|  | $\mathrm{C}_{1}$ _voicing | . 58 | -. 03 |
|  | $\mathrm{V}_{1}$ _height | -. 39 | -. 15 |
|  | V1_backness | . 07 | . 12 |

Details of the Sound-Meaning Associations Figure 2 plots the centroids of object points, which indicate how each value of the sound/meaning categories was weighted. Dimension 1 is clearly divided by the two phonetic features " $\mathrm{C}_{1}$ voicing" and " $\mathrm{V}_{1}$ height," with "voiced" and "mid-low" being positive and "voiceless" and "high" being negative. The figure also contains "nasal," "lateral," "rhotic," and "sonorant" in the positive area, suggesting that [n], [1], and [ r ], as in medi, lela, and reso, are strongly connected with small, slow, non-energetic motion. Likewise, the negative domain contains a voiceless glottal fricative (i.e., [h], as in hali), which was associated with large, fast, energetic motion.

Dimension 2 exhibits a wide distribution of the places and manners of articulation. A marked contrast is observed between the two positive phonetic features "glottal" and "affricate" and the three negative ones "palatal," "velar," and "glide." Among these features, "glottal" and "affricate" can be unambiguously identified as [h] (e.g., hopi) and [t]], respectively, which are linked with heavy, non-energetic,
smooth motion. Similarly, the combination of "palatal" and "glide" is synonymous with [j].


Figure 2: Category centroids for individual phonetic values in English (See Table 1 for the explanations for the abbreviations)

## General Discussion

Comparing the detected sound-meaning correlations in Japanese and English shows us the shared and languagespecific sound symbolism. Table 4 summarizes the soundsymbolic mappings found in the two languages, which shows what sound and meaning components have priority in motion sound symbolism of the two languages.
First, the results suggest that the two languages share in a large part a set of "sound-symbolically relevant" phonetic features. For example, both languages utilized the phonetic features "sonorancy," "voicing," "nasality," and "vowel height," and specific phonetic values "glottal," "palatal," "affricate," and "fricative." It should be noted here that some phonetic values, such as "alveolar," "labial," and "stop," did not make a large contribution in the present data. This might reflect the unmarked nature of these sounds in the phonological systems of the two languages. In the present data, alveolar, labial, and stop consonants were found in $66 \%, 16 \%$, and $61 \%$ of all Japanese mimetics and $51 \%, 28 \%$, and $37 \%$ of all English mimetics, respectively.
Second, the two languages share many semantic features in their primary sound symbolism. Most notably, both of them use "weight" and "energeticity" as the most significant semantic features in sound symbolism of manner of motion. The two features are correlated with "size" and "speed" (see Tables 2 and 3).
Thus, speakers of Japanese and English use a similar set of phonetic and semantic features in sound symbolism of locomotion. However, these similarities do not directly mean that English and Japanese speakers mapped these sounds and meanings in the same way. They shared the
most important sound-symbolic mapping: the voicing-speed mapping in the primary dimension. This can be accounted for by the long VOT (voice onset time) of voiced consonants, which appears to be readily mapped to the long duration of slow motion. Phonosemantic descriptions in the literature support this interpretation (Hamano, 1998; Tamori \& Schourup, 1999). Further, the present study revealed that this sound-symbolic effect is especially strong in nasals (i.e., [m], [n]).

Table 4: Sound-meaning associations obtained in the two experiments

| Dimension | Japanese | English |
| :---: | :---: | :---: |
| Dimension 1 | heavy, slow, jerky, large voiced, nasal + sonorant , | non-energetic, <br> slow, small <br> voiced, <br> nasal + sonorant, <br> lateral , <br> rhotics, <br> mid-low |
|  | light, fast, smooth, small Voiceless, palatalized + affricate | energetic, <br> fast, large, <br> voiceless , <br> glottal + fricative, <br> high |
| Dimension 2 | small, <br> slow, non-energetic, smooth <br> glottal + fricative, <br> palatal + fricative, <br> nasal + sonorant, <br> high + back | heavy, <br> non-energetic, smooth glottal, affricate |
|  | large, fast, energetic, jerky central, low vowel, affricate | light, energetic, jerky palatal + glide, velar |

Note: Sound-meaning associations shared by the two languages are given in boldface.

The present results also established the presence of language-specific sound symbolism. Most strikingly, Japanese and English speakers mapped some sounds to opposite meanings. For example, Japanese speakers associated the palato-alveolar affricate [t 5 ] with light, fast, smooth, small motion in the primary dimension, whereas English speakers linked it to heavy, non-energetic, smooth motion in the secondary dimension. Likewise, the high back vowel /u/ was connected to slow, non-energetic motion in Japanese, but to fast, energetic motion in English.
These disagreements may be explained by the crosslinguistic differences in the phonological status of these sounds. First, in Japanese, the phone [t $\int$ ] often appears secondarily, in a palatalized environment (i.e., /ty/), whereas this is not the case in English. Moreover, another affricate in Japanese (i.e., [ts]) is analyzed into [t] and [s] in English. Second, /u/ is realized as [ w ] (unrounded) in Japanese, but as [ $\mathrm{\sigma}$ ] (rounded) or [ $\mathrm{u}(:)$ ] (rounded) in English. This crosslinguistic contrast in the roundedness of high back vowels suggests an articulation-based link between roundedness and energetic (hence, fast) motion. Thus, the present comparative observation illustrates the possibility that at least some parts of language-specific sound symbolism may
be accounted for in terms of phonological typology. This possibility has been assumed widely in the literature, yet has not been much investigated.
Our study provides some important insights for theories of sound symbolism. We revealed that the sense of sound symbolism is realized in a complex system, which involves both universality and language-specificity. Sound symbolism is often alluded as "phonetic iconicity," but despite the name, this linguistic phenomenon is subject to a certain degree of arbitrariness, which originates from our language experience (Brenner et al., 2013 for a similar discussion). Our holistic and exploratory approach greatly contributes to the clarification of the complexity of iconic and arbitrary mappings in sound symbolism.

## Acknowledgements

We are grateful to Shigeto Kawahara for his insightful and helpful comments from the perspective of phonetics and phonology. We also thank Farzana Bhaiyat, Alicia Griffiths, Junko Kanero, and Yayoi Tajima, who participated in the current project as annotators. Remaining inadequacies are ours. This research was supported by MEXT/JSPS KAKENHI to Saji (\#232913), Akita (\#24720179), Imai (\#15300088, \#22243043, \#23120003), and Biological Sciences Research Council'[curly apostrophe]s Research Development Fellowship [BB/G023069/1] to Kita..

## References

Bailey, T., \& Hahn, U. (2005). Phoneme similarity and confusability. Journal of Memory and Language, 52, 339362.

Brenner, A. J., Caparos, S., Davidoff, J., Fockert, J., Linnell, K. J., \& Spence, C. (2013). "Bouba" and "kiki" in Namibia? A remote culture make similar shape-sound matches, but different shape-taste matches to Westerners. Cognition, 126, 165-172.
Davis, R. (1961). The fitness of names to drawings. British Journal of Psychology, 52, 259-268.
de Saussure, F. (1916/1983). Course in General Linguistics (trans. R. Harris). London: Duckworth.
Firth, J. R. (1935/1957). The use and distribution of certain English sounds. In Papers in Linguistics 1934-1951, 3446. London: Oxford University Press.

Hamano, S. (1998). The Sound-Symbolic System of Japanese. Stanford: CSLI Publications.
Haryu, E., \& Zhao, L. (2007). Understanding the symbolic values of Japanese onomatopoeia: Comparison of Japanese and Chinese speakers. Japanese Psychological Research, 78, 424-432.
Holland, M., \& Wertheimer, M. (1964). Some physiognomic aspects of naming, or maluma and takete revisited. Perceptual and Motor Skills, 19, 111-117.
IBM Corp. (2012). IBM SPSS Statistics for Windows, Version 21.0. Armonk, NY: IBM Corp.

Iwasaki, N., Vinson, D. P., \& Vigliocco, G. (2007). What do English speakers know about gera-gera and yota-yota? A cross-linguistic investigation of mimetic words for laughing and walking. Japanese-Language Education around the Globe, 17, 53-78.
Kawahara, S., Kazuko S., \& Uchimoto, Y. (2008). A positional effect in sound symbolism: An experimental study. Proceedings of the Eighth Annual Meeting of the Japanese Cognitive Linguistics Association, 8, 417-427.
Kita, S. (1997). Two-dimensional semantic analysis of Japanese mimetics. Linguistics , 35, 379-415.
Köhler, W. (1929/1947). Gestalt Psychology: An Introduction to New Concepts in Modern Psychology. New York: Liveright.
Sapir, E. (1929). A study in phonetic symbolism. Journal of Experimental Psychology, 12, 225-239.
Tamori, I., \& Schourup, L. (1999). Onomatopoeia: Its form and meaning. Tokyo: Kurosio Publishers.
Thompson, B. (2000). Canonical correlation analysis. In L. Grimm \& P. Yarnold (eds), Reading and Understanding More Multivariate Statistics, 285-316. Washington, DC: American Psychological Association.
Van der Burg, E. (1988). Homogeneity analysis with $k$ sets of variables: An alternating least squares method with optimal scaling features. Psychometrika, 53, 177-197.
Voeltz, F. K. E., \& Kilian-Hatz, C. (eds) (2001). Ideophones. Amsterdam/Philadelphia: John Benjamins.

# Polarity correspondence does not explain the SNARC effect 

Julio Santiago (santiago@ugr.es)<br>Department of Experimental Psychology, University of Granada<br>Campus de Cartuja, 18071 Granada, Spain<br>Daniel Lakens (D.Lakens@tue.nl)<br>School of Innovation Sciences, Eindhoven University of Technology<br>Postbus 513, 5600 MB Eindhoven, The Netherlands


#### Abstract

Conceptual metaphor congruency effects have been interpreted as evidence for the notion that the representation of abstract conceptual dimensions (e.g., power, evaluation) rests on more concrete dimensions (e.g., space, brightness). However, an alternative account based on the notion of polarity correspondence has recently received empirical support from studies about the mapping between affective evaluation and morality on vertical space. We tested the polarity correspondence account in the domain of number, which shows well-known congruency effects with lateral left-right space (the SNARC effect). Response polarity was manipulated by varying keyboard eccentricity in both parity (odd-even) and quantity (larger-smaller than 5) tasks. Response eccentricity did not modulate the SNARC effect. In a final experiment, the orthogonal Simon effect was modulated by the manipulation of response eccentricity. We conclude that polarity correspondence does not provide an adequate explanation of conceptual congruency effects in the domain of number.


Keywords: conceptual metaphor; polarity correspondence; SNARC; number; space; response eccentricity; orthogonal Simon.

## Introduction

Recent years have witnessed a strong interest in the possibility that the mental representation of abstract concepts relies in a deep sense on more concrete concepts (see Landau, Meier, \& Keefer, 2010, and Santiago, Román, \& Ouellet, 2011, for reviews). Under this view, an abstract concept or conceptual domain imports structure and content from a better understood, more clearly delineated, more concrete conceptual domain. For example, time is understood as physical motion from one location to another (Clark, 1973), and power is linked to physical size (Sorokowski, 2010). Such a view suggests that the mental representation of concepts is hierarchically structured, such that more concrete concepts are more directly linked to perceptual-motor experiences, and these in turn are used to support the understanding of more abstract levels. Therefore, the whole human conceptual structure is anchored to, or founded upon our embodied interaction with the external world, which is why Santiago, Román, and Ouellet (2011) called it the Solid Foundations view.

An important source of evidence for such a view comes from conceptual congruency tasks. In these tasks, a concrete and an abstract dimension are factorially crossed.

Participants' main task requires the processing of the abstract dimension, and the effects of the concrete, task-irrelevant dimension are measured. Typically, both dimensions interact, such that a particular combination of concrete and abstract conditions shows better performance. This metaphorical congruency effect is often interpreted as revealing the use of underlying concrete representations to support the abstract judgment. A well-known example is the Spatial-Numerical Association of Response Codes (SNARC) effect (Dehaene et al., 1993). In a typical SNARC task, the participant has to make a numerical discrimination, such as deciding whether a number is odd or even, by means of key presses. The responding hand (left or right) is the task-irrelevant concrete dimension: in some blocks the "odd" response is given by a left-hand key press and the "even" response by a right-hand key press. In other blocks, the mapping is reversed. The standard result, now widely replicated, consists in better performance when responding to a small number with the left hand and to a large number with the right hand versus using the reverse mapping (see Wood, Willmes, Nuerk, \& Fischer, 2008, for a review). The SNARC effect has most often been interpreted as evidence for the use of a spatial left-right line to mentally represent number magnitude.

However, Lakens and coworkers (Lakens, Semin, \& Foroni, 2012; Lakens, 2012) have proposed that it may not be necessary to resort to concrete representations of any kind in order to account for many of the published metaphorical congruency effects. Their view rests on purely structural features of dimensional concepts based on the concept of markedness and on the principle of polarity correspondence proposed by Proctor and Cho (2006), which applies when two or more dimensions are simultaneously manipulated in a task. The concept of markedness has a long standing tradition in linguistics (Greenberg, 1963) and psycholinguistics (Clark, 1969). The two poles of most conceptual dimensions (e.g., happiness or tallness) do not seem to enjoy the same representational status. One of them, which we will refer to as the + pole, is used to refer to the whole dimension, whereas the other, the -pole, is used to refer only to itself (e.g., compare "how tall is John?" versus "how short is John?": the former does not presuppose that John's height is in any specific range, whereas the second question suggests that John is short). The +pole is more frequent in language and enjoys a processing advantage compared to the -pole (Clark, 1969). Proctor and Cho
(2006) proposed a polarity correspondence principle that predicts an extra processing advantage in those conditions where the pole signs match. Thus, a polarity correspondence account of conceptual congruency effects does not require the postulation of concrete mental representations.

One key piece of evidence for Proctor and Cho's argument (2006) relies on how response eccentricity modulates the orthogonal Simon effect. In a typical orthogonal Simon task, participants are presented with a stimulus in one of two vertical locations (e.g., above or below fixation) and are asked to discriminate its location by means of a left or right key press or a leftward or rightward response. The standard result is that the mapping of the upper location with the right response and down with left produces better performance than the up-left down-right mapping (e.g., Proctor \& Cho, 2003). Proctor and Cho (2006) proposed that the up-right advantage is due to polarity correspondence, being up and right the +poles of the vertical and lateral spatial dimensions, respectively.

Supporting this conclusion, response eccentricity interacts with the up-right advantage. Response eccentricity refers to the lateral displacement of the response set. Placing the response box, keyboard, or joystick to the right of the screen makes the up-right advantage to grow stronger, while it turns into an up-left advantage when the response set is located on left space (Proctor \& Cho, 2003). Proctor and coworkers suggested that response eccentricity changes the saliency of the right and left poles of the lateral spatial dimension, effectively turning the left pole into the +pole when the responses are placed on left space and thus generating the up-left advantage through polarity correspondence.

These results illustrate an important characteristic of the polarity correspondence account: polarities are not fixed, but can be changed by attentional and saliency factors, which opens the possibility of manipulating them experimentally. Lakens (2012) and Lakens et al. (2012) applied this perspective to conceptual congruency effects between the concrete dimensions of vertical location (up-down) and brightness and the abstract dimensions of power and affective evaluation. They showed that those conceptual congruency effects require the simultaneous presence of the two contrasting poles in the task and that it is possible to change the effect by changing the frequency of use of each pole.

To summarize the argument so far, there seems to be good support for the idea that conceptual congruency effects are of a flexible and contextual nature (Lakens et al., 2012; see also Santiago, Ouellet, Román, \& Valenzuela, 2012), thus contradicting their interpretation as indexes of stable semantic memory mappings favoured by the Solid Foundations view. There is also evidence that suggests that polarity differences and cross-dimensional polarity correspondence can account for some conceptual congruency effects without resorting to underlying concrete dimensions, or even to any internal alignment of the relevant dimensions. However, the relevant evidence so far
has concentrated on a small set of conceptual dimensions, namely morality, power and affective evaluation. Can these conclusions be generalized to other abstract dimensions? Such is the question that we seek to answer in the present experimental series.

In order to increase the contrast with already available studies, and therefore the generalizability of our results, we decided to test the abstract dimension of numerical magnitude. In contrast to morality, power and evaluation, which are thought to be associated to vertical space, numbers have been linked to lateral left-right space. Proctor and Cho (2006) explicitly argued that polarity correspondence might explain the standard SNARC effect that obtains in parity tasks. They argued that number processing in parity judgment does reveal markedness effects. Responding "odd" is slower than responding "even", especially in contexts that foster a comparison between odd and even numbers (Hines, 1990). Moreover, responding "odd" with the left hand and "even" with the right hand is faster than vice versa (Nuerk, Iversen, \& Willmes, 2004, called it the Markedness Association of Response Codes, or MARC effect). The SNARC effect would arise because large numbers would be +polar and small numbers -polar and these polarities would match with the +polar right response and the -polar left response. Therefore, Experiment 1 used a parity judgment task on Arabic numerals.

In contrast to parity tasks, Proctor and Cho (2006) suggested that magnitude comparison tasks (e.g., to say whether the number is smaller or larger than 5) induce a continuous representation and, as a result, neither a SNARC nor MARC effect are observed in them (they cite Ito \& Hatta, 2004, Experiment 3, as a relevant case). However, other studies have successfully reported SNARC effects in magnitude judgments, including some which have interpreted their results as support for the polarity correspondence hypothesis (Santens \& Gevers, 2008). We believe that, if polarity correspondence effects underlie the SNARC effect in parity judgments, they should be even clearer when the task explicitly asks participants to categorize the stimuli into the two polar opposites "small" and "large". So, we decided to extend our observations to a magnitude judgment task in Experiment 2.

The rationale of the current set of experiments relies on the manipulation of response eccentricity, thereby influencing the polarities of the left or right hand response alternatives. We manipulated response eccentricity in a procedure that closely followed the procedure used by Proctor and Cho (2003, Experiment 1), but instead of presenting stimuli in upper or lower locations, digits were presented in the centre of the screen. Participants discriminated their parity (Experiment 1) or magnitude (Experiment 2) by means of left or right key presses. Response set location was manipulated within-participants by placing the input device left, centre, or right of the computer monitor. If the left-right numerical and/or temporal congruency effects are due, in all or in part, to
polarity correspondence, then changing the polarity of the response dimension should influence the observed reaction time pattern. The polarity correspondence account predicts standard congruency effects when the response set is either central or to the right. Crucially, it predicts a reduced or inverted effect when the response set is located on the left side. Under this condition, the left response would be +polar and the right response -polar. Therefore, the polarity correspondence principle should induce an advantage for up-left and down-right mappings. To preview the results, response eccentricity failed to interact with the SNARC effect in any task. We concluded this series of experiments by showing that it is possible to replicate the interaction between response eccentricity and the orthogonal Simon effect when upper and lower stimuli are used instead of numbers (Experiment 3).

## Experiment 1

Experiment 1 used centrally presented single digits from 1 to 9 (with the exception of 5) in a parity judgment task. Responses were left and right key presses in a keyboard which could be located at either the left, centre, or right of the screen.

## Method

Participants Twenty Psychology students (17 women, all right-handed, age range $18-39 \mathrm{y}$.) from the University of Granada volunteered to participate, and received course credit in return.
Materials and Procedure The single digits 1 to 4 and 6 to 9 were centrally presented on a computer screen. A single digit was presented in each trial, and the participant's task was to decide whether the digit was odd or even. The participant responded by pressing the keys F (left) or J (right) on a standard computer keyboard. Each trial consisted of a central fixation cross ( 1000 ms ) followed by the target digit (until response). Incorrect trials were followed by the word "Incorrecto" for 500 ms in red font, also at the same location. Each was followed by a 1000 ms blank screen. Experimental trials were divided into six blocks of 54 trials. All digits were presented once every eight trials. The mapping of responses (odd-even) to keys (left-right) was kept constant during three blocks, and then reversed in the following three blocks. The order of presentation of the two mappings was counterbalanced over participants. Keyboard location was varied within-participants. In the left and right locations, the keyboard was moved 30 cm to each side (as in Proctor \& Cho, 2003). Half the participants experienced the three locations in the order left, centre, and right, and the other half in the reversed order. The sequence of keyboard locations was repeated twice over the six blocks.
Design and Analysis Data were analyzed using a factorial ANOVA with the following factors and levels:

Parity (odd-even) X Magnitude (smaller-larger than 5) X Response (left-right) X Keyboard location (left-centre-right). Counterbalance group was also included as a factor in the design in order to exclude noise due to order of conditions, but because its effect and interactions are theoretically uninteresting, they will not be reported below (but note that the inclusion of this factor in the analyses leads to a reduction in the degrees of freedom of the error). Markedness effects would be evident in the main effects of Parity (with faster responses to even than odd numbers) and Magnitude (with faster responses to larger than smaller numbers). A main effect of Response (with faster responses with the right than left hand) could also be interpreted as a markedness effect, although it could just be due to greater perceptuo-motor fluency of the preferred hand. Potential polarity correspondence effects would consist in significant interactions between Parity and Response (MARC effect) and Magnitude and Response (SNARC effect). Three-way interactions of either the MARC and/or the SNARC effects with Keyboard location would support the conclusion that polarity correspondence is indeed their underlying cause.

## Results

Errors occurred in 280 trials (4.32\%). Latencies in correct trials were trimmed by means of fixed cut-off points, set at 300 and 1300 ms after inspection of the RT distribution, which led to the rejection of 100 trials $(1.54 \%)$ as outliers. Average latency and accuracy were analyzed independently.

The analysis of latency showed a very clear pattern. There were main effects of Parity $(F(1,16)=13.52, p<.01)$ and Response $(F(1,16)=9.30, p<.01)$, both in the direction predicted by a markedness effect (faster latencies for even and right-handed responses). There was a clear interaction between Magnitude and Response $(F(1,16)=8.12, p=.01)$, which took the standard form of the SNARC effect. No other effect had a probability level smaller than 0.10 . In particular, the interaction between Parity and Response (the MARC effect) and any second order interactions between either the SNARC and the MARC effects with Keyboard location were far from significance (all $F \mathrm{~s}<1$ ). Figure 1 (upper panel) shows the main results. These results were not qualified by the analysis of accuracy.

## Discussion

Experiment 1 showed that keyboard location did not modulate the SNARC effect, what contradicts the predictions from a polarity correspondence account. The data also showed main effects of parity and response side such that the + pole (even and right) was easier to process. However, there were no interactions between these dimensions and keyboard location. Moreover, there was neither an interaction between parity and response side
(MARC effect) nor a modulation of this interaction at different keyboard locations.


Figure 1: Average latencies (in ms) in the congruent vs. incongruent conditions in Experiment 1 (SNARC, parity task), Experiment 2 (SNARC, magnitude task), and Experiment 3 (orthogonal Simon task). Congruency in the SNARC experiments is defined with respect to numerical magnitude (smaller or larger than 5) and response (left or right), assuming a standard SNARC effect (smaller-left, larger-right). Congruency in the orthogonal Simon experiment is defined with respect to vertical location (up or down) and response (left or right), assuming a standard orthogonal Simon effect (up-right, down-left).

Thus, the full pattern of results shows no support for a polarity correspondence account. Experiment 2 extended these results to a magnitude judgement task.

## Experiment 2

Experiment 2 was an exact replication of the Experiment 1 with a single difference: participants judged whether the central digit was smaller or larger than 5.

## Method

Participants Twenty new participants (17 women, 1 left-hander, age range $18-39$ y.) from the same population took part in the study and received credit course in return.
Materials and Procedure Everything was kept identical to Experiment 1 with the main exception of the task: participants judged whether the digit was smaller or larger than 5. Additionally, there were 56 trials in each block (exactly 7 presentations of each digit).
Design and Analysis Data were analyzed using a factorial ANOVA comprising Parity (odd-even) X Magnitude (smaller-larger than 5) X Response (left-right) X Keyboard location (left-centre-right) X Counterbalance group (not reported).

## Results

There were errors in 216 trials (3.32\%). Latencies of correct trials were trimmed by fixed cut-off points ( 300 and 1300 $\mathrm{ms})$ : 120 trials ( $1.78 \%$ ) were rejected as outliers.

The analysis of latencies once again rendered a clear pattern. Only the main effect of Parity was significant $(F(1,16)=20.34, p<.001)$, with faster responses to even numbers. Out of all possible two-way interactions, only Magnitude by Response was significant $(F(1,16)=6.55$, $p=.02$ ), taking the shape of a standard SNARC effect. These were the only findings that reached standard reliability levels. Keyboard location did not modulate the SNARC effect at all. Figure 1 (middle panel) shows the main results.

The accuracy measure showed a clear effect of Parity $(F(1,16)=10.15, p<.01)$ which went in an unexpected direction: even numbers had more errors than odd numbers. However, an inspection of accuracy means for each number indicated that this is due to a distance effect that concentrates on the two numbers that surround 5 (4 and 6), which are both even. Because this reduced accuracy was associated to faster latencies, it may be revealing a speed-accuracy trade-off. There was also a significant interaction between Parity and Response $(F(1,16)=7.32$, $p=.02$ ), which was in the direction expected for a MARC effect: greater accuracy for the even-right and odd-left mapping. No other effect added to or qualified the findings from the latency analysis.

## Discussion

Experiment 2 found a clear SNARC effect. In contrast to the prior experiment, there was also a MARC effect on accuracy. None of those potential polarity correspondence effects where modulated by the location of the keyboard. There was also a main effect of parity, both in latency and accuracy, but its interpretation is complicated because it went in opposite directions (even numbers were both faster and less accurate), what may suggest a speed-accuracy trade-off.

Summing up, Experiments 1 and 2 found SNARC effects which did not show any trace of being modulated by keyboard location. Other significant main effects and interactions also failed to provide clear indications of being related to markedness or polarity correspondence. Before turning to discuss the general implications of the present results, a final possibility must be discarded: that eccentricity effects on orthogonal Simon tasks cannot be replicated. Thus, our final experiment used a procedure that mirrored Proctor and Cho (2003, Experiment 1).

## Experiment 3

Experiment 3 was an exact replication of our two first experiments with the single difference that the stimulus was a rectangle made of asterisks (as in Proctor \& Cho, 2003, Experiment 1), which could be presented either above or below fixation. Participants' task was just to discriminate its location by pressing the right or left key.

## Method

Participants Participants were 18 Psychology students from the University of Granada (all female, 2 left-handers, age range $18-30 \mathrm{y}$.), who received course credit for their participation.
Materials and Procedure The target stimulus was an array of $3 \times 3$ asterisks that looked like a rectangle. The target was presented horizontally centred midway between fixation and either the upper or lower border of the screen. Participant's task was to judge whether the target appeared above or below fixation.
Design and Analysis The design included Vertical location (up-down) X Response (left-right) X Keyboard location (left-centre-right) X Counterbalance group (not reported).

## Results

There were errors in 137 trials (2.26\%). Cut-off points were set at 250 and 1250 ms , which led to the rejection of 119 (1.85\%) outliers.

The analysis of latency revealed an interaction between Vertical location and Response $(F(1,14)=11.05, p<.01)$. Unexpectedly, this interaction took the form of an up-left advantage (possible causes are discussed below). However, the crucial aspect of the data is that such interaction was strongly modulated by Keyboard location $(F(2,28)=25.73$,
$p<.001$; see Figure 1, lower panel). The up-left advantage was present when the keyboard was located on the left and on the centre, and turned into a slight up-right advantage when the keyboard was moved to the right. Other significant findings of less theoretical importance were the interaction between Vertical location and Keyboard location $(F(2,28)=5.56, p<.01)$ and the main effect of Keyboard location $(F(2,28)=6.81, p<.01)$ due to slower latencies with the keyboard on the left. Accuracy data supported the findings of the latency measure.

## Discussion

Experiment 3 allows very clear conclusions: an up-left advantage was observed both in latency and accuracy when the keyboard was placed at midline and on the left, and this turned into a small up-right advantage when the keyboard was placed to the right of the computer monitor. Proctor and Cho (2003) found an up-left advantage with the keyboard on the left, a very small up-right advantage with keyboard on the centre, and an up-right advantage with the keyboard on the right side. Therefore, we take the present results to constitute a successful replication of their findings (as well as those by Cho, Proctor, \& Yamaguchi, 2008): keyboard eccentricity affects the saliency of the side of space where the keyboard lies, and the most salient side of the left-right dimension attracts the mapping of the +pole of the vertical dimension (up).

The main contrast between present results and those reported by Proctor and Cho (2003) and others is the finding of an up-left (instead of up-right) advantage when the response set is placed at midline. As a post-hoc speculation, we think that the cause may be related to the spatial arrangement of the experimental equipment with respect to the whole room. The computer and keyboard were located on a corner of the lab, with a window to the right of the participant, and the room extending to the left. This may have made the participant to conceptualize the equipment as being located to the left of the wall. Both Weeks, Proctor, and Beyak (1995) and Proctor and Cho (2003) have shown that environmental factors can increase the saliency of either left or right space: placing an unused response apparatus to the right of the response keyboard was enough to turn the up-right advantage into a (very small) up-left advantage. A similar phenomenon may have occurred in the present experiment with the highly salient window located on the right side of the participant. Alternatively, the wall on the right may have provided a clear boundary to lateral space, which may have made the unbounded left space +polar, conforming existing explanations for the asymmetry in vertical space (Clark, 1973; Tversky, Kugelmass, \& Winter, 1991).

## General Discussion

The present experiment series clearly showed that the SNARC effect, both in parity and magnitude judgements, is not modulated by response eccentricity. This occurred in the context of a successful modulation of the orthogonal Simon
effect by our manipulation of response eccentricity. Proctor and coworkers (Cho et al., 2008; Proctor \& Cho, 2003, 2006; Weeks et al., 1995) accounted for eccentricity effects in the orthogonal Simon task as a consequence of a change in polarity in the spatial left-right dimension: Placing the response set on one side increases the saliency of that side, turning it effectively into the +pole. Then, that side now matches the +pole of the vertical dimension (up). If this interpretation is correct, and the SNARC effect is due to polarity correspondence between larger numbers and right responses, placing the response set on the left should reverse the SNARC. However, we found no traces of any influence of response eccentricity on the SNARC effect. This complete absence of eccentricity effects suggests that polarity correspondence is not affecting this particular congruency effect whatsoever, not even as an additional source of influence that acts together with other factors on number tasks. Many questions remain. As a first step, current work in our labs is focusing on extending present results to a different conceptual dimension which is also known to generate congruency effects with left-right space: the dimension of time.

## Acknowledgements

This research has been funded by grants P09-SEJ-4772 (Consejería de Innovación, Ciencia y Empresa, Andalusian Government) and PSI2012-32464 (Spanish Ministery of Economy and Competitivity).

## References

Cho, Y. S., Proctor, R. W., \& Yamaguchi, M. (2008). Influences of response position and hand posture on the orthogonal Simon effect. Quarterly Journal of Experimental Psychology, 61(7), 1020-1035.
Clark, H. H. (1969). Linguistic processes in deductive reasoning. Psychological Review, 76(4), 387-404.
Clark, H. H. (1973). Space, time, semantics, and the child. In T. E. Moore (Ed.), Cognitive development and the acquisition of language (pp. 27-63). New York: Academic Press.
Dehaene, S., Bossini, S., \& Giraux, P. (1993). The mental representation of parity and number magnitude. Journal of Experimental Psychology: General, 122(3), 371-396.
Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. Universals of language (Vol. 2, pp. 73-113). MIT Press.
Hines, T. M. (1990). An odd effect: Lengthened reaction times for judgments about odd digits. Memory \& Cognition, 18, 40-46.
Ito, Y., \& Hatta, T. (2004). Spatial structure of quantitative representation of numbers: evidence from the SNARC effect. Memory \& Cognition, 32(4), 662-73.
Lakens, D. (2012). Polarity correspondence in metaphor congruency effects: Structural overlap predicts categorization times for bipolar concepts presented in
vertical space. Journal of Experimental Psychology: Learning, Memory, and Cognition, 38(3), 726-736.
Lakens, D., Semin, G. R., \& Foroni, F. (2012). But for the bad, there would not be good: Grounding valence in brightness through shared relational structures. Journal of Experimental Psychology: General, 141(3), 584-594.
Landau, M. J., Meier, B. P., \& Keefer, L. A. (2010). A metaphor-enriched social cognition. Psychological Bulletin, 136(6), 1045-67.
Nuerk, H.-C., Iversen, W., \& Willmes, K. (2004). Notational modulation of the SNARC and the MARC (linguistic markedness of response codes) effect. The Quarterly Journal of Experimental Psychology, 57A(5), 835-63.
Proctor, R. W., \& Cho, Y. S. (2003). Effects of response eccentricity and relative position on orthogonal stimulus-response compatibility with joystick and keypress responses. The Quarterly Journal of Experimental Psychology, 56A(2), 309-327.
Proctor, R. W., \& Cho, Y. S. (2006). Polarity correspondence: A general principle for performance of speeded binary classification tasks. Psychological Bulletin, 132(3), 416-442.
Santens, S., \& Gevers, W. (2008). The SNARC effect does not imply a mental number line. Cognition, 108(1), 263-270.
Santiago, J., Ouellet, M., Román, A., \& Valenzuela, J. (2012). Attentional factors in conceptual congruency. Cognitive Science, 36(6), 1051-1077.
Santiago, J., Román, A., \& Ouellet, M. (2011). Flexible foundations of abstract thought: A review and a theory. In A. Maass \& T. W. Schubert (Eds.), Spatial dimensions of social thought (pp. 41-110). Berlin: Mouton de Gruyter.
Sorokowski, P. (2010). Politicians' estimated height as an indicator of their popularity. European Journal of Social Psychology, 40(7), 1302-1309.
Tversky, B., Kugelmass, S., \& Winter, A. (1991). Cross-cultural and developmental trends in graphic productions. Cognitive Psychology, 23, 515-557.
Weeks, D. J., Proctor, R. W., \& Beyak, B. (1995). Stimulus-response compatibility for vertically oriented stimuli and horizontally oriented responses: Evidence for spatial coding. Quarterly Journal of Experimental Psychology, 48A(2), 367-383.
Wood, G., Willmes, K., Nuerk, H.-C., \& Fischer, M. H. (2008). On the cognitive link between space and number: A meta-analysis of the SNARC effect. Psychology Science Quarterly, 50(4), 489-525.

# Dependencies First: Eye Tracking Evidence from Sentence Production in Tagalog 

Sebastian Sauppe (sebastian.sauppe@mpi.nl)<br>Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525XD Nijmegen, Netherlands; International Max Planck Research School for Language Sciences, Wundtlaan 1, 6525XD Nijmegen, Netherlands<br>Elisabeth Norcliffe (elisabeth.norcliffe@mpi.nl)<br>Agnieszka E. Konopka (agnieszka.konopka@mpi.nl)<br>Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525XD Nijmegen, Netherlands

Robert D. Van Valin, Jr (vanvalin@phil-fak.uni-duesseldorf.de)<br>Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525XD Nijmegen, Netherlands; Heinrich Heine University Düsseldorf, Universitätsstr. 1, 40225 Düsseldorf, Germany; University at Buffalo, The State University of New York, 609 Baldy Hall, Buffalo, NY 14260, USA

## Stephen C. Levinson (stephen.levinson@mpi.nl)

Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525XD Nijmegen, Netherlands; Radboud University Nijmegen, Erasmusplein 1, 6525HT Nijmegen, Netherlands


#### Abstract

We investigated the time course of sentence formulation in Tagalog, a verb-initial language in which the verb obligatorily agrees with one of its arguments. Eye-tracked participants described pictures of transitive events. Fixations to the two characters in the events were compared across sentences differing in agreement marking and post-verbal word order. Fixation patterns show evidence for two temporally dissociated phases in Tagalog sentence production. The first, driven by verb agreement, involves early linking of concepts to syntactic functions; the second, driven by word order, involves incremental lexical encoding of these concepts. These results suggest that even the earliest stages of sentence formulation may be guided by a language's grammatical structure.


Keywords: eye tracking; sentence production; incrementality; Austronesian; verb-initial word order.

## Introduction

In the process of transforming thoughts into speech, speakers begin with a preverbal message, which must then be encoded linguistically. In English, this process may proceed in a highly lexically incremental manner: for example, when describing events like the one shown in Figure 1, speakers may have encoded as little as the first element (the syntactic subject, e.g., "the boy") of the to-be-uttered sentence prior to speech onset (Gleitman, January, Nappa, \& Trueswell, 2007). The encoding of additional event participants (e.g., "the ball") and the relation between them (e.g., "kicking") may be delayed until after speakers finish encoding the first element. This type of incremental planning is compatible with English morphosyntax, arguably in part because full noun phrases do not morphologically mark dependencies with other elements in the sentence. For many sentence types, speakers therefore do not have to commit to a particular syntactic structure upon beginning to encode one of the event participants as the syntactic subject. However, not all languages offer this flexibility: in some languages the first word is overtly marked for a dependency
with word(s) occurring only later in the sentence. In such cases, is there an effect of dependency marking on early sentence encoding as speakers begin to map the preverbal message onto linguistic structure?

One such language that exhibits dependency marking on the first word of a sentence is the Austronesian language Tagalog. The predicate is in sentence-initial position and agrees with one of its arguments. Thus, the grammatical properties of Tagalog allow us to test whether and how linguistic structure influences the earliest phases of sentence production; specifically, we test whether the overt dependency marking on the first word in a sentence leads to differences in the time course of sentence formulation in Tagalog compared to languages with no overt dependency marking on the first word (such as English).

In the following, we first sketch the relevant grammatical properties of Tagalog and then report the results of a picture description experiment in which eye-tracked speakers described pictures of simple transitive events.

## Tagalog

Tagalog is spoken by approx. 21.5 million speakers in the Philippines; it belongs to the Western Malayo-Polynesian branch of the Austronesian language family. We provide a brief overview of the morphosyntactic properties that are relevant for the reported experiment. For more comprehensive descriptions of Tagalog morphosyntax, see Himmelmann (2005), Kroeger (1993), and Schachter and Otanes (1972).

Basic declarative Tagalog sentences are predicate ${ }^{1}$-initial, i.e., predicates are followed by their arguments. One argu-

[^202]ment phrase in each sentence hosts the case marker ang. The semantic relation between the ang-marked argument and the predicate is signaled by affixes on the predicate.

The ang-marked argument will henceforth be referred to as the privileged syntactic argument (PSA). It is morphosyntactically prominent in being the only argument with which the predicate agrees in semantic role (see sentences (1-4)) and also in being the target of many syntactic operations (e.g., Kroeger, 1993) ${ }^{2}$. Arguments marked by $n g$ do not exhibit these properties and are therefore referred to as non-privileged syntactic arguments (NPSA).
$s<u m>$ isipa
$<$ AV $>$ kick $^{3,4}$
predicate
undergoer
The child kicks the ball.'
(2)
$<A V>$ kick $\quad$ PSA $=$ child $\quad$ NPSA $=$ ball
predicate actor
"The child kicks the ball."
(3)
$s<$ in>ispa $\quad n g=b a t a$
$<\mathrm{UV}>$ kick $\quad$ NPSA $=$ child
predicate actor
"The child kicks the ball."
(4)

| $<$ UV $>$ kick | PSA=ball | NPSA=child |
| :--- | :--- | :--- |
| predicate | undergoer | actor |

"The child kicks the ball."
The sentences in (1-4) illustrate three properties of Tagalog grammar that are relevant for this study. First, the predicate always agrees in semantic role with the PSA in basic sentences. In sentences (1) and (2), the PSA denotes the actor ${ }^{5}$ of the event so the predicate takes actor voice marking (AV); in sentences (3) and (4) the PSA denotes the undergoer of the event so the predicate takes undergoer voice

[^203]marking (UV) ${ }^{6}$. NPSAs marked by the case marker $n g$ do not agree with the predicate. Second, there are no syntactic constraints on the ordering of arguments for the constructions dealt with in this paper. Sentences (1) and (2) have the same meaning but they differ in their word order: in (1) the PSA is sentence-final, whereas it is sentence-medial in (2); the same holds for (3) and (4), respectively. However, canonically, the PSA is in sentence-final position (as in sentences (1) and (3)). Third, the Tagalog voice system is a socalled "symmetrical voice system" (Foley, 2008): sentences in which the undergoer is selected as PSA and sentences with an actor PSA are equally transitive. This contrasts with languages with asymmetrical voice systems such as English in which valency-changing operations, such as passivization which detranstivizes the verb, are required to allow the patient/undergoer argument to be the PSA (syntactic subject in English). Detransitivization is the key part of the asymmetrical voice system in contrast to Tagalog. Thus, all Tagalog sentences analyzed in this paper are transitive (exhibiting one PSA and one NPSA phrase), regardless of the semantic role of the PSA.

## How Do Speakers Plan Sentences in Tagalog?

The sentence-initial position of the predicate in a Tagalog sentence means that speakers must encode enough information about the relationship between the two discourse entities ("boy" and "ball") to select a suitable predicate ("kick") very early in the formulation process. The predicate's agreement in semantic role with the PSA also means that very early in the formulation process one discourse entity from the preverbal message has to be selected to be the PSA and linked to that syntactic function so that appropriate voice marking for the predicate can be selected. Importantly, speakers can produce the PSA immediately after the predicate (as in sentences (2) and (4)) or may delay its production until the end of the sentence (as in sentences (1) and (3)).

To what extent do speakers then have to encode the PSA at the outset of formulation in sentences like (1-4)? We tested whether the processing of the overt dependency between the predicate and the PSA is temporally separate from lexical encoding of the character selected to be the PSA by comparing the time course of formulation for sentences differing in voice and word order.

Native speakers of Tagalog performed a picture description experiment similar to Griffin and Bock (2000) while their eye movements and speech were recorded. The pictures showed events with one actor and one undergoer (Figure 1). We compared the distribution of fixations to the actor and the undergoer in these pictures for different sentence

[^204]

Figure 1: example target stimulus picture (eliciting the example sentences (1-4))
types. First, we compared sentences differing in voice marking and word order, i.e., the actor voice sentence type in (1) and the undergoer voice sentence type in (3), to investigate whether the semantic role of the PSA has an early influence on sentence planning. Second, comparisons between sentences with different voice marking but the same word order (such as in (2) vs. (3)) and sentences with the same voice marking but different word order (such as in (1) vs. (2)) were carried out to investigate whether a possible PSA effect on planning is solely due to the planning of the dependency between the predicate and the PSA (i.e., voice marking) or whether it is also influenced by word order.

## Experiment

## Method

Participants 53 native speakers of Tagalog ( 13 male; mean age $=17.5$ years; range $=15-28$ years) were recruited at De La Salle University in Manila and participated for payment. All participants reported that they spoke Tagalog a total of at least five hours a day and to at least one of their parents.

Materials and Design Target pictures were 44 cartoon drawings of transitive events (see Figure 1). They were interspersed among 76 filler pictures of intransitive events, with at least one filler separating any two target pictures. Two versions of each target picture were created by mirrorreversing the picture. Pictures were then arranged in four lists created by randomizing the order of the target and filler pictures and counterbalancing the two mirror-reversed versions of each target picture.

Equipment The experiment was run with a Tobii T120 eye tracker ( 120 Hz sampling frequency) on a Panasonic CF-F9 computer. Participants' responses were recorded with a microphone.

Procedure Each experimental session lasted approx. 40 minutes. Participants first read instructions for the experiment in Tagalog and completed a questionnaire about their linguistic background. The experimenter (a native speaker of Tagalog) then explained the procedure and repeated the
instructions: participants were asked to describe the events shown in the pictures with one sentence that named all event participants as accurately and as quickly as possible.

Stimuli were presented in two blocks, each block lasting approx. 10-15 minutes. Calibration was performed before each block. The experiment began with a practice phase in which participants saw 11 pictures presented one at a time and heard a recorded description of each depicted event; these example sentences had predicate-initial word order and were mostly PSA-final. After presentation of the example descriptions, participants saw the same pictures again and were asked to describe them themselves. The experimenter provided feedback after each training picture if participants produced non-predicate-initial structures (e.g., existential constructions) or started speaking very late after picture onset.

In the experimental phase, each picture trial was preceded by a display showing a fixation dot at the top of the screen. Participants were asked to look at the fixation dot and the experimenter initiated the trial with a mouse click. Participants completed the experiment without further instructions from the experimenter; however, the experimenter monitored the entire experimental session and repeated the instructions if participants started consistently using non-predicate-initial structures or dropping arguments.

## Results

Picture descriptions Speakers produced 384 sentences with actor voice marking and predicate ${ }_{\mathrm{AV}}$-undergoer $\mathrm{NPSA}^{-}-$actor $_{\mathrm{PSA}}$ word order (as in sentence (1)), 67 sentences with actor voice marking and predicate ${ }_{\mathrm{AV}}$-actor $_{\mathrm{PSA}}-$ undergoer $_{\mathrm{NPSA}}$ word order (as in sentence (2)), 787 sentences with undergoer voice marking and predicate ${ }_{\mathrm{UV}}$-actor NPSA -undergoer ${ }_{\mathrm{PSA}}$ word order (as in sentence (3)), and 26 sentences with undergoer voice marking and predicate ${ }_{\mathrm{UV}}$-undergoer $\mathrm{PSA}_{\mathrm{PA}}-$ actor $_{\mathrm{NPSA}}$ word order (as in sentence (4)). Analyses were limited to the first three sentence types.

First Fixations The majority of first fixations ${ }^{7}$ (58.3\%) across all trials fell on the actor in the event. Contrary to earlier work on English (e.g., Gleitman et al., 2007), first fixations did not predict choice of voice or word order (both $z$ 's $<1.4$, n.s.).

Time Course of Fixations Consecutive fixations to each character were aggregated into "runs" of fixations directed to those characters. The distributions of fixations directed to the actor and to the undergoer were then compared across the three most frequent sentence types in this dataset (i.e., the sentence types in (1-3)) with quasi-logistic regressions (Barr, 2008, for details about random effects). We selected three time windows for analysis $(0-600 \mathrm{~ms}, 600-1600 \mathrm{~ms}$, and $1600-2600 \mathrm{~ms}$ after picture onset). Selection of time windows was based on three theoretically important distinc-

[^205]

Figure 2: graphs showing fixation proportions to actor and undergoer characters over time for three sentence types
tions and to facilitate comparisons across the three sentence types, as explained below.

In all three sentence types included in the analysis (Figure 2), speakers directed more fixations to the character selected to be PSA than to the NPSA character in an early time window ( $0-600 \mathrm{~ms}$ ). The $0-400 \mathrm{~ms}$ time window is argued to correspond to a period of event apprehension (Griffin \& Bock, 2000), and here we extended this window to 600 ms based on the distribution of fixations in all three sentence types in Figure 2. Fixations to the two characters in this time window were aggregated into 50 ms time bins. Griffin and Bock (2000) propose that after the initial period of event apprehension, speakers begin fixating the two characters in the order of mention in order to retrieve their names; indeed, the distribution of fixations after 600 ms in this dataset largely shows that speakers fixated characters in the order of mention. Thus between 600 ms and speech onset (approx. 1600 ms after picture onset), speakers preferentially fixated the character that was mentioned immediately after the predicate (independently of syntactic function - i.e., whether it was the PSA or the NPSA - or semantic role). After speech onset, speakers then began shifting their gaze to the second character, and we compared the distribution of fixations to the two characters up to 1 second after speech onset (i.e., between 1600 and 2600 ms ). Fixations were aggregat-
ed into 200 ms time bins for the analysis of the $600-1600$ ms and $1600-2600 \mathrm{~ms}$ time windows.

Three analyses were performed to compare the distribution of actor-directed fixations in sentences differing in voice and word order in more detail. All analyses included time bin and sentence type as predictors. All models included random slopes for the two predictors. In the text, we report only the interactions between these factors from the byparticipant analyses using the full random structure. ${ }^{8}$ Effects were considered to be reliable at $p<0.05$ (most effects were also reliable according to the more conservative pMCMC estimates calculated for models without random slopes; in cases of discrepancy between $p$ values calculated for models with random slopes and pMCMC values calculated for models without random slopes, we used the more conservative criterion to indicate significance and provide the corresponding pMCMC value). ${ }^{9}$

[^206]First, to test whether differences in voice marking and word order predict differences in early encoding of the PSA and NPSA characters, we compared actor voice sentences with predicate ${ }_{A V}$-undergoer ${ }_{\text {NPSA }}-$ actor $_{\text {PSA }}$ word order (example (1), Figure 2a) and undergoer voice sentences with predicate ${ }_{\mathrm{UV}}-$ actor $_{\mathrm{NPSA}}-$ undergoer ${ }_{\text {PSA }}$ word order (example (3), Figure 2c). Between 0 and 600 ms , speakers fixated the actor character more often and more quickly when it was selected to be the PSA (i.e., in the actor voice sentences with predicate ${ }_{\mathrm{AV}}$-undergoer $\mathrm{NPSA}^{-}$actor $_{\text {PSA }}$ word order, Figure 2a) than when it was not (Figure 2c; sentence type $\times$ time bin: $\beta=-0.92, t=-7.16$ ). After 600 ms , speakers fixated the characters in the order of mention. Between 600 and 1600 ms they fixated the actor character more often when the actor argument phrase immediately followed the predicate, i.e., in the undergoer voice sentences with predicate $\mathrm{UV}^{- \text {actor }_{\mathrm{NPSA}^{-}}}$ undergoer ${ }_{\text {PSA }}$ word order (Figure 2c), then when it was sen-tence-final (Figure 2a; sentence type $\times$ time bin: $\beta=1.14$, $t=49.48$ ). They shifted their gaze to the second character between 1600 and 2600 ms , and thus fixated the actor character less often when the actor argument phrase was not sentence-final (Figure 2a; sentence type $\times$ time bin: $\beta=-1.70$, $t=-73.93$ ). These results suggest early encoding of the PSA character only for the purposes of selecting the appropriate agreement marking on the predicate; lexical encoding of the PSA character occurred either before or after encoding of the NPSA character, according to word order.

Second, we tested whether differences in voice marking alone can influence the time course of formulation by comparing sentences with the same word order, i.e., actor voice sentences with predicate ${ }_{\mathrm{AV}}$-actor $\mathrm{PSA}_{\mathrm{PA}}$-undergoer ${ }_{\mathrm{NPSA}}$ word order (example (2), Figure 2b) and undergoer voice sentences with predicate ${ }_{\mathrm{UV}}$-actor ${ }_{\mathrm{NPSA}}$-undergoer $\mathrm{r}_{\mathrm{PSA}}$ word order (example (3), Figure 2c). Speakers fixated the actor character more often and more quickly between 0 and 600 ms when it was selected to be the PSA (Figure 2b) than when it was not selected to be the PSA, i.e., in the undergoer voice sentences (Figure 2c; sentence type $\times$ time bin: $\beta=-1.76, t=-7.46$ ). However, there were no differences between the actor voice and the undergoer voice sentences in the overall likelihood of speakers to fixate the actor character or direct fixations to it over time between 600 and 1600 ms (sentence type: $\beta=0.13, t=4.40, \mathrm{pMCMC}=0.65$; sentence type $\times$ time bin: $\beta=-0.03, t=-0.56, \mathrm{pMCMC}=0.93)$ and between 1600 and 2600 ms after picture onset (sentence type: $\beta=-0.06, t=-1.64$, $\mathrm{pMCMC}=0.84$; sentence type $\times$ time bin: $\beta=-0.18, t=-3.09$, $\mathrm{pMCMC}=0.70$ ) because speakers produced the actor character first and the undergoer character second in both sentence types. This confirms that differences in the time course of early sentence formulation reflect encoding of features of the PSA character relevant only for agreement marking.

Finally, we compared actor voice sentences with predica-$\mathrm{te}_{\mathrm{AV}}$-undergoer ${ }_{\text {NPSA }}$-actor ${ }_{\text {PSA }}$ word order (example (1), Figure 2 a ) and actor voice sentences with predicate $\mathrm{AV}^{- \text {-actor }_{\mathrm{PSA}^{-}}}$ undergoer $_{\text {NPSA }}$ word order (example (2), Figure 2b) to test whether word order influences fixations to picture characters when voice marking is kept constant. Comparing fixa-
tions to the actor character between 0 and 600 ms showed no reliable differences between the actor voice sentences with predicate ${ }_{\mathrm{AV}}$-undergoer $\mathrm{NPSA}^{-}$actor $_{\text {PSA }}$ word order (Figure $2 a)$ and the actor voice sentences with predicate $\mathrm{AV}^{- \text {actor }_{\mathrm{PSA}^{-}}}$ undergoer $_{\text {NPSA }}$ word order (Figure 2b) in this time window (sentence type: $\beta=-0.36, t=-3.15, \mathrm{pMCMC}=0.23$; sentence type $\times$ time bin: $\beta=0.88, t=3.00, \mathrm{pMCMC}=0.16)$. The distribution of fixations to the actor and undergoer characters did, however, differ after 600 ms because the linear order of these characters in the two sentence types was different. Thus in the 600 and 1600 ms time window, speakers fixated the actor character more often if the actor argument phrase immediately followed the predicate (predicate $\mathrm{AV}^{- \text {actor }_{\mathrm{PSA}^{-}}}$ undergoer ${ }_{\text {NPSA }}$ word order) then if it was sentence-final (sentence type $\times$ time bin: $\beta=1.13, t=15.79$ ), whereas speakers fixated the actor character less often between 1600 and 2600 ms if the actor argument phrase directly followed the predicate than if it was sentence-final (sentence type $\times$ time bin: $\beta=-1.75, t=-18.39)$.

Speech onsets Speech onsets are shown in Figure 2 for each sentence type. Onsets were somewhat shorter in actor voice sentences with predicate ${ }_{\mathrm{AV}}$-undergoer ${ }_{\mathrm{NPSA}}$-actor ${ }_{\mathrm{PSA}}$ word order than in actor voice sentences with predicate ${ }_{A V}$ actor $_{\text {PSA }}$-undergoer ${ }_{\text {NPSA }}$ word order and undergoer voice sentences with predicate ${ }_{\mathrm{UV}}$-actor ${ }_{\mathrm{NPSA}}$-undergoer ${ }_{\text {PSA }}$ word order (word order $\times$ voice: $\beta=-0.19, \quad t=-1.99$, $\mathrm{pMCMC}=0.052$ ). More importantly, we note a difference from results obtained with English speakers (Griffin \& Bock, 2000): here, speech onsets occurred while speakers were still fixating the character that was mentioned first, suggesting that they had only fully encoded the predicate before initiating production, whereas English speakers begin their sentences only after encoding the first character. The sentence-initial position of the predicate may have allowed speakers to begin their sentences before completing the encoding of the first-mentioned character.

## Discussion and Conclusion

We interpret the results of this experiment as evidence for linguistic guidance in the earliest stages of sentence production in Tagalog and for a temporal dissociation of the mapping of message-level concepts to syntactic functions and the lexical encoding of these concepts.

Linguistic guidance in early sentence production is suggested by differences in fixation patterns in the $0-600 \mathrm{~ms}$ time window across the three sentence types we analyze here: a depicted character was fixated more often if it was to become the sentence's PSA than when it was not. Specifically, speakers fixated the actor character more often than the undergoer character before 600 ms if the actor character was selected as the PSA, regardless of the position of the actor argument in the sentence (i.e., this pattern held for both, actor $_{\text {PSA }}-m e d i a l$ or actor ${ }_{P S A}-$ final word orders). In contrast, there was no difference in early fixations directed to the actor character in actor voice sentences with different word orders (i.e., in actor ${ }_{\mathrm{PSA}}-$ medial or actor $_{\mathrm{PSA}}$-final sentences).

In other words, the results suggest that differences in voice marking (signaling differing semantic roles of the PSAs) but not differences in word order have an effect on fixation patterns in the earliest stage of sentence planning.

Early fixations of the PSA character suggest that the PSA effect is a reflex of linking message-level concepts of discourse entities to prominent syntactic functions. Speakers select a participant of the depicted event to be the PSA and encode its semantic role in order to produce an appropriate voice affix at the predicate. We propose that this process happens very early during formulation as speakers begin encoding information about the relationship between the two characters in the event.

Comparisons of the fixation patterns in the two later time windows ( $600-1600 \mathrm{~ms}$ and $1600-2600 \mathrm{~ms}$ ) suggest that the PSA effect, i.e., the linking of a discourse entity concept to a prominent syntactic function, and the lexical encoding of the PSA are temporally dissociated. Whereas speakers are more likely to fixate the character selected to be the PSA before 600 ms , fixations to the two characters after 600 ms are contingent on word order. In the $600-1600 \mathrm{~ms}$ time window, the character that is to be mentioned immediately after the predicate is fixated more often by speakers than the character that is to be mentioned sentence-finally. Specifically, in actor voice sentences with predicate $\mathrm{AV}^{-}$ undergoer ${ }_{\text {NPSA }}$-actor $_{\text {PSA }}$ word order, speakers shift their gaze from the actor character (the PSA) to the undergoer character (the NPSA) after 600 ms , and similarly, in undergoer voice sentences with predicate ${ }_{\mathrm{UV}}-$ actor $_{\mathrm{NPSA}}$-undergoer ${ }_{\mathrm{PSA}}$ word order, the speakers' gaze shifts from the undergoer character (the PSA) to the actor character (the NPSA) after 600 ms . Finally, in actor voice sentences with predicate $\mathrm{AV}^{-}$ actor $_{\text {PSA }}$-undergoer ${ }_{\text {NPSA }}$ word order speakers continue looking at the actor character (the PSA) because it is to be mentioned directly after the predicate. In the $1600-2600 \mathrm{~ms}$ window, speakers then fixate the character to be mentioned sentence-finally more often than the other character (i.e., the actor character in the first mentioned sentence type and the undergoer character in the two latter types). We interpret this as incremental encoding of the two character names in the order of mention that is distinct from the early phase of linking concepts to syntactic functions ( $0-600 \mathrm{~ms}$ ).

The results suggest that there are two observable phases in the sentence production process in Tagalog: an early phase of sentence planning that includes the planning of the dependency relation between the predicate and the PSA (i.e., the voice marking), which is neither influenced by the actual semantic role of the PSA nor the word order of the to-be-uttered sentence, and a later phase that involves the incremental lexical encoding of the two arguments of the predicate.

Importantly, these analyses of the time course of sentence formulation in Tagalog provide insight into a process that is not easily observable in a language like English, namely the linking of conceptual discourse entities to prominent syntactic functions. The rigid subject-initial word order of English prevents dissociating the linking of concepts to syntactic
functions from planning and encoding of the subject argument; thus our results on Tagalog highlight the need for controlled studies on typologically diverse languages that allow dissociations between different processes at the interface of thinking and speaking.

Ultimately, more fine-grained models of early message and sentence formulation are needed to address the relationship between formulation of a preverbal message and the mapping of this message onto language, and it is important for the development of such models to consider languages with grammatical properties that support investigations of these phenomena (Jaeger \& Norcliffe, 2009).

## Acknowledgments

This research was supported by a PhD fellowship from the International Max Planck Research School for Language Sciences to the first author and by the Language \& Cognition Department of the MPI for Psycholinguistics, Nijmegen. We thank Danilo Dayag for administrative help, Philip Rentillo and Erwin Roxas for help with conducting the experiment and transcriptions, and Ronald Fischer for technical assistance.

## References

Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. Journal of Memory and Language, 59, 457-474.
Foley, W. A. (2008). The place of Philippine languages in a typology of voice systems. In Austin, P. K. \& Musgrave, S. (Eds.), Voice and grammatical relations in Austronesian languages (22-44). Palo Alto: CSLI Publications.
Foley, W. A. and Van Valin, R. D. (1984). Functional syntax and universal grammar. Cambridge: Cambridge University Press.
Gleitman, L. R., January, D., Nappa, R., and Trueswell, J. C. (2007). On the give and take between event apprehension and utterance formulation. Journal of Memory and Language, 57, 544-596.
Griffin, Z. M. and Bock, K. (2000). What the eyes say about speaking. Psychological Science, 11, 274-279.
Himmelmann, N. P. (2005). Tagalog. In Adelaar, A. \& Himmelmann, N. P. (Eds.), The Austronesian languages of Asia and Madagascar (350-376). Oxon: Routledge.
Himmelmann, N. P. (2008). Lexical categories and voice in Tagalog. In Austin, P. K. \& Musgrave, S. (Eds.), Voice and grammatical relations in Austronesian languages (247-293). Palo Alto: CSLI Publications.
Jaeger, T. F. and Norcliffe, E. J. (2009). The cross-linguistic study of sentence production. Language and Linguistics Compass, 3, 866-887.
Kroeger, P. (1993). Phrase structure and grammatical relations in Tagalog. Palo Alto: CSLI Publications.
Schachter, P. \& Otanes, F. T. (1972). Tagalog reference grammar. Berkeley: University of California Press.

# Learning hierarchical category structure in deep neural networks 

Andrew M. Saxe (asaxe@stanford.edu)<br>Department of Electrical Engineering<br>James L. McClelland (mcclelland@stanford.edu)<br>Department of Psychology<br>Surya Ganguli (sganguli@stanford.edu)<br>Department of Applied Physics<br>Stanford University, Stanford, CA 94305 USA


#### Abstract

Psychological experiments have revealed remarkable regularities in the developmental time course of cognition. Infants generally acquire broad categorical distinctions (i.e., plant/animal) before finer ones (i.e., bird/fish), and periods of little change are often punctuated by stage-like transitions. This pattern of progressive differentiation has also been seen in neural network models as they learn from exposure to training data. Our work explains why the networks exhibit these phenomena. We find solutions to the dynamics of error-correcting learning in linear three layer neural networks. These solutions link the statistics of the training set and the dynamics of learning in the network, and characterize formally how learning leads to the emergence of structured representations for arbitrary training environments. We then consider training a neural network on data generated by a hierarchically structured probabilistic generative process. Our results reveal that, for a broad class of such structures, the learning dynamics must exhibit progressive, coarse-to-fine differentiation with stage-like transitions punctuating longer dormant periods.


Keywords: neural networks; hierarchical generative models; semantic cognition; learning dynamics

## Introduction

Our world is characterized by a rich, nested hierarchical structure of categories within categories, and one of the most remarkable aspects of human semantic development is our ability to learn and exploit this structure. Experimental work has shown that infants and children acquire broad categorical distinctions before fine categorical distinctions (Keil, 1979; Mandler \& McDonough, 1993), suggesting that human category learning is marked by a progressive differentiation of concepts from broad to fine. Furthermore, humans can exhibit stage-like transitions as they learn, rapidly progressing through successive levels of mastery (Inhelder \& Piaget, 1958; Siegler, 1976).

Many neural network simulations have captured aspects of these broad patterns of semantic development (Rogers \& McClelland, 2004; Rumelhart \& Todd, 1993; McClelland, 1995; Plunkett \& Sinha, 1992; Quinn \& Johnson, 1997). The internal representations of such networks exhibit both progressive differentiation and stage-like transitions. However, the theoretical basis for the ability of neuronal networks to exhibit such strikingly rich nonlinear behavior remains elusive. What are the essential principles that underly such behavior? What aspects of statistical structure in the input are responsible for driving such dynamics? For example, must networks exploit nonlinearities in their input-output map to detect higher order statistical regularities to drive such learning?


Figure 1: The three layer network analyzed in this work.
Here we analyze the learning dynamics of a linear three layer network and find, surprisingly, that it can exhibit highly nonlinear learning dynamics, including rapid stage-like transitions. Furthermore, when exposed to hierarchically structured data sampled from a hierarchical probabilistic model, the network exhibits progressive differentiation of concepts from broad to fine. Since such linear networks are sensitive only to the second order statistics of inputs and outputs, this yields the intriguing result that merely second order patterns of covariation in hierarchically structured data contain statistical signals powerful enough to drive certain nontrivial, high level aspects of semantic development in deep networks.

We outline our approach here in brief. We begin by decomposing the training set to identify important dimensions of variation using the singular value decomposition (SVD), which will turn out to be fundamental to our analysis. Next, we examine the equations governing gradient descent learning and show that they can be solved in terms of the SVD of the training set. This solution analytically expresses the weight values of the neural network at any point in time during learning as a function of the input training set. Finally, we consider generating the training set from a hierarchical probabilistic generative model. We analytically calculate the SVD of training sets so generated, which in combination with our previous results gives a formal grounding for how neural networks will learn about hierarchical categorical structure. We show that networks must exhibit progressive differentiation of categorical structure and stage-like transitions for any training set generated by a class of hierarchical generative models.

## Decomposing the training set

Our fundamental goal is to understand the dynamics of learning in neural networks as a function of the training set. Toward this goal, in this section we introduce the singular
value decomposition, which identifies important dimensions of variation in the training set. The SVD will turn out to be fundamentally linked to learning dynamics, a connection we develop in the next section. We wish to train a neural network to learn a particular input-output map from a set of $P$ training examples $\left\{x^{\mu}, y^{\mu}\right\}, \mu=1, \ldots, P$. These $P$ pairs of vectors constitute the training set. In the model of semantic development introduced by Rumelhart and Todd (1993), for instance, elements of $x^{\mu}$ correspond to input units representing items such as Canary or Rose. The elements of $y^{\mu}$ correspond to output units representing possible predicates or attributes such as can Fly or has Petals that may or may not apply to each item. Hence each example links a particular item to a set of properties, and the training set contains the semantic content in the world to be learned by the network.

For concreteness, we consider a simple example dataset with four items (Canary, Salmon, Oak, and Rose) and five properties. The two animals share the property that they can Move, while the two plants cannot. In addition each item has a unique property: can Fly, can Swim, has Bark, and has Petals, respectively. In a more natural data set, the plantanimal, bird-fish, and tree-flower distinctions are based on clusters of covarying properties, for which the single properties identified here serve as proxies.

An important function of the training set is the input-output correlation matrix

$$
\begin{equation*}
\Sigma^{31} \equiv \sum_{\mu=1}^{P} y^{\mu} x^{\mu} \equiv E\left[y x^{T}\right] \tag{1}
\end{equation*}
$$

For our example dataset, this matrix is shown in Fig. 2. Each column corresponds to an item, and denotes the properties possessed by that particular item.

Our example dataset contains important shared structure. The Canary and Salmon, for instance, both can Move, and hence may naturally be grouped together. Intuitively, they are both animals, and as a consequence have certain properties in common that are typical of animals. How can we identify these coherently covarying groups of items and their properties? We will show that the singular value decomposition of the input-output correlation matrix accomplishes exactly this.

The singular value decomposition (SVD)

$$
\begin{equation*}
\Sigma^{31}=U^{33} S^{31} V^{11^{T}}=\sum_{\alpha=1}^{N_{1}} s_{\alpha} u^{\alpha} v^{\alpha T} \tag{2}
\end{equation*}
$$

decomposes any matrix into the product of three matrices. Each of these matrices has an important real world interpretation. We call the $N_{1} \times N_{1}$ orthogonal matrix $V^{11}$ the object analyzer-it determines the position of a particular item along a number of important dimensions of the training set. The first row of $V^{11^{T}}$, for instance, determines where items sit on an animal-plant dimension, and hence has positive values for the Canary and Salmon and negative values for the plants. In our example dataset, the three dimensions identified by the SVD are animal-plant, bird-fish, and flower-tree.


Figure 2: First three modes of the singular value decomposition of a toy dataset. Left: The learning environment is specified by an input-output correlation matrix. Right: The SVD decomposes $\Sigma^{31}$ into modes that link a set of coherently covarying items (object analyzer vectors in the rows of $V^{T}$ ) to a set of coherently covarying properties (feature synthesizer vectors in the columns of $U$ ). The overall strength of this link is given by the singular values lying along the diagonal of $S$. In this toy example, mode 1 distinguishes plants from animals; mode 2 birds from fish; and mode 3 flowers from trees.

The $N_{3} \times N_{3}$ orthogonal matrix $U^{33}$ can be interpreted as a feature synthesizer-it contains those features typical of a particular dimension in each column. Hence the feature synthesizer associated with the animal-plant dimension has positive values for can Move, since animals typically can move while plants cannot.

Finally the $N_{3} \times N_{1}$ association strength matrix $S^{31}$ captures the overall strength of the association between an input dimension and output dimension. It is nonzero only on the diagonal; these elements are the singular values $s_{\alpha}, \alpha=$ $1, \ldots, N_{1}$ ordered so that $s_{1} \geq s_{2} \geq \cdots \geq s_{N_{1}}$. The large association strength for the animal-plant dimension reflects the fact that this one dimension explains more of the training set than the finer-scale dimensions like bird-fish and flower-tree.

In a larger training set, the SVD will extract modes that capture patterns of coherent covariation in the properties of items in the training set. The quantities defining each mode, $\left\{s_{\alpha}, u^{\alpha}, v^{\alpha}\right\}$, are connected to the learning dynamics of neural networks in the next section.

## Gradient descent dynamics in multilayer neural networks

We examine learning in a three layer network (input layer 1 , hidden layer 2, and output layer 3) with linear activation functions, simplifying the network model of Rumelhart and Todd (1993). Let $N_{i}$ be the number of neurons in layer $i, W^{21}$ be an $N_{2} \times N_{1}$ matrix of synaptic connections from layer 1 to 2 , and similarly, $W^{32}$ an $N_{3} \times N_{2}$ matrix of connections from layer 2 to 3 . The input-output map of the network is $y=W^{32} W^{21} x$, where $x$ is an $N_{1}$ dimensional column vector representing inputs to the network, and $y$ is an $N_{2}$ dimensional column vector representing the network output (see Fig. 1).

Training is accomplished in an online fashion via stochas-
tic gradient descent; each time an example $\mu$ is presented, the weights $W^{32}$ and $W^{21}$ are adjusted by a small amount in the direction that minimizes the squared error $\left\|y^{\mu}-W^{32} W^{21} x^{\mu}\right\|^{2}$ between the desired feature output, and the network's feature output. This gradient descent procedure yields the standard back propagation learning rule

$$
\begin{align*}
\Delta W^{21} & =\lambda W^{32^{T}}\left(y^{\mu}-\hat{y}^{\mu}\right) x^{\mu T}  \tag{3}\\
\Delta W^{32} & =\lambda\left(y^{\mu}-\hat{y}^{\mu}\right) h^{\mu T} \tag{4}
\end{align*}
$$

for each example $\mu$, where $\hat{y}^{\mu}=W^{32} W^{21} x^{\mu}$ denotes the output of the network in response to input example $x^{\mu}, h^{\mu}=W^{21} x^{\mu}$ is the hidden unit activity, and $\lambda$ is a small learning rate. Here $W^{32^{T}}\left(y^{\mu}-\hat{y}^{\mu}\right)$ in (3) corresponds to the signal backpropagated to the hidden units through the hidden-to-output weights. These equations emphasize that the learning process works by comparing the network's current output $\hat{y}^{\mu}$ to the desired target output $y^{\mu}$, and adjusting weights based on this error term.

By a substitution and rearrangement, however, we can equivalently write these equations as

$$
\begin{align*}
\Delta W^{21} & =\lambda W^{32^{T}}\left(y^{\mu} x^{\mu T}-W^{32} W^{21} x^{\mu} x^{\mu T}\right)  \tag{5}\\
\Delta W^{32} & =\lambda\left(y^{\mu} x^{\mu T}-W^{32} W^{21} x^{\mu} x^{\mu T}\right) W^{21^{T}} \tag{6}
\end{align*}
$$

This form emphasizes two crucial aspects of the learning dynamics. First, it highlights the importance of the statistics of the training set. In particular, the training set enters only through two terms, one related to the input-output correlations $y^{\mu} x^{\mu T}$ and the other related to the input correlations $x^{\mu} x^{\mu T}$. Indeed, if $\lambda$ is sufficiently small so that weights change only a small amount per epoch, we can rewrite these equations in a batch update form by averaging over the training set to obtain the mean change in weights per learning epoch,

$$
\begin{align*}
\tau \frac{d}{d t} W^{21} & =W^{32^{T}}\left(\Sigma^{31}-W^{32} W^{21} \Sigma^{11}\right)  \tag{7}\\
\tau \frac{d}{d t} W^{32} & =\left(\Sigma^{31}-W^{32} W^{21} \Sigma^{11}\right) W^{21^{T}} \tag{8}
\end{align*}
$$

where $\Sigma^{11} \equiv \sum_{\mu=1} x^{\mu} x^{\mu T} \equiv E\left[x x^{T}\right]$ is an $N_{1} \times N_{1}$ input correlation matrix, $\Sigma^{31}$ is the $N_{3} \times N_{1}$ input-output correlation matrix defined previously, and $\tau \equiv \frac{P}{\lambda}$. Hence we see that linear networks are sensitive only to the second order statistics of inputs and outputs. In general the learning process is driven by both the input and input-output correlation matrices. Here we take the simplifying assumption that these input correlations are insignificant; formally, we assume $\Sigma^{11}=I$, the identity matrix. Concretely, this assumption corresponds to the supposition that input representations for different items are highly differentiated from, or orthogonal to each other. While this is unlikely to hold exactly in any natural domain, we take this assumption for two reasons. First, it was used in prior simulation studies (Rogers \& McClelland, 2004), and hence our attempt to understand their results is not limited by this assumption. Second, Rogers and McClelland (2004)
have shown that relaxing this assumption to incorporate more complex input correlations leaves intact the basic phenomena of progressive differentiation and stage-like transitions in learning. Nevertheless, understanding the impact of input correlations is an important direction for further work.

Second, the form of Eqns. (7)-(8) highlights the coupling between the two equations: to know how to change $W^{21}$ we must know $W^{32}$, and visa versa, since each appears in the update equation for the other. This coupling is the crucial element added by the addition of a hidden layer, and as we shall see, it qualitatively changes the learning dynamics of the network compared to a "shallow" network with no hidden layer. Intuitively, this coupling complicates the learning procedure since both weight matrices must cooperate to produce the correct answer; but crucially, it enables knowledge sharing between different items, by assigning them similar hidden unit representations. Without this coupling, the network would learn each item-property association independently, and would not be sensitive to shared structure in the training set.

The temporal dynamics of learning To understand the connection between learning dynamics and training set statistics, then, we can solve Eqns. (7)-(8). We have found a class of exact solutions (whose derivation will be presented elsewhere) that describe the weights of the network over time during learning, as a function of the training set. In particular, the composite mapping at any time $t$ is given by

$$
\begin{equation*}
W^{32}(t) W^{21}(t)=\sum_{\alpha=1}^{N_{2}} a\left(t, s_{\alpha}, a_{\alpha}^{0}\right) u^{\alpha} v^{\alpha T} \tag{9}
\end{equation*}
$$

where the function $a\left(t, s, a^{0}\right)$ governing the strength of each input-output mode is given by

$$
\begin{equation*}
a\left(t, s, a_{0}\right)=\frac{s e^{2 s t / \tau}}{e^{2 s t / \tau}-1+s / a_{0}} \tag{10}
\end{equation*}
$$

That is, the network learns about the $N_{2}$ strongest inputoutput modes identified by the singular value decomposition, progressively incorporating each mode into its representation. The coefficient $a\left(t, s^{\alpha}, a_{0}\right)$ describes how strongly input-output mode $\alpha$ has been learned by time $t$, starting from some small initial value of $a_{0}$. As can be seen from Fig. 3, this function is a sigmoidal curve, capturing the fact that the network initially knows nothing about a particular dimension (the animal-plant dimension, say), but over time learns the importance of this dimension and incorporates it into its representation, ultimately reaching the correct association strength $s^{\alpha}$. At this point the network correctly maps items onto the animal-plant dimension using the object analyzer vector $v^{\alpha T}$, and generates the corresponding correct features using the feature synthesizer vector $u^{\alpha}$.

Eqns. (9)-(10) describe the fundamental connection between the structure of a training set and learning dynamics. In particular, the dynamics depends on the singular value decomposition of the input-output correlation matrix of the


Figure 3: Close agreement between theoretically predicted time course and numerical simulations. Simulations were performed with a dataset sampled from the hierarchical diffusion process described in detail in a later section, with $D=3$ hierarchical levels, binary branching, flip probability $\varepsilon=0.1$, and $N=10,000$ sampled features. This data set had 3 unique singular values. Red traces show ten simulations of the singular value dynamics of $W^{32}(t) W^{21}(t)$ in Eqns. (7)-(8) starting from different random initializations, and blue traces show theoretical curves obtained from (10).
training set. Further, they reveal important properties of these learning dynamics.

First, each input-output mode is learned on a different time scale, governed by its singular value $s_{\alpha}$. To calculate this time scale, we can assume a small initial condition $a_{0}=\varepsilon$ and ask when $a(t)$ in (10) rises to within $\varepsilon$ of the final value $s_{\alpha}$, i.e. $a(t)=s_{\alpha}-\varepsilon$; then the timescale of learning in the limit $\varepsilon \rightarrow 0$ is

$$
\begin{equation*}
t(s, \varepsilon)=\frac{\tau}{s_{\alpha}} \ln \frac{s_{\alpha}}{\varepsilon} \tag{11}
\end{equation*}
$$

Hence up to a logarithmic factor, the time required to learn an input-output mode is inversely related to its association strength, quantified through its singular value.

Second, these dynamics reveal stage-like transitions in learning performance. Intuitively, this property arises from the sigmoidal transition in (10) from a state in which the network does not represent a particular input-output relation at all, to a state in which the network fully incorporates that relation. Because of the sigmoidal shape, the solution can remain very small for a long period of time before rapidly transitioning to mastery. To formalize this, we note that the time it takes to reach half mastery (i.e. $a\left(t_{\text {half }}\right)=s / 2$ ) is

$$
\begin{equation*}
t_{\text {half }}=\frac{\tau}{2 s} \log \left(\frac{s}{a_{0}}-1\right) \tag{12}
\end{equation*}
$$

In contrast, the duration of the transition period in which the weights change rapidly is $t_{\text {trans }}=\frac{2 \tau}{s}$ (using a linear approximation). Thus, by starting with a very small initial condition for the weights (i.e. $a_{0} \approx 0$ ), it is clear that one can make the ratio $t_{\text {trans }} / t_{\text {half }}$ arbitrarily small, i.e., the transition period can be very brief relative to the long initial period of dormancy. Hence the learning dynamics of (7)-(8) exhibit
sharp stage-like transitions. Importantly, we can prove that networks with only direct input-output connections and no hidden layer are not capable of such stage-like transitions. Their existence is an emergent property of nonlinear learning dynamics in deep networks with at least one hidden layer.

The result in (9) is the solution to (7)-(8) for a special class of initial conditions on the weights $W^{21}$ and $W^{32}$. However this analytic solution is a good approximation to the time evolution the network's input-output map for random small initial conditions, as confirmed in Fig. 3.

Summary of learning dynamics The preceding analyses have established a number of crucial features of gradient descent learning in a simple linear network, making explicit the relationship between the statistical structure of training examples and the dynamics of learning. In particular, the learning dynamics depend crucially on the singular values of the inputoutput correlation matrix. Each input-output mode is learned in time inversely proportional to its associated singular value, yielding the intuitive result that stronger input-output associations are learned before weaker ones.

## The singular values and vectors of hierarchically generated data

In this section we introduce a hierarchical probabilistic generative model of items and their attributes that, when sampled, produces a dataset that can be supplied to our neural network. By analytically calculating the SVD of this data, we will be able to explicitly link hierarchical taxonomies of categories to the dynamics of network learning. A key result in the following is that our network must exhibit progressive differentiation with respect to any of the underlying hierarchical taxonomies allowed by our generative model.
Hierarchical feature vectors from a branching diffusion process We propose a simple generative model of hierarchical data $\left\{x^{\mu}, y^{\mu}\right\}$, and compute for this model the inputoutput modes $\left(s_{\alpha}, u^{\alpha}, \nu^{\alpha}\right)$ which drive learning. The hierarchical structure in the generative model is represented by a tree (see e.g. Fig. 4). Each leaf node of this tree corresponds to an item in the dataset. Our generative process assigns features to these items such that items with more recent common ancestors are more likely to share features. For instance, our example dataset might have been generated by a three level binary tree with four leaf nodes. The top level would separate the animals from the plants, while the next level would separate the birds from the fish and the flowers from the plants.

In detail, to sample one feature's value across items, the root node is randomly set to $\pm 1$ with equal probability $\frac{1}{2}$; next this value diffuses to children nodes, where its sign is flipped with a small probability $\varepsilon$. This process continues until the leaf nodes have been assigned values. These assignments yield the value of this feature on each item.

Under this process, the can Move feature, for example, might have arisen as follows: randomly the root node of the three level binary tree was assigned a value of 1 . This value
diffused down to the two second level nodes, maybe in this instance changing sign to -1 for the parent node of the plants, but not changing for the parent node of the animals. Then these values diffused down to the leaf nodes representing the individual items, perhaps not flipping sign for any of them. Hence the ultimate feature assignment would be +1 on the Canary and Salmon and -1 on the Flower and Tree. This is just one possible sample from the generative model, but serves to illustrate how hierarchical structure arises from the feature generation process. To generate more features, the process is repeated independently $N$ times.

For simplicity, we consider trees with a regular branching structure. The tree has $D$ levels indexed by $l=0, \ldots, D-$ 1 , with $M_{l}$ total nodes at level $l$. Every node at level $l$ has exactly $B_{l}$ descendants. Thus $M_{l}=M_{0} \Pi_{k=0}^{l-1} B_{l}$. The tree has a single root node at the top $\left(M_{0}=1\right)$, and again $P$ leaves at the bottom, one per example in the dataset $\left(M_{D-1}=P\right)$.

We have thus far described the output feature vectors $y^{\mu}$. To complete the specification of the training set, we assume that the input vectors $x^{\mu}$ are simply chosen to be highly distinct (i.e., orthogonal). One such choice is a localist coding scheme in which a different element is active to represent the presence of each item.

Input-output modes of hierarchical data How will our neural network learn about training sets generated as just described? To understand this, we calculate the SVD of such training sets. We will see that the input-output modes identified by the SVD exactly mirror the tree structure used to generate the dataset. The feature generation process described in the previous section generates a training set with $N$ features. In the limit of large numbers of features, we obtain the following (the full derivation to be presented elsewhere):

The object analyzer vectors exactly mirror the tree structure, as shown in Fig. 4. One mode will correspond to a broad, high level distinction (e.g., animal-plant) near the root of the tree, while another will correspond to a more detailed distinction (e.g., bird-fish). For binary trees, each object analyzer vector will have positive weights on all items on one side of a binary distinction, and negative weights on all items on the other side. The rest of the entries will be zero. Hence this object analyzer vector will only be able to tell items apart with respect to this one distinction. It contains no information about higher or lower level distinctions in the tree. For trees with other branching factors, the situation is the same: additional object analyzer vectors are introduced to permit distinctions between more than two options, but these vectors contain no information about distinctions at other levels in the tree.

The association strength or singular value $s_{l}$ associated with level $l$ of the binary tree is

$$
\begin{equation*}
s_{l}=\sqrt{N P\left(\sum_{k=l}^{D-1} \frac{\Delta_{l}}{M_{l}}\right)} \tag{13}
\end{equation*}
$$

where $q_{k}=(1-4 \varepsilon(1-\varepsilon))^{D-1-k}$ and $\Delta_{l} \equiv q_{l}-q_{l-1}$, with the caveat that $q_{-1} \equiv 0$.


Figure 4: Statistical structure of hierarchical data. (a) Example hierarchical diffusion process with $D=4$ levels and branching factor $B=2$. (b) Analytically derived input singular vectors, or modes, (up to a scaling) of the resulting data, ordered top-to-bottom by singular value. Besides mode 0 , each mode, or object analyzer, can discriminate objects, or leaves of the tree, whose first common ancestor arises at a given level of the tree. This level is 0 for mode 2,1 for modes 3 and 4 , and 3 for modes 5 through 8 . Singular modes corresponding to broad distinctions (higher levels) have larger singular values, and hence will be learned earlier. (c) The covariance matrix between pairs of objects in the output feature space consists of hierarchically organized blocks.

While this equation gives the correct quantitative value for the association strength in terms of the parameters of the generative process, its most important property is its qualitative behavior: it is a decreasing function of the hierarchy level $l$ (see, e.g., Fig. 5). Crucially, this means that the input-output modes corresponding to broader distinctions like animal-plant have a stronger association strength than those corresponding to finer distinctions like bird-fish. Since we have previously shown that modes with stronger association strengths are learned more quickly, this immediately implies that broader distinctions among examples will be learned faster than fine-grained distinctions among examples.


Figure 5: Agreement between theoretically computed singular values in the limit of large numbers of features (obtained from (13)) and simulation for hierarchically structured data. The simulations show singular values arising from sampling 200 features from a hierarchical generative model with six levels, binary branching, and $\varepsilon=0.1$. The singular values are a decreasing function of the hierarchy level, implying that finer distinctions among examples will be learned more slowly.

Summary of the statistics of hierarchical data Thus we have shown that the singular vectors of data from a hierarchical diffusion process correspond exactly to the hierarchical distinctions in the underlying tree, and furthermore, that singular vectors corresponding to broader hierarchical distinctions have larger singular values than those corresponding to finer distinctions (Fig. 4AB). In combination with the preceding analysis of neural network learning dynamics, this result shows that our deep neural network must exhibit progressive differentiation on any dataset generated by an instance of this class of hierarchical, branching diffusion processes.

## Discussion

Our results explore the rich dynamics arising from gradient descent learning in a deep neural network, despite a completely linear input-output mapping. We have shown that these dynamics, driven solely by second order statistics, identify coherently covarying input and output modes in the learning environment, and we expressed the full time course of learning in terms of these modes. Finally, we moved beyond particular datasets to extract general principles by analyzing the covariance structure of hierarchical probabilistic models, showing that progressive differentiation is a general feature of learning hierarchically structured training data in deep neural networks.

We have focused our analysis on a few notable features of the learning dynamics-progressive differentiation and stagelike transitions-but our framework yields insights (to be presented elsewhere) into many other phenomena in semantic development such as, erroneous "illusory correlations" early in learning, familiarity and typicality effects, inductive property judgements, and the impact of perceptual correlations on learning dynamics. Moreover, this approach enables quantitative definitions of important intuitive notions like "category coherence", and yields precise theorems delineating how category coherence controls network learning rates.

By connecting probabilistic models and neural networks, our framework quantitatively links structured environments to learning dynamics. In future work, it will be important to compare the features of our neural network learning model with those of structured probabilistic learning models (e.g., Kemp and Tenenbaum (2008). Like structured models, neural networks can learn a range of different structure types, but unlike structured models, networks can learn without prior enumeration of such structures. Furthermore, networks can easily learn to represent data that are approximations or hybrids of different structure types-features that, we believe, characterize natural domains, such as the domain of living things considered here.

## Acknowledgments

S.G. thanks DARPA, BioX, Burroughs-Wellcome, and Sloan foundations for support. J.L.M. was supported by AFOSR. A.S. was supported by a NDSEG Fellowship and MBC Traineeship. We thank Juan Gao and Jeremy Glick for useful discussions.

## References

Inhelder, B., \& Piaget, J. (1958). The growth of logical thinking from childhood to adolescence. New York: Basic Books.
Keil, F. (1979). Semantic and conceptual development: An ontological perspective. Cambridge, MA: Harvard University Press.
Kemp, C., \& Tenenbaum, J. B. (2008, August). The discovery of structural form. Proceedings of the National Academy of Sciences of the United States of America, 105(31), 1068792.

Mandler, J. M., \& McDonough, L. (1993). Concept Formation in Infancy. Cognitive Development, 8, 291-318.
McClelland, J. L. (1995). A Connectionist Perspective on Knowledge and Development. In T. Simon \& G. Halford (Eds.), Developing cognitive competence: New approaches to process modeling. Hillsdale, NJ: Erlbaum.
Plunkett, K., \& Sinha, C. (1992). Connectionism and developmental theory. British Journal of Developmental Psychology, 10(3), 209-254.
Quinn, P., \& Johnson, M. (1997). The emergence of perceptual category representations in young infants: A connectionist analysis. Journal of Experimental Child Psychology, 66, 236-263.
Rogers, T., \& McClelland, J. (2004). Semantic cognition: A parallel distributed processing approach. Cambridge, MA: MIT Press.
Rumelhart, D., \& Todd, P. (1993). Learning and connectionist representations. In D. Meyer \& S. Kornblum (Eds.), Attention and performance xiv: Synergies in experimental psychology, artifical intelligence, and cognitive neuroscience. Cambridge, MA: MIT Press.
Siegler, R. (1976). Three aspects of cognitive development. Cognitive Psychology, 8, 481-520.

# Hierarchical Bayesian Modeling: Does it Improve Parameter Stability? 

Benjamin Scheibehenne (benjamin.scheibehenne@unibas.ch)<br>Economic Psychology, Department of Psychology, Missionsstrasse 62a 4055 Basel, Switzerland<br>Thorsten Pachur (pachur@mpib-berlin.mpg.de)<br>Center for Adaptive Rationality, Max Planck Institute for Human Development, Lentzeallee 94<br>14195 Berlin, Germany


#### Abstract

Fitting multi-parameter models to the behavior of individual participants is a popular approach in cognitive science to measuring individual differences. This approach assumes that the model parameters capture psychologically meaningful and stable characteristics of a person. If so, the estimated parameters should show, to some extent, stability across time. Recently, it has been proposed that hierarchical procedures might provide more reliable parameter estimates than nonhierarchical procedures. Here, we examine the benefits of hierarchical parameter estimation for assessing parameter stability using Bayesian techniques. Using the transfer-of-attention-exchange model (TAX; Birnbaum \& Chavez, 1997), a highly successful account of risky decision making, we compare parameter stability based on hierarchically versus non-hierarchically estimated parameters. Surprisingly, we find that parameter stability for TAX is not improved by using a hierarchical Bayesian as compared to a nonhierarchical Bayesian approach. Further analyses suggest that this is because the shrinkage induced by hierarchical estimation overcorrects for extreme yet reliable parameter values. We suggest that the benefits of hierarchical techniques may be limited to particular conditions, such as sparse data on the individual level or very homogenous samples.


Keywords: cognitive modeling; parameter consistency; risky choice; hierarchical Bayesian modeling; transfer-of-attentionexchange model

## Introduction

In cognitive science, a highly popular approach to describing and understanding behavior is to develop models with adjustable parameters that can be fitted to data. As parameters of cognitive models are usually supposed to represent meaningful aspects of cognitive processing, they are often used to study, measure, and describe individual differences between people. For illustration, consider cumulative prospect theory (CPT; Tversky \& Kahneman, 1992), one of the most prominent models of decision making under risk. According to CPT, responses to a risky alternative (which lead to different outcomes with particular probabilities) are a function of several factors including a person's sensitivity to probability information and her relative overweighting of losses as compared to gains ("loss aversion"). In the model, both probability sensitivity and loss aversion can be quantified by adjustable parameters, and several studies have employed CPT to investigate how
differences in age (Harbaugh, Krause, \& Vesterlund, 2002), gender (e.g., Fehr-Duda, Gennaro, \& Schubert, 2006), or personality (Pachur, Hanoch, \& Gummerum, 2010) affect risky decision making. Cognitive modeling thus allows individual differences in behavior to be decomposed into underlying cognitive components.

Using individually fitted model parameters to measure individual differences relies on the assumption of parameter stability-that is, that the parameter values estimated for a person remain relatively invariant across time (Yechiam \& Busemeyer, 2008). This applies in particular when modeling risky decision making, where it is often assumed that people's choices and their cognitive underpinnings reflect stable preferences (Yechiam \& Ert, 2011). In principle, however, it is possible that differences in parameter estimates between people simply reflect unsystematic variability (i.e., noise) rather than stable characteristics. In that case, fitting parameters of cognitive models would not be very useful because the results obtained would not generalize beyond a given task or situation.

Glöckner and Pachur (2012) found some evidence for temporal stability of the parameters of CPT: parameters fitted to individual participants' choices at each of two separate experimental sessions were (moderately) correlated. But does this finding also extend to other models of risky decision making? And-more importantly-do conclusions regarding a model's parameter stability depend on how the parameters are estimated? Whereas parameters are traditionally estimated independently for each single participant, it has recently been proposed that more reliable estimates might be achieved by using hierarchical Bayesian procedures, which exploit group-level distributions to inform individual-level estimation (e.g., Gelman \& Hill, 2007; Lee \& Webb, 2005).

Our goal is to examine whether conclusions regarding the parameter stability of a cognitive model are affected by the statistical method used to obtain the estimates. In particular, we compare hierarchical Bayesian techniques against nonhierarchical Bayesian procedures in a decision-making context. We investigate this issue for the transfer-of-attention-exchange model (TAX; Birnbaum \& Chavez, 1997), which has been claimed to provide a better account of decision making under risk than CPT (Birnbaum, 2008). For example, Birnbaum (2008) showed that TAX can correctly account for several violations of CPT, such as violations of gain-loss separability, coalescing, and
stochastic dominance, while being able to also accommodate apparent loss aversion and risk aversion.

## Hierarchical Bayesian Parameter Estimation

The application of hierarchical Bayesian techniques is becoming an increasingly popular tool to estimate cognitive models, including models of judgment and decision making (Lee \& Wagenmakers, 2013; Nilsson, Rieskamp, \& Wagenmakers, 2011; Scheibehenne, Rieskamp, \& Wagenmakers, 2013). Hierarchical Bayesian techniques are attractive because the approach naturally lends itself to the hierarchical data structure inherent in many psychological experiments, where a single individual provides many observations and researchers aim to draw conclusions on the aggregate group level. The alternative, "independence" approach, by contrast, is to first estimate the parameters of each individual participant separately and then aggregate these measures in a second step (Gelman \& Hill, 2007). While feasible, this approach ignores possible similarities between individuals and does not take into account that some participants might allow more precise and reliable estimates than others. Bayesian hierarchical techniques account for these differences and thus promise to yield more consistent and accurate estimates (Rouder \& Lu, 2005).
The Bayesian approach achieves this because the imposed hierarchical structure simultaneously informs the individual level, such that potentially unreliable individual estimates can borrow strength from the other estimates (Gelman, Carlin, Stern, \& Rubin, 2004). Furthermore, parameter estimates that are deemed unlikely given the distribution of the remaining parameter values (i.e., because they are located at the extremes of the distribution) are pulled closer towards the group mean and implicitly receive less weight. This property is referred to as "shrinkage." For these reasons, it has been argued that hierarchical methods often provide a more thorough evaluation of models in cognitive science (Shiffrin, Lee, Kim, \& Wagenmakers, 2008).
Though increasingly popular, Bayesian hierarchical implementations have been developed for only relatively few cognitive models of decision making under risk (but see Nilsson et al., 2011; Wetzels, Vandekerckhove, Tuerlinckx, \& Wagenmakers, 2010). Below we develop, to our knowledge for the first time, a hierarchical model for estimating individual participants' TAX parameters.

## Transfer-of-Attention-Exchange Model

TAX is a model of how people evaluate risky alternatives that can lead to certain outcomes $x$, each of which occurring with probability $p$. For instance, consider whether you would prefer to play a lottery with a $90 \%$ chance of winning $\$ 100$ (otherwise nothing) or a lottery with a $10 \%$ chance of winning $\$ 1000$ (otherwise nothing). According to TAX, the valuation of a lottery is a weighted average of the utilities of the outcomes; the weight that each outcome receives depends on its rank among all possible outcomes (the $n$ outcomes being ordered such that $x_{1}<x_{2}<x_{3} \ldots x_{\mathrm{n}}$ ) and its probability. To account for the typically found risk aversion
(risk seeking) in gains (losses), is the model assumes that attention (i.e., weight) is "transferred" from better (worse) to worse (better) outcomes. Specifically, the valuation, $V$, of a lottery $A$ is calculated as

$$
\begin{equation*}
V(A)=\frac{\sum_{i=1}^{n}\left[t\left(p_{i}\right)+\frac{\delta}{n+1} \sum_{j=1}^{i-1} t\left(p_{j}\right)-\frac{\delta}{n+1} \sum_{j=i}^{n} t\left(p_{i}\right)\right] u\left(x_{i}\right)}{\sum_{i=1}^{n} t\left(p_{i}\right)} \tag{1}
\end{equation*}
$$

where $\delta$ is a free parameter governing the attention shift from higher to lower outcomes (or vice versa); with $0<\delta<$ 1 attention is shifted from higher (lower) to lower (higher) outcomes in gains (losses), with $0>\delta>-1$ the opposite would occur. The function $u(x)$ is the utility function, $u(x)=$ $x^{\beta}$, transforming objective outcomes into subjective utilities. The free parameter $\beta$ indicates the curvature of the value function and reflects the decision maker's sensitivity to outcome information (with lower values of $\beta$ indicating lower sensitivity). $t(p)$ is the probability weighting function, transforming objective into subjective probabilities, and equals $t(p)=p^{\gamma}$. $\gamma$ is a free parameter reflecting the decision maker's sensitivity to probability information (with lower values of $\gamma$ indicating lower sensitivity). To derive the predicted probability that lottery $A$ is preferred over lottery B, we used an exponential version of Luce's choice axiom:

$$
\begin{equation*}
p(A, B)=\frac{e^{\theta \cdot V(A)}}{e^{\theta \cdot V(A)}+e^{\theta \cdot V(B)}}, \tag{2}
\end{equation*}
$$

where $\theta$ is a choice sensitivity parameter, indicating how sensitively a decision maker reacts to differences in the valuation of lotteries $A$ and $B$. To summarize, TAX as implemented here has four free parameters: attention shift $(\delta)$, outcome sensitivity ( $\beta$ ), probability sensitivity $(\gamma)$, and choice sensitivity $(\theta)$.

Data We applied TAX to model the data reported in Glöckner and Pachur (2012). In this study, 63 participants ( 25 male, mean age 24.7 years) indicated their preference between two-outcome monetary lotteries at two experimental sessions that were one week apart. At each session, the participants were presented (on a computer) with 138 lottery problems that contained pure gain, pure loss, and mixed lotteries, all drawn from sets of lottery problems used in previously published studies; 38 of the problems were shown at both sessions (see Glöckner \& Pachur for details). The outcomes of the lotteries ranged from $-€ 1000$ to $€ 1200$. At the completion of each session, one of the chosen lotteries was picked randomly, played out, and the participant received an additional payment proportional to the outcome.

## Parameter Estimation

To estimate the free parameters of TAX, we implemented two Bayesian versions of the model-a hierarchical version and an independent (i.e., non-hierarchical) version. Bayesian modeling requires a detailed specification of the likelihood function and the prior probability distributions of all model parameters. For the independent version, we
specified the likelihood function based on Equations (1) and (2). The priors for the free parameters were set to uniform probability distributions that span a "reasonable" range that excluded theoretically implausible values and allowed for ample space to include parameter values found in previous research (Michael Birnbaum, personal communication). In particular, the priors ranged from -2 to 2 for the $\delta$ parameter and from 0 to 5 for the $\beta, \gamma$, and $\theta$ parameters.
In the hierarchical version, we utilized the same functions as in the independent version but partially pooled the individual parameters using normally distributed grouplevel distributions. Uninformative priors were assigned to the respective means and standard deviation of these grouplevel distributions. The group-level means were assumed to be normally distributed with mean 0 and variance 1 . The prior on the group-level standard deviation was uniformly distributed ranging from 0 to 10 . To ensure proper parameter scaling, the group-level parameters were linked onto the individual level through a probit transformation (Rouder \& Lu, 2005). As this transformation yields a parameter range from 0 to 1 on the individual level, an additional, linear linkage function was interposed that stretched the parameter range to match the scale used in the independent model implementation outlined above (i.e., a range from -2 to 2 for the $\delta$ parameter, and a range from 0 to 5 for the $\beta, \gamma$, and $\theta$ parameters).

For both the individual and the hierarchical model we estimated the joint posterior parameter distributions using Monte Carlo Markov Chain methods implemented in JAGS, a sampler that utilizes a version of the BUGS programming language (Lunn, Spiegelhalter, Thomas, \& Best, 2009; Plummer, 2011) that was called from the R statistics software (version 2.14.0; R Core Team, 2012). A total of 10,000 representative samples were drawn from the posterior distributions after a "burn-in" period of 1,000 samples. The sampling procedure was efficient as indicated by a low autocorrelation of the samples, the Gelman-Rubin statistic, and visual inspection of the chains.

## Quantifying Parameter Stability

To the extent that the parameters of a cognitive model capture stable characteristics of an individual, the parameters should be invariant across time-at least for relatively short time intervals and under comparable measurement conditions (Bland \& Altman, 1986). One way to quantify parameter stability (or reliability) is to correlate individual parameter estimates between two points in time (i.e., test and re-test). Higher correlations indicate higher parameter stability.

As outlined above, one rationale for using hierarchical Bayesian techniques for parameter estimation is to obtain more reliable estimates. Thus, one might expect a higher test-retest correlation when parameters are estimated hierarchically than when they are estimated for each participant independently. To test this prediction, we calculated correlations between the parameter values estimated for each participant at the two measurement
points ( tl and t 2 ), separately for the individual model and the hierarchical model.

Correlations were calculated based on the mean posterior parameter estimates for each measurement point, using Bayesian techniques implemented in BUGS. A Bayesian approach to calculating correlations allows correlation coefficients to be compared based on their posterior distributions. This avoids many problems inherent in traditional frequentist statistical procedures that rely on nullhypothesis significance testing (Kruschke, 2011).

## Results

Table 1 reports the best-fitting TAX parameter values on the group level, obtained from the hierarchical model. As indicated by the $\delta$ parameter being larger than 0 , participants displayed risk aversion in gains and risk seeking in losses, and some reduced sensitivity to outcomes ( $\beta$ being smaller than 1) and probabilities ( $\gamma$ being smaller than 1). Overall, the parameter values obtained are within the range of values obtained or used in previous applications of TAX (e.g., Birnbaum, 2008).

Table 1: Best-fitting group-level TAX parameters and their $95 \%$ highest density intervals $\left(\mathrm{HDI}_{95}\right)$.

|  | TAX parameters |  |  |  |
| :---: | :---: | :---: | :---: | :---: |
|  | $\delta$ | $\beta$ | $\gamma$ | $\theta$ |
| t1 |  |  |  |  |
| M | . 33 | . 65 | . 64 | . 14 |
| $\mathrm{HDI}_{95}$ | [.25,.40] | [.62,.68] | [.57,.71] | [.11,.16] |
| t2 |  |  |  |  |
| M | . 35 | . 63 | . 61 | . 16 |
| $\mathrm{HDI}_{95}$ | [.27,.43] | [.60,.65] | [.51,.72] | [.13,.20] |

Figure 1 shows Pearson's product-moment correlations (across participants) between t 1 and t 2 for each of the four TAX parameters. As can be seen, the mean correlation coefficient for the $\delta$ and the $\gamma$ parameters is slightly higher when they are estimated hierarchically than when they are estimated independently. However, this difference is not credible, as the $95 \%$ highest posterior density interval $\left(\mathrm{HDI}_{95}\right)$ includes zero. For the $\beta$ parameter, the correlation is slightly higher when parameters are estimated independently, and for the $\theta$ parameter the test-retest correlation is clearly lower for the hierarchical than for the independent estimates. A similar picture emerges based on Spearman's rank correlations (not shown).

## Why Does Hierarchical Estimation Fail to Improve Parameter Stability?

The results indicate that applying a hierarchical TAX model does not yield higher parameter stability on the individual level. At first sight, this seems surprising given the supposed advantages of hierarchical techniques that "borrow strength" from distributional information on the group level to improve estimations on the individual level.


Figure 1: Stability of each TAX parameter as indicated by the mean product-moment correlation across participants between t 1 and t 2 . Circles indicate independent estimates, triangles indicate hierarchical estimates. Error bars $=\mathrm{HDI}_{95}$.

To explore the reasons for this result, it is instructive to take a closer look at the distribution of the parameter estimates obtained. For illustration, Figure 2 displays the posterior means for the independently estimated $\beta$ parameter values at t 1 and t 2 (upper and lower row, respectively) as well as the hierarchically informed estimates at t 1 (middle row) for a subset of 20 representative participants; for each person the estimates are connected by a line. As could be expected, given the partial pooling enforced through the introduction of the higher level group distribution in the hierarchical model, the hierarchical estimates show a lower dispersion than the individually estimated parameters (the same holds for the hierarchical estimates at t 2 , which are not shown). This shrinkage is particularly pronounced for extreme parameter estimates, that is, those that are far away from the group-level mean. The reason is that these estimates appear rather unlikely with respect to the grouplevel distribution and are thus implicitly treated as outliers in the hierarchical model.

Unwarranted Shrinkage Importantly, however, Figure 2 further shows that the shrinkage of the hierarchical method is not necessarily warranted: for the independently estimated parameter values there is rather good correspondence between tl and t 2 even for participants with rather extreme parameter values. That is, individuals who have a high $\beta$ value at t 1 also tend to have a high $\beta$ value at t 2 ; the same applies for small $\beta$ values. Thus, our analysis shows that in the context on hand extreme estimates often reflect meaningful and reliable characteristics of individuals. The partial pooling enforced by the hierarchical modeling somewhat distorts the individual parameter estimates by pulling them too much towards the group-level mean.


Figure 2: Mean posterior estimates of the $\beta$ parameter of TAX separately for each individual at t 1 and t 2 (upper and lower row) and the hierarchically estimated parameters at tl (middle row) for a representative subset of 20 participants.

Diminished test-retest correlation The unwarranted shrinkage imposed by hierarchical modeling does not inevitably lead to lower test-retest correlations. After all, it could be that the compressed hierarchical estimates are nevertheless more reliable and thus stable over time than the (more dispersed) parameter values estimated on the individual level. As we will outline next, however, that does not seem to be the case.
Figure 3 displays a scatterplot for the $\theta$ parameter separately for the independent and the hierarchical estimates. The $\theta$ parameter provides an instructive example because here the difference between the correlations for the individual and the hierarchical estimates is particularly large (Figure 1). Figure 3 shows that the high correlation for the independent estimates is partly due to some individuals having high values on the $\theta$ parameter at both measurement points. As mentioned above, although these values are much higher than for most individuals in the group, they nevertheless seem to be reliable in the sense that they are equally high at both measurement points. In contrast, the range of the hierarchically estimated parameters is much narrower (note that the axis scales in the figure were adjusted to best display the data). Furthermore, the hierarchical model seems to affect the individual parameter estimates to different degrees. This occurs because the influence of the group-level depends, among other aspects, on the variance and the mean of the individual estimates. As indicated by the shape of the scatterplot in the lower panel of Figure 3, this effect pulls the parameter estimates towards the mean and thus leads to a lower (linear) relationship between the two measurement points. In that sense, the hierarchical method also induces shrinkage on the
correlation coefficients. In situations where the correlation of the individually estimated parameters is reduced due to unreliable outliers, however, applying hierarchical techniques will shrink these outliers and may then yield higher parameter stability.


Figure 3: Scatter plot of the mean posterior estimates for the $\theta$ parameter at tl and t 2 . Each point represents one participant. The upper panel shows the parameter values obtained by individual estimation; the lower panel shows the parameter values obtained by hierarchical estimation. Note that the value ranges on the axes are much smaller in the lower panel.

## Discussion

The psychological content and generalizability of a cognitive model hinges on the extent to which its parameters reflect stable characteristics of an individual. Conclusions regarding a model's parameter stability may be affected by the statistical procedures used to estimate these parameters. Specifically, researchers must decide whether to employ hierarchical techniques or to estimate each person individually.

Our analyses show that the free parameters of the TAX model are rather consistent across time, indicating that the model captures stable aspects of decision makers' risk attitude and their outcome and probability sensitivity. This finding parallels previous results obtained for CPT based on the same data using maximum likelihood estimates (Glöckner \& Pachur, 2012). Most importantly-and rather unexpectedly-our analysis provided no evidence that hierarchical Bayesian techniques yield more stable parameter estimates than the alternative approach of estimating each participant independently from the others.

Why did the shrinkage of the hierarchical procedure yielding distorted estimates? In principle, one possibility is that the distribution of the individual parameter values is bimodal, which would render group-level means futile. As indicated by visual inspection, however, the parameter distributions for our data were mostly unimodal in shape, so this cannot explain why the hierarchical procedure distorted the estimates.

Another possibility could be the prior distribution used for shrinkage. To achieve an optimal balance between complete pooling and complete independence, the degree of shrinkage in the hierarchical model is represented by a free parameter (representing the variance of the group-level distributions) estimated from the data. In principle, the choice of prior on the variance could impose an unwarranted amount of shrinkage (i.e., a low variance), for instance, if much weight is put on low variances, or if the prior does not allow for large variances in the first place. For the current data, however, the posterior estimates for the group-level standard deviations were far away from the upper boundaries of the uniform prior distributions on the grouplevel. The choice of prior on the variance of the group-level distributions is thus an unlikely reason for the undue amount of shrinkage.

Generalizability Although our demonstration focused on one particular cognitive model, we suspect that the conclusions will hold for other models as well-particularly in the domain of judgment and decision making; here, people often rely on different strategies (e.g., Pachur \& Olsson, 2012; Scheibehenne et al., 2013) and parameter heterogeneity thus reflects genuine differences between people. In such a case, the parameter estimates will not regress towards the mean if more data or more precise measures are collected.

Advantages of Hierarchical Approaches The case on hand may be different from situations in which hierarchical Bayesian techniques have been shown to outperform independent parameter estimates. In a classic example, Efron and Morris, (1975) predicted the success rate of professional baseball players for an entire season based on their success rate early in the season. This prediction was greatly improved through the application of hierarchical techniques. Presumably, this improvement occurred because the differences in the success rates of professional baseball players are rather small (they are all pretty good players) and random noise will equal out throughout the season.

Under this condition, there will be regression towards the mean, which benefits hierarchical Bayesian techniques.

Another situation in which hierarchical Bayesian estimates presumably provide more accurate results than independent estimates is when only very little data is available for each individual, yielding high uncertainty on the individual level. Here, the unreliability on the individual level might be reduced through partial pooling.

Finally, hierarchical modeling techniques might be beneficial for comparisons on the group level (Gelman \& Hill, 2007), where the goal is not to improve the reliability on the individual level but to derive robust estimates for the group as a whole. As a consequence, the implicit weighting imposed through hierarchical estimation methods might also be advantageous for making out-of-sample predictions for new group members.

## Summary

Our results indicate that hierarchical Bayesian techniques do not necessarily improve the reliability of individual parameter estimates. Therefore, researchers aiming to predict individual behavior may be better advised to rely on individual estimates instead. As discussed above, hierarchical models might have specific strengths in situations in which very little information is available on the individual level, when the group is very homogenous, or when the goal is to describe and compare groups as a whole.

## References

Birnbaum, M. H. (2008). New paradoxes of risky decision making. Psychological Review, 115, 463-501.
Birnbaum, M. H., \& Chavez, A. (1997). Tests of theories of decision making: Violations of branch independence and distribution independence. Organizational Behavior and Human Decision Processes, 71, 161-194.
Bland, J. M., \& Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. Lancet, i, 307-310.
Efron, B., \& Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. Journal of the American Statistical Association 70, 311-319.
Fehr-Duda, H., de Gennaro, M., \& Schubert, R. (2006). Gender, financial risk, and probability weights. Theory and Decision, 60, 283-313.
Gelman, A., Carlin, J. B., Stern, H. S., \& Rubin, D. B. (2004). Bayesian data analysis (2nd ed.). Boca Raton, FL: Chapman \& Hall/CRC.
Gelman, A., \& Hill, J. (2007). Data analysis using regression and multilevel/hierarchical models. Cambridge: Cambridge University Press.
Glöckner, A., \& Pachur, T. (2012). Cognitive models of risky choice: Parameter stability and predictive accuracy of prospect theory. Cognition, 123, 21-32.
Harbaugh, W. T., Krause, K., \& Vesterlund, L. (2002). Risk attitudes of children and adults: Choices over small and
large probability gains and losses. Experimental Economics, 5, 53-84.
Kruschke, J. K. (2011). Doing Bayesian data analysis: A tutorial with R and BUGS. Amsterdam: Elsevier.
Lee, M. D., \& Wagenmakers, E.-J. (2013). Bayesian cognitive modeling: A practical course. Cambridge University Press.
Lee, M. D., \& Webb, M. R. (2005). Modeling individual differences in cognition. Psychonomic Bulletin and Review, 12, 605-621.
Lunn, D., Spiegelhalter, D., Thomas, A., \& Best, N. (2009). The BUGS project: Evolution, critique and future directions. Statistics in Medicine, 28, 3049-3067.
Nilsson, H., Rieskamp, J., \& Wagenmakers, E.-J. (2011). Hierarchical Bayesian parameter estimation for cumulative prospect theory. Journal of Mathematical Psychology, 55, 84-93.
Pachur, T., Hanoch, Y., \& Gummerum, M. (2010). Prospects behind bars: Analyzing decisions under risk in a prison population. Psychonomic Bulletin and Review, 17, 630-636.
Pachur, T., \& Olsson, H. (2012). Type of learning task impacts performance and strategy selection in decision making. Cognitive Psychology, 65, 207-240.
Plummer, M. (2011). JAGS Version 3.1.0. Retrieved from $\mathrm{http}: / / \mathrm{mcmc}$-jags.sourceforge.net/
R Core Team (2012). R: A language and environment for statistical computing. $R$ Foundation for Statistical Computing. Vienna, Austria.
Rouder, J. N., \& Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. Psychonomic Bulletin and Review, 12, 573-604.
Scheibehenne, B., Rieskamp, J., \& Wagenmakers, E.-J. (2013). Testing adaptive toolbox models: A Bayesian hierarchical approach. Psychological Review, 120, 39-64.
Shiffrin, R. M., Lee, M. D., Kim, W., \& Wagenmakers, E.J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. Cognitive Science, 32, 1248-1284.
Tversky, A., \& Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and Uncertainty, 5, 297-323.
Wetzels, R., Vandekerckhove, J., Tuerlinckx, F., \& Wagenmakers, E.-J. (2010). Bayesian parameter estimation in the Expectancy Valence model of the Iowa gambling task. Journal of Mathematical Psychology, 54, 14-27.
Yechiam, E., \& Busemeyer, J. R. (2008). Evaluating generalizability and parameter consistency in learning models. Games and Economic Behavior, 63, 370-394.
Yechiam, E., \& Ert, E. (2011). Risk attitude in decision making: Search for trait-like constructs. Topics in Cognitive Science, 3, 166-186.

# Subjective Awareness during Cross-Language Speech Perception Attending Unattended Regions of an Acoustic Continuum 

Jordan Schoenherr (psychophysics.lab@gmail.com) John Logan (john_logan@carleton.ca)<br>Department of Psychology, Carleton University<br>1125 Colonel By Drive, Ottawa, ON K1S5B6 Canada


#### Abstract

Linguistic experience attenuates adult listeners' attention to acoustic differences that are not phonemic in the listener's language. In the present study we found that acoustic information is available to listeners after acoustic cues have been processed to identify phonemic categories. Moreover, we also found that listeners maintained an awareness of these differences by comparing the identification function to typicality ratings and confidence reports.


Keywords: speech perception, category boundaries, confidence processing

## Introduction

When adult listeners are presented a continuum of speech stimuli varying along an acoustic dimension, they divide the continuum into distinct categories defined by sharp boundaries (Liberman, Harris, Hoffman, \& Griffith, 1957), a phenomenon known as categorical perception (Liberman, Cooper, Shankweiler, \& Studdert-Kennedy, 1967). Categorical perception was originally taken as evidence that adult listeners are no longer capable of detecting stimulus differences in within-category regions of a speech continuum (Eimas, 1975). A developmental account of categorical perception assumes that when infants acquire phonemes they learn to segment acoustic information into discrete categories. Consistent with this view, research indicates that by the end of their first year, infants develop a reduced sensitivity to differences between stimuli within a given phoneme category (Werker, 1989), with category boundaries becoming more distinct as they become older (e.g., Hazan \& Barrett, 2000). Rather than processing all stimulus dimensions, infants learn to attend to specific acoustic cues that determine category membership (Jusczyk, 2000). The question that the present study attempts to answer is whether dults are subjectively aware of attended and unattended acoustic properties.

Extending the concept of categorical perception to the perception of non-native phonemes implies that adult listeners might experience difficulties in perceiving these speech sounds due to a desensitization to acoustic information. In effect, non-native speech sounds are thought to be perceptually assimilated into existing native phonemic categorical representations (cf. Best, 1995, Flege, 1995). Although previous research indicates that listeners can learn to perceive non-native phoneme categories (e.g., Pisoni, Aslin, Perey, \& Hennessy, 1982), limited work has been done to examine the metacognitive awareness associated with learning a non-native phoneme category. In the present
study, we examined whether feedback could be used to allocate attention to a previously unattended region of an acoustic continuum that corresponded to a non-native phoneme category and whether attention was accompanied by awareness.

## Categorical Perception of Speech Sounds

A variety of acoustic cues are used by listeners to define phoneme categories. Voice-onset time (VOT), the interval between aspiration and the vibration of the vocal cords is one such cue. Lisker and Abramson (1967) presented English and Thai listeners with synthetic speech sounds ranging from -150 VOT (prevoiced) to +150 VOT (voiced). The Thai listeners' identification performance resulted in two category boundaries corresponding to $/ \mathrm{p} /$, $/ \mathrm{b} /$, and $/ \mathrm{p}^{\mathrm{h}} /$ whereas English listeners' identification performance yielded only one category boundary corresponding to the phonemes $/ \mathrm{b} /$ and $/ \mathrm{p} /$. These findings illustrate how nonnative speech sounds are assimilated into existing phonemic categorical structure.

## Learning to perceive non-native speech sounds

When categorizing stimuli, participants must attend to stimulus properties along the physical continuum that defines the stimuli (Nosofsky, 1986). With training, participants become sensitized to specific regions along the continuum that can affect performance in other tasks (e.g., discrimination tasks; Goldstone, 1994). Thus, once a phonemic category has been acquired listeners will limit their attention to only those psychophysical characteristics where attention has been directed (Jusczyk, 1992).

In order to promote attention to previously unattended acoustic characteristics, listeners typically require some form of training. For example, in order to examine the effectiveness of feedback, Pisoni et al. (1982) provided English listeners with three exemplars of speech sounds from the three regions of the VOT continuum corresponding to voiceless unaspirated, voiced aspirated, and voiceless aspirated stops (i.e, /p/, /b/, and $/ \mathrm{p}^{\mathrm{h}} /$, respectively). Following a short period of laboratory training, listeners were capable of identifying and discriminating speech sounds from the non-native $/ \mathrm{p}^{\mathrm{h}} /$ category.

Pisoni et al.'s (1982) results have several implications. First, acoustic information must be accessible to listeners in order to classify stimuli into a non-native phoneme category. Further support for the availability of acoustic information also comes from studies using typicality ratings wherein gradedness is exhibited in response functions (e.g., Miller \& Volaitis, 1989). Second, when compared to
previous studies (e.g., Mackain et al., 1981) the amount of exposure to non-native phonemes is less important than the method used to present the non-native phonemes. Finally, methods that allocate selective attention to previously unattended regions can facilitate acquisition of these phonemic categories. Although attention is drawn to these stimulus features, it remains unclear whether participants have a subjective awareness of these acoustic properties.

## Confidence reports and subjective awareness

One method frequently used to quantify subjective awareness in a perceptual discrimination task is to have participants provide a subjective probability that they have provided a correct answer. The use of numeric response scales is particularly useful given that a direct comparison can be made with the proportion of correct responses obtained from the primary task. In this case, $50 \%$ represents a response associated with a guess whereas $100 \%$ represents absolute certainty in a response. Thus, listeners are well calibrated when they assign a level of confidence (e.g., $60 \%$ ) that corresponds to their obtained accuracy (e.g., $p(c o r)=0.6)$. Listeners frequently demonstrate failures in assigned appropriate subjective probabilities, evidencing miscalibration.

Rather then miscalibration being random, systematic deviations have been observed which some argue represent a differential ability to assess the performance of specific cognitive operations (Dawes, 1980), with still others arguing that these measures represent contributions of implicit and explicit knowledge (e.g., Dienes \& Berry, 1997; cf. Lichtenstein \& Fischhoff, 1977). For instance, underconfidence has been observed in perceptual tasks (e.g., Bjorkman, Juslin, \& Winman, 1993) whereas overconfidence is typically observed in tasks requiring assess to general knowledge (e.g., Gigerenzer, Hoffrage, \& Kleinbolting, 1991). Dawes (1980) has argued that these findings are the result of participants' uncertainty in their perceptual processes and certainty about information stored in long-term memory. This suggests that confidence processing uses a second set of operations to assess the content and output of the primary decision process (e.g., stimulus classification).

An alternative class of confidence models assumes an additional set of operations wherein primary decision information is rescaled (e.g., for a review see, Baranski \& Petrusic, 1998). These accounts are supported by the observation that the requirement of confidence reports increase primary decision response time (Baranski \& Petrusic, 1998) and the dissociable effects of nondiagnostic information on the primary decision response selection and confidence reports (e.g., Schoenherr, Leth-Steensen, \& Petrusic, 2010). In the present study, we assume that both acoustic and phonemic information will affect confidence reports and that the correspondence of identification accuracy and mean confidence will demonstrate whether
participants have an awareness of acoustic information when identifying phonemes.

## Present Study

In the present study, English listeners were asked to classify stimuli from a VOT continuum into phoneme categories corresponding to either two categories, /p/ and /b/ (both found in English), or three categories, /p/, b/, plus the prevoiced category $/ \mathrm{p}^{\mathrm{h}} /$ (found in Thai). If listeners are aware of acoustic differences, their subjective confidence should differ across regions of the VOT continuum as evidenced by miscalibration. If underconfidence is evidenced, it suggests that listeners did not have subjective awareness of a well-defined phonemic category whereas if overconfidence is evidenced, it suggests that listeners believed that they had a better understanding of the phonemic category then they in fact did.

## Experiment 1

Previously, Schoenherr, Logan, and Winchester (2012) observed slight overconfidence for responses made for $/ \mathrm{b} /$ stimuli located at the category boundary on the VOT continuum. In Experiment 1, we sought to see whether we could increase uncertainty in this region by presenting stimuli from the prevoiced region. The addition of stimuli with these novel acoustic properties should increase uncertainty for stimuli from the $/ \mathrm{p}^{\mathrm{h}}-\mathrm{b} /$ portion of the continuum if acoustic properties are being used to identify stimuli whereas the participants should remain certain in their judgments of stimuli associated with the /p/ category.

## Method

## Participants

Nine Carleton University students participated in the study for course credit; all were native speakers of English or had extensive experience with English and reported normal hearing and no speech pathologies.

## Materials

Fifteen synthetic speech stimuli were used, obtained from the Haskins Laboratories website (HL, 2011; Lisker \& Abramson, 1967). These stimuli varied along the VOT continuum from -70 to 70 ms VOT. As per the method used by Pisoni et al. (1982), listeners were presented with stimuli which corresponded to the voiceless unaspirated, voiced aspirated, and voiceless aspirated stops. The latter categories are present in English while the former is not. The sounds were originally recorded on reel-to-reel tape and later converted into AIFF format at Haskins Laboratories. Stimuli were pre-processed using a DC offset correction to eliminate clicks present in the AIFF versions and then converted into WAV files.

Figure 1a. ID Function without the Requirement of Confidence


## Procedure

Modelled after the procedure used by Pisoni et al. (1982), listeners were presented with a brief training block in which they heard three stimuli prior to the identification tasks, one from each region of the VOT continuum ( $-70,0$, and 70 ms VOT, corresponding to the $/ \mathrm{p} / \mathrm{h} / \mathrm{h} /$, and $/ \mathrm{p} /$ categories). Ten replications of these stimuli were presented in the order $/ \mathrm{p}^{\mathrm{h}} /, / \mathrm{b} /$, and $/ \mathrm{p} /$. Following this initial training, listeners also were trained using an identification (ID) task with feedback. They were presented with a stimulus and then reported whether it was a $/ \mathrm{p}$ ' $/ \mathrm{/} / \mathrm{b} /$, or $/ \mathrm{p} /$ using the ' V ', ' B ', or ' N ' keys, labeled as '_B', 'B', and 'P', respectively. After they had indicated their response, 'Correct' or 'Incorrect' was presented visually on the screen. Listeners completed a total of 80 trials in the training task.

Following training, listeners again identified the stimulus presented as a $/ \mathrm{p} /$ /, $\mathrm{b} /$, or $/ \mathrm{p} /$ using the keyboard but no feedback was provided. In the first block, after they completed each ID trial they also indicated their level of confidence in their response using the 'E' through 'I' keys, on a 6 -point scale with $50 \%$ representing a guess and $100 \%$ representing certainty. In the second block, confidence was not reported. Each block was composed of a total of 150 trials.

The duration of the experiment was approximately 30 minutes. Replicating the procedure of Pisoni et al. (1982), stimuli were presented randomly except for the sequential presentation of the training stimuli. Listeners were presented with the stimuli over headphones at a comfortable listening level using PsychoPy software (Peirce, 2007).

## Results

Proportion Identification. Unlike studies that have examined two-category identification performance using confidence reports (e.g., Schoenherr et al., 2012), only the $/ \mathrm{p} /$ phoneme category showed a sharp identification function (Figures 1a and 1b). In general, however, listeners could consistently identify stimuli associated with the $/ \mathrm{p}$ h/ and $/ \mathrm{b} /$ category with greater than chance accuracy (i.e., stimuli with VOTs of $-70,-60,-50,0$, and 10 ), indicating that even with a brief period of training, listeners can begin to acquire a non-native speech category. Supporting this, we obtained a significant effect for VOT stimulus, $F(14,112)=7.389$, $M S E=.435, p=.001, \eta^{2}=.480$. Given that we did not

Figure 1b. Identification Function with the Requirement of Confidence

obtain a main effect or interaction of confidence reports, it suggests that confidence reports did not significantly affect ID performance thereby permitting a straightforward interpretation of the remaining results.
(In an alternate version of Experiment 1 the order of confidence and no confidence blocks was reversed [see Schoenherr and Logan, in preparation]. Under these conditions, participants' ( $n=12$ ) identification performance for the non-native phoneme category was reduced in the first block when confidence was required. Participants generally paired neighbouring stimuli together in alternating clusters of 2-4 speech sounds [e.g., VOTs 10 and 0 ms were identified as $/ \mathrm{b} /$, VOTs -10 and -20 ms identified as $/ \mathrm{p}^{\mathrm{h}} /$, and VOTs -30 to -60 ms identified as $/ \mathrm{b} /$ ]. Such a finding suggests that the requirement of producing confidence reports can interfere with the primary task.)

Table 1. Mean identification RTs (ms) along the critical regions of the VOT continuum with standard error reported in parentheses.

| Condition | $/ \mathbf{p}^{\mathrm{h}} /$ | $/ \mathbf{p}^{\mathrm{h}}-\mathbf{b} /$ | $/ \mathbf{b} /$ | $/ \mathbf{b}-\mathbf{p} /$ | $/ \mathbf{p} /$ |
| :--- | :--- | :--- | :--- | :--- | :--- |
| No Conf. | 888 | 910 | 893 | 933 | 796 |
|  | $(45)$ | $(51)$ | $(54)$ | $(75)$ | $(43)$ |
| Conf. | 1,009 | 1,137 | 1,072 | 992 | 855 |
|  | $(47)$ | $(124)$ | $(113)$ | $(83)$ | $(41)$ |

Identification Response Time. Prior to conducting an analysis of the response time (RT) data, we collapsed stimuli into regions five regions along the VOT continuum corresponding to two category boundaries (CBs) $/ \mathrm{p}^{\mathrm{h}}-\mathrm{b} /$ and $/ \mathrm{b}-\mathrm{p} /$ corresponding to $\mathrm{CB}_{1}(-30,-20)$ and $\mathrm{CB}_{2}(20,30)$, respectively, and equivalent within-category pairs corresponding to $/ \mathrm{p}^{\mathrm{h}} /(-70,-60), / \mathrm{b} /(0,10)$, and $/ \mathrm{p} /(60,70)$, respectively. Using the criterion of 3 standard deviations, $4.3 \%$ of the responses were identified as outliers and removed from the final analysis.

An analysis of the remaining responses times revealed a main effect of VOT region, $F(4,32)=4.45, M S E=.041, p=$ $.025, \eta^{2}=.357$. Table 1 indicates that response latencies were longer at category boundaries as well as for the nonnative $\left(/ \mathrm{p}^{\mathrm{h}} /\right)$ and modified native ( $/ \mathrm{b} /$ ) categories relative to the native $/ \mathrm{p} /$ category. A main effect of the requirement of confidence report was also obtained, $F(1,18)=14.55, M S E$
$=.026, p=.005, \eta^{2}=.645$. Again, Table 1 demonstrates longer latencies with the requirement of confidence relative to the no confidence condition. Given that the confidence block always followed the no confidence block, this finding cannot be attributed to automaticity. The interaction of confidence condition and VOT region was only marginally significant, $F(4,32)=2.724, M S E=.019, p=.099, \eta^{2}=$ . 254.
Confidence Reports. Figure 1a and 1 b also demonstrate the effect of confidence measures. Listeners expressed less confidence in their responses to stimuli located within the $/ \mathrm{p}$ / and /b/ categories. As was the case with ID accuracy, we observed a main effect of the stimulus location along the VOT continuum on mean confidence, $F(14,112)=6.931$, $M S E=1011.371, p=.018, \eta^{2}=.464$. Our comparison of over/underconfidence bias did not reveal any significant effects, $F(14,112)=2.146, M S E=.0354, p=.133, \eta^{2}=$ .212. Unlike previous studies that have observed graded responses indicative of perception of acoustic properties (e.g., McMurray, Tanenhaus, \& Aslin, 2002), our findings suggest that listeners are not fully aware of the processes allowing them to identify stimuli.

## Discussion

An instructive comparison can also be made between the results of the present study and conditions in which participants identify two native phonemes. Schoenherr et al. (2012) observed a small decrease in confidence around the /b-p/ category boundary when only the voiced and voiceless portions of the continuum were presented to listeners. When the prevoiced portion of the continuum was additionally presented, Schoenherr et al. (2012) observed lower confidence in the $/ \mathrm{p}^{\mathrm{h}}-\mathrm{b} /$ portion of the continuum. When compared to identification accuracy, this pattern of responses leads to underconfidence in comparison to the overconfidence observed in the present study. Taken together with our results, this suggests that training does result in the allocation of attention to newly relevant acoustic properties of the stimuli in this region of the VOT continuum, thereby reducing certainty. Although listeners are somewhat more conservative in their confidence reports, they are still overconfident suggesting that the phonemic representations that they are subjectively aware of are less accurate than the acoustic information necessary to identify the stimuli.

## Experiment 2

The pattern of overconfidence observed in Experiment 1 suggests two features of a listener's awareness of acoustic properties. First, the observation that mean confidence exceeded identification accuracy in the $/ \mathrm{p}^{\mathrm{h}}-\mathrm{b} /$ region of the VOT continuum indicates that listeners did not have complete access to acoustic properties on a trial-to-trial basis. Second, mean confidence in identification responses to stimuli within the $/ \mathrm{p}^{\mathrm{h}}-\mathrm{b} /$ region was lower relative to stimuli from the $/ \mathrm{p} /$ region. These findings suggest that
while participants confidence reports are influenced by their two native phoneme categories, they might have some awareness of the acoustic properties of stimuli within the $/ \mathrm{p}^{\mathrm{h}}-\mathrm{b} /$ region which cause uncertainty in their responses. To assess the extent to which acoustic and phonemic properties are available to subjective awareness, Experiment 2 required participants to additionally provide typicality ratings that have previously been shown to reflect acoustic properties (Miller \& Volaitis, 1989).

## Method

## Participants, Materials, and Procedure

Fifteen Carleton University students participated in the study. Other participant and stimulus characteristics were the same as in Experiment 1.

We replicated the methods of identification block in Experiment 1 and required post-decisional confidence reports. In a subsequent another block, listeners rated the typicality of each stimuli on a scale of 1 through 9 where ' 9 ' represented highly typical of a category and ' 1 ' represented a highly atypical member of a category. Listeners were presented with the same number of trials as the identification block.

## Results

Proportion Identification. Replicating the general results of Experiment 1, the location of the stimuli along the VOT continuum significantly affected identification performance, $F(14,112)=9.149, M S E=.124, p=.005, \eta^{2}=$ .533. Figure 1b demonstrates, participants had a sharp category boundary between stimuli for the $/ \mathrm{b} /$ and $/ \mathrm{p} /$ categories. A noticeable difference was evident in the location of the boundary. Whereas in Experiment 1 the boundary was located between VOT 20 ms and 30 ms , a shift such that the boundary was now located at VOT 20 ms with a resulting decrement in performance for VOT 10 ms stimuli. We assume that these results are due to range effects. In general, these findings permit a straightforward interpretation of the remaining results.

Confidence Reports. Figure 2 also demonstrates the effect of confidence measures. Like ID accuracy, we found that subjective confidence varied along the VOT continuum, $F(1,14)=6.55, M S E=44.11, p=.008, \eta^{2}=.319$. Relative to Experiment 1, we did observe greater underconfidence in the negative portion of the VOT continuum.

Typicality Task. Typicality ratings also varied significantly as a function of stimulus location along the VOT continuum, $F(14,112)=5.820, M S E=.3 .295, p=.009, \eta^{2}=$ .421. Unlike accuracy, but like mean confidence, typicality ratings appeared to be more responsive to the acoustic properties of the stimuli. Participants considered stimuli in the $/ \mathrm{b} /$ and $/ \mathrm{p}^{\mathrm{h}} /$ range as less typical then stimuli in the $/ \mathrm{p} /$ range even though they exhibited equal accuracy. Moreover, within-category ratings exhibited more graded responses.

Figure 2. ID, Subjective Ratings, and Interpolated Function


Interpolated Function. The similarities in patterns observed in confidence and typicality suggested a potential relationship between these two functions. As Figure 2 indicates, mean confidence ratings are situated between accuracy in the identification task and typicality ratings in the typicality task. Pearson's correlations revealed the strongest relationship between confidence and typicality ratings, $r^{2}=.960, p<.001$. The correlation between identification responses and mean confidence was also significant, $r^{2}=.446, p=.007$, although the correlation between identification and typicality was only marginally significant, $r^{2}=.261, p=.051$. These findings suggest that confidence is associated with both identification accuracy and typicality ratings but that identification accuracy and typicality ratings are only weakly related.

In order to examine the relationship between accuracy, typicality, and confidence ratings we converted typicality to a proportion, summed it with proportion correct, and produced an interpolated function. A paired-samples t-test revealed that the mean confidence function and the interpolated function did not significantly differ from one another, $t(14)=.309, p=.762$. This suggests that confidence reports were closely associated with information from both identification accuracy (associated with phonemic representations) and typicality ratings (associated with acoustic information). All other paired-sample $t$-tests were significant (all $t \mathrm{~s}>3.283, p \mathrm{~s}<.005$ ) indicating that different sources of information contributed to response selection for each dependent measure.

## Discussion

The typicality results provide strong support for the accessibility of acoustic information during speech perception. When normalized along a common scale, typicality ratings were lower in the $/ \mathrm{p}^{\mathrm{h}}-\mathrm{b} /$ region of the VOT continuum relative to identification responses, suggesting that participants had an awareness of acoustic properties. This finding replicates previous studies (e.g., Miller \&

Volatis, 1989). A similar pattern was again observed in confidence ratings. More importantly, when typicality and identification functions were interpolated, a near perfect fit was obtained with the confidence function. This finding is of particular interest as it suggests that confidence is affected by both phonemic representations stored in longterm memory and acoustic properties available in short-term memory.

## Conclusion

The findings of the present study have several implications. First, we replicated results from early studies (e.g., Pisoni et al., 1982) showing that participants can learn a non-native speech category using only three exemplars selected from the VOT continuum. Second, our results indicate that the acquisition of non-native phonemes likely resulted from the allocation of attention to regions along the VOT continuum that were previously unattended (e.g., Jusczyk, 1992). In comparison to tasks where participants must identify two categories, such as in Experiment 2 and Schoenherr et al. (2012), participants exhibited well-defined $/ \mathrm{b} /$ and $/ \mathrm{p} /$ phonemic categories. The induction of a novel category boundary requires attention to acoustic properties. In the present study, support for the detection of acoustic properties of stimuli from the same phonemic category comes from both confidence reports and typicality ratings. Both subjective confidence and typicality decreased in the $/ \mathrm{p}^{\mathrm{h}}-\mathrm{b} /$ region of the VOT continuum relative to ratings within the $/ \mathrm{b}-\mathrm{p} /$ region.

As briefly noted above in reference to Experiment 1, this pattern appears to hold provided that participants do not need to concurrently monitor their performance (i.e., provide confidence reports). Thus, although feedback directs attention toward previously unattended regions of the VOT continuum, additional top-down monitoring might emphasize differences between exemplars to the detriment of the formation of novel phonemic categories. Similar findings of top-down interference have been observed in visual search tasks wherein activation of object-level representations in long-term memory creates interference in detection of stimuli based on visual features (e.g., Zhaoping \& Firth, 2011).

Finally, our study also has implications for models of confidence. The identification function and typicality ratings are believed to reflect the detection of phonemic and acoustic information, respectively. When interpolating between these two functions, we obtained a function that approximated that obtained from confidence reports. This suggests that additional processing is required to create a confidence report and that such a process can integrate information from short- and long-term memory.

## References

Abramson, A., \& Lisker, L. (1965). Voice onset time in stop consonants: Acoustic analysis and synthesis. Proceedings
of the Fifth International Congress on Acoustics, Liege, A51.
Baranski, J. V., \& Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgements. Perception \& Psychophysics, 55, 412-428.
Baranski, J. V., \& Petrusic, W. M. (1998). Probing the locus of confidence judgments: Experiments on the time to determine confidence. Journal of Experimental Psychology: Human Perception and Performance, 24, 929-945.
Best, C. T. (1995). A direct realist view of cross-language speech perception: New Directions in Research and Theory. In W. Strange (Ed.) Speech perception and linguistic experience: Theoretical and methodological issues (pp. 171-204). Baltimore: York Press.
Dawes, R. M (1980). Confidence in intellectual vs. confidence in perceptual judgments. In E. D. Lantermann \& H. Feger (Eds.) Similarity and Choice: Papers in Honor of Clyde Coombs (pp. 327-345). Bern Hans Huber.
Eimas, P. D. (1975). Auditory and phonetic coding of the cues for speech: Discrimination of the [r-1] distinction by young infants. Perception \& Psychophysics, 18, 341-347.
Flege, J., (1995). Second language speech learning: Theory, findings and problems. In W. Strange (Ed.). Speech Perception and Linguistic Experience: Theoretical and Methodological Issues (pp. 233-277). Baltimore: York Press.
Gigerenzer, G., Hoffrage, U., \& Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. Psychological Review, 98, 506-528.
Goldstone, R. (1994). Influences of categorization on perceptual discrimination. Journal of Experimental Psychology: General, 123, 178-200.
Haskins Laboratories (2011). Abramson/Lisker VOT Stimuli. Retrieved 01/12/2011. From http://www.haskins.yale.edu/featured/demoliskabram/index.html/.
Hazan, V. \& Barrett, S. (2000). The development of phonemic categorization in children aged 6-12. Journal of Phonetics, 28, 377-396.
Jusczyk, P. (1989). Developing phonological categories from the speech signal. In C. A. Ferguson, L. Menn, \& C. Stoel-Gammon (Eds.), Phonological Development: Models, Research, Implications. Monkton, MD: York.
Jusczyk, P.W. (2000). The Discovery of Spoken Language. Cambridge: MIT Press.
Kvidera, S., \& Koustaal, W. (2008). Confidence and decision type under matched stimulus conditions: overconfidence in perceptual but not conceptual Decisions. Journal of Behavioral Decision Making, 21, 253-281.
Liberman, A.M., Cooper, F.S., Shankweiler, D.P., \& Studdert-Kennedy, M. (1967). Perception of the speech code. Psychological Review, 74, 431-461.

Liberman, A. M., Harris, K. S., Hoffman, H. S., \& Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. Journal of Experimental Psychology, 54, 358-368.
Lichtenstein, S., \& Fischhoff, B. (1977). Do those who know more also know more about how, much they know? Organizational Behavior and Human Performance, 20, 159-183.
Lisker, L, \& Abramson, A. S. (1967). The voicing dimension: Some experiments in comparative phonetics. Proceedings of the 6th International Congress of Phonetic Sciences. Prague: Academia.
Mackain, K. S., Best, C. T., \& Strange, W. (1981). Categorical perception of English /r/ and /l/ by Japanese bilinguals. Applied Psycholinguistics, 2, 369-390.
McMurray, B., Tanenhaus, M., \& Aslin, R. (2002). Gradient effects of within-category phonetic variation on lexical access. Cognition, 86, B33-B42.
Miller, J. L., \& Volaitis, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. Perception \& Psychophysics, 46, 505-512.
Nosofsky. R. (1986). Attention, similarity, and the identification-categorization relationship. Journal of Experimental Psychology: General, 115, 39-57.
Pisoni, D. B., Aslin, R. N., Percy, A. J., \& Hennessy, B. L. (1982). Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants. Journal of Experimental Psychology: Human Perception and Performance, 8, 297-314.
Pisoni, D. B., \& Tash, J. B. (1974) Reaction times to comparisons within and across phonetic categories. Perception \& Psychophysics, 15, 285-290.
Schoenherr, J. R., Leth-Steensen, C., \& Petrusic, W. M. (2010). Selective attention and subjective confidence calibration. Attention, Perception \& Psychophysics, 72, 353-368.
Schoenherr, J. R. \& Logan, J. (in preparation). Attending Unattended Regions of an Acoustic Continuum: Evidence for Acoustic and Phonemic Representations in Speech Perception.
Schoenherr, J. R., Logan, J. \& Winchester, A. (2012). Subjective confidence of acoustic and phonemic representations during speech perception. Proceedings of the $34^{\text {th }}$ Annual Meeting of the Cognitive Science Society, Sapporo, Japan.
Vickers, D., \& Packer, J. S. (1982). Effects of alternating set for speed or accuracy on response time, accuracy, and confidence in a unidimensional discrimination task. Acta Psychologica, 50, 179-197.
Zhaoping, L. \& Firth, U. (2011). A clash of bottom-up and top-down processing visual search: The reversed letter effect revisited. Journal of Experimental Psychology: Human Perception \& Performance, 37, 997-1006.

# Virtues and Vices in Monetary Tradeoffs 

Marc Scholten (scholten@ispa.pt)<br>ISPA University Institute<br>Lisbon, Portugal<br>Daniel Read (daniel.read@wbs.ac.uk)<br>Warwick Business School<br>Coventry, United Kingdom


#### Abstract

Monetary intertemporal tradeoffs are a restricted, yet underexplored, domain. In this extended abstract, we provide an integrative analysis of monetary tradeoffs involving single dated outcomes, unmixed sequences, virtues (schedules of investment), and vices (schedules of debt). Results include debt aversion, aversion to vices (which adds to debt aversion) and relative vices, and attraction to virtues and relative virtues. The results motivate a comparative mental accounting model, which includes direct comparisons between the outcomes delivered by the options at consecutive delays. The model accommodates not only the results reported in this extended abstract, but also other puzzling phenomena in choices involving sequences.


Keywords: Intertemporal choice, discounting; virtues; vices; sequences; mental accounting.

Intertemporal choices are those in which outcomes of choice are traded off against their timing. One example is the choice between a chocolate mousse and a fruit salad for dessert, where immediate gratification may favor the former, but future health may favor the latter. Another example is the decision of whether to consume on credit now and pay off debt in the future or invest income now and consume more in the future. Again, there is a tradeoff between what is best now and what is best in the future.

The above choices can be viewed as choices between a relative vice and a relative virtue (Wertenbroch, 1998), where the relative vice is better in the short run but worse in the long run or overall, whereas the relative virtue is worse in the short run but better in the long run or overall. Many intertemporal choices fit this definition, and also elementary choices between single dated outcomes. Consider the choice between $\$ 150$ today and $\$ 200$ in 1 year. By the above definition, the smaller-sooner outcome is a relative vice, and the larger-later outcome is a relative virtue. The notion of relative virtues and vices is inherently comparative in nature. However, current models of intertemporal choice do not consider the possibility that people actually make the relevant comparisons, and frame the options as relative virtues and vices. For instance, the above choice between single dated outcomes may be represented as a decision of whether to
accept or reject receiving $\$ 150$ less today and $\$ 200$ more in 1 year. This is a relative virtue: Less money in the short term (-\$150), but more in the long term (\$200) or overall (\$50). Alternatively, the choice may be represented as a decision of whether to accept or reject receiving \$150 more today and $\$ 200$ less in 1 year. This is a relative vice: More money in the short term (\$150), but less in the long term (-\$200) or overall (-\$50). These mental operations involve direct comparisons between the options: Comparisons between the outcomes available today (\$150 and $\$ 0$ ) and in 1 year ( $\$ 0$ and $\$ 200$ ). The question is whether and when people perform these operations, and how it affects the decisions they make.

While most experimental research of intertemporal choice has focused on single dated outcomes, many reallife choices involve prospects of multiple outcomes. A distinction can be made between mixed and unmixed sequences. Unmixed sequences can be goods, which are composed of only positive outcomes, and bads, which are composed of only negative ones. Mixed sequences include, but are not restricted to, absolute virtues, or virtues in short, which exchange sooner costs for largerlater benefits, and absolute vices, or vices in sort, which exchange sooner benefits for larger-later costs.

With the exception of Prelec and Loewenstein's (1998) work on the mental accounting of investment and debt, experimental research on virtues and vices has focused almost exclusively on consumption, such as consumption of healthy or unhealthy food items (Read \& van Leeuwen, 1998, and thereafter), as in the dessert example given above, and consumption of highbrow or lowbrow movies (Read, Loewenstein, \& Kalyanaraman, 1999, and thereafter). In that research, there is no rigorous control over whether the options are treated as absolute or relative virtues and vices. Monetary tradeoffs, often the focus of experimental research on single dated outcomes and unmixed sequences, lend themselves perfectly for that purpose. One goal of this extended abstract is, therefore, to conduct an integrative analysis of choices involving single dated outcomes, unmixed sequences, virtues, and vices in monetary tradeoffs, where monetary virtues are schedules of investment, and monetary vices are schedules of debt.

As a whole, our results cannot be accommodated by
any current model of intertemporal choice, so that a new approach is needed. We propose an extension of Prelec and Loewenstein's (1998) double-entry mental accounting model of preferences for schedules of investment and debt. Essentially, the extension is that, in choices involving sequences, people do make direct comparisons between the outcomes available at consecutive delays, which often means that they reframe the options as relative virtues and vices. We discuss how this comparative mental accounting model accommodates the results reported in this extended abstract, and also results reported elsewhere.

We collected data from many samples in three nations (the United States, the United Kingdom, and Portugal), sometimes with paid, sometimes with unpaid participants, as we went along perfecting the experimental comparisons in order to counter, as much as possible, explanations offered by the current models of intertemporal choice. The comparisons that we report are the most challenging ones. This extended abstract, however, can only cover a few. All choices reported were part of surveys including a larger set of intertemporal choices, the order of which was randomized across participants.

## Debt aversion

A basic assumption underlying models of intertemporal choice is positive time preference: People would prefer a gain sooner rather than later (impatience), and a loss later rather than sooner (procrastination). To test this assumption, we presented 36 participants with different timings of receiving $€ 100$ and 78 participants with different timings of paying $€ 100$. The results were as follows:

## Set 1

W Receive $€ 100$ in 1 year (11\%)
B Receive $€ 100$ today (89\%)

## Set 2

W Pay €100 today (65\%)
B Pay $€ 100$ in 1 year (35\%)
Here and elsewhere, B denotes the best option in the long run, whereas W denotes worst option in the long run. An overwhelming chose B among receipts (positive time preference), $\chi^{2}(1)=21.78, p<.005$ (Pearson's $\chi^{2}$ ), but a smaller yet significant majority chose W among payments, $\chi^{2}(1)=7.38, p<.05$ (negative time preference).

The observed pattern of results can be explained by combining the discounting of delayed outcomes with an aversion to delayed losses, or debt aversion. Discount-ing favors immediate gains over delayed ones, and delayed losses over immediate ones. Debt aversion, however, favors immediate losses over delayed ones, thus countervailing discounting. Therefore, while choice is not conflicted for different timings of a gain, because
discounting unambiguous-ly favors the immediate gain, it is conflicted for different timings of a loss, because discounting, which favors the delayed loss, is countervailed by debt aversion, which favors the immediate loss. In this study, discounting was outweighed by aversion to delayed losses.

Debt aversion operates in addition to loss aversion, which is that the pain of loss is greater than the pleasure of an equal gain (Kahneman \& Tversky, 1979). We next report an aversion to vices, which operates in addition to debt aversion.

## Aversion to Vices

We asked 429 participants to choose from the following pairs of options:

## Set 3

Referent pair
W Pay \$600 in 1 year (26\%)
B Pay $\$ 450$ today (74\%)
Target pair
W Receive $\$ 50$ today and pay $\$ 600$ in 1 year (20\%)
B Pay $\$ 450$ today (80\%)
W in the target pair is obtained from W in the referent pair by adding an immediate $\$ 50$. Because W in the target pair dominates W in the referent pair, it should be more popular. However, the opposite was true, $\chi^{2}(1)=7.72, p<$ .05 (McNemar's $\chi^{2}$ for dependent samples), suggesting that a later payment, or a debt, hurts more when it is the cost of a sooner benefit than when it is an uncompensated loss. This is aversion to vices.

One possible explanation is offered by Loewenstein and Prelec's (1993) sequences model, according to which people have a preference for improvement tempered by a preference for spreading. A vice, however, exhibits deterioration, which decreases preference for it. Another possible explanation is offered by Prelec and Loewenstein's (1998) double-entry mental accounting model: The pleasure of the immediate benefit is attenuated by the pain of the delayed cost (debt), and the experience of the immediate benefit may, through attenuation, change into a negative one.

## Attraction to Virtues

Two principles of outcome valuation are loss aversion and diminishing sensitivity (Kahneman \& Tversky, 1979): The impact of a loss is greater than that of an equivalent gain, and the marginal impact of an outcome decreases with its magnitude. In the following set, we see both principles being violated. We asked 435 participants to choose from the following option pairs:

## Set 4

Referent pair
W Receive $€ 450$ today (54\%)
B Receive $€ 600$ in 1 year (46\%)

Target pair
W Receive €300 today (46\%)
B Pay $€ 150$ today and receive $€ 600$ in 1 year (54\%)
The target pair is obtained from the referent pair by subtracting a common amount (\$150) from both options in period 1. This does not change the interest rate implied by the options (33\%), so that, objectively, the preference between W and B should not change. Moreover, by loss aversion and diminishing sensitivity, the value difference between 300 and -150 in the target pair is more strongly in favor of W than the value difference between 450 and 0 in the referent pair, so that W should be more popular, and B less popular, in the target pair than in the referent pair. Instead, B was more popular in the target pair than in the referent pair, $\chi^{2}(1)=4.90, p<.05$, suggesting that the same receipt is more appealing when it is the benefit of an investment than when it is an uncompensating gain. This is attraction to virtues.

One possible explanation is offered by the sequences model: Preference for improvement. Another possible explanation is offered by the mental accounting model: The pain of the immediate cost (investment) is buffered by the pleasure of the delayed benefit, and the experience of the immediate cost may, through buffering, change into a positive one.

## Unmixed Sequences

We asked the same 435 participants from the section on attraction to virtues to choose from the following option pairs:

## Set 5

Referent pair
W Receive $€ 75$ today (68\%)
B Receive €100 in 1 year (32\%)
Target pair
W Receive €300 today (57\%)
B Receive $€ 225$ today and receive $€ 100$ in 1 year (43\%)

The target pair is obtained from the referent pair by adding a common amount ( $€ 225$ ) to both options in period 1. This does not change the interest rate implied by the options (33\%), so that, objectively, the preference between W and B should not change. However, B was more popular in the target pair than in the referent pair, $\chi^{2}(1)=9.33, p<.005$.

One possible explanation for the above result is diminishing sensitivity: The value difference between 300 and 225 in the target pair is less strongly in favor of W than the value difference between 75 and 0 in the referent pair, so that W should be less popular, and B more popular, in the target pair than in the referent pair. However, diminishing sensitivity is being violated by the results below:

## Set 6

Referent pair
W Receive $€ 300$ today (58\%)
B Receive €400 in 1 year (42\%)
Target pair
W Receive $€ 300$ today and receive $€ 300$ in 1 year (47\%)
B Receive $€ 700$ in 1 year (53\%)
The target pair is obtained from the referent pair by adding a common amount ( $€ 300$ ) to both options in period 2. This does not change the interest rate implied by the options (33\%), so that, objectively, the preference between W and B should not change. However, B was more popular in the target pair than in the referent pair, $\chi^{2}(1)=21.59, p<.005$. By diminishing sensitivity, the value difference between 700 and 300 in the target pair is less strongly in favor of $B$ than the value difference between 400 and 0 in the referent pair, so that $B$ should be less popular in the target pair than in the referent pair.

The above results are incompatible with the sequences model: In Set 5, B deteriorates and yet it gained popularity, and, in Set 6, W neither deteriorates nor improves, and yet it lost popularity. The results cannot be explained by the mental accounting model either, because, in the absence of mixed sequences, i.e., schedules of costs and benefits, this model reduces to a standard delay discounting model.

The results are consistent with the notion that choice involving sequences promotes comparative accounting. In the choice between a single immediate outcome and a sequence (Set 5), the sequence is framed as a relative virtue (' $€ 75$ less today and $€ 100$ in 1 year’), and attraction to virtues increases the preference for this option. In the choice between a single delayed outcome and a sequence (Set 6), the sequence is framed as a relative vice (' $€ 150$ today and $€ 200$ less in 1 year'), and aversion to vices decreases the preference for this option.

## A Core Anomaly

We asked the same 429 participants from the section on aversion to vices to choose from the following option pair:

## Set 7

Receive \$500 in 1 year and receive \$500 in 3 years (29\%) Receive $\$ 1,000$ in 2 years (71\%)

A large and significant majority preferred the single delayed receipt to the flat sequence of delayed receipts, $\chi^{2}(1)=74.69, p<.005$. We call this a core anomaly, because no model of intertemporal choice accounts for it. As to standard delay discounting models, such as Loewenstein and Prelec's (1992) hyperbolic discounting model, discounting per se contributes to a preference for the sequence, which is compounded by hyperbolic discounting and diminishing sensitivity. As to the
sequences model, the sequence neither improves nor deteriorates, which contributes to indifference between the sequence and the single delayed receipt. Finally, the mental accounting model reduces to a standard delay discounting model, because the choice does not involve mixed sequences. In the next section, we try to account for the whole set of results.

## Theory

Our theory is an extension of Prelec and Loewenstein's (1998) mental accounting model of investment and debt. In this model, sooner benefits are attenuated by later costs, and sooner costs are buffered by later benefits. This, by itself, accounts for aversion to vices and attraction to virtues, as observed in Sets 3 and 4. The model incorporates loss aversion, in that negative experiences are augmented relative to positive ones. Negative experiences include sooner benefits when their attenuation by later costs results in a sign reversal, and positive experiences include sooner costs when their buffering by later benefits results in a sign reversal. Experiences in each period are discounted as a function of the delay to the experiences, and the option with the highest discounted value is chosen.

Our extension of the mental accounting model draws on two considerations. First, operating in addition to loss aversion is debt aversion, meaning that delayed costs are augmented relative to immediate ones. This accommodates the preference observed in Set 2. It also increases the aversion to vices observed in Set 3.

Second, the option that has the longest interval between its soonest and latest outcome, i.e., the longest duration, becomes the target option, the outcome of which in any given period is compared with the outcome of the referent option in that period. Thus, for instance, in the choice between a sequence and a single dated outcome, the sequence becomes the target option, and the single dated outcome becomes the referent option. In the choice between two single dated outcomes, neither option has duration, and so there is no targeting and referencing. In the choice between options of equal duration, either option can become the target option.

From the vantage point of the extended mental accounting model, the preference pattern observed in Set 5 shows attraction to relative virtues. In the target pair, the sequence is the target option and the single immediate receipt is the referent option. Thus, the choice is framed as whether to accept or reject the prospect of 'receiving $€ 75$ less today and receiving $€ 100$ in 1 year.' To the degree that the immediate comparative loss is buffered by the delayed receipt, possibly resulting in a positive experience of the immediate comparative loss, the tendency will be to accept this prospect.

The preference pattern observed in Set 6 shows aversion to relative vices. In the target pair, the sequence is the target option and the single delayed receipt is the referent option. Thus, the choice is framed as whether to
accept or reject the prospect of 'receiving $€ 300$ today and receiving $€ 400$ less in 1 year.' To the degree that the immediate receipt is attenuated by the delayed comparative loss, the tendency will be to reject this prospect.
Finally, our explanation of the preference observed in Set 7 is that, the sequence was framed as two gains interleaved with a comparative loss, and that, due to attenuation of the gain in period 1 and aversion the comparative loss in period 2, the tendency was to reject the mixed prospect, notwithstanding a buffering of the comparative loss in period 2 by the gain in period 3.

## Some Implications

The comparative mental accounting model resolves several puzzles. Consider, for instance, the widely investigated preference for improving sequences over deteriorating ones. Loewenstein and Prelec (1993) discuss a number of explanations of this phenomenon, which all invoke within-option operations. One explanation is adaptation and loss aversion. People adapt to ongoing stimuli over time, and evaluate ensuing stimuli relative to their adaptation level. An improving sequence becomes a series of positive departures (gains) from the adaptation level, while a deteriorating sequence become a series of negative departures (losses) from the adaptation level. Preference for improving sequences over deteriorating ones then follows from loss aversion (Kahneman \& Tversky, 1979).

Our explanation, in contrast, invokes between-option operations. When people focus on the improving sequence and compare it with the deteriorating one, they experience an increasing series of comparative losses and gains. Attraction to relative virtues increases the attractiveness of this option, making choice of improvement more likely. Alternatively, when people focus on the deteriorating sequence and compare it with the improving one, they experience a decreasing series of comparative gains and losses. Aversion to relative vices decreases the attractiveness of this option, making choice of deterioration less likely. According to our explanation, preference for improvement over deterioration is fundamentally a choice-related phenomenon, because, without direct comparisons between options, there would be no mental construction of relative virtues and vices. Indeed, it has been shown that preference for improvement over deterioration evaporates in elicitation tasks other than choice, in which other motives and mental operations come to the fore (Frederick \& Loewenstein, 2008).

Another puzzle is the hidden-zero effect (Magen, Dweck, \& Gross, 2008), which is that the preference for B over W increases when two single dated receipts are changed into sequences by explicating the zero receipt. Thus, for instance, choice of ' $\$ 0$ today and $\$ 400$ in 1 year' over ' $\$ 300$ today and $\$ 0$ in 1 year' is more likely than choice of ' $\$ 400$ in 1 year' and ' $\$ 300$ today.' The comparative mental accounting model explains the
hidden-zero effect as follows. Both sequences, each with duration of 1 year, can become the target option. When B is the referent option, W becomes a relative vice, and, by aversion to relative vices, the preference for B over W increases. When W is the referent option, B becomes a relative virtue, and, by attraction to relative virtues, the preference for B over W increases.

Yet another puzzle is the mere token effect (Urminsky \& Kivetz, 2011), which is a violation of independence in which the preference for B over W increases when two single dated receipts are changed into sequences by adding a common consequence before both receipts. For instance, choice of ' $€ 50$ tomorrow and $€ 400$ in 1 year' over ' $€ 50$ tomorrow and $€ 200$ in 1 week’ is more likely than choice of ' $€ 400$ in 1 year' over $€ 200$ in 1 week.' The comparative mental accounting model can explain the mere token effect as well. With the introduction of the token, the sequence of longer duration, B , becomes the target option, whereas the sequence of shorter duration, W , becomes the referent option. As a result of the comparison process, the choice between $B$ and $W$ is framed as a decision of whether to accept the relative virtue ' $€ 200$ less in 1 week and $€ 400$ in 1 year.' By attraction to relative virtues, the tendency will be to accept this prospect.

In our article, we provide a much more exhaustive analysis of recently discovered anomalies in choices involving sequences. The comparative mental accounting accommodates most.

## Conclusion

Our results show an interesting pattern: People are extremely impatient in gains, with many declining to receive a $33 \%$ interest rate, much and much higher than riskless market rates, but they become more farsighted when faced with other intertemporal arrangements. First, their impatience in gains decreases when future benefits are preceded by immediate costs (attraction to virtues). Furthermore, they are averse to procrastination in losses (debt aversion), and become even more farsighted when future costs are preceded by immediate benefits (aversion to vices).

Our theoretical reconstruction suggests that, people make direct comparisons between options. Specifically, the outcomes of the option with the longest duration are compared, period by period, with the outcomes of the options with the shortest duration. The result is that even sequences are cognitively represented as relative virtues, relative vices, or, more generally, mixed prospects. This proposal of comparative framing greatly increases the scope of a mental accounting approach to intertemporal choice.

It also opens avenues toward a better understanding of real-life decisions. The paradigmatic example of intertemporal choice is whether to get a job and earn a living now or go to college and earn a better living later. How is such a complex decision made? Plausibly, people
would make direct comparisons between the features of the options under consideration. In this case, comparisons between studying (or partying) and working, between the jobs available with and without a college degree, between prospective earnings, and between incurring and foregoing tuition debt. Current models of intertemporal choice are notably ignorant of such comparisons in decision making: Each option receives its discounted value, regardless of how it compares to other options, and the option with the highest value is chosen. Our analysis suggests that intertemporal choice is comparative in a carefully crafted choice environment, and we would be surprised if people suddenly ceased to make comparisons in the wild.

## References

Frederick, S., \& Loewenstein, G. (2008). Conflicting motives in evaluations of sequences. Journal of Risk and Uncertainty, 37, 221-235.
Kahneman, D., \& Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, 47, 363-391.
Loewenstein, G., \& Prelec, D. (1992). Anomalies in intertemporal choice: Evidence and an interpretation. Quarterly Journal of Economics, 107, 573-597.
Loewenstein, G., \& Prelec, D. (1993). Preferences for sequences of outcomes. Psychological Review, 100, 91108.

Loewenstein, G., \& Prelec, D. (1993). Preferences for sequences of outcomes. Psychological Review, 100, 91108.

Magen, E., Dweck, C. S., \& Gross, J. J. (2008). The hidden-zero effect. Representing a single choice as an extended sequence reduces impulsive choice. Psychological Science, 19, 648-649.
Prelec, D., \& Loewenstein, G. (1998). The red and the black: Mental accounting of savings and debt. Marketing Science, 17, 4-28.
Read, D., Loewenstein, G. \& Kalyanaraman, S. (1999). Mixing virtue and vice: Combining the immediacy effect and the desire for diversity. Journal of Behavioral Decision Making, 12, 257-273.
Read, D., \& van Leeuwen, B. (1998). Predicting hunger: The effects of appetite and delay on choice. Organizational Behavior and Human Decision Processes, 76, 189-205.
Urminsky, O., \& Kivetz, R. (2011). Scope insensitivity and the "mere token" effect. Journal of Marketing Research, 48, 282-295.
Wertenbroch, K. (1998). Consumption self-control by rationing purchase quantities of virtue and vice. Marketing Science, 17, 317-337.

# Comparing Model Comparison Methods 

Holger Schultheis (schulth@informatik.uni-bremen.de)<br>Cognitive Systems, University of Bremen, Enrique-Schmidt-Str. 5, 28359 Bremen, Germany

Ankit Singhaniya<br>Computer Science and Engineering, NIT Nagpur, Nagpur 440010, India

## Devendra Singh Chaplot

Computer Science and Engineering, IIT Bombay, Mumbai 400076, India


#### Abstract

Comparison of the ability of different computational cognitive models to simulate empirical data should ideally take into account the complexity of the compared models. Although several comparison methods are available that are meant to achieve this, little information on the differential strengths and weaknesses of these methods is available. In this contribution we present the results of a systematic comparison of 5 model comparison methods. Employing model recovery simulations, the methods are examined with respect to their ability to identify the model that actually generated the data across 3 pairs of models and a number of comparison situations. The simulations reveal several interesting aspects of the considered methods such as, for instance, the fact that in certain situations methods perform worse than model comparison neglecting model complexity. Based on the identified method characteristics, we derive a preliminary recommendation on when to use which of the 5 methods. Keywords: computational cognitive models, model comparison, model mimicry, model generalization


When computationally modeling cognition, often several different models are available or conceivable as explanations for the cognitive ability in question. In such a situation, the aim is to select the best of these candidate models according to a set of criteria. Among others (e.g., falsifiability or interpretability) the extent to which the different models are able to simulate observed human behavior is usually considered a key criterion for selecting from the candidate models.

A naïve approach to gauge the models' ability to simulate the existing observations is to fit each model to the available data and choose the model that provides the tightest fit as indicated, for instance, by the models' Root Mean Squared Error (RMSE). Such an approach is problematic, because it does not take into account the complexity of the compared models. As a result, there is a tendency for overfitting and for selecting more complex models even if simpler models provide the better explanation of the considered cognitive ability (Pitt \& Myung, 2002).

Several methods taking into account model complexity have been proposed to avoid the pitfalls of the naïve approach (see Shiffrin, Lee, Kim, \& Wagenmakers, 2008, for an overview). However, common use of such more sophisticated model comparison methods is partly hampered by the fact that many properties of the different methods are insufficiently investigated. Only very few studies (e.g., Cohen, Sanborn, \& Shiffrin, 2008) have systematically examined different comparison methods with respect to their differential advantages and disadvantages. Consequently, when
faced with a situation that requires comparing models regarding their ability for simulating human behavior, modelers are often faced with the problem that it is unclear which model comparison methods could reasonably and should ideally be employed in a given situation.

In this contribution we present the results of a systematic comparison of 5 model comparison methods. The methods are examined with respect to their ability to select the model that actually generated the data across 3 pairs of models and a number of contextual variations (e.g., tightness of fits, amount of noise in the data). The obtained results highlight important properties of the different comparison methods. Together with the fact that all 5 considered methods are general in the sense that they place no restrictions on the type of models that can be compared, these results are, we believe, conducive to increasing the frequency with which more sophisticated comparison methods instead of the naïve approach will be employed for model evaluation and comparison.

The remainder of this article is structured as follows. First, we list and briefly describe all considered methods. Second, the employed models, contextual variations, and procedural details of the method comparison are described. Subsequently, comparison results are presented and discussed before we conclude our considerations and highlight topics for future work.

## Methods

The 5 methods we compared are the bootstrap, the bootstrap with standard error (SE) and confidence interval (CI), the data-uninformed parametric bootstrap cross-fitting method, henceforth called cross-fitting method (CM), the simple holdout, and the prediction error difference method (PED). Each of these was applied to 3 pairs of models and will be described in turn below.

## Bootstrap

Given a set of $n$ observations, the bootstrap method of model comparison proceeds as follows (see Efron \& Tibshirani, 1993, for an overview of bootstrapping procedures). First, an arbitrary but fixed number $B$ of bootstrap samples is generated. A bootstrap sample is a set of $n$ data points randomly drawn with replacement from the $n$ original observations. Due to sampling with replacement, most bootstrap samples will contain only a subset of all original observa-
tion (but some of these more than once). Second, each of the to-be-compared models is fitted to each bootstrap sample. Third, for each bootstrap sample, the fitted models are used to predict those data points that were not in the bootstrap sample and the deviation of the predictions from the original data points is measured (e.g., by the mean squared error). Fourth, the measures of deviation are combined for each model across all bootstrap samples to obtain an overall measure for the prediction error ( $\overline{E r r}$ ) of each model. The model that has the lowest $\overline{E r r}$ is assumed to be the best approximation to the process that actually generated the $n$ original data points.

Due to the randomness in generating the bootstrap samples as well as the noise that is likely included in the original observations, $\overline{E r r}$ only constitutes an estimate of the models' true prediction error. Accordingly, the model showing the lowest $\overline{E r r}$ may do so because of chance and not because it is the best model. Knowing the variability, that is, the SE, of the error estimates can potentially help alleviating this problem. Given the standard error, CIs on the true prediction error can be derived. If the CIs of the models' error estimates do not overlap, one may conclude with more confidencedepending on the confidence level employed to construct the intervals-that the model with the lower $\overline{E r r}$ in fact provides the better approximation to the process that generated the $n$ original data points.

In our simulations we assess both the bootstrap considering the SE and the bootstrap not considering the SE for deciding which of the two models is more appropriate. We construct the CIs by (a) computing the SE as proposed in Efron and Tibshirani (1997), (b) employing a confidence level of $99 \%$, and (c) assuming that the prediction error estimates are distributed approximately normal. The runtime complexity of both bootstrap variants is $O(B *$ fitCost $)$, where $B$ is the number of bootstrap samples and fitCost is the time complexity of estimating model parameters.

## CM

The CM was proposed by Wagenmakers, Ratcliff, Gomez, and Iverson (2004) as a way to assess to what extent two models are able to mimic each other's behavior. Since model complexity and the ability to mimic other models are often related, the obtained mimicry information potentially allows reducing the bias towards selecting more complex models.

The following steps are involved in the CM: First, for one of the models (say, model 1) a certain number, $N D S$, of sets of parameter values are randomly drawn from the feasible range of the model's free parameters. Second, model 1 is used to generate $N D S$ data sets employing each of the $N D S$ parameter value sets, respectively. Third, both models are fitted to each of the NDS data sets yielding NDS measures of goodness of fit (GOF, e.g., the mean squared error) for both models. Fourth, the pairwise GOF differences are computed for all datasets. By repeating these four steps for the second model (model 2), one obtains two distributions of GOF differences, one for data generated from model 1 and one for data generated from model 2.

Given a set of observations, these two distributions can be utilized to decide which of the two models provides the better account of the observations. Both models are fitted to the observations and the difference in the models' GOFs are computed. If the resulting difference is classified to more likely come from the distribution resulting from data generated from model 1 , model 1 is assumed to be more appropriate; otherwise the model 2 is assumed to be more appropriate.

Based on the results reported in Schultheis and Singhaniya (accepted), we employed a variant of the k-Nearest Neighbor algorithm $(k=10)$ for classification. The runtime complexity of the CM is $O(N D S *$ fitCost $)$.

## Simple Hold-Out

This method gauges the to-be-compared models by repeatedly splitting the set of available $n$ observations into a training and test set. For each of these splits, both models are fitted to the respective training set. The fitted models are then used to generate predictions for the data points in the test set and the corresponding prediction error is determined. Accordingly, using $I$ different splits results in $I$ prediction error values for each of the two models. The model that has the lower median prediction error is selected as the more appropriate model. The runtime complexity of the simple hold-out method is $O(I *$ fitCost $)$.

## PED

Similar to the simple hold-out the PED (van de Wiel, Berkhof, \& van Wieringen, 2009) employs $I$ splits of the original data set into training and test set to compare models. For both models the prediction error is computed for each point in the test set after fitting the models to the corresponding training set. Subsequently, pairwise differences between prediction errors for model 1 and model 2 are calculated. These signed error differences are subjected to signed rank tests to derive the probability of the observed distribution of signed ranks under the null hypothesis that the models do not differ in predictive accuracy.

Thus, the PED yields $I$ probability values. If the median of these values is below or equal to a pre-specified significance level $\alpha$, the models are assumed to be significantly different in their predictive accuracy and the model with the smaller prediction error is assumed to be the more appropriate model. In our simulations we used the Wilcoxon signed rank test with $\alpha=0.05$. The runtime complexity of the PED is $O(I *$ fitCost $)$.

## Method Properties

The procedural details of the methods described above imply a number of (differences in) crucial properties of the methods regarding model comparison.

First, the methods apply different criteria for judging the suitability of the compared models for a given data set. Both bootstrap variants, the PED, and the simple hold-out judge the models based on their ability to generalize to new data points, that is, these methods attempt to optimize what has
been called the generalization criterion (Busemeyer \& Wang, 2000). In contrast, the CM has been argued to be optimal "under the validation criterion of selecting the generating model" (Cohen et al., 2008, p. 698). Since our simulations check the methods ability to recover the generating model, they test the conjecture of (Cohen et al., 2008) or, more generally, examine to what extent methods employing different criteria perform (dis)similarly in model recovery.

Second, only the bootstrap without SE and the simple holdout method can straightforwardly be extended to the simultaneous comparison of more than two models. All other methods are (currently) restricted to comparing pairs of models.

Third, the bootstrap with SE and the PED are the only methods that explicitly take into account the statistical variability and reliability during comparison. This renders these methods potentially superior to the other methods, because statistically reliable decisions between models can be assumed to be more accurate. On the other hand this property comes with the potential disadvantage that no decision may be possible in certain situations ${ }^{1}$. Accordingly, the overall quality of the bootstrap with SE and the PED will depend on the precise tradeoff between how accurately a decision between models can be taken and the number of situations in which a decision is reached.

## Approach

Three hypothetical models of memory decay, $M 1, M 2$, and M3, were used to assess the model comparison methods. Each of these models predicts the probability of recall in dependance on the time $t$ that has passed since the to-beremembered items have been learned. The models are defined by the following formulas (see Pitt \& Myung, 2002):

$$
\begin{aligned}
& M 1:(1+t)^{-a}, a \in[0,2] \\
& M 2:(b+t)^{-a}, a \in[0,2], b \in[1,2] \\
& M 3:(1+b t)^{-a}, a \in[0,2], b \in[0,2]
\end{aligned}
$$

Note that $M 1$ is nested in both $M 2$ and $M 3$, but nesting is different in the two cases. Since, furthermore, $M 2$ and $M 3$ are not nested, the three models allowed to examine the comparison methods regarding their ability to cope with different types of nesting as well as non-nested models.

Each method was applied to all three possible pairs of models, $M 1$ vs. $M 2, M 1$ vs. $M 3$, and $M 2$ vs. $M 3$ using the following general procedure. Given one of the three models, first, a set of parameter values was randomly drawn according to a uniform distribution from the range of parameter values specified above. Second, probabilities for this set of parameter values were generated from the model. Third, these probabilities were used to randomly sample the number of successful

[^207]recalls from a binomial distribution assuming a certain number learned items $(N L)$. Fourth, this set of numbers of successful recalls was treated as if it was a set of empirical observations for which to identify the most appropriate model. Accordingly, the comparison method in question was applied as described above to the model pair and the set of observations. Fifth, which (if any) of the two compared models was found to be more appropriate was noted. This procedure was repeated $R=100$ times for each model in each model pair. Across all model pairs and methods the measure to assess model fits and prediction error was always the mean squared error and the models were fit using a variant of the Metropolis algorithm (Madras, 2002).

Following this general procedure, our simulations varied 5 factors that potentially impact the performance of the comparison methods. Besides allowing to assess the importance of each of these factors for method performance, factor variation ensured a more general view on the methods accuracy in model recovery, that is, a view that is not specific to only one particular combination of factor levels. The considered factors are tightness of fit, strength of noise, number of data points, number of samples, and split ration and are described in the following.

Tightness of fit Fitting a model to a set of observations is a specific instance of a general type of optimization problems: Find the optimal set of parameter values for the given observations. It is well known that one is rarely guaranteed to find the optimum in such optimization problems. Thus, model fits may often be suboptimal to greater or lesser extent. This raises the question how susceptible the different comparison methods are to suboptimal model fits. To investigate this, we considered 3 levels of tightness of fits by varying how thoroughly the Metropolis algorithm searches the models' parameter space. More precisely, we varied the number of sets of parameters that were sampled (called swaps) for model fitting, using swaps $=100,1000$, and 10000. Simulations looking at the rates for recovering the generating model when fitting to the probabilities directly (i.e., looking at model behavior without adding sampling noise) corroborated that these numbers of swaps realized increasingly accurate model fits.

Strength of noise Since the only noise in the data is sampling noise, the amount of noise in the data is determined exclusively by the number of learned items: The higher $N L$ is the lower is the influence of sampling noise. Accordingly, employing $N L=5,50$, and 1000 allowed to examine the methods' capability to cope with noisy data.

Number of data points The information about the process that has generated a set of data can be assumed to increase with the number of available observations in the data set. To what extent the different methods require few or many data points for performing well was explored by varying the number of data points $(N D P)$. Levels of $N D P=5,20$, and 100
were employed and the corresponding data points were generated for $t$ distributed equidistantly in the range $[0.1,8.1]$.

Number of samples All of the methods come with a parameter that controls the amount of resources that are invested for model comparison. For PED and simple holdout this parameter is the number of splits that are considered (I), for both bootstrap variants this parameter is the number of bootstrap samples $(B)$, and for the CM this parameter is the number of GOF difference samples (NDS) each GOF difference distribution consists of. By using $I=10,100,1000$, $B=100,1000$, and $N D S=100,1000$ we gauged the models resource-performance trade-offs.

Split ratio Application of the PED and the simple hold-out requires splitting the set of observations into training and test sets and the relative sizes of the two sets is potentially crucial for comparison performance. If the training set is too small, insufficient information about the generating process may be available. If the training set is too large, the danger of over fitting may arise and the test set may become too small to obtain a reliable estimate of generalization performance. In our simulations we investigated splits with $Q=0.2,0.4$, and 0.6 , where $Q$ indicates the fraction of the original observations that are used for the training set.

To assess the methods' ability to outperform less elaborate approaches to model comparison, our simulations comprise the Akaike Information Criterion (AIC, Akaike, 1973) as the sixth method and a seventh method that we term simple recovery. Following the same general procedure as described above, simple recovery compares models by only considering the GOF of each model on the given data set: The model that provides the tighter fit is assumed to be the more appropriate model. Simple recovery and AIC simulations involve the same variations of the factors tightness of fit, strength of noise, and number of data points as employed for the 5 more sophisticated methods.

## Results

To characterize the methods' performance we computed, for each method, model pair, and situation, the sum of the percentages of cases in which both (a) a clear decision between the two models of pair could be taken and (b) the actually generating model was correctly recovered. If, for example, for the model pair M1-M2, M1 was correctly recovered $90 \%$ of the time and M2 was correctly recovered $43 \%$ of the time, the performance measure was computed to be $90+43=133$. Similarly, for BSSE and PED the percentages of cases where no model could be recovered with certainty was computed as the sum of the percentages of such cases for each of the two compared models. ${ }^{2}$. From the such obtained values the first,

[^208]

Figure 1: Quartiles of performance for the three considered model pairs and the seven considered methods. $A I C=$ Akaike Information Criterion, $B S S E=$ bootstrap with standard error, SHO = simple hold-out, $C M=$ cross-fitting method, $B S=$ bootstrap, $S R=$ simple recovery,$P E D=$ PED method.
second (median), and third quartiles (and associated standard errors) were determined for each method and model pair across all situations. Figure 1 and Figure 2 display the quartiles for the different methods.

As is evident from Figure 1, there are marked performance differences between model pairs and comparison methods. As one may have expected, the nested model pairs generally prove more difficult than the non-nested model pair, with M1M3 being even more difficult than M1-M2. It is mainly in the nested pairs that the less elaborate methods, AIC and simple recovery perform worse than all of the 5 more elaborate methods. Of the 5 more elaborate methods, PED, simple hold-out, and BSSE generally outperform BS and CM. In sum, PED, simple hold-out, and BSSE tend to perform best, AIC and simple recovery perform worst, and CM and BS show intermediate performance, but are only better than AIC and simple recovery for nested model pairs. As Figure 2 shows, the superior performance of PED and BSSE comes at the cost of a substantial number of cases in which the two methods do not allow to take a clear decision for one or the other model.

Several aspects of this pattern of results seem noteworthy. In contrast to the assumption that the CM is optimal for recovering the generating model (Cohen et al., 2008), the CM performs comparatively bad. On average, the CM is only better than SR for nested models, and generally worse in avoiding misclassifications than the PED, BSSE, and the simple hold-out. In fact, given its comparative simplicity, the simple hold-out performs remarkably well. While providing a decision for $100 \%$ of the cases, these decision are correct in more than $90 \%$ of the cases on average. This set of results also provides further evidence for dissimilarity in model recovery performance depending on whether a generalization criterion or a model recovery criterion is instantiated by the employed comparison method. Comparing the simple hold-out and CM indicates performance differences depending on which criterion is used and, more interestingly, that a method using the generalization criterion can outperform a method using the recovery criterion in model recovery.

In addition to the results across all factor combinations, considering the impact each factor has on method performance yields a number of interesting insights.

Tightness of fit Across all methods, the influence of the tightness of fit (if present at all) is only considerable between loose fits $($ swaps $=100)$ and moderate to tight fits (swaps $=1000$ and 10000). In simple recovery, the tendency to select the more complex models increases with tightness of fits such that for moderate and tight fits the nesting model is selected more often even if the nested model generated the data. Except for pair M1-M3, performance of AIC increases considerably with tighter fits. In comparison, the bootstrap with SE and the CM, exhibit less (but still noticeable) susceptibility to tightness of fit in the sense that with tighter fits for nested model pairs the overall correct recovery rate increases by selectively increasing the correct recovery rate of


Figure 2: Quartiles of the number of cases for which BSSE and $P E D$ do not allow to take a decision.
the nesting model. Put differently, for loose fits, the CM and the bootstrap with SE tend to erroneously favor the simpler model; a problem that is mitigated when using tighter fits. The remaining three methods are largely insensitive to tightness of fits indicating that, for these methods, it may not be the absolute but the relative tightness of fit that matters.

Strength of noise Not surprisingly, all methods get consistently better with decreasing strength of noise. Furthermore, all methods encounter severe difficulties with the highest noise level $(N L=5)$ that leads to near chance performance for most model pairs and methods. The methods differ, however, regarding the level of noise from which they start to show good or very good performance. While the simple holdout, the bootstrap with SE and the PED achieve high accuracy already for $N L=50$, the CM and the bootstrap tend to do so only for $N L=1000$.

Number of data points Although all methods but the AIC tend to improve with an increase in the number of data points, there are marked differences with respect to the strength of the influence of this factor. The PED and the bootstrap with SE are impacted severely by the number of data points improving considerably - especially for nested models - with an increase from $N D P=5$ to $N D P=20$ as well as from $N D P=20$ to $N D P=100$ both regarding accuracy and the percentage of decision that can be made. The other four methods are much less sensitive to $N D P$ levels, but exhibit a tendency for a reduction in erroneously selecting a nested model when the data was generated from a nesting model. Interestingly, the performance of the AIC drops with increasing NDP due to an increased tendency to erroneously pick the nested model.

Number of samples Effects of increasing the number of samples are mixed across the methods. This factor has virtually no effect on the bootstrap. Yet, for the bootstrap with SE increasing the number of samples leads to a decrease in the percentage of cases in which a decision can be made and to a tendency to more often select the simpler of the two compared models. Both PED and simple hold-out perform better with increased $I$, but this trend is largely due to the difference between $I=10$ and $I=100$. Similar to the bootstrap with SE, the PED allows (slightly) fewer decision with increasing $I$. The CM exhibits a shift towards more often selecting the more complex model with increased numbers of samples.

Split ratio The split ratio has only little impact on the performance of the PED and the simple hold-out. While the number of cases that cannot be decided by the PED slightly increases with an increase in $Q$, the accuracy remains generally high. Only for comparing $M 1$ and $M 3$ do higher values of $Q$ lead to pronounced performance decrements. Similarly, the simple hold-out becomes slightly but consistently worse in correctly recovering the nested model in the two nested model pairs with an increase in $Q$.

## Conclusion

Our simulation studies revealed a number of interesting properties of the considered comparison methods. First, methods employing a generalization criterion for model comparison (e.g., simple hold-out) can outperform methods supposedly optimal for model recovery (the CM) in model recovery. Second, although all 5 considered methods can substantially improve on less elaborate approaches (as instantiated by the AIC and the simple recovery method), the less elaborate methods may perform better under certain conditions. Thus, whether the use of one of the examined methods is advantageous will depend on the precise nature of the model comparison situation at hand (e.g., how many data points are available and how noisy the data is). Third, the considered methods differ noticeably in the degree to which their performance depends on the characteristics of the comparison situation. The comparatively low quartiles of the bootstrap and the CM indicates that these methods outperform the less elaborate approaches only in comparatively few particular settings. Fourth, the highest accuracies were achieved by the PED, but this method allows decisions about which of the compared models is more appropriate only in very few cases. Furthermore, performance of the PED breaks down if only few data points are available. Fifth, despite its comparable simplicity, the simple hold-out method achieves high accuracies while allowing to select one of the models in $100 \%$ of all cases. In addition, the simple hold-out is the only method that can be easily extended to comparing more than two models.

Against this background our results suggest to employ the PED if only pairs of models have to be compared and if accuracy is more important than being able to reach a decision. The simple hold-out appears to be a good choice if more than
two models need to be compared and / or if it is important to reach a decision on which of the compared models to select.

Although this initial assessment already highlights important properties of the comparison methods, it is best viewed as a first glimpse on the methods' characteristics. Further research considering a range of different (types of) models is required to provide a more comprehensive picture of the strengths and weaknesses of available comparison methods. Besides taking up this task we intend to explore modifications of the CM, PED, and bootstrap with SE that renders them applicable to comparing more than two models in our future work.

## Acknowledgments

The authors gratefully acknowledge support by the German Academic Exchange Service (DAAD) and the German Research Foundation (DFG) through the project R1[ImageSpace], SFB/TR 8 Spatial Cognition

## References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov (Ed.), Proceedings of the Second International Symposium on Information Theory (p. 267-281).
Busemeyer, J. R., \& Wang, Y.-M. (2000). Model comparisons and model selections based on generalization criterion methodology. Journal of Mathematical Psychology, 44, 171-189.
Cohen, A. L., Sanborn, A. N., \& Shiffrin, R. M. (2008). Model evaluation using grouped or individual data. Psychonomic Bulletin \& Review, 15, 692-712.
Efron, B., \& Tibshirani, R. J. (1993). An introduction to the bootstrap. New York: Chapman \& Hall.
Efron, B., \& Tibshirani, R. J. (1997). Improvements on crossvalidation: The . 632+ bootstrap method. J Am Stat Assoc, 92, 548-560.
Madras, N. N. (2002). Lectures on monte carlo methods. Providence, Rhode Island: American Mathematical Society.
Pitt, M. A., \& Myung, J. (2002). When a good fit can be bad. TRENDS in Cognitive Sciences, 6, 421-425.
Schultheis, H., \& Singhaniya, A. (accepted). Decision criteria for model comparison using cross-fitting. In 22nd Annual Conference on Behavior Representation in Modeling \& Simulation (BRiMS 2013).
Shiffrin, R. M., Lee, M. D., Kim, W., \& Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical bayesian methods. Cognitive Science, 32, 1248-1284.
van de Wiel, M. A., Berkhof, J., \& van Wieringen, W. N. (2009). Testing the prediction error difference between two predictors. Biostatistics, 10, 550-560.
Wagenmakers, E.-J., Ratcliff, R., Gomez, P., \& Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. Journal of Mathematical Psychology, 48, 2850.

# Rational preference shifts in multi-attribute choice: What is fair? 

Pradeep Shenoy (pshenoy@cs.washington.edu)<br>Microsoft AdCenter, Bangalore, India<br>Angela J. Yu (ajyu@ucsd.edu)<br>Department of Cognitive Science, UC San Diego, 9500 Gilman Dr. MC 0515<br>La Jolla, CA 92093 USA


#### Abstract

Humans exhibit certain systematic context-dependent preference reversals when choosing among options that vary along multiple attribute dimensions. For instance, the attraction, similarity, and compromise effects each involves a change in relative preference between two options when a third option is introduced. Previously, such effects have been attributed to irrationality or suboptimality in decision-making, or to specific architectural or dynamical constraints on cognition. We use a Bayesian model of multi-attribute choice to demonstrate that these effects naturally arise from three basic assumptions: (1) humans assess options relative to "fair market value" as inferred from prior experience and available options; (2) attributes are imperfectly substitutable, and scarce attributes are relatively more valuable; (3) uncertainty about market conditions and option values contributes to stochasticity in choice behavior. This work provides both a novel normative explanation for contextual modulation of choice behavior, and a means to predict choice as a function of past experiences and novel contexts.


Keywords: multi-attribute decision-making; preference shift; context effects; attraction effect; compromise effect; similarity effect

## Introduction

Everyday decision-making often involves choosing among options that differ in multiple attribute dimensions. For example, should you buy a house that is more spacious or one that is better located? Understanding how humans make these multi-attribute decisions, and how their choices depend on the context, is an important problem in cognitive science.

Multi-attribute decision-making is particularly challenging because there is often no universal or intrinsic way to assign relative values to the different attributes. This is especially true in contexts where the decisionmaker has limited experience (and thus significant uncertainty about market conditions), such as with big-ticket items like houses, or new technology like smart phones. Human choice behavior in multi-attribute problems exhibits certain systematic shifts due to context changes, such as when the relative preference between two options shift or even reverse when a third option, known as a decoy, is added, leading to suggestions of underlying irrationality or suboptimality (Kahneman \& Tversky, 1979; Kahneman, Slovic, \& Tversky, 1982; Tversky \& Simonson, 1993).

In the attraction effect (Fig. 1A), given two similarly preferred options, $A$ and $B$, the introduction of a third option $Z$ that is similar to $B$, but also clearly inferior to $B$ in one or both attribute dimensions, results in an increase in relative preference for $B$ over $A$ (Huber, Payne, \& Puto, 1982; Heath \& Chatterjee, 1995). In the compromise effect (Fig. 1B), when $B>A$ in one attribute and $B<A$ in another attribute, and $Z$ has the same tradeoff but is even more extreme than $B$, then $B$ becomes the "compromise" option and becomes preferred relative to $A$ (Simonson, 1989). In the similarity effect (Fig. 1C), the introduction of a third option $Z$, that is very similar and comparable to $B$ in both attribute dimensions, shifts the relative preference away from $B$ to A (Tversky, 1972).


Figure 1: Three classical contextual effects in multiattribute choice: (A) attraction effect, (B) compromise effect, (C) similarity effect. $A$ and $B$ are two equally preferable choices that differ in two attribute dimensions. The introduction of a third option $Z$ induces a preference shift between $A$ and $B$ (indicated by arrows). Solid and dashed lines illustrate model-inferred "fair value" indifference curve before and after introducing $Z$.

Two broad classes of models have previously been proposed for contextual effects in multi-attribute choice behavior: (1) normative models (Marr, 1982) that are built on behavioral constraints/goals and delineated in terms of internal beliefs and assumptions (Luce, 1959; Thurstone, 1954; Luce, 1965; Tversky, 1972; Tversky \& Simonson, 1993); (2) algorithmic or implementational models that explain behavioral phenomena as arising from specific architectural and dynamical constraints on neural processing (Busemeyer \& Townsend, 1993; Usher \& McClelland, 2004; J. S. Trueblood, 2012).

The first class of models are related to bounded rationality (Simon, 1955), but have so far been unable to explain all three contextual effects, leading to sugges-
tions that such preference shifts reflect biases or suboptimalities in human decision-making. For example, the discovery of the similarity effect invalidated Luce's early ratio-of-strength model (Luce, 1959), and other related models that follow the simple scalability principle (Tversky, 1972). Tversky proposed the elimination-byaspects model (Tversky, 1972) to explain the similarity effect, but it was invalidated by the discovery of the attraction effect, which violates the regularity principle, thus ruling out a large class of random utility models (Luce, 1965), including Thurstone's preferential choice theory (Thurstone, 1954). The compromise effect presented further complication, as no previous model could account for it, and a new context-dependent preference model (Tversky \& Simonson, 1993) was only able to account for it, along with the attraction effect, by letting slip the similarity effect (Roe, Busemeyer, \& Townsend, 2001).

The second class of models can account for all three effects, but are based on rather detailed and specific assumptions about neural dynamic and architecture, which have thus far not been verified experimentally, and whose computational provenance and consequences are unclear.

Here, we propose a novel rational account of multiattribute decision-making that explains all three contextual effects. The model is grounded in three basic, empirically motivated assumptions: (1) humans make preferential choices based on relative values anchored with respect to what is perceived "fair" in the marketplace (Ariely, 2008), which is inferred from observed data, including the set of available options (Wernerfelt, 1995; Sher \& McKenzie, 2011); (2) different attributes are imperfect substitutes for one another (Hicks, 1932), in particular one unit of a scarce attribute is more valuable than an abundant one; (3) uncertainty in posterior belief about "market conditions" contributes to stochasticity in preference on repeated presentations of the same options (see e.g., Debreu, 1958). We formalize these assumptions using a Bayesian generative model, and demonstrate that all three contextual effects are consequences of rational (Bayesian) inference of relative value, conditioned on the available options. In contrast to previous models, we view each decision as not only an expression of choice, but also as an opportunity for learning about the marketplace based on the set of options given. Thus, an individual's preference can differ in different contexts, not because of arbitrary contextdependent factors (Hsee, Zhang, Yu, \& Xi, 2004; Srivastava \& Schrater, 2012), but because of normative evolution of an individual's internal beliefs about the option landscape. Moreover, our model provides a means to predict individual and group preferences in novel contexts given past choices. In the following, we first describe the Bayesian model, followed by a comparison of simulated model behavior and empirically observed contextual ef-
fects found in the literature, and finally conclude with a discussion.

## Bayesian model of relative value inference

We begin with an intuitive explanation for contextual effects before delving into the technical details of the model. While we explain the phenomena primarily in terms of consumer decision-making here, in the Discussion we will extend the model and explanation beyond choices among consumer products.

We model presented options as being drawn from a shared landscape of options, which implies that the options are representative of the market in some sense, and are useful for inferring general market conditions. In fact, humans often use available context to infer a reference point for valuation-for instance, in the framing effect, humans evaluate the quality of an outcome differently based on whether it is described in terms of success rates or failure rates (Sher \& McKenzie, 2006). In the case of multi-attribute valuation, "fair market value" could potentially be inferred by fitting an equi-preference contour, or indifference curve (Pareto, 1927), through the presented options, where points above the line would be a "good deal", while ones below would be a "bad deal." Given the formal relationship between regression and inference (Bishop, 2006), this process is equivalent to inferring mean market value and relative attribute importance based on the samples. We first use this general intuition to explain the three effects, and subsequently present a precise generative model and inference procedure for multi-attribute choice.

In the attraction effect, $A$ and $B$ both initially lie on the inferred "fair value" indifference curve. Introducing $Z$, which is close to $B$ but clearly inferior in one or both attribute dimensions, drives down the inferred "fair value" indifference line (dashed line) near $B$, making $B$ appear to be a good deal (while $A$ is still fair, and $Z$ is worse than fair). The compromise effect arises from imperfect substitutability and diminishing marginal utility (Hicks, 1932) - e.g. the value of a small house in a good location would increase much more with a small increase in size than it would with a slight improvement in location. Thus, the indifference curves, including the fair value curve, should be strictly convex rather than linear. The compromise effect then naturally arises when $Z$ is introduced, because the convex line corresponding to "fair" passes between $B$ and $Z$, making $B$ appear to be better than fair (and $A$ fair or worse than fair). To account for the similarity effect, we adopt a stochastic decision policy that reflects posterior uncertainty about both market conditions and option values: the model samples from the joint posterior distribution over option values and chooses the maximally valued option. Because of the proximity of $B$ and $Z$ in the attribute
space, inferred values of $B$ and $Z$ are highly correlated. For each possible setting of market conditions (family of indifference curves), $B$ and $Z$ tend to be both better or worse than $A$. This gives an overall probability of choos$\operatorname{ing} A$ with $1 / 2$ probability, and choosing $B$ (and also $Z$ ) with $1 / 4$ probability.

## Model

$$
\begin{aligned}
& \mu \sim \Gamma\left(k_{\mu}, t_{\mu}\right) \\
& \gamma \sim \beta\left(a_{\gamma}, b_{\gamma}\right) \\
& \theta_{i} \sim \mathcal{N}\left(0, \sigma_{\theta}^{2}\right) \\
& v_{i} \sim \Gamma\left(k_{v}, \frac{\mu}{k_{v}}\right) \\
& o_{i} \sim \mathcal{N}\left(f\left(v_{i}, \theta_{i}, \gamma\right), \sigma_{o}^{2} \mathbf{I}\right) \\
& f\left(v_{i}, \theta_{i}, \gamma\right)=\left(v_{i}^{\frac{1+\theta_{i}}{2 \gamma}}, v_{i}^{\frac{1-\theta_{i}}{2(1-\gamma)}}\right)
\end{aligned}
$$



Figure 2: Bayesian generative model of relative value inference. Each two-attribute option $o_{i}=\left(x_{i}, y_{i}\right)$ has an underlying scalar value $v_{i}$, parameterized by $\gamma, \theta_{i}$. The value $v_{i}$ itself is generated from a prior distribution with mean $\mu$, which corresponds to "fair" value.

The critical assumptions in our model are that subjects use available options to infer about the utility function and "fair market value." We assume subjects do so by inverting a hierarchical Bayesian generative model (Fig. 2), where: (1) values $\left\{v_{i}\right\}$ for the set of options $\{i\}$ are drawn from a prior distribution with mean $\mu$, and (2) 2-d attribute values for each option, $o_{i}$, is generated from $v_{i}$ according to a common utility function and then corrupted by observation noise. For simplicity, we use the classical Cobb-Douglas utility function (Douglas, 1976), parameterized by $\gamma, v_{i}=x_{i}^{\gamma} y_{i}^{(1-\gamma)}$. While more complex utility functions can be used, for example to take into account variability in the relative scaling of the two attributes, the contextual effects are not dependent on the choice of utility function, and thus not dealt with further here. To model observation noise, we first map value into an indifference curve in the attribute space by inverting the utility function, then add Gaussian noise along the indifference curve (parameterized by $\theta_{i}$ ) and isotrophic 2-D Gaussian noise (parameterized by $\sigma_{0}$ ). We expect the main results to hold independent of the specific choices of model parameterization.

Subsequent to doing posterior inference, we assume humans choose an option by first sampling from the joint posterior $P(\mathbf{v} \mid \mathbf{o})$, and then (always) choosing the option with the highest sampled value. The computation of the posterior requires marginalizing over unceratinty about market conditions through a series of steps:

$$
P(\mathbf{v}, \mathbf{o}, \mu, \gamma)=p(\mu) p(\gamma) \Pi_{i}\left[\int_{\theta_{i}} p\left(\theta_{i}\right) P\left(v_{i} \mid \mu\right) P\left(o_{i} \mid \theta_{i}, v_{i}, \gamma\right)\right]
$$



Figure 3: Preference shifts as rational inference. (A) Model chooses $A$ and $B$ equally when there are only two options. (B) Attraction: introducing the inferior $Z$ makes $B$ more preferable to $A$ ( $Z$ is almost never chosen). (C) Compromise: introducing an extreme option $Z$ makes $B$ more preferable to $A$. (B) Similarity: introducing $Z$, highly similar to $B$, makes $B$ less preferable to $A$.

$$
P(\mathbf{v} \mid \mathbf{o}) \propto \int_{\mu, \gamma} P(\mathbf{v}, \mathbf{o}, \mu, \gamma)
$$

## Simulation details

The parameter settings for our simulations were as follows: $\left(k_{\mu}, t_{\mu}\right)=(1,100) ;\left(a_{\gamma}, b_{\gamma}\right)=(2,2) ; \sigma_{\theta}=20 ; k_{v}=$ $20 ; \sigma_{o}=2$ (see Fig. 2). The Gamma distributions were parametrized using parameters for shape ( $k_{\mu}$ ), and scale $\left(t_{\mu}\right)$, and the mean of the corresponding distribution is given by their product (e.g., $k_{\mu} \cdot t_{\mu}$ ). Accordingly, the mean of the prior distribution over $\mu$ is 100 , and the shape parameter encodes a broad uncertainty about the true value of $\mu$ (see Fig. 4).

We finely discretized each of the variables in our model to calculate the relevant posterior distributions numerically (analytical solutions do not exist). The option values used (see Fig. 1) were as follows: $A=(40,60), B=$ $(60,40)$; attraction: $Z=(30,50)$, compromise: $Z=$ $(80,20)$, similarity: $Z=(65,35)$.

## Results

## Preference shift as option-based value inference

As Fig. 3 shows, simulations of our model reproduces all three contextual effects: attraction, compromise, and similarity. In particular, the model reproduces violation of regularity in attraction effect that is also seen in human data. In all three cases, although options $A$ and $B$ are equally preferred when presented as a pair (Fig. 3A),
the presence of a third (decoy) option $Z$ changes this relative preference (Fig. 3B-D). This shift in preference depends on the relationship between the precise attribute values of the decoy relative to those of the two original two options (see Fig. 1). All three contextual effects were obtained using the same model setting, except for the position of the decoy $Z$. Thus, these contextual shifts in preference can indeed be direct consequences of normative inference about relative values, conditioned on both prior beliefs and the available options. Note that the main results hold over a wide range of parameter settings and are not sensitive to the particular parameterization of the model.


Figure 4: Posterior distributions over model variables for the compromise effect. (A) Marginal posterior distribution over what constitutes "fair" in the market, parameterized by $\mu$, before (red) and after (green) introducing Z. (B) (A, B) Marginal posterior distribution over the shape of the family of indifference curves, parameterized by $\gamma$, before (red) and after (green) introducing $Z$.

We explore the compromise effect in more detail to illustrate the inner workings of our model. The joint inference over $(\mu, \gamma)$ is reflected in the shape of the equipreference contours and the probability of each contour being "fair" (Fig. 5): colored bands represent indifference curves for the MAP estimate of $\gamma$ and a range of values of $\mu$, and the color indicates the probability of that band representing fair market value. When only options $(A, B)$ are presented (Fig. 5A), the fair market value contour passes through both $A$ and $B$; when $Z$ is introduced (Fig. 5B), the contours shift so as to make $B$ better than fair (and $A$ fair).

Next, we examine the properties of the inferred joint posterior distribution $P(\mathbf{v} \mid \mathbf{o})$, illustrated in Fig. 6. Shown in panel A are the marginal value distributions for the 3 options $A, B, Z$ in the compromise effect. Consistent with Fig. 5, the inferred value distributions show a clear ordering, with option $B$ having the highest expected value (Fig. 6B). However, the marginal distributions (panel A), and expected values (panel B) do not capture implicit correlations among inferred values induced by the shared (marginalized) variables $\mu$ and $\gamma$. In fact, our model samples from the joint posterior value distribution and selects the highest value in each sample. Fig. 6C shows the the empirical probability of each


Figure 5: Joint marketplace and value inference in the compromise effect: (A) given only $A$ and $B$ as options, (B) given $A, B, Z$. Each colored band represents an equipreference contour (indifference curve) corresponding to the MAP estimate of $\gamma$, with its color indicating the probability of its being the mean market value.


Figure 6: Value inference and sampling in compromise effect. (A) Posterior distributions over option values. (B) Mean posterior value. (C) Empirical choice distribution based on samples $(\mathrm{n}=1000)$ from the joint posterior distribution over option values.


Figure 7: Value inference and sampling in similarity effect. See Fig. 6.
option being chosen. In the joint distribution, $A$ and $Z$ are positively correlated in inferred value, and, as a result, our model strongly prefers the compromise option $B$ over $A$, stronger than would be suggested by the marginal distributions alone.

Correlations in the joint posterior value distribution is particularly important also for generating the similarity effect (Fig. 7), where the marginal distributions and mean values for $A$ and $B$ are indistinguishable from each other, but the sampled preference for $A$ is much higher, due to a strong positive correlation between the inferred values of $B$ and $Z$.

## Model predictions

Our model makes a number of experimentally testable predictions about multi-attribute choice behavior. Since presented options not only influence the immediate choice but also general beliefs about general market conditions, our model predicts systematic consequences in future choice behavior based on experienced choice history. For instance, subjects exposed to a number of choices between options generally higher in one attribute may correspondingly learn a $\gamma$ that discounts this attribute more - resulting in a smaller attraction effect for a decoy that is inferior to $B$ in this attribute dimension compared to the other. There is some evidence that subjects show such "context-dependent utility functions" (Drolet, Simonson, \& Tversky, 2000).

Another arena for experimental exploration suggested by this work is the transition among the different effects due to the precise positioning of the options in the attribute space: for instance, the "similarity" decoy in Fig. 1C could well turn into a "compromise" decoy in Fig. 1B, if it were far enough from $B$. Thus, one prediction of our model is that as the decoy $Z$ is moved away from the option $B$, while maintaining a rough tradeoff between the two attributes, the contextual effect changes from similarity effect to compromise effect. That is, if the decoy were exactly the same as $B$, preference should shift away from $B$, but as the decoy is moved further apart, preference should shift toward $B$. In an analogous manner, we expect to see a smooth transition between the similarity and attraction effects as the decoy is moved away in the orthogonal, dominated direction. Fig. 8 shows that model simulations conform to these expectations: as the decoy is moved further along nondominated (panel A) or dominated (panel B) directions, the model predicts a gradual evolutation from a similarity effect to the compromise and attraction effects, respectively.

## Discussion

We presented a normative Bayesian model for why human subjects exhibit apparently irrational choice behavior in multi-attribute decision-making. We showed that violations of the simple scalability and regularity principles need not be reflections of an irrational or suboptimal decision or valuation process, but rather rational consequences of a decision-maker who is trying to optimize choice in a relativistic system anchored to what is perceived to be fair. We used a normative, hierarchical Bayesian generative model to demonstrate how the set of options themselves can be used to infer about the landscape of available options, such as how value is distributed in the market, how the multi-dimensional observed attribute space is mapped to the scalar value representation, and the distribution of observation noise. Although the language of this paper primarily focuses on


Figure 8: Transitions in relative preference. The three effects are related to each other by the magnitude of the distance between options. (A) When a third option $Z$, initially identical to $B$, is moved away in a nondominated direction, relative preference changes from favoring $A$ (similarity effect) to favoring $B$ (compromise effect). (B) When $Z$ is moved away from $B$ in an orthogonal, dominated direction, preference changes from favoring $A$ (similarity effect) to favoring $B$ (attraction effect).
consumer decision-making, the model can be extended to a much broader range of multi-attribute choice behavior, whenever the observer has uncertainty about how to combine two attributes in order to compare the options. In future work, we plan to extend the current model to explore some non-consumer choice tasks known to exhibit context effects (Choplin \& Hummel, 2005; J. S. Trueblood, 2012; J. Trueblood, Brown, Heathcote, \& Busemeyer, n.d.).

Our approach contrasts with the class of models that explain contextual effects based on specific architectural or dynamic constraints on neural processing. One example is the decision field theory (DFT) model (Busemeyer \& Townsend, 1993), which assumes that the strength of preference for each option is driven by a noisy, accumulative input and dynamical switching of "attention" among different attribute dimensions, as well as "lateral inhibition" between the different units. A related model (J. S. Trueblood, 2012), an extension of the multiattribute linear ballistic accumulator model (Brown \& Heathcote, 2008), employs attentional switching, a contrast mechanism (related to lateral inhibition), and sensitivity to indifference/dominance. A third model, the competing accumulator model (Usher \& McClelland, 2004), assumes loss aversion in addition to attentional switching and lateral inhibition. The various overlapping and nonidentical assumptions of these process models are difficult to verify experimentally, and their computational provenance/constraints are not well understood. This is not to say that such mechanistic models are not useful. Ultimately, to understand how the brain implements multi-attribute choice, we need multiple levels of analysis (Marr, 1982) that integrate both normative and mechanistic explanations. In this vein, our work comple-
ments existing work by helping to frame and constrain mechanistic models.

Although the model presented here succinctly and rationally accounts for contextual effects in multi-attribute choice behavior, it is clearly not a complete theory of human preference choice. In particular, the simple model presented here has no means of accounting for individual differences according to taste. A natural way this arises is when people bring in different previous experiences and thus prior beliefs about the market. However, this cannot be the whole story, as any prior difference would be overwhelmed by sufficient data, and yet people who have repeated exposure to the same choices do not always converge in their preferences (e.g. office workers who eat out at the same set of neighborhood restaurants day after day). An important line of future enquiry is how individual differences in preference may arise and persist in the face of mounting, common experiences.

## References

Ariely, D. (2008). Predictably irrational. Harper Collins.
Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.
Brown, S., \& Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. Cognitive psychology, 57(3), 153-178.
Busemeyer, J., \& Townsend, J. (1993). Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. Psychological review, $100(3), 432-459$.
Choplin, J., \& Hummel, J. (2005). Comparison-induced decoy effects. Memory $\xi^{\mathcal{G}}$ Cognition, 33(2), 332-343.
Debreu, G. (1958). Stochastic choice and cardinal utility. Econometrica, 26, 440-444.
Douglas, P. (1976). The Cobb-Douglas production function once again: Its history, its testing, and some new empirical values. The Journal of Political Economy, 84, 903-915.
Drolet, A., Simonson, I., \& Tversky, A. (2000). Indifference curves that travel with the choice set. Marketing Letters, 11:3, 199-209.
Heath, T., \& Chatterjee, S. (1995). Asymmetric decoy effects on lower-quality versus higher-quality brands: Meta-analytic and experimental evidence. Journal of Consumer Research, 22, 268-284.
Hicks, J. R. (1932). The Theory of Wages. London, UK: Macmillan.
Hsee, C., Zhang, J., Yu, F., \& Xi, Y. (2004). Lay rationalism and inconsistency between predicted experience and decision. Journal of Behavioral Decision Making, 16.

Huber, J., Payne, J., \& Puto, C. (1982). Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. Journal of Consumer Research, 9(1), 90-98.

Kahneman, D., Slovic, P., \& Tversky, A. (1982). Judgment Under Uncertainty: Heuristics and Biases. New York, NY: Cambridge Univ Press.
Kahneman, D., \& Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica: Journal of the Econometric Society, 47(2), 263-292.
Luce, R. (1959). Individual choice behavior: a theoretical analysis. New York: Wiley.
Luce, R. (1965). Preference, utility, and subjective probability. In R. Luce, R. Bush, \& E. Galanter (Eds.), Handbook of mathematical psychology (Vol. 3, pp. 249410). New York: Wiley.

Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information. San Francisco: Freeman.
Pareto, V. (1927). Manual of political economy Tr. by Ann S. Schwier. London: Macmillan.
Roe, R. M., Busemeyer, J. R., \& Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionst model of decision making. Psychological Review, $108(2), 370-392$.
Sher, S., \& McKenzie, C. (2006). Information leakage from logically equivalent frames. Cognition, 101 (3), 467-494.
Sher, S., \& McKenzie, C. R. (2011). Levels of Information: A Framing Hierarchy. In Perspectives on framing. Psychology Press.
Simon, H. (1955). A behavioral model of rational choice. The quarterly journal of economics, 69(1), 99-118.
Simonson, I. (1989). Choice based on reasons: the case of attraction and compromise effects. Journal of Consumer Research, 16(158-74).
Srivastava, N., \& Schrater, P. (2012). Rational inference of relative preferences. Advance in Neural Information Processing Systems, 25.
Thurstone, L. (1954). The measurement of values. Psychological Review, 61, 47-58.
Trueblood, J., Brown, S., Heathcote, A., \& Busemeyer, J. R. (n.d.). Not just for consumers: Context effects are fundamental to decision-making. Psychological Science.
Trueblood, J. S. (2012, October). Multialternative context effects obtained using an inference task. Psychonomic Bulletin $\mathcal{E}^{3}$ Review, 19(5), 962-8.
Tversky, A. (1972). Elimination by aspects: a theory of choice. Psychological review, 79(4), 281-299.
Tversky, A., \& Simonson, I. (1993). Context dependent preferences. Management Science, 39, 1179-89.
Usher, M., \& McClelland, J. (2004). Loss aversion and inhibition in dynamical models of multialternative choice. Psychological review, 111(3), 757-760.
Wernerfelt, B. (1995). A rational reconstruction of the compromise effect: Using market data to infer utilities. The Journal of Consumer Research, 4(2), 627-633.

# Automatic and Instructed Attention in Learned Predictiveness 

Lauren T. Shone (Isho0771@uni.sydney.edu.au)<br>Evan J. Livesey (evan.livesey@sydney.edu.au)<br>School of Psychology, Brennan McCallum Building, University of Sydney, NSW 2006 AUS


#### Abstract

In novel situations, learning is biased towards information that has a degree of prior predictive utility. In human learning, this is termed the learned predictiveness effect and has proved critical in theorising about the role of attention in learning. Two experiments are reported in which the relative contribution of controlled and automatic processes to learned predictiveness are investigated. Experiment 1 showed that while learned predictiveness is susceptible to instructional manipulation, this effect is partial. Experiment 2 manipulated predictive utility and instruction orthogonally in order to test the potential involvement of automatic processes. It was found that even when cues were explicitly instructed as causal, learning was biased in favour of previously predictive over previously non-predictive cues. Interestingly, this was reversed for cues instructed as irrelevant. This suggests that learned predictiveness benefits attentional control, whereby information is both easier to attend and ignore.


Keywords: human learning, attention, controlled processing, automatic processing

## Introduction

An important question facing theories of associative learning is the nature of the relationship between learning and attention. Accordingly, many associative theories (e.g., Kruschke, 2001; Mackintosh, 1975; Pearce \& Hall, 1980) accept that stimulus selection is influenced by attentional processes. Such theories share the basic assumption that the attention devoted to a stimulus is flexible, and governed by its past utility in predicting events. Importantly, this will subsequently influence the rate at which a stimulus enters into future associations.

Evidence in favour of learned attention originates from experiments in which past predictive utility biases learning in a novel situation. A robust example, first reported by Le Pelley and McLaren (2003; see also Lochmann \& Wills, 2003), is the learned predictiveness effect. The basic experimental design used to demonstrate the effect is shown in Table 1. Participants are initially exposed to a scenario in which they are required to learn a causal relationship between cues and outcomes. Each trial consists of the presentation of a compound of two cues, leading to one of two outcomes. Critically, each compound consists of one perfectly predictive cue (represented by A - D), and one non-predictive cue ( $\mathrm{W}-\mathrm{Z}$ ). For example, A is consistently paired with the outcome O1, and therefore has perfect predictive utility. Alternatively, W has no predictive utility because it is paired equally often with both outcomes O1 and O2.

Once these relationships have been learned, a novel scenario is introduced. The same cues, in novel
combinations, are then employed in order to predict different outcomes. Importantly, although the cues are again presented in compound, this time neither component has superior predictive utility. That is, both A and W are perfect predictors as they share the same objective relationship with outcomes O3 and O4 respectively. What differs between the components of the new compounds is their status as a predictive or non-predictive cue in the initial stage of learning. Subsequent tests reveal that more is learned about the relationship between previously predictive cues and the new outcomes compared to previously non-predictive cues.

Table 1. A typical learned predictiveness design.

| Phase 1 | Phase 2 | Test |
| :---: | :---: | :---: |
| $\mathrm{AW}-\mathrm{O} 1$ | $\mathrm{AY}-\mathrm{O} 3$ | AD |
| $\mathrm{AX}-\mathrm{O} 1$ | $\mathrm{BZ}-\mathrm{O} 4$ | XY |
| $\mathrm{BW}-\mathrm{O} 2$ | $\mathrm{CW}-\mathrm{O} 4$ | BC |
| $\mathrm{BX}-\mathrm{O} 2$ | $\mathrm{DX}-\mathrm{O} 3$ | WZ |
| $\mathrm{CY}-\mathrm{O} 1$ |  |  |
| $\mathrm{CZ}-\mathrm{O} 1$ |  |  |
| $\mathrm{DY}-\mathrm{O} 2$ |  |  |
| $\mathrm{DY}-\mathrm{O} 2$ |  |  |

Note. Letters indicate individual cues. O1-O4 refer to four outcomes.

Traditionally, this bias, consistently replicated across various scenarios (see Le Pelley, 2010, for a recent review), has been interpreted to suggest that attention is modulated by the difference in predictive validity during initial stages of learning. According to this logic, attention to $\mathrm{A}-\mathrm{D}$ will be high following phase 1 and will therefore have an advantage when entering into new associations during the second phase. This effect has proved critical in theorising about the reciprocal nature of the relationship between human learning and attention.

The learned predictiveness effect is consistent with models of associative learning that assume attention changes according to mechanisms of associative competition (e.g., Mackintosh, 1975; Le Pelley, 2004; Pearce \& Mackintosh, 2010). For example, Mackintosh (1975) proposed that changes in the association between a cue and an outcome are governed by both attention paid to the cue and the discrepancy between the occurrence of the outcome and the extent to which it is already predicted on the basis of that cue, that is, the prediction error for an individual cue. Critically, attention to the cue changes according to a
comparison between its prediction error and the prediction error for other cues available at the same time. The cues with smaller individual prediction errors (i.e. those with higher predictive utility) will command more attention as learning proceeds. Higher attention, in turn, drives faster learning.

Despite its replicability, the exact nature of the learned predictiveness effect has only recently been questioned. Indeed, the concept of attention is associated with a variety of cognitive mechanisms (see Pashler, 1998; Wright \& Ward, 2008, for a review), raising the question of which processes critically characterise the effect. For example, in demonstrations of learned predictiveness there is often a high degree of conceptual similarity between scenarios. One possibility, therefore, is that the effect is governed by a simple heuristic arising from inferential reasoning. That is, it is possible that participants make the explicit assumption that the predictive utility of cues A - D will transfer across similar contexts (Mitchell, Griffiths, Seetoo, and Lovibond, 2012).

According to this explanation, learned predictiveness should be susceptible to manipulations of inferred beliefs. Indeed, Mitchell, et al., (2012) have provided evidence in support of this view. In their Experiment 2, inferences were directly manipulated across phases by way of instruction. At the onset of the second phase, participants in the continuity condition were explicitly instructed that the same cues would be relevant. Alternatively, those in the change condition were instructed the opposite, that previously predictive cues were now irrelevant. Critically, this condition revealed a complete reversal of the effect. That is, more was learned about the relationship between previously irrelevant cues and the novel outcomes. That learned predictiveness is sensitive to variations in explicit reasoning suggests a role for controlled, volitional attentional processes in explaining the effect.

However, there is evidence to suggest that the presence of the inference alone is not sufficient to produce the learned predictiveness effect. For example, Le Pelley et al. (2010a) investigated the expression of learned predictiveness adopting a procedure in which the critical relationships were embedded in text form. Interestingly, they failed to observe the effect; the attentional bias was only observed when the relevant information was presented in trial and error form across multiple trials. This is contrary to what would be expected if explicit causal attribution was the sole mechanism responsible for this bias. Similarly, related paradigms have found opposing influences of training and instruction on learned attentional responses (Le Pelley, Mitchell, \& Johnson, 2013). Taken together these findings raise the possibility that learned predictiveness reflects the operation of a combination of inferential and non-inferential processes.

As noted previously, learned predictiveness has taken an important role in theorising about learned attention. A common feature of such theories is the assumption that attentional changes are automatic in response to the
formation of associations between events (e.g., Kruschke, 2001; Le Pelley, 2004; Mackintosh, 1975; Pearce \& Mackintosh, 2010). According to this view, because associations between predictive cues and outcomes increase rapidly during phase 1 of a learned predictiveness experiment, these cues are automatically attended. Thus, previously predictive cues will capture attention at the start of phase 2, such that associations between these cues and novel outcomes are facilitated. Importantly, this process does not rely on a deliberate attempt by the individual to control attention in a biased fashion according to the nature of the phase 1 relationships.

While the results of Mitchell et al. (2012) appear to oppose this explanation, there is reason to suggest that their experimental design did not provide the conditions under which the presence of automatic processes could be adequately detected. For example, their demonstration relies on a definitive manipulation: Non-predictive cues were explicitly emphasised as important. If it is assumed that controlled attention is capable of modulating the expression of automatic processes, given the appropriate conditions, then it is possible that the manipulation was too strong, overriding the influence of automatic attention. Thus, although this manipulation demonstrates that learned predictiveness is susceptible to voluntary control via instruction, it does not test whether automatic processes also contribute to the effect under uninstructed conditions.

Further, the scenario employed, in which fictitious seeds grow different trees, potentially favours a more categorical inferential process whereby the outcome is most likely attributable to only one of the cues and not the other. This aspect of the design may have facilitated a complete reversal based on conceptual aspects of the scenario in addition to the manipulation of interest.

Therefore, the relative contribution of controlled and automatic processes to the learned predictiveness effect remains to be fully specified. The aim of the present experiments was to investigate this relationship.

## Experiment 1

Experiment 1 made use of the same instructional manipulation employed by Mitchell et al. (2012), albeit with a different cover scenario, in order to replicate their original result. The allergist scenario, employed in numerous demonstrations of learned predictiveness (e.g., Le Pelley \& McLaren, 2003) was used in which participants were asked to play the role of a doctor who must discover the allergies of a fictitious patient. The cues consisted of different foods, which predict the occurrence of various allergic reactions, serving as outcomes. At the start of phase 2, a new patient was introduced who consumed the same foods, but suffered novel reactions. As before, participants were required to discover which foods were leading to which reactions. The structure of the training phases is shown in Table 1 and reflects the standard learned predictiveness design. At the start of phase 2, one group of participants (the "same"
condition) were told that it was likely that both patients were allergic to the same foods, whereas those in the "change" condition were instructed that their two patients likely suffered from allergies to different foods.

In line with the findings of Mitchell et al. (2012), we anticipated that the bias in learning observed in learned predictiveness would proceed according to the instructions issued at the start of phase 2 training.

## Method

Participants Forty-eight University of Sydney students (27 female, 21 male; age $18-24$ ) participated in the experiment.

Apparatus and Stimuli All experiments were conducted on Apple Mac Mini computers attached to a $17-\mathrm{in}$. monitor, and programmed in PsychToolbox for Matlab (Brainard, 1997; Pelli, 1997). Foods were randomly allocated for each participant to serve as cues $\mathrm{A}-\mathrm{Z}$ in the experimental design, and consisted of: Coffee, Fish, Lemon, Cheese, Eggs, Garlic, Bread, and Peanuts. Similarly, four allergic reactions were randomly allocated to serve as the four outcomes, and were: Headache, Nausea, Rash, and Fever.

Procedure After being randomly allocated to either the same or change conditions, participants were instructed that their task was to learn which foods were causing which allergic reactions in a fictitious patient. They were told that on every trial, two foods that the patient had eaten would be presented. On being shown the foods, participants were required to predict which of two allergic reactions would occur.

Phase 1 consisted of the eight trial types shown in Table 1. Each of these was presented once in each of 16 blocks of trials. The order of trials was randomised across blocks. Each trial was followed by feedback stating whether their prediction was correct, as well as providing the actual allergic reaction experienced.

At the start of phase 2, participants were told that they now had a new patient and, as before, would be required to learn which foods were causing which allergic reactions. Those in the same condition were told that their new patient was allergic to the same foods as their previous patient, whereas those in the change condition were instructed that their new patient was allergic to different foods.

Phase 2 consisted of 16 blocks, each of which contained one of the four trial types shown in Table 1. As before, trial order was randomised within blocks and feedback was provided after each trial.

A test phase was administered immediately following phase 2. All cues were presented individually and in a randomised order throughout this phase. On each test trial, a cue would appear and participants were asked to indicate whether the cue had been paired with outcome 3 or outcome 4. This was done by making a rating on a linear analogue scale, labelled "Definitely goes with [outcome 3]" on the
left anchor, and "Definitely goes with [outcome 4]" on the right anchor.

Finally, a manipulation check was included to ensure that participants had remembered the instructions at the start of phase 2. Participants were presented with both sets of instructions and required to report which of those applied to their patient. There were no exclusions on the basis of this check.

## Results

Phase 1 For each block, accuracy was averaged across the eight compound trials to gauge acquisition. Accuracy increased consistently across training. A mixed-measures analysis of variance (ANOVA) with block (1-16) and condition (same vs. change) as factors revealed a significant main effect of block, $F(15,690)=40.1, p<.001$, but no significant effect of condition, $F<1$, and no block $\times$ group interaction, $F<1$, suggesting that the two groups learned at an equivalent rate in phase 1 .

Phase 2 A mixed-measures ANOVA examining phase 2 acquisition showed a significant effect of block, $F(15,690)$ $=44.13, p=<.001$, as well as a significant block $\times$ group interaction, $F(15,690)=2.1, p<.05$. The effect of condition did not reach significance, $F(1,46)=3.95, p=$ . 053 .


Figure 1. Learning scores for the same and change conditions for previously predictive and previously nonpredictive cues

Test data A learning score for each cue was calculated by combining accuracy for memory of the cue-outcome pairings in the test phase with the magnitude of the rating. This yielded a score out of 100 for each cue, with higher scores indicating better retention. Scores could range between 100 and -100 . Scores were averaged according to whether they were predictive ( $\mathrm{A}-\mathrm{D}$ ) or non-predictive (W -Z ) in phase 1 . These are shown for the same and change conditions in Figure 1.

Scores were subjected to a mixed-measures ANOVA with group (same vs. change) and cues (predictive vs. nonpredictive) as factors. Averaged over cue, there was no significant difference between the same and change conditions, $F<1$. Similarly, there was no effect of cue, $F<$ 1. However, as suggested by Figure 1, this resulted from a significant cue $\times$ group interaction, $F(1,46)=8.79, p<.05$.

This was further investigated with a simple effects analysis, which revealed that learning scores for predictive cues was higher than non-predictive cues in the same condition, $F(1,23)=11.51, p<.05$. The difference between predictive and non-predictive cues did not differ significantly in the change condition, $F<1$.

## Discussion

Our data provide a partial replication of Mitchell et al. (2012). While the same condition showed a standard learned predictiveness effect, this was abolished rather than reversed in the change condition. That is, there was no difference between previously predictive and previously non-predictive cues when participants were told that non-predictive cues were informative for the second phase.

Overall, a clear effect of instruction was observed making use of a scenario in which it is less likely that causal attribution is biased towards categorical reasoning. This suggests that the result of Mitchell et al. (2012) is not entirely a consequence of the conceptual structure of their scenario, further validating the influence of voluntary control on learned predictiveness.

However, it is important to note that our reversal was incomplete in the critical condition. On the basis of the current design, it is unclear why this should be the case. It is possible that the results from the change condition reflect competition between opposing inferential and automatic processes. While automatic processes would bias learning in favour of previously relevant cues, explicit inference favours irrelevant cues.

Alternatively, there may be added difficulty in the change condition. If more is learnt about the predictive cues in phase 1 , this means that they may be required in order to confirm the new object of attention, that is, the previously irrelevant cue. That is, if the explicit identity of the previously irrelevant cues is uncertain due to the fact that little learning has proceeded to these cues, then previously relevant cues may be actively used to guide responding. This is an additional process that is not necessary in the same condition.

Given that the reversal design does not allow the contribution of automatic processes to be assessed, Experiment 2 used an orthogonal manipulation of predictiveness in phase 1 and instruction to further test the relative contribution of voluntary and automatic processes.

## Experiment 2

Experiment 1 confirmed that learned predictiveness is susceptible to the manipulation of inferred beliefs. In Experiment 2, we aimed to further test the involvement of automatic processes. This was done by orthogonally manipulating the predictive status of cues in the first phase and the instructional manipulation. The design of Experiment 2 is shown in Table 2. The first phase of training was identical to that seen in Experiment 1. At the end of the initial training phase, all participants were told explicitly which foods the new patient was allergic to. However, two of those cues were previously predictive, while two were previously non-predictive. That is, they were told that the new patient was allergic to cues A and C , and X and Z .

This means that there were two cues ( A and C ) that were predictive in phase 1 , and known to cause allergies in the new patient, and two previously predictive cues ( B and D ) known not to be allergens. Similarly, of the previously nonpredictive cues, two ( Z and X ) were now known to cause allergies, and the remaining two ( Y and W ) known to be safe. The design therefore creates the condition in which an unambiguous instructional manipulation is present without removing the opportunity to observe an automatic influence of phase 1 training, if indeed it is present.

Table 2. Design of Experiment 2.

| Phase 1 | Phase 2 | Test |
| :--- | :--- | :--- |
| $\mathrm{AW}-\mathrm{O} 1$ | $\underline{\mathbf{A} Y-\mathrm{O} 3}$ | A |
| $\mathrm{AX}-\mathrm{O} 1$ | $\mathrm{~B} \underline{\mathbf{Z}}-\mathrm{O} 4$ | B |
| $\mathrm{BW}-\mathrm{O} 2$ | $\underline{\mathbf{C W}}-\mathrm{O} 5$ | C |
| $\mathrm{BX}-\mathrm{O} 2$ | $\mathrm{D} \underline{\mathbf{X}}-\mathrm{O} 6$ | D |
| $\mathrm{CY}-\mathrm{O} 1$ |  | W |
| $\mathrm{CZ}-\mathrm{O} 1$ |  | X |
| $\mathrm{DY}-\mathrm{O} 2$ |  | Y |
| $\mathrm{DY}-\mathrm{O} 2$ |  | Z |

Note. Letters indicate individual cues. Underlined letters indicate cues instructed as informative for phase 2. O1-O6 refer to six outcomes.

If, as suggested by the findings in Experiment 1, controlled processes are in operation, then a clear influence of instruction should be observed whereby more will be learned about cues $A, C, X$, and $Z$ in the second phase. However, if automatic attention favouring predictive cues is also present, then a difference should also be observed between instructed cues according to whether they were relevant ( A and C ) or irrelevant ( X and Z ) in the first phase. Given the advantage conferred by predictive utility, this predicts that more should be learned about A and C compared to X and Z .

## Method

Participants Participants comprised twenty-four University of Sydney students ( 20 female, 4 male; age $18-23$ ).

Apparatus and Stimuli Experimental stimuli remained the same as that employed in Experiment 1, with the exception that two additional allergic reactions were introduced to account for added outcomes in the design. These were Coughing and Sweating.

Procedure Phase 1 training and instructions remained identical to that used in Experiment 1. Following phase 1, participants were told that they were now observing the allergies of a new patient, but that they would be provided with a set of foods that the patient was allergic to. They were shown the names of four foods, corresponding to cues A, C, X, and Z and were informed that they would need to learn which of these corresponded to the various reactions that the patient was experiencing.

Given that foods were named explicitly, a shorter phase 2 with fewer trials per cue was employed. Participants completed four blocks, each block consisting of one of the four trial types shown in Table 2. On each trial, participants were now required to predict which of four allergic reactions would occur.

During test, each cue was displayed individually in random order. The four outcomes were displayed on screen and participants were asked to indicate which of these the cue had been paired with. This was followed by the appearance of a rating scale, asking how confident they were in their response. The left anchor was labelled "Not at all confident", and the right anchor labelled "Very confident".

Finally, the manipulation check required participants to report the instructed allergens of the second patient. Five participants were excluded, having failed to report this content, leaving 19 participants in the analysis.

## Results

Phase 1 Acquisition across blocks increased steadily for phase 1. A repeated-measures ANOVA showed a significant main effect of block on accuracy, $F(15,270)=13.01, p<$ . 01 .

Phase 2 Overall, accuracy increased during phase 2, resulting in a significant main effect of block on accuracy, $F(3,54)=12.95, p<.01$. However, acquisition varied according to whether a compound contained an instructed component that was previously predictive or an instructed component that was previously non-predictive, such that accuracy was significantly higher for the former (AY/CW higher than BZ/DX), $F(1,18)=7.25, p<.05$. The interaction was not significant, $F<1$.

Test data Accuracy scores, shown in Figure 2, were subjected to a repeated-measures ANOVA with
predictiveness (predictive vs. non-predictive) and instruction (instructed vs. ignored) as factors. This revealed a significant main effect of instruction, $F(1,18)=18.28, p<$ .01 , as well as a significant instruction $\times$ predictiveness interaction, $F(1,18)=10.6, p<.01$. The effect of predictiveness failed to reach significance, $F<1$.

A simple effects analysis investigating the interaction showed that for instructed cues, accuracy was significantly higher for previously predictive cues, $F(1,18)=5.7, p<$ .05. Interestingly, this was reversed for the remaining cues, such that accuracy was significantly higher for previously non-predictive cues, $F(1,18)=6.4, p<.05$.


Figure 2. Accuracy scores for previously predictive and previously non-predictive cues at test in Experiment 2 for the instructed and ignored conditions.

## Discussion

Consistent with the findings in Experiment 1, there was clearly an effect of instructional manipulation. However, the learned predictiveness effect was still evident amongst cues known to be allergenic. That is, more was learned about the previously predictive cues compared to previously nonpredictive cues, despite the explicit knowledge that both sets of cues were allergens. This is consistent with the involvement of automatic processes transferred from initial learning.

However, it is interesting to note that the opposite pattern emerged for cues that were not instructed as allergens, and would presumably be ignored by participants. Thus it appears that previously predictive cues were easier to ignore when known to be irrelevant. This may reflect a general benefit of prior predictive utility whereby attention is more easily directed either towards or away from stimuli in novel situations.

Alternatively, the difference in acquisition during phase 2 between compounds that contained instructed components that were previously predictive (AY and CW) and nonpredictive (BZ and DX) raises the possibility that some sort of automatic interference from phase 1 means that less is learned in general about phase 2 compounds in which participants have to attend to the previously non-predictive cue and ignore the previously predictive cue. If these
compounds were indeed more difficult to learn, despite explicit instruction, this would result in the observed lower accuracy for instructed, yet previously irrelevant cues at test.

## General Discussion

The experiments reported above suggest that a purely inferential account of learned predictiveness is insufficient to fully characterise the effect. However, it is clear that proposing an additive influence of inferential reasoning and automaticity is similarly inadequate as the results reported here suggest an interaction between the two.

For example, in phase 2 of Experiment 2, participants were given information that directly informed them which cues the patient was and was not allergic to. Even though participants could have ignored the non-causal cues completely, some learning of the cue-outcome relationships was evident. The result of interest regarding these noncausal cues was that previously predictive stimuli were learned about more poorly than previously nonpredictive stimuli. If the effects of the prior predictive history of the cues simply added or subtracted from selective attention in an automatic fashion then one would expect the opposite result for this incidental learning. That is, the predictive cues should be learned about more readily than the nonpredictive. This result suggests an interaction between control of attention and the effects of prior predictive history, which is not explained by either an inferential account nor the conventional associative account of learned predictiveness.

Accordingly, there are a growing number of studies that show that the learned predictiveness effect does not operate via the competitive associative algorithms of attentional change described by Mackintosh (1975; Le Pelley, 2004; Pearce \& Mackintosh, 2010). For instance, Le Pelley et al., (2010b) found that competition between cues in compound was not necessary for learned predictiveness to occur, and Livesey et al. (2011) found no evidence that direct comparison between predictive and nonpredictive cues affected the magnitude of learned predictiveness at all. The current study demonstrates another way in which the automatic allocation of attention appears to behave differently from model predictions. Although there appears to be a relatively automatic influence of the previous history of the cues, that influence only matches the predictions of associative learning theories for cues that are deliberately attended and not those that are deliberately ignored.

Clearly an important step in implementing attentional processes within models of human learning will require further investigations into the mechanisms responsible for biases in learning related to past predictive utility. Such biases remain to be fully specified with regards to how information is attended and ignored.

## References

Brainard, D. H. (1997). The Psychophysics Toolbox. Spatial Vision, 10, 433-436.

Kruschke, J. K. (2001). Toward a unified model of attention in associative learning. Journal of Mathematical Psychology, 45, 812-863.
Le Pelley, M. E. (2004). The role of associative history in models of associative learning: A selective review and a hybrid model. The Quarterly Journal of Experimental Psychology, 57(B), 193-243.
Le Pelley, M. E. (2010). Attention and human associative learning. In C. J. Mitchell \& M. E. Le Pelley (Eds.), Attention and associative learning: From brain to behaviour (pp. 187-215). Oxford: Oxford University Press.
Le Pelley, M. E., \& McLaren, I. P. L. (2003). Learned associability and associative change in human causal learning. Quarterly Journal of Experimental Psychology, 56B, 68-79.
Le Pelley, M. E., \& Mitchell, C. J., \& Johnson, A. M. (2013). Outcome Value Influences Attentional Biases in Human Associative Learning: Dissociable Effects of Training and Instruction. Journal of Experimental Psychology: Animal Behavior Processes, 39, 39 - 55.
Le Pelley, M. E., Reimers, S. J., Calvini, G., Spears, R., Beesley, T., \& Murphy, R. A. (2010a). Stereotype formation: Biased by association. Journal of Experimental Psychology: General, 139, 138-161.
Le Pelley, M. E., Turnbull, M. N., Reimers, S. J., Knipe, R. L., \& Murphy, R. A. (2010b). Learned predictiveness effects following single-cue training in humans. Learning \& Behavior, 38, 126-144.
Livesey, E. J., Thowart, A., De Fina, N. L., \& Harris, J. A. (2011). Comparing Learned Predictiveness Effects Within and Across Compound Discriminations. Journal of Experimental Psychology: Animal Behavior Processes, 37, 446-465.
Lochmann, T., \& Wills, A. J. (2003). Predictive history in an allergy prediction task Proceedings of EuroCogSci 03: The European Conference of the Cognitive Science Society (pp. 217-222). Mahwah, NJ: Erlbaum.
Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement.Psychological Review, 82, 276-298.
Mitchell, C. J., Griffiths, O., Seetoo, J., \& Lovibond, P. F. (2012). Attentional mechanisms in learned predictiveness. Journal of Experimental Psychology: Animal Behavior Processes, 38, 191-202.
Pashler, H. (1998). The psychology of attention. Cambridge, MA: MIT Press.
Pearce, J. M., \& Hall, G. (1980). A model for Pavlovian conditioning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. Psychological Review, 87, 532-552.
Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. Spatial Vision, 10, 437-442.
Wright, R. D., \& Ward, L. M. (2008). Orienting of attention. New York: Oxford University Press.

# Communication Leads to the Emergence of Sub-optimal Category Structures 

Catriona Silvey (C.A.Silvey@sms.ed.ac.uk), Simon Kirby, Kenny Smith<br>Language Evolution and Computation Research Unit, School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Dugald Stewart Building, 3 Charles Street, Edinburgh, EH8 9AD, UK


#### Abstract

Words divide the world into labeled categories. Languages vary in the categories they label, sometimes to the point of making cross-cutting divisions of the same space. Previous work suggests two opposing hypotheses about how communication contributes to category emergence: 1) these spaces lack an objective shared similarity structure, and communication dynamically creates one of a number of optimally shareable category structures; 2) the category structures resulting from communication are not necessarily optimal, but diverge from a shared similarity space in language-specific ways. We had participants categorize images drawn from a continuous space in two conditions: a) non-communicative, by similarity, b) communicative, dynamically creating categories when playing a partnered communication game. The memory demands of communication lead to reliance on salient images and early conventions, resulting in non-optimal category structures compared to non-communicative participants. This supports the hypothesis that communication leads to categories that diverge non-optimally from a shared similarity space.


Keywords: communication; category structure; category emergence; language evolution

## Introduction

Words divide the world into labeled categories. Languages vary in the categories they label, with some languages making coarser, finer, or even cross-cutting distinctions relative to how other languages carve up the same space (Bowerman \& Choi, 2001; Malt, Sloman, \& Gennari, 2003). Work is ongoing to quantify and classify this variation (Majid, Jordan, \& Dunn, in progress). The mechanism by which a set of labeled categories emerges in a given language is however unclear. One hypothesis is that at least for some domains (e.g. spatial relations, containers), there is no one perceptually obvious way to divide the space into categories: there are several potential ways an individual observer could draw category boundaries (Bowerman, 2000). Some researchers have built on this idea to suggest that the process of communication itself structures a previously unstructured space, making categories that are optimally shareable between communicators (Freyd, 1983; Markman \& Makin, 1998; Steels \& Belpaeme, 2005; Voiklis \& Corter, 2012). However, cross-linguistic work by Barbara Malt and colleagues on similarity perception versus labeling shows that, while the labeled categories of different languages do indeed diverge from each other, speakers of different languages still perceive the similarities between the objects in comparable ways (Malt, Sloman, Gennari, Shi, \& Wang, 1999). This suggests that the categorization systems of different languages can in fact superimpose a range of divergent structures on a space that has a shared underlying similarity structure. These two accounts suggest radically different roles for communication in the emergence of categories.

The current experiment contributes to this debate by investigating how humans categorize a set of images designed to have unclear category boundaries. The participants categorize the images in one of two conditions: a noncommunicative condition, where solo participants divide the images into categories according to similarity, and a communicative condition, where pairs of participants play a communication game with the images. The results shed light on the effect of communication on category structure, suggesting that the categories created by communication can and do diverge from a relatively shared similarity space, even in a stimulus set designed to have ambiguous boundaries.

## Method

Participants were assigned to two conditions. In the noncommunicative condition, participants divided a continuous space of images into labeled categories on the basis of similarity. In the communicative condition, pairs of participants played a communication game using the same continuous space of images. Participants in this condition produced labeled categories via the words they used to communicate each target image in the last two rounds of the experiment. The category systems the participants produced in the two conditions were then compared.

## Stimuli

The set of images used in the experiment is shown in Figure 1. The four corner images were generated using PsychoPy software (Peirce, 2007). For each image, a random number generator assigned x and y positions for the five vertices, and the resulting shape was drawn. Morphs between these images were then generated by shifting the vertices towards each of the corners, according to a weight defined by inverse Euclidean distance (Matthews, 2009), to create a total set of 25 images. The 'objective' Euclidean distance between the images in the space may of course not correspond to perceptual similarity (see, e.g., Smith \& Heise, 1992); however, in pilot experiments, participants showed variation in where they drew the category boundaries, making these stimuli suitable for the current study.

## Labels

To control for any effects on participants' categorizations arising purely from the use of labels (Lupyan, Rakison, \& McClelland, 2007), words to label the categories were provided in both the non-communicative and communicative conditions. Lists of 25 CVCV nonsense words were generated by combining consonants and vowels randomly selected


Figure 1: The stimuli used in the study (lines thickened for clarity).
from the whole alphabet (e.g., zipi, gisa, wada). Since we expected that participants would use known crossmodal associations between attributes of words and attributes of the images in assigning category labels (e.g. voiceless stops and spikiness, Nielsen \& Rendall, 2011), we assigned the same wordlist to a yoked triple of two non-communicative participants and one communicative pair, so that in the analyses, any peculiar effects of a particular wordlist would apply equally across the conditions.

## Participants

Participants were 42 students at the University of Edinburgh (30 female, median age 23). 20 took part in the noncommunicative condition. The non-communicative experiment took 15 minutes. Participants were paid $£ 2.22$ participants (randomly assigned into 11 pairs) took part in the communicative condition. The communicative experiment took an hour. Participants were paid $£ 7$, and each member of the pair with the highest communication score was awarded a $£ 10$ Amazon voucher. One pair failed to complete the experiment within an hour and so was excluded from analyses.

## Procedure

Non-Communicative Condition Participants were presented with a randomized onscreen array of all 25 images and a set of words to label categories. To avoid cueing the participants to produce a particular number of categories, only one word was initially shown on screen: participants could reveal new words at any time, and were told that a) they could use as few or as many words as they wanted, and b) they did not have to use all the words they had revealed. Participants could reveal a new word at any stage, up to 25 words. They were instructed to label similar images with the same word and different images with different words.

Communicative Condition Participants communicated via computer terminals in separate cubicles. In a communication trial, one participant was assigned as the sender and one as the receiver. The sender was presented with a randomized onscreen array of all 25 images, one of which was selected with a red box to indicate it was the target. The sender was also presented with one initial word. The sender could reveal a new word at any stage, up to 25 words. Any words they had revealed on a previous trial remained visible on their screen for all subsequent trials. The participant was instructed to choose a word that would help the receiver pick out the target from the array of images.

Once the sender had picked a word, the receiver was presented with a randomized onscreen array of all 25 images and the word the sender had chosen. The receiver was instructed to select the image the sender had wanted to communicate.

Once the receiver selected an image, both participants were presented with a feedback screen. The feedback screen showed the word the sender had used, the target image, the image the receiver had selected, the score for the trial, and the running score for the whole experiment. The score for each trial was calculated on the basis of the inverse Euclidean distance between the target and the image the receiver selected, from a minimum of 1 up to a maximum of 15 (for correctly picking the target).

After each communication trial the sender and the receiver swapped roles. The experiment consisted of 100 communication trials divided into 4 rounds. Each round featured the 25 images as targets in a randomized order. The randomized lists were balanced such that each participant was the sender for every target image once in the first half of the experiment, and once in the second half.

The first two rounds of the experiment were not incorporated into the categorization analysis, as it was expected that at this stage a system would still be emerging. Participants' categories were therefore taken from the last two rounds of the experiment. Success scores were taken from the whole experiment.

## Dependent Variables

Number of Categories The number of categories each participant produced was recorded.

Variation in Category Size To achieve a measure of variation in category size that took the number of categories into account (since more categories would generally contain fewer images each), the number of images in each category was divided by the expected number of images in each category, if images were distributed equally. For example, if a participant had 5 categories, an equal distribution would be to place 5 images in each category: if one of their categories in fact had 10 images, this would produce a value for that category of $10 / 5=2$. The range of these values was then taken as a measure of variation in category size adjusted for the number of categories (with a minimum value of 0 in the case of perfectly balanced categories).

Category Alignment Two measures were taken to compare participants' categories and quantify their alignment. The first, the Rand index (Rand, 1971), consists of a pairwise comparison of whether participants tended to place images in the same category or different categories. The calculation produces a value bounded from 0 to 1 , where 1 is perfect alignment. The second, V-Measure (Rosenberg \& Hirschberg, 2007), is based on variation of information between the groupings, normalized to compensate for differences in number of categories. This measure also ranges from 0 to 1 where 1 is perfect alignment. Two further measures, the Variation of Information measure on which V-Measure is based (Meilă, 2003) and an adjusted version of Cramer's phi (Wills \& Mclaren, 1998) were considered, but were found to produce incongruent results when applied to groupings with divergent numbers of categories. Since the variable of interest was participants' categories rather than the words they used, the alignment measures were taken irrespective of whether participants used the same words: i.e., if two participants put the same set of images in a labeled category but used different labels, they would count as fully aligned for this category.

## Hypotheses

For the non-communicative participants, there is no particular incentive to divide the images into more or fewer categories (beyond the minimal assumption that, in being asked to sort the images, the participants are unlikely to place them all in one category). This condition therefore functions as a baseline for assessing the variability of the participants' categorization of the images without communication. The expectation is that with no strong motivation to behave in any particular way, participants' categorization performance will vary.

For the communicative participants, the pressures on their emergent categorization systems are more complex. The only way to attain a perfect communication score with this stimulus space and scoring system is to have a unique label for each image, i.e. 25 words in total, with 25 corresponding categories containing one image each. However, participants' memory constraints will likely prevent this from happening in the experiment. More generally, then, for a given number of words, the optimal strategy is to apply each word to an equal number of images in the space, in a contiguous region (Gärdenfors, 2000). Participants who converge on a system like this would maximize their possible score across all rounds of communication. Figure 2A shows an example of this kind of optimal system. When the sender uses a word corresponding to one of the categories, the receiver can adopt the strategy of picking a central member of the category, thus ensuring their response is a maximum of 1.4 Euclidean distance units (or one diagonal step) from the target. Figure 2B shows, by contrast, a non-optimal system with the same number of categories. This system is non-optimal for two reasons. 1) The number of images in each category is less balanced (one category contains only two images, while another contains ten). This means that when the sender uses the word for the bigger category, the probability of the receiver selecting


Figure 2: A) An example of a category system optimally structured for communicative success. B) A non-optimal system with the same number of categories.
an image close to the target is lower. 2) The images belonging to some categories are spread across different regions of the space and do not form contiguous regions. This raises the probability of a receiver selecting an image some distance away from the target, even if she shares this set of categories with the sender. It is worth noting that the spaces we categorize in the real world may not have this kind of smooth continuous structure, and so the regular contiguous regions of Figure 2A may be more difficult to achieve. However, in the context of this experiment, if communication does give rise to optimally structured categories, this is the kind of system we would expect to see emerging.

## Results

A linear trend ANOVA found that communicative success increased over the 4 rounds of the experiment, $F(1,9)=18.66$, $p=.002$ (Figure 3). Participants' overall success was significantly above chance, $t(9)=4.21, p=.002$.

Participants in the communicative condition used significantly more labeled categories $(M=9.95, S D=3.98)$ than participants in the non-communicative condition $(M=6$, $S D=1.37$ ), Mann-Whitney $U=60, z=-3.54, p<.001$. Communicative participants also showed significantly more variance in how many labeled categories they used, Levene's test $(1,36)=16.47, p<.001$. Pairs who communicated together, however, showed no significant difference in the number of categories they used, $t(18)=-0.38, p=.7$, showing that this effect came from differences between, rather than within, communicative pairs. Thus, even though the noncommunicative participants had less motivation to converge on a particular number of labeled categories, they were more consistent in the number they produced than the communicative participants.

Participants in the communicative condition also varied significantly more in the size of their categories, when number of categories was taken into account (category size variation as described in Methods $M=1.54, S D=0.35$, compared to non-communicative participants, $M=1.17, S D=0.4$ ). That is, images were more unevenly distributed across categories in the communicative condition, $t(38)=3.13, p<$ .005. Surprisingly for the communication-as-alignment hypothesis, communicative pairs' groupings did not align sig-


Figure 3: Average communicative success over rounds in the experiment. Dotted line shows chance. Error bars show 95\% confidence intervals.
nificantly more than non-communicative participants' (bylanguage analysis: paired-samples $t$-test $0.47<t(9)<0.63$, $p>.5$, by-subjects analysis: independent $t$-test $-0.42<$ $t(18)<0.63, p>.4)$. Neither did communicative success correlate significantly with either of the alignment measures, $r<.51, p>.14$.

To test the hypothesis that communicative participants within a pair were more aligned than communicative participants who were not paired with each other, an analysis was run comparing the alignment scores for the true pairs with alignment scores for shuffled pairs (participant 2 paired with participant 3, etc.). A similar analysis was run for the non-communicative pairs, comparing alignment of those who shared the same wordlist with those who had different wordlists. Non-communicative participants displayed equivalent levels of alignment whether or not they used the same wordlist, $t(9)<0.8, p>.58$. For communicative participants, one of the alignment measures (Rand index) tended towards being significantly higher for participants who communicated in a pair than participants who did not, $t(9)=1.88$, $p=.093$, suggesting that communicative participants were marginally more aligned within-pair than between-pair in terms of which pairs of images they categorized together. For the second alignment measure, V-Measure, no significant difference was found, $t(9)=1.22, p>.25$.

## Discussion

The results are somewhat surprising for the hypothesis that communication creates optimal structure in previously variably structured spaces. Communicative participants produced categorizations that were generally non-optimal for maximizing communicative success, as defined in Hypotheses above. This is not merely a property of how humans perceive this particular space, as shown by the contrast with the non-communicative condition, where participants' categories


Figure 4: A) A typical non-communicative participant's categories. B) A typical communicative participant's categories.
were generally more balanced in size, carving up the space in a way that would actually be more optimal by this definition. Figure 4 shows a typical example of A) a non-communicative participant's categories and B) a communicative participant's categories. It is notable that several categories in B are also non-optimal in that they cover non-contiguous regions of the space (e.g. red and yellow categories). The heatmaps in Figures 5 and 6 show more generally how communicative participants' categories were more dispersed (Figure 6) compared to non-communicative participants, who tend to clump more around certain pairings or groups to form their categories (darker regions in Figure 5).

Why did communicative participants divide up the space so differently from non-communicative participants? As mentioned in Hypotheses above, the communicative task exerts a considerable memory demand on participants: although they are presented with the full image space on each trial, they still have to remember which word applies to which image or group of images over the course of the experiment. This exerts a pressure to create a system that is optimized not just for communicative success, but also for learnability.

Aids to learnability in this experiment might include particularly salient words, images, and pairings between them, or felicitous early successes that lead to the forming of conventions. These conventions, once established, may then prove too valuable to shift in favor of more optimally structured categories. Both of these aids to learnability (salient images/words and early successes) are mentioned by participants in the post-experiment questionnaire. Typically, when asked to draw the images they remember, participants could draw from memory two to five salient images and their associated words, but were unclear on other regions of the space. Thus the memory demands of the task, and the fact that participants have to establish a system from scratch, make the salience of individual images and early established conventions important factors determining the shape of each participant's final categorization system.

The possibility that different images in the set had differing salience is also supported by the success heatmaps in Figure 7. The heatmap in Figure 7A shows which target images led to higher success scores for participants. The pattern here is at odds with Figure 7B, which shows the relative expected


Figure 5: Heatmap visualizing how often non-communicative participants placed pairs of images in the same category. Darker areas indicate pairs more often categorized together.
chance levels of success for each image: images in the middle have more low-ED neighbors, so the probability of a higher score goes up when they are the target. The fact that panels A and B differ shows that participants' success with particular images is boosted by some other factor.

Panel C shows a heatmap of this boost - darker images are those whose overall communicative success rate is highest compared to what the chance-based map in panel B would predict. The likely explanation is that these images have higher salience for participants, making them act as Schelling points between sender and receiver. The striking finding that communicative success is not correlated with overall alignment could therefore be explained by participants consolidating success on a few images, leaving other areas of the space more sparsely covered.

While Figure 7 suggests that the salience of particular images may be shared across all pairs, early conventions are


Figure 6: Heatmap visualizing how often communicative participants placed pairs of images in the same category. Darker areas indicate pairs more often categorized together.
more likely to vary between pairs due to the randomized presentation of targets. This could explain the tendency towards higher pairwise (Rand index) alignment within pairs than between pairs, as reported in the Results. Despite the low levels of alignment overall, communicative pairs' language-specific early conventions may bring them more into agreement on how they categorize specific small groups of images.

As mentioned above, the pressures on the participants in the two conditions were substantially different: participants in the non-communicative condition interacted with the stimuli more briefly and without memory constraints, as well as lacking the pressure to create more categories imposed by the communicative task. Future work could investigate how participants divide up the space non-communicatively under the same time and memory constraints as the communicative participants, thus disentangling the effects of these constraints from the effects of communication. The noncommunicative condition in this study still serves as a useful

A


B


C


Figure 7: Heatmaps showing which target images produced higher per-round success scores. Darker images produced higher scores. A) Map of overall success per image in the experiment. B) Map of expected chance success rates per image. C) Difference between maps A and B. Darker images are those whose success rates are boosted highest beyond expected.
baseline, however, for participants' perceptually based divisions of the space.

The outcome of this study - that communication does not necessarily optimize category structures, but can create uneven and suboptimal structures compared to noncommunicators' division of the same space - is reflected in our experience of real language, where words vary widely in whether they specify small regions of semantic space or broad undifferentiated regions. The existence of the latter kind of word does not necessarily mean the users of the language do not perceive the differences between sub-parts of the region it covers: only that, for reasons of salience, or constraints imposed by the history and development of conventions in the language, these internal differences lack category labels. An important additional pressure in real language, not modeled in this study, is that different regions of semantic space may also have different functional importance, motivating coarser or finer-grained distinctions in different regions. However, these results show that even in the absence of functional reasons for uneven division of a space, communication can lead to the establishment of categories that may not align with noncommunicative similarity perception.

## Conclusion

Communication is not a simple process of mapping words onto pre-shared perceptual categories. Even if communicating partners agree on the underlying structure of the space they are talking about, the categories that emerge from communication can diverge in surprising ways, both from the underlying similarity space and from the category structure that would be most optimal for communicative success. Constraints on learning, salience effects, and the impact of early conventions on a language's development all contribute to shaping an emergent system of labeled categories.

## Acknowledgments

CS is supported by an AHRC PhD studentship. Thanks to Christos Christodoulopoulos for help with alignment measures and with networking for the communication experiment, and to Mark Atkinson and Andrea Ravignani for help with piloting.

## References

Bowerman, M. (2000). Where do children's word meanings come from? Rethinking the role of cognition in early semantic development. In L. Nucci, G. Saxe, \& E. Turiel (Eds.), Culture, thought and development (pp. 199-230). Mahwah, NJ: Lawrence Erlbaum.
Bowerman, M., \& Choi, S. (2001). Shaping meanings for language: Universal and language-specific in the acquisition of spatial semantic categories. In M. Bowerman \& S. C. Levinson (Eds.), Language acquisition and conceptual development (pp. 475-511). Cambridge: Cambridge University Press.
Freyd, J. (1983). Shareability: The social psychology of epistemology. Cognitive Science, 7(3), 191-210.

Gärdenfors, P. (2000). Conceptual spaces: The geometry of thought. Cambridge, MA: MIT Press.
Lupyan, G., Rakison, D. H., \& McClelland, J. L. (2007). Language is not just for talking: Redundant labels facilitate learning of novel categories. Psychological Science, 18(12), 1077-83.
Majid, A., Jordan, F., \& Dunn, M. (in progress). Evolution of semantic systems. http://www.mpi.nl/departments/other-research/research-consortia/eoss. (Online; accessed 22 April 2013)
Malt, B. C., Sloman, S. A., Gennari, S., Shi, M., \& Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. Journal of Memory and Language, 40(2), 230-262.
Malt, B. C., Sloman, S. A., \& Gennari, S. P. (2003). Universality and language specificity in object naming. Journal of Memory and Language, 49(1), 20-42.
Markman, A. B., \& Makin, V. S. (1998). Referential communication and category acquisition. Journal of Experimental Psychology: General, 127(4), 331-54.
Matthews, C. (2009). The emergence of categorization: Language transmission in an iterated learning model using a continuous meaning space. Unpublished master's thesis, University of Edinburgh.
Meilă, M. (2003). Comparing clusterings by the variation of information. In B. Schölkopf \& M. K. Warmuth (Eds.), Learning theory and kernel machines (pp. 173187). Berlin: Springer-Verlag.

Nielsen, A., \& Rendall, D. (2011). The sound of round: Evaluating the sound-symbolic role of consonants in the classic Takete-Maluma phenomenon. Canadian Journal of Experimental Psychology, 65(2), 115-24.
Peirce, J. W. (2007). PsychoPy-psychophysics software in Python. Journal of Neuroscience Methods, 162(1), 8-13.
Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, 66(336), 846-850.
Rosenberg, A., \& Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL) (pp. 410-420).
Smith, L. B., \& Heise, D. (1992). Perceptual similarity and conceptual structure. In B. Burns (Ed.), Percepts, concepts and categories (pp. 233-272). Amsterdam: Elsevier B.V.
Steels, L., \& Belpaeme, T. (2005). Coordinating perceptually grounded categories through language: A case study for colour. Behavioral and Brain Sciences, 28, 469-529.
Voiklis, J., \& Corter, J. E. (2012). Conventional wisdom: Negotiating conventions of reference enhances category learning. Cognitive Science, 36(4), 607-634.
Wills, A. J., \& Mclaren, I. P. L. (1998). Perceptual learning and free classification. The Quarterly Journal of Experimental Psychology, 51B(3), 235-270.

# Encoding time and allocation of attention in analogical development 

Nina Simms (ninasimms@ northwestern.edu)<br>Department of Psychology, 2029 Sheridan Road<br>Evanston, IL 60202 USA

Dedre Gentner (gentner@ northwestern.edu)<br>Department of Psychology, 2029 Sheridan Road<br>Evanston, IL 60202 USA


#### Abstract

The aim of the current studies was to explore encoding time differences in objects and relations and to investigate whether these differences lead to differences in allocation of attention to object similarity. Using a match-to-sample paradigm with 5 - to 6 -year-olds and adults, we found that (1) objects were encoded faster than relations for both adults and children, and that (2) children, but not adults, preferentially allocated attention to object similarity. Ultimately, these questions are aimed at identifying the factors responsible for the development of adult-like analogical reasoning. We suggest that changes in selective attention over development may account for the pattern of results seen across these two studies.


Keywords: analogical development; relational reasoning; selective attention; encoding time

## Introduction

Reasoning by analogy is a fundamental and powerful aspect of human cognition (Gentner, Holyoak \& Kokinov, 2001). Analogical reasoning is based on relational similarity. That is, two situations are analogous if they share a common relational structure (e.g., lava lamps and plate tectonics are both characterized by a system of convection); superficial commonalities like the perceptual features of the objects involved are generally irrelevant. However, reasoning on the basis of relational similarity is not trivial. Two analogous situations may share superficial commonalities that conflict with an alignment based on relational similarity. For example, to appreciate that $1: 3$ :: 3:9, one must understand that the relationship that holds between 1 and 3 is the same relationship that holds between 3 and 9 . Based on this shared relationship, the two smaller numbers in each proportion correspond $(1 \rightarrow 3)$ and the two larger numbers correspond $(3 \rightarrow 9)$. In this case, the identity match between the two 3s must be disregarded, since this correspondence (3 $\rightarrow 3$ ) is inconsistent with the overall relational match.

Although 1:3 :: 3:9 may not seem like a particularly challenging analogy for adults, instances where relational similarity conflicts with object similarity can be very challenging for young children (Gentner, 1988; Richland, Morrison \& Holyoak, 2006) In cases like these, children will often reason on the basis of object similarity rather than relational similarity. This tendency is referred to as the object bias, but over development (with age as well as experience), a relational shift occurs whereby children
become increasingly adept at reasoning on the basis of relational rather than object similarity (Gentner \& Rattermann, 1991).

For example, Gentner and Toupin (1986) gave 6-year-old children a simple story and asked them to reenact it with new characters. They performed well when the corresponding characters were highly similar between the two stories, but performed very badly when similar characters played different roles across the two stories (the cross-mapped condition). Further studies have corroborated this finding that when relational similarity is pitted against object similarity children tend to be highly influenced by object matches and less able to attend to relational matches. For example, Richland and colleagues (2006) found the same pattern of results in a picture-matching task. The pictures depicted the same event structure, and the task was to point out correspondences based on the event patterns. When the object matches were inconsistent with the relational match, younger children were greatly impeded in choosing the correct relational match. This pattern of results, in which object similarity disrupts young children's analogical reasoning, has been found repeatedly in a variety of analogical tasks (Gentner \& Rattermann, 1991) and even across cultures (Richland, Chan, Morrison, \& Au, 2010).

The object bias is a robust and well-documented phenomenon, but a clear understanding of why it occurs is still lacking. Most accounts of analogical development that address the object bias implicitly or explicitly appeal to some processing difference (or differences) between objects and relations to explain the bias. Some of these differences include representational complexity, familiarity or fluency, salience, and automaticity. Improvements in relational reasoning over development are then explained by changes in this processing difference, and/or by improvements in some additional capacity that tempers the effects of these processing differences. For example, accounts of analogical development that emphasize the role of relational knowledge suggest that children are familiar with more object concepts than relational concepts, but as children gain more relational knowledge, and as this knowledge becomes more fluent, they become able to focus on relational similarity (Gentner, 1988, 2003). Other accounts appeal to the idea that object similarity is more salient than relational similarity, and that improvements in inhibitory control over development allow children to combat the
influence of salient object similarity when reasoning about relations (Richland et al., 2006).

Without a clear understanding of the processing differences between objects and relations, it is difficult to precisely explain the object bias and to pinpoint what changes over development to decrease this bias. Thus, the goals of the present studies are to: (1) investigate one operationalization of processing differences in objects and relations, namely encoding time; (2) explore how encoding time differences impact analogical reasoning, in particular how it affects allocation of attention; and (3) examine how these patterns change over development.

Encoding time differences are a promising potential difference to investigate for a number of reasons. First, of the many proposed processing differences between objects and relations, faster encoding of objects than relations would be predicted by - or at least consistent with - nearly all of them. Second, prior research suggests that objects are encoded faster than relations by adults (Goldstone \& Medin, 1994; Sagi, Gentner \& Lovett, 2012), and it is likely that such a difference exists for young children as well. Finally, encoding time differences may have important consequences for how analogical reasoning unfolds.

During analogical reasoning, two representations are aligned so that elements from each are placed into correspondence with one another. According to StructureMapping Theory (SMT - Gentner, 1983; and modeled by SME - Falkenhainer, Forbus \& Gentner, 1989), alignment is an incremental, multi-stage process. The mapping process begins with individual identical elements from each representation - including features, objects, and relations being placed into correspondence with one another. These initial correspondences are promiscuous; individual elements may map to multiple other elements (e.g., both cat $\rightarrow$ cat and cat $\rightarrow$ boy). In successful analogical reasoning, the final one-to-one correspondences are based upon shared relational structure (e.g., chase $\rightarrow$ chase; therefore, catchaser $\rightarrow$ boy-chaser, mouse-fleer $\rightarrow$ cat-fleer).

We hypothesize that this incremental mapping process is interleaved with encoding, which is also incremental (Lovett, Gentner \& Sagi, 2009). Pieces of the representations become available at different times, and correspondences between analogues are forged as these pieces become available. Correspondences made early in the mapping process may be particularly influential during analogical reasoning. For example, they may guide attention for further encoding and mapping (Kubose et al., 2002). Early correspondences may also be privileged if initial, incomplete mappings, rather than a full alignment, are used to make a decision (i.e., "satisficing", which young children may be especially likely to do, cf. Thibaut, French, \& Vezneva, 2010). If objects are encoded faster than relations, then object correspondences should also be found earlier, resulting in attention initially being allocated to object similarity. With a "satisficing" strategy, earlier object correspondences would also lead to object-based (rather than relation-based) reasoning.

The present studies use a match-to-sample task to explore encoding time differences between objects and relations and the impact of such differences on the allocation of attention in children and adults. In Study 1 we ask whether objects are encoded more quickly than relations (at least for the stimuli used in these studies). In Study 2 we ask whether encoding time differences predict attention to object similarity. Integrating across these studies, we then ask what might change over development to yield this pattern of results.

## Study 1

If objects are encoded faster than relations, then participants should require less time to encode a stimulus in order to find an object match and more time in order to find a relational match. Thus, this study manipulated the amount of time participants were given to encode a sample stimulus before they were asked to find either the object match or the relational match (a methodology used by Sloutsky and Yarlas, as cited by Lovett et al., 2009).

## Method

Participants Thirty-two adults and 41 5- and 6-year-olds participated in this study. Adult participants came from the undergraduate subject pool and received partial course credit for their participation, or they were recruited from the university area and given monetary compensation. Children were recruited from an existing developmental research database and given a book and t-shirt for participating.

Stimuli and Procedures This study used a match-to-sample task administered on a touchscreen laptop. Stimuli consisted of three shapes arranged in one of three patterns: ABA, AAB, or BAA (Figure 1). On each trial, participants were shown a sample stimulus, which disappeared and was replaced by two choices. All participants completed two versions of the task: object-matching, in which they had to find the stimulus with the same shapes, and relationmatching, in which they had to find the stimulus with the same pattern. In both versions, the incorrect foil did not share objects or a relational pattern with the sample. Participants selected their choice by touching it on the screen.

Within each version of the matching task, there were three sections: practice, long-encoding-time (LET) test trials, and short-encoding-time (SET) test trials. The order of version and short and long trials was fully counterbalanced. In the practice sections, participants were shown an example triad and the matching criterion was explained ("Find the one with the same shapes/pattern"). Then, participants completed several practice trials with feedback to ensure that the task instructions were clear. Practice trials were followed by a block of LET test trials or SET test trials. LET trials displayed the sample stimulus for 1000 ms . For
adults, SET trials displayed the sample stimulus for 50 ms , and for children, 150 ms .

We hypothesize that for both adults and children in this study, objects will be encoded faster than relations. Therefore, we expect high accuracy for object matching on both LET and SET trials. In contrast, we expect high accuracy for relational matching only on LET trials, in which participants have had sufficient time to encode the relational pattern in the sample; SET trials should not provide enough time to encode the relational pattern, and therefore participants should not be able to reliably select the relational match in this case. In sum, we predict a Version by Encoding Time interaction, with a larger effect of Encoding Time for relation matching.


Figure 1: Example stimuli for Studies 1 and 2.

## Results and Discussion

Results for Study 1 are shown in Figure 2.
Adults' mean proportion correct was entered into a 2(Version) x 2(Encoding Time) repeated-measures ANOVA. Overall, adults were more accurate on objectmatching than relation-matching, $F(1,31)=63.07, p<.001$, and more accurate on LET than SET trials, $F(1,31)=51.27$, $p<.001$. However, these main effects are best interpreted in light of their interaction, $F(1,31)=25.27, p<.001$. As predicted, adults showed a larger decrement from shorter encoding time for relation-matching than for objectmatching.

Children's mean proportion correct was entered into a 2(Version) $x$ 2(Encoding Time) repeated-measures ANOVA. Like adults, children were overall more accurate on object-matching than relation-matching, $F(1,40)=$ 182.97, $p<.001$, and more accurate on LET than SET trials, $F(1,40)=60.03, p<.001$. The predicted Version x Encoding Time interaction was marginally significant, $F(1,40)=3.46, p<.10$. As with adults, short encoding times were more disruptive for relation-matching than for objectmatching.

These results support the hypothesis that for both children and adults, objects are encoded faster than relations. Although this may not be true in all cases, for the stimuli used in this task, both groups needed less time to encode the object information than the relational information. These findings echo prior research suggesting that adults encode objects more quickly than relations (Goldstone \& Medin, 1994). To our knowledge, this is the first time this processing difference has been shown for children as well.

What consequences might this difference have on analogical reasoning? Assuming incremental and interleaved encoding and mapping processes (Lovett et al., 2009), information encoded early (i.e., object information) could influence the allocation of attention during alignment. Specifically, early-available object information should initially direct attention toward object similarity. This would predict that conflicting object similarity should disrupt relational matching by diverting attention to the object match, potentially leading to more errors (i.e., selecting the object match instead of the relational match) and longer latencies to correctly select the relational match (because attention to the relational match should be delayed) (Sloutsky \& von Spiegel, 2004). However, for object matching, conflicting relational matches should not disrupt accuracy or response times (RTs). In Study 2, we explore these predictions.


Figure 2: Adults' and children's match-to-sample accuracy in Study 1 by task version and encoding time.

## Study 2

If object similarity captures attention as a result of early object encoding, then the presence of conflicting object similarity on a relation-matching task should result in more errors and slower correct RTs compared to cases without conflicting object similarity. In contrast, object matching should be largely unaffected by the presence of conflicting relational similarity. However, if this asymmetrical pattern is not seen, it would suggest that object similarity is not preferentially commanding attention, despite differences in encoding time.

## Method

Participants Thirty-two adults and 37 5- and 6-year-olds participated in this study. Participants were recruited and compensated as in Study 1.

Stimuli and Procedures This study used the same basic match-to-sample task used in Study 1, with some modifications. As in Study 1, stimuli consisted of three shapes arranged in one of three patterns: ABA, AAB, or BAA (Figure 1). Also as in Study 1, all participants completed an object-matching and relation-matching version, counterbalanced for order.

There were two primary differences between the tasks used in Study 1 and Study 2. First, the encoding time manipulation was removed. In Study 2, the sample was displayed for 1500 ms on all trials. Second, another type of trial - a conflict trial - was added. On conflict trials, the foil matched the sample on the non-relevant dimension. That is, on the object-matching version, the incorrect foil was a relational match, and on the relation-matching version, the foil was an object match. These trials were interspersed with no conflict trials, where the incorrect foil did not match the sample at all (the type used in Study 1) (Figure 1).

If object similarity preferentially captures attention - one possible consequence of faster object than relation encoding - we expect participants to make more errors and respond more slowly to conflict trials than no conflict trials on the relation-matching task. However, participants should not show this difference on the object-matching task. Alternatively, if object similarity does not capture attention (despite encoding time differences), we should not see this asymmetrical pattern. In sum, a Version by Trial Type interaction would suggest preferential attention to object similarity.

## Results and Discussion

Results for Study 2 are shown in Figure 3.
Adults' mean proportion correct and RTs on correct trials were entered into separate 2(Version) x 2(Trial Type) repeated-measures ANOVAs. Adults were significantly more accurate on the object-matching version than the relation-matching version, $F(1,31)=10.92, p<.01$. This main effect of Version was modulated by an interaction with Trial Type, $F(1,31)=4.31, p<.05$. Adults showed a small but reliable decrement in performance on the relationmatching task when a conflicting object match was present, but no such difference on the object-matching task.

Consistent with faster encoding of objects than relations found in Study 1, adults showed a main effect of Version in their RTs. They made significantly faster correct responses on the object-matching task than the relation-matching task, $F(1,31)=7.94, p<.01$. However, this did not interact with Trial Type. That is, on both the object- and relationmatching versions, adults were equally fast to respond to conflict and no-conflict trials.

Parallel analyses were carried out for children's accuracy and correct RTs. Children were significantly more accurate on the object-matching version than the relation-matching version, $F(1,36)=42.09, p<.001$, and significantly more accurate on no conflict than conflict trials, $F(1,36)=10.74$, $p<.01$. These factors also interacted, $F(1,36)=15.79, p<$


Figure 3: Adults' and children's accuracy and RT data from Study 2.
.001. The effect of conflict trials was larger for the relationmatching task than the object-matching task.

Children were also faster to respond correctly on the object-matching task than the relation-matching task, $F(1,36)=11.30, p<.01$. Trial Type did not significantly interact with Version, though the data do qualitatively follow a pattern consistent with the accuracy results. Specifically, compared to no conflict trials, conflict trials showed longer latencies on the relation-matching task than the object-matching task.

Though somewhat mixed, the pattern of accuracy and RT results in Study 2 suggest a difference in adults' and children's attention to object similarity. Children's performance resembles the pattern that would be predicted if early object encoding led to preferential to object similarity. Conflicting matches led to more errors and longer RTs on the relation-matching task than the objectmatching task.

In contrast, adults' performance suggests resilience against conflicting object similarity. Conflicting object
matches did not increase response latencies on the relational matching task, and the decrement in accuracy on relational conflict trials was quite small. Altogether, the results from Study 2 suggest that encoding time differences between objects and relations may lead to preferential attention to object similarity for children but not adults.

## General Discussion

The aim of the current studies was to explore encoding time differences in objects and relations and to investigate whether these differences lead to differences in allocation of attention to object similarity. Ultimately, these questions are aimed at identifying the factors responsible for the development of adult-like analogical reasoning.

In Study 1, we found that both children and adults encoded objects faster than relations. Study 2 found a pattern of object- and relation-matching that suggested children, more than adults, preferentially allocated attention to object similarity (but see Sloutsky \& von Spiegel, 2004). Thus, we see continuity in a basic processing difference
between objects and relations, but the consequences of this difference for analogical reasoning behavior changed over development. Together, the results of Studies 1 and 2 implicate a change in an additional factor as key in explaining the difference in children's and adults' susceptibility to conflicting object similarity.

We think our results are best explained by a change in selective attention over development. Although adults encoded objects faster than relations, object similarity did not capture their attention when relational similarity was the relevant matching criterion. Thus, it seems that adults were able to selectively attend to the relational information, despite the earlier availability of object information, whereas children were not. Further work is needed to build a strong case for the role of selective attention in analogical development. However, this account is consistent with other studies of analogical reasoning in children (e.g., Thibaut, et al., 2010).

Future research will also need to explore how selective attention interacts with other factors implicated in analogical development. For example relational language has been proposed to aid relational thinking in a number of ways (Gentner, 2010). Performance on difficult analogical reasoning tasks can often be improved by providing children with language to describe the relevant relations (e.g., Loewenstein \& Gentner, 2005). With regards to selective attention, an intriguing possibility is that relational language may strengthen top-down control of attention to relational information.

## Acknowledgments

This research was funded by SILC (the Spatial Intelligence and Learning Center), an NSF Science of Learning Center, grant SBE-0541957. The authors would like to thank David Uttal, Steve Franconeri, and the Cognition and Language Group for invaluable feedback on this project. Many thanks also go to Laura Willig, Ashley Poltermann, and the RAs of the Project on Children's Thinking for their assistance in conducting this research.

## References

Gentner, D. (2010). Bootstrapping children's learning: Analogical processes and symbol systems. Cognitive Science, 34, 752-775.
Gentner, D., Holyoak, K. J., \& Kokinov, B. N. (Eds.). (2001). The analogical mind: Perspectives from cognitive science. Cambridge, MA: MIT Press.
Gentner, D., \& Rattermann, M. J. (1991). Language and the career of similarity. In S. A. Gelman \& J. P. Byrnes (Eds.), Perspectives on thought and language: Interrelations in development. London: Cambridge University Press.
Goldstone, R. L., \& Medin, D. L. (1994). Similarity, interactive activation, and mapping. Journal of Experimental Psychology: Learning Memory and Cognition, 20, 3-28.

Gentner, D., \& Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. Cognitive Science, 10, 277-300.
Kubose T .T., Holyoak K .J., Hummel J. E. (2002). The role of textual coherence in incremental analogical mapping. Journal of Memory and Language, 47, 407-435.
Loewenstein, J. \& Gentner, D. (2005). Relational language and the development of relational mapping. Cognitive Psychology, 50, 315-353.
Lovett, A., Gentner, D., Forbus, K., \& Sagi, E. (2009). Using analogical mapping to simulate time-course phenomena in perceptual similarity. Cognitive Systems Research, 10, 216-228.
Richland, L. E., Chan, T. K., Morrison, R. G., \& Au, T. K. F. (2010). Young children's analogical reasoning across cultures: Similarities and differences. Journal of Experimental Child Psychology, 105, 146-153.
Richland, L.E., Morrison, R.G., \& Holyoak, K.J. (2006) Children's development of analogical reasoning: Insights from scene analogy problems. Journal of Experimental Child Psychology, 94, 249-273.
Sagi, E., Gentner, D. \& Lovett, A. (2012). What difference reveals about similarity. Cognitive Science, 36, 10191050.

Sloutsky, V. M. \& von Spiegel, J. (2004). Automatic processing of elements interferes with processing of relations. In K. Forbus, D. Gentner, \& T. Regier (Eds.), Proceedings of the XXVI Annual Conference of the Cognitive Science Society. Mahwah, NJ: Erlbaum.
Thibaut J., French R., \& Vezneva M. (2010). The development of analogy making in children: Cognitive load and executive functions. Journal of Experimental Child Psychology, 106, 1-19.

# Parent-Child Screen Media Co-Viewing: Influences on Toddlers' Word Learning and Retention 

Clare E. Sims (Clare.Holtpatrick@Colorado.Edu)<br>Department of Psychology \& Neuroscience, 345 UCB<br>Boulder, CO 80309-0345 USA

Eliana Colunga (Eliana.Colunga@Colorado.Edu)<br>Department of Psychology \& Neuroscience, 345 UCB<br>Boulder, CO 80309-0345 USA


#### Abstract

Screen media, such as television and videos, are a common part of young children's lives. Yet infants and toddlers have been shown to learn less effectively from screens than from interactions with another person. Using a quasi-experimental design we explored how social factors of screen media coviewing impact children's learning outcomes. We observed parents co-viewing a novel word training video with their children, then tested children for immediate and delayed word learning. We then investigated the links between parental speech during co-viewing and children's subsequent word learning. Parental speech that encouraged children to produce the novel words predicted better retention of word learning, whereas speech that focused more on the video itself rather than the content was negatively associated with learning.


Keywords: Screen media; co-viewing; word learning.

## Introduction

Screen media, such as television, videos, computers, and hand-held devices like smartphones and tablets, are an increasingly common part of young children's lives. Although the American Academy of Pediatrics recommends that children under the age of two not watch any screen media (AAP, 2010), survey data reveal that $43 \%$ of children in this age group watch TV every day (Rideout \& Hamel, 2006). This may be partly due to the recent proliferation of screen media content aimed specifically at infants and toddlers (e.g., Baby Einstein). Much of this content is presented with educational claims, including in the domain of language development (Fenstermacher et al., 2010). However, research that actually compares children's learning from screen media to their learning from face-toface interactions with another person reveals a video deficit effect (Anderson \& Pempek, 2005). That is, across varied experimental paradigms, infants and toddlers learn less effectively from a screen than from a person.

Is it possible to effectively facilitate infants' and toddlers' learning from screen media? Factors that may impact this learning include the complexity of the information on screen, whether key information is highlighted or repeated, or the presence of social cues that help children direct attention to what is important. In the work presented here we employed a quasi-experimental approach to investigate the role of a common social context for screen mediated learning: co-viewing a video with a parent. We explored the
relationship between parental behavior during active coviewing and children's learning outcomes. The results suggest meaningful links between characteristics of the screen media viewing context and children's learning of video content. Exploring the attributes and effects of coviewing is vital for guiding further research on children's screen mediated learning and applying this work in practice. In the remainder of the introduction we will briefly review the evidence for the video deficit effect in the domain of language as well as the existing research on co-viewing with young children. Many questions on this topic remain, and our approach of linking co-viewing to word learning in the moment contributes to this emerging literature.

## The Video Deficit in Language Learning

From birth, infants are constantly exposed to language, both from live and mediated sources. By manipulating whether information to be learned is presented through screen media or through face-to-face interactions, researchers have identified a video deficit effect in the domain of language development-infants and toddlers more effectively learn language from a person than from screen media. The video deficit has been demonstrated in two types of tasks in this domain: phoneme distinction and word learning.

One study took advantage of a change in phoneme discrimination that occurs in infancy. At six months of age infants can discriminate phonetic speech contrasts from their native language as well as from a non-native language to which they have never been exposed (e.g., Werker, Gilbert, Humphrey, \& Tees, 1981). However, over the first year of life, infants become attuned to distinctions between speech sounds in their native language, leading to an inability to recognize speech contrasts that do not exist in that language. Kuhl, Tsao, and Liu (2003) tested what kinds of exposure to non-native speech distinctions can prolong 9-month-old infants' ability to perceive those distinctions. These researchers found a video deficit in maintaining phonetic distinctions: infants exposed to non-native speech on video lost their sensitivity to non-native speech distinctions, whereas those exposed to live speech were still able to discriminate between non-native phonemes. In fact, infants in the video exposure group performed equivalently to a control group with no exposure; screen mediated speech led to the same outcome as no speech at all.

Another major developmental step in language acquisition is learning words. Researchers have looked at how young children learn unfamiliar or novel words from various live and screen mediated sources. In a typical word-learning task, children are shown an object and told a label for that object. To test for learning, the child is presented with an array of objects, including the trained object and distractor objects, and is asked to identify the object corresponding to the trained label. One study found a U-shaped developmental progression of the video deficit in word learning: while children between 13 and 20 months of age showed a significant deficit, younger children (aged 6 to 12 months) and older children (aged 21 to 24 months) did not show a deficit (Krcmar, 2010). Another word learning study found evidence for a video deficit among children between 15 and 24 months of age, which was actually stronger after children were exposed to a commercial video compared to a lab-created video (Krcmar, Grela, \& Lin, 2007). While the studies mentioned so far were conducted in a lab setting, another study confirmed the video deficit in word learning among 12- to 18 -month-olds exposed to a commercial video in their own homes (DeLoache et al., 2010). One study using a slightly older age group did not find any evidence for a video deficit in a word learning task among 30-montholds (O'Doherty et al., 2011). However, recent work in our lab suggests that the video deficit may persist depending on how word learning is assessed. We found that 30 - to $36-$ month-olds learned and retained one-to-one word-referent mappings from a screen, but the same children showed a video deficit in their retention of lexical categories (Sims \& Colunga, in preparation). When toddlers must generalize what they previously learned from a screen, inferring categories based on the words they learned, screen mediated learning still seems to be at a disadvantage.

In sum, studies that compare language learning from screen media to learning from a person show an early emerging video deficit effect that diminishes with age and depending on the type of task used. But why do children struggle with screen mediated learning when they can easily learn the same information in person? Some have proposed that social factors of screen media, or lack thereof, drive the video deficit effect (Richert, Robb, \& Smith, 2011). The screen mediated environment typically lacks the kind of rich social interactions and contingencies that children get when learning directly from a person. The video deficit may be rooted in the socially impoverished nature of screen media itself, particularly in the fundamentally social domain of language learning. Therefore, the social context of screen mediated learning is the focus of a related area of research, including work on co-viewing.

## Screen Media Co-Viewing

Most parents watch TV with their child either all or most of the time (Rideout \& Hamel, 2006). Parental co-viewing may provide a social context for children's screen media use. By actively co-viewing with their children, parents have opportunities to scaffold children's learning from the screen.

Most of the work linking parent-child co-viewing and children's learning from a screen has looked at slightly older age groups than studies of the video deficit effect. However, more recent research is beginning to explore how coviewing impacts infants' and toddlers' attention to and learning from screen media.

Some studies of preschoolers have experimentally tested how different types of co-viewing interactions influence children's subsequent learning. For example, one set of studies tested co-viewing in a context familiar to many young children: Sesame Street episodes. In one study, Reiser, Tessmer, and Phelps (1984) manipulated whether or not adults asked content-specific questions and provided feedback and encouragement while co-viewing segments teaching letters and numbers. Children learned the content of the video more effectively in the experimental condition compared to the control condition, in which adults did not provide any commentary. Reiser, Williamson, and Suzuki (1988) added to this result by showing that asking questions, with or without providing feedback, resulted in better learning than simply directing children's attention to the screen when educational content was being shown. Together these studies show that adult commentary and questions during co-viewing can directly facilitate children's learning from real screen media content.

Fewer studies have linked co-viewing interactions to outcome measures among infants and toddlers. Some studies have linked qualities of parental behavior or parentchild interactions to a precursor for children's screen mediated learning: attention to the screen. Results show that parents' eye-gaze to a screen modulated 12- to 21-montholds' attention to screen media (Demers, Hanson, Kirkorian, Pempek, \& Anderson, 2012). Further, sensitive and reciprocal parent-child interactions during co-viewing predicted 6- to 18 -month-olds' looking time to an infantdirected video (Fidler, Zack, \& Barr, 2010). Another study classified parental interaction styles based on co-viewing behaviors (Barr, Zack, Garcia, \& Muentener, 2008). Parents were classified into different levels of scaffolding, and these clusters of co-viewing behaviors predicted looking time and responsiveness to infant-directed videos among 12-, 15-, and 18 -month-old infants. The high-scaffolding parents tended to use verbalizations that oriented their children to the video and focused on the content therein. Together these studies show that parents who used eye gaze, high-quality, responsive interactions, and content-focused verbalizations were most effective in establishing joint attention to the screen and getting their children actively involved in coviewing.

Only one study that we are aware of has started to link observations of co-viewing to measures of learning specifically among this younger age group. In this study, parents co-viewed a video with their 12 - to 25 -month-old children that was intended to teach words (Fender, Richert, Robb, \& Wartella, 2010). The authors observed parental coviewing behaviors as well as child verbalizations, and, importantly, also asked parents which of the words in the
video their child was unfamiliar with. In this way, the authors measured learning by observing how often children produced words that they had been unfamiliar with prior to seeing the video. Parents tended to cluster into different groups depending on how much their co-viewing behavior was focused on teaching the words in the video to their children. Children produced more words that they were previously unfamiliar with when their parents had a higher teaching focus during co-viewing. Further analyses showed that these parents tended to focus specifically on the words that they knew their children were unfamiliar with. This result is particularly interesting in light of the literature on the video deficit effect in word learning. This study shows that sensitive parental scaffolding during co-viewing with infants and toddlers may be able to reduce the video deficit.

## Rationale and Predictions

In the current study, we investigated the link between parental co-viewing behavior and toddlers' word learning. Parents and their $2 \frac{1}{2}$ - to 3 -year-old children watched a video of a person teaching novel words for novel objects. Children were subsequently tested on their word learning, both immediately after watching the video and after a week-long delay. Different kinds of parental speech during co-viewing were coded, analyzed using principal component analysis, and examined as predictors of children's word learning.

This approach offers several contributions to the literature. First, by training and testing children on novel words for novel objects, we controlled for any prior knowledge or exposure children may have had to the content of the task. Second, by testing children on the content of the video, both immediately and after a delay, we were able to make direct links between parental speech during co-viewing and children's word learning and retention. Third, we included an older age group compared to most studies of the video deficit in word learning to see what behaviors impact learning once children are becoming better able to learn from a screen. Fourth, this work uses an interdisciplinary approach, drawing on methodologies from the fields of linguistics and psychology. The results will provide a first step in identifying specific co-viewing behaviors that facilitate, or possibly inhibit, word learning and retention in young children.

Based on the work reviewed above, the extent to which parents focus their speech on the key information to be learned on screen should predict children's word learning performance. Parents who are responsive and help their children focus on the content of the video should promote their children's attention to and learning from the screen. In this study that means talking about specific objects shown and novel words presented in the video. Further, parents who focus more on the novel labels being taught in the task should help children better learn the correct word-object mappings. An emphasis on the novel labels concurrent with the presentation of objects in the video should help children establish and retain these mappings.


Figure 1. a. Novel target objects taught to children in the word learning task. b. Example word learning trial.

## Method

## Participants

Fifty children were recruited for participation from the Boulder, CO area. Six subjects were excluded from analyses due to missing or inadequate co-viewing data (two children) or co-viewing speech being primarily in a language other than English (four children). Therefore, the final sample included here consisted of 44 children ( $M_{\text {age }}=32.1 \mathrm{mo}$., $S D$ $=1.3$ mo., 25 girls).

## Materials

Children were taught six novel words (elg, ife, nork, gub, $z e b$, and $l u g$ ) for six novel objects (see Figure 1a). The novel objects and words were presented in a lab-made training video of a research assistant whom the children did not meet during the study. In the video, the assistant presents one novel object at a time, placing it on a table in front of her and rotating it as she speaks. Addressing the camera directly, she labels the object in three ways: "This is a/an __. Do you see the __? This is my __." After an object is labeled, the video cuts to a 3 second still close-up image of the object. Each object is presented two times each, and thus labeled six times total, for a total video duration of 2 minutes and 50 seconds.

Two standardized vocabulary measures were used to assess language development. Children were tested on the Peabody Picture Vocabulary Test 4 (PPVT; Dunn \& Dunn, 2007), a test of receptive vocabulary. Parents were given the MacArthur-Bates Communicative Development Inventory (CDI-III; Fenson et al., 2007), a checklist on which they indicated words their children knew. Parents also completed a survey on their children's screen media use at home.

Parent-child interactions during training video co-viewing were recorded and later transcribed in ELAN (Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands: http://tla.mpi.nl/tools/tlatools/elan/; Brugman \& Russel, 2004). Parental verbalizations were subsequently coded using a scheme adapted from Barr et al. (2008). We used codes from this study and also developed codes specifically related to labeling in our task (see Results section for detailed description of the included codes).

Table 1: Factor loading values for the four components resulting from the factor analysis.

|  | Describing <br> Objects |  <br> Feedback | Narrating | Open-Ended Questions vs. <br> Explicit Labeling |
| :--- | :---: | :---: | :---: | :---: |
| Tag questions | $\mathbf{. 6 6 4}$ | .389 | -.057 | .125 |
| Descriptions | $\mathbf{. 8 5 3}$ | -.341 | -.060 | -.040 |
| Label elicitation questions | .189 | .777 | .024 | -.066 |
| Confirmations | -.270 | .739 | -.112 | .007 |
| Evaluations | .141 | -.146 | $\mathbf{8 3 1}$ | -.061 |
| Interactive verbalizations | -.245 | .020 | -.76 | $\mathbf{- 7 7 4}$ |
| Wh- questions | -.411 | -.199 | -.183 | $\mathbf{9 0 3}$ |
| Labels | -.179 | -.190 | -.198 |  |

## Procedure

Training At the beginning of their first visit to the lab, parents and children watched the novel word training video. Parents were encouraged to actively co-view the video with their children. The experimenter explained that the video was meant to teach some new words and that the parent could teach their child about the words in the video as they would while watching at home. The experimenter left the room for the remainder of training, and parent-child coviewing was videotaped for later analysis. When the training video had ended, the experimenter re-entered the room.
Testing Children were tested both immediately after watching the training video and again a week later. To test word learning, children were presented with pairs of trained novel objects and asked to identify one by name (e.g., "Which one is an elg?"; see Figure 1b). Each object was asked for once, for a total of six testing trials. At the end of their first visit to the lab children were given the PPVT vocabulary test.

Children were given the same word learning test at their second visit. Importantly, children were not re-trained on any of the novel words or objects at this time. Therefore, the delayed testing session captured retention of the novel words.

## Results

The first question to assess was how well children learned and retained the novel word-object mappings. Children's proportions of correct target object choices at each testing session were first compared to chance performance. Children were accurate at above-chance levels both at immediate $(M=.57, S D=.22, t(43)=2.20, p=.03)$ and delayed testing $(M=.58, S D=.20, t(43)=2.58, p=.01)$. Next, a paired t-test showed that accuracy did not differ between visits $(t<1, p>.05)$, confirming that children retained the word-object mappings they had learned initially. It is worth noting that accuracy performance at the group level was not particularly high, and yet varied a fair amount across individual subjects. This may suggest that something about the learning context of individual children influenced their accuracy in the task.

The next question was how parent speech during coviewing related to children's word learning performance. We began with two common measures of the quantity of parental speech: total word count and mean length of
utterance (MLU). These variables were entered as predictors of children's word learning outcomes at each visit in two multiple regression analyses. These measures of parental speech quantity did not explain a significant proportion of the variance in either immediate or delayed child word learning performance $\left(R^{2}<.18, F(2,43)<1, p>.05\right.$ for both models). Further, neither total word count nor MLU were significant predictors of word learning outcome at either visit $(t(41)<1.50, p>.05$ for all coefficients). Because these parent speech quantity variables did not predict learning outcome, we next explored the quality of parental speech during co-viewing.

First, to reduce the dimensionality of the qualitatively coded parent speech data, we conducted a principal component analysis (PCA). After removing several codes that were used by very few parents in the sample, eight coding categories were entered into the PCA and resulted in four components with Eigenvalues above 1.0. The first component explained $19.47 \%$ of the variance, the second $18.94 \%$, the third $18.74 \%$, and the fourth $18.01 \%$ for a total explained variance of $75.15 \%$. An orthogonal Varimax rotation was used to facilitate interpretation of the components.

The four components are shown in Table 1 with factor loadings on each included coding category. The first component included tag questions and descriptions of the items shown in the video. Tag questions, which are statements with a question appended at the end, were typically used by parents to talk about the items on screen. Many of the observed instances of tag questions were also descriptions (e.g., "it's a green lug, huh?"). This component will be referred to as describing objects because it captures parents' focus on the individual objects depicted on the screen. The next component includes label elicitation questions and confirmations, which also appeared together often in co-viewing speech. This component will be referred to as label elicitation and feedback because it captures how often parents explicitly asked children to produce labels, including giving positive feedback for doing so. The third component includes evaluations and interactive verbalizations. Parents often used evaluations to make general comments about the video or item shown on screen (e.g., "that's a cool one"), and interactive verbalizations were comments about the video itself (e.g., "it says let's see those again"). This component will be called narrating because it captures parental speech about the video and

Table 2: Multiple regression output for the novel word learning task at immediate and delayed testing.

| Predictors | Immediate Word Learning |  |  | Delayed Word Learning |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Unstandardized Coefficients |  | Standardized Coefficients | Unstandardized Coefficients |  | Standardized Coefficients |
|  | B | SE | Beta | B | SE | Beta |
| Age | -. 017 | . 030 | -. 102 | . 001 | . 025 | . 010 |
| Gender | -. 091 | . 075 | -. 209 | . 051 | . 063 | . 130 |
| Vocabulary Percentile | -. 001 | . 001 | -. 076 | . 001 | . 001 | . 169 |
| Screen Time | . 000 | . 001 | . 077 | . 000 | . 001 | -. 104 |
| PCA Components |  |  |  |  |  |  |
| Describing Objects | . 021 | . 036 | . 096 | -. 020 | . 031 | -. 101 |
| Label Elicitation \& Feedback | . 025 | . 038 | . 114 | . 069 | . 032 | .335* |
| Narrating | . 005 | . 036 | . 022 | -. 052 | . 031 | $-.269^{+}$ |
| O-E Questions v Explicit Labeling | . 046 | . 035 | . 211 | . 015 | . 030 | . 079 |

Note. ${ }^{\dagger} p<.10 .{ }^{*} p \leq .05$.
screen viewing context more broadly. The fourth and final component includes wh- questions and explicit labeling. These codes loaded in opposite directions onto this component, so the component will be referred to as openended questions vs. explicit labeling. In the positive direction this component captures the extent to which parents provided the novel labels being taught on screen, and in the negative direction it captures the extent to which parents asked their children open-ended questions about the video (e.g., "what is that?").

Hierarchical multiple regression analyses were conducted on children's word learning outcomes at each visit. In each analysis, the independent variables were entered in two blocks. Demographic and standardized test variables were entered in the first block. These included child age in months, child gender (dummy coded), vocabulary percentile score (averaged over the CDI-III and PPVT), and average screen use per day in minutes. The second block included the four co-viewing components from the factor analysis. This method of analysis allowed for evaluating the predictive value of the co-viewing components over and above the other included variables.

Table 2 displays unstandardized and standardized regression coefficients after entry of both independent variable blocks for immediate and delayed testing. None of the included variables significantly predicted immediate word learning accuracy $\left(R^{2}=.37, F(4,43)=.68, p>.05\right.$ after entry of all independent variable blocks). Although the full model did not reach significance $\left(R^{2}=.48, F(4,43)=\right.$ 1.31, $p=.27$ ), delayed word learning accuracy was predicted by two of the independent variables entered into the regression.

Parental label elicitation and feedback was a positive predictor of learning, indicating that the more parents asked children to produce novel labels and provided positive feedback, the better children retained word-object mappings. Specifically, the model predicts that for every standard deviation increase in parents' use of label elicitations and feedback, children's delayed word learning performance should increase by 0.34 standard deviations. On the other hand, narrating was a marginally negative predictor of retained novel word learning. This suggests that parental speech that was focused on the video itself and the content only in a general way actually inhibited children's correct retention of novel word-object mappings in this task. Specifically, the model predicts that for every standard
deviation increase in parents' use of narrating speech, children should be 0.27 standard deviations worse at retaining the novel word-object mappings.

## Discussion

The results of the current study show that although toddlers as a group learned novel words from a video, certain aspects of parental speech were associated with differences in this learning. An exploratory factor analysis revealed several variables of parent speech quality that characterized the coviewing linguistic environment and that predicted children's learning outcomes. Of note, the quality of parent speech only predicted children's retention of learning. This suggests that toddlers' immediate word learning from screen media may be relatively robust, and less influenced by the co-viewing environment. Yet the retention of this information may be sensitive to parental intervention.

The first key result showed that the extent to which parents elicited labels from their children and provided feedback while watching the training video predicted children's retention of word-object mappings. Although the finding that labels during co-viewing facilitated word learning is in line with prior work and with our predictions, the specific form of this labeling is informative. Children's retention was predicted not by hearing parents label the items on screen, but by parents cuing children to produce the labels themselves and providing responsive feedback.

Another key result was that the extent of parents' narrating during co-viewing was negatively associated with children's retention of word learning. This shows that parent speech about the video itself or general, non-specific speech about what is shown on screen is not conducive to novel word retention. This result suggests a negative impact on learning due to focusing on the form rather than the content of screen media. Further, this also suggests that general evaluative speech about the content on screen is not much more informative than talking about the video itself, and both of these together may actually impede learning outcomes.

Together these results are consistent with prior work indicating that responsive behavior during co-viewing promotes children's attention to and learning from screen media. The results build on prior work by demonstrating specific co-viewing behaviors that are responsive and thus scaffold children's learning, as well as behaviors that are
linked to detrimental learning outcomes. Together they provide new insight into the role of co-viewing in learning.

The results of this study represent a first step in linking specific co-viewing behaviors to children's learning outcomes in the context of screen media. There are various ways to refine and build on this work. One future direction would be to incorporate measures of child behavior and parent-child interaction quality in the kinds of analyses presented here. The relative timing of utterances and responses between a child and parent may be particularly predictive of word learning. For example, children may learn most effectively when parents respond promptly and provide information about the item that is the focus of the child's attention in that moment. The co-viewing data collected for this study could be coded for contingencies in interactions between parents and children. This kind of analysis would also resonate with research on the social aspects of screen mediated learning (e.g., Richert et al., 2011). Although social information was not manipulated directly in video training in this study, it could be informative to test how different extents of social contingency in co-viewing link to learning outcomes.

Another future direction for this work would be to guide experimental investigations of co-viewing. The kinds of analyses presented here can be used to develop experimental manipulations of the linguistic environment surrounding screen media co-viewing. For example, the current results suggest that the type and extent of labeling during coviewing may impact learning in different ways. This could be tested by manipulating whether labels are provided to children or elicited from them and how many labels are used during co-viewing. This would allow for greater control of other characteristics of the co-viewing context, randomized assignment of children to conditions, and causal conclusions about the role of labels in screen mediated word and category learning. Similar experiments could be designed to test the effects of specific, content-focused speech compared to broad, screen-focused speech during co-viewing. Future work could also investigate co-viewing and learning outcomes from real, professionally produced child-directed media. In conclusion, the current study makes novel contributions to the emerging literature on screen mediated language learning in young children and highlights directions for future research on facilitating this learning.

## References

American Academy of Pediatrics, Council on Communications and Media (2010). Media education. Pediatrics, 126(5), 1012-1017.
Anderson, D. R., \& Pempek, T. A. (2005). Television and very young children. Am. Behav. Sci., 48(5), 505-522.
Barr, R., Zack, E., Garcia, A., \& Muentener, P. (2008). Infants' attention and responsiveness to television increases with prior exposure and parental interaction. Infancy, 13, 30-56.
Brugman, H., \& Russel, A. (2004). Annotating multimedia/ multi-modal resources with ELAN. Proc. of LREC 2004.

DeLoache, J. D., Chiong, C., Sherman, K., Islam, N., Vanderborght, M., Troseth, G. L., Strouse, G. A., \& O'Doherty, K. (2010). Do babies learn from baby media? Psych. Science, 21(11), 1570-1574.
Demers, L. B., Hanson, K. G., Kirkorian, H. L., Pempek, T. A., \& Anderson, D. R. (2012). Infant gaze following during parent-infant coviewing of baby videos. Child Dev.
Dunn, L. M., \& Dunn, D. M. (2007). PPVT-4: Peabody picture vocabulary test. Minneapolis, MN: Pearson Assessments.
Fender, J. G., Richert, R. A., Robb, M. B., \& Wartella, E. (2010). Parent teaching focus and toddlers' learning from an infant DVD. Infant and Child Dev., 19, 613-627.
Fenson, L. et al. (2007) MacArthur-Bates communicative development inventories: User's guide and technical manual. Baltimore, MD: Paul H. Brookes Pub. Co.
Fenstermacher, S. K., Barr, R. R., Salerno, K., Garcia, A., Shwery, C. E., Calvert, S. L., \& Linebarger, D. L. (2010). Infant-directed media: An analysis of product information and claims. Infant and Child Dev., 19, 557-576.
Fidler, A.E., Zack, E., \& Barr, R. (2010). Television viewing patterns in 6- to 18 -month-olds: The role of caregiver-infant interactional quality. Infancy, 15, 176196.

Krcmar, M. (2010). Can social meaningfulness and repeat exposure help infants and toddlers overcome the video deficit? Media Psychology, 13(1), 31-53.
Krcmar, M., Grela, B., \& Lin, K. (2007). Can toddlers learn vocabulary from television? An experimental approach. Media Psychology, 10, 41-63.
Kuhl, P. K., Tsao, F., \& Liu, H. (2003). Foreign-language experience in infancy: Effect of short-term exposure and social interaction on phonetic learning. PNAS, 100(15), 9096-9101.
O’Doherty, K., Troseth, G. L., Goldenberg, E., Akhtar, N., Shimpi, P. M., \& Saylor, M. M. (2011). Third-party social interaction and word learning from video. Child Dev., 82(3), 902-915.
Reiser, R.A., Tessmer, M.A., \& Phelps, P.C. (1984). Adultchild interaction in children's learning from "Sesame Street." Edu. Comm. \& Tech. Journal, 32, 217-233.
Reiser, R.A., Williamson, N., \& Suzuki, K. (1988). Using "Sesame Street" to facilitate children's recognition of letters and numbers. Edu. Tech. Res. and Dev., 36, 15-21.
Richert, R. A., Robb, M. B., \& Smith, E. I. (2011). Media as social partners: The social nature of young children's learning from screen media. Child Dev., 82(1), 82-95.
Rideout, V. J., \& Hamel, E. (2006). The media family: Electronic media in the lives of infants, toddlers, preschoolers, and their parents. Henry J. Kaiser Family Foundation.
Sims, C. E., \& Colunga, E. (In preparation). Toddlers' word and concept learning from screen media.
Werker, J. F., Gilbert, J. H. V., Humphrey, K., \& Tees, R. C. (1981). Developmental aspects of cross-language speech perception. Child Dev., 52(1), 349-355

# Investigating the Other-Race Effect of Germans towards Turks and Arabs using Multinomial Processing Tree Models 

Henrik Singmann (henrik.singmann@psychologie.uni-freiburg.de)<br>David Kellen (david.kellen@psychologie.uni-freiburg.de)<br>Karl Christoph Klauer (christoph.klauer@ psychologie.uni-freiburg.de)<br>Institut für Psychologie, Sozialpsychologie und Methodenlehre, Engelbergerstr. 41<br>Albert-Ludwigs-Universität Freiburg, 79085 Freiburg, Germany


#### Abstract

The other-race effect (ORE) refers to the phenomenon that recognition memory for other-race faces is worse than for ownrace faces. We investigated whether White Germans exhibited an ORE towards Turkish or Arabic faces using a multinomial processing tree model (MPT), the two-high threshold model of recognition memory with three response categories (old, skip, and new). Using an MPT enabled us to adequately disentangle memory and response processes using the Fisher information approximation, a minimum description length based measure of model complexity. Results showed that participants exhibited an ORE on the memory parameters but not on the parameters representing response processes.


Keywords: Recognition Memory; Other-Race Effect; Multinomial Processing Tree Model; Face Recognition; Minimum Description Length

## The Other-Race Effect

The other-race effect (ORE, also known as own-race effect, own-race bias, or cross-race effect; e.g., Meissner \& Brigham, 2001; Hugenberg et al., 2010) describes the phenomenon that people tend to have a recognition memory advantage for own-race faces compared to other-race faces. A typical experiment consists of two phases, the study phase and the test phase. In the study phase, participants are asked to memorize a list of faces of at least two different races (e.g., white and arabic faces). In the subsequent test phase, participants are presented with old (i.e., presented during the study phase) and new (i.e., not presented during the study phase) faces of the two races and have to decide for each face separately if it was presented during the study phase by responding either "old" or "new".

The typical data pattern observed in such an experiment is a mirror effect, namely that participants produce more hits for own-race faces than for other-race faces (i.e., $P$ ("old"|old) is higher for own-race faces) and simultaneously more falsealarms for other-race faces (i.e., $P$ ("old"|new) is higher for other-race faces). A meta-analysis by Meissner and Brigham (2001) of 39 studies with almost 5,000 participants showed $P$ (hit) was 1.4 times higher for own-race faces than for otherrace faces and $P$ (false alarm) was 1.56 times higher for otherrace faces than for own-race faces, indicating a substantial ORE.

Recent findings have qualified this overall effect. For example, in a study by Gross (2009) Asian, Black, Hispanic, and White participants performed a recognition memory experiment with Asian, Black, Hispanic, and White faces. Furthermore note that the study was performed in Southern California (USA) where the majority of the population is Hispanic
( $42 \%$ versus $38 \%$ Whites). For all participants the best performance (at least descriptively) was found for faces of the participants' own race. When analyzing participants based on their race, an interesting pattern emerged. White participants had the best performance for white faces followed by Hispanic faces followed by Asian and Black faces. Hispanic participants had the best performance for Hispanic and White faces (so no significant advantage for own-race faces) followed by Asian faces followed by Black faces. These results indicate that the ORE does not generalize to all "other-races", but its magnitude depends on which other-race is the target.

Sporer and Horry (2011) have conducted a study with a similar design that is of special importance for the current paper. Their participants were White and Turkish participants living in Germany which were tested on faces of African Americans, Turks, White Americans and White Germans. White German participants exhibited an ORE only for African-American faces, there was no reliable difference in the memory for the other three target races. Turkish participants had a comparable performance for Turkish and White German faces, which was better than the performance for White- and African-American faces.

Taken together, these results indicate that people do not display an ORE towards all other-races, rather it is an empirical question which seems to be depend on factors such as facial features of the target race (e.g., White participants in Germany did not show and ORE towards non-German White faces whereas Turks in Germany did) and also on social-cognitive factors such as the majority/minority or ingroup/outgroup status of the target race (as e.g. shown by the study of Gross, 2009). The answer to the question whether an ORE is displayed towards a specific other-race may also have severe practical implications, for example in the domain of eyewitness identification, as and ORE can lead to the wrongful accusation or conviction of innocent individuals or to an acquittal of guilty individuals.

## The Present Experiment

In this experiment we investigate whether White Germans exhibit an ORE towards people of Middle Eastern descent such as Turks and Arabs. We selected Turks and Arabs as "otherrace" as (a) the only published study we know of investigating this (Sporer \& Horry, 2011) surprisingly did not find an ORE, (b) Turks are the biggest ethnic minority in Germany (around 3 million of a 82 million population, Statistisches


Figure 1: The 2HTM for recognition memory. On the left side are the two different item types, old and new items, respectively, each represented by one tree. On the right side are the observed responses "Old", "Skip", and "New". In between are the assumed latent states with the probabilities leading to these states. $D_{o}=$ Detect an old item as old, $D_{n}=$ detect a new item as new, $g_{i}=$ guessing state.

Bundesamt, 2011), and (c) increasing prejudices towards people with Middle Eastern descent have been observed in the aftermath of the terrorist attacks on 9-11 (e.g., Morgan, Wisneski, \& Skitka, 2011).

Furthermore, our experiment employed a novel methodology that enabled us to disentangle two types of cognitive processes that contribute to performance in recognition memory, memory and response processes. According to the current state of knowledge (Meissner \& Brigham, 2001), the ORE is an effect that affects both memory and response processes. However, in the next part of this manuscript we will argue that the usual employed methodology of disentangling memory and response processes is flawed and offer an alternative.

## Measuring Recognition Memory Performance

It is generally agreed upon that there are at least two different kinds of cognitive processes that contribute to an observed pair of $P($ hit $)$ and $P$ (false alarm) in a recognition memory experiment (Snodgrass \& Corwin, 1988): memory processes (e.g., how good the memory for the studied items is) and response processes (e.g., the tendency to respond with "old" instead of "new"). For example, better memory for the studied items should increase $P$ (hit) but decrease $P$ (false alarm), whereas differences in the response tendencies should affect $P($ hit $)$ and $P($ false alarm $)$ in the same direction.

So far, the study of the ORE has relied on simple performance indices such as $d^{\prime}$ or $A^{\prime}$ (see Macmillan \& Creelman, 2005) to measure the aforementioned processes. Based on these indices Meissner and Brigham (2001) concluded that the ORE is mainly a memory effect and that there are also
differences in response bias (i.e., participants are more inclined to respond with "old" for other race faces) which are considerably smaller than the effect for memory processes.

However, the ability of these measures to accurately characterize an individual's performance is known to be quite limited, as they make simplifying assumptions (e.g., homoscedasticity of the evidence distributions) which, if violated, seriously compromise any conclusions drawn from the analysis (see Verde, Macmillan, \& Rotello, 2006). In particular, differences in memory and response tendencies tend to be grossly confounded. Unfortunately the data usually gathered does not allow to test these assumptions, encouraging the use of measurement models which are based on richer data sets.

A model that has been extensively used in the literature is the Two-High-Threshold Model (2HTM; Snodgrass \& Corwin, 1988). The 2HTM assumes that studied and non-studied items at test can be in either a detection or uncertainty state: When an item is detected its true status (studied or nonstudied) is known, and a correct response is invariably given. On the contrary, when an item is not detected, it is in an uncertainty state where no information regarding its true status is available, leading to a response based on guessing. The 2HTM can be represented as a multinomial processing tree (MPT; Riefer \& Batchelder, 1988), as depicted in Figure 1.

Figure 1 describes the 2 HTM for a recognition-memory task, which we have enriched by introducing a third response option "skip" in addition to "old" and "new". The studied items can be detected with probability $D_{o}$, which is assumed to invariably lead to response "old". In the absence of detection (which occurs with probability $1-D_{o}$ ), the individual merely guesses, responding "old", "skip", or "new" with probabilities $g_{1},\left(1-g_{1}\right)\left(1-g_{2}\right)$, and $\left(1-g_{1}\right) g_{2}$ respectively. Non-studied items can be detected with probability $D_{n}$, invariably leading to response "new" and in the absence of detection (which occurs with probability $1-D_{n}$ ) the individual guesses using the above-stated probabilities. Parameters $D_{o}$ and $D_{n}$ attempt to capture memory retrieval processes as well as memory-based metacognitive judgments (e.g., Strack \& Bless, 1994), while $g_{1}$ and $g_{2}$ capture response tendencies.

The advantages of using this model are threefold: First, the 2HTM is a simple yet validated model (e.g., Bröder, Kellen, Schütz, \& Rohrmeier, in press; Snodgrass \& Corwin, 1988) that is used as a building block in more complex measurement models (e.g., Klauer \& Kellen, 2010). Second, the use of 2HTM allows for different independent parameter estimates for German and Turkish face-stimuli, while other popular measurement models based on Signal Detection Theory (Macmillan \& Creelman, 2005) suffer from identifiability issues that compromise their use (see Wickens \& Hirshman, 2000). Third, the 2HTM is a member of the MPT model class, a class for which model selection under the Minimum Description Length principle is well documented and available (Singmann \& Kellen, in press; Wu, Myung, \& Batchelder, 2010). The latter point is discussed in greater detail below.

## Model Selection in the MPT Model Class: A Minimum Description Length Approach

One of the advantages when employing an analysis based on cognitive models is that model parameters capture entities of primary interest such as the probability with which a certain cognitive state is reached (e.g., the probability of remembering a face). A direct consequence of this is that psychological hypothesis directly correspond to relationships among model parameters. For example, the absence of an ORE corresponds to the identity of parameters for German and Arabic faces. Hence, in our analysis different versions of the MPT model described above are used corresponding to different hypothesis (e.g., no ORE, an ORE only based on memory processes, etc.). To establish which hypothesis is the most plausible given the experimental result therefore entails an assessment of the performance of the different models, a process known as model selection.

Model selection requires a weighting between the ability of the model to account for the observed data (via goodness-offit statistics) and the ability of the model to account for data in general (model complexity or flexibility), as more flexible models provide a better fit to data a priori (Roberts \& Pashler, 2000). The established goal is to find the model with the best trade-off between fit and flexibility, with different methods and approaches being proposed in the literature (e.g., Vanderkerckhove, Matzke, \& Wagenmakers, submitted).

One prominent approach in model selection is based on the Minimum Description Length principle (MDL; Grünwald, 2007). According to the MDL approach, both models and data are understood as codes that can be compressed. The goal of MDL is to assess models in terms of their ability to compress data. The greater the compression, the better the account of the underlying regularities that are present in the data. One of the indices emerging from the MDL principle is the Fisher Information Approximation (FIA), which combines a model's goodness of fit with model-flexibility penalties. Specifically, FIA takes the number of parameters and the sample size into account, but additionally contains a term that accounts for the flexibility of the model due to its functional form by integrating over the determinant of the Fisher information matrix of the model for a sample of size 1 (see Wu et al., 2010). An algorithm that computes FIA for members of the MPT model class was developed by Wu et al. (2010), and made available in an open-source package by Singmann and Kellen (in press).

While common model-selection indices such as AIC and BIC (Burnham \& Anderson, 2002) use the number of free parameters as a proxy for the relative flexibility of a model, FIA is able to capture the model's ability to account for data in general. Because of this MDL indices such as FIA usually outperform AIC and BIC (and null-hypothesis testing) when attempting to identify the most suitable model (e.g., Klauer \& Kellen, 2011).

One further advantage of FIA is the ability to incorporate order restrictions being imposed on the parameters (e.g.,
$D_{o} \geq D_{n}$ ), allowing for the testing of informative hypotheses (Kellen, Klauer, \& Bröder, in press). This means that one can restrict the flexibility of the models to patterns that are theoretically plausible, and directly test whether this flexibility is sufficient to account for the observed data.

## Methods

## Participants

A total of 42 White German psychology students (mean age $=21.4$ years, $S D=2.7$ ) participated in the experiment for partial fulfillment of course credits.

## Materials

The pictures were taken from publicly accessible websites of sports team of lower leagues (mostly soccer teams) from Central European countries (e.g., Germany, Belgium), Turkey, and Middle Eastern countries. In total we gathered 123 White and 125 Turkish/Arabic pictures (henceforth we will refer to these as Arabic pictures). We digitally removed the background and eye-catching features and colorized the shirts uniformly black. The pictures were then pretested in an online study on four 7-point scales: two ethnicity dimensions (German/Central European and Turkish/Arabic), distinctiveness ("How easy it is to spot the face in a crowd?", Valentine \& Bruce, 1986), and valence (positive to negative). We obtained a mean of 20.6 ratings per picture. The ethnicity dimensions were subtracted from each other to form a racial extremity score (i.e., German minus Turkish rating for White pictures and vice versa for Arabic pictures).

Based on the pretest data we selected 100 pictures from each category (avoiding extreme ratings on any dimension) that were comparable (albeit significantly different) in their ratings. On the racial extremity dimension the Arabic pictures were somewhat less extreme than the White pictures, 3.7 versus 4.6 . On the distinctiveness and valence dimension ratings were comparable, 3.5 versus 3.0 and 4.3 versus 4.6, respectively. Additionally, we randomly selected another 10 pictures from the remaining pictures to serve as primacy and recency items in the study phase. More details on the pretest can be found here: http://www.psychologie.unifreiburg.de/Members/singmann/pubs/cogsci13supp

## Procedure

Participants were tested individually. They were informed that they were to take part in a memory experiment consisting of two study phases in which they had to memorize a set of faces and a subsequent test phase. No reference was made to race or related concepts. In the first study phase, 50 White and 50 Arabic faces (randomly selected) were presented in random order (plus 5 primacy faces at the beginning and 5 recency faces at the end) each for 2 seconds with a 0.5 seconds inter-trial interval (ITI). To increase memory for the pictures, the study phase was repeated with the same items (plus primacy and recency items) presented in a new random order. Directly after the second study phase, participants were introduced to the test phase in which they had to
judge for each of 200 faces ( 50 White old, 50 White new, 50 Arabic old, and 50 Arabic new) whether or not it was presented during the study phase by selecting one of three options: "old" ["altes Gesicht"], "skip" ["überspringen"], or "new" ["neues Gesicht"]. We implemented a 0.5 seconds ITI in the test phase.

## Results

## Response Proportions

Table 1 presents the response proportions obtained for White and Arabic faces and $p$-values of $t$ tests comparing those (without controlling for multiple testing). As can be seen, we did not find a mirror effect. Rather, we found slightly higher proportions of hits and higher proportions of false alarms for the Arabic faces. Additionally, we found differences in the use of the "skip" option in that participants used "skip" more often for Arabic than for White new faces.

## MPT Analysis

All analyses were performed using MPTinR (Singmann \& Kellen, in press).

The Unrestricted Model. We fitted the unrestricted 2HTM model to each individual dataset separately using the maximum likelihood method. The model seemed to provide a good fit to the data (as the unrestricted model was saturated we couldn't formally test this). Of the 42 participants only 6 participants had a $G^{2}>1$, of those only two participants had a $G^{2}>2$ (3.70 and 4.95). The summed $G^{2}$ was 19.19.

The mean parameter estimates and the underlying distributions plus additional information are depicted in Figure 2. As can be seen, there were no big differences for parameters $D_{o}$, $g_{1}$, and $g_{2}$. Only $D_{n}$ showed a difference in the expected direction, $D_{n}$ was smaller for Arabic than for White faces. This results was also supported by significance testing (Table 2), only for $D_{n}$ did the parameters for White and Arabic faces differ. Somewhat unexpectedly, $D_{o}$ tended to be slightly higher for Arabic faces than for White faces, although this result did not reach significance.
Model Selection. To test for the existence of an ORE we fitted eight models implementing different sets of parameter

Table 1: Mean Response Proportions (SD)

|  | White faces | Arabic faces | $p$ |
| ---: | :---: | :---: | :--- |
| $P($ hit $)$ | $.67(.16)$ | $.71(.14)$ | .06 |
| $P\left(\right.$ skip $\left._{\text {old }}\right)$ | $.07(.08)$ | $.07(.09)$ | .66 |
| $P($ miss $)$ | $.26(.16)$ | $.21(.12)$ | .03 |
| $P($ fa $)$ | $.16(.11)$ | $.27(.17)$ | $<.001$ |
| $P\left(\right.$ skip $\left._{\text {new }}\right)$ | $.09(.12)$ | $.13(.13)$ | .01 |
| $P(\mathrm{cr})$ | $.75(.16)$ | $.60(.19)$ | $<.001$ |

Note. Column $p$ contains $p$-values from paired $t$ tests comparing response proportions for White and Arabic faces. $P(\mathrm{fa})=P($ false alarm $) ; P(\mathrm{cr})=P($ correct rejection $)$.

Table 2: Mean parameter values (SD) of model without parameter restrictions

| parameter | White | Arabic | $p$ |
| ---: | :---: | :---: | :---: |
| $D_{o}$ | $.45(.23)$ | $.50(.23)$ | .10 |
| $D_{n}$ | $.53(.28)$ | $.24(.23)$ | $<.001$ |
| $g_{1}$ | $.39(.22)$ | $.38(.22)$ | .79 |
| $g_{2}$ | $.78(.24)$ | $.74(.25)$ | .09 |

Note. Column $p$ contains $p$-values of paired permutation tests comparing parameters across races using 100.000 bootstrap samples (Hothorn et al., 2006). p-values of paired $t$ tests are identical up to the second decimal (up to the fourth decimal for $p<.001$ ).
restrictions and furthermore calculated the FIA for each of those models using 200,000 Monte Carlo samples (see Table 3 for the results). The different models correspond to the different hypothesis regarding the nature of the ORE we could capture with the 2 HTM . The first model is the model without any parameter restrictions reflecting the possibility that an ORE is driven by both differences in memory processes and differences in response tendencies. In models two to four, only memory parameters (i.e., $D_{o}$ and $D_{n}$ ) were restricted to be equal across the races, but response tendencies were allowed to vary. Model five only assumes differences in the memory parameters but no differences in the guessing parameters. Models six to eight implement different versions of a memory ORE with the guessing parameters restricted. Note that for all but the first model we implemented an inequality restriction on the memory parameters so that $D_{o}$ and $D_{n}$ for White faces were equal or larger than those for Arabic faces (unless they were restricted to be equal).

The model with the best performance was model 6 (Ta-

Table 3: Model selection results for models with different parameter restrictions across face races

| $\#$ | restricted | $d f$ | $G^{2}$ | $p$ | FIA | best |
| :--- | ---: | ---: | ---: | :--- | :---: | ---: |
| 1 | none | 0 | 19.19 |  | 516.41 | 0 |
| 2 | $D_{o}$ | 42 | 48.29 | .23 | 486.55 | 3 |
| 3 | $D_{n}$ | 42 | 71.07 | .003 | 497.69 | 2 |
| 4 | $D_{o}, D_{n}$ | 84 | 96.91 | .16 | 503.66 | 3 |
| 5 | $g$ | 84 | 182.71 | $<.001$ | 477.10 | 1 |
| 6 | $D_{o}, g$ | 126 | 210.97 | $<.001$ | $\mathbf{4 5 4 . 4 6}$ | 16 |
| 7 | $D_{n}, g$ | 126 | 356.87 | $<.001$ | 527.66 | 3 |
| 8 | $D_{o}, D_{n}, g$ | 168 | 385.13 | $<.001$ | 504.25 | 14 |

Note. The results are summed across participants. The lowest FIA value is printed in bold. Column "best" contains the number of times each model provided the best performance (using FIA as the criterion). If not restricted, $D_{o}$ and $D_{n}$ are inequality restricted to be equal or larger for German faces than for Arabic faces (except for the "none" model in which all parameters are free).


Figure 2: Violin plots (Hintze \& Nelson, 1998) of the parameter estimates for the unrestricted 2HTM. The outer area depicts the density of each variable. The boxplot depicts the first, the second (i.e., the median), and the third quartile and the area 1.5 times the interquartile range outside those values. The mean is depicted as a $\times$.
ble 4) with $D_{o}$ and the $g$ parameters restricted across races and the only differences being $D_{n \text { White }} \geq D_{n A r a b i c}$ (this model also had the lowest FIA value in an analysis without the inequality restrictions on the memory parameters), indicating that we found a memory based ORE. This model did not only have the lowest FIA value, but also provided the best FIA performance for the largest number of participants (16 of 42 participants).

Inspecting the best performing models per individual datasets revealed a surprising spread. Each model (with the exception of the unrestricted model) provided the best performance for at least one dataset. Furthermore, the model assuming no ORE (model 8) provided the best fit for 14 participants, indicating that quite a substantial subgroup did not show an ORE.

## Discussion

The goal of this experiment was to investigate whether White Germans exhibit an ORE towards people of Middle Eastern descent such as Turks or Arabs. In contrast to the only other published study we know of investigating this (Sporer \& Horry, 2011), we indeed found evidence for an ORE of our participants. More specifically, the analysis using the 2HTM

Table 4: Mean parameter values (SD) of best performing model with parameters $D_{o}$ and $g$ restricted

| parameter | White | Arabic | $p$ |
| ---: | :---: | :---: | :---: |
| $D_{o}$ | $.50(.23)$ |  |  |
| $D_{n}$ | $.52(.24)$ | $.23(.24)$ | $<.001$ |
| $g_{1}$ | $.35(.18)$ |  |  |
| $g_{2}$ | $.76(.23)$ |  |  |

Note. See Table 2.
revealed that the majority of the participants were less able to detect the correct status of new items for Turkish and Arabic faces (i.e., lower $D_{n}$ ) than for White faces, hence our ORE was a pure memory effect. There were no reliable differences on the other memory parameter (i.e., $D_{o}$ ) nor on the response bias parameters (i.e., $g$ ). Additionally, our analysis revealed that not all participants exhibit an ORE. In fact, although most of the participants did show this effect, 14 of 42 participants did not show any ORE. This latter finding may in part be responsible for the failure of Sporer and Horry to find an ORE towards Turks as their analysis strategy might have not have been as powerful as ours, as it may have suffered from problems of the employed performance index $A^{\prime}$ (see also below).

When looking at the response proportions only, we did not find the expected mirror effect (higher hit rate for own-race faces and higher false alarm rate for other-race faces; Meissner \& Brigham, 2001) which is the usual data pattern in the ORE. One possible explanation for this is that our decision to enrich the data base by introducing a "skip" option may have hidden the mirror effect. Alternatively, the mirror effect, which is usually found in studies when the other-race is Black, could be absent for Arabic faces. Future research should try to disentangle these two explanations. The absence of the mirror effects also indicates that, although we did find an ORE, our finding regarding the underlying processes may not simply generalize to different own- or other-races.

Enriching the data base by introducing the "skip" option and thereby allowing to employ a fully identified two-high threshold multinomial processing tree model, has proven to be a useful tool in investigating the ORE. It not only overcomes limitations of the often-used performance indices such as $d^{\prime}$ or $A^{\prime}$ (Verde et al., 2006), it is also able to overcome identifiability issues when using two different stimuli classes (i.e., White and Arabic faces) in a signal-detection framework
(Wickens \& Hirshman, 2000). We hope that this new tool may help in answering some of the open questions regarding the ORE (see Hugenberg et al., 2010).

The adopted model selection strategy was also successful in uncovering interesting individual differences. Theories of ORE have highlighted that differences in ORE can be explained by social-cognitive variables such as attitudes towards other-races (Hugenberg et a., 2010). Combining our methodology with relevant individual differences measures within the MPT framework, such as the new MPT model for the implicit association test (IAT; F. Meissner and Rothermund, in press), could therefore be fruitful.

## Acknowledgments

We thank Alina Arnhold, Felicitas Flade, and Hannah Kammüller for their help in collecting the data and Jasmyn Touchstone and Johannes Falck for preparing the stimuli.

## References

Broder, A., Kellen, D., Schutz, J., \& Rohrmeier, C. (in press). Validating a two-high-threshold measurement model for confidence rating data in recognition. Memory.
Burnham, K. P., \& Anderson, D. R. (2002). Model selection and multi-model inference: A practical informationtheoretic approach. Springer.
Gross, T. F. (2009). Own-ethnicity bias in the recognition of black, east asian, hispanic, and white faces. Basic and Applied Social Psychology, 31(2), 128-135.
Grunwald, P. D. (2007). The minimum description length principle. Cambridge, Mass.: MIT Press.
Hintze, J. L., \& Nelson, R. D. (1998). Violin plots: A box plot-density trace synergism. The American Statistician, 52(2), 181-184.
Hothorn, T., Hornik, K., Wiel, M. A. van de, \& Zeileis, A. (2006). A lego system for conditional inference. The American Statistician, 60(3), 257-263.
Hugenberg, K., Young, S. G., Bernstein, M. J., \& Sacco, D. F. (2010). The categorization-individuation model: An integrative account of the other-race recognition deficit. Psychological Review, 117, 1168-1187.
Kellen, D., Klauer, K. C., \& Bröder, A. (in press). Recognition memory models and binary-response ROCs: a comparison by minimum description length. Psychonomic Bulletin \& Review, 1-27.
Klauer, K. C., \& Kellen, D. (2010). Toward a complete decision model of item and source recognition: A discrete-state approach. Psychonomic Bulletin \& Review, 17, 465-478.
Klauer, K. C., \& Kellen, D. (2011). The flexibility of models of recognition memory: An analysis by the minimumdescription length principle. Journal of Mathematical Psychology, 55(6), 430-450.
Macmillan, N. A., \& Creelman, C. D. (2005). Detection theory : a user's guide. New York: Lawrence Erlbaum associates.
Meissner, C. A., \& Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-
analytic review. Psychology, Public Policy, and Law, 7(1), 3-35.
Meissner, F., \& Rothermund, K. (in press). Estimating the contributions of associations and recoding in the implicit association test: The ReAL model for the IAT. Journal of Personality and Social Psychology.
Morgan, G. S., Wisneski, D. C., \& Skitka, L. J. (2011). The expulsion from disneyland: The social psychological impact of 9/11. American Psychologist, 66(6), 447-454.
Riefer, D. M., \& Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. Psychological Review, 95(3), 318-339.
Roberts, S., \& Pashler, H. (2000). How persuasive is a good fit? a comment on theory testing. Psychological Review, 107(2), 358-367.
Singmann, H., \& Kellen, D. (in press). MPTinR: analysis of multinomial processing tree models in r. Behavior Research Methods, 1-16.
Snodgrass, J. G., \& Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. Journal of Experimental Psychology: General, 117(1), 34-50.
Sporer, S. L., \& Horry, R. (2011). Recognizing faces from ethnic in-groups and out-groups: Importance of outer face features and effects of retention interval. Applied Cognitive Psychology, 25(3), 424-431.
Statistisches Bundesamt (Ed.). (2012). Bevölkerung mit migrationshintergrund: Ergebnisse des mikrozensus 2011; [population with migration background: results of the microcensus 2011] (Vol. Reihe 2.2) (No. 1). Wiesbaden: Statistisches Bundesamt.
Strack, F., \& Bless, H. (1994). Memory for nonoccurrences: Metacognitive and presuppositional strategies. Journal of Memory and Language, 33(2), 203-217.
Valentine, T., \& Bruce, V. (1986). The effects of distinctiveness in recognising and classifying faces. Perception, 15(5), 525-535.
Vandekerckhove, J., Matzke, D., \& Wagenmakers, E.-J. (submitted). Model comparison and the principle of parsimony. In J. R. Busemeyer, Z. Townsend, J. Wang, \& A. Eidels (Eds.), Oxford handbook of computational and mathematical psychology. Oxford: Oxford University Press.
Verde, M. F., Macmillan, N. A., \& Rotello, C. M. (2006). Measures of sensitivity based on a single hit rate and false alarm rate: The accuracy, precision, and robustness of, A z , andA. Perception \& Psychophysics, 68(4), 643-654.
Wickens, T. D., \& Hirshman, E. (2000). False memories and statistical design theory: Comment on miller and wolford (1999) and roediger and McDermott (1999). Psychological Review, 107(2), 377-383.
Wu, H., Myung, J. I., \& Batchelder, W. H. (2010). On the minimum description length complexity of multinomial processing tree models. Journal of Mathematical Psychology, 54(3), 291-303.

# Semantics: Intuition, Experiment or Illusion? 

Peter Slezak (p.slezak@unsw.edu.au)<br>Department of Philosophy, School of Humanities<br>University of New South Wales, Sydney NSW 2052 Australia


#### Abstract

Externalist theories in natural language semantics have become the orthodoxy since Kripke is widely thought to have refuted descriptive theories involving internal cognitive representation of meaning. This shift may be seen in developments in philosophy of language of the 1970s - the direct reference "revolution against Frege" (Wettstein 2004, 66). Almog (2005, 493) writes of the "uprising against Frege's doctrines" that "spread like fire" based on the work of Kripke, Donnellan, Putnam and Kaplan. However, I consider Fodor's (2004) heretical thought that something has gone "awfully wrong" in this philosophical consensus, perhaps confirming Chomsky's (1992) view that the whole field of philosophical semantics is "utterly wrongheaded" and "crazy" by virtue of its non-naturalist assumptions and "methodological dualism." I suggest that the externalist orthodoxy is a kind of cognitive illusion seen elsewhere in philosophy and cognitive science.


Keywords: semantics; externalism; meaning; intuition.

## Externalist Orthodoxy

Externalism is widely acknowledged to be the orthodoxy in the theory of mental content and psychological states. However, despite its subjective force, externalism may be undermined by attending to its aetiology and showing how the intuitions evoked arise from deceptive mechanisms. Instead of defending internalism directly, we may ask: Why does externalism seem so convincing? This is a cognitive science of biases and illusions among philosophers.

Kripke (1972) is regarded as having "ushered in a new era in philosophy" (Soames 2005, 1) by refuting a widely held descriptive conception of proper names. In the philosophy of language, this was part of the 1970s direct reference "revolution against Frege" (Wettstein 2004, 66). Frege held that something about the speaker's cognitive state must explain the difference between sentences such as "Hesperus is Phosphorus" and "Hesperus is Hesperus." The first is cognitively significant but the second is knowable $a$ priori as necessarily true, even though the substituted terms are co-referential. However, Kripke's externalist doctrine of "rigid designators" has become the orthodoxy - essentially the view of J.S. Mill that proper names have no meaning other than the name's denotation, and a name refers to the same individual in any possible situation.

In Putnam's slogan, the externalist orthodoxy holds that "meanings ain't in the head" since mental content is individuated by referents in the world. This view rests on intuitions elicited by thought-experiments such as Putnam's (1975) famous Twin Earth story, characterized as "a sort of paradigm in the philosophies of language and mind" (Segal

2000, 24). On another planet, Twin Earth, the only difference is that the clear, potable liquid in rivers and lakes has chemical structure XYZ rather than $\mathrm{H}_{2} \mathrm{O}$. An atom-foratom replica of an Earth person might have identical internal psychological/brain states and yet not have the same waterthoughts since Twin Earth thoughts are about XYZ. Also influential has been Kripke's (1979) puzzle about Pierre who believes both that Londres est jolie and also that London is ugly, not realizing that London is the same city as Londres. Kripke says "I know of no answer" to the question "Does Pierre, or does he not, believe that London is pretty?" Kripke regards the puzzle as comparable to the Liar Paradox (1979, 904). On this point, Salmon (2011, 236) endorses Kripke's "sound methodology" quoting Tarski's classic discussion of the Liar antinomy and its intellectual challenge.

Kripke's "primary moral" is that "the puzzle is a puzzle" (1979) and he insists that it can not be resolved by redescribing the problem, but this conception is open to challenge. A re-description need not avoid the problem but rather it may show how a pseudo-problem arises. After all, the indeterminacy of Pierre's belief about London is not like the contradictory state of Shrödinger's cat or the quantum wave/particle duality. To be sure, in another case, restating Zeno's paradox of Achilles and the Tortoise (e.g. with a distance/time graph) is to sidestep the puzzle rather than solving it since the re-description doesn't expose the flaw in Zeno's reasoning. Kripke is right to say that talk of 'what is really going on' doesn't answer his original question, but it does show clearly what's wrong with the original question and why the puzzle isn't a puzzle, after all. With Kripke, we can point out that "No answer has yet been given" to the question of whether Lois Lane loves Clark Kent, but we understand why.

Or, seeing the Necker Cube on two different occasions, Pierre might not recognize it as the same geometrical figure. Adapting Kripke's (1979) words, we may ask "Does Pierre, or does he not, believe that the Figure (not the shape satisfying such-and-such descriptions, but the Figure) is facing upwards to the left? No answer has yet been given." Fodor $(2008,76)$ pointedly asks "But why on earth should we suppose that the question [concerning Pierre] has a definite right answer when it's phrased that way? And, once one sees why it doesn't, why does it matter that it doesn't?" However, while sharing Fodor's dismissive attitude, we may go further to ask why the puzzle should have such a firm grip on philosophical imagination.

Thus, Devitt $(1984,385)$ has made a salutary distinction: "Thoughts are one thing, their ascription another." Devitt
warns "it is a common practice ... to use 'belief', for example, where what one means to refer to is belief ascription" (1984, 389). The failure to respect Devitt's distinction is to blame for Kripke's puzzle in which we seem forced to describe the hapless Frenchman as holding contradictory beliefs about London. The relevance of Devitt's distinction should be clear: "a difference in sorts of thought ascription does not entail a difference in the sorts of thought object ascribed" (1984, 389). In this case, the question concerning Pierre's belief about London involves thought ascription about the thing itself or de re, using our own reference, like Putnam's thought ascriptions about $\mathrm{H}_{2} \mathrm{O}$ and XYZ. The intuition that we can be induced to share is simply the idea that we can ascribe de re beliefs from our own perspective independently of the beliefs of the subject in question. Brandom $(1994,503)$ explains, "expressions that occur within the scope of the 'that' [in de dicto contexts] serve to specify how things are represented by the one to whom the belief is ascribed."

## Little Choice?

Significantly, Kripke (1972, 42) acknowledges that he was led by his "natural intuition" to his view of proper names and that there could not be "more conclusive evidence one can have about anything, ultimately speaking." However, Farkas (2003) characterizes the "deeply rooted" intuitions as "baffling" and a "vexatious problem" that "poses a serious challenge for any attempts to give an internalist analysis." Accordingly, we may ask why philosophers feel that the "intuitive responses to a certain kind of thought-experiment appear to leave them little choice," as Boghossian (1998, 273) puts it. Fodor (1987a) has noted that the Twin-Earth Problem is not a problem but "just a handful of intuitions together with a commentary on some immediate implications of accepting them" (1987a, 208). Significantly, he says: "it is very plausible that all these intuitions hang together. The question is: What on earth do they hang on?" (Fodor 1987, 202). I offer an answer that gains a distinctive, if not decisive, strength from the fact that the intuitions in this domain "hang on" the same biases and illusions to be seen operating elsewhere throughout cognitive science.

## Giving Intuitions a Bad Name

In different guises, under such headings as 'conceptual analysis' (Jackson 1998) or 'conceivability' (Chalmers 2002), intuitions have played a central role in philosophy (DePaul \& Ramsey eds. 1998). Hintikka (1999, 127) suggested intuitions "came into fashion in philosophy" as philosophers' attempted to "get on the bandwagon of transformational grammar" that they took to provide a methodological model for research into cognition. Hintikka (1999, 127,8) specifically cites Kripke's (1972) Naming and Necessity as an influential case in point, suggesting "Unfortunately" his doctrines are "apt to give intuitions a bad name." Even a sympathetic account by Hughes (2004) makes a damaging admission: He confesses "blindness" to Dummett's (1973) alternative reading of key sentences but
takes "comfort" from the fact that the same defect is very widespread among philosophers. However, the Müller-Lyer illusion is very widespread too. As Sosa $(2001,26)$ notes, the phenomenon of ambiguity is widespread in the English language and the "shiftiness" of linguistic constructions containing modal expressions is akin to lexical ambiguity of words such as "bank." Closer are the structural ambiguities familiar to linguists and the basis for jokes such as Groucho Marx's remark: "One morning I shot an elephant in my pajamas." Failure to appreciate the humour through blindness to the ambiguity is a psychological defect rather than theoretical criticism.

Contrary to Hintikka (1999, 132), Chomsky's use of intuitions in linguistics has nothing to do with being a "selfacknowledged Cartesian" or innate ideas. Nevertheless, Hintikka (1999, 133) correctly notes, in contrast to linguists' use of intuition, "philosophers' intuitions do not pertain to the supposed faculty of intuition itself but to the truths about which this faculty is supposed to provide knowledge." For an egregious example, Bealer $(1998,202)$ argues that intuitions have a "strong modal tie to the truth" which he suggests "is a philosophical (conceptual) thesis not open to empirical confirmation or refutation." In the same vein, Chalmers (2002) challenges the systematic scientific picture asking "Does Conceivability Entail Possibility?" However, we need not agree that "Philosophical intuition is epistemologically useless" (Cummins 1998, 125). If philosophers' intuitions are taken properly on the model of generative grammar, they may be seen as psychological evidence rather than intimations of truth. The PutnamKripke intuitions might be explained like the Müller-Lyer illusion as deceptive in spite of its subjective force.

## Omniscient Philosopher-Narrator

The model for this kind of inquiry into intuitive judgements is the 'heuristics and biases' program of Tversky and Kahneman (1974). This work has demonstrated the systematic unreliability of compelling intuitions resulting in a wide range of cognitive illusions to which we are prone. Seen from this perspective, I suggest externalist theories of reference involve a generic pseudo-explanatory mistake that is not confined to any one domain. For example, Chomsky has explained the need for a fully explicit grammar that avoids the unwitting dependence on the linguistic knowledge of the theorist. Of course, the potential for this error is not unique to linguistic explanation and its very seductiveness means we should expect to find it elsewhere. Generally, it seems difficult to avoid invoking internal representations which have their meaning because we, as theorists, can understand them. This has been the charge against pictorial theories of imagery by Pylyshyn (2003) and was precisely anticipated by Descartes. In this case external representations are taken as a model for internal representations and, therefore, relying on the theorist's intelligence and invoking the notorious homunculus. Chomsky (1962) notes that a grammar may produce the illusion of explanatory completeness, but in fact have
"serious limitations so far as linguistic science is concerned" because the success of the grammar depends on being "paired with an intelligent and comprehending reader." Chomsky explains: "Reliance on the reader's intelligence is so commonplace that is significance may be easily overlooked" (Chomsky 1962, 528). In a different guise of interest here, the theorist posits mental representations based on his own knowledge of the truth rather than the subject's beliefs. In this case, philosophical intuitions arise from tacitly adopting the perspective of an invisible narrator - the illusion of the omniscient story-teller, the literary device that Mario Vargas Llosa (1975) aptly refers to as the "philosopher-narrator."

## Residue of Commonsense

Pietroski (2003) suggests that despite a considerable literature on reference, "no one has shown that names do bear any interesting and theoretically tractable relation to their bearers." If he is correct, we are owed an explanation of how so many philosophers could have been so misguided. Chomsky characterizes the commonsense conception of semantics as a kind of illusion and points to the kind of diagnostic, aetiological concern I wish to pursue: "Here, I think, philosophers and linguists and others who are in the modern intellectual tradition are caught in a kind of trap, namely, the trap that assumes that there is a reference relation" $(2012,28)$. That is, "there is no wordthing relations." This is undoubtedly a shocking remark that flies in the face of the most obvious, taken-for-granted facts about language. Of course, that's just the point. Chomsky suggests that we may suffer from a "residue of commonsense," some deeply persuasive, but illegitimate, "distorting" picture of the world (see also Egan 1999, 188). Word-thing relations are "mythical" by contrast with the question of "how the person's mental representations enter into articulation and perception" $(1996,23)$, but this is syntax. Chomsky $(2000,148)$ suggests that we can have no intuitions about such questions as whether an identical replica of ourselves uses the word "water" to refer to something, XYZ, which is not $\mathrm{H}_{2} \mathrm{O}$ because the key terms such as "extension" and "reference" are technical inventions. In the same way, it would be pointless to explore our intuitions about "tensors" or "undecidability." However, there can be no doubt that certain intuitions may be consistently induced in philosophers and others by the notorious thought experiments. These are not random in the way that intuitions about tensors might be among the uninitiated. The vast philosophical literature attests to the existence of systematic, robust and widely shared intuitions that are at the heart of externalism.

## Who Cares What the Mayans Think?

Recently, the question has been illuminated from a new angle by empirical inquiries into the cross-cultural variation in intuitions on which philosophers have relied (Machery et al., 2004). These studies have challenged the universality of the evidence on which philosophical puzzles have relied.

For example, Segal $(2004,339)$ says "we should not trust those intuitions" because Putnam and Kripke "mistakenly think that their intuitions are 'ours', that they are representative of those of all sensible, reflective humans" (2004, 340). Segal reports studies "designed to tap relevant twin-Earth intuitions among tribespeople" such as the Mayans of the Yucatan in Mexico. The data are mixed, but Segal says "surely these data ... should be given considerably more weight than Putnam's intuitions about Oscar's "water" concept and Kripke's intuitions about medieval "unicorn" concepts (Segal 2004, 343). In the same vein, Machery et al. (2004, B7) found that "Chinese subjects tended to have descriptivist intuitions, while Westerners tended to have Kripkean ones" and these data suggest "significant philosophical conclusions." The authors conclude:


#### Abstract

We find it wildly implausible that the semantic intuitions of the narrow cross-section of humanity who are Western academic philosophers are a more reliable indicator of the correct theory of reference ... than the differing semantic intuitions of other cultural or linguistic groups. (2004, B9)


## Competence or Incompetence?

Devitt (2011) rejects the challenge of cross-cultural evidence to semantic theory because they tested the wrong subjects. The intuitions of ordinary folk are unreliable by comparison with intuitions of "experts," namely, "metaphysicians and other philosophers." However, we need not accept philosophers' intuitions as authoritative divinations to treat them, instead, as diagnostic evidence of illusion among those who suffer from it - data for the development of a theory of 'tacit knowledge' or "competence" (i.e. incompetence).

The point has been missed in the ongoing controversy about empirical inquiries into intuitions. Recently, Nagel (2012) has argued that epistemic intuitions do not, after all, vary in ways that pose a challenge, but Stich (2012) has defended such research and its threat to philosophical reliance on intuition. He cites evidence that even the MüllerLyer visual illusion is not universally shared among all human cultures. Kalahari San foragers apparently do not judge the familiar lines as differing in length. So what? Devitt, Stich and Nagel miss the point that it remains a matter of psychological interest to explain why we do suffer from the illusion. The only difference with the case of philosophical intuitions is that we don't take our visual perceptions as veridical.

That is, it is no help to be told that someone else doesn't share your puzzlement. Who cares what the Mayans or Chinese think? Their failure to be puzzled doesn't help resolve our problems. If I am the only one who is guilty of confirmation bias or base rate neglect, I need diagnosis and a cure, not anthropology. Even if it is parochial to Western departments of analytic philosophy, the central problem remains for Kripke and those who do, as a matter of fact, share the intuitions in question. Moreover, the anthropological evidence of cross-cultural variation does not
illuminate the fundamental question because, even if the Kripke-Putnam intuitions were universally shared, their credentials are not thereby established as guides to scientific or metaphysical claims.

## Who is in the Know?

Putnam $(1975,11)$ explains that internally identical "water" thoughts are said to have different meaning on Earth and Twin-Earth, although the chemistry of H 2 O or XYZ may never be discovered by people on either planet. That is, externalism depends on intuitions arising from the theorist's knowing the truth. Indeed, defending externalism, Burge (1988) confirms this diagnosis saying "We take up a perspective on ourselves from the outside." The conception of an "Omniscient Observer" is explicitly embraced as unproblematic by Donnellan (1974), a perspective Kaplan (2012, 156), too, has endorsed as "description from above." This is an understanding "in which one surveys another's thought" from a point of view "independent of whether the subject's thought corresponds to reality." These are remarkable confirmations of my diagnosis of the illusion of the "philosopher-narrators" omniscience.
In Crane's (1996) useful phrase, the question of who is "in the know" is central to untangling the intuitions at the heart of puzzles concerning externalism. Crane's question recalls Putnam's $(1981,50)$ question "From whose point of view is the story being told?" The invisibility of our own role and our own knowledge creates the illusion that it is the relational fact about how the world really is that determines the thought or belief in question. As Crane $(1996,293)$ notes, "the Twin Earth cases are meant to demonstrate that the world itself can, as it were, fix the meanings of some of our words." Crane's apt characterization captures the paranormal or clairvoyant conception of meanings which somehow link the mind directly with its objects in the world.

## Philosophers, autistics \& three year olds

Burge $(2012,119)$ recently explains the nature of de re belief in terms that are suggestive of other philosophical problems: "One can have a de re belief that is successfully referential and meets all other conditions on being de re, which nevertheless fails to count as knowledge." Consider the case in which someone is looking at a chair which he can see in a certain position apparently in the next room. However, he doesn't notice that he is looking at a large mirror and, therefore, sees the reflection of a chair that is actually nearer to him in the same room. As it happens, there is an identical chair in the next room behind the mirror, exactly where the reflection appears to be. It is evident that this circumstance is precisely Burge's scenario of de re belief and it is also exactly the Gettier (1963) case of justified, true belief that doesn't count as knowledge. Burge doesn't mention Gettier, but these parallels suggest the Problem has a wider interest beyond the epistemological issues it has been directly concerned with. Accordingly, it is interesting to notice Fodor's comment about the semantics
of mental representations applies to Gettier too: "we need it [broad or externally individuated content] to make sense of the fact that thoughts have the truth conditions that they do" (1994, 50). As if describing the Gettier Problem, in an entirely different context, Fodor gives a diagnosis that is apt for this puzzle:

It is, to put the point starkly, the heart of externalism that semantics isn't part of psychology. The content of your thoughts (/utterances), unlike for example, the syntax of your thoughts (/utterances), does not supervene on your mental processes. (Fodor 1994, 38)

In the Gettier case, too, the wide contents of your thoughts construed transparently as knowledge do not supervene on your mental processes, being merely justified beliefs. Fodor had made the same point where he said "truth, reference and the rest of the semantic notions aren't psychological categories" $(1980,253)$.

In response to the semantic orthodoxy, Farkas' (2003) argues that "external features are important only if they are incorporated into the internal cognitive or experiential perspective of cognizers." Schantz, too, explains, "As far as psychological explanation is concerned, what counts is how the world is internally represented as being, not how the world really is (2004, 23; emphasis added)." This is essentially the formula with which Fodor (1998, 20) characterized externalism, the view that "what you are thinking depends on what world you're in." This diagnosis of externalist semantic intuitions is precisely appropriate to the notorious Gettier (1963) Problem. In Chisholm's (1966) classical version, the subject sees a sheep-like bush and acquires a perceptual belief "There is a sheep in the field." Although this belief is justified by the evidence, it is true only by accident because, unbeknownst to him, there is a sheep elsewhere in the field. The classical criteria for knowledge - justified, true belief - appear to be met, but the belief does not count as knowledge. Hetherington (2012) has recently given an analysis of "Gettiered beliefs", being cases in which "truth remains essential." His diagnosis is that philosophers' intuitions are evidence of their "being infallibilists, without realizing this about themselves." This seems to be another way of making my point about puzzles that arise from the "narrator's" omniscience. Putnam's Twin Earth example, too, is a case in which mental content is ascribed to someone on the basis of truths that are not represented internally by the subject just as in the Gettier Problem.

Schantz' prescription for psychological explanation what counts is how the world is internally represented as being, not how the world really is - is apt also for capturing the mistaken "theory of mind" in a different domain. We see a striking analogue known to clinical psychologists in the Wimmer and Perner (1983) "false belief" task: Autistics and three year-olds ascribe beliefs to others based on their own knowledge of the truth rather than on the other's justified beliefs. Switching the candy when the character isn't looking in the experiments of Wimmer and Perner is
analogous to Gettier's substitution of bushes for sheep, or Fodor's substitution of shrews for mice in cases of misrepresentation. Putnam's substitution of XYZ for $\mathrm{H}_{2} \mathrm{O}$, like Dretske's (1986) disoriented microbes, are various ways that have been devised to make 'the world go wrong.' The truth-making facts are unconnected with the grounds for belief which are known only to the philosopher-narrator. By ascribing beliefs in this way, it appears that philosophers make the same mistake that autistics commit and children grow out of by the age of four. Ralph's belief about Ortcutt (Quine 1960) just like Twin Oscar's thoughts about water (Putnam 1975) and Pierre's thoughts of London (Kripke 1979) are essentially ascriptions of belief based on the philosophers' knowledge of the truth (see Slezak 2011).

## Obscurantist Intentional Magic

"Object-dependent" referential thoughts called de re are taken to be "singular thoughts" about a particular object or person that the speaker has in mind. This is the strong intuition expressed by Brian C. Smith that symbols somehow "reach out and touch someone" (1987, 215). Kripke has placed these issues in his framework of 'rigid designators' that denote the same individual in all 'possible worlds.' However, Stalnaker (2003) emphasizes that Kripke's claims rest on intuitive grounds, and poses a revealing question: "Doesn't this presuppose that the same individuals can be found in different possible worlds? Searle (1969, 93), too, argues that if an expression has no descriptive content as Kripke and 'direct reference' theorists claim, "then there could be no way of establishing a connection between the expression and the object." He asks "What makes this expression refer to that object?" Kripke's preferred answer is that a chain of historical, causal, connections back to a baptismal event fixes the reference. However, this account utterly fails to explain how a particular individual acquires the competent use of a name. The point is precisely analogous to Putnam's $(1967,18)$ attempt to rebut Chomsky's "innateness" claims by citing the common historical origin of all human languages. But this response fails to address the problem of language acquisition - the question of how each individual child must accomplish the task of becoming a competent speaker. The common origin of all human languages is irrelevant to this question, just as the supposed historical-causal chain is irrelevant to an individual's understanding and use of proper names. Stalnaker $(2003,178)$ captures the problem aptly, speaking of the only alternative to descriptive accounts which seems to be "some kind of obscurantist intentional magic." In Searle's $(1969,87)$ suggestive words, the idea that we can mean or intend a particular object and not another inclines us to think "that it is a movement of the soul." In the same vein, Putnam (1981) suggests that externalist intuitions are a "magical theory of reference" that assumes occult "noetic rays" connecting words with their referents. Indeed, these referential intuitions are suggestive of widely held, compelling misconceptions concerning visual perception that are thought to involve emanations
from the eyes - the so-called "extramission theory of perception" maintained by early Greek philosophers. Remarkably, following Piaget, Winer et al. (2002) report evidence that belief in extramission remains widespread, deeply ingrained and resistant to educational efforts. I don't mean to suggest that such theories are literally believed by philosophers, but the compelling conceptions are very suggestive of intuitions underlying the most widely held externalist semantic theories.

## References

Almog, J. 2005. Replies. Philosophy and Phenomenological Research, 70, 3, 717-730.
Bealer, G. 1998. Intuition and the Autonomy of Philosophy, in M.R. DePaul and W.M. Ramsey eds., Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry. Lanham: Rowman and Littlefield.
Boghossian, P. 1998. What the Externalist Can Know A Priori, in C. Wright, B.C. Smith \& C. Macdonald eds., Knowing Our Own Minds, Oxford: Oxford Univ Press.
Brandom, R. 1994. Making It Explicit, Cambridge, Mass.: Harvard University Press.
Burge, T. 1988. Individualism and Self-Knowledge, The Journal of Philosophy, 85, 649-63.
Burge, T. 2012. Referring De Re, in J. Almog \& P. Leonardi eds., Having in Mind: The Philosophy of Keith Donnellan, Oxford: Oxford Univ Press.
Chalmers, D. 2002. Does Conceivability Entail Possibility? In T. Gendler and J. Hawthorne eds., Conceivability and Possibility, Oxford: Oxford Univ Press, 145-200.
Chisholm, R.M. 1966. Theory of Knowledge. Englewood Cliffs: Prentice-Hall.
Chomsky, N. 1962. Explanatory Models in Linguistics, in E. Nagel. P. Suppes \& A. Tarski, eds., Logic, Methodology and Philosophy of Science, Stanford Univ Press, 528-550.
Chomsky, N. 1992. Language and Interpretation: Philosophical Reflections and Empirical Inquiry, in J. Earman, ed., Inference, Explanation, and Other Frustrations, Berkeley: Univ. of California Press, 99-128.
Chomsky, N. 1996. Power and Prospects, Sydney: Allen \& Unwin.
Chomsky, N. 2000. New Horizons in the Study of Language and Mind, Cambridge, Cambridge Univ Press.
Chomsky, N. 2012. The Science of Language, J. McGilvray ed. Cambridge: Cambridge Univ Press.
Crane, T. 1996. All the Difference in the World, in A. Pessin \& S. Goldberg eds., The Twin Earth Chronicles, New York: M.E Sharpe, 1996.
Cummins, R. 1998. Reflection on Reflective Equilibrium, in M.R. DePaul and W.M. Ramsey eds., Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry. Lanham: Rowman and Littlefield.
DePaul, M.R. \& Ramsey W. eds., 1998. Rethinking Intuition, New York: Rowman and Littlefield.
Devitt, M. 1984. Thoughts and their Ascription, Peter A. French, Theodore E. Uehling and Howard K. Wettstein
eds., Midwest Studies in Philosophy, Vol. IX, 385-420: Minneapolis: Univ of Minnesota Press.
Devitt, M. 2011. Experimental Semantics, Philosophy and Phenomenological Research, 82, 2, 418-435.
Donnellan, K.S. 1974. Speaking of Nothing, The Philosophical Review, 83, 1, 3-31.
Dretske, F. 1986. Misrepresentation. In R.J. Bogdan ed., Belief: Form, Content and Function. Oxford: Oxford University Press.
Dummett, M. 1973. Frege: Philosophy of Language, New York: Harper \& Row.
Egan, F. 1999. In Defence of Narrow Mindedness, Mind \& Language, 14, 2, 177-194.
Farkas, K. 2003. Does Twin Earth Rest on a Mistake? Croatian Journal of Philosophy, III/8, 155-69.
Fodor, J.A. 1987a. Individualism and Supervenience, in in A. Pessin \& S. Goldberg eds., The Twin Earth Chronicles, New York: M.E Sharpe, 1996.
Fodor, J.A. 1987b. Psychosemantics: The Problem of Meaning in the Philosophy of Mind, Cambridge, Mass.: MIT Press.
Fodor, J.A. 1998. Concepts: Where Cognitive Science Went Wrong. Oxford: Oxford University Press.
Fodor, J.A. 2004. Water's Water Everywhere, London Review of Books, 21, October.
Fodor, J.A. 2008. LOT2: The Language of Thought Revisited, Oxford: Oxford University Press.
Gettier, E.L. 1963. Is Justified True Belief Knowledge? Analysis, 23, 121-3.
Hetherington, S. 2012. The Gettier Illusion, Synthese, 188, 2, 217-230.
Hinitkka, J. 1999. The Emperor's New Intuitions, The Journal of Philosophy, 96, 3, March, 127-147.
Hughes, C. 2004. Kripke: Names, Necessity and Identity, Oxford: Clarendon Press.
Jackson, F. 1998. From Metaphysics to Ethics, Oxford: Clarendon Press
Jackson, F. 2010. Language, Names and Information, Malden: Wiley-Blackwell.
Kaplan, D. 2012. An Idea of Donnellan, in J. Almog \& P. Leonardi eds., Having in Mind: The Philosophy of Keith Donnellan, Oxford: Oxford Univ Press.
Kripke, S. 1972. Naming and Necessity, reprinted 1980, Cambridge: Harvard University Press.
Kripke, S. 1979. A Puzzle About Belief, in A. Margalit ed., Meaning and Use, Dordrecht: Reidel, 239-283.
Machery, E., Mallon, R, Nichols, S, Stich, S.P. 2004. Semantics, cross-cultural style, Cognition, 92, B1-B12.
Nagel, J. 2012. Intuitions and Experiments: A Defense of the Case Method in Epistemology, Philosophy and Phenomenological Research, xx, xx, pp-pp.
Pietroski, P.M. 2003. The Character of Natural Language Semantics in A. Barber ed., Epistemology of Language, Oxford: Oxford University Press.
Putnam, H. 1967. The 'Innateness Hypothesis' and Explanatory Models in Linguistics, Synthese 17, 12-22.

Putnam, H. 1975. The Meaning of 'Meaning'. In K. Gunderson ed. Language, Mind and Knowledge: Minnesota Studies in the Philosophy of Science, Volume 7, Minneapolis: University of Minnesota Press.
Putnam, H. 1981. Reason, Truth and History, Cambridge: Cambridge University Press.
Pylyshyn, Z. 2003. Seeing and Visualizing: It's Not What You Think, Cambridge, Mass.: MIT Press.
Quine, W.V. 1960. Word and Object. Cambridge, Mass.: MIT Press.
Salmon, N. 2011. A Note on Kripke's Puzzle about Belief, in A. Berger ed., Saul Kripke, Cambridge: Cambridge University Press.
Schantz, R. ed. 2004. The Externalist Challenge, New York: Walter de Gruyter.
Searle, J. 1969. Speech Acts: An Essay in the Philosophy of Language, Cambridge: Cambridge University Press
Segal, G.M.A. 2000. A Slim Book About Narrow Content, Cambridge, Mass.: MIT Press.
Segal, G.M.A. 2004. Reference, Causal Powers, Externalist Intuitions and Unicorns, in R. Schantz eds., The Externalist Challenge, New York: Walter de Gruyter.
Slezak, P. 2011. Philosophers, Three Year Olds and Autistics, in L. Carlson, C. Hoelscher \& T. F. Shipley. Proceedings of the 33rd Annual Meeting of the Cognitive Science Society, Boston, Massachusetts. Austin, TX: Cognitive Science Society.
Smith, Brian C. 1987. The Correspondence Continuum. CSLI Report 87-71.
Soames, S. 2005. Reference and Description: The Case Against Two-Dimensionalism, Princeton: Princeton University Press.
Sosa, D. 2001. Rigidity in the Scope of Russell's Theory, Nous, 35,1, 1-38.
Stalnaker, R. 2003. Ways A World Might Be, Oxford: Oxford University Press.
Stich, S. 2012. Do Different Groups Have Different Epistemic Intuitions? A Reply to Jennifer Nagel, Philosophy and Phenomenological Research, xx xx 1-28.
Tversky, A. \& Kahneman, D. 1974. Judgment under Uncertainty: Heuristics and Biases, Science, 185, 11241130.

Vargas Llosa, M. 1975. La Orgia Perpetua, Translator Helen Lane, New York: Farrar Straus Giroux.
Wettstein, H. 2004. The Magic Prism: An Essay in the Philosophy of Language, Oxford: Oxford University Press.
Wimmer, H. and Perner, J. 1983. Beliefs About Beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition, 13, 103-28.
Winer, G.A., Cottrell, J.E., Gregg, V., Fournier, J.S. \& Bica, L.A., 2002. Fundamentally Misunderstanding Visual Perception: Adults' Belief in Visual Emissions, American Psychologist, 57, 6/7, 417-424.

# Physical predictions over time 

Kevin A Smith (k2smith@ucsd.edu), ${ }^{1}$ Eyal Dechter (edechter@mit.edu), ${ }^{2}$ Joshua B Tenenbaum (jbt@mit.edu), ${ }^{2}$ Edward Vul (evul@ucsd.edu) ${ }^{1}$<br>1. University of California, San Diego, Department of Psychology, La Jolla, CA 92093<br>2. MIT, Department of Brain and Cognitive Sciences, Cambridge, MA 02139


#### Abstract

In order to interact with the world, people must be able to predict how it will unfold in the future, and these predictions must be updated regularly in light of new information. Here we study how the mind updates these predictions over time. Participants were asked to make ongoing predictions about the destination of a simulated ball moving on a 2 D bumper table. We modeled these decisions by assuming people simulate the world forward under uncertainty. This model fit participants' behavior well overall, suggesting that people continuously update their physical simulations to inform their decisions. In some specific scenarios participants' behavior is not fit well by the simulation based model in a manner suggesting that in certain cases people may be using qualitative, rather than simulation-based, physical reasoning.


Keywords: intuitive physics; forward simulation

## Introduction

Changing lanes while driving seems like a simple and ordinary task - millions of people do it everyday. But to do so safely requires sophisticated predictions. Drivers must judge where their own car and those around it will be during the lane change, and, crucially, they must update these predictions with new information: if a car in the adjacent lane accelerates, a driver may abort her lane change to avoid a collision.

This scenario demonstrates how people typically plan their actions: prediction is updated as new information is gathered. Research spanning decades has investigated how people predict future object movement while objects are hidden (Faisal \& Wolpert, 2009; Rosenbaum, 1975; Runeson, 1975; Smith \& Vul, 2013; Téglás et al., 2011), but in most natural cases, observers continue to see objects while updating their predictions. In this study we investigate how people change their instantaneous predictions about objects over time: are ongoing predictions the result of online simulation?

Recent research provides evidence that people use 'Noisy Newtonian' models of physics to simulate the world (Sanborn, Mansinghka, \& Griffiths, 2013): peoples' internal physical models are based on correct assumptions about physics, but uncertainties in object position, movement, and latent variables can cause biases and variability in prediction. This framework has been used to predict peoples' judgments about the stability of a tower of blocks (Hamrick, Battaglia, \& Tenenbaum, 2011), the movement of hidden objects (Smith \& Vul, 2013), and even judgments about physical causality (Gerstenberg, Goodman, Lagnado, \& Tenenbaum, 2012). These works, however, solicited predictions at single instances in time. In this paper, we
investigate whether a model that assumes faithful physics under uncertainty is also consistent with how peoples' predictions evolve over time. We show that people's decisions are often consistent with online forward simulation, but we also find that people can use qualitative reasoning about the world (e.g., Forbus, 1994) when this is more informative than simulations.

## Experiment

We asked participants to play a game in which they make predictions about the path of a ball bouncing around a computerized table. The ball can reach one of two targets on the table, and participants earn points for predicting which target it reaches first. Crucially, they make this prediction continuously throughout the trial, earning points while predicting the correct target but losing points while predicting the incorrect target. In this way, we could capture how uncertainty (decisions whether to choose a target) and choices (which target) evolved over the course of each trial.

## Methods

Sixty-six UC San Diego undergraduates participated in this experiment for course credit. ${ }^{1}$

On each trial, participants saw a ball moving around a 'table' on the computer screen that contained blocks and both a red and a green target. The ball bounced perfectly elastically off of the edge of the table and blocks, ending when the ball reached one of the two targets. While the trial progressed, participants were asked to predict whether the ball would hit the red target or the green target first, indicating their guess by holding down either the ' $z$ ' or the ' $m$ ' key (each key counterbalanced for red and green between participants). If they were unsure, participants could press neither key, and if their prediction changed midtrial, they were encouraged to switch keys. Holding down a key would fill a bar of the associated color, and at the end of the trial, the score would be determined by the difference between the proportion of time the keys for each target were held down:
(1) Score $=20+100 *($ Prop $($ Correct $)-\operatorname{Prop}($ Incorrect $))$

After each trial, participants were notified of their score and could continue to the next trial by pressing the spacebar.

Participants were each given the same 400 trials in a random order. Of these, 370 trials were randomly generated, and 30 were designed to consider various extreme scenarios.

[^209]Of special note in the hand-crafted trials are five trials in which the configuration of the walls made it impossible for the ball to ever reach one of the two targets; these are called the 'qualitative' trials as they were meant to differentiate between simulation-based intuitive physics and a qualitative assessment of the table configuration.

Each trial lasted between 2.0 s and 10.2 s . Target colors were randomly swapped for each trial to avoid color bias effects. ${ }^{2}$ Responses were polled and recorded once every tenth of a second.


Figure 1: Illustration of an ongoing trial. Participants would see the ball travel along the dotted line and would predict green or red during its motion (neither blue line was visible).

## Results

We analyzed participants' aggregate performance across trials via their total score (eq. 1). Participants showed low variability in their average trial scores (mean $=56.0$, $\mathrm{sd}=$ 5.1) and scores for each trial were very consistent across participants (split half correlation, $r=0.96$ ).

We investigated whether there were surface-level features of trials that make them more or less difficult. Here we use average score as a proxy for difficulty; high scores indicated that most participants could accurately predict the path of the ball easily (and early), while low scores indicated uncertainty and mis-prediction.

The features we considered as possible predictors included: (1) trial duration, (2) the number of blocks on the table, (3) the number bounces before the ball hit the target, (4) the initial deviation of the ball's path from a horizontal or vertical direction, (5) the proportion of the table clear of walls or targets, (6) the ratio of the area of the correct target to the incorrect target, (7) the ratio of the average distance of the ball to the correct target versus the incorrect target, and (8) the closest the ball ever was to the incorrect target.

We found four predictive features: trials were easier when their trajectory was on average closer to the final target, involved fewer bounces, when the initial motion was along a cardinal direction, and when the ball never approached the incorrect target. These four predictors together explained $31.9 \%$ of the variance in scores across trials.

[^210]Table 1: Predictors of trial difficulty. Partial correlation is correlation between predictor and trial score accounting for all other predictors.

| Predictor | $r$ | $\mathrm{r}_{\text {partial }}$ |
| :--- | ---: | ---: |
| (7) Distance ratio | -0.51 | -0.24 |
| (3) Bounces | -0.25 | -0.19 |
| (4) Direction deviation | -0.21 | -0.14 |
| (8) Nearness to incorrect | 0.42 | 0.09 |

Although, aggregate metrics of ball trajectory accounted for some of the variation in difficulty across trial, such an analysis fails to capture the rich predictions individuals make over time and how those change in light of the details of a given table configuration and ball trajectory.

To further delve into how people make online physical predictions and explain why some of these features might make trials more difficult, we compared human behavior to predictions made via stochastic physical reasoning.

## Physical Prediction Model

## Description

The model we used to predict behavior on this task has two parts: the physical simulator, which provides possible paths that the ball can take, and the decision policy, which uses the output of the physical simulations to decide which target to choose (if either).

Physical simulator The part of the model that simulates the trajectory of the ball is based in large part on the model of Smith \& Vul (2013). This model assumes that people base their physical models on real physics but must incorporate uncertainty about the world into their physical judgments.

This model captures two sources of uncertainty: 1) perceptual uncertainty arises from the noisiness of inferring the position and movement of objects, and 2) dynamic uncertainty is uncertainty about the roughness and elastic properties of the table and walls that could cause the ball's path to deviate from idealized Newtonian physics over time.

The physical simulator produces 500 simulation paths ${ }^{3}$ every tenth of a second for each trial to replicate the polling frequency in the experiment. These simulation paths were produced using the same uncertainty parameters and fits as Smith \& Vul (2013). ${ }^{4}$ Each path terminates when the simulated ball reaches either the red or the green target, or when 10 seconds of simulated time has passed. ${ }^{5}$

[^211]The physical simulator outputs a set of proportions: how many paths reached the green target, how many reached the red target, and how many did not reach a target within the maximum simulation time ('uncertain' paths).

Decision policy The decision policy takes the output of the physical simulator and assigns belief to two decisions: 1) is there sufficient certainty about which target the ball will hit to offer any guess at all? (analogous to participants' decision whether or not to press any button at all), and if so, 2) which target should be guessed?

The decision of whether any prediction should be made is based on the proportion of simulation paths that reached either target. These are combined and fit with two parameters: a parameter $\alpha$ representing a Luce choice softmax weighting, and a parameter $\gamma$ to capture a bias towards making any guess at all:

$$
\begin{equation*}
P(\text { Any })=\frac{\text { Sim }(\text { Red or Green })^{\alpha}+\gamma}{\text { Sim }(\text { Red or Green })^{\alpha}+\text { Sim }(\text { Uncertain })^{\alpha}+\gamma} \tag{2}
\end{equation*}
$$

Conditioned on the decision to make any guess at all, the decision whether to guess red or green is based on the relative proportion of simulated paths that reached each color target. Here, there is a single Luce choice soft-max weighting parameter ( $\beta$ ), but because experimental trial colors were randomized, we assumed no bias:

$$
\begin{equation*}
P(\text { Red } \mid \text { Any })=\frac{\operatorname{Sim}(\text { Red })^{\beta}}{\operatorname{Sim}(\text { Red })^{\beta}+\operatorname{Sim}(\text { Green })^{\beta}} \tag{3}
\end{equation*}
$$

Finally, we assume there is a decision offset: people cannot immediately use simulation information, but instead must take time to process it, come to a decision, and move their hand to push a button. Thus we fit a single parameter $t$ to determine how long the model should wait to use simulation information. ${ }^{6}$

These four parameters were optimized to fit the empirical data at each polling time point for each trial. At each point, we created a vector of the empirical probabilities of pressing each of the two keys at time point i of trial $\mathrm{j}:\left[\operatorname{prop}(\operatorname{Red})_{i j}\right.$, $\left.\operatorname{prop}(\text { Green })_{i j}\right]$. We then calculated a similar vector of model predictions: $\left[P(\text { Red })_{i j}, P(\text { Green })_{i j}\right]$. The parameters were fit to minimize the total Euclidean distance between these points over all time points of all trials:
(4) $\sum_{j} \sum_{i} \sqrt{\left(\text { prop }(\text { Red })_{i j}-p(\text { Red })_{i j}\right)^{2}+\left(\text { prop }(\text { Green })_{i j}-p(\text { Green })_{i j}\right)^{2}}$

## Model performance

How do predictions change over time? Participants often changed their decisions as trials progressed, either from uncertain to certain or one color to the other. We first ask whether we can capture changes in participants' predictions over time. We compared model decision probabilities to the distribution of participants' choices at each time point: what

[^212]proportion of participants guessed the ball would end in the red target or the green target, or were too uncertain to offer a guess. Although there were differences in individual predictions, we believe aggregation is appropriate given the low variability in total scores and high consistency within trials (split half correlation, $\mathrm{r}=0.96$ ).


Figure 2: Joint histogram of model (x-axis) and human (yaxis) decisions. (Top) The probability of making any guess (pushing a button). (Bottom) The probability of choosing 'red' given a decision. Colors indicate log-frequency of time points in each bucket, with hotter colors indicating more observations. Observations along the diagonal indicate the model is accurately capturing the exact proportions of participants making decisions.

If people make decisions based on similar uncertainty and decision policies to those of the model, then the model should be able to predict both (a) when people make any decision and (b) which choice they make (red or green) when they do. Figure 2 shows the correlation between model predictions and participants' behavior. In the top panel, we see that participants' decisions whether to push either button are well predicted by the model ( $\mathrm{r}=0.84$ ): at most time points either our model believes that no guess should be made and nearly no participants offer a guess
(bottom-left), or our model believes that a guess should definitely be made and nearly all participants offer a guess (top-right). Moreover, even when participants are not unanimous in their decision to offer a guess, the model captures the variation in the proportion of people making guesses. In the bottom panel, we see that the choice people make (red or green) at time points when they do offer a guess, is well explained by the model $(r=0.92)$ : again much of the time participants are nearly unanimous in their choice of one of the targets, as is the model. But again the model also captures the gradations in beliefs when participants are split on which target to choose

What makes trials difficult? We also wanted to know if we could better explain what makes trials easier or harder. To do so, we calculated the average model score for each trial in an equivalent way to participants' scores:
(5) ModelScore $=20+100 * \sum_{t}\left(P(\text { correct })_{t}-P(\text { incorrect })_{t}\right)$

The model predicts the difficulty of the trials better $\left(\mathrm{R}^{2}=\right.$ 0.675 ; see Figure 3) than using superficial trial features, as we first investigated $\left(\mathrm{R}^{2}=0.319\right)$. It is noteworthy that our model was never explicitly informed about how the trial would unfold, which characteristics should make a trial more difficult, or even how scoring works; nor was the model fit to capture trial scores. Instead simply by considering variations in moment-by-moment physical predictions, we could capture variation in trial difficulty.


Figure 3: Modeled versus empirical trial scores. Each point represents a single trial, where bars are $95 \%$ confidence intervals on empirical scores.

There remains reliable variability in participants' average scores that is not explained by the model. To investigate what might be causing this, we again predicted participants’ scores on each trial, using the same features that we had before, but physical model's score included as a predictor. With the model score added, no new features became significant predictors (indicating that the model is unbiased with respect to those features), and two features - the number of bounces and the smallest distance to the incorrect target - were no longer good predictors (indicating that the model accounts for these difficulties well). Two features did remain though: the deviation of the path from the horizontal or vertical, and the ratio of the average distance of the ball
to each of the targets. Including these two predictors did provide a statistically significantly better fit $(F(2,393)=12.9$, $p<0.001$ ), but only explained slightly more variability in participants' scores $\left(\mathrm{R}^{2}=0.695\right)$.

The remaining feature predictors inform us about what aspects of human cognition the model is not capturing. First, the model does not capture the additional difficulty introduced when the ball is traveling at an angle. The physical model assumes that directions of movement are equally difficult to simulate, but this indicates that people may find it difficult to predict the path of objects travelling at an angle, in the same way people have difficulty discriminating oblique motion (Matthews \& Quin, 1999). The ratio of the distances between the targets also remains as a predictor, but the correlation is attenuated as compared to the relationship without the model predictions, perhaps suggesting that there is a bias to believe the ball will hit the nearer target beyond simulations.

Table 2: Predictors of trial difficult including the physical model. Partial correlation is correlation between predictor and trial score after all other predictors have been included. Prior partial correlation is partial correlation of prediction not including the physical model (see Table 1)

| Predictor | $r$ | $\mathrm{r}_{\text {partial }}$ | Prior $\mathrm{r}_{\text {partial }}$ |
| :--- | ---: | ---: | ---: |
| Physical model | 0.82 | 0.65 | $\mathrm{~N} / \mathrm{A}$ |
| (4) Direction deviation | -0.21 | -0.17 | -0.14 |
| (7) Distance ratio | -0.51 | -0.15 | -0.24 |

Why do human and model predictions differ? Participants' predictions were overall consistent with the model, but this fit varied by trial. To explain how people might be consistent with or deviate from simulation using noisy physics, we investigated how well the model fit empirical data on each trial. Our metric of trial fit was the average deviation between participants' decisions and model predictions over the trial, similar to eq. 4:
(6) Dev $_{j}=\sum_{i} \sqrt{\left(\text { prop }(\text { Red })_{i j}-p(\text { Red })_{i j}\right)^{2}+\left(\text { prop }(\text { Green })_{i j}-p\left(\text { Green }_{i j}\right)^{2}\right.} / t_{j}$


Figure 4: Model deviation from empirical decisions as a function of trial difficulty.

Well-fit trials had lower average deviations, whereas poor fits were characterized by high deviation. As can be seen in

Figure 4, the model predicted participants' decisions better on trials that the participants found easy, with a slight reduction in performance as the trials became more difficult.

Many of the highest scoring and best fitting trials were straightforward, as people quickly decided on the correct target, and the model captured this behavior. We first review two trials with average model fit (the green points, Figure 4) to discuss the strengths of the model, then review where the model deviates from human performance: two trials that people find difficult but the model does not (the purple points), and the five 'qualitative' trials that the model finds difficult but people do not (the red points). Information on all other trials can be found online at experiments.evullab.org/physovertime/trials.html


Figure 5: (Left) Image of average fitting trials. Numbers represent the time in seconds when the ball passes that point. (Right) Associated proportion of red, green, or undecided decisions by participants (top) or the model (bottom). Time on the x -axis is matched to the associated point in the trial diagram. The model tends to capture human behavior, erring only in confidence (top trial) or timing (bottom trial).

Moderately difficult trials were those in which the correct target was not immediately obvious, requiring participants to either resolve their belief over time (see Figure 5-top), or change their beliefs when more information arrived (see Figure 5-bottom). The model predicted these types of decisions well, typically erring only in either the amount of uncertainty or the timing of decision changes. This suggests that in aggregate, the model captures the way that people resolve uncertainty: certainty increases as the ball travels or when people see the outcome of a bounce.

There were also trials that participants did poorly on that the model did not fit well. These were typically trials where (a) the ball was traveling at a steep angle, and (b) small changes in the perceived layout of the table could cause a difference in the ending target. For instance, the top trial in Figure 6 was the trial for which the model performed the worst. Here, if the wall just below the green target were
slightly larger, the ball would miss that target and hit red. The model assumes perfect knowledge of the table, but it is likely that people have uncertainty about the area of the bumpers and target-areas as well - uncertainty that the model does not have.


Figure 6: Image of trial path (left) and associated empirical and model predictions (right) for difficult, poor-fitting trials. Slight changes in the layout of the table would have significant consequences for both trials, which the model does not capture.

We also specifically created 'qualitative' trials in which one target was difficult for the ball to get to, but the other was unreachable (see Figure 7). These trials comprised a large portion of the trials that the model found difficult but participants found easy. If participants are deciding between the two targets based solely on the output of a physical simulator, then they should show large amounts of uncertainty until near the end of the trial. On the other hand, if people can qualitatively analyze the structure of the table and determine that one target is unreachable, then they should quickly show high confidence for the possible but unlikely target. On these trials, participants tended to be much more confident than model predictions starting early in the trial, suggesting that they were performing a qualitative, topological analysis of the possible trajectories and outcomes on the table.

## Discussion

In this study we demonstrated that human online predictions about the world can be well captured by a model that continually simulates the world forward using noisy physical principles. This physical simulation model could better predict which trials were easy and which were difficult than a simple analysis of trial features. Likewise, it well predicted how often people would decide on one of the two target, and which target they would decide on. Together, these results suggest that people are performing forward physical simulation online as the trial unfolds.

Furthermore, we found aspects of human behavior that the model missed, for instance difficulty with balls traveling at oblique angles or uncertainty about the state of the table. These factors are consistent with a forward simulation model, and suggest refinements to our knowledge of how people simulate object movement.

The qualitative trials, on the other hand, are inconsistent with a pure simulation account. On these trials, simulations would have difficulty reaching the incorrect target, and thus pure simulation would predict uncertainty throughout the trial, as our model does. Instead, people quickly grow certain, reasoning that if the ball cannot reach one target, it must eventually reach the other. This suggests that people may use qualitative spatial reasoning (e.g. Forbus, 1994)


Figure 7: Image of trial path (left) and associated empirical and model predictions (right) for qualitative trials.
Participants can quickly discover that the ball cannot reach one of the two targets and select the other, but the model must use simulations to make this determination.
when it provides a clear answer about their environment, but otherwise use simulations to reason about the future. Further investigation is required to understand how and when people decide to switch between modes of prediction.

Finally, our model made predictions by generating a new set of simulated paths at each time step. However, people probably conserve computation and only slightly update a single set of simulations at each time step (e.g., using a particle filtering algorithm). Details of the exact prediction algorithm are a fruitful area for future research.

We can predict the future state of the world and integrate new information to make better predictions. This study suggests these predictions often come from continuously updating our vision of how the world will unfold, but also leaves tantalizing clues that we can overlay our simulations with qualitative reasoning about our environment as well.

## Acknowledgments

KS and EV were supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of the Interior (DOI) contract D10PC20023. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI, or the U.S. Government.

## References

Faisal, A. A., \& Wolpert, D. M. (2009). Near optimal combination of sensory and motor uncertainty in time during a naturalistic perception-action task. Journal of Neurophysiology, 101, 19011912.

Forbus, K. D. (1994). Qualitative spatial reaonsing: Framework and frontiers.
Gerstenberg, T., Goodman, N., Lagnado, D. A., \& Tenenbaum, J. B. (2012). Noisy Newtonians: Unifying process and dependency accounts of causal attribution. Paper presented at the Proceedings of the 34th Annual Meeting of the Cognitive Science Society, Sapporo, Japan.
Hamrick, J., Battaglia, P., \& Tenenbaum, J. (2011). Internal physics models guide probabilistic judgments about object dynamics. Paper presented at the Proceedings of the 33rd Annual Meeting of the Cognitive Science Society, Boston, MA.
Matthews, N., \& Quin, N. (1999). Axis-of-motion affects direction discrimination, not speed discrimination. Vision Research, 39, 2205-2211.
Rosenbaum, D. A. (1975). Perception and extrapolation of velocity and acceleration. Journal of Experimental Psychology: Human Perception and Performance, 1(4), 395.
Runeson, S. (1975). Visual prediction of collision with natural and nonnatural motion functions. Perception and Psychophysics, 18(4), 261-266.
Sanborn, A. N., Mansinghka, V. K., \& Griffiths, T. L. (2013). Reconciling intuitive physics and Newtonian mechanics for colliding objects. Psychological Review, 120(2), 411-437.
Smith, K. A., \& Vul, E. (2013). Sources of uncertainty in intuitive physics. Topics in Cognitive Science, 5(1), 185-199.
Téglás, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., \& Bonatti, L. L. (2011). Pure reasoning in 12-month-old infants as probabilistic inference. Science, 332, 1054-1059.

# Linguistic structure is an evolutionary trade-off between simplicity and expressivity 

Kenny Smith (kenny@ling.ed.ac.uk), Monica Tamariz \& Simon Kirby<br>Language Evolution and Computation Research Unit, School of Philosophy, Psychology \& Language Sciences, University of Edinburgh, Dugald Stewart Building, 3 Charles Street, Edinburgh, EH8 9AD, UK


#### Abstract

Language exhibits structure: a species-unique system for expressing complex meanings using complex forms. We present a review of modelling and experimental literature on the evolution of structure which suggests that structure is a cultural adaptation in response to pressure for expressivity (arising during communication) and compressibility (arising during learning), and test this hypothesis using a new Bayesian iterated learning model. We conclude that linguistic structure can and should be explained as a consequence of cultural evolution in response to these two pressures.


Keywords: language; structure; cultural evolution; learning; communication

## Introduction

Human language is unique among the communication systems of the natural world in that it is exhibits a rich combinatorial and compositional structure: language provides a generative system for productively combining meaningless elements (e.g. speech sounds) to form meaning-bearing units (morphemes), which are further recombined to yield complex units (phrases) whose meaning is derived in a predictable manner from the meaning of their component parts and their manner of composition. This allows us massive expressive potential: at least at a first approximation, anything you can think you can express in language. No other species has a communication system providing anything approaching this expressive power: why do humans?

One explanation for the presence of structure ${ }^{1}$ in human language appeals to biological evolution under natural selection (Pinker \& Bloom, 1990): language is fundamentally a biological trait, being underpinned by some innate languagespecific apparatus; the ability to communicate propositions which a structured language provides is adaptive, since it facilitates social interaction and ultimately increases fitness; therefore, structure in language represents a biological adaptation to facilitate communication. A second account explains structure in language as a consequence of cultural, rather than biological, evolution (Christiansen \& Chater, 2008). Rather than language structure reflecting an evolved domain-specific learning apparatus, the idea is that languages have adapted over repeated episodes of learning and production (a process sometimes called iterated learning) in response to weaker, domain-general constraints arising from the biases of language learners. We have previously termed this evolutionary process cultural selection for learnability (Brighton, Kirby, \& Smith, 2005). A range of models and experiments show

[^213]that cultural selection for learnability leads to the evolution of structure, under certain assumptions about the nature of transmission and the biases of language learners. Under a strong interpretation of this account, language's function for communication could be seen as an epiphenomenon: structured language provides a powerful medium for communication, but language structure is not 'for' communication.

Here we present a new model of iterated learning, motivated by recent experimental work, which goes some way to reconcile these two viewpoints. We draw from the biological account the insight that the alignment between language's apparent function as a system for expressing propositions and its structure, tailor-made for just such a purpose, is unlikely to be a fortuitous coincidence. We draw from cultural evolutionary account two insights: 1) biological evolution is not the only evolutionary process which might act to shape language: cultural evolution, a necessary consequence of the fact that language is socially learned, is a second such mechanism; 2) selection for learnability will impact on language during its transmission. This model suggests that structure arises from cultural evolution when language is under pressure to be expressive and learnable: pressure for expressivity arises from language use in communication, language learning by naive individuals introduces a pressure for simplicity arising from domain-general preferences for compressibility in learning. Crucially, both must be in play: pressure for expressivity or simplicity alone does not lead to structure. Structure in language is a linguistic adaptation, not a biological adaptation, in response to competing pressures for expressivity and learnability (Kirby, Cornish, \& Smith, 2008; Steels, 2012).

## Cultural evolution of structure: previous work

Models of the cultural evolution of linguistic structure typically emphasise the role of learnability constraints in driving the evolution of compositionality. Specifically, language is under pressure to be compressible: to allow the formation of compressed mental representations, i.e. simple grammars. This pressure for compressibility is inherent in learning (Chater \& Vitanyi, 2003), and can be amplified by other constraints acting on language transmission (e.g. the mismatch between the infinite expressivity of languages and the finite set of data from which such languages must be learned). Learning and transmission therefore favour languages which admit to compressed representations, i.e. which permit generalisations. Recursive compositionality is one such generalisation (e.g. Steels, 1998; Kirby, 2002; Brighton, Smith, \& Kirby, 2005), and therefore represents an adaptation by lan-
guage in response to pressures inherent in transmission and learning. However, compressibility is not the only constraint on learnability in these models: they typically include some learner bias in favour of languages which embody a one-toone mapping between meaning and form (which happen to be communicatively functional mappings), e.g. by implementing indirect (Kirby, 2002) or direct (Steels, 1998) competition between meanings which map to a single form. Brighton, Smith, and Kirby (2005) show that, if this bias against one-toone mappings is absent, the pressure for compressibility acting in isolation leads to degenerate, not structured, languages, where all meanings map to a single maximally-ambiguous form. While this might suggest that such a one-to-one bias would be adaptive, Smith (2004) shows that such a bias is unlikely to evolve for its (eventual) communicative payoff, and concludes that the one-to-one bias must be a product of domain-general cognitive biases. Again, this suggests that the utility of language for communication might be a side-effect of learnability pressures alone.

Diffusion-chain experiments with adult human participants have also been used to investigate the impact of cultural selection for learnability. Kirby et al. (2008) report two experiments in which participants are trained on a miniature language which provides labels for objects (coloured moving shapes, e.g. a red square bouncing), and are then prompted to produce labels for a further set of objects. Participants are organised into a diffusion chain, such that the labels produced by the $n$th participant in a given chain provide the training data for participant $n+1$ in that chain. The first participant in each chain is trained on a unstructured holistic system, where each object is associated with a unique random label (and therefore shared elements of meaning do not map to shared components of form).

Across two experiments, Kirby et al. (2008) show that languages change as a result of their transmission to be more learnable: the languages produced later in a chain of transmission are learnt with greater accuracy. In their Experiment 1 , this is achieved by the languages becoming simple: the languages lose distinctions. In the most extreme case, this results in a degenerate language in which all objects (with one exception) are associated with a single, highly-ambiguous label. Simplification facilitates learning at the expense of expressivity: while the initial holistic languages have high expressive potential, the languages which ultimately emerge allow only a few contrasts between objects to be signalled linguistically. However, there is no pressure for expressivity in this experiment: the language is under pressure to be learnable, but given the lack of a communicative task, under very little pressure to provide distinct labels for distinct objects.

In their Experiment 2, an artificial pressure for expressivity was introduced: homonyms (labels paired with multiple objects) were eliminated during the process of sampling from the $n$th participant's productions to yield the training data for participant $n+1$. As in Experiment 1, the languages became more learnable, but this was achieved by the development of
compositional structure: colour, shape and motion came to be encoded in separate 'morphemes' of multi-morphemic words. These structured languages are both learnable and expressive, allowing all distinctions between objects to be encoded linguistically.

Garrod, Fay, Lee, Oberlander, and MacLeod (2007) present a task in which participants are required to communicate a set of pre-specified concepts using drawings. Participants who repeatedly play the game together develop an expressive system of symbol-like graphical representations to communicate these concepts. This system of communication is holistic: each symbol is an idiosyncratic, stand-alone entity. Theisen-White, Kirby, and Oberlander (2011) present a modified version of this paradigm, integrating the dyadic context for communication with the diffusion-chain method from Kirby et al. (2008). An initial pair play a variant of the communication game from Garrod et al., using a modified set of concepts designed to provide a basis for systematic structure (e.g., teacher, school, teaching; firefighter, fire station, fire-fighting). The drawings produced by that pair during communication are then observed by a fresh pair of participants, who go on to communicate together, and so on. The system of communication is therefore under pressure to be both expressive (communicatively functional) and learnable (by the naive individuals during the observation phase). Theisen-White et al. find that the sets of drawings become more structured over these chains of transmission: the drawings develop component parts which refer to the domain (e.g. teaching, fire-fighting) and the category (e.g. person, building, activity).

These experimental results are therefore consistent with the modelling literature reviewed above and suggest a three-way contrast: pressure for compressibility alone results in degenerate languages (Kirby et al., 2008, Experiment 1); pressure for expressivity but not learnability (Garrod et al., 2007) leads to holistic systems; pressure for expressivity (from artificial filtering or, better, communication) leads to structure (Kirby et al., 2008; Theisen-White et al., 2011). However, no one model or experimental paradigm completely decouples learnability and expressivity: below, we present a model which does this, and which demonstrates this link between expressivity, learnablity and structure more conclusively.

## The model

We model individuals as rational learners who infer a distribution over possible languages (meaning-form mappings), and use those languages to communicate. Learners have a (parameterised) prior preference for simple, compressible languages, and during interaction a (parameterised) tendency to avoid utterances which are ambiguous in context.

## Model of languages

A language consists of a system for expressing meanings using forms. We consider the simplest possible meanings and forms which are nonetheless capable of evidencing systematic structure: meanings are vectors of length $v$, where each
element in the vector takes one of $w$ possible values. Similarly, forms are of length $l$, where each character is drawn from some alphabet $\Sigma$. We take $v=w=l=|\Sigma|=2$, which yields a set of meanings $\mathcal{M}=\{00,01,10,11\}$ and a set of forms $\mathcal{F}=\{a a, a b, b a, b b\}$. This gives a space of 256 possible languages, including degenerate, compositional and holistic mappings: see Table 1 for examples. ${ }^{2}$

## Hypotheses

Learners infer a distribution over languages: the space of hypotheses is therefore the space of possible distributions over all 256 languages. ${ }^{3}$ Following Burkett and Griffiths (2010), we use a Dirichlet process prior (Ferguson, 1973), characterised by concentration parameter $\alpha$ and base distribution $G_{0}$. The parameter $\alpha$ determines how many languages feature in this distribution: low alpha (we use $\alpha=0.1$ ) corresponds to an a priori belief that the majority of the probability mass will be on a single language. The base distribution is a distribution over languages, and would be the prior if learners only considered single-language hypotheses.

Our base distribution encodes a preference for simplicity, operationalised as a preference for languages whose description is compressible. Intuitively, degenerate languages permit more compressed descriptions than compositional languages; holistic languages are, by definition, incompressible. The prior used in Kirby, Dowman, and Griffiths (2007) captures this intuition: it assigns higher probability to languages in which fewer forms are used to convey a given set of meanings. We simply apply this metric both over the full set of meanings and specific feature values (see Appendix). This prior splits the space of 256 possible languages into 12 language classes, based on the number of forms in the language and the regularity with which feature of meaning are mapped to components of form: the priors for individual languages are depicted in Fig. 1, with example languages from some pertinent classes in Table 1. The prior yields the desired ranking of languages: more compressible languages (i.e. with fewer forms) are preferred, but within those languages with a given number of forms, there is a preference for languages which consistently map feature values in the meaning to a single character in the corresponding position in the form.

## Likelihood

We sample a form $f$ from the distribution $P(f \mid h, C, t)$, which specifies the probability of $f$ given hypothesis $h$, a context of utterance (a set of meanings) $C$, and topic $t \in C$, which the speaker attempts to discriminate from the other meanings

[^214]Table 1: Example languages from three important classes.

|  | Form |  |  |
| :---: | :---: | :---: | :---: |
| Meaning | degenerate | holistic | compositional |
| 00 | aa | aa | aa |
| 01 | aa | ba | ab |
| 10 | aa | ab | ba |
| 11 | aa | bb | bb |



Figure 1: Probability in $G_{0}$ for individual languages in each class, arranged by number of distinct forms, and (within a given number of forms) increasing compressibility. Annotations give the number of languages per class, all of which have equal prior probability.
in $C$. $P(f \mid h, C, t)=P(l \mid h) \cdot P(f \mid l, C, t)$ : we simply sample a language $l$ from the speaker's hypothesis, then given that language and the context, sample an utterance. We include a parameterisable preference to avoid ambiguity during this latter step, following the model of pragmatics provided by Frank and Goodman (2012). Assuming some small probability of error on production $\varepsilon$ :

$$
P(f \mid l, C, t) \propto \begin{cases}\left(\frac{1}{a}\right)^{\gamma}(1-\varepsilon) & \text { if } t \text { is mapped to } f \text { in } l \\ \frac{\varepsilon}{|S|-1} & \text { if } t \text { is not mapped to } f \text { in } l\end{cases}
$$

where we normalise over all possible forms from $\mathcal{F}$. $a$ is ambiguity, the number of meanings in $C$ that map to form $f$ in $l$, and $\gamma$ specifies the extent to which utterances which are ambiguous in context are penalised. If $a=1$ ( $f$ is unambiguous in this context) and/or $\gamma=0$ then this yields a model of production where the 'correct' form is produced with probability $1-\varepsilon$. However, when $\gamma>0$ and $f$ is ambiguous in context (i.e. $a>1$ ), then the 'correct' mapping from $t$ to $f$ is less likely to be produced, and the remaining probability mass is spread equally over the other possible forms. Therefore, $\gamma>0$ introduces a penalty for languages whose utterances are ambiguous in context. We use $\varepsilon=0.05,|C|=3$, and vary $\gamma$.

## Inference

Exact inference over this hypothesis space in intractable: instead, following Burkett and Griffiths (2010), we use a Gibbs
sampler based on the Chinese Restaurant Process to sample a hypothesis direct from the posterior. As described below, learners acquire an expanding set of observed utterances during their lifetime: we run the inference over the most recent $r=80$ observations, in order to improve simulation runtimes.

## Transmission in populations

Following the experimental methods employed by TheisenWhite et al. (2011) and Garrod et al. (2007), we compare two types of population: in chains, simulated agents are organised into pairs, are trained on data produced by the previous pair (see below), and then interact, producing data which the next generation in the chain (a new, naive pair of simulated individuals) are trained on. In dyads exactly the same regime of training and interaction is observed. However, naive individuals are not introduced at each generation: rather, the same individuals are trained on their own productions from the previous phase of interaction. ${ }^{4}$ The contrast between chains and dyads allows us to manipulate the pressure for learnability: in chains, where naive individuals are introduced at every generation, the pressure for learnability (i.e. the influence of the prior preference for simplicity) is likely to be relatively strong. In dyads, in contrast, there is only one episode of transmission to naive individuals (at generation 1), and consequently the pressure for simplicity arising from the prior is substantially diminished.
Training During training, the pair are presented with a shared set of 20 form-meaning pairs, produced by the preceding pair during interaction or (for the first generation only) a shared set of 20 form-meaning pairs generated from a randomly-selected fully-expressive holistic language (this initialisation with holistic languages is inspired by the experimental work discussed above). This data is added to each agent's memory (which will be empty for individuals in chains), and then a hypothesis is sampled from the posterior.
Interaction After training, the pair interact for 40 rounds. At each round of interaction, one individual acts as teacher and the other as learner. The teacher is prompted with a randomly-selected context and topic, and samples a form from their hypothesis. The learner adds the observed form-meaning-context triple to its memory, and samples an updated hypothesis. The roles of teacher and learner then switch, and a new round is played.
Transmission The 20 form-meaning pairs produced by one randomly-selected member of the pair at generation $n$ is used

[^215]as the training data for the pair at generation $n+1 .{ }^{5}$

## Results

The results (Fig. 2) match the predictions of our hypothesis, and are consistent with the experimental results described above. When there is pressure for learnability arising from transmission to naive individuals, but no pressure for expressivity (achieved by using the chain population model and setting $\gamma$ for interaction to 0 ), the final distribution is dominated by degenerate languages, as in Kirby et al. (2008), Experiment 1. Note that this preference for degenerate languages is even stronger than that seen in the prior: given the parameters of the model, in particular the low concentration parameter for the Dirichlet process prior, this exaggeration of the prior is as predicted by Burkett and Griffiths (2010). In contrast, in the condition where there is expressivity pressure but little pressure for learnability (dyads, $\gamma=3$ ), the initial holistic languages, (expressive but not compressible) persist. Members of the dyad constantly replenish their own evidence that the language is holistic: consequently, the initial holistic language is locked in. This matches the experimental results obtained for dyads (Garrod et al., 2007): due to the lack of transmission to new individuals, there is little pressure for compressibility to counteract lock-in and expressivity requirements during interaction, and structure does not emerge. Note also that this result holds despite the fact that we set a fairly low memory limit for individuals $(r=80)$. Finally, when there is pressure for both learnability and expressivity (chains, $\gamma=3$ ), we see structured languages emerge: the final distribution is dominated by a priori unlikely expressive languages, but among these it is the a priori most likely languages, the compositional languages, that dominate. Again, this matches our hypothesis and the experimental data from Theisen-White et al. (2011).

## Discussion

Our model shows that pressure for expressivity or simplicity alone does not lead to the emergence of structure: only when both pressures are at play does structure emerge. Furthermore, only cultural evolution is required for structure: we can explain why language is structured without recourse to invoking an evolved, domain-specific faculty of language.

As well as corresponding closely with existing modelling and experimental data, these findings make sense of the distribution of structure in the communication systems of nonhuman animals. Many small but expressive communication systems exist in nature, a classic example being alarm calling systems, which allow the discrimination of several referents (predators), but do so using vocalisations which are holistic and unlearned (Fitch, 2000). Learned vocal communication systems are witnessed in many species of bird, as well as being patchily distributed among mammals (Fitch, 2000): strik-

[^216]Learnability, no expressivity


Expressivity, no learnability


## Learnability and expressivity




Figure 2: Time courses (left: mean probability of each language class in the last sampled hypothesis of each individual in multiple chains/dyads) and final distributions (right: mean probability as in time courses, averaged over the final 50 generations of those same simulation runs). When learnability is the only pressure (top: average of 20 simulation runs), degenerate languages dominate the final distribution. When expressivity is the only pressure (middle: 50 runs), the original expressive but holistic languages are preserved. When both pressures are at play (bottom: 50 runs), expressive but compositionally-structured languages emerge.
ingly, song, the classic example of (combinatorial, not compositional) structure in animal communication, occurs in precisely these species, whose communication system is under cultural selection to be learnable but expressive. This is entirely consistent with the predictions of our model, although we would suggest that the expressivity pressures inherent in communication in these species must be rather different from the expressivity pressure in language, with a focus on signalling e.g. individual quality, rather than communicating propositions.

## Conclusions

The results from our model support the hypothesis drawn from our review of the modelling and experimental literature on the evolution of communication systems: structure emerges when a system of communication is under pressure to be both expressive (due to communicative interaction) and simple (due to domain-general preferences for compressibility imposed during language learning). Crucially, both these pressures must be in play: pressure for expressivity or simplicity alone does not lead to structure. Linguistic structure therefore can and should be explained as a consequence of cultural evolution: structure in language is a linguistic adaptation, not a biological adaptation, and it is an adaptation in response to competing pressures for expressivity and learnability inherent in language transmission and use.

## References

Brighton, H., Kirby, S., \& Smith, K. (2005). Cultural selection for learnability: Three principles underlying the view that language adapts to be learnable. In M. Tallerman (Ed.), Language origins: Perspectives on evolution (pp. 291-309). Oxford: Oxford University Press.
Brighton, H., Smith, K., \& Kirby, S. (2005). Language as an evolutionary system. Physics of Life Reviews, 2, 177-226.
Burkett, D., \& Griffiths, T. L. (2010). Iterated learning of multiple languages from multiple teachers. In A. D. M. Smith, M. Schouwstra, B. de Boer, \& K. Smith (Eds.), The Evolution of Language: Proceedings of the 8th International Conference (p. 58-65). Singapore: Word Scientific.
Chater, N., \& Vitanyi, P. (2003). Simplicity: a unifying principle in cognitive science. Trends in Cognitive Science, 7, 19-22.
Christiansen, M., \& Chater, N. (2008). Language as shaped by the brain. Behavioral and Brain Sciences, 31, 489-509.
Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. The Annals of Statistics, 1, 209-230.
Fitch, W. T. (2000). The evolution of speech: a comparative review. Trends in Cognitive Science, 4(3), 258-267.
Frank, M. C., \& Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. Science, 336, 998.
Garrod, S., Fay, N., Lee, J., Oberlander, J., \& MacLeod, T. (2007). Foundations of representation: where might graphical symbol systems come from? Cognitive Science, 31(6), 961-987.

Kirby, S. (2002). Learning, bottlenecks and the evolution of recursive syntax. In E. Briscoe (Ed.), Linguistic evolution through language acquisition: Formal and computational models (pp. 173-203). Cambridge: Cambridge University Press.
Kirby, S., Cornish, H., \& Smith, K. (2008). Cumulative cultural evolution in the laboratory: an experimental approach to the origins of structure in human language. PNAS, 105(31), 10681-10686.
Kirby, S., Dowman, M., \& Griffiths, T. L. (2007). Innateness and culture in the evolution of language. PNAS, 104, 52415245.

Pinker, S., \& Bloom, P. (1990). Natural language and natural selection. Behavioral and Brain Sciences, 13(4), 707-784.
Smith, K. (2004). The evolution of vocabulary. Journal of Theoretical Biology, 228(1), 127-142.
Steels, L. (1998). The origins of syntax in visually grounded robotic agents. Artificial Intelligence, 103, 133-156.
Steels, L. (2012). Self-organization and linguistic selection in language evolution. In L. Steels (Ed.), Experiments in cultural language evolution. Amsterdam: John Benjamins.
Theisen-White, C., Kirby, S., \& Oberlander, J. (2011). Integrating the horizontal and vertical cultural transmission of novel communication systems. In L. Carlson, C. Hoelscher, \& T. F. Shipley (Eds.), Proceedings of the 33rd Annual Conference of the Cognitive Science Society (pp. 956-961). Austin, TX: Cognitive Science Society.

## Appendix

For a given set of forms $F$ (where members of the set are either complete forms, e.g. $a a$, or partially-specified forms, e.g. $\quad a *$, indicating a string-initial $a$ ) and a set of meanings $M$, we can count the number of mappings in a language for which forms from $F$ are associated with meanings from $M$ : we denote this quantity $n(M, F)$. For instance, $n(\{00,10,11,10\}, a a)=4$ for the degenerate language in Table 1 , since the form $a a$ is associated with all 4 of these meanings, but 1 for the compositional and holistic languages; $n(0 *, a *)=2$ for the degenerate and compositional languages but 1 for the holistic language, since there is only a single mapping where meaning-initial 0 maps to form-initial a. Our base probability for language $l$ characterised is then:

$$
G_{0}(l) \propto P(l, \mathcal{M}, \mathcal{F}) . \prod_{m \in\{0 *, 1 *\}} P(l, m,\{a *, b *\}) \cdot \prod_{m \in\{* 0, * 1\}} P(l, m,\{* a, * b\})
$$

where we normalise over all possible languages and $P(l, M, F)$ is the prior from Kirby et al. (2007),

$$
P(l, M, F)=\frac{\Gamma(|F| \sigma)}{\Gamma(\sigma)^{|F|} \Gamma(m+|F| \sigma)} \prod_{f \in F} \Gamma(n(M, f)+\sigma)
$$

where $\Gamma(x)=(x-1)$ ! and $m$ is the number of meanings from $\mathscr{M}$ that unify with $M$. The parameters $\sigma$ determines the strength of the preference for simplicity: low $\sigma$ (we use $\sigma=1$, the lowest possible value) strengthens the preference for more compressible languages, higher $\sigma$ leads to a weaker preference for such languages.

# Modelling the Supervisory System and Frontal Dysfunction: An Architecturally Grounded Model of the Wisconsin Card Sorting Task 

Mariam R. Sood<br>(msood01@mail.bbk.ac.uk)

Richard P. Cooper<br>(R.Cooper@bbk.ac.uk)

Department of Psychological Sciences, Birkbeck, University of London
Malet Street, London WC1E 7HX, UK


#### Abstract

We present a model of the Wisconsin Card Sorting Test, a classical neuropsychological test frequently used to assess deficits in executive functioning. The model is grounded in a cognitive architecture based on the Supervisory System theory of Norman and Shallice (1986) and evaluated against data from control subjects and several groups of neurological patients as reported by Stuss et al. (2000). The model is able to account for control performance across a range of dependent measures. When damaged in theoretically motivated ways it is also able to capture the behaviour of the different patient groups. Specifically, the model supports the association by Shallice et al. (2008) of the function of tasksetting to left lateral prefrontal cortex, of the function of attentiveness to inferior medial prefrontal cortex, and of the function of monitoring to right lateral prefrontal cortex. The implication of these results for the supervisory system architecture and the localisation of function within prefrontal cortex are discussed.


Keywords: Cognitive architecture; Supervisory system; Wisconsin card sorting task; Frontal dysfunction.

## Introduction

Several theories of the organisation of cognitive processes have been proposed over the last 25 years. These cognitive architectures generally comprise complex production systems, and normally have their roots in behaviours in specific cognitive domains (e.g., problem solving, as in, Soar: Newell, 1990; associative memory, as in ACT-R: Anderson, 2007; or immediate response tasks as in EPIC: Meyer \& Kieras, 1997). While such architectures have been highly successful at accounting for a range of behavioural effects, they are not well suited to modelling the behaviour of neurological patients with focal brain damage. This is largely because it is unclear how the functional components of such architectures might be impaired without causing complete breakdown of the system. The cognitive architecture sketched by Norman and Shallice (1986) and elaborated by Shallice et al. (2008), in contrast, provides a modular view of cognition in which functional components may operate more or less efficiently, and hence neurological deficits might be more directly accounted for.

The Norman/Shallice theory draws a primary distinction between routine behaviour, which is generated by a lower level scheduling system - Contention Scheduling - and nonroutine behaviour, which is effected by a higher level system - Supervisory System. This higher level system operates indirectly on behaviour by modulating the functioning of Contention Scheduling. When initially
described (Norman \& Shallice, 1986), the situations requiring Supervisory System input were clearly enumerated but the subsystem's functioning was specified only in abstract terms. Those functions include what have since come to be known as executive functions such as tasksetting, monitoring and working memory maintenance.

In a somewhat separate line of work, Shallice, Stuss and colleagues (e.g. Stuss et al., 2000; Shallice et al., 2008) have attempted to account for the deficits of several groups of patients with focal frontal lobe lesions in terms of deficits affecting specific executive functions which, they argue, are effected by different regions of the prefrontal cortex. Thus, the deficits of patients with left lateral prefrontal lesions across a range of tasks are interpreted as reflecting impaired task-setting, while the deficits of right lateral prefrontal patients are interpreted as reflecting impaired monitoring. Similarly, the deficits of patients with focal lesions affecting inferior medial prefrontal regions are interpreted as reflecting an impaired ability to sustain attention to a task, while the deficits of patients with focal lesions affecting superior medial prefrontal cortex are interpreted as reflecting an impairment in "energisation", i.e., mobilisation of cognitive resources, corresponding phenomenologically to cognitive effort.

Shallice et al. (2008) relate the four executive functions discussed in the previous paragraph to the Supervisory System, with a specific focus to how the two accounts relate within a simple task-switching study. However these authors provide only an informal characterisation of the functions. They do not provide a precise computational instantiation of the ideas. The goal of this paper is to provide and evaluate such an instantiation. More specifically we present a computational account of the heterarchical organisation of the Supervisory System. The account is grounded in a model of a specific task - the Wisconsin Card Sorting Test (WCST). This widely used test of executive function provides multiple dependent measures that are sensitive to frontal lobe damage (Milner, 1963). We report simulations of the behaviour of control subjects and of four patient groups, comparing our results with those of Stuss et al. (2000), who tested patients and controls on the task.

The following sections briefly discuss the cognitive architecture in which the model is framed, the Wisconsin card sorting test and the neuropsychological group study that provides the target data. Following this, we present the model itself, the methodology for modelling control and


Figure 1: The proposed functional organisation of the Supervisory System architecture. Hexagonal boxes represent processes while rounded rectangles represent buffers or storage systems. Arrows show hypothesised connectivity between components.
patient performance, and the respective simulation results. We conclude by considering the implications of this work for the computational specification of the Supervisory System and more generally for the functional organisation of higher cognition.

## The Supervisory System Architecture

The Supervisory System proposed by Shallice, Stuss and colleagues (e.g., Shallice et al., 2008) is a heterarchical system comprising, amongst other things, four core subprocesses: task-setting, active monitoring, energisation and attentiveness. The evidence for this organisation is drawn from neuropsychological case studies where the nature of deficits exhibited by frontal patients show subtle differences based on the lesion location. For example, the impairment exhibited by left prefrontal patients may be understood as resulting from inefficient task strategy formation while right prefrontal patients make errors that suggest poor ability to monitor internal and external events. The deficits of inferior medial prefrontal patients may stem from a characteristic lack of attention while superior medial prefrontal patients exhibit a longer $(30 \%)$ start up delay in task execution compared to other groups (for a review, see Shallice \& Cooper, 2011).

The cognitive architecture of the model described in this paper is derived from the Contention Scheduling / Supervisory System theory and is depicted in figure 1. Processing within the Contention Scheduling components of architecture is as follows: perceptual input enters Sensory Stores. Potential responses are generated from this by Apply Set subject to application of the current stimulus-response mapping set. These responses are passed to a Response Buffer before being generated as actions. The Generate Response process also maintains Forward Model, which represents the anticipated sensory feedback of the system's actions. The Supervisory System modulates the behaviour
of Contention Scheduling by two key processes: a) Monitoring, which compares sensory feedback with anticipated sensory feedback and rejects the current stimulus-response mapping if there is a mismatch (i.e., an unanticipated sensory input) by clearing Current Set, and b) Task Setting, which sets a stimulus-response mapping when Current Set is empty. Two other supervisory processes, Attentiveness and Active Maintenance, work to counteract decay which is assumed to operate on elements within Current Set and Working Memory. With the exception of Energisation, the model adequately represents all other subprocesses of the Supervisory System theory.

## The Wisconsin Card Sorting Test

## The Task

In order to evaluate the Supervisory System architecture we consider its application to a specific task: the Wisconsin Card Sorting Test (WCST). The WCST exists in various forms. The version simulated here is the 64A version used by Stuss et al. (2000). In this version of the task, subjects are required to sort a deck of cards, 64 in total presented one at a time, into four groups. Each card has a picture of a specific shape in variable numbers and colours (e.g., one red triangle or four blue squares; see figure 2). Four "target" cards, differing with respect to the number, colour and shape of items they depict, are provided and subjects are required to place each successive card from the main deck under one of the four target cards. In the 64A version, subjects are informed of the three possible sorting criteria - sort by colour, sort by number or sort by shape - prior to the test. After each card is sorted, the subject is given feedback. Based on the feedback, the subject should attempt to infer the correct sorting rule and use it for subsequent sorts. Once the subject correctly sorts 10 cards consecutively, the experimenter changes the rule without warning. The ideal


Figure 2: The Wisconsin Card Sorting Test, after two cards have been sorted according to the colour of their symbols and as a third card (two blue triangles) is presented for sorting.
subject will detect this and select a new rule, based on the feedback after each sorting attempt.

Neurologically healthy subjects have little difficulty in this task. However patients with frontal lesions are prone to perform poorly, frequently showing incapacity to change the rule when the feedback is negative, i.e. they tend to 'persevere', but also showing 'set loss' errors, where they appear to correctly infer a rule, but fail to follow that rule for ten consecutive sorting trials, even with positive feedback.

## Neuropsychological Evidence

The motivation behind choosing the WCST for evaluation of the supervisory system architecture over other executive tasks is the availability of detailed empirical data published by Stuss et al. (2000) on patients categorised with focal lesions on the four brain regions of theoretical interest. The empirical study carried out by Stuss et al. (2000) tested seven groups of subjects. Four groups had focal frontal lesions on left/right dorsolateral prefrontal cortex (LDL/RDL), superior medial (SM) and inferior medial (IM) prefrontal regions. The fifth and the sixth patient groups had lesions affecting left/right non-frontal brain regions and the seventh group comprised neurologically healthy subjects. The subjects were tested on three versions of WCST: 128, 64 A and 64B. In the 128 version, subjects were not provided any instructions on how to perform the test. In 64A version, subjects were informed of the three possible sorting criteria beforehand, while in 64B version subjects were also alerted when the rule was about to change. In each case, the errors made by subjects were classified into four categories: perseveration of preceding category (PPC: an incorrect response that matches the preceding sorting criterion), perseveration of preceding response (PPR: an incorrect response that matches exactly the features on the preceding trial), set-loss (an incorrect response following
three or more consecutively correct responses) and other errors. Patients with frontal lesions, compared to those with non-frontal lesions and controls, exhibited more PPC, PPR and set-loss errors. The error patterns exhibited by the different frontal groups revealed subtle differences. For example, in the 64A condition, all frontal groups except IM showed significantly more PPC and PPR errors than controls. In contrast, the IM group made significantly more set-loss errors than patients from other frontal groups.

## Modelling the WCST

## Model Assumptions and Description

The model discussed here is an elaboration of the heterarchical Supervisory System theory (figure 1), with its components configured for the WCST 64A condition of the empirical study by Stuss et al. (2000). Consider first the three buffers and two processes that make up Contention Scheduling. When a card is to be sorted, a propositional representation of the card appears in Sensory Stores. Apply Set then consults Current Set for a representation of the current sorting rule (e.g., sort by colour) and uses this in conjunction with the representation in Sensory Stores (e.g., two blue triangles) to generate a putative response (e.g., place the card on the right-most pile) which is stored in Response Buffer. Generate Response then produces the actual response (storing a copy in Working Memory), together with a representation of the anticipated consequences of the response - the Forward Model. (In the current implementation Forward Model is ignored, since the anticipated consequence of any action is positive feedback.)

Processing within the Contention Scheduling components is modulated by the Supervisory System components. First, Monitoring may detect negative feedback in the sensory store (or more generally, a mismatch between the contents of Forward Model and Sensory Store). In such situations, Monitoring will clear Current Set (on the assumption that the current sorting rule is inappropriate). Second, Task Setting may generate a putative sorting rule and place a representation of that rule in Current Set. This occurs when Current Set is empty (e.g., because the representation of the previous sorting rule in Current Set has either decayed or been explicitly deleted by Monitoring). Generation of a putative sorting rule depends on the contents of Sensory Buffer and recent responses stored in Working Memory.

Elements in the two supervisory buffers (Current Set and Working Memory) have activation values that decay over time. If the activation values fall below a threshold, the buffer contents cannot be accessed. The supervisory processes of Attentiveness and Active Maintenance work in direct opposition to decay, exciting buffer elements so as to prevent their loss

The model's behaviour may be summarised as follows: At the beginning of the task, the first sorting schema is generated at random from among the three possible schemas: sort-by-colour, sort-by-number and sort-by-form. When a card is presented, the Contention Scheduler sorts

Table 1: Parameters of the model.

| Parameter | Range | Description |
| :--- | ---: | :--- |
| Monitoring $_{\text {exogenous }}$ | $0-1$ | If impaired, feedback is not acted upon |
| Monitoring $_{\text {endogenous }}$ | $0-1$ | If impaired, drop in attention is not monitored |
| Taskset $_{\text {none }}$ | $0-1$ | If impaired, unable to switch strategy | Taskset $_{\text {random }}$ Attentiveness ${ }_{\text {persistence }}$ If impaired, unable to produce efficient strategy, a random strategy is chosen

the card according to the sorting criterion stored in Current Set. Feedback is monitored by Monitoring, a supervisory process that clears Current Set in the event of negative feedback. When Current Set is empty, Task Setting is invoked. This process accesses Working Memory to gather details of previous unsuccessful sorting attempts (if any can be recalled) and generates a new potential sorting rule that has not been recently used. If there is more than one possible choice of rule consistent with current evidence, Task Setting chooses at random from the available choices.

The model is implemented in the C programming language. In order to ensure their independent nature, the supervisory sub-processes and Contention Scheduling are implemented as separate 'threads'. Results are scored according to the criteria followed by Stuss et al. (2000).

## Behaviour of the Model

The model's behaviour is dependent on a number of parameters, which essentially determine the efficiency of processing of the various sub-processes. Table 1 provides a brief description of what these parameters represent and the range of values they can take. There are essentially two types of parameters: activation-related parameters (thresholds, activation persistence and activation boost parameters) and efficiency-related parameters. The latter specify the probability of a subsystem functioning. For instance, a value of 0.10 for Monitoring exogenous specifies that monitoring is active roughly $10 \%$ of the time. The remaining $90 \%$ of the time, the process does not function.

All parameters have optimal or ideal values. Thus, when all efficiency parameters are set to 1.00 , activation boost rates are set to the reciprocals of corresponding persistence rates (so that maintenance exactly counteracts decay), and thresholds are set to 0.5 , the model sorts optimally, correctly sorting approximately 56 out of 64 cards, achieving 5 categories (i.e., correctly sorting 10 cards according to 5 different rules) and making errors only when it is attempting to discover a rule following negative feedback.

## Modelling Control Performance

When neurologically healthy subjects attempt the WCST they generally do not perform at the optimal level. Thus the control subjects of Stuss et al. (2000) achieved on average 3.9 categories, made occasional perseverative errors, where they continued sorting by a rule even given negative feedback, and also occasionally produced set-loss errors, where they appeared to correctly infer the sorting rule, only to forget it even though the feedback was positive (see figure 3, right-most bars). In order to model control performance, normally distributed random noise was added to activation values of Working Memory and Current Set elements, the persistence of activation values was decreased, and the efficiency of supervisory processes was decreased. Systematic exploration yielded performance similar to controls when these parameters were set as follows: noise standard deviation $=0.05$; Monitoring exogenous $=$ Monitoring $_{\text {endogenous }}=$ Taskset $_{\text {none }}=$ Taskset $_{\text {random }}=0.90$; Memory $_{\text {persistence }}=0.70$; and Attentiveness ${ }_{\text {persistence }}=0.76$. With these values, the model generates perseveration and set-loss errors at rates comparable to those of the control subjects of Stuss et al. (2000) - see figure 3. The values indicate that control performance can be modelled by introducing slight imperfections to the Supervisory System.

## Modelling Frontal Dysfunction

Based on the arguments of Shallice et al. (2008), we associate exogenous monitoring (Monitoring ${ }_{\text {exogenous }}$ ), task setting ( Taskset $_{\text {none }}$ ) and attentiveness (Attentiveness ${ }_{\text {persistence }}$ ) with right dorsolateral, left dorsolateral and inferior medial prefrontal patients respectively. Although endogenous monitoring (Monitoring endogenous ) and task random setting (Taskset ${ }_{\text {random }}$ ) are important elements of monitoring and task setting processes, analysis of the model's behaviour revealed that they do not contribute significantly to the dependent measures and hence they have been excluded from the analysis of frontal dysfunction. Moreover we do


Figure 3: Model performance versus that of the patients of Stuss et al. (2000). Error bars represent one standard error from the mean.
not attempt to account for the behaviour of the superior medial prefrontal patients as the model does not have an explicit representation of the energisation process.

We adopt the methodological approach of modelling patient performance by reducing the efficiency of the process held to be impaired in the corresponding patient group. Specifically, we adjust the relevant parameter so that the model accurately captures the mean number of categories achieved by each set of patients in the Stuss et al. study ( 0.6 categories for RDL patients, 1.3 categories for LDL patients and 2.6 categories for IM patients), and then compare the model's behaviour on the three dependent measures described above (PPC, PPR and set-loss errors).

Thus an impairment level of 0.00 in Monitoring exogenous,, 0.10 in Taskset ${ }_{\text {none }}$, and 0.74 in Attentiveness ${ }_{\text {persistence }}$ produced a mean category measure comparable to RDL, LDL and IM patients respectively. When setting these parameters to model the impairments of the three patient groups, all other parameters were fixed at the levels used to simulate control subjects. Simulation data on three dependent measures PPC, PPR and set-loss errors - for each patient category obtained in this way and averaged over 10 runs of the model is shown in figure 3, plotted against the corresponding patient data published by Stuss et al.

## General Discussion

As shown in figure 3, the model of WCST behaviour, embedded within the broader Supervisory System / Contention Scheduling architecture, is able to provide a good account of control subject behaviour across four dependent measures: categories obtained, PPC errors, PPR errors and set-loss errors. This provides support - albeit weak support - both for the Supervisory System / Contention Scheduling architecture and for the model of WCST within it. However, equally important for the current work is the behaviour of the model when damaged and its relation to that of neurological patients. When damaged in theoretically motivated ways, the model reproduces several key features of the behaviour of neurological patients. Most critically, an impairment of exogenous monitoring leads to elevated levels of PPC and PPR errors, as seen in right dorsolateral prefrontal patients. An impairment of task setting leads to a similar error profile, as seen in left dorsolateral prefrontal patients. Finally, an impairment of attentiveness leads to elevated set loss errors, as seen in patients with inferior medial prefrontal lesions. This provides further support for both the model and the association of these supervisory functions with the different regions of prefrontal cortex.

The results must be interpreted with caution, however. First, the model performs similarly with impairments to either exogenous monitoring or task setting. While this is consistent with patient behaviour, it supports an argument originally made by Stuss et al. (2000) that the erroneous behaviour of their right dorsolateral and left dorsolateral groups, whilst qualitatively and quantitatively similar, may in fact be due to different functional impairments. The model demonstrates that the WCST is unable to discriminate between these functional impairments (at least with respect to PPC and PPR errors), and that empirical studies of the two patient groups on other, more discriminating, tasks is necessary if one is to make the argument that the functions of (exogenous) monitoring and task setting are indeed supported by different regions of prefrontal cortex.

The reverse side of this argument, however, derives from the fact that inferior medial patients produce elevated numbers of set-loss errors but not of PPC or PPR errors. This pattern of behaviour is produced by an impairment to the effectiveness of the attentiveness sub-process. Thus the
model supports the treatment of 'attentiveness' as a functionally and structurally distinct sub-process, as well as the 'impaired attentiveness' account of inferior medial prefrontal patient performance.

A second caution regarding the results concerns the rate of set-loss errors in simulation of RDL patient performance, which is lower than that seen in patient behaviour. This is in part because the model must sort a minimum number of consecutively presented cards correctly (and hence demonstrate that it is following a rule) before an error can be counted as a set-loss error. With severe impairments in the model, this is rare. Hence the opportunity for set loss errors is rare. We have simulated the RDL patient group by setting Monitoring ${ }_{\text {exogenous }}$ to 0.00 in order to match performance on the number of categories correctly sorted. Perhaps this level of impairment is too severe. This is an issue to be addressed in future work.

The issue of severity relates to the methodology employed in simulating patient behaviour. Patient performance was modelled by choosing one parameter value to match the number of categories achieved by the model to that of the relevant patient group. This does not take account of the heterogeneity of each patient group - not all patients were equally severely impaired - and a more appropriate methodology would be one that attempted to match the varying severity of individual patients, rather than of each group as a whole. This is an issue for further research, though in the absence of individual subject data, a plausible strategy may be to sample different levels of severity, as used by Cooper et al. (2005) in modelling the action errors of neurological patients.

Two more general questions concern the nature of supervisory processes and the Contention Scheduling / Supervisory System architecture within which the WCST model is embedded. Considering first the architectural issue, the model demonstrates that the functional decomposition of Contention Scheduling and the Supervisory System is able to support behaviour on a complex task, and so the Contention Scheduling / Supervisory System architecture provides a viable alternative to production system architectures such as ACT-R, Soar and EPIC. At the same time, the architecture remains relatively underspecified and substantial elaboration of the architecture and its subcomponents (e.g., through application to other tasks) is required before it can be fully compared with these alternatives.

With regard to the nature of supervisory processes, many theorists appear to assume, at least implicitly, a distinction between supervisory and non-supervisory processes. There is, however, some debate about whether the supervisory system is most appropriately viewed as a unitary system (e.g., Duncan, 2010) or as functionally heterogeneous (e.g., Stuss, \& Benson, 1986; Shallice, \& Cooper, 2011), and whether prefrontal cortex is better viewed as functionally hierarchical (e.g., Badre, 2008) or functionally heterarchical (Shallice \& Burgess, 1996; Shallice et al., 2008). The model described in this paper substantiates the theoretical stand of

Shallice and colleagues, but again further work is required. A possible extension to the present model is to generalise it to other executive tasks such as Tower of Hanoi, Tower of London, Stroop test etc. Applying the architecture to other executive tasks will allow for better validation of the theoretical hypotheses, which are not adequately and independently assessed by WCST.

## References

Anderson, J. R. (2007). How Can the Human mind Occur in the Physical Universe? Oxford University Press, Oxford, UK.
Badre, D. (2008). Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes. Trends in Cognitive Sciences, 12(5), 193-200.
Cooper, R.P., Schwartz, M.F., Shallice, T. \& Yule, P. (2005). The simulation of action disorganisation in complex activities of daily living. Cognitive Neuropsychology, 22, 959-1004.
Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. Trends in Cognitive Science, 14, 172-179.
Kieras, D. E, \& Meyer, D. E. (1997). An overview of the EPIC architecture for cognition and performance with application to human-computer interaction. HumanComputer Interaction, 12, 391-438.
Milner, B. (1963). Effects of different brain lesions on card sorting. Archives of Neurology, 9, 90-100.
Newell, A. (1990). Unified Theories of Cognition. Harvard University Press, Cambridge, MA.
Norman, D. A. \& Shallice, T. (1986). Attention to action: willed and automatic control of behaviour. In R. Davidson, G. Schwartz \& D. Shapiro (eds.) Consciousness and Self Regulation, Volume 4. NY: Plenum.
Shallice, T., \& Burgess, P.W. (1996). The domain of supervisory processes and temporal organisation of behaviour. Philosophical Transactions of the Royal Society of London, B351, 1405-1412.
Shallice, T., \& Cooper, R.P. (2011). The Organisation of Mind. Oxford: Oxford University Press.
Shallice, T., Stuss, D. T., Picton, T. W., Alexander, M. P., \& Gillingham, S. (2008). Mapping task switching in frontal cortex through neuropsychological group studies. Frontiers in Neuroscience, 2, 79-85.
Sood, M. R. (2012). Executive Function Deficits of Neurological Patients with Frontal lobe injury: Analysis using Computational Modelling. Unpublished master's thesis. Birkbeck, University of London, United Kingdom.
Stuss, D. T., \& Benson, D. F. (1986). The Frontal Lobes. New York: Raven Press
Stuss, D. T., Levine, B., Alexander, M. P., Hong, J., Palumbo, C., Hamer, L., Murphy, K.J. \& Izukawa, D. (2000). Wisconsin Card Sorting Test performance in patients with focal frontal and posterior brain damage: Effects of lesion location and test structure on separable cognitive processes. Neuropsychologia, 38, 388-402.

# Foreign accent does not influence cognitive judgments 

André L. Souza (andre.souza@crdh.concordia.ca)<br>Concordia University, Department of Psychology<br>7141 Sherbrooke St. W.<br>Montreal, QC H4V 1N3 Canada

Arthur B. Markman (markman@psy.utexas.edu)<br>The University of Texas at Austin, Department of Psychology<br>1 University Station, A8000<br>Austin, TX 78712 USA


#### Abstract

A recent paper by Lev-Ari and Keysar (2010) reported that the processing fluency associated with non-native speech causes non-native speakers to sound less credible. The authors found that the same trivia statements were rated as less truthful when spoken by a non-native speaker of English. The present paper reports the results of three studies that attempted to replicate the findings of Lev-Ari and Keysar (2010) by focusing on processing fluency manipulations other than accent. Although we used virtually the same methodology as Lev-Ari and Keysar (2010), we failed to replicate the key finding that foreign-accented speech is less credible than native-accented speech. The implications of this finding is discussed.


Keywords: fluency, foreign accent, credibility.

## Introduction

The U.S. Census Bureau (2010) reported that 38.5 million people (around $12.5 \%$ of the nation's population) have as mother tongue a language other than English. The increasing number of non-native speakers of English in the U.S. suggests that a significant amount of daily interactions involve a non-native speaker communicating in English with some sort of foreign accent.

The social psychological literature on language attitudes has documented considerable amount of evidence showing that, compared to their nonstandard, accented counterparts, listeners evaluate standard, non-accented speakers more favorably across different traits, such as competence, status, intelligence, confidence, guilt and success (Ryan \& Giles, 1982).

It is not entirely clear which cognitive mechanisms underlie this phenomenon. There is research suggesting that accent serves as a signal for the speakers' social group and that any negative attitude towards non-native speakers is caused by in-group biases and not by the accent itself. Alternatively, there is research showing an individual's actions and attitudes towards others are heavily dependent on how that person processes the information provided by them. The subjective ease with which individuals process incoming information influences them in a variety of cognitive tasks and domains (Gilbert, 1991; Schwarz, 2004; Alter \& Oppenheimer, 2009) such as estimates of familiarity (Jacoby \& Whitehouse, 1989), clarity (Whittlesea, Jacoby, \& Girard, 1990), riskiness (Song \& Schwarz, 2009), location and abstractness (Alter \& Oppenheimer, 2008), truthfulness (Reber \& Schwarz, 1999;

Unkelbach, 2007), liking (Winkielman \& Cacioppo, 2001) and even confidence (Koriat, 1993). Thus, one plausible hypothesis is that the negative impressions and judgments towards non-native speakers are triggered by the difficulty associated with processing accented speech.

A recent paper by Lev-Ari and Keysar (2010) directly explored this possibility. They asked native speakers of English to listen to a series of trivia statements such as Ants don't sleep and then indicate the degree of veracity of each statement. Participants listened to statements spoken by both native and non-native speakers of American English. The accented speech varied in terms of the degree: either mildly or heavily accented. They found that the statements spoken by non-native speakers were reliably rated as less truthful compared to the same statements spoken by native speakers.

The authors argued that their findings could not be explained in terms of stereotypes of prejudice signaled by the accent because participants were told that the non-native speakers were only reciting statements provided by a native speaker, and therefore were not displaying their own knowledge. Based on these findings, Lev-Ari and Keysar (2010) claimed that people misattribute the processing difficulty associated with non-native accented speech with the level of credibility they attribute to the content of the speech.

We began this project with the aim of exploring this issue further. The core idea is that if processing fluency influences people's judgments of the veracity of statements, then other manipulations of the speech signal such as adding background noise would also influence judgments of truth. We hoped to explore this issue both for statements of the kind used by Lev-Ari and Keysar (2010) as well as other kinds of judgments like consumer preference judgments. To presage our results, though, we were unable to replicate the initial findings.

This paper reports results for 3 studies. Study 1 explored the claim that inducing processing difficulty with mechanisms other than foreign accent (i.e., white background noise - Study 1a - and speech babble noise - Study 1b) affects judgments of truth. Studies 2 and 3 are attempts to replicate the findings of Lev-Ari and Keysar (2010). In Study 2, we asked participants to judge the truthfulness of trivia statements spoken by native and non-native speakers of English. In Study 3 we explore whether accent influences participant's
perception of the price of a product.

## Study 1

Study 1 investigated the claim that inducing processing difficulty with mechanisms other than foreign accent affects judgments of truth.

## Study 1a

Participants Twenty-six native speakers of English participated in Study 1a. Participants were undergraduate students at The University of Texas at Austin and participated for course credit.

Materials A female native English speaker recorded 70 trivia statements such as $A$ rat can last longer without water than a camel in a sound-attenuated booth. To obtain equivalent overall amplitude level for all statements, the sound files were equated for RMS amplitude. Each sound file was mixed with white noise at a four different Sound-to-Noise Ratios (SNR): level 0 corresponded to +17 dB SNR ( 68 dB of speech and 51 dB of noise), level 1 corresponded to +12 dB SNR, level 2 to +6 dB SNR and level 3 to 0 dB SNR. In the SNR notation used in this paper, the smaller the dB SNR, the louder the background noise. The mixed files were presented to participants using E-prime 2.0.

Procedure Study 1a used a within-subject design. Each participant heard all 70 trivia statements (48 experimental ones and 22 fillers) randomly mixed with one of three levels of noise (level 0, level 1 and level 2). Participants sat in front of a computer screen with headphones and were asked to indicate the truthfulness of each statement, using a scale between 0 (definitely false) and 10 (definitely true). Participants were also asked to rate whether they knew for a fact that the statement was true.

Manipulation Check To ensure that the noise manipulation made the trivia statements more difficult to process, a different group of 24 participants were asked to listen and rate the degree of difficulty to understand the statements. Each participant heard 25 randomly selected statements (five for each level of noise: no noise, level 0, level 1, level 2 and level 3).

Manipulation Check Results A one-way repeated measures ANOVA, with the difficulty level as dependent variable and noise level as independent variable, revealed a statistically reliable main effect of noise, $F(4,80)=60.59$, $p<.0001, \eta^{2}=.75$, suggesting that the overall distribution of the mean perceived difficulty across the five different noise levels significantly differed from each other. Post-hoc Bonferroni-corrected $t$-tests showed that, except for the level 0 vs. level 1 comparison, all other pairwise comparisons reliably differed from each other, $p^{\prime} s<0.05$.

Truthfulness Ratings Results and Discussion Because we wanted to avoid participants suspicion about the noise ma-
nipulation, we decided to present participants only with statements mixed with some level of noise, excluding therefore the sentences with no noise. Also, because the difficulty ratings for level 3 noise was extremely high ( $M=8.30, S D=1.89$ ), we decided to exclude this level, to avoid the possibility that participants would simply be unable to hear the statements completely.

To verify whether white noise affected the truthfulness ratings of the trivia statements, we ran a one-way repeatedmeasures ANOVA, with the truthfulness ratings as the dependent variable and the noise levels as the independent variable. Contrary to what we expected, the mean truthfulness ratings were very similar across all three different levels of white noise. The ANOVA showed that the means did not differ reliably from each other, $F(2,50)=.81, p=.45$.

The pattern of results suggests that the presence of white noise in speech does not affect judgments about the content of the speech. These findings go against the robust literature that shows that processing fluency affects cognitive judgments. On the other hand, because the overall truthfulness ratings across all levels of noise was $M=4.80(S D=2.85)$, one might claim that participants were just not engaging properly in the task, given that, in general, people are not used to hearing speech against this particular type of noise. In fact, Kozou et al. (2005) shows that speech competitors have a different effects on speech recognition and performance compared to non-speech competitors, such as white noise. Study 1 b addresses this point by presenting the statements against a speech competitor (i.e., babble speech) which is more common in people's environments and is found to affect speech differently than white noise (Kozou et al., 2005).

## Study 1b

Participants Twelve native speakers of English participated in Study 1b. Again, participants were undergraduate students of Psychology enrolled in a The University of Texas at Austin and participated for course credit. None of the participants from Study 1a participated in Study 1b.

Materials The materials were the same as in Study 1b, however, each sound file was mixed with speech babble noise at the same four different SNR's $(+17 \mathrm{~dB},+12 \mathrm{~dB},+6 \mathrm{~dB}$ and 0 dB ). Similarly to Study 1a, the mixed files were presented to participants using E-prime 2.0. The procedure was identical to Study 1a.

Manipulation Check As we did for Study 1a, a different group of 21 participants listened and rated the degree of difficulty to understand the statements. The procedure for the manipulation check was identical to Study 1a.

Manipulation Check Results A one-way repeated measures ANOVA showed a statistically significant main effect of noise, $F(4,80)=41.14, p<.0001, \eta^{2}=.67$. Slightly
different from what was found for Study 1a, post-hoc Bonferroni-corrected $t$-tests revealed that, level 3 significantly differed from all other levels ( $p^{\prime} s<0.009$ ). However, level 0 , level 1 and level 2 did not differ significantly from each other.

Truthfulness Ratings Results and Discussion Similarly to the findings from Study 1a, the mean truthfulness ratings did not differ significantly across all three different babble noise levels, $F(2,22)=.14, p=.86$. Although, the results for Study 1a and Study 1b suggest that noise (both white and speech babble) does not influence judgments of truth, one might claim that the failure to show differences in truthfulness ratings in Study 1b is easily explained by the fact that the various levels of noise were not perceived as different in terms of difficulty. To address this point, we re-ran Study 1b, but this time with different levels of SNR's. This time, level 0 corresponded to +8 dB SNR , level 1 corresponded to +2 dB SNR, level 2 to 0dB SNR and level 3 to -2dB SNR (negative SNR means that noise signal is louder than the speech signal). Using the same manipulation check procedure as before, 17 participants were asked to rate the degree of difficulty associated with listening the statements. A one-way repeated measures ANOVA showed a significant main effect of noise, $F(4,64)=45.21, p<.0001, \eta^{2}=.74$. Post-hoc Bonferronicorrected $t$-tests revealed that, except for the pairs level 1 vs . level 2 and level 2 vs. level 3, all other levels reliably differed from each other, $p^{\prime} s<0.05$.

For the truthfulness ratings of this novel noise level manipulation, a group of 13 native speakers of English were asked to rate the degree of truthfulness of the statements (procedure identical as before). Once again, the speech babble noise did not influence the judgments of truth, $F(2,24)=0.43, p=n s$. More importantly, the pairwise combinations that did differ in terms of difficulty level (i.e., level 0 vs. level 1, level 0 vs. level 3 and level 1 vs. level 3) did not show any reliable difference in terms of truthfulness ratings.

The results of Study 1a and 1b combined suggest that neither white noise nor speech babble noise seem to influence judgments of truth. More broadly, processing fluency associated with these auditory stimuli does not affect judgments about the content of the sentences. These findings go directly against Lev-Ari and Keysar (2010)'s claim that processing fluency associated with understanding foreignaccented speech directly influences judgments of truth. Study 2 and Study 3 are direct attempts to replicate Lev-Ari and Keysar (2010)'s findings with foreign-accented speech.

## Study 2

Participants Sixty-five native speakers of English participated in Study 2. Participants were undergraduate students at The University of Texas at Austin and participated for course credit. None of the participants from the previous studies participated in this one.

Materials A female native English speaker, two female native speakers of Brazilian-Portuguese and two female native speakers of Korean recorded the same 70 trivia statements used in the previous studies. As before, all sound files were equated for RMS amplitude. To ensure that the speech was perceived as accented, a separate pool of 28 participants rated the degree of foreign-accentedness of the statements (both the native and non-native speech). A repeated-measures ANOVA showed a main effect of language, that is, both the BrazilianPortuguese and the Korean speech were perceived as significantly more accented than the native speech, $F(2,54)=307.6$, $p<0.001, \eta^{2}=.91$. Brazilian-Portuguese and Korean did not differ from each other, although the Brazilian speakers were perceived as slightly more accented.

Procedures To test for the effect of accent on credibility judgments, participants sat in front of a computer and listened to 48 trivia statements in English. Sixteen of these statements were spoken by a native speaker of Brazilian-Portuguese, 16 by a native speaker of Korean and 16 by the native English speaker. All statements were recited in English. After listening to each statement, participants were asked to indicate how truthful they thought the statements were. For this, they used a Likert scale ranging from 1 (definitely false) to 10 (definitely true). Each participant heard additional 20 fillers statements read by two additional native speakers of English.

Results and Discussion To investigate the effect of foreign accent on the judgments of truth, we ran a repeated-measures one-way ANOVA, with language (accented vs. native) as independent variable and the truthfulness rating as the dependent variable. Our results failed to replicate the findings reported by Lev-Ari and Keysar (2010). There was no reliable main effect of language on the truthfulness ratings, $F(2,128)=.18, p=.83$ (Figure 1). Contrary to what Lev-Ari and Keysar (2010) claimed, although the foreign speech is perceived as accented relative to the native speech, the accent did not change people's perceptions of truthfulness. However, the main claim of Lev-Ari and Keysar (2010) is that the difficulty associated with foreign-accented speech, and not necessarily the accent itself, is what drives the misattribution effect. It is reasonable to assume that although the foreign speech is accented, it might not necessarily be difficult to understand. On top of that, it might be that given that the content of the trivia statements are too opaque, participants in our study just did not engage in the task properly.

To further explore these points, we ran Study 3 using a more engaging decision-making task. We also assessed the level of difficulty on top of the level of accentedness for the non-native speakers. Study 3 used a design similar to (Shah \& Oppenheimer, 2007) who showed that people weigh fluent information more heavily than they weigh disfluent information. Using a similar design, we hypothesized that if accented speech is indeeed more difficult to process (i.e., dis-


Figure 1: Credibility ratings as a function of accent.
fluent) compared to native speech, participants would weight consumer reviews provided by non-native speakers less heavily than the same review provided by a native speaker.

## Study 3

Participants Sixty native speakers of English participated in Study 3. Participants were undergraduate students of psychology at The University of Texas at Austin and participated in exchange for course credit. None of the participants in Study 3 participated in the previous studies.
Materials Three female native speakers of English and three female non-native speakers of English (a BrazilianPortuguese speaker, an Iranian speaker and a Korean speaker) recorded both positive and negative reviews for six different products. To obtain equivalent overall amplitude levels for all recordings across the two speakers, the sound files were equated for RMS amplitude.
Difficulty and Accentedness Manipulation Check To ensure that the non-native speech was indeed perceived as more difficult to understand, a different group of 24 participants rated the degree of difficulty of the reviews. Each participant randomly heard a review for each of the six products. Three of these reviews were positive and three were negative. Three were from a native speaker and three were from a nonnative speaker (one for each non-native language). Participants rated the level of difficulty using a Likert scale ranging from 1 (easy) to 7 (difficult).

A repeated-measures two-way ANOVA, with language


Figure 2: Accentedness and Difficulty
(native vs. non-native) and valence (positive vs. negative) as independent variable and the difficulty ratings as dependent variables, revealed a reliable main effect of language, $F(1,22)$ $=91.51, p<0.001, \eta^{2}=.98$, but no statistically significant main effect of valence or language $x$ valence interaction. Notably, the main effect of language suggests that the non-native speech was perceived as reliably more difficult ( $M=3.54, S D$ $=1.66)$ than the native counterpart $(M=1.05, S D=0.23)$.

The same participants were also asked to rate how accented they perceived the reviews to be. Similarly to the results for the difficulty ratings, a repeated-measures two-way ANOVA, with language (native vs. non-native) and valence (positive vs. negative) as independent variable and the accent ratings as dependent variables, showed only a statistically significant main effect of language, $F(1,22)=260.5, p<0.001, \eta^{2}=$ .92 , suggesting that the non-native speech was perceived as significantly more accented ( $M=4.86, S D=1.44$ ) than the native counterpart ( $M=1.04, S D=0.2$ ). Overall, the pattern shows that the more accented, the more difficult to understand (see Figure 2).

Procedures Study 3 used a 2 (valence: positive vs. negative) X 2 (language: native vs. foreign) fully within-subject design. Each participant completed a total of 12 trials (six fillers and six experimental trials) that were presented to them in random order. Three of the trials were negative reviews and three were positive reviews. Language was also balanced per participant: three native speakers and three non-native speakers.

In each trial, participants were presented with a series of specifications about a product (e.g., this camera has 14.0 megapixels of resolution). The specifications were presented in the written format and were the same across conditions. After reading a product specifications they listened to a consumer review about the product. After listening to each review, participants were asked to estimate how much they


Figure 3: Price Estimates as a function of Accent and Valence
think the product should cost. For each product, participants were given a range or prices to estimate from.
Results and Discussion As the price intervals were different for each product, we standardized the estimates to be amounts between 0 and 1 . To investigate whether valence and language affected the prices estimates, we ran a repeatedmeasures two-way ANOVA with valence and language as independent variables and the standardized price estimate as dependent variable. We found a reliable main effect of valence, $F(1,58)=22.64, p<.001, \eta^{2}=.28$, suggesting that participants were indeed attentive to the content of the reviews, providing higher price estimates for the positive reviews ( $M=$ $.44, S D=.28$ ) than for the negative reviews ( $M=.17, S D$ $=.20$ ). However, no reliable main effect of language or interaction of language and valence were found (see Figure 3). Again, this pattern of results suggests that processing fluency associated with processing foreign accented speech does not affect cognitive judgments such as price estimation, $F(1,58)$ $=.14, p=.71$.

## General Discussion

In the present paper, we ran three studies to further explore the idea that the difficulty associated with foreignaccented speech affects cognitive judgments Lev-Ari and Keysar (2010). In both Studies 1a and 1b, the presence of noise made the statements significantly harder to understand than the statements spoken in quiet. This finding is consonant with several studies showing that processing speech in adverse conditions imposes an extra cognitive burden on listeners (Lane, 1962; Munro, 1998). Yet, this manipulation did not affect participants' judgments of truth.

For Study 2, even though the non-native speech was perceived as accented, they did not affect judgments of truth. This result is consonant with other research showing no rela-
tionship between degree of accent and credibility (De Meo, Vitale, Pettorino, \& Martin, 2011).

In Study 3, although the reviews spoken by non-native speakers were perceived as accented and difficult, they did not influence participants' price estimations. Taken together, these findings fail to support the claim that the processing difficulty associated with understanding non-native accented speech influences cognitive judgments.

The lack of effect of accent on cognitive judgment can be explained in terms of the kinds of masking (energetic versus informational) that accent and background noise causes to the speech signal. Energetic masking (also known as perceptual masking) occurs when there is a degradation of the acoustic signal in shared spectro-temporal regions. Because the energy of a speech signal is concentrated in a few spectrotemporal regions of high informational value, if masking takes place in other regions, little impact on speech processing will be observed (Cooke, 2006). On the other hand, informational masking (also known as conceptual masking) occurs when there is a reduction of speech intelligibility even after any energetic masking has been accounted for (Cooke, 2006). Generally, informational masking refers to distractions that directly competes with the listener's attentional resources when processing the speech (e.g., the presence of an unrelated task.)

Studies on speech processing and speech segmentation (Mattys, White, \& Melhorn, 2005; Mattys, Carroll, Li, \& Chan, 2010) have demonstrated that depending on the type of masking (energetic or informational), people will attend to different cues to process and segment the speech. Energetic masking (e.g. white noise) tends to favor the listener's reliance on lexical-semantic information whereas informational masking tends to favor the listener's reliance on sub-lexical, acoustic information. Related to our current findings, it might be that the presence of white noise (i.e., energetic masking) made our participants focus closely on the conceptual aspect of the message other than the acoustic features. Therefore, instead of producing a metacognitive feeling of disfluency, the presence of the white noise made it easier for people to focus on the declarative information of the speech.

Another alternative is linked to the evidence that listeners normalize accented speech before engaging in any sort of conceptual processing with the content of the speech (Lahiri \& Marslen-Wilson, 1991; Floccia, Goslin, Girard, \& Konopczynski, 2006). According to this view, the acoustic signal is cleaned of all distortions and deviant information and a "clean" signal is processed instead. This normalization process happens after short periods of exposure to accented speech. In fact, there is evidence that after sufficient information on the accent is gathered, comprehension strategies return to baseline levels (Floccia et al., 2006), making people less tuned to the acoustic properties of the signal. It is possible that the participants in our study normalized the accented speech after a short period of exposure and then neglected to attend to sub-lexical acoustic features of the speech.

## Conclusion

Our findings suggest that (Lev-Ari \& Keysar, 2010) findings might have been a case of a false positive. Although scientists always aim at publishing accurate and replicable effects, errors are inevitable (Simmons, Nelson, \& Simonsohn, 2011). In fact, the standard alpha level widely adopted in science (i.e., $5 \%$ ) means that about $5 \%$ of the time, when scientists look for an effect that is not there, they will find a statistically significant difference. The only way to spot such Type I Errors is by reducing the publication bias (Pashler \& Wagenmakers, 2012), that is, by giving more space in the literature for publications attempting to replicate previous findings. In that sense, replications and failures to replicate effects play an important role in the scientific arena (Makel, Plucker, \& Hegarty, 2012). The present paper contributes to the growing body of research interested in unvailing and understanding more systematically psychological phenomena.

## References

Alter, A., \& Oppenheimer, D. (2008). Effects of fluency on psychological distance and mental construal (or why new york is a large city, but new york is a civilized jungle). Psychological Science, 19(2), 161-167.
Alter, A., \& Oppenheimer, D. (2009). Uniting the tribes of fluency to form a metacognitive nation. Personality and Social Psychology Review, 13(3), 219-235.
Cooke, M. (2006). A glimpsing model of speech perception in noise. The Journal of the Acoustical Society of America, 119, 1562.
De Meo, A., Vitale, M., Pettorino, M., \& Martin, P. (2011). Acoustic-perceptual credibility correlates of news reading by native and chinese speakers of italian. Proceedings of the 17th International Congress of Phonetic Sciences.
Floccia, C., Goslin, J., Girard, F., \& Konopczynski, G. (2006). Does a regional accent perturb speech processing? Journal of Experimental Psychology: Human Perception and Performance, 32(5), 1276.
Gilbert, D. (1991). How mental systems believe. American psychologist, 46(2), 107.
Jacoby, L., \& Whitehouse, K. (1989). An illusion of memory: False recognition influenced by unconscious perception. Journal of Experimental Psychology: General, 118(2), 126.
Koriat, A. (1993). How do we know that we know? the accessibility model of the feeling of knowing. Psychological review, $100(4), 609$.
Kozou, H., Kujala, T., Shtyrov, Y., Toppila, E., Starck, J., Alku, P., \& Näätänen, R. (2005). The effect of different noise types on the speech and non-speech elicited mismatch negativity. Hearing research, 199(1), 3139.

Lahiri, A., \& Marslen-Wilson, W. (1991). The mental representation of lexical form: A phonological approach to the recognition lexicon. Cognition, 38(3), 245-294.

Lane, H. (1962). Foreign accent and speech distortion. The Journal of the Acoustical Society of America, 34(12), 1996-1996.
Lev-Ari, S., \& Keysar, B. (2010). Why don't we believe nonnative speakers? the influence of accent on credibility. Journal of Experimental Social Psychology, 46(6), 1093-1096. doi: 10.1016/j.jesp.2010.05.025
Makel, M., Plucker, J., \& Hegarty, B. (2012). Replications in psychology research how often do they really occur? Perspectives on Psychological Science, 7(6), 537-542.
Mattys, S., Carroll, L., Li, C., \& Chan, S. (2010). Effects of energetic and informational masking on speech segmentation by native and non-native speakers. Speech Communication, 52(11), 887-899.
Mattys, S., White, L., \& Melhorn, J. (2005). Integration of multiple speech segmentation cues: a hierarchical framework. Journal of Experimental Psychology: General, 134(4), 477.
Munro, M. (1998). The effects of noise on the intelligibility of foreign-accented speech. Studies in Second Language Acquisition, 20(2), 139-154.
Pashler, H., \& Wagenmakers, E. (2012). Editors' introduction to the special section on replicability in psychological science a crisis of confidence? Perspectives on Psychological Science, 7(6), 528-530.
Reber, R., \& Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. Consciousness and cognition, 8(3), 338-342.
Ryan, E., \& Giles, H. (1982). Attitudes towards language variation: Social and applied contexts. E. Arnold.
Schwarz, N. (2004). Meta-cognitive experiences in consumer judgment and decision making. Journal of Consumer Psychology, September.
Shah, A., \& Oppenheimer, D. (2007). Easy does it: The role of fluency in cue weighting. Judgment and Decision Making, 2(6), 371-379.
Simmons, J., Nelson, L., \& Simonsohn, U. (2011). Falsepositive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science, 22(11), 1359-1366.
Song, H., \& Schwarz, N. (2009). If it's difficult to pronounce, it must be risky fluency, familiarity, and risk perception. Psychological Science, 20(2), 135-138.
Unkelbach, C. (2007). Reversing the truth effect: Learning the interpretation of processing fluency in judgments of truth. Journal of Experimental Psychology: Learning, Memory, and Cognition, 33(1), 219.
Whittlesea, B., Jacoby, L., \& Girard, K. (1990). Illusions of immediate memory: Evidence of an attributional basis for feelings of familiarity and perceptual quality. Journal of Memory and Language, 29(6), 716-732.
Winkielman, P., \& Cacioppo, J. (2001). Mind at ease puts a smile on the face: psychophysiological evidence that processing facilitation elicits positive affect. Journal of personality and social psychology, 81(6), 989.

# Picture-Word Interference with Masked and Visible Distractors: Different Types of Semantic Relatedness Inhibit Lexical Selection 

Katharina Spalek (katharina.spalek@staff.hu-berlin.de)<br>Department of German Language and Linguistics, Humboldt-University Berlin, Unter den Linden 6, 10099 Berlin, Germany

Markus F. Damian (m.damian@bristol.ac.uk)<br>University of Bristol, School of Experimental Psychology, 12a Priory Road, Bristol BS8 1TU, United Kingdom


#### Abstract

One question in word production is how the presence of a semantically related word affects the naming process. It has been suggested that semantic effects in picture-word interference tasks are a net result of both inhibitory and facilitatory processes that take place at different processing levels. Finkbeiner and Caramazza (2006) argued that masking distractor words removes the inhibitory component, leaving only lexical facilitation. We investigated this claim by comparing different types of semantic relationship categorical relatedness, associative relatedness, and a combination of both - in picture-word interference with masked and visible distractors. We observed inhibitory effects in all conditions. In the visible condition, semantic category coordinates exerted the strongest inhibition, while in the masked condition, associatively related distractors interfered most. These findings are not easily reconciled with previous findings on polarity shifts of semantic effects with masked distractors. We discuss how all present findings could be explained within the same framework.


Keywords: lexical access, competition, response exclusion, picture-word interference, unconscious access

## Introduction

In the last decade, models of speech production that assume a competitive process of lexical selection (e.g., Levelt, Roelofs, \& Meyer, 1999) have been subjected to strong and sometimes heated criticism and equally passionate defense (see, e.g., Spalek, Damian, \& Bölte, 2012, for a summary of the arguments). The majority of empirical findings for (and against) the assumption of competitive lexical selection comes from experiments using the picture-word interference paradigm (e.g., Rosinski, 1977): Participants have to name pictures presented on the screen. Pictures are presented together with to-be-ignored distractor words (either in written or in spoken form). An often-repeated finding is that participants' responses are slower when the distractor word belongs to the same semantic category (e.g., fruit) as the target word (e.g., Schriefers, Meyer, \& Levelt, 1990) than when it belongs to an unrelated category. This has been taken as evidence for lexical competition: Target and distractor word are connected at the conceptual level through a common category node and prime each other. This results in two strongly activated representations, making the selection of the target representation more difficult and hence, more time-consuming (e.g., Roelofs, 1992).

An alternative explanation for the effects observed in the picture-word interference paradigm has been formulated in the so-called response exclusion hypothesis (e.g., Mahon et al., 2007): There is no competition between the lexical entries of a target word (the picture name) and a coactivated competitor (the distractor word). Interference arises at a later, post-lexical, processing level: Before a word can be pronounced, it occupies a single-channel output buffer. If the element in the buffer is the target word, it can be articulated; if it is the distractor, it has to be removed from the buffer before the target can enter the buffer and, eventually, be produced. According to Mahon and colleagues, the buffer knows about basic semantic properties of its entries. A word which is relevant to the experimental (or communicative) goal is more difficult to remove from the buffer than a word that is irrelevant to the task. Therefore, if the task is to name the picture of an animal, for example "dog", a distractor like "mouse" will be more difficult to remove from the buffer than a distractor like "pear".

The idea that interference occurs at a post-lexical processing level has received some support from findings with masked distractor presentation in picture-word interference studies: Finkbeiner and Caramazza (2006) had participants name pictures with visible distractors, replicating the semantic interference effect. When they presented the same stimuli but masked distractors such that participants weren't consciously aware of the distractors' identity, the semantic inhibition effect turned into a strong and reliable facilitation effect. Finkbeiner and Caramazza argue that the unconscious presentation prevented the distractor word from occupying the response buffer. Therefore, no competition effect was observed. However, the distractor words were still active enough to prime semantically related items in the mental lexicon, causing a net effect of facilitation. The finding that masking a distractor word turns inhibition into facilitation has been replicated by Dhooge and Hartsuiker (2010).

As noted by several researchers (e.g., LaHeij, Dirkx, \& Kramer, 1990), studies on semantic inhibition effects usually do not report the degree of association between target and distractor word. Pairs such as cat and dog and cat and horse are both related because they belong to the category animals. However, cat and $d o g$ are also associatively related because they often co-occur in the
language, and if people are asked to freely associate words in response to cat, dog is often one of the first associates produced. Finkbeiner and Caramazza (2006) don't provide a list of their materials, but Dhooge and Hartsuiker (2010) do. Perusal of their Appendix shows that they used categorically related picture-distractor pairs that were only weakly associated (e.g., spoon - knife; monkey - bear), but also pairs that were strongly associated (lion - tiger; apple pear), and, most critically, pairs that can be thought of as part-whole-relationships (farm - shed; pot - lid). The last type of relationship has been shown to cause facilitatory effects even in visible picture-word interference paradigms (Costa et al., 2005). While the data pattern observed by Dhooge and Hartsuiker is clear cut and shows a 15 ms interference effect in visible naming and a 12 ms facilitation effect in masked naming, it is possible that different items are responsible for the effects observed in visible and masked naming: If the inhibition observed with visible distractors is driven by the categorically related items, then a manipulation that makes them less salient competitors might allow the facilitation caused by the associated and part-whole relations to come to the fore.

In order to investigate this possibility, we used three different types of semantic relationship in our study, categorically related target-distractor pairs, associatively related target-distractor pairs, and categorically and associatively related (in the following: combined) targetdistractor pairs. Crucially, unlike in the study by Dhooge and Hartsuiker (2010), the categorically related items never were in a part-whole relationship, and strongly and weakly associated pairs were distributed across two different conditions. Before turning to our study, we will briefly review the literature on the effects of categorically and associatively related context words in picture naming.

Studies investigating categorical and associative relationships at SOA 0 (with written distractors) mainly found an effect of the former: Lupker (1979) used categorically related distractors and associatively related distractors in a picture-word interference study. He found that while the former caused interference, the latter had no effect. In a second experiment, he used distractors that were either categorically related or categorically and associatively related. He found that the inhibitory effect was exactly the same for both types of distractors. He concluded that categorical relatedness inhibits word production, and that this effect is not modulated by the association strength of the two category coordinates.

A study by LaHeij, Dirkx, and Kramer (1990) provides a different finding: They selected categorically related targetdistractor pairs that were either highly associated or weakly associated and used these items in a picture-word interference paradigm with different SOAs. At SOA 0, they observed inhibition for weakly associated category coordinates but not for highly associated ones. They argue that in the case of highly associated category coordinates the inhibitory effect is offset by an associative priming effect.

Investigating the time-course of these effects more closely, Alario, Segui, and Ferrand (2000) carried out an experiment on picture naming primed by pre-exposed words (in essence a picture-word interference paradigm with negative SOA). They discovered that associatively related words facilitate picture naming, but only if they are presented around 200 ms before picture onset. By contrast, categorically related words inhibit picture naming, but only if they are presented 100 ms (or less) before picture onset. So, it seems that associative relationships prime a target word whereas categorical relationships compete with a target word. However, these two mechanisms also seem to have a different time-course.

In contrast to the findings by Alario et al. (2000), Abdel Rahman and Melinger (2007) observed inhibition with categorically related distractor words and facilitation with associatively related distractor words with the same timecourse. In their study, spoken distractor words were presented 150 ms before the target pictures.

To sum up, the findings on associative distractor words in picture-word interference, while somewhat inconsistent, support the assumption that an associative relation between target and distractor facilitates target naming.

Given the observation that there is a facilitatory component to both associative and categorical distractors and that masking a distractor enhances the facilitatory component, we wanted to investigate how masking affects picture naming with categorically related, associatively related, and combined distractors. Participants named the pictures both with visible distractors and with masked distractors. For visible distractor presentation we predict an interference effect for categorically related distractor words. For associatively related distractor words, we expect to see either a facilitatory effect or a null effect. Finally, for combined items, we expect to see either an effect of equal size as for the categorically related items (as Lupker, 1979, did) or an attenuation of the effect as in LaHeij et al. (1990).

If masking the distractor effectively prevents it from entering a response buffer, no inhibitory effects are expected in the masked condition. Instead, categorically related and associatively related distractors should yield facilitation which should be greatest for combined distractors.

A second aim of the study was to address a concern formulated by Kouider and Dupoux (2004). They question whether previous studies on unconscious priming truly presented words in a subliminal manner, and argue that participants are typically at least partially aware of a masked stimulus and that this partial awareness causes the priming effect. We carried out a lexical decision task on the distractor words after the picture naming study. Distractors were masked in the same way as during the picture naming study. Assessing participants’ performance in the lexical decision task gave us a tool to investigate in how much (partial) awareness of the distractors modulated the effects.

## Method

## Participants

Forty-eight native speakers of German (thirty-five women) were recruited from the participant database of the Institute of Psychology at the Humboldt-University Berlin. Their mean age was 24.2 years. Participants received monetary compensation for their participation.

## Materials

Twenty pictures of animals and objects were chosen as targets. For each of the pictures (e.g., picture LEMON), three distractor words were selected: a semantically related word (i.e., a category coordinate, e.g., kiwi), an associatively related word (i.e., a word from a different category, e.g., vitamin), and a semantically and associatively related word (e.g., orange). Distractor words were matched on length and frequency. The associative relation was determined pre-hoc by the intuitions of two native speakers of German and backed up post-hoc by associative relatedness ratings of the participants. Participants were asked to rate the strength of the association of two words on a scale from 1 (not associated) to 7 (very strongly associated). The categorically related items had an association strength of 2.93, the associatively related items an association strength of 4.00, and the combined items had an association strength of 5.57. As intended, the categorically related items were less strongly associated than both the associatively related items $(t(19)=4.09, p<.001)$ and the combined items $(t(19)=$ 4.62, $p<.001$ ). What was not intended was that the association strength was also higher for combined items than for associatively related items $(t(19)=11.92, p<.001)$.

We created three unrelated conditions by recombining the related distractors with different pictures. Therefore, in each of the three conditions (categorically related, associatively related, combined), the same pictures and the same words were used in both the related and the unrelated condition.

Each participant saw a target word in all six conditions (three critical conditions and three control conditions). A different randomization was created for each participant to avoid order effects.

For the lexical decision task (see below), 20 non-words were created by using existing words and replacing one or two letters. These letter changes could occur in any position in the word. Care was taken to change each position equally often. Non-words were matched in length to the word targets.

## Procedure

Participants carried out three different tasks: the pictureword interference study, a lexical decision task and a questionnaire. Order of presentation for the picture-word interference studies (visible vs. masked) was counterbalanced across participants. The questionnaire contained all related target-distractor pairs. Participants were asked to indicate how strong the association between
the two concepts is, using a scale from 1 (not associated) to 7 (strongly associated). The experiments were programmed and run with Presentation (NeuroBehavioral Systems).

Visible Distractor Presentation. Participants were instructed to name the pictures on the screen and to ignore the superimposed distractor words. A trial started with a fixation cross that was presented for 500 ms . The word was presented centered on the screen for 53 ms . Picture and word were presented together for 2000 ms . Participants' responses triggered a VoiceKey and were recorded.

Delayed Distractor Presentation Participants were instructed to name the pictures on the screen and to ignore the superimposed distractor words. A trial started with a forward mask (\#\#\#\#\#\#\#\#\#\#) that was presented for 500 ms . The word was presented centered on the screen for 53 ms . It was replaced by the picture and a non-pronounceable mask consisting of a string of 10 consonants presented in the same location as the distractor word. The use of a consonant string as a backward mask was motivated by Finkbeiner and Caramazza (2006) who refer to findings having shown its particular effectiveness in eliminating phonological priming effects. Picture and mask were presented together for 2000 ms. Participants' responses triggered a VoiceKey and were recorded.

Lexical Decision Task A forward mask (\#\#\#\#\#\#\#\#\#\#) was presented for 500 ms centered on the screen. It was followed by a letter string that was presented for 53 ms . The letter string was replaced by the same mask as in the masked picture-word interference paradigm. The mask stayed in place until the participant had made a response. Participants were instructed to decide whether the briefly presented word had been an existing word of their language or not. They were encouraged to make a guess if they felt they had not seen a word at all. The results of the lexical decision task will not be analysed in the present paper, we merely used participants' overall accuracy in order to split the group in a "high-recognition" and a "low-recognition" group (see below).

## Results

We carried out an ANOVA on the mean reaction times and error rates with the within-subject and within-item factors Type of Relationship (Categorical, Associative, Combined) and Relatedness (Related vs. Unrelated).

Table 1 presents the mean reaction times and error rates in the visible distractor condition. Table 2 presents these measurements in the masked distractor condition.

In the visible condition, for the reaction times, the effect of Type of Relationship was highly significant $\left(F_{1}(2,94)=\right.$ 16.61, $\mathrm{MSE}=433, p<.001 ; F_{2}(2,38)=5.31, \mathrm{MSE}=537, p$ $<.01)$, as was the effect of Relatedness $\left(F_{1}(1,47)=17.54\right.$, MSE $=412, p<.001 ; F_{2}(1,19)=6.75$, MSE $\left.=498, p<.05\right)$, showing faster reaction times for unrelated distractors than
for related distractors. The interaction of the two factors was not significant (both Fs $<1$ ).

For the error rates, the effect of Type of Relationship was not significant $\left(F_{1}(2,94)=1.57\right.$, MSE $=0.04, p=.21, F_{2}<$ 1). The effect of Relatedness was marginally significant with slightly higher error rates for related distractors $\left(F_{1}(1,47)=3.65, \mathrm{MSE}=0.05, p=.06, F_{2}(1,19)=1.82, p=\right.$ .19). The interaction was not significant (both $F$ s $<1$ ).

Table 1: Reaction times and error rates (in brackets) in the visible condition.

|  | Categorical | Combined | Associative |
| :--- | :--- | :--- | :--- |
| Related | $646(1.5)$ | $631(1.8)$ | $625(1.5)$ |
| Control | $631(0.7)$ | $622(1.5)$ | $619(0.9)$ |
| Effect | $15(0.8)$ | $9(0.3)$ | $6(0.6)$ |

Table 2: Reaction times and error rates (in brackets) in the masked condition.

|  | Categorical | Combined | Associative |
| :--- | :--- | :--- | :--- |
| Related | $622(1.1)$ | $613(0.5)$ | $615(0.7)$ |
| Control | $616(0.5)$ | $610(0.4)$ | $608(0.6)$ |
| Effect | $6(0.6)$ | $3(0.1)$ | $7(0.1)$ |

In the masked condition, for the reaction times, the effect of Type of Relationship was significant $\left(F_{1}(2,94)=4.05\right.$, $\left.\operatorname{MSE}=414, p<.05 ; F_{2}(2,38)=4.62, \mathrm{MSE}=166, p<.05\right)$, as was the effect of Relatedness $\left(F_{1}(1,47)=6.96\right.$, MSE $=$ 297, $\left.p<.05 ; F_{2}(1,19)=6.00, \mathrm{MSE}=137, p<.05\right)$, again showing inhibition for related distractors. The interaction of the two factors was not significant (both Fs < 1).

For the error rates, the effect of Type of Relationship was not significant $\left(F_{1}(2,94)=1.44\right.$, MSE $=0.02, p=.24, F_{2}<$ 1). The effect of Relatedness was significant by items $\left(F_{1}(1,47)=1.54, \mathrm{MSE}=0.03, p=.22, F_{2}(1,19)=4.75\right.$, MSE $=0.004, p<.05$ ). The interaction was not significant $\left(F_{1}(2,94)=1.04, \mathrm{MSE}=0.02, p=.36, F_{2}(2,38)=2.02\right.$, MSE $=0.004, p=.15)$.

In order to investigate if the masking manipulation affected the critical effects, we pooled the data of both experiments and carried out an ANOVA with the factors Experiment, Type of Relationship, and Relatedness. We observed a significant effect of Experiment with faster reaction times for masked distractors $\left(F_{1}(1,47)=5.43\right.$, MSE $\left.=5877, p<.05, F_{2}(1,19)=17.61, \mathrm{MSE}=772, p<.001\right)$. The factors Type of Relationship $\left(F_{1}(2,94)=19.01\right.$, MSE $=$ 406, $p<.001, F_{2}(2,38)=7.51$, $\left.\mathrm{MSE}=426, p<.001\right)$, and Relatedness $\left(F_{1}(1,47)=21.37\right.$, MSE $=398, p<.001$, $F_{2}(1,19)=12.25$, MSE $\left.=306, p<.01\right)$ also had a significant effect. Importantly, we observed a marginally significant interaction of Experiment and Type of Relationship by participants $\left(F_{1}(2,94)=2.62, \mathrm{MSE}=442, p=.08, F_{2}(2,38)\right.$ $=1.52$, $\mathrm{MSE}=277, p=.23$ ).

Because the interaction, albeit rather weak, suggests that the effects for the different types of relationship might differ
in the visible and in the masked condition, we carried out paired $t$-tests for all three types of relationship in the two visibility conditions.

In the visible condition, the only reliable inhibition effect (by participants) was observed with categorically related distractors, $t_{1}(47)=2.76, p<.01, t_{2}(19)=1.92, p=.07$. In the combined condition, there was only a trend by participants, $t_{1}(47)=1.97, p=.054 ; t_{2}(19)=1.47, p=.16$. Finally, the effect for the associatively related condition was not significant, $t_{1}(47)=1.63, p=.11, t_{2}(19)=1.04, p=.31$.

In the masked condition, the pattern was reversed : The categorical relatedness effect was not significant, $t_{1}(47)=$ $1.50, p=.14, t_{2}(19)=1.47, p=.16$, and neither was the combined effect, both $t \mathrm{~s}<1$. By contrast, the associative relatedness effect was significant by participants $\left(t_{1}(47)=\right.$ 2.33, $p<.05$ ) and approached significance by items ( $t_{2}$ (19) $=1.83, p=.08$ ).

Finally, we split the subjects in two groups, based on their accuracy in the masked lexical decision task. We used a median split, with the "low-recognition group" being correct on $49 \%-73 \%$ of all trials and the "high-recognition group" being correct on $73 \%-93 \%$ of all trials. We reanalyzed the data set with the additional between-subjects variable "Recognition" (high vs. low). There was no main effect of recognition (both $F s<1$ ) and no higher-level interactions of Type of Relationship and Relatedness with Recognition (all ps > .20). Even though the ANOVA showed that Recognition did not affect the data pattern, we present the descriptive data for the high- and low-recognition group in the masked condition in Tables 3 and 4.

Table 3: Results for the high-recognition group in the masked condition

|  | Categorical | Combined | Associative |
| :--- | :--- | :--- | :--- |
| Related | 621 | 609 | 615 |
| Control | 617 | 610 | 609 |
| Effect | 4 | -1 | 6 |

Table 4: Results for the low-recognition group in the masked condition

|  | Categorical | Combined | Associative |
| :--- | :--- | :--- | :--- |
| Related | 622 | 616 | 616 |
| Control | 615 | 609 | 607 |
| Effect | 7 | 7 | 9 |

The descriptive data show that, if anything, those participants who recognized the masked distractors less well showed the stronger inhibitory effects in the picture-word interference study.

## Discussion

Overall, we found inhibitory effects of semantically related distractors in both masked and visible distractor presentation in a picture-word interference paradigm.

The findings for the visible distractor presentation are in line with many previous findings. We observed semantic interference. The lack of a significant interaction between Type of Relationship and Relatedness suggests that the effect was equally strong for all types of relationship. The post-hoc $t$-tests, by contrast, hint at a possible difference: The interference is strongest for categorically related distractors and statistically absent for associatively related distractors, with categorically and associatively related distractors patterning in-between. This finding is generally in accordance with the result obtained by LaHeij et al. (1990).

The observation that interference persists for masked distractors is at odds with the two previous studies (Dhooge \& Hartsuiker, 2010; Finkbeiner \& Caramazza, 2006) discussed in the Introduction. Ignoring for the moment the issue of how different types of semantic relationship might affect the results and simply focussing on the categorical condition, the results do not indicate the predicted polarity reversal from inhibitory (visible) to facilitatory (masked) effects. At a general level, our data show that the polarity reversal for semantic effects dependent on distractor visibility is less universal than suggested by the previous studies.

A possibility is that in our "masked" experiment, the specific masking procedure was not exactly identical to the previous studies in terms of distractor visibility. Although an effort was made to keep all relevant parameters (e.g., distractor duration, backward mask, etc.) as similar as possible, relatively minor variations in, e.g., contrast or display size could potentially affect distractor visibility. It is also the case that our study used a different language from the original studies (English and Dutch), and that words perhaps had slightly different properties. For instance, German words tend to be longer on average than English words, therefore, the amount of information that can be extracted from a word presented under masked conditions might differ among languages. Hence, perhaps our masked distractors were either too heavily masked, or not masked well enough. The first scenario - masking of distractors was too stringent - is refuted by the simple fact that we did observe a significant effect of relatedness in that experiment. Could it therefore be that our distractors were not masked well enough, i.e., that they acted in the same way as visible distractors, and hence induced similar inhibitory effects? The strongest argument against this possibility comes from the comparison of participants who recognized more words during the visibility test with those who recognized fewer words. If the inhibition effect observed in the masked condition is due to the fact that participants recognized the masked distractors too well, then the inhibitory effect should be strongest for those participants who recognized the distractors the best. Contrary to this prediction, there was no significant difference in the data pattern for "good" and "poor" recognizers; indeed, descriptively the inhibitory effect was larger for the "poor" than the "good" recognizers. This
renders it unlikely that heavier masking (or perhaps, a reduction in distractor duration) would have resulted in the predicted facilitatory effect of semantic relatedness.

While we cannot say at this point which differences in the experimental procedure have caused the differences in results, it is clear that the semantic facilitation effect with masked distractors is much more susceptible to such procedural differences than the semantic inhibition effect with visible distractors. Therefore, caution is needed when using this effect for theory-building and it is necessary to better understand the experimental conditions that allow for a polarity shift.

A second important finding is that while inhibition occurred in both presentation conditions and was not statistically modulated by the exact type of semantic relationship, there were still some crucial differences. In the visible condition, the categorically related condition caused the greatest inhibition, whereas in the masked condition, the associatively related condition caused the greatest inhibition. Explanations for this pattern remain at present speculative. One possible scenario derived from earlier work on such relationships (Alario et al., 2000; La Heij et al., 1990) is that associative pairings represent direct interlexical links, perhaps at a "peripheral" level (i.e., the orthographic or phonological lexicon). If so, it is conceivable that links at such "shallow" processing levels would be more dominant with masked distractor presentation, compared to visible distractors whose effect might emerge more clearly at "deeper" (i.e., lexicalsemantic or conceptual) processing levels. To our knowledge, our study represents the first attempt to address the possible dependency of effects of various types of semantic relationships on distractor visibility, and more research is clearly needed.

From a broader perspective, our data, combined with the earlier studies in the literature reporting a polarity reversal, contest the assumption that the inhibitory component in speech selection is binary in the sense that either a distractor will enter the competition or not. Rather, inhibition and facilitation can be relatively stronger or weaker, modulating the net outcome. Roelofs, Piai, and Schriefers (2011) suggest that masking a distractor word results in this word receiving a smaller weight in the competition process. Such a mechanism, depending on the magnitude of the weight change, could accommodate the entire continuum of effects. That is, for clearly visible distractors, the distractor will receive a high competition weight, resulting in an inhibitory effect. As visibility decreases, the competition weight will decrease, too, reducing the inhibitory component of the effect. With a very low competition weight, the facilitatory component of the effect (i.e., the target is primed by the distractor) will result in a facilitation effect. The challenge for future experiments would then be to precisely predict the size of the competition weights in different contexts. The response exclusion hypothesis is less able to explain such a smooth transition from inhibition to facilitation. Intermediate effects could be explained by the response
exclusion hypothesis as an experimental artifact, if either facilitation or inhibition is observed on a trial-by-trial basis (i.e., if a participant observes a word in a given trial, it will enter the buffer and interfere, if (s)he does not observe a word in a different trial, it will not enter the buffer and therefore, facilitation will result). While the sum of trial-bytrial inhibition and facilitation might result in anything from facilitation to inhibition, too, this explanation is refuted by our finding that, numerically at least, the inhibition effects in masked distractor presentation were larger for those participants, who perceived the masked words less well.

In conclusion, previous studies have reported a polarity reversal of semantic effects in picture-word interference tasks, such that clearly visible distractors which are semantically related to the picture name generate interference, whereas visual masking of such distractors results in facilitation. This pattern was taken as supporting different loci of the facilitatory and interfering components. In our own experiments we were unable to replicate this polarity reversal; instead, our findings suggest that significant semantic interference can prevail even under masked conditions, but that the precise pattern might depend on the exact form of semantic relationship between distractor and target.

## Acknowledgments

The research presented in this paper was supported by DFGgrant SP1126/4-1 to KS. We thank Julia Knoepke and Carsten Schliewe for their help in programming the experiment and Julia Knoepke for testing the participants.

## References

Abdel Rahman, R., \& Melinger, A. (2007). When bees hamper the production of honey: Lexical interference from associates in speech production. Journal of Experimental Psychology: Learning, Memory and Cognition, 33, 604-614.
Alario, F.-X., Segui, J., \& Ferrand, L. (2000). Semantic and associative priming in picture naming. Quarterly Journal of Experimental Psychology, 53A, 741-764.
Costa, A., Alario, F.-X., \& Caramazza, A. (2005). On the categorical nature of the semantic interference effect in the picture-word interference paradigm. Psychonomic Bulletin and Review, 12, 125-131.
Dhooge, E., \& Hartsuiker, R. J. (2010). The distractor frequency effect in picture-word interference: evidence for response exclusion. Journal of Experimental Psychology: Learning, Memory, and Cognition, 36, 878891.

Finkbeiner, M., \& Caramazza, A. (2006). Now you see it, now you don't: On turning semantic interference into facilitation in a Stoop-like task. Cortex, 42, 790-796.
Kouider, S., \& Dupoux, E. (2004). Partial awareness creates the "illusion" of subliminal semantic priming. Psychological Science, 15, 75-81.

La Heij, W., Dirkx, J., \& Kramer, P. (1990). Categorical interference and associative priming in picture naming. British Journal of Psychology, 81, 511-525.
Levelt, W. J. M., Roelofs, A., \& Meyer, A. S. (1999). A theory of lexical access in speech production. Behavioral and Brain Sciences, 22, 1-75.
Lupker, S. J. (1979). The semantic nature of response competition in the Picture-Word Interference Task. Memory \& Cognition, 7, 485-495.
Mahon, B., Costa, A., Peterson, R., Vargas, R. A., \& Caramazza, A. (2007). Lexical selection is not by competition: A reinterpretation of semantic interference and facilitation effects in the picture-word interference paradigm. Journal of Experimental Psychology: Learning, Memory, and Cognition, 33, 503-533.
Roelofs, A. (1992). A spreading-activation theory of lemma retrieval in speaking. Cognition, 42, 107-142.
Roelofs, A., Piai, V. M., \& Schriefers, H. (2011). Selective attention and distractor frequency in naming performance: Comment on Dhooge and Hartsuiker (2010). Journal of Experimental Psychology: Learning, Memory, and Cognition, 37, 1032-1038.
Rosinski, R. R. (1977). Picture-word interference is semantically based. Child Development, 48, 643-647.
Schriefers, H., Meyer, A. S., \& Levelt, W. J. M. (1990). Exploring the time course of lexical access in production: Picture-word interference studies. Journal of Memory and Language, 29, 86-102.
Spalek, K., Damian, M. F. \& Bölte, J. (2012). Is lexical selection in spoken word production competitive? Introduction to the special issue on lexical competition in language production. Language and Cognitive Processes, DOI: 10.1080/01690965.2012.718088.

# Implications of Polychronous Neuronal Groups for the Nature of Mental Representations 

William Benjamin St. Clair (wst.clair@ucmerced.edu) David C. Noelle (dnoelle@ucmerced.edu)<br>Cognitive and Information Sciences<br>University of California, Merced<br>5200 North Lake Road; Merced, CA 95343 USA


#### Abstract

How are concepts represented in the human mind? Vector space representations based on the instantaneous firing rates of neurons have been used with great success. However, there is growing evidence, both empirical and computational, that relevant information is encoded in spatiotemporal patterns of spikes called polychronous neuronal groups (PNGs). In this paper, we consider the philosophical implications of PNG representations with regard to their temporal extension, grounding, compositionality, and similarity. We suggest that the temporally extended nature of PNGs implies that conceptual-level dynamics may only be coherent at coarse time scales. We introduce the notion of PNG trigger sets as a way to ground the meaning of PNG representations, and we discuss potential approachs to compositionality. Finally, we identify the open problem of how to define an appropriate similarity metric for PNG-based mental representations.


Keywords: Philosophy of Cognitive Science; Neuroscience; Representation; Dynamical Systems.

## Introduction

How are concepts represented in the human mind? One highly productive approach to this question has involved the application of a continuous dynamical systems perspective to the problem (Spivey, 2008). From this perspective, the currently active concepts of a cognitive system (or subsystem) are jointly encoded as a point in a high dimensional vector space (Churchland, 1989). Nearby points in this space, according to some distance metric, are seen as representing similar conceptual states, allowing regions and manifolds within this space to capture more general concepts or categories (Gärdenfors, 2000). The evolution of mental states over time becomes a trajectory in this vector space (Yoshimi, 2012), driven by mechanistic cognitive processes (Churchland, 1996).

This vector space approach to conceptual representation has many strengths. It supports accounts of the biological basis of cognition by viewing the dimensions of the vector space as the activity of neural units, such as their firing rates, offering a framework for mapping from the physical state space of the brain to the conceptual state space of the mind (Spivey, 2008). The approach provides a straightforward way to discuss the cognitive state of a system at any instant in time, as well as how those states change over time. Issues surrounding the grounding of representations are well addressed from the perspective of conceptual role semantics (Greenberg \& Harman, 2006), with the "meaning" of a representation being a function of the inputs that activate it and the other representations that it produces through causal relationships,
eventually making causal contact with the world through sensory and motor processes. When using the vector space approach, these cognitive processes are well described in neurocomputational terms, with nearby points in a vector space tending to produce similar results when presented to models of downstream neural circuits. While the vector space approach has been criticized as lacking support for compositional and structured conceptual representations (Fodor \& Pylyshyn, 1988), and there continues to be extensive work on addressing this critique (Gayler \& Levy, 2011), highly promising approaches to compositionality have been proposed, making use of vector space operations of superposition, convolution, and sparse coding (Smolensky, 1990; Plate, 2003; OReilly, Bhattacharyya, Howard, \& Ketz, 2011). In general, the vector space approach to conceptual representation has been very productive.

Past challenges to the vector space approach have come from above: from more abstract and symbolic characterizations of cognitive content and cognitive processing. More recently, a challenge has arisen from below: from insights into the neural coding of information (Rieke, Warland, Steveninck, \& Bialek, 1999). There is increasing empirical evidence that, in at least some neural systems, relevant information is encoded in the spatiotemporal pattern of discrete action potentials, or spikes, produced by neurons in a given nucleus (Rolston, Wagenaar, \& Potter, 2007; Madhavan, Chao, \& Potter, 2007; Pasquale et al., 2008). While information, in some cases, may be carried by synchronous or coherent firing of neurons (Fries et al., 2005), computational considerations have suggested that content may frequently be encoded in complex asynchronous patterns of spikes (Izhikevich, 2006). These complex spike patterns have been called polychronous neuronal groups (PNGs). The PNG approach to representation differs substantially from the vector space approach. A PNG is a temporally extended pattern of discrete spiking events over a collection of neurons, and it is not clear how such a pattern could be mapped to a point in a continuous vector space so as to preserve relevant aspects of similarity between representations. In the vector space approach, temporally extended trajectories capture dynamic changes in cognitive content, while the PNG account encodes individual conceptual states in such trajectories. A PNG need not be oscillatory, so it does not make sense to extract features like frequency or phase to map a PNG into a vector space. In some important ways, the PNG approach is fundamentally
different than the vector space approach.
It may be tempting to view the vector space account as supervening on the PNG account, with PNGs implementing vector space states at some lower level of analysis. In this brief article, we argue that such a view either is untenable, with no mapping from the complexities of PNG representations to points in a vector space being possible, or, at least, is in dire need of an explanation of how such a reduction might be accomplished. Specifically, we raise four problematic issues that arise when shifting from a vector space approach to a PNG approach: (1) Temporal Extension - If conceptual representations involve temporally extended PNGs, to what degree can a conceptual state be said to be active at a particular instant or actively maintained over an interval? How does the use of the PNG framework change the characterization of the evolution of conceptual states over time? (2) Grounding - If conceptual role semantics is to be used to understand the "meaning" of a PNG representation, what are the biologically realistic causal mechanisms that link PNGs in an inferential cascade? (3) Compositionality - Does the PNG approach fare better or worse than the vector space approach in accommodating compositional or structured representations? (4) Similarity - Is there a distance measure for PNG representations that could be used to capture conceptual similarity while reflecting the way in which downstream neural circuits would naturally generalize across disparate PNGs?

Each of these four issues is elaborated in the following sections, and preliminary insights into how these issues could be addressed are provided. The goal is to highlight how the PNG approach challenges the prevailing vector space account of mental representation, while offering some clues concerning how this challenge might be met.

We begin by offering a review of PNG representations, providing a foundation for exploring each of the four issues that we find problematic. We then conclude with a brief discussion of open questions in this domain.

## Polychronous Neuronal Groups

Polychronous Neuronal Groups (PNGs) have been proposed as a possible unit of representation in the human brain (Izhikevich, 2006). A PNG is a reproducible, timelocked, spatiotemporal spike-timing pattern over a collection of neurons. They are reproducible in the sense that the sequence of spike times tends to replay when the input conditions experienced by the neural network are repeated. They are time-locked in the sense that, once the PNG begins, the times between the spikes within the spiking pattern are the same whenever the PNG is triggered. They are spatiotemporal in the sense that they are defined in terms of a specific set of neurons that participate in the pattern (spatial) as well as the specific times at which spikes appear in the pattern (temporal). PNGs spontaneously emerge in spiking neural networks that incorporate variance in the amount of time it takes for an action potential to reach its receiving neurons (conductance delays), and they are reinforced by mechanisms of spike
timing dependent plasticity (STDP) (Izhikevich, 2006).
To understand the information-bearing properties of PNGs, it is important to understand how they are generated and propagated. An individual neuron remains at its resting potential until it receives, or "observes", a sufficient number of spikes in a short enough period of time, at which point this coincident input causes the neuron to generate an action potential of its own. This action potential is then, in turn, observed by the neurons to which this neuron projects. However, since it takes time for action potentials to propagate down axonal connections, there is a delay between when a spike is generated and when it is received. For example, in the cat brain, this delay can be as short as 0.1 ms , or as long as 44 ms (Swadlow, 1992). Since a cortical neuron may project to anywhere between 1,000 and 10,000 other neurons, a single action potential will be received at many different times. Thus, spikes that are synchronized on generation will not necessarily be synchronized on their receipt.

Typically, a single input spike is insufficient to drive the receiving neuron to fire an action potential, and the membrane potential of such a neuron is constantly decaying toward its resting potential. Within just a few milliseconds after receiving a single spike, the membrane potential of a neuron will return to its equilibrium state, removing the electrical effects of the spike (Cessac, Paugam-Moisy, Viéville, et al., 2010). This highlights the need for synchrony in the arrival of spikes to initiate firing, but it is important to remember that spikes that are synchronized at the time of receipt will not necessarily be synchronized at the time of their initiation, due to variance in conductance delays.


Figure 1: A small neural network with time delays.

Consider the network portrayed in Figure 1. If neurons $a$, $b$, and $c$ spike at the same time, Time 0 , those spikes will be received by neuron $x$ at Times 1,5 , and 9 , respectively, and those same spikes will be received by neuron $y$ at Times 8 , 5 , and 1 , respectively. In this case, neither $x$ nor $y$ receive the coincident spikes needed to fire. The difference in arrival times are caused by differences in axonal propagation times. If, instead, neurons $a, b$, and $c$ spike at Times 8,4 , and 0 , respectively, neuron $x$ will receive all three of these spikes at Time 9, potentially allowing the cell to fire. In contrast, neuron $y$ would receive the three spikes at Times 16, 9, and

1 , respectively, providing it with no coincident spikes to drive an action potential. Alternatively, if neurons $a, b$, and $c$ fire in the reverse order, neuron $y$ will may spike, while neuron $x$ will remain silent. Thus, the effects of spikes from neurons $a$, $b$, and $c$ on the firing of neurons $x$ and $y$ is critically dependent on the timing of the spikes.

In larger, more connected, networks, like those found in mammalian brains, a particular stimulus will cause a chain reaction of spikes over time. This group of neurons firing with precise timing is what forms a corresponding PNG. Importantly, PNG patterns can be strengthened with repetition, with the strengthening well explained by spike timing dependent plasticity (STDP) (Izhikevich, 2006). Synapses that exhibit STDP are strengthened whenever the post-synaptic neuron fires just after it receives evidence of a pre-synaptic spike. Conversely, whenever the post-synaptic neuron fires just before it receives evidence of a pre-synaptic spike, then the synapse is weakened (Dan, Poo, et al., 2004). Thus, as a PNG unfolds, STDP strengthens the synapses participating in the PNG's generation and weakens the synapses that were active but did not facilitate the firing of neurons participating in the PNG. Thus, every time a particular PNG unfolds, and hence becomes strengthened via the mechanism of STDP, it becomes easier for that PNG to be reproduced.

To restate, a PNG is a reproducible, time-locked, spatiotemporal spike-timing pattern. A PNG is reproducible in the sense that, when the neurons participating in a PNG are stimulated in a similar way, the PNG will unfold in a similar way. Furthermore, each reproduction causes a PNG to become more stable through STDP, making the PNG increasingly robust to timing noise (i.e., some input spikes may be omitted or added without substantially effecting the generation of the PNG). A PNG is time-locked due to the fact that the conductance delays between the participating neurons are fixed by the anatomy of the network. A PNG is spatiotemporal in the sense that it necessarily occurs at many times (polychronous) and involves many neurons. Once stabilized via STDP, subtle variations in spike timing due to noise do not lead to unpredictablely different PNGs, but generate a member of a family of related PNGs (Izhikevich, 2006). Also, it is important to note that many PNGs may be simultaneously active in a common neural network without interacting, due to the low probability that two arbitrary PNGs will overlap substantially in their precise spike times. In addition to these properties, PNGs also minimize redundancy through the weakening of synapses via STDP, and they are more energy efficient than vector space representations that depend on neural firing rate (Levy \& Baxter, 1996). It is also interesting to note that the "small world" connectivity structure of the mammalian brain gives rise to stable PNGs much more readily than networks of neurons that are connected uniformly at random (Sporns \& Zwi, 2004; Vertes \& Duke, 2010).

Since their introduction, PNGs have been utilized extensively in computational neuroscience models of cognitive information processing. The intricate dynamics of PNGs have
been used in combination with models of NMDA receptors and neurotransmitter reuptake to produce a promising account of working memory function (Szatmáry \& Izhikevich, 2011). PNGs have been incorporated into a formal account of the dopamine system in order to produce a candidate model of neural reinforcement learning that addresses the problem of temporally distal reward (Izhikevich, 2007). In addition to their use in computational neuroscience models, empirical evidence for PNGs has been reported, with reproducible, time-locked, spatiotemporal patterns of spikes being observed in cortical slices (Rolston et al., 2007). It is clear that PNGs show great promise as a form of representation in the brain.

As this review of PNGs shows, the information carried by a PNG in a neural network is critically dependent on the timing of individual action potentials. This contrasts with vector space accounts of mental representation which map vector space dimensions onto the instantaneous firing rates of neurons. The PNG approach highlights the way in which individual spike times can carry information, with spiking rates lacking sufficient spatiotemporal detail to discriminate between different representational states. It is this shift that gives rise to a number of potential problems with viewing PNGs as the foundation of mental representation.

## Temporally Extended Representations

## The Challenge

In vector space representation schemes that make contact with biology, each dimension corresponds to the instantaneous activity of a neural element, such as the instantaneous firing rate of a neuron. This provides us with a natural way to capture the mental state of the agent at any given point in time. The active maintenance of a conceptual state involves a relative lack of change in these firing rates, and the evolution of mental states over time are captured in trajectories through the vector space.

In contrast, the PNG approach inherently involves temporally extended representations. A PNG is a spatiotemporal pattern of spikes. If mental representations are to be associated with PNGs, to what degree can any concept be seen as active at any given instant? Can a representation be actively maintained over time if the physical substrate of the representation is changing over time? How does the use of the PNG framework change the characterization of the evolution of conceptual states over time?

## Addressing the Challenge

While some philosophical work may be needed to fully appreciate the nature of temporally extended mental representations, we do not see this challenge as insurmountable.

The activation level of a particular PNG at an arbitrary point in time, in the midst of a sequence of spikes, does not have a clear definition. We can identify, however, the degree to which recently produced spikes match portions of a PNG, as well as the propensity for the neural network to continue
with the production of further spikes in the PNG. Thus, the notion of the activation of a concept is only coherent at time scales that match the time scale of the PNG. For example, if the PNG $\pi$ refers to a 30 ms long spike sequence, it may be asked if, over the last $30 \mathrm{~ms}, \pi$ appeared. For that same $\pi$, it may also be asked at what times over the last $60 \mathrm{~ms} \pi$ was present. In this way, activation of a mental state represented by a PNG only makes sense at relatively coarse time scales in comparison to the time scales used for common vector space representations. The PNG approach does not admit to a coherent sense of a truly instantaneous mental state.

The mental state encoded by a PNG may be actively maintained for a period of time longer than the duration of the spike sequence that makes up the PNG. For example, PNG models of working memory have involved the repeated activation of a PNG, with the spike pattern being sequentially reinitiated, allowing it to persist for arbitrarily long periods, as needed (Szatmáry \& Izhikevich, 2011). Once again, this notion of active maintenance is limited to a time scale corresponding to the temporal length of the PNG, but the PNG approach does not rule out the possibility of persisting in a mental state for a longer period of time.

## Grounding Of Representations

## The Challenge

The symbol grounding problem highlights the need for representational schemes to provide some account of the meaning of mental representations (Harnad, 1990). Understanding in an ungrounded representational system is analogous to the content of a dictionary, defining words in an ultimately circular fashion, in terms of other words. It has been argued that this problem can be overcome by grounding internal representations in reflections of the world, mediated by iconic representations associated with direct sensations (Harnad, 1990). Similarly, the meaning of an internal representation can be seen as arising from the role it plays in a cognitive inferential process, as described by the theory of conceptual role semantics, with causal and inferential chains eventually connecting to the world through sensation and action (Greenberg \& Harman, 2006).

In the vector space approach, representations may be grounded in the causal processes in which they participate, both in terms of inputs that give rise to a representation (eventually leading back to iconic representations) and the effects of that representation on downstream neural circuits. These causal relationships can be characterized in terms of functional mappings between vector spaces. For example, if the transduction of sensory information from the world directly results in a pattern of neural firing rates, this pattern corresponds to a point in a sensory vector space, and neural circuits can be seen as mapping this point to corresponding points in the vector spaces for other neural populations, encoding the corresponding conceptual content. In this way, the mappings implemented by neurocomputational mechanisms ground internal representations.

The activation of a PNG and the downstream effects of the initiation of a PNG are mechanistically and computationally quite different than standard neurocomputational mechanisms that can be easily cast as functions between vector spaces. Given this difference in the causal structure of PNGs, how can PNG representations be grounded?

## Addressing the Challenge

We assert that the PNG approach to mental representation requires only a slightly different understanding of the nature of the relevant causal relationships. Rather than being characterized as functional mappings between vector spaces, we posit that the causal relationships between PNGs are best described in terms of trigger sets. Let us first consider the definition of a $\sigma$-triggered polychronous neuronal group.

Definition A $\sigma$-triggered polychronous group refers to the set of neurons that can be activated by a chain reaction whenever trigger neurons $N_{k}(1 \leq k \leq \sigma)$ fire according to the timing pattern $t_{k}(1 \leq k \leq \sigma)$, where $\sigma$ is the size of the stimulus required to trigger the PNG (Martinez \& PaugamMoisy, 2009).

Here, we recognize that a given PNG can have more than one stimulus trigger. This distinction motivates the definition of a trigger set.

Definition For a given PNG $\pi$, its trigger set, $\tau_{\pi}$, is the set of spike-time patterns that trigger the existence of $\pi$. Each spike-time pattern in the trigger set of $\pi$ will give rise to $\pi$ when presented in the absence of interfering spikes.

Note that each element of a trigger set may be a PNG or a portion of a PNG. Thus, the presence of a PNG spike-time pattern may trigger, or help trigger, other PNGs. The set of PNGs that have the potential of being triggered by a given PNG can form the core of a formal characterization of the causally grounded meaning, $\mu_{\pi}$, of that PNG.

Definition If a PNG, $\pi_{0}$, is seen as a set of spikes (with each spike indexed by the identity of the spiking neuron and the time of the spike), and $\mathbb{P}$ is the set of all possible PNGs in the neural network, then the meaning of $\pi_{0}$, called $\mu_{\pi_{0}}$, is defined as the set of PNGs whose trigger set, $\tau_{\pi}$, contains an element with a nonempty intersection with $\pi_{0}$, or

$$
\mu_{\pi_{0}}=\left\{\pi: \pi \in \mathbb{P}, \exists \pi_{t} \in \tau_{\pi}, \pi_{t} \cap \pi_{0} \neq \emptyset\right\} .
$$

By this definition, $\mu_{\pi_{0}}$ includes any PNG for which $\pi_{0}$ contributes some spikes that may contribute to the triggering of the PNG. Thus, $\mu_{\pi_{0}}$ includes PNGs that may only be triggered by $\pi_{0}$ in the context of other spike-time patterns. In this way, the meaning of a PNG is context sensitive. While constrained by the network's topology, the size of $\mu_{\pi_{0}}$ may be very large.

These definitions describe the causal relationships between PNGs, providing the basis for a conceptual role semantics approach to symbol grounding in the PNG framework. PNGs
triggered by sensory neurons can be considered iconic representations, as can PNGs that trigger motor responses. PNGs triggered by other neurons in a network form internal representations that are ultimately grounded in these iconic representations through the causal connections of their trigger sets.

## Compositionality Of Representations

## The Challenge

A representation that exhibits compositionality, simply put, is one where "the meaning of a compound expression is a function of the meanings of its parts" (Janssen, 1996). In the vector space approach, compositional representations have historically been difficult to capture (Fodor \& Pylyshyn, 1988), but progress has been made (Van Gelder, 1990; Gayler \& Levy, 2011). The most common solutions involve either representational components being maintained in subspaces of a parent vector space, or components being superimposed or convolved to form compound representations like tensor product codes, holographic reduced representations, or sparse codes (Smolensky, 1990; Plate, 2003; OReilly et al., 2011).

Do PNG representations suffer from the same problems of compositionality as vector space representations? Are current approaches to compositionality in the vector space framework also appropriate for PNG representations? How might compositionality be captured in PNG representations?

## Addressing the Challenge

The PNG approach offers two ways to capture compositional representations that are impossible, or at least of limited utility, in the vector space approach. These two methods include sequential concatenation of PNG component representations and the superposition of PNG components.

A PNG representation might be seen as being composed of subsequences of spikes, giving rise to representations at multiple time scales. In essence, a PNG may be seen as containing many smaller PNGs within it, or, inversely, it may be a part of a sequence of other PNGs. Consider the meaning of a particular PNG, $\mu_{\pi_{0}}$. If a PNG $\pi_{1} \in \mu_{\pi_{0}}$ has an element of its trigger set, $\tau_{\pi_{1}}$, contained completely within $\pi_{0}$, then $\pi_{0}$ will reliably trigger $\pi_{1}$, in the absense of interfering spikes. Note that $\mu_{\pi_{0}} \neq \mu_{\pi_{1}}$, so these two PNGs do not have the same meaning. These two PNGs may be combined by simple concatenation, producing a new PNG, $\pi_{2}=\pi_{0} \cup \pi_{1}$. Importantly, the meaning of this compound representation, $\mu_{\pi_{2}}$, is a simple function of the meanings of its parts: $\mu_{\pi_{2}}=\mu_{\pi_{0}} \cup \mu_{\pi_{1}}$. (Note that this is the case even if $\pi_{0}$ does not reliably trigger $\pi_{1}$.)

An alternative approach to compositionality involves directly superimposing PNGs over the same time interval. Since spikes may be sparse over time, the probability of superposition producing interference between PNGs is relatively small. In this way, the composition of PNG representations may simply involve the simultaneous activation of the component PNGs.

## Similarity Of Representations

## The Challenge

Distances in a vector space have been fruitfully used to capture dissimilarity between representations, providing a useful mechanism for generalization. Common distance metrics, like inner-product distance (related to angular distance), are tightly related to the kinds of functional mappings between vector spaces that are are easily implemented by neural circuits. How might similarity between representations be captured in the PNG approach?

## Addressing the Challenge

We see this as an important open question. There are many existing metrics for evaluating the similarity of spike trains (Victor \& Purpura, 1996; Naud, Gerhard, Mensi, \& Gerstner, 2011). Developing a good similarity metric has proven difficult, however, as spike sequences are inherently non-Euclidian (Aronov \& Victor, 2004). The metrics that have been presented focus on comparing spike trains generated by a single neuron, recorded over many trials. Current approaches make the assumption that the significance of a spike can be treated as independent from other spikes produced in the same neural network. Because of these and other issues, existing similarity metrics for comparing groups of spike trains do not predict well whether two PNGs will have a similar effect on downstream neurons.

It is our suspicion that similarity metrics based on trigger sets may overcome some of the obstacles described here. Determining a meaningful similarity metric for PNG representations is a focus of our future work.

## Conclusions and Future Work

Ongoing work in computational neuroscience is uncovering the powerful capabilities of polychronous neuronal groups, and empirical studies are starting to find evidence for this kind of encoding in biological neural circuits. If PNGs emerge as a dominant means of representation in the brain, the vector space account of conceptual states will need to be reconsidered. While many of the strengths of the vector space account appear to transfer, with some modifications, to the PNG framework, there remain challenges for viewing complex spike patterns as conceptual representations. Perhaps the most substantial challenge faced by the PNG approach to conceptual representation involves the nature of representational similarity in this framework. This is the primary focus of our ongoing research program.

Much philosophical work has been done in order to clarify how the vector space account of mental representation provides a bridge between brain processes and cognitive processes. If further neuroscientific investigations suggest that polychronous neuronal groups carry conceptual content in at least some brain systems, similar work will need to be pursued for this representational framework. By highlighting several potentially problematic issues with PNG representations, and sketching promising solutions for some of these
issues, we hope to have helped launch this philosophical effort.

## Acknowledgments

The authors thank Jeff Yoshimi, Michael Spivey, and other members of the University of California, Merced (UCM) Cognitive \& Information Sciences (CIS) graduate group for their helpful comments.

## References

Aronov, D., \& Victor, J. (2004). Non-euclidean properties of spike train metric spaces. Physical Review E, 69(6), 061905.

Cessac, B., Paugam-Moisy, H., Viéville, T., et al. (2010). Overview of facts and issues about neural coding by spikes. Journal of physiology, Paris, 104(1), 5.
Churchland, P. (1989). The neurocomputational perspective: Philosophical essays on the mind, brain, and science. Cambridge, USA: MIT Press.
Churchland, P. (1996). The engine of reason, the seat of the soul: A philosophical journey into the brain. Cambridge, MA: MIT Press.
Dan, Y., Poo, M., et al. (2004). Spike timing-dependent plasticity of neural circuits. Neuron, 44(1), 23-30.
Fodor, J., \& Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. Cognition, 28(1), 3-71.
Fries, P., et al. (2005). A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. Trends in cognitive sciences, 9(10), 474-480.
Gärdenfors, P. (2000). Conceptual spaces-the geometry of thought. Cambridge, MA: Bradford Books, MIT Press.
Gayler, R., \& Levy, S. (2011). Compositional connectionism in cognitive science ii: the localist/distributed dimension. Connection Science, 23(2), 85-89.
Greenberg, M., \& Harman, G. (2006). Conceptual role semantics. In Oxford handbook of philosophy of language. Oxford: Oxford University Press.
Harnad, S. (1990). The symbol grounding problem. Physica D: Nonlinear Phenomena, 42(1), 335-346.
Izhikevich, E. (2006). Polychronization: Computation with spikes. Neural computation, 18(2), 245-282.
Izhikevich, E. (2007). Solving the distal reward problem through linkage of stdp and dopamine signaling. Cerebral Cortex, 17(10), 2443-2452.
Janssen, T. (1996). Compositionality. In Handbook of logic and linguistics. North Holland, Amsterdam: University of Amsterdam.
Levy, W., \& Baxter, R. (1996). Energy efficient neural codes. Neural Computation, 8(3), 531-543.
Madhavan, R., Chao, Z., \& Potter, S. (2007). Plasticity of recurring spatiotemporal activity patterns in cortical networks. Physical biology, 4(3), 181.
Martinez, R., \& Paugam-Moisy, H. (2009). Algorithms for structural and dynamical polychronous groups detection. Artificial Neural Networks-ICANN 2009, 75-84.

Naud, R., Gerhard, F., Mensi, S., \& Gerstner, W. (2011). Improved similarity measures for small sets of spike trains. Neural computation, 23(12), 3016-3069.
OReilly, R., Bhattacharyya, R., Howard, M., \& Ketz, N. (2011). Complementary learning systems. Cognitive science.
Pasquale, V., Massobrio, P., Bologna, L., Chiappalone, M., Martinoia, S., et al. (2008). Self-organization and neuronal avalanches in networks of dissociated cortical neurons. Neuroscience, 153(4), 1354-1369.
Plate, T. (2003). Plate, holographic reduced representation: Distributed representation for cognitive structures. CSLI Publications, Stanford, CA.
Rieke, F., Warland, D., Steveninck, R. van, \& Bialek, W. (1999). Spikes: exploring the neural code. Cambridge, MA: MIT press.
Rolston, J., Wagenaar, D., \& Potter, S. (2007). Precisely timed spatiotemporal patterns of neural activity in dissociated cortical cultures. Neuroscience, 148(1), 294.
Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. Artificial intelligence, 46(1), 159-216.
Spivey, M. (2008). The continuity of mind (Vol. 40). USA: Oxford University Press.
Sporns, O., \& Zwi, J. (2004). The small world of the cerebral cortex. Neuroinformatics, 2(2), 145-162.
Swadlow, H. (1992). Monitoring the excitability of neocortical efferent neurons to direct activation by extracellular current pulses. Journal of neurophysiology, 68(2), 605619.

Szatmáry, B., \& Izhikevich, E. (2011). Spike-timing theory of working. BMC Neuroscience, 12(Suppl 1), O3.
Van Gelder, T. (1990). Compositionality: A connectionist variation on a classical theme. Cognitive Science, 14(3), 355-384.
Vertes, P., \& Duke, T. (2010). Effect of network topology on neuronal encoding based on spatiotemporal patterns of spikes. HFSP journal, 4(3-4), 153-163.
Victor, J., \& Purpura, K. (1996). Nature and precision of temporal coding in visual cortex: a metric-space analysis. Journal of Neurophysiology, 76(2), 1310-1326.
Yoshimi, J. (2012). Active internalism and open dynamical systems. Philosophical Psychology, 25(1), 1-24.

# Testing theories of skill learning using a very large sample of online game players 

Tom Stafford (t.stafford@sheffield.ac.uk)<br>Department of Psychology, University of Sheffield Western Bank, Sheffield, S10 2TP, UK

Michael Dewar (michael.dewar@nytimes.com)<br>The New York Times R\&D Lab, 620 Eighth Avenue<br>New York, NY 10018, USA


#### Abstract

We analyse data from a very large $(n=854064)$ sample of players of an online game involving rapid perception, decisionmaking and motor responding. This data set allows us to connect full details of training history with measures of performance, for participants who are engaged for a sustained amount of time in effortful practice. We show that lawful relations exist between practice amount and subsequent performance, and between practice spacing and subsequent performance. This confirms results long established in the literature on skill acquisition. Additionally, we show that higher initial variation in performance is linked to subsequent higher performance, a result we link to the exploration-exploitation tradeoff from the computational framework of reinforcement learning. We discuss the benefits and opportunities of behavioural datasets with very large sample sizes and suggest that this approach could be particularly fecund for studies of skill acquisition.


Keywords: skill acquisition; learning; game.

## Introduction

The investigation of skill learning suffers from a dilemma. One horn of the dilemma is this: experts in real-world skills can be brought into the lab and their performance tested, but it is difficult to reliably recover comprehensive details of their training. This makes it impossible to be certain of exactly how features of the history of their practice are related to the skilled performance you can observe. The other horn of the dilemma is this: you can test different training regimes rigorously, but you are restricted to measuring performance on trivial or unnatural skills, and often without extended training of the order that experts in complex real-world skills engage in.

Computer games offer a partial resolution to this dilemma. Even simple computer games are not trivial in terms of the cognitive abilities which they test. In fact, these abilities are often the staples of cognitive science: perception, decision making and motor responses. Computer game playing is a real-world skill in which many people choose to become expert, devoting hundreds of hours of practice. Unlike most skills, computer games allow a potential record every action in the history of that practice - allowing for the first time detailed investigation of the connection between features of practice and level of final performance. This is what the current investigation sets out to do. We take detailed records of practice activity from an online game and relate amount of practice and features of practice to levels of eventual performance. In doing this we are able to confirm and quan-
tify established findings from experimental studies of learning. In addition we provide a confirmation of a recent result based on the theoretical framework of reinforcement learning (Stafford et al., 2012). Use of online games to collect very large samples offers a new method for the investigation of skill acquisition, we argue, and the work here showcases just some of the possibilities opened up by this approach.

## Practice amount and spacing

We first consider two well established results against which we will validate our data set as a model of skill acquisition: the effects of practice amount and of practice spacing on performance. Studies of learning have shown a lawful relation between practice amount and performance. If performance is gauged in terms of some measure of efficiency (e.g. time taken to make cigars by experienced cigar manufacturers Crossman, 1959), then it is possible to express the relation between practice extent and performance in a power law of learning (Newell \& Rosenbloom, 1981; Ritter \& Schooler, 2001).

For practical reasons studies of the effect of extensive practice have typically looked at different learners possessing differing amounts of practice rather than the same learners at different stages (i.e. cross-sectional rather than longitudinal designs). Experimental studies of learning which do follow learners longitudinally have predominantly focussed on labbased tasks which can be mastered in one or a small number of sessions (although there are, of course, honourable exceptions such as the work looking at the automatisation of visual search performance (e.g. Neisser, Novick, \& Lazar, 1963; Czerwinski, Lightfoot, \& Shiffrin, 1992).

Highlighting the importance of practice quantity in skill development, Ericsson and colleagues stress that the highest levels of performance are never reached without an amount of practice on the order of ten thousand hours (Ericsson, 2006; Ericsson, Krampe, \& Tesch-Rmer, 1993). Additionally, they report that the nature of that practice matters - effortful, directed, 'deliberate' practice is what distinguishes elite performers, even among those who appear to have performed similar quantities of practice.

Experimental studies of learning have focussed on another factor which defines the nature of practice - spacing. The distributed practice effect denotes the finding that if time devoted to practice is separated out rather than massed, or if the spacing is larger rather than smaller, retention improves
(Cepeda, Pashler, Vul, Wixted, \& Rohrer, 2006; Delaney, Verkoeijen, \& Spirgel, 2010). The distributed practice effect is surely one of the most solid findings in learning and memory research. It holds for both motor skill and declarative learning (Adams, 1987). Due to the limitations of experimental methods there is a dearth of evidence on longer spacing intervals (Cepeda et al., 2006), a dearth which we hope the present study offers a method of addressing.

Next we review an area where the approach adopted in this paper affords particular traction for looking at how the history of skill acquisition affects performance.

## Exploration versus exploitation

The computational framework of reinforcement learning (Sutton \& Barto, 1998), outlines a fundamental trade off in decision making: every decision forces us to choose between taking the action which we estimate will yield the best long term consequence (highest 'value'), or trying out an action of unknown or less certain value. This is known as the 'exploration-exploitation dilemma'. Every choice is an opportunity to receive the outcome from only one action, and so also to update our estimate of the value of only one option. Too much exploitation leads an agent to rely on suboptimal actions, seldom discovering better valued actions. Too much exploration, on the other hand, leads to an agent wasting time exploring the space of actions without garnering the reward of frequently choosing the highest known-valued action. The implications for skill learning are that non-maximising performance during early practice may allow superior subsequent performance. Indeed we might even expect that 'expert learners' would adopt an early exploration strategy in order to maximise final performance.

We have already found evidence for this in humans and rats using an experimental task (Stafford et al., 2012). There is other evidence that variability in practice conditions can aid final performance (Roller, Cohen, Kimball, \& Bloomberg, 2001), as well as generating benefits in learning which crosstask (Seidler, 2004) (which has been termed 'structural learning' by some). This is somewhat in tension with accounts which emphasise the need for transfer-specificity in skilled performance (e.g. Logan, 1988). There is not a direct contradiction, merely we are emphasising the benefit of training off the to-be-tested skill.

## Method

Game designers Preloaded produced a game for the Wellcome Trust called 'Axon', which can be played here http://axon.wellcomeapps.com/. They inserted tracking code which recorded a machine identity each time the game was loaded and kept track of the score and date and time of play. The game was played over 3.5 million times in the first few months of release (Batho, 2012).
The game involved guiding a neuron from connection to connection, through rapid mouse clicks on potential targets. A screenshot can be seen in Figure 1 (see figure caption for
description of game dynamics). Cognitively the game involved little strategic planning, testing rapid perceptual decision making and motor responses.


Figure 1: Screenshot of the game Axon. Players control the axonal branching of the white neuron. At each point, possible synaptic contacts (the other dots) are those within the zone of expansion (the larger transparent circle), which shrinks rapidly after each new contact is made. Non-player neurons (in red here) compete for these synaptic opportunities. Score is total branch length in micrometers (shown bottom left).

The analysis was approved by the University of Sheffield, Department of Psychology Ethics Sub- Committee, and carried out in accordance with the University and British Psychological Society (BPS) ethics guidelines. The data was collected incidentally and so did not require any change the behaviour of game players, nor impact on their experience. No information on the players, beyond their game scores, was collected and so the data set was effectively anonymised at the point of collection. For these reasons the institutional review board waived the need for written informed consent from the participants.

Because the data we record is indexed by machine identity, which is derived from the web browser used to access the game, it is not possible to guarantee that a single individual is responsible for all the scores recorded against an single identity. Nor is it possible to guarantee that a single individual is responsible for only one set of scores. These uncertainties add noise to our analysis, but the data set is large enough to accommodate this. It is not clear what, if any systematic distortions these caveats would introduce. For the remainder of this paper we will use the term 'player(s)' to refer to the set of scores associated with a single machine identity.

The data was extracted from Google Analytics using a Python library by Clint Ecker (2009). Data from between 14th of March and 13th of May 2012 was downloaded and compiled into the source data set for the analyses presented here. This data set comprised a total number of 854064 players. Most played only a small num-
ber of times (the modal number of plays is 1 ), but some played up to 1000 times. The data and code for producing the analysis and plots presented here are available from https://github.com/tomstafford/axongame.

## Results

## Practice amount

On average, scores are higher with each consecutive play for up to 100 plays (Figure 2). At around 80 plays the levels of variation between scores, combined with the drop off of number of players reaching that number of attempts, begin to be seen in the loss of the smooth curve and larger error bars.


Figure 2: Average score for each play attempt. Standard errors shown (n.b. some error bars not visible at this scale).

Taking only those who played more than 9 times ( $n=$ 45672), we can calculate a 'high score' for each players (i.e. the highest score they achieved, irrespective of which play it occurred on). The criterion of 9 or more plays for subset selection is arbitrary, an attempt to balance size of subset (which drops with a higher criterion) against likelihood that practice effects will be reliable (which should be greater for higher criterion values). For this, and all other analyses presented in this paper, the results are not contingent on the particular values used to divide up the data (i.e. here we get similar results if greater than $8,10,5$ or 20 plays are used as the criterion. To confirm this we invite interested readers to run the analysis with altered parameters themselves, by visiting the data and analysis code repository referenced above).

From this subset players are then grouped into 5 groups based on the percentile ranking of their high score, and the average score is calculated for each attempt for all players in each percentile group. This shows that the difference between higher and lower scorers is not merely the amount of practice. The difference in average score is present from the very first plays (Figure 3).


Figure 3: Average score against attempt number for different groupings according to maximum score. Standard errors shown.

## Practice spacing

Taking only those who played more than nine times, we divide players into percentile groups according to their highest score, regardless of on which play it was obtained. We also calculate the separation in time between their first and last play. The result shows a clear upward trend (Figure 4, red dots), with players who score most highly spreading their first and last plays further apart. This is unsurprising, however, since even if there was no relation between practice and scoring, and scores were simply random on each attempt, those players who played had more attempts would tend to collect higher scores and have first and last attempts which were more separated in time. We use bootstrapping to estimate confidence intervals as if this were the case. Keeping the number of players and the number and time of the attempts constant, we generate 2000 simulated datasets, sampling with replacement at random from the total record of all scores for all players. The observed data falls below this bootstrap data for low maximum score percentiles and above for high maximum score percentiles, suggesting that the scores really are distributed non-randomly and according to the spread in time of participant's plays (Figure 4).
It is possible to interrogate this result further by a finer slicing of the data. Taking only players who played more than 14 times $(\mathrm{n}=21575)$, we calculate the spread in time between the first play (or second play where this data was missing) and their tenth play (or ninth, where this data was missing). We also identify their best score on plays 11 to 15. We then divide them into two groups, those who played their first ten times within a 24 hour period ("goers"), and those who split their first ten plays over more than 24 hours ("resters"). Resting between first and tenth plays appears to have a benefit on your subsequent performance (Figure


Figure 4: Players graded according to their maximum score percentile against the delay between their first and last plays. Standard errors shown.
5). The difference between the groups is highly significant $(t(20354)=6.219, p<0.00001)$, albeit for a small effect size (Cohen's d $=0.11$ ).

## Exploration versus exploitation

The variance of scores for each player in the first five plays was calculated, and this statistic for each player ranked and so percentile groups created. The same was done for the average on plays six to ten. Plotting one against the other we see a clear correlation - with higher early variance associated with higher subsequent performance (Figure 6, the very high number of individuals made a scatterplot impractical at this scale, so we present a heatmap).

The Pearsons's $r$ correlation coefficient was 0.59 and significantly different from zero at a high probability ( $p<$ 0.0001 ). Randomising the scores for each attempt within the structure of the number of players and the number of attempts per players, it is possible to generate a bootstrap data set which gives a confidence interval for this correlation - in other words, answers the question "to what extent is a correlation between high early variance and high late scoring inherent in the distribution of scores and the structure of how players accumulate scores from that overall distribution". These bootstrapped confidence intervals, at the $95 \%$ level were 0.009 to -0.009 . Thus we can conclude with a high degree of confidence that the correlation is both significantly different from zero and not a trivial consequence of the distribution of scores. Instead, the correlation results from the particular way individual player's early scores are related to their later scores.

## Discussion

These results confirm, but also quantify, results from experimental psychology regarding the effects of practice quantity


Figure 5: Average maximum score following first ten plays, for those who group their first ten plays within one day ('goers') and for those who split their first ten plays over two or more days ('resters'). Standard errors shown.
and quality on performance. As players practice their average score improves. Dividing the players into percentile groups according to high scores appears to show that practice alone does not allow most players to achieve the highest scores. The best players have an advantage from the very first plays. This advantage is consolidated with practice, in that not only do they score more on their first plays, but their rate of improvement is faster. This is in marked contrast to some popular (e.g. Gladwell, 2008) and academic (e.g. Ericsson et al., 1993) accounts of high performance which have denigrated the importance of talent with respect to practice. We regard this result as provisional. It needs to be replicated with another data set so we can assess if it generalises to other skills. Replication would also assuage worries that some specific confound of the present data set has produced the result. For example, we have no way of controlling for the prior game experience or hardware set-up of the players of the Axon game. It is possible that it a certain amount game experience is required for individuals to get high learning rates with this specific game (we thank an anonymous reviewer for pointing out this potential 'thresholding of performance improvement by prior experience' confound).

The analysis of practice spacing confirms the wisdom from experimental studies of learning and memory that distributed practice is better than massed practice. It remains to be seen if there is an optimal amount of spacing, as has been reported for semantic knowledge (Cepeda, Vul, Rohrer, Wixted, \& Pashler, 2008), or an optimal timing of spacing (Goedert \& Miller, 2008).

The exploration-exploitation result confirms a preliminary result from a recent experimental study (Stafford et al., 2012). Although bootstrapping confirms that this finding is


Figure 6: Heatmap made from scatterplot of variance of scores on first five plays versus average score on plays six to ten.
not an incidental result of the distribution of scores, it still isn't clear if the level of exploration (operationalised as score variance on early plays) per se causes the higher level of performance ('exploitation', characterised as score average on later plays). It is doubtful that low scoring attempts in themselves cause higher subsequent performance. Rather low scores may be the impetus for players to shift their playing style or tactics in ways which unlock higher subsequent performance (similar to the postulated freeing and freezing of degrees of freedom which have been thought to characterise changes in motor skill (Berthouze \& Lungarella, 2004; Bernstein, 1967). The ultimate test exactly if and how early exploration affects subsequent performance will be to intervene to make players explore and see how this affects later scores. In other domains there have been suggestions that introducing guided mistakes or deliberate failure into early training may have benefits for overall performance (something for which there is some evidence: Lorenzet, Salas, \& Tannenbaum, 2005).

## Games

Games are a great opportunity for the cognitive science of learning. They provide participants in high numbers who are engaged willing to undertake extensive practice. Games can provide large amounts of detail on training conditions and actions, in ways that other paradigms cannot. In the future it may even be possible to introduce experimental manipulations into engaging games through partnership with games designers.

## 'Big Data'

The method of study adopted here means we lose experimental control over the factors involved in learning. However,
advantages stem from the very large sample size we are able to collect. Some of the emphasis on the importance of experimental control in cognitive science is due to the loss of statistic power than can result from uncontrolled measurement. With large sample sizes, loss of statistical power is not an issue. We need only concern ourselves with the ways in which lack of experimental control introduces systematic confounds into our data set. As well as large statistical power, very large sample sizes mean we can interrogate data in new ways. One of these is 'slicing' by which we mean identifying individuals who meet certain conditions and comparing within that group. This is a substitute for the conventional experimental method of creating individuals that meet certain conditions in low numbers. In experimental design you control potential confounds in advance (by attempting to remove them). With slicing you attempt to account for potential confounds post hoc by selecting multiple different sub-datasets, each of which controls statistically for a potential confound - and thus by a process of elimination gathering support for your hypothesised causal variables. This is a less powerful method than experimental control, but it does offer some advantages.

Bootstrapping provides a way of testing observed patterns against sophisticated null hypotheses. Both bootstrapping and slicing are illustrated in this paper in the analysis of spacing effects.

Two modern crises of psychology are the apparent low replicability of effects (Pashler \& Wagenmakers, 2012) and the use of inappropriate statistics (Wagenmakers, Wetzels, Borsboom, \& Maas, 2011; Simmons, Nelson, \& Simonsohn, 2011). Very large sample sizes can side-step both of these. With a large enough sample size you do not need to use inappropriate statistical techniques - small effects are easy to find. Furthermore, you have enough data to use techniques such as cross validation to guard against false-positives.

Analysed in detail, very large data sets provide an observational playground in which we can not just detect effects, but compare the size of different effects against each other. For example, in the present data set it can be seen that the benefit of 24 hours spacing is about 3000 points (Figure 5). This is comparable to about 5 plays, in the $10-15$ play range (Figure 2), or equivalent to an extra $50 \%$ practice at this stage of experience.

Obviously, nothing will replace the controlled experiment in terms of causal inference. For hypothesis testing the controlled experiment must remain the the gold-standard. However, there is space within the scope of investigation for studies with purposes other than theory-driven hypothesis testing (Rozin, 2009). This paper has focussed on characterising the data and confirming effects discovered in traditional controlled experiments. We believe the approach illustrated here can be complementary to experimental studies, and has the potential to open up new avenues for investigation in the study of skill acquisition.

## Acknowledgements

Thanks to Tony Barnes for an introduction, to Phil Stuart, Charles Batho and Cameron Yule at Preloaded, to the Wellcome Trust for allowing the data from their game to passed to us, to Stuart Wilson for help with python, Ashvin Shah for discussion of reinforcement learning and four anonymous reviewers. Special thanks to Edith Mary Cameron, whose late arrival and post-birth disposition allowed TS to carry out the bulk of the analysis and writing for this paper.

## References

Adams, J. A. (1987). Historical review and appraisal of research on the learning, retention, and transfer of human motor skills. Psychological Bulletin, 101(1), 41-74.
Batho, C. (2012). Axon - a game for science. http://preloaded.com/blog/2012/07/05/axon-gamescience/.
Bernstein, N. A. (1967). The co-ordination and regulation of movements. Oxford: Pergammon.
Berthouze, L., \& Lungarella, M. (2004). Motor skill acquisition under environmental perturbations: On the necessity of alternate freezing and freeing of degrees of freedom. Adaptive Behavior, 12(1), 47-64.
Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., \& Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. Psychological bulletin, 132(3), 354-380.
Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., \& Pashler, H. (2008). Spacing effects in learning a temporal ridgeline of optimal retention. Psychological Science, 19(11), 10951102.

Crossman, E. (1959). A theory of the acquisition of speedskill. Ergonomics, 2(2), 153-166.
Czerwinski, M., Lightfoot, N., \& Shiffrin, R. M. (1992, July). Automatization and training in visual search. The American Journal of Psychology, 105(2), 271-315.
Delaney, P. F., Verkoeijen, P. P. J. L., \& Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. Psychology of learning and motivation, 53, 63-147.
Ecker, C. (2009). Python library for Google Analytics API. https://github.com/clintecker/python-googleanalytics Accessed 17 May 2012.
Ericsson, K. A. (2006). The influence of experience and deliberate practice on the development of superior expert performance. In N. Charness, P. J. Feltovich, \& R. R. Hoffman (Eds.), The cambridge handbook of expertise and expert performance (p. 683-703). Cambridge University Press.
Ericsson, K. A., Krampe, R. T., \& Tesch-Rmer, C. (1993). The role of deliberate practice in the acquisition of expert performance. Psychological Review, 100(3), 363-406.
Gladwell, M. (2008). Outliers: The story of success. Little, Brown.
Goedert, K. M., \& Miller, J. (2008). Spacing practice sessions across days earlier rather than later in training im-
proves performance of a visuomotor skill. Experimental Brain Research, 189(2), 189197.
Logan, G. (1988). Toward an instance theory of automatization. Psychological Review, 95(4), 492-527.
Lorenzet, S. J., Salas, E., \& Tannenbaum, S. I. (2005). Benefiting from mistakes: The impact of guided errors on learning, performance, and self-efficacy. Human Resource Development Quarterly, 16(3), 301-322.
Neisser, U., Novick, R., \& Lazar, R. (1963). Searching for ten targets simultaneously. Perceptual and Motor Skills, 17(3), 955-961.
Newell, A., \& Rosenbloom, P. (1981). Mechanisms of skill acquisition and the law of practice. In J. Anderson (Ed.), Cognitive skills and their acquisition (pp. 1-55). Lawrence Erlbaum.
Pashler, H., \& Wagenmakers, E.-J. (2012, November). Editors introduction to the special section on replicability in psychological science a crisis of confidence? Perspectives on Psychological Science, 7(6), 528-530.
Ritter, F., \& Schooler, L. (2001). The learning curve. In N. Smelser \& P. Baltes (Eds.), International encyclopedia of the social \& behavioral sciences (pp. 8602-8605). New York: Elsevier.
Roller, C., Cohen, H., Kimball, K., \& Bloomberg, J. (2001). Variable practice with lenses improves visuo-motor plasticity. Cognitive brain research, 12(2), 341-352.
Rozin, P. (2009). What kind of empirical research should we publish, fund, and reward?: A different perspective. Perspectives on Psychological Science, 4(4), 435-439.
Seidler, R. (2004). Multiple motor learning experiences enhance motor adaptability. Journal of Cognitive Neuroscience, 16(1), 65-73.
Simmons, J. P., Nelson, L. D., \& Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science, 22(11), 1359-1366.
Stafford, T., Thirkettle, M., Walton, T., Vautrelle, N., Hetherington, L., Port, M., et al. (2012). A novel task for the investigation of action acquisition. PloS one, 7(6), e37749.
Sutton, R. S., \& Barto, A. G. (1998). Reinforcement learning: An introduction. Cambridge University Press.
Wagenmakers, E.-J., Wetzels, R., Borsboom, D., \& Maas, H. L. J. van der. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). Journal of Personality and Social Psychology, 100, 426-432.

# When seeing isn't believing: Influences of prior conceptions and misconceptions 

Eliane Stampfer (stampfer@cs.cmu.edu)<br>Kenneth R. Koedinger (krk@cs.cmu.edu)<br>Human Computer Interaction Institute, 5000 Forbes Ave<br>Pittsburgh, PA 15213 USA


#### Abstract

Instruction often employs visual representations to support deep understanding. However, students' prior misconceptions may override the meaning in these scaffolds. We investigate fraction bars, a common representation intended to promote sense-making. Our prior work found that students often did not use the fraction bars effectively. This difficulty factors assessment compares four scaffold types: pictures only, two forms of pictures with numbers, and numbers only, to assess which interpretation steps were difficult. On equivalence items, students performed equally well with all scaffolds that included pictures, but worse with the numbers-only scaffold, indicating that fraction bars improved scores for equivalence. However, including numbers with the pictures decreased performance for fraction addition. Although students demonstrated competence with fraction bars in fraction equivalence, they did not transfer this knowledge to addition. These results suggest caution in designing and teaching representations for sense-making.


Keywords: graphical representation; fraction addition; symbolic fractions.

Many researchers strive to identify ways to support deep understanding, as it is thought to promote robust and adaptable learning. One strategy has been to use multiple representations, particularly ones that connect to students' prior knowledge and aid sense-making. However, there is little data on what representations will make sense to the students. Singapore textbooks and the NCTM standards, for example, advocate using concrete visual representations in mathematics as a bridge to more formal, abstract thinking (NCTM, 2013; Leinwand \& Ginsburg 2007). But, perhaps we should question the benefits of these representations (Rittle-Johnson \& Koedginer, 2001; Booth \& Koedinger, 2012): Are they actually easy entry points for students?

Our tutors for $5^{\text {th }}$ graders aim to support sensemaking by providing conceptual representations as feedback, a strategy that appears effective with adults (Mathan \& Koedinger, 2005; Nathan, 1998). In our fraction-addition tutor interfaces, equally-divided rectangles, or fraction bars, provide immediate feedback by dynamically
showing the fractions that students enter numerically. We hypothesized that fraction bars would be a more intuitive representation than symbolic fractions, and having students input symbolic fractions and get feedback from fraction bars would prompt thinking on how the two representations were related. Also, we thought it would show students that the common mistake of adding both numerators and denominators was incorrect. We termed this feedback grounded feedback because it was grounded in student's prior knowledge, and grounded an unfamiliar representation (fraction symbols) in a more intuitive one (fraction bars). An initial think aloud study showed promise. The $5^{\text {th }}$ grade participants seemed to understand what the fraction bars meant, and used them to find and correct fractionaddition errors (Stampfer, Long, Aleven, \& Koedinger, 2011). An experimental study found learning benefits with a fraction bar tutor (interface in Figure 1) (Stampfer \& Koedinger 2012). This tutor does not indicate explicitly if an individual step is right, but students cannot advance to the next problem until all steps have been solved correctly.

Although students learned from the tutor (improved from pre-test to post-test), process measures show incorrect interpretations of the fraction bars. Students often indicated they were done solving the problems even though the fraction bars did not line up. They clicked the "done" button on the tutor screen an average of about 2.5 times per problem (rather than the one necessary click). This finding revealed that one of our key assumptions about this form of grounded feedback for these students was not fully satisfied. It appeared that the fraction bar representations of addition were not as meaningful to all students as the think-aloud results suggested. Thus, we were led to investigate more deeply the cognitive


Figure 1: Fraction Addition Tutor. Top row of fractions and fraction bars are given, second row reflects students' inputs, typed in the boxes at the bottom. Text hints appear below when requested.
mechanisms required for processing these representations and, in particular, to attempt to identify the sticking points where student processing deviates from expectation. This difficulty factors assessment (cf., Koedinger, Alibali, \& Nathan, 2008) examines how students understand fraction bars in the context of the fractions they represent; if this process changes depending on the topic (addition vs. equivalence); and how each processing step affects performance.

## Difficulty Factors: Pictures and Numbers

Using a theoretical cognitive task analysis, we identified three likely skills needed to understand the fraction bar representations for fraction addition: 1) equal areas represent equal amounts; 2) the rectangular bars represent the symbolic fractions written above or below them; 3) if two shaded areas are equal, the fractions they represent are equal. We developed matched test items intended to isolate those skills (Figures 2-5). Fraction addition items presented a fully solved problem and students indicated whether it was solved correctly (true or false). Fraction equivalence items presented two fractions and students indicated if the first fraction was bigger than, equivalent to, or smaller than the second fraction. The four question presentations are intended to isolate the skills needed to make sense of the tutor interface in Figure 1. The pictures format (Figure 2) assesses if students know that the shaded rectangles use area to represent quantity, such that two rectangles with equal-sized


Figure 2: Pictures. Does area equal quantity?


Figure 3: Pictures and Numbers. Are images comprehensible as fractions?
Compare, then circle the correct answer

a) $\frac{2}{3}$ is bigger than $\frac{12}{18}$
b) $\frac{2}{3}$ is equivalentto $\frac{12}{18}$
c) $\frac{2}{3}$ is smaller than $\frac{12}{18}$

True or False:

$$
\frac{3}{4}+\frac{1}{7}=\frac{4}{11}
$$

Circle the correct answer: True False

Figure 4: Half Pictures and Numbers. Can students map relationships from images to symbols?

Compare, then circle the correct answer
a) $\frac{1}{3}$ is bigger than $\frac{8}{19}$
b) $\frac{1}{3}$ is equivalent to $\frac{8}{19}$

True or False: $\frac{2}{11}+\frac{1}{2}=\frac{15}{22}$

Circle the correct answer: True False
c) $\frac{1}{3}$ is smaller than $\frac{8}{19}$

Figure 5: Numbers-Only Control. Can students evaluate solved problems?
shaded areas represent equal quantities. Pictures-and-numbers items (Figure 3) include fraction symbols with the fraction bars, to test if students can understand the fraction bars as representations of fractions. Half-pictures-and-numbers items (Figure 4) also include both fraction bars and fraction symbols, but only presents the fraction bars as the hint at the top of the problem. This determines if students can find the relationship between the two fraction bars, map that relationship to the symbolic fractions represented, and then select the relationship that the symbolic fractions have to each other. Numbers-only (Figure 5) provides a baseline for how well students can evaluate the equivalence and addition problems without fraction bars. Another pair of questions gives a baseline for translating a single fraction bar to a fraction symbol (e.g., when shown a rectangle divided in 5 parts with 3 of them shaded, the student should write $3 / 5$ ).

## Methods and Participants

155 fifth grade students from a local public school participated in this study during their normal school day. They were given 20 minutes for a 30item assessment. The school tracked these classes, with 57 students in the highest track, 61 in the middle track and 37 in the lowest track.
Each test included 8 equivalence items and 8 addition items (one correctly solved and one incorrectly solved for each scaffold type). All addends in these items had unlike denominators. The sums in the incorrect addition items followed the popular misconception of adding both numerators and both denominators. Tests also included two single fraction bar items, one with numbers for how many pieces were shaded and how many total. Item presentations were counterbalanced with the specific numbers in the problems to avoid confounding. Item order was determined randomly and half of the tests were given with the order reversed. Questions were scored as 1 if correct and 0 otherwise.

## Results: Scaffold Type Affects Performance

Scores on the single-fraction-bar items were near perfect ( $94 \%$ correct). Figure 6 shows the mean scores for the equivalence and addition items by


Figure 6: Mean scores (max. 1) on equivalence and addition items
scaffold type. Mean scores on the fraction equivalence items were high, with 81-83\% correct for all scaffold types with pictures, and $50 \%$ for the numbers-only presentation. Equivalence items offered three options (bigger, equivalent, or smaller) so even the numbers-only score is well above $1 / 3$ chance. Mean scores on the fraction addition items were lower ( $21 \%$ to $79 \%$ ). These scores steadily decreased as the saliency of the numbers increased. Lower-than-chance results indicate that instead of guessing randomly on the more difficult scaffolds, students answered based on a systematic misconception. Blank answers were scored as 0 and they could reduce performance below the 50\% chance rate. However, students were no more likely to skip the numbers-only addition items than the other addition items that included numbers (numbersonly addition was skipped 13 times, while half-pictures-and-numbers and pictures-and-numbers were skipped 14 times each).
There is a strong interaction effect between question type and scaffold type. We ran an ANOVA on the item scores: 3 (class tracking level: high, middle, low) x 4 (scaffold type: pictures, pictures and numbers, half pictures and numbers, numbers only) x 2 (item: equivalence or addition) with repeated measures for the scaffold type and item. With the Huynh-Feldt correction (since sphericity could not be assumed), results showed significant within-subjects effects for scaffold type and item, and a significant scaffold
by item interaction (all $\mathrm{p}<.0005$ ). Results also showed significant between-subjects effects for class tracking level, with parameter estimates indicating that higher-tracked students got higher scores.

The patterns in figure 6 suggest that all scaffold types with pictures have a similar effect for equivalence, but each scaffold type has a different effect for addition. To verify these hypotheses statistically, we ran separate ANOVAs on each tracking level for equivalence and addition scores, with scaffold type as a fixed factor and student as a random factor. For each of those analyses on the equivalence scores, scaffold was significant ( $\mathrm{p}<.0005$ ) and post-hoc Tukey tests showed that the numbers-only scaffold was significantly different from the other three ( $\mathrm{p}<.0005$ ). For each of those analysis on the addition scores, scaffold was again significant ( $\mathrm{p}<.0005$ ). Tukey tests for the middle track show significant differences among all scaffold types ( $\mathrm{p}<.01$ ). The lowest track did not have significant differences between half-pictures-and-numbers and numbers-only, likely a floor effect. The highest track did not have significant differences between pictures and pictures-and-numbers, likely a ceiling effect.

Figure 6 also suggests that addition with the pictures-only scaffold is no more difficult than equivalence with the pictures-only scaffold. To test this, we ran an ANOVA on the item scores for the pictures-only scaffold: 3 (class tracking level: high, middle, low) x 2 (item: equivalence or addition) with repeated measures for item. Results showed no significant difference for scores on the two question types ( $p=.2$ with the Huynh-Feldt correction). Subsequent ANOVAs on each of the other scaffold types showed significant differences for scores on the two question types (all p<. 0005 with the Huynh-Feldt correction).

Finally, we examined the effect of spatial reasoning on scores. One may hypothesize that when pictures are present, students would be more accurate when there is a large disparity in the area of the quantities being compared. To test this hypothesis, we calculated a disparity measure for each question where the two fractions were not equivalent or the two addends did not equal the sum. For the equivalence items, the disparity is
the absolute value of the first fraction minus the second fraction. For the addition items, the disparity is the true sum of the addends minus the sum in the question. We ran separate ANOVAs for each question type, with scaffold type and disparity as fixed factors and student ID as a random factor. For both addition and equivalence, between-subject main effects were significant for scaffold type and student ID ( $\mathrm{p}<.0005$ ) but not for disparity ( $p=.141$ for addition, $p=.888$ for equivalence), and there was no scaffold*disparity interaction ( $\mathrm{p}=.257$ for addition, $\mathrm{p}=.136$ for equivalence). This indicates that disparity did not affect scores, and the effect of disparity did not change with scaffold type. Additionally, the equivalence questions all had smaller disparities than the addition questions (means: . 06 for equivalence, .39 for addition), yet the equivalence questions were as easy or easier, further evidence that disparity did not affect scores.

## Discussion: Fraction Bar Skills are Context-Based

Students were at ceiling for writing the symbolic fraction represented by a single fraction bar. Students were quite good at comparing two fractions and indicating if the first was greater than, equivalent to, or smaller than the second. Further, scores on these equivalence items were equally high for all scaffold types that included pictures.

On the equivalence items, students demonstrate competence with the three skills identified in the cognitive task analysis: equal areas represent equal quantities (pictures), the bars represent fractions (pictures and numbers), the relationship between the bars maps to the relationship between the fractions they represent (half pictures and numbers). Students were likely not solving these equivalence problems with the numbers alone, since numbers-only performance is much lower.

Surprisingly, these skills are not consistently demonstrated with fraction addition. Pictures-only scores are just as high with addition as they are with equivalence, indicating that the knowledge that equal areas represent equal quantities does transfer to addition. However, performance decreases steadily across pictures-and-numbers and half-pictures-and-numbers, suggesting
difficulty both with understanding the bars in the context of fractions and mapping the relationship between the fraction bars to the relationship between the fraction symbols. Yet, the bars still increase performance above the numbers-only control (which has worse-than-chance scores).

We hypothesize that the temptation of the incorrect add-both-numerators-and-denominators strategy overrides the area-as-quantity reasoning that students demonstrate when the numbers are not shown. A cognitive-load hypothesis may predict that fraction symbols are distracting because they visually clutter the problem. In that case, scores with half pictures and numbers should be higher than pictures and numbers, since there is less information and less visual clutter. Yet, scores decrease, indicating that performance is not correlated with cognitive load.

Byrnes and Wasik (1991) discuss a theory that conceptual knowledge will prevent students from making certain procedural errors. In this theory, a "self-critic" (our name), evaluates procedural outcomes for conceptual errors. For example, if a student adds $3 / 4$ and $1 / 7$ and gets $4 / 11$, their "self-critic" may reason that $4 / 11$ cannot be right because it is less than half while $3 / 4$ is greater. With the picture scaffolds, these steps are easier instead of numeric mental operations, students can compare the fraction bars. Scores on the equivalence and the pictures-only addition items demonstrate students' skill in comparing fraction bars, yet they still seem to not use their "critic" on the fraction addition items with numbers.

Interestingly, Byrnes and Wasik argue against the self-critic theory, claiming that conceptual and procedural knowledge are not activated simultaneously in problem solving. Further, conceptual knowledge may precede procedural skill, so in some stages of learning conceptual knowledge would not be correlated with procedural performance. Instead, procedural skills improve through proper discrimination and generalization. To test these theories, they compared three instructional techniques for LCD fraction addition. One was procedural, and stressed that "you can't add fractions the way you add ordinary numbers." The other techniques added conceptually-based instruction (one with
paper fraction bars) to that procedural instruction. Results showed that the conceptual methods did not improve learning above the purely procedural one. These findings suggest that aiding discrimination will improve procedural skill, and that skill is not enhanced further with brief conceptual instruction. These findings and the results from the fraction equivalence items suggest that students will not benefit from more conceptual instruction on fraction bars, even though they performed poorly on fraction addition items with fraction bars. Instead, they may benefit from support for separating whole-number and fraction addition. Alternatively, students may benefit from practice and support in invoking their "self-critic." However, these critics may be stifled by a misconception unrelated to fractions: the meaning of the equals sign.

McNeil et al. (2006) found that $6^{\text {th }}-8^{\text {th }}$ grade students looking at a problem such as $3+4=7$ were more likely to interpret the equals sign to mean 'write answer here' than 'both sides are equivalent.' Perhaps this misinterpretation of the equals sign in equations with operations interfered with students' internal "critic" in the addition items. Even when the pictures show the sum to be smaller than one of the addends, the student may not realize that the two sides of the equal sign are supposed to be equivalent. A "critic" that interprets " $=$ " as 'write output of procedure here' may simply verify that the add-both-numerators-and-denominators strategy was executed well. In other words, the presence of numbers may not only prompt over-generalization of whole-number addition, but also interfere with students' interpretation of the equals sign and thus throw off the "critic."

## Conclusion

These data imply that the usefulness of the fraction bar scaffold is dependent on the topic for which it is employed, and the specific combination of images and numbers. When naming fractions represented by individual fraction bars and solving equivalence problems with fraction bars, students were equally proficient whether the numeric symbols were present or not. However, for fraction addition, the
fraction symbols were detrimental. The picturesonly addition problems may invite reasoning based on conceptual understanding (the sum of two areas cannot be smaller than either addend), while the presence of fraction symbols may invite procedural problem solving that is initially divorced from the underlying concepts.

This DFA study suggests that students' difficulty with dynamic fraction bars in a tutoring system was due to the specific addition context. More broadly, it suggests caution in the design and use of conceptual scaffolds for math problems. Students may demonstrate proficiency with a scaffold in one domain without being able to transfer those skills, even to a closely related domain. Procedural misconceptions may override the conceptual reasoning these scaffolds attempt to induce. Perhaps students need instruction to support their "self-critics" in checking procedural outcomes against conceptual knowledge. Or, perhaps students require certain domain-specific knowledge before their "self-critics" are triggered. Our future work with fraction bars will explore the effect of metacognitive "self-critic" training and domain-specific instruction on the meaning of the equals sign.

## Acknowledgments

We thank our participants, their teachers, Jo Bodner, and Stephanie Butler. This work was supported in part by the Pittsburgh Science of Learning Center through NSF award SBE0836012, and by a Graduate Training Grant awarded to Carnegie Mellon University by the Department of Education (\# R305B090023).

## References

Booth, J. L., \& Koedinger, K. R. (2012). Are diagrams always helpful tools? Developmental and individual differences in the effect of presentation format on student problem solving. British Journal of Educational Psychology 82(3) 492-511.
Byrnes, J.P., \& Wasik, B.A. (1991). Role of conceptual knowledge in mathematical procedural learning. Developmental Psychology 27(5) 777-786.

Koedinger, K.R., Alibali, M.W., \& Nathan, M.J. (2008). Trade-offs between grounded and abstract representations: Evidence from algebra problem solving. Cognitive Science 32(2) 36697.

Leinwand, S., \& Ginsburg, A.L. (2007). Learning from Singapore Math. Making Math Count 65 (3) 32-36.

Mathan, S., \& Koedinger, K. R. (2005). Fostering the Intelligent Novice: Learning From Errors With Metacognitive Tutoring. Educational Psychologist, 40(4), 257-265.
McNeil, N.M., Grandau, L., Knuth, E.J., Alibali, M.W., Stephens, A.C., Hattikudur, S., \& Krill, D.E. (2006). Middle-School Students’ Understanding of the Equal Sign: The Books They Read Can't Help. Cognition and Instruction 24(3) 367-385.
Nathan, M. J. (1998). Knowledge and Situational Feedback in a Learning Environment for Algebra Story Problem Solving. Interactive Learning Environments 5, 135-159.
National Council of Teachers of Mathematics (2001). Navigating Through Algebra in Grades 6-8, as cited in http://illuminations.nctm.org/ LessonDetail.aspx?ID=L605
Rittle-Johnson, B., \& Koedinger, K. (2001). Using cognitive models to guide instructional design: The case of fraction division. In J. Moore \& K. Stenning (Eds.), Proceedings of the 23rd Annual Conference of the Cognitive Science Society. (857-862). Mahwah, NJ, Erlbaum.
Stampfer, E., \& Koedinger, K.R. (2012) Tradeoffs between immediate and future learning. Presented at the European Association for Research on Learning and Instruction Special Interest Groups 6\&7 Conference. Bari, Italy.
Stampfer, E., Long, Y., Aleven, V., \& Koedinger, K.R. (2011) Eliciting intelligent novice behaviors with grounded feedback in a fraction addition tutor. In G. Biswas, S. Bull, J. Kay, \& A. Mitrovic, (Eds.), Proceedings of the 15th International Conference of Artificial Intelligence in Education (560-562) Auckland, New Zealand: International AIED Society.

# The Importance of Nouns in Text Processing 

Lauren M. Stuart, Julia M. Taylor, and Victor Raskin (\{lstuart, jtaylor1, vraskin\}@ purdue.edu)<br>Purdue University, 656 Oval Drive<br>West Lafayette, IN 47907-2086 USA


#### Abstract

In the area of computational processing of natural language texts, advances toward simpler yet more accurate models of meaning are desirable. As syntax is a major component of semantic analysis, we explore how a long-term institutional bias towards the verb as the main determiner of syntactic (and semantic) structure may underserve some kinds of information. We introduce an analysis paradigm that restores the noun to some importance in syntactic analysis. A noundriven syntax representation has been developed and evaluated, and implications of its use in further processing and in better modeling of natural language meaning are investigated.


Keywords: Linguistics, Language Understanding, HumanComputer Interaction, Representation, Syntax, Semantics

## Introduction

Text interpretation by computers is highly desirable and arguably necessary as we continue to produce and analyze text. One major benefit to the improvement of natural language understanding (NLU) for text is more intuitive natural interaction with highly structured or, conversely, loosely associated, large stores of information.

Text processing may proceed sequentially, on the assumption that only full (or major) analysis of the surfaceward structure yields the next deepest structure, where deepest structure is some formulation of the text's meaning, possibly applicable to other meanings of other texts, and the surface structure is the written or spoken input for a computer. This linear process encourages the development of incremental processing modules; that is, given some intermediate representation of something going on in the text, the module will produce a further refined model according to its internal rules and heuristics. For an example, take a phonetic processor that processes speech data and outputs a series of symbols for use in phonological and morphological analysis.

The process does not have to be linear; an alternative approach may parallelize the different analyses to some degree, even to the maximum possible (for instance, we cannot process any syntactic data if it has not yet been furnished). In this approach, iterations of processing may "clear up" a map of the sentence's interpretation (meaning) incrementally. Easy or simple rule applications start the process and such selections provide feedback for further selections in those areas of the map that are not yet clear, for some threshold of "clear" that is dependent upon the form and eventual use of the data.

Linear or not, any processing of a text from its surface form to some model of its meaning relies on various stages
of language processing. We wish to explore how a bias in one of these stages, and its correction, affects processing in another.

## Sentence Processing

As a sentence is analyzed, much importance is given to the verb(s): modal verbs modify the main verb; noun phrases participate as subjects or objects; any noun phrase not directly related to the verb may be a complement of a preposition, which is itself associated with the verb or a noun phrase, or of some other clause or phrase whose meaning props up the meaning of the verb (the event said to be encoded in this particular sentence) and whose place or expression in the syntactic structure of the sentence is as much (or more) dictated by the verb as the head of the phrase. Nouns generally remain building blocks of arguments to be fed into a verb.


Figure 1: Representation of a phrase structure parse from Berkeley Parser


Figure 2: Representation of a dependency parse from Stanford Typed Dependencies

In syntactic representations, this privileging of the verb is typically expressed as a shorter distance between the root of the tree and the main verb; some representations go so far as to shorten the paths to the other verbs in the sentence. See Figure 1 for a constituency tree representation, and note the comparative height of the verb and its nouns. Then see

Figure 2 for a dependency tree, noting how the removal of some phrase structure also removes some steps to access any particular word, but verbs are still comparatively close to the root and can be accessed from one another. This reveals an intuition that a sentence is primarily "about" its verb and that composition of multiple sentences is "about" lining up the verbs together.

In the process of computational understanding of natural language, a computer may be given a syntax tree from which to construct a meaningful model of the events and objects that the sentence describes (with "meaningful", here, determined by the eventual use to which the model will be put: are we looking for frequencies of events? a reconstruction of the actions of one particular object? similarities in locations, origins, or attributes of events or objects?). Then, to maintain the appropriate interpretation of meaningfulness, the distance of a particular word or word category from the root of the sentence must correlate to its need, or incorporation, in the construction of this model. Given syntax trees that are verb-centered, it is most efficient to construct verb-centered semantic representations. Processing may not need to be sequential (as in, phonological then morphological then syntactic then semantic), but we will leave that possibility alone for right now.

A mapping may be observed between simple sentences and logical expressions in first-order predicate logic: The images show a landscape can be formulated as a function show() with the arguments images and landscape. The proposition show(images, landscape) is held to be equivalent to the sentence - that is, if it is tested for a binary truth value, it evaluates to "true" in all the situations in which the sentence from which it comes would be considered true. More complex statements can be generated using such rules as well, e.g., show(images, landscape) \& is_on(landscape, Mars). By mapping natural language sentences to this restricted logical form, we arrive at a sort of semantic notation that is easier, somehow, for a computer to use. Its close resemblance to the syntax of many programming languages suggests that, if only we can translate all sentences into such expressions, we can execute the program obtained by concatenation (in accordance with rules for coordination, negation, etc.) of these expressions and thereby arrive at some truth value for the sentences taken together. We may not just want an answer (true or false) but a model of the meaning in the text; tweaking the execution of these formulae may allow us to build that model.

However, first-order predicate logic is not entirely adequate. Luuk (2009) extends the mapping to a less strict system and theorizes about the possible evolution of argument-like concepts (nouns) before predicate-like concepts (verbs, among others).

Still, all of this analysis is predicated on the idea that the verb (or the event it describes) is the central element of analysis, from which all other considerations flow. However, there are some natural and regular instances in
which the verb is no help, or possibly even absent. Take a copular sentence: The Curiosity is the Mars rover. While the existential senses of is are large and have many implications, they are only manageable when knocked down to the scope of is the Mars rover. In practice, the verb in this sentence is demoted to purely technical predicate status and the predicative nominal elevated in its place. Now we are to analyze a noun phrase as a predicate; there is plenty of precedent, as we can talk about noun-expressed events taking arguments structured similarly to those their verbexpressed counterparts accept. Compare We celebrated the launch today with The celebration of the launch was today.

However, the question must be asked: why is our analysis so verb-centered, and to such a degree that we must postulate verbs, i.e., essentially to create dummy verbs, where there are none?

## Towards a Noun-Driven Paradigm

We propose an alternative, perhaps complementary analysis paradigm: center the noun. Such a paradigm might include analysis of concepts as informational objects - for events to be frames - and events as actions somehow intrinsic or controlled by the objects they involve. This paradigm may open up a world of gains in processing different flavors of information sources, particularly those that have been traditionally managed by computers, with different degrees of naturalness in the language used to interact with them. A noun-driven paradigm may then boost ease of interaction with these sources via natural language understanding by simply not introducing an unneeded event structure for analysis.

To this effect, we have proposed a noun-driven syntax representation (Stuart, 2012). It inherits from the class of dependency grammars by formulating syntax rules as directed binary relationships between nodes. For instance, a preposition may be eliminated entirely and encoded in the syntactic tree as a directed relation, carrying the meaning of the preposition, between the elements it used to connect. For an example, see Figure 3.


Figure 3: Prepositional collapse in Stanford Typed Dependencies Output

There may be tradeoffs between dependency grammars and constituency grammars - Nivre (2006) considered the tradeoffs favorable - but some may be more important in
the noun-driven paradigm (for instance: should it perhaps be the noun-phrase-driven paradigm?) and only further investigation will reveal these.

A noun-driven representation has been developed, starting in Stuart, et al. (2012a) and Stuart, et al. (2012b). The structure links all nouns from the root, so a parallelized meaning-scaffolding program may have several starting points and begin to converge upon an intermediate structure (towards a model of meaning) as it traverses the nodes held in common between noun-rooted subsets. The corresponding parallelization applied to verb-driven dependency grammar representations does not result in the same gains: there are typically two noun phrases to every verb phrase in English (Baker, 2005). The number of nouns relative to verbs only gets larger when we consider noun chains ("crater rim"-see also Taylor et al 2010, 2012) and prepositional phrases (the objects of which are always noun phrases).

The number of prepositional phrases also contributes to the complexity of syntactic analysis. Some prepositional phrases have ambiguity in attachment; some may attach somewhere in syntactic analysis but be restricted from that attachment during more meaning-directed evaluation. Consider The images show a landscape on Mars vs. The images show a landscape of Mars. In the first, on may attach to both show (the event of showing could occur on Mars) and to landscape (the landscape is specifically one on Mars). In the second, of may only attach to landscape, but can sometimes attach to other verbs (as a particle, for instance, in to think of), so the evaluation of the attachment as valid may take several steps of analysis. Multiple prepositions can also attach to the same elements; as they all will have noun-phrase objects, the centering of the noun phrase in analysis hikes the prepositional phrase up in the hierarchy of importance as well.

## Implications for Semantic Processing

The "object-oriented" nature of noun-driven syntax may also align sufficiently well with object-oriented semantic languages to collaborate easily in parallelized processing, allowing groups of objects at certain "stages" to be swapped out with "higher-stage" interpretations or representations of them.

Take the sentence introduced before: "The images show a landscape" and its partial processing, as shown in Figures 2 and 3. In the latter, "show" has been tagged as a noun rather than a verb. Possibly these are candidates with the highest confidence due to internal simplicity, some rules about sentence formation, the topic of the information, or from previously-mentioned information: "the images" likely refers to some set of images that have already been introduced. In a pass that has produced some semantic or intermediate representations for parts of the map-the intermediate conclusions made in the other parts of the sentence contribute to analysis of the other parts. Analogously, in a greedy meaning algorithm, the subsets of
the sentence which are simplest to compute or represent drive the interpretation of the rest of the sentence.

As in many syntactic analysis algorithms, "steps" of processing can be reverted (or previous states saved) to enable backtracking or the output of multiple best candidates if appropriate. The processing shown in Figures 4 and 5 presupposes object-oriented semantic-processing modeling. For those systems with event-oriented semantic processing, verb-driven syntactic approaches (linear or parallel) are just as useful.



Figure 4: Processing steps when "show" tagged as a verb


Figure 5: Processing steps when "show" tagged as a noun
Regardless of whether the semantic modeling used in process is object-oriented, event-oriented, or some hybrid, we find it prudent to also consider the possible object- or event-biased information sources. Suppose we have an encyclopedia article about the geographical features of Mars. While many sentences will perhaps explain the dates and methods of discovery of certain facts about Mars and its geography, we can expect a large amount of discussion on the features themselves, their attributes, positions, and the larger classes to which they belong. In the example sentence A peak sits to the south of the crater rim, we receive information that there is some peak (perhaps in a larger range of mountains) somewhere south of a crater that has been (we hope) previously introduced. This is position information; the verb is sit but if it were for some reason left out of the sentence or replaced with is, is located, or is situated, we would still get the gist.

Consider, then, a general article about geography, the content of which an NLU system will have to make use of in order to understand the instantiations (in the Mars geography article) of types and classes described in the general article. We can expect some events here: various geographers may have contributed to the development of some concepts. However, we argue that the most useful part of the article (again, for understanding something about the features of Mars) is that part which defines the types, attributes, and relative locations of geographical features. This area of the article is likely to be strongly nouncentered: positions, lists of items, and weak verbs may
dominate. (A volcanic crater is a circular depression in the ground.) This is not to say that the verb-driven paradigm is useless in a very noun-dominated information space. However, a verb-driven syntactic analysis may waste some resources by trying to identify and promote predicates. As well, a verb-driven semantic analysis may find that most of its "events" are existential and not temporally-bounded instances of some action.

Now consider some more rigidly-defined information spaces. The most rigid might be the company relational database, which strictly encodes lists of relationships between objects - Li et al. (2008) developed an engineering ontology and lexicon for processing natural language search queries, in a supply chain application, to just such a database, which also included information in unstructured (that is, natural language) descriptive documents. The user queries (and therefore the ontology) emphasize parts' shapes, materials, purposes, and origins. Such domains are fundamentally concerned with objects and how they are related to each other. The "translation" from machinereadable language to natural language must rely heavily on nouns, adjectives, quantities, and the relationships and attributes of semantic concepts representing "things" rather than "happenings". The basic semantic structure of the information is already there: the database schema gives us relations by which to connect the objects (event instances, object-parts, etc.) that are its subject.

Take now, for example, a computer program written in some programming language. Improvement of NLU also underwrites the improvement of translation between natural language and programming languages (for instance, in specifying and evaluating privacy constraints, as in Brodie et al. (2006)). The language itself does not need to be conceived of as object-oriented (Java, C++) because a reductive view of a computer program is that of an informational object which takes informational objects as inputs and gives more information objects as outputs. The transformations that these input objects undergo are events, the transformations may be affected partially or wholly by some qualities of the objects to which they are applied. This is a strength of inheritance and polymorphism in those object-oriented languages: one prototypical "event" can be expressed in terms of the classes of objects to which it is applied, and many events are specific, intrinsic, and unique to a (class of) informational object (data structure). Thus, "translation" of the event relies ultimately upon the objects involved; that is, Object.do() is specified at least in part by the implementation of the class Object. Even in a language that is not specifically object-oriented, the specific actions undertaken in the execution of the method process(thing) depend on the content or nature of the object thing.

Finally, consider the sentence "Airborne geomagnetic surveys showed a strange pattern of symmetrical magnetic reversals on opposite sides." Our main verb here is "show", but the important events (the act of surveying, and the occurrence of magnetic reversal) appear in noun form, and some important attributes appear as adjectives and a
prepositional phrase. Verb-centered processing (purely syntactic or as a step towards semantic processing) prioritizes "show"; it stands in for a fuller explanation that investigating scientists learned about the reversals by reviewing data from the surveys, and thus does have some importance (for instance in auditing the assertion "Many magnetic reversals have occurred", as one question could be "How do we know that?"). In an article about the scientific process, its many forms, and its contribution to knowledge, this is a salient detail. However, in reading an article about the geomagnetic history of the earth, we may be much more interested in the apparent occurrence of magnetic reversals, and the source of the data analyzed in order to reach that conclusion.

## Experimental Evaluation

Stuart et al. (2012b) began evaluating the performance of the noun-driven syntax in a small query context: assuming that most queries to syntax trees take the form of traversing the tree to or from a certain node or a node of a certain category, the node's depth is used as a rough measure of accessibility for further analysis. The initial experiment was carried out by hand and in comparison with outputs from Stanford Dependencies, Stanford PCFG, and Berkeley parsers. The dataset consisted of only 30 sentences; the noun-driven syntax representation performed at least as well, or better, than the dependency grammar trees, and both much better than the phrase-structure trees. As well, the dependency and noun-driven trees were evaluated from a parallelized perspective.

A larger experiment used 600 sentences, chosen from six different articles, each of which fell under one of three categories. The categories - "noun-heavy", "verb-heavy", and "neither" - attempted to capture a meaning-motivated difference in syntax between information sources, as well as to test intra-category variance to a degree. The performance of the noun-driven syntax was compared to that of a developed hybrid phrase structure representation - the latter considerably "flattened" phrase structures. Direct comparison with a dependency parser's output could not be obtained for this larger experiment due to technical difficulties, but it is planned for the future.

The noun-driven representation was generated by a phrase-structure parser integrated with Ontological Semantics Technology (OST), a natural language understanding framework under current development (Taylor et al., 2012). The parser used a partial lexicon - a set of word entries with associated syntactic information, intended eventually to include semantic, morphological, and phonological information useful for word sense disambiguation and construction of semantic meaning representations. The parser used a modified chart-parsing algorithm, similar to that presented in Allen (1987) but organized around heads of phrases rather than building phrase structure from left to right. The parser generated parses in phrase-structure representations then converted those to noun-driven trees. An example representation of the
output appears below in Figure 6; compare with Figures 1 and 2.


Figure 6: Representation of a noun-driven dependency parse
A testing program counted the depths for each word in the word categories of interest (determiner, noun, verb, adverb, preposition, adjective). One metric for appropriateness of the categories is in measuring the ratio of nouns to verbs; this data is shown in Table 1. A "fingerprint" for each of the subcategories (counts by word class) appears in Figure 4; note some similarity within each of the categories. However, these findings are complicated by unknown effects of style (the two "noun-heavy" articles were from Wikipedia, the two "neither" articles from the New York Times website, and the last two from Safety.gov and a recipe book, respectively), though there may be an association between style and subject matter that does uphold the categorization. As well, these results do not distinguish between parses with differing levels of correctness or acceptability; work in progress does mark parses on a spectrum from non-grammaticality to candidacy as the parse most compatible (with some semantic processing) with the rest of the article.


Figure 7: Word class counts per parse, by subcategory

| Table 1: Noun/Verb Ratios |  |
| :--- | :---: |
| Sentence subcategory(article subject) | N/V |
| Neutral 1 (Mars Rover) | 3.82 |
| Neutral 2 (2012 Election) | 3.65 |
| Noun-heavy 1 (Plate Tectonics) | 4.82 |
| Noun-heavy 2 (Mathematics professions) | 5.23 |
| Verb-heavy 1 (Safety Tips) | 3.59 |
| Verb-heavy 2 (Recipes) | 4.56 |

A sample of 64 sentences, randomly selected from the 600 , was tested for parser correctness and accuracy - the
distinction arises from a difference between transforming a string of grammar symbols into all possible syntactically correct parses of the symbols, regardless of their content, and obtaining the correct parse, as the syntax parser has no ability to determine which of the correct strings of symbols is actually completely grammatically correct. This is a result of an inexpressive tag set and a lack of semantic parsing integration. For the 64 -set, in 6 cases were the correct parses not included in the output; this was due to a lexicon gap, verb particles not correctly accounted for, or an oversight in vetting the data set for conformance to the limited grammar template that the parser was designed.
For the 600 sentences, some measures were taken of possible syntactic ambiguity: if for one sentence the parser turned out more than one parse and could be counted on (according to an interpretation of the outcome of the 64 -set results, which is not entirely dependable) to be correct, if enthusiastic, in its determination of good syntactic parses, then the sentence was determined to carry syntactic ambiguity. Of the 600,241 sentences were given only 1 parse; an equal number had 2.16 sentences had 10 or more different parses turned out; some of these were due to prepositional phrase attachment ambiguity, and some to possibilities that, for instance, the form of a verb in the $3^{\text {rd }}$ person-present-singular is identical to that of a plural noun, or vice versa.

Table 2: Depth Counts for Nouns

|  | Noun-Driven |  |  | Phrase-Structure |  |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Group | Average | Min | Max | Average | Min | Max |  |
| All | 1 | 1 | 1 | 3.38 | 2 | 9 |  |
| Neutral 1 | 1 | 1 | 1 | 3.29 | 2 | 9 |  |
| Neutral 2 | 1 | 1 | 1 | 3.34 | 2 | 8 |  |
| Noun-heavy 1 | 1 | 1 | 1 | 3.48 | 2 | 9 |  |
| Noun-heavy 2 | 1 | 1 | 1 | 3.66 | 2 | 9 |  |
| Verb-heavy 1 | 1 | 1 | 1 | 3.27 | 2 | 8 |  |
| Verb-heavy 2 | 1 | 1 | 1 | 2.98 | 2 | 7 |  |

Table 3: Depth Counts for Verbs

|  | Noun-Driven |  | Phrase-Structure |  |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: |
| Group | Average | Min | Max | Average | Min | Max |
| All | 2 | 2 | 2 | 2 | 2 | 2 |
| Neutral 1 | 2 | 2 | 2 | 2 | 2 | 2 |
| Neutral 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| Noun-heavy 1 | 2 | 2 | 2 | 2 | 2 | 2 |
| Noun-heavy 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| Verb-heavy 1 | 2 | 2 | 2 | 2 | 2 | 2 |
| Verb-heavy 2 | 2 | 2 | 2 | 2 | 2 | 2 |

Table 4: Depth Counts for Prepositions

|  | Noun-Driven |  | Phrase-Structure |  |  |  |
| :--- | :---: | ---: | :---: | :---: | :---: | :---: |
| Group | Average | Min | Max | Average | Min | Max |
| All | 2.38 | 2 | 3 | 3.60 | 2 | 8 |
| Neutral 1 | 2.40 | 2 | 3 | 3.58 | 2 | 8 |
| Neutral 2 | 2.47 | 2 | 3 | 3.68 | 2 | 7 |
| Noun-heavy 1 | 2.34 | 2 | 3 | 3.69 | 2 | 8 |
| Noun-heavy 2 | 2.33 | 2 | 3 | 3.59 | 2 | 8 |
| Verb-heavy 1 | 2.56 | 2 | 3 | 3.64 | 2 | 7 |
| Verb-heavy 2 | 2.32 | 2 | 3 | 3.26 | 2 | 6 |

Tables 2 and 3 show results for depth counts between the two syntax representations evaluated, over the 600 -sentence set, for the two main categories, noun and verb. We include the depth counts for prepositions as well (Table 4) because of the complexity that prepositional phrases add to syntactic structure.

Examination of these tables reveals that, even when compared with a "flatter" (verb-driven) phrase structure syntax representation, the noun-driven representation does at least as well, if not better. Notice as well that in the phrase-structure trees, nouns have a wider range of "float" because, while a single noun may be the subject and thus be found in a shallower position, prepositional phrases and noun-chaining bury nouns further.

Preliminary results from further evaluation also reveal data that may be usable for characteristic profiles of prepositional attachment. As well, we may investigate whether the addition of some phrase structure features to the dependency-like representation would provide better information for prepositional attachment and other local operations influenced by concepts or structure use.

## Conclusion

If computational processing of information is (even sometimes) object-centered, then an object-centered approach aligns with it. Given that we have started at the syntax level, and that most objects (as well as some events) are typically expressed as nouns, the noun-driven syntax representation, and an eventual development of a parsing approach, begins the building of a noun/object-centered paradigm for the analysis of natural language text.

There are information spaces that are not so noun-biased, or even further verb-biased - with the full field of verbdriven syntactic and semantic analysis, these will not be left behind. A dual analysis, using both perspectives, may produce some gains in efficiency and effectiveness.

## Acknowledgments

This research has been supported in part by the National Science Foundation CISE/TC Research Grant \#1012208.

## References

Allen, James. 1987. Natural language understanding (2 $2^{\text {nd }}$ edn).Addison-Wesley.
Baker, Mark. 2005. "Lexical Categories: Verbs, Nouns, and Adjectives." Cambridge: Cambridge University Press.
Brodie, C. A., Karat, C., \& Karat, J. (2006). An empirical study of natural language parsing of privacy policy rules using the SPARCLE policy workbench. SOUPS '06 Proceedings of the second symposium on Usable privacy and security, 1.
Klein, Dan and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. Proceedings of the 41st Meeting of the Association for Computational Linguistics 423-430.
Luuk, Erkki. 2009. The noun/verb and predicate/argument structures. Lingua 119 (2009): 1707-1727.
de Marneffe, Marie-Catherine and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation 1-8. Manchester, UK.
de Marneffe, Marie-Catherine, Bill MacCartney, and Christopher D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. 5th International Conference on Language Resources and Evaluation (LREC 2006) 449-454.
Li, Zhanjun, Victor Raskin, and Karthik Ramani. 2008. Developing Engineering Ontology for Information Retrieval. Journal of Computing and Information Science in Engineering 8(1):011003-13.
Nivre, J. 2005. Dependency Grammar and Dependency Parsing. MSI report 05133. Växjö University: School of Mathematics and Systems Engineering, Växjö, Sweden
Raskin, V. 1983. A Concise History of Linguistic Semantics Pt 1. Department of English and Program in Linguistics, Purdue University, West Lafayette, IN.
Stuart, Lauren M. 2012. Computational Evaluation of a Noun-Driven Syntax Representation. Unpublished Master's Thesis, Linguistics Program, Purdue University, West Lafayette, IN.
Stuart, Lauren M., Julia M. Taylor, and Victor Raskin. 2012a. Towards Noun-Driven Parsing. In Proceedings of 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC) 1984-1989. Seoul, South Korea.
Stuart, Lauren M., Julia Taylor, and Victor Raskin 2012b. "Noun-Driven Syntactic Parsing for Natural Language Interfaces to Object-Centered Information Stores." In Proceedings of Society for Design and Process Science Conference (SDPS). Berlin, Germany.
Taylor, Julia M., Victor Raskin, Christian F. Hempelmann, and Max S. Petrenko 2010. "Multiple Noun Expression Analysis: An Implementation of Ontological Semantic Technology, Computational Linguistics - Applications Workshop, Wisla, Poland.
Taylor, Julia M., Victor Raskin, and Lauren M. Stuart (2012). Machine Human Understanding: Syntax and Semantics Revisited. IEEE-SMC Conference Seoul, S. Korea.

# Working Memory, Cognitive Miserliness and Logic as Predictors of Performance on the Cognitive Reflection Test 

Edward J. N. Stupple (e.j.n.stupple@derby.ac.uk)<br>Centre for Psychological Research, University of Derby Kedleston Road, Derby. DE22 1GB

Maggie Gale (m.gale@derby.ac.uk)<br>Centre for Psychological Research, University of Derby<br>Kedleston Road, Derby. DE22 1GB

Christopher R. Richmond (c.richmond1@unimail.derby.ac.uk)<br>Centre for Psychological Research, University of Derby<br>Kedleston Road, Derby. DE22 1GB


#### Abstract

The Cognitive Reflection Test (CRT) was devised to measure the inhibition of heuristic responses to favour analytic ones. Toplak, West and Stanovich (2011) demonstrated that the CRT was a powerful predictor of heuristics and biases task performance - proposing it as a metric of the cognitive miserliness central to dual process theories of thinking. This thesis was examined using reasoning response-times, normative responses from two reasoning tasks and working memory capacity (WMC) to predict individual differences in performance on the CRT. These data offered limited support for the view of miserliness as the primary factor in the CRT. The strongest predictor of CRT in both experiments was WMC. It is argued that while cognitive miserliness has been implicated in CRT performance, participants must also possess the requisite WMC and mindware to successfully complete it. Therefore, the psychological and psychometric properties of the CRT require continued study.


Keywords: Cognitive Reflection Test, Heuristics and Biases, Dual-process Theory, Belief-bias, Matching-bias, Reasoning, Cognitive Misers.

## Introduction

Dual-process theories of reasoning and judgment dissociate fast and frugal 'snap' judgments from slow, effortful and methodical analyses (e.g., De Neys, 2012; Evans, 2007; Stanovich, 2004) with the latter being viewed as being more likely to lead to normatively sanctioned answers in a variety of reasoning tasks. These contrasting processes are captured by heuristic-analytic tasks that involve a conflict between these processes (see Kahneman, 2011 for a recent review) and are referred to as Type 1 (heuristic) and Type 2 (analytic) (e.g., Evans, 2011).

Frederick (2005) devised the Cognitive Reflection Test (CRT) to examine the ability of participants to resist intuitive, tempting answers in favour of deeper, more analytic ones. By way of illustration, an example item from the CRT is "A bat and a ball cost $\$ 1.10$ in total. The bat costs $\$ 1$ more than the ball. How much does the ball cost?"

Most participants respond that the answer is 10 cents; however, a slower and more analytic approach to the problem reveals the correct answer to be 5 cents.
The CRT has been a spectacular success, attracting more than 100 citations in 2012 alone (Scopus). This may be in part due to the ease of administration; with only three items and no requirement for expensive equipment, the practical advantages are considerable. There have, moreover, been numerous correlates of the CRT demonstrated, from a wide range of tasks in the heuristics and biases literature (Toplak et al., 2011) to risk aversion and SAT scores (Frederick, 2005). Its publication was also timely as it coincided with the recent boom in dual process theories of thinking and reasoning (e.g., De Neys, 2012; Evans, 2007; Stanovich, 1999). The CRT and its items have been adopted as a testbed for the predictions of these theories (BourgeoisGironde, \& Vanderhenst, 2009; Campitelli \& Labollita, 2010; De Neys, Rossi, \& Houdé, 2013; Toplak et al., 2011). Bourgeois-Gironde and Vanderhenst (2009) have also highlighted the advantage that the CRT offers in terms of testing dual process predictions against arithmetic norms rather than the more controversial normative standards in logic or probability (see Elqayam \& Evans, 2011).
Toplak et al. (2011) presented perhaps the most comprehensive examination of the CRT, demonstrating considerable evidence for it as a predictor of non-normative responses to a battery of heuristics and biases tasks (each explicable) by dual process theories. Based on their findings Toplak et al. argued that the CRT predicts variance in rational thinking independently of intelligence, executive function and thinking dispositions, and that this variance is not insubstantial. Furthermore, Toplak et al. advance the CRT as a promising metric to tap into "What Intelligence Tests Miss" (Stanovich, 2009a) by accounting for rational thinking tendencies that are not captured by standard IQ tests (Stanovich suggests Dysrationalia as a term for people with higher IQ scores who fail on heuristics and biases tasks because they lack these thinking tendencies).

Stanovich (2009b) describes these rational thinking tendencies in a rational thinking taxonomy. Important categories for the CRT include, cognitive miserliness - the well-documented tendency to expend as little cognitive effort as is necessary to complete a task (first coined by Fiske \& Taylor, 1984); and, 'mindware gaps' - whereby the necessary cognitive rules, strategies, or belief systems are lacking, corrupted or are not applied.

Moreover, De Neys, Rossi, and Houdé (2013) presented evidence in support of cognitive miserliness as an explanation of performance on the CRT, based on confidence ratings that demonstrated diminished confidence ratings for participants who give the '10 cents' response to the 'Bat and Ball' question. De Neys et al. argue that even though the participants had an intuitive sense of the correct response they still responded incorrectly. They explicitly argue that their data indicate that, while they appear to be cognitive misers, participants are not offering erroneous responses in blissful ignorance.

In further support of this position, Campitelli and Labollita (2010) investigated how individual differences in cognitive reflection impacted on decision-making. They argue that cognitive reflection is indicative of a thinking disposition related to Baron's (1988) proposals about Actively Open Minded Thinking. This thinking tendency is an obvious contrast with cognitive miserliness. Active Open Minded Thinking is associated with enhanced performance on a range of heuristics and biases tasks including the generation of alternatives and belief based reasoning tasks (Stanovich \& West, 1999).

It would appear that the case for the CRT as a measure of cognitive miserliness is compelling. However, Thompson et al., (in press), examined the CRT as part of a paper testing the influence of perceptual fluency (Alter, Oppenheimer, Epley \& Eyre, 2007) and answer fluency in priming deliberative thinking (Thompson, Prowse Turner, \& Pennycook, 2011). They demonstrated that a degraded presentation of the CRT slowed participants down (conducive to analytic thinking and the converse of cognitive miserliness), but that this failed to facilitate correct responses among all but the most cognitively able participants (those in the uppermost quartile for IQ). These data suggest that increased response times to the CRT which potentially ameliorate cognitive miserliness by encouraging greater cognitive effort - are not universally beneficial. These data, moreover, suggest that there is an important role for cognitive capacity (or working memory) in gaining the benefits of slower Type 2 processing.

In studies of syllogistic reasoning, response-times are predictive of normative responding, but this is not universal across problem types (Stupple, Ball, Evans and KamalSmith, 2011). Stupple, et al. demonstrated that inflated response times predicted normative responding where there were conflicts of belief and logic, and that this effect on normative responding was particularly associated with response times for invalid-believable problem types.

Further support for the utility of response times as a predictor of normative responding in tasks with a dual process conflict was reported by Stupple Ball and Ellis (2013), who created a heuristic-analytic conflict using matched and non-matched surface features in syllogistic reasoning problems (Stupple \& Waterhouse, 2009). Stupple et al. (2013) noted that increased response times for invalid matching problems in a syllogistic reasoning task were associated with an increase in the overall normative responding. In contrast, increased response-times for valid non-matching problems were associated with decreased normative responding. These data demonstrate that it is not just the avoidance of miserliness that is important, but also that being sensitive to normative responses, perhaps by possessing the required mindware is important ${ }^{1}$. In short, a successful use of cognitive resources requires possession of the right mindware or the application of a sound strategy to be successful. Increased time deriving a response may indicate that Type 2 processing has occurred, but it is a fallacy to assume that the correct or normative answer will follow. A slow, effortful, but erroneous process cannot be characterized as the response of a miserly participant.

It is, nonetheless, argued that response times are vital to unpacking the predictions of dual process theories and that willingness to engage in time-consuming Type 2 processing on a syllogistic reasoning task should be predictive of willingness to engage in such processing on the CRT. A disposition to devote cognitive resources to a task coupled with the right mindware, however, may not be enough to find the correct answer if a participant has insufficient cognitive resource to reach the correct or normatively sanctioned conclusion.

Working memory capacity (WMC) has been shown to be important to reasoning performance, and to the process of analytic thinking (Bacon, Handley, Dennis \& Newstead, 2008; Copeland, \& Radvansky, 2004). Working Memory is, moreover, central to measures of intelligence (e.g., Kyllonen \& Christal, 1990). Frederick (2005) makes sound arguments to differentiate the CRT from intelligence measures, but there is yet to be a detailed examination of the importance of WMC in solving the CRT. Detecting the error in the heuristic response to the CRT is arguably only the first step towards solving the problems in the CRT. Working out the correct response is likely to involve working memory demand, for example, when participants consider the candidate values for the ball and then concurrently calculate the total value of the bat and the ball. This argument is supported by the finding from Thompson et al. (in press) that Type 2 processing may only benefit the most cognitively able (and by implication the highest WMC) participants on the CRT. Toplak et al. (2011) argue that the items on the CRT are not insight problems (see Gilhooly \& Murphy, 2005) - which do not incur significant working memory load - and are instead analytic problems, which do.

[^217]Toplak et al. (2011) also acknowledge the influence of WMC and examine the role of CRT in predicting performance on heuristics and biases tasks, with the influence of WMC factored out. The focus here is instead upon the extent that WMC is predictive of the CRT in conjunction with Cognitive Miserliness, and sensitivity to normative considerations.

Mean response-times to syllogistic reasoning problems were used as an index of cognitive miserliness, a logic index (e.g., Stupple et al., 2013) was calculated to generate a measure of normative responding and a composite working memory score derived from Operation Span, Symmetry Span and Reading Span measures (Unsworth et al., 2005) was used as a measure of working memory capacity.

It is argued that Toplak et al.'s (2011) miserliness account of the CRT predicts that participants who devote the longest times to solving syllogisms would also be those who were most successful in solving CRT items. It was predicted that this would be the strongest predictor of CRT performance and the first factor included in the regression model by the stepwise procedure. It was also predicted that normative sensitivities and WMC would be significant predictors of CRT performance, but that these would account for less variance in CRT performance than the miserliness measure.

## Method Experiment 1

Design Predictor variables were generated from the working memory span tasks (Unsworth et al., 2005) and the beliefbias reasoning task. Mean response-times to belief bias problems were calculated to generate an index of miserliness; acceptance rates for belief bias problems were used to generate a logic index. The dependent variable was the score on the CRT.
Participants Sixty-five undergraduates from the University of Derby, aged 18-45, were recruited via opportunity sampling. Participants had no training in formal logic and had not previously studied the psychology of reasoning or encountered the CRT. Each received a voucher (value £5) for participating.
Materials and procedure Participants received 16 target syllogisms counterbalanced for figure and mood. Belieforiented contents were those employed by Stupple and Ball (2008). There were equal numbers of valid and invalid problems, and believable and unbelievable conclusions.

Logic index was calculated by adding acceptance rates for Valid Believable and Valid Unbelievable problems and subtracting total acceptance rate for Invalid Believable and Invalid Unbelievable problems (Valid Believable + Valid Unbelievable - Invalid Believable - Invalid Unbelievable).

Syllogisms and instructions were presented with Authorware 6.5 on a PC. Problems were counterbalanced, with contents rotated through them. WMC was measured using three complex span tasks (Unsworth et al., 2005) in EPrime Version 2.0. These consisted of Automated Operation Span, Automated Symmetry Span and Automated Reading Span (see Unsworth et al., 2005 for details). The three
measures of working memory capacity were combined to form a composite working memory score (Bartlett, 1937), derived from the three absolute span scores (defined as the sum of all sets of items that are recalled without error, Unsworth et al, 2005). The CRT was a pen and paper task.

## Results Experiment 1

A Stepwise Multiple Regression tested the relative predictive strength of response-times and logic index in a belief-bias reasoning task and WMC for performance on the CRT. The Mean CRT score for the sample in Experiment 2 was 1.32 ( $\mathrm{SD}=1.11$ ) which is well within the range described by Frederick (2005) ${ }^{2}$.

Data indicated that WMC reliably accounted for $27 \%$ of the variability in CRT scores with participants with higher WMC scores performing better on the CRT than those with lower scores. Surprisingly, no further steps in the regression analysis significantly increased the variance accounted for as neither the Logic index nor the Response times were reliable predictors. Response-times demonstrated a nonsignificant correlation with CRT scores close to zero.

Table 1: Stepwise Regression Analysis of Working Memory Capacity, Logic index and Reasoning Responsetimes for Belief Bias problems as predictors of the CRT
Predictors

| Model 1 | $\mathrm{R}^{2}=.28, \mathrm{R}_{\text {add }}^{2}=.27$ |
| :--- | :--- |
|  | $\mathrm{~F}(1,64)=24.87, \mathrm{p}<.001$ |
| WMC | $\beta=.529, \mathrm{p}<.001$ |

Excluded
Logic index $\quad \beta=.156, p=.146$
Response-times $\beta=.052, p=.641$
Durbin Watson= 1.72, VIFs ranged from 1.01 to 1.08

## Interim Discussion

These findings were contrary to the expectation as there was no reliable relationship shown for response times to syllogistic reasoning problems. Moreover, the variance explained by the composite measure of WMC was by far the most substantial predictor.
These results were surprising and may be specific to the syllogistic reasoning task employed. While there are similarities between belief bias problems and the CRT, in that some items may require the inhibition of an initial heuristic response, it is not the case that the CRT involves belief inhibition per se. A second experiment utilizing the same methodology, but employing an alternative set of reasoning problems (Roberts, 2005) that are also known to induce a heuristic-analytic conflict - the matching bias problems used by Stupple et al. (2013), was conducted.

[^218]Consistent with the CRT these matching problems feature conclusions, which are tempting to endorse, or reject based on their surface features. For example, in the case of the second item in the CRT: "If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? $\qquad$ minutes?" The most frequent erroneous response is 100 minutes, whereby participants may be matching their answer to the surface features of the problem. Similarly, performance on matching-bias syllogisms requires the inhibition of an inclination to respond based on whether surface features of conclusions and premises match, (and possessing the mindware to eliminate a double negation). It was hypothesized that (1) working memory capacity would again be a significant predictor of CRT scores, and, (2) that logic index and response times would predict CRT scores. However, these predictions were made with reduced confidence in the light of the findings from Experiment 1.

## Method Experiment 2

Design Response times and conclusion acceptance rates from the matching bias reasoning task were used as predictors and the three Working Memory Span measures (Operation span, Reading span and Symmetry span (Unsworth et al., 2005) were again used to derive a composite WMC score. The dependent variable was the CRT scores.
Participants Forty-nine undergraduates from the University of Derby aged 18-45 were recruited via opportunity sampling. Participants had no training in formal logic and had not previously studied the psychology of reasoning or encountered the CRT. Each received a voucher (value $£ 5$ ) for participation.
Materials and Procedure Sixteen one-model syllogisms were presented. Conclusions either matched the premises (premises and conclusions were traditional affirmative or both were double negated), or were not matched with the premises - traditional affirmative premises were presented with double negated conclusions or double negated premises were presented with traditional affirmative conclusions. For non-conflict problems, analytic and heuristic strategies produced the same response, whereas for conflict problems analytic and heuristic matching strategies were in competition. Syllogism content involved combinations of professions and pastimes. These were rotated through the different problems. Reasoning problems, WMC measures and the CRT were administered identically to Experiment 1.

## Results Experiment 2

The Mean CRT score for the sample in Experiment 2 was 1.12 ( $\mathrm{SD}=1.14$ ) which is well within the range described by Frederick (2005), although for this experiment it was below the overall average reported by Frederick (2005).

A Stepwise Multiple Regression was conducted to test the relative predictive strength of response-times, and logic in a
matching bias reasoning task and WMC for performance on the CRT. Data indicated that WMC reliably accounted for $23 \%$ of the variability in CRT scores in the first model, with participants with higher composite WMC scores demonstrating better performance on the CRT than those with lower scores. In a second model, the variance explained increased to $34 \%$ with the addition of the Logic index predictor. As with the first experiment, responsetimes did not reliably account for variance in CRT scores. More surprising, was the fact that the response-times correlated negatively (albeit unreliably) with CRT scores in the opposite direction to that predicted.

Table 2: Stepwise Regression Analysis of Working Memory Capacity, Logic index and Reasoning Responsetimes for Matching Bias problems as predictors of the CRT

| Predictors |  |
| :--- | :--- |
| Model 1 | $\mathrm{R}^{2}=.245, \mathrm{R}_{\text {adj }}^{2}=.229$ |
|  | $\mathrm{~F}(1,48)=15.54, \mathrm{p}=.001$, |
| WMC | $\beta=.495, \mathrm{p}<.001$ |
|  |  |
| Model 2 | $\mathrm{R}^{2}=.365, \mathrm{R}_{\text {adj }}^{2}=.338$ |
|  | $\mathrm{~F}(2,47)=13.52, \mathrm{p}=.001$, |
|  | F change, $\mathrm{p}=.004$ |
| WMC | $\beta=.426, \mathrm{p}=.001$ |
| Logic index | $\beta=.354, \mathrm{p}=.004$ |
|  |  |
| Excluded | $\beta=-.149, \mathrm{p}=.203$ |
| Response-times |  |

Durbin Watson=1.70, VIF $=1.04$

## Discussion

The experiments presented here tested the relative contributions of Response times to reasoning tasks (as an index of cognitive miserliness), Logic Index (as a measure of sensitivity to normative responses) and WMC to predicting variance in the CRT. Consistent with predictions, WMC was a reliable predictor of performance on the CRT in both experiments - and was a substantially stronger predictor than expected. Moreover, the unexpected null finding for response times, suggested that if the CRT is conceptualized as a measure of cognitive miserliness then it might not convincingly generalize beyond the arithmetic based problems to standard dual processing tasks such as belief bias or matching bias syllogisms. If the CRT is a general measure of cognitive miserliness then those participants responding primarily with the Type 2 answers to the CRT should engage in more Type 2 processing on syllogistic reasoning tasks as indexed by increased response times. These data suggest that this was not reliably the case.

WMC correlating most strongly with performance on the CRT is somewhat problematic for the use of the test as a measure of miserliness. Individuals with lower WMC may expend a great deal of effort in attempting to solve heuristic-
analytic problems, but lack the capacity to maintain their representation of, for example, possible ball costs relative to the bat as they work through the alternatives. Participants with higher WMC may find the cognitive costs less expensive and, thus be more willing to pay them ${ }^{3}$.

Cognitive miserliness could be argued to be relative to the cognitive resource of the participant. A participant with a high WMC who provides heuristic responses to the CRT would be categorized appropriately as a cognitive miser as they had the necessary cognitive resources, but chose not to apply them to the task. In contrast, a participant with lower WMC who devotes considerable time and effort, but arrives at a heuristic answer would be inappropriately described as miserly (perhaps they could be considered cognitive wastrels instead). It may be that those participants with greater WMC can engage in the deliberative thought required to avoid the heuristic response with relatively less effort when compared to those with lesser WMC. This reduced cognitive cost may become affordable to the participants with more miserly tendencies, but who also have ample working memory resources available ${ }^{4}$.

De Neys et al. (2013) suggest that participants are aware of the incongruity of answering 10 cents to the bat and ball question, but often fail to engage the deliberative processing required for the correct ' 5 cents' answer. We would add to this claim that while cognitive miserliness is almost certainly a factor, our data indicate that, for a proportion of participants at least, they may not have the cognitive resources to pursue their metacognitive uncertainty about their intuitive response. Alternatively, the intuitive response may offer a cognitive escape hatch, if processing demands are too great (cf. Quayle \& Ball, 2000). Similarly, with regard to Thompson et al.'s (in press) findings - that only the most able participants benefitted from the Type 2 processing that dis-fluent stimuli encouraged in terms of the accuracy of their responding demonstrating that increased response times may be important to success on the CRT, but they are not sufficient. Further investigation is required to understand the nuanced interplay between miserliness and cognitive ability/working memory capacity on the CRT.

It was notable that there was not a reliable relationship between normative responding in a belief bias task with performance on the CRT, but that normative responding on the matching bias task was a highly reliable predictor of CRT performance. A possible account of the discrepancy between studies could be based on the manner in which the heuristic-analytic conflict is resolved. Optimal performance

[^219]on belief bias problems requires an ability to inhibit belief driven responses while searching for alternative models (Stupple et al., 2011), whereas the matching bias problems required an explicit awareness of the logic of double negatives - such that errors could be characterized as the result of missing or corrupted 'mindware' (Stanovich, 2009b). This difference in the source of the heuristicanalytic conflict could potentially account for the discrepancy between problem types. This further contrasts with the proposal of the CRT as an index of cognitive miserliness. The absence of the appropriate mindware for double negations among those participants who score lowest on the CRT would appear to indicate a lack of an understanding of logic or rule based thinking, rather than, an unwillingness to engage in the requisite cognitive effort. This is inconsistent with the arguments from Toplak et al. (2011), who suggest that knowledge gaps represent a major class of reasoning error but that: "The potency of the CRT as a predictor of performance on heuristics-and-biases tasks certainly does not derive from its ability to assess knowledge gaps, because it clearly does no such thing." (Toplak et al, 2011, p. 1284). The variance in CRT scores explained by normative responding to matching-bias syllogisms cannot reasonably be claimed as a causal link, but suggests an association between possessing the necessary cognitive rules or strategies for detecting matching bias conflicts and the heuristic-analytic conflicts that are implicated in success on the CRT. We would argue that examination of the CRT as an index of conflict detection also warrants further investigation.

Therefore, it is advocated that self-report measures such as the Need for Cognition (Cacioppo \& Petty, 1982) or Rational-Experiential Inventory (Epstein, 1994) continue to be used alongside the CRT to quantify the subjective experience of miserliness. This subjective experience is likely to co-vary with cognitive capacity - relative to the task demands. The CRT as an index of cognitive miserliness presupposes a degree of equality in our cognitive wealth. Self-report measures may supplement the CRT by offering insight into the experience of how effortful the task was and by quantifying self-perceptions of cognitive miserliness.

Nonetheless, we agree with Toplak et al. (2011) that the CRT captures variability in performance on heuristics and biases tasks that are not captured by IQ tests and the CRT remains a promising measure to explore in this regard. What remains clear from these data is that explaining the psychological properties of the CRT is not a simple task, and while it is undoubtedly an influential task that will remain popular among dual process theorists, the precise nature of its psychometric and psychological properties require continued study.

## Acknowledgments

This research was funded through the Research Inspired Curriculum Fund at the University of Derby.

## References

Alter, A. L., Oppenheimer, D. M., Epley, N., \& Eyre, R. N. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. Journal of Experimental Psychology: General, 136, 569-576.
Bacon, A.M., Handley, S.J., Dennis, I. \& Newstead, S.E. (2008) Reasoning strategies: the role of working memory and verbal-spatial ability. European Journal of Cognitive Psychology 20 (6), 1065 - 1086
Baron, J. (1988). Thinking and deciding. New York: Cambridge University Press. (4th edition).
Bartlett, M. S. (1937). The statistical conception of mental
factors. British Journal of Psychology, 28, 97-104.
Bourgeois-Gironde, S., \& Vanderhenst, J. B. (2009). How to open the door to System 2: Debiasing the Bat and Ball problem. In S. Watanabe, A.P Bloisdell, L. Huber, A. Young (Eds.), Rational animals, irrational humans. Tokyo: Keio University Press.
Cacioppo, J. T. \& Petty, R. E. (1982). The need for cognition. Journal of Personality and Social Psychology, 42, 116-131.
Campitelli, G., \& Labollita, M. (2010) Correlations of cognitive reflection with judgments and choices. Judgment and Decision Making, 5, 182-191.
Copeland, D. \& Radvansky, G. (2004). Working memory and syllogistic reasoning, The Quarterly Journal of Experimental Psychology Section A, 57:8, 1437-1457
De Neys, W. (2012). Bias and conflict: A case for logical intuitions. Perspectives on Psychological Science, 7 (1), 28-38.
De Neys, W., Rossi, S., \& Houdé, O. (2013). Bats, balls, and substitution sensitivity: Cognitive misers are no happy fools. Psychonomic Bulletin \& Review
Elqayam, S., \& Evans, J. St. B. T. (2011). Subtracting "ought" from "is": Descriptivism versus normativism in the study of human thinking. Behavioral \& Brain Sciences, 34, 233-248.
Epstein, S. (1994). Integration of the cognitive and psychodynamic unconscious. American Psychologist, 49, 709-724
Evans, J. St. B. T. (2007). On the resolution of conflict in dual process theories of reasoning. Thinking \& Reasoning, 13, 321-339.
Evans, J. St. B. T. (2011). Dual-process theories of reasoning: Contemporary issues and developmental applications. Developmental Review, 31, 2, 86-102
Fiske, S. T., Taylor, S. E. (1984). Social cognition. MA: Addison-Wesley Pub. Co
Gilhooly, K. .J., \& Murphy, P.(2005). Differentiating insight from non-insight problems. Thinking \& Reasoning 11 (3), 279-302.
Kahneman, D. (2011) Thinking, Fast and Slow. London: Allen Lane.
Kuhl, J. (2000). A functional-design approach to motivation and self-regulation: The Dynamics of personality systems and interactions. In M. Boekaerts \& P.R. Pinrich (Eds.), Handbook of Self-regulation. San Diego, Academic Press.

Kyllonen, P. C., \& Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! Intelligence, 14, 389-433.
Quayle, J. D., \& Ball, L. J. (2000). Working memory, metacognitive uncertainty, and belief bias in syllogistic reasoning. Quarterly Journal of Experimental Psychology, 53A, 1202-1223.
Rips, L. (1994). The psychology of proof. Cambridge, MA: MIT Press.
Roberts, M. J. (2005). Expanding the universe of categorical syllogisms: A challenge for reasoning researchers. Behavior Research Methods, 37, 560-580.
Stanovich, K. E. (2004). The Robot's Rebellion: Finding Meaning in the age of Darwin. Chicago: Chicago University Press.
Stanovich, K. E. (2009a). What intelligence tests miss: The psychology of rational thought. New Haven: Yale University Press.
Stanovich, K. E. (2009b). Distinguishing the reflective, algorithmic and autonomous minds: Is it time for a triprocess theory? In J.St.B.T.Evans \& K. Frankish (Eds.), In two minds: Dual processes and beyond. Oxford: Oxford University Press.
Stupple, E. J. N., \& Ball, L. J. (2008). Belief-logic conflict resolution in syllogistic reasoning: Inspection-time evidence for a parallel-process model. Thinking \& Reasoning, 14, 168-181.
Stupple, E J. N., Ball, L. J., Evans, J. St. B. T., \& KamalSmith, E. N. (2011). When logic and belief collide: Individual differences in reasoning times support a selective processing model. Journal of Cognitive Psychology, 23, 931-941.
Stupple, E.J.N. Ball, L.J. \& Ellis, D. (2013). Matching bias in syllogistic reasoning: Evidence for a dual-process account from response times and confidence ratings. Thinking \& Reasoning, 19(1):54-77
Stupple, E. J. N., \& Waterhouse E. F. (2009). Negations in syllogistic reasoning: evidence for a heuristic-analytic conflict. Quarterly Journal of Experimental Psychology, 62, 1533-1541.
Thompson, V. A., Prowse Turner, J. A., \& Pennycook, G. (2011). Intuition, metacognition, and reason. Cognitive Psychology, 63, 107-140.
Thompson, V. A., Prowse Turner, J. A., Pennycook, G., Ball, L. J., Brack, H., Ophir, Y., \& Ackerman, R. (In press) The role of answer fluency and perceptual fluency as metacognitive cues for initiating analytic thinking. Cognition
Toplak, West and Stanovich, K.E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. Memory \& Cognition, 39, 7, 1275-1289
Unsworth, N., Heitz, R.P., Schrock, J. C., \& Engle, R. W. (2005) An automated version of the operation span task. Behavior Research Methods, 37, 498-505.

# Syllable Frequency and Stress Priming Interact in Reading Italian Aloud 

Simone Sulpizio (simone.sulpizio@unitn.it)<br>Department of Psychology and Cognitive Science, University of Trento, Corso Bettini 31<br>38068 Rovereto (Tn), Italy<br>Remo Job (remo.job@unitn.it)<br>Department of Psychology and Cognitive Science, University of Trento, Corso Bettini 31<br>38068 Rovereto (Tn), Italy


#### Abstract

In current theories of word reading the structure and operations of the phonological buffer are quite underspecified. We investigated this issue by running a reading aloud experiment in Italian. We adopted a priming paradigm, with three-syllabic words as primes and targets and we jointly manipulated two effects ascribed to the stage of phonological and phonetic encoding, that is stress priming and syllable frequency. Target words varying for the frequency of their initial syllable were preceded by words congruent or incongruent for the stress pattern. The results showed an interaction between syllable frequency and stress prime, with the stress congruency effect larger for the targets with low-frequency first syllable. This result suggests that, in reading aloud, stress assignment and syllable computation have a tight time dynamics in the phonological output buffer, and that the process at the level of phonology-to-phonetic interface operates interactively.


Keywords: Lexical stress; syllable frequency; phonological-to-phonetic interface; phonological buffer; reading aloud.

## Introduction

Reading aloud requires the execution of multiple operations, e.g., perceiving the stimulus, converting the printed information in a speech signal, and articulating the word's sounds, taking into account both segmental (e.g., sounds) and suprasegmental (e.g., stress) information. While many reading studies have investigated the operations involved in word recognition, the phonological encoding of a word and its phonetic realization have received less attention. The same happens with computational models of reading aloud: They usually implement in a detailed way the procedures readers use to recognize words, but they are less specific about those phenomena related to the production stages (see, e.g., Coltheart, Rastle, Perry, Langdon, \& Ziegler, 2001), and the very few that have attempted to implement procedures for stress assignment differ in the solutions they propose (see, e.g., Arciuli, Monaghan, \& Seva, 2010; Perry, Ziegler, \& Zorzi, 2010; Rastle \& Coltheart, 2000; Sibley, Kello, \& Seidenberg, 2010).

The investigation of the production stage of reading aloud can benefit from the speech production literature, as it has been argued that speech production and reading aloud may share the last stages of processing, specifically the phonological and phonetic encoding of the word (Roelofs, 2004). In the model developed by Levelt and colleagues (Levelt, Roelofs, \& Meyer, 1999) it is assumed that during
phonological encoding speakers retrieve in parallel the segmental material and the metrical structure - number of syllables and word's stress pattern - and combine them into the phonological word (see also Roelofs \& Meyer, 1998). At this point, the phonological word is phonetically encoded and it is then translated into its phonetic realization.

A detailed architecture of the phonological and phonetic encoding, however, has never been proposed by any model of word reading and how the reading system converts abstract phonological information into phonetic representations is still an open issue. An effort in this direction has been done by Perry and colleagues (2010): In their CDP++ model of reading, at the level of phonological output buffer, the authors implement a double process analogous to the one proposed for word production, with two different loci for stress and phonemes activation. In particular, the model presents stress-output nodes, i.e. nodes specifying the position of the stress within the lexical string. Such nodes are activated autonomously from the segmental information, although full processing of the latter is conditional upon the former: Articulation of the word phonemes cannot be initiated until the word stress has been fully determined. However, despite the improvement of the phonological output buffer, nothing is said about how segmental and suprasegmental information are assembled together, and how the selected phonological information is converted into a phonetic representation.

Recent empirical data that can help to better understand how the phonological and phonetic encoding work within the reading system. Some studies run in Italian (Colombo \& Zevin, 2009; Sulpizio, Boureux, Burani, Deguchi, \& Colombo, 2012a; Sulpizio, Job, \& Burani, 2012b), support the view that metrical and segmental information are autonomously involved in planning and assembling an utterance, both when stress is sub-lexically computed (Colombo \& Zevin, 2009; Sulpizio et al., 2012a) or lexically retrieved (Sulpizio et al., 2012b). In particular, the latter study showed an effect of stress position priming for segmentally different prime-target pairs. Specifically, readers are faster in reading a word when it is preceded by a word with the same stress, e.g., TESsera (card) - BUfala (hoax), than when in is preceded by a word with a different stress, e.g. cuGIno (cousin) - BUfala (hoax) ${ }^{1}$. The pattern was interpreted as showing that stress priming affects the

[^220]stage of phonological word encoding in the phonological buffer.

An effect that has also been ascribed to the later stages of reading aloud is that of syllable frequency. Researches in different languages have shown that participants are faster in producing a word that starts with a high-frequency syllable than one with a low-frequency syllable (see, e.g, for Dutch: Cholin, Levelt, \& Schiller, 2006; English: Cholin, Dell, \& Levelt, 2011; French: Laganaro \& Alario, 2006; Italian: Sulpizio \& Job, 2010; Spanish: Carreiras \& Perea, 2004) and there is consensus on the claim that such effect is attributed to the phonetic encoding, when readers convert the abstract phonological word into abstract motor programs.

Jointly considering the effects of stress assignment and of syllable frequency in reading aloud may allow us to better articulate the operations involved in the phonological-to-phonetics interface, the rather neglected and oversimplified component of reading models. Both stress priming and syllable frequency are assumed to affect the latest stages of reading process, when readers (a) spell out segmental and metrical information and (b) plan the articulation of the word, with syllable frequency affecting the word's phonetic encoding (Carreiras, Mechelli, \& Price, 2006; Laganaro \& Alario, 2006). Thus, an additive pattern of syllable frequency and stress priming would be consistent with the proposal of two separate consecutive stages for the two effects, or with the assumption of a threshold of activation for one component before the other may start its computations (Perry et al., 2010): In such a view, word phonetic encoding can start only after the processing of stress assignment ends, with the consequence that a delay in the computation of stress would affect the phonetic encoder independently from how fast its content might be computed. Differently, an interaction between syllable frequency and stress priming would suggest that both the effects may concurrently affect the same stage of processing, i.e. the phonological-to-phonetic interface. If this is the case, it would suggest that: a) there is no reason to postulate a threshold setting the timing of either segmental or suprasegmental activation; b) the mapping of the phonological word into phonetic codes may occur through an interactive process.

## Experiment

Three-syllabic Italian words were used as stimuli as stress position for these words is not always predictable. Indeed, Italian three-syllabic words have two main stress patterns (Thornton, Iacobini, \& Burani, 1997): Antepenultimate stress (i.e., the first syllable bears stress, e.g., TAvolo 'table'), and penultimate stress (i.e., the second syllable bears stress, e.g., coLOre 'color'). Although their distribution differs - $80 \%$ of three-syllable words bear penultimate stress and $18 \%$ bear antepenultimate stress ${ }^{2}-$

[^221]the two patterns are lexically stored within the phonological lexicon and the asymmetry does not affect lexical reading (Paizi, Zoccolotti, \& Burani, 2011).

By jointly manipulating stress priming and syllable frequency we aimed at investigating the operations involved in the phonological-to-phonetic interface that take place during the later stages of word reading. Specifically, if stress priming and syllable frequency originate at two separate stages of processing or the former is governed by a threshold mechanism, then the stress priming effect should be of similar size for both words starting with a high- and words staring with a low-frequency syllable. Differently, if stress priming and syllable frequency may concurrently affect the phonological-to-phonetic interface, an interaction between the two effects should be expected.

## Method

## Participants

Twenty-four students (14 male, mean age: 24, sd: 3.8) of the University of Trento. They were all Italian native speakers and they had normal or corrected-to-normal vision. They received credit course for their participation.

## Materials and Design

Four sets of three-syllabic words were used as targets. The sets were selected by combining two variables: Frequency of the first syllable (high or low) and stress pattern (penultimate or antepenultimate). Each set was composed of 22 low-frequency words selected from the CoLFIS database (Bertinetto et al., 2005). Stimuli were matched on length in letters, orthographic neighborhood size, orthographic neighbors' summed frequency, frequency of the second and third syllable, mean bigram frequency, orthographic complexity, initial phoneme (Table 1), and had a stress neighborhood composed mainly of stress friends (Burani \& Arduino, 2004).

Table 1. Summary statistics: means (and standard deviations) for the three-syllabic target words.

|  | First Syllable Frequency |  |  |  |
| :--- | :---: | :---: | :---: | :---: |
|  | High |  | Low |  |
|  | Pen. | Antep. | Pen. | Antep. |
| stress | stress | Stress | Stress |  |
| First Syllable | 690 | 720 | 28 | 41 |
| Frequency | $(561)$ | $(505)$ | $(25)$ | $(30)$ |
| Second+third | 1588 | 1711 | 2088 | 2228 |
| Syllable | $(847)$ | $(809)$ | $(919)$ | $(769)$ |
| Frequency |  |  |  |  |
| Word frequency | 4.5 | 6.5 | 7.1 | 6.05 |
|  | $(4.9)$ | $(11.2)$ | $(12)$ | $(7.3)$ |
| Length in letters | 7 | 6.8 | 7.1 | 7 |
|  | $(0.6)$ | $(0.4)$ | $(0.3)$ | $(0.2)$ |
| Mean Bigram | 11.6 | 11.5 | 11.4 | 11.5 |
| frequency | $(0.2)$ | $(0.2)$ | $(0.2)$ | $(0.4)$ |
| N of orthographic | 1 | 1 | 1.1 | 1 |
| neighbors | $(1.2)$ | $(1.1)$ | $(0.9)$ | $(1)$ |


| Neighbors' | 4.7 | 8.1 | 2.1 | 6 |
| :--- | :---: | :---: | :---: | :---: |
| summed frequency | $(9.9)$ | $(22.9)$ | $(2.8)$ | $(14.9)$ |

Note: Pen. = penultimate stress; Antep. $=$ antepenultimate stress; syllable frequency measures are calculated out of 1 milion occurrences (Stella \& Job, 2001); word frequency measures are calculated out of 1 million occurrences (Bertinetto et al., 2005); mean bigram frequency is $\log$ transformed on the basis of the natural logarithm.

Targets were pre-tested to ensure that none of the initial syllables was a probabilistic orthographic cue for stress (Arciuli, Monaghan, \& Ševa, 2010). Thus, syllable frequency was not expected to interact with word's stress pattern. To further rule out such possibility, we ran a pilot experiment asking 18 university students to read aloud all targets. Stimuli appeared in capital letters in the center of the screen, after a fixation cross displayed for 400 ms . Each stimulus remained on the screen until the participant began to read or for a maximum of 1500 ms . The presentation order was randomized between participants. Mean RTs for correct responses were submitted to a 2 (high- vs. lowfrequency syllable) x 2 (penultimate vs. antepenultimate stress) ANOVA. The analysis showed an effect of syllable frequency $\left(F_{1}(1,17)=22.19, M S E=1246, p<.01 ; F_{2}(1,84)\right.$ $=17.29, M S E=2033, p<.01$ ), with faster reaction time for words with a high-frequency syllable. Neither stress type $\left(F_{l}(1,17)=1.60, M S E=246 ; F_{2}<1\right)$ nor the interaction were significant $\left(F_{l}(1,17)=3.60, M S E=217 ; F_{2}<1\right)$. No effect was significant in the analysis of errors (4.8\%). Results of the pilot experiment suggest that targets' first syllables are not preferentially associated with a certain stress pattern, as suggested by the absence of a syllable frequency by stress type interaction.

Two sets of 44 high frequency three-syllabic words were used as primes. One set included penultimate stress words and the other antepenultimate stress words, all selected from CoLFIS (Bertinetto et al., 2005). The two sets were matched on: Length in letters, orthographic neighborhood size, orthographic neighbors' summed frequency, mean bigram frequency, and initial phoneme (Table 2). Primes were paired with target words in such a way that neither semantic relation nor orthographic overlapping existed between prime and target. Targets were divided between the two prime stress conditions (congruent and incongruent) and each prime word was paired with both a congruent (e.g., niPOte 'nephew' - laSAgna 'lasagna') and an incongruent stress target (e.g., niPOte 'nephew' - MUscolo 'muscle').

Table 2. Summary statistics: means (and standard deviations) for the three-syllabic prime words.

|  | Stress Type |  |
| :--- | :---: | :---: |
|  | Pen. | Antep. |
| Word frequency | 216 | 228 |
|  | $(118)$ | $(127)$ |
| Length in letters | 6.9 | 6.7 |
|  | $(0.7)$ | $(0.7)$ |


| Mean Bigram frequency | 11.5 | 11.4 |
| :--- | :---: | :---: |
|  | $(0.4)$ | $(0.3)$ |
| N of orthographic neighbours | 1.9 | 1.8 |
|  | $(1.7)$ | $(1.4)$ |
| Neighbors' summed frequency | 51.5 | 52.6 |
|  | $(68.7)$ | $(65.1)$ |

Note: Pen. = penultimate stress; Antep. = antepenultimate stress; syllable Word frequency measures are calculated out of 1 million occurrences (Bertinetto et al., 2005); mean bigram frequency is log transformed on the basis of the natural logarithm.

The Experiment had a 2 (congruent $v s$. incongruent stress pattern) x 2 (high- $v s$. low-syllable frequency) design. Following the procedure adopted by Sulpizio and colleagues (2012b), prime-target pairs were divided in 4 pure blocks (prime and target sharing the stress pattern $\&$ target with a high-frequency initial syllable; prime and target sharing the stress pattern \& target with a low-frequency initial syllable; prime and target with different stress patterns \& target with a high-frequency initial syllable; prime and target with different stress patterns \& target with a low-frequency initial syllable). Furthermore, in each block, half of the targets had penultimate stress and half had antepenultimate stress, and in no case prime and target shared the initial phoneme. The order of stimuli was randomized within blocks and block order was counterbalanced across participants. Primes and targets were paired in such a way that for half of the participants a target was in a congruent stress condition (prime and target having same stress), and for the other half the same target was presented in the incongruent stress position (prime and target having different stress).

## Apparatus and procedure

Participants were tested individually. They were instructed to read the targets as quickly and accurately as possible. Each trial started with a fixation cross, centered on the screen, for 400 ms . The prime was then presented in lowercase letters just above the center of the screen for 86 ms and it was followed by a 86 ms blank; then, the target stimulus was displayed in upper-case letters just below the center of the screen. The target remained on the screen until the participant began to read it or for a maximum of 1500 ms . The inter-stimulus interval was 1500 ms . A practice session with 8 trials preceded the experiment. Naming times were recorded by means of E-Prime software. The experimenter noted the naming errors.

## Results

Responses shorter than 250 ms or longer than 1500 ms ( $2.4 \%$ of all data points) were excluded from the analyses. Naming errors, including both phonemic and stress errors, summed to $2.7 \%$ of all data points and were not analyzed. Results are reported in Figure 1.

A $2 \times 2$ analysis of variance with syllable frequency (high- vs. low-frequency syllable) and condition (congruent vs. incongruent stress) was conducted on the reaction times (RTs) of correct responses. The factors were within
participants $\left(F_{1}\right)$ and between items $\left(F_{2}\right)$. The main effect of condition was significant, with congruent target words read faster than incongruent target words $\left(F_{1}(1,23)=10.49\right.$, $M S E=3771, p<.01, \eta^{2}=.27 ; F_{2}(1,176)=51.49, M S E=$ 1558, $p<.001, \eta^{2}=.23$ ). The main effect of syllable frequency was also significant, showing that targets with an initial high-frequency syllable were read faster than targets with a low-frequency syllable $\left(F_{1}(1,23)=8.73, M S E=995\right.$, $p<.01, \eta^{2}=.31 ; F_{2}(1,176)=10.24, M S E=1558, p<.01$, $\left.\eta^{2}=.15\right)$. Finally, there was a significant interaction between congruency condition and syllable frequency, $\left(F_{1}(1,23)=\right.$ $4.39, M S E=675, p<.05, \eta^{2}=.16 ; F_{2}(1,176)=4.26, M S E=$ $\left.1558, p<.05, \eta^{2}=.12\right)$ : LSD post-hoc comparisons showed that the 55 ms stress priming effect ( $\mathrm{p}<.005, \eta^{2}=.31$ ) for targets with a low-frequency initial syllable was significantly different from the 31 ms effect ( $\mathrm{p}<.05, \eta^{2}=$. 23 ) for the targets with a high-frequency initial syllable.


Figure 1. Reaction times and percentage of errors by condition

The results of the present experiment are clear. Word targets preceded by stress-congruent primes were read faster than targets preceded by stress-incongruent primes. Moreover, words with a high-frequency first syllable were read faster than words with a low-frequency first syllable. Finally, the priming effect was larger for targets with a lowfrequency first syllable than for those with a high-frequency first syllable.

## Discussion

The main finding of our study is that syllable frequency and stress priming interact: Reading times are longer to incongruent prime-target stress pairs for both high- and lowfrequency syllable targets, but for the latter the difference is larger than for the former. Thus, syllable frequency modulates the impact of stress priming. The findings allow us to better understand some aspects of the mechanics of phonological and phonetic encoding during reading. They also provide hints on the relative timing of the operations underlying stress retrieval and word articulation in reading aloud.

The effect of syllable frequency has been generally
ascribed to the phonetic encoding level by assuming that speakers are facilitated in articulating those syllables they produce frequently. Specifically, Levelt et al. (1999) argue that high-frequency syllables can be retrieved from a mental syllabary, while low-frequency syllables are assembled using the phonological word as input. The assumption of a mental syllabary has been accepted by most of the reading literature which reported effects of syllable frequency in word and pseuwdoword reading tasks (Carreiras \& Perea, 2004; Carreiras et al., 2006; Laganaro \& Alario, 2006; Perea \& Carreiras, 1998; Sulpizio \& Job, 2010). Thus, also in word reading, the effect of syllable frequency can be located at the stage of phonetic encoding, that is in the phonological output buffer.

The effect of stress priming has been ascribed to mechanisms operating at the level of the phonological buffer (Sulpizio et al. 2012b; see also Colombo \& Zevin, 2009; Sulpizio et al., 2012a). In such a view, the preactivation of metrical information by a prime word would affect the component of the phonological buffer responsible for metrical encoding by affecting the timing of the operations the system performs to retrieve and assign the correct stress pattern to the target word.

The interaction we report suggests that syllable frequency and stress assignment may affect a common locus, and that such locus is the phonological output buffer, where the phonological word is phonetically encoded and thus realized. One might argue - contra Levelt at al. (1999) - that syllable frequency may affect reading during the orthography-to-phonology conversion. If that were the case, the syllable frequency effect would have emerged only in the congruent stress condition; in the incongruent stress condition, in fact, the time needed to account for the stress mismatch would have delayed the assembling of segmental and metrical information, with the result of allowing enough time for fully computing low-frequency syllables. This being the case, the syllable frequency effect would have been greatly reduced or even annulled. This is not the case, and our results support the proposal that syllable frequency effect arises at the phonetic encoding (Levelt et al., 1999).

The difference in speed of processing between high- and low-frequency first syllables for congruent and incongruent stress targets seems to be the critical factor in the pattern we obtained and the interaction suggests that, at the level of phonology-to-phonetic interface, words with a highfrequency initial syllable are less prone to interference from the stress mismatch. Although the present data do not allow us to finely specify the nature of such interaction - that is, it is hard to establish whether the nature of the stress priming effect is facilitatory or an inhibitory - a possible interpretation of our finding can still be sketched by referring to the different procedures for syllabification of high- and low-frequency syllables. According to the mental syllabary theory (Levelt et al.'s, 1999), the former are retrieved from the repertoire of syllables while for the latter a composition from their constituent phonemes is postulated. The syllable stored in the repertoire are used to
drive the motor programs, that is they allow the speakers to map the abstract phonological syllabic representations into phonetic packages, which are still partially abstract representations of the to-be-performed articulatory gestures, and each syllable can thus be still prone to be articulated in different ways (e.g., with longer or shorter duration, with more or less force, or with different kinds of pitch modulation; Levelt, 1989). Therefore, in case of a stimulus with a high-frequency first syllable, the reading system may start the phonological-to-phonetic mapping by processing the segmental information up to the syllable repertoire and independently of how fast the computation of the suprasegmental information occurs; then, as soon as the stress system determines the correct stress pattern the activated phonetic syllable is specified in terms of stress. In such a view, the phonetic code retrieval of a high-frequency syllable is weakly affected by the prime computation as the former can proceed independently from the latter. Thus, for words starting with a high-frequency syllable, the difference between targets in the congruent- and incongruent-stress prime condition would be only due to the different timing required for the specification of the correct stress pattern of the target in the two conditions.

A different process, however, can be postulated for words starting with low-frequency syllables as they do not have a stored representation in the syllabary and are thus assembled on-line. As a consequence, to map the abstract phonological constituents of a syllabic unit into a corresponding phonetic-detailed representation, the reading system needs all the relevant information - the phonemes and the specification of stress (i.e., if the syllable is either stressed or unstressed) - to be both in the phonological output buffer, as a partial or incomplete activation of either segmental or suprasegmental information would make the buffer unable to assemble a well-formed phonetic unit. In such a view, the large priming effect reported for lowfrequency syllables may emerge because, for such stimuli, the operations of stress assignment and phonetic syllable computation are sequential. The implication of such assumption is that the time required to assemble a lowfrequency syllable is a function of the time required to correctly assign the stress pattern to the word, as the latter can speed up or delay the initiation of phonetic encoding of the former.

The CDP++ model of reading (Perry et al., 2010), which was recently implemented for English bisyllables, explicitly assumes that the start of articulation is conditional to stress retrieval, and thus may be used to frame our interpretation of the pattern of results. In the Perry et al.'s (2010) model, the phonological output buffer includes two distinct mechanisms for segmental and suprasegmental computation, i.e., phonological output nodes and stress output nodes, with the latter nodes being responsible for stress assignment. This is consistent with our claim that the locus of the interaction is the phonological output buffer. However, the functional architecture of the model seems to be still underspecified to be able to fully account for our
results since in the model the timing of the operations in the phonological output buffer is such that only after the relevant stress pattern has been activated the word constituent phonemes, structured in their syllabic constituents, can be overtly articulated. Such architecture would predict an additive effect of stress assignment and syllabification, with the consequence that a delay in the processing of stress assignment should equally affect both word with a high-frequency first syllable and words with a low-frequency first syllable. Although our data support the view that stress assignment is essential for articulation to take place, they also suggest different procedures for highand low-frequency syllables, i.e. an interactive process at the level of phonology-to-phonetic interface (for a similar proposal, see Perret, Schneider, Dayer, \& Laganaro, 2012).

To conclude, our findings show that words with an initial low-frequency syllable are more strongly affected by manipulation of incongruent stress priming than words with a high-frequency initial syllable. This is the first evidence showing that, in word reading, the processes of stress assignment and syllable computation may interact within the phonological output buffer. The finding is consistent with the view that the phonological buffer acts as the locus of phonological-to-phonetics interface, where the abstract phonological word is converted into its phonetic representation, and where stress and syllable information may interact.

## References

Arciuli, J., Monaghan, P., \& Seva, N. (2010). Learning to assign lexical stress during reading aloud: Corpus, behavioral, and computational investigations. Journal of Memory and Language, 63, 180-196.
Bertinetto, P. M., Burani, C., Laudanna, A., Marconi, L., Ratti, D., Rolando, C., \& Thorton, A. M. (2005). Corpus e Lessico di frequenza dell'italiano scritto (CoLFIS). [CoLFIS. Corpus and Frequency Lexicon of Written Italian],http://www.istc.cnr.it/material/database/colfis/ind ex_eng.shtml.
Burani, C., \& Arduino, L. S. (2004). Stress regularity or consistency? Reading aloud Italian polysyllables with different stress patterns. Brain and Language, 90, 318-25.
Carreiras, M., Mechelli, A., \& Price, C. J. (2006). Effect of Word and Syllable Frequency on Activation During Lexical Decision and Reading Aloud. Human Brain Mapping, 27, 963-972.
Carreiras, M., Perea, M. (2004). Naming pseudowords in Spanish: Effects of syllable frequency. Brain and Language, 90, 393-400.
Cholin, J., Dell, G. S., \& Levelt, W. J. M. (2011). Planning and Articulation in Incremental Word Production: Syllable-Frequency effects in English. Journal of Experimental Psychology: Learning, Memory and Cognition, 37, 109-122.
Cholin, J., Levelt, W. J. M., \& Schiller, N. O. (2006). Effects of syllable frequency in speech production. Cognition, 99, 205-235.

Colombo, L., \& Zevin, J. (2009). Stress priming in reading and the selective modulation of lexical and sub-lexical pathways. PLoS ONE, 4, e7219.
Coltheart, M., Rastle, K., Perry, C., Langdon, R., \& Ziegler, J. (2001). DRC: A Dual Route Cascaded model of visual word recognition and reading aloud. Psychological Review, 108, 204-256.
Ferrand, L., Segui, J., \& Grainger, J. (1996). Masked priming of words and picture naming: The role of syllabic units. Journal of Memory \& Language, 35, 708-723.
Laganaro, M., Alario, F. X. (2006). On the locus of the syllable frequency effect in speech production. Journal of Memory and Language, 55, 178-196.
Levelt, W. J. M., Roelofs, A., \& Meyer, A. S. (1999). A theory of lexical access in speech production. Behavioral and Brain Sciences, 22, 1-75.
Paizi, D., Zoccolotti, P., \& Burani, C. (2011). Lexical stress assignment in Italian development dyslexia. Reading and Writing: An Interdisciplinary Journal, 24, 443-461.
Perret, C., Schneider, L., Dayer, G., \& Laganaro, M. (2012): Convergences and divergences between neurolinguistic and psycholinguistic data in the study of phonological and phonetic encoding: A parallel investigation of syllable frequency effects in brain-damaged and healthy speakers. Language and Cognitive Processes.
Perry, C., Ziegler, J. C., \& Zorzi, M. (2010). Beyond single syllables: Large-scale modeling of reading aloud with the Connectionist Dual Process (CDP++) model. Cognitive Psychology, 61, 106-151.
Rastle, K., \& Coltheart, M. (2000). Lexical and Nonlexical Print-to-Sound Translation of Disyllabic Words and Nonwords. Journal of Memory and Language, 42, 342364.

Roelofs, A. (2004). Seriality of phonological encoding in naming objects and reading their names. Memory \& Cognition, 32, 212-222.
Roelofs, A., \& Meyer, A. S. (1998). Metrical structure in planning the production of spoken words. Journal of Experimental Psychology: Learning, Memory and Cognition, 24, 922-939.
Sibley, D. E., Kello, C. T., Seidenberg, M. S. (2010). Learning orthographic and phonological representations in models of monosyllabic and bisyllabic naming. Journal of Cognitive Psychology, 22, 650-668.
Stella, V., \& Job, R. (2001). Le sillabe PD/DPSS. Una base di dati sulla frequenza sillabica dell'italiano scritto [The syllables PD/DPSS. A database for syllable frequency of written Italian]. Giornale Italiano di Psicologia, 3, 633642, http://dpss.psy.unipd.it/files/strumenti.php.
Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. Acta Psychologica, 30, 267-315.

Sulpizio, S., \& Job, R. (2010). L'intervento delle sillabe nei processi di produzione e riconoscimento visivo delle parole [The role of syllable in word production and visual word recognition]. Giornale Italiano di Psicologia, 3, 731-738.
Sulpizio, S., Boureux, M., Burani, C., Deguchi, C., \& Colombo, L. (2012a). Stress assignment in the development of reading aloud: Nonword priming effects on Italian children. In N. Miyake, D. Peebles, \& R. P. Cooper (Eds.), Proceedings of the 34th Annual Conference of the Cognitive Science Society (pp. 23692374). Austin, TX: Cognitive Science Society.

Sulpizio, S., Job, R., \& Burani, C. (2012b). Priming lexical stress in reading Italian aloud. Language \& Cognitive Processes, 27, 808-820.
Thornton, A. M., Iacobini, C., \& Burani, C. (1997). BDVBD. Una base di dati sul vocabolario di base della lingua italiana [BDVDB: A database for the Italian basic dictionary]. Roma, Bulzoni.

# How Does Prospect Theory Reflect Heuristics' Probability Sensitivity in Risky Choice? 

Renata S. Suter (suter@mpib-berlin.mpg.de)<br>Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany

Thorsten Pachur (pachur@mpib-berlin.mpg.de)<br>Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany

Ralph Hertwig (hertwig@mpib-berlin.mpg.de)
Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany


#### Abstract

Two prominent approaches to describing how people make decisions between risky options are algebraic models and heuristics. The two approaches are based on fundamentally different algorithms and are thus usually treated as antithetical, suggesting that they may be incommensurable. Using cumulative prospect theory (CPT; Tversky \& Kahneman, 1992) as an illustrative case of an algebraic model, we demonstrate how algebraic models and heuristics can mutually inform each other. Specifically, we highlight that CPT describes decisions in terms of psychophysical characteristics, such as diminishing sensitivity to probabilities, and we show that this holds even when the underlying process is heuristic in nature. Our results suggest that algebraic models and heuristics might offer complementary rather than rival modeling frameworks and highlight the potential role of heuristic principles in information processing for prominent descriptive constructs in risky choice.


Keywords: cumulative prospect theory; probability sensitivity; computational modeling; heuristics; risky choice.

## Introduction

How can risky decision making-in which people have to choose between options offering different outcomes with certain probabilities-best be modeled? Two prominent approaches in decision research are algebraic models and heuristics (e.g., Brandstätter, Gigerenzer, \& Hertwig, 2006; Payne, 1973; Payne, Bettman, \& Johnson, 1993). Algebraic models follow the principle of expectation maximization and use an algorithm that integrates (some function of) probability and outcome information multiplicatively to describe people's risky choices. Arguably the most prominent model in this tradition is cumulative prospect theory (CPT; Tversky \& Kahneman, 1992). According to CPT, options are evaluated independently of each other. The model invokes psychophysical constructs such as probability sensitivity and loss aversion to account for characteristic phenomena in choice, and quantifies them using adjustable parameters. Heuristics, by contrast, are based on simple principles of information processing, such as
sequential and limited search, dimensional comparison, and aspiration levels; in contrast to algebraic models, heuristics often go without integrating information, and ignore part of the information (e.g., Payne et al., 1993; Thorngate, 1980). Models of heuristics for risky choice include the semiorder rule (Luce, 1956), the similarity heuristic (Leland, 1994; Rubinstein, 1988), elimination-by-aspects (Tversky, 1972), and the priority heuristic (Brandstätter et al., 2006).

Algebraic models and heuristics are often treated as antithetical (e.g., Brandstätter et al., 2006; Payne, 1973; Svenson, 1979). As pointed out by Lopes (1995), however, this opposition may be unnecessary: "Some models focus on the algebraic pattern of people's risk preferences, others on the content of their choice processes [models of heuristics]. Although one might suppose that these two kinds of accounts are alternate ways of describing the same thing-indeed, that one kind of model might eventually be reducible to the other-the approaches have tended to be disjoint" (p. 177). To date, however, the relationship between algebraic models and heuristics has yet to be elaborated.

To close that gap, we use CPT (Tversky \& Kahneman, 1992) as an illustrative case highlighting that algebraic models offer a tool for describing characteristics of the decision process in psychophysical terms; here, we focus on the sensitivity to differences in probabilities. We argue that diminished sensitivity to probability information-as captured in CPT's weighting functions-can result from lexicographic and noncompensatory processing of heuristics. As such, CPT may offer a useful framework to represent heuristic decision making in terms of established constructs such as sensitivity to probability information. Conversely, as heuristics are explicit with regard to the informationprocessing steps underlying a decision, elaborating the relationship between heuristics and CPT might contribute to a better understanding of the cognitive mechanisms potentially underlying the characteristic shapes of CPT's weighting and value functions (cf. Hogarth \& Einhorn, 1990). Overall, we thus suggest that the relationship between the algebraic and heuristic models of risky choice is complementary rather than adversarial.

In the following, we first briefly describe CPT's parametric framework and how its weighting functions reflect sensitivity to probability information. Second, we elaborate for one specific heuristic, the priority heuristic (Brandstätter et al., 2006), how heuristic choices may be reflected in CPT's parameters. Specifically, we take advantage of the fact that the degree to which the priority heuristic attends to probability information depends on the choice environment; using computer simulations, we examine how this translates into differences in CPT's weighting function.

## Probability Sensitivity in CPT

CPT assumes that decisions are made to maximize expected return. More specifically, choices between risky options are based on a person's subjective valuation of these options and then maximization. In CPT, the overall valuation $V$ of an option $A$ is defined as
$V(A)=\sum_{j=1}^{k} v\left(x_{j}\right) \pi_{j}^{-}+\sum_{i=k+1}^{n} v\left(x_{i}\right) \pi_{i}^{+}$.
$v(x)$ is the value function, describing how an objective outcome $x$ is translated into a subjective value, and $\pi^{+}$ $\left(\pi^{-}\right)$is the weight given to a positive (negative) outcome $x$ (Tversky \& Kahneman, 1992) and depends on the probability of the outcome.

CPT assumes a rank-dependent transformation of the outcomes' probabilities into decision weights. Specifically, with outcomes $x_{I} \leq \ldots \leq x_{k} \leq 0 \leq x_{k+1} \leq \ldots \leq$ $x_{n}$, the weight $\pi^{+}\left(\pi^{-}\right)$given to a positive (negative) outcome $x$ is the difference between the probability of receiving an outcome at least as good (bad) as $x$ and the probability of receiving an outcome better (worse) than $x$ :
$\pi_{i}^{+}=w^{+}\left(p_{i}+\ldots+p_{n}\right)-w^{+}\left(p_{i+1}+\ldots+p_{n}\right)$ for $k<i<n$
$\pi_{j}^{-}=w-\left(p_{1}+\ldots+p_{j}\right)-w-\left(p_{1}+\ldots+p_{j-1}\right)$ for $1<j \leq k$.
$w^{+}$and $w^{-}$are the probability weighting functions for gains and losses, respectively. They are assumed to have an inverse S-shaped curvature. Different types of weighting functions have been proposed (for an overview, see Stott, 2006). We use the following twoparameter version that separates the curvature of the weighting function from its elevation (e.g., Goldstein \& Einhorn, 1987; Gonzalez \& Wu, 1999):
$w^{+}(p)=\frac{\delta^{+} p^{\gamma^{+}}}{\delta^{+} p^{\gamma^{+}}+(1-p)^{\gamma^{+}}}$.
$w^{-}(p)=\frac{\delta^{-} p^{\gamma^{-}}}{\delta^{-} p^{\gamma^{-}}+(1-p)^{\gamma^{-}}}$
The parameters $\gamma^{+}$and $\gamma^{-}$(both varying between 0 and 1) govern the amount of curvature of the function in the gain and loss domains, respectively, and indicate how sensitive decisions are to differences in probabilities (with smaller values of $\gamma<1$ resulting in more S-shaped weighting functions, reflecting lower sensitivity to
differences in probabilities). The elevation of the weighting functions for gains and losses is controlled by the parameters $\delta^{+}$and $\delta^{-}$(both $>0$ ), respectively.

CPT has repeatedly been shown to be highly successful in describing risky choices between monetary gamble problems (e.g., Glöckner \& Pachur, 2012; but see Birnbaum, 2004; Brandstätter et al., 2006). As a description of the underlying cognitive process, however, CPT's implied algebraic calculus and its commitment to a multiplicative framework have not been unchallenged (e.g., Brandstätter et al., 2006; Lopes, 1995). One such challenge has been put forth by proponents of heuristics. We turn to this modeling approach next.

## Probability Sensitivity Resulting From Heuristic Information Processing

In contrast to the integrative approach of CPT, heuristics often ignore part of the information and do not integrate information. They are based on simple principles of information processing, such as sequential and limited search, dimensional comparison, and aspiration levels (e.g., Payne et al., 1993; Thorngate, 1980). Lexicographic strategies, for instance, proceed through several dimensions sequentially and stop at the first dimension that enables a decision to be made (Fishburn, 1974; Gigerenzer, Todd, \& the ABC Research Group, 1999; Thorngate, 1980). The priority heuristic (Brandstätter et al., 2006), which is related to lexicographic semi-orders (Luce, 1956; Tversky, 1969), belongs to this class. Its architecture is based on established principles of bounded rationality (e.g., Simon, 1955), such as sequential search, stopping rules, and aspiration levels, and it assumes that probabilities and outcomes are compared between gambles, rather than integrated within gambles (as assumed by CPT). For choices between two-outcome gambles involving gains, the priority heuristic entails the following steps:

1. Priority rule. Go through dimensions in the order of minimum gain, probability of minimum gain, and maximum gain.
2. Stopping rule. Stop examination if the minimum gains differ by $1 / 10$ (or more) of the maximum gain; otherwise, stop examination if probabilities differ by $1 / 10$ (or more) of the probability scale.
3. Decision rule. Choose the gamble with the more attractive gain (probability).
(For losses, "gains" are replaced by "losses"; for mixed gambles, "gains" are replaced by "outcomes.")

Due to its stopping rule, the priority heuristic considers probability information depending on the structure of a gamble problem. The heuristic first examines the (minimum) outcomes of the gambles. If this reason discriminates, then probabilities will not be examined. If, however, the outcomes fail to discriminate, probabilities will be examined. That is, the priority heuristic attends to probability information only when the minimum outcomes do not differ. The heuristic's
sensitivity to probability information is thus dependent on the structure of the choice environment.

## Heuristics' Probability Sensitivity as Captured in CPT's Parametric Framework

These two approaches to model risky choice-CPT and heuristics-are based on fundamentally different algorithms. Whereas CPT considers all outcome and all probability information, the priority heuristic considers the reasons sequentially, and stops information search as soon as a reason discriminates. Moreover, although CPT may be a relatively flexible model due to its several adjustable parameters (e.g., Gonzalez \& Wu, 1999), it still has important constraints: both the value and the weighting function are restricted to be monotonic, the value function is concave for gains and convex for losses, and the weighting function is constrained to have an inverse S-shaped curvature. Can CPT, given these constraints and given its starkly different algorithmic structure, nevertheless accommodate choices generated by the priority heuristic and accurately reflect the degree to which the heuristic attends to probability information?

In addressing this question, we strive to contribute to a better understanding of the relationship between algebraic models and heuristics. One crucial aspect of our argument is that diminished sensitivity to probability information may be due not only to psychophysical regularities in magnitude evaluation, but also to the limited attention that a heuristic devotes to probabilities. More specifically, the weighting function's $\gamma$ parameter (Equation 3), which reflects sensitivity to probabilities, should differ systematically as a function of whether the heuristic makes a choice based on the first reason (outcome) or the second reason (probability). The less frequently the priority heuristic considers probabilities in a set of gamble problems, the lower the resulting value of the $\gamma$ parameter should be. Slovic and Lichtenstein (1968) made a similar proposal more than 40 years ago, suggesting that "increases [in] the saliency of the money dimensions and decreases [in] the relative importance of the probabilities" should lead to "relatively flat [i.e., more strongly S-shaped] subjective probability functions" (p. 16). Next, we test this suggestion using a computer simulation.

## Computer Simulation

We created three sets of two-outcome gamble problems, each including 180 randomly generated problems with similar expected values: 60 gain, 60 loss, and 60 mixed problems (cf. Rieskamp, 2008). Across the three sets, we varied the percentage of problems in which the minimum gains (losses) discriminated between the gambles (i.e., that differed by at least $10 \%$ of the highest gain or loss). In the first set, the minimum gains (losses) discriminated in $75 \%$ of the cases, and the priority heuristic therefore only proceeded to the second reason-the probability of the minimum gains (losses)-in $25 \%$ of the cases; in the
second set, the minimum gains (losses) discriminated in $50 \%$ of the gamble problems, and the heuristic therefore proceeded to the probability information in the remaining $50 \%$ of the cases; in the third set, the minimum gains (losses) differed in only $25 \%$ of the cases, and the heuristic therefore proceeded to the probabilities in $75 \%$ of the cases. The gambles were constructed such that if the heuristic proceeded to the probability information, this reason always discriminated. We predicted that CPT's probability sensitivity parameter $\gamma$ fitted to the priority heuristic's choices would increase across the problem sets.

We simulated the choices of the priority heuristic in all three problem sets and subsequently fitted CPT's weighting functions and value functions, respectively, to these choices, separately for each set. Our implementation of CPT had six adjustable parameters: $\alpha$ $(=\beta)$ and $\lambda$ for the value function, $\gamma, \delta^{+}$and $\delta^{-}$for the weighting function, and $\varphi$ for the choice rule necessary to derive predicted choice probabilities (see below). ${ }^{1}$ To reflect CPT's main assumptions (e.g., an inversely Sshaped probability weighting function, a concave value function for gains, and a convex value function for losses; see Tversky \& Kahneman, 1992), in the parameter estimation procedure the parameter values were restricted as follows (see Rieskamp, 2008): $0<\alpha \leq 1 ; 0<\lambda \leq 5 ; 0<$ $\gamma \leq 1 ; 0<\delta^{ \pm} \leq 4 ; 0<\varphi \leq 5$. The deviation between CPT's predictions and the heuristic's choices was quantified using the likelihood measure $G^{2}$ (e.g., Sokal \& Rohlf, 1994), with a smaller $G^{2}$ indicating a better fit:

$$
\begin{equation*}
G^{2}=-2 \sum_{i=1}^{N} \ln \left[f_{i}(y \mid \theta)\right] \tag{4}
\end{equation*}
$$

where $N$ refers to the total number of choices, and $f(\mathrm{y} \mid \theta)$ refers to the probability with which CPT, given a particular set of parameter values $\theta$, predicts an individual choice $y$. If gamble A was chosen, then $f(\mathrm{y} \mid \theta)$ $=p_{i}(\mathrm{~A}, \mathrm{~B})$, where $p_{i}(\mathrm{~A}, \mathrm{~B})$ is the predicted probability that gamble A is chosen over gamble B ; if gamble B was chosen, then $f(\mathrm{y} \mid \theta)=1-p_{i}(\mathrm{~A}, \mathrm{~B})$. To determine $p_{i}(\mathrm{~A}, \mathrm{~B})$, we applied an exponential version of Luce's (1956) choice rule (also known as softmax):

$$
\begin{equation*}
p_{i}(A, B)=\frac{e^{\varphi V(A)}}{e^{\varphi V(A)}+e^{\varphi V(B)}}, \tag{5}
\end{equation*}
$$

where $V(\mathrm{~A})$ and $V(\mathrm{~B})$ represent the subjective valuation of the gambles A and B according to CPT, and $\varphi>0$ specifies how sensitively the predicted choice probability reacts to differences between the gambles' subjective valuations $V(\mathrm{~A})$ and $V(\mathrm{~B})$, with higher values indicating higher sensitivity. In the fitting procedure, we first

[^222]Table 1: Parameter estimates obtained when fitting CPT to the choices of the priority heuristic where the decision was made on the first reason (minimum gain/loss) in $25 \%, 50 \%$, or $75 \%$ of the cases, respectively, and on the second reason (probability of the minimum gain/loss) otherwise.

| \% <br> Decisions <br> on first <br> reason | $\gamma$ | $\delta^{+}$ | $\delta^{-}$ | $\alpha$ | $\lambda$ | $\varphi$ | $G^{2}$ |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $75 \%$ | 0.32 | 0.01 | 0.15 | 1.0 | 2.06 | 0.34 | 84.25 |
| $50 \%$ | 0.50 | 0.05 | 0.49 | 1.0 | 1.30 | 0.15 | 153.45 |
| $25 \%$ | 1.00 | 0.13 | 0.45 | 1.0 | 1.18 | 0.21 | 108.54 |

Note. $G^{2}$ assuming random choice is 249.53 .
implemented a grid search to identify the parameter values that minimize $G^{2}$; the 20 best-fitting combinations of grid values were then used as starting points for subsequent optimization using the simplex method (Nelder \& Mead, 1965), as implemented in MATLAB.

## Results

Table 1 shows the best-fitting CPT parameters when fitted to the simulated choices of the priority heuristic in gamble problems where the decision was made on the minimum gain (loss) in $75 \%, 50 \%$, or $25 \%$ of the cases, respectively, and on the probability of the minimum gain (loss) otherwise. As can be seen-and as predicted-the probability sensitivity parameter $\gamma$ increased across the sets; it was lowest in the set where the priority heuristic decided on the first reason (minimum outcome) in the majority of cases and it was highest in the set where the heuristic decided on the second reason (probability of the minimum outcome) in the majority of cases. In other
words, CPT accurately reflected the different degrees to which the priority heuristic attended to probability information across the three sets.

Panels A and B of Figure 1 plot the weighting functions based on the best-fitting parameters, separately for the gain and loss domains. Irrespective of domain, for choices that only considered probabilities in $25 \%$ of the cases, the weighting function was most strongly Sshaped, indicating low sensitivity to probability information; for choices that considered probabilities in half of the cases, it was comparatively less S-shaped; and for choices that considered probabilities in $75 \%$ of the cases, it was least S-shaped. (Note that the differences in shapes of the weighting function between the gain and loss domains were due to differences in the elevation parameters; i.e., $\delta^{+}$and $\delta^{-}$).

The best-fitting parameter values of the priority heuristic's choices in the three problem sets are summarized in the parameter profiles in Panel C in Figure 1.

Interestingly, CPT did not fit equally well to the choices across the three problem sets (see $G^{2}$ in Table 1). Specifically, the fit was best when most choices (75\%) were made on the first reason, worsened when $50 \%$ considered probabilities, and improved again when only $25 \%$ of the choices were made on the first reason. CPT is thus apparently better able to fit choice sets where a substantial proportion of choices stop examination on the same reason than when choices are based on different reasons (as in the $50 \%$ choice set).

## Discussion

CPT models decisions based on a compensatory algorithm where outcomes and probabilities are integrated multiplicatively and summed up separately within each option. The priority heuristic, in contrast, models decisions based on sequential information


Figure 1: Panels A (gains) and B (losses) plot the weighting functions obtained when fitting cumulative prospect theory to the choices of the priority heuristic in gambles where the decision was made on the minimum gain (loss) in $75 \%$, $50 \%$, or $25 \%$, respectively, and on the probability of the minimum gain (loss) otherwise. Panel C shows the parameter profiles of CPT's value and weighting function parameters fitted to the choices of the priority heuristic in the three gamble sets (as the parameters differed in their scale, they were normalized for this graph; for the exact parameter values, see Table 1).
processing and compares outcomes and probabilities between the options. Despite these stark differences, our result is that CPT is able to represent choices generated by the priority heuristic in a psychologically meaningful manner: the weighting function's curvature reflects differences in the heuristic's sensitivity to probability information between the three choice environments that differed in how frequently choices were decided based on the probability dimension.

Taken together, our results thus demonstrate that although CPT is based on a different algorithmic architecture than heuristics, its parametric framework might offer a useful tool for characterizing heuristic processes in terms of prominent descriptive constructs such as probability sensitivity (for a discussion of other constructs of CPT , such as risk aversion, loss aversion, and outcome sensitivity, see Suter, Pachur, \& Hertwig, 2013a). Conversely, the integration of the two approaches might enable hypotheses to be derived about the processes generating the characteristic shapes of CPT's functions.

Our finding that specific values of CPT's $\gamma$ parameter can reflect the processing steps of a lexicographic heuristic-that is, whether probability information was called upon or not-has important implications for the use of CPT in empirical investigations of risky choice: CPT's parameters might help to identify the interaction of a heuristic with its environment; moreover, they might help to identify the use of different heuristics by different individuals within the same environment, or of different heuristics by the same individual across different environments.

The demonstrated relationship between CPT, the information processing architecture of a heuristic, and the structure of the environment could explain apparent inconsistencies in empirical investigations (see also Hertwig \& Gigerenzer, 2011)-for instance, why the same person's sensitivity to probability information appears low at some times and high at others. Such observations of variability need not mean that CPT's parameters cannot be measured reliably, or that different heuristics are used. They could arise from the interaction of a heuristic's lexicographic architecture with various choice environments. If decision problems are constructed such that a user of the priority heuristic is always able to terminate search after examining the options' minimum outcomes, the person's probability sensitivity will appear low. If they are constructed such that the same person must always move beyond the minimum outcomes and examine their probabilities, the person will seem to be highly sensitive to probabilities.

Relatedly, the elaborated relationship between CPT and heuristic processing not only allows the interactions of a heuristic to be tracked across different environments, but it may also allow differences in strategy selection between individuals within the same environment to be identified. It therefore suggests an alternative interpretation of the observed link between CPT's
parameters and variables that influence risky choice, such as gender. Fehr-Duda, De Gennaro, and Schubert (2006), for instance, concluded that women tend to be less sensitive to probability changes than men (see also Booij \& van de Kuilen, 2009). To the extent that CPT reflects differences in terms of probability sensitivity also between strategies, this finding might indicate that men and women rely on different strategies that differ with regard to their probability sensitivity.

Moreover, CPT's parameters might not only reveal differences between individuals, but also within an individual. For instance, a decision maker might use different strategies for different contexts. In a study on the difference between affect-rich and affect-poor risky choice, Suter, Pachur, and Hertwig (2013b) found that people's choices in affect-rich tasks were consistent with a more strongly inverse $S$-shaped weighting function relative to choices in affect-poor tasks. However, in a model comparison, they found that in affect-rich choices the majority of the participants were better described by the minimax heuristic, a choice strategy that neglects probabilities and only decides based on the minimum outcomes, than by CPT; in affect-poor tasks, in contrast, participants were better described by a strategy that is sensitive to probabilities. Thus, the differences apparent on the weighting function could indicate the selection of a different strategy. Similarly, Abdellaoui, Diecidue, and Öncüler (2011) reported that, relative to lotteries with immediate outcomes, people's responses to lotteries with delayed outcomes are consistent with a less inverse Sshaped curvature (indicating higher probability sensitivity). The authors hypothesized that this difference might be due to a decreased anticipated emotional reaction the more delayed lotteries are (cf. Rottenstreich \& Hsee, 2001). Again, the impact observed on the weighting function might thus reflect the use of different strategies.

Thus, rather than merely describing contextual or individual differences in prospect theory's concepts, such as differences in probability sensitivity, one could go one step further and use differences on CPT's parameters to hypothesize about individual differences in terms of information processing and strategy use. By better understanding how information processing as embodied in heuristics manifests in CPT's parameters, we can gain a more cognitive perspective on CPT and its parametric framework (for an ecological account of the shape of CPT's functions, see Stewart, Chater, \& Brown, 2006).

## Acknowledgments

This research was supported by Swiss National Science Foundation grants to Thorsten Pachur (100014 125047/2) and Ralph Hertwig (100014 126558/1), and by a grant from the German Research Foundation (DFG) as part of the priority program "New Frameworks of Rationality" (SPP 1516) to Ralph Hertwig and Thorsten Pachur (HE 2768/7-1).

## References

Abdellaoui, M., Decidue, E., \& Öncüler, A. (2011). Risk preferences at different time periods: An experimental investigation. Management Science, 57, 975-987.
Birnbaum, M. H. (2004). Tests of rank-dependent utility and cumulative prospect theory in gambles represented by natural frequencies: Effects of format, event framing, and branch splitting. Organizational Behavior and Human Decision Processes, 95, 40-65.
Booij, A. S., \& van de Kuilen, G. (2009). A parameterfree analysis of the utility of money for the general population under prospect theory. Journal of Economic Psychology, 30, 651-666.
Brandstätter, E., Gigerenzer, G., \& Hertwig, R. (2006). The priority heuristic: Making choices without tradeoffs. Psychological Review, 113, 409-432.
Fehr-Duda, H., De Gennaro, M., \& Schubert, R. (2006). Gender, financial risk, and probability weights. Theory and Decision, 60, 283-313.
Fishburn, P. C. (1974). Lexicographic orders, utilities and decision rules: A survey. Management Science, 20, 1442-1471.
Gigerenzer, G., Todd, P. M., \& the ABC Research Group. (1999). Simple heuristics that make us smart. New York, NY: Oxford University Press.
Glöckner, A., \& Pachur, T. (2012). Cognitive models of risky choice: Parameter stability and predictive accuracy of prospect theory. Cognition, 123, 21-32.
Goldstein, W. M., \& Einhorn, H. J. (1987). Expression theory and the preference reversal phenomena. Psychological Review, 94, 236-254.
Gonzalez, R., \& Wu, G. (1999). On the shape of the probability weighting function. Cognitive Psychology, 38, 129-166.
Hertwig, R., \& Gigerenzer, G. (2011). Behavioral inconsistencies do not imply inconsistent strategies. Frontiers in Cognition, 2:292.
Hogarth, R. M., \& Einhorn, H. J. (1990). Venture theory: A model of decision weights. Management Science, 36, 780-803.
Leland, J. W. (1994). Generalized similarity judgments: An alternative explanation for choice anomalies. Journal of Risk and Uncertainty, 9, 151-172.
Lopes, L. L. (1995). Algebra and the process in the modeling of risky choice. The Psychology of Learning and Motivation, 32, 177-220.
Luce, R. D. (1956). Semiorders and a theory of utility discrimination. Econometrica, 24, 178-191.
Nelder, J. A., \& Mead, R. (1965). A simplex method for function minimization. Computer Journal, 7, 308-313.
Nilsson, H., Rieskamp, J., \& Wagenmakers, E. (2011). Hierarchical Bayesian parameter estimation for cumulative prospect theory. Journal of Mathematical Psychology, 55, 84-93.
Payne, J. W. (1973). Alternative approaches to decision making under risk: Moments versus risk dimensions. Psychological Bulletin, 80, 439-453.
Payne, J. W., Bettman, J. R., \& Johnson, E. J. (1993).

The adaptive decision maker. Cambridge, UK: Cambridge University Press.
Rieskamp, J. (2008). The probabilistic nature of preferential choice. Journal of Experimental Psychology: Learning, Memory, and Cognition, 34, 1446-1465.
Rottenstreich, Y., \& Hsee, C. K. (2001). Money, kisses, and electric shocks: An affective psychology of risk. Psychological Science, 12, 185-190.
Rubinstein, A. (1988). Similarity and decision-making under risk (Is there a utility theory resolution to the Allais-paradox?). Journal of Economic Theory, 46, 145-153.
Simon, H. A. (1955). A behavioral model of rational choice. Quarterly Journal of Economics, 69, 99-118.
Slovic, P., \& Lichtenstein, S. (1968). Relative importance of probabilities and payoffs in risk taking. Journal of Experimental Psychology. Monograph, 78, 1-18.
Sokal, R. R., \& Rohlf, F. J. (1994). Biometry: The principles and practice of statistics in biological research. New York, NY: Freeman.
Stewart, N., Chater, N., \& Brown, G. D. A. (2006). Decision by sampling. Cognitive Psychology, 53, 1-26.
Stott, H. P. (2006). Cumulative prospect theory's functional menagerie. Journal of Risk and Uncertainty, 32, 101-130.
Suter, R. S., Pachur, T., \& Hertwig, R. (2013a). Heuristics of risky choice and cumulative prospect theory: Toward a theory integration. Manuscript submitted for publication.
Suter, R. S., Pachur, T., \& Hertwig, R. (2013b). How to model the impact of affect on risky choice: Distorted probability weighting or probability neglect? Manuscript submitted for publication.
Svenson, O. (1979). Process descriptions of decision making. Organizational Behaviour and Human Performance, 23, 86-112.
Thorngate, W. (1980). Efficient decision heuristics. Behavioral Science, 25, 219-225.
Tversky, A. (1969). Intransitivity of preferences. Psychological Review, 76, 31-48.
Tversky, A. (1972). Elimination by aspects: A theory of choice. Psychological Review, 79, 281-299.
Tversky, A., \& Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and Uncertainty, 5, 297323.

Wakker, P. P. (2010). Prospect theory for risk and ambiguity. Cambridge, UK: Cambridge University Press.

# Scope of Real Beliefs in Belief Revision 

Alexander B. Swan (swan@psych.ucsb.edu) Alexandra Y. Chambers (chambers@psych.ucsb.edu)<br>Russell Revlin (revlin@psych.ucsb.edu)<br>Department of Psychological and Brain Sciences, University of California, Santa Barbara<br>Santa Barbara, CA 93106-9660 USA


#### Abstract

The present study examines the decisions made by reasoners when they are asked to revise their beliefs in the face of new, counterfactual information. Participants indicated the Scope (the degree of set inclusion) of semantic generalizations about real categories in a Pretest. In subsequent experiments, these Scope values were used to predict the willingness of participants to retain statements in their existing knowledge sets. When those sets were logically compatible with a Modus Tollens (MT) structure, participants were more likely to retain the general statements, but not when the sets were logically compatible with a Modus Ponens (MP) structure. However, the MP retention rates increased when locatives were added to the generalizations. These findings are inconsistent with several prevailing proposals of belief revision but do support the concept of belief revision as following Possible Worlds logic.


Keywords: belief revision; counterfactual reasoning; scope.

## Introduction

The process of belief revision is known by many terms (e.g., belief updating, belief change or belief dynamics). Simply put, belief revision involves the possibility of changing a previously held belief in light of new and assumed true information. At its very basic level, this involves updating knowledge and resolving inconsistencies within a pre-existing knowledge structure.

True belief revision must involve commitment to true beliefs. The earliest belief revision studies employed dictionary definitions without proper verification that those definitions had merit or that the students were committed to those definitions (Revlis \& Hayes, 1972). Later studies examined belief revision with artificial categories or groups, of which a participant might have had no previous knowledge (e.g., Byrne \& Walsh, 2002; Elio \& Pelletier, 1997; Politzer \& Carles, 2001; Revlin, Cate, \& Rouss, 2001). Revlin, Calvillo, and Ballard (2005) specifically created a fantasy world with Lego figures and various arbitrary rules about knights and kings (see also Van Hoeck, Revlin, Dieussaert, \& Schaeken, 2012). In each of these cases, real beliefs were not tested. It is difficult to assess the process of belief updating when the epistemic system is limited to arbitrary or unverified knowledge. The findings from such studies have supported conflicting theoretical treatments (e.g., Byrne \& Walsh, 2002; Wolf, Rieger, \& Knauff, 2012; Revlin et al., 2001; Revlis \& Hayes, 1972). The focus of the present study is to identify the basic
cognitive processes in true belief revision while still employing an established paradigm.

A basic paradigm for studying natural belief revision has been borrowed from the philosophical treatment of beliefcontravening problems (Rescher, 1964): It consists of a set of beliefs that are relevant to a counterfactual assumption, whose introduction requires a revision of the belief set. For example,
(1a) All whales are mammals
(1b) This creature is not a mammal
(1c) This creature is not a whale
(1d) Assume that this creature is a whale
A typical adult reasoner with the current knowledge of statements (1a-1c) would appreciate the inherent consistency of these statements. This collection of statements follows the logical form of Modus Tollens (MT; $p \rightarrow q, \sim q, \therefore \sim p)$. However, if someone is faced with statement (1d), an inconsistency is introduced to the knowledge structure and the revision process requires the reestablishment of a consistent epistemic set that entails the retention of (1d). To accomplish this, the reasoner notices that statement (1d) is in direct contradiction to statement (1c), which can be easily eliminated. However, there is a larger issue. The remaining statements ( $1 \mathrm{a} \& 1 \mathrm{~b}$ ) jointly create a contradiction with (1d). To resolve this inconsistency, a choice must now be made: Does the individual accept statement (1a), retaining the previously held belief that all whales are mammals, and that it cannot be true that there is an creature that is classified as a whale that is not a mammal? Or does the individual accept statement (1b), and claim that it could be true that there are creatures classified as whales that are not mammals? The revision process requires the elimination of one of the two statements. Either path is equally logical, although standard logic fails to encourage a preference, only indicating to the reasoner that there is an inconsistency (Chisholm, 1946). While an individual can reject both statements, the goal is generally to retain the maximum number of statements that already exist within the epistemic set. A second logical form typically used in belief-contravening problems is that of Modus Ponens (MP; $p \rightarrow q, p, \therefore q$ ), illustrated in (2) below:
(2a) All whales are mammals
(2b) This creature is a whale
(2c) This creature is a mammal
(2d) Assume that this creature is not a mammal

The assumption (2d) introduces the same direct and indirect contradictions as (1) above, and the reasoners must revise the belief-set to resolve the inconsistencies that are created. Notice that the MP assumption (2d) undermines the credibility of the generality (2a) by changing the properties of the specific instance (2b) that is already within a category-thereby making the category incoherent. In contrast, the MT assumption (1d) adds a new member to a category (1b) with seemingly different group membership or properties, while not changing the credibility of the generality as in the case of the MP assumption.

Despite the fact that standard logic is unable to guide the selection made by reasoners, they have shown distinct patterns of resolution for each of the logical forms, which vary with the content of the studies. In some studies, reasoners show a distinct preference to retain the generalities in problems like (1) above (e.g., Revlin et al., 2005; Revlin, Calvillo, \& Mautone, 2003). In other studies, preference has been shown for creating disabling conditions (or exceptions to the generality or rule), which allows them to be supported with caveats (Khemlani \& Johnson-Laird, 2011). In some studies, no preference among reasoners has been shown, especially for problems like (2) above (Byrne \& Walsh, 2002). We propose here that these differential findings may be a consequence of the degree of reality of the beliefs to be revised.

In addition to examining the role of real beliefs, the present study will focus on the importance of Scope in belief revision. The Scope of a quantified statement specifies the instances of the generality (or rule) that are subsumed within it across time and space. To demonstrate, consider the statements: all the coins in my pocket on VE Day are silver (Goodman, 1954) and all whales are mammals (Ryle, 1949). While these statements are both universally quantified generalities, the first is considered an accidental generality (it just so happens that all the coins in the pocket are silver), and the second is a scientific law (which spans space and time). A reasoner would prefer to retain the second statement because of its law-like quality; such statements are intended to act as inference tickets in new situations (Ryle, 1949). A reasoner should regard the Scope of the first statement about the silver coins to be quite small. In contrast, a reasoner should recognize that the second generality regarding mammalian whales has a large Scope (imagine all the whales that have existed in the past, present, and future and classify them into the superordinate category of mammals). The Scope of these relationships is generally an important proxy for knowledge preservation and credibility. It is possible that belief revision with artificial categories employed generalities with restricted Scope, which impacted revisions.

In the experiments in this study, Scope is either presented implicitly (with statements from the Pretest) or explicitly (the inclusion of numbers in the generality expression), and the goal is to determine whether Scope affects belief revision. To gain a sense of a statement's Scope at a granular level, we asked participants to indicate the number
of instances of a large category that possess a critical property-e.g., the number of whales that are mammals.

A pretest was used to assess participants' commitment to Scope values of general statements that were derived from theories of semantic knowledge (e.g., Collins \& Quillian, 1969; Quillian, 1968; Rips, 1989). In Experiment 1, these generalities were included in a basic belief revision paradigm where problems took the logical forms seen in example problems (1) and (2). In Experiment 2, locatives were added to the statements to constrain the law-like quality of the statements and to de-couple real world categories from the reasoning context. In Experiment 3, the Scope of the statements was explicitly manipulated into small and large proportions of a given set (Scope) to determine the effects of explicitly stated Scope on the reasoning process.

## Pretest

In order to create test materials for the belief revision task (Experiments 1-3), a pretest was developed to measure the implicit Scope values of various general statements.

Ninety-one undergraduates volunteered to participate in this pretest condition for course credit. They viewed 24 universally quantified statements. There were four conditions of statements created by crossing two levels of Ontology (Definitionally true or Empirically true statements) by two levels of Relation (Class-Inclusion or Property-Assignment).

Students identified the Scope of each statement on an 8point scale, where each point corresponded to a power of 10 . The scale was anchored by " 0 " and " 7 ". For example, a Scope of " 4 " encompassed a Scope size from 1,000 to 10,000 . The order of the statements was randomized in the booklet.

As anticipated, Scope values were greater for Definitional statements ( $M=6.26, S D=.90$, e.g., All trees are plants) than Empirical statements ( $M=5.65, S D=.82$, e.g., All professors are teachers). Class Inclusion statements ( $M=$ $6.09, S D=.88$, e.g., All oranges are fruit) received greater Scope values than Property-Assignment statements ( $M=$ 5.82, $S D=.83$, e.g., All mammals have hair). These findings are in keeping with the importance of these variables for semantic verification (e.g., Collins \& Quillian, 1969; Quillian, 1968; Rips, 1989)

## Experiment 1

The goal for Experiment 1 was to examine belief revision in the context of statements that have verified believability and Scope. Using the Scope values recorded from the Pretest, we aimed to compare those values with retention rates to determine the role of Scope in a statement's retention. We predicted that for both MT and MP problems, retention rates would increase as the Scope of the generality increased because Scope reflects the law-like aspect of the statements.

## Method

Seventy-seven undergraduate students participated in this experiment for course credit. Two booklets of belief revision problems were created from the Pretest materials. One booklet contained MT belief revision problems and the second contained MP belief revision problems. Within each booklet, there were four types of problems (Problem Type) that result from the crossing of Ontology (Definitional or Empirical relations and Relation (Class-Inclusion or Property-Assignment). There were three exemplars for each type of problem, chosen randomly from the Pretest, creating booklets of 12 problems each.

The problems were randomly ordered in each booklet and the booklets were randomly assigned to participants. For each problem, participants were asked to accept the assumption as true, and discard (by crossing-out) the statements that contradicted the assumption.

## Results and Discussion

Participants' preference to retain the generality was scored for each problem and compared with chance (50\%).

Table 1: Mean (SD) Retention Rates for MT and MP Revisions in Experiment 1

|  |  | MT |  |
| :--- | :--- | :---: | :---: |
| Problem Type |  | MP |  |
| Definitional | Class | $91 \%(.21)^{* * *}$ | $56 \%(.40)$ |
|  | Property | $93 \%(.18)^{* * *}$ | $67 \%(.41)^{*}$ |
| Empirical | Class | $80 \%(.25)^{* * *}$ | $46 \%(.39)$ |
|  | Property | $71 \%(.36)^{*}$ | $48 \%(.38)$ |
|  | Binomial analysis: ${ }^{*} \mathrm{p}<.05,{ }^{* *} \mathrm{p}<.01,{ }^{* * *} \mathrm{p}<.001$ |  |  |  |

Table 1 displays the rate of retention of generalities and shows that participants prefer to retain the generalities in all MT problems significantly more often than chance, but preference for generalities in MP problems was not reliably different from chance (MT: $M=.84, S D=.25$ and MP: $M=$ $.54, S D=.40 ; F(1,66)=18.44, p<.001)$. Analysis also revealed a significant difference in retention rates for Problem Type, $F(3,198)=13.98, p<.001$. There was no significant interaction between Problem Type and Logical Structure, with both MT and MP following similar trends, $F(3,198)=1.28, p=.28$.

Simple regression analyses were conducted to determine if retention rates changed in relation to the Scope of the generality (derived from the Pretest). Overall, Scope ( $\beta=$ $.81, p=.001$ ) significantly predicted overall retention rates $\left(F(1,10)=19.37, p=.001, R^{2}=.66\right)$. Specifically, for MT contradictions, retention of the generalities increased with Scope $\left(\beta=.89, p<.001 ; F(1,10)=37.39, p<.001, R^{2}=\right.$ .80). However, for MP contradictions, Scope ( $\beta=-.55, p=$ .06) was a negative predictor of commitment to generalities, $F(1,10)=4.40, p=.06, R^{2}=.31$.

When reasoners seek to revise true beliefs in order to return consistency to a set of statements, they show a stronger commitment to the generalities when the logical
structure was expressed as an MT argument than as an MP argument even though the generalities are identical in the two conditions.

The Scope of the generality correlates positively and strongly with the tendency of reasoners to retain them in a MT argument structure, but Scope was negatively related to retention of generalities in MP arguments.

## Experiment 2

In this experiment, we once again assess the importance of Scope for belief revision. Here we try to constrain the Scope of the generalities by including a location in each epistemic set that implicitly constrains the generalities' Scope in space and time. When a locative is introduced, does it undermine the impression that the generalities are true across space and time? For example, consider "All snakes slither." This statement entails a large Scope (shown in Experiment 1). However, if a special desert is referred to (e.g., Rich lives in a desert where all the snakes slither), it invites the question whether there could be something peculiar about the location or why would it be introduced? Here we assess whether a specific location limits the Scope of the generalization and therefore the pattern of belief revision. Alternatively, perhaps the cognitive processes employed to uncouple the generalization from the location would result in enhancing the reasoner's commitment to the generality.

## Method

Seventy-eight undergraduate students participated in this experiment for course credit. Participants solved the same belief revision problems used in Experiment 1, but those in Experiment 2 introduced novel locatives for each problem. Participants were either given MT contradictions or MP contradictions of the same problem set to solve (betweensubjects). Participants solved 12 problems in total.

## Results and Discussion

Table 2 shows the retention of generalities for Logic Structure and Problem Type. It reveals that participants preferred to revise beliefs by retaining the general statements significantly more often than would be expected by chance $(50 \%)$ for each condition. As the table reveals, reasoners show a preference in all conditions to revise beliefs by retaining generalities. Logical Structure was important to revisions: participants who solved MT contradictions ( $M=.88, S D=.33$ ) were more likely to retain the generality of the problem than those who solved MP contradictions $(M=.71, S D=.34 ; F(1,63)=7.69, p=$ .007). The Ontology of the statements' relations was also critical, with participants retaining the Definitional generalities $(M=.82, S D=.25)$ more often than Empirical ones $(M=.77, S D=.27 ; F(1,63)=3.62, p=.06)$. There was no effect of Relation (Class vs. Property) and no interaction among the variables in this study.

Scope was not found to be a significant predictor of retention rates for either MT contradictions $(\beta=.27, p=$
.39) or MP contradictions ( $\beta=.21, p=.51$ ). In fact, the trend line seen in Experiment 1 for MP problems reverses direction (from negative to positive). This shows that providing a locative altered the importance of Scope.

Table 2: Mean (SD) Retention Rates for MT and MP Revisions in Experiment 2

|  | MT |  |
| :--- | :---: | :---: |
| Problem Type | $M(S D)$ | $M(S D)$ |
| Definitional | $90 \%(.35)^{* * *}$ | $74 \%(.36)^{* *}$ |
| Empirical | $85 \%(.36)^{* * *}$ | $69 \%(.38)^{*}$ |
| Binomial analysis: ${ }^{*} \mathrm{p}<.05,{ }^{* *} \mathrm{p}<.01,{ }^{* * *} \mathrm{p}<.001$. |  |  |

Overall, the presence of a locative still allows an effect of Logic: participants tended to retain the generality for MT problems more often than those participants evaluating MP problems. However, comparing across experiments, the retention of MP generalities increased at a much higher rate $(t(64)=-2.19, p=.03)$ than those observed in Experiment 1, while no change was observed in the retention rate of generalities in the MT structures.

## Experiment 3

The aim of Experiment 3 was to examine belief revisioning in artificial environments (e.g., the locatives of Experiment 2), but with real categories, whose Scope has been modified. Scope was expressed as either a small proportion or a large proportion of the total members of the reasoning categories (e.g., Kelly has a hive where 5 bees out of 104 insects have wings vs. Logan has a honeycomb where 91 bees out of 104 insects have wings.). In Experiments 1 and 2, Scope was implied. Here, it is explicitly stated. We anticipated that explicitly stating the Scope of the generalities would-along with the locatives-decouple the categorical expressions from their normative senses and therefore make the belief revision context more artificial.

## Method

Fifty-four undergraduate students participated in this experiment for course credit. In addition to the variables present in the task for Experiment 2 (Logic, Ontology, and Relation), a new between-subjects variable was added to the problem set: Scope proportion (Small or Large). There were four total conditions: MT Large, MT Small, MP Large, and MP Small. Participants solved 12 belief revision problems where Scope information was given either in a Small (5\%) or Large ( $87 \%$ ) Proportion. The instructions and procedure for this task were the same as Experiments 1 and 2.

## Results and Discussion

Table 3 shows that participants prefer to revise their beliefs by retaining the generality (Binomials, $p<.05$ ). Overall, reasoners who solved MT contradictions ( $M=.89, S D=$ .29) were slightly more likely to accept the generality of the problem than those who solved MP contradictions ( $M=.77$,
$S D=.29 ; F(1,50)=3.99, p=.05)$. Compared with previous experiments, the retention rates of generalities in MT contradictions were at ceiling and those of MP contradictions were higher than previous. Such increased retention rates overshadowed any effect of Scope, which was not a significant predictor for MT or MP problems in either Small or Large Proportion conditions.

Table 3: Mean (SD) Retention Rates for MT and MP Revisions in Experiment 3 for Small and Large Scope Conditions

|  |  | MT | MP |
| :--- | :--- | :---: | :---: |
| Problem Type | $M(S D)$ | $M(S D)$ |  |
| Small | Definitional | $89 \%(.42)^{* * *}$ | $77 \%(.40)^{* *}$ |
|  | Empirical | $85 \%(.49)^{* *}$ | $70 \%(.48)$ |
| Large | Definitional | $95 \%(.42)^{* * *}$ | $87 \%(.40)^{* * *}$ |
|  | Empirical | $86 \%(.49)^{* *}$ | $74 \%(.48)^{*}$ |
|  | Binomial analysis: ${ }^{*} \mathrm{p}<.05,{ }^{* *} \mathrm{p}<.01,{ }^{* * * \mathrm{p}<.001 .}$ |  |  |  |

The locatives appear to influence the decision-making by de-coupling the artificial context from the implicit Scope of the categories, thus allowing increased retention of generalities.

## General Discussion

Revising our beliefs when we are confronted with conflicting information is ubiquitous. Yet, the cognitive processes underlying this kind of reasoning are poorly understood because the prevailing research has not studied belief revision with consistent content. Some tasks have used artificial content, with no relational structure among the beliefs. In others, the artificial beliefs have been part of simple assertions or immersed in stories. In cases where presumed beliefs have been used, they are often not verified. The present study created belief revision conditions where the meaning of the statements and the degree to which they could be interpreted as scientific laws-their Scope-have been independently verified along with reasoners’ commitment to them.

By controlling the Scope of statements, we were able to identify the importance of the logical structure in which the belief revisions are contained. In an MT structure, reasoners prefer to revise their beliefs by retaining the most law-like generalities and by eliminating the particular statements that are inconsistent with the generalities. For the MT structure, the Scope of the generalization predicts the tendency to retain these true statements: the greater the Scope, the more likely will the generalization be retained. In contrast, when the epistemic structure is cast as an MP argument, reasoners do not show a preference for generalities or facts nor does Scope play an appreciable role.

Belief on its own is not critical to the belief revision process. We know this because the MT problems employed the same generalities as the MP problems, yet the former were retained significantly more often than the latter. These
findings show that critical to belief revision are the dual factors of (a) a statement's Scope and (b) the argument's structure in which the statement is immersed. Neither factor alone is sufficient to account for how people revise a set of beliefs. This leaves the question of why these two factors should dictate the belief revision process.

Pursuant to David Lewis' theory of Possible Worlds (Lewis, 1973, 1986), we propose that when revising their epistemic system, reasoners imagine an organized possible world, closest to the current one. To accomplish this, they rank-order the beliefs in terms of degrees of necessity, with the law-like propositions given the highest ranking. In MT, the general statement tends to be the one with the greatest Scope and is the starting point in the revision process with the deletion of any statement inconsistent with it. In contrast, in MP, the assumption statement challenges the modal status of the generality, diminishes its ranking, and rending all statements equivalent (Rescher, 1963, 2007). As a result, no preference for retaining any statement is revealed. In this case, the generalizations in MP problems do not possess the same commitment post-assumption as they do pre-assumption. Hence, Scope is less predictive of the decisions to retain statements in these problems. However, when the context is rendered artificial (as with a locative) it enhances sensitivity to the implicit modal status of the generalization even in the MP logical structures.

The relation between believability of statements and the logic of the belief revision context can be explained by the Conditional Probability Hypothesis (e.g., Evans, Handley, \& Over, 2003; Wolf et al., 2012). It states that for MP problems, the probability of the truth of the generality is zero in the face of the counterfactual assumption. To understand this, recall MP example (2) above:
(2a) All whales are mammals
(2b) This creature is a whale
(2c) This creature is a mammal
(2d) Assume that this creature is not a mammal
The conditional probability of the generality is stated as $P(\mathrm{q} \mid \mathrm{p})=$ probability of "mammal" given the rule stated conditionally as "if whale, then mammal". The assumption statement (2d) states that the probability of being a mammal is zero. Therefore, $P(\mathrm{q} \mid \mathrm{p})=0$; therefore the probability that the rule is true is also zero. This leads to the expectation that there will be no discernible preference for retaining either the generality or the particular statement in MP. In contrast, consider the case of the MT argument repeated below:
(1a) All whales are mammals
(1b) This creature is not a mammal
(1c) This creature is not a whale
(1d) Assume that this creature is a whale
In the case where the a priori belief in the generality is greater than zero, the assumption does not alter that. Therefore, the preference for the generalities in this MT structure will typically be greater than what is found for MP. This will be true even though the general statements are syntactically identical in the two logical structures. In brief, the conditional probability hypothesis is able to account for
the typical finding that belief revision varies with logical structure, all things being equal.

However, the contrasting preferences shown in Experiments 1 and 2 are not readily explained by the conditional probability hypothesis. The retention of the generalities in Experiment 2 is reliably greater than those in Experiment 1 and this is especially the case for MP. The problems differ in the presence of a locative in Experiment 2, which is intended to reduce the law-like properties of the generality by reminding the reasoner that the truth of these statements may be limited in space and time. These locatives should also reduce the a posteriori probability of the inclusive category (e.g., "mammal") and reduce the retention of generalities especially in MP. The procedure produced the opposite result. So, while the conditional probability hypothesis has much to recommend it as an account for belief revision, more work needs to be done to understand the cognitive processes contributing to the retention of beliefs.

A second account of belief revision has focused on Disabling Conditions. The claim has been made that in the face of the counterfactual assumption, reasoners construct explanations for the inconsistencies (Khemlani \& JohnsonLaird, 2011). These explanations focus on the generalizations because they contain many component elements and reasoners imagine that one of these elements has been disabled, thereby allowing for an inconsistency. The degree to which such disabling conditions are contrived is indirectly related to the strength of belief in the generality. This approach makes the following predictions: (a) generalities will be rejected in order to retain consistency in the epistemic set; (b) since rejection of the generalities are based on believability, the disabling conditions (and therefore rejection of the generality) will be equivalent across logical structure.

These predictions are not consistent with the present findings. Overall, generalities are retained more often than would be expected by chance. Generalities whose Scope is artificially low (Experiment 3) should show the effect of disabling conditions more so than when the Scope is artificially high, yet no difference in retention is shown for these types of statements. While the presence of potential disabling conditions may play a role in some aspects of belief-revision, it is clearly not the underlying mechanism employed for confronting the counterfactual assumptions.

## Conclusion

We are obliged to revise our system of beliefs when we accept a new piece of information that introduces an inconsistency into our knowledge structure. Here we are faced with the task of retaining some old information and rejecting others. The present study examined how people perform this task when dealing with real beliefs and facts.

Scope and Logical Structure were jointly important predictors of whether students would retain a statement when required to revise their epistemic sets. When the statements fit within an MT structure, reasoners organized
their revisions around the law-like generalizations. In contrast, when the statements fit within an MP structure, the participants did not show a preference in how they organized their beliefs. This suggests that the importance of belief strength is influenced by the structure in which they are immersed.

These findings lend empirical support to the philosopher David Lewis' view that belief revision is characterized by Possible Worlds logic in which reasoners structure their revisions by organizing their epistemic systems so as to give priority to the most law-like statements.

## Acknowledgments

The authors would like to thank Desiree I. Garcia for all her hard work during data collection.

## References

Byrne, R. M. J. \& Walsh, C. R. (2002). Contradictions and counterfactuals: Generating belief revisions in conditional inference. In W. Gray \& C. Schunn (Eds.), Proceedings of the 24th Annual Conference of the Cognitive Science Society (pp. 160-165). Austin, TX: Cognitive Science Society.
Chisholm, R. M. (1946). The contrary-to-fact conditional. Mind, 55, 389-407.
Collins, A. M. \& Quillian, M. R. (1969). Retrieval time from semantic memory. Journal of Verbal Learning and Verbal Behavior, 8, 240-247.
Elio, R., \& Pelletier, F. J. (1997). Belief change as propositional update. Cognitive Science, 21, 419-460.
Evans, J. St. B. T., Handley, S. J., \& Over, D. E. (2003). Conditionals and conditional probability. Journal of Experimental Psychology: Learning, Memory, and Cognition, 29, 321-335.
Goodman, N. (1954). Fact, fiction, and forecast, $I^{\text {st }}$ edition. Cambridge, MA: Harvard University Press.
Lewis, D. K. (1973). Counterfactuals. Cambridge, MA: Harvard University Press.
Lewis, D. K. (1986). On the plurality of worlds. New York: Basil Blackwell.
Khemlani, S. S., \& Johnson-Laird, P. N. (2011). The need to explain. The Quarterly Journal of Experimental Psychology, 64(11), 2276-2288.
Politzer, G., \& Carles, L. (2001). Belief revision and uncertainty reasoning. Thinking and Reasoning, 7, 217234.

Quillian, M. R. (1968). Semantic memory. In M. Minsky (Ed.), Semantic Information Processing, Vol. 2. Cambridge, MA: MIT Press.
Rescher, N. (1964). Hypothetical reasoning. Amsterdam: Elsevier, North-Holland.
Rescher, N. (2007). Conditionals. Cambridge, MA: MIT Press.
Revlin, R., Calvillo, D., \& Ballard, S. (2005). Counterfactual reasoning: Resolving inconsistency before your eyes. Psychologica Belgica, 10, 47-56.
Revlin, R., Calvillo, D.P., \& Mautone, P. (2003).

Counterfactual reasoning: How to organize a possible world. In R. Alterman \& D. Kirsh (Eds.), Proceedings of the 25th Annual Conference of the Cognitive Science Society (pp. 991-999). Austin, TX: Cognitive Science Society.
Revlin, R., Cate, C. L., \& Rouss, T. S. (2001). Reasoning counterfactually: Combining and rending. Memory \& Cognition, 29, 1196-1208.
Revlis, R. \& Hayes, J.R. (1972). The primacy of generalities in hypothetical reasoning. Cognitive Psychology, 3, 268290.

Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou \& A. Ortony (Eds.), Similarity and Analogical Reasoning (pp. 21-59). New York: Cambridge University Press.
Ryle, G. (1949). The concept of mind. New York, NY: Hutchinson's University Library.
Van Hoeck, N., Revlin, R., Dieussaert, K. \& Schaeken, W. (2012). The development of counterfactual reasoning in belief revision. Psychologica Belgica, 52, 407-434.
Wolf, A. G., Rieger, S. \& Knauff, M. (2012). The effects of source trustworthiness and inference type on human belief revision. Thinking and Reasoning, 18, 417-440.

# 8-month-olds Know Which Face is Reliable 

Kristen Swan Tummeltshammer (k.swan@bbk.ac.uk)<br>Centre for Brain \& Cognitive Development, Birkbeck, University of London Malet Street, London WC1E 7HX UK

Rachel Wu (rwu@bcs.rochester.edu)<br>Dept. of Brain and Cognitive Sciences, University of Rochester, Meliora Hall<br>Rochester, NY 14627 USA

Natasha Z. Kirkham (n.kirkham@bbk.ac.uk)
Centre for Brain \& Cognitive Development, Birkbeck, University of London Malet Street, London WC1E 7HX UK


#### Abstract

By 8 months of age, infants use statistical regularities and perceptual cues to orient attention (e.g. Kirkham et al., 2007; Wu \& Kirkham, 2010). However, it is unclear whether infants are sensitive to the reliability of individual attentional cues. In this eye-tracking study, 8 -month-olds were familiarized with a reliable face, which always looked to a box where an animation appeared, and an unreliable face, which looked only $25 \%$ of the time to the box containing the animation. At test, when the animations did not appear, infants searched longer in the corner cued by the reliable face, but did not search longer in the corner cued by the unreliable face. These results suggest that even young infants can track the the reliability of potential informants and use this information to distribute attention in support of early learning.


Keywords: Psychology; attention; spatial cognition; infancy; eye-tracking

## Introduction

For young infants, the natural world is a constant stream of dynamic, multi-modal sensory experiences. In a short time, they are able to parse this sensory overload into discrete and recognizable objects, faces, and events. Selective attention plays a critical role in this early learning, as infants must focus on items that contain useful information while ignoring random variation and meaningless noise. A number of studies have demonstrated that infants can allocate attention selectively in support of task-relevant learning (Mareschal \& Johnson, 2003; Richardson \& Kirkham, 2004; Tummeltshammer \& Kirkham, in press; Wu \& Kirkham, 2010). However, the selection process by which they are able to filter relevant information from noise is less well understood.

Given that the natural world contains a high degree of statistical redundancy, showing considerable consistency across space and time (Field, 1994), and there is evidence that the developing response properties of some visual neurons exploit the statistical nature of the input (Olshausen \& Field, 1996), it would be advantageous for the system to selectively attend to statistically reliable and coherent events. Research with young infants robustly shows that they are sophisticated statistical learners, tracking
probabilistic events across multiple instances and updating their representations of the world based on incoming data (Fiser \& Aslin, 2002; Kirkham et al., 2002; Kirkham et al., 2007; Saffran, Aslin \& Newport, 1996; Smith \& Yu, 2008; Wu, Gopnik, Richardson, \& Kirkham, 2011).

Recent studies have demonstrated that infants distribute attention selectively based on statistical information (Kidd, Piantadosi, \& Aslin, 2012; Tummeltshammer \& Kirkham, in press), which may guide early learning of events and features that are reliably linked. For example, new evidence from Kirkham and colleagues (2012) shows that young infants prefer to look at objects with correlated rather than uncorrelated parts and are surprised when statistically coherent parts split apart (Wu, et al., 2011). Infants also deploy attention with the influence of external cues, including bottom-up perceptual salience and even abstract cue-target associations (Cohen, 1972; Colombo, 2001; Johnson, Posner, \& Rothbart, 1991; 1994; McMurray \& Aslin, 2004). If these cues contain reliable information, then they may guide the infant toward learnable content; however, a mismatch between external cues and statistical coherence may drive infants to distraction and prevent them from encoding the critical stimulus events. A few studies have shown that young infants will use central cues to orient attention to peripheral locations when individual cues and targets are perfectly correlated (Johnson, Posner \& Rothbart, 1991; McMurray \& Aslin, 2004). At present, however, there is little evidence to address whether infants use statistical information to evaluate the reliability of salient attentional cues.

Faces offer a good opportunity to test whether attention to salient cues is mediated by statistical reliability. From birth, infants are drawn to faces, particularly those expressing eye contact (Senju \& Johnson, 2009), and very young infants will orient faster to visible targets in the direction of an adult's gaze (Farroni, Massaccesi, Pividori, \& Johnson, 2004; Hood, Willen, \& Driver, 1998). Infants follow faces from 4 months of age, and are sensitive to the relationship between an adult's gaze and the locations of objects (D’Entremont, 2003; Senju, Csibra, \& Johnson, 2008). Indeed, there is recent evidence that infants learn better
from faces than other attention-directing cues (e.g., flashing lights; Wu \& Kirkham, 2010). It is unclear whether infants follow faces as a category of salient attentional cues or perhaps have a general expectation that faces will provide information. It remains an empirical question whether infants track the statistical coherence of associations between cues and their targets, and further, whether they can update their expectations of individual face cues to guide attention optimally.

Research on 'selective trust'/‘source monitoring' with young children has demonstrated that they take an informant's knowledge into account when soliciting or accepting new information. Preschoolers prefer to engage with informants who are knowledgeable rather than ignorant (Koenig \& Harris, 2005), and will extend labels to novel objects when they were provided by a reliable rather than an unreliable adult (Clement, Koenig, \& Harris, 2004; Koenig, Clement, \& Harris, 2004). This work has recently been extended down to older infants: In a recent study, Begus and Southgate (2012) found that 16 -month-olds point more to solicit information from adults who had previously labeled objects correctly than from those who had mislabeled objects. In addition, across two studies, Poulin-Dubois and colleagues found that 14 -month-olds were sensitive to an adult's reliability in a search task, and were more likely to follow a reliable adult's gaze behind an occluder (Chow, Poulin-Dubois, \& Lewis, 2008) and to imitate a reliable adult's actions (Poulin-Dubois, Brooker, \& Polonia, 2011). These studies suggest that infants as young as 14 months can make an association between an informant's actions and the true state of the world and use it to guide their own responses.

There are, however, some reasons to suspect that young infants may have difficulty tracking the reliability of face cues and allocating attention accordingly. First, while young children may be sensitive to the reliability of an informant, young infants may not attend to the relationship between a salient cue and its target outcome. This could be due to a general bias to follow faces, or the inability to simultaneously attend to the face cue and keep track of its reliability over trials. Second, young infants may have difficulty making within-category distinctions; even if they could successfully track the reliability of a category of attentional cues (e.g. 'Faces offer reliable information'), infants may fail to make separate inferences for individual instances of the same category (e.g. 'Face A is reliable, but Face B is not'). Third, young infants may form initial associations between cues and targets that are difficult to update in light of noisy data. In all of the studies described with young children, the unreliable or ignorant adults always provided false or incongruent information, so children may have simply represented those adults as 'wrong' or 'unsuccessful' without having to update their inferences.

The present eye-tracking study aimed to investigate whether 8 -month-old infants are sensitive to the statistical
reliability of attentional cues. Infants were familiarized with four audio-visual animations of animals that appeared within four boxes in each corner of the screen. On separate trials, the locations of the animals were cued by either a reliable or an unreliable face. The reliable face always looked in the box where an animal would then appear, while the unreliable face looked in the box containing an animal only $25 \%$ of the time (and rather, looked in an empty box $75 \%$ of the time). Following familiarization, infants viewed test trials in which the faces looked in the previously-cued boxes and the animal sounds played, but the animations did not appear. If infants had learned to expect an animation in the cued box, then we hypothesized that they should search longer in the cued box than in the uncued boxes. In addition, infants viewed generalization trials in which the faces looked to boxes that were never cued before and novel animal sounds played, but again no animations appeared. If 8 -month-old infants were able to track the reliability of the individual faces across trials, then we hypothesized that they should follow the reliable face to a new box, but abstain from following the unreliable face.

## Methods

## Participants

Twenty-four 8 -month-old infants (11 females, $M=8$ months 13.4 days, range: $7 \mathrm{~m} 12 \mathrm{~d}-9 \mathrm{~m} 7 \mathrm{~d}$ ) participated in the experiment, with an additional four infants tested but not included due to fussiness, inattention and/or failure to calibrate. Infants were recruited on a voluntary basis via local advertisements. Informed consent was received from all caregivers, and babies received a small gift.

## Apparatus and Stimuli

Infants were eye-tracked using a Tobii TX300 eye-tracker (www.tobii.com) with a 23 " built-in monitor. Stimuli were presented using Tobii Studio presentation software, and sounds were played through stereo external speakers. Throughout testing, infants were monitored via a built-in video camera and their eye movements through the Tobii Studio Live Viewer display. Two female actors were filmed in controlled settings and their footage was edited into face cue stimuli in Final Cut Express HD3 (Apple Inc., CA). The animated clips were created using Macromedia Director MX 2004 and combined with the face cues in Final Cut Express.

Infants saw a full-screen display (1920 X1080 pixels) comprised of four white boxes in the four corners of a black screen. Within each box, an animated animal appeared: a barking dog in Box 1, a croaking frog in Box 2, a gurgling fish in Box 3, and a chirping bird in Box 4. The animations were preceded by centrally presented face cues. On each trial, one of two female faces appeared in the center, smiled at the infant and said "Wow, look!". She then turned to one of the boxes and froze. An animal sound played and after a 500 ms delay, the corresponding animal appeared in its box. The animated animal moved within the box for 3.5 seconds, while the face remained frozen, as shown in Figure 1.


Figure 1. Examples of four familiarization trials with a reliable face cue (left) and four familiarization trials with an unreliable face cue (right). While the reliable face always looked to the correct box, where an animal would appear, the unreliable face only looked to the correct box on one out of four trials.

## Design and Procedure

All infants were tested individually in a quiet room, seated on their caregiver's lap approximately 60 cm away from the monitor. A 5-point calibration sequence (the four corners and center of the screen; for details, please refer to von Hofsten, Dahlström, \& Fredricksson, 2005) was used to obtain a reliable signal. Infant needed to fixate each point before the experimenter manually advanced the calibration sequence; if fewer than four points were accurately calibrated, the sequence was repeated.

Following successful calibration, infants were familiarized with a reliable face and an unreliable face on separate blocks (order counter-balanced across infants). The reliable face always looked at the box in which an animal animation would appear, reliably cueing two different boxes on separate trials. The unreliable face also cued two different boxes on separate trials, but only looked $25 \%$ of the time at the box containing an animation; that is, for the unreliable face, the animals often appeared in boxes that did not correspond to where the face had looked. For example, if the reliable face looked in Boxes 1 and 2 on four separate trials, either the $\operatorname{dog}$ (Box 1) or the frog (Box 2) would appear to match where the face had cued (see Figure 1A). However, if unreliable face looked in Boxes 1 and 3 on four separate trials, either the frog (Box 2) or the fish (Box 3) would appear, so that the cue and animation only matched on one of the four trials (see Figure 1B). Critically, one box was only cued by the reliable face, a second box was cued by both faces on separate trials, a third box was only cued by the unreliable face, and the last box was never cued.

Following familiarization, infants viewed test trials and generalization trials. On a test trial, the face looked to the box it had previously cued (whether reliably or unreliably) and the animal sound played; however, the animation did not appear. Instead, all four white boxes flashed briefly (200 ms ) to encourage infants to make a saccade. On a
generalization trial, the face looked to the box it had never looked at before and a new animal sound played. Again, no animation appeared, but all four white boxes flashed briefly to encourage saccades.

Infants viewed four blocks of four familiarization trials, with the reliable and unreliable faces on alternating blocks, followed by the two test blocks. This sequence was then repeated, for a total of 40 familiarization ( 20 reliable, 20 unreliable), 4 test, and 4 generalization trials ${ }^{1}$.

## Data Analysis

Eye movements were recorded and filtered into discrete fixations using a spatial filter of 30 pixels and a temporal filter of 100 ms . On test and generalization trials, when all four boxes flashed but no animations appeared, accumulated looking times (i.e. the summed durations of all fixations) to each of the four boxes were measured as a proportion of total looking time.

## Results

## Familiarization Trials

There were no differences in infants' attention to the faces (i.e. proportion of total accumulated looking time spent on the face) across familiarization trials, suggesting that infants looked equally to the reliable face $(M=0.609, S E=0.017)$ and the unreliable face $(M=0.621, S E=0.016)$, paired $t(23)=1.02$, $p=n s$.

[^223]
## Test Trials

Proportions of looking time to the four boxes during test trials, displayed in Figure 2, were analyzed with a 2 (Reliability) x 4 (Box) repeated measures ANOVA ${ }^{2}$. Results show a significant main effect of Box, $F(3,66)=3.64$, $p=0.017, \eta_{p}{ }^{2}=0.142$, as well as a significant Reliability x Box interaction, $F(3,66)=3.55, p=0.019, \eta_{p}{ }^{2}=0.139$. This interaction was unpacked using separate univariate ANOVAs for test trials with reliable and unreliable face cues. On reliably cued trials, a significant main effect of Box was apparent, $F(3,66)=8.32, p<0.001, \eta_{p}{ }^{2}=0.274$, and post-hoc comparisons indicated that infants looked longer at the cued box than at any other box, $p<0.040$ (Bonferronicorrected). However, on unreliably cued trials, no effect of Box emerged, $F(3,66)=0.21, p=0.888$, indicating that infants did not look longer at the cued box, nor at any other single box. Finally, a planned comparison across reliably and unreliably cued test trials confirmed that infants looked more to the cued box when it was cued by a reliable face than by an unreliable face, $t(22)=2.66, p=0.014$.


Figure 2. Mean proportions of looking time to the four boxes on test trials with the reliable and unreliable face cues.

## Generalization Trials

Similarly, proportions of looking time to the four boxes during generalization trials, shown in Figure 3, were analyzed with a 2(Reliability) x 4(Box) repeated measures ANOVA ${ }^{3}$. Results show a slight main effect of Box, $F(3,63)=2.70, p=0.053, \eta_{p}{ }^{2}=0.114$, as well as a significant Reliability x Box interaction, $F(3,63)=9.83, p<0.001$, $\eta_{p}{ }^{2}=0.319$. This interaction was explored using separate univariate ANOVAs for generalization trials with reliable and unreliable face cues. On reliably cued trials, a significant main effect of Box emerged, $F(3,63)=12.39$, $p<0.001, \eta_{p}{ }^{2}=0.379$, and post-hoc comparisons indicated that infants followed the cue to the new box, looking longer

[^224]at the new box than at any other box, $p<0.024$ (Bonferronicorrected). However, on unreliably cued trials, no effect of Box was apparent, $F(3,63)=0.40, p=0.754$, indicating that infants did not follow the cue to the new box, nor did they look longer at any other single box. Finally, a planned comparison across reliably and unreliably cued generalization trials confirmed that infants followed the cue to the new box more when it was a reliable face cue than an unreliable face cue, $t(21)=4.20, p<0.001$.


Figure 3. Mean proportions of looking time to the four boxes on generalization trials with the reliable and unreliable face cues.

## Discussion

Previous research has demonstrated that young infants are sensitive to statistical and perceptual cues and can use them to allocate attention in their busy, multisensory world. The present study suggests that infants can also integrate these sources of information to infer the reliability of individual cues and modify their responses. In the current study, infants searched consistently in the box cued by the reliable face, and even followed it to search in a box where no animation had appeared before. At the same time, infants did not follow the unreliable face, and rather searched at chance among all four boxes. These differences in looking behavior could not be accounted for by mere differences in global attention, as infants looked equally long at both reliable and unreliable face cues during familiarization trials.

Cue reliability also appeared to have important consequences for infants' audio-visual learning. Infants correctly predicted where a reliably cued animal would appear, but did not learn to localize the animal that had been unreliably cued. This study adds to a growing body of research suggesting that appropriate cues can enhance infants' processing and learning of cued events (Reid, Striano, Kaufman, \& Johnson, 2004; Senju, Csibra, \& Johnson, 2008; Yoon, Johnson, \& Csibra, 2008; Wu \& Kirkham, 2010). For example, Wu and Kirkham (2010) found that 8 -month-olds were better able to remember the spatial locations of audio-visual targets preceded by social cues compared to uncued targets. It is possible that infants' sensitivity to reliable cues may act as a driving force for
early learning, with cued attention helping the learner gather information and integrate it over time (Smith, Colunga, \& Yoshida, 2010). Indeed, enhanced detection and processing of cued stimulus events are well-documented in studies of selective attention with adults and children (Goldberg, Mauer, \& Lewis, 2001; Mackintosh, 1975; Posner, 1980).

While the present results suggest that infants are sensitive to the reliability of attentional cues, it remains unknown whether this sensitivity is face-specific or would extend to other types of cues. A few studies have shown that infants struggle to direct attention with a non-social central cue (Varga et al, 2009), though they seem to succeed in learning the cue-target relationship when the cue is perceived as social (Corkum \& Moore, 1998; Deligianni, Senju, Gergely, \& Csibra 2011; Johnson, Slaughter, \& Carey, 1998; Wu \& Kirkham, 2010). However, in these studies, cues have been used to direct infants' attention to one of multiple objects, with the result that infants look equally to both cued and uncued objects. Perhaps, then, infants need to learn the function of an abstract, non-social cue with a singular target (as in McMurray \& Aslin, 2004) before it can be used to disambiguate multiple targets. Future experiments will aim to evaluate whether infants consider the statistical reliability of attentional cues more broadly.

The mechanisms driving statistically cued attention are also unclear and worth investigating in future research. A modelling approach, using infants' own trial-by-trial data as input (cf. Piantadosi, Kidd, \& Aslin, in press; Yurovsky, Hidaka, \& Wu, 2012), may help to characterize the multiple processes involved in statistical learning of cued events, such as selective attending to cues and targets, tracking the correspondence between them, and deciding which cues to follow. Further, it would be interesting to distinguish whether infants' selective attention to cued events is motivated by the prospect of an exciting reward, or if there may be something intrinsically motivating about the predictive information itself. Bromberg-Martin and Hikosaka (2009) found that macaque monkeys prefer to have predictive cues rather than unpredictive cues, even when the ensuing rewards were identical. In the present study, infants received audio-visual animations on both reliably cued and unreliably cued trials, but did not develop a preference for the reliable (or unreliable) face. This may be due, in part, to the salience of the faces, or perhaps because infants were not trained to make a choice between cues as the monkeys were in Bromberg-Martin and Hikosaka (2009). Nevertheless, future research will aim to explore interactions in cued attention between the reliability of the cue and the salience of the reward.

## Conclusions

The present study demonstrates that 8 -month-olds can distinguish reliable and unreliable faces and use this inference to modify attention to cued targets. These results extend the existing literature on 'selective trust'/ssource monitoring' to young infants, suggesting that a sensitivity to the reliability of potential informants may be present early
in development. Selective trust, like selective attention, is influenced by statistical regularity, external cues, and the extent to which these factors are weighted in a particular context. This study has provided evidence that 8 -month-old infants can track the reliability of individual cues to deploy attention optimally in support of early learning.

## Acknowledgments

We thank Irati Rodriguez Saez de Urabain, Leslie Tucker, \& Marian Greensmith for their help with testing and recruitment. This work was funded in part by a grant from the Nuffield Foundation to NZK, and by Grant G0701484, awarded to Mark Johnson by the UK Medical Research Council.

## References

Begus, K., \& Southgate, V. (2012). Infant pointing serves an interrogative function. Developmental Science, 15, 611-617.
Bromberg-Martin, E.S., \& Hikosaka, O. (2009). Midbrain dopamine neurons signal preference for advance information about upcoming rewards. Neuron, 63, 119126.

Chow, V., Poulin-Dubois, D., \& Lewis, J. (2008). To see or not to see: infants prefer to follow the gaze of a reliable looker. Developmental Science, 11, 761-770.
Clément, F., Koenig, M., \& Harris, P. (2004). The ontogenesis of trust. Mind and Language, 19, 360-379.
Cohen, L.B. (1972). Attention-getting and attention-holding processes of infant visual preferences. Child Development, 43, 869-879.
Colombo, J. (2001). The development of visual attention in infancy. Annual Review of Psychology, 52, 337-367.
Corkum, V. \& Moore, C. (1998). Development of joint visual attention in infants. In C. Moore and P. J. Dunham (Eds.), Joint attention: Its origins and role in development (pp. 61-85). Hillsdale, NJ: Erlbaum.
Deligianni, F., Senju, A., Gergely, G., \& Csibra, G. (2011). Automated Gaze-Contingent Objects Elicit Orientation Following in 8 -Month-Old Infants. Developmental Psychology, 47, 1499-1503.
D'Entremont, B. (2000), A perceptual-attentional explanation of gaze following in 3- and 6-month-olds. Developmental Science, 3, 302-311.
Farroni, T., Massaccesi, S., Pividori, D., \& Johnson, M.H. (2004). Gaze following in newborns. Infancy, 5, 39-60.

Field, D. J. (1994). What is the goal of sensory coding? Neural Computation, 6, 559-601.
Fiser, J., \& Aslin, R.N. (2002). Statistical learning of new visual feature combinations by infants. Proceedings of the National Academy of Sciences of The United States of America, 99, 15822-15826.
Goldberg, M., Maurer, D., \& Lewis, T. (2001). Developmental changes in attention: The effects of endogenous cueing and of distractors. Developmental Science, 4, 209-219

Hood, B.M., Willen, J.D., \& Driver, J. (1998). Adults' eyes trigger shifts of visual attention in human infants. Psychological Science, 9, 131-134.
Johnson, M.H., Posner, M.I., \& Rothbart, M.K. (1991). Components of visual orienting in early infancy: Contingency learning, anticipatory looking, and disengaging. Journal of Cognitive Neuroscience, 3, 335344.

Johnson, M.H., Posner, M.I., \& Rothbart, M.K. (1994). Facilitation of saccades toward a covertly attended location in early infancy. Psychological Science, 5, 90-93.
Johnson, S.C., Slaughter, V., \& Carey, S. (1998). Whose gaze will infants follow? Features that elicit gaze following in 12 month olds. Developmental Science, 1, 233-238.
Kidd, C., Piantadosi, S.T., \& Aslin, R.N. (2012) The Goldilocks Effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. PLoS One, 7, e36399
Kirkham, N. Z., Richardson, D. C., Wu, R., \& Johnson, S. P. (2012). The importance of 'what': Infants use featural information to index objects. Journal of Experimental Child Psychology, 113, 430-439.
Kirkham, N.Z., Slemmer, J.A., \& Johnson, S.P. (2002). Visual statistical learning in infancy. Cognition, 83, B35B42.
Koenig, M., Clément, F., \& Harris, P.L. (2004). Trust in testimony: children's use of true and false statements. Psychological Science, 15, 694-698.
Koenig, M.A., \& Harris, P.L. (2005). Preschoolers mistrust ignorant and inaccurate speakers. Child Development, 76, 1261-1277.
Mackintosh, N.J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. Psychological Review, 82, 276-298.
Mareschal, D. \& Johnson, M. H. (2003). The "What" and "Where" of infant object representations. Cognition 88, 259-76.
McMurray, B., \& Aslin, R.N. (2004). Anticipatory eye movements reveal infants' auditory and visual categories. Infancy, 6(2), 203-229.
Olshausen, B. A., \& Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature, 381, 607-609.
Piantadosi, S.T, Kidd, C., \& Aslin, R. (in press). Rich analysis and rational models: Inferring individual behavior from infant looking data. Developmental Science.
Posner, M.I., 1980. Orienting of attention. Quarterly Journal of Experimental Psychology, 32, 3-25.
Poulin-Dubois, D., Brooker, I., \& Polonia, A. (2011). Infants prefer to imitate a reliable person. Infant Behavior and Development, 34, 303-309.
Reid, V.M., Striano, T., Kaufman, J., \& Johnson, M.H. (2004). Eye gaze cueing facilitates neural processing of objects in 4-monthold infants. NeuroReport, 15, 25532555.

Richardson, D. C., \& Kirkham, N. Z. (2004). Multi-modal events and moving locations: Eye movements of adults and 6 -month-olds reveal dynamic spatial indexing. Journal of Experimental Psychology: General, 133, 4662.

Saffran, J.R., Aslin, R.N., \& Newport, E.L. (1996). Statistical learning by 8-month-old infants. Science, 274, 1926-1928.
Senju, A., Csibra, G., \& Johnson, M. (2008). Understanding the referential nature of looking: Infants' preference for object-directed gaze. Cognition, 108, 303-319.
Senju, A., \& Johnson, M. H. (2009). The eye contact effect: Mechanisms and development. Trends in Cognitive Sciences, 13, 127-134
Smith, L.B., Colunga, E. \& Yoshida, H. (2010). Knowledge as process: Contextually-cued attention and early word learning. Cognitive Science, 34, 1287-1314.
Smith, L.B. \& Yu, C. (2008). Infants rapidly learn wordreferent mappings via cross-situational statistics. Cognition, 106, 1558-1568.
Striano, T., \& Reid, V.M. (2006). Social cognition in the first year. Trends in Cognitive Sciences, 10, 471-476.
Tummeltshammer, K.S., \& Kirkham, N.Z. (in press). Probabilistic variation and noise facilitate visual learning in infancy. Developmental Science.
Varga, K., Frick, J.E., Stansky, L., Beck, H., Dengler, M.J., \& Bright, M.A. (2009). I saw the sign: The development of directional understanding. In Poster presented at the biennial meeting of the Society for Research in Child Development, Denver, CO.
von Hofsten, C., Dahlstrom, E., \& Fredricksson, Y. (2005). 12-month old infants' perception of attention direction in static video images. Infancy, 8, 217-231.
Wu, R., Gopnik, A., Richardson, D.C., \& Kirkham, N.Z. (2011). Infants learn about objects from statistics and people. Developmental Psychology, 47 (5), 1220-1229.
Wu, R., \& Kirkham, N.Z. (2010). No two cues are alike: Depth of learning during infancy is dependent on what orients attention. Journal of Experimental Child Psychology, 107, 118-136.
Yoon, J.M.D., Johnson, M.H., \& Csibra, G. (2008). Communication-induced memory biases in preverbal infants. Proceedings of the National Academy of Sciences, 105, 13690-13695.
Yurovsky, D., Hidaka, S., \& Wu, R. (2012). Quantitative linking hypotheses for infant eye movements. PLOSOne, 7(10), e47419.

# Using intrinsic complexity of turn-taking games to predict participants' reaction times 

Jakub Szymanik (jakub.szymanik @gmail.com)<br>Institute for Logic, Language and Computation, University of Amsterdam<br>Ben Meijering (b.meijering@rug.nl )

Rineke Verbrugge (L.C.Verbrugge@rug.nl )
Institute of Artificial Intelligence, University of Groningen


#### Abstract

We study structural properties of a turn-based game called the Marble Drop Game, which is an experimental paradigm designed to investigate higher-order social reasoning. We show that the cognitive complexity of game trials, measured with respect to reaction time, can be predicted by looking at the structural properties of the game instances. In order to do this, we define complexity measures of finite dynamic two-player games based on the number of alternations between the game players and on the pay-off structure. Our predictions of reaction times and reasoning strategies, based on the theoretical analysis of complexity of Marble Drop game instances, are compared to subjects' actual reaction times. This research illustrates how formal methods of logic and computer science can be used to identify the inherent complexity of cognitive tasks. Such analyses can be located between Marr's computational and algorithmic levels.


Keywords: cognitive difficulty; strategic games; higher-order social reasoning; theory of mind

## Introduction

In recent years, questions have been raised about the applicability of logic and computer science to model cognitive phenomena (see, e.g., Frixione, 2001; Stenning and Van Lambalgen, 2008; Van Rooij, 2008). One of the trends has been to apply formal methods to study the complexity of cognitive tasks in various domains, for instance: syllogistic reasoning (Geurts, 2003), problem solving (Gierasimczuk et al., 2012), and natural language semantics (Szymanik and Zajenkowski, 2010). It has been argued that with respect to its explanatory power, such analysis can be located between Marr's (1983) computational and algorithmic levels.

More recently, there has also been a trend to focus on similar questions regarding social cognition, more specifically, theory of mind. Especially, higher-order reasoning of the form 'I believe that Ann knows that Peter thinks ...' became an attractive topic for logical analysis (Verbrugge, 2009). Here, the logical investigations often go hand in hand with game theory (see, e.g., Osborne and Rubinstein, 1994). In this context, one of the common topics among researchers in logic and game theory has been backward induction (BI), the process of reasoning backwards, from the end of the game, to determine a sequence of optimal actions (Van Benthem, 2002). Backward induction can be understood as an inductive algorithm defined on a game tree. The BI algorithm tells us which sequence of actions will be chosen by agents that want to maximize their own payoffs, assuming common knowl-
edge of rationality. In game-theoretical terms, backward induction is a common method for determining sub-game perfect equilibria in the case of finite extensive-form games. ${ }^{1}$

Games have been extensively used to design experimental paradigms aiming at studying social cognition (Camerer, 2003), recently with a particular focus on higher-order social cognition: the matrix game (Hedden and Zhang, 2002), the race game (Gneezy et al., 2010; Hawes et al., 2012), the road game (Flobbe et al., 2008; Raijmakers et al., 2013), and the Marble Drop Game (henceforth, MDG) (Meijering et al., 2010, 2011, 2012). All the mentioned paradigms are actually game-theoretically equivalent. They are all finite extensiveform games that can be solved by applying BI. As an example in this paper we will consider MDG (see Fig. 1).

Many studies have indicated that application of higherorder social reasoning among adults is far from optimal (see, e.g., Hedden and Zhang, 2002; Verbrugge and Mol, 2008). However, Meijering et al. $(2010,2011)$ report on a near ceiling performance of subjects when their reasoning processes are facilitated by, for example, a step-wise training. Still, an eye-tracking study of the subjects solving the game suggests that backward induction is not necessarily the only strategy used (Meijering et al., 2012).

We still do not know exactly what reasoning strategies ${ }^{2}$ the subjects are applying when playing this kind of dynamic extensive form games. One way to use formal methods to study this question has been proposed by (Ghosh et al., 2010; Ghosh and Meijering, 2011): to formulate all reasoning strategies in a logical language, and compare ACT-R models based on each reasoning strategy with a subject's actual performance in a sequence of games, based on reaction times, accuracy and eye-tracking data. This corresponds to a study between the computational and algorithmic levels of Marr's (1983) hierarchy.

[^225]

Figure 1: Examples of a zero-, first-, and second-order Marble Drop game. The blue marbles, on the left-hand side in the bins, are the participant's payoffs and the orange marbles, on the right-hand side, are the computer's payoffs. The marbles can be ranked from the lightest to the darkest. For each player, the goal is to get the white marble to drop into the bin with the darkest possible marble of their color. The participant controls the blue trapdoors (i.e., blue diagonal lines) and the computer controls the orange ones (the second set of trapdoors from the left). The participants are told that the computer aims at maximizing its pay-off. The dashed lines represent the trapdoors that both players should remove to attain the darkest possible marble of their color. See http://www. ai.rug.nl/~meijering/marble_drop.html for an interactive demo. Backward induction reasoning proceeds from the last decision, which in 1c is Player 1's decision between the blue marbles in payoff-pairs C and D. Player 1 would decide to remove the left trapdoor because $C$ contains the darker blue marble. Backward induction would then proceed with the second-to-last decision, which is Player 2's decision between the orange marbles in payoff-pairs B and C. Player 2 would decide to remove the left orange trapdoor, because B contains the darker orange marble. Backward induction reasoning stops at the third-to-last decision, which is Player 1's decision between the blue marbles in payoff-pairs A and B. Player 1 would remove the right blue trapdoor, because B contains the darker blue marble.

Here, we aim to tackle the problem from a somewhat more generic, complexity-theoretic viewpoint: we propose to study the problem on the computational level. Specifically, we will identify inherent, structural properties of the game that make certain MDG trials harder than others.

## Alternation type

Every instance of a finite extensive form game can be presented as a decision tree. The second-order trials of MDG have the abstract tree form presented in Fig. 2.

How to approximate the complexity of a single instance of MDG? In the worst-case scenario, the backward induction algorithm, based on breadth-first search from the leaves of the tree upwards, will have to travel through all the nodes of the decision tree. Thus, it will find the rational solution (Nash Equilibrium) in time and space proportional to the number of nodes plus the number of edges in the tree, $O(|V|+|E|)$. However, the size of the tree does not seem to be a psychologically plausible complexity measure. To see this, consider two trees of equal size, but in the first one all the nodes are


Figure 2: Nodes $s$ and $u$ are controlled by Player 1. $t$ is controlled by Player 2. If a player controls a node then in that node he can choose whether to go left, $l$, or right, $r$. Every leaf is labeled with the pay-offs for Players 1 and 2.
controlled by Player 1 while in the second tree, the players alternate. Obviously, the problem posed by the second tree is much more complex. This suggests that one of the key aspects of the problem is the structure of the move alternation in the game tree. Let us then categorize game trees with respect to such alternations. In the following, we restrict the analysis to two-player games, although it would be possible to extend the ideas to finite dynamic games for more than two players.

Definition 1 Let us assume that the players $\{1,2\}$ strictly alternate in the game; Let player $i \in\{1,2\}$. Then:

- In a $\Lambda_{1}^{i}$ tree, all the nodes are controlled by Player $i$.
- $A \Lambda_{k+1}^{i}$ tree, a tree of $k$-alternations for some $k \geq 0$, starts with a Player i node. ${ }^{3}$

For instance, the tree in Fig. 2 is $\Lambda_{3}^{1}$, a 1-game tree of 2 alternations, because Player 1 has the first move at the root, followed by an alternation of Player 1 to Player 2 and another alternation of Player 2 to Player 1.

## Pay-off structure and cognitive difficulty

From the psychological perspective, it seems really crucial to take pay-offs into account when comparing the difficulty of particular MDG tasks. For instance, the two trees from Fig. 3 are $\Lambda_{3}^{1}$, because they both start with Player 1 and both have two alternations, from Player 1 to Player 2 and back again. However, clearly, the first game, represented by $T_{1}$, is much easier for Player 1 than the second game, represented by $T_{2}$. In the first game it is enough for Player 1 to realize that 999 is the highest possible pay-off, and then he can instantly move left and finish the game.

To explain the eye-tracking data of the subjects solving the Marble Drop game, Meijering et al. (2012) suggest that subjects may be using forward reasoning with backtracking (henceforth FRB), based on statistical analysis of eye gaze sequences. For instance, in the game from Fig. 1c, Player 1 will find out that B contains the darkest blue marble. He has to ask himself whether that marble is attainable. In other words, he

[^226]

Figure 3: Two $\Lambda_{3}^{1}$ trees.
has to reason about whether Player 2 would remove the left orange trapdoor. Therefore, Player 1 has to look at the orange marbles in bins $\mathrm{B}, \mathrm{C}$ and D to find out that bin D contains Player 2's darkest orange marble. The reasoning continues with Player 1 asking himself whether Player 2 thinks that her orange marble in bin D is attainable. In other words, Player 1 has to reason about whether she thinks that he would remove the right blue trapdoor of the rightmost set of trapdoors. Player 1 knows that he would not remove that trapdoor, but that he would remove the left one instead. He also knows that she is aware of this, as both players are aware of each other's goals. Therefore, Player 1 knows that Player 2 knows that her darkest orange marble in D is unattainable. Therefore, Player 1 has to go back to the second decision point (i.e., the orange trapdoors). There, Player 2 would compare the orange marbles in B and C and decide to remove the left orange trapdoor, because the orange marble in $B$ is the darkest orange marble that she can still attain. To conclude, Player 1 knows that his darkest blue marble in $B$ is attainable, and will thus remove the right blue trapdoor of the leftmost set of trapdoors.

As it is relatively hard to conclude from the eye-tracking data whether subjects apply exactly the above described forward reasoning with backtracking, we propose an orthogonal idea. We aim to identify the properties of the games that make certain trials harder than others and see whether such an explanation is congruent with forward reasoning plus backtracking. In order to do that, we put forward the following definitions. The idea here is that subjects may be looking for the highest possible pay-off and then try to reach it.

Definition 2 A game $T$ is generic, if for each player, distinct end nodes have different pay-offs.

Note, for instance, that the game in Figure 1c is generic: the four bins contain marbles of four different hues of blue
and four different hues of orange.
Definition 3 Suppose $i \in\{1,2\}$. If $T$ is a generic game tree with the root node controlled by Player $i$ and $n$ is the highest possible pay-off for Player $i$, then $T^{-}$is the minimal subtree of $T$ containing the root node and the node with pay-off $n$ for Player i.

To illustrate this definition, Figure 4 shows the restricted $T^{-}$trees for the two trees shown in Figure 3.
Hypothesis 1 Let us take two MDG trials $T_{1}$ and $T_{2} . T_{1}$ is easier for participants than $T_{2}$ if and only if $T_{1}^{-}$is lower in the tree alternation hierarchy than $T_{2}^{-}$.

Hypothesis 1 takes into account pay-off structures. According to it, the first tree from Fig. 3, $T_{1}$, should be easier for participants than the right tree, $T_{2}$, as $T_{1}^{-}$is a $\Lambda_{1}^{1}$ tree while $T_{2}^{-}$is still $\Lambda_{3}^{1}$, see Fig. 4. Moreover, it is possible that some subjects may try to apply the procedure iteratively: check if the maximum pay-off is reachable, if not then check for the second-best pay-off, and so on.


Figure 4: The maximum pay-off restricted trees corresponding to the trees in Fig. 3.

As an additional question, we ask whether the following predictions agree with the proposal of Meijering and colleagues (Meijering et al., 2012) that the subjects in the game are applying forward reasoning, with backtracking when necessary (FRB). First of all, why would subjects ever apply FRB?

Hypothesis 2 For an average random game with 3 decision points structured as the $\Lambda_{3}^{1}$ game of Figure 2, the forward reasoning plus backtracking algorithm needs fewer computation steps to yield a correct solution than backward induction.

Furthermore, if subjects used forward reasoning, then we could observe the following by running FRB algorithm on the game trees:
Hypothesis 3 Let us take two MDG trials $T_{1}$ and $T_{2}$. The forward induction with backtracking algorithm yields a correct solution for $T_{1}$ faster than for $T_{2}$ if and only if $T_{1}^{-}$is lower in the tree alternation hierarchy than $T_{2}^{-}$.

## Experimental results

To experimentally corroborate our hypotheses, we analyzed performance and reaction time data from (Meijering et al.,
2012). Twenty-three first-year psychology students (14 female) with a mean age of 20.8 years (ranging from 18 to 24 years) participated in the experiment and were asked to solve Marble Drop trials, in the sense that they had to make a decision 'left' or 'right' at the first decision point. All experimental game trials had payoff structures that required Player 1 to reason about the decision at each of the three decision points, structured as the $\Lambda_{3}^{1}$ game of Figure 2. Therefore, the experiment was constructed in a way to be diagnostic for second-order theory of mind (see Meijering et al., 2012, for more information on the experimental design).

We divided experimental trials into two sets: Accessible ones, in which the highest possible pay-off for Player 1 is obtainable for him and Inaccessible ones, where his highest possible pay-off is not obtainable. For example, the game of Figure 1c is accessible, because Player 1 can reach the marble of the darkest hue of blue, which is located in bin B, by opening the right trapdoor; after all, Player 2 will also choose to stay there. Note that in general, if $T_{1}$ represents an accessible game and $T_{2}$ an inaccessible one, then $T_{1}^{-}$is lower in the alternation hierarchy than $T_{2}^{-}$.

Therefore, according to Hypothesis 1, our prediction was that the shortest reasoning times will be recorded in the condition "Accessible", where the highest pay-off was obtainable for Player 1.

Furthermore, by simulating forward reasoning with backtracking on experimental trials and computing the number of reasoning steps, we investigated hypotheses 2 and 3. Again, our prediction was that the number of steps should be smaller in "accessible" cases, where the highest-possible pay-off for Player 1 was obtainable.

## Hypothesis 1: pay-offs and alternation type

To investigate the first hypothesis, we compared reaction times (RTs) in games in which the highest payoff was accessible against RTs in games in which the highest payoff was not accessible. The RTs were log-transformed to approximate the normal distribution.

A paired-samples t-test indicated a significant (withinsubjects) difference, $\mathrm{t}(12)=4.07, \mathrm{p} ; .01$. The RTs decrease if the maximum payoff is accessible, which can be seen in Figure 5.

## Hypothesis 2: simulating the algorithms

When looking at all possible payoff-structures in Marble Drop games with two alternations (or three decision points), we implemented the forward reasoning plus backtracking algorithm as a set of heuristics based on several cases that can occur in the Marble Drop game; we used the same algorithm that we derived in (Meijering et al., 2012) from the participants' eye-tracking data. ${ }^{4}$

When using the algorithm on all 576 possible pay-off structures, we see that forward reasoning with backtracking in

[^227]

Figure 5: Players' reaction times with respect to accessibility, namely the attainability of the highest payoff for Player 1.
general requires fewer steps than backward induction, e.g., in 288 cases only 1 step is enough. More specifically, forward reasoning with backtracking requires on average 3 steps, whereas backward induction would always require 6 steps, irrespective of payoff structure. Table 1 provides a cross-table of payoff structures and number of steps. This simulation supports our Hypothesis 2.

Table 1: Cross-table of payoff structures and the necessary number of steps when using forward reasoning with backtracking.

| \# of steps | 1 | 2 | 4 | 5 | 6 | 8 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| \# of payoff structures | 288 | 72 | 48 | 56 | 16 | 96 |

These simulation results imply that, on average, it pays off to use a forward reasoning strategy. In fact, Meijering et al. (2012) found a strong prevalence of forward reasoning with backtracking, even though participants were presented with a subset of hard-to-solve games in which backward induction would actually be more efficient on average. However, participants did not know that they were presented with this particular subset of very difficult games.

## Hypothesis 3: FRB and structural complexity

The implementation of the forward reasoning plus backtracking (FRB) algorithm was applied to the subset of actually presented experimental games to determine the number of reasoning steps required for each game. In the following analyses, number of steps was included as a predictor of the reaction times. We label the factor simply as 'forward reasoning with backtracking'.

The log-RTs were analysed by means of linear mixedeffects (LME) models (Baayen et al., 2008) to account for random effects of participants and unequal numbers of observations across all experimental conditions. Traditional (repeated measures) ANOVAs could not be performed as they
require equal numbers of observations.
Fitting LMEs on the log-transformed reaction times, we see that forward reasoning plus backtracking (FRB) is a good predictor. The model with FRB cannot be rejected in favor of a simpler model without FRB as a predictor, $\chi^{2}(1)=8.4, p=$ 0.004 . We discuss the best model below.

Again, the reaction times significantly decrease if the maximum Player 1 payoff is accessible (Table 2a). In case of games in which the maximum payoff is not accessible, the reaction times do not significantly increase with each additional reasoning step (Table 2 b ). Those games require in between 6 and 8 reasoning steps, which is too small a difference to find a significant effect on the RTs. In contrast, the RTs do significantly increase with each additional reasoning step in games in which the maximum payoff is accessible (Table 2c).

Table 2: Output of full-factorial linear mixed-effects model with factors Accessibility (A), Steps of forward reasoning with backtracking (FRB).

| Parameter | Estimate | St. Error | t-value | p-value |
| :--- | :--- | :--- | :--- | :--- |
| a) Accessible | -0.689147 | 0.271256 | -2.54 | .000 |
| b) FRB | 0.008767 | 0.034930 | 0.25 | .418 |
| c) A:FRB | 0.084336 | 0.037277 | 2.26 | .000 |

## Discussion

We have investigated the structural properties of the Marble Drop Game, an experimental paradigm designed to study higher-order social reasoning. Using theoretical approaches from logic and complexity theory, we identified inherent properties of the game trials responsible for the cognitive difficulty of the task. Meijering and colleagues’ (2012) reaction time data can be explained by looking at the alternation type and pay-off distribution of the particular game items. It turned out that the game items are harder if the maximum possible pay-off for Player 1 is not accessible for him. This observation is consistent with the assumption that participants were mostly applying forward reasoning with backtracking to solve the games. By simulating forward reasoning with backtracking on the experimental items, we have shown that the reaction times and the number of necessary comparisons significantly decrease if the maximum Player 1 payoff is accessible. As MDG is game-theoretically equivalent to many other experimental paradigms making use of turn-based games (see, e.g., Hedden and Zhang, 2002; Gneezy et al., 2010; Hawes et al., 2012; Flobbe et al., 2008; Raijmakers et al., 2013), we would expect that our results generalize to those cases.

One could wonder why the subjects did not use backward induction in the first place, as it is the method that always delivers the optimal pay-off (Osborne and Rubinstein, 1994). One possible answer is that they avoided backward induction in order to simplify the underlying reasoning. Recall, that
while backward induction reasoning always takes 6 steps in the Marble Drop game with 3 decision points, forward reasoning and backtracking takes on average only 3 steps, corresponding with the phenomenon that $T^{-}$is usually lower in the tree alternation hierarchy than $T$ itself. Moreover, iterating the forward reasoning strategy by backtracking in case the highest pay-off is not obtainable will finally lead to the optimal solution. Therefore, some subjects may choose to use that strategy to avoid higher-order reasoning, even though keeping the intermediate results in mind during backtracking is expected to tax working memory more than applying backward induction.

Subjects may as well use other heuristics that do not guarantee reaching the prescribed backward induction result, namely a Nash equilibrium of the game. For instance, as suggested by Hedden and Zhang (2002), subjects may assume that their opponents are playing according to some fixed patterns. Instead of assuming that the opponent is rational and correctly predicts Player 1's choice at the last decision point, Player 1 may take his opponent to be risk-averse or risk-taking. Such heuristics, essentially based on considering sub-trees of the initial game-tree, will also lead to simplified reasoning.

Of course, assuming that the opponent is of some specific type changes the game drastically and can lead to a very bad outcome, in case of wrong judgement of the other player's type. Still, people notoriously apply similar heuristics in strategic situations, for example, when joining a poker table, many players try to evaluate whether the opponents play 'loose' or 'tight'. ${ }^{5}$ An important question is what are the good alternative strategies. They should be not only easy to compute for people but also relatively safe to apply. It seems that the forward reasoning plus backtracking strategy in MDG might be a cognitively attractive strategy for people asked to solve turn-based games. First of all, it does not ask for the second-order social reasoning that is known to be very hard even for many adults (Verbrugge, 2009), and moreover, on average it demands fewer comparisons. One may even think that competent players know a collection of various strategies and their strategic abilities could be partially equated with the skill of choosing the right one, i.e., a strategy that may be safely applied in a given context to simplify the underlying reasoning.

## Outlook

Inspired by the logical study of backward induction and the cognitive science experiments with the Marble Drop Game, we investigated structural properties of turn-taking dynamic games and we provided a more refined analysis of the complexity of particular game trials, which takes into account alternation type of the game and pay-off distribution. We com-

[^228]pared our predictions to actual reaction time data from (Meijering et al., 2012).

Of course, there are many further topics to be resolved. For instance, it would be interesting to extend our analysis to account for imperfect information games. Also it would be fruitful to explore connections with various related logical formalisms and to investigate further epistemic phenomena. In parallel, we would like to confront Hypotheses 1 and 3 with the available eye-tracking data from (Meijering et al., 2012), as well as with eye-tracking data to be gathered from a wider class of turn-based two-player games. Moreover, we plan to investigate other reasonable reasoning strategies that subjects may successfully adapt in game-plays.

## Acknowledgments

The authors are grateful for the support of Vici Grant NWO-277-80-001awarded to RV. JS also acknowledges NWO Veni Grant 639.021.232.

## References

Arora, S. and Barak, B. (2009). Computational Complexity: A Modern Approach. Cambridge University Press, New York, NY, USA, 1st edition.
Baayen, R., Davidson, D., and Bates, D. (2008). Mixedeffects modeling with crossed random effects for subjects and items. Journal of Memory and Language, 59:390-412.
Van Benthem, J. (2002). Extensive games as process models. Journal of Logic, Language and Information, 11(3):289313.

Camerer, C. (2003). Behavioral Game Theory: Experiments in Strategic Interaction. Princeton University Press, New Jersey.
Flobbe, L., Verbrugge, R., Hendriks, P., and Krämer, I. (2008). Children's application of theory of mind in reasoning and language. Journal of Logic, Language and Information, 17(4):417-442.
Frixione, M. (2001). Tractable competence. Minds and Machines, 11(3):379-397.
Geurts, B. (2003). Reasoning with quantifiers. Cognition, 86(3):223-251.
Ghosh, S. and Meijering, B. (2011). On combining cognitive and formal modeling: A case study involving strategic reasoning. In Van Eijck, J. and Verbrugge, R., editors, Proceedings of the Workshop on Reasoning About Other Minds: Logical and Cognitive Perspectives (RAOM-2011), Groningen. CEUR Workshop Proceedings.
Ghosh, S., Meijering, B., and Verbrugge, R. (2010). Logic meets cognition: Empirical reasoning in games. In Boissier, O., Fallah-Seghrouchni, A. E., Hassas, S., and Maudet, N., editors, MALLOW, volume 627 of CEUR Workshop Proceedings, Lyon. CEUR-WS.org.
Gierasimczuk, N., van der Maas, H., and Raijmakers, M. (2012). Logical and psychological analysis of deductive Mastermind. In Szymanik, J. and Verbrugge, R., editors,

Proceedings of the Logic \& Cognition Workshop at ESSLLI 2012, Opole, Poland, 13-17 August, 2012, volume 883 of CEUR Workshop Proceedings, pages 1-13, Opole. CEURWS.org.
Gneezy, U., Rustichini, A., and Vostroknutov, A. (2010). Experience and insight in the race game. Journal of Economic Behavior and Organization, 75(2):144-155.
Hawes, D. R., Vostroknutov, A., and Rustichini, A. (2012). Experience and abstract reasoning in learning backward induction. Frontiers in Neuroscience, 6(23).
Hedden, T. and Zhang, J. (2002). What do you think I think you think?: Strategic reasoning in matrix games. Cognition, 85(1): 1 - 36.
Marr, D. (1983). Vision: A Computational Investigation into the Human Representation and Processing Visual Information. W.H. Freeman, San Francisco.
Meijering, B., Van Maanen, L., Van Rijn, H., and Verbrugge, R. (2010). The facilitative effect of context on secondorder social reasoning. In Catrambone, R. and Ohlsson, S., editors, Proceedings of the 32nd Annual Conference of the Cognitive Science Society, pages 1423-1428, Austin (TX). Cognitive Science Society.
Meijering, B., van Rijn, H., Taatgen, N. A., and Verbrugge, R. (2012). What eye movements can tell about theory of mind in a strategic game. PLoS ONE, 7(9):e45961.
Meijering, B., Van Rijn, H., and Verbrugge, R. (2011). I do know what you think I think: Second-order theory of mind in strategic games is not that difficult. In Proceedings of the 33rd Annual Meeting of the Cognitive Science Society, pages 2486-2491, Boston. Cognitive Science Society.
Osborne, M. J. and Rubinstein, A. (1994). A Course in Game Theory. The MIT Press, Cambridge, MA.
Raijmakers, M. E., Mandell, D. J., Es, S. E., and Counihan, M. (2013). Children's strategy use when playing strategic games. Synthese.
Van Rooij, I. (2008). The tractable cognition thesis. Cognitive Science: A Multidisciplinary Journal, 32(6):939-984.
Stenning, K. and Van Lambalgen, M. (2008). Human Reasoning and Cognitive Science. The MIT Press, Cambridge, MA.
Szymanik, J. (2010). Computational complexity of polyadic lifts of generalized quantifiers in natural language. Linguistics and Philosophy, 33:215-250.
Szymanik, J. and Zajenkowski, M. (2010). Comprehension of simple quantifiers. Empirical evaluation of a computational model. Cognitive Science: A Multidisciplinary Journal, 34(3):521-532.
Verbrugge, R. (2009). Logic and social cognition: The facts matter, and so do computational models. Journal of Philosophical Logic, 38(6):649-680.
Verbrugge, R. and Mol, L. (2008). Learning to apply theory of mind. Journal of Logic, Language and Information, 17(4):489-511.

# A neural network model of working memory for episodes 

Martin Takac (takac@ii.fmph.uniba.sk)<br>Centre for Cognitive Science, Comenius University, Slovakia

Alistair Knott (alik@cs.otago.ac.nz)<br>Department of Computer Science, University of Otago, New Zealand


#### Abstract

We present a neural network model of the storage of episode representations in working memory (WM). Our key idea is that episodes are encoded in WM as prepared sensorimotor routines: i.e. as prepared sequences of attentional and motor operations. Our network reproduces several experimental findings about the representation of prepared sequences in prefrontal cortex. Interpreted as a model of WM episode representations, it has useful applications in an account of long-term memory for episodes and in accounts of sentence processing.


Keywords: working memory, episodic buffer, neural network models of language

## Introduction: working memory for episodes

The classical model of working memory (WM; Baddeley and Hitch, 1974) posits two representational media: one for visual material (the visuospatial sketchpad) and one for phonological material (the phonological buffer). Baddeley (2000) revised the model to include a third medium, holding semantic material-specifically, semantic representations of episodes-called the 'episodic buffer'. This medium stores semantic representations of actions, or events, or stative propositions. Our paper is about the episodic buffer.

Baddeley argues for the episodic buffer on several grounds. Some relate to models of language processing. When a sentence is being generated, the message which it is to express is standardly assumed to be maintained in the speaker's WM (see e.g. Levelt, 1989). When a sentence is being interpreted, several theorists envisage a set of competing episode representations being activated in the hearer's WM, with one of these eventually being chosen as the winner (see e.g. Mayberry and Miikkulainen, 2008). In each case we must assume a WM medium which stores semantic episode representations. Baddeley (2000) postulates bidirectional links between the episodic buffer and the phonological buffer, to support sentence-processing tasks. But in fact his primary argument for the episodic buffer has nothing to do with language processing. This argument concerns the neural mechanisms through which episodes are stored in long-term memory. The long-term neural storage of an episode is widely agreed to involve the hippocampus: specifically, the creation of links between hippocampal assemblies representing the various semantic components of the episode. But associations between hippocampal assemblies can only be learned if they are active almost simultaneously, within around 100 ms of one another (Abraham et al., 2002). Experiencing an episode often takes much longer than this. So we must envisage that episode representations are initially buffered in WM, and only relayed to the hippocampus when they are complete. This buffering
mechanism is likely to predate language, since apes are able to store episodes in long-term memory (see e.g. Schwartz and Evans, 2001). One interesting possibility is that evolution found a new use for the buffering mechanism in linguistic communication (see Knott, 2012; Takac et al., 2012). In this paper we present a connectionist model of WM storage which supports not only language processing, but also the prelinguistic role of the episodic buffer mediating transmission of episode representations to the hippocampus.

## WM episode representations as prepared sensorimotor routines

Our model is founded on the assumption that WM episodes provide an interface between the sensorimotor (SM) mechanisms through which episodes are apprehended and the hippocampal structures in which they are stored. On this assumption, we expect the structure of WM episode representations to reflect both the structure of SM processes and the structure of hippocampal representations. A strong commonality in the structures of these two domains is sequential organisation.

SM processing is strongly sequential at certain timescales, because it involves sequential deployments of the agent's sensory and motor apparatus. (For instance, saccades deploy the agent's fovea sequentially to targets in the world.) Ballard et al. (1997) propose that SM processing is organised into sequentially structured routines, whose atomic elements are discrete sensory or motor actions. These actions are termed deictic operations, and a sequence of such actions is termed a deictic routine. Through a case study of episodes involving reach-to-grasp actions, Knott (2012) argues that the SM processes through which concrete episodes are apprehended take the form of sequentially structured deictic routines.

The hippocampus stores associations between stimuli of many different kinds. But an emerging idea is that it is specially good at storing associations between sequentially structured items (Wallenstein et al., 1998). One recent finding which strongly supports this idea is that the hippocampus actively replays sequences of representations evoked during SM experience (see e.g. Lee and Wilson, 2002). The key result is that sequences of hippocampal place cells activated when a rat navigates a maze are replayed later when the rat is asleep. (Sequences are replayed at much higher speeds, perhaps consistent with the hippocampus' natural dynamics.) Since episodes are apprehended through well-defined sequences of SM operations, and sequences appear to be a natural unit of storage in the hippocampus, an interesting pos-
sibility is that WM episodes are also stored as sequences. Our model of WM episodes basically implements this idea.

We make two main proposals. First, we propose that a concrete episode is stored in WM as the sequence of SM operations through which it was experienced. We suggest that the order of SM operations in a deictic routine implicitly identifies the roles played by participants in the observed episode. Specifically, the object attended to first plays the role of the 'proto-agent': the entity which is most agentlike, animate or active (Dowty, 1991), and the object attended to next is the 'proto-patient'. This idea is motivated in detail in Knott (2012). Second, we propose that the sequence of SM operations is stored as a prepared deictic routine: i.e. as a prepared sequence of attentional and motor operations. Humans (indeed all primates) can prepare complex sequences of sensory and/or motor operations. If episodes are stored as prepared SM sequences, then there is a natural model of how they are transmitted to the hippocampus: they are simply replayed, at a speed commensurate with the associative learning mechanism in the hippocampus. Naturally, in replay mode the prepared attentional and motor operations are simulated rather than actually executed. (In fact, this proposal about the format of WM episode representations can be seen as a way of implementing 'simulationist' accounts of semantic representations; see e.g. Barsalou, 2008.) In summary: in our proposal episodes are experienced as sequences, stored in WM as prepared sequences, and then replayed to the hippocampus where they are stored more permanently as sequences.

## Representation of prepared sequences in prefrontal cortex

A bonus of the above model of WM episodes is that the neural mechanisms supporting preparation of SM sequences have been extensively studied, in single-cell recording experiments in monkeys. The principal mechanisms supporting sequence preparation are in dorsolateral prefrontal cortex (dlPFC; see e.g. Barone and Joseph, 1989; Averbeck et al., 2002). Several schemes for encoding prepared sequences have been found. In one scheme, individual neurons encode specific movements in particular contexts. For instance, Barone and Joseph (1989) found neurons which were active when a monkey prepared movement $A$, but only when it was followed by another movement $B$. In another scheme, neurons encode individual movements, and their position in the prepared sequence is given by their activation levels. For instance, in a monkey preparing a sequence of three movements $A B$ and $C$, Averbeck et al. (2002) found neurons representing each prepared action which were active in parallel, with the neuron encoding $A$ most active and that encoding $C$ least active. Interestingly, when the prepared sequence is executed, neurons encoding specific actions are inhibited just after their associated action is produced. Averbeck et al.'s (2002) findings strongly support a 'competitive queueing' model of sequence preparation, in which PFC assemblies encoding different actions compete against one another, with the winner triggering the associated
action, but also an operation to inhibit itself, so the next-most active assembly wins the competition at the next time point (see Rhodes et al., 2004). In competitive queueing, the representation of a prepared sequence is destructively updated in the medium in which competition occurs. We will call the sequence representations in this medium 'dynamic'. However, there is also evidence that prepared sequences are represented in a WM medium which is not destructively updated when a sequence is replayed. Perhaps most obviously, a given prepared sequence can be executed several times: each time, the sequence representation in the dynamic medium must somehow be restored from some more enduring medium. We will call representations in the enduring medium 'static'.

There is also evidence that a monkey can represent multiple alternative prepared sequences in dlPFC, in a medium which allows competition between candidate sequences and the selection of a winner. This evidence comes from a study by Averbeck et al. (2006), in which monkeys were trained to perform two sequences in response to two cues. Each day different cues were chosen to represent the two sequences. Halfway through the day, the mapping from cues to sequences was reversed, so the monkeys had to gradually learn the new mapping. During this period, dlPFC assemblies could be identified representing each prepared sequence, and the relative activation of the two assemblies after presentation of a cue could be used to predict the sequence which the monkey actually performed.

In summary, the prefrontal mechanism implementing sequence preparation appears to involve four distinct media. There is a medium holding representations of individual operations in a sequence, which encodes the context in which they appear. There is a medium holding distributed representations of whole sequences, in assemblies whose components encode individual actions, whose order is determined by their level of activation. Sequence representations in this medium are destructively updated when a prepared sequence is executed. But there is also a medium holding sequence representations which are not destroyed. Finally there is a medium in which alternative candidate sequence representations are active in parallel and compete with one another. If episodes are stored in WM as prepared SM sequences, then this mechanism would allow for WM episodes to be stored and replayed, and also for alternative WM episodes to compete amongst one another, with the winner being selected.

## A network for storing and selecting WM episodes

In this section we introduce a neural network which implements the sequence-preparation mechanism described above. One part of the network allows the storage and replay of experienced sequences in WM. However, another part of the network learns about commonly-occurring sequences, so it can make predictions about how a sequence being experienced will be completed, and or about which sequences are associated with reward for the agent. (We envisage the network
being used to control the process of 'experiencing an episode' both when the experiencer is acting and when he is watching an external episode.)

Our key aim for the network is that it learns the kind of representations of prepared sequences which are found in monkey PFC, as discussed above. However, there are also two other design criteria. Firstly, there should be a medium in which candidate SM operations compete with one another at every stage in the execution of a sequence. At any point, the operation which an agent executes is dictated partly by what is planned or expected, but also partly by bottom-up stimuli. We want a medium which allows competition between alternative operations from both these sources. Secondly, in the medium holding alternative possible SM sequences, there must be no scope for binding errors, whereby an item belonging to one sequence is falsely identified as part of a different sequence. Given that this medium must represent multiple sequences simultaneously, this is a difficult requirement. To address both these criteria, a key design decision is to use self-organising maps (SOMs; Kohonen, 1982). A SOM is a two-dimensional map of units fully connected to a layer of input units. When presented with training inputs, it learns to represent input vectors as localist units in the map, but also learns to represent similar inputs in similar regions of the map. It thus encodes similarities between its training input vectors even though it represents these in a localist scheme.

The architecture of our network is shown in Figure 1. The


Figure 1: Architecture of the network
network takes as input a sequence of SM signals at successive time points, evoked in the input SM signal area. Input SM signals can be thought of as representing either the agent's own actions (attentional or motor) or external stimuli in the world (objects or observed actions).

SM input signals are fed through an aggregate SM signal
area (see below) to a signal-encoding SOM. This SOM has recurrent connections, as described in Strickert and Hammer (2005): it takes as an additional input a context vector combining the weight and the context vector of the winner at the previous time point. When trained on a sequence of inputs, a recurrent SOM organises itself so that individual units encode signals occurring in particular sequential contexts, very much like the PFC units identified by Barone and Joseph (1989).

Units in the signal-encoding SOM represent signals in a localist way, so that alternative signals compete with one another. The winning signal at each time step is copied to an area which is isomorphic with the recurrent SOM called the dynamic episodic buffer. This area accumulates representations of each signal in an input sequence, with the first signal represented most strongly and subsequent signals being stored with decreasing activation, as in the prefrontal area studied by Averbeck et al (2002). When an input sequence is encoded in the dynamic episodic buffer, it can be replayed immediately by iteratively sending the dynamic episodic buffer's most active unit to the signal-encoding SOM (via the 'WTA' link) and then inhibiting this winning unit. To support repeated execution of a sequence, it can be stored in a static episodic buffer, which has the same structure as the dynamic one, and later reloaded.

At the highest level in the network there is another SOM called the candidate episodes buffer. This area encodes the distributed representations in the dynamic episodic buffer as localist units. During training it learns to represent episodes with similar encodings in the dynamic episodic buffer in neighbouring positions in the SOM. At every time point during presentation of a sequence this area represents a probability distribution over complete episodes. (If the network is being used to control the agent's own actions, this distribution represents action sequences which lead to reward; if it is being used to support observation of external episodes, it represents likely action sequences.) The distribution changes as new items arrive in the sequence and become encoded in the dynamic episodic buffer.

The winning unit in the candidate episodes buffer provides top-down activation to the static episodic buffer, through weights which are copies of those delivering input to the candidate episodes buffer. Since the winning unit always encodes a complete episode, the static episodic buffer likewise always encodes a complete episode, but in the same distributed format used by the dynamic episodic buffer. During presentation of a sequence, activity in the static episodic buffer is fed back to the signal-encoding SOM. This top-down input, when combined with the current context representation, produces a pattern of activity biased towards a representation of the next SM signal. The pattern is passed back to the aggregate SM signal area at the next time point. Thus the aggregate area receives both bottom-up inputs from the input SM signal and top-down ones from the static episodic buffer.

We conclude by reporting some details of the network architecture. Different SM operations are encoded in the 'input

SM signal' layer with 1-hot localist coding, i.e. one unit for each possible SM operation. The 'aggregate SM signal layer' is isomorphic with the input layer. The signal-encoding SOM is a 2-dimensional Merge SOM (Strickert and Hammer, 2005) with 400 units $(\alpha=0.4, \beta=0.5$, constant learning rate 0.1 and Gaussian neighbourhood decreasing from 10 to 0.5 ).

The static and dynamic episodic buffers are both isomorphic with the signal-encoding SOM, i.e. have 400 units each. Experiencing a sequence of SM operations creates a temporal pattern of active units in the signal-encoding SOM. This pattern is recorded in the dynamic episodic buffer as a 'trace' of the isomorphic units with exponentially decaying activity (the $n$th unit in the sequence has activity $\delta^{n-1}$ where $\delta=0.8$ and all unused units have zero activity). To prevent confusion of elements in the trace, we force the signal-encoding SOM to select a new winner in each step of the sequence (i.e. winners from previous steps of this sequence are excluded from competition). After completing the whole sequence, the 400dimensional vector representing its trace serves as a training input to the candidate episodes buffer, which is a standard SOM with 900 units, constant learning rate 0.9 and Gaussian neighbourhood decreasing from 10 to 0.5 .

Once a winner is selected in the candidate episodes buffer, activity is propagated back through the network, a process we call 'top-down reconstruction'. This process uses the property of SOMs that the memory of each unit is in its weights. During reconstruction, the weights of the winning unit in the candidate episodes buffer are copied back to the static and then dynamic episodic buffer. Destructive iterative updating of the dynamic episodic buffer causes a temporal sequence of activations of units in the signal-encoding SOM, which in turn project their weight vectors back to the aggregate SM signal layer where they represent top-down expectations.

## Experiments and results

Training We trained the model on sequences of SM operations, representing the SM routines through which different episodes are experienced. The SM sequences were built from 35 SM operations, e.g. MAN SNEEZE (intransitive episode), MAN CUP GRAB (transitive), MAN WALK HOUSE INTO (intransitive with PP complement), MAN CUP CAUSE BREAK (simple causative) and DOG BONE CAUSE ROLL TABLE UNDER (causative with PP). (For detailed justification of the orderings in these sequences, see Knott, 2012.) We repeated each simulation 10 times with different random initializations of connection weights in the model and different training sets (stochastically generated by the same set of transcription rules). Each training set consisted of 500 sequences, out of which on average 13.1 were of length $2,86.4$ of length $3,126.1$ of length 4 and 274.4 of length 6 . Sequences could contain duplicates: in all, $19.1 \%$ of sequences contained two copies of a single signal and $0.9 \%$ contained 3 . The training took 200 epochs; in each epoch the training sequences were presented in random order and the Merge SOM context was reset after each sequence. After training we tested the net-

| sequence fragment: DOG BALL |  |
| :--- | :--- |
| activity | reconstructed sequence |
| 0.30 | DOG BALL PUSH |
| 0.27 | DOG BALL SEE |
| 0.27 | DOG BALL GRAB |
| 0.26 | DOG BALL KICK |
| 0.25 | DOG BALL HIT |
| Sequence fragment: DOG BALL CAUSE |  |
| 0.33 | DOG BALL CAUSE GO |
| 0.32 | DOG BALL CAUSE STOP |
| 0.32 | *DOG BALL CAUSE GO CAT BALL CAT CAUSE GO |
| 0.29 | DOG BALL CAUSE HIDE DOG NEAR |
| 0.29 | DOG BALL CAUSE HIDE MAN UNDER |

Table 1: Probability distributions of episodes predicted in the candidate episodes buffer from two initial sequences. (The asterisk denotes an 'ill-formed' episode representation.)
work in three ways (all tests were repeated for the 10 different simulation runs and averaged).

Immediate serial recall The basic requirement for our network is that it can store and replay individual behavioural sequences. This capability relies on interactions between the signal-encoding SOM and the dynamic episodic buffer. We presented the trained network with 200 sequences of input signals: 100 taken from the training data and 100 new ones not seen before. Each sequence was coded in the dynamic episodic buffer; then the signal-encoding SOM's context was reset and the winning unit in the dynamic buffer was iteratively sent to the SOM and then inhibited. $99.4 \%$ ( $\mathrm{SD}=0.49 \%$ ) of training sequences were correctly replayed, and $98.6 \% ~(\mathrm{SD}=0.92 \%)$ of unseen sequences.

Predicted completions of sequences The network is also designed to generate top-down predictions about sequences, through activity in the candidate episodes buffer. The prediction is actually a retrieval of the most similar past episode as remembered in the weights of this buffer. The weights of the winning candidate are copied to the static episodic buffer and replayed in the signal-encoding SOM where they generate top-down biases for SM elements. To test this ability, we exposed the trained network to the 500 sequences encountered during training element by element, and examined the prediction about the possible completion of the sequence. At the beginning of an exposure, after seeing a short fragment of an episode, its completion is inherently ambiguous, as there may be many possible continuations (see Table 1). Later the number of candidates narrows down and the prediction can be more accurate. ${ }^{1}$ We can evaluate the retrieval from fragments of an episode of various lengths, up to complete episodes. The results are summarized in Table 2.

Note also that the network is not confused by sequences containing duplicate items. A regular competitive queueing network has problems representing duplicate items, because after the first instance of the item is presented it is inhibited in the competitive medium. But since the dynamic episodic

[^229]| Fragment length | $0-25 \%$ | $25-50 \%$ | $50-75 \%$ | $75-100 \%$ | $100 \%$ |
| :--- | :--- | :--- | :--- | :--- | :--- |
| Matches (avg) | $0.0 \%$ | $0.1 \%$ | $26.0 \%$ | $92.0 \%$ | $94.2 \%$ |
| Matches (SD) | 0.0 | 0.1 | 1.2 | 2.8 | 2.9 |

Table 2: Percentage of correct sequence completions from fragments of different relative length.

|  |  | MAN3 |  |  |  |
| :---: | :--- | :--- | :--- | :--- | :--- |
|  |  |  |  |  |  |
| MAN1 |  |  |  |  |  |
|  | MAN5 |  |  |  |  |
|  |  |  |  |  |  |
| MAN4 |  | MAN2 | MAN6 |  |  |

Figure 2: Position of the winning unit in the signal-encoding SOM for occurrences of the SM signal MAN in six different contexts. (Only a fragment of the $20 \times 20$ SOM is shown.)
buffer receives inputs from the signal-encoding SOM where we forced a unique winner selection, different instances of a given input are represented differently, and it does not suffer from this problem. To verify this, we tested the prediction on a set consisting of 95 sequences with 2 repeating elements and 5 sequences with 3 repeating elements and the results were similar to those presented above (the average success in prediction from fragments of more than $75 \%$ of the sequence length was $91.8 \% ~(\mathrm{SD}=3.5 \%)$.

Relation to neural activation data As discussed above, PFC stores prepared sequences in several different ways. We examined the properties of representations in the trained network to see how they corresponded to representations identified in monkey PFC.

Some PFC units encode individual operations in a prepared sequence in a way which takes into account their sequential context (see e.g. Barone and Joseph, 1989). Inspecting units in the signal-encoding SOM shows that they have this property. We presented the trained signal-encoding SOM with five input sequences featuring six instances of the signal MAN in different serial positions. The SOM unit which represents mAN is different in each case, as shown in Figure 2.

Some PFC units encode individual operations in a prepared sequence in a format where relative activation levels indicates the serial order in which operations will be executed (see Averbeck et al., 2002). Of these units, some have activity which changes dynamically during execution of a prepared sequence, being maximal before execution of the action they encode and being inhibited thereafter. Others are invariant during execution of a planned sequence. Units in the dynamic episodic buffer have the former property, and units in the static episodic buffer have the latter property.

Finally, some areas of PFC provide a medium in which alternative prepared sequences can compete against one another (Averbeck et al., 2006). The candidate episodes buffer acts as such a medium. Table 1 shows the five most active candidates in the candidate episodes buffer as a response to the presentation of DOG BALL and DOG BALL CAUSE fragments. ${ }^{2}$

[^230]
## Summary and discussion

This paper contains two main proposals. Most concretely, we propose a network model of WM for behavioural sequences. We also propose a more far-reaching idea: that episodes are represented in semantic WM as prepared behavioural sequences. Specifically, we propose our model of prepared sequences as a model of the episodic buffer argued for cogently by Baddeley (2000). We now assess these proposals.

WM for sequences There are numerous network models of WM for sequences. However, most of these are explicitly models of phonological WM. We follow Baddeley (2000) in distinguishing between phonological WM and WM for episodes. This means our model does not directly compete with the best-known models of WM for sequences, for instance Burgess and Hitch (1999). It does not have to reproduce the classic effects found in immediate recall of phonological sequences, such as primacy and recency effects. Empirically, our focus is on modelling the neural sequencepreparation mechanisms found in monkeys, which it does quite successfully. There are some computational models which propose the same mechanism both for phonological WM and prepared action sequences-in particular Rhodes et al. (2004). We certainly envisage similarities between the mechanisms subserving these tasks. (In particular they both appear to involve competitive queueing.) But our suggestion is that they are separate, although, as Baddeley suggests, there are links between them, which support sentence processing. We will discuss some ideas about these links below.

Episode representation As a model of representation of episodes in WM, our network is just a first step. An obvious issue for discussion is our localist representation of episodes in the candidate episodes buffer. Since episode representations can have other episode representations nested within them, it is clearly infeasible to have a single assembly in this medium for each possible episode. However, we should distinguish episode representations from sentence representations. Our conception of epsiodes as stored SM sequences means that there are several kinds of nestedness in sentences which we do not have to model declaratively. For instance, to model The dog [which chased Mary] barked we can initially rehearse just the matrix episode The dog barked: when $d o g$ is activated we can temporarily evoke the subordinate episode The dog chased Mary in the candidate episodes buffer, so it can be rehearsed, and then inhibit it, so the matrix episode once again becomes dominant. This device of interrupting processing is not available to schemes which represent episodes declaratively in a static pattern of neural activity: we see this as a strong advantage of representing episodes as sequences. Sequentially structured episode representations also permit an interesting representation of nested sentential complements; see Caza and Knott (2012).

[^231]Sentence processing As regards sentence processing, the network can be extended in several interesting directions. These all enlarge on Baddeley's (2000) proposal that sentence processing involves interactions between two separate WM buffers, one for phonological material and one for episodes.

There is a natural way of extending the network to support sentence generation. A detailed model of sentence generation incorporating the current model of WM episodes is given in Takac et al. (2012). In this model, generating a sentence involves replaying a WM episode stored as a prepared sequence, in a special mode where SM signals can trigger learned articulatory motor plans. During this replay process, an interesting mixture of sustained and transient signals is evoked: in particular, there are tonically active representations of each action in the planned sequence in the static episodic buffer throughout the replay process. These tonic representations permit a neat account of the extended syntactic domain of verbs. Verbs can appear at various different positions in the structure of a clause, and they can carry inflections signalling agreement with arguments at distant positions in the clause (for instance subjects). The neural basis for this non-locality is currently a complete mystery. But if sentences are produced by replaying a prepared SM routine, and if verbs and their inflections are produced from planned motor and attentional action representations which are tonically active during replay, we have a promising explanation of this non-locality: the semantic representations from which inflected verbs are generated are active throughout the generation process, and can be produced at any time.

The WM episode network also has interesting uses in models of sentence interpretation. Neural models of sentence interpretation take sequences of words as input, and use various types of recurrent network to produce output semantic representations. Such a network could deliver episode representations directly to the candidate episodes buffer. After training, this buffer would activate a distribution of possible sentence meanings and a winner could be picked. In our network, this winner could then be simulated as a SM sequence, in line with embodied theories of meaning.

## Acknowledgements

This research was supported by a VEGA grant (1/0439/11) and a SAIA travel grant (both for Martin Takac). We are grateful to Lubica Benuskova for helpful discussions.

## References

Abraham, W., Logan, B., Greenwood, J., and Dragunow, M. (2002). Induction and experience-dependent consolidation of stable long-term potentiation lasting months in the hippocampus. Journal of Neuroscience, 22, 9626-9634.
Averbeck, B., Chafee, M., Crowe, D., and Georgopoulos, A. (2002). Parallel processing of serial movements in prefrontal cortex. PNAS, 99(20), 13172-13177.
Averbeck, B., Sohn, J., and Lee, D. (2006). Activity in
prefrontal cortex during dynamic selection of action sequences. Nature Neurosceience, 9(2), 276-282.
Baddeley, A. (2000). The episodic buffer: A new component of working memory? TICS, 4(11), 417-423.
Baddeley, A. and Hitch, G. (1974). Working memory. In G. Bower, editor, The psychology of Learning and Motivation, pages 48-79. Academic Press.
Ballard, D., Hayhoe, M., Pook, P., and Rao, R. (1997). Deictic codes for the embodiment of cognition. Behavioral and Brain Sciences, 20(4), 723-767.
Barone, P. and Joseph, J.-P. (1989). Prefrontal cortex and spatial sequencing in macaque monkey. Experimental Brain Research, 78, 447-464.
Barsalou, L. (2008). Grounded cognition. Annual Review of Psychology, 59, 617-645.
Burgess, N. and Hitch, G. (1999). Memory for serial order: A network model of the phonological loop and its timing. Psychological Review, 106, 551-581.
Caza, G. and Knott, A. (2012). Pragmatic bootstrapping: A neural network model of vocabulary acquisition. Language Learning and Development, 8, 1-23.
Dowty, D. (1991). Thematic proto-roles and argument selection. Language, 67(3), 547-619.
Knott, A. (2012). Sensorimotor Cognition and Natural Language Syntax. MIT Press, Cambridge, MA.
Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. Biological Cybernetics, 43, 59-69.
Lee, A. and Wilson, M. (2002). Memory of sequential experience in the hippocampus during slow wave sleep. Neuron, 36, 1183-1194.
Levelt, W. (1989). Speaking: From Intention to Articulation. MIT Press, Cambridge, MA.
Mayberry, M. and Miikkulainen, R. (2008). Incremental nonmonotonic sentence interpretation through semantic selforganization. Technical Report AI08-12, Department of Computer Sciences, The University of Texas at Austin.
Rhodes, B., Bullock, D., Verwey, W., Averbeck, B., and Page, M. (2004). Learning and production of movement sequences: Behavioral, neurophysiological, and modeling perspectives. Human Movement Science, 23, 699-746.
Schwartz, B. and Evans, S. (2001). Episodic memory in primates. American Journal of Primatology, 55(2), 71-85.
Strickert, M. and Hammer, B. (2005). Merge SOM for temporal data. Neurocomputing, 64, 39-71.
Takac, M., Benuskova, L., and Knott, A. (2012). Mapping sensorimotor sequences to word sequences: A connectionist model of language acquisition and sentence generation. Cognition, 125, 288-308.
Wallenstein, G., Eichenbaum, H., and Hasselmo, M. (1998). The hippocampus as an associator of discontiguous events. Trends in Neurosciences, 21, 317-323.

# An Experimental Investigation of Adaptive Algorithm Understanding 

Kazunori Terada (terada@gifu-u.ac.jp)<br>Department of Information Science, Gifu University<br>1-1 Yanagido, Gifu, Japan 501-1193

Seiji Yamada (seiji@nii.ac.jp)<br>National Institute of Informatics / SOKENDAI / Tokyo Institute of Technology<br>2-1-2 Chiyoda, Tokyo, Japan 101-8430

Akira Ito (ai@gifu-u.ac.jp)<br>Department of Information Science, Gifu University<br>1-1 Yanagido, Gifu, Japan 501-1193


#### Abstract

There have been few studies on a cognitive model for algorithm understanding in a human-computer cooperative situation. In the present study, we conducted an experiment with participants to investigate the cognitive process of higher level abstraction (algorithm understanding) performed in a humancomputer collaboration task. The most recently used (MRU) algorithm, known to be one of the simplest adaptive algorithms, and probabilistic MRU algorithm were used to test the human capability to understand an algorithm. The experimental results showed that inductive reasoning in which participants observed the history of computer action, and they updated a statistical model while restricting their focus on a certain history with deteministic bias and Markov bias played key role to correctly understand the MRU algorithm. The results also showed that deductive reasoning was used to understand algorithms when participants rely on prior knowledge, and that there was a case in which the algorithm, even known to be the simplest one, was never understood.


Keywords: algorithm understanding; inductive reasoning; deductive reasoning; adaptive user interface;

## Introduction

The number of situations in which humans collaborate with computers has been increasing with the advance of information technology. Although user-adaptive systems that adapt to a user, including adaptive user interfaces, have been a main topic in the human-computer interaction community and artificial intelligence machine learning community (Findlater \& McGrenere, 2004; Oviatt, Swindells, \& Arthur, 2008; Bigdelou, Schwarz, \& Navab, 2012), an adequate design policy for implementing useful user-adaptive systems still remains unclear (Shneiderman \& Maes, 1997; Lavie \& Meyer, 2010; Gajos, Everitt, Tan, Czerwinski, \& Weld, 2008). Furthermore, there have been few studies on a cognitive model for algorithm understanding in the context of human-computer collaboration tasks.

In a human-human collaboration task, mutual intention understanding plays the key role in accomplishing successful work (Byrne \& Whiten, 1988; Call \& Tomasello, 2008). However, in a collaboration task with a computer, the abstraction level of behavior necessary to understand a collaborator's behavior is lower than that used in a human-human collaboration task (Dennett, 1987). Behavior abstraction in terms of
goal (intention) is not required in human-computer collaboration because a goal is given and explicitly shared with both a human and a computer. Instead, algorithm level abstraction is needed. In a human-computer collaboration task, understanding a computer's algorithms in order to accomplish the given goal is quite important because a human relies on the computer's underlying mechanisms in order to predict its behaviors and to adapt to it.

One way to predict the future behavior of a target is to use input-output association acquired on the basis of sequence learning (Clegg, DiGirolamo, \& Keele, 1998; Sun \& Giles, 2001a; Winkler, Denham, \& Nelken, 2009). In a typical sequence learning problem (Nissen \& Bullemer, 1987), humans learn a recurring loop of action sequences from given examples, and as a result, their reaction time for the given examples decreases. This learning is done both explicitly and implicitly (sensory-motor learning), and currently, implicit sequence learning is actively studied (Sun \& Giles, 2001b). The situation in which humans observe only the action sequences given to them is the same in both sequence learning and algorithm understanding. However, the learning target of algorithm understanding is procedures with variables that describe the internal states of computers, and this target is quite different from that of sequence learning (i.e., sequence patterns of values). Obviously, the number of hypotheses in algorithm understanding is far more than that in sequence learning, and this makes algorithm understanding very hard. Hence, algorithm understanding requires quite strong biases to find adequate algorithms. Another difference between understanding cooperative algorithms and sequence learning is the type of interactivity in the tasks. In algorithm understanding in a cooperative situation, a computer's behaviors change depending on the behaviors of humans because it adapts to them. In sequence learning, sequences are given to humans as physical stimuli.

The research objective of this study is to build a cognitive model to describe the human capability to understand computer algorithms in the context of a human-computer collaboration task. We introduce one of the simplest humancomputer collaboration tasks, in which a computer adapts to humans who are asked to try and understand the computer
algorithms. Concretely, we investigated how humans understand the most recently used (MRU) algorithm (Lee et al., 1999; Findlater \& McGrenere, 2004; Gajos et al., 2008). The MRU algorithm is well known to be one of the simplest adaptive algorithms in which a computer's current statement simply corresponds to the user's last one. Examples of the implementation of the MRU algorithm are the most recently used files (Amer \& Oommen, 2006), which lists the user's most recently accessed files in an application, and the most recently used menu (called adaptive menи (Arcuri, Coon, Johnson, Manning, \& Tilburg, 2000)), which lists the user's most recently used menu.

The MRU algorithm has succeeded in contributing to making useful interactive software that includes adaptive user interfaces (Findlater \& McGrenere, 2004). One reason is that it can be easily understood by users. If users can not find any meaning (regularity or rules for computer's behaviors) from a list in which the order of the items is frequently changed, the list causes the user stress. The reason the MRU algorithm is easily understood is that there are explicit descriptions of the algorithm, i.e., there may be a description such as "most recently used file." In this work, we investigate the human ability to understand an algorithm in a situation without such explicit knowledge.

One preferable explanation of algorithm understanding is induction because rule finding is considered to be an inductive process (Haverty, Koedinger, Klahr, \& Alibali, 2000; Simon \& Kotovsky, 1963; Verguts, Maris, \& Boeck, 2002; Schmid \& Kitzelmann, 2011). In general, induction needs to be done only with a small number of examples. It is hard to induce adequate rules with finite examples that can cover infinite facts because there is a huge number of hypotheses of rules that can be induced from the examples. Thus, we need heuristics (called inductive biases) to sufficiently restrict the hypothesis of rules for practical induction. In algorithm understanding, since humans have to induce computer algorithms only with tens of examples, we consider they have a strong bias for algorithm understanding. In this paper, to investigate human algorithm understanding, we hypothesize biases on algorithm understanding and verify them in experiments with participants.

## Cognitive Model of Adaptive Algorithm Understanding

Adaptive algorithm understanding is a subclass of algorithm understanding. An adaptation in human-computer interaction refers to a feature of algorithms in which strategies of a computer dynamically change according to user's input in order to pursue given goals. The goals refer not only corporation but also competition (Hampton, Bossaerts, \& O'Doherty, 2008). In the present study, we focus on a cooperative situation. We introduce a cooperative mark-matching game as a simplified and generalized task of human-computer adaptation in which a user adapts to a user-adaptive system.

## Cooperative Mark-Matching Game

The cooperative mark-matching game is a repeated game with two players. Each player has the same marks (e.g., $\boldsymbol{巾}$, $\diamond, \oslash$ ) and must secretly choose one of the marks. The players then reveal their own choices simultaneously. If the marks match each other, both players obtain a certain score, and if not, nobody obtains a score. In our experiments, the two players were a human and a user-adaptive system.
In a situation of the human-computer adaptation, a system predicts the user's next action (e.g., a menu item that will be chosen next by a user in an adaptive menu (Findlater \& McGrenere, 2004)) and adapts to him/her by modifying the user interface (e.g., changing the menu item positions (Findlater \& McGrenere, 2004)). If the prediction is correct (i.e., the two marks of the human and user-adaptive system matched in the game), the user and system obtain efficiency together. The number of the mark corresponds to the number of menu items in the adaptive menu. The key difference between a cooperative mark-matching game and human-computer adaptation with AUIs is that a user can freely choose his/her next action by him/herself in the game in contrast to the user's action sequence being determined to achieve a task with AUIs.

While the simplest strategy for a cooperative game is for participants in each trial to simply choose the action that in the recent past gave the most rewards (known as reinforcement learning), a more sophisticated strategy is to try to predict the system's next actions by taking into account a statistical model constructed on the basis of the history of prior actions. Studies on game theory (Fudenberg \& Levine, 1998)(Berger, 2005) and sequence learning (Sun \& Giles, 2001a) with an opponent player (a user-adaptive system) in a game situation suggest that opponent strategy is identified on the basis of a mixed strategy, which is defined as a probability distribution over the alternative actions available to each player.

## Statistical model

We hypothesize that, as mentioned earlier, a higher level abstraction, i.e., algorithm identification, for a computer's strategy is carried out on the basis of biases. We set the starting point of our discussion to statistics in which a human updates the conditional probability distribution of the system's next choice over time.

$$
\begin{equation*}
p\left(a_{t}^{s} \mid a_{t-1}^{s}, \cdots, a_{j}^{s}, a_{t-1}^{h}, \cdots, a_{k}^{h}\right) \tag{1}
\end{equation*}
$$

, where $a^{h}, a^{s} \in A$, and $A$ are available choices for both the system and human and $a_{t-1}^{s}, \cdots, a_{k}^{s}$ and $a_{t-1}^{h}, \cdots, a_{j}^{h}$ are the past choices of the system and human, respectively. Indices $j$ and $k$ denote the length of the history, which the human takes into account, and vary depending on focus. However, detecting the computer's algorithm on the basis of only observed behaviors is an ill-posed inverse problem because humans do not know how to restrict their focus to a certain history, and in addition, different strategies sometimes produce the same

Table 1: Conditional probability distributions correspond to most recently used and probabilistic most recently used algorithm

history. Thus, we consider that a human does sufficiently restricted exploration with inductive biases.

The MRU algorithm is formalized as the following distribution.

$$
\begin{equation*}
p\left(a_{t}^{s} \mid a_{t-1}^{h}\right) \tag{2}
\end{equation*}
$$

The actual distribution produced by the MRU algorithm in the cooperative mark-matching game is shown in Table 1(a). The system's choice $\left(a_{t}^{s}\right)$ depends only on the human's most recent choice ( $a_{t-1}^{h}$ ) and is independent from any other history of choices. If the human's most recent choice is heart, for example, the system's next choice will be heart, represented as $p\left(a_{t}^{s}=ऽ \mid a_{t-1}^{h}=\varnothing\right)=1$. Infinite numbers of trials are, theoretically, required to convince a human that the probability is 1 . Hence, one reasonable strategy for this problem is to use inductive biases to adequately control the inference process. As such inductive biases, we consider deterministic bias and Markov bias. If a human has a deterministic bias that assumes computer's behaviors are deterministic, not probabilistic, only one piece of evidence is necessary to estimate the probability distribution. Markov bias, in which the conditional probability distribution of the next choice depends only upon the present choice, not on the sequence of events, is also necessary to ignore any unnecessary history of choice.

## Experiments

We conducted an experiment with participants to investigate the cognitive process of higher level abstraction (algorithm identification) performed in the context of a human-computer cooperation task. The MRU algorithm and probabilistic MRU algorithm was used to test the human capability of algorithm understanding. Participants were asked to play a cooperative game with a computer, and after that they were asked to answer the computer's algorithm.

A 50-round repeated cooperative mark-matching game with different statistical profiles of the MRU algorithm was used. We used the following two conditions.
Deterministic (D) condition Computer's choice is completely the same as the human's most recent choice (deterministic MRU algorithm, see Table 1(a)).
Probabilistic (P) condition Although $90 \%$ of the computer's choices are the same as the human's most recent choices, $10 \%$ differs (probabilistic MRU algorithm). The


Figure 1: Interface of on-line experimental system: 1) history of both players' choices, 2) choice marks (marks are clickable), 3) round number and remaining time, 4) place for unveiling players' choice and scores for both players
actual distribution produced by the probabilistic MRU algorithm is shown in Table 1(b).

The P condition was prepared to contrast the effect of noise on the inductive reasoning performed to understand the MRU algorithm. In particular, we expected that the deterministic bias was strongly affected by the noise and performance deteriorated in the P condition. It was also expected that the score of those who participated in the P condition was at most $10 \%$ worse than that of the D condition if the participants merely estimated the probability distribution and did not use any biases to identify an algorithm.

## Experimental setup and measurement

The game was implemented with JavaScript and HTML and played in a Web browser (Firefox). Figure 1 shows the game interface. The computer's choices were automatically controlled by a JavaScript program. Participants were instructed to click the mark corresponding to his/her choice within 10 seconds for every round. Scores for both players were shown in the interface. The choices of the past five rounds for both players remained displayed so that the participant was able to recognize the computer's strategy.

A single-factor two-level between-subject experimental design was used. Fifty people ( 9 female) aged 19 to 47 (mean $=28$ ) recruited via direct e-mail participated in the experiment. All participants had moderate to high experience using computers. Participants were randomly assigned to either a deterministic or probabilistic condition. Participants were informed of an ostensible goal of the experiment - that the point of the experiment was to assess the usability of an on-line game system. They were also informed that "the computer was cooperative." Participants were told that they would win a PC gadget as a prize according to the score (under 20 points: around $\$ 5,21$ to 44 points: around $\$ 15,45$ to 50 : around \$30).

In the P condition, a 50 -round sequence with $10 \%$ random noise, which corresponds to 5 rounds in which MRU rules


Figure 2: Percentage of participants who won each round (solid line) and percentage of participants who started to take a "fixed choice strategy" (correct solution to the game) throughout the remaining rounds (dotted line)


Figure 3: Computer's algorithm identified by participants
are violated, is generated, and sequences that do not fit the following criteria are omitted: 1) errors do not appear in the first and last 5 rounds and 2) five errors appear within the remaining 40 rounds. The computer's choice for the first round was selected not to match the participant's choice in both conditions.
The outcomes of all 50 rounds were recorded. The round in which participants became aware of the correct solution to the game was identified by detecting the round in which participants started to continue to select the same mark throughout the remaining rounds. After the game, participants were asked to answer 7-point Likert scale questions, such as $Q$. Did the computer make its choices strategically?, and one openended question if participants gave a rating of 5 to 7 (positive) to this question - Describe the computer's strategy.

## Results

The average scores were $43.7(\mathrm{SD}=7.0)$ in the D condition and 31.4 ( $\mathrm{SD}=7.5$ ) in the P condition. ANOVA revealed that there was statistically significant difference $(F(1,48)=$ $33.99, p<0.01$ ) between the two conditions. The difference of the average scores between the two conditions was 12.3. A difference of more than $5(10 \%)$ indicates that participants used deterministic bias to accomplish the game. This gap is explained by the difference in the increasing rate of the winning percentage. While the winning percentage of the D condition rapidly reached a high value (e.g., $80 \%$ at the sixth round), that of the P condition slowly increased (e.g., $80 \%$ at the 35 th round). The slower increase of the winning
percentage in the P condition indicates that the $10 \%$ noise in the MRU algorithm caused the computer's algorithm to be difficult to identify and made the participants require longer rounds to identify it.

Figure 2 shows the percentages of the participants who won the round plotted against the round numbers (solid line). The dotted line in Figure 2 represents the percentage of participants who took a "fixed choice strategy," indicating the percentage of participants who became aware of the correct solution to the game. Note that the correct solution is found not only by identifying the MRU algorithm, but also by merely choosing the same mark without thought.

Figure 3 illustrates the computer's algorithm identified by participants. While $72 \%$ of participants in the D condition correctly identified the MRU algorithm after 50 rounds, only $52 \%$ in the P condition succeeded in identifying it. However, a chi-square test revealed that there was no statistically significant difference in the distribution of the identified algorithm between the two conditions $\left(\chi^{2}(4)=3.41, p=0.49\right)$.

## Discussions

In the present study, we investigated the human capability to understand the MRU algorithm. In particular, we expected that inductive biases such as deterministic and Markov bias are used to understand the algorithm. In the succeeding subsections, we will discuss whether these biases were applied to accomplish the game.

## Inductive algorithm understanding

The red dotted line in Figure 2 reveals that $60 \%$ of participants ( 15 participants) in the D condition found the correct solution to the game. The result of the questionnaire revealed that while 13 of the 15 participants inferred the computer's algorithm as the MRU, one inferred no strategy, and one inferred a fixed choice. A typical behavioral pattern for these kinds of participants is shown in Figure 4(a). They observed the history of the choices and might have inferred the MRU algorithm on the basis of the obtained statistical model. However, while detecting a statistical model of the computer's strategy essentially requires an infinite number of trials, they rapidly identified certain algorithms. One explanation for this rapid identification is the deterministic bias and Markov bias. If the algorithm was assumed to be deterministic, the participants did not need to take into account the six cases filled out as zero in Table 1(a) and required at least three trials to determine the computer's strategy. Without Markov bias, participants could not focus only on the one round past choice and required longer rounds.
The deterministic bias also accounts for the worse performance of those who participated in the P condition. If the participants merely estimated the probability distribution, as expected, an optimal strategy against a mixed strategy would have been taken, and performance would have been at most $10 \%$ worse than in the D condition.
The lowest score for all 50 participants was 19 , which was higher than the theoretically calculated score (16.67) when

#   

(a) Understanding algorithm on the basis of inductive reasoning (correct identification). After eight trials of active learning phase, the participant realized the algorithm was the MRU one.
(b) Understanding algorithm on the basis of inductive reasoning (wrong identification). The detected algorithm was "the computer increased the number of times by repeating the same choice."

(c) Understanding algorithm on the basis of deductive reasoning. The participants used a heuristic from the beginning: "Adaptive system $\Rightarrow$ MRU algorithm."


(d) The participant did not detect any algorithm.

Figure 4: Examples of typical behavioral pattern in the D condition. C : computer, H : human.
participants did not take any strategy, i.e., a random strategy. This implies that almost all of the participants arbitrarily attributed some kind of strategy to the computer's choice. In fact, the rules of the game allowed the participants to attribute strategies other than the MRU algorithm, such as "the computer simply selected the same mark" (fixed choice strategy) and "the computer changed its choice alternatively" or "the computer increased the number of times by repeating the same choice such as $\diamond \boldsymbol{\wedge} \cap \triangle \varnothing^{\prime \prime}$ (increasing number strategy), see Figure 4(b)). Three participants in the D condition answered that the computer's algorithm was "increasing number strategy." Interestingly, they did not aware that the timing to change the mark was determined by themselves. They completely unaware of the rule in which the computer changed its output according to their input.

## Deductive algorithm understanding

The results also indicated that some participants understood the algorithm on the basis of deductive reasoning. Sixteen percent of participants (four participants) in the D condition and four percent (one participant) in the $P$ condition fixed their choice in the first round and never changed during the game (see Figure 4(c)). Surprisingly, all of them described their identified computer algorithm as the MRU. The prior knowledge given to the participants in the instruction phase lead them to deduct the following logic:

$$
\begin{equation*}
\text { Adaptive system } \Rightarrow M R U \text { algorithm } \tag{3}
\end{equation*}
$$

In the instruction phase, participants were explicitly informed that the goal of the task was to get as much points as possible in cooperation with the partner computer. This top down adaptive bias might have enabled them to identify the algorithm immediately without exploring the computer's strategies. They might have logically inferred that the cooperative system acted adaptively to humans and that the most
efficient algorithm for human-computer cooperation was the MRU algorithm. In fact, their MRU algorithm hypothesis was confirmed by the computer's succeeding choice. The confirmation bias (Klayman \& Ha, 1987) was used to convince them that the computer used the MRU algorithm. They marked the highest score 49 (all participants were sure to lose the first round because of the game setting). There was no incentive to explore another strategy and gather evidence to test another hypothesis unless their hypothesis was violated because their goal was to get as many points as possible and not to detect the algorithm exactly. Indeed, while three participants in the P condition started to fix their choice in the first round, two of the three changed their choice after the noise pattern in the computer's choice appeared, indicating that their confirmation bias was destroyed by the noise (falsification).

## Algorithm detection fail

Surprisingly, even though the MRU has been supposed to be one of the most predictable adaptation algorithms, the result showed that two participants in the D condition and three in the P condition failed to identify any strategy in the 50 rounds (see Figure 4(d)). The visual cue shown in the history area in the game's interface might have been a strong cue indicating that the computer's choice was the same as participant's choice one round before. However, they could not detect the algorithm. Further investigation will be required to account for this failure.

## Summary

To the best of our knowledge, this is the first study to investigate the human capability to understand adaptive algorithm in a human-computer collaboration task. In the theoretical model of a human cognitive process for algorithm understanding, a user identifies a computer's algorithm by estimating the conditional probability distribution associated with a
particular strategy and restricting his/her focus on certain history data by using inductive biases. The most recently used (MRU) algorithm, known to be one of the simplest adaptive algorithms, was used to test the human capability to understand an algorithm. The probabilistic MRU algorithm was also used to contrast the effect of noise on the inductive reasoning performed to understand the MRU algorithm. The experimental results indicated that most participants correctly identified the MRU algorithm and used deteministic bias and Markov bias in their inductive reasoning for algorithm identification. The results also indicated that some participants understood the algorithm on the basis of deductive reasoning. Surprisingly, few participants failed to identify any algorithm within 50 rounds.

The present findings implies that designed behavior of computers is not necessarily understood correctly, suggesting that both an understandable algorithm and transparency of the internal state of a computer might be important for designing effective adaptive systems.

## Acknowledgments

This work was supported by MEXT KAKENHI Grant Number 24118703 .

## References

Amer, A., \& Oommen, B. J. (2006). Lists on lists: A framework for self-organizing lists in environments with locality of reference. In Proceedings of the 5th International Workshop on Experimental Algorithms (pp. 109-120).
Arcuri, M., Coon, T., Johnson, J., Manning, A., \& Tilburg, M. van. (2000). Adaptive menus. (US Patent: 6,121,968).

Berger, U. (2005). Fictitious play in $2 \times \mathrm{N}$ games. Journal of Economic Theory, 120, 139-154.
Bigdelou, A., Schwarz, L., \& Navab, N. (2012). An adaptive solution for intra-operative gesture-based human-machine interaction. In Proceedings of the 17th International Conference on Intelligent User Interfaces (pp. 75-84).
Byrne, R. W., \& Whiten, A. (1988). Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys, apes, and humans (R. W. Byrne \& A. Whiten, Eds.). Oxford Science Publications.
Call, J., \& Tomasello, M. (2008, May). Does the chimpanzee have a theory of mind? 30 years later. Trends in Cognitive Sciences, 12(5), 187-192.
Clegg, B. A., DiGirolamo, G. J., \& Keele, S. W. (1998). Sequence learning. Trends in Cognitive Sciences, 2(8), 275281.

Dennett, D. C. (1987). The intentional stance. Cambridge, Mass, Bradford Books/MIT Press.
Findlater, L., \& McGrenere, J. (2004). A comparison of static, adaptive, and adaptable menus. In Proceedings of the 11th international conference on intelligent user interfaces (pp. 89-96).
Fudenberg, D., \& Levine, D. K. (1998). The theory of learning in games. MIT Press, Cambridge, MA.

Gajos, K. Z., Everitt, K., Tan, D. S., Czerwinski, M., \& Weld, D. S. (2008). Predictability and accuracy in adaptive user interfaces. In Proceeding of the 26th annual sigchi conference on human factors in computing systems (pp. 12711274).

Hampton, A. N., Bossaerts, P., \& O'Doherty, J. P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. Proceedings of the National Academy of Sciences, 105(18), 6741-6746. (fictitious play, Reinforcement Learning)
Haverty, L. A., Koedinger, K. R., Klahr, D., \& Alibali, M. W. (2000). Solving inductive reasoning problems in mathematics: Not-so-trivial pursuit. Cognitive Science, 24(2), 249-298.
Klayman, J., \& Ha, Y. won. (1987). Confirmation, disconfirmation, and information in hypothesis testing. Psychological Review, 94(2), 211-228.
Lavie, T., \& Meyer, J. (2010). Benefits and costs of adaptive user interfaces. International Journal of Human-Computer Studies, 68, 508-524.
Lee, D., Choi, J., Kim, J.-H., Noh, S. H., Min, S. L., Cho, Y., et al. (1999). On the existence of a spectrum of policies that subsumes the least recently used (lru) and least frequently used (lfu) policies. In Proceedings of the 1999 acm sigmetrics international conference on measurement and modeling of computer systems (pp. 134-143). New York, NY, USA: ACM.
Nissen, M., \& Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. Cognitive psychology, 19(1), 1-32.
Oviatt, S., Swindells, C., \& Arthur, A. (2008). Implicit useradaptive system engagement in speech and pen interfaces. In Proceedings of the 26th annual sigchi conference on human factors in computing systems (pp. 969-978).
Schmid, U., \& Kitzelmann, E. (2011). Inductive rule learning on the knowledge level. Cognitive Systems Research, 12(34), 237-248.

Shneiderman, B., \& Maes, P. (1997). Direct manipulation vs. interface agents. Interactions, 4(6), 42-61.
Simon, H. A., \& Kotovsky, K. (1963). Human acquisition of concepts for sequential patterns. Psychological Review, 70(6), 534-546.
Sun, R., \& Giles, C. (Eds.). (2001b). Sequence learning: Paradigms, algorithms, and applications (Vol. 1828). Springer.
Sun, R., \& Giles, C. L. (2001a). Sequence learning: From recognition and prediction to sequential decision making. IEEE Intelligent Systems, 16(4), 67-70.
Verguts, T., Maris, E., \& Boeck, P. D. (2002). A dynamic model for rule induction tasks. Journal of Mathematical Psychology, 46(4), 455-485.
Winkler, I., Denham, S. L., \& Nelken, I. (2009). Modeling the auditory scene: predictive regularity representations and perceptual objects. Trends in Cognitive Sciences, 13(12), 532-540.

# A Discussion on the Consistency of Driving Behavior across Laboratory and Real Situational Studies 

Hitoshi Terai ${ }^{* 1,2}$ (terai@is.nagoya-u.ac.jp) Kazuhisa Miwa ${ }^{* 1}$ (miwa@is.nagoya-u.ac.jp) Hiroyuki Okuda*3(h_okuda@nuem.nagoya-u.ac.jp) Yuichi Tazaki*4 (tazaki@nuem.nagoya-u.ac.jp) Tatsuya Suzuki* ${ }^{* 4}$ (t_suzuki@nuem.nagoya-u.ac.jp) Kazuaki Kojima*5 (koj@aoni.waseda.jp) Junya Morita* ${ }^{* 6}$ (j-morita@jaist.ac.jp)<br>Akihiro Maehigashi ${ }^{* 1}$ (mhigashi@cog.human.nagoya-u.ac.jp)<br>Kazuya Takeda*1 (kazuya.takeda@nagoya-u.jp)<br>${ }^{* 1}$ Graduate School of Information Science, Nagoya University, Nagoya, Aichi, 464-8601 Japan<br>${ }^{* 2}$ CREST, Japan Science and Technology Agency, Chiyoda, Tokyo, 102-8666 Japan<br>${ }^{* 3}$ Green Mobility Collaborative Research Center, Nagoya University, Nagoya, Aichi, 464-8601 Japan<br>${ }^{* 4}$ Graduate School of Engineering, Nagoya University, Nagoya, Aichi, 464-8601 Japan<br>${ }^{* 5}$ Faculty of Human Sciences, Waseda University, Tokorozawa, Saitama, 359-1192 Japan<br>${ }^{* 6}$ School of Knowledge Science, Japan Advanced Institute of Science and Technology, Nomi, Ishikawa, 923-1211 Japan


#### Abstract

This study investigated the degrees of consistencies in driving behavior when operating a real system (real car), a virtual system (high fidelity driving simulator), and a laboratory system (computer driving game). The same tendency of behavioral consistencies was confirmed among the three systems: i.e., the steering operation demonstrated the highest behavioral consistencies, followed by the acceleration and braking operations, respectively. The individuality of driving behavior emerged more in the braking and acceleration operations than in the steering operation. The same tendency for behavioral consistencies of braking, acceleration, and steering operations was confirmed in each of the three systems.


Keywords: behavioral consistency; driving behavior; individual differences; virtual environments

## Introduction

In studies of human factors, analyses of human behavior are usually conducted in actual environments using observational methods. However, advances in computer technology are now facilitating experiments on human factors by using various simulators because they provide a convenient and safe method for assessing human behavior. Thus many studies about human behavior in serious situations that may lead to accidents have been performed, such as people driving cars, operating airplanes, controlling industrial plants (e.g., dos Santos et al., 2008; Kemeny, 2003; J. D. Lee et al., 2002; Metzger \& Parasuraman, 2001; Parasuraman et al., 1996; Wickens \& Alexander, 2009). Driving simulators in particular have played an important role in automobile human factors research for more than three decades. Various studies using driving simulators have examined not only basic characteristics of driving behavior but also applied investigations of those effects of drinking and aging that relate to social problems because using automobiles is a major part of our daily lives (e.g., H. C. Lee et al., 2003; Mets et al., 2011; Pradhan et al., 2005; Rizzo et al., 1997).

However, virtual systems cannot simulate real systems completely. Therefore, many researchers agree that an examination of their validity is a crucial component in any study. The validity of driving simulators has previously been evaluated through a comparison of behavior when driving real cars and simulators (e.g., Törnros, 1998; Godley et al., 2002; Underwood et al., 2011; Shechtman et al., 2009; Mayhew et al., 2011). Previous studies have discussed both commonalities and specificities in the distributions of specific errors or characteristics of specific behaviors when operating real and virtual systems. Such discussions have an essential assumption of the consistency of behavioral characteristics when driving vehicles. However, we do not know to what human driving behavior is consistent. In the present study, we examined behavioral consistency (BC) when driving vehicles on road and using simulators.

The purpose of this study is to reveal the degree of BC by analyzing three basic operations of driving behavior: braking, acceleration, and steering operations. First, we investigate the BCs for the three operations when driving a real car. Then, we study the BCs in two other types of systems: a virtual system as a high fidelity driving simulator and a laboratory system as a low fidelity driving simulator (similar to a computer driving game). The following outlines our basic strategies for the investigation.

Imagine a situation in which drivers repeatedly drive on a specific course. The BC within each participant shows the degree of consistency in individual behavior when repeatedly driving on the same course. We also calculate the BCs across participants, demonstrating the degree of consistency in the general characteristics of human behavior independent of each participant's individuality. We refer to the former as the intrapersonal BC and the later as the interpersonal BC.

In our analyses, the interpersonal BC is treated as the baseline because it reflects the generality of BCs across partici-
pants. The intrapersonal BC is predicted to exceed the interpersonal BC. In this study, by comparing the inter- and intrapersonal BCs, we attempt to answer the following two research questions.

RQ1 To what degree is driving behavior, characterized by the three basic operations of braking, acceleration, and steering, consistent across individuals in the real system? Is the tendency observed in the real system confirmed in the two types of simulation systems?

RQ2 To what degree is individual behavior more consistent than behavior across individuals in the real system? Is the greater consistency of individual behavior in the real system confirmed in the two simulator systems? In other words, to what degree are the intrapersonal BCs greater than the interpersonal BCs in each system?

## Multi-layered experimental platform

In this study, to determine BCs within various systems, we constructed an innovative experimental platform consisting of three different types of systems: the real system using an electric vehicle, the virtual system using a high fidelity driving simulator, and the laboratory system implemented as a driving game(Figure 1).

## The systems

Real system We used an instrumented vehicle called COMS from Toyota Auto Body as the real system (Figure 1(a)). We equipped COMS with various sensors to record participant behavior, car dynamics, and environmental data. For participant behavioral data, manipulations of the steering wheel and brake/acceleration pedals were recorded. The car dynamics data were obtained from speed, acceleration, and angular velocity triaxial sensors. These data were collected at 2000 Hz . Three video cameras were mounted on the COMS in three different positions: front, downward, and face views. The front view camera captured the road conditions. The downward view camera was directed at the road surface and recorded road tags that identified where and when COMS passed specific course points. The face view camera captured the participants' facial expressions and steering control. Time codes were synchronized with the logged sensor and video data.

Virtual system A vehicle motion simulator called carSim from Mechanical Simulation Corporation was used as the virtual system (Figure 1(b)). The virtual system shared many characteristics with the real system, such as the front field of view that was $180^{\circ}$ on three screens and the driver's cockpit with the same interior as a real car. The manipulations of the steering wheel and brake/acceleration pedals were recorded as participant behavioral data. These data were collected at 100 Hz .

Laboratory system In the laboratory system, stimuli were presented to the participants on a 21 -inch computer screen
similar to a typical laboratory setting (Figure 1(c)). The laboratory system was different from the other two systems in many ways. For example, the road configuration was shown from a top-down view and the vehicle controlled by the participants was depicted as a black dot. The car dynamics provide simple reactions for participant inputs. The participants controlled the black dot using a gaming pad controller. When the participants input right or left on the steering control, the dot moved in the corresponding direction by pixels on the basis of the input time. Furthermore, the accelerator/braking operations increased/decreased the dot velocity. The participant operation data were collected at 25 Hz .


Figure 1: Multi-layered experimental platform

## Driving course

The participants controlled their vehicles on an experimental driving course. The course consisted of three physical configurations: sharp curves, gentle curves, and straight lines (Fig-


Figure 2: Overview of the driving course
ure 2). The driving courses used for each vehicle were similar and based on the vehicle's size.

## Method

## Participants

Study participants included twenty-one adults (11 males and 10 females) whose ages ranged from 31 to 55 (mean $=49$, SD $=3.0$ ). For the study to capture stable vehicle control, they were required to have over ten years of driving experience and currently drive a car more than ten days a month.

## Task

The experimental task assigned to the participants was to drive the vehicles toward the finish line using each system. They were instructed to drive as rapidly as possible and improve their lap times across the trials while maintaining driving safety.

## Procedure

The participants engaged in the task using each system as a within-participants design. The order of the experiments was counterbalanced between participants whenever possible.

For each system, the participants were involved in a practice and an experimental session. The practice session comprised of eight trials and the experimental sessions had six trials.

## Data treatment

In this study, we analyzed the BCs quantitatively. We defined the BCs of the braking, acceleration, and steering operations as similarities between feature vectors of each operation. In the real system, behavioral data of two participants were treated as missing values because of equipment trouble. In the virtual system, three participants could not participate for personal reasons. Furthermore, in the experimental session using the virtual system, all trials of two participants and four trials of four participants were treated as missing values due to 3D sickness.
Feature vectors Here, we summarize the definitions of feature vectors. For example, the feature vector of a braking operation is calculated as follows.
(1) The time-series data of a braking operation in a trial were divided into 26 sections. Each section corresponded to the region between two pairs of adjoining red pylons (see Figure 2 ).


Figure 3: Examples of feature vectors in the real system
(2) The average amount of a braking operation in each section was calculated.
(3) The series of 26 data points corresponded to a feature vector of a braking operation in each trial (see examples in Figure 3 (a)).

We calculated 6 vectorial data for each operation from all participants.
Behavioral Consistency as Similarity between Feature Vectors The BCs in each operation were calculated as an average of the similarity between two feature vectors using Pearson's product-moment correlation coefficient (Expression 1). The vector components of x and y in Expression 1 correspond to 26 data points each, as shown in Figure 3. Combinations of the feature vectors x and y are as follows from the viewpoint of the participant factor.

$$
S_{p}(\mathbf{x}, \mathbf{y})=\frac{\sum_{i=1}^{26}\left(\begin{array}{ll}
x_{i} & \bar{x}
\end{array}\right)\left(\begin{array}{ll}
y_{i} & \bar{y}
\end{array}\right)}{\sqrt{\sum_{i=1}^{26}\left(\begin{array}{ll}
x_{i} & \bar{x}
\end{array}\right)^{2}} \sqrt{\sum_{i=1}^{26}\left(\begin{array}{ll}
y_{i} & \bar{y} \tag{1}
\end{array}\right)^{2}}}
$$

Interpersonal BCs To discuss RQ1, the interpersonal BCs within each system were calculated to determine to what degree driving behaviors are consistent across individuals in the real system and whether such tendency observed in this system is confirmed in the other two systems. Specifically, first, one participant (Participant 1) was selected, and the correlation coefficients were calculated between the feature vectors of Participant 1 and those of another participant (Participant 2). Each had six feature vectors in each of the three operations of braking, acceleration, and steering; therefore, 18 (= ( $6 \times 6$ )/2) combinations were considered for the calculation. The average of the correlation coefficients among the 18 combinations was calculated. Second, in a similar manner, by repeating the calculation of the average of correlation coefficients between Participant 1 and the others, the average amount, defined as the correlation coefficient of Participant 1, was calculated. Finally, the interpersonal BCs within each system were calculated, defined as the average of the correlation coefficients of all participants (Participants 1-21).

Intrapersonal BCs To discuss RQ2, we calculated the intrapersonal BCs within each system to determine the degree to which individual behavior is more consistent than behavior across individuals in the real system and whether the greater consistency of individual behavior in this system is confirmed in the other two systems. Specifically, the correlation coefficients of one participant were calculated using 15 (= (6 x $5) / 2$ ) combinations and the average of the correlation coefficients among the 15 combinations was calculated. The intrapersonal BCs within each system were calculated, defined as the average of the correlation coefficients of all participants (Participants 1-21).

## Results

## Behavioral consistencies within the real system

Figure 4 shows the results of the inter- and intrapersonal BCs when using the real system. A two-way within-participants ANOVA for the operations (braking, acceleration, steering) and participants (interpersonal, intrapersonal) factors showed significant main effects of the operation and participant factors $(F(2,36)=76.35, p<.01 ; F(1,18)=14.47, p<.01$; respectively). Moreover, a significant interaction was noted between these factors $(F(2,36)=9.90, p<.01)$. The detailed results of the simple main effects tests are presented in Figure 4.

These results are summarized as follows: (1) In the interpersonal BCs, significant differences were found among the three operations, with the interpersonal BC of the steering operation as the highest, followed by those for acceleration and braking operations, respectively. (2) In the intrapersonal

BCs , this tendency was confirmed, and the intrapersonal BCs of the braking and acceleration operations were higher than their interpersonal BCs.


Figure 4: Behavioral consistencies within the real system


Figure 5: Behavioral consistencies within the virtual system


Figure 6: Behavioral consistencies within the laboratory system

## Behavioral consistencies within the virtual and laboratory systems

Figures 5 and 6 illustrate the results of the inter- and intrapersonal BCs when using the virtual and laboratory systems, respectively. The two-way ANOVAs showed significant main effects for the operations and participants factors in each system (virtual: $F(2,30)=103.71, p<$ $.01 ; F(1,15)=32.31, p<.01$, respectively, and laboratory: $F(2,40)=61.56, p<.01 ; F(1,20)=24.37, p<.01$, respectively). Moreover, significant interactions were observed between these factors (virtual: $F(2,36)=9.90, p<.01$, laboratory: $F(2,40)=7.97, p<.01)$. The results of the simple main effect tests are shown in Figures 5 and 6.

These results for the virtual and laboratory systems were similar to those for the real system. In the interpersonal BCs, significant differences were noted among the three operations, with the interpersonal BC of the steering operation being the highest, followed by those of the acceleration and braking operations, respectively. This tendency was confirmed in the intrapersonal BCs, with the braking and acceleration operations higher than their interpersonal BCs. Only in the laboratory system was a significant difference found between the inter- and intrapersonal BCs of the steering operation. However, the effect size (Cohen's $d$ ) in the steering operation was relatively smaller than in the brake and acceleration operation (braking: 0.89 , acceleration: 0.78 , steering: $0.57)$.

## Discussion

In this study, we constructed a multi-layered experimental platform to determine the BCs of driving behavior on the bases of two factors: the operations (braking, acceleration, steering) and participants (interpersonal, intrapersonal).

In this section, we summarize the results of the experiments from the viewpoint of each research question and then discuss them.

## Summary of Experimental Results

RQ1 asks to what degree driving behaviors are consistent across individuals in the real system and whether such a tendency is confirmed in the two different simulation systems. The results indicate that in the real system, the interpersonal BC of the steering operation was the highest, followed by those of the acceleration and braking operations, respectively This tendency was confirmed in the virtual and laboratory systems.

RQ2 is as follows: To what degree is individual behavior more consistent than behavior across individuals in the real system, and is the greater consistency of individual behavior in the real system confirmed in the other two systems? The analyses demonstrate that the intrapersonal BCs of both the braking and acceleration operations were lesser than that of the steering operation but they were higher than the interpersonal BCs for each system.

## Environmental Constraints

Experiment results reveal that the interpersonal BCs were different among the three operations in all systems: the interpersonal BC of the steering operation was the highest, followed by those of the acceleration and braking operations, respectively. This result suggests that each operation is regulated by different environmental constraints.
The higher environmental constraint on the steering operation than on the braking and acceleration operations might be caused by the arrangement of the driving course. Constraints based on driving course are recognized not only in the experimental setting but also in our daily driving situations. Our steering operations are strictly regulated by road configurations, whereas both acceleration and braking operations have high flexibility. That is, we usually do not out of traffic lanes, whereas the gas pedal or brakes can be used comparatively freely.

Additionally, there might be an interactive relation between the braking and acceleration operations. In some literature regarding the computational model of driver behavior based on cognitive architecture, the manipulation of vehicle controls has been defined as consisting of both lateral and longitudinal controls (e.g., Salvucci, 2006). The longitudinal control, or speed control, is achieved through coordination between the braking and acceleration operations, whereas lateral control is achieved by the steering operation. The mutually dependent relation between the braking and acceleration operations leads to an increase in their degrees of freedom. As a result, the BCs of the braking and acceleration operations might decrease more than that of the steering operation. Moreover, the velocity of the vehicle was mainly controlled by the gas pedal and not by the braking operation, causing different BCs for the braking and acceleration operations. In fact, as seen in the examples of feature vectors presented in Figure 3 (a) and (b), the frequency of the braking operation was substantially lower than that of the acceleration operation. Even though the participants typically controlled the vehicle velocity by using the gas pedal, in some accidental situations, they also had to press the brake to reduce the speed. As a result, the BC of the braking operation was lower than that of the acceleration operation.

## Individuality and Behavioral Consistencies

In studies of driving behavior, an important research topic has been to identify the uniqueness of individual driving behavior in order to develop intelligent driver assistance systems customized for individual drivers. More specifically, identifying an individual's deviation from ideal behavior leads to predicting accidents and possibly preventing them (e.g., Igarashi et al., 2004; Wakita et al., 2005; Okuda et al., 2009).
The results of our experiments imply that the braking and acceleration operations are useful measures for identifying individual driver behavior because substantial differences were noted between the inter- and intrapersonal BCs. On the other hand, only a small difference was found in the steering op-
eration because this operation was strongly regulated by the environmental constraints. Our experiments suggest that considering the generality and individuality of the environmental constraints in each operation is important when using behavioral data as personally identifying information.

## Conclusion

In this research, we discussed the behavioral consistencies (BCs) within multiple systems-real, virtual, and laboratory systems-on a multi-layered experimental platform. The results showed that the BCs of the steering operation were the highest, followed by those of the acceleration and braking operations, respectively. The intrapersonal BCs (BCs within individuals) of the braking and acceleration operations were higher than the interpersonal BCs (BCs across individuals) in all systems. Further, this tendency was consistent in all three systems. In this paper, we discussed the behavioral consistencies within each system and the similarity of their tendencies among the three different types of systems. These findings lead to another research question: To what degree is human behavior similar across the different types of systems when comparing behavioral characteristics directly? This question can be investigated in future studies.

## Acknowledgments

This research was conducted as a part of "Modeling and Detecting Overtrust from Behavior Signals," a study led by Kazuya Takeda. This took place within the "Creation of Human-Harmonized Information Technology for Convivial Society" project under the direction of Yoh'ichi Tohkura, Research Supervisor in the Core Research of Evolutional Science and Technology (CREST) program at the Japan Science and Technology Agency (JST).

## References

dos Santos, I. J. A. L., Teixeira, D. V., Ferraz, F. T., \& Carvalho, P. V. R. (2008). The use of a simulator to include human factors issues in the interface design of a nuclear power plant control room. Journal of Loss Prevention in the Process Industries, 21(3), 227-238.
Godley, S. T., Triggs, T. J., \& Fildes, B. N. (2002). Driving simulator validation for speed research. Accident Analysis \& Prevention, 34(5).
Igarashi, K., Miyajima, C., Itou, K., Takeda, K., Itakura, F., \& Abut, H. (2004). Biometric identification using driving behavioral signals. In 2004 ieee international conference on multimedia and expo (icme) (pp. 65-68). IEEE.
Kemeny, A. (2003). Evaluating perception in driving simulation experiments. Trends in Cognitive Sciences, 7(1), 31-37.
Lee, H. C., Cameron, D., \& Lee, A. H. (2003). Assessing the driving performance of older adult drivers: on-road versus simulated driving. Accident Analysis \& Prevention, 35, 797-803.

Lee, J. D., McGehee, D. V., Brown, T. L., \& Reyes, M. L. (2002). Collision warning timing, driver distraction, and driver response to imminent rear-end collisions in a highfidelity driving simulator. Human Factors, 44(2), 314334.

Mayhew, D. R., Simpson, H. M., Wood, K. M., \& Lonero, L. (2011). On-road and simulated driving: Concurrent and discriminant validation. Journal of Safety and Research, 42.

Mets, M. A. J., Kuipers, E., de Senerpont Domis, L. M., Leenders, M., Olivier, B., \& Verster, J. C. (2011). Effects of alcohol on highway driving in the STISIM driving simulator. Human Psychopharmacology, 26, 434-439.
Metzger, U., \& Parasuraman, R. (2001). The role of the air traffic controller in future air traffic management: An empirical study of active control versus passive monitoring. Human Factors, 43(4), 519-528.
Okuda, H., Suzuki, T., Nakano, A., Inagaki, S., \& Hayakawa, S. (2009). Multi-hierarchical modeling of driving behavior using dynamics-based mode segmentation. IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences, E92-A(11), 2763-2771.
Parasuraman, R., Mouloua, M., \& Molloy, R. (1996). Effects of adaptive task allocation on monitoring of automated systems. Human Factors, 38(4), 665-679.
Pradhan, A. K., Hammel, K. R., DeRamus, R., Pollatsek, A., Noyce, D. A., \& Fisher, D. L. (2005). Using eye movements to evaluate effects of driver age on risk perception in a driving simulator. Human Factors, 47(4), 840-852.
Rizzo, M., Reinach, S., McGehee, D., \& Dawson, J. (1997). Simulated car crashes and crash predictors in drivers with alzheimer disease. Archives of Neurology, 54(5), 545-551.
Salvucci, D. D. (2006). Modeling driver behavior in a cognitive architecture. Human Factors, 48(2), 362-380.
Shechtman, O., Classen, S., Awadzi, K., \& Mann, W. (2009). Comparison of driving errors between on-the-road and simulated driving assessment: A validation study. Traffic Injury Prevention, 10(4), 379-385.
Törnros, J. (1998). Driving behaviour in a real and a simulated road tunnel—a validation study. Accident Analysis \& Prevention, 30(4), 497-503.
Underwood, G., Crundall, D., \& Chapman, P. (2011). Driving simulator validation with hazard perception. Transportation Research Part F: Traffic Psychology and Behaviour, 14(6), 435-446.
Wakita, T., Ozawa, K., Miyajima, C., \& Takeda, K. (2005). Parametric versus non-parametric models of driving behavior signals for driver identification. In Audio- and video-based biometric person authentication (pp. 739747). Berlin, Heidelberg: Springer Berlin Heidelberg.

Wickens, C. D., \& Alexander, A. L. (2009). Attentional tunneling and task management in synthetic vision displays. The International Journal of Aviation Psychology, 19(2), 182-199.

# Is Lexical Access Driven by Temporal Order or Perceptual Salience? Evidence from British Sign Language 

Robin L. Thompson (robin.thompson@ucl.ac.uk)<br>David P. Vinson (d.vinson@ucl.ac.uk)<br>Neil Fox (neil.fox@ucl.ac.uk)<br>Gabriella Vigliocco (g.vigliocco@ucl.ac.uk)<br>Deafness, Cognition and Language Research Centre, Department of Cognitive, Perceptual and Brain Sciences<br>University College London, 26 Bedford Way, London, WC1H 0AP, UK


#### Abstract

While processing spoken language, people look towards relevant objects, and the time course of their gaze(s) can inform us about online language processing (Tanenhaus et al, 1995). Here, we investigate lexical recognition in British Sign Language (BSL) using a visual world paradigm, the first such study using a signed language. Comprehension of spoken words and signs could be driven by temporal constraints regardless of modality ("first in, first processed"), or by perceptual salience which differs for speech (auditorialy perceived) and sign (visually perceived). Deaf BSL signers looked more often to semantically related distracter pictures than to unrelated pictures, replicating studies using acoustically-presented speech. For phonologically related pictures, gaze increased only for those sharing visually salient phonological features (i.e., location and movement features). Results are discussed in the context of language processing in different modalities. Overall, we conclude that lexical processing for both speech and sign is likely driven by perceptual salience and that potential differences in processing emerge from differences between visual and auditory systems.


Keywords: lexical access; sign language; semantics, phonology, visual world; modality

## Introduction

General theories of language processing have developed on the basis of extensive data from spoken, but not signed languages, making it impossible to tease apart those aspects of language processing that are truly general from those dependent on the oral-aural language modality. While spoken language processing happens through aural perception of sounds, sign language processing occurs through visual perception which allows for more simultaneous input of information; spoken languages make use of mouth and vocal tract, while signed languages use slower manual articulators (hands, as well as eyes, mouth and body). An understanding of the processing differences that arise from these differing language modalities is critical for understanding the interaction of language processing with other cognitive systems such as perception and action. Here we take advantage of these physical differences in language processing for signed languages compared to spoken languages to investigate the nature of lexical processing and lexical access.

For spoken languages, it is generally uncontroversial that information is processed almost immediately as it comes in (e.g., Rayner \& Clifton, 2009). Such incremental moment-by-moment language processing is likely necessary to keep up with the incredibly fast rate of speech input (estimated to be between 150-190 words per minute, Marslen-Wilson, 1973). However, during incremental processing listeners, processing even a single word, are faced with many possible alternatives that match the current acoustic-phonetic input. Empirical evidence suggests that instead of waiting until temporary ambiguities are resolved, partial activation of possible words (i.e., lexical competitors) that match current phonological information proceeds, with potential words being eliminated across time as more information becomes available (e.g., McClelland and Elman, 1986; Gaskell \& Marslen-Wilson, 1997).

Evidence for incremental activation of lexical competitors during spoken language processing comes from the "visual world" paradigm (language presented simultaneously with related pictures; Allopena, Magnuson, \& Tanenhaus, 1998; Altman \& Kamide, 2004; Huettig \& Altmann, 2005; Yee \& Sedivy, 2006). For example, in Allopena et al. (1998), subjects heard an utterance like "Pick up the beaker" while viewing a display with four pictures including: 1) an object matching the noun (the target; e.g. "beaker"), 2) an object with a name beginning with the same phoneme (e.g. "beetle"), 3) an object with a name sharing the same rhyme (e.g., "speaker") and, 4) an unrelated object (e.g., carriage). The probability of fixating the target and onset competitor were identical immediately after word onset (when the two could not be distinguished from each other), and fixations to these picture types were higher than fixations to the rhyme or unrelated competitors. Immediately after reaching a phoneme differentiating the target and onset competitor, the probability of fixating the target rose sharply while the probability of fixating the related competitor fell. A weaker, but significant effect was also observed for rhyme competitors compared to unrelated competitors, indicating that activation is not restricted to words sharing onsets but is continuous (see for example McClelland and Elman, 1986).

A question of interest, then, is why words that share onsets make the strongest lexical competitors. One possibility is that strong activation of onset competitors compared to word rhymes is due to temporal considerations: i.e., word onsets occur earlier in time. This view about the
activation of onset competitors can be called a 'first in, first processed' account. However, onsets also tend to be salient, particularly in languages such as English (used in the majority of visual world studies) in which stress has the effect of lengthening the first syllable as well as adding both intensity and pitch change: all of which serve to make the first part of a word more salient. Evidence that stress is important to lexical access comes from Reinisch, Jesse, and McQueen (2010). In a visual world study they found that participants use lexical stress information to direct eye gaze such that upon hearing a word with initial stress (e.g., octopus) fixations on printed target words with first-syllable stress (e.g., octopus) were more frequent than fixations on differently stressed competitors (e.g., October, with stress on the second syllable). Thus, an alternate account of the strong activation of onset competitors observed in visual world studies is that word onsets are the most auditorily salient part of a word and that auditory salience drives lexical access for processing efficiency. However, because spoken word onsets tend to be both temporally early and auditorily salient, it is difficult to tease apart these alternate accounts based on previous studies.

Interestingly, unlike spoken words, for visually processed signs there is evidence that the phonological features that form the onset of a sign (i.e., the first features to be formed as a sign moves through time) may not coincide with the most visually salient features (i.e., the features that can be seen most easily, for example, under visually noisy conditions). Just as in spoken languages, signed languages have sub-lexical units (phonological features) that combine in rule-governed ways to form words/signs. Signs are made up of phonological features from three major parameters (handshape, movement, and location [place of articulation]; see Sandler \& Lillo-Martin, 2006 for discussion, and Figure 1 for examples of signs sharing these features). In terms of sign onsets, results from early gating studies (single frame presentation of a sign, with subsequent presentations increasing in length; Grosjean, 1981, Emmorey \& Corina, 1990) suggest that handshape and location features are recognized first across time. In Emmorey \& Corina (1990) subjects' initial responses tended to share the handshape and location of the target sign but differed in movement features. Once the movement of the sign was identified, the target sign also tended to be identified. The authors suggest that lexical recognition in a signed language is a two-stage process such that handshape and location are identified almost from the start of the sign (i.e., form the onset of the sign) followed by movement which coincides with sign recognition.

In terms of sign salience, Corina \& Hildebrandt (2002) used a sign similarity judgement task and found that subjects preferred to pair non-signs with other non-signs sharing location and movement features more frequently than pairing non-signs with matching handshape and location features, or handshape and movement features, suggesting that they are paying attention to these feature pairings. Further support for the salience of movement and location features is found in Corina \& Knapp (2006) who
used a picture sign interference task (subjects named a picture in ASL while trying to ignore a superimposed image of a related distracter) and found that distracter signs sharing both movement and location with the target sign resulted in significant facilitation effects at all stimulus onset asynchronies ( $-130,0,130 \mathrm{~ms}$ ), while signs sharing handshape and location, or handshape and movement features did not affect picture naming.


Figure 1: Examples of phonological minimal pairs in BSL. Top: car and robot share location and movement (up and down) parameters, but differ in handshape. Middle: saxophone and computer share handshape and movement (finger wiggle) features, but differ in location. Bottom: mouse and nose share handshape and location features, but differ in movement (mouse, with a twisting movement and nose with a tapping movement).

While location features are available early in sign perception, movement features only emerge later and are therefore crucial in teasing apart whether lexical access (at least for signs) is driven by temporal constraints or by perceptual salience. If temporal constraints drive lexical access in sign, signers should pay attention to handshape and location features which are available at the start of a sign and ignore movement features which emerge later. Movement features have been argued to be the most sonorous or salient part of a sign (Perlmutter, 1992). Thus,
if perceptual salience is instead key to lexical access, then signers may pay attention to movement features.

Here we investigate lexical recognition in BSL using a visual world paradigm and asking whether or not the nature of access in a dynamic visual language is also incremental and graded with alternate possible words considered simultaneously over the time-course of processingFurther, we consider the nature of activation of lexical competitors (if any) and whether sign access supports a first in, first processed pattern, or a pattern driven by visual salience.

We include two critical conditions. First, a semantic condition will determine if a visual world paradigm can be a successful methodology using sign language which must be presented visually. Previously subjects have been found to look towards semantically related competitor pictures during spoken language visual world studies (Huettig, \& Altmann, 2005). Here we explore whether eye movements are drawn to a semantically related object in a signed visual world in the absence of a phonological or visual relationship. The Visual World paradigm has never been used with sign language stimuli and a semantic condition (see methods) serves as our test case, under the assumption that semantic relationships should hold regardless of language. If the visual world methodology is successful with BSL, we should expect subjects to look more frequently to distracter pictures that are semantically related to a given target sign. Secondly, we examine the nature of sign recognition in real time using pictures that have phonologically related signs. If temporal information is most important, and signers process information primarily through sequential, incremental, first-in first-processed order, then signs sharing handshape and location features should be particularly salient for them. Alternatively, if perceptual salience is more relevant then signers may instead look more frequently to distracters that share movement and location features.

## Method

## Subjects

24 Deaf signers ( 13 women, 11 men, mean age 34.8 ) were recruited from deaf communities in England and took part in the experiment. Of these, eleven were native signers (born to deaf signing parents), four began signing by the age of five (early signers) and 9 learned BSL after age five. All subjects use BSL as their preferred and primary language.

## Materials

For each trial, four pictures of objects were presented simultaneously with a centrally located video clip (see Figure 2). In each video clip, a native BSL signer produced the carrier phrase, "I see...", followed by the target sign. Subjects were asked to indicate (with button press, "yes" or "no") as quickly and accurately as possible whether the target BSL sign matched one of the pictures. "Target Present" trials ( $\mathrm{n}=79$ ) in which a picture of the target sign was present constituted our fillers. On critical "Target

Absent" trials ( $\mathrm{n}=28$ ), three unrelated distractor pictures with no semantic, phonological or visual relationship to the target sign were presented along with a related distracter picture. Related distracter pictures had signs that were either semantically (e.g., target: banana, distracter: strawberry, target: zipper, distracter: button) or phonologically related to the target. Phonologically related pictures were minimal pairs that shared two out of three parameters (see Figure 1 for examples). Semantically related distracter pictures were not phonologically related to the target, and phonologically related pictures were not semantically related.




Figure 2: Example of a single trial. Areas of interest for gaze analyses were set directly around the ( $250 \times 250$ pixels) pictures and the ( $320 \times 240$ pixels) video.

## Procedure

After giving consent to participate, subjects were presented with video-recorded instructions in BSL (signed by N.F., a native BSL signer) and invited to ask clarification questions. Subjects were then fitted with a head-mounted eye-tracker (SR Research, EyeLink II) and initial calibration was performed (9 fixation points). Subjects were seated 50 cm from the monitor with the tracker positioned in front of the right eye. There were four practice trials before the experiment began. Another calibration check was performed after these practice items and then again after every 36 trials (the final set had only 35 trials), at which time subjects took a self-paced break (total 107 trials, 3 sets). Additionally, drift correction on a single centrally located fixation point was performed at the start of each trial. Responses were recorded using a hand-held joypad with buttons that can be located tactilely without the need to look at keys. The entire experiment (with instructions and calibration) took approximately 20 minutes to complete. In order to ensure that all pictures were familiar to the subjects as well as to obtain naming data, subjects named all of the pictures used in the visual world experiment before we began.
The location of the pictures was balanced so that each picture type (related distracter, unrelated distracter [filler]) occurred a roughly equal number of times in each location within a given condition. Additionally, we created two sets of stimuli such that half the subjects saw any one picture in one location and half of the subjects saw it in a different location. The order of trial presentation was randomized
throughout. Pictures were presented simultaneously with the sign video.

## Results

First we analyzed signs produced during picture naming to ensure that signs for target and related pictures in the phonological conditions were indeed phonologically related in subjects' lexicons. Individual trials were excluded when subjects produced a sign (for either target or related pictures) that did not have the intended phonological relationship (6\%). Error trials, in which participants mistakenly indicated that the target sign matched a picture on the screen $(12.4 \%)$ were also excluded from analyses of response latencies and eye gaze. The number of trials by condition along with average correct response latencies and percent of correct answers across different conditions are reported in Table 1. A one-way repeated measures ANOVA by subjects revealed no significant differences for accuracy between conditions: $\mathrm{F}(3,69)=1.686, \mathrm{p}=.178$. However, a significant difference was found between conditions for response latencies $\left(F(3,66)^{1}=3.202, p=.029\right)$. Post-hoc tests revealed that responses were slower for handshapemovement trials than other conditions.

Table 1. Average correct response time (standard deviation by subjects in brackets) and percent correct as a function of relatedness type. Sem: related picture sharing a semantic relationship to the target; HS-MV: related picture sharing the handshape and movement of the target sign; LOC-MV: sharing the location and movement of the target sign; HSLOC: sharing both the handshape and location of the target sign.

|  | Items | $\boldsymbol{R T}$ (SD) | \%Correct |
| :--- | :--- | :---: | :---: |
| SEM | $\mathrm{n}=11$ | $2792(462)$ | 88.3 |
| LOC-MV | $\mathrm{n}=5$ | $2730(421)$ | 91.1 |
| HS-MV | $\mathrm{n}=6$ | $2887(421)$ | 87.9 |
| HS-LOC | $\mathrm{n}=6$ | $2662(446)$ | 83.2 |

Five areas of interest within each time period were identified: the location of the signer in the middle of the screen (displayed as video), and one corresponding to each of the pictures displayed (coded as target, related, unrelated and matching in size to the actual pictures). The dependent measure of interest was dwell time (summed gaze duration in a given area, measured in milliseconds). Not surprisingly, across all trial types, gaze was primarily directed to the signer in the video ( $\mathrm{M}=86.9 \%$ ). This led to fewer looks towards pictures than would be expected in a study with auditory stimuli, so we started with a broad analysis. Specifically, for each trial, we identified two time windows. The early period, began at the start of the trial and ended when the carrier phrase "I see..." finished. Because the

[^232]target sign was not yet produced during the early period, gaze could not yet be informed by the target. The late period was defined as the period from the start of the target sign until the button was pressed. Gaze during the late period should provide information about processing of the target sign.
In the first set of gaze analyses across the different pictures, we tested whether subjects looked longer at related pictures than unrelated pictures in the late time period, once the meaning of the target sign could be processed. We conducted hierarchical linear regressions, treating subjects and target signs as random effects, including picture relatedness (considering only related vs. unrelated pictures) and time period (early vs. late) as predictors, and dwell time (in milliseconds) as the dependent measure. Separate models were fit for each relatedness condition (semantic, location-movement, handshape-movement, handshapelocation) ${ }^{2}$. Across all conditions there was a main effect of time period indicating longer gaze overall in the late period: semantic (95\% CI [183.7, 221.1], $p_{M C M C}<.001$ ); locationmovement ( $95 \%$ CI [134.0, 199.3], $p_{\text {МСМС }}<.001$ ); handshape-movement ( $95 \%$ CI [135.8, 190.0], $p_{\text {MCMC }}$ $<.001$ ); and handshape-location (95\% CI [121.3, 172.4], $p_{M C M C}<.001$ ). The main effect of picture relatedness was not significant in any of the four conditions (all $p_{M C M C}>.67$ ).
The crucial effect is the interaction between picture relatedness (related vs. unrelated) and time period (early vs. late) on dwell times, as increased looks to related pictures should only start to occur once the target sign is being produced. For semantic trials, the picture by time period interaction was significant ( $95 \%$ CI of relative increase for related pictures in the late period [63.5, 107.9], $p_{M C M C}$ $<.001$ ) reflecting longer gaze to related compared to unrelated pictures in the later time period (that is, after the carrier phrase was complete and the target sign was being produced). A significant interaction of picture by time period was also observed in location-movement trials ( $95 \%$ CI of relative increase for related pictures in the late period [20.9, 96.6], $p_{M C M C}=.001$ ), again reflecting longer gaze to related than unrelated pictures in the later time period when information about the target sign becomes available. However, for the other two phonological conditions (handshape-movement and handshape-location) the interaction of picture and time period did not reach significance (both $p_{M C M C}>.3$ ).
We next conducted a Wilcoxon signed-rank test comparing looks to related and unrelated distracter pictures to explore possible differences in gaze across time, beginning at target sign onset. Cumulative fixations, analyzed as arcsine transformations, were grouped into

[^233]100ms bins starting from 400 ms after the target onset and continuing until 1000 ms (see Figure 3 for time course plots). 100 ms bins were used to ensure the presence of sufficient fixations to each area of interest during each time period for statistical analyses. Additionally, analyses began at 400 ms after the start of the target period because during the first 300 milliseconds of the target period across all trials, subjects fixated the sign video almost exclusively. For semantic trials, cumulative gaze toward related pictures differed significantly from the unrelated pictures across all bins from $400-1000 \mathrm{~ms}$ (range of Z from -2.20 to -3.59 , all $\mathrm{p}<.03$ ). This same pattern of results was observed for location-movement trials (range of Z from -2.31 to -3.63 , all $\mathrm{p}<.02$ ). There was no difference between related and unrelated picture gaze for handshape-location trials across all bins (all $p_{>} .24$ ). However, there were significantly more looks to related pictures compared to unrelated pictures for the handshape-movement condition, but this difference was found only from $800 \mathrm{~ms}-1000 \mathrm{~ms}$ ( $\mathrm{p}<.05$, between $800-$ 100 ms ; all $\mathrm{p}>.2$ up to 800 ms ).


Figure 3: Time course of eye gaze from onset of the target sign for 1000 ms for target-absent trials across the four conditions (from left to right: semantic, location-movement, handshape-location, handshape-movement).

## Discussion

Overall, we found both semantic and phonological effects during online processing of BSL using a visual world paradigm. Once information about the target sign became available, subjects looked at related pictures longer than unrelated pictures during the semantic condition. During the production of the target sign, related pictures also attracted more looks than unrelated pictures for one phonological condition (location-movement) but not for the others (handshape-location and handshape-movement). Importantly, in the early period of each trial (i.e. before the target sign was produced), there was no difference in gaze patterns to the different picture types (related and unrelated) confirming that the results are not driven by visual characteristics of the related pictures.

In the semantic condition, subjects looked at semantically related distracter pictures more frequently than unrelated pictures, the first time such findings have been demonstrated for a signed language. This result is predicted under the view that activation of semantically related lexical competitors should not be affected by the modality in which a language occurs. The results from the semantic condition reveal that despite the need for split visual attention to both visual linguistic stimuli and pictures related to that stimuli, it is possible to investigate sign language processing using visual world and related paradigms.
The results from the three phonological conditions pairing different phonological parameters produced differing results. Phonological competitors that shared information occurring at the onset of the sign (handshape and location features) did not draw more looks either at the onset of the period in which the target sign was produced (as evidenced by our analysis of the time period from $400-1000 \mathrm{~ms}$ after the target sign onset) or during the entire time period from the target sign onset until a button press decision was made (the late time period). This finding suggests that onsets may not be as relevant to sign language processing as has been suggested for spoken language processing (e.g., Gaskell \& Marslen-Wilson, 1997).
Crucially, in the location-movement condition, subjects looked significantly more toward the phonologically related picture than unrelated pictures in the late time period. This finding parallels the Corina and Knapp (2006) study that found effects only for signs sharing location and movement features. Further, for location-movement trials, looks to the related and unrelated pictures differed significantly from 400 ms after the onset of the target sign, a time comparable with that found in spoken language studies (e.g., Allopena et al, 1998). Finally, competitor pictures that shared handshape and movement features with the target did not draw more looks than unrelated pictures during the late time period. However, there was a significant, but short-lived difference such that looks to related and unrelated pictures differed between $800-1000 \mathrm{~ms}$ after the start of a target sign. Because a difference between related and unrelated pictures was not observed in the overall late period analyses we conclude that, while subjects may be aware of the phonological similarity of signs sharing handshape and movement features, they are likely not making use of these feature pairs during online processing. Instead, looks to related pictures occurring between 800 and 1000 ms (relatively late after the onset of the target sign) appears to be a post-lexical effect in which subjects consider alternate competitor pictures before determining that the target picture is not shown. Crucially, the pattern of gaze in the handshapemovement condition differs from the location-movement condition for which gaze to the related competitor picture is early and consistent. Thus, we suggest that gaze toward related competitors in location-movement trials is indicative of active online lexical processing as has been found in spoken language studies.
In the introduction, we offered two explainations for why onsets play a special role in auditory lexical access (i.e.,
either temporal constraints or salience). In terms of sign language processing, if temporal constraints are driving lexical access, then signers in our study should have paid attention to handshape and location features because these features are available at the start of a sign. Instead, the sign language results suggest that sign onsets are not similarly privileged to spoken word onsets, which in turn suggests that lexical processing is not temporally driven.
Alternatively, the data support a view under which perceptual salience drives sign access. Specifically, our data show that signers pay attention to movement features which are visually salient, but which occur relatively late in sign production: only trials with related distracters that share movement features show differences between looks to unrelated filler pictures and related competitor pictures. Further, the pairing of location-movement features appears to be of greatest importance during online processing.
The nature of acoustically perceived speech in languages such as English makes it impossible to determine why word onsets seem to have privileged status in lexical access: either due to temporal characteristics (perceived first) or perceptual salience. Investigating signed languages such as BSL allows us to clearly tease these apart, because the most salient perceptual properties (e.g. movement) are not available at the onset but only become available later. Thus, increased looks towards related distracter pictures in the location-movement condition may provide insight, not only into the nature of online sign processing but online speech processing as well. It is important to note that there is no a priory reason to assume that (visual) signs and (auditory) words will be processed similarly and therefore different strategies might be used.

Overall, the results here reveal important characteristics of lexical access concerning the role of lexical variables (semantic condition) and relative time course of access to different phonological parameters (phonological condition) for sign language processing. More broadly, our current understanding of language processing is intimately tied to oral-aural modality of spoken languages. The current work clearly shows that language processing interacts with modality, and that the key to lexical access for both signed and spoken languages may be perceptual saliency, instead of temporal recency.

## Acknowledgments

This work was supported by the Economic and Social Research Council of Great Britain (Grant RES-620-286001), Deafness, Cognition and Language Research Centre (DCAL).

## References

Allopena, P. D., Magnuson, J. S., \& Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. Journal of Memory and Language, 38, 419- 439.
Altmann, G. T. M., \& Kamide, Y. (2004). Now you see it,
now you don't: Mediating the mapping between language and the visual world. In J. Henderson \& F. Ferreira (Eds.), The interface of language, vision, and action. New York: Psychology Press.
Corina, D. P., \& Knapp, H. (2006). Lexical retrieval in American Sign Language production. In L. M. Goldstein, D. H. Whalen, \& C. T. Best (eds.). Papers in laboratory phonology. Berlin: Mouton de Gruyter, 213-240.
Emmorey, K., \& Corina, D. (1990). Lexical recognition in sign language: Effects of phonetic structure and morphology. Perceptual and Motor Skills, 71: 1227-1252.
Emmorey, K. (2002). Language, cognition, and the brain: Insights from sign language research. Lawrence Erlbaum and Associates, Mahwah, NJ.
Gaskell, M.G. \& Marslen-Wilson, W.D. (1997). Integrating form and meaning: A distributed model of speech perception. Language and Cognitive Processes, 12: 613-656.
Grosjean, F. (1981). Sign and word recognition: A first comparison. Sign Language Studies, 32: 195-219.
Corina, D. P., \& Hildebrandt, U. (2002). Psycholinguistic investigations of phonological structure in American Sign Language. In R. P. Meier, K. Cormier \& D. Quinto-Pozos (Eds.), Modality and Structure in Signed and Spoken Languages. Cambridge: Cambridge University Press.
Huettig, F., Altmann, G. (2005). Word meaning and the control of eye fixation: semantic competitor effects and the visual world paradigm. Cognition, 96(1): B23-B32.
Marslen-Wilson,W.D. (1973). Linguistic structure and speech shadowing at very short latencies. Nature, 244:522-523.
McClelland, J.L. \& Elman, J. (1986). The TRACE model of speech perception. Cognitive Psychology, 18:1-86.
Perlmutter, David M. (1992): Sonority and syllable structure in American Sign Language. Linguistic Inquiry, 23, 3, 407-442.
Rayner, K. \& Clifton C. Jr. (2009). Language processing in reading and speech perception is fast and incremental: Implications for event related potential research. Biological Psychology, 80: 4-9.
Reinisch, E., Jesse, A., \& McQueen, J. M. (2010). Early use of phonetic information in spoken word recognition: Lexical stress drives eye-movements immediately. Quarterly Journal of Experimental Psychology, 63(4), 772-783.
Sandler, W. \& Lillo-Martin, D. (2006). Sign Language and Linguistic Universals. Cambridge: Cambridge Univ.Press.
Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., \& Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. Science, 268(5217), 1632-1634.
Yee, E. \& Sedivy, J.C. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. Journal of Experimental Psychology: Learning, Memory \& Cognition, 32: 1-14.
Vinson, D.P., Cormier, K., Denmark, T., Schembri, A. \& Vigliocco, G. (2008). The British Sign Language (BSL) norms for age of acquisition, familiarity and iconicity. Behavior Research Methods, 40: 1079-87.

# A Natural Cure for the Pet Fish Problem: Feature Emergence as Classificatory Composition 

Chris Thornton Informatics<br>University of Sussex<br>Brighton<br>BN1 9QH<br>UK<br>c.thornton@sussex.ac.uk


#### Abstract

Where do emergent features come from? This has long been an intriging puzzle. The concept of pet fish illustrates the difficulty. Most people expect pet fish to live in bowls, even though this is not something either pets or fish normally do. The inference that pet fish have the feature of living in bowls cannot be explained purely in terms of the constituents themselves. The feature seems to emerge. The present paper aims to explain this effect using notions of classificatory composition. Adjoined concept references are taken to construct classifications rather than combinations; a pet fish is taken to be a fish classified as a pet rather than a combination of a pet a fish. It is also shown that, where concepts have a compositional representation, feature emergence can be accounted for in terms of compositional accommodation.


## Introduction

The concept of pet fish is one of the best known examples of a conceptual combination that produces difficult to explain 'emergent features.' People expect pet fish to live in bowls, even though this is not expected of either pets or fish (Murphy, 1988). The feature lives-in-bowls somehow leaps into existence when the concepts pet and fish are combined. But how and why? The effect seems to be something to do with the way the combination is explained. But giving a precise and general account is not straightforward (Rosch, 2011). This is the so-called Pet Fish problem, also known as the Guppy problem (Osherson and Smith, 1981)

Simple cases of concept combination are dealt with relatively easily. Let's say we combine the concept brown with the concept cow to form the concept, brown cow. Linguistically, the effect is to apply an adjective to a noun. In conceptual terms, the process is said to involve attachment of a modifier concept (brown) to a head concept (cow). Various explanations can then be set out. Assuming a schematic, slot/filler type of representation, the effect can be seen as one in which the modifier concept brown becomes the new filler for the color slot in the cow representation (cf. Hampton, 2011).

Simple explanations of this sort run into problems quite quickly, however. One relates to typicality effects. Consider the concept of a road bridge. For residents of the UK, a highly typical case of a road bridge is the Forth Road Bridge in Scotland. Unfortunately, this entity is unlikely to be considered typical of either roads or
bridges, on which basis there is a mystery about how the combination acquires typicality attributions not given to either constituent. This is known as the conjunction effect (Smith and Osherson, 1984).

Schematic theories of representation (Rumelhart and Ortony, 1977; Rumelhart, 1980) seem to offer a way of dealing with such problems. Let's say an apple concept is represented by a schema with slots for color and size. The fact that apples are typically green may be captured by placing a high typicality value on green as a color value (Murphy, 2002, p. 447). If constructing the concept of a red apple has the effect of placing red into the color slot of the apple schema, while also giving it a high typicality value, the expected typicality effect is reproduced. A red apple is then modeled as more typical of the concept red apple than it is of the concept red, or the concept apple. This approach to typicality values is at the heart of the selective modification proposal of Smith and Osherson (1984; Smith et al., 1988). ${ }^{1}$

The immediate difficulty with this idea is that it depends on there being a suitably modifiable slot in the head schema (Machery, 2009). In many cases, this seems to be ruled out. Consider Murphy's example of a 'party dog'. A plausible idea is that a party dog is a dog that does tricks. It is much less plausible that a dog schema will have a does-tricks slot, however. Combinations in which the head concept seems to lack any usefully modifiable slot abound. As Murphy notes, it is simply 'not the case that an adjective can automatically pick out a single dimension to modify' (Murphy, 2002, p. 450).

The general difficulty is the way in which concept combination goes beyond what can plausibly be conjured from constituent representations. The concept of pet fish is the classic illustration but 'Harvard-educated carpenter' is also revealing. People described as instances of this concept are likely to be seen as non-materialistic (Kunda et al., 1990). This seems to involve reasoning with relations that have nothing to do with the constituent concepts. Kunda et al., describe the process as the development of a 'causal narrative.' A simpler ex-

[^234]ample is 'big dog'. This entity seems to have features that cannot be explained in terms of representations for the concept big and the concept dog. A big dog is presumably small in relation to a house, for example. Any theory that models the combinational process purely in terms of constituent representations cannot explain such effects.

Explaining how concept-combination goes beyond processing of constituent representations is thus a key part of most proposals. In the concept specialization model of (Cohen and Murphy, 1984; Murphy, 1988) the process is understood to involve placing a representation of the modifier concept into a slot selected on the basis of background knowledge. This is then interpreted and elaborated taking 'outside knowledge' into account (Murphy, 1988, p. 533). ${ }^{2}$ The model allows scope for emergence of features through explanation. It has also been extended by Wisniewski to incorporate mechanisms of property-mapping, concept hybridization and relationlinking (Wisniewski, 1997; Wisniewski and Love, 1998).

Other proposals envisage ways in which processes of reasoning might directly mediate the combinational process. In Thagard's Amalgam theory (Thagard, 1984) application of procedural rules regulates slot/value selection in ways that promote an interpretation that reconciles 'the conflicting expectations contained in the candidate concepts' (Thagard, 1984, p. 4). Hampton's composite prototype model (Hampton, 1987, 1988, 1991) also deems combination to involve processes of reasoning, although here the process is based on theory-driven relations connecting slots of the original representations. Another proposal is the CARIN model of Gagne and Shoben (1997; Gagne, 2000; Gagne and Shoben, 2002). Here, the key idea is that combinational processes access a small library of fundamental thematic relations, including Cause, Has, About, Make, For, Use and Located-at. The model envisages combination to involve constructing an integrated slot/filler representation based around one or more of these foundational relations.

A variety of combinatorial mechanisms are proposed in the literature, then. Each has its pros and cons from the explanatory point of view (Ran and Duimerang, 2009; Murphy, 2002). The present paper aims to add a new explanatory proposal but does not hypothesise any new form of combinatorial mechanism. Rather it aims to explain featural effects in terms of the compositional properties that concepts inherently possess. It aims to show the sense in which feature emergence is a natural outcome of conceptual compositionality.

The hypothesis, more specifically, is that feature emergence can be explained in terms of natural concept com-

[^235]position. Conceptual structure assembled by this process has the distinguishing feature of being held together purely by the classificatory properties of concepts in the structure. There is no superimposed, compositional framework. The features which emerge from combining two concepts, it is argued, can be explained as the constructs that grow out of extending a natural compositional structure to accommodate constituent representations. However, it is to emphasized that the examples set out are intended to illustrate possible mechanisms. They are not intended to be realistic models of cognitive structure. The proposal is set out in two main sections. The section immediately to follow introduces the idea of natural concept composition. The second section shows how this type of conceptual structure can be a way of explaining cases of feature emergence.

## Natural concept composition

Although the classificatory functions of concepts are normally viewed as applying to non-conceptual entities, they can also be seen as applying to concepts themselves. A fork concept is the means of classifying something as a fork. A plastic concept is the means of classifying something as plastic. But some forks are plastic, on which basis the plastic concept can also be a way of classifying the fork concept. What is picked out is the extensional intersection - the set of all plastic forks. This is just the extension of the concept of plastic forks. Where extensions intersect in this way, the effect of using one concept to cross-classify another can thus be a new concept: it is the concept whose extension is the intersection arising. Cross-classification of one concept by another has the potential to yield an implied concept, with an extension that is derived by intersection.

The more complex case of this is where we have a concept that classifies an $n$-tuple of entities (i.e., a relation). For example, let's say we have a male-celebrity concept, a female-celebrity concept and a married-couple concept. There is the potential for a male and female celebrity to be classified as a married couple. The married-couple concept is a potential classification of an $n$-tuple comprising the male-celebrity and female-celebrity concepts, then. The result is a celebrity-couple concept, whose extension is the set of couples made up of a male and a female celebrity.

Classifications of this type exist whenever the extension of the classifying concept contains at least one of the possible permutations of instances. To formalize this, we have to work in terms of cartesian products. A crossclassification exists if the extension of the classifying concept intersects the cartesian product of the extensions of constituents in the classified $n$-tuple. ${ }^{3}$ Provided we allow singleton $n$-tuples, this deals with all cases, since the

[^236]cartesian product of a single set is just the set itself． A general rule can be stated accordingly．
－Cross－classification rule：concept $x$ classifies some $n$－tuple of concepts（not including $x$ ）if the cartesian product formed from their extensions intersects the extension of $x$ ．

It is then possible to define the situation where a set of concepts gives rise to an implied concept by means of classificatory composition．
－Composition rule：if within some set of concepts there is a valid cross－classification that gives rise to an extension that is new for the set，this composition is an implied concept．

The idea of＇compositional completion＇can then be set out．A set of concepts may be said to be compositionally complete just in case it has no implied concepts．This either means all of them have already been added，or there were none to begin with．The compositional com－ pletion of a set of concepts is the set of implied concepts given rise to，then．The composition rule specifically re－ quires that an extension be new for the original set of concepts．This accommodates the possibility of having two or more implied concepts with the same extension．

Notice the rules allow for cumulative effects：the iden－ tification of one implied concept can give rise to another． The result in such cases is a compositional structure of two levels－one classification inside another．There is no limit on the number of times this can happen，and thus no limit on the structural complexity that may emerge．The compositional completion of a set of con－ cepts may comprise conceptual structures of arbitrary compositional complexity．These structures have the distinguishing feature of being held together purely by the classificatory properties of the concepts they contain． There is no superimposed，compositional formalism．For this reason，the process is called natural concept compo－ sition．

## Couples example

The dynamics of natural concept composition can be illustrated by extending the celebrity－couples example． Let＇s say we start with a set of five concepts defined as follows．

```
male-celebrity = {Brad, David, George}
female-celebrity ={Ange, Posh, Rita}
pets ={Fido, Twinkle, Rover}
couples = {\langle\mathrm{ Brad Ange }\rangle,\langle\mathrm{ David Posh }\rangle,\langle\mathrm{ Fido Qi}\rangle}
pet-owners = {\langleBrad Ange Fido\rangle, \langleJo Sam Rover }\rangle
```

These definitions should be self－explanatory．The exten－ sion of the male－celebrity concept is defined to be $\{\mathrm{Brad}$ ，

David，George\}. The extension of the couples concept is $\{\langle\operatorname{Brad}$ Ange $\rangle,\langle$ David Posh $\rangle,\langle$ Fido Qi $\rangle\}$ ，and so on． All elements of extensions are understood to be $n$－tuples， but where $n=1$ ，the angle brackets are omitted．

Examination of the definitions reveals that initially there is just one compositional implication．The exten－ sional cartesian product of male－celebrity and female－ celebrity intersects the extension of the couples con－ cept，with the intersection being $\{\langle$ Brad Ange $\rangle,\langle$ David Posh〉\}. The classification of 〈male-celebrity femalecelebrity $\rangle$ by the couples concept is an implication of the set，then．For purpose of notating this，the conven－ tion used here is to enclose the classifying concept and its constituents in square brackets，with the classifying concept placed first．The implied construction is thus written

## ［couples male－celebrity female－celebrity］

The concept is that of a celebrity couple：it is referred to as the celebrity－couple concept below．The extensional definition is $\{\langle$ Brad Ange $\rangle,\langle$ David Posh $\rangle\}$ ．
Once this implied concept has been identified，there is a knock－on effect resulting from the ability of the pet－ owners concept to classify a composite of celebrity－couple and pets．The cartesian product derived from this in－ cludes $\langle$ Brad Ange Fido〉，which is within the extension of pet－owners．A second implied concept then exists，in which the initial construction plays the role of a con－ stituent．This is a structure of two levels：

```
[pet-owners
    [couples male-celebrity female-celebrity]
    pets ]
```

The implied concept is that of a celebrity－couple classi－ fied as pet owners．In other words，it is the concept of a pet－owning celebrity－couple．This is the final implica－ tion in the present case．The compositional completion of this set comprises just two concepts，then．

## Feature emergence

The second part of the proposal can now be set out．The hypothesis is that feature emergence results from natu－ ral concept composition．Viewing concept combination as classificatory composition has the effect of making fea－ ture emergence an expected outcome．Where a feature is seen to emerge as a result of combining two concepts， the process can be modeled in terms of compositional conceptions brought into existence by activating compo－ sitional representations．

In the simplest cases，concept combination can be ex－ plained purely in terms of schematic representation，and the ways in which activating one schema modifies an－ other．Such accounts are straightforwardly translated into the present framework．From the compositional point of view，a slot／filler schema is a construction in
which the classifying concept is that of combination, and each constituent is a one-level construction in which a filler value classifies a slot value. An apple schema has something like the following form, then.
[combination [green color] [round shape]]

Any account in which concepts are deemed to be combined by means of schema modification can then be expressed in terms of conceptual integration. Taking the concept of red to have the form [red color], an integration of the apple and red concepts would produce this conception of a red apple, for example.
[combination [red color] [round shape]]

This is the sense in which compositional processes can model schema-updates. ${ }^{4}$ But notice the classificatory viewpoint that is imposed. On the assumption of the combination being represented compositionally, a red apple is an apple that is classified as red. It is not a red thing that is also an apple.

Combinations of more relevance to the Pet Fish problem are ones which yield emergent features via inferential explanation. The account in the case of 'pet fish' has already been noted: people expect pet fish to live in bowls (even though this is not the usual behavior of either pets or fish) because this is a way of explaining how fish can be both kept, and kept alive (cf. Murphy, 1988). The process is seen to involve the construction of an inferential explanation in which living in bowls is inferred to be the only way of meeting the requirement for fish to be kept in water.

The procedure for translating an explanatory process into a compositional one involves treating each inferential step as a conception. Inferences involving a schematic idea are seen as combinatorial conceptions (i.e., classifications based on the concept of combination). Inferences involving a categorical idea are seen as unifying conceptions (i.e., classifications based on the concept of unity). The process of connecting one inference to another is conceptual construction. Such connections are established by making one conception a constituent of another.

The explanatory process prompted by 'pet fish' integrates the schematic inference that fish need to live in water with the schematic inference that pets are kept in enclosures. Assume the construction of the pet conception is
[combination [habitat enclosure] [role amusement]],

[^237]The corresponding conception of fish is
[combination [habitat water] [activity swimming]].

The two schematic inferences are then straightforwardly accomplished-they result directly from activating the constituent representations. Their integration, on the other hand, requires a combining conception. Given the understanding that a pet fish is a fish classified as a pet, this must impose the habitat classification from the pet conception on the habitat classification in the fish conception. The construct obtained is then

```
[combination
    [habitat [enclosure water]]
    [role amusement]
    [activity swimming]].
```

On the assumption that [habitat [enclosure water]] constructs a conception of a habitat containing water, something akin to the lives-in-bowls feature is reproduced.


Figure 1: 'Pet fish'.

This compositional story is set out schematically in Figure 1. In this diagram, triangles represent conceptions: the name of the conception appears at the apex, with the classifying concept placed above the lower edge, and the classified constituents below it. Where we have a classification with a single constituent, a stack arrangement is used. The classifying conception is placed immediately above the classified constituent, with a line between them.

More complex cases of feature emergence give rise to more complex interpretations. But the principles of translation remain the same. Consider the case of 'Harvard-educated carpenter'. The emergent feature in this case relates to an attitude: Harvard-educated carpenters are inferred to be 'non-materialistic' (Kunda et al., 1990). The combination of high earning power and modest remuneration in a single individual is taken to imply that the individual must have a care-less attitude to money. The feature that emerges is a known classification of a combination of features. But these features are inferred rather than given.


Figure 2: 'Harvard-educated carpenter'.

Modeled as natural concept composition, the explanatory process has the form depicted in Figure 2. The inference that a Harvard education produces high earning power comes from activating the relevant compositional representation (which incorporates the conception [elevation earning-power]). The inference that a carpenter is likely to have limited remuneration is produced in a similar same way, by activating the relevant compositional conception. An attitude of non-materialism is then inferred by means of an existing combinatorial conception for non-materialism based on these particular constituents. Again, the interpretation enforces a classificatory understanding. A Harvard-educated carpenter is not taken to be a combination of a carpenter and a Harvard education. It is seen to be the classification of a carpenter as Harvard-educated.

More complex still is the idea of an apartment dog. This phrase is found to suggest the idea of a dog that is small, even though this is not normally a property of either apartments or dogs (Murphy, 1988). The complication is the involvement of an intermediate idea. An apartment dog is not seen to be a dog classified as an apartment, let alone a dog that is also an apartment. It is expected to be a dog that lives in an apartment. Part of the explanatory process entails the idea of a particular type of agent residing in a particular type of dwelling.

The inferential steps in the explanation are seen to be essentially as follows (cf. Murphy, 2002). There is the categorical inference that apartments are types of dwelling, and the categorical inference that dogs and occupants are both types of agent. There are also three schematic inferences: the inference that apartments offer limited scope for exercise, the inference that dwellings have residents that are suitable, and the inference that suitability of an occupant for a dwelling requires correspondence between the occupant's size and exercise requirement. The structure labeled A in Figure 3 shows these five steps as compositional conceptions. The concepts utilized for classification include that of combination, of unity, and of correspondence.

Given this model, the conclusion that apartment dogs are small can be seen to result from an interaction be-


Figure 3: 'Apartment dog'.
tween two conceptual processes. Reference to an apartment activates a dwelling conception in which exercisescope is classified as restricted. A conception of residence is then realized, incorporating a suitability conception that classifies the size of the occupant as corresponding to the (restricted) exercise-scope of the dwelling. The only way of integrating the conception activated by the dog reference is then by inferring the dog to be the occupant in the residence conception. There is then an implicit classification (via the suitability conception) of the size of the dog as corresponding to the exercise-scope of the dwelling. A restriction is inferred to apply to the size of the dog. Emergence of the infered feature of smallness is obtained by means of a sequence of interacting, compositional constructions.

Modeling concept combination in this way has the advantage of parsimony. Notions of schematic representation and explanatory inference are collapsed to the idea of compositional conceptualization. The approach can deal with simple cases involving direct transfer of features as well as more complex cases involving explanation. But it cannot be emphasized too strongly that the examples set out are not intended to be realistic models of human cognition. It is not claimed that these particular conceptions are the ones people use. Given a free choice of what concepts are given, and no restriction on the number of constructive levels that can be brought into play, there are infinitely many ways in which any conception can be constructed. The examples set out are purely illustrative of the way in which feature emergence might result from classificatory composition.

## Conclusion

Adjoined concept references are normally viewed as constructing combinations. The phrase 'toy vehicle' is considered to combine the concepts toy and vehicle, the phrase 'pet fish' is considered to combine pet and fish, and the phrase 'apartment dog' is considered to combine apartment and dog. Hence the name of the area
of study. But as the present paper shows, these phrases can also be seen as classifications. A 'toy vehicle' can be a vehicle classified as a toy, a pet fish can be a fish classified as a pet, and an apartment dog can be a dog classified as living in an apartment. This leads to a new way of explaining feature emergence. On the assumption that concepts are represented as classificatory compositions (rather than as one-leveled schemata), explanatory feature emergence can be seen to grow out of compositional conceptualization. This avoids the need to think in terms of dedicated mechanisms of combination and explanation. The phenomenon is seen to result from activation of naturally constructed, compositional representations.

## References

Cohen, B. and Murphy, G. L. (1984). Models of concepts. Cognitive Science, 8 (pp. 27-58).

Gagne, C. L. and Shoben, E. (1997). Influence of thematic relations on the comprehension of modifiernoun combinations. Journal of Experimental Psychology: Learning, Memory, and Cognition, 23 (pp. 71-87).

Gagne, C. L. and Shoben, E. (2002). Priming relations in ambiguous nounnoun combinations. Memory Cognition, 30 (pp. 637646).

Gagne, C. L. (2000). Relation-based versus propertybased combinations: a test of the CARIN theory and dual-process theory of conceptual combination. Journal of Memory and Language, 42 (pp. 365-389).

Hampton, J. A. (1987). Inheritance of attributes in natural concept conjunctions. Memory and Cognition, 15, No. 1 (pp. 55-71).

Hampton, J. A. (1988). Overextension of conjunctive concepts: evidence for a unitary model of concept typicality and class inclusion. Journal of Experimental Psychology: Learning, Memory and Cognition, 14 (pp. 12-32).

Hampton, J. A. (1991). The combination of prototype concepts. In Schwanenflugel (Ed.), The Psychology of Word Meanings. Hillsdale, N.J: Lawrence Erlbaum Associates

Hampton, J. (2011). Conceptual combination and fuzzy logic. In Belohlavek and Klir (Eds.), Concepts and Fuzzy Logic (pp. 209-232). MIT Press

Kunda, Z., Miller, D. T. and Claire, T. (1990). Combining social concepts: the role of causal reasoning. Cognitive Science, 14 (pp. 551-577).

Machery, E. (2009). Doing without Concepts. Oxford: Oxford University Press

Murphy, G. L. (1988). Comprehending complex concepts. Cognitive Science, 12 (pp. 529-562).

Murphy, G. L. (2002). The Big Book of Concepts. London, England: The MIT Press

Osherson, D. N. and Smith, E. E. (1981). On the adequacy of prototype theory as a theory of concepts. Cognition: International Journal of Cognitive Psychology, 9 (pp. 35-58).

Ran, B. and Duimerang, P. R. (2009). Conceptual combination: models, theories, and controversies (chapter 2). In Weingarten and Penat (Eds.), Cognitive Psychology Research Developments. Nova Science Publishers, Inc

Rosch, E. H. (2011). "slow lettuce": categories, concepts, fuzzy sets and logical deduction (chapter 4). In Belohlavek and Klir (Eds.), Concepts and Fuzzy Logic (pp. 89-120). MIT Press

Rumelhart, D. E. and Ortony, A. (1977). The representation of knowledge in memory. In Anderson, Spiro and Montaque (Eds.), Schooling and the Acquisition of Knowledge. Hillsdale, NJ: Erlbaum

Rumelhart, D. E. (1980). Schemata: the building blocks of cognition. In Spiro, Bruce and Brewer (Eds.), Theoretical Issues in Reading Comprehension. Hillsdale, NJ: Erlbaum

Smith, E. E. and Osherson, D. N. (1984). Categories and Concepts. Cambridge, MA: Harvard University Press

Smith, E. E., Osherson, D. N., Rips, L. J. and Keane, M. (1988). Combining prototypes: a selective modification model. Cognitive Science, 12 (pp. 485-527).

Thagard, P. (1984). Conceptual combination and scientific discovery. In Asquith and Kitcher (Eds.), PSA 1984 vol. 1 (pp. 3-12). East Lansing: Philosophy of Science Association

Wisniewski, E. J. and Love, B. C. (1998). Relations versus properties in conceptual combination. Journal of Memory and Language, 38 (pp. 177-202).

Wisniewski, E. J. (1997). When concepts combine. Psychonomic Bulletin Review, 4 (pp. 167-183).

# Learning the language of time: Children's acquisition of duration words 

Katharine A. Tillman (katillman@ucsd.edu)<br>University of California San Diego, 9500 Gilman Drive<br>La Jolla, CA 92093 USA

David Barner (barner@ucsd.edu)<br>University of California San Diego, 9500 Gilman Drive<br>La Jolla, CA 92093 USA


#### Abstract

Before children acquire the precise definitions of time words, like minute and hour, how do they interpret them? And how are such proto-meanings acquired in development? Here we present three experiments, and assess children's early understanding of seven time words: second, minute, hour, day, week, month, and year. Our findings indicate that children first learn time words as a lexical class, then learn their ordinal relations, but initially have little to no knowledge of their relative durations. This understanding emerges late in development - many years after children first start using time words in speech - and in many children does not emerge until they have acquired formal definitions for the words.


Keywords: abstract word learning; time perception; language acquisition; number-line estimation

## Introduction

Understanding the nature of time is a hard problem, not only for physicists and philosophers who debate its status in the universe, but especially for young children who are exposed to artifacts and linguistic representations of time from early in life. We rely on clocks, calendars, and words like second, minute, and hour to measure and keep track of time, and to coordinate our activities with others. Interestingly, although children begin using time words relatively early in life - by as young as 2 - and 3 -years of age, most do not receive formal instruction regarding the meanings of these words until much later, when they enter school. This raises the question of how children interpret these words prior to formal instruction, and how these words are initially related to their subjective experience of time, and the relative durations of events. In the present study, we explored this question, and asked what types of information children use to make sense of early time words, and thus how they begin to acquire their meanings in early development.

Duration words like time, day, and year, are among the most frequent nouns in English (Kucera and Francis, 1967). In addition to duration words, which we focus on in the present study, time is also conveyed through verb tense, through temporal adverbs such as yesterday and tomorrow, through spatiotemporal metaphor (e.g., "a long meeting"), and through the sequential structure of narrative itself. The rich and varied ways in which language encodes the dimension of time make it possible to reason and communicate about events that are not currently happening.

Despite this abundance of temporal language, acquiring the meanings of time words presents a considerable challenge to the early language learner. Time can neither be
seen nor heard. Unlike concrete nouns referring to whole objects that can be easily pointed out, and even more challenging abstract terms like color words (referring to properties of objects) and number words (referring to sets of objects), there is no static perceptual stimulus to which a duration word like minute refers. Word-learning principles such as "fast mapping" and mutual exclusivity, which describe useful strategies for learning the names of new objects or object properties in the context of familiar ones, do not easily apply. Rarely in everyday life (in the absence of clocks and timers) are there explicit perceptual markers denoting when events or specified temporal periods start and end, further complicating the task of figuring out the proper referents for time words.

Children are not typically taught the formal definitions of duration words (e.g., one minute equals sixty seconds) until they reach school age, but they begin hearing and even producing these words much earlier, albeit with very low accuracy. In child-directed speech, mothers of preschoolers use time words less often, but in a wider variety of contexts, than color and number words (Tare et al., 2008). While over $80 \%$ of children produce duration terms, including minute(s) and $\operatorname{hour}(s)$, by age 5 , only $22 \%$ of 5 -year-olds reportedly use hour(s) appropriately (Grant and Suddendorf, 2011). Here we are interested in whether, during these years of inaccurate production, before learning the adult definitions (e.g., that an hour is 60 minutes), children acquire naive meanings based on other information, and, if so, what information they use to do so.

There are two broad sources of information children could use in forming intuitive definitions of duration words. One source is their capacity to perceive and represent the durations of experienced events, and the other is their linguistic input. Children's ability to use and combine information from these two sources leads to three possible hypotheses characterizing the extent of their early learning, each increasingly sophisticated.

By the first account, which we call the Nominal hypothesis, children rely upon linguistic input to construct a lexical category for time words, thus understanding only that hour and minute belong within a common class of words. Consistent with this, Shatz and colleagues (2010) observed that, when asked "how long" or "how much time" an event takes, a much higher proportion of preschool-aged children are able to respond appropriately (using a quantity word and a duration word) than are able to respond accurately (Shatz et al., 2010). Children apparently understand what kinds of words can answer a question
about time before they can map those words onto specific durations.

Second, children might learn the ordinal relations among time words. This requires an additional inference: duration words vary along a common scale. Linguistic input could also be used to support this level of understanding. For instance, if a child hears an adult utterance such as, "We're leaving for the zoo in an hour, so you only have ten minutes to finish eating lunch," without knowing the precise definitions of either duration word, he could still use the linguistic context to conclude that an hour must be longer than a minute, if he understands that both those words denote amounts of time. By the Ordinal hypothesis, beyond simply learning that time words share a nominal class, children also learn the ordinal relationships among their list of known time words, e.g., year $>$ month $>$ week $>$ day $>$ hour $>$ minute $>$ second.

Third, children might learn the approximate ratios between the durations encoded by time words. How could this most knowledge be acquired before explicit instruction on time words? The Ratio hypothesis relies on duration perception, as understanding of relative temporal magnitudes requires that duration words be associated with nonverbal representations of duration. By the Ordinal account, above, a child will know only that a minute is 'bigger' than a second, but by the Ratio account, he would also know approximately how much bigger than a second a minute is (a ratio of 60:1).

We experience duration, thus children might be able to map this dimension onto language. Experimental work has shown that even nonverbal animals use temporal information to guide behaviors such as seeking food or avoiding shocks that come at predictable intervals. The human mind must have means of representing elapsed time, and many cognitive models have been proposed describing the operation of mental clocks and pacemakers. By four months, babies habituate to the temporal pattern of a flashing visual stimulus, and react when a flash is omitted at a prescribed time, revealing a very early sensitivity to elapsed duration (Columbo \& Richman, 2002). Basic psychophysical tasks have also measured the precision with which adults and children can estimate and compare the durations of auditory and visual stimuli, usually on the order of milliseconds or seconds. Although temporal sensitivity does not reach adult levels until around age 8, even the youngest children tested are able to discriminate stimuli on the basis of duration (Droit-Volet et al., 2004).

If the duration representations are available to children, how would the mapping between duration and language be formed? Perhaps a child hearing adult speech about time may associate unfamiliar duration words with the familiar events they describe or in whose context they are uttered, resulting in associative mappings between duration words and perceived temporal magnitudes. Evidence that children have knowledge of the durations of familiar events that they are not currently experiencing (and which extend beyond a few seconds in temporal extent) comes from a study by

William Friedman (1990). Friedman first taught children that a spatial array of nine boxes, much like a number-line, represented duration, from a very short time (the leftmost box) to a very long time (the rightmost box). He then had children indicate how long familiar events, such as drinking a glass of milk or watching a cartoon show, took, by placing a cube in the appropriate box. Four-year-old children correctly ranked-ordered the activities by duration, and by 5 years their mean placements on the 9-point scale were wellcorrelated with adult-estimated durations of the activities. Friedman's tasks did not utilize any conventional duration terms such as minute or hour. Our Experiment 3 asks whether children are able to use a number-line paradigm to estimate the durations represented by conventional time terms as well as by familiar events.

Few prior studies of language acquisition have assessed children's early comprehension of time words. Such studies probe what children know about time words before they can produce them accurately, for instance by requiring a forced choice. In Shatz et al. (2010)'s Study 2, children were introduced to a puppet "from far away" who "didn't know very much," and were asked show him which of two pictures represented an activity taking a specific amount of time, such as 10 minutes. Five-year-olds performed above chance overall, and 6 -year-olds were near $70 \%$ correct. This study suggests that 5 -year-olds have a rudimentary understanding of the meanings of duration words and how they relate to familiar activities. However, the results are difficult to interpret because each prompt combined duration words, number words, and events. Children could succeed (or fail) at the task based on their level of understanding in any of these three areas. Though Shatz et al. interpreted their results as favoring a lexical domain hypothesis, they do not rule out the possibility that children may rely on quantitative representations of duration as well.

Here we present three experiments designed to assess whether children understand time words at the Nominal, Ordinal, or Ratio level. Experiment 1 uses a forced-choice procedure to ask whether children can make time quantity comparisons on the basis of duration words alone (Nominal hypothesis predicts failure, Ordinal and Ratio hypotheses predict success). Experiment 2 introduces number words into the forced-choice, asking whether children can combine their knowledge of time words with their understanding of number (only Ratio hypothesis predicts success on critical trials). Experiment 3 uses number-line estimation to assess children's ability to map time words and events onto a spatial scale representing duration, providing data that can be analyzed both by ordinality (testing the Ordinal hypothesis) and by relative distance (testing the Ratio hypothesis). Finally, we assess children's explicit knowledge of the formal definitions of duration words, and use this as a predictor of their number-line estimation performance.

## Materials and Methods

## Participants

For Experiment 1, we recruited 89 children from the San Diego area, including 253 -year-olds, 264 -year-olds, 205 -
year-olds, and 18 6-year olds. For Experiment 2, 85 children participated, including 254 -year-olds, 225 -year-olds, 226 -year-olds, and 16 7-year-olds. Fifty-two children participated in Experiment 3, including 225 -year-olds, 17 6 -year-olds, and 137 -year-olds. 36 young adults (Mean age $=20.6$ years) also participated in Experiment 3. An additional 16 children also participated but were excluded from analysis due to failure to complete the task (8), failure to comprehend the task (4), being outside the age range of interest (3), and experimenter error (1).

## Procedure, Experiments 1 and 2: Forced-choice

Two action figures, Farmer Brown and Captain Blue, were placed on a table in front of the child. On each trial, the experimenter read a short scenario such as, "Farmer Brown [jumped] for [a minute]. Captain Blue [jumped] for [an hour]." This was followed by a two-alternative forced choice, "Who [jumped] more, [Farmer Brown or Captain Blue]?" If the child was reluctant to give a verbal response, she was encouraged to point to the character that did the action more. Procedures for Experiment 2 were identical to those of Experiment 1, but the time words were modified by number words. For example, "Farmer Brown [jumped] for [two] [minutes]. Captain Blue [jumped] for [three] [hours]." Each child completed a total of 26 trials in the Experiment 1, or 30 trials in Experiment 2.
Trials and coding, Experiment 1. Children completed two blocks of thirteen duration comparisons involving seven time words: second, minute, hour, day, week, month, and year. The comparisons tested were: week vs. month, day vs. week, month vs. year, hour vs. day, day vs. month, week vs. year, minute vs. hour, second vs. minute, hour vs. week, day vs. year, minute vs. day, second vs. hour, and second vs. day. Six action verbs, all of which were high-frequency words denoting activities that could be done for variable lengths of time, were used: jumped, slept, cried, played, danced, and talked. Within each block, trials were conducted in quasi-random order. Verbs were randomly assigned to duration comparisons, with the stipulation that the same verb was never used in two consecutive trials. Trials were counterbalanced with respect to whether the larger duration word came first, which character represented the correct answer, and which character was prompted first. Half the participants received one item-order, and the other half received the reverse order. For analysis, the child's response on each trial was coded as correct (1) or incorrect(0). These numbers were then converted into proportions correct.
Trials and coding, Experiment 2. Trials in Experiment 2 included the same six verbs from Experiment 1. However, only five time-word comparisons were used in Experiment 2: minute vs. hour, week vs. year, day vs. year, day vs. week, and second vs. hour. For each of those five time-word pairs, 7 different types of number-word comparisons were made (Table 1). One trial included no numbers (identical to Experiment 1,), 3 included "small" numbers (2 and/or 3), and 3 three included "big" numbers (6 and/or 9). Each comparison was designated Same, Congruent ,or

Incongruent, depending on whether the larger number word was paired with the larger time word (see Table 1).All 30 trials were conducted in quasi-random order. Half the participants received one item-order while the other half received the reverse order.

Table 1: Experiment 2 trial types

| Number <br> comparison | Number size | Example |
| :--- | :--- | :--- |
| No numbers | None | a minute vs an hour |
| Same | Small | 2 minutes vs 2 hours |
|  | Big | 6 minutes vs 6 hours |
| Congruent | Small | 2 minutes vs 3 hours |
|  | Big | 6 minutes vs 9 hours |
| Incongruent | Small | 3 minutes vs 2 hours |
|  | Big | 9 minutes vs 6 hours |

## Procedures, Experiment 3: Number-line estimation

Participants were given a sheet of 8.5 'x11' paper with four horizontal, $17-\mathrm{cm}$ lines printed in a vertical column down the center of the page. Each line had circles on both endpoints and no other markings. Children were told that the top line was a number-line going from 0 to 100 . "Each number has its own place on the line," said the experimenter. "You're going to show me where certain numbers go on the number-line. Look, 0 goes here [experimenter draws vertical mark at left endpoint] and 100 goes here [experimenter marks right endpoint]." For each of four number stimuli (see Table 2), the experimenter instructed the child, "The [first] number is [4]. Can you show me where [4] goes? Can you draw a line with the [blue] pencil?" The first line was intended to give a baseline measure of children's ability to perform an estimation task using a number-line. For each of the next three tasks, the line represented duration rather than numerical quantity. This was explained to the participants as follows: "Now, this line is different. It shows how much time things take to do. It goes from a very short amount of time to a very long amount of time. Each amount of time has its own place on the line, and the further you go over here [gesturing along the line], the more time something takes. You're going to show me how long certain things take to do on the line. Something very short, like blinking your eyes, goes here [experimenter marks left endpoint]. Something very long, like the time from waking up in the morning to going to bed at night, goes here [experimenter marks right endpoint]. For each stimulus (see Table 2), the child was instructedto think about how long the activity takes to do and to mark theline accordingly. Participants were reminded that each subsequent line represented duration and what the endpoints represented (blinking eyes, morning to night) in between the remaining tasks and if confused.

## Trials and coding, Experiment 3.

Stimuli for Experiment 3 are shown in Table 2. Each participant estimated number on the first line, familiar event durations on the second, conventional time word durations
on the third, and combinations of time words and number on the fourth. Within each line, half the participants received the four stimuli in the order shown in Table 2, while the other half received the reverse order. As in Experiments 1 and 2, participants were presented with time word stimuli (lines 3 and 4) in the context of events that could take variable amounts of time, e.g. "[jumping] for a minute."

Table 2: Experiment 3 Number-line stimuli

| Number | Event | Time word | Num + time |
| :---: | :---: | :---: | :---: |
| 4 | Watching movie | Hour | 2 hours |
| 45 | Washing hands | Second | 6 hours |
| 18 | Trip to zoo | Minute | 9 min |
| 61 | Eating lunch | Day | 3 min |

Explicit knowledge. Following completion of the four number-line tasks, the participant was asked 3 follow-up questions: how minutes are in an hour, how many hours are in a day, and how many seconds are in a minute. Responses were coded as either correct (1) or incorrect (0), and were converted to proportions correct.
Estimation. To analyze the number-line data, we measured the distance (in cm , to the nearest tenth) from the left endpoint of the line to the intersection of the number-line with each of the participant's pencil marks. Marks falling exactly on the left endpoint were recorded as 0.1 cm (to avoid divide-by-zero errors) and those falling exactly on the right were recorded as 17.0 cm . To assess knowledge of relative durations, we computed ratios between each pair stimuli (e.g., $\mathrm{min} / \mathrm{sec}$, hour/sec, hour/min, day/sec, day/min, day/hour). Children's estimation performance was assessed by comparing their distances and ratios with corresponding means from the adult participant group. We focus on the results from the time word numberline task, which most directly bear on the Ordinal and Ratio hypotheses.
Ordinality. Responses to each trial were also coded for ordinality. To do this, each of the four stimuli for each line was rank-ordered by increasing magnitude or duration. In the case of line 2, the correct (adult-estimated) rank order was: 1 . washing your hands, 2 . eating lunch, 3 . watching a movie, 4. going on a trip to the zoo. The participant's marks were also ranked by increasing distance from zero. For each estimated item which fell in the correct rank, the participant was awarded a 1 , for each incorrectly ranked item, the participant was given a 0 , which were converted into proportions correct for each child and each age group.

## Results and Discussion

We began with three alternative hypotheses for how to characterize children's early knowledge of duration words prior to learning their definitions. The Nominal hypothesis is that children simply understand that durations words belong to a common lexical category, the Ordinal hypothesis is that children have knowledge of the ordinal relations among the words within this category, and the Ratio hypothesis is that children have knowledge both of the
ordinal relations and of the relative lengths of the durations to which the words refer. Of these three possibilities, only the Ratio hypothesis requires that children form associations between duration words and nonverbal representations of duration.

## Experiment 1

The primary goal of Experiment 1 was to distinguish between Nominal and Ordinal/Ratio understanding of time words, by asking whether children are able to compare two lengths of time strictly on the basis of the conventional duration terms used to describe them. In order to succeed at this two-alternative forced choice task, children must possess some understanding of the ordinal relations among the various time words. Unlike in prior forced-choice studies of time word comprehension (Shatz et al., 2010), here participants could not rely on their knowledge of number or of familiar events in order to succeed. Measuring overall accuracy in Experiment 1, we found that while our youngest group of participants, the 3 -year-olds, did not perform better than $50 \%$ accuracy, as predicted by chance ( $M \pm$ SEM $=0.48 \pm 0.02 p=0.2$,n.s.), the 4 -, 5 -, and 6 -year-old groups all performed significantly better than chance ( $M \pm$ SEM, respectively, $=0.57 \pm 0.02 ; 0.67 \pm 0.04 ; 0.81 \pm 0.03$, all $p \prime \mathrm{~s}<0.005$ ). Furthermore, each age group performed significantly better than each younger group (all $p$ ' $s<0.05$ ). While the question of whether 3-year-olds have nominal understanding of some or all of the terms is left open, these data reject the possibility that children 4 years and older know only that time words belong to a common category. It is also noteworthy that our oldest age group, the 6-year-olds, while performing quite well, were not at ceiling, despite the simplicity of the task and the likelihood that this sample had already received some formal instruction on duration words.

We were also interested in possible comparison effects or time-word effects in the data, as these may provide important clues into the order in which duration words are acquired. We hypothesized that, if these words are truly associated with durations, we might observe patterns such as greater accuracy on comparisons between more distant terms (sec. vs. day $>$ sec. vs. min.), or greater success on comparisons involving shorter, and thus easier-to-represent durations, such as second and minute, than comparisons involving longer terms, such as month and year, which may be harder to represent nonverbally. Though a mixed logistic regression predicting the probability of making the correct choice as a function of the participant's age and the timeword comparison type did find significant effects of each (Age: $\quad \mathrm{c}^{2}(3)=142.7, \quad p<0.001$, TrialType: $\quad \mathrm{c}^{2}(13)=59.0$, $p<0.001$ ), as well as an interaction between them ( $\mathrm{c}^{2}(36)=71.2, p<0.001$ ), there was no evidence indicating that the relative durations encoded by the two words being compared were driving the effect. Furthermore, collapsing the data across all comparisons involving each time word so as to compare overall accuracy for each word revealed no differences in performance $(F(6,595)=1.2, p=0.3$, n.s.). As accuracy improved from age group to age group, it improved across the board, with equal improvement on each
tested word, as would be expected if these words are being learned as a set, with performance on each word being limited by overall understanding of the ordinal relations among the words in the list, without direct associations between each individual term and duration per se (consistent with the Ordinal, rather than the Ratio hypothesis).

## Experiment 2

Experiment 2 assessed children's ability to integrate their knowledge of number with their understanding of time words, pitting the Ordinal and Ratio hypotheses against one another by probing the specificity of children's knowledge of the relative lengths of time referred to by conventional duration terms. In Congruent trials (e.g., 3 hours vs. 2 minutes), the numbers provide an additional cue to the correct answer. Even a child with no idea how long either an hour or a minute is might still choose correctly, based solely on his understanding of 3 vs. 2, thus improving overall performance on Congruent relative to Same/No Number trials. We expect the children with the least precise understanding of time words to show the greatest increase in performance in Congruent relative to Same trials. However, in Incongruent trials (e.g., 2 hours vs. 3 minutes), basing the choice on number alone would lead the child to make the wrong choice. While a qualitative understanding that an hour is more time than a minute is sufficient to succeed in the Same or Congruent trials, only a quantitative understanding will suffice on Incongruent trials. Making the correct choice requires sufficient understanding of the relative durations encoded by time words to realize that their ratio far exceeds that of the number words, 3:2. Knowing the order of the time words alone is insufficient, so the Ordinal hypothesis predicts lower performance on Incongruent trials. Only the Ratio hypothesis, in which time words are mapped onto representations of duration, predicts equal success on Incongruent and Same trials.

Overall accuracy in Experiment 2 was similar to that found in Experiment 1 for those age groups represented in both. All groups performed significantly above chance. Proportions correct ( $M \pm$ SEM) for the 4-, 5-, 6-, and 7-yearolds groups, respectively, were $0.55 \pm 0.02,0.71 \pm-0.03$, $0.81 \pm 0.04,0.97 \pm 0.02$. The critical comparison between Same, Congruent, and Incongruent trials is shown in Figure 2. Data were collapsed across time-word comparison types and number sizes, as neither was a significant predictor of children's performance in Experiment 2. Performance in the Same number case was not significantly different from that in the No Number case.

While the 4 -year-old group was both helped by number word congruency and hindered by incongruency, as predicted by the Ordinal hypothesis, the 7-year-olds were near ceiling on the task in all conditions, with no cost to incongruency or benefit to congruency, as predicted by the Ratio hypothesis. The intermediate age groups show different patterns, with the 5 -year-olds showing a cost of incongruency and no benefit to congruency, and the 6-year-
olds showing no cost to incongruency and a benefit to congruency. Strikingly, these results suggest that there are children who know both that 3 is greater than 2 and that an hour greater than a minute, but fail to accurately compare 3 minutes with 2 hours.


Figure 1: Effect of congruency of time word comparisons and number word comparisons in Experiment 2.

Taken together, the results of Experiments 1 and 2 suggest that children learn duration words as a lexical class, and they begin to learn the ordinal structure of that class by age 4, prior to mapping them onto nonverbal representations of duration. Further, children do not have a full understanding of how these words encode relative duration, consistent with the Ratio hypothesis, until at least the age of 7, after they've encountered time words in school. One possibility is that children do not map these words onto specific durations until they learn their definitions. Another possibility is that younger children do associate these words with durations, perhaps relying on their experience hearing them used in relation to familiar events to make these associations, but these representations are imprecise, not easily combined with number knowledge.

A limitation of the forced-choice methodology employed in the first two experiments is that each trial probed knowledge of two different duration words, conflating the participant's knowledge of them which may have precluded finding differences in the acquisition of individual words. To further probe children's ability to estimate the durations encoded by individual time words, and to obtain a more precise measure of participants' ability to rank-order a set of time words, we used the number-line method in Experiment 3. This also allowed us to compare children's ability to estimate the durations of familiar events and conventional time words, and to ask whether overt knowledge of the definitions of the duration words predicted better duration estimation performance.


Figure 2: Proportion of ordinal estimates in Experiment 3 and proportion of correct definitions of duration words

## Experiment 3

Estimation data were analyzed in terms of their distance from 0 along the line representing elapsed duration. Overall time word estimation performance for the three age groups was assessed by plotting each child participant's estimated duration ratios (see Methods) as a function of adults' mean ratios and fitting the data for each age group with a linear model. The closer the slope of that line approximates 1 , the more adult-like the estimation. Slopes for the 5, 6, and 7-year-old groups, respectively, were $0.14,0.57$, and 0.86 . These data confirm that children have essentially no quantitative understanding of the relative durations encoded by these words at the age of 5 (despite their above-chance performance in Experiments 1 and 2), but obtain this understanding in the early school years.

Results from the ordinality measure (see Methods) are shown in Figure 2, alongside results from the follow-up questions testing overt knowledge of the duration words definitions. Comparing time word and event estimation, the five-year-olds perform better with familiar events, lending moderate support to the idea that young children extract duration information from familiar activities and use that knowledge to aid them in learning duration words, via associative mappings. By six, however, children are estimating better overall with conventional time words than without. The probability of successfully rank-ordering the 4 time words is correlated with having explicit knowledge of their definitions. Almost no 5-year-olds but most 7 -yearolds know these definitions. Sorting the 6 -year-old data according to whether each child knows the formal definitions of the words reveals that those who know them perform like 7 -year-olds while those who do not perform like 5-year-olds, highlighting the importance of this factor.

An intriguing possibility is that learning duration words not only improves our ability to estimate the lengths of events described in those terms, but also provides a useful cognitive framework for encoding and estimating the durations of perceived events in general. However, by this account we expect explicit knowledge of time words to
improve performance on duration estimation in both the Event and Time word tasks. However, while we find that accuracy on the follow-up questions (e.g., How many seconds are in a minute?"), when added to a model including age group, was a significant predictor of children's proportions of ordinal responses in the time word task, it did not account for additional variance in event estimation performance.

In conclusion, the three experiments presented here suggest that, prior to acquiring their adult definitions, children learn the nominal category of time words as well as the ordinal structure of that category. However, we find no evidence that children map these terms onto precise representations of duration until after they learn their formal definitions.

## Acknowledgments

We thank Eleanor Chestnut, Katherine Kimura, Saima Malik Moraleda, and Karen Bejar for their assistance.

## References

Colombo, J., \& Richman, W. A. (2002). Infant timekeeping: Attention and temporal estimation in 4-month-olds. Psychological Science, 13, 475-479.
Droit-Volet, S., Tourret, S., \& Wearden, J. (2004). Perception of the duration of auditory and visual stimuli in children and adults. The Quarterly Journal of Experimental Psychology, 57A, 797-818.
Friedman, W. J. (1990). Children's representations of the pattern of daily activities. Child Development, 61, 13991412.

Grant, J. B. \& Suddendorf, T. (2011). Production of temporal terms by 3-, 4-, and 5-year-old children. Early Childhood Research Quarterly, 26, 87-96.
Shatz, M., Tare, M., Nguyen, S. P., \& Young, T. (2010). Acquiring non-object terms: The case for time words. Journal of Cognition and Development, 11, 16-36.
Tare, M., Shatz, M., \& Gilbertson, L. (2008). Maternal uses of non-object terms in child directed speech: Color, number, and time. First Language, 28, 87-100.

# Exploring the Role of Verbal Category Labels in Flexible Cognition 

Jackson Tolins (jtolins@ucsc.edu)<br>Department of Psychology, 1156 High St. Santa Cruz, CA 95064 USA

Eliana Colunga (colunga@psych.colorado.edu)<br>Department of Psychology and Neuroscience, 345 UCB<br>Boulder, CO 80309 USA


#### Abstract

Research under the paradigm of the label feedback hypothesis has proposed a causal role for verbal labels in the online learning and processing of categories. Labeled categories are learned faster, and are subsequently more robust. The present study extends this research paradigm by considering the relationship between verbal labels and flexible categorization. Flexibility is a key trait of human cognition, and flexible categorization is important in a number of tasks. Participants learned to categorize 'friendly' and 'unfriendly' aliens either with or without names, followed by a transfer task. While selective attention to a particular dimension slowed relearning, no effect of label was found for either category learning or relearning with one exception; labels facilitated flexibility when selective attention was not involved in the transfer. The inability to replicate effects of verbal labels in category learning using similar methodologies raises interesting theoretical issues, questioning the extent to which this relationship applies.


Keywords: Categorization; Label Feedback Hypothesis; Flexible Cognition; Selective Attention

## Introduction

Language, along with use in communication, provides a symbolic system of representation through which a speaker contemplates the world around them. The emergence of the capacity for symbolic representation transformed human cognition (Deacon, 1997; DeLoache, 2004), permitting abstract thought and making possible cultural transmission of knowledge. Yet the relationship between language and other cognitive processes is still controversial. For many who view language as a distinct mental module (Gleitman \& Papafragou, 2005; Pinker, 1995), language is merely a formal medium that is used to describe mental representations, while remaining independent of the concepts they express (Li \& Gleitman, 2002). Recent work in understanding the relationship between language and thought has provided evidence against this disassociation. Instead, it has been suggested that language is best understood as built upon domain general cognitive processes, and thus potentially in a mutually transformative relationship with these processes (Bowerman \& Choi, 2001; Gumperz \& Levinson, 1996).

With habitual use of the specific set of conceptual symbolic representations afforded by a language, an individual may be biased towards these representations in problem-solving and other cognitive tasks. How a language may accomplish this is not well understood. One possibility
is that language reduces the ability to flexibly adjust categories outside the structure provided for by the words of a particular language. As such, it is important to consider the influence of language on the ability to dynamically activate and modify the cognitive process of categorization in response to changing task demands. The ability to think and act adaptively, while not a uniquely human trait, is a mental capacity uniquely well developed in human cognition and intelligent behavior (Deák, 2003). For the purposes of the current study, flexible cognition will be defined as a property of the cognitive system, rather than a specific mechanism or process (Deák, 2003; Ionescu, 2012). This definition allows for the consideration of flexible cognition in the interaction of interest; that between categorization and language, specifically verbal labels.

Recent work lead by Gary Lupyan and colleagues on the role of labels in categorization has demonstrated a special status afforded to verbal labels (see e.g. Lupyan, Rakison, \& McClelland, 2007; Lupyan \& Thompson-Schill, 2011). Verbal labels participate in the learning of categories, facilitating learning, creating mental categories that are more robust than when the categories are learned without words (Lupyan, Rakison, \& McClelland, 2007), and encouraging selective attention (Brojde, Porter, \& Colunga, 2011). However, no study has looked directly at the influence of verbal labels on the perceptual and attentional processes that underlie flexibility after learning. Similarly, while a number of studies have looked at how language aides in an individual's ability to flexibly adjust the level of categorization, or switch from taxonomic to thematic (Blaye, Bernard-Peyron, Paour, \& Bonthoux, 2006), no previous research has investigated how individuals flexibly adjust their categorization strategies in regards to the same domain, on the same level. The present investigation seeks to illuminate further the relationship between verbal labels and the cognitive processes underlying categorization. In developing an understanding of the role that verbal labels play in the construction and maintenance of categories, we further our understanding of the relationship between language and the domain general cognitive processes, such as categorization, upon which language is built.

## Background

## Flexibility in Categorization

Categorization, the process by which discriminably different things are classified into groups and therefore responded to in kind, is a ubiquitous cognitive operation relevant to all aspects of human life. How categories are learned is a key issue in understanding the relationship between verbal category labels and flexibility in cognition. A number of studies have demonstrated that the relationship between perceptual descriptions, how the category or concept is defined, and conceptual representations, such as verbal labels, are mutually influential (Goldstone, 2000; Lin \& Murphys, 1997). It is widely accepted that adults tailor the categories they form to the current demands of the task or situation (Barsalou, 1983), and can spontaneously group objects in several ways (Ross \& Murphy, 1999). Categorical flexibility is thus a within-subject variable corresponding to the ability to switch, (or relearn), between different representations of a given object or set of objects.

Related work has focused on the way that categorization influences perceived similarities (e.g. Goldstone, Lippa, \& Shiffrin, 2001). According to these studies, conceptual and categorical flexibility must be accompanied by flexibility in perceptual and attentional processes (Goldstone 1998). Two mechanisms are considered key to perceptual category learning and flexibility: selective attention and differentiation of dimensions (Goldstone \& Steyvers, 2001). Selective attention refers to the process by which, in categorization learning, individuals learn to attend to some features of the objects and ignore irrelevant features. Selective attention is key to models of categorization such as Nosofksy's (1986) exemplar model, in which an object is measured in similarity compared to a stored category member in a multidimensional space. The distances between points along dimensions within this space compress and expand depending on the attention given to particular dimensions. Dimensional differentiation refers to the psychological process by which previously unified dimensions become perceptually and cognitively distinct. For example, in developing categories for circles and squares one must first learn to separate the dimension of shape from task-irrelevant dimensions such as color or size. In order to study these mechanisms, Goldstone \& Steyvers (2001) applied a learning/transfer task, wherein subjects first learned to distinguish between two categories, and then at transfer had to relearn the categories based on altered relevance of dimensions. By making dimensions that were previously diagnostic for categorization unimportant, or the reverse, allows for a measure of the role of selective attention in categorical flexibility. Similarly, new dimensions may exist in the transfer stimuli set that did not exist in the training set, allowing a separate measure of dimensional differentiation.

## Categorization and Verbal Labels

The processes of selective attention and dimensional differentiation in categorization lead stimuli to be considered more similar when in the same category, and more easily distinguishable when in different categories (Harnad, 1987). Recent studies have demonstrated that verbal labels influence categorization, speeding up the attentional processes that focus in on diagnostic properties of categorized objects. It has been suggested that simply sharing a label, defined as a name for a category, causes two objects to be perceived as more similar than those that do not (Lupyan et al., 2007).

There are a number of explanations for this relationship. Researchers have provided evidence that labels offer more maximally informative feedback during categorization learning, making rule-based categories, those categories that are learned explicitly with diagnostic rules that are easily verbalized (Ashby \& Maddox, 2005), easier to learn (Maddox et. al, 2008). Others consider labels as physical, external symbols upon which our categories are hung (Clark, 2006; Lupyan et al., 2007). In this sense, language is viewed as a self-constructed cognitive niche, with words providing the material scaffolding required to promote abstract thought and reason, by providing a target for more basic capacities such as statistical and associative learning (Clark, 2006). These latter theories have been generalized by Lupyan within the Label Feedback Hypothesis framework (Lupyan, 2007).

Labels have been implicated in the learning of categories, but what of their maintenance and adjustment? Lupyan, Rakison, and McClelland (2007) provided evidence that categories associated with verbal labels are not only learned faster, but are maintained more robustly after initially training. If one of the main uses of language is the creation of associations between concepts and words in such a way that the labeled concepts are learned fast and remain more robust, it is possible that a verbal label will also reduce the categorical flexibility by strengthening selective attention to a diagnostic dimension. In contrast, if labels, as suggested by Maddox et al. (2008), simply aid in categorization of rule-based categories by providing a more maximally informative feedback mechanism, it is possible that labels may also positively affect categorical flexibility.

## The Current Investigation

The present study seeks to add to the literature on labels and categorization by investigating the rigidity of categorization both with verbal labels and in their absence. When an individual needs to restructure the categorical divisions of a particular domain, especially when this restructuring requires a shift in attention to a previously non-diagnostic dimension, having verbal labels for categories already established could slow down the relearning curve. The influence of verbal labels on learned sensitivity to dimensions was tested using a category-learning paradigm in which participants received an initial category learning followed by a relearning transfer task, in which either the
diagnostic dimension changed, requiring a shift in selective attention, or the behavioral response but not the diagnostic changed (see Transfer Procedure below).

## Methodology

Subjects 192 participants were drawn from the undergraduate psychology subject pool at CU, Boulder, in exchange for course credit. Subjects were randomly assigned to either label or no label training conditions and one of three transfer conditions, giving six total conditions.

Materials Categories were organized based on the kind of eyes "aliens" exhibited. To this end, 36 gabor patches were created, varying along the dimensions of frequency and orientation (figure 1); these patches were embedded in the stimuli as the aliens' eyes.


Figure 1: Example stimuli demonstrating the range of frequencies across and orientation downward.

Training Procedure Following the procedure from Lupyan, Rakison, \& McClelland (2007), participants were told that they were to take part in a NASA training program before traveling to a newly found planet. In training, it was explained that previous explorers to the planet had discovered two aquatic alien species, one of which was friendly and could be approached, and one that was dangerous and should be avoided. In the label conditions, the participants were told that the explorers had decided to name the aliens, and that the friendly aliens were named 'Gowachi', while the dangerous aliens were named 'Caleba'. Thus, participants were asked to learn to distinguish between two categories within a set of novel stimuli. This distinction was based on either the orientation or the frequency of the alien's giant eye. Individual trials began with a fixation marker in the middle of the screen, presented for 500 milliseconds. For each trial, an alien was presented briefly, ( 500 ms ), before a scuba diver appeared in one of four locations; above, below, or on either side of the alien. The participant then decided whether to approach or escape the alien using the directional keys on a standard keyboard. For example, if a scuba diver appeared on the left
of a friendly alien, the participant should press the "right" key to move the scuba diver closer. After a response was made, feedback was provided in either minimal (a chime for correct, a buzz for incorrect) or maximal (minimal feedback + correct category label) conditions. If the participant waited for longer than 3 seconds, feedback was given without response. After the feedback, the alien and scuba diver remained on the screen for additional 800 ms before the start of the next trial and the representation of the fixation marker. Each unique alien + diver trial was presented once in random order, for a total of 144 trials of training ( 36 alien exemplars x 4 diver locations). All subjects received the same number of categorization learning trials and had equal exposure to the stimuli across conditions.

Transfer Procedure After training was complete the participants were told that they were now ready to travel to the Planet Teeb. In all but the control, or 0 degree, transfer conditions, upon arrival on the planet the participants were alerted that something has gone wrong, and that the aliens are not behaving as expected. Participants in these conditions faced two distinct relearning tasks. In the 90 degree transfer condition, the diagnostic dimension changed, requiring a modulation in selective attention. Participants who learned during training that the friendly aliens had thick bands in their eyes, and the unfriendly aliens thin ones, here had to learn to categorize the friendly and unfriendly aliens based on the steepness of the orientation of the bands, ignoring thickness. This meant that half of each category learned during the first phase subsequently became part of the new category structure learned during transfer testing, or that half of the Gowachi must now be considered Caleba and the reverse. For the 180 degree transfer condition, the diagnostic dimension remained the same, but the escape/approach responses were switched. Here, participants who first learned that aliens with steeply oriented bands in their eyes were friendly now had to learn to treat them as unfriendly, or that the Gowachi and Caleba were opposite what had been learned. These two transfer conditions were compared to the 0 degree transfer condition, in which no change between the training and testing occurred.

Having all conditions transfer to the same categorization allowed for a clear relationship between initial categorization and participants' ability to relearn categorization strategies flexibly (see e.g. Goldstone \& Steyvers 2001). The post-transfer phase consisted of a second set of 144 randomized trials. During the transfer phase trials only minimal feedback (chime or buzz) were given in all conditions, whether label or no label.

## Results

Trials were grouped into blocks of 36 , giving four blocks each for training and transfer phases. Each correct trial was scored as 1 , each incorrect trial as 0 , and each trial in which the participant did not answer was dropped. Accuracy across
block was then calculated. The data from those participants who did not reach at least $50 \%$ accuracy by the end of training were not included (13 participants in total). Data was then entered into a mixed factor ANOVA.

First, we tested for an effect of label on training, collapsing across transfer type, to see if previous findings on the advantage of having a label would replicate, (Lupyan, Rakison \& McClelland, 2007). However, while participants did learn to categorize correctly $F(3,438)=103.42(p<$ .001 ), there was no main effect of label type on this learning trajectory $(p=.312)$. A similar pattern was seen in the testing phase, with a significant effect of block $(F(3,438)=$ $13.140 p<.001$ ), without an effect of label, or a label by block interaction $(F(3,438)=1.263, p=.287)$, (See Figure 2). There was one significant four-way interaction involving label that will be discussed below.


Figure 2: Average accuracy by block for label and no label conditions collapsing across transfer type.

Turning from label to transfer type, while there was no main effect of transfer type $(F(2,146)=.104, p=.901)$, there as a significant interaction between phase (whether training or transfer) and transfer type $(F(2,146)=80.553, p$ $<.001$ ), with accuracy worse when transfer required a switch in selective attention, (see figure 3).


Figure 3: Learning (blocks $1-4$ ) and transfer (blocks $5-8$ ) trajectories for the 0 degree, 90 degree, and 180 degree transfer conditions.

Of final interest was a significant four-way interaction between phase, block, label type, and transfer type ( $F(6$, $438)=2.18, p<.05$ ). As this was the only significant interaction involving label, this interaction was pursued further, with the analysis first involving separating out each transfer type. For the 0 and 90 degree transfer conditions, there were the expected effects of block and phase (all $p \mathrm{~s}<$ .01 ), but no main effects or interactions involving label (all $p \mathrm{~s}>.05$ ). In the 180 degree transfer condition, however, a significant interaction of phase*block*label type was found, $(F(3,132)=4.527, p<.05$. Using a general linear model to explore this interaction further, we found that for the first two blocks of transfer in the 180 degree transfer condition, there was an interaction between block and label type ( $F(1$, $44)=11.595, p<.001$ ), (see figure 4). Thus, there was evidence for an effect of label on transfer learning in the condition that required not a shift in attention, but a shift in the behavioral response from what had been learned in training.


Figure 4: Average accuracy by block for participants in the label and non-labeled conditions of the 180 degree transfer.

## Discussion

The results of the present experiment did not find support for a general advantage for learning categories with labels over categories without labels as seen in previous similar experiments (Lupyan et al., 2007). One important difference between the stimuli used here and that used by Lupyan and colleagues is that their aliens were categorized by shape features, whether the ones in the present study were categorized by what could be seen as textural features. The lack of a label advantage in learning is in line with previous work showing that the effect of labels depends on the sort of categorization being learned (Brojde et al., 2011).

More interesting to the question of this paper, however, is the way labels influenced performance at transfer. In the 180 degree condition, when participants had to relearn that those aliens who had been approachable were now not approachable and vice versa, there was a significantly faster recovery after transfer for those participants who were provided with labels during training. Our results suggest that labels play a positive role in the relearning of
categorization when the boundaries of the categories do not change, and the relevant dimension does not change, but the categorical behavioral responses do, (i.e. whether the astronaut should approach or retreat from the alien). Having verbal labels for the categories allowed the participants to more flexibly adapt to the changing task demands. It is possible that since verbal labels become attached to the categories which they are used to express (Lupyan \& Thompson-Schill, 2011), that when the categories themselves do not change, but only the responses change, these labels continue to act as more easily computed symbolic abstractions of the categories for which they stand. It then becomes possible for the participants in the 180 degree transfer condition to switch from 'Gowachi' and 'Caleba' to 'not Gowachi' and 'not Caleba'.

The visibility of this effect of label on transfer flexibility seems to be made possible by the low cost of transfer when the transfer does not involve modulation of selective attention. The cost of transfer, however, was much larger for those who had to relearn their categorization strategies based on a previously unimportant dimension. Those participants who learned during training to categorize based on frequency of the lines of the eyes and discovered on the planet that the aliens were either friendly or unfriendly based on the orientation demonstrated reduced ability to flexibly adjust to this new categorization strategy. While selective attention is an important process in the development of accurate categorization (Goldstone, 1998), it also reduces the degree of flexibility present in categorization cognitive processes.

At transfer, these participants must not only learn to pay more attention to the previously ignored dimension, they must also inhibit attention to the previously diagnostic features (Goldstone \& Steyvers, 2001). This is demonstrated by the comparison of the four blocks of training for the identification condition with the four blocks of transfer for those participants whose transfer included a change in the diagnostic dimension, despite having had 144 trials more experience than those approaching the task for the first time. This is a clear indication of the cost that comes with increased attention to one historically predictive dimension combined with decreased attention to all other dimensions. This is in contrast with Goldstone and Steyvers (2001), who found that when the categorization rules are orthogonal, participants do no differently than those learning a completely new set. Their analysis of this finding was to posit an equalizing effect of negative transfer from selective attention with positive transfer from dimensional differentiation, meaning that regardless of the type of transfer, it helps to have practice in separating the two perceptual dimensions of the stimuli. By matching the same categorization strategy across training, taken as a control, and transfer, rather than having participants relearn a completely new category during transfer, we demonstrate that the positive effect of dimensional differentiation is not large enough to make the performance of those participants
who transferred across dimensions on par with those coming to the same task without any previous experience.

This role of selective attention in reducing flexibility was not, however, modulated by the presence of verbal labels corresponding to the categories being learned. While participants did learn the correct categories over the course of training, across all conditions this learning trajectory was not modulated by the presence or lack of label as feedback on individual trials. Similarly, transfer-learning trajectories were not significantly affected for those participants whose initial training included verbal labels, for better or for worse. The inability of the current data to replicate previous findings on the influence of verbal labels in category learning draws into question the extent to which the Label Feedback Hypothesis can be extended into categorization.

Previous studies that have demonstrated a positive influence of verbal labels have focused mostly on shapebased categories, including the study upon which the present study is based (e.g. Lupyan et al., 2007; Lupyan \& Thompson-Schill, 2011). Very early in language learning, English-speaking children develop a bias towards categorizing labeled object categories based on shape (Yoshida \& Smith, 2005; Colunga \& Smith, 2005). It's possible that, as shape-based categories are based on dimensions that are historically predictive for English language speakers, the effect of labels during this type of categorization would be stronger than for other types of learning. This is supported by findings from Brojde, Proter, and Colunga (2011), who demonstrated that verbal labels hinder category learning defined by texture or brightness. They argue that the advantage of label comes about only when the relevant dimension aligns with the relevant dimensions in previous similar tasks, which in the case of our English-speaking participants would be shape over features such as orientation and frequency of line.

## Conclusion

The purpose of the current investigation was to assess the effect of verbal labels on the ability to flexibly adjust categorization strategies when faced with changes in the environment. Previous literature (Lupyan et al., 2007; Lupyan \& Thompson-Schill, 2011) has demonstrated that verbal labels influence category learning, improving both speed of learning and strength of representation. Some have argued that this effect of verbal labels is a demonstration of the top-down modulation of labels during learning and therefore shows that verbal labels are directly involved in the learning of concepts and categories (Lupyan, 2009). In this theory, labels work as material symbols upon which categories are attached (Clark, 1996), and so take part in the category learning process, possibly by modulating selective attention (Goldstone, 1998; Goldstone \& Steyvers, 2001). Others, however, have argued that verbal labels are simply a more maximal form of feedback, and are therefore simple a form a facilitation, separate from the categories themselves (Maddox et al., 2008). In order to tease these two views apart, the present study considered the role of verbal labels
in flexible cognition, more specifically the ability of individuals to flexibly adjust their categorization strategies.

Despite a replication of the effect of selective attention across stimuli dimensions, the previous finding of the positive effect of labels as feedback for category learning was not replicated. The failure to replicate a positive effect of label on category learning raises questions as to the generalizability of the label feedback hypothesis. Given the issues raised in the current study above, it appears that not all types of category learning benefit from the presence of verbal labels (see also Brojde et al., 2011). Similarly, there is no evidence that labels modulate selective attention in a way that would either help or hinder flexibly adjusting one's categorization strategies. There was however, an effect in a single transfer condition that demonstrates that labels may aid in recovery from transfer when the type of transfer does not involve a change in selective attention. In the 180 degree transfer condition, while labels did not have a positive effect on learning during training, labels did interact with accuracy immediately after transfer, allowing those who learned with labels to recover faster. Future endeavors could continue to develop an understanding of the relationship between concepts, categories, and the words we use to invoke them.

## Acknowledgments

We would like to thank Bhuvana Narasimhan, Claire Sims for their comments on an earlier draft, and the members of the CU language project for assistance in collecting data.

## References

Ashby, F. G. \& Maddox, W. T. (2005) Human category learning. Annu. Rev. Psychol., 56, 149-178
Barsalou, L. W. (1983). Ad hoc categories. Memory and Cognition, 11, 211-227
Blaye, A., Bernard-Peyron, V., Paour, J., \& Bonthoux, F. (2006) Categorical flexibility in children: Distinguishing response flexibility from conceptual flexibility, the protracted development of taxonomic representations. European journal of devel. psych., 3, 163-188
Bowerman, M. \& Choi, S. (2001) Shaping meaning for language: Universal and language-specific in the acquisition of spatial semantic categories. In: M. Bowerman \& S. C. Levinson (Eds.) Language acquisition and conceptual development. Cambridge University Press, pp. 475-522
Brojde, C. L., Porter, C., \& Colunga, E. (2011) Words can slow down category learning. Psychon Bull Rev.
Clark, A. (2006) Language, embodiment, and the cognitive niche. TRENDS in cognitive sciences, 10, 8, 370-374
Colunga, E. \& Smith, L. B. (2005) From the lexicon to expectations about kinds: A role for associative learning. Psychological Review, 112, 347-382
Deacon, T. W. (1997) The symbolic species: The coevolution of language and the brain. W.W. Norton
Deák, G. (2003) The development of cognitive flexibility and language abilities. In: Kail, R. (Ed.) Advances in child
development and behavior. Sand Diego: Academic Press. Pp. 271-327
DeLoache, J. S. (2004) Becoming symbol-minded. Trends in Cog. Sci., 8, 66-70
Gleitman, L. \& Papafragou, A. (2005) Language and thought. In: K. Holyoak \& B. Morrison (Eds.) Cambridge handbook of thinking and reasoning. Cambridge, University Press, pp. 633-661
Goldstone, R. L. (1998) Perceptual learning. Annu. Rev. Psychol., 49, 585-612
Goldstone, R. L. (2000) Unitization during category learning. Journal of Experimental Psychology: Human perception and performance, 26, 86-112
Goldstone, R. L. \& Steyvers, M. (2001) The sensitization and differentiation of dimensions during category learning. Journal of experimental psychology: General, 130, 1, 116-139
Goldstone, R. L., Lippa, Y., \& Shiffrin, R. (2001) Altering object representations through category learning. Cognition, 78, 27-43
Gumperz, J. J., \& Levinson, S. C. (1996) Rethinking linguistic relativity. Cambridge, Endlang: Cambridge University Press
Ionescu, T. (2012) Exploring the nature of cognitive flexibility. New ideas in psychology, 30, 190-200
Li, P., \& Gleitman, L. (2002). Turning the tables: Language and spatial reasoning. Cognition, 83, 265-294
Lin, E. L., \& Murphy, G. L. (1997). The effects of background knowledge on object categorization and part detection. Journal of Experimental Psychology: Human Perception and Performance, 23, 1153-1169
Lupyan, G. (2007). The label feedback hypothesis: Linguistic influences on visual processing. Unpublished doctoral dissertation, Carnegie Mellon University, Pittsburgh, Pennsylvania.
Lupyan, G., Rakison, D. H., \& McClelland, J. L. (2007) Language is not just for talking: Redundant labels facilitate learning of novel categories. Psychological Science, 18, 12, 1077-1083
Lupyan, G. \& Thompson-Schill, S. L. (2011). The evocative power of words: Activation concepts by verbal and nonverbal means. Journal of Experimental Psychology
Maddox, W. T., Love, B. C., Glass, B. D., \& Filoteo, J. V. (2008) When more is less: Feedback effects in perceptual learning. Cognition, 108, 578-589
Nosofsky, R. M. (1986) Attention, similarity, and the identification-categorization relationship. Journal of Experimental Psychology: General, 115, 39-57
Pinker, S. (1995) The language instinct: The new science of language and mind. Penguin Books
Ross, B. H. \& Murphy, G. L. (1999) Food for thought: Cross-classification and category organization in a complex real-world domain. Cognitive Psychology, 38, 495-553
Yoshida, H. \& Smith, L. B. (2005) Linguistic cues enhance the learning of perceptual cues. Psychological Science, 16, 564-577

# Diverse Evidence for Dissociable Processes in Inductive Reasoning 

Eoin Travers (etravers01@qub.ac.uk)<br>School of Psychology, Queen's University Belfast<br>University Road, Belfast BT7 1NN

Aidan Feeney (a.feeney@qub.ac.uk)<br>School of Psychology, Queen's University Belfast<br>University Road, Belfast BT7 1NN


#### Abstract

Previous work suggests that inductive and deductive reasoning may be accomplished by different processes. Here, we examine whether different phenomena of inductive reasoning, previously explained in the same way, may rely on different types of processes. In Experiment 1 we show that trials which examine sensitivity to sample size in inductive reasoning have greater effects on secondary task performance than do trials examining sensitivity to the diversity of the sample. In Experiment 2 we show that in a surprise recognition memory test, participants have significantly better memory for the content of diversity trials than for sample size trials. Both findings are consistent with the suggestion that some phenomena of inductive reasoning may be rule-based, whereas others may depend on feature-level processing.


Keywords: Reasoning; induction; diversity effect; law of large numbers.

## Introduction

Not all thinking is the same. Because the same experimental manipulations affect them differently, it has been claimed that inductive and deductive thinking are dissociated (see Rips, 2001; Heit \& Rotello, 2010). Heit and Rotello argue that deductive reasoning calls more on processes that are sensitive to logical validity, whereas inductive reasoning relies more on associative processes. However, a background assumption appears to be that inductive reasoning, for example, consistently draws on the same processes, and most theories of inductive reasoning attempt to capture different experimental phenomena in the same way (see Osherson et al, 1990; Sloman, 1993; Rogers \& McClelland, 2004). Here, we will consider whether different processes underlie different phenomena that have been observed in people's inductive reasoning. Specifically, we will examine whether sensitivity to the size of the sample upon which an inductive inference is based may be due to rule-based processes, whereas sensitivity to the diversity of the evidence may call on more feature-based processing.

## Sensitivity to the size and diversity of samples

Despite claims made by Kahneman and Tversky (1972), there is much evidence that adults, and sometimes children, are sensitive to sample size (see Piaget \& Inhelder, 1975; Nisbett, Krantz, Jepson \& Kunda, 1983). In experiments on category-based induction, where participants are typically taught that members of certain categories possess a novel property and are asked whether members of some other
category also possess that property, the tendency to prefer arguments based on a larger sample of categories is known as the monotonicity effect (see Osherson et al., 1990). However, not everyone displays the monotonicity effect in such experiments (see Feeney, 2007).

Whereas sensitivity to sample size has been intensively studied in the literature on judgment and decision making, sensitivity to sample diversity has most often been studied in the literature on category-based inductive reasoning (for a review, see Heit, Hahn \& Feeney, 2005). Although preference for more diverse evidential samples has been informally advocated by a variety of philosophers of science (e.g. Bacon, 1878; Carnap, 1950; Popper, 1963), attempts to formally justify a diversity principle are rarer and there are arguments against the existence of a general principle (see Lo, Sides, Rozelle \& Osherson, 2002; Medin et al., 2003). Nonetheless, there are numerous demonstrations in experiments on category-based induction in which a majority of people consider arguments with more diverse premises to be stronger. People are sensitive to the diversity of the evidence, at least some of the time.

## Accounts of sensitivity to sample size and diversity

Accounts of sensitivity to sample size are to be found in a variety of literatures whereas sensitivity to sample diversity is accounted for only by theories of category-based induction. Fong, Krantz \& Nisbett (1986) claimed that sensitivity to sample size occurs because people possess intuitive but abstract rules that correspond to the law of large numbers, and showed that sensitivity to the law of large numbers can be enhanced by training. This account is similar in some respects to Piaget's (Inhelder \& Piaget, 1975). In particular, both accounts stress the centrality of sensitivity to sample size to reasoning about probability more generally. Stanovich and West (1999) offer a dual process account of sensitivity to sample size, where such sensitivity when it is observed, is the result of effortful processes that draw on working memory in order to apply normatively justified rules or principles for reasoning.

Sensitivity to sample size, or adherence to the monotonicity principle, is explained very differently in models of category-based induction. For example, the similarity-coverage model (Osherson et al., 1990) holds that arguments are strong to the extent that the conclusion category is "covered" by the categories in the premises.

That is, to the extent that instances sampled at random from the conclusion category are similar to the categories in the sample. As a larger sample is more likely to better cover the conclusion category than a smaller sample, people judge arguments based on larger samples to be strong. Sloman (1993) predicts that arguments will be judged strong to the extent that there is overlap in the features of the conclusion category and the features of the categories in the sample. This account predicts sensitivity to sample size on the grounds that larger samples, on average, lead to greater featural overlap. Notably, all accounts of category-based induction, including Bayesian models (Tenenbaum, Kemp \& Shafto, 2007) explain sensitivity to sample size and diversity in the same way.
In summary, different explanations of sensitivity to sample size posit different types of process. Early developmental and decision making accounts posit the existence of abstract and intuitive rule-like representations which, according to some accounts (see Stanovich \& West, 1999) are effortfully applied. On the other hand, accounts of sensitivity to sample size in the literature on category-based induction appeal to processes operating over the relations between specific members of the sample. Some accounts (e.g. Sloman, 1993; Rogers \& McClelland, 2004) hold that the application of these processes is relatively effortless. Accounts of sensitivity to sample diversity appear only to be found in the literature on category-based induction, and are similar to the accounts of sensitivity to sample size that are to be found in the same literature.

## Dissociating the effects: Two different paradigms

The goal of the experiments to be described below was to examine whether similar or different processes underlie the sample diversity and sample size effects in induction. To achieve this goal we derived hypotheses about possible differences between the two phenomena in terms of the effort required by each and about the side effects of the underlying reasoning processes.

Effort and secondary tasks To the extent that models of category-based inductive reasoning are correct in assuming that sensitivity to sample size and diversity require the operation of the same processes, we should expect to find no differences between the effort required in order to demonstrate each effect. However, if sensitivity to sample size requires the operation of a rule-based process (Fong et al, 1986) that draws on working memory (Stanovich \& West, 1999) then we might expect to be able to show that sample size trials require more cognitive effort than do diversity trials. To test this hypothesis we presented reasoning trials (the primary task) concurrently with a memory task (the secondary task). Such designs have previously been employed to test hypotheses about the effort required by particular types of thinking (see De Neys, 2007). If sample size materials require more effort than diversity materials, we should expect to observe (a) a greater effect of the secondary task on sensitivity to sample
size than on sensitivity to diversity; or (b) greater effects of the sample size task than the diversity task on the secondary task; or (c) both effects. The first experiment to be described below tested these hypotheses.

Induction then recognition $A$ contentious claim in the literature is that the processes applied during reasoning may have consequences for the type of representation which reasoners construct of the problem material, and hence for their ability to accurately recognize the materials they reasoned about (see Sloutsky \& Fisher, 2004). There is evidence that following a simple inductive reasoning task, children have better recognition memory for the problem materials than to adults, although they perform equally well on the reasoning task. Sloutsky and Fisher (2004) claimed that this recognition memory effect was a consequence of children and adults using different processes to reason. They claim that adults reason on the basis of category membership and therefore construct category-level or gist (Brainerd, Reyna \& Forrest, 2002) representations of the reasoning stimuli. Children, on the other hand, reason on the basis of correspondences or similarities between the entities in the reasoning problem. This leads them to construct a verbatim (Brainerd et al., 2002) representation of the materials. When all participants are subsequently presented with old pictures and new critical lures, it is children with their more detailed representation of the original materials who are better able to discriminate between old and new items.
Although there has been disagreement about whether the original induction-then-recognition experiments necessitate conclusions about developmental changes in reasoning processes (Wilburn \& Feeney, 2008; Hayes, McKinnon \& Sweller, 2008), the paradigm may be a very useful tool for determining whether different reasoning phenomena are caused by different reasoning processes. For example, if sensitivity to sample size in category-based induction is due to the application of an intuitive rule, then we would not expect participants to encode verbatim representations of the reasoning stimuli. On the other hand, if sensitivity to diversity requires representation of the relations between the entities in the reasoning problems, then participants should be more likely to construct verbatim representations of the entities in those reasoning materials. This difference in the type of representation that is constructed might have consequences for participants' ability to subsequently recognize the entities that they have previously reasoned about. Specifically, participants may have better recognition memory for diversity materials than for sample size materials. On the other hand, if the same processes are involved in sensitivity to both phenomena, we would expect no differences due to reasoning phenomena in recognition accuracy. Experiment 2 below will test these hypotheses.

## Experiment 1

The aim of this experiment was to test for differential effects of a secondary task on sensitivity to sample size and
diversity in inductive reasoning, and to test for differential effects of these reasoning phenomena on performance of a secondary task.

To facilitate the use of the Induction then Recognition paradigm in Experiment 2, across both experiments we adopted a paradigm recently used to test for diversity effects in children (Rhodes, Brickman \& Gelman, 2008) in which participants are asked to select between a diverse and nondiverse sample of category members in order to help them decide whether all members of the category possess a novel property.

## Method

Participants Sixty students (29 males) were recruited in a quiet area of the library at QUB, and paid $£ 2$ each to take part in the study. The mean age was 28.63 years.

Materials In each reasoning task, participants were told about a novel property that might be possessed by all members of a category, alongside two samples of members of that category, and were asked which sample they would like to test in order to decide whether all members of the category possess the property. On the five trials assessing sensitivity to diversity, the diverse sample consisted of pictures of two category members of different coloration, species, or breeds (in the case of dogs), while the nondiverse sample consisted of one of the diverse sample members, and another similar category member. On the five trials assessing sensitivity to sample size, the small sample consisted of two category members, and the large sample consisted of the same two category members plus one additional member. Unique categories, images, and properties were used on each trial.

Because of the possibility that participants might complete the sample size trials without processing the content of the images, we included five control trials at the end of the experiment which asked participants to choose between a small diverse sample and a larger non-diverse sample. If some participants complete the sample size trials without processing the content of the images in the samples, we should find that participants choose the large sample in the control trials as often as in the sample size trials. In addition, there should be a strong correlation between the tendencies to choose the large sample in both types of trial.

The secondary task (see De Neys, 2006) required participants to memorize an array of dots on a $3 x 3$ matrix before each reasoning task, and recreate it immediately afterwards.

Procedure All participants completed the experiment on a laptop computer running E-Prime software. They were told before beginning that the experiment would investigate how people make judgments about category members and their properties. On each trial, participants were presented with a statement at the top of the screen, with the two possible samples below it on either side. They were instructed to press the ' 1 ' button to choose the left sample, and the ' 2 '
button to choose the right sample. There were two practice reasoning trials before the experimental trials began. The first ten trials tested for sensitivity to sample size and diversity and their order was randomized separately for each participant. The final five trials pitted a two-member diverse sample against a three-member homogenous sample.

Before the beginning of each trial, participants were presented with a $3 x 3$ grid for 1000 ms , containing either four dots in random positions (complex condition), or three dots in a straight or diagonal line across the grid (simple condition). After given a response in each reasoning trial, they saw a blank grid, and were required to recreate the pattern seen previously. Participants were instructed to remember the dot pattern as well as they could, while still paying attention to the reasoning task.


Figure 1: Timeline for trials in Experiment 1. At (a) participants were presented with a simple or complex spatial array to memorize; after 1000 ms they were presented with a reasoning problem (b) requiring them to choose one of two samples; and (c) once they chose a sample they recreated the spatial array.

## Results

Primary task performance Across secondary task conditions, participants showed sensitivity to diversity on only $52.3 \%$ of trials ( $\mathrm{SD}=24 \%$ ), and sensitivity to sample size on $72 \%$ of trials ( $\mathrm{SD}=28 \%$ ). A 2 (secondary task: complex vs simple) x 2 (trial type: montonicity, diversity, \& control) mixed ANOVA revealed a main effect of trial type only, $\mathrm{F}(1,58) 21.44, \mathrm{p}<.001$. Neither the effect of load nor the interaction between trial type and load achieved statistical significance.

Secondary task performance Participants' ability to correctly recall the dot arrays broken down by complexity condition and trial type is to be seen in Figure 2. A 2x2 mixed ANOVA revealed a significant main effect of complexity condition, $\mathrm{F}(1,58)=50.90, \mathrm{p}<.001$, and a significant interaction between complexity condition and
trial type, $\mathrm{F}(1,58)=12.88, \mathrm{p}=.002$. Post hoc tests on the means involved in this interaction revealed that reasoning about diversity trials had a significantly greater effect on simple secondary task performance than did reasoning about sample size trials, $\mathrm{t}(29)=2.92, \mathrm{p}<.01$. However, performance on the complex secondary task was affected to a greater degree by sample size trials than by diversity trials, $\mathrm{t}(29)=2.48, \mathrm{p}<.02$.


Figure 2: Interactive effect of trial type and secondary task on secondary task performance in Experiment 1.

Control performance One potential issue with interpreting the results of this experiment and the next is that participants may complete the sample size trials by simply counting the number of images in each sample without processing the content of the samples. One finding that suggests this did not happen is that the mean inspection time for sample size trials was almost identical ( 6377 ms ) to the mean inspection time for the diversity trials ( 6380 ms ). In addition, analysis of the control trials revealed that participants selected the large sample in the control trials $60 \%$ ( $\mathrm{SD}=30 \%$ ) of the time which is significantly less often, $\mathrm{t}(59)=2.36, \mathrm{p}<.03$, than in the sample size trials. If participants had not been processing the content of the samples but only their size, we would have expected the rate at which the large sample was chosen to be virtually identical in these two conditions. In addition, there was almost no association between the tendency to select the large sample in the sample size and control trials, $r(60)=$ . 02.

## Discussion

Participants in Experiment 1 demonstrated less sensitivity to sample size than to sample diversity, and they performed better on the simple than on the complex secondary task. Furthermore, performance on the complex secondary task was significantly worse when the primary task required sensitivity to sample size than when it required sensitivity to sample diversity. On the other hand, performance on the simple secondary task was worse when the primary task required sensitivity to diversity. These results are consistent
with the claim that different processes underlie the sample size and diversity effects. The findings for the complex secondary task, in particular, suggest that participants who are sensitive to sample size may possess a simple rule. Because the operation of such a rule requires general cognitive processes related to working memory (see Stanovich \& West, 1999), performance of a complex secondary task which also requires working memory, is particularly impaired. Fong et al. (1986) suggested that the sample size rule is abstract but intuitive. Its intuitiveness may explain why sensitivity to sample size was observed on a relatively high proportion of trials, and why performance on the simple secondary task was barely impaired when the primary task required sensitivity to sample size.

Notably, performance on the primary task was not affected by the nature of the secondary task and it is not clear why this was the case. One possibility is that participants prioritized the reasoning task.

## Experiment 2

The aim of Experiment 2 was to provide further evidence for dissociation between sensitivity to the size and diversity of the sample in inductive reasoning. To do this we asked participants to complete a surprise recognition memory test once they had completed the reasoning items. If sensitivity to sample size involves the application of an intuitive rule, then we might expect participants to build a gist rather than a verbatim representation of the content of the samples. This representation should lead to relatively poor recognition memory for the entities in the samples. Memory for the entities presented in the diversity trials should be more accurate, if sensitivity to diversity depends on more featurebased processing of the images in the samples. Such processing should be more likely to result in verbatim representations of the pictures in the samples which will better support accurate recognition of those entities.

## Method

Participants 59 QUB students ( 25 males) were tested in a quiet area of the university library, and paid $£ 2$ each to take part in the study. The mean age was 26.5 years.

Materials Materials were the same as used in Experiment 1, except that there were seven diversity and seven sample size trials. We did not include control trials in this experiment. The recognition memory task consisted of 63 images: 28 pictures previously seen in the reasoning tasks (2 from each trial, one of which was featured twice in the trial), 28 previously unseen pictures of members of the previously featured categories, and 7 pictures of categories not featured at any stage in the experiment.

Procedure The procedure for the reasoning part of the experiment was broadly similar to the procedure followed in Experiment 1. However, the secondary task was omitted, trial type was blocked and block order was counterbalanced.

The order in which trials were presented within blocks was randomized.

Once they had completed the reasoning trials, participants were told that the second part of the experiment would consist of a surprise recognition test, and instructed to try to identify which images had been seen previously in the reasoning tasks. Images were presented one at a time and participants pressed the ' 1 ' button for pictures seen before, and the ' 2 ' button for new pictures.

Materials check Our hypothesis is that recognition memory for the contents of diversity trials will be better than for the contents of sample size trials. We carried out a check to ensure that the materials used in the diversity trials were no more memorable than those used in the sample size trials. We presented the materials used in the reasoning part of the experiment to 34 participants. The information about properties was not included and instead of asking participants to make a choice between the samples, we instructed them to memorize the images for a subsequent memory test.

## Results

Reasoning task Participants selected the diverse sample on $73.6 \%$ of trials (SD = 25\%), and the larger sample 81.8\% of the time ( $\mathrm{SD}=26 \%$ ). Participants were significantly more sensitive to sample size than they were to sample diversity, $\mathrm{t}(58)=2.105, \mathrm{p}=.04$.

Recognition memory Performance on the recognition memory test was analyzed with the A' statistic (Snodgrass \& Corwin, 1986), a non-parametric analogue of the d' signal detection measure. An A' of .5 corresponds to chance


Figure 3: A' scores from Experiment 2 broken down by trial type and whether participants reasoned about the materials or studied them for memory.
discrimination between old and new stimuli, while a score approaching 1 indicates perfect discrimination. In Figure 3, A' scores for the main experiment are presented alongside scores from the materials check. While A' scores for diversity and sample size materials were almost identical for
participants in the baseline memory condition ( $\mathrm{A}^{\prime}=.82$, and .81 respectively, SDs $=.09$ and .07$), \mathrm{t}(33)=.177$, amongst participants in the main reasoning condition, recognition was much better for the diversity materials $\left(\mathrm{A}^{\prime}=.78, \mathrm{SD}=\right.$ .13) than the sample size materials ( $\mathrm{A}^{\prime}=.66, \mathrm{SD}=.13$ ), $\mathrm{t}(58)=6.343, \mathrm{p}<.001$.

Inspection times We measured the time between presentation of each reasoning item and participants' responses. The average of this inspection time was 5579 ms (SD $=2055 \mathrm{~ms}$ ) for diversity trials, and 5256 ms ( $\mathrm{SD}=$ 2273 ms ) for sample size trials. This difference was nonsignificant, $\mathrm{t}(58)=1.003$. Thus, the difference due to reasoning phenomenon in the recognition memory data cannot be attributed to differences in how long participants looked at the materials for each trial type.

## Discussion

As we predicted, participants had better recognition memory for the entities they reasoned about in the diversity trials than they did for the entities in the sample size trials. Additionally, the results of our materials check confirmed that the diversity entities were not more memorable than the sample size entities. These results suggest that different processes underlie sensitivity to diversity and sensitivity to sample size. Whereas the former requires feature-based processing of the entities, resulting in a verbatim-type representation which supports accurate recognition memory, the latter is driven by the application of a rule, which leads to a gist representation of the samples and significantly less accurate recognition memory.

## General Discussion

Both experiments reported here show evidence of a disassociation between the processes underlying sensitivity to the sample diversity and size. In Experiment 1, sensitivity to sample size and to diversity differentially impacted upon the secondary task, indicating a dissociation of the underlying mental processes. Similarly, in Experiment 2 materials used in diversity trials were remembered significantly better, suggesting a greater degree of feature-based processing. Taken together, these findings are problematic for single-process accounts of inductive reasoning (e.g. Osherson et al, 1990; Sloman, 1993; Rogers \& McClelland, 2004).

Recent findings have challenged the classical view that inductive inference is the product of similarity-based or associative processes, while deduction relies on the application of abstract logical rules (Evans, 2012). On one hand, similarity-driven processes have been shown to underlie some deductive phenomena (Sloman, 1998). On the other, Heit and Rotello (2010; see also Rips, 2001) have shown that both similarity and logical validity determine inductive and deductive argument strength, but with induction drawing more heavily on similarity-based or associative information. With the blurring of the boundaries
between the processes underlying the two forms of reasoning, it has become somewhat unclear what is distinctive about induction. Heit (2007) offers two views on defining induction: the process view, which relates to the processes by which we make an inference, and the problem view, relating to the structure of the inference to be made. While from the problem view deduction and induction remain discrete, our findings suggest that, from the process view, reasoning cannot be so easily partitioned. Our results, from two diverse paradigms, suggest that there is a disassociation between the processes underlying sensitivity to sample size and to sample diversity in category-based induction, and by extension, that inductive reasoning cannot be captured by single process accounts.

## Acknowledgements

Eoin Travers is funded by a postgraduate studentship from the Department of Enterprise and Learning in Northern Ireland. We are grateful to Michelle Leckey for her help with the development of the materials used in these experiments.

## References

Bacon, F. (1878). Novum organum. Clarendon press.
Brainerd, C. J., Reyna, V. F., \& Forrest, T. J. (2003). Are young children susceptible to the false-memory illusion? Child Development, 73, 1363-1377.
Carnap, R. (1950). Logical foundations of probability. Chicago: University of Chicago Press.
De Neys, W. (2006). Automatic-heuristic and executiveanalytic processing in reasoning: Chronometric and dual task considerations. Quarterly Journal of Experimental Psychology, 59, 1070-1100.
Evans, J. S. B. (2012). Questions and challenges for the new psychology of reasoning. Thinking \& Reasoning, 18, 531.

Feeney, A. (2007). How many processes underlie categorybased induction? Effects of conclusion specificity and cognitive ability. Memory \& Cognition, 35, 1830-1839.
Fong, G. T., Krantz, D. H., \& Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. Cognitive Psychology, 18, 253-292.
Hayes, B. K., McKinnon, R., \& Sweller, N. (2008). The development of category-based induction: Reexamining conclusions from the induction then recognition (ITR) paradigm. Developmental Psychology, 44, 1430-1441.
Heit, E. (2007). What is induction and why study it? In A. Feeney \& E. Heit (Eds.), Inductive reasoning (pp. 1-24). Cambridge: Cambridge University Press.
Heit, E. \& Rotello, C. M. (2010). Relations between inductive reasoning and deductive reasoning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 36, 805-812.
Heit, E., Hahn, U. \& Feeney, A. (2005). Defending diversity. Categorization inside and outside of the laboratory: Essays in honor of Douglas L. Medin, (pp. 87-99). Washington: APA.

Inhelder, B., \& Piaget, J. (1975). The origin of the idea of chance in children. New York: NW Norton and Company.
Kahneman, D., \& Tversky, A. (1972). Subjective probability: A judgment of representativeness. Cognitive Psychology, 3, 430-454.
Lo, Y., Sides, A., Rozelle, J., \& Osherson, D. (2002). Evidential diversity and premise probability in young children's inductive judgment. Cognitive Science, 26, 181-206.
Medin, D. L., Coley, J. D., Storms, G. \& Hayes, B. L. (2003). A relevance theory of induction. Psychonomic Bulletin \& Review, 10, 517-532.
Nisbett, R. E., Krantz, D. H., Jepson, C., \& Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. Psychological Review, 90, 339-363.
Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A. \& Shafir, E. (1990). Category-based induction. Psychological Review, 97, 185-200.
Popper, K. R. (1963). Conjectures and refutations: The growth of scientific knowledge. London: Routledge.
Rhodes, M., Brickman, D. \& Gelman, S. A. (2008). Sample diversity and premise typicality in inductive reasoning: Evidence for developmental change. Cognition, 108, 543556.

Rips, L. J. (2001). Two kinds of reasoning. Psychological Science, 12, 129-134.
Rogers, T. T., \& McClelland, J. L. (2004). Semantic cognition: A parallel distributed processing approach. Cambridge, MA: MIT press.
Sides, A., Osherson, D., Bonini, N., \& Viale, R. (2002). On the reality of the conjunction fallacy. Memory \& Cognition, 30, 191-198.
Sloman, S. A. 1993. Feature-based induction. Cognitive Psychology, 25, 231-280.
Sloman, S.A. (1998). Categorical inference is not a tree: The myth of inheritance hierarchies. Cognitive Psychology, 35, 1-33.
Sloutsky, V. M., \& Fisher, A. V. (2004). When development and learning decrease memory: Evidence against category-based induction in children. Psychological Science, 15, 553-558.
Snodgrass, J. G. \& Corwin, J. (1988). Pragmatics of measuring recognition memory: applications to dementia and amnesia. Journal of Experimental Psychology: General, 117, 34-50..
Stanovich, K. E., \& West, R. F. (1999). Discrepancies between normative and descriptive models of decision making and the understanding/acceptance principle. Cognitive Psychology, 38, 349-385
Wilburn, C. \& Feeney, A. (2008). Do development and learning really decrease memory? On similarity and category-based induction in adults and children. Cognition, 106, 1451-1464.

# Automatic and Strategic Search During Analogical Retrieval 

Máximo Trench (MAXTRENCH@Gmail.Com)<br>Universidad Nacional del Comahue \& Universidad Abierta Interamericana<br>Quintral 1250, Bariloche, Rio Negro, Argentina<br>Valeria Olguín (VALITAO@Yahoo.Es)<br>Department of Psychology, Universidad Nacional del Comahue<br>Irigoyen 2000, Cipolletti, Río Negro, Argentina

Adrián Margni (ADRIANMARGNI@Gmail.Com)
Department of Psychology, Universidad Nacional del Comahue Irigoyen 2000, Cipolletti, Río Negro, Argentina

Ricardo A. Minervino (MINERVINOR@Jetband.Com.Ar)<br>Department of Psychology, Universidad Nacional del Comahue Irigoyen 2000, Cipolletti, Río Negro, Argentina


#### Abstract

The present study investigates two key aspects of analogical retrieval: (1) whether other activities different from problem solving automatically elicit a search for analogical sources, and (2) whether strategic search can overcome the superficial bias observed in classical experiments. In Experiment 1, participants had to generate persuasive arguments for a target situation under three experimental conditions: without indication to use analogies, with instruction to use analogies, and with indication to search for sources within four predefined domains: health, human relations, housekeeping, and breeding of animals and plants. Responses from the first condition showed that argumentation rarely triggers spontaneous analogical retrievals, a result that is at odds with most studies on problem solving. Results from the remaining conditions demonstrated that the superficial bias can be strategically reversed when participants are suggested to focus on specific domains. Experiment 2 replicated this last result with the simple instruction to search within domains different from that of the target (i.e., without being provided with a list of specific domains). The theoretical and educational implications of these findings are discussed.


Keywords: Analogy, retrieval, similarity, transfer.

## Introduction

Analogical reasoning consists in acknowledging that the objects of two situations are organized by an identical system of relations (Gentner, 1983). Across activities as diverse as problem solving, instruction or argumentation, finding the right analogical correspondences allows transferring knowledge from a known situation (the base analog: BA) to novel situation (the target analog: TA) in order to improve the representation of the latter. A traditional taxonomy distinguishes between intradomain analogies (i.e., when BA and TA pertain to the same thematic domain) and interdomain analogies (i.e., when BA and TA belong to thematically separate domains). In intradomain analogies, the compared analogs maintain superficial similarity, as corresponding objects and relations tend to be semantically similar.

A number of empirical studies have demonstrated that people can easily understand analogies even in the absence of superficial similarity (e.g., Gentner, Rattermann, \& Forbus, 1993, see Holyoak, Novick \& Melz, 1994, for a review). In contrast with the relative easiness of finding the right mapping between a BA and a TA that are simultaneously active in working memory (WM), the process of retrieving interdomain BAs from Long Term Memory (LTM) turns out to be rather taxing. As in most studies in the memory literature, the standard paradigm for investigating the conditions that foster analogical retrieval comprises two different phases. During the learning phase, participants receive the BAs embedded in tasks aimed at enforcing a proper encoding of the BAs in LTM. During the retrieval phase, sometimes temporally and/or contextually separated from the first, participants receive the TAs embedded in target tasks for which retrieving the BAs becomes crucial, and experimenters assess whether the processing of the TA triggers the retrieval of the critical BA. Studies using this paradigm showed that intradomain BAs are retrieved between two and four times more frequently than interdomain BAs (Holyoak \& Koh, 1987; Keane, 1987). These findings led researchers to conclude that superficial similarity represents a crucial precondition for analogical retrieval. On the other hand, computational modelers of analogical retrieval agree that the computational cost implicated in carrying out a structural mapping between a TA and every potential BA in LTM would be psychologically implausible (Forbus, Gentner \& Law, 1994; Thagard, Holyoak, Nelson \& Gochfeld, 1990).

Under these considerations, proponents of the structure mapping theory (Gentner, 1983) developed MAC/FAC (Forbus, Gentner \& Law, 1994), an algorithm designed to mimic human patterns of analogical retrieval through psychologically realistic computations. MAC/FAC, for Many Are Called, Few Are Chosen, divides retrieval into two phases: MAC, a fast superficial filter, and FAC, a structural matcher.

The MAC phase begins by generating content vectors for the TA and every representation in LTM, with each content vector being generated by assigning a position in an ordered series to all concepts in LTM, and counting how many times each concept appears in each BA. Upon taking the vector products between the content vector of the TA and the vector of all situations in LTM, the MAC stage submits the winning BAs (most of them superficially similar to the TA) to the FAC stage. For each BA, FAC starts by creating all possible local mappings between elements of the same formal type, with the added restriction that mapped relations must have identical meaning. The program then incrementally coalesces local matches into global mappings that satisfy the constraints of parallel connectivity (if two predicates are mapped, their arguments must also be mapped) and one-toone mapping (elements in one analog must map to only one element in the other analog). Finally, FAC scores the quality of global mappings as a function of their size, their depth, and the semantic similarity of their corresponding objects. This last criterion amplifies MACs' bias towards BAs bearing superficial similarity with the TA.

LISA (Learning and Inference with Schemas and Analogies; Hummel \& Holyoak, 1997) is the latest matcher developed by proponents of the multiconstraint theory of analogy (Holyoak \& Thagard, 1989, 1995). Its architecture aims at encompassing retrieval, mapping, inference and schema abstraction by a unified set of core processes that are more neurally plausible than in earlier attempts (e.g. ARCS; Thagard, Holyoak, Nelson \& Gochfeld, 1990). LISA's architecture is a system for representing dynamic role-filler bindings in WM and encoding them in LTM for later retrieval. When a proposition unit (P) like John loves Mary gets activated, it propagates top-down activation to subproposition units (SPs) that represent bindings between each of the case roles of the proposition and its corresponding filler. During the lapse while each SP unit remains active, it transfers top-down activation to two independent structure units representing a case role and its filler (e.g., John and lover) which fire in synchrony with each other and out of synchrony with the units of the complementary SP (i.e. Mary and beloved). Case roles and their fillers-which represent the lower level in the structural hierarchy-in turn activate a collection of semantic units representing their meaning. Therefore, when a proposition such as John loves Mary is selected, the semantic primitives of lover (e.g., emotion1, positivel, and strongl) fire in synchrony with the semantic primitives of John (e.g., human, male and adult), while units representing the beloved role (e.g., emotion2, positive 2 and strong2) fire in synchrony with units representing Mary (e.g., human, female and adult). When the semantic primitives of a given role-filler binding in the TA fire in WM, predicate, object and SP units from one or various BAs compete in responding to this array as a function of the extent to which their semantic units overlap. As in MAC/FAC, LISA's reliance on semantic similarities between BAs and TAs leads to a majority of superficial remindings.

In contrast with the emphasis placed in justifying the appropriateness of the representational and computational assumptions incorporated in each of the above models (e.g MAC/FAC uses serial operations on symbolic representations while LISA uses connectionist computation on distributed representations), the proponents of these models are ambiguous as to whether the models account for spontaneous remindings, voluntary remindings, or both. Given the importance of this distinction within current memory research (see Mace, 2010, for a review), the first objective of the present study is thus to investigate to what extent the search for BAs in LTM is automatically triggered by the processing of the TAs. A second objective of the present study concerns whether voluntary retrieval of BAs is invariably biased towards superficial matches, as in current implementations of the above models, or if search for BAs can be strategically circumscribed to areas of knowledge different from that of the target-a central preoccupation of psychologists and educators (see, e.g., Loewenstein, 2010). Before presenting our study, the available evidence bearing on these two questions is briefly reviewed.

## Automatic vs. Voluntary Search for Base Analogs

It is a common experience to be spontaneously reminded of analogous cases while carrying out thoughtful activities like science teaching, explanation, and persuasive argumentation. However, a sensible question to be asked concerns to what extent being engaged in the above activities automatically initiates a search for BAs in LTM. Even though no single study has yet manipulated whether or not participants are explicitly invited to "think of analogous problems", acrossstudies comparisons within the problem-solving literature suggest that participants' attempts to find a solution automatically elicit a search for BAs. For instance, using roughly comparable stimuli, Keane (1987) and Holyoak and Koh (1987) assessed the retrieval of a BA during a temporally and contextually separated problem solving activity. Even though the former study (but not the latter) explicitly asked participants to look for analogous problems prior to attempting a solution, both obtained comparable rates of retrieval, which suggests that the mere disposition to find solutions suffices to trigger a search for BAs. Other studies of spontaneous analogical retrieval during problem solving (e.g., Chen, Mo \& Honomichl, 2004) point in the same direction. With these antecedents in mind, the specific question that concerns us here is whether other thoughtful activities such as those listed above also trigger a search for analogous cases in a reliable manner.

A likely candidate task for automatic analogical retrieval is persuasive argumentation. A series of naturalistic studies (e.g., Blanchette \& Dunbar, 2000; Trench, Oberholzer \& Minervino, 2009; Trench, Olguín \& Minervino, 2011) have shown that when being asked to generate analogies to convince somebody of performing an action, people easily retrieve BAs from their autobiographical memory. As in these studies, the procedure followed by one of the groups of the first experiment reported in the present study consisted
in presenting participants with a target situation admitting two alternative lines of action, and asking them to provide as many analogies as they could in favor of one of such actions. In order to shed light on whether the activity of finding persuasive arguments automatically triggers a search for relevant BAs, we had another group receive the same TA and the same instructions to argue in favor of the intended action, but without any hint to base their arguments on analogous situations.

## Voluntary Analogical Search: Fixed or Strategic?

As stated above, a wealth of laboratory studies demonstrate that search for BAs yields mostly superficial matches to the TA. Even though most retrieval algorithms were specially engineered to mimic such pattern of results, some of them left open the question of whether such superficial bias could be "tuned" by the analogizer, be it by means of adjusting the weight given to object attributes by the structural component of the system (FAC), or by having the whole retrieval algorithm run on a subset of MLP selected via other general mechanisms of memory such as spreading activation or indexing (Gentner \& Forbus, 1991, p. 4).

Consistent with this last possibility, Ripoll (1998) postulated the psychological reality of a synthetic level of representation that specifies the thematic domain to which a given problem/story belongs, and demonstrated how these "domain tags" operate during the time-course of analogical retrieval. The procedure consisted in coupling superficially similar and superficially dissimilar target problems with a heading intended to activate a domain tag (e.g., "a learning problem"), which could match (or not match, depending on the condition) the domain tag of the base problem. Using concurrent measures of retrieval, Ripoll (1998) found that the presence of shared surface features facilitated retrieval, but only when the domain tags of the problems matched.

In the first experiment of the present study, the second and third groups received a TA and an instruction to search for potential analogous situations that could be used to convince the main character of the TA to pursue a given action. However, while participants of the second group were not given any indication to focus search in any particular direction, participants of the third group were provided with domain tags representing domains thematically distant from the TA, and were asked to search for potential situations within such domains. The comparison between the types of analogies provided by these two groups seeks to extend Ripoll's (1998) findings in two ways. On the one hand, they test the psychological reality of domain tags outside the realm of analogical problem-solving. Most importantly, though, they explore whether these tags can be strategically exploited by the analogizer during voluntary analogical reminding.

## Experiment 1

## Method

Participants and Design One hundred and twenty undergraduate students at Universidad Nacional del Comahue
volunteered to participate in the experiment (Mean age $=$ $21.49, S D=3.42$ ). An even number of participants was randomly assigned to the argumentation condition $\left(\mathrm{G}_{\mathrm{AR}}\right)$, the analogical argumentation condition $\left(\mathrm{G}_{\mathrm{AN}}\right)$, and the analogical argumentation with predefined domains condition $\left(\mathrm{G}_{\mathrm{AN}+\mathrm{D}}\right)$. The variables indication to use analogies (two levels: with and without explicit indication to use analogies) and provision of search domains (two levels: with and without indication to search within particular domains) received between subjects manipulation. The dependent variables were the number and type (intra/interdomain) of the proposed BAs.

Materials and Procedure Before advancing to the target task, participants of all groups received an instructional material on argumentation. The material handed to the $\mathrm{G}_{\mathrm{AR}}$ covered general features of arguments, but did not describe specific types of arguments (e.g., analogies). The material handed to the $G_{\text {AN }}$ and the $G_{\mathrm{AN}+\mathrm{D}}$ described the use of analogies in persuasion, illustrating with two examples the distinctions between intra and inter-domain analogies, as well as between analogies based on situations retrieved from memory and analogies based on invented situations. Once the 10 min allotted to reading the instructional material had elapsed, participants of all groups were given TA describing the situation of a family that was accumulating an important debt in the balance of their credit card. All groups had to generate as many arguments as they could to persuade them to cut expenses immediately in order to cancel the debt, on the grounds that otherwise it would grow so big that future cuts would need to be even more dramatic. Whereas instructions given to the $\mathrm{G}_{\mathrm{AR}}$ did not mention the convenience of including analogies to prior cases among the persuasive arguments, participants of the $G_{A N}$ and $G_{A N+D}$ were asked to base their arguments on analogies to known situations. The difference between $G_{\text {AN }}$ and $G_{\text {AN }+\mathrm{D}}$ was that while participants of $\mathrm{G}_{\mathrm{AN}}$ received no instructions concerning the domains of the BAs, participants of the $\mathrm{G}_{\mathrm{AN}+\mathrm{D}}$ were asked to sequentially focus their search within four domains different from economy: health, human relations, housekeeping, and breeding of animals and plants. In order to prevent participants of the $G_{\text {AN+D }}$ and $G_{A N+D}$ from reporting BAs not originated in retrieval processes they were encouraged to base their analogies on past episodes which had happened to them or to others, or that were learned from verifiable sources such as newspapers, books, movies, etc. Participants of $G_{A R}$ and $G_{A N}$ were given 20 min to complete the argumentation task. In the case of $\mathrm{G}_{\mathrm{AN}+\mathrm{D}}$, participants were allotted 5 min for each of the suggested domains. Once this time had elapsed, participants of all groups were allotted 5 more minutes to report all other arguments (or analogies, depending on the group) that had come to mind during the previous phase but were not reported for whatever reasons. This question was intended to neutralize an eventual conscious editing of retrieved BAs, (cf. Trench, Olguín \& Minervino, 2011), like when a BA is rejected for not being persuasive, or for not pertaining to the specific domain that was requested (e.g., in $\mathrm{G}_{\mathrm{AN}+\mathrm{D}}$ ).

Data analysis Two judges received instruction on the concept of analogy, as well as on the general distinction between intra and interdomain analogies. For the particular TA at stake, they were instructed to regard as "analogical responses" all proposals including the following elements: (1) a problem of increasing magnitude, (2) a delay in the attempts to solve it, and (3) a consequent increase in the cost of solving it. Regarding the intra/interdomain distinction, judges were instructed to score as intradomain all situations where the problem of increasing magnitude was economic (e.g., a public debt) and to score as interdomain all instances in which the problem of increasing magnitude was not of economic nature (e.g., an illness or addiction). Given that we sought to detect all the BAs that were retrieved from LTM in response to the target task, judges were handed all responses produced by the participants, regardless of whether they were reported during the argumentation phase, or during later requirement to list all other situations that had come to mind during the first task, but were not included among the final proposals. Judges agreed in $82 \%$ of the cases regarding the analogical status of proposals, and in $94 \%$ of the cases regarding their intra/interdomain nature. Cases of disagreement were resolved by open discussion.

## Results and Discussion

Across conditions, participants proposed a mean of 2.10 responses ( $S D=.94$ ), out of which $44 \%$ were rendered "analogical" by the judges. Further comparisons and statistical analyses were restricted to analogical proposals. Our first empirical question concerned whether the task of generating persuasive arguments would reliably elicit a search for BAs in LTM, as observed within the literature on analogical problem solving. Taking together intra and interdomain proposals, participants of $\mathrm{G}_{\mathrm{AR}}$ retrieved a total of 7 BAs in response to the TA $(M=.18, S D=.45)$. This level of analogical retrieval is markedly lower than that of $\mathrm{G}_{\mathrm{AN}}$ ( $M=.73, S D=.60$ ), where participants were explicitly asked to base their arguments on analogies to known situations, $t(72.14)=4.658, p<.01$. Given the performance exhibited by the $G_{A N}$, the disappointing number of BAs retrieved by participants of $G_{\text {AR }}$ cannot be attributed to a lack of BAs available in LTM for retrieval. Rather, it indicates that that the pragmatic of generating persuasive arguments for a realworld target situation does not reliably elicit a spontaneous search for relevant analogs in LTM. A likely explanation for the difference between our results and those obtained with problem solving tasks might lie in the fact that the types of problems typically used (e.g. the tumor problem) do not admit direct methods of solution (e.g., means-ends analysis). Perhaps with other types of problems, the spontaneous use of analogies would be less frequent, as it happened in the present study.

Our second empirical question dealt with whether the search mechanisms underlying voluntary analogical retrieval are invariably set to favor superficially similar BAs. Judges' analysis of analogical proposals reported by the $\mathrm{G}_{\mathrm{AN}}$ showed that $62 \%$ of the retrieved BAs were semantically similar to
the TA, and $38 \%$ of retrieved TAs were semantically dissimilar from the TA, a result that reproduces the pattern typically obtained in the literature. In contrast with this standard pattern of retrievals, judges' analyses of analogies generated by the $G_{\text {AN }+\mathrm{D}}$ showed that whereas $35 \%$ of the retrieved BAs came from the same domain of the TA, $65 \%$ of the retrieved BAs were interdomain, a result that goes against the superficial bias typically obtained in the literature on analogical retrieval. A comparison between the $\mathrm{G}_{\mathrm{AN}}$ and the $\mathrm{G}_{\mathrm{AN}+\mathrm{D}}$ in terms of their relative proportions of superficially similar vs. superficially dissimilar retrievals thus demonstrates that the participants can strategically favor the retrieval of one or the other type of BAs, $Z=-2.54, p<.05$.


Figure 1. Mean number or retrievals, Experiment 1
An intriguing question raised by the possibility of shifting search away from the target domain concerns whether the increased number of distant matches comes at the expense of missing a number of intradomain BAs that would be retrieved under a non strategically-oriented search, as a "shift of focus" metaphor might suggest. A comparison between the $G_{\mathrm{AN}}$ and the $\mathrm{G}_{\mathrm{AN}+\mathrm{D}}$ in terms of the mean number of superficially similar and superficially dissimilar BAs showed that whereas the mean number of distant BAs generated by the $\mathrm{G}_{\mathrm{AN}+\mathrm{D}}(M=1.23, S D=1.17)$ clearly surpassed the mean number of distant BAs retrieved by the $\mathrm{G}_{\mathrm{AN}}(M=.28, S D=.45), t(50.48)=4.806, p<.001$, participants of $\mathrm{G}_{\mathrm{AN}+\mathrm{D}}$ did not retrieve a lesser amount of superficially similar BAs than participants of $\mathrm{G}_{\mathrm{AN}}(M=.65$, $S D=.83$, vs. $M=.45, S D=.50$, respectively), $t(64.14)=$ $1.299, p>.05$. Rather than simply shifting the focus towards interdomain retrieval, it seems that participants of $\mathrm{G}_{\mathrm{AN}+\mathrm{D}}$ are broadening the scope of their search, an operation that boosts access to distant analogs while still retaining baseline levels of intradomain analogizing.

Even though the educational implications of the observed increment in interdomain analogizing are easy to foresee, an interesting question to be asked concerns whether a reasonable increase in the absolute and relative amounts of interdomain retrievals can still be obtained without providing participants with a set of promising domains within which to search for useful BAs. In Experiment 2, we tested this possibility by comparing the analogical argumentation condition (i.e., the former $\mathrm{G}_{\mathrm{AN}}$ ) against a pro-interdomain argumentation condition without the provision of predefined search areas ( $\mathrm{G}_{\text {ANint }}$ ).

## Experiment 2

## Method

Participants and Design Fourty students at Universidad Nacional del Comahue (Mean age $=20.71, S D=2.05$ ) were randomly assigned to two experimental conditions. The variable type of argumentation (two levels: with indication to use analogies vs. with indication to use interdomain analogies) received between subjects manipulation. The dependent variables were the number and type of the proposed BAs.

Materials and Procedure The materials and procedure applied to the $\mathrm{G}_{\mathrm{AN}}$ were a replication of those followed with the $\mathrm{G}_{\mathrm{AN}}$ of Experiment 1. The materials and procedure employed with the $G_{\text {ANint }}$ were identical to those of the $G_{\text {AN }}$ with the sole difference that participants were asked to base their persuasive analogies on episodes pertaining to domains different from that of the TA (i.e., economy). Data analysis was identical as in Experiment 1, with judges' agreement reaching $85 \%$ with regards to the analogical status of proposals, and $96 \%$ regarding their intra/interdomain nature.

## Results and Discussion

Experiment 2 was carried out to assess whether an increase in the absolute and relative amounts of interdomain retrieval could still be obtained without providing participants with a set of interdomain search areas to look for analogous situations. A comparison between the $G_{A N}$ and the $G_{\text {ANint }}$ showed that the relative proportion of interdomain analogies proposed by the $\mathrm{G}_{\text {ANint }}$ was higher than in the $\mathrm{G}_{\mathrm{AN}}, Z=-$ 1.97, $p<.05$. Whereas the analogies proposed by the $\mathrm{G}_{\mathrm{AN}}$ were $63 \%$ intradomain and $37 \%$ interdomain, the analogies proposed by the $G_{\text {ANint }}$ were $40 \%$ intradomain and $60 \%$ interdomain. Though not as strong as in Experiment 1, this reversal demonstrates that participants can voluntarily alter the superficial bias classically obtained in experiments of analogical retrieval with the mere intention to search for thematically distant sources in LTM.


Figure 2. Mean number or retrievals, Experiment 2
As in Experiment 1, the augmented proportion of interdomain retrievals in the pro-interdomain condition was not obtained at the expense of missing a number of intradomain retrievals. A comparison between $G_{A N}$ and $G_{\text {ANint }}$ in terms of the mean number of intra and interdomain retrievals showed that whereas $\mathrm{G}_{\text {ANint }}(M=.65, S D=.98)$ clearly surpassed the $\mathrm{G}_{\mathrm{AN}}(M=.30, S D=.46, t(55.80)=2.05, p<.05)$ in the
number of interdomain retrievals, both groups retrieved similar amounts of intradomain BAs $(M=.43, S D=.59$ vs. $M=.50, S D=.60$, respectively, $t(78)=0.562, p>.05$ ). Once again, it seems that a strategic search for interdomain BAs can powerfully boost access to distant analogs, while still retaining baseline levels of intradomain retrieval.

## General Discussion

In order to reproduce human patterns of analogical retrieval, extant computational models have specified in great detail a number of assumptions about the types of representations and computations implied in retrieving BAs from LTM. In contrast to this long-lasting preoccupation, their presentations are ambiguous as to whether the postulated mechanisms account for the processes of spontaneous reminding, strategic retrieval, or both. Albeit unsystematic, the evidence related to this matter comes mainly from studies of analogical problemsolving, and suggests that both types of search yield similar results, since the mere disposition to find a solution to a problem reliably elicits a search for analogous BAs in LTM.

The first experiment of the present study tackled two interrelated issues. The first one was concerned with spontaneous analogical retrieval, and had to do with whether other thoughtful activities different from problem solving (in this case, persuasive argumentation) can also elicit spontaneous remindings reliably. Results of Experiment 1 showed that when participants are not explicitly asked to base their arguments on analogies to prior cases, this activity seldom occurs spontaneously. In light of the performance of a second group that was explicitly asked to use analogies, the low level of spontaneous retrieval obtained by the first group cannot be attributed to a lack of available BAs in LTM. Rather, it shows that the pragmatics of generating persuasive arguments does not reliably elicit a search for BAs in LTM. These results have implications for models of analogical retrieval, since they can help specify the conditions under which the proposed mechanisms operate.

Our second concern was related to strategic analogical retrieval, and dealt with whether this second type of process can potentially reverse the superficial bias observed in behavioral studies of analogical retrieval, and simulated by computational models. Results showed that when participants are asked to base their arguments on analogies to known situations, they retrieve more superficial than distant matches. However, when provided with a series of distant domains to focus their search, this proportion reverses-a pattern of results that claims for an extension of Ripoll's (1998) domain tags to the arena of voluntary analogical retrieval. It should be noted, however, that our conclusions were based on the use of a single TA. In future studies, it would be desirable to replicate these results with a wider set of materials.

Albeit never implemented, the developers of MAC/FAC left open the possibility of relaxing its superficial bias by either suspending FAC's computation of object attributes, or by having the system run on a subset of LTM selected via mechanisms of spreading-activation or indexing. Given the strong superficial constraints imposed by the MAC stage, it
seems that only by running on a subset of LTM (e.g., on a subset defined by thematic search areas, or domains) the program might have a chance of coping with the pattern of interdomain analogizing elicited during strategic analogical retrieval. In relation to this possibility, the fact that in Experiments 1 and 2 participants of the pro-interdomain conditions still retrieved a significant amount of intradomain matches suggests that, at least with our materials, strategic search can be somewhat demanding, leading to recurrent cycles of non strategic retrieval attempts.

The present results on strategic analogical retrieval also suggest interesting instructional applications, since educators and researchers have long strived to find ways of facilitating cross-domain transfer. In recent times, attention has shifted from promoting an abstract encoding of BAs (e.g., Catrambone \& Holyoak, 1989) to improving the encoding of TAs at retrieval time, such as providing participants with two structurally identical TAs, and asking them to compare such TAs prior to attempting a solution (Gentner, Loewenstein, Thompson \& Forbus, 2009; Kurtz \& Loewenstein, 2007). As Loewenstein (2010) points out, the appeal of this approach lies in its potential to foster retrieval of BAs which might have been encoded in suboptimal ways. However, a practical limitation of the target-comparison method used in the above studies lies in the fact that participants need to be provided with a second TA for every new TA. Even though participants of our $\mathrm{G}_{\mathrm{AN}+\mathrm{D}}$ were also provided with targetspecific information at retrieval time (a set of promising domains to search for BAs), the $G_{\text {ANint }}$ of our second experiment retrieved mostly interdomain BAs with the sole instruction to search within domains different from the TA, that is, without receiving target-specific information. We believe that the austerity of this last intervention opens up encouraging perspectives for the flexible use of analogy in educational environments.

## Acknowledgments

This work was supported by Grant PICT 1461 from the ANCyT of Argentina, Grant 076 from the National University of Comahue, Grant PIP 0266 from CONICET, and grant PIB from Universidad Abierta Interamericana.

## References

Blanchette, I., \& Dunbar, K. (2000). How analogies are generated: The roles of structural and superficial similarity. Memory \& Cognition, 28, 108-124.
Catrambone, R., \& Holyoak, K. J. (1989) Overcoming contextual limitations on problem-solving transfer. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15, 1147-1156.
Chen, Z., Mo, L., \& Honomichl, R. (2004). Having the memory of an elephant: Long-term retrieval and the use of analogues in problem solving. Journal of Experimental Psychology: General, 133, 415-433.
Forbus, K., Gentner, D., \& Law, K. (1994). MAC/FAC: A model of similarity-based retrieval. Cognitive Science, 19, 141-204.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. Cognitive Science, 7, 155-170.
Gentner, D. (1989). The mechanisms of analogical learning. In S. Vosniadou \& A. Ortony (Eds.), Similarity and analogical reasoning. London: Cambridge University Press.
Gentner, D., \& Forbus, K. D. (1991). MAC/FAC: A model of similarity-based access and mapping. Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society (504-509).
Gentner, D., Loewenstein, J., Thompson, L., \& Forbus, K. (2009). Reviving inert knowledge: Analogical abstraction supports relational retrieval of past events. Cognitive Science, 3, 1343-1382.
Gentner, D., Rattermann, M. J., \& Forbus, K. D. (1993). The roles of similarity in transfer: Separating retrievability from inferential soundness. Cognitive Psychology, 25, 431-467.
Holyoak, K., \& Koh, K. (1987). Surface and structural similarity in analogical transfer. Memory \& Cognition, 15, 332-340.
Holyoak, K. J., Novick, L. R., \& Melz, E. R. (1994). Component processes in analogical transfer: Mapping, pattern completion, and adaptation. In K. J. Holyoak \& J. A. Barnden (Eds.), Advances in connectionist and neural computation theory, Vol. 2, Norwood, N.J.: Ablex
Holyoak, K., \& Thagard, P. (1989). Analogical Mapping by Constraint Satisfaction. Cognitive Science, 13, 295-355
Holyoak, K., \& Thagard, P. (1995). Mental leaps: Analogy in creative thought. Cambridge, MA: MIT Press.
Hummel, J. E., \& Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. Psychological Review, 104, 427-466.
Keane, M.T. (1987). On retrieving analogues when solving problems. Quarterly Journal of Experimental Psychology, 39, 29-41.
Kurtz, K., \& Loewenstein, J. (2007). Converging on a new role for analogy in problem solving and retrieval: When two problems are better than one. Memory \& Cognition, 35, 334-341
Loewenstein, J. (2010). How one's hook is baited matters for catching an analogy. In B. Ross (Ed.), Psychology of Learning and Motivation, Volume 53. Elsevier.
Ripoll, T. (1998). Why This Makes Me Think of That. Thinking and Reasoning, 4,15-43.
Thagard, P., Holyoak, K., Nelson, G., \& Gochfeld, D. (1990). Analog retrieval by constraint satisfaction. Artificial Intelligence, 46, 259-310.
Trench, M., Oberholzer, N., \& Minervino, R. (2009). Dissolving the analogical paradox: Retrieval under a production paradigm is highly constrained by superficial similarity. In B. Kokinov, K. Holyoak y D. Gentner (Eds.), New frontiers in analogy research. Sofia: NBU Press.
Trench, M., Olguín, V., \& Minervino, R. (2011). Mammoth cloning reminds us of "Jurassic Park" but storm replication does not: Naturalistic settings do not aid the retrieval of distant analogs. Proceedings of the $33^{\text {rd }}$ Annual Conference of the Cognitive Science Society (pp. 2649-2654). Austin, TX: Cognitive Science Society

# A Dynamic Dual-Process Model of Decision-making Under Uncertainty 

Jennifer S. Trueblood (jstruebl@uci.edu)<br>Department of Cognitive Sciences, University of California, Irvine<br>Irvine, CA 92697 USA


#### Abstract

Current dynamic models of decision-making assume that a unitary system is responsible for forming preferences. However, extensive research has shown that decision-making and behavior result from the interaction of two separate systems of reasoning - one that is fast, automatic, and experiential and one that is slow, deliberative and rational. This paper develops the first dynamic dual-process model of decision-making that can account for choice, response times, and prices. The model is applied to several phenomena from the risky decisionmaking literature including enhancements in preference by small losses, preference reversals due to response mode, and the influence of price and affect on preference.


Keywords: Decision-making, dual-process theory, preference reversal, dynamic models

## Introduction

Existing dynamic models of decision-making such as Decision Field Theory (Busemeyer \& Townsend, 1993) assume there is a single system of thought that produces preferences. However, there is substantial research supporting the idea that preferences are formed from a dual process of reasoning. This paper introduces a dynamic dual-process model of risky decision-making. The model generalizes a previous static two systems model developed by Mukherjee (2010) to account for choice, response times, and prices.

Dual-process theory postulates that there are two fundamentally different systems that can process information. One system is described as automatic, intuitive, fast, and experiential. The other is labeled as deliberative, analytical, slow, and rational. The former system is typically referred to as System 1 and the latter System 2 (Stanovich \& West, 2000). (In this paper, System 1 is labeled the affective system and System 2 the deliberative system following the terminology of Mukherjee (2010).) Research has shown that dual-process accounts are often more successful at explaining behavioral phenomena than unitary approaches (Hogarth, 2001; Kahneman, 2003; Sanfey, Loewenstein, McClure, Cohen, et al., 2006). There is also evidence from the neuroscience community for two separable systems in the brain that contribute to decision-making (Damasio, 1994; Sanfey et al., 2006).

The paper begins by describing the Dynamic Dual-Process (DDP) model for choice and response times which is later extended to also account for pricing elicitation methods. The model is used to make new predictions about the relationship between response time and the involvement of the affective system. It is shown that in some situations greater involvement of the affective system leads to faster decisions as expected. However, the model also predicts that in other situations the two systems can compete with one another resulting in longer response times. DDP is also applied to three
phenomena from the risky decision-making literature: the enhancement effect which occurs when a small loss is added to a positive gamble, preference reversals due to response mode, and the influence of price and affect on preference.

## A Static Two Systems Model

The DDP model generalizes the Dual System Model (DSM) by Mukherjee (2010). DSM is a utility model of risky decision-making where the overall utility for a gamble is composed of two components: the utility of the gamble with respect to the deliberative system and the utility of the gamble with respect to the affective system. Mathematically, the overall utility can be written as $V(G)=V_{A}(G) \oplus V_{D}(G)$ where $V_{A}(G)$ is the evaluation due to the affective system and $V_{D}(G)$ is the evaluation due to the deliberative system.

Based on experimental results by Hsee and Rottenstreich (2004), Mukherjee made two assumptions about the evaluation of outcomes. The first assumption was that the deliberative system evaluates outcomes linearly so that $V_{D}(x)=k x$ where $x$ is an outcome and $k$ is a free parameter. The second assumption was that the affective system evaluates outcomes with respect to a concave value function in the gain domain and a convex value function in the loss domain similar to the value function in prospect theory (Kahneman \& Tversky, 1979). Mukherjee postulates that for positive outcomes this value function can be approximated by $V_{A}(x)=x^{m}$ for $m<1$.

Mukherjee made two additional assumptions about the perception of probabilities based on experimental work by Rottenstreich and Hsee (2001). He assumed that the deliberative system perceives probabilities directly without distortion so that the probability weighting function is $w(p)=p$ as in Expected Utility theory. He also assumed that the affective system is insensitive to probabilities and only recognizes whether or not an outcome is possible. Thus, each possible outcome receives equal weight so that for $n$ outcomes the probability of any single outcome is $1 / n$.

Let $\left(p_{1}, x_{1} ; \ldots ; p_{n}, x_{n}\right)$ be the gamble $G$ with $n$ positive outcomes $x_{i}$. Using the four assumptions above, Mukherjee defined the utility for the deliberative system as $V_{D}(G)=$ $\sum_{i}^{n} p_{i}\left(k x_{i}\right)$ and the utility for the affective system as $V_{A}(G)=$ $1 / n \sum_{i}^{n} x_{i}^{m}$. The overall utility $V(G)$ is simply the convex combination of the two utilities for the two different systems:

$$
\begin{align*}
V(G) & =\gamma V_{A}(G)+(1-\gamma) V_{D}(G) \\
& =\gamma \frac{1}{n} \sum_{i}^{n} x_{i}^{m}+(1-\gamma) k \sum_{i}^{n} p_{i}\left(x_{i}\right) \tag{1}
\end{align*}
$$

where $\gamma$ is the weight given to the affective system. DSM can account for a wide range of choice phenomena including
violations of nontransparent stochastic dominance, ambiguity aversion, common consequence effect, and the common ratio effect (to name a few).

## The Dynamic Dual-Process Model

While DSM has been very successful in accounting for a variety of phenomena, it does not describe the dynamic process underlying decisions. Like most models of judgment and decision-making, DSM is a descriptive model concerned with theorizing at the highest level. Because DSM is a static utility model, it cannot make predictions about response times. On the other hand, DDP is a processing model aimed at explaining the mechanisms that produce behavior. In this way, it is similar to other dynamic models of decision-making such as Decision Field Theory (Busemeyer \& Townsend, 1993). DDP is the first two systems model that can account for both choice and response times.

Another drawback to DSM is that it has not been applied to experiments using pricing elicitation methods such as asking how much money one is willing to pay (WTP) for an option or how much money one is willing to accept (WTA) to sell an option. Many of the experiments in affective decisionmaking use these pricing procedures including the majority of the experiments by Rottenstreich and Hsee (2001) and Hsee and Rottenstreich (2004) that served as motivation for DSM. The DDP model can easily be extended to account for pricing elicitation methods as described in a later section.

The DDP model is formulated with respect to the typical risky decision task of choosing between two gambles. Let $G_{1}=\left(p_{1}, x_{1} ; \ldots ; p_{n}, x_{n}\right)$ and $G_{2}=\left(q_{1}, y_{1} ; \ldots ; q_{m}, y_{m}\right)$ be gambles with outcomes $x_{i}$ and $y_{i}$ and probabilities $p_{i}$ and $q_{i}$ respectively. As a decision-maker considers the two gambles, his or her preference evolves across time. Let $P(t)$ be the individual's preference at time $t$ where positive preference states represent momentary preference for gamble $G_{1}$ and negative preference states represent momentary preference for gamble $G_{2}$. A new preference state $P(t+1)$ is formed at each moment in time from the previous preference state according the linear stochastic difference equation:

$$
\begin{equation*}
P(t+1)=P(t)+d+\varepsilon(t) \tag{2}
\end{equation*}
$$

where $\varepsilon(t)$ is the stochastic error term and $d$ is the difference in the evaluations of the gambles. The evaluation of each gamble is determined by evaluations from the affective and deliberative systems as in DSM. The difference $d$ is given by

$$
\begin{equation*}
d=V\left(G_{1}\right)-V\left(G_{2}\right) \tag{3}
\end{equation*}
$$

where $V\left(G_{1}\right)$ and $V\left(G_{2}\right)$ are calculated as in equation 1.
The preference state starts at an initial state $P(0)=z$ reflecting an initial bias for one gamble over the other. Specifically, if $z>0$, then there is an initial bias for $G_{1}$ and if $z<0$, then there is an initial bias for $G_{2}$. The preference state evolves until it reaches a threshold. There are two thresholds for the model, a positive threshold $\theta$ associated with $G_{1}$ and a negative threshold $-\theta$ associated with $G_{2}$. When the
preference state reaches the positive threshold, $G_{1}$ is selected. When it reaches the negative threshold, $G_{2}$ is selected.

In total, DDP has six parameters. Three parameters, $k, m$, and $\gamma$, are used in the evaluation of the gambles given in equation 1. Parameters $\theta$ and $z$ define the threshold and initial bias respectively. There is an additional variance parameter $s$ used to define the amount of noise in the accumulation process.

DDP is a Wiener diffusion process (the continuous-time, continuous-state version of the random walk). Link and Heath (1975) derived equations for choice probabilities and the conditional mean response time for the Wiener process. Thus, DDP is computationally tractable and easy to apply.

## Response Time and Affect

The affective system is typically characterized as automatic and fast as compared to the deliberative system which is analytic and slow. As such, when the affective system plays a larger role in the decision-making process, decisions should be quick. In terms of the DDP model, this implies that as $\gamma$ increases, response times should decrease.

To test the relationship between response time and the involvement of the affective system, two gambles of equal expected value were analyzed: $G_{1}=(4 / 10, \$ 9 ; 6 / 10, \$ 0)$ and $G_{2}=(9 / 10, \$ 4 ; 1 / 10, \$ 0)$. Choice and response times from DDP were examined for all possible values of $\gamma$. The $m$ parameter used to specify the curvature of the affective valuation function was also allowed to vary from $0.1 \leq m \leq 0.9$. The scaling parameter $k$ in the deliberative value function was fixed to $k=1$ so that $V_{D}(x)=x$ without distortion. The initial bias was fixed to $z=0$ reflecting no bias towards one gamble over the other. The threshold parameter was fixed to $\theta=10$, and the variance parameter was fixed to $s=1$ as is common in response time modeling.

The top left panel of Figure 1 shows the choice probability for gamble $G_{1}$ over $G_{2}$ for different values of $\gamma$ and $m$. As $\gamma$ increases, the probability of selecting $G_{1}$ increases. This reflects the assumption that the affective system ignores probabilities when evaluating gambles. As the affective system becomes more involved, the gamble with the highest payoff is viewed more favorably. As $m$ decreases and the value function for the affective system becomes more concave, the two gambles are viewed as indifferent. Thus, favorability for $G_{1}$ by the affective system is moderated by increasing risk aversion (i.e., decreasing $m$ ). The top right panel of Figure 1 shows the mean response time conditional on selecting $G_{1}$. As predicted, increases in $\gamma$ lead to faster response times. In general, for gambles of equal expected value, as the affective system becomes more involved, decisions become quick and high payoff options are preferred.

When gambles have unequal expected value, it is not necessarily the case that response times decrease with increased involvement of the affective system. The bottom panels of Figure 1 show choice probabilities and response times for the gamble $G_{1}^{*}=(3 / 10, \$ 9 ; 7 / 10, \$ 0)$ as compared to $G_{2}$. In this situation, $G_{2}$ has a greater expected value than $G_{1}^{*}$. When the


Figure 1: Top panel: choice probability (left) and conditional mean response time (right) for gambles with equal expected value for various values of $\gamma$ and $m$. Bottom panels: choice probability (left) and conditional mean response time (right) for gambles with unequal expected value. Red values indicate larger choice probabilities and longer response times.
deliberative system is more involved as indicated by small values of $\gamma, G_{2}$ is preferred. However, as $\gamma$ increases and the affective value function becomes less concave (corresponding to an increase in $m$ ), $G_{1}^{*}$ is preferred. Response times are fast for small values of $\gamma$ paired with small values $m$ and for large values of $\gamma$ paired with large values of $m$. In other words, quick decisions can be made for $G_{2}$ by the deliberative system and for $G_{1}^{*}$ by the affective system, but response times increase when there is conflict between the two systems. The influence of system conflict on response time is a new prediction by DDP which could be tested in future experiments.

## Enhancement by Small Loss

Slovic, Finucane, Peters, and MacGregor (2002) found that adding a small loss to a positive gamble can increase its attractiveness rating and choice probability. When asked to rate the attractiveness of gamble $G_{1}=(7 / 36, \$ 9 ; 29 / 36, \$ 0)$ and $G_{2}=(7 / 36, \$ 9 ; 29 / 36,-\$ 0.05)$, participants rated gamble $G_{1}$ with no loss lower (mean $=9.4$ on a $0-20$ scale) than gamble $G_{2}($ mean $=14.9)$. The gambles were then each paired
with a sure gain of $\$ 2$ (denoted by $S$ ). Half of the participants were asked to choose between $G_{1}$ and $S$ and the other half were asked to choose between $G_{2}$ and $S$. Only $33.3 \%$ chose $G_{1}$ over the gain whereas $60.8 \%$ chose $G_{2}$ over the gain. Slovic et al. (2002) explained these findings by the affect heuristic. The inclusion of a small loss enhances the perceived benefit of $\$ 9$ producing a positive affective feeling for $G_{1}$ leading to higher attractiveness ratings and choice probabilities.

The enhancement in choice probability by the inclusion of a small loss can also be explained by DDP. Based on the idea that $G_{2}$ produces a more affective response than $G_{1}$, it is assumed that the affective system is more activated by $G_{2}$ than $G_{1}$. Mathematically, this implies that $\gamma_{2}>\gamma_{1}$ where $\gamma_{1}$ is associated with the choice between $G_{1}$ and $S$ and $\gamma_{2}$ is associated with the choice between $G_{2}$ and $S$.

To test the enhancement effect, the difference in probabilities $\operatorname{Pr}\left(G_{2} \mid\left\{G_{2}, S\right\}\right)-\operatorname{Pr}\left(G_{1} \mid\left\{G_{1}, S\right\}\right)$ was examined for different values of $\gamma_{1}, \gamma_{2}$ and $m$. For this analysis, $\gamma_{1}$ was allowed to vary from 0.1 to 0.5 and $\gamma_{2}$ was defined in terms of $\gamma_{1}$ by
the equation $\gamma_{2}=\gamma_{1}+\alpha$ where $\alpha$ varied from 0.1 to 0.3 in increments of 0.05 . As in the previous demonstration, $k=1$, $\theta=10, z=0$, and $s=1$. Because the affective value function postulated by Mukherjee (2010) only applies to positive outcomes, the function was generalized to $V_{A}(x)=-|x|^{m}$ for negative outcomes. Figure 2 plots the difference in probabilities for different values of $\gamma_{1}$ given along the $x$-axis. The different curves in the figure are associated with different values of $\gamma_{2}$ The $m$ parameter was fixed to 0.3 in the top panel and 0.5 in the bottom panel.


Figure 2: Enhancement effect shown as the difference in choice probabilities for $G_{2}$ and $G_{1}$ for different values of $\gamma_{1}$. The top panel shows the effect when $m=0.3$ and the bottom panel shows the effect when $m=0.5$. An enhancement effect occurs when the difference is greater than zero.

In the figure, the enhancement effect occurs when the difference in probabilities is greater than zero. When $m=0.5$, this happens for all values of $\gamma_{2}$. When $m=0.3$ implying greater risk aversion, the enhancement effect only occurs for
large values of $\gamma_{2}$ (when $\alpha>0.25$ ) suggesting greater involvement in the affective system is needed in order to produce the effect. In sum, the figure shows that the DDP model can easily account for the enhancement in choice probability by small losses. Further, the model makes new predictions about the magnitude of enhancement with respect to risk aversion and affect.

## Extending DDP for Pricing Elicitation Methods

In many decision tasks, participants are asked to report a value such as a price that they are willing to assign to a particular option. For example, participants might be asked how much they are willing to pay (WTP) to play a certain gamble. DDP can be extended to account for such elicitation methods. The approach taken is similar to the one developed by Busemeyer and Goldstein (1992) and Johnson and Busemeyer (2005). The basic idea is that when an individual is determining price equivalence, they search through a range of possible prices. When a particular price is being considered, the individual can decide that it is too low, too high, or equivalent to the gamble. In the case when the price is too low, the individual increases the price. When the price is too high, the individual decreases the prices. If the price is equivalent, then it is reported.

Mathematically, this search process can be formulated as a discrete Markov chain as illustrated in Figure 3. The states in the chain correspond to possible prices increasing from left to right. The range of possible states is determined by the problem. For example, if an individual is asked how much they would pay to play the gamble $(7 / 36, \$ 9 ; 29 / 36, \$ 0)$, the the range of possible prices would be $\$ 0$ to $\$ 9$. The search process is assumed to begin near the middle of the candidate prices. A step to the right in the chain corresponds to increasing the price. The probability $p$ of stepping to the right is the choice probability from DDP of choosing the gamble over a sure gain of $\$ x$ where $\$ x$ is the candidate price. A step to the left in the chain corresponds to decreasing the price. The probability $q$ of stepping to the left is the choice probability from DDP of choosing a sure gain of $\$ x$ over the gamble. The probability of exiting the search process and reporting a price occurs with probability $r$ whenever DDP enters a neutral state. The neutral state is the point of indifference between the gamble and a sure gain and corresponds to $P(t)=0$ in DDP. Details about implementing Markov chain models can be found in a Diederich and Busemeyer (2003) and Johnson and Busemeyer (2005).

## Response Mode Preference Reversals

A puzzling phenomenon in decision-making is the occurrence of preference reversals with changes in response mode (Lichtenstein \& Slovic, 1971, 1973). For example, Slovic et al. (2002) asked subjects to rate the attractiveness of gambles $G_{1}=(29 / 36, \$ 2 ; 7 / 36, \$ 0)$ and $G_{2}=(7 / 36, \$ 9 ; 29 / 36, \$ 0)$ on a $0=20$ scale. On average, participants rated $G_{1}$ (mean 13.2 ) as more attractive than $G_{2}$ (mean 7.5). Yet, when asked how much they would be willing to pay to play the gambles,


Figure 3: Markov chain model for pricing elicitation methods. Transition probabilities $p$ and $q$ are determined by the DDP model.
participants were willing to pay more to play $G_{2}$ (mean \$2.11) than $G_{1}$ (mean \$1.25).

The DDP model explains this preference reversal by assuming that a high attractiveness rating is associated with a strong affective response. Because $G_{1}$ had a higher attractiveness rating than $G_{2}$, it is hypothesized that the affective system is more involved with decisions about $G_{1}$ than $G_{2}$. Mathematically, this implies that the $\gamma$ parameter for $G_{1}$ should be greater than the $\gamma$ parameter for $G_{2}$. To test this hypothesis, a grid search was performed over the $\gamma$ and $m$ parameters to find the ranges of these parameters that produce price equivalences similar to those in the experiment. Specifically, parameter pairs that produced prices within $\$ 0.20$ of the mean prices from the experiment were examined. For this analysis, $k=1, z=0$, and $s=1$ as before. Matrix methods (Diederich \& Busemeyer, 2003) were used to determine the transition probabilities from DDP rather than using analytical solutions. This was done to accommodate the inclusion of the exit probability $r$. Because the matrix methods only provide an approximation to the choice probabilities, the threshold was fixed to $\theta=50$ to improve the estimates. For gamble $G_{1}$, the states of the Markov chain ranged from $\$ 0$ to $\$ 2$ in increments of $\$ 0.10$. Similarly, for gamble $G_{2}$, the states of the Markov chain ranged from $\$ 0$ to $\$ 9$ in increments of $\$ 0.10$. The exit probability was set to $r=0.01$.

Figure 4 plots the $\gamma$ and $m$ parameter pairs that produce prices in the given ranges. The blue region shows the parameters that yield prices between $\$ 1.05$ and $\$ 1.45$ for $G_{1}$ and the red region shows the parameters that yield prices between $\$ 1.91$ and $\$ 2.31$ for $G_{2}$. From the figure, it is clear that the $\gamma$ parameter for $G_{1}$ must be greater than the $\gamma$ parameter for $G_{2}$ to produce prices in these ranges. Thus, the DDP model can explain preference reversals by greater involvement of the affective system for more attractive gambles.

## Influence of Probability and Affect on Price

Rottenstreich and Hsee (2001) found that the amount of money participants were willing to pay to play a gamble depends on both the probability of winning and whether the outcome is affect-rich or affect-poor. In their experiment, participants were asked how much they were willing to pay to play a gamble offering a $\$ 500$ coupon for a European vacation or a $\$ 500$ coupon for tuition at their university. The European vacation coupon was designed to be affect-rich whereas tu-


Figure 4: Results of a grid search over $\gamma$ and $m$ for WTP prices. The blue region shows the parameters that produce prices for $G_{1}$ within $\$ 0.20$ of the mean price of $\$ 1.25$. The red region shows the parameters that produce prices for $G_{2}$ within $\$ 0.20$ of the mean price of $\$ 2.11$. As predicted, $\gamma$ for $G_{1}$ is greater than $\gamma$ for $G_{2}$.
ition coupon was designed to affect-poor. Rottenstreich and Hsee (2001) also manipulated the probability of winning the coupons. Some participants were told they had a $1 \%$ chance of winning and others were told that they had a $99 \%$ chance of winning. They found that even though the coupons were worth the same redemption value, the median price for the European coupon was $\$ 20$ as compared to $\$ 5$ for the tuition coupon when there was only a $1 \%$ chance of winning. However, when there was a $99 \%$ chance or winning, participants were willing to pay more for the tuition coupon (median price $\$ 478$ ) than for the European coupon (median price $\$ 450$ ).

Rottenstreich and Hsee (2001) explained this preference reversal by greater involvement of the affective system in the European coupon gamble than the tuition coupon gamble. For the low probability gamble, the affective system overweights the probability of winning and thus the affect-rich European coupon is valued more than then affect-poor tuition coupon. However, for the high probability gamble, the affective system underweights the probability of winning and the European coupon is valued less than the tuition coupon.

Following the intuition of Rottenstreich and Hsee (2001), the DDP model explains this preference reversal by using a larger $\gamma$ parameter for the affect-rich European coupon as compared to the affect-poor tuition coupon. For this analysis, the $\gamma$ parameter for the European coupon gamble was set to $\gamma=1$ implying complete involvement of the affective system and no involvement of the deliberative system. For the tuition coupon gamble, $\gamma=0$ implying complete involvement of the deliberative system and no involvement of the affective system. These parameter settings reflect the extreme case when only one system is involved in the decision-making process. In reality, it is more likely that both systems are involved in
both gambles with $0<\gamma_{T}<\gamma_{E}<1$ where $\gamma_{T}$ is associated with the tuition coupon and $\gamma_{E}$ is associated with the European coupon. For ease of demonstration, the extreme $\gamma$ values were used, but the results also hold for more intermediate values of $\gamma$.

As before, $k=1, z=0, s=1, \theta=50$, and $r=0.01$. For the European coupon, the $m$ parameter was allowed to range from 0.1 to 0.9 . For the tuition coupon, $m$ has no impact because $\gamma=0$. Table 1 shows the WTP prices from DDP for the two different coupons and two different chances of winning. In the table, a range of prices is given for the European coupon showing the maximum and minimum prices as $m$ is varied. Because $m$ does not play a role in the tuition coupon, a single price is shown. From the table, the DDP model produces the same pattern of results as Rottenstreich and Hsee's experiment. Namely, the price for the European coupon is greater than the price for the tuition coupon when there is a low probability of winning and the price for the tuition coupon is greater than the price for the European coupon when there is a high probability of winning.

Table 1: WTP prices from the DDP model for the European and tuition coupons for two different probabilities of winning.

| Coupon | $1 \%$ chance | $99 \%$ chance |
| :--- | :--- | :--- |
| European $(\gamma=1)$ | $\$ 32.59-\$ 39.38$ | $\$ 303.48-\$ 491.94$ |
| Tuition $(\gamma=0)$ | $\$ 5.14$ | $\$ 492.97$ |

## Discussion

The DDP model synthesizes ideas from several lines of research in decision-making and cognitive modeling. DDP draws upon the static DSM model developed by Mukherjee (2010) to explain how dual systems of reasoning evaluate options. DDP formalizes the the formation of preferences as an accumulation of information over time similar to other dynamic models such as Decision Field Theory (Busemeyer \& Townsend, 1993). DDP also employs a Markov chain model to account for pricing elicitation methods as in Busemeyer and Goldstein (1992) and Johnson and Busemeyer (2005). In sum, DDP provides a unified theory of how dual systems interact to produce choices, response times, and prices that is grounded in decades of research.

Future work will develop new experiments to rigorously test DDP and to investigate novel predictions from the model. In particular, DDP makes new predictions about the interaction between systems and response times. The affective system is typically conceived as being fast and automatic. Thus, when it is engaged in a task, responses should be quick. DDP suggests that the relationship between response time and the affective system is not this simple. It predicts that response times are influenced by many factors including conflict between the two systems, risk attitudes, and the options themselves.

## References

Busemeyer, J. R., \& Goldstein, W. M. (1992). Linking together different measures of preference: A dynamic model of matching derived from decision field theory. Organizational Behavior and Human Decision Processes, 52(3), 370-396.
Busemeyer, J. R., \& Townsend, J. T. (1993). Decision field theory: A dynamic cognition approach to decision making. Psychological Review, 100, 432-459.
Damasio, A. (1994). Descartes' error: Emotion, reason, and the human brain. Putnam New York.
Diederich, A., \& Busemeyer, J. R. (2003). Simple matrix methods for analyzing diffusion models of choice probability, choice response time, and simple response time. Journal of Mathematical Psychology, 47(3), 304-322.
Hogarth, R. (2001). Educating intuition. University of Chicago Press.
Hsee, C., \& Rottenstreich, Y. (2004). Music, pandas, and muggers: on the affective psychology of value. Journal of Experimental Psychology: General, 133(1), 23.
Johnson, J., \& Busemeyer, J. (2005). A dynamic, stochastic, computational model of preference reversal phenomena. Psychological Review, 112(4), 841.
Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. American economic review, 1449-1475.
Kahneman, D., \& Tversky, A. (1979). Prospect theory: An analysis of decision making under risk. Econometrica, XLVII, 263-291.
Lichtenstein, S., \& Slovic, P. (1971). Reversals of preference between bids and choices in gambling decisions. Journal of experimental psychology, 89(1), 46.
Lichtenstein, S., \& Slovic, P. (1973). Response-induced reversals of preference in gambling: An extended replication in las vegas. Journal of Experimental Psychology, 101(1), 16.

Link, S., \& Heath, R. (1975). A sequential theory of psychological discrimination. Psychometrika, 40(1), 77-105.
Mukherjee, K. (2010). A dual system model of preferences under risk. Psychological review, 117(1), 243.
Rottenstreich, Y., \& Hsee, C. (2001). Money, kisses, and electric shocks: On the affective psychology of risk. Psychological Science, 12(3), 185-190.
Sanfey, A., Loewenstein, G., McClure, S., Cohen, J., et al. (2006). Neuroeconomics: cross-currents in research on decision-making. Trends in cognitive sciences, 10(3), 108116.

Slovic, P., Finucane, M., Peters, E., \& MacGregor, D. G. (2002). The affect heuristic. In T. Gilovich, D. Griffin, \& D. Kahneman (Eds.), Heuristics and biases: The psychology of intuitive judgment (p. 397-420). Cambridge University Press.
Stanovich, K., \& West, R. (2000). Individual differences in reasoning: Implications for the rationality debate? Behavioral and brain sciences, 23(5), 645-665.

# Expert marker of Chinese character recognition: Left-side bias versus holistic processing? 

Ricky Van-yip Tso (rickytso@hku.hk)<br>Center for Advancement of Chinese Language Education and Research, University of Hong Kong 608, Meng Wah Complex, Pokfulam Road, Hong Kong<br>Terry Kit-fong Au (terryau@hku.hk)<br>Janet Hui-wen Hsiao (jhsiao@hku.hk)<br>Department of Psychology, University of Hong Kong<br>627 Jockey Club Tower, Centennial Campus, Pokfulam Road, Hong Kong SAR


#### Abstract

Holistic processing and left-side bias are both behavioral markers of expert face recognition. In contrast, expertise in Chinese character recognition involves left-side bias but reduced holistic processing (Hsiao \& Cottrell, 2009). Here we examine whether the reduction in holistic processing associated with expert Chinese character recognition can be better explained by writing rather than reading experience. Compared with non-Chinese readers (novices), Chinese readers who had limited writing experience (Limited-writers) showed increased holistic processing, whereas Chinese readers who could also write characters fluently (Writers) showed reduced holistic processing. These results suggest that writing/sensorimotor experience can modulate holistic processing effects, and that the reduced holistic processing observed in expert Chinese readers may depend on writing rather than reading experience. By contrast, both Writers and Limited-writers showed a similar level of left-side bias in processing symmetric Chinese characters, left-side bias may therefore be a consistent expertise marker for object recognition uninfluenced by motor experience.


Keywords: Chinese character recognition, holistic processing, reading, writing, left-side bias

## Introduction

Holistic processing (HP) is the tendency to process separate features of an object as a single whole unit (Richler, Wong, \& Gauthier, 2011), and it is shown to be a behavioral marker of face recognition expertise. Some have speculated that HP applies to other types of expert-level object recognition because it facilitates within-category discrimination by incorporating featural and configural information beyond individual parts (e.g., Bukach et al., 2006; but for a contrasting view, see McKone et al., 2007). For example, training participants to recognize novel artificial objects, Gauthier and colleagues (1998) found a positive correlation between HP and expertise in within-category object recognition. Wong and colleagues (2009) also showed that participants had an increase in HP when trained to individualize an artificial object type.

Left-side bias (LSB) is also consistently reported in face perception; it refers to the effect that a chimeric face made from two left half-faces is usually judged more similar to the original face compared with one made from two right half-faces from the viewer's perspective (Brady, Campbell, \& Flaherty, 2005; Fig. 1), perhaps due to right hemisphere (RH) involvement in face recognition (Burt \& Perrett, 1997).


Fig. 1. Examples of chimeric face stimuli. Two left halves of an original face (middle) were combined to form the left chimeric face (left), and the two right halves formed the right chimeric face (right).

Chinese characters, sharing many visual properties with faces, may induce similar processing effects for expert readers in face recognition (McCleery et al., 2008). More specifically, the Chinese writing system is logographic; Chinese characters have a homogenous, square configuration, and each character is a grapheme that maps onto a morpheme (Shu, 2003). Strokes are the basic units of Chinese characters which combine to form more than 200 basic stroke patterns in the Chinese writing system (Hsiao \& Shillock, 2006); these stroke patterns in turn form the characters. A typical literate recognizes 3,000 to 4,000 characters. In addition, Chinese characters are generally recognized regardless of variations in font and handwriting style, similar to face recognition regardless of differences in facial expressions (Hsiao \& Cottrell, 2009), and experts recognize Chinese characters individually like faces (Wong \& Gauthier, 2006).

Indeed, similar to face recognition, Hsiao and Cottrell (2009) showed that expert Chinese readers demonstrated left side bias when viewing mirror-symmetric Chinese characters, whereas novices did not. Their finding suggests that LSB is an expertise marker for both face and Chinese character recognition and was consistent with research suggesting a RH involvement in Chinese orthographic processing (e.g. Yang \& Cheng, 1999). However, unlike face perception, the expertise marker for Chinese character recognition turned out to be reduced HP (Hsiao \& Cottrell, 2009). Experienced Chinese readers engage in less HP than novices in perceiving Chinese characters; perhaps they are more sensitive to the constituent components of Chinese characters and can more readily ignore some configural information unimportant for character recognition, such as exact distances between features (Ge et al., 2006). Such constituent components may not look easily separable to novices, probably because novices are less able to distinguish individual features and
components in Chinese characters (Ho, Ng, \& Ng, 2003). Hsiao and Cottrell (2009) have therefore suggested that HP is not a general expertise marker for object processing; it depends on the features of the stimuli and the tasks typically performed on the stimuli (see also Wong et al., 2009).

Note however that learning to read Chinese characters is different from learning to recognize faces-for instance, while a typical Chinese reader can read and write characters proficiently, one is not expected to draw out all the faces seen every day. Thus, the reduced HP effect in expert Chinese character processing, in contrast to expert face processing, may be related to expert readers' writing rather than reading experience. Unlike writing alphabetic words, which only requires recalling a few dozens of letters in an alphabet together with the specific combinations corresponding to their sounds, writing Chinese characters requires retrieving more than a thousand pieces of script information from long term memory. One may have to attend analytically to detailed stroke patterns of individual Chinese characters in order to memorize and write them. Perhaps expert Chinese readers in Hsiao and Cottrell's (2009) study had reduced HP because of their writing experience. Indeed, Zhou and colleagues (2012) found that artists with face drawing experiences had reduced holistic face processing compared with ordinary observers.

In Hong Kong, although the internal structures of Chinese characters are not explicitly emphasized in formal lessons, Chinese children acquire better orthographic awareness as they progress to higher grades (Ho et al., 2003). One explanation has to do with motor programming through extensive copying and reading at school (Tan et al., 2005). Copying performance (McBride-Chang et al., 2011; Tan, et al., 2005), and dictation performance (McBride-Chang et al., 2011) is correlated with reading performance. Writing performance may predict reading performance because children may consolidate knowledge of orthographic structures of characters with graphomotor memory of strokes as they copy the stroke sequences (Tan et al., 2005). Learning to write indeed seems to strengthen Chinese character recognition (Guan et al., 2011); writing experience also seems to shape the neural representation specialized for reading (e.g. James \& Atwood, 2009; Longcamp et al., 2003). Together, these results suggest a close relationship between increasing sensory-motor integration through writing practice and reading skills development.

However, Tso, Au and Hsiao (2011) identified some Chinese readers who have high reading proficiency but far poorer writing ability - whom we will call "Limited-writers (LW)". They are usually students or graduates of international schools who have learned to "write" in Chinese using computer software that converts input in a phonic alphabet (e.g., the Pinyin system) into Chinese characters, expatriates living in Chinese speaking countries, or overseas Chinese immigrants who learned to read in Chinese from environmental prints including Chinese mass media. Because writing in Chinese is more complex and resource-intensive than writing in an alphabetic language (Chan et al., 2006;

Chung \& Ho, 2010), marked discrepancy between reading and writing performance in Chinese is possible. With limited writing practice but plenty of reading experience, LW may recognize the holistic structures of characters similarly to face recognition, with limited analysis of the constituent structures. Thus, the cognitive processes involved in Chinese reading for LW may be different from readers who have received intensive character writing training (Writers). Without extensive writing experiences, these LW may still process Chinese characters holistically.

Here we aim to investigate whether perceptual expertise effects such as holistic processing (HP) and left-side bias (LSB) effects can be modulated by motor experience through examining how novices, Chinese Writers and limited-writers (LW) process Chinese characters. We first examine whether Writers perceive characters less holistically than LW, and whether the reduced HP effect is related to their reading and writing performance. Since writing practice may enhance orthographic awareness of characters and de-emphasize configural information in character recognition, Writers may perceive characters less holistically than LW, and this effect may be related to their difference in writing rather than reading performance - contrary to what the research literature suggests. The ability to perceive characters analytically (less holistically) may also be the underlying mechanism for how writing experience enhances Chinese character recognition. We also predict that compared to novices, increase in HP marks expert Chinese character recognition in LW whereas Writers show reduced HP.

We then examine whether LW and Writers have a similar LSB effect in Chinese character perception. Brady et al. (2005) showed that the LSB effect in face perception was stronger when viewing familiar faces compared with unfamiliar faces; this phenomenon suggests that the LSB effect may be related to familiarity with the stimuli. Since both Writers and LW are proficient readers and thus are familiar with Chinese characters, we predict that Writers and LW will have a similar degree of LSB in perceiving Chinese characters, while no LSB is shown in novices.

## Methods

## Participants

60 participants in Hong Kong participated in our study. 20 participants reported having no prior experiences in reading Chinese characters (i.e. novices); the remaining 40 were Cantonese native-speaking Chinese readers: 20 of them had always attended traditional local schools and reported to have fluent reading and writing proficiency (i.e., Writers), while 20 had either studied overseas or at international schools and had not received formal Chinese lessons that prepared students for the local public Chinese examinations (i.e., Limited-writers, LW). All LW reported being capable of reading Chinese but with limited writing ability. Writers' and LW's reading and writing abilities were tested by a word-naming and a dictation task respectively (see Procedures); their performance was used to corroborate their self-reports. That is, LW were expected to have similar
performance in the word-naming task as Writers, but have poorer performance in the dictation task (see Results). They were all right-handers, had normal or corrected-to-normal vision and similar college-level education background.

## Procedures

## Test for holistic processing

We adopted procedures from Hsiao and Cottrell (2009). 80 pairs of medium to high frequency Chinese characters in Ming font were chosen (character frequency information was obtained from Ho and Kwan, 2001). In each trial, participants were presented with two characters and instructed to attend to only half (either top or bottom) of each character and judge whether they were the same or different. Twenty pairs were presented in each of the four conditions (Fig. 2a): same in congruent trials, different in congruent trials, same in incongruent trials, and different in incongruent trials. The complete composite paradigm (Gauthier \& Bukach, 2007) was adopted so that in congruent trials, the attended and irrelevant halves corresponded to the same response (i.e., both were the same or different) while in incongruent trials, the attended and irrelevant halves corresponded to different responses. Holistic processing was operationalized as the performance difference between the congruent and incongruent trials; it reflected the amount of interference from the irrelevant parts in the matching of the attended parts. This paradigm was adopted to avoid response biases that may occur in the partial composite design in which the irrelevant halves are always different (see Richler, Cheung, \& Gauthier, 2011).


Fig. 2. (a) Illustration of stimulus pairs in the complete composite paradigm; the attended components are shaded in red. (b) Trial sequences.

After $1,000 \mathrm{~ms}$ of central fixation in each trial, participants were cued with a symbol that directed their attention to the particular halves of the stimuli. The pair of characters was then presented, with one above and one below the initial fixation point, followed by a mask. During the 500 ms presentation time, participants looked at each character once and responded as quickly and accurately as possible by pressing corresponding buttons to judge if the character parts were the same or different (Fig 2b). We measured participants' discrimination sensitivity $A^{\prime}$ as:

$$
A^{\prime}=0.5+\left[\operatorname{sign}(H-F) \frac{(H-F)^{2}+|H-F|}{4 \max (H, F)-4 H F}\right]
$$

H and F are the hit rate and false alarm rate, respectively. $A^{\prime}$ is a bias-free nonparametric measure of sensitivity; we did not use d' because response biases may affect its measurement when assumptions of normality and homogeneity of variance are not met(Stanislaw \& Todorov,
1999). The $A^{\prime}$ difference between incongruent and congruent trials (i.e., Holistic $A^{\prime}$ ) measures HP—a more positive value marks a stronger HP effect.

## Test for left-side bias

We adopted the procedure from Hsiao and Cottrell (2009). 80 Chinese mirror-symmetric characters of high frequency were selected (Ho \& Kwan, 2001). There were a total of 160 trials with each character presented twice: once in Ming font (a common font in print) and once in Feng font (an unfamiliar font that simulates handwriting; Fig. 3). For characters presented in each font, mirror images were used in half of the trials; if a character was presented in Ming font, then the mirror image of the character was presented in Feng font, and vice versa; this was to counterbalance any differences between the two sides of each character. For each character, we counterbalanced the fonts used for the original and mirror-image characters across participants.


Fig. 3.An example of a Ming font (a) and a Feng font (b) character.


Fig. 4. (a) Examples of the stimuli, and (b) the test sequence in the LSB experiment(note that the chimeric characters are still legal Chinese characters).

For each character image, the left chimeric character was created from two left halves and the right chimeric character was created from two right halves of the character (Fig. 4a), similar to chimeric faces. Each character spanned about 6.7 degree of visual angle with a viewing distance of 55 cm . In each trial, after $1,000 \mathrm{~ms}$ of a central fixation, the original character was presented randomly either on the left or right side of the screen, at about 7.2 degree of visual angle away from the center. The left and right chimeric characters were presented along with the original image, with one above and one below an arrow at the center; the arrow directed the location of the original character at which participants were told to look first. Each character was about 3 degree of visual angle away from the center. The stimuli stayed on the screen until participants' response. Participants judged which of the two chimeric characters looked more similar to the original one by pressing the corresponding buttons (Fig. 4b). We measured the LSB effect as the percentage of trials in which the left chimeric character was selected.

## Tests for reading and writing performance:

Naming and dictation tasks were administered to assess, respectively, reading and word recalling/writing abilities.

Naming task: Participants read aloud 40 two-character words arranged from high to low frequency (According to Taiwan Ministry of Education, 1997) as quickly and accurately as possible. Each trial started with a central fixation cross for 500 ms , followed by the character presentation. After the response, the screen turned blank and the experimenter pressed a button to record the accuracy and to start the next trial. The response time was measured as the time difference between the stimulus onset and the onset of the pronunciation, detected by a microphone.

Dictation task: Participants wrote down 40 two-character words (same words as in the naming task) as quickly and as accurately as possible when they heard each word said in a female voice presented by a computer. Two-character words were used instead of characters to reduce ambiguity due to the many homophonic characters in the Chinese lexicon. Each trial started with the words "Get ready" on the screen for 500 ms . After hearing the word, participants pressed buttons to indicate whether they could recall the word or not, before they started writing. After they finished writing, the experimenter pressed a button to indicate accuracy and to reveal the next word. Accuracy was recorded.

These experiments were all conducted using E-prime v2.0 (Psychology Software Tools, Pittsburgh, PA).

## Results

Chinese reading and writing proficiency (Writers vs. LW) ANOVA revealed that Writers and LW did not differ in word naming accuracy, $F(1,38)=.471$, n.s., suggesting that both groups had high reading proficiency for words. Nevertheless, Writers had significantly shorter response times (RT) in word naming than $\mathrm{LW}, \mathrm{F}(1,38)=12.365, \mathrm{p}<.01$. In the dictation task, Writers were significantly more accurate than LW, $F(1$, 38) $=140.53, p<.001$. Fig. 5a contrasts the discrepancy between dictation (word writing) and word naming accuracy in Writers and LW (i.e., they had similar word reading accuracy but differed in dictation/writing accuracy).

## Holistic Processing

Repeated-measures ANOVA was used to investigate HP effects (congruency: congruent vs. incongruent x group: novices vs. LW vs. Writers). On $A^{\prime}$, we found a main effect of congruency, $F(1,57)=21.83, p<.001$, and an interaction between congruency and group, $F(2,57)=5.421, \mathrm{p}<.01$, but no main effect of group, $F(2,57)=.433$, n.s. Both novices and LW had a significantly smaller $A^{\prime}$ in incongruent trials than in congruent trials $(t(19)=3.592, p<.01$, and $t(19)=$ 5.001, $p<.001$, respectively), while this difference was not significant for Writers, $t(19)=0.390$, n.s. In a post-hoc analysis, novices had a larger Holistic $A^{\prime}$ than Writers, $t(38)=$ 2.160, $p<.05$, but a marginally small Holistic A' than LW, $t(38)=1.58, p=0.089$. LW had a larger Holistic $A^{\prime}$ than Writers $t(38)=2.832, p<.01$ (Fig. 5b). For RT, we found a main effect of congruency, $F(1,57)=13.05, p<.01$, and an interaction between congruency and group, $F(2,57)=4.18, p$ $<.05$, but no main effect of group, $F(1,57)=2.26$, n.s. LW responded significantly more slowly in incongruent trials ( $M$ $=592 \mathrm{~ms}$ ) than in congruent trials $(M=499 \mathrm{~ms}), t(19)=5.489$,
$p<.001$, while both Writers and novices recorded similar response times in congruent ( $M=476 \mathrm{~ms}$ and $M=569 \mathrm{~ms}$ respectively) and incongruent trials ( $M=488 \mathrm{~ms}$ and $M=$ 611 ms respectively), $t \mathrm{~s}(19)=0.894$, n.s.

These results reveal an inverted U-shape pattern in which Writers perceived Chinese characters less holistically than LW, while novices perceived Chinese characters more holistically than Writers ${ }^{1}$ but less holistically than LW.


Fig.5. (a) Accuracy rate of Limited-writers and Writers for the dictation and word naming task (***p $<.001$ ). (b) $A^{\prime}$ of Limited-writers and Writers in congruent and incongruent trials of the holistic processing task ( ${ }^{* *} p<.01$ ).

## Left-side bias

We found that both Writers and LW had a stronger LSB effect in Ming font than in Feng font $(t(19)=2.111, p<.05$ : and $t(19)=2.778, p<.05$, respectively), while this font effect in novices was not significant $(t(19)=.693$, n.s.). There was a significant LSB effect in Ming font in both Writers, $\mathrm{t}(19)=$ 2.378, p < .05, and LW, $\mathrm{t}(19)=2.271, \mathrm{p}<.05$, whereas no significant LSB was found in Feng font in either Writers or LW. Novices neither showed LSB in Ming font nor Feng font (Fig 7).When we compared Writers with LW, there was no group or font effect, nor interaction between group and font; this showed that both Writers and LW had a similar degree of LSB in perceiving Chinese characters in either font. On the other hand, when we compared novices with either Writers or LW, novices had a smaller LSB in Ming font than Writers, $t(38)=2.394, p=.022$ and LW, $t(38)=2.396, p=.022$. These results suggested that expert readers exhibited LSB for Chinese characters only in a familiar font (Ming) but not in

[^238]an unfamiliar font (Feng), and LSB is a consistent expertise marker for Chinese character recognition unaffected by writing experience.


Fig. 6. Preference for left chimeric characters in Novices, Writers and Limited-writers in Ming and Feng fonts ( ${ }^{*} \mathrm{p}<.05$ ).

## Discussion

Here we investigated how different learning experience modulates perceptual expertise effects, including HP and left side bias, through examining whether the following groups differ in how they process Chinese characters: Chinese Writers, who read and write Chinese proficiently; LW, who had similar Chinese reading proficiency as Writers (as measured by word naming accuracy) but much poorer writing performance than Writers (as measured by word dictation accuracy); and novices of Chinese characters. Compared with novices, LW processed Chinese characters more holistically, whereas Writers processed Chinese character less holistically. This U-shape pattern suggests that the reduced HP observed in expert Chinese readers (i.e., Hsiao \& Cottrell, 2009) may be related to writing rather than reading performance, or more specifically, the ability to recall and write Chinese characters. These results are consistent with Zhou et al.'s (2012) findings that artists with face drawing experiences had reduced holistic face processing compared with ordinary observers. These effects suggest a close relationship between writing/motor experience and reduced HP in the recognition of Chinese characters/faces/visual stimuli. LW perceived Chinese characters more holistically than novices, consistent with the expertise hypothesis. It seems that HP is still an expertise marker for Chinese character recognition for experts with little or no writing experiences with Chinese characters. Consistent with previous evidence for sensorimotor learning influencing perception (James \& Atwood, 2009; Longcamp, et al., 2003), here we showed how writing experiences could be associated with reduced HP in Chinese character recognition. Note however that LW had slower naming time for Chinese words compared with Writers; thus, they were not as expert at reading Chinese as Writers, given that naming RT has been frequently used as a measure of perceptual expertise (e.g., Tanaka, Curran, \& Sheinberg, 2005). A larger HP observed in LW may indicate an intermediate perceptual change from novices to high-performing experts in Chinese character recognition. Future work will further examine the relationship between HP and writing/motor experience by training novices to
recognize Chinese characters/visual stimuli under different instruction conditions and observe their changes in HP.

Our study also showed that both Writers and LW had a significant left side bias effect in perceiving characters in Ming font (a familiar font) but not those in Feng font (an unfamiliar font); while novices showed no LSB effects. The LSB in Chinese character perception seems to depend on font familiarity. Since both Writers and LW exhibited a similar degree of LSB, writing/motor experience does not seem to modulate the LSB effect. The font familiarity effect is consistent with Brady et al.'s (2005) finding that people showed stronger perceptual asymmetries for familiar faces than for unfamiliar faces; however, their participants showed LSB even in the perception of unfamiliar faces, whereas in our study, the participants did not have significant LSB in an unfamiliar font. This may be due to processing differences between face and Chinese character recognition. In particular, configural information, i.e., distances between parts have been shown to be important in face recognition (Farah et al., 1998) but not in Chinese character recognition, since changes in distance among character components do not change the character identity (e.g., Ge et al., 2006). Recent literature has also suggested a link between configural processing and RH lateralization (see Hsiao \& Cheung, 2011). Thus, face recognition may involve more RH processing than Chinese character recognition, and this involvement of RH configural processing may be transferable to the processing of unfamiliar faces/novel exemplars of a category. In contrast, the LSB/RH lateralization of Chinese character processing may be specific to familiar stimuli; this effect is consistent with the literature showing that the left visual field/RH advantage in tachistoscopic Chinese character identification was only found in real characters, but not in non-existing characters such as pseudo- or non-characters (Cheng \& Yang, 1989). This difference between face and Chinese character recognition also suggests that the RH lateralization in face and Chinese character processing may involve different cognitive processes (Hsiao \& Cheung, 2011).

In conclusion, our study is the first to report on the community of proficient Chinese readers with limited writing ability and to suggest a close relationship between writing experience - rather than reading experience as suggested by prior research - and reduced HP in Chinese character recognition. We uncovered an inverted U-shape pattern: compared with novices, increased HP marked the expertise in LW, while reduced HP marked a higher level of expertise in Writers. In contrast, the LSB effect of Chinese characters depended on font familiarity and is uninfluenced by writing experiences. Our results offer a window on HP and LSB in relation to expertise of complex object recognition by showing that HP can be modulated by both visual and motor experiences, whereas the LSB seems to be a reliable expertise marker not affected by motor experience.

## Acknowledgments

We are grateful to the Research Grant Council of Hong Kong (project code: HKU 745210H to J.H. Hsiao) and the HKU Foundation (Seed Grant \#10401359 to J.H. Hsiao).

## References

Bukach, C.M., Gauthier, I., \& Tarr, J.M. (2006). Beyond faces and modularity: The power of an expertise framework. T.I.C.S., 10, 156-166.
Brady, N., Campbell, M., \& Flaherty, M. (2005). Perceptual asymmetries are preserved in memory for highly familiar faces of self and friend. Brain Cognit., 58, 334-342.
Chen, Y.P., Allport, D.A., \& Marshall, J.C. (1996). What are the functional orthographic units in Chinese word recognition: The stroke or the Stroke pattern? . Q. J. Exp. Psychol-A, 49, 1024-1043.
Cheng, C.M., \& Yang, M.J. (1989). Lateralization in the visual perception of Chinese characters and words. Brain Lang., 36, 669-689.
Chung, K.K.H., \& Ho, C.S. (2010). Second language learning difficulties in Chinese children with dyslexia: What are the reading-related cognitive skills that contribute to English and Chinese word reading? J. Learn Disabil., 43(3), 195-211.
Farah, M.J., Wilson, K.D., Drain, H.M., \& Tanaka, J.N. (1998). What is "special" about face perception? Psychol. Rev., 105, 482-498.
Ge, L., Wang, Z., McGleery, J.P., \& Lee, K. (2006). Activation of face expertise and the inversion effect. Psychol. Sci., 17, 12-16.
Guan, C.Q., Liu, Y., Chan, D.H.L., Ye, F., \& Perfetti, C.A. (2011). Writing strengthens orthographic and alphabetic-coding strengthens phonology in learning to read Chinese. J. Educ. Psychol., 103(3), 509-522.
Gauthier, I., Williams, P., Tarr, M.J., \& Tanaka, J. (1998). Training "Greeble" experts: Aframework for studying expert object recognition processes. Vision Res., 38, 2401-2428.
He, A.G., Tan, L.H., Tang, Y., James, A., Wright, P., \& Eckert, M. A. (2003). Modulation of neural connectivity during tongue movement and reading. H.B.M., 18, 222-232.
Ho, C.S., \& Kwan, T.W. (2001). Hong Kong, Mainland China \& Taiwan: Chinese character frequency-A trans-regional, diachronic survey. Retrieved December 30, 2011, from http://arts.cuhk.edu.hk/Lexis/chifreq/
Ho, C.S., Ng, T., \& Ng, W. (2003). A radical approach to reading development in Chinese: The role of semantic radicals and phonetic radicals. J. Lit. Res., 35, 849-878.
Hsiao, J.H., \& Cheung, K.C.F. (2011). Computational exploration of the relationship between holistic processing and right hemisphere lateralization in featural and configural recognition tasks. Proceedings of the 33rd Annual Conference of the Cognitive Science Society.
Hsiao, J.H., \& Cottrell, G. (2009). Not all visual expertise in holistic, but it may be leftist: The case of Chinese character recognition. Psychol. Sci., 20(4), 455-463.
Hsiao, J.H., \& Shillock, R. (2006). Analysis of a Chinese phonetic compound databse: Implications for orthographic processing. J. Psycholinguist. Res., 35, 405-426.

James, K.H., \& Atwood, T.P. (2009). The role of sensorimotor learning in the perception of letter-like forms: Tracking the causes of neural specialization for letters. Cognitive Neuropsychology, 26, 91-110.
Longcamp, M., Anton, J.L., Roth, M., \& Velay, J.L. (2003). Visual presentation of single letters activates a premotor area involved in writing. NeuroImage, 19, 1492-1500.
McBride-Chang, C., Chung, K.K.H., \& Tong, X. (2011). Copying skills in relation to word reading and writing in Chinese children with andwithout dyslexia. J. Exp. Child Psychol., 110, 422-433.
McCleery, J.P., Zhang, L., Ge, L., Wang, Z., Christiansen, E. M., Lee, K. (2008). The roles of visual expertise and visual input in the face inversion effect: Behavioral and neurocomputational evidence. Vision Res., 48, 703-715.
McKone, E., Kanwisher, N., \& Duchaine, B. C. (2007). Can generic expertise explain special processing for faces? T.I.C.S., 11, 8-15.

Richler, J. J., Wong, Y. K., \& Gauthier, I. (2011). Perceptual expertise as a shift from strategic interference to automatic holistic processing. Curr Dir. Psychol. Sci., 20(2), 129-134.
Shu, H. (2003). Chinese writing system and learning to read. International Journal of Psychology, 38, 274-285.
Siok, W.T., Perfetti, C.R., Jin, Z., \& Tan, L.H. (2004). Biological abnormality of impaired reading is constrained by culture. Nature, 431, 71-76.
Stanislaw, H., \& Todorov, N. (1999). Calculation of signal detection theory measures. Behav. Res. Meth. Inst. C., 31, 137-149.
Taiwan Ministry of Education (1997). Report on frequently used vocabulary in 1997 (in Chinese). Retrieved December 30, 2010 from www.edu.tw/files/site_content/M0001/86news/ch2.html
Tan, L.H., Spinks, J.A., Eden, G.F., Perfetti, C.A., \& Siok, W.T. (2005). Reading depends on writing, in Chinese. PNAS, 102, 8781-8785.
Tanaka, J.W., Curran, T., \& Sheinberg, D. (2005). The training and transfer of real-world, perceptual expertise. Psychol. Sci., 16(141-151).
Tso, R.V.Y., Au, T.K., \& Hsiao, J.H. (2011). The influence of writing experiences on holistic processing in Chinese character recognition. Proceedings of the 33rd Annual Conference of the Cognitive Science Society, Boston.
Wong, A.C., \& Gauthier, I. (2006). An analysis of letter expertise in a levels-of-categorization framework. Visual Cognition, 15, 854-879.
Wong, A.C., Palmeri, T.J., \& Gauthier, I. (2009). Conditions for face-like expertise with objects: Becoming a ziggerin expert-but which type? Psychol. Sci., 20, 1108-1117.
Zhou, G., Cheng, Z., Zhang, X., Wong, A.C.-N. (2012). Smaller holistic processing of faces associated with face drawing experience. Psychon. B. Rev., 19(2), 157-162.
Yang, M.J., \& Chen, C.M. (1999).Hemisphere differences in accessing lexical knowledge of Chinese characters. Laterality, 4, 149-166.

# A distributional learning account of the acquisition of the locative alternation: Corpus analysis and modeling 

Katherine E. Twomey (k.twomey@liverpool.ac.uk)<br>Franklin Chang (franklin.chang@liverpool.ac.uk)<br>Ben Ambridge (ben.ambridge@liverpool.ac.uk)<br>University of Liverpool, Department of Experimental Psychology, Eleanor Rathbone Building, Bedford St South, Liverpool, L69 7WW, UK


#### Abstract

Early in acquisition children overgeneralize verbs to ungrammatical structures. The retreat from overgeneralization is linked to the acquisition of verb classes, the semantics of which constrain the structures in which a verb can appear (e.g., Pinker 1989; Ambridge, Pine \& Rowland, 2012). How children learn these classes remains unclear. Some argue that distributional regularities in linguistic input provide sufficient evidence for verb classes to emerge and become linked to particular structures. A corpus analysis of the English locative construction (e.g., the woman sprayed water onto the wall/the wall with water) demonstrated that children have similar verb classes to adults. A correspondence analysis revealed that distributional regularities in the input could support these verb classes. Finally, a connectionist simulation was able to model early overgeneralization and retreat through distributional learning of verb classes. These results support a distributional learning account of verb semantics.


Keywords: Child language acquisition; locative construction; distributional learning; naturalistic speech; Dual-Path model.

Children's acquisition of grammatical constructions remains one of language acquisition's most puzzling phenomena. Even young children are readily able to generalize known constructions to new verbs, and indeed, early in acquisition children overgeneralize these constructions to contexts in which they are ungrammatical (e.g., *she filled the juice into the glass). However, over time children learn to constrain their choice of structure, despite the lack of evidence for ungrammaticality in the adult speech they hear (Ambridge, Pine, Rowland, Jones \& Clark, 2009).

This remarkable ability has been linked to the semantics of the verbs that appear in these constructions. According to this semantic verb class hypothesis (Pinker, 1989), verbs fall into discriminable classes based on the semantics of the event they describe. These verb classes then constrain the structures in which verbs can appear. For example, pour and drip both describe the movement of an object/substance into/onto a location via gravity, and both appear in the theme-locative (TL) structure only (e.g., the woman poured water into the bucket vs. *the woman poured the bucket with water). In contrast, cover and coat both describe a location being completely covered with a layer of substance or object(s) (e.g., the woman covered the table with the blanket vs. *the woman covered the blanket onto the table). Finally, spray and squirt both describe ballistic motion of a
liquid along a trajectory, are in a class that does not emphasize theme or location, and therefore alternate between both LT and TL structures (e.g., the woman sprayed the wall with water vs. the woman sprayed water onto the wall). Although Pinker's verb classes motivated early work on such verb-structure mappings, recent work has questioned whether it is these particular classes which guide structure choice (Ambridge, Pine \& Rowland, 2012). Nonetheless, it is widely agreed speakers categorize verbs, and that these classes help to guide structural choices and explain overgeneralization.

One account of the acquisition of semantics is based on using situational information in the world (St. Augustine, 397/2001). For verbs, this might involve learning that verbs like coat refer to events where an object is completely covered with some substance. Since the object is more saliently changed by the action than the substance, the situational meaning would classify this verb as belonging to a verb class that prefers the LT structure, which emphasizes the LOCATION (e.g., the man coated the $[\text { table }]_{\text {LOCATION }}$ with paint). However, this situational account of verb meaning is limited: it may not be the case that every child hears every verb that they know in an unambiguous situational context that classifies it precisely (e.g., Harris, Jones \& Grant, 1983).

An alternative, distributional account states that verb semantics are learned from distributional regularities in the linguistic input (Fisher, Gleitman, \& Gleitman, 1991). For example, if spray is always followed by a noun phrase which specifies a liquid (e.g., spray the water), then children might categorize spray with other verbs that take similar arguments (e.g., splash the water). This verb class might then be associated with certain locative structures (e.g., he sprayed/splashed the water onto the wall). The advantage of this account over the situational account is that it allows verb classes to be acquired from heard input without any situational context.

To examine how these accounts relate to behavioral data, we performed a corpus analysis of the British English locative. Then, to explore whether distributional information is sufficient to support the emergence of verb classes, we performed a correspondence analysis on the corpus data. Finally we investigated whether distributional regularities can influence structural choice in a connectionist model of sentence production which learns both syntactic
structures and their verb class associations from the input alone.

## A corpus analysis of the English locative

We queried all UK corpora in CHILDES (MacWhinney, 2000) for the 142 locative verbs listed in Pinker (1989). A main sample of 38,231 utterances was retrieved. A subsample of 2,685 morphologically-tagged utterances was coded by hand for construction type (see Table 1). Importantly, we divided transitives into those where the post-verbal noun was a THEME (spray the water, T transitive) or a LOCATION (spray the wall, L transitive).

Table 1: Constructions and handcoded frequency

| Structure | Example | N |
| :--- | :--- | :---: |
| LT locative | brushing me with it? | 100 |
| TL locative | spread it on your biscuit | 232 |
| L transitive | for him filling the glasses | 271 |
| T transitive | you dump the lady's toys | 294 |
| Ambiguous transitive | brushed it | 608 |
| L intransitive | tape on | 24 |
| T intransitive | it attached on like that | 20 |
| Ambiguous | I brushing | 205 |
| intransitive |  |  |

The subsample was generated by selecting all utterances containing candidate locative verbs tagged as a verb or a participle. Utterances were separated by speaker (adult or child) and post-verbal preposition (with, indicating a candidate LT locative; into, onto or over, indicating a candidate TL locative; or no preposition, indicating some other construction). For the 13 categories containing more than 50 utterances, a random sample of 50 utterances was coded. Since there were very few tagged utterances for conversations involving children over 40 months, we extracted an extra $20 \%$ of child utterances and $5 \%$ of adult utterances from the untagged data for this age group. We excluded 221 non-verb utterances from the analyses. As can be seen in Table 1, only $6 \%$ of the coded locative verbs occurred in the LT locative structure, and $13 \%$ occurred in the TL locative structure.

The situational account predicts that verb classes in children and adults can differ, especially if children do not always understand events in adult-like ways. On the other hand, the distributional account predicts that children learn their verb classes from adult linguistic utterances, so there should be a close match between the two. To examine these predictions, we calculated the proportion of LT out of total locative constructions in the hand-coded data. This controls for raw frequency of each form and variation due to other structures, providing a measure of the bias for these two structures. Figure 1 depicts proportion LT for both children and adults for a range of verbs that occur more than 10 times in the hand-coded adult data. There seems to be a class of non-alternating LT verbs like fill, a class of nonalternating TL verbs like pour, and several alternating verbs
like $r u b$ and splash. Importantly, children have similar verb classes to adults, even though these data come from a range of different situations, which children may or may not understand in adult-like ways.


Figure 1: Proportion LT structures, adult and child speakers

To see whether these structural preferences change over development, we calculated proportion of adult and child transitive and locative utterances with location as object (L or LT) for each age of child included in the sample (Figure 2). A linear model was fit to the data with Structure (locative $=1$, transitive $=0$ ), Child Age (months) and Speaker Group (adult vs. child) crossed. There was no main effect of age, suggesting that $L$ transitive proportion did not vary over age. LT production was lower than L transitive production (beta $=-0.9, t(116)=3.1, p<.003)$. An interaction between Age and Structure (beta $=0.02, t(116)=$ $3.1, p<.003$ ) provided evidence for an increase in the production of LT structures over development. The lack of any interaction with Speaker Group suggests that adults adapted their structural choices to fit children's preferences. The TL bias in young children is also seen in experimental studies, where children reproduce TL structures at a higher rate than LT structures (Gropen et al., 1991a, 1991b), and in diary studies, where TL overgeneralizations are more frequent (Bowerman, 1982).

In our data set, transitives are more frequent than locatives - we found 565 transitive utterances, but only 332 locative utterances. Of the 50 verbs that appeared in a transitive, a locative or both structures, 35 appear in the locative while 44 appear in the transitive. Using transitive contexts to learn about locative verbs provides a better coverage of these verbs, suggesting that children may be able to learn about locative verbs from their appearance in $L$ and T transitive structures.

To examine this hypothesis, we used adult ratings of locative verbs' LT and TL preference from Ambridge et al. (2012) and Bidgood, Ambridge, Pine \& Rowland (under
review) to create a graded LT rating measure for each verb. This was correlated with the proportion LT and L structures in the adult input to children from the corpora. There was a significant correlation between the LT rating and LT corpus distribution $(r(30)=.53, p=.002)$, indicating that the LT/TL biases of verbs in locative structures in child-directed speech match adult ratings of those verbs' LT/TL preference. There was also a significant correlation between LT rating and L corpus distribution ( $r(38)=.32, p=.04$ ), which shows that the L/T biases of verbs in transitive structures in child-directed speech also reflect LT/TL preference ratings.


Figure 2: Proportion L/LT locatives by child age

## Correspondence analysis

The corpus analysis shows that children learn biased verb classes in their first five years. We used correspondence analysis (CA; Greenacre, 2007) to investigate whether distributional regularities can support the development of these classes. CA is a technique for clustering categorical data in a low-dimensional space based on the similarity in how elements co-occur.

We performed a CA on adult utterances including utterances that were not handcoded. Similarity was calculated based on the co-occurrence of each verb with the two post-verbal words, however if a preposition (on, to, with, in, into, onto, over) occurred post-verbally, the
preposition and all following words were excluded. Thus, verbs that occurred in the same context (e.g., the woman poured water and a boy dripped water) were classified as more similar than verbs that occurred in different contexts. In order to have a range of words to support categorization, we only included verbs which occurred more than 40 times. The CA generated six factor scores for each verb.

To determine which factors in the CA best predicted the LT/TL rating measure used earlier, we submitted factor scores for each verb to a regression with LT rating as the dependent variable, and Factor (1, 2, 3, 4, 5, 6) as independent variables. LT rating was significantly predicted by Factor 4 (beta $=0.03, t(27)=2.06, p=.05)$ and Factor 6 (beta $=0.06, t(27)=3.72, p<.001$ ), explaining $39 \%$ of the variance $(F(6,27)=2.85, p=.03)$.

Figure 3 illustrates the relationship between Factors 4 and 6 and LT rating. Verbs rated as more grammatical in LT structures (e.g., the woman poured water into the bucket) cluster in the top right hand corner of the figure, and verbs rated as more grammatical in TL structures (e.g., the woman filled the bucket with water) cluster in the bottom left hand corner. Furthermore, there is evidence of clusters reflecting Pinker's (1989) narrow semantic verb classes; for example, fill and cover are both members of Pinker's cover-type class and they are clustered together. In line with existing studies (e.g., Ambridge, Pine \& Rowland, 2012), however, not all classes generated by our data conform to these narrow verb classes. For example, fill and pack are similar on Factor 4, despite being members of different classes according to Pinker's classification.


Figure 3. CA results for Factors 4 and 6.

Overall, the distributional regularities with which verbs and post-verbal words co-occurred generated two factors in a CA which predicted adult ratings of verbs' grammaticality in LT structures. Substantial evidence exists that children can track such statistical co-occurrences in their environment (Fisher, Gertner, Scott \& Yuan, 2010). Thus, if children record the distributional regularities of locative verbs, they may also create verb classes, helping them to constrain their choice of structure and retreat from overgeneralization. The CA provides an explicit model of how a distributional learning mechanism can create these classes.

## A connectionist model of locative acquisition

The corpus analysis suggested that children may learn about the semantics of locative verbs from transitive structures. Further, the correspondence analysis showed that distributional regularities in the input could support the emergence of verb classes that constrain structural choice. However, a full account requires a mechanism that can learn not only verb classes but also syntactic structures, and link them appropriately. Critically, if this mechanism is similar to the language acquisition system in children, then it should create a TL bias early in development.

To explore the nature of such a mechanism, we adapted the Dual-Path model, a connectionist model of sentence production (Chang, 2009, Chang, Dell \& Bock, 2006). The model learns linguistic representations from messagesentence pairs and can acquire different languages (e.g., English, Japanese). We trained the model with two simple grammars that included both transitive and locative structures. Grammars contained five classes of verbs, one of which was L/LT -only and one of which was T/TL -only. Our principal manipulation was the frequency with which the remaining "alternating" verb classes were associated with the various structures.

## Model

The Dual-Path model's architecture includes separate sequencing and meaning systems. The sequencing system is a simple recurrent network that allows the model to learn syntactic representations (Elman, 1990). The model's internal representations are acquired through error-based learning, in which the model predicts the next word in a sentence, then uses the difference between the prediction and target (error) to modify its internal representations. Since categories are useful for prediction, the sequencing system acquires syntactic categories that support syntactic structures. Verb classes can be seen as a refinement of verb categories to incorporate their biases. The meaning system encodes the message as a set of weights between role and concept units (e.g., AGENT=WOMAN). The simple recurrent network selects appropriate concepts by activating their role at particular positions in sentences. Detailed motivation for the model's architecture is provided in Chang (2002; 2009).

Messages include roles for the various concepts in the sentence as well as a role that selects the verb. In the
current architecture, structural alternations were modeled by associating different structures with the same message. For example, the TL locative sentence the woman sprays water onto the wall and LT locative sentence the woman sprays the wall with water have approximately the same meaning and hence share the message ACTION=SPRAY AGENT=WOMAN THEME=WATER LOCATION=WALL. When given this message, the model must decide whether to activate to LOCATION role after the verb and produce the LT structure, or the THEME role and produce the TL structure. Critically, because they are located in the meaning system, the SPRAY semantics can only be used for selecting the word spray and do not directly influence structural choices in the sequencing system. Thus, the model must acquire a syntactic representation of the verb within the sequencing system using distributional learning.

## Testing different input distributions

Verb classes in the simple recurrent network were shaped by the distributional properties of the input. To see if these classes reflected the behavior of human children's verb classes, we compared two different input distributions: Full and Transitive (Table 2).

Table 2: L/LT ratios in input grammars (LT:TL)

| Structure | Full Distribution | Transitive Distribution |
| :--- | :---: | :---: |
| Locative A | $100: 0$ | $100: 0$ |
| Locative B | $90: 10$ | - |
| Locative C | $50: 50$ | - |
| Locative D | $10: 90$ | - |
| Locative E | $0: 100$ | $0: 100$ |
| Transitive A | $100: 0$ | $100: 0$ |
| Transitive B | $90: 10$ | $100: 0$ |
| Transitive C | $50: 50$ | $50: 50$ |
| Transitive D | $10: 90$ | $0: 100$ |
| Transitive E | $0: 100$ | $0: 100$ |

Our input distributions contained five classes of verbs (A, $\mathrm{B}, \mathrm{C}, \mathrm{D}, \mathrm{E}$ ), each paired with LT and/or TL locative structures (e.g., the woman sprayed the wall with water / water onto the wall) and L and/or T transitive structures (e.g., the woman sprayed the wall/water). Verbs in class A occurred only in L/LT structures, while verbs in class E occurred only in T/TL structures. In the "Full Distribution" simulation, verbs occurred more frequently in L/LT structures in Class B, equally often in either structure type in Class C, or more frequently in T/TL structures in Class D. The "Transitive Distribution" simulation tested the idea suggested by the corpus and correspondence analysis that transitive uses of verbs might be the basis for verb classes. Classes B, C, and D were therefore only biased in transitive structures (and more strongly than in the Full Distribution) and hence tested whether the transitive distribution could influence locative generalization. The raw frequency of TL structures was also higher than that of LT structures in the Transitive Distribution to simulate the data in Table 1.

Ten randomly generated training sets were used to create ten model subjects. Each model was trained for 40,000 epochs on 2,000 message-sentence pairs. Every 1,000 epochs the model was tested on a set of 1,000 grammatical and ungrammatical locative sentences ( 100 per verb class).

Results. Figure 4 depicts proportion location-based sentences for each of the five verb classes, for each of the simulations.


Figure 4: Proportion location-based sentences
Results from the Full Distribution (Figure 4 top panel) suggest that distributional regularities in the input do indeed support the emergence of verb classes. First, the Full Distribution simulation learned nonalternating classes A and E. To establish whether the model distinguished classes B, C and D according to the frequency of $\mathrm{L} / \mathrm{LT}$ and $\mathrm{T} / \mathrm{TL}$ structures for these classes in the input, a mixed effects model was fit to Proportion LT (empirical logit transformed) with Verb Class and Epoch crossed. Verb Class was coded numerically ( $\mathrm{B}=1, \mathrm{C}=0, \mathrm{D}=-1$ ). Model subject was included as a random variable and there were by-subject slopes for Verb Class crossed with Epoch (Barr, Levy, Scheepers \& Tily, 2013). There were no main effects of Verb class or Epoch on LT production, however an interaction (beta $=0.016, \chi^{2}(1)=14.49, p<.001$ ), confirmed that verb classes differentiated as epoch increased. The negative logit intercept coefficient captures the model's initial TL bias ( beta $=-0.16$ ).

The Transitive Distribution simulation (Figure 4 bottom panel) also distinguished classes $\mathrm{B}, \mathrm{C}$ and D in the locative, despite receiving no locative input for these classes. It did so based on the frequency with which these verb classes occurred in L and T structures. Class D, for example, was a class that preferred THEME-type objects rather than

LOCATION-type objects in the post-verbal position. To test this, we fitted a mixed effects model to the Transitive Distribution data. There were no main effects of Verb Class or Epoch on LT production, but again there was an interaction (beta $=0.01, \chi^{2}(1)=6.90, p=.009$ ), demonstrating that verb classes differentiate over development. Finally, this model also exhibited an early TL bias, as denoted by the negative logit intercept coefficient (beta $=-0.54$ ). The preferences in the transitive naturally generalize to locative structures because these structures have similar argument preferences (i.e., TL locatives prefer theme-type objects in the post-verbal position). Thus, the model is able to acquire locative verb classes from transitive distributions.

Our corpus analysis revealed that children have an early preference for the TL locative. This was captured by the model. Like the children, LT production increases over development, but remains outstripped by TL production (overall proportion LT for Transitive Distribution at epoch $10,000=0.36$, epoch $40,000=0.44$; Fig. 4). This simulation captures the early TL bias because the sequencing system does not have access to verb meaning and is biased to learn syntactic categories that fit the frequent TL structure. This structure becomes the default means to express three arguments (i.e., AGENT, THEME, LOCATION). Slowly the model develops verb classes that help predict the nouns that follow verbs. These classes later help support the differential use of LT and TL structures.

## General Discussion

The current study presents naturalistic data which show that children are biased to use TL structures early in acquisition. We hypothesized that the more frequent TL overgeneralization errors in diary studies are due to this general TL bias. The increase in use of LT between two and five years can therefore be interpreted as a metric of an underlying retreat from overgeneralization. If, over development, children acquire verb classes that predict which structure to use, then they will become more likely to produce these structures in verb-specific ways, and consequently less likely to make errors. To explain how children acquire verb classes when the target classes and structures are not explicitly provided, we suggest that they perform a distributional analysis on the words they hear, in a similar manner to our CA. Here, we have shown that an unsupervised clustering algorithm can create component factors that predict structural choices.

Bayesian and other statistical mechanisms have also been applied to verb-structure learning. These systems often assume that the initial state of the learner includes syntactic structures, semantic features (e.g., cause, change-of-state), or even abstract verb classes (Alishahi \& Stevenson, 2008; Niyogi, 2002; Perfors, Tenenbaum \& Wonnacott, 2010). A clear example of this approach can be seen in Niyogi (2002), who developed a Bayesian model of the locative alternation. The model uses manner and path features from scene and syntactic frame regularities to select among verb
class hypotheses. A key feature of this model is that it can assign verb classes quickly from a few exposures; it therefore does not explain the protracted development of the LT structure in children. Furthermore, the model cannot learn locative classes from transitive input, because locative verb classes are not an optimal fit for transitive structures (superordinate motion classes are a better fit). The limitations of these models highlight the fact that language development is not just the fast, optimal weighting of syntax and semantic cues.

In contrast with these models, the Dual-Path model does not start with syntactic structures or verb class hypotheses. It learns its structures with a slow learning algorithm designed to mimic the slow biological changes that support learning in the brain (cell growth). Due to its inability to predict when to use LT and TL structures, the frequent TL initially dominates. As in the CA, the model develops verb classes from the distribution of post-verbal words in frequent transitives. Later, these verb classes become the basis for distinguishing TL and LT structures and thus condition their use.

Taken together, the current studies are the first to characterize how children use a wide range of verbs in locative structures between two to five years of age. We showed that locative production is asymmetrical during this period, with the TL structure dominating. This dominance can causes children to place a newly-learned verb into the TL structure regardless of its own bias, producing an overgeneralization error. We also showed that distributional regularities in child-directed speech are useful for creating classes that predict adult ratings of locative structure preference for this diverse set of verbs. Finally, we showed how distributional learning in the Dual-Path model can be used to learn both the syntactic categories that support structures, as well as the verb classes that bias structural choice. By trying to simultaneously solve both of these difficult learning problems, the model can explain both the early TL bias and slow development of verb classes that support the retreat from that bias.

## Acknowledgments

This research was supported by Leverhulme Trust Research Project Grant RPG-158.

## References

Alishahi, A., \& Stevenson, S. (2008). A Computational model of early argument structure acquisition. Cognitive Science, 32(5), 789-834.
Ambridge, B., Pine, J. M., \& Rowland, C. F. (2012). Semantics versus statistics in the retreat from locative overgeneralization errors. Cognition, 123(2), 260-279.
Ambridge, B., Pine, J. M., Rowland, C. F., Jones, R. L., \& Clark, V. (2009). A semantics-based approach to the "No Negative Evidence" problem. Cognitive Science, 33(7), 1301-1316.
Barr, D. J., Levy, R., Scheepers, C., \& Tily, H. J. (2013). Random effects structure for confirmatory hypothesis
testing: Keep it maximal. Journal of Memory and Language, 68(3), 255-278.
Bidgood, Ambridge, Pine \& Rowland (under review). The retreat from locative overgeneralisation errors: A novel verb grammaticality judgment study.
Bowerman, M. (1982). Reorganizational processes in lexical and syntactic development. In E. Wanner \& L. R> Gleitman (Eds.). Language acquisition: The state of the art (pp 319-346). New York: Cambridge University Press.
Chang, F. (2002). Symbolically speaking: A connectionist model of sentence production. Cognitive Science, 26(5), 609-651.
Chang, F. (2009). Learning to order words: A connectionist model of heavy NP shift and accessibility effects in Japanese and English. Journal of Memory and Language, 61(3), 374-397. doi:10.1016/j.jml.2009.07.006
Chang, F., Dell, G. S., \& Bock, K. (2006). Becoming syntactic. Psychological Review, 113(2), 234.
Elman, J. L. (1990). Finding structure in time. Cognitive Science, 14(2), 179-211.
Fisher, C., Gertner, Y., Scott, R., \& Yuan, S. (2010). Syntactic bootstrapping. Wiley Interdisciplinary Reviews: Cognitive Science 1(2), 143-149.
Fisher, C., Gleitman, H., \& Gleitman, L. R. (1991). On the semantic content of subcategorization frames. Cognitive Psychology, 23(3), 331-392.
Greenacre, M. (2007). Correspondence analysis in practice. Boca Raton, FL: Chapman \& Hall/CRC.
Gropen, J., Pinker, S., Hollander, M., \& Goldberg, R. (1991a). Syntax and semantics in the acquisition of locative verbs. Journal of Child Language, 18(1), 115151.

Gropen, J., Pinker, S., Hollander, M., \& Goldberg, R. (1991b). Affectedness and direct objects: The role of lexical semantics in the acquisition of verb argument structure. Cognition, 41(1), 153-195.
Harris, M., Jones, D., \& Grant, J. (1983). The nonverbal context of mothers' speech to infants. First Language, 4(10), 21-30.
MacWhinney, B. (2000). The CHILDES Project: Tools for analyzing talk. Third Edition. Mahwah, NJ: Lawrence Erlbaum.
Niyogi, S. (2002). Bayesian learning at the syntax-semantics interface. In Proceedings of the 24th Annual Conference of the Cognitive Science Society (Vol. 36, p. 58)
Perfors, A., Tenenbaum, J. B., \& Wonnacott, E. (2010). Variability, negative evidence, and the acquisition of verb argument constructions. Journal of Child Language, 37, 607-642.
Pinker, S. (1989). Learnability and cognition: The acquisition of argument structure. Cambridge, MA: The MIT Press..
St. Augustine (379/2001). The confessions of Saint Augustine. C. White (Trans.) London: Frances Lincoln.

# Moves in the World Are Faster than Moves in the Head: Interactivity in the River Crossing Problem 

Frédéric Vallée-Tourangeau, Lisa G. Guthrie and Gaëlle Villejoubert<br>Department of Psychology, Kingston University<br>Kingston-upon-Thames UNITED KINGDOMKT1 2EE<br>f.vallee-tourangeau/l.guthrie/g.villejoubert@kingston.ac.uk


#### Abstract

In solving a variety of problems people interact with their external environment, often using artefacts close at hand to supplement and augment their problem solving skills. The role of interactivity in problem solving was investigated using a river-crossing problem. All participants performed the task twice, once in a high interactivity condition and once in a low interactivity condition. Moves to completion were higher in the high interactivity condition but latency per move was much shorter with high than with low interactivity. Moves in the world were easier to implement than to simulate mentally and acted as epistemic actions to facilitate thinking. In addition, when participants experienced the low interactivity version of the task second, their performance reflected little learning. However, when the high interactivity version was completed second, latency to solution and latency per move were substantially reduced. These results underscore the importance of investigating problem solving behaviour from a distributed cognition perspective.


Keywords: Problem solving, interactivity, epistemic actions, distributed cognition

## Introduction

Scientists and lay people alike naturally create and build artefacts or recruit existing ones to help them solve problems. To be sure, artefacts such as calculators, data management software, computers can facilitate complex computations. But others, of more modest complexity, such as pen and paper, can help articulate and structure thinking. Space itself is a tool that can facilitate thinking, that is it can be structured, designed (and redesigned) such as to make thinking easier (Kirsh, 1995, 1996, 2010). Thus solving jigsaw puzzles involves physically juxtaposing different pieces to gauge their fit; in Scrabble, letter tiles are physically rearranged to facilitate word production; in Tetris, tetrominoes are physically rotated to determine their optimal place along a line. And beyond puzzles and games, experts structure an external environment to support thinking. Scientists use physical objects and their arrangement in space to formulate and test hypotheses: Watson (1968, pp. 123-125) describes how he cleared his desk, cut out shapes corresponding to the four nucleobases, and manipulated them until he saw
which ones could be paired to hold the double helix together. Artefacts recruited in thinking are rich, varied and modifiable. Their recruitment is at times strategic, such that their users actively engage in their design and engineer their function, and at others, opportunistic, that is they are picked up from the environment in an ad hoc fashion to help solve a problem, capitalizing on a fortuitous interaction.

From a distributed cognition perspective, thinking is the product of a cognitive system wherein internal and external resources are coupled to create a dynamic, fluid, and distributed problem representation (Villejoubert \& Vallée-Tourangeau, 2011; Weller, Villejoubert, \& ValléeTourangeau, 2011). The nature of the external resources recruited in thinking and their functional role are guided by principles of cognitive economy, effort and efficiency (Clark, 1989; Kirsh, 2010). Actions complement and augment thinking by providing new information, unveiling new affordances, and can sometimes serve to create a more cognitively congenial problem presentation (Kirsh, 1996). Through the creation, recruitment and manipulations of artefacts, new perspectives are gained, encouraging the development or retrieval of problem solving strategies, and improving the prospect of solving the problem (Magnani, 2007). As the environment shoulders some of the representational and computational burden, valuable cognitive resources such as working memory capacity and executive functions may be freed to draw on stored knowledge or develop new solutions (Magnani, 2007). For example, recent work on mental arithmetic indicates that people are more accurate, more efficient, and create more congenial interim totals when they can manipulate number tokens that configure the problem presentation, than when they perform the mental arithmetic without (Vallée-Tourangeau, in press).

## River Crossing

Transformation problems have been the focus of research in cognitive psychology for the past 50 years. In these problems, a well-defined space connects an initial and a goal state. Legal moves are defined in terms of simple rules and enacted with simple operators. Participants must reach the goal state by transforming the initial state
through a series of intermediate states. A well-studied class of transformation problems are river-crossing problems. In these problems, objects (people, animals, or things) must be carried from one "riverbank" to another on a "boat" but with a set of constraints on moves that can be selected to reach the goal. A common version involves three missionaries and three cannibals (Reed, Enrst, \& Banerji, 1974; or three hobbits and three orcs, Thomas, 1974). In transporting all cannibals and missionaries from one bank to the other, cannibals must not outnumber missionaries or either bank. The boat can take at most two passengers, and at least one. The problem space is relatively narrow since illegal moves cannot produce blind alleys of any depth (Reed et al., 1974) and can be completed in 11 steps. In different versions, problem difficulty is a function of the rules that constrain the number of objects that can be moved at any one time, which combinations of objects are allowed on the boat, and which combinations can be left on either bank. The number of objects and the rules that govern their transport map out a problem space that links the initial state with all objects on one side of the river to a goal state with all objects on the other riverbank. Cognitive psychologists have used this task as a window onto problem solving, particularly planning (Greeno, 1978), search and move selection (Reed et al., 1974; Simon \& Reed, 1976). As such river crossing problems have been used as a testing platform for a number of process models of search and move selection, strongly influenced by developments in AI (Greeno, 1978; Simon \& Reed, 1976).

The river-crossing task involves moving people or things across a surface and as such foregrounds the importance of interacting with an external task representation. However, interactivity in river crossing problem solving has never been the explicit focus of investigation. The manner with which the river-crossing task has been implemented varies a great deal across studies. For example, Reed et al. (1974) used different types of coins to represent missionaries and cannibals. Jeffries et al. (1976) developed a basic computer interface where participants typed in the objects they wanted to put in the boat on a given crossing. The interface accepted only legal moves and updated the simple representations (often with letters and numbers, such as ' 3 M ' for three missionaries) on either side of the riverbank. Participants kept on typing in their moves until they managed to transport all objects from one bank to the other. Knowles and Delaney (2005) designed a more realistic interface with icons representing travellers against a backdrop of a river with two banks and a boat. Participants selected moves by clicking on the travellers, which then appeared next to the boat on the screen. In all these instances participants were never offered a three-dimensional work surface on which objects transparently corresponding to the scenario protagonists are manipulated and moved by hand. In contrast, developmental psychologists who
worked with the river crossing task, being less sanguine about 'formal operations' presumably, have taken care to design rich interactive thinking environments with physical materials representing the boat, the river, and figurines corresponding to the cover story characters (e.g., Gholson, Dattel, Morgan, \& Aymard, 1989).

A more explicit experimental focus on interactivity may unveil interesting aspects of problem solving performance. For example, there is evidence that in other transformation problems interactivity substantially transformed problem solving behaviour. ValléeTourangeau, Euden and Hearn (2011) reported that mental set is significantly reduced in Luchins's volume measurement problems when participants interact with an actual physical presentation of the problem. The manipulation of water jars created a dynamic problem representation revealing solutions that were not simulated mentally. The selection of moves was guided and governed by three-dimensional perceptual feedback and participants were less likely to persevere using a more complicated solution for the test problems. In a rivercrossing task, interactivity may help participants work out the quality of different moves not by simulating their consequences mentally, but rather by simply completing the move and observing the results. Such moves then are 'epistemic actions' (Kirsh \& Maglio, 1994)—moves that may not, in themselves, necessarily help narrow the gap with the goal state, but rather provide information as to what to do next. As such, move selection can be opportunistic, although not necessarily mindless; rather the strategic consequences of a certain move can simply be observed. Kirsh and Maglio (1994) demonstrated that it is faster and easier to physically rotate the tetrominoes in Tetris than to simulate their rotation mentally, leading to better and more efficient problem solving behaviour. In a similar vein, moves in the world, rather than moves in the head, may help participants solve river-crossing problems more efficiently as the reduced cognitive costs of physical moves will enable them to select more moves more quickly, than they would if they were completing the task with a non-interactive problem presentation.

## The Present Experiment

The present experiment examined performance in the river crossing problem when presented with or without artefacts as an aid to solution. This was measured in terms of number of moves, latency to completion and latency per move. In a high interactivity condition, the problem was presented with a board, a raft and six figurines: Participants had to move the raft and the figurines across the board to register a move until they had moved all six figurines from one bank to the other. In a low interactivity version, the problem was described on a piece of paper and participants were asked to verbalise the moves they would make to reach the goal. They performed the problem twice, once with the high interactivity version
and once with the low interactivity version; the order was counterbalanced across participants. This experiment employed a mixed design with interactivity level as the repeated measures factor and order-low interactivity first, high interactivity first-as the between subjects factor. As moves can act as epistemic actions, we predicted that participants would produce more moves, would solve the problem more quickly and that hence latency per move would be shorter in the high compared to the low interactivity condition. We also predicted that participants would complete the second presentation of the task more quickly than the first since they would be familiar with the procedure and may well exploit an episodic record of their trajectory to help them select better moves, and select them more quickly. A high interactivity problem solving environment may more clearly showcase evidence of learning because of the ease and speed with which moves can be made in the world.

## Method

## Participants

Sixty-four university undergraduates participated in the experiment in return for course credits. Due to testing errors the data from three participants were incomplete, therefore unsuitable for analysis. Of the remaining sixtyone participants, nine did not complete the river crossing problem and were excluded from further analyses. The final sample was composed of 52 participants ( 45 females, 7 males, $M_{\text {age }}=21.4, S D=5.1$ )

## Procedure

Chickens and wolves were the protagonists in the rivercrossing scenario used for this experiment. The objective was for the six animals to be transported from the left riverbank to the right one. The selection of a move had to comply with the constraints and rules of the problem. The same instruction sheet explaining the objective of the task and the rules of the problem was used for both conditions and could be read by the participants throughout the duration of the task. The sheet read:

Three wolves and three chickens on the left bank of a river seek to cross the river to the right bank. They have a boat which can carry only two animals at a time, but there must always be an animal on the boat for it to move.
However if at any time the wolves outnumber the chickens on either bank the wolves will eat the chickens. Thus you cannot move the animal(s) in a manner that will result in the wolves outnumbering the chickens on either bank.
The goal of the task is to move all the animals from the left bank to the right bank.

In the low interactivity version of the task, the researcher transcribed each move as verbalised by the participant onto a record sheet. The record sheet was a simple representation of the raft between the left and right banks of the river, with slots to record the nature and number of the animals on either side (which was denoted with a 'C' for chickens and 'W' for wolves; see Fig. 1); each page represented only one move. At any one time, participants could only inspect their previous move as they dictated their next move to the experimenter. As soon as the next move was dictated, the sheet with the previous move was turned over. Thus participants could not inspect a historical record of previous moves. Illegal moves proposed were noted, but the experimenter did not transcribe the move on the recording sheet. Rather, participants were invited to re-read the task instructions to discover why such a move was not allowed.


Figure 1: Record sheet for the river crossing moves in the low interactivity condition.

The high interactivity version of the task involved the use of six plastic figurines, three wolves $(9 \mathrm{~cm} \times 7 \mathrm{~cm} \times$ 2 cm ) and three chickens ( $4 \mathrm{~cm} \times 5 \mathrm{~cm} \times 1.5 \mathrm{~cm}$ ), one popstick raft ( $9 \mathrm{~cm} \times 6 \mathrm{~cm}$ ) and a painted board ( $60 \mathrm{~cm} \times 45 \mathrm{~cm}$ ) representing the river and banks (see Fig. 2). As the participants interacted with the artefacts, the experimenter recorded the moves, but this record was never shown to the participants. An illegal move prompted the experimenter to instruct participants to move the raft and the animals back to the previous state and, as in the static condition, they were invited to re-read the instruction sheet to determine which moves were possible.


Figure 2: Board, raft and figurines in the high interactivity condition.


Figure 3: Mean latency to completion (left panel), mean number of legal and illegal moves (middle panel), mean latency per move (right panel) as a function of order (completed first or second) in the low interactivity (light grey) and high interactivity condition (dark grey). Error bars are standard errors of the mean.

The river crossing task was embedded in a testing session during which participants completed a number of other problem solving tasks unrelated to the present experiment. In the low interactivity version, the boxes on the first record sheet were completed with three C's and three W's on the left bank. Prior to the selection of a move, the researcher would draw an arrow above the raft to represent the direction in which it was travelling. The participants were discouraged from touching or pointing to the record sheet; they could not sketch out a move using pen and paper beforehand.

In the high interactivity condition, the board was placed on a table in front of the participant with the researcher placing all animals on the bank closest to the participant and positioning the raft on the river. This ensured all participants commenced the task with the same presentation. A move was defined as completed when whichever wolf (wolves) or chicken(s) being transported for that particular move were removed from the raft onto the other bank. Illegal moves were identified before they were completed, with animals and raft returned to the previous position on the board. In both conditions participants were given 15 minutes to complete the river crossing problem.

A 20 -minute interval was designed between the two presentations of the river crossing problem during which participants completed a number of non-verbal puzzles, including finding similarities and differences between series of pictures, and identifying the odd picture in a series of thematically related pictures. Finally, the river crossing task was presented again in the alternate condition (either low or high interactivity) to that which was presented first; the order was counterbalanced across participants. Thus, the independent variables manipulated were condition (static, interactive) and order (static first, interactive first) in a $2 \times 2$ mixed design. Performance in
both conditions was measured in terms of latency to solution, number of legal and illegal moves, and latency per move.

## Results

## Latency

Latencies to solution, displayed in the left panel of Figure 3, suggest that order had little effect on participants in the low interactivity condition but the problem was completed much quicker in the high interactivity condition when it was experienced second. A $2 \times 2$ mixed analysis of variance (ANOVA) revealed that the main effect of interactivity condition was not significant, $F(1,49)=2.14, p=.150$, while the main effect of order was significant $F(1,49)=$ 4.20, $p=.046$, as well as the condition by order interaction $F(1,49)=5.32, p=.025$. Post hoc tests indicated that latencies in the low interactivity condition did not decrease significantly from the first to the second presentation, $t(49)$ $=0.090, p=.929$. In turn, participants were quicker in the second than in the first presentation of the problem in the high interactivity condition, $t(49)=3.744, p<.001$.

## Moves

The mean number of legal and illegal moves are plotted in the middle panel of Figure 3. The high interactivity condition elicited a higher number of legal moves to solve the river crossing problem compared to the low interactivity condition and this was observed for both orders. In a $2 \times 2$ ANOVA the main effect of condition was significant $F(1,49)=11.63, p=.001$, while the main effect of order was not significant, $F<1$, nor was the condition by order interaction, $F(1,49)=1.26, p=.267$.

In turn, the mean number of illegal moves was greater in the high interactivity condition when it was experienced
first, but the frequency of illegal moves was relatively stable in the low interactivity condition across both presentations. In a 2X2 ANOVA the main effect of condition was significant, $F(1,49)=7.16, p=.010$, while the main effect of order was not significant, $F(1,49)=$ 3.34, $p=.074$ nor was the condition by order interaction, $F(1,49)=2.69, p=.108$.

## Latency per Move

The latency per move data are shown in the right panel of Figure 3. Latency per move in the low interactivity condition was unaffected by order, however participants appeared faster at enacting moves in the high interactivity condition, especially the second time the participants engaged with the task. In a $2 \times 2$ mixed ANOVA the main effect of condition was significant, $F(1,49)=20.0, p<$ .001 , but the main effect of order was not, $F(1,49)=2.33$, $p=.133$; the condition by order interaction was significant $F(1,49)=11.4, p<.001$. Post hoc tests revealed that the mean latency per move in the low interactivity condition did not decrease significantly from the first to the second presentation, $t(49)=0.858, p=.395$; in turn moves were selected faster in the high interactivity condition when that condition was experienced second, $t(49)=4.60, p<.001$.

## Discussion

This experiment investigated the impact of interactivity on problem solving performance for a river crossing problem. All participants were required to solve the problem twice, once in a low interactivity context in which move selection could only be simulated mentally and once in a high interactivity context where moves could be implemented in the world with a three-dimensional manipulable presentation of the problem. The repeated measures design eliminated random variance arising from between-subjects differences: Any performance improvement emerging in the high interactivity condition could not be attributed to a different group of participants with a differing pool of internal resources.

A high level of interactivity encouraged participants to make more moves in reaching a solution than when they completed the problem in the low interactivity condition; however, the order in which participants completed the task had no effect on the number of moves. In turn, the order in which the conditions was experienced had an effect on latency to solution. More important still, the main effect of order was qualified by a significant interaction: solution latencies in the low interactivity condition were similar whether this was completed first or second, while latencies dropped substantially when the high interactivity condition was experienced second. The latency per move data indicated that participants were always quicker to select a move in the high interactivity condition, and were generally quicker to select a move during the second presentation of the problem. However, the more important
pattern in these data was the condition by order interaction: Latency per move dropped precipitously when the second presentation of the problem occurred in the interactive condition.

As Kirsh (2010, p. 442) puts it: "Cognitive processes flow to wherever it is cheaper to perform them". Moves were cheap in the high interactivity condition - it is easier to move the pieces in the world than to simulate their movement in the head. More moves were made when the participants were given the freedom to transport the artefacts around the board to reach the solution than when moves were simulated mentally.

## Learning Manifest Through Interactivity

The second presentation of the problem offered the opportunity to gauge the degree of learning and transfer. There was much evidence of learning, when the second opportunity to solve the problem took place in a context that favoured a high level of interactivity: Participants completed the problem in less time and selected moves at a faster rate than when the second presentation of the problem was in the low interactivity condition. In fact, when the low interactivity condition was experienced second, performance reflected little learning and transfer. This pattern of results suggests two competing explanations: (i) the process and quality of knowledge acquisition is different as a function of the level of interactivity or (ii) interactivity is a performance facilitator and a high level of interactivity more clearly showcases learning. Let's take each in turn.

First exposure to the problem without much interactivity might have fostered the acquisition of a sounder and more actionable representation of the task and appreciation of an efficient sequence of moves to solution. In contrast, experiencing the problem in a context that fosters a high level of interactivity might not be accompanied by the same investment in cognitive effort, proceeding primarily on the basis of procedural learning, which in turn might interfere with the development of an accessible and transferable conceptual representation of the problem. As a result, when the problem is encountered for the second time in a condition without much interactivity, the procedural knowledge does not facilitate performance; however, when the second presentation occurs in the high interactivity condition, performance substantially benefits from the knowledge acquired on the basis of the experience in the low interactivity condition.

Alternatively, the substantial improvement in the high interactivity condition when participants are presented the problem a second time might not reflect differences in the type and quality of experience but rather release from a performance bottleneck. In other words, interactivity is a performance facilitator. Cognitive efforts and task demands are more exacting with low interactivity-as evidenced by the significantly longer latency per move. When participants encounter the problem a second time
but this time can indulge in cheaper move selection by moving artefacts on the board, they experience a release from the cognitive demands of the low interactivity condition and are quicker at producing moves.

The moves data offer some support for the performance facilitating interpretation. The number of legal moves to completion increased from an average of 15.8 when the high interactivity condition was experienced first to 17.2 when it was experienced second. And while this was not a significant increase, the pattern suggests that participants did not acquire an appreciation of a more efficient path to solution-which would lead to the selection of fewer moves - on the basis of their experience with the low interactivity condition. The release from the cognitively demanding experience with the low interactivity condition coupled with familiarity with the problem lead participants to select more moves, and interactivity enabled them to do so quickly. Moves provide information, and as participants produced more moves, they were able to reach the goal state faster.

A higher level of interactivity led to improved performance in the river-crossing problem, when preceded with the experience of solving the problem in a context that did not afford the physical manipulation of the problem presentation. Learning from previous experience with the problem, coupled with the reduction in the mental cost of making moves through interactivity provided the solver with the freedom to experiment with more moves. Through the interaction with artefacts, individuals were provided with the opportunity to extend the process of thinking beyond the mind and into the physical world. These data underscore the importance of pursuing a program of research that explicitly contrasts performance when participants can manipulate a physical problem presentation and when they cannot. In addition, we would argue that such research efforts offer a more representative window onto problem solving behavior observed outside the psychologist's laboratory.

## Acknowledgments

We would like to thank Natalie Dorman and K'Dee Bernard for their assistance with the recruitment and running of the participants, and Chris Askew for helpful suggestions.

## References

Clark, A. (1989). Microcognition: Philosophy, cognitive science, and parallel distributed processing. Cambridge, MA: MIT Press.
Gholson, B., Dattel, A. R., Morgan, D., \& Eymard, L. A. (1989). Problem solving, recall, and mapping relations, in isomorphic transfer and nonisomorphic transfer among preschoolers and elementary school children. Child Development, 60, 1172-1187.

Greeno, J. G. (1974). Hobbits and orcs: Acquisition of a sequential concept. Cognitive Psychology, 6, 270-292.
Jeffries, R., Polson, P. G., Razran, L., \& Atwood, M. E. (1977). A process model for Missionaries-Cannibals and other river-crossing problems. Cognitive Psychology, 9, 412-440.
Kirsh, D. (1995). The intelligent use of space. Artificial Intelligence, 73, 31-68.
Kirsh, D. (1996). Adapting the environment instead of oneself. Adaptive Behavior, 4, 415-452.
Kirsh, D. (2009). Problem solving and situated cognition. In P. Robbins \& M. Aydede (Eds.), The Cambridge handbook of situated cognition (pp. 264-306). Cambridge: Cambridge University Press.
Kirsh, D. (2010). Thinking with external representations. AI \& Society, 25, 441-454.
Kirsh, D., \& Maglio, P. (1994). On distinguishing epistemic from pragmatic actions. Cognitive Science, 18. 513-549.

Knowles, M. E., \& Delaney, P. F. (2005). Lasting reductions in illegal moves following an increase in their cost: Evidence from river crossing problems. Journal of Experimental Psychology: Learning, Memory, and Cognition, 31, 670-682.
Magnani, L. (2007). Semiotic brains and artificial minds: How brains make up material cognitive systems. In R. Gudwin \& J. Queiroz (Eds.), Semiotics and intelligent systems development (pp. 1 - 41). Hershey, PA: Idea Group Inc.
Reed, S. K., Ernst, G. W., \& Banerji, R. (1974). The role of analogy in transfer between similar problem states. Cognitive Psychology, 6, 436-450.
Simon, H. A., \& Reed, S. K. (1976). Modelling strategy shifts in a problem-solving task. Cognitive Psychology, 8, 86-97.
Thomas, J. C., Jr. (1974). An analysis of behavior in the Hobbits-Orcs problem. Cognitive Psychology, 6, 257269.

Vallée-Tourangeau, F. (in press). Interactivity, efficiency, and individual differences in mental arithmetic. Experimental Psychology
Vallée-Tourangeau, F., Euden, G., \& Hearn, V. (2011). Einstellung defused: Interactivity and mental set. Quarterly Journal of Experimental Psychology, 64, 1889-1895.
Villejoubert G, \& Vallée-Tourangeau, F. (2011) Constructing preferences in the physical world: A distributed-cognition perspective on preferences and risky choices. Frontiers in Psychology, 2, 302.
Watson, J. D. (1968). The double helix. London: Penguin.
Weller, A., Villejoubert, G., \& Vallée-Tourangeau, F. (2011). Interactive insight problem solving. Thinking \& Reasoning, 17, 429-439.

# A Dynamical Model of Risky Choice 

Marieke M. J. W. van Rooij (vanroomm@mail.uc.edu)<br>Luis H. Favela (favelalh@mail.uc.edu)<br>MaryLauren Malone (malonemo@mail.uc.edu)<br>Michael J. Richardson (richamo@ucmail.uc.edu)

Center for Cognition, Action, and Perception<br>Department of Psychology, University of Cincinnati<br>Cincinnati, OH 45221, USA.


#### Abstract

Individuals make decisions under uncertainty every day based on incomplete information concerning the potential outcome of the choice or chance levels. The choices individuals make often deviate from the rational or mathematically objective solution. Accordingly, the dynamics of human decisionmaking are difficult to capture using conventional, linear mathematical models. Here, we present data from a twochoice task with variable risk between sure loss and risky loss to illustrate how a simple nonlinear dynamical system can be employed to capture the dynamics of human decision-making under uncertainty (i.e., multi-stability, bifurcations). We test the feasibility of this model quantitatively and demonstrate how the model can account for up to $86 \%$ of the observed choice behavior. The implications of using dynamical models for explaining the nonlinear complexities of human decisionmaking are discussed, as well as the degree to which nonlinear dynamical systems theory might offer an alternative framework for understanding human decision-making processes.


Keywords: Decision-making; Complex Systems; Dynamical Systems Modeling; Risky Choice; Multi-stability; Phase Transitions.

## Introduction

Decision-making is part of almost everything humans do. Decisions can be commonplace or trivial but can also have lifelong consequences. Therefore, it is important to understand how individuals make decisions and how various factors play a role in decision-making processes. One such factor is uncertainty, which occurs in situations where there is limited information, ambiguous information, or unreliable information. Another factor is risk, which is different from uncertainty and can be defined as 'probabilized' uncertainty (Etner, Jeleva, \& Tallon, 2010).

Johnson and Busemeyer (2010) distinguish three major streams of development in decision theory: normative research, descriptive research, and the computational approach. While the normative approach defines what would be the optimal decision in a given situation, descriptive research describes how humans actually decide. For example, this approach has lead to the insight that individuals are sensitive to framing. When a decision is framed in terms of potential loss, the majority of participants avoid taking risk, but when the same decision is framed in terms of potential gain, the majority of
participants do take risk (Tversky \& Kahneman, 1974). In another study, Kahneman and Tversky (1979; 1983) showed that risks with low probabilities are either grossly overweighed, or completely neglected, and that there is large heterogeneity among individuals. Specifically, individuals show more variability in deciding about potential loss than potential gain (Tversky \& Kahneman, 1981). These examples suggest that human decision-making behavior under uncertainty can well be described using a nonlinear, dynamic narrative; individual decision behavior is highly context-specific, unstable, and heterogeneous.

The aim of this article is therefore to investigate the feasibility of extending current efforts in decision science towards a nonlinear, dynamical approach.

## Decision-Making and Multi-Stability

Heterogeneity, multi-stability, and context-sensitivity in general, are all strong indications that decision-making under uncertainty is characterized by nonlinear dynamics. A multi-stable system can, for the same input, settle in more than one possible internal stable state. A possible consequence of multi-stability is hysteresis, which is the phenomenon whereby a system's immediate history influences the current state of the system. Sir James Alfred Ewing first coined the term hysteresis while observing the phenomenon in magnetic materials (Ewing, 1881).

Figure 1A displays hysteresis in the magnetization and demagnetization of a magnet as a result of varying strength of the magnetic force. Depending on the direction of change of the magnetic field, the change from magnetization in one direction to the opposite direction occurs at a different moment. The system has a primitive form of memory, and remains in an existing stable state longer than expected. The opposite of hysteresis, reversed hysteresis, can also occur in multi-stable systems. Rather than remaining in the existing stable state longer (as with hysteresis), the system changes to another stable state sooner.

Hysteresis and reversed hysteresis are important indications of nonlinearity (Kelso, 1995). Hysteresis in behavioral dynamics has been found in body-scaled transitions like grasping of objects (Richardson, Marsh, \& Baron, 2007; Lopresti-Goodman, Turvey, \& Frank, 2011), speech categorization (Tuller, Case, Ding, \& Kelso, 1994), perception of whether a slanted surface supports upright
standing (Fitzpatrick, Carello, Schmidt, \& Corey, 1994), and problem-solving (Stephen, Boncoddo, Magnuson, \& Dixon, 2009).


Figure 1: Hysteresis in magnets (A) and risky choice (B). A) A magnet is magnetized by a magnetizing force $H$, into direction $B$ (state I ). If the strength of $H$ is then slowly decreased, the saturation of the magnet will change until it becomes fully magnetized into the opposite direction $-B$
(state II). If $H$ is increased again, the change towards saturation in the positive direction $B$ happens at a different value for the strength of the magnetic force $H$. B) See text.

In order to test for hysteresis and reversed hysteresis in decision-making, we will adopt a standard model of risky decision behavior with the implicit assumption that realworld decisions under uncertainty have the same properties as a monetary gamble (Hertwig \& Erev, 2009). Figure 2 displays a typical example of the type of monetary gamble researchers use to study risky decision behavior; the choice between a sure option A, and a risky option B (Kahneman et al., 1981). Choice A and B have the same expected values, thus from a rational choice perspective, they are equivalent.

## Choose between:

A. a sure loss of $\$ 750$
B. $75 \%$ chance to lose $\$ 1000$, and $25 \%$ change to lose nothing

Figure 2: Example of a risky choice.

This kind of gamble, hereafter called risky choice, can be formulated in terms of potential loss (as in the example above) or in terms of potential gain. For the remainder of this article, we will focus on loss, as potential loss is expected to maximize the variability among participants. The parameters in a risky choice are the probability to lose $P$, and the values of $R$ and $S$. The outcome is either a riskseeking choice for $R$ or a risk-avoiding choice for $S$.

Finding hysteresis or reversed hysteresis in risky choice behavior will provide evidence that decision-making under uncertainty is indeed characterized by nonlinear dynamics.

## Sequential Risky Choice

Two key components to finding hysteresis or reversed hysteresis in risky choice are to (1) change the context in two opposite directions, and (2) do this in a systematic way. It is necessary to find an input parameter for which, at different values, the system's output can have opposite, or at least, qualitatively different values. In risky choice, the key parameter that drives the choice between risk-seeking and risk-avoiding behavior is the amount of risk that is present in $R$. There are several ways to vary the amount of risk in $R$; we have opted to manipulate the value of the risky loss (in $\$$, a high value of $R$ corresponds with a high risk). Only when the value of $R$ is first increased and then decreased or vice versa, there will be an opportunity to observe hysteresis and/or reversed hysteresis. A sequential risky choice task is therefore a sequence of consecutive risky choices between $S$ and $R^{1}$, in which the value of $R$ is either increased or decreased in a step-wise fashion.

In a sequential choice task, hysteresis looks like this: A decision-maker is presented with a risky choice where the risk in $R$ is minimal (relative to $S$ ), and chooses $R$. Next, the decision-maker is presented with a second risky choice, in which the risk in $R$ is slightly higher. Next, another risky choice occurs that is even riskier, and so on. All the while the decision-maker continues choosing $R$. Then, at some switch-point (see definition below), when the risk in $R$ has become too high, the decision-maker will switch to choosing $S$ and continue to do so until the risk in $R$ is maximal (relative to $S$ ). Then, the whole process is reversed by decreasing the risk in $R$ again, causing the decisionmaker to switch back from choosing $S$ to choosing $R$ at another switch-point. If the second switch occurs for a lower risk in $R$ than the first, we have found an indication of hysteresis. If the second switch occurs for a higher risk in $R$ than the first, we have found an indication of reversed hysteresis (see also figure 1B).

## Method

Participants and Design Thirty-six undergraduate students from the University of Cincinnati were presented with three

[^239]sets of sequential risky choices between a risky loss $R$ and a sure loss $S$. In the first and third set, the amount of risk in $R$ was systematically varied, either in increasing, and then decreasing order (ID), or vice versa ( $D I$ ). The second set contained the same choices in randomized order to mediate carry-over effects between the first and third sets. Half of the students were presented first with the $I D$ set, followed by the random set and the $D I$ set. The other half started with the $D I$ set. The value of $R$ ranged from $\$ 1500$ to $\$ 525$, with increments of $\$ 25$. The probability to lose this amount $P=$ $75 \%$, and $S=\$ 750$. The total amount of choices was 238 . After completion of the sequential risky choice task, the students participated in a short money-free version of the balloon analogue risk task (BART), (Lejuez et al., 2002).

Stimulus/Apparatus All stimuli were variations of the example in Figure 2, and contained the values for $P, R$, and $S$. In total, 40 different values of $R$ (ranging from $\$ 525$ to $\$ 1500$ with increments of $\$ 25$ ) were presented either on the left side of the screen, with the value of $S$ on the right, or vice versa. The stimuli were presented on an iMac, and a cordless computer mouse (Apple Inc. ${ }^{2}$ ) was used to select the choices, both were run using PsychToolbox software (Brainard, 1997). The BART stimuli were presented on a different computer monitor ( $\mathrm{Dell}^{\mathrm{TM}}$ ) and responses made using a standard computer mouse (Logitech ${ }^{\mathrm{TM}}$ ) were recorded using BART software made available online.

Procedure Participants provided their written consent and received instructions about the sequential risky choice task. Participants were seated in front of the computer screen that displayed the various choices and were instructed to indicate their choice preferences using the mouse. After completion of the sequential choice task, participants received instruction about the BART. They again sat in front of a computer screen on which the stimuli were displayed and were instructed to respond using the mouse.

## Results

Choice outcomes of one-fourth ( $22 \%$ ) of the participants showed no change at all. This is consistent with an earlier experiment with a smaller range of risk in $R$ (from \$725$\$ 1175$ ), in which $27 \%$ of the participants showed no change.


Figure 3: Model changes between choices for $R$ and $S$. Critical change is defined as the situation where a participant switches from $S(R)$ to $R(S)$ for the same amount

[^240]of risk in the first and second half of an $I D$ or $D I$ sequence. Hysteresis is defined as the situation where a participant switches from $S(R)$ to $R(S)$ later in the second half on an $I D$ or $D I$ sequence. Reversed hysteresis is defined as the situation where a participant switches from $S(R)$ to $R(S)$ earlier in the first half on an $I D$ or $D I$ sequence.

The remaining 28 participants switched between riskseeking and risk-averse choices at least once per sequence ( $M=3.8$ fluctuations ${ }^{3}, S D=3.4$ ). Using an automated search algorithm, two switch-points ${ }^{4}$ per $I D$ and $D I$ sequence were determined for each participant. Based on the locations of the switch-points, most participants (48\%) showed critical change, followed by reversed hysteresis (39\%), and hysteresis (13\%), see Figure 3 for details. The average value of the risk in $R$ for switches from $R$ to $S$ was $\$ 1000(S D=\$ 215)$, and from $S$ to $R, \$ 941(S D=\$ 174)$ indicating that overall, participants were risk-averse ( $p<$ $0.0001)$. The distance between the two switch-points for the $D I$ and $I D$ sequences was significantly larger compared to the random sequences $t(27)=3.61, p=.001, d=.95$.

## Switching under time-constraint

$22-27 \%$ of participants in a sequential risky choice task do not show any change at all. A closer look revealed that all of these participants were presented with the $D I$ sequence first, and consistently chose $R$. One explanation could be that for about one-fourth of participants, the attractor for $S$ is nonexistent. Another explanation is that the initial conditions strengthen the attractor for $R$ relative to $S$ such that the changing constraints provide too little perturbation to the system. A small follow-up study $(N=16)$ was therefore conducted with the only difference being that participants were instructed to decide as quickly as possible while still using the available information on the screen. It was hypothesized that this speed manipulation would destabilize the initial strength of the attractor for $R$. All 16 participants switched at least once between $S$ and $R(M=10.8$ fluctuations, $S D=12.5$ ), and the relation between the speed manipulation and the absence of 'no change' participants is significant, $\chi(1, N=52)=4.20, p=.04$. The speed manipulation increased variability and caused participants to be more sensitive to changing risk constraints. This is consistent with observations that time pressure influences decision-makers' strategy selection (see Edland \& Svenson, 1993 for a review).

[^241]
## Varying increments of $\mathbf{R}$

Increasing the value of $R$ in increments of $\$ 25$ results in a high predictability of the choices in the $D I$ and $I D$ sequences. This could have mediated the amount of reversed hysteresis in our sample. A follow-up study was therefore conducted in which the increments were sampled from an $N(25,1), N(25,2), N(25,4), N(25,8)$, and $N(25,26)$ distribution respectively. The maximum and minimum values of $R$ ( $\$ 525$ and $\$ 1500$ ) were maintained. Figure 4 shows the distribution of types of choice behavior for the fixed increments ( $N=36$ ), and varying increments ( $N=50$; 10 each).


Figure 4: Distribution of types of choice behavior for varying increments of the value of $R$.

There is a main effect of sequence type (ID or $D I ; p<.001$ ), and order ( $D I$ or $I D$ first, $\mathrm{p}<.001$ ) on the difference between the two switch-points, but not of the amount of variability. However, the distribution of the four types of change behavior did differ by the amount of variation in the increments of $R, \chi^{2}(12, N=171)=28.09, p<.01$, with a positive trend for the amount of participants that showed hysteresis and reversed hysteresis.

## Nonlinear Dynamical Modeling

Multi-stability in switching behavior is problematic for most linear models but can be accounted for by a nonlinear dynamical system (e.g. Cho, Jones, Braver, Holmes, \& Cohen, 2002; Roxin \& Ledberg, 2008). A dynamical system is a mathematical concept where the time dependence of a state variable (a variable that describes a certain quantity of a system that we are interested in, like position or concentration) is described using a fixed rule. In a nonlinear dynamical system, this fixed rule is nonlinear, and the system therefore does not satisfy the additivity and homogeneity properties that are necessary for linearity.

Examples of applications of (nonlinear) dynamical modeling to human behavior are vision (for example Fürstenau, 2006), speech (Kelso, Saltzman, \& Tuller, 1986; Tuller et al., 1994), language (for example Spivey, Grosjean, \& Knoblich, 2005), motor and neural dynamics (Haken, Kelso, \& Bunz, 1985; Schöner \& Kelso, 1988, Kelso, et al., 1992), and cognition (Bressler \& Kelso, 2001). Applications of dynamical models to decision-making under uncertainty have focused on either micro-level or macro-
level behavioral observations. For example Brown \& Holmes (2001) modeled a simple choice task using a dynamical model of firing rates of neurons. On a macrolevel, we find examples of dynamical models of multi-agent decision-making processes (for a brief overview, see Lu, Chen \& Yu, 2011).

## A One-Dimensional Model of Multi-Stability and Hysteresis in Risky Choice

To model the observed switching between $R$ and $S$, we propose a nonlinear dynamical system that has previously been applied to other cases in which individuals switched between two different behaviors, and where nonlinear phenomena like hysteresis and reversed hysteresis informed the use of a nonlinear dynamical model (e.g., Tuller et al., 1994). Equation 1 gives the potential function of the onedimensional model.

$$
\begin{equation*}
V(x)=k x-\frac{x^{2}}{2}-\frac{x^{4}}{4}+\xi \tag{1}
\end{equation*}
$$

where $x$ is the observed choice, $k$ the control parameter, and a noise term $\xi$ is added to each choice.
A potential function is the integral of the differential equation describing the evolution of the state variable $x$ (in our case, the observed choice), which means that a minimum or maximum of the potential function corresponds to a stable state of the system. Our system's potential function therefore reveals the attractor and repeller states, to which the system is attracted to or repelled from (see Kelso, 1995 and Strogatz, 2000 for more background on dynamical systems). The behavior of our dynamical system is driven by a control parameter $k$.


Figure 5: Potential landscape for five different values of $k$. Depending on the direction of change, a phase transition occurs between the two possible attractors for a critical value of $k, \pm k_{c}$.

Figure 5 shows some examples of the shape of the potential function, or attractor landscape, for different values of $k$. For a critical value of $k$, a bifurcation occurs (for both $k=k_{c}$, and $k=-k_{c}$ ), causing a phase transition between risk-seeking and risk-avoiding choices or vice versa. A phase transition occurs for a different value of $k$, depending on the direction of change, which explains hysteresis. By defining the two attractor states as the choice for $R$ and $S$ respectively, this model thus explains switches
between risk-seeking and risk-avoiding choices, as well as multi-stability through hysteresis (although not reversed hysteresis; see below for a more detailed discussion of reversed hysteresis).

Parameter Selection and Optimization The potential function offers a way to simulate sequential choice data. The key to modeling the risky choice phenomena is the control parameter $k$, which has to reflect the changing risk in $R$. We propose $k$ as a simple linear function of the risk in $R$ at choice $j$ and a baseline individual value, $k_{0}$, such that

$$
\begin{equation*}
k_{j}=k_{0}-R_{j}, \tag{2}
\end{equation*}
$$

By sampling $k_{0}$ from a uniform distribution spanning all possible values of $k$ between two extremes, and using Eq.
(1) and (2), we simulated an entire range of possible choice data. The lower boundary for $k_{0}$ corresponds to the case where only the attractor for $S$ exists, regardless of the value of the risk in $R$, and the upper boundary corresponds with only one attractor for $S$. Using a bootstrapped optimization with respect to the difference between the simulated and empirical choices on the $D I$ and $I D$ sequences of our main experiment (no variability in step-size, no speed manipulation), we were able simulate $86 \%$ of the observed choices. The differences in switch-points for reversed hysteresis are relatively small compared to the total range of values for $R(M=\$ 170.45, S D=\$ 183.08)$. This explains that, although the model does not account for reversed hysteresis, it generates a high proportion of correct choices.

Individual Risk Sensitivity A frequently reported result in research on decision-making under uncertainty is that people have relatively static personality characteristics that determine their risk-taking behavior (e.g. Mishra \& Lalumière, 2011). Accordingly, we hypothesize that $k_{0}$, the individual baseline value of the control parameter $k$ reflects risk sensitivity or propensity, and should therefore closely relate to participants' BART scores. Correlation between the participants' BART scores and the optimal values of $k_{0}$ however is very low, $r(33)=-.15, p=.36$ for the $I D$ sequences, and $r(33)=-.04, p=.81$ for the $D I$ sequences.

Modeling Reversed Hysteresis The current model does not account for reversed hysteresis, while up to $42 \%$ of participants show reversed hysteresis in their choice behavior. Lopresti-Goodman, Turvey, \& Frank (2012) provide a way to extend nonlinear dynamical models that includes reversed hysteresis using an auto-regulated control parameter. Negative auto-regulation forces the dynamical system to remain close to the bifurcation line and may reflect habituation to the amount of risk presented in the choices; rendering the choice for $S$ or $R$ unstable. This would also explain why the amount of hysteresis relative to reversed hysteresis increases with more variability in the increments of $R$ (Figure 4). Larger variability interferes with
the habituation and diminishes the effect of negative autoregulation.

## Discussion

There are many models of risky choice (see Glöckner \& Pachur, 2012 for a review). However, in order to account for multi-stability, nonlinearity is a necessary assumption. The results presented here show multi-stability in risky choice, for which we have provided a basic nonlinear dynamical model. The model provides a way to explain decisionmaking under uncertainty within the framework of complexity theory; a relative newcomer to the social sciences that offers a promising new perspective on human cognition (Van Orden, Holden, \& Turvey, 2003). Although the current model does not explain reversed hysteresis, it does provide a blueprint for a nonlinear dynamical model that can capture the entire range of observed choice behavior.

The aim of modeling was to provide a formal description of the observed decision-making behavior. Moreover, our hope is that identifying the right kind of nonlinear models will eventually lead to insights into the underlying processes or mechanisms. One of the strengths of the model is that multi-stability is an inherent behavior of the nonlinear dynamical system, pre-empting the need for weight functions or exceptions. The model also provides a starting point for theorizing about the psychological processes underlying the behavior. The control parameter is a single parameter that captures the switching between risk-seeking and risk-averse choices. Unexpectedly, however, there was no correlation between participants' BART scores and the baseline value of the control parameter, $k_{0}$. Upon reflection, this result is not as surprising after all. Nonlinear dynamical systems are especially useful in capturing change and the phenomena that are associated with change, like hysteresis. The BART however assumes individual risk preference is a temporarily static personality trait. The current results therefore indicate that risk preference is a highly complex and multi-dimensional construct and that the dynamics of subsequent risky choice behavior cannot be captured in a single measure of risk sensitivity.

## Acknowledgements

We acknowledge financial support from the National Science Foundation Grant BCS-0843133 to John G. Holden and Guy Van Orden.

## References

Brainard, D. H. (1997). The Psychophysics Toolbox. Spatial Vision 10, 433-436.
Bressler S. L., \& Kelso J. A. S. 2001. Cortical coordination dynamics and cognition. Trends. Cogn. Sci. 5, 26-36.
Brown E., \& Holmes, P. (2001). Modeling a simple choice task: stochastic dynamics of mutually inhibitory neural groups. Stochastics Dynamics 1, 159-191.

Cho, R.Y., L.E. Jones, T.S. Braver, P.J. Holmes, \& Cohen, J.D. (2002). Mechanisms underlying dependencies of performance on stimulus history in a two-alternative forced-choice task. Cognitive, Affective, \& Behavioral Neuroscience, 2(4), 283-299.
Edland, A., \& Svenson, O. (1993). Judgment and decision making under time pressure: Studies and findings. In O. Svenson \& A. J. Maule (Eds.), Time pressure and stress in human judgment and decision making. New York: Plenum.
Etner, J., Jeleva, M., \& Tallon, J. (2012). Decision theory under ambiguity. Journal of Economic Surveys, 26, 234270.

Ewing, J. A. (1881). Effects of Stress on the Thermoelectric Quality of Metals. Part I. Proceedings of the Royal Society of London (1854-1905) 32, 399-402.
Fitzpatrick, P., Carello, C., Schmidt, R. C., \& Corey, D. (1994). Haptic and visual perception of an affordance for upright posture. Ecological Psychology 6, 265-287.
Fürstenau, N. (2006). Modelling and simulation of spontaneous perception switching with ambiguous visual stimuli in augmented vision systems. Lect. Notes Artific. Intell. 4021, 20-31.
Glöckner, A., and Pachur, T. (2012). Cognitive models of risky choice: parameter stability and predictive accuracy of prospect theory. Cognition 123, 21-32.
Haken H., Kelso J. A. S., \& Bunz, H. 1985. A theoretical model of phase transitions in human hand movements. Biol. Cybern. 51, 347-356
Hertwig, R., \& Erev, I. (2009). The description-experience gap in risky choice. Trends in Cognitive Sciences, 13, 517-523.
Johnson, J. G., \& Busemeyer, J. R. (2010). Decision making under risk and uncertainty. Wiley Interdisciplinary Reviews: Cognitive Science, 1, 736-749.
Kahneman, D., \& Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. Econometrica 47, 263292.

Kelso, J. A. S., Saltzman, E. L. \& Tuller, B. (1986) The dynamical perspective on speech production: data and theory. Journal of Phonetics, 14, 29-59.
Kelso, J. A. S., Bressler S. L., Buchanan S., DeGuzman G. C., Ding M., Fuchs A., \& Holroyd, T. (1992). A phase transition in human brain and behavior. Phys. Lett. A 169, 134-144.
Kelso, J. A. S. (1995). Dynamic patterns: The selforganization of brain and behavior. Cambridge, MA: MIT Press.
Lejuez, C. W., Richards, J. B., Read, J. P., Kahler, C. W., Ramsey, S. E., Stuart, G. L., . . Brown, R. A. (2002). Evaluation of a behavioral measure of risk taking: The balloon analogue risk task (BART). Journal of Experimental Psychology: Applied, 8, 75-84.
Lopresti-Goodman, S. M., Turvey, M. T., Frank, T. D. (2012). Negative hysteresis in the behavioral dynamics of the affordance "graspable". Manuscript submitted for publication.

Lopresti-Goodman, S. M., Turvey M. T., \& Frank, T. D. (2011). Behavioral Dynamics of the Affordance 'Graspable'. Attention, Perception, \& Psychophysics, 73, 1948-1965.
Lu, J., Chen, G., \& Yu, X. (2011). Modelling, analysis and control of multi-agent systems: A brief overview. Paper presented at the Proceedings - IEEE International Symposium on Circuits and Systems, 2103-2106.
Mishra, S., \& Lalumière, M. L. (2011). Individual differences in risk-propensity: Associations between personality and behavioral measures of risk. Personality and Individual Differences, 50, 869-873.
Richardson, M. J., Marsh, K. L., \& Baron, R. M. (2007). Judging and actualizing intrapersonal and interpersonal affordances. Journal of experimental psychology. Human perception and performance, 33, 845-59.
Roxin, A. \& Ledberg, A. (2008). Neurobiological models of two-choice decision making can be reduced to a onedimensional nonlinear diffusion equation. PLoS Computational Biology, 4(3), 1-13.
Schöner, G., \& Kelso, J. A. S. (1988). Dynamic pattern generation in behavioral and neural systems. Science, 239, 1513-1520.
Spivey, M. J., Grosjean, M., \& Knoblich, G. (2005). Continuous attraction toward phonological competitors. Proceedings of the National Academy of Sciences, USA, 102, 10393-10398.
Stephen, D. G., Boncoddo, R. A., Magnuson, J. S., \& Dixon, J. A. (2009). The dynamics of insight: Mathematical discovery as a phase transition. Memory \& Cognition 37(8), 1132-1149.
Strogatz, S. (2000). Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering. Cambridge, MA: Westview Press.
Tuller, B., Case, P., Ding, M., \& Kelso, J. A. S. (1994). The nonlinear dynamics of speech categorization. Journal of Experimental Psychology: Human Perception and Performance, 20, 3-16.
Tversky, A., \& Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. biases in judgments reveal some heuristics of thinking under uncertainty. Science, 185, 1124-1131.
Tversky, A., \& Kahneman, D. (1981). The framing of decisions and the psychology of choice. Science, 211, 453-458.
Tversky, A., \& Kahneman, D. (1983). Extension versus intuitive reasoning: The conjunction fallacy in probability judgment. Psychological Review 90, 293-315.
Van Orden, G. C., Holden, J. G., \& Turvey, M. T. (2003). Self-Organization of Cognitive Performance. Journal of Experimental Psychology: General, 132, 331-350.

# Evidence for nonconscious behavior-copying in young children 

Johanna E. van Schaik (J.E.vanSchaik@donders.ru.nl)<br>Donders Institute for Brain, Cognition, and Behavior, Radboud University<br>Montessorilaan 3, 6525HR, Nijmegen, the Netherlands

# Rick van Baaren (R.vanBaaren@psych.ru.nl) 

Behavioral Science Institute, Radboud University
Montessorilaan 3, 6525HR, Nijmegen, the Netherlands

Harold Bekkering (H.Bekkering@donders.ru.nl)<br>Donders Institute for Brain, Cognition, and Behavior, Radboud University Montessorilaan 3, 6525HR, Nijmegen, the Netherlands

Sabine Hunnius (S.Hunnius@donders.ru.nl)<br>Donders Institute for Brain, Cognition, and Behavior, Radboud University<br>Montessorilaan 3, 6525HR, Nijmegen, the Netherlands


#### Abstract

Behavioral mimicry is the nonconscious copying of an interaction partner's behavior and is affected by social dynamics. Whereas it has been studied extensively in adults, little is known about the development of mimicry. The aims of this study were twofold, first to identify whether young children demonstrate mimicry and, second, to investigate whether young children's mimicry displays sensitivity to social dynamics. Using a video-based paradigm, 40-monthold children observed six types of behaviors (i.e. yawning, laughing, frowning, cheek-scratching, mouth-rubbing and head-wiggling) performed by a model which they had previously seen either helping or hindering another model. Results indicate that children carried out five of the six behaviors more often while watching the behavior videos than during baseline. However, no differences were found between the two social manipulations. We conclude that young children demonstrate mimicry like that reported in adults and discuss the possible causes of the absence of a social effect.


Keywords: behavioral mimicry; development; action; social dynamics; social interaction.

## Introduction

An often unnoticed component of social interactions is behavioral mimicry. Mimicry can be defined as nonconsciously adopting the behaviors of an interaction partner (van Baaren et al., 2009). In one of the first comprehensive studies of mimicry, participants were exposed to foot-shaking or face-rubbing confederates with smiles or neutral expressions on their faces. Chartrand and Bargh (1999) showed that participants were more likely to carry out the modeled behaviors and expressions than the non-modeled behaviors and expressions. Importantly, replicating these behaviors occurred outside of the participants' awareness (Chartrand \& Bargh, 1999).

In contrast to the extensive adult literature on mimicry (for a review see Chartrand \& van Baaren, 2009), exceptionally few studies have investigated the development
of mimicry. Some authors have documented neonatal imitation (e.g. Meltzoff \& Moore, 1977; Meltzoff \& Moore, 1983). Others, however, note the lack of breadth of these behaviors and have been unable to replicate original findings with older infants and young children (e.g. Anisfeld, 1996; Jones, 2007). Additionally, in such studies, infants and young children are encouraged to replicate modeled behaviors (e.g. Jones 2007), which stands in contrast to the uninstructed mimicry reported in adults. In one study that did not give replication instructions, children saw video stimuli in which someone often yawned, but children under the age of five did not demonstrate instances of yawning (Anderson \& Meno, 2003). In a live paradigm, only three out of 40 children under the age of four demonstrated contagious yawning (Helt et al., 2010). Similarly, Over and Carpenter (2009) report that, in a pilot study, 5-year-old children who interacted with an adult who repetitively touched her face failed to mimic this behavior. Notably, the authors posited that there was little evidence to suggest that children under the age of five exhibit mimicry of the sort found in adults (Over \& Carpenter, 2009).

Not only do adult studies indicate the uninstructed nature of mimicry, but they also bring to light its sensitivity to social dynamics. For example, liking one's interaction partner has been shown to increase mimicry rates, both when liking was preexistent and manipulated (Likowski et al., 2008; McIntosh, 2006). Although there is no evidence of uninstructed mimicry in young children, a form of imitation has been shown to be affected by social dynamics. Overimitation (also called affiliative imitation) is the replication of actions shown during a task demonstration that are unrelated to achieving the desired end-state of the task (Over \& Carpenter, 2012). In a conceptual replication of an adult study by Lakin, Chartrand and Arkin (2008) which showed that being socially excluded lead to higher mimicry rates, Over and Carpenter (2009) found that priming 5-year-olds with social exclusion increased overimitation rates (Over \& Carpenter, 2009), indicating
that non-mimicry forms of behavior replication are sensitive to social factors in young children.

Children's sensitivity to social dynamics is also manifest in other behavioral measures. One study showed that 3-yearolds helped helpful adults more than destructive adults (Vaish, Carpenter, \& Tomasello, 2010). Kenward and Dahl (2011) demonstrated that, when given an uneven number of biscuits, 4.5-year-olds distributed more biscuits to puppets they saw helping another puppet than to puppets they saw violently hindering the other puppet. Three-year-olds did not distinguish in their biscuit-distribution but the authors suggest this was because they were shocked by the violent nature of the events and were not sure which puppet was which (Kenward \& Dahl, 2011).

Thus far, no studies have reliably found uninstructed mimicry during early childhood, and it is hence also unknown if children's mimicry is affected by social dynamics. In the present study, we first aimed to identify whether young children demonstrate mimicry like that found in adults. Importantly, we incorporated a range of behaviors, such as facial expressions and manual behaviors, to investigate the generality of young children's mimicry. Also, as past adult studies have successfully used videos to elicit mimicry (e.g. Lakin \& Chartrand, 2003; Platek et al., 2003), we chose to present the stimuli as videos to ensure that all children saw identical behaviors. Moreover, this provided the children with a 'task', namely to watch TV, which is in line with the contention of van Baaren and colleagues (2009) that during mimicry experiments the focus should not be on the behaviors specifically. We incorporated a baseline measure so as to compare natural behavior rates with those elicited by observation within participants, because past studies indicate that individual differences influence mimicry rates (e.g. Chartrand \& Bargh, 1999; Platek et al., 2003; Sonnby-Borgström, 2002). We hypothesized that children would demonstrate the behaviors at greater frequencies while watching the behavior videos than during baseline.

The second aim was to address whether mimicry is sensitive to social dynamics at three years of age. As past studies demonstrated that children around three and four years of age show differential treatment of helpers versus hinderers (Kenward \& Dahl, 2011; Vaish et al., 2010), we used a similar paradigm to manipulate the social dynamics. We designed the models’ interactions such that the helper would come across as a nice individual whereas the hinderer would be seen as a mean but not violent individual. In this manner, we aimed to implement a similar effect as in the manipulated-liking designs of adult mimicry studies (Likowski et al., 2008; McIntosh, 2006). Due to possible carry-over effects from previous interactions (e.g. Lakin \& Chartrand, 2003), we used this social manipulation as a between-participants factor, such that half of the children were randomly assigned to the helper condition and half to the hinderer condition. We hypothesized that children would mimic helpers more than hinderers, replicating the pattern of higher mimicry rates for liked individuals in adult studies.

## Methods

## Participants

Participants were recruited through the database of volunteer families of the Baby Research Center Nijmegen. Signed consent was obtained from parents beforehand. Thirty-three children participated in this study (mean age: 39.7 months, range: 39.2-40.2; 23 girls). Seven children were excluded due to not wanting to watch the videos ( $\mathrm{N}=1$ ), technical error ( $\mathrm{N}=1$ ), and not meeting the inclusion criteria of having attended to at least $40 \%$ of the behavior videos ( $\mathrm{N}=3$ ) or having watched each behavior video at least once ( $\mathrm{N}=2$ ). Thus, the final sample consisted of 26 children (19 girls).

## Stimuli

The stimulus videos for the experiment were made using a digital video camera (Sony Handycam, DCR-SR190E) and were digitally muted. Two types of videos were recorded, social manipulation videos and behavior videos.

Figure 1.2 shows the final scene of the helper video, and gives an indication of the scene composition used in the social manipulation videos. In both the helper and hinderer videos, a stuffed animal was initially positioned in the left, front corner of the table, and the helper or hinderer (H) walked in from the left and the neutral model ( N ) from the right, each sitting down at their respective sides of the table. After N failed to reach the stuffed animal from her position, H reached over to get the stuffed animal and held it out to N who reached for it. At this point the videos differed; in the helper videos, H passed the stuffed animal to N who held it as in Figure 1.2, whereas in the hinderer videos, H pulled the stuffed animal back and held it to her chest.

Three adult female models were used. Two models were used for H (i.e. H 1 and H 2 ), who each played both the helper and the hinderer in order to control for possible idiosyncrasies of each model. The model for $H$ was kept consistent within participants, such that children who saw H 1 during the social manipulation video also saw the behavior videos of H 1 , and the same for H 2 . The H models wore a colored shirt to aid subsequent identification while N wore black. Since N never reappeared in the behavior videos, only one model played her role.

Six different behavior videos were made. The first, yawning, was selected for its contagious qualities (Figure 1.3; Platek et al., 2003). Two emotional facial expressions, laughing and frowning (i.e. a sad facial expression), were used as they have successfully elicited mimicry in adult studies (Lakin \& Chartrand, 2003; Moody \& McIntosh, 2011; Sonnby-Borgström, 2002) and recently also in schoolaged children (Deschamps et al., 2012). Two manual behaviors were loosely based on those used in interactive adult studies (e.g. Chartrand \& Bargh, 1999; Lakin \& Chartrand, 2003), namely using the fingertips to scratch the cheek (i.e. cheek-scratching) and rubbing the fingertips back and forth across sealed lips (i.e. mouth-rubbing; Figure 1.5). Finally, in the head-wiggling clip the model moved her head
from side to side while looking forwards. Each behavior video showed the model in a neutral position for the first and last 500 milliseconds. Pilot data indicated that children of this age were capable of replicating all behaviors.


Figure 1: Experimental design.

## Design

This experiment consisted of three types of stimuli: the baseline, the social manipulation and the behavior videos. For the baseline, a non-social video ( 73.7 sec .) from an unrelated experiment was shown displaying a single racecar driving through a racetrack (Figure 1.1; Immens, 2011). Next, the social manipulation video (average duration 23sec.), depending on the condition the participant was assigned to, was shown twice (Figure 1.2). The behavior videos (average duration 7 sec .) were presented after the social manipulation videos (Figure 1.3 and 1.5). Each of the six behaviors was presented five times, resulting in 30 behavior videos in total, and after every 5 behavior videos an attention grabber video ( 2 sec .) was shown. After half of the behavior videos were played, the same social manipulation video was shown a third time (Figure 1.4) and was announced via a recording of a voice saying in Dutch, "Look! Again this video." Children’s behavior during the third repetition of the social manipulation video was not included in the behavior rate calculations. Together, the baseline, the three repetitions of the social manipulation video and the 30 behavior videos lasted approximately six minutes. At the end of the experiment, the experimenter asked the children if they remembered the social manipulation video, if they could describe what had happened and whether the model was nice or mean, as well as whether the child remembered copying the model's behaviors.

Randomization and counterbalancing. The (pseudo)randomizations were done using Mix (van Casteren \& Davis, 2006). Participants were randomly assigned to one of the two conditions (i.e. helper or hinderer) and one of the two models (i.e. H1 or H2); hence there were four groups, one for every combination of condition and model. For each
group there were two presentation orders of behavior videos (i.e. eight in total), which were constrained such that at least three different behavior types had to be presented before the same behavior could be shown again, and these presentation orders were counterbalanced across participants.

## Procedure

Following a short play session, the child and parent were led to the experiment room. Children were seated in front of an eye-tracker (T120, Tobii Technology, Stockholm, Sweden, Tobii Studio software) either alone or on their parent's lap. A video camera (Sony Handycam, DCR-SR190E) was positioned to the side of the child such that it was not in her direct visual field but still obtained the most frontal recording angle possible. The only instructions given were to watch the videos. Upon conclusion, the participants were allowed to select a storybook or were given 10 Euros for participating in the experiment.

## Coding and Reliability

The children's behavior was coded using ELAN Linguistic Annotator (4.3.3, http://tla.mpi.nl/tools/tla-tools/elan, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands; Lausberg \& Sloetjes, 2009). The coder was blind to condition and the order of stimulus presentation.

Although the experiment was presented on an eye-tracker so that attention could be measured precisely, the percentage of looking time according to the output was often considerably lower than the amount of time that the child actually attended the screen (for comparable eye-tracking discrepancies, see Morgante, Zolfaghari, \& Johnson, 2012). For this reason, attention was coded by hand. If the child looked away for more than five seconds, turned to interact with the parent or experimenter, or was not clearly visible on the video, that duration was coded as not-attending.

Pilot data was used to create the coding scheme for the behaviors so as to accommodate how children carry out each behavior. If the child verbally labeled a behavior right before, during or after carrying it out, it was not coded as mimicry. Also, behaviors that started while the child was not attending were not coded as these might have been externally triggered. The exact coding scheme is available from the first author, with the required characteristics as follows. Yawns were coded when the lips were parted forming an O-shape. For laughing, the corners of the mouth needed to be turned upwards (i.e. smiles were also counted) while for frowns they needed to be turned downwards. A cheek scratch was coded if the child brought her hand to her cheek or forehead and made scratching movements with her fingers. If the child rubbed her fingers over her mouth or chin it was coded as a mouth rub. Lastly, the head-wiggle was coded when the child tilted her head to the left or right and then to the other side at least once.

To ensure coding-reliability, a random sample of 20 percent of the participant videos was re-coded. The mean intraclass correlation coefficient between behavior rates of the first and second coding was $\mathrm{r}=.98$.

## Behavioral Measures

The timing of all events (e.g. onset and offset times of stimuli and the participant's behaviors) were synchronized and rounded to the nearest 100 milliseconds. The baseline and behavior videos period were separated; the baseline consisted of the duration of the racecar animation and the behavior videos period was defined as starting when the first behavior video started and ending after the last behavior video, but with the social manipulation video in between excluded. Participant's behaviors that occurred during the behavior videos period but before the first attended behavior video of that type were excluded.

Behavior rates. Per participant, it was counted how often each behavior was carried out, and rates were calculated separately for the baseline and behavior videos period. Total behavior rates were calculated by dividing the total behavior count by the duration in minutes that the screen was attended. Similarly, behavior rates were calculated per behavior type using the count of just one behavior. For these separate behavior rates, the duration attended in minutes for the behavior videos period was adjusted to start from the beginning of the first behavior video of that behavior type, resulting in the separate behavior rates being lower than the overall behavior rate. Hence, per participant, per baseline or behavior videos period, seven behavior rates (i.e. behaviors per minute attended) were calculated: the overall rate and one rate for each of the six behavior types.

## Analysis

Several comparisons were run to check that the models and the presentation orders did not have an effect on behavior rates during the behavior videos period and were run separately for the two conditions. The helper condition consisted of 12 participants, five of whom saw the videos of model H 1 , while the hinderer condition had 14 participants, 7 of whom saw model H1. Independent-samples $t$-tests and Mann-Whitney U-tests compared the effect of model (e.g. H1 or H2) on total behavior rates and separate behavior rates, respectively, and Kruskal-Wallis H-tests compared the effect of the presentation orders on both total behavior rates and separate behavior rates. There were no effects of model or presentation orders for total or separate behavior rates in either condition (all ps >.1). Therefore, the models and presentation orders were collapsed in the subsequent analyses. Additionally, Mann-Whitney U-tests revealed no differences in behavior rates between children sitting on their parents' laps and those sitting alone on the chair during either the baseline or the behavior videos period (all $p s>.2$ ).

## Results

Out of the 26 participants, 25 participants demonstrated at least one of the six behaviors during either the baseline or the behavior videos period, and 23 participants carried out the behaviors more often while watching the behavior videos than during baseline.

Since it first needed to be investigated whether the two conditions (i.e. groups of participants) differed, the hypothesized difference between the helper and hinderer condition during the behavior videos period was tested. However, a Mann-Whitney U-test revealed no significant difference in total behavior rates between conditions ( $p>.4$ ). Hence, for the subsequent comparisons the participant groups were collapsed across conditions.

To investigate whether behavior rates differed between baseline and the behavior videos period, a paired-samples $t$ test was used to compare total behavior rates. Children carried out the behaviors significantly more often during the behavior videos period ( $M=2.38$ behaviors per minute, $S E=0.24$ ) than during the baseline ( $M=0.92$ behaviors per minute, $S E=0.33$; $t(25)=-4.3, p<.001, r=.65)$.

Subsequently, each separate behavior was investigated using Wilcoxon signed-rank tests, and alpha was corrected for multiple comparisons using a Bonferroni correction ${ }^{1}$ (Figure 2). During the behavior videos, the rates of yawning, frowning, mouth-rubbing and head-wiggling, were significantly higher than the baseline rates of yawning ( $z=3.18, r=.44$ ), frowning ( $z=2.74, r=.38$ ), mouth-rubbing ( $z=2.61, r=.36$ ) and head-wiggling ( $z=2.93, r=.41$; all $p s<.008$ ), respectively. Cheek-scratching occurred more often during the behavior videos period than during the baseline at a level of marginal significance ( $p=.011$ ). Laughing did not differ significantly between the two periods.


Figure 2: Mean behavior rates of each behavior type for the baseline and behavior videos period. Error bars indicate one standard error above the mean; ${ }^{* *} p<.008,{ }^{*} p=.011$.

For the five behaviors with significant and marginally significant effects, it was investigated post hoc whether any one behavior was more likely to be replicated than the other behaviors. A Friedman's ANOVA was used to compare the difference in behavior rates between baseline and behavior videos period (i.e. behavior videos period behavior rate

[^242]minus baseline behavior rate) between the behaviors. No differences between the behaviors were found (all ps>.7).

A Mann-Whitney U-test showed that the children's answers to the question of whether they consciously replicated the model's behaviors were not predictive of their behavior rates during the behavior videos period ( $p>.6$ ).

## Discussion and Conclusion

This study aimed to identify and investigate mimicry in 40-month-old children. We found that children carried out the behaviors significantly more often while watching the behavior videos than while watching the baseline video. This was evident across individuals, as 23 out of 26 participants showed higher behavior rates during the behavior videos period than during baseline, and across behavior types, as five of the six behaviors were mimicked. Yawning, frowning, mouth-rubbing, and head-wiggling all occurred at significantly greater rates during the behavior videos than during baseline and cheek-scratching showed this effect at a level of marginal significance. Of the mimicked behaviors, no one behavior was more likely to be mimicked than others, while controlling for baseline rates.

Mimicry of these behavior types have, to the best of our knowledge, not been tested during early childhood before, with the exception of yawning. Helt and colleagues (2010) report very low rates of yawning in live paradigms under the age of four and Anderson and Meno (2003) did not find any instances of yawning during video watching in three-yearolds. In their video-based study, children were instructed to clap whenever they saw a yawn; as also suggested by Helt and colleagues (2010), the disparity between their findings and ours may be a result of the assigned tasks, since our simple instructions to watch the videos better resemble the uninstructed nature of adult mimicry studies. Indeed, the behavior rates during the behavior videos period of our study are similar to the behavior rates measured during live interactions in adults. For example, Chartrand and Bargh (1999) found an average rate of .57 face-rubs per minute, which closely corresponds to the children's average behavior rate of .51 for mouth-rubs.

The only behavior that did not demonstrate a mimicry effect in the current study was laughing. This was likely caused by the children's enjoyment of the baseline video, as average laughing rates during the baseline far exceeded those of the other behaviors' baseline rates. Although the baseline video was selected for its neutrality and non-social nature, the animation still needed to be, and in fact was, attractive enough for children to attend to it.

An important characteristic of mimicry is that it occurs outside of the awareness of both the individual mimicking and the individual being mimicked (Chartrand \& van Baaren, 2009). Children were asked at the end of the experiment whether they copied the model while watching the behavior videos, and their answers were not related to their actual mimicry rates. Additionally, during a pilot study children were instructed to copy the behaviors, but it became apparent that they found it unusual to consciously
replicate the behaviors of a non-responsive model, even when encouraged by their parents. Furthermore, our coding scheme ensured that the few cases in which children verbally labeled a carried-out behavior, indicating that they were focusing on doing that behavior, were not counted as mimicked behaviors. Anecdotally, several parents remarked that they were surprised to see their child replicate the behaviors seemingly automatically. Altogether, there is sufficient evidence to indicate that the children nonconsciously replicated the behaviors, in line with the definition of behavioral mimicry.

This study further investigated whether children's mimicry is sensitive to social dynamics. To influence the social dynamics, a helper-hinderer manipulation was used in a between-participants design. However, no significant differences between the conditions were found. Given that past studies have linked mimicry with social perspective taking skills (e.g. Chartrand \& Bargh, 1999; Platek et al., 2003), it might be that the sensitivity of mimicry to social factors gradually develops during childhood as an effect of increasing social cognition and experience. However, it should be considered whether the social manipulation could have been ineffective. A limitation of the present study was that the social manipulation and behaviors were recorded as separate video clips with different background settings. Since Kenward and Dahl (2011) reported that their participants had difficulty later identifying the puppets, we allocated the helper and hinderer models a colored shirt to aid later identification. Nonetheless, the different setting of the two video types may have prevented children from making the link between the model in the social manipulation video and the model in the behavior videos. More support for this notion comes from recent pilot data with $51 / 2$-year-olds, which indicated that children older than those in this study often failed to relate the model in the behavior videos to the model in the social manipulation video seen before. A similar limitation was that video presentation prevented participants from actually affiliating with the model, thereby possibly preventing an affiliationdriven social effect, as suggested by Over and Carpenter (2012) regarding an overimitation study by Nielsen, Simcock and Jenkins (2008).

The findings of this study highlight avenues for further research into the neural and cognitive underpinnings of mimicry. Whereas a perception-action matching system founded in imitation research has been suggested to also underlie mimicry (Chartrand \& van Baaren, 2009), it is unclear whether neural differences exist between nonconscious mimicry and instances of conscious motor observation and replication. Additionally, cognitive mechanisms have been suggested to contribute to imitative behaviors (e.g. Meltzoff, 2007; Woodward et al., 2009), and future studies should investigate whether similar mechanisms, and the development thereof, are involved in mimicry's reported social sensitivity.

In conclusion, this study is the first to identify uninstructed behavioral mimicry in 40-month-old children.

The spectrum of behaviors for which this was the case reflects the repertoire of mimicked behaviors in the adult literature (Chartrand \& van Baaren, 2009), and provides a basis for future research investigating the underlying neural and cognitive processes. It is unclear whether the lack of social modulation of mimicry was a result of experimental design or an effect of social-cognitive development, and this posits further investigation.

## Acknowledgments

We would like to thank the volunteer families of the Baby Research Center, Nijmegen; Angela Khadar; Loes Immens; Claudia Lüttke, Lianne Scholtens, and Ricarda Braukmann; and Claire Monroy, Sarah Gerson, and the anonymous reviewers for valuable comments on the manuscript.

## References

Anderson, J. R. \& Meno, P. M. (2003). Psychological influences on yawning in Children. Current Psychology Letters: Behaviour, Brain and Cognition, 11, http://cpl.revues.org/document390.html.
Anisfeld, M. (1996). Only tongue protrusion modeling is matched by neonates. Developmental Review, 16, 149161.

Chartrand, T. L., \& Bargh, J. A. (1999). The chameleon effect: The perception-behavior link and social interaction. Journal of Personality and Social Psychology, 76, 893-910.
Chartrand, T. L. \& van Baaren, R. B. (2009). Human mimicry. Advances in Experimental Social Psychology, 41, 219-273.
Deschamps, P. K. H., Schutte, I., Kenemans, J. L., Matthys, W., \& Schutter, D. J. L. G. (2012). Electromyographic responses to emotional facial expressions in 6-7 year olds: A feasibility study. International Journal of Psychophysiology, 85, 195-199.
Helt, M. S., Eigsti, I.-M., Snyder, P. J., \& Fein, D. A. (2010). Contagious yawning in autistic and typical development. Child Development, 81, 1620-1631.
Immens, L. (2011). Do people predict more rationality for an agent who is in a hurry? (Unpublished Bachelor's thesis). Radboud University, Nijmegen.
Jones, S. S. (2007). Imitation in infancy: The development of mimicry. Psychological Science, 18, 593-598.
Kenward, B., \& Dahl, M. (2011). Preschoolers distribute scarce resources according to the moral valence of recipient's previous actions. Developmental Psychology, 47, 1054-1064.
Lakin J. L., \& Chartrand, T. L. (2003). Using nonconscious behavioral mimicry to create affiliation and rapport. Psychological Science, 14, 334-339.
Lakin, J. L., Chartrand, T. L., \& Arkin, R. M. (2008). I am too just like you: The effects of ostracism on nonconscious mimicry. Psychological Science, 19, 816822.

Lausberg, H., \& Sloetjes, H. (2009). Coding gestural behavior with the NEUROGES-ELAN system. Behavior

Research Methods, Instruments, \& Computers, 41, 841849.

Likowski, K. U., Muhlberger, A., Seibt, B., Pauli, P., \& Weyers, P. (2008). Modulation of facial mimicry by attitudes. Journal of Experimental Social Psychology, 44, 1065-1072.
McIntosh, D. N. (2006). Spontaneous facial mimicry, liking and emotional contagion. Polish Psychological Bulletin, 37, 31-42.
Meltzoff, A. N. (2007). 'Like me’: A foundation for social cognition. Developmental Science, 10, 126-134.
Meltzoff, A. N., \& Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. Science, 198, 75-78.
Meltzoff, A. N., \& Moore, M. K. (1983). Newborn infants imitate adult facial gestures. Child Development, 54, 702709.

Morgante, J. D., Zolfaghari, R., \& Johnson, S. P. (2012). A critical test of temporal and spatial accuracy of the Tobii T60XL Eye Tracker. Infancy, 17, 9-32.
Moody, E. J., \& McIntosh, D. N. (2011). Mimicry of dynamic emotional and motor-only stimuli. Social Psychological and Personality Science, 2, 679-686.
Nielsen, M., Simcock, G., \& Jenkins, L. (2008). The effect of social engagement on 24-month-olds’ imitation from live and televised models. Developmental Science, 11, 722-731.
Over, H., \& Carpenter, M. (2009). Priming third-party ostracism increases affiliative imitation in children. Developmental Science, 12, F1-F8.
Over, H., \& Carpenter, M. (2012). Putting the social into social learning: Explaining both selectivity and fidelity in children's copying behavior. Journal of Comparative Psychology, 126, 182-192.
Platek, S. M., Critton, S. R., Myers, T. E., \& Gallup Jr., G. G. (2003). Contagious yawning: The role of selfawareness and mental state attribution. Cognitive Brain Research, 17, 223-227.
Sonnby-Borgström, M. (2002). Automatic mimicry reactions as related to differences in emotional empathy. Scandinavian Journal of Psychology, 43, 433-443.
Vaish, A., Carpenter, M., \& Tomasello, M. (2010). Young children selectively avoid helping people with harmful intentions. Child Development, 81, 1661-1669.
Van Baaren, R., Janssen, L., Chartrand, T. L., \& Dijksterhuis, A. (2009). Where is the love? The social aspects of mimicry. Philosophical Transactions of the Royal Society B, 364, 2381-2389.
Van Casteren, M., \& Davis, M. H. (2006). Mix, a program for pseudorandomization. Behavior Research Methods, 38, 584-589.
Woodward, A. L., Sommerville, J.A., Gerson, S., Henderson, A. M. E., \& Buresh, J. S. (2009). The emergence of intention attribution in infancy. In B. Ross (Ed.) The Psychology of Learning and Motivation, Volume 51. Academic Press.

# Joint Action Coordination through Strategic Reduction of Variability 

Cordula Vesper (vesperc@ ceu.hu)<br>Department of Cognitive Science, Central European University<br>Frankel Leó út 30-34, Budapest 1023, Hungary<br>Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen<br>P.O. Box 9104, 6500 HE Nijmegen, The Netherlands<br>Laura Schmitz (laschmit@uos.de)<br>Department of Cognitive Science, University of Osnabrück Albrechtstraße 28, 49076 Osnabrück, Germany<br>Natalie Sebanz (sebanzn@ceu.hu) Günther Knoblich (knoblichg@ceu.hu)<br>Department of Cognitive Science, Central European University Frankel Leó út 30-34, Budapest 1023, Hungary<br>Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen<br>P.O. Box 9104, 6500 HE Nijmegen, The Netherlands


#### Abstract

How do people coordinate actions with others? We tested the hypothesis that pairs of participants strategically reduce the variability of their action performance to achieve synchronicity in the absence of visual feedback about each other's actions. Consistent with this prediction, participants moved faster and less variably in a condition where they could not see their task partner's movements compared to a condition in which visual information was available. The accuracy of the resulting coordination was the same in both conditions. These findings are interpreted as evidence for general strategic adaptation in the service of real-time action coordination when only minimal perceptual information is available.


Keywords: Joint action; coordination strategy; cooperation; social cognition.

## Introduction

Whenever people coordinate their actions with other people, they are engaged in a 'joint action' (Clark, 1996; Marsh et al., 2009; Sebanz, Bekkering, \& Knoblich, 2006). Depending on the specific task and the presence or absence of an explicit joint action goal, different mechanisms and processes will make coordination of multiple people's actions possible. For instance, a couple might discuss through verbal or non-verbal communication who is responsible for preparing dinner and who will set the table (Clark \& Wilkes-Gibbs, 1986). Or a group of friends might help push-start a car by using perceptual cues and haptic information to predict when everyone else will push (van der Wel, Knoblich, \& Sebanz, 2011; Wilson \& Knoblich, 2005). In yet other cases, coordination might arise without prior planning as when two strangers unintentionally walk in synchrony (van Ulzen et al., 2008).

While people in these and many other everyday examples make use of visual, auditory or haptic information to guide their joint efforts, this is not always possible. Sometimes coordination is required in contexts where only little or even
nothing is known about the coordination partner and how or when the partner will perform a particular action. In these cases, all that might be represented is one's own action part ('ME'), the fact that someone will take care of another action part (' X ') required to achieve the joint goal and the joint action goal ('ME +X ') achieved by combining the individual action parts (Vesper et al., 2010). Thus, a precise representation about the partner's task might not be available. We claim that in these cases, coordination is supported by very general mechanisms and processes that are not required to the same extent if more information about a task partner is available. The present study addressed the mechanisms and processes allowing people to achieve coordination in this kind of minimal joint action situation.

More specifically, we investigated whether people who intend to coordinate their actions under real-time constraints and with no access to visual information about a task partner's actions adapt their own actions in a way that will make interpersonal coordination most likely. Such a coordination strategy (Vesper et al., 2010) reliably simplifies coordination in a general way, i.e. it is a modulation of one's own behavior that does not directly depend on how a task partner's particular action will unfold.

One example of strategic adaptation is to behave in a way that will make one's own actions predictable. When timing is not critical, this could involve relying on shared or conventional knowledge (Clark, 1996). For example, someone might decide to wait at the Brandenburg Gate to meet a friend in Berlin when they forgot to agree on a precise meeting point in advance (Schelling, 1960). Similarly, if each member of a group has to guess a number such that the sum of all numbers matches a randomly selected target number, providing consistent and therefore predictable guesses can be beneficial to achieve the desired group outcome (Roberts \& Goldstone, 2011).

In situations in which actions need to be coordinated in real-time, making actions predictable can involve mini-
mizing the variability of one's own performance. Recent empirical evidence for this claim is provided by a study in which pairs of participants performed a simple two alternative forced choice (2-AFC) reaction time task next to each other with the goal of synchronizing the timing of their response button presses (Vesper et al., 2011). An analysis of mean reaction times and the trial-by-trial variability of reaction times indicated that participants responded faster and with less variability in joint action compared to individual baseline performance. This in turn positively affected coordination such that pairs whose members responded fast and with little variability were on average better synchronized.

Critically, the study showed that it was the reduction in variability that predicted how successful coordination was, as demonstrated by a correlation of variability and asynchrony that persisted when controlling for the potential effects of mean reaction time. Thus, the more predictable actions were, the more successful interpersonal coordination was. Given that performing tasks at higher speed tends to reduce temporal variability (Wagenmakers \& Brown, 2007), participants most likely used speeding as a means to reduce their action variability. A second experiment demonstrated that speeding and predictability were only correlated with asynchrony when task partners intended to synchronize their button presses, but not in an experimental condition where the two people merely performed the task next to each other without a coordination goal. This suggests that the coordination strategy of making oneself predictable is used specifically to achieve intentional joint action coordination.

The aim of the present study was to extend these earlier findings (Vesper et al., 2011) by addressing three predictions following from the concept of a coordination strategy (Vesper et al., 2010) - generalizability, specificity, and independence. The first prediction was that the link between response speed, response variability, and asynchrony of task partners' actions would also be useful for coordination of more complex, temporally extended joint actions. Therefore, we instructed pairs of participants to each use a computer mouse to move a cursor on a screen from a start location towards a target with the joint goal of reaching the target at the same time (Figure 1). Thus, the task required two people to synchronize the endpoints of two-dimensional aiming movements. When they reached the target auditory feedback informed participants about their coordination accuracy.

The second prediction was that a coordination strategy will predominantly be used in situations in which no or only little information about a task partner's actions is available. In this 'minimal' case, all someone can do is to adapt his or her own actions in a general way to make coordination most likely. In contrast, when task or perceptual information is available, other mechanisms and processes will support coordination. For instance, co-actors can monitor (Malfait et al., 2009; Schuch \& Tipper, 2007) or predict (Graf et al., 2007; Knoblich \& Jordan, 2003) when and how another person will perform a particular action. Consequently, in
many situations, perceptual information is beneficial for joint action coordination. As an example, when two people build a toy model together such that one person (the director) verbally instructs another person (the builder) which parts to assemble, coordination is more successful if the director can see what the builder is doing (KrychAppelbaum et al., 2007). Similarly, two people who jointly search a shared workspace for a target object are more efficient in their search if they receive information about where each of them is currently looking at (Brennan et al., 2008). To test the specificity of coordination strategies, we compared an experimental condition in which co-actors did not receive visual information about each other (Other Hidden) with one in which they could see each other and each other's ongoing action performance (Other Visible). We hypothesized that a speeding and predictability strategy would predominantly be employed in the Other Hidden condition, whereas for Other Visible, we expected that the additional perceptual information would allow co-actors to use a different mechanism for coordination. This could involve monitoring and anticipating the partner's computer mouse movements. Therefore, we expected reaction times and movement variability to be smaller in Other Hidden compared to Other Visible. Given that perceptual information often positively influences coordination, we also hypothesized that asynchronies between co-actors' actions in Other Visible would be smaller, indicating better coordination accuracy when more information is available.

The third prediction was that a coordination strategy is used in a general way, independently of how the task partner actually performs an action. This means that the partner's particular action performance is not directly relevant for one's own strategic adaptation. In contrast, when other mechanisms such as monitoring and prediction are used, one's own action performance should be directly related to the task partner's action performance. One way to address this prediction is to compare the actually measured asynchronies between task partners' actions with asynchronies that are calculated after co-actors' reaction times have been shuffled and randomly matched. This method effectively treats the data as if each person's actions were not targeted towards a corresponding action of the co-actor because their actions now come from different trials. We hypothesized that this procedure would affect coordination in Other Hidden to a lesser extent than in Other Visible, indicating that co-actors in the former case adapt in a general way that is independent of the task partner's particular action performance, whereas in the latter case, co-actors make use of the given perceptual information and take into account how the partner's action unfolds on a trial-by-trial basis.

## Method

## Participants

Twenty-four students (14 women) participated in pairs. They were between 19 and 25 years old (mean 21.1 years)
and right-handed. They gave prior informed consent and received monetary compensation for their participation.

## Material and Apparatus

A "space mission" scene was created on two computer screens placed next to each other (Figure 1). The scene contained three elements presented on a dark blue background. First, close to the outer margin of each screen, a yellow "spaceship" was drawn (ca. $2.5 \mathrm{~cm} \times 1.9 \mathrm{~cm}$; position centrally on the vertical axis), indicating the starting position for each trial. Second, on the inner margin of each screen, a blue half circle was drawn on one of three possible locations, indicating a "planet" as the target (radius ca. 2.0 cm or 3.8 cm ; position at $20 \%$, $50 \%$ or $80 \%$ from the upper screen margin). When both screens were visible (Other Visible), the two half circles together formed a complete "planet". Finally, centrally between "spaceship" and "planet", on one of five possible locations, an array of small differently-sized white dots was drawn to represent an "asteroid belt" (ca. $1.9 \mathrm{~cm} \times 9.3 \mathrm{~cm}$; position at $20 \%$, $35 \%$, $50 \%, 65 \%$ or $80 \%$ from upper screen margin). It served as a potential obstacle between start and target locations.


Figure 1: The "space mission" scene (example layouts). A) In Other Hidden, each participant only saw one half of the scene due to an occluder placed between the participants. B) In Other Visible, both participants saw the complete scene.

The stimuli were presented on two 17 "-screens (resolution $1280 \times 1024$ pixel, refresh rate 60 Hz ). In individual baselines and in Other Hidden, a black card board (70 x 100 cm ) was set up between the two participants and between the two screens. The experiment was run on two Dell OptiPlex computers that were connected through a nullmodem cable to allow online data exchange. For data
collection, two special gaming computer mice (Logitech G500) were used that were sampled at 100 Hz and that had automatic acceleration turned off. Matlab version 2012a was used for controlling the experiment and for data analysis.

## Procedure

There were four experimental parts: the Other Hidden condition (Figure 1A), the Other Visible condition (Figure 1B) and two individual baselines. Each participant first performed four practice trials and then the first individual baseline, while the task partner waited in another room. After both participants had finished their first individual parts, they performed the two joint conditions together. The order of Other Hidden and Other Visible was counterbalanced. Finally, each participant separately performed another individual baseline. Each of the four parts consisted of six experimental blocks à 16 trials with short breaks in between. The overall duration of the experiment was about 1.5 hours.

At the beginning of a trial, the start location ("spaceship") was presented for 600 ms . Next, the target ("planet") and obstacle ("asteroid belt") appeared at a location that was randomly chosen from the possible locations. The frequency of target and obstacle locations and the target size were counterbalanced within each block. The relation of target and obstacle locations determined whether the direct path between start and target location was blocked by the obstacle or not. At the same time when target and obstacle appeared, the spaceship briefly flashed for 200 ms by showing flames at the rear engine. The purpose was to redirect participants' attention to the start location where a mouse cursor (a yellow circle) was now visible.

Participants were instructed to move the mouse cursor to the target without moving over the obstacle. A short feedback tone ( 100 ms ) was played as soon as they moved into the target area, i.e. no button press was required. The feedback tones for the left-seated and the right-seated participants differed in frequency so that they could be distinguished ( $1100 \mathrm{~Hz}, 1320 \mathrm{~Hz}$ ). Additionally, visual feedback about the accuracy of the trial was given: The planet turned red indicating negative task performance 1) if participants' movements were too slow (movement onset > 600 ms or reaction time > 1600 ms ), 2) if they moved over the obstacle area, 3) if the task partner had made any of these mistakes or 4) if co-actors did not reach the target synchronously (absolute asynchrony $>400 \mathrm{~ms})^{1}$. In all other cases, trials were successful and the planet turned into a bright green. Participants then returned to the start position and the next trial started.

Participants were told to think of the task as a space contest that requires securing planets from an alien nation by landing on a planet before them. According to this background story, in some areas of the universe (individual baselines), this could be achieved alone, whereas in other

[^243]areas of the universe (Other Hidden, Other Visible), they would have to arrive at the planet at the same time as the task partner in order to win. Thus, task instructions explicitly mentioned that participants should be as fast and as accurate as possible, while strongly focusing on arriving at the same time. Co-actors were not allowed to talk.

## Results

For the purpose of the present paper, we only report analyses of mean reaction times (RT; measured as the time from the start signal until the target was reached), standard deviation of reaction times (STD) and absolute asynchrony between participants' reaction times (ASYNC) ${ }^{2}$. These dependent variables were acquired by averaging over all trial types within a condition, i.e. we did not differentiate between different target and obstacle locations or target sizes. All trials in which participants' own RT was slower than 1600 ms or in which they moved over the obstacle area were excluded from further analyses (1.1 \% in Other Hidden, 0.7 \% in Other Visible).


Figure 2: Results. A) Mean RT. B) Mean trial-by-trial variability measured as STD. The dotted lines show individual baseline performance before (upper line) and after joint action (lower line). Error bars display withinsubject confidence intervals (Loftus \& Masson, 1994).

We first tested whether participants made use of a coordination strategy predominantly in the case where they did not receive visual information about the task partner's action. Confirming this hypothesis, RTs in Other Hidden were significantly faster, $F(1,23)=14.01, p<.01$ (Figure 2A) and less variable, $F(1,23)=5.36, p<.05$ (Figure 2B) than in Other Visible. Moreover, as described in more detail below, coordination between co-actors was equally good in the two conditions.

To investigate the hypothesized relation of RT, STD and ASYNC, we performed zero-order and partial correlations.

[^244]For Other Hidden, these analyses indicated that both RT and STD significantly influenced ASYNC such that shorter and less variable RTs led to better coordination between coactors (for exact results, see Figure 3A). Crucially, however, when controlling for RT in a partial correlation, STD still predicted ASYNC, whereas when controlling for STD, the relation between RT and ASYNC did not persist. Thus, as predicted, participants' response variability was critical in determining how well coordinated co-actors were when no online perceptual information about the task partner's actions was available. In contrast, in Other Visible, RT and STD did not predict ASYNC, although RT and STD were correlated (for exact results, see Figure 3B).


Figure 3: Zero-order and partial (in parentheses) correlations for A) Other Hidden and B) Other Visible. The thick arrow in A indicates that the relation between STD and ASYNC still holds when controlling for the influence of RT with a partial correlation. * $p<.05$, ** $p<.01$, *** $p<.001$.

Finally, we tested the hypothesis that the speeding and predictability strategy in Other Hidden is general in the sense that it depends only to a certain extent on how the task partner actually performed his or her actions. In contrast, coactors in Other Visible should take the other person's actual movements into account by monitoring and predicting the other's action. Based on this reasoning, we hypothesized that the coordination outcome from a general strategy should depend less on a trial-by-trial match of task partners' actions, whereas when using perceptual information this should be relevant. To test this prediction, we compared the originally measured asynchronies with asynchronies in which the specific trial-by-trial relation between co-actors’ actions was destroyed by randomly shuffling the order of trials from one member in each pair (separately for the different trial types) and re-calculating the asynchrony between the two persons' response times.

In line with our hypothesis, a comparison of original and shuffled asynchronies in the two conditions indicated an unequal effect of the shuffling: Although asynchronies increased in both conditions, shuffling co-actors' trial order had a significantly stronger effect for Other Visible than for Other Hidden. This was demonstrated statistically by an interaction of the factors Condition (Other Hidden, Other

Visible) and Trial Order (original, shuffled), $F(1,23)=$ 10.59, $p<.01$ (Figure 4). There was also a main effect of Trial Order, $F(1,23)=16.88, p<.001$, but no significant effect of Condition, $F(1,23)=.01, p>.9$. Thus, co-actors reached the same level of coordination performance in the two conditions.


Figure 4: Asynchrony measured (original) and re-calculated after randomly matching different trials from the members within a pair (shuffled). Error bars display within-subject confidence intervals (Loftus \& Masson, 1994).

## Discussion

The aim of the present study was to systematically test predictions following from the hypothesis that people strategically increase the speed and predictability of their actions to achieve real-time coordination with another person given only minimal perceptual information (Vesper et al., 2010, 2011). Pairs of participants performed mouse movements towards a target displayed on a computer screen. The joint goal was to reach the target at the same time as the task partner. Short feedback tones when arriving at the target informed participants about the accuracy of their joint coordination.

The present results confirm our hypothesis that co-actors strategically reduced the variability of their movements through speeding (Wagenmakers \& Brown, 2007) whereby action variability contributed directly to the coordination outcome. In particular, when controlling for the impact of reaction times, response variability still predicted asynchrony, whereas when controlling for the impact of response variability, there was no longer a correlation between reaction time and asynchrony. This not only replicates earlier findings (Vesper et al., 2011) but also demonstrates that this coordination strategy can support coordination in complex, temporally extended actions.

Furthermore, the present study shows that the predictability strategy is predominantly used in situations in which little or no information about the task partner is available. To test this hypothesis, we compared an experimental condition in which co-actors could not see each other (Other Hidden) with one in which visual
information was available (Other Visible). Consistent with our predictions, participants' movements were significantly faster and less variable without visual information and the relation between reaction time, variability, and coordination accuracy was present only in the Other Hidden condition. This confirms that coordination strategies are specific such that they are predominantly employed when other mechanisms like monitoring and predicting another's actions cannot be used.

A third hypothesis was that the employment of strategic behavior modulations would not depend on how the task partner actually performs his or her actions. Therefore, we compared two types of asynchronies: One that we had actually measured (original) and one that we calculated after shuffling the order of task partners' reaction times and randomly matching them again (shuffled). This resulted in a measure of how much each person's action was related to the task partner's actual action performance. Confirming our hypothesis, shuffling the trial order affected coordination significantly less in Other Hidden compared to Other Visible. Thus, when making oneself predictable one's own actions do not or only to a small extent depend on how exactly the task partner performs his or her action.

A possibly surprising result of the current study is that coactors were on average equally well-coordinated in the two joint conditions. Although perceptual information is often beneficial for joint action (e.g. Brennan et al., 2008; Knoblich \& Jordan, 2003; Krych-Appelbaum et al., 2007), this suggests that using a coordination strategy can compensate for a lack of perceptual information. Moreover, given that participants' actions were overall faster and less variable when no perceptual information was available, one might even argue that action performance was better without visual feedback. This is consistent with other evidence that having 'redundant' information potentially impairs joint action coordination. Specifically, when two people who jointly search a workspace not only receive visual information about each other's looking behavior, but can also talk to each other about the task, their search is considerably less efficient than when verbal communication is restricted (Brennan et al., 2008). As an alternative, the present findings might indicate that taking away perceptual information requires co-actors to put in extra effort in order to achieve the same degree of coordination.

How then did co-actors approach the task in the Other Visible condition? Although investigating this in detail is beyond the scope of the present paper, participants most likely used the available visual information to guide their actions either reactively (monitoring the task partner's action, then acting oneself) or predictively (anticipating when the task partner will reach the target and acting in accordance with this prediction). Further experiments could distinguish these two cases, e.g., by measuring participants’ eye movements to determine at what time during the interaction they track their task partner's movements.

The present study has implications beyond human joint action. For instance, an important research topic in the
cognitive sciences is how to implement real-time interaction of a robot and a human user. To that end, mechanisms observed in human social interaction are currently being transferred to robot platforms, including natural-language discourse (Salem et al., 2010), action prediction (Bicho et al., 2011; Dindo, Zambuto, \& Pezzulo, 2011) and continuous movement synchronization (Mörtl et al., 2012). Considering also general strategic behavioral adaptions for human-robot interaction can be beneficial for this endeavor. First, human users might employ a strategy such as making oneself predictable also when interacting with a robot so that coordination would improve if the robot used the same strategy. Second, human users might expect the robot to adapt its movements strategically so that robots that do so would appear more 'human-like' and thereby are more easily accepted as an interaction partner.

Taken together, this study provides evidence that general strategic adaptations of one's own actions can effectively support coordination with other people in situations in which precise representations about the partner's task might not be available. The concept of a coordination strategy therefore complements other approaches towards joint action like those focusing on communication (Clark, 1996), action prediction (Wilson \& Knoblich, 2005) or dynamic perception-action coupling (Marsh et al., 2009).

## Acknowledgments

The research reported in this manuscript was supported by the ESF program EuroUnderstanding. We thank Veronika Herendy for her help in collecting the data.

## References

Bicho, E., Erlhagen, W., Louro, L., \& Costa e Silva, E. C. (2011). Neuro-cognitive mechanisms of decision making in joint action: A human-robot interaction study. Human Movement Science, 30, 846-868.
Brennan, S. E., Chen, X., Dickinson, C. A., Neider, M. B., \& Zelinsky, G. J. (2008). Coordinating cognition: the costs and benefits of shared gaze during collaborative search. Cognition, 106, 1465-1477.
Clark, H. H. (1996). Using language. Cambridge, England: Cambridge University Press.
Clark, H. H., \& Wilkes-Gibbs, D. (1986). Referring as a collaborative process. Cognition, 22, 1-39.
Dindo, H., Zambuto, D., \& Pezzulo, G. (2011). Motor simulation via coupled internal models using sequential Monte Carlo. Proceedings of the International Joint Conference on Artificial Intelligence, 2113-2119.
Graf, M., Reitzner, B., Corves, C., Casile, A., Giese, M., \& Prinz, W. (2007). Predicting point-light actions in realtime. NeuroImage, 36, T22-T32.
Knoblich, G. \& Jordan, J. S. (2003). Action coordination in groups and individuals: learning anticipatory control. Journal of Experimental Psychology: Learning, Memory, and Cognition, 29(5), 1006-1016.
Krych-Appelbaum, M., Law, J. B., Jones, D., Barnacz, A., Johnson, A., \& Keenan, J. P. (2007). "I think I know what
you mean" - The role of theory of mind in collaborative communication. Interaction Studies, 8(2), 267-280.
Loftus, G. \& Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. Psychonomic Bulletin and Review, 1, 476-490.
Malfait, N., Valyear, K. F., Culham, J. C., Anton, J.-L., Brown, L. E., \& Gribble, P. L. (2009). FMRI activation during observation of others' reach errors. Journal of Cognitive Neuroscience, 22(7), 1493-1503.
Marsh, K. L., Johnston, L., Richardson, M. J., \& Schmidt, R. C. (2009). Toward a radically embodied, embedded social psychology. European Journal of Social Psychology, 39, 1217-1225.
Mörtl, A., Lorenz, T., Vlaskamp, B., Gusrialdi, A., Schubö, A., \& Hirche, S. (2012). Modeling inter-human movement coordination: synchronization governs joint task dynamics. Biological Cybernetics, 106, 241-259.
Roberts, M. E. \& Goldstone, R. L. (2011). Adaptive group coordination and role differentiation. PLoS ONE 6(7), e22377.
Salem, M., Kopp, S., Wachsmuth, I., \& Joublin, F. (2010). Towards an integrated model of speech and gesture production for multi-modal robot behavior. Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication, 649-654.
Schelling, T. C. (1960). The strategy of conflict. Harvard University Press, Cambridge.
Schuch, S., \& Tipper, S. P. (2007). On observing another person's actions: influences of observed inhibition and errors. Perception and Psychophysics, 69(5), 828-837.
Sebanz, N., Bekkering, H., \& Knoblich, G. (2006). Joint action: Bodies and minds moving together. Trends in Cognitive Sciences, 10, 71-76.
van der Wel, R., Knoblich, G., \& Sebanz, N. (2011). Let the force be with us: Dyads exploit haptic coupling for coordination. Journal of Experimental Psychology: Human Perception and Performance, 37, 1420-1431.
van Ulzen, N. R., Lamoth, C. J. C., Daffertshofer, A., Semin, G. R., Beek, P. J. (2008). Characteristics of instructed and uninstructed interpersonal coordination while walking side-by-side. Neuroscience Letters, 432, 88-93.
Vesper, C., Butterfill, S., Knoblich, G. \& Sebanz, N. (2010). A minimal architecture for joint action. Neural Networks, 23, 998-1003.
Vesper, C., van der Wel, R., Knoblich, G., \& Sebanz, N. (2011). Making oneself predictable: Reduced temporal variability facilitates joint action coordination. Experimental Brain Research, 211(3-4), 517-530.
Wagenmakers, E. J. \& Brown, S. (2007). On the linear relation between the mean and the standard deviation of a response time distribution. Psychological Review, 114, 830-841.
Wilson, M., \& Knoblich, G. (2005). The case for motor involvement in perceiving conspecifics. Psychological Bulletin, 131(3), 460-473.

# Incidental and Non-Incidental Processing of Biological Motion: Orientation, Attention and Life Detection 

Peter Veto (peter.veto@univr.it),<br>Department of Neurological, Neuropsychological, Morphological and Movement Sciences<br>Section of Physiology and Psychology University of Verona, Strada Le Grazie, 8, 37143 Verona - Italy

Serge Thill (serge.thill@his.se) and Paul Hemeren (paul.hemeren@his.se)<br>Interaction Lab, Informatics Research Centre<br>University of Skövde<br>54128 Skövde, Sweden


#### Abstract

Based on the unique traits of biological motion perception, the existence of a "life detector", a special sensitivity to perceiving motion patterns typical for animals, seems to be plausible (Johnson, 2006). Showing motion displays upsidedown or with changes in global structure is known to disturb processing in different ways, but not much is known yet about how inversion affects attention and incidental processing. To examine the perception of upright and inverted point-light walkers regarding incidental processing, we used a flanker paradigm (Eriksen \& Eriksen, 1974) adapted for biological motion (Thornton \& Vuong, 2004), and extended it to include inverted and scrambled figures. Results show that inverted walkers do not evoke incidental processing and they allow high accuracy in performance only when attentional capacities are not diminished. An asymmetrical interaction between upright and inverted figures is found which alludes to qualitatively different pathways of processing.


Keywords: biological motion perception; point-light walker; incidental processing; inversion effect; life detector

## Introduction

An important feature of the visual processing of the dynamic human gestalt in point-light displays is the "automatic" nature of the perceptions. As Johansson (1973) points out, "... we have found that it seems to be a highly mechanical, automatic type of visual data treatment that is most important." While Johansson's use of the term "automatic" points more to the early processes involved in establishing hierarchies of locally rigid perceptual units, there is a case to be made for the automatic processing of biological motion at a higher cognitive level under favorable circumstances, i.e., given an appropriate task. Phenomenally, Johansson's own demonstrations point to the immediateness and vividness of viewing point-light displays of biological motion. Observers are fast and accurate in their identifications when not disrupted by dynamic masking. They appear to have direct access to a level of meaning that facilitates the identification and recognition of actions depicted in the point-light displays. In contrast to upright displays, inverted point-light displays lead to impaired recognition, identification, detection and priming (e.g.,

Bertenthal \& Pinto, 1994; Daems \& Verfaillie, 1999; Hemeren, 2008; Shiffrar \& Pinto, 2002; Troje, 2003).

Previous results have also shown that point-light walkers (PLWs) trigger attention mechanisms (Thornton \& Vuong, 2004). Using a flanker paradigm, Thornton and Vuong (2004) demonstrated incidental processing of upright oriented PLW flankers during judgments of the walking direction of the displays. Upright point-light walkers can elicit incidental processing while static, scrambled or chimeric ones cannot, but in these studies they did not address the question of inversion. Incidental processing was indicated by an increase in the time it took to make direction judgments when the direction of the flankers was incongruent with the direction of the target. The task irrelevant flankers interfered with the visual processing of the target.

Recently, Shi et al. (2010) reported effects of upright PLWs on the accuracy of reporting the perceived direction of a Gabor patch. Accuracy was significantly lower when the walking direction of an upright PLW was incongruent with the orientation of the Gabor patch. Importantly, no such effect was found for inverted PLWs. This suggests the existence of a perceptual cue that triggers reflexive attentional orienting for upright, but not for inverted, PLWs. This is consistent with previous results regarding a general inversion effect and evidence for a "life detector" (see e.g., Johnson, 2006; Troje \& Westhoff, 2006).

In our study, we investigate the differential effects of the orientation of PLWs within the framework of Hochstein and Ahissar's Reverse Hierarchy Theory (2002). The idea here is that the visual quality of biological motion perception for upright displays is indicative of global processing as well as quick access to semantic level representations. Consistent with Hochstein and Ahissar (2002), the perception of inverted displays could be characterized as an example of illusory conjunctions. The perception of inverted displays could be said to demonstrate the effects of top-down processing in the sense that the default value is an upright orientation and this creates false conjunctions in the perception of inverted displays.

This line of reasoning is consistent with the reasoning in Shiffrar, Lichtey \& Heptulla Chatterjee (1997) where they show that global processes are involved in the perception of
upright biological motion displays across apertures but that this global processing is impaired when inverted biological motion displays are viewed across apertures. Their findings show that global processing is associated with viewing upright displays and that local processing is associated with viewing inverted displays.

Given this, we suggest that upright PLWs are incidentally processed on the basis of initial explicit perception as 'vision at a glance,' and it also reflects the activity of large receptive fields of high cortical areas and spread attention of initial perception (Hochstein \& Ahissar, 2002). At the other (low-level) end of the processing continuum, inverted PLWs constitute 'vision with scrutiny' which involves focused attention and the activation of small receptive fields in lower cortical areas.

By extending the flanker paradigm in Thornton and Vuong (2004) to include inverted walkers, we can further address the issue of incidental processing of flankers while performing a direct visual task on a central target. It may be the case that inverted walkers can be incidentally processed when the target is also an inverted walker. This condition can be directly contrasted with upright flankers and an upright target, for which there is already evidence of incidental processing (Thornton \& Vuong, 2004). This study will therefore include a replication of those results. Directional congruence will also be included in this study in order to assess the potential interference or facilitation effects of similar or different walking directions of the targets and flankers.

Here we can investigate the orthogonal pattern of interaction between upright and inverted displays under conditions of orientation congruence and direction congruence. This allows us to potentially see asymmetrical interactions in the way upright and inverted flankers modulate the visual processing of upright and inverted targets.

One obvious prediction from previous findings of the inversion effect is that inverted flankers will have no effect on reaction time or accuracy in detecting the walking direction of an upright target. This is due to the relatively fast and automatic processing of an upright and biologically relevant moving human. The structure of the information in the inverted flankers is not sufficient to modulate the processing of an upright target.

The potential effect of inverted flankers on inverted targets is less obvious to predict. Previous evidence (Hemeren, 2008) shows that inverted PLWs can prime (repetition priming) themselves as well as other inverted point-light actions. Given this evidence, inverted flankers may be incidentally processed because the visual system is active in scanning the available information for clues to resolve the conflicts (false conjunctions) in the inverted targets. This, however, entails that the information in the inverted flankers is relevant. If there is no relevant information, then there will be no incidental processing of inverted displays.

Since upright flankers convey biologically relevant information and are visually processed relatively automatically, we expect them to be incidentally processed when presented with an inverted target. This incidental processing will likely lead to significant interference when judging the walking direction of an inverted target.

The effect of upright flankers on upright targets will likely depend on the congruence of walking direction for the target. Based on results from Thornton and Vuong (2004) we expect the incidental processing of upright flankers to interfere with upright targets when they are walking in different directions (directional incongruence). When they are walking in the same direction (directional congruence), the question is whether the incidental processing of upright flankers will speed up (facilitate) the ability to correctly detect the walking direction of the target in relation to inverted flankers or whether there will be no difference between the effects of upright and inverted flanker on detection time for upright targets.

## Methods

## Participants

Ten right-handed subjects (5 male and 5 female, aged 20 to 49 years, $\mathrm{M}=28.9$, with normal or corrected-to-normal vision) participated in the experiment. Participants were selected from colleagues and the student population of the University of Skövde. All participants were naive to the purpose of the experiment and only the students received monetary compensation (approximately \$15). Participants provided written informed consent. Experiment protocol conformed to Swedish law and the World Medical Association Declaration of Helsinki.

## Stimuli

The stimuli consisted of the target walker $\left(1.26^{\circ} \times 0.74^{\circ}\right)$ displayed in the center surrounded by five distractors evenly placed at a fixed distance from the target $\left(1.89^{\circ}\right)$ with a randomly defined angular offset for the five flankers together. To compensate for the smaller cortical representation of peripheral stimuli, flankers were scaled by a cortical magnification factor (Goolkasian, 1997), thus having the size of $2^{\circ} \times 1.17^{\circ}$. The total size for the whole display was $5.78^{\circ}$. Figures were depicted in profile by 13 dots based on the 3-dimensional coordinates of the action "Walk" from a stimulus set of human point-light actions created by Vanrie \& Verfaillie (2004). For presentation of the stimuli and recording of the answers MatLab R2010a was used on an HP EliteBook 8440p laptop computer. An HP L2245wg monitor ( $1440 \times 900$ pixels, 60 Hz ) displayed the stimuli at 100 cm viewing distance.

In every condition the same figure was mirrored so that there were two possible directions of translation (left and right) and two possible orientations (right-side-up and upside-down). In addition to these four variations, scrambled flankers were created for control by mixing dots randomly chosen from the four conditions and displaying
them with random starting positions, while the total size of the figure was kept equal to the regular flankers. This way, scrambled flankers as a whole did not contain directional or orientation information, while the local motion patterns of the dots remained the same as in the globally intact figures. The four variations of the target and five variations of the flankers resulted in twenty conditions in total (see examples in Figure 1).


Figure 1: Upright target with translation to left and inverted flankers with translation to right (panel A); Inverted target with translation to right and scrambled flankers (panel B).

## Procedure

Participants were informed about that they would see a centrally located PLW (target) either upright or inverted and that flankers would surround the target. They were instructed to just focus on the target and to indicate the walking direction of the target by pressing one of two keys.

Left or right responses were given by key presses with the corresponding index fingers, indicating the direction of translation of the target regardless of its orientation. Participants were instructed as to the importance of the speed and accuracy of their responses. Stimuli were played from a randomly chosen starting frame (randomized between figures as well) in a continuous loop at 30 FPS (stride frequency: 2/s) until the participant responded. Every trial was preceded by an ISI of $500-800 \mathrm{~ms}$, during which a fixation cross $\left(0.23^{\circ}\right)$ marked the center of the display.

Each participant started with a training session of 32 trials. After that, 1440 trials were recorded, divided by arbitrary breaks into three sessions of 480 trials. In one session out of the three, $50 \%$ of the trials contained the scrambled conditions, while the other two sessions were made up of only non-scrambled trials. The order of the sessions varied between participants. This design was necessary to avoid a possible novelty effect of scrambled trials, since their total number was less than the total number of non-scrambled trials. Altogether, 1200 nonscrambled and 240 scrambled trials were completed by each subject, which means 75 trials per each non-scrambled condition and 60 trials per each scrambled condition.

The design of this experiment consisted of four independent variables; Target orientation (upright vs.
inverted), Flanker orientation (upright vs. inverted), Target direction (left vs. right) and Flanker direction (left vs. right). In addition, four conditions of scrambled flankers were created by pairing scrambled flankers with levels of target orientation and direction. Dependent variables are reaction time and accuracy.

## Results

## Reaction Times

Reaction time (RT) data were analyzed only for correct responses, with all outliers exceeding 2 SDs above and below the mean eliminated. Errors (accounting for 2.42 \% of all answers) were analyzed separately. Means generated for each condition and subject were used in a repeatedmeasures ANOVA analysis. Individual conditions in relevant cases were compared with t-tests.

Since the task of the participants was to respond to targets and to ignore the flankers, the amount of influence of the flankers on RTs and accuracy can be accounted for incidental processing of these figures. This is expressed in the walker congruency effect (WCE, Thornton \& Vuong, 2004) which in our case is positive for upright flankers but missing when flankers are inverted, i.e., responses to all targets are faster $(t(9)=5.46, p=0.000)$ when upright flankers have congruent direction of translation ( $M=644.2$ $\mathrm{ms}, S D=79.10$ ) compared to responses on incongruent trials with upright flankers ( $M=683.75 \mathrm{~ms}, S D=76.61$ ), while this difference cannot be found ( $M_{\text {Congruent }}=650.05$, $S D=78.98 ; M_{\text {Incongruent }}=650.55, S D=78.38 ; t(9)=0.15, p=$ 0.886 ) when flankers are inverted (Figure 2).


Figure 2: Walker congruency effect. (RSU = right-side-up, INV $=$ inverted. Asterisks indicate significant differences at $\mathrm{p}<0.05$, error bars show $95 \%$ confidence intervals.)

Regarding the interaction between the processing of upright and inverted biological motion, the RT means for direction judgments as a function of direction congruence and orientation congruence between flankers and targets are presented in Figure 3.

The pattern of results shows that inverted flankers have no effect on reaction times to the target and that these responses are also similar to conditions where the target was surrounded by scrambled flankers (Figure 3).

A $2 \times 2 \times 2 \times 2$ repeated measures ANOVA was carried out on the means. The relevant differences here are between right-side-up and inverted targets, where the inversion effect leads to a significant increase in RTs. The main effect for target inversion was significant, $F(1,9)=40.90$, partial $\eta 2=$ $0.82, p<0.001$. In relation to the error rates displayed in Figure 4, this main effect was not accompanied by a difference in accuracy. Upright flankers had no effect on the accuracy of judging the direction of upright or inverted targets. The difference of flanker influence in this case is restricted to reaction time, not accuracy.

The effect of flanker orientation on reaction time is limited to upright flankers. These distractors show a highly significant interference on inverted targets when their directions are incongruent ( $M=728.30, S D=86.82 ; t(9)=$ $6.51, p<0.001$ ) compared to the effect of inverted flankers in the respective condition $(M=669.45, S D=81.10)$. This effect is smaller but still significant when comparing the same conditions ( $M_{R S U}$ flankers $=682.90, S D=88.41 ; M_{I N V}$ flankers $=670.25, S D=82.10$ ) with congruent direction $(t(9)=$ $2.42, p=0.039$ ).

The effect of upright flankers is different when targets are also upright. In this case we do not see any increase in reaction times even with incongruent directions $\left(M_{R S U}\right.$ ${ }_{\text {flankers }}=639.20, S D=69.60 ; M_{\text {INV flankers }}=631.65, S D=78.00$; $t(9)=1.55, p=0.156)$, although the higher error rates show that flankers are processed and they affect the accuracy of responses ( $M_{R S U}$ flankers $=3.71, S D=3.49 ; M_{I N V}$ flankers $=1.89$, $S D=2.87 ; t(9)=2.29, p=0.047)$. When directions are congruent however, the effect of upright flankers on upright targets becomes facilitative: RTs are significantly lowered compared to the corresponding condition with inverted flankers $\left(M_{R S U}\right.$ flankers $=605.50, S D=72.92 ; M_{I N V}$ flankers $=$ $629.85, S D=79.21 ; t(9)=7.90, p=0.000)$. This relative facilitation however does not appear when compared to the effect of scrambled flankers $\left(M_{\text {Scrambled flankers }}=624.45, S D=\right.$ $76.14 ; t(9)=1.45, p=0.182$ ). This shows that upright flankers are processed to an extent that seems to occur in parallel to the upright targets. This is only the case when the direction of upright flankers and targets is congruent.

In contrast to the reaction times, the error rates in the case of relative facilitation stayed unchanged ( $M_{R S U}$ flankers $=1.24$, $S D=1.79$ ), compared to conditions either with inverted ( $M$ $=1.29, S D=2.52 ; t(9)=0.14, p=0.89)$ or scrambled flankers $(M=1.58, S D=2.31 ; t(9)=0.65, p=0.533)$. Accuracy is therefore not affected by flankers when directions are congruent, only in incongruent conditions and with upright flankers.


Figure 3: Conditions across orientation and direction of translation. (RSU = right-side-up, INV = inverted, $\mathrm{Sc}=$ scrambled. Asterisks indicate significant differences at $\mathrm{p}<0.05$, error bars show $95 \%$ confidence intervals.)


Figure 4: Errors. (RSU = right-side-up, INV = inverted, Sc $=$ scrambled. Asterisks indicate significant differences at $\mathrm{p}<0.05$, error bars show $95 \%$ confidence intervals.)

## Discussion

Our results show that inverted biological motion does not elicit incidental processing and even more importantly, upright and inverted point-light walkers have substantially different attention demands. Subjects are only required to respond to targets, and thus the processing of flankers happens without active top-down control. Nevertheless, upright flankers have significant effects on the responses to targets, the effect of which is different depending on the
orientation of the target. When both the target and the flankers are upright, reflexive attention seems to be drawn to both, thus resulting in either a relative facilitation in RTs (seen in the case of congruent directions) or in interference leading to higher error rates (incongruent directions). This interference shows the processing of both the target and flankers reach the level of subtracting directional information by the starting of the response. However, this does not lead to higher RTs - which may be due to the reflexive manner of the response to the target as well, allowing the two processes to run in a parallel manner.

When targets are inverted, they require more top-down control, and there is no incidental processing which indicates that attention is directed to them in a reflexive manner. The amount of attention incidentally drawn by the upright flankers leads to faster processing of the distractors, thus always interfering with the processing of inverted targets and leading to higher RTs. Furthermore when the two processes involve handling incongruent direction between flankers and targets, error rates become higher as well.

In terms of Hochstein and Ahissar's RHT (2002), these results suggest that the visual processing of upright human point-light walkers is consistent with vision at a glance since upright flankers not only interfere with the visual processing of inverted displays but that they can also modulate the visual processing of upright centrally displayed targets. In addition to the results for the speed of visual processing (RT-results), this interpretation is strengthened by the increase in the error rate for upright targets when the direction of upright flankers is in conflict with the target.

It is important to emphasize that RHT is not restricted to perceptual learning as such but applies to perception in general (Ahissar \& Hochstein, 2004). PLWs are salient examples of dynamic gestalt figures, which also include other action categories (e.g., Hemeren, 2008). The original findings from Johansson (1973) demonstrate that when presented with a static form, people have difficulty in identifying the figure and action. However, once the figure starts to move, people see the action that the person is performing. Much previous research (see e.g., Shiffrar \& Pinto, 2002) demonstrates the holistic/global processing involved in the visual perception of point-light displays of biological motion. From a perceptual learning perspective, biological motion perception is an example of the Eureka effect (Ahissar \& Hochstein, 2004), in which learning is governed by top-down control and single exposures and has long-lasting effects.

In contrast to upright targets and flankers, the results for inverted flankers in relation to inverted and upright targets indicate visual processing consistent with vision with scrutiny. There appears to be no access to high-level perceptual meaning that would trigger reflexive attention and lead to incidental processing. Although there is evidence of perceptual learning for inverted displays of biological motion (Grossman, Blake \& Kim, 2004), learning
is relatively difficult in terms of time taken and the ability to discriminate between different action categories depicted in the point-light displays (Hemeren, 2008). Inverted displays are perceptually difficult to resolve, i.e., they have no perceptually obvious ecological relevance. When confronted with this situation, visual processing is guided down the processing hierarchy where more local processing of the stimulus is carried out in order to find a solution to the perceptual problem. This requires the activation of small receptive fields in lower cortical areas. This takes additional time and is also prone to an increase in perceptual errors. If incidental processing of upright PLWs is evidence of a life detector (Johnson, 2006; Troje \& Westhoff, 2006), then it also shows that it also occurs at a high cortical level.

Recently, Ikeda, Watanabe and Cavanagh (2013) used a horizontal flanking paradigm to investigate the effect of the distance of upright PLWs and scrambled PLWs to upright PLW targets. Consistent with our results, they found that direction discrimination became more difficult with smaller distances between the flankers and the target. It is important to note that the conditions in their experiment were all directionally incongruent between flankers and targets. Ikeda et al. (2013) assert that their results show that the "crowding" effect occurs at a high-level of motion perception since the effect was absent when scrambled flankers were used.

An additional perspective on our results can be seen in the work of Vicario and Kiritani (1999) where the issue can be described as a matter of a vertical organization of visual events, i.e., determining what rules apply if simultaneous stimuli are perceived as one object (integration) or as different objects (segregation). In our case the relevant traits of the point-lights influencing this judgment seem to be: the amount and direction of displacement, speed and acceleration patterns and the variability in distance from neighboring dots. These traits are comparable with the Gestalt principles; however, it is not clear how they can unequivocally explain the inversion effect. One possibility is through the congruency or incongruency between acceleration patterns of the dots presented and acceleration patterns normally determined by gravity (this approach is discussed in detail by Chang and Troje, 2009). Another possibility is that an additional Gestalt principle is playing a role here, which leads to an effortless and fast form-frommotion perception of human figures, when the emerging form matches the usual human body configuration (i.e. it is walking with the right side up). Presumably both mechanisms are important in the inversion effect with pointlight walkers (Hirai, Chang, Saunders \& Troje, 2011).

Our results suggest that people have implicit access to initial high-level meaning for upright PLWs and that this access can be reflexively triggered when a visual target is difficult to perceptually resolve or when simultaneously presented PLWs are incongruent with regard to a relevant visual task.

## Acknowledgments

Thanks to the ERASMUS Student Mobility for Placement program for financing the stay of Peter Veto at the University of Skövde.

## References

Ahissar, M. \& Hochstein, S. (2004) The reverse hierarchy theory of visual perceptual learning. TICS, 8(10), 457464.

Bertenthal, B. I. \& Pinto, J. (1994) Global processing of biological motions. Psychological Science, 5, 221-225.
Chang, D. H. F. \& Troje, N. F. (2009) Acceleration carries the local inversion effect in biological motion perception. Journal of Vision, 9(1):19, 1-17.
Daems, A. \& Verfaillie, K. (1999) Viewpoint-dependent priming effects in theperception of human actions and body postures. Visual Cognition, 6(6), 665-693.
Gilaie-Dotan, S., Bentin, S., Harel, M., Rees, G. \& Saygin, A. P. (2011) Normal form from biological motion despite impaired ventral stream function. Neuropsychologia, 49, 1033-1043.
Goolkasian, P. (1997) Size scaling and spatial factors in visual attention. American Journal of Psychology, 110(3), 397-415
Grossman, E. D., Blake, R. \& Kim, C-Y. (2004) Learning to see biological motion: Brain activity parallels behavior. Journal of Cognitive Neuroscience, 16(9), 1669-1679.
Hemeren, P.E. (2008) Mind in Action: Action recognition and the perception of biological motion. Lund University Cognitive Studies 140, Lund University, Sweden.
Hirai, M., Chang, D. H. F., Saunders, D. R. \& Troje, N. F. (2011) Body configuration modulates the usage of local cues to direction in biological-motion perception. Psychological Science, 1(22):12, 1543-1549.
Hochstein, S. \& Ahissar, M. (2002) View from the top: Hierarchies and reverse hierarchies in the visual system. Neuron, 36, 791-804.
Ikeda, H., Watanabe, K. \& Cavanagh, P. (2013) Crowding of biological motion stimuli. Journal of Vision, 13(4):20, 1-6.
Johansson, G. (1973) Visual perception of biological motion and a model for its analysis. Perception and Psychophysics, 14, 201-211.
Johnson, M. H. (2006) Biological motion: A perceptual life detector? Current Biology, 16(10), 376-377.
Jokisch, D., Daum, I., Suchan, B. \& Troje, N. F. (2005) Structural encoding and recognition of biological motion: Evidence from event-related potentials and source analysis. Behavioural Brain Research, 157, 195-204.
Shi, J., Weng, X., He, S. \& Jiang, Y. (2010) Biological motion cues trigger reflexive attentional orienting. Cognition, 117, 348-354.
Shiffrar, M., Lichtey, L. \& Heptulla Chatterjee, S. (1997) The perception of biological motion across apertures. Perception and Psychophysics, 59(1), 51-59.

Shiffrar, M. \& Pinto, J. (2002) The visual analysis of bodily motion. In W. Prinz \& B. Hommel (Eds.) Common mechanisms in perception and action: Attention and Performance. Oxford: Oxford University Press.
Thornton, I. M. \& Vuong, Q. C. (2004) Incidental processing of biological motion. Current Biology, 14, 1084-1089.
Troje, N. F. (2003) Reference frames for orientation anisotropies in face recognition and biological-motion perception. Perception, 32, 201-210.
Troje, N. F. \& Westhoff, C. (2006) The inversion effect in biological motion perception: Evidence for a "life detector"? Current Biology, 16, 821-824.
Vanrie, J. \& Verfaillie, K. (2004) Perception of biological motion: a stimulus set of human point-light actions. Behavior Research Methods, Instruments, \& Computers, 36, 625-629.
Vicario, G. B. \& Kiritani, Y. (1999) Slow-motion tunnel effect: An inquiry into vertical organization of perceptual events. Gestalt Theory, 21(2), 100-121

# The Role of Circadian Variations and Socially Distributed Thinking in Belief Perseverance 

Gaëlle Villejoubert (G.Villejoubert@kingston.ac.uk)<br>Madiha Khan (k0801532@kingston.ac.uk)<br>Frédéric Vallée-Tourangeau (F.Vallee-Tourangeau@kingston.ac.uk)<br>Department of Psychology, Penrhyn road<br>Kingston upon Thames, KT1 2EE United Kingdom


#### Abstract

From the moment they make up their mind, people are reluctant to change it. We tested the hypothesis that people disposing of more cognitive resources-through circadian variations or socially distributed thinking-would engage in deliberative thinking and would consequently be less likely to exhibit belief perseverance. Perseverance was measured by the change in judgments related to a suspect in a criminal case, following the presentation of an offender profile that was at odds with the suspect's description. Individuals tested at a compatible circadian time exhibited less perseverance in the face of contradictory evidence compared to individuals tested at an incongruent time. Individuals deliberating on their own also tended to show more belief perseverance compared to those who worked in groups. There was, however, no interaction effect between circadian timing and condition of deliberation on belief change. The implications for our understanding of the mechanisms that underpin belief perseverance are discussed.


Keywords: Belief perseverance; Dual-process accounts of cognition; Circadian variations; Socially distributed cognition.

## Introduction

Despite the public release of his birth certificate, a 2011 Gallup poll published by USA Today revealed that only $38 \%$ of Americans definitely believed President Obama was born in the US (Adams, 2011). Although the President was reportedly "puzzled" by the persistence of these rumors, this incident illustrates a well-established finding in psychological science: it is easier to get people to believe something new than to get them to abandon an existing belief even in the face of indisputable evidence to the contrary-a phenomenon known as belief perseverance (Ross, Lepper \& Hubbard, 1975).

Belief perseverance may concern beliefs about one's own skills and abilities, those of others, as well as naïve theories about stereotypical traits and behaviors (Anderson, 2007). In their seminal study of perseverance in social perception, Ross et al. (1975) presented a series of cards containing a real and a fictitious suicide note and asked participants to decide which note was written by a patient with suicidal ideation. Participants first received false feedback on their performance and were later debriefed on the arbitrary nature of the feedback they received. Yet, when participants were asked to reassess their performance after the thorough debriefing, those judgments remained strongly influenced by the initial spurious test results they had received.

A combination of three types of cognitive processes have been implicated in the perseverance of beliefs (Anderson \& Lindsay, 1998): the availability heuristic, its associated illusory correlation effect, and the anchoring and adjustment heuristic (Tversky \& Kahneman, 1974). Events such as a performance appraisal or the observation of a person's behavior are presumed to initiate the generation of a causal explanation (e.g., "I am (un)skilled at this task"). This new belief thereafter remains available in memory, independently of the inceptive evidence such that when the evidence is discredited, the belief remains intact (Anderson, Lepper, \& Ross, 1980).

Attempts to alleviate belief perseverance have been met with mitigated success. The most effective approach has been to encourage individuals to consider alternative causal explanations-counterexplanations-in an effort to reduce the influence of the inceptive belief (Anderson, 1982, 2007). Yet, considering alternatives may sometimes backfire. When individuals are given the opportunity to discount negative evidence, for example by plausibly discrediting it, the availability of counterexplanations no longer reduces belief perseverance (Vallée-Tourangeau, Beynon, \& James, 2000). Moreoever, perseverance is aggravated when evidence for the alternative explanation is difficult to elicit or when evidence for the target hypothesis is easily accessible (Nestler, 2010).

The mitigated success of the counterexplanation account, we wish to argue, can be best accounted for within a dualprocess framework. The dual-process view of cognition has gained considerable influence in the past decade in research examining judgment, decision-making or reasoning (e.g., see Darlow \& Sloman, 2010 for a review) but has yet to be applied to the study of belief perseverance. According to this view, two families of cognitive processes may underpin judgments and decisions: an intuitive mode of cognitive functioning-where judgments originate from rapid and automatic processes-and a second, more deliberate and effortful mode of thinking engaging processes which can either be at the origin of the judgment provided or simply monitor its quality (Kahneman, 2003). The type of heuristic processing that is taken to underpin belief perseverance is a trademark of the intuitive heuristic mode of thinking. This suggests that any situation either favoring deliberative thinking or augmenting deliberative thinking capacity should lead to a decrease in belief perseverance. We examine this possibility by investigating the role of two
variations in cognitive resources on belief perseverance: circadian preferences and socially distributed thinking.

Several studies have demonstrated that the efficiency of executive control-a key feature of deliberative thinkingis contingent upon the synchronicity between people's peak period of circadian arousal and the time of testing (for a review, see Schmidt, Collette, Cajochen, \& Peigneux, 2010). For example, West, Murphy, Armilio, Craik and Stuss (2002) used a choice reaction time task placing variable demands upon working memory: individuals were asked to identify and respond to the spatial location of a target presented in a previous display either with or without a distractor that they had to ignore. Time-of-day variations had an effect on performance, but only when more controlled processes were involved (e.g., when a distractor had to be inhibited); performance on simpler trials requiring automatic processes was unaffected by circadian variations. Schmidt et al. (2010) reviewed this and other studies pointing to similar results and concluded that cognitive functioning at nonoptimal time of day was typically associated with failure to clear or suppress irrelevant information and difficulties to resist predominant responses even if they are incorrect. For example, Bodenhausen (1990) showed that people were less likely to rely on stereotypic preconceptions when rendering judgments at a time of day that was congruent with their circadian preferences.

From a dual-process perspective, these results suggest that circadian congruence fosters the optimal deployment of cognitive resources, enabling people to engage in more effortful deliberative thinking. In turn, circadian incongruence encourages more heuristic and less effortful thinking, and hence lead people to unquestionably rely on established beliefs and stereotypic preconceptions. In light of the importance of circadian congruence for effortful deliberation, we hypothesized that belief change should be greater in a task where people are asked to revise a prior belief in light of new conflicting evidence at a time congruent with their circadian preference.

If depleted cognitive resources lead to belief perseverance, augmenting those resources may counteract the effect of circadian variations. We tested this possibility by examining the impact of socially distributed cognition on belief perseverance. From a distributed cognition perspective, cognitive functioning is conceived as taking place in a system including resources and operations that are distributed across time, material artefacts as well as people (Hollan, Hutchins, \& Kirsh, 2000; Villejoubert \& ValléeTourangeau, 2011). From this perspective, interactivity acts as a cognitive scaffold, resulting in improved performance. For example, we showed that when participants could interact with physical matchsticks in a matchstick algebra problem, they were more likely to achieve insight compared to those for whom matchsticks were drawn on paper (Weller, Villejoubert, \& Vallée-Tourangeau, 2011; see also Vallée-Tourangeau, Euden, \& Hearn, 2011).

Unlike material distribution, the contribution of socially distributed resources for performing a cognitive task remains debated, however; with wealth of evidence showing that performance may be both weakened or strengthened when cognition is shared in a group (Larson, 2010). Group performance may vary depending on the fit between members' cognitive resources and the cognitive demands of the task, how resources are distributed, and process costs arising from group interactions (Steiner, 1972). On the one hand, groups may benefit from the potential to generate a more diverse range of interpretations and counterexplanations than would individuals (Hutchins, 1991). For example, multiple-cue judgments were shown to be more accurate when they originated from dyads rather than individuals (Olsson, Juslin, \& Olsson, 2006). Yet, it is not clear that groups will always be in a better position to engage in deliberative thinking as the superiority of groups may also depend on the cognitive resources they have at their disposal (Hutchins, 1991). This suggests, for example, that groups composed of individuals with limited cognitive resources may function worse than individuals, exhibiting more heuristic thinking and "groupthink" (Janis, 1982) whereas groups composed of members at the peak of their cognitive functioning may outperform individuals.

Our study was therefore also designed to examine whether the amount of cognitive resources available to individuals in a group (manipulated through circadian variations) would affect group performance. We expected that groups would exhibit less belief perseverance than individuals when they were made of individuals tested at their best period of circadian arousal. Conversely, we expected belief perseverance to be more pronounced in groups than in individuals when groups were composed of individuals tested at an incongruent circadian time.

## The Present Study

The present study investigated the cognitive processes that underpin belief perseverance by examining the moderating role of circadian variations and socially distributed thinking. It used a forensic scenario where participants were asked to revise their initial judgment of the extent to which a stereotypical suspect was guilty of an offense, after being presented with counterevidence in the form of an atypical offender profile written by an expert profiler. Guilt judgments were produced either by small groups of three participants or by individual participants. Half of the individual participants and groups were tested at a time that was congruent with their optimal circadian preference; the rest were tested at an incongruent time. Building upon previous research on the role of circadian variations on thinking mode, we expected that people who were tested at an incongruent time would exhibit greater belief perseverance because their limited cognitive resources should favor a heuristic mode of thinking. We also explored the role of socially distributed thinking on belief perseverance and hypothesized that the relative performance of groups compared to individuals would depend on
whether or not groups were composed of members functioning at their optimal time of circadian arousal: groups of individuals tested at a congruent time were expected to revise their guilt judgment to a greater extent than individuals while groups of individuals tested at an incongruent time were expected to persevere more in their belief of guilt compared to individuals.

## Method

## Participants

A total of 129 students and administrative staff were recruited on the campus of Kingston University. They were either tested individually ( $N=32$, Mean age $=29$ years, $S D$ $=12.85,22$ women) or in one of 32 small groups made of three to four individuals (see Table 1 for group demographics).

Table 1: Group demographics.

| Group type | $N$ | Mean age $(S D)$ |
| :--- | ---: | :---: |
| Women only | 5 | $25(9.33)$ |
| Men only | 3 | $22(2.46)$ |
| Mixed $(2 \mathrm{~W} / 1 \mathrm{M})$ | 12 | $23(6.65)$ |
| Mixed $(1 \mathrm{~W} / 2 \mathrm{M})$ | 12 | $23(5.60)$ |
| Total | 32 | $23(6.48)$ |

## Design

The experiment used a $2 \times 2 \times 2$ mixed design. The between-subject independent variables were the time of testing (circadian-congruent or circadian-incongruent) and testing condition (individually or in small groups). The within-subject independent variable was the time of judgment (before or after the presentation of disconfirming information). Participants were randomly allocated to one of the resulting four conditions.

## Procedure

Participants were invited to take part in a study examining how jurors make decisions in various circumstances. Participants' circadian type was assessed using the abridged English Version of Morningness-Eveningness questionnaire (rH\&O, Chelminki et al., 2000). They were categorized as either Morning (M) types or Evening (E) types on the basis of a median split of their scores. M-types scored significantly higher on the Morningness-Eveningness dimension compared to E-types; $M_{\text {M-types }}=17.00(S D=$ 3.36), $N=34, M_{\mathrm{E}-\mathrm{types}}=11.57(S D=1.36), N=30, t(64)=$ 7.31, $p<.001$. All participants were then reconvened to complete a small questionnaire. Half of the participants were tested at a time that was congruent with their circadian preferences (M-types tested between 10am and 12noon and E-types tested between 1 pm and 5 pm ) while the remaining half was tested at an incongruent time. Participants were asked, on their own or in a small group, to read a brief description of a criminal case involving a series of sexual assaults against young girls, followed by a stereotypical
description of a suspected child molester (e.g., a 44-year-old white male, unemployed, lonely and morally deviant). Lastly, they read an atypical offender profile by a forensic expert, which listed characteristics informed by actual statistics for this kind of offender although at odds with the stereotypical suspect description (e.g., "In most crimes of this nature the offender is employed in some form of skilled or office job"; see Marshall and Alison, 2007, for the complete descriptions). Participants tested in groups were invited to read the case information on their own and thereafter discuss the case between themselves before reporting a unique group estimate for each of the variables measured.

## Measures

Participants were asked to consider the case and rate the degree to which the suspect may be guilty and the degree to which they felt confident that their judgment was correct ( $1=$ not at all, $10=$ completely ) both before and after the presentation of the atypical profile. Finally, participants were asked to rate their level of involvement in the case ( $1=$ none, $10=$ greatly) and how difficult it was to make a decision ( $1=$ not at all, $10=$ very ) before they were debriefed and thanked for their participation.

## Results

## Manipulation Checks

Participants tested at a congruent circadian time reported a slightly higher level of involvement in the task, $M_{\text {congruent }}=6.55, S D=2.36$ vs. $M_{\text {incongruent }}=5.88, S D=2.56$, as well as higher levels of difficulty, $M_{\text {congruent }}=6.23$, $S D=2.35$ vs. $M_{\text {incongruent }}=5.74, S D=2.21$. Possibly due to lack of statistical power, however, neither difference reached statistical significance; $t(62)=1.08, p=.14$, onetailed, Cohen's $d=0.27$ for involvement, and $t(62)=0.87, p$ $=.19$, one-tailed, Cohen's $d=0.22$ for difficulty.

## Guilt Judgments

The theoretically important patterns in these data are (i) the effect of time of judgment (before or after the presentation of the profile), which captured the degree to which participants, regardless of circadian congruence or grouping, changed their guilt ratings after seeing the atypical offender profile; (ii) the interaction between circadian congruence and time of ratings (see Fig. 1); (iii) the interaction between group and time of ratings (see Fig. 2); and finally (iv) the interaction between circadian congruence and group (see Fig. 3).

Guilt judgment data were analyzed with a 2 (circadiancongruent vs. circadian-incongruent) $\times 2$ (individual vs. group) $\times 2$ (before vs. after the atypical profile presentation) mixed analysis of variance (ANOVA). The main effect of time of testing (congruent or incongruent with circadian preferences) was not significant, $F<1$, nor was the main effect of testing condition (individually or in small group), $F(1,60)=1.25$. However, there was a significant main
effect of the time of judgment, $M_{\text {before }}=7.26(S D=1.52) \mathrm{vs}$. $M_{\text {after }}=5.44(S D=2.07), F(1,60)=47.23, M S E=2.35, p<$ $.001, \eta_{\mathrm{p}}^{2}=.44$. There was also a significant interaction between circadian time of testing and the time of judgment, as Figure 1 illustrates, $F(1,60)=4.65, p=.04, \eta_{\mathrm{p}}^{2}=.07$. Guilt judgments were significantly revised downwards both at circadian-congruent times, $t(29)=6.02, p<.001$, and at circadian-incongruent times, $t(33)=3.46, p=.002$. However, as anticipated, the difference in pre- and postprofile judgments was significantly smaller, indicative of more belief perseverance, under circadian-incongruent time of testing, $t(62)=-2.13, p=.02$, one-tailed.


Figure 1: Mean judgments of the extent to which the suspect may be guilty before and after the presentation of the atypical offender profile, as a function of time of testing (circadian congruent or incongruent)


Figure 2: Mean judgments of the extent to which the suspect may be guilty before and after the presentation of the atypical offender profile, as a function of testing condition (individual or group)

The interaction between time of judgment (before or after the presentation of the profile) and testing condition was also marginally significant, $F(1,60)=3.44, p=.07, \eta_{\mathrm{p}}^{2}=$ .05. Both judgments made by individuals, $t(31)=3.20, p=$ .003 , and those made by groups, $t(31)=6.36, p<.001$, were significantly revised downwards after the presentation of the atypical profile (see Fig. 2). However, the amount of belief revision was larger in group judgments, $t(62)=1.81, p=$ .04, one-tailed.

Of less theoretical interest, the effect of testing condition on guilt judgments collapsed across time of testing was also moderated by circadian congruency, $F(1,60)=4.50, M S E=$ $3.89, p=.04, \eta_{p}^{2}=.07$ (see Fig. 3). Unplanned post-hoc tests (with a Bonferroni-corrected $\alpha$ set at .0125) revealed that testing condition did not affect overall guilt judgments when participants were tested at a congruent time, $t(28)=-$ .61. However, guilt judgments made individually were significantly higher than those made in groups when participants were tested at an incongruent time, $t(32)=2.72$. These results show that individual and group guilt judgments collapsed across time of testing were indistinguishable when produced at circadian congruent times whereas individual judgments of guilt were more pronounced than group judgments at incongruent circadian times.


Figure 3: Mean overall judgments of the extent to which the suspect may be guilty as a function of time of testing and testing condition

Finally, the three-way interaction term was not significant, $F<1$. As Figure 4 illustrates, the effect of circadian time congruency on belief revision was the same for individuals and for groups, albeit groups tended to revise their judgments to a greater extent compared to individuals, in line with the findings reported above.


Figure 4: Mean proportion of belief change as a function of time of testing and testing condition

## Discussion

This study aimed to shed new light on the mitigated success of the counterexplanation approach to reduce belief perseverance. We proposed that belief perseverance originated from a heuristic mode of cognitive processing whereas belief revision demanded an effortful, deliberative processing of the task information, more taxing in cognitive resources. This assumption led us to hypothesize that belief perseverance would be less likely to occur in situations where cognitive resources were unconstrained. We examined two such situations: times when individuals' period of circadian arousal was at its peak and situations when individuals' cognitive resources may be augmented by socially distributed thinking.

Our findings confirmed that the degree of belief persistence in the suspect's guilt after the profile presentation was moderated by circadian congruency: when tested at a congruent time, participants exhibited less belief persistence-that is, they revised their guilt judgment to a greater extent-than when they were tested at an incongruent time. Contrary to what one might expect from the groupthink perspective, but in line with predictions from the distributed cognition perspective, group judgments tended to exhibit less belief perseverance in the suspect's guilt after presentation of the atypical profile-judgments made in groups were revised more substantially than individual judgments. There was, however, no interaction between circadian time of testing and testing condition: group judgments exhibited less belief perseverance than individual judgments, both at congruent and incongruent circadian times of testing.

These findings contribute to our understanding of the cognitive mechanisms that may underpin belief perseverance. They suggest that belief perseverance is not only a consequence of the content of thought-for example, the availability of reasons for or against a target belief or a counterexplanation (Anderson, 1982, 2008)-but is also influenced by the cognitive resources available to individuals, which in turn determine the mode of thought they can apply to the task. As such, this study provides empirical support for the claim that belief perseverance arises from a heuristic mode of thinking since perseverance was more marked when cognitive resources were limited, thus inhibiting a more effortful and deliberative processing of the task. More importantly, simply increasing the pool of cognitive resources available to process the task-either by testing individuals at their peak circadian time or by allowing them to distribute cognitive resources in a social system-was sufficient to significantly reduce belief perseverance.

The fact that we found no evidence of "groupthink" when groups were tested at incongruent circadian times may suggest that the increased pool of cognitive resources offered by socially distributed thinking remained sufficient to counteract heuristic thinking. Distributed thinking enhances cognitive power by, notably, lowering the cost of sense making. According to Kirsh (2010), distributed
thinking involves three cost structures: the cost of mental operations, the cost of outer operations-in the present research, most typically exemplified by the speech acts (Austin, Urmson, \& Sbisà 1975) performed by the group members-and the coupling cost of coordinating these inner and outer processes. This suggests in turn, that the superior performance of groups tested at incongruent circadian times occurred because the benefits of socially distributing thinking continued to outweigh the cost of inner cognitive processes and the coupling costs. Future research may shed light on this possibility by increasing coupling costs. One strategy to do so could be to distribute the information about the suspect and the profile between group members as opposed to present all information to all members, as was the case in this study; this would require group members to engage in the coordination of evidence, and this added cost might eliminate the superior performance of groups in incongruent circadian conditions.

Finally, our findings also have implications for past accounts of belief perseverance. They suggest that instructions to consider counterexplanations may succeed by inviting deeper processing of the belief revision task and may fail when the cognitive cost of this type of processing is either too high or when individuals' cognitive resources are depleted. Alternative accounts (e.g., Nestler, 2010; Sanna, Schwarz, \& Stocker, 2002) have suggested that judgments are mediated by metacognitive feelings of difficulty. For instance, Nestler (2010) suggested that individuals infer the likelihood of the truth of an outcomebe it the target belief or the counterexplanation-from the difficulty they experience in generating many reasons for (or against) this outcome. This metacognitive explanation does not seem to be supported by our present results: if anything, individuals tended to perceive the task as more difficult when they were tested at congruent circadian times and yet, this was also when they exhibited less belief perseverance.

Future research may thus benefit from disentangling the respective impact of metacognitive feelings of difficulty and mode of thinking. Specifically, an alternative account of Nestler's (2010) findings could be made in terms of the cognitive demands of the task and their impact on the mode of thinking elicited. Prompting for a few reasons in favor of the target belief may be the least demanding and thus unlikely to engage a deliberative mode of thinking. Prompting for many reasons supporting the target belief or prompting for a few reasons supporting the alternative hypothesis might be sufficient to engage more deliberative processes and, as a result, reduce belief perseverance. But prompting for many reasons supporting the alternative explanation may be too taxing in cognitive resources: it entails holding both the target and alternative hypothesis in working memory while also exerting efforts to find a large amount of evidence for the alternative hypothesis. Unless they are motivated to do so, people will more naturally consider multiple evidence in support of one hypothesis rather than establishing the diagnostic value of a single
piece of information for two hypotheses (Villejoubert \& Vallée-Tourangeau, 2012). The increasing demand on cognitive resources in this instance could thus have led to cognitive overload, causing people to revert to a heuristic mode of thinking and, as a result, exhibit more belief perseverance.

To conclude, employing circadian congruence as a proxy for the cognitive resources available to perform a judgment task, this research suggests that belief perseverance results from an intuitive mode of thinking. Contrary to what might be anticipated by the groupthink perspective, however, depleted cognitive resources in a group setting did not affect belief perseverance. This suggests that socially distributed thinking may help to counteract the detrimental effect of cognitive depletion.

## Acknowledgments

We thank Rupreet Gurm for her assistance in collecting data.

## References

Adams, R. (2011, April 27). Obama birth certificate released to put end to 'carnival barkers': US president says in TV statement he is puzzled at how questions over his place of birth have persisted. The Guardian. Retrieved from
http://www.guardian.co.uk/world/2011/apr/27/barack-obama-birth-certificate-released?intcmp=239
Anderson, C. A. (1982). Inoculation and Counterexplanation: Debiasing Techniques in the Perseverance of Social Theories. Social Cognition, 1, 126-139.
Anderson, C. A. (2007). Belief Perseverance. In Encyclopedia of Social Psychology (pp. 110111). Thousand Oaks, CA: SAGE Publications.

Anderson, C. A., \& Lindsay, J. J. (1998). The development, perseverance, and change of naive theories. Social Cognition, 16, 8-30.
Anderson, C. A., Lepper, M. R., \& Ross, L. (1980). Perseverance of social theories: The role of explanation in the persistence of discredited information. Journal of Personality and Social Psychology, 39, 1037-1049.
Austin, J. L., Urmson, J. O., \& Sbisà, M. (1975). How to do things with words. Cambridge, MA: Harvard University Press.
Bodenhausen, G. V. (1990). Stereotypes as judgmental heuristics: Evidence of circadian variations in discrimination. Psychological Science, 1, 319-322.
Darlow, A. L., \& Sloman, S. A. (2010). Two systems of reasoning: Architecture and relation to emotion. Wiley Interdisciplinary Reviews: Cognitive Science, 1, 382-392.
Hollan, J., Hutchins, E., \& Kirsh, D. (2000). Distributed cognition: Toward a new foundation for human-computer interaction research. ACM Transactions on ComputerHuman Interaction, 7, 174-196.
Hutchins, E. (1991). The social organization of distributed cognition. In L. B. Resnick, J. M. Levine, \& S. D. Teasley (Eds.), Perspectives on Socially Shared Cognition (pp.

283-287). Washington, DC: American Psychological Association.
Janis, I. L. (1982). Groupthink: Psychological studies of policy decisions and fiascoes (2nd ed.). Boston: Houghton Mifflin.
Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. American Psychologist, 58, 697-720.
Kirsh, D. (2010). Thinking with external representations. $A I$ \& society, 25, 441-454.
Larson, J. R. (2010). In search of synergy in small group performance. New York: Psychology Press.
Marshall, B. C., \& Alison, L. J. (2007). Stereotyping, congruence and presentation order: Interpretative biases in utilizing offender profiles. Psychology, Crime \& Law, 13, 285-303.
Nestler, S. (2010). Belief perseverance: The role of accessible content and accessibility experiences. Social Psychology, 41, 35-41.
Olsson, A.-C., Juslin, P., \& Olsson, H. (2006). Individuals and dyads in a multiple-cue judgment task: Cognitive processes and performance. Journal of Experimental Social Psychology, 42, 40-56.
Ross, L., Lepper, M. R., \& Hubbard, M. (1975). Perseverance in self-perception and social perception: Biased attributional processes in the debriefing paradigm. Journal of Personality and Social Psychology, 32, 880892.

Schmidt, C., Collette, F., Cajochen, C., \& Peigneux, P. (2007). A time to think: circadian rhythms in human cognition. Cognitive Neuropsychology, 24, 755-789.
Steiner, I. D. (1972). Group process and productivity. New York: Academic Press.
Tversky, A., \& Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185, 11241131.

Vallée-Tourangeau, F., Beynon, D. M., \& James, S. A. (2000). The role of alternative hypotheses in the integration of evidence that disconfirms an acquired belief. European Journal of Cognitive Psychology, 12, 107-129.
Vallée-Tourangeau, F., Euden, G., \& Hearn, V. (2011). Einstellung defused: Interactivity and mental set. The Quarterly Journal of Experimental Psychology, 64, 18891895.

Villejoubert, G., \& Vallée-Tourangeau, F. (2012). Relevance-driven information search in "pseudodiagnostic" reasoning. The Quarterly Journal of Experimental Psychology, 65, 541-552.
Villejoubert, G., \& Vallée-Tourangeau, F. (2011). Constructing preferences in the physical world: a distributed-cognition perspective on preferences and risky choices. Frontiers in Psychology, 2, 1-3.
Weller, A., Villejoubert, G., \& Vallée-Tourangeau, F. (2011). Interactive insight problem solving. Thinking \& Reasoning, 17, 424-439.

# How does emotional content affect lexical processing? 

David Vinson, Marta Ponari \& Gabriella Vigliocco<br>\{d.vinson, m.ponari, g.vigliocco\}@ucl.ac.uk<br>Cognitive, Perceptual and Brain Sciences, University College London<br>Gower Street, London WC1E 6BT United Kingdom


#### Abstract

It is now generally accepted that words' emotional content plays a role in lexical processing, but the literature offers incompatible findings concerning what this role may be. Here we use a large sample of lexical decision data (British Lexicon Project, Keuleers et al., 2012) and we carry out a series of analyses differing in the way emotional variables are treated. A variety of statistical approaches yielded common conclusions: when confounding variables are taken into account, emotional words, whether positive or negative, are processed faster than neutral words. This effect is categorical rather than graded; is not modulated by emotional arousal; and is not limited to words explicitly referring to emotions. We discuss this in terms of internally grounding words' meanings in emotional experience, akin to the manner in which concepts may be grounded in perception and action.


Keywords: emotion; valence; lexical decision.

## Introduction

In mainstream lexical processing studies, emotional content has been largely ignored, whether considered irrelevant to the core meanings of words, or as properties of narrowly defined sets of words explicitly referring to emotions (e.g. Altarriba \& Bauer, 2004). Recently, however, a number of studies of lexical processing effectively demonstrated that emotional content plays a role even in shallow tasks involving single words such as lexical decision (e.g. Estes \& Adelman 2008a,b; Kousta, Vinson \& Vigliocco, 2009; Kousta, et al., 2011; Larsen, et al., 2008).

As a result, language processing researchers have begun to acknowledge the interplay between emotion and language processing systems, discussing emotional effects in language processing as due to the embodied nature of linguistic representations (e.g. Kousta et al., 2011; Moseley, et al., 2012; Vigliocco et al., 2009), just as researchers in other domains of cognition have posited embodied emotional effects (e.g. Pistoia et al., 2010). However, precisely which mechanisms are involved in emotional processing is unclear at the present. This is because different studies of lexical processing have found different and apparently incompatible results even when the same task (e.g., lexical decision) is used.

It has been shown that previously reported effects of emotional valence (i.e. numeric ratings indicating the extent to which a word is positive, neutral or negative) can change dramatically once confounding variables such as length, frequency and orthographic neighbourhood size are taken into account (Larsen, Mercer \& Balota, 2006). However, even after controlling for non-emotional variables, results are conflicting. Estes \& Adelman (2008a,b) and Larsen et al (2008) reported slower lexical decision reaction times (RTs)
for negative than positive words. This has been interpreted in terms of attentional vigilance: heightened and/or extended attention to negative stimuli (e.g. Fox et al., 2001; Pratto \& John, 1991) which would slow any decision (such as lexical decisions) on other aspects of the stimuli. In contrast, Kousta et al. (2009) found a processing advantage for both negative and positive over neutral words, which they explain in terms of greater motivational relevance of emotionally loaded stimuli (e.g. Lang, Bradley \& Cuthbert, 1997). Kousta et al argued that the discrepancy in findings was due to the relative lack of neutral words in the data sets tested previously, but especially due to the lack of control of potentially confounding variables, such as ratings of familiarity and age of acquisition (AoA) in previous studies.

In addition, Larsen et al (2008) found that the effect of valence was modulated by the arousal of words such that a negative disadvantage was present for medium-low arousing words, but no effect was observed for highly arousing negative words, Estes and Adelman (2008b) argued for a far more constrained role of arousal, and Kousta et al. (2009) argued that valence effects could not be explained in terms of arousal (although these authors did not explicitly test for valence $\times$ arousal interactions).

All of these previous studies were conducted using lexical decision data from a single source: the English Lexicon Project (ELP, Balota et al., 2007), so in addition to questions about the different assumptions and approaches taken by previous authors, one may also wonder about the extent to which the findings may be related to quirks of that particular item set. Here we take advantage of an entirely independently obtained large-scale set of lexical decision data (British Lexicon Project (BLP): Keuleers, Lacey, Rastle \& Brysbaert, 2012), to try and resolve these questions. Our analyses compare models based on different a priori theoretical assumptions concerning the role of valence in word processing, controlling non-emotional variables known to affect lexical decision RTs. We begin by fitting baseline models in which all the non-emotional predictors mentioned above are taken into account, then add specific terms embedding different assumptions about the role of valence. Such an approach is essential in order to test theoretical accounts of emotion effects in lexical processing.

After assessing how well different measures of valence perform after taking baseline variables into account, we move on to evaluating the role of other aspects of emotional content besides just valence, assessing the extent to which valence effects may instead be explained or modulated in terms of arousal. Finally, we test whether the effects of emotional valence differ for words specifically referring to emotional experience, vs. words that are only valenced.

## Materials and methods

## Data

From the full set of words in the BLP, we selected those 1374 words for which valence ratings were available from ANEW (Bradley \& Lang, 1999), or from the additional ratings described in Kousta et al., 2009, 2011). Next, we filtered out those words for which BLP participants were extremely inaccurate on making lexical decisions: those with overall accuracy less than $67 \%$ in the BLP ( $n=56$, e.g. larkspur, dryad, godhead). This is an important step not employed in previous studies, as widely unfamiliar words are likely to elicit slow RTs, and to receive neutral valence ratings from participants. Finally, we removed five words for which concreteness and imageability ratings were not available, leaving 1313 words for analysis. Of these, 856 were in common with the set from the ELP which Kousta et al. (2009) analysed.

## Measures of emotional valence

We centred the scale of the original valence ratings which ranged from 1-9, so as to range from -4 (most negative) to +4 (most positive) with 0 reflecting neutrality. We then created the following measures that embed different theoretical assumptions concerning valence. The most essential distinction concerns the direction of valence effects as this differentiates between highly distinct accounts of emotion processing. If the crucial distinction is between negative words and other words, this would favour attentional vigilance or other negativity-bias accounts of emotion processing; but if instead the crucial distinction is between emotionally valenced and neutral words, it would favour motivational accounts of emotion processing.
In addition, we compare models in which valence is considered as a continuous measure, vs. models in which it is discretized, as a test of previous claims that effects of emotion should be considered all-or-nothing (e.g. Estes \& Adelman, 2008a,b).

Continuous valence These measures treat valence as a graded measure varying from most negative (-4) through neutral (0) to most positive ( +4 ).
Linear measure includes only the linear relationship between valence and RT. If negative words are slower than other words (e.g. Estes \& Adelman, 2008a,b; Larsen et al., 2008), we expect to find a negative slope (RTs decrease with increasing valence).
Polynomial measure includes linear and quadratic components of valence. ${ }^{1}$ If valenced words are faster than neutral words with no difference between positive and

[^245]negative (e.g. Kousta et al., 2009) we expect a negative quadratic coefficient while the linear coefficient would offer no further benefit.

Discrete valence These measures treat valence as categorical rather than continuous/graded.
Negative/positive measure includes two discrete valence classes: negative (valence $<0$ ) and positive (valence $>=0$ ) valence levels. If negative words are slower than other words, these two categories should differ. This model is the simplest discrete counterpart to the linear measure above, and was preferred by Estes and Adelman (2008a) as more complex measures tested did not account for the data better. Valenced/neutral measure treats positive and negative as a single class, compared to neutral (emotional: |valence| > 1.5; neutral: |valence| <= -1.5). If emotional words are faster than neutral words we expect to find differences between these two categories (as we would for the quadratic term of the polynomial measure).

## Design and analysis

We fit a variety of linear mixed-effects models described in more detail below, in each case testing for a partial effect of valence on lexical decision latencies, using any of the four proposed valence measures. We conducted our analyses on log-transformed RT (excluding error trials) then replicated the same analyses on untransformed RT to be sure that log-transformation did not produce anomalous patterns of results. Additionally, we fit models not only to trial-level data but also to item averages, for both $\log (\mathrm{RT})$ and untransformed RT. ${ }^{2}$

Analysis of trial level data was carried out using linear mixed-effects models (packages lme4: Bates \& Maechler, 2009; and languageR: Baayen, 2009; cf. Baayen, 2008) in the R programming environment ( R Development Core Team, 2009). Model fits included random intercepts for both subjects and items, as well as random slopes by subjects (for emotional predictors only, which are constant for each item). Analysis of item averages was carried out using ordinary least squares regression. Below we focus upon the results for trial-level analyses of $\log (R T)$ but across the board the findings are comparable for analyses of untransformed RT and/or item averages.

In all of the analyses we conduct upon valence measures, we always begin with a baseline model in which we consider the following non-emotional factors that were

[^246]controlled in all the previous studies we have mentioned: number of letters; $\log (H A L$ frequency), orthographic neighbourhood size (all from Balota et al, 2007); We also included additional non-emotional predictors controlled by Kousta et al. (2009) and which those authors argued to be essential in order to unambiguously interpret effects as emotional in nature: mean positional bigram frequency (Balota et al., 2007); ratings of concreteness, imageability and familiarity (Coltheart, 1981) and age of acquisition ratings (Stadthagen-Gonzalez \& Davis, 2006). As a result, our tests of emotional variables provide results that can be unambiguously attributed to emotion rather than other characteristics of words with which emotional properties may be confounded.

## The role of arousal

Some previous studies have shown that effects of valence are modulated by arousal (Estes \& Adelman, 2008b; Larsen et al., 2008, but see Kousta et al. 2009). Using a similar modelling approach as above, we test the role of arousal in two ways. First, we treat arousal as a categorical measure (high arousal words vs. low arousal words), testing valence $\times$ arousal interactions for any of the valence measures described previously which turn out to be significant predictors of lexical decision RT. If arousal modulates the effect of valence we should see such an interaction. Second, we treat arousal as a control variable, testing in a different set of models whether unique effects of valence can be observed after variation related to arousal is taken into account. This is particularly important for models distinguishing valenced from neutral words (i.e. quadradic term of the polynomial measure, and valenced/neutral measure) as valenced words also exhibit a strong tendency to be more arousing as well (Bradley \& Lang, 2000).

## Emotion words vs. emotionally valenced words

One essential factor that has been neglected so far in large-scale studies of emotion in lexical processing is whether any valence effect is being driven by a specific, limited set of words: those referring explicitly to emotion (e.g. love, shame, joy, hate in contrast to valenced words not directly referring to emotions, e.g. prison, cake, justice, cheat). For example, Altarriba and Bauer (2004) argue that emotion words are sufficiently different to other types of words that we ought to consider words as falling into three categories: concrete, abstract and emotion words. Moreover, it has been argued that emotion words may be embodied not only internally (via emotional experience, as argued by Kousta et al., 2011) but also due to body states associated with emotional experience (such as facial expression, posture etc., Moseley, et al., 2012). Such words tend to be prevalent in our vocabulary and even a cursory inspection of valence norms reveal many such items among the set. If these words alone are responsible for emotion effects, one cannot conclude that valence is relevant to lexical processing in general, as it may play a role only in the specific, tightly constrained domain of emotion words.

To address this issue, we used Wordnet-Affect (Strapparava \& Valitutti, 2004) to identify emotion words. Wordnet-Affect classifies words according to their organisation in Wordnet. Any word with an emotional sense is considered "emotional", thus this is a conservative classification. We hand-classified a few additional words as potentially emotional (e.g. courage, craven, stern) with 193 of the 1313 words classified as emotion words. To test whether emotion words alone are responsible for valence effects we fit models as above, testing the interaction between valence and emotion-word classification. If emotion words drive the effects observed we should see an interaction such that the valence effects are restricted to emotion words (or at least differ between emotion and nonemotion words).

## Results

## Fitting baseline models

It is no surprise that many of the non-emotional variables were significant predictors of lexical decision latencies, consistent with a wealth of previous studies. For the purposes of the present study we simply note here that higher-order polynomial transformations offered significant improvement in performance over linear-alone components for several of the predictors. Moreover, although some factors were not significant predictors in the baseline model (i.e., concreteness, imageability and summed positional bigram frequency) we retained them as (linear) predictors ${ }^{3}$ along with the following predictors that were significant in the (reduced) baseline model: 3-order polynomial transformations: (log frequency, number of letters, number of orthographic neighbours, familiarity); linear terms (age of acquisition) (see Figure 1).


Figure 1: Predictors in the baseline model (logRT; triallevel data). Dashed lines depict 95\% highest posterior density CI (parameter estimates). Similar patterns were observed for item-level analyses and untransformed RT.

[^247]
## Measures of emotional valence

We tested the effects of valence by adding each of the valence measures described above to the best-fit baseline model above, thus always allowing us to evaluate the partial effect of valence only after non-emotional variables were taken into account. We also added that same valence term in each model as a random slope by subjects (in analysis of trial-level data).

Table 1: Partial effects of the different valence measures (logRT, trial-level), taking non-emotional variables into account. The same patterns were observed for item-level analyses and for untransformed RT

| Valence measure | Estimate (Std err.) | $t$ statistic |
| :--- | ---: | :--- |
| Linear | $.00066(.00161)$ | 0.57 |
| Polynomial $^{4}$ |  |  |
| $\quad$ (linear term) | $.00052(.00116)$ | 0.45 |
| $\quad$ (quadratic) | $-.00158(.00076)$ | -2.07 |
| Negative/positive | $.00166(.00286)$ | 0.58 |
| Valenced/neutral | $-0.0067(.00357)$ | -1.87 |



| Linear | Quadratic | Positive- | Valenced- |
| :---: | :---: | :---: | :---: |
| (continuous) | (continuous) | Negative | Neutral |

Figure 2: Graphical depiction of parameter estimates of the different valence measures reported in Table 1. In all cases a value of 0 indicates no effect. For continuous measures, the values plotted reflect the slope (linear measure) or quadratic coefficient (polynomial measure). For categorical measures they represent the estimate of the difference between the two conditions (logRT). Horizontal
line $=$ mean parameter estimate. Box depicts 50\% confidence interval of the parameter estimate; whiskers depict 95\% confidence interval.

Only those measures in which valenced words differ from

[^248]neutral words (quadratic term of polynomial continuous measure, and valenced/neutral measure) were reliable predictors of lexical decision RT. Instead, the linear continuous measure did not predict RT once confounding factors were taken into account, and the same was true for the negative/positive measure. To assess whether the continuous (quadratic) measure offers sufficient additional explanatory power beyond the simplest categorical measure contrasting valenced to neutral words, we fit one additional set of models, in which we entered 2-order polynomial valence along with the categorical measure, and we compared the models using likelihood ratio tests. There was no significant improvement gained by adding this additional term (log likelihood ratio for valenced/neutral model = 7629.1; log likelihood ratio for combined model $=7630.0$; $\chi^{2}(5)=1.7972, p=.876$ with comparable results for itemlevel analyses and analyses of untransformed RTs.

At this stage the data suggest that the effect of valence is best described as a simple, categorical contrast between words with emotional associations and those without. Thus, when non-emotional variables are taken into account, we see that a categorical measure of valence, regardless of polarity, is sufficient to account for emotional effects in word processing.

## The role of arousal

Here we focus upon those valence measures which were reliable predictors in the previous section (i.e., 2-order Polynomial and Valenced/Neutral), assessing whether they can be accounted for, or modulated, by arousal.

First, we tested the interaction between arousal and each of the two valence measures (continuous and categorical). For these analyses we discretized arousal, using a median split to characterise words as low or high arousal (contrast coded). For trial-level analyses we included both main effects and the interaction as random slopes by subjects. We found that the main effect of valence persisted, with no effect of arousal category and no interaction between the two. For $\log$ RT and trial level analysis: quadratic coefficient estimate $=-.00338(\mathrm{SE}=.00110), t=-3.067$, arousal main effect and interaction $|t|<1.2$; categorical coefficient estimate $=-.0135(\mathrm{SE}=.0050), t=-2.725$, arousal main effect and interaction $|t|<1$ (with item level analyses and analyses of untransformed RTs showing the same pattern).

Next, we instead considered arousal as a continuous measure of arousal into the models, testing whether a partial effect of a valence measure could still be seen after arousal was taken into account. For trial-level analysis this meant including random slopes by subject for arousal as well. We started by adding arousal to the baseline model described above. When arousal was the only emotional variable included, its effects were significant (estimate of the slope $=$ -. 0050 ( $\mathrm{SE}=-.0020, t=-2.518$ ): more arousing words elicited faster responses. We then added a valence measure to this baseline+arousal model. For both the polynomial and the categorical valence measure, effects persisted once
arousal was taken into account. For log RT and trial level analysis: quadratic coefficient estimate $=-.00261$ ( $\mathrm{SE}=$ .00144), $t=-2.627$; categorical coefficient estimate $=-.0112$ (SE = .0046), $t=-2.410$ ), with the partial effect of arousal not reaching significance in either case ( $|t|<1$ ), a finding replicated in item-level analyses and analyses of untransformed RTs. These effects of emotion can thus be attributed to valence rather than arousal.

## Emotion words vs. emotionally valenced words

As in the second set of analyses considering the role of arousal, we tested whether the effects of valence described above were different for emotion words and those not referring to emotional states (using Wordnet-Affect, Strapparava \& Valitutti, 2004), by testing for statistical interactions.

Just like our analyses involving arousal, the main effect of valence was unchanged, with no effect of Wordnet-Affect category and no interaction. For log RT and trial level analysis: quadratic coefficient estimate $=-.00197$ ( $\mathrm{SE}=$ .00085), $t=-2.31$, Wordnet-Affect category main effect and interaction $|t|<1$; categorical coefficient estimate $=-.00969$ (SE = .00419), $t=-2.31$, Wordnet-Affect category main effect and interaction $|t|<1.02$ (again with item level analyses and analyses of untransformed RTs showing the same pattern).

## Discussion

Our analyses show a reliable, consistent and rather simple pattern of emotion effects in lexical processing. Once potentially confounding variables are taken into account, lexical decisions to emotionally valenced words are recognised faster than those to neutral words. This finding differs from some previous studies (Estes \& Adelman, 2008a,b; Larsen et al., 2008): those investigating ELP data, using a more limited set of words (from ANEW, Bradley \& Lang, 1999) and crucially, for which some important control variables are unavailable. Those studies also conducted analysis over item averages only, allowing the possibility that valence effects observed there may have been magnified or distorted as a consequence of treating these values as point estimates rather than varying by subjects. Our results also differ from those reported by Kousta et al. (2009) although consistent with their overall conclusions. There appears to be no benefit in considering valence as a continuous measure: the 2-order polynomial valence model is no better than the simplest categorical model (valenced vs neutral). In fact when we reanalyze their data set, there too we find that a categorical measure contrasting valenced to neutral words is comparable to the continuous measure they favoured.

We also found this categorical effect of valence to be general in nature: it is not modulated by arousal, and it is not driven by words specifically referring to emotional experience. This finding resonates with recent neuroimaging evidence using a highly controlled set of words, in which activation in rostral anterior cingulate cortex (an area
associated with emotion processing) is modulated by valence (regardless of whether it is positive or negative) and not by arousal (Vigliocco et al., 2013).

Why would emotional content, regardless of polarity, facilitate lexical processing? Under general motivational accounts of processing (Lang, Bradley \& Cuthbert, 1997) both negatively and positively valenced items are relevant to survival and wellbeing albeit for different reasons. Crucial in this regard is the involvement of emotion processing systems even for lexical stimuli which do not exhibit obvious low-level visual characteristics argued to be evolutionarily linked to positive or negative emotions (vs. emotional expressions or visual properties of dangerous entities). In a recent proposal, the involvement of emotional systems has been argued to provide a means for grounding abstract concepts in internal experience, just like concrete concepts are accepted to be grounded in sensory-motor experience (Kousta, Vigliocco, Vinson, Andrews \& Del Campo, 2011; Vigliocco, Meteyard, Andrews \& Kousta, 2009). Under this view, emotional content of words would facilitate their processing, in a manner akin to the way in which sensory experience (operationalised as imageability or concreteness; Kousta et al. 2011) facilitates processing.

## Acknowledgments

Supported by UK Economic and Social Research Council grant RES-062-23-2012 to Gabriella Vigliocco.

## References

Altarriba, J., \& Bauer, L. M. (2004). The distinctiveness of emotion concepts: A comparison between emotion, abstract, and concrete words. American Journal of Psychology, 117, 389-410.
Baayen, R. H. (2008). Analyzing linguistic data: A practical introduction to statistics using $R$. Cambridge University Press.
Baayen, R.H. (2009). languageR: Datasets and functions with "Analyzing Linguistic Data: A practical introduction to statistics". http:/cran.r-project.org/package=languageR .
Baayen, R.H., Davidson, D.J. and Bates, D.M. (2008) Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language, 59, 390-412.
Balota, D.A., Yap, M.J., Cortese, M.J., Hutchison, K.A., Kessler, B., Loftis, B., Neely, J.H., Nelson, D.L., Simpson, G.B., \& Treiman, R. (2007). The English Lexicon Project. Behavior Research Methods, 39, 445459.

Bates, D. \& Maechler, M. (2009). lme4: Linear mixedeffects models using S4 classes. http:/cran.rproject.org/package=lme4.
Bradley, M.M. \& Lang, P.J. (1999). Affective Norms for English Words (ANEW): Stimuli, instruction manual and affective ratings (Tech Report C-1). Gainesville: University of Florida, Center for Research in Psychophysiology.

Bradley, M. M., \& Lang, P. J. (2000). Measuring emotion: Behavior, feeling, and physiology. Cognitive Neuroscience of Emotion, 25, 49-59.
Coltheart, M. (1981). The MRC Psycholinguistic Database, Quarterly Journal of Experimental Psychology, 33A, 497505.

Estes, Z., \& Adelman, J. S. (2008a). Automatic vigilance for negative words in lexical decision and naming: Comment on Larsen, Mercer \& Balota (2006). Emotion, 8, 441-444
Estes, Z., \& Adelman, J. S. (2008b). Automatic vigilance for negative words is categorical and general. Emotion, 8, 453-457.
Fox, E., Russo, R., Bowles, R. \& Dutton, K., (2001). Do threatening stimuli draw or hold visual attention in subclinical anxiety? Journal of Experimental Psychology: General, 130, 681-700.
Keuleers, E., Lacey, P. , Rastle, K., \& Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. Behavior Research Methods, 44, 287-304
Kousta, S.-T., Vigliocco, G., Vinson, D.P., Andrews, M. \& Del Campo, E. (2011). The representation of abstract words: Why emotion matters. Journal of Experimental Psychology: General, 140, 14-34.
Kousta, S.-T., Vinson, D. \& Vigliocco, G. (2009). Emotion words, regardless of polarity, have a processing advantage over neutral words. Cognition, 112, 473-481.
Lang, P.J., Bradley, M.M. \& Cuthbert, B.N. (1997). Motivated attention: Affect, activation and action. In P.J. Lang, R.F. Simons \& M.T. Balaban (Eds.), Attention and orienting: Sensory and motivational processes. Hillsdale NJ: Erlbaum.
Larsen, R.J., Mercer, K.A. \& Balota, D.A. (2006). Lexical characteristics of words used in emotional Stroop experiments. Emotion, 6, 62-72.
Larsen, R.J., Mercer, K.A., Balota, D.A. \& Strube, M.J. (2008). Not all negative words slow down lexical decision and naming speed: Importance of word arousal. Emotion, 8, 445-452.
Moseley, R., Carota, F., Hauk, O., Mohr, B. \& Pulvermüller, F. (2012). A role for the motor system in binding abstract emotional meaning. Cerebral Cortex, 22, 1634-1647.
Pistoia, F., Conson, M., Trojano, L., Grossi, D., Ponari, M., Colonnese, C., Pistoia, M.L., Carducci, F \& Sarà, M. (2010). Impaired conscious recognition of negative facial expressions in patients with locked-in syndroms. Journal of Neuroscience, 30, 7838-7844.
Pratto, F. \& John, O.P. (1991). Automatic vigilance: The attention-grabbing power of negative social information. Journal of Personality and Social Psychology, 61, 380391.

Stadthagen-Gonzalez, H., \& Davis, C.J. (2006). The Bristol Norms for Age of Acquisition, Imageability and Familiarity. Behavior Research Methods, 68, 598-605.
Strapparava, C. \& Valitutti, A. (2004). WordNet-Affect: An affective extension of WordNet, Proceedings of the 4th

International Conference on Language Resources and Evaluation (LREC 2004), 1083-1086.
Vigliocco, G., Meteyard, L., Andrews, M., \& Kousta, S.-T. (2009). Toward a theory of semantic representation. Language and Cognition, 1, 219-247.
Vigliocco, G. Vinson, D., Della Rosa, P., Kousta, S.T., Devlin, J. \& Cappa, S.F. (2013) The neural representation of abstract words: A role of emotion. In press, Cerebral Cortex.

# Does the Influence of Stress on Financial Risk Taking Depend on the Riskiness of the Decision? 

Bettina von Helversen (bettina.vonhelversen@unibas.ch)<br>University of Basel, Department of Psychology, Missionsstrasse 62a<br>4055 Basel, Switzerland<br>Jörg Rieskamp (joerg.rieskamp@unibas.ch)<br>University of Basel, Department of Psychology, Missionsstrasse 62a<br>4055 Basel, Switzerland


#### Abstract

Many decisions under risk and uncertainty are made under physical or emotional stress. Recent research suggests that stress influences decisions between risky options, but that the direction of the influence depends on the characteristics of the gambles. For instance, stress increases risk taking for loss gambles, but decreases risk taking for gain gambles. In the current project we investigate: (1) whether the riskiness of gambles influences the direction of the stress effect and (2) whether changes in risk taking can be linked to changes in attention. Participants who gave relatively more attention to gains than to losses, as indicated by eyetracking data, were more risk seeking than participants who gave less attention to gains. Stress did not influence participants' attention. However, stressed participants became more risk seeking when considering gambles with relatively low risk, but less risk seeking for gambles with relatively high risk.


Keywords: risk; decision making; stress; cortisol; variance

## Introduction

Every day we make decisions involving risk and uncertainty ranging from buying a gamble ticket to investing in stocks, gold, or real estate. Many of these decisions are not made in cold blood, but under physical or emotional stress. How stress and stress-related release of hormones such as cortisol influence risk preferences, however, is far from clear. Research has found that men, but not women, tend to become more risk seeking under stress (Lighthall, Mather, \& Gorlick, 2009; Preston, Buchanan, Stansfield, \& Bechara, 2007; Starcke, Wolf, Markowitch, \& Brand, 2008). Similarly, studies on financial risk taking have found divergent results. For instance, offering participants choices between risky and relatively safe options, Porcelli and Delgado (2009) found that participants became more risk seeking under stress when choosing between options involving losses, but less risk seeking when choosing between options involving gains. In a similar vein, Carr and Steele (2010) found that stereotype threat reduced risk taking in women. Von Dawans, Fischbacher, Kirschbaum, Fehr, \& Heinrichs (2012), however, found no influence of stress on decisions between gambles involving gains and losses.

Porcelli and Delgado (2009) argue that stress enhances decision biases such as the reflection effect (i.e., people are more risk seeking in the loss domain than in the gain domain, Kahneman \& Tversky, 1979). In the current research we follow up on this result, suggesting that stress enhances preexisting preferences for risk. That is, in decision situations in which people usually are risk seeking, they should become even more risk seeking under stress, whereas in decision situations in which people behave risk averse, they should become even more risk averse. We test these hypotheses in a financial risktaking task.

As a second goal we aimed to examine the mechanism underlying changes in risky decision making under stress. One mechanism by which stress could enhance preexisting preferences is by narrowing the focus of attention to the piece of information that is considered as most important. In line with this idea stress has been shown to reduce cognitive resources and narrow the focus of attention as well as the amount of information that can be processed (Friedman \& Förster, 2010; Kelly, Ashleigh, \& Beversdorf, 2007; Wichary \& Rieskamp, 2011). Thus, stress could influence risky decision making by changing the amount of attention given to the attributes of the choice options such as the possible outcomes (gains or losses) and the probability of the outcome (Ben Zur \& Breznitz, 1981).

## Variability in Outcomes as a Measure of Risk

The vast majority of research on financial risk taking involves the choice between gambles; that is, options with various outcomes that occur with a specific probability and that differ in valence (e.g., gains or losses). The risk of a gamble is commonly defined by the variability of the outcomes, with higher variability implying higher risk. For instance, finance models such as the capital-assespricing model equate risk with outcome variance (Sharpe, 1964). However, other variability measures such as the coefficient of variation, a measure based on the relative variance of a gamble, have been proposed to measure risk (Weber, Shafir, \& Blais, 2004). In sum, if stress amplifies people's risk preferences by narrowing their attention to the subjectively important aspect of the decision situation,
then stress should lead to more risk-taking behavior for gambles with little outcome variability and to less risk taking for gambles with high outcome variability.

## Attention in Risky Decision Making

The attention given to positive and negative attributes is an important predictor of decisions under risk. For instance, Ben Zur and Breznitz (1981) found that how often people looked at information about how much they could win or lose was related to their choices. This suggests that if stress narrows attention to the information the participant considers most important, increase in risky choices could be related to more attention being given to gains than to losses, whereas choice of safe options may be related to increased attention to losses over gains.

A non-intrusive way of measuring the relative attention given to gains or losses is by recording eye movements. In general, visual attention is strongly coupled to eye movements (e.g., Hoffman, 1998) and has been successfully used to understand the processes underlying decision making (Glaholt \& Reingold, 2011). In particular, two measures of eye movement have been successfully used to predict decisions. First, the time spent looking at an option is positively related to choosing this option (Glaholt, Wu, \& Reingold, 2009). Similarly, the time spent looking at specific pieces of information has been linked to the importance assigned to it (Rehder \& Hoffman, 2005). Secondly, choices are often reflected by gaze cascade effects; that is, over time attention wanders to the preferred option (Glaholt \& Reingold, 2011; Fiedler \& Glöckner, 2012). In particular, the last focus is related to choice; that is, the option fixated last before making a decision is chosen more frequently than other options (Krajbich, Armel, \& Rangel, 2010). Thus in the current study, we considered the time that gains and losses were looked at as well as the last information that was fixated before making a decision.

## The Study

We investigated the influence of stress on risk taking with a financial decision-making task consisting of 40 decisions between two gambles that contained positive and negative outcomes. Mixed gambles present an interesting problem, because increased risk taking with cortisol has been shown in particular when high gains and high losses were at stake (Putman, Antypa, Crysovergi, \& van der Does, 2010). Within the 40 gambles we varied the variability in the outcomes.

## Method

Participants. 70 participants ( 40 in the stress condition and 30 in the no stress condition, $M_{\text {Age }}=24.4, S D_{\text {Age }}=$ 5.3) were recruited at the University of Basel. We expected that for a substantial number (approximately one third) of the participants the cold pressor task would not result in an increase in cortisol. Therefore we collected more participants in the stress condition, to ensure a
sufficient sample size in the stress condition. 48 were females. Participants received a participation fee of 20 CHF per hour (approx. 22 US-\$). Additionally one of the participant's decisions was randomly chosen and the preferred gamble was played. Participants received/paid $10 \%$ of the gamble's outcome. One participant was excluded from the analysis because he always chose the reference gamble. Overall, testing took 1 h and 30 min .

Financial Decision-Making Task. The financial decision-making task consisted of 40 decisions between two gambles. In each trial participants chose between a reference gamble (Gamble A), in which participants could win 15 Swiss Francs (CHF) or lose 5 CHF with a probability of $.5(\mathrm{EV}=5 \mathrm{CHF})$, and a target gamble (gamble B). The reference gamble was the same in each decision, but there were 40 different target gambles structured in two sets: (1) high outcome gambles (e.g., win/lose 60 with a probability of .5 ) and (2) low outcome gambles (e.g., win/lose 30 with a probability of .5). For each gamble type (high or low outcome) we created sets of gambles by varying the expected value of the target gamble from -5 to $15 / 30$ in steps of 5 . The expected value was varied by changing either (1) the amount that could be won, (2) the amount that could be lost, or (3) the probability with which each outcome could occur (see Table 1 for an overview).

Table 1: Overview of the target gambles

| No | $p($ win $)$ | Gain | $p($ loss $)$ | Loss | EV | Set |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| 1 | .50 | 60 | .50 | -70 | -5 | high |
| 2 | .50 | 60 | .50 | -60 | 0 | high |
| 3 | .50 | 60 | .50 | -50 | 5 | high |
| 4 | .50 | 60 | .50 | -40 | 10 | high |
| 5 | .50 | 60 | .50 | -30 | 15 | high |
| 6 | .50 | 60 | .50 | -20 | 20 | high |
| 7 | .50 | 60 | .50 | -10 | 25 | high |
| 8 | .50 | 30 | .50 | -40 | -5 | low |
| 9 | .50 | 30 | .50 | -30 | 0 | low |
| 10 | .50 | 30 | .50 | -20 | 5 | low |
| 11 | .50 | 30 | .50 | -10 | 10 | low |
| 12 | .50 | 30 | .50 | -0.1 | 15 | low |
| 13 | .50 | 50 | .50 | -60 | -5 | high |
| 14 | .50 | 60 | .50 | -60 | 0 | high |
| 15 | .50 | 70 | .50 | -60 | 5 | high |
| 16 | .50 | 80 | .50 | -60 | 10 | high |
| 17 | .50 | 90 | .50 | -60 | 15 | high |
| 18 | .50 | 100 | .50 | -60 | 20 | high |
| 19 | .50 | 110 | .50 | -60 | 25 | high |
| 20 | .50 | 120 | .50 | -60 | 30 | high |
| 21 | .50 | 20 | .50 | -30 | -5 | low |
| 22 | .50 | 30 | .50 | -30 | 0 | low |
| 23 | .50 | 40 | .50 | -30 | 5 | low |
| 24 | .50 | 50 | .50 | -30 | 10 | low |
| 25 | .50 | 60 | .50 | -30 | 15 | low |
| 26 | .50 | 70 | .50 | -30 | 20 | low |
| 27 | .50 | 80 | .50 | -30 | 25 | low |
| 28 | .50 | 90 | .50 | -30 | 30 | low |
| 29 | .46 | 60 | .54 | -60 | -5 | high |
|  |  |  |  |  |  |  |


| 30 | .54 | 60 | .46 | -60 | 5 | high |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| 31 | .58 | 60 | .42 | -60 | 10 | high |
| 32 | .63 | 60 | .37 | -60 | 15 | high |
| 33 | .67 | 60 | .33 | -60 | 20 | high |
| 34 | .71 | 60 | .29 | -60 | 25 | high |
| 35 | .75 | 60 | .25 | -60 | 30 | high |
| 36 | .42 | 30 | .58 | -30 | -5 | low |
| 37 | .58 | 30 | .42 | -30 | 5 | low |
| 38 | .67 | 30 | .33 | -30 | 10 | low |
| 39 | .75 | 30 | .25 | -30 | 15 | low |
| 40 | .83 | 30 | .17 | -30 | 20 | low |

Note: $p($ win $)=$ probability of receiving the positive outcome (Gain); $p$ (loss) = probability of receiving the negative outcome (Loss). EV = gamble's expected value.

The order in which the target gambles were presented was randomized. For half of the participants gains appeared on the left side and for the other half on the right side. Reference and target gambles were presented sequentially to allow better measures of the relative attention given to each attribute (win, loss, probability) of each gamble (see Figure 1 for a screenshot). Each trial started with a fixation cross $(100 \mathrm{~ms})$. Then the reference gamble was presented until participants pressed the return key. The target gamble appeared until participants made a choice by pressing " 1 " for the reference gamble or " 2 " for the target gamble. The task was implemented in Presentation.


Figure 1. Screenshot of a target gamble.
Stress Manipulation. In the stress condition, we induced stress with the cold pressor task (CPT; Lovallo, 1975). The CPT is a standard method to induce a stress response and has been shown to reliably increase subjective stress and cortisol levels (McRae et al., 2006). In the CPT participants immersed their right hand in ice water $\left(0^{\circ}-4^{\circ} \mathrm{C}, M=1.86^{\circ} \mathrm{C}, S D=0.67\right)$ for as long as possible, up to 3 minutes. In the no stress condition participants immersed their hand in warm water ( $37^{\circ}-$ $40^{\circ} \mathrm{C}, M=38.98^{\circ} \mathrm{C}, S D=0.81$ ).

Measurement of Mood, Arousal, and Stress. We measured mood and arousal with the Self Assessment Mannequins (SAM; Hodes, Cook, \& Lang, 1985). To measure the physiological stress response we took saliva samples collected using Salivettes (Sarstedt, Nuembrecht, Germany) to determine cortisol levels. Saliva samples
were analyzed at the laboratory of the Technical University Dresden. Salivary free cortisol levels were determined using a chemoluminescence immunoassay (IBL, Hamburg, Germany) with intra- and interassay precision of $2.5 \%$ and $4.7 \%$, respectively.

Procedure. After participants arrived we determined whether they met the inclusion criteria for the study and gave them approximately 8 fl . oz. of water to drink. Then, we took the first saliva sample and measured mood and arousal (T1). Following the measurements, participants immediately proceeded with the first session of the financial decision-making task. After that participants gave the second saliva sample and again completed the mood and arousal measures (T2). Next, participants proceeded with the stress manipulation. 15 min after the stress manipulation, so that cortisol levels had time to rise, we took the third saliva sample and measured mood and arousal (T3). Immediately afterwards, participants performed the financial decision-making task again (Session 2). After that we again measured mood and arousal and took the fourth saliva sample (T4). Figure 2 provides a schematic overview of the experimental design.


Figure 2. Overview of the experimental design. The abbreviations are explained in the text.

Eye Movements. We recorded participants’ eye movements while they solved the financial decisionmaking task by using a remote eye-tracking device (SensoMotorics Instruments LLC), using the iViewX software and a remote binocular sampling rate of 120 Hz . The stimulus material was presented on a screen with a resolution of $1680 \times 1050$ pixels and a refresh rate of 60 Hz . The eye tracker was calibrated before the decisionmaking task and calibration was checked and if necessary repeated after each decision (20 pixel tolerance). Further analysis was done in Matlab. Fixations were identified using a 20 pixel tolerance (i.e., added max-min deviation for $x$ - and $y$-coordinates) and a minimum fixation time threshold of 50 ms (see Fiedler \& Glöckner, 2012 for a similar procedure).

We defined areas of interest (AOI) as circles with a radius of 120 pixels around each piece of information;
that, is the potential loss and gain, and the probability with which a loss or gain would occur (see Figure 1).

## Results

Mood, Arousal, and Stress Response. First we analyzed whether the stress manipulation influenced participants' mood, arousal, and cortisol levels (for means and $S D$ see Table 2).

Table 2: Descriptive statistics (Means and SDs) for mood, arousal, and cortisol by stress condition

| Measure | T 1 | T 2 | T 3 | T 4 |
| :--- | :---: | :---: | :---: | :---: |
|  | $M(S D)$ | $M(S D)$ | $M(S D)$ | $M(S D)$ |
| Stress |  |  |  |  |
| Mood | 2.85 | 3.08 | 2.95 | 3.03 |
|  | $(1.00)$ | $(1.19)$ | $(1.11)$ | $(1.10)$ |
| Arousal | 6.38 | 6.38 | 6.03 | 6.9 |
|  | $(1.29)$ | $(1.63)$ | $(1.95)$ | $(1.28)$ |
| Cortisol | 12.50 | 10.16 | 17.92 | 14.21 |
| (nmol/l) | $(10.04)$ | $(7.13)$ | $(10.86)$ | $(11.12)$ |


| No Stress |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: |
| Mood | 3.17 | 3.07 | 3.10 | 3.03 |
|  | $(1.10)$ | $(1.16)$ | $(1.32)$ | $(1.10)$ |
| Arousal | 6.03 | 6.10 | 6.86 | 6.76 |
|  | $(1.57)$ | $(1.70)$ | $(1.62)$ | $(1.86)$ |
| Cortisol | 12.09 | 9.95 | 8.12 | 7.40 |
| (nmol/l) | $(7.50)$ | $(5.68)$ | $(3.79)$ | $(3.53)$ |

Note: $\mathrm{N}_{\text {stress }}=40 ; \mathrm{N}_{\text {no stress }}=29$; lower numbers indicate more positive mood and higher arousal

Mixed analyses of variance (ANOVAs) on mood, arousal and cortisol with measurement time (T1-T4) as within-subject factor and stress condition as betweensubjects factor showed that arousal and cortisol levels increased in the stress group but not in the no stress group. In the no stress group cortisol and arousal decreased, suggesting that participants' initial excitement decreased during participation. This was indicated by significant interactions of measurement time and stress condition, Arousal: Greenhouse-Geisser corrected $F(3,166)=6.15, p=.01$; Cortisol: Greenhouse-Geisser corrected $F(3,90)=18.23, p<.001$. We did not find an effect of stress on mood, that is there was no interaction between measurement time and stress condition, $F(3$, 201) $=0.87, p=.46$, nor main effects of time or stress condition (all $p s>.65$ ).

The Influence of Stress on Decisions under Risk. We measured risk taking as the proportion of trials in which the risky option (i.e., the target gamble) was chosen. On average participants chose the risky option in $44 \%$ of the choices in the first session and in $46 \%$ in the second session, indicating that participants were rather risk averse (a risk neutral decision maker who always chose the
option with the higher expected value should have chosen the target gamble in $67.5 \%$ of the trials). In a first step we analyzed whether stress influenced the proportion of risky choices with a mixed ANOVA with session (before/after the stress induction) as within-subject factor and stress condition as between-subjects factor. We did not find a main effect of session $(F(1,67)=0.89, p=.35)$ or stress $(F(1,67)=0.43, p=.51)$, nor an interaction between them $(F(1,67)=0.10, p=.75$; for means and $S D$ see Table 3).

In the next step we tested whether the difference in outcomes of the gambles influenced how stress affected risky decision making. We focused on the choices where target gambles offered a higher expected value than the reference gamble (i.e., $E V>10$ ), to account for participants' overall risk aversion. A mixed ANOVA with session and gamble type as within-subject factors and stress condition as between-subjects factor showed that participants chose the target gamble more frequently for the low outcome gambles than the high outcome gambles, $F(1,67)=70.96, p<.001$. Additionally we found a threeway interaction between stress condition, session and gamble type, $F(1,67)=5.87, p=.02$. As illustrated in Figure 3, repeated measurement ANOVAs for stressed and not stressed participants separately showed an interaction between time and gamble type for participants in the stress condition, $F(1,39)=7.53, p=.01$, but not in the no stress condition, $F(1,28)=0.76, p=.39$. This suggests that in the second session compared to the first session participants in the stress condition-but not participants in the no stress condition-took more risks with low outcome gambles, but less risk with high outcome gambles.


Figure 3. Proportion of risky choices for gambles with high and low outcomes in the stress and no stress group

Because previous literature has shown that men and women react differently to stress (e.g., Lighthall et al., 2009), we ran additional analyses including gender as a further between-subjects factor. We found a main effect
of gender in that women chose the risky option less frequently than men ( $M_{\text {men }}=.79, S E=.05, M_{\text {women }}=.63$, $S E=.03, F(1,65)=7.15, p=.01)$. However, gender did not interact with the gamble type, nor affect the results of stress on high and low outcome gambles.

Table 3: Descriptive statistics for risky decision making and measures of eye movement by stress condition

| Measure | Stress |  | No stress |  |
| :--- | :---: | :---: | :---: | :---: |
|  | $M$ | $S D$ | $M$ | $S D$ |
| Risky Choices (Session 1) | 0.43 | 0.17 | 0.46 | 0.19 |
| Risky Choices (Session 2) | 0.45 | 0.18 | 0.47 | 0.22 |
| FixationGainLoss (High 1) | 0.08 | 0.09 | 0.08 | 0.10 |
| FixationGainLoss (High 2) | 0.07 | 0.11 | 0.08 | 0.10 |
| LastGainLoss (High 1) | 0.20 | 0.27 | 0.19 | 0.31 |
| LastGainLoss (High 2) | 0.19 | 0.37 | 0.18 | 0.30 |
| FixationGainLoss (ow 1) | 0.08 | 0.10 | 0.08 | 0.11 |
| FixationGainLoss (Low 2) | 0.06 | 0.11 | 0.08 | 0.11 |
| LastGainLoss (Low 1) | 0.17 | 0.36 | 0.25 | 0.26 |
| LastGainLoss (Low 2) | 0.15 | 0.38 | 0.27 | 0.31 |

Eye Movements. Can the influence of stress on risktaking be explained by the relative attention given to gains and losses? To answer this question, we considered two measures of eye movements: (1) the relative duration with which gains were fixated compared to losses (FixationGainLoss) and (2) the relative proportion of trials on which the last fixation before making a decision was to the gain information or the loss information (LastGainLoss). We calculated the measures for high and low outcome gambles separately. Because the reference gamble was always the same, we focused on the target gambles. The FixationGainLoss was calculated by measuring the duration of fixations in each AOI (gains, losses and probabilities) for each trial. Next, we computed how long gains were fixated relative to losses and calculated the average for trials with high and low outcome gambles with an expected value of 10 or higher. The LastGainLoss was calculated by taking the difference between the proportion of trials with high and low outcome gambles with an expected value of 10 or higher in which the last focus was to the gain AOI relative to the loss AOI; see Table 3 for means and SD.

We then investigated whether the two measures of eye movements were related to the proportion of risky choices. Correlations indicated that the longer gains were fixated compared to losses and the more often the last fixation was to the gain AOI relative to the loss AOI, the more participants chose the risky option, particularly in the high outcome gambles (see Table 4).

To investigate whether the attention to gains and losses changed with stress, we conducted two mixed ANOVAs with session and variance as within-subject factors and stress condition as between-subjects factor.

We did not find an effect of session, stress condition or variance for the relative time gains and losses were
looked at (all $p \mathrm{~s}>.18$ ). The analysis on the location of the last fixation before making a decision also showed no main effects of session, gamble type or stress condition (all $p \mathrm{~s}>.45$ ), but indicated a significant interaction between gamble type and stress condition, $F(1,67)=4.09$, $p=.05$.

Follow-up analyses for participants in the stress condition and the no stress condition separately showed an effect of gamble type in the no stress condition, $F(1,28)=4.24, p=.05$, indicating that the participants more frequently looked to gains compared to losses for the low outcome gambles than for the high outcome gambles (see Table 3). In the stress condition, however, we did not find an effect of session or gamble type (all $p$ s $>.37$ ).

Table 4: Correlation between measures of eye movement and risk taking

|  | Risky Choice <br> (H1) |  | Risky Choice <br> (H2) |  |
| :--- | :---: | :---: | :---: | :---: |
|  | $r$ | $p$ | $r$ | $p$ |
| FixationGainLoss (H1) | .33 | .005 | .37 | .002 |
| FixationGainLoss (H2) | .35 | .003 | .34 | .004 |
| LastGainLoss (H1) | .34 | .004 | .38 | .001 |
| LastGainLoss (H2) | .30 | .01 | .30 | .02 |

Note: $N=69 ; \mathrm{H} 1=$ high outcome gambles, session 1; H2 $=$ high outcome gambles, session 2

## Discussion

The effect of stress on decisions under risk seems to depend on the risk the decision involves. Whereas we did not find an overall influence of stress on Although taking, a detailed analysis showed that the influence of stress depended on the variability in the gambles' outcomes. After immersing their hand in ice-cold water, participants chose the risky gamble more frequently when the difference between outcomes was relatively low, but less frequently when the difference between outcomes was high. This suggests that the influence of stress on risk taking depends on the riskiness of the decision-making task, resonating with research showing that stress increases risk taking in the loss domain but decreases risk taking the gain domain (Porcelli \& Delgado, 2009). These results can help reconcile the diverse effects of stress on risky decision making in the literature by showing that to understand the influence of stress it is necessary to take task characteristics such as the involved risk of a decision into account.
A second goal of the research was to investigate whether the relative attention given to gains and losses is a potential mechanism underlying the influence of stress. Overall, participants who gave relatively more attention to gains than to losses tended to choose the risky option more frequently. This resonates with previous work suggesting that the time spent on information is related to its importance for the choice (e.g., Ben Zur \& Breznitz,

1981; Glaholt et al., 2009). Additionally we found that the last fixation before making a choice was related to risk taking, dovetailing with research on gaze cascade effects in risky decision making (e.g., Fiedler \& Glöckner, 2012). Moreover, it suggests that gaze cascade effects extend to the attribute that was most important in determining choice.

We did not find any evidence, however, that stress changed the relative attention given to gains over losses or the last information looked at. This could suggest that the influence of stress is not mediated by the attention given to gains and losses. On the other hand, the effect of stress could have been masked by noise given to the relatively few gambles in our task.

In sum, our results suggest that stress changes how risky decisions are made. Although the mechanism by which stress exerts its influence requires further research, it becomes clear that the effect of stress can only be understood when considering the characteristics of the decision task.

## Acknowledgments

This research was supported by research grants from the Swiss National Science Foundation (100014_130192/1 and CRSII1_136227).

## References

Ben Zur, H., \& Breznitz, S. J. (1981). The effect of time pressure on risky choice behavior. Acta Psychologica, 47, 89-104.
Carr, P. B., \& Steele, C. M. (2010). Stereotype threat affects financial decision making. Psychological Science, 21, 1411-1416.
Fiedler, S., \& Glöckner, A. (2012). The dynamics of decision making in risky choice: an eye-tracking analysis. Frontiers in Psychology, 3, 335.
Friedman, R. S., \& Förster, J. (2010). Implicit affective cues and attentional tuning: An integrative review. Psychological Bulletin, 136, 875-893.
Glaholt, M. G., \& Reingold, E. M. (2011). Eye movement monitoring as a process tracing methodology in decision making research. Journal of Neuroscience, Psychology, and Economics, 4, 125-146.
Glaholt, M. G., Wu. M., \& Reingold, E. M. (2009). Predicting preference from fixations. PsychNology Journal, 7(2), 141-158.
Hodes, R. L., Cook III, E. W., \& Lang, P. J. (1985). Individual differences in autonomic response: conditioned association or fear? Psychophysiology, 22, 545-560
Hoffman, J. E. (1998). Visual attention and eye movements. In H. Pashler (Ed.), Attention (pp. 119-153). Hove, United Kingdom: Psychology Press.
Kahneman, D., \& Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, 47, 263-291.

Kelly, I., Ashleigh, H., \& Beversdorf, D. Q. (2007). Effect of the cold pressor test on memory and cognitive flexibility. Neurocase, 13, 154-157.
Krajbich, I., Armel, C., and Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. Nature Neuroscience, 13, 1292-1298.
Lighthall, N. R., Mather, M., \& Gorlick, M. A. (2009). Acute stress increases sex differences in risk seeking in the balloon analogue risk task. Plos One, 4, e6002.
Lovallo, W. (1975). The cold pressor test and autonomic function: a review and integration. Psychophysiology, 12, 268-82.
McRae, A. L., Saladin, M. E., Brady, K. T., Upadhyaya, H., Back, S. E., \& Timmerman, M. A. (2006). Stress reactivity: biological and subjective responses to the cold pressor and Trier Social stressors. Human Psychopharmacology, 21(6), 377-385.
Porcelli, A. J., \& Delgado, M. R. (2009). Acute stress modulates risk taking in financial decision making. Psychological Science, 20, 278.
Preston, S. D., Buchanan, T. W., Stansfield, R. B., \& Bechara, A. (2007). Effects of anticipatory stress on decision making in a gambling task. Behavioral Neuroscience, 121, 257-263.
Putman, P., Antypa, N., Crysovergi, P., \& van ver Does, W. A. J. (2010). Exogenous cortisol acutely influences motivated decision making in healthy young men. Psychopharmacology, 208, 257-263.
Rehder, B., \& Hoffman, A. B. (2005). Thirty-Something categorization results explained: Selective attention, eyetracking, and models of category learning. Journal of Experimental Psychology: Learning Memory, and Cognition, 31, 811-829.
Sharpe, W. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. Journal of Finance, 19, 425.
Starcke, K., Wolf, O. T., Markowitsch, H. J., \& Brand, M. (2008). Anticipatory stress influences decision making under explicit risk conditions. Behavioral Neuroscience, 122(6), 1352-1360.
Von Dawans, B., Fischbacher, U., Kirschbaum, C., Fehr, E., \& Heinrichs, M. (2012). The social dimension of stress reactivity: acute stress increases prosocial behavior in humans. Psychological Science, 23, 651-660.
Weber, E. U., Shafir, S., Blais, A-R. (2004). Predicting risk sensitivity in humans and lower animals: Risk as variance or coefficient of variation. Psychological Review, 111, 430-445.
Wichary, S., \& Rieskamp, J. (2011). Probabilistic inferences under emotional stress. In L. Carlson, C. Hoelscher, \& T. F. Shipley (Eds.), Proceedings of the 33rd Annual Conference of the Cognitive Science Society (pp. 1872-1877). Austin, TX: Cognitive Science Society.

# Negative observations, induction and the generation of hypotheses 

Wouter Voorspoels (wouter.voorspoels@ppw.kuleuven.be)<br>Chayenne Van Meel (chayenne.vanmeel@student.kuleuven.be)<br>Gert Storms (gert.storms@ppw.kuleuven.be)<br>KU Leuven, Department of Experimental Psychology, Tiensestraat, 102<br>Leuven, 3000, Belgium.


#### Abstract

In category-based induction tasks, it is a robust finding that positive observations raise the judged likelihood of a conclusion and negative observations lower judged likelihood. We present evidence that negative observations can raise the judged likelihood. In particular, we asked participants to judge the likelihood of a conclusion after introducing them to different sets of premises either containing one positive observation or the same positive observation and a negative observation. We found that when the negative observation is dissimilar to the positive observation, willingness to accept a conclusion is raised. Moreover, results from a simultaneous hypothesis generation task suggest that the rise in judged conclusion likelihood is due to a peculiar shift in the hypothesis space of the reasoner, in that the hypothesis with the largest extension, yet still consistent with all premises gains disproportionate popularity when introducing dissimilar negative observations.


Keywords: induction; non-monotonicity; reasoning; sampling assumptions;

## Introduction

People often find themselves in situations that require judgments based on incomplete knowledge, derived from an incomplete set of observations. From experience with traffic lights, we can conclude that red is diagnostic for dangerous situations (positive observations), and we will refrain from crossing the road. However, we have also encountered traffic lights on lonely nights, when there is no traffic. In that situation, red does not necessarily indicate danger (negative observation). How do we combine these observations to make a decision about crossing the road? The world is not sufficiently friendly to provide us with an exhaustive set of observations. But we do not want to stay on the same side of the road all our lives. We want to see the other side of the road! Therefore inference to uncertain conclusions, generally referred to as induction, is omnipresent in everyday life and almost equally widely studied in cognitive science (Heit, 2000).

A common paradigm to study induction a is the categorybased induction task: Participants are asked to infer the presence of a feature in a conclusion category on the basis of a set of observations. The observations are presented as premises of the argument. For example:

## Premise: Tigers have sesamoid bones

## Conclusion: Lions have sesamoid bones

A number of regularities have been reported regarding how people respond to such problems, one of which forms the
topic of the present paper. Following intuition, but also according to the main theories of inductive reasoning (see, e.g., Heit, 2000), there exists a monotonic relation between the number of observations and the strength of an argument: As more objects displaying the property are observed, a conclusion will be judged more likely (see, e.g., Osherson, Smith, Wilkie, Lpez, \& Shafir, 1990). Similarly, as objects are observed that do not have the property, the judged likelihood of a conclusion decreases (e.g., Heussen, Voorspoels, Verheyen, Storms, \& Hampton, 2011). We will refer to this general finding as the monotonicity principle.

Put differently, the monotonicity principle predicts that a positive observation ${ }^{1}$ raises argument strength, and a negative observation ${ }^{2}$ lowers argument strength. In case the observation is extremely dissimilar, and thus irrelevant to the conclusion, argument strength remains the same. For example, the likelihood of the conclusion 'Lions have sesamoid bones' is raised by the observation that tigers have sesamoid bones, is lowered by the observation that leopards do not have sesamoid bones and remains the same following the observation that tea cups do not have sesamoid bones. Additional relevant positive and negative observations will respectively raise and lower the argument strength further (asymptotically, obviously).

Recently we have presented evidence that suggests that negative observations can in some cases increase argument strength, contrary to what the monotonicity principle predict (Heussen et al., 2011). In a forced choice paradigm, participants showed a preference for an argument of the following form as compared to an argument without the second, negative premise:

Mozart's music elicits alpha waves in the brain
Metallica's music does not elicit alpha waves in the brain

## Bach's music elicits alpha waves in the brain

A potential explanation for the results of (Heussen et al., 2011) is that negative observations point the reasoner to a relevant dimension to base inference upon (Medin, Goldstone, and Gentner (1993)). In the above argument, the negative observation highlights a commonality between Mozart and Bach, not shared by Metallica, i.e., that being classical music is the crucial feature to base inference upon. As a conse-

[^249]quence, the argument will be considered stronger. In addition, by explicitly contradicting some of the potential hypotheses (e.g.,all music elicits alpha waves), negative evidence clearly helps in reducing the number of hypotheses. This is expected to raise believe in the hypotheses that remain consistent with the premises after the negative observation.

## Outline

The present study aims at further investigating the rise in argument strength following a negative observation. In particular, our aim is twofold. First, we want to replicate the effect in a rating task, in which we compare generalization judgments of participants who were not presented with negative observations to judgments by participants who were. This differs considerably from Heussen et al. (2011), where a forced choice paradigm was used. Second, we ask participants to generate hypotheses after introducing the observations. In this way, we can examine how the hypothesis space of people confronted with negative observation changes and how this relates to their generalization judgments.

Following Heussen et al. (2011), we hypothesize that a negative observation will raise the willingness to accept a conclusion whenever it points to a dimension that can be used to make the required generalization. In effect, we expect that the projection of a feature from Mozart's music to Bach's music is facilitated when a negative observation excludes other types of music, and points to classical music as the correct extension of the novel feature. Similarly, when projecting a property from Bach's music to Nirvana's music, the projection is expected to be facilitated by a negative observation outside the category that entails both subcategories (music). Adding the premise that the sound of a falling rock does not have the property, is thus expected to increase the willingness to project the property to Nirvana's music. Moreover, we expect that the hypothesis space of participants will vary accordingly.

## Experiment

## Method

Participants Participants were 172 bachelor students psychology who volunteered for course credits.

Materials We used 12 argument topics taken from Heussen et al. (2011) (music, painters, public figures, types of ships, types of glass, types of displays, water bodies, fruit, water birds, insects, polar animals). In each topic, a hierachical structure is present, comprising of a category (e.g., music), two subcategories (A: classical music and B: rock music) and a supordinate category ( C : sound).

Each of the topics has a base argument built from one premis from subcategory A (e.g., Mozart's music has X). Depending on the condition, negative premises are added from the other subcategory (B), or a different category (C). Thus, either the additional premise contains information regarding a member of subcategory A (e.g., Vivaldi's music), or the premise contains information on a member of subcategory

B (e.g., Metallica's music) or the premise presents information from category C (e.g., the sound of a waterfall). Table 1 presents an overview of the premises for the topic music.

In Heussen et al. (2011), only one conclusion from subcategory A was used. For the present experiment, we added two conclusion categories to each topic: One from subcategory B (e.g., Nirvana's music), and one from category C (e.g., the sound of a falling rock). The properties that were to be generalized from premises to conclusions, were intuitively realistic characteristics that participants were likely to have very little knowledge about (e.g., contain lycopene; create conversion currents; elicit alpha waves in the brain).

Table 1: an overview of the simtulus material for the topic 'music'. Entries in bold refer to items that are presented in every condition (e.g., the base premise "Mozart's music elicits alpha waves").

| Type | Premise | Conclusion |
| :--- | :--- | :--- |
| subcategory (A) | Mozart, Vivaldi | Bach |
| subcategory (B) | Metallica | Nirvana |
| superordinate (C) | falling rock | waterfall |

Procedure The experiment had the form of a web-based survey. On each trial, participants were presented with a short scenario describing that specialists in the domain of interest (e.g., neuroscientists) had recently made novel discoveries. This was followed with the premise (or premisses) of an argument. For example, participants were given following premise:

Mozart's music elicits alpha waves in the brain.

After reading the information, participants received two successive tasks. First, in the hypothesis generation task, they were asked to come up with a rule underlying the observations, (e.g., "classical music elicits alpha waves in the brain"). They were asked to type their hypothesis in a textbox in one or two sentences. Second, in the generalization task, participants were asked to judge how likely the three conclusions associated with the argument were by moving a bar on a continuous scale running from 1 to 100 for each of the conclusions.

For each topic, we constructed six premise sets, varying the type of observations, and the "'sign"" (positive or negative). For each premise set the exact same three conclusions were judged for likelihood, but the premise set varied across condtions. Here, we will discuss only three conditions that allow crucial comparisons to test for the effect of negative observations. In the base condition, referred to as posA, participants were presented with the base premise, as in (1). In condition posAnegB, a negative observation from a different subcategory is added to the base premise:

## Mozart's music elicits alpha waves in the brain. Metallica's music does not elicit alpha waves in the brain.

In a third condition, posAnegC a negative observation was added to the base premise, disclosing information on a member of the same superordinate category:

## Mozart's music elicits alpha waves in the brain. The sound of a falling rock does not elicit alpha waves (3) in the brain

In total, $12 \times 6$ arguments were constructed. The 72 arguments were distributed across 6 lists so that each list contained each of the twelve topics (so participants did not see the same topic twice), and a list contained 2 arguments for every type of premise set (so each participant got two arguments from every condition). The lists were distributed randomly across participants. The order of arguments within a list was random for each participant, as well as the order of the conclusions in each argument. The same two practice items preceded each list in order to familiarize the participants with the procedure. These two items were not included in the analyses. The experiment took no longer than 20 minutes

Premise sets posA, posAnegB and posAnegC form the object of the present examination. The structure of premise sets 4 to 6 is listed in Table 2, but will not be discussed in the present paper. As can be seen in table 2, these premise sets do not contain negative observations, except the "completely saturated" premise set 6 .

Table 2: Schematic overview of the experiment. ' + ' refers to a positive observation, '-' to a negative observation. '++' means that two premises from the same subcategory were presented in the corresponding condition. In the present paper we focus on the first three premise sets.

| Cond | subcat A | subcat B | cat C | \# premises |
| :--- | :--- | :---: | :---: | :---: |
| posA | + |  |  | 1 |
| posAnegB | + | - |  | 2 |
| posAnegC | + |  | - | 2 |
| 4 | ++ |  |  | 2 |
| 5 | + | + | 2 |  |
| 6 | ++ | - | 3 |  |

## Results

## Generalization

To recapitulate, participant were shown a set of premises (observations), and asked to judge the likelihood of three conclusions. One conclusions concerned a member of the same subcategory as the base premise (subcategory A), a second conclusion concerned a member of a different subcategory (subcategory B) and a third conclusion a member of the shared superordinate category (category C). In this section, we examine the manner in which these generalization judgments
vary as a function of the premise set that is presented, and in particular, whether adding a negative observation to the premise set can raise argument strength. In what follows, it is informative to keep in mind that, according to the monotonicity principle, negative observations are expected to lower the likelihood of a conclusion (or leave it unaltered).

Figure 1 presents the average scores of all three conclusion likelihood judgments, averaged across participants and items, as a function of the premise set. PosA introduces only the base premise, posAnegB adds a negative observation of subcategory $B$ to the base premise and posAnegC adds a negative observation of category C. In the two following sections, statistical analyses are presented to test for the effects of adding negative observations to a premise set.


Figure 1: Average judged conclusion likelihood for the three types of conclusions, as a function of the premise set.

Generalizing to the same subcategory For the conclusions that concern a member of subcategory A (e.g., conclusions to Bach's music if Mozart's music is the base premise), we are interested in two conditions that should lead to a raise in conclusion likelihood as compared to the base argument, despite the negative observation in a premise set. As in Heussen et al. (2011), we expect to observe a difference between premise set posA and posAnegB, in which a negative observation from subcategory $B$ is added to the premise set. Additionally, we hypothesize that adding negative observation from an entirely different category, as in premise set posAnegC, also raises argument strength. In Figure 1 the left bars present the average judged likelihood of the conclusion to a member of subcategory A as a function of the premise set preceding the conclusion, and visual inspection confirms our hypotheses.

We performed a mixed-effects model analyses with two random effects (participants and topics), and two fixed effects (list and premise set), and their interaction ${ }^{3}$. Premise set is a within subjects factor and list is a between subject variable.

Table 3 gives an overview of the main (fixed) effects of

[^250]premise set, and can be interpreted as follows: For premise set posA, participants on average judge the likelihood of a conclusion to a member of subcategory A to be 75.14. For premise set posAnegB, in which a negative observation from subcategory B is added, the judged likelihood drops with 2.26 according to the model ${ }^{4}$, a change that is not significant. For premise set posAnegC, in which a negative observation from a different category C is added, the judged likelihood is significantly higher 11.12 points $(p=.016)$.

Table 3: Effects of premise set on generalizing to a member of subcategory A.

| premise set | MCMC estimate | MCMC p-value |
| :--- | ---: | ---: |
| posA (base level) | 75.14 | $<.001$ |
| posAnegB | -2.26 | .72 |
| posAnegC | 11.12 | .016 |

In sum, we only find partial support for our hypothesis. In particular, premise sets as used in (Heussen et al., 2011), adding a negative observation from a different subcategory, do not lead to a significant rise in argument strength. We do, however, observe a strong rise in argument strength, when a more distant negative observation - from a different category - is added to the premise set.

Generalizing to a different subcategory For the conclusion to a member of subcategory B (e.g., Nirvana has $X$; the base premise is Mozart), we hypothesize that a negative observation from a different category (but shared superordinate category, e.g., the sound of a falling rock) can raise judged conclusion likelihood. In Figure 1 the average judged likelihood of conclusions to subcategory B for the relevant premise sets is presented in the middle bar of every group, and a rise in mean judged likelihood from premise set posA to premise set posAnegC can be observed. A quantitative test of the difference was performed using mixed model analyses with two random effects (items and participants) and two fixed effects (list and premise set). As in the previous section, this model was preferable to alternative models in terms of AIC and log likelihood deviance.

Table 4: Effects of premise set on generalizing to a member of subcategory B.

| premise set | MCMC estimate | MCMC p-value |
| :--- | ---: | ---: |
| posA (base level) | 33.93 | $<.001$ |
| posAnegB | -16.55 | .08 |
| posAnegC | 33.61 | $<.001$ |

Table 4 summarizes the effects of premise set. When adding a negative observation from subcategory B , judged

[^251]likelihood of the conclusion is lowered by 16.55 , nearly significantly ( $\mathrm{p}=.08$ ). Note that in this case the premise set contains a negative observation form the same subcategory as the conclusion. More interestingly, when adding a negative observation from a different category (the sound of a falling rock does not have X), judged likelihood increases an impressive 33.61 .

Conclusions Our analyses of the judged likelihood of the conclusions have revealed convincing evidence that negative observations can raise argument strength in some circumstances. In particular, we found that a negative observation from a seemingly irrelevant category, can substantially raise the judged likelihood of the conclusion to a member of the same subcategory as the base premise as well as to a member of a different subcategory.

Contrary to (Heussen et al., 2011), we do not find support for a rise in judged likelihood of a conclusion to a member of the same subcategory when a negative observation from a different subcategory is introduced. Note that Heussen et al. used a forced choice paradigm, and report effects that, while significant, were very subtle. Perhaps our methodology was not able to identify these effects.

## Hypothesis generation

Before making the generalization judgments, participants were asked to generate a hypothesis that they believed explained the observations in the premises. This allows us to peak at the type of hypotheses participants entertained when confronted with negative observations

We differentiated between four types of hypotheses: First, a hypothesis can state that the property is only applicable to the base premise (e.g., "only Bach has X"). Second, a hypothesis can generalize the property to the subcategory of which the base premise is a member (e.g., "all classical music has X "), or, the third type, to the entire category ("all music has $X^{\prime \prime}$ ) or, in the fourth case, to the entire superordinate category ("all sound has X"). We classified each rule according to its consequential region following this scheme. Hypotheses that did not fit the scheme, for example due to reporting another subcategory, an unspecified subcategory (e.g., "some types of music") or a causal explanation, were coded as "other" 5 .

Figure 2 presents the relative frequencies of each type of hypothesis being generated as response to premise sets posA, posAnegB and posAnegC. To quantify and test differences in hypothesis generation between premise sets, we performed logistic regressions with premise set and list as predictors and a binary variable indicating whether the type of hypothesis was generated as dependent variable. The regressions were performed seperately for each type of hypothesis.

Figure 2 reveals three patterns. First, hypotheses that restrict the property to the base premise are significantly less frequent when a negative observation is introduced (with comparison level posA, for posAnegB: Wald $=-2.4, p=$

[^252]

Figure 2: Relative frequency of generating a certain type of hypothesis as a function of the premise set.
.01; and for posAnegC: Wald $=-4.7, p<.001$ ). Second, hypotheses that attribute the property to the entire subcategory A are more frequent when a negative observation from subcategory B is introduced (but not significantly, Wald $=$ $.76, p=.45$ ), and significantly less frequent when a negative observation from category C is introduced (Wald $=$ $-3.74, p<.001$ ). Third, hypotheses that project the feature to the entire category are less common when introducing a negative observation from a subcategory (Wald $=-6.23, p<$ .001), yet more common when introducing a negative observation from a different category (Wald $=6.02, p<.001$ ).

Conclusions The hypothesis space following arguments with negative observations from subcategory B or category C is substantially altered as compared to the one premise arguments (posA). Moreover, the shifts in generation frequency seem to follow results in the generalization tasks. In particular the increase in subcategory $B$ conclusion likelihood when premise set posAnegC is observed, is clearly associated to an increase in the hypothesis that music is the relevant category.

Not only do some hypotheses become more frequently generated (as can be expected when a number of hypotheses are excluded due to a negative observation), the relative differences between types of hypotheses change considerably across different conditions. In particular, for premise set posAnegB, hypotheses that refer to subcategory A are disproportionately more popular. For premise set posAnegC, hypotheses that refer to the entire category are disproportionately more frequent. Note that for these premise sets, other hypotheses that are also consistent with the observations become less popular. This suggests that more is going on than evenly redistributing the belief of excluded hypotheses across remaining hypotheses. We will come back to this issue in the discussion.

## General discussion

We have presented evidence against the universality of the monotonicity principle in inductive reasoning. Negative ob-
servations can indeed raise argument strength when they come from a different category than the one shared by the positive observation and the conclusion. Moreover, we found a clear relation with the type of hypotheses that are generated to account for the premise observations. In general, there seems to be a dramatic rise in the weight of the largest hypothesis that is consistent with both positive and negative observations in the premise set. In what follows, we will discuss the relation of these findings to earlier violations of monotonicity in inductive reasoning, and in relation to the sampling assumptions that people have that is, ideas about how the observations are presented to them.

## Relation to positive non-monotonicity

For positive observations, a violation of the monotonicity principle has already been documented Medin, Coley, Storms, and Hayes (2003), in that under some circumstances positive observations can lower conclusion likelihood. For example, consider following two arguments:

## Brown bears have X

## Goats have X

## Brown bears have $X$ <br> Polar bears have $X$

## Goats have X

Medin et al. (2003) report that participants judge argument (4) stronger than argument (5). According to Medin et al., the addition of the positive observation in (5) reinforces a property that is shared among the premises but is not applicable to the conclusion. Put differently, by adding a positive observation, more weight is given to the hypothesis that the being a bear is crucial for the novel property, and since this property is not shared by the conclusion, it is judged less likely.

The non-monotonicity from adding a negative observation is strikingly symmetric to the non-monotonicity reported by (Medin et al., 2003). Consider following two arguments:

## Mozart's music has X

## Metallica's music has $X$

## Mozart's music has $X$ <br> The sound of a falling rock does not have $X$

## Metallica's music has X

In the present study, argument (6) was judged stronger than (5). Following our analyses of the hypotheses generated by participants, the addition of the negative observation from outside the music category, drives people to think that "being music" is the most likely hypothesis, rather than, e.g., classical music or Mozart's music. By virtue of giving more weight to the music hypothesis, belief is raised that Metallica's music has X as well.

In sum, whereas in the positive case, a reasoner's hypothesis "tightens" to a small subcategory (e.g., bears) by introducing an observation that is very similar to the base premise (e.g., another type of bear), in the negative case, a reasoner's hypothesis seems to "broaden" to a large category, by introducing an observation that is very dissimilar to the base premise.

## Sampling assumptions and non-monotonicity

The question then is how reasoners arrive at weighting exactly these hypotheses more. From a naive probability point of view, excluding certain hypotheses by adding negative observations will automatically lead to a redistribution of the probability mass from the excluded hypotheses to remaining hypotheses. In effect, it makes sense that other hypotheses would indeed become more likely, and as a consequence a particular conclusion could also become more likely. However, it is important to appreciate that, in the naive case, the probability mass would be distributed evenly across the remaining hypotheses, so relative differences between different hypotheses remain. This does not seem to hold in the present results. Indeed, when a negative observation from a different category is observed, as in (6), participants generated the category hypothesis (all music has X) disproportionally more frequently. While consistent with the observations in (6), the subcategory hypothesis (all classical music has X) and the base premise hypothesis (Mozart has X ) experience a substantial drop in generation frequency, contrary to what one would expect on the basis of naive probability theory.

Interestingly, the manner in which Bayesian models of induction (e.g., Tenenbaum \& Griffiths, 2001) cope with the positive non-monotonicty effect, is by reweighting the remaining hypotheses when an observation is made. More precisely, depending on assumptions on how the particular observation is sampled from the environment, a Bayesian model would predict that reasoners give more weight as a consistent hypothesis is smaller (e.g., Navarro, Dry, \& Lee, 2011; Tenenbaum \& Griffiths, 2001.

While technical adjustments to Bayesian inference in nature ${ }^{6}$, sampling assumptions also represent a psychological reality and implement what the reasoners' assumptions are on how the observations are presented to him. For example, if a reasoner assumes that the observations in (5) are sampled from the correct hypothesis (for example, because he or she thinks the experimenter intentionally is trying to reveal the correct hypothesis), it is rational to attribute more believe to the hypothesis that it is about bears. Yet, if the reasoner believes the observations are made randomly in the world, and he or she might as well have observed a refridgerator (presumably not having the property in that case) instead of a polar bear, the hypothesis that it is about bears does not gain relative importance (for a more elaborate discussion, see Navarro et al., 2011). Indeed, it would be bad luck on part

[^253]of the reasoner that he or she did not encounter a more informative observation. The non-monotonicity effects, both positive and negative, suggest that reasoners do not share that assumption.

While the specific implementation of sampling assumptions discussed does not yet apply to negative evidence, a similar reweighting mechanism might be at work when a reasoner is presented with negative evidence. Perhaps reasoners assume that a negative observation is intentionally presented, in such a way that it does not only exclude inconsistent hypotheses, but is informative as to which hypothesis is the correct one.

## Acknowledgments

Wouter Voorspoels is a postdoctoral fellow at the Research Association - Flanders. We thank Steven verheyen for useful comments.

## References

Baayen, R. H., Davidson, D. J., \& Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language, 59, 390-412.
Bates, D. M., \& Sarkar, D. (2007). lme4: Linear mixedeffects models using s4 classes, r package version 0.998756. [Computer software manual].

Heit, E. (2000). Properties of inductive reasoning. Psychonomic Bulletin \& Review, 7, 569-592.
Heussen, D., Voorspoels, W., Verheyen, S., Storms, G., \& Hampton, J. A. (2011). Raising argument strength using negative evidence: A constraint on models of induction. Journal of Memory /\& Cognition, 39, 1496-1507.
Medin, D., Coley, J. D., Storms, G., \& Hayes, B. K. (2003). A relevance theory of induction. Psychonomic Bulletin \& Review, 10(3), 517-532.
Medin, D., Goldstone, R., \& Gentner, D. (1993). Respects for similarity. Psychological Review, 100, 254-278.
Navarro, D. J., Dry, M. J., \& Lee, M. D. (2011). Sampling assumptions in inductive generalization. Cognitive Science, 36, 187-223.
Osherson, D., Smith, E., Wilkie, O., Lpez, A., \& Shafir, E. (1990). Category-based induction. Psychological Review, 97, 185-200.
Tenenbaum, J. B., \& Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. Behavioral and Brain sciences, 24, 762-778.

# Explaining to Others Prompts Children to Favor Inductively Rich Properties 

Caren M. Walker ${ }^{1}$, Tania Lombrozo ${ }^{1}$, Cristine H. Legare ${ }^{2}$ \& Alison Gopnik ${ }^{1}$<br>(caren.walker@berkeley.edu, lombrozo@berkeley.edu, legare@psy.utexas.edu, gopnik@berkeley.edu)<br>${ }^{1}$ Department of Psychology, University of California, Berkeley, 3210 Tolman Hall, Berkeley, CA 94720 USA<br>${ }^{2}$ Department of Psychology, University of Texas at Austin, 1 University Station \#A8000, Austin, TX 78712 USA


#### Abstract

Three experiments test the hypothesis that engaging in explanation prompts children to favor inductively rich properties when generalizing to novel cases. In Experiment 1, preschoolers prompted to explain during a causal learning task were more likely to override a tendency to generalize according to perceptual similarity and instead extend an internal feature to an object that shared a causal property. In Experiment 2, we replicated this effect of explanation in a case of label extension. Experiment 3 demonstrated that explanation improves memory for internal features and labels, but impairs memory for superficial features. We conclude that explaining can influence learning by prompting children to favor inductively rich properties over surface similarity.


Keywords: Explanation; causal learning; category labels; non-obvious properties; inductive inference

## Introduction

The world has a complex structure, and the challenge of causal learning is to discover the nature of this structure to facilitate prediction and action. This is not a trivial task; it is sometimes impossible to predict how an object will behave based on its appearance. In fact, perceptually similar objects can be endowed with very different causal properties: Poison hemlock may look identical to wild carrot, but it is certainly not good to eat. Learning to override perceptual features in favor of non-obvious but inductively rich properties is thus an important achievement.

Previous research has examined the role of obvious (perceptual) properties versus non-obvious (internal or abstract) properties in children's inferences. Young children can use both perceptual and non-perceptual properties in categorizing objects (e.g., Gelman \& Markman, 1987; Gopnik \& Sobel, 2000), but adults often group objects according to common internal properties, labels, and causal affordances (regardless of perceptual similarity) in cases where young children tend to group objects based on salient perceptual similarity.

To illustrate this shift, consider the findings from Nazzi and Gopnik (2000). Children observed four objects placed on a toy, one at a time. Two of these objects were shown to be causally effective - they made the toy play music - and two were inert. One of the causal objects was then held up and labeled (e.g., "This is a Tib."), and children were asked to give the experimenter the other "Tib." In no-conflict trials, perceptual and causal properties were always correlated. However, in conflict trials, the same perceptual properties appeared across causal and inert objects. All children were more likely to choose the causal object in the no-conflict trials than in the conflict trials, but analyses of
conflict trials revealed a developmental shift: when generalizing the novel label, 3.5 -year-olds relied on perceptual cues over causal cues, while 4.5-year-olds relied on causal cues over perceptual cues.

Subsequent work has demonstrated a comparable shift in generalizing internal parts (as opposed to a category label). Sobel et al. (2007) used a similar procedure to demonstrate that 4-year-olds, but not 3-year-olds, are more likely to infer that objects have shared internal parts when they share causal properties than when they share external appearance.

These examples - and many others (see Keil, 1989; Gelman, 2003) - demonstrate that by 5 years of age, children begin to favor inductively rich but subtle cues, such as category membership and internal parts, over perceptual similarity when generalizing from known to unknown cases. But how is this transition achieved? Here we explore the hypothesis that the process of generating explanations is an important mechanism in scaffolding this transition.

## Explanation and Causal Learning

Accounts of explanation from both philosophy and psychology suggest an important relationship between explanation and causal learning: By explaining past observations we uncover information likely to support future judgments and interventions (e.g., Lombrozo, 2012; Walker, Williams, Lombrozo, \& Gopnik, 2012).

Consistent with this idea, research with adults finds that prompts to explain can improve learning (e.g., Chi et al., 1994) and promote the discovery and application of broad generalizations underlying what is being explained (e.g., Williams \& Lombrozo, 2010). Prior research also suggests that even young children's explanations have characteristics that make them well suited to highlighting inductively rich properties: they often invoke broad generalizations (Walker et al., 2012) and go beyond appearances (Legare, 2012).

For example, Walker et al. (2012) found that prompting preschool-aged children to explain causal events made them more likely to favor broad patterns in generalizing causal properties to novel objects. In the first of these studies, children were presented with evidence that was consistent with two candidate causes (e.g., "green objects make the toy go" versus "yellow objects make the toy go"), where one accounted for more observations. Children who were prompted to explain were more likely than controls to generalize according to the candidate cause that accounted for more of the data. In the second study, the cause that accounted for more of the data was contrasted with an alternative cause that was more consistent with children's
prior knowledge (e.g., "large blocks make the toy go"). In this case, those who explained were less likely to generalize according to the cause that accounted for more of the evidence, and instead privileged their prior knowledge.

Additionally, young children's explanations often appeal to non-perceptual properties, including unobservable causes (Legare, Wellman \& Gelman, 2009) and labels (Legare, Gelman, \& Wellman, 2010), and studies find that prompting children to explain can lead them to favor causal over perceptual learning (Legare \& Lombrozo, under review).

We therefore predict that by encouraging learners to consider broad generalizations, explaining can encourage learners to go beyond appearances to favor non-obvious but inductively rich properties as a basis for generalization.

In the following experiments, we use a method similar to Nazzi and Gopnik (2000) and Sobel et al. (2007) to examine whether generating explanations prompts children to favor generalizing internal parts (Experiment 1) and labels (Experiment 2) on the basis of causal over perceptual similarity. In Experiment 3, we examine whether the effects of explanation derive from a special relationship between explanation and inductively rich properties, or from a global boost in performance, as might be expected if explaining simply increased attention. Together, these experiments shed light on the role of explanation in the construction of generalizations that support causal inference.

## Experiment 1

Experiment 1 examined whether explanation influenced preschoolers' extension of a hidden, internal property to other objects that shared either perceptual or causal properties. Children observed four sets of three objects individually placed on a toy that played music when "activated" (see Gopnik \& Sobel, 2000). Each set contained three objects: one that activated the toy (target object), one that was perceptually identical to the target object, but failed to activate the toy (perceptual match), and one that was perceptually dissimilar to the target object, but successfully activated the toy (causal match). After each outcome was observed, children were asked to either explain (explain condition) or report (control condition) that outcome. Next, children received additional information about the target object: an internal part was revealed. Children were asked which one of the two other objects in the set (i.e., the perceptual match or causal match) shared the internal property with the target object. This method pit highly salient perceptual similarity against shared causal properties; children could base their generalizations on either one, but not both. We hypothesized that children who were asked to explain each outcome would be more likely than children in the control condition to consider the property with the greatest inductive richness and therefore select the causal match over the perceptual match.

## Method

Participants A total of 108 children were included in Experiment 1, with 363 -year-olds ( $M=40.9$ months; $S D=$
3.7, range: $35.8-47.7$ ), 364 -year-olds ( $M=53.3$ months; $S D=3.6$, range: $48.5-59.8$ ), and 365 -year-olds ( $M=64.4$ months; $S D=3.0$, range: $60.1-70.4$ ). Eighteen children in each age group were randomly assigned to each of the two conditions (explain and control).

Materials The toy was similar to the "blicket detectors" used in past research on causal reasoning (Sobel \& Gopnik, 2000), and consisted of a 10 " $\times 6$ " $\times 4$ " opaque cardboard box containing a wireless doorbell that was not visible to the participant. When an object "activated" the toy, the doorbell played a melody. The toy was in fact surrepticiously activated by a remote control.

Twelve wooden blocks of various shapes and colors were used (see Fig. 1). A hole was drilled into the center of each block. Eight blocks contained a large red plastic map pin glued inside the hole; the remaining four blocks were empty. All of the holes were covered with a dowel cap, which covered the opening to conceal what was inside. Each of the four sets of blocks was composed of three individual blocks. Two were identical in color and shape. One of these blocks (the target object) contained a map pin. The other of these blocks (the perceptual match) did not. The third block (the causal match) was perceptually dissimilar to the other two, and, like the target object, it contained a map pin.


Figure 1: Sample set of objects used in Experiment 1 (top) and Experiment 2 (bottom). Each row corresponds to a single set of items. There were a total of four sets.

Procedure Children participated in a brief warm-up game with the experimenter. Following this warm-up, the toy was placed on the table. The child was told, "This is my toy. Some things make my toy play music and some things do not make my toy play music." Then the first set of three blocks was brought out and placed in a row on the table. The order of presentation of the three blocks was randomized. One at a time, the experimenter placed a block on the toy. Two of the three blocks in each set (the target object and the causal match) caused the toy to activate and
play music. The perceptual match did not. After children observed each outcome, they were asked for a verbal response. In the explain condition, children were asked to explain the outcome: "Why did/didn't this block make my toy play music?" In the control condition, children were asked to describe the outcome (with a yes/no response): "What happened to my toy when I put this block on it? Did it play music?" After all three responses had been recorded, the experimenter demonstrated each of the three blocks on the toy a second time to facilitate recall.

Next the experimenter pointed to the set of objects and said, "Look! They have little doors. Let's open one up." The experimenter selected the target object and removed the cap to reveal the red map pin that had been hidden inside. The experimenter said, "Look! It has a little red thing inside of it. Can you point to the other one that also has something inside?" Children were then encouraged to point to one of the two remaining objects (i.e., the perceptual match or the causal match) to indicate which contained the same inside part, and this selection was recorded. Children could either select the block that was perceptually identical to the target or the object that shared the causal property, but not both.

Following their selection, children were not provided feedback, nor were they allowed to explore the blocks. Instead, all blocks were removed from view, and the next set was produced. This procedure was repeated for the three remaining sets. Each child participated in a total of four sets. Children were given a score of " 1 " for selecting the causal match and a " 0 " for selecting the perceptual match; children thus received $0-4$ points across the four sets.

## Results and Discussion

Data were analyzed with a 2 (condition) x 3 (age group) ANOVA, which revealed main effects of condition, $F(1$, $102)=50.70, p<.001$, and age, $F(2,102)=7.34, p<.01$ (see Fig. 2), with no significant interaction. Overall, children who were asked to explain $(M=2.98, S D=1.23)$ were more likely than controls $(M=1.61, S D=1.58)$ to generalize the internal part of the target object to the causal match rather than the perceptual match. Pairwise comparisons revealed no difference in performance between 3 - and 4 -year-olds, $p=.86$. However, 3- and 4-year-olds each selected the causal match significantly less often than 5-year-olds, $p<.01$.

One-sample $t$-tests were performed to assess whether explaining prompted children to override a preference for perceptual similarity and select the causal match. The 3-year-olds and 4-year-olds in the control condition selected the perceptual match significantly more often than chance, $t(17)=-3.69, p<.01$, and $t(17)=-2.53, p<.05$, respectively. Those in the explain condition selected the causal match significantly more often than chance, $t(17)=$ $3.01, p<.01$, and $t(17)=2.48, p<.05$, respectively. Five-year-olds in the control condition performed no differently from chance $(M=2.61, S D=1.72), t(17)=1.51, p=.15$, and 5 -year-olds in the explain condition selected the causal
match significantly more often than expected by chance, $t(17)=4.57, p<.001$.

These data suggest that in the absence of an explanation prompt, children relied primarily on information about the target object's salient perceptual features to predict whether a novel object would share an internal property. However, when children of the same age were asked to generate an explanation for the evidence that they observed, they instead privileged the target object's causal efficacy in making inferences about internal properties.



Figure 2: Average responses in explain and control conditions for Experiment 1 (top) and Experiment 2 (bottom). Higher numbers indicate a larger number of trials (of 4) on which an internal part (Experiment 1) or a label (Experiment 2) was generalized in line with a shared causal property over perceptual similarity.

Qualitative Data Explanations for the first object most often appealed to appearance (38\%), with a minority (5\%) appealing to internal properties. Explanations for the second set of objects, which occurred after observing the first set, appealed to appearance (33\%) and internal properties (32\%) equally often. By the final set, explanations most often appealed to internal parts $(38 \%) .{ }^{1}$

## Experiment 2

Experiment 2 examined whether the influence of explanation on children's inferences was restricted to consideration of internal parts, or whether these effects generalize to other inductively rich properties as well. A similar method was used to examine children's generalization of a novel label from a target object to either a perceptually-matched or a causally-matched object.

[^254]
## Method

Participants A total of 108 children were included in Experiment 2, with 36 3-year-olds ( $M=42.1$ months; $S D=$ 3.8, range: $35.9-48.0$ ), 364 -year-olds ( $M=54.0$ months; $S D=3.0$, range: $48.4-59.9$ ), and 365 -year-olds ( $M=65.0$ months; $S D=3.8$, range: $60.6-70.9$ ). Eighteen children in each age group were randomly assigned to each of the two conditions (explain and control).

Materials The same toy was used in Experiment 2. Twelve wooden blocks of various shapes and colors were also used. There were a total of four sets of objects, each containing three blocks. As in Experiment 1, two of these blocks (the target object and the perceptual match) were perceptually identical (same color and shape) and one of these blocks (the causal match) was distinct (see Fig. 1).

Procedure The procedure for Experiment 2 was identical to Experiment 1, with one major exception: Instead of revealing a hidden internal property, the experimenter held up the target object and labeled it, saying, "See this one? This one is a blicket! Can you point to the other one that is also a blicket?" Again, children received a total of four sets of objects, and could receive $0-4$ points.

## Results and Discussion

Data were analyzed with a 2 (condition) x 3 (age group) ANOVA, which revealed a main effect of condition, $F(1$, 102 ) $=13.51, p<.001$ (see Fig. 2), and no other significant effects. Overall, children who were asked to explain ( $M=$ $1.91, S D=1.83$ ) were more likely than controls ( $M=.72$, $S D=1.47$ ) to generalize the label to the causal match.

We next considered the data against chance responding. One-sample $t$-tests revealed that 3 -, 4 -, and 5 -year-olds in the control condition selected the perceptual match significantly more often than chance, $t(17)=-2.93, p<.01$, $t(17)=-3.69, p<.01$, and $t(17)=-3.10, p<.01$, respectively. In the explanation condition, the average of children's selections did not differ significantly from chance, $t(17)=.12, p=.90, t(17)=-1.26, p=.23$, and $t(17)$ $=.375, p=.712$, respectively. However, the data for this condition were distributed bimodally, with approximately half the children providing a majority of causal choices and half perceptual choices. The percentage of children selecting the causal match on 3 or 4 trials was $50 \%$ for 3 -year-olds, $44 \%$ for 4 -year-olds, and $56 \%$ for 5 -year-olds. The distribution of children's selections did differ significantly from that expected by chance in all age groups, $\chi^{2}(4)=$ 84.26, $p<.001, \chi^{2}(4)=66.49, p<.001$, and $\chi^{2}(4)=83.97, p$ $<.001$, respectively.

Like the younger children in Experiment 1, children in the control condition relied primarily on information about a target object's salient perceptual features to predict whether a novel object would share a category label. However, when children of all ages were asked to generate an explanation for the evidence that they observed, they
considered the target object's causal efficacy significantly more often in making inferences about shared labels.

Qualitative Data Appearance explanations were most common overall ( $28 \%$ ); however, there was an increase in the proportion of explanations appealing to kind or explicitly mentioning a label, with $7 \%$ in the first set and $19 \%$ in the final set.

## Experiment 3

The findings from Experiments 1 and 2 confirm our prediction that explanation encourages children to favor inductively rich properties (i.e., causality) as a basis for generalization. However, the findings are also consistent with an alternative explanation: that prompts to explain increased children's overall attention or engagement, resulting in "better" performance. Experiment 3 tests this alternative.

In Experiment 3, children were asked to explain or report causal outcomes after observing four unique objects, two of which activated the toy. Because we did not observe relevant age differences in Experiments 1-2, Experiment 3 was restricted to 4 -year olds. After each object was placed on the toy, three properties were revealed: an internal part, a label, and a sticker (added to the object). The internal parts and the labels correlated with the toy's activation (i.e., all and only objects that activated the toy had a particular inside part and label) while the sticker did not. Children then completed a memory task.

The purpose of Experiment 3 was to assess whether the effects of explanation observed in Experiments 1-2 could be due to a global and indiscriminate boost in attention. Based on our interpretation of Experiments 1 and 2, we predicted that a prompt to explain would improve memory for inside parts and labels, but not for the sticker, which was neither correlated with other properties nor plausibly inductively potent. If the effects of explanation can instead be attributed to a global increase in attention or engagement, one would predict improved memory for all features.

## Method

Participants A total of 364 -year-olds were included in Study 3 ( $M=53.8$ months; $S D=3.7$ months; range $=47.9-$ 59.7). Eighteen children were randomly assigned to one of two conditions (explain and control).

Materials Experiment 3 used the same toy as in the previous experiments. Four unique wooden blocks (distinct colors and shapes) were also used (see Table 1). As in Experiment 1, all blocks had a hole drilled into the center. Two of the blocks had a red, round plastic map pin glued inside and two of the blocks had a white, square eraser glued inside the hole. Four stickers were used during the experiment (two heart stickers and two star stickers). Several small cards were constructed as memory aids during the test phase of the experiment. Half of the cards had an image of a black music note (placed in front of the objects that children believed activated the toy), and half of the
cards had an image of a black music note crossed out with a red " $X$ " (placed in front of the objects that children believed did not activate the toy). Four additional cards were constructed: one with a red circle, one with a white square, one with a heart sticker, and one with a star sticker. These cards were used to facilitate the forced choice test.

Table 1. List of properties for objects used in Experiment 3.

|  | Object 1 | Object 2 | Object 3 | Object 4 |
| :--- | :---: | :---: | :---: | :--- |
| Causal | Yes | No | Yes | No |
| Internal | Red | White | Red | White |
| Label | Toma | Fep | Toma | Fep |
| Sticker | Heart | Heart | Star | Star |

Procedure As in Experiments 1 and 2, the experimenter introduced the toy. The experimenter then produced a single block and placed it on the toy. The child observed as the block did or did not cause the toy to play music. As before, children in the explain condition were asked to explain the outcome for each of the blocks and children in the control condition were asked to report the outcome (with a "yes/no" response). After the response was recorded, the experimenter repeated the demonstration a second time.

The experimenter provided three additional pieces of information about the object: the type of internal part was revealed ("Look! It has a little door on it! Let's open it up. Look, there is a [red]/[white] thing inside."), a label was provided ("See this one? This one here? This one is a [Fep]/[Toma]!"), and a sticker was placed on the bottom ("Now I am going to put a sticker on it! I am going to put a [heart]/[star] sticker on the bottom, see?"). The experimenter repeated each property twice, and then the block was removed from view. The entire procedure was repeated for the three remaining blocks, one at a time. All children observed the causal property first. The order of the remaining three properties was counterbalanced.

Next, the experimenter placed all four objects on the table in front of the child in random order, and told the child that they would now play a "memory game." To assess recall for the causal property of each object, the experimenter produced two cards - one with a music note, and one with a crossed out music note. The experimenter asked the child to point to the card that indicated whether the block did or did not play music. The child responded once for each of the four objects. Depending upon the child's response, the experimenter would then place an additional card (with a music note or a crossed-out music note) in front of the object, which would remain throughout.

To assess recall for the internal part, the experimenter produced two cards - one with a red circle and one with a white square. The experimenter asked the child to point to the card that indicated the type of thing inside the block. The child responded once for each of the four objects. To assess recall for the label, the experiment said, "Some of these blocks were called 'Tomas' and some of these blocks
were called 'Feps'. What was this one called, a 'Toma' or a 'Fep'?" The child responded once for each object. The order of presentation was counterbalanced across trials.

Finally, to assess recall for the type of sticker added to the block, the experimenter produced two cards - one with a heart sticker and one with a star sticker. The experimenter asked the child to point to the card that indicated the type of sticker added to the bottom of the block. The child responded once for each of the four objects.

Memory for internal parts, labels, and stickers was solicited in the same order as the corresponding properties were presented to that child in the demonstration phase of the experiment. For each property, children were given a score of " 1 " for accurate recall and a " 0 " for inaccurate recall. Because there were a total of four objects, each child could receive between 0 and 4 points for each property.

## Results and Discussion

Memory for the objects' causal properties was analyzed with a one-way ANOVA, which revealed that children in the explain condition were significantly more accurate ( $M=$ 3.93) than controls $(M=3.39), F(1,34)=8.42, p<.01$.

Next, a repeated measures ANOVA with the other object properties (internal part, label, sticker) as the repeated measure and condition (explain, control) as the between subjects variable revealed a main effect of object properties, $F(2,68)=6.96, p<.01$, and the predicted interaction between object properties and condition, $F(2,68)=8.30$, $p<.01$ (see Fig. 3). Children who were prompted to explain were significantly more accurate than controls in reporting the labels, $F(1,34)=9.34, p<.01$, but less accurate than controls in recalling the sticker type, $F(1,34)=5.16, p<.05$.


Figure 3: Average memory score (out of 4 trials) for each property assessed in Experiment 3. Error bars correspond to one SEM in each direction.

These data provide evidence against the possibility that engaging in explanation simply improves overall attention to the task. Instead, children who explained were more likely to recall the properties that were inductively rich, while ignoring a superficial, perceptual property that did not correlate with other features.

## General Discussion

In each of these experiments, prompting young children to explain made them more likely to privilege inductively rich, non-obvious properties over salient surface similarity in
making novel inferences. Children in the control condition, who were not prompted to explain, based their judgments on perceptual similarity. These effects were not restricted to a particular property or domain, as comparable effects were observed across two quite different properties: internal parts (Experiment 1) and novel labels (Experiment 2).

Although explanation led to fewer perceptual responses in Experiment 1 than in 2, this difference parallels the disparity in children's baseline performance in the control condition (see Fig. 2). In other words, children were more willing to privilege internal parts over appearances than labels over appearances, in line with previous research (Gopnik \& Sobel, 2000; Sobel et al., 2007). For our purposes, the critical finding is that explanation decreased perceptual responding in both cases.

Finally, the results of Experiment 3 provide additional support by demonstrating improved memory for a correlated cluster of inductively rich properties in children prompted to explain. Importantly, Experiment 3 also provides evidence that effects of explanation are selective: Children who explained had impaired memory for the superficial property. This provides evidence against the idea that explanation produces a general benefit for learning by globally and indiscriminately increasing engagement or motivation.

Why might explaining lead children to favor inductively rich properties? Wellman and Liu (2007) suggest that explaining makes an occurrence sensible by reference to a larger framework: The explainer appeals to the interplay between evidence and current theories to construe the phenomenon as an instance of a larger, coherent system. In line with this idea, recent computational approaches to cognitive development have proposed that the formation of generalizations at multiple levels of abstraction enables learners to learn quickly and generalize effectively to novel cases (Kemp, Perfors \& Tenenbaum, 2007). By prompting children to favor inductively rich regularities, explanation could play a role in pushing children beyond immediate observations to consider higher-order generalizations that support abstract knowledge.

Similarly, we have argued that engaging in explanation constrains the learner to consider an event as an instance of a broad generalization (see also Lombrozo, 2012). Recall that previous research found that explaining magnified effects of prior knowledge in the service of broad generalization (Walker et al., 2012). In the current study, the belief that internal parts and labels are inductively rich is itself an important instance of higher-order prior knowledge. We propose that explaining contributes to the formation of causal theories by constraining learners to consider properties that are most likely to generalize to novel cases. In the experiments presented here, this constraint improved children's ability to override highly salient perceptual cues.

## Acknowledgments

Thanks to Rosie Aboody, Dustin Bainto, Ela Banerjee, Brynna Ledford, Donna Ni, Gabriel Poon, UC Berkeley Early Childhood Centers, and Lawrence Hall of Science.

Research supported by the McDonnell Foundation and NSF grant DRL-1056712 to T. Lombrozo.

## References

Chi, M. T. H., DeLeeuw, N., Chiu, M.H., \& LaVancher, C. (1994).Elicitingself-explanations improves understanding. Cog Sci, 18, 439-477.
Gelman, S. \& Markman, E. (1987). Young children's induction from natural kinds: The role of categories and appearances. Child Dev, 58: 1532-1541.
Gelman, S. (2003). The essential child: Origins of essentialism in everyday thought. Oxford: Oxford Press.
Gopnik, A. \& Sobel, D. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction, Child Dev, 71(5): 12051222.

Keil, F.C. (1989). Concepts, kinds, and cognitive development. Cambridge, MA: MIT Press.
Kemp, C., Perfors, A., \& Tenenbaum, J.B. (2007). Learning overhypotheses with hierarchical Bayesian models. Dev Sci, 10, 307-321.
Legare, C.H. (2012). Exploring explanation: Explaining inconsistent information guides hypothesis-testing behavior in young children. Child Dev, 83, 173-185.
Legare, C.H., Gelman, S.A., \& Wellman, H.M. (2010). Inconsistency with prior knowledge triggers children's causal explanatory reasoning. Child Dev, 81, 929-944.
Legare \& Lombrozo (under review). Unique and selective benefits of explanation and exploration on learning in early childhood.
Legare, C.H., Wellman, H.M., \& Gelman, S.A. (2009). Evidence for an explanation advantage in naïve biological reasoning. Cog Psych, 58, 177-194.
Lombrozo, T. (2010). Explanation and abductive inference. In K.J. Holyoak and R.G. Morrison (Eds.), Oxford Handbook of Thinking and Reasoning (pp. 260-276), Oxford, UK: Oxford Univ. Press.
Nazzi, T. \& Gopnik, A. (2000). A shift in children's use of percetual and causal cues to categorization. Dev Sci, 3(4): 389-396.
Sobel, D.M. Yoachim, C.M. Gopnik, A. Meltzoff, A.N. \& Blumenthal, E.J. (2007). The blicket within: Preschoolers' inferences about insides and causes. J. of Cog Dev, 8(2): 159-182.
Walker, C.M., Williams, J.J., Lombrozo, T. \& Gopnik, A. (2012). Explaining influences children's reliance on evidence and prior knowledge in causal induction. In N. Miyake, D. Peebles, \& R.P. Cooper (Eds.), Proceedings of the $34^{\text {th }}$ Ann Conf of the Cog Sci Soc, pp. 1114-1119. Austin, TX: Cog Sci Soc.
Wellman, H.M. \& Liu, D. (2007). Causal reasoning as informed by the early development of explanations. In Causal Learning: Psychology, Philosophy, \& Computation (eds.) L. Schulz \& A. Gopnik, pp. 261-279.
Williams, J.J. \& Lombrozo, T. (2010). The role of explanation in discovery and generalization: Evidence from category learning. Cog Sci, 34: 776-806.

# 24-Month-Olds Engage in Relational Causal Reasoning 

Caren M. Walker (caren.walker@berkeley.edu)<br>Department of Psychology, University of California, Berkeley, 3210 Tolman Hall, Berkeley, CA 94720 USA

Alison Gopnik (gopnik@berkeley.edu)<br>Department of Psychology, University of California, Berkeley 3210 Tolman Hall, Berkeley, CA 94720 USA


#### Abstract

Children make inductive inferences about the causal properties of individual objects from a very young age. When can they infer higher-order relational properties - a task that has proven difficult for non-human primates? In two experiments, we examined 18 -24-month-old infants' relational inferences using a causal version of a relational match-to-sample task. Results suggest that by 21-24 months of age, infants are able to infer a relational causal principle from just a few observations and use this inference to guide their own subsequent actions and bring about a novel causal outcome. Findings are considered in light of recent discussion about the nature of relational and causal reasoning, and their evolutionary origins.


Keywords: Cognitive development; infancy; relational reasoning; causal learning; inference

## Introduction

Learning about causal relationships is one of the most important and challenging problems young humans face. Causal knowledge allows you to act on the world - if you know A causes B, you can act on A to bring about B. Recent research shows that children as young as 19 to 24 months of age can quickly learn causal properties of objects from patterns of statistical contingency and can act on that knowledge to bring about effects (Gopnik, 2012; Gopnik \& Wellman, 2012; Sobel \& Kirkham, 2006; Meltzoff, Waismeyer \& Gopnik, 2012). At 20 months, children can infer the desires of others from sampling patterns (Kushnir, $\mathrm{Xu} \&$ Wellman, 2010) and at 16 months, they can use contingency information to determine whether an effect was caused by their own actions (Gweon \& Schulz, 2011). Other lines of research suggest that infants can infer abstract linguistic structure from statistical data (e.g. Saffran, Newport \& Aslin, 1996; Lany \& Gomez, 2008).

However, little is known about the development of children's ability to infer higher-order relational causal principles from data. In particular, an effect might be caused by an object property (e.g., red blocks activate a toy) or by a higher-order relationship between properties (e.g. two blocks that are the same, regardless of their color, activate a toy). Inferring higher-order relations is essential for building abstract knowledge (Kemp, Perfors \& Tenenbaum, 2007; Dewar \& Xu, 2010) and reasoning about concepts that are not tied to perceptual properties. The ability to form generalizations about higher-order relations
from limited data allows children to make principled abstractions that go beyond the particular properties they have observed.

To investigate this, we used a causal version of Premack's (1983) match-to-sample task. In this task, animals observe an abstract relational pattern - AA', BB', and CC' all lead to a reward. Then they are given a choice between AB (object match) and DD' (relational match). Although A and B have each individually been associated with the reward, an animal who has inferred the more abstract relational pattern ("same") should choose DD'. Premack found that chimpanzees could not solve this relational task without hundreds of trials (Premack, 1988) or explicit training to use linguistic symbols for "same" (Premack, 1983; Premack \& Premack, 1983; 2002). Additional comparative studies have confirmed this pattern for non-human primates and other animals (Penn, Holyoak, \& Povinelli, 2008). These observations have led some researchers to conclude that abstract relational reasoning may be uniquely human.

Research examining the origins of relational reasoning using looking-time measures suggests that human infants, like primates, may be able to recognize relational patterns of data (Dewar \& Xu, 2010; Tyrell, Stauffer \& Snowman, 1991; Ferry, Hespos \& Gentner, 2012). However, there is no evidence that infants can use those patterns to make causal inferences or guide actions. In fact, earlier studies have concluded that even preschoolers have difficulty with relational tasks (Christie \& Gentner, 2010; Gentner, 2010). Not unlike chimpanzees, children succeeded only when given linguistic coaching to point out the pattern of similarity between two simultaneously presented cases. Even when explicitly instructed to compare objects, 3-yearolds' performance on these relational tasks was rather tenuous, dropping significantly below chance when the test items were presented sequentially, rather than simultaneously (Christie \& Gentner, 2010). This might lead to the conclusion that, even in humans, learning higherorder relations and using them to guide actions is a relatively late-developing ability, which depends on direct instruction, language, and cultural scaffolding.

However, the striking success of young children on causal tasks suggests that placing these problems in a causal context might enhance performance. For example, recent evidence suggests that by 24 months, toddlers readily learn
novel causal relations by observing others acting causally on the world, and use this information to fashion their own actions to achieve the same causal outcomes (Meltzoff, Waismeyer, \& Gopnik, 2012). In the current study, we used a similar observational learning paradigm to examine whether infants as young as 18 - to 24 -months could abstract a relational property from their observations in a manual causal task. If infants succeed, this would suggest that the human ability to learn abstract relations is in place earlier than previously thought. It would also suggest that these abilities might be responsible for the impressive learning of very young children. Could infants solve these relational problems spontaneously, and without linguistic cues or explicit directives to compare, if evidence for the existence of a relational property were provided in a causal context?

## Experiment 1

In Experiment 1, the experimenter introduced 18- to 24-month-old infants to a novel toy that played music when "activated," and 3 unique pairs of identical blocks AA', BB', and CC'. Infants observed as the experimenter placed blocks on the toy, one at the time. In 3 demonstrations, infants observed that while individual blocks failed to activate the toy alone, pairs of identical blocks did produce the effect. Immediately after this brief training, we examined whether these infants learned the novel relational property (i.e., pairs of identical objects make the toy play music) by placing a novel block on the machine, asking the infants to generate the effect on their own and observing their first selection.

## Methods

Participants A total of 46 18- to 24-month-old infants participated in Experiment 1 ( $M=20.9$ months; $S D=2.0$ months; range $=18.0-24.4$ months; 22 girls). Five additional children were tested but excluded for fussiness during the training phase or for failing to respond to the experimenter during test trials. Children were recruited from daycare centers and museums, and a range of ethnicities resembling the diversity of the population was represented.

Materials The toy was designed to be similar to the "blicket detectors" used in past research (see Gopnik \& Sobel, 2000). The toy consisted of a 10 " $\times 6$ " $\times 4$ " opaque box constructed from cardboard and painted white with blue borders. The box contained a wireless doorbell that was not visible to the participant. When a block "activated" the toy, the doorbell played a novel melody. The toy was in fact surrepticiously activated by a remote control that was held out of view by the experimenter. Six painted wooden blocks in assorted colors and shapes ( 3 unique pairs of 2 identical blocks) were placed on the toy during the training phase in Experiment 1. Six additional blocks were used during the test phase in Experiments 1 and 2, including 2 novel pairs of identical blocks and 2 unique individual blocks.

Procedure The procedure for Experiment 1 is illustrated in Figure 1. Following a warm-up period in which the child was familiarized with the experimenter, the toy was placed on the table. The experimenter said, "This is my toy. Some things make my toy play music and some things do not make my toy play music." Children then observed while the experimenter placed 6 blocks ( 3 unique pairs: AA', BB', $C^{\prime}$ ') on the table in front of the toy. The experimenter said, "Let's try!", selected a block (A), and placed it on top of the toy. No effect was produced. After a brief pause, the experimenter again said, "Let's try!" and selected the paired block (A') and placed it next to the first block (A) on top of the toy. This pair of objects (AA') activated the toy, which played a novel melody. The experimenter smiled and said, "Music!" Both blocks were removed from the toy and returned to the pile of 6 blocks. This procedure was repeated with the two remaining pairs ( $\mathrm{BB}^{\prime}$ and $\mathrm{CC}^{\prime}$ ). The order of the pairs was randomized. Following all three demonstrations, the 6 training blocks were removed from view. Blocks were placed on the toy one at a time due to the causal nature of the task: In order to provide evidence for the conjunctive causal relation (that both blocks were necessary to activate the toy), infants must observe a single block fail to activate the toy on its own.

Immediately following the training phase, the experimenter produced 3 test blocks ( 1 novel paired block (D), 1 familiar block (A), and 1 novel distractor block (E) and placed them in a row on the table. The order of presentation was randomized. The experimenter said, "Let's try!", produced the target block (D'), and placed it on top of the toy. No effect was produced. The experimenter then pushed the toy and a tray with all 3 test blocks towards the child, and asked, "Can you pick one of these (pointing to the row of test blocks) to make my toy play music?" The first test block that the child placed on the toy was recorded. The toy activated if the child correctly selected the novel paired block (D). If the child selected the familiar block (A) or the novel distractor block (E), the toy failed to activate. After this feedback was provided, this procedure was repeated a second test trial with a new set of test blocks.

If infants were acting based upon the previous association between the block and the effect, they should choose the familiar block (A). If they simply preferred to try novel blocks they should pick the novel distractor block (E) as often as the novel paired block (D). However, if infants were able to learn the relational causal property, then they should select the novel paired block (D) to produce the effect.

Coding Children received 1 point for selecting the novel paired block and 0 points for selecting either of the other two blocks. Therefore, children in the Experiment 1 could receive up to 2 points for their performance across the two test trials. Children's responses were recorded by a second researcher during the testing session, and all sessions were video recorded for independent coding by a third researcher who was naïve to the the hypotheses of the experiment.

Interrater reliability was very high; the two coders agreed on $99 \%$ of the children's responses to the test questions. Two minor discrepancies were resolved by a third party.


Test Trials $1 \& 2$


Figure 1: Schematic representation of training and test trials in Experiment 1. On each training trial, the experimenter first placed a single block on the toy (no activation) and then added an identical block, activating the toy. The procedure was repeated for all 3 training pairs. On each test trial, 3 test blocks (novel distractor block [ND], familiar block [F], novel paired block [NP]) were presented. The experimenter then placed the target block on the toy, yielding no effect. The child was asked to select one test block to activate the toy.

## Results \& Discussion

Across the two test trials, infants inferred the relational property and selected the novel paired block (D) more often than expected by chance $(M=.91, S D=.69$; chance $=.66)$, $t(45)=2.47, p<.02$ (Fischer exact test revealed no order effects for test trials, $p=.39$ ).

Linear regression revealed a significant developmental change in performance on test trials between 18 and 24
months of age, $F(1,44)=8.23, p<.01$. The regression model predicts that while the youngest children in our sample (18-month-olds) perform just above chance values (chance $=.66$ ), by 21 months, children select the novel paired block on at least half of the test trials.

To further investigate this change, we divided infants into two age bins: 18-21 months and 21-24 months. Older infants performed significantly better than chance, $(M=$ 1.13, $S D=.82$ ), $t(22)=2.77, p<.02$, and significantly better than younger infants, $F(1,44)=4.91, p<.05$ who performed at chance, $(M=.70, S D=.47), t(22)=.36, p=$ .72. Older infants chose the novel paired block significantly more often than the novel distractor block (binomial, $p<.05$ for both trials 1 and 2) and the familiar block (binomial, $p<$ .01 for both trials 1 and 2) .

Previous proposals have suggested that children are unable to reason relationally because they tend to focus on the individual objects which have been previously associated with the outcome, thus interfering with their ability to detect the relation (e.g., Gentner, 2010). We show no evidence of this. In fact, only $33 \%$ of infants who answered incorrectly on a given trial selected the familiar block over the novel distractor block. This is particularly surprising, given that this block had been associated with the effect during the training trials. Instead, significantly more incorrect selections were due to infants' choice of the novel distractor block ( $60 \%$ ) over the familiar block, $p<$ .05. This suggests that the younger infants' failure may have been due to a preference for exploring the novel block.

Results indicate that by 21-24 months of age, infants are able to infer a relational causal principle from a few pieces of evidence, and use this inference to bring about a novel causal outcome. However, these data do not rule out some alternative interpretations: Infants may have succeeded on this task by "matching" the experimenter's selection or because they preferred to create pairs on the toy, regardless of training. Experiment 2 was designed to address these alternatives.

## Experiment 2

The procedure for Experiment 2 was identical to Experiment 1, except that infants did not observe the training trials. Infants were therefore given no evidence for the relational property. Instead, after being introduced to the toy, infants were immediately presented with a test trial. If infants were simply matching the experimenter or had a preexisting preference for pairs of blocks, then performance should not differ significantly from the infants in Experiment 1. However, if these alternatives are insufficient to explain the infants' success, then infants should perform at chance.

## Method

Participants Twenty-two 21-24-month-olds participated ( $M$ $=22.8$ months; $S D=1.3$ months; range $=21.5-24.8$ months; 10 girls). Two additional children were tested but
excluded for failing to respond. Recruitment procedures and demographics were identical to Experiment 1.

Materials \& Procedures Materials and procedures were identical to Experiment 1. However, infants did not observe the training trials. Instead, after infants were introduced to the toy, they were given a single test trial. Only one test trial was administered, in order to avoid providing feedback. Therefore, infants could receive 0 or 1 point. Interrater reliability for Experiment 2 was $100 \%$.

## Results \& Discussion

In the absense of evidence for the relational principle, only $36 \%$ ( 8 out of 22) of infants selected the paired block, [binomial test, $p=.72, \mathrm{~ns}$, which was significantly different from the infants of the same age on their first trial in Experiment $1, p<.05$ by Fischer's exact test. These results demonstrate that the findings from Experiment 1 could not have been the result of imitation or a preexisting bias to prefer pairs.

## General Discussion

These findings suggest that the differences in relational reasoning between humans and non-human primates may be in place very early, and that human infants can succeed on match-to-sample tasks in a causal context without explicit linguistic cues or instruction. On the other hand, the failure of younger infants may suggest that the ability to use language may play a role. Alternatively, failure may be due to other factors that make it difficult for younger infants to display competence in manual tasks, such as a general impulse to explore novel objects. Additional research is needed to examine whether relational abilities are supported by the development of linguistic capacities - and language production in particular. To this end, we are currently examining the relationship between infants' performance on the causal match-to-sample task and their general language skills. In particular, we are examining infants' comprehension and production of relational words (e.g., "more").

The method outlined in this paper provides a novel and powerful paradigm for assessing relational reasoning in a causal context. Importantly, this method minimizes the need for verbal guidance, and is thus suitable for very young children. Earlier "blicket detector" studies using very similar methods have confirmed that children's inferences in these tasks go beyond simple associative learning and have the distinctive profile of causal inferences. For example, children will use inferences about the causal relation of the block and machine to design novel interventions - patterns of action they have never actually observed - to construct counterfactual inferences and to make explicit causal judgments, including judgments about unobserved hidden features of the objects (e.g. Gopnik \& Sobel, 2000; Schulz, Gopnik, \& Glymour, 2007; Sobel, Yoachim, Gopnik, Meltzoff, \& Blumenthal, 2007).

However, due to the constraints of the particular causal context (i.e., the need to provide evidence for the conjunctive relation over the disjunctive relation, noisy OR ), we opted to present evidence one block at a time, rather than in simultaneously presented pairs. In the earlier primate studies, the canonical relational tasks presented the pairs simultaneously, so that the animals had to choose between pairs of AA and BB . This difference in procedure may have led to the divergent results between infants and primates.

We have recently completed an additional follow-up experiment exploring this possibility (Walker \& Gopnik, under review). In this study, 18- to 30 -month-old infants ( $M=25.7$ months) were divided into one of two conditions: same or different. In the same condition, infants were given two pieces of evidence that pairs of "same" objects (AA', BB', CC') simultaneously placed on the toy produce the effect. In order to provide evidence for a conjunctive causal relationship, we also provide two pieces of evidence that pairs of "different" objects (DE, FG, HI) fail to produce the effect. In the different condition, infants were given the same four pieces of evidence, with the causal pattern reversed: "different" pairs (i.e., DE) produced the effect, while "same" pairs (i.e., AA') failed to do so. By combining positive and negative evidence, we were able to use a similar causal method to demonstrate that infants are able to learn the relational properties "same" and "different." Results of this study provide strong evidence that 2-yearolds are able to quickly learn relational causal principles, with $81 \%$ of children selecting the relational match in both the same and different conditions.

Clearly, toddlers are able to rapidly learn abstract relational causal principles from minimal evidence and use them to guide their subsequent actions in the world. This ability appears to be in place surprisingly early in human development. It emerges only a few months after the first evidence of the ability to learn about specific causal properties from contingency. This may help explain how children acquire the impressively general and abstract causal knowledge evident in early "intutive theories" (Gopnik \& Wellman, 2012; Carey, 2010).

## Acknowledgments

Research was funded by the James S. McDonnell Foundation and the National Science Foundation (BCS1023875) to A. Gopnik. We thank the parents and children who participated and the University of California, Berkeley Early Childhood Centers, Lawrence Hall of Science, and Habitot. We are grateful to Fei Xu and Tania Lombrozo for their feedback. Finally, we thank Rosie Aboody, Anna Akullien, Ela Banerjee, Sierra Eisen, and Brynna Ledford for facilitating data collection.

## References

Byrne, R.W. (1995). The thinking ape: Evolutionary origins of intelligence. New York: Oxford University Press.

Carey, S. (2010). The origin of concepts. USA: Oxford University Press
Christie, S. \& Gentner, D. (2010). Where hypotheses come from: Learning new relations by structural alignment. Journal of Cognition \& Development, 11, 356-373.
Dewar, K.M. \& Xu, F. (2010). Induction, overhypothesis, and the origin of abstract knowledge: Evidence from 9-month-old infants. Psychological Science, 43, 1227-1238.
Ferry, S., Hespos, S., \& Gentner, D. (2012). Prelinguistic relational concepts: Investigating the origins of analogical reasoning in infants. Poster presented at the $18 t h$ Biennial International Conference on Infant Studies, in Minneapolis, MN, June 7, 2012.
Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. Cognitive Science, 34, 752-775.
Gentner, D., Anggoro, F.K., \& Klibanoff, R.S. (2011). Structure mapping and relational language support children's learning of relational categories. Child Development, 82, 1173-1188.
Gopnik, A. (2012). Scientific thinking in young children: Theoretical advances, empirical research, and policy implications. Science, 337, 1623-1627.
Gopnik, A. \& Sobel, D. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction, Child Development, 71(5): 1205-1222.
Gopnik, A. Wellman, H. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. Psychological Bulletin, 138, 1085-1108.
Gweon, H. \& Schulz, L. (2011). 16-month-olds rationally infer causes of failed actions. Science, 332, 1524.
Heyes, C. \& Frith, U. (Eds.) (2012). New thinking: The evolution of human cognition (Theme Issue). Philosophical Transactions of the Royal Society, B, 367.
Kemp, C., Perfors, A., \& Tenenbaum, J.B. (2007). Learning overhypotheses with hierarchical Bayesian models. Developmental Science, 10, 307-321.
Kushnir, T., Xu, F., \& Wellman, H.M. (2010). Young children use statistical sampling to infer the preferences of other people. Psychological Science, 21, 1134-1140.
Lany, J.A. \& Gómez, R.L. (2008). Twelve-month-olds benefit from prior experience in statistical learning. Psychological Science, 19, 1247-1252.

Meltzoff, A., Waismeyer, A., \& Gopnik, A. (2012). Learning about causes from people: Observational causal learning in 24-month-olds infants. Developmental Psychology, Online First Publication.
Penn, D.C., Holyoak, K.J., \& Povinelli, D.J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. Behavioral and Brain Sciences, 31, 109-178
Premack, D. (1983). The codes of man and beasts. Behavioral \& Brain Sciences, 6, 125-167.
Premack, D. (1988). Minds without language. In L. Weiskrantz, (Ed.), Thought without language (pp. 46-65). Oxford: Clarendon Press.
Premack, D. \& Premack, A.J. (1983). The mind of an ape. New York: W.W. Norton.
Premack, D. \& Premack, A.J. (2002). Original intelligence: Unlocking the mystery of who we are. New York: McGraw-Hill.
Saffran, J.R., Newport, E.L., \& Aslin, R.N. (1996). Statistical learning by 8-month-old infants. Science, 274, 1926-1928.
Schultz, L. Gopnik, A., \& Glymour, C. (2007). Preschool children learn about causal structure from conditional interventions. Dev Sci. 10(3): 322-332.
Sobel, D. \& Kirkham, N.Z. (2006). Blickets and babies: The development of causal reasoning in toddlers and infants. Developmenal Psychology, 42, 1103-1115.
Sobel, D.M. Yoachim, C.M. Gopnik, A. Meltzoff, A.N. \& Blumenthal, E.J. (2007). The blicket within: Preschoolers' inferences about insides and causes. J. of Cog Dev, 8(2): 159-182.
Taylor, H.N., Miller, R., \& Grey, R.D. (2012). New Caledonian crows reason about hidden causal agents. Proceedings of the National Academy of Science. Online Sept. 17, 2012.
Tomasello, M. \& Call, J. (1997). Primate cognition. New York: Oxford University Press.
Tyrrell, D.J., Stauffer, L.B., \& Snowman, L.G (1991). Perception of abstract identity/difference relationships by infants. Infant Behavior \& Development, 14, 125-129.
Walker, C.M. \& Gopnik, A. (under review). Infants infer higher-order relational principles in causal learning.

# Programming of saccades to double-step targets in scene viewing: A test of assumptions present in the CRISP model 

R. Calen Walshe (r.c.walshe@sms.ed.ac.uk)<br>Antje Nuthmann (antje.nuthmann@ed.ac.uk)<br>University of Edinburgh, Psychology Department, 7 George Square, Edinburgh EH8 9JZ


#### Abstract

Several computational models explaining fixation durations in scene viewing (Nuthmann, Smith, Engbert, \& Henderson, 2010) and in reading (Engbert, Nuthmann, Richter, \& Kliegl, 2005; Reichle, Pollatsek, Fisher, \& Rayner, 1998) assume that saccade programming is completed in two stages: an initial, labile stage that is subject to cancellation and an subsequent, non-labile stage in which the program can no longer be cancelled. This distinction is motivated by findings from doublestep experiments that used much simpler situations than scene viewing or reading. Here, we adopt a classic double-step paradigm to a scene-viewing context. In a Static condition targets are presented to the left or right of a central fixation cross along a horizontal axis while in a Scene condition targets are presented in a gaze contingent manner along a trajectory defined by the location of recent fixations. We found evidence in support of the claims that saccade cancellation occurs within a naturalistic scene-viewing context and that saccade cancellation can account for increases in observed fixation duration distributions. The duration of the non-labile stage was estimated to be longer in the Scene condition compared to the Static condition.


Keywords: Double-step; Scene viewing; Saccade programming; Mixed-effects modelling

## Introduction

There is a long history of utilizing the double-step paradigm to explore the lower level details of the programming and execution of eye-movements (Westheimer, 1954). Classic variations of the double-step paradigm involve presenting participants with two targets along a horizontal axis with a varying inter-stimulus interval separating the two targets. For instance, in one classic study of saccade programming that utilized double-step stimuli, Becker and Jürgens (1979) had a condition in which a first target was presented at $15^{\circ}$ to the left or right of fixation with a second target presented at $30^{\circ}$ in the same direction at delays of $50,100,150$ and 200 ms . The participants task was to fixate the target as quickly as possible, which meant that in order to fixate the more distal target, a saccade program initiated to the first target was put in competition with a program to the second target.

By studying the characteristics of the response pattern, the paradigm affords numerous avenues to investigate the processes underlying the programming of saccades. One method of formalizing double-step data has been to produce what is called an amplitude transition function (ATF) (Becker \& Jürgens, 1979). The ATF provides a measure of the saccade amplitude resulting from the stimuli as a function of the delay $(D)$ which measures the time elapsed between the onset of the second target step and the first measured response saccade. Therefore, in this analysis only those trials in which both targets appeared prior to the first response saccade are analysed.
$D$ can therefore be thought of as the amount of time available to the saccadic system to reprogram an eye-movement to the second target. Frequently replicated results demonstrate that when reprocessing time in low (short values of $D$ ) saccades are typically directed towards the first target step, and when reprocessing time is high (high values of $D$ ) then saccades compensate for the updated target position and move to the second target step (Ludwig, Mildinhall, \& Gilchrist, 2007).

From inspection of the ATF it is also apparent that there is a point at which the appearance of the second target step can no longer have an influence on saccade programming. This region of the distribution represents a "point of no return" in the saccade processing to the first target and as such the program to the first target is executed despite the availability of countermanding information from the second target. The point in processing at which a saccade program can no longer be modified by a second target is also referred to as saccadic dead time (SDT) and has been estimated at approximately 80 ms prior to the execution of a saccade (Ludwig et al., 2007).

The double-step paradigm has been a fruitful one in elucidating the basic properties of the occulomotor system. The principles derived from such investigations have formed the basis of several models of eye-movement control in a variety of fields. These investigations have proven particularly useful in models that attempt to explain the mechanisms that control how long aspects of the visual environment are fixated. For instance, Nuthmann et al. (2010)'s CRISP model which explains fixation durations in scene viewing, utilized a two stage saccade programming mechanism. In the first labile stage of programming a saccade could be cancelled and reprogrammed, while a program that had moved into the nonlabile stage could no longer be cancelled. In the CRISP model architecture saccade cancellation acts as a causal mechanism that accounts for systematic delay in fixation durations. The theoretical dichotomy between a labile and nonlabile stage of programming was first introduced in Reichle et al. (1998)'s EZ Reader model of eye-movement control in reading. While these models borrow the distinction from classic double-step results, it has never been formally tested within the domains to which the models apply. In the current study, a classic approach to studying double-step stimuli is adapted to a naturalistic scene viewing context.

In summary, the scene-based double step experiment has several concrete aims. Firstly, the assumption that is inherent in several influential models both in scene viewing and in reading is that delays in the latency of fixations can be
partially accounted for by the time required to cancel and reinitiate a saccade program to a novel stimulus. These models often cite basic research into saccadic programming although little work has been done to verify these findings for the more naturalistic case. Therefore, our study investigates saccade programming within scene viewing by adapting a classic double-step paradigm to the scene viewing context. Secondly, by including a classic version of the double-step paradigm as an experimental condition we are able to directly compare performance across tasks.

## Method

Participants were all University of Edinburgh undergraduate students that were paid $£ 7$ in compensation for their time. Each participant was presented with 100 trials in each condition (Static vs Scene). In the scene condition 100 unique colour photographs were presented at a resolution of $800 \times 600$. Stimuli were presented on a 21 -inch CRT monitor and participants were seated at a distance of 67 cm from the monitor. Eye-movements were recorded with an SR Research Eyelink 1000 desktop system operating at 1000 Hz . Out of the 16 subjects tested, 4 were rejected for poor data quality. Of the remaining 12 participants the mean age was 23 and 10 participants were female and 2 were male.

## Experiment Overview

Double-step experiments typically involve having a participant fixate to a location while a stimulus is displayed at a distal location. At varying delays, this target is then shifted to a new location. The participant is instructed to make a fixation to the final location of this double-stepped stimuli. At short delays, the participant is frequently able to interrupt whatever processing may have been made to the first target, and instead program a saccade to the second target location. Important aspects of the saccade motor system can be derived by looking at the time course of the response. In order to investigate double-step performance in a more naturalistic environment we adapted a single experimental condition from a classic double-step experiment (Becker \& Jürgens, 1979) to a context in which participants received the double-step stimulus while they were actively viewing a natural scene. Furthermore, a replication of the Static double-step condition was included as a baseline measure.

## Static Condition

Participants fixated a cross located in the centre of the screen. The first target step was presented after a variable delay of between 2000-3000 ms. The first target step was presented to either the left, or to the right of the fixation cross. The target step delay (TSD), the duration elapsed between presentation of the first and second targets, was either 50, 100, 150 , or 200 ms . The presentation side and delay durations were counterbalanced. Furthermore, in order to ensure that participants did not simply postpone their responses and wait until the presentation of the second target step had appeared before making a response, $20 \%$ of trials were single step trials
in which only a single target was presented to the first target location. These single steps also provide a baseline for saccade response parameters that can be compared to those on the double-step trials. The first target was always presented at exactly $7^{\circ}$ on the horizontal axis and the second target was always presented at $14^{\circ}$ along the same axis.


Figure 1. Target steps in the Static condition move in the horizontal plane either to the left or right of fixation cross. The first target step moves to an eccentricity of $7^{\circ}$ and the second target step moves in the same direction to an eccentricity of $14^{\circ}$. The second step follows the first at a delay of either $50,100,150$ or 200 ms . In no-step trials, the first target step to $7^{\circ}$ is not followed by a movement to the second target step at $14^{\circ}$.

The instructions to the participant were that they were to "chase the pink box" with their eyes. Their task was to fixate the box as quickly as possible.

After 1000 ms the trial was terminated and a new trial was initiated once the fixation detection procedure had assured that the participants' gaze was directed towards the central fixation cross.

## Scene Condition

The instructions to the participant in the scene condition were that they would be required to memorise a scene for a later recall test. However, this recall test was never applied. Furthermore, participants were instructed that they would see pink boxes appear while memorising the image, and that when they see these pink boxes that they should "chase the pink box" with their eyes. The memorisation task was included in order to observe performance in a more naturalistic and cognitively demanding context. All temporal characteristics of the double-stepped stimuli were identical to those of the Static condition. In the scene condition $20 \%$ of trials only had a single step at an eccentricity of $7^{\circ}$ in order to avoid participants making anticipatory saccades to the second target location at $14^{\circ}$. The first target step was presented once 15 saccades had been made and the scene had been viewed for at least 4 seconds. The first target step was always presented while a fixation was in progress, and this fixation could be at any possible location on the screen. A further difference from the Static double-step condition is that steps were not simply placed on the central horizontal plane as was done in the Static condition. In order to maximize the similarity be-
tween the Static and Scene conditions, while also adapting the study to a naturalistic context, the double-step manipulation trajectory was determined by the line intersecting the current fixation and the last recorded fixation. The first target was presented at $7^{\circ}$ along this line in the same direction as the eye-movement plotted from the nth and $n-1$ th saccade. The second target was presented at $14^{\circ}$ on the same line. As was done in the Static condition the second target was presented in the same direction as the first (See Figure 2 for details). In circumstances such that projecting the targets along the line of presentation would result in a target being presented off the dimensions of the screen, the presentation procedure was delayed until a fixation occurred such that the presentation of the targets would not occur off screen.

The decision to place the targets along any trajectory intersected by the most recent two fixations was done for two reasons. Firstly, we wanted to control for the effect that angular changes of successive saccades has on resulting fixation durations (Tatler \& Vincent, 2008). Furthermore, it is known that saccades in scene viewing are primarily executed along a horizontal axis (Nuthmann \& Henderson, 2010). Due to such a bias it was expected that manipulations would primarily be placed along the horizontal axis and this was confirmed with a post-hoc analysis.


Figure 2. In the scene condition targets are presented at $7^{\circ}$ and $14^{\circ}$. Unlike the Static condition targets can be placed on any axis within the image. The angle at which the boxes are presented is determined by the location of the current and previous fixations and was presented in the direction of the eye-movement. The delay between target presentations is either $50,100,150$ or 200 ms . As in the Static condition $20 \%$ of trials consisted of only a single step to $7^{\circ}$.

## Gaze contingent fixation detection

In order to present targets to participants within the scene condition it was necessary to accurately detect the presence of a fixation with as much temporal precision as possible. The native Eyelink gaze contingent algorithms were used in order to detect the onset of fixation. Once the conditions for presenting the first target had been met, and the Eyelink detected a fixation, the first target was presented to the participant. Delays in the online detection of fixations resulted in the targets being presented after the onset of fixation at a delay (ms) of
$\mu=45.2 ; \sigma=19.0$.

## Results

## Amplitude transition function in the Scene condition

The aim of the first analysis is to provide evidence in support of the hypothesis that saccade cancellation does indeed occur within a more natural scene viewing context than is typically studied with double-step stimuli. In order to construct an ATF, only trials in which both the first and second target steps occurred during a single critical fixation were analysed. Trials were also rejected when the response saccade was not made in a direction consistent with the target steps. Therefore, in the Scene condition 33 trials were removed due to movement of the eyes prior to presentation of the second target, and 6 were removed due to detection of a misdirected saccade. In the Static condition 22 and 2 trials were removed, respectively.

The amplitude transition function for the scene condition was constructed by fitting a four parameter logistic function with a form:

$$
\begin{equation*}
y=a+\frac{b-a}{1+e^{c(d-x)}} \tag{1}
\end{equation*}
$$

where $a$ represents a lower bound for the sigmoid, $b$ represents an upper bound, $c$ scales the response to x (Delay) about the midpoint and $d$ is the inflection point of the sigmoid.
Model fitting The data were fit with a nonlinear mixed effects model (NLME) (Pinheiro \& Bates, 2000). Firstly, we added fixed effects which allow us to directly estimate the effect of experimental condition (Static vs Scene) on the parameters of the nonlinear response function described in (1). Secondly, random effects were included in the model in order to reduce the effect of unreliable differences between participants due to unbalanced observations and individual variability in task performance.

In the analysis of the Scene condition we fit a model which included the effect of only a single condition on the parameters $a, b, c$ and $d$. Random effects of participant on the parameters $a, b, c$ and $d$ were also included in the model. For the comparison between the Static vs Scene conditions the model was extended to include a fixed effect of condition on the four model parameters. The R statistical programming language ( R Development Core Team, 2008) and the nlme package (Pinheiro, Bates, DebRoy, Sarkar, \& R Core Team, 2013) were used to conduct the analysis.

Effect of Scene on model parameters The parameters of the best fitting model are reported in Table 1. From the scatter the typical ATF evoked by double-step stimuli is observed with a characteristic sigmoidal shape. The horizontal dotted lines indicate the location of target 1 at $7^{\circ}$ and target 2 at $14^{\circ}$. Furthermore, the scatter confirms that within the Scene condition lower values of $D$ are typically associated with saccades directed towards the location of the first target step whereas at larger values of $D$ reprogramming occurs and saccades are directed towards the second target location.

Table 1. Mixed effects model parameters

| Effect | $M$ | $S E$ | $t$ | $p$ |
| :--- | :--- | :--- | :--- | :--- |
| $a$ | 6.47 | .160 | 40.48 | $(<0.01)$ |
| $b$ | 12.35 | .351 | 35.23 | $(<0.01)$ |
| $c$ | 0.12 | .027 | 4.33 | $(<0.01)$ |
| $d$ | 113.18 | 5.46 | 20.72 | $(<0.01)$ |

Summary of the estimated values of the fixedeffects parameters along with their means (M), standard errors (SE), $t$ and $p$ values, units of the parameters are reported in milliseconds. The parameters $a$ and $b$ are respectively the lower and upper asymptote of the sigmoid while $c$ scales the response about the midpoint and $d$ is the inflection point of the sigmoid.


Figure 3. Amplitude transition function constructed from responses in the Scene double-step condition. $D$ represents the amount of time elapsed between the onset of the second target stimulus and detection of the first response saccade. Horizontal lines represent the locations of target steps 1 and 2.

The results of the model fits estimated that the lower and upper bounds of the saccadic endpoints were $6.47^{\circ}$ ( $S E=$ $.160, t=40.48)$ and $12.35^{\circ}(S E=0.35, t=35.23)$ respectively. While there was undershoot for saccades targeting both the initial and final target steps, the undershoot to the final step was larger ( $0.53^{\circ}$ vs $1.65^{\circ}$ ).

## Cumulative distribution function of saccade latencies

Saccade programming latencies were compared on trials in which there was no target step (no-step) with trials in which there was a target step ( $50,100,150,200 \mathrm{~ms}$ ). This analysis aims to investigate whether trials in which a saccade was reprogrammed from the first target to the second target require longer latencies when compared to no-step trials in which no such reprogramming occurs.

In no-step trials saccade latency was calculated as the
elapsed time (ms) from the appearance of the $7^{\circ}$ target and the first observed response saccade. For trials of all other delays, latencies were analysed for saccades in which the first response saccade occurred after the second target step had appeared and in addition that the first response saccade compensated for the second target step. Compensation was defined such that the saccadic endpoint was within a distance $2^{\circ}$ of visual angle from the second target location. Latency was calculated as the elapsed duration between the onset of the first target step and the onset of first response saccade.

A cumulative distribution function (CDF) was fit to compare latencies between delays of different lengths. Latencies from 200 ms TSD trials were excluded as too few compensated saccades were observed. Latencies for compensated saccades are clearly longer than those of saccades elicited by no-step stimuli indicating that in order to incorporate the second target step into the response, increased latencies are required. Furthermore, we observe that as the TSD increases a corresponding increase in latency is also observed.


Figure 4. Cumulative distribution function of latencies at target step delays of $50,100,150 \mathrm{~ms}$ compared to the no-step latencies. The no-step latencies are constructed from latencies on trials in which there was only a step to the first target response.

## Comparison of Static vs Scene conditions

While the primary aim was to provide evidence that results from static double-step conditions generalize to a more dynamic scene based context, our dataset also offers an opportunity to directly compare performance differences between the Static and Scene conditions. A description of the model used to fit the Static vs Scene data can be found in the Model fitting section of the Scene only analysis.

Comparing the scatter in the Static vs Scene condition (Fig. 5) it is apparent that there is considerably more variability in the data that comprise the ATF in the Scene as compared to Static condition. Due to the more dynamic nature of the Scene task this is to be expected. For instance, in the Static condition participants stay fixated on a central cross while they wait for the target stimuli to appear. It is therefore likely that any anticipatory processes preparing future eyemovements are suppressed. In contrast, during the Scene con-

Table 2. Mixed effects model parameters

| Effect | $M$ | $S E$ | $t$ | $p$ |
| :--- | :--- | :--- | :--- | :--- |
| $a$ (Intercept) | 6.63 | 0.14 | 48.62 | $(<0.01)$ |
| $a$ (Scene) | -0.17 | 0.16 | -1.1 | $(=0.27)$ |
| $b$ (Intercept) | 13.0 | 0.11 | 119.9 | $(<0.01)$ |
| $b$ (Scene) | -0.74 | 0.20 | -3.82 | $(<0.01)$ |
| $c$ (Intercept) | 0.14 | 0.02 | 7.57 | $(<0.01)$ |
| $c$ (Scene) | -0.05 | 0.02 | -2.18 | $(=0.03)$ |
| $d$ (Intercept) | 76.14 | 2.30 | 32.91 | $(<0.01)$ |
| $d$ (Scene) | 33.86 | 2.51 | 13.44 | $(<0.01)$ |

Summary of the estimated values of the fixed-effects parameters along with their means (M), standard errors (SE) and $t$ and $p$ values, units of the parameters are reported in milliseconds. The parameters $a$ and $b$ are respectively the lower and upper asymptote of the sigmoid while $c$ scales the response about the midpoint and $d$ is the inflection point. The intercept indicates the estimated parameter in the Static condition, while (Scene) indicates the influence of condition Scene.
dition participants are actively engaged in search, the display and measurement of their double-step response is likely to incorporate processes involved in preparing an eye-movement prior to the presentation of the double-step stimuli.


Figure 5. Comparison of the amplitude transition functions in Static vs Scene conditions. D represents the amount of time elapsed between the onset of the second target stimulus and detection of the first response saccade. Horizontal lines represent the locations of target steps 1 and 2.

Effect of experimental condition (Scene vs Static) on model parameters Model parameters are summarized in Table 2. The lower and upper bounds of the fitted functions measure the saccadic endpoints of responses to either the first (lower bound) or second (upper bound) targets. No effect of condition was observed on the lower bound $(t=-1.1$, $p=0.27$ ) however, there was an effect of condition on the
amplitude of responses targeting the final location $(t=-3.82$, $p<0.01$ ), indicating greater undershoot in the Scene condition.

A final observation can be made regarding the markedly slower compensation response in the Scene condition. The ATF in the Scene condition appears to be shifted to the right, and this reflects that increasing values of $D$ are required to make a response of corresponding amplitude to that of the Static condition. Furthermore, we observe a significant effect of condition on $d$ (inflection point) $(t=13.44, p<0.01)$ supporting the observation that responses of comparable amplitude require longer values of $D$ in the Scene condition as compared to the Static condition.

Ludwig et al. (2007) have referred to saccadic dead time (SDT) as the last moment at which a new stimulus can modify a saccade program currently under preparation. They describe that the SDT may be extracted from the ATF by estimating the point at which the compensation function begins to incorporate the location of the second target step. The SDT was extracted from our ATF by deriving the point on the curve which represents a cumulative increase of $5 \%$ from the lower asymptote. We estimated this point in the Static condition as 55 ms and in the Scene condition as 77 ms .

## Discussion

The research question that this paper addresses is whether established results utilizing double-step stimuli to explore saccade programming can be extended to scene viewing. It has been argued that the ATF constructed from double-step responses provides evidence for a distinction between a labile and nonlabile stage of saccadic programming. A target stimulus is only able to modify the current goals of a saccade while it is in the labile stage of programming and can no longer have an influence once the program becomes nonlabile. The sigmoidal shape of the ATF (Fig. 3) reveals that when the target stimulus is presented shortly before the saccade (low values of $D$ ) that processing had reached the nonlabile threshold and therefore had no influence on the resulting saccade. When the second target is presented in earlier stages of saccade preparation (higher values of $D$ ) we see saccades that compensate for the second target location due to programming still being within the labile stage. These results have been previously established in double-step studies utilizing static conditions (Becker \& Jürgens, 1979; Ludwig et al., 2007), and our study provides evidence for an analogous process occurring within a condition more akin to naturalistic scene viewing.

Cancellation has also been suggested as a causal mechanism for the systematic increase in observed fixation durations. For instance, in the CRISP model, saccade cancellation accounts for the increase in fixation durations that is observed directly following a delay of stimulus onset (see Figure 7, Nuthmann et al., 2010 for details). The CDF (Fig. 4) illustrates that latencies are increased in trials in which a reprogramming of a saccade is likely to have occurred. However, caution must be taken when assigning a causal interpre-
tation to the role of saccade cancellation in observed latency increases. The TSD trials analysed are specifically those for which a saccade was not executed prior to the appearance of the second saccade target. Therefore, we expect to see a complementary increase in latency alongside increases in TSD. One possibility is that increased latencies are observed specifically because compensated saccades are those in which the programming to the first target progressed slowly enough to wait out and incorporate the appearance of the second target. While this analysis does provide confirmatory evidence that saccade cancellation is consistent with increased fixation durations, it does not necessarily shed light on the causal connection between cancellation and increased latencies.

The comparison between the Static and Scene condition also indicate the presence of several notable differences. Ludwig et al. (2007) analyse a concept termed saccadic dead time (SDT) which corresponds closely to the concept of a nonlabile stage of programming. The SDT corresponds to the last point in time at which a saccadic eye-movement may be modified. We estimated SDT as 55 ms in the Static condition and 77 ms in the Scene condition. Differences in the SDT across experimental conditions have been observed in prior work (Ludwig et al., 2007). An important implication of this result with regard to models of gaze control in naturalistic scene viewing is that it provides an empirical bound for the duration of the nonlabile stage. In CRISP for instance, a mean duration of 40 ms was assigned to the duration of the nonlabile stage. This value was determined from classic double-step results conducted under static conditions. This value is also roughly consistent with the duration of the nonlabile stage estimated in our own Static condition ( 55 ms ) but represents an underestimate when compared to the nonlabile duration in the Scene condition. As CRISP is a model of fixation durations in scene viewing it is likely that the estimated mean duration of the nonlabile stage in our Scene condition represents an improvement over the corresponding Static estimate.

It should be noted that the comparison reported here may still reflect important differences not solely attributable to the influence of scene processing. For instance, the Scene but not Static condition double-step targets were presented on any axis. Future work may consider including a task in which performance in our Scene condition is compared directly to a similar task but one in which the scene is replaced by a noise filtered image.

Further comment is warranted on the applicability of the data reported here to models such as CRISP that claim a causal interpretation for saccade cancellation in observed fixation duration delays. In CRISP, when a saccade program is within the labile stage of programming a cancellation signal may interrupt the current program. The time required to reinitiate a new saccade program results in a delay to the current fixation duration. In an alternative formulation aimed at explaining saccade latencies within a Static double-step context, Camalier et al. (2007) suggest that cancellation occurs
due to a race between a process initiated to execute a saccade (GO) and process initiated to cancel that saccade (STOP). While there is some similarity in the assumptions between these two models the race model does not insist on a nonlabile/labile dichotomy. Rather, the race model accounts for double-step performance with reference to the timing of the GO and STOP accumulation processes. In order to further explore the role of saccade cancellation it may be of interest to directly compare the predictions of the saccade programming mechanisms in the CRISP model with those of the race model described in Camalier et al. (2007).

## References

Becker, W., \& Jürgens, R. (1979). An analysis of the saccadic system by means of double step stimuli. Vision Research, 19(9), 967-983.
Camalier, C., Gotler, A., Murthy, A., Thompson, K., Logan, G., Palmeri, T., et al. (2007). Dynamics of saccade target selection: Race model analysis of double step and search step saccade production in human and macaque. Vision Research, 47(16), 2187-2211.
Engbert, R., Nuthmann, A., Richter, E., \& Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. Psychological Review, 112(4), 777-813.
Ludwig, C. J. H., Mildinhall, J. W., \& Gilchrist, I. D. (2007). A population coding account for systematic variation in saccadic dead time. Journal of Neurophysiology, 97(1), 795-805.
Nuthmann, A., \& Henderson, J. M. (2010). Object-based attentional selection in scene viewing. Journal of Vision, 10(8), 1-19.
Nuthmann, A., Smith, T., Engbert, R., \& Henderson, J. (2010). CRISP: A computational model of fixation durations in scene viewing. Psychological Review, 117(2), 382-405.
Pinheiro, J., \& Bates, D. (2000). Mixed-effects models in $S$ and $S$-PLUS. New York: Springer.
Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., \& R Core Team. (2013). nlme: Linear and nonlinear mixed effects models [Computer software manual]. (R package version 3.1-108)

R Development Core Team. (2008). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from http://www.R-project.org (ISBN 3-900051-07-0)
Reichle, E., Pollatsek, A., Fisher, D., \& Rayner, K. (1998). Toward a model of eye movement control in reading. Psychological Review, 105(1), 125-157.
Tatler, B. W., \& Vincent, B. T. (2008). Systematic tendencies in scene viewing. Journal of Eye Movement Research, 2(2), 1-18.
Westheimer, G. (1954). Eye movement responses to a horizontally moving visual stimulus. Archives of Ophthalmology, 52(6), 932-941.

# New Empirical Tests of a Quantum Model for Question Order Effects 

Zheng Wang (wang.1243@osu.edu)<br>School of Communication, Derby Hall, The Ohio State University, Columbus, OH 43210 USA<br>Tyler Solloway (solloway.3@osu.edu)<br>School of Communication, Derby Hall, The Ohio State University, Columbus, OH 43210 USA<br>Jerome Busemeyer (jbusemey@indiana.edu)<br>Psychological and Brain Sciences, Indiana University, Bloomington, IN 47405 USA


#### Abstract

Recent findings show that human inferences and decisions interfere in ways analogous to incompatible quantum observables, and conceptual judgments are inseparable in ways similar to entangled quantum states. This discovery has led a group of physicists and psychologists to form a new field called "quantum cognition," which uses mathematical principles of quantum theory to explain human cognitive behavior. The power of this new theoretical approach is illustrated here by testing an a priori and precise prediction derived from quantum theory regarding question order effects commonly observed in survey research. The test of quantum theory was statistically satisfied across a set of 26 national surveys on presidential job approval and country satisfaction in past 10 years. These results suggest that quantum theory, initially invented to explain order effects on measurements in physics, provides a powerful prediction for measurement order effects in social and behavioral sciences too.


The human brain is a powerful and massively complex neural system. It provides the biological substrate for an emergent mind capable of producing highly intelligent cognitive behaviors, such as inferences and decisions. How this happens remains a topic of intense investigation in cognitive neuroscience. The possibility that the brain's tremendous power arises from parallel computations of quantum physical neuronal interactions has been raised (Hameroff \& Penrose, 1996; Hagan, Hameroff \& Tuszynski, 2002) but strongly criticized (Tegmark, 2000; McKemmish, Reimers, McKenzie, Mark \& Hush, 2009). However, what if it is our behavior - rather than our brains - that follows quantum rules?

Supporting this idea, latest evidence shows that human inferences and decisions interfere in ways analogous to incompatible quantum observables (Pothos \& Busemeye, 2009; Busemeyer, Wang \& Lambert-Mogiliansky, 2008), and conceptual judgments are inseparable in ways similar to entangled quantum states (Aerts \& Sozzo, 2011). Formal principles that quantum theorists invented to deal with properties of complex physical systems provide a powerful mathematical description of human behavior (Busemeyer \& Bruza, 2012; Khrennikov, 2010). This discovery has led a
group of physicists and psychologists to work together and form the new field of "quantum cognition," which uses mathematical principles of quantum theory to explain human cognitive behavior. It has successfully accounted for various puzzling findings in psychological literature, ranging across perception (Atmanspacher, Filk \& Romer, 2004), associative memory (Bruza, Kitto, Nelson \& McEvoy, 2009), conceptual reasoning (Aerts, 2009), probability judgments (Busemeyer, Pothos, Franco \& Trueblood, 2011), decision making (Yukalov \& Sornette, 2011), and strategic game behavior (Lambert-Mogiliansky \& Busemeyer, 2012). It is plausible that the underlying neural systems follow classical dynamic laws, but the emergent cognitive behaviors are coarse "quantized" descriptions (Atmanspacher \& Graben, 2007). In fact, more than half a century ago, founding fathers of quantum theory speculated that fundamental quantum principles have implications outside of physics to human cognitive behavior (Pauli, 1950; Bohr, 1958).

Here we tested a new, a priori and precise prediction derived from quantum theory regarding question order effects commonly observed in survey research. This type of exact prediction is rare in social and behavioral sciences. The prediction was statistically supported across a set of 26 national surveys in past 10 years on two important public opinion questions in the U.S.: presidential job approval and country satisfaction. This surprisingly accurate test illustrates the theoretical power of our new approach to use quantum theory as a mathematical tool to explain and predict human cognitive behaviors. We show that quantum theory, initially invented to explain order effects of measurements in physics, provides a powerful prediction for order effects of measurements in psychology.

## Measurement Order Effects

One of the prime paradoxes of physics explained by quantum mechanics is that the order of measurements affects the observed statistics. For example, when testing the direction of spin $1 / 2$ particles, the results depend on whether the "up-down" direction is tested before versus after the
"left-right" direction (Sakurai, 1994). In the terminology of quantum theory, observables like these are defined as incompatible, and the theory was built on a noncommutative algebra of operators (Von Neumann, 1932).

Order effects of measurements are not unique to physics. It has long been recognized that the order of questions can influence human judgments and decisions (Schuman \& Presser, 1981; Sudman \& Bradburn, 1974). For example, the Pew Research Center conducted a telephone survey experiment during June $10-14,2009$ with a nationally representative sample of 1,502 U.S. adults. A random half of the sample was asked, "Do you approve or disapprove of the way Barack Obama is handling his job as President?" followed by "All in all, are you satisfied or dissatisfied with the way things are going in this country today?" The other half was asked the exact same questions but in the opposite order. It turns out that the presidential job approval rate was $63.38 \%$ when it was asked first and dropped to $58.58 \%$ when asked second.

Gauging public opinions is an enormously important task in any democracy. Among many challenges that survey researchers must manage, question order effects are one of the most important (Schuman \& Presser, 1981; Moore, 2002). A common practice is rotating question orders between randomly-split samples to balance out question order effects. Whether the order of two questions produces significant effects can be easily tested. Denote $p(\mathrm{AyBn})$ as the probability of agreeing ("yes") to question A and then disagreeing ("no") to question B , and $p(\mathrm{BnAy})$ as the probability of the same answers when the questions were asked in the opposite order. Similarly, probabilities of the remaining response combinations, $p(\mathrm{AnBy})$ and $p(\mathrm{ByAn})$, are defined. The two order conditions produce a pair of $2 \times 2$ contingency tables, which, according to the null hypothesis, should be equivalent except for sampling error (e.g., $p(\mathrm{AyBn})=p(\mathrm{BnAy}))$. Discrepancy from the null hypothesis is measured by $\chi^{2}$. If the null hypothesis is correct, the $\chi^{2}$ statistic should have a $\chi^{2}(3)$ distribution.

Table 1 shows $\chi^{2}$ results for two Gallup survey experiments reported in a seminal article on question order effects (Moore, 2002). Each sampled around 1,000 U.S. adults using the split sample paradigm. In the first poll, people were asked whether Bill Clinton was honest and trustworthy, and whether Al Gore was honest and trustworthy. In the second poll, people were asked whether white people dislike black people, and whether black people dislike white people. Each $2 \times 2$ contingency table in Table 1 summarizes the observed proportions for the four response combinations in one question order. As shown by the $\chi^{2}$ test on the order effects, both experiments produced large order effects with strikingly different patterns. Now we come backto the presidential job approval and country satisfaction questions. Is there a robust order effect for this pair of important public opinion questions? To examine this, we obtained from the Pew Research Center all its survey experiments that included this pair of questions in past 10

Table 1: Observed proportions for each order condition,
and $\chi^{2}$ tests for testing order effects and the QQ equality. See Appendix on the $\chi^{2}$ tests.

| Observed proportions in the two different question orders |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Clinton-Gore |  |  | White-Black |  |  |
|  | Gy | Gn |  | By | Bn |
| Cy | . 4899 | . 0447 | Wy | . 3987 | . 0174 |
| Cn | . 1767 | . 2886 | Wn | . 1612 | . 4227 |
| Gore-Clinton |  |  | Black-White |  |  |
|  | Cy | Cn |  | Wy | Wn |
| Gy | . 5625 | . 1991 | By | . 4012 | . 0597 |
| Gn | . 0255 | . 2130 | Bn | . 1379 | . 4012 |
| Discrepancy tests |  |  |  |  |  |
| Order effects |  | $\chi^{2}(3)=10.14, p<.05$ | $\chi^{2}(3)=73.04, p<.001$ |  |  |
| QQ equality |  | $\begin{aligned} & q=-.003 \\ & \chi^{2}(1)=.01, p=.91 \\ & \hline \end{aligned}$ | $\begin{aligned} & \hline q=-.02 \\ & \chi^{2}(1)=.56, p=.46 \end{aligned}$ |  |  |

years. There are 26 surveys in total, with a nationally representative sample between 815 and 3,006 U.S. adults ( $M$ $=1,644, S D=422.24$ ). Of each sample in each survey, a random half was asked the presidential job approval question first while the other half was asked the country satisfaction question first. The $\chi^{2}$ test indicates significant question order effects across the 26 surveys (see Figure 1).


Figure 1: $\chi^{2}$ frequency distributions for testing the order effects and the QQ equality. The navy bars show the observed frequencies of $\chi^{2}$ values for order effects distributed across 10 categories separated at 9 deciles (.1, .2, $.3, .4, .5, .6, .7, .8, .9)$; the green bars show those for the QQ equality test; the dotted line shows the expected frequency by the null hypothesis. The observed frequency distribution of order effects significantly differs from the expected frequency $\left(\chi^{2}(9)=37.675, p<.0001\right)$, but that of the QQ equality is not $\left(\chi^{2}(9)=9.5485, p=.3935\right)$. So, as predicted by the quantum model, there is a significant measurement order effects but the QQ equality holds across the 26 nationals surveys. See Appendix on the $\chi^{2}$ tests.

## A Quantum Model for Question Order Effects

It would be a speculative leap, however, to think that quantum theory can be applied to human behavior simply because the behavior displays measurement orders effects. Indeed, quantum models of cognition need to be rigorously tested. A precise and empirically testable prediction has been derived from a quantum model for the question order
experimental paradigm (Busemeyer \& Bruza, 2012; Wang \& Busemeyer, in press). The model is simple, intuitive, but general. First, as illustrated in Figure 2, a person's prior belief state is represented by a unit length vector (denoted by S ) within an $N$-dimensional vector space. This use of feature vectors to represent belief or knowledge is consistent with many other cognitive models of memory. Second, each answer to a question is represented by a subspace within the vector space. Each subspace corresponds to a projector (see Figure 2). Denote $P_{X}$ as the projector corresponding to agreement to a question, and $I-P_{X}$ is the projector corresponding to disagreement to the question, where $I$ is the identity operator. Third, how to compute response probabilities in quantum models? For example, following quantum probability rules, the probability of agreeing to question $A$ and then disagreeing to question $B$ equals the squared length of the result obtained by sequentially projecting the prior belief state on the subspace for agreeing to $A$ and then on the subspace for disagreeing to $B$, that is, $p(\mathrm{AyBn})=\left\|\left(I-P_{B}\right) P_{A} \mathrm{~S}\right\|^{2}$. If the subspaces for the two questions are incompatible (i.e., not spanned by a common basis), then their projectors are non-commutative (i.e., $P_{B} P_{A}$ $\neq P_{A} P_{B}$ ), and question order effects are predicted to occur.


Figure 2: An illustration of basic quantum principles used in the question order model. The figure illustrates a simple 3-dimensional vector space, but the space can be arbitrarily high-dimensional. The probability of agreeing to question $X$ is the squared length of the projection $\mathrm{Px} * \mathrm{~S}$ obtained by projecting the belief state $S$ to the $\mathrm{X}-\mathrm{Y}$ plane representing the subspace for agreeing to question X . If question X was asked after another question, the belief state would have already been changed by answering the preceding question, and the probability of agreeing to question X (conditioned on the preceding answer) becomes the squared length of the result obtained by projecting the adjusted belief state on the subspace for question X .

This model makes an a priori and precise prediction, named the Quantum Question (QQ) equality (see Appendix
for proof $):[p(\mathrm{AyBn})+p(\mathrm{AnBy})]-[p(\mathrm{ByAn})+p(\mathrm{BnAy})]=0$. Intuitively, this means, the probability of having different responses to the two questions (e.g., saying "yes" to one and "no" to the other) should remain the same across the two question orders. As shown in the proof, this equality must hold for any belief state and any pair of projectors in any high-dimensional vector space. This precise prediction can be easily tested empirically: if it holds, the difference in observed proportions on the left hand of the QQ equality, defined as $q$, should not statistically differ from zero as tested by $\chi^{2}$ for difference in proportions.

The QQ equality prediction was tested using the aforementioned two Gallup data sets and was supported with high accuracy (see Table 1). To generalize the results, it was further tested using the 26 Pew national survey experiments. If it holds, the observed frequency distribution of $\chi^{2}$ (shown as the green bars in Figure 1) should be distributed according to a $\chi^{2}$ (1) distribution. Indeed as predicted, the observed distribution is not significantly different from the expected distribution (see Figure 1). In summary, although the 26 Pew studies exhibit significant questions order effects, there are not significant deviations from the predicted QQ equality.

## Can a Classical Brain Give Rise to Quantum Cognitive Behaviors?

The surprisingly accurate predictions generated by the quantum model for question order effects is one example of an accumulating body of evidence supporting the general applicability of quantum theory for explaining a wide range of human cognitive behavior findings that are paradoxical from a classical probability perspective (Busemeyer \& Bruza, 2012). This, however, leaves a question: can a classical brain give rise to behavior that follows quantum principles? Recently, mathematical physicists have provided a mathematical answer to this puzzle. Essentially, coarse measurements of a classical dynamic system typically generate incompatible observables that result in unresolvable uncertainty relations and entangled correlations (beim Graben \& Atmanspacher, 2006; beim Graben, Filk \& Atmanspacher, in press). According to quantum theory, order effects occurs for incompatible observables.

A key idea is to distinguish "ontic" states (e.g., states of a dynamic neural network) in a classical phase space from "epistemic" states (e.g., discrete choices or judgments across time) obtained from an observable. The mapping from ontic to epistemic states usually is many to one, where the epistemic states generated by an observable form a partition of the ontic phase space. Knowing the epistemic state does not completely determine the ontic state, but a sequence of measurements across time refines the partition of the phase space. In the limit, the partition reaches a "finest dynamic refinement," denoted by $\wp$. Now suppose two observables ( $\boldsymbol{f , g}$ ) produce different finest dynamic refinements ( $\wp_{f} \neq \wp_{g}$ ) and neither converge to the identity
partition, as illustrated in Figure 3. This means that no ontic state is accessible by a sequence of measurements from either observable. Then there exists an epistemic state (a set of ontic states) $F_{a} \in \wp_{f}$ that determines the value $a$ produced by the observable $f$, but the value of the observable $g$ must remain dispersive. Likewise, there exists an epistemic state


Figure 3. An illustration showing how uncertainty relations are generated by coarse descriptions of classical dynamic systems. The underlying classical phase space $X$ is inconsistently partitioned by two different observables, $f$ and $g$. The cell within $X$ that always assigns a value $a$ to the observable $f$ assigns a range of different possible values ( $w, u, z$ ) to the observable $g$. In this case, there exists an epistemic state that determines the value $a$ produced by the observable $f$, but the value of the observable $g$ must remain dispersive.
(a set of ontic states) $G_{v} \in \wp_{g}$ that determines the value $v$ produced by the observable $\boldsymbol{g}$, but the value of the observable $\boldsymbol{f}$ remains dispersive. It is impossible to simultaneously determine the value $a$ from observable $\boldsymbol{f}$ and the value $v$ from the observable $g$ with arbitrary precision, so that the two observables are incompatible. Consequently, the partitions generated by the two incompatible observables produce incompatible Boolean algebras of events, and the entire collection forms a partial rather than a complete Boolean algebra. Quantum theory is specifically suitable to assign probabilities to events defined on a partial Boolean algebra.

## Discussion

Scientists are still far from understanding how mental states emerge from the neural substrates. It is too early to conclude whether or not quantum physics plays a significant role in neural processing. Nevertheless, even if the brain is classical, the ubiquitous nature of incompatible observables provides a good reason to consider using quantum theory as a mathematical tool for predicting human behavior (Busemeyer \& Bruza, 2012; Khrennikov, 2010). As our quantum question order model encapsulates and illustrates, at least four motivations drives the development of this new field of quantum cognition. (a) Judgments and decisions are
not simply read out from memory, but rather, they are constructed from the cognitive state for the question at hand; and (b) drawing a conclusion from one judgment or decision changes the context and disturbs the cognitive system, which then (c) affects the next judgment or decision, producing order effects, so that (d) human judgments and decisions do not obey the commutative rule of Boolean logic. If we replace "judgments or decisions" with "physical measurements" and replace "cognitive system" with "physical system," then these are exactly the same reasons that forced physicists to develop quantum theory in the first place. Traditionally, quantum theory has rarely been applied outside of physics, but now a growing number of researchers are successfully using it to explain human cognitive behavior.

## Acknowledgments

We thank Pew Research Center for assistance with data acquisition. P. T. beim Graben and H. Atmanspacher for comments on part of the manuscript. This work was supported by NSF (SES 1153846 and 1153726) and AFOSR (FA9550-12-1-0397).

## Reference

Aerts, D. (2009). Quantum structure in cognition. Journal of Mathematical Psychology, 53, 314-348.
Aerts, D. \& Sozzo, S. (2011). Quantum structure in cognition: Why and how concepts are entangled. In P.D. Bruza (Ed), Quantum interaction: Lecture notes in computer science (116-127). Springer.
Atmanspacher, H. \& Graben, P. B. (2007). Contextual emergence of mental states from neurodynamics. Chaos and Complexity Letters, 2, 151-168.
Atmanspacher, H., Filk, T. \& Romer, H. (2004). Quantum zero features of bistable perception. Biological Cybernetics, 90, 33-40.
beim Graben, P. \& Atmanspacher, H. (2006). Complementarity in classical dynamical systems. Foundations of Physics, 36, 291-306.
beim Graben, P., Filk, T., \& Atmanspacher, H. (in press). Epistemic entanglement due to non-generating partitions of classical dynamical systems. International Journal of Theoretical Physics.
Bohr, N. (1958). Atomic physics and human knowledge. New York: Wiley.
Bruza, P. D., Kitto, K., Nelson, D. \& McEvoy, C. (2009). Is there something quantum like in the human mental lexicon? Journal of Mathematical Psychology, 53, 362377.

Busemeyer, J. R. \& Bruza, P. D. (2012). Quantum models of cognition and decision. Cambridge University Press.
Busemeyer, J. R., Wang, Z. \& Lambert-Mogiliansky, A. (2009). Empirical comparison of Markov and quantum models of decision making. Journal of Mathematical Psychology, 53, 423-433.
Busemeyer, J. R., Pothos, E. M., Franco, R. \& Trueblood, J. S. (2011). A quantum theoretical explanation for
probability judgment errors. Psychological Review, 118, 193-218.
Hagan, S., Hameroff, S. R. \& Tuszynski, J. (2002). Quantum computation in brain microtubles: Decoherence and biological feasibility. Physical Review E, 65, 061901.
Hameroff, S. R. \& Penrose, R. (1996). Conscious events as orchestrated spacetime selections. Journal of Consciousness Studies, 3, 36-53.
Khrennikov, A. Y. (2010). Ubiquitous quantum structure: From Psychology to finance. New York: Springer.
Lambert-Mogiliansky, A. \& Busemeyer, J. R. (2012). Quantum type indeterminacy in dynamic decisionmaking: Self-control through identity management. Games, 3, 97-118.
McKemmish, L. K., Reimers, J. R., McKenzie, R. H., Mark, A. E. \& Hush, N. S. (2009). Penrose-Hameroff orchestrated objective-reduction proposal for human consciousness is not biologically feasible. Physical Review E, 80, 021912.
Moore, D. W. (2002). Measuring new types of question order effects. Public Opinion Quarterly, 66, 80-91.
Pauli, W. (1950). Die philosophische Bedeutung der Idee der Komplentarit. Experientia, 6, 72-75.
Pothos, E.M. \& Busemeyer, J. R. (2009). A quantum probability model explanation for violations of "rational" decision making. Proceedings of the Royal Society B, 276(1665), 2171-2178.
Sakurai, J. J. (1994). Modern quantum mechanics. New York: Pearson Education.
Schuman, H. \& Presser, S. (1981). Questions and answers in attitude surveys. New York: Academic Press.
Tegmark, M. (2000). Importance of quantum decoherence in brain processes. Physical Review E, 61, 4194-4206.
Von Neumann, J. (1932). Mathematical foundations of quantum theory. Princeton University Press.
Wang, Z. \& Busemeyer, J. R. (in press). Explaining and predicting question order effects using a quantum model. Topics in Cognitive Science.
Yukalov, V. \& Sornette, D. (2011). Decision theory with prospect interference and entanglement. Theory and Decision, 70, 283-328.

## Appendix

## 1. Proof of the QQ equality.

Here we briefly introduce the basic axioms of quantum theory and then derive the QQ equality. We use the Dirac bracket notation so that $\langle S \mid T\rangle$ represents the inner product between two vectors. According to quantum theory, events represented as subspaces of a Hilbert space. Corresponding to each event A there is a orthogonal projector $\boldsymbol{P}_{\mathrm{A}}$. The state of a quantum system is represented by a unit length vector $S$ within the Hilbert space. The probability of event A equals the squared length of the projection $p(\mathrm{~A})=\left\|\boldsymbol{P}_{\mathrm{A}} S\right\|^{2}$. If event A is observed, then the state is updated according to Lüder's rule $S_{\mathrm{A}}=\boldsymbol{P}_{\mathrm{A}} S /\left\|\boldsymbol{P}_{\mathrm{A}} S\right\|$.

Define $S$ as the initial state. Denote the projector for saying yes to question C as $\boldsymbol{P}_{\boldsymbol{C}}$ and denote $\boldsymbol{P}_{G}$ as the
projector for saying yes to question G. We start by expanding the probability for answering "yes" to question C:

$$
\begin{aligned}
&\left\|\boldsymbol{P}_{\boldsymbol{C}} \cdot S\right\|^{2}=\left\|\boldsymbol{P}_{\boldsymbol{C}} \cdot \boldsymbol{I} \cdot S\right\|^{2}=\left\|\boldsymbol{P}_{\boldsymbol{C}} \cdot\left(\boldsymbol{P}_{\boldsymbol{G}}+\left(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{G}}\right)\right) \cdot S\right\|^{2}=\| \boldsymbol{P}_{\boldsymbol{C}} \cdot \boldsymbol{P}_{\boldsymbol{G}} \cdot S \\
&+\boldsymbol{P}_{\boldsymbol{C}} \cdot\left(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{G}} \cdot S \|^{2}\right. \\
&=\left\|\boldsymbol{P}_{\boldsymbol{C}} \cdot \boldsymbol{P}_{\boldsymbol{G}} \cdot S\right\|^{2}+\left\|\boldsymbol{P}_{\boldsymbol{C}} \cdot\left(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{G}}\right) \cdot S\right\|^{2}+\langle S| \cdot \boldsymbol{P}_{\boldsymbol{G}} \cdot \boldsymbol{P}_{\boldsymbol{C}} \cdot \boldsymbol{P}_{\boldsymbol{C}} \\
& \cdot\left(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{G}}\right)|S\rangle+\langle S|\left(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{G}}\right) \cdot \boldsymbol{P}_{\boldsymbol{C}} \cdot \boldsymbol{P}_{\boldsymbol{C}} \cdot \boldsymbol{P}_{\boldsymbol{G}}|S\rangle \\
&=\left\|\boldsymbol{P}_{\boldsymbol{C}} \cdot \boldsymbol{P}_{\boldsymbol{G}} \cdot S\right\|^{2}+\| \boldsymbol{P}_{\boldsymbol{C}} \cdot\left(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{G}} \cdot S \|^{2}+\langle S| \boldsymbol{P}_{\boldsymbol{G}} \cdot \boldsymbol{P}_{\boldsymbol{C}} \cdot \boldsymbol{P}_{\boldsymbol{C}}\right. \\
& \cdot\left(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{G}}\right)|S\rangle+\langle S| \cdot \boldsymbol{P}_{\boldsymbol{G}} \cdot \boldsymbol{P}_{\boldsymbol{C}} \cdot \boldsymbol{P}_{\boldsymbol{C}} \cdot\left(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{G}}\right)|S\rangle * \\
&=\left\|\boldsymbol{P}_{\boldsymbol{C}} \cdot \boldsymbol{P}_{\boldsymbol{G}} \cdot S\right\|^{2}+\left\|\boldsymbol{P}_{\boldsymbol{C}} \cdot\left(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{G}}\right) \cdot S\right\|^{2}+2 \cdot \operatorname{Re}\left[\langle\mathrm{~S}| \boldsymbol{P}_{\boldsymbol{G}} \cdot \boldsymbol{P}_{\boldsymbol{C}} \cdot \boldsymbol{P}_{\boldsymbol{C}}\right. \\
&\left.\quad\left(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{G}}\right)|S\rangle\right] \\
&\left.\left\|\boldsymbol{P}_{\boldsymbol{C}} \cdot \boldsymbol{P}_{\boldsymbol{G}} \cdot S\right\|^{2}+\| \boldsymbol{P}_{\boldsymbol{C}} \cdot\left(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{G}}\right)|S\rangle\right],
\end{aligned}
$$

and the latter follows from the idempotent property of projectors. (The symbol $x^{*}$ used in the above derivation refers to the complex conjugate of $x$.) Define the total probability to say yes to question C when G was asked first as

$$
\begin{aligned}
T P_{C} & =\left\|\boldsymbol{P}_{\boldsymbol{G}} \cdot S\right\|^{2} \cdot\left\|\boldsymbol{P}_{\boldsymbol{C}} \cdot S_{G}\right\|^{2}+\left\|\left.\boldsymbol{P}_{\sim \boldsymbol{G}} \cdot S\right|^{2} \cdot\right\| \boldsymbol{P}_{\boldsymbol{C}} \cdot S_{\sim G} \|^{2} \\
& =\left\|\boldsymbol{P}_{\boldsymbol{C}} \cdot \boldsymbol{P}_{\boldsymbol{G}} \cdot S\right\|^{2}+\left\|\boldsymbol{P}_{\boldsymbol{C}} \cdot\left(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{G}}\right) \cdot S\right\|^{2} .
\end{aligned}
$$

An order effect for question C when G was asked first expressed as
$C_{C}=T P_{C}-\left\|\boldsymbol{P}_{\boldsymbol{C}} \cdot S\right\|^{2}=-2 \cdot \operatorname{Re}\left[\langle S| \boldsymbol{P}_{\boldsymbol{G}} \cdot \boldsymbol{P}_{\boldsymbol{C}} \cdot\left(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{G}}\right)|S\rangle\right]$.
Immediately we see that if $\boldsymbol{P}_{\boldsymbol{G}}$ and $\boldsymbol{P}_{\boldsymbol{C}}$ commute so that $\boldsymbol{P}_{\boldsymbol{G}} \cdot \boldsymbol{P}_{\boldsymbol{C}}=\boldsymbol{P}_{\boldsymbol{C}} \cdot \boldsymbol{P}_{\boldsymbol{G}}$ then $\boldsymbol{P}_{\boldsymbol{G}} \cdot \boldsymbol{P}_{\boldsymbol{C}} \cdot\left(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{G}}\right)=\boldsymbol{P}_{\boldsymbol{C}} \cdot \boldsymbol{P}_{\boldsymbol{G}} \cdot\left(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{G}}\right)=0$ and we predict NO order effect. Thus non-commuting projectors are a necessary condition for order effects. Now let us re-examine

$$
\begin{aligned}
C_{C} & =T \boldsymbol{P}_{C}-\left\|\boldsymbol{P}_{\boldsymbol{C}} \cdot S\right\|^{2}=-2 \cdot \operatorname{Re}\left[\langle S| \boldsymbol{P}_{\boldsymbol{G}} \cdot \boldsymbol{P}_{\boldsymbol{C}} \cdot\left(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{G}}\right)|S\rangle\right] . \\
& =-2 \cdot \operatorname{Re}\left[\langle S| \boldsymbol{P}_{\boldsymbol{G}} \cdot \boldsymbol{P}_{\boldsymbol{C}}|S\rangle-\langle S| \boldsymbol{P}_{\boldsymbol{G}} \cdot \boldsymbol{P}_{\boldsymbol{C}} \cdot \boldsymbol{P}_{\boldsymbol{G}}|S\rangle\right] \\
& =-2 \cdot \operatorname{Re}\left[\langle S| \boldsymbol{P}_{\boldsymbol{G}} \cdot \boldsymbol{P}_{\boldsymbol{C}}|S\rangle-\left\|\boldsymbol{P}_{\boldsymbol{C}} \cdot \boldsymbol{P}_{\boldsymbol{G}} \cdot S\right\|^{2}\right] \\
& =-2 \cdot \operatorname{Re}\left[\langle S| \boldsymbol{P}_{\boldsymbol{G}} \cdot \boldsymbol{P}_{\boldsymbol{C}}|S\rangle\right]+2 \cdot\left\|\boldsymbol{P}_{\boldsymbol{C}} \cdot \boldsymbol{P}_{\boldsymbol{G}} \cdot S\right\|^{2} \\
& =2 \cdot\left\|\boldsymbol{P}_{\boldsymbol{G}} \cdot S\right\|^{2} \cdot\left\|\boldsymbol{P}_{\boldsymbol{C}} \cdot S_{G}\right\|^{2}-2 \cdot \operatorname{Re}\left[\langle S| \boldsymbol{P}_{\boldsymbol{G}} \cdot \boldsymbol{P}_{\boldsymbol{C}}|S\rangle\right] .
\end{aligned}
$$

In general, the inner product is a complex number which always can be expressed as $\langle S| \boldsymbol{P}_{\boldsymbol{G}} \quad \cdot \boldsymbol{P}_{\boldsymbol{C}}|S\rangle \quad=$ $\left.\left|\langle S| \boldsymbol{P}_{G} \cdot \boldsymbol{P}_{C}\right| S\right\rangle \mid \cdot[\cos (\phi)+i \cdot \sin \phi]$. The real part equals $\operatorname{Re}\left[\langle S| \boldsymbol{P}_{G}\right.$ $\left.\left.\cdot \boldsymbol{P}_{C}|S\rangle\right]=\left|\langle S| \boldsymbol{P}_{G} \cdot \boldsymbol{P}_{C}\right| S\right\rangle \mid \cdot \cos (\phi)$. By defining the ratio

$$
\left.R=\left|\langle S| \boldsymbol{P}_{\boldsymbol{G}} \cdot \boldsymbol{P}_{\boldsymbol{C}}\right| S\right\rangle \mid /\left(\left\|\boldsymbol{P}_{\boldsymbol{C}} \cdot S\right\| \cdot\left\|\boldsymbol{P}_{\boldsymbol{G}} \cdot S\right\|\right)
$$

then according to the Cauchy-Schwarz inequality, $0 \leq R \leq 1$. Finally we can express

$$
\begin{aligned}
C_{C}= & T P_{C}-\left\|\boldsymbol{P}_{\boldsymbol{C}} \cdot S\right\|^{2}=2 \cdot\left\|\boldsymbol{P}_{\boldsymbol{C}} \cdot \boldsymbol{P}_{\boldsymbol{G}} \cdot S\right\|^{2}-2 \cdot R \cdot \cos (\phi) \cdot \| \boldsymbol{P}_{\boldsymbol{C}} \\
& \cdot S\|\cdot\| \boldsymbol{P}_{\boldsymbol{G}} \cdot S \| \\
= & 2 \cdot\left\|\boldsymbol{P}_{\boldsymbol{G}} \cdot S\right\|^{2} \cdot\left\|\boldsymbol{P}_{\boldsymbol{C}} \cdot S_{G}\right\|^{2}-2 \cdot \theta \cdot\left\|\boldsymbol{P}_{\boldsymbol{C}} \cdot S\right\| \cdot\left\|\boldsymbol{P}_{\boldsymbol{G}} \cdot S\right\|,
\end{aligned}
$$

with $\theta=R \cdot \cos (\phi)$ and $-1 \leq \theta \leq+1$, which is the similarity index referred to in the main text. Similarly, the order effect for question G when C was asked first equals

$$
\begin{aligned}
C_{G}= & T P_{G}-\left\|\boldsymbol{P}_{\boldsymbol{G}} \cdot S\right\|^{2}=2 \cdot\left\|\boldsymbol{P}_{\boldsymbol{G}} \cdot \boldsymbol{P}_{\boldsymbol{C}} \cdot S\right\|^{2}-2 \cdot \operatorname{Re}\left[\langle S| \boldsymbol{P}_{\boldsymbol{C}}\right. \\
& \left.\cdot \boldsymbol{P}_{\boldsymbol{G}}|S\rangle\right], \text { but } \operatorname{Re}\left[\langle S| \boldsymbol{P}_{\boldsymbol{C}} \cdot \boldsymbol{P}_{\boldsymbol{G}}|S\rangle\right]=\operatorname{Re}\left[\langle S| \boldsymbol{P}_{\boldsymbol{G}} \cdot \boldsymbol{P}_{\boldsymbol{C}}|S\rangle\right]
\end{aligned}
$$

so that

$$
C_{G}=2 \cdot\left\|\boldsymbol{P}_{C} \cdot S\right\|^{2} \cdot\left\|\boldsymbol{P}_{G} \cdot S_{C}\right\|^{2}-2 \cdot \theta \cdot\left\|\boldsymbol{P}_{C} \cdot S\right\| \cdot\left\|\boldsymbol{P}_{G} \cdot S\right\|
$$

These two order effects share the same term, $2 \cdot \theta \cdot\left\|\boldsymbol{P}_{C} \cdot S\right\| \cdot \| \boldsymbol{P}_{G}$ $\cdot S \|$, and therefore together they imply the relation

$$
\begin{aligned}
0= & \left(2 \cdot\left\|\boldsymbol{P}_{\boldsymbol{C}} \boldsymbol{P}_{\boldsymbol{G}} \cdot S\right\|^{2}-C_{C}\right)-\left(2 \cdot\left\|\boldsymbol{P}_{\boldsymbol{G}} \boldsymbol{P}_{\boldsymbol{C}} \cdot S\right\|^{2}-C_{G}\right) \\
= & \left(2 \cdot\left\|\boldsymbol{P}_{\boldsymbol{C}} \boldsymbol{P}_{\boldsymbol{G}} \cdot S\right\|^{2}-\left\|\boldsymbol{P}_{\boldsymbol{C}} \boldsymbol{P}_{\boldsymbol{G}} \cdot S\right\|^{2}-\left\|\boldsymbol{P}_{\boldsymbol{C}} \boldsymbol{P}_{\sim \boldsymbol{G}} \cdot S\right\|^{2}+\left\|\boldsymbol{P}_{\boldsymbol{C}} \cdot S\right\|^{2}\right)- \\
& \left(2 \cdot\left\|\boldsymbol{P}_{\boldsymbol{G}} \boldsymbol{P}_{\boldsymbol{C}} \cdot S\right\|^{2}-\left\|\boldsymbol{P}_{\boldsymbol{G}} \boldsymbol{P}_{\boldsymbol{C}} \cdot S\right\|^{2}-\left\|\boldsymbol{P}_{\boldsymbol{G}} \boldsymbol{P}_{\sim \boldsymbol{C}} \cdot S\right\|^{2}+\left\|\boldsymbol{P}_{\boldsymbol{G}} \cdot S\right\|^{2}\right)
\end{aligned}
$$

$$
\begin{aligned}
& =\left(\left\|\boldsymbol{P}_{C} \boldsymbol{P}_{\boldsymbol{G}} \cdot S\right\|^{2}-\left\|\boldsymbol{P}_{C} \boldsymbol{P}_{\sim G} \cdot S\right\|^{2}+\left\|\boldsymbol{P}_{C} \cdot S\right\|^{2}\right)-\left(\left\|\boldsymbol{P}_{G} \boldsymbol{P}_{C} \cdot S\right\|^{2}-\right. \\
& \left.\left\|\boldsymbol{P}_{G} \boldsymbol{P}_{\sim} \cdot S\right\|^{2}+\left\|\boldsymbol{P}_{\boldsymbol{G}} \cdot S\right\|^{2}\right) \\
& =\left[\left\|\boldsymbol{P}_{C} \boldsymbol{P}_{G} \cdot S\right\|^{2}-\left\|\boldsymbol{P}_{C} \boldsymbol{P}_{\sim G} \cdot S\right\|^{2}+\left(\left\|\boldsymbol{P}_{G} \cdot S_{C}\right\|^{2}+\left\|\boldsymbol{P}_{\sim} \cdot S_{C}\right\|^{2}\right) \cdot \| \boldsymbol{P}_{C}\right. \\
& \left.\cdot S \|^{2}\right]-\left[\left\|\boldsymbol{P}_{G} \boldsymbol{P}_{C} \cdot S\right\|^{2}-\left\|\boldsymbol{P}_{G} \boldsymbol{P}_{\sim} C \cdot S\right\|^{2}+\left(\left\|\boldsymbol{P}_{C} \cdot S_{G}\right\|^{2}+\right.\right. \\
& \left.\left.\left\|\boldsymbol{P}_{\sim} \cdot S_{G}\right\|^{2}\right) \cdot\left\|\boldsymbol{P}_{G} \cdot S\right\|^{2}\right] \\
& =\left(\left\|\boldsymbol{P}_{C} \boldsymbol{P}_{G} \cdot S\right\|^{2}-\left\|\boldsymbol{P}_{C} \boldsymbol{P}_{\sim G} \cdot S\right\|^{2}+\left\|\boldsymbol{P}_{G} \cdot S_{C}\right\|^{2}\left\|\boldsymbol{P}_{C} \cdot S\right\|^{2}+\right. \\
& \left.\left\|\boldsymbol{P}_{\sim} \cdot S_{C}\right\|^{2}\left\|\boldsymbol{P}_{C} \cdot S\right\|^{2}\right)-\left(\left\|\boldsymbol{P}_{G} \boldsymbol{P}_{C} \cdot S\right\|^{2}-\left\|\boldsymbol{P}_{G} \boldsymbol{P}_{\sim} \cdot S\right\|^{2}+\right. \\
& \left.\left\|\boldsymbol{P}_{C} \cdot S_{G}\right\|^{2}\left\|\boldsymbol{P}_{G} \cdot S\right\|^{2}+\left\|\boldsymbol{P}_{\sim} \cdot S_{G}\right\|^{2}\left\|\boldsymbol{P}_{G} \cdot S\right\|^{2}\right) \\
& =\left(\left\|\boldsymbol{P}_{C} \boldsymbol{P}_{G} \cdot S\right\|^{2}-\left\|\boldsymbol{P}_{C} \boldsymbol{P}_{\sim G} \cdot S\right\|^{2}+\left\|\boldsymbol{P}_{G} \boldsymbol{P}_{C} \cdot S\right\|^{2}+\left\|\boldsymbol{P}_{\sim G} \boldsymbol{P}_{C} \cdot S\right\|^{2}\right) \\
& -\left(\left\|\boldsymbol{P}_{G} \boldsymbol{P}_{C} \cdot S\right\|^{2}-\left\|\boldsymbol{P}_{\boldsymbol{G}} \boldsymbol{P}_{\sim} \cdot S\right\|^{2}+\left\|\boldsymbol{P}_{C} \boldsymbol{P}_{\boldsymbol{G}} \cdot S\right\|^{2}+\| \boldsymbol{P}_{\sim C} \boldsymbol{P}_{\boldsymbol{G}}\right. \\
& \left.\cdot S \|^{2}\right) \\
& \begin{array}{l}
=\left(\left\|\boldsymbol{P}_{\sim G} \boldsymbol{P}_{C} \cdot S\right\|^{2} \quad-\left\|\boldsymbol{P}_{C} \boldsymbol{P}_{\sim G} \cdot S\right\|^{2}\right)-\left(\left\|\boldsymbol{P}_{\sim C} \boldsymbol{P}_{G} \cdot S\right\|^{2}-\right. \\
={ }_{\|}^{\left\|\boldsymbol{P}_{G} \boldsymbol{P}_{c} \cdot S\right\|^{2} \|^{2}}\left(\left\|\boldsymbol{P}_{\sim G} \boldsymbol{P}_{C} \cdot S\right\|^{2}+\left\|\boldsymbol{P}_{G} \boldsymbol{P}_{\sim C} \cdot S\right\|^{2}\right)-\left(\left\|\boldsymbol{P}_{\sim C} \boldsymbol{P}_{G} \cdot S\right\|^{2}+\right. \\
\left.\left\|\boldsymbol{P}_{C} \boldsymbol{P}_{\sim G} \cdot S\right\|^{2}\right)=0 . \quad \text { Q.E.D. }
\end{array}
\end{aligned}
$$

The last line is the QQ equality expressed as quantum probabilities.

## 2. $\chi^{2}$ tests used in Table 1.

First we present the $\chi^{2}$ test for order effects. Define $n_{Y N}$ as the frequency of saying "yes" to question C when C was asked first and saying "no" to question G when G was asked second, and the other combinations of answers are defined similarly. Define $n=n_{Y Y}+n_{Y N}+n_{N Y}+n_{N N}$. Define $m_{Y N}$ as the frequency of "yes" to question G when G was asked first and "no" to question C when C was asked second, and the other combinations of answers are defined similarly. Define $m=m_{Y Y}+m_{Y N}+m_{N Y}+m_{N N}$. The log likelihood for the unconstrained model that allows order effects is defined by

```
\(\mathrm{G}_{\mathrm{U}}=\left[n_{Y Y} \cdot \ln \left(n_{Y Y} / n\right)+n_{Y N} \cdot \ln \left(n_{Y N} / n\right)+n_{N Y} \cdot \ln \left(n_{N Y} / n\right)+\right.\)
    \(n_{N N} \cdot \ln \left(n_{N N} / n\right)+m_{Y Y} \cdot \ln \left(m_{Y Y} / m\right)+m_{Y N} \cdot \ln \left(m_{Y N} / m\right)+\)
    \(\left.m_{N Y} \cdot \ln \left(m_{N Y} / m\right)+m_{N N} \cdot \ln \left(m_{N N} / m\right)\right]\).

The \(\log\) likelihood for the constrained model that assumes no order effects is defined by
\[
\begin{align*}
\mathrm{G}_{\mathrm{C}}= & {\left[\left(n_{Y Y}+m_{Y Y}\right) \cdot \ln \left(\left(n_{Y Y}+m_{Y Y}\right) /(n+m)\right)+\left(n_{Y N}+\right.\right.} \\
& \left.m_{N Y}\right) \cdot \ln \left(\left(n_{Y N}+m_{N Y}\right) /(n+m)\right)+\left(n_{N Y}+m_{Y N}\right) \cdot \ln \left(\left(n_{N Y}\right.\right. \\
& \left.\left.+m_{Y N}\right) /(n+m)\right)+\left(n_{N N}+m_{N N}\right) \cdot \ln \left(\left(n_{N N}+\right.\right. \\
& \left.\left.\left.m_{N N}\right) /(n+m)\right)\right] . \tag{1b}
\end{align*}
\]

The \(\chi^{2}\) statistic is defined by the difference \(\chi^{2}=-2 \cdot\left(\mathrm{G}_{\mathrm{C}}-\right.\) \(\mathrm{G}_{\mathrm{u}}\) ). The unconstrained model involves (4-1) \(+(4-1)=6\) free parameters and the constrained model involves 4-1 \(=3\) free parameters, and so the \(\chi^{2}\) statistic has \(d f=3\).

Next we define the \(\chi^{2}\) test for the QQ equality. The log likelihood for the unconstrained model is defined as
\[
\begin{align*}
\mathrm{G}_{\mathrm{U}}= & {\left[\left(n_{Y N}+n_{N Y}\right) \cdot \ln \left(\left(n_{Y N}+n_{N Y}\right) / n\right)+\left(n_{Y Y}+n_{N N}\right) \cdot \ln \left(\left(n_{Y Y}+\right.\right.\right.} \\
= & \left.\left.n_{N N}\right) / n\right) \\
= & \left(m_{Y N}+m_{N Y}\right) \cdot \ln \left(\left(m_{Y N}+m_{N Y}\right) / m\right)+\left(m_{Y Y}+m_{N N}\right) \cdot \ln \left(\left(m_{Y Y}\right.\right. \\
& \left.\left.\left.+m_{N N}\right) / m\right)\right] . \tag{2a}
\end{align*}
\]

The log likelihood for the model constrained by the QQ equality equals
\[
\begin{align*}
\mathrm{G}_{\mathrm{C}} & =\left[( n _ { Y N } + n _ { N Y } + m _ { Y N } + m _ { N Y } ) \cdot \operatorname { l n } \left(\left(n_{Y N}+n_{N Y}+m_{Y N}+\right.\right.\right. \\
& \left.m_{N Y} /(n+m)\right)+\left(n_{Y Y}+n_{N N}+m_{Y Y}+m_{N N}\right) \cdot \ln \left(\left(n_{Y Y}+n_{N N}\right.\right. \\
& \left.\left.\left.+m_{Y M}+m_{N N}\right) /(n+m)\right)\right] . \tag{2b}
\end{align*}
\]

The \(\chi^{2}\) statistic is defined by the difference \(\chi^{2}=-2 \cdot\left(\mathrm{G}_{\mathrm{C}}-\right.\) \(\mathrm{G}_{\mathrm{U}}\) ). The unconstrained model involves \((2-1)+(2-1)=2\)
free parameters and the constrained model involves \(2-1=1\) free parameter, and so the \(\chi^{2}\) statistic has \(d f=1\).

\section*{3. \(\chi^{2}\) tests used in Figure 1.}

First we describe the \(\chi^{2}\) test for order effects. The \(\chi^{2}\) statistic for testing an order effect for each of the 26 data sets was computed using Equations 1a and 1 b defined above, producing 26 observed \(\chi^{2}\) values. If the null hypothesis is correct, these should be distributed according to a \(\chi^{2}\) distribution with \(d f=6\). Ten categories were constructed by computing the 9 category bounds: . 5844 equals the \(10^{\text {th }}\) percentile, 1.0052 equals the \(20^{\text {th }}\) percentile, 1.4237 equals the \(30^{\text {th }}\) percentile, 1.8692 equals the \(40^{\text {th }}\) percentile, 2.3660 equals the \(50^{\text {th }}\) percentile, 3.9462 equals the \(60^{\text {th }}\) percentile, 3.6649 equals the \(70^{\text {th }}\) percentile, 4.6416 equals the \(80^{\text {th }}\) percentile, and 6.2514 equals the \(90^{\text {th }}\) percentile. (For example, \(\operatorname{Pr}\left[\chi^{2}(6)<6.2514 \mid \mathrm{H}_{0}\right]=.90\).) These category bounds divide the expected frequency distribution (under the null hypotheses) into two 10 equally likely categories, with 2.6 expected frequency within each of the 10 categories using these cutoffs. Then frequency of the 26 observed \(\chi^{2}\) values were counted for each category. Denote \(f_{i}\) as the observed frequency for category \(i=1,10\). The log likelihood for the unconstrained model equals
\[
\begin{equation*}
\mathrm{G}_{\mathrm{U}}=\sum_{i} f_{i} \cdot \ln \left(f_{i} / 26\right) . \tag{3a}
\end{equation*}
\]

The log likelihood for the expected frequencies according to the null hypothesis equals
\[
\begin{equation*}
\mathrm{G}_{\mathrm{C}}=\sum_{i} f_{i} \ln (2.6 / 26) . \tag{3b}
\end{equation*}
\]

The \(\chi^{2}\) statistic is defined by the difference \(\chi^{2}=-2 \cdot\left(\mathrm{G}_{\mathrm{C}}-\right.\) \(\left.\mathrm{G}_{\mathrm{U}}\right)\). The unconstrained model involves \(10-1=9\) free parameters and the constrained model has no free parameters, and so the \(\chi^{2}\) statistic has \(d f=9\).

Next we describe the \(\chi^{2}\) test for the QQ equality. The \(\chi^{2}\) statistic for testing the QQ equality for each of the 26 data sets was computed using Equations 2a and 2b defined above, producing 26 observed \(\chi^{2}\) values. If the null hypothesis is correct, these should be distributed according to a \(\chi^{2}\) distribution with \(d f=1\). Ten categories were constructed by computing the 9 category bounds: . 0158 equals the \(10^{\text {th }}\) percentile, .0642 equals the \(20^{\text {th }}\) percentile, .1485 equals the \(30^{\text {th }}\) percentile, .2750 equals the \(40^{\text {th }}\) percentile, .4549 equals the \(50^{\text {th }}\) percentile, .7083 equals the \(60^{\text {th }}\) percentile, 1.0742 equals the \(70^{\text {th }}\) percentile, 1.6424 equals the \(80^{\text {th }}\) percentile, and 2.7055 equals the \(90^{\text {th }}\) percentile. (For example, \(\operatorname{Pr}\left[\chi^{2}(1)<2.7055 \mid \mathrm{H}_{0}\right]=.90\).) These category bounds divide the expected frequency distribution (under the null hypotheses) into two 10 equally likely categories, with 2.6 expected frequency within each of the 10 categories using these cutoffs. Then frequency of the 26 observed \(\chi^{2}\) values were counted for each category. Denote \(f_{i}\) as the observed frequency for category \(i=1,10\). Then Equations 3 a and 3 b were used to compute the \(\log\) likelihoods of the unconstrained and constrained models. Once again, the \(\chi^{2}\) statistic is defined by the difference \(\chi^{2}=\) \(-2 \cdot\left(\mathrm{G}_{\mathrm{C}}-\mathrm{G}_{\mathrm{U}}\right)\). The unconstrained model involves \(10-1=9\) free parameters and the constrained model has no free parameters, and so the \(\chi^{2}\) statistic has \(d f=9\).

\title{
Identifying Predictive Collocations
}

\author{
Silas Weinbach (silasw@coli.uni-saarland.de) \\ Department of Computational Linguistics \\ Campus C7.2, 66123 Saarbrücken, Germany
}

\author{
Vera Demberg (vera@coli.uni-saarland.de) \\ Cluster of Excellence, Saarland University, \\ Campus C7.4, 66123 Saarbrücken, Germany
}

\begin{abstract}
Idioms and common multi-word expressions are often argued to be stored as chunks of words or fixed configurations in the mind, and to therefore be accessed faster and interpreted more easily than fully compositional word combinations. Experimental research has furthermore shown that a specific "recognition point" can be identified in such expressions, at which enough information is present to access the meaning of the whole expression and predict the remaining words of the collocation.
In this paper, we suggest measures for automatically identifying those multi-word expressions where the first part is particularly predictive of the rest, and evaluate our measures against human association data collected in a cloze test.
\end{abstract}

Keywords: Predictivity; Multi-Word Expressions; Collocations; Entropy

\section*{Introduction}
"When her boyfriend proposed to her, she was in seventh heaven." "After jogging, he quenched his thirst with some nice orange juice." The above sentences contain collocations where the first part of the collocation (e.g., "in seventh", "quench") is very predictive of the second part ("heaven" and "thirst", respectively). Such predictive collocations can be idiomatic (as in the first example), or literal, fully compositional configurations. Previous studies observing human processing of idioms have argued that there exists a "recognition point", at which comprehenders have identified the idiom and can predict the rest. Some also argue that not only idioms, but also frequent collocations, may be stored in the lexicon.

However, by far, not all collocations are predictive, consider for example light verb constructions where a very unpredictive verb is combined with a sense-carrying noun. Being able to pick out predictive collocations among the set of all collocations, and automatically identifying the recognition point in idioms could be very useful for psycholinguistic models of language processing: Processes of predicting specific upcoming words, and accessing idiomatic meaning could then potentially be captured in a broad-coverage model.

This paper takes a first step in this direction by proposing a number of alternative statistical methods for identifying predictive collocations and evaluating them with respect to a cloze task where people were asked to complete verbs with the argument they associated most strongly. This evaluation captures the predictive strength of a verb in the absence of further predictive context, and is supposed to compare which of the measures works best at identifying good candidates for predictive collocations.

\section*{Background and Related Work}

Collocations are commonly used phrasal expressions which have become characteristic for a language or jargon (Smadja,
1993). They are idiosyncratic because there is no rule which can tell us why a some specific lexemes (e.g., "strong tea" instead of "powerful tea") are combined to express a particular concept (McKeown \& Radev, 2000).

\section*{Representation of Collocations in Humans}

Idioms are a special type of collocations whose semantic meaning is not compositional of the meaning of the words it contains, but are more idiosyncratic such as "give a whirl" (meaning to try) or "spill the beans". The status of these expressions in the lexicon is still under debate. It has often been argued (Swinney \& Cutler, 1979) that these idiomatic expressions should be part of the lexicon. Some have even argued that non-idiomatic collocations may likewise be stored as chunks in longterm memory (Ellis, 2001; Ellis, Frey, \& Jalkanen, 2009).

An alternative model was proposed by Cacciari and Tabossi (1988) and holds that both decomposable and idiomatic expressions are represented in the lexicon "as configurations" and that these configurations can get activated during processing as soon as enough information has been perceived to render the collocation recognizable. Tabossi, Fanari, and Wolf (2009) present evidence that both idiomatic and literal collocations may be stored in memory as such configurations.

On the other hand, Vespignani, Canal, Molinaro, Fonda, and Cacciari (2010) find in an ERP experiment which compares the processing of idiomatic expressions with literal phrases that language comprehenders have categorial templates for idioms in their lexicon, and that these can be activated at a specific recognition point after which a prediction process is initiated. Their results suggest that this prediction process can be distinguished from non-idiomatic predictive mechanisms. If such effects are to be modelled in a computational model, it is necessary for the model to have access to a set of idiomatic expressions and their recognition points.

The goal of this paper is not to answer the question concerning which types of collocations may be stored in memory and which ones may be processed compositionally. Instead, we evaluate statistical measures for automatically identifying predictive collocations. The methods and measures are generally applicable and may later be used in combination with a filter for identifying idiomatic expressions.

An important point for our study however is the relevance of a recognition point and the notion of predictability of a multi-word expression. Tabossi, Fanari, and Wolf (2005) showed that only the meanings of predictable idioms, but not of all idioms, become available early on in idiom processing. Such a prediction process may be beneficial to language understanding because, as Tabossi et al. (2005) finds, recog-
nizing an initial fragment of a predictable idiom inhibits the recognition of the literal meaning of the rest of the expression and hence facilitates comprehension by reducing ambiguity.

\section*{Automatically Identifying Collocations}

The basic idea in automatically identifying collocations (for a good overview, see (Manning \& Schütze, 1999)) is to count how often a set of words occur together within a specific distance of one another (e.g., always adjacent) or within a syntactic relationship (e.g., verb-argument). Many word-pairs or multi-word expressions with frequent co-occurrence however aren't interesting collocations (like "in the") because the reason for their high co-occurrence frequency is the high frequency of each of the words and the syntactic constraints with which they occur.

Two strategies are commonly used to ignore such cases: the first one is to use statistical tests that indicate whether two words were observed together more often than would be expected otherwise. Collocation candidates are then ranked with respect to significance scores. Note though that only the ranking, but not the exact significance level, is usually considered interesting, as most co-occurrences are significant simply due to the fact that language has some regular patterns due to syntactic rules (Manning \& Schütze, 1999). Another common approach is to calculate the pointwise mutual information (PMI) between two words.

The second strategy is to specify what types of collocations should be found by specifying POS tag patterns or dependency relations between words (e.g., only considering adjective-noun pairs or only considering modifiers of nouns).

Finally, automatic methods developed for detecting idiomatic collocations often also use semantics to identify these expressions: in non-compositional expressions, the meaning of the words in the idiom are less likely to be semantically related to the rest of the context (Katz \& Giesbrecht, 2006).

The following paragraphs are going to explain the most commonly used measures for detecting collocations, as well as the word patterns used in this work.

Association Measures for Identifying Collocations To assess whether a pair of words \(w_{1} w_{2}\) is a collocation, we can count how often these words can be observed together \(O_{1,1}\), and calculate how often we would expect to see them together given their unigram frequencies and the size \(N\) of our data set: \(E_{1,1}=\frac{f r e q\left(w_{1}\right)}{N} \times \frac{f r e q\left(w_{2}\right)}{N} \times N\). If we observe them together much more often than would be expected given their unigram frequencies, we conclude that they are strongly associated and represent a collocation.

The most commonly used association measures (AMs) are the following: In a t-test (see for example (Manning \& Schütze, 1999)), the higher the \(t\)-value, the more likely that the observed co-occurrence of the words \(w_{1}\) and \(w_{2}\) would not have happened by chance.
\[
\mathrm{t}-\text { Test }: \quad t=\frac{\frac{O_{1,1}-E_{1,1}}{N}}{\sqrt{O_{1,1}}}
\]

An alternative is the z-score (variant suggested by Evert (2008)). The formula below estimates the mean of the distribution as \(E_{1,1}\) and its standard deviation as \(\sqrt{E(1,1)}\)
\[
\mathrm{z}-\text { score : } \quad z=\frac{O_{1,1}-E_{1,1}}{\sqrt{E_{1,1}}}
\]

Pearson's \(\chi^{2}\) test (for a more detailed description, see (Manning \& Schütze, 1999)) is very similar to the \(z\)-score, except it uses the square of the z -values and takes into account not only the probability of the words occurring together ( \(O_{1,1}\) ), but compares also the estimated and observed frequencies of a \(w_{1}\) not occurring with \(w_{2}, w_{2}\) not occurring with \(w_{1}\) and the co-occurrence of words different from both \(w_{1}\) and \(w_{2}\).
\[
\chi^{2}=\sum_{i, j} \frac{\left(O_{i, j}-E_{i, j}\right)^{2}}{E_{i, j}}
\]

Finally, the \(\log\) likelihood ratio \(\lambda\), similarly to \(\chi^{2}\), uses weighted on the similarity of the words \(w_{1}\) and \(w_{2}\) occurring together or with different words.
\[
\lambda=2 \sum_{i, j} O_{i, j} \log \frac{O_{i, j}}{E_{i, j}}
\]

Pointwise Mutual Information (PMI; Church and Hanks (1989)) is and information-theoretic concept and measures how much information is shared between words \(w_{1}\) and \(w_{2}\) - it is a symmetric measure. If there are two words with only occur in the context of each other, then one of the words conveys all the information that the two of them convey and their mutual information is maximal.
\[
P M I=\log _{2} \frac{O_{1,1}}{E_{1,1}}
\]

Filters Previous work on collocation extraction has shown (Seretan \& Wehrli, 2009; Fazly, Cook, \& Stevenson, 2009; Lin, 1998) that result quality depends also on choosing good patterns in which to observe collocation candidates. These have been defined via windows of observation, via fixed POS tag sequences or via syntactic dependencies, as for example from a dependency parser. The present study focusses on verb-argument pairs as extracted from a large text resource using a dependency parser.

Asymmetric Association Measures While there is a large body of literature on the topic of automatic recognition of multi-word expressions and idioms, there is almost no work on asymmetric association measures. An exception is Michelbacher, Evert, and Schütze (2007, 2011), who use conditional probability (see below), as well as a number of rank measure which are based on the traditional association measures explained above. As we found out after first submitting this paper, a related proposal for developing directional association measures has been made by Gries (to appear). A comparison between our measures and the associative-learning based approach should be addressed in future work.

\section*{Proposed Predictive Measures}

One way of capturing how predictive one word is of another word is to calculate the conditional probability ( CP ; also suggested by Michelbacher et al., 2007; 2011) of the second word given the first word. High CP indicates that the first word is highly predictive of the second word.
\[
C P\left(w_{1}, w_{2}\right)=P\left(w_{2} \mid w_{1}\right)
\]

A straightforward approach to predictive collocations is to use conditional probability as an association measure, or to combine existing measures for association between two words with the conditional probability of the second word given the first word. Different ways of combining the measures are possible, such as for example weighted additive combination \((a \times C P+b \times A M)\), or multiplicative combination \((C P \times A M)\).

In our preliminary experiments, it turned out that the additive models (which essentially represent a form of averaging between the measures) do not perform well. While they boost the score of collocation candidates which are both strongly associated and predictive, they do usually not boost it enough to achieve rankings higher than those of candidate collocations which are extremely good on just one of the measures, such that the resulting highest ranked candidates still contain a lot of highly associated but non-predictive word pairs.

Multiplicative combination, on the other hand, can be thought of as a filter that ranks down any collocation candidates which are highly associated but not predictive, and boost highly predictive candidates, resulting in a cleaner list of predictive collocations. Based on this observation, we propose the following new measures: \(\mathrm{CP}, \mathrm{CP} \times \chi^{2}, \mathrm{CP} \times \mathrm{PMI}\) and \(\mathrm{CP} \times \lambda\), which we will evaluate in the remainder of this paper.

\section*{Comparison of Association Measures}

It is instructive to inspect how similar the alternative association measures are to one another. To this end, we sorted 3.6 million adjective-noun pairs from the ukwac corpus (Ferraresi, Zanchetta, Baroni, \& Bernardini, 2008) according to each of our association measures and calculated the correlations between these sorted lists. Table 1 shows that four of our measures, \(\chi^{2}, Z, \lambda\) and PMI actually result in very similar rankings, with correlations \(\rho>.9\). Only rankings by \(t\)-value look a bit more dissimilar, and relatively more similar than other measures to the overall frequency of word pairs (indicated as FRQ in Table 1). It is also important to observe that conditional probability (CP) leads to a very different ranking and is only correlated at \(0.28<\rho<0.4\) with the other measures.

\section*{Identification of Predictive Collocations}

We dependency-parsed the Gigaword Corpus \({ }^{1}\) using the Stanford parser (Marneffe, MacCartney, \& Manning, 2006). From

\footnotetext{
\({ }^{1}\) http://www.ldc.upenn.edu
}

Table 1: Correlation (Spearman's rho) between different association measures for top 500 ranks.
\begin{tabular}{r||l|lllll|l}
\(\rho\) & FRQ & T & Z & \(\chi^{2}\) & \(\lambda\) & PMI & CP \\
\hline \hline FREQ & 1 & .62 & .28 & .29 & .46 & .06 & .2 \\
\hline T & .62 & 1 & .86 & .83 & .88 & .72 & .28 \\
Z & .28 & .86 & 1 & .97 & .91 & .96 & .38 \\
\(\chi^{2}\) & .29 & .83 & .97 & 1 & .97 & .93 & .4 \\
\(\lambda\) & .46 & .88 & .91 & .97 & 1 & .82 & .4 \\
PMI & .06 & .72 & .96 & .93 & .82 & 1 & .33 \\
\hline CP & .2 & .28 & .38 & .4 & .4 & .33 & 1
\end{tabular}
the Gigaword's 1.7 billion tokens, we extracted all dependency triples of the type "VB*:dobj:NN*" (i.e., verbs and their direct arguments), for which the verb occurred to the left of the argument in the text. Verb-argument pairs which occurred less than 16 times in the corpus were excluded from the analysis, as some of the association measures are not applicable when counts are too low. Furthermore, we removed all verb-argument pairs containing words which were not in WordNet under the correct POS tag. This later step filtered out POS-tagging errors like "unsalted butter" or "quantum mechanic" where "unsalted" and "quantum" were tagged as verbs, or "smile slyly" where "slyly" was tagged as a noun, as well as foreign language material.

\section*{Cloze Task}

The goal of our experiment is to evaluate whether combining one of the established measures for collocation extraction with conditional probabilities will lead to a good measure for identifying predictive collocations, and which of the proposed measures works best. For the evaluation, we use a simple task which is independent of any sentential context: we ask human participants to complete a list of verbs with a noun they associate first, and then compare which of our measures predicts best the cloze probabilities of each verb. A reason for evaluating with a completion experiment instead of simply comparing to a verb's entropy or conditional probability on the corpus itself is that many of the highly ranked collocations in our measures are in fact not necessarily generally valid predictive colloctions - some are very domain-specific, such as rise percent (from "rise 20 percent") and tell reporter.

\section*{Experimental Materials}

For evaluating predictive collocations, we were looking for a set of verbs which contains a good portion of potentially predictive verbs. We therefore selected verbs for our completion experiment by first calculating ranked lists of some of our target measures that we want to compare: \(\mathrm{CP} \times \chi^{2}, \chi^{2}\) and \(\mathrm{CP} \times \lambda\), and randomly chose 50 verbs out of the 200 bestranked verb-object pairs of each measure. This procedure left us with a set of 118 verbs for our completion experiment. The rationale behind choosing verbs this way instead of just selecting a random set of verbs is that we wanted to avoid ending up with only a very small number of predictive verbs.

Table 2: Arguments filled in for the verb "heal" during our completion experiment. We also collected completion times for each response.
\begin{tabular}{lr|lr} 
Answer.w2 & Seconds & Answer.w2 & Seconds \\
\hline the sick & 7 & wounds & 55 \\
a wound & 19 & bodies & 8 \\
the sick & 5 & a wound & 8 \\
the wound & 37 & yourself & 15 \\
a wound & 13 & a sore & 9 \\
a wound & 18 & the wound & 10 \\
sores & 4 & wounds & 4 \\
a wound & 7 & the wound & 7
\end{tabular}

\section*{Procedure}

We ran our experiment via Amazon Mechanical Turk (Paolacci, Chandler, \& Ipeirotis, 2010). In order to explain the task to our subjects, we gave them three examples of completed verb-argument pairs, using verbs which were not part of the 118 verbs that we wanted to collect completion data for: "to quench thirst", "to rob a bank" and "to feed the dog". We restricted our subjects to people living in the U.S. and instructed them to only take part in the experiment if they qualified as native speakers of English. Furthermore, we also restricted our pool of workers to ones that had in the past gotten \(>95 \%\) of their HITs \({ }^{2}\) approved and had successfully completed at least 1000 HITs. We collected a total of 1888 verbargument associations (i.e., 16 associations for each verb). Each worker was allowed to complete as many verbs as they wanted (but, of course, each verb only once). The 1888 associations were completed by 40 separate workers.

\section*{Collected Data}

For each verb, we collected 16 argument-associations. For example, see completions for the verb "heal" in Table 2. We lemmatized all answers, and dealt with typos (e.g., havok instead of havoc), orthographic variants (e.g., judgment vs. judgement) using minimum edit distance.

To assess the predictive strength of a verb, we calculated the entropy of each verb given the types of responses (after clustering them by lemma and dealing with typos etc, as described above). For example, the entropy of "heal" would be 1.53 . As we collected at most 16 associations per verb, entropy ranges between 0 and 4 for our data set. We can then use the entropy to classify our verbs into highly selective verbs (such as "grit", "honk", "flex", "sing", "twiddle"), less selective verbs (e.g., "pay", "fire", "attend") and nonselective ones (e.g., "quote", "shout", "request"). In a linear mixed effects regression analysis with random intercept and random slope for verb entropy under subject, we found that verb entropy is a significant positive predictor of completion times ( \(p<0.01\) ), i.e., when an argument of a verb is less predictable, people take longer to fill in the slot.

\footnotetext{
\({ }^{2}\) HIT stands for "Human Intelligence Task" and is used as the official term for tasks in Amazon Mechanical Turk.
}

\section*{Evaluation}

We evaluate our measures of predictive collocations in two ways. A good measure should rank highly those collocations where the first part is highly predictive of the second part.

Identifying Predictive Collocations We select a group of highly predictive verbs (determined by their entropy in the experiment) and generate verb-noun pairs by selecting the most common completion for those verbs in the experiment. This results in a list of verb-noun collocations where the verb is highly predictive of the noun. Next, we calculate the average rank of these verb-noun pairs for each of our measures, see table in Figure 1.

An important note to keep in mind when interpreting the average ranks in the table in Figure 1 is that the set of verbs was originally randomly chosen from among the top-ranked 200 verb noun pairs of the measures \(\mathrm{CP} \times \mathrm{CHI}, \mathrm{CHI}\) and \(\mathrm{CP} \times \lambda\); note also that Z is almost identical to CHI in the ranking it generates - these measures are therefore marked in bold in the table.

The newly proposed measure \(\mathrm{CP} \times \mathrm{CHI}\) clearly outperforms the other measues. It has the lowest average rank, meaning that the verb-noun pairs which we have identified as being particularly predictive are ranked highest in this measure. Note that 44 verb-noun pairs satisfied the criterion of the verb entropy in the experiment being under the threshold of 1.5 . This gives us an average rank of 22.5 as the best possible ranking which could possibly be achieved. Of course, not all possible verbs were tested in our experiment, hence direct comparison to this value is not meaningful. More important is the comparison to the average ranks of other measures. Clearly, the combined measure \(\mathrm{CP} \times \mathrm{CHI}\) is much better than either of its parts, and also clearly outperforms \(\mathrm{CP} \times \lambda\).

It is also fair to compare the measures which were not part of constructing the evaluation verb set (not in bold) to one another. Clearly, combining CP with the association measures improves identification of predictive collocations, in particular there is an interesting boost in the performance of the \(t\)-test measure when combined with conditional probabilities. We also conclude that \(\lambda\) is not a useful measure for identifying predictive collocations.

Additional insight comes from plotting average ranks for all verb-noun pairs with identical cloze probability, see Fig. 1. For a measure which is good at identifying predictive collocations, we expect there to be a linear relationship between cloze probabilities and \(\log\) rank (log rank makes sense because there are by definition more different noun pairs when cloze probability is lower). Monotonicity in the trend of the log rank indicates that the measure correctly distinguishes between different levels of cloze predictability. Furthermore, average log rank for verb noun pairs with cloze probability 1 should be close to 0 . The plots show that \(\mathrm{CP} \times \mathrm{CHI}\) comes closest to the described ideal correlation. The r squared measure given in the title of each plot in Figure 1 quantifies the fit between the plotted data points are from the regression line.
\begin{tabular}{lr}
\multicolumn{2}{l}{ verb entropy threshold 1.5} \\
\hline measure & rank \\
\hline \(\mathbf{C P x C H I}\) & 87.275 \\
CHI & 152.864 \\
\(\mathbf{Z}\) & 152.948 \\
CPxPMI & 237.692 \\
CPxT & 258.124 \\
CP & 291.342 \\
CPx \(\lambda\) & 404.471 \\
CPxFRQ & 709.715 \\
PMI & 1037.151 \\
\(\lambda\) & 2403.245 \\
Z & 6863.703 \\
T & 9378.727 \\
\hline ceiling & 22.5
\end{tabular}




Figure 1: Table at left: Average rank in lists ranked according to association measures; set of predictive verb-noun pairs defined based on different thresholds for verb entropy in experiment.
Plots: Average rank for \(\mathrm{CPxCHI}, \mathrm{CPx} \lambda\) and CHI grouped by cloze probabilities as obtained from MTurk experiment.

While measure \(\mathrm{CP} \times \lambda\) also follows a clear linear relationship, it does not locate the items with high predictability in its lowest ranks, indicating that it might be a good measure for quantifying collocations in general but not for predictiveness given the first part.

Correlation with human associations Our second evaluation compares the association values from all measures to the cloze probabilities obtained in the experiment. We again evaluate on average association values for each set of verb-noun pairs with a given cloze probability, see Figures 2 and 3. A good measure should increase monotonically with increasing cloze probabilities.

Among previously existing measures, PMI values can explain the largest amount of the variance in terms of average PMI values compared to cloze probabilities from our experiments. It is clear from Figure 2 that the common frequency of the two words, as well as the log likelihood measure are very poor predictors of predictive collocations.

Among traditional measures combined with conditional probabilities, \(\mathrm{CP} \times \mathrm{FRQ}, \mathrm{CP} \times \lambda\) and \(\mathrm{CP} \times \mathrm{T}\) perform very poorly. The problem for these measures is that they reflect strongly the overall frequency of a word pair. On the other hand, average \(\log \mathrm{CP} \times \mathrm{CHI}\) values have the strongest linear relationship with cloze probabilities, with few atypical points, as also reflected in the high \(R^{2}\). This result is thus consistent with the rank analysis in the first evaluation.

\section*{Conclusions and Outlook}

Our experiments indicate that the combination of conditional probability and the \(\chi^{2}\) measure might work best for identifying collocations where the first word is highly predictive of the second one. While this paper went a first step in devoting some attention to the problem of identifying predic-
tive collocations, suggesting possible measures and evaluating these measures on a cloze task of verb-argument associations, the next important step is to evaluate these methods on a more specific task such as automatically identifying recognition points of idioms. Furthermore, this paper has only dealt with one type of collocation (verb-argument pairs) and has focussed on collocations consisting of only two words.

In future work, we furthermore plan to evaluate the usefulness of predictive collocations by including them in a model of language processing in the form of lexical configurations.

\section*{Acknowledgments}

We would like to thank Stefan Thater for providing the Stanford-parsed version of Gigaword.

\section*{References}

Cacciari, C., \& Tabossi, P. (1988). The comprehension of idioms. Journal of memory and language, 27(6), 668-683.
Church, K. W., \& Hanks, P. (1989). Word association norms, mutual information and lexicography. In \(A C L\) (Vol. 27, p. 76-83).

Ellis, N. (2001). Memory for language. Cognition and second language instruction, 33-68.
Ellis, N., Frey, E., \& Jalkanen, I. (2009). The psycholinguistic reality of collocation and semantic prosody (1). Studies in Corpus Linguistics (SCL), 89.
Evert, S. (2008). Corpora and collocations. Corpus Linguistics. An International Handbook, 2.
Fazly, A., Cook, P., \& Stevenson, S. (2009). Unsupervised type and token identification of idiomatic expressions. Computational Linguistics, 35(1), 61-103.
Ferraresi, A., Zanchetta, E., Baroni, M., \& Bernardini, S. (2008). Introducing and evaluating ukwac, a very large web-derived corpus of english. In Proceedings of the WAC4 Workshop at LREC08.


Figure 2: Average association values for each measure grouped by cloze probabilities from MTurk experiment.

Gries, S. T. (to appear). 50-something years of work on collocations: what is or should be next. International Journal of Corpus Linguistics, 18(1).
Katz, G., \& Giesbrecht, E. (2006). Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In Proceedings of the workshop on multiword expressions (pp. 12-19).
Lin, D. (1998). Extracting collocations from text corpora.
Manning, C. D., \& Schütze, H. (1999). Foundations of statistical natural language processing. In (second ed., chap. 5). The MIT Press.
Marneffe, M.-C. de, MacCartney, B., \& Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. LREC.
McKeown, K. R., \& Radev, D. R. (2000). Collocations. In A handbook of natural language processing (pp. 507-523).
Michelbacher, L., Evert, S., \& Schütze, H. (2007). Asymmetric association measures. Proceedings of the Recent Advances in Natural Language Processing (RANLP 2007).
Michelbacher, L., Evert, S., \& Schütze, H. (2011). Asymmetry in corpus-derived and human word associations. Corpus Linguistics and Linguistic Theory, 7(2), 245-276.


Figure 3: Average association values for each combined measure grouped by cloze probabilities from MTurk experiment.

Paolacci, G., Chandler, J., \& Ipeirotis, P. (2010). Running experiments on amazon mechanical turk. Judgment and Decision Making, 5(5), 411-419.
Seretan, V., \& Wehrli, E. (2009, March). Multilingual collocation extraction with a syntactic parser. Language Resources and Evaluation, 43(1), 71-85.
Smadja, F. (1993). Retrieving collocations from text: Xtract. Association for Computational Linguistics.
Swinney, D., \& Cutler, A. (1979). The access and processing of idiomatic expressions. Journal of verbal learning and verbal behavior, 18(5), 523-534.
Tabossi, P., Fanari, R., \& Wolf, K. (2005). Spoken idiom recognition: Meaning retrieval and word expectancy. Journal of psycholinguistic research, 34(5), 465-495.
Tabossi, P., Fanari, R., \& Wolf, K. (2009). Why are idioms recognized fast? Memory \& cognition, 37(4), 529-540.
Vespignani, F., Canal, P., Molinaro, N., Fonda, S., \& Cacciari, C. (2010). Predictive mechanisms in idiom comprehension. Journal of cognitive neuroscience, 22(8), 1682-1700.

\title{
The Cognitive Reflection Test: how much more than Numerical Ability?
}

\author{
Matthew B. Welsh, Nicholas R. Burns \& Paul H. Delfabbro \\ [\{matthew.welsh\}; \{nicholas.burns\}; \{paul.delfabbro\} @adelaide.edu.au] \\ University of Adelaide, North Terrace \\ Adelaide, SA 5005 Australia
}

\begin{abstract}
Frederick's (2005) Cognitive Reflection Test (CRT) is a 3item task shown to predict susceptibility to decision-making biases better than intelligence measures. It is described as measuring 'cognitive reflection' - a metacognitive trait capturing the degree to which people prefer to reflect on answers rather than giving intuitive responses. Herein, we ask how much of the CRT's success can be explained by assuming it is a test of numerical (rather than general) intelligence. Our results show CRT is closely related to numerical ability and that its predictive power is limited to biases with a numerical basis. Although it may also capture some aspect of a rational cognition decision style, it is unrelated to a metacognitive, error-checking and inhibition measure. We conclude that the predictive power of the CRT can, largely, be explained via numerical ability without the need to posit a separate 'cognitive reflection' trait.
\end{abstract}

Keywords: cognitive reflection; heuristics and biases; individual differences; numerical ability; intelligence.

\section*{Introduction}

Frederick's (2005) Cognitive Reflection Task (CRT) asks people to solve three, mathematically-simple problems on which intuitive answers are wrong. Frederick explains CRT performance as reflecting a person's preference for using either System 1 (intuitive) or System 2 (rational) processes (Stanovich \& West, 2000). Given the ease with which one can check whether intuitive answers are incorrect, the score on CRT shows how likely a person is to reflect on their answer rather than respond intuitively. Frederick's (2005) data shows that CRT is superior to intelligence measures in predicting susceptibility to various cognitive biases or errors made due to inherent, cognitive processes (see, e.g., Tversky \& Kaheman, 1974); a conclusion supported by Toplak, West and Stanovich's (2011) recent work.

Given the surprising finding - that a 3-item test better predicts decision-making ability than intelligence tests, Frederick's work has been influential (cited over 600 times). Its results, however, are in line with previous findings which show that, while intelligence is useful in predicting some decision-making biases, in other cases intelligence and bias susceptibility seem independent (Stanovich \& West, 2008).

These findings have led to suggestions that decision style (or a person's preference for thinking rationally or intuitively) may be more important than intelligence for predicting bias susceptibility. CRT shares variance with a number of decision style measures (Frederick, 2005) and 'cognitive reflection' is thought to be central to the metacognitive processes underlying the relationship between System 1 and System 2 thinking. The latter, System 2
processes, inhibit the automatic and frequently incorrect answers generated by System 1 thinking. It is reasonable, then, that intelligence might determine how efficiently a person uses System 2 reasoning but whether they use it may be determined by a separate, metacognitive process, thereby weakening the observed relationship between intelligence and bias susceptibility.

A potential criticism of Frederick's (2005) paper - and other work in this area - however, lies in the choice of intelligence measures. For example, a commonly used intelligence measure is self-reported SAT scores (see, e.g.: Frederick 2005; Stanovich \& West, 1998). Another is the Wonderlic Personnel Test (Wonderlic, 1973 - used in Frederick, 2005; and Furnham, Boo \& McClelland, 2012). Finally, Toplak et al. (2011), use the Vocabulary and Matrix Reasoning scales from the Wechlser Abbreviated Scale of Intelligence (WASI, Wechsler, 1999).

While all of these do measure 'intelligence' - and WASI divides this into Verbal and Non-verbal ability - none take into account the current understanding of the hierarchical nature of intelligence described by the Cattell-Horn-Carroll model (see, e.g., McGrew, 2005), which recognizes at least ten, related, cognitive abilities. By focusing on the relationship between general intelligence and bias susceptibility, it is, therefore, possible to underestimate the relevance of specific intelligences to specific biases.

A key omission is of numerical ability \(-G q\) or quantitative ability in CHC terms. Given that the CRT, and many decision-making problems, rely on numerical calculation to determine the correct response, it seems strange to report correlations between biases and general intelligence rather than the type of intelligence most likely to influence such tasks. Thus, it seems possible that the low predictive power of intelligence on bias susceptibility results from poor measure selection.

The way forward, then, is to incorporate measures of the specific abilities most likely to relate to the biases under consideration - thereby establishing an accurate baseline for the strength of the relationship before positing additional constructs like cognitive reflection. Concerning metacognition, this work has already begun, with Toplak et al. (2011) including measures of metacognitive abilities (e.g., working memory; Baddeley \& Hitch, 1974) that seem likely to be implicated in recognizing errors in intuition and thus switching from System 1 to System 2 reasoning.

\section*{CRT, Heuristics and Biases}

Given the numerical basis of the CRT questions, a key question is whether it predicts numerical biases better than
less numerical ones. For example, a between-subjects framing task such as the Asian Disease Problem (Tversky \& Kahneman, 1981) is structured so that a person can calculate the expected value of the options and recognize that the values of the options do not change with the frame reversal.
By comparison, the conjunction fallacy (Tversky \& Kahneman, 1983) requires an understanding of the logical rule of conjunction - and numerical ability per se may not assist in avoiding the bias. Similarly, while the anchoring bias (Tversky \& Kahneman, 1974) seems numerical - with a seen number affecting a subsequent estimate - numerical ability can not help a person calculate the correct response.
Other tasks are even less clear cut in this aspect. For example, delay discounting tasks like that used by Frederick (2005) can be regarded as a bias measure indicating the extent to which people misjudge the time value of money. This calculation, however, requires the inclusion of nonnumerical factors such as immediate need for money and degree of trust in the person offering the delayed reward. Recent work, however, has suggested that these actually measure a distinct personality trait - impulsivity (Odum, 2011) - and, thus, one might expect less covariance between numerical ability and delay discounting. Similarly, a base rate neglect task (see, e.g. Bar-Hillel, 1982) can be answered using a variety of distinct response strategies (Welsh \& Navarro, 2012) and, for this reason, it is not necessarily the case that estimates closer to the Bayesian solution actually reflect better numerical skills (Welsh, Burns, Delfabbro \& Begg, 2013) as has traditionally been assumed.

\section*{Method}

\section*{Participants}

Participants were 58 university students and 44 non-students (22 graduates and 22 who had never attended university), recruited via posters and research participation lists, aged between 18 and \(46(M=22.5, S D=4.9)\); sixty-eight were female and all received \(\$ 50\) for participating.

\section*{Materials \& Procedure}

Participants completed an online questionnaire, including demographic details, and the decision style measures described below prior to attending the lab for cognitive and metacognitive tests. The bias measures were included in the online questionnaire - excepting the anchoring task.

\section*{Cognitive Reflection Task}

Frederick's (1995) CRT was used to measure cognitive reflection. This test asks three questions requiring numerical responses with CRT score being the number answered correctly. For example, the first question asks:

A bat and a ball together cost \(\$ 1.10\). The bat costs \(\$ 1.00\) more than the ball. How much does the ball cost?

\section*{Bias Measures}

Anchoring. Anchoring bias refers to the unwarranted effect that presented numbers have on subsequent estimates
(Tversky \& Kahneman, 1974). The measure used here was derived from a computerized card game in which participants estimated the probability that they would win, given the hand they had been dealt (for details, see, Welsh, Delfabbro, Burns \& Begg, in press). Prior to this, they were asked whether their chance of winning was greater or less than a randomly generated number (the anchor) between 0 and \(100 \%\). The anchoring measure was the partial correlation (controlling for the true chance of winning) between the anchor and the person's estimate - measured across 140 hands. Higher values thus reflect greater influence of the anchor on estimates (i.e., more bias).

Base Rate Neglect. The Taxi Cab problem (Bar-Hillel, 1982) requires people to integrate base rate and reliability information to determine the probability of a taxi involved in an accident actually being the color a witness describes. As previously noted (Welsh, Burns, Delfabbro \& Begg, 2013), responses to such problems form distinct categories. We scored responses as Mathematical, Non-Mathematical and Unclassifiable according to whether the person: mathematically combined the probabilities given in the task (i.e., either the Bayesian solution or an incorrect calculation); selected one probability as their response; or did something other than either of these.

Conjunction Fallacy. The Linda Problem (Tversky \& Kahneman, 1983) asks participants to judge whether Linda, a woman described as politically active, is more likely to be a "feminist bank teller" or a "bank teller", with the former indicating the conjunction fallacy - as the conjunction can never be more likely than the simple probability of her being a "bank teller".

Delay Discounting. A series of questions asked how long a person would delay taking a smaller amount of money in order to receive a larger amount. The smaller amount varied from \(\$ 500\) to \(\$ 900\) while the delayed amount was always \(\$ 1000\). The maximum delay a participant would tolerate was indicated on an 8 point scale: 1) 6 hours; 2) 1 day; 3) 1 week; 4) 2 months; 5) 6 months; 6) 1 year; 7) 5 years; 8) 25 years). The average of a person's responses from five such questions was used as their overall score.

Framing. The Asian Flu problem (Tversky \& Kahneman, 1981) asks people to select a treatment schedule for dealing with a disease outbreak - with either certain (200 alive, 400 dead) or uncertain ( \(1 / 3\) chance of all alive, \(2 / 3\) chance of all dead) outcomes. The manipulation lies in the framing of the options. Positive framing describes the treatments in terms of the number of people who live, with the result that more people select the certain option. In contrast, negative framing describes the treatments in terms of the number of people who die, with the result that more people select the uncertain option. Our task included both versions and we categorized people according to whether their responses changed with the frame (displaying the framing bias) or were invariant to framing (unbiased).

\section*{Cognitive and Metacognitive Measures}

Numerical Abilities Test (NAT). A computerized, 12-item version of the 48 -item Numerical Abilities scale from the Differential Aptitudes Test (Bennett, Seashore \& Wesman, 1989), measuring quantitative ability ( \(G q\) ).

Symbol-Digit Test (SD). A computerized measure of cognitive processing speed (Gs) similar to the Wechlser IQ test's Digit-Symbol (see McPherson \& Burns, 2005).

Dot Matrix Task (DM). A computerized version of the Dot Matrix working memory measure (Law, Morrin \& Pelligrino, 1995).

Sustained Attention to Response (SART). A computerized test of executive function - requiring the identification and inhibition of a habituated response (Robertson, Manly, Andrade, Baddeley \& Yiend, 1997).

\section*{Decision Style Measures}

Need for Cognition (NfC). The 10 -item International Personality Inventory Pool (IPIP; Goldberg et al, 2006) version of Cacioppo \& Petty's (1982) scale measuring people's engagement and enjoyment of cognitive activities.

Decision Outcomes Inventory. A 20 -item version of Bruine de Bruin, Parker and Fischoff's (2007) test examining whether people have made various, poor decisions (e.g., bought things they did not use, et cetera). The version we used removed US-specific questions.

Rational Experiential Inventory. A 30-item test of risk style (Epstein, Pacinin, Denes-Raj \& Heier, 1996) yielding four measures distinguishing between 'Ability' and 'Engagement' for two different cognitive styles - Rational (conscious, analytical) and Experiential (intuitive, holistic).

Intellect. A 20-item inventory from the IPIP (Goldberg et al, 2006) combining Cattell's (1973) and Costa and McCrae's (1992) approaches. This measures openness to new ideas in an intellectual context - a facet of 'openness-to-experience' from Costa and McCrae's (1992) NEO-PI-R.

Rationality. Measured by a 14 -item test from the IPIP (Goldberg et al, 2006), high Rationality reflects high Conscientiousness and low Agreeableness in NEO-PI-R (Costa \& McCrae, 1992) terms.

Stimulating Instrumental Risk Inventory. A 28-item test yielding two risk attitude measures (Zaleskiewicz, 2001) Stimulating (positive arousal, short-term and impulsive) and Instrumental (negative arousal, long term and reflective).

\section*{Results}

\section*{CRT and Demographic Measures}

Age did not co-vary significantly with CRT (or with any
measures other than the SRT and IRT scales from the Stimulating Instrumental Risk Inventory). An independent samples \(t\)-test, however, showed that males ( \(M=1.53, S D=\) 1.11) scored significantly higher than females ( \(M=1.01, S D\) \(=0.98\) ), on CRT, \(t(57)=2.18, p=.033\), Cohen's \(d=.51-\) in line with Frederick's (2005) observation.

CRT also varied with level of education, with participants who had never attended university scoring lowest ( \(M=\) \(0.73, S D=0.94\) ), then current university students ( \(M=1.24\), \(S D=1.05\) ) and graduates the highest ( \(M=1.50, S D=1.06\) ) . A one-way ANOVA confirmed these differences were statistically significant, \(F(2,99)=3.31, p=.041\), with posthoc Bonferroni testing indicating that the no-university group differed significantly from both others.

\section*{CRT and Biases}

Table 1 summarizes the relationships between the CRT and the five bias measures - noting those that have numerically calculable correct responses and which showed significant relationships with the CRT.

Looking at the table, one sees an interesting pattern of responses. While CRT has relationships in the expected directions with the non-calculable biases (Anchoring and the Conjunction Fallacy) these are very weak. By comparison, its relationships with Framing and Discount Delay measures, where the unbiased answer can be calculated, are statistically significant if moderate and weak, respectively. The more complex, near significant relationship between CRT and Base Rate Neglect is discussed more fully below.

\section*{CRT and Numerical Ability}

Scores on the Numerical Ability Test (NAT) correlated significantly with CRT, \(r(102)=0.44, \mathrm{p}<.001-\) comparable to the correlations observed between CRT and cognitive ability measures in Fredrick (2005) and Toplak et al. (2011). This correlation is the strongest that CRT has with any measure in our analyses.

The relationships between NAT and the demographic variables were also calculated - to determine whether the pattern of responses matches those of the CRT. Welch's \(t\) test revealed a non-significant relationship between NAT and Sex in the same direction as the significant relationship shown by the CRT measure, \(t(57)=1.28, p=.204\).

The relationship between NAT and Education, by contrast, was significant, as indicated by a one-way ANOVA, \(F(2,99)=9.41, p<.001\). As with CRT, a Bonferroni post hoc test indicated that the no-university group's lower scores drove the significant result and the groups were ordered in the expected manner: no-university; current student; and, then graduates.

\section*{Factor Analysis}

To assess relationships between CRT and the individual differences measures, an exploratory factor analysis (minres extraction with geomin oblique rotation) was run, revealing the 4 -factor solution seen in Table 2. (NB - 2- and 3-factor
solutions were also considered; these did not appreciably alter the loadings of the CRT on the first two factors.)

Table 1. Summary of Bias Task Characteristics and Results
\begin{tabular}{|c|c|c|c|}
\hline Task & \begin{tabular}{l}
范 \\

\end{tabular} &  & Results \\
\hline Framing & Yes & Yes & \begin{tabular}{l}
\(t(64)=2.97, p=.004\), Cohen's \(\boldsymbol{d}=\mathbf{. 6 2}\). \\
People whose responses varied with the frame scored lower than those whose responses were invariant to the frame (CRT \(=0.76\) vs 1.39 ).
\end{tabular} \\
\hline \begin{tabular}{l}
Discount \\
Delay
\end{tabular} & Yes & Yes & \begin{tabular}{l}
\[
r(102)=0.25, p=.010
\] \\
Higher CRT accompanied a greater willingness to wait for the larger reward.
\end{tabular} \\
\hline Base Rate Neglect & Yes & No & \begin{tabular}{l}
\[
F(2,98)=2.79, p=.07
\] \\
People whose responses were classified as Mathematical (CRT=1.28) did not score better than the Non-Mathematical group (1.33) but both scored better than the Unclassified group (0.71).
\end{tabular} \\
\hline Anchoring & No & No & \begin{tabular}{l}
\[
r(102)=-0.11, p=.255
\] \\
People more susceptible to anchoring bias scored slightly lower on CRT
\end{tabular} \\
\hline Conjunction Fallacy & No & No & \begin{tabular}{l}
\[
t(56)=0.37, p=0.71
\] \\
People committing fallacy scored slightly lower on CRT (1.30 vs 1.21 ).
\end{tabular} \\
\hline
\end{tabular}

Looking at Table 2, one can see that a sensible structure emerges. The first factor captures the decision style measures relating to people's tendencies toward 'rational cognition'. The second seems to be an intelligence factor. The third has only the experiential measures from the Rational-Experiential Inventory (REI) loading on it reflecting a tendency toward intuitive thinking. Finally, the fourth factor reflects attitudes to risk as captured by both measures form the Stimulating-Instrumental Risk Inventory.
Only one variable, the Rational Ability measure from the REI, loads on more than one factor at the conventional 0.3
level and only the Sustained Attention to Response Task fails to load on any factor - indicating the metacognitive measure differs from both decision style and intelligence.

Table 2. Factor loadings of CRT, cognitive and decisionstyle measures.
\begin{tabular}{lrrrrr}
\hline & \multicolumn{5}{c}{ Factors } \\
Variable & 1 & 2 & 3 & 4 & \(h^{2}\) \\
\hline Intellect & \(\mathbf{. 9 2}\) & -.06 & .13 & .00 & .94 \\
Need for Cognition & \(\mathbf{. 9 1}\) & .02 & .03 & .05 & .85 \\
Rational Engagement & \(\mathbf{. 7 5}\) & -.06 & -.02 & .01 & .56 \\
Rationality & \(\mathbf{. 7 4}\) & .04 & .00 & -.05 & .55 \\
Rational Ability & \(\mathbf{. 6 2}\) & .32 & -.12 & -.01 & .48 \\
Dot Matrix & -.08 &. \(\mathbf{. 7 7}\) & -.03 & .12 & .61 \\
Symbol-Digit & .00 & \(\mathbf{. 6 8}\) & .16 & -.06 & .44 \\
Numerical Ability & .00 & \(\mathbf{. 6 6}\) & .02 & .01 & .43 \\
Cognitive Reflection & .26 & \(\mathbf{. 5 0}\) & .01 & -.05 & .32 \\
Exper. Engagement & -.01 & .02 & \(\mathbf{1 . 0 0}\) & .00 & .00 \\
Experiential Ability & .16 & -.02 & \(\mathbf{. 6 8}\) & .02 & .56 \\
Stimulating Risk & -.02 & -.02 & .04 & \(\mathbf{. 9 9}\) & 1.00 \\
Instrumental Risk & .04 & .24 & -.17 & \(\mathbf{. 4 5}\) & .26 \\
Sustained Attention RT & -.01 & -.01 & .05 & .18 & .05 \\
\hline Prim
\end{tabular}

Primary factor loadings are in bold. \(h^{2}=\) communality, the variance in each variable captured by the four factors.

\section*{Discussion}

The above results suggest that 'cognitive reflection', as measured by CRT, shares much in common with numerical ability - although there remains additional, unshared variance to account for. Key, individual results are discussed below, along with caveats and potential future research.

\section*{Cognitive Reflection and Sex}

An interesting result is the relationship between CRT and Sex - and the lack of similar relationships between Sex and the other measures loading on the 'intelligence' factor in Table 2. The sex difference on CRT was observed by Frederick (2005), who noted that it was unrelated to differences in intelligence and suggested that it might be related to differences in mathematical ability. This was not, however, supported by our data where no significant relationship was seen between numerical ability and sex.

The only variable with which both Sex and CRT shared a relationship was the Rational Ability scale of the RationalExperiential Inventory. CRT correlated with RA significantly, \(r(102)=0.33, p<.001\) and men's scores (22.7) were higher than women's (20.9) - significantly according to Welch's \(t\)-test, \(t(83)=2.34, p=.022\) suggesting that the sex difference in CRT may partly reflect a difference in Rational Ability - a person's self-reported ability to think analytically (Epstein et al, 1996).

\section*{Cognitive Reflection and Numerical Biases}

The pattern of bias results described above fit with a conception of the CRT as a primarily numerical measure.

On those bias tasks where numerical skill has no obvious role in arriving at the correct response - anchoring and the conjunction fallacy, the CRT has no predictive value.
By comparison, in the framing problem, where the irrelevance of the frame can be demonstrated numerically, CRT proved a good predictor of performance. Similarly, there is a significant effect for the delay discounting problem. Despite the complexity of the problem (in terms of potential, contextual factors) it appears that numerical ability pushes participants towards the economically rational choice. This is an interesting addition to Baumann and Odum's (2012) finding that delay behavior relates to temporal perception - potentially arguing for a relationship between numerical and temporal skills under the broad \(G q\) 'quantitative ability' umbrella.

Complexity is added by the base rate neglect task, where the results were somewhat unexpected - although not significant. As noted above, the groups using mathematical and non-mathematical strategies did not differ statistically from one another on the CRT. Instead, both groups outscored participants whose responses were unclassifiable. As noted by Welsh et al. (2013), however, the base rate neglect task differs from many numerical bias tasks in that the calculation of the correct solution is dependent on knowing how to undertake Bayesian updating. That is, while a person with high cognitive reflection or numeracy may realize that their intuitive response is wrong and activate System 2 thinking, they may not have the knowledge required to calculate the correct answer having done so. Given that CRT only requires very simple mathematical skills - as do numerical ability tests - this task's failure to predict response types on a base rate neglect task is less surprising than it first seems.

\section*{Cognitive Reflection and Intelligence}

The factor analysis shown in Table 2 indicates that the CRT is, primarily, an intelligence measure - loading on the second factor along with the three cognitive variables. It does, however, have the weakest loading of the four on this factor at 0.50 . Numerical ability is, however, the variable with the most similar loading - reflecting the strength of the relationship between these two measures. This is unsurprising as the Dot Matrix and Symbol-Digit tasks require learning a novel task, whereas the NAT and CRT require prior knowledge - of how to undertake mathematical operations. (CRT scores could also be affected by prior experience of questions similar to those used in the task making people wary of too-easy answers.)
CRT shows virtually no relationship (loadings of .01 and .05) with the third and fourth factors ('intuition' and 'risk attitude') but its relationship to the first factor bears some scrutiny. While not reaching the 0.3 level conventionally required to be included amongst the variables loading on a factor, its loading of 0.26 on the 'rational cognition' decision style factor approaches this level and is the second highest secondary loading in Table 2 - after Rational Ability's 0.32 secondary loading on the 'intelligence' factor.

This could be taken as offering some support for Frederick (2005) and Toplak et al.'s (2011) conclusions that CRT measures something more than cognitive ability although the factor loadings suggest that the cognitive aspect is more central.

\section*{Cognitive Reflection and Metacognition}

A final observation from the above results is the lack of any relationship between the CRT and the Sustained Attention to Response Task (SART), which measures executive functioning - specifically, a person's ability to monitor their performance for errors and to inhibit incorrect responses.

Given the description of the CRT as a measure of a person's preference for activating rational thinking and thus recognizing errors in intuitive responses, its failure to correlate with the SART seems strange. In light of our results, it thus seems possible that the CRT is measuring only a person's ability to recognize errors in intuitive, numerical results rather than the more general metacognitive function.

\section*{Caveats and Future Research}

While including measures not previously used in studies of cognitive reflection, the present analyses remain limited in their scope. The Cattell-Horn-Carroll model includes ten specific types of intelligence (acknowledging the possibility of more; McGrew, 2005). Of these, only two (plus the nonCHC working memory) were measured herein - Gs (processing speed) and \(G q\) (quantitative or numerical ability). Similarly, while five biases were included here, further effects from the biases literature could improve understanding of what CRT does and does not predict.

A further concern is the sample size. While 102 participants is sufficient to find most large or moderate effects, small effects may still be missed. Frederick's (2005) study, for example, involved more than 3000 participants, allowing statistical significance for even very weak relationships. Given this, the obvious direction for future research is a larger study of participants from a wide range of educational backgrounds, utilizing the widest possible range of biases and cognitive abilities in order to pin down exactly what the CRT is. Including a number of tasks measuring quantitative (numerical) ability would also allow further factor analyses to decide conclusively whether CRT is, as suggested here, primarily a numerical task.

Another key direction is to determine what role metacognitive abilities do play in the divide between System 1 (intuitive) and System 2 (analytic) reasoning and whether CRT is capturing any of these. Specific measures of impulsivity, as discussed by Baumann and Odum (2012), could inform this - as this seems likely to relate to the likelihood of a person relying on System 1.

Finally, additional work could address whether CRT scores are affected by prior experience of similar questions.

\section*{Conclusions}

The above results support the idea that CRT is, at heart, a
numerical task, correlating with quantitative ability and predicting bias only on tasks with a calculable, correct answer. It may, however, measure some aspect of a 'rational cognition' decision style. The CRT does not, however, relate to the executive functioning measure included here, suggesting that 'cognitive reflection' may not be metacognitive as Frederick (2005) describes but, rather, measure a person's ability to quickly recognize bad math.

\section*{Acknowledgments}

MBW and SHB thank ExxonMobil and Santos for their support of this research through the CIBP at the Australian School of Petroleum. The authors thank Karina Burns for her assistance with data collection and Reidar Bratvold and Steve Begg for discussions re the CRT's nature and use.

\section*{References}

Baddeley, A. D., \& Hitch, G. J. (1974). Working memory. The psychology of learning and motivation, 8, 47-89.
Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. Acta Psychologica, 44, 211-233.
Baumann, A. A., \& Odum, A. L. (2012). Impulsivity, risk taking, and timing. Behavioural Processes, 90, 408-414.
Bennett, G. K., Seashore, H. G., \& Wesman, A. G. (1989). Differential Aptitude Test. Marrickville, NSW: Psychological Corporation.
Bruine de Bruin, W., Parker, A.M., \& Fischhoff, B. (2007). Individual differences in adult decision-making competence. Journal of Personality and Social Psychology, 92, 938-956.
Cacioppo, J. T., \& Petty, R. E. (1982). The need for cognition. Journal of Personality and Social Psychology, 42(1), 116-131.
Cattell, R.B. (1973). Personality and mood by questionnaire. San Francisco: Jossey-Bass.
Costa, P. T. J., \& McCrae, R. R. (1992). NEO PI-R Professional Manual. Odessa, FL: Psychological Assessment Resources.
Epstein, S., Pacini, R., Denes-Raj, V., \& Heier, H. (1996). Individual differences in Intuitive-Experiential and Analytical-Rational Thinking Styles. Journal of Personality and Social Psychology, 71(2), 390-405.
Frederick, S. (2005). Cognitive reflection and decision making. Journal of Economic Perspectives, 19(4), 25-42.
Furnham, A., Boo, H.C. \& McClelland, A. (2012). Individual differences and the susceptibility to the influence of anchoring cues. Journal of Individual Differences, 33(2), 89-93.
Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., \& Gough, H. C. (2006). The International Personality Item Pool and the future of public-domain personality measures. Journal of Research in Personality, 40, 84-96.
Law, D.J., Morrin, K.A., Pellegrino, J.W. (1995). Training effects and working memory contributions to skill acquisition in a complex coordination task. Learning and Individual Differences, 7, 207-234.

McGrew, K. S. (2005). The Cattell-Horn-Carroll theory of cognitive abilities: Past, present, and future. In D. P. Flanagan, J. L. Genshaft, \& P. L. Harrison (Eds.), Contemporary intellectual assessment: Theories, tests, and issues (pp.136-182). New York: Guilford.
McPherson, J.L \& Burns, N.R. (2005). A speeded coding task using a computer based mouse response. Behavior Research Methods, Instruments \& Computers, 37, 538544.

Odum, A.L. (2011). Delay discounting: trait variable? Behavioural Processes, 87, 1-9.
Robertson, I.H., Manly, T., Andrade, J., Baddeley, B.T., Yiend, J. (1997). 'Oops!': performance correlates of everyday attentional failures in traumatic brain injured and normal subjects. Neuropsychologia, 35, 747-758.
Stanovich, K. E., \& West, R. F. (1998). Individual differences in rational thought. Journal of Experimental Psychology: General, 127(2), 161-188.
Stanovich, K. E., \& West, R. F. (2000). Individual differences in reasoning: implications for the rationality debate. Behavioral and Brain Sciences, 23, 645-726.
Stanovich, K. E., \& West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. Journal of Personality and Social Psychology, 94(4), 672695.

Toplak, M.E., West, R.F. \& Stanovich, K.E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. Memory and Cognition, 39, 1275-1289.
Tversky, A., \& Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185, 11241131.

Tversky, A., \& Kahneman, D. (1981). The framing of decisions and the psychology of choice. Science, 211(4481), 453-458.
Tversky, A., \& Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. Psychological review, 90(4), 293.
Wechsler, D. (1999). Wechsler abbreviated scale of intelligence (WASI). San Antonio: Harcourt Brace, Psychological Corp.
Welsh, M.B., Burns, N.R., Delfabbro, P.H. \& Begg, S.H. (2013). Individual differences in base rate neglect responses and susceptibility. Manuscript in preparation. www.adelaide.edu.au/directory/matthew.welsh.
Welsh, M.B., Delfabbro, P.H., Burns, N.R., Begg, S.H. (in press). Individual differences in anchoring: traits and experience. Learning and Individual Differences: Special Issue on Metacognition. Accepted 9 \(9^{\text {th }}\) January 2013.
Welsh, M. B., \& Navarro, D. J. (2012). Seeing is believing: priors, trust and base rate neglect. Organizational Behavior and Human Decision Processes, 119(1), 1-14.
Wonderlic, E. F. (1973). Wonderlic Personnel Test. Los Angeles: Western Psychological Services.
Zaleskiewicz, T. (2001). Beyond risk seeking and risk aversion: Personality and the dual nature of economic risk taking. European Journal of Personality, 15, S105-S122.

\title{
How do you know that? Sensitivity to statistical dependency in social learning
}

\author{
Andrew Whalen (awhalen@gmail.com) \\ Daphna Buchsbaum (daphnab@berkeley.edu) \\ Thomas L. Griffiths (tom griffiths@berkeley.edu) \\ Department of Psychology, University of California, Berkeley, Berkeley, CA 94720 USA
}

\begin{abstract}
Social learning has been shown to be an evolutionarily adaptive strategy, but can be implemented via many different cognitive mechanisms. Sensitivity to statistical dependency in the behavior of other people is a factor that discriminates between possible mechanisms: simple rule based strategies may be unaffected by dependency, while more sophisticated social learning strategies should take it into account. We use a Bayesian model to determine how rational agents should incorporate the effects of statistical dependency when learning from other people, conducting two experiments that examine whether human learners behave similarly. We find that people are sensitive to two different patterns of dependency, supporting the use of a sophisticated strategy for social learning.
\end{abstract}

\section*{Introduction}

Social learning is a key factor in the human ability to adapt to a wide variety of environments and plays an important role in cultural transmission of information (Boyd \& Richerson, 1988, 2005). Formal models have shown that social learning is an evolutionarily adaptive strategy, able to outcompete individual learning (Laland, 2004). However, there are many possible mechanisms by which social learning could be implemented, ranging from blind imitation to making sophisticated inferences about the beliefs that underlie that behavior. While evolutionary models tell us that social learning should be favored, they don't tell us which mechanism human learners might be using. This question is particularly important given results showing that both adults and children sometimes "overimitate", reproducing another's unnecessary actions (e.g., Lyons, Young, \& Keil, 2007; Nielsen \& Tomaselli, 2010; McGuigan, Makinson, \& Whiten, 2011).

In this paper, we explore the mechanisms behind human social learning by examining how sensitive people are to statistical dependency in the behavior of other people. For example, imagine hearing from two friends that they visited a particular restaurant. Taken at face value, this seems like strong evidence that the restaurant might be a good place to eat. But if you discover that one friend went there after finding out that the other had been, the two pieces of information are no longer statistically independent and the evidence they provide about the quality of the restaurant is reduced. And if one friend had taken the other there, it is reduced even further.

Examining whether human social learning is sensitive to statistical dependency provides an opportunity to discriminate between social learning strategies. Simpler rule-based approaches such as "imitate the majority" should be insensitive to the subtleties of how other people's behavior is generated, focusing just on the behavior itself. In contrast, if social learning is based on rational inferences from the available data, the way in which those data are generated should
matter a great deal (for an example see Buchsbaum, Gopnik, Griffiths, \& Shafto, 2011). Determining the consequences of dependencies in other's behavior involves reasoning about their mental states and the factors that contribute to their decisions, requiring a sophisticated approach to social learning.

To assess whether people are appropriately sensitive to statistical dependency in the behavior of others, we developed a Bayesian model for a simple social learning task. The model indicates what inferences a rational agent should draw when statistical independence is violated in different ways. We ran two behavioral experiments using this social learning task, finding that people are sensitive to two forms of dependency. These results support the idea that human social learning is based on reasoning about the mental states of other people, rather than simpler strategies such as imitating the majority.

\section*{Learning from others}

Before presenting our model and experiments, we will summarize some of the key theoretical and experimental results on cultural evolution and social learning. These results break down into three areas of research. At the largest scale, models of cultural evolution have examined how the learning strategies adopted by individuals impact the spread of different behaviors between generations. Within generations, models of what are called "information cascades" have been used to analyze the rapid spread of novel innovations among populations. Finally, a number of studies have explored how individual people learn from informant testimony.

\section*{Cultural evolution}

Theoretical studies of cultural evolution have shown that social learning has adaptive advantages (Boyd \& Richerson, 1988, 2005; Laland, 2004). However, many of these studies analyze systems where individuals are faced with the choice of either learning from the environment or learning socially (Rogers, 1988). In reality, learners are likely to combine both environmental and social information when making a decision. Perreault, Moya, and Boyd (2012) modeled agents who choose a behavior based on a Bayesian learning algorithm which integrates social and environmental information. In this model, agents assumed that the social cues provided by other agents were independent from one another. This assumption was justified by the fact that all behavioral transmission happened between generations where the probability that the informants learned from each other was low. However, many behaviors are transmitted within generations, where informants are likely to share information.

\section*{Information cascades}

Unlike the cultural evolution literature, the literature on information cascades, a within-generation model of social decision-making developed by economists, takes into account the statistical dependency between agents (Bikhchandani, Hirshleifer, \& Welch, 1992). The basic scenario has a sequence of agents each making a decision by combining the information provided by the decisions made by previous agents with that provided by a small amount of private data. An information cascade occurs when agents adopt the majority belief, regardless of their own private information. The cascade persists as more agents enter the population and adopt the majority belief. Bikhchandani et al. (1992) analyzed how rational agents who took into account dependencies in previous responses would act in this situation, and showed that information cascades are surprisingly common. This result provides a potential explanation for the adoption and spread of fads and fashions, as well as boom-bust cycles in the economy. Information cascades have been tested in the laboratory using a simple scenario that provided the inspiration for the experiments we present later (Anderson \& Holt, 1997), but this previous work did not examine the consequences of manipulating people's beliefs about statistical dependency.

\section*{Individual decision making}

Social learning, and imitation in particular, have been studied extensively by psychologists. This work has generally demonstrated that adults and even young children are sensitive to many aspects of the knowledge and mental states of their social informants (for some well known examples see Meltzoff, 1995; Gergely, Bekkering, \& Kiraly, 2002; Carpenter, Call, \& Tomasello, 2005). Related work on how children learn from testimony has similarly found that children take many factors into account, including the prior accuracy (e.g., Koenig \& Harris, 2005; Corriveau, Meints, \& Harris, 2009), expertise (e.g. Jaswal, 2006; Sobel \& Corriveau, 2010) and certainty (Jaswal \& Malone, 2007; Tenney, Small, Kondrad, Jaswal, \& Spellman, 2011) of informants. However, other work has found that in some situations, people appear to simply copy the beliefs of others. Adults often disregard their own judgments when socially pressured (for a review see Cialdini \& Goldstein, 2004), and both adults and children may sometimes conform to a majority opinion that conflicts with their own direct perceptions (Asch, 1956; Corriveau \& Harris, 2010). Looking at the effect of statistical dependency can help us determine if this conformity is the result of a simple rule-based strategy, or a more sophisticated inference process.

\section*{Rational social learning}

Many inferences that people make rely upon a combination of their own experience and the behavior of other people. In order to determine how a sophisticated agent should combine these forms of information, we developed a Bayesian model that can incorporate different patterns of dependency. This
model makes direct predictions that we can test experimentally, having no free parameters.

We assume that agents receive some directly observed data about the state of the world, \(d\), and testimony from \(n\) informants \(t_{1}, \ldots, t_{n}\). To make a decision, learners evaluate a potential hypothesis, \(h\), using Bayes' rule,
\[
\begin{equation*}
p\left(h \mid d, t_{1}, \ldots, t_{n}\right) \propto p\left(t_{1}, \ldots, t_{n} \mid d, h\right) p(d \mid h) p(h) \tag{1}
\end{equation*}
\]
where \(p\left(h \mid d, t_{1}, \ldots, t_{n}\right)\) is the posterior probability of \(h\), the degree of belief assigned to \(h\) after receiving the data and testimony, and \(p(h)\) is the prior probability of \(h\), the degree of belief assigned to \(h\) before receiving any evidence.

In order to estimate the probability of the testimony, \(p\left(t_{1}, \ldots, t_{n} \mid d, h\right)\), the learner should consider the sources of information that each informant had access to. If the data the learner observes, \(d\), is unknown to the informants, then \(p\left(t_{1}, \ldots, t_{n} \mid d, h\right)=p\left(t_{1}, \ldots, t_{n} \mid h\right)\). We will assume that this is the case, since it simplifies calculations and is consistent with the task we use in our experiments. The form of \(p\left(t_{1}, \ldots, t_{n} \mid h\right)\) depends on how the informants generate their testimony. We first consider the case of independent testimony, and then discuss two different patterns of dependency.

\section*{Independent testimony}

If the informants' testimonies are independent of one another given \(h\), then the probability of a series of testimonies is equal to the product of the probability of the individual testimonies:
\[
\begin{equation*}
p\left(t_{1}, \ldots, t_{n} \mid h\right)=\prod_{i=1}^{n} p\left(t_{i} \mid h\right) \tag{2}
\end{equation*}
\]

If the testimony produced by the informants is based on their own experiences, this needs to be taken into account in calculating the probability that they would produce their testimony. More formally, if we assume that informant \(i\) received private data \(d_{i}\), we obtain \(p\left(t_{i} \mid h\right)\) by marginalizing over \(d_{i}\),
\[
\begin{equation*}
p\left(t_{i} \mid h\right)=\sum_{d_{i}} p\left(d_{i} \mid h\right) p\left(t_{i} \mid d_{i}\right) \tag{3}
\end{equation*}
\]
where \(p\left(t_{i} \mid d_{i}\right)\) is the probability that the informant produces testimony \(t_{i}\) after observing \(d_{i}\). One possibility is that informants deterministically give testimony that supports the hypothesis with the highest posterior probability, with \(p\left(t_{i}=\right.\) \(\left.h_{i} \mid d_{i}\right)=1\) for \(h_{i}=\arg \max _{h} p\left(d_{i} \mid h\right) p(h)\). This is typically assumed in models of information cascades (e.g., Bikhchandani et al., 1992). Alternatively, empirical (Vulkan, 2000) and theoretical (Luce, 2005; Shepard, 1958) results in psychology suggest that in many cases people "probability match", so that informants would give testimony in support of a hypothesis proportional to the informant's posterior probability of the hypothesis, with \(p\left(t_{i}=h_{i} \mid d_{i}\right) \propto p\left(d_{i} \mid h_{i}\right) p\left(h_{i}\right)\). We evaluate both the maximizing and probability matching models.

\section*{Dependent testimony}

If multiple informants give testimony based on shared information, then the probability of any single testimony is not
independent from the others. We consider two cases: where informants give their testimony sequentially, with each informant hearing the preceding testimony, and where informants base their testimonies on shared private data.

Sequential testimony Much of the theoretical work on information cascades assumes that informants give their testimony sequentially. Each informant uses their own private information, and the testimony of previous informants to make a decision as to which option to support. Formally, the testimony of informant \(i\) is based on the previous testimony of the previous informants, \(t_{1}, \ldots, t_{i-1}\), and their own private data, \(d_{i}\). The probability of \(t_{1}, \ldots, t_{n}\) is then
\[
\begin{equation*}
p\left(t_{1}, \ldots, t_{n} \mid h\right)=p\left(t_{1} \mid h\right) \prod_{i=2}^{n} p\left(t_{i} \mid t_{1}, \ldots, t_{n}, h\right) . \tag{4}
\end{equation*}
\]

The value \(p\left(t_{i} \mid t_{1}, \ldots, t_{i-1}, h\right)\) can be found recursively by finding the values for \(p\left(t_{1} \mid h\right)\) up to \(p\left(t_{i-1} \mid t_{1}, \ldots, t_{i-2}, h\right)\) :
\[
\begin{equation*}
p\left(t_{i} \mid t_{1}, \ldots, t_{i-1}, d_{i}, h\right) \propto\left(\prod_{j=1}^{i-1} p\left(t_{j} \mid t_{1}, \ldots, t_{j-1}, h\right)\right) p\left(d_{i} \mid h\right) p(h) . \tag{5}
\end{equation*}
\]

As in the case of independent informants, we can find \(p\left(t_{i} \mid t_{1}, \ldots, t_{i-1}, h\right)\) by marginalizing over the private information, \(d_{i}\), and assuming informants apply Bayes' rule and then either maximize or probability match.

Shared private data If the informants all provided testimony based on a single piece of data (e.g., they all went to the restaurant together), then the probability of this testimony can be found by marginalizing over this shared private data. Denoting the shared data \(d^{\prime}\), we have
\[
\begin{equation*}
p\left(t_{1}, \ldots, t_{n} \mid h\right)=\sum_{d^{\prime}} p\left(d^{\prime} \mid h\right) \prod_{i} p\left(t_{i} \mid d^{\prime}, h\right) \tag{6}
\end{equation*}
\]
where the probabilities \(p\left(t_{i} \mid d^{\prime}, h\right)\) are calculated by applying Bayes rule and assuming either maximizing or probability matching to the posterior, as above.

\section*{Reasoning about balls and urns}

The consequences of different forms of dependency for rational social learning can be hard to imagine in abstract, so we will work through a concrete example in detail. One of the simplest examples that illustrates these consequences is the "ball and urn" scenario used in the information cascade experiment conducted by Anderson and Holt (1997). This scenario is also the basis for our own experiments.

Imagine there are two colored urns. One of the urns is colored red, the other urn is colored blue. An experimenter explains that in the red urn \(\frac{5}{6}\) of the balls are red, and the rest of the balls are blue. In the blue urn the proportions are reversed. In secret, the experimenter pours one of the urns into a bag. She then shows a ball to each of three informants, and one to the participant. The informants say which urn they think was used to fill the bag, based on the information available to them. The experimenter then asks the participant to decide


Figure 1: Probability of agreeing with the majority opinion in the simple balls and urns task used in Experiment 1, for both the Bayesian model and human participants.
which urn was used to fill the bag, based on the testimony of the informants and the ball seen by the participant.

If all three informants agreed with each other and thought the bag was filled from the red urn, but the participant got a blue ball, what should the participant say? We will analyze three conditions, corresponding to the three cases presented in the previous section. The predictions for the three conditions are shown in Figure 1(a) for the maximizing model and in Figure 1(b) for the probability matching model, using the true probabilities of red and blue balls for \(p(d \mid h)\) and assuming both hypotheses are equally likely for \(p(h)\).

\section*{Independent testimony}

Imagine that the three informants are all in separate rooms and each receive a different ball sampled from the bag, making their testimony completely independent. In this case, the model predicts that the participant should agree with the social testimony, picking the red urn. The model infers that all three informants all probably received red balls and three red balls outweigh the participant's single blue ball.

\section*{Sequential testimony}

In this case, all three informants might be sitting at the same table and each receive a different ball, but have the opportunity to hear the answer given by the previous informants before providing their testimony. This is the situation that was analyzed in Anderson and Holt's (1997) experiment. If the informants give their testimony sequentially, the model again predicts that the participant should agree with the social testimony. However, the model places less weight on the hypothesis that the red urn was used to fill the bag. The model takes into account the fact that the three informants shared information. If the first two people received red balls and the third person received a blue ball, they may still all agree that the red urn was used to fill the bag even if the third person goes against the private evidence she received - a mere blue ball against two likely red balls - and votes in favor of the majority. This possibility makes the model less sure of its decision compared to the independent condition.

\section*{Shared private data}

Now, consider what happens if all three informants are sitting at the same table and all observe exactly the same ball, rather than each seeing a separate ball drawn from the bag. If all three informants saw the same ball, the model is evenly
split between the two urns. On the one hand, the three informants probably received a single red ball, but the participant received a blue ball. With one red ball and one blue ball on the table, the balls provide equal evidence for either urn being used to fill the bag.

\section*{Summary}

Even in a simple scenario with two hypotheses and three informants, a rational social learner should act differently in response to different patterns of statistical dependency. To compare our model with human behavior, we ran an experiment to see how people incorporate their own understanding of the information each informant used to give their testimony.

\section*{Experiment 1: Consistent informants}

Experiment 1 used the scenario presented in the previous section, with three informants providing consistent testimony that went against the private data received by the participant. There were three conditions corresponding to the independent testimony, sequential testimony, and shared private data.

\section*{Methods}

Participants A total of 120 participants were recruited through Amazon Mechanical Turk (http://www.mturk.com). Participants were compensated \(\$ 0.25\) for their time. They were randomly assigned to one of three experimental groups: the independent condition ( \(n=37\) ) or the shared testimony \((n=41)\), or shared-data ( \(n=45\) ). No participants were dropped from the analysis.

Stimuli The experiment was a web-administered survey involving text and pictures. A cartoon of a brown haired woman was the experimenter. Three cartoon women were the informants. The informants differed in terms of hair color, hair style, skin color, and shirt color. Each urn was a picture of a red or blue opaque urn. The balls were colored red and blue.

Procedure First a woman named Jane (the experimenter) introduced the urns. She explained that five-sixths of the balls in the red urn were red, and one-sixth were blue. The opposite was true for the blue urn. She introduced her three friends (the informants), and explained that she was going to pour one of the urns into a bag and give a ball from the bag to each of her three friends. The friends would then tell the participant which urn they think the bag was filled from. In all three conditions the three informants agreed that the bag was filled from the red urn. The participant then saw a blue ball. \({ }^{1}\)

In the independent testimony condition the participant was shown three doors, and was told that one informant was waiting in each room. Inside, each informant sat behind a desk.

In the sequential testimony condition the informants sat behind a long table. The informants gave their testimony in order down the table and acknowledged that they had used their own ball and the testimony of previous informants to make

\footnotetext{
\({ }^{1}\) The actual colors were randomized, so half the participants received testimony favoring the blue urn and then saw a red ball.
}
their decision, but did not see anyone else's ball. Each informant agreed with the previous informants' testimony.

The shared private data condition was the same as the sequential testimony condition, except that a single ball was shared between the informants, and each informant said that they saw the same ball as the other informants. The experimenter then showed the participant a single blue ball, contrary to the three informants' testimony.

Finally, the experimenter asked participants to rate how likely it was that the bag was filled each urn. Participants responded to the survey on an 11-point scale, with 0 corresponding to "definitely the blue urn", 10 to "definitely the red urn", and 5 to "equally likely the blue urn or red urn".

\section*{Results and Discussion}

Ratings were placed on a consistent scale, corresponding to agreement with the majority, by recoding a rating \(x\) to \(10-x\) if testimony favored the blue urn. The mean rescaled ratings for all conditions are shown in Figure 1(c). Overall, participants sided with the informants' testimony over their own private information most in the independent condition, second in the sequential testimony condition, and least in the shared private data condition. The ordering of the means are consistent with the model predictions. The matching model provided a good model fit to the data (Pearson's \(r=.90\) ).

We analyzed the effect of condition on participant responses using an ANOVA. The effect of condition was significant \((F(2,99.1)=7.749, M S E=49.56, p<0.001)\). We explored the differences between conditions using planned t-tests. The difference between the independent and shared private data conditions was significant (two sample t-test, \(t(80)=3.88, p<.001)\) as well as the difference between the sequential testimony and shared private data conditions (two sample t-test, \(t(84)=2.66, p<.01\) ). The difference between the sequential testimony and independent testimony conditions was not (two sample t-test, \(t(76)=0.96, p=.34\) ).

The difference between the shared private data condition and the independent condition suggests that participants were able to use their knowledge of what information informants received to evaluate the informants' testimony. Because the three informants received the same ball and gave the same testimony, participants were able to weigh their judgments against their own conflicting ball.

However, both the maximizing and the probability matching models predict that in the shared private data condition the probability of the bag having mostly red balls is approximately \(50 \%\) : less than the participant's average value of \(60 \%\). Even though this difference was not significant (one sample t-test, \(t(44)=1.31, p>.05\) ), participants may place more weight on informant testimony than the model predicts.

At first glance, the null result between the sequential testimony and independent testimony conditions suggests that people respond similarly in the cases of independent testimony and sequential testimony. However, the magnitude of the difference between these two conditions predicted by the model is relatively small. This suggests instead that the sce-


Figure 2: Probability of agreeing with the majority opinion in the more complex task used in Experiment 2, for both the Bayesian model and human participants.
nario used in previous experiments on information cascades may not be sufficient to distinguish between how people use independent testimony over sequential testimony, a limitation that we address in our second experiment.

\section*{Experiment 2: Dissenting informant}

The ball-and-urn scenario presented above does not result in situations where there is a large expected difference between the independent testimony and sequential testimony conditions. In order to assess whether people are sensitive to the difference between these two patterns of dependency, we changed the scenario by having the third informant dissent from the previous two informants. To give a reason why the informant would dissent, a single diagnostic ball (either white or black) was added to each of the two urns. Since each diagnostic ball was present in only one of the two urns, any informant who received the diagnostic ball would know exactly which urn was used to fill the bag. We also made two other changes. First, the participant did not receive their own ball, having to make a judgment based purely on the testimony of the informants. Second, in the shared private data condition only the first two informants received the same ball. This was done so that the dissenter received her own ball, providing an explanation for why she might dissent.

The model predictions are given in Figure 2(a), for maximizing, and Figure 2(b), for probability matching. The addition of a low-probability diagnostic ball does not substantially change the model predictions in the independent condition. However, it makes an important change to the predictions in the sequential testimony condition, most dramatically in the maximizing model. The model predicts that the last informant will dissent only if she received a diagnostic ball. Since she does dissent, she most likely received a diagnostic ball and so the participant should side with the dissenter over the majority (a similar but somewhat more subtle effect occurs for the probability matching model). Finally, in the shared private data condition, the dissenter probably received a different colored ball than the two informants in the majority. This provides equal evidence for either urn.

\section*{Methods}

A total of 124 participants were recruited through Amazon Mechanical Turk. Participants were compensated \(\$ 0.25\) for their time. They were randomly assigned to one of three experimental groups: the independent condition \((n=41)\) or the
shared testimony ( \(n=41\) ), or shared private data \((n=42)\). No participants were dropped from the analysis.
Stimuli The stimuli were identical to those in Experiment 1, except for the urns shown. Instead of using opaque colored urns, the urns were replaced with a picture of a clear jar filled with a mix of red and blue balls. A single diagnostic ball (either white or black) was placed in each urn. Each urn was labeled either "Jar A" or "Jar B".
Procedure The procedure was the same as Experiment 1, except for the following changes. References to the "red urn" and the "blue urn" were replaced by references to "Jar A" and "Jar B". In all three conditions the last informant dissented from the previous informants, and supported the belief that the bag was filled from the other urn. In the shared private data condition, the first two informants received the same ball. The last informant received a different ball. At the end of the experiment the participant did not see their own ball and made their judgments based solely on the informants' testimonies. Responses were made on the same 11-point scale as in Experiment 1, changing the names of the urns appropriately.

\section*{Results and Discussion}

Ratings were rescaled as in Experiment 1. The mean rescaled ratings are shown in Figure 2(c). Participants sided with the majority testimony most in the independent testimony condition, second in the sequential testimony condition, and least in the shared private data condition. The means and order of the results are consistent with the probability matching model predictions, but not the maximizing model predictions. The probability matching model provides a good fit for the experimental data (Pearson's \(r=.94\) ).

We analyzed the effect of condition on participant responses using a one-way ANOVA. The effect of condition was significant \((F(2,54.3)=5.561, M S E=27.13, p<\) 0.005 ). We explored the differences between the conditions using planned t -tests. The difference between the independent testimony and sequential testimony conditions was significant (two sample t-test, \(t(80)=3.12, p<.005\) ) as well as the difference between the independent testimony and shared private data conditions (two sample t-test, \(t(81)=3.16, p<\) .001 ). The difference between the sequential testimony and shared private data conditions was not significant (two sample t-test, \(t(81)=0.22, p>.82\) ).

The difference between the independent testimony and sequential testimony conditions suggests the learning mechanism that participants use is sensitive to social information that is shared between informants. The difference between the shared private data condition and the independent testimony condition supports our conclusion from Experiment 1 that people are sensitive to non-social shared information.

Qualitatively, participants' performance resembles the probability matching model more than the maximizing model used in earlier work on information cascades. However, in both the sequential testimony and the shared private data con-
ditions participants sided with the majority slightly more than the probability matching model predicts, suggesting that even though individuals are able to utilize shared information that informants use to make their judgments, they may place more trust in the informants' testimony. This difference was significant in the Shared Private Data condition (one sample \(t\)-test, \(t(41)=2.867, p<.01)\), but not significant in the sequential testimony condition (one sample t-test, \(t(40)=.54, p>.05\) ).

\section*{General Discussion}

The goal of this research is to determine whether human social learning is based on a sophisticated strategy that appropriately takes into account dependencies in the behavior of other people. To answer this question, we developed a Bayesian model that indicates how patterns of dependency should affect social learning. The model makes clear predictions about two kinds of dependency - sequential testimony and shared private data - which we tested through two experiments. Experiment 1 showed that people are sensitive to shared private data, using a task that has been employed in previous experiments on information cascades. Experiment 2 showed that people are sensitive to sequential testimony, using a task that is more sensitive to this kind of dependency. However, in both experiments people's judgments were influenced by dependency less than they should have been.

Our results have implications for models of cultural evolution and information cascades. For models of cultural evolution, they offer insights into the mechanisms that underlie social learning, and suggest that patterns of dependency should be taken into account in contexts where agents might encounter dependent social cues. While models of information cascades typically assume sequential testimony, our results show that people are sufficiently sensitive to patterns of dependency that information cascades will be even more probable if it is assumed that informants provide independent testimony. In addition, the matching model provided a closer qualitative and quantitative fit to human performance than the maximizing model. This empirical evidence conflicts with the assumption that informants maximize their posterior used in previous work on information cascades (e.g. Bikhchandani et al., 1992) and helps explain some of the patterns of "errors" observed in the experiments by Anderson and Holt (1997).

Taken together, our findings suggest that human social learning mechanisms are fairly sophisticated. People do not just use simple rule-based imitation strategies. Instead they are able to integrate their own private information with informants' testimony, and take into account how each informant decided upon their testimony. This implies that human cultural evolution is not simply a result of individuals making a trade-off between acquiring their information socially or through trial-and-error learning, but is instead the result of complex decisions that draw on beliefs about informants' sources of information. When learning from testimony, learners are asking themselves the question: "and just how do you know that?"

Acknowledgments. This work was supported by grant number IIS-0845410 from the National Science Foundation (NSF), a NSF Graduate Research Fellowship, and a SURF/Rose Hills Fellowship.

\section*{References}

Anderson, L., \& Holt, C. (1997). Information cascades in the laboratory. The American economic review, 847-862.
Asch, S. E. (1956). Studies of independence and conformity: I. a minority of one against a unanimous majority1. Psychological Monographs: General and Applied, 70(9), 1-70.
Bikhchandani, S., Hirshleifer, D., \& Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. Journal of political Economy, 992-1026.
Boyd, R., \& Richerson, P. (1988). Culture and the evolutionary process. University of Chicago Press.
Boyd, R., \& Richerson, P. (2005). The origin and evolution of cultures. Oxford University Press, USA.
Buchsbaum, D., Gopnik, A., Griffiths, T. L., \& Shafto, P. (2011). Children's imitation of causal action sequences is influenced by statistical and pedagogical evidence. Cognition, 120(3), 331-340.
Carpenter, M., Call, J., \& Tomasello, M. (2005). Twelve- and 18-month-olds copy actions in terms of goals. Developmental Science, 8(1), F13-F20.
Cialdini, R., \& Goldstein, N. (2004). Social influence: Compliance and conformity. Annu. Rev. Psychol., 55, 591-621.
Corriveau, K. H., \& Harris, P. L. (2010). Preschoolers (sometimes) defer to the majority in making simple perceptual judgments. Developmental Psychology, 46(2), 437-445.
Corriveau, K. H., Meints, K., \& Harris, P. L. (2009). Early tracking of informant accuracy and inaccuracy. British Journal of Developmental Psychology, 27(2), 331-342.
Gergely, G., Bekkering, H., \& Kiraly, I. (2002). Rational imitation in preverbal infants. Nature, \(415,755\).
Jaswal, V. K. (2006). Preschoolers favor the creator's label when reasoning about an artifact's function. Cognition, 99, B83-B92.
Jaswal, V. K., \& Malone, L. S. (2007). Turning believers into skeptics: 3-year-olds' sensitivity to cues to speaker credibility. Journal of Cognition and Development, 8(3), 263-283.
Koenig, M., \& Harris, P. (2005). Preschoolers mistrust ignorant and inaccurate speakers. Child Development, 76(6), 1261-1277.
Laland, K. (2004). Social learning strategies. Learning \& behavior, 32(1), 4-14.
Luce, R. (2005). Individual choice behavior: A theoretical analysis. Dover publications.
Lyons, D., Young, A., \& Keil, F. (2007). The hidden structure of overimitation. Proceedings of the National Academy of Sciences, 104(50), 19751-19756.
McGuigan, N., Makinson, J., \& Whiten, A. (2011). From overimitation to super-copying: Adults imitate causally irrelevant aspects of tool use with higher fidelity than young children. British Journal of Psychology, 102(1), 1-18.
Meltzoff, A. N. (1995). Understanding the intentions of others: Re-enactment of intended acts by 18 -month-old children. Developmental Psychology, 31, 838-850.
Nielsen, M., \& Tomaselli, K. (2010). Overimitation in kalahari bushman children and the origins of human cultural cognition. Psychological Science, 21(5), 729-736.
Perreault, C., Moya, C., \& Boyd, R. (2012). A bayesian approach to the evolution of social learning. Evolution and Human Behavior.
Rogers, A. (1988). Does biology constrain culture? American Anthropologist, 90(4), 819-831.
Shepard, R. (1958). Stimulus and response generalization: Tests of a model relating generalization to distance in psychological space. Journal of Experimental Psychology, 55(6), 509.
Sobel, D. M., \& Corriveau, K. H. (2010). Children monitor individuals' expertise for word learning. Child Development.
Tenney, E. R., Small, J. E., Kondrad, R. L., Jaswal, V. K., \& Spellman, B. A. (2011). Accuracy, confidence, and calibration: How young children and adults assess credibility. Developmental Psychology, 47(4), 1065-1077.
Vulkan, N. (2000). An economist's perspective on probability matching. Journal of Economic Surveys, 14, 101-118.

\title{
A quantum probability perspective on the nature of psychological uncertainty
}

\author{
Lee C. White (l.c.white.517813@swansea.ac.uk) \\ Department of Psychology, University of Swansea, Singleton Park, Swansea, SA2 8PP UK.
}

Emmanuel M. Pothos (e.m.pothos@gmail.com)
Department of Psychology, City University London London, EC1V 0HB UK.

\author{
Jerome R. Busemeyer (jbusemey @indiana.edu) \\ Department of Psychological and Brain Sciences, Indiana University, Bloomington 47468 Indiana, USA.
}

\begin{abstract}
Making a choice between alternatives can influence our subsequent evaluation of the selected option (e.g. Sharot, Velasquez \& Dolan, 2010). Thus, in resolving psychological uncertainty, the act of making a judgment itself appears to have a constructive role in subsequent related decisions. This study focuses on emotional ambivalence and the development of affective evaluations over two stages, such that (just) making an intermediate evaluation in the first stage is shown to influence the overall affective evaluation in the second stage. Models based on classical probability theory, which assume that an intermediate evaluation simply reads off an existing internal state, cannot accommodate this result in a natural way. An explanation is offered with a quantum probability model, which, under specific circumstances, requires the measurement of an internal state to have a constructive role. The predictions of the quantum probability model were supported by the empirical results.
\end{abstract}

Keywords: Quantum probability; Interference effects; Affective uncertainty.

\section*{Introduction}

One basic fact about cognition is that it reflects uncertainty. In fact cognition appears to involve several kinds of uncertainty. As well as uncertainty regarding future events, there is uncertainty about internal states, an inevitable consequence of the fact that life events are often agglomerations of pleasant and unpleasant components. For example, consider 'emotional ambivalence', the apparent ability of the cognitive system to concurrently represent positive and negative affect. Emotional ambivalence is reflected in e.g., students' thoughts about graduation day or advertisements with mixed emotional appeals (Larsen, McGraw \& Cacioppo, 2001; Williams \& Aaker, 2002).

Understanding how the cognitive system resolves affective uncertainty presents challenges (e.g. Brehm \& Miron, 2006). For example, is positive and negative affect experienced sequentially or simultaneously? What happens when people are asked to make a judgment about their affective state whilst experiencing affective uncertainty? Does this judgment resolve uncertainty or does the act of
making the judgment itself influence their affect in the same way that choice has been shown to have a constructive influence on preference (e.g. Sharot, Velasquez \& Dolan, 2010)? Our objective in this paper is to propose an ambitious new perspective on this question, based on quantum probability (QP) theory (note that in this work by QP theory we simply mean the rules for how to assign probabilities to events from quantum theory; for more specific proposals see Aerts, 2009, or Atmanspacher, Romer \& Wallach, 2006).

We can acquire some preliminary intuition from models for response times in choice problems, such as random walk models (e.g., Ashby, 2000; Busemeyer \& Townsend, 1993; Ratcliff \& Smith, 2004; Usher \& McClelland, 2001). In this influential class of models, discriminating between two options involves an accumulation of evidence, so that, on every step, the weight for a particular option is increased. Crucially, at any time point, the system is assumed to be in a specific state. This state may reflect large or little weight for a particular option, but, regardless, it has to be at a specific state. Classical approaches must assume that the system is always at a particular state, even if knowledge of this state is uncertain. Such an assumption seems straightforward. How else could we build a model?

Yet, there is an alternative, intriguing possibility, which emerges from the recent uses of QP theory in cognitive modeling (Busemeyer \& Bruza, 2012; Pothos \& Busemeyer, in press). QP theory is a framework for assigning probabilities to observables and, therefore, potentially relevant wherever there is a need to formalize uncertainty. QP cognitive models often have the same intentions (Griffiths et al., 2010; Oaksford \& Chater, 2007) as classical probability models. But, classical and QP frameworks are founded on different axioms. QP models incorporate certain unique features, such as superposition and the capacity for interference, and there has been growing interest in exploring the relevance of such features for cognitive modeling (e.g., Aerts, 2009; Atmanspacher, Filk \& Romer, 2004; Blutner, 2009; Busemeyer, Pothos, Franco \& Trueblood, 2011; Khrennikov, 2010; Pothos \& Busemeyer, 2009; Wang et al, in press).

In QP theory, a superposition state has amplitude (weight) across more than one possibility. Suppose we are interested in representing whether a stimulus induces a positive or negative affect. Classically, the situation is straightforward: if we are uncertain about a person's state, we assign probabilities to the person having a positive or negative affect. Perhaps there is a dynamic process which evolves (reshuffles) the person's state, until a final state is reached. But, the person is always assumed to be at a particular state. The situation with a QP approach is markedly different: as long as there is weight for both possibilities, the person is in a superposition of possibilities, and it is impossible to interpret the person as being at a particular state, rather, there is a potentiality for each possibility. That this has to be the case is not obvious and it is the result of the famous Kochen-Specker theorem in QP theory. The key implication, which is fundamental to QP theory, is that a transition from a superposition to a definite state must have a constructive role.

The QP perspective enables a simple, but surprising, empirical prediction. The relevant difference between the classical and the QP approach is that in the former the system is always assumed to be at a particular state, while in the latter there is a distinction between particular states and superposition states. Therefore, consider a situation in which affective evaluation is developed over two steps, such that each step involves a stimulus presentation. Classically, it should not matter whether the person is asked to provide an affective evaluation just after the second step, or after the first step as well. In the latter case, the intermediate evaluation would simply 'read off' the existing state and so this should not affect the overall outcome of the affective evaluation. However, in the QP model, an action of affective evaluation (a "measurement") can have a profound impact on the state of the system and, therefore, the intermediate evaluation influences the eventual outcome of the second evaluation. Note that a classical model could incorporate the possibility that an evaluation (or rating etc.) has a constructive role, but this could only be done with additional assumptions, which are not part of classical probability theory.

In the current study stimuli were hypothetical advertisements, appearing as static images. In the positivenegative (PN) condition, a single positively valenced 'positive image' was presented, followed by a mixed advert, including the same positive image presented concurrently with a 'negative image', and vice versa for the negativepositive condition (NP). In the 'single rating' condition, participants viewed the single image advert and then provided an overall affective evaluation for the mixed advert. In the 'double rating' condition the same participants provided an intermediate rating to the single advert, before viewing the mixed advert and rating it. Note that the relative order of the images is likely to impact on the final evaluation. Moore (2002) demonstrated order effects in Gallup poll questionnaires (see also Bergus et al., 1998; McKenzie, Lee, \& Chen, 2002). Relatedly, Vlaev et al.
(2009) argued that pain perception depends on recent pain experiences.

Hogarth \& Einhorn's (1992) research on order effects in belief updating has obvious similarities with the current research, although as we shall see there are also some important differences. Their belief-adjustment model describes order effects as arising from the interaction of key variables including the complexity of stimuli, length of the sequence of items and whether participants respond using a Step-by-Step (SbS) procedure, where they report their judgment after integrating each piece of evidence, or an End-of-Sequence (EoS) procedure where they report their judgment only after they have viewed all stimuli in the sequence.

In a review of previous research as well as their own experiments they argue that in the case of short sequences (i.e. between 2 and 12 items) requiring simple judgments (i.e. a single item for each stimulus in the sequence) the majority of studies employing a SbS procedure result in a recency effect whereas the majority of studies using an EoS procedure result in a primacy effect. The belief-adjustment model describes a sequential anchoring-and-adjustment process in which the current belief is adjusted by the subsequent pieces of evidence.

Although there are similarities between our research and that described by Hogarth \& Einhorn, there are also several important differences. Most results described in Hogarth \& Einhorn's paper, including their own experiments, involved 3 or more items in a sequence, whereas the current experiment is concerned with two items. The effects they described were concerned with items that were related to each other whereas in the current experiment items were chosen to be unrelated. In our experiment participants are required to evaluate each individual item in its own right whereas in the studies described in the Hogarth \& Einhorn paper each subsequent piece of evidence is evaluated with respect to an overall judgment about a person or object (e.g. trait adjectives used to make social judgments about someone's "likeableness").

However, our objective here is not to demonstrate order effects in affective evaluation, but rather to understand the potential role of an intermediate evaluation on the final one. In Hogarth \& Einhorn's terms, whether there is a difference between evaluations produced using EoS and SbS procedures for the same items viewed in the same order. For this, we need to consider some elementary QP principles.

The states of a system are represented by vectors, \(\psi\), within a multidimensional space. Different subspaces represent possibilities for \(\psi\) and a projection of \(\psi\) onto a subspace involves laying \(\psi\) onto the subspace. The squared length of this projection gives the probability that \(\psi\) is the possibility represented by the subspace (cf. Sloman, 1993). Finally, the angles between subspaces correspond to the relation between the corresponding possibilities. This can be easily understood by noting that, if a state vector is
consistent with one possibility, then we want it to have large projections on related possibilities.


Figure 1: An illustration of how the state for the mixed image is created in the PN condition.

In Figure 1 we represent various possibilities for the affective state of the participant in the PN condition. The positive and negative affect subspaces correspond to purely positive and negative affect respectively; they are orthogonal, since a state in the positive affect subspace must have a zero projection onto the negative affect one. The positive and negative image subspaces represent the affective impact of seeing a positive and negative image during the experiment, respectively. These two subspaces are also nearly orthogonal since the images were chosen to be unrelated. Note that, in this example, the positive image subspace is close to the positive affect one, since perceiving a positive image is more likely to lead to a positive affect (the state vector created as a result of perceiving the positive image has a large projection to the positive affect subspace).

In the single rating condition, after perceiving the positive image, the state vector is aligned with the positive image subspace. The impact of introducing the negative image is represented by a rotation of the state vector, denoted as U , which leads to the state labeled as 'PN single rating'. The subsequent projection to the negative affect subspace is a measure of how negative we expect the resulting rating to be (the thick line along the bad feeling subspace). Specifically, the squared length of the projection of the mixed image state onto the negative affect subspace determines the probability of a negative rating; it is natural to assume that the higher this probability, the more negative the rating for the mixed image.

In the double rating condition, after perceiving the positive image, the intermediate rating is assumed to lead to a transition to the positive affect subspace. This is the critical difference between the single rating and the double rating conditions, which can lead to a prediction about behavioral differences depending on the presence of the intermediate rating or not. In the double rating case, once
the rating for the first image is completed, as before, the impact of introducing the negative image leads to the same angle rotation \({ }^{1}\). But, in this case the starting state is different (it is aligned with the positive affect subspace), therefore the resulting state corresponding to the mixed image is different too (labeled as 'PN double rating'). The resulting state is now closer to the negative affect subspace, which predicts a more negative rating.

Finally, in both the single and double rating PN conditions the final state is assumed to be closer to the negative affect subspace, than to the positive affect one, to reflect a recency effect in the importance of the negative, final image on the affective state (Trueblood \& Busemeyer, 2011).

For the NP case (Figure 2), we are led to the converse prediction, namely that the final evaluation in the double rating condition will be more positive than the one in the single rating condition. Thus, a quantum approach predicts a striking interaction in the final affective evaluation in the PN vs. NP conditions, depending on whether single ratings or double ratings are solicited, only on the basis of the role of measurement in QP theory.


Figure 2: An illustration of how the state for the mixed image is created in the NP condition.

\footnotetext{
\({ }^{1}\) A consistency consideration determines the direction of rotation. In the double rating PN condition, in transforming the state vector from the positive affect subspace to the state corresponding to the mixed image, an intermediate state cannot be aligned with the positive image subspace.
}

\section*{Method}

\section*{Participants}

Fifty-four Swansea University students participated in the experiment for course credit ( 45 women, 9 men, average age 21.74 years).

\section*{Stimuli}

Realistic-looking adverts were created, so that the positive and negative versions would make sense together. Different advertised products were used for the PN condition (insurance; see Figure 3 for example) and the NP one (smartphone), so as to avoid interference between conditions. For the PN set there were three positive images individually presented, and three mixed images with each of the positive images joined with a negative one, and analogously for the NP set. All images were piloted to confirm their intended affective response and their unrelatedness. The images were randomly presented with 24 adverts for a camera, which were included for a different study and acted as fillers in the current experiment.


Figure 3: Sample advert used in PN condition and procedure for presentation of single and double rated adverts.

\section*{Procedure}

Participants first completed a six-item current mood questionnaire. They were then told that they would see several adverts and that for each advert, when asked, they should answer the question 'how does this advert make you feel?', responding on a nine-point scale, with anchors "very unhappy/very happy". Each trial involved the presentation of a single image, followed by a request for rating (double rating condition) or not (single rating condition), followed by the mixed image and a final request for rating (Figure 3). Trials were organized into two blocks. One block contained the six single rating PN adverts and six double rating NP ones, together with 12 filler adverts (which were also rated). The other block contained the same adverts, but switching the requirement for single vs. double rating. Block order and trial order within blocks were randomized across participants.

\section*{Results}

As we had previously established the valence of the images (with the pilot study), we excluded four participants because their ratings for the single image adverts were over one standard deviation below (for positive adverts) or above (for negative adverts) the mean.

We conducted a two (advert type condition: PN vs. NP) \(\times\) two (rating condition: single vs. double) repeated measures ANOVA on the participants' ratings for the mixed adverts. There was a main effect of advert type \((F(1,49)=7.98\), \(p=.007\) ), but not of rating condition \((F(1,49)=0.04\), n.s. \()\). Crucially, the advert type x rating condition interaction was significant \((F(1,49)=10.96, \mathrm{p}=.002)\). Paired samples t-tests further showed that double rated PN adverts ( \(M=4.04\), \(S D=1.17\) ) were rated significantly lower (i.e. unhappier), compared to single rated adverts \((M=4.34, S D=1.43\); \(t(49)=2.18, p=.02\), two tailed; \(d=.31)\). For the NP adverts, double rated adverts ( \(M=4.94, S D=1.21\) ) were rated significantly higher (i.e. happier) than single rated adverts ( \(M=4.60, S D=1.22 ; t(49)=-2.39, p=.01\), two tailed; \(d=.34\) ).


Figure 4: Mean participant ratings of single and double rated PN and NP adverts (error bars represent standard deviations).

We also considered a plausible alternative explanation, that it is the availability of a rating after the first advert, rather than the act of measurement as such, drives the observed result, a possibility consistent with anchoring (Tversky \& Kahneman, 1974). In the double rating condition, the more readily accessible rating for the first advert is perhaps a reference point, against which the rating for the second advert is computed. However, there was no evidence for such an anchoring effect, as there were low, non-significant correlations between participant ratings for the first and second advert in the PN ( \(r=.26\), n.s.) and NP ( \(r=.18\), n.s) conditions.

\section*{Discussion}

The results of the experiment confirmed our predictions. For the NP mixed adverts, an intermediate rating led to a higher final evaluation and, for the PN mixed adverts, an intermediate rating led to a lower final evaluation. Such a finding is difficult to reconcile with a classical probability perspective, without additional assumptions, since an intermediate rating should simply read off an existing internal state. However, the QP approach can predict a change of the state of the system, as a result of a measurement, and so is able to predict how an intermediate evaluation could affect the final one. Our finding resonates with the uncertainty intensification hypothesis (Bar-Anan, Wilson \& Gilbert, 2009), according to which uncertainty about an event will prolong and intensify how people feel about it. In single rating trials the impact of the first image on the final evaluation is higher, than in the double rating ones (e.g., in PN trials with a single rating, the final evaluation is more positive than with double rating). Perhaps uncertainty about the internal state, after viewing the first image, intensifies its effect on the final rating, but in the double rating condition, reducing this uncertainty results in a greater impact on the final rating from the second image (that is, the component of the mixed image which is novel). The QP model can be seen as a formalization of such ideas.

We can elaborate on the intuition of why the QP approach works. The critical point concerns the state prior to the second, final rating. In, e.g., the PN condition, the single rating case, the state prior to the second rating reflects the impact of seeing the mixed image, that is, the original positive image, together with the new negative one. The impact of this mixed imaged would be slightly weighted in favor of the negative image, since this is shown last. In the double rating case, the intermediate evaluation (which produces a result of positive affect) can be understood as a process of abstracting away some information from the first image, but emphasizing its positive affective qualities. Therefore, this makes the introduction of the negative image produce a more contrasting affective impression. In other words, accepting that the first image is positive, creates a 'perspective' of positive affect for processing the subsequent negative image, which makes it look, well, more negative. The result of the intermediate rating is thus a larger negative final rating in the PN condition and exactly vice versa in the NP one.

The intuition that a measurement is not simply a record of an existing state, but rather it creates a state, is not alien to psychology. Notably, Shafer and Tversky (1985, p.337) proposed "A probability judgment depends not just on the evidence on which it is based, but also on the process of exploring that evidence." Quantum theory provides a formal framework within which to express this intuition.

There are plenty of possible extensions to the present work. First, it would be worth exploring more the putative role of reference points in such experiments, perhaps with the provision of an external rating. Second, we would like to explore this paradigm with changes in procedure and
materials, so as to establish its robustness. Finally, there have been recent interesting analyses on the putative constructive role of decisions (e.g. Sharot et al., 2010) and it would be worth exploring further the conceptual links between these ideas and the quantum approach.

\section*{References}

Aerts, D. (2009). Quantum structure in cognition. Journal of Mathematical Psychology, 53(5), 314-348.
Ashby, F. G. (2000). A stochastic version of general recognition theory. Journal of Mathematical Psychology, 44(2), 310-329.
Atmanspacher, H., Filk, T. \& Romer, H. (2004). Quantum Zeno features of bistable perception. Biological Cybernetics, 90(1), 33-40.
Atmanspacher, H., Romer, H., \& Wallach, H. (2006). Weak quantum theory: formal framework and selected applications. Weak quantum theory: complementarity and entanglement in physics and beyond. Foundations of Physics, 32, 379-406.
Bar-Anan, Y., Wilson, T.D. \& Gilbert, D.T. (2009). The feeling of uncertainty intensifies affective reactions. Emotion, 9(1), 123-127.
Bergus, G. R., Chapman, G. B., Levy, B. T., Ely, J. W., \& Oppliger, R. A. (1998). Clinical diagnosis and the order information. Medical Decision Making, 18(4), 412-417.
Blutner, R. (2009). Concepts and bounded rationality: An application of Niestegge's approach to conditional quantum probabilities. In L. Accardi, G. Adenier, C. Fuchs, G. Jaeger, A.Y. Khrennikov, J. Larsson, \& S. Stenholm (Eds.), Foundations of probability and physics5 (Vol. 1101).
Brehm, J.W. \& Miron, A.M. (2006). Can the simultaneous experience of opposing emotions really occur? Motivation and Emotion, 30(1), 13-30.
Busemeyer, J. R. \& Townsend, J. T. (1993). Decision field theory: A dynamic cognition approach to decision making. Psychological Review, 100(3), 432-459.
Busemeyer, J. R. \& Bruza, P. (2011). Quantum models of cognition and decision making. Cambridge, UK: Cambridge University Press.
Busemeyer, J. R., Pothos, E., Franco, R., \& Trueblood, J. S. (2011). A quantum theoretical explanation for probability judgment errors. Psychological Review, 118(2), 193-218.
Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., \& Tenenbaum, J. B. (2010). Probabilistic models of cognition: exploring representations and inductive biases. Trends in Cognitive Sciences, 14(8), 357-364.
Hogarth, R.M., \& Einhorn, H.J. (1992). Order effects in belief updating: The belief-adjustment model. Cognitive Psychology, 24, 1-55.
Khrennikov, A. Y. (2010). Ubiquitous quantum structure: From psychology to finance. Berlin: Springer-Verlag.
Larsen, J.T., McGraw, A.P. \& Cacioppo, J.T. (2001). Can people feel happy and sad at the same time? Journal of Personality and Social Psychology, 81(4), 684-696.

McKenzie, C. R. M., Lee, S. M., \& Chen, K. K. (2002).When negative evidence increases confidence: change in belief after hearing two sides of a dispute. Journal of Behavioral Decision Making, 15(1), 1-18.
Moore, D. W. (2002). Measuring new types of questionorder effects. Public Opinion Quarterly, 66(1), 80-91.
Oaksford, M. \& Chater, N. (2007). Bayesian rationality: The probabilistic approach to human reasoning. Oxford: Oxford University Press.
Pothos, E. M. \& Busemeyer, J. R. (2009). A quantum probability explanation for violations of 'rational' decision theory. Proceedings of the Royal Society B: Biological Sciences, 276(1665), 2171-2178.
Pothos, E. M. \& Busemeyer, J. R. (in press). Can quantum probability provide a new direction for cognitive modeling? Behavioral \& Brain Sciences.
Ratcliff, R. \& Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. Psychological Review, 111(2), 333-367.
Shafer, G., \& Tversky, A. (1985). Languages and designs for probability judgment, Cognitive Science, 9, 309-339.
Sharot, T., Velasquez, C.M., \& Dolan, R.J. (2010). Do decisions shape preference? : Evidence from blind choice, Psychological Science, 21(9) 1231-1235.
Sloman, S. A. (1993). Feature-based induction. Cognitive Psychology, 25(2), 231-280.
Trueblood, J. S. \& Busemeyer, J. R. (2011). A quantum probability account of order effects in inference. Cognitive Science, 35(8), 1518-1552.
Tversky, A., \& Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185, 11241130.

Usher, M. \& McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. Psychological Review, 108(3), 550-592.
Wang, Z., Busemeyer, J. R., Atmanspacher, H., \& Pothos, E. (in press). The potential of using quantum theory to build models of cognition. Topics in Cognitive Science.
Williams, P. \& Aaker, J.L. (2002). Can mixed emotions peacefully coexist? Journal of Consumer Research, 28(4), 636-649.
Vlaev, I., Seymour, B., Dolan, R.J. \& Chater, N. (2009). The price of pain and the value of suffering. Psychological Science, 20(3), 309-317.

\title{
Learning nonadjacent dependencies in thought, language, and action: Not so hard after all...
}

\author{
Jon A. Willits (jwillits@indiana.edu) \\ Indiana University, Department of Psychological and Brain Sciences, \(1101 \mathrm{E} .10^{\text {th }} \mathrm{St}\). Bloomington, IN 47405 USA
}

\begin{abstract}
Learning to represent hierarchical structure and its nonadjacent dependencies (NDs) is thought to be difficult. I present three simulations of ND learning using a simple recurrent network (SRN). In Simulation 1, I show that the model can learn distance-invariant representations of nonadjacent dependencies. In Simulation 2, I show that purely localist SRNs can learn abstract rule-like relationships. In Simulation 3, I show that SRNs exhibit facilitated learning when there are correlated perceptual and semantic cues to the structure (just as people do). Together, these simulations show that (contrary to previous claims) SRNs are capable of learning abstract and rule-like nonadjacent dependencies, and show critical perceptual- and semantics-syntax interactions during learning. The studies refute the claim that neural networks and other associative models are fundamentally incapable of representing hierarchical structure, and show how recurrent networks can provide insight about principles underlying human learning and the representation of hierarchical structure.
\end{abstract}

Keywords: hierarchical structure; recurrent connectionist networks; nonadjacent dependencies

\section*{Background}

Human concepts, languages, goals, and patterns of action are all describable in terms of complex hierarchical structures, but our experience of them as inputs, and our production of them as outputs, is often arranged in linear strings that unfold over time. A necessary consequence of this transformation of complex structure into linear strings is that most human knowledge involves many nonadjacent dependencies, where one element predicts another element, but at a distance. These nonadjacent dependencies, whether in thought, language, or action, enormously expand the computational complexity of representing the structure of the world.

In several subfields of cognitive science, difficulty learning and representing nonadjacent dependencies has generated considerable theoretical controversy. In linguistics, the limitation of simple associative structures has been a cornerstone of arguments for abstract syntactic structures (Chomsky, 1957). In cognitive psychology, researchers argued that associative mechanisms cannot learn the vast range of nonadjacent dependencies in the world, and thus rule-
based representations are necessary for human cognition (Bever et al., 1968). In early artificial intelligence, arguments about the limitations of associative systems led to a focus on symbolic, rulebased systems (Newell \& Simon, 1961).
However, recent research has questioned the need for rule-based representations of nonadjacent structure. A number of studies have demonstrated or modeled simple learning of nonadjacent structure in memory (Cleeremans \& McClelland, 1991), goals and event structure (Botvinick \& Plaut, 2004; visual sequences (Fiser \& Aslin, 2002), and artificial grammars using linguistic stimuli (Gomez, 2002; Newport \& Aslin, 2004). These results have changed the nature of the debate concerning the extent to which knowledge of nonadjacent dependencies requires a rule-based or an association-based explanation. Although there are many specific examples of learning or failing to learn in particular situations, what is lacking is a general account of nonadjacent dependency learning. As a result, the many subfields of cognitive science (such as linguistics, cognitive psychology, and artificial intelligence) continue working on the problem separately, without a clear theory or explanation for some of the most foundational human behaviors.
The current work aims to make progress toward a general account by examining whether a fairly simple neural network model, the simple recurrent network (SRN; Elman, 1990) can provide a general model of nonadjacent dependency learning. An SRN was used because previous research (Botvinick \& Plaut, 2006; Cleeremans \& McClelland, 1991; Elman, 1991) suggests that SRNs and other recurrent networks are capable of learning nonadjacent structure. However, there is controversy about whether they can serve as general solution for all cases, especially those involving abstract, rule-like relationships (Marcus, 2000) or complex interactions between structure and meaning (Fodor \& Pylyshyn, 1988).

In the service of testing the viability of SRNs, the current work had two distinct sub-goals. First, to be a general model of nonadjacent dependency learning, SRNs ought to be able to learn nonadjacent dependencies of the types that exist in the natural
world. This includes abstract, rule-like nonadjacent dependencies, such as learning "distance-invariant" representations (for example, learning the link between the and a noun, independent of how many adjectives come between them). Second, SRNs ought to capture behavioral phenomena observed in laboratory experiments, such as facilitated learning in the presence of perceptual (Newport \& Aslin, 2004) and semantic (Willits, Lany, \& Saffran, 2013) cues. Close analysis of model behavior can then shed light on the bases of the empirical effects. The following three studies test SRNs' abilities to satisfy these criteria.

\section*{General Methodology}

The three studies shared three core features common in connectionist-modeling approaches (Rumelhart \& McClelland, 1986). First, all simulations used sets of interconnected units and weights specifying how strongly each unit was connected to each other unit. The units in the model were divided into an input group, used to specify the input stimulus in each sequence; an output group, used to specify the output response (which also served as a prediction about the next item in the sequence); and a hidden group that mediated between the input and output groups. Second, the models featured recurrent connectivity, allowing the model to feed back information about its own previous internal state in ways critical to forming internal representations of sequential structure. Third, the models all made use of weight-based encoding, where the network's knowledge was encoded in the weighted connections between units.

The goal of the network was to learn a set of weights such that, for any given input, the model's weights led to activation in the output layer that was a correct prediction of the next item in the sequence. During training, a model was given an input, its output activation was treated as a prediction of what the next input would be. This prediction was compared to the target output, and divergence error was calculated across each unit and was used to adjust the weights of the model, using a version of recurrent backpropagation through time. For each simulation, 30 different randomly initialized models were trained. Each model was trained until it reached a predetermined level of overall error, corresponding to optimal prediction performance in the task. The critical test in each simulation was the relative rate of learning across the different conditions in that study.

\section*{Study 1: Distance Invariance}

In experiments on nonadjacent dependencies using artificial grammars, the distance between dependent
items is usually fixed, with one intervening item separating dependent items. However, in many realworld cases (such as the distance between nonadjacently related events in the world, or words in language) the distance between dependent items varies. In fact, learning a "distance-invariant" representation of a nonadjacent dependency has been considered a critical phenomenon, proving the need for a rule-based mechanism.

In Simulation 1, I attempted to train an SRN to learn distance-invariant representations of nonadjacent dependencies by exposing them to the same nonadjacent dependency at multiple spans of distance between the related items. A second issue of interest was whether SRNs would show facilitation in learning longer-distance dependencies if they also had experience with the dependency at a shorter distance, a learning effect that has been demonstrated in both infants and adults (Lany \& Gomez, 2008).

\section*{Stimuli and Design}

The models in Study 1 were trained on sequences where the first element (hereafter the \(A\) item) perfectly predicted the last element in each sequence (hereafter the \(B \mathrm{item}\) ), with the sequences having a number of items (hereafter the \(X\) items) intervening between them. The sequences were of lengths 2 to 5 , resulting in distances between the \(A\) and \(B\) items spanning from zero (adjacent dependencies) to three. There were two \(A B\) pairs ( \(\mathrm{A}_{1} \& \mathrm{~B}_{1}, \mathrm{~A}_{2} \& \mathrm{~B}_{2}\) ) and six possible intervening X -items \(\left(\mathrm{X}_{1} \ldots \mathrm{X}_{6}\right)\). The x -items were distributed across trials such that they provided zero predictive value for which \(B\) would occur. The only way to predict the correct \(B\left(\mathrm{~B}_{1}\right.\) or \(\left.\mathrm{B}_{2}\right)\) was to have stored which \(A\left(\mathrm{~A}_{1}\right.\) or \(\left.\mathrm{A}_{2}\right)\) initiated the sequence. The full set of stimuli used in Study 1 is shown in Table 1.

Thirty different networks (starting from different randomly initialized weights) were trained in each of six different training conditions: (1) only Span 0 trials; (2) only Span 1 trials; (3) only Span 2 trials; (4) only Span 3 trials; (5) a mixture of all Span trials; (6) a mixture of all Span trials except Span 3.

Over the course of training, networks from all six conditions were tested on stimuli from all Span conditions (without updating the network weights during those test trials), to assess the network's performance on strings of various spans. Networks were compared at points where they had experienced the same number of trials, controlling for the amount of experience the networks had with each \(A B\) pair.

\section*{Network Architecture}

The network had 10 input and output units (one for
each \(A, B\), and \(X\) ) and 25 hidden units. A simplification of the network architecture is shown in Figure 1.

Table 1. Stimulus inputs used in Study 1.
\begin{tabular}{|c|c|c|}
\hline Span 0 & \multicolumn{2}{|l|}{Span 1} \\
\hline \(\mathrm{A}_{1} \mathrm{~B}_{1}\) & \multicolumn{2}{|l|}{\(\mathrm{A}_{1} \mathrm{x}_{1} \mathrm{~B}_{1} \quad \mathrm{~A}_{2} \mathrm{x}_{1} \mathrm{~B}_{2}\)} \\
\hline \[
\mathrm{A}_{2} \mathrm{~B}_{2}
\] & \multicolumn{2}{|l|}{\begin{tabular}{l}
\[
\mathrm{A}_{1} \mathrm{x}_{2} \mathrm{~B}_{1} \quad \mathrm{~A}_{2} \mathrm{x}_{2} \mathrm{~B}_{2}
\] \\
Span 3
\end{tabular}} \\
\hline \(\mathrm{A}_{1} \mathrm{x}_{1} \mathrm{x}_{3} \mathrm{~B}_{1}\) & \(\mathrm{A}_{1} \mathrm{x}_{1} \mathrm{X}_{3} \mathrm{x}_{5} \mathrm{~B}_{1}\) & \(\mathrm{A}_{2} \mathrm{x}_{1} \mathrm{X}_{3} \mathrm{x}_{5} \mathrm{~B}_{2}\) \\
\hline \(\mathrm{A}_{1} \mathrm{X}_{1} \mathrm{X}_{4} \mathrm{~B}_{1}\) & \(\mathrm{A}_{1} \mathrm{x}_{1} \mathrm{X}_{4} \mathrm{x}_{5} \mathrm{~B}_{1}\) & \(\mathrm{A}_{2} \mathrm{X}_{1} \mathrm{X}_{4} \mathrm{x}_{5} \mathrm{~B}_{2}\) \\
\hline \(\mathrm{A}_{1} \mathrm{X}_{2} \mathrm{x}_{3} \mathrm{~B}_{1}\) & \(\mathrm{A}_{1} \mathrm{x}_{2} \mathrm{x}_{3} \mathrm{x}_{5} \mathrm{~B}_{1}\) & \(\mathrm{A}_{2} \mathrm{X}_{2} \mathrm{x}_{3} \mathrm{x}_{5} \mathrm{~B}_{2}\) \\
\hline \(\mathrm{A}_{1} \mathrm{x}_{2} \mathrm{x}_{4} \mathrm{~B}_{1}\) & \(\mathrm{A}_{1} \mathrm{x}_{2} \mathrm{X}_{4} \mathrm{x}_{5} \mathrm{~B}_{1}\) & \(\mathrm{A}_{2} \mathrm{x}_{2} \mathrm{X}_{4} \mathrm{x}_{5} \mathrm{~B}_{2}\) \\
\hline \(\mathrm{A}_{2} \mathrm{x}_{1} \mathrm{x}_{3} \mathrm{~B}_{2}\) & \(\mathrm{A}_{1} \mathrm{x}_{1} \mathrm{X}_{3} \mathrm{x}_{6} \mathrm{~B}_{1}\) & \(\mathrm{A}_{2} \mathrm{x}_{1} \mathrm{x}_{3} \mathrm{x}_{6} \mathrm{~B}_{2}\) \\
\hline \(\mathrm{A}_{2} \mathrm{x}_{1} \mathrm{x}_{4} \mathrm{~B}_{2}\) & \(\mathrm{A}_{1} \mathrm{x}_{1} \mathrm{X}_{4} \mathrm{x}_{6} \mathrm{~B}_{1}\) & \(\mathrm{A}_{2} \mathrm{x}_{1} \mathrm{x}_{4} \mathrm{x}_{6} \mathrm{~B}_{2}\) \\
\hline \(\mathrm{A}_{2} \mathrm{x}_{2} \mathrm{x}_{3} \mathrm{~B}_{2}\) & \(\mathrm{A}_{1} \mathrm{x}_{2} \mathrm{x}_{3} \mathrm{x}_{6} \mathrm{~B}_{1}\) & \(A_{2} \mathrm{x}_{2} \mathrm{x}_{3} \mathrm{x}_{6} \mathrm{~B}_{2}\) \\
\hline \(\mathrm{A}_{2} \mathrm{X}_{2} \mathrm{x}_{4} \mathrm{~B}_{2}\) & \(\mathrm{A}_{1} \mathrm{X}_{2} \mathrm{X}_{4} \mathrm{X}_{6} \mathrm{~B}_{1}\) & \(\mathrm{A}_{2} \mathrm{X}_{2} \mathrm{X}_{4} \mathrm{x}_{6} \mathrm{~B}_{2}\) \\
\hline
\end{tabular}


Figure 1. A simplified depiction of the network architecture used in Study 1. The actual model had \(8 X\)-units \(\left(\mathrm{X}_{1} \ldots \mathrm{X}_{8}\right)\) and 25 units in the hidden layer.

\section*{Hypotheses}

Three main hypotheses were under investigation. First, do networks trained on longer-distance dependencies (bigger Spans) take longer to learn the dependency, as people do? Second, do networks trained in more variable conditions (Conditions 5 \& 6) learn more slowly due to increased variability and noise? Or do they, like people (e.g. Lany \& Gomez, 2008) show facilitated learning of more distant dependencies due to experience with shorter dependencies? Third, are SRNs capable of learning a distance-invariant representation? Specifically, do the networks that are trained only on Spans of 0,1 , and 2, predict the correct B item on Span 3 trials, even though they have never before experienced the dependency at that distance?

\section*{Results \& Discussion}

Figure 2 shows the average SRN performance predicting the correct B (the network's activation level for the correct B output, on X trials) for networks
trained on only a single Span, when tested on the same Span. Networks showed strong effect of taking longer to learn, as the distance between the dependent items increased. Figure 3 shows the average performance on items of Span3 distance, for networks (1) trained on Span3, compared to (2) networks trained on a mixture of all the spans (SpanX) and (3) to networks trained on all the spans except Span3 (SpanX-3). At the earliest stages of training (trials \(0-1000\) ), the networks that experienced more variability showed slight decrements in performance on Span3 test items, relative to networks trained on Span3 alone. However, at later stages of training, both SpanX and SpanX-3 networks outperformed the Span3 network on Span3 items.


Figure 2. Average SRN performance for networks trained on a single span between nonadjacently dependent items, when tested on items of the same span. The \(y\)-axis is the network's softmax activation level of the correct B unit, when the network was presented with the preceding X item.


Figure 3. Average SRN performance for networks trained on Span 3, a mixture of all Span conditions (SpanX), or all Span conditions except Span3 (SpanX-3).
Thus, in Study 1 I show that SRNs display three critical features of human learning: (1) they show increased difficulty with longer dependencies; (2) they show facilitated learning when they have had experience with shorter-distance variations of that dependency; (3) they learn distance-invariant representations of nonadjacent dependencies, making
the correct prediction for Span3 items even in the SpanX-3 condition, where they had no training with dependencies of that span. This evidence that SRNs can learn a distance-invariant representations of nonadjacent dependencies is a critical finding, as it undercuts one of the fundamental arguments against association-based representations of knowledge, and in favor of rule-based explanations of cognition.

\section*{Study 2: Abstract Rules}

Marcus et al. (1999) performed a learning study with infants, where the infants where played sequences of syllables following either an ABB repetition pattern (e.g. "go-la-la") or an ABA alternation pattern (e.g. "go-la-go"). After hearing many examples repeated multiple times, infants then heard novel test sequences that either followed or violated that rule, and showed evidence of discriminating the legal and illegal sequences. Marcus argued that because no items were co-present at training and test, associative accounts were inadequate and only rule-based models could explain behavior. Marcus (2000) further argued that SRNs (like in Figure 4), could not in principle account for this finding. A number of researchers (Altmann \& Dienes, 1999; Christiansen \& Curtin, 1999) presented distributed SRN models of this phenomenon, where microfeatures (but not items) were co-present at training and test. Marcus, however, argued that resorting to such microfeatures was proof that SRNs and other network models are fundamentally incapable of learning abstract, algebraic rules, which some believe to be fundamental to human cognition.
In Study 2, I show that a simple, localist SRN without any distributed microfeature information learns to represent abstract, rule-like structure. Marcus's (2000) characterization of SRNs was correct; a localist SRN trained in the manner he described cannot show transfer of the rule-like knowledge. That is because the network learns (during the initial training) that the elements in the test items never occur, and thus their weights are set to zero, making them unable to make use of any information about the previous items' sequential structure that may have been learned and stored in the network's recurrent or output connections. However, there is no reason to restrict training in this way; one could instead allow the model to continue learning during the test phase, and again determine whether the model learns about the rule consistent test strings more quickly than the rule-violating ones.

\section*{Stimuli and Design}

The models in Study 2 (using the architecture in


Figure 4. A depiction of the architecture in Study 2. The actual model had 12 A - and B-units and 25 hidden units.

Figure 4) were trained on the exact design from Marcus et al., shown in Table 2. During the first training phase, models were trained in one of two conditions: (1) an ABA condition, where the first item perfectly predicted the last item, and predicted that it would be repetition of itself; (2) an ABB condition, where the middle item perfectly predicted the last item, again a repetition of itself. These ABA and ABB strings were composed of six possible A's and B's, which all occurred in all possible combinations, thus making all transition probabilities uninformative, and leaving the item-independent ABA or ABB rule as the only way to correctly predict whether the final element should be an A or B. The models were then given a second training phase, where they were trained on a new \(A B A\) or \(A B B\) sequences using new \(A\) and \(B\) items, and tested to see if they learned these sequences more quickly if the new rule was consistent with the rule on which they had been trained in phase 1.

Table 2. Stimulus inputs used in Study 2.
\begin{tabular}{|c|c|c|c|}
\hline ABA1 & ABB1 & ABA2 & ABB2 \\
\hline \(\mathrm{A}_{1} \mathrm{~B}_{1} \mathrm{~A}_{1}\) & \(\mathrm{A}_{1} \mathrm{~B}_{1} \mathrm{~B}_{2}\) & \(\mathrm{A}_{7} \mathrm{~B}_{7} \mathrm{~A}_{7}\) & \(\mathrm{A}_{7} \mathrm{~B}_{7} \mathrm{~B}_{7}\) \\
\hline \(\mathrm{A}_{1} \mathrm{~B}_{2} \mathrm{~A}_{1}\) & \(\mathrm{A}_{1} \mathrm{~B}_{2} \mathrm{~B}_{2}\) & \(\mathrm{A}_{7} \mathrm{~B}_{8} \mathrm{~A}_{7}\) & \(\mathrm{A}_{7} \mathrm{~B}_{8} \mathrm{~B}_{8}\) \\
\hline \(\mathrm{A}_{1} \mathrm{~B}_{3} \mathrm{~A}_{1}\) & \(\mathrm{A}_{1} \mathrm{~B}_{3} \mathrm{~B}_{3}\) & \(\mathrm{A}_{7} \mathrm{~B}_{9} \mathrm{~A}_{7}\) & \(\mathrm{A}_{7} \mathrm{~B}_{9} \mathrm{~B}_{9}\) \\
\hline \(\ldots\) & \(\ldots\) & \(\cdots\) & \(\cdots\) \\
\hline \(\mathrm{A}_{1} \mathrm{~B}_{6} \mathrm{~A}_{1}\) & \(\mathrm{A}_{1} \mathrm{~B}_{6} \mathrm{~B}_{6}\) & \(\mathrm{A}_{7} \mathrm{~B}_{12} \mathrm{~A}_{7}\) & \(\mathrm{A}_{7} \mathrm{~B}_{12} \mathrm{~B}_{12}\) \\
\hline \(\ldots\) & ... & \(\ldots\) & \(\cdots\) \\
\hline \(\mathrm{A}_{6} \mathrm{~B}_{6} \mathrm{~A}_{6}\) & \(\mathrm{A}_{6} \mathrm{~B}_{6} \mathrm{~B}_{6}\) & \(\mathrm{A}_{12} \mathrm{~B}_{12} \mathrm{~A}_{12}\) & \(\mathrm{A}_{12} \mathrm{~B}_{12} \mathrm{~B}_{12}\) \\
\hline
\end{tabular}

\section*{Results \& Discussion}

The results from Study 2 are shown in Figure 5. When the model was allowed to continue learning during the second training phase, it shows facilitated learning if the new items follow the same structural sequence as the items in the first phase. Follow-up
analyses of the network's weight configurations show this is because the network's recurrent and output weights are effectively learning the abstract structural order of the sequence. Because of this, if the new set of items are following the same structural rule, all the network needs to do is learn to adjust the input weights for the new items so that they work well with the already-learned recurrent and output weights.


Figure 5. Average SRN performance during the second stage of learning in Study 2.

These findings have very significant implications, as they (along with the findings in Simulation 2), refute claims that associative models are not capable of learning the abstract and rule-like knowledge that seems fundamental to human cognition.

\section*{Study 3: Perceptual/Semantic Bootstrapping}

Previous research on nonadjacent dependencies has mainly focused on learning to represent sequences of events, actions, or words independent of other cues about those entities, such as perceptual or semantic features or similarity. Learning structure in such a purely symbolic way would be hard. However, there is no reason to limit attention to this type of impoverished input, which is uncharacteristic of naturalistic conditions.. Studies that have examined the use of correlated perceptual cues (Newport \& Aslin) or semantic cues Willits et al.), have found that under these circumstances nonadjacent dependencies are significantly easier to learn. For example, Willits et al. found that when the items to be learned are from the same category (e.g. nonadjacently related items both foods), both infants and adults learn the dependency more easily. Learners even learn the nonadjacent dependency if the two words form a consistent mapping between categories (e.g. across set of nonadjacent pairs, foods are always paired with an animals). These findings are critical, because many of the nonadjacent dependencies people need to learn have these kinds of correlated perceptual and semantic attributes.

The question, then, is whether SRNs also exhibit facilitated learning from correlated cues, thus broadening their appeal as a general model of dependency learning, and whether they provide any insights as to why learning might be easier under these circumstances. This was investigated in Study 3.

\section*{Stimuli and Design}

The models in Study 3 were trained using the architecture in Figure 4. This architecture allowed for tests of whether correlated similarity structure affected learning by allowing each input to activate two units: (1) one item-specific unit (either an \(\mathrm{A}_{\mathrm{N}}, \mathrm{X}_{\mathrm{N}}\), or \(\mathrm{B}_{\mathrm{N}}\) ), where the letter refers to which category the item is from); (2) a category-specific unit (either Category \({ }_{\mathrm{A}}\), Category \({ }^{2}\), or Category \({ }_{B}\) ), where the category unit turned on for all inputs that came from that category.

Output Layer


Figure 6. A depiction of the architecture used in Study 3. The actual model had 25 hidden units.

The models were trained in one of three conditions (shown in Table 3). In the Consistently Same condition, the nonadjacently dependent items were always from the same category (e.g. the first item in each sequence would activate the \(\mathrm{A}_{1}\) unit and the Category \(_{\mathrm{A}}\) unit, and third item would activate the \(\mathrm{A}_{3}\) unit and the Category \({ }_{A}\) unit). In the Consistently Different condition, the nonadjacently dependent items were consistently from opposite A \& B categories. In the Inconsistent condition, the dependent items'

Table 3. Stimulus inputs used in Study 3
\begin{tabular}{ccccc}
\begin{tabular}{c} 
Consistently \\
Same Category
\end{tabular} & \begin{tabular}{c} 
Consistently Different \\
Categories
\end{tabular} & & \begin{tabular}{c} 
Inconsistent \\
Categories
\end{tabular} \\
\cline { 1 - 2 } & & \(\mathrm{A}_{1} \mathrm{X}_{\mathrm{n}} \mathrm{X}_{3} \mathrm{~B}_{3}\) \\
\(\mathrm{~A}_{2} \mathrm{X}_{\mathrm{n}} \mathrm{A}_{4}\) & & & \(\mathrm{~A}_{1} \mathrm{X}_{\mathrm{n}} \mathrm{A}_{3}\) \\
\(\mathrm{~A}_{\mathrm{n}} \mathrm{B}_{4}\) & & \(\mathrm{~A}_{2} \mathrm{X}_{\mathrm{n}} \mathrm{B}_{4}\) \\
\(\mathrm{~B}_{1} \mathrm{X}_{\mathrm{n}} \mathrm{B}_{3}\) & & \(\mathrm{~B}_{1} \mathrm{X}_{\mathrm{n}} \mathrm{A}_{3}\) & & \(\mathrm{~B}_{1} \mathrm{X}_{\mathrm{n}} \mathrm{B}_{3}\) \\
\(\mathrm{~B}_{2} \mathrm{X}_{\mathrm{n}} \mathrm{B}_{4}\) & & \(\mathrm{~B}_{2} \mathrm{X}_{\mathrm{n}} \mathrm{A}_{4}\) & & \(\mathrm{~B}_{2} \mathrm{X}_{\mathrm{n}} \mathrm{A}_{4}\)
\end{tabular}
categories were not predictable in terms of the other unit in the dependency. Across training trials, the
models were compared to see if any of the conditions showed facilitated learning.

\section*{Results \& Discussion}

The results for Study 3 are shown in Figure 7. SRNs showed facilitated learning in both consistent conditions, but not the inconsistent conditions, results similar to behavioral experiments with infants and adults. Follow-up analyses of network behavior show this is because the network has an easier time learning the category sequences, an intriguing hypothesis to test in future work with human learners.


Figure 7. Average SRN performance for the three training conditions in Study 3.

\section*{Conclusions}

Nonadjacent dependencies are a necessary consequence of experiencing a hierarchically structured world though a linear sequence of inputs and actions. The current studies support the notion that SRNs and other recurrent networks are viable models of the representation of hierarchical knowledge. They are capable of learning to represent abstract, rule-like structure (Study \(1 \& 2\) ), and they show critical learning effects that people exhibit, such as the interaction between structure and similarity (Study 3 ).

In addition, these simulations also provide evidence for the hypothesis that many learning situations that are thought to be especially difficult (of which the learning of nonadjacent dependencies is but one example) are only difficult because the problem has been underepresented. Many cues learners might use are stripped away in overly controlled experiments, making the problem harder than it is in the real world. Complexity is not the same thing as noise, if that complexity provides learners with useful cues to the structure of the world.

\section*{Acknowledgments}

This work was funded by NICDC F31-DC00936-02 to Jon Willits. The work received much useful input from Mark Seidenberg, Jenny Saffran, Maryellen MacDonald, and Timothy Rogers, and Jessica Montag.

\section*{References}

Altmann, G.T.M. and Dienes, Z. (1999). Rule learning by seven-month-old infants and neural networks. Science. 284, 875.
Bever, T. G., Fodor, J., A., \& Garrett, M. (1968). A formal limitation of associationism. . In T. R. Dixon \& D. L. Horton's Verbal Behavior and General Behavior Theory. Englewood Cliffs: Prentice Hall.
Botvinick, M., \& Plaut, D.C. (2004). Doing without schema hierarchies: A recurrent connectionist approach to normal and impaired routine sequential action. Psychological Review, 111, 395-429.
Chomsky, N. (1957). Syntactic Structures. The Hague.
Christiansen, M. H., \& Curtin, S. L. (1999). The power of statistical learning. In Proceedings of the twentyfirst conference of the Cognitive Science Society.
Cleeremans, A., \& McClelland, J. L. (1991). Learning the structure of event sequences. Journal of Experimental Psychology: General, 120, 235-253
Elman, J. (1990). Finding structure in time. Cognitive Science, 14, 179-211.
Elman, J. (1991). Distributed representations, simple recurrent networks, and grammatical structure. Machine Learning, 7, 195-225.
Fiser, J., \& Aslin, R.N. (2002). Statistical learning of higher-order temporal structure from visual shapesequences. Journal of Experimental Psychology: Learning, Memory, and Cognition, 28, 458-467.
Fodor, J., \& Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. Cognition, 28, 3-71.
Gómez R.L. (2002). Variability and detection of invariant structure. Psychological Science, 13, 431-36.
Lany, J., \& Gómez, R. (2008). Twelve-month-old infants benefit from prior experience in statistical learning. Psychological Science, 19, 1247-1252.
Marcus, G. F., Vijayan, S., Bandi Rao, S., and Vishton, P. M. (1999). Rule-learning in seven-month-old infants. Science, 283, 77-80.
Marcus, G. F. (2000). Rethinking eliminative connectionism. Cognition, 37, 243-282.
Newell, A. \& Simon, H. A. (1972). Human Problem Solving. Prentice Hall.
Newport, E. L., \& Aslin, R. N. (2004). Learning at a distance: I. Statistical learning of nonadjacent dependencies. Cognitive Psychology, 48, 127-162.
Rumelhart, D. E., McClelland, J. L., \& the PDP research group. (1986). Parallel distributed processing: Volume I. Cambridge, MA: MIT Press.
Willits, J. A., Lany, J., \& Saffran, J. R. (in review). Semantic cues facilitate nonadjacent dependency learning in infancy.

\title{
Acquisition of Phrase Structure in an Artificial Visual Grammar
}

\author{
Sarah T. Wilson (sarahtwilson@gatech.edu) \\ School of Psychology, 654 Cherry St., J.S. Coon Building, Atlanta, GA 30332 \\ Carla L. Hudson Kam (Carla.HudsonKam@ubc.ca) \\ Department of Linguistics, 2613 West Mall, Vancouver, BC, V6T 1Z4
}

\begin{abstract}
Recent studies showing learners can induce phrase structure from distributional patterns (Thompson \& Newport, 2007; Saffran, 2001) suggest that phrase structure need not be innate. Here, we ask if this learning ability is restricted to language. Specifically, we ask if phrase structure can be induced from non-linguistic visual arrays and further, whether learning is assisted by abstract category information. In an artificial visual grammar paradigm where co-occurrence relationships exist between categories of objects rather than individual items, participants preferred phrase-relevant pairs over frequency-matched non-phrase pairs. Additionally, participants generalized phrasal relationships to novel pairs, but only in the cued condition. Taken together these results show that learners can acquire phrase structure in a nonlinguistic system, and that cues improve learning.
\end{abstract}

Keywords: statistical learning, language learnability, syntax, modality independence

\section*{Introduction}

Theories of syntax differ, however, most contain two important elements: words are members of categories (traditionally nouns, verbs, determiners, etc.) and these categories are related to each other in higher-order patterns, e.g., phrases or sentences. To give an example in English take the sentence "The cat chased the dog." The word "cat" is a member of the word class, or category, noun and it has a relationship with "the" - its determiner - forming a noun phrase. A similar relationship exists between "the" and "dog." The verb phrase is comprised of "chased" plus "the dog." Thus, the sentence consists of several phrases defined over categories, arranged hierarchically.

In the traditional view, these elements of language are not learned, but rather considered to be innate by necessity (e.g. Crain, 1992; Wexler, 1991). A number of recent studies have begun to challenge the notion that these aspects of language are unlearnable, however, particularly with respect to categories (see, e.g., Mintz, 2002). The other basic properties of syntax, namely phrases (the property of interest in the current study) and their hierarchical organization have proved more challenging for a learning account. Saffran (2001) created a miniature artificial language, based on one used by Morgan, Meier, and Newport (1987), that was defined by a grammar over classes of words. Phrase structure in this language was defined by a number of rewrite rules over a basic or
canonical sentence type: \(\mathrm{S} \rightarrow \mathrm{AP}+\mathrm{BP}+(\mathrm{CP})\), where AP , BP , and CP are phrases, and CP is an optional phrase. The phrase rewrite rules were: \(\mathrm{AP} \rightarrow \mathrm{A}+(\mathrm{D}) ; \mathrm{BP} \rightarrow \mathrm{CP}+\mathrm{F}\) or \(\mathrm{BP} \rightarrow \mathrm{E}\); and \(\mathrm{CP} \rightarrow \mathrm{C}+(\mathrm{G})\). Learning of this grammar was statistically above chance; however, it was only marginally so, leaving open the question of whether phrase structure is an innate component of human knowledge.

More recently, Thompson and Newport (2007) used an adapted version of the same language with stronger cues to phrase boundaries - in particular, phrases tended to hang together in perfectly predictive relationships, while various language-like sentential manipulations created dips in predictive dependencies across phrase boundaries that were relatively low - and found greatly enhanced learning.

More specifically, the Thompson and Newport (2007) language had a phrase structure where phrases were composed of pairs of categories of words. There were 6 categories (labeled here, for simplicity: A, B, C, D, E, and F) which formed three phrases: AB, CD, and EF. Categories were distributionally defined. That is, the only way in which words were in the same category was that they occurred in the same locations both absolutely (their place in the sentence) and relatively (their adjacency to other elements). There were a total of 18 monosyllabic words in the language, 3 per category. Phrases could take part in a variety of operations: (1) movement, (2) repetition, (3) omission, and (4) insertion, thereby creating a set of sentences where the probability of a transition between categories within phrases was high (perfect 1.0) and the probability of a transition between categories that occurred across phrase boundaries was low. Importantly, the probability of a transition between individual words was also low, both within and across phrases. Therefore, the only indicator of structure was the transitional probabilities between categories of words - a higher-order relationship. At test, adult participants selected pairs of words which comprised a grammatical phrase more often than pairs of words which had co-occurred equally often in the input but which did not form a phrase, demonstrating they had acquired an understanding of category-level relationships. That is, they had learned categories as well as which categories formed phrases and which did not.

We investigate whether higher-order category relationships of this type are learnable in a non-linguistic system, something that might be expected if such learning is domain general. We exposed participants to visual stimuli constructed to have the same properties as the auditory language used by Thompson and Newport (2007). Simple
two-dimensional objects were organized into categories, then arranged into visual arrays according to a phrase structure grammar based on how categories of objects cooccurred. After exposure, participants were tested to see if they had learned the category-based grammar governing the combination of the items in the array.

We also assessed whether and how learning was affected by the presence and reliability of (nondistributional) cues to category membership. In previous work on larger versions of auditory languages (i.e., languages with a greater number of words per category than Thompson \& Newport, 2007) we found that phrase learning is affected by the presence and reliability of cues to category membership (Wilson \& Hudson Kam, 2009, 2013). Presumably, the cue makes it easier for people to identify the categories, thereby facilitating the tracking of probabilities over the categories necessary for phrase learning. We were interested in whether this would also be true of learning in the context of a non-linguistic visual system, and so included subtle visual cues to category membership in varying degrees in different conditions.

The visual array paradigm used here is based on that originally developed by Fiser and Aslin (2001). In their third and final experiment, Fiser and Aslin exposed adult participants to a set of visual arrays in which the adjacency relationships had a specific statistical structure irrespective of absolute spatial location. There were 12 uniquely-shaped black objects. Pairs of objects formed base pairs, always appearing together, in one of three possible alignment types: (1) vertical, (2) horizontal, or (3) oblique (diagonal). Additionally, the frequencies of some base pairs and crosspair, non-base pairs of items were equated. Therefore, the lower order, joint probability of these base pairs and cross pairs were equal (i.e., \(\mathrm{P}(\) object1, object2) \(=\mathrm{P}(\) object2, object3)), but the higher-order relative statistic, their conditional probabilities, differed (i.e., \(\mathrm{P}(\) object \(2 \mid\) object 1\()=\) 1.0 vs. \(\mathrm{P}(\) object2|object3) ~ low). At test, participants reliably chose the base pairs over cross pairs, suggesting they understood the higher order conditional probability relationship. (See Figure 1 for a schematic of a sample exposure scene.)


Figure 1. Schematic of example scene from Fiser and Aslin (2001), composed of three base pairs (one vertical, one horizontal, one oblique)

Their paradigm was modified here to investigate the acquisition of a phrase structure, where statistical relationships occur across pairs of categories, as opposed to pairs of individual items. To implement these ideas in the visual array paradigm, we expanded base pair relationships to include categories of objects which were adjacent in relevant configurations, while equating the co-occurrence of individual items within and across phrase boundaries. If our hypothesis is correct, that the learning processes that contribute to learning phrase structure are domain general, then we expect learning outcomes in the visual system to be commensurate with those found in previous auditory artificial language learning work, namely that it is possible to learn from dips in transitional probability that occur between categories of items in order to understand category relatedness (i.e. phrases) and that this learning is facilitated by non-distributional cues to category membership.

\section*{Methods}

\section*{Participants}

A total of 60 adults ( 20 per condition) participated in this study for course credit in Psychology courses at the University of California, Berkeley.

\section*{Stimuli}

Twenty-four unique objects were used, each with a unique color (properties of the color to be discussed later). Objects were assigned to one of eight categories (A, B, C, D, E, F, G, and H), with three objects per category. Pairs of categories were then grouped into phrases (much like the previous experiments), in one of two forms: vertical or horizontal. Phrases were then arranged into one of 16 distinct arrays in a five by five grid, with each array containing one example of each phrase. The 16 arrays, or category constructions, are much like sentence types. As such, the arrays constitute the 'grammar' of the visual system. Four distinct example arrays are shown in Figure 2.


Figure 2. Four examples of the 16 construction types or arrays with category placement labels.

This design resulted in conditional probabilities of adjacent co-occurrence of categories within phrases being perfect (1.0). Adjacent co-occurrence of pairs of categories that were possible but not necessary - i.e, which crossed a phrase boundary - had much lower conditional probabilities: each occurred exactly once over the exposure set, and therefore with \(p=.0625\). The complete set of adjacent cooccurrence relationships, for both the vertical and horizontal dimensions appear below in Tables 1 and 2.

Table 1. Adjacent co-occurrence conditional probabilities, vertical from top category to bottom category (phrase transitions in bold)
\begin{tabular}{ccccccccc} 
& A & B & C & D & E & F & G & H \\
A & - & \(\mathbf{1 . 0}\) & - & - & - & - & - & - \\
B & - & - & .06 & - & .06 & .06 & .06 & .06 \\
C & - & - & - & \(\mathbf{1 . 0}\) & - & - & - & - \\
D & .06 & - & - & - & .06 & .06 & .06 & .06 \\
E & .06 & - & .06 & - & - & - & .06 & .06 \\
F & .06 & - & .06 & - & - & - & .06 & .06 \\
G & .06 & - & .06 & - & .06 & .06 & - & - \\
H & .06 & - & .06 & - & .06 & .06 & - & -
\end{tabular}

Table 2. Adjacent co-occurrence conditional probabilities, horizontal from left category to right category (transitions in bold)
\begin{tabular}{ccccccccc} 
& A & B & C & D & E & F & G & H \\
A & - & - & .06 & .06 & .06 & - & .06 & - \\
B & - & - & .06 & .06 & .06 & - & .06 & - \\
C & .06 & .06 & - & - & .06 & - & .06 & - \\
D & .06 & .06 & - & - & .06 & - & .06 & - \\
E & - & - & - & - & - & \(\mathbf{1 . 0}\) & - & - \\
F & .06 & .06 & .06 & .06 & - & - & .06 & - \\
G & - & - & - & - & - & - & - & \(\mathbf{1 . 0}\) \\
H & .06 & .06 & .06 & .06 & .06 & - & - & -
\end{tabular}

The adjacent co-occurrence frequencies (or joint probabilities) of some within-phrase pairs and cross-phrase pairs of objects were equated. In order to accomplish this, some object pairs (pairing of particular objects either within or across phrases) were highly frequent (occurring 26 times) and some were less frequent (occurring 6 times). In this way, the less frequent within-category object pairs had equal joint probability as some cross-phrase object pairs (those that occurred adjacently in the 6 examples of any given scene) and served as test items. Additionally, some object pairs, both within phrase and across phrase boundaries, were
reserved from the exposure set also for test purposes.
The exposure set contained 96 unique scenes total, 6 of each construction type. (An example scene appears in Figure 3.) The exposure set was seen a total of four times, and so each scene appeared four times per session. All cross-phrase object pairs occurred 24 times per exposure session. Within-phrase object pairs occurred either 24 or 104 times per exposure session. Each individual object occurred exactly 32 times in the exposure set, and so occurred exactly 128 times per exposure session.

Each slide was seen for 2.5 seconds, and was interspersed with 1 second fixation slides. Additionally, there was a 2 minute break at the halfway point. The total exposure session lasted for approximately 25 minutes.


Figure 3. Example visual array (of construction type 1 from Figure 2), with phrases outlined

Note that the visual displays merely appear as complex designs; there is nothing in the visual arrays themselves that indicates the phrasal structure. If anything, Gestalt principles (Palmer, 1999) might lead participants to 'missegment' individual arrays into components larger than the phrases. In Figure 2 array 3, for example, participants might perceive two squares rather than four phrases, or in the display in Figure 3 participants might see an archway.

\section*{Experimental Manipulation}

This study also addressed the contribution of a subtle non-distributional cue to category membership in acquisition of the phrase structure. The visual cue to category was an aspect of the color of the objects irrespective of hue. Colors for objects were selected from levels of brightness and saturation available in Microsoft Powerpoint - three hues from each level. In the cuepresent version of the visual arrays, objects from the same category were of different hues from the same brightness and saturation level. In the without cue condition, objects were randomly assigned to categories, therefore, color could not serve as a cue to category membership. A third version of the arrays contained a partially predictive cue to category membership, where two of the three objects in the category were of the same brightness and saturation level.


Figure 4. All 24 objects, shown in respective color assignment, organized into 8 levels of brightness and saturation, (category shown at bottom of column).

\section*{Tests}

There were two types of tests in this experiment designed to test whether participants understood the phrases or units of the visual grammar - very much like the phrase tests from Thompson and Newport (2007). Both tests required participants to compare two pairs of objects: one with a high category-level conditional probability and one with a low category-level conditional probability. The two comparison pairs were displayed to the left and to the right of the center square of the \(5 \times 5\) grid, as shown in Figure 5 .

Phrase Test. Some pairs of objects in the exposure set were matched for frequency - that is, had the same joint probabilities of appearing together - either within or across a phrase boundary. However, the pairs differed in that some had high category-level conditional probability (i.e., they were within a phrase) while others had a category-level conditional probability that was low (i.e,, they were not within a phrase). The first test compared these two types of pairs. There were 12 such items total, six on the first day and six on the second day.

Generalization Test. The second test was a generalization test, in which participants were tested using pairs of objects that had been reserved from the exposure set. One test pair was a novel object pair with high category-level conditional probability. The comparison pair of objects was also novel, but with a low category transitional probability (but not zero or absent). There were 12 of these items, six on the first day and six on the second day.


Figure 5. Sample test item, within-phrase object versus frequency matched objects crossing a phrase boundary (vertical phrase).

\section*{Procedure}

Participation in this study spanned two days, with each day involving an exposure session and a test session. Unlike earlier experiments that tested strictly end-state performance outcomes, we also were interested in the trajectory of learning - whether we could capture an intermediary stage of having learned some aspects, but not all, of the grammar.

On each day, participants saw the exposure set a total of eight times: four times through, followed by a two-minute break, then another four times through, for a total exposure session of about 25 minutes. Across both days, participants saw the exposure set 16 times. After exposure on both days, participants were given the two-alternative, forced choice test.

The phrase test items were always given first, followed by the generalization test items. Prior to test, participants were shown a practice comparison that contained objects that had not appeared in the scenes, first in the vertical then the horizontal orientation. Participants were instructed that they were going to indicate which of the pairs of objects they thought more likely came from the scenes they had been learning about. Responses were recorded by the experimenter, who was also advancing the test-item slides. Participants were given as much time as they needed to make a response.

\section*{Results}

First, it is of interest to compare performance on the initial phrase test both across the two days and across conditions. This test compared pairs of objects with either high or low category-level conditional probability, with test pairs in the comparison having appeared with the same frequency in the exposure set. Importantly, successful performance on this test cannot be accounted for by simple adjacency since both pairs in the comparison had occurred an equal number of times in the exposure. Mean performance outcomes on this test appear in Figure 6.

An overall, 2 x 3 (day x cue-condition) ANOVA revealed a significant interaction between the two factors in the analysis \((F(5,119)=3.93, p=.022\). (An examination of main effects, day and cue condition, revealed that there were no significant differences \((\mathrm{F}(1,119)=.5452, \mathrm{p}=.463\) and \(\mathrm{F}(2,119)=.094, \mathrm{p}=.910\) respectively). This was also true for simple main effects of condition on both days ( \(\mathrm{F}(2\), 59) \(=1.640, \mathrm{p}=.203\) ) and \(\mathrm{F}(2,59)=1.936, \mathrm{p}=.154)\).) The interaction reflects the difference in performance patterns for the groups by day, which was, interestingly, not significant for either day. However, given that it was our expectation that all or some of the cue groups would demonstrate learning of the phrases, given results from previous work with auditory languages where this type of distinction was possible, while allowing for differences in performance, we did performance comparisons for each cue group against chance level performance. On Day 1, Without Cue participants performed significantly above chance, \(\mathrm{M}=\) \(63.3 \%, \mathrm{SD}=48.4 \%(\mathrm{t}(19)=2.707, \mathrm{p}=.014)\), while With


Figure 6. Mean percent correct on the first phrase test. Dashed line indicates chance level performance.

Cue participants performed at chance level, \(\mathrm{M}=52.5 \%\), \(\mathrm{SD}=50.1 \%(\mathrm{t}(19)=.529, \mathrm{p}=.603)\) as did Partially Predictive Cue participants, \(\mathrm{M}=53.3 \%, \mathrm{SD}=50.1 \%\) ( \(\mathrm{t}(19\) ) \(=.748, \mathrm{p}=.464\) ).

These means from the second day were also tested against chance performance. Without cue participants performed at chance level, \(\mathrm{M}=50.0 \%, \mathrm{SD}=50.2 \%(\mathrm{t}(19)=\) \(.000, \mathrm{p}=1.000\) ), while With Cue participants performed above chance, \(\mathrm{M}=65.0 \%\), \(\mathrm{SD}=47.9 \%(\mathrm{t}(19)=2.932, \mathrm{p}=\) .009) as did Partially Predictive Cue participants, \(\mathrm{M}=\) \(63.3 \%, \mathrm{SD}=48.4 \%(\mathrm{t}(19)=2.320, \mathrm{p}=.032)\).

We also tested participants' ability to generalize to novel phrases. This test asked participants to compare novel base pairs that had been reserved from the exposure set, but which again differed in that one had a high category-level conditional probability and one had a low category-level conditional probability. Mean performance scores on this test can be seen in Figure 7. An overall, \(2 \times 3\) (day x cuecondition) ANOVA did not reveal a significant interaction \((F(5,119)=.173, p=.841)\). Nor were there main effects of day
or cue-condition \((\mathrm{F}(1,119)=.640, \mathrm{p}=.425\) and \(F(2,119)=2.404, p=.095)\). Simple main effects of condition, additionally, were null for each day \((\mathrm{F}(2,59)=1.862, \mathrm{p}=.165\) and \(F(2,59)=.646, p=.528)\). Nonetheless, there were some intriguing patterns in the data that we pursued further with individual group analysis. As before, we performed planned comparisons to chance. With Cue participants performed significantly above chance on the first day ( \(M=62.5 \%\), SD \(=48.6 \%(\mathrm{t}(19)=2.380, \mathrm{p}=.028)\) ) while Without Cue performed at chance \(\mathrm{M}=49.2 \%, \mathrm{SD}=50.2 \%(\mathrm{t}(19)=-.188\), \(\mathrm{p}=.853\) ), as did the Partially Predictive Cue participants ( M \(=51.7 \%\), \(\mathrm{SD}=50.2 \%(\mathrm{t}(19)=.302, \mathrm{p}=.766))\).

We also compared performance on the generalization test for the second day. On this day, With Cue, Without Cue, and Partially Predictive Cue participants all scored at chance level \((\mathrm{M}=55.8 \%, \mathrm{SD}=49.9 \%(\mathrm{t}(19)=1.234, \mathrm{p}=\) \(.232) ; \mathrm{M}=48.3 \%, \mathrm{SD}=50.2 \%(\mathrm{t}(19)=-.302, \mathrm{p}=.766)\); and \(\mathrm{M}=49.2 \%, \mathrm{SD}=50.2 \%(\mathrm{t}(19)=-.165, \mathrm{p}=.871)\), respectively.


Figure 7. Mean percent correct on the second phrase test

\section*{Discussion}

This experiment was designed to assess whether category relatedness or phrases can be inferred in a nonlinguistic system, or is instead a property only of linguistic systems. In addition, we asked whether cues to category membership would function similarly in the auditory and visual domains. Participants were exposed to visual arrays comprised of phrases defined over categories, arranged so that the within-phrase category-level conditional probabilities were higher than those of categories that co-occurred but did not form phrases. Participants were then tested to see if they had acquired the phrases of the visual grammar. The hypothesis was that general purpose learning processes would enable the acquisition of phrase structure in the visual system as in the auditory language, and that these learning processes would be improved by cues that facilitated the matching of items in categories. If this is the case, the relative statistics in the input should inform judgments about category relatedness that contrast pairs of objects that are a phrase-relevant pair to pairs that cross phrase boundaries.

We found some evidence of this. On the first day, Without Cue participants performed above chance on the first phrase test, demonstrating that they had learned something about the category-level co-occurrence probabilities. Surprisingly, performance in this group dropped on the second day - potentially the result of looking for further patterns in the stimuli that were not present. In contrast, With Cue and Partially Predictive Cue participants performed at chance level initially on Day 1 and went on to improve on the first phrase test on Day 2. These groups may have taken longer precisely because of the presence of distributional cues that were correlated with the color cue - they were figuring out that relationship first (as demonstrated by their above-chance performance on the second test), then having attained some (albeit shaky) knowledge of the color relationships, they went on to learn the relationships between categories. The With Cue participants were the only group to demonstrate above chance learning on the second test at all, on the first day -a result that may just be due to chance. However, since the relative pattern of performance (With Cue participants doing better) was consistent on this test across the two days we think that the fact that they performed better than the other two groups of participants on this test (even if not significantly so) is a real, if small, effect.

Given that the effects are sometimes present, sometimes absent, it may bring up the question as to whether there were particular aspects of our test stimuli that could have skewed the pattern of the data. There were a number of controls in place to minimize this possibility. While the particular test items were different for all three cueconditions, the number of pairings that incidentally paired objects of the same hue (albeit different brightness and saturation - as the cue dictated) were the same across all three conditions and all tests and were a very low number.

Additionally, each test had two versions: an A version as well as a B version, and those versions were randomized as to whether a particular participant received the A version on Day 1 or the B version. Thus, the pattern of results seems unlikely to have occurred due to particular test stimuli.

Ultimately, these general learning results should be replicated with different participants and stimuli if possible. The explanation for learning being sometimes present, sometimes absent should be explored and tested, possibly by looking at more individual learning trajectories. This project was intended to provide a visual analogue of both our previous work and work by Thompson and Newport (2007) - all of which provided a much longer input period. And so, this work may benefit from equivalent time on task to see if learning improves and generalization ability ever emerges and remains persistent in this paradigm.

\section*{Acknowledgments}

This work was supported by the National Institutes of Health Grant HD 048572 and a Discovery Grant (Individual) from the Natural Sciences and Engineering Research Council of Canada to CLHK.

\section*{References}

Crain, S. (1992). Language acquisition in the absence of experience. Behavioral and Brain Sciences, 14, 597-650.
Fiser, J. \& Aslin, R. N. (2001). Unsupervised Statistical Learning of Higher-Order Spatial Structures from Visual Scenes. Psychological Science, 12, 499-504.
Mintz, T.H. (2002). Category induction from distributional cues in an artificial language. Memory and Cognition, 30, 678-686.
Palmer, S. E. (1999) Vision science: Photons to phenomenology. Cambridge, MA: Bradford Books/MIT Press.
Saffran, J. R. (2001). The use of predictive dependencies in language learning. Journal of Memory and Language, 44, 493-515.
Thompson, S.P \& Newport, E.L. (2007). Statistical learning of syntax: The role of transitional probability. Language Learning and Development, 3, 1-42.
Wexler, K. (1991). On the Argument from the Poverty of the Stimulus. In Kasher, A. (Ed.) The Chomskyan Turn (252-70), Cambridge: Blackwell Publishers.
Wilson, S. T. \& Hudson Kam, C. L. (2013). Learning
Syntax Through Statistics: When do transitional probabilities need a boost? Unpublished manuscript, Georgia Institute of Technology, Atlanta, GA.
Wilson, S. T. \& Hudson Kam, C. L. (2009). Learning Syntax Through Statistics: When do transitional probabilities need a boost? Paper presented at the Boston University Conference in Language Development, Boston, MA

\title{
The Role of Scene Gist and Spatial Dependency among Objects in the Semantic Guidance of Attention
}

\author{
Chia-Chien Wu (chiachie@cs.umb.edu) \\ Hsueh-Cheng Wang (hchengwang@gmail.com) \\ Marc Pomplun (marc@cs.umb.edu) \\ Department of Computer Science, University of Massachusetts at Boston \\ 100 Morrissey Boulevard, Boston, MA, 02125-3393, USA
}

\begin{abstract}
A previous study (Hwang et al., 2011) found evidence for semantic guidance of visual attention during the inspection of real-world scenes, i.e., an influence of semantic relationships among scene objects on overt shifts of attention. In particular, the results revealed an observer bias toward gaze transitions between semantically similar objects. However, these results are not necessarily indicative of semantic processing of individual objects but may be confounded by knowledge of the scene gist, which does not require object recognition (Torralba et al., 2006), or by known spatial dependency among objects (Oliva \& Torralba, 2007). To examine the mechanisms underlying semantic guidance, in the present study, subjects were asked to view a series of displays with the scene gist removed and spatial dependency varied. Our results confirm the previous finding of semantic guidance and show that it is not entirely due to either the effect of scene gist or the spatial dependency among objects. Even without scene gist or spatial dependency, subjects still retrieved semantic information to guide their attention. This strategy may facilitate scene understanding and object memorization.
\end{abstract}

Keywords: Attention, semantics, eye movements, visual guidance, real-world scenes.

\section*{Introduction}

Real-world scenes contain rich information, which usually is not thoroughly processed during natural viewing. Therefore, the way in which the visual system deploys the limited attention resources is crucial for effective vision and has drawn huge interest over the last two decades. The guidance of attention based on the features of stimuli in the visual environment has been well investigated in both its bottom-up (Itti \& Koch, 2001; Koch \& Ullman, 1985) and top-down aspects (Hayhoe et al, 2003; Hwang, Higgins \& Pomplun, 2009; Pomplun, 2006).

Visual attention is not only affected by factors based on the overt visual appearance, but also by inherent factors, such as meaning and semantic relations among objects. Hwang, Wang and Pomplun (2011) found that during natural scene viewing, humans tend to bring their gaze to the objects that are semantically similar either to the currently fixated one or to the specified search target. This result, however, may have been confounded by the observers' knowledge of the global scene context. That is, instead of considering the semantic relation between the currently fixated object and the objects located in the extrafoveal visual field, observers may simply use their knowledge about the scene type to decide where to look
next. For example, if observers are aware that the viewed image is a kitchen, they may only attend the regions nearby the counter or sink, where most of the kitchenware is likely located.

The ways in which people acquire such global contextual information is not well understood. Torralba, Oliva, Castelhano and Henderson (2006) found that observers could extract some global scene properties - referred to as scene gist - without recognizing individual objects and use this information to guide their attention and eye movements.

Even when the global context, which usually comes from visual background information, is missing, it is still possible to learn some context of the scene. Chun (2000) showed that some contextual information can be learned merely by the typical arrangement of elements and affect the deployment of attention. Oliva and Torralba (2007) also found that spatial dependency among objects could provide different contextual information about a scene. For example, a chair may be expected to be located behind a table, or a fork may be expected to be next to a spoon.

In summary, both the scene gist and the spatial dependency among scene objects may have caused a bias in observers' gaze patterns that could explain the results of Hwang et al. (2011) without the need for semantic analysis of extrafoveal scene objects. If gist, object dependency, or both were entirely responsible for the effect observed in that study, the concept of semantic guidance would not be a new phenomenon but rather a bias introduced by already known factors. The aim of the present study was to discern the contributions of scene gist, object dependency, and semantic object analysis to semantic guidance in order to address this problem.

To study the influence of spatial dependency among scene objects, we employed the LabelMe object annotated image data base (Russell, Torralba, Murphy, \& Freeman, 2008) in which scene images were manually segmented into annotated objects by volunteers. In addition, the locations of objects are provided as coordinates of polygon corner and all objects are labeled with English words or phrases. It provides an excellent opportunity for not only segregating each object from its scene, but also shifting the object's coordinates to any desired location in the image.

One way to eliminate potential influence of scene gist on attentional guidance is to remove all background information and only keep segregated objects in the scene. We used the resulting images in an experimental condition
referred to as 'fixed condition'. This procedure effectively removes the relation between scene and objects as defined by Torralba et al. (2006). For example, it is easier to predict where a plate is located in a scene when the plate is shown on a dining table than when it is shown by itself. The spatial dependency among objects, however, is still retained when the background information is excluded. For instance, it is possible to predict the likely location of a glass merely based on the location of a seen plate in a scene, even when no context is provided. To remove the spatial dependency among objects as well, we created another set of stimuli ('scrambled condition') by generating scenes without background as in the fixed condition and then randomly shifting the objects within the scene.

If the semantic guidance found in the previous study (Hwang et al., 2011) were due to the spatial dependency among objects, this effect should be eliminated once the background information and spatial arrangement are removed. On the other hand, if observers are able to use conceptual semantic information between objects to guide attention, their gaze transitions should still show an abovechance semantic relevance.

\section*{Method}

\section*{Subjects}

Ten subjects, aged between 19-40 years old, were tested. All had normal or corrected to normal vision and were naïve as to the purpose of the study. Each subject received a \(\$ 10\) honorarium.

\section*{Apparatus}

Eye movements were tracked and recorded using an SR Research EyeLink-2k system. Its sampling frequency was set to 1000 Hz . Stimuli were presented on a 22-inch ViewSonic LCD monitor. Its refresh rate was set to 75 Hz and its resolution was set to \(1024 \times 768\) pixels. Participant responses were entered using a keyboard.

\section*{Stimulus display}

A total of 60 images ( \(1024 \times 768\) pixels) were generated. Each image was composed of 13 to 15 objects selected from a real-world scene from the LabelMe database (http://labelme.csail.mit.edu). The selected scenes included home interiors, landscapes and city scenes. Objects of extreme size (small or large) were not chosen as scene objects. To remove the scene gist or other global regularity from the scene, all objects were segregated from the image and were pasted on a grey canvas. Each object was placed at either the same coordinates as in the original scene, which was referred to as 'fixed condition', or at randomly selected locations on the canvas, referred to as 'scrambled condition'. In the scrambled condition, different objects were placed manually to avoid overlap and clutter (see Figure 1 for an example).


Figure 1: Original scene (top) and a sample trial (bottom). The upper panel shows the original scene used to generate stimulus displays. The scene would be used to generate an image with objects at same coordinates (fixed condition) and an image with objects at randomly selected locations (scrambled condition). During each trial, the created image was presented for 5 seconds. After the stimulus image disappeared, a word was presented and subjects had to report whether the indicated object had been shown in the previous display.

\section*{Procedure}

Subjects were instructed to inspect the scenes and memorize them for the subsequent object recall test (see Figure 1, bottom panel). Each image was presented for 5 seconds. After the image had disappeared, an English word was shown and subjects were asked whether the object indicated by the word had been shown in the previous scene. Subjects responded by pressing one of two possible keys on a keyboard. If they believed the indicated object was shown in the previous image, they would press the left arrow key. Otherwise, they would press the right arrow key. The next trial would begin once subjects made a response. Subjects performed a total of 60 trials ( 30 trials each in the fixed and scrambled conditions). Each scene was only presented once to each subject, either in the fixed condition or in the scrambled condition.

\section*{Data Analysis}

\section*{Assigning fixations to objects}

Since all images excluded the global contextual information by only leaving the selected objects on a grey canvas, some fixations may land on the blank area rather than on any object in the image. When this happened, we assumed this fixation was aimed at the nearest object, i.e., the one whose center had the shortest Euclidean distance to the current fixation location.

\section*{Latent Semantic Analysis}

Similar to the original semantic guidance study (Hwang et al., 2011), we used Latent Semantic Analysis (referred to as LSA; Landauer \& Dumais, 1997) to serve as a quantitative measure of semantic similarity between objects. LSA is able to extract and represent the contextual usage-meaning of words by statistical computations applied to a large corpus of text. The basic premise in LSA is that the aggregate contexts in which a word does or does not appear provide a set of mutual constraints to deduce the word's meaning (Landauer, Foltz, \& Laham, 1998). The greater the cosine value, the higher is the semantic similarity. Since annotated objects in LabelMe have descriptive text labels, their semantic similarity can be estimated by calculating cosine values for the labels of object pairs.

LSA similarity computation can be described as follows: First, an occurrence matrix is constructed from a large corpus of text, where each row typically stands for a unique word, and each column stands for a document, which is typically a collection of words. Each cell contains the frequency with which the word occurred in the document. Subsequently, each cell frequency is normalized by an information-theoretic measure. However, it is computationally inefficient to operate with this very highdimensional matrix. Therefore, a form of factor analysis called Singular Value Decomposition (SVD; see Berry, Dumais, \& Obrien, 1995) is applied to reduce the matrix to a lower-dimensional vector space called 'semantic space'. LSA can still estimate the semantic similarity of two words even when they never co-occur in the same document (Jones \& Mewhort, 2007; Landauer \& Dumais, 1997).

Every term, every document, and every novel collection of terms has a vector representation in the semantic space. Thus, the pair-wise semantic similarity between any of them can be calculated as the cosine value of the angle between the two corresponding vectors, with greater cosine value indicating greater similarity. Table 1 shows examples of LSA cosine values for various object labels used in the LabelMe scene image "Child4" (see Figure 1) in terms of the reference object label "AIRPLANE". This label has, for instance, a higher cosine value (greater semantic similarity) with "HELICOPTER" (0.62) than with "PILLOW"' \((0.03)\). This difference indicates that in the text corpus, "AIRPLANE" and "HELICOPTER" occur in more similar contexts than "AIRPLANE" and "PILLOW". One of the nice features of LSA is that it can quantify
higher-level conceptual semantic similarity, regardless of any geometrical relation, functional relation or visual relation.

Table 1: Sample LSA cosine values
\begin{tabular}{lll}
\hline Label 1 & Label 2 & Cosine \\
\hline- & - & - \\
AIRPLANE & HELICOPTER & 0.62 \\
AIRPLANE & TOY TRAIN & 0.28 \\
AIRPLANE & PICTURE & 0.14 \\
AIRPLANE & PILLOW & 0.03 \\
- & - & - \\
\hline
\end{tabular}

To compute semantic similarity for each pair of object labels in our experiment, a web-based LSA tool, LSA@CU (http://lsa.colorado.edu), developed at the University of Colorado at Boulder, was used. This tool was set to create a semantic space from general readings up to 1st year college with 300 dimensions. Based on this space, we computed semantic similarity as the LSA cosine value, ranging between 0 and 1, for each object label compared to all other objects' labels for the same image.

\section*{Measuring semantic guidance}

In this study, the semantic guidance effect was defined as the extent to which the semantic relation/similarity between the currently fixated object and the other objects in the scene influences the choice of the next fixated object. In order to compute this effect quantitatively, the computation had to follow each subject's eye movements. Since we were interested in the effect of semantic similarity on gaze transitions, i.e., which object would be inspected next, only eye movements that transitioned between distinct objects were analyzed. For the starting point of each of these transitions, a semantic landscape was generated based on the LSA cosine value between the labels of the currently fixated object and each other object in the scene, as shown in Figure 2. The semantic landscapes, excluding the area occupied by the currently fixated object, were normalized so that the sum of all activation was one. With the normalized semantic landscape, the Receiver Operating Characteristic (ROC) value was computed in a similar way as it was done in previous studies (Hwang et al., 2009; Tatler, Baddeley \& Gilchrist, 2005). Overall, each fixation would build its own semantic landscape as a predictor of the target point of the next transition. All ROC values computed along scan paths were averaged across scenes to obtain the extent of semantic guidance during the inspection of a scene. If eye movements were exclusively guided by semantic information, this average ROC value should be close to one. If there were no semantic effect on eye movements at all, the average ROC value should be close to 0.5 , indicating prediction at chance level.


Figure 2: Example of semantic landscapes. The currently fixated object is marked with an orange square. (a) The original image that subjects inspected. (b) Semantic landscape during gaze fixation on the object labeled as "AIRPLANE". (c) Semantic landscape during gaze fixation on the object labeled as "GLOBE". (d) Semantic landscape during gaze fixation on the object labeled as "STORAGE BOX". As shown above, objects with conceptually higher relevance - measured as greater sematic similarity to the currently fixated object - receive higher activation (brightness), for example, the helicopter in (a) shows a higher activation due to the fixated object labeled as 'AIRPLANE'.

\section*{Excluding potential confounds by computing control analyses}

Following Hwang et al. (2011), to control for possible confounds in the measurements of semantic guidance, subjects' ROC values computed from their empirical gaze transition data were compared with two control data sets: (1) random fixations and (2) dissociated fixations. The random fixations were generated by replacing subjects' fixation positions with randomly positioned coordinates in the scene. This data set served as an unbiased test of ROC values. That is, since gaze transitions of the random data set were not affected by any other factor, we should always receive a chance level ROC value ( \(\mathrm{ROC}=0.5\) ).

Furthermore, it is likely that any above-chance ROC value was simply caused by the proximity effect. This effect is due to the previous finding (Hwang et al., 2011) that semantically similar objects tend to be located closer to each other and subjects' saccades tend to be shorter than gaze transitions in the random data set. To examine this possible confound, subjects' data were also compared with a "dissociated" data set. The dissociated data were analyzed using the eye movement data recorded in scene n against
object data from scene \(n+1\), and the eye movement data recorded from the last scene against the object data from the first scene. This mismatch conserved the spatial distribution of both the scene objects and the observers' fixations and therefore the proximity effect (at least in the fixed condition in which the coordinates of selected objects were not changed). This method allowed us to examine whether any observed above chance level ROC value for the empirical data was simply caused by proximity, which would be indicated by ROC values in the dissociated case being similar to the actual ROC values).

\section*{Experimental Results}

Results showed that subjects recall performances were above chance level in both the fixed and scrambled conditions (Recall performance in the fixed condition, 79\%, \(t(9)=21.50, p<0.05\); recall performance in the scrambled condition: \(70 \%, t(9)=4.36, p<0.05)\).
As mentioned earlier, in order to examine semantic guidance, we computed ROC values for the two experimental conditions (fixed vs. scrambled) for all three data sets (empirical, random and dissociated). Figure 3 shows that the transitional semantic guidance values of random fixations were close to 0.5 in both the fixed and scrambled conditions. This result shows that the ROC computation was applied properly and the normalized semantic landscapes used in our analysis were unbiased.

The ROC value in the fixed condition (ROC \(=0.704 \pm\) 0.14 ) was significantly higher than that in the scrambled condition ( \(\mathrm{ROC}=0.65 \pm 0.19\) ), \(t(9)=4.76, p<0.05\). ' \(\pm\) ' here indicates a mean value and its standard error.

This result suggests that the spatial dependency preserved in the fixed condition provided additional semantic information and facilitated semantic guidance. The ROC value decreased when this spatial dependency was eliminated by shuffling the locations of objects.

Interestingly, in the scrambled condition, where the spatial dependency among objects was destroyed, the ROC value of empirical transition between distinct objects was still substantially greater than both ROC values for the other two control cases. A one-way ANOVA showed that the effect was significant, \(F(2,27)=51.61, p<0.05\). A post hoc Tukey test indicated that there was no difference between the dissociated and random cases, \(p=0.53\). This finding shows that the proximity effect, at least in our experiment, had no impact on semantic guidance.

Overall, the results indicate that, even without scene gist and the spatial dependency among objects, subjects were still able to extract the semantic relevance between objects to guide their attention.


Figure 3: Transitional semantic guidance as measured by the ROC method in the fixed condition and the scrambled condition. The red dashed line represents the chance level \((\) ROC \(=0.5)\) and errors represent \(+/-1\) standard error of the mean.

\section*{Conclusions}

Hwang et al. (2011) found that, during scene inspection, observers tend to bring the line of sight to objects that are semantically relevant to the currently fixated object. Based on these previous data alone, it cannot be ruled out that the high semantic relevance of gaze transitions was contributed by scene gist information or by subjects' prediction of local scene context based on the spatial layout of objects. In other words, observers may not actually evaluate the semantics of peripheral objects for saccade target selection, and consequently, semantic guidance could not be considered a new phenomenon but rather an effect caused by other known mechanisms.

Our present results clarify the influence of these possible confounds in the previous findings. In the fixed condition in which the scene gist was removed, observers still showed strong semantic guidance. This result demonstrates that semantic guidance of visual attention in scene inspection is not entirely due to the scene gist. In fact, semantic guidance in the present study was even higher than that measured in Hwang et al. (2011), suggesting that scene gist only plays a marginal, if any, role in semantic guidance.

In the scrambled condition, in which both scene gist and possible spatial dependency among objects were removed, the effect of semantic guidance was slightly decreased but remained substantially higher than chance level. This finding shows that the spatial arrangement of objects only makes a small contribution to semantic guidance. Moreover, these data reveal that even when the scene gist was excluded and the spatial dependency was removed, subjects could still retrieve semantic information to guide their attention.

Moreover, Hwang et al. (2011) also found an even greater effect of semantic guidance in a visual search task. That is, observers tend to fixate on the objects which are semantically similar to the specified target. Instead of using any verbal probe and search paradigm as they did, the present study used a natural viewing and memory task which was less constrained by cognitive goal and we still found a substantial effect of semantic guidance.

Consequently, the question becomes how observers obtained this semantic information and how it influenced the guidance of attention. It is likely that extrafoveal visual processing may play a crucial role in enabling the semantic effect since observers had to recognize, at least partially, the objects in peripheral vision and processed the semantic relevance in the context of the currently fixated object. Kotowicz, Rutishauser and Koch (2010) found that, during visual search, observers already identified the extrafoveal target before fixating on it. We do not claim that in our task, observers were able to recognize the objects in the extrafoveal field. At the very least, extrafoveal perception may be used to increase the belief of what this object could be. Therefore, when contextual information was removed, people could still learn semantic information to help them determine where to fixate next by using the immediately acquired information from the current fixation and the information accumulated from extrafoveal vision. Such a strategy may facilitate scene understanding and memorization. It is also possible that, instead of using extrafoveal information, observers may construct their own scene representation merely based on the currently fixated object and update it during later fixations. This strategy may become useful when the extrafoveal information is not available or when the cost of processing it is too high.

Overall, the current study showed that semantic guidance of visual attention during the inspection of real-world scenes, as reported by Hwang et al. (2011), is a novel phenomenon that cannot be explained by effects of scene gist (e.g., Torralba et al., 2006) or spatial dependency among scene objects (e.g., Oliva \& Torralba, 2007) alone. It has been known that, in addition to the visual saliency from low-level features, attention could be also driven by other particular classes of objects, such as faces (Judd, Ehinger, Durand, \& Torralba, 2009) and texts (Wang \& Pomplun, 2012). Our result provides a new alternative class of features and suggests that the conceptual semantic effects may need to be considered in the present model of attentional guidance. Further research on semantic guidance, its underlying mechanisms, and its function is necessary before this concept can be integrated into existing models of visual attention.

\section*{Acknowledgments}

This research was supported by Grant Number R01 EY021802 from NIH to M.P.

\section*{References}

Berry, M., Dumais, S., \& O'Brien, G. (1995). Using linear algebra for intelligent information retrieval. SIAM review, 37(4), 573-595.
Chun, M. (2000). Contextual cueing of visual attention. Trends in Cognitive Sciences, 4(5), 170-178.
Hayhoe, M. M., Shrivastava, A., Mruczek, R., \& Pelz, J. B. (2003). Visual memory and motor planning in a natural task. J Vis, 3(1), 49-63.
Hwang, A. D., Higgins, E. C., \& Pomplun, M. (2009). A model of top-down attentional control during visual search in complex scenes. J Vis, 9(5), 25.1--2518.
Hwang, A. D., Wang, H.-C., \& Pomplun, M. (2011). Semantic guidance of eye movements in real-world scenes. Vision Res, 51(10), 1192-1205.
Itti, L., \& Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Res, 40(10-12), 1489-1506.
Jones, M., \& Mewhort, D. (2007). Representing word meaning and order information in a composite holographic lexicon. Psychological Review; Psychological Review, 114(1), 1.
Judd, T., Ehinger, K., Durand, F., \& Torralba, A. (2009). Learning to predict where humans look, IEEE International Conference on Computer Vision (ICCV), Kyoto, Japan, 2106-2113.
Koch, C., \& Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. Hum Neurobiol, 4(4), 219-227.
Kotowicz, A., Rutishauser, U., \& Koch, C. (2010). Time course of target recognition in visual search. Frontiers in Human Neuroscience, 4, 12.
Landauer, T. K., Laham, D., \& Derr, M. (2004). From paragraph to graph: latent semantic analysis for information visualization. Proc Natl Acad Sci U S A, 101 Suppl 1, 5214-5219.
Landauer, T., \& Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review; Psychological Review, 104(2), 211.
Landauer, T., Foltz, P., \& Laham, D. (1998). An introduction to latent semantic analysis. Discourse processes, 25(2-3), 259-284.
Oliva, A., \& Torralba, A. (2006). Building the gist of a scene: the role of global image features in recognition. Prog Brain Res, 155, 23-36.
Oliva, A., \& Torralba, A. (2007). The role of context in object recognition. Trends Cogn Sci, 11(12), 520-527.
Pomplun, M. (2006). Saccadic selectivity in complex visual search displays. Vision Res, 46(12), 1886-1900.
Russell, B. C., Torralba, A., Murphy, K. P., \& Freeman, W. T. (2008). LabelMe: A database and web-based tool for image annotation. International Journal of Computer Vision, 77, 157-173.
Tatler, B. W., Baddeley, R. J., \& Gilchrist, I. D. (2005). Visual correlates of fixation selection: effects of scale and time. Vision Res, 45(5), 643-659.

Torralba, A., \& Oliva, A. (2003). Statistics of natural image categories. Network, 14(3), 391-412.
Torralba, A., Oliva, A., Castelhano, M. S., \& Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychol Rev, 113(4), 766-786.
Wang, H.C. \& Pomplun, M. (2012). The attraction of visual attention to texts in real-world scenes. Journal of Vision, 12(6):26, 1-17,

\title{
Framing effects in evaluation of accuracy of others' predictions
}

\author{
Saiwing Yeung (saiwing.yeung@gmail.com) \\ Institute of Education, Beijing Institute of Technology, China
}

\begin{abstract}
Most predictions can be partitioned into two components: the predicted outcome, and the chance that one considers the outcome will happen. We studied how people evaluate predictions with binary outcomes. These predictions can be conveyed in two equivalent ways: one predicting an outcome with some probability, and the other predicting the other outcome with the probability of the complement of the first outcome. Although these two ways of stating the predictions are mathematically interchangeable, we hypothesized that people would judge the congruently stated prediction, one that has the same qualitative component as the actual outcome, as more accurate. We tested this hypothesis in four experiments. Results suggested that this effect is consistent across a number of domains; depends on the frame in which the prediction is stated; is robust regardless of whether the ratings were elicited in positive or negative terms; holds for both rating and choice tasks.
\end{abstract}

Keywords: framing effects; probabilistic judgment; decision making.

Probabilistic predictions are frequently encountered in everyday life. For example, weather forecasts are often made in probabilistic terms (e.g. "chance of rain is \(80 \%\) "). By comparing these statements against the actual outcomes, we can assess the predictors' skills at predicting these events. It is important to be able to accurately evaluate other people's predictions because it would then allow us to learn how good the predictors are in making predictions, to judge whether or to what degree should we trust the predictions, and to make decisions accordingly. For example, if a certain investment analyst predicts that there is a \(99 \%\) chance that Acme Company will declare bankruptcy, and that we consider this analyst to be a good predictor, then it would be advantageous to sell stocks of Acme Company that we are holding.

In this paper, we focus on one particular aspect of evaluating predictions - how framing of predictions affect people's evaluations. Framing effect is an extremely well-researched topic and has led to numerous scholarly work. It refers to a phenomenon in which people's judgment, decisions, and actions are influenced by frames, or presentation of information and its context.

Framing effects have been found to influence people in various ways in different contexts. Levin, Schneider, and Gaeth (1998) proposed a typology that categorized them into three main types. The first type, risky choice framing effect, induces a choice reversal effect between two logically equivalent gambles (Tversky \& Kahneman, 1981). In a prototypical setup, participants see one of the two gambles: either choosing between a sure gain and a risky gain, or choosing between a sure loss and a risky loss. Previous research has found that a majority of the people would prefer the sure gain choice in the gain condition, and risky loss choice in the loss condition.

The second type of framing effects was called attribute framing effects, as a single attribute within a given context presented in two logically equivalent frames has been shown
to change people's evaluations about the subject. For example, in Levin and Gaeth (1988), beef that was labeled as "75\% lean" was rated as better tasting and less greasy than beef that was labeled as " \(25 \%\) fat."

Goal framing effects is the third type in Levin et al.'s typology. Here negatively framed messages are found to be more persuasive than positively framed messages. Works by Meyerowitz and Chaiken (1987) demonstrated a typical setup of this problem. They found that women are more likely to perform breast self-examination (BSE) if they are told of the negative consequences of not performing BSE, compared to being told of the positive consequences of performing one.

In the present study we report a new type of framing effect, in which people's evaluation of a prediction with respect to the outcome is influenced by the frame in which the prediction is presented. We will focus on predictions in which there are clearly two possible outcomes (e.g. coin flips) and are stated with the subjective probability of said event happening (e.g. " \(80 \%\) "). Because there are exactly two outcomes, any predictions can be stated in two ways that are logically equivalent. For example, to say that there is a \(99 \%\) chance that the world will be destroyed at end of 2012 is equivalent to a \(1 \%\) chance that the world will not be destroyed at end of 2012.

We argue, however, that people evaluate these predictions differently. As demonstrated by the framing effects literature described earlier, people's judgments are often influenced by how information is presented. In the context of prediction evaluation, we suggest that people would overweight the qualitative component of the prediction (the stated outcome), relative to its quantitative component (the chance that one considers the outcome will happen). To differentiate this from previously discovered types of framing effects, we will call this probabilistic statement framing effect (PSFE). We will next describe four experiments that were carried out to investigate this hypothesized effect.

\section*{Pilot Experiment}

The main objective of the Pilot Experiment was to establish initial evidence about PSFE. To ensure the realism of the stimulus, we used a cover story about the 2012 U.S. presidential election which had just ended a few weeks prior.

\section*{Methods}

The participants were recruited using Amazon Mechanical Turk (MTurk). Only workers who were residing in the U.S., were at least 18 years old, and had a lifetime acceptance rate with MTurk of \(95 \%\) or over were allowed to participate \({ }^{1}\).

\footnotetext{
\({ }^{1}\) The same requirements applied to all experiments in this paper. Moreover, we disallowed participants from participating in more than one experiment in this paper (except for two participants who
}

In order to detect participants who might have been bored or inattentive during the experiment, an attention check (AC) was employed in the experiment (Oppenheimer, Meyvis, \& Davidenko, 2009). The AC took place before the actual experiment, and consisted of a paragraph of instruction followed by a question. The instruction began by asking participants to enter their favorite sports in the space below. However, at the end of the instruction we asked the participants to enter a different response: "To show that you have read this far, please enter candle below. To repeat, enter the word candle no matter what your favorite sports is." If participants had read the entire instruction, then they should have responded with the target word ("candle"). The other experiments in this paper employed AC's with exactly the same format with the exception of different target words.

The key content of the experiment would be next. The instructions were as follows, with the conditions marked by parentheses, and the differences between conditions marked by double brackets ( \(\llbracket\) and \(\rrbracket)\) and vertical lines \((\|\|)\) :

Acme inc. is a company that conducts public opinion polls about the 2012 presidential election between Barack Obama and Mitt Romney. Before the election it had predicted that \(\llbracket\) (congruent) Mitt Romney had a 20 percent chance of winning \(||\mid\) (incongruent) Barack Obama had an 80 percent chance of winning 】.
All participants were then asked "If Romney had won, was Acme inc. wrong?" in a forced-choice question. The two conditions in this experiment represented the different ways in which predictions were framed. In the congruent condition, the qualitative component of the prediction was the identical to the hypothetical result stated in the stimuli (Romney winning), whereas it was the opposite in the incongruent condition.

We then asked participants to rate the prediction using a 9point Likert scale on "How accurate was the prediction?" and "How useful was the prediction?" The participants then filled in a demographics survey, which included a question about their political orientation.

\section*{Results}

There was a total of 93 responses. Eleven of them failed the attention check question and their data were discarded. Out of the resulting 82 data points, \(56.1 \%\) were female, \(81.7 \%\) had at least some college education. We recorded age information in brackets. Almost half of the participants were in the youngest bracket of under 25 ( \(48.8 \%\) ), but there were also significant portion of the participants in older brackets ( \(22.0 \%\) between 26 and \(35 ; 18.3 \%\) between 36 and \(50 ; 11.0 \% 51\) or over).

We first examined the forced-choice question on whether the participants regarded the prediction as wrong. Relatively fewer participants in the congruent group rated the prediction as wrong \((16 / 40=40 \%)\) than in the incongruent group \((22 / 42=52.4 \%)\). However, the differences were not significant \(\left(\chi^{2}(1, N=82)=1.26, p=0.26, \phi=0.12\right)\).
participated in two experiments because of a programming error). This ensures a broader representativeness of our samples.

The congruent group rated prediction accuracy ( \(M=\) 4.58 , s.d. \(=2.21\) ) significantly higher than the incongruent \(\operatorname{group} \operatorname{did}(M=3.17\), s.d. \(=1.83 ; t(80)=3.15, p<0.01\), Cohen's \(d=0.71\) ). The congruent group also rated prediction usefulness \((M=4.30)\) higher than the incongruent group \(\operatorname{did}(M=3.45)\), although the difference was only marginally significant \((t(80)=1.75, p=0.09\), Cohen's \(d=0.39)\).

As the stimuli in this experiment involved a question in politics, we also tested whether subjects' political orientation influenced their responses. There were more self-reported Democrats than Republicans, with 22 (26.8\%) self-identified as strongly Democrat and 35 (42.7\%) as moderately Democrat. Nonetheless, the participants' political orientations had a low correlation with their evaluation of accuracy at \(r=\) 0.084 and was insignificant \((t(80)=0.75, p=0.45)\).

\section*{Discussion}

In this experiment we found initial evidence supporting PSFE: Participants in the congruent frame rated the prediction as more accurate, although they did not consider the prediction less wrong. Nonetheless, there remains a number of unresolved issues. The two conditions represent differences at multiple attributes, including the prediction frame (whether the prediction was described in terms congruent with the actual result), framing of the result (whether the results were described using the same agent as the prediction frame), and valence of the evaluation (whether the evaluation is elicited in positive or negative terms). It remains to be established which of these attributes underlie this phenomenon. Moreover, the scenario used was based on a real event and this might have interfered with people's reasoning, especially because most of our participants self-identified as liberal. Therefore, we conducted the next two experiments to tease apart the pathways involved in bringing about this phenomenon.

\section*{Experiment 1}

The first objective of Experiment 1 was to investigate PSFE using an artificial cover story in which the participants do not have a preference towards one of the two possible outcomes. The second objective was to investigate whether PSFE is driven by the prediction frame or result frame.

\section*{Methods}

Participants were again recruited from MTurk. The experiment used a between-subject \(2 \times 2\) design, crossing the prediction frame and the result frame. The stimuli in this experiment used the cover story of a college (American) football game. The stimuli were as follows:

Imagine that you have just arrived a little early for a new class on the first day of the semester. Another student was already there. The two of you started talking and the conversation turned to an upcoming college football game between universities A and B. The other student predicted that \(\llbracket\) (Prediction frame: congruent) University B has a \(30 \%\) chance of winning \(\| \mid\) (Prediction frame: incongruent) University A has a 70\% chance of winning \(\rrbracket\).

The game took place later that week and \(\llbracket\) (Result frame:
A) University A lost to University B \(\|\|\) (Result frame: B) University B defeated University A】.
The conditions in the two prediction frames are so named because if we ignore the confidence levels in the predictions, the prediction in the congruent prediction frame ( B winning) is congruent with the result ( B won in all conditions in this experiment), while the prediction in the incongruent prediction frame (A winning) is incongruent. The conditions in the result frames are simply named after the agent in the frame.

The participants were then asked to state whether the predictions wrong, and how accurate was the prediction ( 9 -point Likert scale). Finally the participants answered a demographics survey similar to that in the Pilot.

\section*{Results}

There were a total of 112 participants ( \(41.1 \%\) female), after discarding data from eight others for failing the attention check \((6.7 \%)\). Average age was \(28.62(\) s.d. \(=11.95)\) and \(84.8 \%\) had at least some college education.

The main objective of Experiment 1 was to investigate whether the prediction frame or the result frame is driving the PSFE, and whether there is interaction. To examine the effect of the prediction frame, we performed a \(t\)-test to compare the evaluation of prediction accuracy between the two prediction frames. The mean rating in the congruent prediction frame was 4.91 ( \(s . d .=1.79\) ), higher than that of the incongruent prediction frame at \(3.34(s . d .=1.47)\), and the difference was significant \((t(110)=5.08, p<0.01\), Cohen's \(d=0.97)\). This replicated the results from the Pilot.

In the Pilot, there was no significant difference between the two conditions in whether participants consider the predictions were wrong. Interestingly, this was significant in Experiment 1 , in which 12 of 56 ( \(21.4 \%\) ) participants in the congruent condition judged the prediction as wrong, compared to 29 of \(56(51.8 \%)\) of those in the incongruent condition did so \(\left(\chi^{2}(1, N=112)=11.12, p<0.01, \phi=0.32\right)\).

One alternative hypothesis is that the differences were caused by the different result frames. We found the mean accuracy ratings to be \(4.25(s . d .=1.96)\) for frame A and 4.00 \((s . d .=1.66)\) for frame B, respectively. There were no significant difference \((t(110)=0.74, p=0.46\), Cohen's \(d=0.14)\). For the question on prediction wrong-ness, there were no significant difference between different result frames either \(\left(\chi^{2}(1, N=112)=0.20, p=0.66, \phi=0.04\right)\). Moreover, there were no interaction between prediction framing and result framing \(\left(F(1)=0.32, p=0.57, \eta^{2}=0.00\right)\). Figure 1 plots the results from Experiment 1.

Another alternative hypothesis is that having the same agent in the prediction frame and result frame would lead to higher accuracy ratings. We found this to not be the case. Mean accuracy ratings for participants who had the same agent in both frames was lower (4.05, s.d. \(=1.69\) ) than those with different agents \((4.20\), s.d. \(=1.94)\), and the differences were not significant \((t(110)=0.43, p=0.67\),


Figure 1: Results of Experiment 1. Each graph plots a comparison of accuracy ratings for a different factor. Error bars represent s.e.

Cohen's \(d=0.08\) ). There were no significant differences in the forced-choice question either \(\left(\chi^{2}(1, N=112)=0.00\right.\), \(p=0.96, \phi=0.01\) ).

Self-reported football knowledge was evenly spread over the 4 -point scale. There were 32, 33, 23, and 24 responses, from the least knowledgeable to the most knowledgeable. To investigate whether there is an interaction between football knowledge and prediction frame, we carried out an ANCOVA analysis. The results indicated that there was no significant interaction \((F(1)=0.696, p=0.41)\).

\section*{Discussion}

Results of Experiment 1 suggested that framing of predictions significantly changes people's evaluation of predictions, whereas framing of results and whether the same agent is used in both frames has little effect. This not only replicated the results of the Pilot, but also suggested that prediction frame is what underlies the difference in how people evaluate how accurate predictions are. The results of this experiment correspond to the compatibility effects (Slovic, Griffin, \& Tversky, 1990), which states that stimuli attribute that is compatible with the response mode would be overweighted.

\section*{Experiment 2}

In both the Pilot and Experiment 1, the forced-choice question on evaluations were elicited in negative terms, i.e. we asked the participants whether the predictions were wrong. Therefore in Experiment 2 we tested whether PSFE also holds when the evaluations had a positive valence.

\section*{Methods}

Participants were again recruited from MTurk and the procedures were mostly the same as the previous two experiments. The instructions were:

Imagine that you have just arrived a little early for a new class on the first day of the semester. Another student was already there. The two of you started talking and the conversation turned to an upcoming college football game between universities \(A\) and \(B\). The other student predicted that \(\llbracket\) (congruent) University B has a 30\% \|| (incongruent) University A has a \(70 \%\) § chance of winning.

The game took place later that week and \(\llbracket\) (congruent) University B defeated University A ||| (incongruent) University A lost to University B \(\rrbracket\).

As can be seen from the instructions, there were two conditions: congruent and incongruent. The major departure of Experiment 2 from the previous two lies in how we elicited the forced-choice response on about the prediction: we asked "Was the prediction made by the other student right?"

Note that in both condition, the agent remains the same in both the prediction frame and the result frame.

\section*{Results}

Experiment 2 had a total of 78 participants ( \(34.6 \%\) female), after discarding data from nine of them for failing the attention check \((10.3 \%)\). Mean age was \(29.03(s . d .=12.73)\) and \(88.5 \%\) had at least some college education.

The main objective of this experiment was to test whether PSFE could be replicated when evaluations were elicited in positive terms. We first analyzed results of the forced-choice question in which the participants were asked whether the prediction was right. In the congruent condition, 22 of 40 ( \(55.0 \%\) ) responded affirmatively; whereas in the incongruent condition, 8 of 38 participants ( \(21.1 \%\) ) responded affirmatively. \(\chi^{2}\)-squared test showed that the difference was significant \(\left(\chi^{2}(1, N=78)=9.49, p<0.01, \phi=0.35\right)\). The quantitative accuracy ratings for the two conditions reflected a similar picture. The mean accuracy rating for the congruent condition was \(5.05(s . d .=2.06)\), compared to that of the incongruent condition of \(3.39(s . d .=1.72)\). The difference was significant \((t(76)=3.84, p<0.01\), Cohen's \(d=0.88)\).

\section*{Discussion}

Experiment 2 focused on whether PSFE holds when the people are asked to evaluate the predictions in positive terms. Results indicated that this is indeed the case, suggesting that PSFE to be robust regardless of the valence in which evaluations were elicited.

\section*{Experiment 3}

The first three experiments in this paper demonstrated that when people give accuracy ratings to predictions, predictions presented in a congruent frame as the actual result would be rated as more accurate. Experiment 3 investigated whether this phenomenon could be extended to choice tasks - when the two frames (congruent and incongruent) are presented at the same time as two choices and people are asked to judge which one is the more accurate one.

We also tested two factors that might shed light on the mechanism of PSFE. First, one potential reason that people rated predictions in the incongruent condition as less accurate might have been that the quantitative components of these predictions involve higher numerical probabilities (compared to those in the congruent condition), and this might have been perceived as being overconfident, which in turn led to participants down-adjusting their accuracy ratings. Second, many prior works have suggested that numeracy plays an important role in judgment and decision making. For example, Peters et al. (2006) found that participants who are higher in numeracy are less susceptible to attribute framing effects. To investigate
the influences of these two factors, we also assessed perception of overconfidence and participants' numeracy.

\section*{Methods}

Similar to the previous three experiments, all participants were recruited through MTurk. However, because this experiment is slightly longer than the previous three, we increased the reward from US \(\$ 0.15\) to US \(\$ 0.20\).

In the previous experiments, predictions in the two frames were given logically equivalent probability estimates (e.g. \(75 \%\) vs. \(100 \%-75 \%=25 \%\) ). However, in Experiment 3 the participants would see both frames side-by-side, and therefore such a setup might seem contrived. Moreover, we wanted to test whether the congruent frame would be favored even when it is logically inferior. Hence we parameterized the congruent frame with a probability estimate of \(15 \%\) (in the direction of the actual result), and the incongruent frame with \(80 \%\) (opposite the direction of the actual result). The congruent frame is now logically superior because it predicts the outcome that turns out to be correct with \(100 \%-80 \%=20 \%\) confidence, compared to \(15 \%\) in the congruent frame. Additionally, in order to make the scenarios more realistic, we added two detractor predictions to each option. The instruction for one of the conditions was as follows:

Imagine that you are an analyst at an investment firm. Currently you are evaluating predictions made a year ago by two of your subordinates concerning a technology company called Acme Corp.
Analyst A predicted that in the coming year:
- Acme would buy out their supplier SuperTech Company.
- Acme would expand into the European Union.
- There was an \(80 \%\) chance that Acme would become a public company.
Analyst B predicted that in the coming year:
- Acme would license crucial technology patents from their competitor CompX Company.
- Acme would build another manufacturing plant within the U.S.
- There was a \(15 \%\) chance that Acme would not become a public company.

The probabilistic prediction shared by both analysts was whether Acme would become public or not. Each of the two analysts also made two detractor predictions additionally.

The participants then read about what actually happened. There were five total predictions: two unique detractors from each analyst, plus the common prediction. In all conditions, Acme would not become public. However, one of the two detractors from each of the analysts would come true.

In this counter-balance condition shown above, Analyst A predicted that there was an \(80 \%\) chance of the target event (Acme became a public company) happening. Analyst B, in contrast, predicted that there was a \(15 \%\) chance of the target
event not happening. If probabilistic statements could be inverted algebraically, it would mean Analyst B predicted that there was an \(85 \%\) chance of the target event happening. As the target event did not happen, Analyst A should be evaluated as being more accurate, if prediction frames have no influence on people's judgment.

There were two counter-balancing conditions. First, the order of the congruent and incongruent options was randomized between subjects. Second, the detractors that came true were counter-balanced. For roughly half of the participants the supplier buy out and new U.S. manufacturing plant turned out to be true, while for the other half it was the opposite.

We then asked participants "Which analyst do you think made the better predictions?" and "Which analyst do you think was more confident about the predictions?" This was followed by a memory test. We asked the participants to indicate whether each of the five events happened in the actual outcome. Then to investigate the influence of participants' numeracy on their judgments, we added the 8 -item abbreviated numeracy scale from Weller et al. (2012). After the numeracy section, participants answered a few demographics questions, including two questions about their level of knowledge concerning stock trading and technology.

\section*{Results}

There were a total of 85 participants ( \(60 \%\) female; one declined to self-identify). We discarded data from 29 (25.4\%) participants: 27 for failing the AC and 2 for leaving over \(80 \%\) of the answers blank \({ }^{2}\). Mean age was \(33.1(s . d .=12.87)\) and \(87.1 \%\) had at least some college education.

The portion of workers who failed the AC was higher than the previous experiments. We ran a 4 (experiment) \(\times 2\) (number of AC pass/failure) \(\chi^{2}\)-squared test of independence and the results were significant \(\left(\chi^{2}(3, N=414)=19.33, p<0.01\right.\), \(\phi=0.22)\). However, there was no a priori reason to suspect that the workers in this experiment were different from those in the previous ones. In all four experiments, the AC was the second question in the entire experimental procedure, after only the question that elicited their MTurk ID. Therefore up to the AC , the experimental procedures of all four experiments were essentially the same. The monetary reward was the only difference between this experiment (US\$0.20) and the previous ones (all three at US\$0.15). However, Mason and Watts (2009) have found that financial incentives do not significantly impact the quality of MTurk experiments, even for amounts that differ by as much as 10 times. To further confirm the quality of the data, we checked the result of the memory test. The range of the memory score was from 0 to 5 (remembered perfectly). The mean memory score across all participants was 4.25 , indicating that the participants remembered the details of the experiment well. Hence, we attribute the high AC failure rate to coincidence.

The main objective of this experiment was to test whether the PSFE could be extended to a choice task. More partic-

\footnotetext{
\({ }^{2}\) No other participants left more than one of the nondemographic answers blank.
}
ipants ( \(56 ; 65.9 \%\) ) chose the analyst in the congruent condition \((15 \%)\) as more accurate, compared to the one in the incongruent \((80 \%)\) condition ( \(29 ; 34.1 \%\) ). A \(\chi^{2}\)-squared test indicated that it was significantly different from chance \(\left(\chi^{2}(1, N=85)=8.58, p<0.01, \phi=0.32\right)\).

We then examined whether perception of overconfidence was related to PSFE. There were 35 ( \(41.2 \%\) ) and 50 ( \(58.8 \%\) ) participants who judged the congruent and incongruent option, respectively, as more confident. The result was close to reaching significance \(\left(\chi^{2}(1, N=85)=2.65, p=0.10\right.\), \(\phi=0.18\) ). This suggests that perception of predictors' overconfidence might play a small part in this effect and deserves further investigation.

The order of presentation had a big effect on choice. In conditions where the incongruent option was presented first, there were about the same number of participants who chose the congruent option \((N=21)\) as those who chose the incongruent option \((N=20)\) as more accurate. However, if the congruent option was presented first, 35 (vs. 9) participants judged the congruent option as more accurate. The interaction was significant \(\left(\chi^{2}(1, N=85)=7.58, p<0.01\right.\), \(\phi=0.30\) ). This suggests that order of presentation significantly influenced evaluation of accuracy. However, order of presentation did not have a significant effect on evaluation of confidence \(\left(\chi^{2}(1, N=85)=0.24, p=0.62, \phi=0.05\right)\). The other counter-balancing condition - which pair of distractors turned out to be correct - had no significant effect on evaluation of accuracy ( \(p=0.37\) ) nor confidence ( \(p=0.79\) ).

We also investigated the effect of numeracy on people's judgments. As there are eight questions in Weller et al.'s numeracy scale, the range of the numeracy scores is from 0 to 8. No participant answered the mammogram question correctly. In fact, no answer came within 3 percentage point of the correct answer. This is not surprising because this question has been found to be a very difficult question (see Weller et al., 2012). The percentage of participants who answered each question correctly (Table 1) was in fact quite close to the result obtained by Weller et al. (2012). This suggests that the numeracy and motivation of the participants in this experiment were comparable to those in their experiment. This result also partly mitigated the concern raised by the high percentage of participants failing the AC.

The mean (and s.d.) of the numeracy score for participants who chose the congruent or incongruent options as more accurate were 4.43 (1.45) and 5.41 (1.45), respectively. We fitted a logistic model using the numeracy score as the independent variable, and participants' choices as dependent variable. Results indicated that the influence of numeracy was significant ( \(\beta=0.75, z=2.74, p<0.01\) ). This suggested that participants who were lower on numeracy are more likely to consider the analyst in the congruent option - the normatively less accurate of the two - the more accurate predictor.

The effect of self-reported knowledge about stock trading and technology on choice of more accurate prediction was not significant in either case ( \(p=0.80\) and \(p=0.56\) ).

Table 1: Percentage of participants correctly answering each item of the numeracy scale in Experiment 3 (E3), compared to the results from Weller et al. (2012).
\begin{tabular}{rrrrrrrrr}
\hline & Q1 & Q2 & Q3 & Q4 & Q5 & Q6 & Q7 & Q8 \\
\hline E3 & 0.0 & 27.1 & 45.9 & 57.6 & 80.0 & 81.2 & 96.5 & 88.2 \\
W & 11.0 & 26.5 & 39.8 & 42.1 & 60.4 & 70.2 & 73.8 & 84.8 \\
\hline
\end{tabular}

\section*{Discussion}

The key objective of this experiment was to investigate whether PSFE would hold in a choice task. We also put PSFE to a stronger test because the congruent option was presented vis-à-vis a logically superior option. Our results found that significantly more participants would choose the congruent option, suggesting the robustness of PSFE. We also found that perception of overconfidence did not explain PSFE. However, numeracy was found to be a moderating factor. Like Peters et al. (2006), we found that people who are higher in numeracy to be less susceptible to framing effects.

\section*{General Discussion}

We proposed a new phenomenon, probabilistic statement framing effect (PSFE), that occurs when predictions made in congruent frames (relative to eventual outcomes) are judged as more accurate, compared to logically equivalent or even superior predictions made in incongruent frames. Across four experiments, we found that this effect holds regardless of real world based event (Pilot Experiment) or hypothetical events (Experiments 1 to 3), and rating (Pilot and Experiment \(1 \&\) 2) or choice (Experiment 3) task. The effect held even when the congruent option was logically inferior (Experiment 3). Finally, we found numeracy to be a moderating factor.

The results from these experiments suggest that a majority of people do not evaluate the goodness of predictions in a normative manner. They overweight the qualitative component of a prediction while underweighting its quantitative component. This is especially true for people who are low in numeracy. The findings in this paper might have important implications in domains such as personal finance, medical decision making, and corporate strategic planning.

Among the three major types of framing effects, PSFE might be most closely related to the attribute framing effects. However, we argue that it is distinct for one major reason. Levin et al. (1998) demonstrated that attribute framing effects occurs because positive frames evoke favorable associations in memory; and vice versa for negative frames. However, PSFE can favor evaluations of negative frames (e.g. losing a game in sports), as long as the predictions are congruent to the outcome. This cannot be explained using the above framework and therefore we suggest that PSFE should be regarded as a separate phenomenon.

Although the effect seems to be robust across a broad range of conditions, its causal mechanism and cognitive processes are not well understood. Moreover, prior research has suggested that important personal decisions are less influenced by frames (Marteau, 1989). We are currently examining what
roles information leakage (Sher \& McKenzie, 2006), selective attention (Levin, 1987), and encoding of information (Levin \& Gaeth, 1988), might play in relation to this effect.

All experiments here have been carried out through MTurk. This enabled us to collect data from a subject pool more diversified than one that of a university sample. Moreover, MTurk has been found to be able to yield high quality data (Buhrmester, Kwang, \& Gosling, 2011), and be able to replicate a number of classical findings (Crump, McDonnell, \& Gureckis, 2013). However, it might be interesting in the future to study this phenomenon in lab-based and field studies.

The findings in this paper demonstrate the psychological impact of prediction frames on how people evaluate predictions with respect to outcomes. When predictions are described in congruent frames as the eventual result, people consider them as more accurate than if they were described in incongruent frames. This observation is not captured by the previous literature on framing effects and highlights the need for a better understanding of the processes that underlies this phenomenon.
Acknowledgments. We thank the very helpful comments from four anonymous reviewers.

\section*{References}

Buhrmester, M., Kwang, T., \& Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? Perspectives on Psychological Science, 6(1), 3-5.
Crump, M. J. C., McDonnell, J. V., \& Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. PLOS ONE, 8(3), e57410.
Levin, I. P. (1987). Associative effects of information framing. Bulletin of the Psychonomic Society, 25(2), 85-86.
Levin, I. P., \& Gaeth, G. J. (1988). How consumers are affected by the framing of attribute information before and after consuming the product. Journal of Consumer Research, 15(3), 374-378.
Levin, I. P., Schneider, S. L., \& Gaeth, G. J. (1998). All frames are not created equal: A typology and critical analysis of framing effects. Organizational Behavior and Human Decision Processes, 76 (2), 149-188.
Marteau, T. M. (1989). Framing of information: Its influence upon decisions of doctors and patients. British Journal of Social Psychology, 28(1), 89-94.
Mason, W., \& Watts, D. J. (2009). Financial incentives and the performance of crowds. In Proceedings of the ACM SIGKDD workshop on human computation (Vol. 15, pp. 77-85).
Meyerowitz, B., \& Chaiken, S. (1987). The effect of message framing on breast self-examination attitudes, intentions, and behavior. Journal of Personality and Social Psychology, 52(3), 500-510.
Oppenheimer, D. M., Meyvis, T., \& Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. Journal of Experimental Social Psychology, 45(4), 867-872.
Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., \& Dickert, S. (2006). Numeracy and decision making. Psychological Science, 17(5), 407-413.
Sher, S., \& McKenzie, C. R. M. (2006). Information leakage from logically equivalent frames. Cognition, 101(3), 467-494.
Slovic, P., Griffin, D., \& Tversky, A. (1990). Compatibility effects in judgment and choice. In Insights in decision making: A tribute to Hillel J. Einhorn (pp. 5-27). University of Chicago Press.
Tversky, A., \& Kahneman, D. (1981). The framing of decisions and the psychology of choice. Science, 211(4481), 453-458.
Weller, J. A., Dieckmann, N. F., Tusler, M., Mertz, C. K., Burns, W. J., \& Peters, E. (2012). Development and testing of an abbreviated numeracy scale: A Rasch analysis approach. Journal of Behavioral Decision Making. doi: 10.1002/bdm. 1751

\title{
Cognitive Science, Aesthetics, and the Development of Taste
}

\author{
William W. York (wwyork @indiana.edu) \\ Indiana University \\ Center for Research on Concepts \& Cognition \\ 512 N. Fess Avenue \\ Bloomington, IN 47408 USA
}

\begin{abstract}
Aesthetics and the arts have garnered more attention within cognitive science in recent years. Despite this increasing interest, "scientists of art" often focus on one of two areas: the formal properties of artworks themselves, or the mental processes involved in perceiving these works in an isolated, one-on-one encounter. In this paper, I review some representative examples of such work before suggesting some alternative ways that cognitive science might approach aesthetics and the arts-ways that would complement the isolationist approaches that have predominated to this point. In doing so, I draw on the observations and arguments of various philosophers of art, highlighting some of the socially and culturally situated factors that are important in shaping the development of our taste and sensibilities.
\end{abstract}

Keywords: Aesthetics; Culture; Social Cognition; Art.

\section*{Introduction}

What can cognitive science tell us about aesthetic experience? Given the putative aims of cognitive scienceroughly, "to provide a cogent scientific account of how human beings achieve their most remarkable symbolic products" (Gardner, 1987, p. 391)—artistic and aesthetic phenomena fit within the scope of the field. While they have long been fringe topics within the field, they have been garnering increasing attention in recent years. This work ranges from so-called neuroaesthetics (Ramachandran \& Hirstein, 1999), which seeks to uncover the evolved neural underpinnings of our aesthetic responses, to computational aesthetics (Hoenig 2005), which employs sophisticated mathematical tools to analyze the formal properties of various aesthetic objects. Meanwhile, others have focused more on the representational and/or computational processes involved in perceiving and appreciating works of art

Despite this recent interest, there remains a deep-seated tension between the aims of science, which prizes generality, laws, and quantification; and the arts, which we experience qualitatively through encounters with particular works (songs, paintings, films, etc.). Furthermore, these experiences take place amid a complex background of social, cultural, and historical influences. Thus, one could argue that the goals of science are simply incompatible with the kind of understanding we seek when it comes to aesthetics and the arts. Such misgivings have been voiced by many philosophers (Dickie, 1962; Morgan, 1950; Wittgenstein, 1967) and even some psychologists (e.g., Arnheim, 1991). Others view these misgivings as stubbornly anti-science, insisting that the problems faced by
"scientists of art" are merely very difficult, not fundamentally intractable or ill-conceived. If the results of their efforts have been meager, they argue, it is because of this difficulty, together with the fact that it's still early-and after all, one must start somewhere (cf. Berlyne, 1971; Birkhoff, 1932; Meyer, 1957; Rigau, Feixas \& Spert, 2008).

Instead of trying to resolve this longstanding debate, I want to focus on the picture of aesthetic experience that has tended to emerge from cognitive science's encounters with the arts. According to this picture, the artwork (or other aesthetic object) is treated as an isolated stimulus, while the viewer or listener is treated as a sort of idealized receiver of the information encoded in the work. With this picture in mind, researchers typically either focus on (a) the intrinsic properties of artworks (or other aesthetic objects), as with much of computational aesthetics; or (b) on the mental processing involved in perceiving and appreciating art (Kintsch, 2012; Leder, Belke, Oeberst \& Augustin, 2004).

In this paper, I want to look more closely at some of this research, with the dual aim of showing what we can learn from it as well as what its limitations are. In keeping with the theme of the conference ("Cooperative Minds: Social Interaction and Group Dynamics"), I also want to suggest some alternative ways of approaching aesthetics and the arts-ways that would complement the isolationist approaches that have predominated to this point. In doing so, I draw on the observations and arguments of various philosophers of art and aesthetics, highlighting some of the socially and culturally situated factors that are important in shaping the development of our taste and sensibilities.

There are three main sections in this paper. The first looks at research (both recent and not-so-recent) on the perception, appreciation, and value of visual art; the second looks at some parallel work on music (in particular, on musical meaning); and the third focuses on outstanding questions and possible future directions for research.

\section*{Order, Complexity, and Value in Visual Art}

The scientific study of aesthetics date back to at least the 1870s and Gustav Fechner (cf. Arnheim 1985). However, the work of mathematician George Birkhoff's Aesthetic Measure (1933) remains an important landmark in this pursuit. Birkhoff's quest to formalize beauty yielded a succinct mathematical equation, \(M=O / C\), where \(M\) is the aesthetic measure (or value) of the stimulus in question, \(O\) is the order, and \(C\) is the complexity. This equation was thought to crystallize Fechner's notion of "unity in variety" while providing a "logical tool in order to answer aesthetic
questions by purely mathematical (logical) reasoning" (p. 46). Birkhoff asked, and sought to answer, questions such as, "Which is the most beautiful of all polygonal forms?"

Birkhoff's approach was grounded in two key assumptions. First, it is assumed that the formal properties of the aesthetic object (e.g., symmetry, equilibrium, number of components, etc.) can be isolated from its connotative (i.e., referential or associative) properties. Second, he believed that the same kinds of methods employed in simplified domains (e.g., geometric forms) could be applied to other, more complex domains such as visual art, poetry, and music. The difference between the two was thought to be merely one of degree, not of kind.

\section*{Optimal Complexity, Pleasure, and Arousal}

Birkhoff's conception of aesthetic measure influenced subsequent efforts by psychologists (e.g., Eysenck 1942) and information theorists (e.g., Moles 1966). The latter would recast Birkhoff's order and complexity as redundancy and entropy, respectively. The informational and psychological approaches were brought together by Berlyne (1971). Berlyne conceptualized the link between the two in terms of the Wundt curve (Fig. 1). The idea was that people prefer stimuli of moderate-but not excessivenovelty and complexity. The greater the stimulus complexity, the greater the arousal potential, which in turn correlated with a more pleasurable aesthetic experience, so long as the subject was not overwhelmed by the stimulus. Berlyne acknowledged that "what constituted novelty and complexity would vary from person to person" (Margulis \& Beatty, 2008, p. 66), but maintained "that his adapted Wundt curve could apply to both of them; it would simply shift along the x -axis to reflect the experience level" of the perceiver (ibid.).


Figure 1: The Wundt curve. From Margulis and Beatty (2008); adapted from Berlyne (1971).

Surely, there is some truth to this sort of "Goldilocks theory" of complexity and optimal arousal. However, there is also a clear tradeoff between (1) the level of generality sought in such a theory and (2) the degree of fidelity one
would hope for in a genuinely enlightening account of musical experience. In order to achieve the latter, it becomes necessary to reincorporate just those factors-listener background, experience, personality, mood, et cetera-that must be subtracted out in order for the information-theoretic approach to get off the ground. What I want to suggest is that the factors influencing this "shift along the x -axis" are where much of the interest lies. (I will return to this point below.)

\section*{From Aesthetic Measure to Computational Aesthetics}

Computational aesthetics (Hoenig, 2005) is the most recent offspring of Birkhoff's aesthetic measure and its subsequent reformulations in terms of information theory. Various researchers have picked up on these threads, including Koshelev, Kreinovich, and Yam (1998), who recast aesthetic measure as a joint function of (a) the length of the shortest program required to generate a given visual design and (b) the running time of this program; and Machado and Cardoso (1998), who recast it as0020the ratio between image complexity and processing complexity.

More recently, Rigau, Feixas, and Sbert (2008) created several reformulations of Birkhoff's aesthetic measuremost notably, as the ratio between algorithmic reduction of uncertainty and initial information content, which correspond, respectively, to order and complexity. Essentially, this ratio measures "the degree of order created from a given palette" (2008, p. 131), with the "palette" construed as "the range of colors selected by the artist with a given probability distribution" (p. 128). This and other measures are applied to paintings by Mondrian, Pollock, and van Gogh, resulting in is a series of rank-orderings that, not surprisingly, show Mondrian's works to possess a higher degree of order than those of the other two painters.

These proposed metrics are intended to "help us ... quantify the aesthetic experience" (p. 124), but it is questionable what we are to make of them. Are these formulas being proposed as measures of aesthetic value (in which case Mondrian trumps van Gogh and Pollock)? Would these metrics be able to discern the difference between a genuine Pollock and an imitation, or between a Mondrian painting and some generic arrangement of primary-colored geometric forms? Do the authors themselves draw a distinction between artistic value and mere pleasantness? It is not quite clear what lessons we are to take from this work.

\section*{Meaning, Information, and Entropy in Music}

The quest to formalize aesthetic value in the visual arts has parallels in the attempt to quantify meaning in music. The work of Leonard Meyer (1957) is a touchstone here. Meyer linked musical meaning to expectation, uncertainty, probability, observing that "the rules of musical grammar and syntax found in textbooks on harmony, counterpoint, and theory in general" are "almost invariably stated in terms of probability" (p. 414). Conceived of in this light, the
"meaning" of a musical event-a note, a chord, or a phrase-is inversely proportional to its probability: low probability events (such as the sounding of a \(D b\) in the key of C ) are more surprising and thus more meaningful. Meyer summarized, "Both meaning and information are thus related through probability to uncertainty" (p. 416). The greater the probability, the lower the information, or entropy.

By operationalizing musical meaning in terms of information theory, Meyer lent the former a newfound precision. However, the tradeoff is that this precise characterization does not intuitively capture what we typically mean when we talk about "meaning," whether in music or in more general terms. "So much for intuitions," one might reply-except that Shannon and Weaver themselves warned against conflating information (in the information-theoretic sense of the term) with meaning. For example, Weaver (1949) stressed that "the rather strange way in which, in this theory, the word 'information' is used ... must not be confused at all with meaning" (p. 12). He added, "It is surprising but true that, from the present viewpoint, two messages, one heavily loaded with meaning and the other pure nonsense, can be equivalent as regards information" (ibid.).

Similar objections were raised by Vermazen (1971) and Sherburne (1966) in direct response to Meyer. One objection, later referred to as the "Information Theory Paradox" (cf. Titchener \& Broyles, 1973), holds that if meaning were tied to uncertainty, then repeated listening to the same piece would yield less and less "meaning" each time-which surely runs counter to experience. The second objection is that the most meaningful music would be that in which all of the musical events within the piece were equally likely to occur at any moment.

\section*{Cultural Noise and Distance}

Meyer anticipated some of the aforementioned objections in his 1957 paper. In order to circumvent them, he appealed to the related notions of cultural distance and cultural noise. Writing during the heyday of serialism, Meyer acknowledged the general public's disdain for modern classical music: "Here 'noise' is the result of a time-lag between the habit responses which the audience actually possesses and those which the more adventurous composer envisages for it" (p. 420). He added that "in their zeal to 'pack' music full of meaning some contemporary composers have perhaps so over-loaded the channel capacity of the audience that one meaning obscures another in the ensuing overflow" (p. 420). Thus, too much meaning (in the information-theoretic sense) can essentially render a work meaningless (in the pre-theoretic sense), at least to lay audiences.
Cultural noise and distance are also invoked to explain why audiences struggle to make sense of music from unfamiliar cultures. In a nutshell, "[T]he more distant a culture is from our present set of habit responses, the greater the amount of cultural noise involved in communication" ( p .
420). While there is truth to this statement, it overlooks that (a) the listener always brings something to the table, even if it is simply a lack of familiarity with the musical style in question; and (b) there is always some degree of distance between listener and work. Meyer's way of factoring out this distance was to take for granted the notion of an "Ideal Auditor"-that is, someone who "knows the style of the piece and the styles of the period and thus has an experiential basis for the expectations which Meyer's theory requires" (Titchener \& Broyles 1973, p. 17). But how do we come to know the style of a piece or the style of a period? As with the confounding factors that caused Berlyne's modified Wundt curve to shift along the x -axis, these factors are worthy of exploration in their own right.

\section*{Meyer Rehabilitated? Huron's Sweet Anticipation}

The most thorough and ambitious attempt to bring Meyer's work up to date can be found in Huron (2006). Huron maintains Meyer's emphasis on listener expectations-and the ways in which they are "exploited" by composers-as the key to a systematic understanding of how music works on the mind/brain. One could debate the "composer as manipulator" characterization that emerges throughout this work, along with the idea that the chief aim of music is to evoke specific emotions in listeners. However, I will instead look briefly at Huron's effort to incorporate cultural context into an account that is otherwise rooted in evolutionary psychology, statistical learning, and information theory.

A specific example comes from a study that compared Balinese and American musicians' predictions of successive notes in a melodic line. The melody was composed in a \(10-\) tone Balinese scale but was unfamiliar to participants in each group. Huron and his associates found that while the Americans performed better than chance, they were outpaced by their Balinese counterparts in terms of both predictive accuracy and confidence (as opposed to uncertainty) in their guesses.
Thus, in contrast to Meyer, Huron does try to account for "cultural noise," or at least one aspect of it. However, we should keep in mind that cultural background plays a more significant role in musical understanding and experience than merely imparting a set of statistical expectations for melodic or harmonic development. For example, what does it really mean to be an "American musician" (or "American listener")? Of course, there are certain melodic, harmonic, and rhythmic norms that most American (and, more generally, Western) listeners are accustomed to. However, underneath the broad umbrella of "Western music," there is a vast array of musical subcultures-pop, jazz, punk, classical, rap, electronic, noise, drone, Tin Pan Alley-most of which can be further subdivided into sub-subcultures. Each subculture (or subgenre) has its own norms, its own aesthetic values. To know a genre (or subgenre, or artist) extends far beyond possessing a matrix of transition probabilities of the sort used to model melodic expectation.

Another way to put it is this: Huron's emphasis on generalities-transition probability matrices, statistical
learning tendencies, and our (mostly) shared evolutionary heritage as human beings-lends itself to a study of what is universal about music cognition (or at least "universal" within a particular culture). This is fine as far as it goes, but this sort of account is not going to supplant the kind of understanding that comes from engaging with particular works and understanding them in particular contextswhether that's the context of a genre, a historical period, an individual artist's work, or whatever else. There is value in the sort of research documented in Huron's book. It's just that the gains made in understanding the psychology of expectation through the study of music are likely to far exceed the gains made in understanding music via the psychology of expectation.

\section*{Sketching an Alternative Approach}

In this section, I highlight some important points and arguments from philosophers of art and aesthetics, with the goal of suggesting alternative ways for cognitive science to engage with aesthetics and the art.

\section*{Aesthetic sensibility and personal development}

Despite David Hume's (1757) ingenious arguments to the contrary, the notion of a fixed "standard of taste" is unrealistic. This is true whether we seek this standard in the form of a group of ideal critics or judges, as Hume suggested, or whether we follow the Birkhoffs of the world in searching for quantifiable measures of aesthetic value. Regardless, the lack of a fixed or objective standard doesn't stop us from seeking to improve our taste and encounter more rewarding aesthetic experiences. The development of taste and aesthetic sensibility is an ongoing process. But how do we know where to look for these more rewarding experiences as we undertake this process of developing our taste? Herwitz (2008) offers some useful suggestions, arguing that "taste is a circular and constructivist enterprise. We are led by others because they elevate our taste to their level, and this because we already have taste" (p. 52). Even so, we are left to ask how we are able to gain an initial foothold in this process.

One suggestion comes from neuroaesthetics (Ramachandran \& Hirstein, 1999), which seeks to uncover the evolutionarily hardwired tendencies that shape our preferences. Certainly, our preferences and tastes are constrained by our biological makeup, but they are not rigidly determined by them. What we know about an artwork or other aesthetic object affects our appreciation of it. This seems like a truism, but it poses problems for nativist accounts of aesthetic preference.

A vivid example comes from Saito (2010), in which the author discusses the example of a lavishly kept green lawn in Arizona. Superficially, the lawn might be visually appealing-the kind of lawn that would make any suburban homeowner jealous. Yet once we come to understand what goes into maintaining such a lawn in the middle of the desert-in particular, the burden it places on the local environment-it is likely to lose some of its appeal. It might
even be perceived as garish or tacky, in much the same way that a previously admired painting loses its luster when it turns out to be a forgery. In other words, we do not just respond automatically and passively to aesthetic stimuli. Furthermore, our differential responses to artworks and other aesthetic objects cannot be simply a matter of differences in processing fluency (think back to Berlyne's modified Wundt curve). How can we better understand the effects of such background knowledge on our aesthetic responses? This is another underexplored question for cognitive science to consider.

\section*{Getting Outside the Frame}

It is a given that scientific research must make certain simplifying assumptions in order to get off the ground. This is especially true when the subject matter is as complex as human aesthetic and artistic experience. That said, many of the assumptions taken for granted by Birkhoff, Berlyne, and their followers have been (indirectly) called into question by the work of philosophers of art, on issues ranging from originality and forgery (Dutton, 1979) to the very distinction between works of art and "mere real things" (Danto, 1992). A unifying thread among these arguments is that neither artistic value nor even an object's status as an artwork can be predicated on mere appearances-that is, by an exclusive concern with what lies "inside the frame." Here we find a basic difference between works of art and psychological stimuli such as Birkhoff's geometric forms. The latter are designed to be context-invariant, perceived and experienced in isolation; the former are not and, in fact, cannot be if they are to be genuinely understood and appreciated. As philosopher Garry Hagberg (2011) recently put it, "Art that we see or hear or read is to a large part constituted by relational interconnections." These connections involve not just other works of art, but the art world itself (Danto, 1992), as well as the broader culture in which art works (and worlds) exist.

Take the case of Mondrian, who has long been a favorite of aesthetic formalists, since his work might initially appear to consist of nothing but pure form. Even here, though, there is more to the story. Kieran (2005) describes a visit to a Mondrian exhibit in which the artist's work was presented in chronological order, allowing for an understanding of the way his style and approach evolved over the years, become increasingly abstract but always "trying to get at the underlying structure of the naturalistic world of appearances" (p. 38). Kieran adds that "unless one is concerned with what Mondrian was striving to capture and express in his artistic development, one will fail to understand and properly appreciate his art" (p. 40). Danto (1992), in his discussion of the work of avant-garde sculptor Eva Hesse, makes a similar point about the role of arthistorical (or "art world") context in criticism.

\section*{Expertise and the "Feeling for the Rules"}

In addition to knowledge about artworks, aesthetic appreciation also draws on less explicit, more tacit forms of
knowledge. This tacit knowledge can be likened to the kind of know-how that Dreyfus and Dreyfus (1988) emphasize in their five-stage model of skill acquisition, the last stage of which is expertise. It makes sense to think of the development of aesthetic sensibility within a given field or genre as tracing a similar arc of development.

In discussing Huron's research on melodic expectation, I suggested that there is more to learning a style of music than, say, internalizing a frequency distribution matrix. But I also raised the question of how we come to know a novel genre or style of music in the first place. Yes, listening is important, but trying to grasp a foreign style of music can be as bewildering as trying to learn a new language without so much as a dictionary. Wittgenstein (1967) offers some illuminating, if occasionally cryptic, hints on this process. As he argues, coming to know a style amounts to developing a "feeling for the rules"-rules that are largely social and cultural in character. This feeling for the rules, when fully "internalized," constitutes a kind of expertise. Novitz (2004) helpfully elaborates on Wittgenstein's terse remarks that "to have a 'feeling for the rules' that are embodied or instantiated in a work or a category of art is to understand the role that they play within the 'culture of a period.' It is to understand their cultural or their social significance" (p. 61). Developing this feeling for the rules goes hand in hand with overcoming the barriers Meyer spoke of in his discussion of cultural noise and cultural distance.

But why should we bother trying to overcome such barriers? One kind of argument suggests that we should try because doing so is intrinsically valuable. As Cooper (2010) puts it, "[A]ppreciation of new beauty is educative, for it requires initiation into traditions, practices and cultural contexts that allow for beauty of a certain kind to become visible" (pp. 63-64). He adds, "[T]his appreciation is an achievement or acquirement that, typically, calls for effort, imagination, and intelligence. Finally, the appreciation is, typically, edifying or improving" (p. 64). In other words, aesthetic appreciation-especially when it comes to "new beauty"-is not an automatic, facile accomplishment, but is often the product of much cognitive "work."

If, as Cooper suggests, this sort of achievement is intrinsically valuable, it seems that it is worth trying to understanding it better. What kinds of imagination and intelligence are involved? Why are some people more open to pursuing such experiences than others? What kinds of barriers-social, cultural, biological, or otherwise-prevent those others from pursuing the kind of "initiation" Cooper describes? At this point, there are more questions than answers, but I believe they are worthwhile questions for us to ask, even if doing so is bound to raise further difficult questions about the scope and methods of cognitive science.

\section*{Conclusion}

The drive to bring aesthetics and the arts under the umbrella of cognitive science is understandable. They are important aspects of our (mental) lives, so to simply ignore them
would be to limit the scope of the field to a perhaps depressing extent. On the other hand, it is still unclear how best to go about studying them. However, as I have emphasized in this paper, there are limits to what can be understood via many of the tried-and-true methods of the past. Artworks differ from garden-variety psychological stimuli in important ways, such that many of the "complicating," noise-like factors-social, cultural, and otherwise - that have traditionally been removed from the equation are actually quite important, and possibly even essential to accurately understanding the phenomena in question.

\section*{Acknowledgments}

Thank you to the Center for Research on Concepts and Cognition for funding and support. Additional thanks to Hamid Ekbia, Michel Chaouli, Matthew Kieran, Anna Ichino, and Peter Lamarque for helpful comments on related work.

\section*{References}

Arnheim, R. (1985). The other Gustav Theodor Fechner. In S. Koch \& D. Leary (Eds.), A Century of Psychology as Science. New York: McGraw-Hill.
Arnheim, R. (1991). But is it science? In To the Rescue of Art. Berkeley: Univ. of California Press.
Berlyne, D. (1971). Aesthetics and Psychobiology. New York: Appleton-Century-Crofts.
Birkhoff, G. (1933). Aesthetic Measure. Cambridge, Mass.: Harvard Univ. Press.
Cooper, D. (2010). Edification and the experience of beauty. In W. Keping, (Ed.), Diversity and Universality in Aesthetics: International Yearbook of Aesthetics, Vol. 14. Beijing.
Danto, A. (1992). Beyond the Brillo Box. Berkeley: Univ. of California Press.
Dickie, G. (1962). Is psychology relevant to aesthetics? Philosophical Review, 71(3), 285-302.
Dreyfus, H. L, \& Dreyfus, S. E. (1988). Mind over Machine. New York: Simon \& Schuster.
Dutton, D. (1979). Artistic crimes. British Journal of Aesthetics, 19, 302-341.
Gardner, H. (1987). The Mind's New Science. New York: Basic Books.
Hagberg, G. (2011). Wittgenstein, music, cultural context. London Aesthetics Forum (public lecture). Accessed from http://www.londonaestheticsforum.org/?p=1000
Herwitz, D. (2008). Aesthetics. London: Continuum.
Hoenig, F. (2005). Defining computational aesthetics. Computational Aesthetics 2005: Eurographics Workshop on Computational Aesthetics in Graphics, Visualization, and Imaging. Aire-la-Ville, Switzerland: Eurographics.
Hume, D. (1757). Of the standard of taste. In Selected Essays (1993). Oxford: Oxford Univ. Press.
Huron, D. (2006). Sweet Anticipation. Cambridge, Mass.: MIT Press.
Kieran, M. (2005). Revealing Art. London: Routledge.

Kintsch, W. (2012). Musings about beauty. Cognitive Science, 36(4), 635-654.
Koshelev, M., Kreinovich, V., \& Yam, Y. (1998). Towards the use of aesthetics in decision making: Kolmogorov complexity formalizes Birkhoff's idea. Bulletin of the European Association for Theoretical Computer Science.
Leder, H., Belke, B., Oeberst, A., \& Augustin, D. (2004). A model of aesthetic appreciation and aesthetic judgments. British Journal of Psychology, 95, 489-508.
Machado, P., \& Cardoso, A. (1998). Computing aesthetics. Proceedings of XIV \({ }^{\text {th }}\) Brazilian Symposium on Artificial Intelligence. Springer-Verlag.
Margulis, E. H. \& Beatty, A. P. (2008). Musical style, psychoaesthetics, and prospects for entropy as an analytic tool. Computer Music Journal, 32(4), 64-78.
Meyer, L. B. (1957). Meaning in music and information theory. Journal of Aesthetics and Art Criticism, 15(4), 412-424.
Moles, A. (1966). Information Theory and Esthetic Perception. Urbana: Univ. of Illinois Press.
Morgan, D. N. (1950). Psychology and art: a summary and critique. Journal of Aesthetics and Art Criticism, 9(2), 81-96.
Novitz, D. (2004). Rules, creativity and pictures: Wittgenstein's lectures on aesthetics. In P. Lewis, (Ed.), Wittgenstein, aesthetics, and philosophy. Burlington, Vermont: Ashgate.
Ramachandran, V.S., \& Hirstein, W. (1999). The science of art. Journal of Consciousness Studies, 6, 15-51.
Rigau, J., Feixas, M., \& Sbert, M. (2008). Informational aesthetics measures. IEEE Computer Graphics and Applications, 28(2), 24-34.
Saito, Y. (2010). Everyday Aesthetics. Oxford: Oxford Univ. Press.
Sherburne, D. W. (1966). Meaning and music. Journal of Aesthetics and Art Criticism, 24(4), 579-583.
Titchener, J. M. \& Broyles, M. E. (1973). Meyer, meaning, and music. Journal of Aesthetics and Art Criticism, 32(1), 17-25.
Vermazen, B. (1971). Information theory and musical value. Journal of Aesthetics and Art Criticism, 29(3), 367-370.
Weaver, W. (1949). The mathematics of communication. Scientific American, 181(1), 11-15.
Wittgenstein, L. (1967). Lectures and Conversations on Aesthetics, Psychology, and Religious Belief. Oxford: Blackwell.

\title{
An Embodied Perspective of Early Language Exposure
}

\author{
Hanako Yoshida (yoshida@uh.edu) \\ Joseph M. Burling (jmburling@uh.edu) \\ Department of Psychology, University of Houston, 126 Heyne Building \\ Houston, TX 77204-5022 USA
}

\begin{abstract}
Advances in developmental research has made it clear that word learning has a long beginning. Recent work has demonstrated that infants learn words at 6 months of age - that is, before the traditional "first word" milestone in productive language - which is a full year before the usual "naming explosion" in productive vocabulary. Before infants talk, walk, or even point, how can the earliest stage of word learning take place at all? We used recent technology that allowed us to zoom in on the point of view of infants and also the traditional roomview observations to document how infants' visual input is dynamically synchronized with their own participation, as well as from social input in the context of parent-child word learning play. The parents' task was to play with the child with a set of toys as they taught the toys to them. To specifically document the child's dominant view and their participation, we coded the size of the toy object on which the child was focused and who was manipulating the toy at the moment. The results reveal systematic and dynamic links between infants' view and their level of participation.
\end{abstract}

Keywords: embodied perception; word learning context; child-centered view

\section*{Introduction}

Recent advances have made it clear that word learning has an early beginning. A new study has demonstrated that infants comprehend at least some words at 6 months of age - that is, before the traditional "first word" milestone in productive language (Bergelson \& Swingley, 2012) and a full year before the usual "naming explosion" in productive vocabulary (Goldfield \& Reznick, 1990). What is the nature of this very early stage of word learning? What are the experiences when such infants hear words, and are they fundamentally different from the experiences of older infants? During the first 2 years of life, infants learn and refine a whole set of new motor skills that dramatically change the ways in which the body moves and interacts with the environment, and their social interactions consequently also change. Six-month-old infants are relative novices at reaching for objects and do not sit steadily, and so for the most part, objects are brought to them or perceived from a distance; 12-month-olds, in contrast, walk and bring themselves to objects, and 18-month-olds are mobile, socially skilled, and capable of physically achieving their own desires. Do these changes in the ways infants physically interact in the world also change the way they socially interact, and determine the nature of their visual experiences that support word learning?

Needless to say, there are a number of factors that contribute to the process of later word learning, including
social cues such as eye gaze and gesture, prosody, language structure, input frequency, pragmatics, and many others. Yet recent evidence demonstrating effective word learning at a much earlier stage than previously thought suggests the need for investigation of the language learning environment at the earliest stage. Moreover, learners, even very young babies, actively engage in the world by contingently responding to the social gestures of others. In doing so, infants-and perhaps in different ways at different ages - distort regularities and carve up the input in systematic ways. This means that one cannot really consider the input separately from the learners' own actions, because the learner selects and creates the input. Importantly, word learning takes place at any stage of early parent-child interactions, and most often co-occurs with infants' active exploration of objects where their head, hands, or body, and eyes coordinate and shape "input" and possibly optimize their view for initial learning of labels.

Recent work that motivated this present study investigated how early learners create their own visual input by observing the first-person-view during toy play (Smith, Yu, \& Pereira, 2011; Yoshida \& Smith, 2008; Yu, Smith, Shen, Pereira, \& Smith, 2009). These studies typically used a small camera attached to young children's foreheads and documented how this child-centered view provides insight into factors relevant for early word learning. In the work presented here, we sought to study developmental changes in this child-centered view by longitudinally following children from 6 to 18 months in order to detail the quality of their visual input, how it changes over time, and the potential dynamic relation to their rapidly growing physical capacities in the context of parent-child play. Documentation of precise changes in visual experiences mediated through children's own physical growth is essential to studying how infants' changing motor skills serve as strong filters of their early learning experiences, because effective visual attention determines effective learning.

\section*{Emergence of Language Learning}

One crucial question in language learning is how children overcome referential ambiguity - that is, how children selectively extract the referents to map the corresponding words. Effective learning of such a complex skill requires a highly selective process of sampling information from the environment; observing infants' eye gaze reveals a great deal about what they are processing and learn-
ing. Infants' language learning, however, takes place in a dynamically changing context-that of physical growth and motor development. Indeed, acting and knowing are inseparable aspects of human life, and early learning heavily depends on a person's physical capacities and environment. The potential link between language learning and bodily constraints may be uniquely different for different developmental stages. Yet, there are few studies looking directly at the how early bodily experiences influence language learning in relation to an individual's task involvement. As a first step, we document how the early visual input in a language learning tasks relates to the physical constraints of the child's sensory-motor engagement during the task.

\section*{Object Size and Language Learning}

If infants shape the effective view of objects through their own physical growth, consequential actions, and mature social partners' participation, the degree of referential ambiguity or perceptual accessibility may be partially addressed at the level of sensory motor coordination. In a recent study of 18-month-old infants (even some as young as 17 months) in a naturalistic parent-child play context, both parent and infant physically participated in the play (holding objects) at the same level, but objects dominated the child's view (much bigger, thus occluding other objects) when the object was held by the child (Pereira, Smith, \& Yu, under revision; Yu \& Smith, 2012). Infants of this age (18 months) are capable of producing smooth head turnings and can manage a stable posture, suggesting that their own body coordination may help optimize their focus, and bodily movements having a relatively stronger role for the optimal view at this age. This quality of visual experiences was also related to learning object names (Yu \& Smith, 2012), suggesting that how well 18-month-olds can zoom in on objects has important implications for their successful word learning. What happens if younger infants, who seem capable of learning words (Bergelson \& Swingley, 2012), do not yet have the physical capacity to coordinate their bodies to support their optimal view? We specifically targeted younger infants to further our understanding of the role of physical development in the organization of early visual input. We looked at the size of the object in the infants' view at the moment of it being held by the infant or parent.

\section*{Visual Attention, Motor Skills, and Social Development}

Infants and young children's ability to follow eye gaze and pointing directed toward objects is considered a major milestone in the development of joint attention. This type of attention emerges in infants as young as 3 months. It helps early learners identify word meaning and serves a potential communicative function. For infants to develop this ability, they have to coordinate their head and
body so that they can capture the adult's head and eyes as well as the target. As they become more efficient at coordinating their view, they can produce head turnings to follow the adult's movement to maximize the input (Butterworth \& Grover, 1990; D'Entremont, Hains, \& Muir, 1997). If a stable view containing all the important elements for effective attention and learning is important, then one might think that 6 -month-olds, whose head movements are less active (thus much more stable) and views more distal, would be in a better position to capture all the relevant components than more advanced infants who are capable of producing more movements. However, there is one study using a peripheral distracter to measure the focus of 7 - to 10 -month-olds' attention which found infants were more attentive - and showed potentially more effective attentional shifts-when engaged in active play than when receptively observing the task stimuli (Oakes, 1994). Another previous observation points to an interesting potential role of a social partner, suggesting that caregivers naturally respond to children's motor skill changes in order to support the transformation (e.g. Zumbahlen, 1997). These studies raise the question of how, exactly, an infant's own physical participation shapes the development of visual experiences and possibly influences the social partner's physical participation. If infants are not capable of bringing toys to their view, does this motivate the parent to help the infant gain a better view by bringing and holding the objects themselves? Or is the quality of the infants' view tightly linked to their own physical growth (with minimal parental support), thus emerging poorly yet dramatically improving as a function of their physical growth and advances in body coordination?

In the present study, we investigated two components that together influence an infant's visual focus: (a) who brought the object to the child's view (parent/social partner or the child), and (b) how parents' participation reflects their child's physical constraints. Studying how visual and bodily experiences support language learning reveals what contributes to effective visual attention, such as joint attention, and how such effective attention may emerge through both the child's own experiences and parental participation.


Figure 1: Snapshots showing the first-person (right) and third- person (left) view from Yoshida and Smith (2008).


Figure 2: Snapshot of third-person views of the same child across five sessions.

\section*{Child-centered View}

In a head camera study of parent-child play, Yoshida and Smith (2008) recorded 18 -month-old infants' perspective in the context of a parent teaching a set of early learned words while the parent and child sat naturally at a table. The results revealed that the child's view was much more constrained (captured fewer items) compared to the room view (see Figure 1) and provided evidence of the coupling of head and eye movements. Yoshida and Smith (2008) independently measured eye gaze direction (frame by frame via a camera fixated on the infant's eyes) and head direction and found that eye and head directions were highly correlated, such that \(87 \%\) of head camera frames coincided with independently coded directions of eye gaze.

In this study, we used the same procedure to capture younger infants' point of view. We monitored the development of their view by following a set of infants from 6 months to 18 months, testing them every 3 months to document how their focus changes over time, and how their own object exploration and parent's support (in bringing objects) relates to their view (demonstrated by Figures 2 and 3).

Because of the current focus on much younger children than have been studied in the past, we used a slightly different setup - the infant was supported in a child's chair and the parent sat diagonal to the infant-yet we maintained the relational position between the infant and parent used in the previous study. Moreover, a previous head camera study (Smith et al., 2011) evaluating head camera images from different seating arrangements (sitting naturally in a chair or on the floor) found no differences in any aspect of infant or parent behavior as a function of the task geometry, suggesting minimal impact from the current modification to the sitting arrangement. Another modification is that we used the head-mounted portable eye-tracking device (Figure 4) instead of a head camera to ease the difficulty of calibrations. With the head-mounted eye-tracking device, eye gaze can be directly measured and calibration issues can be better addressed without having the infants point to the object of their fixation (typically camera adjustments are made by asking infants to point to or touch what they are looking at).

To run the experiment, we placed an infant in the chair, and then one experimenter put a light weight headmounted eye-tracking device on the infant's forehead while another experimenter distracted the infant by intro-


Figure 3: Images corresponding to a distinct developmental time point, which demonstrates the changes in object size over time and parent-child interactions.
ducing him or her to a set of attractive toys (which made noise and had colorful moving lights). A standard camera recorded the play scene from a corner of the room, so that we collected synchronized third- and first-person-views, as in Yoshida and Smith (2008). In the present study we used naturalistic word learning in the context of playing with toys. The parent was instructed to teach the infant a set of words by selecting the toys the parent thought were most appropriate to play with from a collection of available toys.


Figure 4: Snapshot of an infant (at 6 months) with a headmounted eye-tracking device.

\section*{Study}

With a setup similar to that of Yoshida and Smith (2008), we observed seven infants, starting when they were 6 months old and ending when they were 18 months old, in the context of a parent-child word-learning play session (see Figs. 2 and 3). We specifically investigated the infants' moment-to-moment quality of object view by measuring the size of the focal object (as percentage of image pixels), that is, the image size of the largest object in the infant's view, and determining if the optimal visual moments for an object relate to who is holding it (the infant or parent), and whether this contributes to the infants' development of visual experiences.

\section*{Method}

Participants Seven parent-infant dyads were brought to the laboratory five times (3-month intervals). There were three male and four female infants. Three additional infants began the study but did not contribute to the data because of refusal to wear the measuring equipment. The mean age of the infants was 6.4 months (first visit), 9.2 months (second visit), 12.4 months (third visit), 15.3 months (fourth visit), and 18 months (fifth visit), with each range less than 13 days.

Stimulus and Materials There were eight unique toy objects in a box, located on the floor beside the parent's chair for easy access to them by the parent. Each toy was a naturalistic toy whose name is listed on a developmental vocabulary inventory, McArthur Child Development Inventory (Fenson et al., 1993). Figure 5 shows eight toy objects that were used for parents to teach the typical early-learned words (open, bunny, car, bottle, cookie, eat, drink, put). The order and the duration of playing with them were controlled by the instructions given through the experimenter.


Figure 5: Toy objects used for the word-teaching play sessions.

Procedure Prior to entering the experimental room, the parents were presented with a set of instructions on their type of interaction. Parent and infant then entered the experimental room and were asked to sit in the designated chairs. The experimenter started one camera attached to the wall ( 5 meters away from the chairs) for recording the third-person view, and then put the headmounted eye-tracking device on the child's forehead. The calibration procedure took place to adjust synchronization between camera placement and eye position. The experimenter left the room as she instructed the parents to pace their play and teaching according to the guidance from the audio prompt, which said the name of one of the objects every 40 seconds. The entire session took approximately 6 minutes.

\section*{Results}

The results suggest that a child's view changes dramatically over the course of 12 months of development. At 6 months, the child's view was moderately selective. The focus of the object is then reduced in size at 12 months, then becomes most selective with a sharp increase in object size by 18 months. These changes may be the development of optimal focus, and the results from the room view suggest that the infant's frequency of reaching and holding of objects has a systematic influence on the infant's own view. Furthermore, the parent's participation level also reflected the development of this optimal view.

We first focus on two variables for the present study in order to address the nature of developmental growth in visual experiences. These variables of interest are the
quality of the focused object (how the object looks to the child in terms of the size of the object from their perspective) and also the participation level of the parent and child interacting with the object, determined by the frequency in which either manipulates, holds, or touches the object within a given trial. Eye gaze was measured and tracking data was overlayed onto the video to capture the infant's central view. This was used to aid the coding process and to help determine to which object the child was currently attending.

The coding process involved evaluating individual frames of video to determine the size of the object in view and whether or not the parent and child were interacting with the object. For every 5 seconds of video at 30 fps , multiple data entry fields were randomly generated, giving a total of 250 rows of data per video. Each video corresponds to a single subject and experimental session at a specified developmental time period. Research assistant coders were thoroughly trained to identify the object in view based on the tracking data, and to determine the area of the object in pixels as a proportion of the entire child-centered view. The mean and standard errors of the size proportions are based on sampled frame evaluations and then averaged across all the frames for each 320 second session. A separate set of coders were trained to enter information about the individual that was manipulating the object at each of the sampled frames (if any). Coders made judgments about who was interacting with the object and the type of object seen on screen. The frequency in which parent or child manipulated the object throughout the session was calculated as a proportion based on the total number of cells randomly generated for that session.

The first result, which is evident in Figure 6, shows that objects within an infant's view during the first session take up a moderately sized proportion of their view, yet at 12 months (their 3rd visit), the average size of objects focused within their view is reduced dramatically, and reaches its lowest point during this age. The proportion of all objects in terms of pixel area was greatest at 18 months of age (5th and final visit). There were significant developmental differences in object size between 6 months and 9 months, \((t=2.56 ; p<.01)\), and between 15 months and 18 months, \((t=2.28 ; p<.05)\). There were no significant differences found between ages 9 and 12 months, nor between 12 and 15 months.

The developmental shifts in focus is interesting and suggests that these optimal visual experiences do not come as easy, and do not appear to be a straight linear growth. At the 6,9 , and 12 month points, the size of objects in their visual field seem to fluctuate, and the dip at 12 months may be of developmental significance and contain important information about their physical growth and parental involvement. Initial parental interactions with the infant may help set the stage for determining


Figure 6: The average size of all objects as a proportion of the camera view at each longitudinal time point.
optimal visual experiences while infants learn to properly handle and manipulate objects later on their own.

An infant's optimal view has been characterized by very few objects being focused upon and dominating the infant's visual field (Smith et al., 2011; Yoshida \& Smith, 2008; Yu, Smith, et al., 2009). But, the magnitude of this type of focus appears to change dramatically in one year (see Figure 3), where at least by 6 months, infants show initial steps toward the development of language learning based on both physical growth and environmental interactions. To address this question, we analyzed the variable where coders made a judgment about who was interacting with the object. In absence of a longitudinal perspective, results might suggest that across all developmental time points ( 6 to 18 months) objects were held equally frequent between infants and parents. Yet, as can be seen in Figure 7, as early as 6 months, parents held the object of focus reliably more than infants \(\left(\chi^{2}=6.25 ; p<.05\right)\). This contrasts with later periods, in which children are interacting with objects much more frequently than adults. At 15 months, infants began to reliably hold the objects more often than adults ( \(\chi^{2}=4.9 ; p<.05\) ), and by 18 months, the magnitude of this difference in object interaction was most dramatic ( \(\chi^{2}=7.79 ; p<.01\) ). There were no significantly reliable differences found in object manipulation during the 9 month and 12 month time periods.

During the early stages of development, parental involvement seems to account for a majority of the visual experiences gained by infants, while at later stages, infants are responsible for their own experiences as they hold and manipulate objects with greater frequency. At the 12 month time point, there was no reliable difference in who was holding the object. Interestingly, this is the point where their center view quality drops to as small as \(12.1 \%\). This may be the point at which the parent begins to demonstrate less involvement in shaping the child's visual experiences-and thus a drop in object


Figure 7: Frequency in which participants interact with objects over the course of 5 longitudinal sessions shown as a proportion of the number of coded frames.
focus-which allows for the infant to independently start shaping what comes into their view and determine their own optimal focus, an ability that they apparently are not immediately successful in controlling during the early months.

\section*{General Discussion}

The present findings suggest the contexts for early and later word learning are quite different, and that they shift from greater parental control to greater infant control, with parents ensuring early on that the named object is visually dominant, and the infant plays a more active role later in development. Clearly these findings are just a first step toward understanding the possible, developmentally changing pathways through which infants learn words. Recent advances make clear that infants start learning words much earlier than previously thought (as young as 6 months), before the traditional "first word" milestone in productive language (Bergelson \& Swingley, 2012). The present results suggest that parents may isolate objectsfocusing on objects one at a time, zooming in on a single object-to help this learning. Later, this isolation may be most effectively done when the child actively engages in object manipulation (holding and bringing objects) as suggested by the Yu and Smith (2012) results.

The inflexion point in the visual size of the named object at 12 -months is intriguing, and suggests a possible transition from more parent control toward more infant controlled learning, and the parent following of the child's interests when naming. The 12 -month mark has been noted by others as a period of change in social interactions. For example, whereas very young infants appear to automatically follow the eye-gaze of another (Farroni, Massaccesi, Pividori, \& Johnson, 2004; Hood, Willen, \& Driver, 1998) and follow head movement-not eye-gaze - when head and eye direction are in competition (Brooks \& Meltzoff, 2005; Gergely, Nádasdy, Csibra,
\& Bíró, 1995), 12-month olds appear to require more coherent cues as well as contingent interactions to follow these cues (Tomasello, Hare, Lehmann, \& Call, 2007; Moore \& Corkum, 1998; Johnson, Ok, \& Luo, 2007). In brief, as infants' motor and cognitive skills make them more independent, social interactions and the structure of word learning may change in systematic ways.

The present finding that 12-month-old infants do not seem to experience this optimal view suggests that both parents and infants may be working out this transition. This finding does not mean that the infant and parent are not involved with each other or have no interest in playing. Rather, infants and parents appeared to participate relatively equally. This leads to new insight into how action coordinated through dominant participation (one agent or the other, but not both) maximizes support for the development of optimal view, and it raises a number of novel developmental predictions about the role of action coordination such as emergence of join attention and social contingencies in learning.

Linking word learning to the changing sensory-motor skills of infants may also be key to understanding the co-morbidity of language delay and motor deficits in developmental disorders such as autism (Iverson, 2010). Studying changes in visual experiences mediated through a child's own physical growth and experiences is a way to gain new perspectives on the nature of embodied language learning.

\section*{Acknowledgments}

This research was supported by the National Institutes of Health grant (R01 HD058620), the Foundation for Child Development, and University of Houston's Grants to Enhance and Advance Research (GEAR) program. We especially wish to thank the families who participated in this study. We also thank the undergraduate research students in the Cognitive Development Lab for their support in coding for the present study.

\section*{References}

Bergelson, E. \& Swingley, D. (2012). At 6 to 9 months, human infants know the meanings of many common nouns. Proceedings of the National Academy of Sciences of the USA, 109, 3253-3258.
Brooks, R. \& Meltzoff, A. N. (2005). The development of gaze following and its relation to language. Developmental Science, 8(6), 535-543.
Butterworth, G. \& Grover, L. (1990). Joint visual attention, manual pointing, and preverbal communication in human infancy. In M. Jeannerod (Ed.), Attention and performance XIII: motor representation and control (pp. 605-624). Lawrence Erlbaum Associates, Inc.
D'Entremont, B., Hains, S. M. J., \& Muir, D. W. (1997). A demonstration of gaze following in 3- to 6-month-olds. Infant Behavior \& Development, 20(4), 569-572.

Farroni, T., Massaccesi, S., Pividori, D., \& Johnson, M. H. (2004). Gaze following in newborns. Infancy, 5(1), 3960.

Fenson, L., Dale, P. S., Reznick, J. S., Thal, D., Bates, E., Hartung, J. P., . . . Reilly, J. S. (1993). The MacArthur Communicative Development Inventories: User's guide and technical manual. Baltimore: Paul H. Brokes Publishing Co.
Gergely, G., Nádasdy, Z., Csibra, G., \& Bíró, S. (1995). Taking the intentional stance at 12 months of age. Cognition, 56(2), 165-193.
Goldfield, B. A. \& Reznick, J. S. (1990). Early lexical acquisition: rate, content, and the vocabulary spurt. Journal of Child Language, 17, 171-183.
Hood, B. M., Willen, J. D., \& Driver, J. (1998). Adult's eyes trigger shifts of visual attention in human infants. Psychological Science, 9(2), 131-134.
Iverson, J. M. (2010). Developing language in a developing body: the relationship between motor development and language development. Journal of Child Language, 37, 229-261.
Johnson, S. C., Ok, S.-J., \& Luo, Y. (2007). The attribution of attention: 9-month-olds' interpretation of gaze as goal-directed action. Developmental Science, 10(5), 530-537.
Moore, C. \& Corkum, V. (1998). Infant gaze following based on eye direction. British Journal of Developmental Psychology, 16(4), 495-503.
Oakes, L. M. (1994). Development of infants' use of continuity cues in their perception of causality. Developmental Psychology, 30(6), 869-879.
Pereira, A., Smith, L. B., \& Yu, C. (under revision). A bottom-up view of toddler word learning. Psychological Bulletin and Review.
Smith, L. B., Yu, C., \& Pereira, A. (2011). Not your mother's view: the dynamics of toddler visual experience. Developmental Science, 14, 9-17.
Tomasello, M., Hare, B., Lehmann, H., \& Call, J. (2007). Reliance on head versus eyes in the gaze following of great apes and human infants: the cooperative eye hypothesis. Journal of Human Evolution, 52(3), 314320.

Yoshida, H. \& Smith, L. B. (2008). What's in view for toddlers? Using a head camera to study visual experience. Infancy, 13, 229-248.
Yu, C. \& Smith, L. B. (2012). Embodied attention and word learning by toddlers. Cognition, 125(2), 244-262.
Yu, C., Smith, L. B., Shen, H., Pereira, A., \& Smith, T. (2009). Active information selection: visual attention through the hands. IEEE Transactions on Autonomous Mental Development, 2, 141-151.
Zumbahlen, M. R. (1997). The role of infant locomotor onset in shaping mother-infant communication (Doctoral dissertation, University of Illinois at UrbanaChampaign).

\title{
Online Processing of Speech and Social Information in Early Word Learning
}

\author{
Daniel Yurovsky \\ yurovsky@stanford.edu \\ Department of Psychology \\ Stanford University
}

\author{
Anna Wade \\ wadea@neurosurg.ucsf.edu \\ Department of Neurological Surgery \\ University of California, San Francisco
}

\author{
Michael C. Frank \\ mcfrank@stanford.edu \\ Department of Psychology \\ Stanford University
}

\begin{abstract}
Although word learning unfolds over days, weeks, and months, individual naming events are over in a matter of seconds. To benefit from a naming event, children must at least hear the label and see the referent. We tested 1 -, 2- , 3-, and 4 -year old children in a naturalistic word learning task with two conditions: one that taxed both speech processing and rapid gazefollowing, and one in which a social cue-to-reference was available for an extended time. The development of wordlearning in the extended condition paralleled the development of speech processing, but learning in the brief condition lagged behind. However, learning from both the brief and extended cues was predicted by individual differences in speech processing and cue-following together. Thus, even through the 4th year, real-time processing of social and linguistic information are a critical bottleneck for word learning.
\end{abstract}

Keywords: Language acquisition, word learning, attention, social cues, development

\section*{Introduction}

Language learning is a fundamentally social endeavor - it relies critically on input from social partners. This is because many aspects of natural languages, like the mappings between words and their referents, are conventions that can vary from community to community (Chater \& Christiansen, 2010). To learn their first words, children must track the relationship between the sounds that speakers produce and the things in their world (Pinker, 1984; Siskind, 1996).

Not everything about a word and its referents needs to be learned in a single shot. Instead, it is likely that this relationship is refined over multiple exposures, over a period of days, weeks or months (Carey \& Bartlett, 1978; Smith \& Yu, 2008; McMurray, Horst, \& Samuelson, 2012). Nonetheless, each individual naming event occurs in real-space and realtime (Samuelson, Smith, Perry, \& Spencer, 2011; Spencer, Perone, Smith, \& Samuelson, 2011). If a child does not hear the label, or does not see the target referent, the information "available" in the naming event is effectively unavailable to the child (Yu \& Smith, 2012; Yurovsky, Smith, \& Yu, in press). Thus, a critical bottleneck in language acquisition is the ability to process the right information at the right time (Fernald \& Marchman, 2012).

Young children are slow processors of both speech and visual input (Kail, 1991; Fernald, Pinto, Swingley, Weinberg, \& McRoberts, 1998). They are also slow to re-deploy their attention in response to changing visual information (Colombo, 2001). Thus, it is perhaps unsurprising that the properties of child-directed speech and child-directed actions seem welldesigned to scaffold slow processors. Child-directed utterances are slower, shorter, and have larger pitch contours than do utterances spoken to adults. Repetition is common in
child-directed speech, and key words are made more salient through minor local variations (Onnis, Waterfall, \& Edelman, 2008). In addition, new labels are often introduced through a small set of common naming phrases that facilitate attention tom and learning of, new words (Fernald \& Hurtado, 2006; Yurovsky, Yu, \& Smith, 2012). Exposure to these kind of structure makes a real difference, predicting individual differences in vocabulary development (Hoff, 2003). Childdirected actions are similarly exaggerated, marked by bigger, simpler, and more repetitive movements (Brand, Baldwin, \& Ashburn, 2002). Indeed, multi-modal synchrony between these exaggerated visual and auditory inputs may be precisely the the information that young children use to to learn their first words (Gogate, Bahrick, \& Watson, 2000).

However, while social partners may sometimes scaffold young language learners, simplifying their speech and timing their naming events so that labels coincide with the focus of children's visual attention, this kind of pedagogical naming likely accounts for a minority of the relevant input from which words and their referents could be learned. For instance, while isolated words facilitate speech processing and word learning, they make up less than \(10 \%\) of all speech to children (Brent \& Siskind, 2001). Similarly, referential expressions produced while children are already looking at the target object facilitate word learning, but make up only a portion of the naming events produced to young children (Tomasello \& Farrar, 1986; Frank, Tenenbaum, \& Fernald, 2013). In the remaining naming events, successful learning requires the child to check with the speaker to determine the target of her reference (Baldwin, 1991). Consequently, the ability to quickly process speech, and to quickly follow a speaker's social cues, should both give learners access to more and more useful information.

Over the first two years, children make rapid gains in the rate at which they process auditory input, picking referential words out of continuous speech (Fernald et al., 1998). During this time, individual differences in rate of spoken language processing predict individual differences in vocabulary size (Fernald, Perfors, \& Marchman, 2006). Over the same two year period, children also improve in their abilities to attend to and use social cues indicating the target of a speaker's reference (Scaife \& Bruner, 1975; Hollich, Hirsh-Pasek, \& Golinkoff, 2000). As with speech processing, individual differences in children's gaze- and point-following predict their language development over these two years (Brooks \& Meltzoff, 2006).

But, while the second year is marked by an increase in the rate at which children learn new words, this acceleration in


Figure 1: Example training and test trials from the Experiment. On Extended Cue trials (a), the speaker picked up and interacted with the target toy over the course of \(10-20\) seconds, providing consistent social information about her target reference. On Brief Cue trials (b), the speaker made a quick glance at the target toy, and then looked forward into the camera for the remainder of the trial. Determining her target of reference on these trials thus required rapid social-cue following. On test trials (c), children saw two of the toys from training and heard the speaker's voice ask them to find the target toy.
vocabulary growth continues into the third and fourth years and beyond (Bloom, 2000). Do these same skills in spoken language processing and rapid social-cue following continue to develop and continue to predict word learning over this extended period of vocabulary acceleration?

To determine how speech processing, social-cue following, and word learning co-develop over the first four years, we tested a large cross-sectional sample of children in age from \(1-5\) years in a short, naturalistic word-learning task. Over the course of approximately four minutes, we measured each child's ability to process speech containing a known referent, learn a new word when an Extended social-cue continuously provided disambiguating information about the target referent, and learn a new word the speaker gave only a Brief social-cue to indicate her target referent, requiring rapid gazefollowing. We subsequently fit a linear mixed-effects model to children's looking times to determine how speech processing and social-cue following predicted word learning in both conditions over the course of development.

\section*{Method}

Children's eye movements were tracked while they watched a series of naturalistic word-learning videos. In each, children saw a speaker seated at a table between two novel toys. She introduced them to one of the toys, providing a label and several interesting facts about it. Crucially, on some of the trials she provided an Extended Cue indicating the target of her reference - picking up the object and interacting with it over the course of the video. On other trials, she provided Brief Cue - only a quick glance to the target object when she first labeled it. After each learning trial, children were tested for their knowledge of the referent for the new word using the Looking While Listening procedure (Fernald et al., 1998). In addition, similar test trials were administered for known objects to measure children's processing of familiar words. Because eye movements were recorded during both learning and test, we were able to analyze the relationship between children's behavior during learning and test trials.

\section*{Participants}

Parents and their 1-5 year-old children were invited to participate in a short language learning study while they visited the San Jose Children's Discovery museum. All-together, we collected demographic and experimental data from 114 children, 39 of whom were excluded for one or more of the following reasons: abnormal developmental issues \((N=7)\), failure to calibrate ( \(N=26\) ), less than \(75 \%\) exposure to English ( \(N=13\) ), and fussiness or inattention \((N=22)\). The final sample consisted of \(181-2\) year olds ( \(M_{\text {age }}=1 \mathrm{yr}\).; 7 mo., 9 girls), 25 2-3 year olds ( \(M_{\text {age }}=2\) yr.; 6 mo., 9 girls), \(213-4\) year olds ( \(M_{\text {age }}=3 \mathrm{yr}\).; 6 mo., 8 girls ), and 11 children over the age of 4 ( \(M_{\text {age }}=4\) yr.; 8 mo., 5 girls).

\section*{Stimuli}

The experiment consisted of two kinds of trials: learning and test. Learning trials were \(10-20\) second video clips in which a speaker first introduced herself to the child, and then produced a short monologue about one of the two toys on the screen, labeling it three times. The script for the first learning trial, for example, was "Hey there, can I show you my friend's toys? This is a manu. I really like the manu. The manu is fun to play with." The exact script varied from trial, but always followed this general format. On Extended Cue trials, the speaker picked up the target toy and engaged with it over the course of labeling (Figure 1a). In contrast, on the Brief Cue trials, the speaker indicated her target of reference only with a quick glance to the toy when she first produced it's label. She looked straight into the camera for the rest of the trial (Figure 1b). Thus, learning from Brief Cue trials required children to rapidly follow her gaze.

Test trials followed the standard Looking While Listening protocol (Fernald et al., 1998). On each test trial, children saw two objects - one on each side of the screen - heard a short audio clip of the speaker from the learning trials asking them to find a target object (Figure 1c). Each test trials was 7.5 seconds long. On Familiar test trials, both the target and distractor were common objects familiar to young children (e.g. book vs. dog). On Novel test trials, both the target and distractor were novel objects from the previous learning trial.

Finally, the experiment ended with a calibration check: a short video in which small dancing stars appeared in four places on the screen. Because eye-tracker calibration can be imprecise, especially with younger children (Morgante, Zolfaghari, \& Johnson, 2011), this check allowed us to adjust initial calibration settings to minimize the discrepancy between the behavior children produced and the behavior we analyzed (for details, see Frank, Vul, \& Saxe, 2012).

\section*{Design and Procedure}

The experiment began with a 4-point calibration and then proceeded into a series of learning/test blocks. In each block, children first watched a learning trial in which a speaker labeled one of two on-screen toys. Following this learning trial, children were given a Looking While Listening test trial in which they saw both of these toys and were asked to find the toy labeled on the previous learning trial (e.g. "Can you find the manu?"). Each block consisted to two such learning/test combinations: one for a toy indicated by an Extended Cue, and one for a toy indicated by a Brief Cue (Figure 1a and b). The same toys and the same label (тапи) were used for all Extended Cue trials, and a different set of toys and a different label (bosa) were used for all Brief Cue trials. The entire experiment consisted of three such blocks, and two Familiar test trials were inserted between each block. Thus, in total, each child participated in three learning and test trials in each Cue condition, and four Familiar test trials.

\section*{Data Analysis}

Children's eye movements during both learning and testing were analyzed using a Regions of Interest (ROI) approach. On learning trials, bounding-box ROIs were drawn by a human coder frame-by-frame for the speaker's face and for the two objects. On test trials, a bound-box ROI was drawn for each of the two static images. To ensure that recorded eye movements were mapped to the correct ROIs, children's calibrations were first adjusted by fitting a robust linear regression for their fixations during the calibration check video and using this model to transform eye movements during the rest of the experiment (Frank et al., 2012).

Children's learning and test behaviors were quantified by measuring their proportion of looking to each ROI on each trial. To ensure that proportions were representative, individual test trials were excluded from analysis if eye gaze data was missing for more than half of their duration. To compute age-group looking proportions, proportions were computed first for each individual trial, averaged at the individual-child level, and then averaged across children.

Window-of-analysis selection began by coding the point of disambiguation for each trial. This was the onset of the target label for test trials, and the rotation of the speaker's head for learning trials (marked ' 0 ' in the graphs in the Results section). The window for each trial began 500 ms after this point of disambiguation to allow children of all ages enough time to process. The window ended at the end of test trials, and


Figure 2: Children's probabilities of fixating the correct target of each label over the course of each test trial. The point labeled 0 indicates the onset of the label, and different colors indicate different age groups. Each line indicates the mean proportion of looking for one age group, and shaded areas repent \(\pm 1 \mathrm{SE}\). A proportion of .5 indicates chance performance.
the point at which the label was heard for a second time on learning trials: 2.5 seconds after the point of disambiguation.

\section*{Results and Discussion}

Children's patterns of fixation provide a continuous record of their moment-by-moment visual attention over the course of both learning and test trials. We first present an analysis of word learning and familiar word recognition over development. We then connect test behavior to children's patterns of looking during learning. Figure 2 shows gaze trajectories over the course of both Familiar and Novel test trials for each age group. To quantify children's learning with standard analyses, we aggregated these patterns of looking over time to compute the aggregate proportion of looking at the target object on each test trial.

\section*{Test Trials}

Overall, children in each age range showed evidence of recognizing familiar words - looking at the correct target on Familiar trials for a greater proportion of time than expected by chance \(\left(M_{1-2}=.60, t(16)=2.27, p<.05 ; M_{2-3}=.76\right.\), \(t(22)=10.31, p<.001 ; M_{3-4}=.78, t(20)=7.95, p<.001\); \(\left.M_{4+}=.82, t(9)=8.46, p<.001\right)\). A linear model showed that familiar word recognition improved significantly across development ( \(\beta_{\text {age }}=.07, t(67)=4.12, p<.001 ; r=.46\) ).

When tested for their knowledge of the word from Extended Cue trials, children in the youngest age group did not show evidence of learning \(\left(M_{1-2}=.48, t(17)=-.41, p=\right.\) .68), but children in the older age groups did \(\left(M_{2-3}=.64\right.\), \(t(20)=3.14, p<.01 ; M_{3-4}=.71, t(20)=5.22, p<.001\); \(\left.M_{4+}=.75, t(8)=8.46, p<.001\right)\). Learning from Extended Cue trials also improved significantly across development ( \(\beta_{\text {age }}=.09, t(69)=4.28, p<.001 ; r=.45\) ).

Finally, when tested for knowledge of the word from Brief Cue trials, 1-2 year olds did not show evidence of learning ( \(M_{1-2}=.48, t(17)=-.44, p=.67\) ), 2-3 year olds showed marginal evidence of learning \(\left(M_{2-3}=.59, t(21)=1.96\right.\), \(p=.06\) ), and the older two age groups showed significant evidence of learning; \(M_{3-4}=.60, t(20)=2.41, p<.05 ; M_{4+}=\) \(.74, t(9)=6.57, p<.001)\). As with the other conditions, learning from Brief Cue trials improved significantly across development \(\left(\beta_{\text {age }}=.08, t(69)=3.64, p<.01 ; r=.40\right)\).

Together, these results provide clear evidence of a developmental trajectory in both word learning and word recognition (Figure 3). Word recognition and learning from the Extended Cue improved particularly rapidly over early development, consonant with previous work examining the link between speech recognition and early word learning (Fernald et al., 2006; Fernald \& Marchman, 2012). Because the speaker in the Extended Cue condition continued to provide a social cue-to-reference over the course of labeling, the primary hurdle to learning in this condition was speech processing rather than referential ambiguity. The Brief Cue condition, however, presented an additional challenge: children needed to rapidly follow the speaker's social gaze to determine her target of reference. In this condition, the biggest jump in performance came much later in development. While 2-3 year olds showed marginal evidence of learning from the Brief Cue, and learning in 3-4 year olds was statistically significant, only the oldest children showed robust evidence of learning. Because children in the middle age groups are well into the stage of development at which they attend to and learn from social cues, it is likely that Brief Cue condition was difficult for them precisely because it required rapid processing of referential information (Baldwin, 1991; Hollich et al., 2000).

These results suggest that a critical bottleneck in early word learning may be attention in-the-moment: children need to process speech and social information quickly enough to determine the label and target of reference.


Figure 3: Children improved in their abilities to recognize familiar words, and to learn from both the Extended and Brief Cues over the course of development. Individual lines indicate different age groups and error bars indicate \(\pm 1\) SE.

Table 1: Predicting Learning of Novel Words.
\begin{tabular}{lrrr}
\hline Predictor & Estimate (SE) & \(t\) Value & Significance \\
\hline Intercept & \(.34(.09)\) & 4.04 & \(p<.001\) \\
Age (yrs) & \(.06(.02)\) & 3.18 & \(p<.001\) \\
Brief Cue & \(.04(.04)\) & 1.00 & \(p=.34\) \\
Familiar Test & \(.25(.12)\) & 2.09 & \(p<.05\) \\
Face Prop. & \(-.23(.08)\) & -2.91 & \(p<.01\) \\
\hline
\end{tabular}

\section*{Connecting Learning and Test}

In addition to recording children's patterns of looking on test trials, we also captured their looking behavior during learning. This allowed us to chart the developmental trajectories of looking at the caregiver's face and at the target of reference. Figure 4 shows the time course of looking for each age group in both Cue conditions around the point of disambiguation.

In the Extended cue condition, looking patterns were qualitatively similar across development. At all ages, children oriented to the speaker's face as she began speaking, and then switched their attention to her target of reference between 500 ms and 1.5 s after she produced the label. They continued to look predominantly at this target object for the next several seconds. There were apparent quantitative differences for instance the youngest children were slowest to disengage form the face, but the Extended cue scaffolded children at all ages into finding the target of reference and sustained their attention on it.

In contrast, looking patterns in the Brief cue condition changed qualitatively across development. Children in the youngest two age groups generally maintained fixation on the speaker's face long after the point of disambiguation, and were relatively unlikely to attend to the target referent. Thus, they were not able to process the speaker's social gaze quickly enough to use it for disambiguation. In contrast, the 3-4 year olds, and especially children over the age of four, showed evidence of disengaging from the face and following the speaker's gaze to find her intended referent. These data provide evidence of children's developing abilities to track and use social information in real-time at a rapid rate.

To determine whether these developing abilities to process speech and social cues contribute to word learning, we fit a linear mixed-effects model to the data (Baayen, Davidson, \& Bates, 2008). This model used children's age, their accuracy on Familiar test trials, and their looking during learning trials to predict their test accuracy for both Extended and Brief cue trials. Table 1 shows coefficient estimates and their significance for each of these predictors. While Cue type was not a significant predictor, age and Familiar test accuracy were both significant positive predictors of test accuracy, and looking to the speaker's face was a significant negative predictor. No interaction terms approached significance. \({ }^{1}\)

\footnotetext{
\({ }^{1}\) When looking to the target was included instead of looking to the face, this term was a significant positive predictor \((\beta=.18, t=\)
}


Figure 4: Children's looking patterns during learning for both Extended Cue and Brief Cue trials. The top row shows looking to the target referent and the bottom row shows looking to the speaker's face. Dotted lines at ' 0 ' indicate the point of disambiguation. Looking patterns were qualitatively similar in the Extended Cue condition across development, but diverged significantly in the Brief Cue condition. Only the oldest two groups of children were able to rapidly follow the speaker's social cue.

Thus, children who were fast at picking the label out of the speaker's utterance, and fast to follow her social cue in both Cue conditions were the most likely to learn the mapping between the word and its target referent. Because age was also a significant predictor, even after accounting for speech processing and cue-following, there must be additional changes in cognitive processing across development that moderate to the connection between real-time attention and ultimate word learning (e.g. working memory). Nonetheless, these data provide strong evidence that children's abilities to process both speech and social signals change over the course of the first four years, and that changes in these skills are important contributors to word learning.

\section*{Conclusion}

Although children may learn words by aggregating information across a number of naming events (Pinker, 1984; Smith \& Yu, 2008), their success must ultimately be constructed from the information they acquire in each individual event. Because both speech and social cues to reference are rapid, serial channels, getting the most out of each naming event requires processing words and identifying social referents quickly and accurately. Our data suggest that the ability to do both of these things develops significantly over the course of childhood, and that both of these abilities are related to the ability to learn novel labels for novel objects.

While a large body of work has established the relationship between children's language processing speed and their later language outcomes (Fernald et al., 1998, 2006), our study

\footnotetext{
\(1.996, p<.05)\). However, this model gave a slightly poor fit to the data, so we report the version including looking to the face.
}
adds to this literature by suggesting that processing speed is important in social understanding as well. Much of the early social input that children receive from their caregivers is the social equivalent of child-directed speech: slow, clear, and focused on accessible referents. But as children develop and begin to interact with others, they may encounter an increasing proportion of situations in which they need to track a fleeting glance or a subtle reference. Being able to apprehend these brief social signals may play an important role in allowing children to learn across a range of environments.

More generally, becoming a better word learner is about getting more information out of less input. Many developments that are linked to better word learning - the emergence of mutual exclusivity, the shape bias, and increased speed in language processing (Yurovsky, Bion, Smith, \& Fernald, 2012; Smith, Jones, Landau, Gershkoff-Stowe, \& Samuelson, 2002; Fernald \& Hurtado, 2006) - have their effects because they allow children to glean information about word meanings from their environment more effectively. The work in this paper suggests that children's developing understanding of the social environment may have a similar role in early word learning.

\section*{Acknowledgments}

The authors are grateful to Middy Casillas and to all of the members of the Language and Cognition Lab for feedback on the project and manuscript. This work was supported by a John Merck Scholars Fellowship to MCF.

\section*{References}

Baayen, R. H., Davidson, D. J., \& Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language, 59, 390-412.
Baldwin, D. A. (1991). Infants' contribution to the achievement of joint reference. Child Development, 62, 875-890.
Bloom, P. (2000). How children learn the meanings of words. Cambridge: MA. MIT Press.
Brand, R. J., Baldwin, D. A., \& Ashburn, L. A. (2002). Evidence for 'motionese': Modifications in mothers' infantdirected action. Developmental Science, 5, 72-82.
Brent, M. R., \& Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. Cognition, 81, B33-44.
Brooks, R., \& Meltzoff, A. N. (2006). Infant gaze following and pointing predict accelerated vocabulary growth through two years of age: A longitudinal, growth curve modeling study. Journal of Child Language, 35, 207-220.
Carey, S., \& Bartlett, E. (1978). Acquiring a single new word. Papers and Reports on Child Language Development, 15, 17-29.
Chater, N., \& Christiansen, M. H. (2010). Language acquisition meets language evolution. Cognitive Science, 34, 1131-1157.
Colombo, J. (2001). The development of visual attention in infancy. Annual Review of Psychology, 52, 337-367.
Fernald, A., \& Hurtado, N. (2006). Names in frames: Infants interpret words in sentence frames faster than words in isolation. Developmental Science, 9, F33-F40.
Fernald, A., \& Marchman, V. A. (2012). Individual differences in lexical processing at 18 months predict vocabulary growth in typically developing and late-talking toddlers. Child Development, 83, 203-22.
Fernald, A., Perfors, A., \& Marchman, V. A. (2006). Picking up speed in understanding: Speech processing efficiency and vocabulary growth across the 2nd year. Developmental Psychology, 42, 98-116.
Fernald, A., Pinto, J. P., Swingley, D., Weinberg, A., \& McRoberts, G. W. (1998). Rapid gains in speed of verbal processing by infants in the 2nd year. Psychological Science, 9, 228-231.
Frank, M. C., Tenenbaum, J. B., \& Fernald, A. (2013). Social and discourse contributions to the determination of reference in cross-situational word learning. Language, Learning, and Development, 9, 1-24.
Frank, M. C., Vul, E., \& Saxe, R. (2012). Measuring the development of social attention using free-viewing. Infancy, 17, 355-375.
Gogate, L. J., Bahrick, L. E., \& Watson, J. D. (2000). A study of multimodal motherese: The role of temporal synchrony between verbal labels and gestures. Child Development, 71, 878-94.
Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary de-
velopment via maternal speech. Child Development, 74, 1368-1378.
Hollich, G. J., Hirsh-Pasek, K., \& Golinkoff, R. M. (2000). Breaking the language barrier: An emergentist coalition model for the origins of word learning. Monographs of the Society of Research in Child Development, 65.
Kail, R. (1991). Processing time declines exponentially during childhood and adolescence. Developmental Psychology, 27, 259-266.
McMurray, B. A., Horst, J. S., \& Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. Psychological Review, 119, 831-877.
Morgante, J. D., Zolfaghari, R., \& Johnson, S. P. (2011). A critical test of temporal and spatial accuracy of the Tobii T60XL eye tracker. Infancy, 17, 9-32.
Onnis, L., Waterfall, H. R., \& Edelman, S. (2008). Learn locally, act globally: Learning language from variation set cues. Cognition, 109, 423-430.
Pinker, S. (1984). Learnability and cognition: The acquisition of argument structure. Cambridge, MA.: MIT Press.
Samuelson, L. K., Smith, L. B., Perry, L. K., \& Spencer, J. P. (2011). Grounding Word Learning in Space. PLoS ONE, 6, e28095.
Scaife, M., \& Bruner, J. S. (1975). The capacity for joint visual attention in the infant. Nature, 253, 265-266.
Siskind, J. M. (1996). A computational study of crosssituational techniques for learning word-to-meaning mappings. Cognition, 61, 39-91.
Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., \& Samuelson, L. (2002). Object name learning provides on-the-job training for attention. Psychological Science, 13, 13-9.
Smith, L. B., \& Yu, C. (2008). Infants rapidly learn wordreferent mappings via cross-situational statistics. Cognition, 106, 1558-1568.
Spencer, J. P., Perone, S., Smith, L. B., \& Samuelson, L. K. (2011). Learning words in space and time: Probing the mechanisms behind the suspicious-coincidence effect. Psychological Science, 22, 1049-1057.
Tomasello, M., \& Farrar, M. J. (1986). Joint attention and early language. Child Development, 57, 1454-63.
Yu, C., \& Smith, L. B. (2012). Embodied attention and word learning by toddlers. Cognition, 125, 244-262.
Yurovsky, D., Bion, R. A. H., Smith, L. B., \& Fernald, A. (2012). Mutual exclusivity and vocabulary structure. In N. Miyake, D. Peebles, \& R. P. Cooper (Eds.), (pp. 11971202). Austin, TX: Cognitive Science Society.

Yurovsky, D., Smith, L. B., \& Yu, C. (in press). Statistical word learning at scale: The baby's view is better. Developmental Science.
Yurovsky, D., Yu, C., \& Smith, L. B. (2012). Statistical speech segmentation and word learning in parallel: Scaffolding from child-directed speech. Frontiers in Psychology, 3, 374.

\title{
Cheap but Clever: Human Active Learning in a Bandit Setting
}

\author{
Shunan Zhang Angela J. Yu \\ (s6zhang, ajyu@ucsd.edu) \\ Department of Cognitive Science, University of California, San Diego \\ 9500 Gilman Drive, La Jolla, CA 92093-0515
}

\begin{abstract}
How people achieve long-term goals in an imperfectly known environment, via repeated tries and noisy outcomes, is an important problem in cognitive science. There are two interrelated questions: how humans represent information, both what has been learned and what can still be learned, and how they choose actions, in particular how they negotiate the tension between exploration and exploitation. In this work, we examine human behavioral data in a multi-armed bandit setting, in which the subject choose one of four "arms" to pull on each trial and receives a binary outcome (win/lose). We implement both the Bayes-optimal policy, which maximizes the expected cumulative reward in this finite-horizon bandit environment, as well as a variety of heuristic policies that vary in their complexity of information representation and decision policy. We find that the knowledge gradient algorithm, which combines exact Bayesian learning with a decision policy that maximizes a combination of immediate reward gain and longterm knowledge gain, captures subjects' trial-by-trial choice best among all the models considered; it also provides the best approximation to the computationally intense optimal policy among all the heuristic policies.
\end{abstract}

Keywords: Bandit problems; human decision making; human active learning; knowledge gradient

\section*{Introduction}

How humans achieve long-term goals in an imperfectly known environment, via repeated tries and noisy outcomes, is an important problem in cognitive science. The computational challenges consist of the learning component, whereby the observer updates his/her representation of knowledge and uncertainty based on continual observations, and the control component, whereby the observer chooses an action that somehow balances between the need to obtain immediate reward and to obtain information that assists long-term reward accumulation.

A classical task setting used to study sequential decisionmaking under uncertainty is the multi-armed bandit problem (Robbins, 1952). The bandit problems are a family of reinforcement-learning problems where the decision maker must choose among a set of arms on each trial: the reward gained on each trial both has intrinsic value and informs the decision maker about the relative desirability of the arms, which can help with future decisions. In the basic bandit setting, each arm has an unknown probability of generating a reward on each trial. The problem is called finite horizon if the total number of trials is finite; it is called infinite horizon if the number of trials is infinite, in which case one either discounts future rewards or tries to maximize average reward per unit time. In this work, we focus on stationary, non-discounted, finite-horizon bandit problems, where the underlying reward rates are independent and identically (iid) distributed across the arms.

Bandit problems elegantly capture the tension between exploration (selecting an arm about which one is ignorant) and exploitation (selecting an arm that is known to have relatively high expected reward), which is manifest in many real-world decision-making situations involving noise or uncertainty. Bandit problems have been well studied in many fields, including statistics (Gittins, 1979), reinforcement learning (Kaebling, Littman, \& Moore, 1996; Sutton \& Barto, 1998), economics (Banks, Olson, \& Porter, 2013, e.g.), psychology and neuroscience (Daw, O'Doherty, Dayan, Seymour, \& Dolan, 2006; Cohen, McClure, \& Yu, 2007; Steyvers, Lee, \& Wagenmakers, 2009; Lee, Zhang, Munro, \& Steyvers, 2011). There is no analytical solution to the general bandit problem, though properties about the optimal solution of special cases are known (Gittins, 1979). For relatively simple, finite-horizon problems, the optimal solution can be computed numerically via dynamic programming (Kaebling et al., 1996), but its computational complexity grows exponentially with the number of arms and with the time horizon. In the psychology literature, a number of heuristic policies, with varying levels of complexity in the learning and control processes, have been proposed as possible strategies used by human subjects (Daw et al., 2006; Cohen et al., 2007; Steyvers et al., 2009; Lee et al., 2011). Most models assume that humans either adopt simplistic policies that retain little information about the past and sidestep longterm optimization (e.g. win-stay-lose-shift and \(\varepsilon\)-greedy), or switch between an exploration and exploitation mode either randomly (Daw et al., 2006) or discretely over time as more is learned about the environment (Steyvers et al., 2009).

Here, we analyze a new model for human bandit choice behavior, based on the knowledge gradient (KG) algorithm, which has been developed by Frazier, Powell, and Dayanik (2008) to solve problems in operations research. At each time step, the KG policy chooses, conditioned on previous observations, the option that maximizes future cumulative reward gain. It is based on the myopic assumption that the next observation is the last exploratory choice, used to learn about the environment, and all remaining choices will be exploitative, choosing the option with the highest expected reward by the end of the next trial. Note that this myopic assumption is only used in reducing the complexity of computing the predicted value of each option, and not actually implemented in practice - the algorithm may end up executing arbitrarily many non-exploitative choices. Despite a certain greedy aspect to the KG control policy, it is not completely short-sighted. In particular, it tends to explore more when the number of trials left is large, because finding an
arm with even a slightly better reward rate than the currently best known one can lead to a large cumulative advantage in future gain; on the other hand, when the number of trials left is small, KG tends to exploit and stay with the currently best known option, because it knows that finding a slightly better option will not lead to large improvement, while the risk of wasting time on a bad option is high. KG is also known to be exactly optimal in certain special cases (Frazier et al., 2008), such as when there are only two arms.
KG has several advantages over previously proposed algorithms. Unlike the simple heuristic algorithms such as win-stay-lose-shift and \(\varepsilon\)-greedy, and in common with the other Bayesian learning algorithms (Daw et al., 2006; Steyvers et al., 2009; Lee et al., 2011), KG uses a sophisticated Bayesian posterior distribution as its belief state at each time step. Unlike the other Bayesian learning algorithms, KG gracefully and gradually transitions from primarily exploring to primarily exploiting over the course of a finite-horizon bandit experiment. Also unlike previously proposed algorithms, which typically assumes that the stochastic component of action selection is random or arbitrary, KG also provides a more sophisticated and discriminating way to explore, by normatively combining immediate reward expectation and long-term knowledge gain. On the other hand, in contrast to the optimal algorithm, which scales exponentially in computational complexity with respect to the number of remaining timesteps, KG is computationally much simpler, incurring a constant cost regardless of the number of timesteps left.
In the following, we first describe the experiment, then describe all the learning and control models that we consider. We then compare the performance of the models both in terms of agreement with human behavior on a trial-to-trial basis, and in terms of computational optimality.

\section*{Data}

\section*{Participants}

A total of 451 participants completed a series of bandit problems as part of 'testweek' at the University of Amsterdam.

\section*{Experimental procedure}

Each participant completed 20 bandit problems in sequence, all problems had 4 arms and 15 trials. The reward rates for all games were generated independently from a \(\operatorname{Beta}(2,2)\) distribution, and were all done prior to data collection. All participants thus played games with the same sets of reward rates, but the order of the games was randomized. Participants were aware that the reward rates in all games were drawn from the same environment, but they were not told its form, i.e. Beta(2,2). A representation of the basic experimental interface is shown in Fig 1.

\section*{Modeling Methods}

There exist multiple levels of complexity and optimality in both the learning and the decision components of decision making models of bandit problems. For the learning component, we examine whether people learn any abstract representation of the environment at all, and if they do, whether


Figure 1: Experiment interface. The four panels correspond to the four arms, each of which can be chosen by pressing the corresponding button. In each panel, successes from previous trials are shown as green bars, and failures as red bars. At the top of each panel, the ratio of successes to failures, if defined, is shown. The top of the interface provides the count of the total number of successes to the current trial, index of the current trial and index of the current game.
they only keep a mean estimate (running average) of the reward rate of the different options, or also uncertainty about those estimates, or indeed more complex meta-information, such as the general abundance/scarcity of rewards. The decision component can also differ in complexity in at least two respects: the objective the decision policy tries to optimize (e.g. reward versus information), and the time-horizon over which the decision policy optimizes its objective (e.g. greedy versus long-term). In this section, we introduce models that encompass different combinations of learning and decision policies.

\section*{Bayesian Learning in Beta Environments}

The observations are generated independently and identically (iid) from an unknown Bernoulli distribution for each arm. We consider two Bayesian learning scenarios, either subjects have a fixed belief about the distribution from which the Bernoulli rates are drawn ("basic learning"), or they do meta-learning about the parameters of that distribution over time ("meta learning"). We explore the two scenarios below. In either case, we assume the distribution that generates the Bernoulli rates is a Beta distribution, Beta \((\alpha, \beta)\), which is a conjugate prior, and whose two hyper-parameters, \(\alpha\) and \(\beta\) are determined by the total number of rewards and failures experienced so far, plus any pseudo-counts associated with the prior.

Basic Learning Suppose we have \(K\) arms with reward rates, \(\theta_{k}^{g}, k=1, \cdots, K\), which are independent and identically drawn from \(\operatorname{Beta}(\alpha, \beta)\) for the \(g\) th game. On the \(t\) th trial, if the \(k\) th arm is chosen, a reward is attained with a

Bernoulli distribution, \(R_{k}^{t, g} \sim \operatorname{Bernoulli}\left(\theta_{k}^{g}\right)\). Let \(\mathbf{S}^{t, g}\) and \(\mathbf{F}^{t, g}\) be vectors of the number of successes and failures attained from each arm at the \(t\) th trial of the \(g\) th game. The model learns the individual reward rates using Bayes' Rule:
\[
\begin{aligned}
\operatorname{Pr}\left(\theta^{g} \mid \alpha, \beta, \mathbf{S}^{t, g}, \mathbf{F}^{t, g}\right) & \propto \operatorname{Pr}\left(\mathbf{S}^{t, g}, \mathbf{F}^{t, g} \mid \theta^{g}\right) \operatorname{Pr}\left(\theta^{g} \mid \alpha, \beta\right) \\
\theta^{g} & \sim \operatorname{Beta}(\alpha, \beta) \\
S_{k}^{t, g} & \sim \operatorname{Binomial}\left(S_{k}^{t, g}+F_{k}^{t, g}, \theta_{k}^{g}\right)
\end{aligned}
\]

The learner's belief state at the trial \(t\) of the game \(g, \mathrm{~B}^{t, g}\), is the set of posterior Beta distributions for each arm, and the mean reward on each arm, based on the observed sequence, is \(\hat{\theta}^{t, g}=\left(\alpha+S_{k}^{t, g}\right) /\left(\alpha+\beta+S_{k}^{t, g}+F_{K}^{t, g}\right)\).
Meta Learning We also consider the case that subjects may use observations to learn about the true environmental reward distribution (the true Beta distribution), corresponding to the general abundance/scarcity of resources in the environment. In this case, observing an outcome on any arm will affect the posterior distribution on all arms because of the correlation induced by shared hyper-parameters of the environment (Gelman, Carlin, Stern, \& Rubin, 2004):
\(\operatorname{Pr}\left(\theta^{g}, \alpha, \beta \mid \mathbf{S}^{t, g}, \mathbf{F}^{t, g}\right) \propto \operatorname{Pr}\left(\mathbf{S}^{t, g}, \mathbf{F}^{t, g} \mid \theta^{g}\right) \operatorname{Pr}\left(\boldsymbol{\theta}^{g} \mid \alpha, \beta\right) \operatorname{Pr}(\alpha, \beta)\)
The belief state on trial \(t\) of game \(g, \mathrm{~B}^{t, g}\), is a joint posterior distribution over the reward rates and environmental parameters, conditioned on the observed sequence.

\section*{Decision Policies}

We consider five different decision policies. We first describe the optimal model, and then the four heuristic models with increasing levels of complexity.
The Optimal Model The learning and decision problem for bandit problems can be instantiated as a Markov Decision Process with a finite horizon (Kaebling et al., 1996). Due to the low dimensionality of the bandit problem here (i.e. small number of arms and number of trials per game), the optimal policy, up to a discretization of the belief state, can be computed numerically according to Bellman's dynamic programming principle. Let \(V^{t}\left(\mathbf{S}^{t}, \mathbf{F}^{t}\right)\) be the expected total future reward on trial \(t\). The optimal policy should satisfy the following iterative property:
\[
V^{t, g}\left(\mathbf{S}^{t, g}, \mathbf{F}^{t, g}\right)=\max _{k} \mathbb{E}\left[V^{t+1, g}\left(\mathbf{S}^{t+1, g}, \mathbf{F}^{t+1, g}\right)\right]+\hat{\theta}_{k}^{t, g}
\]
and the optimal decision, \(D^{t, g}\), is decided by
\[
D^{t, g}\left(\mathbf{S}^{t, g}, \mathbf{F}^{t, g}\right)=\operatorname{argmax}_{k} \mathbb{E}\left[V^{t+1, g}\left(\mathbf{S}^{t+1, g}, \mathbf{F}^{t+1, g}\right)\right]+\hat{\theta}_{k}^{t, g}
\]

We solve the equation using dynamically programming, backward in time from the last time step, whose value function and optimal policy are known for any belief state, i.e. any setting of posterior Beta distribution for each of the arms: it always choose the arm with the highest expected reward, \(\hat{\theta}^{T, g}\), and the value function is just that expected reward. In the simulations, we compute the optimal policy offline, for any conceivable setting of belief state on each trial
(up to a fine discretization of the belief state space), and then apply the computed policy for each sequence of choice and observations that each subject experiences. We use the term "the optimal solution" to refer to the specific solution under \(\alpha=2\) and \(\beta=2\), which is the true experimental design.

Win-Stay-Lose-Shift WSLS does not learn any abstract representation of the environment, and has a very simple decision policy. It assumes that the decision-maker continues to choose an arm following a reward, but shifts to other arms (with equal probabilities) following a failure to gain reward.
\(\varepsilon\)-Greedy The \(\varepsilon\)-greedy model assumes that decisionmaking is driven by a parameter \(\varepsilon\) that controls the balance between random exploration and exploitation inherent in bandit problems. On each trial, with probability \(\varepsilon\), the decision-maker chooses randomly (exploration), otherwise chooses the arm with the greatest estimated reward rate (exploitation). \(\varepsilon\)-Greedy keeps simple estimates of the reward rates, but does not track the uncertainty of the estimates. It is not sensitive to the horizon, maximizing the immediate gain with a constant rate, otherwise searching for information by random selection \({ }^{1}\).
We call the situation \(k \in \operatorname{argmax}_{k^{\prime}} \hat{\theta}_{k}^{t, g}\) 'case 1 ', and the \(\varepsilon\)-greedy model is implemented as
\[
\operatorname{Pr}\left(D^{t, g}=k \mid \varepsilon, \hat{\theta}^{t, g}\right)= \begin{cases}(1-\varepsilon) / M^{t, g} & \text { if case 1 } \\ \varepsilon /\left(K-M^{t, g}\right) & \text { otherwise }\end{cases}
\]
where \(M^{t, g}\) is the number of arms with the greatest estimated value at the \(t\) th trial of the \(g\) th game.
\(\varepsilon\)-Infomax The \(\varepsilon\)-infomax model is similar to the \(\varepsilon\)-greedy model in that it chooses the arm with the greatest estimated reward rate with probability \(1-\varepsilon\), and explores with probability \(\varepsilon\). The difference is that, instead of random selection for exploration, it selects the arm that results in the largest reduction in the expected total entropy. In our study, the arms are independent given the same environmental distribution, and the policy reduces to choose the arm with the largest uncertainty. We use \(S_{k}^{t, g}+F_{k}^{t, g}\) as an approximate, simple measure of the uncertainty associated with arm \(k\) given the state of the game. In this model, an arm may be chosen when one of the two cases applies: in case 1 , it has the greatest estimated reward rate; in case 2 , it does not have the greatest estimated reward rate, but has the least number of times being chosen. We implement \(\varepsilon\)-infomax as
\[
\operatorname{Pr}\left(D^{t, g}=k \mid \varepsilon, \hat{\theta}^{t, g}\right)= \begin{cases}(1-\varepsilon) / M^{t, g} & \text { if case 1 } \\ \varepsilon / N^{t, g} & \text { if case 2 } \\ 0 & \text { otherwise }\end{cases}
\]
where \(M^{t, g}\) and \(N^{t, g}\) are the number of arms that satisfy case 1 and 2 , respectively, at the \(t\) th trial of the \(g\) th game.
The \(\varepsilon\)-infomax model uses both the mean estimates and measure of uncertainty as criteria for action selection. It is a

\footnotetext{
\({ }^{1}\) The \(\varepsilon\)-Greedy model has a variant, \(\varepsilon\)-decreasing, where the probability of random selection decreases over trials. However, previous studies found that \(\varepsilon\)-decreasing had a poor fit to the same data when compared with the \(\varepsilon\)-greedy model (Zhang \& Lee, 2010), so we only consider the \(\varepsilon\)-greedy model in this study.
}
greedy heuristic, maximizing the immediate reward gain at a constant rate.

Knowledge Gradient The knowledge gradient (KG) algorithm (Ryzhov, Powell, \& Frazier, 2012) is an approximation to the optimal policy, by pretending only one more exploratory measurement is allowed, and assuming all remaining choices will exploit what is known after the next measurement. It evaluates the expected change in each estimated reward rate, if a certain arm were to be chosen, based on the current belief state. Its mathematical expression is
\[
v_{k}^{\mathrm{KG}, t}=\mathbb{E}\left[\max _{k^{\prime}} \hat{\theta}_{k^{\prime}}^{t+1} \mid D^{t}=k, B^{t}\right]-\max _{k^{\prime}} \hat{\boldsymbol{\theta}}_{k^{\prime}}^{t}
\]

The first term is the expected largest reward rate on the next step if the \(k\) th arm were to be chosen, with the expectation taken over all possible outcomes of choosing \(k\). The KG decision rule is
\[
\begin{equation*}
D^{\mathrm{KG}, t, g}=\underset{k}{\arg \max _{k}} \hat{\theta}_{k}^{t, g}+(T-t-1) v_{k}^{\mathrm{KG}, t, g} \tag{1}
\end{equation*}
\]

The first term of Equation 1 denotes the expected immediate reward by choosing the \(k\) th arm at \(t\) of the \(g\) th game, whereas the second term reflects the expected gain of total remaining reward from \(t+1\) to the last trial of the current game. The formula for calculating \(v_{k}^{\mathrm{KG}, t, g}\) for the binary bandit problems can be found in Chapter 5 of Powell and Ryzhov (2012).

\section*{Model Implementation and Agreement Calculation}

We used model agreement as a measure of how well it captures experimental data, which was calculated as the average per-trial likelihood, conditioned on the observed game states. We fit the models and calculated model agreement across all participants.
WSLS is a fully deterministic paradigm, so the per-trial likelihood is 1 for a win-stay decision, \(1 / 3\) for a lose-shift decision, and 0 otherwise. All other models have at least two free parameters, \(\alpha\) and \(\beta\), and the \(\varepsilon\)-greedy and the \(\varepsilon\)-infomax models have one additional parameter, \(\varepsilon\). We implemented the \(\mathrm{KG}, \varepsilon\)-greedy and \(\varepsilon\)-infomax models as Bayesian graphical models under both learning frameworks. We used a vague prior for the environmental parameters, \(\operatorname{Pr}(\alpha, \beta)=(\alpha+\beta)^{5 / 2}\), as suggested by Gelman et al. (2004), because it is uniform on the psychologically interpretable reparameterization, \(\alpha /(\alpha+\beta)\) and \((\alpha+\beta)^{-1 / 2}\). We used uniform prior for \(\varepsilon\). Model inference used combined sampling algorithm, with Gibbs sampling of \(\varepsilon\), and Metropolis sampling of \(\alpha\) and \(\beta\). All chains contained 3000 steps, with a burn-in size of 1000. All chains converged according the R-hat measure (Gelman et al., 2004). We calculated the model agreement as the proportion of same choices between the model and the data, based on the full posterior predictive distribution of choices given each observed state of the game. For this study, we implemented the optimal model only with basic learning because of the heavy computational load.


Figure 2: Model agreement with data simulated by the optimal solution under the correct prior of the environment. Each bar shows the agreement of a model combining the corresponding decision policy and the learning framework. For the \(\varepsilon\)-greedy (eG), \(\varepsilon\)-infomax (eINFO) and the KG models, the error bars show the standard errors of the average agreement based on a 4 -fold cross-validation. WSLS has no parameters to fit and does not rely on any learning framework.


Figure 3: Model agreement with human data. The figure is generated in the same way as for Figure 2, except for that the optimal model was only implemented with basic learning for this study.

\section*{Results}

\section*{Model Agreement with the Optimal Solution}

As shown in Figure 2, the KG algorithm, under either learning framework, is able to approximate the optimal solution well in terms of the average number of correct predictions. In this sense, the KG policy is 'process optimal'. \(\varepsilon\)-infomax outperforms the \(\varepsilon\)-greedy model, which implies that smarter exploration for information gain increases the optimality of the heuristic. The simple WSLS model achieves model agreement well above \(60 \%\). In fact, both WSLS and the optimal model do win-stay with probability 1 . The only situation that WSLS does not resemble the optimal behavior is when it shifts away from an arm that the optimal solution would otherwise stay with.


Figure 4: Behavioral patterns in the human data and the simulated data from a selection of the best- and worst-performing models. The four panels show the trial-wise probability of win-stay, lose-shift, choosing the greatest estimated value, and choosing the least known when it is an exploration trial, respectively. Probabilities are calculated based on simulated data from each model at their MAP estimate, and are averaged across all games and all participants. The optimal solution shown here uses the correct prior \(\operatorname{Beta}(2,2)\).

\section*{Model Agreement with the Human Data}

Figure 3 shows the average model agreement with human data. Overall, the type of decision policy, other than the learning framework, makes significant differences in the model agreement. However, a decision policy tends to do better under the meta learning framework - the \(\varepsilon\)-greedy model and the KG model have significantly greater model agreement with meta learning.

We next break down the overall behavioral performance into four finer measures: how often people adopt win-stay and lose-shift, how often they exploit, and whether they use random selection or search for the greatest amount of information during exploration. We compare three of our models that have the highest agreement with human data on these additional behavioral criteria. Figure 4 shows the model analysis results. We show the patterns of the human subjects, the optimal solution, the best performing decision policy (KG) under both learning frameworks, and the simplest WSLS.

The first panel probably contains the most interesting results. It shows the trialwise probability of staying with the same arm following a previous success. People show clear sub-optimality by not staying with the same arm after an immediate reward. In fact, obtaining a reward from any arm should always increase the estimated value of the chosen arm. Under the basic learning framework where unchosen arms do not change in value, this means the optimal decision process should always do win-stay. This is consistent with the curve of the optimal solution. As implied by Equation \(1, \mathrm{KG}\) considers the likelihood of an arm surpassing the known best value upon chosen, and weights this knowledge gain more heavily in the early stage of the game. In general, during the early trials, it chooses the second-best arm with a certain probability, not necessarily depending on the previous outcome. This explains the drop of the win-stay probability of KG during the early trials. When the learner is also updating its knowledge of the environment, a previous suc-
cess will cause the environment to appear more rewarding, making other arms more likely to surpass the current best arm.

The second panel shows the trialwise probability of shifting away following a previous failure. People, the optimal solution, and KG show a decline in this probability over trial. When the horizon is approaching, it becomes increasingly important to stay with the arm that is known to be reasonably good, even if it may occasionally yield in a failure, because it is increasingly important to maximize the reward on the current trial.

In general, the KG model with meta learning matches the second-order trend of human data. However, there still exists a big difference on the absolute scale, especially regarding the probability of staying with 'good' arms - in fact, the KG policy does win-stay and exploitation more often, and resembles the optimal solution more than the human data.

\section*{Model Performance in Cumulative Reward Collection}

Fig 5 shows a comparison of the distribution of average reward per trial achieved by the participants, the optimal solution, and the knowledge gradient model. When playing at their best fit parameterization based on the human data, KG with meta learning and WSLS achieve nearly identical reward distributions as the participants. Moreover, if we let KG with meta learning forward play under the correct prior knowledge of the environment, i.e. Beta \((2,2)\), it is able to achieve a nearly identical distribution as the optimal solution.

\section*{Discussion}

Our analysis supports the KG decision policy under the meta learning framework as a good fit to human data in bandit problems. Our result implies that people might learn the individual reward rates as well as the general environment,


Figure 5: Average reward achieved by the KG model forward playing the bandit problems with the same reward rates. KG achieves similar reward distribution as the human performance, with KG playing at its maximum a posteriori probability (MAP) estimate, \(\alpha=.1\) and \(\beta=.8\). KG achieves the same reward distribution as the optimal solution when playing with the correct prior knowledge of the environment.
and the shared, latent environment induces a special type of correlation among the bandit arms. The meta learning framework is a psychologically sensible improvement to basic learning, because correct knowledge of the environment can be critical for achieving the best performance, especially when the environment can change over time or contexts. For the decision component, our results support the KG policy, which optimizes the semi-myopic goal of maximizing future cumulative reward while assuming only one more time step of exploration and strict exploitation thereafter (but does not actually ever carry out that policy). The KG model under the more general learning framework has the largest proportion of correct predictions of human data, and can capture the trial-wise dynamics of human behavioral reasonably well. KG achieves similar behavioral patterns as the optimal model, and is computationally tractable, making it a plausible algorithm for human learning and decision-making

One remaining puzzle why human subjects tend to explore more often than policies that optimize the specific utility of the bandit problems. One possibility is that people believe the task environment can undergo stochastic changesand exhibit sequential effects due to recent trial history, as in many other psychological task contexts (Yu \& Cohen, 2009) . This would be an interesting line of future inquiry.

\section*{Acknowledgments}

We thank Mark Steyvers and Eric-Jan Wagenmakers for sharing the data on bandit problems for our analysis, and three anonymous reviewers for their valuable comments that helped improve the paper.

\section*{References}

Banks, J., Olson, M., \& Porter, D. (2013). An experimental analysis of the bandit problem. Economic Theory, 10, 55-77.
Cohen, J. D., McClure, S. M., \& Yu, A. J. (2007). Should I stay or should I go? Exploration versus exploitation. Philosophical Transactions of the Royal Society B: Biological Sciences, 362, 933-942.
Daw, N. D., O’Doherty, J. P., Dayan, P., Seymour, B., \& Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. Nature, 441, 876-879.
Frazier, P., Powell, W., \& Dayanik, S. (2008). A knowledgegradient policy for sequential information collection. SIAM Journal on Control and Optimization, 47, 24102439.

Gelman, A., Carlin, J. B., Stern, H. S., \& Rubin, D. B. (2004). Bayesian data analysis (2 ed.). Boca Raton, FL: Chapman \& Hall/CRC.
Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. Journal of the Royal Statistical Society, 41, 148-177.
Kaebling, L. P., Littman, M. L., \& Moore, A. W. (1996). Reinforcement learning: A survey. Journal of Artificial Intelligence Research, 4, 237-285.
Lee, M. D., Zhang, S., Munro, M., \& Steyvers, M. (2011). Psychological models of human and optimal performance in bandit problems. Cognitive Systems Research, 12, 164-174.
Powell, W., \& Ryzhov, I. (2012). Optimal learning (1 ed.). Wiley.
Robbins, H. (1952). Some aspects of the sequential design of experiments. Bulletin of the American Mathematical Society, 58, 527-535.
Ryzhov, I., Powell, W., \& Frazier, P. (2012). The knowledge gradient algorithm for a general class of online learning problems. Operations Research, 60, 180-195.
Steyvers, M., Lee, M. D., \& Wagenmakers, E.-J. (2009). A bayesian analysis of human decision-making on bandit problems. Journal of Mathematical Psychology, 53, 168-179.
Sutton, R. S., \& Barto, A. G. (1998). Reinforcement learning: An introduction. Cambridge, MA: MIT Press.
Yu, A. J., \& Cohen, J. D. (2009). Sequential effects: Superstition or rational behavior? In Advances in neural information processing systems (Vol. 21, p. 1873-1880). Cambridge, MA.: MIT Press.
Zhang, S., \& Lee, M. D. (2010). Cognitive models and the wisdom of crowds. In N. Taatgen \& H. van Rijn (Eds.), Proceedings of the 32th annual conference of the cognitive science society. Austin, TX.

\title{
Knowledge tracing and cue contrast: Second language English grammar instruction
}

\author{
Helen Zhao (helenz@cuhk.edu.hk) \\ Department of English, The Chinese University of Hong Kong Shatin, Hong Kong
}

\author{
Kenneth R. Koedinger (koedinger@cmu.edu) \\ Human-Computer Interaction Institute and Department of Psychology, Carnegie Mellon University 5000 Forbes Avenue, Pittsburgh, PA 15213 USA \\ John Kowalski (jkau@andrew.cmu.edu) \\ Department of Psychology, Carnegie Mellon University \\ 5000 Forbes Avenue, Pittsburgh, PA 15213 USA
}

\begin{abstract}
This paper introduces a cognitive tutor designed for second language grammar instruction. The tutor adopted Corbett and Anderson's (1995) Bayesian knowledge tracing model and provided adaptive training on the English article system. We followed the Competition Model (MacWhinney, 1997) and understood the article system as a galaxy of cues determining article usage on the basis of form-function mapping. Cues are in competition during language acquisition; hence cue contrast is predicted to be an effective instructional method. Seventy-eight students were randomly assigned to four article training conditions (to learn 33 cues) and a control condition (to write essays). We found that article-training groups significantly outperformed the control group in an immediate posttest and a delayed posttest. Specifically, our result also suggested that there was a significant interaction between cue contrast and cue type (definite vs. indefinite). Cue contrast promoted more learning on the indefinite cues (more difficult for learners). Knowledge tracing did not demonstrate such an interactional effect with cue types. Instead, it boosted the instructional effect promoted by cue contrast.
\end{abstract}

Keywords: knowledge tracing; cue contrast; cognitive tutor; second language acquisition; English article.

\section*{Introduction}

Since the mid-1990s Corbett \& Anderson's (1995) Bayesian knowledge tracing model has been widely used to model student knowledge in learning systems of various domains, including tutors for mathematics, computer programming, and reading skills (Baker et al., 2010). In recent years, there has been an emergence of tutoring systems designed to facilitate second language learning (MacWhinney, 1995; Pavlik \& Anderson, 2005). Among them we rarely find learning systems adopting Bayesian knowledge tracing to promote robust language learning (Koedinger, Corbett \& Perfetti, 2012).

This paper introduces a Bayesian tutorial system of grammar instruction applied in an English as a Foreign Language (EFL) context. The primary goal of this research has been to develop an adaptive vehicle for testing the efficacy of Bayesian Knowledge Tracing in this domain. Another feature of the tutor, which presents grammatical cues in contrast, is informed by the cognitive linguistic
theories of the Competition Model (MacWhinney, 1997). This paper discusses how these two areas of thought are blended to shape the design of the tutor and how they interact to influence learning effects. Specifically, the tutor targets the English article system, a difficult grammatical category for second language learners (Butler, 2002; CelceMurcia \& Larsen-Freeman, 1999).

\section*{Bayesian Knowledge Tracing}

Corbett and Anderson's (1995) Bayesian knowledge tracing assumes that at any given opportunity to use a rule within the software, there exists a probability that a student knows the rule and may either give a correct or incorrect response. A student who does not know a skill generally will give an incorrect response, but there is a certain probability (called \(p(G)\), the Guess parameter) that the student will give a correct response. Correspondingly, a student who does know a skill generally will give a correct response, but there is a certain probability (called \(p(S)\), the Slip parameter) that the student will give an incorrect response. Each rule has an initial probability \(\left(p\left(L_{0}\right)\right)\) of being known by the student, and at each opportunity to practice a skill, the student has a certain probability \((p(T))\) of learning the skill. Once these four parameters are set, the model can be used to predict student performance. Figure 1 illustrates the relationship between the four parameters in the Bayesian Knowledge Tracing Model.


Figure 1: Bayesian Knowledge Tracing Model (Corbett and Anderson, 1995)

The system's estimate that a student knows a rule at time \(\mathrm{n}\left(\mathrm{P}\left(L_{n}\right)\right)\) is continually updated every time the student responds (correctly or incorrectly) to a problem step. First,
the system calculates the probability that the student knew the rule before making the attempt, using the evidence from the current step. Then, taking this into account, it computes the probability that the student learned the rule during the problem step. The equations for these calculations are:
\[
\begin{aligned}
& P\left(L_{n-1} / \text { Correct }_{n}\right)=\frac{P\left(L_{n-1}\right) *(1-P(S))}{P\left(L_{n-1}\right) *(1-P(S))+\left(1-P\left(L_{n-1}\right) *(P(G))\right.} \\
& P\left(L_{n-1} / \text { Incorrect }_{n}\right)=\frac{P\left(L_{n-1}\right) * P(S)}{P\left(L_{n-1}\right) * P(S)+\left(1-P\left(L_{n-1}\right)\right) *(1-P(G))} \\
& P\left(L_{n} / \text { Action }_{n}\right)=P\left(L_{n-1} / \text { Action }_{n}\right)+\left(\left(1-P\left(L_{n-1} / \text { Action }_{n}\right)\right) * P(T)\right)
\end{aligned}
\]

To set the initial \(\mathrm{P}(\mathrm{S}), \mathrm{P}(\mathrm{G}), \mathrm{P}(\mathrm{T})\), and \(\mathrm{P}\left(L_{0}\right)\) parameters for each skill, we used data from a previous English article study to train each model (Zhao, 2012). During this previous study, we collected 10,523 student attempts at choosing the correct article, with each attempt labeled with the article rule applied. We used the "brute force" method (Baker et al, 2010) to utilize this data and arrive at the most likely parameter values. This method tries every possible combination of the four parameters in the grain size of 0.01 and for each combination, computes the sum of squared residuals (SSR). The parameter value combination that gives the best SSR for that rule are the ones we use to model the rule in the tutor.

\section*{The Competition Model}

An important feature of the tutorial system in this paper is its theoretical ground in cognitive linguistics. We adopted the Competition Model (MacWhinney, 1997) and integrated one of its fundamental principles (cue competition) into the design of grammar instruction.

The model presents a functionalist rather than nativist view of language acquisition and understands the linguistic sign as a set of mappings between forms and functions. Forms are the external phonological and word order patterns that are used in words and syntactic constructions. Functions are the communicative intentions or meanings that underlie language usage. Each lexical item or syntactic construction can be understood as a form-to-function mapping.

In the context of discussion of the English article system, there are four forms: the, \(a, a n\), and the zero article (0). The form the specifies the definite article; the forms \(a\) and an encode the indefinite article ( \(a\) is followed by noun phrases starting with consonants; an is followed by noun phrases starting with vowels); the form 0 is commonly known as the zero article or null article.

How about functions in the English article system? Adopting the model, we carried out a functional linguistic analysis and found that the four article forms are mapped with approximately 90 different functions or usages (more information see Zhao, 2012). Some functions are syntactic and semantic properties (e.g., countability, singularity, plurality); some functions are discourse-based properties
(e.g., first mention, second mention, immediate situation); many functions are idiosyncratic surface features whose usage is highly conventional (e.g., names of rivers, lakes, malls, parks, bridges, theatres); and some functions combine both syntactic and idiosyncratic properties (e.g., names of singular mountains or plural mountains). So many functions are mapped with only four forms. This complex formfunction mapping is one of the critical reasons why English articles are difficult to acquire.

The Competition Model understands one form-function mapping as a unit or a cue. E.g., the tag "the - river names" represents a cue since it maps the association between the form the and the semantic property of names of rivers. But the tag "the - river names and second mention in the discourse" is not a valid cue because one form is mapped with two functions. In this case, the form-function mapping needs to be broken down to the smallest unit.

The basic claim of the Competition Model is cue competition. It considers cue competition as vital for language acquisition. Sentences (1-2) illustrate how cues compete and one gains dominance over another. In sentence (1), the zero article is required because "wealth" is a noncountable mass noun and is used alone with no modifiers. When this noncountable noun is modified by the prepositional phrase (PP) "of her parents" in sentence (2), the noun "wealth" becomes concrete and identifiable. Hence, the PP structure (strongly associated with the) overrides noncountability (strongly associated with 0 ) in sentence (2).
(1) Alice is interested in \(\underline{0}\) wealth.
(2) Alice is interested in the wealth of her parents.

Early stages of language acquisition focus on obtaining input on individual cues from the learning environments. Learners may not know when a cue can or cannot override another cue. As their language proficiencies increase, learners develop their skills of interpreting cue conflict. Some cue competitions are easier to interpret than others. For example, the competition between the cue " \(a\) - first mention" and the cue "the - second mention" is relatively easier to interpret and are among the earliest acquired cues in the article system. But the competition between the two cues in Sentence (1-2) is relatively harder to interpret. Learners need to know the grammatical properties of a PP structure, its strong association with the definite article, and how a PP structure typically interacts with mass nouns.

This paper proposes an instructional invention of cue contrast that originates from the basic claim of the Competition Model. There are two theoretical justifications for the proposal. First, regularities, and heuristics are always good for learning a complex problem space (Reber et al., 1980; Ellis, 2011). Instead of understanding the article system as a space with almost 90 unrelated usages, learners formulate a more organized mental space with more than 40 contrasting pairs of usages. That leads to the most important advantage of cue contrast: it significantly reduces learners'
memory load and storage cost, and consequently increases their learning capacity. One major theoretical commitment made in the Competition Model is to a capacity-limited model of language processing. This account treats sentence interpretation as a constraint satisfaction process that balances the limitations imposed by verbal memory against the requirements of conceptual interpretation. Our raw memory for strings of nonsense words is not more than about four. However, when words come in meaningful groups, we can remember dozens of words, even when the message is unfamiliar. The most likely candidate for this additional storage is some form of conceptual representation. By presenting article usages to learners as meaningful groups, we help learners form the conceptual representation of the article system. In turn, storage cost is reduced.

\section*{Method}

\section*{Participants}

The participants of the current experiment were 78 (31 males, 47 females) Chinese intermediate-advanced learners at a public university in Beijing that specialized in foreign language education and research. Their average years spent learning English was 7.8 years.

\section*{Materials and Design}

A 2 (Contrast: yes vs. no) \(\times 2\) (Knowledge tracing: yes vs. no) \(\times 2\) (Cue type: the vs. \(0 / a / a n)^{1}\) mixed model design was used in this experiment. Cue contrast and knowledge tracing are two between-subject variables, whereas cue type is a within-subject variable.

There were five conditions in the experiment (Table 1): four experimental groups who received article training; one control group who did not receive article training. The four experimental conditions were manipulated based on the two between-subject variables. The control group spent roughly the same amount of training time as the four experimental groups. Instead of learning articles, they were asked to write four English essays during the training sessions and weren't given feedback.

Table 1: Experiment conditions.
\begin{tabular}{ccc|c}
\hline & \multicolumn{2}{c|}{ Knowledge Tracing (KT) } & \\
\cline { 2 - 3 } \begin{tabular}{c} 
Cue \\
Contrast \\
(CC)
\end{tabular} & KT-CC & noKT-CC & \begin{tabular}{l} 
Control \\
group
\end{tabular} \\
\hline
\end{tabular}

Article training was provided as a sentence-level cloze task. Figure 2 exemplifies training received by the two cue contrast groups. Students were given a prompt question and two sentence items that belonged to a pair of contrasting cues. They used a pull down menu to make choices and

\footnotetext{
\({ }^{1}\) We grouped indefinite article cues with zero article cues due to the small number of \(a / a n\) cues.
}
were given immediate feedback. The feedback included correct/wrong, cue name, explicit cue explanations, and examples. Training pages in the two no-cue-contrast conditions also included two sentences. But they cannot belong to a contrasting pair.


Figure 2: A training page of cue contrast conditions.
This experiment trained 33 article cues that were grouped into 18 pairs of contrasting cues. Definite article cues were paired with indefinite article or zero article cues primarily based on structural distinction. Table 2 exemplifies five representative cue pairs. In the first and second pairs, article choices are distinguished based on the existence of a relative clause as a post-modifier. The distinction in the third pair is due to the "of ..." prepositional phrase as a post-modifier. The contrast in the four pair originates from the singular vs. plural distinction of mountain names. When we could not rely on structural distinction to create cue pairs, we relied on semantic distinction. The 'hall' and 'building' cues in the fifth pair, for example, are idiosyncratic cues whose article choices can only be explained by historical conventions. Hence, we relied on their semantic distinction to manipulate the contrast.

Table 2: Exemplar Cue Pairs (With Examples)
\(\left.\begin{array}{ccc}\hline 1 & a-\text { singular countable } \\ (a \text { store })\end{array} \begin{array}{c}\text { the }- \text { singular countable }+ \text { clause } \\ \text { (the store she bought the dress) }\end{array}\right]\)

\section*{Implementing Bayesian Knowledge Tracing}

We adopted Corbett and Anderson's (1995) Bayesian knowledge tracing algorithm and used Baker et al.'s (2010) approach to train the model parameters using learner data from a previous but similar article tutor study (Zhao, 2012).

With the models for each article skill trained this way, the tutor updates \(\mathrm{P}\left(L_{n}\right)\) after observing correct/incorrect attempts at each skill and uses \(\mathrm{P}\left(L_{n}\right)\) for the item selection criteria. If we simply chose the next item to present with the lowest \(\mathrm{P}\left(L_{n}\right)\), the tutor would frequently show the same item back-to-back. To introduce some variety while still giving more practice on the items least learned, we use \(\mathrm{P}\left(L_{n}\right)\) to set the selection criteria of training items as: random selection in proportion to percent unlearned. Percent unlearned for a rule is calculated by taking the probability the rule is "unlearned": \(\left(1-\mathrm{P}\left(L_{n}\right)\right)\), divided by the sum of the probabilities each other rule is "unlearned". A difficult cue will have a higher percent unlearned compared to other cues and will thus be more likely to be chosen next. But it is still possible (but less likely) that a better-acquired cue will be chosen next. This selection criterion avoids over-training of unlearned cues and under-training of better-acquired cues.

\section*{Procedure}

The study was composed of three sessions. Session I included a pretest ( \(25-\mathrm{min}\) ) and the first training session (1 hour). Session II (2 days later) included the second training session ( 1 hour) and an immediate posttest ( \(25-\mathrm{min}\) ). Session III (2 weeks after Session 2) included a delayed posttest ( \(25-\mathrm{min}\) ). All sessions were administered online. The tests were also in the format of a sentence-level cloze.

\section*{Results}

\section*{Instructional Effects}

A univariate analysis indicates that there was no significant difference of mean accuracy between the article training groups and the control group in the pretest ( \(F=.207, p=\) .651). These two groups showed a significant difference of mean accuracy in the immediate posttest ( \(F=37.836, p<\) .001). The article training groups gained a mean accuracy of \(.154(\mathrm{SD}=.071)\), whereas the control group only gained a mean accuracy of \(.026(\mathrm{SD}=.076)\). Figure 2 illustrates the learning trajectories of the above conditions.

A paired samples \(t\)-test shows that the article training groups significantly improved mean accuracy from pretest to immediate posttest ( \(T=17.156, p<.001\) ). Though they had a significant drop of accuracy from immediate-posttest to delayed-posttest ( \(T=-3.774, p<.001\) ), their mean accuracy in the delayed posttest was significantly higher than their pretest level ( \(T=13.374, \mathrm{p}<.001\) ). This suggested that the article training groups retained learning two weeks after training.

A paired samples t-test suggests that the control group did not show improvement from pretest to immediate posttest ( \(T=1.359, p=.196\) ). Neither did they improve mean
accuracy from immediate-posttest to delayed-posttest ( \(T=\) 1.361, \(p=.195\) ).


Figure 2: Mean test accuracies of conditions.

\section*{Cue Contrast, Knowledge Tracing, Cue Type}

Figure 3 compares the four article training conditions (KTCC, KT-noCC, noKT-CC, noKT-noCC) in terms of mean accuracies of all the cues in the pretest, posttest, and delayed posttest. In the pretest, univariate analysis indicates no significant difference of mean accuracy among the four conditions when all the cues are examined ( \(F=.468, p=\) \(.759, \eta 2=.025\) ). Meanwhile, univariate analyses of pretest accuracy also show no significant difference among the four conditions regarding the acquisition of the-cues \((F=1.083\), \(p=.371\) ) and \(0 / a / a n\)-cues ( \(F=.441, p=.778\) ).


Figure 3: Mean accuracies of three tests of four conditions
In the immediate posttest, univariate analyses suggest no significant difference between contrast and no-contrast ( \(F=\) \(.110, p=.741, \eta 2=.002\) ) and between knowledge-tracing and no-knowledge-tracing ( \(F=2.435, p=.124, \eta 2=.038\) ), when all the article cues are examined. The same pattern was found in the delayed posttest. We then delved into four conditions and looked for variations. Again, when all article cues were considered, we could not find significant differences among the four conditions in the immediate
posttest ( \(F=2.165, p=.102, \eta 2=.099\) ) and in the delayed posttest ( \(F=1.464, p=.234, \eta 2=.069\) ).

Our next step was to explore interaction between the two primary variables (KT, CC) and cue type (the vs. 0/a/an). A repeated measure ANOVA suggested that there was a significant interaction (Figure 4) between cue contrast and cue type ( \(F=9.744, p<.001, \eta 2=.138\) ). Bonferroni pairwise comparisons indicated that within each cue type, the differences between cue contrast and non-cue contrast were significant: the-cues ( \(p<.001\) ) and 0/a/an-cues ( \(p<\) .001 ). Cue contrast promoted significantly more learning ( \(p<.001\) ) of 0/a/an cues than of the cues. In contrast, noncue contrast created a more balanced instructional effect. Its instructional effect on the cues and \(0 / a / a n\) cues were not significantly different ( \(p=.844\) ).


Figure 4: Contrast versus non-contrast
Yet, we did not find significant interaction between knowledge tracing and cue type. Bonferroni pairwise comparisons suggested no significant differences between knowledge tracing and non-knowledge tracing within the cues ( \(p=.557\) ) or \(0 / a / a n\) cues ( \(p=.385\) ).
\begin{tabular}{|c|c|c|}
\hline \multirow[t]{9}{*}{} & & \\
\hline & & \\
\hline & & \\
\hline & & \\
\hline & & \\
\hline & & \\
\hline & & \\
\hline & & \\
\hline & the & 0/a/an \\
\hline - KT-CC & -0.069 & 0.291 \\
\hline - KT-noCC & 0.17 & 0.115 \\
\hline - noKT-CC & 0.064 & 0.269 \\
\hline \(\cdots\) noKT-noCC & 0.128 & 0.225 \\
\hline
\end{tabular}

Figure 5: Gained accuracy (immediate posttest - pretest) of the-cues and \(0 / a / a n\)-cues in four training conditions

Figure 5 illustrates an interaction between cue types (the vs. O/a/an) and the four training conditions. The Y-axis is gained mean accuracy (immediate posttest-pretest). As we can see, the cue contrast conditions (KT-CC, noKT-CC) pushed for more learning of the \(0 / a / a n\) cues. In particular, KT-CC had a more tilted slope leaning towards more learning on \(0 / a / a n\), which sharpened the interactional effect between cue contrast and article forms. Meanwhile, the noncue contrast conditions (KT-noCC, noKT-noCC) were suggested to promote more learning on the cues. This trend appears to be more obvious in the KT-noCC condition than in the noKT-noCC condition.

The above findings indicated that cue contrast played the primary role of determining patterns of results. Knowledge tracing did not change the interactional effects between cue contrast and article type. It gave the interactional effects a boost and made the patterns sharper.

\section*{Discussion and Conclusion}

The intermediate-advanced students in this study were given a difficult task of learning 33 article cues within two hours of training. The article training groups managed to show significant learning in the immediate posttest and retained learning two weeks later. This positive instructional effect confirms two principles that MacWhinney (1995) suggested to be important in designing and evaluating foreign language tutoring systems: 1) practice makes perfect and 2) feedback promotes learning (p. 318-319). Our study clearly demonstrated that given a difficult learning task, students would learn when they are given enough practice trials, accurate and digestible feedback, and an effective instructional method that helped to reduce memory and learning loads. The study implications for a third principle of MacWhinney (1995), that 3) cue conflicts are crucial for learning, are less clear.

As indicated by the interaction with type of item ('the' vs. \(0 / \mathrm{a} / \mathrm{an}\) ), the cue contrast manipulation had a clear impact, raising performance on 0/a/an-cues but lowering performance on the-cues. By analyzing contrasting cue pairs, it was hypothesized that learners in the cue contrast condition would formulate a new understanding of the article system and develop a more systematic knowledge space. Knowledge tracing effects, in comparison, would be less visible. Indeed, KT did not change how learners conceptualized the article system. It mainly functioned to escalate learning.

Why did only cue contrast show interaction with cue type (the vs. O/a/an)? Because that was the way the tutor was designed. Contrasting pairs were created because they shared similar features (e.g., mountain names) but required different article forms (e.g., 0-single mountain names, theplural mountain names). But knowledge tracing was not manipulated based on article type.

And why did cue contrast promote more learning on O/a/an cues than the cues? At first we suspected an ordering effect behind this interactional effect, i.e., cue contrast made learners pay more attention to the first item on a training
screen and consequently paid less attention to the second item. But in fact there were more the items than O/a/an items at the top of the training screen. Meanwhile, a repeated measure ANOVA also suggested no significant interaction between cue contrast and item sequence ( \(F=\) 2.427, \(p=.124\) ) and no main effects of item sequence ( \(F=\) 1.311, \(p=.991\) ) or of cue contrast ( \(F=.049, p=.825\) ). Also we suspected that cue contrast groups got a higher frequency of exposure to \(0 / \mathrm{a} / \mathrm{an}\) items. Yet this was not confirmed either. Frequency of exposure to the-items and \(0 / a /\) an-items was balanced between contrast and no-contrast conditions.

Then the most plausible account was that cue contrast illuminated weak areas of learning. 0/a/an cues were poorly acquired by learners. Their pretest mean accuracy of \(0 / a / a n\) cues ( \(M=.472\) ) was significantly lower ( \(p<.001\) ) than the cues \((M=.761)\). We found that a particular problem associated with the zero article acquisition was due to a misunderstanding made by students \({ }^{2}\). They thought that all the so-called "proper nouns" (e.g., Lake Michigan, the Colorado River, Baker Hall, the Tepper Building) were unique and therefore had to be used with the definite article. They did not know that some of such noun phrases required the zero article. Therefore, it became enlightening moments for students to see two proper nouns being contrasted in one screen. They allocated more time and attention to the zero article proper noun. As we can see, the mechanism behind cue contrast was cue focusing that directed students' attention to the right areas.

Due to the interaction between contrast and cue type, there was a trade-off that cancelled the overall instructional effect of cue contrast when all cues were considered.

The last question to discuss is the reason why knowledgetracing groups did not outperform non-knowledge-tracing groups. The most plausible reason was the relatively short duration of instruction. Two hours might not be enough for knowledge tracing to demonstrate its full advantages. Learners in the knowledge tracing conditions were in the middle of tackling the most difficult cues when training ended. They did not have enough time to work on the less difficult cues. The posttest mean accuracy (.744) of the article training groups was far from the mastery level (.950). This sent us a stronger signal that a longer training time was needed for knowledge tracing to be more effective.

In short, this study demonstrated a successful application of cognitive psychology and human-computer interaction theories in second language grammar instruction. We found that cue contrast was an effective method in teaching English article usages to adult second language learners. In particular, contrast allowed learners to become aware of and shift focus to problematic knowledge domains. Knowledge tracing boosted instructional effects of cue contrast. More research is needed to further specify duration of instruction so that we can make the best use of knowledge tracing in second language grammar instruction.

\footnotetext{
\({ }^{2}\) We carried out semi-structured interviews with selected students. Due to space limit, the interview data is not reported here.
}

\section*{Acknowledgments}

We thank Brian MacWhinney, Naoko Taguchi, Yasuhiro Shirai, Like Li, and Howard Seltman for their help in conducting this research. This work is supported by the Pittsburgh Science of Learning Center which is funded by the National Science Foundation award number SBE0354420 .

\section*{References}

Baker, R. et al. (2010). Contextual Slip and Prediction of Student Performance After Use of an Intelligent Tutor. Proceedings of the 18th Annual Conference on User Modeling, Adaptation, and Personalization, 52-63.
Butler, Y. G. (2002). Second language learners' theories on the use of English articles. Studies in Second Language Acquisition, 24, 451-480.
Celce-Murcia, M. \& Larsen-Freeman, D. (1999). The Grammar Book: An ESL/EFL Teacher's Course (2 \({ }^{\text {nd }}\) Edition). U.S.: Heinle \& Heinle Publishers.
Corbett, A. T., \& Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. User Modeling and User-Adapted Interaction, 4, 253-278.
Ellis, N. (2011). Implicit and explicit SLA and their interface. In Sanz, C. \& Leow, R. P. (Eds.), Implicit and Explicit Language Learning: Conditions, Processes, and Knowledge in SLA and Bilingualism. Washington, DC: Georgetown University Press.
Koedinger, K. R., Corbett, A. C., \& Perfetti, C. (2012). The Knowledge-Learning-Instruction (KLI) framework: Bridging the science-practice chasm to enhance robust student learning. Cognitive Science, 36, 757-798.
MacWhinney, B. (1995). Evaluating foreign language tutoring systems. In V. M. Holland, J. D. Kaplan \& M. R. Sams (Eds.), Intelligent Language Tutors: Theory Shaping Technology. Mahwah, NJ: Lawrence Erlbaum.
MacWhinney, B. (1997). Second language acquisition and the Competition Model. In J. Kroll \& A. De Groot (Eds.), Tutorials in Bilingualism. Mahwah, NJ: Lawrence Erlbaum.
Pavlik Jr., P. I. \& Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: An activationbased model of the spacing effect. Cognitive Science, 29, 559-586.
Reber, A. S. et al. (1980). On the relationship between implicit and explicit modes in the learning of a complex rule structure. Journal of Experimental Psychology: Human Learning and Memory, 6, 492-502.
Zhao, H. (2012). Explicitness, Cue Competition, and Knowledge Tracing: A Tutorial System for Second Language Learning of English Article Usage. Unpublished PhD Dissertation, Carnegie Mellon University.

\title{
Word Reading Practice Reduces Stroop Interference in Children
}

\author{
Laoura Ziaka (laoura.ziaka@yahoo.gr) \\ Despoina Moirou (moimina1984@yahoo.gr) \\ Graduate Program in Basic and Applied Cognitive Science, University of Athens \\ Ano Ilissia University Campus, GR-15771 Zografos, Greece \\ Eleni L. Vlahou (eleni.vlahou@uni-ulm.de) \\ Clinical and Biological Psychology, Institute of Psychology \& Education University of Ulm, 89069 Ulm, Germany
}

Athanassios Protopapas (aprotopapas@phs.uoa.gr)
Department of Philosophy \& History of Science (MITHE)
Ano Ilissia University Campus, GR-15771 Zografos, Greece

\begin{abstract}
Stroop interference is often explained by an automaticity account, according to which it arises due to more extensive practice in reading than in color naming. Here we investigated the effect on interference of isolated practice in color naming (of incongruent and neutral stimuli) and in word reading (of color names) in adults and children in Grades 4-5. In both groups interference was reduced after practicing color naming of incongruent stimuli. For children, interference was also reduced after practice in word reading of color names. In neither group was interference diminished after practice in color naming of neutral stimuli. These findings are consistent with a negative relationship between reading ability and interference and challenge the automaticity account.
\end{abstract}

Keywords: Stroop Interference; Reading; Color Naming; Training; Automaticity

\section*{Stroop Interference and Reading Ability}

One of the most familiar, most studied and most cited phenomena in cognitive psychology is Stroop interference. It is a robust finding that it takes longer to name the color of the ink in which a word is printed when the stimulus is a word denoting a different color (e.g. the word "red" printed in green ink) than when the stimulus is a string of colored letters (e.g. XXX) or a plain rectangular patch. MacLeod (1991) reviewed the evidence for variants of this task and addressed three issues that are crucial for understanding the causes of the Stroop effect: practice, integration, and the relation between facilitation and interference.

For the purposes of our analysis the concepts of practice and automaticity are of major importance. The notion of automaticity has been central for understanding and explaining the Stroop effect, since it is generally considered obligatory to read the word but not to name the color. This imbalance is thought to arise from our extensive practice in reading but not in color naming. The automaticity account considers Stroop interference as a consequence of a betterpracticed skill, namely reading, that dominates color naming without regard of attention (Logan, 1997; MacLeod, 1991). Cohen, Dunbar, and McClelland (1990) proposed a connectionist model whereby practice in a task such as
reading strengthens its connections and allows it to interfere with other tasks that have weaker connections, like color naming. Practice is seen as influencing the relative level of automaticity of two dimensions, resulting in interference. From this point of view reading and color naming differ along a continuum of practice and the degree of interference in the Stroop task may reflect the degree of word reading automaticity (Logan, 1997; Samuels, 1999).

Consequently, the degree of interference is often taken as a marker of word reading automaticity. This has significant practical implications, taking into account that word reading automaticity is been considered a fundamental element of reading development and as an essential background for more complex processes like reading comprehension (Kuhn \& Stahl, 2003; Wolf \& Katzir-Cohen, 2001).

One prediction that derives from the automaticity account is that poorer and less skilled readers will exhibit less interference in the standard Stroop task than good readers as a result of less practiced and therefore less automatized reading. This prediction stands in contrast to empirically observed data. A number of studies have showed that poor readers produce robust interference (Alwitt, 1966; Everatt, Warner, \& Miles, 1997; Helland \& Asbjørnsen, 2000; Kelly, Best, \& Kirk, 1989; van der Schoot, Licht, Jorsley, \& Sergeant, 2000). Everatt et al. (1997) found that children with dyslexia exhibit more interference than age-matched controls. More recently, Protopapas, Archonti, and Skaloumbakas (2007) showed that reading ability is negatively related to Stroop interference. In a first study they compared children with dyslexia to age-matched controls in the Stroop task and reported greater interference for the children with dyslexia. In a second study they examined the relationship between interference and reading skills in the general school population and found that poorer reading skills were associated with more interference. Furthermore, interference was found to be primarily associated with reading speed. Protopapas et al. suggested that reading ability and interference are directly linked, without mediation from executive functions such as attention and inhibition.

In line with these observations, Faccioli et al. (2008) also found that 7 - to 12 -year-old Italian children with dyslexia exhibited larger interference than a control group. The same pattern of results was reported by Kapoula et al. (2010) for French teenagers with dyslexia. These findings run counter to the automaticity account and challenge the notion that interference can be used to assay word reading automaticity.

In place of automaticity, a blocking mechanism, as implemented in the computational model WEAVER++ (Roelofs, 2003), may account for the empirical findings. According to this model, word reading can directly activate lemma retrieval and word-form encoding, whereas colornaming must pass through conceptual identification before activating lemma retrieval and word-form encoding. Stroop interference is assumed to occur because color naming must wait until the incorrect response (from word reading) is suppressed. As a consequence of the basic assumptions of the model, greater reading speed leads to less Stroop interference, consistent with the data.

\section*{Training}

Although the effects of practice on the Stroop task have been of concern since the very beginning, only a small number of studies have examined this relationship. Stroop (1935) himself tried to examine the development of interference through practice in color naming and found that interference can be present even after 8 days of practice.

MacLeod and Dunbar (1988) trained participants to respond to shapes with familiar color names. Prior to practice, "shape naming" suffered from interference from incompatible colors, an effect that was reversed 20 days later. After training, participants showed interference in naming the colors but not in naming the shapes.

Other studies have used mixed or inconguent-only stimuli and have succeeded to reduce but not eliminate interference (Davidson, Zacks, \& Williams, 2003; Dulaney \& Rogers, 1994; MacLeod, 1998). However, so far no study has examined the effects of practice on the individual task dimensions involved in Stroop interference, namely plain color word reading and color naming of neutral stimuli.

\section*{Rationale of the Present Study}

In light of the aforementioned findings the aim of the present study was to examine how interference is affected by practice not only in color naming but also in word reading. According to the automaticity account, practice in color naming should make this otherwise unpracticed dimension more competitive and thereby decrease interference, while practice in word reading might strengthen an already practiced dimension and thereby further increase interference, at least when reading skill has not yet reached ceiling performance. However, if the relationship between reading ability and interference is negative, as proposed by Protopapas et al. (2007), practice in color naming should have little effect on interference, because the bottleneck causing interference does not involve the processing of color but the delay in rejecting the word.

In contrast, practice in word reading could lead to a reduction of interference insofar as there is any potential for increase in the speed of reading the color words.

Adults are skilled readers and are considered to have achieved a high level of word reading automaticity. Therefore training in word reading should not affect interference in this population. In contrast, word reading automaticity has only partially developed in children and can improve further through practice, leading us to the prediction of a reduction in Stroop interference through practice in word reading.

\section*{Method}

\section*{Participants}

The study included adults and children from the general school population. The adult sample comprised 92 volunteers 18-40 years old, including 25 in the incongruent color group, 23 in the neutral color group, 22 in the word group, and 22 in the control (no-training) group. The school sample consisted of 105 children attending Grades 4-5, including 22 in the incongruent color group, 26 in the neutral color group, 31 in the word group, and 26 in the control group.

\section*{Materials}

The Greek words for red (ко́ккıvo /kocino/), green ( \(\pi \rho \alpha ́ \sigma\) vo /prasino/), and yellow (кít \(\rho \imath v o /\) citrino/) were used, because they have the same number of letters and syllables, comparable written frequency, and begin with voiceless stops, which facilitate response time triggering. The corresponding colors are familiar and easily distinguishable. The color word condition included these three words in white font.

Stimuli for the neutral color condition were made up of 7 repetitions of the letter X (no spaces) in red, green, and yellow color (RGB \#FF0000, \#00FF00, and \#FFFF00, respectively). For the incongruent condition the Greek words for red, green and yellow appeared in a non-matching color. All stimuli were presented on a black background.

\section*{Procedure}

Testing. On Day 1 the first interference measurement was taken. Participants were asked to name the color of the ink as quickly as possible and to try to avoid errors. A blocked design was implemented to minimize errors. The neutral condition was administered first (24 stimuli, including 8 in each color), followed by the incongruent condition (24 stimuli, including 4 in each mismatching word-color combination). The number of test trials was determined in a pilot study, to minimize learning due to testing. Each stimulus appeared on the screen for up to 2 s . Responses were recorded via a headset under the control of DMDX (Foster \& Foster, 2003). Four practice trials preceded data collection. The entire testing session lasted about 3 minutes. An identical measurement was made on Day 5, following practice.


Figure 1: Response times in milliseconds (left) and proportions of incorrect responses (right) for each sample (adults and children) in each condition (neutral and incongruent) and time point (pre- and post-practice). Each row of panels displays data for one experimental group (top to bottom: D , no practice; A , incongruent; B , color naming; C , reading). Each colored box contains the middle two quartiles of individual participant logarithmic means; the thick horizontal line marks the median; error bars extend to the full range.

Practice. Participants were assigned randomly into one of four conditions to practice for three consecutive days. Group A practiced color naming of incongruent stimuli (e.g. red), Group B color naming of neutral stimuli (e.g. XXXXXXX), and Group C practiced word reading of neutral stimuli (e.g. "red"). Children were required to complete one block of 144 trials per day. Adults completed one block of 192 trials per day. A fourth group (D) of control participants included no practice, to serve as a reference baseline.

\section*{Results}

Responses were examined with CheckVocal (Protopapas, 2007) to determine accuracy and placement of the timing marks. Response times were subsequently logarithmically transformed to bring their distribution closer to normal.

To examine the differential effects of practice on interference between groups we tested a triple interaction between group, time, and condition using function lmer of
the lme4 package (Bates, Maechler, \& Bolker, 2012) in R ( R Core Team, 2012). Each training group was thus compared for effects of practice against the no-practice group. We employed linear mixed-effects models with maximal random structures (Barr, Levy, Scheepers, \& Tily, 2013), that is, including random intercepts for participants as well as random slopes for time, condition, and their interaction. For response time, in R notation the model formula was specified as:
```

logRT~time*cond*group+(time*cond|subject)

```

Here, "time" refers to before vs. after practice and "cond" refers to neutral vs. incongruent (condition coding). Models were estimated with full maximum likelihood. The significance of the critical triple interaction was then tested via likelihood ratio test against a simpler model excluding the interaction from the fixed effects.

For children, as expected, the group practicing color naming in the incongruent condition significantly reduced
their time interference relative to the no-practice group ( \(\hat{\beta}=\) \(-.14, t=-3.83 ; \chi^{2}=12.84, d f=1, p<.0005\) ). The group practicing color naming in the neutral condition did not differ from the no-practice group ( \(\hat{\beta}=-.03, t=-.04 ; \chi^{2}=\) \(.49, d f=1, p=.482\) ). In contrast, the group practicing color word reading significantly reduced their time interference ( \(\hat{\beta}=-.07, t=-2.05 ; \chi^{2}=4.05, d f=1, p=.044\) ).

For adults, the group practicing color naming in the incongruent condition significantly reduced their interference relative to the no-practice group ( \(\hat{\beta}=-.13, t=\) -3.44; \(\chi^{2}=10.56, d f=1, p=.001\) ). However, there was no significant change in interference either for the group practicing color naming in the neutral condition ( \(\hat{\beta}=-.03\), \(t\) \(=-.83 ; \chi^{2}=.69, d f=1, p=.408\) ) or for the group practicing word reading \(\left(\hat{\beta}=-.02, t=-.45 ; \chi^{2}=.20, d f=1, p=.652\right.\) ).

To analyze error rates, we employed generalized linear mixed-effects models with binomial responses modeled via a logit link. Again, a maximal random structure was used.

For children, the group practicing color naming in the incongruent condition significantly reduced their accuracy interference relative to the no-practice group ( \(\hat{\beta}=-1.89, z=\) \(-2.34, p=.019\) ). The group practicing color naming in the neutral condition did not differ from the no-practice group ( \(\hat{\beta}=-.43, z=-.68, p=.494\) ), in contrast the group practicing color word reading significantly reduced their accuracy interference ( \(\hat{\beta}=1.46, z=2.04, p=.041\) ).

For adults, there were no significant effects in the accuracy analyses (all \(p>.5\) ).

\section*{Discussion}

The results of the present study show that just three days of practice in word reading suffice to produce a reduction in Stroop interference in children. In contrast, interference did not diminish with practice in color naming of neutral stimuli in either population. Even though color naming is considered to be insufficiently automatized (indeed this is the usual explanation for the interference) and therefore presumably amenable to improvement through practice, this dimension of performance did not seem to have much effect on interference. In contrast, we found that interference can be reduced in children through practice in word reading of color names, a highly counterintuitive finding from the point of view of automaticity but a predicted outcome on the basis of the blocking hypothesis. It is strengthened by the fact that it was not replicated in adults, whose reading is skilled and presumably not amenable to improvement.

Our findings cannot be explained by the automaticity account. If Stroop interference derives from the imbalance between reading and color naming due to extensive practice in reading, then we should have observed the opposite pattern of results. Specifically, practice in color naming of neutral stimuli should strengthen the unpracticed dimension and thereby decrease interference, while practice in word reading should have made reading even more dominant and further increase interference. Both of these predictions were contradicted by the results.

Our main finding that word reading practice led to a reduction of Stroop interference is consistent with a negative relationship between reading ability and Stroop interference, as proposed by Protopapas et al. (2007). This relationship is due to the dependence of interference primarily on the speed of processing of the irrelevant stimulus, that is, of reading the word. The time needed to retrieve the irrelevant word response is a crucial factor in the process of interference because it sets a lower limit on the time taken to reject (inhibit) this response. Therefore the speed of reading puts severe constraints on interference. In contrast, neutral color naming may be related to incongruent color naming, because they are both color naming tasks (cf. di Filippo \& Zoccolotti, 2012), but it is not specifically predictive of interference because there is ample time for the color naming response to build up while the inappropriate one (word reading) is retrieved and subsequently inhibited. This suggestion links word reading and neutral color naming, the two component tasks involved in Stroop interference, with the phenomenon of interference directly, via a blocking mechanism as proposed by Roelofs (2003).

Specifically, according to the computational model WEAVER++ (Roefols, 2003) a blocking mechanism prevents obligatory responses from being produced while allowing their processing speed to determine the nondominant response latency. In this theoretical context, interference does not occur because of the relative processing strength of reading and color naming but because a fundamental architectural distinction forces color naming responses to wait until word reading responses are activated and then suppressed. This approach can explain our results in children, assuming that word lemma activation can benefit through reading practice by strengthening the direct connections between visual word stimuli and phonological word forms. This leads to a faster suppression of the word and ultimately to a faster color naming response.

Interference was also reduced in both groups with practice in color naming of incongruent stimuli. Presumably, in this condition participants learned to apply cognitive control more effectively and suppress the irrelevant response faster.

As our data were not constrained by any participant selection criteria and the participants were randomly assigned to the experimental conditions, there is no factor to attribute the differential development of interference other than the experimental manipulation, namely practice. Moreover, our implementation of an individual-item computerized version of the Stroop task, instead of the oftemployed sheet form, alleviates concerns related to task demands and spatial context that might impose consideration of attentional allocation factors (Lachter, Forster, \& Ruthru, 2004; Risko, Stolz, \& Besner, 2005). Without diminishing the significance of executive functions and attentional processes in interference, our data are consistent with the idea of a direct link between reading ability and Stroop interference.

One objection that may be raised against our interpretation is that three days of practice in color naming
of neutral stimuli may not have been enough to counterbalance years of experience in word reading in children, and especially in adults, so that a reduction of interference can be observed. However, if we take into account the effect of word reading practice in children, we see that three days of practice were enough to reduce interference despite years of previous experience with words. Thus it seems that a few hundred trials may be sufficient for these kinds of effects to emerge. Certainly, it cannot be precluded that additional factors modulate color and word processing and act differentially in the naming and reading tasks, but this is an issue beyond the scope of our current analysis, to be addressed in future study. However, from the point of view of a graded automaticity account, it still cannot be explained why practice in word reading did not increase Stroop interference but, instead, decreased it.

Close examination of the raw response times for children (Figure 1, left) reveals that post-training performance tended to be slower than pre-training performance, a trend visible also in the no-training group. This is an additional reason to use a control group rather than directly comparing pre- to post-training times. As it turned out, significant reduction of interference was associated with a lack of increase in posttraining incongruent color naming times. So, in comparison with the control group, this amounts to faster responses.

To the best of our knowledge, this is the first training study of the simple skills involved in Stroop interference, namely reading and color naming, carried out with children. Therefore there is no closely related literature to compare our findings to. Because of the theoretical importance of these results and the implications for cognitive theories of conflict resolution processing, further research will be required to confirm and extend these findings to additional populations and tasks.

\section*{References}

Alwitt, L. F. (1966). Attention in a visual task among nonreaders and readers. Perceptual and Motor Skills, 23, 361-362.
Barr, D. J., Levy, R., Scheepers, C., \& Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language, 68, 255-278.
Bates, D., Maechler, M., \& Bolker, B. (2012). lme4: Linear mixed-effects models using S4 classes. R package version 0.999999-0. http://CRAN.R-project.org/package=lme4.

Cohen, J. D., Dunbar, K., \& McClelland, J. L. (1990). On the control of automatic processes: a parallel distributed processing account of the Stroop effect. Psychological Review, 97, 332-361.
Davidson, D. J., Zacks, R. T., \& Williams, C. C. (2003). Stroop interference, practice, and aging. Aging, Neyropsychology, and Cognition, 10, 85-98.
Dulaney, C. L. \& Rogers, W. A. (1994). Mechanisms underlying reduction in Stroop interference with practice for young and old adults. Journal of Experimental Psychology: Learning, Memory, Cognition, 20, 470-484.

Everatt, J., Warner, J., \& Miles, T. R. (1997). The incidence of Stroop interference in dyslexia. Dyslexia, 3, 222-228.
Faccioli, C., Peru, A., Rubini, E., \& Tassinari, G. (2008). Poor readers but compelled to read: Stroop effects in developmental Dyslexia. Child Neuropsychology, 14, 277-283.
Helland, T., \& Asbjørnsen, A. (2000). Executive functions in dyslexia. Child Neyropsychology, 6, 37-48.
Kapoula, Z., Lê, T.-T., Bonnet, A., Bourtoire, P., Demule, E., Fauvel, C., Quilicci, C., \& Yang, Q. (2010). Poor performance in 15-year old dyslexic teenagers. Experimental Brain Research, 203, 419-425.
Kelly, M. S., Best, C. T., \& Kirk, U. (1989). Cognitive processing deficits in reading disabilities: A prefrontal cortical hypothesis. Brain and Cognition, 11, 275-293.
Kuhn, M. R., \& Stahl, S. A. (2003). Fluency: A review of developmental and remedial practices. Journal of Educational Psychology, 95, 3-21.
Logan, G. D. (1997). Automaticity and reading: Perspectives from the instance theory of automatization. Reading and Writing Quarterly, 13, 123-146.
MacLeod, C. M., \& Dunbar, K. (1988). Training and Stroop-like interference: Evidence for a continuum of automaticity. Journal of Experimental Psychology: Learning, Memory, and Cognition, 14, 126-135.
MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. Psychological Bulletin, 109, 163-203.
MacLeod, C. M. (1998). Training on integrated versus separated Stroop tasks: The progression of interference and facilitation. Memory \& Cognition, 26, 201-211.
Protopapas, A. (2007). CheckVocal: A program to facilitate checking the accuracy and response time of vocal responses from DMDX. Behavior Research Methods, 39, 859-862.
Protopapas, A., Archonti, A., \& Skaloumbakas, C. (2007). Reading ability is negatively related to Stroop interference. Cognitive Psychology, 54, 251-282.
R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.
Roelofs, A. (2003). Goal-referenced selection of verbal action: Modeling attentional control in the Stroop task. Psychological Review, 110, 88-125.
Samuels, S. J. (1999). Developing reading fluency in learning-disabled students. In R. J. Sternberg \& L. SpearSwerling (Eds.), Perspectives on learning disabilities: Biological, cognitive, contextual. Boulder, CO: Westview.
Stroop, J. R. (1935). Studies of interference in serial verbal reactions. Journal of Experimental Psychology, 18, 643662.
van der Schoot, M., Licht, R., Jorsley, T. M., \& Sergeant, J. A. (2000). Inhibitory deficits in reading disability depend on subtype: Guessers but not spellers. Child Neuropsychology, 6, 297-312.
Wolf, M., \& Katzir-Cohen, T. (2001). Reading fluency and intervention. Scientific Studies of Reading, 5, 211-239.

\title{
Context-freeness Revisited
}

\author{
Willem Zuidema (zuidema@uva.nl) \\ Institute for Logic, Language and Computation, University of Amsterdam \\ PO Box 94242, 1090GE Amsterdam, the Netherlands
}

\begin{abstract}
A series of papers have appeared investigating the ability of various species to learn context-free languages. From a computational point of view, the experiments in this tradition suffer from a number of problems concerning the stimuli used in the training phase of the experiments, the controls presented in the test phase of the experiments, and the motivation for and the conclusions drawn from the experiments. This paper discusses in some detail the problems with the existing work in this domain before presenting a new design for this type of experiments that avoids the problems identified in existing studies. Finally, the paper presents results from a small study demonstrating the benefits of the new design.
\end{abstract}

Keywords: Context-free languages, formal language theory, artificial language learning, animal experiments

\section*{Introduction}

Since the publication of (Fitch \& Hauser, 2004), a small but highly visible literature has emerged investigating the ability of various species to learn and process a context-free language (e.g., Friederici, 2004; Perruchet \& Rey, 2004; Gentner, Fenn, Margoliash, \& Nusbaum, 2006; Hochmann, Azadpour, \& Mehler, 2008; van Heijningen, de Visser, Zuidema, \& ten Cate, 2009; Abe \& Watanabe, 2011; ten Cate \& Okanoya, 2012). It is not difficult to see why the questions addressed in this literature appeal to a wide audience: the grammars generating context-free languages are context-free by virtue of their ability to generate hierarchical structures and to implement center-embedding. Hierarchy and centerembedding are, since (Chomsky, 1957), widely recognized to be hallmark features of human language. Hence, experimentally establishing whether non-human animals can handle a context-free language \({ }^{1}\) seems to address a prime candidate in the search for uniquely human, and perhaps uniquely linguistic, cognitive skills.

However, on a closer look, there are many problems with this literature, and almost a decade of investigation and debates have not brought the clarity about this issue that we might have hoped for. In this paper I will first discuss in some detail the problems with the existing work in this domain before presenting a new design for this type of experiments. I will present results of a small experiment with this design, that show it is workable. For lack of space, I will not review elementary formal language theory here; see (O'Donnell, Hauser, \& Fitch, 2005) for an introductory and (Jäger \& Rogers, 2012) for a more advanced discussion of the formal background of the experiments discussed here.

\footnotetext{
\({ }^{1}\) Throughout this paper I will use the phrase "a context-free language" as denoting a member of the subset of the context-free languages that is not also in the set of regular languages.
}

\section*{Problems with the experimental record}

From a computational point of view, the experimental record suffers from a number of problems concerning the stimuli used in the training phase of the experiments, the controls presented in the test phase of the experiments, and the motivation for and the conclusions drawn from the experiments.

The first major problem is a lack of clarity about which ability is really investigated: the ability to implement a context-free language, the ability to learn a context-free language, or a preference for selecting a context-free strategy from the set of strategies adequate for solving the task. Much of the rhetoric seems to be about the ability to implement, but all existing experiments that I am aware of really at best address the weaker hypothesis that non-human animals lack the human preference.

This problem is exacerbated as all existing studies allow a great deal of ambiguity in the training phase about which strategies are adequate. Some ambiguity is unavoidable: all real-world experiments can only present a subset of the infinite stringsets that make up context-free languages, leaving the learner fundamentally uncertain about whether or not sub- or supersets of the intended context-free language are the target (see figure 4). Moreover, studies using familiarization/habituation paradigms can only present positive stimuli in the training phase. However, in a reinforcement paradigm some ambiguity is avoidable, but existing studies using such a paradigm fail to provide learners with the information that some plausible alternatives are not intended. For instance, (Gentner et al., 2006) presented their starlings with in the order of 300000 stimuli with positive and negative feedback to learn to distinguish \(A^{n} B^{n}\) from \((A B)^{n}\), but not with a single stimulus that would help the birds exclude \(A^{n} B^{m}\). If the question we want to ask is whether these birds can learn the context-free language at all, it would be better to avoid unnecessary ambiguity about the task (desideratum 1).

A second unclarity in existing work comes from unnecessary variation in the syllables of (song) elements used to compose the stimuli. Thus, when testing whether subjects can learn \(A^{n} B^{n}\), all studies I am aware off use multiple instances of \(A\) 's and \(B\) 's. This means the subjects are really confronted with two tasks at the same time: the task to categorize \(a_{1}, a_{2}, \ldots\) as instances of class \(A\), and the task to learn sequencing rules. While the interaction between categorization and sequence learning is certainly interesting, this interaction has in fact not been explicitly addressed in this paradigm. In most studies the categorization task is made rather trivial because \(A\) 's and \(B\) 's are carefully selected to be acoustically very similar within one category and very dissimilar between categories. In these studies the variation in stimuli probably has
little impact on the results and just makes describing the experiments unnecessarily complicated; in other cases, it introduces confounds. It would be better, therefore, to avoid these complications and start with experiments using an alphabet with just 2 items: \(a\) and \(b\) (desideratum 2).

A third major problem with the existing literature concerns inadequate controls (see also Beckers, Bolhuis, Okanoya, \& Berwick, 2012). (Fitch \& Hauser, 2004) present no data on controls for alternative strategies (although the supplementary material states - without presenting details - that various alternative explanations have been controlled for). Unlike many other experiments, (Gentner et al., 2006) did test a number of these alternative strategies, and presented results that seemed to exclude all except for the most "heavily contrived" finite-state grammar hypotheses. It turns out that even their quite elaborate efforts to control for various alternative strategies are insufficient, as I will discuss below. It would seem necessary, therefore, to work out better ways to evaluate the plausibility of alternative explanations for the results (desideratum 3).

\section*{A case-study: Gentner et al. 2006}

To make these problems concrete, I will here discuss them in the context of (Gentner et al., 2006). This is not because this paper has more methodological problems than others; on the contrary, in fact, this paper probably represents one of the most serious efforts to control for alternative explanations among the experimental papers in this domain. As I will show below, however, the results from this study have nevertheless little to say about the ability or inability of song birds to learn a context-free language.

Training Gentner et al. studied whether starlings (Sturnus vulgaris) are able to learn a context-free language. As in many other studies, the stimuli in this experiment were strings of elements that fall into two easily distinguishable categories, \(A\) and \(B\), each with a small number of members, i.e. \(A=\left\{a_{1}, a_{2}, a_{3}, \ldots, a_{8}\right\}\) and \(B=\left\{b_{1}, b_{2}, b_{3}, \ldots, b_{8}\right\}\). Gentner et al. extracted these stimuli from the starling's own song, where the \(A\) 's were "rattle" motifs and the \(B\) 's were "warble" motifs. Stimuli in the training phase consisted of strings of length 4 from two patterns \({ }^{2}\) : (i) \((A B)^{n}\) and (ii) \(A^{n} B^{n}\). I will refer to string sets defined by these patterns as the FINITE-STATE-0 and the CONTEXT-FREE language (the 0 indicating that this is just the first of many finite-state languages that I will consider).

The birds were trained in a go-nogo operant conditioning procedure to respond selectively to stimuli from one or the other pattern. In the experiment, birds did indeed learn to distinguish the stimuli sets, at levels far exceeding chance, also when new \(A\) - and \(B\)-category elements were used. This in

\footnotetext{
\({ }^{2}\) I use a conventional shorthand notation for sets of strings of a given pattern, where \(A\) 's and \(B\) 's indicate any elements from these classes, \(X^{n}\) indicates \(n\) repetitions of \(X\), and brackets are used to disambiguate the scope.
}
itself is not enough to prove context-freeness, as Gentner et al note. For instance, the two groups of birds could have internalized +FINITE-STATE-0 and -FINITE-STATE-0 instead \({ }^{3}\). I.e., they could do with a model for the finite-state stimuli set, and only accept/reject stimuli that do not/do conform to it. Or, because the string length is set to \(4, A^{n} B^{n}\) is indistinguishable from \(A^{2} B^{2}\) (which, again, doesn't need context-free power to be recognized). Worse even: there are many other alternative strategies to distinguish the training stimuli-sets, the simplest of which are based on detecting specific element-to-element transitions, or memorizing the beginning or end of strings. For instance, the \(B A\) transition, and the \(A B\) beginning, are diagnostic, because they both only occur in the +FINITE-state-0 set.

If one could show, in the test phase, that the birds have learned a context-free language, this ambiguity in the training phase is not a problem. However, if the birds turn out to choose one of the simpler strategies that also suffice to distinguish the two classes, we are left almost empty-handed. We cannot make plausible, then, that birds cannot learn contextfree language, because we haven't tried very hard to force them to.

Testing In the test phase, Gentner et al did consider a relevant set of alternative strategies.

The first test is whether subjects generalize from \(A^{2} B^{2}\) to a larger subset of \(A^{n} B^{n}\). Of course, in formal language theory the language \(A^{n} B^{n}\) contains an infinite number of strings, where \(n\) can be any integer. Gentner et al. argue, quite reasonably, that in an experimental setting we should be concerned about whether subjects generalize to unseen \(n\). (This is completely analogous to the use of formal language theory in the study of natural language: if we can demonstrate the right generalization mechanisms on necessarily finite data, we can reason about an infinite competence under a hypothetical lifting of performance constraints.) Gentner et al. report, for birds trained with \(A^{2} B^{2}\), a strong preference for \(A^{3} B^{3}\) and \(A^{4} B^{4}\) strings over \((A B)^{3}\) and \((A B)^{4}\) respectively (and an inverse preference for birds trained on \(\left.(A B)^{2}\right)\). This rules out the -FINITE-STATE-0 (or +FINITE-STATE-0) strategy.

There remain, however, still many alternative hypotheses that predict successful discrimination of \(A^{n} B^{n}\) and \((A B)^{n}\) strings. It is useful to define the following simple, but effective strategies for positively responding to the +CONTEXTFREE stimuli \({ }^{4}\) :
+ AnBn: \(A^{n} B^{n}\), with \(n \geq 1\)
+ AnBm: \(A^{+} B^{+}\), the set of strings that consist of 1 or more \(A\) 's followed by 1 or more \(B\) 's;
-BIGRAM-BA: . \({ }^{*} B A \cdot{ }^{*}\), strings containing transition \(B A\);

\footnotetext{
\({ }^{3}\) A strategy is defined by a PATTERN, written in smallcaps, and a + or \(\mathrm{a}-\) in front of it; the + indicates that strings that conform to the pattern are treated as positive stimuli; the - indicates that strings that conform to the pattern are treated a negative stimuli.
\({ }^{4}\) I will use more or less standard regular expression notation, where a dot \(\cdot\) means any symbol, \({ }^{*}\) means repeated any number \((\geq 0)\) of times, and \({ }^{+}\)means repeated any number \((\geq 1)\) of times.
}
\begin{tabular}{|c|c|c|c|c|c|c|c|c|c|c|c|c|c|c|}
\hline & A2B2 & A3B3 & A4B4 & AB2 & AB3 & AB4 & A1B3 & A3B1 & A2B3 & A3B2 & A4 & B4 & ABBA & BAAB \\
\hline +ANBN & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline +ABN & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline +ANBM & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\
\hline +AA-PRIMACY & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\
\hline +BB-RECENCY & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\
\hline +AB-RECENCY & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\
\hline +AA-BIGRAM & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 \\
\hline +BA-BIGRAM & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\
\hline +BB-BIGRAM & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\
\hline \(+\cdot\{1,4\}\) & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 \\
\hline \(+\cdot\{1,6\}\) & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
\hline
\end{tabular}

Table 1: Predicted response of various hypothesized pure strategies to probe stimuli. See the main text for descriptions of the top 9 strategies; the bottom two strategies check the length of a string, and accept strings up to length 4 and 6 respectively.
+BIGRAM-AA: \(\cdot{ }^{*} A A \cdot^{*}\), strings containing transition \(A A\);
+BIGRAM-BB: . \({ }^{*} B B \cdot^{*}\), strings containing transition \(B B\);
-PRIMACY-AB: \(A B \cdot^{*}\), the set of strings that start with \(A B\);
+PRIMACY-AA: \(A A \cdot^{*}\), the set of strings that start with \(A A\);
-RECENCY-AB: \(\cdot{ }^{*} A B\), the set of strings that end with \(A B\);
+RECENCY-BB: . \({ }^{*} B B\), the set of strings that end with \(B B\);
Any of these strategies (together with their complements when considering the birds that should NOT respond to \(A^{n} B^{n}\) ) suffices to distinguish positive from negative samples in the experimental set-up (and all listed alternative strategies are in the finite-state class). But these nine hypotheses do make different predictions on the behavior of the subjects for previously unseen patterns. For instance, the +PRIMACY-AA strategy classifies all of the +CONTEXT-FREE stimuli as positive, but in addition, for string length 4 , also includes AAAA, AAAB, AABA. Table 1 gives for (the + variety of) each of these strategies the predicted response ( 1 is a GO-response, 0 is a NOGO-response).

There are, in fact, still many other alternative strategies that we could consider, such as memorizing non-adjacent pairs (e.g., \(A \cdot B \cdot^{*}\) ), or requiring a specific number of a particular transition (e.g., requiring two \(A B\) transitions, as in \(B^{*} A^{+} B^{+} A^{+} B^{+}\), or exactly one, as in \(A^{+} B^{+} A^{*}\) ). Gentner et al. appeal, quite reasonably again, to considerations of parsimony to ignore such alternatives.

To rule out the 9 remaining alternative explanations, Gentner et al. presented birds with a number of diagnostic strings. For instance:
- AAAB, which is incorrectly predicted to give a positive response by -FINITE-STATE-0, -BIGRAM-BA, +BIGRAMAA, and +PRIMACY-AA;
- BBBB, which is incorrectly predicted to give a positive response by -FINITE-STATE-0, -BIGRAM-BA, +BIGRAMBB, -PRIMACY-AB, -RECENCY-AB and +RECENCY-BB.

In an experimental setup, however, checking these predictions needs to be buffered to unavoidable noise in the data. (It would be unreasonable to reject an hypothesis based on a single unexpected classification by a bird.) Hence, we need
to use statistics, but how statistical methods for data analysis are combined with formal language theory is a non-trivial issue that both theoreticians and experimentalists have so far largely ignored. (Even the review by (Jäger \& Rogers, 2012), which presents a major effort to bridge formal language theory and artificial language learning experiments, ignores this issue).

Gentner et al. chose to use the \(d^{\prime}\)-statistic, which is a measure for discrimination between stimuli classes that corrects for response bias (the tendency to prefer a GO or a NOGOresponse regardless of the stimulus). They show that the \(d\) ' between AAAA and ABBA is significantly lower than the \(d\) ' between A2B2 and AB2, and argue this rules out the +AAPRIMACY strategy. Similarly, they find lower \(d\) ' for BBBB vs BAAB , and for \(\mathrm{BAAB} / \mathrm{ABBA}\) vs. \(\mathrm{AAAA} / \mathrm{BBBB}\), and argue this rules out + BB-RECENCY and -BA-BIGRAM. Similar analyses can be given for the remaining alternative strategies.

However, it turns out that this approach for ruling out alternatives is only valid if the population is homogeneous all members follow the same strategy - and if each individual follows a pure strategy. If individuals or populations can mix multiple strategies, Gentner et al.'s method leads to invalid conclusions. The data of Gentner et al., reproduced as the blue bars in figure 2, clearly show that the assumption of pure strategies is false: the \(d^{\prime}\)-statistic for the \(A^{n} B^{n}\) vs. \(A B^{n}\) constrast decreases with increasing \(n\), and the \(d^{\prime}\) 's for the primacy and recency strategies differ significantly. A study on zebra finches, by (van Heijningen et al., 2009), reported major individual differences between birds, further strengthening the case against the pure strategy assumption.

Simulated Data To show how the \(d\) '-statistic can lead to wrong conclusions, I will now present some artificial data that shows a qualitatively similar pattern of \(d^{\prime}\)-scores for both a model that involves an underlying context-free grammar (model I: CFG) and a model that is just a mix of finite-state strategies (model II: MIX).

To generate the CFG data, I assume a population where \(70 \%\) of the individuals have internalized the +ANBNstrategy, \(10 \%\) follow a strategy to reject long strings


Figure 1: Simulated data, generated from the CFG-model (left) and the MIX-model (right). In both graphs, the first group of 3 bars represent the number of go-responses to \(A^{n} B^{n}\)-stimuli (out of 100); the second to \((A B)^{n}\)-stimuli; the third group to \(A^{n} B^{m \neq n}\)-stimuli; the fourth to the remaining control-stimuli.


Figure 2: The \(d^{\prime}\)-statistic calculated for the \(A^{n} B^{n}\) vs. \((A B)^{n}\) distinction (left) and for various controls (right). Blue: Gentner et al, Red: CFG, Yellow: MIX.
\((+.\{1,6\}), 10 \%\) reject medium and long strings (+. \(\{1,4\}\) ), and \(10 \%\) randomly choose GO or NOGO. This is equivalent to assuming a mixed strategy with the same proportion, or a combination of intra- and interindividual variation.

To generate the MIX data, I assume a mix of finite-state strategies in the following proportions (roughly based on the findings of (van Heijningen et al., 2009)): \(25 \%\) +AAPRIMACY, \(15 \%\) +BB-RECENCY, and \(10 \%\) for each of the other strategies in table \(1+\) RANDOM.

A given mix of strategies defines for each stimulus a specific number \(f\) of GO-responses, from a fixed number of 100 presentations. To generate some randomness, I assume each of the \(f\) stimuli that should be classified as a GO-response has a fixed probability \(p=0.03\) to receive a NOGO-response, and similarly, that each of the \(100-f\) remaining stimuli have a probability \(p\) of receiving a GO-response.

Hence, the final number of GO-responses is \(g=x+y\), where both \(x\) and \(y\) are sampled from a binomial distribution:
\[
x \sim B(f, p), y \sim B(100-f, 1-p)
\]

This generates a dataset as in figure 1 (left) for the CFG model, and (right) for the MIX model (note that the datasets
are the result of a single run of the model).
I subsequently calculate the \(d^{\prime}\)-statistic in the same way as (Gentner et al., 2006). This statistic is simply the difference between the z-transform of the counts: \(d^{\prime}(x, y)=z(x)-z(y)\) where the z -transform in turn is a way to express the magnitude of the score in terms of how many standard deviations it is away from the mean: \(z(x)=(\bar{d}-x) / \sigma_{d}\) where \(d\) is the complete data vector from which \(x\) is one value (or the average of several values), and \(\sigma_{d}\) is the standard deviation over that vector.

Applying these formulas to the contrasts between \(A^{n} B^{n}\) and \((A B)^{n}\) for \(n \in\{2,3,4\}\) we obtain, in figure 2(left) qualitatively similar results to Gentner et al: significant discrimination for all, but a decrease in discriminability with increasing \(n\). Although unsurprising, this result points to a problem with existing studies that fail to show generalization: this could be due to length effects. It would be better if tests for generalization do not only test on longer strings in the test phase than were offered in the training phase (desideratum 4).

Applying these formulas to the contrasts between AAAA and ABBA (labeled "primacy" in (Gentner et al., 2006)), BBBB vs BAAB (labeled "recency"), and BAAB/ABBA vs. AAAA/BBBB (labeled "bigram B/A") we observe, in figure 2(right) the exact same pattern of results as Gentner et al reported: overall much lower \(d^{\prime}\)-values than for the baseline, with primacy receiving the second highest score and recency the lowest.

Hence, for both the context-free (CFG) and mix of finitestate strategies (MIX) we see the same pattern of \(d^{\prime}\)-values, showing that when mixed strategies are possible, these values are uninformative about whether or not a context-free language is learned by any individual in the population. Hence, we need better tools to assess which strategies individuals are using and whether there is significant individual variation in a population (desideratum 5).

\section*{A new design}

The problems I discussed with the Gentner et al. study are symptomatic for many studies in this domain. Confusion
about exact goals and methodology are of course typical for the early phase of a new research field. It is now time, however, for an experimental design to emerge that is both methodologically sound and capable of generating useful results. In the following I will present an attempt to give such a design and a first experiment to assess its usefulness. The design follows desiderata 1-4 discussed above:
1. The goal of the design is to test whether or not the subjects can learn the context-free language \(A^{n} B^{n}\) from the type of data that can be used with animals as well as human infants and adults; i.e., not too long strings, possibly with positive and negative feedback. Some strings are reserved for the test phase only, to assess generalization, but otherwise any training regime is allowed within these constraints. In practice, I choose for a two-stage training phase: a familiarization phase where only positive stimuli are presented, and a feedback phase where positive and negative stimuli are presented with positive and negative feedback. The negative stimuli are not just from \((A B)^{n}\), but also from other plausible, but incorrect, alternative languages.
2. To make the task as simple and unambiguous as possible, I define the patterns over an alphabet of just two different sounds: \(a\) and \(b\), selected to be short and acoustically clearly distinct.
3. To be able to test for generalization, I reserve 2 values of \(n\) for strings from \(A^{n} B^{n}\) and \(A^{n} B^{n}\) for the test phase only. I further reserve a number of \(n, m\) combinations for strings in \(A^{n} B^{m \neq n}\) for the test phase, to be able to exclude primacy, recency and ANBM-strategies.
4. To make sure the test strings are not much longer than the strings seen at training, I use \(n \in\{2,3,5,6\}\) for \(A^{n} B^{n}\) and \((A B)^{n}\) strings at training, and \(n \in\{3,4,6,7\}\) at test.

The stimuli presented to subjects in the various phases are thus as follows:
\begin{tabular}{l|l} 
Phase & Stimuli \\
\hline Familiarization & \(\mathrm{a} 2 \mathrm{~b} 2, \mathrm{a} 3 \mathrm{~b} 3, \mathrm{a} 5 \mathrm{~b} 5, \mathrm{a} 6 \mathrm{~b} 6\) \\
\hline Feedback & \begin{tabular}{l} 
Positive: \(\mathrm{a} 2 \mathrm{~b} 2, \mathrm{a} 3 \mathrm{~b} 3, \mathrm{a} 5 \mathrm{~b} 5, \mathrm{a} 6 \mathrm{~b} 6\) \\
Negative: \(\mathrm{ab} 2, \mathrm{ab} 3, \mathrm{ab5}, \mathrm{ab} 6, \mathrm{a} 3 \mathrm{~b} 2, \mathrm{a} 5 \mathrm{~b} 4\)
\end{tabular} \\
\hline Test & \begin{tabular}{l} 
Positive: \(\mathrm{a} 3 \mathrm{~b} 3, \mathrm{a} 4 \mathrm{~b} 4, \mathrm{a} 6 \mathrm{~b} 6, \mathrm{a} 7 \mathrm{~b} 7\) \\
Negative: \(\mathrm{ab} 3, \mathrm{ab} 4, \mathrm{ab} 7, \mathrm{a} 3 \mathrm{~b} 2, \mathrm{a} 4 \mathrm{~b} 3, \mathrm{a} 2 \mathrm{~b} 3\)
\end{tabular}
\end{tabular}

Experimental data Ultimately we need a lot of data and new analysis tools to meet desideratum 5 and exclude mixed strategies between and within individuals. However, a first important check on the design is to evaluate whether we can replicate the findings (Hochmann et al., 2008) that humans adults can, at the population level, (i) learn to distinguish strings from \(A^{n} B^{n}\) from several finite-state alternatives, and (ii) generalize to unseen \(n\). We therefore carried out a small experiment with the design above, to test whether its new features 1-4 stand in the way of successful learning.

The experiment was implemented as a simple internetbased applet. Subjects were instructed that they were going to do an experiment that looked a bit like a computer game, where they would have learn an alien language. At the computer screen, subjects were presented with written instructions, and, once they started the game, presented with a space background and UFO's moving over the screen. Subjects were asked to click on disks and listen to the sounds produced. After hearing the sounds they would decide to either shoot the UFO or save the aliens inside. In the familiarization phase (4 exposures to each stimulus), they were told all aliens were 'good aliens' and shooting was disabled. In the feedback phase ( 1 exposure to each stimulus) feedback was provided in the form of happy or sad face on the screen. In the test phase ( 2 exposures to each stimulus) no feedback was given.

54 subjects were recruited in the Amsterdam Science Museum (Nemo) in August 2012, and volunteered for the experiment without payment. The experiment last only about 5 minutes per subject. Subject ages ranged from around 10 to around 80; native languages included several major European languages. We thus worked with a very heterogeneous group of subjects and obtained very little data per person. Hence, the experiment was not useful (nor intended) for settling the question of whether human adults can learn a context-free language in such a setup, but to assess whether the experimental design defined above is able to generate useful data.

Figure 3 (left) gives the overall response rates for each of the stimuli presented in the test phase. As can be seen, the response rates for the \(A^{n} B^{n}\) stimuli on the left are higher than for the \(A^{n} B^{n}\) stimuli on the right, which in turn are higher than the \(A B^{n}\) stimuli in the middle. All three pairwise between-group difference are highly significant ( \(p<0.01\), Kolmogorov-Smirnov test). Crucially, responses to a4b4 are indistiguishable from other positive stimuli, indicating subjects have, at the population level, generalized to unseen \(n\).

These data thus, roughly, replicate earlier results. But can we check for all relevant alternative explanations, including a mix of finite-state strategies, as I argued above would be necessary? Unfortunately, with so little data per subject, we cannot sensibly estimate individual strategies. To get some idea about individual variation I split the data in two based on performance during the feedback phase. 20 subjects were classified as low-performers, with an accuracy in the feedback phase of less than \(70 \%\) (similar to the criterion used in van Heijningen et al., 2009). The other 34 subjects were classified as high-performers. Figure 3 (right) shows the \(d^{\prime}\) statistic for the three pairwise contrasts. One striking value is a \(d^{\prime}\) of approximately zero for the low performers on the \(A^{n} B^{n}\) vs. \(A^{n} B^{m}\) contrast, showing that they did not distinguish between the two classes and clearly had not learned a context-free language.

These data thus suggest that the experimental design presented above is very workable and can be used to obtain useful data to address the question of whether subjects can


Figure 3: Experimental results.
learn \(A^{n} B^{n}\), provided it is applied in more controlled circumstances and more data per person is gathered. With enough data per subject, we can apply the model selection approach of (van Heijningen et al., 2009) to estimate the most likely strategy for each individual subject.

\section*{Conclusions}

I have discussed some major problems with existing studies attempting to show that nonhuman animals can or cannot learn a context-free language. In that discussion, I identified four desiderata for a new design of an experiment, and a fifth desideratum for data analysis, to properly address that question. I have shown that these desiderata for the design can be satisfied, and presented some experimental results that suggest there are no major obstacles to apply the new design in animal experiments. I hope the interdisciplinary community that tries to bring formal language theory, artificial language learning and animal cognition experiments together will apply this new design in future experiments, such that the search for uniquely human cognitive skills can be based on a more sound foundation. I have not, in this paper, discussed the difficult question of whether context-freeness is really the most important property to investigate in the search for a biological basis for language (Zuidema, 2013). Even if it is not - and I suspect it isn't - it is essential that the methodological errors in the experimental record on context-freeness get corrected.

Acknowledgments I thank Vanessa Ferdinand, Richard Kunert and Sander Latour for their help in designing the experiment and creating the software for stimulus presentation and data gathering, and the team at Nemo for running the experiment.

\section*{References}

Abe, K., \& Watanabe, D. (2011). Songbirds possess the spontaneous ability to discriminate syntactic rules. Nature neuroscience, 14(8), 1067-1074.
Beckers, G., Bolhuis, J., Okanoya, K., \& Berwick, R. (2012). Birdsong neurolinguistics: songbird context-free grammar claim is premature. Neuroreport, 23(3), 139.

Chomsky, N. (1957). Syntactic structures. The Hague: Mouton.
Fitch, W. T., \& Hauser, M. D. (2004). Computational constraints on syntactic processing in a nonhuman primate. Science, 303(5656), 377-380.
Friederici, A. D. (2004). Processing local transitions versus long-distance syntactic hierarchies. Trends in Cognitive Sciences, 8(6), 245-247.
Gentner, T. Q., Fenn, K. M., Margoliash, D., \& Nusbaum, H. C. (2006). Recursive syntactic pattern learning by songbirds. Nature, 440, 1204-1207.
Hochmann, J.-R., Azadpour, M., \& Mehler, J. (2008). Do humans really learn anbn artificial grammars from exemplars? Cognitive science, 32(6), 1021-1036.
Jäger, G., \& Rogers, J. (2012). Formal language theory: refining the chomsky hierarchy. Philosophical Transactions of the Royal Society B: Biological Sciences, 367(1598), 1956-1970.
O’Donnell, T. J., Hauser, M. D., \& Fitch, W. T. (2005). Using mathematical models of language experimentally. TRENDS in Cognitive Sciences, 9(6).
Perruchet, P., \& Rey, A. (2004). Does the mastery of centerembedded linguistic structures distinguish humans from nonhuman primates? Psychonomic Bulletin and Review. (In press)
ten Cate, C., \& Okanoya, K. (2012). Revisiting the syntactic abilities of non-human animals: natural vocalizations and artificial grammar learning. Philosophical Transactions of the Royal Society B: Biological Sciences, 367(1598), 1984-1994.
van Heijningen, C., de Visser, J., Zuidema, W., \& ten Cate, C. (2009). Simple rules can explain discrimination of putative recursive syntactic structures by a songbird species. Proceedings of the National Academy of Sciences, 106(48), 20538-20543.
Zuidema, W. (2013). Language in nature: on the evolutionary roots of a cultural phenomenon. In P. Binder \& K. Smith (Eds.), The language phenomenon (p. 163-190). Berlin: Springer.

\title{
The Effects of Overt Head Movements on Valenced Image Recognition and Recall in Bulgarians*
}

\author{
Meryl Varadinov (meryl.varadinov@gmail.com) \\ Department of Cognitive Science and Psychology, New Bulgarian University, 21 Montevideo Street Sofia 1618, Bulgaria
}

Lyuben D. Laskin (lyubenlaskin@gmail.com)
Department of Cognitive Science and Psychology, New Bulgarian University, 21 Montevideo Street Sofia 1618, Bulgaria

\begin{abstract}
Vertical and horizontal head movements (universally associated with nodding and shaking, respectively) have frequently been demonstrated to affect cognitive processes. Two experiments were conducted to test the hypothesis that overt head movements can influence memory for valenced images. In the first experiment, participants were instructed to perform either vertical or horizontal head movements while viewing a slideshow of 76 randomized positive and negative images, which they later had to recognize from a set containing \(50 \%\) of the same target images and \(50 \%\) distractor images. No interaction between head movement type and image valence was obtained. In the second experiment, participants were told to remember as many images as possible from a slideshow of 60 randomized valenced images, which they were later asked to freely recall. A significant interaction was obtained, with a higher rate of recall for positive images when vertical head movements (VHM) were performed and a higher rate of recall for negative images when horizontal head movements (HHM) were performed.
\end{abstract}

Keywords: overt head movements; image recall and recognition; embodiment.

\section*{Introduction}

Previous work has demonstrated that inducing overt head movements can influence certain cognitive processes due to their positive or negative association with a given cognitive activity. For example, Wells \& Petty (1980) showed that the manipulation can have an effect on persuasion. Participants in their study performed either vertical or horizontal head movements while listening to a simulated radio broadcast, containing either a message in agreement or disagreement with participants’ attitudes (proattitudinal or counterattitudinal messages). Those participants who performed VHM agreed with the content of the broadcast more than those who performed HHM, regardless of the content of the message. Participants also found it more difficult to perform head movements that were incompatible with the message (VHH during a counterattitudinal broadcast vs. HHM during a proattitudinal broadcast). Consistent results were obtained by Briñol \& Petty (2003).

This manipulation has produced similar effects in various other domains. Tom et. al. (1991) showed that VHM led to an increased preference for neutral objects, while HHM led
to a decline in preference for the same. In another study, it was demonstrated that overt head movements can affect product choice and price perception (Tom et. al., 2006). Eppley \& Gilovich (2002) examined the effects of the same manipulation on the anchoring and adjustment heuristic, demonstrating that participants induced to accept values by nodding adjusted to self-generated anchors less than those induced to deny values by shaking their heads.

Apart from attitudinal effects, overt head movements have also been demonstrated to influence memory. Förster \& Strack (1996) induced overt head movements in participants while they listened to valenced adjectives and found that VHM led to better recognition for positive adjective and HHM led to better recognition for negative adjectives in a surprise recognition task. They refer to this as a motor-compatibility effect, suggesting that the process of learning valenced words was more effective when accompanied by compatible head movements during the encoding phase, which led to better recognition during the test phase. They point to natural and socially-learned cooccurrences of overt and covert responses as a possible explanation of this effect. An example of the former is the co-occurrence of basic emotions with specific facial expressions, while learned nonverbal responses like nodding and shaking our heads to indicate agreement and disagreement, respectively, are an example of a sociallylearned co-occurrence of overt and covert responses.

Following Förster \& Strack's methodology, in this study we explore the influence of overt head movements on the recognition and free recall of valenced images. We performed two experiments in order to investigate whether a similar motor-compatibility effect would be obtained, by inducing either HHM or VHM while participants viewed positive and negative images, which they later had to recognize among distractor images (Experiment 1) or freely recall (Experiment 2).

There were two main motivations behind this study. First and foremost, we wanted to see if the manipulation extends to memory for images. It is well known that people's memory capacity for images is far superior to that for words (Shepard, 1967; for an old but good review of the literature, see Landauer, 1986). Hence, observing the same effect

\footnotetext{
* This paper was accepted for CogSci 2012 in Sapporo but not included in those proceedings by accident.
}
would be a non-trivial finding, further confirming the validity of the effect. Second, Bulgaria is a unique place to test this manipulation, given that nodding and shaking mean the exact opposite to what they mean in the rest of the world - shaking denotes agreement, whereas nodding indicates disagreement. If the mechanism of this effect is social in nature, as Förster \& Strack suggest, we should expect the exact opposite results: facilitation for encoding negative images while performing VHM and facilitation for encoding positive images while performing HHM. Thus, the results would give insight into the extent to which culture mediates this effect.

\section*{Experiment 1}

\section*{Method}

\section*{Participants}

Nineteen New Bulgarian University students (11 men and 8 women) were given course credit for participation in what they were told was a marketing study to test the comfort and quality of a headphone set.

\section*{Stimulus Material and Apparatus}

Seventy six target images (38 positive and 38 negative) were selected from The International Affective Picture System (IAPS; Lang, Bradley, \& Cuthbert, 2008) for the encoding part of the procedure, along with half as many positive and negative distractor images (closely resembling half of the target images thematically) for the recognition part of the procedure. The reason why such distractor images were used for the retrieval procedure is that people have a naturally high memory capacity for images (Brady, 2008; Schacter, Israel, \& Racine, 1999; Shepard, 1967), which can lead to ceiling effects. One of the most impressive demonstrations of this is a study by Standing (1973), who, using a single-trial learning task with a delayed recognition test, tested subjects' memory capacity for groups of pictures ranging from 20 to 10,000 and found that, even though percentage of retention gradually declined, the absolute number of stimuli retained increased as the set of learning material increased. In an attempt to avoid losing the effect of the head movement manipulation because of such a ceiling effect, we increased the difficulty of the recognition task by introducing interference with distractor images thematically related to half of the target images (the presentation of similar distractor images has been shown to reduce accuracy in image recognition; Goldstein \& Chance, 1970). These distractor images were presented in the recognition task instead of the target images that they resembled in order to confused participants and prevent them from getting perfect accuracy scores. For samples of positive and negative target-distractor pairs, see Figure 2.

Images were selected for similar levels of arousal (approx. 5 on a 1-9 scale). The average valences of the negative and positive stimuli were approximately 3 and 7, respectively, also on a 1-9 scale. For the first part of the procedure, images were presented on a 17 " monitor
(1024×760) in Microsoft PowerPoint© and then in EPrime \({ }^{\circledR} 2.0\) during the recognition task.


Figure 1. A visualization of the procedure in Experiment 1. On the left: Encoding phase, consisting of 76 positive (+) and negative (-) target images (top left and bottom left blocks). On the right: Recognition task, in which half (38) of the original (=) targets (top right block) are presented
with 38 distractor images (bottom right block), corresponding ( \(\cong\) ) to 38 of the original targets excluded from the retrieval phase.

\section*{Design and Procedure}

The experimental design was a \(2 \times 2\) mixed-model factorial comparing head movement type (horizontal vs. vertical) between subjects and image valence (positive vs. negative) within subjects. The dependent measure was the accuracy of recognition during the test phase.

Participants were introduced to the experiment by signing a consent form and being told that they were participating in a marketing study to test the comfort and quality of a headphone set (this is the same cover story as the one used by Wells \& Petty, 1980). The entire session consisted of an encoding phase, a distractor task, and a recognition task (see Fig. 1 for a visualization of the procedure). Each of the 76 positive and negative images appeared on screen for 3 seconds as Astor Piazolla's tango "Adios Nonino" played in the background (the same type of music used by Förster \& Strack, 1996). Stimulus order was pseudo-randomized (same order for each participant). The set of images was preceded and followed by a 6 second long blank slide and the entire slideshow lasted for 4 minutes. Ten of the participants were instructed to perform HHM while listening to the music and viewing the images (presumably to test the sound quality of the headphones under more realistic conditions, namely, during movement), while the remaining nine performed VHM. The experimenters would demonstrate the movement and instruct participants to
maintain one head movement per second for the entire duration of the slideshow.

After the encoding procedure was over, participants were asked to fill out a feedback form regarding comfort of the headphones, sound quality, difficulty of the head movements, and likability of the music as part of the cover story. They were then given a distractor task for 15 minutes, during which they had to assemble two different wooden puzzle cubes (similar to the classic Soma puzzle cube).

Once those 15 minutes were up, the experiment proceeded to a surprise recognition task. Participants were presented with only half (38) of the target images from the original slideshow mixed with 38 distractor images. For each image that appeared on the screen, participants had to press an "old" button to indicate that they had seen the image before or a "new" button if they hadn't seen it. The images remained on the screen until both responses were given. Once done, participants were thanked for their participation and dismissed. All were debriefed via e-mail once the study was over.


Figure 2. Sample target and distractor images. Target images (top and bottom left) were presented during the encoding procedure and replaced by their matching distractors (top and bottom right) in the recognition procedure.

\section*{Results and Discussion}

We performed a repeated measures ANOVA with valence as a within-subject and head movement as a betweensubject factor. There was a main effect of valence, \(F(1,17)\) \(=8.253, p<0.05\), with better discrimination for negative images, \(\operatorname{Pr}=0.53\), as opposed to positive images, \(\operatorname{Pr}=\) 0.43 . \({ }^{1}\) There was no main effect of head movement type,

\footnotetext{
\({ }^{1}\) Due to the fact that several participants made no false alarms during the recognition task, our first preference, \(d^{\prime}\), could not be calculated and \(\operatorname{Pr}\) was used instead. \(\operatorname{Pr}\) is a coefficient that describes participants' discrimination performance; the higher its value, the better participants' discrimination between targets and
}
\(F(1,18)=0.59, p=0.812\). Contrary to what was hypothesized, there was no significant interaction between head movement type and image valence: \(F(1,17)=0.001, p\) \(=0.975\). Average accuracy for participants performing HHM was \(\operatorname{Pr}=0.4\) for positive images and \(\operatorname{Pr}=0.5\) for negative ones, while participants performing VHM had an average accuracy score of \(\operatorname{Pr}=0.46\) and \(\operatorname{Pr}=0.56\) for positive and negative images, respectively (see Fig. 3).


Figure 3. The average discriminability of images as a function of head movement type and image valence.

Due to the fact that accuracy scores are generally high on image recognition tasks because participants rely on their implicit memory, we tried to deliberately deteriorate their performance by presenting them with distractor images closely resembling half of target images from the encoding procedure, which explains why they made errors in the first place. However, even though it prevented a ceiling effect, this manipulation could have caused other problems. By increasing the difficulty of the task with thematically similar distractor images, we may have suppressed an existing HMimage valence interaction due to the strong interference between the similar target-distractor pairs. A reason for this suspicion is that some studies have demonstrated constructive memory effects in image recognition (e.g., Foley \& Foy, 2008 and Miller \& Gazzaniga, 1998 demonstrated this using the DRM paradigm). \({ }^{2}\)
We decided to conduct a second experiment, but instead of increasing the memory task's difficulty by introducing interference, we replaced the recognition task with a free recall task, which requires participants to rely on their explicit memory. We expected this manipulation to prevent a ceiling effect without introducing additional strong memory effects, as in Experiment 1.
distractors, and vice versa. It is calculated using the following formula: \(\operatorname{Pr}=\mathrm{H}-\mathrm{FA}\), where H (hits) is the proportion of correct identifications of targets, and FA (false alarms) is the proportion of incorrect identifications of distractors (Snodgrass \& Corwin, 1988).
\({ }^{2}\) For more on the DRM paradigm, see Roediger \& McDermott (1995).

\section*{Experiment 2}

\section*{Method}

\section*{Participants}

Twenty-six volunteers and New Bulgarian University students ( 6 men and 20 women), took part in Experiment 2, the latter of whom were given course credit for participation. The cover story from Experiment 1 was maintained.

\section*{Stimulus Material and Apparatus}

Thirty positive and 30 negative images were selected from IAPS (Lang et. al., 2008). They were selected according to the same criteria used in Experiment 1.

\section*{Design and Procedure}

The experimental design and stimulus presentation were the same as in Experiment 1. The changes made in this experiment's procedure were as follows: Participants were explicitly told to try to remember as many of the images from the slideshow as possible, as they would be asked to recall them later, after filling out the feedback form. Following the learning phase, participants were asked to write down all the images they could remember, listing each one on paper (free recall test phase) and using the minimum number of words to accurately describe them. They were given 20 minutes to describe as many images as they could remember. Upon completion of this task, they were thanked for participating and dismissed.

\section*{Results and Discussion}

Two independent experts were used to evaluate the correspondence between the descriptions of recalled images given by participants and the images used in the slideshow. Only answers which gained unanimous consent by the experts were counted as correctly recalled images. Overall agreement between them was \(98 \%\).

The average number of correctly recalled images was 20.04. Similar to Experiment 1, on average, more negative (10.73) than positive (9.31) images were recalled for a main effect of valence, \(F(1,24)=5.03, p<0.05\). There was no significant main effect of head movement type, \(F(1,24)=\) \(0.207, p=0.653\). Contrary to Experiment 1, the crucial interaction between head movement type and valence was significant, \(F(1,24)=4.5, p<0.05\). That is, more positive images were recalled by participants who performed VHM (9.69), compared to participants who performed HHM (8.92). In contrast, more negative images were recalled by participants who performed HHM (11.7), compared to those performing vertical ones (9.77; see Fig. 4). It is evident that the difference between vertical and horizontal head movements is greater for negative \((d=1.14)\) than for positive ( \(d=0.4\) ) images. It is curious to note that this is the opposite of what Förster \& Strack (1996) observed for valenced adjectives and may be due to some type of negativity bias. Many previous studies have failed to demonstrate memory effects for images as a function of
valence, but according to Ochsner (2000), this may have been a result of using memory measures that lack sensitivity to differences in the experience of past events. Using the remember/know paradigm, the author obtained more remember responses for negative images compared to positive ones, whereas positive images evoked more know responses (i.e. they were just familiar, rather than remembered). Kensinger et. al. (2007) have similarly shown a general recognition advantage for negative stimuli.


Figure 4. The average number of recalled images as a function of head movement type and image valence.

\section*{General Discussion}

Despite the positive results obtained in Experiment 2, the findings of Experiment 1 did not show a significant interaction between head movement type and image valence. We hypothesized that our attempt to prevent a ceiling effect during the recognition task by introducing strong interference between target and distractor images may have concealed an otherwise existing effect of the head movement manipulation. In Experiment 2, we employed a different method for eliminating the possible ceiling effect by changing the recognition task to a free recall task, which proved successful in finding the significant interaction.

A finding consistent across both experiments was that more negative images were remembered than positive images, even though they were specifically selected for equal arousal ratings. One possible explanation for this is that negative events leave stronger memory traces compared to positive events. Studies on memory have tended to focus more on arousal levels of stimuli than on valence, but recent research has shown it to be an important factor in encoding and retrieval processes. Multiple studies show that negative information is remembered more vividly than positive information, that it is remembered in more detail, and that people are better at remembering whether they saw or only imagined negative stimuli, whereas with positive stimuli, they are more likely to be confused (for a review, see Kensinger, 2009). This body of research and the results of our experiments are in line with the general trend observed in various domains in psychology (e.g. impression
formation, learning, judgments, information processing, memory, etc.) that bad is stronger than good (for an extensive review of the converging evidence across research domains, see Baumeister, 2001).

Our study extends the overt head movement paradigm to the domain of memory for images. The significant interaction found between head movement type and image valence in Experiment 2 is consistent with the results of Förster \& Strack (1996), who found the same interaction between head movement type and adjective valence. Obtaining a significant interaction between head movement type and image valence despite people's exceptional memory capacity for images offers further support for the motor-compatibility effect. The main idea that information is better encoded while compatible head movements are performed, as opposed to when incompatible head movements are performed, is in line with the findings of other researchers in many other domains (see the Introduction of this paper for a brief review) and is also consistent with findings of other research on stimulusresponse compatibility (Romaiguere et. al., 1993; Solarz, 1960).

Interestingly, the reversed meaning of head movements in Bulgarian culture seems to have no influence on the effect observed in previous studies. We found that upon asking how they gave nonverbal responses for "yes" and "no", oftentimes participants demonstrated nodding and shaking, respectively, whereas when implicitly tested about this right after the explicit test (the experimenter would informally ask the participant a yes/no question, the answer to which was known in advance), they would often respond in the traditional Bulgarian way (shaking for "yes" and nodding for "no"). Our findings may suggest that the motorcompatibility effect is independent of culture, but it may also be due to the strong influence of Western culture (e.g., through media, mass communication, globalization, etc.). The discrepancy suggests that the issue of which nonverbal head gestures Bulgarians perform to denote agreement/disagreement is not clear cut and is an interesting empirical question on its own.

\section*{Summary}

We conducted two experiments to test the effect of overt head movements on the encoding of positive and negative images. The main findings of our study are the following:
1) Negative images were remembered more than positive images, despite equal levels of arousal;
2) Head movement type interacted significantly with image valence during the free recall task, but not during the recognition task;
3) Contrary to what was expected, the direction of the interaction was the same with Bulgarian participants as has been reported in previous studies with nonBulgarians, despite the cultural difference in meaning of vertical and horizontal head movements.

Although interesting, these results require further investigation in future studies.

\section*{Acknowledgments}

We would like to thank the people who volunteered to participate in our experiment and Oksana Itkes for her assistance in finding some of the references.

\section*{References}

Baumeister, R., Bratslavsky, E., Finkenauer, C., \& Vohs, K. (2001). "Bad is Stronger than Good". Review of General Psychology, 5, 323-370.
Brady, T. F., Konkle, T., Alvarez, G. A., Oliva, A. (2008). Visual Long-term Memory Has a Massive Storage Capacity for Object Details. Proceedings of the National Academy of Sciences, 105, 14325-4329.
Briñol, P. \& Petty, R. E (2003). Overt head movements and persuasion: A self-validation analysis. Journal of Personality and Social Psychology, 84, 1123-1139.
Epley, N., \& Gilovich, T. (2001). Putting adjustment back in the anchoring and adjustment heuristic: Differential processing of self-generated and experimenter-provided anchors. Psychological Science, 12, 391-396.
Foley, M., \& Foy, J. (2008). Pictorial encoding effects and memory confusions in the Deese-Roediger-McDermott paradigm: Evidence for the activation of spontaneous imagery. Memory, 16, 712-727.
Förster, J, \& Strack, F. (1996). Influence of overt head movements on memory for valenced words: a case of conceptual-motor compatibility. Journal of Personality and Social Psychology, 71, 421-30.
Goldstein, A. G., \& Chance, J. E. (1970). Visual recognition memoryfor complex configurations. Perception \& Psychophysics, 9, 237-241.
Kensinger, E. A. (2009). Remembering the details: Effects of emotion. Emotion Review, 1, 99-113.
Kensinger, E. A., Garoff-Eaton, R. J., \& Schacter, D. L. (2007). Effects of emotion on memory specificity in young and older adults. Journal of Gerontology: Psychological Sciences, 62, 208-215.
Landauer, T. K. (1986). How much do people remember? Some estimates of the quantity of learned information in long-term memory. Cognitive Science, 10, 477-493.
Lang, P. J., Bradley, M. M., Cuthbert, B.N. (2008). International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical Report A-8. University of Florida, Gainesville, FL.
Ochsner, K. N. (2000). Are affective events richly recollected or simply familiar? The experience and process of recognizing feelings past. Journal of Experimental Psychology: General, 129, 242-261.
Roediger, H. L., III, \& McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. Journal of Experimental Psychology: Learning, Memory, \& Cognition, 21, 803-814.
Romaiguere, P., Hasbroucq, T., Possamai, C, A., \& Seal, J. (1993). Intensity to force translation: A new effect of
stimulus-response activity revealed by analysis of response time and electromyographic activity of a prime mover. Cognitive Brain Research, 1, 197-201.
Schacter, D. (1999). Suppressing False Recognition in Younger and Older Adults: The Distinctiveness Heuristic. Journal of Memory and Language, 40, 1-24.
Shepard, R. N. (1967). Recognition memory for words, sentences, and pictures. Journal of Verbal Learning and Verbal Behavior, 6, 156-163.
Snodgrass, J. G., \& Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. Journal of Experimental Psychology: General, 117, 34-50.
Solarz, Andrew K. (1960). Latency of Instrumental Responses as a Function of Compatibility with the Meaning of Eliciting Verbal Signs. Journal of Experimental Psychology, 59, 239-45.
Standing, L. (1973). Learning 10000 pictures. Quarterly Journal of Experimental Psychology, 25, 207-222.
Tom, G., Ramil, E., Zapanta, I., Demir, K., Lopez, S. (2006). The Role of Overt Head Movement and Attention in Persuasion. The Journal of Psychology, 140, 247-253.
Tom, G., Pettersen, P., Lau, T., Burton, T., Cook, J. (1991). The Role of Overt Head Movement in the Formation of Affect. Basic and Applied Social Psychology, 12, 28189.

Wells, G. L., \& Petty, R. E. (1980). The effects of overt head movements on persuasion: Compatibility and incompatibility of responses. Basic and Applied Social Psychology, 1, 219-230.
Wolfe, J. M. (1998). Visual Memory: What Do You Know about What You Saw? Current Biology, 8, 303-304.

\title{
Sequential Structure Suffices to Solve Nativist Puzzles
}

\author{
Rens Bod (rens.bod@uva.nl) \\ ILLC, University of Amsterdam \\ Science Park 107, Amsterdam, Netherlands
}

Keywords: language modeling; computational models of language learning.

\section*{Nativism versus Empiricism}

The debate between hierarchical versus sequential structure in language acquisition has recently flared up again (cf. Frank, Bod \& Christiansen 2012; Pesetsky 2013). Roughly, the nativist view on language endorses that human language acquisition is guided by innate rules that operate on hierarchical structures. The empiricist view assumes that language acquisition is the product of abstractions from empirical input but leaves it as an open question whether sequential or hierarchical structure is needed. Some empirical models use sequential structure (e.g. Reali \& Christiansen 2005) while other models are based on hierarchical structure (Bod 2009; Bod \& Smets 2012).

Much work in empirical language acquisition has focused on a relatively small set of phenomena such as auxiliary fronting. For example, Reali \& Christiansen (2005) argued that auxiliary fronting could be learned by linear models based on sequential structure, though Kam et al. (2008) showed that the success of these models depend on accidental English facts. Other empiricist approaches have taken the notion of structural dependency together with a combination operation as minimal requirements (e.g. Bod 2009), which overcomes the problems raised by Kam et al. (2008).

In Bod and Smets (2012) it was shown that a much larger set of phenomena can be learned by an empiricist computational model. These phenomena are well-known in the generativist literature (Ross 1967; Adger 2003) and are related to wh-questions, relative clause formation, topicalization, extraposition and left dislocation. It turned out that these hard cases can be learned by an unsupervised tree-substitution grammar induction algorithm that returns the sentence with the best-ranked derivation for a particular phenomenon, using only a very small fraction of the input a child receives.

However, Bod and Smets (2012) also observed that these nativist cases were learned by using relatively shallow structures with little or no hierarchy. This raised the question as to how much structure is actually needed to learn these syntactic constraints. In the current paper, we present a very simple model that reduces all syntactic structuring to concatenations of substrings without any hierarchy. We show that almost all results obtained by the hierarchical grammar in Bod \& Smets (2012) can also be learned by means of a sequential grammar using substringconcatenation only.

It should be stressed that the essence of the debate between nativism and empiricism lies often in the relative contribution of prior knowledge and linguistic experience (cf. Lidz et al. 2003; Ambridge \& Lieven 2011; Clark and Lappin 2011). Following the nativist view, the linguistic evidence is so hopelessly underdetermined that innate components are necessary. This Argument from the Poverty of the Stimulus can be phrased as follows (see Pullum \& Scholz 2002 for a detailed discussion):
(i) Children acquire a certain linguistic phenomenon
(ii) The linguistic input does not give enough evidence for acquiring the phenomenon
(iii) There has to be an innate component for the phenomenon

In this paper we falsify step (ii) for a number of linguistic phenomena that have been considered "parade cases" of innate constraints (Crain 1991; Crain and Thornton 2006). We will show that even if a linguistic phenomenon is not in a child's input, it can be learned by a sequential model using only a tiny fraction of child-directed utterances, i.e. the Adam corpus in Childes (MacWhinney 2000).

\section*{Methodology}

Our methodology is very simple: by means of concatenations of substrings (of parts of speech) of any length from the Adam corpus, we compute from the alternative sentences of a syntactic phenomenon (reported in the generativist literature) the sentence with the most probable shortest concatenation. Next, we check whether this sentence corresponds with the grammatical sentence. The shortest concatenation is defined as consisting of the minimal number of substrings (smoothed by the n-shortest concatenations, similar as in Bod and Smets 2012). In case there is more than one shortest concatenation, the most probable one is computed by multiplying the (smoothed) relative frequencies of these substrings in the corpus. For example, given a typical nativist problem like auxiliary fronting, the question is: how do we choose the correct sentence from among the alternatives (0) to (2):
(0) is the boy who is eating hungry?
(1) \(*\) is the boy who eating is hungry?
(2) \(*\) is the boy who is eating is hungry?

According to Adger (2003), Crain (1991) and others, this phenomenon is regulated by an innate principle. In our approach, instead, we produce all concatenations of
substrings that generate (the pos-sequences corresponding to) those sentences. Next, the sentence generated by the most probable shortest concatenation is compared with the grammatical expression.

\section*{An Example and Overvie w of the Results}

As an example we will look into the Left Branch Condition (Ross 1967; Adger 2003). This condition has to do with the difference in grammaticality between (3) and (4):
(3) which book did you read?
(4) *which did you read book?

When we let our model generate these two sentences by the shortest combinations of substrings from Adam, we get the respective concatenations ( \(3^{\prime}\) ) and (4'), where for reasons of readability we substituted the pos-tags with the words:
(3') [which book] + [did you read]
(4') [which] + [did you] + [read book]
In this case the shortest concatenation already breaks ties, thus we do not have to compute the most probable shortest concatenation (the latter actually being the typical case).

Table 1 gives an overview of the syntactic constraints/phenomena we have tested so far, and whether these can be successfully explained by the most probable shortest concatenation. The table shows that with only a tiny fraction of a child's input (i.e. just the sentences from the Adam corpus) the correct sentence can be predicted by our simple model for all but two of the phenomena. Our result approaches Bod and Smets (2012) which missed only one phenomenon rather than two, but which relied on a much more complex hierarchical model that induced full-fledged probabilistic tree-substitution grammars. In the future we will therefore also test with larger corpora in Childes.

Table 1: overview of phenomena tested
\begin{tabular}{lc} 
Phenomenon & Succesful? \\
\hline \hline Subject Auxiliary Fronting & yes \\
\hline \hline WH-Questions & \\
\hline Unbounded Scope & yes \\
Complex NP Constraint & yes \\
Coordinate Structure Constraint & no \\
Left Branch Condition & yes \\
Subject WH-questions & yes \\
WH in situ & yes \\
Superiority & yes \\
Extended Superiority & yes \\
Embedded WH-questions & yes \\
WH-islands & yes \\
\hline \hline Relative Clause Formation & \\
\hline Complex NP Constraint & yes \\
Coordinate Structure Constraint & yes \\
Sentential Subject Constraint & yes \\
Left Branch Condition & yes \\
\hline \hline Extraposition from NP & yes \\
\hline \hline
\end{tabular}

Topicalization
\begin{tabular}{ll}
\hline Complex NP Constraint & yes \\
Coordinate Structure Constraint & yes \\
Sentential Subject Constraint & yes \\
Left Branch Condition & yes \\
\hline \hline Left Dislocation & \\
\hline Coordinate Structure Constraint & no \\
Restriction & yes \\
\hline
\end{tabular}

\section*{References}

Adger, D. (2003). Core syntax: A minimalist approach. Oxford University Press..
Bod, R., Hay, J. \& Jannedy, S. eds. (2003). Probabilistic Linguistics, The MIT Press.
Bod, R. (2006). Exemplar-Based Syntax: How to Get Productivity from Examples. The Linguistic Review, 23(3), 291-320.
Bod, R. (2009). From Exemplar to Grammar: A Probabilistic Analogy-based Model of Language Learning. Cognitive Science, 33(5), 752-793.
Bod, R. \& Smets, M. (2012). Empiricist Solutions to Nativist Problems using Tree-Substitution Grammars, Proceedings Cognitive Models of Language Acquisition and Loss, EACL 2012, 10-18.
Clark, A. and Lappin, S. (2011). Linguistic Nativism and the Poverty of the Stimulus, Wiley-Blackwell.
Crain, S. (1991). Language acquisition in the absence of experience. Behavorial and Brain Sciences, 14, 597-612.
Crain, S. \& Thornton, R. (2006). Acquisition of syntax and semantics. In M. Traxler and M. Gernsbacher, editors, Handbook of Psycholinguistics. Elsevier.
Frank, S., Bod, R. \& Christiansen, M. (2012). How Hierarchical is Language Use? Proceedings of the Royal Society B, 297(1747), 4522-4531.
Frank, S. \& Bod, R. (2011). Insensitivity of the Human Sentence-Processing System to Hierarchical Structure, Psychological Science, 22(6), 829-834.
Kam, X., L. Stoyneshka, L. Tornyova, J. Fodor \& W. Sakas (2008). Bigrams and the Richness of the Stimulus. Cognitive Science, 32, 771-787.
Lidz, J., S. Waxman and J. Freedman (2003). What infants know about syntax but couldn't have learned: experimental evidence for syntactic structure at 18 months. Cognition, 89, B65-B73.
MacWhinney, B. (2000). The CHILDES project: Tools for analyzing talk. Mawah, NJ: Erlbaum.
Pesetsky, D. (2013), What is to be done?, Plenary talk, Linguistic Society of America, 1/4/2013.
Pullum, G. \& Scholz, B. (2002). Empirical assessment of stimulus povery arguments. The Linguist Review, 19(2002), 9-50.
Reali, F. \& Christiansen, M. (2005). Uncovering the richness of the stimulus: structure dependence and indirect statistical evidence. Cognitive Science, 29, 1007-1028.
Ross, J. (1967). Constraints on variables in syntax. PhD thesis, Massachusetts Institute of Technology.

\title{
Whorf for the \(21^{\text {st }}\) century: From interactive processing to linguistic relativity
}

\author{
Gary Lupyan (lupyan@wisc.edu) \\ Department of Psychology \\ University of Wisconsin-Madison \\ Madison, WI 53706
}

The crux of the Whorfian thesis is that our thought and behavior are influenced in deep ways by the language we use. In recent years we have seen a wave of rigorous and creative investigations of this thesis (Boroditsky, 2010; Wolff \& Holmes, 2011 for reviews). Yet, many researchers remain highly skeptical of findings purporting to support Whorfian claims (Gleitman \& Papafragou, 2005), and much confusion remains about how to integrate these findings into existing theories of cognition. A major barrier to understanding the degree to which various aspects of human cognition may be affected by speaking different languages is understanding the relationship between language-any language - and the rest of cognition. To remove this barrier we need to address a fundamental question: To what degree is normal human cognition actually language-augmented cognition? I will argue that a surprising variety of behavior previously assumed to be "nonverbal" shows signs of being influenced by linguistic factors and I will outline a theory of language-augmented thought that offers a mechanistic account of where we might expect to find effects of language on "nonverbal" cognition (Lupyan, 2012a, 2012b, for reviews).

One of the core features of language is using words to denote categories, e.g., using the word "dog" to refer to dogs. Words are commonly seen as a kind of "pointer" to concepts, the content of which is independent of language. In recent work, we have argued for an alternative: verbal labels do not simply point or refer to nonlinguistic concepts, but rather actively modulate conceptual representations that are brought online during "nonverbal" tasks. For example, Lupyan \& Thompson-Schill (2012) showed that hearing referential labels such as "dog" consistently enhanced picture recognition compared to equally familiar, predictive, and unambiguous nonverbal cues such as a barking sound. This label advantage extended to newly learned labels and sounds. Despite participants' equivalent facility in learning what a novel object is called and what sound it makes, newly learned verbal labels were subsequently more effective in activating the concept than nonverbal sounds. In particular, hearing a label appeared to activate more category-typical information than hearing equally predictive nonverbal cues. This and related findings that verbal labels selectively activate category-typical features is hypothesized to underlie detrimental effects of labeling on visual memory such as the ability to remember not just that one saw a chair, but what kind of chair it was (Lupyan, 2008a).

As a further example of the kinds of powerful and surprising effects that category labels have on putatively nonverbal tasks, consider the following results (summarized in Lupyan, 2012a): When asked to draw a figure with three
sides, all participants predictably drew triangles: \(50 \%\) were isosceles/equilateral and \(50 \%\) were parallel to the bottom of the page. When a separate group was asked to draw a "triangle," \(91 \%\) drew isosceles or equilateral triangles and \(82 \%\) drew triangles with bases parallel to the bottom of the page (the canonical horizontal orientation). These differences do not stem solely from pragmatics. In a speeded recognition task, participants were faster to verify isosceles than scalene triangles, and horizontally-oriented than oblique triangles, but only on trials on which they actually heard the word "triangle" and not on trials on which they viewed the same shapes after hearing "three-sided" (all factors withinsubjects). Finally, in an untimed visual-reasoning task, participants were asked to estimate the angle of the base of various three-sided polygons. These shapes were referred to as "triangle" or a "three-sided shape" (between subjects). As shown in Fig. 1, when the shape was referred to as a "triangle," its tilt was perceived as deviating more from the canonical (horizontal) as steeper than when the category name was omitted. On one interpretation, these results support the hypothesis that the representation activated by the word "triangle" are a better match to more "canonical" triangles than a formally equivalent cue. Despite denotative equivalence between "triangles" and "three-sided polygons" the category label "triangle" seems to reliably activate a more "canonical" triangle as measured by both explicit and implicit tasks_prima facie evidence of category labels augmenting underlying representations in systematic ways.


Figure 1: Left: Perceived orientation of shapes is systematically affected by whether they are called "triangles" or "three-sided". Right: Participants take longer to look at scalene (atypical) triangles when they hear the word "triangle."

Thus, referring to an object by its name appears to activate a different representation than when ostensibly the same concept is activated without using the name (Lupyan \& Thompson-Schill, 2012; Lupyan, 2008b).

Why do labels have these effects? On the present account these effects are a product of (1) the association history of the discrete label with numerous category exemplars and (2) the feedback of the label on conceptual/perceptual representations. Under the influence of this feedback, the representations of various entities (objects, relations, etc.) become more categorical. This account can explain, for example, findings of pervasive effects of language on color perception (e.g., Regier \& Kay, 2009; Lupyan, 2012a for discussion). Stated simply: the association of a label such as "green" with a range of colors means that when one sees a greenish color, the label is rapidly activated, temporarily warping the perceptual space. Viewing a green object becomes a hybrid visuo-linguistic experience.

To better understand this account, a simulation of how feedback label-feedback can augment conceptual and perceptual representations will be presented. Fig 2A shows a schematic of an interactive neural network trained on a bidirectional mapping between bit-vectors (representing featurebased object representations) and category labels (i.e., learning to label chairs as "chairs" and learning to activate a likely visual representation of a chair given the label). After training, when the network is presented with a perceptual input, the label becomes automatically activated, and then feeds back to affect representations as they unfold


Labels prevented from feeding back

Labels allowed to feed back
Figure 2. (A) Schematic of a neural network for studying the role of label feedback on object representations. (B) Activation dynamics in the "conceptual" layer. Each line represents an activation of a category exemplar over the course of a single trial. Color represents category membership.
in time in the "perceptual" and "conceptual" layers. We can then examine what role the label is playing in the activation and maintenance of the representation of a particular category exemplar by directly perturbing the activation of the label or its feedback onto these layers. Feedback from labels (whether activated by the network on its own, or provided externally) provides much more categorical (clustered) representations (Fig. 2b), leading to improved categorizationa prediction confirmed by overt categorization tasks (Lupyan, 2009; Lupyan, Rakison, \& McClelland, 2007; Lupyan \& Thompson-Schill, 2012). Additional evidence for verbal labels augmenting "nonverbal" representations comes from their apparent effects on basic visual perception. Visual rep-
resentations activated by verbal means appear to be differ-ent-specifically, more categorical-than ostensibly the same representations activated by nonverbal means (Lupyan \& Spivey, 2008, 2010; Lupyan, 2008b).

Given that small linguistic manipulations affect how perceptual and conceptual information is brought online even within the same language community, we may expect that the substantial cross-linguistic differences in human languages should have substantially larger consequences on "thought," but there seems to be fewer such differences than expected. I will argue that this curious absence of evidence is due to a dichotomy made by researchers between verbal and nonverbal processes (e.g., "thinking for speaking") with the consequence that investigators may of linguistic relativity may have been looking in the wrong places.

\section*{References}

Boroditsky, L. (2010). How the languages we speak shape the ways we think: the FAQs. In M. J. Spivey, M. Joanisse, \& K. McRae (Eds.), The Cambridge Handbook of Psycholinguistics (p. forthcoming). Cambridge: Cambridge University Press.
Gleitman, L., \& Papafragou, A. (2005). Language and thought. In K. Holyoak \& B. Morrison (Eds.), Cambridge Handbook of Thinking and Reasoning (pp. 633-661). Cambridge: Cambridge University Press.
Lupyan, G. (2008a). From chair to "chair:" A representational shift account of object labeling effects on memory. Journal of Experimental Psychology: General, 137(2), 348-369.
Lupyan, G. (2008b). The conceptual grouping effect: Categories matter (and named categories matter more). Cognition, 108(2), 566-577.
Lupyan, G. (2009). Extracommunicative Functions of Language: Verbal Interference Causes Selective Categorization Impairments. Psychonomic Bulletin \& Review, 16(4), 711718.

Lupyan, G. (2012a). Linguistically modulated perception and cognition: the label-feedback hypothesis. Frontiers in Cognition, 3(54).
Lupyan, G. (2012b). What do words do? Towards a theory of lan-guage-augmented thought. In B. H. Ross (Ed.), The Psychology of Learning and Motivation (Vol. 57, pp. 255297). Academic Press.

Lupyan, G., Rakison, D. H., \& McClelland, J. L. (2007). Language is not just for talking: labels facilitate learning of novel categories. Psychological Science, 18(12), 1077-1082.
Lupyan, G., \& Spivey, M. J. (2008). Perceptual processing is facilitated by ascribing meaning to novel stimuli. Current Biology, 18(10), R410-R412.
Lupyan, G., \& Spivey, M. J. (2010). Redundant spoken labels facilitate perception of multiple items. Attention, Perception, \& Psychophysics, 72(8), 2236-2253.
Lupyan, G., \& Thompson-Schill, S. L. (2012). The evocative power of words: Activation of concepts by verbal and nonverbal means. Journal of Experimental PsychologyGeneral, 141(1), 170-186. doi:10.1037/a0024904
Regier, T., \& Kay, P. (2009). Language, thought, and color: Whorf was half right. Trends in Cognitive Sciences, 13(10), 439-446. doi:10.1016/j.tics.2009.07.001
Wolff, P., \& Holmes, K. (2011). Linguistic relativity. Wiley Interdisciplinary Reviews: Cognitive Science, 2(3), 253-265.

\title{
How many features does it take to change a lightbulb?
}

\author{
David M W Powers (powers@computer.org) \\ Beijing Municipal Lab for Multimedia \& Intelligent Software Beijing University of Technology, Beijing, China \\ Brain Signals Laboratory \& \\ CSEM Centre for Knowledge \& Interaction Technology \\ Flinders University, Adelaide, South Australia
}

Keywords: Computational Psycholinguistics, Sensor Fusion, Epistemic Robotics, Cognitive Linguistics.

\section*{Introduction}

In the 1970s and 80s Cognitive Science and Cognitive Linguistics and Computational Psycholinguistics emerged as the boxes around our disciplines started to become straight-jackets, and research out of one discipline would start to make waves in others. The toy systems of Artificial Intelligence were reaching limits, and introspection by programmers and engineers was reinventing square wheels without any biological plausibility and in ignorance of relevant work across the cognitive sciences, while conversely, work in other fields often lacked the understanding of computability and complexity necessary to ensure that models were realistic and computationally plausible.

This is the starting point for the research program I have been undertaking for the last 35 years, seeking to build intelligent computer systems and computational cognitive models. The idea has been to try to build an intelligent system modelled on the way a baby learns about the world, culture, society and language. Conversely, the idea has been to explore theories from psychology, linguistics and neuroscience through the medium of computational models. The primary focus and agenda of our research program are summed up in Powers and Turk (1989): language and ontology are learned together through multimodal association.

\section*{Language \& Ontology}

Over the years, the breadth of the both the "Language" and "Ontology" learning aspect of the research has grown to include audio-visual speech, gesture and emotion recognition and synthesis, as well as robots both simulated and physical. The earliest models (Powers, 1983; 1984) selforganized with a clear dependency on closed class lexemes as the basic for syntactic structuring, and later work extended this to the levels of phonology and morphology (Powers, 1991;1997abc). In parallel, the same learning models, including both statistical and neurally based coclustering models, were also used to learn noun, verb and preposition semantics in the context of a robot world simulation, and remain of major importance in our research (Pfitzner et al. 2009; Leibbrandt \& Powers, 2010;2012).

The physical models ranged from a robot baby that turned and looked at you if your talked to it or touched it
(Powers, 2002), whilst wheeled robots took on a life of their own (Powers et al., 2012) with simulated Teaching Head applications becoming a major focus (e.g. Milne et al., 2011-12) as an outcome of a major ARC/NHMRC Thinking Systems initiative that not only funded our "Thinking Head" project, but our colleagues' "Thinking Hand" and "Thinking Feet" projects. \({ }^{1}\) Whereas we concentrated on hands and feet and wheels for locomotion, with fairly conventional path planners for navigation, and made use of conventional robotic grippers for grasping, or much safer simulated grasping for our Hybrid World (Newman et al., 2010), this Thinking Hand team concentrated on such matters as how to hold a glass or a light bulb without breaking it, whilst the Thinking Feet team looked at spiking models for navigation.

One of the core driving forces for our work at this point is the realization that our "five senses" actually hide a multitude of specific sensors and percepts each. For example the fingerprints on the hand distinguish the transverse motion of slip vs the normal force of pressure, in ensuring we neither drop nor crush the light bulb. Our two eyes and four types of visual transducer, and the different afferents and efferents involved in controling convergence and focus and aperture, combine with our inner and outer ears to direct our gaze and focus sound, with two different Nyquist tradeoffs of time vs frequency, with 3D balance and inertial sensing. Much of our focus is combining together different senses or subsenses, or discriminating out the different features from our combined senses that have particular value in the tasks we attempt.

This combination of multiple sensory or feedback inputs is called fusion (Lewis and Powers, 2000;2003) and complements processes of signal deconvolution (Li et al., 2003) and feature selection (Atyabi et al., 2012). Computationally it is not effective to learn by throwing all the mass of sensory input together into one big vector and trying to make sense of it ('early fusion'), but nor is it effective to try to deal with each sense or sensor on its own to do the task, and at the last minute vote to fuse sources or models ('late fusion'). Rather we need to look at the similarities and correlations (e.g. whose lips are moving to know who is talking to us) and dissimilarity and independence (viz. we don't want a committee of yes-men, but of independent thinkers, so we search a large space of potential solutions). The first step is often to figure out how many independent components, or clusters or features there are - or we can use algorithms that decide on the fly.

\section*{References}

Ali, H. B. \& Powers, D. M. W. (2011), Comparison of Regionbased and Weighted Principal Component Analysis and Locally Salient ICA in terms of Facial Expression Recognition, Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, 368: 81-89.
Atyabi, A., Luerssen, M.H., Fitzgibbon, S.P., \& Powers, D.M., 2012. Evolutionary feature selection and electrode reduction for EEG classification. IEEE Congress on Evolutionary Computation (CEC).
Fitzgibbon, S. P., Powers, D. M. W., Pope, K., \& Clark, C. R. (2007). Removal of EEG noise and artefact using blind source separation. Journal of Clinical Neurophysiology 24(3):232-243.
Fitzgibbon, SP, Lewis, T, Powers, DMW, Whitham, E, Willoughby, J \& Pope, K. (2013), Surface Laplacian of Central Scalp Electrical Signals is Insensitive to Muscle Contamination, IEEE Transactions On Biomedical Engineering, 60:1, 4-9.
Jia, X., Han, Y., Powers, David M.W., Zheng, Z., Sha, 1., Bao, X. and Fei, J., (2012). Spatial and temporal visual speech feature for Chinese phonemes, Journal of Information and Computational Science 9:14, 4177-4185.
Leibbrandt, R. E. \& Powers, D. M. W. (2010), Frequent frames as cues to part-of-speech in Dutch: why filler frequency matters. 32nd Annual Meeting of the Cognitive Science Society (CogSci2010), 2680-2685.
Leibbrandt, R. E. \& Powers, D. M. W. (2012). Robust Induction of Parts-of-Speech in Child-Directed Language by Co-Clustering of Words and Contexts, Joint Workshop of ROBUS and UNSUP, Conference of the European Chapter of the Association for Computational Linguistics, 44-54.
Lewis, T. W. \& Powers, D. M. W. (2000). Lip Feature Extraction using Red Exclusion. In P. Eades and J. Jin (Eds), Visualisation 2000, CRPIT 2:61-70.
Lewis, T. W. and D. M. W. Powers (2003). Audio-Visual Speech Recognition using Red Exclusion and Neural Networks. Journal of Research and Practice in Information Technology 35, 1:41-64.
Li, Y., Powers, D. M. W. \& Pope, K. (2003). A new approach to blind signal deconvolution using recurrent neural networks. International Journal of KnowledgeBased Intelligent Engineering Systems 7:2, 62-69.
Milne, M., Leibbrandt, R. E., Lewis, T. W., Luerssen, M. H. \& Powers, D. M. W. (2010). Development of a virtual agent based social tutor for children with autism spectrum disorders. Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN2010), 1555-1563.
Milne, M., Luerssen, M. H., Lewis, T. W., Leibbrandt, R. E. \& Powers. D. M. W. (2011). Designing and Evaluating Interactive Agents as Social Tutors for Children with Autism Spectrum Disorder. In Diana Perez-Marin and Ismael Pascual-Nieto (Eds.) Conversational Agents and Natural Language Interaction: Techniques and Effective Practices, Madrid, Spain: IGI Global.
Newman, W., Franzel, D., Matsumoto, T., Leibbrandt, R. E., Lewis, T. W., Luerssen, M. H., and Powers, D. M. W.
(2010). Hybrid world object tracking for a virtual teaching agent. Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN2010), 2244-2252.
Pfitzner, D. M., Leibbrandt, R. E. \& Powers, D. M. W. (2009) Characterization and Evaluation of Similarity Measures for Pairs of Clusterings, Knowledge and Information Systems 19:361-394.
Powers, D. M. W. (1983). Neurolinguistics and Psycholinguistics as a Basis for Computer Acquisition of Natural Language," SIGART, 84, 29-34.
Powers, D. M. W. (1984) Natural Language the Natural Way, Computer Compacts, 100-109.
Powers, D. M. W. \& Turk, C. C. R. (1989). Machine Learning of Natural Language, Berlin: Springer-Verlag.
Powers, D. M. W. (1991) Goals, Issues and Directions in Machine Learning of Natural Language and Ontology, SIGART Bulletin 2:1, 101-114.
Powers, D. M. W., Clark, C. R., Dixon, S. E. \& Weber, D. L. (1996). Cocktails and Brainwaves: Experiments with Complex and Subliminal Auditory Stimuli, Proceedings of the Australian and New Zealand Conference on Intelligent Information Systems, 68-71.
Powers, D. M. W. (1997). Learning and Application of Differential Grammars, CoNLL97: Computational Natural Language Learning, 88-96.
Powers, D. M. W. (1997). Perceptual Foundations for Cognitive Linguistics, International Conference on Cognitive Linguistics, 173.
Powers, D. M. W. (1997). Metaphor and Learning: The Phonology of Syntax, International Conference on Cognitive Linguistics, 260-265.
Powers, D. M. W. (1997). Unsupervised learning of linguistic structure: an empirical evaluation, International Journal of Corpus Linguistics 2\#1, 91-131.
Powers, D. M. W. (2002). Robot babies: what can they teach us about language acquisition?" In J. Leather and J. Van Dam, (Eds) The Ecology of Language Acquisition, Kluwer Academic
Powers. D. M. W. (2003). Recall and Precision versus the Bookmaker. International Conference on Cognitive Science, 529-534.
Powers, D. M. W. (2011), Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness \& Correlation, Journal of Machine Learning Technologies 2:1, 37-63.
Powers, D. M. W. (2012). The Problem of Kappa", 13th Conference of the European Chapter of the Association for Computational Linguistics, 345-355.
Powers, D. M. W., Atyabi, A. \& Anderson, T. (2012). MAGIC Robots - Resilient Communications, Spring World Congress on Engineering \& Technology, IEEE:591-594
Yang, D. Q. and Powers, D. M. W. (2010), Using Grammatical Relations to Automate Thesaurus Construction, Journal of Research and Practice in Information Technology 42\#2:129-146

\footnotetext{
\({ }^{1}\) http://www.arc.gov.au/ncgp/sri/TS sumapps 06.htm
}

\title{
Experiments in dynamic group action and decision making: How crowds of people can walk a tightrope together and survive a zombie attack
}

\author{
Daniel C. Richardson (dcr@eyethink.org) \\ Peter Riefer (peter.riefer.11@ucl.ac.uk) \\ Bradley Love (b.love@ucl.ac.uk) \\ Cognitive, Perceptual \& Brain sciences, University College London \\ Gower Street, London WC1E 6BT, United Kingdom \\ Beau Lotto (lotto@ucl.ac.uk) \\ Richard C. Clarke (richard.clarke@ucl.ac.uk) \\ Institute of Ophthalmology, University College London, 11-43 Bath Street, London EC1V 9EL, United Kingdom
}

\author{
Rick Dale (rdale@ucmerced.edu) \\ Cognitive \& Information Sciences \\ University of California, Merced, CA 95343, USA
}

\author{
John Rogers (john@delosis.com) \\ James Ireland (james@delosis.com) \\ Delosis, 8 Grosvenor Road, Twickenham, Middlesex TW1 4AE, United Kingdom
}

Keywords: joint action; wisdom of crowds; group behaviour, situated cognition; decision making; public goods games

We present results from a new paradigm: mass participation games. In our experiments, hundreds of people can play a computer game simultaneously using audience response handsets. We can collect responses from a lecture hall full of people with the precision of a laboratory cubicle. We have studied two games: continuous, action games where participants cooperate to achieve a goal; and decisionmaking paradigms in which participants make repeated choices to maximise their own or the group's rewards. We address a range of theoretical questions with experimental manipulations and computer modelling. Do participants play as if they were alone, or as a group? If so, do they represent the group as a single entity, or a collection of other agents? What are the dynamics of these behaviours, with learning across many trials? Lastly, what does it feel like to act in concert, or in competition, with a room full of people?

There is wisdom in a crowd. The averaged response of a crowd usually betters any of the individual guesses, whether they are guessing the weight of a cow (Galton, 1907), predicting the stock market or making bets about geopolitical events (Surowiecki, 2004). But is the superiority of crowds restricted to single, static decisions? There may be wisdom in a crowd but what happens when they have to act together? What happens when they have to make decisions - pervasive in society - that trade off their own interests with those of the group? For these actions and decisions to be made, how do they learn to predict the behaviour of the group? We addressed these questions by developing mass participation games, in which participants cooperate or compete, maximising their own rewards or those of the group \({ }^{1}\).

\section*{Tightrope walking: a cooperative action game}

In our first, action-based game participants saw on a projection screen a picture of a man holding a pole, balanced on rope (Figure 1). Each participant held a handset and pressed one of two buttons. A laptop computer collected the responses and controlled the movements of the tightrope walker. Each time one of the participants pressed a button, it sent a small nudge to the tightrope walker, sending him left or right. A game ended when the tightrope walker fell.

Analysing the time he stayed aloft and tracking individual responses, we found that on successful games participants were able to anticipate and compensate for the behaviour of the group. This conclusion was supported by agent based simulations. In later conditions we instructed the participants to vocalise their button presses as they made them. The evidence is that they made use of this information, and were better able to predict and compensate for each others' actions.


Figure 1. The tightrope walker game, the participants and the correlation between performance and anticipations

\footnotetext{
\({ }^{1}\) If this paper is presented as a talk at the Cognitive Science conference, then the audience will of course be invited to play these games
}

\section*{Zombie attack: A public goods game}

Our second mass participation game studied decisionmaking in a public goods game. A contribution tin in an office kitchen is a public goods game. If everyone contributes money each week, it will pay for everyone's coffee. But a single person could chose not to contribute: they would get free coffee. If everyone followed this strategy, then the whole system would collapse and there would be no coffee. The trade-off between private and public gain is at the heart of public goods games. They are a standard tool in economic theories, used as a model for a huge range of activity from traffic patterns to tax returns.

Most empirical studies of public goods games use a small number of participants or have a small number of trials. We believe these miss the essential character of decisions made outside of the laboratory. People make these choices continuously throughout their lives, within the context of a social interaction. To understand how they learn to anticipate the actions of others, to see how social forces sway the choice to be selfish or cooperative, we argue it is vital to study the behaviour of a large group of people, present in a room together, playing repeated trials over time.

In our Zombie attack game, participants made a binary choice on each trial (Figure 2). They pressed one button to hide from the zombies, and another to fight. Rewards for the individual and the group decreased as more people chose to hide, but increased as more people chose to fight. Feedback on groups' decisions and rewards were shown. In one condition, they were told how many people had changed their choice from the previous game. Over the course of 25 games, we found that the switching feedback influenced strategies: without it, more people chose to fight. Though fighters always score less than hiders individually, group scores increased across the no feedback group. This was because the higher number of fighters increased the scores of those who chose to hide too. This pattern of results shows that participants are not just making a rational decision about the relative rewards, but also making a social decision that is shaped by the perceived actions of others.

\section*{Future Directions}

We continue to collect data with these paradigms to answer a range of questions. Does the size of the group influence the group dynamics? How are participants learning about each others' behaviour and shaping their actions? We are answering these question by manipulating the information participants have about each other, and by developing computer models of the process. Lastly, we are interested in the social phenomenology of group dynamics. What social forces might shape group cohesion? How does it feel to be part of a successful, coordinated group? Anecdotal evidence suggest that there are complex issues at play. During one zombie game, a small group stood up and shouted, 'Everyone fight! Stand up if you are fighting!'. But later, in debriefing, members of the group admitted they were all choosing to hide and maximise their individual rewards. Clearly, though the games themselves are simple, more research is required to understand the choices made when they are played in the context of mass participation.


Figure 2. The zombie feedback screen, participants playing the game, and the relationship between feedback on switching behaviour and group decisions.

\section*{Relevant publications of authors}

\section*{Decision making and modelling}

Knox, W.B., Otto, A.R., Stone, P., \& Love, B.C. (2012). The Nature of Belief-Directed Exploratory Choice in Human Decision-Making. Frontiers in Psychology,2,398.
Otto, A.R., \& Love, B.C. (2010). You Don't Want To Know What You're Missing: When Information about Forgone Rewards Impedes Dynamic Decision Making. Judgment and Decision Making, 5, 1-10.
Otto, A.R., Gureckis, T.M., Markman, A.B., \& Love, B.C. (2009). Navigating through Abstract Decision Spaces: Evaluating the Role of State Generalization in a Dynamic Decision-Making Task. Psychonomic Bulletin \& Review, 16, 957-963.
Gureckis, T. M., \& Love, B. C. (2009). Short Term Gains, Long Term Pains: Reinforcement Learning in Dynamic Environments. Cognition, 113, 293-313.
Gureckis, T. M., \& Love, B. C. (2009). Learning in Noise: Dynamic Decision-Making in a Variable Environment. Journal of Mathematical Psychology, 150, 180-193.

\section*{Interpersonal coordination}

Richardson, D.C., Dale, R., \& Kirkham, N.Z. (2007). The art of conversation is coordination: Common ground and the coupling of eye movements during dialogue. Psychological Science, 18(5), 407-413.
Richardson, D.C., Dale, R., \& Tomlinson, J.M. (2009). Conversation, gaze coordination, and beliefs about visual context. Cognitive Science, 33(8), 1468-1482.
Richardson, D. C., Street, C. N. H., Tan, J. Y. M., Kirkham, N. Z., Hoover, M. A., \& Ghane Cavanaugh, A. (2012). Joint perception: gaze and social context. Frontiers in Human Neuroscience, 6
Dale, R., Kirkham. N.Z., \& Richardson D.C. (2011). The Dynamics of Reference and Shared Visual Attention, Frontiers in Cognition.
Richardson, D. C., Dale, R., Rogers, J., \& Ireland J. (2011). How do 100 people walk a tightrope together? An experiment in large scale joint action. In L. Carlson, C. Hölscher, \& T. Shipley (Eds.), Proceedings of the 33rd Annual Conference of the Cognitive Science Society, Austin, TX: Cognitive Science Society.

\title{
Mental Simulation for Grounding Object Cognition
}

\author{
Haline E. Schendan (haline.schendan@plymouth.ac.uk) \\ Cognition Institute, School of Psychology, University of Plymouth, Drake Circus \\ Plymouth, Devon PL4 8AA UK
}

\begin{abstract}
Grounded (embodied) theories of cognition propose that memory, including knowledge and meaning, is grounded in sensorimotor and mental state processes. The main proposed mechanism for how memory is grounded is mental simulation. Simulation occurs when neural activity in modal association cortex triggers time-locked, recurrent and feedback activity across multiple lower-level modal processing areas from which the memory was initially constructed. Through this distributed multi-regional activity, seeing an object or reading its name (e.g., "dog") re-enacts associated features that were stored during earlier learning experiences (e.g. its shape, color, motion, actions with it), thereby constructing cognition, memory, and meaning. This paper reviews convergent evidence from cognitive neuroscience of mental imagery, object cognition, and memory that supports a multi-state interactive (MUSI) account of automatic and strategic mental simulation mechanisms that can ground memory, including the meaning, of objects in modal processing of visual features.
\end{abstract}

Keywords: Embodiment, grounded cognition; category; concept; meaning, memory; shape; object; vision; brain.

The MUSI account of the brain dynamics of visual object cognition proposes that posterior object processing areas activate at different times in at least 3 states performing distinct functions (Schendan \& Ganis, 2012; Schendan \& Kutas, 2007). In state 1 , initial activation of object processing cortex feeds forward from occipital to anterior temporal cortex. In this state, from \(\sim 120\) to 200 ms , an object is perceptually categorized coarsely for the first time (Ganis \& Schendan, 2008; Ganis, Smith, \& Schendan, 2012; Schendan \& Ganis, 2013; Schendan, Ganis, \& Kutas, 1998). An event-related potential (ERP) that localizes to occipitotemporal cortex, the N170/VPP, shows the first clear object-sensitivity (i.e., greater for intact objects than scrambled versions depicting no figure), which is a hallmark of this cortex using functional magnetic resonance imaging (fMRI) (Schendan \& Lucia, 2010; Schendan \& Stern, 2007). However, cognitive factors (e.g., mental imagery, category decision success, meaning, semantic context) modulate this cortex sensitively in fMRI studies but do not likewise affect the N170/VPP (e.g., Ganis \& Kutas, 2003; Schendan \& Lucia, 2009; Schendan \& Lucia, 2010; Schendan \& Maher, 2009). Thus object information activated in state 1 supports categorical perception, but cognition that enables complex behavior (e.g., deciding the object is a member of the dog category) does not start until later, in a second state.

State 2 operates from \(\sim 200\) to 500 ms when occipitotemporal cortex is activated again but in a sustained, interactive manner dominated by feedback and recurrent
processing among these areas and with ventrolateral prefrontal cortex (VLPFC). The frontal N3(00) complex is the first ERP to reflect activity in occipitotemporal cortex related to the success of visual object cognition. Like the N170/VPP (state 1) and fMRI activation, the N3 is objectsensitive, category-specific, and shows adaptation effects (Ganis \& Schendan, 2008; Schendan \& Ganis, 2012; Schendan \& Lucia, 2010). However, unlike state 1 but like occipitotemporal and VLPFC activity in fMRI, the N3 varies dramatically with mental imagery and factors affecting category decision success, such as stimulus typicality and impoverishment, implicit memory, knowledge, and meaning (Ganis \& Kutas, 2003; Philiastides \& Sajda, 2007; Schendan \& Ganis, 2012; Schendan \& Kutas, 2002, 2003, 2007; Schendan \& Lucia, 2009, 2010; Schendan \& Maher, 2009; Schendan \& Stern, 2008; Voss, Schendan, \& Paller, 2010). Later from 300 to 500 ms , the centroparietal N400 reflects semantic memory activation related to processing word-related information in anterior temporal cortex and VLPFC (Kutas \& Federmeier, 2011). Intriguingly, N3 effects start and peak before those on the N400, placing the N3 in a temporal position to reflect processes supporting mental simulation of object information that constructs the meaning analyses indexed by the N400. State 2 reflects decision, implicit memory, knowledge, and meaning processes distinct from earlier state 1 and later state 3 processes.

State 3 operates from \(\sim 400\) to 900 ms during complex cognitive tasks and evaluates internally the accuracy of earlier and ongoing decision processes and executes verification processes, including effortful, strategic, conscious mental simulations. These brain dynamics are reflected in a centroparietal late positive complex (LPC) that distinguishes between correct and wrong decisions but does not vary with how well the stimulus matches memory, which, by contrast, sensitively modulates the N3 (Schendan \& Kutas, 2002; Schendan \& Maher, 2009). The LPC varies with episodic recollection, as when recalling details of the learning experience during recognition and mental imagery tasks (Rugg \& Curran, 2007; Schendan \& Ganis, 2012), as does a default mode network that connects strongly with the mediotemporal system for episodic memory and is associated with episodic simulation and strategic, conscious mental imagery (Ganis \& Schendan, 2011; Schacter, Addis, \& Buckner, 2008). Such late processes, however, may also support complex semantic analysis (e.g., Sitnikova, Goff, \& Kuperberg, 2009). Thus the LPC reflects internal evaluation and verification processes that also support strategic, conscious, goal-driven mental simulation that can contribute
to grounding cognition in more abstract and complex ways than earlier automatic mental simulation.

The MUSI account can explain object cognition as well as the brain mechanisms of mental simulation to ground cognition in visual object processing, positing two functionally-distinct states of mental simulation: Earlier automatic simulation and later strategic simulation (Schendan \& Ganis, 2012). Crucially for grounded cognition theory, the pattern of mental imagery findings on the N3, N400, and LPC resembles that for repetition priming of perceived pictures, implicating these ERPs as markers of mental simulation. Following initial categorical perception of objects in state 1 , interactive, top-down and reflexive feedback, and recurrent processes in state 2 support automatic mental simulation to ground knowledge and meaning (Barsalou, 2009) in modal processing of visual features in occipitotemporal cortex (N3) and word-related semantic processes in anterior temporal cortex (N400). The second type of mental simulation is strategic, goal-directed, and conscious and recruited when the task demands internal evaluation of cognition, as in mental imagery and episodic memory tasks. This simulation reflects intentional top-down processes directed by lateral prefrontal and posterior parietal networks for attention, cognitive control, and working memory (LPC). These neural markers of automatic and strategic mental simulation should be the focus of needed research into the brain mechanisms for how modal information processing grounds cognition.

\section*{Acknowledgments}

EU FP7 PCIG09-GA-2011-294144-COGNITSIMS.

\section*{References}

Barsalou, L. W. (2009). Simulation, situated conceptualization, and prediction. Philos Trans \(R\) Soc Lond B Biol Sci, 364(1521), 1281-1289.
Ganis, G., \& Kutas, M. (2003). An electrophysiological study of scene effects on object identification. Brain Res Cogn Brain Res, 16(2), 123-144.
Ganis, G., \& Schendan, H. E. (2008). Visual mental imagery and perception produce opposite adaptation effects on early brain potentials. Neuroimage, 42(4), 1714-1727.
Ganis, G., \& Schendan, H. E. (2011). Mental imagery. Wiley Interdisciplinary Reviews: Cognitive Science, 2(3), 239-252.
Ganis, G., Smith, D., \& Schendan, H. E. (2012). The N170, not the P1, indexes the earliest time for categorical perception of faces, regardless of interstimulus variance. Neuroimage, 62(3), 1563-1574.
Kutas, M., \& Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). Аппи Rev Psychol, 62, 621-647.
Philiastides, M. G., \& Sajda, P. (2007). EEG-informed fMRI reveals spatiotemporal characteristics of perceptual decision making. J Neurosci, 27(48), 13082-13091.

Rugg, M. D., \& Curran, T. (2007). Event-related potentials and recognition memory. Trends Cogn Sci, 11(6), 251-7.
Schacter, D. L., Addis, D. R., \& Buckner, R. L. (2008). Episodic simulation of future events: concepts, data, and applications. Ann \(N\) Y Acad Sci, 1124, 39-60.
Schendan, H. E., \& Ganis, G. (2012). Electrophysiological potentials reveal cortical mechanisms for mental imagery, mental simulation, and grounded (embodied) cognition. Frontiers in Psychology, 3(Article 329), 1-22.
Schendan, H. E., \& Ganis, G. (2013). Face-Specificity Is Robust across Diverse Stimuli and Individual People, Even When Interstimulus Variance Is Zero. Psychophysiology.
Schendan, H. E., Ganis, G., \& Kutas, M. (1998). Neurophysiological evidence for visual perceptual categorization of words and faces within 150 ms . Psychophysiology, 35(3), 240-251.
Schendan, H. E., \& Kutas, M. (2002). Neurophysiological evidence for two processing times for visual object identification. Neuropsychologia, 40(7), 931-945.
Schendan, H. E., \& Kutas, M. (2003). Time course of processes and representations supporting visual object identification and memory. Journal of Cognitive Neuroscience, 15(1), 111-135.
Schendan, H. E., \& Kutas, M. (2007). Neurophysiological Evidence for the Time Course of Activation of Global Shape, Part, and Local Contour Representations during Visual Object Categorization and Memory. J Cogn Neurosci, 19(5), 734-749.
Schendan, H. E., \& Lucia, L. C. (2009). Visual object cognition precedes but also temporally overlaps mental rotation. Brain Research, 1294, 91-105.
Schendan, H. E., \& Lucia, L. C. (2010). Object-Sensitive Activity Reflects Earlier Perceptual and Later Cognitive Processing of Visual Objects between 95 and 500 ms . Brain Research, 1329, 124-141.
Schendan, H. E., \& Maher, S. M. (2009). Object knowledge during entry-level categorization is activated and modified by implicit memory after 200 ms . Neuroimage, 44(4), 1423-1438.
Schendan, H. E., \& Stern, C. E. (2007). Mental rotation and object categorization share a common network of prefrontal and dorsal and ventral regions of posterior cortex. Neuroimage, 35(3), 1264-1277.
Schendan, H. E., \& Stern, C. E. (2008). Where Vision Meets Memory: Prefrontal-Posterior Networks for Visual Object Constancy during Categorization and Recognition. Cerebral Cortex, 18(7), 1695-1711.
Sitnikova, T., Goff, D., \& Kuperberg, G. R. (2009). Neurocognitive abnormalities during comprehension of real-world goal-directed behaviors in schizophrenia. \(J\) Abnorm Psychol, 118(2), 256-277.
Voss, J. L., Schendan, H. E., \& Paller, K. A. (2010). Finding meaning in novel geometric shapes influences electrophysiological correlates of repetition and dissociates perceptual and conceptual priming. Neuroimage, 49(3), 2879-2889.

\title{
Approximating Bayesian inference with a sparse distributed memory system
}

\author{
Joshua T. Abbott (joshua.abbott@berkeley.edu) \\ Jessica B. Hamrick (jhamrick @berkeley.edu) \\ Thomas L. Griffiths (tom_griffiths@berkeley.edu) \\ Department of Psychology, University of California, Berkeley, CA 94720 USA
}

\begin{abstract}
Probabilistic models of cognition have enjoyed recent success in explaining how people make inductive inferences. Yet, the difficult computations over structured representations that are often required by these models seem incompatible with the continuous and distributed nature of human minds. To reconcile this issue, and to understand the implications of constraints on probabilistic models, we take the approach of formalizing the mechanisms by which cognitive and neural processes could approximate Bayesian inference. Specifically, we show that an associative memory system using sparse, distributed representations can be reinterpreted as an importance sampler, a Monte Carlo method of approximating Bayesian inference. This capacity is illustrated through two case studies: a simple letter reconstruction task, and the classic problem of property induction. Broadly, our work demonstrates that probabilistic models can be implemented in a practical, distributed manner, and helps bridge the gap between algorithmic- and computationallevel models of cognition.
\end{abstract}

Keywords: Bayesian inference, importance sampling, rational process models, associative memory models, sparse distributed memory

\section*{Introduction}

Probabilistic models of cognition can be used to explain the complex inductive inferences people make every day, such as identifying the content of images or learning new concepts from limited evidence (Griffiths, Chater, Kemp, Perfors, \& Tenenbaum, 2010; Tenenbaum, Kemp, Griffiths, \& Goodman, 2011). However, these models are typically formulated at what Marr (1982) called the computational level, focusing on the abstract problems people have to solve and their ideal solutions. As a result, they explain why people behave the way they do, rather than how cognitive and neural processes support these behaviors. This approach is thus quite different from previous work on modeling human cognition, which focused on Marr's algorithmic and implementation levels, and has been criticized because it seems to imply that human minds and brains need to solve intractable computational problems and use structured representations (Gigerenzer \& Todd, 1999; McClelland et al., 2010).

Understanding the actual commitments that computational-level accounts of human cognition based on probabilistic models make at the algorithmic and implementation level requires considering how these levels of analysis could be bridged (Griffiths, Vul, \& Sanborn, 2012). Identifying specific cognitive algorithms and neural architectures that can approximate Bayesian inference is a key step towards knowing whether it really poses an intractable problem for human minds, or whether structured representations need to be used to implement models that involve structured probability distributions. In this paper,
we take on this challenge by showing that an associative memory using sparse distributed representations can be used to approximate Bayesian inference, producing behavior consistent with a structured statistical model while using distributed representations of the kind normally associated with artificial neural networks.

The associative memory that we use to approximate Bayesian inference implements a Monte Carlo algorithm known as importance sampling. Previous work has shown that this algorithm can be implemented in a common psychological process model - an exemplar model (Shi, Griffiths, Feldman, \& Sanborn, 2010). Shi and Griffiths (2009) further demonstrated that importance sampling can be implemented with a radial basis function neural network. However, this neural network used a localist representation, in which each hypothesis considered by the model had to be represented with a single neuron - a "grandmother cell." While this might be plausible for modeling aspects of perception in which a wide range of neurons prefer specific stimuli, it becomes less plausible for modeling complex cognitive tasks in which hypotheses correspond to structured representations. For example, having separate neurons for each concept or causal structure we consider seems implausible.

We demonstrate that an associative memory that uses distributed representations - specifically, sparse distributed memory (SDM) (Kanerva, 1988, 1993) - can be used to approximate Bayesian inference through importance sampling. The underlying idea is simple: we use the associative memory to store and retrieve exemplars, allowing us to build on the equivalence between exemplar models and importance sampling. The critical advance is that this is done using distributed representations, meaning that arbitrary hypotheses can be represented, and arbitrary distributions of exemplars encoded by a single architecture. We show that the SDM naturally implements one class of Bayesian models, and explain how to generalize it to implement a broader range of models.

The plan of the paper is as follows. First, we give a brief overview of performing Bayesian inference with importance sampling and summarize how the sparse distributed memory system is implemented. Next, we formalize how importance sampling can be performed using a SDM. We provide two case studies drawn from existing literature in which we use the SDM to approximate existing Bayesian models. The first case study is a simple example involving reconstructing English letters from noisy inputs, and the second is a more sophisticated model of property induction. We conclude the paper with a discussion of implications and future directions.

\section*{Background}

Our results depend on two important sets of mathematical ideas: approximating Bayesian inference by importance sampling, and sparse distributed memories. We introduce these ideas in turn.

\section*{Bayesian inference and importance sampling}

Probabilistic models of cognition provide rational solutions to problems of inductive inference, where probability distributions represent degrees of belief and are updated as more data becomes available. Beliefs are updated by applying Bayes' rule, which says that the posterior probability of a hypothesis, \(h\), given observed data, \(d\), is proportional to the probability of observing \(d\) if \(h\) were the correct hypothesis (known as the likelihood) multiplied by the prior probability of that hypothesis:
\[
\begin{equation*}
p(h \mid d)=\frac{p(d \mid h) p(h)}{\int_{\mathcal{H}} p(d \mid h) p(h) d h} \tag{1}
\end{equation*}
\]

Unfortunately, computing the integral in the denominator is computationally expensive and often intractable. This has resulted in the development of many algorithms for approximating Bayesian inference.

For the sake of illustration, consider the case in which we have noisy observations \(x\) of a stimulus \(x^{*}\). To recover the value of \(x^{*}\), we use Bayes' rule to compute the posterior distribution over \(x^{*}\) :
\[
\begin{equation*}
p\left(x^{*} \mid x\right)=\frac{p\left(x \mid x^{*}\right) p\left(x^{*}\right)}{\int_{x^{*}} p\left(x \mid x^{*}\right) p\left(x^{*}\right) d x^{*}} \tag{2}
\end{equation*}
\]

It is often desirable to compute the expectation of the posterior distribution over some function \(f\left(x^{*}\right)\) :
\[
\begin{equation*}
E\left[f\left(x^{*}\right) \mid x\right]=\int f\left(x^{*}\right) p\left(x^{*} \mid x\right) d x^{*} \tag{3}
\end{equation*}
\]
where the choice of \(f\left(x^{*}\right)\) depends on the task. However, evaluating this expectation still requires computing the full posterior distribution.

To approximate expectations over posterior distributions, we can use a Monte Carlo method known as importance sampling (see, e.g., Neal, 1993), in which a finite set of samples are used to represent the posterior. These samples are drawn from a surrogate distribution \(q\left(x^{*}\right)\) and assigned weights proportional to the ratio \(p\left(x^{*} \mid x\right) / q\left(x^{*}\right)\) :
\[
\begin{equation*}
E\left[f\left(x^{*}\right) \mid x\right]=\int f\left(x^{*}\right) \frac{p\left(x^{*} \mid x\right)}{q\left(x^{*}\right)} q\left(x^{*}\right) d x^{*} \tag{4}
\end{equation*}
\]

Given a set of \(K\) samples \(\left\{x_{k}^{*}\right\}\) distributed according to \(q\left(x^{*}\right)\), this integral can be approximated by:
\[
\begin{equation*}
E\left[f\left(x^{*}\right) \mid x\right] \approx \frac{1}{K} \sum_{k=1}^{K} f\left(x_{k}^{*}\right) \frac{p\left(x_{k}^{*} \mid x\right)}{q\left(x_{k}^{*}\right)} \tag{5}
\end{equation*}
\]
with the approximation becoming more precise as \(K\) becomes larger.


Figure 1: An illustration of the basic read and write operations over SDMs. The outer dotted line represents the space of \(2^{N}\) possible addresses while the squares with labels \(\mathbf{A}_{\mathbf{m}}\) represent the \(M\) sampled hard addresses used for storage. The address being requested for operation is the \(\mathbf{x}\) in the center of the blue circle of radius \(D\). The hard addresses selected for operating correspond to the blue squares within the Hamming radius of \(\mathbf{x}\).

One possible choice for \(q\left(x^{*}\right)\) is the prior, \(p\left(x^{*}\right)\), which yields importance weights proportional to the likelihood, \(p\left(x \mid x^{*}\right)\). Formally,
\[
\begin{align*}
E\left[f\left(x^{*}\right) \mid x\right] & \approx \frac{1}{K} \sum_{k=1}^{K} f\left(x_{k}^{*}\right) \frac{p\left(x_{k}^{*} \mid x\right)}{p\left(x_{k}^{*}\right)} \\
& =\frac{1}{K} \sum_{k=1}^{K} f\left(x_{k}^{*}\right) \frac{p\left(x \mid x_{k}^{*}\right) p\left(x_{k}^{*}\right)}{p\left(x_{k}^{*}\right) p(x)} \\
& =\alpha(x) \sum_{k=1}^{K} f\left(x_{k}^{*}\right) p\left(x \mid x_{k}^{*}\right) \tag{6}
\end{align*}
\]
where we assume \(x_{k}^{*}\) is drawn from the prior, \(p\left(x^{*}\right)\), and \(\alpha(x)\) is a constant of proportionality that depends only on \(x\). Returning to the general case of data \(d\) and hypotheses \(h\), we can use the same approximation to compute the expectation of a function \(f(h)\) given observed data, with
\[
\begin{equation*}
E[f(h) \mid d] \approx \alpha(d) \sum_{k=1}^{K} f\left(h_{k}\right) p\left(d \mid h_{k}\right) \tag{7}
\end{equation*}
\]
where \(h_{k}\) is drawn from the prior \(p(h)\), and \(\alpha(d)\) is a constant of proportionality that depends only on \(d\).

\section*{Sparse distributed memory}

Sparse distributed memory (SDM) was developed as an algorithmic-level model of human memory, designed to encapsulate the notion that distances between concepts in memory correspond to distances between points in highdimensional space (Kanerva, 1988, 1993). In particular, it has a natural interpretation as an artificial neural network that uses distributed representations.

SDMs preserve distances between items in memory by storing them in a distributed manner. Assume that items enter memory as two strings of bits, with \(N\) bits indicating a
location and \(L\) bits indicating its content. A conventional approach would be to sequentially enumerate the set of \(2^{N}\) locations (more technically, addresses), storing items by setting the content bits at the appropriate address in turn. In contrast, a SDM samples \(M \gg N\) addresses \(\mathbf{a}_{j}\) to use as registers from the space of \(2^{N}\) possible addresses. Items are then stored by changing the bits that encode the content associated with multiple addresses, according to how close those addresses are to the target location. The algorithms for writing to and reading from a SDM are given below and are illustrated in Figure 1.

Writing A SDM stores an \(L\)-bit vector, z, associated with an \(N\)-bit location \(\mathbf{x}^{*}\) by storing the pattern at multiple addresses \(\mathbf{a}_{j}\) near \(\mathbf{x}^{*}\). Since the set of \(M\) available addresses does not enumerate the total space of \(2^{N}\), there may be very few addresses near \(\mathbf{x}^{*}\). Consequently, \(\mathbf{z}\) is written to all addresses \(\mathbf{a}_{j}\) that are within a Hamming distance \(D\) of \(\mathbf{x}^{*}\) (i.e. those \(\mathbf{a}_{j}\) that differ from \(\mathbf{x}^{*}\) in \(D\) or fewer bits). The contents of these selected addresses are modified to store the pattern \(\mathbf{z}\) such that each bit in the contents is increased or decreased by 1 depending on whether or not that bit in \(\mathbf{z}\) is a 1 or 0 , respectively.

A SDM can be constructed as a neural network with \(N\) units in the input layer, a hidden layer with \(M\) units for each sampled address, and an output layer with \(L\) units. The weights between the input and hidden layer correspond to the \(M \times N\) matrix \(\mathbf{A}=\left[\mathbf{a}_{1} ; \mathbf{a}_{2} ; \ldots ; \mathbf{a}_{M}\right]\) of hard addresses, and the weights between the hidden and output layer correspond to the \(M \times L\) matrix \(\mathbf{C}\) of contents stored at each address. The rule for writing \(\mathbf{z}\) to memory address \(\mathbf{x}^{*}\) is expressed as:
\[
\begin{align*}
\mathbf{y} & =\Theta_{D}\left(\mathbf{A} \mathbf{x}^{*}\right)  \tag{8}\\
\mathbf{C} & =\mathbf{C}+\mathbf{z y} \tag{9}
\end{align*}
\]
where \(\Theta_{D}\) is a function that converts its argument zeros and ones, with \(\Theta_{D}(w)=1\) if \(\frac{1}{2}(w-N) \leq D\) and 0 otherwise. \(\mathbf{y}\) is thus a selection vector that picks out a particular set of addresses. The expected number of addresses selected in \(\mathbf{y}\) is a function of \(N, M\), and \(D\) :
\[
\begin{equation*}
E\left[\mathbf{y}^{T} \mathbf{y}\right]=\frac{M}{2^{N}} \sum_{d=1}^{D}\binom{N}{d} \tag{10}
\end{equation*}
\]

Reading To read a pattern out of memory from address \(\mathbf{x}\), the SDM again computes a \(M\)-bit selection vector \(\mathbf{y}\) of addresses within Hamming distance \(D\) of \(\mathbf{x}\). The contents of each address selected by this vector are summed, resulting in a vector \(\widehat{\mathbf{z}}\) of length \(L\). The rule for reading \(\widehat{\mathbf{z}}\) from memory address \(\mathbf{x}\) is expressed as:
\[
\begin{align*}
& \mathbf{y}=\Theta_{D}(\mathbf{A} \mathbf{x})  \tag{11}\\
& \widehat{\mathbf{z}}=\mathbf{C}^{T} \mathbf{y} \tag{12}
\end{align*}
\]

The output \(\widehat{\mathbf{z}}\) can then be passed through a non-linearity to return a binary vector if desired.

\section*{SDMs as importance samplers}

Previous work formalizing a probabilistic interpretation of SDMs (Anderson, 1989) lays the groundwork for using

SDMs to perform Bayesian inference. Here, we show that the output of SDMs approximates the expectation of a function \(f\left(\mathbf{x}^{*}\right)\) over the posterior distribution \(p\left(\mathbf{x}^{*} \mid \mathbf{x}\right)\) by linking its behavior to that of the importance sampler in Equation 6.

Writing Let \(w\left(\mathbf{a}_{j}, \mathbf{x}^{*}\right)\) be the probability of writing to address \(\mathbf{a}_{j}\) given an input address \(\mathbf{x}^{*}\). In the standard SDM, this is 1 if the Hamming distance between \(\mathbf{a}_{j}\) and \(\mathbf{x}^{*}\) is less than or equal to \(D\) and 0 otherwise. In the limit, the number of addresses increases to the point where we will always be able to write to exactly \(\mathbf{x}^{*}\) (i.e., to set \(D=0\) ). Thus, this writing probability must satisfy the following constraint:
\[
\begin{equation*}
\lim _{M \rightarrow 2^{N}} w\left(\mathbf{a}_{j}, \mathbf{x}^{*}\right)=\prod_{i=1}^{N} \delta\left(x_{i}^{*}-a_{j i}\right) \tag{13}
\end{equation*}
\]

After writing \(K\) (address, data) pairs \(\left(\mathbf{x}_{k}^{*}, \mathbf{z}_{k}\right)\), the value of the counter associated with bit \(i\) of address \(\mathbf{a}_{j}\) will be:
\[
\begin{equation*}
\mathbf{c}_{j}=\sum_{k=1}^{K} w\left(\mathbf{a}_{j}, \mathbf{x}_{k}^{*}\right) \mathbf{z}_{k} \tag{14}
\end{equation*}
\]

Reading We are given a location \(\mathbf{x}\), which as before is a corrupted version of \(\mathbf{x}^{*}\). Let \(r\left(\mathbf{x}, \mathbf{a}_{j}\right)\) be the probability that we read from address \(\mathbf{a}_{j}\) given input \(\mathbf{x}\). In the standard SDM, this is 1 if the Hamming distance between \(\mathbf{a}_{j}\) and \(\mathbf{x}\) is less than or equal to \(D\) and 0 otherwise. Then, the output of the SDM for a particular set of addresses \(\mathbf{A}\) is:
\[
\begin{equation*}
\widehat{\mathbf{z}}=\sum_{j=1}^{M} \mathbf{c}_{j} r\left(\mathbf{x}, \mathbf{a}_{j}\right) \tag{15}
\end{equation*}
\]
where \(\mathbf{c}_{j}\) is defined in Equation 14.
To see how this output behaves for any SDM, we consider the expected value of \(\widehat{\mathbf{z}}\) over sampled sets of addresses \(\mathbf{A}\). We first substitute Equation 14 into Equation 15 and simplify according to linearity of expectation:
\[
\begin{align*}
E_{\mathbf{A}}[\widehat{\mathbf{z}} \mid \mathbf{x}] & =E_{\mathbf{A}}\left[\sum_{j=1}^{M}\left(\sum_{k=1}^{K} w\left(\mathbf{a}_{j}, \mathbf{x}_{k}^{*}\right) \mathbf{z}_{k}\right) r\left(\mathbf{x}, \mathbf{a}_{j}\right)\right]  \tag{16}\\
& =\sum_{k=1}^{K} \mathbf{z}_{k} \cdot E_{\mathbf{A}}\left[\sum_{j=1}^{M} w\left(\mathbf{a}_{j}, \mathbf{x}_{k}^{*}\right) r\left(\mathbf{x}, \mathbf{a}_{j}\right)\right] \tag{17}
\end{align*}
\]

As our address space grows larger (as in Equation 13), this approaches:
\[
\begin{equation*}
\lim _{M \rightarrow 2^{N}} E_{\mathbf{A}}[\widehat{\mathbf{z}} \mid \mathbf{x}]=\sum_{k=1}^{K} \mathbf{z}_{k} \int_{\mathbf{a}_{j}} \delta\left(\mathbf{x}_{k}^{*}-\mathbf{a}_{j}\right) r\left(\mathbf{x}, \mathbf{a}_{j}\right) \mathrm{d} \mathbf{a}_{j} \tag{18}
\end{equation*}
\]

Thus, in the limit, the expected value of \(\widehat{\mathbf{z}}\) read from the SDM will be:
\[
\begin{equation*}
E_{\mathbf{A}}[\widehat{\mathbf{z}} \mid \mathbf{x}]=\sum_{k=1}^{K} \mathbf{z}_{k} r\left(\mathbf{x}, \mathbf{x}_{k}^{*}\right) \tag{19}
\end{equation*}
\]

Comparing Equation 19 to Equation 6 yields our main result: SDMs can perform importance sampling by defining a
reading density proportional to a likelihood function, approximating the posterior expectation of the items stored in memory. More formally, the expectation of \(\widehat{\mathbf{z}}\) given \(\mathbf{x}\) is proportional to the output of the importance sampling approximation of the expectation of \(f\left(\mathbf{x}^{*}\right)\) with respect to \(p\left(\mathbf{x} \mid \mathbf{x}^{*}\right)\) :
\[
\begin{equation*}
E_{\mathbf{A}}[\widehat{\mathbf{z}} \mid \mathbf{x}] \propto \sum_{k=1}^{K} f\left(\mathbf{x}_{k}^{*}\right) p\left(\mathbf{x} \mid \mathbf{x}^{*}\right) \tag{20}
\end{equation*}
\]
provided \(\mathbf{z}_{k} \propto f\left(\mathbf{x}_{k}^{*}\right)\) and \(r\left(\mathbf{x}, \mathbf{x}_{k}^{*}\right) \propto p\left(\mathbf{x} \mid \mathbf{x}^{*}\right)\).
The utility of this result is limited with the standard formulation of the SDM, as it only holds in the limit where the address size becomes large and \(D\) becomes small, meaning that \(r\left(\mathbf{x}, \mathbf{x}^{*}\right)\) becomes a delta function. Instead, we can consider generalizations in which we use different values of \(D\) for reading and writing ( \(D_{r}\) and \(D_{w}\), respectively), or where we choose \(r\left(\mathbf{x}, \mathbf{x}^{*}\right)\) more freely. These modifications allow us to approximate a variety of Bayesian models using SDMs.

We make a further note, which is that in most practical applications, the address space will not approach the limit (i.e. \(M \ll 2^{N}\) ). For any sampled set of addresses, we would still expect the value of \(\widehat{\mathbf{z}}\) to be near the posterior mean \(E_{k}\left[f\left(\mathbf{x}^{*}\right) \mid \mathbf{x}\right]\), but we cannot make any statement about how close. We leave it as an area for future work to place analytic bounds on the accuracy of the SDM's approximation. For the case studies we present here, we estimate the variance of the SDM using Monte Carlo approximations.

In the following two sections, we evaluate SDMs as a scheme for approximating Bayesian inference in two tasks: one where the Bayesian likelihood matches the standard SDM read rule, and one where the SDM read rule is adjusted to match the Bayesian likelihood. The second case further illustrates how SDMs can be applied to more general problems of Bayesian inference that go beyond simply removing noise from a stimulus.

\section*{Letter reconstruction}

As a simple illustration of approximating Bayesian inference with a SDM, we consider the task of recovering images of English letters, \(\mathbf{x}^{*}\), from noisy observations \(\mathbf{x}\). To solve this problem, we set up a Bayesian model loosely based on that presented in Rumelhart and Siple (1974). Each letter of the alphabet is encoded as a binary feature vector of length \(N=\) 14 based on the Rumelhart-Siple font template (Figure 2).

\section*{SDM approximation}

In the Bayesian model, we wish to reconstruct the original letters \(\mathbf{x}^{*}\) from the exemplars \(\mathbf{x}\) by computing the mean of the posterior distribution over original letters, \(p\left(\mathbf{x}^{*} \mid \mathbf{x}\right)\), i.e. \(f\left(\mathbf{x}^{*}\right)=\mathbf{x}^{*}\). Each of the original letters is associated with a prior probability, \(p\left(\mathbf{x}^{*}\right)\), set to be the relative letter frequency of English text (Lewand, 2000). Following the generative model, a noisy image \(\mathbf{x}\) is produced from the original letter \(\mathbf{x}^{*}\) via a noise process in which at most \(B\) bits are flipped (uniformly). Given a noisy image vector \(\mathbf{x}\), we define the likelihood \(p\left(\mathbf{x} \mid \mathbf{x}^{*}\right)\) to be uniformly distributed over the number of possible bit strings in a hypersphere of radius \(D_{r}\).


Figure 2: The Rumelhart-Siple font feature map and the Rumelhart-Siple representation for the letter " \(A\) " along with its binary feature pattern.

In the SDM, we sample exemplars \(\mathbf{x}^{*}\) of the original letters from the prior \(p\left(\mathbf{x}^{*}\right)\) and write them to the SDM at inputs \(\mathbf{x}^{*}\) (i.e., \(\mathbf{z}=\mathbf{x}^{*}\) ). The likelihood defined for the Bayesian model is naturally compatible with the standard read rule of a SDM, where we only consider hypotheses \(\mathbf{x}^{*}\) which are within Hamming distance \(D_{r}\) of \(\mathbf{x}\). Thus, we can set the read function \(r\left(\mathbf{x}^{*}, \mathbf{x}\right)\) of the SDM to be a variation of this likelihood:
\[
p\left(\mathbf{x} \mid \mathbf{x}^{*}\right) \approx r\left(\mathbf{x}, \mathbf{x}^{*}\right)= \begin{cases}{\left[\sum_{d=1}^{D_{r}}\binom{N-d+1}{d}\right]^{-1}} & \left|\mathbf{x}-\mathbf{x}^{*}\right| \leq D_{r} \\ 0 & \text { otherwise }\end{cases}
\]

The corrupted images \(\mathbf{x}\) are the inputs that we attempt to read from and \(D_{r}\) is the SDM's read radius; we additionally define \(D_{w}\) to be the write radius. The intuition is that we write the original letters \(\mathbf{z}=\mathbf{x}^{*}\) to input \(\mathbf{x}^{*}\), and read from \(\mathbf{x}\) outputs \(\widehat{\mathbf{z}}\) which are similar to the mean of the Bayesian posterior.

\section*{Analysis}

To evaluate how well the SDM approximates the posterior mean, we sampled 1000 exemplars from the prior distribution \(p\left(\mathbf{x}^{*}\right)\). We then created three images \(\mathbf{x}\) for each letter in the alphabet, each with two bits of corruption, yielding a total of 72 test images. We repeated this simulation - sampling from the prior and generating test images - 20 times for each SDM with parameters \(D_{r}, D_{w}\), and \(M\), where each simulation used a different set of sampled addresses \(A\).

We determined the appropriate settings of the read and write radii, \(D_{r}\) and \(D_{w}\), by considering the constraints imposed by the SDM and the specific problem of letter reconstruction. The mean Hamming distance between pairs of letters was 5.4615 , indicating that the read radius must lie somewhere between 2 (the amount of corruption) and 5. Choosing a radius outside these bounds would have the effect of returning an inaccurate expectation, either because the noise was greater than the signal \(\left(D_{r}<2\right)\) or because too many hypotheses were considered \(\left(D_{r}>5\right)\). So, we chose \(D_{r}=2\), thus ensuring that the true \(\mathbf{x}^{*}\) would always fall within this radius, and also minimizing the number of incorrect hypotheses considered. For the write radius, we considered three different values, \(D_{w}=\{0,1,2\}\).

We stored all 1000 letters in each SDM, varying the number of hard addresses, \(M\), among eight evenly spaced values


Figure 3: The average correlations between the Bayesian posterior mean and pre-thresholded SDM outputs for the task of recovering a Rumelhart-Siple letter from a noisy observation. \(D_{r}=2\) for each of the 4 values of \(D_{w}\) presented.
between 2048 and \(2^{N}=16384\). We then read the value \(\widehat{\mathbf{z}}\) for each corrupted input \(\mathbf{x}\), for different address spaces \(A\). For the Bayesian model, we analytically calculated the posterior mean by evaluating the full posterior for each \(\mathbf{x}^{*}\) and then calculating the average \(\mathbf{x}^{*}\) weighed by its posterior probability. We then computed the average correlation between the SDM values of \(\widehat{\mathbf{z}}\) and the Bayesian posterior mean across sampled addresses \(A\). These results are displayed in Figure 3.

The SDMs with \(D_{w}=0\) performed similarly to the Bayesian model only when \(M=2^{N} \quad(\rho=0.9628\), se \(=\) 0.0058 ), reflecting the intuition that it's highly unlikely to find an exact address from a random sample of \(2^{N}\). Conversely, the SDMs with \(D_{w}=2\) and \(D_{w}=3\) had near-constant correlations with the Bayesian model ( \(\rho \approx 0.88\) for \(D_{w}=2\) and \(\rho \approx 0.79\) for \(D_{w}=3\) ), regardless of the size of the hard address space. This behavior was also in line with our expectations: with \(D_{w}=2\) and \(D_{w}=3\), the probability of having no hard addresses within \(D_{w}\) of \(\mathbf{x}^{*}\) is extremely low for \(M=2048\); this probability only decreases as \(M\) increases.

In summary, these results show SDMs can naturally approximate Bayesian models of noisy stimulus reconstruction. Here, we set the likelihood function to follow the standard SDM read rule. In the next section we consider a more general example of Bayesian inference and explore the consequences of adjusting the SDM read rule to match the likelihood in question.

\section*{Property induction}

If you are told that a horse has a particular property, protein \(K\), what other animals do you think share this protein? Questions of this type are considered problems of property induction, where one or more categories in a domain are observed to have some novel property and the task is to determine how the property is distributed over the domain. For our analyses we explore the category-based induction task introduced by

Osherson, Smith, Wilkie, Lopez, and Shafir (1990), where judgments are made as to whether certain animals can get a disease knowing other animals that can catch it.

\section*{SDM approximation}

We solve this problem with a Bayesian model of property induction based on Kemp and Tenenbaum (2009). Here, we observe a set of examples \(d\) of a concept \(C\) (known to have property \(K\) ) and we aim to calculate \(p(y \in C \mid d)\), the probability that object \(y\) is also a member of \(C\). Thus, averaging over all possible concepts, we compute:
\[
\begin{equation*}
p(y \in C \mid d)=\sum_{h \in \mathcal{H}} p(y \in C \mid h) p(h \mid d) \tag{21}
\end{equation*}
\]
where the hypothesis space \(\mathcal{H}\) is the set of all possible concepts in a domain, \(p(y \in C \mid h)=1\) if \(y\) is in hypothesis \(h\), and 0 otherwise. The posterior probability \(p(h \mid d)\) can be computed via Bayes' rule from Equation 1.

We explore two variations of likelihood functions based on assumptions of how the data were generated. If the data are generated uniformly at random, the likelihood follows from weak sampling: \(p(d \mid h)=1\) if all examples in \(d\) are in \(h\) and \(p(d \mid h)=0\) otherwise. If the data are generated at random from the true concept \(C\), the likelihood follows from strong sampling, where \(p(d \mid h)=1 /|h|^{n}\) if all \(n\) examples in \(d\) belong to \(h\), and \(p(d \mid h)=0\) otherwise. This likelihood function incorporates the size principle, where hypotheses with fewer items are given more weight than hypotheses with more items for the same set of data (Tenenbaum \& Griffiths, 2001).

In the SDM, the location \(\mathbf{x}^{*}\) corresponds to a hypothesis \(h\), and the content \(\mathbf{z}\) corresponds to data \(d\). For both weak and strong sampling assumptions, we set the read function of the SDM, \(r(d, h)\), to be proportional to the Bayesian likelihood by weighting the selection vector \(\mathbf{y}\) by \(p(d \mid h)\).

\section*{Analysis}

To evaluate the performance of this modified SDM, we used the category-based induction dataset from Osherson et al. (1990), consisting of 36 two-premise arguments and 45 threepremise arguments ( e.g., "Cat, Dog, Horse can get disease X ") for a domain of 10 animals. Thus, each observation \(d\) is encoded as a binary feature vector of length \(N=10\). We used a taxonomic hypothesis space following Kemp and Tenenbaum (2009), constructed from human similarity judgments for each possible pairing of animals. As before, we stored 1000 exemplars sampled from \(p(h)\) in the SDM, and varied the number of hard addresses among eight evenly spaced values between 128 and \(2^{N}=1024\). We evaluated this modified SDM against the Bayesian model of property induction for 3 values of \(D_{w}\) over a constant reading radius \(D_{r}=0\). The results are presented in Figure \(4^{1}\).

We find the SDM approximates Bayesian inference for a variety of write radii and, as predicted, matches Bayes when

\footnotetext{
\({ }^{1}\) We note that these correlations are not strictly monotonic as one might intuit, due to sampling error in the address space (most notable in the case when \(D_{w}=0\).)
}


Figure 4: The average correlations between the Bayesian posterior and pre-thresholded SDM outputs assuming weak sampling (left panel), and assuming strong sampling (right panel) for the property induction task. \(D_{r}=0\) for each of the 3 values of \(D_{w}\) presented.
\(D_{w}=0\) and \(M=2^{N}\). Furthermore, the SDMs that use a weak or strong sampling read rule correlate equally well with the Bayesian models that use weak or strong sampling likelihoods, respectively. This nicely illustrates the correspondence between the SDM's read rule and a Bayesian likelihood, and implies that SDMs can approximate a broad range of Bayesian models by adjusting the read rule to match the likelihood function.

\section*{Conclusion}

What constraints do the algorithmic and implementation levels impose on probabilistic models of cognition? We explored this question by considering whether the computations employed by a distributed representation of associative memory could approximate Bayesian inference. By choosing the SDM read rule to appropriately match the the likelihood function, we showed in two separate scenarios that SDMs can accurately implement a specific form of Bayesian inference called importance sampling.

Future work will take our analyses one step further and investigate whether SDMs can approximate Bayesian inference without modifying their read rule. Specifically, can we encode the data in a manner such that reading it from a standard SDM is still proportional to the likelihood? If so, it would make SDMs an even more general and appealing approach to performing Bayesian inference. Regardless, the results presented in this paper are an important step towards an explanation of how the structured representations assumed by probabilistic models of cognition could be expressed in the distributed, continuous representations commonly used in algorithmic-level models such as neural networks.

Acknowledgments. This work was supported by grant number FA-9550-10-1-0232 from the Air Force Office of Scientific Research and a Berkeley Fellowship awarded to JBH.

\section*{References}

Anderson, C. H. (1989). A conditional probability interpretation of Kanerva's sparse distributed memory. In International joint conference on neural networks (pp. 415-417).
Gigerenzer, G., \& Todd, P. (1999). Simple heuristics that make us smart. Oxford University Press, USA.
Griffiths, T., Chater, N., Kemp, C., Perfors, A., \& Tenenbaum, J. (2010). Probabilistic models of cognition: exploring representations and inductive biases. Trends in Cognitive Sciences, 14(8), 357-364.
Griffiths, T., Vul, E., \& Sanborn, A. (2012). Bridging levels of analysis for probabilistic models of cognition. Current Directions in Psychological Science, 21(4), 263-268.
Kanerva, P. (1988). Sparse Distributed Memory Systems. MIT Press.
Kanerva, P. (1993). Sparse Distributed Memory and Related Models. In M. Hassoun (Ed.), Associative neural memories: Theory and implementation (pp. 50-76). Oxford University Press.
Kemp, C., \& Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. Psychological Review, 116(1), 20-58.
Lewand, R. (2000). Cryptological mathematics. Mathematical Association of America.
Marr, D. (1982). Vision. San Francisco, CA: W. H. Freeman.
McClelland, J., Botvinick, M., Noelle, D., Plaut, D., Rogers, T., Seidenberg, M., et al. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cognition. Trends in Cognitive Sciences, 14(8), 348-356.
Neal, R. (1993). Probabilistic inference using Markov chain Monte Carlo methods (Tech. Rep. No. CRG-TR-93-1). Department of Computer Science, University of Toronto.
Osherson, D., Smith, E., Wilkie, O., Lopez, A., \& Shafir, E. (1990). Category-based induction. Psychological Review, 97(2), 185.
Rumelhart, D., \& Siple, P. (1974). Process of recognizing tachistoscopically presented words. Psychological Review, 81(2), 99.
Shi, L., \& Griffiths, T. (2009). Neural implementation of hierarchical Bayesian inference by importance sampling. Advances in Neural Information Processing Systems, 22, 1669-1677.
Shi, L., Griffiths, T. L., Feldman, N. H., \& Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. Psychonomic Bulletin \& Review, 17(4), 443-64.
Tenenbaum, J., \& Griffiths, T. (2001). Generalization, similarity, and Bayesian inference. Behavioral and Brain Sciences, 24(4), 629-640.
Tenenbaum, J., Kemp, C., Griffiths, T., \& Goodman, N. (2011). How to grow a mind: Statistics, structure, and abstraction. Science, 331(6022), 1279-1285.

\title{
The Effect of Physical Load on the Cognitive Process of Estimation
}

\author{
Keiga Abe (keiga.abe@gmail.com) \\ Department of Education, 1-1 Takakuwa-Nishi, Yanaizu-cho \\ Gifu, Japan
}

\begin{abstract}
The purpose of this study was to examine the process of embodied cognition in distance estimation. According to recent cognitive science studies, our intelligent behavior that ranges from perception to inference is not accomplished in only a closed mental process, but is affected by body and action. However, previous studies do not clarify whether these effects were derived from physical load or subjective heaviness. In order to examine the question, two experiments were conducted using the "size-weight illusion". Performance on the distance estimation task was not affected by subjective heaviness but by physical load.
\end{abstract}

Keywords: embodied cognition; size-weight illusion; distance estimation

\section*{Introduction}

We examined the contribution of the physical body on higher-order cognitive processing. Recently, in cognitive science, studies have reported that a wide range of intellectual behavior, from perception to inference, is not only a closed mental process but is also subject to influences of the physical body and its actions/motions(Wilson, 2002; Gibbs, 2005; Proffitt, 2006). Since physical loading is known to exert effects on mental processes, Narukawa, et al. (2010) reported changes in gustatory sensation that accompany the degree of fatigue. Krishna \& Morrin (2008) showed that the sense of hardness of the bottle affected the evaluation of mineral water. Bhalla \& Profitt (1999) demonstrated experimentally that different estimates are made of the inclination of a sloped path under the conditions of carrying a load on the back versus being empty-handed. In addition, in the study by Ackerman, Nocera \& Bargh (2010), the curriculum vitae of a fictitious person bound to two types of clipboards that differed in heaviness were handed to the subjects, who were asked to make evaluations of the person. The evaluations made by those of the group handed the heavier clipboard were higher than that of the group handed the lighter clipboard. The results of these prior studies suggest that mental processes are influenced by loading and fatigue of the physical body of the subject. However, it has not been clarified whether these effects were due to the amount of actual physical load or due to the amount of the subjective load. In this study, this issue was examined using a distance estimation task adopted from a prior study. If the effects were due to the amount of the physical load, then physical/non-overt processes, which are separate from the subjective view of the subject, are
expected exert an effect on the inference. Conversely, if they are due to the amount of the subjective load, it may be considered that the subjective view of the subject and overt processes exert the effects on the inference.

To examine these physical and subjective loads separately, the "size-weight illusion" (Charpentier, 1891) was used in this study. This illusion occurs when if the weights of two objects are the same, the larger object is sensed as being lighter. Utilizing this illusion, distance estimation tasks under conditions of being subject to different subjective loads while being subject to the same physical load (Experiment 1) and distance estimation tasks under conditions of being subject to different physical loads while being subject to the same subjective load (Experiment 2) were conducted to examine the effect of the physical and subjective amount of the physical load.

\section*{Experiment 1}

In Experiment 1, experimental manipulations were conducted to generate the subjective view that loads with different weights were being exerted while the same weight physically was exerted, and distance estimation tasks were conducted under conditions of a divergence between the amount of subjective and physical load. This was used to examine how the perceived load of the weight exerted on the body is processed.

\section*{Method}

Subjects Ninety-two college students participated in the experiment. Of them, 24 were assigned to the 10 L group, in which the subject held a 5 kg tank with capacity of 10 L as the number of steps of a stairway was estimated; 33 were assigned to the 20 L group, in which the subject held a 5 kg tank with a capacity of 20 L as the estimation was made; and 35 were assigned to the control group, in which they made the estimation without holding any weight. A singlefactor between-subjects design was used in this experiment.

Task A revised form of the distance estimation task published by Bhalla \& Profitt (1999) was employed. In the revised form, a picture of the up-bound steps of the Atago Shrine (Fig. 1) was presented for 5 s , and the subject was


Figure 1. The up-bound steps of the Atago Shrine
instructed to estimate the number of the steps. The picture was displayed on a 17-in XGA display placed at a height of 160 cm . With regard to the physical loads to be exerted on the subjects, we prepared a reference weight with 5 kg of water in a polyethylene tank with a capacity of 5 L , and the weights for those assigned to the groups other than the control group ( 5 kg with a size of 10 L , or 5 kg with a size of 20 L ).

Procedure For each group, the subject was first handed the reference weight and was told that its weight was 5 kg . Next, the weight assigned to each group for the distance estimation task was given, and the task of estimating the number of steps was performed with subject holding the reference weight with both hands. The up-bound steps of the Atago Shrine were present to the subject for 5 s ; then the estimation was given orally while holding the weight. The subjects were told to respond immediately without thinking deeply when providing their oral response. Following this, they were asked to estimate the weight of the tanks.

\section*{Results}

First, the values of the weight of the tanks used for the estimation task predicted by the 10 L group and by the 20 L group are discussed. The value was 5.94 kg for the 10 L group and 4.06 kg for the 20 L group. A significant difference was found between the two groups \((t(26)=2.74\), \(p<0.05, r=0.470\), Fig. 2). This result confirmed that a size-weight illusion effect had occurred for members of the 10 L and 20 L groups. he estimates of the distance of the steps was 52.79 steps for the 10 L group and 49.88 steps for the 20 L group.


Figure 2: The values of the weight of the weights used in Experiment 1


Figure 3: The estimated distance of the steps in Experiment 1

A one-way ANOVA revealed a main effect for the weight factor \(\left(F(2,89)=14.82, p<0.01, \eta^{2}=0.25\right.\), Fig. 3). Multiple comparisons (Bonferroni's method) revealed significant differences between the control group and the other two groups ( \(p<0.01\) ) but not between the 10 L and the 20 L groups.

\section*{Experiment 2}

The results of Experiment 1 suggested that a difference in perceived weight did not affect the estimate of the number of steps but instead the physical loading affected the distance estimations. However, the conclusion that the subjective amount did not affect the estimates, runs contrary to the finding that no significant difference was observed between the 10 L and 20 L groups. Therefore, in Experiment 2, the self-adjustment of the amount of loading by the subject was performed using the point of subjective equality(PSE) measurement procedure, and comparisons were made for cases in which physically different loads were exerted whereas the subjective load was the same.


Figure 3: The values of the weight of the tanks used in Experiment 2


Figure 4: The subjective weight of the tanks in Experiment 2

As in Experiment 1, if the effects on the estimates were due to the physical load and not due to the subjective load, significant differences should be observed between the estimates.

As, in Experiment 1, the subjects were asked to estimate the number of steps, but it was difficult to predict whether the number is over-estimated or under-estimated by the physical loading in comparison to the previous studies since the correlations between the number of steps and the distances and inclinations reported by those studies cannot be guaranteed. Thus, in Experiment 2, they were instructed to estimate the distance, and not the number, of the steps.

\section*{Method}

Subjects Twenty-seven college students participated in the experiment. Of them, 14 were assigned to the experimental group and 13 were assigned to the control group. A singlefactor between-subjects design was used for the experiment.


Figure 5: The estimated distance of the steps in Experiment 2

Task The same distance estimation task was employed as in Experiment 1. With regard to the physical loads to be exerted on the subjects, after having the reference stimulus of a polyethylene tank with a capacity of 5 L containing 5 kg of water presented to them, they were asked to, by themselves, adjust the amount of physical loading by using the PSE measurement procedure. In accord with this procedure, the subjects put water into a polyethylene tank with a capacity of 20 L for the stimulus weight until they thought it to be identical in weight as that of the reference stimulus. The average weight of the stimulus weights set by the subjects was 6.36 kg , which was more than 1 kg heavier than the reference weight (Fig. 3). The subjects were not informed that the reference weight was 5 kg , and were only aware that the adjusted weights have the same weight as the reference weight.

Procedure In each group, the subject was first asked to perform the adjustment of the weight in accordance with the PSE procedure. Subsequently, tasks of estimating the length of the steps were performed with the self-adjusted weight, in the case of the experimental group, and with the reference weight, in the case of the control group, held in both hands. As in Experiment 1, the up-bound steps of the Atago Shrine were presented to the subject for 5 s . Next, each of the subjects provided their estimates orally while still holding the weight. Following this, the weight used for the distance estimation task was also estimated.

\section*{Results}

First, the predicted values of the weight of the reference weight are discussed. The mean value was 4.11 kg for the control group and 4.25 kg for the experimental group. No significant difference was found between the two groups \((t(25)=0.26, p=0.79\), n.s., \(r=0.05\), Fig. 4).
Based on this result, the possibility that the sense of weight was significantly different between those in the control group and those in the experimental group was rejected.

The estimate of the distance of the steps was 31.79 m for the control group and 70.77 m for the experimental group. A significant difference was found between the two groups \((t(25)=2.65, p<0.05, r=0.47\), Fig. 5).

\section*{Discussion}

In the present study, the result of experiments using distance estimation tasks accompanied by physical loading using the size-weight illusion showed that the effects of physical loading on mentation are due to physical load rather than subjective load. The subjects in the control group in Experiment 2 estimated the length of the steps while holding a 5 kg weight that they thought weighted 4 kg , and the subjects in the experimental group while holding a 6 kg weight that they thought was 4 kg . The difference between their length estimates of the two groups suggests that their estimations were being influenced, not by how heavy they thought the weights were, but rather by the actual physical load exerted on the body. This also suggests that a load that is exerted on the body may play an implicit role in making inferences and judgments.

Unlike previous studies, a picture of stairs was used in the present study instead of an actual environment. Nevertheless, physical loading affected the participants' distance estimation. This shows robustness of the previous studies, and suggests that participants mentally simulate action with reference to physical load.

\section*{Future Issues}

Although physical loading exerted effects in a manner that did not reach the subjective level, it is not possible to conclude that the amount of subjective load did not exert any effect at all. Further, many aspects of the process by which physical loading influenced the estimation have yet to be elucidated. It is necessary to consider the mutual relationship of and processing between the subjective load and the physical load.

In the present study, the estimates were made when the individuals were subjected to a physical load, but it is not known to what extent the effects were sustained. As indicated by previous studies, fatigue influences inference. The degree of fatigue and the extent of recovery are expected to also influence the duration for which the effects are sustained.

In addition, the duration of the presentation of the object to be estimated warrants some discussion. In this study, the stimulus was presented for a limited duration of 5 s . The subjective level could become dominant in processing when the object of the estimation is presented for longer duration. Moreover, many issues remain with regard to the mutual relationship between the top-down processing of the subjective view and the bottom-up processing from the physical body.

In the literature of the cognitive process of metaphor, it is thought that metaphors enable us to think about concepts on the basis of concrete sensorimotor experiences. Previous
studies suggested that a conceptual representation was linked to some somatic and physical state by some metaphorical concept (Clark, 1973; Lakoff \& Johnson, 1980). Researchers who investigated the haptic priming effect found evidence to support that idea (Williams \& Bargh, 2008; Ackerman, Nocera \& Bargh, 2010). In future studies, the effects of physical load on conceptual representation should be examined.

\section*{References}

Ackerman, J. M., Nocera, C.C., \& Bargh, J. A. (2010). Incidental Haptic Sensations Influence Social Judgements and Decisions. Science, 25, 1712-1715.
Bhalla, M., \& Proffitt, D. R. (1999). Visual-Motor recalibration in geographical slant perception. Journal of Experimental Psychology: Human Perception \& Performance. 25(4), 1076-1096.
Charpentier A. (1891). Analyse experimentale: De quelques elements de la sensation de poids. Arch Physiol Norm Pathol, 3, 122-135.
Clark, H. (1973). Space, time, semantics, and the child. In T. Moore(Ed.), Cognitive development and the acquisition of language(pp. 27-63). New York, NY:Academic Press.
Gibbs, R.W. Jr.(2005). Embodiment and Cognitive Science. New York: Cambridge University Press.
Krishna, A. and Morrin, M. (2008), Does touch affect taste? The perceptual transfer of product container haptic cues, Journal of Consumer Research, 34, 807-818.
Lakoff, G. \& Johnson, M. (1980), Metaphors we live by. Chicago: University of Chicago Press.
Narukawa, M., Ue, H., Uemura, M., Morita, K., Kuga, S., Isaka, T., and Hayashi, Y. (2010). Influence of prolonged exercise on sweet taste perception. Food Science and Technology Research, 16(5), 513-516.
Proffitt, D. R. (2006). Embodied perception and the economy of action. Perspectives on Psychological Science, 1(2), 110-122.
Williams L. E., \& Bargh, J. A. (2008). Keeping one's distance: The influence of spatial distance cues on affect and evaluation. Psychological Science, 19, 302-308.
Wilson, M. (2002) Six views of embodied cognition. Psychonomic Bulletin \& Review, 9(4), 625-636.

\title{
Automatic Labeling of Phonesthemic Senses
}

\author{
Ekaterina Abramova (e.abramova@ftr.ru.nl) \\ Department of Philosophy, Radboud University Nijmegen \\ Raquel Fernández (raquel.fernandez@uva.nl) \\ Institute for Logic, Language \& Computation, University of Amsterdam \\ Federico Sangati (federico.sangati@gmail.com) \\ Institute for Logic, Language \& Computation, University of Amsterdam
}

\begin{abstract}
This study attempts to advance corpus-based exploration of sound iconicity, i.e. the existence of a non-arbitrary relationship between forms and meanings in language. We examine a number of phonesthemes, phonetic groupings proposed to be meaningful in the literature, with the aim of developing ways to validate their existence and their semantic content. Our first experiment is a replication of Otis and Sagi (2008), who showed that sets of words containing phonesthemes are more semantically related to each other than sets of random words. We augment their results using the British National Corpus and the Semantic Vectors package for building a distributional semantic model. Our second experiment shows how the semantic content of at least some phonesthemes can be identified automatically using WordNet, thereby further reducing the room for intuitive judgments in this controversial field.
\end{abstract}

Keywords: Iconicity; Phonesthemes; Corpus analysis; Distributional Semantics; WordNet.

\section*{Introduction}

The claim that the relationship between forms and meanings in language is not always arbitrary is controversial. However, evidence for non-arbitrary relationships comes at multiple levels of language, from phonology to syntax (Perniss, Thompson, \& Vigliocco, 2010). Here we focus on the phonetic level and investigate the association of particular sounds with aspects of word meaning. Such sound iconicity has been described in a variety of non-Indo-European languages (see the studies in Hinton, Nichols and Ohala Hinton, Nichols, \& Ohala, 2006b) and its existence in English suggested by a number of authors (Firth, 1930; Marchand, 1969), and exploited in commercial settings (Shrum \& Lowrey, 2007).

Phonesthemes (a technical term for meaningful sound patterns) are sub-morphemic units that play a role of morphemes but have been traditionally distinguished from them by being non-compositional (but see Rhodes Rhodes, 2006 for an opposite view). The most oft-cited example is the English phonestheme \(g l\) which occurs in a large number of words related to light or vision (glitter, glisten, glow, gleam, glare, glint, etc.). Once the phonestheme is taken out, the remainder of the word is not a morpheme (-itter, -isten, -ow etc.) and one does not attach \(g l\) to other words to make them light-related. Still, the extent and the nature of this phenomenon is not clear.

Traditionally, the evidence for the existence of phonesthemes and their proposed meaning consisted in listing a number of words that share a given sound and attempting to find the semantic core that unites them. Popular expla-
nations for the phenomenon would rest on the intuited association between sound production and meaning. For example, Reid (1967) states that "The explosive nature of the letter \(b\) is intensified when it is combined with \(l\) before the breath is released. Consequently words beginning with \(b l\) are found generally to indicate a 'bursting-out' or the resultant swelling or expansion" (p. 10). More recent accounts view them rather as a matter of statistical clustering. According to such "snowballing effect" theory, a group of phonemes in related words (for example, by common etymology) becomes over time associated with the meaning of these words and given the right conditions starts to attract other words with the same phoneme into a cluster, through semantic change or influencing the creation of new words (Blust, 2003; Hinton, Nichols, \& Ohala, 2006a). \({ }^{1}\)

Dissociating these competing explanations would require a combination of historical and cross-linguistic research but, arguably, there is a wealth of more basic questions that need to be addressed first. The nature of iconicity is such that it is easy to see the connection between form and meaning once we are aware of both elements but such intuition is not always a reliable guide for discovering the connections. Just as it is difficult to interpret an iconic sign from American Sign Language when its meaning is unknown (Bellugi \& Klima, 1976), we might miss the connection that is in fact present. On the other hand, we might over-estimate the connection by listing only the light-related \(g l\)-words and forgetting the amount of \(g l\)-words that have nothing to do with light (glide, glucose, globe, glove, etc.). In other words, if we want to validate the existence of phonesthemes or explain their origin, we need to apply more falsifiable and unbiased methods in all stages of investigation: identifying them in a given language, quantifying their scope and establishing their meaning.

So far, the reality of phonesthemes has been demonstrated in behavioral experiments (Bergen, 2004; Hutchins, 1998) and corpus studies (Drellishak, 2006; Otis \& Sagi, 2008). Our aim is to contribute to the second current of this research. We believe that this is a valuable way of objectively addressing large-scale linguistic phenomena that can refine our understanding of sound iconicity and lead to further testable hy-

\footnotetext{
\({ }^{1}\) This is not to say that there are no universal sound features underlying certain cases of sound iconicity, such as words for small and large objects usually associated with high and low acoustic frequency respectively (Ohala, 1994).
}
potheses with respect to its cognitive underpinnings.
Otis and Sagi (2008) conducted the first corpus-based analysis of phonesthemes. They examined 47 groups of words containing phonesthemes using Project Gutenberg texts and a method for calculating word similarity based on Latent Semantic Analysis (LSA), in particular its Infomap \({ }^{2}\) variety (Schütze, 1997). The analysis performed by Otis and Sagi showed that semantic relatedness of clusters of words that share a phonestheme is higher than that of clusters composed of randomly chosen words. This method, therefore, can be used to examine the validity of conjectured phonesthemes. However, as the authors admit, it "does not identify what specific semantic content is carried by the identified phonestheme" (p. 68). Our first aim is replicating the study of Otis and Sagi using (1) a more recent and balanced corpus - the British National Corpus (BNC), and (2) a newer and more versatile and efficient tool for calculating semantic relatedness, Semantic Vectors \({ }^{3}\) (Widdows \& Cohen, 2010).

Our second aim is attempting to develop a method for automatically identifying the semantic content associated with a particular phonestheme-a task that, to our knowledge, has not previously been addressed in the literature. Otis and Sagi (2008) suggest that methods designed to identify the topic of a given text could be used to that end. We think, however, that a more straightforward method lies in analogy with the task of unsupervised ontology acquisition: placing a word within a hierarchy of concepts based on its semantic relationship with the rest of the words in the hierarchy: for example, pear being placed close to apple and banana under fruit. In the case of phonesthemes, it is conceivable that a group of \(g l\) words would be assigned a vision-related higher class. Whether this can be done automatically and applied to a variety of phonesthemes is one of the questions we pose in this study.

In sum, our hypotheses are the following:
Hypothesis 1: Words that share a phonestheme are on average more semantically related than random words.

Hypothesis 2: The core semantic import conjectured in the literature for a phonestheme can be derived automatically from a set of phonestheme-bearing words.

\section*{Experiment 1: Semantic Relatedness}

\section*{Methods}

To explore our Hypothesis 1, we used the British National Corpus (Burnard, 2007), a 100 million word collection of written and spoken English language compiled from a wide variety of sources and genres. We pre-processed the entire corpus using the Natural Language Toolkit (NLTK) (Bird, Klein, \& Loper, 2009). In particular, we used NLTK to extract the content words in the corpus (nouns, adjectives, and lexical verbs) and to lemmatize them, i.e. to reduce a family of inflected words (such as walk, walks, walked, walking) to a single word type or lemma (e.g. walk). This resulted in a subcorpus of about 43 million words, which we used as input to

\footnotetext{
\({ }^{2}\) Freely available at http://infomap-nlp.sourceforge.net.
\({ }^{3}\) Freely available at http://code.google.com/p/semanticvectors/.
}
construct a distributional semantic model with the Semantic Vectors package (Widdows \& Cohen, 2010).

Semantic Vectors allows us to use a corpus to build a high-dimensional vector space where words are represented as vectors that record their frequency of co-occurrence with other words or other documents in the corpus. We can then use well-defined methods to measure how similar the meanings of two words are, such as computing the cosine of the angle formed by their corresponding vectors. As Otis and Sagi (2008) indicated in their pioneering corpus study, this methodology can be of great value to investigate the claims behind phonesthemes in an objective, data-driven way, since we can use the distributional model to test whether words sharing a hypothesized phonestheme exhibit higher semantic similarity than random words.

Like Otis and Sagi, we built a term-term model where each word vector records the co-occurrence of that word with other words in the context (rather than recording occurrence in particular documents like LSA), but unlike them, who used the traditional singular value decomposition method for reducing the dimensions in the matrix, we used Random Projection, a more computationally efficient algorithm. \({ }^{4}\) We experimented with the settings of two parameters in the Semantic Vectors package: the minimum frequency of the word types considered for building the model (as we may not be able to construct reliable distributional semantic representations for low frequency words) and the window size, i.e. the context window of \(n\) words to left and right of the target word where the model looks for co-occurrences of other words. McDonald and Ramscar (2001) claim that "the best fit to psychological data is typically achieved with word vectors constructed using context window sizes between \(\pm 2\) and \(\pm 10\) words." Otis and Sagi used \(n=15\), which is the default setting in Infomap.

We focused on the 22 prefix phonesthemes conjectured by Hutchins (1998). Our statistical analysis followed the procedure proposed by Otis and Sagi (2008). For each phonestheme, we first extracted all the vectors of the phonesthemebearing word types in our distributional semantic model. \({ }^{5}\) We shall refer to the resulting set of words (and vectors) as a phonestheme cluster. We then performed two Monte Carlo analyses. In the first analysis, we computed the average semantic similarity of each phonestheme cluster by forming 1000 random pairs and averaging the semantic distance obtained. In addition, we did the same for similarly-sized clusters of random words and performed an independent samples \(t\)-test for the resulting two groups of values. In the second analysis we took 50 random pairs within each phonestheme cluster and a corresponding group of pairs of random words and run 100 independent sample \(t\)-tests noting whether the mean of phonestheme cluster distances was significantly higher than the distances obtained for pairs of random words (with \(\alpha=0.05\) ). Based on the binomial distribution, we

\footnotetext{
\({ }^{4}\) See Sahlgren (2005) and Widdows and Cohen (2010) for a comparison of these methods.
\({ }^{5}\) Since we are dealing with a written corpus, this is done on the basis of an orthographic match with the phonetic grouping.
}
judged the number of significant t -tests as higher than 15 to lend statistical support to our Hypothesis 1. We performed the procedure 5 times and took the mean to be the final result.

We used the results obtained with the \(g l\) phonestheme cluster (which obtained the highest statistical support in the Otis and Sagi study) to optimize the minimum frequency and the window size parameters of our distributional semantic model. The model produced the most qualitatively sensible and most statistically stable results when setting the minimum frequency to 100 and the window size to 10 . This resulted in a model containing a set of 22292 vectors. This vector space was used in all subsequent parts of our study.

\section*{Results}

The results obtained with our parameter optimizing test on the \(g l\) phonestheme showed that the semantic relatedness of the words in the \(g l\) cluster was significantly higher than that of clusters of random words, as measured by our t-test procedure. On average, 26.4 t -tests produced a significant result (recall that the threshold of significance of the binomial test was for at least 15 out 100 t -tests to turn out as significant).

We used the same vector space (with the parameters fixed) to analyze the remaining 21 prefix phonesthemes. The results obtained are reported in Table 1. For each phonestheme, the table shows the number of word types in the phonestheme cluster (\# Tokens), the average degree of semantic relatedness amongst those words (Sim) calculated according to our first Monte Carlo analysis, the number of significant t-tests (\# Sig) calculated according to our second Monte Carlo analysis, and the mean effect size (Effect) of these t-tests. As can be seen, the model did not only confirm the semantic similarity of the words in the \(g l\) phonestheme (for which it had been optimized), but produced significant results for 16 different prefix phonesthemes out of the 22 considered. The

Table 1: Phonestheme semantic relatedness results
\begin{tabular}{|l|rccc|}
\hline Prefix & \#Tokens & Sim & \#Sig & Effect \\
\hline \hline\(b l-\) & 105 & 0.4607 & 56.8 & 0.2845 \\
\hline\(c l-\) & 156 & 0.4295 & 29.4 & 0.2570 \\
\hline\(c r-\) & 197 & 0.3921 & 7.40 & 0.2327 \\
\hline\(d r-\) & 99 & 0.4504 & 63.6 & 0.2849 \\
\hline\(f l-\) & 137 & 0.4340 & 34.2 & 0.2536 \\
\hline\(g r-\) & 197 & 0.4050 & 25.2 & 0.2617 \\
\hline\(s c-/ s k-\) & 167 & 0.4031 & 10.2 & 0.2443 \\
\hline\(s c r-\) & 32 & 0.5174 & 68.4 & 0.3093 \\
\hline\(s l-\) & 83 & 0.4275 & 42.4 & 0.2734 \\
\hline\(s m-\) & 42 & 0.4803 & 51.4 & 0.2817 \\
\hline\(s n-\) & 40 & 0.4650 & 52.2 & 0.2909 \\
\hline\(s p-\) & 161 & 0.4127 & 14.4 & 0.2392 \\
\hline\(s p l-\) & 11 & 0.5224 & 59.0 & 0.2723 \\
\hline\(s p r-\) & 24 & 0.3950 & 4.80 & 0.2373 \\
\hline\(s q u-\) & 24 & 0.4916 & 67.0 & 0.3205 \\
\hline\(s t-\) & 298 & 0.4307 & 25.0 & 0.2465 \\
\hline\(s t r-\) & 89 & 0.4525 & 61.8 & 0.2899 \\
\hline\(s w-\) & 67 & 0.5138 & 91.2 & 0.3396 \\
\hline\(t r-\) & 249 & 0.3912 & 5.80 & 0.2318 \\
\hline\(t w-\) & 22 & 0.4304 & 17.4 & 0.2408 \\
\hline\(w r-\) & 38 & 0.5155 & 90.6 & 0.3915 \\
\hline
\end{tabular}
average semantic relatedness of phonestheme clusters ( \(M=\) \(0.446, S D=0.044\) ) was highly correlated with the number of significant t-tests ( \(r=0.95, p<0.0001\) ) and was furthermore significantly higher than the average semantic relatedness of random words clusters ( \(M=0.397, S D=0.018\) ), as shown by an independent samples t-test \((t(42)=4.83, p<0.0001)\).

In line with the findings of Otis and Sagi (2008), we were thus able to obtain support for Hypothesis 1 for 16 conjectured phonestheme prefixes. Using the BNC - a more general, balanced, and modern corpus of English than Project Gutenberg - our study yielded higher support for the hypothesis than Otis and Sagi's previous study, which had found only 12 phonestheme prefixes as reaching statistical significance.

\section*{Experiment 2: Phonestheme Cluster Labeling}

\section*{Methods}

After establishing significant differences between the semantic relatedness scores for phonestheme word clusters and clusters of random words, we turned to our second experiment, whose aim is to investigate possible methods to automatically detect the core semantic content carried by a phonestheme. To our knowledge, this is the first attempt to address this issue by objective means. To test our Hypothesis 2 , we selected a number of prefix phonesthemes based on the amount of statistical support obtained in our first experiment and on how unambiguous and generally agreed upon were the sense definitions proposed in the literature. We selected 10 phonesthemes with high semantic relatedness scores and compiled a list of definitions based on the descriptions given by Hutchins (1998), Marchand (1969) and Reid (1967). The resulting list of phonesthemes together with their conjectured semantic import is presented in Table 2.

Table 2: Phonesthemic senses
\begin{tabular}{|l|l|l|}
\hline Prefix & Definition & Example \\
\hline \hline\(b l-\) & swelling, explosion, extension, broadness & bloating \\
\hline\(g l-\) & light, vision, look, brightness, shine & glitter \\
\hline\(g r-\) & threatening noise, anger, grip & growl \\
\hline scr- & unpleasant sound, irregular movement & screech \\
\hline sn- & nose, mouth, smell, snobbish person & sneeze \\
\hline spl- & divergence, spread, splash & splash \\
\hline squ- & discordant sound, softness, compression & squeeze \\
\hline str- & linear, forceful action, effort & strike \\
\hline\(s w-\) & rhythmical movement & swing \\
\hline\(w r-\) & irregular motion, twist & wring \\
\hline
\end{tabular}

In order to automatically assign a semantic class label to a phonestheme cluster, we used WordNet (Fellbaum, 2005), a cognitively motivated ontology of words and concepts linked by different semantic relations commonly used in computational linguistics. The main semantic relation connecting words that express different concepts in WordNet is the super/subordinate relation (also called hypernymy/hyponymy), which establishes a hierarchy of concepts from more general concepts like animal to increasingly specific ones like mammal or whale. Since hypernymy is a transitive relation, for
each word we can construct its hypernymy chain: the set of all its superordinate concepts or hypernyms connecting the word in question to the root node in the hierarchy (entity in the case of WordNet), ordered by their level of specificity.

WordNet is made up of independent hierarchies for different parts of speech: nouns, verbs, and adjectives. Given this (which prevents the possibility of assigning a crosscategorical semantic label) and the fact that the hierarchies for verbs and adjectives are far less complete than the noun hierarchy, we focused on the common nouns \({ }^{6}\) within each phonesthemic cluster amongst those listed in Table 2. This resulted in eliminating \(37 \%\) of words over all clusters.

For each common noun \(w\) in a phonesthemic cluster, we computed a set \(H(w)\) containing all superordinate concepts in the hypernym chain of \(w\), and derived a set \(\mathcal{H}\) of potential class labels for that cluster by taking the union of all sets \(H(w)\) for each noun \(w\) in the cluster. We then considered several methods for selecting the most optimal semantic class labels from \(\mathcal{H}\). Our methods were inspired by the approach to unsupervised ontology acquisition proposed by Widdows (2003) according to which "the most appropriate class-label for the set [of words] \(S\) is the hypernym \(h \in \mathcal{H}\) which subsumes as many as possible of the members of \(S\) as closely as possible in the hierarchy" (p. 278). Widdows offers a general scheme for defining an affinity score function \(\alpha(w, h)\) between a word \(w\) and a candidate label \(h\), which generates a ranking of all the potential class labels for a cluster of words:
\[
\alpha(w, h)= \begin{cases}f(\operatorname{dist}(w, h)) & \text { if } h \in H(w) \\ -g(w, h) & \text { if } h \notin H(w)\end{cases}
\]
where \(\operatorname{dist}(w, h)\) is a distance measure between a given word and a hypernym, \(f\) is a reward function that gives points to \(h\) if it subsumes \(w\) and the more points the closer this relationship, while \(g\) is a penalty function that subtracts points if \(h\) does not subsume \(w\). The best class-label is the hypernym \(h \in \mathcal{H}\) that has the highest affinity score summed over all the elements in the cluster.

Following Widdows, we chose as our distance measure the number of intervening levels in the WordNet hierarchy and set the rewarding function to \(f=1 / \operatorname{dist}(w, h)^{2}\). As for the penalty function \(g\), we tested constant values of \(0.25,0.1\) and 0.01. This particular variant of the scoring function thus magnifies the credit given to classes that are close to the words they subsume while giving a very small penalty to potential labels that miss out words in the cluster. The expected result is thus a ranking of class labels with a strong preference for specificity. This seems congruent with the nature of phonestheme clusters, which may contain a relatively large number of words that due to, for example, etymological factors are unlikely to be all related to the phonesthemic meaning. In fact, it is acknowledged in the literature that the soundmeaning associations are likely to be probabilistic (Hutchins,

\footnotetext{
\({ }^{6}\) We discarded proper nouns, which in WordNet are always terminal leaf nodes representing concrete instances rather than types.
}
1998) and that phonesthemic meaning can fall into related but separate groups. For example, \(g r\) is taken to be related to both angry noises (growl, grunt) and grabbing actions (grab, grasp). Given this, we also considered an approach whereby we first run a Gaussian Expectation-Maximization clustering algorithm on each phonestheme cluster to obtain more refined subsets of words and then run our scoring function algorithm on each of the resulting sub-clusters.

Finally, to counterbalance the preference for high specificity but potentially low coverage of the words in the phonesthemic clusters, we experimented with a different labeling algorithm that fixed a minimum coverage threshold. The algorithm examines all hypernyms \(h \in \mathcal{H}\), selects those that subsume a minimum percentage \(\theta\) of words in the cluster and then ranks them according to their specificity (the number of intervening levels to the root node entity). We tested the percentage values \(\theta=10\) and \(\theta=20\) and run the algorithm both on complete cluster phonesthemes and on the unsupervisedly derived sub-clusters. \({ }^{7}\)

\section*{Results}

Our results show that successful labeling of phonesthemic clusters can be performed but success depends on a number of factors. First, it is necessary to clarify what we mean by successful labeling. A labeling outcome of phonesthemic senses was deemed successful when the top 10 labels fulfilled the following criteria:
1. the topmost label is not the WordNet root node (entity);
2. the top 5 labels do not all have specificity score \(m \leq 2\);
3. at least \(50 \%\) of the top 10 labels carry meaning predicted for a given phonestheme;
4. the top 10 labels together subsume at least \(50 \%\) of the words in the cluster.
These heuristics mean that if it is possible to establish the semantic core of a phonesthemic cluster using WordNet hypernym trees, the top labels will be both specific and in the direction predicted by the literature. It is always possible to subsume all the words in the cluster, independently of their semantic relationship, under the root, just due to the WordNet structure. Such a label, however, would not be very informative. By the same reasoning, we excluded the next two levels of the hierarchy which contain concepts such as physical entity, abstraction, matter or relation. On the other hand, specificity needs to be balanced out by coverage, i.e. it is possible to have very specific labels as top results but covering only a small portion of words in the cluster. Finally, the labels need to at least intuitively relate to the domain specified in the literature for a given phonestheme.

Given these criteria, we obtained clear positive results for one phonestheme \((g l)\) out of 10 examined; moderately successful results for two phonesthemes ( \(s n\) and \(s t r\) ); and negative results for the remaining 7 phonesthemes. We present

\footnotetext{
\({ }^{7}\) Assigning a higher value to \(g\) would also increase coverage. However, for consistency \(g\) would have to be dependent on the size of the cluster. We instead choose a simple approach here which resorts to a percentage.
}

Table 3: Top 5 WordNet labels for \(g l-\), \(s n-\), and \(s t r\) -
\begin{tabular}{|c|c|ccc|}
\hline Prefix & Label & Score & Spec & Cov \\
\hline \hline \multirow{4}{*}{\begin{tabular}{c} 
gl- \\
\((N=56)\)
\end{tabular}} & brightness & 4.82 & 6 & \(23.7 \%\) \\
\cline { 2 - 5 } & flash & 4.67 & 5 & \(13.2 \%\) \\
\cline { 2 - 5 } & radiance & 3.92 & 7 & \(13.1 \%\) \\
\cline { 2 - 5 } & light & 2.23 & 5 & \(26.3 \%\) \\
\cline { 2 - 5 } & look & 2.16 & 8 & \(10.5 \%\) \\
\hline \hline sn & laugh & 1.71 & 5 & \(6.5 \%\) \\
\cline { 2 - 5 } & unpleasant person & 1.71 & 8 & \(6.5 \%\) \\
\cline { 2 - 5 } & photograph & 1.71 & 7 & \(6.5 \%\) \\
\cline { 2 - 5 } & smell & 1.71 & 8 & \(6.5 \%\) \\
\cline { 2 - 5 } & piece & 1.71 & 4 & \(6.5 \%\) \\
\cline { 2 - 5 } & noise & 1.71 & 6 & \(6.5 \%\) \\
\hline \hline str & effort & 2.22 & 8 & \(11.3 \%\) \\
\cline { 2 - 5 } & motion & 1.14 & 7 & \(11.3 \%\) \\
\cline { 2 - 5 } & labor & 0.39 & 7 & \(11.3 \%\) \\
\cline { 2 - 5 } & change & 0.98 & 6 & \(20.5 \%\) \\
\hline
\end{tabular}
the top labels for the 3 successfully labeled phonesthemes in Table 3, together with the scores calculated by the affinity score function with penalty set to a constant \(g=0.01\) (Score), specificity of each label (Spec) and the proportion of words in the cluster subsumed (Cov).

The \(g l\) phonestheme received light- and vision-related labels in all labeling algorithms that we tested. As can be seen in Table 3, they are clearly specific, cover a large proportion of words and all carry the predicted meaning. Similar results are obtained using other two settings of the \(g\) function and the coverage-based algorithm, although a small percentage of high-level labels like entity does appear in these lists.

Our moderately supported phonesthemes sn and str obtain labels in the predicted direction only with algorithms that reward specificity over coverage. For \(s n\) (related to nose, mouth and snobbism according to the literature), the best result is obtained with the distance measure and penalty function \(g=0.01\), while for the str phonestheme, related to forceful action - with the coverage-based algorithm of \(\theta=10\).

Similar conclusions can be drawn from phonesthemes that did not lead to clear tendencies in their labels or to specific enough labeling. Such lack of success is evident in either only one label out of the top 10 being relevant to the predicted meaning or all of the labels being very general. In the first case, for example, both \(g r\) and \(s c r\) words are subsumed under noise. In fact, this label appears in all instances of scr scores as a top label, covering \(26.9 \%\) of words in the cluster. However, the rest of the top labels are either of a general kind (entity, change) or not related to sound or movement (handwriting, wound) and therefore we cannot consider the labeling result to be very strong. In other cases, the words are primarily subsumed by labels like entity and abstraction.

As explained in the Methods, we considered the possibility that clusters might be composed of several groups of words that do not all share the same semantic content. This is especially likely for numerous clusters (e.g. \(g r\) cluster even with proper nouns removed contained 158 words). To counteract this problem we examined how prior sub-clustering affects the labeling results. The EM algorithm we used detected the
presence of two clusters for 4 out of 10 phonesthemes we considered (bl, gl, gr and str). \({ }^{8}\) Again, however, the most interesting result was obtained for the \(g l\) phonestheme. According to the labels we obtained for its two sub-clusters, only one of them was light-related. The second sub-cluster contained words like gluten, glucose and glycoprotein, which were placed under labels such as protein, macromolecule and organic compound, indicating a clear presence of chemistryrelated words. No clear patterns were obtained for the other 3 phonesthemes with two sub-clusters. We do not exclude the possibility that this result was due to the quality of the vectors given to our clustering algorithm or to the algorithm itself. For example, the presence of a large proportion of kinship concepts in one of the \(g r\) sub-clusters (grandfather, grandchildren etc.) led to it being assigned labels such as grandparent and ancestor, while the same sub-cluster contained words such as growl and grunt. Therefore, whether the sub-clustering step is theoretically sound and if so how it should be accomplished requires further study.

\section*{Discussion}

The results of our first experiment are largely in line with those of Otis and Sagi (2008), but we also see a number of differences. On the one hand, we obtain higher support for phonestheme clusters overall and show statistical significance for several phonesthemes previously unsupported. On the other hand, our support for the strongest phonesthemes in the original study ( \(g l\) and \(s p r\) ) is weaker. These differences can be due to several factors. First, we use a different, more modern and balanced, BNC corpus and our resulting phonestheme clusters are larger. Second, we use a different method for building our distributional model - both a different algorithm (Random Projection) and a smaller window size.

It is worth noting that our tuning experiments with the \(g l\) phonestheme show that the kind of pre-processing that we apply to the corpus and the window size parameter do make a difference to the statistical results that can be obtained from the model. However, while pre-processing could be viewed as merely a methodological challenge common to all types of corpus analyses, there might be a theoretical significance behind the impact of the window size. Sahlgren (2008), for example, suggests that a small window size is preferable for detecting paradigmatic relationships between words (those that hold between words that do not co-occur themselves but occur in similar contexts, e.g. dog and cat) and at the same time there is evidence (Peirsman, Heylen, \& Geeraerts, 2008) that larger context is beneficial for picking out syntagmatic relationships that hold between words that often occur together (e.g. "crystal clear"). To our knowledge, the kinds of relationships that hold between phonesthemic words (in general or depending on a given phonestheme) have not been systematically investigated using such distinctions and further work on the influence of the kind of context useful for detecting phonesthemic relatedness, in conjunction with experimental

\footnotetext{
\({ }^{8}\) For the other phonesthemes, no sub-clusters were detected.
}
work on similarity, could offer clues on this issue.
Our labeling results are somewhat disappointing but, given the novelty of our approach, still highly informative. The fact that we obtain better results for \(s n\) and \(s t r\) phonesthemes with algorithms that favor specificity over coverage and that labeling is not fully successful with the remaining phonesthemes are puzzling given the high support that we obtain for these phonesthemes in our first experiment. We believe that there are two possibilities that can explain this.

The first possibility is that our WordNet-based methodology is not fully suited to discover the common semantic content that is present. WordNet does not allow for integrating hypernymy tree chains across different parts of speech, which might be vital for phonesthemes, a large proportion of which are verbs. In addition, it does not make all the distinctions that would be useful for phonesthemic studies, e.g. both \(s c r\) and \(g r\) are associated with kinds of sound but one is "unpleasant" and the other one "threatening" - a distinction which is not part of the WordNet taxonomy. On a more general note, hypernymy might not be the most appropriate relation for all phonesthemes, e.g. snout and sneezing are not similar because they are both a type of nose. Therefore, perhaps better labeling results could be achieved using a different semantic network, such as ConceptNet \({ }^{9}\), which allows for exploiting other than merely "is a" relations.

The second possibility, which we cannot reject, is that there is in fact no semantic core that unites phonestheme clusters and that the statistical support obtained by Otis and Sagi and in our first experiment is a result of a particular methodology. This interpretation is suggested by the fact that in our second experiment the overall coverage of phonesthemic words by the semantic labels is relatively low. Qualitative examination of the clusters also seems to show that they contain a lot of variability. In the future, we plan to design stricter tests for example, comparing phonestheme clusters to clusters that share a particular (non-phonesthemic) sub-string rather than simply to a group of random words.

Ultimately, the aim of automatically detecting phonesthemes and their semantic content in a more objective, falsifiable way is, on the one hand, to help researchers interested in iconicity to validate the existence of phonesthemes previously reported in the literature, to possibly discover new phonesthemes, and to settle disputes over their particular meaning; and on the other hand, to open the door to investigating further the cognitive nature of the semantic relationships that unite phonestheme clusters. This study constitutes a step in this research programme.

\section*{References}

Bellugi, U., \& Klima, E. (1976). Two faces of sign: Iconic and abstract. Annals of the New York Academy of Sciences, 280, 514-538.
Bergen, B. (2004). The psychological reality of phonaesthemes. Language, 80(2), 291-311.

\footnotetext{
\({ }^{9}\) Freely available from http://conceptnet5.media.mit.edu.
}

Bird, S., Klein, E., \& Loper, E. (2009). Natural language processing with python: Analyzing text with the natural language toolkit. O'Reilly Media.
Blust, R. A. (2003). The phonestheme n-in austronesian languages. Oceanic Linguistics, 42(1), 187-212.
Burnard, L. (Ed.). (2007). Reference Guide for the British National Corpus (XML Edition). Oxford Universtity Computing Services.
Drellishak, S. (2006). Statistical techniques for detecting and validating phonesthemes. Unpublished master's thesis, University of Washington.
Fellbaum, C. (2005). Wordnet and wordnets. In Encyclopedia of language and linguistics (Second ed.). Oxford: Elsevier.
Firth, J. (1930). Speech. London: Oxford University Press.
Hinton, L., Nichols, J., \& Ohala, J. J. (2006a). Introduction. In Sound symbolism. Cambridge University Press.
Hinton, L., Nichols, J., \& Ohala, J. J. (Eds.). (2006b). Sound symbolism. Cambridge University Press.
Hutchins, S. S. (1998). The psychological reality, variability, and compositionality of english phonesthemes. Unpublished doctoral dissertation, Emory University.
Marchand, H. (1969). The categories and types of presentday english word-formation. Munich: C. H. Beck.
McDonald, S., \& Ramscar, M. (2001). Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. In Proc. 23rd CogSci.
Ohala, J. J. (1994). The frequency codes underlies the sound symbolic use of voice pitch. In Sound symbolism. CUP.
Otis, K., \& Sagi, E. (2008). Phonaesthemes: A corpus-based analysis. In Proc. 30th CogSci.
Peirsman, Y., Heylen, K., \& Geeraerts, D. (2008). Size matters: Tight and loose context definitions in English word space models. In Proc. ESSLLI Workshop on Distributional Lexical Semantics.
Perniss, P., Thompson, R., \& Vigliocco, G. (2010). Iconicity as a general property of language: Evidence from spoken and signed languages. Frontiers in Psychology, 1, 1-15.
Reid, R. (1967). Sound symbolism. T\&A Constable.
Rhodes, R. (2006). Aural images. In Sound symbolism. CUP.
Sahlgren, M. (2005). An introduction to Random Indexing. In Proc. Methods \& Applications of Semantic Indexing.
Sahlgren, M. (2008). The distributional hypothesis. Italian Journal of Linguistics, 20(1), 33-54.
Schütze, J. (1997). Ambiguity resolution in language learning. Stanford: CSLI Publications.
Shrum, L., \& Lowrey, T. (2007). Sounds convey meaning: The implications of phonetic symbolism for brand name construction. In Psycholinguistic phenomena in marketing communications. Mahwah: Erlbaum.
Widdows, D. (2003). Unsupervised methods for developing taxonomies by combining syntactic and statistical information. In Proc. HLT-NAACL.
Widdows, D., \& Cohen, T. (2010). The semantic vectors package: New algorithms and public tools for distributional semantics. In Proc. 4th Int'l Conf. on Semantic Computing.

\title{
Can Playing Portal Affect Spatial Thinking and Increase Learning in a STEM Area?
}

\author{
Deanne M. Adams (adams@psych.ucsb.edu) \\ Richard. E. Mayer (mayer@psych.ucsb.edu) \\ Department of Psychological and Brain Sciences, University of California, Santa Barbara, CA 93106.
}

\begin{abstract}
Spatial skills have been associated with learning in STEM areas and some research has shown that playing video games could facilitate the development of spatial skills. This study examines whether playing a game that uses a realistic physics engine and places spatial demands on the players could facilitate learning a subsequent physics lesson. Fifty-eight participants viewed a brief lesson on Newton's laws of motion after either playing the puzzle game Tetris or the firstperson perspective puzzle game Portal, which incorporates aspects of physics such as momentum. The groups did not differ on subsequent tests of learning outcomes involving physics, but the Portal group scored significantly higher on a perspective taking test \((d=0.57)\). This study shows that playing a commercial game that incorporates Newtonian physics does not prepare students to learn physics but does improve an important spatial cognition skill related to physics.
\end{abstract}

Keywords: video games; physics learning; spatial orientation

\section*{Objectives}

The goal of this study is to examine whether playing an off-the-shelf first-person perspective puzzle game based on physics principles (i.e., Portal) can help prepare students to learn physics concepts and improve their spatial skills as measured by the perspective taking task. In the present study, students studied a brief lesson on Newton's laws of motion after spending an hour playing Portal or the puzzle game Tetris. Examining the effects of playing an off-the shelf computer game can be called cognitive consequences research and constitutes one of three major experimental methodologies for game research (Mayer, 2011). In short, the goal is to determine the cognitive consequences of playing Portal on (a) improving a spatial skill that is related to learning in physics and (b) enabling students to learn physics concepts on a subsequent physics lesson.

\section*{Learning Physics and Video Games}

Learning physics can often be difficult because many learners already have misconceptions about how the physical world works. White (1993) argued that one of the problems with physics education is the top-down approach in which abstract formulas are taught first, which students later have trouble applying to every-day phenomenon. Instead White (1993) argued that physics should be taught using an approach in which students are presented with concrete versions of these models in the form of computer
simulations. While the real world can be overly complex with multiple forces acting simultaneously, a simulation can control for these factors and allow for students to make predictions, then test them, and to try to explain the results. White (1993) used a group of microworlds called "ThinkerTools" with 6th graders. The curriculum was developed so that the initial microworlds had simple situations (no friction and only one dimension of motion) so that learners could develop intuitive knowledge before dealing with more sophisticated causal relationships. White (1993) found that, compared to high school students who were taught using traditional methods, \(6^{\text {th }}\) graders who received the "ThinkerTools" curriculum performed better on simple force and motion problems, better retained what they learned, and transferred what they learned to new contexts.

Similar to White's (1993) computer simulation, some off-the-shelf video games have been developed to depict realistic movement based on Newtonian physics and provide simplified environments to make game play easier. In a study by Masson, Bub, and Lalonde (2011) participants completed 6 one-hour game training sessions playing the video game Enigmo or the control game Railroad Tycoon 3. During Enigmo the player must alter the trajectories of falling droplets so that the drops land in target receptacles. The authors proposed that the Enigmo group would benefit from game play because the game gives repeated exposure to the movement of falling objects and this may benefit students by priming them to learn from formal physics instruction. The pretest/posttest consisted of a test of knowledge about the motion of objects with 15 items involving objects moving freely through space based on physics. Participants in the Enigmo group increased their ability to produce realistic trajectories but only in terms of the general parabolic shapes of those trajectories. After the posttest, participants then completed a PowerPoint tutorial on physics after which they completed 13 test problems based on the tutorial. Masson et al. (2011) found that students in the Enigmo group did not show a higher improvement after viewing the tutorial compared to the Railroad Tycoon 3 group.

Masson et al. (2011) were not able to show that experience playing a game that uses realistic physics motion prepares students to benefit from direct instruction in physics, but video games may benefit science learning through improvements in visuospatial ability. Previous research has shown that playing video games such as firstperson shooters (Feng, Spence, \& Pratt, 2007), and spatial puzzle games (Okagaki \& Frensch,1994; Subrahmanyam \&,

Greenfield,1994; Terlecki et al., 2008) can increase different spatial cognition skills, such as mental rotation. Work by Kozhevnikov and colleagues has shown a relationship between spatial ability and physics problem solving (Kozhevnikov, Hegarty and Mayer, 2002; Kozhevnikov, Motes, and Hegarty, 2007). When looking at a factor analysis of spatial ability tests and different types of kinematic problems, Kozhevnikov, Hegarty, and Mayer (2002) found that spatial ability loaded on the same factor as problems which involved determining an object's trajectory based on combining two motion vectors and using a different frame of reference to determine the characteristics of an object's motion. In an additional study in which participants were classified as being either high or low spatial, students classified as having high spatial ability were: (1) more successful at integrating several motion parameters versus only considering one at a time; (2) could interpret a object's motion based on kinematic graphs versus seeing the graphs as picture-like representations; and (3) understood the connection between different representations of spatial problems versus using multiple uncoordinated representations of the same problem (Kozhevnikov et al. 2007). Kozhevnikov et al.’s (2007) results with eye movements also suggest that high spatials actually visualize the movement of objects based on integrating motion components while low spatial individuals do not. Thus, there is evidence that certain spatial skills are related to success in STEM subjects.

Sanchez (2012) showed that playing games can also have a benefit on learning in science areas through priming these visuospatial abilities. Participants either played 25 minutes of the first-person shooter game, Halo: Combat Evolved or the word anagram game Word Whomp before reading a lesson on plate tectonics. Participants did not significantly differ on prior knowledge in the subject area or spatial skills, as measured by the first section of both the card rotation task and the paper folding task. After playing the game participants then read a complex text about plate tectonics. They then completed an essay task in which they were asked to write a causal essay about "What caused Mt. St. Helens to erupt?" After the essay task they completed the second part of both the card rotations task and the paperfolding task. The results found that playing the action video game had a significant positive effect on essay quality and rotation task performance. Sanchez (2012) proposed that the first-person shooter game requires visuospatial skills that are important for learning in some science areas. The present study parallels Sanchez's methodology, but explores the domain of physics learning.

\section*{Current Study}

In the fall of 2011, the game company Valve introduced an educational program called Learn With Portals, which proposed using their games Portal and Portal 2 to help teach students critical-thinking skills and physics (http://www.learnwithportals.com/). The games, depicted in Figure 1, incorporate elements of physics, such as
momentum, into a problem solving game. Portal is intended to benefit physics learning because it applies realistic physics principles into the game experience, therefore allowing the player to build experience with physics concepts in a controlled environment.

It is unclear whether Portal has any effect on spatial cognition skills similar to previous research with firstperson shooters and Tetris. If Portal does facilitate cognitive ability development it could help students learn physics similar to Sanchez's (2012) work with plate tectonics. Playing Portal requires the participant to imagine what a room may look like from a different perspective. Placing the portals in order to solve the puzzles within the game may therefore require the use of the spatial skill known as spatial orientation or the ability to visualize what a different perspective may look like from another location (Hegarty \& Waller, 2004). Kozhevnikov, Hegarty, and Mayer (2002) found that performance on a spatial orientation test correlated with performance on a kinematics questionnaire, which included items from the physics test known as the Force Concept Inventory (FCI) (Hestenes, Wells, \& Swackhamer, 1992). In this study, Tetris is used as the control condition because although Tetris has been found to increase performance on mental rotation under certain training regimes (Okagaki \& Frensch,1994; Terlecki et al., 2008), Kozhevnikov et al. (2002) showed that mental rotation was not associated with kinematic problem solving.

Therefore, the objective of this experiment is to: (1) determine whether playing Portal can increase performance on a spatial cognition task; (2) determine whether an hour's worth of playing Portal versus Tetris can increase learning from a subsequent lesson on physics; (3) determine whether there is a relationship between spatial cognition skills and performance on physics problem solving.

Participants and Design Participants were 63 (39 male, 24 female) students from the University of California, Santa Barbara. Ages ranged from 17-23 years old with a mean age of \(19.03(S D=1.28)\). Participants received class credit for their participation. Thirty-four participants served in the Portal group and 29 served in the Tetris group.

Materials The pre-game paper-based materials consisted of a participant questionnaire and pretest. The participant questionnaire contained basic demographic items concerning the participant's gender, year in school, age, and also asked participants to rate their spatial cognition ability (i.e. being able to visualize objects or imagine rotating items) on a 5-point scale ranging from "Very Poor" to "Very Good". Participants were also asked how many hours they played video games, excluding card games and text based games, during a typical week ranging from "I do not play video games" to "More than 10 hours per week". Participants were also asked whether they had played Portal or Tetris before. To examine prior knowledge, participants were asked to indicate whether they had previously taken
physics courses during high school or college, or if they were in the process of taking a physics course.

The pretest asked participants to try to recall Newton's three laws of motion. Participants could receive a total of 6 points on this section, 2 points for each law if all of the elements were correct. For example, for the \(1^{\text {st }}\) Law, the Law of Inertia, participants had to state both that a body in motion will stay in motion while a body at rest will stay at rest and that the object's state will not change unless acted upon by an external force. Excluding either the "at rest" or "in motion" element would result in the student only receiving one point for the \(1^{\text {st }}\) law. The pretest also included 4 multiple-choice questions dealing with naïve physics. The first two were the cliff problem and the ball problem from McCloskey (1983). The cliff problem asks the learner to determine what path a person will take if they run at a constant rate of speed off the edge of a cliff. The correct answer to this problem is based on the \(1^{\text {st }}\) law of motion, while some of the incorrect options are consistent with impetus theory or the idea that objects contain force that runs out. The ball problem asks the learner to determine where a heavy ball will land if you dropped the ball while running forward at a constant speed. The last two questions came from White's (1993) testing materials and asked participants about two balls falling from different heights. This question was used to examine the participant's understanding of gravity. Students received one point for each correct answer in this section. Overall, the pretest scores could range from 0 to 10 .

The control game used for this study was the puzzle game Tetris. During Tetris the player must make lines of blocks using 6 different block shapes. Every time a line is completed the line disappears from the rectangular play area and the player receives points. The more lines that are completed at once, or the larger the combo, the higher the points the player receives. The player can press a button on the keyboard to rotate the blocks in increments of 90 degrees in order to best fit them into the available spots at the bottom. The block shapes fall from the top of the play area at a constant rate and as players gather more points the falling rate increases therefore increasing the level difficulty of the game. In the marathon mode version of the game,


Figure 1: Sample screen shot from Portal game play. Chamber 13.
play continues until the player fills the rectangular play area with incomplete lines.

The target game used in this study was Portal (2007), a first-person perspective puzzle game. The narrative of the game is that you are a test subject named Chell that has woken up in a facility in which you must navigate through testing chambers using portals. The player is given advice and feedback from a computer named GLaDOS who promises cake upon the completion of the testing regimen. During the game the player acquires the use of a portal gun, which shoots two portals, a blue and an orange one, which are linked to the left and right mouse buttons respectively. The two portals can be fired on specific surfaces during the game and can link those two locations so that when you enter one portal you will exit the other. The game sometimes requires the participant to make use of momentum so that the player can traverse large horizontal distances. To do this a player can place one portal at the bottom of a pit and another on a vertical wall so that falling into the portal at the bottom of the pit will increase their momentum using gravity and they will exit the opposite portal with enough speed to travel horizontally over pits and other obstacles (Chamber 10 of the game requires this solution). Solutions become progressively harder as the chambers continue requiring the use of more and more portals. There are a total of 19 levels/chambers in the game. In this experiment, participants started on the \(10^{\text {th }}\) chamber of the game since it is the first one that deals with momentum to solve the puzzle. The chamber also starts with GLaDOS explaining momentum, in which she states that portals do not affect forward momentum. She also informs the player that momentum is a function of mass and velocity. Participants were encouraged to get as far through the chambers as possible until the hour of game play was over.

The physics lesson consisted of an 18-slide presentation on Newton's three laws of motion and the law of conservation of momentum. The presentation also addressed the incorrect impetus theory and how it is a common misconception in physics. The lesson included the basic rules along with examples for each of the laws such as a canon recoiling after firing a cannonball for Newton's \(3^{\text {rd }}\) law or "for every action this is an equal and opposite reaction."

There were four paper-based posttests: a retention test, a shorten, adapted version of the Force Concept Inventory (FCI) (Hestenes, Wells, \& Swackhamer, 1992; Hestenes \& Halloun, 1995), a Portal based scenario test, and a spatial orientation test. The retention test asked the participant to recall the three laws of motion. This question was used to determine whether there were basic recall differences between the two groups. Once again, students could receive a total of 6 points for this section, 2 points for each law with all of the components correctly defined.
The adapted version of the FCI consisted of 24 multiplechoice items. Only items dealing with the first three laws were included since the short physics lesson only dealt with
these topics. This test was chosen because many of the items deal with the movement of objects and often includes items that could be answered incorrectly based on impetus theory instead of using Newtonian physics. The learner must apply what he or she knows about the three laws and momentum in order to select the correct answer. For example, one item asks participants to imagine that a bowling ball had been dropped out of the cargo bay of an airliner traveling horizontally and the participant must pick the correct path that the ball will fall from the plane to the ground below. There was only one correct answer for each item with a total of 24 possible points.

The scenario test contained two questions about scenarios taken straight from the Portal game and asked participants to determine whether the law of conversation of momentum had been violated. In one example, the direction of the individual changes (from traveling vertically to horizontally) while in the other the direction is kept constant (vertical to vertical). Participants are asked to justify their answers and must have the correct explanation to receive full marks on the two items with one point for correctly selecting whether the law had been violated or not and one point for justifying their reason, for example, explaining how momentum is a vector (speed and direction). The total score could range from 0 to 4.

To determine whether playing Portal affected the spatial skill known as spatial orientation, Hegarty and Waller's (2004) Perspective Taking/Spatial Orientation Test was used. During this task participants are given an array of 7objects including a house, a cat, a tree, etc. For each question, participants are asked to imagine that they are standing at one object facing the direction of another. They are then asked to "point" to the direction of a third object. To respond, below the picture array, participants are given a circle in which the first direction (i.e. cat facing the flower) is given and they must then draw a line indicating which direction the third object is relative to the other two. Participants are given 5 minutes to complete as many items as possible with a total of 12 possible items. Hegarty and Waller (2004) showed that the spatial ability known as spatial orientation is highly correlated with mental rotation but there is a disassociation between the two, suggesting two separate abilities.

Apparatus Both games were run on Dell computers with 17 inch color monitors, with ATI Radeon HD 2600 XT video cards. The lesson was also administered using the computers. All the testing materials, including the spatial orientation task, were given using paper and pencil.

Procedure Participants were randomly assigned to groups and tested in individual cubicles. Upon entering the lab participants were seated at separate computer cubicles. Participants were first asked to fill out the participant questionnaire sheet and the pretest, at their own pace. Participants were then informed that they were going to play their respective game for an hour followed by a lesson on
physics, a posttest, and the spatial orientation task. Each cubicle also had instructions for how to play the participant's particular game. Participants in the Tetris condition played on "marathon" mode in which the game becomes progressively harder as the player acquires points. For Tetris the experimenters recorded the scores and level reached for each of the completed games. At the end of the hour, the Portal group had their game progress saved, which was later accessed by the experimenter to determine how many chambers the participant had completed.

Next, the physics lesson was initiated on the participant's computer. Participants were told that they had a minimum of 8 minutes to review the physics lesson and could have more time if they wished. Upon completing the lesson the participants were given a packet including the retention test, FCI items, and the Portal scenario questions and told that they had as much time as they wanted to answer the questions. After turning in the packet, participants were then given the spatial orientation test. They had 5 minutes to complete as many items as possible.

\section*{Results}

For the analysis, only participants who were actively engaged during game play were included. The reasoning behind this decision is that only active participants who had Portal full exposure to all the elements within the game were of interest. Therefore participants were excluded from the analysis if they did not get past Chamber 11 while playing Portal or if they did not get beyond level 5 in Tetris. Using these criteria, 4 Portal participants and 2 Tetris participants were removed from the analysis, leaving 30 participants in the Portal group and 28 in the Tetris group.

The two groups did not differ significantly in the proportion of males and females, \(X^{2}(1, N=58)=.009, p\) \(=.92\), the proportion of individuals who were familiar with the game Portal, \(X^{2}(1, N=58)=1.62, p=.20\), and the proportion of individuals who were familiar with the game Tetris, \(X^{2}(1, N=58)=.283, p=.595\). The participants also did not differ on their self-ratings of spatial cognition ability, \(t(56)=-.431, p=.67\), and reported hours of video game playing, \(t(56)=.037, p=.97\). There was no significant difference on pretest performance, \(t(56)=-1.15\), \(p=.26\), or prior knowledge with physics, \(t(56)=.82, p=\) . 42 .

\section*{Does playing Portal improve students' spatial cognition?}

The perspective taking task was scored so that any item in which the participant was within 15 degrees of the correct angle was scored as correct and awarded 1 point while anything beyond 15 degrees and items that were not attempted were not awarded any points. Participants in the Portal condition significantly outperformed participants in the Tetris condition on the spatial orientation test, \(t(55)=-\) \(2.12, p=.04, d=0.57\). This is the major new positive finding in the study.

Table 1: Means and standard deviations for all posttest measures.
\begin{tabular}{lcccc}
\hline & \multicolumn{4}{c}{ Group } \\
\cline { 2 - 5 } & Portal & Tetris & & \\
Measure & \(M(S D)\) & \(M(S D)\) & \(p\) & \(d\) \\
\hline Retention & \(5.27(1.44)\) & \(5.39(1.06)\) & 0.71 & -0.09 \\
FCI & \(13.07(5.37)\) & \(12.75(4.45)\) & 0.81 & 0.06 \\
Portal Scenerio & \(1.77(1.50)\) & \(1.39(1.42)\) & 0.34 & 0.26 \\
Spatial Orientation & \(7.07(3.03)\) & \(5.52(2.41)\) & 0.04 & 0.57 \\
\hline
\end{tabular}

Importantly, there was a significant positive correlation between performance on the adpated FCI (which measures physics intuitive knowledge) and performance on the spatial orientation test, \(r(57)=.323, p=.014\). This finding suggests that spatial cognition skills such as spatial orientation may be related to success in physics learning.

\section*{Does playing Portal help students learn physics? Table 1} shows the means (and standard deviations) of the two groups on each of the four tests. There were no significant differences on recall of the three laws of motion in the retention test, \(t(56)=.378, p=.71\); applying what they had learned to answer the FCI items, \(t(56)=-.242, p=.81\); or answering questions involving conservation of momentum through portals on the scenario test, \(t(56)=-.972, p=.34\). Therefore, there was no evidence that playing Portal facilitated learning about the laws of motion.

\section*{Discussion}

On the negative side, playing Portal did not improve learning of physics content, paralleling the results of Masson et al.'s (2011) research with Enigmo. On the positive side, playing the first-person perspective puzzle game Portal for an hour resulted in higher performance on an important spatial cognition skill (i.e., spatial orientation) compared to playing the 2D puzzle game Tetris. In addition the results showed a significant correlation between performance on a measure of spatial cognition (i.e., the spatial orientation test) and a measure of physics knowledge (i.e., the adapted FCI), paralleling the results from Kozhevnikov et al. (2002) showing a connection between spatial skills and success in STEM learning.

This study provides evidence that spatial orientation is a learnable skill. Games such as Portal, which require participants to imagine taking different viewpoints, may facilitate the development of this skill. In contrast, a game like Tetris which can utilize mental rotation under certain circumstances, does not tax spatial orientation therefore causing no improvement. Overall, the results support the idea that training of spatial skills is domain specific, such that different kinds of computer games can promote
different kinds of spatial skills rather than improving spatial cognition in general.

These findings support the idea that if educators want students to improve in spatial orientation skill, they can benefit from playing a first-person perspective puzzle game like Portal. Improving in this skill appears to be related to STEM learning, so in order to help students that might be struggling in areas such as physics, perhaps developing their spatial orientation skills could facilitate learning. Educational physics games could incorporate both direct instruction and spatial components to increase learning.

\section*{Limitations and Future Directions}

One limitation for this study is the lack of a pretest measure for spatial orientation. None of the pretest or demographic measures showed any significant differences between the two testing groups; therefore random assignment should have balanced spatial ability between the two groups. The spatial orientation task only has one form with 12 items and dividing the task into 2 sections may have weakened the power of the measure. In the future, a second version of this test with an alternative array of objects could be used a pretest to determine spatial orientation ability before game play.

Although playing the off the shelf version of Portal for a brief period of time did not benefit students when learning physics, perhaps playing either for longer or playing chambers created to teach specific principles would result in higher learning gains. Our study found that there was a significant benefit on spatial orientation scores for playing Portal as well as a significant correlation between performance on the spatial orientation task and performance on the modified FCI. Perhaps with further game play participants could increase their spatial skills, therefore facilitating learning physics problems dealing with motion. Previous research with video games has shown that different cognitive skills can be improved by playing games (Green and Bavelier, 2003). While Tetris can improve mental rotation under some circumstances but not others (Terlecki, et al., 2008, Sims and Mayer, 2002) it is important to consider what skills are improved by a particular game and what skills are associated with success in a particular STEM
area. For example, Sanchez (2012) found improvements in mental rotation and learning about plate tectonics from playing a first-person shooter but no improvement in the paper-folding task. Spatial orientation has been found to correlate with performance on kinematic tasks, therefore a game which trains these skills could help participants with solving these problems.

In addition, the game company Valve has released a tool in which players can create their own testing chambers with the Portal 2 game software. Similar to White’s (1993) highly controlled simulations, if the Portal 2 software could be used to create lessons in which students build up prior knowledge through playing the game, then perhaps physics learning could be improved. One issue with Portal is that participants view the game from the first-person perspective so they are unable to see the falling trajectories of their game avatar caused by differences in momentum. Therefore, misconceptions about how objects fall can not be correctly addressed. By creating special testing chambers, other objects could be used to show how physics behaves in a controlled environment. Further research must be done to determine under what circumstances a lesson using the Portal game environment could facilitate learning and the development of spatial skills.

\section*{Acknowledgments}

This project was supported by a grant from the Office of Naval Research. We would also like to thank the game company Valve, specifically Dr. Mike Ambinder and Leslie Redd, for the electronic copies of Portal donated for research purposes.

\section*{References}

Chinn, C.A., \& Malhotra, B.A. (2002). Children's responses to anomalous scientific data: How conceptual change impeded? Journal of Educational Psychology, 94 , 327343.

Feng, J., Spence, I., \& Pratt, J. (2007). Playing an action video game reduces gender differences in spatial cognition. Psychological Science, 18(10), 850855.

Green, C.S., \& Bavelier, D. (2003). Action video game modifies visual selective attention. Nature, 423, 534-538.
Hegarty, M., \& Waller, D. (2004). A dissociation between mental rotation and perspective-taking spatial abilities. Intelligence, 32, 175-191.
Hestenes, D., \& Halloun, I. (1995). Interpreting the FCI.
The Physics Teacher, 33, 502-506.

Hestenes, D., Wells, M., \& Swackhamer, G. (1992). Force concept inventory, The Physics Teacher, 30, 141-151.
Kozhevnikov, M., Hegarty, M., \& Mayer, R. E. (2002). Visual/spatial abilities in problem solving in physics. In M. Anderson, B. Meyer, \& P. Olivier (Eds.), Diagrammatic Representations and Reasoning SpringlerVerlag.
Kozhevnikov, M., Motes, M.A., \& Hegarty, M. (2007). Spatial visualization in physics problem solving. Cognitive Science, 31, 549-579.
Learn With Portals (2011). Retrieved from http://www.learnwithportals.com/
Masson, M.E.J., Bub, D.N., Lalonde, C.E. (2011). Videogame training and naïve reasoning about object motion. Applied Cognitive Psychology,25 (1), 166-173.
Mayer, R. E. (2008). Learning and instruction (2 \({ }^{\text {nd }}\) ed). Upper Saddle River, NJ: Pearson.
Mayer, R. E. (2011). Multimedia learning and games. In S. Tobias and J. D. Fletcher (Eds.), Computer games and instruction (pp. 306). Charlotte, NC: Information Age Publishing.
McCloskey (1983). Intuitive Physics. Scientific American, 248(4), 122-130.
McCloskey, M., Washburn, A., Felch, L. (1983). Intuitive physics: The straight-down belief and its origin. Journal of Experimental Psychology: Learning, Memory, and Cognition, 9(4), 636-649.
Okagaki, L., \& Frensch, P. A. (1994). Effects of video game playing on measures of spatial performance: Gender effects in late adolescence. Journal of Applied Developmental Psychology, 15, 33-58.
Portal (computer software) (2007). Bellevue, WA: Valve
Sanchez, C.A. (2012). Enhancing visuospatial performance through video game training to increase learning in visuospatial science domains. Psychono Bull Rev, 19, 5865.

Sims, V.K. \& Mayer, R.E. (2002). Domain specificity of spatial expertise: The case of video game players. Applied Cognitive Psychology, 16, 97-115.
Subrahmanyam, K. \&, Greenfield, P.M. (1994). Effect of video game practice on spatial skills in girls and boys. Journal of Applied Developmental Psychology, 15, 13-32.
Terlecki, M.S., Newcombe, N.S., \& Little, M. (2008). Durable and generalized effects ospatial experience on mental rotation: Gender differences in growth patterns. Applied Cognitive Psychology, 22, 996-1013.
White, B.Y. (1993) ThinkerTools: Causal Models, Conceptual Change, and Science Education. Cognition and Instruction, 10(1), 1-100.

\title{
Dancing With Myself: The effect of majority group size on perceptions of majority and minority robot group members
}

\author{
Henny Admoni, Bradley Hayes, David Feil-Seifer, Daniel Ullman, and Brian Scassellati \\ Department of Computer Science, Yale University \\ 51 Prospect Street, New Haven, CT USA
}

\begin{abstract}
While social psychology has identified characteristics of intergroup dynamics, few studies have looked into the perceptions of robot group dynamics. In this experiment, we separate robots into majority and minority groups based solely on their behavior in a simple dance routine. We attempt to understand how people's perceptions of robots within those groups change based on group size and features of behavior. Participants viewed the robot dances and rated one robot from each group on a variety of characteristics. We find that being from the minority versus majority group has a significant impact on perceptions of a robot's creativity, interestingness, anti-sociality, dancing ability, and how much of a team player it is. At the same time, individual behaviors (leading the dance, following the dance, or performing an entirely unique dance) have no statistical effect on participants' ratings of robot characteristics. From these results, we conclude that group size has a larger effect than behavior on subjective evaluations of robots in majority and minority groups.
\end{abstract}

Keywords: Group dynamics; majority group; minority group; intergroup relations; robotics; human-robot interaction

\section*{Introduction}

The tendency to categorize people into established groups is an automatic social behavior, and influences much of how we perceive the world. Psychologists have studied the dynamics of intergroup relations, particularly when the group sizes are unequal, for many decades (Tajfel, 2010). Among many other findings, the literature reveals that minority groupsthose with fewer members-are susceptible to being influenced by a unanimous majority group (Asch, 1956), even when the majority's assertions are incorrect on a point of fact. Furthermore, people are biased toward perceiving their own group (an in-group) with more positive characteristics than a group of "others" (an out-group) (Tajfel, 1974).

In this paper, we pursue a systematic, empirical analysis of how simple manipulations of group dynamics can affect the perception of group members' personal qualities, some of which are not directly related to the behaviors being demonstrated. To accomplish this, we analyze peoples' judgments of a group of simple robotic agents in terms of characteristics such as interestingness, creativity, and sociality. We vary the agents' behaviors in systematic ways to tease apart which elements of their behavior affect these character judgments. In particular, we compare peoples' perceptions of a majority group robot-one member of a group of robots all exhibiting nearly identical behaviors-to their perceptions of a minority group robot-a single robot exhibiting a distinct set of behaviors different from the majority group.

In the last century, Michotte investigated the appearance of animacy by systematically varying the behavior of simple
moving shapes (Michotte, 1963). The current investigation is inspired by this and, in some sense, continues Michotte's work by investigating the characteristics of groups of agents by systematically varying the behavior of the group. Through this work, we hope to contribute a greater understanding of social attributions and the dynamics of groups.

Another benefit of this work is a greater understanding of the perception of robot groups. As technology improves, groups of humans are being joined by increasingly agentic technologies such as medicine-delivery robots in hospitals and package-retrieval robots in warehouses. Research in the field of human-robot interaction (HRI) has begun to examine intergroup relations in terms of including or excluding robots from human in-groups. Studies have shown that people respond more favorably to robots in their in-group than one in an out-group (Eyssel \& Kuchenbrandt, 2011; Kim, Kwak, \& Kim, 2010; Kuchenbrandt, Eyssel, Bobinger, \& Neufeld, 2011; Wang, Rau, Evers, Robinson, \& Hinds, 2009). However, much of this research investigates a single robot interacting with one or more humans. This paradigm is historically reasonable, because social robots have been typically designed for and deployed in single-robot environments. As robots become less expensive and more socially accepted, however, groups of robots may become more common. In these cases, understanding the dynamics of robot groups will be important for robot designers and users.

Few studies have explored intergroup dynamics for groups comprised exclusively of robots. Most importantly, little research exists that addresses the perception of groups of robots that are distinguished solely based on the behavior of the group members. We are interested in how the behavioral categorization of robots into majority and minority groups affects perceptions of group members' characteristics.

In this paper, we describe an experiment that attempts to identify the effect of majority group size on people's characterizations of majority group and minority group robots. We use a basic robot behavior-a simple dance-as the distinguishing feature between groups. In our experimental manipulation, we vary both majority group size (one, three, or seven robots in the majority group versus a single-robot minority group) and the type of dance performed by the minority group robot (same as or different from the majority group dance). We ask participants to rate the robots on a number of characteristics both related to and unrelated to dancing, in an attempt to understand how groups distinguished only by behavior are perceived. Our hypotheses are:

H1 The minority group robot will be rated more highly in in-
dividualistic characteristics (such as "creative" and "antisocial") than a majority group robot. Because its behavior will set it apart from the majority group, we expect the single minority robot to appear more independent than the group of similiarly-behaving majority group robots. We expect to see these characterizations with both positive and negative connotations.

H2 As the majority group size increases, the differences between minority and majority group members will increase. Research has demonstrated a positive correlation between group size and conformity (Tanford \& Penrod, 1984). Although participants in this study are not conforming themselves, they are making a character judgment about the conformity (and conversely, the individuality) of the minority group agent, so we expect this trend to hold.
H3 Dance type will affect how the minority robot is perceived relative to the majority. For instance, a dance in which the minority robot performs the motions one second before the majority group robots will lead to perceptions of the minority robot as the leader.

\section*{Related Work}

In this paper, we ask participants to make judgments of majority and minority group members from an outside perspective; the participant is not a member of either group. In contrast, in-group and out-group studies typically involve the participant being a member of one of the groups. Though we are analyzing majority versus minority groups in this paper, we are still influenced by research about in-group and out-group dynamics, and we briefly summarize the literature here.

Intergroup relations among human groups have been studied for over four decades, and there is a well-established body of literature describing the effects of in- and out-group membership on perception and behavior of people (e.g., (Tajfel, 1982; Turner, Hogg, Oakes, Reicher, \& Wetherell, 1987; Brewer \& Brown, 1998)). People exhibit positive responses to conformity with in-group norms, and they tend to be protective of the in-group stereotype (Castano, Paladino, Coull, \& Yzerbyt, 2002; Christensen, Rothgerber, Wood, \& Matz, 2004). People also tend to view an out-group as having less variety than their in-group (Boldry, Gaertner, \& Quinn, 2007). Similarly, members of a group of size one tend to be stereotyped by their distinguishing characteristic, such as gender (Wolman \& Frank, 1975).

Studies of group pressure between minority and majority groups have found that both groups can be swayed by the other, but that minority groups are particularly susceptible to being overridden by majority opinion (Asch, 1956). The amount of influence exerted by either a majority or a minority group is mostly affected by the size of the competing groups, and is less affected by the task or group characteristics (Tanford \& Penrod, 1984).

While human groups have been studied extensively, less research exists on peoples' perceptions of robot groups. Studies have shown that robots perceived as in-group members
appear more compelling, familiar, reliable, and anthropomorphic (Kim et al., 2010; Kuchenbrandt et al., 2011; Wang et al., 2009). Participants report more positive interactions with an in-group robot than an out-group robot, even when that grouping was established only implicitly by changing the robot's name and country of origin (Eyssel \& Kuchenbrandt, 2011). Participants exhibit greater perception of secondary emotions to in-group virtual agents (Besmann \& Rios, 2012). Responses to in-group membership may depend on culture: one study found that Chinese participants (i.e., from a more collectivist culture) were more comfortable with a robot than US participants (i.e., from a more individualistic culture) when that robot was presented as a strong in-group member (Evers, Maldonado, Brodecki, \& Hinds, 2008).

\section*{Experiment Design}

We used survey responses to videos of groups of dancing robots to investigate how in-group size and out-group behavior affected perceptions of group members. Surveys were completed using an interactive web page provided by SocialSci, an online social science research utility. Respondents were recruited from the SocialSci survey pool, a vetted collection of survey takers similar to Amazon's Mechanical Turk. Respondents were rewarded for their time with SocialSci credits which can be exchanged for gift cards and other rewards (SocialSci, 2012).

In all, 89 respondents were recruited. All survey respondents were over 18 years of age, with a mean age of 29 years (SD 10 years). About half ( \(54 \%\) ) of the respondents were male. Sixty-two percent of respondents identified themselves as white, \(6 \%\) were African-American, \(12 \%\) Hispanic, \(10 \%\) Asian, 2\% Pacific Islander, and 6\% identified as "Other."

\section*{Stimuli}

We employed a \(3 \times 4\) within-respondents design. The two factors in this design are group size (one majority versus one minority robot, three majority robots versus one minority robot, and seven majority robots versus one minority robot) and dance type (similar, unique, follower, and leader). The condition with a single majority robot was used as a control. Participants saw one video from each combination of group size and dance type for a total of twelve videos. We randomized presentation order of these twelve videos. To add to the perception of dancing, each video had an audio track containing one of five songs taken from a current top-40 pop music list. The robots' dance rates were altered with video processing to match the tempo of the current song. Song presentation order was also randomized.

Each video clip was 30 seconds long. Videos were created by stitching together multiple videos of the same robot (for majority-group members) or a different robot (the minoritygroup member) side-by-side (see Figure 1 for a still image from one video). The position of the minority-group robot was randomized across trials. In order to elicit a sense of realism, small delays were added to some majority-group members' videos at random; these delays were 10 frames
(about 160 milliseconds) at maximum, and simply served to strengthen the illusion that the video presented to participants showed several individual robots instead of one robot replicated several times.

To test how behavior affects characterizations of majority and minority group members, we created four distinct dance types. Figure 2 shows these four conditions. Dance types are described as a comparison of the minority robot's dance to the majority group's dance, and they can vary across two dimensions. First, the minority robot's dance can have the same available dance moves (repertoire) or a disjoint set of dance moves than the majority group's dance. Second, the minority robot's dance can have the same sequence of dance moves or a different sequence of dance moves than the majority group's dance. In the similar condition, the minority robot's dance has the same repertoire but a different sequence of moves. In the unique condition, the minority robot's dance has a completely disjoint repertoire. In the follower condition, the minority robot's dance has the same repertoire and the same sequence of moves, but the minority robot's sequence is one second behind that of the majority group members, eliciting the appearance that the minority robot is following the majority robots and struggling to catch up. Finally, in the leader condition, the minority robot's dance again has the same repertoire and sequence as the majority group's dance, but this time it is the minority robot that performs its sequence ahead of the majority robots by one second, in an attempt to simulate a single leader that is being followed by the other robots. These dances were designed to span the range of possible group dynamics, to identify how a robot's behavior with respect to other robots affects peoples' perceptions of it.

\section*{Survey}

The survey consisted of one page for each of the twelve videos and a final page to collect demographic information. Each video page contained a video that was 140 pixels tall and either 222 pixels wide (in the control, one-versus-one condition), 444 pixels wide (in the three-versus-one condition), or 888 pixels wide (in the seven-versus-one condition). Each robot was numbered with an overlay on the video which did not obstruct a view of the robot's motions (see Figure 1).

Following the video stimulus, respondents were asked to rate two of the robots (a random member of the majority group and the sole minority group member) on nine attributes using a seven-point scale:
- How well did robot \(x\) dance?
- How entertaining is robot x ?
- How likable is robot x ?
- How lifelike is robot \(x\) ?
- How interesting is robot \(x\) ?
- How much is robot x a team player?
- How mindless is robot x ?
- How creative is robot \(x\) ?
- How anti-social is robot \(x\) ?

Four survey questions were carefully selected for their positive and negative connotations of conformity or nonconformity. "Team player" and "mindless" represent positive and negative conformist words, respectively. Similarly, "creative" and "anti-social" have positive and negative connotations for non-conformity, respectively. Other survey questions address a range of attributes, including those having to do with the relevant behavior of dancing ("good dancer," "entertaining") and those not specifically related to dancing ("likable," "lifelike," "interesting").

\section*{Results}

In order to investigate our three hypotheses, we conducted within-subjects ANOVAs with group size (one-versus-one, three-versus-one, or seven-versus-one), dance type (similar, unique, follower, or leader), group affiliation (majority or minority), and their interactions as fixed factors, participant as a random factor and each of the nine ratings of the robot as dependent variables.

Dance type showed no statistically significant effect on ratings between minority and majority robots for any of the group size conditions, disproving hypothesis H3. Song type also showed no statistically significant effect, so in our subsequent analysis we combine all trials of a single group size together, regardless of dance type or song.

Hypothesis H1 predicts that the minority group robot will be rated more highly than the majority group robot in individualistic characteristics and less highly than the majority group robot in conformist characteristics. The survey attributes dealing with individualism and conformism are: creative, anti-social, mindless, and team player.

For perceptions of how creative the robot was, the majority group robot was rated lower than the minority group robot \((F(1,2128)=9.148, p<0.01)\). The interaction was also significant \((F(2,2128)=6.973, p<0.01)\). Individual group comparisons using Tukey HSD confirm significant differences between the minority and majority robots in the three-versus-one and seven-versus-one conditions (Figure 3a).

For perceptions of how anti-social the robot was, the majority group robot was rated lower than the minority group \(\operatorname{robot}(F(1,2128)=26.146, p<0.001)\). The interaction was significant as well \((F(2,2128)=8.706, p<0.01)\). Individual group comparisons using Tukey HSD confirm significant differences between minority and majority robots in the three-versus-one and seven-versus-one conditions (Figure 3b).

For perceptions of the robot as a team-player, the majority group robot was rated higher than the minority group robot \((F(1,2128)=158.06, p<0.001)\). The interaction was significant as well \((F(2,2128)=39.67, p<0.001)\). Individual group comparisons using Tukey HSD confirm significant differences between minority and majority robots in the three-versus-one and seven-versus-one conditions (Figure 3c).

No statistical differences were found for mindlessness. Analysis therefore shows significant differences for three of the four attributes dealing with individualism (creative, anti-


Figure 1: A still image from video of the seven-versus-one group size condition, showing the minority group robot (number 4) performing a different dance than the seven majority group robots.
\begin{tabular}{|c|c|c|c|}
\hline Condition Name & Available Repertoire majority minority & \begin{tabular}{l}
Sample Sequence \\
time
\end{tabular} & Explanation \\
\hline similar &  & \begin{tabular}{l}
maj: \\
min: \(-5-3-2\)
\end{tabular} & Same repertoire, different sequence \\
\hline unique &  & \begin{tabular}{l}
maj: \\
min: -
\end{tabular} & Different repertoire \\
\hline follower &  & maj:
\[
\min :
\] & Same repertoire, minority sequence is one second behind \\
\hline leader &  & \begin{tabular}{l}
maj: \\
min :
\end{tabular} & Same repertoire, minority sequence is one second ahead \\
\hline
\end{tabular}

Figure 2: A comparison of the four dance types used for the dance manipulation. Each row represents one dance type. The "available repertoire" column shows a graphical representation of the repertoire for majority and minority groups, where each shape is a distinct move (e.g., rocking back and forth). The "sample sequence" column shows example dances for each group.
social, and team player), partly confirming H 1.
Hypothesis H2 predicts that an increase in group size will increase the impact of being a minority. There were significant differences between the group size conditions on ratings of the team player attribute, positively correlated with group size \((F(1,2128)=14.64, p<0.001)\). Individual comparisons with Tukey HSD confirm significant differences in the minority robot's ratings between the three-versus-one and the seven-versus-one conditions. No other differences were observed across group sizes. Therefore, H2 is only upheld for the team player attribute.

Analysis of the additional attributes reveals statistical differences between majority and minority robot ratings for interestingness and dance ability as well. For perceptions of how interesting the robot was, the majority group robot was rated lower than the minority group robot \((F(1,2128)=\) \(4.128, p<0.05)\). The interaction was marginally significant as well \((F(2,2128)=3.611, p=0.058)\). Individual group comparisons using Tukey HSD confirm significant differences between minority and majority robots in the three-versus-one and seven-versus-one conditions (Figure 3 d ). For perceptions of how well the robot danced, the majority group robot was rated higher than the minority group robot \((F(1,2128)=30.35, p<0.01)\). The interaction was
significant as well \((F(2,2128)=6.22, p<0.05)\). Individual group comparisons using Tukey HSD confirm significant differences between minority and majority robots in the three-versus-one and seven-versus-one conditions (Figure 3e).

We also found a weak but highly significant correlation between positive and negative conformist words, "team player" and "mindless" (Pearson's \(r=0.165, p<0.001\) ), as well as between the positive and negative non-conformist words, "creative" and "anti-social" (Pearson's \(r=0.235, p<0.001\) ).

There were no significant differences in ratings of how mindless, lifelike, likeable, or entertaining the robots were. There was also no significant influence of demographic information such as age, gender, or ethnic background.

\section*{Discussion}

Hypothesis H1 was confirmed in the three-versus-one and seven-versus-one case for creative, anti-social, and teamplayer attributes. Furthermore, we found statistical differences between majority and minority robots for interestingness and how well the robot danced. In the control case-that is, in the one-versus-one condition-there was no statistical difference in ratings for either robot in these five traits, suggesting that the appearance of majority and minority groups elicited these differences in evaluation.


Figure 3: Graphs of participant ratings of minority and majority group robots for each of the group size conditions. The dotted line indicates an average rating on the seven-point scale. For brevity, we only show ratings that achieved statistically significant results (.: \(\mathrm{p}<0.1 ;{ }^{*}: \mathrm{p}<0.05 ;{ }^{* *}: \mathrm{p}<0.01 ;{ }^{* * *}: \mathrm{p}<0.001 ;\) NS: not significant).

We also predicted that majority group size would have an effect on the magnitude of differences between minority and majority robots (H2). For the team-player condition, as the group size increased, the differences between the majority and minority group robots also increased. Therefore, H2 was confirmed for one of the attributes. This result is interesting because it shows that group size can affect the intensity of "otherness" exhibited by the out-group member. Our finding is in line with other psychology research, which reports that group size affects the intensity of pressure to conform to group behavior (Tanford \& Penrod, 1984). This result also suggests that teams are identified not just by coordination of action but by size of population, because a larger majority group increased the difference in team-player ratings between the majority and minority robots. It would be interesting to identify whether there is an upper bound in majority group size for this effect. Future work might also analyze minority group sizes greater than one, to see if the effect of group size on the intensity of "otherness" is mitigated when the minority group grows proportionally to the majority group.

The dance types used in this experiment were carefully designed to account for many possible group interactions: one group leading another, one group following another, the groups having completely different behaviors, or the groups having the same behaviors but in a different pattern. Regardless, we found no significant differences due to dance type, disproving hypothesis H3. This suggests that it may not mat-
ter what the minority group robot is doing, as long as it is not doing exactly what the majority robot does. This supports Asch's majority effect (Asch, 1956), and is an interesting and novel finding in multi-robot interactions.

We chose the phrases "team player" and "mindless" as terms that have positive and negative connotations of conformity, respectively. Similarly, we chose "creative" and "antisocial" for their positive and negative connotations with nonconformity. Results indicate that the out-group robot was perceived as less conformist (team-player) and more nonconformist (creative, anti-social) overall. When comparing positive and negative connotations of those categories, we see weak but highly significant correlations between the two conformist words and between the two non-conformist words. Thus, if participants saw a robot as conformist, they were more likely to rate it highly for both positive and negative conformist words (and similarly with non-conformist words).

In this study, we measured participants' subjective evaluations, but they were not part of either group. The lack of situatedness of the robots, coupled with the fact that respondents did not participate in the task, make it unlikely that participants identified with either the majority or minority groups. Assigning participants to one of the groups might reveal further perceptions relating to group membership, and would allow comparisons to existing research on intergroup dynamics.

This paper investigates a social phenomenon-group dynamics-using controlled, precise stimuli and quantitative
metrics for evaluation. The general methods presented here, such as using recorded stimuli to manipulate a single social behavior within a group, can be useful for exploring many other social phenomena as well. The use of robots as stimuli further benefits researchers by allowing them to leverage the flexibility, precise control, and repeatability of movements, facial expressions, and other socially relevant behaviors, which may be impractical to achieve with human actors.

\section*{Conclusion}

Social psychologists have long known that group membership and size affect characterizations of group members, but there is a dearth of studies on this topic within the field of human-robot interaction. In this paper, we analyze the dynamics of robot groups: how are members of robot majority and minority groups perceived, and how does this perception change with differences in group size? We created majority and minority robot groups based on robot behaviors, by having robots perform simple dances which varied between the two groups in either timing (the minority group led or followed the majority group) or content (the minority group performed a different set of behaviors from the majority group). The minority group always contained one robot, while the majority group size varied from one robot (a control condition) to three robots to seven robots. We found a significant effect of group size on five characteristics: how creative, interesting, anti-social, much of a team player, and good a dancer the robot was. In contrast to this group size effect, there was no significant influence of behavior on any character attributions. This research suggests that group size has a strong effect on the evaluation of robot characteristics.

\section*{Acknowledgments}

Thanks to anonymous reviewers for helpful comments, and to Geoffrey Litt, Caroline Jaffe, and Marty Keil for early exploration of this idea. This material is based upon work supported by National Science Foundation grants number 1139078 and 1117801, as well as Office of Naval Research grant number N00014-12-1-0822.

\section*{References}

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. Psychological Monographs: General and Applied, 70(9), 1-70.
Besmann, A., \& Rios, K. (2012). Pals in power armor: Attribution of human-like emotions to video game characters in an ingroup/outgroup situation. Cyberpsychology, Behavior, and Social Networking, 15(8), 441-443.
Boldry, J. G., Gaertner, L., \& Quinn, J. (2007). Measuring the measures: A meta-analytic investigation of the measures of outgroup homogeneity. Group Processes Intergroup Relations, 10, 157-178.
Brewer, M. B., \& Brown, R. J. (1998). Intergroup relations. In D. T. Gilbert, S. T. Fiske, \& G. Lindzey (Eds.), The handbook of social psychology (Vol. 1 and 2, pp. 554594). New York, NY, USA: McGraw-Hill.

Castano, E., Paladino, M.-P., Coull, A., \& Yzerbyt, V. Y. (2002). Protecting the in-group stereotype: Ingroup identification and the management of deviant ingroup members. British Journal of Social Psychology, 41, 365-385.
Christensen, P. N., Rothgerber, H., Wood, W., \& Matz, D. C. (2004, October). Social norms and identity relevance: A motivational approach to normative behavior. Personality and Social Psychology Bulletin, 30, 1295-1309.
Evers, V., Maldonado, H., Brodecki, T., \& Hinds, P. (2008). Relational vs. group self-construal: Untangling the role of national culture in HRI. In Proceedings of the 3 rd ACM/IEEE international conference on human robot interaction (HRI '08) (pp. 255-262). New York, NY, USA: ACM.
Eyssel, F., \& Kuchenbrandt, D. (2011). Social categorization of social robots: Anthropomorphism as a function of robot group membership. British Journal of Social Psychology.
Kim, Y., Kwak, S. S., \& Kim, M.-s. (2010). Effects of intergroup relations on people's acceptance of robots. In Proceedings of the 5th ACM/IEEE international conference on human-robot interaction (HRI '10) (pp. 107-108). Piscataway, NJ, USA: IEEE Press.
Kuchenbrandt, D., Eyssel, F., Bobinger, S., \& Neufeld, M. (2011). Minimal group-maximal effect? Evaluation and anthropomorphization of the humanoid robot NAO. In B. Mutlu, C. Bartneck, J. Ham, V. Evers, \& T. Kanda (Eds.), Social Robotics (Vol. 7072, pp. 104-113). Springer Berlin / Heidelberg.
Michotte, A. (1963). The perception of causailty. Oxford, England: Basic Books.
SocialSci. (2012, September). Available from http://www.socialsci.org/
Tajfel, H. (1974). Social identity and intergroup behaviour. Social Science Information, 13, 65-93.
Tajfel, H. (1982). Social psychology of intergroup relations. Annual Review of Psychology, 33, 1-39.
Tajfel, H. (Ed.). (2010). Social identity and intergroup relations. New York, USA: Cambridge University Press.
Tanford, S., \& Penrod, S. (1984, March). Social influence model: A formal integration of research on majority and minority influence processes. Psychological Bulletin, 95, 189-225.
Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., \& Wetherell, M. S. (1987). Rediscovering the social group: A self-categorization theory. Cambridge, MA, USA: Basil Blackwell.
Wang, L., Rau, P.-L. P., Evers, V., Robinson, B., \& Hinds, P. (2009). Responsiveness to robots: effects of ingroup orientation \& communication style on HRI in China. In Proceedings of the 4th ACM/IEEE international conference on human robot interaction (HRI '09) (pp. 247-248). New York, NY, USA: ACM.
Wolman, C., \& Frank, H. (1975, January). The solo woman in a professional peer group. American Journal of Orthopsychiatry, 45, 164-171.

\title{
Gesture and Language Production in Communication through Bar Graphs
}

\author{
\({ }^{1}\) Özge Alaçam (alacam@informatik.uni-hamburg.de) \\ \({ }^{1}\) Christopher Habel (habel@informatik.uni-hamburg.de) \\ \({ }^{2}\) Cengiz Acartürk (acarturk@metu.edu.tr) \\ \({ }^{1}\) Department of Informatics, University of Hamburg, Hamburg/Germany \\ \({ }^{2}\) Cognitive Science, Middle East Technical University, Ankara/Turkey
}

\begin{abstract}
Bar graphs and line graphs are commonly used ways of graphical communication. Due to the difference in their perceptual visuo-spatial properties, they facilitate comprehension of different events. Bar graphs are commonly used in the domain of precipitation although the data intrinsically carry information that is averaged over long time spans. In this study, we investigate how the presence of incongruence between consecutive graph pairs influences conceptualization of the represented information about precipitation. For this, we analyzed gestures and verbal descriptions produced by the participants as indicators of event conceptualizations. The results of the experimental investigation reveals that when incongruent graph pairs are presented, the participants show tendency to produce directional gestures that accompany the verbal descriptions of the specific regions represented by one/two bars, indicating that bar graphs presented in consecutive order facilitates comprehension of trend information as well as of discrete entities. Additionally, the presence of incongruence seems to enhance the production of comparative words accompanied with non-directional gestures.
\end{abstract}

Keywords: Gesture production; language production; bar graph comprehension; multimodal communication

\section*{Visualization - Bar and Line Graphs}

The primary goal of visualizing data is to (re-)present them in a format more suitable for using them in thinking, problem solving and communication (This view is taken implicitly or explicitly in most seminal publications on graphs, as well as on visualization during the last decades, see, e.g., Tufte 1983, Kosslyn 1989, 2006, Hegarty 2011). Line graphs and bar graphs are successful means to present data, both in the task of analyzing the data and in the task of communicating the results of data analysis. Communicating visualized data using bars or lines is used extensively in scientific publications, textbooks, magazines and newspapers; Zacks, Levy, Tversky, \& Schiano's (2002) study on the use of graphs in the print media shows that line graphs and bar graphs are the dominant, i.e. most frequently used, types of graphs in addressing non-experts in communication through graphs.

The primary gain in using graphs is not to make individual data points visible but to provide visual access to relations between data points ('xl-yl has a larger \(y\)-value than \(x 2-y 2\) ') or to second-order entities as 'trends'. This advantage can be ascribed to humans' pattern perception processes, in particular visual chunking (see, Shah, Mayer
\& Hegarty, 1999). Beyond these commonalities, there seem to be functional differences between bar graphs and line graphs. Zacks and Tversky (1999) investigate the bar-line message correspondence, which considers the systematic relations between the type of graph used and the type of message intended to be communicated. Zacks and Tversky point to a preferred "use of bar graphs to depict comparisons among discrete data points, and line graphs to depict trends" (p. 1073). On the other hand, participants in their experiments had a strong tendency for relational descriptions (e.g., "A is higher than B ") after comprehending bar graphs and for process-oriented secondorder descriptions as 'trends' (e.g., "X increases from A to B") in the of line graphs (p. 1078). Shah, Meyer and Hegarty (1999) report-with respect to these tendencies-a comparable view, but in presenting their perceptual organization hypothesis they lay an additional focus on Gestalt principles realizable in the graph types in question.

In addition to text-graphics documents, in many professional communication settings as well as in classroom settings, graphs, spoken language, and often gestures, accompany each other forming multimodal communication. In dynamic communication of this type, often recipients have to integrate messages communicated by a sequence of graphs. The present study investigates participants' verbal descriptions of pairs of succeeding bar graphs and the gestures produced during these descriptions. The first graph of each pair depicts averages (monthly precipitation over three decades) whereas the second graph depicts instances (monthly precipitation of a specific year). Due to the average-instance constellation, commonalities and differences, which we regard as 'incongruences', between the graphs play a major role in comprehending the graphs and in following production of verbal descriptions; in this setting the within-the-bar bias (Newman \& Scholl, 2012) did not occur.

\section*{Gesture and Language}

The studies on gesture-language interaction are mainly based on the assumption that concepts are sensorimotor, by emphasizing that they are grounded in physical world and based on perceptual experience (Barsalou, 1999; Garbarini \& Adenzato, 2004). There are several frameworks that investigate gestures from various perspectives, but all of them agree on that gestures rely on spatial representations. According to the GSA framework ('Gesture-as-simulated-
action', Hostetter \& Alibali, 2008), one of the frameworks that focus on how gestures are produced, gestures are byproduct of speech. In particular, linguistic planning involves simulation of visuo-spatial events; this activation during articulation is considered as a source of speech accompanying gestures. Another framework, that is closely aligned with the GSA framework and that focuses on how gesture and language production are integrated is the "Interface Hypothesis" (Kita \& Özyurek, 2003). The preparation for language production requires organization of rich and comprehensive information into small packages that contain appropriate amount of informational complexity within a processing unit. According to the "Interface Hypothesis", this processing unit may correspond to a clause for speech production, and the contents of a representational gesture are affected by the organization of these information-processing units, which are prepared for speech production. Therefore this close relationship between the gestures and language makes gestures an effective tool in the assessment of the reader's conceptualization of event, which is simultaneously described verbally (GoldinMeadow \& Beilock, 2010).

Although the interaction between language and gesture has been investigated for the past several decades in a variety of domains (Goldin-Meadow, 2003; Hostetter \& Alibali, 2008; Hostetter \& Sullivan, 2011; McNeill, 1992; 2005) specific investigations of graph comprehension in interaction with language and gesture, has been one of the scarce topics in the field of multimodal interaction. Gestures and graphical communications are visuo-spatial modalities, and they share similar perceptual visuo-spatial features to convey meaning such as quantity, direction and relations. (Tversky, 2011). Therefore during describing a visualization with an accompanying gesture, the places (or punctual events in the domain of our interest) become "fleeting positions" while marks and forms on the visualization become "fleeting actions" (Tversky et. al., 2009). Following this idea originated from the resemblance between two modalities, the vocabularies of gestures, speech and diagrams can be considered as parallel (Tversky, 2011).

For instance, within the context of communication through graphs, a fluctuating increase in a line graph may be verbally described by the term "increase" and it may be simultaneously accompanied by a gesture that represents the fluctuation in the increase. One of the studies focused on communication through line graphs (Acartürk \& Alaçam, 2012), showed that the perceptual features of the annotation that highlights the event presented in the sub-region of the graph (e.g., a graphical cue such as an arrow) have an effect on the conceptualization of the event, and this effect is observable in the gestures produced by graph readers. The results of this study indicated that in order to emphasize processes (e.g., increase, decrease) more vertical and diagonal gestures were produced by humans, whereas more pointing gestures were produced for emphasizing punctual states (e.g., a peak).

To sum-up, gestures can be used as a tool to assess how the graph reader interprets the graph and conceptualizes the event represented by the graph, because gestures provide additional information which is aligned with the visuospatial aspects of the graphical communication. Therefore gesture analysis helps to detect the hard-to-encode information and disambiguates, that are generally highlighted with the presence of accompanying gestures.

In the domain of bar graph comprehension and in communication through bar graphs, differences in gesture production are expected due to perceptual properties of bar graphs that contrast to those of line graphs. Bar graphs enhance comprehension of discrete events, since each bar on the graph perceptually corresponds to a single entity in the domain of discourse, while line graphs facilitate comprehension of trends. On the other hand, although the perceptual properties of graphs are crucial in the conceptualization, the comprehension is still highly dependent on their conceptual properties too (Zacks \& Tversky, 1999). Our goal in this study is to investigate the conceptualization of events that belong to average data (in the domain of precipitation, which is frequently represented with bar graphs), by analyzing the gestures produced during the description of the represented events.

We hypothesize that relations between events of the same domain that are represented with the same graph type (bar graphs) may be conceptualized differently when a perceptual change regarding small areas on the graph (in the case of incongruence) is introduced. In our experimental design, comparisons between regions of two consecutive graphs are required, rather than a comparison between two discrete entities in one graph. Therefore, in addition to discrete comparisons, trend evaluation may also play a major role during comprehension. Moreover, the differences in event conceptualization are examined by analyzing the speech accompanied gestures produced by the graph readers during the verbal description.

\section*{Experiment}

\section*{Participants, Materials and Design}

Twelve participants (university students at the Department of Human Computer Interaction, University of Hamburg, 4 female, Mean age \(=24.2, \mathrm{SD}=3.21\) ) participated in the study. The experiment was conducted in German, the native language of all participants.

Each participant was presented six precipitation graph pairs (two additional pairs of the graphs were employed for the familiarization part). The graphs represented average precipitation data of various cities. In the first graph of each graph pair, a bar graph that represented the monthly precipitation data average for the time period between 1970 and 2011 was shown for 10 seconds on a computer screen (the data were retrieved from Turkish State Meteorological Service). After the graph disappeared, the participant was asked to present a single-sentence verbal description of the first graph to a hypothetical audience. After then, the second
graph of the graph pair was presented. The second graph represented monthly precipitation data for the specific year (2011 for all stimuli) for the same city presented before, again for 10 seconds. The participant was asked to give a verbal description by taking into account both the first graph and the second graph. This procedure was applied for 6 graph pairs. The first graph in each graph pair was always the representation of the monthly precipitation data averaged over 1970-2011, whereas the second graph was always the representation of the monthly precipitation data for 2011 only (see Figure 1 and Figure 2). Participants' spontaneous gestures for 6 precipitation-graph pairs were video-recorded. The participants were informed only about producing verbal descriptions, therefore the gestures produced by the participants were spontaneous gestures.
The second within-subject condition in the experiment design was the congruency between the two graphs in each graph pair. In three graph pairs, the second graph was the same as the first graph, thus leading to a congruent graph pair (Figure 1). In the other three graph pairs, the second graph involved deviant bars (compared to the first graph), thus leading to an incongruent graph pair (Figure 2). The deviant bars were obtained by either increasing or decreasing the value of two/three bars drastically. The motivation for testing the congruency effect was to investigate how conceptualization differed when the congruency between the two related stimuli was systematically changed, even in the same domain and same graph type.


Figure 1: Sample graph for the average data (left) and the data for "instance" year with congruent graph (right)


Figure 2: Sample graph for the average precipitation data (left) and the data for "instance" year with anomaly in the distribution (right)

\section*{Coding}

The main experiment session consisted of six pairs of stimuli, all presented to each participant. Twelve
participants produced 144 sentences in verbal descriptions, 547 time period phrases in the sentences and 165 gestures that accompanied the verbal descriptions.

Gesture Annotation. The coding scheme was based on both McNeil's (2005) semantic gesture classification and syntactic features. The ANVIL software tool was employed for gesture annotation. In the first classification, the gestures were categorized according to their semantic classifications, such as beat gestures and representational gestures. Then each representational gesture was classified in terms of its directionality: non-directional, and directional (vertical/diagonal/horizontal). According to this classification, the hand movements conducted in small space without having any directed trajectory were categorized as non-directional gesture, whereas the hand movements with aimed trajectory on the air were classified as directional gestures.

Spoken Language Transcription. The sentences produced by the participants were transcribed and then the parts of the sentences were segmented into phrases. After this process, the phrases, which referred to temporal information on the graph, were classified into two categories in terms of the size of time interval. The time phrases that referred to multiple bars (such as in "summer" and "towards to winter") were classified into the "longterm" category. The second group covered the time phrases for specific time intervals (such as "in May" or "in July and August"). Finally, the phrases that referred to the previous graph in the comparative context were classified into the "comparatives" category.

\section*{Results}

The results revealed similar time spent for the description for the overall-data graph (i.e., the graph that represented monthly data averaged over 1970-2011, \(M=30.2\) seconds, \(S D=10.6\) ) and for the specific-year graph ( \(M=30.6\) seconds, \(S D=11.3 ; t=-0.38, p>.05\). As for the congruency, the participants spent similar time both for the congruent graphs ( \(M=28.4\) seconds, \(S D=10.9\) ) and for the incongruent graphs ( \(M=32.8, S D=11.7 ; t=-0.93, p>\) .05). Sample pairs of participants' verbal descriptions are presented in Table 1 and Table 2 below.

Table 1: Sample description for a congruent graph (translated from German)

\footnotetext{
Precipitation averaged over 30 years:
Looking at the past 30 years in Antalya there was almost no rain in the months of summer, but instead very very much in winter, it falls and then rises from the winter to the summer very strongly, in August I believe no rain at all, in the adjacent months only very very little.
}

\section*{The specific year:}

Also in Antalya 2011 reflects the past 30 years, because here we also have relatively little rain in the summer and in contrast very much in the winter and it is a quite steady decline and increase in the months in between.

Table 2: Sample description for an incongruent graph (translated from German)

\section*{Precipitation averaged over 30 years:}

In this graph, again in June, July and August, in the months of summer we have the least precipitation and it increases from August to September/December and January, where the highest point is reached and from January it decreases again slightly until it reaches the lowest point.

The specific year:
In this graph it is striking, that does not look like the average at all, because at the point where the lowest point should be we now have a little deflection upwards with much precipitation and also have less precipitation than expected in the months where a lot of rain falls.

In order to understand the underlying differences and similarities induced by the congruency, the speech parts accompanied by the gestures were focalized. The gestures were classified according to temporal information ("specific time" and "long term time" interval) referred in the accompanied speech parts as explained in the "Coding" section. Eight of 12 participants ( 2 female, Mean age \(=\) 24.3, \(S D=0.98\) ) produced gestures during verbal description of the graphs. Five of those eight participants produced representational gestures \((\mathrm{N}=146)\) classified according to scheme presented above. Two coders analyzed and classified the data. Interrater reliability between coders was calculated by Cohen's kappa. The results revealed an agreement value of .77. According to Landis and Koch (1977), a value above .61 indicates substantial interrater agreement. The results of Chi-square test revealed that during the congruent graph description, the gestures accompanied to "specific time" phrases ( \(\mathrm{N}=25\) ) were observed more than that for "long term" phrases ( \(\mathrm{N}=7\) ), \(\chi 2(1)=10.1, \mathrm{p}<.05\). On the other hand, in the description of the incongruent graphs, similar usage of gesture accompanied "specific" ( \(\mathrm{N}=30\) ) and "long-term" phrases \((\mathrm{N}=20)\) was observed \((\chi 2(1)=2.0, \mathrm{p}>.05)\). The production of non-directional and directional gestures were similar within congruency conditions, see Table 3.

Table 3: Number of gestures classified w.r.t. temporal information (NDir: Non-Directional, Dir.: Directional)
\begin{tabular}{ccccc}
\hline & \multicolumn{2}{c}{ Congruent } & \multicolumn{2}{c}{ Incongruent } \\
\cline { 2 - 5 } & NDir. & Dir. & NDir. & Dir. \\
\cline { 2 - 5 } Specific & 10 & 15 & 19 & 11 \\
Long-Term & 4 & 3 & 9 & 11 \\
\hline
\end{tabular}

The overall results that focuses on the difference between overall and instance graphs showed that the participants tend to produce the same amount of gesture for the first stimuli corresponding to overall precipitation amount for 30 years \((\mathrm{N}=64)\) and for the second stimuli corresponding to specific year \(2011(\mathrm{~N}=82)\), \(\chi 2(1)=2.22, \mathrm{p}>.05\). Additionally, the number of non-directional gestures ( \(\mathrm{N}=78\) ) and directional gestures ( \(\mathrm{N}=68\) ) produced during the course of verbal descriptions were similar. see Table \(4(\chi 2(1)=\) \(.65, \mathrm{p}>.05)\).

Table 4: Number of gestures produced during the verbal description for "Overall" and "Instance" Graphs (NDir: Non-Directional, Dir.: Directional)
\begin{tabular}{ccccc}
\hline & \multicolumn{2}{c}{ Congruent } & \multicolumn{2}{c}{ Incongruent } \\
\cline { 2 - 5 } & NDir. & Dir. & NDir. & Dir. \\
\cline { 2 - 5 } Overall & 18 & 18 & 18 & 10 \\
Instance & 14 & 18 & 28 & 22 \\
\hline
\end{tabular}

Since the congruency is always presented in the second stimulus ("Instance" graph), more detailed analysis was conducted on the scores of second stimulus (see Table 4). The results of a Chi-Square test, conducted to compare the overall number of gesture accompanied time phrases across different congruency groups, showed that more gestures were produced during incongruent graph description ( \(\mathrm{N}=\) 50) than that during congruent graph description ( \(\mathrm{N}=32\) ), \(\chi 2(1)=3.95, p<.05\). On the other hand, the results of the test, which compared the number of directional gestures ( N \(=40)\) and non-directional gestures \((\mathrm{N}=42)\) that accompanied the phrases, revealed no significant difference, \(\chi 2(1)=.05, \mathrm{p}>.05\). Similarly, there was also no difference within the incongruent graphs \((\chi 2(1)=.72, \mathrm{p}>.05)\) and congruent graphs \((\chi 2(1)=.50, \mathrm{p}>.05)\) in terms of the directionality of the gesture. However, in the description of the incongruent graphs, the participants produced more directional gestures compared to their previous description about overall data, indicating that incongruence on the data had a positive effect on the directional gesture production, \((\chi 2(1)=4.5, \mathrm{p}<.05\), while it had no effect on the production of non-directional gestures (see Figure 3). For the description of the congruent graphs, no such a significant difference in the production of non-directional and directional gestures was observed.


Figure 3: Gestures that accompany the description of the overall graph and the incongruent graph

More detailed analysis on the sub-regions with the incongruent data revealed the source of the increase in the directional gestures. The results of Chi-square test showed that for the description of the incongruent regions, more directional gestures \((N=12)\) were observed than nondirectional gestures \((N=3), \chi^{2}(1)=5.40, p<.05\). See Table 5 for the examples of produced sentences and Figure 2 for the corresponding graphs. These regions represented with one or two bars on the graph, their descriptions were accompanied with directional gestures.

Table 5: Directional Gestures that accompany the descriptions for the regions that present incongruence (translated from German)

\section*{Overall Rainfall over 30 years (see Figure 2 - Left for the corresponding graph):}

Ok, so here we have seen in the first months the rainfall was quite ... (non-directional), while in the months of summer relatively fast, quite low at about 20 and within the year the rainfall increased really significantly to about 30 I assume (directional).

\section*{The specific year (see Figure 2 - Right for the corresponding graph) :}

We have seen, that in this year the rainfall came off a little more steady (horizontal). They have decreased from month to month and overall in the first months came off smaller (diagonal) in general but in the months of summer the rainfall increased a little, but not so much, as now the rainfall in the former months have won (non-directional), especially not this time relatively steady decreasing from the average rainfall (diagonal) and in the end again there was a little increase.

Additionally, the verbal data that belongs to twelve participants (regardless of the accompaniment of the gesture) were also analyzed in order to examine the use of comparatives across the two congruency conditions. For each participant, the number of descriptions that referred to the overall graph at least once was counted for each congruency condition. The analysis showed that the participants tend to refer to the first graph in the graph pair, the overall precipitation graph, more in the description of the incongruent graph \((M=2.4, S D=1.0)\) than in the description of congruent graph \((M=1.9, S D=1.1), Z=\) 2.12, \(p<.05\), indicating that the incongruence between the overall graph and the specific-year graph enhanced the production of the comparative phrases. A similar pattern was observed for the comparatives accompanied by gestures: more comparative phrases in the incongruent graph description \((N=14)\) were accompanied by a gesture than in the congruent graph description \((\mathrm{N}=5), \chi^{2}(1)=\) 4.26, \(p<.05\). Additionally, the comparative speech parts are mainly accompanied with the non-directional gestures \((\mathrm{N}=14), \chi^{2}(1)=5.56, p<.05\).

\section*{Discussion}

The goal of the experimental investigation was to analyze the role of congruency between two graphs in a set of graph pairs in conceptualization of bar graphs. The first graph in the graph pair was always a monthly representation of 30 year average of precipitation data. The second graph was always a monthly representation of precipitation for a specific year. The results were analyzed in terms of the analysis of verbal descriptions of the participants, as well as the gestures produced by the participants. The results of the experimental investigation revealed that, in general, the participants spent similar amount of time to describe the congruent and incongruent situations with respect to overall graph (i.e., the graph which represented monthly precipitation data averaged over 30 years). On the other hand, a more frequent use of comparatives was observed in the incongruent condition during the course of the description of the second graph in the graph pair. This finding indicates that the participants noticed the difference between the overall graph and the specific-year graph and they found this anomaly worth mentioning in their verbal descriptions. Moreover, the comparatives were accompanied by non-directional gestures that aimed at referring to the previous graph in the graph pair.

As for the production of gestures, there was no difference in the type of the gestures that accompanied the speech parts. However, while gestures during congruent graph's description mainly were correlated with specific time phrases, during incongruent graph's description, gestures for "specific" and "long term" phrases were similar. This may indicate that unlike congruent graphs' description, description of two relational but incongruent events represented with bar graphs requires "as-a-whole" comprehension of the events as well as focusing on the specific regions of the graphs. Additionally, the number of gestures produced during incongruent graph description was higher than that during congruent graph description. Furthermore, when the incongruence was presented, differences in the event description between the incongruent graph and the overall graph were observed in the production of directional gestures: the participants produced more directional gestures for the specific-year graph compared to the overall graph, whereas there was no significant difference in the production of non-directional gestures. The increase in the number of directional gestures was considered as a likely indicator of a different conceptualization. Therefore, a more detailed analysis was conducted on the small region in the graph where the incongruence was presented. In the descriptions of those regions, the graph readers tended to use more directional gestures, indicating that those regions were interpreted as a trend, although those regions referred to a specific time period on the bar graph representation.

\section*{Conclusion and Future Work}

In this study, we investigated the conceptualization of events by focusing on the gesture production and verbal descriptions in the precipitation domain represented by bar graphs. Although the previous research on graph comprehension provides evidence that bar graphs are preferred to emphasize discrete entities, rather than trends, experts in specific domains, in our case meteorology, frequently use them. As the current study demonstrates as well, bar graphs are highly effective to communicate trends. The specific regions, where the incongruence is presented, are conceptualized as trends and the descriptions are accompanied by directional gestures. The perceptual properties of bar graphs that emphasize the entities may be helpful to catch the incongruence, but it also seems that the events are interpreted as "processes", similar to typical comprehension of the events represented by line graphs (Zacks \& Tversky, 1999). In order to understand the underlying mechanism in more detail, our future research will address the preference of the terms used to emphasize two different events, "process" such as increase or fluctuating and "state" such as peak, maximum, and their co-existence with the gestures in the case the congruency was systematically changed. Moreover, applying same experimental design with line graphs will also shed light into the effect of graph type on the conceptualization of the event that requires extrinsically comparison and also requires intrinsically trend evaluation.

Furthermore, the analysis of gestures seems as an effective tool to assess the graph reader`s comprehension and to obtain the important aspects considered as worth to mention in verbal descriptions. In addition to the rich data provided by verbal descriptions, the gestures point out the hard-to-encode information and conceptually salient points, as well as perceptually salient regions and entities of the graph.

Acknowledgments. We thank our student assistants Lena Andreeßen and Neele Stoeckler for their valuable effort. The study reported in this paper has been partially supported by DFG (German Science Foundation) in ITRG 1247 'Cross-modal Interaction in Natural and Artificial Cognitive Systems’ (CINACS).

\section*{References}

Acartürk, C., Alaçam, Ö. (2012). Gestures in communication through line graphs. In N. Miyake, D. Peebles, \& R. P. Cooper (Eds.), Proceedings of the 34th Annual Conference of the Cognitive Science Society (pp. 66-71). Austin, TX: Cognitive Science Society.
Barsalou, L.W. (1999). Perceptual symbol systems. Behavioral and Brain Sciences, 22, 577-660.
Garbarini, F., \& Adenzato, M. (2004). At the root of embodied cognition: Cognitive science meets neurophysiology. Brain and Cognition, 56, 100-106.

Goldin-Meadow, S. (2003). Hearing gesture: How our hands help us think. Cambridge, MA: Harvard University Press.
Goldin-Meadow, S. \& Beilock, S. L. (2010). Action's influence on thought: The case of gesture. Perspectives on Psychological Science, 5(6) 664-674.
Hegarty, M. (2011). The cognitive science of visual-spatial displays: Implications for design. Topics in Cognitive Science, 3, 446-474.
Hostetter, A.B., \& Alibali, M.W. (2008). Visible embodiment: Gestures as simulated action. Psychonomic Bulletin and Review, 15, 495-514.
Hostetter, A. B., \& Sullivan, E. L. (2011). Gesture production during spatial tasks: Its not all about difficulty. In L. Carlson, C. Hoelscher \& T. F. Shipley (Eds.), Proceedings of the 33rd Annual Meeting of the Cognitive Science Society. Austin, TX: Cognitive Science Society
Kita, S., \& Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. Journal of Memory and Language, 48, 16-32.
Kosslyn, S. M. (1989). Understanding charts and graphs. Applied Cognitive Psychology, 3. 185-226.
Kosslyn, S. M. (2006). Graph design for the eye and mind. New York: Oxford University Press.
McNeill, D. (1992). Hand and mind: What gestures reveal about thought. Chicago: University of Chicago Press.
McNeill, D. (2005). Gesture and thought. Chicago: University of Chi- cago Press.
Newman, G. E., \& Scholl, B. J. (2012). Bar graphs depicting averages are perceptually misinterpreted: The within-the-bar bias. Psychonomic Bulletin \& Review, 19, 601-607.
Shah, P., Mayer, R. E., \& Hegarty, M. (1999). Graphs as aids to knowledge construction: Signaling techniques for guiding the process of graph comprehension. Journal of Educational Psychology, 91, 690-702.
Tufte, E. R. (1983). The visual display of quantitative information. Cheshire CT: Graphic Press.
Tversky, B. (2011). Visualizing Thought. Topics in Cognitive Science, 3(3), 499-535.
Tversky, B., Heiser, J., Lee, P., \& Daniel, M.P. (2009). Explanations in gesture, diagram, and word. In K. R. Coventry, T. Tenbrink, \& J. A. Bateman (Eds.), Spatial language and dialogue (pp. 119131). Oxford, England: Oxford University Press.

Zacks, J., Levy, E., Tversky, B., \& Schiano, D. (2002). Graphs in print. In M. Anderson, B. Meyer, \& P. Olivier (Eds.), Diagrammatic representation and reasoning. (pp. 187-206). London: Springer-Verlag.
Zacks, J., \& Tversky, B. (1999). Bars and lines: A study of graphic communication. Memory \& Cognition, 27(6), 1073-1079.

\title{
Reactivity effects of concurrent verbalisation during a graph comprehension task
}

\author{
Nadia Ali (n.ali@hud.ac.uk) \\ David Peebles (d.peebles@hud.ac.uk) \\ Department of Behavioural and Social Sciences, University of Huddersfield, Queensgate, Huddersfield, HD13DH, UK
}

\begin{abstract}
We report an experiment investigating how concurrent verbalisation during a task can affect performance (a so-called "reactivity" effect). Participants studied three-variable line graphs while (a) concurrently thinking aloud or (b) silently studied the graphs and provided an interpretation once they felt they had understood it. Results showed that verbalisation hindered performance significantly compared to the silent condition. To support the claim that the act of verbalising was hindering performance, competing explanations were also tested, which confirmed thinking aloud as the most likely cause. This contradicts claims by Ericsson and Simon (1993) that thinking aloud reflects but does not affect performance and provides further evidence that verbalising thought processes can hinder performance.
\end{abstract}

\section*{Introduction}

Arguably one of the most important advances to have occurred during the cognitive revolution has been the attempt to develop theoretical justifications and rigorous methods for obtaining information about cognitive processes through the analysis of verbal reports produced during their execution.

Although the use of personal reports to infer mental processes has a long history, the approach has always remained controversial, with critics arguing that data obtained from them may be unreliable or that the methods themselves distort or react with the cognitive processes under investigation. The employment of verbal reports was given a cognitive basis and justification by Ericsson and Simon (1993) and their theory of protocol generation.

As a result of their analysis of the different types of verbalisation, the use of verbal protocol methods is now considered a legitimate approach for tracing thought processes and being a valid source of data about the steps involved in problem solving and decision making (Wilson, 1994). Since the original proposal, the think aloud method has been widely adopted, resulting in a large body of research into the processes underlying decision making, problem solving, text comprehension, diagrammatic reasoning, writing, and various other tasks (Crutcher, 1994).

The method considered by Ericsson and Simon (1993) as being the most valuable and rigorous is the concurrent think aloud method in which experiment participants are asked to simply verbalise their thoughts while carrying out a task. Ericsson and Simon (1993) claim that if appropriate instructions are given and followed carefully, the reports participants provide are an accurate reflection of the thought sequence that would have been followed if participants performed the task silently.

Questions remain however concerning the possible reactivity effects of thinking aloud while performing a task and a number of recent studies have revealed that verbalising while performing a task can hinder performance, challenging Ericsson and Simon's (1993) claim that verbal reports are nonreactive.

This area of research has primarily focused on tasks such as face recognition and insight problem solving where the processes involved in reaching a solution are not accessible to the individual to report (Chin \& Schooler, 2008). Schooler and Engstler-Schooler (1990) investigated whether describing a previously seen face would later hinder participants' ability to correctly recognise the face later. They found that compared to a control group who did not describe the face, those who did performed significantly worse in the recognition test. Schooler and Engstler-Schooler (1990) called this effect "verbal overshadowing", proposing that verbal overshadowing occurs when attention is directed to information that can easily be verbalised and so eclipses information that cannot easily be put into words.

Ericsson and Simon (1993) have challenged findings such as these arguing that the method employed did not adhere to guidelines of how protocols should be elicited. They distinguish between different techniques employed to elicit verbalisations-when participants are asked only to report their thoughts (Type 1 verbalisations) and when participants are asked to explain them (Type 3 verbalisations). Numerous studies (e.g., Wilson and Hodges (1992); Wilson, Hodges, and LaFleur (1995)) have reported reactivity effects but have required participants to provide a reason for their decision (e.g., "why do you prefer this painting over the other one?") which would elicit Type 3 verbalisations, a technique Ericsson and Simon accept is prone to reactivity effects. This is because when researchers ask "why?" questions, participants are required to process information which they would normally not need to, thus altering their thought processes and making the method susceptible to reactivity effects (Ericsson \& Simon, 1993). Therefore, there are only a handful of studies reporting reactivity effects when employing this method which adhere to the criteria outlined for eliciting valid protocols.

Despite this, there is a growing consensus that under certain circumstances, employing the verbal protocol method may result in reactivity effects (e.g., Schooler, Ohlsson, and Brooks (1993); Wilson (1994)). However, studies conducted to investigate this research question have focused on tasks
where verbal overshadowing of information is likely to occur (e.g., insight problems). Although this research is useful for identifying particular instances in which the think-aloud method may be susceptible to reactivity effects, the question remains whether these results generalise to tasks where information is more readily available for verbalisation. There is a need therefore for a greater range of tasks to be tested to determine whether reactivity effects are limited to tasks where information may be difficult to verbalise, especially considering the growth in areas this method has been applied to (Wilson, 1994).

In addition, it has been assumed that it is the act of thinking aloud itself which results in reactivity effects. There is the possibility of an alternative explanation however. When employing the think aloud method, the experimenter must be present with the participant, which may affect performancethe widely investigated "social facilitation/inhibition effect" (Zajonc, 1965; Rosenthal, 1976; Huguet, Galvaing, Monteil, \& Dumas, 1999).

One study which potentially indicates this could be the case was conducted by de Vet and de Dreu (2007), who studied the effects of concurrent verbalisation on creativity in a group setting. They found that thinking aloud impaired performance, particularly in individuals who were sensitive to other people's opinions of them. Although the authors concluded that the presence of others played a role in the performance impairments, it is difficult to generalise these results because the large groups used in the study are not typical of the scenarios used in the majority of verbal protocol studies.

\section*{Current experiment}

The aim of this study is twofold. Firstly, we seek to determine whether reactivity effects found in the literature are due to the demands of thinking aloud or whether potential competing explanations (e.g., experimenter presence) could account for this effect. Secondly, we also wish to investigate whether reactivity effects are limited to the types of tasks investigated in previous verbal overshadowing experiments by requiring people to think aloud while performing a task in which information is readily available for verbalisation.
To do this we employed a graph comprehension task for a number of reasons. First, in such tasks the information being processed is readily available at all times, thereby reducing the burden on working memory and freeing up resources for the interpretive task (Pinker, 1990). Second, previous research into graph comprehension employing verbal protocols (e.g., Ratwani, Trafton, \& Boehm-Davis, 2008) has demonstrated that such methods are able to provide a reliable trace of the problem solving processes undertaken by users.

Finally, the graphs we employed, although widely used in statistics to depict relationships between more than two variables, are relatively simple and constrain the interpretative processes available to users. Evidence for this assertion is provided by Halford, Baker, McCredden, and Bain (2005) who manipulated the number of graphically displayed statis-
tical interactions participants were required to process. They found performance for \(2 \times 2\) problems (the type employed in our task) was near perfect but a steep drop in performance emerged when the graphical representation depicted 3 or 4 way interactions; consistent with processing capacity constraints.

Based on these criteria, one might expect no effect of concurrent verbalisation to be found in this task. However, in a previous study in which we compared graph comprehension assessed by written and verbal reports, we found the written interpretation to be superior in terms of accuracy and detail (Ali \& Peebles, 2011). These findings did not reveal whether the differences were a result of a facilitation produced by the act of writing or a detriment from verbalising. The previous study laid open the possibility that this task may be susceptible to reactivity effects. Therefore this study will attempt to determine whether this is the case.

\section*{Assessing potential reactivity effects in a comprehension task}

To measure reactivity effects, the output from thinking aloud is compared to that of a "silent" condition using dependent measures such as number of correct responses and this is the method adopted in this study. In problem solving tasks the output of the silent condition may be simply a solution to the problem, e.g., \(29 \times 4=116\). In a graph comprehension task however, the output is a series of statements expressing the participants interpretation of the data depicted.

In the think-aloud condition this will result in participants verbalising their interpretation of the graphs until they complete the task. If reactivity effects are not an issue then performance will not differ between the think-aloud and silent condition, i.e., the demands of verbalisation will have no effect on the ability of participants to successfully apply the processes involved in graph comprehension. If however performance is superior in the silent condition to the think aloud condition then the act of verbalising is interfering in the processes involved in graph comprehension.

However, the silent and think-aloud condition is not comparable with these types of tasks because the silent condition involves two stages: an initial silent stage in which the participant constructs the interpretation and a second stage where this interpretation is reported to the experimenter. As this task is split into two stages it could be argued that improvement in performance could occur for a number of reasons other than remaining silent. For example, being explicitly required to communicate understanding to someone else could perhaps result in an improvement.
However, this effect can be balanced by including a further control condition where the second stage of the silent condition is incorporated into the think aloud condition. Therefore, in order to test whether communicating understanding affects performance, a third "summary" condition was included. If it is the act of communicating understanding (and not performing the task silently) which alters performance this condition
will be on par with the silent condition. If however the findings are similar to the think aloud condition then the silent and think aloud condition are comparable. This condition acted as a further control condition allowing for comparisons between the silent and think aloud condition.
Finally, the fourth condition tests any potential influence of experimenter presence on performance by including a "solitary" condition. These manipulations result in three conditions where participants are required to think aloud throughout the task and one condition where participants remain silent. If it is the demands posed by verbalisation resulting in reactivity effects performance should be superior in the silent condition than the other three conditions tested.

\section*{Method}

\section*{Participants}

Sixty undergraduate psychology students (41 female, 19 male) from the University of Huddersfield were paid \(£ 5\) (approximately \(\$ 8\) ) in grocery store vouchers to take part in the experiment. The age of participants ranged from 18.1 to 29.7 years with a mean of 22.2 years \((S D=2.1)\). The participants were in their first year of a three year psychology degree and were randomly allocated to the experiment conditions.

\section*{Design}

The experiment was an independent groups design with four between-subject variables: whether participants were in the think aloud, silent, solitary or summary condition. 15 participants were allocated to each of the graph conditions.

\section*{Materials}

The stimuli used were six three-variable line graphs depicting a wide range of (fictional) content. The graphs were generated using the PASW Statistics software package (produced by SPSS Inc.). Stimuli were printed in colour (with the levels of legend variable in blue and green) on white A4-sized paper. Examples of the stimuli used are depicted in Figure 1. The variables in the graphs were chosen so that no prior knowledge of the domain or relationships would influence interpretation.

\section*{Procedure}

In the first think-aloud control condition participants were informed that they were to be presented with a sequence of six three-variable line graphs and that their task was to try to understand each one as fully as possible while thinking aloud. The nature of the task was further clarified by telling participants that they were being asked to try to understand the relationships between the variables (rather than simply describing the variables in the graph), to try to comprehend as many relationships as possible, and to verbalise their thoughts and ideas as they did so. During the experiment, if participants went quiet, the experimenter encouraged them to keep talking. If participants stated that they could not understand the graph, it was suggested that they attempt to interpret the parts
of the graph they could understand. If they still could not do this, they were allowed to move on to the next trial.

In the second silent condition participants were informed there were two stages to the task. In the first "quiet" stage they could take as long as they wanted to understand the graph they were viewing as much as possible. In the second "talking" stage they were required to tell the experimenter what they had understood about the graph.

In the summary condition participants were instructed that the experiment consisted of two stages-in the first "think aloud" stage they were to think aloud whilst interpreting the graph. In the second "talking" stage they were to tell the experimenter what they had understood about the graph.

In the solitary condition instructions were identical to the think-aloud condition except participants were told they would be left alone throughout the experiment but it was important they remember to think aloud throughout the task.

The instructions were designed to be consistent with Type 1 verbalisations, where participants are required to think aloud throughout the task, but not explain or justify the statements they made. According to Ericsson and Simon (1993) eliciting protocols in this manner should result in no reactivity effects.

Stimuli were presented in random order and all participants were informed that there was no time limit to the task. Verbal protocols were recorded using a portable digital audio recorder.

\section*{Data analysis}

The verbal protocols participants produced while interpreting the graph were transcribed and their content analysed. Only statements in which a sufficient number of concepts could be identified were included for analysis. For example, the statement "low nitrogen levels have no effect on maize yield whether plant density is sparse or compact" was included whereas "low nitrogen affects. . . um. . . I'm not sure" was not. Data analysis was conducted according to the procedure and criteria employed in our original study (Peebles \& Ali, 2009; Ali \& Peebles, 2013). For each trial, the participant's statements were analysed against the state of affairs represented by the graph. If a participant made a series of incorrect statements that were not subsequently corrected, then the trial was classified as an incorrect interpretation. If the participant's statements were all true of the graph or if an incorrect interpretation was followed by a correct one, however, then the trial was classified as a correct interpretation. An example of a correct interpretation for the line graph in Figure 1a is "Whether nitrogen level is low or high when plant density is sparse, maize yield is two. When plant density is compact for low nitrogen level, maize yield is still at two but this increases to seven when nitrogen level is high".

In addition to this trial-level performance analysis, we also analysed the nature of the errors made in incorrectly interpreted trials. When participants made an erroneous interpretation that was not subsequently corrected, in addition to classifying the trial as an incorrect interpretation, we coded


Figure 1: Two of the six line graphs used in the experiment.
the type of error against the trial. As these graphs depict a relationship between three variables, if participants failed to incorporate all three variables into their interpretation the trial was coded as an error. The nature of the fault was categorised according to which of the variables had been ignored or misrepresented or whatever other error had occurred. Errors followed a similar pattern to the original experiment. An example of an incorrect interpretation for the line graph in Figure 1a is "When plant density is sparse, nitrogen levels remain low. When plant density is compact, nitrogen levels increase". In this instance the graph viewer is ignoring the dependent variable, maize yield. Verbal protocol evidence revealed participants were unable to provide an interpretation incorporating all three variables. One participant providing this interpretation stated "I don't understand how maize yield fits into it. I can understand the graph if I focus on plant density and nitrogen level" and then proceeded to ignore the dependent variable. The occurrences and explanations for why these errors occur are explained in greater depth in Ali and Peebles (2013).

In this way, each participant's trials were coded as being either correctly or incorrectly interpreted. The verbal protocol for each trial was initially scored as being either a correct or an incorrect interpretation by the first author and a sample (approximately 20\%) of trials were independently coded by the second author. The level of agreement between the two coders was approximately \(90 \%\). When disagreements were found, the raters came to a consensus as to the correct code.

\section*{Results}

Figure 2 displays the number of correct trials in each verbal protocol condition. The silent condition resulted in a higher
number of correct trials compared to the other three conditions. A comparison of the number of correct trials between the think aloud, solitary, silent and summary conditions revealed that the silent condition resulted in a significant increase (Kruskal-Wallis \(H=7.93, d f=3, p<.05\) ) in the number of correctly interpreted trials (mean rank \(=40.83\) ) compared to the think aloud (mean rank \(=24.60\) ), solitary ( mean rank \(=27.0\) ) and summary ( mean rank \(=29.57\) ) conditions.

Three post-hoc Mann Whitney U tests (with alpha levels Bonferroni adjusted to .017 ) revealed the significant difference to be between the silent condition and the think-aloud condition ( \(p=.005\) ), but not the solitary condition ( \(p=.713\) ) nor the summary condition ( \(p=.595\) ).

\section*{Discussion}

The results of this experiment reveal that participants who attempted to verbalise their interpretation of graphs were significantly less likely to provide a correct interpretation than subjects who interpreted the graphs silently before verbalising their interpretation. Additional control conditions revealed that it was not experimenter presence (the solitary condition) or the act of communicating understanding to someone else (the summary condition) which resulted in the performance differences between the think aloud and silent condition. These results definitively demonstrate that verbalisation results in reactivity effects; in this case a detriment in observed performance.

Although previous research has found that requiring participants to think aloud can result in reactivity effects, these findings have been challenged based on how the method was employed. For example, Cook (2006) required participants


Figure 2: Mean number of correct trials for the experiment conditions. Error bars indicate standard error.
to solve a series of algebra tasks with problems presented by computer in the silent condition but with cards in the verbalisation condition. This introduced a potential confound of verbalisation condition and stimulus format.

In a recent meta-review, Fox, Ericsson, and Best (2011) identified 95 studies employing verbal protocols. Studies were excluded from the meta-analysis if they did not include a comparison to a control condition, if findings were considered suspect because of potential confounding variables, or if effect sizes were not reported. Based on this analysis, the authors concluded that "Studies with confounds are common because few studies with verbal report and silent conditions are designed explicitly to test directly for reactivity" (p. 323).

The experiment reported here directly addresses these issues. We carefully followed Ericsson and Simon's guidelines for eliciting protocols and explicitly tested the think-aloud condition by comparing output to a silent condition as well as ruling out potential competing explanations for the difference observed between the think-aloud and silent condition.

These findings provide a strong demonstration that reactivity effects can emerge even when participants are asked only to report their thoughts (Type 1 verbalisations) and are not asked to explain them (Type 3 verbalisations). In addition, the task used in this experiment does not fall into a category where information is difficult to verbalise, demonstrating that reactivity effects are not limited to such tasks.

Although this task did not reveal any effect of experimenter presence, this issue deserves further investigation as tasks in the social psychology literature which can generate self-
presentation concerns may reveal findings which corroborate those of de Vet and de Dreu (2007). The increasing use of the verbal protocol method in the social psychology literature indicates further research is required to establish this method is appropriate for these types of research questions (Wilson, 1994).

Our knowledge of why reactivity effects emerge when employing the verbal protocol method is limited primarily because of the lack of studies explicitly testing for such effects. Based on previous findings, it appears that this effect is most likely due to a number of interacting factors and so such findings will not emerge consistently. Our findings demonstrate reactivity effects occur due to the demands of verbalisation and this effect is not restricted to tasks where information is difficult to verbalise.

One potential explanation which could account for the effect observed in our experiment is a competition for processing resources explanation. Russo, Johnson, and Stephens (1989) argued that the additional demands for processing resources (which occurs when individuals are required to verbalise whilst performing a task) can explain deterioration in performance. In order to deal with additional demands of verbalisation, participants can draw upon any unused resources which are not being employed. When the demands of the task exceed processing resources however, reactivity effects can occur, resulting in a detriment in performance due to the resources being divided between completing the task and verbalising throughout (Russo et al., 1989).

However, it is difficult to predict a priori whether or not
performance will be distorted by the generation of a concurrent protocol. Even when a task adheres to established guidelines for when the think aloud method is appropriate to use reactivity effects can emerge (Russo et al., 1989). Control conditions as standard practice when employing this method would help establish the conditions under which reactivity effects emerge; a necessary precursor for a theory of protocol generation which can account for reactivity effects. This practice would also allow more confidence in findings employing this method.

\section*{References}

Ali, N., \& Peebles, D. (2011). The different effects of thinking aloud and writing on graph comprehension. In L. Carlson, C. Holscher, \& T. Shipley (Eds.), Proceedings of the twentieth annual conference of the cognitive science society. Mahwah, NJ: Lawrence Erlbaum Associates.
Ali, N., \& Peebles, D. (2013). The effect of Gestalt laws of perceptual organisation on the comprehension of threevariable bar and line graphs. Human Factors, 55(1), 183203.

Chin, J. M., \& Schooler, J. W. (2008). Why do words hurt? Content, process, and criterion shift accounts of verbal overshadowing. European Journal of Cognitive Psychology, 20, 396-413.
Cook, J. L. (2006). College students and algebra story problems: Strategies for identifying relevant information. Reading Psychology, 27, 95-125.
Crutcher, R. J. (1994). Telling what we know: The use of verbal report methodologies in psychological research. Psychological Science, 5, 241-244.
de Vet, A. J., \& de Dreu, C. K. W. (2007). The influence of articulation, self-monitoring ability, and sensitivity to others on creativity. European Journal of Social Psychology, 37, 747-760.
Ericsson, K. A., \& Simon, H. A. (1993). Protocol analysis: Verbal reports as data (Revised ed.). Cambridge, MA: MIT Press.
Fox, M. C., Ericsson, K. A., \& Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. Psychological Bulletin, 137, 316-344.
Halford, G. S., Baker, R., McCredden, J. E., \& Bain, J. D. (2005). How many variables can humans process? Psychological Science, 16(1), 70-76.
Huguet, P., Galvaing, M., Monteil, J., \& Dumas, F. (1999). Social presence effects in the stroop task: Further evidence for an attentional view of social facilitation. Journal of Personality and Social Psychology, 77, 1011-1025.
Peebles, D., \& Ali, N. (2009). Differences in comprehensibility between three-variable bar and line graphs. In N. Taatgen, H. van Rijn, J. Nerbonne, \& L. Schoemaker (Eds.), Proceedings of the thirty-first annual conference of the cognitive science society (pp. 2938-2943). Mahwah, NJ: Lawrence Erlbaum Associates.

Pinker, S. (1990). A theory of graph comprehension. In R. Freedle (Ed.), Artificial intelligence and the future of testing (pp. 73-126). Hillsdale, NJ: Lawrence Erlbaum Associates.
Ratwani, R. M., Trafton, J. G., \& Boehm-Davis, D. A. (2008). Thinking graphically: Connecting vision and cognition during graph comprehension. Journal of Experimental Psychology: Applied, 14(1), 36-49.
Rosenthal, R. (1976). Experimenter effects in behavioural research. New York: Irvington.
Russo, J. E., Johnson, E. J., \& Stephens, D. L. (1989). The validity of verbal protocols. Memory and Cognition, 17, 759-769.
Schooler, J. W., \& Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. Cognitive Psychology, 22, 36-71.
Schooler, J. W., Ohlsson, S., \& Brooks, K. (1993). Thoughts beyond words: When language overshadows insight. Journal of Experimental Psychology: General, 122, 166-183.
Wilson, T. D. (1994). The proper protocol: Validity and completeness of verbal reports. Psychological Science, 5(5), 249-252.
Wilson, T. D., \& Hodges, S. D. (1992). Attitudes as temporary constructions. In A. Tesser \& L. Martin (Eds.), The construction of social judgment (pp. 37-65). Hillsdale, NJ: Erlbaum.
Wilson, T. D., Hodges, S. D., \& LaFleur, S. J. (1995). Effects of introspecting about reasons: Inferring attitudes from accessible thoughts. Journal of Personality and Social Psychology, 69, 16-28.
Zajonc, R. B. (1965). Social facilitation. Science, 149, 269274.

\title{
A Question of Timing: The Impact of Label Synchrony on Infants' Categorisation
}

\author{
Nadja Althaus (nadja.althaus@psy.ox.ac.uk) \\ Department of Experimental Psychology, University of Oxford \\ South Parks Road, Oxford OX1 3UD, United Kingdom
}

Kim Plunkett (kim.plunkett@psy.ox.ac.uk)
Department of Experimental Psychology, University of Oxford
South Parks Road, Oxford OX1 3UD, United Kingdom

\begin{abstract}
Recent research into the impact of labelling on infants' visual category formation has led to controversial results, with some findings indicating a beneficial role and others pointing to interference effects in the presence of labels. Here we present an eye tracking study with 12 -month-olds investigating the impact of the label's timing on categorisation. We find that synchronous presentation of words and objects leads to a decreased novelty preference, creating the impression of a dramatic detrimental impact on learning. Asynchronous presentation of the word one second after the image onset does not appear to interfere with processing. Detailed analyses of infants' gaze patterns with respect to object parts reveal that even synchronous labels do not hinder learning but slow down infants' shift from familiarity to novelty preference. Besides offering detailed insight into the effects of labelling on infants' attention our findings offer the potential to reconcile previous contradictory results.
\end{abstract}

Keywords: Categorisation; cognitive development; language development; eye tracking; attention.

\section*{Introduction}

The idea of linguistic influences on cognitive processes has been a heavily debated subject over the past century, with very extreme positions like Whorfian determinism (Whorf, 1956) gradually being replaced by less radical points of view (e.g. Boroditsky, 2001). From a developmental perspective, the question is fundamental: do infants use language, and words in particular, as cues to learn about the complex structure of the world? The almost universal presence of labels in an infant's environment, both in speech directed at the infant and in conversation between adults overheard by the infant, makes the hypothesis that labels may serve as meaningful cues very compelling. Shared labels indicate, after all, that dissimilar looking things may share attributes or function (e.g., a bonnet and a boater may both simply be called a "hat"). However, labels are not always readily identified by infants in their first year of life: language development is a gradual process involving learning about relevant dimensions in rhythm, prosody and phonetics before individual words are segmented from the speech stream and mapped onto referents (Tincoff \& Jusczyk, 1999; Bortfeld, Morgan, Golinkoff, \& Rathbun, 2005; Bergelson \& Swingley, 2012). In spite of this, several studies in the past 20 years have found facilitative effects of labelling on categorisation in pre-linguistic infants between six and twelve months (e.g., Waxman \& Markow,

1995; Balaban \& Waxman, 1997; Waxman \& Braun, 2005; Fulkerson \& Waxman, 2007), and more recently even in infants as young as three months (Ferry, Hespos, \& Waxman, 2010). This work suggests that even infants who are just at the beginning of language development can make use of labels when learning about objects and similarities between them. In fact, Plunkett, Hu, and Cohen (2008) and Althaus and Westermann (in prep.) have demonstrated that labels can serve to guide the formation of category boundaries when the structure of visual space is ambiguous, i.e. labels can cause infants to merge or split visual clusters depending on whether the visual exemplars are encountered with identical or differing labels. However, contradicting results which report "auditory overshadowing" effects in the presence of labels (as well as other auditory stimuli) have also been reported (Robinson \& Sloutsky, 2007), calling into question whether labelling has uniformly beneficial effects. It is as of yet unclear under what circumstances labels can facilitate learning, and what factors may contribute to labels attenuating learning.

From an information-processing point of view, labels provide additional information that may help learning - e.g., by increasing perceived similarity between objects that share labels (Sloutsky, Lo, \& Fisher, 2001), or by highlighting commonalities (Waxman \& Markow, 1995). However, processing this additional signal comes at a cost: attention and processing resources have to be allocated to two modalities rather than one - a feat which may be particularly problematic for young infants. This factor in particular makes it seem likely that the exact circumstances of how labels are encountered will play a role in whether they are going to interfere with, or aid, processing.

Here we explore the hypothesis that whether or not labels and objects can both be processed depends on the timing of the label: if both are presented in exact synchrony, i.e. image and label occur simultaneously, this may impose high cognitive load, and processing in one or both modalities may be attenuated. By contrast, if there is a delay between visual onset and auditory onset, this may allow infants to process both stimuli equally well - simply because some visual object recognition processes will already have been completed by the time the label occurs (Quinn, Westerlund, \& Nelson, 2006; Grossmann, Gliga, Johnson, \& Mareschal, 2009).

Whereas synchronous label onsets have been used in experimental scenarios (e.g., Robinson \& Sloutsky, 2007), delayed labelling scenarios appear to be much more likely to occur in a young child's experience (for example, a caregiver asking "Do you like the ball?" when the child is already attending to the toy). Even though some researchers have claimed that synchrony is beneficial for word-object association (Gogate, Bahrick, \& Watson, 2003), and cross-modal synchrony has been demonstrated to facilitate discrimination of amodal signals such as tempo or rhythm (Bahrick \& Lickliter, 2000), it is likely that synchronous picture-word pairings are unusual and surprising to infants at one year of age. These infants, after all, are at a stage in development where they may have learned some things about words (e.g., they often occur together with their referents, but not generally in synchrony like "causal" sounds, such as a hammer hitting a wall), but are far from being experts at processing speech sounds as phonetic units and mapping them to words.

In order to examine the impact of audiovisual synchrony on the interaction of labelling and categorisation, we familiarised three groups of 12 -month-olds with a novel category either in silence (Silent condition), with labels presented one second after the picture onset (Asynchronous Label condition), or with labels and pictures occurring simultaneously (Synchronous Label condition). In order to gain further insight into how infants process objects, the target category was constructed such that each exemplar consisted of two spatially separate object parts (a shell and a leaf, see Figure 1 for example objects). The shells were highly variable across exemplars, whereas the leaves were quite similar. This allowed us to track infants' encoding of features with different similarity structure. On test, infants were presented with an out-of-category item, as well as an object consistent with the familiarised category. In familiarisation/novelty preference paradigms, novelty preference for the out-of-category object has been established as an indicator of category formation. However, the two out-of-category items occurring in the first two test trials were constructed by replacing just one of the two parts (shell or leaf) with an item that differed from the familiarisation exemplars. We were therefore able to track on a very fine-grained, almost featural, level whether infants responded to novelty. Our hypothesis was that if infants were able to learn the category in silence, but labels (asynchronous or synchronous) interfered with learning, then infants should not exhibit novelty preference in the relevant conditions. A difference between Asynchronous and Synchronous conditions with regard to novelty preference would further indicate that the timing of the label plays an important role in infants' ability to process and integrate both stimuli.

\section*{Methods}

\section*{Participants}

A total of 87 infants participated in this study (mean age: 372 days, 43 girls). Eight additional infants were not included in the analysis due to failure to reach the looking time criterion.

Infants were recruited shortly after birth at the local maternity ward and English was the main language spoken in their home.

\section*{Stimuli}

A novel category was constructed by assembling 11 "objects" from images of a shell, a leaf and a pipe-cleaner in the Gnu Image Manipulation Program (see Figure 1 for example stimuli). Across the different objects, the leaves were very similar, the shells highly variable, and the invariable pipe cleaner served as a connecting limb between these two parts. In addition, three "out-of-category" objects were constructed (see Figure 2): Test object 1 contained a shell consistent with the category, but an inconsistent type of leaf ("novel leaf"), Test object 2 contained a leaf consistent with the category but an inconsistent shell ("novel shell"), and Test object 3 contained a sea urchin and a starfish instead of shell and leaf. All images were depicted against a medium grey background. Objects subtended approximately \(14 \times 10^{\circ}\) visual angle. On test displays, there was a gap of approximately \(5^{\circ}\) visual angle between out-of-category and within-category objects. A recording of the novel label "timbo", pronounced by a female British-English speaker in an infant-directed voice, served as the auditory stimulus in the Asynchronous and Synchronous Label conditions.


Figure 1: Example familiarisation stimuli.


Figure 2: Test objects 1, 2 and 3.

\section*{Procedure}

After a short warm-up phase during which written consent was obtained from the caregiver, infants were seated on the caregiver's lap at 75 cm distance from the eye tracker. A ninepoint calibration sequence was performed up to three times or until all points had been calibrated successfully.

Infants were presented with eight familiarisation images in pseudo-randomised order which were on the screen for 6000 ms each. Four of the familiarisation images appeared on the left half of the screen, and four on the right, in no
predictable order. Every image was preceded by an attention getter, a small animation at the centre of the screen (with a medium grey background) accompanied by an attractive chiming sound. Animation and sound lasted about \(1.5 \mathrm{sec}-\) onds, with the next trial beginning 2 seconds after the onset of the attention getter. In the Asynchronous Label condition, the sound file containing the label "timbo" was played 1000 ms after the picture onset. In the Synchronous Label condition, the label started at picture onset. Familiarisation was followed by three test trials, lasting 10000 ms each. On the test trials, the three test objects described above were paired with one of the three remaining objects from the familiar category. Test trials in all conditions were silent. Infants' looking was recorded using a Tobii eye tracker sampling at 120 Hz throughout the familiarisation and test phase.

\section*{Results}

We first focus on global measures of looking during familiarisation and test (i.e. with respect to whole objects), and then turn to a more detailed analysis of looking directed at individual object parts. In order to analyse gaze patterns with regard to object parts, areas-of-interest (AOIs) were defined to contain the area covered by the images of shell and leaf, respectively, plus a 30 -pixel margin around the image outline (corresponding roughly to the eye tracker's 0.5 degree visual angle accuracy). Recorded gaze data were analysed using custom Matlab code.


Figure 3: Looking time during familiarisation.

Looking time during familiarisation Looking time for each familiarisation trial was calculated as the sum of fixation time falling on the leaf and shell AOIs. Average looking times for Blocks 1 (Trials 1-4) and 2 (Trials 5-8) are shown in Figure 3. A mixed ANOVA with within-subjects factor Block (Block 1, Block 2) and between-subjects factor Condition (Silent, Asynchronous Label, Synchronous Label) revealed a near-significant main effect of Block \((F(1,84)=3.639, p=.06)\). Neither the Block x Condition interaction \((F(2,84)=2.231, p=.114)\) nor the main effect of condition were significant \((F(2,84)=1.76, p=.178)\). Planned comparisons (paired t-tests) showed that infants' looking decreased in the Silent condition \((t(28)=2.864, p=.008)\),
but not in either of the Label conditions (Asynchronous Label: \(t(28)=.46, p>.64\); Synchronous Label: \(t(28)=1.246\), \(p>.22\) ). This is consistent with previous research showing that auditory input causes infants to maintain their looking (e.g. Baldwin \& Markman, 1989; Robinson \& Sloutsky, 2007; Plunkett et al., 2008).

Part-based looking during familiarisation In order to investigate whether synchronous or asynchronous labels affected infants' processing of individual parts during familiarisation, we calculated a mean looking proportion for the "leaf" part by dividing the amount of looking at the leaves accumulated during familiarisation by the total looking time accumulated during familiarisation. These data were subjected to a one-way ANOVA with factor Condition, which revealed no significant effect \((F(2,84)=1.226, p=.299)\). Across all three conditions, infants spent less time looking at the leaves than at the shells, indicating that they were sensitive to the greater variability of the shells (Proportion of looking at leaf, collapsed across conditions: \(M=.33, S E=.01\); \(t(86)=13.1, p<.001)\).


Figure 4: Novelty preference scores in Test 1.

Preferential looking at test (object-based preferences) Object-based novelty preference scores were obtained for all test trials by dividing the amount of looking at the out-ofcategory object by the total looking time accumulated for the trial (within-category and out-of-category object).

Separate one-way ANOVAs on novelty preference scores with factor Condition were conducted for all three test types, which did not reveal any significant differences between conditions (all \(F \mathrm{~s}<1.4, p \mathrm{~s}>.2\) ). However, planned comparisons were conducted to test each group's performance against the chance level of 0.5 . If infants failed to form a category and

Table 1: Novelty preference scores for test trials 2 and 3 in all conditions. * indicates significance at the .05 -level, \({ }^{* *}\) indicates significance at the .005 level, \({ }^{* * *}\) indicates significance at the .0005 level.
\begin{tabular}{lll}
\hline & Test 2 & Test 3 \\
Condition & \(M(S E)\) & \(M(S E)\) \\
\hline Silent & \(.63(.05)^{*}\) & \(.66(.05)^{* *}\) \\
Async. & \(.64(.04)^{* *}\) & \(.65(.04)^{* *}\) \\
Sync. & \(.65(.04)^{* *}\) & \(.68(.04)^{* * *}\) \\
\hline
\end{tabular}
did not discriminate between the two objects, we would expect them to spend \(50 \%\) of their looking directed at each object. By contrast, if they successfully formed a category we would expect them to reliably prefer the out-of-category object.

The results for Test 1 are illustrated in Figure 4. Here, the out-of-category object contained a novel, inconsistent type of leaf and consistent shell. Infants in the Silent and Asynchronous Label conditions exhibited significant preference for the out-of-category object (Silent: \(t(28)=2.679, p=.01\); Asynchronous Label: \(t(28)=4.04, p<.001)\). Infants in the Synchronous Label condition, by contrast, failed to prefer the out-of-category object systematically \((t(28)=1.067, p=.29)\). Results for Test 2, where the out-of-category object contained a novel shell, and Test 3, where the out-of-category object contained two entirely novel parts, are provided in Table 1. In both Test 2 and 3, all infants exhibited systematic novelty preference, even those familiarised with synchronous labels.

In order to further understand the pattern of results in Test trial 1, we obtained the number of infants in each condition who spent more than \(50 \%\) of looking time on the novel object. This confirmed that infants in the Silent and Asynchronous Label condition were mostly successful at learning the category (Silent: \(N=20, p=.06\), two-tailed binomial test; Asynchronous label: \(N=22, p<.01\) ), whereas the number of successful infants was at chance level in the Synchronous Label condition ( \(N=16, p>.7\) ).

Infants' failure to recognise the out-of-category stimulus on Test 1 as novel in the Synchronous condition suggests a detrimental impact of the synchronous label on learning: it seems as though infants in this condition have not encoded the category equally well as infants in the other conditions. While this is in line with the hypothesis that synchronous labels impose greater processing load and therefore visual stimuli may be processed in less detail, it is possible to achieve a more fine-grained insight into infants performance at test. The out-of-category stimulus in Test 1 was designed to be novel owing to the presence of a different type of leaf its shell part, by comparison, is relatively consistent with the familiarisation category. Examining infants' looking patterns with regard to the two object parts and corresponding parts in the within-category object should therefore provide more information as to how infants process the object.


Figure 5: Looking proportions for individual object parts during Test 1: (a) Leaves belonging to out-of-category object (OOCO), here the "novel" part, and within-category object (WCO); (b) Shells belonging to OOCO and WCO. The inset shows an example test display with both objects. * indicates a statistically significant difference, \(\left({ }^{*}\right)\) indicates a trend.

Part-based looking at test Figure 5 (a) shows looking proportions directed at the leaves belonging to the two ob-
jects in Test 1, i.e. out-of-category leaf and within-category leaf, for every condition. All infants in fact had a clear preference for the novel leaf in comparison to the familiar leaf, even those in the Synchronous Label condition (paired t-test: \(t(28)=5.157, p<.001\), two-tailed). In contrast to what the standard analysis of global looking indicates, infants in the Synchronous Label condition as a group did not fail to encode the distributional properties of the "leaf", as they did not fail to perceive the novel leaf as unfamiliar. In fact, the only difference between the looking patterns across conditions is a trend for the "familiar shell" to be looked at more in comparison to the "novel shell" (illustrated in Figure 5 (b) ) in the Synchronous condition ( \(t(28)=1.76, p=.09\), paired t-test, 2-tailed) - which does not exist in the other conditions.

\section*{Discussion}

We familiarised infants with a novel object category, either in silence (Silent condition) or with novel labels. Labels were either presented one second after image onset (Asynchronous Label condition) or simultaneously with image onset (Synchronous Label condition). Global preferential looking results on three subsequent test trials indicated that infants in the Silent and Asynchronous Label conditions were highly successful at learning the target category, exhibiting novelty preference on all test trials. Infants in the Synchronous Label condition, however, did not exhibit novelty preference until the second test trial, pointing to a disruptive role of synchronous labels. Detailed looking patterns indicated, however, that infants in the Synchronous Label condition still preferred the novel leaf within the out-of-category object over the familiar leaf in the consistent test object.

While our present analysis of part-based looking during familiarisation did not reveal an effect of labeling, future analyses will focus on more fine-grained measures of looking (such as the time course of processing within a trial) in order to establish a link between individual infants' gaze patterns during familiarisation and their performance on test.

Preferential looking on the test trials shows that both infants in the Silent and the Asynchronous Label condition learned the category and showed novelty preference on Test 1 and 2 , where either the highly variable or the less variable object parts were replaced. This indicates that they successfully encoded the feature distribution of both parts. By contrast, infants in the Synchronous condition did not exhibit novelty preference until Test Trial 2. Using the established measure of object-based novelty preference as a marker of successful category formation this appears to indicate that learning was attenuated in comparison to silence by the presence of the label at the start of the trial. However, analysing infants' looking patterns at the level of individual object parts revealed that infants in the Synchronous condition did appear to be sensitive to the replacement of the leaf part, even though they spent more time inspecting the within-category object compared to
the other groups. In addition, on the subsequent test trial, in which the highly variable shell part was replaced by a novel type of shell, even infants in the Synchronous Label condition exhibited significant preference for the novel object. Taken together these data imply that rather than signifying a failure to learn, the looking patterns reflect a delay in the progression to novelty preference, or a lingering familiarity preference (Hunter \& Ames, 1988). This finding is consistent with the hypothesis that synchronous labelling increases cognitive load, but it is inconsistent with the hypothesis that visual processing as such is compromised as a result. Infants hearing synchronous labels during familiarisation have learned about the category in question - but they have a tendency to prefer looking at familiar elements.

This finding has implications for the interpretation of studies investigating the impact of labelling on categorisation. First of all, a decrease in novelty preference scores at the object level does not necessarily imply a decrease in visual learning, but can potentially be explained by changes in the speed in which the shift from familiarity to novelty preference is obtained. Null preferences therefore have to be interpreted with caution, specifically when comparing conditions that inherently differ in terms of cognitive load, such as a silent condition vs. one that includes auditory stimuli.

Secondly, this result highlights the role of timing in infants' early learning: even if it is not the categorisation success per se that has been compromised due to the synchrony of the label, a lingering familiarity preference still indicates greater difficulty processing the stimuli. This could potentially play a role in reconciling previous contradictory findings regarding the impact of labelling on categorisation. Most studies cited above (e.g. Fulkerson \& Waxman, 2007; Ferry et al., 2010; Plunkett et al., 2008; Althaus \& Westermann, in prep.) used a delayed label onset similar to our Asynchronous label condition, but Robinson and Sloutsky (2007), who found auditory overshadowing in the presence of labels, presented labels at picture onset. Further research will be necessary in order to determine whether the timing of the label can indeed explain the discrepancies between the findings.

Finally, the fact that infants can deal with asynchronous labels and may even benefit from them, whereas synchronous presentation seems to cause problems with cognitive load, is an important cue to how words and images are processed. Clearly stimuli are processed on-line, rather than stored in short term memory and processed separately and independently of their presentation time. While, in the context of cross-modal processing, synchrony is often claimed to be beneficial at least for young infants (Gogate et al., 2003), the increased load due to synchronous presentation here appears to slow down learning. The actual processing load of any coupling of visual and auditory stimuli is also dependent on the visual and auditory complexity or novelty. However, the fact remains that the likelihood of a label in real life occurring at the same time as an object comes into view is rather small. In terms of learning, this may be an ecological advantage rather
than a shortcoming. Perhaps having the opportunity to process visual and auditory information sequentially allows the learning of more complex visual structures. Specifically, the grouping of similar (but not identical) items into categories, for which more abstract visual processing may be needed than just for recognition of individual objects, could be facilitated in this way. Further research is needed in order to shed light on the relationship between exact time course of word-object integration and the ecological circumstances the auditory and visual signals occur in.

\section*{Acknowledgments}

This research was supported by Wellcome Grant 084386/Z/07/Z to Kim Plunkett and the Winkler Career Development Fellowship at St Hugh's College, University of Oxford, to Nadja Althaus. We wish to thank all parents and infants who participated in this study.

\section*{References}

Althaus, N., \& Westermann, G. (in prep.). Labels can cause infants to split a visual category. (In preparation)
Bahrick, L., \& Lickliter, R. (2000). Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. Developmental psychology, 36(2), 190.
Balaban, M., \& Waxman, S. (1997). Do words facilitate object categorization in 9-month-old infants? Journal of Experimental Child Psychology, 64, 3-26.
Baldwin, D., \& Markman, E. (1989). Establishing wordobject relations: A first step. Child, 60, 381-398.
Bergelson, E., \& Swingley, D. (2012). At 6-9 months, human infants know the meanings of many common nouns. Proceedings of the National Academy of Sciences, 109, 32533258.

Boroditsky, L. (2001). Does language shape thought?: Mandarin and english speakers' conceptions of time. Cognitive Psychology, 43, 1-22.
Bortfeld, H., Morgan, J., Golinkoff, R., \& Rathbun, K. (2005). Mommy and me familiar names help launch babies into speech-stream segmentation. Psychological Science, 16(4), 298-304.
Ferry, A., Hespos, S., \& Waxman, S. (2010). Categorization in 3- and 4-month-old infants: An advantage of words over tones. Child Development, 81, 472-479.
Fulkerson, A., \& Waxman, S. (2007). Words (but not tones) facilitate object categorization: Evidence from 6-and 12-month-olds. Cognition, 105, 218-228.
Gogate, L., Bahrick, L., \& Watson, J. (2003). A study of multimodal motherese: The role of temporal synchrony between verbal labels and gestures. Child development, 71(4), 878-894.
Grossmann, T., Gliga, T., Johnson, M., \& Mareschal, D. (2009). The neural basis of perceptual category learning in human infants. Journal of Cognitive Neuroscience, 21, 2276-2286.
Hunter, M., \& Ames, E. (1988). A multifactor model of
infant preferences for novel and familiar stimuli. Advances in Infancy Research, 5, 69-95.
Plunkett, K., Hu, J., \& Cohen, L. (2008). Labels can override perceptual categories in early infancy. Cognition, 106, 665-681.
Quinn, P. C., Westerlund, A., \& Nelson, C. (2006). Neural markers of categorization in 6-month-old infants. Psychological Science, 17, 59-66.
Robinson, C., \& Sloutsky, V. (2007). Linguistic labels and categorization in infancy: do labels facilitate or hinder? Infancy, 11, 233-253.
Sloutsky, V., Lo, Y., \& Fisher, A. (2001). How much does a shared name make things similar? linguistic labels, similarity, and the development of inductive inference. Child Development, 72, 1695-1709.
Tincoff, R., \& Jusczyk, P. (1999). Some beginnings of word comprehension in 6-month-olds. Psychological Science, 10, 172-175.
Waxman, S., \& Braun, I. (2005). Consistent (but not variable) names as invitations to form object categories: new evidence from 12-month-old infants. Cognition, 95, 5968.

Waxman, S., \& Markow, D. (1995). Words as invitations to form categories: Evidence from 12-to 13-month-old infants. Cognitive Psychology, 29, 257-302.
Whorf, B. (1956). In J. B. Carroll (Ed.). Language, thought and reality: Selected writings of Benjamin Lee Whorf, Cambridge, MA: MIT Press.

\title{
Shape categorization in Autism: Does it follow the pattern of typical development?
}

\author{
Joseph L. Amaral (amaralj1@mail.uc.edu) \\ University of Cincinnati Department of Psychology, Mail Location 0376 \\ Cincinnati, OH 45221 USA \\ Susan Collins (collinsu@mail.uc.edu) \\ University of Cincinnati College of Medicine, 231 Albert Sabin Way \\ Cincinnati, OH 45267 USA \\ Shannon O'Connor (oconnosl@mail.uc.edu) \\ University of Cincinnati, Department of Psychology, Mail Location 0376 \\ Cincinnati, OH 45221 USA
}

\author{
Shealan McAlister (mcalissd@mail.uc.edu) \\ University of Cincinnati Department of Psychology, Mail Location 0376 \\ Cincinnati, OH 45221 USA
}

\author{
Heidi Kloos (heidi.kloos@uc.edu) \\ University of Cincinnati Department of Psychology, Mail Location 0376 \\ Cincinnati, OH 45221 USA
}

Carol D. Luzzi (cluzzi@memorialsb.org)
Memorial Hospital of South Bend, 100 Navarre Place, Suite 5550
South Bend, IN 46601 USA

\begin{abstract}
When identifying basic-level categories (e.g., airplane, cow), typically developing (TD) children commonly use the overall shape of objects as basis for their judgments. This so-called shape bias is tied to the size of a child's vocabulary and as such might be a way of adaptively organizing an evergrowing vocabulary. The current study looks at whether the same is true for children with autism spectrum disorder (ASD). A group of participants with ASD and TD controls were asked to categorize objects that differed in the amount of item detail. Results show that vocabulary size was related to success in categorizing objects for TD participants, but not for ASD participants. We discuss the degree to which a link between shape bias and vocabulary size in ASD children may be an indication of differentiated patterns of adaptation.
\end{abstract}

Keywords: categorization; language development; autism
The overall shape of items is important when it comes to learning the words of basic-level categories. Whether we consider a dog, a car, an airplane, or a cake, the most salient difference among these items is their overall shape. Indeed, typically developing (TD) children show a pronounced bias toward the overall shape of objects (e.g., Diesendruck \& Bloom, 2003). For example, when children are asked to group novel objects, they overwhelmingly group objects based on their overall shape rather than other features (e.g., texture, color etc.; Samuelson \& Smith, 2005). The global feature of overall shape supersedes details.

While the "shape bias" has been documented repeatedly, it was found only recently that this bias is linked to the size
of a child's vocabulary (Pereira \& Smith, 2009). In Pereira \& Smith (2009), participants were TD toddlers with varying vocabulary sizes. They participated in multiple trials of a task in which they were asked to decide which of three toy objects (e.g., a car, a plane, a cake) matched the label offered by the experimenter (e.g., "show me the car"). Importantly, the degree of detail of the presented objects differed as a function of trial type. In some trials, objects were highly detailed, providing information about color and fine-grain shape. In others trials color was omitted, leaving details only about the fine-grain shape. And finally, in the third trial type, the objects were mere shape abstractions, missing both color and fine-grain shape. Results indicated that a child's productive count noun vocabulary (as compared to receptive vocabulary or general vocabulary) had a significant effect on performance. While children could categorize the detailed objects equally well, only children with larger count-noun vocabularies could identify the objects represented as shape abstractions.

In a similar vein, research has demonstrated that TD children with small vocabularies benefit from teaching props that focus their attention to the overall shape of items (Son, Smith \& Goldstone, 2008). Son and colleagues (2008) taught children the names of new objects that either had large amount of detail (e.g., texture, color, fine-grain shape) or were mere shape abstractions. In this latter case, shape abstractions approximated the overall shapes of the objects. Results show that training with shape abstractions yielded
better performance later identifying detailed objects from the same category than the training with detailed objects. Focusing children's attention on the overall shape of objects by removing irrelevant information promoted better learning. The ability to see gist, Gestalts, and global features appears integral to how TD children learn, categorize, and identify objects.

\section*{Detail Focus in ASD}

Compared to their TD peers, children with ASD have a tendency to focus on specific details. This may include fixation with the parts of objects (e.g., the wheels of a toy car) or having very limited and particular interests (e.g., former Secretaries of the Interior). Indeed, these types of detail orientation are included in the diagnostic criteria for ASD (American Psychiatric Association, 2000). Theorists have argued that the focus on detail may stem from what they describe as "weak central coherence" (WCC), or a decreased push toward Gestalts (e.g., Happé \& Frith, 2006; Happé \& Booth, 2008). This account of ASD does not postulate that children with ASD are incapable of processing global information, but rather that they tend to gravitate toward details. That is, when absolutely pressed, children with ASD are able to see Gestalts, but all things equal, will focus on detail.
The best example of this difference in focus was established with the classical Navon task, a task in which stimuli consist of many small letters configured in the arrangement of a large letter (cf., Navon, 1977). For TD participants, results show a distinct interference of large letters on the perception of small letters, both in children (Plaisted, Swettenham, \& Rees, 1999) and in adults (e.g., Navon, 1977). In particular, when participants are asked to focus on small letters, reaction time is longer for trials in which large and small letters differ than on trials in which large and small letters match. This global interference is indetectable in participants with ASD: They perform equally fast in both letter-mismatch trials and letter-match trials, in both cases with high accuracy (e.g., Plaisted et al., 1999).

Another example of weak central coherences in ASD comes from face perception tasks. The identity of a face is defined not only by its individual parts (e.g., nose, eyes, mouth), but also by the holistic configuration of these parts, something that appears to be disrupted when faces are presented upside down. For TD children, recognition accuracy decreases when faces are presented upside down, compared to trials in which faces are presented upright (Mondloch, Le Grand, \& Maurer, 2002). In contrast, children with ASD do not perform differently as a function of face orientation (e.g., Tantam, Monaghan, Nicholson, \& Stirling, 1989). Along the same lines, participants with ASD could classify faces better when local rather than global features were exaggerated (through the use of a high-pass vs. low-pass filter; Deruelle, Rondan, Salle-Collemiche,

Bastard-Rosset, \& Da Fonséca, 2008). The inverse pattern of results was obtained for TD children.

Applied to language learning, children with ASD do not show evidence of the same shape bias found in TD children (Tek, Jaffery, Fein, \& Naigles, 2008). While TD children demonstrated movement toward categorizing objects by global shape, ASD children did not. Compared to their TD peers, children with ASD also often have difficulty communicating, frequently displaying atypical language development (American Psychiatric Association, 2000). However, it is not clear whether the connection seen between vocabulary size and object categorization style exists in ASD as it does in TD. The current study aims to explore this explicitly, potentially providing evidence for differentiated patterns of adaptive mental functioning.

\section*{Rationale for the Current Study}

Very little research exists exploring how children with ASD categorize typical objects and what role shape might play. For TD children, the development of a bias toward categorizing objects based on shape may relate to the emergence of an overall tendency to focus on Gestalts, which may have an adaptive function. Compared to TD children, children with ASD tend to focus on details and do not show a natural shape bias. This may indicate an atypical pattern of adaptive functioning. To begin exploring this line of research, the current study aims to compare TD children and children with ASD who possess similar productive count-noun vocabularies on a task in which they are asked to identify objects that afford various degrees of detail.

\section*{Method}

\section*{Participants}

Seventy TD children (39 boys and 31 girls) were recruited from Cincinnati area schools and a local children's museum Ages ranged from 14-29 months ( \(M=20.78, S D=3.67\) ). Twenty-five children with ASD (22 boys and 3 girls) were recruited from Cincinnati area treatment centers and special needs schools. Their ages ranged from 2 years, 9 months -17 years, 5 months ( \(M=5\) years, 11 months; \(S D=3\) years, 7 months) Diagnoses of ASD were confirmed through contacting their pediatricians or therapists after written consent was obtained from their guardians.

\section*{Language Measure}

To assess each child's vocabulary, parents were asked to fill out the MacAurthur-Bates Communicative Development Inventories (MCDI; Bates, Dale \& Thal, 1995), a survey for parents that is widely used in the language development literature. Parental report of language abilities has been demonstrated to be a valid measure of both TD and ASD language abilities (Luyster, Lopez \& Lord, 2007; Tomasello, 1994). Parents completed the entire survey, but for the purposes of the current study, only the sum of items representing count nouns each child could understand and
say were used for comparison and analysis. It was a child's productive count noun vocabulary that Pereira and Smith (2009) tied to his or her ability to categorize objects based on global features.

\section*{Materials}

Stimuli were constructed to represent 12 noun categories commonly known by young children. These categories were: horse, cow, pig, fish, bird, butterfly, turtle, car, airplane, cake, shoe, and hamburger. Categories were represented by three objects, each from a different condition based on how much information they afforded the child.

The first object for each category, referred to as the detailed object, consisted of a toy or model purchased from toy stores. It contained detailed color, texture, and shape information (see Figure 1A). The second object, referred to as the rich-shape object, was constructed using a duplicate of the detailed object covered with black modeling clay. This clay served to remove the color and textural information while maintaining detailed shape (see Figure 1B). The object from the third condition, referred to as the shape abstraction, was made of Styrofoam. It was designed to represent the overall shape of the object category without providing any detailed information (see Figure 1C). Detailed objects served to confirm that children were able to identify the object categories. Shape abstractions, in contrast, provided information about whether children were able to identify global abstractions of objects. The intermediate, rich-shape condition served to help illuminate potential trends in identification abilities.


Figure 1: Examples of (A) a detailed turtle, (B) a rich-shape turtle, and (C) a shape abstraction turtle.

\section*{Procedure}

The procedure for the current study was adapted from Pereira and Smith (2009). To begin, there were four practice trials, carried out on a laptop computer, designed to acclimate the child to the researcher and, for lower functioning children, to ensure that they were able to follow directions. During each practice trial, the researcher showed the child two photographs of easily discriminable, common nouns (e.g., dog, bunny, train, and kitty). The experimenter
labeled the object on one of the photos and asked the child to point to it. Performance did not affect the child's eligibility to participate in the rest of the study.

For the main task, the experimenter told participants that they were going to play with some toys from a toy box. The experimenter then placed a red wooden tray ( 60 cm by 30 cm ) in front of the child so that it was out of reach. This served as a platform for stimuli and to help children focus their attention on the testing space. For each test trial, the experimenter placed three detailed objects, three rich-shape objects, or three shape abstractions on the board (see Figure 2 for an example). One object served as a target, the other two as distractors. The experimenter then asked for the target (e.g., "Give me horse."), and pushed the tray within reaching distance of the child. Clear pointing or picking up the target were considered correct responses. Regardless of whether or not the child was correct, neutral feedback was given. The experimenter recorded responses on a laptop computer. This procedure was repeated for 12 testing trials. Children never saw multiple versions of an object category. For example, if presented with a shape distraction airplane, children would not see the detailed airplane. Types of object were balanced across participants.


Figure 2: Example test trial showing a fish, car and pig in the shape abstraction condition.

\section*{Results}

For both TD children and children with ASD, productive count noun vocabularies ranged from 0-199 words (TD \(M=\) 78.17, \(S D=66.5\); ASD \(M=79.59, S D=66.91\) ). Similar to previous work by Smith (2003), participants from both the TD and ASD groups were divided into subgroups based on their productive count noun vocabulary sizes: Children whose productive count-noun vocabulary was between 0 and 100 words were classified as being in the LowVocabulary group (the largest vocabulary in this group was 92), and children with a vocabulary between 100 and 200 nouns were classified as being in the High-Vocabulary groups (the smallest vocabulary in this group was 102). The vocabulary groups, as well as associated descriptive statistics are shown in Table 1. Because analyses conducted utilizing divisions based on mean and median vocabularies yielded similar results, the above method was utilized to maintain continuity with previous work.

Table 1: Descriptives of TD and ASD Participants, Mean Age and Number of Count-Nouns in Productive Vocabulary Separated by 2 Vocabulary Groups.
\begin{tabular}{ccc}
\hline & \multicolumn{2}{c}{ Vocabulary Group } \\
& \(<100\) nouns & \(>100\) nouns \\
\hline & \(N=45\) & \(N=25\) \\
& \(M\) age \(=19.00\) months & \(M\) age \(=24.00\) months \\
TD & \((S D=2.88)\) & \((S D=2.60)\) \\
& \(M\) vocab \(=34.29\) & \(M\) vocab \(=157.16\) \\
& \((S D=28.46)\) & \((S D=33.95)\) \\
\hline & \(N=14\) & \(N=11\) \\
& \(M\) age \(=5\) yrs, 10 months & \(M\) age \(=6\) yrs, 1 month \\
ASD & \((S D=4\) yrs, 1 month \()\) & \((S D=3\) yrs \()\) \\
& \(M\) vocab \(=30.57\) & \(M\) vocab \(=135.55\) \\
& \((S D=34.48)\) & \((S D=41.23)\) \\
\hline
\end{tabular}

\section*{Categorization Performance: TD Sample}

The average performance on detailed objects, rich-shape objects, and shape abstraction for TD children in the Lowand High-Vocabulary groups are shown in Figure 3.

A 2 X 3 (Vocabulary Group X Object Condition) mixed measures analysis revealed a significant effect of Vocabulary Group, \(F(1,68)=33.31, p<.001\), with better performance for the High- than the Low-Vocabulary Group \(\left(M_{H}=88.46 \%, S D_{H}=34.12 \% ; M_{L}=60.93 \%, S D_{L}=\right.\) 27.25\%). There was also a significant effect of Object Condition, \(F(1,68)=48.50, p<.001\), with highest performance for detailed objects \((M=80.71 \%, S D=\) 22.99\%), followed by rich-shape objects ( \(M=68.57 \%, S D\) \(=27.15 \%\) ), and followed by shape abstractions ( \(M=\) \(58.21 \%, S D=27.16 \%\) ). While the interaction was not significant, \(p>.90\), TD children in the Low-Vocabulary group performed better on detailed object trials than richshape trials, \(t(48)=2.07, p<.05\), and better on rich-shape trials than shape abstraction trials, \(t(48)=2.08, p<.05\). In contrast, performance for TD children in the HighVocabulary group did not differ between rich-shape and shape abstraction trials, \(t(48)=1.31\) (though there was a difference between detailed object and rich-shape trials, \(t(48)=2.07, p<.05)\).

\section*{Categorization Performance: ASD Sample}

The average performance of children with ASD from the Low- and High-Vocabulary groups across the detailed, richshape, and shape abstraction trials is illustrated in Figure 4.

As was done with the TD sample, a \(2 \times 3\) (Vocabulary Group X Object Condition) mixed measures analysis was conducted. Surprisingly, there was only a marginal effect of Vocabulary Group, \(F(1,23)=3.18, p=.063\left(M_{H}=89.39 \%\right.\), \(\left.S D_{H}=25.79 \% ; M_{L}=72.02 \%, S D_{L}=31.33 \%\right)\). However, as was found with TD children, there was a significant effect of Object Condition, \(F(1,23)=7.07, p<.02\). Across vocabulary groups, performance was best for detailed objects ( \(M=88.00 \%, S D=22.96 \%\) ), second best for rich-
shape objects ( \(M=78.00 \%, S D=32.33 \%\) ), and lowest for shape abstractions ( \(M=73.00 \%, S D=32.21 \%\) ).
Importantly, the interaction was not significant, \(p>.50\). Looking at simple effects within vocabulary groups, performance did not differ between detailed object and richshape trials, \(p \mathrm{~s}>.30\). There was also no difference between rich-shape and shape abstraction trials, \(p \mathrm{~s}>.30\).


Figure 3: Mean proportion correct responses across object conditions for TD children as a function of vocabulary size.

\section*{Categorization Performance: Comparing Samples}

Categorization performance was compared between diagnostic groups for both the Low- and High-Vocabulary groups. For children in the Low-Vocabulary groups, performance did not differ between children with ASD and TD children on detailed object and rich-shape trials, \(p s\) > .37. However, children with ASD from the Low-Vocabulary group performed significantly better than TD children with similar vocabularies on shape abstraction trials, \(t(57)=2.43\), \(p<.02\). Children with ASD from the High-Vocabulary group outperformed TD children on detailed object trials, \(t(24)=2.14, p<.05\), but differences were not significant for rich-shape or shape abstraction trials, \(p s>.33\).


Figure 4: Mean proportion correct responses across object conditions for children with ASD as a function of vocabulary size.

\section*{Discussion}

Results provide evidence for a difference in the relation between count-noun vocabulary sizes and categorization abilities in TD children versus children with ASD. In particular, while there was a characteristic difference between TD children in the Low- versus High- Vocabulary group, this difference disappeared for ASD Children. Note that the TD finding is not as robust as previously found (e.g., Pereira \& Smith, 2009; Smith, 2003). Nevertheless, TD children in the High-Vocabulary group performed equally well in the rich-shape and shape abstraction conditions, whereas TD children in the Low-Vocabulary group performed differently across all three object conditions.

In contrast, children with ASD from both vocabulary groups demonstrated similar performance patterns across conditions. When performance patterns between diagnostic groups were compared directly, children with ASD from the Low-Vocabulary group performed better than their TD peers on shape abstraction trials, and equally well in other trials. Performance across the rich-shape and shape abstraction trial types were equivocal for TD children and children with ASD from the High-Vocabulary groups.

Before interpreting how the current results relate to global processing and contextual changes, the current study makes several assumptions. The first assumption is that vocabulary size is a contextual factor. The second assumption is that categorization of shape abstraction objects translates to the ability to process Gestalt information. The third assumption is that Gestalt processing is an adaptive function that arises when contexts make tasks difficult (in this case, as vocabulary size increases). Under these assumptions, the fact that children with ASD did not demonstrate stratified performance seen in TD children may suggest that they do not adapt to contextual changes in the same manner. The generally high performance for children with ASD across object conditions regardless of vocabulary size is in line with the argument that the focus on detail commonly seen in ASD does not equate to an inability to process global information, but rather a preference for local features.

Could alternate claims explain the findings? Most prominent is the issue of how vocabulary was measured. The MCDI only allows parents to indicate words that they believe their child does not understand, words that they understand, but cannot say, and words that they understand and say. Many children with ASD use alternate forms of communication, such as sign language or exchange cards. Thus, the MCDI may not have been an accurate measure of each child's true productive count-noun vocabulary. Furthermore, the MCDI only lists words that TD children tend to learn in the first few years of life. Children with ASD who learn language later in life may not learn the same first words. Thus, it is possible that a greater proportion of their produced words were not captured by this measure. Given the issues with the MCDI, it is possible that a more
detailed assessment of language, including receptive vocabulary, could influence results.

Other issues that could affect the interpretation of results include the current sample size of children with ASD and the methods used to confirm diagnoses. Only 25 children with ASD were included in the final sample. The current study also based diagnosis on physician confirmation, rather than using standardized measures, such as the Autism Diagnostic Observation Schedule (ADOS; Lord et al., 2000). Though best diagnostic practice suggests the use of standardized measures (Ozonoff, Goodlin-Jones, \& Solomon, 2005), the individual practices of clinicians may vary.

Nevertheless, even given the alternate explanations for the results of this study, there is evidence from previous research which supports the current claims. First, previous work that explored how children with ASD see Gestalts demonstrated that they have the ability to identify both global and local features (e.g., Ozonof et al., 1994; Plaisted et al., 1999). Second, there is an indication that the shape bias in children with ASD is different from the shape bias of TD children (Tek et al., 2008). Combined, these studies suggest that there is also a weakened connection between global processing and vocabulary. In the current study, children with ASD tended to perform well on shape abstraction objects regardless of their vocabulary size, indicating that they were, in general, able to categorize objects based on their overall shape.

Research involving children with language delays, socalled late talkers, may also provide evidence for current claims. In similar object categorization tasks, they demonstrate performance patterns are similar to both TD children and children with ASD. Like TD children, late talkers show a developmental trend in their ability to categorize objects by shape (Jones \& Smith, 2005). Specifically, Jones and Smith (2005) found that neither a late talker's receptive count-noun vocabulary nor age was significantly related to their ability to categorize objects based on overall shape. Productive count-noun vocabulary size, instead, was related. However, like children with ASD, they tend to not show an innate tendency to categorize objects by shape, and thus have an atypical shape bias (Jones, 2003). Though this provides evidence that in another clinical group, productive count-noun vocabulary size relates to categorization abilities, it does not address the issue of the accuracy of the MCDI for children with ASD.

The current study is a preliminary step towards better understanding the relationship between vocabulary size and object categorization style in children with ASD. It appears that though they may not have a bias toward global features, even those with poor verbal language skills have the capacity to categorize objects based on overall shape. This may have important clinical implications.

\section*{Acknowledgments}

The authors would like to thank Cathy Walker, Heather Baker and Lori Hood for their help on this project.

\section*{References}

American Psychiatric Association. (2000). Diagnostic and statistical manual of mental disorders (4th ed., text rev). Washington, DC: Author.
Bates, E., Dale, P. S., \& Thal, D. (1995). Individual differences and their implications for theories of language development. In P. Fletcher \& B. MacWhinney (Eds.) The handbook of child language. Oxford, UK: Basil Blackwell.
Deruelle, C., Rondan, C., Salle-Collemiche, X., BastardRosset, D., \& Da Fonséca, D. (2008). Attention to lowand high- spatial frequencies in categorizing facial identities, emotions and gender in children with autism. Brain and Cognition, 66, 115-123.
Driesendruck, G. \& Bloom, P. (2003). How specific is the shape bias? Child Development, 74(1), 168-178.
Happé, F. G. E. \& Booth, R. D. L. (2008). The power of the positive: Revisiting weak coherence in autism spectrum disorders. The Quarterly Journal of Experimental Psychology, 61, 50-63.
Happé, F., \& Frith, U. (2006). The weak coherence account: Detail-focused cognitive style in autism spectrum disorders. Journal of Autism and Developmental Disorders, 36(1), 5-25.
Jones, S. (2003). Late talkers show no shape bias in a novel name extension task. Developmental Science, 6(5), 477483.

Jones, S. \& Smith, L. B. (2005). Object name learning and object perception: A deficit in late talkers. The Journal of Child Language, 32, 223-240.
Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Jr., Leventhal, B. L., DiLavore, P. C., Pickles, A., \& Rutter, M. (2000). The autism diagnostic observation schedulegeneric: A standard measure of social and communicative deficits associated with the spectrum of autism. Journal of Autism Developmental Disorders, 30, 205-223.
Luyster, R., Lopez, K., \& Lord, C. (2007). Characterizing communicative development in children referred for autism spectrum disorders using the MacArthur-Bates Communicative Development Inventory (CDI). Journal of Child Language, 34, 623-654.
Mondloch, C.J., Le Grand, R., \& Maurer, D. (2002). Configural face processing develops more slowly than featural face processing. Perception, 31, 553-566.
Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. Cognitive Psychology, 9, 353-383.
Newport, E. (1990). Maturational constraints on language learning. Cognitive Science, 14(1), 11-28.
Ozonoff, S., Goodlin-Jones, B. L., \& Solomon, S. (2005). Evidence-based assessment of autism spectrum disorders in children and adolescents. Journal of Clinical Child and Adolescent Psychology, 34(3), 523-540.

Pereira, A. \& Smith, L. B. (2009). Developmental changes in visual object recognition between 18 and 24 months of age. Developmental Science, 12(1), 67-83.
Plaisted, K., Swettenham, J., \& Rees, L. (1999). Children with autism show local precedence in a divided attention task and global precedence in a selective attention task. Journal of Child Psychology and Psychiatry and Allied Disciplines, 40(5), 733-742.
Samuelson, L. K. \& Smith, L. B. (2005) They call it like they see it: Spontaneous naming and attention to shape. Developmental Science, 8 (2), 182-198.
Smith, L. B. (2003) Learning to recognize objects. Psychological Science, 14 (3), 244-250.
Son, J., Smith, L. B. \& Goldstone, R. (2008). Simplicity and Generalization: Short-cutting Abstraction in Children's Object Categorizations. Cognition, 108 (3), 626-638.
Tantam, D., Monaghan, L., Nicholson, H., \& Stirling, J. (1989). Autistic children's ability to interpret faces: A research note. Journal of Child Psychology and Psychiatry, and Allied Disciplines, 30(4), 623-630.
Tek, S., Jaffery, G., Fein, D., \& Naigles, L.R. (2008). Do children with autism spectrum disorders show a shape bias in word learning? Autism Research, 1(4), 208-222.
Tomasello, M. M. (1994). The instrument is great, but measuring comprehension is still a problem. Monographs of the Society for Research in Child Development, 59(5), 174-179.

\title{
Gender and Sex: The Experiential Basis of Grammar
}

\author{
Elena Andonova (eandonova@nbu.bg) \\ Department of Cognitive Science and Psychology, 21 Montevideo Street \\ Sofia, 1618 Bulgaria
}

\begin{abstract}
There have been contradictory reports of sex differences in language processing. A novel approach is adopted here which explores the experiential basis of such differences. Two studies examine the auditory processing of grammatical gender in Bulgarian in a gender decision (gender monitoring) task and a cued shadowing (word repetition) task. Reaction times in both experiments reveal significant two-way interactions between the grammatical gender of words (masculine vs. feminine) and the sex of the voice (male vs. female). The sex of participants in the gender decision task also interacted with word gender in terms of decision accuracy. Women were relatively more accurate on their "own reference" word gender (feminine) and less accurate on masculine gender words. A two-way interaction between word gender and participant sex on response latencies in the cued shadowing task supports the view that these effects are not strategic but have a highly automatic nature instead. Findings are interpreted in terms of individual differences in the experience of grammatical gender in such gender-marking languages.
\end{abstract}

Keywords: grammatical gender, participant sex/gender, individual differences.

\section*{Introduction}

Over the past decades a number of studies have claimed to reveal sex differences in language performance, language ability and underlying brain cortical areas and hormonal levels. For example, Shaywitz, Shaywitz, Pugh, Constable, Skudlarski, Fulbright, et al. (1995) have suggested differences in language lateralization, Weiss, Kemmler, Deisenhammer, Fleischhacker, \& Delazer (2003) in verbal fluency tasks, Kramer, Kaplan, Delis, O’Donnell, \& Prifitera (1997) in verbal learning, etc., and theoretical accounts of such differences may attempt to explain them in terms of evolutionary origins and advantages and/or hormone level variation. Although sex differences may disappear in later childhood, large-scale studies have found robust evidence for the effects of gender on early language development, including vocabulary comprehension and production using the MacArthur Bates Communicative Development Inventories (Fenson, Dale, Reznick, Bates, Thal, \& Pethick, 1994). On the other hand, in a critical review of the research presented in numerous studies, Wallentin (2009) concludes that there is not much clear and uncontroversial evidence for sex differences in language
processing with the exception of the early language development advantage for girls and that although certain language-related deficits exhibit clear sex differences, such as stuttering, dyslexia, and autism whose occurrence is much higher in males, a causal link and a good theory of why and how such differences may arise are still lacking.
In this paper, an entirely different aspect of human experience with language is explored, which may lead to sex differences in language processing, more specifically grammar. Even if grammatical categories are learned and used by all typically developing speakers of a language, both men and women, the specific individual experiences with these categories may differ. The use of grammatical gender is a case in point here. Individuals vary on the dimension of sex and in at least some languages this individual characteristic is important in selecting the appropriate grammatical gender form used in reference to that individual. In the relatively poor morphology of the English language, this variation is observed in very few forms such as the personal pronouns in the \(3^{\text {rd }}\) person, Singular (he vs. she), in the richer morphology of the German language it is found in noun phrases referring to both animate and inanimate entities encoded in pronouns and articles (der, die, das), and in Slavic languages even further on word categories such as adjectives, verb forms, numerals, etc. Grammatical gender may furthermore be not as arbitrary as it is habitually seen. In a study of gender processing and lexical access in Bulgarian, Andonova, D'Amico, Devescovi, \& Bates (2004) discovered a significant contribution of semantic gender to processing in Bulgarian in contrast with previous findings for Italian. Particularly interesting, however, was another finding of this research, namely, an interaction between sex of the subject and noun gender, reflecting a bias toward one's own grammatical gender "counterpart" (especially for females) in Bulgarian. Triggered by this novel finding, a reanalysis of data from a prior study in Italian showed a similar interaction. How could such differences emerge? In contrast with biological and/or cultural explanations, we offer here an experiential account.

Both men and women produce gender-marked words (nouns, adjectives, verbs, etc.) in Bulgarian in large quantities on a daily basis, and so both sexes have a largely equivalent experience in terms of frequency of usage of the three genders. However, a woman's individual experience with gender-marked forms matching her own sex, i.e., Feminine gender words, would be different in some ways from the experience of a man with the same feminine gender forms, all else being equal. The difference is both in
the quantity and quality of the experience. Women speaking Slavic and Romance languages (for example) would produce a higher number of Feminine gender forms than men given that they use such forms in situations of selfreference, in which men would have to use Masculine gender forms, and vice versa. For example, a Bulgarian woman may express her feeling of fatigue by saying уморена съм (Eng., I am tired_Fem), an Italian woman would say sono stanca (Eng., I am tired_Fem) while in Italian, a man would have to admit sono stanco (Eng., I am tired_Masc). The individual experience women and men have with gender-marking forms is not only different in quantity (frequency of usage) but also in the degree to which there is personal relevance to the individual (in selfreference). Thus, a lifetime of using gender marked forms in a sex-specific way would lead to the accumulation to differences in the sensitivity to such forms in the two sexes.

\section*{Experiment 1}

This first experiment had the following objectives. First, it aimed at testing for an interaction between participant sex and word gender in the gender decision task and extending previous somewhat limited findings with Bulgarian and Italian speakers performing a gender monitoring (decision) task on nouns in their native language (Andonova et al., 2004).

Second, every noun typically has only one gender, i.e., Masculine, Feminine or Neuter, for example, стол (Eng., chair) is a Masculine noun, маса (Eng., table) is a Feminine noun, and куче (Eng., dog) is a Neuter noun, that is, grammatical gender is invariant. However, adjectives and some verb forms in Bulgarian have all three gender forms, for example, бавен, бавна, бавно are the translation equivalents of the English word slow but in three different forms, one for each of the three genders (slow_Masc, slow_Fem, slow_Neut). The second research objective was to test whether a participant sex by grammatical gender interaction would also emerge in the processing of adjectives and verb forms that can vary across gender categories. Note that in Bulgarian, there is typically a regular and transparent mapping between word form endings and the category of gender. This applies to nouns as well as the adjectives and verb forms included in the stimuli materials for this experiment.

The third objective was to examine the possibility of an interaction between grammatical gender and speaker sex in addition to participant sex. The motivation for this follows a somewhat similar rationale as that concerning the interaction of participant sex and grammatical gender and its possible explanation on the basis of the difference in the quantity and kind of experience the two sexes have with gender-marking word forms, as elaborated in the examples above. Speakers of gender-marking languages such as Slavic and Romance languages not only produce owngender matching word forms more frequently but they also
hear such forms more frequently in verbal interactions, viz., in situations where their interlocutor refers to them, for example, asks them whether they are tired, etc. (Note though later - that it is confounded without situational context - speaker talking about \(1^{\text {st }}\) or \(2^{\text {nd }}\) person etc.) Therefore, one might expect to find such an interaction between speaker sex (the voice for the auditory stimuli) and the grammatical gender of the words. This possibility was tested by presenting stimuli in the auditory modality instead of in writing and recording stimuli in two voices - male and female - to be used as an experimental variable.

\section*{Method}

The design of the experiment included Participant Sex (men vs. women), Grammatical Gender of the word (Masculine vs. Feminine) and Voice Gender (male vs. female voice) as independent variables and mean percent errors and response latencies as the dependent variables.

Participants 40 participants ( 20 men and 20 women) took part in the experiment. They were university students within the 19-30 age range who were paid a modest amount for their participation. All were native speakers of Bulgarian.

Stimuli and Procedure The experimental stimuli consisted of 3 different gender forms each of 62 verbs and adjectives in Bulgarian presented in a different randomized order for each participant in the auditory modality by two speakers of Bulgarian, i.e., in two voices (a male and a female voice). Words of all three grammatical genders in Bulgarian were included in the list of stimuli in order to make the task more natural and the research objectives less obvious so that participants would not be tempted to follow a simple binary choice strategy. The analyses, however, focused on the two critical grammatical genders (Masculine and Feminine) in line with the research hypotheses.

Participants were tested individually in a sound proof booth and were asked to listen to stimuli one at a time and press one of three available buttons to indicate the grammatical gender of the word they have just heard (Masculine, Feminine, or Neuter). Accuracy and speed in completing the task were both emphasized in the experimental instruction.

Presentation of stimuli and registration of participants' responses were controlled by Psyscope and a button box. Before the experiment began, participants did four to six practice trials to familiarize themselves with the task and procedure. In addition, since gender in Bulgarian is a threemember grammatical category, and in order to avoid bias, participants were assigned randomly to one of six spatial configurations of the three buttons in the gender-monitoring task, i.e., m-f-n, m-n-f, f-n-m, f-m-n, n-m-f, or n-f-m.

The 186 word forms were recorded by a female and a male speaker of Bulgarian in a neutral intonation with a falling tone. They were digitized using the Macintosh SoundEdit system, and were placed in a sound file within
the PsyScope experiment preparation package developed by Cohen, MacWhinney, Flatt, and Provost (1993) at Carnegie Mellon University. For each item, reaction times (RT) were measured from the offset of the stimulus word to the participant's response (the button press in the gender decision task).

\section*{Results}

All participants had a mean accuracy higher than \(90 \%\) and a mean RT below 900 msec . RTs for trials on which participants made a decision error were dropped from further analysis. The following outlier procedure was used. The means (M) and standard deviations (SDs) for each participant were computed, and all RTs less than three SDs below the mean of the participant or greater than three SDs above the mean were considered outliers. This resulted in the rejection of \(1.58 \%\) of all RT data. Here first the results of analyses run on participant means will be presented.

The data were analyzed in two 2 (Participant Sex: men vs. women) x 2 (Grammatical Gender of the word: masculine vs. feminine) x 2 (Voice: male vs. female) repeated measures ANOVAs on participant means of mean percent gender decision error and on mean response latencies measured from the end of presentation of the auditory stimuli with participant sex as a between-participants variable and word gender and voice as within-participant variables.

Accuracy In the analysis of the mean percent gender decision error, a main effect of words' grammatical gender was found, \(F(1,38)=6.77, p=.013, \eta_{p}{ }^{2}=.151\). Participants overall \((\mathrm{n}=40)\) had a lower error rate in their gender decision when responding to Feminine gender words (Mean \(=0.89 \%\), SD \(=9.38 \%\) ) than Masculine gender words (Mean \(=1.81 \%, \mathrm{SD}=13.34 \%\) ). There were no main effects of participant sex or voice. There was, however, a statistically significant two-way interaction between participant sex and word gender, \(F(2,38)=6.20, p=.017\), \(\eta_{p}{ }^{2}=.140\). Whereas men's error rate did not differ on the two grammatical genders ( \(M=1.29 \%\) and \(\mathrm{M}=1.25 \%\) for masculine and feminine words, respectively), women produced more inaccurate gender responses on words of masculine grammatical gender than of feminine grammatical gender ( \(\mathrm{M}=2.34 \%\) for masculine and \(\mathrm{M}=\) \(0.52 \%\) for feminine words). This interaction is illustrated in Figure 1. There were no further interactions.


Figure 1: Mean percent error of male and female participants in the gender decision task for words of Masculine and Feminine gender.

Response Times In the analysis of the mean reaction times, a main effect of words' grammatical gender was found again, \(F(1,38)=23.31, \mathrm{p}<.001, \eta_{p}{ }^{2}=.380\). Participants overall responded slower to words of masculine grammatical gender ( \(\mathrm{M}=576 \mathrm{~ms}, \mathrm{SD}=218 \mathrm{~ms}\) ) than to words of feminine grammatical gender ( \(M=525 \mathrm{~ms}\), \(\mathrm{SD}=\) 194 ms ). In addition, there was a main effect of participant sex, \(F(1,38)=6.26, p=.017, \eta_{p}^{2}=.141\). On average, women's decisions took longer ( \(\mathrm{M}=582 \mathrm{~ms}, \mathrm{SD}=201 \mathrm{~ms}\) ) than men's \((\mathrm{M}=512 \mathrm{~ms}, \mathrm{SD}=208 \mathrm{~ms})\). However, there was no participant sex by grammatical gender interaction on the mean reaction times.

Finally, in the analysis of reaction times, a main effect of voice emerged, \(\mathrm{F}(1,38)=125.33, \mathrm{p}<.001, \eta_{\mathrm{p}}{ }^{2}=.767\). Participants' decisions took longer for stimuli pronounced by the male voice ( \(\mathrm{M}=557 \mathrm{~ms}\), \(\mathrm{SD}=205 \mathrm{~ms}\) ) than by the female voice ( \(\mathrm{M}=536 \mathrm{~ms}\), \(\mathrm{SD}=209 \mathrm{~ms}\) ). This gender difference is likely due to acoustic characteristics of the voices such as baseline pitch, for example. More importantly, however, there was a significant two-way interaction between the independent variables of grammatical gender and voice, \(F(1,38)=36.22, \mathrm{p}<.001\), \(\eta_{\mathrm{p}}^{2}=.488\), such that responses to masculine gender words did not differ in latency with respect to the gender of the voice ( \(M=582 \mathrm{~ms}\) for male voice stimuli and \(M=570 \mathrm{~ms}\) for female voice stimuli) but, on the other hand, participants responded slower to the male voice pronouncing feminine gender words ( \(M=556 \mathrm{~ms}\) ) than to the female voice saying feminine gender words ( \(\mathrm{M}=493 \mathrm{~ms}\) ). This interaction is illustrated in Figure 2. There were no further interactions.


Figure 2: Mean response times in the gender decision task for words of Masculine and Feminine gender spoken by a male or a female voice. Note: RTs are measured from the offset of the auditory stimuli.

Analyses on Item Means A second set of analyses examined data patterns in item means. It yielded the same pattern of results as described above and will not be presented here in detail.

\section*{Discussion}

The data from Experiment 1 support the research hypothesis of an interaction between word gender and participant sex in the gender decision task. The results have shown that men and women respond in a different manner to the same stimuli in the Masculine and Feminine grammatical genders. The pattern of errors in this task is particularly revealing whereas men's gender decisions are equally accurate on both Masculine and Feminine gender words, women tend to make more errors on Masculine gender words than Feminine, a pattern that is in line with the expectation that after a lifetime of personal experience in a gender-marking language they have become sensitized to word forms that match their own gender in referential expressions. It remains to be studied further why the pattern is not mirrored clearly in the case of men and Masculine forms. One possibility is that their "own" referential gender overlaps with the unmarked, default member of the grammatical category while Feminine forms are also the marked member of the category and thus more salient generally but at this point it is hard to offer an explanation of sufficient specificity.

Secondly, a two-way interaction between word gender and voice/speaker sex on reaction times was also established. This is a novel and intriguing finding that deserves future investigation. Participants reached their gender decision faster when there was a match between the sex of the female voice and the grammatical gender of the words being heard. It appears that saliency of the feminine forms in combination with a matching 'speaker perspective' facilitated responses in this task.

\section*{Experiment 2}

The first experiment established the interaction of word gender decision times with speaker voice gender and the interaction of gender decision accuracy with the sex of participants. However, one might argue that these findings are task-specific. After all, participants were asked to reveal their linguistic competence by making a judgment on a clearly grammatical aspect of the stimuli. Although the participants in Experiment 1 performed the task with the kind of ease that shows them to be fully competent (low error rates) and highly efficient (low RTs) in making a gender decision, the question remains - would we find an interaction between word gender and each of the two extralinguistic variables of participant sex and speaker voice sex in a different task, especially one that is even more automatic and requires no conscious effort in everyday experience? One such highly automatic experimental task is cued shadowing in which participants listen to and repeat words as fast as they can. This was the task used in Experiment 2.

\section*{Method}

The design of the second experiment was exactly the same as the first experiment but this time the experimental task was cued shadowing. Participant sex (men vs. women) was a between-participant independent variable; voice (male vs. female) and word gender (Masculine vs. Feminine) were within-participant independent variables. No decision was required in this task and error rates were not of interest. The analyses were conducted on participant means and on item means of response times measured from voice onset registered by the Psyscope button box used as in Experiment 1.

Participants Another 40 participants (20 men and 20 women) took part in Experiment 2. They were university students within the 19-30 age range who were paid a modest amount for their participation. All were native speakers of Bulgarian.

Stimuli and Procedure The experimental stimuli consisted of two different gender forms each of 100 verbs and adjectives in Bulgarian presented in a different randomized order for each participant in the auditory modality by two speakers of Bulgarian, i.e., in two voices (a male and a female voice). As the cued shadowing task is a highly automatic and non-strategic one, there was no need to include neuter gender words to make the task more natural and the research objectives less obvious. Again, the analyses examined the two critical grammatical genders (Masculine and Feminine) in line with the research hypotheses.

Participants were tested individually in a sound proof booth and were asked to listen to stimuli one at a time and repeat each word. Accuracy and speed in completing the task were both emphasized in the experimental instruction.

Presentation of stimuli and registration of participants' responses were controlled by Psyscope and a button box. Before the experiment began, participants did four to six practice trials to familiarize themselves with the task and procedure.

The 200 word forms were recorded by a female and a male speaker of Bulgarian in a neutral intonation with a falling tone. They were digitized using the Macintosh SoundEdit system, and were placed in a sound file within the PsyScope experiment preparation package developed by Cohen, MacWhinney, Flatt, and Provost (1993) at Carnegie Mellon University. For each item, reaction times (RT) were measured from the offset of the stimulus word to the participant's response (voice onset as registered by the button box).

\section*{Results}

Hardly any errors were made by participants in this task except for fifty individual trials with false starts or where no response was registered by the button box. Data cleanup was accomplished in a two-step outlier procedure for the RTs of correct responses, following Balota, Yap, Cortese, Hutchison, Kessler, Loftis, et al. (2007). First, all negative response latencies, i.e., where voice onset preceded in time the end of the auditory stimulus, and all latencies longer than \(1,500 \mathrm{msec}\) were identified as outliers. Second, for the remaining RTs, the means and SDs were computed for each participant, and all RTs less than three SDs below the mean of the participant or greater than three SDs above the mean were considered outliers as well. This resulted in the rejection of \(1.87 \%\) of all reaction time data.

Here the results of the statistical analyses of participant means will be presented. The pattern of results from the analyses of the item means was the same and would be redundant to describe.

The data were analyzed in a 2 (Participant Sex: men vs. women) x 2 (Grammatical Gender of the word: masculine vs. feminine) x 2 (Voice: male vs. female) repeated measures ANOVAs on participant means of response latencies measured from the end of presentation of the auditory stimuli with participant sex as a betweenparticipants variable and word gender and voice as withinparticipant variables.

Response Times In the analysis of participants' mean reaction times, a main effect of words' grammatical gender was found again, \(\mathrm{F}(1,38)=93.68, \mathrm{p}<.001, \eta_{\mathrm{p}}{ }^{2}=.711\). Participants overall responded slower to words of masculine grammatical gender than to words of feminine grammatical gender. This result is consistent with the data pattern observed in the gender decision task. Since a decision was not involved in this task at all, however, the main effect of word gender in both experiments is likely due to the measurement of reaction times from the offset of the auditory stimuli.

There was no main effect of participant sex. In addition, a main effect of voice emerged, \(\mathrm{F}(1,38)=214.78, \mathrm{p}<.001\),
\(\eta_{\mathrm{p}}{ }^{2}=\).850. Participants' responses took longer for stimuli pronounced by the male voice than by the female voice. Again here, this gender difference is likely due to acoustic characteristics.

More importantly, however, there were two significant two-way interactions. One of them was an interaction between the independent variables of word gender and participant sex, revealing that women responded particularly fast to Feminine gender words, \(\mathrm{F}(2,38)=5.51, \mathrm{p}=.024\), \(\eta_{\mathrm{p}}{ }^{2}=.127\). This interaction is illustrated in Figure 3.


Figure 3: Mean response times by men and women in the cued shadowing task as a function of word gender. Note: RTs are measured from the offset of the auditory stimuli.

The second significant two-way interaction was between the independent variables of word gender and speaker voice, \(F(2,38)=59.33, \mathrm{p}<.001, \eta_{\mathrm{p}}^{2}=.610\). This interaction is illustrated in Figure 4. Participants responded to Feminine gender words particularly fast again when they were spoken by a female voice.


Figure 4: Mean response times in the cued shadowing task for words of Masculine and Feminine gender spoken by a male or a female voice. Note: RTs are measured from the offset of the auditory stimuli.

\section*{Discussion}

The results of experiment 2 mirror those of experiment 1 remarkably. Even though the two tasks were different in
their nature and task demands, participants' responses were strikingly similar in terms of speed of processing and the emergence of significant interactions between word gender as a grammatical feature and participant sex and speaker voice sex as extra-linguistic information sources.

\section*{Conclusion}

The studies reported here were guided by the research question whether the individual language experience of speakers may influence the way they process language online and offline, and how linguistic and extra-linguistic categories may interact as a result of this experience. In two studies with different language processing tasks and a total of 80 participants, the grammatical category of gender was found to interact with participant sex and with voice sex in the expected direction revealing the impact of individuals' usage of gender-marking forms throughout their lifetime on the speed and accuracy of gender decisions and on the efficiency of processing lexical items in a word repetition task.

It is worth pointing out here that these interaction effects were observed in the analyses of both participant means and item means and in both tasks, making them robust findings. The second task, word repetition, was particularly important in establishing the automatic, non-strategic nature of the effects, a finding that speaks in favor of extensive habitual use implicated in the emergence of these novel phenomena.

These findings have important implications of a methodological nature. Future studies of lexical processing in at least heavily gender-marking languages such as those from the Slavic and Romance language groups need to take into account the possibility of participant sex interactions with the grammatical category of gender, and implement the necessary control and counterbalancing mechanisms.

More importantly, however, the results from the two studies indicate the importance of studying individual differences in language processing and have further theoretical implications in line with the current growth of interest in embodiment and its experiential aspects and in alignment with our understanding of language and language usage being at least partially dependent on specific individual human experience instead of being an entirely arbitrary symbolic system without intrinsic or extrinsic motivation.

\section*{Acknowledgments}

I would like to thank Elizabeth Bates for support and inspiration, Radu Luchianov for assistance with the Psyscope script and Armina Janyan for consultations.

\section*{References}

Andonova, E., D’Amico, S., Devescovi, A., \& Bates, E. (2004). Gender and lexical access in Bulgarian. Attention, Perception, \& Psychophysics, 66(3), 496-507.
Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., et al. (2007). The English Lexicon Project. Behavior Research Methods, 39, 445-459.
Cohen, J., MacWhinney, B., Flatt, M., \& Provost, J. (1993). PsyScope: An interactive graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers. Behavior Research Methods, Instruments, \& Computers, 25, 257-271.
Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., \& Pethick, S. J. (1994). Variability in early communicative development. Monographs of the Society for Research in Child Development, 59, 1-185.
Kramer, J. H., Kaplan, E., Delis, D. C., O’Donnell, L., \& Prifitera, A. (1997). Developmental sex differences in verbal learning. Neuropsychology, 11(4), 577-584.
Shaywitz, B. A., Shaywitz, S. E., Pugh, K. R., Constable, R. T., Skudlarski, P., Fulbright, R. K., et al. (1995). Sex differences in the functional organization of the brain for language. Nature, 373(6515), 607-609.
Wallentin, M. (2009). Putative sex differences in verbal abilities and language cortex: A critical review. Brain and language, 108(3), 175-183.
Weiss, E. M., Kemmler, G., Deisenhammer, E. A., Fleischhacker, W. W., \& Delazer, M. (2003). Sex differences in cognitive functions. Personality and Individual Differences, 35(4), 863-875.

\title{
Why and How to Measure the Association between Choice Options
}

\author{
Sandra Andraszewicz (s.andraszewicz@unibas.ch) \\ University of Basel \\ Department of Psychology, Missionsstrasse 62A \\ Basel, 4055 Switzerland
}

\author{
Jörg Rieskamp (joerg.rieskamp@unibas.ch) \\ University of Basel \\ Department of Psychology, Missionsstrasse 62A \\ Basel, 4055 Switzerland
}

\begin{abstract}
Prominent theories of decision making, such as proportional difference model, priority heuristics, decision field theory and regret theory assume that people do not evaluate options independently of each other. Instead, these theories predict that people compare the options' outcomes with each other. Therefore the theories' predictions strongly depend on the association between outcomes. In the present work, we examine how the association between options can be best described. For options with two outcomes the standard correlation measure between option's outcomes does not provide a meaningful interpretation. Therefore, we propose the standardized covariance between options A and B , denoted as \(\sigma_{A B}^{*}\). We describe the properties and interpretation of this measurement and show its similarities and differences with the correlation measurement. Finally, we show how the predictions of different models of decision making vary depending on the value of the standardized covariance.
\end{abstract}

Keywords: decision making models; covariance; gambles; two-outcome; risky choice

\section*{Introduction}

Standard economic models of decision making like expected utility theory assume that people evaluate choice options independently of each other (Neumann \& Morgenstern, 1944). However, contrary to this basic independence assumption a vast amount of evidence has shown that people evaluate choice option depending on the set of alternative choice options (see Rieskamp, Busemeyer, \& Mellers, 2006 for more details).

For example, choice of between two health insurance offers is a choice between risky options with payouts depending on the occurence of an illness. An illness can occur with a certain probability that can be estimated based on teh patient's age and health history. When deciding between the two options, one would probably compare the insurance coverage in case of specific illnesses of both offers with each other, rather than first evaluate one offer and then another.

Many cognitive models of decision making assume that when people make choices between options they compare the options' outcomes with each other. For instance, the priority heuristics (Brandstätter, Giegerenzer, \& Hertwig, 2006) assumes that people first compare all options with respect to their minimum outcomes. If these outcomes do not allow to discriminate the options, the options are compared with respect to the probability of the minimum outcomes, and so forth. Regret theory (Loomes \& Sugden, 1982), proportional
difference model (González-Vallejo, 2002), and decision field theory (Busemeyer \& Townsend, 1993) are three other prominent models of decision making assuming that people compare options with respect to their outcomes. These comparisons are then accumulated to form an overall preference. Therefore the predictions of all these models depend on the associations between the outcomes of options.

How can these associations be best characterized? Using the covariance of the options' outcomes would be a sound solution, especially because many of the studies of decision making use monetary choices which are an analogy to investments. Indeed, in portfolio management covariance plays an important role for selection of assets (i.e. Pafka \& Kondor, 2003; Disatnik \& Benninga, 2007).

Also, many studies investigating decision making focus on comparing various models. In such case, the selection of the choice options, which are usually presented as gambles, is very important. As highlighted in the work on optimal experimental design, selecting gambles for discriminating between the models, is an essential issue that determines the effectiveness of the experiment (see Cavagnaro, Gonzalez, Myung, \& Pitt, 2013; Myung \& Pitt, 2009; Zhang \& Lee, 2010) The main problem with using covariance is that its value depends on the range of the outcomes' values which makes it hard to interpret.

As an alternative measurement one could use the correlation between the outcomes. However, a large part of the research is done with two-outcome choice problems (e.g. González-Vallejo, 2002; Birnbaum, 2008), for which correlation is either 1 or -1 (see Rodgers \& Nicewander, 1988), so that the correlation measurement does not provide a meaningful interpretation.

Therefore, we propose an alternative measurement, standardized covariance, as a measure of the strength of the association between the options' outcomes, which can be easily interpreted. In the following sections we will first explain how the standardized covariance is determined and how it should be understood. Next, we will present relations between the covariance, variances and expected value of twooutcome gambles, which will clarify the construction of the standardized covariance. Finally, we will show how decision making models make different predictions depending on the strength of the standardized covariance.

Table 1: Twelve examples of choice options with different standardized covariance. In each example, the top row indicates the probability of the occurrence of two outcomes. Two consecutive rows display the possible outcomes of Option A and Option B.
\begin{tabular}{|c|c|c|c|c|c|}
\hline \multicolumn{3}{|c|}{\(\sigma_{A B}^{*}=1\)} & \multicolumn{3}{|c|}{\(\sigma_{A B}^{*}=-1\)} \\
\hline Example 1 & 60\% & 40\% & Example 2 & 60\% & 40\% \\
\hline A & 80 & 55 & A & 80 & 55 \\
\hline B & 80 & 55 & B & 55 & 80 \\
\hline \multicolumn{3}{|c|}{\(\sigma_{A B}^{*}=.80\)} & \multicolumn{3}{|c|}{\(\sigma_{A B}^{*}=-.80\)} \\
\hline Example 3 & 60\% & 40\% & Example 4 & 60\% & 40\% \\
\hline A & 80 & 55 & A & 80 & 55 \\
\hline B & 80 & 30 & B & 30 & 80 \\
\hline \multicolumn{3}{|c|}{\(\sigma_{A B}^{*}=.80\)} & \multicolumn{3}{|c|}{\(\sigma_{A B}^{*}=1\)} \\
\hline Example 5 & 40\% & 60\% & Example 6 & 40\% & 60\% \\
\hline A & 80 & 55 & A & 80 & 55 \\
\hline B & 80 & 30 & B & 70 & 45 \\
\hline \multicolumn{3}{|c|}{\(\sigma_{A B}^{*}=.32\)} & \multicolumn{3}{|c|}{\(\sigma_{A B}^{*}=.05\)} \\
\hline Example 7 & 60\% & 40\% & Example 8 & 60\% & 40\% \\
\hline A & 80 & 20 & A & 42 & 40 \\
\hline B & 50 & 40 & B & 80 & 6 \\
\hline
\end{tabular}

\section*{Standardized Covariance}

We denote standardized covariance of a pair of dependent choice options A and B, with each having two possible outcomes, as \(\sigma_{A B}^{*}\), where the non-standardized covariance is denoted as \(\sigma_{A B}\). Standardized covariance is equal to twice the covariance divided by the sum of the variances \(\sigma_{A}\) and \(\sigma_{B}\) of each of the options (see Equation 1).
\[
\begin{equation*}
\sigma_{A B}^{*}=\frac{2 \sigma_{A B}}{\sigma_{A}^{2}+\sigma_{B}^{2}} \tag{1}
\end{equation*}
\]

For stochastically non-dominant options, \(\sigma_{A B}^{*}\) is a continuous variable ranging from -1 (strong negative association) to 1 (strong positive association). When \(\sigma_{A B}^{*}=0\) either the options are completely unrelated (e.g. they are statistically independent, where two options do not depend on one external event) or the covariance between the options' outcomes is equal to 0 . The second case occurs, when one of the options is a sure thing. When \(\sigma_{A B}^{*}\) approaches 0 , the variances of both options are low, and so is the association between the options.

\section*{Properties of \(\sigma_{A B}^{*}\)}

When \(\sigma_{A B}^{*}\) reaches its maximum at 1 then the sum of the variances equals twice the covariance:
\[
\begin{equation*}
\sigma_{A B}^{*}=1 \Longleftrightarrow 2 \sigma_{A B}=\sigma_{A}^{2}+\sigma_{B}^{2} \tag{2}
\end{equation*}
\]

Analogically, for negatively related gambles the relation is:
\[
\begin{equation*}
\sigma_{A B}^{*}=-1 \Longleftrightarrow-2 \sigma_{A B}=\sigma_{A}^{2}+\sigma_{B}^{2} \tag{3}
\end{equation*}
\]

Situation from Equation 2 occurs when both options are the same (Example 1 in Table 1) and, analogically, by interchanging the outcomes of Option B we can obtain choice options for which \(\sigma_{A B}^{*}=-1\) (Example 2), which shows the symmetricity of options with positive and negative \(\sigma_{A B}^{*}\). Further, as shown in Examples 3 and 4, by altering one outcome from Option B so that the options are not the same any more, we obtain options with slightly lower \(\sigma_{A B}^{*}\). Interrestingly, the probabilities of the outcomes do not influence \(\sigma_{A B}^{*}\) (compare Examples 3 and 5). Also, the "perfect association" does not occur only when the options are identical, but also, when the difference between outcomes of option A and B is the same and this difference is the difference between expected values (Example 6 with the difference in expected values of 10 points). By making the outcomes corresponding to the same probabilities more dissimilar, one can decrease \(\sigma_{A B}^{*}\) (compare Examples 3 and 7). Finally, as shown in Example 8, when outcomes of one option are almost the same (almost a sure thing), while outcomes of the other option are dissimilar, \(\sigma_{A B}^{*}\) is almost 0 .

For stochastically non-dominant options, \(\sigma_{A B}^{*}\) is not higher than 1 or lower than -1 because it is not true that \(2 \sigma_{A B}>\) \(\left(\sigma_{A}^{2}+\sigma_{B}^{2}\right)\). Below, we provide the mathematical proof.

Proof. \(2 \sigma_{A B}>\left(\sigma_{A}^{2}+\sigma_{B}^{2}\right)\) is false
In stochastically non-dominant options the variances of options' outcomes are unequal, thus \(\sigma_{A}^{2}<\sigma_{B}^{2} \vee \sigma_{A}^{2}>\sigma_{B}^{2}\).
\(\sigma_{A}^{2}<\sigma_{B}^{2} \Longleftrightarrow \sigma_{B}^{2}=\sigma_{A}^{2}+s \wedge s \in \mathbb{R}^{+}\)
Then,
\(2 \sigma_{A B}>\sigma_{A}^{2}+\sigma_{B}^{2}\)
\(2 \sigma_{A B}>2 \sigma_{A}^{2}+s\)
\(0>\sigma_{A}^{2}+\frac{s}{2}-\sigma_{A B}\)
\(0>E\left[a^{2}\right]+\frac{E\left[b^{2}\right]-E\left[a^{2}\right]}{2}-E[a b]\)
\(E\left[a^{2}\right]<E\left[b^{2}\right] \Longleftrightarrow b=a+g \wedge g \in \mathbb{R}^{+} \wedge g=\) const.
By expanding the inequality we get
\(0>\frac{g^{2}}{2}\)
Since \(g^{2}>0\) the inequality is false.

\section*{Standardized Covariance vs. Correlation}

Examples presented in Table 1 indicate that there are similarities between correlation and standardized covariance and the "perfect" correlation overlaps with the "perfect" standardized covariance (e.g. Examples 1, 2 and 3 in Table 1). Correlation coefficient \(r\) equals to
\[
\begin{equation*}
r=\frac{\sigma_{A B}}{\sigma_{A} \sigma_{B}} . \tag{4}
\end{equation*}
\]

The relation between correlation coefficient and standardized covariance is as follows:
\[
\begin{equation*}
\sigma_{A B}^{*}=\frac{2 r \sigma_{A} \sigma_{B}}{\sigma_{A}^{2}+\sigma_{B}^{2}} \tag{5}
\end{equation*}
\]
and correlation is equal to standardized covariance when
\[
\begin{equation*}
2 \sigma_{A} \sigma_{B}=\sigma_{A}^{2}+\sigma_{B}^{2} . \tag{6}
\end{equation*}
\]

Because for stochastically non-dominant options with two outcomes correlation is always either -1 or \(1, \sigma_{A B}=\sigma_{A} \sigma_{B}\). Therefore, the relation between the sum of variances and the product of the standard deviation is the same as the relation between the sum of variances and the covariance. Thus, standardized covariance could also be written as
\[
\begin{equation*}
\sigma_{A B}^{*}=\frac{2 \sigma_{A} \sigma_{B}}{\sigma_{A}^{2}+\sigma_{B}^{2}} \tag{7}
\end{equation*}
\]
when the association between the options' outcomes is positive. When the association between the options' outcomes is negative, \(\sigma_{A B}=-2 \sigma_{A} \sigma_{B}\).

\section*{Variances and Covariance of Two-Outcome Options}

In stochastically non-dominant pairs of options which are not identical, one option has higher variance than the other (compare range of outcomes of options A and B in Table 1 in Examples 3, 4, 5, 7 and 8, to Example 6 which contains a stochastically dominant pair of options). Therefore the sum of variances is composed of a smaller and larger variance.

The relation between the smaller variance and covariance is close to linear and is symmetric with respect to the x -axis. In contrast, the relation of the larger variance to covariance takes the shape of a triangular area and is also symmetric with respect to the x -axis, as shown in Figure 1. This figure shows a very interesting pattern in which the graph on the right side fits into the graph on the left side like "key and lock". The data points in Figure 1 were obtained from 100000 randomly generated two-outcome choice options, with various differences between expected values. The outcomes' values ranged between 1 and 100, and probabilities of their occurrence ranged between 1 and \(99 \%\).

The relation between the sum of variances and covariance is symmetric with respect to the x -axis (Figure 2). The gray data points, which lay on the diagonal of the graph in Figure 2, are the only ones, for which standardized covariance is equal to correlation. This is the key property of standardized covariance, as the gray points correspond to the "perfect correlation" between the options, for which equation 6 holds. In contrast, all black points represent the pairs of options whose relation varies between -1 and 1 (not perfect correlation).

\section*{Options with Negative and Mixed Outcomes}

Until now, we have discussed the properties of standardized covariance, covariance and variances of options which generate only positive outcomes. However, some experiments might include choice options which generate only losses or might generate both, gains and losses. As a consequence, we randomly generated stochastically non-dominant options with only negative outcomes ( \(N=100000\) ), to which we will refer as negative options, and options that have one positive


Figure 1: Left: relation between the lower variance (more secure option) and covariance between two-outcome options, Right: relation between the higher variance (more risky option) and covariance between two-outcome options.


Figure 2: Relation of the sum of variances to twice the covariance. Gray points indicate the cases for which standardized covariance overlaps with correlation, such that \(\sigma_{A B}^{*}=r=1\) or \(\sigma_{A B}^{*}=r=-1\).
and one negative outcome ( \(N=100000\) ), which we call mixed options. The outcomes of negative options varied between -100 and -1 points, while the outcomes of mixed options were in range \([-100,1]\) and \([1,100]\) points. The probabilities of these outcomes ranged from \(1 \%\) to \(99 \%\). We repeated the analysis of the relation between variances and covariance, as well as the sum of variances and the product of standard deviations of the options for the two new types of options.

The obtained results for the negative and mixed options were the same as for the positive options. Thus, the properties of the choice options regarding their variances, covariance and the standardized covariance, depicted in Figures 1 and 2, apply to various kinds of choice options. The only difference is that the range of the values of variances and covariance of mixed options is much greater (range: \([-10000,10000]\) ). This is due to the fact that the range of the possible values is twice as big compared to the positive and negative options. In sum, standardized covariance is a stable measure of association between options.

Table 2: Ranges of values of standardized covariance, ratio of the smaller to the larger variance and the amount of pairs of options generated for each of the five differences between expected values between the options.
\begin{tabular}{lccc}
\hline\(\Delta E V\) & \(\sigma_{A B}^{*}\) & \(\frac{\min \left(\sigma_{A}^{2}, \sigma_{B}^{2}\right)}{\min \left(\sigma_{A}^{2}, \sigma_{B}^{2}\right)}\) & N \\
\hline 10 & \(.02-.99\) & \(.00-.72\) & 975021 \\
15 & \(.02-.94\) & \(.00-.49\) & 377750 \\
20 & \(.02-.94\) & \(.00-.49\) & 418251 \\
25 & \(.02-.94\) & \(.00-.49\) & 244734 \\
30 & \(.02-.89\) & \(.00-.36\) & 119241 \\
\hline
\end{tabular}

\section*{Standardized Covariance vs. Expected Value}

Standardized covariance is sensitive to the differences between expected values of two choice options. We generated all possible pairs of options with outcomes and probabilities as previously, for which the difference between the expected values was either \(10,15,20,25\) or 30 . As shown in Table 2, the greater the difference between expected values, the more narrow the range of possible values of \(\sigma_{A B}^{*}\). Thus, when manipulating the difference between the expected value of pairs of options, our analysis shows that this manipulation will most likely also change the covariance of the options outcomes. Thus, when not controlling for this aspect, then variations of the expected value differences will often be confounded with variations of covariance differences. Therefore, the experimenters should keep in mind that the strength of the association between the gambles that they present to the participants may depend on the differences between expected values ( \(\Delta E V\) ).

Interestingly, the greater the difference between the expected values, the fewer choice options could be obtained (see Table 2). Also, the greater the difference between expected values, the more narrow the ranges of possible values of standardized covariance and ratio between the lower and the higher variance within the pair of options (see Table 2). Therefore, in experiments that control for the expected value difference, it might be the standardized covariance between the options that influences people's choice, rather than the expected value.

Further, we selected a group of options for which \(\Delta E V=\) 15. For these options with fixed difference between expected values, we tested the relation between the variances of both options. As shown in Figure 3, the data points create a pattern that is symmetric with respect to the diagonal of the graph. In other words, when \(\Delta E V\) is fixed, the variances of both options are related to each other with respect to a certain ratio, whose ranges we listed in Table 2.

\section*{Options with More than Two Outcomes}

In order to analyze in more detail the relation between the correlation measure and the standardized covariance, one would have to extend the problem to choices with more than two


Figure 3: Relation between variances of two options with two outcomes.


Figure 4: Left: relation between the smaller variance and the covariance between four-outcome options. Right: relation between the larger variance and the covariance between four-outcome options.
possible outcomes. Therefore, we generated 10000 pairs of stochastically non-dominant options with four outcomes. The outcomes varied between 1 and 100 points, and probabilities varied betweeen \(1 \%\) and \(40 \%\).

Firstly, we investigated the relations between the variances and the covariance. As shown in Figure 4, the relations between both variances and covariance do not display the "keylock" pattern as in Figure 1, and the patterns are not symmetrical. Analogically, the relation between the sum of variances and twice the covariance is not symmetric with respect to the x -axis.

Secondly, we looked at the relation between correlation and standardized covariance. Figure 5 shows a strong relation between the correlation measure and standardized covariance. Pearson correlation between these two measures is very strong, \(r=.98, p<.001\). In the current sample of generated pairs of options, covariance and correlation are equal to each other for \(24 \%\) of the cases. Also, the slope of the regression line is high and the intercept very small (see caption of Figure 5). Thus, standardized covariance is a similar measure as correlation, but it has the advantage that it can be


Figure 5: Relation between the correlation coefficient and standardized covariance of options with four outcomes. Gray line indicates the regression line, with the slope of .87 and intercept 0027.
applied to both two-outcome choice options and options with several outcomes.

\section*{Model Predictions Depending on the Standardized Covariance}

In the previous sections, we have described features of stochastically non-dominant options with two outcomes and how one could measure the association between the options with the use of standardized covariance. In this section, we will show why the association between two options is important, based on two prominent theories of decision making, regret theory and decision field theory. Models of decision making generate different predictions for options with various differences between expected values. Thus, we focused on the choice options for which \(\Delta E V=15\).

For all of these options we generated predictions of the two models. Following Pathan, Bonsall, and Jong (2011), we defined the regret function of choosing option A over option B with outcomes \(x_{i}, i \in\{1,2\}\) as
\[
\begin{equation*}
R_{i A}=\ln \left(1+\exp \left(\beta\left(x_{i}-\max \left(x_{i A}, x_{i B}\right)\right)\right)\right) \tag{8}
\end{equation*}
\]

The total regret from choosing option A equals to
\[
\begin{equation*}
R_{A}=\Sigma_{i=1}^{2} R_{i A} \tag{9}
\end{equation*}
\]

Further, the probability of choosing option A over option B is estimated using softmax rule
\[
\begin{equation*}
\operatorname{Pr}(A \mid A, B)=\frac{1}{1+\exp \left(\theta\left(R_{B}-R_{A}\right)\right)} \tag{10}
\end{equation*}
\]
\(\beta\) and \(\theta\) are free parameters of the model. More details regarding regret theory is provided in Loomes and Sugden (1982). A parsimonious version of decision field theory was used, as described by Busemeyer and Townsend (1993).

The models' predictions are expressed as probabilities of choosing option A over B . We converted these results to the prediction that the option with the higher expected value


Figure 6: Average predictions of regret theory and decision field theory. To generate predictions the following parameters were used: regret theory \(\beta=.05, \theta=4.6\), decision field theory \(\theta=1.19\). The parameter of decision field theory was based on Rieskamp (2008) and the parameters of regret theory were adjusted so that the predictions of both models are at the same level. Error bars indicate standard deviations.
would be chosen, which resulted in all predictions ranging between 0.5 and 1 . Next, we grouped the options depending on their standardized covariance, such that group 1: \(\sigma_{A B}^{*}<0.2(21.2 \%)\), group 2: \(0.2<\sigma_{A B}^{*} \leq 0.5\) ( \(34.7 \%\) ), group 3: \(0.5<\sigma_{A B}^{*}(44.2 \%)\). For each of the three groups we calculated the mean prediction and its standard deviation.

As shown in Figure 6, the models' predictions differ among the three groups. This constitutes evidence that some theories of decision making not only assume on a theoretical level that the relation between the options' outcomes play an important role in decision making, but also provide quantitative evidence. Therefore, one should control for the association between the options. This is a crucial property of the standardized covariance, because in experiments in which the association between options was not examined, the results might depend on the selected set of choice options.

Furthermore, the models' predictions for choices with the same level of association could differ depending on the difference between expected values. From each group of options with difference between expected values of \(10,15,20\), 25 and 30 points, we picked all options for which \(\sigma_{A B}^{*}=.3\) and we generated models' predictions using the same set of parameters as previously. As shown in Figure 7, decision field theory makes very systematic predictions in which the higher the expected value difference, the higher the probability of choosing the option with the higher expected value. In contrast, regret theory indicates some differences but no trend can be observed.

In sum, predictions of models of decision making result from the interaction between the difference between expected values and the strength of the association between the choice options. This finding is very important, as in most studies


Figure 7: Average predictions of regret theory and decision field theory for choices with \(\sigma_{A B}^{*}=.3\) and various expected values. Error bars indicate standard deviations.
the researchers do not even consider the covariance of the options' outcomes, but only report the differences in expected values.

\section*{Discussion}

The association between two options' outcomes may play an important role in testing models of decision making. As we have shown, models can generate different predictions depending on the combination of the expected value difference and the association between the options. Experiments that control only for the expected value difference may obtain confounded results.

Therefore, a simple measure of the strength of this association is needed. For experiments that employ two-outcome choice options, we proposed the standardized covariance. Its values range between -1 and 1 , where 1 indicates "perfect positive association" and -1 indicates "perfect negative association". When standardized covariance equals to 0 , one of the options is a sure thing. When standardized covariance equals to \(-1,0\) and 1 it overlaps with correlation.

There is a strong association between correlation measure and standardized covariance. This constitutes solid evidence in favor of the reliability of the standardized covariance as a measure of the association between two choice options. Interestingly, there are very clear patterns of relations between variances and covariance of the two-outcome options. In contrast, these patterns are different when there are more outcomes. Therefore, future empirical research is needed to test the applicability of the standardized covariance and its perception by human decision makers. Also, as a future investigation, we suggest that one should test whether the predictions of the aforementioned models of decision making reflect the real human choice behavior.

In sum, this work was based on extensive simulations of random choice options and choice options with specific properties. We have shown that standardized covariance is a ro-
bust measure, with similar properties to the correlation. Finally, we showed that the covariance strongly influences the prediction of different cognitive models of decision making and should be given more attention in empirical work.

\section*{References}

Birnbaum, M. H. (2008). New paradoxes of risky decision making. Psychological Review, 463-501.
Brandstätter, E., Giegerenzer, G., \& Hertwig, R. (2006). The priority heuristic: Making choices without trade-offs. Psychological Review, 113(2), 409-432.
Busemeyer, J. R., \& Townsend, J. T. (1993). Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. Psychological Review, 100(3), 432-459.
Cavagnaro, D. R., Gonzalez, R., Myung, J. I., \& Pitt, M. A. (2013). Optimal decision stimuli for risky choice experiments: An adaptive approach. Management Science, 59(2), 358-375.
Disatnik, D., \& Benninga, S. (2007). Shrinking the covariance matrix. The Journal of Portfolio Management, 33(4), 55-63.
González-Vallejo, C. (2002). Making trade-offs: A probabilistic and context-sensitive model of choice behavior. Psychological Review, 109(1), 137.
Loomes, G., \& Sugden, R. (1982, December). Regret theory: An alternative theory of rational choice under uncertainty. Economic Journal, 92(368), 805-24.
Myung, J. I., \& Pitt, M. A. (2009). Optimal experimental design for model discrimination. Psychological review, 116(3), 499.
Neumann, J. V., \& Morgenstern, O. (1944). Theory of games and economic behavior. Princeton University Press.
Pafka, S., \& Kondor, I. (2003). Noisy covariance matrices and portfolio optimization ii. Physica A: Statistical Mechanics and its Applications, 319, 487-494.
Pathan, A., Bonsall, P., \& Jong, G. de. (2011, June). How the choice of mode is conditioned by information source. In International choice modelling conference, leeds, \(u k\).
Rieskamp, J. (2008). The probabilistic nature of preferential choice. Journal of Experimental Psychology: Learning, Memory, and Cognition, 34(6), 1446.
Rieskamp, J., Busemeyer, J. R., \& Mellers, B. A. (2006, September). Extending the bounds of rationality: Evidence and theories of preferential choice. Journal of Economic Literature, 44(3), 631-661.
Rodgers, J. L., \& Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. The American Statistician, 42, 59-66.
Zhang, S., \& Lee, M. D. (2010). Optimal experimental design for a class of bandit problems. Journal of Mathematical Psychology, 54(6), 499-508.

\title{
Probabilistic Language Modeling with Hidden Stochastic Automata
}

\author{
Mark Andrews \\ Division of Psychology \\ Nottingham-Trent University \\ and \\ Division of Psychology and Language Sciences \\ University College London
}

\begin{abstract}
In this paper, we introduce a novel dynamical Bayesian network model for probabilistic language modeling. We refer to this as the Hidden Stochastic Automaton. This model, while based on a generalization of the Hidden Markov model, has qualitatively greater generative power than either the Hidden Markov model itself or any of its existing variants and generalizations. This allows the Hidden Stochastic Automaton to be used as a probabilistic model of natural languages in a way that is not possible with existing dynamical Bayesian networks. Its relevance to Cognitive Science is primarily as a computational - in the Marr (1982) sense of the term model of cognition, but potentially also as a model of resource bounded cognitive processing, and as a model of the implementation of computation in physical dynamical systems.
\end{abstract}

A probabilistic language model is a hypothetical generative model of a language, where a language is defined most generally as a set of strings concatenated out of a finite set of symbols. By far the most widely used formalisms for specifying probabilistic language models are stochastic grammars, which are symbol rewriting rules with accompanying probabilities. The use of grammars is motivated by the fact that human languages are structurally complex, with properties that place them between the so-called context-free and contextsensitive formal languages (see, e.g., Chomsky, 1956, 1963; Shieber, 1985), and formal grammars are computationally universal in the sense that they can generate any recursively enumerable set (see, e.g., Hopcroft, Motwani, \& Ullman, 2001).

By contrast to the case of language modeling, in probabilistic modeling more generally, the most widely used formalism for specifying probabilistic models is the graphical model (see, e.g., Koller \& Friedman, 2009; Jordan, 2004). Graphical models are directed or undirected graphs whose vertices are identified with random variables and whose edges indicate conditional dependencies. The appeal of graphical models is their flexibility to represent complex relationships between large numbers of variables, and their graph-theoretic properties that afford general and computationally efficient algorithms for probabilistic inference, whether exactly or approximately by, for example, Monte Carlo methods. As a result, graphical models have effectively become a graph-based modeling language for developing and extending probabilistic models. They have had widespread application in fields such as bioinformatics (e.g., Fried-
man, 2004), computer vision (e.g., Oliver, Rosario, \& Pentland, 2000), machine learning (e.g., Bishop, 2006, 2013), expert systems (e.g., Lauritzen \& Spiegelhalter, 1988; Pearl, 1988), information retrieval (e.g., Salakhutdinov \& Hinton, 2009), and in cognitive science (see, e.g., Chater \& Oaksford, 2008; Griffiths, Chater, Kemp, Perfors, \& Tenenbaum, 2010, for overviews).

Despite their breadth of appeal, graphical models have had a rather limited role as language models, if by language models we specifically mean generative models of language. There are at least two important reasons for this. On the one hand, stochastic grammars can not, in general, be represented as graphical models. (In some cases, notably stochastic regular grammars, the terminal and nonterminal variables of the grammar can be identified with vertices of a directed Markovian graph. For the super-regular grammars, however, this is not the case and the variables of the grammar can not be identified with the vertices of any fixed graph). On the other hand, the most widely used graphical models for sequential probabilisitic modeling, including the Hidden Markov model and its extensions, are limited in their generative power to the regular languages (i.e. the Type3 languages in the Chomsky hierarchy). In other words, graphical models have had a relatively limited role as language models because the most widely used probabilistic models that have sufficient generative power to model human languages can not be represented as graphical models, and the most widely used graphical models for sequential structures do not have sufficient generative power to model natural languages.

There is, however, no inherent limitation to the generative power of graphical models. In this paper, we introduce a graphical model, specifically a dynamical Bayesian network, whose generative power is equivalent to that of an arbitrary stochastic grammar. This model, that we will refer to as the Hidden Stochastic Automaton, is based on a novel generalization of the widely used Hidden Markov model. As such, it retains many of the appealing characteristics of the Hidden Markov model while extending its generative power.

\section*{Hidden Stochastic Automata}

To introduce the Hidden Stochastic Automaton (HSA), it is necessary to first briefly describe the Hidden Markov model (HMM). Given a set of \(J\) independent discrete
valued sequences \(\mathbf{w}_{1}, \mathbf{w}_{2} \ldots \mathbf{w}_{j} \ldots \mathbf{w}_{J}\), where the \(j\) th sequence is \(\mathbf{w}_{j}=w_{j 1}, w_{j 2} \ldots w_{j i} \ldots w_{j n_{j}}\), the generative model assumed by the HMM treats each \(w_{j i}\) as drawn from one of \(K\) discrete probability distributions \(\phi_{1}, \phi_{2} \ldots \phi_{k} \ldots \phi_{K}\) over a finite vocabulary of length \(V\). Which distribution is chosen for \(w_{j i}\) is determined by the value of the unobserved variable \(x_{j i} \in\{1 \ldots K\}\) that corresponds to \(w_{j i}\). For all \(j\), each \(x_{j 1}, x_{j 2} \ldots x_{j i} \ldots x_{j n_{j}}\) is a first-order Markov chain, with initial distribution \(\pi\) and a \(K \times K\) transition matrix \(\theta\). More formally, the HMM assumes that for all \(j\),
\[
\begin{aligned}
w_{j i} \mid x_{j i}, \phi & \sim \text { Categorical }\left(w_{j i} \mid \phi_{x_{j i}}\right) & & 1 \leq i \leq n_{j}, \\
x_{j i} \mid \pi & \sim \operatorname{Categorical}\left(x_{j i} \mid \pi\right) & & i=1, \\
x_{j i} \mid x_{j i-1}, \theta & \sim \operatorname{Categorical}\left(x_{j i} \mid \theta_{x_{j i-1}}\right) & & 1<i \leq n_{j} .
\end{aligned}
\]

The graphical model for the HMM is shown below.


Figure 1: The graphical model or dynamical Bayesian network for the Hidden Markov model. The shaded nodes indicate the observed variables. For simplicity, we have omitted the priors on \(\phi, \pi\) and \(\theta\).

Precisely because graphical models naturally afford generalizations and extensions, the HMM has lead to many variants. Most notably, these include the mixed memory HMM (Saul \& Jordan, 1999), the coupled HMM (Brand, Oliver, \& Pentland, 1997), the factorial HMM (Ghahramani \& Jordan, 1997), and the hierarchical HMM (Fine, Singer, \& Tishby, 1998). These extensions are often based on introducing additional chains of latent variables with varying degrees of conditional independence between them. Despite the evident value of these models, they do not qualitatively alter the formal generative complexity of the underlying model. In all of these extensions, the sequences generated are equivalent to regular or Type-3 formal languages

\section*{From HMM's to Hidden Stochastic Automata}

It is possible, however, to generalize the HMM in such a way that its generative complexity is increased. This can
be done by replacing the single valued \(x_{j i}\) in the HMM by a variable sized array or vector. In other words, while in the HMM, each state variable is \(x_{j i} \in\{1 \ldots K\}\), this may be generalized to \(x_{j i} \in\{1 \ldots K\}^{*}\). Here * indicates Kleene star, or the union of all concatenations of the elements from \(\{1 \ldots K\}\) and \(\{\emptyset\}\). This change clearly increases the cardinality of the state space to a countably infinite set. Importantly, however, as we will elaborate, if the set of operations that can increase or decrease the state-vector are limited to a finite set, and if the the conditional dependencies on this state-vector are limited to a finite range of elements, then inference in this generalized model is almost identical in kind to inference in the standard HMM.

For reasons that will be made clear, we will collectively refer to generalizations of the HMM using a state-vector as Hidden Stochastic Automata (HSA). For the purposes of this paper, however, we will mostly concentrate on one specific form of the HSA. For simplicity, we will also refer to this particular case of the model as the HSA, with the understanding that it is but one of many variants based on the same principles.

Just as with the HMM, the HSA is a generative model of discrete valued sequences. It assumes that each variable \(w_{j i}\) in the sequence of observations \(\mathbf{w}_{j}=\) \(w_{j 1}, w_{j 2} \ldots w_{j i} \ldots w_{j n_{j}}\) is drawn from one of \((H+1) \times K\) discrete probability distributions \(\phi_{01}, \phi_{02} \ldots \phi_{h k} \ldots \phi_{H K}\) over a length \(V\) vocabulary. Which of these \((H+1) \times K\) distributions is chosen is determined by the values of two latent or unobserved state variables that correspond to \(w_{j i}\). On the one hand, there is a standard finite state variable \(x_{j i} \in\{1 \ldots K\}\). On the other hand, there is an additional state-vector variable \(z_{j i} \in\{1 \ldots H\}^{*}\), with \(w_{j i}\) being conditional on only the first element of \(z_{j i}\), if \(z_{j i} \neq \emptyset\). In other words, \(w_{j i}\) is sampled from \(\phi_{\left[z_{j i}^{1}, x_{j i}\right]}\), where \(z_{j i}^{1}\) indicates the value of the first element of the state-vector \(z_{j i}\), or else 0 when \(z_{j i}=\emptyset\).

For all \(j\), the sequence \(\left(x_{j 1}, z_{j i}\right),\left(x_{j 2}, z_{j i}\right) \ldots\) \(\left(x_{j i}, z_{j i}\right) \ldots\left(x_{j n_{j}}, z_{j i}\right)\) is a first-order Markov chain of coupled state variables. The distribution over \(x_{j 1}\) is given by the \(K\) valued distribution \(\pi\), and the value of \(z_{j 1}\) is deterministically set to \(z_{j 1}=\emptyset\). For \(1<i \leq n_{j}\), both \(x_{j i}\) and \(z_{j i}\) are conditional on \(x_{j i-1}\) and, if \(z_{j i} \neq \emptyset\), the first element of \(z_{j i}\). The value of \(x_{j i}\) is determined by sampling from the \(K\) dimensional probability distribution specified by \(\theta_{\left[z_{j i-1}^{1}, x_{j i-1}\right]}\), where \(\theta\) is a \((H+1) \times K \times K\) stochastic transition matrix, and \(z_{j i-1}^{1}\) is as above. The value of \(z_{j i}\) is determined by applying one of \(H+1\) different operations to \(z_{j i-1}\), specifically prepending \(z_{j i-1}\) by one symbol from \(\{1 \ldots H\}\) or removing the first element from \(z_{j i-1}\). For example, if \(\sigma_{1} \sigma_{2} \sigma_{3}\) (with each \(\sigma_{l} \in\{1 \ldots H\}\) ) is the value of the state-vector \(z_{j i-1}\), a possible sequence of
operations and their effect on the state-vector could be
\[
\begin{aligned}
& z_{j i-1}=\sigma_{1} \sigma_{2} \sigma_{3} \xrightarrow{\text { remove }} z_{j i}=\sigma_{2} \sigma_{3}, \\
& \quad z_{j i}=\sigma_{2} \sigma_{3} \xrightarrow{\text { prepend } 3} z_{j i+1}=3 \sigma_{2} \sigma_{3}, \\
& z_{j i+1}=3 \sigma_{2} \sigma_{3} \xrightarrow{\text { prepend } 2} z_{j i+2}=23 \sigma_{2} \sigma_{3} .
\end{aligned}
\]

Which of these \(H+1\) operations is applied is determined by sampling from the \(H+1\) dimensional probability distribution specified by \(\Omega\left[z_{j i-1}^{1}, x_{j i-1}\right]\), where \(\Omega\) is a \((H+1) \times K \times(H+1)\) stochastic transition matrix.

More formally, the probabilistic generative model defined by this HSA is, for \(i \leq i \leq n_{j}\),
\[
w_{j i} \mid x_{j i}, z_{j i}, \phi \sim \text { Categorical }\left(w_{j i} \mid \phi_{\left[z_{j i}^{1}, x_{j i}\right]}\right),
\]
and for \(i=1\),
\[
x_{j i} \mid \pi \sim \operatorname{Categorical}\left(x_{j i} \mid \pi\right), \quad z_{j i}=\emptyset,
\]
and for \(1<i \leq n_{j}\),
\[
\begin{aligned}
x_{j i} \mid x_{j i-1}, z_{j i-1}, \theta & \sim \operatorname{Categorical}\left(x_{j i} \mid \theta_{\left[z_{j i-1}^{1}, x_{j i-1}\right]}\right), \\
z_{j i} \mid u_{j i-1}, z_{j i-1} & =O_{\left[u_{j i-1}\right]}\left(z_{j i-1}\right), \\
u_{j i-1} \mid x_{j i-1}, z_{j i-1}, \Omega & \sim \operatorname{Categorical}\left(u_{j i-1} \mid \Omega_{\left[z_{j i-1}^{1}, x_{j i-1}\right]}\right) .
\end{aligned}
\]

Here, we use the auxilary variable \(u_{j i}\) to refer to the operation applied to \(z_{j i}\), and \(O\) is the set of \((H+1)\) functions that map \(z_{j i}\) to \(z_{j i+1}\) when these operations are applied. In other words, this makes clear that the value of \(z_{j i+1}\) is a deterministic function of \(z_{j i}\) when the value of \(u_{j i}\) is known, but this value is stochastically conditional on \(x_{j i}\) and \(z_{j i}\). In terms of the original variables, the graphical model for the HSA is as follows:


Figure 2: The graphical model or dynamical Bayesian network for the Hidden Stochastic Automaton. As with Figure 1, shaded nodes indicate observed variables and we have omitted the priors on \(\phi, \pi, \theta\) and \(\Omega\).

\section*{Generative Power of Hidden Stochastic Automata}

The generative power of the HSA model (as shown in Figure 2) relative to that of the standard HMM (as shown in Figure 1) arises from the fact that the statespace of the state-vector \(z_{j i}\), namely \(\{1 \ldots H\}^{*}\), is a countably infinite set yet the conditional relationships to and from \(z_{j i}\) are finitely specifiable. The consequences of this can be better appreciated by reference to discrete automata of the kind that form the foundations of theoretical computer science (see, e.g., Hopcroft et al., 2001).

As we have described it, the state-vector \(z_{j i}\) is identical to a pushdown stack with a symbol set \(\{1 \ldots H\}\). Prepending an element to the state-vector is equivalent to a push operation, while removing the first element is a pop operation. Assuming known values for \(\Omega\), which operation is applied to \(z_{j i}\) is dependent only on the value of the finite state variable \(x_{j i-1}\) and the first element or head of \(z_{j i-1}\). Likewise, assuming known values for \(\theta\), the value taken by \(x_{j i}\) is also dependent only on \(x_{j i-1}\) and the head of \(z_{j i-1}\). In other words, the HSA model described above is equivalent to a stochastic generative version of a pushdown stack automaton.
If we allow a greater variety of operations on the statevector than just prepending or removing symbols from the left, the computational power of the HSA can be beyond that of a generative pushdown stack automaton. For example, if
\[
\sigma_{1} \sigma_{2} \dot{\sigma}_{3} \sigma_{4} \sigma_{5} \sigma_{6}
\]
is the value of the state-vector, we may treat an arbitrary element - in this cases \(\sigma_{3}\) - as its head. If we allow for the appending of new elements to the left or the right of the head, or for the deleting of the element at the head, followed by the moving of the head pointer to the left or right, then this state-vector is equivalent to a two-way memory tape. As before, assuming known values for \(\Omega\), which of the operations is applied to the state-vector \(z_{j i}\) is again dependent only on the value of the finite state variable \(x_{j i-1}\) and the head of \(z_{j i-1}\). Likewise, as before, assuming known values for \(\theta\), the value taken by \(x_{j i}\) is also dependent only on \(x_{j i-1}\) and the head element of \(z_{j i-1}\). As such, with these changes the HSA is now equivalent to a stochastic generative version of the Turing machine.

\section*{Inference}

As is clear from Figure 2, only the variables \(\mathbf{w}=\left\{w_{j 1} \ldots\right.\) \(\left.w_{j i} \ldots w_{j n_{j}}\right\}_{j=1}^{J}\) are observed. In general, therefore, the problem of inference in the HSA is the problem of inferring the joint posterior
\[
\mathrm{P}(\theta, \phi, \Omega, \pi, \mathbf{x}, \mathbf{z} \mid \mathbf{w}, \alpha, \beta, \gamma, \nu),
\]
where \(\mathbf{x}\) and \(\mathbf{z}\) are the set of finite state and state-vectors variables, and \(\alpha, \beta, \gamma, \nu\) are the Dirichlet priors for \(\theta, \phi\), \(\Omega, \pi\), respectively.

The procedure for inference that we will follow is to use a collapsed Gibbs sampler to draw samples from the posterior
\[
\mathrm{P}(\mathbf{x}, \mathbf{z} \mid \mathbf{w}, \alpha, \beta, \gamma, \nu)
\]
that integrates over the values of \(\theta, \phi, \Omega, \pi\). This Gibbs sampler is identical in nature to the collapsed sampler used in Andrews and Vigliocco (2010) for the case of a hierarchical mixture of Hidden Markov models.

For all \(j \in\{1 \ldots J\}\) and \(i \in\left\{1 \ldots n_{j}\right\}\), the Gibbs sampler iteratively draws samples from the posterior over \(x_{j i}\) and over \(z_{j i}\), conditioned on sampled values for all remaining variables.

The posterior distribution over \(x_{j i}\), conditioned on known values for all the other variables is \({ }^{1}\)
\[
\begin{aligned}
& \mathrm{P}\left(x_{j i} \mid w_{j i}, z_{j i}, x_{\neg j i}, w_{\neg j i}, z_{\neg j i}, \alpha, \beta, \gamma\right) \propto \\
& \int \mathrm{P}\left(w_{j i} \mid x_{j i}, z_{j i}, \phi\right) \mathrm{P}\left(\phi \mid w_{\neg j i}, x_{\neg j i}, z_{\neg j i}, \beta\right) d \phi \times \\
& \int \mathrm{P}\left(z_{j i+1} \mid x_{j i}, z_{j i}, \Omega\right) \mathrm{P}\left(\Omega \mid x_{\neg j i}, z_{\neg j i}, \gamma\right) d \Omega \times \\
& \int \mathrm{P}\left(x_{j i+1} \mid x_{j i}, z_{j i}, \theta\right) \mathrm{P}\left(x_{j i} \mid x_{j i-1}, z_{j i-1}, \theta\right) \mathrm{P}\left(\theta \mid x_{\neg j i}, z_{\neg j i}, \alpha\right) d \theta .
\end{aligned}
\]

This leads to the following closed form:
\[
\begin{aligned}
& \mathrm{P}\left(x_{j i}=k \mid w_{j i}, z_{j i}, x_{\neg j i}, w_{\neg j i}, z_{\neg j i}, \alpha, \beta, \gamma\right) \\
& \quad \propto \frac{S_{h k v}^{\neg j i}+\beta_{v}}{S_{h k}^{\neg j i}+b} \times \frac{Q_{h k q}^{\neg j i}+\gamma_{q}}{Q_{h k}^{\neg j i}+c} \\
& \\
& \quad \times \frac{\left(R_{h k k_{+}}^{\neg j i}+\delta_{k_{-}, k, k_{+}}+\alpha_{k_{+}}\right)\left(R_{h_{-} k_{-k}}^{\neg j i}+\alpha_{k}\right)}{R_{h k \cdot}^{\neg j i}+\delta_{k_{-}, k}+a} .
\end{aligned}
\]

Here, we are assuming that the value of the observed variable at \(j i\) is \(v\), the value of head of the state-vector at \(j i\) is \(h\), its value at \(j i-1\) is \(h_{-}\), the value of the finite state variable at \(j i-1\) is \(k_{-}\)and its value at \(j i+1\) is \(k_{+}\). The \(S, Q\) and \(R\) are rank- 3 arrays of frequencies, with the superscript of \(\neg j i\) indicating that they are based on excluding variables at \(j i\). As such, \(S_{h k v}^{\neg j i}\) is the number of times the observed variable has a value of \(v\) when the finite state variables has the value \(k\) and the head (e.g., the first) element of state-vector takes the value of \(h \in\{0 \ldots H\}, Q_{h k q}^{\neg j i}\) is the number of times that statevector operation \(q\) occurs whenever the head element of the state-vector takes the value of \(k\) and the finite state variable takes the value of \(k\), and \(R_{h k k_{+}}^{\neg j i}\) gives the number of times the finite state variable takes the value \(k_{+}\)whenever its value at the previous index is \(k\) and the value of the head of the state-vector at the previous index is \(h\). The dot in place of the third index, e.g., \(S_{h k}^{\neg j i}\), indicates a sum over the index. The term \(\delta_{k_{-}, k, k_{+}}\)

\footnotetext{
\({ }^{1}\) We will provide the conditional distributions for values of \(x_{j i}\) and \(z_{j i}\) where \(1<i<n_{j}\). The distributions for the cases where \(i=1\) and \(i=n_{j}\) require minor modifications, which we will omit here in the interests in space.
}
takes the value of 1 is \(k_{-}=k=k_{+}\)and takes the value of zero otherwise. Likewise, \(\delta_{k_{-}, k}\) takes the value of 1 when \(k_{-}=k\), and takes the value of 0 otherwise. The terms \(a, b\) and \(c\) are the sums of \(\alpha, \beta, \gamma\), respectively.

For the case of the posterior distribution of the statevector, it is sufficient to infer the distribution over operations applied to it. As mentioned, the value of the state-vector \(z_{j i}\) is deterministic function of \(z_{j i-1}\) when the operation \(u_{j i-1}\) is known. The posterior distribution over \(u_{j i}\) is given by
\[
\begin{aligned}
& \mathrm{P}\left(u_{j i} \mid w_{j i}, z_{j i}, x_{\neg j i}, w_{\neg j i}, z_{\neg j i}, \alpha, \beta, \gamma\right) \propto \\
& \times\left[\int \mathrm{P}\left(w_{j i+1} \ldots w_{j n_{j}} \mid x_{j i+1} \ldots x_{j n_{j}}, z_{j i+1} \ldots z_{j n_{j}}, \phi\right)\right. \\
& \left.\quad \mathrm{P}\left(\phi \mid w_{\neg \overrightarrow{j i}}, x_{\neg \overrightarrow{j i}}, z_{\neg \overrightarrow{j i}}, \beta\right) d \phi\right] \\
& \times\left[\int \mathrm{P}\left(x_{j i+1} \ldots x_{j n_{j}} \mid x_{j i} \ldots x_{j n_{j}-1}, z_{j i} \ldots z_{j n_{j}-1}, \theta\right)\right. \\
& \left.\quad \mathrm{P}\left(\theta \mid x_{\neg \vec{i} i}, z_{\neg \overrightarrow{j i}}, \beta\right) d \theta\right] \\
& \times \mathrm{P}\left(z_{j i+1} \ldots z_{j n_{j}} \mid u_{j i}, z_{j i}\right) \\
& \times \int \mathrm{P}\left(u_{j i} \mid x_{j i}, z_{j i}, \Omega\right) \mathrm{P}\left(\Omega \mid x_{\neg \overrightarrow{j i}}, z_{\neg \overrightarrow{j i}}, \gamma\right) d \Omega
\end{aligned}
\]
where we see that because a change to the operation \(u_{j i}\) deterministically changes the values of \(z_{j i+1} \ldots z_{j n_{j}}\), the likelihood terms for the \(u_{j i}\) variable include the variables \(w_{j i+1} \ldots w_{j i+1}\) and \(x_{j i+1} \ldots x_{j i+1}{ }^{2}\). In the above, the notation \(\neg \overrightarrow{j i}\), e.g., in \(x_{\neg \overrightarrow{j i}}\), indicates the exclusion of variables \(j i \ldots j n_{j}\). This distribution leads to the closed form
\[
\begin{aligned}
& \mathrm{P}\left(u_{j i}=q \mid w_{j i}, z_{j i}, x_{\neg j i}, w_{\neg j i}, z_{\neg j i}, \alpha, \beta, \gamma\right) \propto \\
& \frac{\prod_{\left\{h k v: S_{h k v}^{q}>0\right\}} \prod_{s=0}^{S_{h k v}^{q}-1} S_{h k v}^{\neg \overrightarrow{j i}}+\beta_{v}+s}{\prod_{\left\{h k: S_{h k}^{q}>0\right\}} \prod_{s=0}^{S_{h k}^{q}-1} S_{h k}^{\neg \overrightarrow{j i}}+b+s} \times \\
& \frac{\prod_{\left\{h k q: Q_{h k q}^{q}>0\right\}} \prod_{s=0}^{Q_{h k q}^{q}-1} Q_{h k q}^{\neg \overrightarrow{j i}}+\gamma_{q}+s}{\prod_{\left\{h k: Q_{h k}^{q}>0\right\}} \prod_{s=0}^{Q_{h k .}^{q} .-1} Q_{h k .}^{\neg \overrightarrow{j i}}+c+s} \times \\
& \frac{\prod_{\left\{h k l: R_{h k l}^{q}>0\right\}} \prod_{s=0}^{R_{h k l}^{q}-1} R_{h k l}^{\neg \overrightarrow{j i}}+\alpha_{l}+s}{\prod_{\left\{h k: R_{h k}^{q}>0\right\}} \prod_{s=0}^{R_{h k}^{q}-1} R_{h k .}^{\neg \overrightarrow{j i}}+a+s} .
\end{aligned}
\]

Here, \(S_{h k v}^{\neg \overrightarrow{j i}}, Q_{h k q}^{\neg \overrightarrow{j i}}\) and \(R_{h k l}^{\neg \overrightarrow{j i}}\) have the same meaning as \(S_{h k v}^{\neg \overrightarrow{j i}}, Q_{h k q}^{\neg \overrightarrow{j i}}\) and \(R_{h k l}^{\neg \overrightarrow{j i}}\) with the difference being that the frequencies are calculated excluding variables at the indices \(i j \ldots j n_{j}\). By contrast, the arrays \(S_{h k v}^{q}, Q_{h k q}^{q}\) and \(R_{h k l}^{q}\) are the frequencies of the co-occurrences the values

\footnotetext{
\({ }^{2}\) In graphical model terms, the variables \(w_{j i+1} \ldots w_{j i+1}\), \(x_{j i+1} \ldots x_{j i+1}\) and \(z_{j i+1} \ldots z_{j n_{j}}\) are all children of \(u_{j i}\).
}


Figure 3: Strings generated by the probabilistic context-free grammar \(S \rightarrow 0 S 1(p=0.66), S \rightarrow 01(p=0.34)\) were used as observed data in a HSA. Shown above are samples of the binary strings generated by the HSA model on the basis of estimates of the parameters \(\phi, \theta, \Omega\) and \(\pi\) after \(3,5,10,20,50\) and 100 iterations of the Gibbs sampler. The dark shade codes the value of 1 . It is evident that by over 50 iterations, the HSA has inferred the correct generative model of the probabilistic language. By 100 iterations, it is only generating strings from the language \(L=\left\{0^{n} 1^{n}: n \geq 0\right\}\).
of the variables after operation \(q\) is applied to the statevector \(z_{j i+1}\) and the changes to the subsequent statevectors are deterministically applied.

\section*{Demonstration}

We demonstrate inference of a language from data by using the textbook example of a simple context-free language, namely \(L=\left\{0^{n} 1^{n}: n \geq 0\right\}\). We can generate strings from a probabilistic version of this language using the probabilistic context-free grammar
\[
\begin{array}{rlrl}
S & \rightarrow 0 S 1, & & p=0.66 \\
& \rightarrow 01, & p=0.34
\end{array}
\]

We sample \(J=25\) strings from this language and use them as the data \(\mathbf{w}=\mathbf{w}_{1}, \mathbf{w}_{2} \ldots \mathbf{w}_{j} \ldots \mathbf{w}_{J}\) for a HSA model of the kind described.

Using the collapsed Gibbs sampler, we can sample from the posterior over the finite state and state-vector trajectories conditional on w. From these, we may then draw sample estimates of \(\phi, \theta, \Omega\) and \(\pi\). Shown in Figure 3 are strings generated by the HSA model with parameters \(\phi, \theta, \Omega\) and \(\pi\) as estimated after, from left to right, \(3,5,10,20,50\) and 100 iterations of the Gibbs sampler.

\section*{Relevance for Cognitive Science}

Our initial motivation for the HSA model was put in terms of the computational advantages of graphical models as formalisms for probabilistic modeling. Graphical
models, we have argued, have effectively become a graphbased modeling language for developing and extending probabilistic models. They have had a remarkable influence on the progress of probabilistic modeling in a wide variety of fields, including cognitive science. It is notable, therefore, that graphical models have had a relatively limited role in the probabilistic modeling of natural language. The obvious reason for this is due to the structurally complex nature of natural languages. While this structure is modeled well by probabilistic grammars, grammars can not, in general, be represented by graphical models. By contrast, the graphical models most widely used for modeling sequential data do not have the structural complexity necessary for modeling natural language.

We have introduced the HSA as a dynamical Bayesian network model that is capable of modeling structurally complex sequences. Its principal relevance to cognitive science is therefore as a computational model of cognition, where by computational model we specifically mean the Marr (1982) sense of the term: a model of the abstract nature of problem being faced and of its rational solution. However, the HSA model is potentially as relevant as a model of the resource limited practice, or possibly even the physical implementation, of cognition. For example, the HSA is an incremental statespace model, where inference is naturally modeled by the kind of sequential Monte Carlo methods, particularly particle filters, that have been advocated by, for
example, Griffiths, Vul, and Sanborn (2012); Sanborn, Griffiths, and Navarro (2010); Levy, Reali, and Griffiths (2009) as models of memory and time constrained approximations to rational computational models. On the other hand, from the point of view of physical implementation, the state-vector of the HSA can be represented naturally by a real-valued variable. If the state-vector is \(\sigma_{1} \sigma_{2} \ldots \sigma_{i} \ldots \sigma_{n}\), this can be represented exactly by the real number \(\sum_{i=1}^{n} \sigma_{i}(H+1)^{-i}\) and the operations applied to the state vector correspond to real-valued functions. For example, if the state-vector is binary, prepending a \(\sigma \in\{0,1\}\) to \(\sigma_{1} \sigma_{2} \ldots \sigma_{i} \ldots \sigma_{n}\) is identical to multiplying \(\sum_{i=1}^{n} \sigma_{i} 2^{-i}\) by \(\frac{1}{2}\) and adding \(\frac{\sigma}{2}\). By treating the finite state variable as another real number, this allows us to represent the HSA exactly as a stochastic nonlinear dynamical system that is directly comparable to a recurrent neural network (see, e.g., Tabor, 2000, for related discussion).

\section*{References}

Andrews, M., \& Vigliocco, G. (2010). The Hidden Markov Topics Model: A Probabilistic Model of Semantic Representation. Topics in Cognitive Science, 2, 101-113.
Bishop, C. M. (2006). Pattern Recognition and Machine Learning. New York, NY: Springer.
Bishop, C. M. (2013, Feb 13). Model-Based Machine Learning. Philosophical Transactions of the Royal Society A - Mathematical Physical and Engineering Sciences, 371 (1984).
Brand, M., Oliver, N., \& Pentland, A. (1997). Coupled Hidden Markov Models for Complex Action Recognition. In 1997 IEEE Computer Society Conference On Computer Vision And Pattern Recognition, Proceedings (p. 994-999).
Chater, N., \& Oaksford, M. (2008). The Probabilistic Mind. Oxford, UK: Oxford University Press.
Chomsky, N. (1956). Three models for the description of language. Institute of Radio Engineers Transactions on Information Theory, 2, 113-124.
Chomsky, N. (1963). Formal properties of grammars. In R. D. Luce, R. R. Bush, \& E. Galanter (Eds.), Handbook of mathematical psychology (Vol. 2, p. 323-418). New York and London: John Wiley and Sons, Inc.
Fine, S., Singer, Y., \& Tishby, N. (1998). The Hierarchical Hidden Markov Model: Analysis and Applications. Machine Learning, 32(1), 41-62.
Friedman, N. (2004, Feb 6). Inferring Cellular Networks using Probabilistic Graphical Models. Science, 303(5659), 799-805.
Ghahramani, Z., \& Jordan, M. (1997). Factorial Hidden Markov Models. Machine Learning, 29, 245-273.
Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., \& Tenenbaum, J. B. (2010, Aug). Probabilistic Mod-
els of Cognition: Exploring Representations and Inductive Biases. Trends in Cognitive Sciences, 14(8), 357-364.
Griffiths, T. L., Vul, E., \& Sanborn, A. N. (2012). Bridging Levels of Analysis for Probabilistic Models of Cognition. Current Directions in Psychological Science, 21(4), 26-268.
Hopcroft, J., Motwani, R., \& Ullman, J. (2001). Introduction to automata theory, languages and computation (2nd ed.). Addison Wesley.
Jordan, M. I. (2004). Graphical models. Statistical Science, 19(1), 140-155.
Koller, D., \& Friedman, N. (2009). Probabilistic graphical models: Principles and techniques. Cambridge, MA: MIT Press.
Lauritzen, S., \& Spiegelhalter, D. (1988). Local Computations With Probabilities On Graphical Structures And Their Application To Expert Systems. Journal Of The Royal Statistical Society Series BMethodological, 50(2), 157-224.
Levy, R., Reali, F., \& Griffiths, T. L. (2009). Modeling the Effects of Memory on Human Online Sentence Processing with Particle Filters. In Advances in Neural Information Processing Systems (Vol. 21, p. 937-944).

Marr, D. (1982). Vision. New York, NY: W. H. Freeman \& Company.
Oliver, N., Rosario, B., \& Pentland, A. (2000, Aug). A Bayesian Computer Vision System for Modeling Human Interactions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8), 831-843.
Pearl, J. (1988). Probabilistic reasoning in intelligent systems: Networks of plausible inference. Morgan Kaufmann.
Salakhutdinov, R., \& Hinton, G. (2009, Jul). Semantic Hashing. International Journal of Approximate Reasoning, 50(7), 969-978.
Sanborn, A. N., Griffiths, T. L., \& Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. Psychological Review, 117, 1144-1167.
Saul, L., \& Jordan, M. (1999, Oct). Mixed Memory Markov Models: Decomposing Complex Stochastic Processes as Mixtures of Simpler Ones. Machine Learning, 37(1), 75-87.
Shieber, S. M. (1985). Evidence Against the ContextFreeness of Natural-Language. Linguistics And Philosophy, 8(3), 333-343.
Tabor, W. (2000). Fractal encoding of context-free grammars in connectionist networks. Expert Systems, 17(1), 41-56.

\title{
An evolutionary account of reactions to a wrong
}

\author{
Giulia Andrighetto (giulia.andrighetto@istc.cnr.it) \\ Institute of Cognitive Science and Technology (ISTC-CNR), via San Martino della Battaglia 44, 00185, Rome, Italy European University Institute (EUI), Fiesole, Italy \\ Francesca Giardini (francesca.giardini@istc.cnr.it) \\ Institute of Cognitive Science and Technology (ISTC-CNR), via San Martino della Battaglia 44, 00185, Rome, Italy \\ Rosaria Conte (rosaria.conte@istc.cnr.it) \\ Institute of Cognitive Science and Technology (ISTC-CNR), via San Martino della Battaglia 44, 00185, Rome, Italy
}

\begin{abstract}
In this work, we propose an evolutionary account of reactions to a wrong as an integrated set. Unlike other theories, we are not interested in revenge, punishment or sanction per se, but in their co-existence. We posit that this variety of reactions is needed in order to achieve different goals, but it also implies an increase in cognitive costs that requires to be explained from an evolutionary perspective. Moving from the identification of the psychological traits that uniquely define each reaction, two concurrent hypotheses are suggested and discussed: either the richness of human social life requests a variety of reactions, or the benefits of single reactions at the psychological level allowed these reactions to be maintained in the social life.
\end{abstract}

Keywords: Evolution; punishment; revenge; sanction; cognitive influencing; norms; enforcing institutions social order.

\section*{Introduction}

Human actions are potentially unbounded and much more opportunities are available when other people are involved. When talking about social actions we have to distinguish between actions and reactions, i.e., actions triggered by someone's else previous action. Reactions are a constitutive part of living in societies, and the ability of displaying the appropriate reaction in the right content is extremely important for our "ultra-social" species (Richerson \& Boyd 1998, 2005; Hill et al. 2009). The nature and the intensity of reactions depend on both the actor and the triggering action, and it requires the capacity to forecast further reactions and to plan ahead, among other things. Humans are unique under this respect, and everyone has experienced how many reactions the same individual can display in response to the same action, even in the same context. Animals can modulate their reactions, in some cases they can also decide their behaviour on a cost-benefit analysis, but others' representations do not enter this picture (Clutton-Brock \& Parker 1995; Jensen et al. 2007). Humans react because of what they believe and want, and because of what they want others to believe and of how they want them to behave.

A particularly interesting class of social reactions is that triggered in response to a wrong. Retaliation, revenge,
punishment and sanctions have been a matter of interest since the rise of Western culture, as witnessed by the fact that the need to understand and explain motives for reacting to wrongs never ceased since pre-Homeric Greece to these days (for an analysis of the differences among these reactions see Giardini et al. 2010).

Philosophers, social scientists, political scientists, psychologists, anthropologists have been striving to answer the fundamental question: why do people react to a wrong? In many circumstances reacting is more costly than standing, it requires some kind of planning, and it also implies the possibility of suffering a counter-reaction. Even more striking, people react to wrongs suffered by strangers, intervene in others' disputes, and sanction others when failing to comply with norms that they are not supposed to enforce. Although several scholars have been interested in explaining the evolution of revenge, punishment and sanction (Lorenz, 1966; Hamilton, 1970; Boyd \& Richerson, 1992; Clutton-Brock \& Parker 1995; Gardner \& West, 2004; Jensen, 2010), these phenomena have been usually considered in isolation and not as a rich and complex repertoire. We propose that revenge, punishment and sanction are different reactions that should not be considered in isolation but as interdependent and complementary. If we look at them as an intertwined set, we need to explain the reason why they are different, but we also need to understand why we still have more than one reaction to an offense, and how the related extra cognitive costs are compensated. Our goal is to explain the decision to apply punishment in terms of the complementary decisions to use neither revenge nor sanction, thus understanding the motives behind each and every reaction.

We propose that this variety is necessary because, unlike animals, humans' reactions do not only target the offender's behavior, but also her mental states, as well as the victim's mental states. Comparing different reactions, we highlight an evolutionary trajectory that links revenge, punishment and sanction by explaining costs and benefits of each reaction. Having the opportunity to choose among several responses means higher cognitive costs to select between actions, and to choose the most appropriate one. Therefore, a set of questions arises: Why do we have such a repertoire? Can we identify evolved mechanisms that allowed us to
distinguish among reactions and to selectively apply them depending on the context?

Moving from the identification of the psychological traits that uniquely define each reaction, we propose a complex relationship between the richness of human social life, which requests a variety of reactions (society \(\rightarrow\) individual motivations), and the benefits of single reactions at the psychological level. These benefits favoured the maintenance of reactions in the social life (individual motivations \(\rightarrow\) society).

The rest of the paper is organized as follows: Section 1 outlines the evolutionary model, Section 2 defines the different phenomena and Section 3 introduces the main features of our taxonomy.

\section*{An evolutionary account of reactions}

Revenge, punishment and sanction are superficially similar but deeply different in terms of the evaluation of the wrong suffered (or its interpretation), the intended goal, the consequent cognitive influencing, the temporal dimension and the kind of target. Humans are usually effective in administering punishment, i.e., in selecting the best reaction, taking into account the differences and selecting, through a fast and efficient process, how to react according to the external circumstances and their internal states.
The computational demands associated with the choice are not negligible and the risks of a mismatch between the perceived wrong and the reaction are high. We propose that revenge, punishment and sanction require the evolution of specialized mental mechanisms regulating the activation of different responses to wrongs or rule violations. We suggest that humans have mechanisms designed to produce revenge, punishment and sanction that evolved because of their effectiveness in solving recurrent social problems that humans encountered during evolution (Petersen et al. 2012). Given the richness of human bonds and social life, the need for acquiring social bonds (Dunbar, 1996) and for maintaining them, also remembering who is related to whom, could have favoured the selection of different reactions that have different consequences in terms of relationships (McCullough, Kurzban, Tabak, 2012). Moreover, psychological benefits of reactions may motivate their maintenance at the individual level and thus foster their selection at the social one. On the one hand, restoring the status quo, achieving deterrence or promoting the norms are goals that cannot be achieved through a single reaction, and their related specific benefits at the psychological level may have prompted the maintenance of multiple responses. On the other hand, the costs of selecting among different reactions are not negligible, also because having more choices implies being more prone to errors, with negative consequences arising at both the individual level and the social level. In the latter case, this mismatch between the reaction chosen and the wrong suffered can be extremely dangerous, and it may challenge the social order. Avenging a wrong when there is a social norm and the related sanction, or punishing someone in a context in which
revenge was expected could result in a negative judgment about the reacting agent. Failing in interpreting correctly the situation and thus applying an inappropriate reaction may lead individuals to consider the avenger/punisher/sanctioner as socially inadequate and to avoid interactions with her. At the group level, frequent failures in using the appropriate reaction may undermine the cohesion of the group and make it more vulnerable to turmoil and fights.
In evolutionary terms, the risk associated with the application of the wrong kind of reaction were compensated by the evolution of specific psychological mechanisms for selecting among reactions.
Each and every reaction involves some unintended side effects, which may prevent the agent from achieving her goals and may also make the reaction inappropriate. In revenge, making the other suffer and regaining one's sense of control, together with restoring the status quo, require the agent to evaluate the wrong suffered and to estimate how much sufferance to inflict on the offender. Since there is not any objective criterion to estimate the sufferance experienced, this evaluation can only be subjective, thus exposing the avenger to the risks of damaging his reputation because the reaction was disproportionate (too harsh or too weak), or loosing social ties, or even triggering a feud with escalation of violence. Feuds are especially costly at the group level and they may even lead to the dissolution of the group. The punisher aims at deterring the wrongdoer from further hostility (by making it a costly option). There is not a pre-established and socially shared set of rules that govern how to punish. This lack of explicit and objective regulation can have several negative consequences. If the punishment inflicted is not appropriate in quantity or in kind, this can result in perception of the punishment received as unjustified, not legitimate and unfair. When punishment is perceived inappropriate it may also become ineffective in inducing deterrence, so the punisher is not able to achieve her main goal. In addition, the punisher can acquire a bad reputation for being too harsh, and she can see some social ties severed because of his action with the consequent risk of an escalation of violence, which has consequences for the whole group.
The risks of administering an inappropriate sanction are more limited and they are mainly related to the fact that the normative message is not clearly understood by those who receive the sanction. Therefore, when the normative character of the situation is not recognized, the sanction is ineffective, and the normative belief and the normative goal will not be formed in the mind of those who receive the sanction. An inappropriate sanction may also lead to counter-reactions, either in the form of a further sanction or as a retaliatory behavior.
In what follows we will detail our model of reactions, specifying the cognitive underpinnings and the dimensions of change characteristics of each and every phenomenon, and then supporting our model with a discussion of the relevant literature.

\section*{Distinguishing among reactions}

Although a number of accounts (for some representative work see Bowles \& Gintis 2004; Henrich \& Boyd, 2001; Henrich et al. 2006) have stressed the relevance of punishment in human societies, they suffer the flaw of considering punishment as a single behaviour. In our view, punishing actually consists in a complex behavioral repertoire in which it is useful to disentangle at least revenge, punishment, and sanction. In Giardini, Andrighetto, Conte (2010) it has been argued that this variety of punishing strategies can be differentiated on the basis of 1 . their mental antecedents, 2. the way in which they influence the future conduct of others, and 3. the effects they aim to achieve. Having more than one available strategy allows humans to tailor their reactions and to achieve their goals more easily but, at the same time, this implies higher cognitive and computational costs. In fact, agents must be able to categorize actions in the correct way, meaning that the context has to be interpreted adequately, the most appropriate reaction has to be chosen on the basis of the perceived wrong, of the situation, and of the offender and other agents' mental states. This calculation leads to a significant increase in the computational costs, which should be compensated, by some sort of benefits. Revenge, punishment and sanction result from psychological adaptations that allowed to solve recurrent conflicts that humans encountered during their evolutionary history, but we still do not know why we have more than one mechanism.
If animal societies are able to cope with aggressions by using just one form of reaction (Clutton-Brock \& Parker, 1995), usually termed "punishment", why do we need a collection of counter-reactions? What are the fitness benefits coming from revenge, punishment and sanction?

In our theoretical analysis of reactions to a wrong, we start by providing a preliminary list of the core elements that determine the kind of response that an individual will choose in response to an aggression (see Table 1):
- The wrong suffered, i.e., the cause of the response. The evaluation of the offense depends both on the intentions the aggressor (the offense was intentional vs the offense was not intentional), and on the nature and value of the goal(s) frustrated by the aggressor.
- The goal of the reaction. When deciding how to react to an aggression, individuals consider the goal(s) they want to achieve and then select the appropriate reaction.
- The kind of influencing the agent reacting wants to apply to achieve her goal(s). Our theory is based on the idea that different reactions are aimed to produce different changes in the mind-set of the victim. For example, the avenger is aimed at acting at the epistemic level, by changing the target's and audience's beliefs about herself. The punisher aims to act both at the epistemic and motivational levels, by generating in the victim's mind the goal -
usually under threat of punishment- of abstaining from doing the action that has triggered punishment again. Finally, the sanctioner wants to endow the offender with new normative knowledge and to generate in her mind the goal to comply with the norm in the future.
- The focus of the reaction refers to the agent herself (as it is in revenge), another agent (as it is in punishment), or a norm (as it is in sanction).

It is worth noticing that we do not consider reactions as clear-cut phenomena, but they are overlapping in several respects. In Table 1, we summarize the main features of each reaction, in an attempt to identify the key elements of each phenomenon.

Table 1: Dimensions of change
\begin{tabular}{|l|l|l|l|}
\hline & Revenge & Punishment & Sanction \\
\hline \begin{tabular}{l} 
Wrong \\
suffered
\end{tabular} & \begin{tabular}{l} 
Intentional \\
aggression; \\
Frustration of \\
a personal \\
goal of the \\
agent; \\
Sufferance \\
experienced
\end{tabular} & \begin{tabular}{l} 
Intentional \\
aggression; \\
Frustration of \\
a personal or \\
social goal of \\
the agent.
\end{tabular} & \begin{tabular}{l} 
Norm \\
violation
\end{tabular} \\
\hline Goal & \begin{tabular}{l} 
Making the \\
aggressor \\
suffer; \\
Status quo \\
restoration
\end{tabular} & Deterrence & \begin{tabular}{l} 
Norm \\
recognition; \\
Norm \\
compliance
\end{tabular} \\
\hline \begin{tabular}{l} 
Cognitive \\
influencing
\end{tabular} & \begin{tabular}{l} 
Beliefs
\end{tabular} & \begin{tabular}{l} 
Beliefs \\
Goals
\end{tabular} & \begin{tabular}{l} 
Normative \\
Beliefs \\
Normative \\
Goals
\end{tabular} \\
\hline \begin{tabular}{l} 
Temporal \\
dimension
\end{tabular} & \begin{tabular}{l} 
Backward- \\
looking
\end{tabular} & \begin{tabular}{l} 
Forward- \\
looking
\end{tabular} & \begin{tabular}{l} 
Forward- \\
looking
\end{tabular} \\
\hline Focus & Self & Other & Norm \\
\hline
\end{tabular}

Figure 1 depicts the mental path that triggers to choose a specific reaction. In the following section, a cognitive anatomy of revenge, punishment and sanction will be provided. In section 3, an analysis of the intended and unintended effects of the three reactions is presented. The latter analysis will allow us to sketch an evolutionary explanation of why we have more than one reaction to an offense, and how the cognitive extra-costs resulting from this variety of reactions are compensated.


Figure1: The cognitive path of reactions to a wrong
A further consideration involves the role of emotions. The specific role that anger, but also social emotions, such as moral outrage, pride, shame, guilt, indignation, contempt, disgust, resentment, etc., (e.g. Fessler \& Haley 2003; Frank 1988) play in triggering the reactions under study deserves an attentive theoretical and experimental analysis. Although crucial, this analysis is beyond the scope of the present paper and will be developed in future work.

\section*{Taxonomy of reactions to a wrong: Revenge, Punishment and Sanction}

\section*{Revenge}

Revenge, according to the Merriam-Webster dictionary, is "punishment inflicted in retaliation for an injury or offence". In Elster's terms (1990) it is "the attempt at some cost or risk to oneself, to impose suffering upon those who made one suffer, because they have made one suffer" (p. 862). Broadly speaking, the term 'revenge' refers to two diverse but connected phenomena.
In the first of these phenomena, revenge is a social ritual that requires and prescribes specific behaviours to group members to repair an offence. Ethnographic studies highlighted the transition from tribal to modern societies, in which retributive concepts of law and the creation of institutions replaced vengeance and avoided blood feuds (Boehm 1986). Posner (1980) suggests that revenge and retribution may be partially determined by historical and economic circumstances, such the private enforcement of law and high probabilities of detecting and punishing offences. When these conditions are met, a pure vengeance system may appear, although it is unlikely to be optimal. These systems are not completely extinguished, as the culture of honour in the southern United States (Nisbett 1993; Nisbett and Cohen 1996) and the Kanun in Albania demonstrate. The Kanun, a customary set of laws used mostly in northern Albania and Kosovo, disciplined people's reactions to murder (blood revenge or gjakmarrje) and other offences (hakmarrje), according to the roles and degree of kinship of all the people involved. Shirking revenge or taking it without respecting what is stated in the Kanun leads to the same result: honour cannot be restored and the whole family or clan is to blame. Apparently, the Kanun has not disappeared completely, and in some areas it is still observed, showing how an institution that is preserved in the mind can out-compete another centrally enforced institution, because the latter one is not recognized as such.

The other way of looking at revenge is to consider it as an individual behaviour, which is present both in human societies (Zaibert 2006), and non-human primate groups (Jensen, Call and Tomasello 2007). Turning our attention to individual factors it becomes possible to provide a cognitive anatomy of this reaction. The avenger wants to repay the damage she suffered with an equal or greater offence, no matter how risky or dangerous this retaliation is. In a sense, we can say that the avenger is a backward-looker who revolves around the past and acts in the present to rebalance what happened, with no concern for the future. Unlike other authors (McCullough, Kurzban, Tabak, 2012), we do not see vengeance as a means to affect the likelihood that the wrongdoer will repeat the aggression in the future, inducing her to cooperate next time or deterring her from further aggressions. Long term, strategic planning does not seem to characterize the avenger's mind, although unintended deterrence effects can be obtained.
Revenge is motivated not only by the desire to make the target suffer, but also by the goal to change the target's and audience's beliefs about the avenger, in order to restore the image that has been damaged by the aggression suffered. In this case cognitive influence is aimed at changing the beliefs of the wrongdoer and of the audience: the avenger aims to repay the damage she suffered with an equal or greater offence in order to change the target's and audience's beliefs about himself. Revenge is a way to regain one's position after an offence and this applies also to the symbolic dimension: the avenger wants to restore her image, damaged by the aggression suffered. Revenge is aimed to modify what the others believe about the avenger, her role and status. Presumably, the greater the offence, the more efficacious the image restoration and the effort to restore the status-quo.

\section*{Punishment}

Enforcing institutions have evolved with society: starting out as simple systems of revenge and retribution imposed by the individual, family, or tribe, in modern societies they grew as institutions characterized by a higher concern for deterrence and rehabilitation. Institutions controlling modern societies moved from systems based on revenge to ones based on punishment. In primitive society enforcement was left to the individuals wronged, or their families, and was vindictive or retributive (Boehm 1986): in quantity and quality it would bear no special relation to the character or gravity of the offence. Gradually it arose the idea of proportionate punishment, of which the characteristic type is the lex talionis of early Roman law or in the Old Testament and Koran. Like revenge, also punishment refers to two distinct class of phenomena: punishment is both a social institution and an individual behavior.
As an institution, punishment serves to dissuade people from engaging in activities deemed wrong by law and by the society itself, thus reducing the frequency and likelihood of future offences. Deterrence theory suggests that punishment works by modifying the relative costs and benefits of
situation, so that wrongdoing becomes a less attractive option (Bentham 1962; Becker 1968). Punishment possibly has the effect of preventing blood feuds and giving more stability to the social order.
As an individual behaviour, punishment is a reaction intentionally aimed to minimize the chance that the aggressor will repeat the act again (Giardini, Andrighetto and Conte 2010). Unlike revenge, punishment is not inflicted in retribution for an offence or transgression. The punisher is driven by forward-looking considerations, and deterrence is intentionally pursued.
This enforcing mechanism, controlling modern societies, is not at all easy to distinguish from revenge (Zaibert 2006), at least from a mere behavioural point of view. Cognitive modelling allows us to disentangle them on the basis of their mental antecedents and the way in which they influence the future conduct of others. The punisher and the avenger are aimed at influencing and modifying the target and the audience's minds in different ways: unlike the avenger, the punisher has the explicit goal to deter the wrongdoer from repeating the aggression in the future. To achieve this goal, the punisher should act in such a way that the offender, and possibly the audience, generates in her mind the goal usually under threat of punishment (i.e., by generating the belief in the victim's mind that future aggressions will be punished)- of abstaining from doing the action that has triggered punishment again.

\section*{Sanction}

Social order can be explained as the mere result of the deterrence effect of punishment. However what makes human cooperation so spectacular with respect to all other species is the presence of social norms, efficiently orchestrating social life. When punishing institutions are able to work in tandem with social norms, they are much more viable and effective in achieving and maintaining compliance and are more robust across time (Andrighetto and Villatoro, 2011; Villatoro et al. 2011).
By analyzing a large number of spontaneously emerged institutions in different countries, the political scientist Elinor Ostrom has identified a set of characteristics that make them successful in promoting social order. She suggests that the most effective institutions are those that facilitate norms' elicitation, their spreading, and compliance (Ostrom 2005; see also Casari 2007). Punishment, when properly designed, should tell people which behaviours are acceptable, i.e., the (social) norms regulating society, and which actions will cause punishment.
We refer to punishing institutions enforcing social order through mechanisms intentionally aimed to focus people's attention on social norms and to condemn their violation as sanction institutions. We consider sanction institutions as the last step of the institutional evolutionary process.
As in previous work (Giardini et al. 2010; Andrighetto and Villatoro, 2011; Villatoro et al. 2011), we use sanction to indicate the enforcing individual behaviour that, in addition to imposing a cost for the wrongdoing, as punishment does,
is also intentionally aimed at signalling norms to the offender (and possibly to the audience) so that she will comply with them in the future.
The type of cognitive influencing sanction exerts on the offender is more complex than those in revenge and punishment. In order to deter future norms' violations, the sanctioner endows the offender with (new) normative knowledge. The sanctioner uses scolding to reign in wrongdoers, expresses indignation or blame, or simply mentions that the targeted behaviour violated a norm. Through these actions, the sanctioner aims to focus people's attention on different normative aspects, such as: (a) the existence and violation of a norm; (b) the causal link between violation and sanction: "you are being sanctioned because you violated that norm" (c) the probability that violations will be sanctioned; (d) the fact that the sanctioner is acting as a norm defender. As recent psychological and economic experimental evidence shows (Cialdini et al. 1990; Bicchieri 2006; Galbiati and Vertova 2008; Houser and Xiao 2010), the norm focusing effect of sanction plays an important role in eliciting norm compliance. Thus, despite punishment, we suggest that sanction has the further effect, possibly aimed at by the sanctioner, to encourage the target to ground future decisions on internal evaluative criteria, established by the norm. By facilitating the spreading, recognition and internalization of norms, sanction possibly has the effect of promoting social order in a more stable and less costly way with respect to punishment.

\section*{Concluding Remarks}

In this work, we proposed an evolutionary account of reactions to a wrong as an integrated set. Unlike other theories, we are not interested in revenge, punishment or sanction per se, but in their co-existence. We posited that this variety of reactions is needed in order to achieve different goals, but they also imply an increase in complexity, due to the costs associated with the interpretation of the situation and the selection among reactions. We proposed that the transition from one to the other has been allowed by specific cognitive patterns, and suggesting that these mental mechanisms selected among given social structures, at the same time reinforcing and being reinforced by them.
Modifying others' actions require a set of cognitive skills that allow to represent others' mental states, considering the way in which these are harbored in one's mind, giving rise to social beliefs, namely beliefs about others' mental states (e.g. beliefs, intentions, desires, emotions). In addition, this requires also cognitive influencing, as the willingness to modify others' goals. Having a set of available reactions means that individuals should also be endowed with cognitive mechanisms to recognize which reaction is more appropriate in a given situation. This theoretical analysis will be

\section*{Acknowledgments}

This work was supported by the Institute of Cognitive Science and Technologies, CNR; the European University Institute; the Complexity-Net Project SEMIRA; the ERC project "Willing to Pay"; the European Project GLODERS.

\section*{References}

Andrighetto, G., \& Villatoro, D. (2011). Beyond the carrot and stick approach to enforcement: An agent-based model. In European Perspectives on Cognitive Science, edited by B. Kokinov, A. Karmiloff Smith, N.J. Nersessian. Sofia: New Bulgarian University Press.
Bentham, J. (1962). Principles of penal law. In J. Bowring (Ed.). The Works of Jeremy Bentham. (pp. 365-580) New York: Russell and Russell.
Becker, G. S. (1968). Crime and Punishment: An Economic Approach. Journal of Political Economy 76: 169-217.
Bicchieri, C. (2006). The Grammar of Society: the Nature and Dynamics of Social Norms. Cambridge (MA): Cambridge University Press.
Boehm, C. (1986). Blood Revenge: The Enactment and Management of Conflict in Montenegro and Other Tribal Societies. Philadelphia (PA): University of Pennsylvania Press.
Bowles, S., \& Gintis, H. (2004). The evolution of strong reciprocity: Cooperation in heterogeneous populations. Theoretical Population Biology, 61: 17-28.
Boyd, R. \& Richerson, P. J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. Ethol. Sociobiol. 13: 171-195.
Casari, M. (2007). Emergence of endogenous legal institutions: Property rights and community governance in the Italian Alps. The Journal of Economic History 67(1):191-226.
Cialdini, R. B., Reno, R.R. \& Kallgren, C.A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. Journal of Personality and Social Psychology 58: 1015-1026.
Clutton-Brock, T. H., \& Parker, G.A. (1995). Punishment in animal societies. Nature 373: 209-216.

Elster, J. (1990). Norms of revenge. Ethics 100(4): 862885.

Fessler, D. M. T. \& Haley, K. J. (2003). The strategy of affect: emotions in human cooperation. In P. Hammerstein (Ed.). Genetic and cultural evolution of cooperation (pp. 7 36). Cambridge, MA: MIT Press.

Frank, R. H. (1988). Passions within reason: the strategic role of the emotions. New York, NY: Norton.

Galbiati, R., \& Vertova, P. (2008). Obligations and Cooperative Behaviour in Public Good Games. Games and Economic Behavior 64(1): 146-170.

Gardner, A. \& West, S. A. (2004) Cooperation and punishment, especially in humans. Am. Nat. 164: 753-764.

Giardini, F., Andrighetto, G., Conte, R. (2010). A cognitive model of punishment. Proceedings of the 32nd Annual Conference of the Cognitive Science Society (pp. 1282-1288). Austin, TX: Cognitive Science Society.

Gjeov, S. (1989). Kanuni i leke dukagjinit. Albanian text collected and arranged by Shtjefn Gjeov. New York: Gjonlekaj Publishing Company.

Hamilton, W. D. (1970). Selfish and spiteful behaviour in an evolutionary model. Nature, 228: 1218-1220

Henrich, J., \& Boyd, R. (2001). Why people punish defectors. weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. Journal of Theoretical Biology, 208: 79-89.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., et al. (2006). Costly punishment across human societies. Science, 312(5781), 1767-1770.

Hill, K., Barton, M. \& Hurtado, A. M. (2009). The emergence of human uniqueness: characters underlying behavioral modernity. Evolutionary Anthropology, 18, 187-200

Houser, D., \& Xiao, E. (2010). Understanding context effects. Journal of Economic Behavior \& Organization 73(1): 58-61.

Lorenz, K. (1966). On aggression. London, UK: Methuen.

Janssen, M. \& Bushman, C. (2008). Evolution of cooperation and altruistic punishment when retaliation is possible. J. Theor. Biol., 254, \(541-545\).

Jensen, K. (2010). Punishment and spite, the dark side of cooperation. Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 365, 2635-2650.

Jensen, K., Call, J., \& Tomasello, M. (2007). Chimpanzees are vengeful but not spiteful. Proceedings of the National Academy of Sciences 104: 13046-13050.

McCullough, M.E., Kurzban, R., Tabak, B.A. (2012). A cognitive system for revenge and forgiveness. Behavioral and Brain Sciences, 1-58

Nisbett, R.E. (1993). Violence and U. S. regional culture. American Psychologist 48: 441- 449.

Nisbett, R. E., \& Cohen, D. (1996). Culture of honor. Boulder (CO): Westview Press.

Ostrom, E. (2005). Understanding Institutional Diversity. Princeton (NJ): Princeton University Press.

Petersen, M. B., Sell, A., Tooby, J. \& Cosmides, L. (2012) To punish or repair? Evolutionary psychology and lay intuitions about modern criminal justice. Evolution and Human Behavior. Advance online publication.

Posner, Richard. 1980. "Retribution and related concepts of punishment." Journal of Legal Studies 9: 71-92.

Richerson, P. J., \& Boyd, R. (1998). The evolution of human ultra-sociality. In I. Eibl-Eibisfeldt \& F.Salter (Eds.). Ideology, warfare, and indoctrinability; evolutionary perspectives (pp. \(71-95\) ). London,UK: Berghahn Books.

Villatoro, D., Andrighetto, G., Conte, R., \& Sabater-Mir, J. (2011). Dynamic Sanctioning for Robust and CostEfficient Norm Compliance. Proceedings of the TwentySecond International Joint Conference on Artificial Intelligence (IJCAI 2011)

Zaibert, L. (2006). Punishment and Revenge. Law and Philosophy 25: 81-118.

\title{
The Effect of Left-Hand Gestures on Metaphor Explanation
}

\author{
Paraskevi Argyriou (pxa180@bham.ac.uk) \\ School of Psychology, Frankland Building, University of Birmingham, Edgbaston, B15 2TT, UK \\ Sotaro Kita (S.Kita@bham.ac.uk) \\ School of Psychology, Hills Building, University of Birmingham, Edgbaston, B15 2TT, UK
}

\begin{abstract}
Research suggests that gestures influence cognitive processes, but the exact mechanism is not clear. Additionally, it has been shown that when a linguistic task (metaphor explanation) involves the right brain hemisphere, the left hand becomes more gesturally active. We hypothesized that gestures with a particular hand activate cognitive processes in the contralateral hemisphere. We examined whether gestures with the left hand enhance metaphoricity in verbal responses. Results showed participants produced more metaphoric explanations when instructed to produce gestures with their left hand as compared to the right hand or not gesture at all. In addition, we measured the mouth asymmetry during metaphorical speech to determine individual differences in righthemisphere involvement in metaphor processing. The leftside mouth dominance, indicating stronger right-hemisphere involvement, positively correlated with the left-hand-over-right-hand advantage in gestural facilitation of metaphorical speech. We concluded that left-hand gestures enhance metaphorical thinking in the right hemisphere.
\end{abstract}

Keywords: Metaphor; representational gestures; brain hemispheric lateralization; mouth asymmetry.

\section*{Introduction}

There are many studies providing evidence for the relationship between gestures and cognitive processes, and several theoretical accounts explaining how gestures may determine cognitive processing. However, there is a debate about the type of processes gesture influences (e.g., lexical retrieval, imagery maintenance, and conceptualization; for a review see Kita, 2000), and the mechanism through which gesture influences cognitive processes is not yet clear. In this study, we will focus on the self-oriented functions, that is, the effect that gestures - and in particular representational gestures - have for those who produce them, and we will explore a neural mechanism for gestures' self-oriented functions, which has not been investigated in the literature.

According to the Lexical Retrieval Hypothesis gestures help speakers retrieve the lexical form on a surface level. In particular, it is suggested that gesture related information enters the speech production system to help the grammatical and/or phonological encoding (for a review see Krauss \& Hadar, 2001). Evidence for this hypothesis comes mainly from speech fluency studies. For example, Rauscher, Krauss, and Chen (1996) showed that gesture prohibition
led to more dysfluencies and slower speech rate when talking about spatial concepts. Therefore, it is proposed that gestures promote and facilitate speech production.

Alternatively, according to the Image Maintenance Hypothesis (de Ruiter, 2000) gestures have been thought to help the working memory maintain mental imagery during speech production. In particular, Wesp, Hesse, Keutmann, and Wheaton (2001) have shown that when speakers described images from memory, they used more gestures compared to talking about images they had a physical experience with; thus, indicating that gestures facilitate speakers to represent spatial information and maintain spatial imagery in working memory.

Finally, according to the Information Packaging Hypothesis, gestures help speakers at the conceptualization level; that is to formulate the concept to be uttered. In particular, Alibali, Kita, and Young (2000) showed that speakers gestured differently in two lexically comparable yet conceptually different tasks. Similarly, a gesture prohibition study (Alibali \& Kita, 2010) showed that children who were allowed to gesture could focus more on present perceptual-motor information in their verbal descriptions compared to those who were prohibited from gesturing. Thus, it is suggested that gestures help speakers focus on relevant information, and plan concepts in the way suitable for verbalization.

The above theoretical accounts - which are not necessarily mutually exclusive, rather complementary have attempted to explain how gestures may influence various cognitive processes. However, the mechanism for such effects remains to be explored. The aim of the present study is to determine whether gestures activate cognitive processes in the contra-lateral hemisphere. This is plausible because the hand choice for gesturing is influenced by the brain hemisphere that is predominantly active in a given linguistic task. In particular, Kita, de Condappa, and Mohr (2007) have shown that in right-handers the right-hand over left-hand preference for gesturing is significantly weaker whilst interpreting metaphoric expressions compared to non-metaphoric ones. This finding has been explained in terms of differential hemispheric specialization for linguistic processes, and in particular the key role that the right hemisphere has in the processing of figurative language (following the Right Hemisphere Hypothesis for Metaphor; see for example, Brownell, et al., 2007; for alternative views, see Cardillo et al., 2012 and Rapp, et al., 2007); that
is when a metaphor task activates the right hemisphere, this activation increases the frequency of the left-hand gestures. The present study tested the reverse causality: Do left-hand gestures activate metaphorical processes?

To investigate this hypothesis, we manipulated which hand is used for gesturing and assessed the performance in a metaphor explanation task. More specifically, participants were asked to explain the metaphorical mapping in English idiomatic expressions with metaphorical meaning (e.g., "to spill the beans" meaning "to reveal secrets"). These expressions and task have been previously shown to engage metaphorical thinking, and furthermore to activate the right hemisphere. For example, when participants explain such metaphorical expressions they demonstrate reduced righthand choice for gesturing (Kita et al., 2007), and reduced right-sided mouth dominance (Argyriou \& Kita, in prep.) than when they explain non-metaphorical expressions. Gesture production was manipulated within-participants by asking subjects to gesture with their left hand only, right only, or do not gesture at all. If gestures activate cognitive processes in the contra-lateral hemisphere, then metaphor explanations should demonstrate higher level of metaphoricity when participants gestured with their left hand compared to the other two gesturing conditions.

In addition, in order to further support the hypothesis, mouth asymmetry measurements during metaphor explanation were collected from the same group of participants. Mouth asymmetry has been agreed to indicate relative hemispheric specialization for speech production, and in particular the right-sided mouth asymmetry observed during verbal tasks has been related to the left hemisphere cerebral specialization for language production (for a review see Graves \& Landis, 1990). Moreover, Argyriou and Kita (in prep.) showed that mouth openings are more left-side dominant in a metaphor explanation task than in a concrete phrase explanation task, indicating the right-hemispheric specialization for metaphor. Therefore, we expected that the observed left-side bias in mouth openings during metaphor explanation would be positively correlated with the lefthand gesture advantage on speech metaphoricity.

\section*{Method}

\section*{Participants}

31 right-handed, male, native English speakers and monolinguals before the age of 5 years (age: \(M=20.35, S D=\) 2.86) participated in the experiment for course credit or \(£ 4\). They were all right-handed according to a standardized 12item handedness questionnaire (Oldfield, 1971): a score of " 1 ", " 0.5 " and " 0 " was given for each right-hand, either and left-hand preference respectively. We calculated the mean of the sum of these scores, and defined as right-handed those participants who scored at least 8.5. None of the participants had any previous serious injury to the face or jaw. All of them were recruited at the University of Birmingham. We focused on male speakers because
bilateral representation of language processing in men is less compared to women (McGlone, 1980).

\section*{Stimuli}

For the main descriptive task we used eighteen English expressions with metaphorical meaning. We created three (plus one in case one expression was unknown) additional metaphorical and concrete expressions for the mouth asymmetry task (see Table 1).

Table 1: Complete list of stimuli for the two tasks.
\begin{tabular}{|c|c|}
\hline \multicolumn{2}{|l|}{Metaphorical expressions for main descriptive task} \\
\hline To burst someone's & To sit on the fence \\
\hline bubble & To skate on thin ice \\
\hline To cross that bridge later & To spill the beans \\
\hline To dodge the bullet & To stand your ground \\
\hline To fall back down to earth & To take the bull by the horns \\
\hline To get back in the saddle & To turn a corner \\
\hline To get hot under the collar & To turn the tables \\
\hline To hold all the cards & Water under the bridge \\
\hline \multicolumn{2}{|l|}{To leave a bad taste in the mouth} \\
\hline \multicolumn{2}{|l|}{To look on the bright side} \\
\hline \multicolumn{2}{|l|}{Metaphorical expressions for the mouth asymmetry task} \\
\hline To pour oil onto the fire & To spin a yarn \\
\hline To set your sights higher & (To hit the nail on the head) \\
\hline \multicolumn{2}{|l|}{Concrete expressions for the mouth asymmetry task} \\
\hline To pour oil into the pan To put a shelf higher & To spin a golf ball (To hit someone on the head) \\
\hline
\end{tabular}

\section*{Procedure}

Participants were tested individually. They were seated on a chair, which was located between two tables of the same height ( 71 cm tall). The experimenter was facing the participant, and the video camera (Sanyo HD camera) was placed next to the experimenter. Stimuli were presented one by one on a white sheet of paper (font size 72, Times New Roman), which was held by the experimenter until the participant started the description.

Participants were instructed to explain the meaning of stimuli as if they were explaining it to a non-native English speaker. To encourage metaphorical thinking, participants were instructed to include an explanation as to how the literal meaning can be mapped on to the metaphorical meaning of the expression (e.g., in the expression "to spill the beans", "beans" refer to secrets, and "spilling" refers to spreading them to everybody). During the description, participants were told to place one of their hands on the indicated marks (white sticky dots) on the surface of the table(s), and to keep it still for the whole procedure. For the total prohibition condition, participants were asked to place both their hands on the tables (see Figure 1). For the gesturing conditions, they were instructed to gesture with
their free hand during the description (gesture encouragement instruction followed the paradigm in Chu \& Kita, 2011). Participants were debriefed about the purpose of the hands immobilization after the experiment and the permission to use the data was allowed. Order of stimuli (forward - reverse), and order of hand(s) prohibition was counterbalanced across participants in a within-participants design.


Figure 1: Experimental conditions (from left to right) Right Hand Gesturing, Left Hand Gesturing, No Gesturing.

In the mouth asymmetry task participants were instructed to explain metaphorical expressions, just as in the main task, and concrete expressions whilst both hands were prohibited. The order of the tasks (concrete - metaphor) was counterbalanced across participants. Hand prohibition was a necessary experimental control in order to collect a pure measurement of participants' hemispheric specialization for metaphor without any influence from gesturing. Videorecording zoomed-in on the face area.

\section*{Measures}

The verbal responses from the main task were transcribed and coded for their level of metaphoricity. The level of metaphoricity was measured based on whether the explanations included an explicit link between the literal and metaphorical meanings, and whether participants explicitly referred to the mapping between the source and target domains of the conceptual metaphor underlying each idiomatic expression (adopted from McGlone, 1996). More specifically \({ }^{1}\), a ' 0 '' rating indicated that the explanation did not contain words or phrases referring to the source domain of the relevant conceptual metaphor, therefore there was no

\footnotetext{
\({ }^{1}\) To illustrate how the \(0-2\) metaphoricity coding has been used, consider the following explanations generated for the item "to spill the beans": (a) "To spill the beans is to tell someone a secret or gossip" was coded with 0 because the explanation includes the meaning of the expression only. (b) "To spill the beans means to let something out, to tell someone something perhaps that you shouldn't been telling them; I guess the beans like information make a mess once spilling them" was coded with 1 because there is an implicit reference to the beans representing the information. (c) "To spill the beans is to tell someone something that you were not meant to tell; something which was confidential, private, and the beans represent the information that was private and by spilling them you are telling the news." was coded with 2 because it includes an explicit mapping between the source and target domains, and participant mentions the representation of each concept.
}
metaphorical cross-domain mapping; a rating of " 1 " indicated that the explanation contained words or phrases that might be construed as references to the source domain, but the references were ambiguous, and the mapping between the two domains implicit; a rating of ' 2 '" indicated that the explanation contained words or phrases that clearly refer to the source and target domains, and the mapping was explicit.

One individual "blind" coder was trained and coded 33\% of the total verbal responses in terms of metaphoricity. All answers from 10 randomly selected participants were coded (in total 180 trials were double coded). Coding of metaphoricity matched between the two coders \(76 \%\) of the time (Cohen's weighted kappa \(\kappa_{\mathrm{w}}=.68, p<.001\), kappa maximum \(\kappa_{\max }=.91\) ).

Video recordings from the three gesturing conditions in the main task were analyzed using ELAN software (developed by the Max Planck Institute for Psycholinguists, Nijmegen, the Netherlands). They were coded on a trial-bytrial basis to locate the existence of at least one gesture type; that is representational gestures, palm-revealing gestures, conduit, and other (e.g., beats).

Video recordings from the mouth asymmetry task were analyzed on a frame-by-frame basis using ELAN software. The first ten mouth openings were coded per trial for each participant (sixty mouth openings in total). We measured the laterality at each maximum mouth opening. One maximum opening was defined as the widest point the mouth opens since the lips open to the lips resting or the lips meeting completely. The options for laterality classification were: right-side dominant, left-side dominant, or sides equally open (see Figure 2 for examples). Maximum openings during filled-pauses were included, but not the ones for nonspeaking purposes (e.g., smile), nor the ones whilst participants were repeating the phrase to be explained.


Figure 2: (From left to right) Examples of right, left, equal maximum mouth openings. "Right" and "left" refer to the speaker's right and left.

One individual "blind" coder was trained and coded \(22 \%\) of the data in terms of right, left or equal dominance of mouth openings. All mouth openings from 7 randomly selected participants were coded (in total 414 maximum mouth openings were double coded). Coding of dominance matched between the two coders 79\% of the time (Cohen's kарра \(\kappa=.66, p<.001\) ).

The degree of left-side mouth dominance was computed for each participant based on the laterality (right-R, left-L, equal-E) of their 30 maximum mouth openings for each task (concrete and metaphor), and using the following formula:
(L-R)/(L+R+E) (adopted and adjusted from Holowka \& Petitto, 2002). Thus, a positive and/or low negative mean score indicated more instances of left-side dominant mouth openings during metaphor explanation (right-hemispheric lateralization), and a high negative score indicated more instances of right-side dominant mouth openings (lefthemispheric lateralization).

In addition, we calculated a left-hand gesture advantage index whilst participants gestured and explain metaphors in the main descriptive task. That is, the average metaphoricity per participant when gesturing with the left hand minus the average metaphoricity when gesturing with the right hand. Thus, a high and positive mean score indicated that participants were more metaphoric when gesturing with their left hand compared to the right (= left-hand gesturing advantage on metaphoricity). A negative or low positive mean score indicated that participants were more metaphoric when gesturing with their right hand compared to the left.

\section*{Design \& Analysis}

Out of the 522 trials in total in the main task, \(4 \%\) was excluded as failed trials; that is when the participants did not proceed as instructed (e.g., no gesture production when right or left hand was free), and when they did not know the expressions. The independent variable was which hand was free to gesture: right-hand gesturing, left-hand gesturing, not gesturing. The dependent variable was the level of the metaphoricity of the explanations (see the section Measures).

\section*{Results and discussion}

Out of the 354 gesturing trials, \(99 \%\) included at least one representational gesture; 23\% included at least one palmrevealing gesture; 7\% included at least one conduit gesture; \(18 \%\) included at least one "other" gesture - comprising mainly of beat and metacognitive gestures. Thus, the instruction to produce gesture was effective and we may assume that whatever the gesturing effect is, it will be due to representational gestures during the gesturing trials.

One-way repeated measures ANOVA was conducted to compare the effect of gesturing hand on level of speech metaphoricity in the three gesturing conditions (left-hand gesturing, right-hand gesturing, not gesturing at all). There was a significant effect of the gesturing hand, \(F(2,60)=\) 13.92, \(p<.001, \eta^{2}=.32\). Post hoc comparisons with Bonferroni correction between conditions indicated that level of speech metaphoricity was significantly higher when participants gestured with the left hand than not gesturing ( \(t(30)=2.81, p<.001\) ); metaphoricity was significantly higher when gesturing with the right hand than not gesturing \((t(30)=1.38, p=.028)\); and metaphoricity was significantly higher when gesturing with the left hand compared to right hand \((t(30)=1.43, p=.038)\) (see Figure 3). Thus, the gesturing hand had an effect on the level of metaphoricity in
speech. Specifically, gestures, especially, those by the left hand, improved metaphorical thinking.

We focused on trials in which only representational gestures were produced, and we limited the analysis to individuals who had trials with representational gestures only (2 participants were excluded; \(N=29\) ). Pattern of the results remained the same: left-hand gesturing ( \(M=1.53\), \(S E=.08\) ), not gesturing ( \(M=1.15, S E=.06\) ), right-hand gesturing ( \(M=1.39, S E=.08\) ). Also, the one-way repeated measures ANOVA remained significant \((F(2,56)=14.87\), \(\left.p<.001, \eta^{2}=.35\right)\). Thus, there is evidence that effect of the gesturing hand is due to representational gestures.


Figure 3: Average metaphoricity in speech in the three gesturing conditions. Error bars represent standard errors.

Next we compared the left-side bias in mouth openings during concrete and metaphor explanations. A one-way repeated-measures ANOVA performed on the left-side mouth dominance index with linguistic task as the independent variable yielded significant effect of the task, \(F(1,30)=6.45, p=.016, \eta^{2}=.18\). The left-side bias was higher during metaphor explanations ( \(M=-.11, S E=.08\) ) compared to the concrete ones ( \(M=-.24, S E=.09\) ), thus suggesting a reduced right-sided mouth asymmetry during explanation of metaphorical expressions. More importantly, we assessed the relationship between the left-side bias in mouth openings and the left-hand gesturing advantage during metaphor explanation. The range on the mouth asymmetry measurement was -0.90 to 0.77 , where positive scores indicate a right-hemispheric lateralization (= that is participants open their left side of the mouth wider than the right whilst explaining metaphors). The range on the lefthand gesture advantage index was -0.30 to 0.83 , where higher positive scores indicate that participants were more metaphoric when they gestured with the left than with the right hand. There was a significant positive correlation between the two scores \((r(29)=.38, p=.036)\) (see Figure 4). Thus, the participants for whom the left-hand gesturing advantage was bigger tended to have a stronger righthemisphere involvement in metaphoric speech production. Note further that the mouth asymmetry during explanation
of concrete expressions did not significantly correlate with the left-hand gesture advantage \((r(29)=.32, p>.05)\).


Figure 4: The scatter plot for the correlation between the left-side mouth dominance and left-hand gesture advantage during metaphor explanations.

\section*{General Discussion}

The goal of the present study was to investigate a neural mechanism for gestures' self-oriented functions. We measured the level of metaphoricity in metaphor explanations as a function of the hand used for gesture: the right hand, the left hand, no hands. We found that speakers produced more metaphoric verbal responses when they gestured with either hand compared to not gesturing at all, and when they gestured with the left hand compared to the right. We propose that left-hand gestures led to better performance in metaphor explanation because they activated metaphorical processing in the right hemisphere.

The present findings are in line with the Information Packaging Hypothesis (e.g., Alibali, Kita, \& Young, 2000), indicating that gesture helps the conceptual planning of the speech, and in particular the conceptual mapping for metaphorical speech.

In addition, the present results are compatible with previous studies on gesture and metaphor. For example, the present study found that metaphoricity was higher when gesturing, regardless of the hand, than when not gesturing. This is compatible with the observations that gesture inhibition reduces the use of metaphorical spatial language (Bos \& Cienki, 2011). More importantly, the findings shed new light on the inter-relation between the hand used for gesturing and hemispheric specialization. Kita et al. (2007) showed that hand choice for gesturing can be determined by the relative hemispheric specialization during different linguistic tasks. Thus, right-hand preference is reduced during metaphor explanations compared to concrete or abstract ones. Our findings provide evidence for the reverse causal link. That is, the gesturing hand can determine the level of speech metaphoricity, and in particular left-hand gestures enhance metaphor explanations. So, there seems to be a bi-directional causal relationship between left-hand gestures and metaphorical processing.

Although there are several studies, which manipulate gesturing in order to assess gestures' effect on cognitive processes (e.g., Alibali \& Kita, 2010; Rauscher et al., 1996), as far as we know, this is the first study to explore the neural mechanism for gestures' self-oriented functions, and link it with the hemispheric lateralization of cognitive processes. Crucially, the left-hand gesture advantage for metaphoricity significantly correlated with the left-side dominance in mouth openings for metaphorical expressions, but not for concrete expressions. That is, the left-hand gesture advantage is stronger in speakers who have strong righthemispheric control for metaphor. Thus, it further supports the idea that gesturing activates cognitive processes in the contra-lateral hemisphere.

But, how exactly this neural mechanism works? We may speculate how based on our current findings, and also in light of metaphor theories. Metaphor is considered as a matter of conceptualizing one conceptual domain in terms of another (Lakoff \& Johnson, 1980), and specifically the metaphorical mapping requires speakers to map a concrete concept on to a more abstract one. In addition, this mapping requires the conceptualization of a distant semantic relationship between the source and target domains of the metaphor, and it is considered to be predominantly computed in the right hemisphere, which processes coarse grained semantic information (Jung-Beeman, 2005). For example, in the expression to "spill the beans" participants had to represent the abstract concept of IDEAS (target) in terms of the distant concrete concept of OBJECTS (source). Our findings revealed that left-hand gestures were particularly beneficial compared to the right-hand ones for the metaphorical mapping. Therefore, we suggest that lefthand gestures make the distant semantic relationship between target and source domains of the metaphor to become closer, and then speakers can represent the metaphorical mapping in speech, thus become more metaphoric. It seems that left-hand gestures give some kind of "feedback" to the contra-lateral right hemisphere ("Hemisphere-Specific Feedback Hypothesis") and promote metaphorical processing, which crucially involves the right hemisphere.

The present study did not account for what aspects of gestural hand movement influences metaphorical thinking. More specifically, our findings cannot address the question, "is it the gesture or the arm movement per se which activates the processes in the contra-lateral hemisphere?" Previous studies (Ravizza, 2003) have shown that meaningless arm movements, such as tapping, may facilitate lexical retrieval. However, this is only in tasks where lexical items have been selected by automatic spreading activations but not sufficiently so, and not in tasks where words have to be strategically searched. We may assume that metaphorical mapping requires strategic search of semantic fields, thus arm movement per se may not facilitate the process. Thus, it is the depictive nature of the gestural movement as described above that enhanced participants' performance rather than merely the arm movement. Moreover, even
when the analysis included trials with representational gestures only, the results remained significant and demonstrated the same pattern (= left-hand gesture advantage). Thus, it provided implicit support for the effect of the depictive nature of representational gestures. However, future research to compare the effect of meaningless versus meaningful arm movements on metaphorical thinking would directly assess this issue.

In conclusion, the current study has advanced our knowledge of and enhanced theoretical accounts on a neural mechanism for gestures' self-oriented functions, which have received little attention so far. We propose that gestures activate cognitive process in the contra-lateral hemisphere such that left-hand gestures enhance a right-hemispheric specialized process such as metaphor processing.

\section*{Acknowledgments}

This work was supported by Paraskevi's Argyriou ESRC award for pursuing Ph.D. studies at the University of Birmingham. We would like to thank the financial support from the School of Psychology at the University of Birmingham for the study. We thank Sarah Byfield and Monica Leverton for their help with data coding.

\section*{References}

Alibali, M. W., \& Kita, S. (2010). Gesture highlights perceptually present information for speakers. Gesture, 10(1), 3-28. doi: 10.1075/gest.10.1.02ali
Alibali, M. W., Kita, S., \& Young, A. J. (2000). Gesture and the process of speech production: We think, therefore we gesture. Language and Cognitive Processes, 15(6), 593613.

Argyriou, P., \& Kita, S. (in prep.). Metaphor explanation reduces the right-sided mouth asymmetry during speech production.
Bos, A., \& Cienki, A. (2011). Inhibiting gesture reduces the amount of spatial metaphors in speech. Paper presented at the GESPIN Gesture and Speech in Interaction, Bielefeld, Germany.
Brownell, H., Lundgren, K., Cayer-Meade, C., Nichols, M., Caddick, K., \& Spitzer, J. (2007). Assessing quality of metaphor interpretation by right hemisphere damaged patients. Brain and Language, 103(1-2), 197-198. doi: 10.1016/j.bandl.2007.07.113

Cardillo, E. R., Watson, C. E., Schmidt, G. L., Kranjec, A., \& Chatterjee, A. (2012). From novel to familiar: Tuning the brain for metaphors. Neuroimage, 59(4), 3212-3221. doi: 10.1016/j.neuroimage.2011.11.079
Chu, M. Y., \& Kita, S. (2011). The Nature of Gestures' Beneficial Role in Spatial Problem Solving. Journal of Experimental Psychology-General, 140(1), 102-116. doi: 10.1037/a0021790
de Ruiter, J. P. (2000). The production of gesture and speech. In D. McNeill (Ed.), Language and gesture. Cambridge: Cambridge University Press.

Holowka, S., \& Petitto, L. A. (2002). Left hemisphere cerebral specialization for babies while babbling. Science, 297(5586), 1515-1515. doi: 10.1126/science. 1074941
Jung-Beeman, M. (2005). Bilateral brain processes for comprehending natural language. Trends in Cognitive Sciences, 9(11), 512-518. doi: 10.1016/j.tics.2005.09.009
Kita, S. (2000). How representational gestures help speaking. In D. McNeill (Ed.), Language and gesture. Cambridge: Cambridge University Press.
Kita, S., de Condappa, O., \& Mohr, C. (2007). Metaphor explanation attenuates the right-hand preference for depictive co-speech gestures that imitate actions. Brain and Language, 101(3), 185-197. doi: 10.1016/j.bandl.2006.11.006

Krauss, R., \& Hadar, U. (2001). The Role of SpeechRelated Arm/Hand Gestures in Word Retrieval. In R. Campbell \& L. Messing (Eds.), Gesture, speech, and sign. Oxford: Oxford University Press.
Lakoff, G., \& Johnson, M. (1980). Metaphors We Live By. Chicago: University of Chicago Press.
McGlone, J. (1980). Sex-differences in human-brain asymmetry - a critical survey. Behavioral and Brain Sciences, 3(2), 215-227.
McGlone, M. S. (1996). Conceptual metaphors and figurative language interpretation: Food for thought? Journal of Memory and Language, 35(4), 544-565. doi: 10.1006/jmla.1996.0029

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia, 9(1), 97-113. doi: 10.1016/0028-3932(71)90067-4
Rapp, A. M., Leube, D. T., Erb, M., Grodd, W., \& Kircher, T. T. J. (2007). Laterality in metaphor processing: Lack of evidence from functional magnetic resonance imaging for the right hemisphere theory. Brain and Language, 100(2), 142-149. doi: 10.1016/j.bandl.2006.04.004
Rauscher, F. H., Krauss, R. M., \& Chen, Y. S. (1996). Gesture, speech, and lexical access: The role of lexical movements in speech production. Psychological Science, 7(4), 226-231. doi: 10.1111/j.1467-9280.1996.tb00364.x
Ravizza, S. (2003). Movement and lexical access: Do noniconic gestures aid in retrieval? Psychonomic Bulletin \& Review, 10(3), 610-615. doi: 10.3758/bf03196522
Wesp, R., Hesse, J., Keutmann, D., \& Wheaton, K. (2001). Gestures maintain spatial imagery. American Journal of Psychology, 114(4), 591-600. doi: 10.2307/1423612

\title{
Cognitive Externalism meets Bounded Rationality
}

\author{
Eric Arnau (eric.arnau@gmail.com) \\ Department of Philosophy, Universitat Autònoma de Barcelona \\ Edifici B, Campus UAB, 08193 Bellaterra, Spain
}

\author{
Saray Ayala (sarayayala@gmail.com) \\ Department of Humanities: Philosophy, Language and Literature \\ University Carlos III, Madrid 126, 28903 Getafe, Spain
}

Thomas Sturm (tsturm@mpiwg-berlin.mpg.de)
Department of Philosophy, Universitat Autònoma de Barcelona
Edifici B, Campus UAB, 08193 Bellaterra, Spain

\begin{abstract}
When proponents of cognitive externalism (CE) have turned to empirical studies in cognitive science to put the framework to use, they have typically referred to perception, memory or motor coordination. Not much has been said about reasoning. One promising avenue to explore here is the theory of bounded rationality (BR). In this paper, we try to clarify the potential relationship between these two programs. We start by discussing Andy Clark's interpretation of BR, which we find unconvincing in several respects. Next, we take a closer look at CE in order defend a version of it that stands against mainstream internalism without committing itself to constitutional claims about the mind. We then turn to analyze BR from the CE perspective. Finally, we argue that internalism about cognition cannot explain important aspects of the BR program.
\end{abstract}

Keywords: extended cognition; bounded rationality; heuristics.

\section*{Introduction}

By Cognitive externalism (CE) we refer to the framework that accommodates the initially differentiated challenges to the internalist picture of cognition, developed under the flag of extended cognition and distributed cognition. CE departs from the original proposal introduced by Clark and Chalmers (1998) in which what determines the cognitive status of an extended process is its functional parity with an intracranial cognitive process. The core of this shift is the complementarity of internal and external elements (Menary 2007, Sutton 2010, Wilson and Clark 2009). The argument for extended cognition turns on the way different inner and outer components co-operate so as to yield an integrated system capable of supporting intelligent behavior.

One consequence of this revised CE is a broadening of the span of the putative cases considered, as it points our attention to the many ways in which "the computational power and expertise is spread across a heterogeneous assembly of brains, bodies, artifacts, and other external structures" (Clark 1998, p. 77). Once parity is not required, we can tackle any manipulation of structures or elements that is integrated -e.g. measurement instruments, information storage devices, representational systems, etc.

Typically, proponents of externalist accounts of cognition have sought for empirical support from research in lower cognition, mostly memory (Clark and Chalmers 1989, Sutton et. al. 2010), and perception (Wilson 2010). But seldom the literature has addresses the realm of higher cognition. Here we want to explore the benefits that CE could offer to empirical research on human reasoning, and whether such research can vindicate \(C E\).

\section*{Bounded Rationality}

A promising avenue to explore is the interplay between CE and the program of Bounded Rationality (BR) as it is developed by Gerd Gigerenzer and colleagues into the Fast and Frugal Heuristics theory. The core research question of BR is: "How do people make judgments and decisions in everyday life, when time and information is limited and the future uncertain?" (Gigerenzer 2008, p. 5).

From here, the program unfolds in three dimensions: (1) The core descriptive tenet is that people typically rely on fast and frugal heuristics. There are strategies, conscious or unconscious, that search for minimal information and exploit evolved capacities and environmental structures (Gigerenzer and Brighton, 2009). (2) The core normative tenet is that such reliance on heuristics is at least oftentimes desirable. Heuristics work remarkably well; as good as -or even better than- optimizing strategies. They are not merely second-best options for "wanna-be optimizers". (3) The core prescriptive tenet is that the BR picture of reasoning affords prescriptive guidance, opening a space in which to enhance reasoning by intervening on the environment rather than in the inner processing.

There is a strong prima facie affinity between BR and CE. In the BR literature one can find many claims that echo the externalist tenets. We find claims such as that "the heuristic lets the environment do much of the work" (Gigerenzer and Hutchinson 2005, p. 101), or that "in the ecological view, thinking does not happen simply in the mind, but in interaction between the mind and its environment" (Gigerenzer 2008, p. 17, emphasis added). Somewhat surprisingly, BR has attracted only one discussion within
the externalist camp, by Andy Clark. In the next section we turn to his discussion of BR.

\section*{Clark on Bounded Rationality}

Clark (2001, 2003) addresses the relation between CE and BR guided by his own defense of CE, but also by his favoring of a "naturalistic" or "mechanical" account of rationality, i.e. one that is merely descriptive. Clark claims that there are two ways to develop a "naturalistic" or "mechanical" account of rationality based on a noninternalist approach in cognitive science. First, a "biological cognitive incrementalism" (BCI), according to which "fullscale human rationality is reached, rather directly, by some series of tweaks to basic biological modes of adaptive response" (2001, p. 121). Defenders of BCI view reasoning or rationality as continuous to more "basic" processes, such as perceptual responses to one's environment, that do not require the manipulation of symbols. A second option assumes that rationality is at best indirectly based upon processes of BCI and, instead, strongly based upon "symbiotic relationships with knowledge-rich artifacts and technologies" (2001, p. 122). Not surprisingly, Clark favors the latter view, since it obviously comes down to his version of CE. His argument accordingly should have two sides, one being directed against BCI, the other in favor of CE.

In Clark's view (2001, p. 126), BCI can be characterized by three core assumptions: (1) the thesis of organismenvironment interaction, or anti-representationalism; (2) the modularity thesis (heuristics, short-cuts which are locally rather than globally active); (3) the thesis of distributed cognition. He illustrates these assumptions with research on the wing-flapping of houseflies and phonotaxis in robot crickets (ibid., pp. 126-129). All behave successfully in their environments using extremely simple means. This "breeds skepticism" that "symbols, internal representations and the like play little role even in advanced human problemsolving" (ibid., p. 129). However, as Clark rightly points out, the anti-representationalism entailed by this view must not be taken too far, since typical instances of reasoning, such as drawing an inference or making a choice, are all "representation-hungry" activities (ibid.). Clark does not claim that BR requires that assumption, but he does think that BR is committed to assumptions 2 and 3 . Simple heuristics, after all, are supposed to make us smart because of their joint exploitation of evolved capacities and the structure of (the information in) the environment. To that extent, BR seems to be a naturalistic account of rationality that also subscribes to BCI.

Let us assume for the moment that the program of BR is indeed to be understood as endorsing (2) and (3). What, then, is Clark's objection to these contentions? His claim is that BR, thus understood, still does not grasp what we understand as full-scale human rationality (Clark 2001, p. 131). As he also says, "In much of this recent work, traditional conceptions of thought, reason and action are not so much reworked as by-passed entirely." (ibid., p. 126). But what is missing? Elsewhere he declares: "Rational behavior is, in some sense, behavior that is guided by, or
sensitive to, reasons. Intuitively, this seems to involve some capacity to step back, and assess the options; to foresee the consequences, and to act accordingly." (Clark 2003, p. 314) But this is surely only a necessary condition, and it is surely not denied by Gigerenzer's program. Although he emphasizes that we should often rely on such heuristics, he does in no way think they are to be used blindly. Any reasoner must ask himself or herself, Am I in an environment where this heuristic works? If not, can I use another heuristic, and which one? That obviously involves the ability to step back, assess the options, and so on. We should also note in this context that Gigerenzer has explicitly distanced himself from characterizing heuristics in terms of the modularity thesis: While they are short-cuts that are tuned to specific domains, heuristics need not be viewed as implastic as, say, perceptual processes (see e.g. Gigerenzer 2007, p. 43f.).

Now, what about the argument in favor of Clark's own position, viz. that rationality can better be understood by his extended mind thesis? Clark (2001, pp. 132ff.) points to the frequent close coupling of reasoning processes with artificial aids and scaffolds. But he actually expresses his thesis in two different ways: (1) We do understand rationality only if we see it as a capacity "tuned and applied to the very special domain of external and/or artificial cognitive aids" (2001, p. 131). (2) "human thought and reason is sculpted, enhanced, and ultimately transformed" by technology (ibid.). Certainly it is one thing to view rationality as a capacity tuned to work with certain external technologies, and quite another thing that to say that this capacity is (partly) constituted by technology. If our faculty of reason is fitted towards such aids, it should exist beforehand. Conversely, if it is "sculpted and transformed" by those technologies, it probably does not exist beforehand. Claims (1) and (2) do not easily go together, at least not without much further ado.

Perhaps here there is a mixing of rationality as a faculty and rational processes. That reasoning processes are causally related to such technological aids, or even constituted by them, does not show that rationality as a faculty is constituted by these interactions. Indeed, it cannot: What constitutes any mental faculty is commonly understood to be expressed in terms of laws that guide the behavior of the capacity -in the case of rationality, the usual suspects include rules of logic, probability theory and heuristics. To explain the constitution of a faculty by the constitution of certain processes is simply a category mistake.

Against this, Clark might bring up his view that reasoning involves feedback loops between the potential for refinement through reflection on its own basis (Clark 2001, p. 134). Fair enough. When someone gives us a reason for a belief or a decision, it is common to ask back "Why?", requesting the thinker's grounds and principles. But these may always be questioned too, of course. Moreover, using technological aids or developing new notations (in mathematics or logic, say) has been an essential part of these feedback loops. This has led to the repeated
renovation of the laws of proper reasoning. Assuming that the faculty of reason is constituted by some fixed laws is therefore problematic. If rationality possesses no immutable nature, it may seem that the distinction between the faculty and processes of reasoning cannot be used to undermine Clark's position. By truly engaging in such processes, we are caught in those feedback loops, and create the faculty anew, with no end in sight.

The problem with this rejoinder is that it is inconsistent with Clark's focus on a purely descriptive, "mechanical" account of rationality. Insofar as we ask for better modes and principles of reasoning, we engage in genuinely normative issues. BR doesn't face this problem because, pace Clark's reading, it does not commit to such a purely descriptive account. In contrast, one of the most interesting features of BR is that it constitutes a new attempt for bridging Is and Ought, at least for certain domains of human reasoning (Gigerenzer and Sturm 2012). This bridging is, in part, possible because BR does not subscribe to BCI in the way Clark thinks it does.

To sum up: While Clark thinks there is some good in the BR approach, he also thinks it is too much committed to BCI. This is largely a misunderstanding of the program. Clark's alternative position is expressed by claims that tend to be inconsistent. Now, before getting back to its relation with BR, let's take a closer look at CE.

\section*{Cognition Embedded, Scaffolded or Extended?}

The cognitive externalist wants us to treat cases of complementary integration of internal and external resources as a whole, and not to frame cognition as the working of an organic system that is just causally embedded in external resources. One thing that critics object is that whatever the explanatory benefits of CE are, you can get the same from internalism, so it wins, if only for parsimony concerns. To address this charge, we need to spell out the precise statement of what CE proposes for the putative cases in which environmental resources are involved in cognitive processes. There are three basic options:
(a) Embedded cognition takes the domain of complementarity-motivated putative cases of external cognition as showing the ways in which cognition is causally embedded in features of the environment that surround and supplement real cognition, which remains still located within the organism. Ontologically speaking, it is a claim of mere causal dependence. On the methodological side, the claim is that cognitive sciences should be mostly concerned in studying processes that take place within the organism, and not outside (Sutton et. al. 2010).
(b) Scaffolded cognition is the idea is that (at least some of) our cognitive capacities both depend on and have been transformed by our manipulations of environmental resources. The claim here is not about mere causal dependence but about integrative coupling between internal and external elements. Accordingly, cognitive science should study these processes as they appear distributed across organism and environment, instead of isolating the internal (Sterelny 2010).
(c) Extended cognition claims that sometimes cognitive processes and systems are literally extended, having regions of the environment as proper parts located outside the organism. In other words, sometimes manipulated elements and structures of the environment, material or otherwise, constitute part of the cognitive system. This claim of constitution is held hand in hand with the urge for a revisionary attitude in the cognitive sciences towards the study of such extended processes without isolating its biological parts.

These views can be considered as stretches in a continuum, each best covering a different range of putative cases (Sterelny 2010, Sutton 2010). Adjudicating between the three options might be more a matter of degree and preponderance than a 'winner-takes-it-all' situation. This is so because the range of cognitive phenomena that motivates CE is heterogeneous. Among other dimensions of variation (Sterelny 2010, Wilson and Clark 2009) we can distinguish between (a) Individual artifacts, such as notebooks or sensory substitution devices, that given certain conditions might call for a genuine extended cognition reading à la Clark; (b) Collective resources, in which the external cognitive resources are embedded in a collective activity which involves several coordinated individual agents, as in Hutchins' (1995) case study of the distributed processes that enable ship navigation; and (c) public resources, like symbolic representational systems.

The relationship between agents and public resources is best seen as a process of cognitive niche construction in which humans sculpt their environment so that it affords novel cognitive possibilities (Sterelny 2010). The real issue, then, is not the current synchronic location of the elements that constitute the cognitive system, but the integration between internal and environmental resources. Thus, "resources can be extended in the relevant explanatory sense even when they are not literally external" (Sutton et. al. 2010, p. 535). It is the manipulation of such resources and the transformative effect they have on the individual cognitive profile that provides the explanatory cornerstone.

The common internalist strategy is to conflate embedded with scaffolded cognition and contrast them both with extended cognition. They stress the difference between claims of dependence and claims of constitution. By insisting that the putative examples of scaffolded cognition involve a claim of dependence, they see them as grist on the internalist's mill.

The way to resist this move is to highlight that, beyond ontological qualms, the idea of scaffolded cognition moves cognitive science in practice in the same line as that of extended cognition (Sutton et. al. 2010). Despite the skepticism that the ontological claims of extended cognition can bring about, when it comes to the explanatory tasks of empirical research, we take the most significant divide to be that between embedded cognition on one side and scaffolded and extended cognition on the other. That is the choice between cognitive internalism and cognitive externalism. In the next section we will address BR from the perspective here sketched.

\section*{Unpacking Bounded Rationality}

BR tells us that simple heuristics make us smart by exploiting the environment. But, what does that mean? In order to further assess BR, we need to unpack its vague appeals to the environment.

\section*{The Execution of Reasoning}

Let us focus first on the role that the physical environment has in reasoning. Some of the heuristics discussed by the proponents of BR require for their execution that agents are in current sensorimotor interaction with the relevant physical environment, although it is doubtful that they would qualify, within the broad realm of problem solving, as instances of genuine reasoning. A model example of this kind of heuristics is the gaze heuristic that people use to catch a flying ball: Fix your gaze on the ball, start running, and adjust your running speed so that the angle of gaze remains constant. "A player who relies on the gaze heuristic can ignore all causal variables necessary to compute the trajectory of the ball (...) the player will end up where the ball comes down without computing the exact spot" (Gigerenzer and Brighton 2009, p. 110).

The execution of these heuristics requires that agents are in current physical contact with the relevant environment. The agent relies on the manipulation of the environment to solve the problem. The baseball player alters the relative position of the ball in the egocentric space. This subset of heuristics, although not distinctively characterized in the BR literature, is best seen as epistemic actions. By acting on the environment itself, agents dispense with the need of otherwise required complex internal representations. As they alter the physical and informational structure of the environment, these processes take the agent closer to the solution; thus they are part of the agent's processing of the problem (Menary 2010).

However, just as much of human reasoning, many of the heuristics analyzed by the BR program do operate decoupled from the environment. What role does the environment play in those cases? Consider a much discussed cognitive task: estimating the relative size of two cities. Two suitable strategies provide a good illustration of the paradigmatic kind of heuristic that BR puts forth.

The first is the Recognition Heuristic: if one of the options is recognized and the others are not, infer that the recognized alternative has the higher value on the target criterion. That is, if we recognize one of the cities, we ought to infer that it is the larger. Research suggests this is indeed what people usually do (Todd and Gigerenzer 2007).

The second is Take the Best (TTB); "To infer which of two alternatives has the higher value (a) search through cues in order of validity, (b) stop search as soon as a cue discriminates, and (c) choose the alternative this cue favors" (Todd and Gigerenzer 2007, p. 168). That is, we consider the available cues (airports, tourism, industry, universities, etc.) sequentially, in order according to the degree to which they correlate with the population size, stop at the first cue that discriminates and infer that the city favored by that cue
is the largest. There is ample empirical evidence from behavioral studies that precise models of such heuristics better predict the subjects behavior in different task settings than optimizing models (Gigerenzer and Brighton 2009).

These heuristics typically operate upon internal representations. Construed as algorithmic process models, all that the algorithm requires is supplied by either the contents of our memory (i.e. cues) or by effects produced by their recall (i.e. perceived recognition). Thus, here the environment does not play any direct role in the execution of reasoning. So, apart from those peripheral cases of epistemic action, the bulk of BR's account of the execution of reasoning can be accommodated by internalism.

\section*{The Assessment of reasoning}

Defenders of BR do not merely claim that our reasoning often relies on heuristics; they also argue that this is frequently for the better. One of the major findings of BR is that often, for a given environment, fast and frugal heuristics outperform more costly strategies in terms of accuracy (Gigerenzer and Brighton 2009). Hence the claim that heuristics are rational; only that their rationality is ecological. Ecological rationality defines the rationality of heuristics by the match between internal processing and the environment. But what is meant by environment, here?

First, there are always some constraints on the strategy being applicable. In the case of the recognition heuristic, one of the items must be recognized while the others are not. If this obtains, there are further features of the environment that will determine the performance of the heuristic. Plainly, the recognition heuristic will perform well if and only if the target criterion correlates with the recognition of the item. This is called recognition validity, and it constitutes the relevant structure of the environment for the performance of the heuristic.

The necessary condition for the applicability of TTB is that we have different available cues and we can rank them in order of validity. Then, the main structural properties of the environment that determine its performance are that (i) the more correlated the available cues are, the less it pays to take them all into account (and accordingly, TTB performs as good as or better than optimizing strategies that demand taking all cues into account); (ii) the more the cue validities vary, the more it pays to use a strategy like TTB; and (iii) the smaller the learning sample, the better lexicographic strategies like TTB pay.

What is relevant and needs to be specified in order to evaluate reasoning strategies is the structure of the available information as defined by properties like cue redundancy, variability of cue validities, or size of learning samples. That is the environment that most heuristics exploit.

Reasoning performance is determined by the match between strategies and informational features of the environment that determine the strategy's relative success. Thus, the locus of ecological rationality is not only the internally processed algorithm, but the internal-algorithm-in-specific-environment complex. In this sense, the
environment does play a normative role in reasoning according to the BR program.

At this point, the internalist might object, "Fine, agents implement reasoning algorithms that operate upon stored information, and then it turns out that the performance of these algorithms is dependent upon how things are in the world. So what? This doesn't necessitate CE. There is no close internal-environmental resource integration here." However, this is not the whole story. Some of the features of the environment that play a normative role are not simply encountered out there. They are the product of the agent's ongoing coupling with the environment. Consider the role of uncertainty. As Gigerenzer has argued: "the degree of uncertainty reflects the environment (ontic uncertainty) as well as the mind's limited understanding (epistemic uncertainty); hence, the degree of uncertainty is located in the system mind-environment ... redundancy and variability of cues depend both on what is in the physical environment and on humans who select certain cues and not others, which can result in more or less redundant cues. Physicians, for instance, tend to look at redundant cues, but they could choose to analyze independent cues and thus in part create their environment" (Gigerenzer and Sturm 2012, p. 257).

A narrowly internalistic interpretation of BR takes heuristics as one-shot games. But in order to understand how rationality emerges from the use of such strategies, it is of much importance to consider the way in which the agents' behavior shapes the environment that shape the performance of their available reasoning strategies. Complementarily, the relative performance of a heuristic does affect its actual occurrence, albeit in an unexplained way. Part of the selection of a strategy operates upon internal representations, as memory constrains which heuristics can be applied. A second factor driving the selection of strategies is reinforcement by feedback. Still, beyond memory constrains and in absence of reinforcement, there is evidence that people rely on heuristics when they face those environments in which doing so pays off and not otherwise -i.e. relying on TTB when the validity of the available cues varies highly but not otherwise (Gigerenzer and Brighton 2009). This suggests that people are sensitive to the structure of the environment that determines the performance of heuristics.

The internalist can account for part of this story. It can explain how the heuristic strategy that the agent happens to apply is derived from a given structure of mental representations. It can also evaluate the performance of a strategy given a particular environment. It can even make inventories of heuristic-environment successful pairings. But, when it comes to deeper issues, such as why and how the agent chooses right and correctly applies a heuristic, or how and why certain environments have come to afford simple strategies, or how we have come to have such proficient heuristic tools, internalism seems to fall short. It faces the threat of falling into the "just happens" stance.

The emergence of normativity that results from the interaction between agent and environment is hardly the result of (only) internal computations. Whatever a much
needed further investigation into these frontiers of theorizing delivers, we contend that, at least for the large domain of uncertain reasoning, it is highly plausible that normativity is due to the ongoing process of back-and-forth manipulation (of the environment) and transformation (by the environment) that CE aims to unravel.

\section*{The Enhancement of Reasoning}

Next to the descriptive and the normative, BR also has a prescriptive dimension. Insofar as reasoning performance has to be assessed by the match between internal processing and environment (for some limits, see Sturm 2012), a dual perspective opens with regard to the enhancement of reasoning: we can either change what goes on within our heads, or change the environment. The development of both these prescriptive stances highlights a different point of connection between BR and CE .
A good example of the first line of prescription is fast and frugal decision trees for medical diagnoses. These are sequential trees designed for a very specific situation, like deciding whether a patient in the ER requires immediate attention facing a heart attack. They work as enacted protocols that the physicians must blindly follow. Interestingly, fast and frugal trees do not directly modify doctor's internal capacity of processing. From the doctor's point of view, they are an environmental resource, a cultural artifact they engage with in repeatable cognitive practices. It is part of the setting of the ER, upon which coordination between doctors is optimized. And indeed it has a transformative effect in doctor's cognitive behavior -i.e. leading their attention to certain cues while ignoring others.
The second path for reasoning enhancement that BR pursues is typically illustrated by drawing attention to the alarming lack of competence of physicians to assess risk probabilities (Gigerenzer 2008). Several studies show that when given information in percentages, physicians often perform poorly. They could be further instructed not to neglect base rates. But, as Gigerenzer puts it, "in the ecological view, thinking does not happen simply in the mind, but in interaction between the mind and its environment. This opens up a second and more efficient way to solve the problem: to change the environment" (2008, p. 17). The proposed prescription is to provide information in natural non-normalized frequencies (that is, in terms of " \(n\) out of every 1,000 people have disease \(x\). Of these \(n\) people, \(m\) will have a positive test, etc.").
Interestingly, "the relevant part of the environment is the representation of the information, because the representation does part of the Bayesian computation" (Gigerenzer 2008, p. 18). The object of manipulation of these change-theenvironment proposals is neither the material environment that plays a role in the execution of heuristics nor the information-structural properties of the environment that play a pervasive role in assessing the performance of heuristics. Instead, they point to symbolic and representational resources, which are thereby assumed to play a computational role. That is, cognitive practices carried upon the manipulation of symbolic structures. This
still squares perfectly with the work done by Menary (2010), Sterelny (2010) and Sutton (2010) on cognitive scaffolding, and the engagement with representational tools.

\section*{Conclusion}

Although Clark's characterization of BR is misguided, it is still true that BR's account of reasoning does not fit much with the old-school notion of extended cognition. Its processes and mechanisms do not involve agents in stable and durable couplings with technological resources they trust. But, pace Clark, this does not mean BR is a dead end for CE, since CE needs not be bounded to those cases in which the environmental resources constitute a proper part of cognition. BR's environmentalism does not consist in any ontological claim about the location and boundaries of the mind. It is a claim about explanatory variables and about normative criteria for understanding and improving human rationality. As such, it is best accommodated by the notion of scaffolded cognition.

From this moderate take on CE, we have shed light on the appeals to the environment made throughout the \(B R\) literature. We distinguished the role played by the physical environment in online heuristics qua epistemic actions, where we find instances of reasoning carried upon processes that functionally span through certain physical elements of the environment. But we also acknowledged that the execution of heuristics is typically decoupled from the real, mind-independent environment, and the account offered here by BR still fits with what a traditional internalist account would claim. Then we focused on the normative role played by the informational structure of the environment in determining heuristic performance, and on the role played by symbolic representations in the enhancement of reasoning performance. We believe that this conceptual clarification is quite a beneficial outcome of the whole debate for the development of the BR program.

Beyond that, we also pointed to some blind spots of internalism. The internalist can accommodate the actual processing of the heuristic, but it can hardly sustain the picture of rationality that emerges from the use of heuristics. Its narrow scope leaves important parts of the story on the dark, such as the emergence of the so-called "informational environment" and the co-enhancement process between agents and environments by which we can explain why we use the strategies we use and why is it that these happen to succeed beyond luck.

Those issues require further research that should be focused on the ongoing dynamic of interaction between agents and environments, and we contend that a moderate, scaffold externalism is the most fertile framework for this. In turn, we take these considerations to vindicate our take on CE, which acknowledges and incorporates the heterogeneity of cognitive engagements with the environment, and needs not be too much drawn by ontological claims, but rather to prove its explanatory value for cognitive science.

\section*{References}

Brighton, H., \& Todd, P. (2009). Situating Rationality. In M. Aydede \& P. Robbins (Eds.), Handbook of Situated Cognition. Cambridge: Cambridge University Press.
Clark, A. (1998). Being there: Putting Brain, Body and World together again. Cambridge, MA: MIT Press.
Clark, A. (2001). Reasons, Robots, and the Extended Mind. Mind and Language, 16, 121-145.
Clark, A. (2003). Artificial intelligence and the many faces of reason. In S. Stich \& T. Warfield (Eds.), The Blackwell guide to philosophy of mind. Oxford: Blackwell.
Clark, A. (2010). Memento's revenge: The extended mind extended. In R. Menary (Ed.), The extended mind. Cambridge: MIT Press.
Clark, A. \& Chalmers, D. (1998). The extended mind. Analysis, 58, 7-19.
Gigerenzer, G. (2007). Gut Feelings: The intelligence of the unconscious. London: Viking.
Gigerenzer, G. (2008). Rationality for Mortals. Oxford: Oxford University Press.
Gigerenzer, G. \& Brighton, H. (2009). Homo Heuristicus: Why Biased Minds Make Better Inferences. Topics in Cognitive Science, 1, 107-143.
Gigerenzer, G. \& Sturm, T. (2010). How (far) can rationality be naturalized? Synthese, 187, 243-268.
Hutchins, E. (1995). Cognition in the wild. Cambridge, MA: MIT Press.
Hutchinson, J. \& Gigerenzer, G. (2005). Simple heuristics and rules of thumb. Behavioral Processes, 69, 97-124.
Menary, R. (2007). Cognitive integration: Mind and cognition unbounded. Basingstoke: Palgrave Macmillan.
Menary, R. (2010). Dimensions of mind. Phenomenology and the Cognitive Science, 9, 561-578.
Sterelny, K. (2010). Minds: extended or scaffolded. Phenomenology and the Cognitive Science, 9, 465-481.
Sutton, J. (2010). Exograms and interdisciplinary: history, the extended mind and the civilizing process. In R. Menary (Ed.), The extended mind. Cambridge: MIT Press.
Sutton, J., Harris, C., Keil, P. \& Barnier, A. (2010). The psychology of memory, extended cognition and socially distributed remembering. Phenomenology and the Cognitive Science, 9, 521-560.
Todd, P. \& Gigerenzer, G. (2007). Environments that make us smart. Current Directions in Psychological Science, 16, 167-171.
Wilson, R. (2010). Meaning making and the mind of the externalist. In R. Menary (Ed.), The extended mind. Cambridge: MIT Press.
Wilson, R. \& Clark, A. (2009). How to Situate Cognition: Letting Nature Take its Course. In M. Aydede \& P. Robbins (Eds.), The Cambridge Handbook of Situated Cognition. Cambridge: Cambridge University Press.

\title{
The Role of Visual Coherence in Graphical Passwords
}

\author{
Ülkü Arslan Aydın \({ }^{1}\) (ulku.arslan@gmail.com) Cengiz Acartürk \({ }^{1}\) (acarturk@metu.edu.tr) \\ Kürşat Çağıltay \({ }^{2}\) (kursat@metu.edu.tr) \\ \({ }^{1}\) Cognitive Science, Informatics Institute \\ \({ }^{2}\) Computer Education and Instructional Technology \\ Middle East Technical University, 06800, Ankara, Turkey
}

\begin{abstract}
Graphical password is an alternative method of authentication to alphanumerical passwords. From the perspective of research on human memory, it is yet another novel technology that introduces challenges on human memory components. In this study, we aim to investigate the previous findings in human visual memory in the domain of graphical passwords by analyzing the role of visual coherence in passwords. The results of an experimental study reveal that in terms of memorability, coherent images are better candidates as graphical password images than jumbled images.
\end{abstract}

Keywords: Graphical passwords; visual coherence, visual working memory, eye tracking.

\section*{Knowledge-Based Authentication Systems}

The extended use of human computer interfaces in the past few decades has introduced several challenges on users' working memory. One such challenge is the requirement to memorize numerous passwords for security authentication. From the viewpoint of information security, user access to a security system is granted in three phases: identification, authentication and authorization (Figure 1). After identification, the user supplies the proof of her/his identity in the authentication phase. The proof of identity is usually accomplished by employing methods such as using a smartcard (token-based authentication), using biometric information such as fingerprints (biometric-based authentication), or entering an alphanumeric or a graphical password (knowledge-based authentication).


Figure 1: A taxonomy of authentication methods in information security systems.

Recently, knowledge-based authentication methods-in particular, text-based, alphanumerical passwords-are largely used for information access in information security
systems (Herley et al., 2009). An alternative knowledgebased authentication method, namely graphical passwords, has been recently gaining an increased use.

\section*{Graphical Passwords as an Alternative Method to Alphanumerical Passwords}

Graphical passwords were developed to overcome some of the security issues involved in the use of alphanumeric passwords (Dunphy et al., 2008). Graphical passwords are of different types, such as recall-based, recognition-based, and click-based (Figure 2). In a click-based graphical password system, a pixel-based image acts as a cue for activating user's memory. When creating a password in a click-based system, the user selects a sequence of number of (e.g., four or five) points on the presented image. After then, to login the system, the user reselects the points on the image in the same order by clicking on (or near to) them (Blonder, 1996; Wiedenbeck et al., 2005; Chiasson et al., 2007; Chiasson, 2008).


Figure 2: Sample graphical passwords (Dhamija \& Perrig, 2000, Lashkari et al., 2009; Valentine, 1999).

From the end-user's point of view, the major motivation for the development of graphical passwords was to take the advantage of picture memorability over text while maintaining security (Wiedenbeck, et al., 2005), thus providing a solution to the security-usability dilemma. \({ }^{1}\) The motivation for the development of graphical passwords finds its roots in early studies in cognitive psychology research, which revealed that humans have a tendency to

\footnotetext{
\({ }^{1}\) The security- usability dilemma refers to the observation that "passwords are often either memorable-but-insecure or secure-but-difficult-to-remember" Chiasson, 2008, p. 3. Graphical passwords as a solution to the dilemma are beyond the scope of this study.
}
remember images longer and better than words (cf. the picture superiority effect, Nelson et al., 1976). Accordingly, images are usually expected to be "easier to remember and more secure than words" (e.g., Cranor, \& Garfinkel, 2005; Kirkpatrick, 2002; Suo, Zhu, \& Owen, 2005), thus leading to memorability advantages over alphanumerical passwords.

In addition to offering a more memorable solution for security system authorization, graphical passwords provide a naturalistic environment for research on visual memory in daily life tasks. Although the focus of research has been the security-usability dilemma from an information security point of view, there are many aspects that need further investigation from the perspective of cognitive science, such as the identification of the circumstances under which graphical passwords achieve better memorability. One such factor is visual coherence, as described below.

\section*{Visual Coherence in Graphical Passwords}

Two major aspects of binding of objects in visual working memory are the binding of objects to perceptual features, such as color, shape and orientation, and the binding of objects to locations (Hollingworth \& Rasmussen, 2010). In visual cognition, the concept of coherence has been studied by Biederman (1972) and Biederman, Glass and Stacy (1973), leading to research results which showed that the objects were recognized and identified more efficiently and quickly when the scene image was presented coherent rather than jumbled. \({ }^{2}\) Mandler and colleagues have shown that the presence of a coherent background scene improves memory for both the location and the perceptual features of the object in the scene (e.g., Mandler \& Parker, 1976; Mandler \& Ritchey, 1977; Hollingworth, 2009). The facilitating effect of context in memory retrieval has been observed in both short-term time scale and long-term time scale (see Brady et al., 2011 for a review) (Brockmole et al., 2006; Foulsham et al., 2011). Those findings in visual working memory research suggest that visual coherence in graphical password images would improve memory for graphical passwords. In other words, when used as a graphical password image, a coherent image may reveal advantages over jumbled images. To test this hypothesis, we conducted an experimental study, in which the participants were shown how to create a click-based graphical password and how to login with the password, as described in the following section.

\section*{Experiment}

In the practice session of the experiment, participants were guided by on-screen instructions about how to create a password. During the experiment, participants were presented a visual distraction task and then they were asked to login once. They were asked, however, to click on a black

\footnotetext{
\({ }^{2}\) The jumbled image was created by dividing the coherent image into multiple sections and manipulating the relative positions of the sections without rotating them.
}
screen to login, instead of the previously presented graphical password image. This was the end of the first session. In the second session, three days after the first session, they were asked to login by using their password, again on the black screen. In both sessions, participants' login success and time were recorded. Participants' eye movements were recorded by a 50 Hz . non-intrusive eye tracker, integrated into 17 '" TFT monitor. \({ }^{3}\) The experiment was conducted in an office environment, with a developed application which simulated the interfaces of operating system that the participants were already familiar with. Overall, the experimental setting provided a relatively naturalistic environmental setting.

\section*{Participants, Materials and Design}

Sixty-three participants (29 females, 34 males \(\mathrm{M}=32.1\); \(\mathrm{SD}=0.73\) ) participated in the experiment. All of the participants were employees at a governmental institution and the participation in the study was voluntary. The participants were divided into two groups, according to the type of the graphical password image they were presented in the password creation phase: (1) a coherent image or (2) a jumbled image. Each group was further divided into two groups according to the type of the image presented when participant failed to login on a black screen.: (1) the same image as the image presented in the password creating phase or (2) a shuffled version of the previous image. The base image for the graphical password was a high resolution (2362*2362) image taken in a professional setting. The image was converted into gray scale to reduce visual saliency effects due to color contrast. This image was used as the graphical password for the coherent-image group participants (henceforth, the coherent group). The jumbled image, which was used as the graphical password image for the jumbled-image group participants (henceforth, the jumbled group), was produced out of the coherent image by randomly jumbling the pieces of the coherent image, in the form of a \(3 x 3\) grid (Figure 3). There was at least one identifiable object in each cell of the grid.


Figure 3: The images \((600 * 600)\) for the graphical password in the coherent group (left) and the jumbled group (right).

\footnotetext{
\({ }^{3}\) The participants were seated at a distance of approximately 60 cm to the monitor. Spatial resolution and accuracy of the eye tracker was about \(0.25^{\circ}\) and \(0.50^{\circ}\) degrees respectively.
}

The participants were instructed to choose passwords which they could remember but that would be difficult for others to guess. They were guided by the instruction screens. The experiment consisted of two sessions. The first session was divided into five phases: practice, password generation, questionnaires, mental rotation task and login. In the practice session, the participant was shown how to create a graphical password and how to login with the selected graphical password. After the practice session, the participants picked their passwords by clicking four clickpoints on the provided image. After then, they filled out a demographic questionnaire and a usability questionnaire. A 30 second mental rotation task was then administered to disrupt visual memory. In the last phase of the first session, the participants were asked to use the selected password to login the system. The second session was a login session only; it was administered three days after the first session.

\section*{Results}

In the last phase of the first session, the participants were asked to login the system by clicking on a black grid screen, without the graphical password image on the screen. This screen consisted of nine black squares in the form of a grid. The motivation for using the black screen was to investigate participants' strategy for choosing the password items. If the participant chose password just by memorizing object properties, without memorizing the spatial locations, s/he would not be able to log in without seeing the graphical password image. This was a surprise task for the participants because they were not informed about the black screen beforehand. The results showed that, however, the participants achieved a very high success login ratio on the black screen: Fifty-seven of sixty-three participants were able to login on the black screen, before being presented the graphical image (i.e., in the first, the second and the third attempt). This finding suggests that the participants remembered very well the locations of the click points in the first session. The results also suggested that a comparative analysis between the coherent-group participants (who were presented a coherent image as the graphical password) and the jumbled-group participants (who were presented a jumbled image as the graphical password) would be possible, because the results were similar between the pairgroups and the further division according to the type of the image presented at the login phase (i.e., shuffled vs. same) was no more necessary. Accordingly, the analyses were performed in terms of the measures below
- The time to login, create and confirm the password
- Eye movement parameters (fixation count, duration, Levenshtein distance) and visual saliency
- Password creation strategies

All the analyses were performed on participants' performance on the black grid screen in the first session (i.e., the login test in the same day) and in the second session (i.e., the login test three days after the first session). Additional analyses were also reported below, on visual saliency and on answers to questionnaires. Overall, the
results suggested that coherent-group participants exhibited better memory performance compared to the jumbled-group participants, as presented below.

\section*{Login Success}

The participants were allowed to try to login three times on the black grid screen. We performed a comparative analysis for the login success of the 56 participants in the first login attempt only. A three-way loglinear analysis (Login Success x Session x Group Type) produced a final model that retained login success and group type effects. This indicated that the interaction between login success and group type was significant, independently from the session. \(\chi^{2}(1)=\) \(5.20, p=.02\) (Figure 4). Based on the odds ratio, the odds of success in the first attempt was 2.98 times higher for the coherent-group than the jumbled-group participants.


Figure 4: Login success of the participants in the first try (the numbers show the success and the failure ratios of the participants between 0 and 1)

The overall success ratio on the black screen, including the further attempts (up to three), revealed a similar finding, (i.e. the interaction between login success and group type was significant, independently from the session, \(\chi^{2}(1)=\) 4.96, \(p=.04\) ), showing that coherent-group participants were more successful to login than the jumbled-group participants in the overall login attempts on the black screen.

\section*{Login Duration}

The participants spent time to create the password and then to login on the first day of the experiment (i.e., the first session). In the first session, no difference was observed between the jumbled-group ( \(M=25.2\) seconds, \(S D=15.7\) ) and the coherent-group participants \((M=22.4 \mathrm{~s}, S D=11.6)\) in creating the password, \(t(61)=0.81, p=.42, r=.10\). Moreover, the time to login was not different between the jumbled-group ( \(M=9.75 \mathrm{~s}, S D=6.32\) ) and the coherentgroup participants ( \(M=7.85 \mathrm{~s}, S D=3.82\) ). Although the participants spent approximately the same time to login between the first session ( \(M=8.79 \mathrm{~s}, S D=5.25\) ) and the second session ( \(M=8.56 \mathrm{~s}, S D=5.69\) ), the time spent to login in the second session was different between the jumbled-group ( \(M=10.1 \mathrm{~s}, S D=7.19\) ) and the coherent-
group participants \((M=7.07 \mathrm{~s}, S D=3.19), t(57)=2.02, p=\) .048, with a small effect size of \(r=.26\). To sum up, the analysis of login durations showed that, in the second session of the experiment which was conducted three days after the first phase, the coherent-image group spent less time to login compared to the jumbled-group participants.

\section*{Fixation Counts}

In this study, the term fixation count is used for describing the number of fixations on the black grid screen. The fixation counts were analyzed for a comparison between the jumbled group and the coherent group. The results were similar to the results obtained for login duration: there was no significant difference between the jumbled group ( \(M=\) 18.3, \(S D=12.3\) ) and the coherent group ( \(M=14.4, S D=\) 8.83) in the first session. In the second session, however, the difference in fixation counts between the jumbled group ( \(M\) \(=18.6, S D=16.6\) ) and the coherent group ( \(M=11.1, S D=\) 6.57) was significant, \(t(56)=-2.00, p=.05\), with an effect size of \(r=.26\). There was also a significant main effect of the session in fixation counts, \(F(1,56)=10.16, p=.002\), showing that the participants produced more fixation counts in the first session than they did in the second session. As the final step of the fixation count analysis, we investigated whether each fixation location belonged to the password (i.e., a pass item) or it did not belong to the password (i.e., a non-pass item). The participants in both groups spent more fixations on their pass items than their non-pass items, both in the first session, \(F(1,61)=79.9, p=.00\), and in the second session, \(F(1,56)=111.3, p=.00\). Moreover, in the second session, the coherent group spent less fixations on the non-pass items ( \(M=0.47, S D=0.51\) ) compared to the jumbled group ( \(M=1.63, S D=2.07\) ), \(t(56)=3.17, p=.00\). To sum up, in the second session, the jumbled group produced more frequent fixations compared to the coherent group. Moreover, the coherent group focused more efficiently on their pass items compared to jumbled group, who were focusing on non-pass items as well as pass items.

\section*{Fixation Durations}

The term fixation duration is used in this study for the mean duration of single fixations on the black grid screen. There was no significant difference between the coherent group ( \(M=429.4 \mathrm{~ms}, S D=95.2\) ) and the jumbled group ( \(M=\) \(433.1 \mathrm{~ms}, S D=145.7\) ) in the first session. On the other hand, in the second session, the participants in the jumbledimage group had shorter mean fixation duration on the black \(\operatorname{grid}(M=453 \mathrm{~ms}, S D=171)\) than the participants in the coherent-image group ( \(M=496 \mathrm{~ms}, S D=125\) ). This difference was significant \(t(56)=-2.12, p<.04\) and it did represent small-sized effect \(r=.27\). As the final step of the mean fixation duration analysis, we investigated whether each fixation location belonged to a pass item or it belonged to a non-pass item. The participants in both groups made longer fixations on their pass items than their non-pass items, both in the first session, \(F(1,56)=30.5, p=.00\), and in the second session, \(F(1,56)=50.8, p=.00\). Moreover, in
the second session, mean fixation duration on pass items were similar for both the coherent group ( \(M=562 \mathrm{~ms}, S D=\) 259.8 ) and the jumbled group ( \(M=486 \mathrm{~ms}, S D=152\) ). On the other hand, the coherent group made shorter fixations on non-pass items ( \(M=228 \mathrm{~ms}, S D=173\) ) compared to the jumbled group ( \(M=373 \mathrm{~ms}, S D=250\) ), \(t(56)=3.33, p=.00\) in the second session. To sum up, the analysis of fixation durations revealed that in contrast to the similarities between the two groups in the first session, the jumbled group exhibited shorter fixations compared to the coherent group in the second session. Moreover, in the second session, the coherent group exhibited shorter fixations on the non-pass items.

\section*{Levenshtein Distance (LD)}

The Levenshtein Distance (LD) is a specific application of the string editing analysis, where the distribution of fixations on certain locations (in our case, the grid cells) is coded by letters. The letter strings of each participant are then compared with the password of the participant for similarity. The LD defines the number of modifications (i.e., insertions and deletions) on one string that is necessary to make it the same as the other.

In our study, LD was used a specification of the similarity between the two groups of participants. The results of the LD analysis revealed that, the participants in the first session ( \(M=8.22, S D=7.18\) ) exhibited longer LD compared to the participants in the second session ( \(M=\) \(6.97, S D=8.24), F(1,56)=9.69, p=.00\). In addition, in the first session, no significant difference was obtained in LD between the jumbled-group participants \((M=9.35, S D=\) 7.96) and the coherent-group participants ( \(M=7.08, S D=\) 6.25 ). In the second session, the difference between the jumbled-group participants ( \(M=9.64, S D=10.2\) ) and the coherent-group participants ( \(M=4.30, S D=4.29\) ) was significant, \(t(56)=2.57, p=.01\), with a medium-size effect of \(r=.32\), indicating more search effort in the jumbled-group participants compared to the coherent-group participants.

\section*{Visual Saliency Analysis}

The saliency maps of the coherent image and the jumbled image were computed by using the algorithm provided by Walther and Koch (2006). \({ }^{4}\) Based on this, the percentage distribution of the saliency of each cell in the \(3 \times 3\) grid was calculated. The resulting distribution provided the relative saliency distribution over the password image. The distribution of participants' pass-items was also calculated by analyzing the selected graphical passwords in the experiment. For the jumbled image, no relation was obtained between the saliency values and the ratio of being pass-item, \(\mathrm{r}=-0.11, p=.34\) (Figure 5).

\footnotetext{
\({ }^{4}\) SaliencyToolbox library, http://www.saliencytoolbox.net
}


Figure 5: The saliency distribution in the jumbled graphical password (left) and the percentage of being selected as a pass-item by the participants (right).

A similar analysis was conducted for the coherent image. Again, no significant relation was obtained between the saliency values and the ratio of being pass-item, \(r=0.30, p\) \(=.47\) (Figure 6).


Figure 6: The saliency distribution in the coherent graphical password (left) and the percentage of being selected as a pass-item by the participants (right).

In summary, visual saliency analyses revealed no useful results to account eye movement behavior characteristics in graphical passwords, in line with the findings in relevant domains of visual cognition (Tatler \& Vincent, 2009)

\section*{The Analysis of Password Creation Strategies}

After the participants created the password in the first session of the experiment, they filled in a questionnaire about their strategy for creating the password. Four choices were presented to the participants: (a) I created a pattern that looked like an L-shape or a V-shape (create pattern), (b) I memorized the names of the objects in the password (object recognition) (c) The objects I selected had common visual features (e.g., color, shape) or functional (e.g., cutting) features, (similar features) (d) I created a story (story). The participants were allowed to make multiple choices. The participants' answers were analyzed in terms of the relation between the group type, the session, the adopted strategy and the login success. A main effect was obtained for strategy, \(\chi^{2}(1)=91.7, p=.00\), indicating that there was a significant difference between the adopted strategies. Pattern creation was the most preferred strategy (33 out of 77). Furthermore, the interaction between strategy and group type was significant, \(\chi 2(3)=8.61, p=.03\), indicating that adopted strategy was significantly affected
by group type. On the other hand, no relationship was observed between the selected strategy and the login success, \(\chi 2(3)=4.46, p=.21\).

\section*{Discussion}

The results of the experimental investigation showed that a high majority of the participants (57 of 63 participants) in the first session was able to login the system by clicking on a black screen. This finding indicates that the participants memorized the locations of the pass-items in the graphical password. The rest of the analyses were conducted on those 57 participants. Overall, the coherent-group participants, who were presented a coherent image as the graphical password, achieved better memory performance compared to the jumbled-group participants, who were presented a jumbled image as the graphical password. This finding was obtained in terms of a set of measures, including login success, login time, eye movement parameters and visual saliency, as well as password creating strategies. The analysis of login success showed that the coherent group exhibited higher login success compared to the jumbled group, independent of the session. This difference was obtained both in the first attempt to login and in the analysis of all attempts to login (the participants were allowed to try three times to login). The performance difference between the groups was evident, for some of the measures, in the second session of the experiment, which was conducted three days after the first session. For example, the analysis of the login duration showed that in the first session, there was no difference between the groups. The difference, however, was significant in the second session in favor of the coherent group: the coherent-group participants were able to login in shorter time. These findings have implications for end-users, as well as password system designers. The facilitating role of image coherence suggests that users should be encouraged to select coherent images for graphical passwords rather than jumbled images.

The analysis of fixation counts revealed two major findings: not in the first session but in the second session, the jumbled group fixated more frequently on the black screen compared to the coherent group. Moreover, in the second session, the coherent group spent less fixations on non-pass items, thus exhibiting a higher memory efficiency for the pass items. The analysis of fixation durations revealed that in the second session, the mean fixations of the jumbled group were shorter than the mean fixations of the coherent group. Shorter fixation durations may be indicators for visual search (compared to normal scene viewing, Rayner, 1998). Jumbled-group participants’ higher effort to find the pass-items, as well as the longer Levenshtein distance exhibited by the jumbled group, provide support for our interpretation that they had more difficulty in remembering the pass items compared to the coherentimage group participants. Finally, we observed no relationship between likelihood of the selected pass items and their visual saliency. This may be due to participants' strategies in selecting the pass items. The analysis of the
strategies, however, returned no significant relationship between the selected strategy and the login success, though higher preference of certain strategies (in particular, pattern creation) by the participants over the others.

\section*{Conclusion and Future Work}

Coherence has been a research topic in relevant domains to human cognition. In linguistics, discourse coherence is described as constructing the continuity in context by constructing the meaning between the parts of the written text or spoken utterance (Wolf, 2005). A coherent discourse has comprehension advantages compared to an incoherent discourse. In visual cognition, the studies reveal an improved efficiency in object identification and memory in favor of coherent images. These findings reveal the importance of coherence for cognition in different modalities. The findings in the present study show that the coherence effect is also applicable to practical settings, in this case graphical passwords. The present study also shows that the advantages of visual coherence can be observed in various measures, including login success and duration, as well as eye movement parameters. Future studies will address extending the evaluation by additional eye tracking metrics, such as scan path ratio, the investigation of the role of specific memory components, and a more extensive analysis of users' strategies for creating graphical passwords.
Acknowledgments. Thanks the Capital Markets Board of Turkey for providing the setting for the experimental investigation. Thanks Deniz Zeyrek and our reviewers for their valuable comments and suggestions.

\section*{References}

Biederman, I. (1972). Perceiving real-world scenes. Science, 177, 77-80.
Biederman, I., Glass, A.L., \& Stacy, E.W. (1973). Searching for objects in real-world scenes. Journal of Experimental Psychology, 97(1), 22-27.
Blonder, G., (1996). United States Patent 5559961.
Brady, T.F., Talia, K., \& Alvarez, G.A. (2011). A review of visual memory capacity: Beyond individual items and toward structured representations. J. Vision, 11(5):4, 1-34.
Brockmole J.M., Castelhano, M.S., \& Henderson, J.M. (2006). Contextual cueing in naturalistic scenes: Global and local contexts. Journal of Experimental Psychology: Learning, Memory, and Cognition, 32(4), 699-706.
Chiasson, S. (2008). Usable autentication and click-based graphical passwords. Carleton University dissertation.
Chiasson, S., Biddle, R., \& Van Oorschot, P. (2007). A second look at the usability of click-based graphical passwords. Symposium on Usable Privacy and Security.
Cranor, L.F., \& Garfinkel, S. (2005). Security and usability: Designing secure systems that people can use. L. Cranor \& S. Garfinkel, (Eds.), O'Reilly.
Dhamija, R. \& Perrig, A. (2000). Déjà Vu: A user study using images for authentication. Proceedings of the 9th conference on USENIX Security Symposium.

Dunphy, P., Nicholson, J., \& Olivier, P. (2008). Securing passfaces for description. Proceedings of the 4th symposium on Usable privacy and security. ACM Press.
Foulsham, T., Alan, R. \& Kingstone, A. (2011). Scrambled eyes? Disrupting scene structure impedes focal processing and increases bottom-up guidance. Attention, Perception and Psychophysics, 73 (7), 2008-2025.
Herley, C., Van Oorschot, P., \& Patrick, A. (2009). Passwords: If we're so smart, why are we still using them? In Dingledine \&Golle (Eds.) Financial cryptography and data security (pp. 230-237). Springer.
Hollingworth, A. (2009). Two forms of scene memory guide visual search: Memory for scene context and memory for the binding of target object to scene location. Visual Cognition, 17(1), 273-291.
Hollingworth, A., \& Rasmussen, I.P. (2010). Binding objects to locations: The relationship between object files and visual working memory. Journal of Experimental Psychology: Human Percept. and Perf., 36(3), 543-564.
Kirkpatrick, E.A. (2002). An experimental study of memory. Psychological Review, 1(6), 602-609.
Lashkari, A.H., Saleh, R., Farmand, S., Zakaria, O. (2009). A wide range survey on recall based graphical user authentications algorithms based on ISO and attack patterns. International Journal of Computer Science and Information Security, 6(3), 17-25.
Mandler, J.M. \& Parker, R.E. (1976). Memory for descriptive and spatial information in complex pictures. Journal of Experimental Psychology: Human, Learning, and Memory, 2, 38-48.
Mandler, J.M., \& Ritchey, G.H. (1977). Long-term memory for pictures. Journal of Experimental Psychology: Human Learning \& Memory, 3, 386-396.
Nelson, D.L., Reed, U.S., \& Walling, J.R. (1976). Pictorial superiority effect. Journal of Experimental Psychology: Human Learning \& Memory, 2, 523-528
Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. Psychological Bulletin, 124(3), 372-422.
Suo, X., Zhu, Y., \& Owen, G.S. (2005). Graphical passwords: A survey. 21st Annual Computer Security Applications Conference, (pp. 463-472). IEEE.
Tatler, B. W., \& Vincent, B. T. (2009). The prominence of behavioural biases in eye guidance. Visual Cognition, 17, 1029-1054.
Valentine, T. (1999). An evaluation of the passface personal authentication system. Technical Report Goldsmith College University of London.
Walther, D., \& Koch, C. (2006), Modeling attention to salient proto-objects. Neural Networks, 19, 1395-1407.
Wiedenbeck, S., Waters, J., Birget, J.-C., Brodskiy, A., \& Memon, N. (2005). PassPoints: Design and longitudinal evaluation of a graphical password system. International Journal of Human-Computer Studies, 63(1-2), 102-127.
Wolf, F. (2005). Coherence in natural language: Data structures and applications. MIT Dissertation.

\title{
Conceptual change in proportional reasoning: Effects of collaboration, own / partner reasoning level and hypothesis testing
}

\author{
Christa S.C Asterhan (asterhan@huji.ac.il) \\ Baruch B. Schwarz (msschwar@mscc.huji.ac.il) \\ Noa Cohen-Eliyahu (noacoe@gmail.com) \\ School of Education, Hebrew University of Jerusalem, Mt. Scopus \\ Jerusalem, 91905 Israel
}

\begin{abstract}
Systematic research of instruction-based conceptual change in Mathematics and Science is characterized by examining the effectiveness of a particular instructional principle in isolation. It is suggested that the field could gain from studying how different instructional principles interact when they are combined. The goal of this research was to systematically study the combined effects of collaborative learning and hypothesis testing on cognitive growth. In a randomized experiment, 496 9th graders solved challenging tasks that required fully developed proportional reasoning. Half of them were given the opportunity to test their solutions. Based on individual pretests, each student was assigned to one of three competency levels (low, medium, high), and randomly assigned to either work alone or with a (low, medium, high) peer. The findings show that the effectiveness of hypothesis testing are conditioned by fine-grained differences in the contingencies between the target student's level of competence, the peer partner's level of competence and the feedback they receive from the objective testing device.
\end{abstract}

Most of the early research on cognitive growth through peer collaboration focused on the question of optimal dyad composition (e.g., Messer, Joiner, Loveridge, Light \& Littleton, 1993; Tudge \& Winterhoff, 1993). However, results have overall been inconclusive and research has largely been abandoned in favor of process-oriented investigations, such as peer dialogue (e.g., Asterhan \& Schwarz, 2007, 2009; Schwarz, Neuman \& Biezuner, 2000) or other instructional techniques to elicit cognitive conflict, such as collaborative hypothesis testing. (e.g., Howe, Tolmie, Duchak-Tanner \& Rattay, 2000; Howe, Tolmie \& Rodgers, 1992).Hypothesis testing tasks require learners to translate their conceptual knowledge into hypotheses and subject these to empirical evaluation. When disconfirmed, it may confront learners with compelling evidence that they should reconsider their prior understanding even when two learners agree on their predictions (e.g., Howe et al, 2000). Vice versa, when a prediction is confirmed, it validates the explanation that led to the prediction.

In this paper, we present findings from a new study that examines whether the effects of hypothesis testing techniques depend on dyad compositions. We predict that it is. First of all, hypothesis testing in collaborating dyads may create conflict in W-W dyads (two 'wrong' learners), and settle a social conflict between members W-R dyads (one 'right' and one 'wrong' learner), who each gave different predictions and explanation. The success of hypothesis testing in socio-cognitive conflict tasks, however, hinges on a careful design: only the correct explanation or strategy should lead to a confirmation. If not, the feedback may confirm an individual's naïve, incorrect conception.

If designed carefully, this can then lead to quite powerful learning opportunities: For instance, a 'wrong' (W) student that collaborates with a 'right' \((\mathrm{R})\) student will not only be exposed to a higher level of reasoning during the discussion phase, but will also receive empirical confirmation that this reasoning is correct. That is likely to be a quite powerful combination. Students in a \(\mathrm{W}_{\mathrm{x}}-\mathrm{W}_{\mathrm{x}}\) pair on the other hand, would be expected to reach quick agreement without much discussion, but shown wrong in the hypothesis testing phase, forcing them to generate a new, higher-level explanation for these findings all by themselves. Lastly, in \(\mathrm{W}_{\mathrm{x}}-\mathrm{W}_{\mathrm{y}}\) pairs the outcomes are likely to be contingent on the competency level of the particular student: A lower competency W student \(\left(\mathrm{W}_{1}\right)\) is likely to benefit more from interaction with a slightly more competent W student \(\left(\mathrm{W}_{2}\right)\) when there is no hypothesis testing than with it. The reason for this somewhat counterintuitive expectation is that if the \(\mathrm{W}_{1}\) student will be convinced by \(\mathrm{W}_{2}\) 's reasoning in the discussion phase, this solution will be proven wrong in the hypothesis testing phase. As a result, \(\mathrm{W}_{1}\) students may very well regress back to their prior level of reasoning and \(\mathrm{W}_{2}\) students may regress as well.

Very few studies have examined whether hypothesis testing techniques are more effective in collaborative or individual conditions. Two studies are particularly relevant to ours and are worth mentioning in further detail: The first is a study reported by Ellis, Klahr \& Siegler (1993) that sought to investigate the effects of feedback and collaboration on \(5^{\text {th }}\) graders' use of mathematical rules for decimal fractions. Each of the approximately 120 pupils in this study consistently used one of two incorrect mathematical rules that were equally wrong, but qualitatively different. They were assigned to either work alone or in \(\mathrm{W}_{\mathrm{x}}-\mathrm{W}_{\mathrm{y}}, \mathrm{W}_{\mathrm{x}}-\mathrm{W}_{\mathrm{x}}\) or \(\mathrm{W}_{\mathrm{y}}-\mathrm{W}_{\mathrm{y}}\) pairs. The results demonstrated that children who had the opportunity to collaborate with a partner were more likely to use a correct rule on a posttest than children who worked alone, but only if they were given feedback during the interaction as to whether their answers were correct or not. However, dyadic composition was not found to affect children's understanding on individual tests.

Tudge, Winterhoff and Hogan (1996) also investigated the effects of feedback (hypothesis testing) and dyad composition on early elementary school children's problem solving performance on a balance beams task ( \(N=83\) ). Children in this study either worked alone or with a partner who was equally, less, or more competent and either did or did not receive feedback on the correctness of their predictions. In direct conflict with the findings reported by Ellis et al, the presence of a partner was more effective than
working alone only when children did not receive feedback. When children received feedback, working alone was more effective than working with a partner. Similar to the Ellis et al findings, no differences were found between the different types of dyad compositions.

The findings from these two studies then lead to quite different predictions: Based on the Tudge et al findings, students may be expected to profit more from hypothesis testing when they work alone, whereas based on the Ellis et al study and findings reported by Howe et al students are expected to benefit particularly from the combination of hypothesis testing and collaboration and hypothesis testing.

The main aim of the present study is then to settle the disparate findings with regard to hypothesis testing and dyad composition in collaborative problem solving and address the following caveats in the literature. Moreover, none one of the above-mentioned studies systematically tested the effects of hypothesis testing for the full range of different dyad compositions that specifies the target student's and the partner's competence level. Finally, they did not control for nested effects of the individual within the dyad and reported findings may thus be overestimates.

The topic domain that was chosen for this study is proportional reasoning. Research suggests that students experience difficulty with proportional reasoning problems because they over-extend numerical equivalence concepts to proportional equivalence problems (e.g., Mix, Levine, \& Huttenlocher, 1999; Tourniaire \& Pulos, 1985). Sophisticated tests, such as the Blocks task, have been developed to serve both as instructional interventions as well as assessment tools (e.g., Schwarz \& Linchevski, 2007).

\section*{Method}

\section*{Participants}

Eight public junior high schools from the Jerusalem and Tel Aviv metropolitan areas in Israel agreed to participate in the study. The entire 9th grade population of each school (over 600 students) completed a screening (pretest) questionnaire, to assess each student's use of problem solving strategies. Students that did not complete the questionnaire, did not provide explanations for their answers or based their answers on superficial, visual features of the two target shapes only were excused from participation in the intervention phase (see Coding section for further details). The remaining 496 9th graders ( 301 boys, 195 girls) used either additive ( \(\mathrm{N}=196\) ), proto-proportional \((\mathrm{N}=194)\) or proportional \((\mathrm{N}=105)\) reasoning strategies and participated in the intervention stage of the study. Six students did not complete the post test ( 2 additive and 4 pre-proportional problem solvers, respectively).

\section*{Design}

Participating students within each classroom were randomly assigned to experimental condition within each group of initial level of proportional reasoning: additive (AddS), proto-proportional (ProtoS) and proportional (PropS) strategy. The basic experimental design was 2 (hypothesis testing / no hypothesis testing) * 2 (individual / dyadic work). The dyadic condition was furthermore subdivided into 5 different pairing options: AddS-AddS, AddS-ProtoS, AddS-PropS, ProtoS-PrepS and ProtoS-PropS. The entire study then included a total of 16 different experimental conditions (see Table 1 for a distribution of the participants according to conditions).

Table 1. The 16 experimental conditions of the study
\begin{tabular}{|c|c|c|c|}
\hline \multicolumn{2}{|l|}{Paring condition} & \multicolumn{2}{|l|}{Hypothesis testing condition} \\
\hline Dyad member 1 & Dyad member 2 & Without HT & With HT \\
\hline Adds & - & \(N=22\) & \(N=15\) \\
\hline Adds & AddS & \(N=40\) & \(N=40\) \\
\hline Adds & ProtoS & \(N=44\) & \(N=44\) \\
\hline Adds & PropS & \(N=36\) & \(N=34\) \\
\hline ProtoS & - & \(N=13\) & \(N=18\) \\
\hline ProtoS & ProtoS & \(N=42\) & \(N=40\) \\
\hline ProtoS & PropS & \(N=30\) & \(N=44\) \\
\hline PropS & - & \(N=15\) & \(N=19\) \\
\hline
\end{tabular}

\section*{Tools}

The task that was used for the screening, the posttest and the interaction phase is an adaptation of the Blocks task, originally developed by Harel, Behr, Lesh \& Post (1992). In any given trial in the current version of the Blocks test, students are shown 4 three-dimensional block constructions (blocks A, B, C and D), each made up of a number of bricks. The bricks in C and A are of identical color, and so are the bricks in shapes B and D. Students are told that the weight of each brick in shapes \(A\) and \(C\) is identical, and that the same is true for each brick in B and D . At each trial, students are given information about the relation between the two base block constructions A and B ( A is heavier than \(B, B\) is heavier than \(A\), or they are of equal weight). They are then asked to determine the relation between the two target blocks, C and D . They are given four different options to choose from ( C is heavier than \(\mathrm{D}, \mathrm{D}\) is heavier than C , they are of the same weight, or it is impossible to determine) and are asked to base their choice with appropriate explanations (see Figure 1 for an example).

5BlocksTaskTest. Individual student's proportional reasoning level at pre- and posttests was assessed with a pen-and-paper test compiled of five Block tasks of increasing difficulty, ranging from tasks that could be solved with any strategy correctly with any strategy (e.g.,


Figure 1. Example of a Blocks task item
Task 1) to tasks that could be only solved with S4 (task 4, 5).

Intervention tasks. The two items that were given during the intervention stage were not included in the 5BloksTaskTest and could only be solved correctly with proportional reasoning strategies (S4).

\section*{Coding procedures.}

Students' level of proportional reasoning was assessed with the help of a slightly adapted version of a coding scheme developed by Schwarz \& Linchevski (2007). Each written response to a test item (5 on pretest, 5 on posttest, and 2 during intervention task for each participant) was assigned to one of 3 different and mutually exclusive problem solving strategy categories, in ascending order of reasoning quality:

S2 (additive reasoning, grade: 2). The student takes into account the weight of a single brick in relation to the entire block, compares the target blocks to the base blocks. In this Strategy there is no multiplicative related also there is a hint to the right strategy. For example: If \(A\) and \(B\) have the same weigh then \(C\) and \(D\) have the same weight because we add to \(A\) and \(B 4\) bricks each to get \(C\) and \(D\).

S3 (proto-proportional reasoning, grade: 3). The explanation relates to all four blocks, but only refers to the nominal difference between the number of bricks of two blocks. Example: If blocks A and B have the same weight. But there is one more bricks in B, that mean one brick in \(A\) weights more than one brick in B. so 3 bricks that added from \(A\) to \(C\) are heavier than 3 bricks that added from \(B\) to \(D\). so \(C\) is heavier than \(D\).

S4 (full proportional reasoning, grade: 4). The explanation relates to all four blocks. This strategy is characterized by numerical calculation of the proportion between the four blocks. For example: The rate between \(C\) and \(A\) is \(24 / 10=2.4\) and the rate between \(D\) and \(B\) is 37/16 \(=2.3125\) so if they weight the same and \(A\) is multiplied in a bigger number to get \(C\) so \(C\) is heavier than \(D\).

Ten percent of the entire data set was coded by two independent raters, blind to condition. Inter-rater reliability was high, Cohen's \(\kappa=.925\). The highest strategy level a student used on the pretest version of the 5BlocksTaskTest formed the basis for assessing a student's initial level of proportional reasoning: S2 (S2 on each of the 5 pretest items), S3 (used S3 at least once, but not S4), S4 (used S4 at least once). Students that did not use at least S2 strategies on all five pretest items were excused from further
participation. Performance on pretests and posttests was calculated by the mean grade of the five tasks on each test.

\section*{Procedure}

All data collection and experimental interventions were completed locally in each of the 8 participating schools. Students participated in the following sequence of activities:

Stage 1: Assessment and selection. The 5BlocksTask test was administered in pen-and-paper format to all students in the participating 9th grade classes to assess their initial level of proportional reasoning and lasted between 25-40 min. Trained research assistants read aloud the instructions explaining the task. During each of the five Blocks tasks, the research assistants physically showed the 4 relevant constructions (A, B, C and D) for each task in the front of the classroom.

Stage 2: Intervention. Participating students were called to a separate room during regular school hours, in familiar rooms adjacent to participants' classrooms, either individually or in dyads, according to condition. Trained research assistants informed students that they were going to solve two additional tasks and repeated the Blocks task instructions. Students were shown the 4 physical block constructions during each task (A, B, C and D). Students in the dyadic condition were instructed to solve the tasks together. They were furthermore told that they did not have to reach consensus but that they should share ideas and explanation with each other before writing down a solution on one shared solution sheet. Students in the hypothesis condition additionally received the following instructions: "After writing down the solution you can test whether your solution is right or wrong by placing the two target constructions C and D on a scale. If you were wrong you may re-think [together] your solution and try to explain the outcome you received". The research assistant refrained from intervening, except to remind students of the instructions when this was needed.

Stage 3: Post-test assessment. The 5BlocksTaskTest was administered in pen-and-paper format in each classroom after all participating students had completed the intervention phase. All participating students completed the three stages in less than one month.

\section*{Results}

Analyses were conducted with a mixed model (SAS PROC MIXED) with random effects of dyad within condition and of individual within dyad and condition, on individual students' mean gains from pretest to posttest. Residuals were checked for each model separately and outliers ( \(z<-4\) or \(z>4\) ) were locally trimmed from a data set. In a few cases the kurtosis of a distribution was slightly greater than zero. When this was the case a separate analysis was conducted on the SQRT of the dependent variable (individual learning gains) and its outcomes compared to the model of its non-transformed counterpart. No differences were found in the overall pattern of results, and we therefore only report on the result from untransformed models only.

Table 3. Adjusted mean (and SE) learning gains for 'non-proportional' students by peer pairing and hypothesis testing condition, \(N=456\).
\begin{tabular}{lccccc}
\hline & \multicolumn{4}{c}{ Pairing condition } & \\
\cline { 2 - 5 } & Alone & \begin{tabular}{c} 
Same level W \\
partner
\end{tabular} & \begin{tabular}{c} 
Different level W \\
partner
\end{tabular} & \begin{tabular}{c} 
Proportional R \\
partner
\end{tabular} & Total \\
\hline Hypothesis testing & \(.20(.08)\) & \(.12(.06)\) & \(.16(.08)\) & \(.58(.06)\) & \(.26(.03)\) \\
Without hypothesis testing & \(.22(.07)\) & \(-.02(.06)\) & \(.22(.08)\) & \(.24(.07)\) & \(.16(.03)\) \\
Total & \(.22(.05)\) & \(.05(.04)\) & \(.19(.05)\) & \(.41(.05)\) & \\
\hline
\end{tabular}

\section*{Overall effects of collaboration and hypothesis testing on learning}

Table 2 presents the adjusted mean learning gains of the entire data set, according to pairing condition (working alone or in a dyad) and hypothesis testing condition (with or without weighing apparatus). A significant main effect was found for hypothesis testing, \(F(1,422)=5.10, p=.024\), with students in the hypothesis testing conditions showing larger cognitive gains ( \(M=.25, S E=.04\) ), compare to those who did not ( \(M=.13, S E=.04\) ). No main effect of collaborative condition was found, \(F(2,422)<1, n s\), and the two factors were not found to interact, \(F(2,422)=1.48\), \(n s\).

Table 2.Adjusted mean (and SE) learning gains for collaborative condition (dyadic or individual) and hypothesis testing condition (with or without weighing apparatus), \(N=490\).
\begin{tabular}{llll}
\hline & Individual & Dyadic & Total \\
\hline With HT & \(.22(.06)\) & \(.28(.03)\) & \(.25(.04)\) \\
Without HT & \(.17(.06)\) & \(.10(.03)\) & \(.13(.04)\) \\
Total & \(.19(.04)\) & \(.19(.02)\) & \\
\hline
\end{tabular}

Effect of collaborating with a 'proportional' or 'non-proportional' problem solver.
We then tested whether the lack of effect for collaboration on individual learning gains could be explained by differences between students who were paired with a peer that had employed proportional strategies and students who were paired with a non-proportional peer that (i.e., either additive or proto-proportional). As in the previous model, a main effect was found for hypothesis condition, \(\mathrm{F}(1,326)=\) \(10.40, \mathrm{p}=.001\). In addition, a main effect was found for pairing condition, \(\mathrm{F}(2,315)=6.86, \mathrm{p}=.001\). Post-hoc analyses (with Tukey-Kramer adjustments) showed that students that collaborated with a proportional peer had larger learning gains \((\mathrm{M}=.31, \mathrm{SE}=.04)\) than both students that collaborated with a non-proportional peer ( \(\mathrm{M}=.12\), SE \(=.03), \mathrm{t}(233)=3.70, \mathrm{p}<.001\), as well as those that worked alone \((\mathrm{M}=.19, \mathrm{SE}=.04)\), \(\mathrm{t}(348)=2.00, \mathrm{p}=.046\). No interaction between hypothesis testing and pairing condition was found, \(\mathrm{F}(2,315)=1.97\), ns.

\section*{Effects of dyadic pairing and hypothesis testing for non-proportional students}

Next, we explored the effects of pairing and hypothesis testing amongst 'non-proportional' students only, that is: those students who had not solved any of the five pretest tasks with a full-fledged algebraic strategy. We distinguished between the following four pairing options: working without a partner (alone), being paired with a nonproportional partner of the same strategy level (same level W partner), with a partner of a different non-proportional strategy level (different level W partner) or with a partner of a full proportional strategy level (proportional R partner). Table 3 presents the adjusted mean gain scores for each of these eight conditions.

Similar to the previous models, a main effect was found for hypothesis testing, \(F(1,239)=4.13, p=.043\), such that regardless of whom they were paired with, non-proportional students gained more in the weighing condition ( \(M=.26\), \(S E=.03\) ) than in the non-weighing condition ( \(M=.16, S E=\) .03). A main effect for pairing condition was also found, \(F\) \((2,239)=10.98, p<.001\). Post-hoc analyses (with TukeyKramer adjustments) showed that being paired with a proportional student ( \(M=.31, S E=.04\) ) resulted in larger learning gains than being paired with a same-level, 'nonproportional' peer ( \(p<.001\) ), with a different level, 'nonproportional' peer ( \(p=.016\) ) or working individually ( \(p=\) .040).

In addition, the effect of pairing among nonproportional students was also found to be dependent on hypothesis testing condition, \(F(2,225)=3.22, p=.024\). Judging from Table 3 there are two conditions that stand out in particular: The condition with hypothesis testing and a proportional partner for its comparatively high mean gain score ( \(M=.58, S E=.06\) ), and the condition no hypothesis testing / same-level non-proportional partner for its comparatively low mean gain score ( \(M=-.02, S E=.06\) ). Tukey-Kramer tests for multiple comparisons confirmed these impressions: When students were given the opportunity to test their predictions with a testing device, being paired with a 'proportional' peer indeed led to better learning gains compared to working with a same-level, 'non-proportional' peer ( \(p<.001\) ), with a different level, 'non-proportional' peer ( \(p=.001\) ) or individually ( \(p=.007\) ). There were no differences between being paired with a same-level partner, a different-level wrong partner or working alone. device, \(t(240)=3.53, p=.012\). Comparisons between weighing and non-weighing condition in the other three pairing conditions did not yield any significant differences. Thus, it seems that


Figure 2 The effect of pairing with a non-proportional peer and hypothesis testing on learning gains, for additive (1a) and proto-proportional (1b) problem solvers
for 'non-proportional' learners as a group, neither hypothesis testing nor the pairing with a 'proportional' peer by itself resulted in learning gains, but only the combination of the two. This is further supported by the finding from post-hoc comparisons that 'non-proportional' learners in the hypothesis testing condition who were paired with a 'proportional' partner had significantly higher gains scores than students in each of the other 7 conditions.

However, when 'non-proportional' learners did not have access to a hypothesis testing device, being paired with a 'proportional' student did not have any advantage over any of the other pairing conditions (all comparisons were \(n s\) ). Moreover, 'non-proportional' learners who are paired with a 'proportional' student gain more when they are given the opportunity to test their predictions with a hypothesis testing

Interestingly, when learners did not have access to the hypothesis testing device, 'non-proportional' students gained least when they were paired with a partner from the same level, and significantly less so than when working alone ( \(p=.009\) ), with a 'proportional' partner \((p=.005)\) or with a different level 'non-proportional' partner ( \(p=.012\) ). No differences were found between the latter three conditions.

\section*{Effects of \(W_{x}-W_{y}\) pairing and hypothesis testing for different types of 'non-proportional' students}

The findings reported above seem to indicate that for non-proportional students being paired with a differentlevel, non-proportional partner student ( \(\mathrm{W}_{\mathrm{x}}-\mathrm{W}_{\mathrm{y}}\) pairing) is only preferable when students do not have access to a hypothesis-testing device, but that there is no advantage to this pairing when they have the opportunity to test their
predictions. However, these findings disregard differences in the target student's initial strategy level. The effect of \(\mathrm{W}_{\mathrm{x}}-\mathrm{W}_{\mathrm{y}}\) pairing and hypothesis testing was then separately tested for students that were initially diagnosed as 'additive' problem solvers in the Blocks task and for those that were diagnosed as 'proto-proportional' problem solvers (see Method section).

Figure 2 presents the adjusted mean learning gains for additive (Fig 2a) and for proto-proportional problem solvers (Fig 2b) that are paired with non-proportional peers. In contrast to the previous models, no main effects for hypothesis testing were found, neither for additive problem solvers ( \(F<1\) ), nor for proto-proportional learners ( \(F<1\) ). Among additive problem solvers, a main effect was found for pairing condition, \(F(1,56.7)=6.01, p<.017\), with students who were paired with a proto-proportional peer showing higher learning gains ( \(M=.24, S E=.03\) ) compared to those that were paired with a same-level peer ( \(M=.10\), \(S E=03\) ). Pairing condition was also found to interact with hypothesis testing, \(F(1,56.7)=5.56, p<.022\). Post-hoc analyses (with Tukey-Kramer adjustments) showed that when paired with another additive problem solver, they learning gains were higher with hypothesis testing ( \(M=.18\), \(S E=05)\), than without it ( \(M=.02, S E=05\) ), \(t(55)=2.39, p\) \(=.032\). When they were paired with a proto-proportional peer, on the other hand, additive problem solvers seemed to gain more without the hypothesis device ( \(M=.28, S E=06\) ) than with it ( \(M=.19, S E=06\) ). This apparent difference did not reach statistical significance however, \(t(73)=1.12\), \(n s\).

For the proto-proportional problem solvers, on the other hand, no effect were found for neither pairing condition ( \(F\) (81) \(=1.34, n s\) ), hypothesis testing ( \(F<1\) ), nor their interaction \((F<1)\).

\section*{Discussion}

Previous studies have examined the effects of hypothesis testing and collaborative learning on cognitive growth (e.g., Ellis et al, 1993; Howe et al, 2000; Schwarz et al, 2000; Tudge et al, 1996). Unfortunately, this literature has yielded a mixed pattern of results. In the present study we revisited the major research questions in this field with a controlled experimental design that systematically explored the full range of dyadic compositions and with statistical models that controlled for nested effects. Overall, the findings show that the answer to the question whether hypothesis testingbased interventions for learning are more effective in individual or collaborative settings, really depends on the level of analysis and the comparisons being made.

First of all, when all different types of dyadic compositions are included in the data set but not further specified, hypothesis testing was overall found to improve students' learning gains. This finding is consistent with earlier research on the effectiveness of providing students with feedback that consistently confirms correct predictions and disconfirms predictions based on incorrect understanding (e.g., Tudge \& Winterhoff, 1993; Tudge et al, 1996). Collaboration, on the other hand, was not found to have an overall advantage over individual work. It was neither found to improve learning through hypothesis testing as is often expected and as found in other studies (Ellis et al, 1993). It is often believed that peer collaboration allows learners to discuss different explanations and generate interpretations of the hypothesis testing outcomes (Howe et al 2000). However, such potential benefits of collaboration are not detectable when the full range of dyadic pairings are included but not further specified.

A further dissection of the general construct of 'collaboration' according to the target student's and the partner's competence levels uncovers that interaction with a more competent peer only improves learning under certain specific conditions: For non-proportional ("wrong") students, the combination of hypothesis testing and being paired with a proportional ("right") partner was particularly powerful. However, similar to Ellis et al (1993) we found that when students received no feedback from the equipment (no hypothesis testing), singletons, students paired with proportional peers and students paired with different level non-proportional peers showed only comparable (moderate-to-small) gains. In concordance with previous findings (Ames \& Murray, 1983; Schwarz et al, 2000), students who were paired with a same-level "wrong" peer without the opportunity to receive any feedback through hypothesis testing did not improve at all. The pattern that emerges from these findings seems to underline the importance of the combination of exposure to higherorder reasoning strategies and the confirmation of the correctness of these strategies by an objective test. This is not an additive effect, since neither the exposure to higherorder reasoning strategies, nor the conflict created by the disconfirmation of incorrect predictions alone led to substantive learning gains.

This subtle contingency of, on the hand, the kind of feedback that is obtained from objective testing and, on the
other, the persuasiveness of a higher-order reasoning strategy becomes even more evident when we considered the wrong-wrong pairings only: The benefits of interaction with a more competent peer and hypothesis testing were found to hold only when the test proved that the predictions of this more competent peer were correct. When less competent (using additive strategies) interacted with a more competent peers (using proto-proportional strategies), the former actually gained more without hypothesis testing. When there is no opportunity to test the correctness of predictions, the verbal explanation provided by the slightly more competent peer may convince the lower competence peer to the more sophisticated reasoning strategy, thus improving their performance on posttests. However, with access to hypothesis testing devices, the predictions of the more competent peer will be disconfirmed, and with it the (slightly) more sophisticated reasoning strategy.

\section*{References}

Asterhan, C. S. C. \& Schwarz, B. B. (2009). The role of argumentation and explanation in conceptual change: Indications from protocol analyses of peer-to-peer dialogue. Cognitive Science, 33, 373-399.
Asterhan, C. S. C., \& Schwarz, B. B. (2007). The effects of monological and dialogical argumentation on concept learning in evolutionary theory. Journal of Educational Psychology, 99, 626-639.
Ellis, S., Klahr, D., and Siegler R.S. (1993). The effects of feedback and collaboration on changes in children's use of mathematical rules. Presented at the Biennial Meetings of the Society for Research in Child Development.
Howe, C., Tolmie, A., \& Rodgers, C. (1992). The acquisition of conceptual knowledge in science by primary school children: Group interaction and the understanding of motion down an incline. British Journal of Developmental Psychology, 10, 113-130.
Howe, C., Tolmie, A., Duchak-Tanner, V., \& Rattay, C. (2000). Hypothesis-testing in science: Group consensus and the acquisition of conceptual and procedural knowledge. Learning \& Instruction, 10, 361-391.
Messer, D. J. Joiner, R. Loveridge, N. Light, P. \& Littleton, K (1993). Influences on teh effectiveness of peer interaction: children's level of cognitive development and teh relative abiluty of partners. Social Development, 2, 279 - 294.
Mix, K. S., Levine, S. C., \& Huttenlocher, J. (1999). Early fraction calculation ability. Developmental Psychology, 35, 164-174.
Schwarz, B. B., \& Linchevski, L. (2007). The role of task design and of argumentation in cognitive development during peer interaction. The case of proportional reasoning. , 17(5), 510-531.
Tudge, J. R. H., Winterhoff, P. A. and Hogan, D. M. (1996). The Cognitive Consequences of Collaborative Problem Solving with and without Feedback. Child Development, 67, 2892-2909

\title{
Spatial Gestures Point the Way: A Broader Understanding of the Gestural Referent
}

\author{
Kinnari Atit (kinnari.atit@temple.edu) \\ Ilyse Resnick (ilyse.resnick@temple.edu) \\ Thomas F. Shipley (tshipley@temple.edu) \\ Temple University, 1701 N. 13th St. \\ Philadelphia, PA 19122 USA
}

\author{
Tilbe Goksun (tilbe@mail.upenn.edu) \\ University of Pennsylvania, 3400 Spruce St. Philadelphia, PA 19104 USA
}

\author{
Carol J. Ormand (cormand@carleton.edu) \\ Cathryn A. Manduca (cmanduca@carleton.edu) \\ Carleton College, One North College St. \\ Northfield, MN 55057 USA
}

\author{
Basil Tikoff (basil@geology.wisc.edu) \\ University of Wisconsin-Madison, 1215 W Dayton St. Madison, WI 53706 USA
}

\begin{abstract}
We investigated the use of iconic and deictic gestures during the communication of spatial information. Expert structural geologists were asked to explain one portion of a geologic map. Spatial gestures used in each expert's response were coded as deictic (indicating an object in the conversational space), iconic (depicting an aspect of an object or event), or both deictic and iconic (indicating an object in the conversational space by depicting an aspect of that object). Speech paired with each gesture was coded for whether or not it referred to complex spatial properties (e.g. shape and orientation of an object). Results indicated that when communicating spatial information, people occasionally use gestures that are both deictic and iconic, and that these gestures tend to occur when complex spatial information is not provided in speech. These results suggest that existing classifications of gesture are not exclusive, especially for spatial discourse.
\end{abstract}

Keywords: gesture; deictic; iconic; highlighting

\section*{Introduction}

People communicate and focus the listener's attention to different levels of spatial information in speech and gesture. Spatial information is expressed in gesture both for communicating with others and for individual problem solving (Alibali, 2005). Common spatial activities (e.g. giving directions) (Lavergne \& Kimura, 1987; Allen, 2003) and the communication of complex spatial ideas (e.g. geology) (Liben, Christensen, \& Kastens, 2010) often include gesture.

Communicating three-dimensional spatial relationships using only language is difficult. As most spatial words are qualitative and are not apt for asserting metric spatial information (Tversky \& Lee, 1998), gesture is critical in conveying relations that cannot be easily expressed in speech. Gesture allows one to communicate thoughts that do not easily fit into the categorical system language offers (Goldin-Meadow, 1999). The literature provides a classification for spontaneous gestures made during regular discourse (e.g. McNeill, 1992; Krauss, Chen, \& Chawla, 1996; Ekman \& Friesan, 1969). This study investigates
whether the existing classification is appropriate for gestures that occur during spatial discourse involving complex spatial reasoning.

Extant research indicates that gestures occur more frequently when communicating spatial information, than when communicating non-spatial information (e.g. Alibali, Heath, \& Myers, 2001; Rauscher Krauss, \& Chen, 1996; Lavergne \& Kimura, 1987). For example, Alibali, Heath, and Myers (2001) asked participants to narrate a Tweety and Sylvester cartoon to a naïve addressee, and found that the speakers were nearly twice as likely to produce gestures with units that contained spatial prepositions than with units that did not. Furthermore, gesture frequency varies depending on speech topic. Lavergne and Kimura (1987) asked participants to speak for six minutes each on neutral topics (e.g. describe your typical school day routine), verbal topics (e.g. describe your favorite books and authors), and spatial topics (e.g. describe the route you would take to walk from the university's main library to the main entrance of campus). Participants produced twice as many gestures when speaking about spatial topics than when speaking about verbal or neutral topics.

People convey information using gestures in many different ways. For the purpose of this paper, gestures are defined as movements of the hands and arms that are produced when engaging in effortful cognitive activity (e.g. speaking, problem solving) (Alibali, 2005). Much of the literature has focused on two broad categories of movements: beat and representational gestures. Beats are hand movements that match the rhythm of the associated speech. For example, when reciting his grocery list, the speaker moves his finger up and down for every item on the list, "apples, bananas, cheese, and bread." Within the category of representational gestures (gestures that convey semantic content by virtue of shape, placement, or motion trajectory of the hands - e.g. pointing to the right to mean "right") (Alibali, 2005), gestures can be categorized as iconic or deictic (McNeill, 1992). These two broad types of gesture are the focus of this paper.

Iconic gestures "bear a close formal relationship to the semantic content of speech" (McNeill, 1992, p.12). For
example, when describing a scene from a comic book in which a character bends a tree back to the ground, the speaker makes gripping and pulling gestures as he or she describes the same actions (McNeill, 1992). Deictic gestures indicate entities in the conversational space (the physical space visible to both participants of the conversation). Usually, deictic gestures are pointing gestures that indicate objects and events in the concrete world (McNeill, 1992). For example, when choosing a puppy at the pet store, the child points to the puppy that he wants to buy.

Starting at 9 to 12 months of age, humans use pointing gestures to indicate objects in the environment (Bates, 1976; Bates, Benigni, Bretherton, Camaioni, \& Volterra, 1979). Though pointing is an easy and efficient way of indexing an object in space, and although the distinction between iconic and deictic gestures may be helpful in classifying nonspatial discourse, we propose that in tasks involving the communication of complex spatial information, the current classification may be limiting. The current classification is implicitly mutually exclusive, perhaps for methodological reasons, such that gestures would be classified as either deictic or iconic, but not both. If this assumption is incorrect, researchers could be in danger of missing potentially informative gestures that are both deictic and iconic.

The existing classification does not capture gestures that simultaneously draw the listener's attention to a specific object and represent two or more dimensions of spatial information. Gestures that are not pointing (or tracing) can provide "deictic" information; for example, an "iconic" gesture that resembled an object could be used to refer to the object. Such gestures have been reported anecdotally, Roth (2000) details a middle school science student explaining to his class the mechanism behind a pulley system. In his explanation, the student says, "Pull here," and used a gesture that made salient both the location and direction of the pull (Roth, 2000). The gesture is both iconic and deictic. As in this example, one could use the hand to draw the listener's attention to the form and location of something in the environment. Since we know that listeners make use of the information in a speaker's gestures (e.g. Alibali, Flevares, \& Goldin-Meadow, 1997; Goldin-Meadow \& Sandhofer, 1999) and spatial concepts can be hard to convey in speech alone, we may learn more about the function of gestures by observing their use in discussion of complex spatial settings.

Thus, the current study examines the use of deictic gestures during a spatial task. The results presented here are part of a larger study investigating the communication of spatial information by structural geologists. Structural geology is a spatially complex and cognitively demanding field, where experts gesture extensively when they speak. A reason to begin research in this domain is that these experts' gestures are likely to focus on complex spatial information. In the future, we plan to investigate if the patterns of communication found here are also present in other, more common, spatial situations. In the study, expert structural
geologists were asked to complete a series of tasks, including explaining the geology of two regions using geologic maps. Here we investigate experts' use of pointing gestures versus iconic gestures to indicate one or more objects on a map.

One reason geologists gesture is they are often in a situation where it is not possible to see the entire object of interest. Since the information found on an outcrop is complex and only one face of a structure is usually visible (providing two-dimensional information), experts could use iconic gestures to highlight critical features since the whole three-dimensional structure is not observable (Frodeman, 1995). A geologic map shares some of the same characteristics. The information found on a geologic map is quite complex, and it is a two-dimensional representation of three-dimensional structures. Based on our observations of experts in the field, combined with the complexity and twodimensional quality of a geologic map, we hypothesize that structural geology experts will use iconic gestures, in addition to pointing gestures, to index specific geological entities. We predict that they will use iconic gestures for the following reasons: 1) pointing gestures may be ambiguous as the referent is located within a complex image with overlapping features; and 2) the object of interest is a threedimensional structure - something that is not shown, but needs to be inferred from the map as only a slice through the three dimensional form may be visible at the surface. To test our hypothesis and characterize the gestures experts use, we coded experts' gestures for type (deictic, iconic, or both), and kind of spatial information (point, line, plane, or form, process/event) for responses to one question about one geologic map.

\section*{Methods}

\section*{Participants}

Thirty-four attendees at a Structural Geology and Tectonics Conference participated in the study. To focus on experts' gestures, we restricted analysis to data from those participants with a PhD who were also professors at an academic institution. Thus, data from ten participants were excluded. Data from one additional expert was excluded because he or she was bilingual in English and American Sign Language. Therefore, data from 23 expert structural geologists ( 14 men, 9 women, \(M_{\text {age }}=45.8\) years, age range: 33-60 years) was used for this analysis.

\section*{Materials}

Explanations were recorded with a Canon HD Video Camcorder HV20 (3.1 Megapixels). The map used for this portion of the study was a Geologic Map of the Black Hills Area, South Dakota and Wyoming (DeWitt, Redden, Buscher, \& Wilson, 1989). It was presented on a flat 78 cm high table.

\section*{Design and Procedure}

The study took place in a quiet room where only the experimenter and the participant were present. After arriving, the participant completed the consent process. He or she was then asked to stand behind the empty table placed in the center of the room. Throughout the course of the study, the experimenter stood directly across from the participant, approximately five feet away.

The experimenter explained that this was a study investigating teaching and reasoning about geologic maps. A Geologic Map of the Black Hills Area, South Dakota and Wyoming (DeWitt, Redden, Buscher, \& Wilson, 1989) was then placed on top of the table so the participant could see it. Participants were first asked whether they were familiar with the map before beginning the task. They were then asked to pretend that the experimenter was a geology undergraduate student with some domain knowledge, specifically having completed an introductory course and one or two upper level geology classes. The task was to explain what structures were under the ground along a specific cross-section, and explain how he or she knew. This task would be a familiar one to a structural geologist, and the map was designed to provide this information. After providing the prompt, the experimenter indicated the crosssection region on the map for the participant. Gesturing was never mentioned. Responses were audio and videorecorded.

\section*{Coding}

Each of the experts' spontaneous gestures and accompanying speech was coded by the first-author. Interrater reliability was established by having a second trained coder who independently coded a subset (20\%) of the responses. Each gesture was coded for the following things:

Speech The speech accompanying each gesture was transcribed. Due to the complexity of the information communicated by the experts, participants' speech was used to clarify the information represented in the gesture. Furthermore, speech accompanying each gesture was coded for whether it included information about a structure (e.g. dome, mountain) or provided orientation information (e.g. layers are steeply dipping). Inter-rater agreement for speech was \(\kappa=0.80\) ( \(\mathrm{n}=182\) gestures).

Gesture Using the accompanying speech to clarify, each gesture was coded for whether or not it represented a spatial property (e.g. the spatial relations between two rocks). See Atit, Shipley, and Tikoff (2013) for more information about the spatial properties represented in gesture. Inter-rater agreement for spatial property represented was \(\kappa=0.79\) ( \(\mathrm{n}=182\) gestures). Gestures that represented spatial properties were further characterized as follows.
Spatial Information The categories for the spatial information in a gesture were created based on the dimensional information that the gesture conveyed. A
gesture that conveys 1 D information indicates a point, a gesture that conveys 2D information indicates a line, gestures that convey 3D information indicate planes and forms, and gestures conveying 4D information indicate changes or processes. For more information on this categorization of gestures, see Atit, Shipley, and Tikoff (2013).

Each gesture was coded for one of the following six spatial categories: 1) point: hand-shape used was typically an index finger indicating a location in space, 2) line: hand shape typically was an index finger indicating a line in space, 3) plane: hand shape typically was a flat palm indicating a plane in space (generally providing information about orientation), 4) form: hand formed a threedimensional shape in space (e.g., forming the hand in the shape of a dome or moving the hand to sculpt the shape of a dome), 5) process/event: hand conveyed a process or an event (e.g., hand showing the movement of magma representing an intrusion), and 6) other: all other gestures. Inter-rater agreement for spatial information in gesture was \(\kappa=0.81\) ( \(\mathrm{n}=182\) gestures).
Function Type Each gesture was also categorized into one of the following four types: 1) deictic: if it indicated an entity on the map (e.g. pointing to a specific fault line on the map, or tracing the fault line on the map), 2) iconic: if it "bears a close relationship to the semantic content of speech" (McNeill, 1992, p. 12), and depicted an aspect of an object within the conversational space (e.g. a curved hand used to represent a fold), 3) both: if it simultaneously drew the listener's attention to a specific object on the map while depicting an aspect of it (e.g. using a curved hand to show the shape and location of a fold on the map), or 4) unrelated: if it could not be classified into any of the three type categories. Inter-rater agreement for gesture type was \(\kappa=0.82\) ( \(\mathrm{n}=182\) gestures).

\section*{Results}

\section*{Spatial Gestures}

On average, each participant gestured 51.87 times over the course of the task ( \(S D=26.94\) ). We found no difference in the number of gestures produced by men ( \(M=51.00\), \(S D=29.14\) ) versus women ( \(M=53.22, S D=24.75\) ), n.s; and no difference in the number of gestures produced by participants who were familiar with the map ( \(M=58.60\), \(S D=30.93\) ) versus those that were not familiar with the map ( \(M=46.69, S D=23.37\) ), n.s.

When looking at the information conveyed within gestures, we found that participants gestured more about spatial information (gestures conveying a spatial property) ( \(M=0.73, S D=0.14\) ) than about non-spatial information ( \(M=0.27, S D=0.14\) ), \(t(22)=7.96, p<.001\). All means and standard deviations for the following analyses are provided in Table 1.

Table 1: Means and standard deviations of gestures
\begin{tabular}{c|ccc}
\cline { 2 - 4 } & Point & Line & Plane/Form/Process/Other \\
\hline Deictic & \(M=0.11\) & \(M=0.20\) & \(M=0.06\) \\
& \(S D=0.08\) & \(S D=0.15\) & \(S D=0.07\) \\
Non- & \(M=0.01\) & \(M=0.01\) & \(M=0.37\) \\
Deictic & \(S D=0.02\) & \(S D=0.02\) & \(S D=0.22\) \\
\hline
\end{tabular}

Note. Table presenting means and standard deviations across participants for the different kinds of gesture (point, line, plane/form/process/other) and for different types of gesture (deictic, non-deictic). The descriptives reported here are the means and standard deviations of the proportions of spatial gestures for each category. As there were no differences between the plane, form, process, and other categories, we collapsed across these categories. Gestures that were deictic indexed an object on the map, and could be composed of point, line, plane, form, process, and other gestures. Gestures that were non-deictic did not index an object on the map, but also could be composed of point, line, plane, process, and other gestures.

First, we looked at what proportion of spatial gestures was pointing gestures. On average, \(12 \%\) of each participant's spatial gestures involved pointing ( \(M=0.12, S D=0.09\) ). An overwhelming majority of those were identified as deictic ( \(M=0.11, S D=0.08\) ), with only \(1 \%\) of gestures being iconic and pointing \((M=0.01, S D=0.02), t(22)=5.98, p<.001\). Pointing gestures were mainly used to index an object in the conversational space.

Second, we considered the other cases of deictic gestures. The first notable observation is that while pointing maybe the prototypical deictic gesture, in this context the most frequent deictic gesture was tracing a line in the conversational space. On average, \(21 \%\) of each participant's spatial gestures were of this kind \((M=0.21, S D=0.15)\). More line gestures were classified as deictic \((M=0.20, S D=0.15)\), than non-deictic \((M=0.01, S D=0.02), t(22)=5.83, p<.001\), and the frequency of deictic line gestures \((M=0.20\), \(S D=0.15\) ) was significantly greater than pointing gestures ( \(M=0.11, S D=0.08\) ), \(t(22)=3.24, p<.01\).

Finally, when we consider the spatial gestures that were plane, form, or process/event gestures, we find that most of these were iconic \((M=0.37, S D=0.22)\). However, an intriguing portion was both iconic and deictic ( \(M=0.06\), \(S D=0.07\) ). The proportion of gestures classified as "both" iconic and deictic was significantly different from 0 , \(t(22)=4.25, p<.001\). Thus, most of the gestures made by experts could be classified as iconic or deictic, but there were a significant number of gestures that were both deictic and iconic. Experts in this task used complex iconic gestures to index objects in the conversational space.

\section*{Gestures Classified as Both, Iconic and Deictic}

To further explore the information represented in the gestures that were both deictic and iconic, we categorized them by the spatial information in the gestures. About half
of the \(6 \%\) represented planes \((M=0.03, S D=0.04)\) and half represented forms ( \(M=0.03, S D=0.05\) ). Less than \(1 \%\) of all gestures represented a process or event, or other kind of information. Experts may have used the planar and form gestures because the task was to explain the structures at a line of cross-section where orientation and shape information is not readily visible on the map. Without visible support for this spatial information in the diagram, experts may have employed gestures to ensure the threedimensional referent was clear.

Lastly, we investigated the information conveyed in speech when experts employed these iconic and deictic gestures compared to when they pointed or traced to indicate an object. To make this comparison, we computed the following for each participant: 1) the proportion of pointing deictic gestures paired with spatially complex speech (speech containing structure or orientation information) relative to the total number of pointing deictic gestures; 2) the proportion of line deictic gestures paired with spatially complex speech relative to the total number of line deictic gestures; and 3) the proportion of plane and form gestures classified as both iconic and deictic paired with spatially complex speech relative to the total number of plane and form gestures classified as both iconic and deictic.

We found that a greater proportion of pointing deictic gestures were paired with spatially complex speech ( \(M=0.34, S D=0.36\) ) than planar/form iconic and deictic gestures, \((M=0.04, S D=0.12), t(22)=3.57, p<.01\). Similarly, more line deictic gestures were paired with spatially complex speech ( \(M=0.27, S D=0.23\) ) than planar/form iconic and deictic gestures, \(t(22)=3.91, p<.01\). Thus, experts used gestures classified as both iconic and deictic especially in instances where the complex spatial information was not provided in speech.

\section*{Discussion}

This study investigated the relevance of an existing distinction made in the gesture literature, iconic versus deictic gestures, within the realm of communicating spatial information. Traditionally, deictic gestures are defined as hand movements that indicate entities in the conversational space and usually consist of pointing (McNeill, 1992). Iconic gestures are hand movements that "bear a close formal relationship to the semantic content of speech" (McNeill, 1992, p.12), and generally do depict some aspect of an object in the conversational space.

Data from this study indicates that when asked to explain the structures present in a section of a geologic map, expert structural geologists use gestures that can be classified as deictic or iconic, along with gestures that fall in both categories. When using gestures traditionally classified as deictic, experts tended to trace more than point to draw the listener's attention to an object on the map. When using gestures traditionally classified as iconic, we found that experts used some iconic gestures to indicate an object on the map. Furthermore, gestures that were both iconic and deictic and represented information about planes and forms,
tended to occur when there was no spatially complex information (e.g. structure or orientation information) conveyed in speech.

A number of studies have shown that gestures are useful in separating relevant from irrelevant information (e.g. Roth, 2000; Lozano \& Tversky, 2006; Heiser, Tversky, \& Silverman, 2004). Gestures help organize the conversational space into a salient foreground and an unrepresented, more diffuse background. Researchers in the past have likely focused on pointing because this hand shape is the most common one used to indicate the foreground. In contrast, here we replicate previous work (e.g. Lozano \& Tversky, 2006; Heiser, Tversky, \& Silverman, 2004) that has found that when the referent is complex and includes multiple kinds of information (e.g. maps), the speaker also uses tracing gestures to highlight objects for the listener. For example, Lozano and Tversky (2006) asked participants to explain to a listener how to assemble a piece of furniture, and found that participants used tracing gestures in addition to pointing gestures to draw the listener's attention to individual pieces (Lozano \& Tversky, 2006). Heiser et al. (2004) asked pairs of students to use a campus map to design and produce an optimal emergency rescue route. They found that students also used tracing gestures along with pointing gestures to focus their partner's attention to specific aspects of their sketch and to highlight certain routes (Heiser, Tversky, \& Silverman, 2004). We suspect that speakers used tracing gestures because pointing alone may be ambiguous when the referent has significant spatial extent and when it does not have clear boundaries.

The results from our study reveal an interesting type of gesture that is used to draw the listener's attention to an object. Experts in our study used complex gestures traditionally classified as iconic to highlight objects on the geologic map. A geologic map presents a horizontal crosssection through the three-dimensional topography of a region. Therefore, many objects of interest to a geoscientist will be three-dimensional structures that are not completely visible at the surface. Indeed what is visible at the surface may be a slice through a three-dimensional form. Since the two-dimensional information presented on the map does not directly resemble the actual three-dimensional form of these objects, the expert may use gesture to provide the listener with the missing information. For example, the elliptical outcrop pattern of rock layers presented on the map in this study does not resemble the three-dimensional form of the dome in the Black Hills region. Furthermore, it can be difficult to determine the orientation of the rock layers within a domal structure on a geological map because the inclination of the rock layers is typically represented using a symbol. Thus, the expert uses gestures to depict the shape of the dome and planar gestures to show the orientation of the layers in space. Whether it is the two-dimensional characteristic of the map, or the penetrative nature of geological structures that elicits this special type of gesture is a question for future research. For example, would an
architect use gestures that could be classified as both iconic and deictic when explaining a blueprint of a building?

Finally, gestures classified as both iconic and deictic were used in instances where the spatially complex information is not provided in speech. For example, an expert represents the orientation and location of a layer of rocks on the map while referring to their relative ages in speech. Since complex spatial information is difficult to convey using language (Tversky \& Lee, 1998) and gesture allows one to communicate thoughts that are not easily conveyed in speech (Goldin-Meadow, 1999), perhaps the speaker uses this type of gestures when providing multiple levels of information (e.g. location and orientation information) in language alone becomes difficult. Or, perhaps the expert could not produce the gesture and speech at the same time due to the cognitive load required by the task. A more global analysis of speech in the future can address this question.

One potential limitation of the current study may be that the experts were asked to pretend that the experimenter was a geology undergraduate to whom they were explaining the cross-section. It is possible the experts' language and gestures would have differed if they were providing explanations to real geology undergraduates. This seems unlikely as the experts were obviously engaged in their answers to the questions. Nevertheless, studies collecting and analyzing interactions between experts and novices in the field are in progress.
Supported by the findings of this study, we argue that the existing types of representational gestures (e.g. iconic and deictic) should not be treated as exclusive categories overlap between the two types exists. People do use iconic gestures (e.g. planar gestures) to indicate, or highlight, objects in the world when the important spatial information is three-dimensional. One open question is whether the number of gestures that fall into "both" categories is related to the spatial complexity of the information conveyed. Perhaps the communication of more spatially complex information elicits a greater use of this special type of gestures.

Understanding how different types of gestures are employed in simple communicative contexts to highlight and convey complex spatial information may serve as the foundation for developing pedagogical techniques for conveying spatial information. A richer understanding of the different types of gestures could inform geology professors, and science professors in general, about how to most effectively communicate information to their students.

\section*{Acknowledgements}

This research was supported by a grant to the Spatial Intelligence and Learning Center, funded by the National Science Foundation (SBE-0541957 and SBE-1041707), and by a Fostering Interdisciplinary Research on Education grant, funded by the National Science Foundation (grant number DRL-1138619).

\section*{References}

Alibali, M. W. (2005). Gesture in spatial cognition: Expressing, communicating, and thinking about spatial information. Spatial Cognition \& Computation, 5(4), 307331.

Alibali, M., Flevares, L. M., \& Goldin-Meadow, S. (1997). Assessing knowledge conveyed in gesture: Do teachers have the upper hand?. Journal Of Educational Psychology, 89(1), 183-193.
Alibali, M. W., Heath, D. C., \& Myers, H. J. (2001). Effects of visibility between speaker and listener on gesture production: Some gestures are meant to be seen. Journal Of Memory And Language, 44(2), 169-188.
Allen, G. L. (2003). Gestures accompanying verbal route directions: Do they point to a new avenue for examining spatial representations? Spatial Cognition and Computation, 3(4), 259-268.
Atit, K. A., Shipley, T. F., \& Tikoff, B. (submitted). A framework for classifying spatial gestures. In D. Montello, K. Grossner, \& D. Janelle (Eds.), Space in Mind: Concepts and Ontologies for Spatial Education. Chapter submitted for publication.
Bates, E. (1976). Language and context: The acquisition of pragmatics Academic Press New York.
Bates, E., Benigni, L., Bretherton, I., Camaioni, L., \& Volterra, V. (1979). The emergence of symbols: Cognition and communication in infancy Academic Press New York.
DeWitt, E., Redden, J., Buscher, D., \& Wilson, A. B. (1989). Geologic map of the Black Hills Area, South Dakota and Wyoming.
Ekman, P., \& Friesen, W. V. (1969). The repertoire of non-verbal behavior: Categories, origins, usage, and coding. Semiotica, 1, 49-98.
Frodeman, R. (1995). Geological reasoning: Geology as an interpretive and historical science, GSA Bulletin, 107, 960 968.

Goldin-Meadow, S. (1999). The role of gesture in communication and thinking. Trends in Cognitive Science, 3, 419-429.
Goldin-Meadow, S. \& Sandhofer, C. M (1999). Gestures convey substantive information about a child's thoughts to ordinary listeners. Developmental Science, 2(1), 67-74.
Heiser, J., Tversky, B., \& Silverman, M. (2004). Sketches for and from collaboration. Visual and Spatial Reasoning in Design III, , 69-78.
Krauss, R. M., Chen, Y., \& Chawla, P. (1996). Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us? Advances in Experimental Social Psychology, 28, 389-450.
Lavergne, J., \& Kimura, D. (1987). Hand movement asymmetry during speech: No effect of speaking topic. Neuropsychologia, 25(4), 689-693.
Liben, L. S., Christensen, A. E., \& Kastens, K. A. (2010). Gestures in geology: The roles of spatial skills, expertise,
and communicative context. Spatial Cognition VII, 95111.

Lozano, S. C., \& Tversky, B. (2006). Communicative gestures facilitate problem solving for both communicators and recipients. Journal of Memory and Language, 55(1), 47-63. doi: 10.1016/j.jml.2005.09.002
McNeill, D. (1992). Hand and mind: What gestures reveal about thought University of Chicago Press.
Rauscher, F. H., Krauss, R. M., \& Chen, Y. (1996). Gesture, speech, and lexical access: The role of lexical movements in speech production. Psychological Science, 7(4), 226231.

Roth, W. M. (2000). From gesture to scientific language. Journal of Pragmatics, 32(11), 1683-1714.
Tversky, B., \& Lee, P. (1998). How space structures language. In C. Freksa, C. Habel, \& K. F. Wender (Eds.), Spatial Cognition: An Interdisciplinary Approach to Representing and Processing Knowledge (pp. 157-175). Heidelberg, Germany: Springer-Verlag.

\title{
The Role of Contextual Repetition During Fast Mapping on Word Learning
}

\author{
Emma L. Axelsson (e.axelsson@unsw.edu.au) \\ School of Psychiatry, ICAMHS Research, Mental Health Centre Level 1, Locked Bag 7103 \\ Liverpool Hospital, Liverpool BC NSW 1871 AUSTRALIA
}

\author{
Jessica S. Horst (jessica@sussex.ac.uk) \\ School of Psychology, Pevensey 1 Building \\ Falmer, Brighton, BN1 9QH, UK
}

\begin{abstract}
Recent research suggests that children's ability to learn words via fast mapping is strongly related to the attentional demands of the task. Here we explore whether lowering the attentional demands during the initial fast mapping task facilitates word learning. Three-year-old children completed fast mapping and test trials using a touch screen computer. For half of the children, the non-targets (competitors) repeated across trials and for other children there was no repetition. All children received the same word learning test trials. Only children who had received repeating competitors (lower attentional demands) during the initial fast mapping task demonstrated word learning. Thus, these data suggest that children's ability to learn novel names is strongly influenced by the competition and attentional demands of the initial fast mapping context.
\end{abstract}

Keywords: fast mapping; word learning; competition; attentional demands.

\section*{Does Fast Mapping Enable Word Learning?}

Learning to comprehend the vast number of words that children are bombarded with daily is a daunting task. Arguably, it is equally daunting for researchers to explain the mechanisms responsible for children's developing ability to understand what a speaker is referring to when uttering a new word, but also children's ability to remember the link between the word and referent for later encounters.

Without direct instruction, children effectively determine the referent for a novel word on their own-particularly when the referent is presented in the context of known objects. Typically referred to as fast mapping (Carey, 1978), children appear to determine the referent via process-ofelimination (Halberda, 2003). That is, children use their prior knowledge (i.e., known vocabulary) to rule out the objects they have already linked to a name. Instead, they select a novel object as the likely referent of a novel name. Biases to novelty appear to be key in fast mapping. Recent research further indicates that even in the context of other nameless, novel objects, children will select the most novel object as a referent for a novel name (Horst, Samuelson, Kucker, \& McMurray, 2011; Mather \& Plunkett, 2012).

However, fast mapping is not word learning per se, but rather an initial step in the word learning process (Carey, 1978; Horst \& Samuelson, 2008). Fast mapping appears to be relatively easy for children, yet retention of the nameobject mappings is more difficult. For example, Horst and Samuelson (2008) presented 24 -month-old children with
referent selection (i.e., fast mapping) trials with one novel object and two known objects. Only 5 minutes later, they tested children's retention for the same novel name-object associations by presenting the targets among other novel targets that had been encountered during the initial referent selection trials and foils that had been seen during a preferential looking task. Children had little difficulty selecting the correct referents on the initial referent selection trials. However, children performed at chance levels on the retention trials, suggesting that the initial name-object mappings did not lead to robust representations in long-term memory. Other studies have found similar difficulties in long-term retention of name-object mappings despite demonstrations of minimal difficulty disambiguating novel from known objects (e.g., Bion, Borovsky \& Fernald, 2013; Gureen, Horne \& Erjavec, 2012).

Note, other studies do suggest minimal exposure to a novel word and object may lead to long-term retention of the association, but these have typically involved only naming a single novel object (e.g., Woodward et al., 1994). It remains unclear, therefore, whether children's selection of the target on the delayed test trials are really in response to the specific phonetic content of the new word or because that word was the only new word introduced during training, rendering it unique among the available alternatives (Dollaghan, 1985; Schafer \& Plunkett, 1998).

\section*{Task Demands}

Taken together, these studies suggest that children are adept at forming the initial associations between novel referents and their names, however, children's ability to retain these mappings may suffer when task demands are high. For example, in a recent study exploring the effect of the number of known competitors present during the initial fast mapping task on subsequent word learning, 2-year-old children were only able to retain novel name-object associations when task demands were relatively low (Horst, Scott \& Pollard, 2010). Specifically, children were only able to retain words when they had seen few known competitors across fast mapping trials (eight) but not when they had seen more competitors across trials (12 or 15)-although the number of targets was the same for all groups.

Clearly then, the non-target competitors play an important role in word learning via fast mapping. Horst and Samuelson (2008) found that children could only retain the novel name-object associations when ostensive naming was
provided after each fast mapping trial. That is, when the experimenter picked up a novel object, pointed to it and restated its name. Importantly, the experimenter physically moved the target away from the known competitors. For another group of children, naming was also provided after each trial, however, targets were not moved away from the competitors. These children did not learn the name-object associations. The authors argue that ostensive naming helps children focus on the target while simultaneously drawing their attention away from the competitors.

A follow-up study explicitly tested this explanation of ostensive naming by illuminating targets from below (a light flashed inside the tray the objects were on) and by covering competitors with semi-opaque boxes (Axelsson, Churchley \& Horst, 2012). Children who received this form of ostensive naming learned the name-object associations. The authors argue that ostensive naming helped children better encode the name-object association as their attention to the target object was sustained. Children who only received ostensive naming in the form of pointing (without moving the target away from the competitors) did not demonstrate significant word learning.

In both of these examples, the initial fast mapping task was made easier by decreasing the attentional demands. Specifically, the experimenter helped guide the child's attention away from the competitors, which facilitated processing of the targets.

Similarly, increasing attentional demands can also have a detrimental effect on subsequent word learning. Wilkinson, and colleagues (2003) presented one group of children with fast mapping trials with one novel object and three known competitors. Another group received fast mapping trials where each successive trial included not only a novel object and two known competitors, but also the novel object from the previous trial. Children who faced such increased attentional demands were less successful on subsequent fast mapping trials. Because children typically prefer novelty (Horst et al., 2011; Mather \& Plunkett, 2012), the presence of another novel object likely decreased attention to the target.

\section*{Contextual Repetition}

Similar effects of the reduction of attentional demands during word learning tasks have been demonstrated in other contexts (for a review, see Horst, 2013). In an investigation of word learning via shared storybook reading, Horst, Parsons and Bryan (2011) found that 3-year-old children who were read the same storybooks repeatedly (three times) successfully retained novel words from the books when tested one week later. In contrast, children who were read three different stories performed at chance levels when tested on the same novel words one week later. Importantly, children in both groups were exposed to the same number of novel words the same number of times. The critical difference was that the first group encountered the words in the same context repeatedly whereas the second group encountered them in different contexts. The authors
concluded that children who encountered words in repeated contexts were at an advantage because they had less information to process, facilitating long-term retention of the name-object associations. The question that remains unanswered, then, is whether the same mechanism is at play in other contexts.

\section*{The Current Study}

The aim of the current study is to examine the effect of contextual repetition during fast mapping on subsequent word learning. To examine the effects of the learning context, children either saw novel objects with the same or different known competitors across trials. Specifically, children either encountered the novel object in the same context (with the same competitors) across all three trials with that given target or encountered the novel object in different contexts (with different competitors) across trials. Thus, the attentional demands of the task were either relatively low (repeated contexts) or relatively high (different contexts). Therefore, the degree of attentional demand was expected to differ across the four groups during the fast mapping trials. As in the storybook studies (e.g., Horst et al., 2011), those presented with repeated contexts (competitors) during novel target referent selection trials were predicted to perform better at test. As attentional demands are presumably high during referent selection of novel objects, the repetition of competitors may aid in longterm learning of name-object mappings.

\section*{Method}

\section*{Participants}

Forty-eight typically developing, monolingual, Englishspeaking children aged 36 months ( \(M\) age 36 months, 13 days, \(S D=73.79\) days; range 33 months, 0 days -41 months, 30 days; 24 girls) participated. Children were from predominantly white, middle class homes recruited from southern England. Data from 1 additional child were excluded because she consistently touched the screen before waiting to hear the instructions. Parents were reimbursed for travel expenses and children received a small gift.

\section*{Stimuli}

Digital photographs of known (familiar) and novel objects served as the target stimuli. Specifically, six known objects (ball, cup, train, cow, frog, elephant) and three novel objects (the end of a foam arrow/zorch, a y -shaped rubber dog toy/gaz and a clacker/sprock) served as target stimuli on the referent selection trials. The known objects were chosen because they are highly familiar to 3 -year-old children. The novel objects were chosen because they are unfamiliar to most 3 -year-old children and they do not know names for these objects. Additional photographs of three aliens, a bed and a dresser were used during the experiment. A female, native British English speaker narrated the procedure for the child (henceforth the narrator).

\section*{Procedure and Design}

During the experiment, children were seated on their parents' laps at a table in front of a Dell computer with a touchscreen monitor. The keyboard and mouse were out of the children's reach. Stimuli were presented using ePrime 2.0. Children were asked if they would like to play a computer game. During the experiment, the experimenter stood behind the child to ensure the child remained on task, to answer questions and to minimize parental interference.

Introduction and Warm-up Trials The first screen depicted a red alien who the narrator introduced as Modo. The recorded script was carefully written so that no pronouns were used: thus each child could decide if Modo was male or female. The next screen depicted a bed, a dresser and small versions of each of the known and novel objects that would later be used on the referent selection trials. The narrator explained that Modo was very messy and asked if the child would "help tidy up Modo's room."
Three warm-up trials immediately followed. These trials served to introduce the child to the task and to help the child feel comfortable touching the computer screen. On each trial the child was presented with three additional known objects and asked to select an object by the narrator, e.g., "Can you find the spoon? Touch the spoon." If the child had not yet touched the screen after 1000 ms , the recording repeated and continued to loop with 1000 ms in between requests for up to 30 seconds, or until the child touched the screen. After the child touched a picture, a blank, white screen appeared to give the experimenter a moment to praise the child. The same objects were presented on each warm-up trial, but object positions (left, middle, right) were counterbalanced across trials so that each object appeared once in each position. Thus, the child was asked for a different object in a different position on each warm-up trial and could practice touching an object at each possible position.

Referent Selection Trials Referent selection trials immediately followed the warm-up trials and followed the same procedure except that children did not receive feedback after these trials. On each referent selection trial the child was presented with two known objects (e.g., elephant, cup) and one novel object (e.g., clacker) and asked to choose either a known or novel object, e.g., "Can you find the sprock? Touch the sprock." Children were presented with 18 referent selection trials including 9 known
name and 9 novel name referent selection trials, of which there were 3 trials for each novel name (e.g., 3 sprock trials). Which objects children saw on these trials varied across conditions depending on whether or not the same competitors were repeated across trials (see Table 1).

For half of the children the same competitors were repeated across all trials for a given novel name. For example, each time they were asked for the sprock (clacker), it was displayed with the elephant and the cup. For the other children, different competitors were displayed on each trial for a given novel name. For example, the first sprock trial may have included the elephant and cup, the second sprock trial the train and frog and the third sprock trial the cow and ball. Object animacy was also counterbalanced across trials.

Likewise, for half of the children, the same competitors were repeated across all trials with the same known name. For example, each time they were asked for the elephant, it was displayed with the cup and sprock (clacker). For the other children, different competitors were displayed on each known name referent selection trial. For example, the first elephant trial may have included the cup and sprock, the second elephant trial the ball and zorch, and the elephant was also a competitor along with the gaz on a train trial. Importantly, all children saw the same six known objects and three novel objects across referent selection trials an equal number of times. However, for this to evenly occur some targets (e.g., elephant) were also competitors in the non-repeat conditions.
This resulted in a \(2 \times 2\) design with four conditions: competitors repeat across all trials (i.e., both trial types), competitors repeat across novel trials (but not known trials), competitors repeat across known trials (but not novel trials) and competitors do not repeat on any trials. One can also consider the four conditions of the current study on a continuum of attentional demands from low (competitors repeat across all trials) to intermediate (competitors repeat across novel trials or competitors repeat across known trials) to high (competitors never repeat).

Children were asked for each of the novel targets once during trials 1-6, once during trials 7-12 and once during trials 13-18. Each novel target appeared once in each position (left, middle, right) when it was a target and once in each position when it was a competitor during the known name trials. The same objects were never presented on two consecutive trials and no more than two trials of either type (i.e., known or novel targets) were presented sequentially.

Table 1: Example targets and competitors (comp.) for the Repeat Across All Trials and the No Repeat Conditions.
\begin{tabular}{clllclll}
\hline \multicolumn{4}{c}{ Repeat Across All Trials } & \multicolumn{4}{c}{ No Repeat Across Trials } \\
Trial & Target & Comp. 1 & Comp. 2 & Trial & Target & Comp. 1 & Comp. 2 \\
\hline 1 & Sprock & Elephant & Cup & 1 & Sprock & Elephant & Cup \\
7 & Sprock & Elephant & Cup & 8 & Sprock & Train & Frog \\
14 & Sprock & Elephant & Cup & 11 & Sprock & Cow & Ball \\
4 & Elephant & Cup & Sprock & 5 & Elephant & Frog & Sprock \\
10 & Elephant & Cup & Sprock & 13 & Elephant & Ball & Zorch \\
17 & Elephant & Cup & Sprock & 18 & Train & Elephant & Gaz \\
\hline
\end{tabular}

Word Learning Trials After the final referent selection trial, the screen showed the bed, dresser and the small stimuli pictures from before but displayed in neat rows and the narrator told the child the room was now tidy. The next screen showed Modo with another alien who was blue. The narrator explained that Modo's friend ZeeBee had come over to play and asked if the child could help Modo share toys with ZeeBee. Then, the word learning test trials were presented again following the same procedure. Importantly, the word learning trials were the same for all conditions. On each trial all three novel objects were displayed and the child was asked to choose an object. Children were asked to choose each target once (for a total of 3 trials). Each novel object appeared once in each position (left, middle, right). At the end the narrator told the child "It's fun to share toys!"

\section*{Results}

\section*{Referent Selection Trials}

As can be seen clearly in Figure 1, children in all four conditions performed very well during the initial referent selection task. On the known name referent selection trials, children chose the target object significantly more than would be expected by chance (.33) in the repeat across all trials condition \((t(11)=19.34, p<.001, d=5.53)\), in the repeat across novel trials condition \((t(11)=14.45, p<.001\), \(d=4.12\) ), in the repeat across known trials condition \((t(11)\) \(=44.78, p<.001, d=12.78)\) and the no repeat condition \((t(11)=29.40, p<.001, d=8.40)\), all \(p\) s two-tailed. Children's proportions of target choices on the known name referent selection trials were submitted to an ANOVA with competitors repeat across novel trials (yes, no) and competitors repeat across known trials (yes, no) as betweensubjects factors. The ANOVA yielded no significant effects, indicating that contextual repetition during referent selection did not influence children's ability to select known objects.


Figure 1. Children's proportion correct during the referent selection trials. Dotted line indicates chance (.33), error bars indicate 1 standard error from the mean. \({ }^{* * *} p<.001\).

Similarly, on the novel name referent selection trials, children chose the target object significantly more than would be expected by chance (.33) in the repeat across all trials condition \((t(11)=6.30, p<.001, d=1.82)\), in the repeat across novel trials condition \((t(11)=6.96, p<.001, d\) \(=2.02\) ), in the repeat across known trials condition \((t(11)=\) 18.35, \(p<.001, d=5.23\) ) and the no repeat condition \((t(11)\) \(=11.08, p<.001, d=3.18\) ), all \(p\) s two-tailed. Children's proportions of target choices on the novel name referent selection trials were submitted to an ANOVA with competitors repeat across novel trials (yes, no) and competitors repeat across known trials (yes, no) as between-subjects factors. The ANOVA yielded no significant effects. Thus, contextual repetition during referent selection did not influence children's ability to select novel objects either.

\section*{Word Learning Trials}

In contrast, contextual repetition during referent selection did influence children's learning of previously fast-mapped novel names. As can be seen in Figure 2, children in the repeat across all trials and repeat across novel trials conditions identified referents of novel names at rates significantly greater than expected by chance (.33), both \(t(11)=3.87, p<.01, d=1.12\). However, in contrast, children in the repeat across known trials and the no repeat conditions performed at chance levels, \(t(11)=1.50, n s\), and \(t(11)=1.20, n s\), respectively. Thus, repeating the same competitors across multiple novel name referent selection trials did influence children's ability to learn novel words.

To best understand the individual contributions of repeating the competitors on both novel and known name referent selection trials, children's proportion of target choices on the word learning trials was submitted to an ANOVA with competitors repeat across novel trials (yes, no) and competitors repeat across known trials (yes, no) as between-subjects factors.


Figure 2. Children's proportion correct during the word learning test trials. Dotted line indicates chance (.33), error bars indicate 1 standard error from the mean. \({ }^{* *} p<.01\).

The ANOVA yielded a main effect of repeat across novel trials, \(F(1,44)=6.99 p=.01, \eta_{\mathrm{p}}^{2}=.14\). Children in the conditions where competitors repeated across novel trials learned more words ( \(M=.65, S D=.29\) ) than children in the conditions in which competitors did not repeat across novel name referent selection trials ( \(M=.44, S D=.27\) ). No other significant effects were found. These findings clearly demonstrate that repetition of competitors during novel name referent selection influences later word learning.

\section*{Reaction Times}

Finally, we wanted to assess how the attentional demands of the task influenced children's performance throughout the experiment. To this end, we examined children's reaction times (RTs) on all correct trials over the course of the experiment. RTs were submitted to an ANOVA with competitors repeat across novel trials (yes, no) and competitors repeat across known trials (yes, no) as betweensubjects factors and encounter (first, second, third, fourth) as a repeated-measure (the fourth encounters were the word learning trials). The ANOVA yielded a significant main effect of encounter, \(F(2.57,94.96)=10.39, p=<.001, \eta_{\mathrm{p}}{ }^{2}=\) .22. Overall, children became faster over the four encounters. There was also a significant encounter by competitors repeat across novel trials interaction, \(F(2.57,94.96)=3.63, p=.02, \eta_{\mathrm{p}}^{2}=.09\) (see Figure 3).


Figure 3. Children's reaction times across encounters to the novel targets including during the initial fast mapping trials (First, Second, Third encounters) and the word learning test (Fourth encounter).

Further analyses revealed a significant linear trend for this interaction, \(F(1,37)=6.88, p=.01, \eta_{\mathrm{p}}{ }^{2}=.16\). Specifically, in the conditions where the competitors did not repeat across novel trials RTs decreased from the first to the fourth encounter. In contrast, in the conditions where the competitors did repeat across novel trials RTs became faster during referent selection (first three encounters) but slowed down again on the word learning test (forth encounter).

RTs to the fourth encounter (i.e., word learning test trial) were then compared. In the conditions where the competitors did repeat across novel trials, children were marginally slower on the word learning test trials ( \(M=5408.09 \mathrm{~ms}, S D\) \(=3870.36 \mathrm{~ms}\) ) than children in the other conditions ( \(M=\) \(3909.67 \mathrm{~ms}, S D=1100.29 \mathrm{~ms}), t(25.91)=1.79, p=.084\), two-tailed, \(d=0.60\). This difference in RTs is interesting because it suggests that children in the different groups were processing the names and objects differently. Children in the conditions with high task demands (competitors did not repeat across novel name referent selection trials) may have been simply guessing on these trials-hence the quick RTs. In contrast, children in the conditions with low task demands (competitors repeated across novel name trials) may have been committed to the task and searching for the correct referent-hence the slower RTs. Overall, these data demonstrate that lowering the attentional demands through repeating contexts (competitors) facilitated word learning and influenced processing as measured by reaction times.

\section*{Discussion}

Word learning is typically a gradual process as children experience the statistical regularity of the co-occurance of a name and object across repeated encounters (Smith \& Yu, 2008). Repeated encounters also facilitate retention for novel name-object associations (e.g., Gurteen et al., 2012). In everyday life, children may be exposed to some names and objects multiple times in the same context (e.g., a rubber duck in the bathtub). When a new word and its referent are repeatedly encountered in the same context, the non-targets may become increasingly redundant and predictable, freeing up attentional resources for processing the target referent (for a review and discussion of how this relates to research using storybooks, see Horst, 2013).

The current study investigated the effect of contextual repetition in referent selection trials on word learning. The results demonstrate that children presented with contextual repetition (i.e., the same competitors across multiple novel name referent selection trials) were significantly better at learning the target words than children who did not receive contextual repetition. Repeating competitors across trials appears to facilitate the initial encoding of the novel targets, enabling robust learning of the name-object associations. Children may have also associated the targets with the context of the non-targets (e.g., sprocks occur with elephants and cups), as they likely do with real-life objects (e.g., rubber ducks occur with bathtubs). One could argue that this gave targets on the contextual repetition trials a "figure-ground advantage," however, this is an unlikely
explanation as the objects were all presented on a plain white background as in previous studies (e.g., Schafer, 2005), rather than as part of a real scene.

We also found a difference in reaction times between conditions in which the competitors repeated across novel trials and conditions in which the competitors did not repeat. Although all children responded faster across successive encounters with the novel target, the children who saw the same competitors across novel name trials slowed down on the word learning trials when the novel target was presented with the other novel objects. This group may have been meaningfully considering among the possible test alternatives while those who saw different competitors may have been responding randomly at test, resulting in speedier responses. Although all children received the same number of exposures to the target items, only those for whom the context repeated across exposures demonstrated significant word learning.

Other studies, however, indicate that variability across learning encounters may facilitate learning. (e.g., Thiessen, 2011). An important consideration, however, is what is varying and what is remaining constant (i.e., repeating). Previous studies have typically found an advantage for variability at the target-level across encounters. Here, we demonstrate an advantage for reducing variability at the non-target-level (i.e., the competitors). Future research is needed to investigate when and how variability and reduction of attentional demands work together to facilitate robust word learning over both short- and long-term timescales. Clearly, however, for word learning via fast mapping, reducing the attentional demands of the initial referent selection task facilitates subsequent word learning.

\section*{Acknowledgments}

This research was funded by ESRC grant number RES-000-22-4451 to J.H. We would like to thank Ruth Pearce for assistance in data collection and Emilly J. Scott for serving as our narrator and help with ePrime. We also thank the members of the WORD Lab for their help in recruiting and scheduling participants and the parents and children who participated.

\section*{References}

Axelsson, E. L., Churchley, K., \& Horst, J. S. (2012). The right thing at the right time: Why ostensive naming facilitates word learning. Frontiers in Developmental Psychology, 3(88), 1-8.
Bion, R. A. H., Borovsky, A., \& Fernald, A. (2013). Fast mapping, slow learning: Disambiguation of novel wordobject mappings in relation to vocabulary learning at 18 , 24, and 30 months. Cognition, 126(1), 39-53.
Carey, S. (1978). The child as word learner. In Halle, J. Bresnan \& A. Miller (Eds.), Linguistic Theory and Psychological Reality (pp. 264-293). Cambridge, MA: MIT Press.

Dollaghan, C. (1985). Child meets word: Fast mapping in preschool children. Journal of Speech and Hearing Research, 28(3), 449-454.
Gurteen, P. M., Horne, P. J., \& Erjavec, M. (2011). Rapid word learning in 13- and 17-month-olds in a naturalistic two-word procedure: Looking versus reaching measures. Journal of Experimental Child Psychology, 109, 201-217.
Halberda, J. (2003). The development of a word-learning strategy. Cognition, 87(1), B23-B34.
Horst, J. S. (2013). Context and Repetition in Word Learning. Frontiers in Developmental Psychology, 4 (149), 1-11.

Horst, J. S., Parsons, K. L., \& Bryan, N. M. (2011). Get the story straight: contextual repetition promotes word learning from storybooks. Frontiers in Developmental Psychology, 2, 1-11.
Horst, J. \& Samuelson, L. (2008). Fast mapping but poor retention by 24-month-old infants. Infancy, 13, 128-157.
Horst, J. S., Samuelson, L. K., Kucker, S. C., \& McMurray, B. (2011). What's new? Children prefer novelty in referent selection. Cognition, 118, 234-244.
Horst, J. S., Scott, E. J., \& Pollard, J. A. (2010). The role of competition in word learning via referent selection. Developmental Science, 13, 706-713.
Mather, E., \& Plunkett, K. (2009). Learning words over time: The role of stimulus repetition in mutual exclusivity. Infancy, 14(1), 60-76.
Mather, E., \& Plunkett, K. (2012). The role of novelty in early word learning. Cognitive Science, 36, 1157-1177.
Schafer, G. (2005). Infants Can Learn Decontextualized Words Before Their First Birthday. Child Develeopment, 76, 87-96.
Schafer, G., \& Plunkett, K. (1998). Rapid word learning by fifteen-month-olds under tightly controlled conditions. Child Development, 69(2), 309-320.
Smith, L. B., \& Yu, C. (2008). Infants rapidly learn wordreferent mappings via cross-situational statistics. Cognition, 106, 1558-1568.
Thiessen, E. D. (2011). When variabilty matters more than meaning: the effect of lexical forms on use of phonemic contrasts. Developmental Psychology, 47, 1448-1458.
Wilkinson, K. M., Ross, E., \& Diamond, A. (2003). Fast mapping of multiple words: Insights into when "the information provided" does and does not equal "the information perceived". Applied Developmental Psychology, 24, 739-762.
Woodward, A. L., Markman, E. M., \& Fitzsimmons, C. M. (1994). Rapid word learning in 13- and 18 -month-olds. Developmental Psychology, 30(4), 553-566.

\title{
On The Possibility of an Extended, Yet Reductionist Mind
}

\author{
Saray Ayala (saray@mit.edu) \\ Department of Humanities: Philosophy, Language \& Literature University Carlos III of Madrid - Madrid 126, 28903 - Getafe. Madrid. Spain \\ Department of Linguistics \& Philosophy, Massachusetts Institute of Technology \\ 77 Massachusetts Avenue, 32-d808, Cambridge, MA. 02139-4307. USA
}

\begin{abstract}
In the debate around the extended mind, the special alliance that the extended thesis often has with functionalism usually plays in favor of the former, with functionalism providing support for the extended thesis. Here I want to consider this alliance in the opposite direction: does the extended thesis provide support for functionalism by promoting the need of a level of explanation that is independent of implementational (in particular neural) details? In spite of a seemingly promising line of reasoning for an affirmative answer, I show here that a commitment to the extended thesis or any version of externalism neither paves the way for a functionalist (or any other anti-reductionist) position nor is incompatible with an explanatory reductionism about the mind. I arrive to that conclusion after analyzing an argument by van Eck et al. (2006) meant to conclude the opposite, and showing why it is unsound.
\end{abstract}

Keywords: externalism; reductionism; functionalism; Marr's computational level.

\section*{Extended, therefore abstracted away from implementational details?}

When the thesis that the mind extends beyond the limits of the head (in all its different versions: extended mind, extended cognition, environmentalism, wide computationalism, etc. Extended thesis for short) is developed together with some version of functionalism, often the latter provides support for the former, by guaranteeing the legitimacy of a coarse-grain level of analysis where the implementational particularities do not matter (e.g., Clark \& Chalmers, 1998; Clark, 2010). According to this extended functionalism, the details of implementation (be it inside of outside the head) are not what decide the mental status of a structure or process, but rather the causal role it plays in the total economy of the system. Thus, the functionalist stance paves the way for an extended mind. Sometimes, though, this union becomes a reason for criticism of the Extended thesis. One strong line of criticism against some versions of the Extended thesis that draw on this alliance, claims that there is no such substantive theory about the (extended) location of minds, but a mere consequence of functionalism (see Wheeler, 2010), which, in fact, does not care about where the implementational base is located. Thus, it seems that the

Extended thesis would be better off by elaborating an argument that does not rely on functionalist justifications. In line with this predicament, a second wave of the Extended thesis has been developed in the last years (Sutton, 2010; Wilson, 2010).

Instead of the beneficial or not so beneficial role that functionalism can play for extended theorizing, here I address the opposite possibility. My goal is to test the potential for externalist considerations in general to legitimize the autonomy of a level of explanation above implementational details, and therefore, to provide support to anti-reductionist views like functionalism. My motivation is the apparent appeal of a line of reasoning that takes externalist considerations to justify the autonomy of such a level, and so it justifies a functionalist position where what matters for mentality is not the physical particularities of a state but its functional profile, usually described in computational terms.

That apparently appealing reasoning goes as follows: "if environmental structures play a critical role in mental processes, then our explanations of them cannot be limited to neural descriptions, since we need to account for those external elements. Therefore in order to explain cognition we need to approach it from a higher level of explanation that abstracts away from implementational details. A perfect candidate for this is a computational level."

This argument says that the inclusion of external elements shows the necessity of a coarse-grain level of analysis. And so it can be used to argue for an anti-reductionist position. In Clark's version of the extended thesis (Clark \& Chalmers, 1998), this coarse-grain level responds to the need to account for what makes internal and external elements mental, which is their computational role in the total functional economy of the system. Here I want to show why this apparently appealing reasoning is wrong. My conclusion is that neither the Extended thesis nor in general any version of (vehicle or content) externalism provide support for the autonomy and legitimacy of an independent-of-implementational-details explanation of mental phenomena. I will focus on a particular case where the above argument has been applied, that is the work of van Eck, Looren de Jong \& Schouten (2006).

\section*{The environment in visual perception}
van Eck et al. (2006) provide us with an argument to dismiss explanatory reductionism on the basis of empirical research in visual perception suggesting the significance of the environment in vision, and in general in cognitive processes. Their conclusion is that the inclusion of the environment in cognitive process shows the irreducible nature of psychological explanations operating at the computational level, since only at that level we can account for the environmental elements.

Some relatively recent studies give support to the idea that vision is not exclusively an internal process (e.g., Ballard et al., 1997). The phenomena of change blindness (Rensink et al. 2000; Simons, 2000) and inattentional amnesia (Wolfe, 1999) show, contra what traditional representationalist theories claimed (Marr, 1982), that subjects do not build a complete and detailed representation of the outside world. Instead, they rely on the stability of the environment, exploiting its resources. These recent findings could be said to be sympathetic with Gibson's idea that the ecological context is a central component in visual perception (Gibson, 1972). Although representations are still on the arena and Marr's (1982) classical computational theory of vision is not challenged yet, these findings point towards a more direct relation of the perceiver with her environment: the subject is not representing the whole scene (only part of it), but constantly consulting it, as if the information needed to successfully perceive, was, as Gibson defended, (at least partly) already there in the environment.

Norman (2002) proposes a theory where, apparently, Marr's emphasis on representations and informationprocessing on the one hand, and Gibson's emphasis on action and environment, can be reconciled. Norman's Dual Process Approach exploits the anatomical distinction between two brain pathways serving two different functions. The dorsal pathway is for action, while the ventral pathway is for representation (this approach resembles the perception-action model developed by Milner \& Goodale (1995), where vision is said to have two different functions, carried out by two separate brain structures.)

We are interested in how van Eck et al. examine Norman's approach and conclude it is incorrect. The findings mentioned above, point towards outside the brain, while Norman's approach restricts the action-related aspect of vision to the functioning of the dorsal pathway within the brain. That is, while Norman's proposal recognises that visual perception also serves to guide motor behaviour (perception is here, as Gibson argued, related to action), it is still internalist, situating the ingredients of perception inside the head. Gibson's direct perception proposal, however, includes the organism's environment, and it is this environmental factor what van Eck et al. want to rescue from inside the head.
van Eck et al. draw on Norman's mistake, presenting it as a confusion between different levels of explanation. Norman seems to be missing the externalist implications of
recognizing the importance of the environment. In van Eck et al. proposed clarification I find, however, two mistakes. On the one hand, they confuse ontological and explanatory issues; on the other hand, they make a controversial identification between Marr's computational level and Gibson's emphasis on the environment.

\section*{Internalism vs. Externalism}

Norman is wrong in keeping everything inside the head. Ecological perception (Gibson's proposal) is not the creation of action-based representations. It is not (only) about internal mechanisms. In van Eck et al's opinion, "Norman ignores that environment constitutes an additional level above internal processes" (2006, p. 21). According to them, ecological Gibson-like theories cannot be equated with neurophysiological accounts of the dorsal system. Gibson's view "is much broader than the level of (action) representations, because ecological perception also encompasses the environment" (ibid, p. 21). van Eck et al. are here moving the debate towards internalism/externalism. Are mental processes exclusively determined by intraorganism facts, or also by external factors? Gibson's view and the findings mentioned above seem to support some version of an externalist position. Therefore any account of visual perception solely in terms of internal mechanisms is missing something (e.g. this account will be unable to explain why the phenomenon of, for example, change blindness, happens). The internalist mistake lies in ignoring that other level, the environment. van Eck et al.'s conclusion is that the environment constitutes a different level of explanation, distinct from the internalist one.

This reasoning paves the way for their anti-reductionist conclusion. To account for that extra level we need a discipline that is not confined to lower-level processes; a science that is concerned with mental processes and is not restricted to neural, and therefore, internal processes. That is psychology. Psychology explains mental processes at a higher-level (describing them first at a computational level), where not only neural, but also elements outside the organism can be accounted for. Thus, this goes against any reductionist attempt to explain mental processes with a lower-level science like neuroscience. Thus, findings supporting Gibson's view are said to support an antireductionist project where psychology is necessary. Let's summarize van Eck et al.'s argument in bullet points:
(i) Environment plays a critical role in (visual) perception; in order to understand this we need to consider organismenvironment interactions, and not only what is going on inside the organism.
(ii) To account for organism-environment interactions we need a description at a level wider than a merely internalist one, like Marr's computational level, where the function of (visual) perception is described.
(iii) Psychology provides this kind of description, since it draws on explanations at a level wider than that of neuroscience.
(iv) Psychology is therefore necessary to explain visual perception.

It is now time to consider this argument critically.

\section*{Two Confusions}

\section*{From externalism to anti-reductionism}

In my opinion, van Eck et al. 's argument fails as an antireductionist defence. First, they confuse two different discussions, that is, the internalism vs. externalism debate, with the reductionism vs. anti-reductionism debate. It is true, I believe, that they assume that many practices within neuroscience are internalist (e.g., Bickle, 1998; 2003). So in attacking internalism, their argument can count against internalist reductionists (like Bickle seems to be). But reductionism in general, as a not-necessarily-internalist program, is left untouched.

Once we admit that empirical findings suggest the critical role that environment plays in (visual) perception, we have to account for that environmental factor. This however does not in itself warrant the need for psychology. The need to break the limits of the skull in order to describe the explanandum does not warrant the necessity of a higher level of explanation. That is, premise (ii) is not as obvious as van Eck et al. seem to present it. We can hold an externalist position and at the same time opt for the lower level of explanation. To tell against reductionism van Eck et al. would need, besides empirical findings suggesting that environment is critical in visual perception (and in general in any mental process), a conceptual claim asserting that to account for the organism-environment critical interaction we need to do so in an independent-of-implementation level.

The key assumption in van Eck et al. 's argument, premise (ii) is, in my opinion, a non sequitur. We can recognise that the environment makes a critical contribution to cognitive processes, and we can even say that the environment is a constitutive part of the system (like some versions of the Extended thesis claim), and it would still be a different thing to claim that this contribution constitutes a different level that needs to be accounted for by an autonomous science that is independent of physical details.

Here it is relevant to bring up the distinction between ontological and explanatory questions. The former seeks the constitution of something, while the latter inquires into a proper explanation of something. The ontological question at issue here is what constitutes vision: processes exclusively inside the organism or also states and features outside the organism? The explanatory question asks for how best to explain vision, or what counts as a (good) explanation of it. The explanatory question that concerns us here asks how we should deal (assuming we accept there is such an ingredient) with the outside-the-organism dimension(s) of vision. A particular response to the ontological question does not (automatically) imply a
specific response to the explanatory question. These two questions have to be answered separately.

Thus, from a claim about the constitution of vision, it is a non sequitur to state that as a consequence vision has to be explained in a particular way. Moreover, the (explanatory) demand that an explanation of vision requires reference to the environment still does not entail that a description at Marr's computational level is necessary. In order to assert such an explanatory choice, an additional argument is required. The ontological proof (i.e. that environment is part of visual perception) is not sufficient, and neither is the explanatory demand (i.e. that environment has to be accounted for).

In conclusion, for van Eck et al.'s argument to work, in particular, for their premise (ii) to be true, they need to prove that the environmental factor needs to be accounted for in a different, higher level explanation, where the visual task is described in a functional way as an information-processing task, and without mention of the particular (physical) component parts. But they only provide evidence for premise (i), that is, for the fact that environment plays a significant role in vision. They rely on Marr's theory to claim that this critical role that environment plays has to be described at Marr's computational level, where the function of vision is described. It is debatable whether Marr's theory is internalist or externalist. In the former case, van Eck et al. 's argument fails, since, if visual processes are (taxonomically) located inside the organism, there is no need for this extra level of explanation, and then their next step towards psychology as the science that accounts for that level, does not follow. If we concede that Marr's theory is externalist, according to what I have said above about the two different types of questions (ontological vs. explanatory), they still have not provided any argument for their explanatory conclusion. It is one thing to maintain that environment is part of our visual processes (ontological assertion), and another, to state that to explain vision we need to do it at a higher level (explanatory claim). The former claim (within the internalist vs. externalist debate) does not imply the latter (which is part of the reductionism vs. antireductionism debate).

It is important to notice here that the distinction between ontological and explanatory questions does not map onto the two different debates (internalism/externalism, and reductionism /anti-reductionism). I am not assuming neither that the former debate is exclusively ontological, nor that the latter is solely an explanatory matter. On the one hand, externalist considerations might respond to an explanatory concern, i.e. what is the appropriate unity of analysis. On the other, the debate around reductionism might be presented as an ontological question, e.g. are psychological properties reducible to properties of their implementation basis? The argument I am examining here (i.e., van Eck's et al.'s), however, takes the ontological aspect of an externalist position to conclude against an explanatory reductionism. The reason why I find it misleading is not that this argument
drags externalism towards explanatory fields, but that it forces externalism to a particular explanatory position.

\section*{A wrong appropriation}
van Eck et al.'s argument also fails for another reason: they wrongly equate Gibson's emphasis on the environment with Marr's emphasis on an independent-of-the-physicaldetails description where the cognitive ability to be explained is described as an information-processing task. Let us pay some attention to premise (ii) again.

They claim that the environmental factor has to be explained at a level different from the strictly internal where mechanisms responsible for the task are accounted for. They could then simply rely on Marr's model reputation, as classical cognitive science has done, and claim that a description at Marr's computational level, where the function of the cognitive task is described, is necessary to understand that task. But they go further, and seem to pursue legitimating Marr's computational level. Empirical findings point towards the necessity of including the environment in our understanding of visual perception (what we said to be premise (i)), and this partly supports Gibson's proposal. Interestingly enough, in van Eck et al. 's argument these empirical findings are assumed to legitimate Marr's computational level, since it is at Marr's computational level where the interaction with the environment is approached (that is the level of organism-environment interaction, where the task at issue is described independently of the algorithmic and physical details).

Once they claim that we have an "ecological level" to be accounted for (that is, an environmental ingredient that, according to them, calls for a new, higher level of explanation above a purely neural one), Marr's upper level appears as the proper place for that explanation. After all, Gibson's theory is too simplistic, or so a fan of Marr's computational complexity would say, to account for the complexity of organism-environment interaction. And so, although it makes a good point about the importance of environment, Gibson's theory falls short to account for the information-processing complexity that is involved.

Thus, van Eck et al. go from empirical findings to the necessity of Marr's computational level, via the assimilation of Gibson's ecological level into Marr's three-level model. And this assimilation is what attracts my attention. Gibson's emphasis on the environment does not imply another level above the level of representations, but a completely different anti-representational account of perception. Gibson's ecological level, as van Eck et al. label it, is not meant to provide a competing, alternative account of the informationprocessing mechanisms of vision. Gibson's theory is a completely contrasting explanation of perception where there is no room for internal addition and manipulation of information.

It is this major difference between Gibson's and Marr's theories that invalidates van Eck et al's assimilation of Marr's computational level (where vision is decomposed as an information-processing task) with Gibson's ecological
level (where vision is considered as an interaction between organism and environment and senses are decomposed into their biological component parts). In my opinion, then, van Eck et al's are using Gibson's emphasis on the ecological level in an incorrect way. The ecological level has nothing to do with Marr's computational level, so it is not correct to use Gibson's ecological level (and the evidence supporting the importance of environment) to legitimate Marr's computational level and the necessity of psychology.

We see that premise (ii) hides an incorrect assumption. Proving the significant role that environment plays is not the same as legitimating Marr's computational level. Premise (ii) (i.e. to account for environmental contribution we need a description at Marr's computational level, where the function of -visual- perception is described) being incorrect, premise (iii) (i.e. psychology provides this kind of description) has no meaning, because it is unimportant whether psychology provides the kind of descriptions that are required at Marr's computational level. Premise (i), let's remember, is only concerned with the necessity of including the environment in our explanation, not with the necessity of providing a particular kind of explanation. Only premise (i) in van Eck et al.'s argument, where empirical findings are reported, reveals as arguably correct. And it is easy to see that empirical support for the importance of environment is not evidence for the necessity of psychology. Premise (i) by itself does not support their antireductionist conclusion.

\section*{Conclusion}

By analyzing van Eck et al.'s argument, I tried to show that externalist considerations, and the extended mind thesis in particular (understood as a proposal about the location of the object of study in our explanations of mental processes), neither guarantee nor promote an anti-reductionist methodology and the necessity of a higher level of explanation where mental processes are described at a computational level.

If my considerations above are on the right track, van Eck et al.'s anti-reductionist argument is a poor one. On the one hand, they are mixing explanatory and ontological questions, extracting an explanatory conclusion from an ontological claim. On the other hand, they wrongly equate Gibson's emphasis on the environment with Marr's emphasis on an independent-of-the-physical-details description where the cognitive ability to be explained is described as an information-processing task. The identification of the first mistake, so I have argued, reveals that they only provide evidence for the ontological claim, leaving the explanatory assertion without any defence. The second mistake draws on the former. The empirical findings suggesting that environment plays a decisive role in cognitive processes supports Gibson's emphasis on the environment, but does not necessarily legitimates Marr's computational level of explanation, where organism-
environment interaction is accounted for. By equating Gibson's emphasis on the environment with Marr's computational level, they are trying to provide the functionalist sympathy for a description at Marr's upper level with empirical support. If their strategy were right, empirical findings pointing towards the significance of the environment would give support to the (computational) functionalist claim that visual perception, and mental processes in general, need to be approached as a platformfree information-processing task.

The empirical findings they mention might or might not be said to be evidence for Gibson's theory. What I think is clear is that they do not give any support to the claim that cognitive abilities need to be described at Marr's computational level. Their anti-reductionist conclusion is not well supported by their argument. It would require more and different argumentation to go from ontological claims about the location of cognitive processes to an explanatory anti-reductionism (e.g. some claim linking the inclusion of environment in explanations of mental processes to the need of accounting for it in an independent-of-the-implementational-details way).

An extended mind, and in general an externalist mind where environmental factors are a key ingredient, does not necessarily call for a high-level of explanation á la Marr's computational level. In our quest for cognition, we can go out of the head and still expect to satisfy our explanatory demands at the implementational level.

\section*{Acknowledgements}

This work has been developed with the support of the Alliance 4 Universities Program and the Department of Humanities: Philosophy, Language and Literature at the University Carlos III of Madrid. Special thanks to TECNOCOG Research Group at Universitat Autònoma de Barcelona.

\section*{References}

Ballard, D., Hayhoe, M., Pook \& Rao, R.P. (1997). Deictic Codes for the Embodiment of Cognition, Behavioral and Brain Sciences, 20, 4: 723-767.
Bickle, J. (1998). Psychoneural Reduction: The New Wave. Cambridge, MA: MIT Press.
Bickle, J. (2003). Philosophy of Neuroscience: A Ruthlessly Reductive Approach. Dordrecht, Netherlands: Kluwer Academic Press.
Clark, A. (2010). Memento's revenge: The extended mind extended. In R. Menary (Ed.), The extended mind (pp. 4366). Cambridge: MIT Press.

Clark, A. and Chalmers, D. (1998). The extended mind. Analysis, 58, 7-19.
Gibson, J. J. (1979). The Ecological Approach to Visual Perception. Houghton Mifflin: Boston.
Marr, D. (1982). Vision, San Francisco: W.H. Freeman.

Milner, A. D. \& Goodale, M. A. (1995) The Visual Brain in Action, Oxford: Oxford University Press.
Norman, J. (2002). "Two Visual Systems and Two Theories of Perception: An Attempt to Reconcile the Constructivist and the Ecological Approaches", Behavioral and Brain Sciences, 25: 73-144.
Rensink, R. A., O’Regan, J. K. \& Clark, J. J. (2000). "On the Failure to Detect Changes in Scenes Across Brief Interceptions", Visual Cognition, 7: 127-46.
Simons, D. J. (2000). "Current Approaches to Change Blindness", Visual Cognition, 7:1-15.
Sutton, J. (2010). Exograms and interdisciplinary: history, the extended mind and the civilizing process. In R. Menary (Ed.), The extended mind (pp.33-81). Cambridge: MIT Press.
van Eck, D., Looren de Jong, H. \& Schouten, M. (2006). "Evaluating New Wave Reductionism: The case of Vision", British Journal for Philosophy of Science 57: 167-196.
Wilson, R. (2010). Meaning making and the mind of the externalist. In R. Menary (Ed.), The extended mind (pp. 167-188). Cambridge: MIT Press.
Wolfe, J. M. (1999) "Inattentional Amnesia", in V. Coltheart (ed.), Fleeting memories, Cambridge, MA:MIT Press, pp. 71-94.

\title{
An Oscillator Model of Categorical Rhythm Perception
}

\author{
Rasmus Bååth (rasmus.baath@lucs.lu.se) \\ Erik Lagerstedt (drattans@gmail.com) \\ Peter Gärdenfors (peter.gardenfors@lucs.lu.se) \\ Lund University Cognitive Science \\ Lund University, Kungshuset, Lundagård, 22222 Lund, Sweden
}

\begin{abstract}
Categorical perception is a well studied phenomenon in, for example, colour perception, phonetics and music. In this article we implement a dynamical systems model of categorical rhythm perception based on the resonance theory of rhythm perception developed by Large (2010). This model is used to simulate the categorical choices of participants in two experiments of Desain and Honing (2003). The model is able to accurately replicate the experimental data. Our results supports that resonance theory is a viable model of rhythm perception and they show that by viewing rhythm perception as a dynamical system it is possible to model properties of categorical perception.
\end{abstract}

Keywords: Categorical perception; rhythm perception; dynamical systems; resonance theory

\section*{Introduction}

Categorical perception occurs when categorization is amplified by the perceptual systems so that distances within a category are perceived as being smaller and distances between categories are perceived as larger than they are according to the values of some physical measurement. It is a common phenomenon that is well studied in, for example, colour perception, phonetics and music (Harnad, 1990). A central question for understanding categorical perception is: What is the underlying mechanism? There have been attempts to model categorical perception in terms of neural networks (Damper \& Harnad, 2000). In this article, we will focus on modelling categorical rhythm perception and present a model that is based on oscillators.

In the field of music perception rhythm refers to a temporal pattern of sound onsets. A rhythm in this sense does not have to be periodic or recurrent. This is in contrast with how that word is used in other fields (cf. circadian rhythm or delta rhythm). A related concept that does involve periodicity is beat. When listening to a piece of music a common response is to move one's body with a perceived periodic pulse (Snyder \& Krumhansl, 2001), that pulse is the beat of the corresponding piece of music. It is not common that all beats in a piece of music are perceived as being equally accented and a periodically recurring pattern of strong and weak beats is called a meter. For example, a duple meter would imply that every second beat is perceived as having a stronger accent while every third beat is perceived as having a stronger accent in the case of a triple meter. Rhythm perception and the ability to entrain to a musical beat was long thought to be uniquely human and, while it has recently been shown that some vocal mimicking species are, to some degree, able to move in synchrony with a beat (Schachner, Brady, Pepperberg, \& Hauser,
2009), humans are still unique in their aptitude for rhythmic processing. Already infants have been shown to have a sense of rhythm (Honing, Ladinig, Háden, \& Winkler, 2009) and there exists only one documented case of "beat deafness" (Phillips-Silver et al., 2011), that is, the inability to reliably synchronize to a musical beat.

Desain and Honing (2003) showed in two experiments that listeners reliably experienced rhythms as belonging to rhythmic categories and that categorizations were strongly influenced when the listeners were primed with a metric beat before hearing a rhythm. Furthermore, participants agreed to a large degree on which rhythms belonged to what category and, similar to categorization of other kinds of stimuli (c.f. Jäger (2010) on colour categories), the categories were found to be roughly convex with respect to a temporal performance space (Gärdenfors, 2000). Honing (2013) concludes that: '"It is puzzling, however, that although meter was shown to exert a strong influence on the recognition of rhythm [...] existing computational models of meter can explain this phenomenon only to a small extent". In this article we show that an oscillation based, resonance theory model of rhythm perception (Large, 1996, 2010) can replicate many of the findings of Desain and Honing (2003). Our results support the notion that resonance theory is a viable model of rhythm perception and show that by viewing rhythm perception as a dynamical system it is indeed possible to model the properties of categorical rhythm perception.

\section*{Resonance Theory and Categorical Rhythm Perception}

Modelling of human timing and rhythm perception has a long history. One influential model is the one described by Wing and Kristofferson (1973), which is based on an information theoretic perspective. Like many such models (cf. Repp, 2005), it models a participant's behaviour when attempting to elicit isochronous timing responses. An alternative to this information theoretic perspective is to take a dynamical systems perspective and model time and rhythm perception as an emergent, dynamic phenomenon. A number of models of this kind have been proposed (e.g., Large, 1996; Noorden and Moelants, 1999). Here, the term resonance theory (cf. Large, 2010) will be used to refer to this type of models. Resonance theory does not dictate a specific model but rather incorporates a number of related models. All can be considered as dynamical system models and they consist of one or more resonating oscillatory units. Resonance theory provides a compelling framework since it is biologically plausible, has


Figure 1: (a) and (b) show examples of two possible rhythms and their placement in the triangular performance space (c) defined by Desain and Honing (2003). All one second long, four sound rhythms can be represented as a point in this space. (From Honing, 2013 with permission).
a solid base in dynamical systems theory and is able to model many aspects of meter and rhythm perception (Large, 2000). The general idea of resonance theory is that an external auditory rhythm can be represented by the rhythm of internal oscillatory units. These oscillatory units are coupled to the external rhythm and are by definition periodic while the external rhythm does not have to be periodic. Given a rhythm sequence as input the basic output of a resonance theory model, or resonance model for short, is the amplitude response of the oscillators over time. This high dimensional representation might be difficult to work with directly, however, and a more convenient representation is given by creating an \(a c\) tivation pattern, \(A\), by summing the amplitude responses of each oscillator over time, as in
\[
\begin{equation*}
A_{i}=\sum_{t=t_{s}}^{t_{e}} a_{i, t} \tag{1}
\end{equation*}
\]
where \(a_{i, t}\) is the amplitude for oscillator \(i\) at time \(t\) while \(t_{s}\) and \(t_{e}\) are the start and end time steps for the summation. Before the resonance model is given any input it is in a resting state and it takes a number of time steps before the system is activated by the stimuli. Therefore it is not necessarily desirable to sum over the whole extent of the duration of the rhythm sequence and an activation pattern created by summing over the latter time steps may represent the rhythm sequence better than an activation pattern created by summing over all time steps.

While not all resonance theory models claim biological plausibility, a number of neuroimaging studies have shown connections between neural resonance and rhythm perception
(e.g., Brochard, Abecasis, Potter, Ragot, and Drake, 2003). One persuasive study that clearly showed that rhythm perception involves neural oscillatory activity is that of Nozaradan, Peretz, Missal, and Mouraux (2011). They found that playing a rhythmic beat to a participant elicited a sustained periodic neural response, as measured by EEG, that matched the frequencies of the beat. Resonance theory models differ in their biological plausibility, the number of oscillatory units employed and the type of oscillators used. Eck (2002) constructed a model with a clear biological connection as it used the Fitzhugh-Nagumo model of neural action potential (Nagumo, Arimoto, \& Yoshizawa, 1962) as the oscillatory unit. Other models claim no biological plausibility, for example, the model by Scheirer (1998) that employs a comb filter as the oscillatory unit. Toiviainen and Snyder (2003) modelled participants behaviour when tapping along to excerpts of music composed by Bach using a single oscillatory unit while Large and Kelso (2002) used a bank of 96 oscillators to model participants' tapping to ragtime music.

To our knowledge, resonance theory models have not previously been used to model categorical rhythm perception. One reason for this might be that while the amplitude response of the oscillators in the resonance model reflects, perhaps even represents, the rhythm sequence given as input to the system it does not give rise to a categorization per se. That is, while the state the resonance model arrives at depends on the given rhythm sequence, there is no finite number of discrete states that can be said to constitute categories. Still, the state of the resonance model can be used as the basis of a categorical decision based on learnt prototype states or a discrete partitioning of the system state space. By considering the activation pattern of a resonance model as a point in an \(n\)-dimensional space, \(n\) being the number of oscillatory units, this space can be partitioned into regions corresponding to rhythm categories and used to produce categorical decisions (following the general model of concepts from Gärdenfors (2000)). The relation between the activation pattern of a resonance model and such a rhythm categorization is analogous to the relation between the hue, saturation and lightness of a colour percept and a colour categorization. That is, a colour percept can similarly be viewed as a point in a three dimensional space with dimensions hue, saturation and lightness and this space can be partitioned into regions, each representing a colour category, and used to produce categorical colour decisions.

If the state of the resonance model is viewed as the basis for a categorical decision then two predictions regarding categorical rhythm perception can be made:
(1) More distinct states will facilitate categorization. Here a distinct state refers to a subset of oscillators in a resonance model having a strong amplitude response while most oscillators show a weak amplitude response. This is in contrast to a non-distinct state where most oscillators have a similar amplitude response, that is, there are many competing signals and there is no clear single winning candidate among


Figure 2: Maps over categorization consistency. (a) shows the entropy of the categorical choices for the participant given the same rhythm sequences multiple times from Desain and Honing (2003, used with permission). (b) shows the signal-to-noise measure calculated from the activation patterns generated by the resonance model.
the categories. In an experimental categorical task it would then be predicted that a participant would categorize a rhythm sequence more consistently, and with more confidence, if the sequence resulted in a more distinct state in a resonance model than if the sequence resulted in a less distinct state.
(2) Rhythm sequences resulting in similar states are categorized similarly. That is, different rhythm sequences resulting in similar states when used as the input to a resonance model should be categorized similarly by participants in an experimental task.

In order to test these predictions, data from a rhythm categorization task is needed. A study by Desain and Honing (2003) provides suitable experimental data to do so.

\section*{The Rhythm Categorization Study of Desain and Honing (2003)}

Desain and Honing (2003) employed a novel paradigm where participants were asked to categorize 66 different rhythm sequences by transcribing them into common music notation. The sequences all lasted for one second and consisted of four tone onsets and were therefore uniquely determined by the three interonset intervals (IOI) between the tones. Two such
possible sequences are shown in figure 1 a and 1 b where a possible categorization of 1 b could be \(\lrcorner \curvearrowright . \mathrm{d}\) ( or 1-1-2 when written as an integer ratio). Any possible one second, four tone rhythm sequence can be thought of as a point in a three dimensional triangular performance space that determines the lengths of the three IOIs as shown in figure 1. The 66 rhythm sequences were constructed so that they evenly covered the area in the performance space with the constraint that no IOI would be shorter than 153 ms . The location of these sequences in the performance space can be seen in figure 2 b where each circle marks the position of one of the \(66 \mathrm{se}-\) quences.

In a first experiment, 29 highly trained musicians categorized the rhythm sequences and the result was that even though the rhythms were performed on a more or less continuous scale, the participants tended to stick to a limited number of categories with 1-1-1 being the single most common. Twelve categories, all categories considered, stood out as being the most common and the location in performance space of these categories are shown in figure 3a. One participant was presented with the 66 rhythm sequences multiple times and as a measure of consistency the entropy was calculated of her responses for each rhythm. These entropy values were mapped on to the performance space and the resulting entropy map is shown in figure 2 a .

In a second experiment two meter conditions were added. Duple meter versions of the rhythms were constructed by prepending the rhythms with a repeated, one second long, two sound beat, thus inducing a \(2 / 4\) meter feeling. Triple meter versions of the rhythms were similarly constructed by prepending a three sound beat instead. This resulted in three different meter conditions: The original no meter condition, a duple meter condition and a triple meter condition. Maps over what categories the participants ascribed to the different rhythms, similar to the map shown in figure 3a, were constructed (shown on p. 358 in Desain and Honing, 2003). A main finding was that the participants' categorization in the no meter condition was significantly more similar to the participants' categorization in the duple meter condition than in the triple meter condition.

For the purpose of the current study, data from Desain and Honing was downloaded from a web resource containing supplementary material \({ }^{1}\). The data downloaded was the information regarding which of the twelve most common categories was most often ascribed to each of the 66 rhythm sequences for the no meter condition in experiment one and the duple and triple meter conditions in experiment two. A data point for a rhythm sequence was excluded if none of the twelve most common categories was the most common for that specific rhythm. Information regarding the categorization entropy for the participant presented with the rhythms multiple times was unfortunately not available from the web resource. This information was retrieved manually from figure 2 a .

\footnotetext{
\({ }^{1}\) http://www.mcg.uva.nl/categorization
}


Figure 3: Categorization maps for (a) the experimental data from Desain and Honing (2003, used with permission) and (b) the resonance model. The transparent areas in (a) indicate areas where there was a large amount of disagreement between the participants.

\section*{Resonance Theory and the Data of Desain and Honing}

It is possible to test the two predictions from resonance theory concerning how rhythms are categorized by implementing a resonance model that consists of an array of oscillators (as in Large, 2000). We used the rhythm stimuli from Desain and Honing (2003) as input to such a model and compared the results with the experimental data using the methods outlined below.

Prediction (1) implies that rhythm sequences resulting in distinct states in a resonance model should be the sequences that are categorized more consistently. In Desain and Honing's data, a measure of consistency is the categorization entropy for the participant presented with the rhythm sequences multiple times. The prediction is that this measure of consistency is correlated with a measure of distinctness of the state of a resonance model. Signal-to-noise ratio is a common measure of distinctness of a signal and a modified version of this measure can be used to quantify the distinctness of the state of a resonance model. For a resonance model that has been given a rhythm sequence as input, the activation pattern is first calculated according to equation (1). In this activation pattern, the signal \(A_{s}\) is defined as being the \(A_{i}\) with the highest amplitude. The signal-to-noise ratio is then defined
as:
\[
\begin{equation*}
S N R=\frac{A_{s}}{\sum_{i=1}^{n} A_{i}} \quad i \neq s \tag{2}
\end{equation*}
\]
where the sum in the denominator is over the rest of the \(A_{i}\) oscillator amplitudes. Notice that this measure of consistency should be negatively correlated with the entropy measure of Desain and Honing: As the signal gets weaker relative to the noise, the entropy of the participants choices of category should go up.

Prediction (2) implies that rhythm sequences resulting in similar states when given as input to a resonance model should be categorized similarly in an experimental task. A resonance model does not directly produce a categorization but this is not required for testing this prediction. It is possible to compare the resulting states of two rhythm sequences by calculating the respective activation patterns and comparing these. A suitable similarity measure is given by considering the activation patterns as points in an \(n\)-dimensional space, where \(n\) is the number of oscillators in the resonance model, and then taking the Euclidean distance between these two points, where a shorter distance corresponds to more similar states. Considering the twelve most common rhythm categories chosen by the participants in Desain and Honing's study as prototype categories, it is possible to use the rhythm sequences corresponding to these categories to generate prototype activation patterns. For example, to generate the prototype activation pattern for the category 1-2-1 (as shown in figure 4) the rhythm sequence with IOIs \(0.25 \mathrm{~s}, 0.5 \mathrm{~s}\) and 0.25 s would be used as input to the resonance model. A rhythm sequence's activation pattern can then be compared with these prototypes' activation patterns and the prototype category with the most similar activation pattern can be assigned to that rhythm sequence. In this way, all rhythm sequences can be assigned a category and these categories can be compared with the categories selected by the participants in Desain and Honing's study. Specific hypotheses are then that a resonance model categorization of the stimulus used by Desain and Honing should be similar to the categorization made by the participants in the no meter, duple meter and triple meter conditions. Furthermore, since the participants' categorizations of the rhythm sequences in the duple meter condition was more similar than the triple meter condition to the categorization in the no meter condition the same relation should be present in the categories generated by the resonance model.

\section*{The Setup of the Resonance Model}

The resonance model was implemented in MATLAB \({ }^{2}\) using the Nonlinear Time-Frequency Transformation Workbench (Large and Velasco, in preparation). The model consisted of 145 Hopf oscillators, a type of oscillator that entrains to periodic input and where the amplitude of an oscillator depends on that oscillator's intrinsic frequency and the periodicities

\footnotetext{
\({ }^{2}\) http://www.mathworks.se/products/matlab/
}


Figure 4: An example of an activation pattern generated by feeding the resonance model a rhythm with IOIs \(0.25 \mathrm{~s}, 0.5 \mathrm{~s}\) and 0.25 s .
of the input. The differential equation of the Hopf oscillator used is:
\[
\begin{gather*}
\frac{d z}{d t}=z\left(\alpha+i \omega+\frac{\beta \varepsilon|z|^{4}}{1-\varepsilon|z|^{2}}\right)+\frac{x}{1-\sqrt{\varepsilon} x} \cdot \frac{1}{1-\sqrt{\varepsilon} \bar{z}}  \tag{3}\\
\alpha=-0.1, \quad \beta=-0.1, \quad \varepsilon=0.5
\end{gather*}
\]
where \(\alpha\) is a damping term, \(\beta\) is an amplitude compression factor and \(\varepsilon\) is a scale factor. The last term in equation (3) is the resonant term, which is dependent on the stimulus \(x\). These parameter values and this specific formulation of the Hopf oscillator were not chosen on the basis of any specific theoretical considerations; many other configurations are possible and a more general form of the Hopf oscillator is derived in Large, Almonte, and Velasco (2010). The oscillators' intrisic frequencies were centred around 1 Hz with frequencies logarithmically distributed from 0.25 Hz to 4 Hz . The method used for creating activation patterns was that in equation (1) with \(t_{s}\) set to the time step corresponding to half the stimulus length and \(t_{e}\) set to the last time step \({ }^{3}\).

The 66 rhythm sequences from the no meter condition were encoded and given as input to the model yielding 66 activation patterns. This was repeated for the rhythm sequences from the duple and triple meter conditions. Additionally, the rhythm sequences of the prototype categories were encoded in the same way as the no meter condition sequences yielding twelve prototype activation patterns.

\section*{Results}

The signal-to-noise measure was calculated for all activation patterns in the no meter condition and, as predicted, a negative correlation between Desain and Honing's (2003) entropy measure of consistency and the signal-to-noise ratio (Pearson product-moment correlation, \(r=-0.33, p=0.006\) ) was found. The two measures of consistency are expected

\footnotetext{
\({ }^{3}\) The MATLAB code for the model and both input data and the resulting output are available on request from the first author. The code for the Nonlinear Time-Frequency Transformation Workbench (Large and Velasco, in preparation) has not yet been publicly released and has to be requested separately.
}
to have a reverse relationship, that is, low entropy in the experimental data should indicate high consistency, while a low signal-to-noise ratio in the simulated data should indicate low consistency. A comparison between these two measure of consistency is shown in figure 2 . To facilitate comparison, the colour scales have been matched so that red indicates low consistency while blue indicates high consistency. The measures of consistency are clearly comparable, showing the same broad patterns in both the simulated (figure 2b) and experimental data (figure 2a).

The activation patterns for all the three meter conditions were compared with the prototype activation patterns using Euclidean distance as the similarity measure and each rhythm sequence was assigned the category of the most similar prototype. A comparison with the categories assigned in the experimental task for the no meter condition is shown in figure 3. It is clear that the categorizations to a large extent agree. The \(1-1-1\) category is the most common in both the experimental and the simulated categorizations and both categorizations exhibit roughly convex category regions. A randomized permutation test \({ }^{4}\) also showed that the categorization generated by the resonance model and the categorization from Desain and Honing's data was more similar than would be expected by chance alone for all the three meter conditions. In the no meter condition (shown in figure 3) the agreement was \(71 \%\) ( \(p<0.001\) ) and in the duple and triple meter conditions \(67 \%\) ( \(p<0.001\) ) and \(61 \%(p<0.001)\) respectively.

In the experimental data, the categorization of the duple meter condition was more similar than the triple meter condition to the no meter condition and this was also the case for the simulated categorizations. The agreement between the no meter condition and the duple and triple meter conditions for the simulated categorizations was calculated as being \(77 \%\) and \(71 \%\) respectively with duple meter agreeing with the no meter categorization in 6 percentage points more of the cases ( \(p=0.045\), one-tailed randomized permutation test).

\section*{Conclusions}

Many models of categorical perception have been based on neural networks and there exist several models of rhythm perception based on neural networks (Mozer, 1993; Miller, Scarborough, \& Jones, 1992). We believe that using a dynamical

\footnotetext{
\({ }^{4}\) Randomized permutation tests (Ernst, 2004) were used to compare the categorization of the rhythm sequences from the behavioural data with the categorization from the resonance model. Given two different categorizations of the 66 rhythms a similarity score, is calculated as the number of rhythms that are given the same category by both categorizations. In the cases when the most common categorization of a specific rhythm sequence in the behavioural data is not one of the twelve prototype categories this rhythm sequence is excluded from further analysis. Next, all category labels are randomly assigned to different rhythm sequences and a new similarity score is calculated. This is repeated 10,000 times, yielding a randomized permutation distribution of similarity scores. A p-value is then calculated as the probability of achieving the actual similarity score, or a similarity score being more extreme, given the randomized distribution of similarity scores. The permutation tests were two-tailed (calculated according to the method in Ernst, 2004) in all cases except when noted.
}
system of resonating oscillators provides a more principled way of modelling such phenomena. By modelling rhythm perception in such a system, we have shown that it is possible to explain empirical findings of listeners' categorical perception of rhythm. Our oscillator model has been able to accurately replicate the experimental data from Desain and Honing (2003).

An advantage of oscillator models is that they can be generalized to other kinds of categorical perception. Examples from the domain of music are pitch perception and tonality perception (Large, 2010). Oscillatory models are not confined to temporal processes and can be used for other modalities. The main importance of our model is perhaps that the example of how oscillator models can be constructed for categorical rhythm perception can serve as inspiration for similar models of other cognitive phenomena.

\section*{Acknowledgements}

The authors gratefully acknowledges support from the Linnaeus environment Thinking in Time: Cognition, Communication and Learning (CCL), Swedish Research Council grant number 349- 2007-8695.

\section*{References}

Brochard, R., Abecasis, D., Potter, D., Ragot, R., \& Drake, C. (2003). The "ticktock" of our internal clock : direct brain evidence of subjective accents in isochronous sequences. Psychological Science, 14(4), 362-366.
Damper, R., \& Harnad, S. (2000). Neural network models of categorical perception. Attention, Perception, \& Psychophysics.
Desain, P., \& Honing, H. (2003). The formation of rhythmic categories and metric priming. Perception, 32(3), 341365.

Eck, D. (2002). Finding downbeats with a relaxation oscillator. Psychological research, 66(1), 18-25.
Ernst, M. D. (2004). Permutation methods: a basis for exact inference. Statistical Science, 19(4), 676-685.
Gärdenfors, P. (2000). Conceptual Spaces: The Geometry of Thought. MIT press.
Harnad, S. (1990). Categorical Perception: The Groundwork of Cognition. Cambridge University Press.
Honing, H. (2013). Structure and interpretation of rhythm in music. In D. Deutsch (Ed.), The psychology of music (pp. 369-390). Elsevier.
Honing, H., Ladinig, O., Háden, G. P., \& Winkler, I. (2009). Is beat induction innate or learned? Probing emergent meter perception in adults and newborns using event-related brain potentials. Annals of the New York Academy of Sciences, 1169, 93-6.
Jäger, G. (2010). Natural color categories are convex sets. In Logic, language and meaning: 17th amsterdam colloquium, amsterdam. Springer.
Large, E. W. (1996). Modeling beat perception with a nonlinear oscillator. In Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society: July

12-15, 1996, University of California, San Diego (p. 420). Lawrence Erlbaum.
Large, E. W. (2000). On synchronizing movements to music. Human Movement Science, 19(4), 527-566.
Large, E. W. (2010). Neurodynamics of music. In M. Riess Jones, R. R. Fay, \& A. N. Popper (Eds.), Music Perception (pp. 201-231). Springer.
Large, E. W., Almonte, F. V., \& Velasco, M. J. (2010). A canonical model for gradient frequency neural networks. Physica D: Nonlinear Phenomena, 239(12), 905-911.
Large, E. W., \& Kelso, J. A. S. (2002). Tracking simple and complex sequences. Psychological Research, 3-17.
Large, E. W., \& Velasco, M. (n.d.). Nonlinear TimeFrequency Transformation Workbench: MATLAB software for signal processing using nonlinear oscillators.
Miller, B., Scarborough, D., \& Jones, J. (1992). On the perception of meter. Understanding Music with AI.
Mozer, M. (1993). Neural net architectures for temporal sequence processing. In A. Weigend \& N. Gershenfeld (Eds.), Predicting the future and understanding the past. (pp. 243-264). Addison-Wesley.
Nagumo, J., Arimoto, S., \& Yoshizawa, S. (1962). An active pulse transmission line simulating nerve axon. Proceedings of the IRE, 50(10), 2061-2070.
Noorden, L. van, \& Moelants, D. (1999). Resonance in the perception of musical pulse. Journal of New Music Research, 28(1), 43-66.
Nozaradan, S., Peretz, I., Missal, M., \& Mouraux, A. (2011). Tagging the neuronal entrainment to beat and meter. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 31(28), 10234-40.
Phillips-Silver, J., Toiviainen, P., Gosselin, N., Piché, O., Nozaradan, S., Palmer, C., et al. (2011). Born to dance but beat deaf: A new form of congenital amusia. Neuropsychologia, 49(5), 961-969.
Repp, B. H. (2005). Sensorimotor synchronization: A review of the tapping literature. Psychonomic Bulletin \& Review, 12(6), 969.
Schachner, A., Brady, T., Pepperberg, I., \& Hauser, M. (2009). Spontaneous motor entrainment to music in multiple vocal mimicking species. Current Biology, 19(10), 831-836.
Scheirer, E. D. (1998). Tempo and beat analysis of acoustic musical signals. The Journal of the Acoustical Society of America, 103(1), 588-601.
Snyder, J., \& Krumhansl, C. L. (2001). Tapping to Ragtime: Cues to Pulse Finding. Music Perception, 18(4), 455-489.
Toiviainen, P., \& Snyder, J. (2003). Tapping to Bach: Resonance-based modeling of pulse. Music Perception, 21(1), 43-80.
Wing, A., \& Kristofferson, A. (1973). Response delays and the timing of discrete motor responses. Attention, Perception, \& Psychophysics, 14(1), 5-12.

\title{
Measuring Cultural Relativity of Emotional Valence and Arousal using Semantic Clustering and Twitter
}

\author{
Eugene Y. Bann (eugene@aeir.co.uk) \\ Advanced Emotion Intelligence Research (AEIR) \\ Joanna J. Bryson (J.J.Bryson@bath.ac.uk) \\ Department of Computer Science, University of Bath, United Kingdom
}

\begin{abstract}
Researchers since at least Darwin have debated whether and to what extent emotions are universal or culture-dependent. However, previous studies have primarily focused on facial expressions and on a limited set of emotions. Given that emotions have a substantial impact on human lives, evidence for cultural emotional relativity might be derived by applying distributional semantics techniques to a text corpus of self-reported behaviour. Here, we explore this idea by measuring the valence and arousal of the twelve most popular emotion keywords expressed on the micro-blogging site Twitter. We do this in three geographical regions: Europe, Asia and North America. We demonstrate that in our sample, the valence and arousal levels of the same emotion keywords differ significantly with respect to these geographical regions - Europeans are, or at least present themselves as more positive and aroused, North Americans are more negative and Asians appear to be more positive but less aroused when compared to global valence and arousal levels of the same emotion keywords. Our work is the first in kind to programatically map large text corpora to a dimensional model of affect.
\end{abstract}

Keywords: Semantic Clustering, Emotion Analysis; Twitter; Core Affect Model.

\section*{Introduction}

The question as to whether the experience and expression of emotions is universal or relative to specific cultures has resulted in a wide variety of studies, with theories ranging from the universality hypothesis to culture-specific facial expressions. Here we present evidence that culture is a necessary framework for researchers studying variation in emotions. Independent of the question of biological differences in the experience of emotions, it would be unsurprising if culture shapes our conscious perception, expression and experience of emotions, as has been hypothesised for other cognitive phenomena (Hunt \& Agnoli, 1991; Fuhrman et al., 2011). Here, we use Latent Semantic Clustering on an emotional text corpus mined from Twitter to discern how the primary properties normally attributed to emotional keywords - valence and arousal - differ as the keywords are used in the same language (English) as exploited across different global regions.

\section*{The Conceptualisation of Emotion Qualia}

Emotion qualia refers to the raw feel of an emotion. The actual phenomenon of a particular emotion experienced may differ according to each person's perception or understanding of that emotion, with perception being the result of the individual's past and hypothesised responses, unique to each human being. Barrett (2006) describes the act of conceptualising core affect, or in other words, why people attach emotion labels to the experience of emotion qualia. Since emotion
keywords are constructed from conceptual knowledge about the world, emotions themselves may be concepts that humans begin learning in infancy and continuously extend and revise throughout life (Lindquist \& Barrett, 2008). This repeated experience of labelling a combination of core affect and the context in which it occurs as an emotion provides training in how to recognise and respond to that emotion. In this sense, Barrett describes emotions as simulations. This skill of conceptualising core affect as an emotion could be a core aspect of emotional intelligence, in much the same way as conceptual thinking is core to cognitive intelligence. Each person learns the label in association with their unique experience, thus each person's conceptualisation of their emotional spectrum is unique. Cultures, formed of communicating individuals, may therefore also be unique if individual experiences vary somehow systematically. We base our analysis on this hypothesis. The reader should bear in mind that we are not analysing emotion keywords in particular, rather, we are analysing emotion conceptualisations, or what cultures understand specific emotion keywords to mean, using Latent Semantic Clustering to infer these meanings.

\section*{Core Affect}

Core affect is an emerging paradigm in affective neuroscience, and postulates a continuous approach to defining emotions (Posner et al., 2005). Several core-affect, or circumplex models have been proposed (e.g. Watson \& Tellegen, 1985; Russell, 1980; Cacioppo \& Berntson, 1994), yet all have one thing in common: they represent emotions as a single point in a continuous space defined by two (or rarely three) dimensions. Different labels have been assigned to these two dominant dimensions by various theorists, such as pleasure and engagement, however most commonly, valence and arousal are chosen. Thus far, there has been no attempt to computationally pinpoint emotions or documents within a core affect model using 'online' and 'big' data; to date, research regarding the core affect model has either been theoretical (e.g. Watson \& Tellegen, 1985), or conducted via a limited survey (e.g. Russell, 1980).

Core affect is one of two main theories regarding the representation of emotions, the other being the Basic Emotion model, however, neither has thus far received unequivocal support. Basic emotions could turn out to map to multiple subtypes of coherent emotion networks, but this implies we need to split basic emotion categories into further subtypes to better reflect these emotion networks (Hamann, 2012; Bann,
2012). Here we extend this view and suggest that the core affect model enables us to quantify the properties of the basic emotions themselves.

\section*{Previous Work}

There is growing evidence that aspects of a person's psychology can be predicted from their language usage. In the 1990s, human semantics was shown to be recoverable from linguistic corpra independent of any further grounding (Lowe, 1997; Bryson, 2008). Recent applications to individual psychology include discovering individual differences in personality (Pennebaker \& King, 1999), discovering cultural change in moral beliefs (Bilovich \& Bryson, 2008), as well as for emotion categorization (Fitzpatrick \& Logan, 2011). French discovered that co-occurrence techniques such as LSA does not detect personality from short text samples (French \& Gill, 2007), but do reveal that texts expressing particular emotions have a greater semantic similarity to corresponding exemplar words (Gill et al., 2008).

A recent study by Jack et al. (2012) found significant evidence that facial expressions are indeed culture-dependent; that is, different cultures represent the same emotions differently. However, whether or not this is because they experience different emotion qualia is another question. Using language, rather than facial expressions, as an accessor to emotion will enable a much more detailed and less ambiguous analysis, increasing significance by "throwing more data at the problem" (Recchia \& Jones, 2009, p.3).

Currently, there have been few attempts to analyse cultural differences using language semantics. Language plays a key role in how emotions are conceptualised (and thus perceived); Lindquist states "language can be no more removed from emotion, than flour can be removed from an already baked cake" (Lindquist, 2009, p.1). Recently, Bann \& Bryson (2012) demonstrated how conceptualisations of emotions can be inferred by performing Latent Semantic Analysis on a corpus of self-reported emotional tweets. Their DELSAR algorithm analysed 21,000 tweets each labelled with an emotion, and clustered each document in the corpus to its most similar corresponding emotion label using Latent Semantic Clustering. Here we use the same algorithm as the basis for our analysis.

\section*{Corpus}

Typing emotion keywords into the Internet is increasingly becoming a significant technique for individual expression. There now exists a rich available source of information about emotions on the Internet, because so many people spend time expressing how they feel in blogs, forums, social networking websites and the like. We use data from the microblogging website Twitter to perform large-scale analysis of the language used in thousands of expressions of emotions within tweets. Acquiring a significantly larger corpus than Bann \& Bryson (2012), we use the Gardenhose level of Twitter's
streaming API \(^{1}\) to create a corpus of \(5,625,844\) tweets \(^{2}\) collected between 19th October 2012 and 18th January 2013. Each emotion keyword (see selection criteria below) is given a five-minute streaming window in turn for the duration of the period, ensuring an even temporal distribution of Tweets is collected. Table 1 describes our corpus, split by 'cultural' region. We use the tweet's timezone as an indication of the corresponding user's geographical location; seeing as it is very unlikely that a Twitter user would select a timezone other than that which they reside in, it is somewhat safe to assume that this reflects the cultural origin of each user.

Table 1: Distribution of tweets within our corpus.
\begin{tabular}{lrrrr}
\hline Emotion & Asia & Europe & NA & All \\
\hline Angry & 12194 & 27070 & 61293 & 200024 \\
Ashamed & 1008 & 5097 & 17107 & 46486 \\
Calm & 5975 & 10181 & 36681 & 102827 \\
Depressed & 3078 & 11615 & 43129 & 120473 \\
Excited & 30923 & 100792 & 292822 & 847679 \\
Happy & 149129 & 186709 & 730839 & 2201874 \\
Interested & 3527 & 9728 & 31891 & 86763 \\
Sad & 46351 & 83075 & 341912 & 966165 \\
Scared & 15435 & 42500 & 194130 & 517715 \\
Sleepy & 26031 & 10787 & 120473 & 290666 \\
Stressed & 2587 & 8774 & 41716 & 109295 \\
Surprised & 3032 & 12454 & 56332 & 135877 \\
\hline Total & \(\mathbf{2 9 9 2 7 0}\) & \(\mathbf{5 0 8 7 8 2}\) & \(\mathbf{1 9 6 8 3 2 5}\) & \(\mathbf{5 6 2 5 8 4 4}\) \\
\hline
\end{tabular}

Region definitions. We only include those timezones that have over 5000 tweets within our corpus. The Asia region consists of the timezones Kuala Lumpur, Beijing, Singapore, Jakarta, Bangkok, Hong Kong, Tokyo; the Europe region consists of the timezones London, Amsterdam, Athens, Edinburgh, Dublin, Berlin, Paris; the North American (NA) region consists of the timezones Eastern Time (US \& Canada), Central Time (US \& Canada), Mountain Time (US \& Canada), Pacific Time (US \& Canada).
Selection of emotions. As opposed to strictly using the basic emotions as identified by Bann \& Bryson (2012), we use the most popular emotions that are used on Twitter, that is, those emotions that have the highest stream rate. Twelve emotions were selected that had a high rate and that equally divided into positive/negative and engaged/disengaged theoretical categories (see Table 2).
Subcorpus creation. Each subcorpus is created using a limit of 1000 documents per emotion for all subcorpora to ensure consistency within our results; we chose 1000 as it is the lowest value in Table 1. To mitigate micro-temporal effects, if the number of documents for a particular emotion is significantly greater than 1000, we use a modulus function to extract 1000

\footnotetext{
\({ }^{1}\) https://dev.twitter.com/docs/streaming-apis.
\({ }^{2}\) Having first removed 34,725 duplicate tweets. Corpus and code is available to download at www.aeir.co.uk/code.
}
documents equally spaced across the subcorpus - for example, if a particular emotion in a particular subcorpus has 6000 documents, we take one document every six documents. We also create six control subcorpora so to compare our regionspecific results with a baseline. We use the same modulus function to extract 1000 equally spaced tweets, but without any timezone clause, selecting six random starting points.

\section*{Proposed Method}

We use DELSAR (Bann \& Bryson, 2012, see Algorithm 1) to generate the clustering matrix for each subcorpus - the three regions Asia, Europe and NA, and six random controls.
```

Algorithm 1 DELSAR
Require: Corpus $\mathbf{C}$ and Keyword Set $\mathbf{K}$, where each docu-
ment in $\mathbf{C}$ is mapped to one emotion keyword, emotion, in
$\mathbf{K}$ (through corpus generation)
Generate cosine document similarity matrix of $\operatorname{LSC}(\mathbf{C}, \mathbf{K})$
(document $\times$ document similarity matrix)
for each emotion $\in \mathbf{K}$ do
for each document that has emotion emotion do
delete emotion within the document
Find the closest document nearest where nearest $\neq$
document
Increment the count for the emotion that nearest is
labelled as in emotion_vector
end for each
return emotion_vector
end for each

```

For each subcorpus, DELSAR uses LSA (Landauer \& Dumais, 1997) to create a document-document matrix of cosine similarities (Similarity Matrix), in which similar documents are closer to one (i.e. the cosine of the angle between their vectors). It creates a clustering matrix that represents the corpus as an emotion-emotion matrix, describing how each emotion is similar to each other emotion.

All analysis was performed on a 64-bit Intel Core i5 CPU \(2 \times 2.67 \mathrm{GHz}\) with 4 GB RAM using the GENSIM framework for Python (Řehůřek \& Sojka, 2010) to create LSA spaces. For all tasks, we use a dimension of 36 and use Log-Entropy normalisation as our Association Function, found to generate optimal results (Nakov et al., 2001) and recommended for LSA (Landauer \& Dumais, 1997).

\section*{Valence and Arousal}

Here we take valance to mean the theoretical positive or negative attribution of an emotion keyword, and similarly arousal to mean the implied level of engagement. We should use the keywords theoretical valence and theoretical arousal as we are measuring emotion keywords relative to their generally accepted categorisation, although there does seem to be consistency in these categorisations between theorists. Table 2 shows the theoretical definitions of our keywords, accumulated using several circumplex models of affect (Watson \& Tellegen, 1985; Russell, 1980; Cacioppo \& Berntson, 1994).

Table 2: Valence and arousal categorisation of the twelve emotion keywords analysed.
\begin{tabular}{lll}
\hline Emotion & Valence & Arousal \\
\hline Angry & Negative & Engaged \\
Ashamed & Negative & Disengaged \\
Calm & Positive & Disengaged \\
Depressed & Negative & Disengaged \\
Excited & Positive & Engaged \\
Happy & Positive & Disengaged \\
Interested & Positive & Engaged \\
Sad & Negative & Disengaged \\
Scared & Negative & Engaged \\
Sleepy & Positive & Disengaged \\
Stressed & Negative & Engaged \\
Surprised & Positive & Engaged \\
\hline
\end{tabular}

We calculate the valence and arousal levels of each emotion for each subcorpus as follows. First, we run DELSAR on the subcorpus to generate clustering vectors for each emotion. Each emotion's valence is then calculated as the number of positive elements within its vector, as defined in Table 2 , divided by the total number of documents across all elements (which will always be 1000), or in other words, the percentage of positive elements within its vector. Similarly, each emotion's arousal is calculated as the percentage of engaged elements within its vector, again as defined in Table 2. We then normalise each valence and arousal value by taking away the average valence and arousal value, respectively, for all subcorpora analysed - Asia, Europe and NA regions and the six control subcorpora. This ensures relativity of the resulting circumplex model between these analysed groups; these groups can now be compared to one another to establish similarities and differences between them.

\section*{Results}

Figure 1 shows a plot of our circumplex of selected subcorpora. We can see that some emotions are more tightly packed than others, and interestingly, that low-valence-high-arousal and high-valence-low-arousal emotions are much more universally similar when compared to the other two quadrants of the circumplex. In order to visualise each separate region more clearly we illustrate the aggregate theoretical positivity and engagement for each subcorpus, shown in Figure 2. This clearly illustrates that our three regions do indeed have different conceptualisations of the same emotion keyword; we see that the region Europe is a much more positive and engaged culture; in other words, Europeans find the same emotion keywords to be more positive and engaging when compared to other cultures and indeed our control samples. Also, we discover that Asians find the same emotion keywords to be somewhat more positive, and North Americans somewhat more negative, with negligible arousal differences.

In order to analyse how tightly packed our emotion clus-


Figure 1: Circumplex of three regions and six controls.
ters are in Figure 1, we conducted K-Means cluster analysis to determine the centroids for each emotion, calculating the distances of each emotion to its centroid. We plot our centroids, shown in Figure 3, resulting in a circumplex that could be thought of as a universal emotion circumplex, illustrating what people think emotions to be, relative to each other emotion. We can see that the emotions scared, depressed and sad have a very similar valence, yet varying arousal levels; so too do the emotions sleepy and sad. We can also see, albeit less definitively, that the emotions stressed and surprised have a similar arousal level, but opposite valence; so too do the emotions sad and sleepy.

In order to identify which emotions have the most and least similar conceptualisations across cultures, we calculate the distance of each emotion to its respective centroid for each region, and calculate the sum of these distances for each emotion across all subcorpora, shown in Table 3. We discover that
the emotions sad and stressed have the most similar conceptualisations across all cultures; in other words, people understand these two emotions to mean the same thing independent of culture. Similarly, we find that the emotions surprised and depressed have the most widely varying conceptualisations across cultures; in other words, different cultures have very different valence and arousal attributions towards these two emotions. Note that we do not include the emotion ashamed in the top two due to a strange anomaly in control group 6 which skews an otherwise relatively tight cluster.

\section*{Discussion}

We would expect that the control groups would be tightly clustered around the centre of the circumplex in Figure 2, and for the most part, they are. The exceptions are control groups one and four, possibly due to the fact the the corpus is skewed in favour of tweets originating from NA (see Table 1);


Figure 2: Aggregate theoretical positivity and engagement for each subcorpus.

Table 3: Sum of subcorpus distances to respective centroids.
\begin{tabular}{lr|lr}
\hline Emotion & Distance & Emotion & Distance \\
\hline Sad & \(\mathbf{1 7 . 9 4}\) & Scared & 23.27 \\
Stressed & \(\mathbf{1 9 . 6 6}\) & Happy & 26.10 \\
Calm & 20.86 & Excited & 27.00 \\
Interested & 22.72 & Depressed & \(\mathbf{2 9 . 5 6}\) \\
Angry & 23.18 & Surprised & \(\mathbf{3 2 . 8 9}\) \\
Sleepy & 23.19 & Ashamed & 40.70 \\
\hline
\end{tabular}
this is somewhat verified by their closeness to the NA subcorpus. Other than these anomalous subcorpora, the circumplex does illustrate how different cultures significantly conceptualise emotions differently, in keywords of valence and arousal. Interestingly, there are certain emotions in certain regions that stick out of our analysis. One example is the emotion depression; Asians find this emotion much more negative than all other cultures and control groups. This could be due to cultural differences such as coping strategies (Aldwin \& Greenberger, 1987). Another example concerns the emotions happy and calm; Europeans and Asians find these emotions much more positive than North Americans and all control groups. Another suggests that Asians find interest a very positive and aroused emotion, compared to North Americans who conceptualise the same emotions, relatively, as negative and disengaged.

\section*{Limitations}

We document several limitations of our approach. Firstly, our database may still contain duplicate tweets, as some users duplicate tweets by appending, for example, a number at the end, making them unique from one another. Second, our


Figure 3: Centroid emotion circumplex.
modulus function does not take an even sample for our control groups at the country level, so they may be skewed in favour of countries with a higher frequency of documents within the database (our corpus on the whole is in fact skewed in favour of NA). Thirdly, we assume that the emotion keywords we have selected are in fact emotion qualia as opposed to adjectives. Fourth, our corpus is essentially a snapshot in time and may reflect, for example, the political or economic climate at the time, or skew due to global events such as the US election. Finally, our corpus consists entirely of English tweets, which skews our results in favour of Western cultures; our Asia, and to some extent, Europe subcorpora may not be entirely representative of their respective cultures as we disregard all native languages other than English. In addition, the subpopulations of those regions who choose to use Twitter, and do so in English, may be a biased sample.

\section*{Conclusions}

Emotions are being increasingly expressed online, and being able to understand these emotions is rapidly becoming a concern of AI and Cognitive Science. By mapping culturespecific emotion circumplexes, we hope to be better able to understand culture-specific perceptions or even experience of emotions. From the work presented here we can conclude the following:
Emotional semantics depends on culture. The same emotion keyword in one culture may describe different valence and arousal properties in another. This seems to be more true of some keywords than others, and could be critical where, for example, a significantly differing conceptualisation of the emotion depression would require a different understanding and response.
Emotions vary by geographic region. Europeans are more likely to express positiveness and engagement. Asians are
also more positive than North Americans, both relative to each other and to the control subcorpora. Note that this may reflect cultural differences in the public expression of emotion rather than its actual qualia - our method cannot disambiguate these.
Some emotions do seem to be conceptualised universally. The emotion keywords sad and stressed have the same conceptualisation across cultures, whereas cultures have the most disagreement regarding the conceptualisation of surprised.

We hope that our research paves the way for a better understanding of how language can be used to identify specific properties of emotions, and we encourage the reader to verify our results by downloading our code and corpus at http://www.aeir.co.uk/code.

\section*{References}

Aldwin, C., \& Greenberger, E. (1987). Cultural differences in the predictors of depression. Am J Community Psychol, 15(6), 789-813.
Bann, E. Y. (2012). Discovering basic emotion sets via semantic clustering on a Twitter corpus (Vol. CSBU-201301; Tech. Rep.). University of Bath, Department of Computer Science.
Bann, E. Y., \& Bryson, J. J. (2012, July). The conceptualisation of emotion qualia: Semantic clustering of emotional Tweets. In Proceedings of the 13th Neural Computation and Psychology Workshop (NCPW13).
Barrett, L. F. (2006, February). Solving the emotion paradox: Categorization and the experience of emotion. Personality and Social Psychology Review, 10(1), 20-46.
Bilovich, A., \& Bryson, J. J. (2008, November). Detecting the evolution of semantics and individual beliefs through statistical analysis of language use. In Proceedings of the AAAI Fall Symposium on Naturally-Inspired Artificial Intelligence (pp. 21-26). Arlington, VA: AAAI Press.
Bryson, J. J. (2008, June). Embodiment versus memetics. Mind \& Society, 7(1), 77-94.
Cacioppo, J. T., \& Berntson, G. G. (1994). Relationship between attitudes and evaluative space: A critical review, with emphasis on the separability of positive and negative substrates. Psychological Bulletin, 115(3), 401-423.
Fitzpatrick, J. K., \& Logan, J. (2011). Categorizing emotion in spoken language: An analysis of semantic and prosodic contributions to emotional communication. In Proceedings of the 33rd Annual Cognitive Science Conference. (pp. 762-767).
Fuhrman, O., McCormick, K., Chen, E., Jiang, H., Shu, D., Mao, S., et al. (2011). How linguistic and cultural forces shape conceptions of time: English and Mandarin time in 3D. Cognitive Science, 35(7), 1305-1328.
Gill, A. J., Gergle, D., French, R. M., \& Oberlander, J. (2008). Emotion rating from short blog texts. In Pro-
ceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1121-1124). ACM.
Gill, A. J., \& French, R. M. (2007). Level of representation and semantic distance: Rating author personality from texts. In Proceedings of the 2nd European Cognitive Science Conference.
Hamann, S. (2012). Mapping discrete and dimensional emotions onto the brain: controversies and consensus. Trends in Cognitive Sciences, 16(9), 458-466.
Hunt, E., \& Agnoli, F. (1991). The Whorfian hypothesis: A cognitive psychology perspective. Psychological Review, 98(3), 377.
Jack, R., Garrod, O., Yu, H., Caldara, R., \& Schyns, P. (2012, May). Facial expressions of emotion are not culturally universal. In Proceedings of the National Academy of Sciences of the United States of America (Vol. 109, pp. 7241-7244).
Landauer, T. K., \& Dumais, S. T. (1997). Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. Psychological Review(104), 211-240.
Lindquist, K. A. (2009, January). Language is powerful. Emotion Review (1), 16-18.
Lindquist, K. A., \& Barrett, L. F. (2008). Constructing emotion: The experience of fear as a conceptual act. Psychological Science, 19(9), 898-903.
Lowe, W. (1997, August). Meaning and the mental lexicon. In Proceedings of the 15th International Joint Conference on Artificial Intelligence (pp. 1092-1097). Nagoya: Morgan Kaufmann.
Nakov, P., Popova, A., \& Mateev, P. (2001). Weight functions impact on LSA performance. In Recent Advances in Natural Language Processing (RANLP 2001) (p. 187-193).
Pennebaker, J. W., \& King, L. A. (1999, December). Linguistic styles: Language use as an individual difference. Journal of Personality and Social Psychology, 77(6), 12961312.

Posner, J., Russell, J. A., \& Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. Development and Psychopathology, 17(3), 715-734.
Recchia, G., \& Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. Behavior Research Methods, 41(3), 647.
Řehůřek, R., \& Sojka, P. (2010, May 22). Software Framework for Topic Modelling with Large Corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks (pp. 45-50). Valletta, Malta: ELRA.
Russell, J. A. (1980). A circumplex model of affect. Journal of Personality and Social Psychology, 39(6), 1161-1178.
Watson, D., \& Tellegen, A. (1985, September). Toward a consensual structure of mood. Psychological Bulletin, 98(2), 219-235.

\title{
Modeling the Emergence of an Exemplar Verb in Construction Learning
}

\author{
Libby Barak, Afsaneh Fazly, and Suzanne Stevenson \\ Department of Computer Science \\ University of Toronto \\ \{libbyb,afsaneh,suzanne\} @cs.toronto.edu
}

\begin{abstract}
Using a computational model of verb argument structure learning, we study a key assumption of the usage-based theory: that the acquisition of a construction relies heavily on the existence of a high-frequency exemplar verb that accounts for a large proportion of usages of that construction in the input. Importantly, unlike the psycholinguistic experiments that focus on the learning of an artificial novel construction using novel verbs, here we examine the acquisition of the English sentential complement construction from naturalistic input. Our results provide new insights into exemplar-based learning in the context of naturalistic input with multiple semantic classes, and a diverse set of constructions for the verbs.
\end{abstract}

\section*{Introduction}

Verb argument structure acquisition is a challenging task that children face early in their life. In order to correctly use a verb, children must learn the syntactic structures that the verb appears in, as well as the semantic relations among the arguments of the verb. Nonetheless, children learn the correct usages of many verbs at a young age. Usage-based theories of language acquisition suggest that children learn the argument structure regularities mainly from the input they receive. These theories are supported by observing that children initially learn verb argument structures on an item-by-item basis, and only later generalize their verb-specific knowledge into abstract constructions that map a particular syntactic form to certain semantic properties (Tomasello, 2000).

The distributional properties of verb usages in childdirected input highly affect the developmental path of the acquisition of argument structure constructions. For example, several studies have shown that children tend to learn highfrequency verbs earlier (Naigles \& Hoff-Ginsberg, 1998; Matthews, Lieven, Theakston, \& Tomasello, 2005), and that they are more likely to detect grammatical anomalies in sentences containing such verbs (Theakston, 2004; Ambridge et al., 2008). Moreover, the relative frequency of a verb with a particular syntactic construction has been shown to positively correlate with the ability of young children to recall sentences containing that verb (Kidd, Lieven, \& Tomasello, 2006, 2010). Most importantly, there is evidence that the acquisition of a construction is connected to a high-frequency exemplar verb; e.g., give is the exemplar for the ditransitive construction (Goldberg, 1999; Kidd et al., 2006, 2010). In fact, several studies have shown that it is not just the amount of overall exposure to a construction that affects its acquisition, but instead learning seems to be facilitated by a high number of usages of a particular exemplar verb (Casenhiser \& Goldberg, 2005; Wonnacott et al., 2008).

The above psycholinguistic studies perform experiments on children, and hence are often limited in the number of
items they can investigate, and in how much they can tease apart the various interacting factors that might play a role on the results. For example, the sentence recall tasks performed by Kidd et al. (2010) examine only eight complement-taking verbs (CTVs). Moreover, due to their choice of verbs, they cannot separate the effects of overall frequency and relative construction frequency on their results. Using a computational model of argument structure learning, we extend these investigations into a larger set of CTVs, and also manipulate input in such a way that we can tease apart the effects of the various frequency factors. Our results are consistent with the findings of Kidd et al. \((2006,2010)\), that the relative frequency of a verb with a sentential complement is positively correlated with the ability of young children to recall sentences containing the verb in that construction. However, through computational modeling, we are further able to provide evidence on the interaction of verb frequency and relative construction frequency in accounting for their findings.

Studies examining the effect of a high-frequency exemplar verb in the acquisition of novel constructions often do so in the context of an artificial language learning task, where children are introduced to a novel verb mapped to a novel (or familiar) event semantics (Casenhiser \& Goldberg, 2005; Wonnacott et al., 2008). We use our computational model to investigate the existence and role of an exemplar verb in the acquisition of the English finite sentential complement (SC) syntax - a complex structure that has received less attention in the experimental studies (though see Kidd et al., 2006, 2010). Importantly, the use of a model enables us to vary distributional properties of the input in a way not easily achieved in a human experimental setting. Inspired by the work of Casenhiser and Goldberg (2005), we study the role of a highfrequency exemplar verb (think) for the acquisition of SC, but we do so in the context of a diverse set of verbs and constructions, as is the case in the naturalistic input that children receive. Our results suggest that the acquisition of a construction is facilitated by the relative frequency with which a class of semantically-related verbs appear with the syntactic form associated with the construction.

\section*{The Computational Model}

We use an extended version of the verb argument structure acquisition model of Alishahi and Stevenson (2008), which we have used in studying the acquisition of mental state verbs (Barak, Fazly, \& Stevenson, 2012). This model has appropriate characteristics for our study: (i) it focuses on argument structure learning, and the interplay between syntax and semantics; (ii) it is probabilistic and hence can naturally reflect

Input Utterance: He thinks Mom made pancakes.
Extracted Frame:
\begin{tabular}{|l|l|}
\hline main predicate & think \\
\hline other predicate & make \\
\hline event primitives & \{state, consider, cogitate, action \(\}\) \\
\hline event participants & \begin{tabular}{l} 
\{experiencer, perceiver, \{onsiderer \(\}\) \\
\{agent, animate \(\}\) \\
\{ theme, changed \(\}\)
\end{tabular} \\
\hline syntactic pattern & arg1 verb arg2 verb arg3 \\
\hline argument count & 3 \\
\hline complement type & finite \\
\hline
\end{tabular}

Table 1: An example of an input frame based on the utterance He thinks Mom made pancakes, and the semantic information assumed to be available from the scene (not shown).
the role of the statistical properties of the input in the formation of constructions; and (iii) it is incremental, which allows us to investigate changes in behaviour over time.

The input to our model is a sequence of frames, where each frame is a collection of features that resemble what children can extract from the utterances they hear and the typical learning scenes they preceive from their environments. We use features that include both semantic properties (i.e., event primitives and event participants), and syntactic properties (i.e., syntactic pattern, argument count, and complement type). Table 1 presents an example of an input frame given a child-directed utterance in a typical learning scene.

The model incrementally clusters the input frames into constructions that reflect probabilistic associations of semantic and syntactic features across similar verb usages. Note that a cluster is not simply a set of similar frames, but instead an abstraction over these frames represented as probability distributions over the possible values of each feature.

\section*{Algorithm for Learning Constructions}

The model clusters input frames into constructions on the basis of their overall similarity in the values of their features. Importantly, the model learns these constructions incrementally, considering the creation of a new construction for a given frame if the frame is not sufficiently similar to any of the existing constructions. Formally, the model finds the best construction (including a new one) for a given frame \(F\) as in:
\[
\begin{equation*}
\text { BestConstruction }(F)=\underset{k \in \text { Constructions }}{\operatorname{argmax}} P(k \mid F) \tag{1}
\end{equation*}
\]
where \(k\) ranges over all existing constructions and a new one. Using Bayes rule:
\[
\begin{equation*}
P(k \mid F)=\frac{P(k) P(F \mid k)}{P(F)} \propto P(k) P(F \mid k) \tag{2}
\end{equation*}
\]

The prior probability of a construction \(P(k)\) is estimated as the proportion of observed frames in \(k\), assigning a higher prior to constructions that are more entrenched (i.e., observed more frequently). The likelihood \(P(F \mid k)\) is estimated based on the values of features in \(F\) and the frames in \(k\) :
\[
\begin{equation*}
P(F \mid k)=\prod_{i \in \text { frameFeatures }} P_{i}(j \mid k) \tag{3}
\end{equation*}
\]
\begin{tabular}{|l|l|l|l|}
\hline \begin{tabular}{l} 
Semantic \\
class
\end{tabular} & Verb & \begin{tabular}{l} 
Overall \\
frequency
\end{tabular} & \begin{tabular}{l} 
Frequency \\
with finite-SC
\end{tabular} \\
\hline \hline \multirow{4}{*}{ Belief } & think & 13829 & \(100 \%\) \\
\cline { 2 - 4 } & bet & 391 & \(100 \%\) \\
\cline { 2 - 4 } & guess & 278 & \(76 \%\) \\
\cline { 2 - 4 } & know & 7189 & \(61 \%\) \\
\cline { 2 - 4 } & believe & 78 & \(21 \%\) \\
\hline \hline Desire & wish & 132 & \(94 \%\) \\
\cline { 2 - 4 } & hope & 290 & \(86 \%\) \\
\hline \hline Communication & tell & 2953 & \(64 \%\) \\
\cline { 2 - 4 } & say & 8622 & \(60 \%\) \\
\cline { 2 - 4 } & ask & 818 & \(29 \%\) \\
\hline \hline Perception & hear & 1370 & \(21 \%\) \\
\cline { 2 - 4 } & see & 9717 & \(14 \%\) \\
\cline { 2 - 4 } & look & 5856 & \(9 \%\) \\
\hline
\end{tabular}

Table 2: The Overall frequency of the 13 CTVs in our data, along with their relative frequency with finite-SC. Verbs are grouped by semantic class, and only the 13 verbs that appear with this construction are listed.
where \(i\) refers to the \(i^{t h}\) feature of \(F\) and \(j\) refers to its value. The conditional probability of a feature to have value \(j\) in construction \(k\), is calculated using a smoothed version of:
\[
\begin{equation*}
P_{i}(j \mid k)=\frac{\operatorname{count}_{i}(j, k)}{n_{k}} \tag{4}
\end{equation*}
\]
where \(\operatorname{count}_{i}(j, k)\) is the number of times feature \(i\) has the value \(j\) in construction \(k\), and \(n_{k}\) is the number of frames in \(k\).

\section*{Generation of the Input Corpora}

We generate artificial corpora for our simulations, since we do not have access to sufficient data of actual utterances paired with scene representations. To create naturalistic data that resembles what children are exposed to, we build an inputgeneration lexicon that is based on the distributional properties of actual child-directed speech (CDS). We extracted our verbs and their distributional properties from the CDS to 8 children from CHILDES (MacWhinney, 2000). \({ }^{1}\) We selected 31 verbs from different semantic classes and different frequency ranges, including 11 Physical Action (come, go, fall, eat, play, get, give, take, make, put, sit), 5 Perception (hear, listen, look, see, watch), 5 Communication (ask, say, speak, talk, tell), 5 Belief (think, know, guess, bet, believe), and 5 Desire (want, wish, like, mind, need). For each verb, we manually analyzed a random sample of 100 CDS usages (or all usages if fewer than 100) to extract distributional information about argument structures. Many of these verbs can take a (finite or infinitival) SC. Our focus in this work is on the finiteSC construction, and so we use the term Complement-Taking Verb (CTV) to refer to verbs that appear with the finite SC, following Kidd et al., 2010. Table 2 lists the 13 CTVs in our data, along with their semantic class, their overall frequency, and their relative frequency with the finite SC ).

\footnotetext{
\({ }^{1}\) Corpora of (Brown, 1973; Suppes, 1974; Kuczaj, 1977; Bloom, Hood, \& Lightbowny, 1974; Sachs, 1983; Lieven, Salomo, \& Tomasello, 2009).
}

We construct the input-generation lexicon by listing each of the 31 verbs (i.e. the main predicate), along with its overall frequency, as well as the frequency with which it appears with each argument structure. Each entry contains values for the syntactic and semantic features (see Table 1 for examples). By including these features, we assume that a learner is capable of (i) understanding basic syntactic properties of an utterance, such as syntactic categories (e.g., noun and verb) and word order, and (ii) perceiving and conceptualizing the general semantic properties of events - including mental, perceptual, communicative, and physical actions - as well as those of the event participants. Values for the semantic features (the event primitives and event participants) are taken from several resources, including Alishahi and Stevenson (2008), VerbNet (Kipper et al., 2008), and Dowty (1991). For each simulation in our experiments (explained below), we use the input-generation lexicon to automatically generate an input corpus of frames that reflects the observed frequency distribution in CDS. We perform 100 simulations, each on 20,000 frames, and examine the behaviour of our model over the course of learning.

\section*{Experiment 1: The Imitation Task}

Our goal here is to examine the role of verbs' overall frequency and their frequency with finite-SC, and the interaction of these frequencies, in the acquisition of argument structure constructions. Our simulations are inspired by the imitation task in which participants are asked to repeat a recently-heard utterance. Kidd et al. \((2006,2010)\) use this approach to examine the effect of verb frequency with finite-SC on how well children repeat utterances, in particular focusing on the relation between frequency of a verb with finite-SC and its likelihood of being correctly repeated, or substituted by another verb.

\section*{Experimental Setup}

Following Kidd et al. \((2006,2010)\), we focus on whether our model correctly repeats the verb of a sentence in an imitation task involving CTVs with sentential complements. We present the model with a full frame representing a complete utterance plus its corresponding scene, analogous to the presentation of a sentence with an accompanying picture, as in the psycholinguistic experiments. We then ask the model to predict the best verb in response to that frame, essentially asking it to repeat the just-presented verb.

To consider the responses of the model over a developmental trajectory, we train it on the full corpus, and at periodic fixed points during training, we present it with a test frame for each of the 13 CTVs in our lexicon, to see how it responds to each CTV. All the test frames have the same syntactic features (i.e., syntactic pattern, argument count, and complement type) corresponding to a finite SC that contains a transitive action verb, paired with the appropriate semantic features for the given CTV (see Table 1). For consistency, we use the same physical action verb for the embedded verb
(other predicate) in all 13 test frames, but randomly vary this verb across each of 100 simulations.

As in Kidd et al. \((2006,2010)\), we focus on the patterns of verb repetition and verb substitution among the model's responses. We record for each of the 13 test frames (at each point of testing) which verb the model predicts as its best response to that frame. To do this, we calculate the likelihood of each of our 31 verbs \(v\) given a test frame \(F_{\text {test }}\), as in:
\[
\begin{equation*}
P\left(v \mid F_{\text {test }}\right)=\sum_{k \in \text { Clusters }} P_{\text {main }}(v \mid k) P\left(k \mid F_{\text {test }}\right) \tag{5}
\end{equation*}
\]
where \(P_{\text {main }}(v \mid k)\) is the probability of the main predicate feature having the value \(v\) in cluster \(k\), calculated as in Eqn. (4), and \(P\left(k \mid F_{\text {test }}\right)\) is calculated as in Eqn. (2) (see Section for details). The model's response is taken to be the verb with the highest likelihood; this resembles the single choice of a verb made by the participants in the psycholinguistic experiments.

\section*{Results: Verb Repetition}

Kidd et al. \((2006,2010)\) observe a positive correlation between the frequency of a verb with finite-SC and the proportion of its correct repetitions. We focus on how frequency with finite-SC impacts our model's correct repetition of a verb. Figure 1 presents the proportion of times that each of the 13 CTVs are correctly repeated, which we refer to as the repetition accuracy. According to these results, a high frequency with finite-SC is neither a necessary nor a sufficient condition for a verb to be correctly repeated by our model. For example, although the two verbs with the highest repetition accuracy (i.e., think and say) have high frequencies with finite-SC, other verbs with high frequency with finiteSC (i.e., bet, guess, know, wish, and hope) are not easy for our model to repeat. In addition, see is among the top four verbs to be correctly repeated, although it has relatively low frequency with finite-SC (see Table 2). These results suggest that other factors beyond the frequency with finite-SC examined by Kidd et al. \((2006,2010)\) may play a role here.

Our model enables us to explore some of the possible factors, and to make predictions that could be verified through experiments with children. For example, the overall frequency of the verb affects the model's responses: Out of the four highest-frequency verbs (think, see, say, know), three also have a higher repetition accuracy compared with the other CTVs. However, like frequency with finite-SC, overall frequency alone does not predict the responses: The repetition rate is not in frequency order, and know is high frequency but has a low repetition rate. In fact, we note that, except for the verb think, the model rarely repeats Belief verbs correctly, regardless of their frequencies. These results point to another factor that might affect the performance of our model in repeating a verb: the frequency with which semanticallyrelated verbs appear with the same syntactic pattern as the verb to be repeated. To illustrate, when given a test frame that represents the semantic properties of a Belief verb with SC syntax, the model predicts the Belief verb with the highest frequency since it will have more occurrences in the clusters


Figure 1: Verb repetition accuracy for each of the 13 CTVs.
that the model bases its predictions on (see Eqn. 5). E.g., the verb know will have usages in clusters with many more occurrences of think, and hence the model will mainly produce think in response to test frames containing the semantics of know with the SC syntax. To further understand the interaction of overall frequency and frequency with finite-SC, and distribution over semantic classes, we next look at the patterns of verb substitution by our model.

\section*{Results: Verb Substitution}

Interestingly, Kidd et al. \((2006,2010)\) found that in a large number of cases, children specifically substituted the verb think in place of the verb they heard. They thus suggest that think is an 'exemplar' for the finite-SC construction. Figure 2 presents the proportion of times each of the 13 CTVs is produced by our model in place of the other 12 verbs, which we refer to as the substitution rate. That is, for each verb \(v\), its substitution rate reflects the proportion of times that our model incorrectly produces \(v\) in response to the test frames for the other 12 CTVs (out of 100 simulations). In line with the findings of Kidd et al. \((2006,2010)\), the model substitutes the verb think for the other verbs with a very high likelihood from an early stage (See Figure 2)

Kidd et al. \((2006,2010)\) attribute their finding to the high frequency of the verb think with finite-SC. However, we have observed that think also has the highest overall frequency among the 13 CTVs (see Table 2). In addition, think is a Belief verb, and it is known that people form a strong association between Belief verbs and the finite-SC syntax (Gleitman et al., 2005). It is thus not clear whether the status of think as an exemplar for the finite-SC construction is solely due to its high frequency with finite-SC, or if it is also affected by these other factors: (a) the high overall frequency of think, and/or (b) the overall strong connection of Belief verbs to the construction. We explore these factors in the next set of experiments.

\section*{Interaction of the Different Frequency}

One of the advantages of using a computational model is that we can manipulate the input to study the effects and interactions of the different frequency factors. Here, we manipulate


Figure 2: Verb substitution rate for each of the 13 CTVs.
the input such that we can examine the effects on the substitution patterns in our model of: overall frequency, frequency with finite-SC, as well as the frequency with finite-SC of a verb class as a whole. We perform three new experiments, in each of which we switch the overall frequency of think with one of the following three verbs: guess, believe, and tell. The goal is to change the input such that it is not the case that the verb with the highest overall frequency is also the verb with a frequency of \(100 \%\) with finite-SC (as is the case with think)-that is, we want to tease apart the effect of these two frequencies.

The first interesting finding is comparing the results of making guess vs. believe (other Belief verbs) the highestfrequency verb (in place of think). This explores the impact of a relatively high (but not \(100 \%\) ) frequency with finite-SC (for guess, of \(76 \%\) ) and a low frequency with finite-SC (for believe, of \(21 \%\) ), in the context of a very high overall frequency. We find that, as in the original results with think, guess is substituted for other verbs a very high proportion of the time ( \(75 \%\) ). However this does not hold for believe; when it is the highest-frequency verb, the Belief verb with next highest overall frequency and relatively high frequency with finite-SC (know) becomes the verb most often substituted for others, with a substitution rate of \(43 \%\). This behaviour predicts that both a high overall frequency and a relatively high frequency with finite-SC are required for a verb to be treated as an 'exemplar' of the finite-SC construction.

We also examined the result of making tell, which is not a Belief verb, the highest frequency verb with fintie-SC (again, in place of think). Interestingly, although tell is a verb with a relatively high frequency with finite-SC (like guess above), tell does not become the verb the model most frequently substitutes for other verbs (in contrast to guess). In this case, know - a Belief verb - is the verb most frequently substituted for others. This suggests that the semantics of the verb also plays an important role in determining the substitution behaviour. The strong association of particular (Belief-verb) semantics with the finite-SC syntactic pattern are necessary to the verb substitution behaviour.

In summary, our findings suggest a somewhat different view from that of Kidd et al. (2010), who suggested that think
was an exemplar verb in their experiments mainly because of its \(100 \%\) frequency with finite-SC. The results of the input manipulation with guess and believe predict that for a verb to be the exemplar for a construction (here the finite-SC), it has to have a sufficiently high overall frequency and also appear with the construction with a relatively high frequency. In addition, although a semantically diverse group of verbs appears with finite-SC, the input manipulation involving tell suggests that the exemplar verb will come from the Belief class, since Belief verbs as a whole have an overall high frequency of appearance with the SC syntax.

\section*{Experiment 2: Generalization}

Experiment 2 further examines the role of verb, construction, and semantic verb class frequencies in the acquisition of the finite SC. Given the noted strong association between the finite-SC syntax and Belief semantics, we focus here on the emergence of a 'Belief-SC' construction.

\section*{Experimental Setup}

Following Casenhiser and Goldberg (2005), we focus here on the effect of the distributional pattern of verb usages with a particular construction on the acquisition of that construction. Casenhiser and Goldberg (2005) introduce five novel verbs appearing in a novel construction (a novel syntactic pattern paired with a novel meaning), to 5-to 7 -year-old children in two input conditions: The skewed condition where one verb accounts for half of the occurrences of the construction, and the balanced condition with roughly equal number of usages of each verb. The study used a preferential-looking paradigm to show that participants in the skewed condition were significantly better at generalizing the newly-learned construction to a new novel verb (by looking at the scene with the appropriate semantics), compared to the balanced condition.

Our results in Experiment 1 imply that, in addition to the frequency with finite-SC of the individual verbs, their semantic class also influences the learning and use of verbs in a construction. This interaction of semantic classes is not addressed by the artificial language experiment of Casenhiser and Goldberg (2005), since it includes only a single class. Using a computational model enables us to explore the impact of a skewed vs. balanced distribution in a naturalistic setting, in which verbs from different semantic classes occur in the same syntactic frame under investigation (here, the finite SC), and verbs occur with multiple constructions (not just the one under investigation). Specifically, we examine how strongly our model learns the Belief-SC construction given the skewed input of our CDS-based data, compared to a balanced input, both with the same total exposure to CTVs. We form the balanced input by re-distributing the overall number of occurrences of CTVs so that each CTV would have an equal number of occurrences with finite-SC. Note that, all CTVs have an equal number of occurrences with finite-SC in the balanced input. However, because there is a different number of CTVs in each semantic class, the total number of finite-SC usages still slightly differs across classes.


Figure 3: The 3 highest likelihood values of semantic properties of the event (a) given the CDS-based distribution, (b) given artificially balanced frequencies with finite-SC.

We need to evaluate the ability of the model to generalize its knowledge of the Belief-SC construction in response to a novel verb when training on these two types of input. However, the model is incapable of engaging in preferential looking, as in Casenhiser and Goldberg (2005). Instead, we simulate preferential looking in our model as a choice between possible sets of event primitives, given a test frame. Following the psycholinguistic settings, we construct the test frame with a novel verb in place of the main predicate, where event participants are associated with a belief event, but the semantics of the predicate is missing. In other words, the test frame represents the belief construction used to test the children, and each set of event primitives represents one of the test scenes in a preferential looking task. At each point of testing, over 100 simulations, we record the set of event primitives that the model predicts as its best response to the partial test frame. This prediction corresponds to the selection of the scene with the appropriate action, given the arguments and syntax of the construction (as in Casenhiser and Goldberg).

\section*{Results}

Figure 3(a) and (b) report the proportion of times each of the three most likely sets of event primitives is chosen by our model as the most appropriate one, which we refer to as the event prediction rate. \({ }^{2}\) Figure 3(a) shows that the semantics of Belief events is highly associated with the arguments and syntax of novel Belief verbs from an early stage, given

\footnotetext{
\({ }^{2}\) Other sets of event primitives have lower likelihoods than the likelihoods presented here.
}
the skewed condition. That is, the Belief-SC construction is strongly entrenched given the naturalistically-skewed input.

However, in the balanced condition, as shown in Figure 3(b), only much later in training is the Belief event semantics predicted with the highest rate for the test frames. As in the results of Experiment 1, there is an effect of overall frequency in addition to frequency with finite-SC, of both verbs and classes. In the balanced input, each CTV has the same number of occurrences with finite-SC; hence there is only a small difference in the total number of occurrences of the different classes with this pattern. Recall that, to balance the input in terms of the CTV usages, we had to change the overall frequencies of the verbs and classes. In particular, we note that the overall frequency of the Belief class in the balanced input is much lower than that of the Perception class. The model is thus exposed to many more usages of Perception verbs with the finite-SC compared to the usages of Belief verbs with the same syntax, causing the observed delay in the formation of a strong Belief-SC construction.

\section*{Summary}

We have used a computational model to examine the effect of various distributional properties of the input on the acquisition of argument structure constructions. Specifically, we have examined the interaction of several factors in the emergence of an exemplar verb for the finite SC construction. Our results suggest that exemplar-based learning of a construction (such as the finite SC ) is sensitive to several properties of the input, including overall verb frequency, frequency of each verb with the construction, and the frequency of each semantic verb class with the construction. These results are in line with the psycholinguistic findings (e.g., Naigles \& HoffGinsberg, 1998; Casenhiser \& Goldberg, 2005; Wonnacott et al., 2008; Kidd et al., 2006, 2010). Moreover, they further our understanding of the exemplar-based learning mechanism by providing a broader investigation of the role of each of the above factors in the context of naturalistic input that contains multiple classes of verbs each appearing with multiple constructions. Our findings signify the importance of considering the interaction of the various distributional factors in the design of psycholinguistic experiments.

\section*{References}

Alishahi, A., \& Stevenson, S. (2008). A computational model of early argument structure acquisition. Cognitive Science, 32(5), 789-834.
Ambridge, B., Pine, J., Rowland, C., \& Young, C. (2008). The effect of verb semantic class and verb frequency (entrenchment) on children's and adults' graded judgements of argument-structure overgeneralization errors. Cognition, 106(1), 87-129.
Barak, L., Fazly, A., \& Stevenson, S. (2012). Modeling the acquisition of mental state verbs. In Proc. of the \(3 r d\) Wkshp. on Cognitive Modeling and Computational Linguistics.
Bloom, L., Hood, L., \& Lightbowny, P. (1974). Imitation
in language development: If, when, and why. Cognitive Psychology, 6(3), 380-420.
Brown, R. (1973). A first language: The early stages. Harvard U. Press.
Casenhiser, D., \& Goldberg, A. (2005). Fast mapping between a phrasal form and meaning. Developmental Science, 8(6), 500-508.
Dowty, D. (1991). Thematic Proto-Roles and Argument Selection. Language, 67(3), 547-619.
Gleitman, L., Cassidy, K., Nappa, R., Papafragou, A., \& Trueswell, J. (2005). Hard words. Language Learning and Development, 1(1), 23-64.
Goldberg, A. (1999). The emergence of the semantics of argument structure constructions. The emergence of language, 197-212.
Kidd, E., Lieven, E., \& Tomasello, M. (2006). Examining the role of lexical frequency in the acquisition and processing of sentential complements. Cognitive Development, 21(2), 93-107.
Kidd, E., Lieven, E. V., \& Tomasello, M. (2010). Lexical frequency and exemplar-based learning effects in language acquisition: Evidence from sentential complements. Language Sciences, 32(1), 132-142.
Kipper, K., Korhonen, A., Ryant, N., \& Palmer, M. (2008). A large-scale classification of English verbs. Language Resources and Evaluation, 42(1), 21-40.
Kuczaj, S. (1977). The acquisition of regular and irregular past tense forms. Journal of Verbal Learning and Verbal Behavior, 16(5), 589-600.
Lieven, E., Salomo, D., \& Tomasello, M. (2009). Two-year-old children's production of multiword utterances: A usage-based analysis. Cognitive Linguistics, 20(3).
MacWhinney, B. (2000). The CHILDES project: Tools for analyzing talk (Vol. 2). Psychology Press.
Matthews, D., Lieven, E., Theakston, A., \& Tomasello, M. (2005). The role of frequency in the acquisition of English word order. Cognitive Development, 20(1), 121-136.
Naigles, L., \& Hoff-Ginsberg, E. (1998). Why are some verbs learned before other verbs? Effects of input frequency and structure on children's early verb use. Journal of Child Language, 25(1), 95-120.
Sachs, J. (1983). Talking about the there and then: The emergence of displaced reference in parent-child discourse. Children's language, 4.
Suppes, P. (1974). The semantics of children's language. American psychologist, 29(2), 103.
Theakston, A. (2004). The role of entrenchment in children's and adults' performance on grammaticality judgment tasks. Cognitive Development, 19(1), 15-34.
Tomasello, M. (2000). Do young children have adult syntactic competence? Cognition, 74(3), 209-253.
Wonnacott, E., Newport, E., \& Tanenhaus, M. (2008). Acquiring and processing verb argument structure: Distributional learning in a miniature language. Cognitive Psychology, 56(3), 165-209.

\title{
Social Learning in Complex Networks: The Role of Building Blocks and Environmental Change
}

\author{
Daniel Barkoczi (barkoczi@mpib-berlin.mpg.de) \\ Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, Lentzeallee 94, 14195, Berlin, Germany \\ Mirta Galesic (galesic@mpib-berlin.mpg.de) \\ Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, Lentzeallee 94, 14195, Berlin, Germany
}

\begin{abstract}
We explore the interaction between information sampling and the structure of the social environment in the case of two prominent social learning strategies: imitate-the-best and imitate-the-majority. In a series of simulations a group of agents made repeated choices between options. We varied the building blocks of the strategies used by agents, the structure of the social network and characteristics of the task environment. A key factor influencing strategies' success is the speed with which they are able to respond to environmental change. In general, imitate-the-best provides a faster response compared to imitate-the-majority and larger samples help the former but hurt the latter. Less efficient networks decrease the performance of both, but are more detrimental for imitate-the-majority. Our findings highlight the role of sampling and social structure in the study of social learning, an area not sufficiently explored before.
\end{abstract}

Keywords: Social learning; information sampling; social networks; simple heuristics; simulation; decision-making

\section*{Introduction}

Humans and other animals obtain information via social learning. This is an efficient way to save the time and effort involved in individual trial-and-error learning and is known to underlie our capacity for culture. Despite the diverse list of empirical evidence for its use in the wild (Laland, 2004; McElreath, et al. 2008), theoretical models exploring the adaptive nature of social learning strategies lack sufficient detail to explain when we should expect to observe them. Most models study unstructured groups and focus only on the decision phase of implementing a strategy (e.g. imitate-the-majority), leaving open an important dimension affecting strategy performance: the interaction between information sampling and the structure of the social environment. The present study is an attempt towards filling this gap in the literature.

Social learning is often based on limited samples of the social environment. Most communities consist of sizable groups where an individual cannot survey all other group members within reasonable time before making a decision. Consider migrating animals deciding between multiple directions, individuals in an organization trying to jointly
solve a problem or stock traders trying to predict the best investment option (Couzin, Krause, Franks \& Levin, 2005; March, 1991). In such situations the way information about options is sampled from the social environment is likely to be an important aspect of any strategy. The structure of the social network in which social learning takes place can then in turn affect the options available for sampling. Previous work has shown that different network structures and their efficiency can affect the diversity of options in the population and the time it takes groups to converge on a solution (Lazer \& Friedman, 2007; Mason \& Watts, 2012). How does the performance of different strategies depend on the way they sample information and on the social environment in which they are embedded?

To address this question we study two representative social learning strategies: imitate-the-best and imitate-themajority (Boyd \& Richerson, 1985; Laland, 2004) and model them as decision heuristics that consist of different building blocks: search, stop and decision rules (Gigerenzer, Todd \& the ABC Research Group, 1999). By explicitly modeling these three phases we are able to test their relative contribution to strategy success in different social environmental structures.

Overall, a general characteristic shared by many social learning strategies, including those we study here, is that they alter the structure of the social environment by increasing the frequency of the correct option (i.e. the one with the highest payoff) and simultaneously decreasing the diversity of options in the group. This is a result of their bias towards specific sources (best member, majority) and their selectiveness (e.g. copy only if payoff better) \({ }^{1}\). This property has been extensively studied in the context of biased cultural transmission (Boyd \& Richerson, 1985) and suggests a key factor influencing strategy success in a changing environment: the speed with which they increase the frequency of the correct option in the group and, therefore, their ability to respond to environmental change. Our goal here is to show how this speed can be influenced by the strategy's building blocks (their sampling and decision rule) and by the structure and efficiency of the social network.

In what follows we derive specific expectations, based on previous literature and preliminary analytic calculations,

\footnotetext{
\({ }^{1}\) One can relax this assumption if other selective forces (e.g. natural selection) are at work.
}
about the effects of different building blocks and network structures on imitate-the-best and imitate-the-majority.

We consider a hypothetical situation where a group of agents make repeated choices between two options (one correct, the other incorrect). Whenever the environment changes, the previously correct option becomes incorrect and vice versa.

Effects of decision rules. In general, as long as the correct option is used by the majority of agents in a group and the environment is stable, both imitate-the-best and imitate-themajority will converge to the correct option. However, under the assumption that the best member can be reliably identified within the sample, the imitate-the-best will always converge faster because it requires only a single agent with the correct solution to reach a decision, whereas imitate-themajority requires at least two out of three. As soon as the environment changes, the correct option will be in minority. In this case, imitate-the-best will still be able to find it, however, as predicted by the Condorcet Jury Theorem (CJT), imitate-the-majority will never find the correct option because it requires that the proportion of agents with the correct option be higher than 0.5 (e.g. Grofman, Owen \& Feld, 1983).

Effects of information sampling and sample size. The CJT prediction may no longer hold when sampling is involved. Even if the correct option is in minority, imitate-the-majority may still be able to find it. Sampling as opposed to group-level aggregation can create situations where the correct option is more frequent in one's sample than overall in the group. When agents with such samples choose the correct option, this further increases the correct option's frequency in the group as a result of the environment altering feature of social learning discussed earlier (Boyd \& Richerson, 1985). Smaller samples are more likely to produce such situations, both because they are more likely to be biased and because they require fewer agents with the correct option in order to reach a decision. This suggests two situations where smaller as opposed to larger samples should benefit imitate-the-majority. First, whenever the group is converging towards the incorrect option, smaller samples will delay this process and keep the payoffs of the group higher for the longer time. Second, when the correct option is in minority, smaller samples will make it more likely to accidentally have a majority of agents with the correct option. In contrast, for imitate-the-best larger samples are always more advantageous, because they increase the chance of finding at least one agent with the correct solution.

Effects of network structure. Previous studies have demonstrated that higher network efficiency increases the speed with which information spreads and consequently decreases the diversity of information in the group. More efficient networks should, therefore, favor all strategies. Network efficiency depends on a variety of factors (Mason \& Watts, 2012); here we focus on clustering and average path length. As networks become more clustered and average path lengths increase, their efficiency decreases, and they maintain diversity for a longer time (Lazer \&

Friedman, 2007). We hypothesize that in such networks, the speed with which different strategies can find the correct option will become more important. As a result, the difference in speed between imitate-the-best and imitate-the-majority should become even larger. More clustered networks could have an additional effect by enabling the occurrence of relatively homogeneous clusters using the same option. If this option is incorrect, imitate-the-majority using a sample within that cluster will not be able to find the correct option. In contrast, imitate-the-best should be less affected by diversity of information as it only requires a single agent with the correct option.

\section*{Method}

\section*{Overview}

We simulated a situation where multiple agents \((N=100)\) had to make repeated choices between different number of options by acquiring information from their contacts. The choices they made directly affected their payoffs.
We created three social networks differing in their efficiency (as measured by clustering and average path length). Each agent had the same number of contacts in the network ( \(d=10\) ) and was assigned one of four decision strategies. Each strategy sampled randomly among one's contacts but differed in its stopping and decision rule. The agents' task was to make repeated choices between different number of options (2 or 10) at each time-step using their decision strategy. The environment could change on each time-step \(\left(t_{i}\right)\) with some probability \(\left(p_{c}\right)\) affecting the payoff of options at the next time-step \(\left(t_{i+1}\right)\). The simulation was run for \(t=1000\) time-steps and each condition was replicated 30 times \(^{2}\). To evaluate the performance of different strategies we tested them both in isolation and in an evolutionary competition where better performing strategies could replace worse performing ones. More specifically the simulation consisted of the following steps:
1) at \(t=0\) agents were placed in the networks and randomly assigned a decision strategy and an initial option
2) from \(t=1\) onwards, agents sampled the options and corresponding payoffs at \(t_{i-1}\) of their contacts
3) made a choice between sampled options based on their decision rules
4) only in the evolutionary competition: switched strategies with a small probability (introduced from \(t=50\) )
5) the environment changed with a certain probability, leading to a different option with the highest payoff
6) payoffs for the choice from step 3) were determined

Note that there is a lag between the information acquired from contacts and the realization of the agent's payoff in the sense that information is collected before environmental change occurs, thus allowing for the possibility of acquiring

\footnotetext{
\({ }^{2}\) Sensitivity analyses revealed that running the simulation for 2000 time-steps and 60 replications produced identical results.
}
outdated information when the environment changes to a new state.

\section*{Decision strategies}

We studied four decision strategies that differed in their building blocks (see Table 1). For each strategy we assumed that agents sample among their contacts randomly, and stop after collecting either a small \((n=4)\) or a large sample \((n=10)^{3}\). They then decide to try an option that is either endorsed by the majority of the sample contacts or by the agent that had the best payoff in the last time-step. In all cases agents only switch to a new option if that option's payoff was higher at the previous time-step than the option they are currently using. In situations where these two payoffs are equal or when the majority rule results in ties, agents chose randomly.

Table 1: Decision strategies
\begin{tabular}{cll}
\hline Sampling rule & Stopping rule & Decision rule \\
\hline random & \(n=4\) & imitate-the-majority \\
sample of & \(n=10\) & imitate-the-majority \\
contacts & \(n=4\) & imitate-the-best \\
& \(n=10\) & imitate-the-best \\
\hline
\end{tabular}

In order to keep track of a changing environment any social learning strategy requires that there is some form of individual learning generating novel options, therefore, we allowed new information to enter the population through copying error, a parameter we fixed at \(p_{e}=0.01\). That is, on each step there was a 0.01 chance that the agent does not consider the option used by its contacts, but a randomly selected option, however, agents only switched to this option if it had a higher payoff at the previous time-step. This lies in contrast with other studies which allowed new information to enter the group by assuming that whenever other agents' payoffs are lower or equal, the agent does not stick with its own option but explores other options randomly (Lazer \& Friedman, 2007; Mason \& Watts, 2012). These studies, therefore, allowed for a higher amount of innovation than our model. In this way we explore the performance of social learning strategies when aided with only a minimum amount of individual learning.

\section*{Decision environment}

Two factors affecting the decision environment were varied in different simulations: a) the number of options available and \(b\) ) the rate of environmental change. To manipulate the first factor we assumed that agents choose either between 2 or 10 options with payoffs ranging from 1 to 2 and from 1 to

10 respectively. At any given time, only one option had the highest payoff. On the first time-step agents were assigned options randomly. In conditions with 2 options, we varied the initial proportion of the correct option in the group ( \(p_{\text {init }}=0.2,0.5\) or 0.7 ). For 10 options each option had the same initial proportion. For the second factor we assumed that the payoffs of options can change on each time-step with probability \(p_{c}=0.001,0.01,0.1\) or 0.4 reflecting a discreet scale between slow and fast rates of change. We ran all possible combinations of environmental change on all 3 network structures described below.

\section*{Network structure}

Three different networks were created, ranging from most efficient to least efficient as measured by two standard indicators in the network science literature (Mason \& Watts, 2012): clustering coefficient and average path length. The clustering coefficient measures the extent to which the network is dominated by isolated cliques, which from a communication perspective decreases the efficiency of a network by making it harder for information to spread the higher the clustering. Consider an example where small groups of tightly connected agents exchange information but because groups are isolated from other groups information spreads much slower between these small units.
Another measure of efficiency is average path length, the average number of steps it takes to get from any agent to any other agent in the network. The shorter the path length the more easily information can spread. The efficiency of a network is known to affect how quickly information spreads from one part to another, however, it can also enable maladaptive information to spread more rapidly as in the case of panics following flu pandemics or stock bubbles. Many real-world networks are known to have both high clustering and low average path lengths thus representing an intermediate level of efficiency. These small-world networks (Watts \& Strogatz, 1998) can be mimicked by performing random re-wirings on edges of a lattice. In line with previous studies (e.g. Schwenk \& Reimer, 2007), we started by first generating a random directed lattice and then rewired it with a 0.1 probability to obtain a small-world network \({ }^{4}\). In addition we created a fully-connected network absent of any structural properties to be able to compare to previous studies that focused on unstructured groups (see Table 2). All three networks had a fixed degree of 10 and a total of 100 nodes \((d=10, n=100)\).

\footnotetext{
\({ }^{3}\) Sensitivity analyses with sample sizes \(n=3\) and \(n=9\) produced similar results and we do not report them here.
\({ }^{4}\) Other networks with lower values of rewiring produce similar results, therefore, we omit them.
}

Table 2: Types of networks used in the simulation ( \(\mathrm{n}=100, \mathrm{~d}=10\) )
\begin{tabular}{llll}
\hline Network & \begin{tabular}{l} 
Clustering \\
coefficient
\end{tabular} & \begin{tabular}{l} 
Average \\
path length
\end{tabular} & \begin{tabular}{l} 
Rewiring \\
probability
\end{tabular} \\
\hline Lattice & 0.67 & 5.55 & \(\mathrm{p}=0\) \\
\begin{tabular}{l} 
Small world
\end{tabular} & 0.31 & 2.35 & \(\mathrm{p}=0.1\) \\
\begin{tabular}{l} 
Fully \\
connected
\end{tabular} & 1 & 1 & \(\mathrm{p}=1\) \\
\hline
\end{tabular}

\section*{Evolutionary competition}

In order to properly evaluate each strategy we look at their performance both in isolation (in homogeneous groups using the same strategy) and by directly testing different strategies against each other (heterogeneous groups) in an evolutionary competition. In the former we are interested in isolating the factors contributing to the success of different strategies, whereas in the latter we wish to evaluate them in a competitive setting where the performance of a strategy can depend on the strategies used by other agents in the group. Evolutionary competitions are a popular method in the study of social learning (e.g. Rendell, et al. 2010) where the strategy accumulating the highest payoff has the best chance of reproducing and spreading in the population, while the worst performing strategies die out. The prevalence of a strategy is, therefore, a clear-cut measure of its success in a given environment.

There are many ways to implement an evolutionary dynamic. Here we chose the 'imitation process' (Nowak, 2006) in order to reflect a plausible real-life scenario. We assumed that on each time-step, randomly selected agents change their strategies to one of their contacts' strategy with a probability proportional to the cumulative payoff of that contact. If none of the contacts has a higher payoff, the agent keeps its strategy, and in situations of equal payoff random choice is implemented. We fixed the parameter specifying the probability of strategy change to \(p_{s}=0.02\) thus expecting 2 agents switching strategies on each time-step. Evolutionary dynamics were introduced from the \(t=50\) timestep to allow for a burn-in period.

\section*{Simulation results}

Figure 1A shows the overall performance of the four different decision strategies observed in isolation, measured by their rate of environmental tracking (percentage of agents using the correct option on each time-step). We show the results for 2 options, probability of environmental change \(p_{c}=0.001\), and initial probability of correct option \(p_{\text {init }}=0.5\), averaged across networks \({ }^{5}\). To make the main results easier to view, we focus on the time-steps before and after environmental change occurring at \(t=100\). Figure 1B shows the frequency of different strategies in the evolutionary
competition averaged across networks for the same environmental condition. Overall, imitate-the-best consistently outperforms imitate-the-majority both in homogeneous and heterogeneous groups. This result holds in all network structures and in environmental conditions.


Figure 1. Panel A. Performance of strategies observed in isolation. Panel B. Frequency of strategies in the evolutionary competition. Results are shown for environmental conditions \(p_{c}=0.001\) and \(p_{\text {init }}=0.5\), averaged across networks.

\section*{Effects of information sampling}

From Figure 1A we can see the number of time-steps it takes groups using each of the strategies to converge on the correct solution after the environment has changed. As expected, imitate-the-best benefits somewhat from larger samples, however, even its small sample version outperforms both versions of imitate-the-majority. The opposite is the case for imitate-the-majority, which is hurt by larger samples and actually performs better when it samples fewer people. This result highlights that speed with which different strategies can recover after environmental

\footnotetext{
\({ }^{5}\) Results for 10 options and other rates of environmental change and initial probability of correct option do not change the main conclusions and we do not present them here.
}
change is crucial to their success and demonstrates that different sampling regimes should be adopted depending on the decision rule used.

As mentioned before, without sampling, imitate-themajority will converge on an incorrect option whenever the proportion of agents using the correct option is smaller than 0.5 . As expected, these results do not hold when decisions are based on sampled information as opposed to overall group aggregation. As visible in Figure 1A, imitate-themajority is able to find the correct option even when the proportion of agents using it falls under 0.5 . As a sensitivity check, we reran our simulations with \(p_{\text {init }}=0.2\) and copying error of \(\mathrm{p}_{\mathrm{e}}=0\), allowing no new information to enter the population. Even then, imitate-the-majority can still converge on the correct option, in particular when it uses small samples.


Figure 2. Performance of different strategies in the three network structures. Results are shown for environmental conditions \(p_{c}=0.001\) and \(p_{\text {init }}=0.5\)

\section*{Effects of network structure}

Overall, we find that regardless of strategy, more efficient networks are faster at spreading information and that this helps groups in all conditions. However, we observe an effect for network structure on the relative difference between strategies. Figure 2 shows that the difference between strategies is least pronounced in the fully connected network absent of any structural properties, however, as networks become more structured (thereby decreasing the efficiency and speed with which information flows), the difference between imitate-the-best and imitate-the-majority becomes more pronounced.

The effect of network structure is especially visible immediately after environmental change. In networks with high clustering and long path lengths such as lattice, relatively isolated agents may form homogeneous groups possessing the same information. In these situations, imitate-the-majority has problems finding the correct option. The larger the sample, the more prone is this strategy to get stuck. As expected, the performance of imitate-the-best is less affected by network structure.

\section*{Discussion}

Our goal was to study how information sampling and the structure of the social environment affect the performance of two representative social learning strategies: imitate-thebest and imitate-the-majority. We modeled social learning strategies as heuristics consisting of different building blocks and embedded them in three social networks in a task involving repeated choices between multiple options.

Overall, we find that imitate-the-best consistently outperforms imitate-the-majority and our results suggest that the reason underlying this finding is the speed with which different strategies are able to respond to environmental change. This speed is affected both by different building blocks and the structure of the social environment. Imitate-the-best is always faster at finding the good option because its decision rule requires fewer correct instances in the sample and larger samples are always beneficial. In contrast, sample size has a counterintuitive effect on imitate-the-majority with smaller samples increasing the likelihood and thereby the speed of finding the correct option. The relative difference between imitate-the-best and imitate-the-majority, however, is moderated by network structure. More efficient networks (those with lower clustering and shorter path lengths) benefit all strategies and decrease the difference between them while less efficient networks (with more clusters and longer path lengths) increase the difference by having a worse impact on imitate-the-majority.

Information sampling as opposed to group-level aggregation has an additional effect on imitate-the-majority: it can still converge on the correct option, even if less than \(50 \%\) of the group is using it. This result lies in contrast to the predictions of the Condorcet and related Theorems on full group-level aggregation of information in a single trial (Grofman, Owen \& Feld, 1983).

Both imitate-the-best and imitate-the-majority have been extensively studied both theoretically and empirically (e.g. Conradt \& Roper, 2003; Garcia-Retamero, Takezawa \& Gigerenzer, 2006; Hastie \& Kameda, 2005; Katsikopoulos \& King, 2010; McElreath, Wallin \& Fasolo, 2012). Much of this work has studied small and unstructured groups and focused exclusively on the decision-phase of implementing these strategies (but see Pachur, Rieskamp \& Hertwig, 2005; Schwenk \& Reimer, 2007 for exceptions in other contexts). We believe that this leaves many important details affecting strategy success unaddressed and can be one reason why some studies reach different conclusions. The present study is a first step towards developing a more general framework for capturing the interactions between the building blocks of social heuristics and the structure of the social and task environments that they exploit. We propose that their study can bring novel insight into our understanding of social phenomena including the evolution of different social learning rules, the diffusion of innovations in cultures or the strategy selection process in social domains.

\section*{Acknowledgments}

The authors wish to thank Konstantinos Katsikopoulos and Henrik Olsson for helpful comments on an earlier draft and the Max Planck Institute for Human Development for financial support.

\section*{References}

Boyd, R., \& Richerson, P. J. (1985). Culture and the Evolutionary Process. Chicago: University of Chicago Press.
Conradt, \& Roper, T. J. (2003). Group decision-making in animals. Nature, 421, 155-158.
Couzin, I. D., Krause, J., Franks, N. R. \& Levin, S. A. (2005). Effective leadership and decision-making in animal groups on the move. Nature, 433, 513-516.
Garcia-Retamero, R., Takezawa, M. \& Gigerenzer, G. (2006). How to learn good cue orders: When social learning benefits simple heuristics. In R. Sun \& N. Mikaye (Eds.), Proceedings of the 28th Annual Conference of the Cognitive Science Society (pp. 1352-1357). Mahwah, NJ: Erlbaum
Gigerenzer, G., Todd, P. M., \& the ABC Research Group. (1999). Simple Heuristics that Make Us Smart. New York: Oxford University Press.
Grofman, B., Owen, G. \& Feld, S. C. (1983). Thirteen theorems in search of the truth. Theory and Decision, 15, 261-278.
Hastie, R. \& Kameda, T. (2005). The Robust Beauty of Majority Rules in Group Decisions. Psychological Review, 112, 494-508.
Katsikopoulos, K. \& King, A. J. (2010). Swarm Intelligence in Animal Groups: When Can a Collective Out-Perform an Expert? Plos ONE, 5, 1-5.
Laland, K. N. (2004). Social learning strategies. Learning \& Behavior, 32, 4-14.

Lazer, D., \& Friedman, A., (2007). The Network Structure of Exploration and Exploitation. Administrative Science Quarterly, 52, 667-694.
Mason, W. A., \& Watts, D. J. (2012). Collaborative learning in networks. Proceedings of the National Academy of Sciences, 109, 764-769.
March, J. G. (1991). Exploration and Exploitation in Organizational Learning. Organizational Science, 2, 71-78.
McElreath, R., Bell, A. V., Efferson, C., Lubell, M., Richerson, P. J. \& Waring, T. (2008). Beyond existence and aiming outside the laboratory: estimating frequencydependent and pay-off-biased social learning strategies. Philosophical Transactions of the Royal Society B, 363, 3515-3528.
McElreath, R., Wallin, A. \& Fasolo, B. (2012). The Evolutionary Rationality of Social Learning. In R. Hertwig \& U. Hoffrage \& the ABC Research Group (Eds.), Simple Heuristics in a Social World. New York: OUP.
Nowak, M. A. (2006). Evolutionary Dynamics: Exploring the Equations of Life. Harvard University Press.
Pachur, T., Rieskamp, J., \& Hertwig, R. (2005). The social circle heuristic: Fast and frugal decisions based on small samples. In K. Forbus, D. Gentner, \& T. Reiger. (Eds.), Proceedings of the 26th Annual Conference of the Cognitive Science Society (pp. 1077-1082). Mahwah, NJ: Erlbaum.
Rendell, L., Boyd, R., Cownden, D., Enquist, M., Eriksson, K., Feldman, M. W., Fogarty, L., Ghiranda, S., Lillicrap, T., \& Laland, K. N. (2010). Why Copy Others? Insights fomr the Social Learning Strategies Tournament. Science, 328, 208-213.
Schwenk, G., \& Reimer, T. (2007). Social Influence and Bounded Rationality: Heuristic Decision Making in Complex Networks. In D. S. McNamara \& G. Trafton. (Eds.), Proceedings of the Twenty-Ninth Annual Conference of the Cognitive Science Society (pp. 1497-1484). Mahwah, NJ: Lawrence Erlbaum Associates.
Watts, D. J., \& Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. Nature, 393, 409-410.

\title{
Spoken Words Activate Cross-Linguistic Orthographic Competitors in the Absence of Phonological Overlap
}

\author{
James Bartolotti (j-bartolotti@u.northwestern.edu)
}

\author{
Natalia L. Daniel (nataliadaniel2012@u.northwestern.edu)
}

Viorica Marian (v-marian@northwestern.edu)
Northwestern University
Department of Communication Sciences and Disorders
2240 Campus Drive, Evanston, IL 60208 USA

\begin{abstract}
Related languages, like English and Spanish, often have similar orthographies but use the same letters to represent different sounds. Learning a second language frequently involves learning additional letter-sound mappings that mismatch those in the native language. In the current study, we investigated whether L2 spoken words activate L2 orthography despite conflict with L1 orthography-to-phonology mappings. Participants first learned an artificial language with letter-sound mappings that mismatched English (e.g., the letter 'G' represented the sound \(/ \mathrm{h} /\), and the word /gufo/ was spelled 'hane'). Next, fixations of L1 crosslinguistic orthographic competitors (e.g., 'cane') in response to auditory L2 input (e.g., /gufo/) were assessed using the visual world paradigm. Results showed that participants fixated L1 competitors that overlapped with L2 targets orthographically (but not phonologically) more than unrelated fillers. We conclude that second language learners can rapidly acquire novel letter-sound mappings, and words based on these mappings are integrated into the existing lexicon where they can activate orthographic competitors in the native language.
\end{abstract}

Keywords: Language processing; Language learning; Crosslinguistic competition

\section*{Introduction}

Spoken language processing involves decoding an incoming auditory signal to access words in the mental lexicon. It's not obvious that this process should be affected by orthographic knowledge, because written language is a relatively recent invention, and is learned years after spoken language. Yet, there is evidence that orthography, once acquired, influences performance on phonological tasks (Jakimik, Cole, \& Rudnicky, 1985; Johnston, McKague, \& Pratt, 2004; Salverda \& Tanenhaus, 2010), suggesting tight interconnectivity between orthography and phonology. This interconnectivity may be a source of difficulty during second language acquisition, because the same letters can represent different sounds across languages. For example, the letter 'W' maps onto the phoneme /w/ in English, but /v/ in German (one is a labiovelar approximant, while the other is a voiced labio-dental fricative, which differs on both voicing, place, and manner of articulation). Second language learners thus need to learn and use these novel letter-sound correspondences in the appropriate language context, despite years of experience with a different set of mappings in their native language.

Orthographic knowledge can help or hinder phonological processing, depending on the context. Literate adults perform better than illiterate adults on metaphonological tasks
such as adding or deleting sounds at the beginning of nonwords (Morais, Cary, Alegria, \& Bertelson, 1979), because literate adults can use their orthographic representations as a mental aid. On the other hand, orthography can also distort phonological perception. French speakers are more likely to misperceive the phoneme \(/ \mathrm{p} /\) as a \(/ \mathrm{b} /\) in spoken words when the sound is represented by the letter ' \(B\) ', as in the French word 'absurd,' pronounced /apsyrd/ (Hallé, Chéreau, \& Segui, 2000). Furthermore, orthography can affect online processing of speech. Orthographically-related primes improve auditory lexical decision times (Jakimik et al., 1985), but written words can act as competitors during auditory visual world search tasks (Salverda \& Tanenhaus, 2010). Based on timecourse analyses from research on event-related potentials (ERPs), these orthographic effects occur early in the speech signal and are time-locked to the source of orthographic effects in the word, suggesting that orthography is activated online during speech processing, and not strictly as a postlexical decision process (Perre \& Ziegler, 2008).

The link between orthography and phonology extends to novel words as well. In a recent study (Johnston et al., 2004), monolingual English speakers were taught a series of novel words but only learned the words' phonological forms, and were never presented with orthography. During a subsequent masked priming task, orthographic versions of the trained words showed a significant priming effect, compared to the absence of any effect for completely novel written nonwords. This finding suggests that learners automatically generate orthographic forms for novel auditory words, based on the phonotactics of their native language. When these generated forms are accurate, they can accelerate vocabulary learning and improve reading of previously learned auditory words (McKague, Pratt, \& Johnston, 2001). During second language learning, though, they are more likely to be inaccurate and impair learning. For example, an English-speaking learner of German may hear the auditory German word /vek/ (spelled 'weg') and create an incorrect orthographic representation 'veck' based on their knowledge of English. This incorrect representation may then impair learning to read and write in German, as the learner's internal representations must be inhibited and relearned. Previous work indicates that English speakers are able to learn words and letter-sound mappings in artificial languages with training, even when they include
non-English phonemes (Kaushanskaya \& Marian, 2009a), graphemes (Bitan \& Karni, 2003), or a combination of the two (Kaushanskaya \& Marian, 2008). In the current study, we isolated acquisition of novel letter-sound mappings by recombining familiar English letters and sounds. Even when there are no new letters or sounds to be learned, acquiring novel mappings can be difficult - for example, two of the most challenging letters for English learners of Russian to learn are B (pronounced \(/ \mathrm{v} /\) ), and Y (pronounced \(/ \mathrm{u} /\) ), which are often mispronounced as /b/ and /i/ respectively, representing interference from existing English mappings (Comer \& Murphy-Lee, 2004). Full language acquisition requires learners to form new mappings between orthography and phonology that are appropriate for the target language, and to be able to inhibit their native language mappings during L2 processing.

In sum, there is a tight interconnectivity between orthography and phonology, but letter-sound mappings often conflict across languages, which may lead to second language learning difficulties. The current study was designed to investigate how learners manage these difficulties. The first goal of the study was to assess how well learners are able to acquire vocabulary in a novel language with letter-sound correspondences that mismatch English. The second goal was to determine whether auditory words in the L2 will activate L2 orthography, based on spreading activation from the L2 to words that resemble the L2 orthographic form. These questions are addressed by teaching participants a L2 vocabulary with letter-sound correspondences completely distinct from English. If participants are able to learn the novel orthophono mappings, then presentation of the auditory form of the word will lead to activation of the corresponding orthographic form. In a connectionist model of language processing, activation should then spread to similarly spelled words in the lexicon.

Because the novel language and English have different letter-sound mappings, auditory targets in the new language do not overlap phonologically with their English orthographic competitors (e.g., the novel word /gufo/, spelled 'hane', overlaps orthographically but not phonologically with the English word 'cane'). If participants look at crosslinguistic orthographic competitors upon hearing L2 words, they must have activated the target's L2 orthographic form, which then spread activation to orthographically related items, including the crosslinguistic competitor. The current study thus allows us to simultaneously assess the effects of novel orthophono mappings and cross-linguistic interference on speech processing in a newly learned language.

\section*{Methods}

\section*{Participants}

Twenty monolingual English speakers ( 16 females, 4 males) participated. Eyetracking data was unavailable for one participant due to equipment error. All participants reported current English use at \(99 \%\) of the time or more, and a proficiency
of three or less on a scale of 1 (no knowledge) to 10 (perfect) in a second language (LEAP-Q, Marian, Blumenfeld, \& Kaushanskaya, 2007).

\section*{Materials}

A miniature artificial language named Colbertian \({ }^{1}\) was created using a novel alphabetic system. Thirteen English graphemes (four vowels and nine consonants) were paired with thirteen English phonemes so that the English and Colbertian sounds for each letter differed maximally in voice, place, and manner for consonants, or height, backness, and rounding for vowels (Table 1). Reusing English phonemes ensured that participants needed only to learn the novel lettersound correspondences, but not any new phonetic categories.

Table 1: Colbertian Alphabet
\begin{tabular}{|c|c|c|}
\hline Grapheme & English Phoneme & Colbertian Phoneme \\
\hline a & /eı/ /æ/ & /u/ \\
\hline e & /i/ /e/ & \(1 / 1\) \\
\hline i & /ai/ /I/ & /æ/ \\
\hline o & /ov/ /o/ & /i/ \\
\hline b & /b/ & /s/ \\
\hline d & /d/ & /t \(\mathrm{f} /\) \\
\hline h & /h/ & /g/ \\
\hline k & /k/ & /w/ \\
\hline n & /n/ & /f/ \\
\hline p & /p/ & /z/ \\
\hline r & /.I/ & /h/ \\
\hline t & /t/ & /d3/ \\
\hline v & /v/ & /t/ \\
\hline
\end{tabular}

Twenty-four words were then created using the Colbertian alphabet (Table 2). Each word was recorded by a female speaker of Standard American English, and was associated with an easily-nameable black and white line drawing (naming consistency higher than \(80 \%\) from the International Picture Naming Project database, Bates et al., 2003, or norming with Amazon's Mechanical Turk). Each of the novel words was designed to overlap orthographically, but not phonologically, with an English competitor word in order to isolate the effect of English orthographic knowledge on Colbertian auditory word processing. Target words, competitor words, and filler words were matched on the following variables: phonological neighborhood size (IPhOD; Vaden, Halpin, \& Hickok, 2009), orthographic neighborhood size (N-Watch; Davis, 2005), English lexical frequency (SUBTLEX-US; Brysbaert \& New, 2009), concreteness, imageability, or familiarity (MRC Psycholinguistic Database; Coltheart, 1981), all \(p\) 's \(>0.05\).

\footnotetext{
\({ }^{1}\) The language was named after comedy show wordsmith and Northwestern University alumnus Stephen Colbert to engage participants in the learning task.
}

Table 2：Colbertian Vocabulary
\begin{tabular}{|c|c|c|c|}
\hline Colbertian & Colbertian & English & English \\
\hline Orthography & Phonology & Translation & Competitor \\
\hline vite & ／tæd3コ／ & wig & kite \\
\hline tave & ／dzuta／ & pan & cave \\
\hline eron & ／Jhif／ & tent & iron \\
\hline dipe & ／t「æzコ／ & snake & pipe \\
\hline vope & ／tiza／ & mouse & rope \\
\hline vate & ／tudza／ & ear & gate \\
\hline kire & ／wæhว／ & gun & fire \\
\hline dibe & ／tfæsっ／ & hose & dice \\
\hline rako & ／huwi／ & grapes & rake \\
\hline dova & ／tfitu／ & ax & dove \\
\hline rike & ／hæwo／ & shark & rice \\
\hline nove & ／fito／ & sun & nose \\
\hline rone & ／hifo／ & swan & cone \\
\hline hane & ／gufo／ & ruler & cane \\
\hline bave & ／suts／ & purse & wave \\
\hline nake & ／fuwo／ & bird & cake \\
\hline bine & ／sæfo／ & pants & wine \\
\hline robi & ／hisæ／ & bench & robe \\
\hline tavo & ／d3uti／ & raft & taco \\
\hline vabe & ／tuss／ & owl & vase \\
\hline bika & ／sæwu／ & plate & bike \\
\hline bona & ／sifu／ & cow & bone \\
\hline roke & ／hiws／ & lock & rose \\
\hline tapi & ／d3uzæ／ & cat & tape \\
\hline
\end{tabular}

\section*{Procedure}

Participants learned Colbertian in a single experimental ses－ sion in four steps．In the first step，participants were exposed to each of the Colbertian words＇spellings and pronunciations． A single written word appeared in the center of a computer screen，and the auditory form of the word was pronounced over headphones．The participant repeated the word aloud and clicked the mouse to advance to the next word．Each word was presented once，for a total of 24 exposures．In the second step，participants practiced associating the words and their pronunciations until they reached a \(90 \%\) learning cri－ terion．In a single trial，four Colbertian words were shown on the screen，and the auditory form of the target word was played over headphones．After selecting one of the four words，participants received feedback：the target word turned green，the three foils disappeared，and the word was replayed over headphones．This ensured that participants had an op－ portunity to relearn the words they answered incorrectly．Af－ ter 24 trials，with each word as a target once，the participant was shown their accuracy for the block．Each participant re－ peated blocks of 24 trials until they reached \(90 \%\) accuracy on two consecutive blocks．

In the third step，participants were familarized with the meanings of the words they had just learned．Four pictures appeared on the screen，and after 1500 ms ，a Colbertian word
appeared on the center of the screen and was played over headphones，and the picture it represented was outlined with a red box（nontarget pictures remained visible）\({ }^{2}\) ．In the fourth step，participants practiced associating Colbertian words with their pictures until achieving the \(90 \%\) learning criterion．In each trial，four pictures were displayed on the screen，and the target word was simultaneously presented in written and auditory forms．After selecting a picture，feedback was pro－ vided：the target picture was outlined in a red box（nontarget pictures remained visible）and the target word was replayed over headphones．Each trial，including response time and feedback，lasted exactly six seconds to equate picture view－ ing times across trials．After 24 trials，with each word as a target once，the participant was shown their accuracy for the block．Participants continued doing training blocks until they reached \(90 \%\) accuracy on two consecutive blocks，at which point they were finished learning Colbertian．

After learning Colbertian，participants immediately began a visual world eyetracking task to assess the effect of English orthographic knowledge on Colbertian spoken word process－ ing．Each trial began with a 1000 ms fixation cross to ori－ ent participants＇gaze．Next，the cross disappeared and four pictures appeared in the corners of the screen．After a 500 ms delay，a Colbertian word indicating the target picture was played over headphones（the orthographic form of the target was never shown）．The participant＇s task was to click on the target picture as quickly and accurately as possible．No feed－ back was provided．Trial presentation was controlled by the experimental software（MATLAB with Psychophysics tool－ box），and monocular eye gaze was recorded with an SR Eye－ link 1000 eyetracker at 1000 Hz in order to assess changes in activation of pictured referents over time．In 24 Experimen－ tal trials，the English name of one of the three filler pictures overlapped orthographically（but not phonologically）with the orthographic form of the Colbertian target word in three out of four letters（Targets and Competitors are shown in Table 2）．Twenty－four Filler trials，in which none of the pictures overlapped orthographically or phonologically with the Col－ bertian target，were included to mask the experimental ma－ nipulation．

Finally，participants＇knowledge of Colbertian＇s letter－ sound correspondences was assessed with a novel word gen－ eralization task．In each of 48 trials，four novel Colbertian words，one target and three foils，were presented in the four corners of the screen，and the novel auditory form of the target was played over headphones．The participant selected a word and the next trial began after an inter－trial interval of 1500 ms ．Accuracy and response time were recorded，but no feed－ back was provided．Twenty－four of the trials constituted the Simple Discrimination condition，in which none of the foils used any of the target word＇s letters in the same position（e．g．， Target／suzo／spelled＇bape＇and Foils＇kovi＇，＇vedo＇，＇rina＇）．

\footnotetext{
\({ }^{2}\) To control for picture familiarity，targets，competitors，and fillers from the visual world task were viewed equally during train－ ing．Competitors never appeared with the overlapping Colbertian targets．
}

As such, knowing only one of the letters in Colbertian was sufficient to identify the target. The other 24 trials constituted the Hard Discrimination condition, where one foil overlapped the target in the first consonant, another overlapped in the second consonant, and the third overlapped in both vowels (e.g., Target / wotfæ/ spelled 'kedi', with C1 Foil 'kova', C2 Foil 'nado', Vowel Foil 'beri'). Thus, a correct response required additional knowledge of the target beyond a single letter-sound mapping. Simple and Hard Discrimination trials were presented in an intermixed fashion.

\section*{Results}

\section*{Learning}

Participants reached the \(90 \%\) criterion for whole word learning after \(\mathrm{M}=10.10\) blocks ( \(\mathrm{SD}=7.40\), Range [2, 31]). For learning the semantic meaning of the words, participants reached the \(90 \%\) criterion after only \(\mathrm{M}=3.05\) blocks ( \(\mathrm{SD}=\) 0.69 , Range [2, 4]).

Participants demonstrated high competence with Colbertian orthography on the generalization task. Accuracy was \(92 \% ~(\mathrm{SD}=8)\) in the Simple Discrimination condition, and \(75 \% ~(S D=19)\) in the Hard Discrimination condition (significantly lower, \(t(19)=4.55, p<0.001\) ). Consistent with accuracy, RTs were significantly faster in Simple Discrimination, \(\mathrm{M}=3.56\) seconds \((\mathrm{SD}=0.86)\), compared to Hard Discrimination, \(\mathrm{M}=4.20\) seconds \((\mathrm{SD}=1.39), t(19)=3.04, p<0.01\).

Though the training paradigm equated participants on Colbertian proficiency, learning rate was associated with Colbertian generalization skill. Faster learning rate in wholeword training blocks was associated with increased accuracy in Simple Discrimination, \(R^{2}=-0.23, p<0.05\), and highly associated with increased accuracy in Hard Discrimination, \(R^{2}=-0.49, p<0.001\). Faster learning rate was also associated with longer RTs in Hard Discrimination, \(R^{2}=-0.27, p<\) 0.05 , but not in Simple Discrimination, \(R^{2}=-0.03\), ns.


Figure 1: Proportion of looks to orthographic competitors compared to fillers. Asterisk denotes significance at the . 05 level, error bars indicate standard error.

\section*{Cross-Linguistic Orthographic Interference}

Proportion of Looks Visual fixations lasting at least 200 ms were analyzed (shorter fixations are mostly parts of a preplanned path for rapidly analyzing a newly-presented scene, since eye-movements in visual world tasks take about 200 ms to plan and execute, Viviani, 1990) from auditory target onset to 1600 ms post-target onset, at which point visual fixations reached an asymptote. The proportion of looks to English orthographic competitors was compared to the average of both fillers present on the same display in a one-way repeated measures ANOVA. Participants looked more often at the orthographic competitor pictures than filler pictures, \(F_{1}(1,17)=\) 17.09, \(p<0.001, F_{2}(1,23)=3.98, p=0.05\) (Figure 1).

Fixation Timecourse Proportion of looks to Competitors versus Fillers were analyzed with point-to-point t-tests in 100 ms time bins from -500 ms pre-word-onset to 2000 ms postword onset (Figure 2). Participants looked more often at orthographic Competitors than Fillers from \(0-100 \mathrm{~ms}\) and from \(100-200 \mathrm{~ms}\) post-word onset ( \(p\) 's \(<0.05\) corrected).


Figure 2: Proportion of looks to targets, orthographic competitors, and fillers in 100 ms time windows. Participants fixated orthographic competitors more than fillers from 0-200 ms post-target onset. Asterisks denote significance at the .05 level.

\section*{Discussion}

In this experiment, we examined the role of orthography on novel language learning and auditory processing. We found that participants were successfully able to learn a novel language containing letter-sound mappings that contrasted with English. Learners successfully generalized their knowledge to novel, untrained words, suggesting that they acquired Colbertian's phonetic rules and did not rely on whole-word learning alone. In fact, faster learners were also better at identifying novel words; it may be that these participants were able to
extract and make use of Colbertian's letter-sound mappings early in their training, which accelerated their learning. In contrast, those who struggled to learn the novel words appeared to have learned less about specific ortho-phono mappings, and performed more poorly identifying novel words.

Auditory presentation of learned words activated their corresponding orthographic forms, as evidenced by more frequent visual fixations to cross-linguistic English orthographic competitors from \(0-200 \mathrm{~ms}\) post target word onset. The early timecourse of the effect suggests that orthography affected speech processing online rather than at a post-lexical decision level, a finding that converges with evidence from ERPs (Perre \& Ziegler, 2008). Note that because of the contrasting letter-sound mappings between English and Colbertian, competitor items did not overlap with the target phonologically. By design, this rules out phonological competition, providing strong evidence for automatic activation of crosslinguistic orthographic competitors during spoken word processing. Overall, our findings indicate that not only were participants able to activate orthographic forms of novel words despite conflict with existing letter-sound mappings in their native language, but these words were also able to spread activation to similarly spelled words in the native language, suggesting some integration with the existing lexicon.

These findings suggest that when people hear words in one language, not only do they experience activation of the letters in that language, but that they also experience activation of words in other languages they know or are learning that are spelled similarly. In other words, when an English learner of German hears a German word that is pronounced /za:gə/ and is spelled 'sage', (conjugation of the verb 'sagen,' meaning 'to say'), the English word 'sage' (pronounced /seid3/) becomes activated due to its overlapping orthography, despite having minimal phonological overlap with the actual auditory input. This spreading co-activation of phonology and orthography across languages testifies to the highly interactive and dynamic nature of the human language system.

In the present study, since all the phonemes of the novel language also exist in English, we would expect English orthographic mappings to be more easily accessible based on their greater frequency of use. However, participants were able to activate the novel language's orthographic forms, which suggests that the language system may contain a mechanism to increase activation of newly-learned letter-sound mappings, enabling them to match or exceed mappings in the native language. Although we did find that participants activated orthographic forms of spoken words using L2 lettersound mappings (e.g., 'hane' for the spoken word /gufo/), it's unclear whether an orthographic form based on L1 lettersound mappings, such as 'goofaw' was also activated. The current study was unable to probe for this kind of L1 activation, given that an orthographic competitor, like the word 'goofy,' would also overlap with the target phonologically, obscuring orthographic effects. Overall, it's unlikely that the native language was completely suppressed during the task,
given that we saw fixations to competitors in the visual display based on L1 lexical knowledge, which suggests that both the novel language and the native language remained active to some degree.

The present results indicate that orthographic information plays an important role during second language learning and auditory word processing. Future work should investigate how different types of language experience affect learning and processing of a novel orthography. The English monolinguals in the current study had moderate experience with contrasting letter-sound mappings (e.g., the phoneme /s/ can be represented by either ' \(S\) ' or ' \(C\) '), compared to speakers of a transparent language (low experience) or bilinguals (high experience). Transparent languages with nearly one-to-one mappings between orthography and phonology, like Italian and Finnish, may not prepare speakers well for acquisition of contrasting mappings in a novel language, resulting in more cross-linguistic interference. On the other hand, bilinguals should acquire novel mappings faster and exhibit less interference compared to English monolinguals, since bilinguals already have experience with two sets of letter-sound correspondences. Indeed, bilinguals learn novel languages better than monolinguals (Cenoz, 2003; Cenoz \& Valencia, 1994; Kaushanskaya \& Marian, 2009a, 2009b; Sanz, 2000; Thomas, 1992; Van Hell \& Mahn, 1997) and can control phonological competition more efficiently (Bartolotti \& Marian, 2012), and it's likely that these advantages will extend to acquisition of a novel orthography. In sum, language perception and learning can be shaped by existing language knowledge across modalities, which emphasizes the highly interactive nature of the language system.

In conclusion, our results show that orthography can be activated online during auditory word processing, and furthermore, that the individual links between letters and sounds can be updated as part of learning a second language. Acquiring the orthographic and phonological systems of a new language is an important step in achieving proficiency. Identifying both how previous experience with language may affect acquisition of novel letter-sound correspondences, and the rate at which novel words become integrated in the lexicon, will help uncover the essential components to successful language learning.

\section*{Acknowledgments}

This research was funded in part by grant NICHD RO1 HD059858-01A to the third author. The authors would like to acknowledge Anthony Shook, Scott Schroeder, Sarah Chabal, Jen Krizman, and Tuan Lam for comments on an earlier draft of this paper.

\section*{References}

Bartolotti, J., \& Marian, V. (2012). Language learning and control in monolinguals and bilinguals. Cognitive Science, 36, 1129-1147. doi:10.1111/j.1551-6709.20 12.01243.x

Bates, E., D’Amico, S., Jacobsen, T., Székely, A., Andonova, E., Devescovi, A., ... Tzeng, O. (2003). Timed picture naming in seven languages. Psychonomic Bulletin \& Review, 10(2), 344-380. doi:10.3758/BF03196494
Bitan, T., \& Karni, A. (2003). Alphabetical knowledge from whole words training: Effects of explicit instruction and implicit experience on learning script segmentation. Brain Research, 16(3), 323-337. doi:10.1016/S 0926-6410(02)00301-4
Brysbaert, M., \& New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behavior Research Methods, 41(4), 977-990. doi:10.3758/BRM.41.4.977
Cenoz, J. (2003). The additive effect of bilingualism on third language acquisition: A review. International Journal of Bilingualism, 7(1), 71-87. doi:10.1177/F1367006 9030070010501
Cenoz, J., \& Valencia, J. F. (1994). Additive trilingualism: Evidence from the Basque Country. Applied Psycholinguistics, 15(02), 195-207. doi:10.1017/FS014 2716400005324
Coltheart, M. (1981). The MRC psycholinguistic database. Quartely Journal of Experimental Psychology, 33(4), 497-505. doi:10.1080/14640748108400805
Comer, W., \& Murphy-Lee, M. (2004). Letter-sound correspondence acquisition in first semester Russian. Canadian Slavonic Papers, 46(1), 23-35.
Davis, C. J. (2005). N-watch: A program for deriving neighborhood size and other psycholinguistic statistics. Behavior Research Methods, 37(1), 65-70. doi:10. 375 8/BF03206399
Hallé, P. A., Chéreau, C., \& Segui, J. (2000). Where is the /b/ in absurde [apsyrd]? It is in French listeners' minds. Journal of Memory and Language, 43(4), 618-639. doi:10.1006/jmla. 2000.2718
Jakimik, J, Cole, R., \& Rudnicky, A. (1985). Sound and spelling in spoken word recognition. Journal of Memory and Language, 24(2), 165-178. Retrieved from http://www. sciencedirect. com/science / art icle/pii/0749596X85900221
Johnston, M., McKague, M., \& Pratt, C. (2004). Evidence for an automatic orthographic code in the processing of visually novel word forms. Language and Cognitive Processes, 19(2), 273-317. doi:10.1080/0169096034 4000189
Kaushanskaya, M., \& Marian, V. (2008). Mapping phonological information from auditory to written modality during foreign vocabulary learning. Annals of the New York Academy of Sciences, 1145, 56-70. doi:10.1196/annals.1416.008
Kaushanskaya, M., \& Marian, V. (2009a). Bilingualism reduces native-language interference during novel-word
learning. Journal of Experimental Psychology: General, 35(3), 829-835. doi:10.1037/a0015275
Kaushanskaya, M., \& Marian, V. (2009b). The bilingual advantage in novel word learning. Psychonomic Bulletin \& Review, 16(4), 705-10. doi:10.3758/PBR.16.4.7 05
Marian, V., Blumenfeld, H. K., \& Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. Journal of Speech, Language, and Hearing Research, 50, 940-967. doi:10.1044/1092-4 388 (2007/067)
McKague, M, Pratt, C, \& Johnston, M. B. (2001). The effect of oral vocabulary on reading visually novel words: A comparison of the dual-route-cascaded and triangle frameworks. Cognition, 80(3), 231-62. doi:10 . 101 6/S0010-0277(00)00150-5
Morais, J, Cary, L, Alegria, J, \& Bertelson, P. (1979). Does awareness of speech as a sequence of phones arise spontaneously? Cognition, 7, 323-331. Retrieved from http://www.sciencedirect.com/science/artic le/pii/0010027779900209
Perre, L., \& Ziegler, J. C. (2008). On-line activation of orthography in spoken word recognition. Brain Research, 1188, 132-138. doi:10.1016/j.brainres.2007.10. 084
Salverda, A. P., \& Tanenhaus, M. K. (2010). Tracking the time course of orthographic information in spokenword recognition. Journal of Experimental Psychology: Learning, Memory, \& Cognition, 36(5), 11081117. doi:10.1037/a0019901

Sanz, C. (2000). Bilingual education enhances third language acquisition: Evidence from Catalonia. Applied Psycholinguistics, 21(01), 23-44. doi:10. 1017 / S01427 16400001028
Thomas, J. (1992). Cognitive processing in bilinguals. In Advances in psychology (Vol. 83, pp. 531-545). Advances in Psychology. Elsevier.
Vaden, K., Halpin, H., \& Hickok, G. (2009). Irvine Phonotactic Online Dictionary, Version 2.0. [Data file]. Available from http://www.iphod.com.
Van Hell, J. G., \& Mahn, A. C. (1997). Keyword mnemonics versus rote rehearsal: Learning concrete and abstract foreign words by experienced and inexperienced learners. Language Learning, 47(3), 507-546. doi:10.1111/0023-8333.00018
Viviani, P. (1990). Eye movements in visual search: Cognitive, perceptual and motor control aspects. Reviews of Oculomotor Research, 4, 353-93. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7492533

\title{
Syntactic category disambiguation within an architecture of human language processing
}

\author{
Peter Baumann (baumann@u.northwestern.edu) \\ Northwestern University \\ Department of Linguistics, 2016 Sheridan Road \\ Evanston, IL 60208, USA
}

\begin{abstract}
Syntactic category ambiguities are very frequent in natural languages, and all architectures of language processing need a mechanism for disambiguating syntactic category ambiguities. Corley and Crocker (2000) suggested that syntactic category disambiguation can be assigned its own module within a modular architecture. We will show that the model defined by Corley and Crocker can account for a considerable amount of variance in reading times of naturally occurring texts. In addition, we provide evidence that syntactic category disambiguation may be independent of syntactic top-down expectations, emphasizing the important role of bottom-up processes within an architecture of human language processing.
\end{abstract}

Keywords: sentence processing; reading; eye-tracking; ambiguity; lexical access.

\section*{Introduction}

Successful language processing requires the integration of bottom-up information extracted from the current input and top-down expectations generated from what has been processed so far. When and how bottom-up and top-down processes interact has been a distinguishing feature of different processing architectures. On the one hand, there are constraint-based models (e.g. MacDonald, Pearlmutter, \& Seidenberg, 1994; Trueswell \& Tanenhaus, 1994; Tabor, Juliano, \& Tanenhaus, 1997), which assume one single processing unit, in which all available information is considered simultaneously. Modular architectures, on the other hand, consist of several distinct processing modules (e.g. Frazier, 1987; Frazier \& Clifton, 1996; Corley \& Crocker, 2000). These modules are restricted to each having its own internal representation, and they are independently predictive and informationally encapsulated (Crocker \& Corley, 2002). Assuming this definition of modules in terms of information flow, bottom-up processes are more likely to be modular than top-down processes (Appelbaum, 1998; Fodor, 1983).

One particular process, for which constraint-based and modular models make contradicting predictions, is syntactic category assignment or disambiguation: constraint-based models assume that rich contextual information is utilized to determine the syntactic category (i.e part of speech) of a word, while modular architectures only allow context-independent information. Although previous research may seem to have provided evidence for both positions, Corley and Crocker (2000) (see also Gibson, 2006) have shown that most of the evidence
for constraint-based models can also be accounted for under a modular architecture with a module for bottom-up syntactic category assignment. In this paper, we follow Corley and Crocker's proposal and provide further evidence for the existence of a syntactic category module by showing that Corley and Crocker's model of syntactic category disambiguation is a significant predictor of reading times in naturally occurring texts. In addition, we provide evidence that syntactic category disambiguation may be independent of syntactic top-down expectations, emphasizing the critical role of bottom-up processes within a modular architecture of human language processing.

\section*{Syntactic Category Ambiguity}

Many words in English (and presumably all other languages) are ambiguous, they can have different senses and/or belong to different syntactic categories or part-of-speech (i.e. noun, verb, adjective, etc.). The following example (from Boland, 1997) illustrates these ambiguities:
(1) I saw her duck ...
a. ... under the porch to eat some potato chips.
b. .... under the porch eat some potato chips

In (1), the word duck is ambiguous between its verb and noun readings, and only the following context can disambiguate between the two syntactic categories and senses. Syntactic category ambiguity and lexical ambiguity (in terms of different senses) need not come together like in (1). Lexical ambiguity often occurs within the same syntactic category as in the word cabinet, which as a noun can denote either a group of advisors or a closet. Syntactic category ambiguity, on the other hand, does not require lexical ambiguity, as evidenced by the English verbal system, where for all regular verbs there is only one form for the past-tense and the past-participle. This ambiguity is crucial to many garden-path sentences.

The horse raced past the barn fell.
The horse ridden past the barn fell.
While example (2) is a classical garden-path sentence, which upon first encounter may be nearly impossible to understand, example (3) is unambiguous and relatively easy to process. The fact that example (2) is derived
from (3) only by replacing the ambiguous word raced with the unambiguous ridden demonstrates the important role of syntactic category disambiguation in language processing (cf. Chomsky \& Lasnik, 1977).

\section*{Previous Research}

One particular type of syntactic category ambiguity, which has received considerable attention in research is the noun-verb ambiguity. Based on three experiments, Frazier (1987) suggested that the processor delays resolving the ambiguity until disambiguating information is encountered, as readers spent less time on ambiguous words and more time on disambiguating context than on unambiguous words and their respective contexts. These results were put into question by MacDonald (1993) (see also MacDonald, 1993), who in turn argued that different statistical measures and biases such as semantic biases, syntactic context and word co-occurrences could influence syntactic category disambiguation. Similarly, Tabor et al. (1997) showed that readers are sensitive to syntactic context when resolving syntactic category ambiguities between the determiner and complementizer readings of that: a reading time delay occurred when that following a verb was disambiguated as a determiner or sentence-initial that was disambiguated as a complementizer.

While the results cited so far suggest that syntactic category disambiguation is - at least to some degree - dependent on syntactic or discourse context, Boland (1997) and Boland and Blodgett (2001) demonstrated in a series of experiments that when reading a syntactic category ambiguous word like duck, readers are sensitive to its lexical bias, i.e. the relative frequencies of the lexical entries for this word, independent of the syntactic or discourse context it appears in. In a similar vein, Stolterfoht, Gese, and Maienborn (2010) showed that for German adjectival passives (e.g. closed), whose forms are ambiguous between passive participle and adjective, there is an increase in reading times when preceded by an adjective-copula auxiliary as compared to the passive auxiliary, and as compared to unambiguous adjectives. This suggests that syntactic category disambiguation has a strong bottom-up component, which cannot be overridden by any top-down information. It is thus rather uncontroversial that lexical bias plays an important role in syntactic category disambiguation (cf. e.g. Gibson, 2006).

However, it remains open to what extent additional contextual information is used in this process: Gibson (2006) proposed that in addition to the contextindependent lexical bias syntactic category disambiguation is also affected by context-dependent syntactic expectations, which he broadly formalizes as the probability of a syntactic category in a given 'syntactic environment'. A more restrictive notion of sufficient contextual information in syntactic category disambiguation, which
was proposed by Corley and Crocker (2000), will be introduced in the next section and forms the basis of this paper.

\section*{The Statistical Lexical Category Module}

One curious fact about syntactic category disambiguation is that computers seem to be nearly as good at it as humans are: unlike many other tasks in natural language processing, part-of-speech tagging has been an area in which rather simple models can achieve near-ceiling accuracy (Charniak, 1993). Inspired by this observation, Corley and Crocker (2000) assumed that syntactic category disambiguation is distinct from syntactic parsing. Reasons for this assumption are that syntactic category disambiguation happens extremely locally, that the relevant statistics are different from syntactic parsing, and that syntactic category disambiguation does not involve structure building. This means that syntactic category disambiguation can have its own internal representation, be informationally encapsulated and independently predictive, thus constituting the requirements for a separate module, the Statistical Lexical Category Module \({ }^{1}\) (Corley \& Crocker, 2000).

Corley and Crocker's model for the Statistical Lexical Category Module (SLCM) is based on a simple bigram statistical part-of-speech tagger defined by Equation 1, which expresses the assumption that the joint probability \(P\left(t_{0}, \ldots, t_{n}, w_{0} \ldots w_{n}\right)\) of all part-of-speech tags \(t_{0}, \ldots, t_{n}\) and all words \(w_{0} \ldots w_{n}\) read so far can be reasonably approximated by the product of the lexical bias (i.e. the probability of word \(w_{i}\) given tag \(t_{i}\) ) and the category bigram transitional probability.
\[
\begin{equation*}
P\left(t_{1}, \ldots, t_{n}, w_{0} \ldots w_{n}\right) \approx \prod_{i=1}^{n} P\left(w_{i} \mid t_{i}\right) P\left(t_{i} \mid t_{i-1}\right) \tag{1}
\end{equation*}
\]

Since lexical bias \(P\left(w_{i} \mid t_{i}\right)\) is a property of the word, the category bigram transitional probability \(P\left(t_{i} \mid t_{i-1}\right)\) is the only means to capture context-dependence in this model of syntactic category disambiguation, implying that syntactic context-dependence is in fact only a dependence on the syntactic category of the preceding word.

One may object that limiting context-dependence to the category of only the preceding word is a too restrictive assumption, but Corley and Crocker (2000) (see also Crocker \& Corley, 2002) showed that it is enough to model the results reported by MacDonald (1993) and Tabor et al. (1997).

However, the aim of this paper is not to try to explain all psycholinguistic evidence involving syntactic category disambiguities. Instead, we will evaluate Corley and Crocker's SLCM model on a larger scale as a predictor of reading times in naturally occurring text. While Corley

\footnotetext{
\({ }^{1}\) Corley and Crocker (2000) refer to syntactic category ambiguity as 'lexical category ambiguity'.
}
and Crocker assume a direct link between the probabilities derived from Equation 1 and human processing difficulties, we follow common practice (e.g. Demberg \& Keller, 2008; Pynte, New, \& Kennedy, 2008) and take the logarithm as the linking function between probabilities and reading times.

We thus obtain the following measure \(\log P_{S L C M}\) for a word \(w_{i}\) given its tag \(t_{i}\) and the tag \(t_{i-1}\) of the previous word:
\[
\begin{equation*}
\log P_{S L C M}=\log P\left(w_{i} \mid t_{i}\right)+\log P\left(t_{i} \mid t_{i-1}\right) \tag{2}
\end{equation*}
\]

This measure is evaluated in Experiment 1, where we show that it is a significant predictor of reading times in naturally occurring texts. In Experiment 2, we evaluate both terms in Equation 2 separately and show that lexical bias and category bigram transitional probabilities make independent contributions to the model fit observed in Experiment 1. In the final experiment, we provide evidence that syntactic category disambiguation may be independent of syntactic top-down expectations as measured by surprisal (Hale, 2001) based on a probabilistic context-free grammar.

\section*{Experiments}

In recent years, it has become standard to evaluate computational models of language processing on 'eyetracking corpora', i.e. on eye-tracking data of people reading naturally occurring texts (Pynte et al., 2008; Demberg \& Keller, 2008). The basic idea is to fit two regression models to a measure of readings times. One regression model (baseline model) includes as predictors control variables, which are known to have an influence on reading times. The second regression model includes all those predictors as well, but in addition it also includes our computational model of language processing as a predictor. To test whether our computational model of language processing is a significant predictor we compare the fit of the two regression models to the data by means of a log-likelihood test.

\section*{Methods}

In this section we describe the methodological detail common across all three experiments.

Data and Dependent Variable All three experiments use the Dundee Corpus (Kennedy \& Pynte, 2004), a collection of eye-movement data from 10 participants reading 51,501 words each of the British newspaper The Independent. We approximated lexical categories by part-of-speech (PoS) tags, which were obtained by tagging the Dundee Corpus with the CLAWS tagger (Garside, 1987). Since syntactic category disambiguation is assumed to happen 'early' in processing, we chose first-pass reading times as our dependent variable. Firstpass reading times are calculated for a given word and participant as the sum of all eye fixations on that word
in the first pass, i.e. before leaving the word either to the right or to the left. Data points were removed if a word was not fixated, appeared as the first or last word in a line, or contained any non-letter symbol.

Control Variables All regression models included the following control variables, which are known to have an influence on reading times (c.f. Demberg \& Keller, 2008): number of characters per word, position of word in a sentence, an indicator variable whether the previous word was not fixated, and indicator variable whether the following word was not fixated, the frequency of the word, the frequency of the previous word, the forward transitional probability, i.e. bigram probability \(P\left(w_{i} \mid w_{i-1}\right)\), and the backward transitional probability \(P\left(w_{i} \mid w_{i+1}\right)\). All frequencies and transitional probabilities were obtained by fitting a unigram or bigram model with modified Kneser-Ney smoothing (Chen \& Goodman, 1998) to the British National Corpus (100 million words) using the SRILM toolkit (Stolcke, 2002). All continuous variables were centered and scaled to two standard deviations to minimize collinearity. In addition, all frequencies and transitional probabilities were logtransformed before scaling.
Estimating Probabilities in the SLCM Model The probabilities in Equation 2 were estimated from a corpus obtained by concatenating the CLAWS-tagged versions of the British National Corpus and the Dundee Corpus. The lexical bias \(P\left(w_{i} \mid t_{i}\right)\) was estimated as is, i.e. without any smoothing. For estimating the the category bigram transitional probability \(P\left(t_{i} \mid t_{i-1}\right)\) we again used a bigram model with modified Kneser-Ney smoothing.

Regression Models For the regression models we used linear 'mixed-effects' models (Pinheiro \& Bates, 2000; Gelman \& Hill, 2007) of first-pass reading times with participant, word and text number as random effects, as a generalization of the common by-subject and by-item analyses, thus taking into account that the different words and texts read by the participants are random samples in the same sense as the participants are (cf. Clark, 1973). All models were fit in R (R Development Core Team, 2011) using the lme4 package (Bates, 2005).

\section*{Baseline Model Results}

The coefficients and standard errors of the baseline model are shown in Table 1. The coefficients are as expected based on prior research: e.g. reading times decreases with increasing position in the sentence and increasing word frequency, and increase with an increasing number of characters in a word.

Table 1: Baseline model coefficients
\begin{tabular}{lrrr}
\hline Predictor & Coeff. & Std.Error & \(t\) \\
\hline (Intercept) & 206.34 & 7.31 & 28.22 \\
Position in Sentence & -6.02 & 0.51 & -11.76 \\
Number of Characters & 51.68 & 1.16 & 44.45 \\
Frequency of Word & -23.89 & 1.58 & -15.15 \\
Freq. of Prev. Word & -12.84 & 0.61 & -20.90 \\
Forward Trans. Prob & -10.24 & 0.94 & -10.94 \\
Backward Trans. Prob. & -1.95 & 0.70 & -2.77 \\
No Fixation Next & 10.14 & 0.49 & 20.61 \\
No Fixation Previous & 27.84 & 0.52 & 53.70 \\
\hline
\end{tabular}

\section*{Experiment 1}

The objective of Experiment 1 is to evaluate Corley and Crocker's model of the SLCM as a predictor of reading times. The predictor to be evaluated is the full model as stated in Equation 2.

Figure 1: Partial effect of full SLCM model with all other predictors held constant


Results The coefficient and standard error of the full tagger-based model of syntactic category disambiguation (Equation 2) are shown in Table \(2^{2}\). A log-likelihood test between the regression model with the predictor \(\log P_{S L C M}\) and the baseline model confirmed that Equation 2 is a significant predictor of reading times \(\left(\chi^{2}=\right.\) \(29.955, p<.0001)\). The relation between \(\log P_{S L C M}\) and reading times is plotted in Figure 1, which shows the

\footnotetext{
\({ }^{2}\) Coefficients for the control variables are not listed as they are qualitatively similar to the ones reported for the baseline model.
}

Table 2: Model coefficient of full SLCM model
\begin{tabular}{lrrr}
\hline Predictor & Coeff. & Std.Error & \(t\) \\
\hline \(\log P_{S L C M}\) & -6.85 & 1.05 & -6.54 \\
\hline
\end{tabular}
partial effect of \(\log P_{S L C M}\) with all other predictors held constant at their respective means. It can be seen that reading times increase as \(\log P_{S L C M}\) or \(P_{S L C M}\) decrease.

Discussion The above result shows that the simple model of syntactic category disambiguation in Equation 2 cannot only account for many empirical results in psycholinguistic experiments, but is also a significant predictor of reading times in naturally occurring text. The direction of the effect is in line with experimental evidence modeled by Corley and Crocker (2000) in the sense that a lower probability in Equation 2 leads to higher reading times.

\section*{Experiment 2}

In Experiment 2 we investigate whether lexical bias and category bigram transitional probability are also independently significant as predictors of reading times. To test this hypothesis, we fitted three models, one with only lexical bias (log-transformed \(P\left(w_{i} \mid t_{i}\right)\) ), one with only category bigram transitional probability (logtransformed \(P\left(t_{i} \mid t_{i-1}\right)\) ), and a third one with both terms as additional predictors to the baseline model.

Results The coefficients and standard errors for lexical bias and category bigram transitional probability are shown in Table 3. The negative coefficients indicate that increasing the lexical bias (i.e. making the 'correct' category more likely) and increasing the category bigram transitional probability both lead to shorter reading times. A log-likelihood test confirmed that a model with either lexical bias \(\left(\chi^{2}=7.37, p<.001\right)\) or category bigram transitional probability \(\left(\chi^{2}=22.97, p<.0001\right)\) yields a significantly better fit to the data than the baseline model, and that a model with both predictors significantly improves over a model with only one.

Discussion Our results show that both lexical bias and category bigram transitional probability are significant predictors of reading times. For lexical bias this is in line with the results of Boland (1997) and Boland

Table 3: Model coefficient for lexical bias and category bigram probabilities
\begin{tabular}{lrrr}
\hline Predictor & Coeff. & Std.Error & \(t\) \\
\hline Lexical Bias & -4.86 & 1.57 & -3.09 \\
Category Bigram & -4.28 & 0.69 & -6.18 \\
\hline
\end{tabular}
and Blodgett (2001), who also found a significant effect of lexical bias on reading times. The effect of category bigram transitional probabilities shows that the immediately preceding category contains information beyond what is contained in the corresponding preceding word, as including category bigram transitional probabilities improves over a baseline model, which already contained word bigram transitional probabilities.

\section*{Experiment 3}

In Experiment 3 we test whether the effects of syntactic category disambiguation accounted for by the SLCM model can be ascribed to syntactic top-down expectations. If this were the case, it would provide strong evidence against any modular approach to syntactic category disambiguation. Syntactic top-down expectations are often measured by surprisal (Hale, 2001), which can be calculated from a probabilistic context-free grammar.

We calculated unlexicalized surprisal values for all words in the Dundee Corpus using the top-down parser described in (Roark, 2001) and (Roark, Bachrach, Cardenas, \& Pallier, 2009) and included it as an additional predictor in our baseline model. We than compared this enriched baseline model to a regression model, which contained both surprisal and the log-probabilities of the tagger-based model of syntactic category disambiguation (Equation 2).

Results The coefficients and standard errors for surprisal and the tagger-based model of syntactic category disambiguation are shown in Table 4. As in Experiment 1 , the coefficient of the tagger-based model is negative coefficients indicating that increasing the probability in Equation 1 leads to shorter reading times. The coefficient of surprisal is positive. This is expected as higher surprisal is associated with longer reading times (Hale, 2001; Demberg \& Keller, 2008). A log-likelihood test confirmed that a model with the tagger-based model and surprisal improves significantly over a baseline model with only surprisal \(\left(\chi^{2}=13.18, p<.001\right)\).
Discussion The above results show that SLCM model is a significant predictor of reading times even if surprisal is included in the baseline regression model. Although this does not rule out the hypothesis that the effects of syntactic category disambiguation accounted for by the SLCM model may be reduced to syntactic top-down expectations, it provides strong evidence against such a hy-

Table 4: Model coefficient for surprisal and the full SLCM model
\begin{tabular}{lrrr}
\hline Predictor & Coeff. & Std.Error & \(t\) \\
\hline Surprisal & 2.17 & 0.69 & 3.13 \\
\(\log P_{S L C M}\) & -5.67 & 1.11 & -5.10 \\
\hline
\end{tabular}
pothesis, and suggests instead that syntactic top-down expectations and bottom-up syntactic category disambiguation may be independent processes, as suggested by Gibson (2006) and Corley and Crocker (2000).

\section*{General Discussion}

In our experiments, we have shown that the model of a Statistical Lexical Category Module as formulated by Corley and Crocker (2000) is a significant predictor of reading times in naturally occurring texts. While our results do not necessarily imply that syntactic category disambiguation is a separate module, they provide further evidence for modular models relying on simple context-independent statistics for lexical category disambiguation. The observation that SLCM model is a significant predictor of reading times in addition to syntactic expectations as measured by surprisal indicates that Corley and Crocker's model may indeed account for bottom-up processes in reading, while surprisal accounts for top-down processes.

Since any architecture of language processing needs to integrate bottom-up and top-down processes, one may conclude that the combination of a restricted (or modular) model of bottom-up syntactic category disambiguation with a model of syntactic top-down expectations may ultimately lead to better models of the architecture of human language processing and, more specifically, to a better understanding of syntactic category disambiguation as a phenomenon at interface of lexical access and syntactic processing, as recent experiments have shown that syntactic category ambiguity also plays a crucial rule in lexical-semantic access and disambiguation (Jones, Folk, \& Brusnighan, 2012).

Finally, our results may also contribute to the ongoing debate on lexicalized vs. unlexicalized measures of syntactic expectations and their reflections in reading times (for a review, see Roark et al., 2009): since bigram probabilities are the simplest form of syntactic expectations, our observation that category bigram probabilities are a significant predictor of reading times, even if controlled for word bigram probabilities, suggests that lexicalized and unlexicalized measures of syntactic expectations may have independent contributions to reading times.

\section*{Acknowledgments}

The author would like to thank the four anonymous reviewers for their suggestions and Masaya Yoshida and Daniel Müller-Feldmeth for many fruitful discussions.

\section*{References}

Appelbaum, I. (1998). Fodor, modularity, and speech perception. Philosophical Psychology, 11 (3), 317-330.
Bates, D. M. (2005). Fitting linear mixed models in R: Using the lme4 package. \(R\) News: The Newsletter of the \(R\) Project, 5(1), 27-30.

Boland, J. E. (1997). Resolving syntactic category ambiguities in discourse context: Probabilistic and discourse constraints. Journal of Memory and Language, 36, 588-615.
Boland, J. E., \& Blodgett, A. (2001). Understanding the constraints on syntactic generation: Lexical bias and discourse congruency effects on eye movements. Journal of Memory and Language, 45, 391-411.
Charniak, E. (1993). Statistical Language Learning. Cambridge: MIT Press.
Chen, S. F., \& Goodman, J. (1998). An empirical study of smoothing techniques for language modeling (Tech. Rep.). Center for Research in Computing Technology, Harvard University.
Chomsky, N., \& Lasnik, H. (1977). Filters and control. Linguistic Inquiry, 8(3), 425-504.
Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. Journal of Verbal Learning and Verbal Behavior, 12(4), 335-359.
Corley, S., \& Crocker, M. W. (2000). The Modular Statistical Hypothesis: Exploring Lexical Category Ambiguity. In M. W. Crocker, M. Pickering, \& C. Clifton (Eds.), Architectures and Mechanisms for Language Processing (pp. 135-160). Cambridge: Cambridge University Press.
Crocker, M. W., \& Corley, S. (2002). Modular Architectures and Statistical Mechanisms: The Case from Lexical Category Disambiguation. In P. Merlo \& S. Stevenson (Eds.), The Lexical Basis of Sentence Processing: Formal, computational and experimental issues (pp. 157-180). Amsterdam: John Benjamins Publishing.
Demberg, V., \& Keller, F. (2008). Data from eyetracking corpora as evidence for theories of syntactic processing complexity. Cognition, \(109(2), 193-210\).
Fodor, J. A. (1983). The Modularity of Mind. Cambridge: MIT Press.
Frazier, L. (1987). Theories of Sentence Processing. In J. Garfield (Ed.), Modularity in Knowledge Representation and Natural Language Processing (pp. 291308). Cambridge: MIT Press.

Frazier, L., \& Clifton, C. (1996). Construal. Cambridge: MIT Press.
Garside, R. (1987). The CLAWS Word-tagging System. In R. Garside, G. Leech, \& G. Sampson (Eds.), The Computational Analysis of English: A Corpus-based Approach. London: Longman.
Gelman, A., \& Hill, J. (2007). Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge: Cambridge University Press.
Gibson, E. (2006). The interaction of top-down and bottom-up statistics in the resolution of syntactic category ambiguity. Journal of Memory and Language, 54, 363-388.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In Proceedings of the 2nd Conference of the North American Chapter of the Association for Computational Linguistics (pp. 159-166). Pittsburgh, PA.
Jones, A., Folk, J., \& Brusnighan, S. (2012). Resolving syntactic category ambiguity: An eye- movement analysis. Journal of Cognitive Psychology, 24 (6), 672688.

Kennedy, A., \& Pynte, J. (2004). Parafoveal-on-foveal effects in normal reading. Vision Reserach, 45, 153168.

MacDonald, M. (1993). The interaction of lexical and syntactic ambiguity. Journal of Memory and Language, 32, 692-715.
MacDonald, M., Pearlmutter, N., \& Seidenberg, M. (1994). The lexical nature of syntactic ambiguity resolution. Psychological Review, 101 (4), 676-703.
Pinheiro, J. C., \& Bates, D. M. (2000). Mixed Effects Models in \(S\) and S-Plus. New York: Springer.
Pynte, J., New, B., \& Kennedy, A. (2008). A multiple regression analysis of syntactic and semantic influences in reading normal text. Journal of Eye Movement Research, 2(4), 1-11.
R Development Core Team. (2011). R: A Language and Environment for Statistical Computing [Computer software manual]. Vienna.
Roark, B. (2001). Probabilistic top-down parsing and language modeling. Computational Linguistics, 27(2), 249-276.
Roark, B., Bachrach, A., Cardenas, C., \& Pallier, C. (2009). Deriving lexical and syntactic expectationbased measures for psycholinguistic modeling via incremental top-down parsing. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Singapore.
Stolcke, A. (2002). SRILM - an extensible language modeling toolkit. In Seventh international conference on spoken language processing.
Stolterfoht, B., Gese, H., \& Maienborn, C. (2010). Word category conversion causes processing costs: Evidence from adjectival passives. Psychonomic Bulletin \& Review, 17(5), 651-656.
Tabor, W., Juliano, C., \& Tanenhaus, M. K. (1997). Parsing in a dynamical system: An attractor-based account of the interaction of lexical and structural constraints in sentence processing. Language and Cognitive Processes, 12(211-272).
Trueswell, J., \& Tanenhaus, M. (1994). Toward a lexical framework of constraint-based syntactic ambiguity resolution. In C. Clifton \& L. Frazier (Eds.), Perspectives on Sentence Processing (pp. 155-179). Mahwah: Erlbaum.

\title{
Roles of Diagrams in Computational Modeling of Mechanisms
}

\author{
William Bechtel (bechtel@ucsd.edu) \\ Department of Philosophy and Center for Chronobiology, University of California, San Diego, La Jolla, CA 92093-0119 USA
}

\section*{Adele Abrahamsen (aabrahamsen@ucsd.edu)}

Center for Research in Language, University of California, San Diego
La Jolla, CA 92093-0119 USA

\begin{abstract}
As tools in science, diagrams not only serve as vehicles for communication but also facilitate and constrain scientific reasoning. We identify roles that diagrams play when computational models and synthesized organisms are used to recompose mechanisms proposed to explain biological phenomena. Diagrams not only serve as locality aids for constructing computational models but also help in identifying ways to manipulate these models and interpret the results. Moreover, they serve as blueprints for constructing synthetic organisms and then guide the interpretation of discrepancies between these organisms and computational models.
\end{abstract}

Keywords: diagrams; computational models; mechanistic explanation; circadian rhythms

\section*{Introduction}

Cognitive scientists have contributed analyses and experiments on the roles diagrams play in reasoning and problem solving (e.g., Hegarty, 2004, 2011; Tversky, 2011) and have even designed new diagram formats that facilitate learning in math and science (Cheng, 2002, 2011). However, there have been only a few studies of the roles diagrams play in the natural sciences (Nersessian, 2008; Gooding, 2010). The most obvious role, evidenced by the ubiquity of diagrams in talks and publications, is communication of methods, results, and proposed mechanistic explanations (Perini, 2005). Less visibly, but crucially, diagramming is a tool that scientists use to reason about phenomena (Bechtel \& Abrahamsen, 2012) and the mechanisms that might explain them (Sheredos, Burnston, Abrahamsen, \& Bechtel, in press).

In many fields of biology, such as cell and molecular biology, the primary goal of research in the \(19^{\text {th }}\) and \(20^{\text {th }}\) centuries was to identify and decompose mechanisms to determine their parts (e.g., proteins) and operations (e.g., catalyzing particular chemical reactions). As recognized in the new mechanistic philosophy of science, the organization of these parts and operations must also be determined to arrive at a basic mechanistic explanation of a phenomenon of interest (Bechtel \& Richardson, 1993/2010; Bechtel \& Abrahamsen, 2005; Machamer, Darden, \& Craver, 2000; Thagard, 2003). That is, to understand how the parts and operations contribute to producing the phenomenon, researchers must recompose the responsible mechanism either conceptually or physically. Through most of the \(20^{\text {th }}\) century this involved proposing a simple sequence in which the operations might occur, perhaps using mental simulation to verify its plausi-
bility (Bechtel, 2006). By the last decades of the century, however, the operations of numerous biological mechanisms were understood to display nonlinear, continuous dynamics and complex interactions. As sequential organization broke down, so too did biologists’ ability to mentally track the functioning of the proposed mechanisms. Hence, they turned first to computational models and later to synthetic organisms as tools for recomposing mechanisms, with an emphasis on investigating the complex dynamics and interactions of operations by which a mechanism generates a phenomenon. In this paper we identify some of the roles diagrams play in the design of computational and synthetic models of mechanisms in actual scientific practice.

Computational modeling in biology, in contrast to that in much of cognitive science, has been grounded in considerable knowledge of the physical parts and operations of the mechanisms being targeted (Bechtel \& Abrahamsen, 2010). Diagrams showing how different parts are thought to operate on each other serve as locality aids that "group together information that is used together" in the mechanism itself and hence often in computational models of its dynamics (Jones \& Wolkenhauer, 2012, p. 705). But such diagrams also figure centrally in conceiving how manipulations made to the computational model correspond to possible perturbations of the mechanism, thereby relating experiments on models to experiments on actual mechanisms or to pathologies known to result from damage to actual mechanisms. Moreover, as the efforts to recompose mechanisms increasingly take a step beyond computational modeling to synthesizing organisms, a diagram can serve both as a blueprint for synthesizing an organism and as a medium for adjudicating mismatches in behavior between organism and model.

We focus on one domain of biology, circadian rhythms: the daily oscillations in a variety of physiological and behavioral processes in species ranging from bacteria and fungi to plants and animals. The phenomena of greatest interest involve three characteristics of these rhythms: they are endogenously generated, entrained to the day-night cycle on our planet, and sustained over time (not dampened). Their complex dynamics have made circadian rhythm research a model case for developing computational models and synthesized organisms to determine how a proposed mechanism might account for relevant phenomena. By examining specific exemplars of this research, we show how diagrams can play an important role in the reasoning that goes into computational modeling and synthetic biology.

\section*{Diagrams for Modeling "How-Possibly" Mechanisms}

Computational modeling of circadian rhythms began shortly after behavioral researchers determined that the daily oscillations in organisms are endogenously generated, with a period varying slightly from 24 hours (Bünning, 1960). Since engineers had shown that negative feedback systems generate oscillations, biologists were attracted to the idea that feedback loops are involved in circadian oscillations. But most feedback systems dampen, settling into a steady state. The challenge was to determine how a biological mechanism might generate sustained oscillations, which entailed computational modeling of its dynamics.

Goodwin (1963) accepted this challenge, and took as his starting point one of the first molecular feedback mechanisms identified in biology: the lac operon. Jacob and Monod (1961) had specified how synthesis of the enzymes needed to metabolize lactose could be restricted via negative feedback to occur only when glucose levels are low. Although the molecular parts and operations involved in the circadian mechanism had not yet been identified, Goodwin borrowed the architecture of the better-understood lac operon to construct a diagram depicting a possible circadian mechanism (Figure 1). In it he included not only generic labels for the putative parts and operations but also associated variables and parameters relevant to their dynamics. The mechanism has five types of molecular parts, three of which undergo changes in their concentration. These concentrations are represented by the variables \(X, Y\), and \(Z\). Arrows depict six operations that affect the concentrations: three (labeled) involve aspects of gene expression and three indicate decay of a particular type of molecule, at rates associated with the parameters \(k_{1}, \ldots k_{6}\). Thus, \(X\) is the concentration of mRNA transcribed from the gene, \(Y\) the concentration of the enzyme resulting from translating the mRNA, \(Z\) the concentration of the repressor molecule whose synthesis is catalyzed by the enzyme, \(k_{4}\) to \(k_{6}\) the rates of decay, and \(k_{1}\), to \(k_{3}\) associated with rates of gene expression operations.

There are three equations in the computational model. Each specifies the change in concentration of one molecular component by subtracting a term for its decay from a term


Figure 1. Diagram of the generic mechanism for feedback control of gene expression that Goodwin used as a locality aid in constructing his computational model of circadian rhythms (adapted from Goodwin 1963).
for the impact of one of the operations in the feedback loop. Consulting the diagram, it is easy to see which variables and parameters should be in the same equation. Each variable has one arrow from it (its decay) and one arrow to it from another variable; its equation includes that variable and the parameters on those arrows. By providing these groupings, the diagram does service as a locality aid.
\[
\begin{aligned}
& \frac{d X}{d t}=\frac{k_{1}}{Z^{n}+1}-k_{4} X \\
& \frac{d Y}{d t}=k_{2} X-k_{5} Y \\
& \frac{d Z}{d t}=k_{3} Y-k_{6} Z
\end{aligned}
\]

Five of the terms simply multiply a concentration by a rate parameter. The first term is more complex: since the repressor reduces synthesis of mRNA, its concentration \((Z)\) is in the denominator and raised to the power \(n\); known as the Hill coefficient, \(n\) represents the number of molecules that must interact. As the only nonlinear term, this first term is crucial for generating sustained oscillations.

It is difficult to determine exactly how a mechanism will behave when even one component exhibits nonlinearity and also when appropriate parameter values are not yet known. For both of these reasons, it is important to run simulations by solving the equations with different initial values and parameter settings. Doing so on an analog computer, Goodwin concluded that such a mechanism could generate sustained oscillations when \(n\) equaled 2 or 3 . These are biologically plausible values, but when Griffith (1968) ran simulations on a digital computer he determined that sustained oscillations resulted only when \(n>9\), generally recognized as biologically unrealistic. Accordingly, he concluded that negative feedback with a single gene product operating on a gene could never "give rise in practice to undamped oscillations in the concentrations of cellular constituents" (p. 207). This reasoning highlights an advantage of grounding a computational model in a representation of the associated mechanism. A biologist, having noticed that the term in question relates to molecules interacting to inhibit a biochemical reaction, can draw on knowledge of such reactions to judge the plausibility of different parameter values. Lacking such grounding, the modeler has no independent check on the values obtained from parameter fitting.

\section*{Diagrams for Modeling Known Parts and Operations}

Diagrams continued to serve as locality aids after researchers discovered some of the actual parts and operations of the circadian mechanism, and modelers turned to modeling their specific dynamics. As we will see, the diagrams also supported additional reasoning about the mechanism.

The first component part of a circadian clock was discovered by Konopka and Benzer (1971) through a process of generating mutant fruit flies with short, long, or absent circadian rhythms. They named the gene in which mutations
produced altered rhythms period (per). When cloning techniques became available, Hardin, Hall, and Rosbash (1990) were able to measure the mRNA into which per was transcribed and the protein into which it was translated. They determined that these concentrations oscillated over 24 hours, with the peak concentration of the protein lagging several hours behind that of the mRNA transcript. They thus hypothesized a feedback mechanism whereby the protein PER fed back to inhibit the transcription of the gene per.

This research physically identified some of the parts and operations of the proposed mechanism, but the "feedback hypothesis" left open the question of whether and under what specific conditions it could generate sustained oscillations. Goldbeter (1995) took up this question by developing a computational model, drawing upon Hardin et al.'s empirical discoveries and inspired in part by Goodwin's abstract model. Like Goodwin, he portrayed the mechanism in a diagram (Figure 2) in which each part and operation was accompanied by its corresponding variable or parameter. Shown within the dashed box is the operation occurring in the nucleus in which the PER protein inhibits per transcription. The rest of the diagram shows the operations of transcription and translation and an additional post-translational operation through which the protein PER is phosphorylated


Figure 2. Goldbeter's (1995) diagram that guided his computational model based on the mechanism proposed by Hardin, Hall, and Rosbash (1990).
(a step that had been determined to be necessary before PER could be transported back into the nucleus).

Like Goodwin, Goldbeter then constructed differential equations, each characterizing the change in concentration of one of the molecular components. Again, the grouping of arrows around each variable served as a locality aid in determining the equations. As a result of including additional nonlinearities in the terms representing decay, Goldbeter's model exhibited sustained oscillations using parameter values deemed biologically realistic.

In the same window of time during which Goldbeter was constructing his model, molecular researchers were searching for additional parts to fill known gaps in the mechanism. They recognized, for example, that PER could not directly inhibit its own transcription since it lacked the needed binding region. Mammalian researchers identified a gene, Clock, in which a mutation could eliminate circadian function and whose protein contained a DNA-binding region (Vitaterna, King, Chang, Kornhauser, Lowrey, McDonald, Dove, Pinto, Turek, \& Takahashi, 1994). In short order, it was found that CLOCK forms a dimer with BMAL1 that binds to the promoter region of \(\operatorname{Per}\) (as well as a second gene, Cry) and that by interacting with this dimer, PER and CRY inhibit their own transcription. Realizing that concentrations of BMAL1 oscillate, researchers hypothesized a second negative feedback loop in which it inhibited the transcription of its gene. The introduction of this additional feedback loop raised the question of whether the results of Goldbeter's (1995) simulation were still applicable: would the two loops generate sustained oscillations? To address this question, Leloup and Goldbeter (2003) constructed a diagram (Figure 3) that included a variable for the concentration of each molecular part and a rate parameter for each operation. Again, the grouping of arrows around each variable served as a locality aid. With 16 variables being tracked this time, the computational model consisted of 16 differential equations.


Figure 3. Leloup and Goldbeter's (2004) diagram of the mammalian circadian oscillator in which proteins are represented as ovals (labeled within) and operations as arrows (some identified in adjacent boxes, and all with rate parameters shown).

Leloup and Goldbeter employed their computational model not only to establish that the mechanism could generate sustained oscillations, but to determine as well whether it could account for other circadian phenomena. Of prime importance is the ability of circadian clocks to be entrained by light. Light had been shown experimentally to affect PER expression, and hence Leloup and Goldbeter incorporated light in their diagrams as a black box with an arrow feeding into the box for Per transcription. This in turn guided their strategy for simulating light exposure in the computational model: instead of a setting a single value for the parameter \(v_{s P}\), which set the maximum rate of Per expression, they used a square wave function to alternate between a high value (simulating light) and a low value (simulating darkness). Leloup and Goldbeter were then able to use their model to show that the mechanism's responses to light exposure varied with time of day in ways similar to the responses of mammals.

Leloup and Goldbeter were also interested in whether the proposed mechanism could be perturbed in ways that correspond to known circadian pathologies. Advanced sleep phase syndrome is a condition in which people naturally go to sleep around 7 PM and rise around 3 AM . Genetic studies of families with this pathology had revealed a mutation affecting the interaction of PER with a kinase that phosphorylates it. The diagram includes the parameter \(v_{I P}\) at this location, and Leloup and Goldbeter showed that they could replicate the characteristics of the pathology by altering it.

In a subsequent paper Leloup and Goldbeter (2004) explored the sensitivity of the model to variations in all of the parameters. Here the diagram facilitated identifying which operations in the actual mechanism correspond to those perturbed by varying parameters in the computational model.

A question researchers often ask when they encounter a mechanistic account is whether all of the parts are required for the phenomenon to occur. Leloup and Goldbeter questioned which of the two feedback loops in their diagram were essential for circadian rhythmicity, and explored this by setting the parameter governing PER synthesis to 0 . The model ceased to exhibit oscillation. They then explored whether oscillation could be rescued by increasing parameters regulating the synthesis of BMAL1. This restored oscillation, but with a shorter period of approximately 19 hours.

This question of what different components contribute to the generation of circadian rhythms remains one of great interest to modelers. Some have pursued the question using highly reduced models, but adopting Goldbeter's approach instead, Relógio, Westermark, Wallach, Schellenberg, Kramer, \& Herzel (2011) included in their model all of the currently identified operations in the mammalian circadian mechanism. They developed the diagram in Figure 4 as a locality aid. Like the other diagrams, it includes variables and parameters adjacent to the relevant parts and operations. An innovation is use of a dashed line to differentiate two sub-mechanisms. By running the model with targeted variables set to constant values-first those for concentrations of parts above the line and then those below-they concluded
that it was the feedback loops involving BMAL1 that were crucial to the generation of circadian rhythms.


Figure S 1
PC loop
Figure 4. Relógio et al's (2011) diagram of the mammalian circadian oscillator. They use a dotted line to differentiate two sub-mechanisms investigated in their model.

The diagrams discussed in this section all serve as locality aids in constructing computational models, but then serve additional roles in determining which variables to manipulate in various simulations and in relating simulations back to the hypothesized mechanism.

\section*{Diagrams of Mechanisms to be Synthesized}

Traditionally, biologists have been limited to analyzing extant mechanisms to determine what parts, operations, and organization are responsible for a phenomenon of interest. But the development of techniques for inserting genes into host organisms (typically, E. coli) has generated a new field of synthetic biology, in which researchers use computational models to help design regulatory networks, insert them into organisms, and assess the effects on behavior. As Cookson, Tsimring, and Hasty (2009) make explicit, diagrams play a central role in this research. In the first step "genetic wiring diagrams are translated into equations that can be analyzed." After such analysis, "modern recombinant DNA techniques are used to construct gene-regulatory networks in living cells according to the design specification." In this endeavor, diagrams are not only locality aids for developing mathematical models, but also blueprints for constructing an organism. Once the behavior of the synthesized organism can
be assessed, diagrams play a further role in analyzing that behavior and revising the network design in light of the effects discovered in the synthesized organism.

This practice is illustrated in the efforts of Stricker, Cookson, Bennett, Mather, Tsimring, \& Hasty (2008). They explicitly drew upon the mechanism understood to be operative in the fruit fly circadian clock to construct a synthetic clock in E. coli. Specifically, they added a lacZYA promoter to the naturally occurring \(\operatorname{araBAD}\) promoter and then situated the hybrid promoter on the araC, lacI, and yemGFP genes (the last generates a green fluorescent protein used as a reporter of oscillations). Before inserting this mechanism into the bacterium, Stricker et al. constructed a diagram (Figure 5) from which they developed a computational model. Satisfied that the proposed mechanism would generate sustained oscillations under a limited set of parameter values (especially, of IPTG levels), they then employed the diagram as a blueprint for synthesizing the mechanism and as a guide to what components would have to be fine-tuned to generate sustained oscillations.


Figure 5. Stricker et al.'s (2008) diagram, which they used both to develop a computational model and to synthesize a bacterium that could generate oscillations.

The organism Stricker et al. synthesized did not behave as the model had led them to expect. Most surprising, it generated sustained oscillations under almost all parameter values tested. This led Stricker et al. to return to the mechanism as represented in the diagram and question whether processes that they had not represented in the diagram or in the equations of the model, such as protein folding, multimerization, and DNA-binding, were important to the process. They constructed a new diagram (Figure 6) and computational model that incorporated additional operations. The behavior of this model now corresponded closely to that of the synthesized bacterium. Stricker et al. concluded that the delays introduced into the feedback by these additonal steps were responsible for the oscillations.

In this example from synthetic biology, the diagram serves not only as blueprint for building the mechanism but also as a guide to determining why the mechanism did not behave as expected and then for proposing an alternative account of the mechanism.


Figure 6. Stricker et al. (2008) revised diagram motivated by the discrepencies between the behavior of the synthesized organism and their computational model.

\section*{Conclusion}

We have focused on one of the contexts in which diagrams provide the basis for reasoning in the development of mechanistic explanations-recomposing mechanisms through computational models and synthesized organisms. Through examples we have identified a widespread practice of constructing a diagram of the hypothesized mechanism that includes variables and parameters and using it as a locality aid in constructing equations to model the dynamics of the mechanism. But this is only the start. One of the interests in constructing a computational model is to experiment on it to determine whether the mechanism could explain various identified phenomena. A diagram can help with this, by guiding the selection of parameters to be reset or of variables to be given fixed values. When researchers set out to synthesize organisms, diagrams function both as locality aids in developing the computational models and as blueprints guiding the determination of components to include. When a synthesized organism fails to behave as the computational model suggested, researchers returned to the diagram to explore alternatives.

Our examination of published diagrams is only a first step in understanding researchers' cognitive engagement with diagrams as they seek to recompose mechanisms. Although unlikely, we cannot rule out the possibility that the diagrams we have discussed are epiphenomenal-constructed after developing the computational model as a means of communicating it to others. Given the utility of the diagram for grounding the modeling and the experiments on the model, it seems most likely the scientists would have so used it. Having identified ways diagrams appear to function in recomposing mechanisms, our hope is that other cognitive scientists will contribute to further understanding this aspect of scientific reasoning. One strategy would be ethnographic studies of modelers in which one can observe interactions with the diagrams in the process of developing and experimenting with computational models. Another strategy would involve experiments in which some modelers were allowed to create or consult diagrams while constructing a computational model and others were restricted from doing so. Such studies may help elucidate the cognitive operations that go into the construction of computational models. Further, such studies can also go beyond what we have been
able to do and address the specific features of diagrams that serve the aims of developing computational models and whether different representations, including different diagram formats, might serve these ends better. What we hope to have done is demonstrate a widespread practice of using diagrams in constructing and experimenting with computational models of biological mechanisms.

\section*{Acknowledgments}

This material is based upon work supported by the National Science Foundation under Grant No. 1127640. We thank the other members of the WORking Group on Diagrams in Science (WORGODS), Daniel Burnston and Ben Sheredos, for constructive comments on earlier drafts.

\section*{References}

Bechtel, W. (2006). Discovering cell mechanisms: The creation of modern cell biology. Cambridge: Cambridge University Press.
Bechtel, W., \& Abrahamsen, A. (2005). Explanation: A mechanist alternative. Studies in History and Philosophy of Biological and Biomedical Sciences, 36, 421-441.
Bechtel, W., \& Abrahamsen, A. (2010). Dynamic mechanistic explanation: Computational modeling of circadian rhythms as an exemplar for cognitive science. Studies in History and Philosophy of Science Part A, 41, 321-333.
Bechtel, W., \& Abrahamsen, A. (2012). Diagramming phenomena for mechanistic explanation. Proceedings of the 34th Annual Conference of the Cognitive Science Society (pp. 102-107). Austin, TX: Cognitive Science Society.
Bechtel, W., \& Richardson, R. C. (1993/2010). Discovering complexity: Decomposition and localization as strategies in scientific research. Cambridge, MA: MIT Press. 1993 edition published by Princeton University Press.
Bünning, E. (1960). Opening address: Biological clocks. Cold Spring Harbor Symposia on Quantitative Biology, 25, 1-9.
Cheng, P. C.-H. (2002). Electrifying diagrams for learning: principles for complex representational systems. Cognitive Science, 26, 685-736.
Cheng, P. C.-H. (2011). Probably good diagrams for learning: Representational epistemic recodification of probability theory. Topics in Cognitive Science, 3, 475498.

Cookson, N. A., Tsimring, L. S., \& Hasty, J. (2009). The pedestrian watchmaker: Genetic clocks from engineered oscillators. FEBS Letters, 583, 3931-3937.
Goldbeter, A. (1995). A model for circadian oscillations in the Drosophila Period protein (PER). Proceedings of the Royal Society of London. B: Biological Sciences, 261, 319-324.
Gooding, D. C. (2010). Visualizing Scientific Inference. Topics in Cognitive Science, 2, 15-35.

Goodwin, B. C. (1963). Temporal organization in cells; a dynamic theory of cellular control processes. London: Academic.
Griffith, J. S. (1968). Mathematics of cellular control processes I. Negative feedback to one gene. Journal of Theoretical Biology, 20, 202-208.
Hardin, P. E., Hall, J. C., \& Rosbash, M. (1990). Feedback of the Drosophila period gene product on circadian cycling of its messenger RNA levels. Nature, 343, 536540.

Hegarty, M. (2004). Mechanical reasoning by mental simulation. Trends in Cognitive Science, 8, 280-285.
Hegarty, M. (2011). The cognitive science of visual-spatial displays: Implications for design. Topics in Cognitive Science, 3, 446-474.
Jacob, F., \& Monod, J. (1961). Genetic regulatory systems in the synthesis of proteins. Journal of Molecular Biology, 3, 318-356.
Jones, N., \& Wolkenhauer, O. (2012). Diagrams as locality aids for explanation and model construction in cell biology. Biology and Philosophy, 27, 705-721.
Konopka, R. J., \& Benzer, S. (1971). Clock mutants of Drosophila melanogaster. Proceedings of the National Academy of Sciences (USA), 89, 2112-2116.
Leloup, J.-C., \& Goldbeter, A. (2003). Toward a detailed computational model for the mammalian circadian clock. Proceedings of the National Academy of Sciences, 100, 7051-7056.
Leloup, J.-C., \& Goldbeter, A. (2004). Modeling the mammalian circadian clock: Sensitivity analysis and multiplicity of oscillatory mechanisms. Journal of Theoretical Biology, 230, 541-562.
Machamer, P., Darden, L., \& Craver, C. F. (2000). Thinking about mechanisms. Philosophy of Science, 67, 1-25.
Nersessian, N. (2008). Creating scientific concepts. Cambridge, MA: MIT Press.
Perini, L. (2005). Explanation in two dimensions: Diagrams and biological explanation. Biology and Philosophy, 20, 257-269.
Relógio, A., Westermark, P. O., Wallach, T., Schellenberg, K., Kramer, A., \& Herzel, H. (2011). Tuning the mammalian circadian clock: Robust synergy of two loops. PLoS Computational Biology, 7, e1002309.
Sheredos, B., Burnston, D., Abrahamsen, A., \& Bechtel, W. (in press). Why do biologists use so many diagrams. Philosophy of Science.
Stricker, J., Cookson, S., Bennett, M. R., Mather, W. H., Tsimring, L. S., \& Hasty, J. (2008). A fast, robust and tunable synthetic gene oscillator. Nature, 456, 516-519.
Thagard, P. (2003). Pathways to biomedical discovery. Philosophy of Science, 70, 235-254.
Tversky, B. (2011). Visualizing thought. Topics in Cognitive Science, 3, 499-535.
Vitaterna, M. H., King, D. P., Chang, A.-M., Kornhauser, J. M., Lowrey, P. L., McDonald, J. D., et al. (1994). Mutagenesis and mapping of a mouse gene, Clock, essential for circadian behavior. Science, 264, 719-725.

\title{
Using Perlin Noise to Generate Emotional Expressions in a Robot
}

\author{
Aryel Beck (aryelbeck@gmail.com) \\ Antoine Hiolle (a.hiolle@herts.ac.uk) \\ Lola Cañamero (l.canamero@herts.ac.uk) \\ Embodied Emotion, Cognition and (Inter-)Action Lab, School of Computer Science \& STRI \\ University of Hertfordshire, College Lane, Hatfield, Herts AL10 9AB, UK
}

\begin{abstract}
The development of social robots that convey emotion with their bodies-instead of or in conjunction with their facesis an increasingly active research topic in the field of humanrobot interaction (HRI). Rather than focusing either on postural or on dynamics aspects of bodily expression in isolation, we present a model and an empirical study where we combine both elements and produce expressive behaviors by adding dynamic elements (in the form of Perlin noise) to a subset of static postures prototypical of basic emotions, with the aim of creating expressions easily understandable by children and at the same time lively and flexible enough to be believable and engaging. Results show that the noise increases the recognition rate of the emotions portrayed by the robot.
\end{abstract}

Keywords: Bodily emotional expression; human-robot interaction; affective robotics; Perlin noise.

\section*{Introduction}

Echoing the importance of emotional expression in social interaction and communication among humans, the development of expressive robots that can interact with us in a human-oriented way is nowadays a very active research topic in the field of human-robot interaction (HRI). Interest in using robot's bodies for emotional expression is rapidly increasing. This is partly due to two main factors. On the one hand, an increasing corpus of research in psychology and neuroscience (e.g., (Wallbott, 1998; De Gelder, 2006; Avizer, Trope, \& Todorov, 2012)) is emphasizing the role of the body in conveying emotion-specific information rather than merely nonspecific information related to intensity as it was previously thought. On the other hand, the fact that a number of robotic platforms currently available have complex bodies with a high number of degrees of freedom and/or good motion capabilities, but do not necessarily have articulated faces-that is the case in \(\mathrm{Nao}^{1}\), the robot that we have used in this study.

While researchers typically focus either on the use of expressive postural elements or on expressive aspects of movement (Coulson, 2008)—see (Kleinsmith \& BianchiBerthouze, 2012) for a survey-the combination of both aspects has not received as much attention in robotics. In the study resented here, we combine both elements and produce expressive behaviors by adding dynamic elements to a subset of static postures prototypical of basic emotions. Our underlying motivation from the point of view of \(\mathrm{HRI}^{2}\), as part of the European project ALIZ-E (www.aliz-e.org), was to create a set of expressions easily understandable by children and at

\footnotetext{
\({ }^{1}\) www.aldebaran-robotics.com.
\({ }^{2}\) See (Cañamero, 2002, 2008) for discussions of design issues regarding expressive robots for HRI.
}
the same time lively and flexible enough to be believable and engaging.

\section*{Affect Space}

This study is part of our research investigating the elaboration of an Affect Space for the generation of emotional body language to be displayed by robots. It builds on an Affect Space that was generated using key poses (Beck, Cañamero, \& Bard, 2010; Beck, Hiolle, Mazel, \& Cañamero, 2010). In the context of this paper, a key pose is a posture modeled after an actor performance so that it clearly describes the emotion displayed.

\section*{Static features}

In animation, one of the standard methods for creating convincing and believable displays relies on expressive key poses rather than body language in motion (Thomas \& Johnston, 1995; Vala, Paiva, \& Rui Gomes, 2008). Taking inspiration from this method, in previous work (Beck, Cañamero, \& Bard, 2010; Beck, Hiolle, et al., 2010) we used static key poses as a basis to produce expressive animated behaviors in a humanoid robot. This method presents the advantage of permitting to investigate and model independently postural and motion-related expressive elements. This approach is also consistent with research on affective body expression suggesting that form and movement information are processed by separate pathways in the brain (Kleinsmith \& BianchiBerthouze, 2012). The key poses that we used are consistent with the static features \({ }^{3}\) in (Kleinsmith, Bianchi-Berthouze, \& Steed, 2011).
Our initial experiments (Beck, Cañamero, \& Bard, 2010) showed that it is possible to successfully convey emotions using static key poses displayed by a Nao humanoid robot. Based on these results, we started to develop a continuous Affect Space for our robot by "blending" key poses to generate new expressions (Beck, Hiolle, et al., 2010). The resulting system maps static key poses into a continuous dimensional model of emotion. Empirical results regarding the interpretation of the static key poses generated by this Affect Space can be found in (Beck, Hiolle, et al., 2010). While some of the expressions were clearly recognized, our results also show that some of the generated key poses are ambiguous and do not convey a clear emotion. In addition, feedback from people interacting with the robot indicated that they found it too static, which might have a negative impact on the perception on the

\footnotetext{
\({ }^{3}\) In particular, the collar joint angle was also found to be salient to the expression of emotion through body posture.
}
robot and hence on the interaction. This led us to hypothesize that the addition of dynamic aspects to the key poses could greatly improve the understanding and believability of the expressions

\section*{Animating Emotional Key Poses Using Perlin Noise}

To endow the key poses with a dynamic dimension, we added Perlin noise \({ }^{4}\) (Perlin, 1990) to them. In animation, Perlin noise-a coherent noise that is highly controllable-is a wellknown tool used to procedurally generate movements and increase the lifelikeness of animations. It presents the advantages of being simple and computationally cheap, which are important factors for implementation on a robotic platform. Moreover, the parameters used to generate it can be modulated, resulting in different types of animations. Perlin noise can be used to modify movement but also to create different types of non-repetitive and "idle" behaviors, as well as to generate textures. In robotics, Perlin noise and similar methods have also been used, applied to joint angles, to increase the lifelikeness of robot movements and to generate idle behaviors (Snibbe, Scheeff, \& Rahardja, 1999; Ishiguro, 2005).

Going beyond standard practice, in the work reported in this paper we have used Perlin noise to generate all the movements of the robot, rather to simply modify existing trajectories. The addition of Perlin noise values to the current joint angles produces a Perlin noise-based animation for the current pose of the robot. Although this step has not been validated with formal perceptual studies, the movements generated have been successfully used as idle behavior in empirical interaction studies with children carried as part of the ALIZ-E project (Nalin et al., 2012).

\section*{Using Perlin Noise to Express Emotions}

Following a "deep" approach to emotion modeling (Cañamero, 2008), affective expression in our robot is driven by the dynamics of the internal "affective state" of the robot in its interaction with the world. Consequently, movements produced by Perlin noise can be modulated by the internal state of the robot and used as a tool to express emotions. This novel use of Perlin noise can potentially be a powerful tool to create more subtle expressions in robots, since it permits to procedurally create non-repetitive body movements that convey different emotions or nuances of the same emotion. Another advantage of our approach is that such expression would not be limited to a single platform and could be reused across different robots-both humanoid and non-humanoid.

One of the main challenges posed by the use of Perlin noise to express emotions is to find a mapping between the parameters used to generate the noise and the emotion to be conveyed. In our model, we used the following mappings:
- Velocity was mapped to the time taken by the robot to move, i.e., the shorter the time the higher the velocity.

\footnotetext{
\({ }^{4}\) See http://freespace.virgin.net/hugo.elias/models/ m_perlin.htm for a description of the method used.
}

This mapping was chosen, rather than directly using the speed of the motors, due to constraints imposed by our robot. However, it should be noted that the actual velocity of the movement also depends on the amplitude of the noise, since the time is kept constant but the amplitude varies. Based on the existing literature, we expected that this parameter would have a significant effect on the perception of the emotion as it is related to Quantity of Motion (Camurri, Mazzarino, \& Volpe, 2003), Speed (Roether, Omlor, Christensen, \& Giese, 2009; Bernhardt, 2010) and Activation (Wallbott, 1998; Hartmann, Mancini, Buisine, \& Pelachaud, 2005).
- Jerkiness was introduced by applying random variations to the duration parameter, slightly modifying the interval of update of the joint angle. The literature suggests that jerkiness has a strong effect on the expression of emotion (Hartmann et al., 2005; Lee, Park, \& Nam, 2007; Bernhardt, 2010)

\section*{The Experiment}

To assess the potential of using Perlin noise to express emotions in robots, we designed a study to investigate the relation between characteristics of the movements generated using Perlin noise and the perceived emotion.

Independent Variables: Three independent variables were manipulated: Emotional Key Pose, Velocity and Jerkiness.
- Key Pose had five different values that corresponded to the different emotions tested.
- Velocity had three levels and described how fast the robot moved.
- Jerkiness had two levels. In the Jerky condition, the velocity of each movement (generated using Perlin noise) was multiplied by a random value between 0.5 and 1.5 ensuring that the mean of the velocity remained the same but introducing variation of speed during the animation. In the Regular (non-Jerky) condition, the speed (given by the Velocity condition) remained constant throughout the whole animation.

This resulted in 35(5 KeyPoses \(* 3\) Velocity \(* 2\) Jerk +5 static \()\) animations tested.

Dependent Variables: Perception of emotion was defined in terms of Emotional Label, Valence and Arousal.

\section*{Participants}

20 Participants were recruited, mostly members of staff of the University of Hertfordshire ( 9 females and 11 males) ranging in age from 18 to \(55(\mathrm{M}=29.31, \mathrm{SD}=11.93)\).

\section*{Apparatus}

Five key poses were selected from previous studies (Figure 1): two positive, two negative and one neutral that had been

Figure 1: The five key poses (from left to right: sadness, anger, neutral, pride, happiness)

recognized well above chance level in previous studies (Beck, Cañamero, \& Bard, 2010; Beck, Hiolle, et al., 2010). To ensure stability, the robot was sitting and only the joint angles of the upper body were modified while changing key pose. The animations were generated by adding Perlin noise to the joints of the upper body (as described above).

\section*{Procedure}

The same experimenter tested all participants individually. Once each participant had given consent at the beginning of their session, they were given standardised explanation regarding the questionnaire that they were expected to answer and were instructed to imagine that the robot was reacting to something. In this context, Valence was defined as the extent to which this "something" was positive or negative, and Arousal was defined as the level of energy (low to high energy).

After confirming that they understood all the questions, participants watched and assessed the 35 animations. Each animation was displayed only once in a randomized order different for each participant. A distance was introduced to avoid having the same pose coming twice in a row. Each time, the robot took a pose and displayed an animation during 15 seconds and returned to a non-expressive key pose (a second neutral pose) until the participant answered. For each animation, participants were asked to describe the animation using their own terms and eventually choose an emotion label from a list of six emotions. The list was comprised of Anger, Sadness, Fear, Neutral, Pride, Happiness and Excitement. Participants completed ratings of Valence and Arousal on a 10 -point Lickert scale. After all the poses had been assessed, participants were fully debriefed. Each session lasted approximately 30 minutes.

\section*{Results}

Since this experiment uses a modified set of key poses (unlike in the test of the static key poses, here the robot is sitting), it was necessary to validate the material created for this study.

\section*{Validation of the Sitting Key Poses}

Recognition rates showed that it was possible for participants to correctly identify the different static key poses far above chance level (Chance level would be \(17 \%\) ). Thus, it was possible for participants to identify the static key poses displayed (Table 1).

Table 1: Recognition rate of the Key Poses with and without added movements
added movements
\begin{tabular}{|l|l|l|l|}
\hline Emodion & Recognition Rate Static & Recog. Rate with Movement & Best Condition \\
\hline Sadness & \(84 \%\) & Slow Regular \\
\hline Anger & \(42 \%\) & \(68 \%\) & Fast Regular \\
\hline Pride & \(63 \%\) & \(68 \%\) & Medium Regular \\
\hline Happiness & \(99 \%\) & \(74 \%\) & Fast Jerky \\
\hline Neutral & \(84 \%\) & \(95 \%\) & Medium Regular \\
\hline
\end{tabular}

Figure 2: Effect of Changing the Key Pose on Valence


As part of the validation of the material, a two-ways (static vs. highest recognition rate) Repeated Measures ANOVA was conducted on the total Number of Correct Interpretations comparing the static display and the highest recognition rate with movement for each emotion. This was done to check that it was possible to increase the recognition rate by adding movements generated with Perlin noise in at least one condition for the different key poses. The results show that this was the case \(\left(F(1,18)=9.08, p<0.01, \eta^{2}=0.33\right)\). Table IV also highlights the recognition rates as well as the conditions in which the highest recognition rates were obtained.

In the following sections, the data was analysed using 5(Key Pose)*3(Velocity)*2(Jerkiness) Repeated Measures Anovas on the dependent variables. It should be noted that since they do not have a Jerkiness condition, the static poses were not included in these tests.

\section*{Effect of Changing the Key Pose Displayed}

Effect on the Number of Correct Interpretations As expected, Key Pose had a significant effect on the Number of Correct Interpretations \((F(4,72)=6.89, p<\) 0.01 , partial \(\eta^{2}=0.99\) ). This indicates that overall, when displayed with movements, the key poses were not all equally well recognized. Post-Hoc tests (Least Significant Difference) showed that the poses for Sadness and Pride were recognized better than the others \((p<0.01)\).

Effect on Valence Key Pose had a significant effect on Valence \(\left(F(4,72)=33.26, p<0.01\right.\), partial \(\left.\eta^{2}=0.65\right)\). Post-hoc tests (Least significant Difference) showed that the pose for Sadness was perceived as more negative than the rest of the poses ( \(p<0.01\) for all of them). The key pose for Anger was perceived as more negative than Happiness

Figure 3: Effect of Changing the Key Pose on Arousal

( \(p<0.01\) ) and Pride \((p<0.01)\). There was however no significant difference between Anger and Neutral \((p=0.29)\). Pride was perceived as significantly more positive than the rest of the key pose ( \(p<0.05\) for all of them). Happiness was perceived as significantly more positive than Sadness ( \(p<0.01\) ), Anger ( \(p<0.01\) ) and Neutral ( \(p<0.05\) ) (Figure 2)

These results indicate that participants' perception of Valence was affected by the Key Pose being displayed. Overall, negative key poses were interpreted as such and positive key poses were interpreted as positive (Figure 2).

Effect on Arousal Key Pose had a significant effect on Arousal \(\left(F(4,72)=13.29, p<0.01\right.\), partial \(\left.\eta^{2}=0.42\right)\). Post-Hoc tests(Least Significant Difference) showed that Sadness was perceived as less aroused than Anger ( \(p<0.01\) ), Pride \((p<0.01)\), and Happiness \((p<0.05)\). There was no significant difference between Sadness and Neutral ( \(p=\) 0.21 ). Anger was perceived as more aroused than Neutral ( \(p<0.01\) ). However, there was no significant difference with Happiness \((p=0.26)\) and Pride \((p=0.37)\). Pride was perceived as less aroused than neutral \((p<0.01)\). There was a trend toward Pride being perceived as less aroused than Happiness \((p=0.06)\).
These results indicate that perception of Arousal was affected by the key pose being displayed (Figure 3).

\section*{Effect of Velocity}

Effect on Interpretation Velocity had a significant effect on the number of correct interpretation \((F(2,36)=\) 11.02, \(p<0.01\), Partial \(\left.\eta^{2}=0.98\right)\). This effect was further investigated while looking at the interactions between the dependent variables.

Effect on Valence Although it did not reach significance, there was a trend of Velocity affecting Valence \((F(2,36)=\) \(3.14, p=0.06\), partial \(\eta^{2}=0.15\) ). Post-Hoc tests (Least Significant Difference) showed that there was a trend of Slow movement perceived as less positive than Fast \((p=0.07)\). There was no difference between the Slow and Medium con-

Table 2: Effect of Velocity and Jerkiness on Interpretation per Key Pose
\begin{tabular}{|c|c|c|}
\hline Key Pose & Effect of Velocity & Effect of Jerkiness \\
\hline Sadness & \[
\begin{aligned}
& F(2,34)=5.34, p<0.05, \eta^{2}=0.24 \\
& \text { Slow }>\text { Medium }(p<0.05) \\
& \text { Slow }>\text { Fast }(p<0.01) \\
& \text { Medium }=\text { Fast }(p=0.31) \\
& \hline
\end{aligned}
\] & \[
\begin{aligned}
& F(1,17)=11.73, p<0.01, \eta^{2}=0.41 \\
& \text { Regular }>\operatorname{Jerki}(p<0.01)
\end{aligned}
\] \\
\hline Anger & \[
\begin{aligned}
& F(2,34)=6.21, p<0.01, \eta^{2}=12.43 \\
& \text { Fast }>\text { Medium }(p<0.05) \\
& \text { Fast }>\operatorname{Slow}(p<0.01) \\
& \text { Medium }=\operatorname{Slow}(p=0.45)
\end{aligned}
\] & \(F(1,18)=0.79, p=0.39, \eta^{2}=0.04\) \\
\hline Neutral & \[
\begin{aligned}
& \left.F(2,36)=48.69, p<0.01, \eta^{2}=0.73\right) \\
& \text { Slow }>\operatorname{Fast}(p<0.01) \\
& \text { Medium }>\operatorname{Fast}(p<0.01) \\
& \text { Slow }=\operatorname{Medium}(p=0.1) \\
& \hline
\end{aligned}
\] & \(F(1,18)=0.00, p=1, \eta^{2}=0.00\) \\
\hline Pride & \[
\begin{aligned}
& F(2,36)=17.95, p<0.01 \eta^{2}=0.50 \\
& \text { Slow }>\text { Fast }(p<0.01) \\
& \text { Medium }>\text { Fast }(p<0.01) \\
& \text { Slow }=\text { Fast }(p=0.19)
\end{aligned}
\] & \(F(1,18)=1.09, p=0.31, \eta^{2}=0.06\) \\
\hline Happiness & \[
\begin{aligned}
& \left.F(2,36)=5.36, p<0.01, \eta^{2}=0.23\right) \\
& \text { Fast }>\operatorname{Slow}(p<0.01) \\
& \text { Fast }=\operatorname{Medium}(p=0.09) \\
& \text { Medium }=\operatorname{Slow}(p=0.17) \\
& \hline
\end{aligned}
\] & \(F(1,18)=1.20, p=0.29, \eta^{2}=0.06\) \\
\hline
\end{tabular}
ditions ( \(p=0.34\) ). The Medium condition was perceived as significantly less positive than the Fast condition ( \(p<0.05\) ).

These results indicate that the fast movement condition was perceived as more positive than the other two.

Effect on Arousal Velocity had a significant effect on Arousal \(\left(F(2,36)=93.60, p<0.01\right.\), partial \(\left.\eta^{2}=0.84\right)\). Post-Hoc tests (Least Significant Difference) showed that the Slow condition was perceived as less aroused than the Medium condition \((p<0.01)\) which in turn was perceived as less aroused than the Fast condition \((p<0.01)\).

These results indicate that overall the faster the movement is, the more aroused the expression is perceived.

\section*{Effect of Jerkiness}

Effect on Interpretation There was a trend of Jerky being more correctly interpreted than the same display in the Regular condition \(\left(F(1,18)=4.21, p=0.55\right.\), partial \(\left.\eta^{2}=0.49\right)\). This was further explored while considering the interactions between the dependent variables.

Effect on Valence Jerkiness had no significant effect on Valence \(\left(F(1,18)=0.26, p=0.62\right.\), partial \(\left.\eta^{2}=0.01\right)\). These results indicate that overall, participants' perception of Valence was not affected by the Jerkiness of the movements.
Effect on Arousal Jerkiness had a significant effect on Arousal \(\left(F(1,18)=27.51, p<0.01\right.\), partial \(\left.\eta^{2}=0.60\right)\).

Post-Hoc tests showed that the "Jerky" condition was perceived as more aroused than the Regular one ( \(p<0.01\) ).

\section*{Interaction between the independent variables}

Interpretation There was an interaction between Key Pose and Velocity of movements over the Number of Correct Interpretation \(\left(F(8,144)=13.15, p<0.01\right.\), partial \(\left.\eta^{2}=1\right)\). Similarly, there was an interaction between Key Pose and Jerkiness \(\left(F(4,72)=2.54, p<0.05\right.\), partial \(\left.\eta^{2}=0.69\right)\). This indicates that the interpretation of emotion depended both on the Key Pose being displayed, on the Velocity of movement and on the Jerkiness. This was further investigated using repeated measures ANOVAs on the different Key Pose and Ve-

Table 3: Effect of Velocity on Valence per Key Pose Displayed
\begin{tabular}{|l|l|}
\hline Key Pose & Repeated Anovas \\
\hline Sadness & \(F(2,36)=0.43, p=0.65\), partial \(^{2} \eta^{2}=0.02\) \\
\hline Anger & \(F(2,36)=1.46, p=0.25\), partial \(\eta^{2}=0.08\) \\
\hline Neutral & \(F(2,36)=0.86, p=0.43\), partial \(\eta^{2}=0.05\) \\
\hline Pride & \(F(2,36)=1.57, p=0.22\), partial \(^{2}=0.08\) \\
\hline Happiness & \begin{tabular}{l}
\(F(2,36)=10.24, p<0.01\), partial \(^{2} \eta^{2}=0.36\) \\
Fast \(>\operatorname{Slow}(p<0.01)\) \\
Fast \(>\) Medium \((p<0.01)\) \\
Medium \(=\operatorname{Slow}(p=0.33)\)
\end{tabular} \\
\hline
\end{tabular}
locity conditions (Table 2). Table 2 shows that the highest recognition rate for Sadness was with Slow and Regular movements, for Anger, it was with Fast movements (no effect of jerkiness), neutral was better interpreted with Slow and Medium speed. Pride was better interpreted at Slow and medium speed. For Happiness, it was with Fast and Medium speed.
Valence There was a significant interaction between Velocity and Key Pose on Valence \((F(8,144)=5.85, p<\) 0.05, partial \(\left.\eta^{2}=0.11\right)\). This indicates that the effect of Velocity depends on the Key Pose. This was therefore investigated in details using 3(Velocity) Repeated Measure Anovas on the different Key Pose individually (Table 3).
Arousal There was a significant interaction between Key Pose and Velocity on Arousal \((F(8,144)=5.81, p<\) \(0.01, \operatorname{partial} \eta^{2}=0.24\) ). Repeated Measures Anovas were therefore conducted on the different Key Pose conditions separately. The results of these showed that the pattern were constant for all of them and that the Fast condition was perceived as more aroused than the Medium condition ( \(p<0.01\) for all the Key Poses) which in turn was perceived as more Aroused than the Slow condition ( \(p<0.01\) for all the Key Pose).

\section*{Discussion}

Valence and Arousal As expected, Key Pose had a strong effect on Valence and Arousal. More precisely, the perceived Valence and Arousal were consistent with the respective positions of each Key Pose within the Affect Space (Figures 2 and 3). Moreover, Velocity had a marginal effect on Valence. However, the interactions between Velocity and Key Pose suggest that the difference in Valence was due to the key pose for happiness (Table 3) as it was found that for all the other key poses, Velocity had no effect on Valence. Similarly, Jerkiness did not affect the perceived Valence of the display. This is consistent with existing results in psychology which suggest that Arousal is a formless cue that relates directly to the movement kinematics while Valence seems to be related to the relations between the different limb segments (Pollick, Paterson, Bruderlin, \& Sanford, 2001).
However, both Velocity and Jerkiness were found to increase the perception of Arousal. Taken together, the results suggest that the perceived Valence depended on the Key Pose displayed without taking into account the different dynamic conditions. In contrast, the perceived Arousal depended on all three dependent variables. Hence, participants relied only on the body posture to assess Valence. However, all the inform-
ation available (Key Pose, Velocity and Jerkiness) was used to rate Arousal.

Interpretation Participants were able to correctly identify the different static key poses. Whilst the recognition rate for Anger was lower than for the other key poses, it was still above chance level. This low recognition rate could be due to the modification done to the material as the robot was sitting down. The key pose was better recognised in previous experiment with the robot standing up (Beck, Stevens, Bard, \& Cañamero, 2012) and the lack of significant difference between the key poses for anger and neutral on Valence that was found in this study could be due to the key pose for anger being misinterpreted in most of the conditions. This will have to be investigated in future work.

Moreover, when compared with static poses, the recognition rates for the display with movements clearly show that adding appropriate dynamic elements improves significantly the expressivity of the key pose (Table 1). Although it was not possible to capture this statistically, Velocity seems to have a consistent effect on interpretation. For instance, the key pose for sadness was interpreted as sad in slow motion (resulting in the very high recognition rate in this condition); however, as the Velocity increased, it shifted toward anger and frustration. This is consistent with the results found with regards to the effect of Velocity on Valence (Table 3) which show that, with the exception of happiness, Velocity had not effect on Valence. Thus, these shifts in interpretation can be explained by the effect of Velocity on Arousal. In other words, a negative expression, remains negative, but its level of Arousal increases along with Velocity shifting from sadness to anger and frustration. The interpretations of the key poses were affected by the Velocity and the Jerkiness of the movements. More precisely, the dependence between Key Pose and Velocity with regards to the interpretation shows the importance of matching the Velocity and the Jerkiness of movements to the Key Pose in order to express specific areas of the Affect Space. The drop in recognition for Sadness in the Jerky condition suggests the importance of regular movement for this expression.

Even though pride was correctly labeled, the rating of Arousal was higher than what could have been expected. This was also the case in (Beck, Cañamero, \& Bard, 2010) and could be due to this specific posture. It could also be related to the physical aspect of the Nao robot, as the arm joints are very salient in this key pose.

Limitations and Future Work It is important to highlight that the key poses used for this study are prototypical and were intentionally selected to be expressive. This is appropriate and beneficial for the development of an expressive system. However, it is likely that the use of prototypical expressions had an effect on the results found in this study. Moreover, the Jerkiness condition could have been implemented by manipulating the number of Harmonics and the Frequency of the noise. This could result in different visual
results with different effects on the perception of emotion. It should also be noted that Perlin noise does not capture the relationship that exists between the rotation of one joint and another. This may result in unrealistic animations (Egges \& Magnenat-Thalmann, 2005). Although this did not seem to be the case in this study as the material was carefully checked, it could still have affected the results.

This study did not consider the effect of context on the perception of the body language displayed. However, it can be argued that interpretation of emotion is context dependent and that changing the context could change the perception of the expressions generated by this Affect Space. On the other hand, work on facial expressions of emotion has shown that at least for a few basic emotions, context is not necessary to identify the expressed emotion. In other words, the expression of an emotion is to a certain extent independent from the context, as evidenced by the widespread use of FACS. Similarly, the high recognition rates obtained in this study suggest that these expressions could convey the intended emotion in different contexts. However, people's reaction to the emotional expression are likely to differ. This will be investigated as part of the ALIZ-E project.

\section*{Acknowledgments}

This work was funded by the ALIZ-E FP7 European Project (Grant 248116). The opinions expressed are solely the authors' and not necessarily those of the consortium.

\section*{References}

Avizer, H., Trope, Y., \& Todorov, A. (2012). Body cues, not facial expressions, discriminate between intense positive and negative emotions. Science, 338, 1225-1229.
Beck, A., Cañamero, L., \& Bard, K. (2010, sept.). Towards an affect space for robots to display emotional body language. In Ro-man, 2010 ieee (p. 464-469).
Beck, A., Hiolle, A., Mazel, A., \& Cañamero, L. (2010). Interpretation of emotional body language displayed by robots. In Proc. of the 3rd int. workshop on affective interaction in natural environments (pp. 37-42). Firenze, Italy.
Beck, A., Stevens, B., Bard, K. A., \& Cañamero, L. (2012, March). Emotional body language displayed by artificial agents. ACM Trans.. Interact. Intell. Syst., 2(1), 2:1-2:29.
Bernhardt, D. (2010). Emotion inference from human body. Unpublished doctoral dissertation, University of Cambridge, Computer Laboratory.
Camurri, A., Mazzarino, B., \& Volpe, G. (2003). Analysis of expressive gesture: The eyesweb expressive gesture processing library. Gesture-Based Communication in HumanComputer Interaction, 460-467. (LNAI)
Cañamero, L. (2002). Playing the emotion game with feelix. In K. Dautenhahn, A. Bond, L. Cañamero, \& B. Edmonds (Eds.), Socially intelligent agents (Vol. 3, p. 69-76). Springer US.
Cañamero, L. (2008). Animating affective robots for social interaction. In L. Cañamero \& R. Aylett (Eds.), Animating expressive characters for social interaction (p. 103-121).

Coulson, M. (2008). Expressing emotion through body movement: A component process approach. In L. Cañamero \& R. Aylett (Eds.), Animating expressive characters for social interaction (p. 7186). John Benjamins Publishing Company.
De Gelder, B. (2006). Toward the neurobiology of emotional body language. Nature Reviews, Neuroscience, 7, 242-249.
Egges, A., \& Magnenat-Thalmann, N. (2005). Emotional communicative body animation for multiple characters. In Proc. of the first int. workshop on crowd simulation (pp. 31-40). Lausanne, Switzerland.
Hartmann, B., Mancini, M., Buisine, S., \& Pelachaud, C. (2005). Design and evaluation of expressive gesture synthesis for embodied conversational agents. In Proc. 4th int. joint conf. on autonomous agents and multiagent systems (pp. 1095-1096). The Netherlands.
Ishiguro, H. (2005). Android science: Toward a new crossdisciplinary framework. In Cogsci-2005 worskshop: Toward social mechanisms of android science (p. 1-6).
Kleinsmith, A., \& Bianchi-Berthouze, N. (2012). Affective body expression perception and recognition: A survey. IEEE Trans. on Affective Computing, 99(PrePrints).
Kleinsmith, A., Bianchi-Berthouze, N., \& Steed, A. (2011, August). Automatic recognition of non-acted affective postures. Trans.. Sys. Man Cyber. Part B, 41(4), 1027-1038.
Lee, J.-H., Park, J.-Y., \& Nam, T.-J. (2007). Emotional interaction through physical movement. In Proc. of the 12th int. conf. on human-computer interaction (pp. 401-410). Berlin, Heidelberg: Springer-Verlag.
Nalin, M., Baroni, I., Kruijff-Korbayov, I., Cañamero, L., Lewis, M., Beck, A., et al. (2012). Childrens adaptation in long-term interactions with a humanoid robot. In Proc. of ro-man12.
Perlin, K. (1990). Noise and turbulence. Available from http://www.mrl.nyu.edu/~perlin/doc/oscar.html
Pollick, F., Paterson, H., Bruderlin, A., \& Sanford, A. (2001). Perceiving affect from arm movement. Cognition, 82(2), B51-B61.
Roether, C. L., Omlor, L., Christensen, A., \& Giese, M. A. (2009). Critical features for the perception of emotion from gait. Vision, 9(6).
Snibbe, S., Scheeff, M., \& Rahardja, K. (1999). A layered architecture for lifelike robotic motion. In Proc. of the 9th int. conference on advanced robotics.
Thomas, F., \& Johnston, O. (1995). The illusion of life. NewYork: Abbeville Press.
Vala, M., Paiva, A., \& Rui Gomes, M. (2008). Effective bodies for affective interaction. In L. Cañamero \& R. Aylett (Eds.), Animating expressive characters for social interaction (p. 87-101).
Wallbott, H. G. (1998). Bodily expression of emotion. European J. of Social Psychology, 28(6), 879-896.

\title{
Response direction and sentence-tense compatibility effects: An eye tracking study
}

\author{
Raymond B. Becker (rbecker@cit-ec.uni-bielefeld.de) \\ Bridgette DeCot (bdecot@cit-ec.uni-bielefeld.de) \\ Ernesto Guerra (ernesto.guerra@uni-bielefeld.de) \\ Pia Knoeferle (knoeferl@cit-ec.uni-bielefeld.de) \\ Cognitive Interaction Technology Excellence Cluster and Department of Linguistics, Bielefeld University, Morgenbreede 39, 33615, Bielefeld, Germany \\ Rolf Zwaan (zwaan@fsw.eur.nl) \\ Department of Psychology, Erasmus University \\ Burgemeester Oudlaan 50, 3062 PA, Rotterdam, The Netherlands
}

\begin{abstract}
Recent evidence shows tense-response compatibility effects only when the task relates to sentence tense (Ulrich \& Maienborn, 2010). In two eye-tracking experiments, we investigated tense-response compatibility effects. In our first experiment (E1, where sentence tense was relevant to the task) we found compatibility effects at the beginning of the sentence (e.g., Yesterday versus Tomorrow), which shifted to interference effects by sentence end. Overall, we also found compatibility effects in response times, replicating Ulrich and Maienborn. Both compatibility effects in Experiment 1 (E1) were stronger for low- compared to high-WM readers. In Experiment 2 (E2, where tense was irrelevant), we found compatibility effects for high-WM readers, but only in early reading measures. These results suggest that compatibility effects are weaker depending on the task, but not eliminated; an implication which may help refine a strict view of embodied cognition.
\end{abstract}

Keywords: Mental timeline, embodiment, individual differences, eye-tracking.

\section*{Introduction}

Research over the last decade has continued to refine embodiment theory (Barsalou, 1999; Glenberg, 1997), and this refinement was prodded along by criticism (Machery, 2007; Mahon \& Caramazza, 2008). For example, Mahon \& Caramazza argued that embodiment theory could not adequately explain how JUSTICE and other abstract concepts are understood through bodily experience because they do not reliably correspond to sensory or motor information. However, conceptual metaphor theory has laid out the groundwork for how abstract concepts such as TIME are mapped onto concrete concepts such as SPACE (Lakoff \& Johnson, 1980; 1999). Torralbo, Santiago, \& Lupiáñez (2006) found evidence that corroborated this potential mapping mechanism. In their Experiment 1, participants saw the silhouette of a human head looking either rightward or leftward on a screen. A word with a temporal connotation in a speech bubble was presented either in front of or behind the silhouette. Participants judged whether the person represented via the silhouette was contemplating the past or the future. When a past word appeared on the left, responses were faster than when it appeared to the right; when a future word appeared on the right side, responses were faster than
when it appeared to the left (this interaction of responselocation with tense has been credited to a 'mental timeline', i.e., the use of a spatial left-right line to represent time in our mind). These results suggest that left- and right-hand response preparation interacts with linguistic temporal cues (past and future tense respectively). Thus it appears that abstract concepts such as TIME are grounded in experiential and bodily schemas. Meanwhile, the focus of inquiry in this area has changed from whether grounding effects occur for abstract concepts to how rapidly they occur and whether they are task-dependent. In addition, the role of participants’ working memory in these kinds of congruence effects is unclear. To contribute to these research questions we examined the time course of time- response location congruence effects during sentence comprehension as (low and high working memory) participants planned a right or left hand movement in two different tasks. Below we motivate in more detail the investigation of tense-response location congruence effects are modulated by task and working memory.

\section*{Accommodating tense-response-location congruence effects}

Task appears to play an important role for tense-response location congruence effects. In a recent study, compatibility effects of tense (e.g., past versus future) and left/right response locations were eliminated in a task where tense was irrelevant. When participants paid attention to sentence tense, tense-response location compatibility effects emerged. For example, participants pressed a button labeled Past on the left in response to a past tense sentence more quickly than when the Past button was on the right. A similar compatibility effect was found for future-tense sentences and right-hand responses. However, when the task was time-irrelevant (sentence-sensibility judgments), compatibility effects were eliminated, suggesting that timeresponse location compatibility effects occur only when people pay attention to time. If that were the case, then both embodied (e.g., Barsalou, 1999) and non-embodied accounts such as amodal symbol systems (Collins \& Loftus, 1975, Collins \& Quillian, 1969) could accommodate these results. Non-embodied accounts could accommodate the
results via a traditional spreading activation network composed of disembodied, or amodal, symbols. A similar reasoning has been proposed for emotion and embodiment, but it was ruled out as it was shown that compatibility effects between emotional sentences and facial expression were task-independent (Glenberg, Havas, Becker, \& Rinck, 2005). Nevertheless, the lack of tense-response compatibility effects in a time-irrelevant task (Ulrich and Maienborn, 2010) left the door open for accounts via hybrid embodiment theories (Mahon \& Caramazza, 2008; Louwerse \& Jeuniaux, 2008) or via cross-modal integration (Kemmerer \& Gonzalez-Castillo, 2010).

There are at least two key differences between hybrid embodiment accounts and views of embodiment in which mental representations are strictly composed of perceptual symbol systems (henceforth 'strict embodiment'). First, hybrid accounts argue that the hierarchical processing of amodal symbols occur before additional top-down context from perceptual symbols (Kemmerer \& Gonzalez-Castillo, 2010) and this takes more time than a direct mapping of TIME onto SPACE. Second, strict embodiment proposes automaticity (Glenberg, 1997, p4). Automaticity refers to whether the sensorimotor system is involved in processing the meaning of abstract concepts (temporal cues in a sentence) regardless of task. Strict embodiment would thus have predicted task-independent activation of the mental timeline. Ulrich \& Maienborn's results of task-dependent tense-response location compatibility effects appeared to support a hybrid view of embodiment; however, they did not explicitly address the implications of their findings for this debate. Moreover, we cannot be certain that the lack of tense-response location compatibility effects with timeirrelevant tasks is at least in parts due to the coarse-grained response time measure they used. By monitoring eye movements during reading in addition to response times at sentence end in the same tasks that they used, we can assess the time course of tense-response location compatibility effects and determine whether the null findings are due to the nature of the measure.

\section*{Working memory and embodiment}

Strict embodied cognition draws on attention and memory functions (Barsalou, 1999; Glenberg, 1997). For example, Glenberg and Gallese (2010; Koziol, Budding, \& Chidekel, 2011) argued that higher order processes such as executive function are for motor control, and also part of language comprehension (see Repovš \& Barch, 2012 for a possible link between working memory (WM) and cerebellum function). However, the notion of working memory as an important component of theories in cognitive psychology (see Baddeley, 2012 for a review) and psycholinguistics (Huettig, Olivers, \& Hartsuiker, 2011; Lewis, Vasishth, \&

Van Dyke, 2006), has been studied very little by strict embodiment theorists like Glenberg or Barsalou \({ }^{1}\).

Due to the scarcity of research on embodiment and working memory, we drew from work on temporal order and working memory (Münte, Schiltz, \& Kutas, 1998). This previous research suggested that participants with high-WM used temporal cues such as Before versus After immediately to aid sentence processing. For example, in the following sentences, the initial temporal adverb and verb tell the reader that this sentence describes an event that occurred in the past (1) or the future (2):
(1) Früher in dieser Woche ADV \(\mid\) faltete \(_{\mathbf{V P}} \mid\) Jennifer \(_{\text {NP1 }}\) im Wohnzimmer \({ }_{\mathbf{P P}} \mid\) die Wäsche \({ }_{\mathbf{N P 2}}\).
\({ }^{\prime}\) Earlier this week \({ }_{\mathbf{A D V}} \mid\) folded \(_{\mathbf{V P}} \mid\) Jennifer \(_{\mathbf{V P}}\) in the living room \(_{\mathbf{P P}} \mid\) the laundry \({ }_{\mathbf{N P 2}}\) '.
(literal translation).
(2) Später in dieser Woche \(_{\text {ADV }} \mid\) faltet \(_{\mathbf{V P}} \mid\) Jennifer \(_{\mathbf{N P} 1}\) im Wohnzimmer \(_{\mathbf{P P}} \mid\) die Wäsche \(_{\mathbf{N P} 2}\).
\({ }^{\prime}\) Later this week \({ }_{\mathbf{A D V}} \mid\) folds \(_{\mathbf{V P}} \mid\) Jennifer \(_{\mathbf{V P}}\) in the living \(\operatorname{room}_{\mathbf{P P}} \mid\) the laundry \({ }_{\mathbf{N P 2}}\) '.

It would be consistent with the findings of Münte, Schiltz, and Kutas (1998) if high- (but not low) WM immediately processed the temporal cue 'Earlier / Later'. A question that could be asked with respect to the role of working memory in embodied cognition is whether participants immediate integrate temporal cues in addition to response-location as they make a sensibility judgment about a sentence.

\section*{The present study}

Using eye-tracking, the present studies thus investigated the time course of tense-response location compatibility effects as a function of (a) task (time-focus vs. no time focus), and (b) participants' working memory. The use of eye tracking and a between-experiment task manipulation permitted us to test the strict embodied hypothesis.

In Experiment 1, participants performed a tense evaluation task (was the sentence in the past or in the future)? One gaze pattern in support of strict embodiment would be an early-peaking, quickly decaying Simon-like effect consistent with the action and perception literature (Symes, Tucker, \& Ellis, 2005). This pattern would suggest both a rapid (tied to first-pass measures) and automatic (insensitive to task) tense-response location congruence effects in first-pass times at the verb and potentially also the

\footnotetext{
\({ }^{1}\) We would like to thank an anonymous reviewer for suggesting the Coherent Working Models theory proposed by Santiago and colleagues. Unfortunately, due to time and page limit constraints, we have not integrated their proposal with our current framing or the discussion of our results. However, we will review the data supporting the Coherent Working Models theory, and how well our results fit with it in an extended manuscript in preparation.
}
next sentence region. If it is rapid and automatic but quickly decays, then there may be no effect in a relatively late and course-grained measure such as response times. By contrast, the amodal views would predict no tense-response compatibility effect for time-irrelevant tasks: symbols representing the concepts of FUTURE and PAST would not become bound to right and left procedural symbols and in turn not become activated. Hybrid accounts, would predict task-independent effects but these should occur later than in strict embodiment accounts.

We further hypothesized that high-WM readers would rapidly process the sentence-initial temporal cue (see Münte, Schiltz, \& Kutas, 1998). In order to come up with a hypothesis regarding working memory and tense-response location congruence effects, we drew from one of the more important ideas in embodied cognition which is that people can "offload" cognition to the environment (Clark, 1997; Spivey, 2008). To the extent that cognitive load can be offloaded onto the environment, a left-right mental timeline in which left indexes the past and right the future could assist in processing temporal information. And if the results from Münte et al. generalize to tense-response location congruence effects, then high-WM participants should process tense-response location congruence earlier. We predicted longer first-pass times in the subject noun phrase region (NP1 in sentences (1) and (2)) for incompatible (vs. compatible) tense-response location for high-WM participants because the region is potentially where participants would shift their attention from the tense processing to the sensibility judgment. This effect could extend to the locative prepositional phrase region as well because the attention shift could take time even for highWM readers (see sentences (1) and (2)). By contrast, for low-WM readers congruence effects should emerge at the end of the sentence, because in Münte et al. low-WM readers did show evidence that they were processing the temporal cues but later than the high-WM readers. In response times, both groups should show a compatibility effect replicating Ulrich and Maienborn (2010).

When time is not relevant for the task (Experiment 2), and if the null effect in Ulrich and Maienborn is an artifact of the post-sentence response time measure, then we should see similar yet more subtle effects than in E1 (compatibility effects in first-pass times, potentially also earlier for highWM than low-WM readers). However, these patterns of reading times should not result in compatibility effects in response times based on the findings by Ulrich and Maienborn.

\section*{Experiment 1: Time-relevant task}

In Experiment 1, we replicated the procedure of the first experiment by Ulrich and Maienborn (2010). Participants were asked to pay attention to sentence tense and registered their decision via a button press if the sentence made sense. Thinking about time was part of the task because participants made explicit decisions about sentence tense.

\section*{Method}

Participants 48 members (17 male with mean age 24; \(S D=\) 3 years) of Bielefeld University participated in the experiment. All participants were native German speakers with no second language exposure prior to 6 years of age; had normal or corrected-to-normal vision; were naïve with respect to the purpose of the study; and received \(€ 6\) for their participation or course credit. All gave informed consent.
Materials Items consisted of 48 past and future tense sentences beginning with a temporal adverb and 48 nonsense sentences of the same syntactic structure The nonsense sentences included the same words as those used in Ulrich and Maienborn (2010), but were restructured to be similar to the critical sentences, where the temporal adverb was always at the beginning of the sentence.

Procedure Participants were asked to judge whether the sentence referred to the past or future. However, they were also asked to respond only if the sentence made sense (see Figure 1). For nonsense sentences, participants were instructed to wait until the trial timed out. Their eyemovements were recorded at 1000 Hz using an Eyelink 1000 desktop mounted tracker. Participants' WM was tested by the automated reading span test (Unsworth, Redick, Heitz, Broadway, \& Engle, 2009).


Figure 1: Example trial from E1.

\section*{Analysis}

Prior to analysis fixations were cleaned using a 4-stage procedure. In the first stage fixations less than 80 ms were merged with the nearest neighboring fixation if it was longer than 80 ms and within 0.5 degrees of visual angle away along the \(x\)-axis. Similarly, in the second stage fixations less than 40 ms were merged with the nearest neighboring fixation if it was longer than 40 ms and within 1.25 degrees away along the \(x\)-axis. In stage 3 , every interest area was checked for at least three fixations less than 140 ms and none larger than 140 ms . If an interest area was found that met these criteria, these fixations were merged with the larger ones. Lastly, all fixations less than 80 ms and 1200 ms were
removed. Trials with incorrect answers and nonsense sentences were not analyzed. Participants were split into high- and low-WM groups using a tertile split forming three groups of 16 people each, but only the high- and low-WM groups were included in order to do an extreme groups analysis. We conducted a 2 (WM) x 2 (tense) x 2 (response location) linear mixed effects model analysis to test for tense-response compatibility effects in each sentence region; starting with the full model and removing parameters until we found the most parsimonious model that best fit the data (Baayen, Bates, \& Davidson, 2008) \({ }^{2}\).

\section*{Results}

There were no significant effects in first-pass reading times for any sentence region. In total dwell times in the sentenceinitial temporal adverb region, we found a significant tenseresponse compatibility effect, \(t(30)=-2.16, p<0.05\) (see Fig. 2). Further, low-WM readers showed the compatibility effect, whereas high-WM readers did not, as evidenced by a 3-way interaction, \(t(30)=2.22, p<0.05\) (see Fig. 3). Surprisingly, interference effects emerged at sentence end for both groups, \(t(30)=2.08, p<0.05\) (see Fig. 4). In response times, we replicated the congruence effect from Ulrich \& Maienborn, (2010), \(t(30)=-3.87, p<0.05\). However, the congruence effect was driven by the low-WM group as evidenced by a 3-way interaction, \(t(30)=2.71, p<\) 0.05 (see Fig. 5).


Figure 2: Tense-response location compatibility effects for both WM groups for total dwell times in the temporal adverb region. Error bars indicate the standard error (SE).

\footnotetext{
\({ }^{2}\) Thank you to an anonymous reviewer for pointing out the fact that eye-tracking researchers enjoy many degrees of freedom in their research (e.g., regions of interest, first-pass readings versus total times, etc...). And further that our effects are quite small and potentially would not stand up to Bonferroni correction. Because we are in a crisis in psychology of false positives and failures to replicate, we used linear mixed effect models, backward model selection, and report the pMCMC values (Barr, Levy, Scheepers, \& Tily, 2013).
}


Figure 3: Tense-response location compatibility effects for low- (left) and high-WM (right) groups for total dwell times in the temporal adverb region. Error bars indicate the \(S E\).

\section*{Discussion}

The response time results from Experiment 1 replicate prior tense-response location congruence effects in response times (Experiment 1 in Ulrich and Maienborn, 2010). However, the pattern of the compatibility effects over the course of the sentence, and as a function of working memory showed that there is more to the story. The response time compatibility effects in the low-WM group, but not the high-WM, are similar to the pattern for both groups. The same pattern can also be seen in the total reading times of the sentence-initial region (e.g., 'earlier' / 'later') for low-WM readers only.


Figure 4: Tense-response location interference effects in total times in the sentence-final region. Error bars indicate the \(S E\).


Figure 5: Tense-response location compatibility effects for low- (left) and high-WM (right) groups for response times. Error bars indicate the \(S E\).

Next, we attempted to replicate the procedure of Ulrich and Maienborn's (2010) Experiment 2 to assess whether we would find early compatibility effects undetectable in fullsentence response time, and how that may vary as a function of working memory ability. Participants' task was to make a
sentence-sensibility judgment, and then press a button to indicate their decision. In this case thinking about time is irrelevant to the task, because it is not the decision that participants have been asked to make about the sentence.

\section*{Experiment 2: Time irrelevant task}

\section*{Method}

Participants We tested a further 48 students (11 male with the mean age of 23; \(S D=3\) years) who met the same criteria as those in E1.
Materials and Procedure The materials were identical to those in E1. Participants judged sentence sensibility, thus time, or tense, was irrelevant to the task (Fig. 5). This procedure is identical to the second experiment in Ulrich \& Maienborn (2010). Again, the only difference was that the sentences always had the same word order and always included a prepositional phrase after the verb.


Figure 6: Example trial from Experiment 2.

\section*{Results}

Data filtering and separating of participants into two WM groups was done in the same way as in E1. For first-pass reading times, we found a significant 3-way interaction in the verb region, \(t(30)=-2.07, p<0.05\) (see Fig. 7).


Figure 7: Sentence-tense compatibility effects for low- (left) and high-WM (right) groups for first-pass reading times in the verb region. Error bars indicate the standard error.

Next, both first-pass, \(t(30)=2.11, p<0.05\), and regression-path duration, \(t(30)=2.15, p<0.05\), revealed a significant interaction in the sentence-initial temporal adverb region. For both measures, durations were longer when the adverb indicated a past tense sentence and the participants were planning a left response compared to a right response, whereas for future-indicative adverbs there was no reliable difference between left and right response locations. Lastly, we replicated the absence of reliable
compatibility effects in response times for tense (see Figure 8 and Ulrich \& Maienborn, 2010).


Figure 8: This figure illustrates the absence of sentencetense compatibility effects for low- (left) and high-WM (right) groups in response times. Error bars indicate the \(S E\).

\section*{General Discussion}

Consistent with Ulrich and Maienborn (2010), we found compatibility effects in the response times when participants judged sentence tense (Experiment 1, E1) but not when they judged sentence sensibility (Experiment 2, E2). However, eye tracking revealed additional details about the time course and individual differences of the compatibility effects in E1 and the null finding in E2.
When the task was to decide on sentence tense (E1), we found compatibility effects in the earliest sentence region (e.g., 'Earlier'), but in a relatively "late" measure; in total times. This replicated the region of the sentences where temporal processing occurred, similar to Münte et al. (1998) but it is later in the time course than we initially predicted. In contrast to Rapid Serial Visual Presentation used by Münte et al., total times include re-readings and thus potentially later processes. One possible reason why compatibility effects emerge only at this region and not in the verb region is because that region is central to judging sentence tense, and foregrounds tense processing.

Two further unexpected findings in Experiment 1 were the sentence-final interference effect and that the response time compatibility effect was driven by the low-WM readers. The interference effect could index that as participants' prepare for the tense decision and gauge sensibility at the end of the sentence, they may momentarily inhibit temporal information. Because the temporal information has already been mapped into the environment, freeing up resources from WM needed for further language comprehension may be aided by the suppression of environmental patterns, in this case tense-response mappings (Glenberg, 1997, p4). The compatibility effects for low-WM readers at the sentence-initial temporal region are consistent with this idea: Perhaps low WM-readers are slower than high-WM readers to inhibit tense information, and thus show compatibility effect in total times at the sentence-initial region while these effects are absent for high-WM readers. The assumption here is that low-WM readers are slower because they have more difficulty updating their WM, which according to Glenberg is a conscious and effortful use of memory.

For Experiment 2, when the task did not involve a sentence-tense decision, we replicated the absence of a
tense-response location compatibility effect in response times (Ulrich \& Maienborn, 2010). By contrast, compatibility effects emerged as predicted at the verb region in first-pass reading times, but only for high-WM readers. The null effect for low-WM readers may indicate that unless the temporal information is part of the task, low WM readers do not integrate tense information in relation to response location.

Overall, thus, tense-response location compatibility effects varied as a function of task and comprehenders' working memory. Our findings highlight the importance of using continuous measures: While end-of-sentence response times suggested task can eliminate compatibility effects, these effects were clearly present in gaze measures during sentence reading even when the task did not ask participants to focus on tense cues. With regard to embodiment theory, it seems that tense-response location compatibility effects are not eliminated by tasks in which tense is irrelevant but there is a need to accommodate their variation by task and working memory.

\section*{Acknowledgments}

This research was funded by the Cognitive Interaction Technology Excellence Center (German Research Foundation, DFG).

\section*{References}

Baayen, R.H., Davidson, D.J. and Bates, D.M. (2008) Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language, 59, 390-412.
Baddeley, A. (2012). Working memory: Theories, models, and controversies. Annual Review of Psychology, 63, 129.

Barr D. J., Levy R., Scheepers C. \& Tily, H. (2013). Random-effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language, 68, 255-278.
Barsalou, L. W. (1999). Perceptions of perceptual symbols. Behavioral and Brain Sciences, 22, 637-660.
Clark, A. (1997). Being there: Putting brain, body, and world together again. Cambridge, MA: The MIT Press.
Collins, A. M., \& Loftus, E. F. (1975). A spreadingactivation theory of semantic processing. Psychological Review, 82, 407-428.
Collins, A. M., \& Quillian, M. R. (1969). Retrieval time from semantic memory. Journal of Verbal Learning and Verbal Behavior, 8, 241-248.
Glenberg, A. M., Havas, D., Becker, R., \& Rinck, M. (2005). Grounding Language in Bodily States: The Case for Emotion. In R. Zwaan and D. Pecher (Eds.) The grounding of cognition: The role of perception and action in memory, language, and thinking. Cambridge: Cambridge University Press.
Glenberg, A. M., \& Gallese, V. (2010). Action-based language: A theory of language acquisition, comprehension, and production. Cortex, 48, 905-922.

Huettig, F., Olivers, C. N. L., Hartsuiker, R. J. (2011). Looking, language, and memory: Bridging research from the visual world and visual search paradigms. Acta Psychologica, 137, 138-150.
Kemmerer, D., \& Gonzalez-Castillo, J. (2010). The twolevel theory of verb meaning: An approach to integrating the semantics of action with the mirror neuron system. Brain and Language, 112, 54-76.
Koziol, L. F., Budding, D. E., \& Chidekel, D. (2011). From Movement to Thought: Executive Function, Embodied Cognition, and the Cerebellum. Cerebellum, 11, 505-25.
Lakoff, G., \& Johnson, M. (1980). Metaphors we live by. Chicago: University of Chicago Press.
Lakoff, G., \& Johnson, M. (1999). Philosophy in the flesh. New York: Basic Books.
Lewis, R. L., Vasishth, S., \& Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. Trends in Cognitive Sciences, 10, 447454.

Louwerse, M. M., \& Jeuniaux, P. (2008). Language comprehension is both embodied and symbolic. In M. de Vega, A. Glenberg, \& A. C. Graesser (Eds.), Symbols and embodiment: Debates on meaning and cognition (pp. 309326). Oxford: Oxford University Press.

Machery, E. (2007). Concept empiricism: A methodological critique. Cognition, 104, 19-46.
Mahon, B. Z., \& Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. Journal of PhysiologyParis. 102, 59-70.
Münte, T. F., Schiltz, K., \& Kutas, M. (1998). When temporal terms belie conceptual order. Nature, 395, 7173.

Repovš, G., \& Barch, D. M. (2012). Working memory related brain network connectivity in individuals with schizophrenia and their siblings. Frontiers in Human Neuroscience, 137, 1-15.
Spivey, M. (2008). The continuity of the mind. New York, NY: Oxford University Press.
Spivey, M., \& Geng, J. (2001). Oculomotor mechanisms activated by imagery and memory: Eye movements to absent objects. Psychological Research, 65, 235-241.
Symes, E., Ellis, R., \& Tucker, M. (2005). Dissociating object-based and space-based affordances. Visual Cognition, 12, 1337-1361.
Torralbo, A., Santiago, J.,\& Lupiáñez, J. (2006). Flexible conceptual projection of time onto spatial frames of reference. Cognitive Science, 30, 745-757.
Ulrich, R., \& Maienborn. C. (2010). Left-right coding of past and future in language: The mental timeline during sentence processing. Cognition, 117, 126-138.
Unsworth, N., Redick, T. S., Heitz, R. P., Broadway, J. M., \& Engle, R. W. (2009). Complex working memory span tasks and higher-order cognition: A latent-variable analysis of the relationship between processing and storage. Memory, 17, 635-654.

\title{
Word Learning in the Wild: What Natural Data Can Tell Us
}

\author{
Barend Beekhuizen \\ Leiden University Centre for Linguistics \\ Leiden University \\ b.f.beekhuizen@hum.leidenuniv.nl
}

\author{
Afsaneh Fazly, Aida Nematzadeh, Suzanne Stevenson \\ Department of Computer Science \\ University of Toronto \\ \{afsaneh,aida,suzanne\}@cs.toronto.edu
}

\begin{abstract}
When a child first begins to acquire a lexicon, the sources of word-meanings must be available from the situational context. However, it has been argued that the situational availability of the meanings of relational terms, such as verbs, is lower than that of whole-object labels, such as nouns. In this paper, we present a corpus of child-directed language, paired with situational descriptions, that enables us to explore the situational availability of word-meanings using a computational learner.
\end{abstract}

Keywords: word learning; relational meaning; corpus development; computational modeling

\section*{Introduction}

However the lexical acquisition process in infants develops beyond the earliest stages, the seeds of the first word meanings must be found in the immediate situational context of early linguistic interaction (Gleitman, 1990). Bootstrapping these early meanings across a variety of situations, so-called cross-situational learning (Akhtar \& Montague, 1999), is one of the early cognitive tasks that children need to perform. For cross-situational learning to work, the situational contexts have to contain information that can be extracted and used to determine what the caregiver is likely referring to. However, relatively little is known about the information actually contained in situational contexts.

In this paper, we present a corpus of child-directed language in which the situational context, as found in the accompanying video material, is described in a precise, formalized manner. Not only have the basic-level categories of objects been coded, but also some of their properties and the observable relations among agents and objects. This annotated corpus enables us to explore the situational availability of these various sources of meaning using computational modeling techniques. As such, we demonstrate the use of computational models as a methodological tool to gain insight the information that children have available in their natural learning environment, and that can contribute to cross-situational learning of word meaning.

The process of cross-situational learning has been studied using a multitude of methodologies, each with its limitations. Experimental set-ups must trade off control of the stimuli with the naturalism of the interaction, and thus typically underestimate the complexity of the situations caregiver-child interactions normally take place in (as Medina, Snedeker, Trueswell, and Gleitman (2011) recently noted again). Some computational studies use child-directed language from transcribed child language corpora, which require the researchers to automatically enrich the corpora with artificial meaning representations (Fazly, Alishahi, \& Stevenson, 2010).

Interest has grown in the use of multimodal material for computational studies of word-meaning acquisition, since it contains language embedded in a video of the situation of its use. There have been experiments with virtual environments (Fleischman \& Roy, 2005), natural environments in which participants were asked to label objects and actions (Yu \& Ballard, 2003; Roy \& Pentland, 2002), and natural caregiver-child interaction (Roy et al., 2006; Frank, Goodman, \& Tenenbaum, 2009). Despite the greater potential for naturalistic data, these corpora also suffer from limitations. First, some only code whole-object labels, thus restricting themselves to the meaning of one sort of words, viz. nouns (Roy et al., 2006; Frank et al., 2009). In others, the language is not child-directed (Fleischman \& Roy, 2005; Yu \& Ballard, 2003), or the language and situation are unrealistically temporally aligned (Yu \& Ballard, 2003). In this paper, we also overcome the above limitations by developing a corpus of child-directed language paired with a precise description of the situational context. Unlike other corpora, the restriction of our corpus to a particular structured activity allows us to precisely describe situational aspects that are relevant to the meanings of various sorts of content words, although the resulting corpus is necessarily small.

One topic we explore in detail is the extent to which words with observable relational meanings (i.e., physical actions and spatial relations) can be bootstrapped from cross-situational learning. As Gentner (1978) argues, mapping words to relations is more difficult than to objects because relations can typically be construed in more ways. Gleitman (1990) shows how even observable relations are often not present at the time of uttering a word referring to them. This paper shows, using a different methodology, that relational terms are indeed harder to glean from the situational context.

\section*{A Situated Corpus of Child-Directed Language}

Our goal is to construct a corpus that contains situational information that is available to a learner and that can be used in learning the meaning of a variety of content words. For developing such a corpus, there are two requirements. At a minimum, in very early word learning at least, we assume that the information that contributes to a word's meaning must be situationally available - that is, the information must be reflected in the situation that is perceivable at or near the time that the word is uttered. But it must also be the case that the learner can process this information and understands its relevance to the interaction with an interlocutor-i.e., the information must be cognitively available as well.

In recent work on coding the available whole objects in
video data paired with child-directed language (Roy et al., 2006; Frank et al., 2009), generally only situational availability need be considered, because the cognitive availability of the objects is implicitly assumed. Turning to relational terms, as we do here, we must explicitly argue that the appropriate meanings are cognitively available, because of the evidence that gleaning the appropriate relational meanings from a situation is more difficult (cf., Gentner, 1978). Here we assume that, although child-caregiver interactions take place in the complex world of everyday life, cognitive availability of meanings for the child is eased (again, early on) because of the highly-structured nature of such situations, along with the joint attention caregivers and children share for their objects, relations, goals and consequences, which function to narrow down the set of meanings communicated (cf., Tomasello, 2003). Thus we focus the annotation on those meanings we argue to be cognitively available to the child, which are not all the objects and relations in the situation, but only the subset that pertains to the current activity.

The result is a corpus that provides information on both the situationally and cognitively available objects, properties of objects, and relations between objects. These annotations rely on relatively lean assumptions about the cognitive availability of this information. To the best of our knowledge, this is the first corpus that pairs observed objects, properties and relations with spontaneously produced language. As such annotation is costly, the corpus is necessarily small. It can, however, give us insight into the availability of the sources of lexical meaning in the situational context, and the problems a lack of availability may bring about. In that respect this small but naturalistic corpus complements earlier annotated corpora in enabling us to explore what is and is not available at the time some word is uttered.

The source of our material is a collection of 131 videotaped dyadic interactions (recorded for other purposes) between Dutch-speaking mothers and their 16-month-old daughters, containing activities such as playing games and eating. In the videos, each dyad played a game of putting variously-shaped blocks in a bucket with holes of matching shapes in the lid. A set of 32 block games ( 152 minutes of video) was selected for our annotation. The first author (a native speaker of Dutch) transcribed all speech according to CHAT-guidelines \({ }^{1}\), and two assistants coded the video data for the objects, properties and relations in the situations. The transcriptions contained 7842 word tokens ( 480 types) in 2492 utterances. The language mostly refers to aspects of the game.

The situational coding was done according to guidelines developed by the first author. As the situation consists of just one type of activity (playing the game), the set of objects, properties and relations is relatively limited. The most common objects are the bucket, lid, blocks, holes and the two participants, mother and child. The feature color=\{red, green, yellow, blue \(\}\) was coded for the blocks and the feature shape \(=\{\) square, round,

\footnotetext{
\({ }^{1}\) Available at http://childes.psy.cmu.edu/manuals/CHAT.pdf
}

Table 1: Coded relations. Parentheses denote optionality. Ag \(=\) Agent, \(\mathrm{Pa}=\) Patient, \(\mathrm{In}=\) Instrument, \(\mathrm{Re}=\) Recipient, \(\mathrm{So}=\) Source, \(\mathrm{Go}=\) Goals, \(\mathrm{Fi}=\) Figure, \(\mathrm{Gr}=\) Ground
\begin{tabular}{lll}
\hline type & name & roles \\
\hline action grab, letgo, hit & \(\mathrm{Ag}, \mathrm{Pa},(\mathrm{In})\) \\
action point, show & \(\mathrm{Ag}, \mathrm{Pa}, \mathrm{Re},(\mathrm{In})\) \\
action move, force & \(\mathrm{Ag}, \mathrm{Pa}, \mathrm{So}, \mathrm{Go},(\mathrm{In})\) \\
action position & \(\mathrm{Ag}, \mathrm{Pa}, \mathrm{Gr},(\mathrm{In})\) \\
spatial in, on, off, out, at, near \(\mathrm{Fi}, \mathrm{Gr}\) \\
spatial match, mismatch & \(\mathrm{Fi}, \mathrm{Gr}\) \\
\hline
\end{tabular}
triangular, star\} for blocks and holes. The relations and their roles are in Table 1.

For every three-second interval of video, all coderobserved relations, their associated objects and their properties were coded. \({ }^{2}\) The actions (first four rows of Table 1) denote simple manual behavior, which we assume children can recognize (Baillargeon \& Wang, 2002). The spatial relations reflect basic categories of containment and support (in,on) and their negation (out,off), as well as two relations denoting non-containment and non-support contact (at) and nearness (near). Understanding basic spatial relations precedes the onset of meaning acquisition and can thus be assumed to be in place (Needham \& Baillargeon, 1993; Hespos \& Baillargeon, 2001), although many specifics may be languagespecific (Choi, 2006). \({ }^{3}\) The match or mismatch with a hole was furthermore inferred from these relations. Spatial relations were deemed salient if a change in the relation occurred (e.g., if a block was the Figure of an in-relation in the current interval, when it was not in the previous interval).

The coding procedure was evaluated for inter- and intracoder agreement (Carletta, 1996). All relations were coded reliably both within and between coders (Cohen's \(\kappa>0.8\) ), except position (intercoder: \(\kappa=0.51\), intracoder: \(\kappa=\) 0.47). When the coders disagreed, the first author decided the annotation. A sample of the resulting data is given in Table 2.

\section*{The Computational Model}

We use the probabilistic alignment-based word learning model of Fazly et al. (2010), which has been shown to perform well using naturalistic data. Using a computational model, we can manipulate input, and doing so, explore the situational and cognitive availability of information, as well as how changes in the input affect learning (Experiment 2).

The model incrementally takes as input a pair of an utterance (a set of words) and a situation (a set of primitive meanings). The learning algorithm has two phases. In the alignment phase, the words and meanings in the input are prob-

\footnotetext{
\({ }^{2}\) Using ELAN (Brugman \& Russel, 2004).
\({ }^{3}\) Ideally, one would encode the range of construals of a situation, including 'tightness-of-fit'. As a first attempt at relational coding of situations, we opted for convenient, yet widely known, universal notions like 'containment' and 'support'.
}

Table 2: A sample of the dataset. The dash-separated abbreviations denote blocks and holes and their properties, where for blocks the order is \(\mathbf{b}\) - \(\{\mathbf{r e d}, \mathbf{g r e e n}\), blue, yellow \(\}-\{\) round,star,square,triangular\}, and for holes ho- \(\{\) round,star,square,triangular\}
\begin{tabular}{lll}
\hline time & type & coding/transcription \\
\hline Om0s & \begin{tabular}{l} 
situation \\
utterance \\
translation
\end{tabular} & \begin{tabular}{l} 
< nothing happens> \\
een. nou jij een. \\
one. now you one. "One. Now you try one."
\end{tabular} \\
\hline 0m3s & situation & \begin{tabular}{l} 
position(mother, toy, on(toy, floor)) grab(child, b-ye-tr) \\
move(child, b-ye-tr, on(b-ye-tr, floor), near(b-ye-tr, ho-ro)), mismatch(b-ye-tr, ho-ro) \\
\\
\\
\\
utterance \\
translation
\end{tabular} \\
\hline nee daar. \\
no there. "No, there."
\end{tabular}
abilistically mapped to each other; this process is guided by the conditional probabilities of the meanings given the words ("the learned meanings"). Second, in the update phase the obtained alignments are used to update the word-meaning associations by adding the alignment score to the association. The word-meaning associations, next, are used to calculate the learned meanings, which are then used in the alignment phase of the next input. These probabilities are based on the association mass a meaning has for a word, relative to all other meanings associated with that word. For a formal explanation, we refer the reader to Fazly et al. (2010).

\section*{Experiment 1: Exploring the Corpus}

Using the computational model and the corpus, we aim to gain insight into questions such as: what kind of and how much information is derivable from the situational contexts? And is the information equally valuable for different kinds of words (relational words like verbs and prepositions, and non-relational words like adjectives and nouns)?

\section*{Running the Model}

A set of each utterance's lemmatized word forms is used as the linguistic input. As the model takes a set of primitive meanings as the other part of its input, we considered all content elements from the structured meaning annotation of the interval containing the start of the utterance as the set of situation primitives. An example of an input item is:

> Utterance: \(\{\) nee lieverd hier passen hij niet \}
> Situation: \{ point, mother, hole, triangular, child, position, block, yellow, near, round, mismatch \}

We set the two smoothing parameters of the model to reflect the size of the lexicon, as in Fazly et al. (2010).

\section*{Evaluation}

We need to understand how the model learns various types of words that refer to aspects of the situational context. To

Table 3: A sample of the lexicon of target words
\begin{tabular}{ll}
\hline type & examples \\
\hline action & duwen \(=\) force, halen \(=\{\) move, off, out \(\}\) \\
spatial & in \(=\) in, af \(=\) off, dicht \(=\{\) lid, on, bucket \(\}\) \\
object & gat \(=\) hole, emmer \(=\) bucket \\
property & rood \(=\) red, ster \(=\) star \\
\hline
\end{tabular}
this end we need some sort of gold standard, as well as some measure of how well the model approximates this standard.

Many words in the utterances have no semantic representation in the coded situations (articles, modals, discourse particles). As we cannot expect the model to learn anything about them, we do not consider them in our evaluation. This leaves us with a small subset of lemmas \((n=41)\) that do refer to possible aspects of the situation. These are verbs of manipulation (e.g., pakken 'grab') and placing (e.g., stoppen 'put into'), spatial relations (e.g., op, 'on'), object labels (e.g., blok 'block') and properties (vierkant 'square'). As some words have multiple meanings (stoppen meaning put and in), we have to determine which set of meanings should be associated with each word. Table 3 gives a sample of words and their relevant gold-standard (true) meanings.

We evaluate the learned meanings using two measures. First, we look at the summed meaning probabilities over the set of true meanings (Summed Conditional Probability or \(S C P\) ). This measure tells us what proportion of the probability mass is correctly assigned.
\[
\begin{equation*}
S C P=\sum_{f \in \text { true meanings }(w)} p(f \mid w) \tag{1}
\end{equation*}
\]

Second, we look at how high the true meanings are ranked among all learned meanings, and do so using Average Precision \((A P)\), calculated as follows:
\[
\begin{equation*}
A P=\sum_{k=1}^{n} P(k) \Delta r(k) \tag{2}
\end{equation*}
\]


Figure 1: Development of the lexicon's mean \(S C P\) and \(A P\)
where \(k\) is the rank, \(n\) the total number of ranks, \(P(k)\) is the number of true meanings found up to and including \(k\), divided by the number of meanings found up to and including \(k\), and \(\Delta r(k)\) is the change in recall between \(k-1\) and \(k\), which is the number of true meanings found at \(k\) divided by the total number of true meanings (which is zero in case no true meanings are found at rank \(k\) ). This tells us whether the true meanings are more or less prominent than the irrelevant ones.

\section*{Results}

Table 4 presents the global results, binned per meaning type (properties, objects labels, spatial relations, and actions). We can see that the meanings of non-relational word meanings are ranked higher than those of relational word meanings (compare \(A P=0.81\) and \(A P=0.25\) for properties and object labels, with \(A P=0.19\) and \(A P=0.15\) for spatial relations and action labels), although \(S C P\) does not differ much between the categories. In general, the probability distributions of the learned meanings do not have very strong peaks: the highest ranking meanings rarely have a learned meaning probability of more than 0.20 . Nevertheless, with 78 primitive meanings, the model does learn well beyond a baseline of \(\frac{1}{78}=0.013\).

Looking at the development of the \(S C P\) and \(A P\) values over time (Fig. 1), we see strikingly little development in the \(S C P\), whereas the \(A P\) rises for a time, then shows a slight decline.

Splitting the developmental curves out over some of the words (Fig. 2), we see that the words are learned rather heterogeneously. Looking at the \(A P\) first, some words are acquired instantly, with \(A P=1\) (i.e., the correct meaning ranking first) from early on (groen and rond), others gradually

\section*{Table 4: Results of Experiment 1}
\begin{tabular}{llllll}
\hline & property & object & spatial & action & total \\
\hline\(S C P\) & 0.10 & 0.05 & 0.09 & 0.07 & \(\mathbf{0 . 0 8}\) \\
\(A P\) & 0.81 & 0.25 & 0.19 & 0.15 & \(\mathbf{0 . 3 1}\) \\
\hline
\end{tabular}


Figure 2: Development of \(S C P \& A P\) over time for 9 words
approach \(A P=1(g a t)\), while for most words, the true meanings remain low ranked. There is, however, a development towards a higher \(A P\) for many of these words, except for halen and uit. The \(S C P\) remains low in all cases, even when the true meaning is ranked first (as in groen and rond), although note that for several words there is some improvement in \(S C P\) over time. Recall that the model has only seen 2492 utterances at this point, and that more data may increase the \(S C P\) further.

\section*{Discussion}

In this experiment, the model does not learn most words well. One potential reason is the small data set, representing only three hours of interaction. We observe that many developmental curves seem not to have reached their asymptotes yet, suggesting that further learning could occur with more data. We also, admittedly, have the model discard valuable information from the data. Both the linguistic structure (syntax) and the semantic structure (predicate-argument relations) are currently ignored by the model but could be useful in creating the mapping.

In addition, the highly structured and restricted nature of the data, which we expected to help by focusing the learning, may actually be hindering performance. We observe that some words have a very high ranking for their true meanings (high \(A P\) ), yet have low learned probability mass (low \(S C P\) ). (For example, see the words groen and rond in Figure 2.) On the one hand, the structured and restricted nature of the blocks game entails that a word's true meaning often consistently appears with it. On the other hand, however, the limited nature of the interactions in the data also entails that many irrelevant meanings consistently appear with the word. For example, the object of a grab action is almost always a block, so that the learner cannot rule out block as a possible
meaning of pakken 'grab'. The lack of situational variability in the input is thus an obstacle to cross-situational learning, because it requires a consistent co-occurrence of true meanings with a word coupled with variability in the presence of irrelevant meanings to help rule them out.

A first solution that comes to mind is a corpus representing a wider variety of activities, with less situational uniformity, for true cross-situational learning. The corpus from which we drew our dyads here does have a number of other types of situations we can include in future annotation. Second, even with relatively homogeneous situations, we expect the learner's attentional mechanisms to help filter out irrelevant meanings. Adding attentional mechanisms, such as the ones in Nematzadeh, Fazly, and Stevenson (2012), is a next step

A final issue we observed with the data is that the true meanings for words in an utterance are sometimes not present within the situational interval paired with the utterance. This problem is very salient for relational meanings, which are often displaced in time from the utterance that refers to them (e.g., Go grab that one! or Don't take the lid off now!). This might explain why spatial relation terms and verbs display weaker associations with their true meanings than do words for objects and their properties. In the case of positive imperatives, we do find that the actions are often carried out slightly later than the utterance. In Experiment 2, we explore whether this problem of temporal displacement can be mitigated.

\section*{Experiment 2: Widening the Temporal Scope}

Our hypothesis is that presenting the model with situational meanings only from the time of the utterance impedes the learning of relational terms. Here we explore expanding the temporal scope of the situational input to the model.

\section*{Motivation and Set-up}

Suppose that in word learning, the learner is not narrowly focussed on the situation at exact moment of the utterance, but also considers some of the situational context taking place around that moment. That is: not only the situation at the very moment of the utterance is cognitively available to a learner, but also some of the surrounding situations. To make this notion precise, we assume that the learner may consider as relevant to an utterance \(U_{i}\) any meanings in the situational context starting from the interval of the previous utterance \(U_{i-1}\) up to and including the interval of the next utterance \(U_{i+1}\). (That is, we assume that the relevance of situations overlaps previous and subsequent utterances.) We thus evaluate the model on three possible "windows" \(W\) of situational context for utterance \(U_{i}\) : all video intervals up to and including the previous and next utterance in the corpus ( \(W=U_{i-1}: U_{i+1}\) ); only the interval of \(U_{i-1}\) up to the current interval ( \(W=U_{i-1}: U_{i}\) ), or the current interval up to \(U_{i+1}\left(W=U_{i}: U_{i+1}\right)\).

\section*{Results}

Using the same parameter settings and evaluation metrics as in Experiment 1, we obtain the results in Table \(5\left(W=U_{i}: U_{i}\right.\)

Table 5: Results of Experiment 2
\begin{tabular}{lllllll}
\hline\(W\) & & prop. & object & \multicolumn{3}{c}{ spatial action } \\
\hline\(W\) & total \\
\hline \multirow{2}{*}{\(U_{i}: U_{i}\)} & \(S C P\) & 0.10 & 0.05 & 0.09 & 0.07 & \(\mathbf{0 . 0 8}\) \\
& \(A P\) & 0.81 & 0.25 & 0.19 & 0.15 & \(\mathbf{0 . 3 1}\) \\
& \(S C P\) & 0.10 & 0.04 & 0.09 & 0.07 & \(\mathbf{0 . 0 7}\) \\
\(U_{i-1}: U_{i}\) & \(A P\) & 0.80 & 0.17 & 0.20 & 0.14 & \(\mathbf{0 . 3 1}\) \\
& \(S C P\) & 0.11 & 0.06 & 0.11 & 0.08 & \(\mathbf{0 . 0 8}\) \\
\(U_{i}: U_{i+1}\) & \(A P\) & 0.79 & 0.45 & 0.24 & 0.18 & \(\mathbf{0 . 4 0}\) \\
& \(S C P\) & 0.08 & 0.05 & 0.10 & 0.08 & \(\mathbf{0 . 0 7}\) \\
\multirow{2}{*}{\(U_{i-1}: U_{i+1}\)} & AP & 0.79 & 0.41 & 0.22 & 0.20 & \(\mathbf{0 . 3 9}\) \\
\hline
\end{tabular}
is the window-setting used in Experiment 1). The windowsetting that only draws situational context from the intervals between the previous utterance and the current one ( \(W=\) \(U_{i-1}: U_{i}\) ) does not improve over \(W=U_{i}: U_{i}\). As hypothesized, however, due to utterances that refer to future actions, the results show that having a window that includes meanings from the intervals up to the next utterance enables the model to learn the object, spatial and action words better (at least according to our \(A P\) measure). The trade-off is a negligible decline in the learning of property words.

\section*{Discussion}

Some important information for acquiring the meaning of relational words can be found in the situations unfolding after the utterance has been produced. Clearly, this needs to be interpreted within the context of playing a game, in which the relevant topics of communication (the game goals) often lie in the future w.r.t. the moment of communication. While expanding the situational window adds some irrelevant as well as true meanings, the balance struck by this pragmaticallydefined windowing approach seems to help the model acquire the meaning of relational terms (as well as objects!) somewhat better, with little negative impact on property words. Note that the improvement from adding the post-utterance meanings is found mainly in the \(A P\) metric: the \(S C P\) values remain similar across the simulations. Even though the probability mass of the true meanings is not changed much, they are now more often better than the irrelevant meanings. This means that the probability values are close to each other and a very small change may improve the rankings visibly.

\section*{General Discussion and Future Directions}

In this research, we have developed a corpus of caregiverchild interactions in which video is annotated with transcribed utterances and a precise description of the depicted situational context. Unlike other recent multimodal corpora, our annotation of the situational context includes meaning elements that correspond not only to objects and their properties, but to relations as well. Thus the meaning annotations support the learning of various word types, including nouns, adjectives, prepositions/particles, and verbs. Our initial work has explored how we can use this corpus with a computational
model of cross-situational word learning to explore what information must be available to the child from the situation to support word learning, and to examine the relative ease or difficulty of learning various types of words in early acquisition.

Despite the small size of the target lexicon, the model did not perform robustly in the learning task, revealing a number of potential areas of improvement for both the corpus and the model itself. First, due to the cost of annotation, the size of the corpus (only 8,000 word tokens) almost certainly limits the learning. Nonetheless, even this small corpus can be a complementary source of information to larger corpora that are semantically less naturalistic, or contain only object labels. Second, the corpus seems to lack sufficient crosssituational variability for many words to be learned. In more general child-caregiver interactions, a word occurs across a wider variety of contexts (eating scenes, bed-time procedures and so on), enabling a child to rule out as possible meanings those aspects of the context that are irrelevant to the word. Third, regardless of the uniformity or variability of the data, a realistic model of word learning needs to incorporate an attentional mechanism that helps it focus on those aspects of the situation that are likely to be referred to.

Even with this restricted corpus, we find that relational words (verbs, prepositions) are particularly problematic to learn compared to words for objects and properties, in line with a wealth of psycholinguistic observation to this effect (Gleitman, 1990; Gentner, 1978). Because the situational context to which a relational term refers is often displaced, expanding the temporal window of situational context for each utterance led to an improvement in the learning of relational terms, but surprisingly led to even greater improvement in the learning of words for objects.

Perhaps, following Gleitman, more structured learning is necessary for acquiring the meaning of relational words, but the exact source and nature of this structured learning, and its integration with methods of cross-situational learning, is an exciting open issue. Important to look into, and perhaps problematic, is the high proportion of closed-class items in childdirected utterances (e.g., pronouns, aspectual and modal auxiliaries, and particles) that have received little attention in word-learning models, but may play a crucial role in using the structure of an utterance to help determine the meaning of unknown lexical items. More research into the degree to which this information, as found in actual child-directed language, can help is a question in want of an answer, and modeling techniques combined with good data can help us approach it.

\section*{Acknowledgments}

We gratefully acknowledge the funding of BB through NWO of the Netherlands (grant 322.70.001) and AF, AN and SS through NSERC of Canada, and the Faculty of Arts \& Science, University of Toronto. We would like to thank Marinus van IJzendoorn and Marian Bakermans-Kranenburg for making the video data available, and Arie Verhagen and two anonymous reviewers for helpful comments.

\section*{References}

Akhtar, N., \& Montague, L. (1999). Early Lexical Acquisition: The Role of Cross-Situational Learning. First Language, 19(57), 347-358.
Baillargeon, R., \& Wang, S.-H. (2002). Event Categorization in Infancy. Trends in Cognitive Sciences, 6(2), 85-93.
Brugman, H., \& Russel, A. (2004). Annotating Multimedia/ Multi-modal resources with ELAN. In Proceedings LREC.
Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. Computational Linguistics, 22(2), 249-254.
Choi, S. (2006). Preverbal Spatial Cognition and LanguageSpecific Input: Categories of Containment and Support. In K. Hirsh-Pasek \& R. M. Golinkoff (Eds.), Action Meets Word. How Children Learn Verbs (pp. 191-207). Oxford, UK: Oxford University Press.
Fazly, A., Alishahi, A., \& Stevenson, S. (2010). A Probabilistic Computational Model of Cross-Situational Word Learning. Cognitive Science, 34(6), 1017-1063.
Fleischman, M., \& Roy, D. K. (2005). Why Verbs are Harder to Learn than Nouns. Initial Insights from a Computational Model of Intention Recognition in Situated Word Learning. In Proceedings CogSci.
Frank, M. C., Goodman, N. D., \& Tenenbaum, J. B. (2009). Using Speakers Referential Intentions to Model Early Cross-Situational Word Learning. Psychological Science, 20(5), 578-585.
Gentner, D. (1978). On Relational Meaning: The Acquisition of Verb Meaning. Child Development, 49, 988-998.
Gleitman, L. (1990). Sources of Verb Meanings. Language Acquisition, 1(1), 3-55.
Hespos, S. J., \& Baillargeon, R. (2001). Reasoning about Containment Events in Very Young Infants. Cognition, 78(3), 207-45.
Medina, T. N., Snedeker, J., Trueswell, J. C., \& Gleitman, L. R. (2011). How Words Can and Cannot be Learned by Observation. PNAS, 108(22), 9014-9.
Needham, A., \& Baillargeon, R. (1993). Intuitions about Support in 4.5-Month-Old Infants. Cognition, 47, 121148.

Nematzadeh, A., Fazly, A., \& Stevenson, S. (2012). A Computational Model of Memory, Attention, and Word Learning. In Proceedings CMCL.
Roy, D. K., Patel, R., Decamp, P., Kubat, R., Fleischman, M., Roy, B., et al. (2006). The Human Speechome Project Stepping into the Shoes of Children. In Proceedings CogSci.
Roy, D. K., \& Pentland, A. P. (2002). Learning Words from Sights and Sounds: A Computational Model. Cognitive Science, 26, 113-146.
Tomasello, M. (2003). Constructing a Language: A UsageBased Theory of Language Acquisition. Cambridge, MA: Harvard University Press.
Yu, C., \& Ballard, D. H. (2003). A Multimodal Learning Interface for Grounding Spoken Language in Sensory Perceptions. Proceedings ICMI.

\title{
Visual Support for Instructional Analogy: Context Matters
}

\author{
Kreshnik Nasi Begolli (kbegolli@uci.edu) \\ School of Education, 3200 Education \\ Irvine, CA 92697 USA
}

\author{
Lindsey Engle Richland (lrichland@uchicago.edu) \\ Department of Comparative Human Development, 5730 S. Woodlawn Ave. Chicago, IL 60637
}

\begin{abstract}
Instructional analogies can overload children's executive function and working memory resources (see Richland, Morrison \& Holyoak, 2006), though structure-mapping lies at the core of recommended pedagogy in mathematics instruction (National Mathematics Panel, 2008; NRC, 2001). Videotaped mathematics instruction was manipulated to test the role of visual representations in instructional analogy. Pretest, posttest, and delayed posttest measures assessed 1113 year old children's learning from one of three versions of the same lesson in which three solution strategies (one a misconception) were compared. Analogs were either a) Not Visible (NV) - presented only orally, b) Partially Visible (PV) - only the most recent solution was visible, or 3) All Visible (AV) - all solutions were visible throughout the instruction. Overall, AV students experienced greater learning gains in procedural knowledge, procedural flexibility, and conceptual/ schematic knowledge compared to PV students. These results persist after one-week delay. Apart from procedural knowledge, the same trend is evident when comparing AV students' to NV students' immediate learning gains. Overall, visual representations of analogs within an instructional analogy appear to support schema formation only when they are all visible simultaneously and throughout structuremapping. Showing students visual representations of analogs but not enabling them to be simultaneously visible led to the lowest performance overall, suggesting this may lead to more object-level encoding than schema formation.
\end{abstract}

Keywords: analogy; comparison; mathematics education; video stimulus; misconception; executive function.

Comparing different student solutions to a single instructional problem is a key recommended pedagogical tool in mathematics, however the cognitive underpinnings of successfully completing this task are complex. Students must represent the multiple solutions as relational systems, align and map these systems to each other, and draw inferences based on the alignments (and misalignments) for successful schema formation (see Gentner, 1983; Gick \& Holyoak, 1983; Richland, Zur \& Holyoak, 2007).

Orchestrating classroom lessons in which learners successfully accomplish relational structure mapping is not straightforward, particularly because opportunities for learning through structure mapping often fail in laboratory contexts (e.g., Gick and Holyoak, 1983; Ross, 1989). Specifically, reasoners regularly fail to notice the utility of aligning and mapping two or more available relational structures.

The low success rate with which participants notice and use relational structure mapping, or analogy, within laboratory studies to solve problems may in part reflect limitations in the working memory system (see Waltz, Lau, Grewal \& Holyoak, 2000). Working memory is required to relationally represent systems of objects, in this case steps to solution strategies, to re-represent these systems of relations so that their structures can align and map together, to identify meaningful similarities and differences, and to derive conceptual/ schematic inferences from this structuremapping exercise to better inform future problem solving (see Morrison, Krawczyk, Holyoak et al 2004).

The current study tests the role of visual representations of the source and target analogs within an opportunity for structure-mapping. The manipulation assesses whether 1) making source and target analogs visual (versus oral) increases the likelihood that participants will notice and successfully benefit from structure mapping opportunities, and 2) whether the visual representations must be visible simultaneously during structure-mapping in order to increase the likelihood of future success in problem solving and schema formation. The former is likely to increase the salience of the relational structure of each representation, while the latter is likely to reduce the working memory load and executive function resources necessary for participants to engage in structure-mapping and inference processes.

These are research questions with high ecological validity. A cross-cultural study of \(8^{\text {th }}\) grade mathematics instruction revealed that comparing verbal and visual structured representations is a common practice in U.S. mathematics classrooms as well as in higher achieving regions (Hong Kong and Japan), but that U.S. teachers are less likely to make visual representations visible during a structure-mapping episode than the teachers in higher achieving countries (Richland, Zur \& Holyoak, 2007). Thus findings from this experiment will yield both theoretical insight into the resource load necessary for complex structure mapping and schema formation, and practice relevant implications for everyday mathematics teachers.

Because the study takes ecological validity and the complexity of everyday classrooms as serious constraints, a novel methodology was used to derive rigorous, experimental data that incorporates the complexities of situated cognition. Specifically, the stimuli for the experiment derive from videotapes of a public school
teacher in her naturalistic classroom, teaching a lesson codesigned with the research team. This methodology and its motivation are next explained in more detail, followed by a report of the experiment itself.

\section*{Video-editing as a Tool to Bridge Laboratory and Classroom Settings}

Classrooms are vibrant, complex environments in which the high level of unexpected variability makes experimental control often impossible (Brown, 1992). The overarching commitment to controlled manipulation of experimental contexts within psychological research has led much cognitive scientific study of learning behavior to be conducted in controlled laboratory settings. While in some ways this model leads to the production of data that can be easily interpreted (x behavior derived from y manipulation), the meaningfulness of these results for educational practice have been less clear. Theoretically, this research epistemology has also meant that the search for universal cognitive processes of learning can best be accomplished through the design and examination of cognition within atypical, impoverished environments (see Schweder, 2012). The assumption that cognitive mechanisms underlying classroom learning are not moderated by environmental factors is unexplored.


Figure 1. Still images illustrating the experimental conditions created by video editing the same lesson, from left to right: Not Visible, Part Visible, and All Visible.

The current study does not interrogate that question but rather reduces the assumption by situating the stimuli creation in the naturalistic classroom context itself. A naturally occurring classroom lesson is videotaped using three cameras that capture different features of the lesson, (e.g., teacher only and teacher plus visual representations) though the same classroom discourse, affect, eye gaze, and many other potentially important features of the context are held constant across cameras. The distinct camera angles are then used to create different conditions of a videotape of the
same lesson, which are then shown to a new group of classroom students. This is clarified in the below description of stimuli creation for the current study.

\section*{Experiment: Impact of Visual Support for Instructional Analogy}

\section*{Method}

Participants. Participants were drawn from a suburban public school with a diverse population. Five students that scored in the bottom \(5 \%\) of the participant pool were also excluded from analyses. The final analyses included 78 students ( 46 boys, 32 girls) with ages ranging between 1112 years old. Within classrooms, students were randomly assigned to condition, with 25 students in the All Visible condition, 27 students in the Part Visible condition, and 26 students in the Not Visible condition.

Materials. Materials for the intervention consisted of a worksheet, a netbook, and a pre-recorded video-lesson embedded in an interactive computer program. The lesson used in the current study was developed by the authors in collaboration with a public school teacher. Three cameras were used simultaneously to videotape a classroom lesson on ratio. Ratio was chosen for this study for two reasons: (a) it is part of the common core standards for elementary mathematics instruction and (b) previous research has shown that ratio problems prompt diverse systematic student responses, useful for charting trajectories of reasoning change across the study. One camera was set to a wide shot, captured the teacher, parts of the classroom, and all visual representations of the three solution analogs throughout the lesson (All Visible -AV). A second camera was more tightly focused, capturing the teacher, some of the class, and only the visual representation of a solution as it was being produced (Part Visible - PV). The third camera focused only on the teacher and students, and did not capture any of the visual representations of the solutions written onto the white-board (Not Visible - NV; see Figure 1 for an illustration of each condition).

The video-lesson was made interactive by embedding clips of the video in a computer program. These stimuli were then used experimentally with students in other classrooms. This methodological approach of stimuli creation, provided a rigorous level of experimental control of a highly dynamic context - an everyday classroom. Further, it allowed for randomization within each classroom, which to the authors knowledge has not been previously done using a video teacher's guidance.


Figure 2. Gain scores for immediate and 1-week delayed posttest calculated by subtracting mean pretest score with respective posttest score.

Table 1. Mean Gain Scores by Knowledge Type for Each Construct Calculated by Subtracting Pretest from Respective Posttest.
\begin{tabular}{|c|c|c|c|c|c|c|c|c|c|c|c|c|}
\hline Knowledge Type & \multicolumn{2}{|l|}{Procedural} & \multicolumn{2}{|l|}{Flexibility} & \multicolumn{2}{|l|}{Conceptual} & \multicolumn{2}{|l|}{Negative Transfer} & \multicolumn{2}{|l|}{Misconception} & \multicolumn{2}{|l|}{Efficient Strategy} \\
\hline Time of Test & Immediate & Delayed & Immediate & Delayed & Immediate & Delayed & Immediate & Delayed & Immediate & Delayed & Immediate & Delayed \\
\hline All Visible & 13\% & 20\% & 19\% & 27\% & 23\% & 25\% & -28\% & -24\% & -7\% & -10\% & 35\% & 45\% \\
\hline Part Visible & 2\% & -1\% & -1\% & 7\% & 18\% & 13\% & -10\% & -7\% & 9\% & 12\% & 17\% & 13\% \\
\hline Not Visible & 19\% & 27\% & 9\% & 21\% & 11\% & 18\% & -19\% & -14\% & -2\% & -3\% & 20\% & 33\% \\
\hline
\end{tabular}

Assessment. The assessment was designed to assess schema formation and generalization. Mathematically, the assessment included three constructs, procedural knowledge, flexibility, conceptual knowledge, and negative transfer. The first three constructs were conceptually derived from Rittle-Johnson and Star (2007; 2009), and adapted to the core concepts and procedures underlying ratio problems (Figure 3). Scores for each construct were averaged to yield an overall mean for that particular construct.

Ken and Yoko shot several freethrows in their basketball games. The result of their shooting is shown in the table. Who is better at shooting free throws?
\begin{tabular}{|c|c|c|}
\hline & Shots Made & Total Shots Tried \\
\hline Ken & 12 & 20 \\
\hline Yoko & 16 & 25 \\
\hline
\end{tabular}

Adelina and Marcos have both set up lemonade stands. Adelina's lemonade recipe uses 2 cups of lemon juice and 1 cup of water. Marcos' lemonade recipe uses 3 cups of lemon juice and 2 cups of water.


Figure 3. Procedural/Procedural flexibility problem (left) used in the video-lesson and assessments, and conceptual problem used in assessments. For procedural flexibility students were told to solve a problem similar to the one on the left using two different strategies.

Procedural Knowledge. Procedural problems on the pretest evaluated whether students had the basic skills necessary to solve ratio problems and designed to test students' knowledge of producing solutions of familiar and transfer problems. Cronbach's alpha was .89 at posttest, .92 at delayed posttest, and .86 at pretest.

Procedural Flexibility. The procedural flexibility construct measured: (a) students' adaptive production of solution methods ( \(n=3\) ), (b) their ability to identify the most efficient strategy ( \(\mathrm{n}=1\) ), and (c) students' ability to identify a novel solution method which was related to a taught strategy ( \(\mathrm{n}=1\) ). Cronbach's alpha on the flexibility construct was .67 at posttest, .67 at delayed posttest, and .57 at pretest.

Conceptual Knowledge. The conceptual construct was designed to probe into students' explicit and implicit knowledge of ratio. Cronbach's alpha was .66 at posttest, .64 at delayed posttest, and .42 at pretest.

Negative Transfer. The purpose of the negative transfer construct was to measure whether students will overextend their knowledge of ratio to similar looking problems for which a strategy shown to be invalid during instruction subtraction, is correct. While this construct was expected to assess overextensions of the taught strategy, due to its high similarity with the taught problems, it can also help diagnose whether conditions that do not eliminate the misconception appear to be sensitive to variations in the
problem type. Cronbach's alpha was .68 at posttest, .81 at delayed posttest, and .58 at pretest.

Efficient Strategy. The aim for this measure was to assess learners' ability to utilize the most efficient solution as instructed during the video lesson. This has also been called adaptive choice of strategy (Siegler, 1996). Efficient strategy was assessed by scoring all problems taking the form of the problem taught in the video lesson to evaluate whether students used the most efficient strategy taught - the division method.

Common Misconception. Misconceptions are mistakes that students make, which obstruct learning (Smith, diSessa, Roschelle, 1994). Based on a published lesson (Shimizu, 2003), pilot data, and pretest data, a solution involving subtraction was expected to be the most common misconception participants would bring to the study. This score assessed students' ability to overcome their misconceptions about how to solve rate and ratio problems as well as the conditions under which students confirm invalid biases. The common misconception measure examined students' use of subtraction by scoring problems that looked like the instructed problem in the video lesson.

Design \& Procedure. Students who were not in the original classroom lesson interacted with videotaped lesson clips via computer. The study followed a standard experimental procedure (pretest, intervention, immediate posttest, and 1-week delayed posttest). Students within a classroom were randomly assigned to either watch an instructional video version video-edited so that no solutions were visible on the board - Not Visible ( \(\mathrm{n}=26\) ), a version where the most recent solution was visible - Part Visible (n \(=27\) ), or a version of the video that showed all solutions on the board throughout the lesson - All Visible ( \(\mathrm{n}=25\) ). All students were given a packet on which they recorded their answers to prompts from the videotaped lesson. Students underwent the intervention before being introduced to rate and ratio.

\section*{Results}

First, between-subjects regression analyses revealed no differences between conditions on any of the outcome constructs. Boys and girls also did not differ in performance. Separate univariate one-way between-subjects ANCOVAs were run for each construct with posttest or delayed posttest as a dependent variable and pretest as a covariate. Average gain scores at posttest for the main constructs are shown in Figure 2, and the full set of gain scores immediately and after a delay are provided in Table 1. Table 2 provides all statistics, revealing that the All Visible condition outperformed the Part Visible condition in procedural knowledge, procedural flexibility, conceptual knowledge and efficient strategy both on immediate and

Table 2. Between Group Comparisons for Each Construct on Immediate and Delayed Posttest. p-values \(*>0.10, * *>0.05, * * *>0.01\)
\begin{tabular}{|c|c|c|c|c|c|c|c|c|c|c|c|c|}
\hline Knowledge Type & \multicolumn{2}{|l|}{Procedural} & \multicolumn{2}{|l|}{Flexibility} & \multicolumn{2}{|l|}{Conceptual} & \multicolumn{2}{|l|}{Negative Transfer} & \multicolumn{2}{|l|}{Misconception} & \multicolumn{2}{|l|}{Efficient Strategy} \\
\hline Time of Test & Immediate & Delayed & Immediate & Delayed & Immediate & Delayed & Immediate & Delayed & Immediate & Delayed & Immediate & Delayed \\
\hline All vs. Not Visible & 0.646 & 0.951 & 0.041** & 0.250 & 0.038** & 0.338 & 0.117 & 0.308 & 0.095* & 0.171 & 0.307 & 0.297 \\
\hline All vs. Part Visible & 0.027** & 0.019** & 0.000*** & 0.021** & 0.073* & 0.030** & 0.014** & 0.104 & 0.018** & 0.000*** & 0.001*** & 0.000*** \\
\hline Not vs. Part Visible & 0.079* & 0.026** & 0.176 & 0.150 & 0.750 & 0.325 & 0.377 & 0.530 & 0.498 & 0.027** & 0.117 & 0.014** \\
\hline
\end{tabular}
delayed posttests. The inverse was true for the negative transfer on immediate posttest and common misconception construct on both posttests, which is indicating that most students are being misled by the appearance of the problems (which are similar to the ratio problems) and assume they are solving for a ratio problem, not a simple subtraction problem. Thus, PV students may have used the misconception throughout all problems, regardless of whether it was correct or not. Students in the All Visible condition also outperformed students in the Not Visible condition in the flexibility and conceptual knowledge constructs on the immediate posttest. For AV and NV students the differences do not seem to hold at a delayed posttest. The NV students were better than PV students when measured for use of efficient strategy and common misconception, after a 1-week delay.

\section*{Discussion}

The results of this study clarify the role of visual representations in supporting structure-mapping and generalization from instructional analogies. The manipulation revealed that making source and target analogs visual (versus oral) increased the likelihood that participants notice and benefit from structure mapping opportunities. As noted above, the use of a visual representation of a structured relational analog was hypothesized to be likely to increase the salience of the relational structure of each representation, while maintaining their visibility was predicted to reduce the working memory load and executive function resources necessary for participants to engage in structure-mapping and inference processes.

The data revealed that this variation in visible representations had a great impact on their learning. Seeing all problem solutions on the board simultaneously during structure-mapping led to the most robust and generalizable, flexible knowledge acquisition in the context of this intervention. Having all visual representations available throughout the lesson may provide students with the necessary working memory supports to attend to key elements in the source and target representations, enabling the child to represent the solutions as systems of relations, map these representations together, and correctly identify elements that are in alignment (or misalignment). Thus, children in the AV condition may have successfully accomplished and benefitted from structure-mapping, while children in the PV and NV conditions may have benefited less from the instructional analogy itself, though both groups did show knowledge acquisition. This may explain the AV students' greater gains in flexible use of strategies
and conceptual knowledge, compared to PV students (on both posttests) and NV students (on immediate posttest).

These data coalesce with results from Rittle-Johnson and Star (2007; Star \& Rittle-Johnson, 2009), who administered measures of procedural flexibility and conceptual knowledge with pre-algebra concepts and found that having pairs of students compare representations simultaneously was more effective than sequentially for students with adequate entry level knowledge of estimation strategies. The current study provides specificity to instructional techniques and supports previous findings with more ecologically valid stimuli, but also provides more detailed data on the role of visual representations of source and target analogs.

The data for the implications of constructing visual representations of analogs but not leaving them visible throughout the lesson are quite different. The least flexible learning derived from the PV condition. This may be because the use of the visual representation did draw learners' attention to the structure of the discussed solution representations, but these learners did not have the resources to move beyond these representations to perform structuremapping and schema generalization.

In contrast, the NV participants may not have had the executive function and working memory resources available for complex structure mapping between representations, they may have also encoded less of the lesson and the first solution (a misconception), may have been less instantiated for them. The delayed data support this interpretation. While the difference in learning gains between AV and PV students remained after one-week delay, this was not so when comparing AV and NV students, highlighting that lack of visual information was less detrimental to overall learning than providing one visual representation at a time.

In fact, students in the Not Visible condition outperformed students in the Part Visible condition in procedural knowledge significantly at a delayed test, and this difference approached significance at immediate posttest. Perhaps, keeping only the latest representation visible on the board may be detrimental for teachers looking to challenge students' misconceptions. Students who see a instantiate a solution modeled on the board as valid, particularly if it is easier, (e.g. subtraction is easier than division), despite teachers' efforts to show it is incorrect.

Previous research suggests that students seek to validate their misconceptions (Chinn and Brewer, 1993) and having the misconception visible, but not throughout the entire lesson in which it was compared to two more accurate solutions, may have helped students in doing that, even more than if it was never visible. Understanding the cognitive processes at play that reconcile these results
warrants further investigation. Outcomes in negative transfer and common misconception provides further support that seeing one problem at a time is detrimental for students attempting to learn by drawing connections between solution strategies.

A reverse trend is apparent for students overextending instructed strategies to problems appearing similar to taught problems, but where the common misconception is the correct solution strategy. The results for the negative transfer measure show that PV students outperform AV students on immediate posttest, which may have been a result of the PV students using the misconception as a correct strategy for all problems that appeared like the problem used in the lesson. This is supported by the fact that PV students also used the common misconception more than AV students on both posttests. Recall that in the video lesson, the common misconception was modeled and was discussed only to be exemplified as an invalid strategy. Challenging misconceptions by modeling and discussing them is common practice in higher achieving countries (e.g. Japan and Hong Kong) and recommended by researchers as a way for students to overcome them (Berry and Graham, 2006; Kuhn, 1989). However, students who did not see the misconception compared with valid solution strategies, despite hearing the same verbalization, may have failed to overcome this challenge and instead may have led them to memorize the misconception as a valid strategy.

Zook (1991) conceptualizes two factors that may be responsible for developing misconceptions from analogies. The first is learner-generated and the second is teachergenerated, either leading to misconceptions, which Zook (1991) defines as: (a) difficulties of the learner attending to key elements increases the potential for misconceptions, and (b) difficulties in leading learners' attention to key elements increases the potential for misconceptions. An interplay of these factors may have negatively affected students in the Part Visible condition in their procedural knowledge, because the valid solutions were not visible throughout the lesson. From the teacher's perspective, there were not sufficient visual cues to support the verbal explanations provided in the instruction, so, from a students' perspective, students had difficulty attending to key mathematical ideas necessary to overcome their misconceptions.

Misconceptions are common throughout the curriculum, and researchers focused on the potential of analogies to overcome these through conceptual change have revealed the real challenges of teaching children to reconsider their misconceptions. For example Chinn and Brewer (1993) provide evidence that many students finish high-school and University without giving up pre-Newtonian perspectives of motion (e.g., Clement, 1982).

Overall, teaching through instructionally supported structure-mapping has the potential to enhance students' conceptual knowledge, procedural flexibility, and procedural knowledge in mathematics. Visual representations can augment these benefits, though it is important to note that the overall advantages in procedural
flexibility in this study were driven by students who saw all the solutions on the board at all times, where students who saw only one solution at a time did most poorly. Strikingly, for procedural knowledge, students who only saw one solution at a time performed worse than students who did not see any solutions throughout the lesson. Thus, students in the Part Visible and Not Visible condition may not have learned by analogy due to insufficient supports.

\section*{Implications for Theory and Practice}

The findings from this study have the potential to positively shape U.S. teaching practices as well as contribute to several areas of cognitive scientific literatures. Utilizing teaching by comparison is critical for learning deep mathematical conceptual knowledge (Rittle-Johnson and Star, 2007, 2009; Star and Rittle-Johnson, 2009; National Mathematics Panel, 2008a, 2009b). Teachers in the U.S. rarely scaffold instructional comparisons adequately (Richland, Zur, and Holyoak, 2007; Heibert et al., 2005), this has been partly due to a lack in specificity in recommendations on how to improve these practices (Hiebert et al., 2005). Recent work that has used cognitive science research in the classroom has provided positive evidence in this direction (Rittle-Johnson and Star, 2007, 2009; Star and Rittle-Johnson, 2009), but even these studies do not examine instructional strategies as they unfold in a real classroom lessons.

The current study uses a novel theoretical perspective and methodological approach that bridges analogy research in laboratory settings with studies of instructional practice in classroom environments. From a theoretical standpoint, these findings support previous laboratory-based results indicating that visual representations can support schema formation and learning from analogy (Gick and Holyoak, 1983), and extend them to an applied setting. The current study provides data on a relevant instructional scaffold that facilitates learners' ability to draw connections between mathematical solution strategies. Comparing representations is common to everyday mathematics instructions, and making all source and target representations visible for the length of the analogy requires only a small time investment and modification of current practice. Thus this intervention is highly feasible to integrate into current teaching practices. Using more ecologically valid stimuli to test teaching practices with the use of a videotaped teacher guided lesson, instead of static written learning materials, thus allows for greater generalizability and specificity for instructional recommendations. Further research that uses these experimental methods is underway, and the authors encourage interdisciplinary researchers to consider the use of video stimuli when designing educational studies.

One must note that we cannot interpret these results to indicate that making analogs visible simultaneously will necessarily lead to successful structure-mapping and mathematical schema formation. Key to improving educational practice is certainly ensuring the instruction uses optimal structured analogs, and ensuring that any misconceptions are identified and compared well with an
alternative and more accurate representation.. Much is still unknown about the ideal combination of support for instructional analogy. At present, further studies are being conducted to examine the impact of the following practices: (a) the teacher's gestures when presenting and linking key ideas, (b) the visual organization of solutions on the board, and (c) the sequencing of chosen solutions (i.e. beginning with a common misconception versus a correct strategy). The first of these two practices (a) and (b) were observed by Richland, Zur, and Holyoak (2007) to correlate with the practices used in our experiment, but they remain to be tested experimentally.

Lastly, from a technology perspective, these findings could have implications for current trends to replace classic chalk or white-boards with smartboards, the highly popular electronic innovations that enable teachers to use their board very actively as a dynamic connection to their computer. While there is the potential for rich activity, there is little room to write, since the smart boards are about a third of typical white boards. These data suggest that smartboards may be highly effective at instantiating single visual representations at a time, much as in our part visible condition, which led to the lowest learning gains, greatest rate of misconceptions, and least flexible knowledge. In summary, instructional attention should be paid to carefully considering the role of visual representations in balancing the benefits for improved encoding of relational structure with ensuring that students align, map, and compare these structured representations to ensure broader generalization, misconception revision, and appropriate schema formation.

\section*{References}

Berry, J. and Graham, T. (1991). Using concept questions in teaching mechanics.International Journal of Mathematical Education in Science and Technology, 22:5, 749-157.
Brown, A. L. (1992) Design Experiments: Theoretical and Methodological Challenges in Creating Complex Interventions in Classroom Settings. The Journal of the Learning Sciences, 2(2), 141-178.
Good, T and Brophy, J. (2003). Looking in classrooms ( \(9^{\text {th }}\) ed.) Boston: Allyn and Bacon.
Chinn, C., \& Brewer, W. (1993). The role of anomalous data in knowledge acquisition; A theoretical framework and implications for science instruction. Review of Educational Research, 63(1), 1-50
Clement, J. (1982). Students' preconceptions in introductory mechanics. American Journal of Physics, 50, 66-71.
Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. Cognitive Science, 7, 155-170.
Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. Cognitive Science, 34, 752-775.
Gick, M.L. and Holyoak, K. J. (1983). Schema induction and analogical transfer. Cognitive Psychology, 15, 1-38.
Good, T and Brophy, J. (2003). Looking in classrooms (9 \(9^{\text {th }}\) ed.) Boston: Allyn and Bacon.

Hiebert, J. Stigler, J. Jacobs, J., Givvin, K. B., et al. (2005); Mathematics Teaching in the United States Today (and Tomorrow): Results From the TIMSS 1999 Video Study Educational Evaluation and Policy Analysis.
Holyoak K. J. (2012). Analogy and relational reasoning. In Oxford handbook of thinking and reasoning, pages 234 259. Oxford University Press.

Morrison, R. G., Krawczyk, D. C., Holyoak, K. J., Hummel, J. E., Chow, T. W., Miller, B. L., et al. (2004). A neurocomputational model of analogical reasoning and its breakdown in Frontotemporal Lobar Degeneration. Journal of Cognitive Neuroscience, 16, 1-1 1
Morrison, R. G., Doumas, L. A. A., and Richland, L. E. (2011) A computational account of children's analogical reasoning: Balancing inhibitory control in working memory and relational representation. Developmental Science, 14, 516-529.
National Mathematics Panel, (2008). Foundations for Success: The Final Report of the National Mathematics Advisory Panel, U.S. Department of Education: Washington, DC.
National Research Council. (2001). Adding it up: Helping children learn mathematics.
Richland, L.E., Zur, O., and Holyoak, K.J. (2007). Cognitive supports for Analogies in the Mathematics Classroom. Science, 316, 1128-1129.
Rittle-Johnson, B, and Star, J.R. (2007). Does comparing solution methods facilitate conceptual and procedural knowledge? An experimental study on learning to solve equations. Journal of Educational Psychology, 99, 561574.

Rittle-Johnson, B., and Star, J. R. (2009). Compared to what? The effects of different comparisons on conceptual knowledge and procedural flexibility for equation solving. Journal of Educational Psychology, 101, 529-544.
Ross, B.H. (1989). Distinguishing types of superficial similarities: Different effects on the access and use of earlier problems. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,456-468.
Star, J. R., and Rittle-Johnson. B. (2009). It pays to compare: An experimental study on computational estimation. Journal of Experimental Child Psychology, 102, no. 4: 408-426.
Stigler, J., and Hiebert, J. (1999). The Teaching Gap. NY: The Free Press, Simon and Schuster, Inc.
Smith J.P., III, diSessa A. A., and Roschelle J. (1994). Misconceptions Reconceived: A Constructivist Analysis of Knowledge in Transition. The Journal of Learning Sciences. 3(2), 115-163.
Waltz, J.A., Lau, A., Grewal, S. K., Holyoak, K.J. (2000). The role of working memory in analogical mapping. Memory and Cognition, 28(7):1205-12.
Zook, K. B. (1991). Effects of analogical processes on learning and misrepresentation. Educational Psychology Review, 3, 41-72.

\title{
Priming and Conceptual Pacts in Overhearers' Adoption of Referring Expressions
}

\begin{abstract}
Current theories of communication yield predictions about the expression choice of overhearers as well as primary discourse participants. We discuss three such theories and evaluate them with reference to new data on object naming elicited through a confederate priming paradigm. Our results show that participants adopt primed referring expressions if they are highly involved in the task, but mere exposure to the object labels yields very limited priming effects. Also, common ground is a relatively marginal factor in expression choice here. We interpret these results as supportive of the importance of grounding and challenging for interactive alignment-based accounts of expression choice.
\end{abstract}

\section*{Introduction}

Inter-personal communication is customarily taken to involve processes of cooperation and coordination between interlocutors at a number of levels. At the level of the conversational turn, speakers cooperate with hearers by making their contribution appropriate to the current purpose of the talk exchange, as observed by Grice (1975). They also coordinate with hearers by making their intentions understood (Grice 1957). At a discourse level, speaker and hearer work together to achieve conversational goals, which might involve the sharing of information, the making and satisfying of requests, the formation of joint plans, etc. (Clark 1996).

A diverse range of theories have been proposed to account for how speakers and listeners successfully engage in this process of communication, with particular reference to dyadic interactions such as dialogues. An influential account of dialogue, the interactive-alignment model (Pickering \& Garrod 2004), places low-level processes of priming at the heart of communication. In this account, interlocutors align their representations as a result of dialogue. This alignment commences at a surface level, in that the dialogue participants converge at a lexical and syntactic level, due to the priming effects exerted by the use of particular words and syntactic forms. The resulting alignment then percolates up through the system, eventually reaching the level of situation models. The goal of communication, on this account, is to accomplish the alignment of situation models. However, the fundamental drivers of this are low-level, automatic and unconscious processes, specifically priming processes. Consequently, this account posits little involvement of strategic factors in the success of dyadic communication.

A contrary viewpoint is that interlocutors are highly aware of each other's mental states and that this awareness informs their behaviour. Clark and Schaefer (1989) argue that successful contributions to a discourse requires grounding; that is, speaker and hearer must mutually believe that the speaker's meaning has been understood. The notion of common ground (CG) - the shared knowledge, beliefs and assumptions of the interlocutors (Clark \& Marshall 1981) - thus becomes relevant here. The goal of the interaction involves building and updating

CG, and doing this requires consideration of the existing CG state. A simple example is the use of a referring expression: if a speaker predicates a new property of an entity (e.g. "John is away"), this can only be successful as a discourse contribution if the hearer correctly identifies the entity. This requires the speaker to take account of the hearer's knowledge about how this entity is labelled. Such knowledge may be presumed on the basis of linguistic community membership, but it may also arise from previous referential success, or on the basis of the formation of "conceptual pacts" (Brennan \& Clark 1996). Distinctively, conceptual pacts involve the establishment of partner-specific labels for entities, which can then be successfully used in interaction with that specific partner but are not preferred for general use with other interlocutors.

An intermediate position between these two viewpoints is occupied by Keysar (2007). He argues that "when people communicate, they do not routinely take into account the mental states of others" (ibid., p.72). Instead, drawing upon evidence from Theory of Mind experiments, he argues for the primacy of egocentric processing, and contends that "one's own perspective is dominant...the consideration of others' beliefs is not automatic" (ibid., p.75). Unlike the interactive-alignment model, this approach entails conscious reasoning about the choice of referring expression, but unlike Clark and colleagues, Keysar considers CG to be a relatively peripheral issue, and the role of the hearer and his/her mental state to be a marginal factor in the speaker's choice of expression. In support of this, Barr and Keysar (2002) provide experimental evidence that people (unconsciously) expect new conversational partners to adhere to conceptual pacts that have previously been established, even though the new partner is not privy to this pact. This in turn suggests that conceptual pacts are not triadic relations between two interlocutors and an entity, in which both agree to refer to this entity in a particular way within their interaction, but rather pairs of relations in which both parties separately agree to refer to this entity by a particular label.

Experimental work on dialogue has been conducted from numerous theoretical perspectives, including those outlined above. However, relatively little attention has been paid in this literature to non-dyadic interactions, for instance those in which a third individual is present but not directly engaged in the conversation \({ }^{1}\). In this paper, we aim to extend findings about the choice of referring expression into the domain of non-dyadic interactions. There are several motivations for this move. First, the potential relevance of third parties in conversation has long been acknowledged (see Clark \& Carlson 1982) but the implications of this for expression choice have attracted relatively little attention. Secondly, such

\footnotetext{
\({ }^{1}\) A partial exception to this is Keysar and Henly (2002), but their primary focus is on using overhearers to evaluate the likely communicative success of utterances, rather than on examining the effect of dialogue on overhearers themselves
}
situations are common in everyday interaction, and understanding the dynamics of conversation in such settings is an end in itself. Thirdly, and perhaps most importantly, non-dyadic interactions represent a testing ground in which the factors governing expression choice can be disentangled to a certain extent, thus offering useful insights as to the relative strengths of the competing factors.

In the following section, we consider how the competing theories of Pickering and Garrod, Clark, and Keysar naturally yield distinctive predictions about the behaviour of overhearers in a non-dyadic setting. We then introduce an experiment to test the effects attributable to priming, egocentricity and conceptual pacts, specifically examining whether and under what conditions overhearers select referring expressions according to their status in a preceding dialogue.

\section*{Critical Predictions about Overhearer Behaviour}

Although the specific accounts discussed in the previous section are primarily oriented towards explaining dyadic interactions, the mechanisms that they posit should apply also in non-dyadic interactions. If so, predictions can be drawn about overhearer behaviour, as we articulate in the following paragraphs. Of course, the falsification of such a prediction would not imply the incorrectness of the theory in the dyadic case. However, it would suggest that additional machinery would need to be posited to cover non-dyadic interactions. We would interpret it as favourable for a theory if it makes correct predictions about both types of interaction without further stipulation.

In particular, we focus on a specific scenario of nondyadic interaction. In this scenario, two interactants are playing a game in which they match picture cards that display tangram figures. Both have matching packs of cards. One of them (the 'director') selects a card and describes the figure, and the other (the 'matcher') has to identify which card is being talked about. There is also an overhearer, who does not participate in the game. When the game is completed, the overhearer plays the game, taking the role of director. The question is whether, and to what extent, the overhearer will re-use the descriptions that were used by the original director in the previous phase of the game.

Turning first to Pickering and Garrod's (2004) account, priming is predicted to occur automatically upon exposure to the relevant labels. They predict stronger alignment effects for addressees than overhearers (ibid., 174), on the basis that the former engage their production systems during the interaction (anticipating that they will speak at some point) whereas the latter do not need to. However, overhearers are still expected to exhibit some priming effects. Crucially, this does not depend upon the establishment of full common ground, which is argued only to occur "when radical misalignment becomes apparent" (ibid., 179). Rather, it relies merely on implicit common ground, defined as the information shared between the interlocutors, to which the overhearer might reasonably be supposed to have access. Hence, in this
experimental paradigm (where there are no observable failures in communication), their account predicts priming of overhearers, a possible effect of involvement, and no effect of common ground. It further predicts that priming will be boosted if the overhearer's production system is activated.

Contrastingly, for Clark and colleagues, high-level conscious processes are critical to determining whether the overhearer adopts the referring expressions that have been used. These expressions should be used only if they have been observed to be successful, which entails that the overhearer is sufficiently engaged in the dialogic process to determine whether this is the case: merely hearing the expressions will not do. In particular, where conceptual pacts have been formed, the status of the addressee with respect to these pacts should also be relevant. When addressing someone who was involved in the conceptual pact (in our scenario, someone involved in the first phase of the game), the former overhearer is predicted to re-use the established referring expression to a greater extent than they would if addressing a new individual. So in brief, this account predicts no priming unless the overhearer is sufficiently involved in the discourse, and more priming when common ground is also present.

The predictions arising from Keysar's (2007) account differ from Clark's with respect to common ground. According to Keysar, the choice of expression should be egocentrically motivated in the first instance, and therefore it should be irrelevant whether or not the hearer has a prior conceptual pact about that referent. Whether the overhearer should adopt any of the expressions used by the previous director is not clear on this account: as Keysar does not posit a role for low-level priming, this should not occur automatically, although it might be feasible for the overhearer to learn new labels under certain conditions (see General Discussion). In sum, we take this account to predict no priming unless the overhearer is sufficiently involved in the discourse, but no effect of common ground.

\section*{Establishing Baseline Naming Probabilities}

In order to establish our baseline naming probabilities, we ran an online elicitation study with EFS Survey (http://www.unipark.info). Participants were asked to provide names for 50 configurations of tangram pieces, which were presented in silhouette. One tangram was presented per page, with the task being to give a name to the presented graphical display (no information was given as to whether the display was a picture or representation of a specific object).

331 participants were recruited via the University's mailing list and leaflets around the campus. All were students and native speakers of German. Participants were entered into a prize draw to win \(€ 10\) cash or one of \(10 € 10\) Amazon vouchers.

For the following experiments, we selected 15 tangrams for each of which a specific response had occurred at rates of \(5-15 \%\). These responses could be considered plausible but dispreferred, in that they were neither unique to an individual respondent nor were they the 'obvious' description for the tangram in question. The
use of such items in the following experiments reduces the probability that the participant selects the target expression just by chance. The mean rate of usage for the relevant descriptions across these 15 tangrams was \(8.29 \%\) (278/3353).

\section*{Experiment 1: Effects of Involvement and Common Ground}

Experiment 1 was designed to show whether an overhearer's involvement in the interaction, and the extent to which they shared common ground with their subsequent addressee, influenced their uptake of dispreferred referring expressions.

\section*{Participants}

86 participants ( 47 female), all native speakers of German, were paid for participation in the experiment. They were divided randomly between the four test conditions.

\section*{Materials}

Three sets of 15 white cards ( \(74 \times 105 \mathrm{~mm}\) ) were used, each with a black tangram on the upper half of the card. The confederate director and the matcher each had one set of cards. Each set of cards showed identical figures: the confederate director's cards also showed the names that were to be used for the tangrams. Video and audio recordings were made of each trial.

\section*{Procedure}

For each condition, the experimental setting comprised an interacting dyad of director and matcher, plus an overhearer. All three individuals were separated by opaque screens. They were instructed that they were to play a game in which the object was to match the order of 15 cards. The director's cards were arranged in a stack in the correct order, while the matcher's cards were arranged on the table top and all were visible. Both were instructed not to change the orientation of the cards.

The director was instructed to proceed by naming the card on top of the pile so that the matcher could find the corresponding card, using names that were as short and spontaneous as possible but as long as necessary. Matchers were allowed to ask for additional descriptions but were told that they would lose points for doing so. These instructions were devised to avoid the use of detailed descriptions rather than impressionistic names for the tangrams.

The experiment proceeded in two phases, using the confederate priming paradigm. In the first phase, the director was a confederate and used pre-specified descriptions (chosen from the pre-test results as discussed above). The matcher was also a confederate, and the experimental participant was the overhearer. In the second phase, the game was played again, with the participant now playing the role of director, but having no direct access to the list of descriptions that had previously been used.

A \(2 \times 2\) design was used, within which levels of involvement and common ground were manipulated. In the high-involvement conditions, overhearers were
presented with a sheet displaying all 15 tangram shapes prior to the first matching phase of the experiment. In the low-involvement conditions, overhearers were not shown the shapes that were being discussed. Instead, they were asked to count the number of times that /t/ was uttered during the interaction (cf. Bavelas, Coates \& Johnson 2000), in order to ensure that they were attending to the linguistic material being uttered. In the high common ground condition, the confederate who was the director in phase 1 of the experiment became the matcher in phase 2 of the experiment, whereas in the low common ground condition, a new confederate who had not participated in phase 1 of the experiment was the matcher in phase 2.

The transcript of phase 2 was analysed in order to establish whether the participant preferentially re-used descriptions that had been used by the confederate in phase 1. The participant's descriptions were considered according to two criteria: a strict lexical priming criterion, in which only identical or similar words (modulo morphosyntactic alternations) were considered to 'match', and a more liberal semantic criterion, in which expressions of similar concepts and synonyms were also considered as matches.

\section*{Results}

The results are summarised in Table 1, for semantic priming, and Table 2, for lexical priming.

Table 1: \% semantic priming effects in Experiment 1
\begin{tabular}{|l|l|l|}
\hline Condition & - CG & + CG \\
\hline - involvement & 14.8 & 15.3 \\
& \((49 / 330)\) & \((46 / 300)\) \\
\hline+ involvement & 41.0 & 50.4 \\
& \((129 / 315)\) & \((174 / 345)\) \\
\hline
\end{tabular}

Table 2: \% lexical priming effects in Experiment 1
\begin{tabular}{|l|l|l|}
\hline Condition & - CG & + CG \\
\hline - involvement & 13.0 & 14.3 \\
& \((43 / 330)\) & \((43 / 300)\) \\
\hline+ involvement & 36.2 & 47.0 \\
& \((114 / 315)\) & \((162 / 345)\) \\
\hline
\end{tabular}

In all conditions, the use of primed expressions was significantly higher than their rates of spontaneous use in the pre-study (binomial, all p < 0.001). We applied a logistic mixed model with full random slopes to the semantic priming results. This showed a highly significant main effect of involvement ( \(\beta=1.88, \mathrm{SE}=0.259, \mathrm{Z}=\) \(7.28, \mathrm{p}<0.001\) ), but the main effect of common ground did not reach significance ( \(\beta=0.255\), \(\mathrm{SE}=0.203, Z=\) \(1.26, \mathrm{p}=0.21\) ), despite the numerical trend in the high involvement condition. In a second model we also posited an interaction term, but this did not reach significance ( \(\beta=\) \(0.58, \mathrm{SE}=0.411, \mathrm{Z}=1.4, \mathrm{p}=0.16\) ), while involvement remained significant and common ground non-significant. This pattern of effects was replicated for the lexical priming results.

\section*{Discussion}

The results of Experiment 1 indicate that the degree of the participants' involvement is highly relevant to their uptake of dispreferred referring expressions. In the conditions in which overhearers were allowed to see the set of tangram figures, they were effective at acquiring the labels used in phase 1 of the experiment. When they were not allowed to see the figures, they exhibited much smaller priming effects, using the primed labels only slightly more frequently than would have been expected in spontaneous, unprimed description.

The presence of common ground led to numerically more frequent reuse of primed descriptions, but this effect did not reach significance in our sample after subject and item effects were taken into consideration.

We interpret these results as potentially supportive of the positions of Clark or Keysar. In particular, it is not the case that overhearers frequently use dispreferred descriptions just as a consequence of having heard these object labels; they must also be aware of the referent picked out by the label. In the terminology of Clark and Brennan (1991), the use of the label must be "grounded". It could of course be argued that the overhearers in our experiment do not have the opportunity to ground the labels with certainty, even in the high involvement conditions, as they cannot be sure which referent is picked out by which expression. Nevertheless, the results suggest that our participants were generally adept in solving this mapping problem, and having done so, used this information to inform their choice of referring expression.

These results can be reconciled with the account of Pickering and Garrod (2004) if we assume that the participant in the high involvement condition is sufficiently engaged in the discourse to have an activated production system, making them effectively a discourse participant rather than merely an overhearer. From that perspective, we could see these results as indicative of the degree of involvement that is required in order for the third individual to be subject to substantial priming effects. On this account, although the priming effect still persists in the absence of the referents (in that primed expressions are used at above-baseline rates), it is very much weakened.

The lack of a strong effect of common ground speaks in favour of the egocentric view proposed by Keysar and colleagues. However, the trend towards greater reuse of priming expression to familiar interlocutors in the high involvement condition suggests that some participants may be influenced by the existence of a prior conceptual pact. If this were the case, it would challenge both the egocentric account and the assumption of Pickering and Garrod (2004) that common ground is only consulted when there is some kind of difficulty in the dialogue, such as deceit or extensive repair. Further work is required to confirm or exclude the existence of this trend.

\section*{Experiment 2: Task-Specific Effects}

A question arising from the first experiment is whether the manipulation of involvement also influenced the participants' expectations about their task. Could it be the case that the participants who were presented with a copy of the tangram pictures inferred that their task was to learn
how to describe these images? This could in turn result in greater activation of their production mechanisms, predicted by Pickering \& Garrod (2004) to lead to greater priming effects. To address these possibilities, we conducted a further experiment in which the overhearers were not given access to pictures of the tangrams, similarly to the original low involvement condition, but were told that after the first part of the experiment, they would then be playing the game, in the role of director.

\section*{Participants}

41 participants (27 female), all native speakers of German, and none of whom participated in Experiment 1, were paid for participation in the experiment. They were assigned to the two new test conditions (common ground and no common ground, as in Experiment 1).

\section*{Materials}

The same materials were used as in Experiment 1.

\section*{Procedure}

The same procedure was used as in the low involvement condition of Experiment 1, with the exception that the participants were not asked to perform t-counting, on the basis that this might interfere with their ability to follow the task (and potentially the engagement of their production systems). Instead, they were instructed to listen to what was going on and told that they would be asked about how successful the interaction had been. Before the experiment began, participants were told that they would be taking the role of director in the second part of the experiment.

\section*{Results}

The results are summarised in Table 3.
Table 3: Results of Experiment 2
\begin{tabular}{|l|l|l|}
\hline Condition & \begin{tabular}{l} 
Semantic \\
priming \%
\end{tabular} & \begin{tabular}{l} 
Lexical \\
priming \%
\end{tabular} \\
\hline- CG & 16.0 & 13.0 \\
& \((48 / 300)\) & \((39 / 300)\) \\
\hline+ CG & 15.9 & 15.2 \\
& \((50 / 315)\) & \((48 / 315)\) \\
\hline
\end{tabular}

In both conditions, the use of primed expressions was significantly higher than their rates of spontaneous use in the pre-study (binomial, both \(\mathrm{p}<0.001\) ). Comparing these results with the low involvement conditions of Experiment 1, logistic regression analyses showed no significant main effect of task awareness.

\section*{Discussion}

The results of Experiment 2 suggest that awareness of the potential usefulness of the descriptions that are employed does not suffice, on its own, to enable the overhearer to pick up dispreferred expressions in this paradigm. Without access to depictions of the referents, the participants in this experiment exhibited very limited evidence of priming effects. This suggests that the higher rates of priming
attested in the high involvement condition of Experiment 1 are largely attributable to the perceptibility of the figures, rather than the participants drawing any specific inferences about the way in which they were expected to perform the task.

\section*{General Discussion and Conclusions}

Our experiments strongly suggest that overhearers are able to acquire dispreferred labels for objects, but that they do so to a very limited degree if they do not have perceptual access to the object that is being referred to.

We take these results to point to limitations in the power of 'pure priming' effects; that is, the view that access to the phonetic content of labels will lead to their adoption by hearers, as a consequence of percolation (Pickering \& Garrod 2004). In the confederate priming paradigm, it appears that such access is not enough: the label must also be associated with an object in order for it later to be adopted. This suggests that the process by which speakers align on object labels is not merely bottom-up, but requires the presence of a referent or meaning as well as the verbal label.

Nevertheless, it could be argued that our results do point to non-zero priming effects, with increased uptake of primed expressions even among uninvolved overhearers who do not see the potential referents and are not attending to the dialogic process that is occurring. Such effects could indeed be attributable to the type of processes that Pickering and Garrod (2004) posit. However, at least in this paradigm, these effects are much smaller than the priming effects in the high involvement condition.

A possible explanation of the effect of involvement, within the Pickering and Garrod account, is that the overhearers' production mechanisms are more highly activated in the high involvement condition. Given the results of Experiment 2, we consider this unlikely to be the sole cause of the involvement effect. The results of experiment 2 suggest that, even when participants are explicitly informed that they will later be called upon to describe the same figures, and hence might be assumed to engage their production systems in preparation to participate in a dialogue, they do not exhibit greater uptake of the primed expressions.

Of course, it may be the case that the effect of involvement is a matter of attention, and that overhearers in the low involvement condition are less engaged in the task in general. However, in experiment 1, these overhearers are obliged to attend to the phonetic content of the utterances, which should in principle be sufficient to initiate priming effects via percolation. This explanation might be tenable if we modify the 'pure priming' account to require that lexical items must be heard and understood in their entirety in order to be primed.

The strong effect of involvement is straightforwardly explicable in Clark's approach: according to this view, expressions are re-used as a result of their observable effectiveness in the prior interaction, and it is the highly involved participants who are in a position to discern this. However, it can also be captured by the egocentric approach, articulated by Keysar and colleagues. Here we must also posit that the condition of high involvement -
which presented participants with a visual representation of the tangram figures being talked about - enabled the overhearers to learn the referring expressions corresponding to (some of) these figures. Given that there were no 'right' or 'wrong' answers in our experiment, it is perhaps slightly counter-intuitive that 'egocentric' overhearers should bother to learn the names of tangrams, when they could simply describe them as they saw fit. It is possible that learning the names in this way represents an economical strategy that obviates the need for any decisions about how to describe the tangrams later on (although participants in Experiment 1 were not told that they would need to do so). We cannot, therefore, exclude the possibility that hearing the primed names in the high involvement condition merely shaped the egocentric preferences of the overhearers, and that this was later manifested in their choice of expression.

Our experiments documented a numerical tendency towards common ground effects, but this might be attributable to random variation. If this effect is replicated in further research, it would more seriously challenge Keysar's (2007) claim that the speaker's choice of referent should initially be egocentrically motivated, irrespective of conceptual pacts. Note that, in this experiment, there were no failures of communication (as the matcher was a confederate), hence there was no need for the director to reformulate his or her utterance: purely egocentric behaviour would, to all intents and purposes, have done just as well. The preferential reuse of primed expressions when the matcher was familiar would suggest that awareness of conceptual pacts may, at least for some speakers and on some occasions, be influencing the initial choice of utterance.

In short, our results so far do not permit us to exclude the possibility of egocentrism on the part of our participants, and can be reconciled with a slightly modified version of the form-based priming account of Pickering and Garrod (2004). Nevertheless, the results appear to fit most naturally with the viewpoint articulated by Clark and colleagues. Specifically, in order for expressions to be adopted, it appears to be broadly necessary for overhearers to understand the purpose of the expressions or to experience them being used effectively (our experiments do not distinguish these possibilities). Crucially, it is not sufficient merely to hear expressions that are not grounded, even if you know that you will be called upon to produce similar expressions in the future. However, two aspects of these results are unexpected from the perspective of Clark's approach: firstly, as discussed above, the effect of conceptual pacts is, at most, marginal in these experiments. A definitive absence of such effects would speak in favour of Keysar's view. Secondly, expressions can (occasionally) be picked up spontaneously by uninvolved overhearers without awareness of the current discourse goals, as predicted by the model of Pickering and Garrod. Further investigations might show whether the apparent examples of this in our data are actually attributable to the overhearer attending to the discourse, or whether they should be treated as genuine instances of automatic priming effects that are not predicted by Clark's theory.

\section*{References}

Barr, D. J., \& Keysar, B. (2002). Anchoring comprehension in linguistic precedents. Journal of Memory and Language, 46, 391-418.
Bavelas, J. B., Coates, L., \& Johnson, T. (2000). Listeners as co-narrators. Journal of Personality and Social Psychology, 79, 941-952.
Brennan, S. E., \& Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. Journal of Experimental Psychology: Learning, Memory and Cognition, 22, 1482-1493.
Clark, H. H., \& Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, \& S. D. Teasley (eds.), Perspectives on Socially Shared Cognition. Washington DC: APA Books.
Clark, H. H., \& Carlson, T. B. (1982). Hearers and speech acts. Language, 58, 332-373.
Clark, H. H., \& Marshall. C. (1981). Definite reference and mutual knowledge. In A. K. Joshi, B. L. Webber, \& I. A. Sag (eds.), Elements of Discourse Understanding. New York: Cambridge University Press.
Clark, H. H., \& Schaefer, E. F. (1989). Contributing to discourse. Cognitive Science, 13, 259-294.
Grice, H. P. (1957). Meaning. Philosophical Review, 66, 377-388.
Grice, H.P. (1975). Logic and conversation. In P. Cole \& J. Morgan (eds.), Syntax and Semantics, vol. 3. New York: Academic Press.
Keysar, B. (2007). Communication and miscommunication: The role of egocentric processes. Intercultural Pragmatics, 4, 71-84.
Keysar, B., \& Henly, A. S. (2002). Speakers' overestimation of their effectiveness. Psychological Science, 13, 207-212.
Pickering, M. J., \& Garrod, S. (2004). Toward a mechanistic psychology of dialogue. Behavioral and Brain Sciences, 27, 169-226.

\title{
Gaze strategies in object identification and manipulation
}

\author{
Anna Belardinelli (belardinelli@informatik.uni-tuebingen.de) \\ Department of Computer Science, University of Tübingen,
}

\author{
Martin V. Butz (martin.butz@uni-tuebingen.de) \\ Department of Computer Science, Department of Psychology, University of Tübingen, Sand 14, Tübingen, 72076 Germany
}

\begin{abstract}
Task influence has long been known to play a major role in the way our eyes scan a scene. Interestingly, how the task modulates attention when interacting with objects has been less investigated. Only few studies have contrasted the distribution of eye fixations during viewing and grasping objects. How is attention differently deployed when different actions have to be planned on objects in contrast to a purely perceptual viewing condition? To investigate these issues, we conducted an eyetracking experiment showing participants 2D images of realworld objects. In blocks of trials, participants were asked either to assign the displayed objects to one of two classes (classification task), to mimic lifting the object (lifting task), or to mimic opening the object (opening task). Mean fixation locations and attention heatmaps show different modes in gaze distribution around task-relevant locations, in accordance with previous literature. Reaction times, measured by button release in the manual response, suggest that the more demanding the task in terms of motor planning the longer the latency in movement initiation. Results show that even on simplified, two dimensional displays the eyes reveal the current intentions of the participants. Moreover, the results suggest elaborate cognitive processes at work and confirm anticipatory behavioral control. We conclude with suggesting that the strongly predictive information contained in eye movements data may be used for advanced, highly intuitive, user-friendly brain-computer interfaces.
\end{abstract}

Keywords: Eye-tracking, object interaction, fixation distribution, eye-hand coordination, movement preparation

\section*{Introduction}

Since the early works of Buswell (1935) and Yarbus (1967) top-down, task-related guidance has been shown to strongly influence the way people move their gaze upon pictures. In the second study, depending on the question asked, different patterns of scanning were observed. Such an influence is so critical that, as soon as a specific task is given, low-level, bottom-up visual saliency is basically overridden and plays quite a minor role in explaining eye fixations w.r.t. higherlevel cognitive factors (Henderson, Brockmole, Castelhano, \& Mack, 2007; Einhäuser, Rutishauser, \& Koch, 2008). Similarly, moving from pictures to real-world scenes and to tasks involving motor actions, it is even more striking how eye movements are precisely planned to provide information for the execution of the current piece of action. This has been shown in different settings, from tea-making (Land, Mennie, \& Rusted, 1999) to sandwich-making (Hayhoe, Shrivastava, Mruczek, \& Pelz, 2003) to a wealth of other more or less complex motor tasks (Land \& Tatler, 2009). In this case, anyway, the nature of attention deployment is quite different. The purpose of vision is here indeed less to get sense of the scene and more to direct effectors and coordinate a much slower
and more complex behaviour than scanning. Strategies like 'look-ahead' and 'just-in-time' fixations (Hayhoe et al., 2003; Ballard, Hayhoe, \& Pelz, 1995) support the idea that vision is deeply intertwined with the needs of motion planning and supervising.

Further, in the context of the theory on the duplex nature of vision (Goodale \& Milner, 1992), distinct neural pathways subserving the different functional demands of object categorization and object manipulation were suggested. This dissociation between vision-for-action and vision-for-perception has often been investigated by means of grasping tasks contrasted to perceptual judgement tasks, with visual illusions or in covert attention settings (Goodale, 2011), but contrasting evidence has emerged and it seems reasonable to assume a strict interaction between the two systems.

How the differences between perceptual and motor task are reflected in eye-movements has been less investigated. In a seminal paper for eye-hand coordination, Johansson, Westling, Backstrom, and Flanagan (2001) recorded both eye- and hand movements data during a motor task involving grasping a bar, avoiding an obstacle, touching a goal position and placing the bar back. Subjects almost exclusively fixated landmark positions on the bar or in the experimental set-up, before making contact to them. The preparation of an action upon an object defines an attentional landscape (Baldauf \& Deubel, 2010), (covertly) encoding in parallel locations relevant for the subsequent serial motor execution.

This evidence suggests that visual cues are sought and weighted differently depending if the task is a skilled movement or a perceptual judgement. Gaze behaviour in viewing and grasping was investigated by (Brouwer, Franz, \& Gegenfurtner, 2009) and (Desanghere \& Marotta, 2011). The first ones used simple geometric shapes to be simply viewed or grasped, while in the latter study Efron blocks were used and in the viewing condition a perceptual judgement had to be made. In both cases, the viewing condition produced first fixations closer to the center-of-gravity (COG) of the object (in accordance with (Foulsham \& Underwood, 2009), among others), while the grasping condition was characterized by first fixations closer to the index finger location (or to the more difficult to grasp location).

In this paper, we present an experiment building on that of Brouwer et al. (2009). The main novelty of our approach is the use of real object stimuli (displayed on a monitor) and the comparison of three simple but realistic tasks, one 'passive' (classification) and two 'active' (lifting and opening).

We were interested in investigating to what extent eye movements subserve and anticipate the task demands, in the form of information collection for movement planning, and the relation to affordances (Gibson, 1979). This relation was expected to show in different scanning strategies determined by the different landmarks associated to each task. Even though the interaction with real objects in our daily life heavily relies on depth perception, Westwood, Danckert, Servos, and Goodale (2002) showed how subjects can effectively program actions to 2D pictures, suggesting that the dorsal stream does not critically rely on binocular information for prehension movements (see also (Kwok \& Braddick, 2003)). This turned out to be the case in this study, where indeed familiar objects were used and the scanning patterns were similar to those described for real objects.

\section*{Experiments}

We conducted a main eye-tracking experiment and a parallel experiment aimed at extracting Regions Of Interest (ROI) from every stimulus in every condition. This was done to have an objective measure of the contact point regions that would be chosen for an actual grasp instead of arbitrarily choosing some expected ROIs. Both experiments are detailed in the following subsections.

\section*{Participants}

Eleven participants ( 6 women, 5 men, aged 22-41) carried out the eye-tracking experiment in all 3 conditions (task). One female participant's data was discarded because of very bad quality. All subjects were right-handed with corrected to normal vision. Ten different ( 4 men, 6 women, aged 18-41) participants carried out the ROI extraction experiment. All of them were confirmed right-handed. In both experiments participants were compensated with study credits or money.

\section*{Stimulus material}

Stimuli were chosen from the ALOI dataset (Geusebroek, Burghouts, \& Smeulders, 2005), containing pictures of 1000 daily-use objects in different light/view conditions. 14 objects (plus 2 test objects) were chosen such that all of them could be easily lifted and had an opening part. They are all portrayed in a frontal view against a black background. Six objects are displayed upright, six lie horizontally with the opening part on the right. Two objects present a handle on the right and the opening on the top. All 14 stimuli are showed in Fig.1. Each picture is \(768 \times 576\) pixels. In each condition they were presented at mid-height on the right of the screen.

\section*{Apparatus and Procedure}

Participants sat in front of the screen, where the object stimuli were presented. In the eye-tracking experiment their head was resting on a chin rest, about 70 cm away from the monitor, \(1680 \times 1050\) pixels, subtending \(45.3^{\circ} \times 28.3^{\circ}\) of field of view. Stimulus pictures subtended \(20.7^{\circ}\), with the center of the picture lying at \(12.3^{\circ}\) from the center of the monitor.


Figure 1: Stimuli pictures used in the experiment.

Eye movements were recorded via a binocular remote eyetracker (EyeFollower, LC Technologies) working at 120 Hz . A keyboard was placed between the chin rest and the monitor to record reaction times. Participants had to look at the same stimuli with three different tasks in mind - each in one block. The task order was randomized across participants, so was the stimulus order within each block. For each task, every object was presented five times, resulting in 210 trials per participant. For training purposes, 30 more test trials were conducted on 2 other objects before the main experiment.

In the classify task, participants were asked to look at the presented object and to decide whether it could contain liquid or not. The response was given by a left/right arrow key press. This served the purpose of both having participants looking at the objects each time and making a manual response as in the other conditions. In the lift condition, participants had to reach to the screen and to mimic lifting the presented object in front of the screen. Analogously, in the open condition, they reached to the screen and mimicked opening the object. They were instructed to use only the right hand. To grasp objects, they were asked to always perform a grasp frontally, either with the thumb rightwards or downwards or by the handle, where present. As to the opening, they were told to imagine that the objects were glued to the shelf so they could open them with just one hand. They were asked to execute the movement as naturally as possible and to act on the object according to the perceived size \({ }^{1}\). In each trial, participants were asked to press the spacebar until they were ready to execute the proper response. Each trial proceeded as follows: 1) the task (classify/lift/open) is displayed as a reminder at the center of the screen for \(1.5 \mathrm{~s} ; 2\) ) the fixation cross is presented for at least 1 s (or as long as the space bar is not pressed); 3)

\footnotetext{
\({ }^{1}\) The displayed object stimuli were all of the same size, so that objects were presented larger or smaller than they typically are in reality. However, this scaling was not excessively pronounced so that the action to perform was still plausibly and naturally performable.
}
the stimulus appears on the right side of the screen; 4) Phase A: eye data and reaction times are collected up to the release of the space bar; 5) Phase B: eye data collected during the execution of the motor response; 6) the hand goes back to the spacebar and the next trial starts.

In the ROI extraction experiment, the same objects were presented to different participants. In just 2 blocks (lifting and opening), they were asked to place the tips of their fingers on the object, picturing the requested action. These points were recorded via a touch screen. After each trial, the participant was shown the selected points and, if not satisfied, she could repeat the trial. Every object was presented 3 times per block, resulting in 84 trials total per participant.

\section*{Data Processing and Analysis}

Fixations for the phase A and B were extracted for each trial via the dispersion algorithm (Salvucci \& Goldberg, 2000) with a temporal threshold of 100 ms and a spatial dispersion threshold of \(1.5^{\circ}\). Data collected during phase A are supposed to be indicative of the information extraction and motor planning preceding movement initiation. Still, since in many cases there was just one or even no phase A fixation on the stimulus, quantitative evaluations were done on the first 3 fixations (or up to the third fixation) and on the mean of these first three fixations. This choice was motivated by the consideration that 3 fixations amount to about 1 s of stimulus presentation, sufficient to retrieve necessary visual information and start the movement (according to reaction times), while later fixations could be more arbitrary and dependent on the subjects' preference and interest for the object. For qualitative evaluation and informative visualization, heatmaps were computed from fixation data. These were obtained by placing a Gaussian with \(\sigma=1^{\circ}\), centered on each fixation and height proportional to the duration of the fixation, so that longer fixations would be weighted more in the heatmap surface. Each map was scaled between 0 (not fixated) and 1 (longest fixated) to make maps comparable. Regions of interest were extracted considering the distribution of the finger points in each condition. In the 'open' condition, points were compactly concentrated around the opening region, hence mean and variance of the point coordinates sufficed to identify a rectangle containing the underlying region. In the case of 'lift', points were more evidently multi-modal, resulting in two major clusters one, smaller, for the thumb and one for the rest of the fingers. To include both clusters in the ROI, points were clustered via k -means, and a rectangle containing the region underlying both clusters was identified (see Fig.6, left, for an example of extracted ROIs). In most objects the two ROIs were well-separated. In a few cases, they were slightly overlapping and just in one case there was a major overlap. This, nevertheless, did not hamper the comparison with the heatmaps.

\section*{Results}

\section*{Heatmaps}

As a first qualitative impression of the general patterns of behavior observed in the three examined conditions, we compared heatmaps obtained from fixations of phase A, from first 3 fixations (in total and separated) and for the mean of the first 3 fixations. The same pattern was shown at different extents across all maps and objects, namely a maximum left of the object center in the 'classify' condition, a slightly higherleft of the center maximum in the 'lift' condition and a clear maximum on the opening region in the 'open' condition. Fig. 2 shows the phase A maps for one of the up-right objects and one of the horizontal objects. Already in phase A, taskdependent differences in eye fixations are evident.


Figure 2: Heatmaps of the phase A fixations superimposed on corresponding stimuli. From left to right: 'classify', 'lift' and 'open' condition.


Figure 3: Heatmaps of the first, second and third fixation (left to right). From top to bottom: 'classify','lift' and 'open' condition.

An evolution in time across the first 3 fixations/conditions for one object is presented in Fig.3. If the first fixation is usually close to the COG (with some undershoot) for all conditions, already by the second fixation is possible to infer where the scanpath will lead. The first fixation was a 'phase A' fixation in \(90 \%\) of cases, the second fixation in \(53 \%\), while the third just in \(28 \%\). Of 5733 examined fixations, 3359 were
phase A. While for the first fixation phase A fixations are equally distributed across tasks (1832 A fixations, \(34 \%\) classify, \(32 \%\) lift, \(34 \%\) open), in the second the proportion is in favor of lifting and opening (1037 A fixations, \(24 \%\) classify, \(33 \%\) lift, \(43 \%\) open), by the few third A fixations mostly for the 'active' tasks motion had not yet initiated (490 A fixations, \(17 \%\) classify, \(38 \%\) lift, \(45 \%\) open).

\section*{Average Fixations}

The mean of the first three fixations (or up to 3 ) on each stimulus image was extracted for each trial. Often the first fixation was in the direction of the COG of the object but landed either on the black background or on the edge of the object, hence showing some undershoot along the x -axis (we use image coordinates since the objects are not shown in a completely frontal view but in perspective, hence the center of the object outline would not correspond to the COG). A repeated measures ANOVA on the \(x\) coordinate of the average fixation with task and object as within-subject factors showed a main effect of task \((F(2,18)=36.9, p<.001)\), a main effect of object \((F(13,117)=19.87, p<.001)\), and and interaction effect of object and task \((F(26,234)=13.73, p<.001)\). The mean X coordinates according to object and task are presented in Fig. 4. For most objects, the 'classify' mean position was the most left and the 'opening' the most right. This is of course more extreme for horizontal objects, e.g, the gel tube, the white jar, the juice bottle, while for three up-right objects (yellow tea pot, orange tea pot, and chips tube) the lifting mean position is to the right of the opening position either because the handle was on the right or the plastic lid was best opened by exerting force with the right thumb.


Figure 4: Mean X coordinate of each object across task.
An analogous analysis was performed on the vertical mean location. Again the effect of task was significant \((F(2,18)=\) \(51.58, p<.001)\) as that of object \((F(13,117)=134.02\), \(p<.001)\) and interaction \((F(26,234)=28.13, p<.001)\).

The mean Y coordinates according to object and task are presented in Fig. 5. In this case the ordinate is expressed in image coordinates, with origin in the top left corner. Up-right objects (such as the green can or the chips tube) present of course the most extreme mean vertical location for the 'open' task, while for horizontal objects the mean y location is always at the same height with a slight tendency upwards in the 'lift' condition.


Figure 5: Mean Y coordinate of each object across task. Note that the \(y\) axis is in picture coordinates, hence the lower the value the higher the location in the picture.

\section*{Comparison Heatmaps-ROI}

To gain a more specific insight regarding to what extent the fixation map can predict the region on which the motor action is performed, we compared the ROIs extracted for the two 'active' conditions with the peak of the corresponding heatmaps achieved considering the first three fixations (see Fig.6). The peak of the fixation map (where the map has value 1) consistently falls within the corresponding ROI. The mean distance between the peak and the center of the ROI for the 'lift' condition was \(91.1 \pm 59.52\) pixel, while for the 'open' condition was \(63.2172 \pm 35.53\). In both conditions the distance between the peak and the center of the corresponding ROI was always smaller than that to the center of the other ROI (one-tailed t-test, \(p<.001\) ).

\section*{Reaction Times}

Mean reaction times in releasing the spacebar significantly increase from 'classify' to 'lift' to the 'open' condition. The difference is most pronounced between 'passive' and 'active' conditions (classify: \(0.596 \pm 0.052 \mathrm{~s}\), lift: \(0.805 \pm 0.126 \mathrm{~s}\), open: \(0.826 \pm 0.110 \mathrm{~s})\). A repeated measures ANOVA on the average reaction time with task and object as withinsubject factors showed a main effect of task \((F(2,18)=7.04\), \(p=0.006)\) and a main effect of object \((F(13,117)=2.14\),


Figure 6: Left: touched points and Regions Of Interest extracted for one of the stimuli (green: 'lift' condition; magenta: the 'open condition'). Center: heatmap of the first 3 fixations in the 'lift' condition (in green the center of the corresponding ROI). Right: heatmap of the first 3 fixations in the 'open' condition (in green the center of the corresponding ROI).
\(p=0.016\) ). The mean reaction times for object and task are presented in Fig.7. Three objects (spice bottle, basket, and yellow tea pot) obtained shorter reaction times for opening than for lifting, in contrast to the general pattern - possibly because of the size difference compared to the real object, which made the decision on how to lift the object more difficult, and because of the particularly obvious opening action for all three objects. It must be noted that longer reaction times in the active tasks may be due not only to motion planning and affording points selection but also to the extraction of 3D information in absence of disparity cues.


Figure 7: Mean reaction times of each object across task.

\section*{General Discussion}

The presented experiment was aimed at assessing different eye movement strategies employed in identifying an object in contrast to tasks in which actual interactions had to be performed on the object. The distinct tasks as well as the objectspecific affordance points were expected to strongly influence the distribution of eye fixations on each object. Indeed, we found significant differences in the scanpath behavior in the 3 conditions, suggesting for each one the construction of a spe-
cific attentional landscape around the informative/affording points.

In the classification task, the mean position of the first three fixations was mostly in the direction of the COG of the object. When grasping an object to lift it, fixations concentrated on a position to the left of and slightly higher than the COG. On the one hand, it seems reasonable that instead of fixating both contact points in an alternate fashion, fixating near the center of the object allows both contact points to be in the fovea and para-fovea, as suggested in (Desanghere \& Marotta, 2011). On the other hand, for up-right objects a preference to fixate more on the side of the thumb could be observed, while horizontal objects were on average fixated closer to the rest of the fingers. In the case of the two objects with a handle, there was a smaller peak in the center of the object (suggesting a first brief fixation there) and a higher mode on the handle, where later, longer fixations concentrated. In both cases it is possible that due to the objects' reduced size, subjects first considered lifting them with a power grasp and then went for the handle. In the case of opening, the fixation distribution presented a clear peak well localized on the opening region, which required the most processing for the planning of the finer motor operation (usually performed with a precision grip). Even if the overall distribution of fixations is already indicative, the different patterns in the unfolding of the scanpath are best appreciable when looking at the temporal evolution of the first three fixations. The distributions of the first fixation is hardly distinguishable across tasks, but already by the second fixation (at which point the reaching movement often had not been initiated, yet) the task 'signature' became evident.

These results confirm the general predictive nature of eye movements. Beyond that, however, our data indicate that tracking eye movements may be exploited in even more subtle ways, inferring the exact intention of how a user may interact with an object. Such discriminability of eye scanpaths according to the intended interaction goal may substantially help in devising machine learning algorithms to timely infer the intention of impaired patients and possibly inform assistive interfaces to control prosthetic devices without the need of cumbersome training. The reliability with which the fixation mode consistently fell within the specific ROI supports considerations.

It seems plausible that the general flow of processing is first concerned with locating the object of interest (first fixation close to the COG). Next, it moves towards the most informative points - either for decision making in the case of the classification task, or for the purpose of executing anticipatory behavior control (Hoffmann, 2003; Butz, Sigaud, Pezzulo, \& Baldassarre, 2007) towards interaction-relevant points (for lifting/opening) with proper behavioral interaction routines. In the former case, just the ventral system would be involved, pooling resources for recognition and decision-making. In the latter, 'active' conditions, also the dorsal pathway and premotor cortical regions would be substantially involved. After object localization and recognition, object-relative behavior needs to be planned, which involves reference-frame transformations of position, size, and shape and planning of reaching and grasping motions with properly aligned hand shapes (Jeannerod, Arbib, Rizzolatti, \& Sakata, 1995; Cisek, 2007; Herbort \& Butz, 2011). The consequentially more elaborate motion planning is also confirmed by significantly longer reaction times when an active motor task, different for every object, has to be planned anew.

In conclusion, as for more complex behavior, even for single actions to be performed within the same object, the eyes extract visual information in a goal-oriented, anticipatory fashion, incrementally revealing the interaction intentions.

\section*{References}

Baldauf, D., \& Deubel, H. (2010). Attentional landscapes in reaching and grasping. Vision Research, 50(11), 9991013.

Ballard, D. H., Hayhoe, M. M., \& Pelz, J. B. (1995). Memory representations in natural tasks. J. Cognitive Neuroscience, 7(1), 66-80.
Brouwer, A.-M., Franz, V. H., \& Gegenfurtner, K. R. (2009). Differences in fixations between grasping and viewing objects. Journal of Vision, 9(1).
Buswell, G. T. (1935). How People Look at Pictures. Chicago: University of Chicago Press.
Butz, M. V., Sigaud, O., Pezzulo, G., \& Baldassarre, G. (Eds.). (2007). Anticipatory behavior in adaptive learning systems: From brains to individual and social behavior (LNAI 4520). Springer-Verlag.
Cisek, P. (2007). Cortical mechanisms of action selection: The affordance competition hypothesis. Philosophical Transactions of the Royal Society B: Biological Sciences, 362(1485), 1585-1599.
Desanghere, L., \& Marotta, J. (2011). "graspability" of objects affects gaze patterns during perception and action tasks. Experimental Brain Research, 1-11.
Einhäuser, W., Rutishauser, U., \& Koch, C. (2008). Taskdemands can immediately reverse the effects of sensorydriven saliency in complex visual stimuli. Journal of Vision, 8(2).
Foulsham, T., \& Underwood, G. (2009). Does conspicu-
ity enhance distraction? saliency and eye landing position when searching for objects. Quarterly journal of experimental psychology, 62(6), 1088-1098.
Geusebroek, J.-M., Burghouts, G. J., \& Smeulders, A. W. M. (2005). The amsterdam library of object images. Int. J. Comput. Vision, 61(1), 103-112.
Gibson, J. J. (1979). The ecological approach to visual perception. (Houghton Mifflin)
Goodale, M. A. (2011). Transforming vision into action. Vision Research, 51(13), 1567-1587.
Goodale, M. A., \& Milner, A. D. (1992). Separate visual pathways for perception and action. Trends in neurosciences, 15(1), 20-25.
Hayhoe, M. M., Shrivastava, A., Mruczek, R., \& Pelz, J. B. (2003). Visual memory and motor planning in a natural task. J Vis, 3(1), 49-63.
Henderson, J. M., Brockmole, J. R., Castelhano, M. S., \& Mack, M. (2007). Visual saliency does not account for eye movements during visual search in Real-World scenes. In R. van Gompel, M. Fischer, W. Murray, \& R. Hill (Eds.), Eye movement research: Insights into mind and brain. Elsevier.
Herbort, O., \& Butz, M. V. (2011). Habitual and goal-directed factors in (everyday) object handling. Experimental Brain Research, 213, 371-382.
Hoffmann, J. (2003). Anticipatory behavioral control. In M. V. Butz, O. Sigaud, \& P. Gérard (Eds.), Anticipatory behavior in adaptive learning systems: Foundations, theories, and systems (p. 44-65). Springer-Verlag.
Jeannerod, M., Arbib, M. A., Rizzolatti, G., \& Sakata, H. (1995). Grasping objects: the cortical mechanisms of visuomotor transformation. Trends in neurosciences, 18(7), 314-320.
Johansson, R. S., Westling, G., Backstrom, A., \& Flanagan, J. R. (2001). Eye-Hand coordination in object manipulation. J. Neurosci., 21(17), 6917-6932.
Kwok, R., \& Braddick, O. (2003). When does the titchener circles illusion exert an effect on grasping? two- and threedimensional targets. Neuropsychologia, 41(8), 932-40.
Land, M., Mennie, N., \& Rusted, J. (1999). The roles of vision and eye movements in the control of activities of daily living. Perception, 28(11), 1311-1328.
Land, M., \& Tatler, B. (2009). Looking and acting vision and eye movements in natural behaviour. Oxford University Press.
Salvucci, D., \& Goldberg, J. (2000). Identifying fixations and saccades in eye-tracking protocols. Proc. of the 2000 symposium on Eye tracking research \& applications, 7178.

Westwood, D. A., Danckert, J., Servos, P., \& Goodale, M. (2002). Grasping two-dimensional images and threedimensional objects in visual-form agnosia. Experimental Brain Research, 144, 262-267.
Yarbus, A. L. (1967). Eye movements and vision (1st ed.). Plenum Press. Hardcover.

\title{
The Role of Achievement Goal Motivation in Self-Explanation and Knowledge Transfer
}

\author{
Daniel M. Belenky (dbelenky@andrew.cmu.edu) \\ Human-Computer Interaction Institute, 2602H Newell Simon Hall, 5000 Forbes Avenue \\ Pittsburgh, PA 15213 USA \\ Timothy J. Nokes-Malach (nokes@pitt.edu) \\ Department of Psychology, 818 Learning Research and Development Center, 3939 O'Hara St. Pittsburgh, PA 15260 USA
}

\begin{abstract}
Self-explanation is an important constructive cognitive process that helps students learn in such a way that they can flexibly transfer their knowledge to solve novel problems (Chi, Bassok, Lewis, Reimann, \& Glaser, 1989). However, research has not addressed what leads students to spontaneously self-explain, in the absence of prompting. The present study experimentally manipulates student motivation (in terms of achievement goals) and measures what influence this has on self-explanation and transfer. Participants ( \(N=\) 140) received goal framings that reflected either a masteryapproach goal (striving to develop one's understanding), a performance-approach goal (an aim to outperform others), a performance-avoidance goal (avoid doing worse than others) or a no-goal control. This framing was applied to a set of learning and test tasks on basic statistics, which participants completed while thinking aloud. Results showed a benefit for a performance-avoidance condition in terms of both higher levels of self-explanation and transfer. This unexpected result is discussed in terms of theories of motivation and learning, and their potential impact on educational practice.
\end{abstract}

Keywords: Knowledge Transfer, Motivation, Achievement Goals, Self-Explanation, Education

\section*{Introduction}

A fundamental goal of instruction is to foster learning which leads to successful, flexible, and useful knowledge transfer. Research and theoretical development which elucidates how knowledge transfers has a long history in psychology and cognitive science, and continues to be important for educational psychologists and learning scientists. Continuing to advance our understanding of what sorts of learning activities lead to transfer allows for recommendations on how to improve educational practices.

Evidence has accumulated that a promising method for promoting flexible knowledge transfer is to increase the conceptual quality of the original learning (Pashler et al., 2007). As such, constructive learning processes which promote the acquisition of more abstracted knowledge (e.g., schemas) are likely to promote successful knowledge transfer. A representative example of such a process is selfexplanation (Chi, Bassok, Lewis, Reimann, \& Glaser, 1989), which is the process by which students generate, for themselves, explanations which go beyond the text, inferring underlying principles and highlighting important interrelations. Chi et al. (1989) documented a large
difference in the amount and quality of self-explanations between those students who ultimately go on to flexibly transfer their knowledge and those who do not. A number of studies since have documented that students can be prompted to engage in self-explanation (e.g. Aleven \& Koedinger, 2002) with beneficial effects for learning and knowledge transfer. However, a fundamental question about self-explanation has been left unaddressed; what leads students to engage in self-explanation, in the absence of prompting? It is clear that some students do so, to their benefit, while others do not. It is also likely that whether students are capable of self-explaining profitably is not the sole limitation, given the experimental literature which shows a benefit for self-explanation prompts. In the present research, we address the possibility that student motivation leads to the spontaneous use of self-explanation during learning, and that this can influence the likelihood of successfully transferring. While Chi et al. (1989) conjectured that higher levels of self-explanation are "a natural consequence of wanting to understand the solution example better" (pg. 160), no research has systematically tested this claim. As we will review in the subsequent section, this sort of motivation has been studied extensively by researchers of "achievement goal theory," which has documented just such a desire, labeling it a "masteryapproach" goal. The present study leverages achievement goals as a tool for experimentally investigating how motivation influences self-explanation and transfer.

\section*{Achievement Goal Motivation}

Achievement goals are the reasons people have for engaging in achievement settings, such as school or work. An achievement setting is one that is organized around one's competence in a domain, and an achievement goal describes a goal a person has in relation to this competence, such as wanting to use it demonstrate how good they are in this domain, or wanting to develop their competence so that they can complete more challenging work. A large body of research and theory development has led to a generallyaccepted framework which proposes three main classes of goals (Elliot, McGregor, \& Gable, 1999); mastery-approach, performance-approach, and performance-avoidance. A mastery-approach goal is an aim to improve or develop one's competence or understanding. A performance-
approach goal reflects striving to demonstrate one's competence by doing better than one's peers, while a performance-avoidance goal occurs when one strives not to demonstrate one's incompetence, compared to peers \({ }^{1}\).

Research has shown that these goals produce a characteristic pattern of effects on various learning behaviors, affective states, and measures of performance. Specifically, performance-avoidance goals tend to be associated with negative outcomes, such as lower performance and worse study strategies (e.g., Elliot, Shell, Henry \& Maier, 2005, Elliot, McGregor, \& Gable, 1999). Performance-approach goals have been associated with a mixed pattern of results, such that they are sometimes associated with positive effects on grades, but also with more shallow learning strategies (e.g., Elliot, McGregor, \& Gable 2001). Mastery-approach goals tend to be associated with positive affective outcomes, such as increased interest (Harackiewicz, Barron, Pintrich, Elliot, \& Thrash, 2002). Results relating to grades are inconsistent, with the majority of studies finding no relationship between grades and transfer (Linnenbrink-Garcia, Tyson, \& Patall, 2008). Critically for understanding the present study, masteryapproach goals have been linked to better performance on more difficult tasks (Utman, 1997), as well as self-reported constructive learning processes (e.g., Elliot, McGregor, \& Gable, 1999). Additionally, a small number of studies have documented a link between mastery-approach goals and knowledge transfer (e.g., Belenky \& Nokes-Malach, 2012).

It is important to note that the prevailing method for measuring achievement goals in this research literature is through self-report questionnaires and assessed at the level of academic courses. That is, goals for a particular course are assessed at the beginning of a semester and then are correlated to self-reports of learning behaviors collected during the semester, as well as achievement measures such as grades on a final exam. However, this "course-based" style of measurement may inadvertently measure more than goal motivation. That is, it may reflect other personality characteristics (e.g., Need for Cognition; Cacioppo \& Petty, 1982), beliefs (e.g., Naïve theories of intelligence, Dweck, 1999) and other variables that are not motivation, per se. Developing a strong theory of how motivation influences behavior requires a narrower focus on a task-by-task basis.

As such, in the current study, we focus on "task-based" achievement goals, and draw upon the literature that experimentally manipulates these goals for a given task. In studies of this nature experimenters provide some information to frame the task for participants in such a way that leads to the adoption of a particular achievement goal.

The question addressed in the present research is whether one can produce a change in self-explanation behaviors by manipulating motivation for the task, such that a benefit is

\footnotetext{
\({ }^{1}\) Readers familiar with achievement goal research may note the exclusion of mastery-avoidance goals in this discussion. As these goals are a newer addition to the field, and have less empirical studies to establish their effect on learning and transfer, they are not a focus of the current work.
}
observed for knowledge transfer. Specifically, we test the hypothesis that manipulated mastery-approach goals predict increased self-explanation, compared to performanceapproach or performance-avoidance goals. Additionally, we expect that the mastery-approach condition will produce higher levels of transfer, as observed for course-based goals in prior research using similar materials (Belenky \& NokesMalach, 2012).

\section*{Method}

\section*{Participants}

The participants were 140 undergraduates from the University of Pittsburgh, who participated in exchange for course credit. The first 105 participants were randomly assigned to the mastery-approach, performance-approach, or performance-avoidance conditions ( 35 each). The no-goal control was collected the following academic semester.

\section*{Materials}

The materials were presented to participants in binders. Within each packet was a pre-test, a set of learning activities, activity questionnaires, a post-test, and a final set of questionnaires.
Learning Activities. The learning materials were adapted from the "Tell-and-Practice" materials used in prior research (Belenky \& Nokes-Malach, 2012; Schwartz \& Martin, 2004). These materials comprise worked examples and problems that introduce, model, and give practice problems on two basic statistical concepts; mean deviation and standardized scores.

Specifically, participants first received a worked example on how to calculate mean deviation, which demonstrates the standard procedure. This was followed by a learning activity problem that presented data from four pitching machines and asked the participant to decide which of the four is the most reliable. The datasets are designed in such a way that contrasting between them should help focus participants' attention to the critical features of the mean deviation formula (and their conceptual underpinnings), such as the number of data points, the spread, etc. However, with the tell-and-practice nature of the activity, these aspects could be ignored in favor of a "plug-and-chug" method.

After completing this problem, participants moved on to the next worked example, which described a scenario where two students in different classes want to know who did better on a test, given that their teachers may grade differently. The worked example showed the participant how to draw a histogram for each of the classes, and then how to map the given information about means, mean deviations, and the particular students' scores onto the histogram. Finally, it explains how the participant can use this information to decide which is better. This procedure is roughly equivalent to graphically estimating a standardized score. Immediately following this worked example was another learning activity problem; this one asked students to decide which of two world records, from two different track and field events, was "more shattered." Students were given
a set of scores from two different events and two exceptional values for each, and told to use the procedure they had just learned to help them decide which had a more impressive performance, given the rest of the competitors.
Test Materials. The pre-test consisted of a procedural fluency measure, a transfer problem, and a graphical representation problem. The post-test included, in order, three procedural fluency items, a worked example on standardization, a mean deviation word problem, an openended explanation problem, and a transfer problem. This manuscript will focus exclusively on the transfer measures, which dealt with the target concept of using standardized scores to compare values from two different distributions. These problems presented an exceptional value from each of two different distributions, along with their means and standard deviations, and asked the participant to decide which of the two values was more impressive. One problem dealt with the distance of homeruns, and the other with scores on a driving test. Correctly solving these problems requires calculating a standardized score, so that the degree to which each value is exceptional, compared to the distribution it comes from, can be determined.

In accordance with research on transfer as Preparation for Future Learning, a worked example was embedded in the post-test, a few problems before the transfer. This worked example describes a scenario (with data) in which a standardized score is needed to help determine which of two performances in different athletic events was better, demonstrates the formula for a calculating a standardized score, and uses that formula to solve the problem. This is followed by another very simple problem and a prompt to use the formula to solve it, which all participants solved correctly. Correctly solving the subsequent transfer problem indicates that participants learned this procedure well, and understood that it could help them solve problems where they need to compare across two different distributions.
Activity Questionnaires. Following the learning activities, participants completed two pages of questionnaires. The first page contained a manipulation check modeled on prior research (e.g., Elliot et al., 2005). The second measured their task-based goal adoption during the learning activity, which serves as another measure to ensure that the goal manipulations produced the anticipated effects. The manipulation check asked students "At the beginning of the learning phase, you were asked to focus on just one goal for this study. What was it?" (the manipulation will be described in the next section). It asked participants to check only one of the possible responses, which corresponded to performance-avoidance, performance-approach, and mastery-approach.

The second page was a measure of their achievement goal adoption during the just-completed learning phase. Specifically, this questionnaire had six Likert-scale (1-7) items that each started with the stem, "While completing these activities," which was followed by descriptions of either a mastery-approach (skill development and improved understanding), performance-approach (doing better than
others), and performance-avoidance (not doing worse than others) goal. Two items of each of these goal types was included, and, after ensuring that they had adequate reliability (Cronbach's \(\alpha=.65\) for mastery-approach, .84 for performance-approach, .88 for performance-avoidance) a construct score was created for each by averaging across the two items.
Final Questionnaires. The questionnaires administered at the end of the experiment assessed participants' task-based goal and strategy adoption during the experiment, their course-based achievement goals for mathematics, selfreported strategy usage during the experiment, and demographic information. These will not be addressed in the current manuscript, but the general pattern of results on the goal-related measures was similar to that on the activity questionnaire and manipulation check, which will be presented in the results section.

\section*{Procedure}

The procedure followed the order of the packet, with additional instructions provided by the experimenter on the talk-aloud protocol and the goal manipulation, which both occurred after the pre-test. Specifically, the procedure consisted of: pre-test ( 5 minutes), talk-aloud training (2 minutes), goal manipulations ( 5 minutes), learning activities and activity questionnaires ( 20 minutes), post-test (20 minutes), final set of questionnaires, and a short debriefing ( \(\sim 8\) minutes). Important aspects of the procedure carried out by the experimenter will be described next.
Talk-Aloud Training. After the pre-test, the experimenter informed the participants that they would be recorded as they talked aloud during parts of the experiment. To practice doing so, participants were given a sheet with simple arithmetic problems and asked to talk aloud as they solved them. The experimenter listened and gave corrective feedback if participants were not talking aloud properly (i.e., without reflection, but simply saying what was in their working memory at the time; see Ericsson \& Simon, 1993).
Goal Manipulations. After the talk-aloud training, participants in the experimental conditions received the goal manipulation. The manipulations focused on the reasons that the study was being conducted, and how that should influence the goals participants should adopt as they went through the study. These were constructed based on reported studies (e.g., Elliot et al., 2005) and were delivered verbally by the experimenter in a conversational manner, as additional information about the study. Specifically, all participants first heard a general statement about how people can have different goals in different situations before receiving the manipulation specific to their condition. Among other aspects of the manipulation (see Table 1), the mastery-approach condition was told that their goal should be "to develop your understanding of these materials and your skill in solving these types of problems." The performance-approach condition was told to focus on the

Table 1: Representative Excerpts from the Goal Manipulations
\begin{tabular}{|c|c|c|}
\hline Mastery-Approach & Performance-Approach & Performance-Avoidance \\
\hline \begin{tabular}{l}
... we are interested in developing a set of materials that help students learn this material well. \\
... I really want you to ... try to develop your understanding of these materials \\
....At the end of the study, I will ... give you feedback on how you much you improved from the beginning of the study to the end
\end{tabular} & \begin{tabular}{l}
... we are trying to find those (students) that produce better performance than most of the other participants. \\
. focus on trying to perform better than the majority of other participants throughout the study... \\
.... At the end of the study, I ... give you feedback on how you performed relative to other participants...
\end{tabular} & \begin{tabular}{l}
...we will examine each person's performance and compare it to other students to find instances when people do particularly poorly \\
... try and not perform any worse than the majority of other participants throughout the study \\
... At the end of the study, I will ... give you feedback on how you performed relative to other participants...
\end{tabular} \\
\hline
\end{tabular}
goal "to perform well compared to other participants." The performance-avoidance condition was told their goal should be "to not perform poorly compared to other participants."

\section*{Protocol Coding}

Transcriptions were made of the verbalizations participants produced during the learning phase, the worked example in the test, and the transfer problem. These were then broken down into utterances (defined as one classifiable thought, usually at the level of a sentence). Each utterance was then coded according to a rubric which was developed based on prior research on self-explanation, and refined to reflect the statements made by participants through a process of iterated revisions. Although the full rubric covered a number of categories, this manuscript will focus on a small set that were a priori considered the most theoretically interesting. In particular, we focus on those statements coded as self-explanations and as comparisons between the transfer problem and the earlier worked example, as these are considered strong evidence of constructive learning processes (Renkl, 1997).

\section*{Results}

The first set of analyses deal with the question of how successful the manipulations were in influencing goal adoption. The two measures of goal adoption reported here are the manipulation check - which asked participants in the three experimental conditions to recall the goal they had been asked to focus on - and the activity questionnaire.

The forced-choice manipulation check clearly demonstrated that participants knew which goal they were asked to focus on, \(\chi^{2}(4, N=105)=176.35, p<.001\), with \(94 \%\) of the participants in the experimental conditions correctly choosing their condition. Each of the activity questionnaire goal adoption scores (calculated as described in the methods section) were also analyzed in separate oneway ANOVAs. There were significant differences between the three conditions for each of the mastery-approach, performance-approach, and performance-avoidance goal adoption scores, \(F \mathrm{~s}(2,102)>9.35, p \mathrm{~s}<.001\).

Additionally, a series of planned comparisons was conducted between each pair of conditions on each goal adoption score. On the measure of mastery-approach goal adoption, the mastery-approach condition was significantly higher than the performance-approach, \(t(68)=2.59, p=\)
.012 , and performance-avoidance, \(t(68)=4.13, p<.001\), conditions on mastery-approach goal adoption, but not significantly different from the control condition, \(t(68)=\) 1.53, \(p=.130\), (see Table 2). For performance-approach goal adoption, the performance-approach condition was significantly higher than the mastery-approach, \(t(67)=\) \(9.17, p<.001\), or control condition, \(t(68)=6.36, p<.001\) but not significantly higher than the performance-avoidance condition, \(t(67)=1.90, p=.062\). Finally, performanceavoidance goals were adopted the least by the masteryapproach condition, \(t\) s \((68)>4.42, p \mathrm{~s}<.001\). All other conditions did not differ statistically in their performancevoidance goal adoption, ts \((68)<.86, p \mathrm{~s}>.39\).

All told, the manipulations clearly created a different pattern of results across the mastery, control, and performance conditions. However, the performanceapproach and performance-avoidance conditions did not differentiate as cleanly.

Table 2: Activity Questionnaire Means (and Standard Deviations).
\begin{tabular}{cccc}
\hline & \begin{tabular}{c} 
Activity \\
Mastery- \\
Approach
\end{tabular} & \begin{tabular}{c} 
Activity \\
Performance- \\
Approach
\end{tabular} & \begin{tabular}{c} 
Activity \\
Performance- \\
Avoidance
\end{tabular} \\
\hline \begin{tabular}{c} 
Mastery-
\end{tabular} & \(11.31(2.19)\) & \(5.24(3.20)\) & \(5.51(3.37)\) \\
\begin{tabular}{c} 
Approach
\end{tabular} & & \\
\begin{tabular}{c} 
Performance- \\
Approach
\end{tabular} & \(9.94(2.24)\) & \(11.63(2.57)\) & \(9.89(3.05)\) \\
\begin{tabular}{c} 
Performance- \\
Avoidance \\
Control
\end{tabular} & \(8.74(2.96)\) & \(10.44(2.63)\) & \(9.83(3.62)\) \\
\hline
\end{tabular}

\section*{Goals and Transfer}

Given that the manipulations seemed to produce the desired goal adoption, we turn to the first hypothesis; namely, that mastery-approach goals would lead to better transfer. As the transfer problem was coded dichotomously, correct or incorrect, logistic regression was used to assess differences between the conditions in the likelihood of correctly solving the transfer problem. The logistic regression model predicting the likelihood of transfer based on the categorical variable of condition was significant, \(\chi^{2}(3, N=140)=\) \(11.24, p=.011\). This analysis revealed that both the control (54\%) and mastery-approach (49\%) conditions were significantly less likely to transfer than the performance-
avoidance condition (83\%). There was no significant difference between the performance-approach (69\%) and the other three conditions.

While the mastery-approach condition did not increase the likelihood of transfer, it is possible that the degree to which a student adopts a mastery-approach goal would benefit transfer. To analyze this prediction, goal adoption (as measured by the construct scores from the activity questionnaire) was entered as a predictor of a transfer in a logistic regression model. This model was significantly better than a constant-only model, \(\chi^{2}(3, N=140)=11.92, p\) \(=.008\). Within this model, the only variable which is significantly different from zero is the performanceapproach construct score, Wald's \(\chi^{2}(1, N=140)=9.36, p=\) \(.002, \operatorname{Exp}(B)=1.19\). For every unit increase in adopted performance-approach goals, the likelihood of transfer increased by \(19 \%\).

To summarize, it appears that mastery-approach goals did not lead to an increased likelihood of transfer \({ }^{2}\). Instead, the performance-avoidance condition had the highest levels of transfer, and, in terms of goal adoption, only the degree to which performance-approach goals were endorsed predicted transfer.

\section*{Goals and Self-Explanation}

The second hypothesis being investigated was that masteryapproach goals would lead to more self-explanation. Results for the three experimental conditions are reported here, as coding for the control condition remains ongoing. One participant from the mastery-approach condition is not included in these analyses, as her think aloud data was lost due to a technical error.

There were no significant differences between the experimental conditions in the number of self-explanations made overall, \(F(2,101)=.01, \mathrm{p}=.995\), or on time spent making self-explanation statements, \(\mathrm{F}(2,101)=.93, \mathrm{p}=\) .40. Subsequent analyses examined self-explanations made during each of the particular components of the study that were coded (the learning phase, worked example, and transfer problem). There were no differences between the conditions on self-explanations made during the learning phase, \(F(2,101)=.37, p=.689\). For the worked example on standardization, there was a significant difference between conditions, \(F(2,101)=4.00, p=.021, \eta^{2}=.073\). Post-hoc analysis revealed a higher degree of explanation statements for the performance-avoidance condition ( \(M=\) \(.94, S D=1.49\) ) compared to the performance-approach condition ( \(M=.37, S D=.54\) ) and the mastery-approach condition ( \(M=.38, S D=.49\) ). Finally, for the transfer problem, there was no difference between conditions on self-explanation, \(F(2,101)=2.39, p=.097\).

The other type of constructive learning behavior that was analyzed was comparisons back to the worked example during the transfer problem. For example, statements like

\footnotetext{
\({ }^{2}\) The same pattern of results is observed when controlling for pre-test transfer performance.
}
"So we'll just, um do that procedure we were doing before... like the standardized number" were coded as referencing back to the worked example. There were significant differences in this type of elaboration, \(F(2,101)\) \(=3.15, p=.047, \eta^{2}=.059\). Post-hoc analysis revealed that the performance-avoidance condition ( \(M=.37, S D=.60\) ) produced significantly more of these statements than the performance-approach condition ( \(M=.06, S D=.24\) ), but there were no differences between the mastery-approach condition ( \(M=.26, S D=.67\) ) and either of the other two.

In summary, the performance-avoidance condition generated the most self-explanation statements during the worked example. Additionally, this condition referenced the worked example more during the transfer problem than the performance-approach condition.

\section*{Discussion}

The current study experimentally manipulated achievement goals for a learning and transfer task, and found that a performance-avoidance goal manipulation had a positive effect on transfer. This condition produced the most selfexplanations during the worked example, and made the most references back to the worked example during the transfer problem. This is a different pattern of results than was expected, as course-based performance-avoidance goals are usually associated with negative outcomes, and almost never associated with positive ones. We will discuss possible reasons for these results, and describe how future research can confront the issues this research presents.

One obvious possibility is that task-based and coursebased goals do not reflect the same constructs, and produce different effects. As discussed earlier, measures of coursebased achievement goals may reflect other individual differences, like personality traits and beliefs. These (typically unmeasured) individual differences may in fact be responsible for the observed pattern of results found in prior research, rather than achievement goals themselves. Benefits that have been associated with course-based mastery-approach goals (e.g., interest, better study strategies, transfer) may actually be due to individual differences like need for cognition (Cacioppo \& Petty, 1982) or incremental theories of intelligence (Dweck, 1999), which may lead to both a higher level of mastery-approach goal endorsement, as well as self-reports of positive learning behaviors and grades. If this were so, manipulations that only target task-based goals would be unlikely to aid learning. However, this is somewhat contradicted by the present results, which do find a benefit for performance-avoidance goals. It will be important for researchers to consider in what ways course-based and taskbased goals are similar and in what ways they are different.

While the performance-avoidance condition did seem to perform the best, other measures indicate that it may be premature to tout the utility of this type of goal. Students in both performance goal conditions reported similar goal adoption, and only performance-approach goal adoption was predictive of transfer. The current results indicate that
performance goals may help guide students towards generating explanations and knowledge transfer, but research on the relative impact of each type of goal is needed. This is especially important giving the general consensus in the literature that performance-avoidance goals do not aid learning.

An interesting avenue for future research will be to explore how different types of learning activities can influence motivational effects. It is possible that performance-avoidance goals are particularly beneficial for learning simply material and for nearer transfer, compared to more challenging materials. Participants may have perceived the learning materials in this study as quite straightforward. For example, correctly solving the transfer problem required a direct application of the procedure in the worked example to the transfer problem. The performanceavoidance condition produced more self-explanation during the worked example, and referenced the worked example during the transfer problem more frequently than the other conditions, indicating a relatively direct transfer of knowledge. Mastery-approach goals may benefit "further" transfer than what was required for this task. Using goal manipulations across a variety of learning and performance tasks will allow for a richer picture to emerge. Research on regulatory fit (e.g. Higgins, 2000), for example, has shown that prevention goals (similar to avoidance goals) improve performance in tasks organized around minimizing losses, while promotion goals (similar to approach goals) help with tasks based on maximizing gains. Similar interactions may occur between achievement goals and different task structures encountered while learning.

While the results did not support the hypothesis that mastery-approach goals promote self-explanation and transfer, the study did achieve one of its aims. Specifically, a short motivational intervention produced a change in taskbased goals, which led to different levels of self-explanation and transfer. While important questions were raised as to which goal may be most beneficial in which settings, the fundamental premise (that goals influence learning behaviors and transfer) was supported. This research can provide a basic paradigm for further inquiry.

The study of motivation in academic settings is a fruitful enterprise, as it can inform both cognitive psychology theories of how people learn, as well as educational practice to improve learning outcomes. Future research should continue to address what influence adopted goals have on behavior, as theories of motivation are useful inasmuch as they can be used to predict behaviors. Measures of coursebased goals seem quite predictive of long-term success, and of certain self-reported attitudes, behaviors, and affective experiences, but it is less clear what effect they have on moment-to-moment behaviors. More broadly, incorporating motivation into theories of learning and knowledge transfer (see Nokes \& Belenky, 2011) remains an important goal for cognitive scientists who wish to have an impact on educational practice, while also furthering our understanding of the human cognitive architecture.

\section*{References}

Aleven, V.A.W.M.M., \& Koedinger, K.R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. Cognitive Science, 26, 147-179.
Belenky, D.M., \& Nokes-Malach, T.J. (2012). Motivation and transfer: The role of mastery-approach goals in preparation for future learning. Journal of the Learning Sciences, 21, 399-432.
Cacioppo, J.T., \& Petty, R.E. (1982). The need for cognition. Journal of Personality and Social Psychology, 42, 116-131
Chi, M.T.H., Bassok, M., Lewis, M.W., Reimann, P., \& Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. Cognitive Science, 13, 145-182.
Dweck, C. S. (1999). Self-theories: Their role in motivation, personality, and development. Philadelphia: Psychology Press.
Elliot, A.J., McGregor, H.A., \& Gable, S. (1999). Achievement goals, study strategies, and exam performance: A mediational analysis. Journal of Educational Psychology, 91, 549-563.
Elliot, A.J., Shell, M.M, Henry, K.B., \& Maier, M.A. (2005). Achievement goals, performance contingencies, and performance attainment: An experimental test. Journal of Educational Psychology, 97, 630-640.
Ericsson, K. A., \& Simon, H. A. (1993). Protocol analysis: Verbal reports as data. Boston: MIT Press.
Harackiewicz, J.M., Barron, K.E., Pintrich, P.R., Elliot, A.J., \& Thrash, T.M. (2002). Revision of achievement goal theory: Necessary and illuminating. Journal of Educational Psychology, 94 (3), 638-645.
Higgins, E. T. (2000). Making a good decision: Value from fit. American Psychologist, 55, 1217-1230.
Linnenbrink-Garcia, L., Tyson, D.F., \& Patall, E.A. (2008). When are achievement goal orientations beneficial for academic achievement? A closer look at main effects and moderation factors. International Review of Social Psychology, 21, 19-70.
Pashler, H., Bain, P., Bottge, B., Graesser, A., Koedinger, K., McDaniel, M., \& Metcalfe, J. (2007). Organizing instruction and study to improve student learning (NCER 2007-2004). Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.
Renkl, A. (1997). Learning from worked-out examples: A study on individual differences. Cognitive Science,21,129.

Schwartz, D. L., \& Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. Cognition and Instruction, 22, 129-184.
Utman, C.H. (1997). Performance effects of motivational state: A meta-analysis. Personality and Social Psychological Review, 1 (2), 170-182.

\title{
The Impact of Communicative Constraints on the Emergence of a Graphical Communication System
}

\author{
Till Bergmann (tbergmann@ucmerced.edu) Rick Dale (rdale@ucmerced.edu) \\ Cognitive and Information Sciences, University of California, Merced, Merced, CA 95343 \\ Gary Lupyan (lupyan@wisc.edu) \\ Department of Psychology, University of Wisconsin-Madison, Madison, WI 53706
}

\begin{abstract}
We investigated the behavior of participants tasked with communicating in a novel environment. Participants had to use their mouse to draw graphical representations (termed squiggles in the game) of human faces in order to communicate with fellow players. Experiment 1 investigated the effect of varying features of the input images on the resulting drawings. Experiment 2 introduced varying comprehension conditions that were predicted to produce differences in how features of faces would be graphically represented. In experiment 1 , the features of the different faces significantly shaped the structure of the resulting squiggles. In experiment 2, the structure of the squiggles was influenced by the environment in which they were interpreted.
\end{abstract}

Keywords: social interaction; language evolution; evolution of communication; social feedback; reference; games

\section*{Introduction}

For an individual learner, language can be described as a moving target (Christiansen \& Chater, 2008; Chater, Reali, \& Christiansen, 2009). Languages are commonly assumed to undergo changes at various levels. Recent studies have sought to identify the extra-linguistic factors that may determine the direction of these changes. These perspectives articulate an adaptive approach to our communication system: Any existing variability in a population of language users may have their language move in particular directions because some of those variants are more useful or successful than others at propagating. For example, recent studies have argued that the social environment of languages is a very important factor for these changes and developments (Lupyan \& Dale, 2010; Wray \& Grace, 2007). In addition, humans are able to invent new communication systems on the fly, with the emergence of Nicaraguan Sign Language and Al-Sayyid Bedouin Sign Language being two of the most impressive cases (Meir et al., 2010).

In recent years, a number of studies have been conducted with a focus on the structure of new communication systems that emerge in a given environment. In these studies, participants have to use novel symbols to successfully communicate with other participants. Over time, participants achieve more efficiency in their communication by aligning their symbols (Fay et al., 2010; Galantucci, 2005; Garrod et al., 2007).

In another line of research, the same question has been tackled by the use of different "generations" of participants. The first generation of participants learns simple word-object patterns, and then passes this information on to a new generation of learners. After a couple of generations, struc-
ture emerges that was absent in the original language system (Kirby, Cornish, \& Smith, 2008).

This literature offers new insights, and importantly new techniques, for exploring artificial communication systems as a testbed for the forces that may drive communicative change. The present paper aims to shed light on the ways structure emerges in a new communication system by combining these two experimental frameworks in a large-scale interactive communication game. Specifically, we look to how the structure of a referential environment - the space of things that a community is going to talk about, and how these things are encountered - shapes the manner in which individuals in that community communicate. A first, exploratory experiment investigated the different strategies participants would employ to communicate a given image and how this given image influences the structure of the produced symbol. A further experiment introduced different environmental conditions that affected the way symbols were structured.

\section*{Previous Work}

The current study expands on previous work by Dale and Lupyan (2010), in which players connect via the Internet to a gaming platform, where they had to create and comprehend visual signs for objects in order to communicate with other participants. Successful communication was achieved when players where able to match a previously drawn sign to its corresponding object.

The experiment consisted of two trial types: the production (speaking) and comprehension (listening) trials. In a speaking trial, participants were shown an object, and after a short delay, had three seconds to "squiggle" an image by clicking and dragging their computer mouse on a provided canvas. Such a method is akin to putting a temporal bandwidth on the ability of the user to sketch the image. This was designed to avoid their using English orthography or detailed sketches. Participants had to focus on essential aspects of an image to relay it. A different approach was used by Galantucci (2005), who used a rotating tablet to keep subjects from using orthography. Because this game framework is used online, the temporal bandwidth is a straightforward way of doing this, while also moving the game along more quickly.

In a listening trial, players were shown a previously drawn squiggle, along with two pictures of objects. The participant then had to choose the matching picture by clicking on it. Feedback is given whether they chose the correct picture. In the original work, squiggles in the listening trial were cho-
sen according to an evolutionary algorithm factoring in the novelty of the squiggle and its previous comprehensibility. Squiggles that performed badly in listening trials thus "die out" and are no longer used.

Data from 60 users were collected, which resulted in around 1,400 produced squiggles and around 4,100 listening trials. Feedback by the players was generally very positive and many found it to be entertaining and rather addictive.

Results from this study supported the findings from other studies: a) Squiggles get simplified. The average size of a squiggle shrinks over gameplay; b) The evolutionary algorithm produced stability for most squiggles, despite opportunities for novel squiggles to replace them during listen trials and c) Squiggles were at first drawn highly iconically, but gradually evolved towards more simplified symbols as participants continued playing.

In the current paper, we utilize this framework as a testbed for exploring how environments shape the squiggles as participants play. The central hypothesis guiding the study was that users' squiggles are shaped by their "listening" environments (i.e., distribution over the foil-target pairs). In what follows, we first describe our redesign of the system, and an initial exploration (Experiment 1) of how aspects of objects can predict aspects of squiggles. Following this, we describe a comparison of two conditions (Experiment 2) meant to more subtly explore the role of the distribution of listening trials in determining squiggle composition.

\section*{Experiment 1}

\section*{Method}

Game Design The approach in Dale and Lupyan (2010) was replicated in the present study. However, the game was re-written to take into account new technological developments to ensure maximal compatibility with current browsers. We chose a specific domain of objects for squiggles to make "reference" to: human faces. While the previous study used a non-standardized set of pictures, all pictures were now taken from the Face Database (Minear \& Park, 2004). This had the advantage that all picture used the same light conditions, background, facial expressions and dimensions. Variability on those parts was thus excluded. 120 pictures were chosen with respect to their age and gender, using a \(2 \times 2\) matrix, with 30 pictures in each set (cf. Table 1). In order to reduce the number of salient features in these images, pictures with very distinguishing features such as birthmarks and earrings were not chosen.

Table 1: Summary statistics for the different picture sets
\begin{tabular}{|l|c|c|c|}
\hline Group & M (Age) & SD (Age) & Range (Age) \\
\hline Young Males & 21.35 & 2.35 & \(18-28\) \\
Young Females & 21.53 & 2.85 & \(18-28\) \\
\hline Old Males & 73.37 & 7.47 & \(61-91\) \\
Old Females & 73.80 & 6.69 & \(61-85\) \\
\hline
\end{tabular}

Participants connected to the gaming platform via the Internet. Players were instructed to squiggle the displayed images so that other players are able to match a squiggle to the correspoding image later.

In the speaking trial, one of the 120 pictures in the data set was randomly shown to the participant along with a framed "canvas" on which participants could draw (Fig. 1). Participants had five seconds (without delay) to draw on the canvas before a listening trial appeared. The time limit was increased from the earlier study as the delay until the countdown started was removed. The remaining time for the speaking trial was shown to the user below the canvas.

In the listening trial, a squiggle previously produced by a fellow player is animated on a canvas using the original speed and directions of the user who had drawn the squiggle. The matching image was then placed either above or below this squiggle, with a competitor image taking up the other position (Fig. 2). Players had to click on the matching image in order to successfully communicate, and received visual feedback whether they had made the right choice. The ratio of speaking to listening trials was 1:3.

An important constraint on communication was time - participants were given only 5 seconds to create the squiggle.

Remember: Draw a Squiggle representing
the person you see so that other people can make the connection later!


Figure 1: An example of a speaking trial.

Expectations Experiment 1 mainly served as a first step in testing the impact that different image dimensions have on the behavior of participants. It also served as an exploratory study of this behavior. We were interested in the different strategies employed by players to communicate. Which salient features were chosen and which were the most efficient? The input images were believed to have an effect on the way squiggles were structured by the participants; that is, we expected dimensions such as age and gender to influence squiggle strategies.

It was predicted that users would perform better in the listening trials with more practice, and hypothesized that the

Remember: Click the picture you think the Squiggle is representing!


Figure 2: An example of a listening trial.
complexity of produced squiggles would shrink over time. Players were expected to become more efficient with more exposure to the communication system in both the listening and speaking trials.
Participants 100 participants from the United States were recruited via Amazon Mechanical Turk and compensated financially (\$0.75) for taking part in the study. The use of Amazon Mechanical Turk had the indispensable advantage that players signed up almost instantaneously and simultaneously, which created a community of speakers playing at the same time and communicating with each other. Participants had to produce thirty squiggles and the respective number of listening trials (90) in order to receive compensation.

\section*{Data}

On speaking trials, measurements of coordinates were taken along the paths drawn by the players. Time stamps were collected at each sampled \(x, y\) coordinate. The length in pixels of each squiggle was calculated using the stored coordinates. The number of unique strokes was collected as well.

In the listening trial, unique identifiers for the player, squiggle and corresponding image as well as the competitor image were stored. The comprehension of the result whether the correct picture was chosen - was stored in a binary fashion along with the time it took the player to decide.

Data by participants who produced fewer than thirty squiggles were excluded, leaving data from 96 participants.

\section*{Results}

Communication Strategies (Speaking) We performed a linear mixed-effects model on the relationship between the total length of a squiggle in pixels and the interaction of age, gender and progression of gameplay. Random effects consisted of the subject and item. \(p\)-values were obtained by calculating the likelihood of the corresponding \(t\)-values in a normal distribution. We found main effects for gender and age of the target, as well as for the progression of gameplay. Additionally, there was an effect for the interaction between age and gender. Squiggle length decreased for male input images \((t=-9.47, p<0.0001\) ) by 129.6 pixels \(\pm 13.68\) (standard error). Squiggle length decreased for young input images \((t=-3.19, p=0.001)\) by 43.33 pixels \(\pm 13.68\). Length increased slightly over time by 1.1 pixels \(\pm 0.39(t=2.71\), \(p<0.0001\) ). The interaction of age and gender meant that squiggles for young males were 42.24 pixels \(\pm 19.37\) longer than for old males ( \(t=3.91, p<0.0001\) ), negating the main effect of age for males (Fig. 3).


Figure 3: Average length of squiggles grouped by item age and gender

To account for possible differences in compositionality, we carried out a linear mixed-effects model on the relationship between the number of strokes and the interaction of age, gender and progression of gameplay. \(p\)-values were obtained by calculating the likelihood of the corresponding tvalues in a normal distribution. We obtained main effects for age, gender and progressing gameplay. Squiggles contained more features \((+0.195 \pm 0.08)\) if depicting a male \((t=\) \(-2.42, p=0.015\) ) or young person \((+0.38 \pm 0.08, t=4.80\), \(p<0.0001\) ). The number of strokes also increased slightly by \(0.008 \pm 0.002\) with more practice \((t=3.43, p<0.001)\). An effect of the interaction between age and gender \((t=-3.55\), \(p=0.0004\) ) meant that the main effect of age was negated for males, as old males elicited slightly more strokes (0.02) than young males (Fig. 4).

Comprehension Performance (Listening) Overall, participants chose the correct image \(71 \%\) of the time, demonstrating that successful communication in novel environments is


Figure 4: Average number of strokes grouped by item age and gender.
achievable even in a short period of time.
A linear-mixed effects model on the relationship of comprehension and the interaction of age, gender and progressing gameplay was constructed. An additional fixed effect was the time it took participants to make their decision. This variable was log-transformed to reform skew. Random effects were the users, items and the competitor images. The analysis revealed no significant effect for age, gender and progressing gameplay. The model produced a main effect for decision time \((z=-3.56, p=0.0004)\). Scores decreased by \(0.18 \pm\) 0.05 the longer users contemplated their decision.

\section*{Discussion}

As predicted, squiggles varied in structure and appearance depending on the image drawn. For females, participants focused on the hair form and length, which resulted in longer squiggles for females than males. Players tended to encode several, short features for younger females and fewer, but longer, features for old females. The increased number of strokes for young females can be explained by a preference for drawing hair which required several strokes (cf. Fig. 5a). Both strategies performed with the same level of accuracy in listening trials, which shows that each strategy chose the right aspects to be encoded.

Salient features of young males were often hair or head form, and on average, squiggles contained more features for males than for females. Old males were typically depicted by focusing on wrinkles and mouth form, which resulted in very short squiggles containing several features. The number of strokes did not vary much across age for males, as in both categories several features were encoded. A range of strategies across ages and genders is shown in figure 5 . These were successful strategies, in that the squiggles were easy to be recognized among those who "listened" to it.

Both the number of encoded features and the total length of squiggles increased over time. This does not support our hypothesis that the complexity of squiggles would shrink over time. There are two explanations for this which are not mu-
tually exclusively. First, participants did not play the game long enough to get an effect. Second, there is no explicit pressure on the participants to reduce complexity. The time limit given in the speaking trial is sufficient enough to encode several features, and there is no motivation for users to change their strategy.

The fact that players have ample time in the speaking trial to produce detailed squiggles without much practice means that squiggles are comprehended right away. This leads to a ceiling effect and participants do not get better at comprehending squiggles over time. A decrease of the time limit in the speaking trial will very likely produce a decrease of comprehensibility at the beginning of the game, as players will not be able to produced detailed enough squiggles. Accuracy of comprehension should, however, increase over time in this scenario.

This initial study simply aimed at identifying basic strategies of players, observing the dynamics of gameplay (effect of time on production/comprehension), and whether squiggles can be quantified as having particular properties associated with particular classes of stimuli.


Figure 5: Different strategies employed in successful squiggles

\section*{Experiment 2}

\section*{Method}

Game Design After the initial experiment, a second, more refined experiment was carried out. We introduced a second condition in the listening trial, in which the competitor of the
displayed squiggle was always of the opposite age and gender. While age is normally a continuous variable, we chose the pictures in such a way that made a binary distinction possible (see above). For example, if the squiggle depicted a young male, the competitor would now always be an old female. The new condition provides additional contextual clues to the players: In order to match a squiggle, only age and gender need to be discriminated against instead of focusing on individual features. The other condition remained as in experiment 1, i.e. the competitor was randomly chosen with no regard to age or gender. A simple encoding of age and gender is not be sufficient in this condition.

The ratio of speaking trials to listening trials was changed additionally in both conditions. The higher rate of language comprehension in the early stages of language acquisition was modeled by increasing the number of listening trials to 10 for the first three speaking trials, and then slowly decreasing the number back to 3 . Participants had to go through 20 speaking trials and 110 listening trials in total. No further changes were made to the game setup.
Expectations We predicted that the a changed comprehension environment would result in subsequent changes in the production of squiggles. In particular, we hypothesized that squiggles in the new condition would require less detail and complexity in order to achieve successful communication than squiggles in the control condition. The difference should be clearly visible after a short acquisition time. We expected a difference to occur roughly after the third speaking trial after which the number of listening trials was slowly decreased.

The expected differences in the structure of squiggles were predicted to not impact the performance of players in listening trials. Experiment 1 has shown that players choose the matching item with a high accuracy with little practice, and this accuracy is believed to stay constant in this experiment.
Participants 25 participants were recruited through Amazon Mechanical Turk and compensated financially by receiving \(\$ 0.75\).

\section*{Results}

A linear mixed-effects model was conducted with the number of strokes as the dependent variable, and the interaction of the conditions with progressing gameplay as the fixed factor. Intercepts for subjects and items were used as random effects. We found a main effect of condition \((t=-3.31, p<0.001)\), showing that the number of strokes decreased by \(1.06 \pm 0.32\) in the between condition. The interaction of between condition and the progressing gameplay proved to be significant ( \(t=-2.91, p=0.0036\) ), lowering the number of strokes by \(-0.07+ \pm 0.02\), despite an increase of \(0.07 \pm 0.01\) generally over time ( \(t=5.2, p<0.0001\) ).

We also performed a linear mixed-effects model to check for possible effects on comprehension of the interaction of conditions, progression of gameplay and the time it took the participants to make their choice. Random effects controlled
for participants, items and competitor images. The variable time was log-transformed to remove skew. The model revealed no effects on comprehension.

\section*{Discussion}

As predicted, the number of strokes increased after roughly three speaking trials in the condition in which the competitor image was chosen randomly, but not in the condition in which comprehension only required discriminating between gender and age categories. (Fig. 6). Additionally, the number of strokes continuously dropped in the latter condition as the game progressed. This sudden change after a short acquisition time is in line with other studies, which have shown that novel strategies are picked up rapidly by the majority of the population after a short period of time (Steels, 2011). The number of strokes should stabilize in each condition if players continue playing. Figure 7 shows examples of successful squiggles with differing numbers of strokes and encoded features.

Both conditions performed equally well in the listening trial, which shows that both conditions adapted to their given environment successfully. Importantly, fewer details were needed in the between-category condition to achieve the same accuracy as squiggles with more complexity in the condition in which the competitor was chosen completely arbitrarily.


Figure 7: Different number of encoded features corresponding to the number of strokes a player drew.

\section*{General Discussion}

Experiment 1 used a simple design in order to study the impact that visual and conceptual properties of have on the structure of squiggles drawn to communicate these objects. Input images differed in age and gender, and the produced squiggles varied in their structure according to these factors. Participants were able to adapt quickly to the given environment and performed well in comprehending squiggles from the start. The second experiment examined the impact of different comprehension conditions. An increase of listening trials at the start of the game and the introduction of different conditions led to varying squiggle structures across time and conditions. The referential environment had a profound impact on squiggle production. Specifically, we observed a decrease in squiggle complexity in the environment where the

\section*{Number of Strokes grouped by condition}


Figure 6: The evolution of the number of strokes over time across conditions
referents had to be distinguished categorically (although the participants were never informed of this constraint).

This method developed here to explore the evolution of communication in a simple graphical environment lends itself to numerous experimental manipulations. The impact of several parameters can be explored in this context. For example, a decreased time limit in the speaking trial will increase pressure on participants to draw more efficiently and to focus on fewer features. This may lead to a greater influence of the referential environment: Conditions in which categorical information (age and gender) is sufficient for successful communication lead to squiggles with simpler surfacestructure, a likely starting place for the emergence of morphemes and compositional structure. The emergence of such morphemes could be sped up by using a system of immediate feedback, in which players are repeatedly informed how successful their squiggles perform in the listening trials, i.e. whether other players recognize their squiggles. Immediate feedback allows players to establish more common ground and has been shown to drastically increase communication performance (Garrod et al., 2007).

This study shows that humans rapidly create efficient communication systems in a novel shared environment by integrating contextual clues. Comprehension and production were tightly interconnected in this communication game, and the referential environment significantly affected the structure of the produced squiggles. This lends further support to the idea that communication systems adapt to the environment in which they are learned and used.

\section*{References}

Chater, N., Reali, F., \& Christiansen, M. H. (2009). Restrictions on biological adaptation in language evolution. PNAS, 106(4), 1015-1020.
Christiansen, M. H., \& Chater, N. (2008). Language as
shaped by the brain. Behavioral and Brain Sciences, 31(5), 489-558.
Dale, R., \& Lupyan, G. (2010). Squiggle: Large-Scale Social Emergence of Simple Symbols. In A. D. M. Smith, M. Schouwstra, B. de Boer, \& K. Smith (Eds.), The Evolution of Language: Proceedings of the 8th International Conference (pp. 391-392). New Jersey: World Scientific.
Fay, N., Garrod, S., Roberts, L., \& Swoboda, N. (2010). The interactive evolution of human communication systems. Cognitive Science, 34(3), 351-386.
Galantucci, B. (2005). An experimental study of the emergence of human communication systems. Cognitive Science, 29(5), 737-767.
Garrod, S., Fay, N., Lee, J., Oberlander, J., \& Macleod, T. (2007). Foundations of representation: Where might graphical symbol systems come from? Cognitive Science, 31(6), 961-987.
Kirby, S., Cornish, H., \& Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. PNAS, 105(31), 10681-10686.
Lupyan, G., \& Dale, R. (2010). Language structure is partly determined by social structure. PloS one, 5(1), 1-10.
Meir, I., Sandler, W., Padden, C., \& Aronoff, M. (2010). Emerging sign languages. In M. Marschark \& P. E. Spencer (Eds.), Oxford Handbook of Deaf Studies, Language, and Education, Volume 2 (pp. 267-280). Oxford: Oxford University Press.
Minear, M., \& Park, D. C. (2004). A lifespan database of adult facial stimuli. Behavior Research Methods, Instruments, \& Computers, 36(4), 630-633.
Steels, L. (2011). Modeling the cultural evolution of language. Physics of Life Reviews, 8(4), 339-356.
Wray, A., \& Grace, G. W. (2007). The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. Lingua, 117(3), 543-578.

\title{
Analogy and Arithmetic: An HDTP-Based Model of the Calculation Circular Staircase
}

\author{
Tarek R. Besold (tbesold@uos.de) and Martin Schmidt (martisch@uos.de) \\ Institute of Cognitive Science, University of Osnabrück, 49069 Osnabrück, Germany
}

\author{
Alison Pease (a.pease@ed.ac.uk) \\ School of Informatics, University of Edinburgh, Edinburgh, EH8 9AB, UK
}

\begin{abstract}
Analogical reasoning and its applications are gaining attention not only in cognitive science but also in the context of education and teaching. In this paper we provide a short analysis and a detailed formal model (based on the Heuristic-Driven Theory Projection framework for computational analogy-making) of the Calculation Circular Staircase, a tool for teaching basic arithmetic and insights based on the ordinal number conception of the natural numbers to children in their first years of primary school. We argue that such formal methods and computational accounts of analogy-making can be used to gain additional insights in the inner workings of analogy-based educational methods and tools.
Keywords: Analogy, Education, Teaching, Arithmetic, Formal Model, Computational Analogy-Making, HDTP.
\end{abstract}

\section*{Introduction}

Analogical reasoning is the ability to perceive, and operate on, dissimilar domains as similar with respect to certain aspects based on shared commonalities in relational structure or appearance. This has been proposed as an essential aspect of the ability to learn abstract concepts or procedures (Gentner, Holyoak, \& Kokinov, 2001), and is recognised as ubiquitous in human reasoning and problem solving (Gentner, 1983), representational transfer (Novick, 1988), and adaptation to novel contexts (Holyoak \& Thagard, 1995).

Inherent in the structure of analogical reasoning is its role in education and learning: new ideas can be constructed and explored in relation to familiar concepts. While substantial research has been carried out into the role of analogical reasoning and science education (see, for instance, (Duit, 1991; Arnold \& Millar, 1996; Guerra-Ramos, 2011)), its role in mathematics education has been somewhat less explored although notable exceptions include (Pimm, 1981; English, 1997). These studies support our assumption that analogies can be used for facilitating the understanding of concepts and procedures in abstract and formal domains, such as mathematics, physics or science. The pedagogical use of analogies as a means of triggering, framing and guiding creative insight processes still needs to be widely recognised as part of teaching expertise and incorporated into innovative teacher education schemes (Akgul, 2006).

In this paper, we want to contribute to a deeper understanding of the role and the mode of operation of analogy in an educational context by first providing a description and short analysis of the analogy-based Calculation Circular Staircase used for teaching basic arithmetic to children attending their initial mathematics classes at the beginning of
primary school (Schwank, 2003; Schwank, Aring, \& Blocksdorf, 2005), before showing how a computational analogymaking framework as Heuristic-Driven Theory Projection (HDTP) (Schwering, Krumnack, Kühnberger, \& Gust, 2009) can be used to provide a formal computational reconstruction of the staircase as a prototypical example of analogy-use taken from a real-life teaching situation. We thereby also continue the work started in (Besold, 2013) with a far more complex and deep-rooted case study. By doing so, we aim to show one way (amongst several) of how analogy-engines and their corresponding background theories can fruitfully be applied to modeling and analysis tasks from the field of psychology of learning, education, and didactics.

\section*{Heuristic-Driven Theory Projection (HDTP)}

There is much work on both theoretical and computational models of analogy-making. Heuristic-Driven Theory Projection (HDTP) (Schwering et al., 2009) is one such perspective: this is a formal theory and corresponding software implementation, conceived as a mathematically sound framework for analogy-making. HDTP has been created for computing analogical relations and inferences for domains which are given in the form of a many-sorted first-order logic representation. Source and target of the analogy-making process are defined in terms of axiomatizations, i.e., given by a finite set of formulae. HDTP tries to produce a generalization of both domains by aligning pairs of formulae from the two domains by means of anti-unification: Anti-unification tries to solve the problem of generalizing terms in a meaningful way, yielding for each term an anti-instance, in which distinct subterms have been replaced by variables (which in turn would allow for a retrieval of the original terms by a substitution of the variables by appropriate subterms).

HDTP in its present version uses a restricted form of higher-order anti-unification. In higher-order anti-unification, classical first-order terms are extended by the introduction of variables which may take arguments (where classical firstorder variables correspond to variables with arity 0 ), making a term either a first-order or a higher-order term. Then, antiunification can be applied analogously to the original firstorder case, yielding a generalization subsuming the specific terms. The class of substitutions which are applicable in HDTP is restricted to (compositions of) the following four cases: renamings (replacing a variable by another variable of the same argument structure), fixations (replacing a variable by a function symbol of the same argument structure), ar-


Figure 1: The "big" Calculation Circular Staircase (as depicted in (Schwank et al., 2005)): Numbers from 1 to 9 are represented by orange balls in the inner circle, numbers from 10 to 19 by green and orange balls in the outer one, the white door on the right marks the transition point between circles.
gument insertions, and permutations (an operation rearranging the arguments of a term). This formalism has proven capable of detecting structural commonalities not accessible to first-order anti-unification, as for instance also structural commonalities between functions and predicates within the logical language can be found and exploited (whilst the firstorder formalism would in these be limited to the respective argument positions only), allowing for a better recognition of relational mappings (as opposed to mere attribute mappings). Once the generalization has been computed, the alignments of formulae together with the respective generalizations can be read as proposals of analogical relations between source and target domain, and can be used for guiding an analogybased process of transferring knowledge between both domains. Analogical transfer results in structure enrichment on the target side, which corresponds to the addition of new axioms to the target theory, but may also involve the addition of new first-order symbols.

\section*{The Calculation Circular Staircase}

Dedekind (Dedekind, 1887/1969) argued that ordinal numbers and insights into the basic structure of the natural numbers play a crucial role in understanding (and thus also teaching) the foundations of arithmetic. About a century later, studies by Brainerd (Brainerd, 1979) also showed that a deeper understanding of the ordinals supports and facilitates the learning of basic arithmetic operations in children. Based on this line of thought, the Calculation Circular Staircase (Schwank, 2003; Schwank et al., 2005) has been developed.

\section*{A Teaching Tool for Basic Arithmetic}

Learning by analogy requires conceiving of and performing a transfer mapping of concepts and relational structures from a
better-known base domain into a less familiar target domain. This mapping is typically established by a pairwise matching of individual elements from the respective domains, resulting in a set of systematic correspondences. In the context of mathematics education and mathematical reasoning, children are required to understand abstract relations and operations (such as equality, addition, and subtraction) which can best be taught by drawing parallels between similar examples in less abstract domains (Clement, 1993). Still, the availability of supportive cues for the analogy is crucial for the success of the learning process (Glynn, Duit, \& Thiele, 1995).

The Calculation Circular Staircase (cf. Fig. 1) offers children a means of developing an understanding of the interpretation of numbers as results of transformation operations. This goal shall be achieved by enabling a mental functional motor skill-based way of accessing the foundational construction principles of the number space and the corresponding basic arithmetic operations: The numbers from 0 to 9 are represented in the inner circle by a corresponding number of orange balls, numbers from 10 to 19 are represented in the outer circle at the respective places (corresponding to the inner circle's ordering) by 10 green balls and a corresponding number of orange ones. A little door indicates the point of transition between circles. Arithmetic operations are introduced via "magical" signs (showing " + " or "-") carried by toy figures. When equipped with the respective sign, a toy figure can perform jumps on the staircase - before moving, a decision for a sign has to be taken (involving the child in a responsible and motor active way instead of assigning the role of a passive spectator). Addition corresponds to an ascending movement, subtraction to a movement in descending direction. This enables children to experience subtraction as a proper inverse operation to addition, arising naturally from the "wish" of a toy figure to also descend the staircase.

Decimal structure-based analogies between different computations (e.g. between " \(5-4=1\) " and " \(15-4=11\) ") are made accessible to children's understanding via synchronous movements of two toy figures in the inner and outer circle, respectively: The 10 green balls in the outer circle stay constant, with respect to the orange balls identical movements yield identical results. The door between the circles provides children with a natural "resting point" for simplifying difficult computations involving a decimal transition: If e.g. \(13-5\) shall be computed, the toy figure (carrying the "-" sign) first moves to the column representing the number 10 (being the only column in the outer circle not containing an orange ball and being directly next to the door, naturally corresponding to the 0 -column in the inner circle), losing as many orange balls in height as were initially situated below the figure. The remaining height difference of 2 is now accounted for in a second step, thus transforming the original task into the easy to handle \(10-2\) and making the decimal transition attractive for the children. Also, the number 0 obtains a natural position in the number system of the Calculation Circular Staircase, simply corresponding to the result of performing
another step down from the first stair. Having the caesura between the two circles after the representation of the number 9 becomes meaningful to the children once they write down the corresponding numbers, there also encountering a significant difference between the initial one-digit numbers and the two-digit numbers starting with 10.

At the level of cognitive analysis, the idea underlying the Calculation Circular Staircase is an active recruitment of the "functional thinking" approach to mathematics (cf., e.g., (Schwank, Gelfman, \& Nardi, 1999)). As opposed to a "predicative thinking"-style understanding of relationships within mathematics, which uses equality as ordering principle when conceptualizing mathematical structures (i.e., mathematics being conceptualized on the basis of the repeated applicability of certain predicates), the "functional thinking" perspective bases its conceptualizations on differences between mathematical concepts which can then be used to conceive of a construction process for the respective class of structures (i.e., mathematics being conceptualized on the basis of repeated constructive steps). In their interaction with the Calculation Circular Staircase, the children are naturally led into taking the "functional thinking" approach across repeated stages of play, each time actively becoming aware of the individual steps a toy figure has to take when changing from one position on the staircase to another one. Instead of post hoc merely checking whether a certain distance has been covered between the initial step and the final one (i.e., whether a predicate indicating a certain value for the difference between both steps holds), they experience an active construction process explicating the guided transition from the initial step to the final one.

\section*{An HDTP-Based Model of the Staircase}

We now reconstruct the "big" Calculation Circular Staircase (i.e., the version equipped with two circles or 19 steps) as an analogy-based model for understanding and learning among others important aspects of the ordinal number conception of the natural numbers in the range from 0 to 19 .

The analogy uses the Calculation Circular Staircase as a base domain, transferring the structure and relational conception children acquire by playing with the staircase into their previously acquired knowledge about natural numbers as target domain. The latter domain is most likely initially still very poor as compared to the Calculation Circular Staircase domain as only very little (if any at all) internal structure or relations have been acquired besides the mere ordering of the number terms from one to nineteen that had been committed to memory in previous lessons. And even for this ordering it can be assumed that the ordering has mostly only been developed on basis of isolated neighboring tuples of the form \((n, n+1)\), for each number term only remembering its immediate successor. The arithmetic operations " + " (addition) and "-" (subtraction) are known as abstract concepts (as are their corresponding addition and subtraction tables), but have not yet been developed into a grounded, constructively applicable conceptualization. Table 1 gives a formal HDTP-style
```

Sorts:
steps, sign, circle, caesura, direction, time, natural.
Entities:
one, two, three, $\ldots$, nineteen, zero, $S_{a}, S_{b}, S_{c}, S_{d}$ : steps.
,+- : sign.
up, down, $D_{1}$ : direction
door, $C_{1}$ : caesura.
$T_{1}, T_{2}$ : time.
inner, outer, $\mathrm{Ci}_{1}$ : circle
Functions:
height : steps $\rightarrow$ natural.
interpretation : sign $\rightarrow$ direction
dist : direction $\times$ steps $\times$ steps $\rightarrow$ direction $\times$ natural.
magn : direction $\times$ natural $\rightarrow$ natural.
diff: direction $\times$ steps $\times$ steps $\rightarrow$ natural.
Predicates:
succ : steps $\times$ steps.
higher : steps $\times$ steps
lower : steps $\times$ steps.
inCircle : circle $\times$ steps.
base : steps.
inFocus : steps $\times$ time.
currentSign : sign $\times$ time
between : caesura $\times$ steps $\times$ steps.
move : steps $\times$ natural $\times$ direction $\times$ time.
analogs: steps $\times$ steps.
Facts:
$\left(s_{1}\right) \operatorname{succ}(z e r o$, one $)$.
$\left(s_{2}\right) \operatorname{succ}($ one, two).
( $s_{19}$ ) succ(eighteen, nineteen).
( $s_{20}$ ) inCircle(inner, zero).
$\left(s_{21}\right)$ inCircle(inner, one).
( $s_{29}$ ) inCircle(inner, nine).
$\left(s_{30}\right)$ inCircle(outer, ten).
( $s_{39}$ ) inCircle(outer, nineteen)
( $s_{40}$ ) between(door, nine, ten).
Laws:
$\left(s_{41}\right)$ interpretation $(+)=u p$.
$\left(s_{42}\right)$ interpretation $(-)=$ down
$\left(s_{43}\right)$ interpretation $(+) \neq$ interpretation $(-)$.
$\left(s_{44}\right) \operatorname{higher}\left(S_{a}, S_{b}\right) \leftrightarrow \operatorname{succ}\left(S_{b}, S_{a}\right) \vee \exists S_{c}:\left(\operatorname{succ}\left(S_{b}, S_{c}\right) \wedge \operatorname{higher}\left(S_{a}, S_{c}\right)\right)$
$\left(s_{45}\right)$ higher $\left(S_{a}, S_{b}\right) \leftrightarrow$ lower $\left(S_{b}, S_{a}\right)$.
$\left(s_{46}\right) T_{1}<T_{2}: \operatorname{inFocus}\left(S_{a}, T_{1}\right) \wedge$ currentSign $\left(+, T_{1}\right) \wedge$
move $\left(S_{a}, 1\right.$, interpretation $\left.(+), T_{1}\right) \wedge \operatorname{succ}\left(S_{a}, S_{b}\right) \rightarrow \operatorname{inFocus}\left(S_{b}, T_{2}\right) \wedge$
$\operatorname{dist}\left(\right.$ interpretation $\left.(+), S_{a}, S_{b}\right)=$ (interpretation $\left.(+), 1\right)$.
$\left(s_{47}\right) T_{1}<T_{2}, \forall n \in \mathbb{N}, n>1: \operatorname{inFocus}\left(S_{a}, T_{1}\right) \wedge \operatorname{currentSign}\left(+, T_{1}\right) \wedge$
$\operatorname{move}\left(S_{a}, n\right.$, interpretation $\left.(+), T_{1}\right) \wedge \operatorname{succ}\left(S_{a}, S_{b}\right) \rightarrow \operatorname{inFocus}\left(S_{b}, T_{2}\right) \wedge$
currentSign $\left(+, T_{2}\right) \wedge \operatorname{move}\left(S_{b}, n-1\right.$, interpretation $\left.(+), T_{2}\right) \wedge$
$\operatorname{dist}\left(\right.$ interpretation $\left.(+), S_{a}, S_{b}\right)=($ interpretation $(+), 1)$
$\left(s_{48}\right) T_{1}<T_{2}: \operatorname{inFocus}\left(S_{a}, T_{1}\right) \wedge$ currentSign $\left(-, T_{1}\right) \wedge$
move $\left(S_{a}, 1\right.$, interpretation $\left.(-), T_{1}\right) \wedge \operatorname{succ}\left(S_{b}, S_{a}\right) \rightarrow \operatorname{inFocus}\left(S_{b}, T_{2}\right) \wedge$
$\operatorname{dist}\left(\right.$ interpretation $\left.(-), S_{a}, S_{b}\right)=$ (interpretation $\left.(-), 1\right)$
$\left(s_{49}\right) T_{1}<T_{2}, \forall n \in \mathbb{N}, n>1: \operatorname{inFocus}\left(S_{a}, T_{1}\right) \wedge \operatorname{currentSign}\left(-, T_{1}\right) \wedge$
move $\left(S_{a}, n\right.$, interpretation $\left.(-), T_{1}\right) \wedge \operatorname{succ}\left(S_{b}, S_{a}\right) \rightarrow \operatorname{inFocus}\left(S_{b}, T_{2}\right) \wedge$
currentSign $\left(-, T_{2}\right) \wedge$ move $\left(S_{b}, n-1\right.$, interpretation $\left.(-), T_{2}\right)$
$\left(s_{50}\right) \operatorname{lower}\left(S_{a}, S_{b}\right), T_{1}<T_{2}, \forall n \in \mathbb{N}: \operatorname{dist}\left(\right.$ interpretation $\left.(+), S_{a}, S_{b}\right)=$
(interpretation $(+), n) \wedge \operatorname{inFocus}\left(S_{b}, T_{1}\right) \wedge$ currentSign $\left(+, T_{1}\right) \wedge$
$\operatorname{move}\left(S_{b}, 1\right.$, interpretation $\left.(+), T_{1}\right) \wedge \operatorname{succ}\left(S_{b}, S_{c}\right) \rightarrow \operatorname{inFocus}\left(S_{c}, T_{2}\right) \wedge$
$\operatorname{dist}\left(\right.$ interpretation $\left.(+), S_{a}, S_{c}\right)=($ interpretation $(+), n+1)$.
$\left(s_{51}\right)$ higher $\left(S_{a}, S_{b}\right), T_{1}<T_{2}, \forall n \in \mathbb{N}$ : dist(interpretation $\left.(-), S_{a}, S_{b}\right)=$
(interpretation $(-), n) \wedge \operatorname{inFocus}\left(S_{b}, T_{1}\right) \wedge$ currentSign $\left(-, T_{1}\right) \wedge$
move $\left(S_{b}, 1\right.$, interpretation $\left.(-), T_{1}\right) \wedge \operatorname{succ}\left(S_{c}, S_{b}\right) \rightarrow \operatorname{inFocus}\left(S_{c}, T_{2}\right) \wedge$
$\operatorname{dist}\left(\right.$ interpretation $\left.(-), S_{a}, S_{c}\right)=($ interpretation $(-), n+1)$.
$\left(s_{52}\right) \forall n \in \mathbb{N}: \operatorname{magn}\left(D_{1}, n\right) \rightarrow n$.
( $s_{53}$ ) lower $\left(S_{a}, S_{b}\right)$ : diff(interpretation( + ), $\left.S_{a}, S_{b}\right)=$
$\operatorname{diff}\left(\operatorname{interpretation}(-), S_{b}, S_{a}\right)=\operatorname{magn}\left(\operatorname{dist}\left(\right.\right.$ interpretation $\left.\left.(+), S_{a}, S_{b}\right)\right)$.
$\left(s_{54}\right) \nexists S_{a}$ : lower $\left(S_{a}, S_{b}\right) \rightarrow$ height $\left(S_{c}\right)=\operatorname{diff}\left(\right.$ interpretation $\left.(-), S_{c}, S_{b}\right)$
$\left(s_{55}\right)$ between $\left(C_{1}, S_{a}, S_{b}\right) \vee\left(\operatorname{inCircle}\left(C_{1}, S_{b}\right) \wedge \nexists S_{c}:\left(\right.\right.$ inCircle $\left(\right.$ Ci $\left._{1}, S_{c}\right) \wedge$
$l$
$\left.\left.\operatorname{lower}\left(S_{c}, S_{b}\right)\right)\right) \xrightarrow{\rightarrow} \operatorname{base}\left(S_{b}\right)$.
$\left(s_{56}\right) S_{a} \neq S_{b} \wedge \exists S_{c}, S_{d}:\left(\operatorname{base}\left(S_{c}\right) \wedge \operatorname{base}\left(S_{d}\right) \wedge \operatorname{dist}\left(\right.\right.$ interpretation $\left.(-), S_{a}, S_{c}\right)=$
dist $\left(\right.$ interpretation $\left.\left.(-), S_{b}, S_{d}\right)\right) \rightarrow \operatorname{analogs}\left(S_{a}\right)$
dist(interpretation $\left.\left.(-), S_{b}, S_{d}\right)\right) \rightarrow \operatorname{analogs}\left(S_{a}, S_{b}\right)$.

```

Table 1: Formalization of the Calculation Circular Staircase.
model of the Calculation Circular Staircase, whilst an idealized version (i.e., a version featuring complete addition and subtraction tables, which in reality should be assumed to be rather incomplete or sparse) of the students' initial conceptualization of the natural number domain can formally be represented as shown in Table 2.

We quickly want to focus on some aspects of the respective formalizations. The base domain of the later analogy, i.e., the formalization of the Calculation Circular Staircase,
```

Sorts:
number, sign, operation.
Entities:
one, two, ..., nineteen, zero : number.
,+- : sign
plus, minus : operation
Functions:
apply: operation $\times$ number $\times$ number $\rightarrow$ number.
interpretation : sign $\rightarrow$ operation.
Predicates:
Predicates:
succ : number $\times$ number.
Facts:
$\left(n_{1}\right) \operatorname{succ}($ zero, one).
( $n_{2}$ ) $\operatorname{succ}$ (one, two).
$\left(n_{3}\right) \operatorname{succ}($ two, three $)$.
$\left(n_{4}\right) \operatorname{succ}($ three, four).
$\left.\ddot{( }_{19}\right)$ succ(eighteen, nineteen).
$\left(n_{20}\right)$ apply(interpretation $(+)$, one, one $)=$ two.
$\left(n_{21}\right)$ apply (interpretation $(+)$, one, two $)=$ three.
$\left(n_{22}\right)$ apply (interpretation $(+)$, one, three $)=$ four.
$\left.\dddot{(n}_{37}\right)$ apply (interpretation $(+)$, one, eighteen $)=$ nineteen.
$\left.n_{38}\right)$ apply(interpretation $(+)$, two, one $)=$ three.
$\left(n_{39}\right)$ apply(interpretation $(+)$, two, two $)=$ four.
$\left(n_{54}\right)$ apply (interpretation $(+)$, two, seventeen $)=$ nineteen.
$\left(n_{55}\right)$ apply(interpretation $(+)$, three, one $)=$ four.
$\left(n_{190}\right)$ apply (interpretation $(+)$, eighteen, one $)=$ nineteen.
$n_{191}$ apply (interpretation $(-)$, two, one $)=$ one.
$\left(n_{192}\right)$ apply (interpretation $(-)$, three, one $)=$ two.
$\left(n_{193}\right)$ apply (interpretation $(-)$, three, two $)=$ one.
$\left(n_{360}\right)$ apply(interpretation $(-)$, nineteen, seventeen $)=t w o$
$\left(n_{361}\right)$ apply(interpretation $(-)$, nineteen, eighteen $)=$ one.
Laws
$\left(n_{362}\right)$ interpretation $(+)=$ plus
$\left(n_{363}\right)$ interpretation $(-)=$ minus.
$\left(n_{364}\right)$ interpretation $(+) \neq$ interpretation $(-)$.

```

Table 2: Formalization of an idealized form of the children's initial conception of the number domain.
exhibits a rich structure, both concerning facts and laws alike. The facts represent the easily accessible structure of the staircase, namely the order of succession of the steps, the distinction between the inner and the outer circle, and the placement of the door between steps nine and ten. The laws cover the transformational and constructive insights accessible to the children via interaction with the staircase: For instance ( \(s_{46}\) ) to ( \(s_{49}\) ) encompass the previously described process of having the toy figure move up or down the staircase, and ( \(s_{50}\) ) and \(\left(s_{51}\right)\) then add a counting principle keeping track of the number of steps passed by the figure (which in reality allows children to determine the distance the toy figure may move on the staircase). ( \(s_{52}\) ) and ( \(s_{53}\) ) serve for converting the distance measured in steps into a natural number, i.e., represent the children's mental process when realizing that distances on the staircase correspond to a more abstract number concept, i.e., are not bound to the individual stairs but can be generalized. Concerning the final two laws, \(\left(s_{55}\right)\) introduces the previously mentioned concept of singular steps in the Calculation Circular Staircase which are similar in that they form the base of one closed part of the staircase (namely of the inner or outer circle) or are marked by being preceded by the door as caesura, and ( \(s_{56}\) ) concludingly introduces the concept of structure-based analogs amongst the steps.

The formalization of the target domain of the later analogy, i.e., of an idealized version of the children's initial conception of the number domain, contains mostly facts the children have learned by heart, namely the order of the number terms
```

Sorts:
circle, caesura, time, sign, direction/operation, steps/number, natural/number.
Entities:
one, two,$\ldots$, nineteen, zero, $S_{a}, S_{b}, S_{c}, S_{d}:$ steps/number.
$O$ : direction/operation.
,+- : sign.
(*) door, $C_{1}$ : caesura.
(*) $T_{1}, T_{2}$ : time.
(*) inner, outer, $\mathrm{Ci}_{1}$ : circle.
Functions:
DiffApply: direction/operation $\times$ steps/number $\times$ steps/number $\rightarrow$ natural/number.
interpretation : sign $\rightarrow$ direction/operation.
(*) height : steps/number $\rightarrow$ natural/number.
(*) dist: direction/operation $\times$ steps/number $\times$ steps/number $\rightarrow$ direction/operation $\times$
natural/number.
(*) magn : direction/operation $\times$ natural/number $\rightarrow$ natural/number.
Predicates:
succ : steps/number $\times$ steps/number.
(*) higher : steps/number $\times$ steps/number.
*) lower : steps/number $\times$ steps/number.
(*) inCircle : circle $\times$ steps/number.
* base : steps/number.
*) inFocus : steps/number $\times$ time
*) currentSign $:$ sign $\times$ time.
*) between : caesura $\times$ steps/number $\times$ steps/number.
(*) move : steps/number $\times$ natural/number $\times$ direction/operation $\times$ time.
(*) analogs : steps/number $\times$ steps/number
Facts:
$\left(g_{1}\right) \operatorname{succ}(z e r o$, one $)$
$\left(g_{2}\right) \operatorname{succ}($ one, two).
$\left.\dddot{(g}_{19}\right)$ succ(eighteen, nineteen).
( $g_{20} *$ ) inCircle(inner, zero).
$\left(g_{21} *\right)$ inCircle(inner, one).
( $\left.g_{29} *\right)$ inCircle(inner, nine)
$\left(g_{30} *\right)$ inCircle(outer, ten).
$\left(g_{39} *\right)$ inCircle(outer, nineteen).
$\left(g_{40} *\right)$ between(door, nine, ten).
Laws:
$\left(g_{41}\right)$ interpretation $(+)=0$
$\left(g_{42}\right)$ interpretation $(-)=O$
$\left(g_{43}\right)$ interpretation $(+) \neq$ interpretation $(-)$.
$\left(g_{44 *}\right)$ higher $\left(S_{a}, S_{b}\right) \leftrightarrow \operatorname{succ}\left(S_{b}, S_{a}\right) \vee \exists S_{c}:\left(\operatorname{succ}\left(S_{b}, S_{c}\right) \wedge \operatorname{higher}\left(S_{a}, S_{c}\right)\right)$.
$\left(g_{44 *}^{*}\right)$ higher $\left(S_{a}, S_{b}\right) \leftrightarrow \operatorname{succ}\left(S_{b}, S_{a}\right)$
$\left(g_{45 *}\right)$ higher $\left(S_{a}, S_{b}\right) \leftrightarrow$ lower $\left(S_{b}, S_{a}\right)$.
$\left(g_{45} *\right)$ higher $\left(S_{a}, S_{b}\right) \leftrightarrow$ lower $\left(S_{b}, S_{a}\right)$.
$\left(g_{46} *\right)$
$T_{1}<T_{2}:$ inFocus $\left(S_{a}, T_{1}\right) \wedge$ currentSign $\left(+, T_{1}\right) \wedge$
$\left(g_{46} *\right) T_{1}<T_{2}: \operatorname{inFocus}\left(S_{a}, T_{1}\right) \wedge \operatorname{currentSign}\left(+, T_{1}\right) \wedge$
move $\left(S_{a}, 1\right.$, interpretation $\left.(+), T_{1}\right) \wedge \operatorname{succ}\left(S_{a}, S_{b}\right) \rightarrow \operatorname{inFocus}\left(S_{b}, T_{2}\right) \wedge$
move $\left(S_{a}, 1\right.$, interpretation $\left.(+), T_{1}\right) \wedge \operatorname{succ}\left(S_{a}, S_{b}\right) \rightarrow$ inF
dist $\left(\right.$ interpretation $\left.(+), S_{a}, S_{b}\right)=($ interpretation $(+), 1)$.
$\left(g_{47} *\right) T_{1}<T_{2}, \forall n \in \mathbb{N}, n>1: \operatorname{inFocus}\left(S_{a}, T_{1}\right) \wedge \operatorname{currentSign}\left(+, T_{1}\right) \wedge$
$\operatorname{move}\left(S_{a}, n\right.$, interpretation $\left.(+), T_{1}\right) \wedge \operatorname{succ}\left(S_{a}, S_{b}\right) \rightarrow \operatorname{inFocus}\left(S_{b}, T_{2}\right) \wedge$
currentSign $\left(+, T_{2}\right) \wedge \operatorname{move}\left(S_{b}, n-1\right.$, interpretation $\left.(+), T_{2}\right) \wedge$
$\operatorname{dist}\left(\right.$ interpretation $\left.(+), S_{a}, S_{b}\right)=($ interpretation $(+), 1)$
dist (interpretation $\left.(+), S_{a}, S_{b}\right)=$ (interpretation $\left.(+), 1\right)$.
$\left(g_{48} *\right) T_{1}<T_{2}: \operatorname{inFocus}\left(S_{a}, T_{1}\right) \wedge \operatorname{currentSign}\left(-, T_{1}\right) \wedge$
move $\left(S_{a}, 1\right.$, interpretation $\left.(-), T_{1}\right) \wedge \operatorname{succ}\left(S_{b}, S_{a}\right) \rightarrow \operatorname{inFocus}\left(S_{b}, T_{2}\right) \wedge$
move $\left(S_{a}, 1\right.$, interpretation $\left.(-), T_{1}\right) \wedge \operatorname{succ}\left(S_{b}, S_{a}\right) \rightarrow$ inFo
dist $\left(\right.$ interpretation $\left.(-), S_{a}, S_{b}\right)=($ interpretation $(-), 1)$.
$\operatorname{dist}\left(\right.$ interpretation $\left.(-), S_{a}, S_{b}\right)=$ (interpretation $\left.(-), 1\right)$.
$\left(g_{49} *\right) T_{1}<T_{2}, \forall n \in \mathbb{N}, n>1: \operatorname{inFocus}\left(S_{a}, T_{1}\right) \wedge \operatorname{currentSign}\left(-, T_{1}\right) \wedge$
$\left(g_{49} *\right) T_{1}<T_{2}, \forall n \in \mathbb{N}, n>1: \operatorname{inFocus}\left(S_{a}, T_{1}\right) \wedge \operatorname{currentSign}\left(-, T_{1}\right) \wedge$
move $\left(S_{a}, n\right.$, interpretation $\left.(-), T_{1}\right) \wedge \operatorname{succ}\left(S_{b}, S_{a}\right) \rightarrow \operatorname{inFocus}\left(S_{b}, T_{2}\right) \wedge$
$\operatorname{move}\left(S_{a}, n\right.$, interpretation $\left.(-), T_{1}\right) \wedge \operatorname{succ}\left(S_{b}, S_{a}\right) \rightarrow \operatorname{inFocus}\left(S_{b}, T_{2}\right) \wedge$
currentSign $\left(-, T_{2}\right) \wedge \operatorname{move}\left(S_{b}, n-1\right.$, interpretation $\left.(-), T_{2}\right)$.
$\left(g_{50} *\right)$ lower $\left(S_{a}, S_{b}\right), T_{1}<T_{2}, \forall n \in \mathbb{N}: \operatorname{dist}\left(\right.$ interpretation $\left.(+), S_{a}, S_{b}\right)=$
(interpretation $(+), n) \wedge \operatorname{inFocus}\left(S_{b}, T_{1}\right) \wedge$ currentSign $\left(+, T_{1}\right) \wedge$
move $\left(S_{b}, 1\right.$, interpretation $\left.(+), T_{1}\right) \wedge \operatorname{succ}\left(S_{b}, S_{c}\right) \rightarrow \operatorname{inFocus}\left(S_{c}, T_{2}\right) \wedge$
move $\left(S_{b}, 1\right.$, interpretation $\left.(+), T_{1}\right) \wedge \operatorname{succ}\left(S_{b}, S_{c}\right) \rightarrow$ inFocus
$\operatorname{dist}\left(\right.$ interpretation $\left.(+), S_{a}, S_{c}\right)=($ interpretation $(+), n+1)$.
$\operatorname{dist}\left(\right.$ interpretation $\left.(+), S_{a}, S_{c}\right)=$ (interpretation $\left.(+), n+1\right)$.
$\left(g_{51} *\right) \operatorname{higher}\left(S_{a}, S_{b}\right), T_{1}<T_{2}, \forall n \in \mathbb{N}: \operatorname{dist}\left(\right.$ interpretation $\left.(-), S_{a}, S_{b}\right)=$
$\left(g_{51} *\right)$ higher $\left(S_{a}, S_{b}\right), T_{1}<T_{2}, \forall n \in \mathbb{N}: \operatorname{dist}($ interpretation $(-)$,
interpretation $(-), n) \wedge \operatorname{inFocus}\left(S_{b}, T_{1}\right) \wedge$ currentSign $\left(-, T_{1}\right) \wedge$
(interpretation $(-), n) \wedge \operatorname{inFocus}\left(S_{b}, T_{1}\right) \wedge \operatorname{currentSign}\left(-, T_{1}\right) \wedge$
move $\left(S_{b}, 1\right.$, interpretation $\left.(-), T_{1}\right) \wedge \operatorname{succ}\left(S_{c}, S_{b}\right) \rightarrow \operatorname{inFocus}\left(S_{c}, T_{2}\right) \wedge$
$\operatorname{move}\left(S_{b}, 1\right.$, interpretation $\left.(-), T_{1}\right) \wedge \operatorname{succ}\left(S_{c}, S_{b}\right) \rightarrow$ inFocus
$\operatorname{dist}\left(\right.$ interpretation $\left.(-), S_{a}, S_{c}\right)=($ interpretation $(-), n+1)$.
$\left(g_{52} *\right) \forall n \in \mathbb{N}: \operatorname{magn}\left(D_{1}, n\right) \rightarrow n$.
$\left(g_{53} *\right)$ lower $\left(S_{a}, S_{b}\right):$ DiffApply(interpretation $\left.(+), S_{a}, S_{b}\right)=$
$\operatorname{DiffApply}\left(\right.$ interpretation $\left.(-), S_{b}, S_{a}\right)=\operatorname{magn}\left(\operatorname{dist}\left(\right.\right.$ interpretation $\left.\left.(+), S_{a}, S_{b}\right)\right)$
$\left(g_{54} *\right) \nexists S_{a}:$ lower $\left(S_{a}, S_{b}\right) \rightarrow$ height $\left(\bar{S}_{c}\right)=$ DiffApply (interpretation $\left.(-), S_{c}, S_{b}\right)$.
$\left(g_{54} *\right) \nexists S_{a}:$ lower $\left(S_{a}, S_{b}\right) \rightarrow$ height $\left(S_{c}\right)=\operatorname{DiffApply}\left(\right.$ interpretation $\left.(-), S_{c}, S_{b}\right)$
$\left(g_{55 *}\right)$ between $\left(C_{1}, S_{a}, S_{b}\right) \vee\left(\operatorname{inCircle}\left(\operatorname{Ci}_{1}, S_{b}\right) \wedge \nexists S_{c}:\left(\operatorname{inCircle}\left(C i_{1}, S_{c}\right) \wedge\right.\right.$
$\left(g_{55} *\right)$ between $\left(C_{1}, S_{a}, S_{b}\right)$
lower $\left.\left.\left(S_{c}, S_{b}\right)\right)\right) \rightarrow \operatorname{base}\left(S_{b}\right)$
$\left.\operatorname{lower}\left(S_{c}, S_{b}\right)\right) \rightarrow \operatorname{base}\left(S_{b}\right)$.
$\left(g_{56} *\right) S_{a} \neq S_{b} \wedge \exists S_{c}, S_{d}:\left(\operatorname{base}\left(S_{c}\right) \wedge \operatorname{base}\left(S_{d}\right) \wedge \operatorname{dist}\left(\right.\right.$ interpretation $\left.(-), S_{a}, S_{c}\right)=$
dist $\left(\right.$ interpretation $\left.\left.(-), S_{b}, S_{d}\right)\right) \rightarrow$ analogs $\left(S_{a}, S_{b}\right)$.
dist(interpretation $\left.\left.(-), S_{b}, S_{d}\right)\right) \rightarrow$ analogs $\left(S_{a}, S_{b}\right)$.

```

Table 3: Generalized theory of the Calculation Circular Staircase and the children's number domain, already expanded by the generalized forms of the candidate elements for analogical transfer from base to target domain (marked with \(*\) ).
between zero and nineteen, and addition and subtraction tables within this range. In reality it has to be assumed that the addition and subtraction tables are significantly more sparsely populated than in our formalization, corresponding to incomplete recall of the memorized full tables.

The HDTP mechanism can now be used for computing a
common generalization of both domains, yielding a generalized theory like given in Table 3. The main domain elements defining the alignment of formulae are the matching between the entities of sort steps and number, between the functions diff and apply, the alignment of the respective sign entities, as well as the matching between the direction and operation entities (induced by the alignment of the respective interpretation functions). Here it has to be noted that in order to analogically match the two domains it is not only necessary to generalize facts and laws but in this case also sorts have to be generalized, for two sorts yielding the least general supersort. This is needed, for instance, when pairing up the representation of the staircase's steps (conceived as mere pillars) and the number terms known to the children.

In conclusion, the generalized theory forms the basis for transferring knowledge in an analogy-based way from the (originally richer) Calculation Circular Staircase domain to the children's number domain, resulting in an expanded theory for the numbers as given in Table 4. The important aspect in this expanded version of the domain is the availability of the constructive relations and insights obtained in the interaction with the Calculation Circular Staircase, e.g., providing a means to give meaning to the number terms via the assignment of the corresponding natural number values (using the diff function in \(\left(e_{381} *\right)\) ) or via laws ( \(\left.e_{373} *\right)\) to ( \(\left.e_{378} *\right)\) allowing for the independent computation of parts of the addition and subtraction table that might not be obtainable from memory (i.e., that would not explicitly be present as a fact in a more realistic formalization of the number domain).

\section*{Related Work and Conclusion}

We are not the first to consider the use of formal models and computational analogy-making systems in the context of education and teaching-related topics. Among many others, for example in (Thagard, Cohen, \& Holyoak, 1989), the authors present a theory and implementation of analogical mapping that applies to explanations of unfamiliar phenomena, and (Siegler, 1989) briefly conjectures how the StructureMapping Engine (SME) (Falkenhainer, Forbus, \& Gentner, 1989) as a prototypical analogy-engine could be used to gain insights about developmental aspects of analogy use. General cognitive theories of analogical reasoning and associated computation models are also highly relevant to analogies as learning mechanisms. These include Gentner's structure mapping theory and engine, in which relations between objects are preserved (and relations which contribute to higher order predicates are mapped preferentially), and attributes of objects are not mapped (Gentner, 1983); Holyoak and Thagard's multi-constraint theory, in which mappings are evaluated according to constraints of structural consistency, pragmatic centrality and semantic similarity (Holyoak \& Thagard, 1997); and Hummel and Holyoak's theory of analogy formation, which integrates memory access and structural mapping, implemented in LISA (Hummel \& Holyoak, 2003) (see (Gentner et al., 2001) for a review of computational models of
```

Sorts:
circle, caesura, time, sign, operation, number, natural.
Entities:
one, two, ..., nineteen, zero, $S_{a}, S_{b}, S_{c}, S_{d}$ : number.
,+- : sign.
plus, minus : operation
*) door, $C_{1}$ : caesura.
*) $T_{1}, T_{2}$ : time.
(*) inner, outer, $\mathrm{Ci}_{1}$ : circle.
Functions:
apply: operation $\times$ number $\times$ number $\rightarrow$ number.
interpretation : sign $\rightarrow$ operation.
$(*)$ diff: operation $\times$ number $\times$ number $\rightarrow$ natural.
(*) height: number $\rightarrow$ natural.
(*) dist: operation $\times$ number $\times$ number $\rightarrow$ operation $\times$ natural.
(*) magn : operation $\times$ natural $\rightarrow$ natural.
(*) interpretation : sign $\rightarrow$ operation.
Predicates:
succ : number $\times$ number.
(*) higher: number $\times$ number.
(*) lower: number $\times$ number.
** inCircle: circle $\times$ number.
* base : number.
(*) inFocus : number $\times$ time.
(*) currentSign $:$ sign $\times$ time.
*) between : caesura $\times$ number $\times$ number.
(*) move : number $\times$ natural $\times$ operation $\times$ time.
(*) analogs : number $\times$ number.
Facts:
( $e_{1}$ ) $\operatorname{succ}($ zero, one).
( $e_{2}$ ) $\operatorname{succ}($ one, two).
$\left(e_{19}\right) \operatorname{succ}($ eighteen, nineteen).
$\left(e_{20}\right)$ apply(interpretation $(+)$, one , one $)=$ two.
$\left(e_{21}\right)$ apply (interpretation $(+)$, one, two $)=$ three.
$\left(e_{190}\right) \operatorname{apply}($ interpretation $(+)$, eighteen, one $)=$ nineteen.
$\left(e_{191}\right)$ apply (interpretation $(-)$, two, one $)=$ one
$\left(e_{192}\right) \operatorname{apply}($ interpretation $(-)$, three, one $)=t w o$.
$\left(e_{193}\right)$ apply (interpretation $(-)$, three, two $)=$ one.
$\left(e_{360}\right)$ apply(interpretation $(-)$, nineteen, seventeen $)=$ two
$\left(e_{361}\right)$ apply(interpretation $(-)$, nineteen, eighteen $)=$ one.
(e $e_{362} *$ ) inCircle(inner, zero).
$\left(e_{363} *\right)$ inCircle(inner, one).
$\left(e_{364} *\right)$ inCircle(inner, nine).
(e365*) inCircle(outer, ten).
( $\left.e_{366} *\right)$ inCircle(outer, nineteen).
$\left(e_{367} *\right)$ between (door, nine, ten).
Laws:
$\left(e_{368}\right)$ interpretation $(+)=$ plus.
$\left(e_{369}\right)$ interpretation $(-)=$ minus.
$\left(e_{370}\right)$ interpretation $(+) \neq$ interpretation $(-)$.
$\left(e_{371} *\right)$ higher $\left(S_{a}, S_{b}\right) \leftrightarrow \operatorname{succ}\left(S_{b}, S_{a}\right) \vee \exists S_{c}:\left(\operatorname{succ}\left(S_{b}, S_{c}\right) \wedge \operatorname{higher}\left(S_{a}, S_{c}\right)\right)$
$\left(e_{372} *\right)$ higher $\left(S_{a}, S_{b}\right) \leftrightarrow$ lower $\left(S_{b}, S_{a}\right)$.
$\left(e_{373} *\right) T_{1}<T_{2}: \operatorname{inFocus}\left(S_{a}, T_{1}\right) \wedge \operatorname{currentSign}\left(+, T_{1}\right) \wedge$
move $\left(S_{a}, 1\right.$, interpretation $\left.(+), T_{1}\right) \wedge \operatorname{succ}\left(S_{a}, S_{b}\right) \rightarrow \operatorname{inFocus}\left(S_{b}, T_{2}\right) \wedge$
$\operatorname{dist}\left(\right.$ interpretation $\left.(+), S_{a}, S_{b}\right)=($ interpretation $(+), 1)$.
$\left(e_{374} *\right) T_{1}<T_{2}, \forall n \in \mathbb{N}, n>1: \operatorname{inFocus}\left(S_{a}, T_{1}\right) \wedge \operatorname{currentSign}\left(+, T_{1}\right) \wedge$
$\operatorname{move}\left(S_{a}, n\right.$, interpretation $\left.(+), T_{1}\right) \wedge \operatorname{succ}\left(S_{a}, S_{b}\right) \rightarrow \operatorname{inFocus}\left(S_{b}, T_{2}\right) \wedge$
currentSign $\left(+, T_{2}\right) \wedge$ move $\left(S_{b}, n-1\right.$, interpretation $\left.(+), T_{2}\right) \wedge$
dist (interpretation $\left.(+), S_{a}, S_{b}\right)=$ (interpretation $\left.(+), 1\right)$.
$\left(e_{375} *\right) T_{1}<T_{2}: \operatorname{inFocus}\left(S_{a}, T_{1}\right) \wedge \operatorname{currentSign}\left(-, T_{1}\right) \wedge$
move $\left(S_{a}, 1\right.$, interpretation $\left.(-), T_{1}\right) \wedge \operatorname{succ}\left(S_{b}, S_{a}\right) \rightarrow \operatorname{inFocus}\left(S_{b}, T_{2}\right) \wedge$
$\operatorname{dist}\left(\right.$ interpretation $\left.(-), S_{a}, S_{b}\right)=($ interpretation $(-), 1)$.
$\left(e_{376} *\right) T_{1}<T_{2}, \forall n \in \mathbb{N}, n>1: \operatorname{inFocus}\left(S_{a}, T_{1}\right) \wedge \operatorname{currentSign}\left(-, T_{1}\right) \wedge$
$\operatorname{move}\left(S_{a}, n\right.$, interpretation $\left.(-), T_{1}\right) \wedge \operatorname{succ}\left(S_{b}, S_{a}\right) \rightarrow \operatorname{inFocus}\left(S_{b}, T_{2}\right) \wedge$
currentSign $\left(-, T_{2}\right) \wedge \operatorname{move}\left(S_{b}, n-1\right.$, interpretation $\left.(-), T_{2}\right)$.
$\left(e_{377} *\right)$ lower $\left(S_{a}, S_{b}\right), T_{1}<T_{2}, \forall n \in \mathbb{N}$ : dist(interpretation $\left.(+), S_{a}, S_{b}\right)=$
(interpretation $(+), n) \wedge \operatorname{inFocus}\left(S_{b}, T_{1}\right) \wedge$ currentSign $\left(+, T_{1}\right) \wedge$
move $\left(S_{b}, 1\right.$, interpretation $\left.(+), T_{1}\right) \wedge \operatorname{succ}\left(S_{b}, S_{c}\right) \rightarrow \operatorname{inFocus}\left(S_{c}, T_{2}\right) \wedge$
$\operatorname{dist}\left(\right.$ interpretation $\left.(+), S_{a}, S_{c}\right)=($ interpretation $(+), n+1)$.
$\left(e_{378} *\right) \operatorname{higher}\left(S_{a}, S_{b}\right), T_{1}<T_{2}, \forall n \in \mathbb{N}: \operatorname{dist}\left(\right.$ interpretation $\left.(-), S_{a}, S_{b}\right)=$
(interpretation $(-), n) \wedge \operatorname{inFocus}\left(S_{b}, T_{1}\right) \wedge$ currentSign $\left(-, T_{1}\right) \wedge$
$\operatorname{move}\left(S_{b}, 1\right.$, interpretation $\left.(-), T_{1}\right) \wedge \operatorname{succ}\left(S_{c}, S_{b}\right) \rightarrow \operatorname{inFocus}\left(S_{c}, T_{2}\right) \wedge$
$\operatorname{dist}\left(\right.$ interpretation $\left.(-), S_{a}, S_{c}\right)=($ interpretation $(-), n+1)$.
dist $\left(\right.$ interpretation $\left.(-), S_{a}, S_{c}\right)=(n$.
$\left(e_{379} *\right) \forall n \in \mathbb{N}: \operatorname{magn}\left(D_{1}, n\right) \rightarrow n$.
$\left.\begin{array}{l}\left(e_{379}\right) \\ \left(e_{380}\right)\end{array}\right) \operatorname{lower}\left(S_{a}, S_{b}\right):$ diff(interpretation $\left.(+), S_{a}, S_{b}\right)=$
$\left(e_{380 *}^{*}\right)$ lower $\left(S_{a}, S_{b}\right): \operatorname{diff}\left(\right.$ interpretation $\left.(+), S_{a}, S_{b}\right)=$
diff(interpretation $\left.(-), S_{b}, S_{a}\right)=\operatorname{magn}(\operatorname{dist}($ interpretation
$\operatorname{diff}\left(\right.$ interpretation $\left.(-), S_{b}, S_{a}\right)=\operatorname{magn}\left(\operatorname{dist}\left(\right.\right.$ interpretation $\left.\left.(+), S_{a}, S_{b}\right)\right)$
$\left(e_{381 *}^{*} \nexists S_{a}: \operatorname{lower}\left(S_{a}, S_{b}\right) \rightarrow\right.$ height $\left(S_{c}\right)=\operatorname{diff}\left(\right.$ interpretation $\left.(-), S_{c}, S_{b}\right)$
$\left(e_{382} *\right)$ between $\left(C_{1}, S_{a}, S_{b}\right) \vee\left(\operatorname{inCircle}\left(\operatorname{Ci}_{1}, S_{b}\right) \wedge \nexists S_{c}:\left(\operatorname{inCircle}\left(\operatorname{Ci}_{1}, S_{c}\right) \wedge\right.\right.$
lower $\left.\left.\left(S_{c}, S_{b}\right)\right)\right) \rightarrow \operatorname{base}\left(S_{b}\right)$.
$\left(e_{383 *}\right) S_{a} \neq S_{b} \wedge \exists S_{c}, S_{d}:\left(\operatorname{base}\left(S_{c}\right) \wedge \operatorname{base}\left(S_{d}\right) \wedge \operatorname{dist}\left(\right.\right.$ interpretation $\left.(-), S_{a}, S_{c}\right)=$
$\operatorname{dist}\left(\right.$ interpretation $\left.\left.(-), S_{b}, S_{d}\right)\right) \rightarrow \operatorname{analogs}\left(S_{a}, S_{b}\right)$.

```

Table 4: Analogically enriched formalization of the idealized version of the children's conception of the number domain.
analogy). Our HDTP model of the Calculation Circular Staircase is intended to complement such approaches, showing how formal methods and computational accounts of analogymaking can be used to gain additional insights in the inner workings of analogy-based educational methods and tools. By providing a detailed formal description of the involved domains, also sketching how the domains relate to each other in terms of their joint generalization and how this relation can be used to transfer knowledge from the staircase domain into the number domain, we managed to explicate the structural relations and governing laws underlying the Calculation Circular Staircase as teaching model of the natural number domain, and to point out how the identified constructive and transformation-based conceptualizations then also can provide additional support and a deeper-rooted model for the childrens' initially very flat and sparse conception of the number domain.

We see this work as a first step towards the design of analogy-based teaching material, both specifically in arithmetic and, more generally, in mathematics and other disciplines. Modelling educational analogies provides another perspective on a particular analogy, in terms of which information is transferred, what the limitations of the analogy are, or whether it makes unhelpful mappings; and what potential extensions to the analogy it suggests. We envisage that our model of the Calculation Circular Staircase can be used in order to design a lesson plan on the natural number domain. Its usefulness would then be evaluated via empirical studies on the students, testing the depth of their understanding of cardinal and ordinal numbers and basic operations, and whether the students' understanding (and misunderstanding) mirrors the inferences made by the model.

\section*{References}

Akgul, E. (2006). Teaching Science In An Inquity-Based Learning Environment: What It Means For Pre-Service Elementary Science Teachers. Eurasia Journal of Mathematics, Science and Technology Education, 2(1), 71-81.
Arnold, M., \& Millar, R. (1996). Exploring the use of analogy in the teaching of heat, temperature and thermal equilibrium. In G. Welford, J. Osborne, \& P. Scott (Eds.), Research in Science Education in Europe: Current Issues and Themes. London: Farmer Press.
Besold, T. R. (2013). Analogy Engines in Classroom Teaching: Modeling the String Circuit Analogy. In AAAI Spring 2013 Symposium on Creativity and (Early) Cognitive Development.
Brainerd, C. (1979). The origins of the number concept. New York: Praeger.
Clement, J. (1993). Using bridging analogies and anchoring intuitions to deal with students' preconceptions in physics. Journal of Research in Science Teaching, 30, 1241-1257.
Dedekind, R. (1887/1969). Was sind und was sollen die Zahlen? 2. Ndr.. der 10. Aufl. Braunschweig: Vieweg.

Duit, R. (1991). The role of analogies and metaphors in learning science. Science Education, 75(6), 649-672.
English, L. D. (Ed.). (1997). Mathematical reasoning: Analogies, metaphors and images (studies in mathematical thinking and learning series). Lawrence Erlbaum Assoc.
Falkenhainer, B., Forbus, K., \& Gentner, D. (1989). The Structure-Mapping Engine: Algorithm and Examples. Artificial Intelligence, 41, 1-63.
Gentner, D. (1983). Structure-mapping: a theoretical framework for analogy. Cognitive Science(7), 155-170.
Gentner, D., Holyoak, K., \& Kokinov, B. (Eds.). (2001). The analogical mind: Perspectives from cognitive science. Cambridge, MA: MIT Press.
Glynn, S., Duit, R., \& Thiele, R. (1995). Teaching science with analogies: A strategy for constructing knowledge. In S. Glynn \& R. Duit (Eds.), Learning science in the schools: Research reforming practice. Mahwah, NJ: Lawrence Erlbaum Assoc.
Guerra-Ramos, M. (2011). Analogies as Tools for Meaning Making in Elementary Science Education: How Do They Work in Classroom Settings? Eurasia Journal of Mathematics, Science and Technology Education, 7(1), 29-39.
Holyoak, K., \& Thagard, P. (1995). Mental leaps: Analogy in creative thought. Cambridge, MA: MIT Press.
Holyoak, K., \& Thagard, P. (1997). The analogical mind. American Psychologist.
Hummel, J. E., \& Holyoak, K. J. (2003). A symbolicconnectionist theory of relational inference and generalization. Psychological Review, 110, 220-264.
Novick, L. R. (1988). Analogical transfer, problem similarity, and expertise. Journal of Experimental psychology: Learning, Memory, and Cognition, 14, 510-520.
Pimm, D. (1981). Metaphor and analogy in mathematics. For the Learning of Mathematics, 1(3), 47-50.
Schwank, I. (2003). Einführung in prädikatives und funktionales Denken. Zentralblatt Didaktik der Mathematik, 35(3), 70-78.
Schwank, I., Aring, A., \& Blocksdorf, K. (2005). Betreten erwünscht - die Rechenwendeltreppe. In Beiträge zum Mathematikunterricht. Hildesheim: Franzbecker.
Schwank, I., Gelfman, E., \& Nardi, E. (1999). Mathematical Thinking and Learning as Cognitive Processes. In I. Schwank (Ed.), Proc. of the First Conference of the European Society for Research in Mathematics Education (pp. 16-24).
Schwering, A., Krumnack, U., Kühnberger, K.-U., \& Gust, H. (2009). Syntactic principles of Heuristic-Driven Theory Projection. Journal of Cognitive Systems Research, 10(3), 251-269.
Siegler, R. (1989). Mechanisms of Cognitive Development. Annual Review of Psychology, 40, 353-379.
Thagard, P., Cohen, D., \& Holyoak, K. (1989). Chemical Analogies: Two Kinds of Explanation. In Proc. of the 11th International Joint Conference on Artificial Intelligence (pp. 819-824).

\title{
The Dynamics of Anchoring in Bidirectional Associative Memory Networks
}

\author{
Sudeep Bhatia (sudeepb@andrew.cmu.edu) \\ Department of Social \& Decision Sciences, Carnegie Mellon University, 5000 Forbes Ave. Pittsburgh, PA 15232 USA
}

\author{
Shereen J. Chaudhry (sjchaudh@andrew.cmu.edu) \\ Department of Social \& Decision Sciences, Carnegie Mellon University, 5000 Forbes Ave. Pittsburgh, PA 15232 USA
}

\begin{abstract}
We formalize the biased activation theory of anchoring using a bidirectional associative memory network. Anchors determine the starting state of this network. As the network settles, we show that the nodes representing numerical responses activate and deactivate consecutively, generating sequential adjustment. By demonstrating that anchoring as adjustment emerges naturally from the dynamics of the biased activation process, we are able to unify the two main theories of the anchoring effect, and subsequently provide a parsimonious explanation for a large range of findings regarding anchoring, and its determinants. Although we focus largely on phenomena related to anchoring, the results of this paper apply equivalently to all judgments under the influence of bidirectional processing, including those involving constraint satisfaction.
\end{abstract}

Keywords: Decision Making, Neural Networks, Dynamic Processes, Anchoring Effect, Constraint Satisfaction

\section*{Introduction}

Anchors have a powerful effect on human judgment. Responses to simple questions involving magnitude or time are systematically affected by uninformative numbers, known as anchors, displayed to the decision maker prior to the judgment task. High anchors generate high responses, low anchors generate low responses, and final judgments can be manipulated by selecting the appropriate anchor.

The anchoring effect has been shown to emerge in a large number of domains, and is one of the best studied judgment biases in psychology. Yet despite its importance, the cognitive mechanisms responsible for the anchoring effect are still being debated. In their seminal paper on heuristic choice, Tversky and Kahneman (1974) proposed that anchoring is caused by an imperfect sequential adjustment process. At each step in this process, decision makers evaluate the validity of a particular response. The judgment process terminates if the response in consideration is adequate; otherwise it moves on to the next feasible value. Anchors determine the starting point in this process, and adjustment is insufficient. Subsequently responses are closer to the anchor than optimal.

This explanation for the anchoring effect has been popular for many decades, and formal models of the anchoring effect have assumed that anchoring operates through sequential adjustment (Johnson \& Busemeyer, 2005, but see also Choplin \& Tawney, 2010). A more recent approach, however, claims that anchoring is the product of biased
activation (Chapman and Johnson, 1994, 1999; Mussweiler \& Strack, 1999). Anchors, according to this view, increase the accessibility of cues supporting the anchor. This evidence subsequently generates final responses that are closer to the anchor than optimal.

Is anchoring caused by sequential adjustment or biased activation? Both theories are supported by a large number of empirical findings (discussed in later sections), but neither is able to predict all of these findings by itself. In this paper we provide a simple answer to this question. We show that these processes are not necessarily distinct: sequential adjustment emerges from the dynamics of biased activation. Anchoring, thus, is caused by both these mechanisms simultaneously, and a large range of findings regarding anchoring and its moderators, can be explained within a unitary, parsimonious, theoretical framework.

\section*{Bidirectional Associative Memory}

Consider a very simple judgment task. The decision maker is asked to select one of \(N\) responses based on \(M\) cues stored in memory. We assume, for simplicity, that the relationship between the responses and the cues is binary, with each cue either supporting or opposing each response. We can write a response \(i\) as \(r_{i}\), and a cue \(j\) as \(c_{j}\). If \(c_{j}\) supports \(r_{i}\) then we can write \(s_{i j}=+1\), and if it opposes \(r_{i}\) then we can write \(s_{i j}=-1\).

These responses can be numeric, as in typical anchoring tasks, or non-numeric as in more general judgment tasks. For numeric responses, we assume that the \(N\) nodes are ordered in a sequence \(r_{1}, r_{2}, \ldots, r_{N}\), corresponding to the sequence of available responses. For example, when considering the percentage of African countries in the United Nations, with responses in intervals of \(1 \%, r_{1}, r_{2}, \ldots\), \(r_{100}\) correspond to the responses \(1 \%, 2 \%, \ldots, 100 \%\).

We can implement this structure in a two layer neural network, with the first layer consisting of \(M\) nodes representing the \(M\) different cues, and the second layer consisting of \(N\) nodes representing the \(N\) response options. The activation of the node corresponding to \(c_{j}\), at time \(t\), can be written as \(C_{j}(t)\), and the activation of the node corresponding to \(r_{i}\), at time \(t\), can be written as \(R_{i}(t)\).

The connections from the cue layer to the response layer are equal to the strength of support provided by the cues to the responses. As activated response options (such as anchors) also affect the activation of the available cues, these connections can be assumed to be recurrent. Hence the connections from \(c_{j}\) to \(r_{i}\) and from \(r_{i}\) to \(c_{j}\) are both simply \(s_{i j}\).

At a given time \(t\), the activated nodes in the response layer first send inputs, weighted by \(s_{i j}\), into the cue layer. This affects the activation of the nodes in the cue layer. The activated nodes in the cue layer subsequently send inputs weighted by \(s_{i j}\) into the response layer, affecting the activation of the response nodes at \(t+1\), at which point the process repeats.

In addition to the inputs from the response layer, we assume that the nodes in the cue layer receive constant exogenous inputs with strength \(I=1\). These inputs ensure that evidence nodes are activated even when none of the response nodes are active, and that the judgment process can begin in the absence of a response bias. We also assume that all of the nodes in our network have the same binary activation function, with a threshold at zero. With this assumption we can write the activation functions of any \(r_{i}\) as \(R_{i}(t)=H\left[q_{i}\right]\), and any \(c_{j}\) as \(C_{j}(t)=H\left[b_{j}\right]\) such that \(q_{i}=\sum s_{i j} C_{j}(t-\) 1), \(b_{j}=\sum s_{i j} \cdot R_{i}(t)+1\), and \(H\) as the unit step function with \(H[x]=1\) for \(x>0\) and \(H[x]=0\) for \(x \leq 0\).

We can now formalize the effect anchors have on the judgment process. We assume that anchors determine the starting state of the network. Hence if \(r_{i}\) is the anchor, then at \(t=1\), we have \(R_{i}(1)=1\), and \(R_{k}(1)=0\) for \(k \neq i\). In the absence of an anchor, the network begins with \(R_{k}(1)=0\) for all \(k\). Finally, we assume that responses active once the network stabilizes are the ones that are selected, and that the response time is proportional to the time it takes for the network to settle.

The proposed network is motivated primarily by the memory structure assumed to be at play in anchoring and related judgment tasks: indeed, it is one of the simplest possible cognitive instantiations of the biased activation theory of anchoring, which posits a recurrent relationship between cues and responses. That said, this network is ultimately a special case of the bidirectional associative memory (BAM) network, introduced in Kosko (1988). BAM itself generalizes the Hopfield network, which BAM resembles when node updating is asynchronous.


Figure 1: The BAM network.

\section*{Activation and Stability}

What determines the responses that get activated at any time period, in the BAM network? The answer is cue overlap. Assume that only \(r_{i}\) is activated at time \(t\). This activation causes only the cues that support \(r_{i}\) to be activated at \(t\). Intuitively, the decision maker focuses on the cues that support the activated response and suppresses the cues that oppose the activated response. Once these cue nodes are activated, the activation pattern in the response layer
changes. At \(t+1\), responses supported by most of the cues activated at \(t\) turn on. These include \(r_{i}\), but also other novel responses, that overlap sufficiently with \(r_{i}\) in cue support. Eventually at \(t+2\) these responses activate other responses that they overlap with, and this process continues until the network stabilizes. Stability is always guaranteed: any BAM network with any memory structure, starting at any point, will stabilize in a finite number of time steps (Kosko, 1988).

\section*{Defining Sequential Adjustment}

We hope to show that this settling process of the BAM network in the presence of anchors resembles sequential adjustment. Before we can do this, however, we need to understand what sequential adjustment really is. Sequential adjustment is generally defined as the successive movement through the range of responses available to the decision maker. In the simplest case, this definition imposes a form of serial processing, according to which only one response is considered at any given time. For example, when judging the proportion of African countries in the U.N., decision makers may first consider 1\%. After rejecting this response they would consider \(2 \%\). If this too is inadequate they would move on to \(3 \%\), and so on. We consider the more general (and more realistic) case in which multiple responses can be considered at the same time. This allows decision makers to focus on all the responses within a particular interval, such as \(1-10 \%\), simultaneously, before moving on to the next interval in the sequence.

Such a dynamic is compatible with the general idea underlying sequential adjustment, as long as the responses activated are contiguous. Sequential adjustment does not permit the simultaneous consideration of different, nonneighboring responses. For example decision makers who consider both \(1 \%\) and \(99 \%\) simultaneously, without considering the responses between these two numbers, would not appear to be displaying sequential adjustment.

This then allows us to formalize the first requirement for sequential adjustment. This requirement, titled contiguous activation, states that sequential adjustment must not involve the simultaneous activation of multiple nonneighboring responses. Responses must be considered individually or in contiguous intervals.

Settling dynamics that display contiguous activation do not necessarily resemble sequential adjustment. It is possible for the decision maker to consider responses in contiguous intervals at any given time, but transition across different intervals in a non-sequential manner. For example, when evaluating the proportion of African countries in the U.N., decision makers could begin by considering the interval \(1-10 \%\), and then move to the interval \(20-30 \%\), without considering the interval \(10-20 \%\).

We thus need an additional requirement for our definition of sequential adjustment, in order to rule out these types of dynamics. This requirement, titled sequential transitions, states that sequential adjustment must not involve changes in activation that skip over a set of responses. Changes to response activation must be successive.

\section*{Connected Memory}

Do the dynamics of the anchored BAM network satisfy contiguous activation and sequential transition? Not necessarily. However with a simple assumption about the underlying memory structure, these requirements can indeed be satisfied. This assumption relates to the distribution of cue support for the responses. In numeric judgments, cues can seldom support two disparate responses without supporting intermediate responses. For example, when judging the proportion of African countries in the UN, any cue that supports the \(10 \%\) response, and the \(12 \%\) response, should, in general, support the intermediate \(11 \%\) response. This property, titled connectedness, more formally requires that a cue that supports \(r_{i}\) and \(r_{k}\), also supports \(r_{l}\) for \(i<l<k\). Memory structures displaying this property involve cues with a single, connected, interval of supported responses, where as those that do not display this property have cues with multiple, fragmented, intervals of supported responses.

While connectedness may not be satisfied in all judgment tasks, it is certainly a reasonable assumption when responses are ordered, as with the numerical scales used in anchoring tasks. Cues in these settings generally provide support for "large" responses, or "small" responses, or "medium" responses, or some other connected interval of responses. Very few cues provide support for a set of nonneighboring responses, distributed sporadically across the response scale. Indeed it is quite difficult to think of memory structures with diagnostic cues for numerical responses that do not satisfy the connectedness property.

\section*{The Emergence of Sequential Adjustment}

When memory structures satisfy connectedness, then the resulting BAM network, with the anchored response activated at the start of the decision process, satisfies both contiguous activation and sequential transition. Of course, satisfying these properties does not imply that the decision maker necessarily adjusts away from the anchor. It may be the case that the anchor is stable. If there is adjustment, however, the adjustment is guaranteed to be sequential. Anchors trigger a cascade of activation in the response layer: Neighboring responses activate and deactivate consecutively. There are no jumps in response activation, nor do multiple non-neighboring responses activate, without the activation of the intermediate responses.


Figure 2: The emergence of sequential adjustment.
How does the connectedness property satisfy contiguous activation and sequential transitions? While the proof of this claim is in the appendix, the intuition for it is as follows. Due to connectedness, cues that support both the anchor and
a non-neighboring non-anchored response must also support any intermediate responses, lying between the anchor and the non-neighboring response. Thus if the activation of the anchor activates cues that subsequently activate nonneighboring responses, these cues must also activate all of these intermediate responses. Subsequently, response activation at \(t=2\) must be contiguous, and any transitions that may have happened at \(t=1\) must be sequential. This intuition however also applies for the contiguous interval of responses activated at \(t=2\), implying that any further changes to activation after \(t=2\) must be sequential. Additionally, once a contiguous interval of responses is activated, we can show that connectedness implies that this interval cannot splinter into smaller, non-contiguous activated intervals, implying that contiguous activation must also be satisfied after \(t=2\). Mathematical induction shows that these properties then hold at all times.

Connected BAM memory structures guarantee sequential activation. But can they generate insufficient adjustment? Let us consider the case with one correct response. When the memory structure is such that two nodes lying between the anchor and the correct response do not overlap on an appropriate number of cues, the sequential adjustment process described above will be insufficient: it will stabilize with the activation of response values closer to the anchor than the correct response.

The intuition for this is fairly straight forward. If, for a low anchor, there exist two response nodes between the anchor and the correct response, whose cue support does not overlap sufficiently, then the activation of the lower response node will not lead to the activation of cues that activate the higher response node. As activation must be contiguous and transitions must be sequential, no higher nodes can be activated, the network will stabilize with the activation of incorrectly low responses, and the correct response will remain turned off. The same intuition holds for tasks involving a high anchor, in which the network will stabilize with the activation of incorrectly high responses, and the correct response will remain turned off.

\section*{Demonstrations}

The above sections have shown that the BAM network with connected memory structures satisfies contiguous activation and sequential transition, and can generate insufficient adjustment. While this is an analytical result, proved in the appendix, and guaranteed to hold regardless of any underlying parameters, demonstrations of the types of sequential adjustment generated by connectedness can provide important insights regarding the behavior emerging from the BAM network.

Figure 3 provides one such demonstration. It shows a hypothetical distribution of cue support for a sequence of responses, and the settling dynamics of the corresponding BAM network with a high anchor, low anchor, and without any anchor. The correct response in this network is \(r_{4}\), and this is the stable response in the absence of an anchor. When anchored at \(r_{6}\) (a high anchor), however, the network
stabilizes at \(r_{5}\). Similarly when anchored at \(r_{1}\) (a low anchor), the network stabilizes at \(r_{3}\). These behaviors indicate the presence of the anchoring effect. Additionally, the settling dynamics with these anchors display sequential adjustment: response nodes activate and deactivate consecutively until the network stabilizes.

Why do we observe these behaviors? \(r_{1}, r_{2}\), and \(r_{3}\) overlap on the component cues in such a way that the set of cues supported by \(r_{i}\) also on average support \(r_{i+1}\), for \(i=1,2\). This means that activating \(r_{1}\) leads to the activation of \(r_{2}\), which then activates \(r_{3}\). The set of cues supporting \(r_{3}\) and \(r_{4}\) do not however overlap in this way, implying that the cascade of activation begun by anchoring the network at \(r_{1}\) ends with the stable activation of \(r_{3}\). A similar property holds for \(r_{5}\) and \(r_{6}\). Also note that the network satisfies connectedness, which implies that the activation dynamics generated by the anchor display contiguous activation and sequential transitions, leading to sequential adjustment.


Figure 3: Distribution of cue support, and resulting network dynamics for low, high and no anchors.
These dynamics also emerge with larger, randomly generated memory structures. Consider a setting with \(N=100\) responses and \(M=1000\) cues. Let us randomly generate support or opposition between these cues and these responses. For each cue we can pick a number from the normal distribution with mean 50 and variance 25 , and round it to its nearest integer. We can subsequently take an interval of length 20 around this integer, to generate the set of responses supported by the cue. All other responses are opposed by the cue. Taking an interval of responses around the randomly chosen number generates a "blurring" in the underlying memory structure: it is seldom the case that individual cues support point estimates; rather their support is distributed across an interval of responses.

As the randomly generated memory structure satisfies connectedness, it should be able to generate sequential adjustment. Figure 4 displays the dynamics of the BAM network instantiating this randomly generated memory structure, with a high anchor, \(r_{100}\) and a low anchor, \(r_{1}\). Note that the stable responses for the two anchors are different, with the stable responses for the low anchor lower than the stable responses for the high anchor. Additionally, activation at all points of time is contiguous, and all transitions are sequential: we can observe a cascade of activation in the response layer over time, with intervals of
responses activating and deactivating consecutively before finally stabilizing.

Note that the dynamics observed in figure 4 also emerge with alternate parameters in the model. In general, however, increasing the ratio of total responses to total cues and increasing the blurring in the cue support for the responses generates a higher likelihood of adjustment, as well as longer sequences of adjustment. This subsequently leads to weaker anchoring effects. Overall the anchoring bias is strongest when there are many relevant cues, and each cue supports few neighboring responses.


Figure 4: Network dynamics for high and low anchors, with randomly generated memory.

\section*{Explaining Anchoring Phenomena}

Anchoring is a well-studied phenomenon and the sequential adjustment and biased activation theories of anchoring have a large range of behavioral findings that they must be able to account for. The above sections have shown that these theories are almost identical: the process assumed by one, emerges directly from the process assumed by the other. This section shows how this result can explain most of the findings documented in anchoring research.

Using a lexical decision task, Mussweiler and Strack (2000) find that decision makers identify "cold" related words quicker and more accurately after temperature judgments with low anchors, and identify "hot" related words quicker and more accurately after temperature judgments with high anchors. Sequential adjustment theory is unable to account for this finding, however, the BAM framework allows for both sequential adjustment and anchor dependent cue accessibility biases to emerge simultaneously: once the network settles, the cues that support the stable responses are themselves stable. If the judgment began with a low anchor then stable cues are more likely to support the low anchor than the high anchor. The opposite holds if the judgment began with a high anchor.

The biased activation theory of anchoring also predicts that exogenous factors influencing cue accessibility can affect anchoring. This has been verified by Chapman and

Johnson (1999) and Mussweiler et al. (2000). Unlike sequential adjustment theory, the BAM model can explain these findings. If we assume that exogenous influences on cue attention affect the inputs, \(I\), into the cue layer, then directing attention towards cues that oppose the anchored response \(r_{i}\), leads to stronger inputs, \(I>1\), into these cues. Due to these inputs, these cues are not inhibited by feedback from the activated anchor in the response layer. Subsequently all cues are activated at the start of the decision process, the pattern of activation on the cue layer resembles the pattern observed in the absence of an anchor, and the network stabilizes without an anchoring bias.

According to the traditional sequential adjustment theory, all types of anchors, regardless of underlying cue support, should lead to the anchoring effect. Research by Chapman and Johnson (1994), however, finds that implausible anchors (anchors that are not supported by any cues) have a much weaker effect than plausible anchors. BAM provides a simple explanation for this result. When implausible anchors are activated at the start of the decision process, all cues are suppressed (as these anchors are not supported by any cues). Subsequently none of the response nodes activate in the next time period. This leads the network to a state identical to the starting state of the network in the absence of an anchor. Implausible anchors thus do not generate an anchoring effect.

A fourth finding supporting the biased activation theory of anchoring pertains to the effect of multiple anchors. Sequential adjustment theory predicts that the decision maker adjusts sequentially away from the one anchor presented in the decision task. This theory cannot make predictions for settings with multiple anchors. Switzer and Sniezek (1991) and Whyte and Sebenius (1997), however, demonstrate that multiple anchors affect judgment differently relative to single anchors. Single anchors paired with more extreme anchors generate a stronger anchoring effect than the single anchors alone, whereas single anchors paired with less extreme anchors generate a weaker anchoring effect than the single anchors alone.

BAM can account for the effect of multiple anchors. The activation of multiple response nodes at the start of the judgment process leads to the activation of all the cues supporting these anchors. When a single anchor is paired with a more extreme anchor then the set of cues activated are more likely to support extreme responses, relative to when the single anchor is activated by itself. This can lead to the stable activation of responses close to the extreme anchor, generating a stronger anchoring effect. The opposite happens when a single anchor is paired with a less extreme anchor. Here the activated cues are less likely to support extreme responses. This can lead to the ultimate stable activation of responses close to the moderate anchor, generating a weaker anchoring effect.

The cue accessibility, exogenous attentional influence, implausible anchor and multiple anchor results discussed above present strong evidence for the biased activation theory of anchoring. The standard biased activation theory
cannot however provide a comprehensive account of all the moderators of the anchoring effect. Research by Reitsmavan Rooijen and Daamen (2006), for example, finds that time pressure increases the anchoring effect. This has traditionally seen as providing evidence for the sequential adjustment theory of anchoring, according to which time pressure limits the number of adjustments possible, thereby increasing the strength of the anchoring effect. As the BAM network proposed in this paper generates sequential adjustment, it is able to provide an explanation for these results as well. The BAM network often does not settle at its stable response in one time step; rather its response nodes activate and deactivate consecutively over time, before stabilizing at the final response (as in e.g. figure 4). When the decision maker is faced with time pressure, the network is not allowed to stabilize and the adjustment process generated in this network is curtailed, generating a stronger anchoring effect.

Another finding providing evidence for sequential adjustment theory relates to the role of incentives on anchored judgment. Particularly, Simmons et al. (2010) find that financial incentives reduce the anchoring effect. This cannot be explained by biased activation theory. If, however, we assume that incentivized decision makers send stronger inputs into the cue activation layer (perhaps due to increased attention towards all cues relevant to the decision task) then the BAM network can in fact explain this effect. As discussed above, when \(I>1\), the exogenous inputs override the inhibitory feedback from the anchor in the response layer. Cue activation subsequently resembles the unbiased decision process, and the anchoring effect disappears.

\section*{Anchoring as Constraint Satisfaction}

The bidirectionality assumed in this paper is a property of a general class of models that have been used to explain findings on inference across a variety of domains. These are models of constraint satisfaction (see e.g. Holyoak \& Simon, 1999 for a review). Constraint satisfaction models provide a powerful approach to studying the interrelationships between cues and responses, and the ways that these relationships affect the dynamics of the decision process. Indeed, the anchoring effect can be seen as just a specific instantiation of the general type of starting point sensitivity displayed by these models: if the memory structures in these models satisfy connectedness then these models will also generate sequential adjustment. In this light, the BAM network is not just a model of anchoring, but rather a model of constraint satisfaction; one which provides a tractable framework with which to understand the cognitive dynamics that constraint satisfaction entails, and the behaviors that these dynamics can generate.

\section*{Conclusion}

We have used the bidirectional associative memory network to study the anchoring effect. The BAM network provides a simple model for the biased activation theory of anchoring.

We have shown that the settling dynamics of this BAM network generate sequential adjustment. Anchors trigger a cascade of activation in the response layer of the BAM network, with nodes in this layer activating and deactivating consecutively. This progression of activation is generally insufficient and final responses depend critically on starting anchor values. By reconciling two contrasting theories within one framework, the BAM network is able to provide a parsimonious explanation for a wide range of findings regarding anchoring and its moderators.

\section*{APPENDIX}

Here we shall show that BAM networks with connected memory structures satisfy contiguous activation and sequential transition. Let us define \(\boldsymbol{D}_{\boldsymbol{i}}\) to be the set of cues supporting \(r_{i}, \boldsymbol{D}^{t}\) to be the set of cues activated at \(t, \boldsymbol{E}_{j}\) to be the set of responses supported by \(c_{j}\) and \(\boldsymbol{E}^{\boldsymbol{t}}\) to be the set of responses activated at \(t\). \(|\boldsymbol{X}|\) shall indicate set \(\boldsymbol{X}\) 's cardinality. Now consider the following propositions:

Proposition la: If a contiguous interval of responses, \(r_{i}\), \(r_{i+1}, \ldots r_{k}\) is activated at \(t\) (and all other responses are deactivated at \(t\) ), and for \(l>k, r_{l}\) is activated at \(t+1\), then it is the case that \(r_{k}, r_{k+1} \ldots r_{l-l}\) are activated at \(t+1\). Proof: \(c_{j} \in \boldsymbol{D}^{t}\) implies \(c_{j} \boldsymbol{D}_{\boldsymbol{i}} \cup \boldsymbol{D}_{i+1} \ldots \cup \boldsymbol{D}_{\boldsymbol{k}}\). Since \(r_{l} \in \boldsymbol{E}^{\boldsymbol{t + 1}}\), we have \(\left|\boldsymbol{D}_{l} \cap \boldsymbol{D}^{t}\right|>\left|\boldsymbol{D}^{t}\right| / 2\). Connectedness implies that if \(c_{j} \in \boldsymbol{D}_{i} \cup \boldsymbol{D}_{i+1} \ldots \cup \boldsymbol{D}_{k}\) and \(c_{j} \in \boldsymbol{D}_{l}\) then \(c_{j} \in \boldsymbol{D}_{l}\), for \(l>l^{\prime} \geq k\). Hence if \(\left|\boldsymbol{D}_{l} \cap \boldsymbol{D}^{t}\right|>\left|\boldsymbol{D}^{t}\right| / 2\) we also have \(\left|\boldsymbol{D}_{l}, \cap \boldsymbol{D}^{t}\right|>\left|\boldsymbol{D}^{t}\right| / 2\) for all \(l>l \geq k\), which means that \(r_{l} \boldsymbol{E}^{t+1}\) implies \(r_{l} \in \boldsymbol{E}^{t+1}\) for \(l>l^{\prime} \geq k\).

Proposition \(1 b\) : If a contiguous interval of responses, \(r_{i}\), \(r_{i+l}, \ldots r_{k}\) is activated at \(t\) (and all other responses are deactivated at \(t\) ), and for \(l<i, r_{l}\) is activated at \(t+1\), then it is the case that \(r_{l+1}, r_{l+2}, \ldots r_{i}\) are activated at \(t+1\). Proof: The proof for this is identical to that for proposition 1a.

Proposition 2: If a contiguous interval of responses, \(r_{i}\), \(r_{i+l}, \ldots r_{k}\) is activated at \(t\) (and all other responses are deactivated at \(t\) ), then for any \(p\) and \(q\) with \(k>p>q>i\), if \(r_{q}\) and \(r_{p}\) are activated at \(t+1\) then so is any \(r_{l}\) for \(p>l>q\). Proof: \(c_{j} \in \boldsymbol{D}^{t^{t}}\) implies \(\left|\boldsymbol{E}_{j} \cap \boldsymbol{E}^{t} \geq\left|\boldsymbol{E}^{t}\right| / 2\right.\). As \(\boldsymbol{E}_{j}\) is contiguous (by connectedness), and \(\boldsymbol{E}^{t}\) is contiguous, \(\boldsymbol{E}_{j} \cap \boldsymbol{E}^{\boldsymbol{t}}\) is also contiguous. Hence if \(c_{j} \boldsymbol{D}^{t}\) it supports at least \(\left|\boldsymbol{E}^{t}\right| / 2=(k\) \(i+1) / 2\) contiguous responses in \(\boldsymbol{E}^{t}\). Assume that \(q<(k+i) / 2\). If \(c_{j} \in \boldsymbol{D}_{\boldsymbol{q}} \cap \boldsymbol{D}^{t}\) then as \(c_{j}\) supports at least ( \(k-i+1\) )/2 neighboring responses in \(\boldsymbol{E}^{t}\), we must also have \(c_{j} \boldsymbol{D}_{q+\boldsymbol{1}}\). Hence if \(\left|\boldsymbol{D}_{q} \cap \boldsymbol{D}^{t}\right|>\left|\boldsymbol{D}^{t}\right| / 2\), as is implied by \(r_{q} \boldsymbol{E}^{t+1}\), then we have \(\left|\boldsymbol{D}_{q+1} \cap \boldsymbol{D}^{t}\right|>\left|\boldsymbol{D}^{t}\right| / 2\), which implies that \(r_{q+1} \boldsymbol{E} \boldsymbol{E}^{t+1}\). Now we can use this method again to show that \(r_{q+2} \epsilon \boldsymbol{E}^{t+1}\), and keep iterating it to show that \(r_{l} \boldsymbol{E}^{t+1}\) for all \((k+i) / 2 \geq l \geq q\). Now if \((k+i) / 2 \geq p\) then our proof is done. If not then note that we can use the same logic as above to show that \(r_{l} \in \boldsymbol{E}^{t+1}\) for \(p \geq l \geq(k+i) / 2\). This then gives us our result.

Now, propositions 1a and 1b show that if a contiguous interval of responses is activated at time \(t\) then a response that does not neighbor this contiguous interval, cannot be activated at \(t+1\) without activating all intermediate responses. Proposition 2 shows that if a contiguous interval of responses is activated at time \(t\) then this interval cannot splinter into two or more non-contiguous intervals of
activated responses at \(t+1\). Together these results imply both contiguous activation and sequential transitions.

\section*{References}

Chapman, G. B., \& Johnson, E. J. (1994). The limits of anchoring. Journal of Behavioral Decision Making, 7(4), 223-242.

Chapman, G. B., \& Johnson, E. J. (1999). Anchoring, activation, and the construction of values. Organizational Behavior and Human Decision Processes, 79(2), 115153.

Choplin, J. M. \& Tawney, M. W. (2010). Mathematically modeling anchoring effects. Proceedings of the Thirty-Second Annual Conference of the Cognitive Science Society.

Holyoak, K. J., \& Simon, D. (1999). Bidirectional reasoning in decision making by constraint satisfaction. Journal of Experimental Psychology: General, 128(1), 3.

Johnson, J. G., \& Busemeyer, J. R. (2005). A dynamic, stochastic, computational model of preference reversal phenomena. Psychological Review, 112(4), 841.

Kosko, B. (1988). Bidirectional associative memories. Systems, Man and Cybernetics, IEEE Transactions on, 18(1), 49-60.

Mussweiler, T., \& Strack, F. (1999). Hypothesisconsistent testing and semantic priming in the anchoring paradigm: A selective accessibility model. Journal of Experimental Social Psychology, 35(2), 136-164.

Mussweiler, T., \& Strack, F. (2000). The use of category and exemplar knowledge in the solution of anchoring tasks. Journal of Personality and Social Psychology, 78(6), 1038.

Mussweiler, T., Strack, F., \& Pfeiffer, T. (2000). Overcoming the inevitable anchoring effect: Considering the opposite compensates for selective accessibility. Personality and Social Psychology Bulletin, 26(9), 11421150.

Reitsma-van Rooijen, M., \& L Daamen, D. D. (2006). Subliminal anchoring: The effects of subliminally presented numbers on probability estimates. Journal of Experimental Social Psychology, 42(3), 380-387.

Simmons, J. P., LeBoeuf, R. A., \& Nelson, L. D. (2010). The effect of accuracy motivation on anchoring and adjustment: Do people adjust from provided anchors?. Journal of Personality and Social Psychology, 99(6), 917.

Switzer, F. S., \& Sniezek, J. A. (1991). Judgment processes in motivation: Anchoring and adjustment effects on judgment and behavior. Organizational Behavior and Human Decision Processes, 49(2), 208229.

Whyte, G., \& Sebenius, J. K. (1997). The effect of multiple anchors on anchoring in individual and group judgment. Organizational Behavior and Human Decision Processes, 69(1), 74-85.

\title{
A Neurally Plausible Encoding of Word Order Information into a Semantic Vector Space
}

\author{
Peter Blouw (pblouw@uwaterloo.ca) \\ Chris Eliasmith (celiasmith@uwaterloo.ca) \\ Center for Theoretical Neuroscience, University of Waterloo \\ Waterloo, ON N2L3G1 Canada
}

\begin{abstract}
Distributed models of lexical semantics increasingly incorporate information about word order. One influential method for encoding this information into high-dimensional spaces uses convolution to bind together vectors to form representations of numerous \(n\)-grams that a target word is a part of. The computational complexity of this method has led to the development of an alternative that uses random permutation to perform order-sensitive vector combinations. We describe a simplified form of order encoding with convolution that yields comparable performance to earlier models, and we discuss considerations of neural implementation that favor the use of the proposed encoding. We conclude that this new encoding method is a more neurally plausible alternative than its predecessors.
\end{abstract}

Keywords: semantic memory; convolution; random permutation; vector space models; distributional semantics

\section*{Introduction}

The well-known 'semantic space' approach to modeling word meanings is frequently employed by researchers interested in understanding how the brain represents lexical information. At its most simple, the approach involves encoding word co-occurrence statistics from natural language corpora into a set of high dimensional vectors (e.g. Landauer \& Dumais, 1997; Lund \& Burgess, 1996; Jones \& Mewhort, 2007). The spatial relationships between such vectors are then taken to reflect semantic relationships amongst corresponding words. Experiments involving semantic space models have produced impressive results matching human data from studies of category typicality (e.g., Jones \& Mewhort, 2007) and synonym identification (e.g., Landauer \& Dumais, 1997), amongst other things.

However, one concern with the traditional semantic space approach is that it fails to take into account information about how words are sequentially related to one another (Jones \& Mewhort, 2007). For example, the latent semantic analysis (LSA) model developed by Landauer and Dumais (1997) functions by building a word-document frequency matrix that treats all words occurring in a single document equivalently. Similarly, Lund and Burgess' (1996) hyperspace analog to language (HAL) model simply counts the frequency of words occurring within a multi-word window around a target term. This indifference to sentence structure has led to HAL and LSA being referred to as 'bag of words' models of lexical semantics (Jones \& Mewhort, 2007; Recchia et al., 2010).

More recently, two techniques have been developed to incorporate word order information into semantic vectors. The first, developed by Jones and Mewhort (2007), uses circular convolution (proposed by Plate (2003) as a vector binding operation) to create vector representations of the numerous \(n\)-grams a target word is a part of. The second, developed by Sahlgren, Holst, and Kanerva (2008), uses random vector permutation to index the positions of neighboring words in relation to a target word. Functionally, the two approaches are quite similar, but random permutation is much more computationally efficient than convolution (Sahlgren, Holst, \& Kanerva, 2008). Moreover, a recent analysis indicates that convolution and random permutation offer similar degrees of accuracy during information retrieval, and that they perform comparably on a set of basic semantic tasks involving synonym identification (Recchia et al., 2010).

Given that computational efficiency favors the use of random permutation, the aim of this paper is to develop a simplified version of convolution encoding that can replicate many of the important functional properties of Jones and Mewhort's (2007) method. More specifically, we use convolution with position-indexing vectors to produce a single \(n\)-gram for each occurrence of a target word in a corpus (cf. Sahlgren, Holst, \& Kanerva, 2008). Encoding a single \(n\)-gram per word occurrence is much simpler than Jones and Mewhort's technique of encoding multiple \(n\) grams per word occurrence, and we demonstrate that this simplification provides good model performance on a range of order-specific tasks involving phrase-completion.

In addition, we argue that our encoding is more biologically plausible for two reasons:
1) All of the required vector representations can be instantiated using simulated spiking neurons.
2) All of the required computations on these representations can also be instantiated using simulated spiking neurons.

To substantiate these claims, we rely on prior work. Eliasmith and Anderson (2003) describe a method for representing and transforming high dimensional real-valued vectors in neural systems through a combination of the nonlinear encoding of a signal into a pattern of neural spikes, and the weighted linear decoding of these spikes. Simple operations such as vector addition are easily implemented
using these methods, and Eliasmith (2005) extends such work to describe a neural implementation of the circular convolution operation. Since our encoding method utilizes only circular convolution and vector addition, these remarks indicate that it is therefore a neurally plausible method.

In contrast, the approach of Sahlgren, Holst and Kanerva employs binary vectors, which are not naturally implemented in neural models (Stewart \& Eliasmith, 2012). Moreover, the approach of Jones and Mewhort employs a series of computations that are arguably too complex to scale appropriately if implemented in neurons. Our position encoding approach, on the other hand, has been utilized in a portion of what is currently the world's largest functional brain model (Eliasmith, et al., 2012), capable of a range diverse tasks involving perception, cognition, and action.

In what follows, we first review the convolution-based encoding algorithm presented by Jones and Mewhort (2007), along with the random permutation algorithm presented by Salhgren, Holst, and Kanerva (2008). We then introduce our own encoding algorithm. Next, we report results from a series of simulations conducted to assess model performance. We conclude that convolution with position indices offers an equally useful but more biologically plausible strategy for incorporating order information into semantic space models.

\section*{Two Approaches to Encoding Word Order}

The main challenge facing efforts to encode syntactic information into high-dimensional spaces is to find an appropriate, order-preserving mathematical operation for recursively combining vectors. Given that standard vector operations, such as superposition, are inadequate for this purpose, researchers have proposed a number of multiplicative binding methods instead. Examples include Smolensky's (1990) tensor products, Kanerva's (1994) binary spatter codes, and Plate's (2003) holographic reduced representations. Plate's approach has been particularly attractive to researchers interested in language because of its use of circular convolution, which ensures that all recursively bound vectors are of the same dimensionality. In absence of preserved dimensionality, it becomes difficult to compare vectors representing differently structured linguistic objects (e.g. phrases of different lengths; Jones \& Mewhort, 2007).

Before getting into the details of encoding with convolution and random permutation, it is worth noting that the point of departure for comparing the two methods is Jones and Mewhort's (2007) BEAGLE \({ }^{1}\) model, which assigns each word in a modeled corpus a unique environmental vector ( \(e\) ), along with a zero-valued memory vector \((m)\). Each time a word is encountered in the corpus, its memory vector is updated with context information provided through the superposition of the environmental vectors for every other word in the surrounding sentence.

\footnotetext{
\({ }^{1}\) The acronym stands for 'bound encoding of the aggregate language environment'.
}

Simultaneously, the memory vector is also updated with a vector describing the ordering of the target word in relation to a limited range neighbors. As whole, the process conforms to the following expression:
\[
\begin{equation*}
m_{i}=m_{i}+c_{i}+o_{i} \tag{1}
\end{equation*}
\]
where \(i\) indexes the word being represented, while \(c_{i}\) and \(o_{i}\) refer to vectors describing context and order information for a given word occurrence. \({ }^{2}\) The primary difference, then, between the approaches of Jones \& Mewhort (2007) and Sahlgren, Holst, and Kanerva (2008), is in the calculation of \(o_{i}\). In BEAGLE, \(o_{i}\) incorporates a range of \(n\)-grams that a target word is a part of. To give an example of how this works, consider the sentence 'make hay while the sun shines' and the target word 'hay'. The order vector, o ohay, is then calculated as the sum of various \(n\)-grams that 'hay' is a part of:
\[
\begin{aligned}
& \text { bigram }_{1}=e_{\text {make }} * \Phi \\
& \text { bigram }_{2}=\Phi * e_{\text {while }} \\
& \text { trigram }_{1}=e_{\text {make }} * \Phi * e_{\text {while }} \\
& \text { trigram }_{2}=\Phi * e_{\text {while }} * e_{\text {the }} \\
& \text { ngram }_{i}=\ldots
\end{aligned}
\]
where, * denotes the circular convolution operation, \(\Phi\) denotes a placeholder vector for the target word, and \(n\) sets size of the window around the target word from which order information is drawn. The value of \(n\) is typically set to 7 .

Overall, this method is quite computationally expensive given that each word occurrence prompts the generation of numerous sequences of convolutions, each of which must be computed in \(O(n \log n)\) time (Jones \& Mewhort, 2007). Moreover, because convolution is a commutative operation, permutations are applied to distinguish vectors of the form A * B and B * A. This adds an additionally layer of complexity when encoding large sequences of ordered vectors.

In light of this computational complexity, Sahlgren, Holst, and Kanerva's (2008) proposal is to recursively apply a random permutation to the environmental vectors to indicate their position relative to the target word. The random permutation, \(\Pi\), scrambles the order of the elements in a vector, and its recursive application indexes positions at varying distances from the target word:
\[
o_{\text {hay }}=\Pi^{-1} e_{\text {make }}+0+\Pi^{1} e_{\text {while }} \ldots+\Pi^{4} e_{\text {shines }}
\]

Here, the positive superscripts indicate the number times the permutation is applied to an environmental vector, and the negative superscripts indicate the number of times the inverse of the permutation is applied. One important feature of this method is that each occurrence of a target word in the

\footnotetext{
\({ }^{2}\) The context and order vectors are normalized prior to being combined and incorporated into the memory vector.
}
corpus results in the memory vector being updated with only a single \(n\)-gram containing every word in the order window. The resulting order vector, \(o\), is thus structurally quite different from vectors produced through the summing of multiple \(n\)-grams (Sahlgren, Holst, \& Kanerva, 2008).

For information retrieval in this framework, the inverse of a particular position permutation is applied to a memory vector. This process yields a vector that is most similar to environmental vectors that have been frequently bound into the memory vector in this position. Thus, one can extract information about which words are likely to occur in various positions around a target word. For example, \(\Pi^{-1} m_{\text {hay }}\) would yield a vector most similar to words that have frequently been bound into the first position succeeding 'hay' in various order vectors generated over the course of scanning the corpus. Depending on the statistical properties of this corpus, a comparison (i.e. cosine measure) between \(\Pi^{-1} m_{\text {hay }}\) and the environmental vectors will likely yield an environmental vector such as \(e_{\text {bale }}\) as most similar.

Overall, when comparing these methods for generating memory vectors, three things are important to keep in mind. First, there are a number of further differences between BEAGLE and Sahlgren, Holst, and Kanerva's model beyond the use of random permutation for order encoding. For example, the latter model uses binary environmental vectors, while Jones and Mewhort's model uses environmental vectors whose elements are picked from a Gaussian distribution of a mean of zero and variance equal to \(1 / D .{ }^{3}\) Moreover, Sahlgren, Holst, and Kanerva apply a smaller window for calculating context information that ignores sentence boundaries. These differences limit the ability to conduct performance comparisons based on the use of random permutation alone.

Second, to the extent that such comparisons have been made, they focus almost exclusively on storage capacity measures and performance on simple synonym identification tasks. However, one of the more compelling attributes of the BEAGLE model is its ability to reflect experimental effects involving things like category typicality, priming, and semantic constraints on stem completion. It has not been demonstrated that models built using random permutation have comparable capabilities.

Third, the BEAGLE model is computationally expensive, but uses real-valued vectors (which are efficiently implementable in a biologically plausible network; Eliasmith \& Anderson, 2003), whereas the permutation model is computationally efficient, but uses binary vectors (which have not been demonstrated to be efficient to implement biologically). Past work has not proposed a representation that is both computationally and biologically efficient.

Here, we describe a new representation that is comparable to the BEAGLE model in that it preserves the functional

\footnotetext{
\({ }^{3}\) These properties are needed to ensure that convolution can be used effectively as an operation for binding and unbinding vectors (Plate, 2003).
}
properties of its memory vectors, but it uses a single \(n\)-gram order encoding method that is structurally similar to Sahlgren, Holst, and Kanerva's technique while employing real-valued vectors.

\section*{Convolution with Position Vectors}

Our proposal is to encode order information with a set of reusable, real-valued, unitary, randomly generated 'position vectors'. \({ }^{4}\) These vectors are convolved with environmental vectors and summed to give an order vector of the following form:
\[
\begin{equation*}
o_{i}=\ldots p_{-1} * e_{-1}+0+p_{1} * e_{1}+p_{2} * e_{2} \ldots \tag{2}
\end{equation*}
\]
where \(p_{1}\) is the vector that indexes the first position succeeding the target word, \(p_{-1}\) is the vector that indexes the first position preceding the target word, and so forth. \(e_{1}, \mathrm{e}_{2}\), etc. are the environmental vectors of the words in each position around the target word. Structurally, this approach shares the property of position indexing with the model of Sahlgren, Holst and Kanerva (2008), but computationally, it shares the use of convolution of real-valued vectors with the model of Jones and Mewhort (2007).

To make the proposal clearer, consider again the word 'hay' in the sentence 'make hay while the sun shines'. The order vector produced with our method would be
\[
o_{\text {hay }}=p_{-1} * e_{\text {made }}+0+p_{1} * e_{\text {wwile }} \ldots+p_{2} * e_{\text {sthines }}
\]

Once this order vector is incorporated into the memory vector for 'hay', this memory vector will become slightly more similar to other vectors with have had 'hay' bound into the first position to the right too.

To retrieve order information from a memory vector, we can use one of two methods, both adapted from Jones \& Mewhort (2007). The first is to convolve the inverse of a position vector with a memory vector to extract a representation that is most similar to the environmental vectors that have been frequently bound into the memory vector in this position. For example:
\[
m_{\text {hay }} * p_{1}^{-1} \approx e_{\text {while }}
\]

Note that this method can be used to extract words commonly found in any of the twelve positions for which order information is encoded.

The second form of information retrieval involves constructing a probe corresponding to particular ordering around a target word, and then identifying which memory vectors have most frequently encoded the ordering of

\footnotetext{
\({ }^{4}\) To index position, a single unitary vector could also be selfconvolved multiple times. This would avoid the use of random vectors for each position, but it is functionally equivalent to the present formulation.
}
interest. To give an example, one could construct the following probe vector:
\[
\text { probe }=p_{-1} * e_{\text {make }}+0+p_{1} * e_{\text {wwile }}+\ldots+p_{3} * e_{\text {stimes }}
\]

If this vector is compared to all memory vectors generated from the corpus, it will match most closely with words that have frequently encoded the order sequence 'make while the sun shines'. Provided that the corpus does not contain a multitude of words that repeatedly occupy the blank position in relation to the same the surrounding words, the comparison will return the memory vector \(m_{\text {hay }}\) as the closest match.

Overall, information retrieval is made quite simple when position encoding is conducted via convolution with position vectors. As important, however, is whether or not the encoding enables good model performance.

\section*{Simulations}

We test the effects of the position encoding method for performance on a range of tasks involving semantic similarity and phrase completion. As per Jones and Mewhort (2007), context vectors are calculated as the superposition of environmental vectors in the sentence surrounding a target word, and environmental vectors are randomly generated with elements drawn from a Gaussian distribution. A list of stop words is used to prevent frequently occurring function words from being overrepresented in the context vectors, and order information is calculated using position indices ranging from -6 to +6 . This range is chosen because it captures the same set of words that would be included in order vectors calculated using Jones and Mewhort's original method. Finally, context vectors and order vectors are normalized prior to inclusion in the overall memory vector for a given word.

All simulations are run, for efficiency, on a subset of the same TASA corpus used in tests of both BEAGLE and Sahlgren, Holst, and Kanerva's (2008) random permutation model. Approximately 27,000 unique words are modeled using roughly 110,000 sentences, and words occurring less than twice in the corpus are ignored to exclude misspellings and typographical errors.

\section*{A Nearest Neighbors Task}

As an initial qualitative assessment of model performance, we calculated the nearest neighbors to the memory vectors for four common words found in the TASA corpus. We chose the same four words used in Table 3 of Jones and Mewhort (2007). The results, shown in Table 1 below, indicate that encoding order information with position vectors instead of an array of \(n\)-grams results in plausible model performance for each of the four words. All reported activation values are cosines of the angle between two vectors in the semantic space. The context space is comprised of memory vectors only updated with context information, while the order space is comprised of memory
vectors only updated with order information. The combined space includes memory vectors calculated in accordance with Equation 1.

As with the comparison between BEAGLE and the model of Sahlgren, Holst, and Kanerva (2008), subtle differences in things like the selection of stopwords and the formation of the environmental vectors make quantitative comparisons impractical, so we present these results as an independent demonstration of model performance.

Table 1: Nearest Neighbors in Three Spaces
\begin{tabular}{|c|c|c|c|c|c|}
\hline \multicolumn{2}{|l|}{Context} & Order & \multicolumn{3}{|c|}{Combined} \\
\hline \multicolumn{6}{|l|}{EAT} \\
\hline food & 0.69 & get & 0.89 & get & 0.78 \\
\hline get & 0.65 & buy & 0.87 & make & 0.75 \\
\hline animals & 0.63 & make & 0.86 & take & 0.70 \\
\hline need & 0.62 & keep & 0.86 & keep & 0.69 \\
\hline make & 0.61 & meet & 0.85 & find & 0.69 \\
\hline \multicolumn{6}{|l|}{CAR} \\
\hline came & 0.65 & nation & 0.89 & house & 0.75 \\
\hline back & 0.64 & village & 0.88 & road & 0.73 \\
\hline road & 0.64 & fire & 0.88 & big & 0.73 \\
\hline one & 0.63 & family & 0.88 & little & 0.71 \\
\hline way & 0.63 & story & 0.88 & dog & 0.70 \\
\hline \multicolumn{6}{|l|}{READING} \\
\hline read & 0.66 & writing & 0.72 & writing & 0.68 \\
\hline book & 0.61 & making & 0.67 & that & 0.61 \\
\hline writing & 0.61 & business & 0.64 & your & 0.61 \\
\hline skimming & 0.59 & power & 0.62 & or & 0.61 \\
\hline may & 0.56 & food & 0.62 & this & 0.59 \\
\hline \multicolumn{6}{|l|}{SLOWLY} \\
\hline little & 0.63 & quickly & 0.75 & quickly & 0.62 \\
\hline around & 0.63 & again & 0.67 & and & 0.60 \\
\hline back & 0.62 & ran & 0.65 & down & 0.60 \\
\hline across & 0.60 & to & 0.65 & then & 0.59 \\
\hline move & 0.59 & brought & 0.65 & to & 0.59 \\
\hline
\end{tabular}

\section*{Retrieval with Decoding}

Retrieval through decoding, again, involves convolving a memory vector with the inverse of a position vector, and then comparing the output of this process to a library of environmental vectors to find closest matches. In this simulation, we use the decoding retrieval method to find the most likely word to occur both before and after a particular target word. Results are reported in Table 2.

One point to note about these decoding results is that the activation values for the words in each column indicate nonrandom correspondence with the target word if the similarity value is greater than \(\sim 0.1\) (see Jones and Mewhort, 2007, p. 13). Accordingly, the decoding does a good job of picking out words that are likely to follow before or after a given word.

Table 2: Decoding Around a Target Word
\begin{tabular}{cccc}
\hline Word Before & \multicolumn{3}{c}{ Word After } \\
\hline LUTHER & & \\
martin & 0.29 & king & 0.21 \\
straightening & 0.17 & gravity & 0.17 \\
latest & 0.17 & 1733 & 0.16 \\
coinage & 0.16 & puff & 0.16 \\
so-called & 0.16 & conscience & 0.16 \\
\hline KING & \multicolumn{4}{c}{} \\
the & 0.54 & was & 0.19 \\
experienced & 0.17 & tens & 0.17 \\
boundaries & 0.17 & bowing & 0.17 \\
kites & 0.17 & lawfully & 0.17 \\
donor & 0.16 & pasture & 0.16 \\
\hline
\end{tabular}

\section*{Retrieval with Resonance}

Resonance retrieval, again, involves constructing a probe by superposing a number of bound environmental and position vectors. This probe vector is then compared to all of the memory vectors to find items that have frequently occurred within the sequence of words described by the probe.

Table 3: Resonance Around a Target Word
\begin{tabular}{|c|c|c|c|}
\hline Word Before & \multicolumn{3}{|c|}{Word After} \\
\hline \multicolumn{4}{|l|}{KING} \\
\hline rex & 0.38 & midas & 0.42 \\
\hline luther & 0.22 & tut & 0.42 \\
\hline rumbles & 0.17 & aietes & 0.39 \\
\hline hamlet & 0.17 & farouk & 0.36 \\
\hline oyster & 0.16 & richards & 0.31 \\
\hline \multicolumn{4}{|l|}{PRESIDENT} \\
\hline vice & 0.32 & eisenhower & 0.45 \\
\hline activist & 0.20 & lincoln & 0.31 \\
\hline egypts & 0.19 & coolidge & 0.27 \\
\hline middle-of-the-road & 0.19 & johnson & 0.25 \\
\hline dove & 0.18 & nixon & 0.23 \\
\hline \multicolumn{4}{|l|}{WAR} \\
\hline spanish-american & 0.31 & II & 0.49 \\
\hline civil & 0.29 & bonnet & 0.21 \\
\hline post-world & 0.27 & hysteria & 0.19 \\
\hline pre-civil & 0.26 & whoops & 0.19 \\
\hline post-civil & 0.23 & 1898 & 0.18 \\
\hline \multicolumn{4}{|l|}{SEA} \\
\hline caspian & 0.22 & anemone & 0.38 \\
\hline Aegean & 0.22 & level & 0.27 \\
\hline mediterranean & 0.19 & gull & 0.26 \\
\hline foaming & 0.17 & anenomes & 0.24 \\
\hline sensitivity & 0.16 & captains & 0.24 \\
\hline
\end{tabular}

To assess model performance with resonance, we simulate a task involving retrieval around a set of four target words drawn from Table 4 of Jones and Mewhort (2007). The results from this simulation are presented in Table 3. Despite the intrusion of a few unexpected items into these lists of nearest matches (e.g. 'sensitivity'), the overall trend here provides further evidence that order encoding with position vectors can produce a functioning semantic space model.

\section*{Phrase Completion with Resonance}

To go beyond the retrieval of words either immediately to the left or to the right of a target word, we next simulate a set of tasks in which probe vectors corresponding to short phrases are compared to the memory vectors. Initially, only a limited amount of information is included in the probe vector, but subsequently, the probe is enriched to represent a more and more specific order sequence (see Jones \& Mewhort, 2007). As more information is incorporated into the probe in this way, the model increasingly converges on a single word that best fits the blank region in the probe phrase. We use phrase materials drawn from Jones and Mewhort (2007). Results are reported in Table 4 below.

Once again, the model generally meets performance expectations. Preliminary results also indicate that the model generally performs well with other phrases similar to the ones shown. Further work is ongoing in this area.

\section*{Discussion}

At this point, it seems clear that the method of encoding with position vectors performs well enough to be considered a plausible alternative to earlier methods. However, it is worth considering the criteria by which one might select amongst the three forms of encoding discussed in this paper. Computational efficiency, again, favors the use of a single \(n\)-gram encoding method like random permutation or encoding with position vectors.

Then, to decide between convolution and random permutation, one could look to performance measures of the sort just examined. Here, position vector encoding has the advantage of a demonstrated ability to perform a variety of phrase completion tasks; the performance credentials of random permutation have yet to be comparably established. It is possible that random permutation supports the same degree of functionality as demonstrated here, and future work might bear out such a prediction.

However, even if this is the case, we think that independent considerations of neural implementation favor the use of the position vector encoding method. First, note that vector space models of language have appealed to cognitive researchers in part because they possess certain properties suggestive of neural plausibility (Jones \& Mewhort, 2007; Recchia et al., 2010). Connectionist models, for example, have long been used to implement computations defined over vectors, and one of the main attractions of these models is their use of neurally inspired processing mechanisms. So, because semantic space models

Table 4: Highest Word Activations as an Order Sequence is Filled in Around a Target Position
\begin{tabular}{ccccccc} 
Phrase & \multicolumn{5}{c}{ Activations } \\
\hline \begin{tabular}{c} 
emperor [penguins] \\
[penguins] have
\end{tabular} & \begin{tabular}{c} 
yuan \\
planaria
\end{tabular} & 0.26 & penguins & 0.26 & caligula & 0.20 \\
threepio & 0.27 & astronomers & 0.26 \\
\begin{tabular}{c} 
the emperor [penguins] have come \\
to their breeding grounds
\end{tabular} & penguins & 0.34 & yaun & 0.31 & annelida & 0.27 \\
\hline \begin{tabular}{cccccc} 
although [ostriches]
\end{tabular} & \begin{tabular}{c} 
gauges
\end{tabular} & 0.21 & democratically & 0.20 & tsumanis & 0.18 \\
\begin{tabular}{c} 
although [ostriches] cannot \\
although [ostriches] cannot fly they \\
have other skills
\end{tabular} & pretends & 0.16 & raindrops & 0.16 & democratically & 0.16 \\
\hline
\end{tabular}
are constructed through computations defined over vectors, and connectionist models can implement such computations, it follows that semantic space models can share to some extent in the claim of being consistent with how the brain processes information.

Second, models with a high degree of neural plausibility have been built using vector symbolic architectures that employ real-valued vectors and convolution as a binding operator. The same cannot be said for binary vectors. For instance, neurally implemented convolution operations play a key role in a recent model of working memory (Choo \& Eliasmith, 2010), and more significantly, what is currently the world's largest functional brain model (Eliasmith et al., 2012). So, the argument in favor of using convolution with position vectors to encode word order into semantic space models is straightforward: doing so is consistent with the architectural principles that guide state-of-the-art models of complex cognition. Put simply, there is a good deal evidence from these models that the convolution operation accommodates the computational constraints of neural systems.

Together with the demonstrated functionality of semantic space models built using convolution encoding, we think that these considerations of neural implementation provide a compelling case in favor of the method we demonstrate here. Convolution with position vectors provides an approach to building an order-sensitive semantic vector space that is functional, neurally plausible, and relatively computationally efficient. We leave it to future work to determine whether methods utilizing random permutation can display a similar range of strengths.

\section*{Acknowledgments}

This research was supported by the Social Sciences and Humanities Research Council of Canada.

\section*{References}

Choo, F.X., \& Eliasmith, C. (2010). A spiking neuron model of serial order recall. Proceedings of the \(32^{\text {nd }}\) Annual Conference of the Cognitive Science Society. (pp. 2188-2193)
Eliasmith, C. (2005). Cognition with neurons: A large-scale biologically plausible model of the Wason card task.

Proceedings of the \(27^{\text {th }}\) Annual Conference of the Cognitive Science Society (pp. 624-630).
Eliasmith, C. \& Anderson, C. (2003). Neural engineering: Computation, representation, and dynamics in neurbiological systems. Cambridge, MA: MIT Press.
Eliasmith, C., Stewart, T., Choo, F.X., Bekolay, T., DeWolf, T., Tang, Y., \& Rasmussen, D. (2012). A large-scale model of the functioning brain. Science, 338.6111, 12021205.

Jones, M.N. \& Mewhort, D. (2007). Representing word meaning and order information in a composite holographic lexicon. Psychological Review, 114.1, 1-37.
Kanerva, P. (1994). The spatter code for encoding concepts at many levels. Proceedings of the International Conference on Artificial Neural Networks. (pp. 226-229). Sorrento, Italy: Springer-Verlag.
Landauer, T., and Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychological Review, 104.2, 211-240.
Lund, K. \& Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. Behavioral Research Methods, 28.2, 203-208.
Plate, T.A. (2003). Holographic reduced representations: distributed representations for cognitive structures. Stanford, CA: CSLI Publications.
Recchia, G., Jones, M., Sahlgren, M., \& Kanerva, P. (2010). Encoding sequential information in vector space models of semantics: Comparing holographic reduced representations and random permutation. Proceedings of the \(32^{\text {nd }}\) Annual Conference of the Cognitive Science Society (pp. 865-870).
Sahlgren, M., Holst, A., \& Kanerva, P. (2008). Permutations as a means to encode order in word space. Proceedings of the \(30^{\text {th }}\) Annual Conference of the Cognitive Science Society (pp. 1300-1305).
Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. Artificial Intelligence, 46, 159-216.
Stewart, T. \& Eliasmith, C. (2012). Compositionality and Biologically plausible models. In W. Hinzen, E. Machery, \& M. Werning (eds.) Oxford Handbook of Compositionality (pp. 596-615). Oxford: Oxford University Press

\title{
Visualizing the Invisible: Generating Explanations of Scientific Phenomena
}

\author{
Eliza Bobek (ebobek@packer.edu) \\ The Packer Collegiate Institute \\ Brooklyn, NY 11238 USA
}

\begin{abstract}
This study investigated the ability of learner-generated visualizations to improve learning in science. The hypothesis was tested in two domains, a mechanical system and a chemical system, and the results were analyzed separately to compare low and high spatial ability learners. The production of visual explanations of a mechanical system, a bicycle tire pump, increased understanding of the pump particularly for participants with low spatial ability. In the domain of chemical bonding, visual explanations were more effective than verbal explanations for all participants. Visual explanations often included crucial yet invisible features; their accurate construction requires and provides a check for completeness of explanations.
\end{abstract}

Keywords: learning; drawing; external representation; structure; function; spatial ability; self-generated explanation

\section*{Introduction}

Many topics in science are notoriously difficult for students to learn. Mechanisms and processes that exist on a scale outside student experience, such as gravitational pull, chemical bonding, and cellular processes, present particular challenges. When students attempt to learn these phenomena, they often experience difficulty because they must understand not only the individual components of the process (structure) but also the interactions and mechanisms (function). While instruction often involves visualizations, students typically explain in words, spoken or written. Visualizations have many advantages over verbal explanations, especially for science, so asking student to produce visual rather than verbal explanations should improve their learning.

\section*{Learner-generated Explanations}

When learners make connections between information, knowledge, and experience, by generating headings, summaries, pictures, and analogies, deeper understanding develops (Wittrock, 1990). Mayer and colleagues have conducted several experiments that have shown a learning benefit to generative activities in domains involving invisible components, including electric circuits (Johnson \& Mayer, 2010), lightning formation (Johnson \& Mayer, 2009), and the chemistry of detergents (Schwamborn et al., 2010). Hausmann \& Vanlehn (2007) addressed the possibility that generating explanations is beneficial because learners merely spend more time with the content material. In their study in the domain of physics, provided
explanations were not as effective as generated explanations.

\section*{Learner-generated Explanations in Visual and Verbal Formats}

The cognitive processes underlying the development of understanding may differ for visual and verbal explanations. Language has words for some parts, configurations, actions, and causes, but complex and complete descriptions of spatial and dynamic systems can be difficult to produce. Visualizations can readily depict the parts, shape, and configuration of a system, but it may be more difficult to depict the operation of a system, its functionality, and its causal mechanisms. Of course, the configuration provides clues for the system's operation and causality, and visual information can be supplemented with non-depictive diagrammatic devices, notably arrows (Heiser \& Tversky, 2006; Tversky et al., 2000, Tversky, 2002, 2011). Importantly, visual explanations demand completeness. Like other types of models, all of the essential parts of a system need to be represented in the proper configuration for it to work. In this way, drawings provide a visual check for completeness that verbal descriptions do not require. Inferences can then be made from diagrams that preserve and map the parts and configuration of the represented system or process. In an experiment that asked students to take notes while reading a text that they could later use to answer questions about the text, many students used only language, but those who made diagrams performed better (Schneider et al., 2010). Furthermore, requiring diagrams benefited all students.
Some researchers have demonstrated visual explanations' superiority over written explanations. Gobert \& Clement (1999) investigated the effectiveness of student-generated diagrams versus student-generated summaries on understanding plate tectonics after reading an expository text. Students who generated diagrams scored significantly higher on a post-test measuring spatial and causal/dynamic content, even though the diagrams contained less domainrelated information. Hall, Bailey, \& Tillman (1997) showed that learners who generated their own illustrations from text performed equally as well as learners provided with text and illustrations. Both groups outperformed learners only provided with text. In a study concerning the law of conservation of energy, participants who generated drawings scored higher on a post-test than participants who wrote their own narrative of the process (Edens \& Potter, 2003). In addition, the quality and number of concept units present in the drawing/science log correlated with
performance on the post-test. Van Meter (2001) found that drawing while reading a text about Newton's Laws was more effective than answering prompts in writing. Finally, Witherspoon et al. (2007) showed that generating external representations while studying the circulatory system increased scores compared to re-reading the provided text.

\section*{The Role of Spatial Ability in Learner-generated Explanations}

Developing an ability to visually manipulate a model of scientific processes is complicated. In constructing a visual representation of a scientific process, people may need to first imagine actions. Kozhevnikov, Hegarty, \& Mayer (2002) found that low spatial ability participants interpreted graphs as pictures, whereas high spatial ability were able to construct more schematic images and maniupulate them spatially. Hegarty \& Just (1993) found that the ability to mentally animate mechanical systems correlated with spatial ability, but not verbal ability. In their study, low spatial ability participants made more errors in movement verification tasks. However, Leutner, Leopold, \& Sumfleth (2009) found no effect of spatial ability on the effectiveness of drawing compared to mentally imagining text content.

\section*{Experiment 1: Explaining the Function of a Bicycle Tire Pump}

\section*{Method}

Participants Participants were \(1277^{\text {th }}\) and \(8^{\text {th }}\) grade students, ages 12-14, enrolled in an independent school in New York City. Of the 127 students, 59 were females, and 68 were males.

Materials Each participant was given a 12 -inch Spalding bicycle tire pump, a blank \(8.5 \times 11\) sheet of paper, a 16 question post-test, and the Vandenberg-Kuse Mental Rotation Test (MRT). Half of the participants received instructions to create a verbal explanation in writing; the other half received instructions to create a visual explanation in a drawing.

Procedure On the first of two non-consecutive school days, participants completed the MRT as a whole-class activity. Participants were read aloud the instructions, and were given untimed practice on several items. They were then given three minutes to complete items 1-10, and an additional three minutes to complete items 11-20. On the second day, participants were given the pump and instructions to try to understand how it worked. This segment was untimed. The next set of instructions asked students to verbally explain how the pump worked (in words) or to visually explain how the pump worked (in a drawing). Upon completion of the explanation, participants were given the 16 question post-test.

\section*{Coding}

Coding for Structure and Function. A maximum score of twelve points was awarded for the inclusion and labeling of six structural components: chamber, piston, inlet valve, outlet valve, handle, and hose. Information was coded as functional if it depicted or described the function/movement of an individual part, or the way multiple parts interact. There was no maximum imposed on the number of functional units.
Coding of Essential Features. Explanations were also coded for the inclusion of information essential to its function according to a four-point scale (adapted from Hall, Bailey, \& Tillman, 1997). One point was given if both the inlet and the outlet valve were clearly present in the drawing or described in writing, one point was given if the piston inserted into the chamber was shown or described to be airtight, and one point was given for each valve if they were shown or described to be opening/closing in the correct direction. The maximum score for essential features was four points.
Coding of Invisible Features. The presence of three invisible features (the inlet valve, the outlet valve, and the movement of air) were coded separately, with one point given for the presence of each valve, and three points given for movement of air (entering, moving through, and exiting the pump). The maximum score for invisible features was thus five points.
Coding Visual Elements: Arrows and Multiple Steps. Arrows were coded for three purposes: label for a part or action, to show motion, or to indicate sequence. Each use of arrows was coded for one of these purposes and a score tallied for each use. The use of multiple steps/frames was used to show starting and ending positions, and change in location of parts of the pump and air.

\section*{Results}

Spatial ability. Participants scores' on the MRT were used to divide participants into low and high spatial ability groups based on a median split in the data. Scores on the MRT range from 0-20; the mean score for participants was 10.56, and the median was 11 . Scores were significantly higher for males \((M=13.5, S D=4.4)\) than for females \((M=\) \(8.8, \mathrm{SD}=4.5), \mathrm{F}(1,126)=19.07, \mathrm{p}<.01\).
Structure and Function. Both visual and verbal explanations contained from two to ten structural components. Visual explanations contained a significantly greater number of structural components \((\mathrm{M}=6.05, \mathrm{SD}=2.76)\) than verbal components \((M=4.27, S D=1.54), F(1,126)=20.53\), \(\mathrm{p}<.05\), while there was no difference in the number of expressed functional components between visual and verbal explanations.
Essential Features. Scores for the inclusion of essential information were significantly higher for visual explanations \((\mathrm{M}=1.78, \mathrm{SD}=1.0)\) than for verbal explanations \((\mathrm{M}=1.20, \mathrm{SD}=1.21), \mathrm{F}(1,126)=7.63\), \(\mathrm{p}<.05\). No significant differences were found between low ( \(\mathrm{M}=1.34, \mathrm{SD}=1.04\) ) and high spatial participants \((\mathrm{M}=\)
\(1.45, \mathrm{SD}=1.2\) ). Essential features were also found to positively correlate with delayed post-test scores, \(r=.197\), \(\mathrm{p}<.05\) ).
Invisible Features. Scores for the inlet valve were higher for visual explanations \((\mathrm{M}=.67, \mathrm{SD}=.45)\) than verbal explanations ( \(\mathrm{M}=.51, \mathrm{SD}=.5\) ), however the effect was only marginally significant, \(\mathrm{F}(1,126)=3.13, \mathrm{p}=.07\). Scores for air movement also showed a marginally significant difference, \(F(1,126)=2.93, p=.09\), with visual explanations \((\mathrm{M}=2.35, \mathrm{SD}=1.28)\) containing a greater number than verbal explanations \((M=1.88, \mathrm{SD}=1.45)\). No significant differences between visual ( \(\mathrm{M}=.92\), \(\mathrm{SD}=.43\) ) and verbal explanations \((M=.79, S D=.65)\) were found for the outlet valve. Analysis of the invisible parts between low and high spatial participants also failed to show any significant differences in the inclusion of the inlet valve, the outlet valve, or air movement. Finally a total score for the inclusion of invisible parts was calculated for each participant by totaling the scores for the inlet valve, the outlet valve, and for air movement. The mean score was 3.26, \(\mathrm{SD}=1.25\). The data was analyzed using linear regression, and revealed that the total score for invisible parts significantly predicted scores on the post-test, \(\mathrm{F}(1\). \(118)=3.80, \mathrm{p}=.05\).
Multiple Steps. The number of steps used by participants ranged from one to six. Participants whose explanations contained more than a single step scored significantly higher ( \(\mathrm{M}=.76, \mathrm{SD}=.18\) ) on the post-test than participants whose explanations consisted of a single step \((\mathrm{M}=.67, \mathrm{SD}=.19)\), \(\mathrm{F}(1,126)=5.02, \mathrm{p}<.05\).
Learning Outcomes. Scores on the post-test by group and spatial ability are shown in Figure 1. A test of the overall interaction between group and spatial ability was significant, \(F(1,124)=4.094, p<.01\). In particular, low spatial participants who generated verbal explanations had significantly lower scores \((M=.609, S D=.145)\) than low spatial participants who drew explanations ( \(\mathrm{M}=.716\), \(\mathrm{SD}=\) .121). Analyzing structure and function questions separately on the post-test found no differences in performance between low and high spatial participants on structural questions. However, analyzing performance on functional questions found a significant effect: low spatial participants who generated verbal explanations \((\mathrm{M}=.502, \mathrm{SD}=.194)\) scored significantly lower than low spatial participants that drew ( \(\mathrm{M}=.678, \mathrm{M}=.122\) ), see Figure 2.

\section*{Discussion}

The results of Experiment 1 show that low spatial ability participants were able to learn as successfully as high spatial ability participants when they first generated an explanation in a visual format. Importantly, this result was particularly strong for functional understanding. Visual explanations were more likely to contain certain invisible features of the pump, such as the valves. Including the inlet valve and attempting to explain its function is crucial because then it is performing its function it is inside the chamber and air entering or exiting cannot be felt by the user.


Figure 1: Scores on the post-test, by group and spatial ability.


Figure 2: Scores on functional questions on the post-test, by group and spatial ability.

As mentioned previously, drawing encourages completeness. They force learners to decide on the size, shape, and location of parts/objects, and how the parts are related. Understanding the "hidden" function of the invisible parts is key to understanding the function of the entire system and requires an understanding of how both the visible and invisible parts interact. The visual format may have been able to elicit components and concepts that are invisible and difficult to integrate into the formation of a mental model.

Finally, an analysis of the visual explanations revealed that \(67 \%\) also added written components to accompany their explanation. Arguably, some types of information may be difficult to depict visually, and our verbal language has many possibilities that allow for specificity. Indeed, several studies by Mayer and colleagues have found that
understanding a system is enhanced when text and pictures are presented simultaneously to learners (e.g. Mayer \& Gallini, 1990).

The utility of visual explanations may differ for scientific phenomena that are more abstract, or contain elements that are invisible due to their scale. '

\section*{Experiment 2: Explaining the Process of Chemical Bonding}

\section*{Method}

Participants Participants were \(1268^{\text {th }}\) grade students, ages 13-14, enrolled in an independent school in New York City. Of the 126 students, 58 were females, and 68 were males.

Materials Each participant was given an immediate posttest, a delayed post-test, a blank \(8.5 \times 11\) sheet of paper, and the Vandenberg-Kuse Mental Rotation Test (MRT). Half of the participants received instructions to create a verbal explanation in writing; the other half received instructions to create a visual explanation in a drawing. In addition, the experimenter showed all participants a pre-recorded video lesson on bonding ( 13 minutes, 22 seconds long). The video began with a brief review of atoms and their structure, and introduced the idea that atoms combine to form molecules. Next, the lesson discussed how location in the periodic table affects behavior and reactivity of atoms, and makes atoms more or less likely to gain, lose, or share electrons. Examples of atoms, their valence shell structure, stability, charges, transfer and sharing of electrons, and the formation of ionic, covalent, and polar covalent bonds were discussed. The immediate post-test and delayed post-test each consisted of seven multiple-choice items and three freeresponse items.

Procedure On the first of three non-consecutive school days, participants completed the MRT as a whole-class activity, following the same procedures as Experiment 1. On the second day, participants viewed the recorded lesson on chemical bonding. They were instructed to pay close attention to the material but were not allowed to take notes on material presented in the video. Immediately following the video, participants were administered the immediate post-test of chemical bonding knowledge. Participants were given twenty minutes to complete the test; all participants finished within this time frame. On the third day, the particpants were randomly assigned to either the visual or verbal condition. The next set of instructions asked students to either visually or verbally explain how atoms bond and how ionic and covalent bonds differ). Upon completion of the explanation, participants were given the delayed posttest.

\section*{Coding}

Coding for Structure and Function. Visual and verbal explanations were coded for structural and functional
components. Table 1 and Table 2 show the components that were coded for structure and function, respectively.

Table 1: Coding Guide for Structure
Structural Components (1 pt. each)
Atoms with the correct number of electrons/valence
electrons
Atoms with the correct charges (magnitude,
positive/negative)
Bond between appropriate elements (i.e. between non-
metals for covalent molecules and between a metal and a
non-metal for ionic molecules)
Ionic bonds depicted/described as crystalline structure
Covalent bonds depicted/described as individual
molecules

Table 2: Coding Guide for Function
Functional Components (1 pt. each)

Transfer of electrons in ionic bonds
Sharing between atoms in covalent bonds
Attraction between ions of opposite charges
Outcome of bonding shows atoms with stable valence electron shell configurations.

Outcome of bonding shows molecules with overall neutral charge

Coding system for arrows. Arrows were present in \(92 \%\) of visual explanations. Their use was categorized into the use of arrows as labels and to show movement/action. Each use was tallied for each explanation.
Coding system for the use of specific examples. Explanations were coded for the use of specific atoms, such as NaCl to illustrate ionic bonding.
Coding for the use of multiple representations. Explanations were coded as symbolic (e.g. NaCl ), atomic (showing structure of atom(s), and macroscopic (visible).

\section*{Results}

Spatial ability. As in Experiment 1, participants' scores on the MRT were used to divide participants into low and high spatial ability groups based on a median split in the data. Scores were significantly higher for males \((M=12.5, S D=\) 4.8) than for females \((\mathrm{M}=8.0, \mathrm{SD}=4.0), \mathrm{F}(1,125)=\) 24.49, \(\mathrm{p}<.01\).

Structure and Function. The maximum score for structural and functional information was five points. Visual explanations contained a significantly greater number of structural components \((M=2.81, S D=1.56)\) than verbal
components \((\mathrm{M}=1.30, \mathrm{SD}=1.54), \mathrm{F}(1,125)=13.69\), \(\mathrm{p}<.05\), while there was no difference in the number of expressed functional components between visual and verbal explanations. Structural information was more likely to be depicted in pictures \((M=3.38, S D=1.49)\) than described in words \((\mathrm{M}=.429, \mathrm{SD}=1.03), \mathrm{F}(1,62)=21.49, \mathrm{p}<.05\), but functional information was equally likely to be expressed in pictures \((\mathrm{M}=1.86, \mathrm{SD}=1.10)\) and words \((\mathrm{M}=1.71, \mathrm{SD}=\) 1.87). Functional information in words added to visual explanations significantly predicted scores on the post-test, \(\mathrm{F}(1,62)=21.603, \mathrm{p}<.01\). As in Experiment 1, there were no significant differences in the amount of structural information contained in explanations created by low and high spatial ability participants. However, explanations created by high spatial participants contained significantly more functional components, \(\mathrm{F}(1,125)=7.13, \mathrm{p}<.05\).
Arrows. \(83 \%\) of visual explanations contained arrows. The use of arrows was positively correlated with scores on the post-test, \(\mathrm{r}=.293, \mathrm{p}<.05\).
Specific examples. High spatial participants \((\mathrm{M}=1.6, \mathrm{SD}=\) .69) used specific examples in their explanations more often than low spatial participants \((\mathrm{M}=1.07, \mathrm{SD}=.79)\). The difference was marginally significant \(\mathrm{F}(1.125)=3.65\), \(\mathrm{p}=.06\). There were no significant differences in the use of specific examples between visual and verbal groups. The inclusion of a specific example was positively correlated with scores on the delayed post-test, \(\mathrm{r}=.555, \mathrm{p}<.05\).
Multiple representations. Multiple representations were included in \(65 \%\) of the explanations. Participants generated significantly more when creating visual explanations ( \(\mathrm{M}=\) 1.79 , \(\mathrm{SD}=1.20\) ) compared to verbal explanations ( \(\mathrm{M}=\) \(1.33, \mathrm{SD}=.48), \mathrm{F}(1,125)=6.03, \mathrm{p}<.05\). However, the use of multiple representations did not significantly correlate with delayed post-test scores.

Learning outcomes. The immediate post-test was scored so that the maximum score was ten points. Each of the seven multiple choice questions and three free-response questions was given one point for the correct answer. The mean score (defined by proportion correct) on the immediate post-test was \(.463, \mathrm{SD}=.469\). Scores did not differ significantly between participants in the visual group ( \(\mathrm{M}=.486, \mathrm{SD}=.308\) ) and the verbal group \((\mathrm{M}=.443\), \(\mathrm{SD}=\) \(.260), \mathrm{F}(1,125)=.740, \mathrm{p}>.05\). Scores between high spatial ( \(\mathrm{M}=.532, \mathrm{SD}-.421\) ) and low spatial participants \((\mathrm{M}=.402\), \(\mathrm{SD}=.390)\) also did not differ significantly, \(\mathrm{F}(1,125)=\) \(2.72, \mathrm{p}>.05\).

The mean score on the delayed post-test (after participants generated explanations) was .704 , \(\mathrm{SD}=.299\). Participants in the visual group improved significantly from the immediate post-test \((\mathrm{M}=.822, \mathrm{SD}=.208), \mathrm{F}(1,125)=51.24, \mathrm{p}<.01\), Cohen's \(d=1.27\). Participants in the verbal group also showed significant increases from the immediate post-test \((\mathrm{M}=.631, \mathrm{SD}=.273), \mathrm{F}(1,125)=15.796, \mathrm{p}<.05\), Cohen's \(d=71\).

A comparison of the delayed post-test scores between groups found significant differences. Figure 3 shows scores on the post-test by group and spatial ability. Participants
generating visual explanations \((\mathrm{M}=.822, \mathrm{SD}=.208)\) scored higher on the post-test than participants generating verbal explanations \((\mathrm{M}=.631, \mathrm{SD}=.273), \mathrm{F}(1,125)=19.707\), \(\mathrm{p}<.01\), Cohen's \(d=.88\). In addition, high spatial participants ( \(\mathrm{M}=.824, \mathrm{SD}=.273\) ) scored significantly higher than low spatial participants \((\mathrm{M}=.636, \mathrm{SD}=.207), \mathrm{F}(1,125)=19.94\), \(\mathrm{p}<.01\), Cohen's \(d=.87\) (Figure 4-12). The results of the test of the interaction between group and spatial ability was not significant. A separate analysis comparing performance on multiple choice questions and free response questions did not show any differences between visual and verbal groups or between low and high spatial ability groups.


\section*{Discussion}

The results of Experiment 2 supported those of Experiment 1: learner-generated visual explanations provided an advantage over learner-generated verbal explanations. Visual explanations resulted in higher scores on the post-test for both low spatial and high spatial participants.

No difference was found between low and high spatial participants in the amount of structural information contained in the explanations, but high spatial participants included more function, were more likely to use specific examples, and scored higher on the delayed post-test.

An interesting finding of Experiment 2 was that the use of arrows significantly correlated with scores on the delayed post-test. How does their use lead to greater understanding? Arrows were most often used to label structure, or to label an action. They were also used to differentiate an initial versus and ending state, to show change. Previous research has shown arrows to serve a number of purposes. Notably, studies have shown the addition of arrows able to convey functional information in a structural diagram (Heiser \& Tversky, 2006). While the purpose of this study was to examine student-generated explanations, these results support those of previous work that shows when arrows are used in diagrams in a way that encourages the development of mental models, they become more effective.

\section*{Conclusion}

Experiment 1 showed that learning about a physical bike pump and generating visual explanations was primarily a benefit to low-spatial ability participants. The measures of learning (from a true/false post-test) were of course limited, and it is possible that higher-order learning by high spatial ability participants was not revealed. Experiment 2 showed that viewing a class lesson on chemical bonding and generating visual explanations benefited both low and high spatial ability students, although to different degrees (high spatial ability participants scored significantly higher on the post-test). In the generation of visual explanations, learners use the information they gather from new material to create internal representations that become richer with the integration of verbal and non-verbal representations, forming a mental model that then informs and direct the creation of visual explanations. Learners with high spatial ability are more adept at forming and manipulating mental images; this may make the generation of visual explanations easier for them. Learners with low spatial ability may find the task difficult, but may be able to be more successful with generating visual explanations if support is provided.

Together, the results from the two experiments support the use of learner-generated visual explanations as a learning strategy in science. Future studies should explore how this strategy mediates the comprehension of concepts presented in physical models, experiments, and textbooks, and posttest performance. Students live in a macroscopic world, where objects have mass and occupy space. Understanding "invisible" processes in science, then, presents a challenge. Generating visual explanations through drawing is likely an underused method of monitoring and supporting students' understanding of scientific concepts.

\section*{References}

Gobert, J.D., \& Clement, J.J. (1999). Effects of studentgenerated diagrams versus student-generated summaries on conceptual understanding of causal and dynamic knowledge in plate tectonics. Journal of Research in Science Teaching, 36(1), 39-53.
Hall, V. C., Bailey, J., \& Tillman, C. (1997). Can studentgenerated illustrations be worth ten thousand words? Journal of Educational Psychology, 89(4), 677-681.
Hausmann, R. G. M., \& Vanlehn, K. (2007). Explaining self-explaining: A contrast between content and generation. In Luckin, R., Koediner, K. R., \& Greer, J. (Eds). Artificial intelligence in education: Building technology rich learning contexts that work, 158 (pp. 417424). Amsterdam: IOS Press.

Hegarty, M., \& Just, M. A. (1993). Constructing mental models of machines from text and diagrams. Journal of Memory and Language, 32, 717-742.
Heiser, J., \& Tversky, B. (2006). Arrows in comprehending and producing mechanical diagrams. Cognitive Science, 30, 581-592.

Johnson, C. I., \& Mayer, R. E. (2009). A testing effect with multimedia learning. Journal of Educational Psychology, 101, 621-629.
Johnson, C. I., \& Mayer, R. E. (2010). Applying the selfexplanation principle to multimedia learning in a computer-based game-like environment. Computers in Human Behavior, 26, 1246-1252.
Kozhevnikov, M., Hegarty, M., \& Mayer (2002). Spatial abilities in kinematics problem solving. In M. Anderson, B. Meyer \& P. Olivier (Eds). Diagrammatic Representation and Reasoning (pp. 155-171). Berlin: Springer-Verlag.
Leutner, D., Leopold, C., \& Sumfleth, E. (2009). Cognitive load and science text comprehension: Effects of drawing and mentally imagining text content. Computers in Human Behavior, 25, 284-289.
Mayer, R.E. \& Gallini, J. K. (1990). When is an illustration worth ten thousand words? Journal of Educational Psychology, 82, 715-726.
Schwamborn, A., Mayer, R. E., Thillmann, H., Leopold, C., \& Leutner, D. (2010). Drawing as a generative activity and drawing as a prognostic activity. Journal of Educational Psychology, 102, 872-879.
Tversky, B., Zacks, J., Lee, P. U., \& Heiser, J. (2000). Lines, blobs, crosses, and arrows: Diagrammatic communication with schematic figures. In M. Anderson, P. Cheng, \& V. Haarslev (Eds.), Theory and application of diagrams, (pp. 221-230). Berlin, Germany: Springer.
Tversky, B. (2002). What do Sketches Say about Thinking? 2002 AAAI Spring
Symposium, Sketch Understanding Workshop, Stanford University, AAAI Technical Report SS 02-08, March, 2002, pp. 148-151.
Tversky, B. (2011). Visualizing Thought. Topics in Cognitive Science, 3, 499-535.

\title{
Do You Mean What You Say? The Effect of Uncertainty Avoidance on the Interpretation of Probability Expressions - A Comparative Study between Spanish and German
}

\author{
Franziska Bocklisch (franziska.bocklisch@psychologie.tu-chemnitz.de) Anne Georg (anne.georg@s2009.tu-chemnitz.de) Steffen F. Bocklisch (steffen.bocklisch@etit.tu-chemnitz.de) \\ Josef F. Krems (josef.krems@ psychologie.tu-chemnitz.de) \\ Wilhelm-Raabe-Str. 43, Chemnitz University of Technology, Germany
}

\begin{abstract}
In times of globalization, differences between cultures and in the interpretation of linguistic terms can lead to misunderstanding in communication. The present study focuses on the influence of cultural dimensions, especially uncertainty avoidance, on the interpretation of verbal probability expressions. It is hypothesized that uncertainty avoidance has an effect on the interpretation of uncertainty expressions. Therefore, Spanish and German participants were asked (1) for uncertainty avoidance and (2) to estimate numerical equivalents for 12 verbal probability expressions (e.g., possible). The estimation data were modeled using fuzzy membership functions. Results neither show differences in uncertainty avoidance nor in the interpretation of the probability expressions between these two languages. Possible reasons and future research perspectives are discussed.
\end{abstract}

Keywords: Uncertainty Avoidance, Linguistic Terms, Probability Expressions, Fuzzy Membership Functions.

\section*{Introduction}

In times of the most severe economic crisis the European Union ever experienced the following situation is likely to happen: Imagine economic experts from Spain and Germany discussing about the risk of a financial breakdown of Spain's banks in the near future: Expert 1 argues: "I am sure that we will have a crash next year." and expert 2 responds: "It's possible but I am still optimistic. Probably we can manage it." In many social interactions between persons or groups of different countries the language for communication might not be the native language of the speakers (e.g., Spanish or German) but a foreign one (such as English). Nevertheless, one is likely to "think" in terms of the own language, choose words that express the own intention best and then "translate" them into the foreign language. But given the vagueness of natural languages (e.g., Teigen \& Brun, 2003; Budescu, Karelitz \& Wallsten, 2003; Bocklisch, Bocklisch \& Krems, 2010; 2012) and potential cultural differences (e.g., Hofstede, 1980; 2001; Hofstede, Hofstede, \& Minkov, 2010) we can not necessarily be sure that the translation reflects the meaning precisely and therefore, communication partners may misunderstand each other. The example described above is prototypical for situations in which forecasts and decisions
have to be done under uncertain circumstances. Often, the decision makers express their beliefs using verbal probabilities such as "probable" or "possible". There is a risk of miscommunication because of the considerable variation of people's interpretation of the meaning of linguistic terms (LTs) (e.g., Karelitz \& Budescu, 2004). Further, misunderstandings may lead to wrong decisions with undesirable consequences. Therefore, avoiding misunderstandings and improving interpersonal communication is highly relevant in a globalized world.

The present paper highlights the questions (1) if LTs (probability expressions) of different languages (Spanish vs. German) are interpreted differently by native speakers and (2) whether there exists a cultural influence on the LTs interpretation, namely, concerning the dimension of uncertainty avoidance (Hofstede, 1980). Methodologically, we use fuzzy membership functions (MFs) to formalize and compare the vague meaning of the LTs (e.g., Bocklisch, Bocklisch \& Krems, 2012).

\section*{Culture and Language}

One influential paradigm in intercultural research is Hofstedes model of cultural differences. In its original form (Hofstede, 1980), he differentiated four cultural dimensions on which cultures vary and, therefore, may be compared: individualism/collectivism, power distance, uncertainty avoidance and masculinity/femininity. Concerning the investigation of probability expressions, we consider the third dimension "uncertainty avoidance" (UA) as especially interesting and, therefore, focus on UA in the following. In short, UA is determined by the extent to which members of a society become nervous or insecure about situations that seem to be unpredictable, unstructured and uncertain.

There are already evidences that culture and language are important concerning the interpretation of LTs, for instance, Doupnik and Richter (2003) already found differences in the interpretation of German and English uncertainty expressions between native speaking auditors in two cultural areas, namely, America and Germany. We will focus on two languages, Spanish and German, which developed in the same European cultural region. According to Hofestede (2001), Spain and Germany are both countries with high UA values and, at the same time, differ remarkably in UA \(\left(\mathrm{UA}_{\text {Spain }}>\mathrm{UA}_{\text {Germany }}\right)\). In our study we (1) expect to replicate Hofstedes results and find high UA scores for both
languages as well as significant differences. Furthermore, we hypothesize that the probability expressions should be interpreted differently. We expect that because the context seems to influence the meaning of expressions (e.g., Weber \& Hilton, 1990). We think that culture may be determined as a very global form of context. Therefore, if the scores of \(\mathrm{UA}_{\text {Spain }} \neq\) UA Germany the LTs might also be interpreted differently.

\section*{Numerical Translation of Linguistic Terms}

To determine the meaning of LTs researchers have developed procedures for the numerical translation of verbal expressions (e.g., Simpson, 1944; Beyth-Marom, 1982; Budescu, Karelitz \& Wallsten, 2003; Bocklisch, Bocklisch \& Krems, 2010; 2012). Generally, results of these studies show (1) that mean estimates for linguistic expressions are similar and that they have stable meaning (Simpson, 1944 and Hakel, 1968). At the same time there exists large interindividual variability in the interpretation of verbal expressions (for reviews see Pepper, 1981; Teigen \& Brun, 2003). Another important outcome is (2) that fuzzy set theory (Zadeh, 1965) proved especially useful for describing the vague meaning of LTs by modelling them using fuzzy MFs (e.g., Zimmer, 1984; Budescu \& Wallsten, 1995; Budescu et al., 2003, Bocklisch et al., 2012).

We used the two-step translation procedure outlined in Bocklisch et al. (2012) for the numerical translation of the probability expressions. This procedure also uses fuzzy MFs to model the empirical estimates of participants. MFs are truth value functions. In this study, the membership value \((\mu)\) represents the degree of truth that a numerical estimate fulfils a specific criterion represented by a LT (e.g., the numerical probability "in 70 of 100 cases" belongs to the linguistic probability expression probable). We use a standardized \(\mu\) ranging from 0 (no membership) to 1 (full membership). Furthermore, the procedure has been proven useful for the translation of LTs (e.g., probability expressions: Bocklisch et al., 2010; symptom intensities in medical contexts: Bocklisch, Stephan, Wulfken, Bocklisch \& Krems, 2011; and frequency expressions: Bocklisch et al., 2012), and as basis for evaluating and choosing verbal response labels for questionnaire scales (e.g., for the StateTrait Anxiety Inventory (STAI-T) see Bocklisch, Bocklisch \& Krems 2011). It includes (1) an empirical estimation method in which participants assign typical, minimum and maximum correspondence numbers to presented words, and (2) a fuzzy approach for the analysis of data and the generation of MFs - specifically, parametric MFs of the potential type. This procedure is very efficient as only three numerical values are estimated, and is easily understood because the semantic meaning of estimation points is implicitly clear to participants (e.g., minimum and maximum correspondence values \(=\) borders of LTs meanings). Semantic comprehensibility, as such, makes participants' estimates understandable even if they have no theoretical knowledge of the concept of fuzzy membership.

Such intuitive understanding is also advantageous when an estimation method is used for participants not highly trained in estimation tasks. For more details concerning the theoretical justification or parameter estimation see Bocklisch (1987) or Bocklisch, Groß, Bocklisch and Krems (submitted).

\section*{Method}

\section*{Participants}

We collected data of 147 German participants ( 51 males) with an average age of \(27.5(S D=9.9)\) and 21 Spanish participants (12 males) with a mean age of 34.8 ( \(S D=\) 12.9). All subjects were native speakers. Four (German sample) vs. one (Spanish sample) persons were raised bilingual. Most of the participants were students and the German students received course credits for participation.

\section*{Design, Material and Procedure}

The study was quasi-experimental employing an online questionnaire in two languages (Spanish vs. German). The material was carefully constructed to make sure that both questionnaire versions were equivalent in meaning, especially concerning the 12 probability expressions (see Table 1) that were presented in the questionnaire in a random order. The 12 LTs were chosen from a pool of 47 probability expressions gained from a literature review and dictionaries. These 12 words are rather unambiguous in meaning and are frequently used in both languages. The translation from German to Spanish was done independently by two Spanish that also speak German very well (international C1 and B2 level). The questionnaire had four main parts: (1) a short introduction, (2) the estimation part for the translation of the probability expressions (Bocklisch et al., 2012), (3) the uncertainty avoidance scale including ten items (e.g., "I would like to have more control about the future.") (Mealy, Stephan, Abalakina-Paap, 2006; see also Fahmie, 2012) and (4) a few questions concerning demographic data.

\section*{Results}

\section*{Uncertainty Avoidance}

We found UA scores of \(M_{\text {German }}=38.8\left(S D_{\text {German }}=7.0\right)\) and \(M_{\text {Spanish }}=39.5\left(S D_{\text {Spanish }}=5.8\right)\) and, therefore, no significant differences \((t(166)=-.425, p=0.67)\).

\section*{Meaning of Linguistic Terms}

The results for the meanings of LTs are structured starting with descriptive statistics of the data, and then the fuzzy MFs and discriminatory power values ( \(d p\) ) indicating how similar the MFs and, hence, the meaning of the LTs are. As outlined in Bocklisch et al. (2012, p.148), we present descriptive statistics and MFs for purposes of completeness and comparison, even though we believe that MFs are more suitable for describing the meaning of vague LTs. We have
to emphasize that statistical and fuzzy analyses are two approaches that should be understood independently, because fuzzy MFs, by definition, do not refer to probability theory and statistics. Although some parameters of our MF type can be interpreted statistically (e.g., representative values \((r)=\) the arithmetic mean), a MF is not a probability density function and conventional requirements (i.e., the integral of the variable's density is equal to 1 ) are not valid.
Descriptive Statistics. Table 1 shows a comparison of the descriptive statistics (i.e., means and standard deviations) for the Spanish and German LTs. The minimal and maximal estimates were used for modelling of the MFs and are not reported here.

The order of the means of the typical estimates (equals \(r\) of the MFs, see below) presented in Table 1 is the same for the Spanish and German LTs and even the numerical values are very similar. The largest difference is 12.5 for probable.
(MFs shapes) in both languages. For instance, the MFs for possible (MF 7) are almost congruent.

Furthermore, we find the same pattern as reported in Bocklisch et al. (2010 and 2012) that MFs of LTs at the scales borders (e.g., impossible) are smaller in extent and, therefore, more precise in meaning compared to mid-scale LTs (e.g., possible). MFs 8 and 12 (probable and certain) show the largest differences. This is also confirmed by the \(d p\) s (see Table 1). For MFs 12 (certain) \(d p=0.71\) reaches the threshold of a remarkable difference while all other \(d p\) s are smaller than 0.5.

Table 1: Descriptive Statistics (Typical Estimates) and Discriminatory Power Values ( \(d p\) )
\begin{tabular}{lllllll}
\hline Probability Expressions & \multicolumn{4}{l}{ German } & \multicolumn{2}{l}{ Spanish } \\
English Translation (Original in German / Spanish) & \(M\) & \(S D\) & \(M\) & \(S D\) & \(d p\) \\
\hline 1 & Impossible (Unmöglich / Imposible) & 1.53 & 2.93 & 0.50 & 1.24 & 0.44 \\
2 & Very improbable (Sehr unwahrscheinlich / Muy improbable) & 7.48 & 4.98 & 8.95 & 7.41 & 0.16 \\
3 & Improbable (Unwahrscheinlich / Improbable) & 14.24 & 8.29 & 13.71 & 9.18 & 0.05 \\
4 & A small probability (Wenig wahrscheinlich / Poco probable) & 19.79 & 10.14 & 18.38 & 11.63 & 0.08 \\
5 & Uncertain (Unsicher / Inseguro) & 28.21 & 15.04 & 27.52 & 19.11 & 0.04 \\
6 & Maybe possible (Vielleicht möglich / Tal vez posible) & 36.69 & 16.54 & 35.24 & 15.53 & 0.10 \\
7 & Possible (Möglich / Posible) & 52.04 & 13.49 & 52.86 & 9.43 & 0.05 \\
8 & Probable (Wahrscheinlich / Probable) & 71.64 & 13.62 & 59.14 & 15.26 & 0.35 \\
9 & Rather certain (Ziemlich sicher / Bastante seguro) & 82.89 & 10.99 & 82.86 & 9.43 & 0.09 \\
10 & Very probable (Sehr wahrscheinlich /Muy probable) & 85.96 & 8.76 & 85.60 & 9.42 & 0.17 \\
11 & Highly probable (Höchstwahrscheinlich / Altamente probable) & 87.16 & 10.66 & 85.75 & 7.48 & 0.27 \\
12 & Certain (Sicher / Seguro) & 94.22 & 8.93 & 98.30 & 3.23 & 0.71 \\
\hline
\end{tabular}

Fuzzy Membership Functions. Figure 1 shows the potential MF for both the German and Spanish probability expressions. The MFs' peaks (marked with a vertical line), indicating the highest membership ( \(r\) ), are identical to the means shown in Table 1. Dps of the MFs are shown in Table 1. According to Bocklisch et al. (2012; p.149), \(d p\) values are defined by the overlapping area of two MFs and is standardized by taking values between 0 (MFs are identical) and 1 (no overlap at all). By definition, the larger the overlap, the smaller the \(d p\), and the more similar the meanings of the verbal expressions are. Concerning the interpretation of \(d p\), values \(\geq 0.7\) suggest that the MFs are considerably different (because overlap is \(<30 \%\) ).

For reasons of clarity and comprehensibility, the LTs with odd numbers are shown in the upper part and even numbers in the lower part of Figure 1. The MFs positions and shapes show that the meaning of a certain LT is highly similar regarding its typical meaning ( \(r \mathrm{~s}\) ) as well as vagueness

\section*{Discussion}

The results show that there is (1) no significant difference in the UA between Spanish and German participants. Hence, we could replicate only parts of Hofstedes results (Hofstede, 2001): Spain and Germany are both countries with high UA. But our results do not show remarkable differences between Spain and Germany in UA as we hypothesized according to Hofstede (2001). Concerning (2) the interpretation of probability expressions we did not find differences in the MFs between the Spanish and German sample except for LT certain and, by tendency, for probable. This is a surprise because we did not expect that the MFs of the Spanish and German LTs are so similar because the languages do not origin from the same language family. But in regard to the cultural background it seems reasonable and goes along with the idea that UA and interpretation of probability LTs may be connected. For future studies we suggest to collect data from a more representative sample. Especially the Spanish sample of this study was rather small.

\section*{Membership}


\section*{Membership}


Figure 1: Fuzzy MFs of Spanish and German LTs

Furthermore, a comparison of languages from cultures with larger differences in UA (e.g., Asian vs. European cultures) would be interesting to test the hypothesis whether UA influences the interpretation of LTs because the larger the differences in UA the more differences in LTs MFs should be.

As an additional result it was possible to compare the data of Bocklisch et al. (2010) that used the same translation procedure and the results of the German sample for seven LTs (see Table 2). The \(r\) s (means of the typical values) are almost identical (largest difference is 3.96 for probable) meaning that we could replicate their findings very good using the fuzzy MF approach.

Table 2: Comparison of Representative Values of German Sample and Bocklisch et al. (2010)
\begin{tabular}{lcc}
\hline \begin{tabular}{l} 
Probability Expressions \\
(Original German)
\end{tabular} & \begin{tabular}{c} 
German \\
Sample
\end{tabular} & \begin{tabular}{c} 
Bocklisch, et \\
al. (2010)
\end{tabular} \\
\hline Impossible & 1.53 & 1.44 \\
Very Improbable & 7.48 & 5.53 \\
Improbable & 14.24 & 11.68 \\
Possible & 52.04 & 51.49 \\
Probable & 71.64 & 67.68 \\
Very Probable & 85.96 & 83.95 \\
Certain & 94.22 & 96.28 \\
\hline
\end{tabular}

The results encourage the application of fuzzy MFs in technical, cognitive and language interaction systems. Concerning the methodological possibilities of MFs the implementation in technical systems can easily be done by implementing the parametric description that underlies the MFs. We see a high potential for this methodology for future research and application, for instance, in the field of intercultural communication.

\section*{Acknowledgments}

Special thanks to the students of Chemnitz University of Technology.

\section*{References}

Beyth-Marom, R. (1982). How probable is probable? A numerical translation of verbal probability expressions. Journal of Forecasting, 1, 257-269.
Bocklisch, S.F. (1987). Prozeßanalyse mit unscharfen Verfahren.VEB Verlag Technik Berlin.
Bocklisch, F., Bocklisch, S.F., \& Krems, J.F. (2010). How to translate words into numbers? A fuzzy approach for the numerical translation of verbal probabilities. In Hüllermeier, E., Kruse, R., \& Hoffmann, F. (Eds.), IPMU 2010, Lecture Notes in Artificial Intelligence 6178 (pp. 614-623). Berlin, Heidelberg, Springer.
Bocklisch, F., Bocklisch, S.F., \& Krems, J.F. (2011). The fuzziness of verbal response scales: The STAI-T Questionnaire. In Galichet, S., Montero, J., \& Mauris, G. (Eds.), Proceedings of the \(7^{\text {th }}\) Conference of the European

Society for Fuzzy Logic and Technology (EUSFLAT-2011) and LFA-2011 (pp. 592-597), Atlantic Press.
Bocklisch, F., Groß, S., Bocklisch, S.F., \& Krems, J.F. (submitted to Psychological Methods). Solving Problems of Vagueness and Equidistance in Verbal Rating Scales: a Fuzzy Approach.
Bocklisch, F., Stephan, M., Wulfken, B., Bocklisch, S.F., \& Krems, J.F. (2011). How medical expertise influences the understanding of symptom intensities - A fuzzy approach. In A. Holzinger \& K.-M. Simonic (Eds.), Proceedings of USAB 2011: Information Quality in e-Health, LNCS 7058.(pp. 703-706). Berlin, Heidelberg, Springer.

Bocklisch, F., Bocklisch, S.F., \& Krems, J.F. (2012). Sometimes, often and always - Exploring the vague meanings of frequency expressions. Behavior Research Methods, 44, 144-157.
Budescu, D.V. \& Wallsten, T.S. (1995). Processing linguistic probabilities: General principles and empirical evidences. The Psychology of Learning and Motivation, 32, 275-318.
Budescu, D.V., Karelitz, T.M., \& Wallsten, T.S. (2003). Predicting the directionality of probability words from their membership functions. Journal of Behavioral Decision Making, 16, 159-180.
Doupnik, T. S., \& Richter, M. (2003). Interpretation of Uncertainty Expressions: A Cross-national Study. Accounting, Organizations and Society, 28(1), 15-35. doi:10.1016/S0361-3682(02)00010-7.
Fahmie, A. (2012). Relationships among Uncertainty Avoidance, Individualism-Collectivism, and Usability of Personal Management Information Perceived by German and Indonesian Users (Dissertation). Chemnitz University of Technology, Germany.
Hakel, M.D. (1968). How often is often? American Psychologist, 23, 533-534.
Hofstede, G. (1980). Culture's Consequences: International Differences in Work-Related Values. Beverly Hills: Sage Publications.
Hofstede, G. (2001). Culture's Consequences: Comparing Values, Behaviors, Institutions, and Organizations Across Nations. Thousand Oaks, California: Sage Publications.
Hofstede, G., Hofstede, G. J., \& Minkov, M. (2010) Cultures and Organizations: Software of the Mind. New York: McGraw-Hill.
Karelitz, T.M., \& Budescu, D.V. (2004). You say "probable" and I say "likely": Improving interpersonal communication with verbal probability phrases. Journal of Experimental Psychology: Applied, 10(1), 25-41.
Mealy, M., Stephan, W. G., Abalakina-Paap, M. (2006). Reverence for Mothers in Ecuadorian and Euro-American Culture. Journal of Cross-Cultural Psychology, 37(4), 465484. doi:10.1177/0022022106288481

Pepper, S. (1981). Problems in the quantification of frequency expressions. In D.W. Fiske, (Ed.), Problems with language imprecision - New directions for methodology of social and behavioral science series (pp. 25-41). San Francisco, USA: Jossey-Bass.

Pepper, S., \& Prytulak, L.S. (1974). Sometimes frequently means seldom: context effects in the interpretation of quantitative expressions. Journal of Research in Personality, 8, 95-101.
Simpson, R.H. (1944). The specific meanings of certain terms indicating differing degrees of frequency. Quarterly Journal of Speech, 30, 328-330.
Teigen, K.H. \& Brun, W. (2003). Verbal expressions of uncertainty and probability. In D. Hardman (Ed.), Thinking: Psychological perspectives on reasoning, judgment and decision making (Ch. 7, pp. 125-145), Chichester, Wiley and Sons.
Wallsten, T.S., Budescu, D.V., Rapoport, A., Zwick, R., \& Forsyth, B. (1986). Measuring the vague meaning of probability terms. Journal of Experimental Psychology: General, 115, 348-365.
Wallsten, T.S., Budescu, D.V., Zwick, R., \& Kemp, S.M. (1993). Preferences and reasons for communicating probabilistic information in numerical or verbal terms. Bullet of the Psychonomic Society, 31, 135-138.
Weber, E.U., \& Hilton, D.J. (1990). Contextual effects in the interpretation of probability words: Perceived base rte and severity of events. Journal of Experimental Psychology: Human Perception and Performance, 16(4), 781-789.
Zadeh, L. A. (1965). Fuzzy sets. Information and Control 8, 338-353.
Zimmer, A. (1984). A model for the interpretation of verbal predictions. International Journal of Man-Machine Studies, 20, 121-134.

\title{
Why 100 Once Is Worse Than 10 Times 10: Dread Risks versus "Continuous" Risks
}

\author{
Nicolai Bodemer (bodemer@mpib-berlin.mpg.de) \\ Max Planck Institute for Human Development, Lentzeallee, 94, 14195 Berlin, Germany \\ Azzurra Ruggeri (ruggeri@mpib-berlin.mpg.de) \\ Max Planck Institute for Human Development, Lentzeallee, 94, 14195 Berlin, Germany \\ Mirta Galesic (galesic@mpib-berlin.mpg.de) \\ Max Planck Institute for Human Development, Lentzeallee, 94, 14195 Berlin, Germany
}

\begin{abstract}
People tend to react more strongly to a dread risk, a rare event that kills many people at once, than to a continuous risk, a relatively frequent event that kills many people over a longer period of time, even when both cause the same number of fatalities. This different reaction to the dread risk is often considered a bias, but we show that it is an ecologically rational strategy. In a series of simulations, we found evidence that dread risks affect the population more severely over time than continuous risks causing the same number of fatalities. This holds particularly true when the risks affect children and young adults who would have produced future offspring if they had survived longer.
\end{abstract}

Keywords: dread risk; continuous risk; risk perception; ecological rationality

\section*{Introduction}

Imagine two different risky events: One threatens to kill 100 people at once; the other threatens to kill 10 people every year over a period of 10 years. The first event represents a dread risk, a rare event that kills many people at once, such as a pandemic, an earthquake, or a terrorist attack. The second event represents a continuous risk, a relatively frequent event that kills many people over a longer period of time, such as diabetes, air pollution, or car accidents. Which of the two risks is more severe? Both events kill the same number of people and differ only with respect to the time frame. Yet, people react much more strongly to dread risks than to continuous risks, in terms of both perception and avoidance behavior (Gigerenzer, 2004, 2006; Slovic, 1987).

For instance, in reaction to the 9/11 terrorist attacks (a typical dread risk), many Americans avoided air travel and switched to their cars without considering that the risk of dying in a car accident (a continuous risk) is larger than the risk of an airplane terrorist attack, and even of dying in an airplane accident in general (Sivak \& Flannagan, 2003). The avoidance of flying and the elevated use of cars increased the number of fatal highway crashes after the \(9 / 11\) attacks (Gaissmaier \& Gigerenzer, 2012).

People's higher sensitivity to dread risks compared with continuous risks is often considered a bias: If the continuous risk causes the same number of fatalities, it should not be perceived as less dreadful. In this paper we offer an
alternative explanation to the assumption of biased minds and argue that a stronger reaction to dread risks is ecologically rational, because dread risks actually cause a larger cumulative reduction in the population size.

\section*{Previous Accounts}

Different hypotheses have been proposed to explain why people fear dread risks more than continuous risks. First, the psychometric paradigm (Slovic, 1987) suggests that high lack of control, high catastrophic potential, and severe consequences account for the increased risk perception and anxiety associated with dread risks. Second, people might lack knowledge about the statistical information underlying risks (Gigerenzer, Mata, \& Frank, 2009), in particular about the large number of fatalities caused by continuous risks. Third, because people estimate the frequency of a risk by recalling instances of its occurrence from their social circle or the media, they may overvalue relatively rare but dramatic risks and undervalue frequent, less dramatic risks (Hertwig, Pachur, \& Kurzenhäuser,2005; Lichtenstein, Slovic, Fischhoff, Layman, \& Combs, 1978). Fourth, according to the preparedness hypothesis, people are prone to fear events that have been particularly threatening to survival in human evolutionary history (Öhman \& Mineka, 2001). Given that in most of human evolutionary history people lived in relatively small groups, rarely exceeding 100 people (Hill et al, 2011; Lee \& DeVore, 1968), a dread risk, which kills many people at once, could potentially wipe out one's whole group. This would be a serious threat to individual fitness, as being in a group reduces predation risk, helps with finding food and hunting, and increases survival chances when injured (Dunbar \& Schultz, 2007; Krause \& Ruxton, 2002). In line with this hypothesis, Galesic and Garcia-Retamero (2012) found that people's fear peaks for risks killing around 100 people and does not increase if larger groups are killed.

\section*{A population-based perspective}

A different perspective reveals that dread risks lead to significantly worse short- and medium-term consequences than continuous risks, even if they do not eliminate a whole group. Thus, we focus not only on the overall number of
immediate fatalities, as in previous accounts, but also on (a) the population size over time, and (b) the role of the age group that is affected by the risky event. Note that a fatal event strikes twice: it kills a number of people immediately, and it reduces the number of future offspring by reducing the number of their potential parents. A risk that affects children and young adults will have stronger negative effects on future group growth than a risk that affects group members who are past their reproductive period. Dread risks such as pandemics, terrorist attacks, or nuclear accidents are more likely to strike children and young adults compared to many continuous risks such as diabetes, cancer, heart attack, or household accidents, which affect primarily older people (Statistisches Bundesamt Deutschland, 2012). For example, the H1N1 pandemic in 2009 was more likely to infect younger people, whereas older people were relatively immune, probably due to previous exposure to a similar virus strain (ECDC, 2009).

Hypothesis We hypothesize that dread risks cause larger cumulative losses on the population level than continuous risks. More specifically, we hypothesize that the number of people-years lost because of a dread risk is larger than the number of people-years lost because of a continuous risk, in particular when the event affects the younger age groups. People-years correspond to the number of people who live 1 year in the population. Hence, by killing a large number of children or young adults at once, dread risks not only deprive the society of their contribution in subsequent years, but they also remove the potential contribution of the offspring the victims could have had if they had survived longer.

To illustrate this hypothesis, consider first a very simplified example. Imagine a population of 40 people, uniformly distributed across four age groups:

Children and adolescents, aged 0-19 years: Pre-fertile generation that may produce offspring in the future.

Young adults, aged 20-39 years: Fertile generation that currently produces offspring.

Older adults, aged 40-59 years: Post-fertile generation.
Elderly adults, aged 60-79 years: Post-fertile generation.
Further assume that the population growth is constant and that every year each young adult produces exactly one offspring. This implies that the number of children at time point \(i, t_{i}\), corresponds to the number of young adults at time point \(i-1, t_{i-1}\). Moreover, at every \(t_{i}\) a generation shift takes place, so that the number of young adults at \(t_{i+1}\) corresponds to the number of children at \(t_{i}\), and so on for the other groups. Moreover, all elderly adults at \(t_{i-1}\) will be dead at \(t_{i}\). In the absence of any dread risk or continuous risk, the population is constant over time with \(N_{\text {total }}=40\) (see Figure 1).

What happens if a dread risk occurs at \(t_{1}\) that kills \(50 \%\) of the young adults (i.e., 5 young adults)? At \(t_{1}\), the total population is reduced to \(N_{\text {total }}=35\left(N_{\text {children }}=10, N_{\text {young adults }}=\right.\) \(\left.5, N_{\text {older adults }}=10, N_{\text {elderly adults }}=10\right)\). At \(t_{2}\) the population is further reduced to \(N_{\text {total }}=30\left(N_{\text {children }}=5, N_{\text {young adults }}=10\right.\),
\(N_{\text {older adults }}=5, N_{\text {elderly adults }}=10\) ), because the number of newborn offspring is smaller due to the fewer young adults. Finally, the population size settles at \(N_{\text {total }}=30\), with continuous fluctuation within the respective groups.

What happens if a continuous risk, a disease, occurs at \(t_{1}\) that kills five young adults over a period of five time steps (one young adult at every \(t_{\mathrm{i}}\), from \(t_{1}\) to \(t_{5}\) )? Note that the total number of fatalities directly caused by the risk is the same as in the dread risk scenario (i.e., 5). The total population is reduced to \(N_{\text {total }}=39\) at \(t_{1}\) and continues to decline until \(t_{6}\), where it finally corresponds to the size of the population hit by the dread risk.

In sum, the continuous risk takes five more generations to affect the population as severely as the dread risk. The difference in the cumulative losses caused in the population by the dread versus continuous risk, can be calculated by determining the area between the curves representing the difference in the cumulative population sizes of the two conditions (i.e., the difference in people-years over time). In the example in Figure 1, this integral is 20, meaning that the population hit by the dread risk lost 20 people-years more than the population experiencing the continuous risk.


Figure 1: Development of the population size when no risky event is present (baseline), and when a continuous risk (1 individual killed from t1 to t ) or a dread risk (5 individuals killed at t1) event occurs. A dread risk leads to a more immediate impact on cumulative population size that lasts longer compared with the continuous risk.

\section*{Simulation Set 1}

In the first set of simulations, we assumed a small population size, similar to groups in which people lived throughout most of evolutionary history (Lee \& DeVore, 1968). We manipulated whether the population growth rates
were constant, increasing, or decreasing, and which age group was exposed to a dread or to a continuous risk.

\section*{Method}

We set the total population to 160 people. The individuals were distributed equally across 80 years (i.e., there were 2 individuals for each age at t0) and across four age groups, as in the illustrative example above. Between conditions, we manipulated (a) whether a dread or a continuous risk occurred, (b) the population growth rate, and (c) which age group was hit by the risk. The risk simulated was either a dread risk that immediately killed \(50 \%\) of the population of the age group hit, or a continuous risk that killed the same total number of people in the same age group over a period of 10 years. The population growth rate was manipulated by setting the birth rate to either 0.05 (constant population), 0.075 (increasing population), or 0.025 (decreasing population). All individuals would die naturally after their 79th year. The risk hit only children, only young adults, only older adults, or only elderly adults.

In total there were 24 scenarios. Each scenario was simulated 500 times, and we calculated for every time point the average population size within the simulations. We analyzed each scenario by comparing the log difference in cumulative people-years between the dread risk condition and the continuous risk condition after 25,50, 75 and 100 years.

\section*{Results}

Figure 2 shows the results for the log difference in cumulative people-years depending on the population growth rate and the hit group after \(25,50,75\), and 100 years. A zero value indicates no difference in cumulative peopleyears between the dread risk and continuous risk; a negative value indicates a higher loss in cumulative people-years in the dread risk condition, and a positive value a higher loss in the continuous risk condition.

When children and young adults were hit by the risks, the effect was stronger and lasted for the entire 100-year-range simulated. When older and elderly adults were hit, the difference between dread and continuous risks was weaker, decreased over time, and sometimes even became positive.

Figure 2. Log difference in people-years lost because of continuous and dread risk, by age group hit by the risk, separately for A. constant, B. increasing and C. decreasing populations. The dread risk killed \(50 \%\) of a specific age group at once; the continuous risk the same total number of people over a period of ten years. A negative value of the difference indicates that the loss in people-years is larger for the dread risk; a positive value that the loss is larger for the continuous risk.

In sum, the results show that the dread risk affected the cumulative population size more strongly for most scenarios, particularly when it hit children or younger adults. The objective of this first set of simulations was to evaluate the impact of a dread and a continuous risk on
small samples that would reflect the sample size of social circles. With a second set of simulations we investigated the effects of such risks on a much larger population of the size of the U.S. population in 2010.


Figure 2: Log difference in people-years lost because of continuous and dread risk, by age group hit by the risk, separately for A. constant, B. increasing and C. decreasing populations. The dread risk killed \(50 \%\) of a specific age group at once; the continuous risk the same total number of people over a period of ten years. A negative value of the difference indicates that the loss in people-years is larger for the dread risk; a positive value that the loss is larger for the continuous risk. Results show that dread risks lead to larger losses in people-years across time compared with continuous risks, in particular when children and young adults are affected.

\section*{Simulation Set 2}

\section*{Method}

We set the population size to the actual U.S. population size in 2010 (Howeden \& Meyer, 2011) with the respective age
distributions \({ }^{1}\) and population growth rates. As in Simulation Set 1, we manipulated which age group (children, young adults, older adults, elderly adults) was hit by the risk. The risk killed either \(20 \%\) of the hit group, or the same total number of people over 10 years.

We again ran 500 simulations for each scenario, calculated the averaged population size of the dread risk and continuous risk and plotted the log integrals after 25, 50, 75, and 100 years.

\section*{Results}

Using real U.S. data, we found support for the findings of the previous simulations. The differences between the cumulative population hit by dread versus continuous risks occurred across all conditions and lasted over, at least, 100 years. Independent of which age group was affected, the dread risk led to a higher loss in people-years than the continuous risk (Figure 3). Loss was highest when children and young adults were hit by the risk.

Although our simulations are simplified and ignore death rates for different age groups, fluctuations in population growth rates, immigration and migration, gender differences, and fluctuations in disease, they illustrate the rationale of our hypothesis. Moreover, sensitivity analyses showed that the conclusions do not change when the continuous risk is distributed over a longer period or when the number of fatalities is larger or smaller.


Figure 3. Log difference in people-years lost because of continuous and dread risk, based on the US population. The dread risk killed \(20 \%\) of a specific age group at once; the continuous risk killed the same total number of people over a period of 10 years. Results show that the dread risk leads to a larger loss in people-years over time across all age groups. The loss was largest when children and young adults were affected.

\footnotetext{
\({ }^{1}\) The statistics only provided population size for age groups. For instance, 20,201,362 children \(<5\) years old lived in the United States in 2010. For simplicity, we assumed an equal distribution of the children across \(0-4\) years.
}

\section*{Discussion}

People's stronger reaction to dread risks compared with continuous risks is often perceived as a bias. This result proposes a new perspective against which the current hypotheses accounting for people's perception and reaction to dread risks might be reconsidered.

We showed through two different sets of simulations that this is in fact an ecologically rational strategy. The effect of dread risks compared with continuous risks is amplified twice: First by killing more people at a specific point in time, and second by reducing the number of children and young adults who would have potentially produced offspring. Hence, this effect is particularly strong when children and young adults are hit which is often the case for dread risks (e.g., earthquakes, terrorist attacks, pandemics). This result is also in line with findings suggesting that people are more concerned about risks killing younger, and hence more fertile, groups (Wang, 1996).

There are important practical implications of this finding. For instance, from a public policy perspective, an appropriate reaction to dread risks would be to stimulate increase in birth rates and/or immigration to counterbalance the stronger loss in population size.

In sum, people's fear and stronger risk perception of dread risk, compared to continuous risks, should not be considered an irrational bias, an emotional overreaction to a dramatic event. In fact, people's intuition seems to capture the objective severity of the two different risks.

\section*{References}

Dunbar, R.I.M., Schultz, S. (2007). Evolution in the social brain. Science, 317, 1344-1347.
ECDC (2009). 2009 influenza A(H1N1) pandemic (Version 7-17 December). Stockholm: European Centre for Disease Prevention and Control.
Gaissmaier, W., Gigerenzer, G. (2012). 9/11, Act II: A Finegrained analysis of regional variations in traffic fatalities in the aftermath of the terrorist attacks. Psychological Science, 23, 1449-1454.
Galesic, M., Garcia-Retamero, R. (2012). The risks we dread: A social circle account. PloS One, e32837.
Gigerenzer, G. (2004). Dread risk, September 11, and fatal traffic accidents. Psychological Science,15, 286-287.
Gigerenzer, G. (2006). Out of the frying pan into the fire: Behavioral reactions to terrorist attacks. Risk Analysis, 26, 347-351.
Gigerenzer, G., Mata, J., Frank, T. (2009). Public knowledge of benefits of breast and prostate cancer screening in Europe. Journal of National Cancer Institute, 101, 1216-1220.
Hertwig, R., Pachur, T., Kurzenhäuser, S. (2005). Judgments of risk frequencies: tests of possible cognitive mechanisms. Journal of Experimental Psychology: Learning, Memory, and Cognition, 31, 621-642.
Hill, K.R., Walker, R.S., Bozicevic, M., Eder, J., Headland, T., Hewlett, B.,...Wood, B. (2011). Co-residence patterns
in hunter-gatherer societies show unique human social structure. Science, 331, 1286-1289.
Howeden, L.M., Meyer, J.A. (2011). Age and sex composition: 2010. Washington, DC: US Census Bureau.
Krause, J., Ruxton, G.D. (2002). Living in Groups. New York: Oxford University Press.
Lee, R.B., DeVore, I. (Eds.) (1968). Man the Hunter. Chicago: Aldine.
Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., Combs, B. (1978). Judged frequency of lethal events. Journal of Experimental Psychology: Human Learning and Memory, 4, 551-578.
Öhman, A., Mineka, S. (2001). Fears, phobias, and preparedness: Toward an evolved module of fear and fear learning. Psychological Review, 108, 483-522.
Sivak, M., Flannagan, M.J. (2003). Flying and driving after the September 11 attacks. American Scientist, 91, 6-8.
Slovic, P. (1987). Perception of risk. Science, 236, 280-285.
Statistisches Bundesamt Deutschland. (2012). Gesundheit: Todesursachen in Deutschland (Fachserie 12, Reihe 4). Wiesbaden: Statistisches Bundesamt.
Wang, X.T. (1996). Evolutionary hypotheses of risksensitive choice: Age differences and perspective change. Ethology and Sociobiology, 17, 1-15.

\title{
Benefits of Graphical and Symbolic Representations for Learning and Transfer of Statistical Concepts
}

\author{
David W. Braithwaite (dwbraith@indiana.edu) \\ Robert L. Goldstone (rgoldsto@indiana.edu) \\ Department of Psychological and Brain Sciences, 1101 E. 10th Street \\ Bloomington, IN 47405 USA
}

\begin{abstract}
Past research suggests that spatial configurations play an important role in graph comprehension. The present study investigates consequences of this fact for the relative utility of graphs and tables for interpreting data. Participants judged presence or absence of various statistical effects in simulated datasets presented in various formats. For the statistical effects introduced earlier in the study, performance was better with graphs than with tables, while for the effect introduced last in the study, this trend reversed. Additionally, in the later sections of the study, responses with graphs, but not tables, reflected increasing influence from the presence of stimulus features which had been relevant earlier in the study, but were no longer relevant. The findings suggest that graphs, relative to tables, may better facilitate perception of complex relationships among data points, but may also bias readers more strongly to favor some perspectives over others when interpreting data.
\end{abstract}

Keywords: representations; graphs; tables; mathematics; statistics; human factors

\section*{Introduction}

Humans have devised a variety of different formats for externally representing information. Often, the same information may be represented in multiple representations that are informationally equivalent, in that each may be reconstructed perfectly on the basis of any other. Despite such equivalence, different representations may support performance of specific cognitive tasks at different levels of efficiency. Such differences have important implications for the selection and design of external representations.

The present study explores such differences with respect to graphs and tables, two of the most commonly-employed representational formats for quantitative information in a variety of fields. The relative advantages of graphs and tables have been the subject of extensive research. Tables appear to be at least as effective as graphs with respect to point reading tasks, which require one to estimate or read off individual data points (Meyer, Shamo, \& Gopher, 1999; Porat, Oron-Gilad, \& Meyer, 2009; Vessey \& Galletta, 1991). However, graphs have often shown advantages for tasks involving complex relationships between multiple data points, such as estimating or comparing differences between points (Schonlau \& Peters, 2012; Vessey \& Galletta, 1991), projecting trends (Meyer et al., 1999), and detecting changes in function parameters (Porat et al., 2009).

Models of graph comprehension (Carpenter \& Shah, 1998; Ratwani, Trafton, \& Boehm-Davis, 2008) suggest a possible explanation for the latter findings. According to
these models, spatial configurations of data points are the raw material on which graph comprehension processes operate. Importantly, some configurations may be directly perceived as basic visual features (Pomerantz \& Portillo, 2012), allowing relationships between points to be "read off" directly without first encoding each point separately (Carpenter \& Shah, 1998; Porat et al., 2009). For example, distances between points may be used to determine or estimate differences in the values they represent, without the need to encode those individual values at all (Pinker, 1990). Thus, in graphs, spatial configurations can act as cues for recognizing relationships between data points. Because such cues are unavailable or less salient in tables, this property of graphs can account for their observed advantages in conveying relationships among data points.

Many studies comparing task performance with graphs and tables have employed univariate datasets (Meyer et al., 1999; Porat et al., 2009). Consideration of bivariate data introduces another difference between graphs and tables. In graphs of bivariate data, there is a representational asymmetry between the two independent variables, in that one is often laid out along a spatial axis, typically the x -axis, while the other is typically represented by a non-spatial visual feature such as line color or thickness. For tables, on the other hand, such representational asymmetry is reduced, because the levels of both independent variables are laid out along spatial axes, albeit horizontal in one case and vertical in the other.

Can such representational asymmetries as exist in graphs of bivariate data lead to performance asymmetries in tasks involving one or the other variable? A few studies have provided evidence in the affirmative (Carpenter \& Shah, 1998; Shah \& Freedman, 2011). For example, Shah and Freedman (2011) found that when asked to interpret graphs of bivariate data, participants were more likely to describe main effects of the variable depicted in the legend than of that depicted on the \(x\)-axis, and were more likely to describe interaction effects as moderating effects of the legend variable on the effect of the x-axis variable than vice versa.

Such representational asymmetries in graphs, together with the intuition that these asymmetries are reduced in tables, suggest that performance asymmetries between tasks relating to one or the other independent variable in bivariate data should be greater for graphs than for tables. While a few studies have compared performance with graphs and tables on tasks involving bivariate data (Schonlau \& Peters, 2012; Vessey \& Galletta, 1991), the specific issue of how
display format affects performance asymmetry between tasks has not been directly investigated.

Consideration of multiple tasks introduces the possibility of transfer, in which experience with one task affects subsequent performance on other tasks. Such transfer could be positive or negative, depending on whether previouslylearned skills are correctly adapted for novel tasks, or applied without adaptation despite being inappropriate. Differences in the methods used to comprehend data in different formats, such as greater reliance on spatial configurations in graphs than in tables, could cause differences in ease of adaptation to novel tasks. Consistent with this possibility, Porat et al. (2009) found evidence of greater negative transfer between tasks for tables than for graphs of univariate data. However, it is unclear whether, and to what extent, these findings may generalize to other tasks, and in particular, to tasks involving bivariate data.

A related issue is how best to promote future positive transfer, and reduce negative transfer, when instructing learners to perform particular tasks. Educational theories (e.g. Ainsworth, 2006) suggest that incorporating multiple representations into instruction may be one path to these goals. Learners who integrate knowledge from multiple representations to form unified internal concepts are likely to show more robust and flexible learning. Analogy research suggests that comparison is a powerful tool to facilitate such integration and thus encourage positive transfer. For example, Gentner, Loewenstein, and Thompson (2003) found that management students who compared case studies illustrating a negotiation technique were more likely to apply the technique to novel cases. Considering these two lines of research together suggests that comparing graphs and tables illustrating a concept may encourage learners to learn the concept in a more abstract way, and thus to apply and adapt them more flexibly when faced with novel tasks.

The preceding discussion suggests several questions, which were investigated in the present study. First, for tasks focusing on one or the other variable in bivariate datasets, does graphical presentation lead to greater performance asymmetry than tabular presentation with respect to the depicted variables? Second, do graphs or tables show more positive (or less negative) influence of previous task practice on novel task performance? Third, does comparing graphs and tables during training promote such positive transfer (and/or reduce negative transfer)?

\section*{Method}

Participants received tutorials on different types of statistical effects in the context of \(2 \times 2\) factorial designs with one ex-perimentally-manipulated variable, or "treatment factor," and one observed variable, or "secondary factor." The first two tutorials involved, respectively, main effects of the treatment factor and interaction effects of the two factors. Each tutorial explained how to judge the presence of the given effect in graphs and tables. Each tutorial was followed by a test requiring participants to judge whether the given effect was present in a series of graphs and tables.

The first two tutorials and tests were followed by a third tutorial and test pertaining to main effects of the secondary factor. This test required participants to perform the same task as for main effects of the treatment factor, namely marginalizing over one of the two factors, and differed only in which factor was to be marginalized. Comparing performance across test sections allowed us to tell whether the size of performance asymmetries across tasks differed by representational format. Further, the first two tutorials explained explicitly how to determine whether the given effects were present. By contrast, the third tutorial, regarding main effects of the secondary factor, did not. Thus, the third test provided a measure of transfer to a novel task following practice with other related tasks. The tests following each tutorial also included stimuli in a verbal format which was not shown during training. Performance with these stimuli served as a measure of knowledge transfer to a task involving a novel representation.

\section*{Participants}

Participants were \(\mathrm{N}=127\) undergraduate students from the Indiana University Psychology Department who participated in partial fulfillment of a course requirement.

\section*{Materials}

A set of tables, graphs, and text passages representing possible outcomes of a fictional study were developed for use as test stimuli (Figure 1). The study involved a drink taste test with two binary independent variables, drink flavor and participant age group, and one continuous dependent variable, taste rating. Drink flavor is referred to as the "treatment factor," and age group as the "secondary factor."

Each stimulus represented a dataset comprising one taste


Figure 1. Test stimuli in (a) graph, (b) table, and (c) verbal format for a single dataset.
The pictured dataset shows a treatment effect and a treatment \(\times\) secondary interaction, but no secondary effect.
rating for each combination of factor levels. 2 datasets were generated for each combination of presence or absence of effects of the treatment factor, secondary factor, and their interaction, yielding 16 datasets. Each effect appeared in exactly half of the datasets, and no effect was correlated with any other. 3 stimuli were created for each dataset by presenting the data in each of 3 formats: table (Figure 1a), graph (Figure 1b), and verbal (Figure 1c), yielding 48 stimuli. The secondary factor always appeared on the horizontal axis of the graphs and tables, while the treatment factor was laid out vertically in the tables and the graph legends, but these orientations were reversed in the verbal stimuli.

Another fictional study, involving effects of cognitive enhancement drugs on test scores of males and females, was devised as a basis for examples to be shown in the tutorials. Analogous to our terminology for the test stimuli, drug is referred to as the treatment factor and sex as the secondary factor. As for the test stimuli, 3 effects of these factors were possible: treatment effect, secondary effect, and treatment \(\times\) secondary interaction. For each effect, 2 datasets were developed: a "positive" dataset, which had the effect, and a "negative" one, which did not. Using the same conventions as for the test stimuli, one graph and one table were created for each dataset, yielding 4 examples for each effect.

\section*{Procedure}

The experiment was divided into 3 sections, one for each effect. Each section consisted of a tutorial, followed by a test, for the given effect. The sections were always presented in the same order, namely (1) treatment effect, (2) interaction effect, and (3) secondary effect. The tutorials and tests were presented via a computer interface.

The tutorials for treatment and interaction effects followed the same structure. First, participants were shown a brief description of the cognitive enhancer study, together with 2 of the 4 examples for the given effect shown side-byside, and asked to judge whether or not the examples showed the given effect. Second, they were told that the presence of the effect depended on certain values, i.e. difference in drug scores when marginalizing over sex in the case of treatment effect, or difference of differences between drugs for each sex in the case of interaction effect. They were required to calculate and compare the relevant values, and were then told in which example(s) the effect was present, using the calculated values as justification \({ }^{1}\). Next, participants were asked to compare the two examples. Finally, the above procedure was repeated for the remaining 2 examples for the given effect.

The tutorial for secondary effects followed the same pattern as those for treatment and interaction effects, except that participants were not told which values they should calculate in order to judge the presence of secondary effects. Instead, after selecting which of the example(s) they thought

\footnotetext{
\({ }^{1}\) Participants were informed that normally a statistical test would be required, but for simplicity, they were to make their judgments using the standard that differences were significant if greater than or equal to 5 , and not significant otherwise.
}
showed effects of the secondary factor, they were asked to state how they thought the judgment should be made. They were given no feedback on their responses to this question.

Each participant was assigned randomly to one of three training conditions, which determined which examples were shown together in the tutorials. (1) In the Comparing Representations condition, the two positive examples, i.e. one graph and one table, were shown together first, followed by the two negative examples, again one graph and one table. (2) In the Contrasting Examples condition, the positive and negative examples in table format were shown together first, followed by the positive and negative examples in graph format. (3) In the Control condition, the positive table and negative graph examples were shown together first, followed by the negative table and positive graph examples.

The Comparing Representations condition directly implemented the idea, described in the introduction, of encouraging learners to compare different representations of the same information. The Contrasting Examples condition was intended as a pedagogically plausible alternative approach that employed the same materials, and involved the same amount of training, but did not afford the above opportunity for comparison of different representations. The Control condition was intended as a baseline with the same materials and same amount of variation across examples as the other two conditions, but with the examples paired in a way not expected to be useful for learners. \(\mathrm{N}=42\) participants were assigned to Comparing Representations, \(\mathrm{N}=41\) to Contrasting Examples, and \(\mathrm{N}=44\) to Control.

Each tutorial was followed by a test. Participants were shown a description of the taste test study and told that they would need to judge whether or not the effect about which they had just learned was present for various outcomes of the study. For each trial, one test stimulus appeared and remained onscreen until a response was received. No feedback was given. Each test stimulus was presented once per test section, in random order, for a total of 48 trials.

The experiment may be viewed online at http://perceptsconcepts.psych.indiana.edu/experiments/dwb/ MRIS_02/experiment_demo_live.html.

\section*{Results}

Mean accuracy on test trials was 66\%, and ranged from 25\% to \(100 \%\). Accuracy was significantly higher than chance, i.e. \(50 \%\), for all test sections and stimulus formats.

Accuracy scores were submitted to a \(3 \times 3 \times 3\) mixed ANOVA with training condition as a between-subjects factor, and test section and stimulus format as within-subjects factors. The main effect of training was not significant, \(F(2,124)=1.82, p=.166\), nor were any of its interactions with other factors. The main effect of section was significant, \(F(2,248)=23.67, p \approx .000\), indicating that accuracy was highest in the treatment section (74\%), lower in the interaction section (69\%), and lowest in the secondary section (63\%). There was a marginal main effect of format, \(F(2,248)=2.82\), \(p=.061\), qualified by a significant section \(\times\) format interaction, \(F(4,496)=11.54, p \approx .000\). Accuracy scores by test and
format are shown in Figure 2. In the treatment and interaction sections, accuracy was highest for graphs, lower for tables, and lowest for verbal, while in the secondary section, accuracy showed the opposite trend.


Figure 2: Accuracy by Test Section and Format. Error bars indicate standard errors.

Several of our research questions relate to graphs and tables only. Thus, the above analysis was repeated with the data from verbal stimuli excluded. The interaction of format with section was still significant, \(F(2,248)=3.61, p=.029\). While accuracy decreased across the three sections for both graphs and tables, it decreased more for graphs (treatment: \(77 \%\), interaction: \(71 \%\), secondary: 61\%) than for tables (treatment: 75\%, interaction: 69\%, secondary: 63\%).

Response times for test trials were analyzed using the same ANOVA model structure. The results strongly resembled those for accuracy. No significant effects involving training were found, \(p \mathrm{~s}>.25\). The main effects of test section and format were both significant, \(F(2,248)=63.78\), \(p \approx .000\) for section and \(F(2,248)=40.69, p \approx .000\) for format, as was their interaction, \(F(4,496)=3.04, p=.017\). Response times by section and format are shown in Figure 3. Responses sped up over the course of the three test sections. Responses were, overall, faster for graphs than for tables and verbal, but these differences were more pronounced in the treatment section than in the later sections.


Figure 3: Response Time by Test Section and Format. Error bars indicate standard errors.

Just as for accuracy, the analysis of response time was repeated using for graph and table trials only. The main effect of format was significant, \(F(1,124)=53.41, p \approx .000\), but the interaction of format with section was not, \(F(2,248)=.913\), \(p=.403\). Thus, response times were faster for graphs (6209 ms ) than for tables ( 7230 ms ) across all three sections.

Accuracy scores reflect the differing utilities of graphs and tables for task performance in different test sections, but give little insight regarding the mental processes underlying task performance. One way in which the latter might differ is the degree of influence exerted by different stimulus features. Each test stimulus was determined by presence or absence of treatment, interaction, and secondary effects, which may be viewed as three binary features. In each test section, only one feature was relevant, but the two irrelevant features may also have influenced responses. For example, in the secondary effect section, only the presence/absence of secondary effects was relevant, but a participant who had not adequately differentiated the three effects might give a positive response to a stimulus exhibiting treatment and interaction effects, even if no secondary effect was present. Thus, it could be useful to understand the influences of relevant and irrelevant features on responses for different stimulus formats and test sections.

To this end, a measure of the degree \(I_{x, s}\) to which the presence of effect \(x\) influenced responses in the test section regarding effect \(s\) was calculated as follows:
\[
I_{x, s}=P\left(R=+\mid E_{x}=+, S=s\right)-P\left(R=+\mid E_{x}=-, S=s\right)
\]
\(R=+\) signifies a positive response, \(E_{x}=+\) and \(E_{x}=-\) signify, respectively, the presence and absence of effect \(x\), and \(S=s\) signifies that the test section concerns effect \(s\). Thus, \(I_{x, s}\) represents the difference in probability of a positive response regarding effect \(s\) when effect \(x\) is present, relative to when effect \(x\) is absent. For a perfect responder, we would have \(I_{x, s}=100 \%\) when \(x\) is relevant, i.e. \(x=s\), and \(I_{x, s}=0 \%\) when \(x\) is irrelevant, i.e. \(x \neq s\). In other words, perfect responses would reflect total influence of relevant features and zero influence of irrelevant features.

Influence \(I_{x, s}\) was calculated separately for each participant, stimulus format, effect \(x\), and test section \(s\). The pattern of results for relevant features closely resembled those for accuracy, and thus are not reported here. The results for irrelevant features are shown in Figure 4. The mean of \(I_{x, s}\) in these cases was \(18 \%\), and was significantly greater than \(0 \%\) for all combinations of format and test section. Thus, participants were significantly biased towards positive responses by the presence of irrelevant features.

The data for influence \(I_{x, s}\) over all cases where \(x \neq s\) were analyzed using the same ANOVA model structure as for accuracy and response time. No significant effects involving training condition were found, \(p \mathrm{~s}>.12\), nor was the main effect of test section significant, \(F(2,248)=0.86, p=.423\). However, a significant main effect of format was found, \(F(2,248)=5.68, p=.004\), indicating that irrelevant features had the most influence for graphs (20.5\%), less for tables
(18.7\%), and least for verbal format (17.4\%). This effect was qualified by a format \(\times\) test section interaction, \(F(4,496)=8.61, p \approx .000\). As shown in Figure 4, the influence of irrelevant features increased over test sections for graphs, stayed about the same for tables, and decreased for verbal stimuli. Separate ANOVAs conducted using the data for each format alone found a significant effect of test section on influence \(I_{x, s}\) for graphs, \(F(2,248)=8.41, p \approx .000\), but not for tables or verbal stimuli, \(p s>.38\).


Figure 4. Influence \(I_{x, s}\) for \(x \neq s\), i.e. for irrelevant features. Error bars indicate standard errors.

\section*{Discussion}

In the first two sections of the study, participants were trained to judge whether treatment and interaction effects were present in bivariate data, and then tested on their ability to do so when the data was presented in graphical, tabular, or verbal format. In both sections, responses were faster and more accurate for graphs than for tables. Judging the presence of either effect requires assessing complex relationships between data points, i.e. comparing averages of pairs of data points for treatment effects or differences between pairs of data points for interaction effects. The advantage shown by graphs over tables is thus consistent with the general view that complex relationships among data points are more easily assessed in graphical than in tabular format (Meyer et al., 1999; Porat et al., 2009; Schonlau \& Peters, 2012; Vessey \& Galletta, 1991).

Accuracy was lower in the secondary effect section than in the previous two sections. This result is not surprising, considering that participants were not told how to judge the presence of secondary effects. However, interestingly, the effect of format on accuracy was reversed in this section. What might have caused this reversal? One possible explanation, detailed below, involves transfer. Specifically, low accuracy with graphs in the secondary effect section may have reflected negative transfer from the previous sections that was absent, or reduced, in the case of tables.

To flesh out this possibility, we consider how experience of the earlier sections of the study might have affected performance in later sections. In the earlier sections, participants were trained in explicit calculation methods to judge the presence of treatment and interaction effects. With
graphs, however, their judgments may have relied in part on visual patterns. For example, a sideways " \(v\) " shape in the graphs (Figure 1a) could be a useful cue for the presence of both treatment and interaction effects. Reliance on such visual patterns may have led to the creation of automatic visual routines (Ullman, 1984) that could support quick judgments regarding presence or absence of effects without, or before, performing the relevant calculations. Importantly, such routines, once acquired in the earlier sections of the study, might continue to be used in the later sections.

Thus, visual routines associating responses with visual patterns are one mechanism by which experience of the earlier sections might influence performance in the later sections. Importantly, this account predicts that such influence would be greater for graphs than for tables. Visual patterns are believed to play an important role in graph comprehension (Carpenter \& Shah, 1998; Pinker, 1990), but are much less salient in the case of tables. Moreover, the above mechanism could lead to negative transfer. Because visual patterns that were relevant earlier become irrelevant, even misleading, later, continuing to rely on them could hurt performance. For example, having learned in the first two sections to give positive responses when seeing the sideways "v" shape (Figure 1a), participants might continue to do so in the secondary effect section, even though that shape actually indicates the absence of a secondary effect. In sum, the above account predicts greater negative transfer for graphs than for tables in the later sections of the study.

Support for this explanation comes from our analysis of influence of irrelevant features on responses. In general, such influence was greater for graphs than for tables. More important for our present purpose, such influence increased over the course of the study for graphs, exactly as would be expected if responses in later sections were influenced by visual patterns which had proven useful in earlier sections. By contrast, influence of irrelevant features did not change over the course of the study for tables, as one would expect given the lesser salience of visual patterns in tables.

An alternate explanation for the reversal, in the secondary effect section, of relative accuracies for graphs and tables involves variation in the intrinsic difficulty of recognizing different effects in different formats. Specifically, for graphs, treatment and interaction effects may have been relatively easy to detect, and secondary effects relatively difficult, while for tables, there may have been less variation in the ease of detecting the various effects. This possibility is consistent with the hypothesis, stated in the Introduction, that performance asymmetry between tasks should be greater for graphs than for tables, due to greater representational asymmetry between variables in the former case. It is also consistent with Shah and Freedman's (2011) abovementioned finding that spontaneous interpretations of bivariate graphs tend to focus on main effects of the legend variable (in our study, the treatment factor) rather than the x axis variable (in our study, the secondary factor).

However, two aspects of our results cannot easily be explained in terms of variation in intrinsic task difficulty. The
first is the observed pattern of response times. Although accuracy in the secondary effect section was lower for graphs than for tables, reaction times showed the opposite trend, i.e. faster responses for graphs. These faster responses are consistent with reliance on automatic visual routines, as described above, but less consistent with the assumption that the task was more difficult to perform with graphs. Second, variation in intrinsic task difficulty cannot explain why influence of irrelevant features increased over the study for graphs, but not for tables. However, this effect is predicted by the first account given above.

While available evidence favors the first over the second account, further research could more definitively disambiguate between them by placing the secondary effect section at the beginning, and the treatment effect section at the end. If the first account, in terms of learned visual routines, is correct, then whichever section comes last should show negative transfer for graphs. If the second account, in terms of intrinsic task difficulty, is correct, then performance on the secondary effect section should be worse for graphs regardless of when it is encountered.

Another question investigated in our study was whether comparing graphs and tables of the same data during training, as in the Comparing Representations condition, would facilitate learning and transfer. However, this prediction was not confirmed. Accuracy showed no effect of training condition, suggesting that the Comparing Representations condition was not more effective overall. Nor did accuracy show any interaction of training condition with either format or section, suggesting that the Comparing Representations condition did not produce any particular benefits for transfer, either to a novel format, i.e. verbal, or to a novel effect type, i.e. secondary effect.

Importantly, this negative finding does not address the issue of whether the use of multiple representations during instruction can benefit learners, because multiple representations were included in all of our training conditions. However, our findings do suggest that the specific technique of comparing different representations of the same data may not produce any incremental learning benefit. This finding stands in contrast to the considerable learning benefits that can result from comparing semantically different instantiations of the same concept (Gentner et al., 2003).

In conclusion, our findings are consistent with previous research in finding an advantage for graphs over tables for tasks involving complex relationships between data points. Theories of graph comprehension suggest that salient visual patterns in graphs may underlie this advantage. However, a novel finding of the present study is that such visual patterns may not always be helpful. In particular, when performing novel tasks, graph readers may focus on visual features which were relevant to previous tasks, and have difficulty shifting perspective to focus on features which were previously irrelevant. By contrast, such shifts of perspective may be relatively easier with representational formats in which visual patterns are less salient, such as tables. These considerations suggest that graphical presentation may be pref-
erable for performing well-practiced tasks which are known in advance, while tabular presentation may be most suitable when performing or learning to perform unfamiliar tasks.

\section*{Acknowledgements}

This research was in part supported by National Science Foundation REESE grant \#0910218 and US Department of Education grant \# R305A1100060.

\section*{References}

Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. Learning and Instruction, 16(3), 183-198.
Carpenter, P. A., \& Shah, P. (1998). A model of the perceptual and conceptual processes in graph comprehension. Journal of Experimental Psychology: Applied, 4(2), 75-100.
Gentner, D., Loewenstein, J., \& Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. Journal of Educational Psychology, 95(2), 393-405.
Meyer, J., Shamo, M. K., \& Gopher, D. (1999). Information Structure and the Relative Efficacy of Tables and Graphs. Human Factors: The Journal of the Human Factors and Ergonomics Society, 41(4), 570-587.
Pinker, S. (1990). A theory of graph comprehension. In R. Freedle (Ed.), Artificial Intelligence and the Future of Testing (pp. 73-126). Hillsdale, N.J.: Erlbaum.
Pomerantz, J., \& Portillo, M. (2012). Emergent Features, Gestalts, and Feature Integration Theory. In J. M. Wolfe \& L. Robertson (Eds.), From Perception to Consciousness: Searching with Anne Treisman (pp. 187-192). Oxford University Press.
Porat, T., Oron-Gilad, T., \& Meyer, J. (2009). Taskdependent processing of tables and graphs. Behaviour \& Information Technology, 28(3), 293-307.
Ratwani, R. M., Trafton, J. G., \& Boehm-Davis, D. A. (2008). Thinking graphically: Connecting vision and cognition during graph comprehension. Journal of Experimental Psychology: Applied, 14(1), 36-49.
Schonlau, M., \& Peters, E. (2012). Comprehension of Graphs and Tables Depend on the Task: Empirical Evidence from Two Web-Based Studies. Statistics, Politics, and Policy, 3(2).
Shah, P., \& Freedman, E. G. (2011). Bar and Line Graph Comprehension: An Interaction of Top-Down and Bottom-Up Processes. Topics in Cognitive Science, 3(3), 560-578.
Ullman, S. (1984). Visual Routines. Cognition, 18(1-3), 97159.

Vessey, I., \& Galletta, D. (1991). Cognitive Fit: An Empirical Study of Information Acquisition. Information Systems Research, 2(1), 63-84.

\title{
Entrainment on the Move and in the Lab: The Walking Around Corpus
}

\author{
Susan E. Brennan (susan.brennan@stonybrook.edu) \\ Department of Psychology, Stony Brook University, Stony Brook, NY 11794-2500 USA
}

\author{
Katharina S. Schuhmann (katharina.schuhmann@stonybrook.edu) \\ Department of Linguistics, Stony Brook University, Stony Brook, NY 11794-4376 USA
}

\author{
Karla M. Batres (karla.batres@stonybrook.edu) \\ Department of Psychology, Stony Brook University, Stony Brook, NY 11794-2500 USA
}

\begin{abstract}
We examined lexical choice and variability in referring expressions during direction-giving to pedestrians. The Walking Around Corpus comprises an experimentally parameterized collection of spontaneous spoken dialogues produced by 36 pairs of people communicating by mobile telephone; it provides both a testbed for lexical entrainment "in the wild" as well as a resource for pedestrian navigation applications. A stationary partner (the Giver) directed a mobile partner (the Follower) to walk about 1.8 miles, to 18 destinations on a medium-sized campus. Givers viewed a map marked with target destinations, labels, and photos. Followers carried a cell phone with GPS and a digital camera in order to photograph the destinations they visited; Givers monitored Followers' progress as a cursor on a map display. Immediately after the navigation task, Followers returned to the lab and both were tested individually on their spatial ability and memory for the destinations. Next, the Experimenter attempted to interfere with any conceptual precedents they had established by giving Followers printed copies of the photos they had just taken and prompting them (sometimes with competing labels) to identify each destination. Finally, each pair participated in 6 rounds of a more traditional referential communication lab task to repeatedly match duplicate copies of the Follower's pictures of the destinations. Results include significant rates of lexical entrainment, evidence for partner-specific conceptual pacts, and that joint navigation efficiency is affected by directiongivers' spatial ability. The Walking Around Corpus is available to the research community.
\end{abstract}

Keywords: Referential communication; conceptual pacts; entrainment; collaboration; mobile communication; GPS applications; pedestrian navigation; spatial cognition.

\section*{Introduction}

Speakers make many expressive choices in conversation, leading to enormous variability in spontaneous speech. These choices emerge not only from their individual proclivities and the availability of words in memory, but also from collaboration with conversational partners (Clark \& Wilkes-Gibbs, 1986). There is much less variation within than between conversations; speakers are more likely to continue using the same term with the same addressee than with a new addressee (Brennan \& Clark, 1996). Studies of referential communication in laboratory settings show that
lexical choice is influenced not only by precedent (what a speaker has said previously) and strength of precedent (e.g., how often a referent has been discussed), but also by partner (whether the precedent was established with the current partner or a different one) but that it can be maintained via interactive cues provided during grounding (ibid).

We have proposed that lexical entrainment marks two speakers' beliefs that they are referring to the same thing; in fact, breaking a conceptual pact, such as hearing a speaker inexplicably abandon an entrained-upon expression for a new expression when referring to the same referent, results in slower comprehension than hearing the same new expression from a different (new) partner (Metzing \& Brennan, 2003; Matthews, Lieven, \& Tomasello, 2010). A conceptual pact does not reflect a rigid mapping between expression and referent, but emerges instead from a flexible, temporary agreement arrived at with an interactive partner to take a particular perspective on the referent; as such, it is highly dependent on context. Although there is evidence that the partner associated with a conceptual pact is represented by a cue in memory (Horton, 2007; Horton \& Gerrig, 2005), this does not pre-empt an expression and referent from being easily associated with a different partner (Barr \& Keysar, 2002; Brennan \& Clark, 1996), nor does having an existing conceptual pact with a partner inhibit a new expression-referent association from being encoded with that partner when the pragmatic context changes (ibid).

Lexical entrainment has been demonstrated in adults (Brennan \& Clark, 1996; Clark \& Wilkes-Gibbs, 1986; Garrod \& Anderson, 1986) and in children (Matthews et al., 2010); in experts and in novices (Isaacs \& Clark, 1987); in native and native/non-native conversations (Bortfeld \& Brennan, 1997); and for lexicalized and non-lexicalized referents (Gergle, Kraut, \& Fussell, 2013; Schober \& Carstensen, 2010). It has also been measured in statistical analyses of the transcripts of speech corpora (e.g., Nenkova, Gravano, \& Hirschberg, 2008's studies of the Switchboard and Columbia Games corpora, and Stoyanchev \& Stent, 2009's from Let's Go, a real-world bus information system). In laboratory studies of referential communication and lexical choice, we and others have demonstrated effects of psycholinguistic processing in interactive dialogue contexts,
attempting to gain sufficient control to test hypotheses while modeling spontaneous interaction in conversational contexts. However, the question sometimes arises as to whether dialogue in non-laboratory contexts may be too noisy to show the same sorts of effects (Stoyanchev \& Stent, 2009).

In this project, we examine lexical entrainment and conceptual pacts both outside of and inside the laboratory.

\section*{The Walking Around Project}

\section*{Design and Materials}

Visual Context Manipulation Visual evidence relevant to a collaborative task has been shown to affect task strategies, to incrementally shape the form and content of spontaneous utterances, and to enable partners to shift the initiative or distribute the responsibility for grounding utterances (see Brennan, 2004 for discussion).

The amount of visual context provided to the stationary direction-giving partner during the navigational task was varied as a between-pairs and within-destinations factor. For half of the target destinations, the direction-giver saw only a close-up photo of the target (Figure 1) along with its numbered location on a map; for the other half, they saw the close-ups plus extra visual context (Figure 2). Each pair was assigned to either List 1 or List 2, with Givers seeing minimally illustrated targets in List 1 or extra visual context in List 2, and vice versa.
Because the mobile partner in the navigation task had rich visual information from being present within the local context, while the stationary partner had global information about the target location but less information about its visual appearance and relative local landmarks, we predicted that additional visual context for the stationary partner would help them align their knowledge, making presentation and acceptance phases of the directions more efficient jointly. On the other hand, such knowledge might be unnecessary for direction-givers who knew campus well (assessed during the final questionnaire).

\section*{Methods}

Task Sequence Pairs of volunteers did a sequence of five tasks during a 2 -hour session: (1) a collaborative navigation task in which one partner followed the other's directions in order to visit 18 pre-chosen destinations on campus, (2) a timed mental rotation test of spatial ability, (3) individual recall tests of their memory for the locations they had discussed, (4) showing the Experimenter the photos taken of the destinations (done by the mobile partner only) during which the Experimenter provided competing labels for some of the destinations, (5) 6 rounds of a referential communication task in which pairs matched identical copies of this set of photos (switching director-matcher roles halfway through, with the mobile partner acting as director for the first three rounds and the stationary partner as director for the last three), and (6) a questionnaire, completed individually. This sequence of tasks enabled us
to examine lexical choices within and across different interactive and solitary contexts; the spatial ability test provided the opportunity to examine the impact of this individual difference on strategies and performance.

Status of Corpus Project The corpus being released to the research community includes 36 digitally recorded spoken dialogues (with associated data on individual differences and experimental parameters). The navigational dialogues from these pairs have all been transcribed in detail (including disfluencies and pauses, along with time stamps). A full article in preparation will present data on the content and sequencing of referring expressions and navigation strategies. Here, we report findings from the post-navigation recall test, entrainment data from the 6 -round referential communication task, and effects of spatial ability.


Figure 1: Close-up views of 10 targets seen by Givers ( \(L\) to \(R\), top to bottom): Patriotic faces, Mushroom house, Outside chalkboard, Ship sculpture, Sorority rock with girl, Spaceship label, Goldfarb plaque, Cylindrical cement structure, Cedar plaque, Warning sticker

\section*{Subjects}

54 pairs of Stony Brook students were recruited from the Psychology Subject Pool or via flyers on campus to participate with another person in a task involving a nature hike or walking around campus. All identified as native speakers of English, all but two pairs were strangers, and all gave informed consent. As they arrived, they were asked whether they preferred to walk around or remain indoors; roles were assigned according to their preferences.
Because the sessions involved so many tasks (with the inevitable risks of equipment failure, bad weather, and subjects who either failed to complete the navigation task or who disregarded instructions and ran errands during the
session), and since inclusion of the navigation dialogues in the final corpus required both members of a pair to sign an additional release form, 18 of the 54 sessions could not be included in the Walking Around corpus. However, a total of 49 pairs completed most or all of the matching rounds, so contributed usable data for the current analyses. Subjects received their choice of either 2 research credits that could be used to fulfill a course requirement, or \(\$ 9\) per hour.

The student assigned to the role of direction-Giver (G) remained inside the lab with a landline phone to direct the student assigned as Follower (F) to locations on campus.


Figure 2: Extra visual context provided to Giver (without highlight) for selected targets from Fig. 1: Spaceship label, Warning sticker, Sorority rock with girl, Patriotic faces.

\section*{Procedure}

The sequence of tasks unfolded as follows:
(1) The goal of the collaborative navigation task was for \(F\) to visit and identify each target destination in order, and to take a photograph documenting the destination. F was provided with a digital camera and a mobile phone with GPS, but no information about the destinations, and so had to rely entirely on G's directions and descriptions. G used an interactive web interface displaying a custom Google Map of the relevant portion of Stony Brook's campus, with 18 target destinations marked and numbered in the order in which they were to be visited by F. By clicking on each numbered destination on the map, \(G\) could view the associated pop-up photograph(s) and label (Table 1 lists the labels provided to G). In addition, the same photographs and labels were provided as hard copy in a binder, with the materials for one destination visible per page. The map interface also displayed a cursor generated by F's mobile phone's GPS tracking, providing G with visual evidence of F's location on campus, updated every few seconds.
(2) Immediately following the navigation task, \(F\) returned to the lab, where both F and G were tested individually for their spatial ability via a timed paper and pencil mental rotation test (Card Rotation Test-S-1 [Rev.] by Educational

Testing Services, 1962, 1975).
(3) Then each partner was tested individually for their memory for the destinations. A SuperLab program on a laptop randomly selected and displayed 36 pictures of campus destinations, one at a time; half of these were the close-ups of the eighteen target destinations from the navigational task (as in Figure 1), and the other half depicted destinations the partners had not seen. Each partner was instructed to indicate within five seconds whether the destination was old (discussed or visited in the navigation task) or new. If a partner recognized the destination as old, the program allowed 30 seconds for typing its name into an expanding textbox. This task aimed to probe for the most available conceptualization (presumably, related to the one they had grounded with the partner during the navigational study completed only a few minutes earlier).
(4) Next, the Experimenter attempted to interfere with any conceptual precedents or pacts the pair may have established by giving F printouts of the photos F had just taken and having F identify them one by one; for some, the Experimenter deliberately used different labels than had been originally provided to \(G\) (see Table 1).

Table 1: Influences upon G and F, prior to matching task
\begin{tabular}{|l|l|}
\hline \begin{tabular}{l} 
Label provided to Giver \\
(screen and binder):
\end{tabular} & \begin{tabular}{l} 
Expression addressed to \\
Follower (by Exptr)
\end{tabular} \\
\hline Cedar plaque & Lebanon plaque \\
\hline Goldfarb plaque & Stickball plaque \\
\hline Patriotic faces & Profiles plaque \\
\hline Warning sticker & Watertower sticker \\
\hline Mushroom house & Brown brick structure \\
\hline Outside chalkboard & Physics chalkboards \\
\hline Ship sculpture & Engineering sculpture \\
\hline Spaceship label & SUNYLab label \\
\hline Cylindrical cement structure & T-shaped cement block \\
\hline Sorority rock with girl & Sorority rock with the moon \\
\hline
\end{tabular}
(5) After this, the partners re-joined one another to participate in 6 rounds of a referential communication task in which they repeatedly matched identical sets of pictures of the 18 destinations. The pictures used in this phase consisted of the actual photos taken by F while walking around and were printed in grey-scale, on \(8.5 " \mathrm{X} 11\) " paper. G and F sat at separate tables, back-to-back; F played the role of director for the first 3 rounds with G as matcher, and they switched roles for the last 3 rounds. To begin each round, each partner's set of pictures was shuffled randomly, with the director's arranged as a stack from which to view and describe one picture at a time, and the matcher's spread out on the table so that the matcher could identify and stack each one in the same order as the director's. Partners were told to accurately order their pictures as quickly as possible; the experimenter intervened if they made a matching error or tried to peek at each other's pictures.
(6) Partners filled out a questionnaire asking about their familiarity with campus, whether they knew each other, their length of time at SBU, whether they drive or live on
campus, their confidence in their own sense of direction and ability to read maps, and any other information they wanted to volunteer as relevant. Then they were debriefed.

\section*{Coding and Examples}

Entrainment coding We coded the content of referring expressions that both partners used in the 6 rounds of the matching task, including whether F's and G's expressions had converged by Round 6 ( F was director and G was matcher for Rounds 1-3, with G as director and F as matcher for rounds 4-6). We coded only the 10 target destinations for which the Experimenter provided an expression to F before the matching task, where the expression was distinct enough to compete with G's given label (illustrated in Figure 1 and listed in Table 1).

Lexical convergence (as evidence for entrainment) was coded for each of the 10 pictures using moderately strict criteria (adapted from Brennan \& Clark, 1996 and Bortfeld \& Brennan, 1997), as follows: if all of the content morphemes (nouns, adjectives) in the expressions spoken by the director (G) in Round 6 were included in any expressions from Rounds 1-3 (where F was director), or were also spoken or confirmed (e.g., by a backchannel or verbal acknowledgment) by F (the matcher) in Rounds 4-5, that counted as convergence.

Examples By these criteria, the expression "mushroom house" (used by a particular G as director in Rounds 4-6) was coded as not convergent with "mushroom hut" (which the F partner had used throughout Rounds 1-3) because F never used "house" and G never used "hut").

It has been shown many times that entrained-upon expressions become more efficient with repeated referring, so by this criteria, "mushroom" produced during Round 6 was coded as a convergent, shortened form of "brick mushroom" from Rounds 1-3.

Our coding ignored the order of propositional content, as well as the appearance of proxy terms (e.g., "thingy" in "mushroom thingy" or "blah blah" in "environmental blah blah" were considered to be like a wild card or variable, replaceable with other terms). Note that we did not include prepositions or other function words in the coding, so "moon rock girl" was coded as convergent with "girl on the rock with the moon."

However, when \(G\) introduced a new propositional morpheme in Round 6 that F had not used or acknowledged in Rounds 1-5, we coded this as a distinct expression (and a failure to entrain), such as in the following example, for which the given label was Ship sculpture, and the Experimenter's term was Engineering sculpture. (Note: in this and other examples, matchers did not always verbally respond, but would often just identify and arrange a picture.)

\section*{Round 1 F : boat structure}

G: boat structure?
F: yeah, the- in front of the Engineering building

Round 2 F: structure in front of Engineering building, the boat
Round 3 F: structure in front of Engineering building
Round 4 G : wire ship, *the*
F: *what?*
G : in front of the-
F : Engineering building?
G: yeah
F: okay
Round 5 G: the ship thing, in front of Engineering
Round 6 G: wire ship
(Pair 13, \#13)
In the example above, F's "boat structure" was never taken up by G, and G's "wire" was never confirmed by F.

Note that when pairs failed to entrain on a referring expression by Round 6, this did not mean that the expressions they used were unrelated. In some cases, they had all but converged in the 5 th round, but introduced an element new to their conversation in the 6th round (which could not count as entrainment by our strict criteria). This is the case with the next example (note also that in the navigation task, \(G\) had told \(F\) that he hated this particular object because he had run into it on his bicycle as a child):

Round 1: F: the cylinder you ran into
Round 2: F: the cylinder you ran into
Round 3: F: cylinder you ran into
Round 4: G: cylinder
F: cylinder
Round 5: G: cylinder
Round 6: G: cement
(Pair 54, \#15)
Note that for this destination, both the label given in the binder and the Experimenter's term had included "cement" (see Table 1). Like other behaviorally expressed products of human memory, conceptual pacts are subject to interference.

In other cases where pairs did not entrain, they sometimes appeared to be in the process of converging on a conceptual pact that was still a bit unstable, having included all of the same elements but in partial combinations during the 6 rounds, with F never producing or accepting all of the propositions in G's final expression.

However, at other times, F and G did express distinct, competing conceptualizations (possibly expressing disagreement), such as when F repeatedly used "submarine" in Rounds 1-3 and G subsequently used "spaceship" on Rounds 4-6, with neither comment nor discussion. Such dueling perspectives rarely led to errors in the matching task, but they do illustrate that pairs did not always succeed in entraining on a stable perspective.

Some gave explicit evidence that they were monitoring the sources of the content they expressed, both from walking around and from hearing competing terms from the Experimenter. In this example (coded as entrainment), F refers explicitly to both:

Round 1: F: the thing I couldn't find, that really tall, I think
she \(<\) the Experimenter \(>\) described it as a watertower, the really tall thing that was by G: - the helium (tank)?
F: yeah, yeah, yeah
Round 2: F: the helium tank, I think you said
Round 3: F: helium, tank, or
Round 4: G: the helium tank
Round 5: G: helium tank
F: () tank
Round 6: G: helium tank
(Pair 30, \#6)
Other Influences on Perspective Taking We also coded the extent to which the expressions in Round 6 included content that was (a) originally provided in the label provided to G in the binder for the Walking Around task, (b) spoken to F by the Experimenter after the recall test and before the matching task, or (c) "other", or not provided in (a) or (b). We predict that "other" content of the referring expressions will reflect conceptualizations and expressions that the pair discussed during the navigation task, as here:
Round 1: F: the T-shaped structure that you thought looked like a thumbtack
Round 2: F: T-shaped structure, thumbtack
Round 3: F: the thumbtack scri-, structure
Round 4: F: thumbtack
Round 5: F: thumbtack
Round 6: F: thumbtack
(Pair 49, \#15)
Because we have not yet coded all of the destinationrelevant expressions produced in over 30 hours of recorded navigation sessions, we cannot yet quantify the influence of that task on the expressions pairs used later on. So for the purpose of the current analyses, we use these proxies:
- Often, G began describing a destination in the Walking Around task by first proposing the label given for that target in the destination binder. So for the current analyses, we used the initial labels from G's binder as an estimate of the conceptualization first proposed by G and often elaborated, modified, or renegotiated and replaced as F walked around.
- Likewise, we considered content from alternative expressions directed at F by the Experimenter as potentially interfering with conceptual pacts established while walking around (especially due to the recency of that label, heard immediately before the matching task).
- Finally, content coded as "other" was hypothesized as having as its potential source the perspectives negotiated by the partners while walking around.

\section*{Results and Discussion}

Effects of Spatial Ability The time it took a given pair to complete the Walking Around task was correlated with G's mental rotation ability, \(\mathrm{r}=-.42, \mathrm{p}=.004\). Walking Around time was not reliably correlated with F's mental rotation ability, \(r=-.24, p=.111\).

Entrainment while Matching By Round 6, there was evidence of lexical entrainment \(74.8 \%\) of the time, with individual pairs entraining on perspectives for \(33 \%\) to \(100 \%\) of destinations, and particular target destinations showing entrainment by \(54 \%-95 \%\) of the pairs.

This represents a rather strong degree of entrainment considering that conceptual pacts are theorized to be flexible across contexts, and the multiple contexts in our study differed substantially from one another (e.g., during the matching task, G viewed F's photos for the first time, and now needed to compare and distinguish the destinations alongside one another. Brennan \& Clark's (1996) lab experiments showed \(56 \%-71 \%\) entrainment (this did not include expressions where one was a shortened version of the other, but the task context was far more uniform).

Hybrids When several perspectives are in the air, one highly effective strategy is to hybridize them. In the next example, G first read warning sticker in the binder. While walking around, F proposed a big water tank or something, while G proposed diamond shaped sticker. Later, F heard watertower from the Experimenter. In the matching rounds, each accommodated the other:

Round 1: F: the watertower with the triangle, diamond I mean
Round 2: F: watertower, with the diamond, the colors
Round 3: F: the watertower with the diamond
Round 4: G: the, the watertower sign, triang-, diamond, with the colors
Round 5: G: the watertower sign, warning sign
Round 6: G: the diamond watertower sign (Pair 49, \#6)
Memory for Destinations \& Labels G was more accurate than F in correctly recalling the destinations discussed in the navigation task (responding "yes" to old items), \(90.6 \%\) to \(80.4 \%, t(47)=3.69, \mathrm{p}=.001\). G was also marginally faster to respond than \(\mathrm{F}, t(47)=1.72, \mathrm{p}<.10\). This makes sense, as the contexts in which G encoded the destinations were consistent with the context of recall (memory prompts were the same closeup photos from the binder, and G had ample time to look at them as F walked around).

Convergence of Recalled Terms with Given Label After responding "yes" in the recall test, the G partner produced an expression with the identical content as the F partner \(41.7 \%\) of the time. The fact that this was so much lower than their entrainment rate in the matching rounds highlights the difference between individual memories and collaboratively-achieved perspectives.

The expression \(G\) recalled had all the same propositional content as the given (binder) label \(37.5 \%\) of the time, and a somewhat shorter label \(20.4 \%\) of the time. So \(57.9 \%\) of the time, the label G recalled matched the one G started out with (we do not call this convergence "entrainment" since it does not emerge from interpersonal interaction). G’s recalled label added propositional content to the given label
only \(2.2 \%\) of the time (even though they often created hybrid labels later when interacting with F in the matching rounds). This is striking, as it suggests that when they encoded the given label as precedent, they did so very strongly. The rest of G's recalled expressions (30.9\%) included no content from the given label.


Figure 3: Proportions of matching rounds that included content from labels initially given to G (in G's binder) or spoken to F (by the Experimenter). The discontinuities between Rounds 3 and 4 reflect a change in speaker roles ( \(F\) directed Rounds 1-3, and G directed Rounds 4-6).

Influence of Experimenter's Label Finally, we consider whether \(F\) designed referring expressions with \(G\) in mind by quantifying the interference from expressions addressed by the Experimenter to F just before the matching rounds. Figure 3 shows that when \(F\) was director (R1-R3), some content from those expressions made it into the expressions F proposed to \(G\) (note: this does not yet take into account a baseline for how often such terms were used while walking around). If F chose referring expressions egocentrically or based on recency alone, rather than designing them for G, then the black bars in Figure 3 would be much higher.

\section*{Conclusions}

The Walking Around Corpus and associated data from this project demonstrate how spontaneous referring expressions in conversation are shaped by partner-specific influences and interaction, as well as by other factors that include prior conceptualizations, spatial ability, and recent interference from another speaker. We are analyzing the referring expressions from the corpus to further explore evidence for entrainment in real world contexts, as well as to address applied questions. The corpus may be applicable to spoken dialogue applications; although GPS-based spoken dialogue systems for cars are now standard, such systems have not yet been optimized for pedestrians. This corpus is available to others for research purposes (contact the first author).

\section*{Acknowledgments}

This research was supported by the National Science Foundation under Grant \#IIS-1043665. We thank M. Walker and Z. Hu for helpful discussions, R. Bowmani for technical assistance, and S. Eng, L. Snelling, R. A. Wang, and M. Zhang for assistance with coding and transcription.

\section*{References}

Barr, D. J., \& Keysar, B. (2002). Anchoring comprehension in linguistic precedents. Journal of Memory \& Language, 46, 391-418.
Bortfeld, H. \& Brennan, S. E. (1997). Use and acquisition of idiomatic expressions in referring by native and nonnative speakers. Discourse Processes, 23, 119-147.
Brennan, S. E. (2004). How conversation is shaped by visual and spoken evidence. In J. Trueswell \& M. Tanenhaus (Eds.), Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action tradition. Cambridge MA: MIT Press.
Brennan, S. E., \& Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. Journal of Experimental Psychology: Learning, Memory and Cognition, 6, 14821493.

Clark, H. H., \& Wilkes-Gibbs, D. (1986). Referring as a collaborative process. Cognition, 22, 1-39.
Garrod, S., \& Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic coordination. Cognition, 27, 181-218.
Gergle, D., Kraut, R.E., \& Fussell, S.R. (2013). Using visual information for grounding and awareness in collaborative tasks. Human-Computer Interaction, 28, 1-39.
Horton, W. S. (2007). The influence of partner-specific memory associations on language production: Evidence from picture naming. Language and Cognitive Processes, 22, 1114-1139.
Horton, W. S., \& Gerrig, R. J. (2005). Conversational common ground and memory processes in language production. Discourse Processes, 40, 1-35.
Isaacs, E. A., \& Clark, H. H. (1987). References in conversation between experts and novices. Journal of Experimental Psychology: General, 116, 26-37.
Kosslyn, S. M. (1988). Imagery in learning. In M. S. Gazzaniga (Ed.), Perspectives in Memory Research. Cambridge, MA: MIT Press.
Matthews, D., Lieven, E., Tomasello, M. (2010) What's in a manner of speaking? Children's sensitivity to partnerspecific referential precedents. Devel. Psychology, 46, 749-760.
Metzing, C. \& Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects in the comprehension of referring expressions. Journal of Memory \& Language, 49, 201-213.
Nenkova, A., Gravano, A., \& Hirschberg, J. (2008). High frequency word entrainment in spoken dialogue. In Proc. of the ACL (Short Papers). 169-172.
Schober, M. F., \& Carstensen, L. L. (2010). Does being together for years help comprehension? In E. Morsella (Ed.), Expressing oneself/Expressing one's self: Communication, cognition, language, and identity. London: Taylor \& Francis.
Stoyanchev, S. \& Stent. A. (2009). Lexical and syntactic adaptation and their impact in deployed spoken dialog systems. In Proc. of NAACL HLT (Short Papers), 189192.

\title{
Tracking Persons Over Time is Tracking What?
}

\author{
Andrew Brook (andrew_brook@carleton.ca) \\ Department of Philosophy and Institute of Cognitive Science \\ Carleton University, Ottawa ON K1S 5B6 Canada
}

\begin{abstract}
Tracking persons, that is, determining that a person now is or is not a specific earlier person, is extremely common and widespread in our way of life and extremely important. If so, figuring out what we are tracking, what it is to persist as a person over a period of time, is also important. Trying to figure this out will be the main focus of this paper. (This paper will introduce a theme on tracking persons in Topics in Cognitive Science.)
\end{abstract}

Keywords: tracking persons; personal identity; personal identity - psychological criteria.

Tracking persons, that is, determining that a person now is or is not a specific earlier person, is extremely common and widespread in our way of life and extremely important. If so, figuring out what we are tracking, what it is to persist as a person over a period of time, is also important. Trying to figure this out will be the main focus of this paper.

I will begin with three preliminary points.
1. Philosophers call persisting as the person one is over time, i.e., what we are tracking when we track persons over time, personal identity. This is an unfortunate term, not least because the term 'identity' is now widely used to talk of features of personality, attitude to oneself, and the like. In this usage, it makes sense to talk of a strong identity, diffuse identity, identity crisis, etc., terms that make no sense in the philosophical context. For such reasons, I will generally speak of a person persisting or personal persistence.
2. Because of the possibility (in brain bisection operations, some say the actuality) of one person splitting into two people, by 'personal persistence' I do not mean a relationship that has to be one-to-one. (A charming depiction of one person becoming two is central to the film, To Be, by John Weldon (http://www.nfb.ca/film/to_be).) Likewise, it would appear that there can be degrees of persisting as oneself over time.
3. One might expect a paper on personal persistence to begin with at least a few comments on what a person is. That does not happen. The reason is that figuring out what we track or should track when we track persons also tells us a lot about what persons are.

\section*{1. Where tracking persons is central}

Tracking persons is at the heart of a great many social institutions, including
- Criminal law and punishment. Hence the effort that goes into determining that the person under arrest is the person who committed the crime.
- Obligations. You are now responsible only for obligations (contracts, promises, and the like) that you took on in the past.
- Property. You are now entitled to what you earlier owned. Sometimes, as in the case of educational policies, the changes to the person in the meantime can be massive.
- Credit. You are entitled to use only credit cards and the like approved for your use earlier. Hence photo ID.
- Insurance and benefits. The only benefits that you have now are ones that were assigned to you earlier (a very large issue in medicine in the United States). Likewise, you are entitled to recompense for harm done only if the harm was done to you or what belongs to you.
- Compensation. You are paid only for services you rendered or caused to be rendered earlier.
- Rewards. For example, you get the grade that your work earlier earned.

And so on. There seem to be two general principles behind tracking in these situations:
- Responsibility. A person is responsible only for what s/he, the same person, did (or caused to be done) in the past (a central feature of all western legal systems).
- Entitlement. A person is entitled to praise, benefits (including property), and compensation only for what \(\mathrm{s} / \mathrm{he}\) did in the past.

Tracking persons is even central to
- Interpersonal relationships. If you have just lost someone dear to you, your grief will not exactly dissipate upon being told, 'No problem. Your loved one had an identical twin who can do everything for you that \(\mathrm{s} / \mathrm{he}\) used to do.'

We even assume that we can track ourselves over time. Each of us, for example, has a:
- Special concern for one particular person's past. I might regret something you have done but I will not normally be ashamed of it (not normally because when I have a stake in your actions, by being your parent for example, I can feel ashamed of something you have done). And a,
- Special concern for one person's future, the person whom I believe will be me.

In short, tracking persons is central to much human social activity. (For further discussion of where we track persons in everyday life, see Shoemaker 2012.)

\section*{2. Current Tracking Practices}

The importance of tracking persons in our way of life is not matched by excellence in our tracking practices. The most common tracking practice uses facial similarity, whether in the form of eye-witness testimony in court (a practice that is not entirely reliable) or photo identification almost everywhere else now. Since two different people can look a lot alike, even entirely like in the case of identical twins, especially at a distance, and a single person can look very different in widely separated periods of life, tracking by facial similarity has its limitations.

Nonetheless, prosopagnosia, the inability to recognize faces, demonstrates how large a role tracking by facial appearance in fact plays. People with prosopagnosia cannot recognize people by face, therefore cannot tell whether they know a person before them or not. The result is that their lives are endlessly and embarrassingly complicated. To figure out whether they know a person before them, for example, they might have to get the person to speak.

Almost as common and equally limited is tracking by similarities in hand-writing, particularly signatures, a very common practice with contracts. In the legal system, similarity of fingerprints used to be the gold standard. It has been replaced by DNA sequencing, i.e., looking for similarities in the arrangement of molecule pairs in a particular stretch of DNA.

A common tracking practice with people we know but one that has received little attention is tracking by emotional reaction. We seem to have a distinctive emotional reaction to each person we know well. If a person before us triggers the distinctive reaction that we have to A and there are no countervailing factors (different gender, very different facial appearance and the like), this is a good reason for us to treat the person before us as A, the person to whom we have had the same reaction in the past.

One important piece of evidence for the importance of emotional reactions in tracking familiar people is the Capgras delusion. The Capgras delusion is the delusion that a person before one, a person whom one knows well and would normally care about, is an impostor. Despite the person before one looking like the familiar person, reacting
like the familiar person, expressing full and detailed memories of earlier events in the life of the familiar person - the person of course is the familiar person -, to someone in the throes of the delusion the familiar person is taken to be an impostor. (Capgras is usually accompanied by some form of major cognitive impairment such as severe schizophrenia.)

Neuroscience has not reached a settled view about what is going on the Capgras delusion but one widely held view is that, due probably to damage in the limbic system, the person suffering from the delusion has stopped reacting with the appropriate emotions to familiar, formerly liked or loved people. This is enough to convince the victim of the delusion that the person before him or her is not the person he or she knows and likes or loves. If this explanation is right, it would be evidence for the centrality of emotional reaction in reidentifying familiar people at a later time.

How well do our tracking practices relate to what matters in tracking persons? Not well. Return to fingerprints and DNA sequences. Let us suppose that the claims made for their uniqueness are right and the odds of two people having the same fingerprint or relevant DNA sequence are one in some billions. Would this tell us something valuable about what we want to track when we track a person? Even if fingerprints and DNA sequences were unique to each person, I don't think so.

The reason is that fingerprints and DNA are not what personal persistence consists in. They are merely features correlated with the person in question persisting. Again, suppose that the correlation is nearly perfect. Even here, knowing what a certain fingerprint is like or how a particular DNA sequence goes would tell us almost nothing about what being the bearer of that fingerprint or DNA sequence over time consists in. Indeed, knowing the fingerprint or sequence would tell us almost nothing about either their bearer in particular or what a person persisting in general consists in. Here is another way to put the same point: To know how well similarity of fingerprints or DNA correlate with the person before us being an earlier person, we need an answer to the question, correlate with what? What would it make the later person the earlier person?

\section*{3. Persons Over Time: What Interests Us?}

In the philosophical literature, two approaches to what must persist for a person to persist have dominated, the psychological approach and the somatic or bodily approach. On the psychological approach, the most frequent appeal is to the later person remembering events in the life of the earlier person in a particular way. However, psychological continuity of personality, abilities, and dispositions has also played a role. On the somatic or bodily approach, persistence of a functioning body or sometimes just a functioning brain has been front and centre.

It seems fairly clear that when we judge a person before us to be some earlier person, what primarily interests us are psychological factors. Even when somatic factors such as
facial similarity drive the judgment, it is because we take the somatic factor/s to be a reliable indication that the psychological factors of interest are present.

That psychological factors are what interest us can be shown in a number of ways. When we judge a person before us to be an earlier person, in the absence of special factors such a cognitive injury and dementia, we believe that the person before us will have a host of values, commitments, attachments, abilities, ways of viewing things, and so on so very similar to the same factors in the earlier person. Gaining this assurance is one of the reasons we track people. (We will return to this issue of psychological similarity.)
Another indication that psychological factors dominate is the way in which many people respond to brain bisection patients when, in special laboratory conditions, the body in question simultaneously does and disavows doing certain actions, responds to and disavows having heard certain requests, and so on. Many people (including Roger Sperry, who won a Nobel Prize for his work with brain bisection patients) take that it that, temporarily, two 'centres of consciousness' have appeared in these patients. Since there is only one body and brain throughout (albeit a brain whose corpus callosum has been severed in part or in whole), we cannot be making this judgment on somatic or even brain evidence and have to be making it on psychological factors, specifically, performance/lack of performance pairs.

A third argument is similar. When we learn about the idea of teletransportation (in Stars Trek episodes or whatever), the idea of a person being transported from one location to another without a single molecule of their body being transported makes instant sense to nearly everyone. For the idea to make sense, we have to be conceiving of the transfer as something psychological. (The same short film mentioned earlier, To Be by John Weldon, depicts teletransportation in a charming way.)

Finally, think of Kafka's Metamorphosis. In this story, the central character, Gregor Samsa, goes to sleep a human being and wakes up a 'monstrous vermin'. We have no trouble making sense of the idea that it is him, the very same person, who is now a bug. Yet the two bodies would not share any structure and not much if any matter.

Conclusion: What matters to us about personal persistence is something psychological. The next question is, what?

\section*{4. Memory}

What connects a later person to an earlier person when the earlier person persists as the later? Memory has a property that makes it a prime candidate. Unlike all other factors whether somatic or psychological, memory depicts events in the life of the earlier person. Its intentionality, to use that term of art, is backward-looking.

What kind of memory? Clearly we are not interested in short-term or working memory and we are not interested in procedural memory, memory of how to do things. What interests us is long-term declarative memory of some kind,
memory over substantial periods of time of what was the case. In the literature, three kinds of long-term declarative memory are distinguished:

Semantic memory (memory of facts, whether or not you were there)
Episodic memory (memory of events, usually with a requirement that you had witnessed the event)
Autobiographical memory (memory of events in one's own life, which can be both semantic and episodic)
However, this tristinction is not fine-grained enough for our purposes. For there are at least two kinds of autobiographical memory:
1. Remembering events in one's life 'from the inside', i.e., from the same point of view as the events were originally experienced. Thus one not only remembers an experience, a thought, or whatever, one remembers having the experience, the thought, or whatever. One not only remembers an action, one remembers doing the action. One not only remembers a feeling, one remembers having the feeling. And so on.
2. The rest - all the memories of events in one's life that are not from the standpoint of having lived them.
If I remember having had an experience, thought, feeling, it will appear to me that \(I\) had that experience, etc. I will appear to myself to be that person. And when I remember having had the experiences of an earlier person, or a series of person-stages tied together by a string of such memories, the appearance of the earlier person being me will be correct. Continuities and similarities can run from one person to another (Shoemaker, 2012, p. 12). However, I do not remember having others' experiences, etc. \({ }^{1}\) Absent some countervailing factor (such as reason to think that, for example, a memory transfer has taken place), if I have autobiographical memory 'from the inside' of having, doing, feeling a single earlier person's thoughts, experiences, actions and feelings, I am that person. That person has persisted as me.
Moreover, this suggestion about autobiographical memory 'from the inside' has more than intuitive appeal going for it, considerable though that is. We can use it to generate a nice theory of why we are responsible for earlier things we did and why we have a special concern for the future person who will be oneself.

Ask, why am I responsible for what I am doing right now?

Answer: Because I am the agent of the action - I experience myself from the standpoint of originating and doing the action.
And ask, why do I feel a special concern for me right now?
Answer: Because I will feel my pleasures and pains and other experiences - I experience them 'from the inside'. Likewise with plans and intentions. I put my plan in place, I act on my intentions. I merely observe the experiences, plans, and intentions of others

This suggests that I am responsible for an action of an earlier agent if I remember doing it and similarly for thinking, perceiving and feeling. Similarly, when I project my hopes and plans for my life onto a specific future person, when I feel special concern for a specific future person, I project onto and feel concern for the future person who will remember me as I am now 'from the inside'. A nice account. It flows directly from my account of remembering 'from the inside'.

\section*{5. Memory 'From the Inside' and Unified Consciousness}

There is a certain artificiality in what we have said about memory up to now. Contrary to the way I have written so far, we seldom remember having or doing or feeling individual experiences or actions. Usually what we remember about ourselves is far 'bigger' than that. Memories 'from the inside' are usually a kind of global representation:

Global representation - representing many objects as a single complex object.
What characterizes a global representation is that the representation of the elements of its object is united: One is aware of all the elements together, in a single conscious act, and one is aware of them not just as individual items but as a group.

To see how this works, consider representation of items that could be expressed by these sentences:
1. I am reading the words on the screen in front of me,
2. I am puzzled by your comments
3. I am enjoying the music I hear outside
4. I believe our agreement was to meet at 6:00
5. I thought I understood Kant's notion of the object
6. I wish the world were a fairer place

Here there are three different elements that could be united in a single global representation, (a) what I am representing, (b) the acts (act when unified) of representing them, and (c) myself as the subject doing the representing.

Similarly with memory. When I remember, for example, doing something, I nearly always also remember how I felt at the time, what I experienced at the time, the outcome of the action and how I felt about that, and so on. If so, my memory is a global representation that represents a unified group of earlier experiences and actions (see Raymont and Brook 2006)..

With this fuller description of memory, we can now give a fuller description of the relationship of memory to personal persistence. When we know the contents of a person's current global memory 'from the inside' of earlier experiences had, actions done, etc., and we track back and discover who had the global experience that is depicted in
the global memory, we know which earlier person the current person was.

Unified global experience and unified global memory 'from the inside' are a central part of what it is to be a (normal, cognitively intact) person. A persisting person is a series of global representations, each of which contains or contained memories of having thoughts and experiences, doing actions, feeling feelings 'from the inside', i.e., from the standpoint of having, doing and feeling them. Similarly, mutatis mutandis, for anticipating a future person as oneself. (This paragraph is my response to the wish discussed earlier for an account of what a person is.)

\section*{6. Problems with Memory}

So far, so good. But so far is not far enough. We do not ground judgments of personal persistence entirely on global autobiographical memory 'from the inside' and there are at least three challenges to the idea that we should do so.
1. In some cases, it appears that such memory is not necessary for personal persistence.
2. Some pressure can be put on the idea that it is always sufficient.
3. The kind of memory in question could in principle branch, go back to two or more earlier persons, or merge, two streams of memory becoming one.

Is memory necessary? Consider the most famous case in neuropsychology, Mr. H.M. (Henry Molaison, recently deceased). In the 1960s, to block epilepsy spreading from one hemisphere of his brain to the other, surgeons severed not just his corpus callosum but also the two halves of his hippocampus (and removed some other structures). This made it impossible for him to lay down new memories lasting more than about twenty minutes. Yet he was still taken to be a single, persisting person. No one questioned, for example, whether it was appropriate to continue to call 'him' by the same name or suggest that he was not the beneficiary of a pension plan created during the working life of the earlier person who had his name. Sacks (1970) discusses two cases with similar memory deficits.

To be sure, this attitude can be questioned. From the moment of the operation, HM was very different from people with normal memory. Post-operation, he never again entered a significant human relationship. (Even his care staff had to introduce themselves to him every morning.) He had no idea where he was and could not travel or even take a walk on his own. He had no knowledge of having had a life since the operation and so in one sense did not know who he was. If he had ever done anything that created entitlements or responsibility, he would have had no knowledge of having done so (so what would be the point of holding him responsible?). Thus he had no sense of accomplishment or failure, no pride in himself or guilt or shame, no sense of the trajectory of his life, no ... no ... no ... . And he could not plan a future for himself; his life did not have a planned or
desired trajectory. People with radical amnesia are very different from people with intact autobiographical memory.

Still, radical amnesia is not death. What makes memories 'from the inside' especially pertinent to personal persistence is that such memories depict, refer back to, earlier experiences and actions (and do so from the point of view of the person who had the original experience or did the original action). They do not, as we said, share this feature with any other kind of psychological state. However, they do share something else. Memories were caused by earlier experiences and actions.

Thus, memories are one kind of psychological continuity. When memory is missing, we can back off to other kinds of psychological continuity, ones that do not have backward intentionality. These can include continuitycarried similarities - HM, for example, had the same linguistic and arithmetic skills, the same knowledge of the world, the same manual abilities, and so on after the operation as before and the causes were primarily earlier events in the same body. Because HM is causally continuous with the body on whom the disastrous operation was performed, tracking his current causal continuities would lead us back to that body. We continue to find personal persistence even when memory is absent, I think, because we back off to these other kinds of causal continuity.

One very important non-memory causal continuity is continuity of plans, projects, and intentions. Usually I will have or have acted on much the same plans, life projects, etc., as I laid down for myself earlier and usually the main cause of having those plans, etc., now is that I laid them down for myself earlier. As we will see, continuity and discontinuity of such plans can make a difference in certain cases. I said that continuity of plans, etc., is a non-memory continuity and that is correct. However, they are usually carried from the past into the present in memories. Thus Mr. HM could not form any such plans, any that required him to remember them for more than half an hour at any rate. This is another and highly significant way in which he was radically unlike a person with normal memory.

Now our second question: Is global autobiographical memory 'from the inside' always sufficient for personal persistence? Cases where there has been massive personal change over time put some pressure on the idea. Let me sketch two real cases that certainly give one pause.

In 1941, one of Hitler's lieutenants, Rudolph Hess, flew to Scotland to try to negotiate a non-aggression treaty with England. (This would have left Hitler free to invade eastern Europe and the Soviet Union.) Hess was arrested as soon as he landed - and never lived outside a prison again. For many years, he was the sole inhabitant of last prison of the Allies in Germany, Spandau Prison in Berlin (the Soviet Union would not agree to his release), and so was effectively in solitary confinement. He died a very old man of 92 in 1987 (of either murder or suicide, theories vary). By the end, he was an embittered, cognitively-impaired shadow of his former self.

More recently, in 1998 Karla Faye Tucker was executed in Texas. She had taken part in a drug-fuelled murder at the age of 24 in 1983, so was in prison for close to \(40 \%\) of her life and nearly all her adult life. During her time in prison, she converted to Christianity and was not just a model prisoner but a counselor and mentor to other inmates. She even married the prison chaplain. In short, by the time she was killed, she could hardly have been more different than the out-of-control drug addict who took part in the murder.

Yet in both cases there was autobiographical memory and also psychological continuity and a single history, both psychologically and biologically. So all the tracking mechanisms that we normally use would lead us back to the same earlier person in both cases.

The trouble is, both cases raise the following question. Even though both people retained autobiographical memory and the usual continuities to the end, was there a sound basis at the time they died for taking the earlier person who bore their name, etc., at the time of their arrest, say, to have persisted as them? More directly, was there any justification for holding either of them responsible for what had been done by someone with the same name so many long years before?

Here is a basis for caution about how to answer these questions. Normally personal persistence carries with it persistence of character, life projects, and the like, so that if these things had been vicious earlier, they will be vicious now. And memory 'from the inside' will ensure that the later person knows about the earlier character - or at least the actions to which it gave rise. When character is no longer vicious, projects no longer malign, especially if accompanied by remorse or regret, the fact that the person nonetheless remembers his/her earlier character, projects, and actions 'from the inside' does not seem to matter as much.

And here the two cases differ. Tucker clearly fit the description of the paragraph above - but Hess did not. Tucker's character, life projects, and the like had been transformed. Hess, however, merely lost the power to act on his; he remained an unrepentant Nazi to his death. Thus, there would seem to be a better basis for continuing to hold Hess responsible for the actions of the triumphant young Nazi of old than for continuing to hold Tucker responsible for the actions of the earlier person who bore her name. (If so, it is a source of regret that Gov. Bush, as he was then, did not see things this way. Cases such as Hess and Tucker illustrate vividly that tracking persons can have profound ethical implications.)

Now the third question. What if a memories 'from the inside' lead back to two or more different people? Here there are two kinds of case, one where most of the memories originated in one person but a few originated in another, and one where the split is roughly equal.

In the first case, we could just ignore the aberrant memories, maybe by treating them as transferred somehow from another person. As to the second case, where
memories 'from the inside' originated in two different people about equally, we have grounds to hold that both people have persisted as the single current person with the memories. Because memories have a substructure of causal continuities and in our world the preponderant causal path carrying memory and other psychological continuities is within a body, memory transfer would take some very special technology.

\section*{7. ... And When There Is Little Or No Psychological Continuity Of Any Kind?}

There are also cases in which there is not just no memory but little or no psychological continuity of any kind, where nonetheless we take there to be a persisting person. Vegetative state patients are one kind of case. Newborns are a second. The relationship between me now as an adult and the newborn who was given my name decades ago is a third. (In the first two kinds of case, there is no psychological continuity. In the third, psychological continuity eventually developed but there is little or no psychological continuity running all the way back to the newborn.) How do we track persons in these cases?

Well, psychological continuities are causal continuities and some causal continuities exist in all three kinds of case. If materialism is correct, moreover, psychological continuities are one kind of bodily continuity. So we can, and do, fall back on other causal and bodily continuities such as looking alike, similarity of DNA, and a continuous causal history.

In summary, the pattern is this. When we have memory 'from the inside' and there are no countervailing factors, we stop there. When memories don't exist or have taken an unhelpful form, we back off to other psychological continuities. When psychological continuities are absent, we back off further, to non-psychological bodily continuities.

\section*{8. Practicalities}

Suppose that the story that we have told of what personal persistence consists in is at least roughly right. How would it connect to the tracking practices that we actually use, the ones that we laid out earlier? The answer is: At a conceptual level at least, not very tidily.

The problem is that in real life, it is often hard to gain knowledge of memories and other psychological continuities. To identify someone's memories, we need sophisticated skills in 'mind-reading' (assignment of psychological states to others) and considerable cooperation from the person in question. Such co-operation can, of course, be in short supply when you are trying to track people in the context of the criminal law, fraud, and
the like. So we resort to such things as facial appearance, fingerprints, and DNA pattern.

It may appear to be remarkable that such purely somatic factors work as well as they do for tracking what is, except in the rare special cases that we delineated in Stn. 7, a matter of psychological continuities. In fact, it is not. All the continuities we considered are underpinned by substantial causal continuities. Such continuities do not have to run though a single persisting body, as the very intelligibility of teletransportation shows. But in our world, they invariably do. For this reason, the rule, One persisting person per persisting body, works pretty well - and finding a reliable way to track persisting bodies over time is usually a pretty good way to track persisting persons. So facial similarity, fingerprints, and DNA sequences usually work pretty well. (For an excellent discussion of the relationships between cognition and the body, see Ismael 2007, Chapter 11, especially Section 5 .) \({ }^{2}\)

\section*{References}

Raymont, Paul and Andrew Brook. Unity of Consciousness, Stanford Electronic Encyclopaedia of Philosophy (plato.stanford.edu/unity of consciousness).
Ismael, J. 2007. The Situated Self/ Oxford: Oxford University Press.
Kafka, F. 1915, trans. I. Johnson 2009. The Metamorphosis. <http://records.viu.ca/~johnstoi/stories/kafka-E.htm>.
Parfit, Derek. 1984. Reasons and Persons. Oxford: Oxford University Press.
Sacks, Oliver. 1970. The Lost Mariner and A Matter of Identity. In The Man Who Mistook His Wife for a Hat. New York: Harper \& Row, pp. 23-42 and 108-115.
Shoemaker, D. 2012. Personal Identity and Ethics. Stanford Encyclopedia of Philosophy <plato.stanford.edu/personal identity>
Weldon, John. To Be. National Film Board of Canada <http://www.nfb.ca/film/to_be>

\section*{Endnotes}
\({ }^{1}\) Bishop Joseph Butler claimed in the \(17^{\text {th }}\) century that we cannot by definition remember having an experience had by another. If so, being the same person is a requirement of the kind of memory we are discussing - and, of course, cannot be used to define or analyze it. Here I will just assume that we can define a form of memory that does not presuppose personal persistence.
\({ }^{2}\) Thanks to Ted Lougheed, Dave Matheson, Jordan Dodd, Nicolas Bullot, and audiences at Carleton University and the University of Ottawa for helpful comments.

\title{
Plausibility and Visualizability in Relational Belief Revision
}

\author{
Leandra Bucher (leandra.bucher@psychol.uni-giessen.de) \\ Jelica Nejasmic (jelica.nejasmic@psychol.uni-giessen.de) \\ Sabine Bertleff (sabine.bertleff@psychol.uni-giessen.de) \\ Markus Knauff (markus.knauff @ psychol.uni-giessen.de) \\ Justus-Liebig University, Experimental Psychology and Cognitive Science, Otto-Behaghel-Str. 10F, 35394 Giessen, Germany
}

\begin{abstract}
Belief revision is required when new facts are incompatible with existing beliefs. In the present experiment, participants changed their mind about the spatial and non-spatial relations between objects. The participants received information about relations, which were subsequently contradicted by irrefutable counterfacts. The task was to decide which of the initial relations to retain and which ones to give up. Previous experiments showed that these decisions are guided by the linguistic asymmetry between located (LO) and reference objects (RO). Reasoners have a strong preference to relocate the LO of the counterfactual relation. Our experiment explores whether this robust effect can be overwritten by the plausibility of revised beliefs; and how visualizability of problems affects revision. We found the LO-preference to be robust even when the resulting representation is implausible; and that revision is impeded when problems are easy to visualize. The results shed new light on relational belief revision in humans.
\end{abstract}

Keywords: Relational reasoning; Spatial reasoning; Belief revision; Mental models, Visual impedance

\section*{Relational Reasoning and the Revision of Beliefs}

Imagine you involuntarily put on some weight over the Christmas holidays. That is why, for the next couple of months, in order to get rid of the additional pounds, you consider nutrition which is low in fat and calories. You know that pasta, buckwheat, potatoes, and fruits are all low in fat, and further that potatoes are higher in calories than buckwheat is, and that pasta provides more energy than potatoes and fruits. Your ability to rank these, and even more, different types of food according to the amount of energy they provide enables you to conclude that fruits are a good choice when you want to pursue your aim of weight loss. This little example demonstrates that reasoning with relations is essential in our daily life. In fact, it is ubiquitous and it plays a vital role in higher cognitive processing, for instance, in planning and categorizing (Halford, Wilson, \& Phillips, 1998; 2010; Hummel \& Holyoak, 2005).

Now, imagine you learn about avocado fruits that they contain high amounts of fat. You presumably integrate this fact with ease into your knowledge base, although it is not coherent with what you thought you knew about fruits (that they were low in fat). The process of integrating nonconsistent pieces of information into already existing belief sets is referred to as belief revision (e.g. Gärdenfors, 1988; Elio \& Pelletier, 1997; Wolf, Rieger, \& Knauff, 2012).

Reasoners usually revise their beliefs about the state of the world when confronted with contradicting evidence. Indeed, we frequently encounter new facts that do not cohere with our beliefs. When the source of a new piece of information is reliable and the fact itself somewhat indisputable, we might consider taking it into account. In case we do, it entails that we update knowledge bases and revise current sets of beliefs.

Frequently, there are multiple ways in which the revision could be performed, implicating different decisions about which beliefs to maintain and which ones to discard. Consider your belief that fruits are a good choice when you want to lose weight: do you maintain it in the face of the fact that avocados are high in fat; or will you discard at least avocados from the diet menu? Do you still think of avocados as fruits after all? It is clear that belief revision is often accompanied by uncertainty and ambiguity.

The current study relies on recent work done in the field of relational belief revision. A recent finding in studies that looked at belief revision about spatial relations is that the revision is based on the variation of spatial mental models (Bucher, Krumnack, Nejasmic, \& Knauff, 2011; Krumnack, Bucher, Nejasmic, \& Knauff, 2011; Bucher \& Nejasmic, 2012; Knauff, Bucher, Krumnack, \& Nejasmic, 2013). Often, there are multiple (logically equal) alternatives for variations that would all re-establish consistency. However, human reasoners hold strong preferences for specific alternatives. These preferences can rely on linguistic cues provided by relational statements. The experiment presented here was designed to investigate whether reasoners still rely on these cues during revision, even when the resulting object relations are implausible. Furthermore, we compared reasoners' performance in problems that were easy to visualize and easy to spatially represent.

\section*{Preferences in Spatial Belief Revision}

Our recent experimental studies have focused on the revision of object arrangements. Imagine a person has reason to think that the objects \(\mathrm{X}, \mathrm{Y}\), and Z are arranged in this linear order. The spatial mental model that is constructed can be sketched as:
\[
X-Y-Z
\]

Let us assume the reasoner then learns from a reliable and trustworthy source that as an incontrovertible fact, "object Z is to the left of object \(X\) ". This fact is inconsistent with the
reasoner's model. In order to take the fact into account and - at the same time - keep changes to the model as little as possible, the reasoner can vary the model in two different ways: the X can be relocated; the Z can be relocated. These two alternatives are comparable, from a logical point of view.

The finding of recent studies is that reasoners encounter this ambiguity with clear and robust preferences. Preferred model revisions of the type introduced here are guided by cues provided by the conflicting statements. Binary relations - such as "Z left of X" - feature a functional asymmetry between the two objects, well known as distinction of figure and ground, target and anchor, or (the terminology used in the present context) "located" (LO; the "Z" in "Z left of X") and "reference" object (RO; the " X " in " \(Z\) left of \(X\) "). The asymmetry of LO and RO specifies the location of the LO relative to the location of the RO (Miller \& Johnson-Laird, 1976; Talmy, 1983; Landauer \& Jackendorff, 1993). Reasoners tend to perceive the RO's position as fixed and inflexible while the LO is considered to be more flexible and locatable.
The following example sketches a reasoner's characteristic preference for the revision of a horizontal linear arrangement of the objects \(\mathrm{X}, \mathrm{Y}\), and Z :

Arrangement: \(\mathrm{X}-\mathrm{Y}-\mathrm{Z}\)
Counterfact: Z is left of X ,
with \(Z\) as the \(L O\) of the counterfact and \(X\) as the \(R O\)
Revisions: (1) \(\mathrm{Z}-\mathrm{X}-\mathrm{Y}\)
(2) \(Y-Z-X\)

The revised arrangement (1) results from the relocation of the counterfact's LO relative to its RO and is usually the preferred revision. The logical equivalent but non-preferred alternative (2), results from the relocation of the RO relative to the LO. The LO-preference is a strong effect. Indeed, reasoners apply this principle in around \(90 \%\) of the problems of the described type (Bucher et al., 2011; Krumnack, Bucher, Nejasmic et al., 2011; Bucher \& Nejasmic, 2012; Knauff et al., 2013).
Note that abstract entities such as \(\mathrm{X}, \mathrm{Y}\), and Z are neutral with regard to the position within an arrangement. The same applies for objects such as fruits (apple, mango, orange) and tools (hammer, drill, pliers). Indeed these were the objects used in the experiments so far.

Here, as a novelty, we manipulated two factors: the plausibility of revisions and the visualizability of the statements. We used spatial and non-spatial relations of objects "that make sense", e.g. "an elephant is bigger than a fly". The statements used in the problems differed with regard to their visualizability, i.e. in their extent to which they provoke picture-like representations ("mental images").

The first question is: do reasoners still apply the LOprinciple when the revised model is implausible? In fact, reasoners often base their problem solutions on the plausibility of the content or on prior experiences within a
certain field (Newstead, Pollard, Evans, \& Allen, 1992; Klauer, Musch, \& Naumer, 2000; Evans, 2008, DeNeys, 2006; Knauff, Budeck, Wolf, \& Hamburger, 2010). These content effects show the strong tendency of reasoners to take into account what is meaningful or plausible. On the other hand, the LO-preference is a strong effect.

The second question is: does the visualizability of a problem modulate revision? Relations which are easy to visualize, impede reasoning (Knauff \& Johnson-Laird, 2002; Knauff, Fangmeier, Ruff, \& Johnson-Laird, 2003; Knauff \& May, 2006; Knauff, 2009). Mental images are considered to be irrelevant for reasoning itself but the inspection of the images appears to slow down thinking and makes it more prone to errors. This so-called visual impedance effect occurs complementary to the facilitating effect of spatial relations (Knauff, 2009; Knauff, 2013). Spatial belief revision is conceived as the manipulation of spatial mental models. The assumption for the current experiment is that models which are easy to mentalize as visual images should accordingly be harder to manipulate by a reasoner than models constructed from easy to spatially representable statements.

In order to prepare the manipulation of the experimental problems' visualizability, we conducted a pilot study.

\section*{Pilot study: the Visualizability of Statements}

Participants of the pilot study rated statements with regard to their visualizability. This procedure allowed the allocation of statements to categories: visual, neutral, and spatial.

\section*{Method}

30 volunteers ( 14 male; aged from 19 to 55) participated in the study. Each of them rated individually, 72 binary spatial and non-spatial relational statements according to their visualizability. The statements were accessible online via a link sent by email. They were generated and the data collected, using LimeSurvey, Version 1.92+ software. Example statements are: "Asparagus is thinner than cucumber", "Cucumber is thinner than cabbage"; "Whisper is quieter than speech"; "Speech is quieter than scream".

Participants rated the subjectively perceived visualizability of each statement on a scale with the points: "very easy to visualize"; "easy to visualize"; "easy to visualize and spatially represent"; "easy to spatially represent"; very easy to spatially represent"; and "neither easy to visualize nor to spatially represent". The four most clear-cut rated statements from the three categories, "very easy to visualize", "neither easy to visualize nor to spatially represent", and "easy to spatially represent" were chosen as experimental material. In accordance with these ratings, the relations were allocated to one of three experimental conditions: "visual"; "neutral", "spatial". Table 1 shows example statements.

Table 1: Examples of statements used in the experiment
Visual
The cucumber is thinner than the pumpkin. The asparagus is
thinner than the cucumber.
Neutral
The bird is weaker than the dog. The dog is weaker than the polar
bear.
Spatial
Russia is further east than Poland. Poland is further east than
Germany.

\section*{Discussion of the Pilot Study}

It is clear that many people experience their thinking as inspection of visual images. However, our pilot study indicates that some relations are more "visual" than others. The results show that, on the one hand, the categories "visual", "neutral", and "spatial" have no clear-cut borders. On the other hand, however, the results also clearly show that some relations are experienced as more visual than others while some relations are experienced as more spatial than others. So, we do not have relations that are purely visual or spatial. However, for our main experiment we could identify relations which are more visual or more spatial than other relations.

\section*{Experiment: Plausibility and Visualizability}

For the main experiment, the visualizabilty of the problems and the plausibility of revisions, were manipulated. Regarding plausibility, we relied on common knowledge. We assumed that a statement such as "the father is younger than the grandfather" is regarded as plausible, while the invers relation, "the grandfather is younger than the father" as implausible.

\section*{Method}

Participants A new group of 20 volunteers ( 8 male ; age range from \(20-35\); all native speakers of German) gave written informed consent to participation. They were tested individually in a quiet lab room.

Materials, Procedure, and Design The experiment is based on a \(3 \times 2\) (within-subject) design. We manipulated the factors visualizability (visual, neutral, spatial) and plausibility (plausible, implausible). The experiment consisted of 64 problems in the visual, neutral, and spatial condition, respectively. During the revision phase, participants chose between plausible and implausible revised models.

In the first phase, the description phase, the participant received two statements (premises, \(P\) ) describing the relations between three entities. In half of the problems, P1 was plausible and P2 implausible. In the other half, it was reversed. The premises were presented in a sequential manner, each at one time, by the participants' own speed. See an example problem of the "visual" condition below:

\section*{Description:}

P1: "Asparagus is thinner than cucumber"
P2: "Pumpkin is thinner than asparagus"
The task of the participants was to order the entities according to the description. Subsequently, two "models" were presented on the left and the right side of the monitor. One of the models was "correct", i.e. it was in agreement with P1 and P2, the other one was "incorrect".

Models constructed from the description:
Correct: Pumpkin Asparagus Cucumber
Incorrect: Cucumber Asparagus Pumpkin
Presentation locations of correct and incorrect models on the left and right side of the monitor were counterbalanced across the experiment. Participants were asked to indicate the correct model by pressing a left or right button. This step of the "correct model choice" was implemented in order to warrant that participants constructed the "correct model" before entering the next phase of a problem.

There is evidence that reasoners order objects spatially even when the relations are non-spatial. "Venus shines brighter than the moon but the sun shines even brighter", can easily be reflected by the order: Moon - Venus - Sun. Relations, also non-spatial ones, are thought to be closely linked to space. The argument of many researchers is that mental space is relational (rather than geometrical) space (e.g. Knauff, 1999; Knauff, 2013). This notion is corroborated by many findings, e.g. that spatial distance effects also occur with non-spatial relations (Prado, Van der Henst, \& Noveck, 2008; Prado, Chadha, \& Booth, 2011).

Indeed, participants' performance was very accurate. In more than \(90 \%\) of the cases ( \(M=92.90 \%\); SD=0.26), the correct models were selected. The few incorrect problems were excluded from further analysis.

In the second phase, the participants received a third premise which they were explicitly instructed to treat as an incontrovertible fact (while the instruction included the hint that the participant could not be entirely sure whether the description was true). The "fact" was always plausible. In half of the problems, it was consistent with P1 and P2; in the other half (see the example below) it was inconsistent.

\section*{Counterfact: "Cucumber is thinner than pumpkin"}

The participants decided - using "yes"- and "no"-buttons whether the fact was in agreement with the initial statements or not. Again, participants performed very accurate in this phase. In \(86.20 \%(S D=10.59)\) of the problems, the participants decided correctly. Incorrect problems were eliminated from further analysis, so were the consistent ones.

The third phase, the revision, was the most interesting part of the experiment. This part followed only if the participant recognized a fact as inconsistent with the initial description. Participants were then instructed to revise their
assumption about the objects' relations by taking into account the counterfact. Two alternative revised models, both variations of the initial model, taking into account the fact while preserving as much of the initial information as possible, were presented on the screen. The two revised models were presented on the left and the right side of the computer monitor. The task was to choose among the models the one which matched the participant's assumption about the revised object relations. Choices were indicated by left and right button presses. One of the revised models was plausible; the other one was implausible. The question was whether reasoners still apply the LO-principle or whether they prefer revisions based on the plausibility. The two alternative revised models for the example above were:
(1) Cucumber Pumpkin Asparagus
(2) Asparagus Cucumber Pumpkin

Note that model (1) results from the relocation of the LO of the fact (which is the cucumber) but leads to an implausible order of objects. Model (2), in contrast, results from the relocation of the RO of the fact (which is the pumpkin) but leads to a plausible order of the objects. Over the entire set of problems, in half of the problems the LO-principle led to implausible and the RO-principle into plausible relations of the entities (as in the example above), in the other half of the problems it was reversed.

Revision choices and duration were recorded. The problems were presented in a random order. They were preceded by eight practice trials (not analyzed). All stimuli were generated, presented, and recorded with Superlab 4.0 (Cedrus Corporation, San Pedro, CA, 1999) with an RB-530 response box running on a standard personal computer connected to a 19 '' -monitor.

\section*{Results and Discussion}

In the first analyses, we examined whether revision preferences were based on plausibility. Subsequently, we looked at the effects of visualizability. We also looked at the interactions between plausibility and visualizability. However, none of them reached the level of statistical significance ( \(p \mathrm{~s}>.05\) ).
Plausibility: ANOVAs were calculated, with the factors Plausibility (plausible, implausible) \(\times\) Relocated Object (LO, RO), separately for the frequency (in percent) of the respective revision choices and revision duration (in seconds). Both ANOVAs revealed a main effect of Relocated Object (choices: \(\left[F(1,19)=71.91 ; p<.001 ; \eta_{\text {part }}^{2}\right.\) \(=.79]\); duration: \(\left[F(1,19)=6.53 ; p=.019 ; \eta_{\text {part }}^{2}=.26\right]\); all other \(p \mathrm{~s}>.20)\). LOs were relocated more often and faster compared to ROs. Choices LO vs. RO: \(M=78.77 \% ; S D=\) 14.99 vs. \(M=21.23 \% ; S D=14.99 ; t(19)=8.59 ; p<.001\); duration LO vs. RO: \(M=2.69 \mathrm{~s} ; S D=1.71\) vs. \(M=3.46 \mathrm{~s}\); \(S D=1.74: t(19)=-2.35 ; p=.03)\).


Figure 1. Revisions [\%] and revision durations [s; error bars indicate standard errors] of "located" (LO) and "reference" objects (RO) showed an LO-effect. The preference was not modulated by plausibility

Figure 1 provides a graphically overview of the data. The result suggests that reasoners were guided by the distinction of LO and RO provided by the counterfactual relation. They followed the asymmetry of the objects and relied on the LOprinciple. Plausibility did not overwrite this preference. Next, we examined the impact of the visualizability of the statements. The question was: does the easiness to construct a visual mental image or a spatial representation of the problems affect reasoning and belief revision?

Visualizability: in order to compare the revision duration of visual, neutral and spatial problems, an ANOVA with the within-subject factor Visualizability (visual, neutral, spatial) was calculated. It indicated a significant main effect \(\left[F(2,18)=4.80 ; p=.014 ; \eta_{\text {part }}^{2}=2.02\right]\). When the statements were easy to visualize, the revision duration was significantly higher ( \(M=3.00 \mathrm{~s} ; S D=1.3\) ) compared to neutral and spatial problems (neutral: \(t(19)=-2.70 ; p=\) .014; spatial: \(t(19)=-2.73 ; p=.013\) ). Revision duration for neutral ( \(M=2.60 \mathrm{~s} ; S D=1.60\) ) and spatial problems ( \(M=\) 2.6 s ; \(\mathrm{SD}=1.3\) ) were comparable ( \(p>.85\) ).

Figure 2 provides a graphical overview. The result clearly suggests an impeding effect of statements that are easy to visualize. We also looked at the interaction between visualizability and relocated object, which was nonsignificant ( \(p>.35\) ).


Figure 2. Mean revision durations of different relation types [ s ; error bars indicate standard errors] indicate a visual impedance effect

\section*{General Discussion}

Belief revision is performed in order to re-establish consistency within belief sets (Gärdenfors, 1988). Frequently, there exists ambiguity because there are multiple solutions for revision. The present experiment on relational belief revision agrees with recent work suggesting that reasoners solve this ambiguity with strong preferences. Recent experiments used objects (e.g. fruits) which are "neutral" regarding their position within object arrangements. These objects were also not related to the individuals' prior knowledge or pre-existing beliefs. (e.g. Knauff et al., 2013). The current experiment, in contrast, addressed two novel aspects in reasoning with spatial and non-spatial relations: the plausibility of a relation and the visualizability of the reasoning problems. Both aspects have been shown to affect reasoning in general (e.g. Evans, 2008; e.g. Knauff, 2009).

A powerful theory in cognitive science puts forward that reasoners represent situations and states of the world in "mental models"; and that these models provide the basis for reasoning (Johnson-Laird \& Byrne, 1991; Goodwin \& Johnson-Laird, 2005; Krumnack, Bucher, Nejasmic, Knauff, 2010; Krumnack, Bucher, Nejasmic; Nebel, \& Knauff, 2011). Indeed, the mental model theory is corroborated by many phenomena. Moreover, model-based reasoning rather than the application of formal rules (e.g. Rips, 1994) nicely explains why reasoners often ignore the logical form of an argument. In fact, reasoners often base their problem solutions on the plausibility of the content or on prior experiences within a certain field, rather than on the validity of a conclusion (Newstead et al., 1992; Klauer et al., 2000; Evans, 2008, DeNeys, 2006; Knauff et al., 2010). These content effects show that reasoners have a strong tendency to take into account what is meaningful or plausible to them, even when this entails a trade-off with logic.
Recent findings on spatial belief revision suggest that reasoners vary spatial mental models and that they prefer certain variations above others. The variation of simple
spatial models of "neutral" objects was found to be based on a principle which we call the LO-preference. The first aim of the current experiment was to test whether reasoners hold on to that preference, even when it leads to implausible models. Our data suggest that they do. The LO-preference remained the guiding revision principle even when the resulting model was implausible.

Are there alternative interpretations of this result? One alternative account is that the effect is due to the specific layout of our experiment. In fact, during the construction phase, reasoners were forced to partially "ignore" plausibility of relations in order to construct the correct initial model from plausible and implausible statements. This might have triggered them to do the same in the revision process. Thus, they also ignored the plausibility of the revised model. We think that this might be a possible explanation for the finding that the LO-preference was stronger than the plausibility of the revised model. However, we think that the robustness of the LOpreferences is still an important result. In our future research, we will explore whether the plausibility effect is more powerful in more complex revision tasks. We assume that with more complex problems, the LO-effect on model variation would disappear and "plausibility" would play a more important role.

An important finding in the area of relational reasoning is that the visualizablity of a relation can modulate reasoning performance. Relations which are easy to visualize as mental images impede reasoning (e.g. Knauff \& JohnsonLaird, 2002). Reasoning with relations is best described by the construction and the manipulation of spatial mental models (e.g. Johnson-Laird \& Byrne, 1991; Schaecken, Johnson-Laird, P. N., \& d'Ydewalle, 1996; Goodwin \& Johnson-Laird, 2005; Jahn, Knauff, \& Johnson-Laird, 2007; Nejasmic, Krumnack, Bucher, \& Knauff, 2011). It is likely that problems that are easy to spatially represent accommodate reasoning because of their shared nature with (spatial) mental models. Image-like representations, in contrast, impede reasoning because they hold additional but irrelevant information (Knauff, 2009; 2013). Our results corroborate these assumptions. With the present experiment, we found an influence of the visualizability on revision. Problems that were easy to visualize appeared to impede the revision process. Indeed, visual problems seem to provide an additional effort which slows down the revision process. In contrast, relations that were rated as easy to represent spatially were manipulated faster during the revision phase. This is in line with the assumption that those relations accommodate revision because they share their spatial structure with the spatial model that is varied. In our experiment, spatial and neutral relations were both processed faster than visual relations. This result supports the assumption that spatial and non-spatial relations are both easily integrated into spatial models. Pursuing this thought could possibly reveal more interesting aspects of the mental space as relational space (Knauff, 1999; Prado et al., 2008; Prado et al., 2011).

\section*{Acknowledgments}

This work was funded by a grant from the DFG (KN 465/62). We thank Martin Becker and Pascal Pfendert for assistance with the experiment.

\section*{References}

Bucher, L., Krumnack, A., Nejasmic J., \& Knauff, M. (2011). Cognitive processes underlying spatial belief revision. In L. Carlson, C. Hölscher, T. Shipley (Eds.), Proceedings of the 33rd Annual Conference of the Cognitive Science Society (pp. 3477-3482). Austin, TX: Cognitive Science Society.
Bucher, L. \& Nejasmic, J. (2012). Relocating multiple objects during belief revision. In C. Stachniss, K. Schill, \& D. Uttal (Eds.), Lecture Notes in Artificial Intelligence: Spatial Cognition (pp. 476-491). Berlin, Germany: Springer.
De Neys, W. (2006).Dual processing in reasoning - Two systems but one reasoner. Psychological Science. 17, 428433.

Elio, R., \& Pellitier, F. J. (1997). Belief change as propositional update. Cognitive Science, 2, 419-460.
Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social Cognition. Annual Review of Psychology, 59, 255-278.
Gärdenfors, P. (1988). Knowledge in flux. Cambridge, MA: MIT Press.
Goodwin, G. P., \& Johnson-Laird, P. N. (2005).Reasoning with relations. Psychological Review, 112, 468-493.
Jahn, G., Knauff, M., \& Johnson-Laird, P. N. (2007). Preferred mental models in reasoning about spatial relations. Memory \& Cognition, 35, 2075-2087.
Johnson-Laird, P. N., \& Byrne, R. M. J. (1991). Deduction. Erlbaum, Hove (1991).
Klauer, K. C., Musch, J., \& Naumer, B. (2000). On belief bias in syllogistic reasoning. Psychological Review, 107, 852-84.
Knauff, M. (1999). The cognitive adequacy of Allen's interval calculus for qualitative spatial representation and reasoning. Spatial Cognition and Computation, 1, 261290.

Knauff, M. (2009). A neuro-cognitive theory of deductive relational reasoning with mental models and visual images. Spatial Cognition \& Computation, 9, 109-137.
Knauff, M. (2013). Space to Reason. Cambridge, MA: MIT Press.
Knauff, M., Bucher, L., Krumnack, A., \& Nejasmic, J. (2013). Spatial belief revision. Journal of Cognitive Psychology, 1, 1-10.
Knauff, M., Budeck, C., Wolf, A. G., \& Hamburger, K. (2010). The illogicality of stockbrokers: Psychological experiments on the effects of prior knowledge and belief biases on logical reasoning in stock trading. PLoS One, 5(10), e13483. doi:10.1371/ journal.pone.0013483.
Knauff, M., Fangmeier, T., Ruff, C. C., \& Johnson-Laird, P. N. (2003). Reasoning, models, and images: Behavioral
measures and cortical activity. Journal of Cognitive Neuroscience, 15, 559-573.
Knauff, M., \& Johnson-Laird, P. N. (2002). Visual imagery can impede reasoning. Memory \& Cognition, 30, 363371.

Knauff, M., \& May, E. (2006). Mental imagery, reasoning, and blindness. Quarterly Journal of Experimental Psychology, 59, 161-177.
Krumnack, A., Bucher, L., Nejasmic, J., \& Knauff, M. (2010). Spatial reasoning as verbal reasoning. In S. Ohlsson \& R. Catrambone (Eds.), Proceedings of the 32nd Annual Conference of the Cognitive Science Society (pp. 1002-1007). Austin, TX: Cognitive Science Society.
Krumnack, A., Bucher, L., Nejasmic, J., \& Knauff, M. (2011). Efficiency and minimal change in spatial belief revision. In L. Carlson, C. Hölscher, \& T. Shipley (Eds.), Proceedings of the 33rd Annual Conference of the Cognitive Science Society (pp. 2270-2275). Austin, TX: Cognitive Science Society.
Krumnack, A., Bucher, L., Nejasmic, J., Nebel, B., \& Knauff, M. (2011). A model for relational reasoning as verbal reasoning. Cognitive Systems Research, 11, 377392.

Landau, B. \& Jackendoff, R. (1993) "What" and "where" in spatial language and spatial cognition. Behavioral and Brain Sciences, 16, 217-265.
Miller, G. A. \& Johnson-Laird, P. N. (1976). Language and Perception. Cambridge MA: Havard University Press.
Nejasmic, J., Krumnack, A., Bucher, L., \& Knauff, M. (2011). Cognitive processes underlying the continuity effect in spatial reasoning In L. Carlson, C. Holscher, \& T. Shipley (Eds.), Proceedings of the 33rd Annual Conference of the Cognitive Science Society (pp. 11271132). Austin, TX: Cognitive Science Society.

Newstead, S. E., Pollard, P., Evans, J., \& Allen, J. (1992). The source of belief bias in syllogistic reasoning. Cognition, 45, 257-284.
Prado, J., Chadha, A., \& Booth, J. R. (2011). The brain network for deductive reasoning: a quantitative metaanalysis of 28 neuroimaging studies. Journal of Cognitive Neuroscience, 23, 3483-3497.
Prado, J., Van Der Henst, J.-B., \& Noveck, I. (2008). Spatial associations in relational reasoning: Evidence for a SNARC-like effect. The Quaterly Journal of Experimental Psychology, 61, 1143 - 1150.
Rips, L. J. (1994). The Psychology of Proof. Cambridge, MA: MIT Press.
Schaecken, W., Johnson-Laird, P. N., \& d'Ydewalle, G. (1996). Mental models and temporal reasoning. Cognition, 60, 205-234.
Talmy, L. (1983). How language structures space. In H. L. Pick \& L. P. Acredolo (Eds.), Spatial Orientation: Theory, Research and Application (pp. 225-282). London, UK: Plenum.
Wolf, A.G., Rieger, S., \& Knauff, M. (2012). The effects of source trustworthiness and inference type on human belief revision. Thinking \& Reasoning, 18, 417-440

\title{
Mechanistic Diagrams as Search Organizers
}

\author{
Daniel C. Burnston (dburnsto@ucsd.edu) \\ Department of Philosophy and Center for Chronobiology, University of California, San Diego, La Jolla, CA 92093-0119 USA
}

\begin{abstract}
Many in cognitive science have noted the importance of external visualizations for reasoning and learning, and have suggested that such visualizations play a role in complex reasoning contexts such as scientific investigation. However, what cognitive role diagrams play in scientific reasoning is unclear. I suggest that mechanistic diagrams function as search organizers in active research projects. Diagrams aid in scientific reasoning by being uniquely positioned to coordinate cognitive search through multiple search spaces, both within an individual and within a field. I examine this role using a number of published diagrams from mammalian chronobiology.
\end{abstract}

Keywords: scientific diagrams; cognitive search; chronobiology.

\section*{Introduction}

Diagrams are nearly ubiquitous in biological practice, in which the goal is often to construct an explanation of the mechanisms responsible for complicated phenomena. In journal articles, conference presentations, and whiteboard discussions, research scientists continually engage in the process of constructing, analyzing, and modifying diagrammatic representations. This ubiquity suggests that diagrams play an important role in reasoning about phenomena in biology. However, the specific role that diagrams play in reasoning is not established. Part of the difficulty is the sheer complexity of the reasoning processes involved. Scientists must coordinate a variety of representational resources in constructing mechanistic explanations, and diagrams are often involved in characterizing the phenomena of interest, organizing and presenting obtained data, and conveying the parts, operations, and organization of a proposed mechanism (Gooding, 2010; Sheredos, Burnston, Abrahamsen, \& Bechtel, forthcoming).

The current literature on diagrams has focused largely on the meaning of diagrammatic elements and how they relate (Tversky, 2011), as well as on how diagrams might encode complete explanations (Perini, 2005), or function as learning tools for novices (Cheng, 2011). While these are important analyses, they leave a gap in understanding how diagrams might play a role, even for experts, in constructing explanations of complex phenomena. Pioneering historical studies of episodes of scientific discovery (Cheng, 1992; Nersessian, 2008) have focused on the use of visualizations in the efforts of individual scientists to explain mathematical or physical phenomena. The use of diagrams in an active field of contemporary science presents new challenges, as
diagrams are used, discussed, and appropriated by numerous different researchers, each with different backgrounds, interests, and experimental skills.

In this paper I propose an account of how diagrams aid scientific reasoning in an active research field. The proposal draws upon several approaches in cognitive science, which construe reasoning as involving "cognitive search"-a process of selecting the right representations out of a space of possibilities in order to meet a cognitive goal. The search perspective on diagrams contends that diagrams facilitate scientists' cognitive search through the complex realm of possibilities that are relevant in explaining natural phenomena. Specifically, I contend that diagrams provide an external search space that allows for the coordination of both conceptual and experimental resources, both by individuals and by entire scientific communities.

I focus on diagrams of proposed mechanisms in biology. This type of diagram generally consists of proposed entities or events depicted by shapes and/or linguistic labels, organized in visual space, and related via arrows, lines, and enclosures to convey the structural, causal, functional, and/or conceptual (e.g., categorical) relations between them. Often, these diagrams occur at the beginning or end of a research article as a way of organizing the findings into a model of how physical components might interact to produce phenomena of interest. Moreover, these diagrams are often changed and expanded over a series of publications in order to incorporate new results.

While receiving relatively little attention in analyses of scientific practice, diagrams of this type can be extremely important, as even minor changes to diagrammatic form can have large effects on how the organization of a mechanism is interpreted. Sheredos et al. (forthcoming) have argued that these differences both constrain thought in particular ways and afford particular inferences that are useful for hypothesis construction. The search perspective expands on this viewpoint to further elucidate the use of mechanistic diagrams in a research field.

I will develop this perspective by analyzing the use of mechanistic diagrams in mammalian chronobiology. Chronobiologists study phenomena of circadian rhythmsdaily, roughly 24 hour cycles in biological activity-in a variety of different organisms, at each of the behavioral, physiological, and molecular levels. Many biological processes exhibit circadian rhythms, including gene transcription, cell division, metabolism, and overt sleep and feeding behavior. Over the last 25 years, much progress has been made in understanding how these rhythms are
regulated by internal, molecular clocks that both keep time and entrain the organism to environmental timing cues such as light and temperature.

Due to its extensive use of diagrams in identifying and explaining complex phenomena, chronobiology is a fertile ground for inquiries about diagrammatic reasoning in active science. My aim is to offer a theoretical perspective that can guide further empirical research on scientific reasoning, diagrams, and cognitive search. I begin by discussing how the notion of search has been employed in understanding reasoning in cognitive science, and how it can be applied to diagrams, before turning to discuss particular examples from chronobiology.

\section*{Diagrams and Cognitive Search}

One of the major challenges in science is the complexity of the reasoning processes that are required to grapple with natural phenomena. The notion of "search spaces" has been useful in trying to understand this challenge and how scientists proceed in meeting it. A "space," in this context, is simply a set of possibilities that are relevant to a reasoning task.

In a classic study, Klahr and Dunbar (Klahr \& Dunbar, 1988) asked subjects to discover the function of a particular command in a robot's programming language. There were two relevant spaces in the task: the "hypothesis" space of possible functions for the command, and the "experimental" space of possible manipulations to test given hypotheses. In the study, these spaces were artificially constrained via the experimenters-they chose what to tell subjects about the command, thus limiting the hypothesis space, and they set up the language of the robot, thus limiting what manipulations could be performed. They used this limited space to analyze subjects' reasoning, which allowed them to characterize the difficulty of the task and the (sometimes different) reasoning strategies that individuals used to solve it.

The search perspective has been used in a variety of other investigations into scientific reasoning (Schunn \& Klahr, 1996; Thagard, 1998), and experimental work in nonscientific contexts has begun to elucidate the cognitive and neural mechanisms that underlie search through the space of semantic memory (Hills, Todd, \& Goldstone, 2008). I here apply the search perspective as a way of understanding contributions of diagrams to reasoning in active scientific research.

While different numbers of search spaces have been proposed, for simplicity's sake I begin with two spaces, a conceptual space and an experimental space. The conceptual space consists of a scientist's or group of scientists' knowledge or beliefs about a system-including the particular entities that produce phenomena, their properties, and the kinds of interactions in which they can be involved. So, when a scientist approaches a phenomenon, they do so with an understanding of the
entities involved in producing that phenomenon. This knowledge provides a set of resources for reasoning about the system in question, which is continually modified and updated as investigation proceeds. In constructing and testing explanatory hypotheses, scientists consider this realm of possibilities in a flexible way-coming up with a good hypothesis involves "finding" the right system knowledge to account for the phenomenon of interest.

Experimental space consists of the possible manipulations that can be performed on the system in question given the practical strategies and limitations available to a field at a given time. This knowledge is often "embodied" in the sense that it involves practical know-how about successful manipulations, but it also can involve a theory of the instrument that licenses inferences to be drawn from particular results. To these I add a third, diagrammatic space, which plays the role of flexibly indexing and guiding search through the other two spaces. In employing the notion of search space, I make no claims about the format of the internal representations involved, or about the nature of the search algorithm. So long as such conceptual and experimental knowledge exists, my claim is that diagrams provide a way of indexing those bodies of knowledge.

One of the differences between reasoning tasks posed in psychology experiments and those undertaken by scientists "in the wild" is that in science the search task is often illdefined. That is, there are not clearly constrained solution options for a given reasoning task. Search through diagrammatic space, I propose, allows for flexible constraints on conceptual and experimental search, which allows both for productive investigation within specific models and continual questioning and reconceptualization of those models. Diagrams contain elements which provide directions of search through the diagram-arrows, enclosures, etc. This external search can then serve as a guide to the difficult work of employing one's conceptual space in reasoning about the system, and in using one's experimental space to devise tests of that reasoning. This external search space can be manipulated with relatively little cognitive demand. Moreover, diagrammatic space can be shared in common amongst individuals whose conceptual and experimental spaces differ, thus guiding a field's investigations into phenomena and mechanisms. I will discuss each of these points, with examples from mammalian chronobiology. My discussion will be illustrative. Importantly, I do not claim that all such reasoning must occur through diagrams. I only attempt to characterize the resources that diagrams can provide in active research, and I contend that this can help account for their importance and ubiquity in biology.

\section*{Mechanistic Diagrams in Biological Research}

As Klahr and Dunbar (1998) pointed out, one of the ways to constrain a space is to convey an abstract structure. Diagrams, given their particular elements and arrangements, do this exceedingly well. Consider Figure 1, a diagram
from a relatively early period in the history of mammalian circadian research. The diagram depicts a series of events


Figure 1: Three-stage model of the relationship between the endogenous clock and activity onset; from Welsh, Engle, Richardson, and Dement (1986).
that occurs at the beginning of an organism's "subjective day" (the part of the day during which the animal is active-dawn for diurnal organisms, dusk for nocturnal ones). The organism's internal clock functions to anticipate the external light schedule, and sends a signal to the mechanisms in the organism that govern waking and activity onset.

The diagrammatic space in the figure consists of the enclosed shapes and the arrows connecting them. Importantly, there are two distinct shapes, providing a visible, categorical distinction between the referent of the circle and those of the rectangles. The arrows imply an ordering of some sort between these referents, where this ordering has a directionality (i.e., it goes from the circle to the squares and not vice-versa) and is sensitive to the variables \(w\) and \(a\). This exhausts the purely visual set of constraints present in the diagram, which provide suggested patterns of search through the diagram. However, even these very minimal constraints manage to convey a great deal of abstract structure. It is abstract in that any entities and relations referred to must fit this pattern, if the diagram is taken as correct.

The connections to conceptual and experimental space are provided by the linguistic labels, as well as the instructions for how to interpret the figure. The denotation of the circle as a 'clock signal' indexes researchers' conceptual space regarding the nature of the clock and its relation to observable behavior. At this point, behavioral studies addressing rhythmicity had already established that the clock was endogenous-i.e., that it is an internal mechanism that can run without external input. Lesion studies had also suggested that the central clock in mammals has a particular brain locus in the suprachiasmatic nucleus (which was later conclusively confirmed), but little to nothing was known about the detailed mechanisms. Thus, the abstract model encoded in the diagram suggested that the 'clock signal', presumed to be coming from this central mechanism (whatever its detailed nature), must be related by an unknown process to the observable behavioral events under its control-in this case, waking and activity onset. The distinction between the circle and the rectangles denotes this categorical difference between the presumed mechanism and observable states. The caption, in addition, instructs that \(w\) and \(a\) should be interpreted as the time lag involved
in the transmission of the signal that cues wake processes, and the time lag between waking and activity onset, respectively.

Thus, the abstract structure conveyed in the diagram, along with its indexing of conceptual space and interpretational instructions, expresses a three-stage model of activity onset rhythms, in which a clock signal precedes waking, which precedes activity onset. As Welsh et al. stress, there is also an implied causal order in the diagramthat is, since the timing of activity onset is dependent on both the timing of the clock signal and that of the waking onset, it is suggested that the causal process leading to activity runs through these events.

In addition, the model's structure indexes the experimental space available to researchers. Two important experimental procedures in chronobiology are (i) the statistical analysis of variation in the phase of particular circadian events and (ii) external manipulation of the environmental factors involved in generating rhythms. Welsh et al. kept mice on a constant light schedule (ii), and analyzed the resulting phases of each stage (i) to test the model above. They discovered that there were no significant differences in phase between activity onsets over a period of days, while there were significant differences in wake onset.

Welsh et al. explicitly interpret this result with reference to the diagram in Figure 1. This result, initially, seems to be incompatible with the model, since variation across multiple time lags should produce greater variation at the end of the signaling chain. However, Welsh et al. argue that it is not incompatible, so long as the two signals are anti-correlated, with one becoming longer whenever the other is shorter. They then consider a number of possible mechanisms. The first proposes that a longer lag in \(w\) allows the organism to be better prepared to begin activity, and thus leads to shorter \(a\). The second, more radical proposal, argues that the timing of these events depends on a direct relationship between activity and the clock signal, with waking being indirectly regulated.

Interpreting these possibilities in light of the model further provides suggestions for experimental manipulation via method (ii)-namely, manipulating the availability of activity by controlling access to the running wheel. Several of the authors performed a separate study in which limiting wheel-running to specific times of day was shown to shift the phase of the central clock signal (Edgar, Martin, \& Dement, 1991). This in turn prompted the idea that there is a feedback signal from mechanisms controlling activity to the central clock, which was subsequently widely adopted. The picture that emerges is one in which continual consultation and interpretation of the figure allows for iterative episodes of indexing conceptual and experimental spaces, where the constraints present in the model guide subsequent investigation. Even if the model is eventually expanded (e.g., through the incorporation of a feedback
arrow running from the box on the right to the circle on the left) or overturned, this in no way lessens the potential importance of the figure for reasoning about the system in question. This analysis is at least broadly in tension with views of diagrams as conveyors or communicators of explanations-diagrams can aid cognitive search even if they are not taken as correct or complete explanations of phenomena. If this analysis is right, even extremely simple diagrams can be important reasoning tools. I now go on to discuss how this kind of analysis can be applied to more complex diagrams, which aid search in the construction of explanations involving complex mechanisms with many interacting parts.

\section*{Discovering Parts and Operations: The Function of CRY Proteins}

The central clock mechanism, in many organisms, consists of a "core" molecular clock, which operates via the interaction of multiple feedback processes. In this mechanism, the expression of a "positive" element causes, via DNA binding, the expression of a "negative" element, which in turn inhibits both its own transcription and that of the positive element. After the negative element is degraded, inhibition on the positive element is released and the cycle begins again-the period of the cycle is determined by the time course of the interactions between the relevant components. In the 1990s, several of the key genes involved in these feedback processes were discovered through research on fruit flies and rodents. Figure 2, from Dunlap (1999), represents the state of understanding of the mammalian core clock at the end of the decade.


Figure 2: Diagram of a model of the mammalian core clock, circa the late 1990s; from Dunlap (1999).

The positive elements of the core clock are the protein products of the genes Clock and Bmal, and the negative elements are transcripts of the various paralogs of the Period (Per) gene. The diagram depicts the positive element proteins binding to the promoter region (E-box) of the Per genes, whose transcribed mRNAs move outside of the nucleus and are translated into proteins. The proteins then form complexes of an unknown kind, and then inhibit
their own transcription by binding to the CLOCK::BMAL complex. External light, denoted with yellow arrows, is presumed to affect the clock by interacting with the negative loop.

Dunlap's diagram also included a question mark in the red box at the upper left, to indicate that there were likely other dimeric partners involved in the negative loop. At the same time, it was known that mammalian cells contained the gene Cryptochrome (Cry), a homolog of the Cry gene in drosophila, which serves as a photoreceptor for the light signaling pathway in flies. However, the function of Cry in the mammalian clock was unclear, as manipulation did not have any effect on entrainment to external light. Thus, it was left out of Dunlap's diagram. Further evidence against Cry as part of the entrainment mechanism appeared in the same year; Van der Horst et al. (1999) showed that individual knockouts of the Cryl and Cry2 paralogs had effects on circadian period, and that double knockouts eliminated rhythmicity completely. Kume et al. (1999), while citing Dunlap at several places, immediately proposed that CRY proteins might fit into the model at the point where, in the red box in the upper left, Dunlap had left a question mark-it might replace TIM as the proposed dimeric partner for PER. They performed a variety of manipulations on CRY, showing: (i) that CRY protein quantities are dependent on the functioning of Clock, (ii) that CRY can stop activation due to CLOCK::BMAL, and (iii) that CRY and PER dimerize and are transmitted to the nucleus together. It seems clear that Kume et al. were working from the model encoded in Dunlap's figure.

In this case, a change in conceptual space-the idea that Cry might be involved in the core clock-interacts with the model represented in the diagrammatic space to suggest experiments that can establish whether this new role is correct. Constraints for these experiments are present in the diagram-e.g., by guiding search for interactions with the proposed PER dimer-and its further indexing of the other elements of conceptual space. Once again we can see the potential cognitive benefits of encoding a model in diagrammatic space, and how this further relates to the interaction of conceptual and experimental space.

\section*{Diagrams for a Field: Coordinating Multiple Conceptual and Experimental Spaces}

In my final example, I show how diagrams can be shared search spaces for researchers with different theoretical and experimental backgrounds. The results of Kume et al., among others, are expressed in Figure 3, from Lowrey and Takahashi (2004). The basic organization from the Dunlap diagram is still present (although flipped left to right), and a few further elements have been added, including the additional support loop formed by the Rev-erb gene, which acts as a positive regulator of Bmal transcription. Ye, Selby, Ozturk, Annayev, and Sancar (2011) decided to test a core assumption of this model-that the PER::CRY dimer


Figure 3: Diagram of a model of the mammalian core clock, circa the mid-2000s; from Lowrey and Takahashi (2004).
inhibits transcription of Per and Cry by binding to the CLOCK::BMAL complex while it is on the E-box-via sophisticated biochemical analysis. They employed chromatin immunoprecipitation (ChIP), a technique that allows for the isolating of particular DNA/protein complexes in the nucleus, and the determination of what proteins bind to particular sections of DNA. Importantly, this technique allows for a different trajectory through experimental space than the techniques used to develop the core model, which did not allow for precise localization and analysis of binding within the nucleus. While ChIP had existed in other fields of biology, it began to be used frequently in chronobiology only in the mid-2000s.

Ye et al. found, contrary to the standard model, that only CRY, and not the PER::CRY dimer, bind to the CLOCK::BMAL dimer while it is on the E-box. Moreover, the presence of PER actually inhibits this process of binding. If correct, this forces a relatively major revision of the model. Ye et al. represent a possible revision in Figure 4. This re-coding has significant conceptual ramifications for anyone familiar with the core clock, as it forces revision of the standard assumption about the causal process in the negative loop.

As the diagram suggests, and Ye et al. elucidate, new functional posits are needed to understand the role of PER in the clock mechanism. What does the diagram contribute? First, it emphasizes the difference between the previous model and the current results. Second it provides a functional posit for PER, as being potentially involved inmodulating the Rev-erb loop. This diagram suggests a new course through experimental space-inquiry into the potential binding of PER to the elements of this loop at different times during the circadian day. Moreover, those familiar with the standard model must now adjust their representation of the place of PER in the core loop (this suggests "replacing the top part of Figure 3 with the type of representation given in Figure 4). Despite revisions such as this, the standard diagrams still provide structural and


Figure 4: Diagram of the results of Ye et al. (2011), suggesting a revised role for the PER protein in the core mammalian clock; from Ye et al. (2011).
functional indices that constrain conceptual search for new roles of PER, as we saw in the discussion of Figure 2. We can expect that, should these results gain widespread acceptance, future review articles will incorporate these changes to the standard model.

This example shows a broader role for diagrams than in the reasoning of individuals. The standard model is explicitly targeted by Ye et al. as the source of their analysis-and they cite the Lowrey and Takahashi paper, among others. This suggests that diagrams play the role of organizing different methodological approaches around the same phenomenon and proposed mechanism. As Ye et al. mention, a variety of methods were used in constructing the standard model; however, emerging methods of analyzing protein interactions using biochemistry have the potential to fill in gaps or question particular aspects of models that are standard in the field. Crucially, not all scientists studying a phenomenon possess the same methodological expertisethat is, their experimental spaces differ. Equally important, once the results regarding the PER::CRY dimer have been obtained, they are re-encoded into a diagrammatic form that is common to those across different experimental backgrounds.

Based on these analyses, I propose that a primary function of mechanistic diagrams is to provide an external search space that coordinates search in conceptual and experimental spaces, across both personal and interpersonal contexts.

\section*{Conclusion}

The ubiquity of diagrams, and their seemingly important resources for aiding reasoning about complex systems, resist the interpretation that they are eliminable-i.e., that all of the actual thinking done by scientists is purely internal. To treat diagrams as themselves sufficient to convey scientific theories or explanations, however, seems equally unrealistic, as it fails to account for the vast amounts of detailed conceptual knowledge and experimental expertise that individual scientists, as well as research fields, bring to understanding any particular diagram.

I have argued that a search perspective on diagrams can make sense of their role in active research, and used this perspective to construct a sense-making narrative of important epochs of research in mammalian chronobiology. While the success of the narrative is not proof of the theory, the view I have proposed has a number of potential benefits for experimental studies of scientific reasoning. First, it can relate the search for scientific solutions to the more general literature on cognitive search. Much progress has already been made in understanding how subjects search through a visible space in relation to a task. Do these principles carry over to search in scientific diagrams? Stieff, Hegarty, and Deslongchamps (2011) have conducted an eye tracking study showing that individuals' eye movement patterns while using multiple visualizations (a mechanism diagram, a graph, and an equation) in a problem solving task are related to their particular educational experience. Mechanistic diagrams, on the search perspective, are ripe for this kind of study.

Finally, the search perspective can aid discussions and experimentation on both scientific reasoning and diagrams in general. Diagrammatic form can be used in a variety of tasks, and can help model discovery situations when individuals' conceptual and experimental spaces are shaped by the experimental setup. As shown by Sheredos et al. (forthcoming), changes in diagrammatic form affect interpretation, and thus the experimental manipulation of diagrams across reasoning tasks can shed light on the nature of individuals' search strategies. Taking the search perspective on mechanistic diagrams, then, has promise for helping to overcome the difficult methodological gap between standard psychology experiments and creative scientific reasoning.

\section*{Acknowledgements}

This material is based upon work supported by the National Science Foundation under Grant No. 1127640. I would like to thank the members of the WORGODS (WORking Group on Diagrams in Science) research group for valuable feedback.

\section*{References}

Cheng, P. C. H. (1992). Diagrammatic reasoning in scientific discovery: Modelling Galileo's Kinematic Diagrams. AAAI Technical Report on Reasoning with Diagrammatic Representations (Report No. SS-92-02), 33-38.
Cheng, P. C. H. (2011). Probably Good Diagrams for Learning: Representational Epistemic Recodification of Probability Theory. Topics in Cognitive Science, 3(3), 475-498.
Dunlap, J. C. (1999). Molecular bases for circadian clocks review. Cell, 96, 271-290.
Edgar, D. M., Martin, C. E., \& Dement, W. C. (1991). Activity Feedback to the Mammalian Circadian

Pacemaker: Influence on Observed Measures of Rhythm Period Length. Journal of Biological Rhythms, 6(3), 185199.

Gooding, D. C. (2010). Visualizing Scientific Inference. Topics in Cognitive Science, 2(1), 15-35.
Hegarty, M. (2011). The Cognitive Science of Visual-Spatial Displays: Implications for Design. Topics in cognitive science, 3(3), 446-474.
Hills, T. T., Todd, P. M., \& Goldstone, R. L. (2008). Search in External and Internal Spaces Evidence for Generalized Cognitive Search Processes. Psychological Science, 19(8), 802-808.
Klahr, D., \& Dunbar, K. (1988). Dual space search during scientific reasoning. Cognitive science, 12(1), 1-48.
Kume, K., Zylka, M. J., Sriram, S., Shearman, L. P., Weaver, D. R., Jin, X., et al. (1999). mCRY1 and mCRY2 Are Essential Components of the Negative Limb of the Circadian Clock Feedback Loop. Cell, 98(2), 193-205.
Lowrey, P. L., \& Takahashi, J. S. (2004). Mammalian circadian biology: elucidating genome-wide levels of temporal organization. Annu. Rev. Genomics Hum. Genet., 5, 407-441.
Nersessian, N. J. (2008). Creating scientific concepts. Cambridge, MA: MIT Press.
Perini, L. (2005). Explanation in Two Dimensions: Diagrams and Biological Explanation. Biology and Philosophy, 20(2-3), 257-269.
Schunn, C. D., \& Klahr, D. (1996). The problem of problem spaces: When and how to go beyond a 2-space model of scientific discovery. Paper presented at the Proceedings of the 18th Annual Conference of the Cognitive Science Society.
Sheredos, B., Burnston, D., Abrahamsen, A., \& Bechtel, W. (forthcoming). Why do biologists use so many diagrams? Philosophy of Science.
Stieff, M., Hegarty, M., \& Deslongchamps, G. (2011). Identifying representational competence with multirepresentational displays. Cognition and Instruction, 29(1), 123-145.
Thagard, P. (1998). Ulcers and bacteria I: Discovery and acceptance. Studies in the History and Philosophy of Biology and Biomedical Science, 9(107-136).
Tversky, B. (2011). Visualizing Thought. Topics in Cognitive Science, 3(3), 499-535.
van der Horst, G. T., Muijtjens, M., Kobayashi, K., Takano, R., Kanno, S., Takao, M., et al. (1999). Mammalian Cry1 and Cry2 are essential for maintenance of circadian rhythms. Nature, 398(6728), 627-630.
Welsh, D. K., Engle, E. M. R. A., Richardson, G. S., \& Dement, W. C. (1986). Precision of circadian wake and activity onset timing in the mouse. Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology, 158(6), 827-834.
Ye, R., Selby, C. P., Ozturk, N., Annayev, Y., \& Sancar, A. (2011). Biochemical analysis of the canonical model for the mammalian circadian clock. The Journal of biological chemistry, 286(29), 25891-25902.

\title{
Individual and Strategy Differences in an Allocentric-Heading Recall Task
}

\author{
Heather Burte (burte@psych.ucsb.edu) \\ Department of Psychological \& Brain Sciences, UCSB Santa Barbara, CA 93106 USA
}

\author{
Mary Hegarty (hegarty@psych.ucsb.edu) \\ Department of Psychological \& Brain Sciences, UCSB \\ Santa Barbara, CA 93106 USA
}

\begin{abstract}
Participants estimated allocentric headings using pictures of familiar buildings around a college campus, in an allocentricheading recall task. A weak relationship between sense-ofdirection and accuracy, an alignment effect, and a novel relationship between strategy and accuracy were found. These results demonstrate that sense-of-direction and strategy use differentially affect accuracy across heading disparities. Our findings suggest that individual differences and strategy differences need to be incorporated into current hypotheses regarding allocentric-heading - specifically, into the animalmodel hypothesis.
\end{abstract}

Keywords: allocentric heading; sense of direction; headingrecall; strategy; egocentric and allocentric reference frames.

\section*{Introduction}

People characterize their ability to move effectively through environmental-scale spaces, such as neighborhoods or cities, by referring to their 'sense-of-direction'. Kozlowski and Bryant (1977) found that people's ratings of their sense-ofdirection (or SOD) correlated with accuracy in distance, direction, and time estimation tasks. Since then, research has either focused on how to assess SOD or how SOD correlates with performance in environmental-scale spatial tasks (e.g., Hegarty et al., 2002). However, the field is lacking insights into the underlying causes of the vast individual differences found in environmental-scale spatial cognition.

Research on SOD has typically focused on strategy differences in navigation or on individual differences in learning novel environments. In terms of strategy differences, individuals with a poor SOD tend to prefer route strategies and those with a good SOD tend to prefer survey strategies (Prestopnik \& Roskos-Ewoldsen, 2000). In terms of individual differences, there are large individual differences in the rates and accuracy with which individuals can learn novel environments and these differences are related to self-reported SOD (Ishikawa \& Montello, 2006).

To date, little research has focused on strategy and individual differences in manipulating one's knowledge of familiar environments; however, manipulating one's spatial knowledge is essential for wayfinding and route planning. Research that has used familiar environments tends to focus on two tasks: egocentric pointing (pointing from one's current location towards a landmark) and judgments of relative direction (pointing from an imagined location and orientation towards another landmark) (e.g. Kozlowski \&

Bryant, 1977; Hegarty et al., 2002). These studies found significant correlations between SOD and task performance, but have not investigated strategy differences.

The strategies identified during navigation, namely route and survey strategies, are not necessarily relevant in all spatial knowledge manipulation tasks (such as egocentric pointing and judgments of relative direction tasks). Therefore, research is needed to uncover the strategies used in spatial knowledge manipulation tasks - specifically, in environmental-scale spaces. In contrast, research on strategy differences has tended to utilize small-scale tasks. Kozhevnikov and Hegarty (2001) found that when people complete judgments of relative direction tasks, while viewing a map, they use either perspective-taking or mental rotation strategies. They found individual differences in performance, which were related to SOD, and they identified two strategies, which were separable abilities.

The goal of this paper is to investigate the individual and strategy differences that exist within manipulating one's environmental-scale spatial knowledge. To do so, we will focus on one task, the allocentric-heading recall task (Sholl, Kenny, \& DellaPorta, 2006), which requires participants to manipulate their spatial knowledge of a familiar place.

Sholl et al. (2006) developed the allocentric-heading recall task to reveal the architecture of a proposed human head-direction system. They argued that SOD is a singlefaceted construct related to the performance of a headdirection system in humans, and they assumed that this system operated similarly to the head-direction system found in rats (Ranck, 1984). In rats, each head-direction cell fires maximally to one angle of difference between the rat's facing direction and a reference direction grounded in the environment. In other words, head-direction cells respond to allocentric headings and not directions based on the axis of the body (or egocentric headings). Sholl et al. proposed that the human head-direction system operates similarly to that of rats and that the human head-direction system is the neural mechanism underlying self-reported SOD. We will refer to this the animal-model hypothesis.

In the allocentric-heading recall task, participants view photographs of a familiar environment, identify the direction from which the photographs were taken, and then rotate in their chair to reproduce the direction. Initial studies (Sholl et al., 2006) revealed an alignment effect. Specifically, when a participant is facing the same direction as that from which the photograph was taken, participants
were more accurate (a facilitation effect). However, when participants were \(180^{\circ}\) misaligned from the direction of the photograph (for example, they faced north but the photograph was taken from a south-facing direction), participants were the least accurate (a detrimental effect). The alignment effect was explained as interference between one's current head-direction signals and the retrieval of the head-direction signals, which were activated when the individual viewed the photograph location in the real world. Strong correlations (. 7 or higher) between performance on the allocentric-heading task and self-assessed ratings of SOD supported Sholl et al.'s proposal that SOD solely reflects the operation of the human head-direction system.

The allocentric-heading recall task assesses people's ability to manipulate their spatial knowledge of environmental-scale spaces. According to Sholl et al., when viewing a building, the allocentric-heading of that view is stored in memory and is linked to signals of body-direction. Upon seeing a picture of that building, a person recognizes the building, and then recalls the allocentric-heading from spatial memory. Therefore, Sholl et al. proposed that only one strategy exists, and individual differences reflect differences in the fidelity of head-direction signals, and consequently, the ability to carry out this strategy.

In contrast to this view, Burte and Hegarty (2012) found preliminary evidence for possible strategy differences. During informal debriefing interviews, participants reported a range of strategies, including imaging a walk to the photograph location, and relating the photograph heading to the direction of a local mountain range. However, strategy differences have yet to be systematically investigated in this task.

There is also a possibility that familiarity with the tested environment drives individual differences in performance. Sholl et al. did not investigate familiarity differences, as pretesting had revealed that all their pictures were highly familiar to participants. However, Burte and Hegarty (2012) found significant correlations between familiarity and SOD, as well as between familiarity and accuracy, despite pretesting photos for high familiarity. This suggests that individual differences in this task might be partially due to differences in familiarity with the environment.

The main purpose of this study is to investigate individual and strategy differences within the allocentric-heading recall task. First we will describe the allocentric-heading task in more detail and then consider strategy differences found within a similar task.

\section*{Allocentric-Heading Recall Task}

The allocentric-heading recall task is a four-alternative, forced-choice task, using campus pictures as stimuli. Pictures were taken from magnetic north, east, south or west (to match the intrinsic structure of the environment). However, while cardinal directions will be used for simplicity in writing this article, it should be noted that cardinal directions were never used in written or verbal instructions, as they are not required to complete the task.

First, we will define key terminology used: picture heading is the photographer's orientation when taking the picture; default heading is the orientation of participant before each trial; response heading is the orientation the participant responded with; and heading disparity is the angular disparity between default heading and picture heading. The animal-model hypothesis makes predictions about the relationship between heading disparities and performance - specifically, about the alignment effect; therefore, heading disparity is the main independent measure of interest.

Turning in one's chair (to replicate the picture heading) was used as the response mode in previous studies. Sholl et al. (2006) argued that turning to represent an angle was a natural response for this task, because turning allows for one's current head-direction cells to find a match to the memory of one's head-direction cell firing. A secondary goal of this study was to investigate whether participants could perform this task without rotating in a chair, but only by using a button-press as the response mode. Therefore, we attempted to replicate Sholl et al.'s results with an alternative, less body-based response mode.

\section*{Strategy Differences}

In studies using a judgments of relative direction task (JRDs), a dissociation has been made between two strategies: (1) a perspective-taking strategy whereby participants imagined moving themselves to assume a new orientation, or used directions related to their bodies to assume a new orientation; and (2) a mental rotation strategy whereby participants imagined moving the entire scene around themselves, or imagined rotating angles between locations (Kozhevnikov \& Hegarty, 2001). This suggests that participants can think in terms of a body-based reference frame (egocentric), or a reference frame grounded in the environment (allocentric) while completing the task.

Kozhevnikov, Motes, Rasch, and Blajenkova (2006) found that the perspective-taking strategy resulted in decreased accuracy with increasing heading disparities, a similar pattern to that found by Sholl et al. (2006). However, use of a mental rotation strategy resulted in a significantly weaker alignment effect. Therefore, another goal of this study is to investigate if these strategy differences exist within the allocentric-heading recall task.

In sum, our goals are (1) to investigate if the predictions of the animal-model hypothesis are robust to a new context and to a button-press response mode; (2) to investigate our prediction that individual differences in familiarity are related to task performance; (3) to determine if egocentric and allocentric strategy use exist within this task; (4) to investigate if individual differences in SOD and strategy are related to task performance; and (5) to investigate whether strategy differences are related to SOD.

\section*{Method}

Participants Seventy-four students (39 males and 35 females) participated as part of a research requirement. Two
participants, both males, were excluded from analysis because their mean familiarity with the picture stimuli was 2 \(S D\) s below that of all participants. Participants had spent at least two quarters on campus before participating.

Design The methodology of the study was both experimental and correlational. The experimental factors were picture heading (within subjects) and default heading (between subjects). The correlational factors are familiarity, SOD, and strategy use. Participants were randomly assigned to one of the four default headings (19 participants faced north, 18 east, 17 south, and 18 west) and completed fortyfive trials, one for each picture.

Materials The experiment took place in a room that was aligned with the main axes of the campus (and the cardinal directions). The experimental room had one east-facing window that was open during the experiment. The view directly out that window was of a major pathway and a large (eight storey) building. However, if one stood next to the window, one could see the mountains and ocean (major orientation markers for the campus), and a few major buildings. Therefore, the window afforded excellent views for initial orientation to the campus (when standing near the window), but only basic information while participants completed the experiment.

Experimenters arranged a chair and laptop facing the assigned default heading before each participant arrived. The large table at which participants were seated (but in different orientations) was aligned with the room, the room was aligned with campus, and campus is aligned with the cardinal directions. Therefore, the space was aligned with respect to the default headings and response headings. This alignment was never mentioned to participants.

The photographic stimuli were sourced from the 36 most familiar photographs from a previous experiment (Burte \& Hegarty, 2012), and nine new photographs (two north and seven east), for a total of 45 pictures. A global positioning device (GPS) was used to ensure that photographs were taken facing the cardinal directions. Photographs were taken of highly recognizable building facades and were cropped to exclude surrounding buildings or large-scale landmarks.

A typical trial started with viewing a photograph of campus on a computer, and participants responded by using a keypad with four arrows (front, right, back, and left). The participant determined the direction (with respect to the campus environment) in which the photographer stood to take the photograph (i.e., picture heading) and pressed a button to reproduce that direction. For example, if the photograph was taken facing south, and the participant was facing north, then the participant should press the downward arrow to indicate the direction behind him/her.

Procedure Participants were briefly introduced to the experiment, completed a demographics questionnaire, and then completed the Santa Barbara Sense of Direction (or SBSOD) scale (Hegarty et al., 2002). Next, participants
were asked to orient to the layout of campus while looking out the window. The experimenter asked the participant to point towards six major campus landmarks, to ensure that \(\mathrm{s} / \mathrm{he}\) was oriented to the global layout of the campus. The experimenter provided feedback, if needed, but most participants oriented and pointed correctly.

Participants were then introduced to the allocentricheading recall task and presented with 12 practice trials in a fixed order. Participants were given feedback by being presented with the correct answer after each practice trial, and then completed 45 experimental trials without feedback.

After the allocentric-heading recall task, participants rated their familiarity with each photograph location on a 7-point Likert scale, with 1 being "Very familiar" and 7 being "Not at all familiar". As an objective measure of familiarity, participants were required to place an arrow on an unlabelled map of campus, to indicate the location and direction from which each photograph was taken. This map task and the familiarity task were used to ensure high familiarity with each photograph.

Finally, participants completed a strategy questionnaire that consisted of a free-response question, in which they entered the strategies they used, and then selected the strategies they used from a list of potential strategies. The list was created based on pre-testing and consisted of strategies such as "Using cardinal directions", "Using largescale landmarks to determine orientation (mountains, ocean, Isla Vista, etc.)", and "Imagining travelling to the location".

\section*{Results}

Photograph Familiarity To ensure that participants were sufficiently familiar with the pictures, pictures needed to pass three criteria to be included in the analysis: (1) mean familiarity for each picture could not be \(2 S D\) s lower than grand mean familiarity, (2) less than \(25 \%\) of participants needed to rate their familiarity as " 6 " or " 7 - Not at all familiar" for each picture; and, (3) at least \(25 \%\) of participants needed to correctly identify the orientation and location of the picture, on the map task. Given these criteria, seven photographs were dropped from analysis, resulting in 9 north-facing, 9 east-facing, 10 south-facing, and 10 westfacing pictures. The familiarity grand mean for the 38 photographs was 2.2. Familiarity ranged from 1.0 to 3.7 across participants and from 1.2 to 3.9 across pictures.

Accuracy Heading disparity (angular difference between default and picture heading) served as the main independent measure. For example, if the picture heading was aligned with the default heading for a particular participant and trial, then this trial would be labeled as having a \(0^{\circ}\) heading disparity. A 4 (Heading disparity: \(0^{\circ}, 90^{\circ}, 180^{\circ}, 270^{\circ}\) ) by 2 (Gender) ANOVA comparing mean accuracy indicated a main effect of heading disparity, \(F(3,216)=6.79, M S E=\) \(.13, p<.001\). The mean accuracy by heading disparity is shown in Figure 1. Post hoc tests revealed that the \(180^{\circ}\) condition was less accurate ( \(M=54 \%\) ) than all other conditions, which had similar accuracies ( \(0^{\circ} M=61 \%\); \(90^{\circ}\)
\(M=64 \% ; 270^{\circ} M=63 \%\) ). This can be interpreted as a detrimental effect on performance when one's body is positioned \(180^{\circ}\) away from the memory trace from one's head-direction cells when the location was last viewed. This detrimental effect is predicted by the animal-model hypothesis; however, we failed to replicate the predicted facilitation effect.

The main effect of gender was also significant, \(F(1,70)=\) 6.58, \(M S E=1.32, p<.05\), with males being more accurate ( \(M=67 \%\) ) than females ( \(M=54 \%\) ). The interaction of heading disparity and gender was not significant, \(F(3,210)\) \(=1.07, M S E=.02, p=.36\).


Figure 1: Mean accuracy rate as a function of heading disparity. Error bars are the standard errors of the mean.

Self-Reported Sense-of-Direction The correlation between SBSOD scores and overall accuracy was statistically significant, \(r(70)=.31, p<.01\), indicating that people who rated themselves as having a good SOD were more accurate on the task. This correlation is similar to that found by Burte and Hegarty (2012); but substantially lower than those reported by Sholl et al. (2006). In addition, we failed to find a significant correlation between familiarity and SBSOD scores, \(r(70)=.03, p=.81\), indicating that good SOD participants were not more accurate simply due to being more familiar with the photographs.

To further investigate individual differences in task performance, we compared the performance of good SOD (or GSOD) participants from the top \(25 \%\) of the SBSOD distribution ( \(N=18\) ), and poor SOD (or PSOD) from the bottom \(25 \%(N=19)\). A 2 (GSOD, PSOD) X 4 (Heading disparity: \(0^{\circ}, 90^{\circ}, 180^{\circ}, 270^{\circ}\) ) ANOVA comparing mean accuracy indicated significant main effects and a significant interaction. There was a main effect of heading disparity, \(F(3,105)=3.60, M S E=.08, p<.05\), such that a heading disparity of \(180^{\circ}\) resulted in lower accuracy ( \(M=54 \%\) ) compared to \(90^{\circ}(M=63 \%)\) and \(270^{\circ}(M=64 \%)\). The heading disparity of \(0^{\circ}\) ( \(M=60 \%\) ) was not significantly different from other headings.

As shown in Figure 2, GSOD participants were significantly more accurate ( \(M=71 \%\) ) than PSOD
participants \((M=49 \%), F(1,35)=7.82, M S E=1.78, p<\) .01, and there was a significant interaction of SOD with heading disparity, \(F(3,105)=2.60, M S E=.05, p<.05\). Importantly, the simple effect of heading disparity for GSOD participants was not significant, \(F(3,33)=1.13, p=\) .35, indicating that GSOD participants were equally accurate across all heading disparities. This is a novel finding and has not been found in previous studies (Sholl et al., 2006; Burte \& Hegarty, 2012).

In contrast, the simple effect for PSOD participants indicated a significant difference across heading disparities, \(F(3,33)=4.11, p<.05\). Not only are PSOD individuals less accurate on this task than GSOD participants, but they are significantly less accurate with \(180^{\circ}\) heading disparities compared to other disparities. This indicates that the detrimental effect of having one's body \(180^{\circ}\) misaligned with the picture, primarily affects PSOD individuals.


Figure 2: Mean accuracy rate as a function of heading disparity and SOD. Error bars are the standard errors of the mean.

Strategy Use To examine reported strategy differences, items in the strategy questionnaire were classified as egocentric or allocentric strategies (cf. Kozhevnikov \& Hegarty, 2001). Example items from the egocentric strategy were "imagining myself standing at the photograph location", "imagining traveling to the photograph location using campus walkways", "comparing my current facing direction to the photographer's facing direction at the photograph location", etc. These strategies were labeled as 'egocentric' due to their reliance on thinking about directions in relationship to the participant's body. Strategies that focused on thinking about directions in relationship to external frames of reference were labeled as 'allocentric'. Examples of these items are "using a mental map or imaging a campus map", "using cardinal directions", "using large-scale landmarks", etc.

Participants were classified into strategy groups by calculating z-scores to reflect each participant's tendency to use each strategy compared to that of the entire group. For each participant, the egocentric and allocentric z-scores
were compared and if the two scores differed by more than . 75 SDs, the participant was deemed to have used one strategy more than the other. If the z-scores did not differ by . 75 SDs, the participant was classified as using a mixed strategy. This resulted in 33 participants who used a mixed strategy, 15 who used an egocentric strategy, and 24 who used an allocentric strategy.

Strategy Use and Sense-of-Direction To test the relationship between SOD and strategy, we compared the SBSOD scores of those classified as using egocentric and allocentric strategies. Egocentric strategy use corresponded with lower (or poorer) SBSOD scores ( \(M=3.7, S E M=.2\) ) and allocentric strategy use corresponded with higher (or better) SBSOD scores \((M=4.5, S E M=.2)\), and this difference was statistically significant, \(t(37)=-2.24, p<.05\).

Looking at strategy use across PSOD and GSOD individuals, we see that PSOD individuals used egocentric ( \(N=8\) ), allocentric ( \(N=6\) ), and mixed strategies \((N=5\) ). However, GSOD individuals only reported using allocentric strategies \((N=8)\) and mixed strategies \((N=10)\). A chisquared test revealed a significant relationship between SOD and strategy, \(X^{2}(2, N=37)=9.93, p<.01\) GSOD individuals were less likely to use egocentric strategies than predicted by chance.


Figure 3: Mean accuracy rate as a function of heading disparity and strategy. Error bars are the standard errors of the mean.

Strategy Use and Accuracy A 2 (Strategy: egocentric or allocentric) X 4 (Heading disparity: \(0^{\circ}, 90^{\circ}, 180^{\circ}, 270^{\circ}\) ) ANOVA indicated a significant main effect of heading disparity, \(F(3,111)=6.07, M S E=.12, p<.001\), a significant main effect of strategy, \(F(1,37)=9.50, M S E=\) 1.72, \(p<.01\), and a non-significant interaction, \(F(3,111)=\) \(1.27, M S E=.03, p=.29\). Figure 3 shows the mean accuracy by heading disparity. Participants who tended to use allocentric strategies were significantly more accurate ( \(M=\) \(72 \%\) ) than those who tended to use egocentric strategies ( \(M\) \(=50 \%\) ). While Figure 3 shows a trend for allocentric strategy users to show a weaker alignment effect, this trend
was not statistically significant. The finding that strategy use impacts the accuracy with which participants respond to the allocentric-heading task is a novel finding and is not predicted by the animal-model hypothesis.

Photograph Familiarity and Accuracy Correlations between participants' mean familiarity rating (averaged over the 38 pictures) and their mean accuracy on the headingrecall task were not significant, \(r(70)=.03, p=.81\). This indicates that participants, who rated their familiarity as high, were not more accurate than participants with lower familiarity. One interpretation is that all participants had a level of familiarity high enough, as to not hinder their task performance. However, correlating mean familiarity per picture (averaged over individuals) with mean accuracy per picture resulted in a significant correlation, \(r(36)=-.49, p<\) .01. This suggests that despite pretesting for familiarity, some familiarity differences remained between the pictures. Importantly, as default heading is manipulated between participants, differences in picture familiarity cannot account for the effects of heading disparity on performance.

\section*{Discussion}

We replicated findings that individuals can recall allocentric-directional information from pictures, and that individual performance in the allocentric-heading recall task is related to SOD (Sholl et al., 2006; Burte \& Hegarty, 2012). We also showed that these results replicate with a button-press response rather than the more body-based response of turning in one's chair. Importantly, we provided evidence for the use two strategies in this task, and showed that strategy use was related to self-reported SOD. GSOD participants reported using allocentric or mixed strategies, compared to PSOD participants were equally divided across strategy groups. Furthermore these groups had very different patterns of performance; PSOD individuals showed an alignment effect while GSOD did not. This pattern suggests that allocentric strategy use resulted in better performance, in general, and the alignment effect primarily affects PSOD individuals.

Changing the response mode, from turning in a chair to pressing a button, led to a weakened relationship between heading disparity and accuracy relative to previous studies. Specifically, the facilitation effect at \(0^{\circ}\) was not found. Another weakened relationship was the correlation of SOD with accuracy. Our correlations were noticeably lower than those found by Sholl et al (2006); therefore, this experiment adds doubt to the conclusion that that SOD solely reflects the operation of the human head-direction system. Instead, self-reported SOD might also relate to strategy differences.

Egocentric strategy use resulted in decreased accuracy in general and decreased accuracy for larger heading disparities (an alignment effect). Allocentric strategy use resulted in somewhat more equivalent accuracy across heading disparities. Although the interaction of strategy and heading disparity was not significant in the present study, the trends are notable in that they are similar to trends found
by Kozhevnikov et al. (2006) using JRD tasks.
Although we found both strategy differences and performance differences between those with good and poor sense-of-direction, the relationship between SOD, strategy, and performance remains ambiguous. Participants who used the egocentric strategy were more likely to have a lower (or poor) SOD, and those who used the allocentric strategy were more likely to have a higher (or good) SOD. Good SOD participants were also less likely to use egocentric strategies, than would be predicted by chance. But the causal relationships between strategy differences, individual differences and performance are currently unclear, as having a good SOD could cause people to use the allocentric strategy or repeated use of the allocentric strategy could contribute to having a good SOD. We are investigating the causal relationship between strategy use, SOD, and performance in a current study.

The newly discovered strategy differences between good SOD participants and poor SOD participants might have been due to the replacement of the body-based response with the button-press response. It is possible that the response of turning in one's chair in previous studies forced participants to use an egocentric strategy, which resulted in the alignment effect for participants of all ability levels. Perhaps, pressing a button did not force participants into using an egocentric strategy, so good SOD participants were freed from this restriction to think in terms of their body. This allowed good SOD participants to demonstrate similar performance across differing default headings. While our findings cannot provide support for these ideas, our findings do suggest that individual and strategy differences need to be incorporated into accounts of the performance within the allocentric-heading task. Research is also needed to determine if the response mode change was responsible for the identification of strategy differences.
We found a more nuanced relationship between familiarity and performance than in our earlier study (Burte \& Hegarty, 2012). Specifically, familiarity was correlated with performance, but only when compared across pictures, and overall familiarity with the pictures was unrelated to SOD. It seems that our goal to use only familiar photographs was achieved, as accuracy was not correlated with mean familiarity for all pictures; however, the familiarity rating of individual photographs still impacted accuracy. Since recognition of a location is likely the first step in completing the allocentric-heading task, it follows that familiarity on a picture-by-picture basis would affect accuracy.

Another novel finding of this study is that we found gender differences in task performance. It is possible that females are more tied to their bodies than males, which leads to a greater gender difference with a button-press response than a more body-based response (i.e., turning). Future experiments should continue to monitor gender differences in this task.
In conclusion, we found novel evidence for strategy differences in the allocentric-heading recall task and these
differences are related to level of performance and selfreported sense-of-direction. Given similar findings in navigational tasks, we propose that choice of strategy is a critical element to the understanding of individual differences within spatial tasks. Specifically, this study demonstrates that individual and strategy differences can be found within tasks that are often conceptualized as universal or invariant cognitive processes. Neuroscientific research using animals has provided the foundation for understanding the functional architecture of human spatial abilities. Now, there is a need to incorporate the unique aspects of human cognition - like strategy and individual differences - into the functional architecture.

\section*{Acknowledgments}

This research was supported by the Institute for Collaborative Biotechnologies.

\section*{References}

Burte, H., \& Hegarty, M. (2012). Revisiting the Relationship between Allocentric-Heading Recall and Self-Reported Sense of Direction. In N. Miyake, D. Peebles, \& R. P. Cooper (Eds.), Proceedings of the 34th Annual Conference of the Cognitive Science Society (pp. 162-167). Austin, TX: Cognitive Science Society.
Hegarty, M., Richardson, A. E., Montello, D. R., Lovelace, K., \& Subbiah, I. (2002). Development of a self-report measure of environmental spatial ability. Intelligence, 30, 425-447.
Ishikawa, T., \& Montello, D. R. (2006). Spatial knowledge acquisition from direct experience in the environment: Individual differences in the development of metric knowledge and the integration of separately learned places. Cognitive psychology, 52(2), 93-129.
Kozhevnikov, M., \& Hegarty, M. (2001). A dissociation between object manipulation spatial ability and spatial orientation ability. Memory \& Cognition, 29(5), 745-756.
Kozhevnikov, M., Motes, M. A., Rasch, B., \& Blajenkova, O. (2006). Perspective-taking vs. mental rotation transformations and how they predict spatial navigation performance. Applied Cognitive Psychology, 20(3), 397417.

Kozlowski, L. T., \& Bryant, K. J. (1977). Sense of direction, spatial orientation, and cognitive maps. Journal of Experimental Psychology: Human Perception and Performance, 3, 590-598.
Prestopnik, J. L., \& Roskos-Ewoldsen, B. (2000). The relations among wayfinding strategy use, sense of direction, sex, familiarity, and wayfinding ability. Journal of Environmental Psychology, 20(2), 177-191.
Ranck, J. B., Jr. (1984). Head-direction cells in the deep cell layers of dorsal presubiculum in freely moving rats. Society of Neuroscience Abstracts, 10, 599.
Sholl, M. J., Kenny, R. J., \& DellaPorta, K. A. (2006). Allocentric-heading recall and its relation to self-reported sense-of-direction. Journal of Experimental Psychology: Learning, Memory, and Cognition, 32(3), 516-533.

\title{
Explicit awareness supports conditional visual search in the retrieval guidance paradigm
}

\author{
Daniel R. Buttaccio (buttacciodr@ou.edu) \(\ddagger\) \\ Nicholas D. Lange (ndlange@gmail.com) \(\dagger\). \\ Sowon Hahn (sowon@ou.edu) \(\ddagger\) \\ Rick P. Thomas (Rickey.P.Thomas-1@ou.edu) \(\ddagger\) \\ Eddy J. Davelaar (e.davelaar@bbk.ac.uk) \(\dagger\) \\ \(\ddagger\) University of Oklahoma, Department of Psychology \\ Lindsey Street, Norman, OK, USA 73019 \\ \(\dagger\) Birkbeck, University of London, Department of Psychological Sciences, Malet Street, London, UK, WC1E 7HX
}

\begin{abstract}
In two experiments we explored whether participants would be able to use probabilistic cues to simplify perceptually demanding visual search in a task we call the retrieval guidance paradigm. On each trial a background cue appeared prior to (and during) the search task and the diagnosticity of the background cue(s) was manipulated to provide complete, partial, or non-diagnostic information regarding the target's color on each trial. Only when participants were made aware of the possible relationship between the background cues and target features were they able to utilize the cue information for search. When participants were not made aware of the possible connection, they were only able to use target base rates. In the General Discussion we address how a recent computational model of hypothesis generation (HyGene, Thomas, et al., 2008), provides a useful framework for understanding how long-term memory, working memory, and attention coordinate in visual search.
\end{abstract}

Keywords: attention, memory, visual search, hypothesis generation.

In the present research we examined whether participants would be able to use experience in order to reduce the perceptual demands of visual search. More specifically, we ask whether participants would be able to use cues to retrieve associated target features in service of visual search. We argue that much of our day to day visual search relies on such long-term memory (LTM) retrieval to define an attentional set to support search. To investigate the processes unfolding in such circumstances we have developed a novel visual search paradigm in which participants are provided with cues that probabilistically predict a target feature (its color) in a forthcoming search array. We refer to this procedure as the retrieval guidance paradigm as retrieval of likely target colors given a cue will drastically improve search.

The usefulness of the paradigm lies partly in the ecologically relevant variables it affords control over. Importantly, it allows us to assess people's sensitivity to the probabilistic relationships between cues and targets through 1) the global base rate of a target (raw frequency of occurrence) and an individual cue's diagnosticity (i.e., its predictive ability). Both of these characteristics influence the posterior probability and thus should influence retrieval
from LTM.
Related research examining how people use systematic cue-target associations to support visual search has largely focused on the contextual cueing paradigm. The general trend in these experiments is that targets within repeated visual scenes (having the same target to distracter spatial configurations) are found faster than non-repeated scenes (Chun \& Jiang, 1998). One of the most intriguing aspects of the results emanating from this paradigm is that the facilitation of repeated scenes operates at an implicit level of awareness as participants are unable to distinguish between repeated scenes versus novel scenes in a recognition task. Much of this research has focused on comparing conditions where cues perfectly predict the location of the target and conditions in which cues are non-diagnostic. However, many of the environments that we encounter are probabilistic such that cues have varying degrees of diagnosticity regarding possible target characteristics. Moreover, recent research has suggested that there is minimal attentional guidance in contextual cueing (Kunar, Flusberg, Horowitz, \& Wolfe, 2007), suggesting that the paradigm may be lacking in allowing for the examination of how LTM, working memory (WM) and attention coordinate in a visual search task. In the present study we offer a new paradigm that allows for the promise of: 1) investigation into whether probabilistic relationships can be learned and exploited in a visual search task (c.f. contextual cueing, Zellin, Conci, Muhlenen, \& Muller, 2011), and 2) an examination of how LTM, WM, and attention coordinate in a visual search task. We accomplish this by manipulating both target base rates and cue diagnosticities within the same experimental paradigm (and within subject) by pairing background cues (preceding the onset of a search array) with critical target features (i.e., colors). Thus, the background cues provide complete, partial, or nondiagnostic information regarding the color of the target in the upcoming search array.

To foreshadow the findings of the present study, in Experiment 1 we find that participants are able to find a target faster when its color is associated with a more diagnostic cue. However, Experiment 2 reveals that when participants are explicitly aware of the associations between the cues and the color of the target, their visual search
performance is significantly improved (over the participants of Experiment 1). This difference is explained by participants reliance solely on target base rate or retrieval from LTM to guide search processes.

\section*{Experiment 1}

Experiment 1 was conducted to assess the degree to which participants would use cue information to simplify a difficult visual search task. Participants were asked to respond to a specific orientation of a "T" (rotated 90 degree clockwise or 90 degrees counterclockwise), while ignoring modified "L" letters in a visual array. Each visual array contained 14 different items (13 distracters and 1 target), with each item being unique in color. Note that there was always a target present in each search array and the search array was presented until participants responded (i.e., there was no time limit).

On each trial, a background cue appeared prior to (and during) the search array. The background cues consisted either of circles, squares, or triangles and were positioned randomly on the screen (i.e., one cue consisted of only circles, another only squares, and the last only triangles). The statistical relationship between the background cues and the identity of the upcoming target was manipulated in order to provide complete, partial, or non-diagnostic information regarding the color of the forthcoming target. Table 1 provides the contingency table describing how the backgrounds were paired with the different colors (see Figure 1 for an example background). For example, background 2 was paired with colors 2 and 3 (C2 and C3) such that when background 2 was presented the target's color was C2 on \(60 \%\) of trials and C3 on \(40 \%\) of trials (note that although C1 and C4-C14 were always present in the array when background 2 was presented, they were always "L"s in the search array). Each participant was exposed to each of the different background cues throughout the entire experiment (i.e., a within-subjects design was used). Participants went through 360 trials (i.e., 6 Epochs of 60 trials each. Within each epoch the 3 backgrounds were presented 20 times each and were randomly selected for each trial. For each Epoch C1 was the target 20 times given background 1, C2 was the target 12 times given background 2, C3 was the target 8 times given background 2, and each of the fourteen colors (C1-C14) appeared as the target roughly 1.43 times given background 3 . Thus, the base rate (raw frequency) of each color appearing as the target per epoch was approximately 21.43 for C1, 13.43 for C2, 9.43 for C3, and 1.43 for C4-C14.

Participants were not informed at the beginning of the experiment that a statistical relationship existed between particular backgrounds cues and the color of the target. Because of this, at the conclusion of the experiment they were asked whether they noticed a statistical relationship between particular backgrounds and the target colors. After providing an answer, participants performed a recognition task. For each trial during the recognition task, participants were provided with a search array that contained a target of
a particular color. The color of the target was either valid (the target was paired with that background in the experiment) or invalid. Participants performed the recognition task for the three different backgrounds. The participant's recognition performance was used as a measure of their explicit knowledge concerning the cuehypothesis contingencies.
Table 1: Contingency table, showing how the backgrounds were paired with the different colors (C1-C14) for
Experiments 1 and 2. Note that the last column indicates the contingencies for each of the eleven colors (C4-C14).
\begin{tabular}{|l|cccc|}
\hline & C1 & C2 & C3 & C4-C14 \\
\hline Background 1: 100 & 1.0 & 0 & 0 & 0 \\
Background 2: 60/40 & 0 & 0.6 & 0.4 & 0 \\
Background 3: Random & 0.07 & 0.07 & 0.07 & 0.07 \\
\hline
\end{tabular}

\section*{Method}

Participants Twenty-two participants (10 females; \(M_{\text {age }}=\) 19) from the University of Oklahoma participated in Experiment 1 for course credit. All participants reported normal or corrected-to-normal vision. Four participants were excluded from the analysis due to high error rates (error rates \(\geq 15 \%\) ) during the visual search task, leaving 18 participants for the analysis.
Stimuli and Apparatus Stimuli were presented on a 17 " monitor, controlled by a Dell computer with a 3 GHz Pentium 4 processor. Distance to the monitor was approximately 60 cm . Stimulus presentation and data recording were controlled via E-Prime 2 by PST, Inc. The following 14 colors were used in all of the experiments presented black, blue, brown, cyan, green, lime, magenta, maroon, orange, pink, red, tan, white, and yellow. All the stimuli (the T's and L's) in the visual search array were 22 \(\mathrm{mm} \times 22 \mathrm{~mm}\). For each visual array each of the 14 items were placed randomly at one of 35 possible locations based on 3 ellipses of varying sizes.

Procedure. Each trial started with a fixation followed by a background cue ( 2004 ms ). The search array then followed. Participants were told to find the rotated " T " as quickly as possible and press the F-key when the "T" was rotated 90 degrees counterclockwise and the J-key when it was rotated 90 degrees clockwise. After responding a brief mask was presented ( 68 ms ) followed by a feedback screen ( 500 ms ) that indicated whether the response to the visual search array was correct. Figure 1 provides a schematic of the main components of each trial (i.e., the background cue and visual search array). Following the completion of all 360 visual search trials the participants then performed the recognition task as described above.


Figure 1: Schematic illustration of the main components of each trial of Experiments 1 and 2. The background cue was presented for 2004 ms , while the search array was presented until response.

\section*{Results}

For our main analysis, we examined RTs for each of the cue validity conditions. Trials in which the orientation of the target was mis-reported ( \(4.19 \%\) ), as well as trials with RTs faster than 200 ms or slower than \(10,000 \mathrm{~ms}\) ( \(2.66 \%\) ) were removed prior to analysis. We report the differences in the cue validity conditions collapsed across the last 3 epochs (although see Figure 2 for RT performance throughout the entire experiment). Specifically, the means of the median values (for each cue validity condition at Epochs 3-6) were calculated for each participant for each cue validity condition.

A within subjects repeated measures ANOVA revealed a significant main effect of cue validity on visual search RTs, \(F(2,34)=30.31, p<.01, \eta_{p}^{2}=641\). Pairwise comparisons revealed a significant difference between all three conditions such that the 100 ( \(M=1942.1 \mathrm{~ms}\) ) cue validity condition was significantly different from the 60/40 ( \(M=\) 2667.1 ms ) cue validity condition ( \(p<.001\) ) and the random cue validity condition ( \(p<.001\) ). The 60/40 cue validity condition was significantly different from the random ( \(M=\) 3033.8 ms ) cue validity condition ( \(p=.003\) ).


Figure 2: Reaction time performance as a function of epoch and cue validity in Experiment 1.

Seven out of the 18 participants (38.88\%) indicated that they had noticed a relationship between the background and
the likelihood of the target being of a particular color \({ }^{1}\). To examine participant's accuracy in the recognition task we only considered the 100 and 60/40 cue-validity conditions. The overall accuracy rate for the recognition task was at chance level (50.93\%). Accuracy rates for the different cue validity conditions were \(55.56 \%\), and \(48.61 \%\) for the 100 and \(60 / 40\) cue validity conditions. Participants who indicated that they recognized a relationship between the background and the likelihood of the target being a particular color were no better at the recognition task, with accuracy rates of \(50 \%\) for both the 100 and 60/40 conditions. Note that when the participants that indicated that they noticed a relationship were excluded from the analysis, a similar pattern of RT results as described above was obtained. Additionally, a between-subjects analysis of RTs did not reveal a significant difference between the group that indicated that they noticed a relationship from those that did not, \(F(1,16)=1.355, p=.262\).

\section*{Discussion}

In Experiment 1 we found improved search performance in accordance with the diagnosticity of the background cue. This RT result, coupled with the poor recognition performance suggests that participants may have been using the background cues at an implicit level of awareness, which is common in the contextual cueing literature.

One of the striking aspects concerning the results from many contextual cueing studies is its implicit nature (e.g., Chun \& Jiang, 1998; 1999). That is, although participants are able to find a target faster when aspects of a scene are repeated (as opposed to when they are changed), this occurs at an implicit level as participants are not able to discriminate the visual scenes they have viewed previously from those they have not (at the end of the training). Therefore, it could be an implicit utilization of contextual cues that explains the observed discrepancy between search and recognition performance in the present data. However, it should be pointed out that there is another possible explanation of the results for Experiment 1, which is explored below.

The second possible interpretation of the results holds that participants were essentially ignoring the background cues (and by extension the cue target associations), but were noting the likelihood of the target colors across all trials (i.e., the target color base rates). That is, participants' expectations regarding the color of the target may have been based on the unconditionalized base rate of the target color across all background conditions. For instance, let's assume that the color of the target is red in the 100 cue validity condition. After enough trials participants adopt an

\footnotetext{
\({ }^{1}\) During debriefing some of the 7 participants that answered affirmatively to the question "Did you notice any relationship between the background and the target colors?" were confusing the probability of target color given background with the probability of the target being a particular color (i.e., the base rate of the target colors). Thus, it is likely that even fewer than 7 participants truly noticed the association between target color and background.
}
attentional set for red and begin searching for the red item across all background conditions, even though red is not valid for the other cue validity conditions. Likewise the higher likelihood of the targets in the 60/40 condition makes them suitable targets across all background conditions as well. This explanation is supported by a study by Kunar, Flusberg, \& Wolfe (2006) that demonstrated improved search efficiency for consistently mapped search arrays when the delay between cue and array onset was sufficiently increased and participants were explicitly told about the relationship between cues and the critical aspect of the target. In Experiment 2 we investigate the two possibilities discussed above, by explicitly informing participants about the possible relationships between the background cues and the colors of the target.

\section*{Experiment 2}

Experiment 2 assessed whether participants in Experiment 1 were merely using base rate information to inform search or if they were using the background cues to improve search performance by developing a conditional attentional set. The experiment was nearly identical to Experiment 1 except there was a knowledge test at the end of each epoch, as opposed to the recognition test administered at the end of Exp. 1. The intention of this manipulation was to provide a hint to the participants that the background cues and the color of the targets were systematically associated.

\section*{Method}

Participants Twenty-Seven participants (23 females; \(M_{\text {age }}=\) 20.6) from the University of Oklahoma participated in Experiment 2 for course credit (26 participants) or \(\$ 10\). All participants reported normal or corrected-to-normal vision. Two participants were excluded from the analysis due to a high error rate (error rate \(\geq 15 \%\) ) and one was excluded for having exceptionally long RTs as determined by having a mean RT value more than 3 standard deviations higher than the average for the mean of the collapsed \(100,60 / 40\), and random conditions.

Procedure The same procedure used in Experiment 1 was used in Experiment 2 with the following exception. At the end of each epoch, participants were asked to indicate the most likely colors of the target (up to 4) given the cue.

\section*{Results}

Error trials were excluded from analysis (2.3 \%) as well as trials with RTs faster than 200 ms and slower than 10,000 ms (5.1\%). Errors were not analyzed in any manner.

As in Experiment 1 we collapsed across the last 3 epochs (see Figure 3 for RT performance across all epochs). A within subjects ANOVA revealed a main effect of cue validity, \(F(2,32)=133.302, p<.001, \eta_{p}^{2}=.847\) on search

RTs. Pairwise comparisons revealed a significant difference between the 100 ( \(M=1053.5 \mathrm{~ms}\) ) cue validity condition to the \(60 / 40\) ( \(M=1885.3 \mathrm{~ms}\) ) cue validity condition ( \(p<.001\) ) and the Random ( \(M=3243.3 \mathrm{~ms}\) ) cue validity condition ( \(p<\) .001). The \(60 / 40\) cue validity condition was also significantly different from the Random cue validity condition ( \(p<.001\) ).


Figure 3: Reaction time performance as a function of epoch and cue validity in Experiment 2.

We next performed a between subjects analysis comparing the respective cue validity conditions of Experiment 2 to Experiment 1 (the mean values obtained whilst collapsing across the last epochs) with cue validity (100, 60/40, Random) as a within subjects variable and Experiment as a between subjects variable.


Figure 4: Reaction time performance comparing the participants of Experiment 2 with the participants of Experiment 1. Error bars represent standard errors. Note that the bars represent the means for Epochs 4-6.

A main effect of cue validity was found \((F(2,82)\) \(=134.183, p<.001, \eta_{p}^{2}=.766\) ) as well as a main effect for Experiment such that participants were faster in Experiment 2 than in Experiment 1, \(\left(F(1,33)=15.533, p<.001, \eta_{\mathrm{p}}^{2}=\right.\) .275). An interaction between Experiment and cue validity was observed, \(\left(F(2,82)=18.26, p<.001, \eta_{p}^{2}=.308\right)\). A Post-hoc analysis revealed a significant difference when comparing the 100 cue validity condition in Experiment 2 to the 100 cue validity condition in Experiment 1 ( \(p<.001\) ) and also the 60/40 cue validity condition was significantly different across experiments ( \(p=.012\) ), but the Random cue validity condition was not significantly different across Experiments ( \(p=1.0\) ). Thus, participants were significantly faster to find the target when they were informed of the possible connection between the cues and the color of the
target, suggesting that the participants in Experiment 1 were merely relying on base rate information to guide search \({ }^{2}\). There were also quite large differences in memory performance as well in Experiment 2. To examine whether this had an influence on the visual search process we split the participants into two groups based on their performance in the memory task. If a participant named the 100 color first for background 1 and the 60/40 colors first or second (in any order) for background 2 at one time during testing (out of the 6 times to do so), and only those colors, they were placed into the good memory performance group (17 participants) and the others were placed into the poor memory performance group (8 participants). Figure 5 plots the good and poor memory performers in the knowledge test.


Figure 5: Performance in the knowledge test as a function of epoch, cue diagnosticity, and memory performance classification for Experiment 2. Poor memory performers are plotted in gray and good memory performers are plotted in black. Those that performed well increased dramatically over the course of the experiment whereas the poor memory participants did not.

We performed a between subjects analysis on the visual search RTs to compare the good memory performers from the poor memory performers, with cue validity (100, 60/40, Random) as a within subjects variable and memory performance (good, poor) as a between subjects variable. This analysis revealed a significant main effect of cue validity \(\left(F(2,46)=123.781, p<.001, \eta_{\mathrm{p}}^{2}=.843\right)\) as well as memory performance on search RTs \((F(1,23)=12.881, p\) \(\left.=.002, \eta_{p}^{2}=.359\right)\) such that participants were faster as cue validity increased and participants who performed well in the memory task were faster overall in the visual search task. An interaction between cue validity and memory performance type was also revealed \(F(2,46)=5.007, p=\)

\footnotetext{
\({ }^{2}\) Although this statement is conjecture, we have additional empirical evidence that the participants' performance in Experiment 1 is entirely compatible with the notion that they were only using base rate information. In an additional follow up experiment we eliminated the diagnosticity of the background cue(s) by presenting only one background throughout the entire experiment (i.e., on each trial they saw the same background cue). We found remarkably similar performance in this experiment as compared to Experiment 1. Post hoc analysis revealed no differences between the 100 and 60/40 cue validity conditions across experiments (both \(p\) ' \(=1.0\) ).
}
\(.011, \eta_{\mathrm{p}}^{2}=.179\). Although a between subjects analysis using a post-hoc Bonferonni test did not reveal a significant difference when comparing the 100 cue validity condition of the good memory performers \((M=893)\) to the poor memory performers ( \(M=1394.7\) ) of Experiment 2 ( \(p=.59\) ), a significant difference was observed when comparing the respective \(60 / 40\) cue validity conditions ( \(p=.012\) ). The good memory performers were faster in the 60/40 condition ( \(M=1601.1\) ) relative to the poor memory performers ( \(M\) =2489.2).

Because there seemed to be a qualitative difference in search RTs for the poor memory performers to the good memory performers, we next compared the poor memory performers of Experiment 2 to the participants of Experiment 1 to examine whether these two sets of participants were qualitatively similar. A main effect of cue validity was found \(\left(F(1,48)=51.985, p<.001, \eta_{p}^{2}=.684\right)\). There was no effect for Experiment, \((F(1,24)=1.028, p=\) .321, \(\left.\eta_{p}^{2}=.041\right)\). An interaction was observed, \((F(2,48)=\) 3.568, \(p=.036, \eta_{p}^{2}=.129\) ), however, a post-hoc analysis revealed that the 100 cue validity condition was not significantly different across groups ( \(p=.91\) ) nor was the 60/40 cue validity condition ( \(p=1.0\) ), nor the Random cue validity condition ( \(p=1.0\) ). Thus, it appears that the poor memory performers: 1) were qualitatively different than the good memory performers of Experiment 2, particularly for the 60/40 cue validity condition and 2 ) were qualitatively similar to the participants of Experiment 1 for the 100 and 60/40 cue validity conditions.

\section*{Discussion}

Experiment 2 revealed that the participants who learned the associations between the cues and the critical features of the target were able to leverage that information in the visual search task. Although participants were unable to pick up on the cue-target associations without a suggestion from the experimenter that such a relationship may exist (see Footnote 2 and General Discussion), most participants were able to do so once this relationship was suggested in Experiment 2. We also found that performance of the poor memory performers was similar to that of the participants in Experiment 1. In the General Discussion we posit that an important cognitive mechanism (i.e., attentional selection) was not operating over the cues for the participants of Experiment 1 and the poor memory performers in Experiment 2, thus disabling them from using the cues to improve their search.

\section*{General Discussion}

In two experiments we explored whether participants would be able to use cues to simplify a perceptually demanding visual search task. We found that although participants were sensitive to base rate information in Experiment 1, Experiment 2 revealed that the participants in Experiment 1 were likely not utilizing the background cues. Although this conclusion may be un-warranted given that a cue validity effect was found in Experiment 1, we ran an additional
experiment where the diagnosticity of the background cues was eliminated by using the same background cue for the different cue validity conditions. We found remarkably similar performance to Experiment 1 such that there were no differences for the 100 and 60/40 cue validity conditions (see Footnote 2).

In Experiment 2 we also found a qualitative difference between participants who performed well on the knowledge tests versus those that did not. The good memory performers were faster overall and were particularly faster when the background cue suggested two features (i.e., the 60/40 condition). Thus, it appears that suggestion of the possible cue to target color connection and explicit knowledge of the cue-target associations (as evidenced by good performance in the memory task) are important factors in contributing to the use of cues from the environment to support conditional visual search in the retrieval guidance paradigm (c.f. contextual cueing; Chun \& Jiang, 1998; 1999). The results suggest that the paradigm used allows for an examination of how LTM memory, WM, and attention coordinate in a visual search task.

We argue that the difference between these two sets of participants is due to attentional selection (see TurkBrowne, Jungé, \& Scholl, 2005). We suggest that the poor memory performers of Experiment 2 and the participants of Experiment 1 were not attending to the background cue on each trial and thus were not able to exploit the cues later in the experiment (i.e., Epochs 3-6). The good memory performers, on the other hand, were attending to the cues, thereby allowing them to encode the cue-target color associations into LTM, and exploit their memories as the task progressed. We now turn to a discussion of how a recent model of hypothesis generation called HyGene (short for Hypothesis Generation; Dougherty, Thomas, \& Lange, 2010; Thomas, Dougherty, Sprenger, \& Harbison, 2008) provides a useful framework for theorizing about interactions between visual search and memory within the retrieval guidance paradigm and beyond.

Although HyGene was originally developed to bridge research concerning LTM, WM, and judgment \& decision making, we argue that it provides a useful framework for understanding how WM and LTM interact to support visual search. In this framework, an individual receives information from the environment, such as the background cue in the present experiment, which prompts the generation of hypotheses previously associated with the observed information. In the case of the retrieval guidance paradigm, the hypotheses being retrieved are the likely colors of the forthcoming target on each trial. The retrieved hypotheses are placed into WM and are then available to drive search processes in a top-down manner (e.g., Desimone \& Duncan, 1995), affording the filtering out of perceptual information that is discordant with WM content.

Although we believe that HyGene provides a useful framework for understanding ecological visual search, it currently does not have direct access to the outside world (i.e., visual input). Because of this, the model cannot yet
make specific predictions which are likely important to visual search researchers (e.g., RTs, fixations). Thus, it would be fruitful to integrate HyGene with models of visual search, such as the guided search model (Wolfe, 1994) or the more recent target acquisition model (Zelinsky, 2008). Not only would such models be able to perform visual search, but they would also would generate predictions concerning information foraging (e.g., Hypothesis testing; Dougherty, Thomas, \& Lange, 2010).

\section*{References}

Chun, M., \& Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. Cognitive Psychology, 36, 28-71. doi: 10.1006/cogp.1998.0681

Chun, M. M., \& Jiang, Y. (1999). Top-down attentional guidance based on implicit learning of visual covariation. Psychological Science, 10, 360-365. doi: 10.1111/14679280.00168

Desimone, R., \& Duncan, J. (1995). Neural mechanisms of selective visual attention. Annual Review of Neuroscience, 18, 193-222. doi: 10.1146/annurev.ne.18.030195.001205
Dougherty, M., Thomas, R., \& Lange, N. (2010). Toward a integrative theory of hypothesis generation, probability judgment, and hypothesis testing. In B. H. Ross, B. H. Ross (Eds.) , The psychology of learning and motivation: Advances in research and theory (Vol 52) (pp. 299-342). San Diego, CA US: Elsevier Academic Press.
Kunar, M. A., Flusberg, S., Horowitz, T. S., \& Wolfe, J. M. (2007). Does contextual cuing guide the deployment of attention?. Journal Of Experimental Psychology: Human Perception And Performance, 33(4), 816-828. doi:10.1037/0096-1523.33.4.816
Kunar, M. A., Flusberg, S., \& Wolfe, J. M. (2006). Contextual cuing by global features. Perception \& Psychophysics, 68(7), 1204-1216. doi: 10.1037/00961523.33.4.816

Thomas, R., Dougherty, M., Sprenger, A., \& Harbison, J. (2008). Diagnostic hypothesis generation and human judgment. Psychological Review, 115(1), 155-185. doi: 10.1037/0033-295X.115.1.155

Turk-Browne, N. B., Jungé, J., \& Scholl, B. J. (2005). The Automaticity of Visual Statistical Learning. Journal Of Experimental Psychology: General, 134(4), 552-564. doi: 10.1037/0096-3445.134.4.552

Wolfe, J. (1994). Guided search 2.0: A revised model of visual search. Psychonomic Bulletin \& Review, 1(2), 202-238. doi: 10.3758/BF03200774
Zelinsky, G. J. (2008). A theory of eye movements during target acquisition. Psychological Review, 115(4), 787835. doi: 10.1037/a0013118

Zellin, M., Conci, M., van Mühlenen, A., \& Müller, H. J. (2011). Two (or three) is one too many: Testing the flexibility of contextual cueing with multiple target locations. Attention, Perception, \& Psychophysics, 73(7), 2065-2076. doi:10.3758/s13414-011-0175-x

\title{
Intentionality and Choice
}

\author{
Benignus N. Ndubuisi (ndubuibn@tcd.ie) and Ruth M.J. Byrne (rmbyrne@tcd.ie) \\ School of Psychology \& Institute of Neuroscience, Trinity College Dublin, University of Dublin, Dublin 2, Ireland
}

\begin{abstract}
People judge that harmful side effects are intentional, e.g., a CEO who introduces a new program to increase profits that results in harm to the environment is judged to have intentionally harmed the environment. They judge helpful side effects are unintentional, e.g., a CEO who introduces a new program to increase profits that results in helping the environment is not judged to have intentionally helped the environment. We report two experiments that suggest the effect arises because people believe individuals can make alternative choices in bad situations and not in good ones.
\end{abstract}

Keywords: Intentions, choices, side-effects, inactions.

\section*{Intentionality}
"Nothing hinders one act from having two effects, only one of which is intended, while the other is beside the intention. Now moral acts take their species according to what is intended, and not according to what is beside the intention, since this is accidental." Aquinas (1265-1274)

Intentionality is a core category of mental life, along with space, time and cause (Miller \& Johnson-Laird 1976). Philosophers, psychologists and legal scholars have identified that the accurate assessment of other people's intentions is vital to moral and legal judgment, and to how we understand and explain other people's behavior (e.g., Knobe 2010). Logicians and artificial intelligence researchers have modeled intentions using dynamic doxastic logic and related systems as an important aspect of simulating revisions to beliefs (e.g., Gardenfors 1988). Neuroscientists and psychiatrists have established that the loss of the ability to reason about intentions is catastrophic after prefrontal cortex damage (e.g., Young, Bechara, Tranel, Damasio, Hauser, \& Damasio 2010) and in disorders such as schizophrenia (e.g., Roese, Park, Smallman \& Gibson 2008), just as it is essential to children's proper development of a theory of mind (e,g., Leslie, Knobe \& Cohen 2006). Not surprisingly then, it is of concern that recent evidence indicates that people may make systematic errors in their assessments of other people's intentions, at least in relation to the intentionality of side effects (e.g., Knobe 2010). Consider the following story:

The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment.' The chairman of the
board answered, 'I don't care at all about harming the environment. I just want to make as much profit as I can. Let's start the new program.' They started the new program. Sure enough, the environment was harmed.

Participants judged that the chairman intentionally harmed the environment (Knobe, 2003a). The judgment is puzzling because intentionality implies that the protagonist desires the outcome, has the belief or knowledge that the action will bring it about, and intends to carry it out, as well as that the protagonist is aware of carrying it out and has the skill to do so (Malle \& Knobe, 1997).
The puzzle deepens when participants are told that the program will help rather than harm the environment:

The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, and it will also help the environment.' The chairman of the board answered, 'I don't care at all about helping the environment. I just want to make as much profit as I can. Let's start the new program.' They started the new program. Sure enough, the environment was helped.

Participants judged the chairman did not intentionally help the environment (Knobe 2003a).
The asymmetry between harmful and helpful side-effects occurs in many different sorts of situations. It occurs for different contents, ranging from serious violations such as when an army commander's decision to capture a region in battle has the side effect of affecting the numbers of soldiers killed (Knobe 2003a), to more trivial ones such as when a protagonist's decision to mow the lawn early in the morning affects their neighbor's sleep (Sverdlik 2004). It is observed in languages other than English such as Hindi (Knobe \& Burra 2006) and for an array of linguistic expressions such as 'advocated' and 'decided' (McCann 2005; Pettit \& Knobe 2009). It emerges early in young children (Leslie, et al 2006) and occurs even in people with deficits such as Asperger's (Zalla \& Leboyer 2011).
Why do people judge that others bring about harmful side-effects intentionally but helpful side-effects unintentionally? The issue is hotly debated and several alternative explanations have been proposed to account for it.

\section*{Intentionality and morality}

One influential explanation for the asymmetry in judgments of the intentionality of harmful and helpful side-effects is that people first assess the morality of the side-effect. Their judgment of the side effect as morally good or morally bad infuses their judgment of its intentionality (e.g., Knobe 2006). Against this proposal however, it has been observed that people judge that a protagonist brought about a harmful side-effect intentionally even for non-moral side-effects. For example, when the CEO of a movie company decides to introduce a new program that will increase profits and have the side effect of making movies worse from an artistic standpoint, participants judged that he intentionally brought about the side-effect of harming movies from an artistic standpoint (Knobe 2004). When the story substituted 'help' for 'harm', they judged that he didn't intentionally bring about the side effect of helping movies from an artistic standpoint. Likewise, the effect occurs for non-moral norms, e.g., it occurs for a decision to change a manufacturing process that will have the side effect of creating a product that deviates from an industry standard of 'darker than blue' (Uttich \& Lombrozo 2010).
Most tellingly, badness and intentionality can be 'doubly dissociated'. On the one hand, some harmful side effects are judged intentional even when they are not judged to be bad e.g., a chairman who decides to increase profits in one branch of the company with the side effect of decreasing profits in another branch, is judged to have intentionally harmed the other branch's profits even though harming the other branch's profits is not judged to be bad (e.g., Knobe 2006; Knobe \& Mendlow 2007). On the other hand, some side effects are judged unintentional even though they are judged bad, e.g., a town-planner who introduces a program to clean toxic waste with the side effect of increasing joblessness is judged to have affected joblessness unintentionally even though joblessness is judged to be bad (e.g., Phelan \& Sarkissian 2008; Sverdlik 2004).
A related explanation is that people judge individuals to be blameworthy when their decisions lead to harm. The motivation to express blame leads participants to conclude that the harmful side effect is intentional (e.g., Adams \& Steadman 2004; Alicke 2008; Mele 2003; Nadelhoffer 2004). Against this proposal however, it has been observed that a harmful side-effect is judged intentional even when participants have the opportunity to blame the protagonist, or otherwise to assign responsibility to the protagonist separately (e.g., Knobe 2003b; Pellizzoni, Girotto \& Surian 2010). Again, most tellingly, blame and intentionality can be 'doubly dissociated'. On the one hand, some harmful side-effects are judged unintentional even when the protagonist is blamed, e.g., a driver who goes out of control while drunk and injures a family is judged to be blameworthy, but not to have harmed the family intentionally (Knobe 2003b). On the other hand, some harmful side-effects are judged intentional even when the protagonist is not blamed, e.g., a dentist who
carries out necessary dental surgery and inflicts pain on the patient is judged to have inflicted the pain intentionally but is not blamed for doing so (Sverdlik 2004).
Hence, moral assessments of goodness and badness, or judgments of blameworthiness, do not appear to be the reason why people tend to judge harmful side-effects to be intentional and helpful side-effects to be unintentional. However, one further possibility from this perspective is that the difference in intentionality judgments for harmful and helpful side-effects arises because of the moral disparity between the primary goal (increasing profits) and the side-effect (affecting the environment): the primary goal itself may be perceived to be morally bad in the context of the side effect. The first experiment tests this moral disparity explanation.

\section*{Experiment 1}

We gave participants an 'Aid' story in which the primary goal was elevated to be of equivalent moral status to the side-effect of affecting the environment:

The vice-president of an international aid charity went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us save more people from starvation in Africa, but it will also harm the environment.' The chairman of the board answered, 'I don't care at all about harming the environment. I just want to save as many people as I can. Let's start the new program.' They started the new program. Sure enough, the environment was harmed.

If moral assessments are central to intentionality judgments, then harming the environment should be judged unintentional because its immorality is ameliorated by the morality of saving starving people. We also gave participants a 'Rival' story in which the side-effect was diminished to be of equivalent moral status to the primary goal of increasing profits:

The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, but it will also harm our rival's profits.' The chairman of the board answered, 'I don't care at all about harming our rival's profits. I just want to make as much profit as I can.' Let's start the new program.' They started the new program. Sure enough, the rival's profits were harmed.

\section*{Method}

The participants were 60 students from Trinity College Dublin who took part voluntarily. They were 17 men and 43 women, aged 16 to 58 years, with an average age of 24 years.

Participants were assigned to the Aid or Rival groups ( \(\mathrm{n}=30\) in each). They were each given a harm and help
version of the story. Half received the harm version first and half the help version first (and no effects of order were observed). They completed several tasks, such as a praiseblame assignment task, as well as the key side-effect intentionality judgment task, in response to the question, 'Do you think the chairman intentionally affected the <side-effect>?'. They circled their answer on a 7 point likert-type scale with 6 anchored as 'intentional' and 0 as 'unintentional' and the mid-point anchored as 'neither'.

They were instructed that they would be given two short stories and they were asked to read them carefully, to answer the questions in the order they were given, not to change any of their answers, and to complete all of the questions on one story before moving on to the next.

\section*{Results and Discussion}

Participants judged that the protagonist intentionally affected the side-effect more for the harmful side-effect than the helpful one in the Aid condition, Wilcoxon's \(\mathrm{z}=\) \(-2.828, \mathrm{p}=.005, \mathrm{r}=.365\) as Figure 1 shows. (For clarity, responses were graphed using scores translated from 0 to 6 to -3 to +3 ). The result shows that even when the protagonist's primary goal was elevated to be morally compelling (saving people from starving) participants tended to judge that he brought about the harmful side effect (harming the environment) more intentionally than the helpful side effect (helping the environment).


Figure 1: Judgments of intentionality for harmful and helpful side-effects in the Aid and Rival conditions. Error bars are standard error of the mean.

Participants also judged that the protagonist intentionally affected the side-effect more for the harmful outcome than the helpful one in the Rival condition, Wilcoxon's \(z=\) 2.481, \(p=.013, r=.3203\), as Figure 1 also shows. Even when the side-effect (affecting a rival's profits) was diminished to be as morally unenlightening as the primary
goal (increasing one's own profits), participants judged that the protagonist brought about the harmful side effect (harming the rival's profits) more intentionally than the helpful side effect (helping the rival's profits).
The results suggest that the moral disparity between the primary goal, of increasing profits, and the side-effect, of harming or helping the environment, does not underlie the asymmetry in judgments of intentionality. An alternative explanation is based on the availability of choice.

\section*{Intentionality and choice}

A new explanation for why people judge harmful side effects to be intentional and helpful side effects to be unintentional is that the protagonist is perceived to have a choice when faced with the harmful dilemma but not when faced with the helpful one. We propose that a harmful side-effect poses a genuine dilemma: the goal is positive whereas the side-effect is negative, and in a dilemma a protagonist makes choices between priorities.
This availability of choice explanation proposes that people think about whether the protagonist has other options. They can think about an alternative to the harmful side-effect: the protagonist could have decided not to introduce the program to increase profits, and so not harmed the environment. Because they can think of alternatives, they perceive that the protagonist had a choice and they judge the side-effect to be intentional. In contrast, a helpful side-effect poses no dilemma: the goal and side-effect are positive and the protagonist need not make choices between them: his action will increase profits and help the environment. Participants do not tend to think of an alternative to the helpful side-effect and so they perceive that the protagonist had little choice and they judge the side-effect to be unintentional. This suggestion is consistent with earlier acknowledgments that choice has a potential role in intentionality judgments (e.g., Alicke 2008; Cushman \& Young 2011; Machery 2008; Malle \& Knobe 1997; Phillips \& Knobe 2009; Royzman \& Baron 2002). For example, when the protagonist does not have sufficient knowledge of the outcome, the effect is eliminated (e.g., Nichols \& Ulatowski 2007; Pellizzoni et al 2010).
At the heart of the availability of choice explanation is the idea that thinking about choices requires people to imagine alternatives: they think about the protagonist's choice of pursuing the goal and its harmful side-effect, and they imagine a counterfactual alternative of not pursuing the goal and no harmful side-effect. Evidence to support this suggestion comes from the observation that when participants are required to create 'if only' counterfactual thoughts about how things could have turned out differently prior to making their judgments of intentionality, the side-effect asymmetry is amplified (Byrne 2012). They judged the harmful side effect to have been brought about intentionally more often when they were required to create counterfactuals compared to when they were not, presumably because they could think of
alternative choices the protagonist could have made; they judged the helpful side effect to have been brought about unintentionally more often when they were required to create counterfactuals compared to when they were not, presumably because they could not think of alternative choices the protagonist could have made. The suggestion is consistent with the idea that intentionality judgments may potentially be affected by counterfactual generation (e.g., Adams \& Steadman 2004; Knobe, 2010; McCloy \& Byrne 2000; Pellizzoni et al 2010; Young \& Phillips 2011). Our second experiment tests the availability of choice proposal, by examining intentionality judgments for side-effects that are brought about by actions or inactions.

\section*{Experiment 2}

Outcomes that result from a protagonist's actions may appear to be the result of deliberate choices, more so than outcomes that result from a protagonist's inactions. If so, the asymmetry in judging harmful side-effects to be intentional and helpful side-effects to be unintentional may be diminished when the side-effects result from the protagonist's inaction, rather than from the protagonist's action.
People tend to regret bad outcomes that arise from their actions more than bad outcomes that arise from their inactions. Consider the following scenario:

Mr. Paul owns shares in company A. During the past year he considered switching to stock in company B, but he decided against it. He now finds out that he would have been better off by \(\$ 1,200\) if he had switched to the stock of company B. Mr. George owned shares in company B. During the past year he switched to stock in company A. He now finds out that he would have been better off by \(\$ 1,200\) if he had kept his stock in company B . Who feels greater regret?

Most people judge that the actor, Mr. George, will regret his action more than the individual who did not act, Mr. Paul (Kahneman \& Tversky 1982). Even when their task is to judge the regret that a person experienced without making a comparison to the regret experienced by another person, their estimates of regret for an actor are higher than their estimates of regret for a non-actor (Feeney \& Handley 2006; N'gbala \& Branscombe, 1997). They also judge that actors will feel better about good outcomes that arise from their actions compared to individuals whose inaction leads to a good outcome (Landman, 1987). This 'omission bias' may arise because actions appear to change the status quo more than inactions (Byrne \& McEleney 2000; Ritov \& Baron 1999). Of course, when there are compelling reasons to act, inactions can be seem inexcusable (Gilovich \& Medvec 1995; Zeelenberg, Van den Bos, Van Dijk, \& Pieters 2002).

We gave participants a version of the company scenario which emphasized the protagonist action in switching to a new program:

The vice-president of a company UMT Ltd went to the chairman of the board, Mr. Smith, and said 'We are thinking of switching to a new program, instead of staying with our old one. If we switch to the new program it will help us increase profits, but it will also harm the environment'. The chairman of the board, Mr. Smith answered, 'I have no desire to affect the environment. I just want to make as much profit as I can'. Mr. Smith considered staying with their old program but in the end he said 'Let's switch to the new program.' They switched to the new program and sure enough, the environment was harmed.

We compared this 'Action' version to an 'Inaction' version which indicated instead that the protagonist had not acted:

The vice-president of another company in a different region, OZF Inc went to the chairman of the board, Mr. Jones, and said 'We are thinking of staying with our old program, instead of switching to a new one. If we stay with the old program it will help us increase profits, but it will also harm the environment'. The chairman of the board, Mr. Jones, answered, 'I have no desire to affect the environment. I just want to make as much profit as I can'. Mr. Jones considered switching to the new program but in the end he said 'Let's stay with the old program.' They stayed with the old program and sure enough, the environment was harmed.

We used the phrase 'I have no desire to affect the environment' rather than 'I don't care at all about harming the environment' to obviate any inference that the protagonist was maliciously negligent (Cushman \& Mele 2008; Guglielmo \& Malle 2010).

\section*{Method}

The participants were 40 students from Trinity College Dublin who participated voluntarily. They were 7 men and 33 women, aged 18 to 43 years, with an average age of 25 years.

Participants were assigned to the harm or help groups ( n \(=20\) in each). They each received an action and an inaction version of the story (in that order, see Feeney \& Handley 2006). They received the action and inaction versions for two vignettes, the company scenario and also a parenting scenario in which a county council chairwoman pursued a primary goal of funding basic community services such as maintenance of roads and parks, with the side-effect of harming (or helping) funding for a 'better parenting' program designed to provide skills to vulnerable adults in at-risk families.

Half the participants received the company scenario first and half the parenting one, and order had an effect (participants tended to give different responses to the company scenario when it appeared first rather than second). Hence we report the results for participants' responses to the first scenario they received only. There were no differences in responses to the two contents, company versus parenting when they were received first. Participants completed several tasks, such as a praiseblame assignment task, as well as the key side-effect intentionality judgment task. We eliminated four participants because their response to a final question on 'protected values' (absolute values that people protect from trade-offs) indicated they did not value the environment or parenting programs highly (Ritov \& Baron, 1999; Tanner \& Medin 2004). The procedure was similar to the previous experiment.

\section*{Results and Discussion}

Participants in the Action condition judged that the actor brought about the side-effect intentionally for the harmful side-effect more than the helpful one, Mann Whitney U, z \(=-2.111, \mathrm{p}=.035, \mathrm{r}=.3518\). The difference was eliminated in the Inaction condition, \(z=-1.203, p=.229\), as Figure 2 shows. The result is consistent with the suggestion that the asymmetry in judgments of intentionality for harmful and helpful side-effects arises from the perceived differential availability of choice for harmful and helpful side-effects that result from actions. Side-effects that result from a protagonist's inaction may appear not to be the result of deliberate choice as much as those that result from a protagonist's actions.


Figure 2: Judgments of intentionality for harmful and helpful side-effects in the Action and Inaction conditions. Error bars are standard error of the mean.

However, it is important to note that participants judged the protagonist had brought about a harmful side effect intentionally as often for an inaction as for an action,

Wilcoxon's \(z=-.680, p=.479\); in contrast, they judged the protagonist had brought about a helpful side effect marginally more unintentionally for an action than for an inaction, \(\mathrm{z}=-1.792, \mathrm{p}=.073, \mathrm{r}=.299\). The result indicates that people judge the unintentionality of side-effects differently when they arise from inaction rather than action.

\section*{Conclusions}

Participants tend to judge that an individual brought about a harmful side-effect intentionally, but a helpful side-effect unintentionally (Knobe 2003a). We suggest the phenomenon arises because a harmful side-effect poses a genuine dilemma in which the actor must make choices, whereas a helpful side-effect poses no dilemma and the actor does not need to make choices. People imagine a counterfactual alternative in which the actor made a different choice for a harmful side-effect and the readily available imagined alternatives for a harmful side-effect lead them to infer it was intentional.
The tendency to judge that an individual brought about a harmful side-effect intentionally but a helpful side-effect unintentionally persists even when the goal is of equal moral worth to the side effect, such as saving people from starvation, as the first experiment showed, ruling out an explanation based on moral disparity. The effect is eliminated when the side effects arise from inactions rather than actions, as the second experiment showed, consistent with an explanation based on the availability of choice. The availability of choice explanation is consistent with the view that common mechanisms underlie reasoning about intentionality in moral and non-moral domains (e.g., Byrne 2005; Rai \& Holyoak 2010; Royzman \& Baron 2002; Shenhav \& Greene 2010; Uttich \& Lombrezo 2011).

\section*{Acknowledgments}

Some of the research was funded by the TCD Arts and Social Sciences Benefactions Fund. Thanks to Shane Doherty and Kathryn Lambe for help collecting the data.

\section*{References}

Adams, F., and Steadman, A. (2004). Intentional action in ordinary language. Analysis, 64, 173-181.
Alicke, M.D. (2008). Blaming badly. Journal of Cognition and Culture 8, 179-186.
Byrne, R.M.J. (2012). Reasoining about intentions. Talk at the International Conference on Thinking, London, UK.
Byrne, R.M.J. (2005). The Rational Imagination: How People Create Alternatives to Reality. Cambridge: MIT.
Byrne, R.M.J. \& McEleney, A. (2000) Counterfactual thinking about actions and failures to act. Journal of Experimental Psychology: Learning, Memory, and Cognition, 26, 1318-1331
Cushman, F., \& Mele, A. (2008). Intentional action: Two-and-a-half folk concepts? In J. Knobe \& S. Nichols (Eds.), Experimental Philosophy. Oxford: OUP.

Cushman, F., \& Young, L. (2011). Patterns of moral judgment derive from nonmoral psychological representations. Cognitive Science, 1-24.
Feeney, A., \& Handley, S. J. (2006). Comparisons, mental models, and the action effect in judgements of regret. Memory \& Cognition, 34(7), 1422-1430.
Guglielmo, S. \& Malle, B. F. (2010). Can unintended side effects be intentional? Personality and Social Psychology Bulletin, 36, 12, 1635-1647
Gärdenfors, P. (1988). Knowledge in Flux. Cambridge, MA: MIT Press.
Gilovich, T. \& Medvec, V.H. (1995). The experience of regret: what, when, and why. Psychological Review, 102, 379-395.
Kahneman, D., \& Tversky, A. (1982). The psychology of preferences. Scientific American. 246, 1, 160-173.
Knobe, J. (2003a). Intentional action and side-effects in ordinary language. Analysis, 63, 190-193.
Knobe, J. (2003b). Intentional action in folk psychology. Philosophical Psychology, 16, 309-323.
Knobe, J. (2004). Folk psychology and folk morality. Journal of Theoretical and Philosophical Psychology, 24, 270-279.
Knobe, J. (2006). The concept of intentional action. Philosophical Studies, 130, 203-231.
Knobe, J. (2010). Person as scientist, person as moralist. Behavioral and Brain Sciences. 33, 353-365.
Knobe, J., \& Burra, A. (2006). The folk concept of intention and intentional action. Journal of Culture and Cognition, 6, 113-132.
Knobe, J., \& Mendlow, G. (2004). The good, the bad, and the blameworthy. Journal of Theoretical and Philosophical Psychology, 24, 252-258.
Landman, J. (1987). Regret and elation following action and inaction. Personality and Social Psychology Bulletin, 13, 524-536.
Leslie, A., Knobe, J., \& Cohen, A. (2006). Acting intentionally and the side-effect effect. Psychological Science, 17, 421-427.
Machery, E. (2008). The folk concept of intentional action. Mind and Language, 23, 2, 165-189.
Malle, B.F. \& Knobe, J. (1997). The folk concept of intentionality. Journal of Experimental Social Psychology, 33, 101-121.
McCann, H. (2005). Intentional action and intending. Philosophical Psychology, 18, 6, 737/748.
McCloy, R. \& Byrne, R.M.J. (2000) Counterfactual thinking about controllable actions. Memory \& Cognition, 28, 1071-1078.
Miller, G. \& Johnson-Laird, P.N. (1976). Language and perception. Harvard: Harvard University Press.
Mele, A. (2003). Intentional action. Philosophical Psychology, 16, 325-340.
Nadelhoffer, T. (2004). On praise, side effects, and folk ascriptions of intentional action. Journal of Theoretical and Philosophical Psychology, 24, 196-213.

N'gbala, A., \& Branscombe, N. R. (1997). When does action elicit more regret than inaction and is counterfactual mutation the mediator of this effect? Journal of Experimental Social Psychology, 33, 324343.

Nichols, S., \& Ulatowski, J. (2007). Intuitions and individual differences. Mind \& Language, 22, 346-365.
Pellizzoni, S., Girotto, V., \& Surian, L. (2010). Beliefs and moral valence affect intentionality attributions. Review of Philosophy and Psychology, 1, 2, 201-209
Pettit, D. and Knobe J. (2009). The pervasive impact of moral judgment. Mind \& Language 24, 5, 586-604.
Phelan, M., and Sarkissian, H. (2008). The folk strike back. Philosophical Studies, 138, 291-298
Phillips J., \& Knobe, J. (2009). Moral judgments and intuitions about freedom. Psychological Inquiry, 20, 1, 30-36.
Rai, T.S. \& Holyoak, K.J. (2010). Moral Principles or Consumer Preferences? Cognitive Science 34, 311-321
Roese, N. J., Park, S., Smallman, R., \& Gibson, C. (2008). Schizophrenia involves impairment in the activation of intentions by counterfactual thinking. Schizophrenia Research, 103, 343-344.
Rozyman, E.B. \& Baron, J. (2002). The preference for indirect harm. Social Justice Research, 15, 2, 165-184.
Ritov, I., \& Baron, J. (1999). Protected values and omission bias. Organizational Behavior and Human Decision Processes, 79, 2, 79-94.
Shenhav, A. \& Greene, J.D. (2010). Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. Neuron. 67, 667-677
Sverdlik, S. (2004). Intentionality and moral judgments in commonsense thought about action. Journal of Theoretical and Philosophical Psychology, 34, 224-236.
Tanner, C., \& Medin, D. L. (2004). Protected values: No omission bias and no framing effects. Psychonomic Bulletin \& Review, 11, 1, 185-191.
Uttich, K., \& Lombrozo T. (2010). Norms inform mental state ascriptions: A rational explanation for the sideeffect effect. Cognition 116, 1, 87-100.
Young, L., Bechara, A., Tranel, D., Damasio, H., Hauser, M., \& Damasio, A. (2010). Damage to ventromedial prefrontal cortex impairs judgment of harmful intent. Neuron, 65, 6, 845-851.
Young, L., \& Phillips, J. (2011). The paradox of moral focus. Cognition, 119, 166-178.
Zalla, T. \& Leboyer, M. (2011). Judgment of intentionality and moral evaluation in individuals with high functioning autism. Review of Philosophy and Psychology, 2, 681-698
Zeelenberg, M., Van den Bos, K., Van Dijk, E., \& Pieters, R. (2002). The inaction effect in the psychology of regret. Journal of personality and Social Psychology, 82, 3, 314-327

\title{
Emotional Priming of Sentence Comprehension: Effects of a Speaker's Static Emotional Expression and Listener Age
}

\author{
Maria Nella Carminati (mcarmina@techfak.uni-bielefeld.de) \({ }^{1,3}\)
}

\author{
Pia Knoeferle (knoeferl@cit-ec.uni-bielefeld.de) \({ }^{1,2,3}\) \\ 1 SFB 673 "Alignment in Communication" \\ 2 Cognitive Interaction Technology Excellence Center \\ 3 Department of Linguistics \\ Universitätsstr. 25, Bielefeld University, 33615 Bielefeld, Germany
}

\begin{abstract}
We report two visual-world eye-tracking experiments that investigated how and with which time course emotional information from a speaker's face affects younger ( \(N=32\), Mean age \(=23\) ) and older \((N=32\), Mean age \(=64)\) listeners' visual attention and language comprehension as they processed emotional sentences in a visual context. The age manipulation was aimed at testing predictions by socioemotional selectivity theory of a positivity effect in older adults. After viewing the emotional face of a speaker (happy or sad) on a computer display, participants were presented simultaneously with two pictures depicting opposite-valence events (positive and negative; IAPS database) while they listened to a sentence referring to one of the events. Participants' eye fixations on the pictures while processing the sentence were enhanced when the speaker's face was emotionally congruent with the sentence/picture compared to when it was not. The enhancement occurred from the early stages of sentence-reference disambiguation; importantly, it was modulated by age, in that for the older adults it was more pronounced with positive faces, and for the younger ones with negative faces. These findings demonstrate for the first time that emotional facial expressions, similarly to previously studied speaker cues such as eye gaze and gestures, are rapidly integrated into sentence processing. They also provide new evidence for positivity effects in older adults in online incremental situated sentence processing.
\end{abstract}

Keywords: sentence processing; visual-world paradigm; emotional processing; speaker cues; positivity effect; facial expressions

\section*{Visual Context Effects on Language Processing}

The study of context effects on language processing has been a major topic of investigation in psycholinguistic research, and since the development of the visual world paradigm in the mid-nineties, psycholinguists have had at their disposal a powerful tool to investigate a potentially rich source of context effects on language processing, that of the visual context. Among other things, findings from this research have demonstrated how information such as an object's size, color, or shape, depicted clipart events, realworld action events, action affordances, and the spatial location of objects are all rapidly integrated during sentence comprehension and can affect a listener's visual attention
while processing sentences (for a recent review, see Huettig, Rommers, \& Meyer, 2011).
In recent years the scope of research on the languagevision interaction has been extended to more complex and subtle aspects of naturalistic, visually-situated language events, such as dialogue interactions. One topic has been how visually-perceivable speaker-based cues, for example, speaker gaze and gestures, affect language processing. Results suggest that a speaker's gaze is incrementally integrated into language processing by listeners (e.g., Hanna \& Brennan, 2007).

Another potentially powerful cue that could be used by a listener is a speaker's emotional facial expression. The question of how such a cue is used in language processing is particularly relevant to current psychological research, especially in light of the recent surge in interest in embodied and situated cognition, and the increasingly available evidence supporting a close interaction between emotions and language (e.g., Havas, Glenberg, \& Rinck, 2007). However, to the best of our knowledge there is todate no study that has examined effects of a speaker's facial expression on spoken sentence comprehension. Additionally, evidence for visual context effects in sentence processing comes almost exclusively from studies with young adults (ca. 19-31 years). By contrast, the extent to which visual context affects sentence comprehension in older adults is less clear. The present research addresses these two open issues in two eye-tracking experiments that examined (a) the time course with which a speaker's emotional facial expressions can influence a comprehender's visual attention to target pictures during spoken sentence comprehension; and (b) the nature of this influence in young versus older adults.

\section*{Emotion Processing and Emotional Priming}

Ekman's (1972) proposal of a set of six basic universal emotions associated with distinct facial emotional expression configurations (happiness, sadness, fear, anger, disgust, and surprise) has been widely tested over the years and assumed by many scholars in the field of emotion research (e.g., Lundqvist, Flykt, \& Öhman, 1998). Within this view, the basic facial expressions are associated with a
distinctive meaning, so they could, in principle, be used by a speaker to strengthen the meaning of her utterances. There is evidence that emotional faces such as happy or angry ones, are attended to faster and are processed more deeply than neutral ones (Palermo \& Rhodes, 2007). Generally, the same attention advantage enjoyed by emotional faces (compared to nonemotional, neutral ones) is found for emotional stimuli in different modalities, for example, emotional words, pictures and sounds (e.g., Hermans, De Houwer, \& Eelen, 2001).

Not only do emotional stimuli attract more attention and are remembered better than corresponding neutral ones, they can also influence how other stimuli (e.g., words, pictures) are processed. This influence has been demonstrated in emotional priming studies (Fazio, 2001), where responses to a target stimulus are facilitated (i.e., faster) when prime and target have the same emotional valence (e.g., positivepositive, negative-negative), compared to when they have opposite valence. Interestingly, priming occurs not only when prime and target belong to same modality and category (e.g., when they are both faces, pictures or words) but also across modalities. e.g., from a picture to a face (Carroll \& Young, 2005, Expt 2), or from a picture or facial expression to a word (Carroll \& Young 2005, Expt 1 and 4).

\section*{Emotional Priming of Sentences: Current Study}

With regard to the issue of whether a speaker's emotional facial expression can influence not just lexical but also sentence processing, the just-mentioned findings on emotional priming from faces to words suggest that it should: Just as the perception of a happy face (the prime) produces a faster response to a positive (vs. negative) target word, so a smile on a speaker's face might facilitate a listener's processing of a positive (vs. negative) sentence. To our knowledge, no research has so far investigated the emotional priming of whole sentences (as opposed to isolated words) using emotional facial expressions. In the current study, we used the visual world paradigm to examine the time course of emotional priming in sentence processing. We employed a design typical of many visualworld experiments on sentence comprehension: Participants listened to sentences relating to visual material displayed on a computer screen (see Fig.1, for the sequence of events in an experimental trial). Before hearing the sentences, our participants saw either a smiling or a sad face (see Fig. 1, Display 1). They were told that this was the face of the speaker of the ensuing sentence (thus simulating a speakerhearer scenario). Then two emotional pictures from the International Affective Picture System database (IAPS, Lang, Bradley, \& Cuthbert, 1999), one positive and one negative, were displayed side by side on the screen. After 1500 ms the sentence was played and referenced either one or the other picture; accordingly, the sentence also had a positive or negative emotional content (see Fig. 1, Display 2). Thus, the speaker's facial expression could be emotionally congruent or incongruent with the sentence. Participants' eye movements to the display on the monitor
while they listened to the sentence were recorded. In line with the usual findings from the visual world paradigm, when participants begin processing the sentence, we expect them to look at the target (the IAPS picture described in the sentence) from the time it becomes clear which picture the sentence is about (i.e., a sentence effect).


Figure 1: Sequence of events in an experimental trial
However, for us the more interesting question is whether and how the facial prime affects (i.e., primes) the processing of the sentence, in other words, the face x sentence interaction. In line with findings from emotional priming research, we expect facilitation when the prime (i.e., the face) is emotionally congruent with the target (i.e., the sentence/picture) compared to when prime and target are incongruent. In our experiment the dependent variable is fixations on the pictures while incrementally processing the sentence; so we expect that looks to the target picture should be facilitated when the emotional face is valencecongruent (vs. incongruent) with the sentence. This facilitation should be reflected in more and longer fixations to the target picture in congruent than in incongruent conditions (cf. Arai, Van Gompel, \& Scheepers, 2007).

Furthermore, the timing of this facilitation is of particular interest to us, as it would reveal details about the time course of integrating emotion information into language processing. Earlier findings (see above) suggest that emotional information enjoys privileged attention, so this would predict that facilitation effects should occur from the early stages of processing the sentence. Alternatively, considering the specifics of our experimental task, facilitatory effects may not surface until later or not occur at all during the processing of the sentence. This is because for facilitation to take place perceivers need to integrate cues from the visual, linguistic and emotional modalities and this may be a complex task to perform on the fly. In addition to facilitating the processing of the sentence itself (face \(x\) sentence interaction), the face may affect the fixations that listeners make on the pictures independently of the sentence
valence. This would be reflected in a preference to look at the picture which is emotionally congruent with the face, i.e., a face-picture congruence effect. To be triggered, this face-picture congruence effect does not require linguistic input from the sentence (but only information from the face and the pictures), so it could occur earlier, before sentence disambiguation, as well as later, after disambiguation. Note that a face \(x\) sentence interaction, which is the effect of primary interest to us, cannot be reduced to a face-picture congruence effect, as it requires the additional input from the sentence to occur.

\section*{Emotion Processing in Younger and Older Adults}

The age group manipulation in our study was inspired by research showing that emotion processing changes across the life span (for a review, see Ruffman et al., 2008). According to socio-emotional selectivity theory (Mather \& Carstensen, 2003), as people grow older, they realize that their time is limited and focus more on emotionallysatisfying experiences in the present moment. This change arguably leads to the so-called 'positivity effect', observed in studies where young and older adults were compared on emotional processing. For example, when presented with pairs of pictures (a neutral face, and a positive or negative face), older people spent less time inspecting the negative than positive face; i.e., they displayed an attentional bias away from the negative and towards the happy expression. Younger people, by contrast, showed no preference (Mather \& Carstensen, 2003), or preferred negative faces (Isaacowitz et al., 2006). Positivity effects have also been found in the recall of pictures and facial expressions (e.g., Mather \& Carstensen, 2003), or of long-term life events (Kennedy, Mather, \& Carstensen, 2004). In recent years, researchers have discussed the proper characterization of the positivity effect and the experimental conditions under which it can be observed. This has led to a broadening of the definition of the effect, which now includes, not only an increased focus on positive compared to negative information in older versus younger adults, but also a reduced focus on negative information in older adults (see especially Langeslad \& van Strien, 2009; Scheibe \& Carstensen, 2010).

In light of this, given that our study involves the processing of emotional faces and emotional pictures and sentences, we should see differences in the way younger and older adults integrate the information from a negative or positive face with the processing of a negative or positive target sentence and corresponding picture (i.e., a face \(x\) sentence x age interaction). A prediction is that older people should find it easier to integrate a positive face with a positive sentence than a negative face with a negative sentence (i.e., facilitation only for positive sentences, or greater facilitation for positive than negative sentences). For younger people, on the other hand, one may expect equal facilitation for positive and negative sentences, only facilitation for negative, or greater facilitation for negative than positive sentences. Similar modulations by age are predicted for the face-picture congruence effect.

\section*{Methods}

\section*{Participants}

Thirty-two older ( \(60-72\) years, \(M=64.37, S D=3.57\) ) and 32 younger ( \(19-29\) years, \(M=22.90, S D=2.73\) ) adults took part in the experiment in return for a monetary reward; all gave informed consent.

\section*{Materials}

Materials consisted of photographs of emotional facial expressions, emotional pictures and auditorily presented sentences. There were 28 experimental and 56 filler items. Each experimental item consisted of a facial expression (happy/sad), a display showing a positive and a negative picture taken from the IAPS database (Lang et al., 1999) and a sentence describing either the positive or negative picture. The emotional faces were selected from 15 sets of Bielefeld-University student portraits, each set depicting a neutral, a sad and a happy expression. From these sets we selected the 4 best sets ( 2 male, 2 female) based on the results of a valence-rating study ( \(N=18\) ).

The positive and negative images were selected on the basis of the valence ratings in Lang et al., 2008, (negative images: range \(2.42-5.07, M=3.46, S D=1.69\); positive images: range \(5.51-8.22, M=7.19, S D=1.55\) ). Arousal scores of negative and positive images were similar (paired t-test \(t(27)=-.84, p=.41)\).

The sentences for each of the two images of the 28 experimental item-picture pairs fulfilled constraints specific to length, structure and content. All started with an introductory main clause containing a verb of opinion in the first person singular (e.g., I think/believe/am of the opinion that...). This was followed by a subordinate clause about the event depicted in one of the two IAPS images of an item. The subordinate clause contained a subject noun phrase (N1), followed by an object noun phrase (N2), an adverb (Adv) and a final finite verb (Verb). Examples of the positive and negative sentence for an item are given in Figure 1. Care was also taken to match the sentences, so far as possible, by lemma frequency of nouns and adverbs, using frequency counts from the CELEX database (Baayen, Piepenbrock, \& Gulikers, 1995).

The sentences were recorded by 4 native speakers of German, two female and two male, and the speakers assigned to the faces, with 1 male being used for half of the experimental items (14) and 1 female for the other half (14). The sound files of the two experimental sentences associated with an experimental picture pair were edited using professional sound editing software, to ensure that the onsets of the critical words ( \(\mathrm{N} 1, \mathrm{~N} 2\), Adverb) occurred exactly at the same point in time from sentence start in the positive and corresponding negative sentence (to achieve this, pauses were shortened or between-word breaks were lengthened slightly as necessary). The combination of the experimental faces, pictures and sentences yielded a 2 (Face: positive vs. negative) x 2 (Picture: positive vs. negative) x 2 (Sentence: positive vs. negative) design.

For the 56 distractor item picture-pairs, we constructed a sentence that matched one of the two IAPS pictures. The content of half of the distractor sentences (28) was neutral, while 14 contained at least one positive word (e.g., The talented artist is drawing the nice portrait) and the remaining half one negative word (e.g., It is obvious that today the weather will be unbearable). The 56 filler items further differed from the experimental items as follows: the facial expression was either neutral ( 28 items), positive ( 14 items) or negative ( 14 items); both IAPS images had mid-range valence (3.5-6.5); there was only 1 sentence per filler item.

\section*{Procedure}

The experimental session started with the collection of demographic details from the participants, and with the administration of some cognitive tests and of a mood questionnaire. Eye movements were recorded using an SR Research Eyelink 1000 Desktop head-stabilised eye tracker (SR Research, Mississauga, Ontario, Canada). Participants were told that the study investigated language comprehension in relation to a visual display: They would first see the face of a person who was thinking about something and was about to speak, and after that they would hear him/her utter a sentence which described one of the two pictures shown on the screen. The task was to look, listen and understand the sentence, and decide whether the valence of the face matched the valence of the sentence ("Does the face match the sentence?") by pressing one of two buttons. The sequence of events in an experimental trial is illustrated in Fig. 1.

\section*{Analyses and Results}

The data of interest were the participants' fixations on the pictures during sentence processing, i.e., during the time listeners inspect Display 2 (see Fig. 1). Because the initial part of the sentence (cf. Fig. 1, "Ich meine dass") was neutral between the negative and positive sentence, disambiguation towards the positive or negative picture occurred from the initial NP (N1) of the embedded sentence onwards (cf. Fig. 1, "die Mechaniker/die Vorstadtkinder"). Therefore any possible facilitation in the processing of the sentence (face \(x\) sentence interaction) due to having seen a congruent face can only be expected to occur after N1 onset.

By contrast, a face-picture congruence effect (i.e., looks to the pictures as a function of prime face) can occur both before and after sentence disambiguation. We thus defined two time periods: the Post-N1 (onset) region (from the onset of N1 until sentence end (average duration \(4016 \mathrm{~ms}, S D=\) 456) and the Pre-N1 (onset) region (including, in addition to the initial, neutral part of the sentence, the last 1200 ms of the \(1500-\mathrm{ms}\) picture preview period, for a total duration of 3000 ms ). Because our main focus is the face x sentence interaction (and its possible modulation by age), we will first present the findings for the post-N1 onset region.

The measure we used to analyze fixations on the pictures is the mean \(\log\) gaze probability ratio, i.e., the \(\log\) of the ratio of the probability of looking at the negative picture
divided by the probability of looking at the positive picture ( \(\ln (p\) (neg picture) \(/ p\) (pos picture)). This measure expresses the strength of the visual bias towards the negative versus positive picture. It is particularly suited for eye tracking data analyses with parametric tests (such as ANOVAs) because it violates neither independence nor homogeneity of variance assumptions (cf. Arai et al., 2007). The log ratio is symmetrical around zero: A positive log ratio indicates more looks to the negative than the positive picture; a negative \(\log\) ratio indicates more looks to the positive than the negative picture; and a value of zero means the two pictures get an equal number of looks.

Fig. 2 (a)-(b) plots the time course of fixations for the Post-N1 onset region for the two age groups. These graphs are based on log gaze probability ratios (henceforth 'log ratios') computed on successive \(20-\mathrm{ms}\) time slices.


Figure 2: Mean log gaze probability ratios for young and older participants in the Post-N1 onset region, from the onset of N 1 .

In Fig. 2 the sentence effect can clearly be seen in the two sets of lines moving apart from about 500 ms after the onset of N1: The red lines (for the two negative sentence conditions) rise steadily above zero, indicating an increasing preference for the negative picture, while the black lines (for the positive sentence conditions) go in the opposite direction, indicating an increasing preference for the positive picture in these conditions.

A face-priming effect on the processing of the sentence (i.e., the facilitatory effect occurring from having seen a sentence-congruent emotional face) emerges in the relative distance between the solid and the dotted line of each sentence condition: If having seen a face of the same valence as the sentence facilitates sentence processing, the congruent condition (solid line) should be associated with a greater absolute value than the incongruent condition (dotted line).

The log-ratio means for the post-N1 region were submitted to \(2 \times 2 \times 2\) (Face x Sentence x Age) repeatedmeasures ANOVAs with participants and items as random effects. The ANOVAs revealed a significant effect of facepicture congruence ( \(\mathrm{s} \ll .001\) ), with a negative picture preference when the face was negative ( \(M=.15\) ) and a positive picture preference when the face was positive ( \(M=\) -.11 ). This effect was not modulated by age (both \(F\) 's \(<1\) ).

As expected, there was a highly significant sentence effect ( \(p \mathrm{~s}<.001\) ): When the sentence was negative, there was a preference for looking at the negative picture and the opposite was true when the sentence was positive ( \(M \mathrm{~s}=\) 1.65 vs. - 1.61). Importantly, the sentence effect was significantly modulated by age ( \(p \mathrm{~s}<.002\) ). This Sentence x Age interaction is due to the fact that older adults, when hearing a negative sentence, look less at the negative picture (vs. the positive one) than the younger adults; in other words, they are less "responsive" to the negative sentence than the younger group.

Crucially for our experimental hypotheses, the 3-way Face x Sentence x Age interaction was fully significant by participants ( \(p 1=.025, p 2=.13\) ). For our hypotheses, this interaction reflects the facilitating effect of the face on the processing of the sentence and the modulation of this effect by age. Post-hoc pairwise comparisons on participants and items means of the individual groups (i.e., for each age group) compared the two negative sentence conditions and the two positive sentence conditions (Bonferroni correction for 4 comparisons, \(p=05 / 4=.0125\) ). These comparisons can tell us if younger and older participants show different sensitivities to the negative or positive prime face during the processing of the sentence. In the comparisons for the younger participants, only the difference between the two negative sentence conditions was significant ( \(p \mathrm{~s}<.02\) ), while the corresponding comparisons for the older adults yielded only a significant difference between the two positive sentence conditions ( \(p \mathrm{~s}<.02\) ).

Thus, for younger participants a negative prime face (vs. a positive one) significantly enhanced looks to the negative picture during the processing of a negative sentence. By contrast, a positive face had no enhancing effect on younger adults' processing of a positive sentence. For the older group however, the opposite pattern emerged: A negative face had no effect on the processing of a negative sentence, but a positive (vs. negative) face elicited more looks to the positive picture when the positive sentence was processed.

To assess the time course of this facilitation, we performed pairwise comparisons (similar to the ones
reported above for the whole Post-N1 region) on mean log ratios for the individual word regions, i.e., N1, N2, Adverb and Verb (see Fig. 2). For the N1 region, these comparisons yielded a similar pattern of results as in the previous analyses, i.e., older participants showed a facilitation from the positive prime face in the positive sentence conditions ( \(p s<.01\) ), but not in the negative sentence conditions. By contrast, young participants showed significant facilitation in the negative sentence condition ( \(p s<.002\) ), while in the positive sentence condition facilitation was fully significant only in the item analysis ( \(p 1=.06 ; p 2=.02\) ). The only other (nearly) significant comparison occurred for the positive sentences in the adverb region ( \(p \mathrm{~s}<.05\) ): For older adults a positive (vs. negative) face, facilitated the processing of a positive sentence, but a negative face was of no advantage in processing a negative sentence. The fact that results were significant in the early, N1 region for both age groups suggests that the integration of the visual context with facial and linguistic information occurs early and that the time course of this integration does not substantially differ between the two age groups.

In the pre-N1 onset region, the ANOVA analyses on the \(\log\) ratios revealed a face-picture congruence effect ( \(p \mathrm{~s}>\) .02), not modulated by age: The negative picture was fixated longer when the face was negative and the positive picture, when the face was positive. There was also a significant picture effect, with the negative picture attracting overall more looks than the positive one ( \(p s<.02\) ). However, this general bias for the negative picture was weaker for the older participants (the interaction with age was marginally significant in the participants' analysis ( \(p 1=.069 ; p 2=.18\) ).

\section*{Discussion and Conclusion}

These eye-tracking results are important for the following reasons. First, they demonstrate for the first time that priming from an emotional face occurs during sentence comprehension in a visually-situated task (i.e., when language is about objects and actions in the visual context). Moreover, priming effects were found from the early stages of sentence-reference and valence disambiguation (i.e., N1), showing that the seemingly complex integration of visual information from an emotional face, a picture and a sentence happens rapidly and without particular effort. Previous research in visually-situated comprehension tasks has demonstrated that speaker-based information such as gaze is rapidly integrated into sentence processing (e.g., Hanna \& Brennan, 2007; Knoeferle \& Kreysa, 2012). Importantly, our results provide evidence that a speaker's emotional facial expression also has a rapid influence on sentence interpretation.

Crucially also, our results provide new evidence for age differences in the processing of emotional information. All of the age-based modulations that we observed are compatible with a positivity effect, i.e., either an increased focus on positive compared to negative information by older versus younger adults, or a reduced focus on negative information by older adults (Langeslad \& van Strien, 2009;

Scheibe \& Carstensen, 2010). The fact that positivity effects were found in the early stages of sentence processing using a highly time-sensitive methodology such as eye tracking has also implications for the question of the mechanisms underlying the positivity effect, and the level(s) of processing at which these mechanisms operate.

A central tenet of socioemotional selectivity theory is that emotion regulation improves with age and that the positivity effect occurs because older people are capable (consciously or unconsciously) to selectively regulate their emotions in order to enhance positivity and well-being. According to this view, the positivity effect should be strongest in tasks and situations that require controlled processing with associated exertion of cognitive effort, and less so in tasks that measure automatic or initial processing (Scheibe \& Carstensen 2010). Although evidence from several studies suggests that positivity effects require deliberate use of mood regulation strategies to occur (e.g., Isaacowitz, Toner, \& Neupert, 2009), recent evidence has shown that controlled processing and cognitive effort are not necessary to trigger positivity effects in older adults (e.g., Allard, Wadlinger \& Isaacowitz, 2010). We suggest that the positivity effects found in early processing in the eye tracking measures of our experiment are also more likely to be a result of an early and non-strategic emotion processing mechanism.

\section*{Acknowledgments}

This research was funded by the German Research Foundation (DFG) within the SFB-673 Project. We thank Katja Glados and Katja Münster for their help.

\section*{References}

Allard E.S., Wadlinger H.A. \& Isaacowitz, D.M. (2010). Positive gaze preferences in older odults: Assessing the role of cognitive effort with pupil dilation. Aging, Neuropsychology, and Cognition, 17.3, 296-311.
Arai, M., Van Gompel, R.P.G., \& Scheepers, C. (2007). Priming ditransitive structures in comprehension. Cognitive Psychology, 54, 218-250.
Baayen, R.H., Piepenbrock, R., \& Gulikers, L. (1995). The CELEX lexical database [CD-ROM]. Philadelphia: University of Pennsylvania, Linguistic Data Consortium.
Carroll, N.C., \& Young, A.W. (2005). Priming of emotion recognition. Quarterly Journal of Experimental Psychology, 58A, 1173-1197.
Ekman, P. (1972). Universals and cultural differences in facial expressions of emotion. In J. Cole (Ed.), Nebraska Symposium on Motivation, 1971 (Vol. 19). Lincoln: University of Nebraska Press.
Fazio, R.H. (2001). On the automatic activation of associated evaluations: An overview. Cognition and Emotion, 15, 115-141.
Hanna, J., \& Brennan, S. (2007). Speakers’ eye gaze disambiguates referring expressions early during face-to-
face conversation. Journal of Memory and Language, 57, 596-615.
Havas, D.A., Glenberg, A.M., \& Rinck, M. (2007). Emotion simulation during language comprehension. Psychonomic Bulletin and Review, 14, 436-441.
Hermans, D., De Houwer, J., \& Eelen P. (2001). A time course analysis of the affective priming effect. Cognition and Emotion, 15, 143-165.
Huettig, F., Rommers, J., \& Meyer, A.S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. Acta Psychologica, 137, 151-171.
Kennedy Q., Mather M., \& Carstensen L.L. (2004). The role of motivation in the age-related positivity effect in autobiographical memory. Psychological Science, 15, 208-214.
Knoeferle, P. \& Kreysa, H. (2012). Can speaker gaze modulate syntactic structuring and thematic role assignment during spoken sentence comprehension? Frontiers in Psychology, 3:538.
Isaacowitz, D.M., Wadlinger, H.A., Goren, D., \& Wilson, H.R. (2006). Selective preference in visual fixation away from negative images in old age? An eye tracking study. Psychology and Aging, 21, 40- 48.
Isaacowitz, D.M., Toner, K., \& Neupert, S.D. (2009). Use of gaze for real-time mood regulation: Effects of age and attentional functioning. Psychology and Aging, 24, 989994.

Langeslag, S.J., \& van Strien, J.W. (2009). Aging and emotional memory: The co-occurrence of neurophysiological and behavioral positivity effects. Emotion, 9, 369-377.
Lundqvist, D., Flykt, A., \& Öhman, A. (1998). Karolinska directed emotional faces. Stockholm: Karolinska Institute and Hospital, Section of Psychology.
Mather, M., \& Carstensen, L.L. (2003). Aging and attentional biases for emotional faces. Psychological Science, 14, 409-415.
Palermo R., \& Rhodes G. (2007). Are you always on my mind? A review of how face perception and attention interact. Neuropsychologia, 45, 75-92.
Ruffman, T., Henry, J.D., Livingstone, V., \& Phillips, L.H. (2008). A meta-analytic review of emotion recognition and aging: Implications for neuropsychological models of aging. Neuroscience \& Bio-behavioral Reviews, 32, 863881.

Scheibe S., \& Carstensen L.L. (2010). Emotional aging: Recent findings and future trends. Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 65B.2, 135-44.
Trueswell, J.C., Sekerina, I., Hill, N., \& Logrip, M.L. (1999). The kindergarten-path effect: studying online sentence processing in young children. Cognition, 73, 89134.

\title{
How to present exemplars of several categories? Interleave during active learning and block during passive learning
}

\author{
Paulo F. Carvalho (pcarvalh@indiana.edu) \\ Department of Psychological and Brain Sciences, Indiana University \\ 1101 E. Tenth St., Bloomington, IN 47405 USA \\ Robert L. Goldstone (rgoldsto@indiana.edu) \\ Department of Psychological and Brain Sciences, Indiana University \\ 1101 E. Tenth St., Bloomington, IN 47405 USA
}

\begin{abstract}
Research on how information should be presented during inductive category learning has identified both interleaving of categories and blocking by category as beneficial for learning. Previous work suggests that this mixed evidence can be reconciled by taking into account within- and betweencategory similarity relations. In this paper we present a new moderating factor. One group of participants studied categories actively, either interleaved or blocked. Another group studied the same categories passively. Results from a subsequent generalization task show that active learning benefits from interleaved presentation while passive learning benefits from blocked presentation.
\end{abstract}

Keywords: interleaving; blocking; learning; comparison

\section*{Introduction}

Can the method with which information is presented substantially affect learning? The answer seems to be "yes." Changing the way with which information is presented not only changes what is learned (Schyns, Goldstone, \& Thibaut, 1998) but also how well it is learned (Goldstone, 1996). One example is the order in which instances are presented in a study session and the effect this has for inductive learning. Kornell and Bjork (2008) demonstrated that if participants are given study examples of paintings from several artists' interleaved, participants' later memory and generalization is substantially improved when compared to presenting each artist in a separate block.

The advantage of interleaving over blocking for inductive learning has been repeatedly shown in recent years. Interleaving of categories has been shown to improve learning of naturalistic materials for both young and older adults (Kornell, Castel, Eich, \& Bjork, 2010; Wahlheim, Dunlosky, \& Jacoby, 2011), as well as flashcard learning (Kornell, 2009). It has also been demonstrated in children (Vlach, Sandhofer, \& Kornell, 2008; Vlach \& Sandhofer, 2012).

Notwithstanding the clear benefit of interleaving in some situations, there have also been demonstrations of the advantage of blocking for category learning. For example, Goldstone (1996) presented participants with complex images consisting of 20 line segments. There were two conditions: frequent alternation of categories (interleaving) and infrequent alternation (blocking). The results showed that participants were better at learning the categories in the
infrequent alternation condition. The author associates this advantage with the relative difficulty in finding the common features shared by members of each category (for further evidence of blocking advantages using different kinds of tasks and stimuli see Kurtz \& Hovland, 1956; Whitman \& Garner, 1963).

Given this mixed evidence about the best way to sequentially present information for optimal learning, an important question is: what conditions yield an advantage for interleaving compared to blocking?

This question has received some attention in recent years. For instance, Carvalho and Goldstone (2012) showed that when studying low similarity categories, blocked presentation resulted in improved subsequent generalization performance. This pattern was reversed for high similarity categories (for similar results with category discriminability, see Zulkiply \& Burt, 2013).

Carvalho and Goldstone (2012; see also Goldstone, 1996) have proposed that interleaving categories allows participants to identify the features that distinguish between the categories, while blocked presentation promotes the identification of features that are characteristic among stimuli within a single category. This dichotomy is the result of the same principle: the selective emphasis of categorization-relevant features during comparison of sequentially presented objects.

In the case of interleaved presentation, differences between objects belonging to different categories will be emphasized while for blocked presentation, similarities among objects belong to the same category will be emphasized. This same process will result in improved or hindered learning depending on whether similarities or differences need to be learned. One possible way in which category learning could move from an emphasis on differences towards an emphasis on similarities is by changing the similarity relations within and between categories (Carvalho \& Goldstone, 2012; Zulkiply and Burt, 2013).

However, the characteristics of the categories being studied are not the only factors that can have an influence on how sequential comparison impacts learning. In theory, any property of the learning situation that changes attentional constraints could have similar impact by changing the task
demands from an emphasis on similarities to an emphasis on differences.

Inductive learning can take place in several ways. One such way is active category learning. In this kind of learning task, participants actively try to categorize never-beforeseen stimuli into one of the categories provided. Participants are then given feedback on the accuracy of their responses. Learning takes place via feedback-informed update of perceptual, attentional and decisional processes.

Another kind of inductive learning task can be referred to as passive learning. In this kind of situation, participants study category exemplars along with their correct category assignment. This task can be considered 'passive' in the sense that participants do not actively make responses during learning, and no feedback is provided via which a participant could adjust their judgments.

These two tasks differ in a number of educationally relevant aspects (e.g., motivation, engagement, etc.) but also in their cognitive aspects. It could be argued that if participants actively study the categories, emphasis will be placed on finding the differences between the categories. This is perhaps the most obvious way one can learn to discriminate As from Bs and achieve good performance. By contrast, in passive learning, participants are not tasked with learning how to discriminate between the categories. They are explicitly given the category assignments for the stimuli. Instead participants may search for features that characterize each category or the similarity amongst objects in each category.

Put another way, if participants passively study the stimuli along with their correct category assignments, their self-imposed task may be to identify the features that characterize that category (e.g., that all the 'Zups' have a similar nose shape). By contrast, if participants are not given the category assignment but instead have to try to categorize the stimulus and only then receive feedback, they may focus on finding differences between objects of different categories (e.g., that 'Zups' are round and 'Rikes' are squares; see Markman \& Ross, 2003; Yamauchi \& Markman, 2000).

Following Carvalho and Goldstone's (2012) proposal, combining interleaved study with active study will be beneficial for category learning because they are both compatible with focusing on features that differentiate between the categories being acquired. Likewise, combining blocked study with passive learning should be beneficial because they are consistent in leading participants to find similarities that are useful for successfully learning each category in isolation.

In this paper we aim to extend previous evidence for a comparison and attentional mechanism as the unifying processes behind both blocked and interleaved study, by manipulating the properties of the study session that affect attention allocation. One group of participants completed a passive study session associated with both interleaved and blocked presentations. The other group of participants
completed an active study session, while keeping all other aspects of the task constant between the two groups.

\section*{An Experiment}

In this experiment, participants studied a set of six categories, three presented interleaved and the remaining three presented blocked. Critically, for one group of participants, the study session was set up as a passive learning task. Participants studied each object for a short period of time during which the correct category assignment was also presented on the screen. For the other group of participants, the categories were studied in an active learning task. Both groups completed the same generalization task afterwards.

\section*{Method}

Participants Eighty-one undergraduate students at Indiana University volunteered to participate in return for partial course credit. Data from seven participants in the passive learning group were excluded from analyses due to failure to repeat the label of the object just presented on more than half of the total number of study trials (see bellow for details). All participants in the active learning group reached the criterion of \(34 \%\) correct responses during study and their data were kept for analyses.


Figure 1: Examples of one exemplar of each of the 6 categories used. The top row constitutes one group of categories and the bottom row another group.

Apparatus and Stimuli In this experiment, stimuli were "Fribble" objects (Williams, 1998). Three of the categories were composed of very similar objects differing diagnostically only in one of their parts (see top panel in Figure 1). The other three categories were also very similar and differed diagnostically from each other only in one of their parts, however, they were substantially different from the other three categories (see bottom panel in Figure 1). Random variation existed in each of the categories but was the same across the three categories in each group.

Each category was given a unique label that perfectly predicted the presence of the unique feature that defined that category. At the start of the experiment, one label was
randomly picked for each category from the pool: beme, kipe, vune, coge, zade, and tyfe (Hendrickson, et al., 2012).

Design and procedure This experiment had two conditions manipulated within-subjects (schedule of presentation: blocked category learning and interleaved category learning), and two conditions manipulated between-subjects (study type: active vs. passive). Each of these four conditions was composed of a study and test phase.

Study phase For the passive study group, during this study phase, participants were presented with a stimulus in the center of the screen along with the correct category assignment above the object for 2.5 s . Immediately after the presentation of the stimulus, three buttons were shown on the screen corresponding to the three possible category names for that study session. The participants' task was to press the button corresponding to the category of the object they just saw. This task was introduced to ensure participants' attention to the task and to equate the active and passive learning situations for the presence of a motor response. However, note that in this condition, participants simply needed to repeat the category they had just seen. There was no need for participants to learn a categorization rule. The mapping between the position of the buttons on the screen and the label was randomly shifted each trial.

For the active learning group, participants were presented with a stimulus for 500 ms . without its label. After the stimulus was removed, three buttons were shown on the screen and the participant had to choose the category assignment for that stimulus. After the participant's response, the stimulus along with the correct category assignment was shown on the screen for 2 s . The mapping between the position of the buttons on the screen and the label was also randomly shifted each trial.

For both groups, a 1000 ms inter-trial interval followed the trial and then the next trial began. In the blocked condition, the categories presented alternated \(25 \%\) of the time while in the interleaved condition they alternated \(75 \%\) of the time. That is, in the interleaved condition, the probability of an object being followed by an object of the same category was low, whereas for the blocked condition this probability was high. We used this probabilistic approach rather than creating purely interleaved or blocked conditions in order to diminish the possibility that participants noticed the pattern of alternation in responses, which would affect categorization accuracy. Furthermore, if a purely blocked condition had been used there would be no way to guarantee participants' attention to the task, as there would be no uncertainty as to the correct categorization. This approach has been used before in similar tasks with successful results (Carvalho \& Goldstone, 2012; Goldstone, 1996).

Each study phase was composed of 4 blocks for both groups of participants and the entire study phase took approximately the same amount of time for each group. Each block had 24 trials (4 exemplars of each category
repeated 2 times each). After the 4th block of study a new set of instructions was presented on the screen and the second phase began. Each participant completed two sets of study and test phases (one for each schedule of presentation).

The two schedule conditions (blocked vs. interleaved) differed only in the frequency of category change during study and the species labels. Which condition was presented first was counterbalanced across participants and the allocation of the stimuli to each category and condition was randomized across participants.

Test Phase This second phase was a generalization task during which 36 stimuli were shown in random order - the 12 blobs participants studied during the learning task and 24 new stimuli. The new stimuli were similar to the studied stimuli, with new instantiations of the unique features. Each stimulus was presented in the center of the screen for 500 ms , after which participants were asked to classify it into one of the species just learned, by pressing a key on the screen. After a 1000 ms inter-trial interval, a new trial would begin. No feedback was provided during this phase. Each test phase followed the respective study phase.

\section*{Results}

We started by analyzing participants' performance during the study phase in the active learning group. As can be seen from Figure 2, participants' performance improves across the task for both the interleaved and blocked conditions.

\section*{Study Phase - Active Learning}


Figure 2: Results from the Study Phase for the Active Learning group. Error bars indicate standard errors of the mean.

A within-subjects ANOVA with Block (1 vs. 2 vs. 3 vs. 4) and schedule of presentation (interleaved vs. blocked) as factors confirmed this interpretation. There is a main effect of Block, \(F(3,111)=139.93, p<.0001, \eta_{G}^{2}=0.36\). No
main effect of schedule of presentation, \(F(1,37)<1\) was found but the interaction between the two variables was also significant, \(F(3,111)=4.15, p=0.008, \eta_{G}^{2}=0.03\), indicating a larger improvement for the interleaved condition compared to the blocked condition.

However, the result of greater interest is how well participants are able to generalize this learning to new stimuli. These analyses will not only allow us to test the effect of interleaving vs. blocking and of passive vs. active learning but, more importantly, the interaction between the two.

\section*{Test Phase}


Figure 3: Main results of the Test Phase for both the Active and Passive groups and for each of the schedule of presentation conditions. Error bars indicate standard errors of the mean

The graph depicted in Figure 3 shows performance in the generalization task for participants in the active and passive study conditions and for each study presentation format. Given the overall high level of performance we began by analyzing possible differences in response time. No main effect or interaction between any of the variables was found for RT.

We then proceeded to analyze performance differences. The most obvious result is an interaction between the study condition and the presentation format. While for participants studying the stimuli actively, interleaving presentation of the objects results in better generalization performance, blocking is better for participants in the passive learning condition.

A mixed ANOVA with presentation schedule (blocked vs. interleaved) as a within-subject factor and study type (passive vs. active) as a between-subjects factor confirms this analysis. There is no main effect of presentation schedule, \(F(1,73)<1\) or study type, \(F(1,73)<1\).

However, the interaction between the two variables was reliable, \(F(1,73)=7.30, p=.04, \eta_{G}^{2}=0.03\).

Performance results for each group and schedule of presentation, sub-divided by studied and novel stimulus, are presented in Table 1. As can be seen, no differences in performance between novel and studied stimuli were found. We repeated the ANOVA analyses with stimulus type (studied vs. novel) as another within-subject factor. These analyses revealed the same critical interaction between schedule of presentation and study type, but no differences in performance for novel and studied stimuli or interaction with any of the other variables.

Table 1: Categorization accuracy in the test phase for both groups and schedule conditions, broken down by type of item (novel vs. studied). Standard deviations are presented in parentheses.
\begin{tabular}{ccccc}
\cline { 2 - 5 } & \multicolumn{2}{c}{ Active } & \multicolumn{2}{c}{ Passive } \\
& Interleaved & Blocked & Interleaved & Blocked \\
\hline \multirow{2}{*}{ Novel } & 0.96 & 0.90 & 0.91 & 0.96 \\
& \((0.13)\) & \((0.21)\) & \((0.20)\) & \((0.13)\) \\
Studied & 0.96 & 0.91 & 0.89 & 0.97 \\
& \((0.13)\) & \((0.18)\) & \((0.26)\) & \((0.11)\) \\
\hline
\end{tabular}

\section*{Discussion}

Determining how to order information so that learners can achieve the best learning outcomes is crucial for effective training. In this work we present further evidence that the way information is ordered impacts learning and that this influence is modulated by whether learning is active or passive.

The results of the experiment presented here show that whether interleaving examples of several concepts or blocking examples by category is beneficial is a function of the training task's implicit demands. More specifically, in a task involving discrimination of the concepts being studied by identifying their differences (the active learning situation), interleaved study results in better performance in a subsequent generalization task. However, if the learning situation involves creating a positive, stand-alone representation of the concepts by identifying the similarities among the instances within each category (the passive learning situation), blocked study benefits performance in the generalization task.

Interestingly, both study conditions result in similar performance during the learning task in the active learning condition. This eliminates the possibility that one study condition is more difficult and results in greater cognitive effort, which is a known important factor contributing to improved learning (Bjork, 1994).

In the generalization task, the interaction between the type of study situation and the schedule of presentation of the
exemplars had an effect for both studied stimuli and novel ones. This is an interesting result, suggesting that study benefits go beyond memorization of the whole exemplars. Very likely, participants succeeded at the categorization task by identifying the single defining parts for each category. These defining parts were instantiated identically for studied and novel objects, explaining why novel objects were not more difficult to categorize. Future research will be needed to assure that the interaction between learning activity and presentation schedule generalizes to other category structures.

Overall these results are consistent with the framework proposed by Carvalho and Goldstone (2012; see also Goldstone, 1996) hypothesizing that participants compare successive objects and update attention to stimulus features as a result of these comparisons.

The role of allocating one's attention during category learning has been highlighted before in different models (Kruschke, 1992; Love, Medin, \& Gureckis, 2004; Minda \& Smith, 2002; Nosofsky, 1986) and the use of eye tracking technology has made it possible to study the patterns of overt trial-by-trial, or even within-trial, attention. For example, Blair, Watson, Walshe, and Maj (2009) have demonstrated that in a categorization task, different stimuli can elicit different patterns of attention allocation to their features. Additionally, previous research has also demonstrated that during category learning, participants often take into account information from only the previous few trials to decide whether a stimulus belongs in one category or another (Jones, Love, \& Maddox, 2006; Jones \& Sieck, 2003; Stewart, Brown, \& Chater, 2002; Stewart \& Brown, 2004; Stewart \& Chater, 2002).

Carvalho and Goldstone (2012) propose that when studying a new exemplar in an inductive learning task, participants compare the properties of that object with the properties they recall from the previous ones. However, learners do not remember all the features from all the objects presented. Instead, when studying a new exemplar, learners weight more heavily the information presented in the immediately preceding instances. If the previous trial consisted of an object in one category and the current trial consists of another object in a different category, participants' attention will be directed towards the differences between the two objects. Conversely, if the two objects come from the same category, learners will attend to similarities between the objects.

This framework can aptly account for the results presented here: passive learning requires attending to similarities, while active learning requires attending to differences. When the presentation order also emphasizes those factors, learning will be facilitated.

Finally, although our results do not directly speak to the importance of active vs. passive learning, it is worth noting that this interaction should be taken into account when deciding whether learners should be given worked examples to study or not.

Indeed, we think one of the most important contributions of the present work is the proposal that when deciding how to structure learning, one needs to take into account the entire learning situation and possible interactions between situational factors. So far, we have demonstrated this for interleaving/blocking benefits relative to the training activity (passive vs. active) and the similarity structure of the categories being studied (similar vs. dissimilar; Carvalho \& Goldstone, 2012).

\section*{Acknowledgements}

This research was supported in part by National Science Foundation REESE grant 0910218 and Department of Education IES grant R305A1100060. PFC was also supported by Graduate Training Fellowship SFRH /BD/78083/2011 from the Portuguese Foundation for Science and Technology (FCT). The authors would like to thank the Percepts and Concepts Lab members for discussion and Spenser Benge for his help with data collection. Stimulus images courtesy of Michael J. Tarr, Center for the Neural Basis of Cognition and Department of Psychology, Carnegie Mellon University,
http://www.tarrlab.org/

\section*{References}

Blair, M., Watson, M. R., Walshe, R. C., \& Maj, F. (2009). Extremely selective attention: Eye-tracking studies of the dynamic allocation of attention to stimulus features in categorization. Journal of Experimental Psychology: Learning, Memory, and Cognition, 35(5), 1196-1206.
Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe \& A. P. Shimamura (Eds.), Metacognition: Knowing about knowing. Cambridge, MA: MIT Press.
Carvalho, P., \& Goldstone, R. (2012). Category structure modulates interleaving and blocking advantage in inductive category acquisition. In C. R. P. Miyake N, Peebles D (Ed.), Proceedings of the 34th Annual Conference of the Cognitive Society. Austin, TX: Cognitive Science Society.
Goldstone, R. L. (1996). Isolated and interrelated concepts. Memory \& cognition, 24(5), 608-28.
Hendrickson, A. T., Kachergis, G., Fausey, C. M., \& Goldstone, R. L. (2012). Re-learning labeled categories reveals structured representations. In N. Miyake, D. Peebles, \& R. P. Cooper (Eds.), Proceedings of the 34th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.
Jones, M., Love, B. C., \& Maddox, W. T. (2006). Recency effects as a window to generalization: separating decisional and perceptual sequential effects in category learning. Journal of experimental psychology. Learning, memory, and cognition, 32(2), 316-32.
Jones, M., \& Sieck, W. R. (2003). Learning myopia: An adaptive recency effect in category learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 29(4), 626-640.

Kornell, N. (2009). Optimizing learning using flashcards: Spacing is more effective than cramming. Applied Cognitive Psychology, 23(9), 1297-1317.
Kornell, N., \& Bjork, R. A. (2008). Learning concepts and categories: is spacing the "enemy of induction"? Psychological science, 19(6), 585-92.
Kornell, N., Castel, A. D., Eich, T. S., \& Bjork, R. A. (2010). Spacing as the friend of both memory and induction in young and older adults. Psychology and aging, 25(2), 498-503.
Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. Psychological review, 99(1), 22.
Kurtz, K. H., \& Hovland, C. I. (1956). Concept learning with differing sequences of instances. Journal of experimental psychology, 51(4), 239.
Love, B. C., Medin, D. L., \& Gureckis, T. M. (2004). SUSTAIN: A Network Model of Category Learning. Psychological Review, 111(2), 309-332.
Markman, A.B., \& Ross, B.H. (2003). Category use and category learning. Psychological Bulletin, 129(4), 592613.

Minda, J. P., \& Smith, J. D. (2002). Comparing prototypebased and exemplar-based accounts of category learning and attentional allocation. Journal of Experimental Psychology: Learning, Memory, and Cognition, 28(2), 275-292.
Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. Journal of experimental psychology. General, 115(1), 39-61.
Schyns, P. G., Goldstone, R. L., \& Thibaut, J. P. (1998). The development of features in object concepts. The Behavioral and brain sciences, 21(1), 1-17; discussion 17-54.
Stewart, N., \& Brown, G. D. A. (2004). Sequence effects in the categorization of tones varying in frequency. Journal of experimental psychology. Learning, memory, and cognition, 30(2), 416-30.
Stewart, N., Brown, G. D. A., \& Chater, N. (2002). Sequence effects in categorization of simple perceptual stimuli. Journal of Experimental Psychology: Learning, Memory, and Cognition, 28(1), 3-11.
Stewart, N., \& Chater, N. (2002). The effect of category variability in perceptual categorization. Journal of Experimental Psychology: Learning, Memory, and Cognition, 28(5), 893.
Vlach, H. A., \& Sandhofer, C. M. (2012). Distributing Learning Over Time: The Spacing Effect in Children's Acquisition and Generalization of Science Concepts. Child development, 83, 1137-1144.
Vlach, H. A., Sandhofer, C. M., \& Kornell, N. (2008). The spacing effect in children's memory and category induction. Cognition, 109(1), 163-7.
Wahlheim, C. N., Dunlosky, J., \& Jacoby, L. L. (2011). Spacing enhances the learning of natural concepts: an investigation of mechanisms, metacognition, and aging. Memory \& cognition, 39(5), 750-763.

Whitman, J. R., \& Garner, W. R. (1963). Concept Learning as a Function of Form of Internal Structure. Journal Of Verbal Learning And Verbal Behavior, 2(2), 195-202.
Williams, P. (1997). Prototypes, exemplars, and object recognition. New Haven, CT: Yale.
Yamauchi, T., \& Markman, A.B. (2000). Inference using categories. Journal of Experimental Psychology: Learning, Memory, and Cognition, 26(3), 776-795.
Zulkiply, N., \& Burt, J. S. (2013). The exemplar interleaving effect in inductive learning: Moderation by the difficulty of category discriminations. Memory \& Cognition, 41(1), 16-27.

\title{
An eyetracking study of children's relational thinking: The role of labels and sustained attention
}

\author{
Paulo F. Carvalho, Catarina Vales, Caitlin M. Fausey \& Linda B. Smith \\ (\{pcarvalh, cvales, cfausey, smith4\} @indiana.edu) \\ Department of Psychological and Brain Sciences, Indiana University, 1101 E. Tenth St., Bloomington, IN 47405 USA
}

\begin{abstract}
Relational match-to-sample is a difficult task for young children. However, it has been shown that either presenting two examples of the relation or adding a label to a single presentation can improve children's performance. The role of labels has been seen as increasing the likelihood of comparing the instances available. In this paper, we present sustained attention as an alternative to this view. Children completed a relational match-to-sample task in different conditions while an eyetracker registered their eye movements. When only one instance was available, children benefited from the addition of a label. This benefit was associated with an overall decrease in switching behavior, indicating greater sustained attention. Moreover, in the absence of a label, children who showed greater sustained attention were able to achieve good performance by the end of the task.
\end{abstract}

Keywords: relational matching; comparison; sustained attention; labels; language and cognition; eyetracking.

\section*{Introduction}

Relational thinking is a fundamental activity of human cognition and everyday experience and might be uniquely human (Gentner, 2003). For example, knowing why both "left shoe goes with right shoe" and "left glove goes with right glove" entail a sameness relation transcends the properties of individual shoes and gloves. This kind of thinking involves going beyond superficial properties of stimuli and noticing the underlying commonalities and differences (Gentner, Rattermann, Markman, \& Kotovsky, 1995). Relational thinking, however, is a developmental feat: early in life, children appear to categorize based primarily on perceptual features, and only begin to attend to the relational properties of the objects after four years of age (Gentner \& Namy, 1999; Loewenstein \& Gentner, 2005).

One typical task used to study relational thinking in children is the relational match-to-sample task (see Figure 1 for an example). In this kind of task, children are presented with an object that instantiates a relational property (the sample) and are then presented with two choices where only one matches the relational property instantiated by the sample. The youngest children in this task do not reliably pick the object that matches the relation depicted in the sample. The research reported here concerns two task manipulations that have been shown to increase relational matching in young children.

\section*{How to promote relational thinking: Compare instances}

Multiple instances If given the right amount of support, children can succeed at the initially difficult task of relational match-to-sample. One way this can be achieved is by presenting multiple examples in the same trial (comparison). For instance, Christie and Gentner (2010) showed 3- and 4 -year old children cards depicting the relation of sameness. Children saw either only one card (solo condition), or two cards simultaneously (comparison condition) or two cards sequentially (sequential condition). Only children who saw two sample cards simultaneously reliably picked the relational match between two choice cards. Multiple instances have been proposed to benefit relational reasoning by encouraging comparison of the instances.

The benefits of multiple instances have been shown many times in children in a variety of tasks (Gentner \& Namy, 1999; Loewenstein \& Gentner, 2001; Namy \& Gentner, 2002; Oakes \& Ribar, 2005; Pruden, Hirsh-Pasek, Shallcross, Golinkoff, 2008; Wang \& Baillargeon, 2008) and also in adults (Gentner, Loewenstein, \& Thompson, 2003; Gick \& Holyoak, 1983). In the context of relational matching, "comparison" has two meanings: (1) the name of the task manipulation of the simultaneous presentation of another object that, although perceptually different, instantiates the same relation and (2) the presumed psychological mechanism that leads to better performance, that is, joint (or temporally close) inspection of the instances which fosters the discovery of deep relational similarities.
Label a single instance Another way to improve children's performance in the relational match-to-sample task is by labeling an original instance.

For example, Christie and Gentner (2007) presented 4and 8 -year old children as well as adults with only one sample instantiating the relation of 'sameness' and then asked participants to pick which of two options matched the sample. When the sample was not labeled, only the adults reliably picked another card instantiating sameness in this condition. However, when a label was added during the presentation of the original sample (e.g., "Look, this is a truffet!'"), children reliably picked the relational match and even adults' performance improved. It has been proposed that labeling benefits performance through comparison, this is, that a label prompts people to compare the original
sample to each choice, and through this comparison they discover relations (Christie \& Gentner, 2007; Gentner \& Namy, 1999; Namy \& Gentner, 2002).

In sum, presenting multiple instances, or labeling an instance, is hypothesized to invite comparison at some point during the relational match-to-sample task and this comparison supports successful relational matching.

\section*{A role for sustained attention?}

The fact that two rather different manipulations - adding another instance and labeling one single instance - help children to discover matching relations should provide insight into a more precise specification of the processes (i.e., of comparison) that limit children's relational comparisons. Our working hypothesis is that each of these manipulations change how children visually inspect instances, perhaps when initially presented or during the difficult step of figuring out what choice to make. One possibility is that when multiple instances are available, children may establish links between the two samples by looking back and forth between them. They may also switch between these samples and the choice options, as they try to make their decision, which may also link the instances and choices and thus reveal the common relations (see Vurpillot, 1968). The process of switching among instances could be the critical behavior that highlights relational similarities between the objects and foster relational choices.

But why would a label encourage switching? The addition of a label to a single sample has been interpreted as inviting just this comparison process and more back-and-forth examination of the sample and choices, resulting in more links between the sample and the options and thus the discovery of the underlying common relation (Christie \& Gentner, 2007; Gentner \& Namy, 1999; Namy \& Gentner, 2002). By fostering this sort of sampling, relational similarities can be discovered and children can successfully choose the relational match. Note that this hypothesized mechanism requires two steps: using a label leads to comparison, which then highlights relational features. The power of labels works only through comparison.

However this is not the only way that labels might work to promote relational matching. Using labels has been shown to improve performance across a great number of tasks other than making relational matches in both adults and children (Lupyan, Rakison, \& McClelland, 2007; Waxman \& Leddon, 2011; Vales \& Smith, 2012). One leading possibility on why labels help is that labeling an object increases sustained attention (see Baldwin \& Markman, 1989; McDuffie, Yoder \& Stone, 2006). Sustained attention is generally good for learning in young children (Smith \& Yu, 2013; Yu \& Smith, 2012) and so may be critical to success in challenging tasks. That is, whenever children face challenges, if they can sustain attention to the relevant stimulus information they may be able to move beyond superficial or salient properties to the underlying structure. The power of labels could come, not
from comparisons in the sense of back-and-forth looking, but from more sustained looking to individual stimuli.

A third issue important to understanding how comparison works concerns how performance changes across trials in the task. The two presented hypotheses concern what happens in a single trial. But relational-matching experiments present children with a series of trials that present, successively, instances of the same relation. By definition, successive presentation does not involve direct comparison, but if children remember what they have seen in previous trials, then comparison to items in memory becomes a factor that may affect either back-and-forth comparison or sustained attention. Critically, past research has shown that there can be accrued effects across trials that influence children's relational matching (Gentner, Loewenstein, \& Hung, 2007; Kotovsky \& Gentner, 1996).

The main goal of the current experiment was to test the competing hypotheses about the power of labels to promote relational matching: Either they increase (switch) or decrease (sustain) the rate that children visually sample available information.

If labels work through sustained attention, then we might also see improved performance when sustained attention is present in other ways. In order to detect this, the experiment includes both Multiple and Single instance conditions. The Single No Label instance condition is critical to our analysis as to the best of our knowledge, no research has reported that young children can successfully match relations when they see only one unlabeled instance on each trial. However, labeling this single instance has been shown to dramatically improve performance.

In the case of unlabeled single instances, children might need to accumulate enough evidence to allow them to understand the relation being instantiated across trials. They might only do so if they show sustained attention. Do children who show sustained attention over several trials reach the same level of success as children who got a single labeled sample?

To answer these questions, we designed a novel eyetracking relational match-to-sample task. To examine how labels and sustained attention matter for relational matching over time, we analyzed children's performance across eight trials. To capture how children sample visual information in different conditions, we used eye-tracking technology. We included a full set of four conditions: multiple or single sample, with or without label. As in previous studies, conditions with multiple exemplars (both labeled and unlabeled) should support relational matching. We included these conditions to show that our novel paradigm replicates well-known effects.

By considering learning over time as well as finer-grained measures of visual sampling, this study offers novel insight into the role of labels in children's relational thinking.

\section*{An Experiment}

\section*{Method}

Participants Fifty-eight children ( \(\mathrm{M}=54\) months, range \(=\) 42-68 months) were randomly assigned to one of four conditions: Multiple presentation without label, Multiple presentation with label, Single presentation without label, Single presentation with label. Twenty additional children were recruited but not included in the final sample due to inappropriate calibration, missing video data, refusal to complete the task or eye tracking data missing for more than half of the total number of samples. Children had no known developmental disorders and were reported to have normal (or corrected to normal) visual acuity. Parental consent was obtained for all participants in compliance with the IRB of Indiana University.
Apparatus and procedure Children were seated approximately 211 cm from a 55 '" LED screen. A freestanding Tobii X120 eye tracker (Tobii Technology BA, Stockholm, Sweden) was used to capture children's eye movements at 60 Hz sampling rate. E-Prime software (PST, Pittsburg, PA) was used to control stimuli presentation and to record eye gaze data. Before starting the main experiment, children completed a 9-point eyetracking calibration that was followed by a familiarization to the structure of the task. The main experiment included 8 trials, and each trial consisted of an Exposure phase followed by a Choice phase (see Figure 1). During the Exposure phase, children saw either one exemplar of a same-relation (Single Conditions) or two different exemplars of a same-relation (Multiple Conditions) on the top half of the screen. A prerecorded voice oriented children to the exemplar(s). In the Single No Label condition this prompt was "See this one?" or "This is one". In the Multiple No Label condition we added to the prompts of the Single condition the following prompts: "See this one too? See how they are the same kind of thing?" or "This is one too. They are both the same kind of thing!".

In the Label conditions, on each trial the pre-recorded voice said the name of the Target during the Exposure (e.g. "See this dax?" / "See this dax too? See how they are both daxes?" and "This is a dax" / "This is a dax too. They are both daxes!'"). A different label was used on each trial (dax, ryke, fode, pabe, zup, kiv, mell or cheem), with target-word assignment randomized across participants.

After the original instance(s) were presented, two new exemplars appeared on the bottom half of the screen for the Choice phase: one same-relation (Target) and one different relation (Foil). Children were asked to point to the choice that was "the same kind of thing" (no label conditions) or "another \(\{\) dax \}" (label conditions). Across trials, the Target appeared equally often on the left and right side of the screen. There was no time limit for children's response. The prompts used throughout the experiment were recorded by a female native English speaker at a sample rate of 44.1 KHz .


Figure 1: Schematic representation of the structure of a trial in the multiple and single conditions without label. In the label conditions prompts had a unique label for each trial. The top row represents the Exposure Phase and the bottom row the Choice Phase. Example prompts are presented in the picture for illustration purposes only and were not presented to children.

\section*{Results}

Accuracy Did children successfully find the relational match? In the first four trials, the pattern of responses is consistent with prior findings (see Figure 2). Specifically, children performed above chance when given multiple instances (No Label: \(M=.65, t(14)=3.16, p=.007\); Label: \(M=.62, t(15)=1.94, p=.07)\) and also when a single instance was labeled \((M=.63, t(15)=2.18, p=.04)\). Children who got a single unlabeled instance did not reliably choose the relational match ( \(M=.51, p>.05\) ). Thus as in previous research, both multiple samples and labels support relational matching.

We then analyzed performance during the last four trials of the task to examine if children are able to establish relational matching over several trials. In the second half of the task, only one group of children performed above chance. Children who got single unlabeled instances throughout the task learned over the course of the task and successfully found relations \((M=.63, t(14)=3.16, p=\) .007). Thus, performance did not get better over trials but declined in the multiple samples conditions and improved in the condition usually associated with the poorest relational matching in young children. On the last of the experiment, the single unlabeled condition yielded the highest performance.


Figure 2: Proportion of relational choices for each condition across the task. The dotted line represents chance performance in the task.

Sampling of information How did children sample the visual information that was available to them when making a relational match-to-sample decision? One of the hypotheses for the benefit of labeling in the Single condition is that labels encourage children to compare the sample with each of the choices. An alternative hypothesis is that labels increase sustained attention to each instance. Critically, these two hypotheses make opposite predictions about the number of switches between the sample and choice options: the label either increases switching (more comparing) or the label decreases switching (more sustained attention).


Figure 3: Median number of switches back to the original instances during the choice phase for each condition.

Accordingly, we analyzed how often children switched back to the original sample(s) while they made a decision in each condition (Multiple or Single) in the presence and absence of labels. The results are clear: Children switched less when instance(s) were labeled (see Figure 3).

Adding a label reduced considerably the number of switches for both presentation conditions. An ANOVA looking at the mean number of switches with label condition (Label vs. No Label) and condition (Multiple vs. Single) as between-subject factors revealed a main effect of label, \(F\) \((1,54)=4.74, p=.03\), and no main effect of condition, \(F\) \((1,54)=1.86, p=.18\) or interaction between the two, \(F\) \((1,54)=1.31, p=.26\).

These results are consistent with the sustained attention hypothesis: Less switching would be associated with greater sustained attention on each object. Could sustained attention also be critical for learning over time, as children in the unlabeled Single condition did?

To answer this question we divided children in the Single No Label condition into two groups based on how much children switched from the options back to the sample during the first half of the task. Children who switched back more than the median for the sample were considered "High Switchers" while the remainder was considered "Low Switchers". If the benefit of time in the Single No Label condition were associated with sustained attention during the initial part of the task, Low Switchers would show better performance in the second half compared to High Switchers. Indeed, children who showed sustained attention during the initial learning trials were the children who learned over the task (see Figure 4).

\section*{Switching back to samples - Single Condition}


Figure 4: Proportion of relational choices for the Single No Label conditions across the task as a function of the amount of switching back to the original instances while making the choice. The dotted line represents chance performance in the task.

Although less switching in this condition during the initial part of the task does not result in better performance during
that part of the task it does seem to be associated with improved relational matching performance by the second half of the task. Indeed, only performance of Low Switchers on the second half of the task was significantly above chance levels, \(M=.68, t(7)=3, p=.02\). Moreover, this group's improvement from performance in the first half ( \(M\) \(=.57)\) to the second half is also significant, \(t(7)=2.78, p=\) . 03 .

\section*{Discussion}

Can sustained attention play a role in relational thinking? In answering this question, we need to take into consideration how children sample the information presented and what are the dynamics between sampling and sustained attention.

In the present work we aimed to investigate this question in the context of children's relational thinking - an important cognitive tool in human development. We asked children to match one of two options to a sample relation by visually inspecting the objects. The inclusion of labels, which have been shown to enhance relational matching, did so in the present study on early trials (but not later ones) and also led to more sustained looking and less switching.

Performance in the one-instance condition was particularly informative. This condition has been shown to be particularly challenging to children. Furthermore, to measure children's sampling and attentional behavior we used eyetracking technology. This method allowed us to gain initial knowledge on the dynamics of single presentation of evidence, labeling and relational matching. The evidence from the present work shows that relational match-to-sample performance is related to accumulation of evidence, effective sampling of information and sustained attention.

Behaviorally, the results presented here replicate previous evidence that single presentation of an instance of the relation does not provide enough support for children's relational thinking without the addition of a label. However, if we analyze the progress of children's performance across the task, we see that even in the absence of a label, children in the single condition can achieve above chance performance.
Accumulation of evidence Children who are presented with multiple initial instances of the target relation are more likely to choose the relational match from the beginning. Conversely, children who are only presented with one instance of the target relation in each trial require more trials to achieve this level of performance. Thus, learning the relational structure of objects requires children to relate several instances of the target relation. This can be more easily done when two instances are presented in each trial (Multiple condition).

Our results show that children in the Single No Label condition were able to achieve above change relational match behavior by the end of the task. However, only children who switched less from the choices back to the sample were able to achieve higher levels of performance by the second half of the task. This indicates a role for both
accumulation of evidence and sustained attention on performance in a relational match-to-sample task.
Focal attention and effective sampling Overall, adding a label to the presentation of the original instances decreases the mean number of times children check back to the original instances. This result is in line with previous evidence that labels increase sustained attention to objects (Balaban \& Waxman, 1997; Baldwin \& Markman, 1989; Fulkerson \& Waxman, 2007, McDuffie, Yoder \& Stone, 2006). This increased attention to each object is essential for a better identification of the target relation, particularly when the support from multiple samples is not readily available (the Single conditions).

In sum, the present work demonstrates the important role of sustained attention in learning how to make a relational choice when only one sample is presented. When a label is present in this condition it will increase sustained attention and result in better, faster, learning. In the absence of learning, sustained attention is important in the accumulation of evidence across several individual trials.

The importance of sustained attention is evidently dependent on some switching between the objects to establish relevant links among them. However, it does show an alternative to the hypothesis that comparison involves greater amounts of switching between samples and that labeling improves performance by promoting comparison.

At present, the specific mechanisms that underlie the usual positive effects of simultaneously presented multiple instances, labeling, and accrued effects of repeated trials have all been explained under the rubric of "comparison". The present approach - seeking micro-level behavioral evidence of direct comparison - suggests that all these phenomena are not the same. By pulling these factors that benefit children's relational matching apart, we may get a better handle on the processes that limit relational matching and on just what "comparison" is.

Accordingly, we believe the biggest contributions of the work presented here is the use of a new analysis paradigm to the investigation of comparison benefits to learning across development and the introduction of sustained attention as an important player in relational thinking.

\section*{Acknowledgements}

The work presented here was supported by National Institute of Health National Research Service Award HD007475-17. CV was supported by a Graduate Fellowship from the Portuguese Foundation for Science and Technology. PFC was supported by a Fulbright Research Fellowship and by a Graduate Fellowship from the Portuguese Foundation for Science and Technology. We would also like to thank Char Wozniak for help recording audio stimuli, the Cognitive Development Lab members for discussion and Charlene Tay, Ipek Gencer, Anna Mackinnon, Alyssa Ruffier, Madison Singell, Tracy Kelsey and Blakeley Meyer for help with stimuli creation, recruitment, data collection and coding.

\section*{References}

Balaban, M. T., \& Waxman, S. R. (1997). Do words facilitate object categorization in 9-month-old infants? Journal of Experimental Child Psychology, 64(1), 3-26.
Baldwin, D. a, \& Markman, E. M. (1989). Establishing word-object relations: a first step. Child development, 60(2), 381-98.
Christie, S., \& Gentner, D. (2010). Where Hypotheses Come From: Learning New Relations by Structural Alignment. Journal of Cognition and Development, 11(3), 356-373.
Christie, S., Gentner, D.. (2007). Relational similarity in identity relation: The role of language. In . S. Vosniadou \& D. Kayser (Eds), Proceedings of the Second European Cognitive Science Conference, Routledge.
Fulkerson, A. L., \& Waxman, S. R. (2007). Words (but not Tones) facilitate object categorization: Evidence from 6and 12-month-olds. Cognition, 105(1), 218-228
Gentner, D. (2003). Why we're so smart. In D. Gentner and S. Goldin-Meadow (Eds.), Language in mind: Advances in the study of language and thought. Cambridge, MA: MIT Press
Gentner, D. \& Namy, L.L. (1999). Comparison in the development of categories. Cognitive Development, 14, 487-513.
Gentner, D., Loewenstein, J., \& Hung, B. (2007). Comparison facilitates children's learning of names for parts. Journal of Cognition and Development,8(3), 285307.

Gentner, D., Loewenstein, J., \& Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. Journal of Educational Psychology, 95(2), 393.

Gentner, D., Rattermann, M.J., Markman, A.B., \& Kotovsky, L. (1995). Two forces in the development of relational similarity. In T.J. Simon \& G.S. Halford (Eds.), Developing cognitive competence: New approaches to process modeling. Hillsdale, NJ: LEA.
Gick, M. L., \& Holyoak, K. J. (1983). Schema induction and analogical transfer. Cognitive Psychology, 15(1), 138.

Kotovsky, L., \& Gentner, D. (1996). Comparison and categorization in the development of relational similarity. Child Development, 67, 2797-2822.
Loewenstein, J., \& Gentner, D. (2005). Relational language and the development of relational mapping. Cognitive Psychology, 50, 315-353.
Lupyan, G., Rakison, D.H., \& McClelland, J.L. (2007). Language is not just for talking: labels facilitate learning of novel categories. Psychological Science 18(12): 10771083.

McDuffie, A. S., Yoder, P. J., \& Stone, W. L. (2006). Labels increase attention to novel objects in children with autism and comprehension-matched children with typical development. Autism, 10(3), 288-301.
Namy, L. L., \& Gentner, D. (2002). Making a silk purse out of two sow's ears: Young children's use of comparison
in category learning. Journal of Experimental Psychology, 131(1), 5-15.
Oakes, L. M., \& Ribar, R. J. (2005). A comparison of infants' categorization in paired and successive presentation familiarization tasks. Infancy, 7(1), 85-98.
Pruden, S., Hirsh-Pasek, K., Shallcross, W. L., \& Golinkoff, R. M. (2008). Foundations of verb learning: Comparison helps infants abstract event components. In H. Chan, H. Jacob, \& E. Kapia (Eds.), Proceedings of the 32st Annual Boston University Conference on Language Development (Vol. 2, pp. 402-414).
Smith, L. B., \& Yu, C. (2013). Visual Attention Is Not Enough: Individual Differences in Statistical WordReferent Learning in Infants. Language Learning and Development, 9 (1), 25-49.
Wang, S. H., \& Baillargeon, R. (2008). Can infants be "taught" to attend to a new physical variable in an event category? The case of height in covering events. Cognitive Psychology, 56(4), 284-326.
Waxman, S. \& Leddon, E. (2011). Early word learning and conceptual development: Everything had a name, and each name gave birth to a new thought. In U. Goswami (Ed.) The Wiley-Blackwell Handbook of Childhood Cognitive Development. Malden, MA:Wiley-Blackwell.
Yu, C., \& Smith, L. B. (2012). Embodied attention and word learning by toddlers. Cognition, 125 (2), 244-262.
Vales, C. \& Smith, L.B. Words cue children's attention in a visual search task (2012). Poster presented at the 2012 VSS Annual Meeting, Naples, FL, USA.
Vurpillot, E. (1968). The development of scanning strategies and their relation to visual differentiation. Journal of Experimental Child Psychology, 6(4), 632650.

\title{
Investing Amid Uncertainty: A Test of the Domain Specific Anchoring Hypothesis
}

\author{
Hui Yih Chai (hui.chai@psy.unsw.edu.au) \\ and \\ Ben R. Newell (ben.newell@unsw.edu.au) \\ School of Psychology, University of New South Wales, Sydney, Australia
}

\begin{abstract}
The current study examined reactions to the precision of earnings' forecasts in hypothetical investment decisions. In a forced choice task, participants were found to be indifferent between point (e.g., \$2) or range (e.g., \$1.70-\$2.30) forecast formats when both outcomes were favorable (i.e., above market expectation). When the outcomes were unfavorable (below expectation), participants' preferences were significantly biased towards range estimates. When faced with options which mixed forecast formats and favorability, participants almost always opted for forecasts with a favorable outlook regardless of format. These results are inconsistent with domain specific ambiguity reactions found previously (Du, 2009) and also offer no support for the domain specific anchoring hypothesis (e.g. Du, 2009; Du \& Budescu, 2005). These findings raise some doubts about the generality of domain specific reactions to uncertainty and suggest that such effects might be dependent, in part, on the (financial) sophistication of participants.
\end{abstract}

Keywords: Point; Range; Uncertainty; Ambiguity; Communication; Precision format; Domain specific

Investors often rely on management earnings forecasts when making their investment decisions. However, these forecasts are rarely precise. For example, there is uncertainty surrounding the reliability and credibility of the forecasts, the forecasters might be unsure about the current economic outlook, the market might be too volatile for an accurate forecast to be made and so on. This growing uncertainty that the investors, inevitably, have to face has generated a heated discussion on how best to communicate uncertainty to financial decision makers (Christensen, Glover, Omer, \& Shelley, 2012).

Extant literature has found that managers often communicate uncertainty by incorporating imprecision into their management earnings forecasts (Christensen et al., 2012; Du, 2009; Du \& Budescu, 2005; Du, Budescu, Shelley, \& Omer, 2011). Instead of providing a precise point forecast (e.g. \$1.00), managers issue earnings forecasts in the form of a range estimate (e.g. \$0.60-\$1.40).

These findings are further evidenced in the National Investor Relations Institute survey results (NIRI, 2003). They found that the majority (78\%) of corporate members provided earnings forecasts information regularly and among them, \(75 \%\) use range estimates, whereas only \(11 \%\) choose the precise point formats. In addition, the authors found that, by the end of 2001 , \(55 \%\) of the firms reported their earnings in a range format while only \(23 \%\) issued the earnings forecasts in the form of point estimate (Cotter, Tuna, \& Wysocki, 2006).

\section*{Reactions to Range Forecasts}

Despite the fact that range forecasts are widely used in earnings forecasts, investors' reactions to this format are rather mixed. On one hand, some literature has found that being open about uncertainty by choosing a range disclosure format could not only increase perceived credibility, trustworthiness (Hirst, Koonce, \& Venkataraman, 2008), and investors' confidence in the company (Habicht, 1992) but also reduce the company's legal liability (Hirst et al., 2008).

On the other hand, evidence of range formats not being well-received has also been obtained. Prior research has shown that using the range format has led to reduced investors’ confidence (Hirst, Koonce, \& Miller, 1999), heightened risk perception (Han et al., 2009; Kuhn \& Budescu, 1996), negative affective reactions such as worry or distress related to choice outcomes and avoidance of decision making in choice situations (Camerer \& Weber, 1992; Han, Moser, \& Klein, 2006).

Mixed reactions towards range forecasts were also highlighted in participants' comments that, although they felt that range forecasts were more trustworthy and informative than precise point format, the range format has led them to question the company's competence and ability in estimating the uncertainty and risk involved (Dieckmann, Mauro, \& Slovic, 2010; Johnson \& Slovic, 1995).

In terms of people's understanding of the information conveyed in the range format, past studies reported that participants felt that range forecasts were more complicated and harder to understand (Han et al., 2009; Johnson \& Slovic, 1995), even for highly educated and professional participants (Sheridan \& Pignone, 2002).

\section*{Domain Specific Ambiguity Reactions}

Driven by the mixed findings between the reactions to point and range earnings forecasts, Du (2009) conducted a study to clarify how investors react to forecasts in different presentation formats and with different outcome favorability. In Du's study, MBA students were asked to evaluate four different investment options based on the brief company background information provided and also their respective CEO's forecasts of next year's earnings presented either in a point or a range form. Specifically, the participants were given prediction of the Earnings Per Share (EPS) for next year.

Furthermore, participants received information about the performance of the company relative to a benchmark.

Participants were told whether the forecasts were higher (indicating positive/favorable performance) or lower (negative/unfavorable performance) than market expectation. Having read the descriptions and forecasts information, the participants made a series of judgments, including earnings prediction, confidence, investment, risk and likelihood of investment decisions.

Drawing evidence from the literature on decision making under ambiguity (e.g. Camerer \& Weber, 1992; Du \& Budescu, 2005; Ellsberg, 1961), Du predicted that investors would display ambiguity attitudes which were domain specific. In other words, in the domain of favorable outcomes, participants would seek ambiguity and react positively to range forecasts whereas in the domain of unfavorable outcomes, participants would avoid ambiguity and react negatively to range forecasts.

Du (2009) found that precision format had no effect on investors' earnings predictions, confidence and risk judgments. However, partial support for the domain specific hypothesis was found in that when the outcomes were favorable (i.e. the forecasts were higher than market expectation), participants preferred ambiguity and invested more when the forecasts were presented in a range rather than in a point format. However, no such finding was found when the outcomes presented were unfavorable (investment was equally low in both range and point formats).

Du's (2009) findings are consistent with Viscusi and Chesson's (1999) findings that participants' reactions to ambiguity was driven by the "hope and fear" effects. Viscusi and Chesson found that when the situation generates a 'fear' effect (or a small possibility of loss), people are more averse to ambiguity whereas when the situation generates a 'hope' effect, people are more inclined to seek ambiguity. Du's favorable forecasts could have generated a 'hope' effect, and hence, driven participants to seek ambiguity and invest more in companies with range forecasts.

\section*{Domain Specific Anchoring Hypothesis}

Du (2009) suggested that her findings are consistent with the argument that participants selectively draw information from different focal end points of range estimates when the decision context changes ( Du \& Budescu, 2005). In particular, the domain specific argument posits that participants are more likely to focus on the upper bound of the range and seek ambiguity (or imprecision) when there is a potential for gains, but they are more likely to focus on the lower bound of the range and avoid ambiguity when there is a potential for losses (hereafter referred to as the domain specific anchoring hypothesis). Du's finding that people invested more in companies with range forecasts when the outcomes were favorable could be the result of a preference for the higher earnings values when participants compared the upper end of the range forecasts with the point forecasts.

Furthermore, this domain specific anchoring hypothesis has been linked to goal framing theory. Budescu, Kuhn and colleagues (2002) argue that investors are motivated by the
goals to maximize gains and to minimize losses in the financial market, and thus, anchor their judgments on different focal end points in different contexts. In the domain of gains, participants anchor their judgment on the upper end of the range estimates so as to maximize the potential gains and vice versa in the domain of losses in order to minimize their potential losses.

\section*{Aim and Hypotheses}

Whilst evidence shows that the domain specific argument is well-established, thus far, no research has tested this hypothesis directly in the context of investor decisions. Therefore, the current study examined the prevalence of domain specific ambiguity reactions and tested the extent to which these findings could be explained by the domain specific anchoring hypothesis. The current study adapted a method used by Du et al. (2011) and presented participants with eight forced choice options differing in the forecast format, predicted outcome favorability and forecasts values. After they read the brief descriptions of the two forecasts, participants evaluated the informativeness, accuracy and credibility of each form of presentation.

Prior to indicating their preference of investment in the two options, they were required to predict where the actual earnings would fall. That is, for the point forecasts (e.g., \(\$ 2.00\) ), they estimated the probability of the actual earnings being lower than, exactly at, and higher than the point estimate. For the range forecasts (e.g., \$1.70-\$2.30), participants indicated the probability of the actual earnings to be lower than the lower end, exactly at the lower end, between the range, exactly at the higher end, and higher than the higher end of the range forecasts (hereafter referred to as the anchoring task). Their responses served as an indication of where they anchored their judgments when making the investment decision under different domains.

In accordance with Du (2009), it was predicted that participants would show domain specific ambiguity reactions, that is, participants would show ambiguityseeking behavior when the forecasts were favorable (i.e., prefer range over point estimates) and ambiguity-averse behavior when the forecasts were unfavorable (i.e., prefer point over range).

Furthermore, it was expected that these patterns of responses could be explained by the domain specific anchoring hypothesis. We predicted that the participants' estimates of the actual earnings to be more likely to occur at the higher end of the range estimates when forecasts were favorable and given that this upper end estimate was higher as compared to the point estimate, they would be more inclined to choose range forecasts when they saw favorable outcomes. In a similar vein, we predicted participants would anchor more on the lower end of the range estimates when the forecasts were unfavorable and since this lower bound estimate was lower than the comparative point estimate, participants would prefer point estimates more when the outcomes were unfavorable.

In terms of informativeness, accuracy and credibility, it was predicted that point estimates would be rated as more accurate than range ones ( \(\mathrm{Du}, 2009\) ), and that as the imprecision increased, the perceived informativeness (Kim \& Verrecchia, 1991) as well as its credibility (Longman, Turner, King, \& McCaffery, 2012) would decrease.

\section*{Method}

\section*{Participants}

Thirty Psychology undergraduate students from the University of New South Wales participated (70\% male, \(M_{\text {age }}=19.5\) year-old, \(S D_{\text {age }}=2.2\) ) in return for course credit.

\section*{Experimental Design and Measures}

This study employed a 2 (forecasts format: point vs. range) X 2 (forecasts favorability: favorable vs. unfavorable) X 2 (forecasts values: high vs. low) withinsubject design. Participants were asked to assume the role of an investor and to assess a number of investment options that varied in three dimensions: (1) earnings forecasts format, (2) forecasts favorability, and (3) forecasts values.

Forecasts Format The earnings forecasts (i.e. EPS) were presented either in a point (e.g. \$2.00) or a range (e.g. \(\$ 1.70-\$ 2.30\) ) format. These two formats were considered to be informationally equivalent because the midpoint of the range estimate always matched the point estimate. The width of the range estimates was fixed at \(\$ 0.60\).

Forecasts Favorability Similar to Du (2009), benchmark information for each forecast was provided to indicate the overall performance of the company. Participants were asked to assume that an earnings forecast that was higher than market expectation indicated good performance whereas one that was lower than market expectation indicated poor performance.

Forecasts values Prior research has shown that investors' decisions were affected by the expected earnings values ( Du \& Budescu, 2005). Thus, the absolute values of the current earnings forecasts were also manipulated. Half of the estimates were high in values (with a midpoint of \(\$ 5.00\) or \(\$ 6.00\) ) while the remaining half was low in their absolute amount (with a midpoint of \(\$ 1.00\) or \(\$ 2.00\) ).

\section*{Procedures}

Subsequent to providing informed consent, participants were given earnings forecasts of two companies and asked to indicate which company they would invest in. They could also express indifference between the two options. However, prior to making their investment decision, they were required to evaluate the informativeness, accuracy and credibility of each of the formats on a 6-point scale (1: Not at all, 6: Very). Then, they were instructed to complete the anchoring task for each form of presentation. After
completing the rating and anchoring judgments, the respondents made their investment decision.

The two forecasts formats and two favorability outcomes yielded four possible combinations of the two dimensions:
1. Range favorable vs. Point favorable
2. Range unfavorable vs. Point unfavorable
3. Point favorable vs. Range unfavorable
4. Point unfavorable vs. Range favorable

For example, participants comparing statements in the first combination would evaluate "Company PA with a predicted EPS of \(\$ 2.00\), which is higher than market expectation" (point favorable statement) and "Company QC with a predicted EPS in the range of "\$1.70-\$2.30, which is higher than market expectation (range favorable statement). Altogether, participants completed eight forced choice options, two versions (a high and a low absolute value version) for each combination.

\section*{Results}

Table 1 displays the distribution of investor's preferences of forecasts format and outcome favorability across high and low earnings values. Contrary to prediction, the pattern of responses did not show that participants favored range (point) forecasts when the outcomes were favorable (unfavorable). Instead, Table 1 shows that when the outcomes were favorable (comparison 1), participants were about equally likely to choose the company with point or range forecasts. When the forecasts were unfavorable (comparison 2), participants were biased towards choosing range estimates. However, in situations where there was a mix between forecast format and outcome favorability (comparisons 3 and 4), participants almost always opted for the ones with favorable outlook, regardless of format.

A statistical analysis was conducted to examine if these differences were statistically significant. Participants’ choices were coded into an index of preference for precision. As in Du et al. (2011), a preference for point forecast was given a value of 0 whereas a preference for range forecast was given a value of 1 . Indifferent (or "Either") option was assigned a value of \(0.5^{1}\). A mean preference score was calculated with a mean of less than 0.5 indicating a preference for point estimates and a mean score of more than 0.5 indicating a preference for range forecasts. A one-sample t-test, with a test-value of 0.5 , was conducted and support for the patterns of responses aforementioned was found (see Table 2).

In short, no evidence of domain specific ambiguity reactions was found. As can be seen in Table 2, participants seemed to prefer range estimates more especially when the forecasts were unfavorable, and when they were faced with a choice with mixed forecast format and outcome favorability, their decisions were almost always swayed by the favorability of the outcomes.

\footnotetext{
\({ }^{1}\) A further analysis showed that excluded indifferent responses did not alter the statistical pattern of effects.
}

\section*{Domain Specific Anchoring Hypothesis}

The distributions of participants' predictions about the range and point forecasts are summarized in Figure 1. Figure 1 shows, overall, participants estimated that the actual earnings were more likely to be in between the range estimates and exactly at the point estimate. The participants also appeared to believe that the actual earnings were roughly equally likely to occur at the different focal end points, regardless of decision context.

Recall that the domain specific anchoring hypothesis predicted that participants’ allocation of the probability judgments at the different end points would differ as a function of favorability of the outcomes. Specifically, the hypothesis predicts that participants will anchor their judgments at the lower end of the range forecasts when the outcomes were unfavorable, but at the upper end when the outcomes were favorable. The data plotted in Figure 1a seem to show no support for this anchoring hypothesis.

In order to test this prediction statistically, an average of participants' responses at the lower ends (i.e. both 'exactly at lower bound' and 'lower than lower bound') and the upper ends (i.e. both 'exactly at upper bound’ and 'higher than upper bound') of the range estimates were calculated. T-tests were then carried out to examine if participants’ prediction of the occurrence of the actual earnings at the lower and upper ends would differ as a function of outcome favorability. No significant differences were found. The results indicated that participants believed that the actual earnings were equally likely to occur at the lower and upper ends regardless of outcome favorability, \(t(29)_{\text {lower }}=-.408\), \(p_{\text {lower }}=.686\) and \(t(29)_{\text {higher }}=1.378, p_{\text {higher }}=.179\) respectively.

Similar analyses were also conducted on point forecasts. In accordance with the domain specific anchoring hypothesis, it was predicted that participants would be more optimistic and believe that the actual earnings were more likely to be higher than the forecast when the forecast was favorable, but they would be more pessimistic and consider the actual earnings to be lower than the estimate when the forecast was unfavorable. T-tests revealed that none of the effects were significant, both \(p \mathrm{~s}>.05\).

Collectively, findings from both the point and range analyses showed no support for the domain specific anchoring hypothesis. Participants did not focus on the different end points when the favorability of the outcomes differed.

\section*{Informativeness, Accuracy and Credibility Ratings}

A 2 X \(2 \times 2\) repeated measure ANOVA was carried out on participants' informativeness, accuracy and credibility ratings. Consistent with prior research (Du, 2009), point estimates ( \(M=4.07, S D=1.18\) ) were rated as more accurate than the range forecasts \((M=3.32, S D=0.87), F(1,29)=\)
11.242, \(p=.002\). In terms of credibility ratings, favorable forecasts (or forecasts that were higher than market expectation; \(M=3.26, S D=1.14\) ) were rated as slightly less credible than unfavorable ones (or those that were lower than market expectation; \(M=3.46, S D=1.13), F(1,29)=\) 6.849, \(p=.014\). No other significant effects of forecast formats and favorability on how participants rated informativeness, accuracy and credibility judgments were found.

Table 1. Distribution of preferences of forecast format and outcome favorability across different earnings conditions.
\begin{tabular}{llccc}
\hline & DMs prefer & \begin{tabular}{c} 
Low \\
EPS
\end{tabular} & \begin{tabular}{c} 
High \\
EPS
\end{tabular} & Average \\
\hline 1. Point Fav vs. & Point Fav & 11 & 9 & 10 \\
Range Fav & Range Fav & 15 & 18 & 16.5 \\
& Either & 4 & 3 & 3.5 \\
2. Point Unfav vs. Point Unfav & 5 & 8 & 6.5 \\
Range Unfav & Range Unfav & 17 & 14 & 15.5 \\
& Either & 8 & 8 & 8 \\
3. Point Fav vs. & Point Fav & 17 & 22 & 19.5 \\
Range Unfav & Range Unfav & 8 & 6 & 7 \\
& Either & 5 & 2 & 3.5 \\
4. Point Unfav vs. Point Unfav & 6 & 3 & 4.5 \\
Range Fav & Range Fav & 21 & 23 & 22 \\
& Either & 3 & 4 & 3.5 \\
\hline
\end{tabular}

Note: DM = Decision Makers; EPS = Earnings per share; Low EPS = EPS with low values (with a midpoint of \(\$ 1.00\) or \$2.00); High EPS = EPS with high values (with a midpoint of \(\$ 5.00\) or \(\$ 6.00\) ); Fav = Favorable; Unfav = unfavorable, Either = indifferent between the two options.

Table 2. Comparison of point and range forecasts differing in favorability.
\begin{tabular}{|c|c|c|c|}
\hline & Mean Preference & \(t(29)\) & Sig. (2tailed) \\
\hline 1. Point Fav vs. Range Fav & 0.61 & 1.383 & 0.177 \\
\hline 2. Point Unfav vs. Range Unfav & 0.65 & 2.473 & .019* \\
\hline 3. Point Fav vs. Range Unfav & 0.29 & -3.117 & .004** \\
\hline 4. Point Unfav vs. Range Fav & 0.79 & 4.592 & .000** \\
\hline
\end{tabular}

Note: Ms > 0.5 indicates a preference for range forecasts; \(\mathrm{Ms}<\mathbf{0 . 5}\) indicates a preference for point forecasts
*indicates significance at the .05 level
** indicates significance at the .01 level


Figure 1. Distribution of probabilities of occurrence of actual earnings at different focal points for (a) range and (b) point estimates.

\section*{Discussion}

The current study aimed to test the domain specific anchoring hypothesis directly in the context of financial decision making. An examination of investors' preferences for point or range presentation format in favorable and unfavorable contexts revealed no evidence for domain specific ambiguity reactions. Participants did not show ambiguity seeking or avoidance behavior when the favorability of the outcomes changed. A further investigation of participants' anchoring judgments on where the actual earnings would fall also failed to support the domain specific anchoring hypothesis. Somewhat surprisingly, we found that even though participants believed that point forecasts were more accurate than range forecasts, they still seemed to prefer forecasts in range format more than those in point format, particularly when the forecasts were unfavorable. This finding was not affected by the perceived informativeness and credibility judgments on the presentation format.

Interestingly, favorable forecasts were rated as less credible than unfavorable estimates. Given the growing uncertainty in the current economy, participants may be more cautious and skeptical about the forecasts provided. They may feel that favorable forecasts have not sufficiently incorporated the uncertainty in the current economic conditions, and hence, rated them as less credible than unfavorable ones.

Another interesting finding of the present study is that financial decisions are largely dominated by the favorability of the outcomes. In the third and fourth comparisons where forecast format and outcome favorability were mixed, participants almost always selected the ones with favorable outcomes. Given that participants were told to regard 'higher (lower) than market expectation' as an indication of good (poor) performance, it is perhaps unsurprising that the effects of favorability may overshadow the concern for
forecast format when both dimensions are mixed. It would be interesting in follow up work to see if participants' preferences' differ when favorability of the outcomes is manipulated between subjects.

One possibility for why we did not find domain specific ambiguity reactions could be that the current participant pool was psychology undergraduates rather than the MBA students used in Du (2009). Moreover, we did not take into account participants' prior investment experience or knowledge about the stock market. Limited exposure to the financial settings may have affected participants’ understanding of the earnings prediction presented to them making them less susceptible to the influence of the different forecast formats. Future work could examine this role of 'expertise' in reactions to imprecision.

Another difference between the current research and Du's (2009) is that Du employed range forecasts with variable width whereas we used range forecasts with fixed width. The largest range width used in Du's study (i.e. \$1.80) was three times the value of the current width (i.e. \(\$ 0.60\) ). It could be that the domain differences in ambiguity reactions are only found in forecasts with larger range width.

This assertion is further supported by examination of the range width used in Du and Budescu's (2005) study which also found support for domain specific ambiguity reactions. The size of the range width chosen by the authors ranged from \(\$ 2.00\) to \(\$ 32.00\). Although participants in the current study rated range forecasts as different (in terms of accuracy) from point forecasts, the \(\$ 0.60\) range width may still be too narrow, and hence, too subtle to influence their preferences between the two formats.

On the other hand, previous evidence of domain specific ambiguity reactions (e.g. Du \& Budescu, 2005) is not as direct as Du’s (2009). For example, Du and Budescu (2005) found that domain specific ambiguity reactions was task specific. They only found the reversal of ambiguity attitudes in a certainty equivalent task but not in a pairwise
comparison task. Thus it seems that reactions to ambiguity might be dependent on the particular task used to elicit preferences.

Furthermore, in their study, the authors manipulated the sources of uncertainty - whether it is uncertainty in the estimates or in the outcomes (Du \& Budescu, 2005). It could be that the effect of domain specific ambiguity reactions would be stronger when another source of uncertainty is included. Future research specifically focusing on these issues will help to clarify the exact nature of domain specific ambiguity reactions.

In conclusion, we found no evidence for domain specific ambiguity reactions or for the anchoring hypothesis in the context of investment decisions. However, our results revealed that favorability of the outcomes dominates judgments and its effect may have masked the concern for presentation format when both dimensions were mixed. Future research, controlling for the issues discussed, is required to unpack the domain specific observations more thoroughly.

\section*{Acknowledgements}

This research was supported by an Australian Postgraduate Award Scholarship awarded to the first author, and an Australian Research Council Linkage Project Grant and a Future Fellowship awarded to the second author. The support of the ARC Centre of Excellence in Climate Systems Science is also gratefully acknowledged.

\section*{References}

Budescu, D. V., Kuhn, K. M., Kramer, K. M., \& Johnson, T. R. (2002). Modeling certainty equivalents for imprecise gambles. Organizational Behavior and Human Decision Processes, 88(2), 748-768.
Camerer, C., \& Weber, M. (1992). Recent developments in modeling preferences: Uncertainty and ambiguity. Journal of Risk and Uncertainty, 5(4), 325-370.
Christensen, B., Glover, S., Omer, T., \& Shelley, M. (2012). Does Estimation Uncertainty Affect Investors' Preference for the Form of Financial Statement Presentation? Available at SSRN: Abstract 2163878.
Cotter, J., Tuna, A. I., \& Wysocki, P. D. (2006). Expectations management and beatable targets: How do analysts react to explicit earnings guidance? Contemporary Accounting Research, 23(3).
Dieckmann, N. F., Mauro, R., \& Slovic, P. (2010). The Effects of Presenting Imprecise Probabilities in Intelligence Forecasts. Risk Analysis, 30(6), 987-1001.
Du, N. (2009). Do Investors React Differently to Range and Point Management Earnings Forecasts? Journal of Behavioral Finance, 10(4), 195-203.
Du, N., \& Budescu, D. V. (2005). The Effects of Imprecise Probabilities and Outcomes in Evaluating Investment Options. Management Science, 51(12), 1791-1803.
Du, N., Budescu, D. V., Shelley, M. K., \& Omer, T. C. (2011). The appeal of vague financial forecasts.

Organizational Behavior and Human Decision Processes, 114(2), 179-189.
Ellsberg, D. (1961). Risk, Ambiguity, and the Savage Axioms. The Quarterly Journal of Economics, 75(4), 643669.

Habicht, F. H. (1992). Guidance on risk characterization for risk managers and risk assessors. Retrieved from United States Environmental Protection Agency website: http://www.epa.gov/oswer/riskassessment/habicht.htm
Han, P. K., Klein, W. M. P., Lehman, T. C., Massett, H., Lee, S. C., \& Freedman, A. N. (2009). Laypersons' Responses to the Communication of Uncertainty Regarding Cancer Risk Estimates. Medical Decision Making, 29(3), 391-403.
Han, P. K., Moser, R. P., \& Klein, W. M. (2006). Perceived ambiguity about cancer prevention recommendations: relationship to perceptions of cancer preventability, risk, and worry. J Health Commun, 1, 51-69.
Hirst, D. E., Koonce, L., \& Miller, J. (1999). The Joint Effect of Management's Prior Forecast Accuracy and the Form of Its Financial Forecasts on Investor Judgment. Journal of Accounting Research, 37, 101-124.
Hirst, D. E., Koonce, L., \& Venkataraman, S. (2008). Management Earnings Forecasts: A Review and Framework. Accounting Horizons, 22(3), 315-338.
Johnson, B. B., \& Slovic, P. (1995). Presenting Uncertainty in Health Risk Assessment: Initial Studies of Its Effects on Risk Perception and Trust. Risk Analysis, 15(4), 485494.

Kim, O., \& Verrecchia, R. E. (1991). Trading volume and price reactions to public announcements. Journal of Accounting Research, 29(2), 302-321.
Kuhn, K. M., \& Budescu, D. V. (1996). The Relative Importance of Probabilities, Outcomes, and Vagueness in Hazard Risk Decisions. Organizational Behavior and Human Decision Processes, 68(3), 301-317.
Longman, T., Turner, R. M., King, M., \& McCaffery, K. J. (2012). The effects of communicating uncertainty in quantitative health risk estimates. Patient Educ Couns, 89(2), 252-259.
NIRI. (2003). NIRI survey results on earnings guidance practices: The executive alert: The National Investors' Relations Institute.
Sheridan, S. L., \& Pignone, M. (2002). Numeracy and the medical student's ability to interpret data. Eff Clin Pract, 5(1), 35-40.
Viscusi, W. K., \& Chesson, H. (1999). Hopes and Fears: the Conflicting Effects of Risk Ambiguity. Theory and Decision, 47(2), 157-184. doi: 10.1023/a:1005173013606

\title{
Producing gestures facilitates encoding of spatial relation
}

\author{
Amy Chong (S1155009099@Mailserv.Cuhk.Edu.Hk) \\ Ben Choi (S1155016253@Mailserv.Cuhk.Edu.Hk) \\ Elena Kwong (S0960192@Mailserv.Cuhk.Edu.Hk) \\ Department of Psychology, The Chinese University of Hong Kong, Hong Kong \\ Jennifer Chan (S1155004719@Mailserv.Cuhk.Edu.Hk) \\ Irina Chong (S1155004024@Mailserv.Cuhk.Edu.Hk) \\ Mavis Ip (S1155003387@Mailserv.Cuhk.Edu.Hk) \\ Department of Educational Psychology, The Chinese University of Hong Kong, Hong Kong \\ Christopher Yeung (S1155003299@Mailserv.Cuhk.Edu.Hk) \\ Department of Psychology, The Chinese University of Hong Kong, Hong Kong \\ Wing Chee So (WINGCHEE@Cuhk.Edu.Hk) \\ Department of Educational Psychology, The Chinese University of Hong Kong, Hong Kong
}

\begin{abstract}
This paper examines whether producing gestures would facilitate encoding of spatial relation in a navigation task. In this experiment, we focused on gestures produced without accompanying speech. Adult participants were asked to study spatial sequence of routes shown in four diagrams, one at a time. Participants rehearsed the routes with gestures, actual hand movements (actually drew the routes on papers), or mental simulation. They then were asked to reconstruct the routes with sticks. Participants who moved their hands (either in the form of gestures or actual drawing) recalled better than those who mentally simulated the routes and those who did not rehearse, suggesting that hand movements produced during rehearsal facilitate encoding of spatial relation. Interestingly, participants who gestured the routes in the air recalled better than those who drew them on papers, suggesting that gesture, as a kind of representational action, exerts more powerful influence on spatial relation encoding.
\end{abstract}

Keywords: Gesture; Spatial Cognition; Action; Encoding; Embodied Cognition.

\section*{Introduction}

Spatial knowledge consists of three major skills, including spatial visualization, spatial relation, and spatial orientation (Lohman, 1979). The present study focuses on spatial relation. Understanding relational information enables us to form a spatial representation regarding relation between locations, objects, and paths. Such understanding is particularly useful when we are processing spatial information of how starting points and destinations are considered in relation to one another. Therefore, developing techniques to facilitate encoding of spatial relation has received increasing attention from cognitive and educational psychologists all over the world.

In the present study, we examine whether embodied movements like gestures might be effective in encoding spatial relation. Previous research has shown that producing gestures is directly involved in encoding new information but those studies focused on mathematics domain. Children who were told to gesture when explaining their solutions to a math problem benefited more from the subsequent math lesson, compared to children who were told not to gesture (Broaders, Cook, Mitchell, \& Goldin-Meadow, 2007). Children who were instructed to reproduce teacher's gestures while acquiring new mathematics concepts learnt and memorized mathematics knowledge better than did those who were instructed to reproduce teacher's verbal instructions only (Cook, Mitchell, \& Goldin-Meadow, 2008). However, no experimental work has examined whether gestures strengthen spatial relation encoding. Gestures are spontaneous hand movements. They are produced in space, and thus are inherently spatial (McNeill, 1992; 2005). Therefore, learners can exploit the spatial properties of gestures to encode spatial relation between the starting point and destination. For example, when encoding spatial sequence of a route, learners may trace the steps with an index finger in the air by moving it to the right, upwards, and to the right again.

In fact, gestures and spatial relation are tightly linked. Previous studies have shown that speakers produce cospeech gestures (gestures that are co-occurring with speech) when they convey spatial relation to listeners in speech. For example, they use co-speech gestures to depict spatial layout of an area (Emmorey, Tversky \& Taylor, 2000) and spatial arrangement of objects (Sauter, Uttal, Alman, GoldinMeadow, \& Levine, 2012). In addition, previous studies have reported that speakers produce co-speech gestures frequently when they are identifying spatial relation between two characters in narratives (So, Coppola, Liccidarello, \& Goldin-Meadow, 2005; So, Kita, \& GoldinMeadow, 2009). They also increase gesture production
when encountering difficulty in describing complex spatial patterns (Hostetter, Alibali, \& Kita, 2007; Melinger \& Kita, 2007).

The present study asks whether asking participants to produce gestures while encoding spatial relation information would enhance subsequent spatial recall. We here focus on gestures produced while thinking silently (i.e., co-thought gestures, see Chu \& Kita, 2011). If co-thought gestures merely depict spatial relation, then participants who are told to gesture when rehearsing spatial sequence silently should recall comparable number of steps than those who are told not to gesture. However, if co-thought gestures do more than simply conveying spatial relation, i.e., they are involved in encoding and constructing spatial relation, then participants who are told to gesture during rehearsal should recall more steps than those who are told not to gesture.

In order to test the above hypotheses, adult participants were told to study various routes and to rehearse the routes by producing co-thought gestures (e.g., index finger moves up and then to the right). We then compared their recall performance to another three groups of learners who were instructed, respectively, to rehearse the routes by actually drawing them out on papers, to mentally rehearse the routes while having their hand movements prohibited, and to read letters that prevented rehearsal.

\section*{Method}

\section*{Participants}

One hundred and twelve Chinese-speaking undergraduates (53 men, age range: 19-21 years) were recruited and each of them was paid for HKD \(\$ 30\) for their participation. All of them had correct or correct-to-normal vision. All but one participant were right-handed. They were undergraduates at the Chinese University of Hong Kong. The participants were randomly assigned to one of the four experimental conditions: 1) co-thought gesture; 2) actual drawing; 3) hand movement prohibited; and 4) no rehearsal, with 28 participants in each condition.

\section*{Stimuli}

We designed the stimuli that purposefully examined spatial relation. Four diagrams were created by the software "Edraw Max". In each diagram, there were seven vertical lines and ten strokes that were horizontal, diagonal, or curly connecting or not connecting with the vertical lines. The strokes that were connected to vertical lines formed a route navigating from the starting point to the destination. See Figure 1 for one of the diagrams (top) and its route highlighted in red (bottom). For the sequence of this route, one should move down, then move diagonally downwards, move up, move to the right, move down, move to the right, move down, cross the curly road, move up, move diagonally upwards, move down, cross the bridge, and finally move to the destination. There were thirteen steps in each route.


Figure 1. The top figure shows one of the maps tested in this experiment. The bottom figure shows the route navigating from the starting point to the destination.

\section*{Procedure}

Participants were tested individually. They were asked to study four routes, one at a time, and later on describe the routes to an experimenter. Each time they were presented with a diagram that showed a complete route on an A4-sized paper (see the bottom figure in Figure 1). In order to help them to get familiar with the steps, participants were told to trace the complete route twice with a highlighter. They should trace every step and not to pause at any junctions of the route. Then we removed the diagram.

Participants then received different instructions for rehearsal in different conditions. In the co-thought gesture condition, participants were told to rehearse the route from the starting point to the destination with their hands. In the actual drawing condition, participants were instructed to draw the route from the starting point to the destination once on a piece of blank A4-sized paper. Participants were told that they were not required to draw the route in the same scale as that shown in the previous diagram. They were also told that neatness of their drawing would not be evaluated. In the hand movement prohibited condition, participants were told to visualize or mentally simulate the route sequence from the starting point to the destination once while holding a softball in both hands. Then they informed the experimenter when they finished visualizing a complete route. In the no rehearsal condition, participants were given an A4-sized paper with different alphabets randomly printed on it. They were told to read the alphabets aloud for 20 seconds in order to prevent them from mentally rehearsing
the route. Before this experiment, we conducted a pilot study and found that on average participants spent 20 seconds on rehearsing a complete route in the hand movement prohibited condition. Therefore, we asked participants to read letters aloud for 20 seconds. We also expected that reading letters aloud would not interfere with participants' spatial representations because the letters were randomly printed on an A4-sized paper such that they did not form any clear spatial pattern.

Then all participants recalled the route they had just rehearsed. They were given thirteen sticks with the same length and told to reconstruct the route sequence from the starting point to the destination on a table. They were told that they were not required to reconstruct the route in the same scale as that shown in the diagrams.

Before they studied the second route, participants were required to work on a set of mathematics problems for two minutes in order to prevent proactive interference of the directions from the previous route. Then the second diagram was presented and the aforementioned procedures were repeated. The experiment was complete after all four routes were studied. The order of diagrams was randomized across participants. The whole experiment was videotaped.

\section*{Coding}

We measured the average amount of time (in seconds) each participant spent on rehearsing a complete route (including pauses and self-corrections, if any) across four diagrams in different conditions (except the no rehearsal condition). We also examined the mean number of steps participants rehearsed in the co-thought gesture and actual drawing conditions.

We then assessed the accuracy of recall by considering how many steps (out of thirteen) were correctly reconstructed by sticks for each diagram. A step was considered recalled correctly if the direction and sequence of the corresponding stick matched those in the diagram. The mean proportion of steps correctly recalled in each diagram was calculated for each participant, which was the number of steps correctly recalled, divided by thirteen (i.e., the total number of steps in the diagram). We also measured the average amount of time (in seconds) each participant spent on reconstructing each route (including all pauses and hesitations)

Reliability was assessed by having a second coder coded a subset (20\%) of the data. Inter-rater agreement was \(98 \%\) (Cohen's Kappa \(=.95\) ) for measuring the time spent on rehearsal; \(91 \%\) (Cohen’s Kappa \(=.88\) ) for identifying the number of steps rehearsed in the co-thought gesture in the air, and actual drawing conditions; 95\% for determining the accuracy of steps reconstructed (Cohen's Kappa = .92); and \(100 \%\) for determining the duration of reconstruction (Cohen's Kappa = 1).

\section*{Results}

All participants in the co-thought gesture condition gestured when they were rehearsing the routes and most of them used their index fingers. All but one participant gestured with their right hands. On average, participants in the co-thought gesture condition spent 17.86 seconds ( \(S D=2.32\) ) to rehearse a route. Participants in the actual drawing and hand movement prohibited conditions spent 24.81 seconds ( \(S D\) \(=3.14\) ) and 19.18 seconds ( \(S D=1.19\) ) respectively. Oneway ANOVA showed that there was a significant difference in the rehearsal duration among different conditions, \(F\) (2, 82) \(=12.19, p<.001\). Tukey posthoc tests showed that the time spent on rehearsing a complete route in the actual drawing condition was significantly longer than that in the co-thought gesture, \(\mathrm{p}<.001\), and hand movements prohibited condition, \(\mathrm{p}<.002\). There was no difference between co-thought gesture and hand movement prohibited conditions, \(p s=n s\).

The mean number of steps participants rehearsed in the co-thought gesture was \(11.83(S D=3.54)\) and that in the actual drawing condition was \(11.91(S D=5.43), t(54)=.88\), ns. Thus, participants in both conditions rehearsed comparable number of steps.

We then examined the proportion of steps accurately reconstructed, which was our main interest. Figure 2 shows the mean proportion of steps correctly reconstructed in the four conditions. We conducted ANOVA with condition (cothought gesture, actual drawing, hand movement prohibited, no rehearsal) as the between-subject independent variable and the proportion of steps correctly reconstructed as the dependent variable.


Figure 2. The mean proportion of steps correctly recalled in the co-thought gesture, actual drawing, hand movement prohibited and no rehearsal conditions.

There was a significant effect of condition, \(F(3,107)=\) 12.81, \(p<.001, \eta^{2}=.35\). Planned contrasts using Bonferroni correction showed the proportion of steps correctly reconstructed in the co-thought gesture condition was higher than that in the actual drawing condition, p \(<.001\), that in the hand movement prohibited condition, \(p\) \(<.001\), and that in the no rehearsal condition, \(p<.001\). The proportion of steps correctly reconstructed in the actual
drawing condition was also higher than that in the hand movement prohibited condition, \(\mathrm{p}<.04\), and that in the no rehearsal condition, \(\mathrm{p}<.001\). Participants in the hand movement prohibited condition reconstructed more steps than those in the no rehearsal condition, \(\mathrm{p}<.02\).

On average, participants spent comparable amount of time (in seconds) in reconstructing a route in all conditions: 28.32 seconds \((S D=3.51)\) in the co-thought gesture in the air condition; 29.48 seconds ( \(S D=3.19\) ) in the actual drawing condition; 28.38 seconds ( \(S D=3.29\) ) in the hand movement prohibited condition; and 30.26 seconds ( \(S D=\) \(3.41)\) in the no rehearsal condition, \(F(3,107)=.89\), ns. Therefore, the greater reconstruction accuracy in the cothought gesture conditions was not attributed to the time spent on recall.

\section*{Discussion}

To summarize, participants who were instructed to gesture reconstructed more steps than those who were told to mentally rehearse the routes and those who did not rehearse the routes at all, suggesting that producing co-thought gestures during rehearsal facilitates encoding of spatial relation. Besides gesturing, drawing the routes on a paper also yielded better spatial recall than mentally rehearsing the routes. Therefore, hand movements produced during rehearsal, either in the forms of gestures or actual drawing, enhance spatial relation encoding, which in turn promote subsequent recall.

There are very few studies to date that show the role of gesture in encoding spatial relation or spatial learning in general, despite the fact that gesture itself is spatial in nature (McNeill, 1992) and it often represents visuo-spatial information (e.g., Alibali, 2005; Lavergne \& Kimura, 1987; McNeill, 1992; Kita \& Özyürek, 2003; So, Coppola, Liccardello, Goldin-Meadow, 2005). Of a few studies, Chu and Kita (2011) found beneficial roles of co-thought gestures in mental rotation task; Ehrlich, Levine, and Goldin-Meadow (2006) reported that frequency of cospeech gestures is positively associated with children's performance in the mental rotation task. The findings in the present study contribute to the field of gesture research in a way that producing co-thought gestures while encoding spatial information of route sequence increases recall accuracy.
Our findings converge with the embodied viewpoint of cognition. According to the theories of embodied cognition, our bodily actions are interconnected with mental representation of objects and events (e.g., Barsalou, 1999; Glenberg, 1997; Wilson, 2002). However, most of the previous studies that supported the theories of embodied cognition focused on actions on real objects. For example, Beilock, Lyons, Mattarella-Micke, Nausbaum, \& Small (2008) found that expert ice-hockey players understood hockey-language scenarios better than did hockey novices, suggesting that previous action experience facilitates the comprehension of action-related language. Casasanto and Dijkstra (2010) also reported that participants who were told
to move marbles upward retrieved positive memories more often and faster than did those who moved marbles downward, suggesting a causal link between bodily action and cognition. Previous research has also shown that actual movements improved spatial skills. For example, Weidenbauer, Schmid, and Jansen-Osmann (2007) found that participants who were trained to use a joystick to rotate two-dimensional figures had better performance in the mental rotation task than did those who were not trained. Similarly, Wexler, Kosslyn, and Berthoz (1998) showed that participants who were required to turn a joystick while solving the mental rotation task had faster response rate and higher accuracy when the direction of hand movements was congruent with the direction of mental rotation than when it was not congruent. Our findings provide additional support to the theories of embodied cognition by demonstrating that actual movements like drawing routes on papers facilitate encoding of spatial relation. However, we here take a step further and find that representational actions, i.e., gestures, produced during rehearsal also exert significant influence on encoding and retention of spatial information. Hence, embodied movements, both real and representational, would influence our spatial cognition.

In addition, although it was not part of our prediction, our findings showed that participants in the co-thought gesture condition had better recall than those in the actual drawing condition. Participants who gestured the routes recalled more steps than did those who drew them on paper. As a result, co-thought gestures seem to bring a greater impact on encoding and retaining spatial relation than actually drawing on paper. The better performance in the co-thought gesture condition, as compared to the actual drawing condition, might indicate that less concrete actions provided a greater cognitive benefit in reinforcing the spatial representation. It is possibly because producing co-thought gestures might solidify the spatial information better than drawing on paper. In the actual drawing condition, participants could see the route sequences drawn on paper and they might rely on those sequences for the rehearsal of the subsequent steps. As such, it was not necessary for them to maintain the steps actively in their memory. In contrast, participants in the cothought gesture condition did not leave visible trails. Hence, they might have to keep rehearsing the previous steps in order to proceed to the subsequent ones. As a result, producing co-thought gestures would help participants to maintain and create a richer mental representation of the path than drawing on paper. The findings might advance our understanding about the effects of different kinds of embodied movements on spatial learning. While embodied movements in general enhance spatial relation encoding, representational movements (i.e., gestures) seem to be more effective than actual movements in improving encoding and retrieval of the path information.

However, one might contend that drawing was also involved in facilitating encoding of spatial relation in the cothought gesture condition. This is because participants in this condition traced the complete paths twice while
studying the routes before they produced co-thought gestures during rehearsal. Hence, their recall performance might be attributed to dual encoding of spatial relation by drawing and gesturing (Paivio, 1971). In contrast, participants in the actual drawing condition seemed to repeat what they were doing (i.e., drawing on papers) when they were learning and rehearsing the routes. However, participants did see the complete routes on papers while they were tracing the routes on papers whereas they did not see those routes during rehearsal. As a result, although participants drew on papers when learning and rehearsing the routes, they might use different mental processes to encode spatial relation.

While our results provide strong evidence that cothought gestures play a causal role in encoding spatial relation, they do not tell us how gesture is involved in the encoding process. We propose that co-thought gestures can facilitate encoding of spatial sequence in various ways. First, they provide participants with rich sensori-motor representation of the sequence of steps (Hegarty, Mayer, Kriz, \& Keehner, 2005; Hostetter \& Alibali, 2008). In addition, they help participants to maintain the representation active in memory (de Ruiter, 1998; Wesp, Hesse, Keutmann, \& Wheaton, 2001). They also help participants to offload intermediate representations of the spatial sequence to their hands in order to reduce the chance of forgetting those representations (Chu \& Kita, 2011; Goldin-Meadow, Nasbaum, Kelly, \& Wagner, 2001).

To our knowledge, this study is the first one to test the hypothesis that co-thought gesture is more powerful than actual movement in facilitating spatial relation encoding. Producing co-thought gestures allows us to construct the spatial information and retain it in our memory with relatively little effort. Further research should also address whether gestural encoding can be applied to other spatial tasks and how long its mnemonic effect lasts for. Based on the findings in this study, however, we could start practicing moving our fingers in the air when we are learning a direction in a new environment.

\section*{Acknowledgments}

This research was supported by Faculty Supportive Fund at the Chinese University of Hong Kong.

\section*{References}

Alibali, M. W. (2005). Gesture in spatial cognition:
Expressing, communicating and thinking about spatial information. Spatial Cognition \& Computation, 5, 307331.

Barsalou, L.W. (1999). Perceptual Symbol Systems. Behavioral and Brain Sciences (22) 577-660.
Beilock, S. L., Lyons, I. M., Mattarella-Micke, A., Nusbaum, H. C., \& Small, S. L. (2008). Sports experience changes the neural processing of action language. Proceedings of the National Academy of Sciences, USA, 105, 13269-13273.

Broaders, S., Cook, S.W., Mitchell, Z., \& Goldin-Meadow, S. (2007). Making children gesture reveals implicit knowledge and leads to learning. Journal of Experimental Psychology: General, 136, 539-550.
Casasanto, D. \& Dijkstra, K. (2010). Motor action and emotional memory. Cognition, 115(1), 179-185.
Chu, M., \& Kita, S. (2011). The nature of the beneficial role of spontaneous gesture in spatial problem solving. Journal of Experimental Psychology: General, 140, 102-116.
Cook, S. W., Mitchell, Z., Goldin-Meadow, S. (2008). Gesturing makes learning last. Cognition, 106, 1047-1058.
de Ruiter, J.P. (1998). Gesture and speech production. Doctoral dissertation at Catholic University of Nijmegen, Netherlands.
Ehrlich, S. B., Levine, S., \& Goldin-Meadow, S. (2006). The importance of gesture in children's spatial reasoning. Developmental Psychology, 42, 1259-1268.
Emmorey, K., Tversky, B. \& Taylor, H.B. (2000). Using space to describe space: Perspective in speech, sign and gesture. Spatial Cognition and Computation, 2, 157-180.
Glenberg, A. M. (1997). What memory is for. Behavioral \& Brain Sciences, 20, 1-55.
Goldin-Meadow, S., Nusbaum, H., Kelly, S.D., \& Wagner, S. (2001). Explaining math: Gesturing lightens the load. Psychological Science, 12, 6, 516-522
Hegarty M, Mayer S, Kriz S, \& Keehner M. (2005). The role of gestures in mental animation. Spatial Cognition and Computation, 5:333-56
Hostetter, A. B., Alibali, W. M., \& Kita, S. (2007). I see it in my hand's eye: Representational gestures are sensitive to conceptual demands. Language and Cognitive Processes, 22(3), 313-336.
Hostetter, A. B., \& Alibali, M. W. (2008). Visible embodiment: Gestures as simulated action. Psychonomic Bulletin \& Review, 15(3), 495-514
Kita, S. \& Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. Journal of Memory and Language, 48, 16-32.
Lavergne, J. \& Kimura, D. (1987). Hand movement asymmetry during speech: no effect of speaking topic. Neuropsychologia, 25, 689-693.
Lohman, D. F. (1979). Spatial ability: Review and reanalysis of the correlational literature. Stanford University Technical Report No. 8.
Melinger, A. \& Kita, S. (2007). Conceptual load triggers gesture production. Language and Cognitive Processes, 22, 4, 473-500.
McNeill, D. (2005). Gesture and Thought. Chicago: University of Chicago Press.
Paivio, A (1971). Imagery and verbal processes. New York: Holt, Rinehart, and Winston.
Sauter, M., Uttal, D.H., Alman, A.S., Goldin-Meadow, S., Levine, S.C. (2012).Learning what children know about space from looking at their hands: the added value of
gesture in spatial communication. Journal of Experimental Child Psychology, Apr;111(4):587-606
So, W.C., Coppola, M., Licciardello, V. \& Goldin-Meadow, S. (2005). The seeds of spatial grammar in the manual modality. Cognitive Science, 29, 1029-1043.
So, W. C., Kita, S., \& Goldin-Meadow, S. (2009). Using the Hands to Identify Who Does What to Whom: Gesture and Speech Go Hand-in-Hand. Cognitive Science, 33, 115125.

Weidenbauer, G., Schmid, J., \& Jansen-Osmann, P. (2007). Manual training of mental rotation. European Journal of Cognitive Psychology, 19(1), 17-36.
Wesp, R., Hesse, J., Keutmann, D., \& Wheaton, K. (2001). Gestures maintain spatial imagery. The American Journal of Psychology, 114, 591-600.
Wexler, M., Kosslyn, S. M., \& Berthoz, A. (1998). Motor processes in mental rotation. Cognition, 68, 77-94
Wilson, M. (2002). Six views of embodied cognition. Psychonomic Bulletin \& Review, 9, 625-636.

\title{
Aspectual Coercion in Non-native Speakers of English
}

\author{
Ho Leung Chan (ellchl@nus.edu.sg) \\ Department of English Language \& Literature, National University of Singapore, Singapore
}

\begin{abstract}
This study examined the processing correlates of aspectual coercion among native and non-native speakers of English. For native English speakers, results suggested that the processing delay associated with aspectual coercion is minimal. Aspectual coercion was perhaps cognitively easy to perform. By contrast, non-native speakers of English from unlike first language (L1) backgrounds differed in their reading performance. The differences varied systematically as a function of aspectual contrasts in L1 after controlling for second language (L2) English proficiency. Korean participants showed trends of aspectual coercion despite the absence of significant effects; German participants exhibited indifference across experimental conditions; Chinese participants showed aspectual coercion effects opposite to the predictions specified by the English grammar. A coupling of these data with evidence from the semelfactive progressive (e.g. coughing) in English suggests that the socalled online aspectual coercion effects may arise from a prototype organization of aspectual categories that is prone to L1 influence.
\end{abstract}

Keywords: Aspectual coercion; semelfactive progressive; prototype; L1 transfer

\section*{Introduction}

The study of aspectual coercion in non-native speakers provides an unusual opportunity to understand how aspectual conflicts are recognized and resolved in the course of language processing. Presumably, the challenge for nonnative speakers to process subtle semantic nuances on the fly is far greater than that of native speakers. If aspectual coercion incurs an extra processing cost, it will be more likely to find evidence of that in non-native speakers than in native speakers. This may in turn shed light on aspectual coercion research conducted on native speakers that have reported mixed findings in the literature.

\section*{Aspectual Coercion}

Verbs denote events that take place in time. A semelfactive verb (e.g. cough) denotes a single-stage, atelic situation (Smith, 1991; 1997). When the semelfactive verb cough combines with an adverbial modifier of duration or durative adverbial for an hour in the sentence Sam coughed for an hour, the combination becomes problematic. The semelfactive verb and durative adverbial are aspectually incompatible with each other. However, the sentence is neither ill-formed nor ungrammatical. Often an iterative interpretation is derived, namely Sam coughed repeatedly for an hour. Researchers have hypothesized that a computational process is invoked to resolve the incompatibility and construe a more coherent interpretation.

Such a process is commonly known as aspectual coercion. The discussion of coercion phenomena first appeared in Moens and Steedman (1988).
Empirical studies to date have yet to provide conclusive evidence about the processing consequence of aspectual coercion. Also, it remains unexplored that semelfactive progressive (e.g. coughing) in English derives an iterative interpretation even in the absence of durative adverbials. Whether a construction such as Sam was coughing for an hour incurs an extra processing cost or not, and how nonnative speakers respond to aspectual coercion relative to native speakers, become the twin goals of this study. It is hypothesized that a greater processing cost can be found in non-native speakers than in native speakers if aspectual coercion is computationally costly. Another prediction is that a construction like Sam was coughing for an hour will not incur an extra processing cost, because there is not any aspectual mismatch between the verbal predicate and the adverbial in the first place.

\section*{Psycholinguistic Evidence}

A small number of empirical studies have examined the psycholinguistic evidence of aspectual coercion using behavioral and brain-imaging techniques. The reported findings were mixed. Some studies found longer decision times and/or higher reading latencies in cases of aspectual coercion, while others reported null results.
Task differences may be responsible for these dissimilar findings. Piñango, Zurif, and Jackendoff (1999) first examined the processing load associated with aspectual coercion using a cross-modal lexical decision task. They found longer decision times at the probe positions of coercion sentences. Todorova, Strab, Bedecker, and Frank (2000) employed a self-paced, makes-sense judgment task, and found higher reading latencies in coercion sentences. Piñango, Winnick, Ullah, and Zurif (2006) later reported that online effects of aspectual coercion could only be found when a secondary lexical decision task was administered at a delayed interval of 250 ms . Unlike previous paradigms, Pickering, McElree, Frisson, Chen, and Traxler (2006) employed the self-paced reading and eye-tracking techniques to foster more naturalistic reading in experimental settings. Nonetheless, Pickering et al. found no behavioral differences in terms of reading times and other eye-tracking estimates across conditions. The researchers attributed the null results to an underspecification account, which claimed that native English speakers did not commit to the telicity of situations immediately during normal sentence comprehension.

Another challenge stems from lexical aspect. Lexical aspect (or Aktionsart) refers to the temporal meanings
inherent in verbal predicates. A semelfactive verb (e.g. cough) is, by definition, semelfactive because it conveys a single-stage, atelic situation (Smith, 1991; 1997). The classification of lexical aspect can be tricky. Brennan and Pylkkänen (2008) addressed this by first norming a selection of verbs for punctuality with native English speakers. Only a set of strongly punctual verbs (all semelfactives indeed) was then chosen for their self-paced reading and MEG (i.e. magnetoencephalography) experiments. As illustrated in (1), the critical sentences varied in adverbial type (either a durative adverbial or a punctual adverbial), followed by a genuine semelfactive verb:
(1) a. Throughout the day the student sneezed in the back of the classroom.
b. After twenty minutes the student sneezed in the back of the classroom.

Crucially, Brennan and Pylkkänen reported significantly longer reading times at the inflected verb in 1a than in 1 b in native English speakers from the self-paced reading experiment.

In addition to tasks and experimental control, previous empirical studies have only narrowly examined aspectual coercion in which a punctual situation is interpreted as iterative by means of an interaction with a specific type of temporal modifier, namely durative adverbial. Little is known about how other factors such as grammatical aspect could affect aspectual coercion. Comrie (1976) stated that grammatical aspect encodes different ways of viewing the internal temporal constituency of a situation. One can therefore distinguish what happened from what was happening owing to grammatical aspect, which is often marked via verbal morphology. In this light, semelfactive progressive in English (e.g. coughing) provides an exceptional window to elucidate this issue, precisely because it denotes iterative action-in-progress (Smith, 1991; 1997) without any temporal adverbials. A psycholinguistic investigation along this line shall cast some light on the study of aspectual coercion phenomena at large.

\section*{Experiment}

This study examined the influence of grammatical aspect and temporal adverbials on aspectual coercion in the course of language processing. The two research questions were:
1. Does aspectual coercion incur an extra processing cost in native and non-native speakers of English?
2. Does grammatical aspect mediate the online effects of aspectual coercion? In other words, is there a trade-off between grammatical aspect and temporal adverbial?

\section*{Method}

Participants Participants consisted of native English speakers and non-native speakers of English from Korean, Mandarin Chinese, and German L1 backgrounds. The
profiles of the participants were summarized as follows: native English speakers ( 15 women, 9 men, \(M_{\text {age }}=20.2\) years, age range: 18-25 years); Korean ( 14 women, 1 man, \(M_{\text {age }}=21.7\) years, age range: 18-29 years); Chinese (16 women, 5 men, \(M_{\text {age }}=23.8\) years, age range: 21-30 years). These participants enrolled at the University of Pittsburgh and Carnegie Mellon University in the US. The German participants ( 21 women, 4 men, \(M_{\text {age }}=25.5\) years, age range: 20-41 years) were recruited from the Ruprecht-KarlsUniversität Heidelberg, Germany. All participants took part in the experiment for compensation.

Also, all non-native speakers completed a standardized English proficiency test - the Michigan Test of English Language Proficiency (MTELP) of the Michigan Test Battery (Corrigan, Dobson, Kellman, Spaan, Strowe, \& Tyma, 1979). The maximum score was 100 . At the time of testing, both Korean ( \(M=81.13, S E=3.67\) ) and German ( \(M\) \(=79.88, S E=2.85\) ) participants were more proficient in English than their Chinese \((M=63, S E=3.11)\) counterparts, \(p \mathrm{~s}=.001\).

Stimuli Twenty-four sentences were constructed from seventeen semelfactive verbs. These verbs were selected based on the norming results of punctuality (Brennan \& Pylkkänen, 2008) as well as ratings for telicity (Wulff, Ellis, Römer, Bardovi-Harlig, \& Leblanc, 2009). Participants’ knowledge of English was also taken into consideration. The experiment implemented a \(2 \times 2\) design crossing Grammatical Aspect (SIMPLE, PROG) and Adverbial (Punctual, Durative). Here, SIMPLE means the grammatical aspect is unspecified, whereas PROG denotes the progressive aspect. All critical items were distributed into 4 lists such that each list contained one token of each of the 24 critical items and six items from each of the four conditions. The 4 sets of experimental stimuli were each embedded into a list of 120 filler sentences, plus an additional 84 items from two other experiments (Chan, 2012). Presentation orders were completely randomized. Table 1 summarizes the quadruple design crossing grammatical aspect and adverbial.

Table 1: Conditions and sample stimuli
\begin{tabular}{|c|c|c|}
\hline & Punctual adverbial & Durative adverbial \\
\hline SIMPLE & At noon the kid jumped into the swimming pool. & All day the kid jumped into the swimming pool. \\
\hline PROG & At noon the kid was jumping into the swimming pool. & All day the kid was jumping into the swimming pool. \\
\hline
\end{tabular}

Condition A is a control condition. It serves as a baseline for condition B , in which there is an aspectual mismatch between the durative adverbial all day and the verb jumped. Condition B is an example of aspectual coercion. Previous studies described this as an instance of iterative coercion
(e.g., Brennan \& Pylkkänen, 2008). Because iteration is triggered by an adverbial, condition B is called Adverbial Coercion. This term is not new; it first appeared in Todorova et al (2000). Conditions C and D represent new manipulations, as they have never been tested before. As discussed earlier, Smith \((1991\); 1997) asserted that semelfactive verbs marked in English progressive denote iterative action-in-progress. The combination of the semelfactive progressive jumping with the punctual adverbial at noon therefore creates an aspectual conflict in condition C. Although conditions B and C both involve aspectual mismatch, the iteration in condition C arises via the progressive marking on the semelfactive verb, whereas that in condition B is enforced by a durative adverbial external to the verbal predicate. To differentiate between the two, condition C is therefore called Grammatical (Aspect) Coercion. Lastly, condition \(D\) serves as a baseline to Grammatical Coercion in condition C.

Two predictions are made for native English speakers. First, Adverbial Coercion sentences (Condition B) will incur longer reading times than respective control sentences (Condition A). This is based on previous findings that durative adverbials could trigger coercion (e.g. Todorova et al., 2000). Second, a new prediction for this study is that Grammatical Coercion sentences (Condition C) will take longer time to read than control sentences (Condition D) because of the aspectual mismatch. Therefore, the \(2 \times 2\) design predicts aspectual coercion effects as a trade-off between adverbial and grammatical aspect.
It is generally assumed that native and non-native speakers will have similar reading performance. Of course, one may predict some variations as a result of L1 differences. For example, Korean participants may perform similarly to native English speakers, considering the many meaning overlaps between aspectual systems of Korean and English. Because German lacks a grammatical aspectual system, a reasonable prediction is that German participants may not exhibit any differences between SIMPLE and PROG. The Chinese participants may as well perform similarly to native English speakers. It must be emphasized that Chinese has richer perfective and imperfective contrasts, in addition to the optional marking system. The progressive aspect is obligatory in English, however. Given these cross-linguistic differences in aspectual meaning and grammar, the above predictions are speculative at best.

Procedure The current study involved a computerized selfpaced reading experiment, which was administered individually to participants in a laboratory setting. Participants were instructed to read English sentences as quickly as possible and answer comprehension questions as accurately as possible. Six practice items were given before the actual experiment.

The self-paced reading task was implemented on Linger software (Rohde, 2001), following a word-by-word noncumulative moving window paradigm presentation technique (Just, Carpenter, \& Woolley, 1982). Each
sentence was masked by a series of dashes (-). These dashes were replaced by a word from left-to-right every time the participant pressed the space bar. Only one word was shown on the computer screen at a time.
An optional break was provided in between every 50 trials. To ensure meaningful comprehension, a yes/no comprehension question prompt was presented to each of the 120 filler sentences embedded throughout the experiment. Feedback on accuracy was also provided. The majority of participants finished the experiment in an hour.

\section*{Data Analysis}

The following procedures were employed. First, I ascertained that all participants scored \(90 \%\) or above for the comprehension questions. For native English speakers, the mean accuracy was \(94 \%\) ( \(S D=3.9 \%\) ). For non-native speakers, Korean participants achieved a mean accuracy of \(94.2 \% ~(S D=3.4 \%)\), German 94.6\% ( \(S D=2.9 \%\) ), and Chinese \(91 \%\) ( \(S D=4.8 \%\) ). The overall high accuracy confirms that all participants paid attention and read the sentences carefully.
Next, extreme reading times (RTs), including those shorter than 100 ms or longer than \(2,500 \mathrm{~ms}\) per word, were discarded. These criteria excluded \(0.59 \%, 1.36 \%, 1.26 \%\), and \(2.31 \%\) of data points among the English, Korean, German, and Chinese participants, respectively.
RTs were then transformed logarithmically. A linear regression was performed on the \(\log\) RT data to correct for word length differences across conditions, while taking into account each participant's individual reading speed. This procedure utilized all words from experimental items and fillers for each participant (e.g., Ferreira \& Clifton, 1986; Trueswell, Tanenhaus, \& Garnsey, 1994). The values predicted from the regressions were subtracted from the actual reading times to produce residual reading times for each participant. Thus, word-length adjusted residual \(\log\) RTs became the dependent variable for subsequent statistical tests.
Separate analyses were conducted on RTs at four target word regions: 1) the verb; 2) the first word following the verb \((\mathrm{V}+1)\) to capture spill-over effects; 3 ) the second word following the verb \((\mathrm{V}+2)\) to assess further downstream effects; and 4) the sentence-final (SF) word to investigate sentence wrap-up effects (Just \& Carpenter, 1980).
Furthermore, a number of problematic items were excluded from statistical analyses. All trials containing yesterday, last night, last week and the verb open that were intended to serve as punctual adverbials and semelfactives were excluded. This procedure reduced the entire data set by another \(36.98 \%\).

\section*{Results}

A three-way mixed-design ANCOVA was performed with Grammatical Aspect (SIMPLE, PROG) and Adverbial (Punctual, Durative) as within-participant variables, group (English, Korean, German, and Chinese) as a between-


Figure 1: ANCOVA reading time results. Error bars indicate \(\pm 1\) standard error.
participant variable, and English proficiency as a covariate. A default score of 100 was entered for native English speakers in the covariate for ANCOVA analyses. An \(\alpha\)-level of .05 was used. Figure 1 plots the ANCOVA RT results for each participant group.

Verb An ANCOVA controlling for English proficiency at the verb revealed a significant Adverbial \(\times\) Grammatical Aspect interaction by both participants and items, \(F_{I}(1,80)\) \(=5.773, p=.019 ; F_{2}(1,56)=4.607, p=.036\). The main effect of grammatical aspect was significant by participants, \(F_{l}(1,80)=7.265, p=.009 ; F_{2}<3.072\). No other effects approached significance by either participants or items: adverbial, \(F_{l}(1,80)=3.757, p=.056 ; F_{2}<.071\); language, \(F_{l}(3,80)=1.412, p=.245 ; F_{2}<.892\); all interactions, \(F \mathrm{~s}<\) 1.128.

To explore the Adverbial \(\times\) Grammatical Aspect interaction collapsed across language groups, a follow-up simple main effect of adverbial across levels of grammatical aspect was performed in this word region. However, none of the comparisons reached significance, \(p \mathrm{~s}>.113\).

V+1 An ANCOVA controlling for English proficiency at the first word after the verb revealed a significant Adverbial \(\times\) Grammatical Aspect interaction by participants, \(F_{l}(1,80)\) \(=5.036, p=.028 ; F_{2}<.002\). The main effect of language was significant by both participants and items, \(F_{l}(3,80)=\) 5.456, \(p=.002 ; F_{2}(3,56)=2.893, p=.043\). All other main effects and interactions were not significant by either participants or items: grammatical aspect, \(F_{l}(1,80)=.119\), \(p=.731 ; F_{2}<.559\); adverbial, \(F_{l}(1,80)=1.605, p=.209\); \(F_{2}<.818\); all interactions, \(F \mathrm{~s}<1.025\).

To explore the Adverbial \(\times\) Grammatical Aspect interaction, a follow-up simple main effect of adverbial across levels of grammatical aspect was conducted in this word region. No comparisons approached significance, \(p \mathrm{~s}>\) .492.

In order to understand how different language groups performed in this word region, posthoc pairwise comparisons with Bonferroni adjustment revealed that the reading speed of Chinese participants \((M=.041, S E=.011)\)
was significantly slower than that of native English participants \((M=-.03, S E=.01), p=.001\), as well as German participants \((M=.000, S E=.008), p=.014\), respectively. No other comparisons were significant, \(p \mathrm{~s}>\) . 107.
\(\mathbf{V + 2}\) An ANCOVA controlling for English proficiency at the second word after the verb revealed a significant Adverbial \(\times\) Grammatical Aspect interaction by participants, \(F_{I}(1,80)=11.736, p=.001 ; F_{2}<.39\), and a three-way Adverbial \(\times\) Grammatical Aspect \(\times\) Language interaction by participants, \(F_{l}(3,80)=3.251, p=.026 ; F_{2}<.169\), suggesting that the four language groups may behave differently across levels of adverbial and grammatical aspect. All other main effects and interactions were not significant by either participants or items: grammatical aspect, \(F_{l}(1,80)=.848, p=.36 ; F_{2}<.038 ;\) adverbial, \(F_{l}(1\), \(80)=.923, p=.34 ; F_{2}<.143\); language, \(F_{l}(3,80)=1.025\), \(p=.386 ; F_{2}<.149\); all interactions, \(F \mathrm{~s}<2.206\).

Because of a significant three-way interaction, a followup simple main effect of adverbial across levels of grammatical aspect was performed separately for each language group in this word region. Native English speakers slowed down at Adverbial Coercion sentences ( \(M=-.011\), \(S E=.017\) ) relative to corresponding control sentences \((M=\) \(-.061, S E=.019), p=.052\). Also, they read Grammatical Coercion sentences \((M=-.011, S E=.016)\) marginally slower than respective control sentences \((M=-.047, S E=\) .016 ), \(p=.096\). Although these results were only marginally significant, native speakers in this experiment behaved in accord with the prediction that sentences involving aspectual coercion generally took longer to read than noncoercion sentences. These results provided a reasonable baseline when evaluating non-native speakers' reading performance in the same experiment.
Unexpectedly, Chinese participants read Adverbial Coercion sentences ( \(M=-.071, S E=.018\) ) significantly faster than the respective control sentences \((M=.01, S E=\) .02 ), \(p=.003\), which is opposite to the prediction of adverbial coercion. All other comparisons were not significant in this word region, \(p \mathrm{~s}>.164\).

SF An ANCOVA controlling for English proficiency at the sentence final word revealed a significant main effect of language by both participants and items, \(F_{l}(3,80)=7.122\), \(p<.001 ; F_{2}(3,56)=11.948, p<.001\). All other main effects and interactions were not significant by either participants or items: grammatical aspect, \(F_{I}(1,80)=1.207\), \(p=.275 ; F_{2}<.952 ;\) adverbial, \(F_{l}(1,80)=.155, p=.695 ; F_{2}\) \(<.049\); all interactions, \(F \mathrm{~s}<1.549\).

To explore how different language groups performed in this word region, posthoc pairwise comparisons with Bonferroni adjustment revealed that the reading speed of German participants ( \(M=.162, S E=.016\) ) was significantly slower than that of native English speakers \((M=.066, S E=\) \(.021), p=.004\), and Chinese participants \((M=.077, S E=\) .022 ), \(p=.01\), respectively. German participants also appeared to be slower than Korean participants ( \(M=.098\), \(S E=.02\) ), \(p=.087\). All other comparisons were not significant, \(p \mathrm{~s}>.087\).

Taken together, the ANCOVA analyses revealed several interesting results. At the verb, there were no RT differences by condition collapsed across language groups. The same was generally true for \(\mathrm{V}+1\). At \(\mathrm{V}+2\), native English speakers exhibited marginally significant trends that Adverbial Coercion and Grammatical Coercion sentences took longer time to read. The Chinese participants, however, took significantly longer to read control sentences relative to Adverbial Coercion sentences \((p=.003)\). At the sentence final word, the German participants showed elevated RTs across all conditions when compared to Korean, Chinese, and native English participants.

\section*{General Discussion}

Overall, results provided partial support for the first research question that there is a processing cost for aspectual coercion. Native English speakers had the tendency to slow down at sentences involving Adverbial Coercion (All day the kid jumped...) and Grammatical Coercion (At noon the kid was jumping...), even though the reading time results were only marginally different from their control counterparts. These findings were consistent with the general prediction that aspectual coercion may incur a somewhat greater processing cost. However, it is noted that these results were delayed, and emerged only at the second word after the verb \((\mathrm{V}+2)\). It is unclear why no strong, immediate online effects emerge as other self-paced reading studies have shown. Brennan and Pylkkänen (2008), for example, presented evidence that iterative coercion can produce significant, immediate effects. The highly salient semelfactive verbs used in this experiment may have been responsible for the diminished online effects within native English speakers.

Despite these suggestive findings, one unambiguous result was that aspectual coercion is mediated by the interaction among grammatical aspect, lexical aspect, and adverbial. This is evident in the significant Adverbial \(\times\) Grammatical Aspect interaction effect collapsed across language groups
at three of the four word regions probed, \(p \mathrm{~s}<.028\). As predicted, not only adverbials but also grammatical aspect triggers aspectual coercion. This finding provides a new theoretical insight to aspectual coercion phenomena, as previous studies showed that a durative adverbial is responsible for iteration involved in a semelfactive predicate (e.g., Todorova et al., 2000). Here, results clearly show that there is no reason to believe that temporal adverbials independently cause processing slowdown. Instead, lexical aspect, grammatical aspect, and adverbial conspire to shape the aspectual interpretation of a sentence.
Moreover, the findings here were at odds with the underspecification account put forward by Pickering et al. (2006) to account for their null results. Pickering et al. asserted that readers routinely underspecify aspectual properties of an interpretation during comprehension. The underspecification account seemed to be untenable here because of the strong interaction between adverbial and grammatical aspect. In this light, the current study is more compatible with Brennan and Pylkkänen (2008), among others.
For the first time, this study extended the psycholinguistic investigation of aspectual coercion to non-native speakers. Although native English speakers behaved differently from non-native speakers in general, non-native speakers also differed systematically from one another after removing preexisting differences in L2 English proficiency. For example, the Chinese participants showed significantly shorter reading times in Adverbial Coercion (All day the kid jumped...) sentences than control sentences (At noon the kid jumped...), which was opposite to the prediction ( \(p=.003\) ). These results seemed puzzling at first glance. One potential explanation may rest on the differences in the aspectual systems of English and Chinese. According to Yang (1995), the perfective marker le strongly prefers telic and bounded situations in Chinese. This explains why the semelfactive predicate kesou 'cough' in (2) cannot felicitously co-occur with \(l e\), because semelfactives are by definition atelic (i.e., Activities). However, when a bounded temporal situation is introduced via a verbal classifier phrase yi-sheng 'once' as shown in (3), the utterance becomes felicitous.
(2) *Lisi kesou le

Lisi cough PERF
"Lisi coughed"
(3) Lisi kesou le yi-sheng Lisi cough PERF one-CL "Lisi coughed once"
(Yang, 1995; cited in Xiao \& McEnery, 2004, p. 103)
Xiao and McEnery (2004) adduced native Chinese corpus data to support the idea that the sensitivity of le to boundedness is relative rather than absolute. In their sample, an overwhelming \(89.4 \%\) of all 1138 tokens of le occur in bounded contexts, whereas a meager \(10.6 \%\) occurred in unbounded contexts. Of the 27 tokens of semelfactives taking \(l e\) in the same corpus, 16 are bounded by additional
adverbials that impose a spatially or temporally bounded situation. This distributional pattern suggests that the semelfactive plus durative adverbial combination is quantitatively more common in Chinese. Although Xiao and McEnery did not articulate the underlying reason for such a language-specific bias in Chinese, they maintained that semelfactive verbs taking perfective \(l e\) prefers to be bounded, particularly by means of a verbal classifier phrase, verb reduplication, or by a for-adverbial as shown in (4).
(4) Da-le ni ji-tian
beat-PERF you how-many-day
"For how many days did they beat you?"
(Xiao \& McEnery, 2004, p. 111)
The co-occurrence of a semelfactive verb taking \(l e\) in the presence of a durative adverbial in (4) is equivalent to the Adverbial Coercion construction in English. Accordingly, the processing advantage found in Chinese participants can be attributed to the skewed distribution of \(l e\) in bounded contexts for semelfactives. Although I consider such a possibility using Xiao and McEnery's Chinese corpus data, future Chinese sentence processing experiments will need to independently verify this claim. What is remarkable here is that Chinese participants exhibited a language-specific bias from their L1 Chinese aspectual system even when they were reading in English. If that's the case, results from Chinese participants provided crucial support for L1 transfer.
Korean participants did not show any within-subject differences in terms of their reading performance across experimental conditions. They exhibited trends of aspectual coercion, despite the absence of statistically significant results. The same can be said about the German participants in which the reading times performance were highly comparable across conditions. I reckon that the lack of grammatical aspect (and associated grammaticized meanings) is responsible for their indifference (e.g., Stutterheim \& Carroll, 2006).

\section*{Conclusion}

What is interesting about aspectual coercion is that it involves contextual re-interpretation of aspectual information, rendering it computationally more demanding for non-native speakers. Although this prediction has not been borne out in all non-native speaker groups, the current study revealed important L1-based variations that could not have been exposed otherwise. A psycholinguistic investigation comparing both the performance of native and non-native speakers can thus reveal rather than obscure aspectual coercion operations in the course of language processing. Importantly, I contend that the so-called online aspectual coercion effects arise from a prototype organization of aspectual categories, which is, not surprisingly, prone to L1 influence in a systematic way.

\section*{Acknowledgments}

I thank my advisors Shirai Yasuhiro and Charles Perfetti, three anonymous reviewers, and the Language Learning Dissertation Grant for financial support.

\section*{References}

Brennan, J., \& Pylkkänen, L. (2008). Processing events: Behavioral and neuromagnetic correlates of aspectual coercion. Brain \& Language, 106, 132-143.
Chan, H. L. (2012). Tense-aspect Processing in Second Language Learners. University of Pittsburgh, Pittsburgh.
Comrie, B. (1976). Aspect: An introduction to the study of verbal aspect and related problems. New York: Cambridge University Press.
Corrigan, A., Dobson, B., Kellman, E., Palmer, A., Peterson, J., Spaan, M., et al. (1979). English placement test. Ann Arbor, MI: University of Michigan English Language Institute.
Ferreira, F., \& Clifton, C. (1986). The independence of syntactic processing. Journal of Memory and Language, 25, 348-368.
Just, M. A., \& Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. Psychological Review, 87, 329-354.
Just, M. A., Carpenter, P. A., \& Wooley, J. D. (1982). Paradigms and processes in reading comprehension. Journal of Experimental Psychology: General, 111, 228-238.
Moens, M., \& Steedman, M. (1988). Temporal ontology and temporal reference. Computational Linguistics, 14, 15-28.
Pickering, M. J., McElree, B., Frisson, S., Chen, L., \& Traxler, M. J. (2006). Underspecification and aspectual coercion. Discourse Processes, 42, 131-155.
Piñango, M. M., Winnick, A., Ullah, R., \& Zurif, E. (2006). Timecourse of semantic composition: The case of aspectual coercion. Journal of Psycholinguistic Research, 35, 233-244.
Piñango, M. M., Zurif, E., \& Jackendoff, R. (1999). Real-time processing implications of enriched composition at the syntaxsemantics interface. Journal of Psycholinguistic Research, 28, 395414.

Rohde, D. (2001). LINGER (Version 2.9.4). Retrieved from http://tedlab.mit.edu/~dr/Linger/
Smith, C. S. (1991). The parameter of aspect. Dordrecht: Kluwer Academic Publishers.
Smith, C. S. (1997). The parameter of aspect (2 \({ }^{\text {nd }}\) Ed.). Dordrecht: Kluwer Academic Publishers.
Stutterheim, C. von, \& Carroll, M. (2006). The impact of grammatical temporal categories on ultimate attainment in L2 learning. In H. D. Weger-Guntharp, H. Byrnes \& K. A. Sprang (Eds.), Educating for advanced foreign language capacities: Constructs, curriculum, instruction, assessment. Washington DC, USA: Georgetown University Press.
Todorova, M., Straub, K., Badecker, W., \& Frank, R. (2000). Aspectual coercion and the online computation of sentential aspect. In Proceedings of the twenty-second annual conference of the cognitive science society. Mahwah, NJ: Lawrence Erlbaum Associates.
Trueswell, J. C., Tanenhaus, M. K., \& Garnsey, S. M. (1994). Semantic influences on parsing: Use of thematic role information in syntactic disambiguation. Journal of Memory and Language, 33, 285-318.
Wulff, S., Ellis, N. C., Römer, U., Bardovi-Harlig, K., \& Leblanc, C. J. (2009). The acquisition of tense-aspect: Converging evidence from corpora and telicity ratings. The Modern Language Journal, 93, 354-369.
Xiao, R., \& McEnery, T. (2004). Aspect in Mandarin Chinese: A corpus-based study. Amsterdam: John Benjamins.

\title{
Space, agency and word order: Evidence from Greek
}

\author{
Spyridoula Cheimariou (spyridoula-cheimariou@uiowa.edu) \\ Department of Communication Sciences and Disorders, University of Iowa \\ 250 Hawkins Drive, Iowa City, IA 52242 USA \\ Sonia Loui (sonialoui@gmail.com) \\ Department of Philosophy \& History of Science (MITHE), Ano Ilissia University Campus \\ GR-15771 Zografos, Greece
}

\begin{abstract}
We examined the role of spatial representations and word order on thematic role assignment in Greek. Previous studies suggest that spatial representations influence thematic role assignment; agent is typically depicted on the left, and patient on the right. Here, we address this issue using a language with flexible word order which allows us to manipulate sentence structure (SVO-OVS) orthogonally to thematic role. Greek speakers heard SVO/OVS sentences while viewing depictions of actions involving two characters and they judged whether sentence and picture matched in meaning. The agent's position in the picture was directly manipulated. The results support the effect of left bias on language processing. However, this bias may be better understood when its interaction with other sources of information and languagespecific constraints are taken into account. Theories of prediction may help us illuminate how spatial biases and linguistic factors interactively affect the way we process our world.
\end{abstract}

Keywords: spatial representation, language, thematic role assignment, word order, sentence comprehension, prediction, Greek.

\section*{Introduction}

The combinatory study of language and space aims to shed light on how an analog, geometric and continuous representation is encoded into a propositional algebraic and discrete representation (Jackendoff, 1992; Geminiani, Bisiach, Berti \& Rusconi, 1995; Hayward \& Tarr, 1995; Jackendoff, 1996; Chatterjee, Southwood \& Basilico, 1999; Levinson, 2003; Papafragou, Hulbert \& Trueswell, 2008, among others). Recent research has inaugurated a discussion on how language structures are constrained by spatial biases (Chatterjee, Maher, Gonzales-Rothi, \& Heilman, 1995; Chatterjee, Southwood, \& Basilico, 1999). Researchers (Chatterjee et., al, 1995; Chatterjee, Southwood, \& Basilico, 1999; Chatterjee, 2001; 2008) have entertained the claim that events are conceptualized spatially and prelinguistically proceeding from left towards the right. They assume that language development exploits systems meant for left-toright spatial attention in the left hemisphere. Therefore, the left-to-right directional bias indicates primitive spatial representations and reflects a prelinguistic neural encoding of events and actions.

Evidence for this claim comes from case studies in agrammatic speech (Caplan \& Futter, 1986; Caramazza \& Miceli, 1991). In these studies, the person with
agrammatism systematically assigned agency to the first noun heard or to the one located to the left of the verb. Chatterjee and colleagues (Chatterjee, et al. 1995; Chatterjee, Southwood, \& Basilico, 1999), based on the Jacksonian notion that primitive cognitive functions are overlaid with more complex functions, hypothesized that people with agrammatic aphasia tend to follow a temporal or spatial strategy based on those primitive spatial representations in order to interpret a sentence once their more complex linguistic abilities fail. To test for their assumption they conducted studies in typical population (Chatterjee et al., 1995; Chatterjee, Southwood, \& Basilico, 1999; Barrett \& Craver-Lemley, 2008). In these studies, participants were asked to depict sentences describing an action with two persons involved or to match sentences to pictures. Participants tended to draw the agent of an action closer to the left side of the picture. For example, in the sentence «The girl chased the boy», it was more probable for the participants to depict the girl on the left side of the picture and the boy on the right side. Also, participants responded faster when the agent was located on the left side in sentence-picture matching tasks.

According to an alternative explanation, the left-to-right directional bias is culturally determined by the directionality of the reading/writing system (Maass \& Russo, 2003; Chan \& Bergen, 2005; Dobel, Diesendruck, \& Bolte, 2007). Specifically, Maass and Russo (2003) investigated spatial biases in thematic role assignment in directionally opposite writing system, such as Italian (left-to-right) and Arabic (right-to-left). They found that Italian speakers tended to assign the agent on the left, while Arabic speakers had the reverse tendency. Furthermore, the more years Arabic speakers had spent abroad exposed to the opposite writing system, the more mitigated their right-to-left bias was.

Furthermore, Dobel, Diesendruck, and Bolte (2007) strengthened this argument by showing that the left or right bias in depicting agency is based not only on reading and writing practices, but also on the degree of exposure on those practices. Specifically, they tested spatial biases in German- and Hebrew-speaking adults and preschool children. They found that the writing system influenced thematic role assignment in adults, that is, German-speaking adults had a left-to-right spatial bias, while Hebrewspeaking adults had the opposite bias. However, this was not observed for preschool children who had no exposure to the reading and writing systems of their language.

The previous studies suggest that thematic role assignment is affected conceptually and spatially by the reading and writing systems. The present study extends the role of language on thematic role assignment bias by adding word order, that is, the within-sentence structure, as a possible variable. The methodological paradigm used in previous research (Chatterjee et al., 1995; Chatterjee, Southwood, \& Basilico, 1999) was based on English, a language with highly restrictive word order. To manipulate temporally or spatially the agent or the patient of an action, previous studies either used active and passive voice (Chatterjee et al., 1995), or verbs with different trajectories (e.g. "The circle pushes the square", in which the circle is the agent and the action moves away from the agent, "The circle pulls the square" in which the circle is the agent and the action goes forward to the agent) (Chatterjee, Southwood, \& Basilico, 1999).

A question arising from this manipulation is whether agency will be affected by the within-sentence structure, that is, the order that thematic roles are presented within the sentence. In Greek, which has a left-to right reading and writing system, grammatical information is conveyed through inflection, thus multiple word orders are allowed (SVO, OSV, VOS, VSO, OVS, OSV). The agent or the patient of an action can appear in any order within the sentence independently of whether the sentence is active or passive. For example, in active voice, structures such as "The cook \({ }_{\text {nom }}\) kicks the pirate acc" (the cook=agent) and "The pirate \({ }_{\text {acc }}\) kicks the \(c o o k_{\text {nom }}\) " are both grammatical. Therefore, highly inflectional languages, such as Greek, allow us to manipulate word order without necessarily using the passive voice which is of less frequency. In our study, we assume that spatial biases will be influenced by the within-structure sentence. We predict that participants will be faster in responding to sentences that not only match in terms of agency, but also in terms of characters' position in the sentence independently of thematic role assignment, that is, even in conditions that characters' share the same location spatially and temporally, independently of meaning.

\section*{Methods}

A sentence-picture verification task was used. Pairs of characters were presented in a \(2 \times 2 \times 2\) experimental design. Each picture involved two characters (character \({ }_{1}\) character \({ }_{2}\); e.g. "cook", "pirate") and each sentence contained two nouns corresponding to those characters ( noun \(_{1}-\) noun \(_{2}\) ), one in nominative (agent) and one in accusative (patient). The experimental stimuli were manipulated on three dimensions: 1 . characters' position in the picture (left - right), by flipping the image, 2. characters' position in the sentence (e.g. SVO - OVS) and 3. characters' thematic role (agent - patient) in the sentence by interchanging nominative and accusative case. Therefore, thematic role assignment was actually the variable that produced either matched- or mismatched-in-meaning sentence-picture pairs. Eight experimental conditions were
created (see Table 1). Every pair of characters was presented in each of eight conditions.

Table 1: Experimental design for sentence-verification task
\begin{tabular}{|c|c|c|c|c|}
\hline \multirow[t]{4}{*}{} & \multicolumn{4}{|c|}{Meaning} \\
\hline & \multicolumn{2}{|c|}{Match} & \multicolumn{2}{|c|}{Mismatch} \\
\hline & \multicolumn{4}{|c|}{Position in sentence} \\
\hline & Match & Mismatch & Match & Mismatch \\
\hline  & "The cook \(_{\text {nom }}\) kicks the pirate \(_{\text {acc }}\) " & "The pirate \({ }_{\text {acc }}\) kicks the \(\mathrm{cook}_{\text {nom }}\) " & "The cook \(_{\text {acc }}\) kicks the pirate \(_{\text {nom }}\) " & "The pirate \(_{\text {nom }}\) kicks the cook acc" \\
\hline  & "The pirate \({ }_{\text {acc }}\) kicks the cook \(_{\text {nom }}\) " & "The \(\mathrm{Cook}_{\text {nom }}\) kicks the pirate \(_{\text {acc }}\) " & "The pirate \({ }_{\text {nom }}\) kicks the cook \(_{\text {acc }}\) " & "The cook \(_{\text {acc }}\) kicks the pirate \({ }_{\text {nom }}\) " \\
\hline
\end{tabular}

In half of the experimental conditions thematic role assignment reflected the depicted action resulting in matched-in-meaning pairs, whereas in the other half, thematic roles mismatched the depicted action (mismatched-in-meaning pairs). Furthermore, in half of the matched-inmeaning pairs, nouns' location in sentence matched characters' position in picture, whereas in the other half, nouns' location in sentence mismatched characters' position in picture. The same was true for the mismatched-inmeaning conditions. Our dependent variable was response times and our independent variables were character's position in the picture (left, right), and word order (SVO OVS).

\section*{Participants}

Thirty-three adults, 18-30 year olds, participated in the present experiment. They were all Greek native speakers. They did not receive any compensation for their participation.

\section*{Materials}

Pictures The test stimuli consisted of 9 colored pictures. Each picture depicted two characters taking part in an action, and additional objects and people. The action presented in the picture was always the same, that is, character \({ }_{1}\) was always doing the action and character \({ }_{2}\) was receiving the action. However, there were two conditions. Half of the pictures depicted the agent on the left and half of them depicted the agent on the right. For each target picture, the agent of the action was either depicted on the left or on the right side of the screen by flipping the image. The filler items consisted of 60 colored pictures depicting objects and people. The order of experimental and filler items was pseudo-randomized, with the constraint that each experimental item was separated by a minimum of one filler item. All pictures were part of a larger set of pictures used in

Gennari, Mirkovic, and MacDonald (2012) and were appropriately adjusted for the purposes of this experiment.

Sentences For the test sentences, 9 verbs representing actions were used to construct quadruplets of active sentences, resulting in 36 Greek sentences as test items, ranging in length from 5 to 6 (mean \(=5.5\) ) words. These sentences had two possible word orders (SVO or OVS) and half of them were matching in meaning with the experimental pictures, whereas the other half were not. Examples of sentences are presented in Table 1. Additionally, 120 Greek sentences were used as filler items ranging in length from 3 to 10 words (mean \(=4.9\) ). Half of them matched in meaning the filler pictures and half of them did not. All sentences were recorded by a female native Greek speaker whose instructions were to read each sentence aloud in a natural, clear manner, in normal intonation.

\section*{Procedure}

Each participant, after giving verbal consent to participate in the study was seated in a quiet room and given instructions about the experiment. DMDX (Forster \& Forster, 2003) was used for the presentation of the stimuli. Visual stimuli were presented at the center of a laptop screen. Auditory stimuli were delivered over high quality headphones. In each trial, participants saw a picture and simultaneously heard a sentence corresponding or not to the picture. Participants had to perform a 2AFC task to indicate whether the sentence they heard matched the picture by pressing one of two buttons. Participants were given three practice trials at the beginning of the experiment in order to make sure they had understood the task. There was no feedback. The experiment lasted approximately 30 minutes.

\section*{Results}

Data were analyzed with generalized linear mixed-effects modeling, with random effects for participants and items, employing function lmer of package lme4 (Bates, Maechler, \& Bolker, 2012) in R (R Development Core Team, 2012). Response times were log-transformed. Only accurate responses were included in the analysis ( \(2 \%\) excluded) and only those participants that had lower than \(8 \%\) error rate. None of the participants was excluded from the analysis. Outliers were removed, that is, items with response time values below 500 msec or above two standard deviations from the mean. This resulted in excluding \(4 \%\) of the total data.
We conducted separate analyses for matched- and mismatched-in-meaning stimuli. For the matched-inmeaning condition, a main effect of agents' position was found (t value \(=-5.19, p=0.0001\) ) (i.e. participants' responses were faster when the agent was depicted on the right) and a main effect of word order ( t value \(=-9.29, \mathrm{p}=\) 0.0001). However, agents' position interacted with word order ( t value \(=4.13, \mathrm{p}=0.0004\) ). Contrasts among SVO and OVS conditions revealed that participants' reaction
times did not differ in the SVO condition ( \(\mathrm{t}=0.65\); \(\mathrm{p}=\) 0.5306 ), that is, agent's position did not affect reaction times. However, in the OVS condition, participants were significantly faster in responding to pictures presenting the agent on the right compared to left ( \(\mathrm{t}=-5.56\); \(\mathrm{p}=0.0001\) ), i.e. when within-sentence structure matched in characters' position in the sentence (Figure 1).


Figure 1: Agents' position and word order interaction in reaction times in matched-in-meaning pairs. In all figures, untransformed RTs are presented for ease of interpretation.

Same effects were found for the mismatched-in-meaning condition. A main effect of agents' position, (-6.06, p = 0.0001 ), word order ( t value \(=-10.12, \mathrm{p}=0.0001\) ) and an interaction between agents' position and word order was found ( t value \(=4.13, \mathrm{p}=0.0001\) ). Contrasts among SVO and OVS conditions revealed that participants were slower in responding to pictures that depicted the agent on the right compared to left in the SVO condition (6.07; \(\mathrm{p}=0.0001\) ). However, in the OVS condition, people were faster when the agent was presented on the right compared to left ( \(\mathrm{t}=-\) \(6.24 ; \mathrm{p}=0.0001\) ). That is, participants were faster in rejecting a mismatched-in-meaning pair when characters’ position in the sentence and characters' position in the picture matched (Figure 2).


Figure 2: Agents' position and word order interaction in reaction times in mismatched-in-meaning pairs.

\section*{Discussion}

The interaction between spatial and linguistic representations was investigated in a sentence-picture
verification task. Stimuli were manipulated in order to explore the role of word order in spatial representation of agency in a language with flexibility in word order. Our experimental manipulation allowed us to disentangle sentence structure and thematic role, in that agents could appear in one of two positions in the sentence. We found that when word order matched agents' position latencies dropped. This effect was not only observed in matched-inmeaning pairs, i.e. pairs in which both structure and thematic role represented the depicted action, but also, in mismatched-in-meaning pairs, i.e. pairs that only matched in sentence structure independently of thematic role.

However, a robust finding in literature, that processing of spatial representations is influenced by the directionality of the reading and writing system, was not obtained in this study. Since Greek is a left-to-right language we expected that participants would be faster in responding to pictures representing the agent on the left. In our study, participants were faster responding to pictures presenting the agent on the left only when the paired sentence was presented in SVO structure. In the other conditions, left agency did not facilitate participants' responses. We suggest that this seemingly contradictory result could be explained by the interaction between agents' spatial position and word order.

Specifically, we suggest that when the agent was presented on the left it was consistent with the left-to-right bias. This led to the formation of a strong expectation about the upcoming sentence structure (i.e. SVO). When this expectation was violated (i.e. OVS), reaction times became longer. In contrast, when the expectation was fulfilled, processing was significantly faster and reaction times dropped. However, when the agent was presented on the right side, no strong expectations were formed because the two effects (agent's position and left-to-right bias) partly canceled each other out. Therefore, the differences in latency between SVO and OVS structures should be much smaller in this case. Moreover, since SVOs are more frequent (and therefore easier to process), the mismatching between the agent's position and the sentence structure, should affect to a greater extent the processing of the less frequent OVS structures. Indeed, our results are in accordance with this prediction. In sum, the seeming absence of a left-to-right effect may be due to the violation of a strong left expectation.

Our explanation seems compatible with recent theories of prediction in cognition (Clark, 2012) and in sentence processing (Kamide, Altmann, \& Haywood, 2003; Dikker \& Indefrey, 2007; Altman \& Mirkovic, 2009; Farmer, Brown, \& Tanenhaus, in press). Specifically, a way of explaining the rapid nature of language comprehension stems from the idea of prediction. Comprehenders exploit all available information, integrate contextual constraints rapidly and generate predictions about upcoming stimuli. In our study, stimuli pairs were presented simultaneously. However, auditory stimuli are inherently more dynamic than visual stimuli. Sentences take longer to be presented and thus are processed later than a static picture. Therefore, we assume
that participants had the opportunity to process the picture longer and faster than the sentence, arguably allowing them to formulate predictions about the sentence structure. To test for this hypothesis, a future experimental manipulation could involve pictures and sentences presented not simultaneously but in different time points so that expectations about upcoming stimuli could be enhanced. For example, a sentence presented first in SVO structure may formulate the expectation of left agency, whereas an OVS structure may formulate the reverse expectation. If this turns out to be correct, then language may impose strong constraints and guide the way we conceptualize spatially thematic roles.

To conclude, we found that sentence processing not only reflects generic language characteristics, such the directionality of the writing system, but is also sensitive to frequency-driven effects, such as the occurrence rate of specific syntactic structures (i.e. SVO versus OVS). In addition, online language processing seems to be affected by non-linguistic information, which is in line with other findings (Tanenhaus, Spivey-Knowlton, Eberhard, \& Sedivy, 1995). Crucially, our findings are consistent with the left-bias account according to which language-specific factors may constrain and affect our conceptual representations. In sum, our findings suggest that different sources of information (both linguistic and non-linguistic) are interactively used in forming expectations about upcoming material.

\section*{Acknowledgments}

We would like to thank Silvia Gennari (University of York) for providing the visual stimuli. We also thank Efthymia Kapnoula for useful comments on earlier versions of the manuscript.

\section*{References}

Altmann, G.T.M. \& Mirkovic, J. (2009). Incrementality and prediction in human sentence processing. Cognitive Science, 33(4), 583-609.
Barrett, A.M., \& Craver-Lemley, C.E. (2008). Is it what you see, or how you say it? Spatial bias in young and aged subjects. Journal of the International Neuropsychological Society, 14(4), 562-570.doi:10.1017/s1355617708080764
Bates, D., Maechler, M., \& Bolker, B. (2012). lme4: Linear mixed-effects models using S4 classes. Available from http://CRAN.R-project.org/package=lme4 (R package version 0.999999-0).
Caplan, D., \& Futter, C. 1986. Assignment of thematic roles to nouns in sentence comprehension by an agrammatic aphasic patient. Brain and Language, 27(1), 117-134.
Caramazza, A., \& Miceli, G. (1991). Selective impairment of thematic role assignment in sentence processing. Brain and Language, 41(3), 402-436.
Chatterjee, A. (2001). Language and space: some interactions. Trends in Cognitive Science, 5, 55-61.

Chatterjee, A., (2008). The neural organization of spatial thought and language. Seminars in Speech and Language, 29, 226-38 .
Chatterjee, A., Maher, L.M., Gonzales-Rothi, L.J., \& Heilman, K.M. (1995). Asyntactic thematic role assignment: The use of a temporal-spatial strategy. Brain and Language, 49, 125-139.
Chatterjee, A., Southwood, M.H., \& Basilico, D. (1999). Verbs, events and spatial representations. Neuropsychologia, 37, 395-402.
Chan, T.T. \& Bergen, B. (2005). Writing Direction Influences Spatial Cognition. In Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society, 412-417.
Clark, A. (2012). Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences (in press).
Dikker, S. \& Indefrey, P. (2007). Fine-graining structure prediction in sentence processing: How language mixing reveals an economic parser. 20th Annual CUNY Conference on Human Sentence Processing, La Jolla, CA.
Dobel, C., Diesendruck, G., \& Bolte, J. (2007). How writing system and age influence spatial representations of actions: A developmental, cross-linguistic study. Psychological Science, 18(6), 487-491.
Farmer, T.A., Brown, M., \& Tanenhaus, M.K. (2012). Prediction, explanation, and the role of generative models in language processing. Behavioral and Brain Sciences (Commentary Article) (in press).
Forster, K., \& Forster, J. (2003). DMDX: A Windows display program with millisecond accuracy. Behavior Research Methods, 35(1), 116-124.
Geminiani, G., Bisiach, E., Berti, A., \& Rusconi, M.L. (1995). Analogical representation and language structure. Neuropsychologia, 33, 1565-1574.
Gennari, S.P., Mirkovic, J., MacDonald, M.C. (2012) Animacy and competition in relative clause production: a cross-linguistic investigation, Cognitive Psychology, 65, 141-176.
Hayward, W.G., \& Tarr, M.J. (1995). Spatial language and spatial representation. Cognition, 55, 39-84 .
Jackendoff, R. (1996). The architecture of the linguisticspatial interface. In P. Bloom, M.A. Peterson, L. Nadel, \& M. F. Garrett (Eds.), Language and space, Cambridge, MA: MIT Press.
Kamide, Y, Altmann, G.T.M, Haywood S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. Journal of Memory and Language, 49, 133-159.
Levinson, S.C. (2003). Space in language and cognition: Explorations in cognitive diversity. Cambridge: Cambridge University Press.
Maas, A. \& Russo, A. (2003). Directional bias in the mental representation of spatial events: nature or culture? Psychological Science, 14(4), 296-301.

Papafragou, A., Hulbert, J., \& Trueswell, J. (2008). Does language guide event perception? Evidence from eye movements. Cognition, 108, 155-184.
R Core Team. (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available from http://www.R-project.org/ (ISBN 3-900051-07-0).
Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M., \& Sedivy, J.E. (1995). Integration of visual and linguistic information in spoken language comprehension. Science, 268, 632-634.

\title{
Whether Chinese Speakers Think about Time More Vertically Depends on their Immediate and Lifetime Experience of Reading Horizontal or Vertical Texts: Evidence from Contextual Priming
}

\author{
Jenn-Yeu Chen (PSYJYC@NTNU.EDU.TW) Michael Friederich (MICHAEL_TAIWAN_AG@YAHOO.COM) \\ Department of Chinese as a Second Language, National Taiwan Normal University 162 Heping East Road Section 1, Taipei, Taiwan 106
}

\author{
Hua Shu (SHUH@BNU.EDU.CN) \\ National Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University 19 Xinjiehouwai Street, Beijing, China 100875
}

\begin{abstract}
Do Chinese speakers think about time vertically because they use vertical spatial metaphors to express time? Inconsistent findings have been reported even when the same paradigms were used. The present study examined participants' performance on a temporal judgment task while holding language constant but varying their lifetime and immediate reading experience of horizontal and vertical texts. Chinese participants from Taiwan and China were randomly assigned to a reading task involving horizontally or vertically arranged texts (contextual primes). A temporal judgment task (spatial-temporal association of response codes or STARC) followed the reading task, asking the participants to judge if the event depicted in a second picture occurred earlier or later than that in a first picture. Responses were faster when the left keys represented the 'earlier' responses than when the right keys did, representing a STARC effect. Half of the participants responded with horizontally oriented keys while the rest with vertically oriented keys. For the Taiwan participants, the overall STARC effect was greater when the response keys were vertical than horizontal, but no difference was observed for the China participants. A questionnaire indicates that the two groups of participants had similar lifetime experiences of reading horizontal texts, but the Taiwan participants read vertical texts in their life far more frequently than the China participants. Immediate reading experiences interacted with lifetime experiences in modulating the vertical bias. For the Taiwan participants, the vertical bias was strong following the vertical prime, but disappeared following the horizontal prime. For the China participants, the horizontal prime led to no vertical bias whereas the vertical prime brought about a horizontal bias. We conclude that the directionality of orthography and speakers' immediate and lifetime reading experiences, rather than the use of vertical spatial metaphors, can better explain the vertical bias (or the lack of it) in the Chinese speakers.
\end{abstract}

Keywords: linguistic relativity; temporal reasoning; reading direction

\section*{Introduction}

Once denounced as scientifically unsound (Devitt and Sterelny, 1987; Pinker, 1994), the linguistic relativity hypothesis has regained much attention in the past two decades. The essence of the hypothesis is that the
particular linguistic form in a language can shape the habitual way of thinking by the speakers of the language (Whorf, 1956; Hunt \& Agnoli, 1991). For example, if language A does not distinguish two shades of blue whereas language B does, speakers of language A would not be able to tell apart the two shades of blue as easily as speakers of language B (Davidoff, Davies, \& Roberson, 1999; Gilbert, Regier, Kay, \& Ivry, 2006; Winawer et al., 2007). Similarly, if language \(A\) does not encode the biological gender lexically whereas language \(B\) does, the gender information would become less accessible to speakers of language \(A\) than speakers of language \(B\) (Chen \& Su, 2011). In the temporal domain, it has been observed that Chinese speakers seem to conceptualize time continuously and maintain an "extended present" view that encompasses recent past and near future, whereas English speakers tend to maintain a relatively discrete view of time with distinct present, past and future. This cross-linguistic difference has been attributed to the use of explicit tense and aspect markers in English and the lack of them in Chinese (Chen, Su, Lee, \& O'Seaghdha, 2012; Chen, Su, \& O'Seaghdha, 2013).

While much of recent empirical work has produced evidence consistent with the linguistic relativity hypothesis, there were controversies due to inconsistent findings as well. One particular controversy comes from the study of spatial metaphors of time. An early study employing a spatial priming paradigm found that the frequent use of vertical spatial metaphors to express time in Chinese led to a vertical bias in the Chinese speakers' conception of time whereas the rare use of such metaphors in English led to a horizontal bias in the English speakers (Boroditsky, 2001). However, subsequent studies were unable to confirm such a differential bias (Chen, 2007; January \& Kako, 2007; Tse \& Altarriba, 2008; Sanvido, de Rose, \& Chen, 2011).

More recently, a SNARC-like paradigm (spatial-numerical association of response codes, Dehaene, Bossini, \& Giraux, 1993) applied to temporal processing (spatial-temporal association of response codes or STARC) detected a similar vertical bias in the Chinese speakers relative to the English speakers (Boroditsky, Fuhrman, \& McCormick, 2011; Fuhrman et al., 2011; Miles, Tan, Noble, Lumsden, \& Macrae, 2011). In a STARC task, the participants saw two photographs of an event and had to determine if the second photograph
occurred earlier or later than the first one. In the canonical condition, they pressed a left key to indicate 'earlier' and a right key to indicate 'later' while in the non-canonical condition, the key assignment was reversed. Response times were typically slower in the non-canonical condition relative to the canonical condition, representing a STARC effect. For half of the participants, the keys were placed horizontally while for the other half, the keys were oriented vertically. It was found that Chinese and English speakers displayed similar horizontal STARC effects, but more importantly, the Chinese speakers demonstrated a greater vertical STARC effect than the English speakers. Unfortunately, inconsistent findings were observed with this paradigm as well. Chen and O'Seaghdha (2012 accepted) observed a vertical bias in the Chinese speakers from Taiwan, but no such bias in the Chinese speakers from China. Because horizontal printing of texts is a national policy in China, but not in Taiwan, where vertical texts are fairly common, it was suggested that reading experience of horizontal and vertical texts might have something to do with the participants' performance on the STARC task. The suggestion, however, was inferred from quasi-experimental evidence.
The present study was designed to test the effect of reading experience on Chinese speakers' performance on the STARC task by experimentally manipulating the layout of texts (horizontal or vertical) which participants read before the STARC task. The reading task, serving as a contextual prime, was expected to bias the Chinese participants towards displaying a greater or smaller horizontal or vertical STARC effect depending on the direction of reading. The modulation of immediate reading experience might interact with Chinese speakers’ lifetime reading experience, which was assessed by including participants from Taiwan and China. The participants from Taiwan would have more extensive experience of reading vertical texts than the participants from China.

\section*{Method}

\section*{Participants}

Fifty-six native Mandarin Chinese speakers from Taiwan and the same number from China participated in this study. The participants from Taiwan were graduate or undergraduate students from National Taiwan Normal University and nearby universities in Taipei, while those from China were similar students from Beijing Normal University and nearby universities in Beijing. The age range for the participants was from 18 to 26 . All the participants had normal or corrected-to-normal vision and they were paid 200 TWDs or 20 RMBs for participation.

\section*{Design and Materials}

The Reading Task Seven short essays with 11 comprehension questions were chosen from the Taiwan University Entrance Exams for the reading task. Two versions of the texts (the essays and the questions) were prepared as paper booklets, one arranged horizontally and the other vertically. Participants were randomly assigned to one version. There was no time pressure for taking the
task. On the average, it took approximately 15 to 20 minutes for the participants to complete this task. Upon completion, the participants proceeded immediately to the STARC task.
The STARC Task The design and procedure of the STARC task followed those of Chen \& O'Seaghdha (2013). The materials were 37 action events, each being photographed at three different phases of time (e.g., Time X : a man holding the handle and about to turn the key to open a door, Time Y: door being open with the man stepping half into the room, and Time Z: door being half closed with the man inside the room facing inward with his left hand holding against the closing door). On each trial, a Time Y picture was randomly chosen from the 37 events and shown to the participants. The Time Y picture was followed by a Time X or a Time Z picture. The participants were asked to determine whether the action depicted in the second picture occurred earlier or later than the action depicted in the first. In one condition (the canonical response), the number-4 key on the numeric keypad of a standard keyboard, marked with a blue sticker, was designated as the 'earlier' response, and the number-5 key, marked with a orange sticker, was designated as the 'later' response. In the other condition (the noncanonical response), the key assignment was reversed, i.e., the ' 5 ' key was the blue one designated as the 'earlier' response and the ' 4 ' key was the orange one designated as the 'later' response. Canonicity was a within-subjects factor. The same set of 37 action events was used in the two canonicity conditions, with the Time X and the Time Z pictures appearing exactly once in each condition. The order of the two conditions was counterbalanced across the participants. A between-subjects factor was also included. Half of the participants, randomly determined, responded with the keyboard placed on the desk in a normal horizontal orientation, and the other half responded with the keyboard oriented vertically (propped up against a bookend).

The task was programmed in E-Prime and was run on a desktop (ASUS B53S with an Intel® Core \({ }^{\mathrm{TM}}\) i5 2520 M processor and a 15.6" 16:9 HD 1366x768-resolution LED screen) and a laptop computer (ASUS R500V with an Intel \({ }^{\circledR}\) Core \({ }^{\mathrm{TM}}\) i7 3610QM 2.3 GHz processor and a \(15.6^{\prime \prime}\) 16:9 HD 1366x768-resolution LED screen), both with a separate USB- connected numeric keypad (Kingyo). A trial began with a fixation cross which appeared at the center of the screen for 500 msec . and was followed by a blank screen for 500 ms . Then, the first picture in a pair appeared at the same location for 2000 ms followed by another blank screen for 500 ms . The second picture followed and stayed on until the participants responded. Upon a response, a last blank screen of 500 ms replaced the second picture and the next trial began. Both pictures measured 22.5 cm in width and 17 cm in height. The participants sat at a viewing distance of 70 cm in front of the computer screen. The participants were told to respond with the index finger of their preferred hand as quickly and accurately as soon as the second picture appeared. The index finger was parked at the gulf between the blue and orange keys at the beginning of a
trial. The participants received five practice trials before going on with the experimental trials.

The Reading Experience Questionnaire Upon completion of the reading task and the STARC task, the participants also filled out a questionnaire to indicate how frequently they encountered a vertical text, a horizontal text printed from left to right, and a horizontal text printed from right to left on a 8-point scale ranging from never (0) to very frequently (7). They also reported the sources of the texts (e.g., magazines, newspapers, textbooks, street signs, slogans, advertisements, etc.).

\section*{Results}

\section*{The Taiwan Sample}

For the Taiwan sample, the participants’ comprehension scores in the reading task were close to perfect. Their rated experience of vertical texts, horizontal left-to-right texts and horizontal right-to-left texts was 5.7 ( \(\mathrm{SD}=1.6\) ), 6.5 ( \(\mathrm{SD}=0.5\) ), and 1.4 ( \(\mathrm{SD}=1.6\) ), respectively. Their error rate in the STARC task was on the average \(3 \%\). The analysis of their log-transformed response times in the STARC task shows the pattern in Figure 1. The STARC effect on the Y-axis represents the averaged difference in log-transformed response time of the noncanonical condition minus the canonical condition. The overall STARC effect was significant by the linear mixed-effect analysis: \(F(1,7915)=59.18, p<.0001\). As the figure shows, the STARC effect was greater when the response keys were oriented vertically than when they were oriented horizontally: \(F(1,7915)=11.13, p=.0009\). This indicates an overall vertical bias in temporal judgment by our Taiwan participants. The figure also shows that whereas the vertical bias was fairly strong following the vertical prime (i.e., having read the vertical texts and questions), it was substantially reduced (in fact disappeared) following the horizontal prime. Statistically, the response orientation by canonicity interaction was highly significant under the vertical prime, \(F(1,3952)=\) 18.51, \(p<.0001\), but the same interaction was far from being significant under the horizontal prime, \(F(1,3927)\) \(=.35, p=.5564\).

\section*{The Beijing Sample}

For the Beijing sample, the participants' comprehension scores in the reading task were also close to perfect. Their rated experience of vertical texts, horizontal left-to-right texts and horizontal right-to-left texts was 2.7 ( \(\mathrm{SD}=1.6\) ), 6.9 ( \(\mathrm{SD}=0.7\) ), and 1.3 ( \(\mathrm{SD}=1.3\) ), respectively. Their averaged error rate in the STARC task was \(3 \%\). The analysis of their log-transformed response times in the STARC task shows the pattern in Figure 2. The overall STARC effect was significant by the linear mixed-effect analysis: \(F(1,7857)=50.07, p<.0001\). As the figure shows, the STARC effect interacted significantly with prime and response orientation: \(F(1,7857)=8.01, p\) \(=\).0047. Separate post-hoc analyses show that the response orientation by canonicity interaction was significant under the vertical prime, \(F(1,3895)=4.83, p\) \(=.0280\), showing a greater horizontal STARC effect than
the vertical one; under the horizontal prime, the vertical STARC effect was greater than the horizontal one, but the interaction fell short of the conventional level of significance, \(F(1,3926)=3.22, p=.0728\). Worth noting is no significant difference between the horizontal and vertical STARC effects ( \(p=.57\) ), indicating no overall vertical bias in the Beijing participants. None of the other effects were significant, \(p\) 's > .24.


Figure 1: The STARC effect (difference in log RT of the noncanonical condition minus the canonical condition) as a function of response key orientation and type of contextual prime (the Taipei sample).


Figure 2: The STARC effect (difference in log RT of the noncanonical condition minus the canonical condition) as a function of response key orientation and type of contextual prime (the Beijing sample).

\section*{Discussion}

The Chinese language employs both horizontal and vertical spatial metaphors for expressing time. It has been suggested that the common use of vertical spatial metaphors biases the Chinese speakers to conceptualize time vertically. We hypothesized that reading experience of horizontally and vertically arranged texts might be a potent variable contributing to such a bias. The hypothesis was tested by assigning Chinese participants from Taiwan and China to a reading task involving either horizontally or vertically arranged texts, followed by a STARC task. Although the vertical STARC effect was overall greater than the horizontal one among the Taiwan participants, such a vertical bias was absent among the China participants. These results can be accounted for by the significantly more frequent lifetime experience of encountering vertical texts among the Taiwan participants
than among the China participants: 5.7 vs. \(2.7, \mathrm{t}(110)=\) \(9.9, \mathrm{p}<.0001\). The rated experience of encountering horizontal left-to-right texts was similar between the two groups of participants: 6.5 vs 6.9.
Furthermore, the vertical bias, when present, was modulated by the immediate reading experience such that it disappeared when the Taiwan participants had just read horizontally arranged texts. For the China participants, the immediate reading experience also modulated the vertical bias, but in the opposite direction. The horizontal prime led to no significant vertical bias while the vertical prime brought about a horizontal bias. The different patterns of results between the Taiwan and the Beijing participants indicate that lifetime reading experience interacts with immediate reading experience in its effect on the participants' temporal judgment in the STARC task.
The finding of a causal role of directionality of orthography and reading experience in Chinese speakers’ temporal judgment is consistent with the findings of many studies in the literature showing a relationship between the directionality of orthography and the performance on a space-implicated task (Tversky, Kugelmass, and Winter, 1991; Dehaeneet al., 1993; Zebian, 2005; Chan \& Bergen, 2005; Fuhrman \& Boroditsky, 2010; Ouellet, Santiago, Israeli, \& Gabay, 2010). In conjunction with these findings as well as the facts that (1) horizontal spatial metaphors are used far more frequently than vertical ones in Mandarin Chinese (Chen, 2007) and (2) our Taiwan and China participants speak the same language, the results of the present study suggest that the directionality of orthography and speakers’ (immediate and lifetime) reading experience, rather than the use of spatial metaphors per se, can better explain the vertical bias (or the lack of it) in the Chinese speakers.

Chen and O'Seadhgha (2012 accepted) previously observed a vertical bias in the Chinese participants from Taipei (Taiwan), but a horizontal bias in the Chinese participants from Guangdong (China). The discrepancy was attributed to the fact that China has adopted the national policy of printing all texts horizontally whereas both horizontal and vertical directions are allowed and prevalent in Taiwan. The evidence, however, is indirect due to the quasi-experimental nature of the study. By directly manipulating the participants’ experience of reading horizontal and vertical texts, the present study offers the needed evidence for establishing the causal role of directionality of orthography and reading experience.

Reading experience can also account for the inconsistent findings across studies. Because the Chinese participants in the previous studies came from different regions, their experience of reading horizontal and vertical texts could vary greatly, which was attested to by the rating data in the present study, and thus could contribute to the inconsistency in findings.

Returning to the use of spatial metaphors for expressing time, Chen (2007) has previously reported that horizontal spatial metaphors were actually used far more frequently than vertical spatial metaphors in Chinese. He argued that the usage pattern did not lend the logical support for the hypothesis that Chinese speakers would think about time
more vertically. Boroditsky et al. (2011) countered Chen's argument by maintaining that it was the cross-language difference in the usage of vertical spatial metaphors that predicted the vertical bias in the Chinese speakers. However, without controlling for potent factors such as directionality of orthography and speakers’ reading experience, it is impossible to make certain that Chinese speakers do think about time differently than English speakers and that this is due to the differential usage of vertical spatial metaphors in the two languages.

The hypothesized conceptual link between spatial and temporal reasoning has also been questioned recently with respect to the use of frame of reference. It has been claimed that people reference time onto space, and because different linguistic communities prefer different spatial frames of reference, their temporal references vary as well. Beller, Rothe, Hüther and Bender (2012) examined existing data as well new data, concluding that there is not a close link between referencing preferences across spatial and temporal domains.

Although linguistic relativity has manifested itself in several domains of cognition, whether it extends to the conception of time in relation to the latter's metaphorically projected meaning requires further investigations at best.

\section*{Acknowledgments}

This work was supported by "Aim for the Top University Plan" of the National Taiwan Normal University and the Ministry of Education, Taiwan, R.O.C. Preparation of this manuscript was supported by the NSC 101-2410-H-003 -105-MY2 grant. We thank Yahsin Huang for conducting the experiments and Lijuan Zou and Tian Hong for helping with recruiting participants in Beijing.

\section*{References}

Beller, S., Rothe, A., Hüther, L., \& Bender, A. (2012). Moving forward in space and time: How strong is the conceptual link between spatial and temporal frames of reference? Frontiers in Cultural Psychology, 3:486. doi: 10.3389/fpsyg.2012.00486.

Boroditsky, L. (2001). Does language shape thought? English and Mandarin speakers’ conceptions of time. Cognitive Psychology, 43, 1-22.
Boroditsky, L., Fuhrman, O., \& McCormick, K. (2011). Do English and Mandarin speakers think about time differently? Cognition, 118, 123-129.
Chan, T.-T., \& Bergen, B. (2005). Writing direction influences spatial cognition. In Bara, B., Barsalou, L., \& Bucciarelli, M. (Eds.), Proceedings of the 27th Annual Conference of the Cognitive Science Society (pp. 412-417). Mahwah, NJ: Lawrence Erlbaum Associates.
Chen, J.-Y. (2007). Do Chinese and English speakers think about time differently? Failure of replicating Boroditsky (2001). Cognition, 104, 427-436.

Chen, J.-Y., \& O'Seaghdha, P. G. (2012 accepted). Do Mandarin and English speakers think about time differently? Review of existing evidence and some new data. Journal of Chinese Linguistics.

Chen, J.-Y., \& Su, J.-J. (2011). Differential sensitivity to the gender of a person by English and Chinese speakers. Journal of Psycholinguistic Research, 40, 195-203
Chen, J.-Y., Su, J.-J., Lee, C.-Y., \& O’Seaghdha, P. G. (2012). Linguistically directed attention to the temporal aspect of action events in monolingual English speakers and Chinese-English bilingual speakers with varying English proficiency. Bilingualism: Language and Cognition, 15, 413-421.
Chen, J.-Y., Su, J.-J., \& O'Seaghdha, P. G. (2013). Enduring moments: The extended present in Chinese speakers' orientation to event time. Journal of Pragmatics, 45, 90-103.
Davidoff, J., Davies, I., \& Robertson, D. (1999). Colour categories in a stone-age tribe. Nature, 402(6762), 604-605.
Dehaene, S., Bossini, S., \& Giraux, P. (1993). The mental representation of parity and number magnitude. Journal of Experimental Psychology: General, 122, 371-396.
Devitt, M., \& Sterelny, K. (1987). Language and reality: An introduction to the philosophy of language. Oxford: Blackwell.
Fuhrman, O., \& Boroditsky, L. (2010). Cross-cultural differences in mental representations of time: Evidence from an implicit non-linguistic task. Cognitive Science, 34, 1430-1451.
Fuhrman, O., McCormick, K., Chen, E., Jiang, H., Shu, D., Mao, S., \& Boroditsky, L. (2011). How linguistic and cultural forces shape conceptions of time: English and Mandarin time in 3D. Cognitive Science, 35, 1305-1328.
Gilbert, A. L., Regier, T., Kay, P., \& Ivry, R. B. (2006). Whorf hypothesis is supported in the right visual field but not the left. Proceedings of the National Academy of Sciences of the United States of America, 103(2), 489-494.
Hunt, E., \& Agnoli, F. (1991). The Whorfian hypothesis: a cognitive psychology perspective. Psychological Review, 98(3), 377-389.
January, D., \& Kako, E. (2007). Re-evaluating evidence for the linguistic relativity hypothesis: Response to Boroditsky (2001). Cognition, 104, 417-426.
Miles, L. K., Tan, L., Noble, G. D., Lumsden, J., \& Macrae, C. N. (2011). Can a mind have two time lines? Exploring space-time mapping in Mandarin and English speakers. Psychonomic Bulletin and Review, 18, 598-604.
Ouellet, M., Santiago, J., Israeli, Z., \& Gabay, S. (2010). Is the future the right time? Experimental Psychology, 57, 308-314.
Pinker, S. (1994). The Language Instinct. N.Y.: Harper Perrenial.
Sanvido, G. B., de Rose, J. C., \& Chen, J.-Y. (2011 November). Chinese speakers do not think about time differently than Portuguese speakers. Poster presented at the 52nd Annual Meeting of the Psychonomic Society, Seattle, USA.
Tse, C.-S., \& Altarriba, J. (2008). Evidence against linguistic relativity in Chinese and English: A case study of spatial and temporal metaphors. Journal of Cognition and Culture, 8(3/4), 335-357.

Tversky, B., Kugelmass, S. \& Winter, A. (1991). Crosscultural and developmental trends in graphic productions. Cognitive Psychology, 23, 515-557.
Whorf, B. L. (1956). Language, thought and reality: Selected writings of Benjamin Lee Whorf, ed. J. B. Carroll. Cambridge, MA: The MIT Press.
Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., \& Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. Proceedings of the National Academy of Sciences of the United States of America, 104(19), 7780-7785.
Zebian, S. (2005). Linkages between number concepts, spatial thinking, and directionality of writing: The SNARC effect and the REVERSE SNARC effect in English and Arabic monoliterates, biliterates, and illiterate Arabic speakers. Journal of Cognition and Culture, 5, 165-190.

\title{
Hemispheric Asymmetry in Nonconscious Processing
}

\author{
Jing Chen (jinghku@hku.hk) Janet Hui-wen Hsiao (jhsiao@hku.hk) \\ Department of Psychology, University of Hong Kong \\ 691 The Jockey Club Tower, Centennial Campus, Pokfulam Road, Hong Kong SAR
}

\begin{abstract}
Here we investigated whether hemispheric asymmetry effects can be observed in nonconscious processing with a basiclevel animal categorization (cat/dog) task. We found a significant nonconscious congruency priming effect when the prime was presented in the right visual field/left hemisphere but not when it was presented in the left visual field/right hemisphere when the prime duration was only 10 ms ; the leftlateralized congruency priming effect was consistent with the left hemisphere superiority in processing abstract category information reported in the literature (e.g., Marsolek, 1999). This result thus showed that nonconscious processing can go beyond the sensory level to influence hemispheric asymmetry in the processing of category information. In contrast, this hemispheric difference was not observed when the prime was presented for 50 ms (nonconscious) or 150 ms (conscious). This effect may be because 10 ms subliminal information was insufficient to allow inter-hemispheric transfer/processing, allowing the hemispheric difference to emerge. It also suggests that hemispheric asymmetry may be better observed at subliminal level.
\end{abstract}

Keywords: subliminal priming; hemispheric asymmetry; attention; nonconscious processing;

\section*{Introduction}

It has been well established in the literature that subliminal stimuli are able to elicit subsequent cognitive and behavioral influences (see Kouider \& Dehaene, 2007 for a review). It remains unclear, however, how a subliminally presented stimulus is processed. Here we aim to examine whether a subliminally presented stimulus can induce hemispheric asymmetry effects.

In the literature of hemispheric asymmetry in perception, hemispheric asymmetry effects typically did not emerge until a late perceptual stage (Fendrich \& Gazzaniga, 1990; Hsiao., Cipollini., \& Cottrell., accepted; Sergent, 1982). More specifically, hemispheric asymmetry effects were typically observed in tasks involving high-level perception such as face recognition (e.g., Keenan, Whitman, \& Pepe, 1989), but not in tasks more relevant to early perceptual processes such as grating detection (e.g., Fendrich \& Gazzaniga, 1990). Based on these findings, it has been argued that hemispheric asymmetry effects "must result from processing taking place beyond the sensory level" (Sergent, 1982). Thus, examining hemispheric asymmetry effects in subliminal priming can help us understand whether nonconscious processing can go beyond early
perceptual processes to induce hemispheric asymmetry effects.

Nevertheless, in the literature on subliminal priming, whether hemispheric asymmetry effects can be observed in the processing of subliminal stimuli was rarely studied. Marzouki, Grainger, and Theeuwes (2007) examined subliminal priming effects in a letter/pseudo-letter judgment task. In their experiment, the target stimulus always appeared at the center, while the preceding \(45-\mathrm{ms}\) prime (the same letter as the target) could be in the left visual field (LVF) or the right visual field (RVF). The results showed that in the trials with a letter target, a robust priming effect was found, however, only when the prime was presented in the RVF/left hemisphere (LH), but not when it was presented in the LVF/right hemisphere (RH). While Marzouki et al. (2009) interpreted the results as an attentional bias in favor of the RVF, their results may suggest a hemispheric asymmetry effect in subliminal letter processing as well. However, it is also possible that the RVF priming effect Marzouki et al. (2007) observed was due to perception-response compatibility, as participants always used their right hand to respond to the letter targets. Thus, it remains unclear whether hemispheric asymmetry effects can be observed in subliminal priming.

Using a supraliminal priming paradigm, Marsoleck (1999) found hemispheric asymmetry effects in processing category-related information. In Marsoleck (1999)'s study, participants were asked to name pictures of objects presented for 17 ms in either the LVF or the RVF. Before the target presentation, the same object or another object from the same basic category as the target object was presented centrally for 3 seconds as the prime. While a same-object priming effect was found for both LVF and RVF targets, the same-category different-object priming effect was only observed in the RVF/LH. This result suggested that the LH may be biased towards abstractcategory representations whereas the RH is biased towards exemplar-specific representations. This hemispheric asymmetry in the representation of category knowledge was further supported by several follow-up studies using various stimuli (Laeng, Zarrinpar, \& Kosslyn, 2003; Studer \& Hübner, 2008).

Thus, in the present study, we aimed to examine whether this hemispheric asymmetry in the representation of category knowledge can be induced with subliminal stimuli in a category judgment task (Marsolek, 1999), with perception-response compatibility controlled (cf. Marzouki
et al., 2007). The results of this examination can shed light on whether subliminal information can go beyond the sensory level to influence late perceptual processes.

In the representation of object category information, there are multiple levels of abstraction. A cat can be identified as "cat" (the basic level), or "animal" (the superordinate level), or "Sam" (a specific cat, the subordinate level). The current study focuses on the processing of basic-level category information (i.e., a cat/dog classification task). In a previous study (Finkbeiner \& Palermo, 2009), the authors investigated nonconscious congruency priming effects in an animal/tool classification task (experiment 1a; or an animal/vegetable classification task in experiment 2a), which was at the superordinate level. In their study, the prime was always at the top location and the target was always at the button location. And the top prime can be spatially cued or not. The results showed that a nonconscious congruency priming effect was found only in the condition when the prime was cued. This effect suggests that nonconscious processing of superordinate-level category information relies on the engagement of spatial attention.

Thus, the other aim of the current study was to examine whether nonconscious processing of basic-level category information also relied on spatial attention. Previous studies showed that, compared with superordinate-level category information, basic-level information was more accessible (e.g., Rosch, Mervis, Gray, Johnson, \& Boyes-Braem, 1976), basic-level terms were preferred in naming objects (e.g., Jolicoeur, Gluck, \& Kosslyn, 1984; Rosch et al., 1976) and recognizing objects at the basic level required less reaction time (e.g., Tanaka, 2001). Thus, it is possible that nonconscious processing of basic-level category information is more automatic than that of superordinate-level category information and does not require attention.

Therefore, in the current study, we aimed to 1) examine whether there is a hemispheric asymmetry in the subliminal priming effect by presenting the prime in either the LVF or the RVF, and 2) investigate whether nonconscious processing of basic-level category information relies on the engagement of spatial attention. In addition, we manipulated the duration of the prime ( \(150-\), 50 -, and \(10-\mathrm{ms}\) ) in order to compare hemispheric asymmetry and attentional modulation effects across conditions with different levels of consciousness.

\section*{Experiment 1}

In experiment 1, the prime ( \(150-\mathrm{ms}\) duration) was visible in order to: 1) examine whether the congruency priming effect can be found at the supraliminal level, and 2) investigate whether this congruency priming effect can be modulated by hemispheric asymmetry and spatial attention.

\section*{Method}

Participants 20 students ( 13 females, age mean \(=20.7\), range \(=18-24)\) at the University of Hong Kong participated
in exchange for course credits or payment. All were Asian, with normal or corrected-to-normal vision.

Stimuli Dog and cat images, all frontal-view or close to frontal-view, were selected from the Oxford-IIT Pet Dataset (Parkhi, Vedaldi, Zisserman, \& Jawahar, 2012), with cat images from the species of British Shorthair and dog images from that of Shiba Inu. Only the head parts of the animal images were used and were converted to grayscale, and were further cropped to fit within a black oval. They all had equal size, luminance, and contrast.

Two types of black and white noise images were used as masks. The random noise mask was generated by assigning a random value between 155 and 255 to each pixel, while the animal mask was created by, firstly, combining a cat and a dog image, and then scrambling the combined image. The animal mask was used as a backward mask for the prime as well as the target.

Each stimulus, including both animal images and masks, spanned \(4^{\circ}\) visual angle vertically and \(2.5^{\circ}\) horizontally with a \(60-\mathrm{cm}\) viewing distance.

Design and Procedure The experiment used a 2 (spatial attention: prime-cued vs. uncued) \(\times 2\) (congruence between the prime and target: congruent vs. incongruent) \(\times 2\) (visual field: LVF vs. RVF for the location of the prime) withinparticipant design. Participants were instructed to judge the category of the target pictures, i.e., a dog/cat judgment. They made the binary choice by pressing two buttons ("F" and "J") on a keyboard with their two index fingers simultaneously, or another two buttons ("D" and "K") with their middle fingers simultaneously. The response buttons were counterbalanced across participants.


Figure 1: Sequence of stimuli in a single trial
The basic trial structure was similar to Marzouki, Grainger and Theeuwes (2007)'s study (Figure 1). After a forward mask with a duration randomly selected among 400, 500, 600, and 700 ms , a yellow box cue was presented
either at the left or the right side for 150 ms , followed by the \(150-\mathrm{ms}\) prime image at either the same (Prime-cued) or the opposite (Prime-uncued) side of the cue. Then, the target image appeared for 200 ms at the center together with two animal backward masks at both the left and right sides. Following the target, another animal mask appeared at the center for 1500 ms (See Figure 1). Participants were instructed to make their judgment as quickly and as accurately as possible after they saw the target.

Before the main experiment, a practice section with 20 trials was performed. The procedure of the practice was the same as the main experiment. The images used in the practice never appeared in the main experiment. After the main experiment, in order to confirm that the prime images were visible, participants were instructed to do a primediscrimination task, i.e., to judge the category of the prime images. The procedure was identical to the main experiment except that participants were to make judgments on the prime instead of the target stimulus, and that the backward masks at the end of each trial stayed until participants made their response.

40 images ( 20 cat and 20 dog images) were used as the target as well as the prime stimuli in all four conditions (congruent with prime cued, congruent with prime uncued, incongruent with prime cued, incongruent with prime uncued), with the constraint that the prime and the target could not be the same image. In each condition, all 40 images appeared once as the target, resulting in 160 trials in total. Trials were presented in a random order, with the limits that the same target did not appear in succession, that no more than 3 trials in a row elicited the same response, and that no more than 3 trials in a row were in the same condition. In the prime-discrimination task, there were also 160 trials.

\section*{Results}

In analyzing the response times (RTs), null responses and incorrect responses (6.47\%) were excluded. And any RT outside 2.5 standard deviations from the mean, for each participant, was treated as outliers (2.78\%). In analyzing the accuracy data, only null responses ( \(0.75 \%\) ) were excluded.

The result of the prime-discrimination task confirmed that the \(150-\mathrm{ms}\) prime was visible. The mean accuracy reached \(77.7 \%\), significant higher than the \(50 \%\) chance level, t (19) \(=9.57, p<.001\).

A 2 (spatial attention) \(\times 2\) (congruency) \(\times 2\) (visual field) repeated measures ANOVA on the RTs revealed a significant main effect of congruency, \(\mathrm{F}(1,19)=4.90, p\) \(<.039\), with faster RTs for congruent trials (Mean \(=638 \mathrm{~ms}\) ) than incongruent trials (Mean \(=648 \mathrm{~ms}\) ). No significant main effects of spatial attention or visual field (both \(p>.10\) ) were found. There was no interaction between spatial attention and congruency ( \(p>.70\) ) or between visual field and congruency ( \(p>.29\) ), suggesting that spatial attention and visual field did not modulate the priming effect under this supraliminal condition.

The same \(2 \times 2 \times 2\) repeated measures ANOVA was conducted on the accuracy data; no reliable effects or interactions were observed (all p > .10).

\section*{Discussion}

The \(150-\mathrm{ms}\) prime was long enough for conscious processing of category information, as participants had well above-chance performance in the prime-discrimination task. At the conscious level, a robust congruency priming effect was observed, both when the prime was cued or uncued. This result was consistent with the results of Lachter, Forster and Ruthruff (2004)'s study (experiment 1). In a lexical decision task, Lachter et al. (2004) observed a significant repetition priming effect in conditions where a supraliminal prime ( 110 or 165 ms ) was at a task-relevant or task-irrelevant position. At supraliminal level with a long prime duration, participants were typically able to shift their attention to the prime even when the prime was at an irrelevant location (Lachter, Forster, \& Ruthruff, 2004) or uncued. Thus, no attentional modulation effect was observed.

\section*{Experiment 2}

In experiment 2, the prime duration was changed to 50 ms in order to be rendered invisible, as was in the previous studies (Finkbeiner \& Palermo, 2009; Harry, Davis, \& Kim, 2012). In contrast to the superordinate level recognition task (animal vs. tool; or animal vs. vegetable) used in Finkbeiner and Palermo (2009)'s study (the animal condition); participants in the current study were to categorize images at the basic level (cat vs. dog). Finkbeiner and Palermo (2009) found that nonconscious processing of superordinate-level category information relies on the engagement of spatial attention. Here we aimed to examine whether a similar effect can be found in the processing of basic-level category information.

\section*{Method}

Participants 16 students ( 9 females, age mean \(=21.8\), range \(=18-27)\) at the University of Hong Kong participated in exchange for course credits or payment. All were Asian, with normal or corrected-to-normal vision. None of them participated in experiment 1.

Stimuli, Design and Procedure All settings in experiment 2 were exactly the same as those in the experiment 1, except that the prime duration was changed from 150 ms to 50 ms .

\section*{Results}

In analyzing the RTs, null responses and incorrect responses (6.8\%) were excluded. And any RT outside 2.5 standard deviations from the mean, for each participant, was treated as outliers (3.7\%). In analyzing the accuracy data, only null responses (1.5\%) were excluded.

The result of the prime-discrimination task confirmed that the 50 -ms prime was invisible, with the mean accuracy of
51.6\% being not significantly different from the \(50 \%\) chance level, \(\mathrm{t}(15)=1.64, \mathrm{p}>.10\).

A 2 (spatial attention) \(\times 2\) (congruency) \(\times 2\) (visual field) repeated measures ANOVA on the RTs showed a significant main effect of congruency, \(\mathrm{F}(1,15)=4.88, p\) \(<.043\), with faster RTs for congruent trials ( 668 ms ) than incongruent trials ( 680 ms ). In contrast to the results of Finkbeiner and Palermo (2009), however, we did not find evidence of attentional modulation on the congruency priming effect, \(\mathrm{F}(1,15)=0.65, p=.43\). In addition, followup analysis revealed a significant priming effect when the prime was uncued, F \((1,15)=6.91, \mathrm{p}<.019\); but the priming effect when the prime was cued failed to reach significance, \(p>.50\). There was no main effect of visual field ( \(p>.40\) ) or interaction between visual field and congruency ( \(p>.60\) ).

The same \(2 \times 2 \times 2\) repeated measures ANOVA on the accuracy data revealed a main effect of congruency, \(\mathrm{F}(1,15)\) \(=6.74, p<.02\), with lower accuracy in congruent trials (93.5\%) than incongruent trials (95.7\%). No modulation effects of attention ( \(p>.70\) ) or visual field ( \(p>.50\) ) were found.

Table 1: Mean RTs (SE) in experiment 1 and 2
\begin{tabular}{lll}
\hline \multicolumn{1}{c}{ Condition } & \multicolumn{2}{c}{ RTs (ms) } \\
& Prime-cued & Prime-un-cued \\
\hline 150-ms prime & & \\
Congruent & \(635(14)\) & \(641(15)\) \\
Incongruent & \(647(14)\) & \(648(13)\) \\
Priming & 12 & 7 \\
& & \\
50-ms prime & & \\
Congruent & \(670(17)\) & \(666(18)\) \\
Incongruent & \(676(19)\) & \(685(18)\) \\
Priming & 6 & 19 \\
\hline
\end{tabular}

\section*{Discussion}

The results in experiment 2 showed no evidence suggesting attentional modulation on the nonconscious priming effect in the basic-level categorization task; in addition, a significant nonconscious priming effect was observed when the prime was not cued. This effect suggests that nonconscious processing of basic-level category information does not rely on spatial attention, in contrast to the finding that nonconscious processing of superordinate-level category information depends on spatial attention (Finkbeiner \& Palermo, 2009).

Results in some earlier studies examining nonconscious priming effects with visual word stimuli (Fuentes, Carmona, Agis, \& Catena, 1994; Kiefer \& Brendel, 2006) were consistent with this finding. In Fuentes et al. (1994)'s study, for instance, a lexical decision task was used. In experiment 2 , the two nonconscious prime words, one presented in the foveal and the other in the parafoveal vision, were presented at the same time, followed by masks. The center target
could be semantically related to the foveal prime or the parafoveal prime. The results showed that similar, reliable priming effects were found for both the foveal and parafoveal prime. It was argued that semantic representations stored in the long-term memory were ready to be activated even when no attention was drawn to the nonconscious prime word (Kiefer \& Brendel, 2006). Thus, similar to semantic information, basic-level category information may also be readily accessible even when no attention was drawn to the nonconscious prime picture.

\section*{Experiment 3}

In experiment 3, the prime duration was further changed to 10 ms . As no modulation from attention or VF was observed in experiment 1 and 2 , experiment 3 aimed to explore it further by testing the condition where the awareness of the prime was extremely low.

\section*{Method}

Participants 18 students ( 14 females, age mean \(=22.9\), range \(=18-29)\) at the University of Hong Kong participated in exchange for payment. All were Asian, with normal or corrected-to-normal vision. None of them participated in experiment 1 or 2.

Stimuli, Design and Procedures All settings were exactly the same as those in experiment 1, except that the prime duration was changed to 10 ms and that there was a \(40-\mathrm{ms}\) inter-mask between the prime and the target.

\section*{Results}

In analyzing the RTs, null responses and incorrect responses (4.8\%) were excluded. And any RT outside 2.5 standard deviations from the mean, for each participant, was treated as outliers (3.3\%). In analyzing the accuracy data, only null responses ( \(1.0 \%\) ) were excluded.

The results of the prime-discrimination task confirmed that the \(10-\mathrm{ms}\) prime was invisible, with the mean accuracy of \(50.5 \%\) being not significantly different from the \(50 \%\) chance level, \(\mathrm{t}(17)=0.69, p>.50\).

A 2 (spatial attention) \(\times 2\) (congruency) \(\times 2\) (visual field) repeated measures ANOVA on the RTs did not find any reliable effect or interaction (all \(p>.10\) ).

The same \(2 \times 2 \times 2\) repeated measures ANOVA on the accuracy data, however, showed a significant interaction between VF and congruency, F \((1,17)=5.64, p<.03\), but no main effect of congruency, \(p>.15\), or interaction between attention and congruency, \(p>.70\). Follow-up analysis revealed a significant congruency priming effect when the prime was presented in the RVF, \(\mathrm{F}(1,17)=9.19\), \(p<.01\), with a higher accuracy in the congruent trials (97.3\%) than that in the incongruent trials (95.2\%); in contrast, no congruency priming effect was observed in the LVF, \(p>.60\). This priming effect in the RVF did not interact with spatial attention, \(\mathrm{F}(1,17)=1.41, p>.25\). Also, there was no significant interaction between congruency and spatial attention overall, \(\mathrm{F}(1,17)=0.15, p>.70\).

\section*{Discussion}

With extremely low awareness of the prime (with only 10ms duration), a significant congruency priming effect was obtained on accuracy in the RVF/LH, but not when the prime was presented in the LVF/RH. Marzouki et al. (2007) found similar results using a \(45-\mathrm{ms}\) prime with letter/pseudo letter stimuli. One potential concern in Marzonki et al. (2007)'s study was that, participants always used their right hand to respond to letter targets. Thus it was possible that the RVF superiority effect was due to the compatibility between prime location and response hand. In the present study, the location-response compatibility was counterbalanced across participants, and a strong RVF (LH) advantage was still observed.

The finding of the RVF/LH superiority in congruency priming in the basic-level categorization task here is consistent with Marsolek’s (1999) results, which suggested that the LH was biased towards abstract-category representations whereas the RH was biased towards exemplar-specific representations (e.g., Laeng et al., 2003; Marsolek, 1995, 1999; Studer \& Hübner, 2008). Here our participants were required to make a basic-level category judgment on the target, with a prime presented in either the LVF or the RVF. The target and the prime were always different exemplars from the same category or from different categories. Since the prime was always different from the target, the abstract category representation in the LH activated by the prime was able to facilitate the categorization of the target. Thus, the priming effect was only observed with a RVF prime but not a LVF prime.

A critical difference between the present study and the earlier studies was that visual information projected to the LH and the RH was rendered nonconscious by masks and brief duration. Our findings suggested that the processing of nonconscious information could go beyond the sensory level to induce hemispheric asymmetry effects

\section*{General Discussion}

In the current study, we aimed to examine whether there is hemispheric asymmetry in the processing of nonconscious information. We observed a nonconscious congruency priming effect on accuracy when the prime was presented in the RVF/LH, but not when the prime was presented in the LVF/RH, with a \(10-\mathrm{ms}\) prime, in a basic-level object categorization task.

Previous studies have shown that hemispheric asymmetry effects typically do not emerge at an early visual processing stage, such as findings in grating detection tasks (e.g., Fendrich \& Gazzaniga, 1990). In contrast, hemispheric specialization can typically be observed in high-level perception tasks such as face recognition (e.g., Keenan et al., 1989). Our results of hemispheric asymmetry in subliminal priming suggest that the processing of subliminal information can go beyond the sensory level to influence high-level perceptual processes. It is consistent with previous findings that nonconscious information could
influence semantic processing (e.g., Yeh, He, \& Cavanagh, 2012).

The LH advantage in the present study was in accordance with the abstract-specific representation account of hemisphere asymmetry in the representation of category knowledge proposed by Marsolek (1999). However, surprisingly, this effect was not found under the \(50-\mathrm{ms}\) (nonconscious) and \(150-\mathrm{ms}\) (conscious) prime conditions. The effect that longer prime presentation duration diminished hemispheric asymmetry effects in subliminal priming may be due to interhemispheric communications. It is possible that for a prime appeared in the LVF/RH, 50 ms or longer was long enough for the information to reach the LH through the corpus callosum, and to consequently activate the abstract representation of the corresponding category, resulting in a similar level of priming as compared with a prime presented in the RVF/LH. 10 ms duration, however, might be insufficient for a LVF/RH subliminal prime to reach the LH. Without the activation of abstract representations stored in the LH (Marsolek, 1999), a prime in the LVF/RH failed to facilitate the processing of the target in the basic-level categorization task; consequently, a significant priming effect was observed only with a RVF prime. This effect also suggests that hemispheric asymmetry may be better observed at subliminal level.

Note that in the present study, the prime was followed by a mask. Previous studies have revealed that a backward mask is able to block brain processing of stimuli (BaconMacé, Macé, Fabre-Thorpe, \& Thorpe, 2005). In BaconMacé et al.'s (2005) study, an image (with 6.25 ms duration) was followed by a \(100-\mathrm{ms}\) mask. The SOA between the image and the mask was manipulated, raning from 6 ms to 106 ms . Participants were asked to judge whether there was an animal in the image. The behavirol accuracy decreased as the SOA became shorter, reaching the chance level when the SOA was 6 ms . In addition, brain activities trigerred by the image decreased rapidly with shorter SOAs, indicating that brain processing was impaired significantly by the backward mask. These findings further suggest that in our study the \(10-\mathrm{ms}\) LVF/RVF subliminal prime with a backward mask was too fragile to reach the other side of the brain, allowing hemispheric difference to emerge.

Another major finding in the current study was that no attentional modulation on the nonconscious congruency priming effect in the basic-level categorization task was observed across the 3 experiments with different prime durations; in particular, we observed a significant priming effect when attention was not drawn to the \(50-\mathrm{ms}\), invisible prime (experiment 2). This result was in contrast to the superordinate-level categorization task used in Finkbeiner and Palermo (2009)'s study, in which the nonconscious priming effect depended on attention. Our results provided evidence for the influence from levels of abstraction in a categorization task in nonconscious processing. This is consistent with previous findings that basic-level category information was accessed first in the perception of visual objects and named first by children (Rosch et al., 1976).

Why is basic level information special? It is possible that in the human evolution, people had to recognize "tiger" as a tiger (i.e., the basic level), rather than as an animal (i.e., the superordinate level), in order to detect and escape from potential dangers in the environment. The basic-level categorization was crucial for survival, and thus through evolution we developed the ability to process basic-level category information both nonconsciously and without attention.

\section*{Acknowledgments}

We are grateful to the Research Grant Council of Hong Kong (project code HKU 745210H to J.H. Hsiao) and the HKU Seed Funding Program for Basic Research (project code 201011159124 to J.H. Hsiao).

\section*{References}

Bacon-Macé, N., Macé, M.J.M., Fabre-Thorpe, M., \& Thorpe, S.J. (2005). The time course of visual processing: Backward masking and natural scene categorisation. Vision research, 45(11), 1459-1469.
Fendrich, R., \& Gazzaniga, M. (1990). Hemispheric processing of spatial frequencies in two commissurotomy patients. Neuropsychologia, 28(7), 657-663.
Finkbeiner, M., \& Palermo, R. (2009). The Role of Spatial Attention in Nonconscious Processing A Comparison of Face and Nonface Stimuli. Psychological Science, 20(1), 42-51.
Fuentes, L.J., Carmona, E., Agis, I.F., \& Catena, A. (1994). The role of the anterior attention system in semantic processing of both foveal and parafoveal words. Journal of Cognitive Neuroscience, 6(1), 17-25.
Harry, B., Davis, C., \& Kim, J. (2012). Subliminal access to abstract face representations does not rely on attention. Consciousness and cognition, 21, 573-583.
Hsiao., Janet H., Cipollini., Ben, \& Cottrell., Garrison W. (accepted). Hemispheric asymmetry in perception: A differential encoding account. Journal of Cognitive Neuroscience.
Jolicoeur, P., Gluck, M.A., \& Kosslyn, S.M. (1984). Pictures and names: Making the connection. Cognitive psychology, 16(2), 243-275.
Keenan, P.A., Whitman, RD, \& Pepe, J. (1989). Hemispheric asymmetry in the processing of high and low spatial frequencies: A facial recognition task. Brain and cognition, 11(2), 229-237.
Kiefer, M., \& Brendel, D. (2006). Attentional modulation of unconscious "automatic" processes: evidence from eventrelated potentials in a masked priming paradigm. Journal of Cognitive Neuroscience, 18(2), 184-198.
Kouider, S., \& Dehaene, S. (2007). Levels of processing during non-conscious perception: a critical review of visual masking. Philosophical Transactions of the Royal Society B: Biological Sciences, 362(1481), 857-875.

Lachter, J., Forster, K.I., \& Ruthruff, E. (2004). Forty-five years after Broadbent (1958): still no identification without attention. Psychological review, 111(4), 880-913.
Laeng, B., Zarrinpar, A., \& Kosslyn, S.M. (2003). Do separate processes identify objects as exemplars versus members of basic-level categories? Evidence from hemispheric specialization. Brain and Cognition, 53(1), 15-27.
Marsolek, C.J. (1995). Abstract visual-form representations in the left cerebral hemisphere. Journal of Experimental Psychology: Human Perception and Performance, 21(2), 375-386.
Marsolek, C.J. (1999). Dissociable neural subsystems underlie abstract and specific object recognition. Psychological Science, 10(2), 111-118.
Marzouki, Y., Grainger, J., \& Theeuwes, J. (2007). Exogenous spatial cueing modulates subliminal masked priming. Acta psychologica, 126(1), 34-45.
Parkhi, O.M., Vedaldi, A., Zisserman, A., \& Jawahar, CV. (2012). Cats and dogs. Paper presented at the Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on.
Rosch, E., Mervis, C.B., Gray, W.D., Johnson, D.M., \& Boyes-Braem, P. (1976). Basic objects in natural categories. Cognitive psychology, 8(3), 382-439.
Sergent, J. (1982). The cerebral balance of power: Confrontation or cooperation? Journal of Experimental Psychology: Human Perception and Performance, 8(2), 253-272.
Studer, T., \& Hübner, R. (2008). The direction of hemispheric asymmetries for object categorization at different levels of abstraction depends on the task. Brain and cognition, 67(2), 197-211.
Tanaka, J.W. (2001). The entry point of face recognition: evidence for face expertise. Journal of Experimental Psychology: General, 130(3), 534-543.
Yeh, S.L., He, S., \& Cavanagh, P. (2012). Semantic priming from crowded words. Psychological science, 23(6), 608616.

\title{
Generative Inferences Based on a Discriminative Bayesian Model of Relation Learning
}

\author{
Dawn Chen \({ }^{1}\) (sdchen@ucla.edu) \\ Hongjing Lu \({ }^{1,2}\) (hongjing@ucla.edu) \\ Keith J. Holyoak \({ }^{1}\) (holyoak@lifesci.ucla.edu) \\ Departments of Psychology \({ }^{1}\) and Statistics \({ }^{2}\) \\ University of California, Los Angeles \\ Los Angeles, CA 90095 USA
}

\begin{abstract}
Bayesian Analogy with Relational Transformations (BART) is a discriminative model that can learn comparative relations from non-relational inputs (Lu, Chen \& Holyoak, 2012). Here we show that BART can be extended to solve inference problems that require generation (rather than classification) of relation instances. BART can use its generative capacity to perform hypothetical reasoning, enabling it to make quasideductive transitive inferences (e.g., "If \(A\) is larger than \(B\), and \(B\) is larger than \(C\), is \(A\) larger than \(C ? "\) ). The extended model can also generate human-like instantiations of a learned relation (e.g., answering the question, "What is an animal that is smaller than a dog?"). These modeling results suggest that discriminative models, which take a primarily bottom-up approach to relation learning, are potentially capable of using their learned representations to make generative inferences.
\end{abstract}

Keywords: Bayesian models; generative models; discriminative models; relation learning; transitive inference; deduction; induction; hypothetical reasoning

\section*{Introduction}

\section*{Generative and Discriminative Models}

Bayesian models of inductive learning can be designed to focus on learning either the probabilities of observable features given concepts (generative models) or the probabilities of concepts given features (discriminative models; Friston et al., 2008; Mackay, 2003). Generative models are especially powerful as they are capable of not only classifying novel instances of concepts (using Bayes’ rule to invert conditional probabilities), but also generating representations of possible instances. In contrast, discriminative models focus directly on classification tasks, but do not provide any obvious mechanism for making generative inferences.

In recent years, generative Bayesian models have been developed to learn complex concepts based on relational structures (e.g., Goodman, Ullman \& Tenenbaum, 2011; Kemp \& Jern, 2009; Kemp, Perfors \& Tenenbaum, 2007; Tenenbaum, Kemp, Griffiths \& Goodman, 2011). Representations of alternative relational structures are used to predict incoming data, and the data in turn are used to revise probability distributions over alternative structures. The highest level of the structure typically consists of a
formal grammar or a set of logical rules that generates alternative relational "theories", which are in turn used to predict the observed data. That is, the set of possible relational structures is provided to the system by specifying a grammar that generates them.

Despite their impressive successes, there are some reasons to doubt whether the generative approach provides an adequate basis for all psychological models of relation learning. Since the postulated grammar of relations is not itself learned, the generative approach implicitly makes rather strong nativist assumptions. Moreover, generative models of relation learning do not fit the intuitive causal direction. For example, it seems odd to claim that a binary relation such as larger than somehow acts to causally generate an ordered pair (e.g., <dog, cat>) that constitutes an instantiation of the relation. It seems more natural to consider how observable features of the objects in the ordered pair give rise to the truth of the relation, i.e., to apply a discriminative approach.

\section*{Discriminative Models of Relation Learning}

Recently, discriminative models have also been applied to relation learning. Silva, Heller, and Ghahramani (2007) developed a discriminative model for relational tasks such as identifying classes of hyperlinks between webpages and classifying relations based on protein interactions. Although their model was developed to address applications in machine learning, the general principles can potentially be incorporated into models of human relational learning. One key idea is that an \(n\)-ary relation can be represented as a function that takes ordered sets of \(n\) objects as its input and outputs the probability that these objects instantiate the relation. The model learns a representation of the relation from labeled examples, and then applies the learned representation to classify novel examples. A second key idea is that relation learning can be facilitated by incorporating empirical priors, which are derived using some simpler learning task that can serve as a precursor to the relation learning task.

These ideas were incorporated into Bayesian Analogy with Relational Transformations (BART), a discriminative model that can learn comparative relations from nonrelational inputs (Lu, Chen \& Holyoak, 2012). Given
independently-generated feature vectors representing pairs of animals that exemplify a relation, the model acquires representations of first-order comparative relations (e.g., larger, faster) as weight distributions over the features. Learning is guided by empirical priors for the weight distributions derived from initial learning of one-place predicates (e.g., large, fast). BART's learned relations support generalization to new animal pairs, allowing the model to discriminate between novel pairs that instantiate a relation and those that do not. Moreover, BART's learned weight distributions can be systematically transformed to solve analogies based on higher-order relations (e.g., opposite).

BART has thus demonstrated promise as a discriminative model of relation learning, which does not presuppose an innate grammar of relations. However, the challenge remains to extend the model to tasks requiring generative inferences. For example, people are able to construct actual instantiations of relations, answering questions such as, "What is an animal that is smaller than a dog?" (Although one might suppose that such questions could be answered by undirected trial-and-error, we shall see that people's answers are often systematically guided by their representations of the relation and of the animal provided as a cue.) Another challenging task is purely hypothetical reasoning, which requires making inferences about arbitrary instances of the relation. Comparative relations such as larger exhibit the logical properties of transitivity and asymmetry, supporting deductions such as "If \(A\) is larger than \(B\), and \(B\) is larger than \(C\), then \(A\) is larger than \(C\)." Children as young as five or six years can make such transitive inferences reliably (Halford, 1992; Goswami, 1995; Kotovsky \& Gentner, 1996). In the present paper we describe an extension of the BART model that addresses these challenges of making generative inferences.

\section*{BART Model of Relation Learning}

\section*{Domain and Inputs}

We focus on the same domain and inputs used in the initial BART project (Lu et al., 2012): the domain of comparative relations between animal concepts (e.g., a cow is larger than a sheep). To establish the "ground truth" of whether various pairs of animals instantiate different comparative relations, Lu et al. used a set of human ratings of animals on four different continua (size, speed, fierceness, and intelligence; Holyoak \& Mah, 1981). These ratings made it possible to test the model on learning eight different comparative relations: larger, smaller, faster, slower, fiercer, meeker, smarter, and dumber.

Each animal concept is represented by a real-valued feature vector. In order to avoid the perils of hand-coded inputs (i.e., the possibility that the model's successes may be partly attributable to the foresight and charity of the modelers), we use what we call "Leuven vectors." These representations are derived from norms of the frequencies with which participants at the University of Leuven
generated features characterizing 129 different animals (De Deyne et al., 2008; see Shafto, Kemp, Mansinghka, \& Tenenbaum, 2011). Each animal in the norms is associated with a set of frequencies across more than 750 features. We created vectors of length 50 based on the 50 features most highly associated with the subset of 44 animals that are also in the ratings dataset (Lu et al., 2012). Figure 1 provides a visualization (for 30 example animals and the first 15 of the 50 features) of these high-dimensional and distributed representations, which might be similar to the representations underlying people's everyday knowledge of various animals.


Figure 1: Illustration of Leuven vectors (reduced to 15 features to conserve space) for some example animals. The cell intensities represent feature values (light indicates high values and dark indicates low values).

\section*{Relations as Weight Distributions}

BART represents a relation using a joint distribution of weights, \(\mathbf{w}\), over object features. A relation is learned by estimating the probability distribution \(P\left(\mathbf{w} \mid \mathbf{X}_{\mathbf{s}}, \mathbf{R}_{\mathbf{s}}\right)\), where \(\mathbf{X}_{\mathbf{s}}\) represents the feature vectors for object pairs in the training set, the subscript \(\mathbf{S}\) indicates the set of training examples, and \(\mathbf{R}_{\mathrm{s}}\) is a set of binary indicators, each of which (denoted by \(R\) ) indicates whether a particular object (or pair of objects) instantiates the relation or not. The vector \(\mathbf{w}\) constitutes the learned relational representation, which can be interpreted as weights reflecting the influence of the corresponding feature dimensions in \(\mathbf{X}\) on judging whether the relation applies. The weight distribution can be updated based on examples of ordered pairs that instantiate the relation. Formally, the posterior distribution of weights can be computed by applying Bayes' rule using the likelihood of the training data and the prior distribution for \(\mathbf{w}\) :
\[
\begin{equation*}
P\left(\mathbf{w} \mid \mathbf{X}_{\mathbf{s}}, \mathbf{R}_{\mathbf{s}}\right)=\frac{P\left(\mathbf{R}_{\mathbf{s}} \mid \mathbf{w}, \mathbf{X}_{\mathbf{s}}\right) P(\mathbf{w})}{\int_{\mathbf{w}} P\left(\mathbf{R}_{\mathbf{s}} \mid \mathbf{w}, \mathbf{X}_{\mathbf{s}}\right) P(\mathbf{w})} \tag{1}
\end{equation*}
\]

The likelihood is defined as a logistic function for computing the probability that a pair of objects instantiates the relation, given the weights and feature vector:
\[
\begin{equation*}
P(R=1 \mid \mathbf{w}, \mathbf{x})=\frac{1}{1+e^{-\mathbf{w}^{\mathrm{T}} \mathbf{x}}} \tag{2}
\end{equation*}
\]

The prior, \(P(\mathbf{w})\), is a Gaussian distribution and is constructed using a bottom-up approach in which initial learning of simple concepts provides empirical priors that guide subsequent learning of more complex concepts. Specifically, BART extracts empirical priors from weight distributions for one-place predicates such as large to guide the acquisition of two-place relations such as larger. Lu et al. (2012) trained BART on the eight one-place predicates (e.g., large, small, fierce, meek) that can be formed using the extreme animals at each end of the four relevant continua (size, speed, ferocity, and intelligence).

After learning the joint weight distribution that represents a relation, BART discriminates between pairs that instantiate the relation and those that do not by calculating the probability that a target pair \(\mathbf{x}_{T}\) instantiates the relation \(R\) :
\[
\begin{align*}
& P\left(R_{T}=1 \mid \mathbf{x}_{T}, \mathbf{X}_{\mathbf{s}}, \mathbf{R}_{\mathbf{s}}\right)= \\
& \int_{\mathbf{w}} P\left(R_{T}=1 \mid \mathbf{x}_{T}, \mathbf{w}\right) P\left(\mathbf{w} \mid \mathbf{X}_{\mathbf{s}}, \mathbf{R}_{\mathbf{s}}\right) . \tag{3}
\end{align*}
\]

Although the general framework of the relation learning model is straightforward, the calculations of the normalization term in Eq. (1) and the integral in Eq. (3) are intractable, lacking analytic solutions. As in Silva, Heller, and Gharamani (2007), we employed the variational method developed by Jaakkola and Jordan (2000) for Bayesian logistic regression to obtain closed-form approximations to the posterior weight distribution \(P\left(\mathbf{w} \mid \mathbf{X}_{\mathrm{s}}, \mathbf{R}_{\mathrm{S}}\right)\) and the predictive probability \(P\left(R_{T}=1 \mid \mathbf{x}_{T}, \mathbf{X}_{\mathrm{S}}, \mathbf{R}_{\mathrm{S}}\right)\).

\section*{Extension to Generative Inference}

The goal of the present paper is to endow BART with generative abilities, allowing it (for example) to complete a partially-instantiated relation, answering questions such as, "What is an animal that is smaller than a dog?" We use the weight representation for a relation learned by BART to construct a new generative model for the completion task. When presented with a cue relation (e.g., smaller) and a cue object (e.g., dog), the model produces possible responses for the remaining object (e.g., cat) so that the ordered object pair satisfies the relation. More specifically, given the features of an object \(B, \mathbf{x}_{B}\), and the knowledge that relation \(R\) holds for the object pair \((A, B)\), the model generates a probability distribution for the features of object \(A, \mathbf{x}_{A}\), by making the following inference:
\[
\begin{equation*}
P\left(\mathbf{x}_{A} \mid \mathbf{x}_{B}, R=1\right) \propto P\left(R=1 \mid \mathbf{x}_{A}, \mathbf{x}_{B}\right) P\left(\mathbf{x}_{A} \mid \mathbf{x}_{B}\right) . \tag{4}
\end{equation*}
\]

The likelihood term, \(P\left(R=1 \mid \mathbf{x}_{A}, \mathbf{x}_{B}\right)\), is the probability that relation \(R\) holds for a particular hypothesized object \(A\), \(\mathbf{x}_{A}\), and the known object \(B, \mathbf{x}_{B}\). It is defined using a logistic function, just as in Eq. (2):
\[
\begin{equation*}
P\left(R=1 \mid \mathbf{x}_{A}, \mathbf{x}_{B}\right)=\frac{1}{1+e^{-\boldsymbol{w}_{1}^{T} \mathbf{x}_{A}-w_{2}^{T} x_{B}}} . \tag{5}
\end{equation*}
\]

Relative to Eq. (2), we have only introduced small differences in the notation. The learned relational weights, \(\mathbf{w}\), are written as two separate halves: weights associated


Figure 2: Illustration of the generative model for inferring an animal that is larger than a sheep. Colors annotate probability densities (red indicates high values and blue indicates low values). The top panel shows the prior and posterior distributions with \(\sigma^{2}=7\) (favoring similaritybased completions such as cow), and the bottom panel shows the prior and posterior with \(\sigma^{2}=25\) (favoring "landmark" completions such as elephant). Various animals are represented in the two-dimensional space based on their size and speed ratings. The posterior was generated using the relational weights that BART learned from the full ratings input (i.e., all four dimensions).
with the first relational role ( \(\mathbf{w}_{1}\) ) and weights associated with the second relational role ( \(\mathbf{w}_{2}\) ). Similarly, the feature vector \(\mathbf{x}\) for a pair of objects is separated into the feature vector for object \(A\left(\mathbf{x}_{A}\right)\) and the feature vector for object \(B\) ( \(\mathrm{x}_{B}\) ).
The prior for the features of object \(A, P\left(\mathbf{x}_{A} \mid \mathbf{x}_{B}\right)\), is the conditional distribution given the features of object \(B\). It is defined as the following:
\[
\begin{equation*}
P\left(\mathbf{x}_{A} \mid \mathbf{x}_{B}\right)=N\left(\mathbf{x}_{B}, \sigma^{2} \mathbf{I}\right) . \tag{6}
\end{equation*}
\]

We assume that object \(B\) (the referent) serves a starting point for generating object \(A\), so the means of \(P\left(\mathbf{x}_{A} \mid \mathbf{x}_{B}\right)\) are taken to be the feature values of object \(B\), reflecting a certain degree of semantic dependency between the two objects (i.e., people's tendency to think of \(A\) objects that are similar to \(B\) ). The prior also encodes the assumptions that the features of \(A\) are uncorrelated and have the same variance \(\sigma^{2}\), the value of which is a free parameter reflecting the strength of the model's preference for generating \(A\) objects that are similar to \(B\).
Our generative model infers a feature distribution for object \(A\) that reflects a compromise between (1) maximizing
the semantic similarity of \(A\) and \(B\), which is reflected in the prior term; and (2) maximizing the probability that the relation holds, which is reflected in the likelihood term. We adapted the variational method to estimate the posterior distribution, using the following update rules for the mean \(\boldsymbol{\mu}\) and covariance matrix \(\mathbf{V}\) of the feature distribution, as well as the variational parameter \(\xi\) :
\[
\begin{gather*}
\mathbf{V}^{-1}=\frac{\mathbf{I}}{\sigma^{2}}+2 \lambda(\xi) \mathbf{w}_{1} \mathbf{w}_{1}^{\mathbf{T}} \\
\boldsymbol{\mu}=\mathbf{V}\left(\frac{\mathbf{I}}{\sigma^{2}} \mathbf{x}_{B}+\frac{\mathbf{w}_{1}}{2}-2 k \lambda(\xi) \mathbf{w}_{1}\right),  \tag{7}\\
\xi^{2}=\mathbf{w}_{1}^{\mathbf{T}}\left(\mathbf{V}+\boldsymbol{\mu} \boldsymbol{\mu}^{\mathbf{T}}\right) \mathbf{w}_{1},
\end{gather*}
\]
where \(\lambda(\xi)=\frac{\tanh \left(\frac{1}{2}(\xi+k)\right)}{4(\xi+k)}\) and \(k=\mathbf{w}_{2}^{\mathrm{T}} \mathbf{x}_{B}\).
Figure 2 illustrates the operation of the model in generating an animal \((A)\) that is larger than a sheep \((B)\). The feature distribution for \(A\) is updated from a prior favoring some degree of similarity between the two animals (left panel; top: high similarity, bottom: low similarity) to a posterior distribution after taking into consideration the relation (i.e., larger) instantiated by the animals (right panel). These distributions are shown in a simplified twodimensional feature space (the size and speed ratings for animals; Holyoak \& Mah, 1981).

\section*{Modeling Transitive Inference}

Comparative relations such as larger exhibit the logical properties of transitivity and asymmetry, supporting deductions such as, "If \(A\) is larger than \(B\) and \(B\) is larger than \(C\), then \(A\) is larger than \(C\)." Such hypothetical reasoning seems to depend on the ability to generate arbitrary instantiations of the relation without any guidance from object features (as the object representations are semantically empty). Our first test evaluated whether the generative extension of BART enables transitive inferences on comparative relations using arbitrary hypothetical instances.

\section*{Operation of the Model}

The basic approach to transitive inference is straightforward: The model "imagines" objects \(A, B\), and \(C\) that instantiate the two given premises, as in the example above, and then tests the unstated relationship for the pair \(<A, C>\). If the larger relation that BART has learned is indeed transitive, then any such instantiation will satisfy the conclusion, " \(A\) is larger than \(C\)." This test is done repeatedly, in essence searching for a counterexample. If no counterexample is ever found, the transitive inference is accepted.

Specifically, for each of the eight comparative relations that BART learned, we first let the model "imagine" an animal \(B\) (because the statement " \(A\) is larger than \(B\) " implies that \(B\) is the referent against which \(A\) is being compared) by sampling a feature vector from a distribution representing
the animal category. This is a Gaussian distribution with a mean vector and covariance matrix that were directly estimated from the feature vectors of the 44 animals in the Leuven dataset that are included in the ratings dataset.

Given the sampled animal \(B\), the generative model constructs a distribution for animal \(A\) (e.g., to satisfy the premise that " \(A\) is larger than \(B\) ") by letting \(B\) fill the second role of the relevant relation. Similarly, the model constructs a distribution for animal \(C\) (e.g., to satisfy the premise that " \(B\) is larger than \(C\) ") by letting \(B\) fill the first role of the same relation. Next, the model creates feature representations for specific animals \(A\) and \(C\) by setting their feature vectors, \(\mathbf{x}_{A}\) and \(\mathbf{x}_{C}\), to be the means of the inferred feature distributions for \(A\) and \(C\), respectively. Note that these "imagined" animals are hypothetical: although their features are sampled from the distribution of animal features, the results will seldom correspond to actual animals. To ensure that the premises have actually been satisfied, the model accepts the imagined animal \(A\) only if \(P\left(R=1 \mid \mathbf{x}_{A}, \mathbf{x}_{B}\right)>0.5\) and \(P\left(R=1 \mid \mathbf{x}_{B}, \mathbf{x}_{A}\right)<0.5\), and the imagined animal \(C\) only if \(P\left(R=1 \mid \mathbf{x}_{B}, \mathbf{x}_{C}\right)>0.5\) and \(P\left(R=1 \mid \mathbf{x}_{C}, \mathbf{x}_{B}\right)<0.5\).
Finally, if \(\mathbf{x}_{A}\) and \(\mathbf{x}_{C}\) have been accepted as satisfying the premises, the model calculates both \(P\left(R=1 \mid \mathbf{x}_{A}, \mathbf{x}_{C}\right)\), denoting the probability that \(A\) is larger than \(C\), and \(P\left(R=1 \mid \mathbf{x}_{C}, \mathbf{x}_{A}\right)\), denoting the probability that \(C\) is larger than \(A\). The model concludes that the relation holds for the pair \(\left\langle A, C>\right.\) (and not for \(\left\langle C, A>\right.\) ) if \(P\left(R=1 \mid \mathbf{x}_{A}, \mathbf{x}_{C}\right)>0.5\) and \(P\left(R=1 \mid \mathbf{x}_{C}, \mathbf{x}_{A}\right)<0.5\), implying that a counterexample has not yet been found to refute the transitive inference.

We conducted tests of transitive inference using the relational representations that BART learned based on 100 randomly-chosen training pairs. For comparison, we also tested a baseline model that substituted an uninformative prior for the empirical prior that guides BART's relation learning (see Lu et al., 2012). For each of the eight comparative relations, the relation learning model was run ten times, each time with a different set of training pairs and resulting in a different learned weight distribution. For each of these learned weight distributions, we let the model generate \(100 A-B-C\) triads satisfying the premises, testing the relevant relationship between \(A\) and \(C\) for each triad. To assess the influence of the free parameter in model predictions, the tests were conducted multiple times with different values of \(\sigma^{2}\) ranging from 1 to 1000 . The strongest tests are those in which \(\sigma^{2}\) is set at low values, creating a strong prior preference that \(A, B\), and \(C\) are similar to one another. When the similarity constraint is strong, the model is forced to generate animals that are similar on the relevant dimension, and hence more likely to yield a counterexample. When the value of \(\sigma^{2}\) was reduced below 1, the models produced many instantiations that did not satisfy the required premises (i.e., \(A>B, B>C\), and not vice versa). We therefore treated the value of 1 as the
minimal value of \(\sigma^{2}\) that yields triplets of animals with discriminable values on the relevant dimension.

\section*{Results and Discussion}

Figure 3 shows the mean proportion correct (i.e., the mean proportion of triads that satisfy the conclusion based on transitive inference) for BART and the baseline model as a function of \(\sigma^{2}\). These results are averaged over the eight comparative relations. The critical result is that the BART's accuracy remains constant at \(100 \%\) as \(\sigma^{2}\) is reduced to the effective minimal value of 1 . Thus, BART demonstrates what may be considered an inductive approximation to deduction: despite exhaustive search for a counterexample to the transitive inference, no counterexample is ever found. In contrast, the baseline model often fails to infer that \(A>C\) (and not vice versa) even when the value of \(\sigma^{2}\) is as large as 100 .


Figure 3: Mean proportion correct on the transitive inference task for BART and baseline model, as a function of the variance parameter. These results are averaged across the eight comparative relations.

\section*{Animal Generation Task}

A second evaluation of the model involves predicting the distribution of human responses in an animal generation study conducted using Amazon Mechanical Turk. In this free-generation study, participants typed responses to queries of the form, "Name an animal that is larger than a dog." They were instructed to enter the first animal that came to mind. Four comparative relations (larger, smaller, faster, and slower) and nine cue animals (shark, ostrich, sheep, dog, fox, turkey, duck, dove, and sparrow) were used. At least 50 responses were collected for each of the 36 relation-animal pairs. To minimize learning across trials, we asked each participant to answer only two questions about a single animal: either larger and then slower, slower and then larger, faster and then smaller, or smaller and then faster.

The same relation-animal pairs were presented to the model after it had been trained on the relevant relations. For each question, the model produces a continuous posterior distribution for the feature vector of the missing animal using Eq. (4). This distribution was used to calculate the probability densities for the feature vectors of all animals among the human responses that had Leuven vectors. These probability densities were normalized to produce a discrete


Figure 4: Observed human response proportions and BART's predictions for the queries, "Name an animal that is smaller than a dog" (top), and "Name an animal that is slower than a dog" (bottom).
probability distribution over the animals included in the human responses. The model's predicted probabilities were averaged across the ten runs.

The human results were complex, and here we report only a partial and preliminary attempt to make a comparison with model predictions. Qualitatively, human responses were dominated by two trends: (1) reporting an animal similar to the cue animal and fitting the cue relation (e.g., cat for "smaller than a dog"), or (2) reporting a "landmark" animal at an extreme of the continuum (e.g., turtle for "slower than a dog"). The landmark animal coupled with the cue animal provides an ideal example of the cue relation. This tradeoff between reporting animals that are similar to the cue animal and reporting animals that are landmarks for the cue relation (and usually more dissimilar to the cue animal) is captured by the single free parameter in the generative module, \(\sigma^{2}\). As explained earlier (see Figure 2), a low \(\sigma^{2}\) results in a response distribution that favors animals similar to the cue animal, whereas a high \(\sigma^{2}\) leads to a preference for response animals that are more likely to satisfy the cue relation with respect to the cue animal (i.e., landmark animals for the cue relation).

To reflect the unique pattern of human responses to each question, the variance parameter in the generative model was chosen separately for each question (from the values, 1 , \(5,10,50\), and 100) to maximize the average of Pearson's \(r\) and Spearman's \(\rho\) (rank-order) correlations between the model's predicted probabilities and the observed response proportions for that question. Here we present results for two illustrative questions. The top panel of Figure 4 shows the model's predicted response distribution and the human response distribution for the request, "Name an animal that is smaller than a dog." The human response pattern reveals a strong influence of semantic similarity between the cue
animal and generated animal. The most common human response was cat, followed by mouse (the landmark animal for the smaller relation). With \(\sigma^{2}=10\), the correlation between the model predictions and the human response pattern was \(r=.76\).

The bottom panel of Figure 4 depicts the model predictions and human response pattern for the request, "Name an animal that is slower than a dog." For this question, the most common response was the landmark animal turtle. With \(\sigma^{2}=50\), the correlation between the model predictions and the human response pattern was \(r=\) .72. The higher variance assumed for this question (relative to that for the smaller question) reflects the dominance of the landmark response for the slower question.

Note that even though the two questions use the same cue animal (dog), different sets of animals were generated depending on the cue relation, revealing that humans do take relations into consideration in this free generation task. The model showed a similar pattern of results.

\section*{Conclusions}

These results provide initial evidence that a discriminative model of relation learning, BART (Lu et al., 2012), can be extended to yield generative inferences. These inferences can involve relations between either hypothetical (in the case of transitive inference) or actual (in the case of the animal generation task) objects. In the latter free generation task, preliminary analyses indicate that BART achieves some success in modeling human response patterns.

The model's ability to make transitive inferences based on relations it has learned in a bottom-up fashion from examples illustrates the potential power of the discriminative approach to relation learning. Importantly, BART is not endowed with any notion of what a "transitive and asymmetric" relation is (though like a 6 -year-old child, it is endowed with sufficient working memory to integrate two relations as premises). Rather, it simply uses its learned comparative relations to imagine possible object triads, and without exception concludes that the inference warranted by transitivity holds in each such triad. The model thus approximates "logical" reasoning by a systematic search for counterexamples (and failing to find any), akin to a basic mechanism postulated by the theory of mental models (Johnson-Laird, 2008). The fact that BART achieves errorfree performance in the tests of transitive inference is especially impressive given that its inductively-acquired relational representations are most certainly fallible (e.g., the model makes errors in judging which of two animals close in size is the larger; see Lu et al., 2012). It turns out that imperfect representations of comparative relations, acquired by bottom-up induction, can be sufficiently robust as to yield reliable quasi-deductive transitive inferences.

\section*{Acknowledgments}

Preparation of this paper was supported by grant N000140810186 from the Office of Naval Research.

\section*{References}

De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel,W., Dry, M., Voorspoels, W., \& Storms G. (2008). Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. Behavior Research Methods, 40, 1030-1048.
Friston, K., Chu, C., Mourão-Miranda, J., Hulme, O., Rees, H., Penny, W., \& Ashburner, J. (2008). Bayesian decoding of brain images. NeuroImage, 39, 181-205.
Goodman, N. D., Ullman, T. D., \& Tenenbaum, J. B. (2011). Learning a theory of causality. Psychological Review, 118, 110-119.
Goswami, U. (1995). Transitive relational mappings in 3and 4-year-olds: The analogy of Goldilocks and the Three Bears. Child Development, 66, 877-892.
Halford, G. S. (1992). Analogical reasoning and conceptual complexity in cognitive development. Human Development, 35, 193-217.
Holyoak, K. J., \& Mah, W. A. (1981). Semantic congruity in symbolic comparisons: Evidence against an expectancy hypothesis. Memory \& Cognition, 9, 197-204.
Jaakkola, T. S., \& Jordan, M. I. (2000). Bayesian logistic regression: A variational approach. Statistics and Computing, 10, 25-37.
Johnson-Laird, P.N. (2008) Mental models and deductive reasoning. In L. Rips \& J. Adler. (Eds.), Reasoning: Studies in human inference and its foundations (pp. 206222). Cambridge, UK: Cambridge University Press.

Kemp, C., \& Jern, A. (2009). Abstraction and relational learning. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams \& A. Culotta (Eds.), Advances in Neural Information Processing Systems, 22, 943-951.
Kemp, C., Perfors, A., \& Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. Developmental Science, 10, 307-321.
Kemp, C., \& Tenenbaum, J. B. (2008). The discovery of structural form. Proceedings of the National Academy of Sciences, USA, 105, 10687-10692.
Kotovsky L, \& Gentner D. (1996). Comparison and categorization in the development of relational similarity. Child Development, 67, 2797-2822.
Lu, H., Chen, D., \& Holyoak, K. J. (2012). Bayesian analogy with relational transformations. Psychological Review, 119, 617-648.
Mackay, D. (2003). Information theory, inference and learning algorithms. Cambridge, UK: Cambridge University Press.
Shafto, P., Kemp, C., Mansinghka, V., \& Tenenbaum, J. B. (2011). A probabilistic model of cross-categorization. Cognition, 120, 1-25.
Silva, R., Heller, K., \& Ghahramani, Z. (2007). Analogical reasoning with relational Bayesian sets. In M. Mella \& X. Shen (Eds.), Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics.
Tenenbaum, J. B., Kemp, C., Griffiths, T. L., \& Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. Science, 331, 1279-1285.

\title{
Logical Consistency and Objectivity in Causal Learning
}

\author{
Patricia W. Cheng (cheng@lifesci.ucla.edu) \\ Department of Psychology, UCLA
}

\author{
Mimi Liljeholm (mlil@caltech.edu) \\ Computation and Neural Systems Program, California Institute of Technology
}

\author{
Catherine M. Sandhofer (sandof@psych.ucla.edu) \\ Department of Psychology, UCLA
}

\begin{abstract}
Logical consistency and objectivity are cornerstones of science that distinguish it from cult and dogma. Scientists' concern with objectivity has led to the dominance of associative statistics, which define the basic concept of independence on observations. The same concern with avoiding subjective beliefs has led many scientific journals to favor frequentist over Bayesian statistics. Our analysis here reveals that to infer causes of a binary outcome, (1) the associative definition of independence results in a logical inconsistency -- even for data from an ideal experiment -- for both frequentist and Bayesian statistics, and (2) removing the logical error requires defining independence on counterfactual causal events. The logically coherent causal definition is the one intuitively adopted by humans. Our findings have direct implications for more consistent and generalizable causal discoveries in medicine and other sciences.
\end{abstract}

Keywords: Causal inference, rationality, cognition, statistics.

\section*{Introduction}

Whenever we humans or other animals apply causal knowledge to achieve a desired outcome, we implicitly assume that the future resembles the past. Without the assumption that the course of nature remains invariant, all experience becomes useless (Hume, 1739). But what is the course of nature if not change (e.g., seeds sprout, species evolve, oceans warm, stars implode)? What we assume to remain invariant in nature are -- instead of events -- the forces of change, namely, causation (Kant, 1781; Kitcher, 1995). The fact that we routinely base actions on our causal knowledge (e.g., I strike this match because I expect it to ignite) is indubitable evidence that we hold the causal invariance assumption across the learning and application contexts. The present paper examines a previously unsuspected role that this assumption should play in scientific causal inference, leading to implications for more rational evaluations of hypotheses regarding causes of a binary outcome (e.g., a student graduating or not, an organism being dead or alive).

To test causal hypotheses based on data from experiments or quasi-experiments, the statistics in typical scientific use define invariance (often termed "independence" or "no interaction") on observations (Fienberg, 1980/2007; Jaynes, 2003; Wickens, 1989). Objectivity would seem to dictate this definition, given that the input necessarily consists of
observations only. Because causal relations are inferred and inherently unobservable (Hume, 1739), defining invariance on causal relations seems objectionable.

Thus, for the respective purposes of scientific causal discovery and of justifying the application of causal knowledge, there are two distinct definitions of invariance: the associative and the causal. The associative conception defined on observable events traces its inspiration to the philosophical works of Hume (1739), who questioned the grounds for our compelling belief that causation exists in the world. The causal conception defined on causal influences rests on Kant's (1781) argument that an ontological commitment to causation is essential for a coherent interpretation of the world. We use "causal influence" here in the sense of capacity or power, which when realized explains the occurrence of the outcome.

There is a discrepancy between the two conceptions, but the discrepancy has not seemed problematic: The unspoken consensus is that while causal invariance justifies generalization, it plays no role in causal discovery. Accordingly, using associative statistics to test experimental data is standard practice, and is viewed as appropriate as long as the experimental manipulation, which disambiguates causal direction, succeeds in eliminating confounding.

The consensus opinion, however, is mistaken. Here we show that even in the ideal case in which there are no confounding variables, the definition of invariance incorporated in a measure can affect the statistical output. Moreover, with regard to causes of a binary outcome, a type of outcome prevalent in medical and business research (e.g., a tumor cell being malignant or not, a consumer buying an item or not), only the definition based on counterfactual causal events, the Kantian causal power definition, is logically consistent. Notably, the coherent definition is the seemingly less objective one.

To explain the inferential problem, we step back and examine the nature and definitions of causal invariance from a cognitive-science perspective, in particular, within the broader issue of how an intelligent agent with access to inherently limited information can construct a representation of the world that best enables desired outcomes. From this perspective we examine the implications of conceptions of causal invariance for the experimental sciences and everyday causal inference.

\section*{Causal Invariance and its Implications}

Under the premise that all changes are caused, one way of stating causal invariance is: a cause \(c\) of an effect \(e\) retains the same capacity to affect \(e\) regardless of the temporal or spatial context, in which alternative (often unobserved) causes of \(e\) may occur with different probabilities. That is, the causal power of \(c\) is independent of the occurrence of alternative causes of \(e\), as if those alternative causes were not there. A change in the capacity of a cause to produce its effect is an indication of the causal mechanism operating differently. As we show later, this interpretation of causal invariance applying the concept of independence (i.e., "no interaction") to causal powers enables logical implications of the assumption to emerge, by enabling a mathematical definition of causal invariance (Eq. 4). (Causal invariance is the simpler of two conceptions that are equivalent with respect to our conclusions; the other conception is that although \(c\) interacts with enabling conditions in the background, the enabling conditions occur just as frequently in the learning and application contexts [Cheng, 2000].)

The concept of causal invariance serves two distinct functions. First, as a working hypothesis, a defeasible default assumption, it justifies causal generalization and prediction. By rendering inference compositional, it enables the generation of logically consistent answers to an unlimited variety of questions regarding an outcome's occurrence in an unlimited range of application contexts (Cheng et al., 2007). Second, as a definition of what it means for the nature of a cause to remain the same (rather than as a description of a particular causal mechanism), causal invariance serves as a criterion for hypothesis revision. Thus, if a generalization proves wrong, as would often happen in the dynamic mental construction of our complex causal world, the deviation from expectation signals a need to better capture invariance. In this second role, causal invariance is a navigation device that orients hypothesis testing towards its goal of formulating the simplest explanation of a phenomenon that allows invariance to obtain (Carroll \& Cheng, 2010).

Consider the alternative, the non-uniformity of nature, as the default. Not only would predictions and applications be impossible, so would hypothesis revision -- given no expectation, there is no deviation from expectation to guide revision towards causal invariance. Thus, the choice is a) inapplicable and stagnant causal knowledge or b) risky causal inference with the potential for effective generalization and hypothesis revision. In its two roles, as a default and a criterion for revision, causal invariance embodies the conviction that the world is knowable, that one can tease things apart, comprehend them, and mentally recompose them at will.

\section*{Defining Causal Invariance: Hume versus Kant}

Assuming causal invariance requires two leaps of faith. The first is apparent: faith that the future resembles the past. The second is subtler: faith in the existence of causation, a faith Hume (1739) resisted. Here we show why the second
leap plays a central role in rational causal discovery, in particular, why an associative definition of invariance, omitting this leap, is irrational for causal discovery.

We classify models as causal or associative depending on whether or not they have a definition of independence on causal influences. Whereas the ontological commitment to the existence of causation under the causal view enables this view to define independence on causal influences (e.g., Cartwright, 1989; Cheng, 1997; Lu, Yuille, Liljeholm, Cheng \& Holyoak, 2008; Pearl, 2000; Sheps, 1958; Sloman, 2005; Yuille \& Lu, 2008), the lack of this a priori assumption confines the associative view to defining independence on observations only (e.g., the cross product ratio; Fienberg, 1980/2007; Wickens, 1989). These two views differ most clearly for causes and effects that are represented by binary variables with a "present" value and an "absent" value; our argument therefore uses this variable type. For this variable type, observable events consist of the values of candidate cause \(c\) and of effect \(e\). We denote the "present" and "absent" values by " 1 " and " 0 " respectively.

The Associative View The associative view defines independence on observations of \(c\) and \(e\) (we use \(c\) and \(e\) as variables or values depending on context): if \(c\) occurs independently of \(e\), then
\[
\mathrm{P}(c=1, e=1)=\mathrm{P}(c=1) \cdot \mathrm{P}(e=1)
\]
where \(\mathrm{P}(c=1, e=1)\) is the probability of the joint occurrence of \(c\) and \(e\). This view computes associations, and leaves causal inference to a separate and subsequent interpretation of the associative output, for example, according to scientists' principles of experimental design or as Hume's "habit of mind". To enable predictions, this view typically amends Eq. 1 with additional assumptions, often variations of linearity or additivity. This amendment implicitly extends the definition of independence; deviation from linearity is what signals the need for interaction terms.

We illustrate the linear combination of associative strengths with the \(\Delta P\) model (Jenkins \& Ward, 1965; Salmon, 1965),
\[
\Delta \mathrm{P}=\mathrm{P}(e=1 \mid c=1)-\mathrm{P}(e=1 \mid c=0) \quad \text { (Eq. 2) }
\]
where \(\mathrm{P}(e=1 \mid c=1)\) and \(\mathrm{P}(e=1 \mid c=0)\), respectively, denote the probability that \(e\) occurs given that \(c\) occurs and given that \(c\) does not occur. Eq. 1 is a special case of Eq. 2, the case in which \(\Delta \mathrm{P}=0\). To tease apart the influence of \(c\) from all other influences on \(e\), our analysis partitions all direct causes of \(e\) into \(c\) and \(a\), where \(a\) represents a composite of alternative causes of \(e\) in the context. When \(c\) is absent, the occurrence of \(e\) is explained by \(a\). Let \(w_{\mathrm{c}}\) represent the weight (i.e., strength) of the association between \(c\) and \(e\), and \(w_{\mathrm{a}}\) represent that between \(a\) and \(e . \Delta \mathrm{P}\) has been shown to be a maximum-likelihood estimator of \(w_{c}\) in the Bayesian framework (Griffiths \& Tenenbaum, 2009; Tenenbaum \& Griffiths, 2001).

When there is no confounding (i.e., \(a\) occurs just as often whether or not \(c\) occurs), \(\Delta \mathrm{P}\) estimates \(w_{\mathrm{c}}\). Thus, replacing \(\Delta \mathrm{P}\) with \(w_{\mathrm{c}}\) and \(\mathrm{P}(e=1 \mid c=0)\) with \(w_{\mathrm{a}}\), Eq. 2 can be rewritten and rearranged to give the linear equation:
\[
\mathrm{P}(e=1 \mid c=1)=w_{\mathrm{c}}+w_{\mathrm{a}}
\]
(Eq. 3).
That is, when multiple causes are present, the occurrence of \(e\) according to this model is explained by a sum of the associative strengths of the causes. Bayesian structurelearning models can likewise adopt the linear definition ( Lu et al., 2008; Yuille \& Lu, 2008; Tenenbaum \& Griffiths, 2001).

Similarly, generalized linear models (GLMs [Fienberg, 2007; McCullagh \& Nelder, 1989]), some process models in psychology (e.g., Rescorla \& Wagner, 1972), and prominent causal models in epidemiology (Rothman et al., 2008) also adopt the definition in Eq. 1 amended with variants of linearity. For example, logistic regression, likely the most commonly used model for evaluating causal hypotheses in medical research and widely used in business research as well, amends Eq. 1 with a logistic scale transformation to better justify the linearity. A feature common across the generalizations in GLMs is "the presence in all the models of a linear predictor based on a linear combination of explanatory or stimulus variables" (McCullagh \& Nelder, 1989, p. xvi).

Now, consider a situation in which representation in terms of observable events alone cannot capture the constancy of a causal relation across contexts. When effect \(e\) is binary, a factor's capacity to influence \(e\) may have no observable manifestations, even when there is no confounding. Suppose \(c\) is a cause of \(e\) that does not interact with any other cause of \(e\). Yet, whenever \(e\) is already present (regardless of which other cause produced it), introducing \(c\) will yield no change in the state of \(e\), indistinguishable from introducing a noncausal factor. For example, suppose someone is already dead (the binary outcome in question) from being hit by a car. Being hit by another car will show no change in the outcome (the person is still dead), despite the sameness of the forces underlying car accidents (the second car would have killed the person too). In such occlusion events, unobservable causal capacities lose their mapping onto observable changes. Given the lack of constancy in this mapping, postulating capacities becomes crucial for representing a stable causal world; observable changes, as used in associative models, or even actual causation in an episode, as used in epidemiological causal models (Rothman, Greenland \& Lash, 2008), would be inadequate. Just as objects occluded in the 2-d visual input on our retinas are assumed to continue to exist in the world, so should occluded causal capacities.
The Causal View The causal view builds on Hume's insight - that causal knowledge is induced from noncausal data - but goes beyond it: Intervening between the observable input and the causal output is a causal explanation of the data. This explanation, under Kant's domain-general a priori causal framework, posits the existence of such things as causal relations: theoretical events that yield observed phenomena. We denote "causing" by " \(\rightarrow\) " (e.g., "c \(\rightarrow \mathrm{e}\) " denotes " \(c\) causing \(e\) "). Once causal events are assumed to exist, the definition of their independence analytically follows:
if \(c \rightarrow e\) is independent of \(a \rightarrow e\), then
\[
\mathrm{P}(c \rightarrow e, a \rightarrow e)=\mathrm{P}(c \rightarrow e) \bullet \mathrm{P}(a \rightarrow e)
\] \(\mathrm{P}(c \rightarrow e)\) is the probability of \(c\) causing \(e\); it corresponds to the theoretical probability that \(e\) would occur if \(c\) is present but no other (observed or unobserved, generative or preventive) cause of \(e\) were present. The probability is theoretical because it is impossible to know that a context has no unobserved causes. Note that \(\mathrm{P}(\mathrm{c} \rightarrow \mathrm{e})\) is not a conditional probability involving two random variables, but instead the probability associated with a single random variable. Likewise, \(\mathrm{P}(a \rightarrow e)\) is the probability of \(a\) causing \(e\), and \(\mathrm{P}(c \rightarrow e, a \rightarrow e)\) is the probability of one of the two causes, \(c\) or \(a\), producing \(e\) and the other cause also producing \(e\) if \(e\) had not been already produced. ("No interaction" between the occurrences of \(c\) and \(e\), as defined in Eq.1, is a special case of the independence of causal powers as defined in Eq. 4 when there is no confounding and \(\Delta \mathrm{P}=0\).)
Notice that the definition in Eq. 4 centers on conjunctive causation in an "occlusion" event. The conjunctive causal event (e.g., a dead car-accident victim being killed by a second car) can never occur (rather than happen to not have occurred). Our " \(\rightarrow\) " notation serves as a reminder that the causal events denoted are nonexistent and theoretical.

Although none of the events in Eq. 4 is observable, the intervening causal explanation of the data (e.g., when \(e\) occurred in the presence of \(c\), it occurred because \(c\) caused it or \(a\) caused it) maps observable event frequencies (e.g., how often \(e\) occurred when \(c\) was present) onto their theoretical causal probabilities [e.g., \(\mathrm{P}(c \rightarrow e\) OR \(a \rightarrow e)\) ]. Thus, \(\mathrm{P}(e=1 \mid c=1)\) estimates \(\mathrm{P}(c \rightarrow e\) OR \(a \rightarrow e)\). The latter in turn can be expressed in terms of the constituent events in Eq. 4:
\(\mathrm{P}(c \rightarrow e\) OR \(a \rightarrow e)=\mathrm{P}(c \rightarrow e)+\mathrm{P}(a \rightarrow e)-\mathrm{P}(c \rightarrow e, a \rightarrow e)\) (Eq. 5), where the final term equals the product, \(\mathrm{P}(c \rightarrow e) \cdot \mathrm{P}(a \rightarrow e)\), if \(c\) and \(a\) produce \(e\) independently (Eq. 4).

Under this view, causal interpretation is integral to the computation of the numerical output (e.g., Cheng, 1997; Griffiths \& Tenenbaum, 2009), rather than subsequent to it. Data analysis incorporates causal invariance.

The difference between the two views and its implications for rational scientific causal inference has not received attention. Like frequentist statistics for the experimental sciences, causal Bayes nets adopt the separation of statistics and causal inference. The "generic" parameterization most commonly adopted in causal Bayes nets uses neither the associative nor the causal definition, and the "noisy OR" parameterization in Eq. 5 is used for efficiency rather than rationality. In a similar vein, Bayesian causal models allow both the associative and causal definitions (Griffiths \& Tenenbaum, 2009; Lu et al., 2008; Yuille \& Lu, 2008).

The Rationality of the Two Views Is it rational to define causal invariance on unobservable, imaginary events, as the causal view does? Ceteris paribus, it is objectionable to use unobservable events. What is at stake, however, is logical consistency. What it means for the nature of a cause in our physical world to remain invariant across contexts is nonarbitrary. There is only one way for a causal mechanism in
a coherent world to operate the same way, without change. For binary causes and effects that are "present" or "absent," Eqs. 4 and 5 specify the only logically consistent definition of causal invariance (e.g., so that \(c\) causes \(e\) with indeed the same probability in one context as in another). In other words, systematic deviation from independence as specified in these equations indicates causal interaction. (Note that for other variable types and combinations of variable types, the singular meaning of causal invariance in the world is captured by other mathematical functions.)

We first explain the correlated influences inherent in associative amendments by illustrating how the linear model in Eq. 3 deviates from causal invariance. The additivity in Eq. 3 holds only if the capacities of \(c\) and of \(a\) to cause \(e\) are mutually exclusive [i.e., \(\mathrm{P}(c \rightarrow e, a \rightarrow e)=0\); there are no occlusion events]. But, to define independence as mutual exclusivity (i.e., to define "no correlation" as a negative correlation) is self-contradictory.

To see the self-contradiction without the abstraction of causal inference, consider a simple concrete example involving two events regarding a deck of playing cards: drawing a diamond and drawing a face card. (Assume that the deck has diamonds and face cards, among other cards.) Defining independence between the two events as mutual exclusivity of the events would entail asserting that the chance of drawing a face card is the same for diamonds as for other suits if and only if face cards and diamonds are mutually exclusive: when there are no face cards that are diamonds. The chance of drawing a face card then would be 0 for diamonds but not for other suits. The mutualexclusivity definition therefore implies a logical contradiction: "the chance of drawing a face card is the same across suits only if it is not the same across suits."

Our analysis so far may seem irrelevant to current frequentist statistics: nonlinear GLMs, which avoid a logical shortcoming of linear models for analyzing data with binary outcome variables, have long replaced linear models for that purpose (Fienberg, 1980/2007; Wickens, 1989). But, GLMs in fact do not sidestep the contradiction in other associative models. First, GLMs concur with the \(\Delta \mathrm{P}\) model in adopting the mutual-exclusivity definition for special cases involving data that have the feature of symmetry. We illustrate this agreement presently with a logistic-regression analysis of fictitious data in a story in an experiment designed for preschool children. Second, GLMs more generally carry the broader contradiction of defining independence as interaction. Because \(\mathrm{P}(e=1 \mid c=1)\) estimates \(\mathrm{P}(c \rightarrow e\) OR \(a \rightarrow e)\), Eqs. 3 and 5 can be directly compared. They differ by the final (negative) term in Eq. 5 being omitted in Eq. 3. A scale transformation that would avert the contradiction would therefore need to result in subtracting the product, \(w_{\mathrm{c}}{ }^{\bullet} w_{\mathrm{a}}\), from the right-hand-side of Eq. 3. But this is neither the intent nor the result of the transformations in GLMs. The logistic function, for example, is symmetric (see s-shaped curve in Figure 1), as is characteristic of associative models. In contrast, for every value of \(w_{\mathrm{a}}\), subtracting \(w_{\mathrm{c}} \bullet w_{\mathrm{a}}\) from the sum, \(w_{\mathrm{c}}+w_{\mathrm{a}}\), yields an asymmetric concave function of \(w_{\mathrm{c}}\) (as
\(w_{\mathrm{c}}\) increases, an increasing amount is subtracted from the linear sum).

Without the a priori postulate that causal relations exist, associative models cannot coherently define independence on the missing relations, hence cannot justify the application of causal knowledge. They cannot, even when ideal experiments are concerned, because the error is logical.
An Illustration of the Associative and Causal Views Arriving at Opposite Conclusions We return to the mutual-exclusivity definition of causal invariance in associative statistics. In a story presented to preschoolers in our experiment, two brothers -- a farmer and a zookeeper try to figure out what prevents red dots from appearing on the faces of animals at their farm and at zoo. The candidate preventive causes of red dots are two treats: a grain and a type of leaves. At the farm, the brothers gave the grain to all 10 animals there: 9 of them had red dots before eating the grain, and 6 did so afterwards. At the zoo, the brothers gave both treats -- grain and leaves -- to all 10 animals there: 4 of them had red dots before eating the two treats, and only one had red dots afterwards. The question is: which treat is one's best bet for removing red dots from the faces of farm and zoo animals?

Regardless of how "sameness" is defined, the rationale underlying the choice is: Assuming the grain operates "the same way" across contexts (i.e., farm and zoo), if the influence of the intervention (grain at farm vs. both treats at zoo) remains invariant across contexts, one's best guess would be that leaves had no effect - grain alone would already explain the outcome. But, if the influence of the intervention varies across contexts, one would attribute the difference to leaves.

According to the causal view, the grain operating with the same causal mechanism across contexts implies that for every animal (all 20), grain has the same causal power to remove red dots. We denote the two interventions by "farm_iv" and "zoo_iv" respectively and "red dots on the face" by "red" in the calculations below. The causal power of candidate cause \(c\) to prevent effect \(e, p_{c}\), is estimated according to (Cheng, 1997):
\[
\begin{equation*}
p_{c}=\frac{P(e=1 \mid c=0)-P(e=1 \mid c=1)}{P(e=1 \mid c=0)} \tag{Eq.6}
\end{equation*}
\]

Thus,
\[
\begin{equation*}
p_{\text {farm }-i v}=p_{\text {grain }}=\frac{9 / 10-6 / 10}{9 / 10}=1 / 3 \tag{Eq.7}
\end{equation*}
\]

Likewise,
\[
\begin{equation*}
p_{z o o_{i} i v}=\frac{4 / 10-1 / 10}{4 / 10}=3 / 4 \tag{Eq.8}
\end{equation*}
\]

But, according to causal invariance (Eqs. 4 and 5),
\[
\begin{equation*}
p_{\text {zoo_iv }}=p_{\text {grain }}+p_{\text {leaves }}-p_{\text {grain }} \cdot p_{\text {leaves }} \tag{Eq.9}
\end{equation*}
\]

It follows that
\(3 / 4=1 / 3+p_{\text {leaves }}-1 / 3 \cdot p_{\text {leaves }}\)
Therefore, \(p_{\text {leaves }}=5 / 8\). Because \(5 / 8\) is greater than \(1 / 3\) (i.e., the leaves treat is a stronger cure than grain), the causal view prescribes choosing leaves.

Associative models, whether Bayesian or frequentist, all reach the opposite conclusion, prescribing grain instead. The mutual-exclusivity definition implies that the set of animals with "no red dots" due to grain, 3 out of 10 animals, has no overlap with the set due to the contextual cause at each place: grain should therefore heal 3 animals both at the farm and at the zoo. Because 3 animals indeed had their red dots "go away" at each place, leaves must have no effect. The \(\Delta \mathrm{P}\) model therefore prescribes grain.

Logistic regression is a GLM used for predicting the probability of the occurrence of a dichotomous outcome (e.g., red_dots vs no red_dots) by fitting data to a logistic function of a linear combination of input variables (e.g., grain, leaves, background causes at the farm and at the zoo). For the farm-and-zoo scenario (see Figure 1), because the pattern of events is symmetrical around the probability of .5 , the same reduction in P (red_dots) occurs at the farm and at the zoo (see vertical dashed lines) at symmetrical segments of the logistic curve. Therefore, the grain (see heavy horizontal dashed lines) -- which explains the reduction in the probability of animals with red dots at the farm explains the entire reduction at the zoo as well. That is, logistic regression detects no influence at all from leaves, either by itself or in an interaction, concurring with the \(\Delta P\) model. Increasing sample size does not change this conclusion.


Figure 1. A schematic explanation of the probability of the outcome according to logistic regression: the probability of having red dots at the farm and at the zoo, before and after the respective interventions in the scenario, as a logistic function of the weighted sum of the four predictor variables.

Preschoolers in our experiment chose leaves, in accordance with the causal view. Recall that the causal view avoids the incoherence of the associative view by defining causal invariance on counterfactual causal events. Note that the causal explanations involve no prior domain-specific knowledge; the causal-invariance assumption is domaingeneral and the input consists of data alone. This view thereby achieves objectivity without sacrificing coherence.

If the world happens to be causal, then a leap of faith to assume unobservable causal capacities would be adaptive,
by enabling a coherent definition of causal invariance in our representation. Coherence is essential because there are infinitely many possible representations of the world based on available observations, only some of which support generalization to new contexts. Reasoners use logical consistency and, more generally, parsimony of the represented explanations to prune the vast search space and efficiently converge on truth, if truth exists (Kelly, 2007). Causal discovery should therefore require general-purpose Sherlock Holmeses, who make use of coherence to infer how things work.

\section*{Discussion}

In summary, noting a simple logical consequence of Kant's a priori assumption of causation for rational causal inference, we have shown that -- contrary to the unspoken consensus among scientists -- the causal invariance assumption critically affects causal discovery. To evaluate a causal relation involving a binary outcome variable that is "present" or "absent", only invariance defined on causal capacities is logically consistent and supports generalization to new contexts. Thus, associative statistics, for which invariance is only defined on observations, may arrive at a fallacious conclusion even when applied to data from a perfect experiment.

The potential for the associative and causal views to arrive at opposing recommendations has obvious implications. For example, a critical linear-regression analysis in the influential Seven Countries Study (Keys, 1980), a large longitudinal study on how diet affects coronary heart disease and other health outcomes, shows that controlling for saturated fat, consumption of sugar is unrelated to death (a binary outcome). Medical and publichealth dietary advice in the U.S. based on this and other analyses in the study (Keys et al., 1984; Menotti et al., 2003), using linear models as was common practice, has created a food industry that produces low-fat but high-sugar foods (e.g., fat-free salad dressings with added sugar to compensate for taste). More generally, these associative analyses formed the foundation for three decades of dietary advice to adhere to a low-fat diet, without special attention to sugar intake (as distinct from caloric content). There is currently no causal analogue of logistic regression, which allows predictor variables that are continuous (e.g., consumption of sugar) as well as discrete. As we have shown for binary outcome variables, coherent causal generalization requires a causal framework, and applying causal instead of associative statistics to evaluate the influences of fat and sugar intake could potentially reverse estimates of the magnitude of their harm or change their assessed causal status. The researchers could have found that consumption of sugar causes coronary heart disease, diabetes, cancer and other diseases constituting the metabolic syndrome, as recent evidence indeed suggests. A more rational statistical approach could have profoundly altered the course of the obesity epidemic in the U.S. and worldwide.

Note that one interpretation of associative models that would remove the incoherence we noted is to posit a mediating continuous variable and to assume that the causes operate independently on this continuous variable rather than on the observed binary outcome variable. The linear definition of causal invariance holds for continuous outcome variables, thereby removing the incoherence. Regardless of the plausibility of the revised hypothesis with the mediating variable, note that it is deviation from the criterion of causal invariance that signals the need to revise the simple hypothesis (Carroll \& Cheng, 2010), the tenacious goal being to achieve causal invariance.

\section*{Acknowledgments}

The research reported in this article was supported by AFOSR FA 9550-08-1-0489. We thank Chris Carroll, Tom Wickens, and Ying Nian Wu for helpful discussion.

\section*{References}

Carroll CD, Cheng PW (2010) The induction of hidden causes: Causal mediation and violations of independent causal influence. In Proceedings of the 32nd Annual Conference of the Cognitive Science Society, eds Ohlsson S, Catrambone R (Austin, Texas: Cognitive Science Society), pp. 913-918.
Cartwright N (1989) Nature's capacities and their measurement (Clarendon Press, Oxford).
Cheng PW (1997) From covariation to causation: A causal power theory. Psychological Review, 104: 367-405.
Cheng PW (2000) Causality in the mind: Estimating contextual and conjunctive causal power. In Explanation and Cognition, eds Keil F, Wilson R. (MIT Press, Cambridge), pp. 227-253.
Cheng PW, Novick LR, Liljeholm M, Ford C (2007) Explaining four psychological asymmetries in causal reasoning: Implications of causal assumptions for coherence. In Topics in contemporary philosophy, Volume 4: Causation and explanation, ed O'Rourke M (MIT Press, Cambridge), pp. 1-32.
Fienberg SE (1980/2007) The analysis of cross-classified categorical data (2nd edition (MIT Press, Cambridge; Spring, New York).
Griffiths TL, Tenenbaum JB (2009) Theory-based causal induction. Psychological Review, 116 (4): 661-716.
Hume D (1739/1987) A treatise of human nature (2nd edition, Clarendon Press, Oxford).
Jaynes ET (2003) Probability theory: The logic of science. (Cambridge Univ Press, Cambridge).
Jenkins HM, Ward WC (1965) Judgment of contingency between responses and outcomes, Psychological Monographs: General and Applied, 79 (1, Whole No. 594).

Kant I (1781/1965) Critique of pure reason, trans Smith NK (Macmillan, London).
Kelly K (2007) Ockham's razor, empirical complexity, and truth-finding efficiency, Theoretical Computer Science, 383: 270-289.

Keys A (ed, 1980) Seven countries: a multivariate analysis of death and coronary heart disease (Harvard, Cambridge).
Keys A, Menotti A, Aravanis C, Blackburn H, Djordevic BS, Buzina R, Dontas AS, Fidanza F, Karvonen MJ, Kimura N, Mohacek I, Nedeljkovic S, Puddu V, Punsar S, Taylor HL, Conti S, Kromhout D \& Toshima H (1984) The Seven Countries Study: 2,289 Deaths in 15 Years, Preventive Medicine, 13: 141-154.
Kitcher \(P\) (1995) Revisiting Kant's epistemology: Skepticism, apriority, and psychologism. Noûs, 29: 285315.

Lu H, Yuille A, Liljeholm M, Cheng PW, Holyoak, KJ (2008) Bayesian generic priors for causal learning, Psychological Review, 115: 955-984.
Lustig RH (2012) The toxic truth about sugar. Nature, 482: 27-29.
McCullagh P, Nelder JA (1989). Generalized linear models (2 \({ }^{\text {nd }}\) edition, CRC Press, Boca Raton).
Menotti A, Pudd, PE, Lanti M, Kromhout D, Blackburn H \& Nissinen A (2003) Twenty-five-year coronary mortality trends in the seven countries study using the accelerated failure time model. European Journal of Epidemiology, 18: 113-122.
Pearl J (2000) Causality: Models, reasoning, and inference (Cambridge Univ Press, Cambridge).
Rescorla RA, Wagner AR (1972) A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Classical conditioning II: Current theory and research, eds Black AH, Prokasy WF (Appleton-Century Crofts, New York), pp. 64-99.
Rothman KJ, Greenland S, Lash TL (2008) Modern epidemiology (3rd edition, Lippincott, Williams \& Wilkins, Philadelphia).
Salmon WC (1965) The status of prior probabililities in statistical explanation. Philosophy of Science, 32:137-146.
Sheps MC (1958) Shall we count the living or the dead? The New England Journal of Medicine, 259: 1210-1214.
Sloman S (2005) Causal models: How we think about the world and its alternatives. (Oxford Univ Press, New York).
Spirtes P, Glymour C, Scheines R (1993/2000) Causation, prediction and search (2nd edition, MIT Press, Cambridge).
Tenenbaum JB, Griffiths TL (2001) Structure learning in human causal induction. In TK Leen, TG Dietterich, \& V Tresp (eds), Advances in neural processing systems, 13: 59-65 (MIT Press, Cambridge).
Wickens TD (1989) Multiway contingency tables analysis for the social sciences (Erlbaum Associates, Hillsdale).
Yuille A, Lu H (2008) The noisy-logical distribution and its application to causal inference. In Advances in neural processing systems, vol. 20 (MIT Press, Cambridge).

\title{
Inferring One's Own Prosociality Through Choice: Giving Preschoolers Costly Prosocial Choices Increases Subsequent Sharing Behavior
}

\author{
Nadia Chernyak (nc98@cornell.edu) \& Tamar Kushnir (tk397@cornell.edu) \\ Department of Human Development, Martha Van Rensselaer Hall \\ Ithaca, NY 14853 USA
}

\begin{abstract}
Prosociality emerges early in ontogeny, but the mechanisms driving its early-emergence are not well understood. We propose that the experience of choice is tied to the expression of children's prosocial behavior. In Experiment 1, preschoolers shared with a puppet by either making a Costly Choice (giving a resource they could have kept for themselves), Non-Costly Choice (giving a resource that would otherwise be thrown away), or No Choice. Subsequent prosociality was measured by allowing children to share with a new puppet. While most children shared initially, children who were given costly choices shared more with the new puppet. Experiment 2 replicated this result using a different manipulation for Costly vs. Non-Costly choices. Experiment 3 found that preschoolers were more likely to infer that actions are intentional when they are costly. Results suggest a prosocial construal hypothesis: that children rationally infer their prosociality through making difficult, autonomous choices.
\end{abstract}

Keywords: cognitive development; choice; altruism; preschoolers

\section*{Introduction}

People very rapidly acquire remarkable prosocial tendencies. By the second to third year of life, children help others complete their goals (Warneken \& Tomasello, 2006), share toys (Schmidt \& Sommerville, 2008; Svetlova, Nichols, \& Brownell, 2010), sympathize with those who are harmed (Vaish, Carpenter, \& Tomasello, 2009) or are in distress (Zahn-Waxler, Radke-Yarrow, \& Wagner, 1992), and punish those who harm others (Dunfield \& Kuhlmeier, 2010; Vaish, Carpenter, \& Tomasello, 2010; Vaish, Missana, \& Tomasello, 2011). But how children acquire such tendencies remains an understudied empirical question. Here, we explore the possibility that having and making choices encourages young children's prosocial behavior.

One potential mechanism for the expression of prosocial behavior is through past experience with prosocial action (Staub, 1971). Self-perception theory (see Beaman, Cole, Preston, Klenty, \& Steblay, 1983; Bem, 1967; Eisenberg, Cialdini, McCreath, \& Shell, 1987; Lepper, 1973) suggests that individuals are likely to act in congruence with their past actions because of a desire to stay self-consistent. Thus, through acting prosocially, children may be forming a cognitive representation of what "the self" is like, and acting in accordance with that representation (Freedman \& Fraser, 1966; Grusec, Kuczynski, Rushton, \& Simutis 1978; Grusec \& Redler, 1980).

Importantly, however, children evaluate their own actions not simply by their occurrence, but also by the contexts under which they occur (Warneken \& Tomasello, 2008). Here we explore one important context critical to evaluating prosocial behavior: that of choice. Choice differs from action in that it involves the contrast between actions performed and alternative actions not performed. For example, I evaluate Bob, who gave \(\$ 5\) to charity but could have kept it for himself (had an alternative) more positively than Jim, who accidentally dropped \(\$ 5\) into the hands of a homeless person (had no alternative). In fact, we often go beyond evaluating choice in absolute terms (having vs. not having choice) to also consider degree of costliness of the alternatives. To extend the above example, I would consider Bob more generous if his choice was to give away his last \(\$ 5\) than if his choice was to give away \(\$ 5\) out of his last \(\$ 100\). Thus, both the presence and the costliness of choice influence how we evaluate others.

No study to our knowledge has addressed whether choice plays a causal role in young children's own prosocial behavior. In this work, we asked whether making costly choices increases young children's prosociality. Specifically, we hypothesized that the contrast between actions chosen and alternative actions not chosen influences children's later prosocial behavior above and beyond the prosociality of the actions themselves.

We allowed preschool-aged children (3-4 year-olds) to perform a prosocial action: allocating a limited and desired resource to a puppet who was feeling sad. We systematically manipulated the presence and magnitude of alternative actions (non-prosocial actions) that children could also undertake. We were interested in how the presence and valence of these alternative actions affected children's subsequent prosociality. Subsequent prosociality was measured by allowing children to then make a new prosocial action towards a different puppet.

\section*{Experiment 1}

In Experiment 1, children were presented with an attractive and limited resource: 1 star sticker that they could give to a puppet ("Doggie") who was described as feeling sad. We manipulated children's experience of choice by allowing children to either make a Costly Choice (give the sticker to Doggie instead of keeping it for themselves), Non-Costly Choice (give the sticker to Doggie instead of having the experimenter put the sticker away), or No Choice (instructed to give the sticker to Doggie). As such, all children were given the option between (a) a positive
prosocial action \((+)\), and (b) either a selfish, negative action \((-)\), a neutral action (0), or no action. See table 1 for a summary. Children's actions towards Doggie were recorded. We were then interested in how the contrast between the action chosen (action a) and the action unchosen (the alternative action b) affected children's subsequent prosociality. To measure subsequent prosociality, children were introduced to a new puppet ("Ellie") who was also feeling sad, and given three stickers that they could either keep or share with Ellie.

\section*{Table 1: Summary of Experiment 1}
\begin{tabular}{lll}
\hline Condition & Target Action & \begin{tabular}{l} 
Alternative \\
Action
\end{tabular} \\
\hline Costly Choice & \begin{tabular}{l} 
Give sticker to \\
Doggie \((+)\)
\end{tabular} & \begin{tabular}{l} 
Keep sticker for \\
self \((-)\)
\end{tabular} \\
Non-Costly Choice & \begin{tabular}{l} 
Give sticker to \\
Doggie \((+)\) \\
No Choice
\end{tabular} & \begin{tabular}{l} 
Throw sticker \\
Give sticker to \\
Doggie \((+)\)
\end{tabular} \\
None
\end{tabular}

\section*{Participants}

Seventy-two preschool-aged children (mean: 3.96 years; range: 2.85-4.98) participated. Conditions were fully balanced for age and gender. There were no age differences between conditions, \(F(2,71)=.96, p=.39\). One child was replaced due to parental interference. Participants were tested at a local school or children's museum.

\section*{Materials and Procedure}

Materials were two plush puppets ("Doggie" and "Ellie"), three small wooden boxes: Doggie's box, Ellie's box (which had pictures on the tops and insides of Doggie and Ellie, respectively), and the child's box (no pictures), and a set of small star and smiley face stickers. A schematic of the materials and procedure is shown in Figure 1.

Introduction All children sat at a table facing the experimenter. Children were first shown a plush animal named "Doggie" and told that Doggie was feeling "very sad today". Doggie was then put away. One of the toy boxes was placed on the table and introduced as "Doggie's box."

Choice Manipulation All children were induced to act prosocially. However, we varied the presence and magnitude of the alternative option across conditions. In the Costly Choice Condition, children were presented with the choice of either keeping the sticker for themselves or giving it to Doggie. In the Non-Costly Choice condition, children were presented with the choice of putting the sticker away or giving it to Doggie. Finally, in the No Choice Condition, children's actions were restricted by experimenter instruction ("This star sticker, you have to put in the box for Doggie so that he feels better"). Across all conditions, once
children made their choices, the experimenter said "good job!" and put the box away.

Dependent Measure A new puppet was then shown ("Ellie") who was also feeling sad. Ellie was then put away, Ellie's box was presented along with a second (plain) box on the table, and three smiley-faced stickers placed between the two boxes. The positioning of the two boxes was counterbalanced across participants. The experimenter then said that the three stickers were for the child, but that Ellie also really liked them. The number 3 was chosen to force children to create an uneven distribution (either to prioritize themselves, or to prioritize Ellie).

After counting the stickers, the experimenter then said that the child could either keep all of the stickers for him/herself (and pointed to the plain box) or share some with Ellie (and put them in Ellie's box). Re-prompts were used if children left any stickers on the table ("and what do you want to do with this/that one?"), until a box was chosen for each sticker.

\section*{Results and Discussion}

We first analyzed children's initial prosocial responses: the majority of children chose the prosocial action over the non-prosocial alternative: \(19 / 24\) in the Costly Choice condition, \(23 / 24\) in the Non-Costly Choice condition, and \(23 / 24\) in the No Choice condition (all Binomial \(p\) 's \(<.01\) ).

Next, we analyzed children's prosocial actions subsequent to the choice manipulation (Figure 2). Almost all children gave at least one sticker and shared at least one sticker, confirming that children both liked stickers and were motivated to share. Children were thus divided into two response groups based on whether they distributed unequally in favor of themselves or Ellie: other-prioritizing (giving majority, 2, or 3 , stickers to Ellie), and selfprioritizing (giving the minority, 1 , or 0 , stickers to Ellie). \({ }^{1}\) See Table 2 for details on number of stickers given per condition. A higher proportion (16/24; 67\%) of children in the Costly Choice condition made an other-prioritizing response than those in the No Choice ( \(8 / 24 ; 33 \%\) ) condition, Fisher's exact test \(p<.05\) (see Figure 2), suggesting that having choice influenced children's subsequent sharing. The cost of the choice also affected sharing: a higher proportion of children who made the initial Costly Choice were more likely to be other-prioritizing than those who made the NonCostly Choice (7/24; 29\%), Fisher's exact test \(p<.01\). Making a non-costly choice did not increase subsequent sharing over being instructed to share, \(p>.15\).

The results of Experiment 1 thus provide initial evidence that having made a costly choice to perform a prosocial action increased children's later prosocial behaviors. Why might this be the case? One possibility is that, by

\footnotetext{
\({ }^{1}\) Results remain nearly identical when analyzing only the subset of children who made the initial prosocial choice. For a conservative estimate, we thus include the full set of children across all experiments.
}


Figure 1: Schematic of materials and procedure used in all experiments.
Choices (a) and (b) were presented verbally in the order shown.

Contrasting their chosen actions with non-prosocial alternatives, children inferred their prosociality. The above explanation is consistent with traditional self-perception theories (e.g.,Bem, 1967) which predict that people learn about their own preferences from observing their past actions. There are, however, at least two alternative explanations, also consistent with self-perception theory, which consider the actions but do not take into account whether the action was contrasted with alternatives. One possibility is that the initial costly choice may have led children to believe they had exhibited their dislike for the object ("I gave away the sticker so I must not like stickers). Another possibility is that the initial Costly Choice caused children simply to repeat the initial outcome of distributing more to another than to themselves.

\section*{Experiment 2}

Experiment 2 was designed to rule out these possibilities. Procedures mirrored Experiment 1's Costly Choice condition, with the following modifications. Children were once again introduced to the first puppet, Doggie. This time, however, in the Costly Choice condition, children were given a colorful rubber toy frog, rather than a star sticker. In the Non-Costly Choice condition, children were given a
small white piece of torn paper. All children were told they could choose to keep the object for themselves or give it to Doggie. The dependent measure (and the new puppet, Ellie) remained the same.

It is important to note that unlike in Experiment 1, the objects used were different between the choice manipulation (which involved either a frog or piece of paper) and the dependent measure (which again involved smiley face stickers). Thus, any increased tendencies to share stickers during the dependent measure phase could not be attributed to children's inferences about their preference (or lack thereof) for stickers. Additionally, the choice manipulation of both the Costly And Non-Costly choice conditions required children to undertake the same prosocial action of giving the object to Doggie instead of keeping it for themselves, controlling for the possibility that initial practice with giving away objects causes children to repeat the outcome of giving more to others than to themselves.

\section*{Participants}

Forty-eight preschool-aged children (mean: 3.91 years; range: 2.81-4.96) participated. Conditions were fully balanced for age and gender. There were no age differences


Figure 2: Proportion of children who allocated the majority of their stickers in the dependent measure phase (to Ellie) in Experiments 1 and 2. \({ }^{*} p<05, * * p<.01\).
between conditions, \(t(46)=.41, p=.69\). Four children were replaced due to either experimental error or prior participation. Participants were tested at a local school or children's museum.

\section*{Materials and Procedure}

Materials were the same as those used in Experiment 1, except a set of colorful toy frogs and plain torn pieces of paper were used during the introduction instead of smiley face stickers (see Figure 1).

The procedure also largely followed that of Experiment 1, with the following modifications. In the Costly Choice Condition, children were given an attractive object (a colorful toy frog), and told they could either keep it or give it to Doggie. In the Non-Costly Choice condition, children were given a small torn piece of paper and also told they could either keep it or give it to Doggie. The dependent measures remained the same.

\section*{Results and Discussion}

First, to confirm that giving away the toy frog was in fact a more costly choice than giving away the piece of paper, we showed the two objects (side of object counterbalanced) to an independent sample of age-matched children, and asked them which object they preferred more. Nineteen (of 20) confirmed they preferred the frog (Binomial \(p<.001\) ). Once again, the majority of children in both the Costly Choice (frog) condition (21/24) and the Non-Costly Choice
(paper) condition (24/24) chose the prosocial action over the non-prosocial alternative (Binomial \(p\) 's \(<.01\) ).

Table 2: Number of Stickers Donated Across Conditions (modal responses bolded and underlined)
\begin{tabular}{lllll}
\hline Condition & \multicolumn{3}{c}{\begin{tabular}{c} 
Number of Children Who \\
Made Each Allocation
\end{tabular}} \\
& 0 & 1 & 2 & 3 \\
\hline Exp 1: Costly Choice & 2 & 6 & \(\underline{\mathbf{1 1}}\) & 5 \\
Exp 1: Non-Costly Choice & 3 & \(\underline{\mathbf{1 4}}\) & 3 & 4 \\
Exp 1: No Choice & 3 & \(\underline{\mathbf{1 3}}\) & 4 & 4 \\
Exp 2: Costly Choice & 2 & 6 & \(\underline{\mathbf{1 3}}\) & 3 \\
Exp 2: Non-Costly Choice & 3 & \(\underline{\mathbf{1 3}}\) & 5 & 3 \\
\hline
\end{tabular}

A higher proportion of children in the Costly Choice (16/24; 67\%) condition performed other-prioritizing prosocial behaviors than those in the Non-Costly Choice (8/24; 33\%) condition, Fisher's exact test \(p<.05\), demonstrating once again, that costly choices led to greater subsequent sharing behaviors. Once again, for details on number of stickers given per condition, see Table 2.

Moreover, we confirmed that children's prosociality could not be explained by the child making inferences about their own lack of preference for stickers: sharing rates across the two Costly Choice conditions of Experiments 1
and 2 were identical. Moreover, we ruled out the possibility that children in Experiment 1 simply repeated the outcome of having fewer objects than another agent - children in both conditions of Experiment 2 initially shared an object with Doggie instead of keeping it for themselves.

\section*{Experiment 3}

The results of Experiments 1 and 2 provide initial evidence that making costly prosocial choices plays an important role in children's subsequent prosocial behavior. We suggest that our findings are best explained by a prosocial construal hypothesis (see Cialdini, Eisenberg, Shell, \& McCreath, 1987; Grusec et al., 1978): In making costly prosocial choices, children construe their actions as a signal of their prosociality (e.g., "I shared so I must like to share").

How children perceive costly vs. non-costly situations, however, remains an important question. In Experiment 3, we wished to more closely investigate the differing perceptions that might occur during costly vs. non-costly choice situation. One possibility is that children perceive their own costly choices as intentions (i.e., that their actions were in fact, intentionally and freely chosen, rather than obligatory). On this account, children would encode costly choice situations as choices, and non-costly choices as obligatory acts (e.g., "I chose to give the sticker to Doggie instead of keeping it for myself" vs. "I had to give the sticker to Doggie instead of throwing it out").

In Experiment 3, we tested for this possibility, by once again giving children either a Costly or Non-Costly Choice, and then asking them whether they chose to or had to perform the target action.

\section*{Participants}

Fifty preschool-aged children (mean: 3.37 years; range: 2.84-4.84) participated. There were no age differences between conditions, \(t(48)=98, p=.34\). Five children were replaced due to either experimental error, or because they refused to answer the question. Participants were tested at a local school or children's museum.

\section*{Materials and Procedure}

Materials were the same as those used in Experiment 1, except there was no new puppet (Ellie). See Figure 1.

The Introduction and Choice Manipulation were nearly identical to that of Experiment 1, with the following modifications: a smiley-face sticker was used instead of a star sticker in the Introduction phase. Additionally, because we did not wish to bias children's answers with choice language, we avoided using the phrase "You get to choose", and instead simply presented the two options (e.g., "You can either give this sticker to Doggie or you can keep it for yourself").

Dependent Measure In the dependent measure, children were reminded of the choice they had made ("Do you remember when you put that sticker in Doggie's box?"). Children were then asked a Choice Question ("Did you
choose to do that, or did you have to do that?"). The question was re-asked if children did not initially answer.

\section*{Results and Discussion}

As in Experiments 1 and 2, the majority of children in both the Costly Choice condition (19/25) and the NonCostly Choice condition ( \(25 / 25\) ) chose the prosocial action over the non-prosocial alternative (Binomial \(p\) 's \(<.01\) ).

A greater proportion of children in the Costly Choice condition stated that they chose to perform the target action (18/25; 72\%) than those in the Non-Costly Choice condition (9/25; 36\%), Fisher's exact test \(p<.05\).

These results suggest that one of the inferences children may be making during Costly Choice situations is that their actions were intentional. These results are consistent with work that finds that young children learn about people's intent, both in the moral and non-moral domain, through evaluating the presence and amount of alternative actions available to them (e.g., Kushnir, Xu, \& Wellman, 2010; Zelazo, Helwig, \& Lao, 1996).

\section*{General Discussion}

We began this paper with the hypothesis that allowing children opportunities to make costly choices would influence their subsequent prosocial behavior. In fact, children were more prosocial after making costly than noncostly choices, and after making costly choices than making no choices at all. Moreover, children were sensitive to different types of evidence for what counts as a costly action: they shared more after making costly choices, and also after giving away costly objects. Finally, we found that children were more likely to construe their actions as intentional when making costly vs. non-costly choices.

Further work may examine the specific features of costly choice situations that enable children's subsequent prosocial behavior. One possibility is that children felt positive emotion by making a costly choice, and therefore were motivated to repeat the behavior at the next timepoint. The positive emotion may have occurred because children were subconsciously attuned to their own pride in making a choice that was costly, and were thus motivated to make themselves proud again by being prosocial. Yet another possibility is that in making a costly prosocial choice, children actively self-regulated their own physiological arousal elicited by hearing about a sad puppet (Hepach, Vaish, \& Tomasello, 2012). The coordination of setting goals and watching oneself effectively self-regulate in order to meet those goals may then have empowered children to repeat the self-regulatory prosocial behavior later on (Grolnick, 2009).

It is also important to examine the scope of influence that costly choices have on the development of prosocial behavior and on later-developing altruistic behaviors. Moral self-construction, as well as altruistic behavior, are likely to be the product of a rather complicated process involving emotional, behavioral, and cognitive components (Blasi, 1983; Hardy \& Carlo, 2011; Kochanska, 2002). Our
findings show that costly choices play a causal role in determining the short-term prosocial behavior of very young children. Though more research is needed to investigate how choice interacts with other components of moral development, demonstrating the short-term results underscores previous findings that choice may make a critical contribution to children's emerging understanding of themselves as moral beings through rational inference.

\section*{Acknowledgments}

This work was supported by a Cornell Cognitive Science Fellowship to NC. We would like to thank Bertilia Trieu, Christina Bryce, Chelsea Brite, Andrew Strauss, Lauren St. Victor, Kelly Yang, and Emily Hayko for assistance with data collection and coding. A portion of this work has been accepted for publication.

\section*{References}

Beaman, A. L., Cole, C. M., Preston, M., Klentz, B., \& Steblay, N. M. (1983). Fifteen Years of Foot-in-the Door Research A Meta-Analysis. Personality and Social Psychology Bulletin, 9, 181-196.
Bem, D. J. (1967). Self-perception: an alternative interpretation of the cognitive dissonance phenomena. Psychological Review, 74, 183-200.
Blasi, A. (1983). Moral cognition and moral action: A theoretical perspective. Developmental Review, 3, 178210.

Cialdini, R. B., Eisenberg, N., Shell, R., \& McCreath, H. (1987). Commitments to help by children: Effects on subsequent prosocial self-attributions. British Journal of Social Psychology, 26, 237-245.
Dunfield, K. A., \& Kuhlmeier, V. A. (2010). Intentionmediated selective helping in infancy. Psychological Science, 21, 523-527.
Eisenberg, N., Cialdini, R. B., McCreath, H., \& Shell, R. (1987). Consistency-based compliance: When and why do children become vulnerable? Journal of Personality and Social Psychology, 52, 1174.
Eisenberg, N., Guthrie, I. K., Cumberland, A., Murphy, B. C., Shepard, S. A., Zhou, Q., \& Carlo, G. (2002). Prosocial development in early adulthood: a longitudinal study. Journal of Personality and Social Psychology, 82, 993.

Freedman, J. L., \& Fraser, S. C. (1966). Compliance without pressure: the foot-in-the-door technique. Journal of personality and social psychology, 4, 195.
Grusec, J. E., Kuczynski, L., Rushton, J. P., \& Simutis, Z. M. (1978). Modeling, direct instruction, and attributions: Effects on altruism. Developmental Psychology, 14, 51.
Grusec, J. E., \& Redler, E. (1980). Attribution, reinforcement, and altruism: A developmental analysis. Developmental Psychology, 16, 525.
Grolnick, W. S. (2009). The role of parents in facilitating autonomous self-regulation for education. Theory and Research in Education, 7, 164-173.

Hardy, S. A., \& Carlo, G. (2011) Moral identity: What is it, how does it develop, and is it linked to moral action? Child Development Perspectives, 5, 212-218.
Hepach, R., Vaish, A., \& Tomasello, M. (2012) Young children are intrinsically motivated to see others helped. Psychological Science, 23, 967-972.
Kochanska, G. (2002). Committed compliance, moral self, and internalization: A meditational model. Developmental Psychology, 38, 338-351.
Kushnir, T., Xu, F., \& Wellman, H. M. (2010). Young children use statistical sampling to infer the preferences of others. Psychological Science, 21, 1134-1140.
Lepper, M. R. (1973). Dissonance, self-perception, and honesty in children. Journal of Personality and Social Psychology, 25, 65.
McGrath, M. P., Wilson, S. R., \& Frassetto, S. J. (1995). Why some forms of induction are better than others at encouraging prosocial behavior. Merill Palmer Quarterly, 41, 347-347.
McGrath, M. P., \& Power, T. G. (1990). The effects of reasoning and choice on children's prosocial behavior. International Journal of Behavioral Development, 13, 345-353.
Schmidt, F. H., \& Sommerville, J. A. (2012). Fairness expectations and altruistic sharing in 15-month-old human infants. PLoS ONE, 6, e23223.
Staub, E. (1971). The use of role playing and induction in children's learning of helping and sharing behavior. Child Development, 805-816.
Svetlova, M., Nichols, S., \& Brownell, C. (2010). Toddlers' prosocial behavior: From instrumental to empathic to altruistic helping. Child Development, 81, 1814-1827.
Vaish, A., Carpenter, M., \& Tomasello, M. (2009). Sympathy through affective perspective-taking and its relation to prosocial behavior in toddlers. Developmental Psychology, 45, 534-543.
Vaish, A., Carpenter, M., \& Tomasello, M. (2010). Young children selectively avoid helping people with harmful intentions. Child Development, 81, 1661-1669.
Vaish, A., Missana, M., \& Tomasello, M. (2011). Three-year-old children intervene in third-party moral transgressions. British Journal of Developmental Psychology, 29, 124-130.
Warneken F., \& Tomasello, M. (2006). Altruistic helping in human infants and young chimpanzees. Science, 311, 1301-1303.
Warneken, F., \& Tomasello, M. (2008). Extrinsic rewards undermine altruistic tendencies in 20-month-olds. Developmental Psychology, 44, 1785-1788.
Zahn-Waxler, C., Radke-Yarrow, M., Wagner, E. W. M. (1992). Development of a concern for others. Developmental Psychology 28, 126-136.
Zelazo, P. D., Helwig, C. C., \& Lau, A. (1996). Intention, act, and outcome in behavioral prediction and moral judgment. Child Development, 67, 2478-2492.

\title{
Estimating Multiple Evoked Emotions from Videos
}

\author{
Wonhee Choe (wonheechoe@gmail.com) \\ Cognitive Science Program, Seoul National University, Seoul 151-744, Republic of Korea \\ Digital Media \& Communication (DMC) R\&D Center, Samsung Electronics \\ Suwon 443-742, Republic of Korea
}

\author{
Hyo-Sun Chun (hschun@bi.snu.ac.kr) \\ Junhyug Noh (jhroh86@gmail.com)
}

School of Computer Science and Engineering, Seoul National University, Seoul 151-744, Republic of Korea

Seong-Deok Lee (Isdlee@samsung.com)
Future IT, Samsung Advanced Institute of Technology (SAIT), Yongin 449-712, Republic of Korea

\author{
Byoung-Tak Zhang (btzhang@bi.snu.ac.kr) \\ Computer Science and Engineering \& Cognitive Science and Brain Science Programs, Seoul National University, Seoul, 151-744, Republic of Korea
}

\begin{abstract}
Video-sharing websites have begun to provide easy access to user-generated video content. How do we find what we want to view among the huge video database? When people search for a video, they may want to know whether the video evokes a certain emotional sensation. The evoked emotion is one of the important factors we consider when we select a video. One of the key concepts of evoked emotions from videos: the evoked emotions are different for each scene and for each viewer. Considering these differences, we obtained humanevoked emotions from 33 videos. We used these emotions to estimate the multiple emotions evoked by each scene of the videos. Using a computational model of emotion estimation based on mid-level visual features, we found that, in individual videos, the same scene evoked multiple emotions. Our results show that a video evoked different emotions from different people. A computational model might deliver probabilistic multiple-evoked emotions from video analyses.
\end{abstract}

Keywords: evoked emotion; visual feature; video retrieval

\section*{Introduction}

Video-sharing websites provide easy access to a wide variety of user-generated video content, including movie clips, television clips, music videos, and amateur content. We can search a huge database of videos for a video that we want to see via the use of keywords or a search by genre. Unfortunately, these search strategies are not sufficient. If we do not have any prior knowledge of a video, how do we find what we want to see? The mood of movie is one of the most important factors we consider when we select a movie. When people search for a movie or TV series, they may want to know whether the video has a mood that is similar to that of a video they have viewed before. Sometimes, they may want to change their mood by watching a video.

Until now, most of the research efforts have focused on a content- or genre-based analysis of videos despite many users' needs for emotion-sensitive video retrieval. Fortunately, some have studied the emotions evoked by images (Wang \& He, 2008; Yanulevskaya et al., 2008; Li, Zhang \& Tan, 2010). Wang and He (2008) focused on emotional semantic image retrieval (ESIR) instead of content-based image retrieval (CBIR) to reduce the "semantic gap." Others studied emotional picture categorization using the International Affective Picture System (IAPS) according to the 10-emotion model (Yanulevskaya et al., 2008; Li, Zhang \& Tan, 2010). However, the application of these approaches to a video sequence may not be appropriate due to the lack of consideration of temporal variations. Various moods change sequentially within a given video. That is, within a given video, different emotions may be evoked across scenes.

Two different studies evaluated the emotions evoked by videos. Canini et al. (2009) evaluated the emotional identity of videos. The research used light source color, motion dynamics, and audio track energy as the temporal features of videos. It was a first attempt to evaluate the temporally changing emotional identity of movies. They used a 3 dimensional (3D) emotional identity space (warm/cold, dynamic/slow, and energetic/minimal) to show the trajectory of one video clip. Unfortunately, the emotional identity space was used to express movies' content changes rather than humans' various emotional changes. Bailenson et al. (2008) focused on a classification algorithm of emotions evoked by videos. Facial feature tracking was used and physiological responses were measured. The study tried to predict 2 emotions (amusement and sadness) and the intensity of each emotion. Unfortunately, this approach is not applicable to the study of video retrieval.

Winter and Kuiper (1997) reported that individual difference factors play an important role in the experience of emotion. In fact, individuals may respond differently depending on their current state of mind. However, most research of the emotion of videos has assumed that an emotion is unified at any given moment.

In the present paper, we propose a new temporally changed emotional analyzer that functions as a probabilistic estimator of multiple emotions evoked by videos. The goal of this study is to generate sequentially changing emotional responses from a video clip. This paper is organized as follows: First, a psychophysical experiment to investigate the evoked emotions by each scene is described. Second, the proposed system is introduced with mid-level visual features, an estimation model, and a performance test. Finally, the conclusions are summarized and future tasks are proposed.

\section*{Method}

\section*{Data Set}

To investigate the evoked emotion of each video clip, stimuli were taken from following movies and TV series: Amélie of Montmartre (2001), Artificial Intelligence: AI (2001), Curse of the Golden Flower (2006), The Amazing Spider-Man (2012), Wuthering Heights (2012), Friends Season8: The One Where Rachel Tells Ross (2001), and CSI: Miami Season8 Episode4: In Plane Sight (2009). The movies and TV series were selected non-intentionally. Each was decomposed to a set of video scenes, the emotional valence of which varied. Thirty-three video clips were selected in order to not exceed one hour per experiment per person. The average run time of the clips was 81 seconds. Each video clip had one main event that occurred in one location. We preferred that each video clip evoke one kind of emotion.

\section*{Evoked Emotions}

To examine the emotions evoked, we used emotional models instead of Canini's 3D emotional identity space (Canini et al., 2009). Ekman's model is defined by 6 basic facial emotional expressions: anger, surprise, disgust, sadness, happiness, and fear (Ekman \& Friesen, 1978). However, this model may not be applicable for videoevoked emotions because it contains more negative emotions than positive ones: it has 4 negative emotions (anger, disgust, sadness, and fear), one neutral emotion (surprise) and one positive emotion (happiness). Thus, the model needs to be balanced. Moreover, the model did not include some key emotions that viewers experience while watching videos. Gross and Levenson (1995) modified the 6 -emotion model to an 8 -emotion model. The 8 -emotion model comprises the following emotions: amusement, anger, contentment, disgust, fear, neutral, sadness and surprise. Unfortunately, this model is also not balanced.

We designed a 12-emotion model that includes Gross and Levenson's (1995) 8-emotion model with 4 emotions (humor, romance, tension, and suspicion). Our emotional space is balanced in that there are 4 positive emotions (humor, romance, contentment, and joy/pleasure), 4 blended emotions (suspicion, tension, surprise, and neutral), and 4 negative emotions (anger, disgust, sadness, and fear).

\section*{Procedure}

20 people participated in the experiment. At the beginning of the test, the 12 emotions were described (Fig. 1) to the participants. The video clips were shown to the participants on a TV monitor. The presenting order of videos was randomized to minimize potential error. After watching each video clip, the participants were asked to choose 1 of the 12 emotions that best represented the emotion evoked by the video clip.


Figure 1 : The names and illustrations of the 12 emotions used in this experiment.


Figure 2 : Emotions evoked by the 33 video clips.

\section*{Results}

\section*{Responses of Evoked Emotion}

The participants' responses are summarized by a 12emotion population for each video clip. The results of the evoked emotional responses for the 33 video clips are illustrated in Figure 2. The figure shows the frequency of the most commonly reported and the second most commonly reported emotion per video clip. Some emotions were not reported often; this may generate some noise in our model. Gross and Levenson (1995) reported that contentment, anger, and fear are more difficult to be evoked by movies than are other emotions. Notably, anger was not evoked by the video clips.

\section*{Statistical Analysis}

To determine whether the response data varied according to the video clip, we analyzed the data statistically using SPSS 1.9. First, we tested independence of the participants' responses for all of the test video clips, through a crosstabulation analysis using chi-square tests. The relationship between the evoked emotion and the video clip presented was significant ( \(\mathrm{p}<0.001\) ).

Second, the independence of participants' evoked emotions was evaluated in the several extracted video clips from the same movie or TV series. The video clips were composed of 3 clips of AI, 5 clips of CSI, 4 clips of Wuthering Heights, 4 clips of Amelie, 4 clips of Friends, and 12 clips of Spider-Man. The evoked emotion was significantly related to the video clip ( \(p<0.001\) ). The results provided in Tables 1 and 2 are examples of our statistical analysis applied to the data collected in response to the 3 clips of \(A I\). While the \(A I\) clips evoked all of the 12 emotions, the other clips evoked as 3 or 4 emotions.

Table 1 : The results of the cross-tabulation analysis for the \(A I\) video clips.
\begin{tabular}{|c|c|c|c|c|c|c|c|c|c|c|c|c|c|}
\hline \multirow[b]{2}{*}{Count} & \multicolumn{12}{|c|}{Emotion} & \multirow[b]{2}{*}{Total} \\
\hline & \multicolumn{4}{|l|}{Hum Rom Con Joy} & Sur & Sus & Ten & Neu & Ang & & Sad & & \\
\hline AI_1 & 0 & 0 & 0 & 1 & 0 & 2 & 13 & 3 & 0 & 1 & 0 & 0 & 20 \\
\hline AI_2 & 1 & 0 & 0 & 0 & 3 & 0 & 0 & 0 & 3 & 9 & 2 & 2 & 20 \\
\hline AI_3 & 0 & 4 & 9 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 4 & 0 & 20 \\
\hline Total & 1 & 4 & 9 & 4 & 3 & 2 & 13 & 3 & 3 & 10 & 6 & 2 & 60 \\
\hline
\end{tabular}

Table 2 : The results of the chi-square test for the data collected in response to the \(A I\) video clips.
\begin{tabular}{cccc}
\hline Chi-Square Tests & Value & df & \begin{tabular}{c} 
Asymp. Sig. \\
(2-sided)
\end{tabular} \\
\hline Pearson Chi-Square & 102.100 & 22 & .000 \\
Likelihood Ratio & 113.195 & 22 & .000 \\
Linear-by-Linear Association & 6.362 & 1 & .012 \\
N of Valid Cases & 60 & & \\
\hline
\end{tabular}

\section*{Estimation Model of Evoked Emotions}


Figure 3 : A block diagram of our proposed estimation system.

Do computers mimic human emotions such as the emotions evoked by the video clips? To answer this question, we proposed a movie-evoked emotion estimator by introducing mid-level visual features. The mid-level visual features were composed by referring to color emotional theory with a 3color combination and emotionally correlated general visual features: contrast by histogram distribution, relative brightness, and motion size. To create the evoked-emotion estimator, we examined the relationship between humans' psychophysical responses and the mid-level visual features extracted from the video clips. The emotion model was trained by supervised learning with humans' psychophysical responses and their features. The proposed approach generated multiple emotional states of the video contents sequentially and is illustrated in Figure 3.

\section*{Visual Feature Extraction}

Psychologists have investigated the emotion-eliciting properties of industrial media (Gross \& Levenson, 1995; Pos \& Armytage, 2007; Kobayashi, 1981). In particular, Gross and Levenson (1995) focused on eliciting the emotion of films. They determined that each film could evoke 8 different emotions from the viewers. The results show that video clips can be categorized by the different emotions that they evoke from the viewers. However, they did not attempt to draw relationships between the visual cues and the elicited emotions. The present study focused on some midlevel visual cues and evoked emotions from some video clips. The mid-level visual cues are reorganized by an analysis of low-level features. Moreover, the mid-level cues were differentiated from low-level cues. The mid-level cues were extracted by color, contrast, brightness, and motion.

Color Color is known to correlate strongly with psychological constructs. Many studies describe the relationships between these variables (Pos \& Armytage, 2007; Kobayashi, 1981; Solli \& Lenz, 2010), but to date, correlations in movie settings have not been studied. Pos and Armytage (2007) investigated the relationships between emotions, facial expressions, and colors. Kobayashi (1981) matched 1170 3-color combinations to 180 adjective words describing emotional appearance. Solli and Lenz (2010)
transformed and classified natural images on the basis of Kobayashi's list of emotional words. Kobayashi's color scale is useful for industrial design. Unfortunately, it is not appropriate to apply the 180 emotional words (i.e, warm, cold, luxury, etc.) to movies. Pos and Armytage (2007) studied the relationships between 3-color combinations and Ekman's 6-emotional facial expressions. The 3-color combination might connote simple feelings (e.g., warm, cold) and emotions (e.g., happy, sad).

We extracted the color distribution of a video clip to estimate evoked emotions. We transformed the colors of the input video clips into Kobayashi's 130 -color scale Hue and Tone System. The transformation helped to reduce the complexity of the analysis \(\left(2^{8 \times 3} \rightarrow 130\right)\) and to classify the input colors into emotionally meaningful representative colors. An input pixel color (RGB) is converted to the Hue and Tone value (HT) as shown in Equation 1.
\[
\begin{equation*}
R G B t o H T(x)=\underset{i}{\arg \min }\left\{R G B_{-} \operatorname{Dist}(x, H T[i])\right\} \tag{1}
\end{equation*}
\]
where \(x\) is 1 of the \(2^{8 \times 3}\) colors used as an input color, \(H T[\mathrm{i}]\) is 1 of the 130 Hue and Tone colors ( \(1 \leq i \leq 130\) ), and \(R G B \_\operatorname{Dist}(x, y)\) is a distance measuring method such as Euclidean distance on RGB space. The HT with the minimum distance was selected as the HT value of the pixel. Then, we extracted 3 colors used most frequently in the frame. The 3 dominant colors were accumulated for all of the frames of a video clip as a probability distribution of hue and tone colors. The entire process is illustrated in Figure 4.
(a)


Figure 4 : Our color extracting method: (a) a video stream; (b) RGB images of each of the frames; (c) converted HT images of each of the frames; (d) three dominant colors of each of the frames; and (e) a probability distribution of HT for all of the frames.


Figure 5 : Contrast classification method (Kim, Choe, \& Lee, 2006).

Contrast Except for the color feature, we can extract many features from visual information. Above all, a high-contrast scene may evoke a different emotion than a low- or mildcontrast scene. We used a contrast-categorization-method based histogram analysis of image intensity (Kim, Choe, \& Lee, 2006). The categorization performance of the method has already been proven by a previous human psychophysical experiment. The method assigns the histogram distribution of every image to 1 of 5 representative contrasts. We merged class \(\mathbf{D}\) and class \(\mathbf{E}\) (see Figure 5) into 1 category since they yield similar contrast.

Brightness Image brightness is 1 of the important cues serving as a connotative feature of videos. Low brightness may not be used by directors to express a hopeful scene. However, it would be inappropriate to use absolute brightness as an estimating tool for eliciting emotion, because the overall brightness of a video depends on the brightness characteristics of the video camera or the director's preference. Consequently, we extracted the relative brightness \((\widetilde{\boldsymbol{Y}})\) to the average brightness of the entire video as described in Equation 2. \(Y_{f}(i)\) is an average brightness of all of the pixels in the \(i\) th frame, \(n\) is the total number of frames for a scene, and \(m\) is the total number of frames for an entire video.
\[
\begin{equation*}
\tilde{Y}=\frac{\sum_{i=0}^{n} Y_{f}(i)}{n}-\frac{\sum_{j=0}^{m} Y_{f}(j)}{m} \tag{2}
\end{equation*}
\]

Motion Motion is an important factor in the evaluation of how dynamic scene is. We used a scale-invariant feature transform (SIFT, Lowe, 2004) to obtain motion information. SIFT features are extracted from the video frames and their trajectory is evaluated to estimate inter-frame motion (Lowe, 2004). In this study considers the distance of their trajectory was considered as the motion size. Then, the motion feature defined the largest motion size of each inter-frame.

\section*{Emotion Estimation}

To estimate emotions evoked in participants while watching videos, we used an evoked-emotion model. The psychophysical experiment was used to develop the emotion model and the multiple emotional responses were the
training data. Then, the emotion model was learned by a supervised learning method using the psychophysical responses.

Training Data As mentioned above, we extracted a motion feature from inter-frame data and 3 other features from intra-frame data. We needed some representative values of the features to train the emotion model. Thus, each feature was analyzed in every frame of a clip, and all feature data were summarized during 1 -second intervals. All video clips were standardized to the same pixel resolution and frame rate. Every clip was composed of 24 frames per a second. The color feature was encoded by using a 130 -variable combination. The variable was calculated by normalization (between 0 and 1) of the accumulated HT probability distribution for all of the frames. The contrast feature was encoded by using a 5 -variable combination and the variable was presented by the normalization of the accumulated probability distribution for contrast categories of all of the frames. Brightness and motion features were calculated by averaging them over all of the frames.

A training set of supervised learning consisted of an input feature vector and an output target vector. The target vector was designed as 6 probability values. The probability values were such that five emotions (contentment, surprise, anger, sadness, and fear) were filtered out to prevent mis-learning by null data (see Fig. 2). Moreover, "neutral" is a broad emotional term with diverse meanings and, therefore, it was also filtered out. Owing to the removal of the emotions, 6 clips in which one of the removed emotions predominated had to be excluded.

Estimation Model The emotions evoked in the participants were the target data used to train the model. Classification and Regression Trees (CART, Breiman, et al., 1984), a kind of decision tree learning, was selected as the learning method to classify the evoked emotion from a video. One advantage of CART is that it can consider misclassification costs in the tree induction, using handling numerical-valued attributes. The model learns a probabilistic response for each emotion and the extracted visual features. The participants' responses are set as the independent variables and mid-level visual features serve as the dependent variables of the model.

Estimation Performance Finally, we conducted supervised learning with CART. The model outputs probability values (between 0 and 1) for every 6 emotions for an input video. We evaluated the performance with a between-clip crossvalidation of 27 video clips.

A summary of our estimation performance is shown in Table 3. For the sake of convenience, the accuracy of performance was calculated by using an emotion with the estimated maximum probability for each clip. That is, each of the 27 videos had a predicted emotion with a majority probability. The predicted emotions were compared with the most common emotion of the target data; the correction rate
was \(56 \%\). In Table \(3,2^{\text {nd }}\) emotion accuracy refers to the percentage of times that either the most common or the second-most-common emotions was the predicted emotion.

Table 3: Details of the accuracy of the evoked emotion model.
\begin{tabular}{cccc}
\hline & \(1^{\text {st }}\) emotion & \(2^{\text {nd }}\) emotion & \(3^{\text {rd }}\) emotion \\
\hline \begin{tabular}{c} 
Prediction \\
accuracy
\end{tabular} & \(56 \%\) & \(63 \%\) & \(70 \%\) \\
\hline
\end{tabular}

Figure 6 shows a partial emotional profile of The Amazing Spider-Man using our model. The profile shows that the movie does not have 1 evoked emotion; instead, it has 3 or more evoked emotions at the same time. In addition, each emotional state is shown as a probability value. The emotional state exhibits variations that are similar to the original characteristics of the movie. For instance, the middle of the movie has many scenes with a heroine and a hero that induce a romantic mood, and the latter part has many episodes in which the hero challenges a powerful villain. Further, the dominant emotion of each second from Figure 6 is illustrated in Figure 7. The pie chart might help one to search a movie for some of the most important information.


Figure 6 : A partial emotional profile of The Amazing SpiderMan using the proposed method.


Figure 7 : The emotional composition of The Amazing Spider-Man.

\section*{Discussion}

Previous attempts to characterize the emotions of videos have used a single emotion, as a representative emotion for an entire video sequence (Gross \& Levenson, 1995; Bailenson et al., 2008; Canini et al., 2009). However, a person's emotions vary every moment (Winter \& Kuiper, 1997).

To take these individual differences into consideration, the present paper proposes a probabilistic model that can estimate multiple emotions evoked by a video over time. The proposed method involves the automatic labeling of videos with the emotions that were evoked and the duration of the evoked emotions. The present study attempted to characterize the evoked emotions by using mid-level features from the video frames, such as dominant colors, contrast, brightness levels, and motion quantification. The characterization was derived from a previous psychophysical experiment using human participants. A classical machine learning methodology was applied in order to build and test the model of the emotional categories targeted.

The present study provides 2 significant advances. First, it is the first to propose a new paradigm for video retrieval, using probabilistic multiple-evoked emotions. This approach may be used to construct an emotion-based video retrieval system, a video recommendation system, or an emotional treatment system using a video. It may also help a machine to generate a new video profile that automatically describes sequentially changing emotions. Second, this study provides a technically new approach. That is, we used emotionally meaningful mid-level visual features and we modeled them to estimate multiple-evoked emotional states from videos.

However, there are limitations to the present research. The experimental video clips were limited and participants made non-various responses (Figure. 2). Having participants choose only a single emotional response in the experiment might lead to inadequate data for generation of a probabilistic emotional estimator. One of the interesting findings was that there were no anger responses. Mikels, et al. (2005) reported the same result with still images. The authors concluded that anger is very difficult to elicit with the passive viewing conditions of static images. Our experiment used some video clips that were a few minutes long. One explanation for the finding of no anger responses is that the video clips were not long enough to evoke the emotion.

Thus, 3 tasks are left for future research to implement our approach in a real system. First, more video clips are required to teach the various emotions to the evoked emotion model. Second, a larger sample is required to get a robust trained model. Third, other machine learning methods should be tried to find an optimal solution.

\section*{Acknowledgments}

This work was supported by the National Research Foundation (2012-0005643, Videome) grant funded by the Korea government (MEST) and was supported in part by the Industrial Strategic Technology Development Program (10044009) funded by the Korea government (MKE).

\section*{References}

Bailenson, J.N. et al. (2008). Real-time classification of evoked emotions using facial feature tracking and physiological responses. International Journal of HumanComputer Studies, 66, 303-317.
Breiman, L. et al. (1984). Classification and Regression Trees, New York, Chapman \& Hall (Wadsworth, Inc.).
Canini, L. et al. (2009). Emotional identity of movies. IEEE International Conference on Image Processing (ICIP). 1821-1824.
Ekman, P. \& Friesen, W. V. (1978). Manual for facial action coding system. Palo Alto: Consulting Psychologists Press.
Gross, J.J., \& Levenson, R.W. (1995). Emotion elicitation using films. Cognition and Emotion, 9, 1, 87-108.
Kobayashi, S. (1981). Aim and method of the color image scale. Color Research \& Application, 6, 2, 93-107.
Kim, I., Choe, W., \& Lee, S.D. (2006). Psychophysical measurement for perceptual image brightness enhancement based on image classification. Proceedings of the SPIE, 6057, 425-431.
Li, S., Zhang, Y.J., \& Tan, H.C. (2010). Discovering Latent Semantic Factors for Emotional Picture Categorization. IEEE International Conference on Image Processing (ICIP). 1065-1068.
Lowe, D. (2004). Distinctive image features from scaleinvariant keypoints. International Journal of Computer Vision, 60, 2, 91-110.
Mikels, J.A. et al. (2005). Emotional category data on images from the International Affective Picture System, Behav. Res. Methods 37, 626-630.
Pos, O., \& Armytage, P. G. (2007). Facial Expressions, Colours and Basic Emotions. Colour: Design \& Creativity 1(1) 2, 1-20.
Solli, M., \& Lensz, R. (2010). Color semantics for image indexing. Proceedings of 5th European Conference on Colour in Graphics, Imaging, and Vision, 353-358.
Yanulevskaya, V. et al. (2008). Emotional valence categorization using holistic image features. IEEE International Conference on Image Processing (ICIP). 101-104.
Wang, W., \& He, Q. (2008). A survey on emotional semantic image retrieval. IEEE International Conference on Image Processing (ICIP). 117-120.
Winter, K. A., \& Kuiper, N. A. (1997). Individual differences in the experience of emotions. Clinical Psychology Review, 17(7), 791- 821.

\title{
How Working Memory Capacity Constrains the Learning of Relational Concepts
}

\author{
Adam Chuderski (adam.chuderski@uj.edu.pl) \\ Institute of Philosophy and Institute of Psychology, Jagiellonian University \\ Grodzka 52, 31-044 Krakow, Poland
}

\author{
Anna Chuderska (ania.chuderska@gmail.com) \\ Institute of Psychology, Jagiellonian University \\ Mickiewicza 3, 31-120 Krakow, Poland
}

\begin{abstract}
We investigated the way in which working memory (WM) constrains the learning of relational concepts - categories defined by the way objects are assigned to roles in the structure of an underlying relation, and not by objects' intrinsic features. By applying to a large sample a novel test of concept learning as well as the battery of WM tasks, we found that WM is a strong predictor of the scores on the test, but the WMlearning correlation decreases as the relational complexity of the to-be-learned concepts increases. Such results support those theoretical models of relational learning, which assume that learning of relational concepts (and relations, in general) consumes more WM resources than just the processing of relations which have already been learned.
\end{abstract}

\section*{Introduction}

The issue of relational thinking - the humankind's ability to acquire, process, and effectively use mental representations of relations - has huge importance in cognitive science (Gentner \& Kurtz, 2005; Halford, Wilson, \& Phillips, 1998; Hummel \& Holyoak, 2003). A relation can be described as an ordered list (a structure, a predicate) of well-defined roles and objects that fulfill them. The key aspect of relations consists of the fact that understanding of them as well as inferring from them depends primarily on the way objects are assigned to roles in the relation's structure, and not necessarily on objects' intrinsic features. Relational representations constitute the core of human complex cognition, including abstraction, reasoning, analogy making, creativity, and language (Halford, Wilson, \& Phillips, 2010).

The extension of an \(n\)-ary relation (where \(n\) is a number of roles in a relation; its arity) is a subset of Cartesian product of \(n\) sets, which includes all lists of objects ( \(n\)-tuples) that can fulfill roles in that relation (i.e., an object from the first set in a tuple fulfills the first role, etc.; Halford et al., 1998). So, each relation can be treated as a relational category/ concept (Gentner \& Kurtz, 2005). Unlike so-called entity categories, that is, categories formed by objects due to their perceptual or/and internal (e.g., genetic) similarity (e.g., natural kinds), relational concepts in the first place organize entity categories (or lower level relational categories), and so their exemplars may drastically vary featurally. For example, the instances of the relational concept of barrier will include: a wall, a river, a person, but also an insult, and loss of support (ibidem). Relational concepts constitute the main part of culture, science, and technology.

A key goal of cognitive science is to understand what relational concepts are, how they are acquired in childhood and adulthood, and how they are used in relational thinking. Consequently, the present paper aims to deal with one specific problem in this domain: it investigates in what way the constraints of human cognitive architecture, particularly its working memory capacity (WMC), influence the learning of relational concepts (from here on, the process/ability referred to as relational learning).

\section*{Working memory and relational learning}

Computational models of relational thinking (e.g., Chuderski, Andrelczyk, \& Smolen, 2013; Doumas, Hummel, \& Sandhofer, 2008; Halford et al., 2010; Hummel \& Holyoak, 2003) as well as psychometric studies on reasoning and analogy making (e.g., Martinez et al., 2011) suggest that processing relations is grounded in working memory (WM). WM is a neurocognitive mechanism responsible for maintenance of a limited, but crucial for the current task/goal, amount of information, in an active and easily available state (Cowan, 2001). It thus allows for flexible manipulation of that information (Hummel \& Holyoak, 2003; Oberauer, Süß, Wilhelm, \& Sander, 2007), including binding of relational roles and corresponding objects, which is a necessary process for a relational representation to be constructed. People can hold in their WM up to, on average, as few as three or four chunks of information (Cowan, 2001; Luck \& Vogel, 1997) and, probably, the similar number of bindings (Chuderski et al., 2013; Oberauer et al., 2007), though these values vary among individuals (approx. from 1 to 6 ). This clearly corresponds to the fact that accuracy of processing relations sharply decreases with increasing arity of relations, and few participants can cope with relations more complex than quaternary ones (Halford et al., 2010).

An interesting research question pertains to a problem of whether similar influence of WMC, as in abovementioned case of processing relations (e.g., during analogical mapping or inference), also takes place in case of relational learning, when people have to discover an (abstract) relation between related objects and construct a mental representation of the relational concept referring to that category.

A widely used paradigm of relational concept learning was proposed by Shepard, Hovland, and Jenkins (1961). They presented to participants series of eight three-feature geometric figures, each of which could take one out of two values on each featural dimension (shape, size, color), and
needed to be classified as belonging or not to a category defined by an arbitrary rule of propositional logic (henceforth named a Boolean concept). Each decision was followed by a feedback information on whether an object was categorized correctly or not. Participants improved on that classification task, thus learning to some extent a hidden Boolean concept. The most important result suggested that the accuracy of categorization decreased as the number of features relevant for a concept increased from one to three. This observation was generalized beyond Boolean concepts domain by Halford et al. (1998), who defined relational complexity in terms of the number of entities (variables or dimensions of a relation, that is, its arity) that must be related in parallel, because their decomposition into a set of less complex relations would lead to the loss of the relation's meaning.

Moreover, accuracy of categorization decreased with increasing complexity (expressed by minimal description length, MDL; Feldman, 2000) of a logical rule associated with the three-feature concepts. However, as the MDL approach led to a problem of which logical operators should count as minimal (e.g., if we include exclusive disjunction, then MDL no longer predicts Shepard et al.'s data; see Goodwin \& Johnson-Laird, 2011), later this approach was disputed. For example, Feldman (2006) proposed an algebraic complexity metric of discrete-value concept learning difficulty, which depends on the sum of a number of constant values of variables and the number of implications which derive values of one variable from values of another variable, into which a given concept can be decomposed. Kemp, Goodman, and Tenenbaum (2008) adopted this approach to describe relational concepts beyond a Boolean domain. Those approaches nicely predicted observed data. Similarly good fit was obtained by a theory predicting that the number of all possible mental models (iconic-like representations precisely corresponding to the structure of themselves roughly represented - elements of a situation) which match a rule describing a concept (Goodwin \& Johnson-Laird, 2011).

In the present paper, we ask whether the effectiveness of relational learning can be predicted by WMC. Moreover, we test whether the link between those two variables, if any is found, depends on the abovementioned complexity of concepts which are learned. Such a test may be very informative regarding the validity of existing models of relational learning, because, as we will see, some of them seem to yield opposite patterns of predictions on the strength of WMC-relational-learning link in the function of complexity.

Although Halford et al. (1998) have not inferred such predictions directly from their theory (instantiated also in a computational model called STAR), closer inspection of this theory leads to the prediction that the critical value of relational complexity for learning relations should be four dimensions. For example, Halford, Baker, McCredden, \& Bain (2005) have shown that accuracy to understand statistical interactions is quite high for two- and three-way interactions, while it radically falls down in the case of fourway ones. As Halford et al. (2010) assume that the same constrains pertain to both processing and acquiring relations (both limits are grounded in the maximal size of a tensor that humans can mentally represent), learning bi- and ter-
nary relations should be relatively easy and not so much constrained by individual WMC, as the mean WMC is about four. In contrast, there should be substantial differences in learning quaternary relations, as people of WMC below four (i.e., one, two, or three) will not be able to learn them fully, while people of WM above that limit (i.e., of four, five or six slots) will have enough capacity to do that. So, the correlation between WMC and relational learning should be the strongest in case of the mean value of WMC.

In contrast, a neurosymbolic model of the discovery of relations proposed by Doumas et al. (2008) assumes that in order to learn a relation, a cognitive system has to represent each role-filler pair as two separate neuronal oscillations, asynchronic, but peaking close in time. This implicates that for learning each dimension of a relation, the system needs two WM chunks, and only after having learned it, both a role and a filler can be compressed into one synchronized oscillation. So, even learning binary relations will consume WMC (i.e., four chunks) of a large part of participants, and their performance on binary relations should be particularly sensitive to individual differences in WMC. Learning ternary (i.e., requiring six WM slots) or quaternary (i.e., occupying eight slots) relations should be difficult for almost everyone's WM, and - if nevertheless effective - will have to rely on mechanisms other than WM (e.g., relational knowledge accretion, compressing relations, etc.).

Interestingly, a recent study by Lewandowsky (2011), who examined correlations between each type of Shepard et al.'s concepts and WMC, has shown that the strength of such a correlation is basically the same in case of unary, binary and ternary concepts of such a kind. This study suggests that a third possibility regarding the pattern of correlations between relational learning and WMC is possible, specifically that WMC influences learning relations of any complexity. However, three disputable aspects of the Lewandowsky's study suggest that more data is needed before a decisive conclusion on WMC-relational learning link can be given.

Firstly, the criterion for a successful learning of Shepard et al.' concepts was that a certain number of correct classifications can be consecutively made by a participant. However, this does not guarantee that he or she really started to represent a relation underlying the concept, because due to a large number of classification trials a complex association, instead of a fully-blown relational representation, may be formed as well. So, in order to prevent such a case, participants should be able to explicitly report a relation to be found, as a necessary criterion for judging that a relational representation has indeed been learned.

Secondly, with the use of Shepard et al.'s concepts, at most ternary relations can be investigated, what does not allow to directly test predictions derived from Halford et al. (1998, 2005, 2010). More complex relations, above and beyond binary features and three dimensions, are needed (e.g., Kemp et al., 2008). Optimally, participants should be required to learn quaternary relations, in which each variable depends on the values of three other variables.

Finally, all existing studies (e.g., Goodwin \& JohnsonLaird, 2011; Halford et al., 2005; Kemp et al., 2008; Lewandowsky, 2011) have investigated relational learning defined
as the rate of success in acquisition of a to-be-learned concept/rule. However, learning is a dynamic process, and its most important indicator is the increase of knowledge one has gained, and not the total amount of knowledge (including the knowledge possessed before the study started) that one can display. So, the examination of the progress in the discovery of relations, and not only how one can discover them in general, as well as the testing of possible associations between the rate of that progress and WMC, can bring a vital insight about relational learning and its WM mechanisms. To our knowledge, no study so far addressed all aforementioned issues in parallel.

In the remaining part of the paper, we present a direct examination of possible predictions on the link between WMC and relational learning, by applying to a large sample of participants a test that requires discovery of relational concepts differing in complexity. Each discovered concept must aptly describe six presented associated exemplars, while excluding three accompanying counterexamples. We also measured participants' WMC with four versions of a well-established WM measure (a complex span task). We investigated the resulting correlations with the use of confirmatory factor analysis (CFA). Firstly, we correlated relational learning accuracy with WMC, in the function of the complexity of the former. Secondly, we tested the link between the latter and the improvement on learning, that is, when structurally identical relations must be discovered for the second time, but now governing new kind of stimuli. That is, we examined if the transfer of the effects of relational learning would be linked to WMC or not.

\section*{A study}

\section*{Participants}

A total of 243 participants ( 142 women, mean age \(=24.3\) years, \(S D\) age \(=5.0\), range \(18-45\) years) were recruited via publicly accessible social networking websites. Each person was paid 15 euro for their participation in the study. Data from six people were discarded because of theirs failure to provide even one elaborate description in the learning test.

\section*{A test of relational concepts discovery}

The DREL (Discovery of RELations) paper-and-pencil test consists of two, letter and digit, parts. Each part includes 15 items. Each item consists of six four-symbol strings, which are governed by a to-be-discovered relation, and another three strings, which form counterexamples for that relation, that is, the discovered relation must exclude all three counterexamples. A participant is required to write down a concise and abstract description of a relation that matches six positive exemplar strings. The counterexamples were introduced in order to prevent describing too general relations (e.g., all strings consist of four symbols). In each part of the test, there are five binary, five ternary, and five quaternary relations, and item positions for each complexity level with regard to the beginning of the test were balanced (the sequence of levels is: 324234342432423 ).

In the first part of the test, symbols in each string are two different letters, and a relation governs the place of each letter relative to some number of remaining letters in a string. We assumed that in binary-relation items, the proper relation can be discovered using only pairwise comparisons of letters, so in each step of analysis of a string, a participant needs to maintain in WM only two representations. One example of a binary-relation item requires to discover a relation the same letters in the middle are different from the same two letters on the extremes:

\section*{OEEO LSSL BVVB \\ ZKKZ NUUN YAAY \\ RRVV AKAK PPLL}

Counterexamples prevent people from proposing relations like there are always two exemplars of one letter and two of the other letter. There is only one mental model corresponding to binary relations (in case of this example: \(a b b a\) ).

In the ternary-relation items, the proper relation can be discovered using comparisons of three letters in parallel, so in each step of analysis three representations have to be maintained in WM. An instance of ternary relation is one and only letter different from three other identical letters is always placed in the middle (corresponding models are: \(a a b a\) and \(a b a a\) ). In the item presented below, a participant is expected to relate: a pair of two identical letters to another identical letter on the opposite, and both of them to one remaining different letter always placed in the middle:
\begin{tabular}{llc} 
ZEZZ & LLUL & NRNN \\
ASAA & JJWJ & PBPP \\
OLLL & KKKN & VVVB
\end{tabular}

In the most difficult, quaternary-relation items, we assumed that all four letters have to be related in one step. An example relation is the first letter is different from the second one or the third one or both, and the third letter is different from the fourth one (three corresponding models: \(a a b a, a b a b\), and \(a b b a\) ). The complexity of this relation is a result of introducing an inclusive disjunction \(x\) or \(y\) or both. A participant in this example is expected to simultaneously relate the first letter to the second, the first one to the third, and the third one to the fourth:


The only difference between the first and the second part of the test is that symbols are digits, and relations pertain to
their evenness or oddness. However, the abstract structure of relations of corresponding items in both tests is identical. For example, the digit version of aforementioned binary relation would be: two digits in the middle are both odd or both even, and in the former case two extreme digits are even, while in the latter case two extreme digits are odd. This part is more difficult, as the crucial feature (evenness/ oddness) is not linked to the appearance of a symbol, while the crucial feature of the letter part (identity/difference) is.

The scoring on the test depended on the abstractness on given descriptions. One point was scored if a described relation was correct and properly abstract (as in the examples), no matter what exact formulation were used by participants. Half point was scored if a description was correct, but it was not abstract enough, instead it was composed of particular subcategories of strings (usually corresponding to possible models), for example, in case of the ternary example, if a description was like there is either (a) one letter, then another is different, and then two last letters are the same as the first one, or (b) there are two same letters, then another is different, and then the last letter is the same as the two first ones. No score was given for incorrect descriptions, no matter if they excluded valid instances of strings or included counterexamples. Such a partial scoring resulted in much better reliability of the test (Cronbach's \(\alpha=.91\) ) than did binary (correct/incorrect) scoring ( \(\alpha=.78\) ). The dependent variables were total scores (in range 0 to 5 ) on each level of relational complexity, and the corresponding differences between scores on the second and first part of the test (i.e., indices of learning).

\section*{Working memory tasks}

Four complex span tasks were designed following Conway et al. (2005). In general, a complex span requires memorizing a sequence of a few stimuli, each of them followed by a simple decision task. In the present versions, each task required memorizing three to seven (set size) stimuli, presented for 1.2 s apiece, out of nine possible ones for that task. After two two-stimuli training trials, three trials for each set size (in increasing order) were presented in each complex span task. The letter span task (sometimes called an operation span task) required memorizing letters, while deciding with a mouse button if intermittent simple arithmetical equations (e.g., \(2 \times 3-1=5\) ? ) are correct or not. The digit span consisted of memorizing digits, while checking if letter strings begin and end with the same letter. The spatial span task required memorizing locations of a red square in the \(3 \times 3\) matrix, while deciding which of two presented bars is larger (the difference was always \(25 \%\) ). In the figural span task, participants were instructed to memorize simple geometric figures, while judging colors to be light (yellow or beige) or dark (brown or navy blue). The dual (decision) task in each WM test aimed to prevent the chunking of stimuli or the extensive use of phonological loop, which could obscure "real" WMC of individuals. The participants were instructed that they should recall as many stimuli as they can (in proper order), but that they should also try to be correct on the decision tasks.

The response procedure in each task consisted of a presentation of as many \(3 \times 3\) matrices as was a particular set size, in the center of the computer screen, from left to right. Each matrix contained the same set of all nine possible stimuli for a task. A participant was required to point with the mouse those stimuli that were presented in a sequence, in the correct order (from left to right). Only choices that matched both the identity and ordinal position of a stimulus were taken as correct answers. The dependent variable for each complex span task was the proportion of correct choices to all stimuli presented in the task. All complex span tasks displayed high reliability ( \(\alpha \mathrm{s}=.85\) to .89 ).

\section*{Procedure}

The presented study was a part of a larger project testing various cognitive abilities (WM, attention, reasoning), which included 17 computerized tasks applied in one fourhr session, and 5 tests of relational thinking applied in another four-hr session (sessions were administered in a random order), with a 1 -hr break between the sessions. Complex span tasks were the \(5^{\text {th }}, 9^{\text {th }}, 13^{\text {th }}\) and \(16^{\text {th }}\) tasks in a row applied in the former session, while the DREL test was the first task in the latter session. Half hour was allowed for each part of the DREL test.

\section*{Results}

Table 1 shows the descriptive statistics and correlations of all dependent variables. No variable deviated from the normal distribution. Correlations ranged from moderate \((r=\) \(.21)\) to strong \((r=.75)\).
\begin{tabular}{lcccccccccc}
\hline \multicolumn{1}{c}{ Task } & 1. & 2. & 3. & 4. & 5. & 6. & 7. & 8. & 9. & 10. \\
\hline 1.DL2 & - & & & & & & & & & \\
2.DL3 & .46 & - & & & & & & & & \\
3.DL4 & .36 & .66 & - & & & & & & & \\
4.DD2 & .42 & .45 & .44 & - & & & & & & \\
5.DD3 & .27 & .46 & .47 & .65 & - & & & & & \\
6.DD4 & .23 & .43 & .52 & .56 & .75 & - & & & & \\
7.LSPAN & .36 & .39 & .34 & .37 & .40 & .31 & - & & & \\
8.NSPAN & .42 & .37 & .30 & .38 & .33 & .26 & .70 & - & & \\
9.SSPAN & .21 & .32 & .23 & .34 & .34 & .25 & .57 & .51 & - & \\
10.FSPAN & .24 & .29 & .23 & .36 & .32 & .28 & .65 & .72 & .59 & - \\
Mean & 4.46 & 2.18 & 1.53 & 3.07 & 1.37 & 0.88 & 0.69 & 0.76 & 0.52 & 0.62 \\
SD & 0.95 & 0.85 & 0.99 & 1.46 & 0.96 & 0.96 & 0.19 & 0.16 & 0.18 & 0.18 \\
Min. & 0 & 0 & 0 & 0 & 0 & 0 & 0.05 & 0.09 & 0.05 & 0.13 \\
Max. & 5 & 4 & 4.5 & 5 & 3.5 & 3 & 0.99 & 1.00 & 0.97 & 1.00 \\
\hline
\end{tabular}

Table 1: Correlation coefficients and descriptive statistics for all dependent variables in the study ( \(\mathrm{N}=237\) ). All correlations were significant at \(p=.001\) level. Note: D DREL test, L or D - its letter or digit version, 2, 3, or \(4-\) relational complexity level. SPAN - versions of complex span task, L - letter, N - number, S - spatial, F - figural.


Figure 1: The general structure of the CFA models linking the discovery of binary, ternary, and quaternary relations to WMC. Ovals represent latent variables (factors), while boxes stand for observed variables (measures). Arrows represent factor loadings, while a line stands for correlation.

The two (test versions) by three (levels of complexity) ANOVA of the DREL test's scores indicated that they were significantly higher in the letter version \((M=2.72)\) than in the digit version \((M=1.77), F(1,236)=316.93, p<.001\), \(\eta^{2}=.57\), and that they decreased with increasing relational complexity \(\left(M_{\mathrm{RC} 2}=3.75, M_{\mathrm{RC} 3}=1.74\right.\), and \(\left.M_{\mathrm{RC} 4}=1.20\right)\), \(F(2,472)=1523.80, p<.001, \eta^{2}=.87\). Also, both factors interacted, \(F(2,472)=43.18, p<.001, \eta^{2}=.15\), as the effect of complexity was more profound in the letter version than in the digit one. These data indicate that the DREL test seems to be a proper tool for measurement of how effectively people discover relations, and that participants were sensitive to the complexity of the test's items.

Then, we tested whether our participants improved at all in the digit version of the DREL test, by comparing their scores on that version to another 79 participants from a similar study, who only attempted the digit version (i.e., they did not "train" on the letter version). This control group scored \(M=1.34\) per condition (comparing to \(M=1.77\) in the experimental group), that is, there was a highly significant learning effect, \(t(314)=3.46, p<.001\).

Next, with CFA, we assessed the strengths of correlations between the latent variable reflecting WMC (loaded by four complex span tasks) and variables representing the effectiveness of the discovery of relational concepts, separately for each level of complexity. The structure common to three calculated models is shown in Fig. 1. Each model had a good fit, as estimated by Bentler's comparative fit index (CFI; its widely accepted criterion value \(=.92\) ) and the standardized root mean square residual (SRMSR; the criterion value \(=.05\) ). For all models, CFIs surpassed .965 , and SRMSRs were below .035. Complex span measures' loadings on WMC variable were high ( \(>.667, p<.001\) ), as well as loadings of DREL measures ( \(>.609, p<.001\) ). This data indicates that the structure of models reflected very well the structure of correlations among variables. The comparison of correlations between both latent variables showed that there was no significant difference between the correlations for binary ( \(r=.663, S E=.068, p<.001\) ) and ternary ( \(r=.631, S E=.065, p<.001\) ) relations ( \(\Delta r=-.028\), n.s.),
while discovery of quaternary ones was more weakly correlated with WMC \((r=.477, S E=.071, p<.001)\) than discovery of both binary \((\Delta r=-.186 ., t[235]=2.70 ; p=.004)\) and ternary relations ( \(\Delta r=-.154 ., t[235]=2.30 ; p=.009)\).

Finally, we tested another CFA model, which related the WMC variable to the index of learning that occurred from the letter to the digit version of the DREL test. Because the scores in quaternary conditions approached floor, and thus the difference between them might have poor psychometric parameters, we decided to aggregate indices of learning of ternary and quaternary relations. The model, presented in Fig. 2, had a very good fit (CFI \(=.979, \mathrm{SRMSR}=.035)\). Most importantly, it suggests that the performance of participants displaying more capacious WM deteriorated less on the more abstract version of the test in comparison to less capacious participants ( \(r=.207, p=.002\) ), most probably due to a more effective process of the transfer of the abstract pattern of relations, which had been introduced in the letter part of the test, to its digit version.

\section*{Discussion}

The newly designed DREL test appeared to be a very reliable tool, and scores on DREL responded well to experimental manipulations. The significant drop of the DREL-WMC correlation only for quaternary relations (in comparison to binary and ternary ones) seems to provide more support for Doumas et al.'s (2008) model than to Halford et al.'s (1998) model. Moreover, not only quaternary relations were very difficult to learn ( \(24.5 \%\) accuracy), as the latter model predicts, but also ternary relations were rarely found (34.8\%), though according to that model they should well fit in WMC of most of participants. In contrast, people displayed fair performance only in cases of binary relations (75.0\%), and that fact better corresponds to Doumas et al.'s (2008) assumption telling that during relational learning (but not when processing relations) even as few as two role-filler representations may occupy the whole available capacity. The study provided data convergent with Lewandowsky (2011) results, though moving beyond ternary relations to newly introduced quaternary condition suggests that relational learning is not uniformly linked to WMC with regard to the complexity of relations being learned.


Figure 2: The CFA model linking WMC to relational learning (a difference in scores between two parts of DREL). The same graphical symbols were used as in Fig. 1.

It must be acknowledged, however, that due to the floor effect in the quaternary condition, a possible alternative explanation of the drop in the value of the DREL-WMC correlation coefficients might appeal to a possible worse psychometric usefulness of scores on quaternary relations. However, this is an unlikely explanation, because of relatively low values of the \(95 \%\) confidence intervals [. 338 .616] for that correlation, comparable to the respective intervals regarding binary and ternary conditions, indicating that all three correlation coefficients have been estimated with a similar precision. Nevertheless, in order to be able to draw firm conclusions on the issue of which model best explains WM contribution to relational learning, the present results should be replicated with a similar method, but one yielding relatively higher scores in the quaternary condition.

Another new result brought by the present study pertains to the fact that not only some general ability to discover relational concepts correlated - though with a varied strength depending on the complexity of those concepts with WMC, but WMC predicted also the amount of transfer of relational knowledge from one task to another. Although whole our test was strongly dependent on WM resources, we accounted for this fact by subtracting the initial (i.e., general) performance on the task, from the final performance, thus measuring the sheer increase in effectiveness of relational thinking during the coping with the test. It appeared that more capacious WM allows for better learning of abstract relational structures and more effective application of them to new, but analogous, situations. This observation seems to be an interesting challenge for existing models of analogy-making and relational learning, and has potentially profound practical (e.g., educational) implications.

\section*{Summary}

This study provided another evidence for the thesis that mechanisms of WM impose substantial constraints on human complex cognition, especially its core component: relational thinking. Understanding those constraints by developing computational models of thinking within WM is one of the crucial current focuses in cognitive science. This study seems to contribute to those efforts by presenting data supporting those models (e.g., Doumas et al., 2008) which predict that WM resources may be exceptionally loaded during the acquisition of relations, in comparison to a lesser load predicted in situations requiring only transformations and manipulations of relational representations which have already been learned.

\section*{Acknowledgments}

This work was sponsored by The National Science Centre (NCN) of Poland (grant no. N106 417140). We thank Krzysztof Cipora, Dominika Czajak, and Jolanta Wojcik for conducting the experiment and gathering the data.

\section*{References}

Chuderski, A., Andrelczyk, K., \& Smolen, T. (2013). An oscillatory model of individual differences in working
memory capacity and relational integration. Cognitive Systems Research. doi:10.1016/j.cogsys.2012.12.005
Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., \& Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. Psychonomic Bulletin \& Review, 12, 769-786.
Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. Behavioral and Brain Sciences, 24, 87-114.
Doumas. L. A. A., Hummel, J. E., \& Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. Psychological Review, 115, 1-43.
Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. Nature, 407, 630-633.
Feldman, J. (2006). An algebra of human concept learning. Journal of Mathematical Psychology, 50, 339-368.
Gentner, D., \& Kurtz, K. (2005). Relational categories. In W. K. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman \& P. W. Wolff (Eds.), Categorization inside and outside the lab (pp. 151-175). Washington, DC: APA.
Goodwin, G. P., \& Johnson-Laird, P. N (2011). Mental models of Boolean concepts. Cognitive Psychology, 63, 34-59.
Halford, G. S., Baker, R., McCredden, J. E., \& Bain, J. D. (2005). How many variables can humans process? Psychological Science, 16, 70-76.
Halford, G. S., Wilson, W. H., \& Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. Behavioral and Brain Sciences, 21, 803-864.
Halford, G. S., Wilson, W. H., \& Phillips, S. (2010). Relational knowledge: the foundation of higher cognition. Trends in Cognitive Sciences, 14, 497-505.
Hummel, J. E., \& Holyoak, K. J. (2003). A symbolicconnectionist theory of relational inference and generalization. Psychological Review, 110, 220-264.
Kemp, C., Goodman, N. D. \& Tenenbaum, J. B. (2008). Learning and using relational theories. Advances in Neural Information Processing Systems, 20.
Lewandowsky, S. (2011). Working memory capacity and categorization: Individual differences and modeling. Journal of Experimental Psychology: Learning, Memory, and Cognition, 37, 720-738
Luck, S. J., \& Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. Nature, 390, 279-281.
Martínez, K., Burgaleta, M., Román, F. J., Escorial, S., Shih, P. C., Quiroga, M \({ }^{\text {a }}\) A., \& Colom, R. (2011). Can fluid intelligence be reduced to 'simple' short-term storage? Intelligence, 39, 473-480.
Oberauer, K., Süß, H.-M., Wilhelm, O., \& Sander, N. (2007). Individual differences in working memory capacity and reasoning ability. In A. R. A. Conway, C. Jarrold, M. J. Kane, A. Miyake, \& J. N. Towse (Eds.), Variation in working memory (pp. 49-75). Oxford: Oxford University Press.
Shepard, R. N., Hovland, C. L., \& Jenkins, H. M. (1961). Learning and memorization of classifications. Psychological Monographs: General \& Applied, 75, 1-42.

\title{
Face age and social status exert different modulatory effects on gaze following behaviour
}

\author{
Francesca Ciardo (francesca.ciardo@unimore.it) \\ Department of Communication and Economics, Viale Allegri, 9 42121 Reggio Emilia, Italy
}

\author{
Barbara F.M. Marino (barbara.marino@unimib.it) \\ Department of Psychology, Edificio U6 Piazza dell'Ateneo Nuovo, 1 20126, Milano, Italy
}

\author{
Angela Rossetti (a.rossetti5@campus.unimib.it) \\ Department of Psychology, Edificio U6 Piazza dell'Ateneo Nuovo, 1 20126, Milano, Italy \\ Rossana Actis-Grosso (rossana.actis@unimib.it) \\ Department of Psychology, Edificio U6 Piazza dell'Ateneo Nuovo, 1 20126, Milano, Italy
}

\author{
Paola Ricciardelli (paola.ricciardelli@unimib.it) \\ Department of Psychology, Edificio U6 Piazza dell'Ateneo Nuovo, 1 20126, Milano, Italy
}

\begin{abstract}
The aim of the present study was to investigate whether face age and social status information associated with faces have different effects on gaze following behaviour as an index of joint attention. Participants were instructed to perform goaldirected saccades towards a peripheral stationary target, while a task-irrelevant face with averted gaze was presented. Faces of three different age groups (younger adults; middle-aged adults; and older adults) were associated with fictional résumés which could describe distracters as high or low social status people. Results showed that face age affected both saccade accuracy and latencies. Social status did not have an effect on accuracy and only affected correct saccades with higher latencies by modulating the face age effect. It is argued that the overt orienting of joint attention could be affected both by perceptual and higher order socio-cognitive factors, but at different stages of processing.
\end{abstract}

Keywords: Joint attention, Social Status, Face Age, Gaze Following, Social cognition, Automaticity.

\section*{Introduction}

Understanding what a co-specific sees is necessary for social cognition. The gaze of others allows the rapid extraction of socially relevant information such as their mental and attentional states (i.e. the focus of their attention; e.g. Baron-Cohen 1995), and allows us to understand and predict their future actions (e.g. Pierno et al. 2006; Innocenti et al., 2012). Several studies have shown that perceiving averted gaze leads the observer to automatically shift his/her attention in the same direction or towards the same object that the other person is looking at (Friesen and Kingstone, 1998; Driver et al., 1999; Frischen, Bayliss and Tipper, 2007). For example, an uninformative cue by a centrally-
presented face gazing to one-location reliably reduces reaction time to targets presented peripherally at the location consistent with the gaze (i.e. gaze cueing effect, e.g. Driver et al., 1999).

The automatic shift of attention in the direction of another person's gaze - known as joint or social attention, can be achieved both overtly, through eye movements (gazefollowing behaviour), and/or covertly without eye movements. Gaze following behaviour (i.e. overt orienting of joint attention), which is considered an early and a direct index of joint attention orienting, is present early in development (Morales et al., 1998; Mundy and Newell, 2007) in humans and studies have shown that many species, including non-human primates, orient gaze in the direction of a co-specific's gaze and use it for interaction (for a review see Shepherd, 2010).

In adults the automatic nature of gaze following behaviour has been shown by Ricciardelli and colleagues (2002) who by using an oculomotor task reported a significant increase in the number of erroneous saccades matching the direction of the distracting gaze (gaze following errors, GFE). This was taken as evidence that perceiving a gaze shift can interfere with the execution of an oculomotor task by affecting oculomotor programming.

However, recent studies have shown that attention orienting driven by gaze is likely to be a product of both stimulus-driven and top-down attentional mechanisms (e.g. Greene et al., 2009). Therefore, modulatory effects on joint attention should be possible. It is more likely, in fact, that some gaze shifts are more important than others depending on face features, environment relevance or current task.

Age is known to be one of the sources of information that is rapidly extracted from faces and is an important dimension that influences how face are attended to (e.g. Slessor et al., 2010), encoded and retrieved from memory (Wiese et al., 2008). Several studies, using different kinds of face processing-task, reported that young adults show an advantage for faces within their own age group compared to elderly faces (for a review see Anastasi \& Rhodes, 2012). In a recent study, Slessor and colleagues (2010) reported a greater gaze cueing effect in younger adults for own-age face distracters than for distracters with elderly faces.

Moreover, recent studies have shown that the automatic and reflexive nature of gaze-mediated attentional orienting can be modulated by a number of high-order cognitive variables, such as the task context (Ricciardelli et al., 2012), social identification (Cesario, 2006; Liuzza et al., 2011), social status (Dalmaso et al., 2012, Shepherd et al., 2006), emotional expression (Tipples, 2006; Bonifacci et al., 2006) and familiarity (Deaner et al., 2007). In particular, Dalmaso and colleagues (2012) showed a greater gaze cueing effect for faces associated with high-status information. Although these authors (2012) did not report differences related to face age (younger vs. older adults) on gaze mediated attention orienting, there is evidence of both differences in interference from emotional faces of different ages (Ebner and Johnson, 2010), and of face-age effects in overt orienting of attention (Ciardo et al., 2012).

A possibility that has not been investigated before is whether they affect overt orienting of joint attention differently when combined together. This stems from the different nature of face age and the information regarding social status associated with the face (perceptual vs. cognitive). In particular, it is reasonable to expect that their distracting/cueing effect may vary as a function of task accuracy and response speed. In other words, given that age is a perceptual feature that is extracted rapidly from the face, one may expect that it affects early stages of saccade programming and the execution of eye movements with lower latencies. By contrast, since the processing of social status information is a more complex and time-consuming higher-order cognitive process, it should play a role later on and its effect should be more evident, for example, in the execution of eye movements with higher latencies.

In the present study we tested this hypothesis by investigating in young human adults the impact of distracting face age and associated social status information on performance in a goal-directed oculomotor task.

\section*{Methods}

\section*{Participants}

Thirty-two right-handed undergraduates ( 23 female, 9 male, mean age \(=22.8\) years, \(\mathrm{SD}=2.0\) ) from the University of Milano-Bicocca participated, in exchange for course credits. All had normal or corrected to normal vision and were unaware of the experiment's purpose. The study was
conducted in accordance with the Declaration of Helsinki and approved by the local ethical committee.

\section*{Stimuli}

Grayscale photographs \((7.98 \times 15.76\) degrees of visual angle) of the faces of 4 younger adults ( 2 females and 2 males, age range: \(18-23\) years), 4 middle-aged adults ( 2 females and 2 males, age range: 34-40 years), and 4 older adults ( 2 females and 2 males, age range: 74-85 years), bearing a neutral expression and a straight gaze, were used. All the photos were taken from the Productive Aging Lab Face Database (Minear and Park, 2004). The gaze direction of each photo was manipulated using Adobe Photoshop, creating face pictures with gaze-averted 0.75 degrees of visual angle both to the left and to the right.

\section*{Procedure}

The experiment was carried out in a sound-attenuated room, dimly illuminated. Participants sat in front of a 19-inch LCD monitor (Samsung Syncmaster 943; \(1280 \times 1024\) pixels; 60 Hz ), with their head supported by a chin rest in order to maintain a stable eye-to-screen distance of 50 cm . At the beginning of the experimental session, participants were invited to read 12 fictional résumés associated with the photographs selected as stimuli. The résumés indicated either a relatively high social status or a relatively low social status. Social status was mainly related to educational/professional information (e.g., high status: She was recently admitted to the faculty of Medicine/ He is a Public Prosecutor of the Supreme Court of Palermo; low status: He did not complete Secondary school/She was dismissed as a worker for incompetence; for a similar manipulation see Dalmaso et al., 2012). The résumés could be considered as brief biographies (hereafter biography). Participants were randomly divided into two groups. For participants in the first group, 6 faces (a male and a female of each age range) were displayed along with biographies indicating a relatively high social status, and the remaining 6 faces to biographies indicating a relative low social status. For participants in the second group, the same faces were displayed along with biographies indicating the opposite social status to that used for the first group. The biography presentation order was randomized. After the biography presentation, the associations created by participants between social status information and faces were tested by means of a true/false questionnaire composed of 12 items (one for each biography). The items were randomly selected from one of two lists (one containing true items and the other containing false items). Item presentation order was randomized. If participants gave a wrong answer to an item, they were immediately presented again with the biography to which the item was related. At the completion of the questionnaire, the biography procedure was restarted for those participants whose accuracy was less than \(90 \%\). Biography presentation and response collection were controlled using the software package E-Prime2 (Psychology Software Tools, Inc.).

Having successfully completed the biography procedure, participants took part in an instructed saccadic eye movement task (Figure 1). Each trial started with the presentation of a black fixation circle (diameter: 0.51 degrees of visual angle) centrally presented on the betweeneyes point of a stimulus face, bearing a straight gaze, on a grey background. The face was flanked by two black target circles (diameter: 0.89 degrees of visual angle), one to the left and the other to the right of the horizontally aligned fixation circle (eccentricity: 10.66 degrees of visual angle). After a delay of 1500 ms , the color of the fixation turned either green or red. 100 ms before the fixation color change, the stimulus face bearing the straight gaze was replaced by the same face with the gaze averted either to the left or to the right. This face replacement created a dynamic gaze, shifting towards the left or the right target. Participants were required to perform a fast and accurate saccade towards the left or right target, depending on the change in color of the fixation. The correspondence between color instruction and saccade direction was inverted for half of the participants. The direction of the dynamic gaze could be congruent or incongruent with the instructed direction. Since it was task irrelevant participants were explicitly instructed to ignore the distracting face. The stimulus face, the fixation and the two targets remained on the screen until a response was given. Immediately after a response was given a new trial was presented.


Figure 1: Experimental procedure.
Participants performed a training block, comprising 12 trials, and a test block, comprising 240 trials, with each of the 12 faces being randomly presented 20 times. In 120 test trials ( 10 replications of the male and the female belonging to each age range and social status), the direction of the dynamic gaze was congruent with the instructed direction. In the remaining 120 test trials ( 10 replications of the male and the female belonging to each age range and social status), the direction of the dynamic gaze was incongruent with the instructed direction.

The participants' eye positions and movements were recorded monocularly in real-time by an infrared video gaze tracking system (EyeLink II, SR Research Ltd., Mississauga, Ontario, Canada). For all participants, we
recorded the movement of the dominant eye. Stimulus generation and presentation were controlled by the SR Research Experiment Builder software (version 1.10.56). Throughout the test block, participants took a break every 80 trials (a total of 3 breaks). During the second break, the biography procedure was repeated to maintain the association between stimulus faces and their fictional social status.

At the end of the experimental session, participants were asked to rate the age of each distracting face with a value between 1 and 99 , in order to verify that they perceived the faces as belonging to the 3 age groups of interest (age manipulation check). Participants were also asked to rate the social status associated during the biography procedure to each distracting face with a value between 1 and 5 (social status manipulation check).

The experiment used a \(3 \times 2 \times 2\) repeated measures factorial design with Distracter Age (younger adults, middle-aged adults, and older adults), Distracter Social Status (high and low), and Congruency between gaze direction and instructed direction (congruent and incongruent) as the within-subjects variables.

\section*{Results}

\section*{Age manipulation check}

Age rating scores were entered in a one-way repeated measures ANOVA with Distracter Face (DF1, DF2, DF3, DF4, DF5, DF6, DF7, DF8, DF9, DF10, DF11, and DF12) as the within-subjects factor. The analysis revealed that the effect of Distracter Face was significant \([F(11,341)=734.9\), \(M S=17215.9, p<.001]\). Post-hoc tests with Bonferroni correction showed significant differences (all \(p s<.05\) ) among distracter faces belonging to different age manipulation levels (i.e. younger adults, middle-aged adults, and older people). Furthermore, post-hoc tests revealed that the faces belonging to each age manipulation level did not differ from each other, confirming our manipulation.

\section*{Social status manipulation check}

A two-way mixed ANOVA, with Distracter Face (DF1, DF2, DF3, DF4, DF5, DF6, DF7, DF8, DF9, DF10, DF11, and DF12) as the within-subjects factor and Subject Group (Group 1 vs. Group 2) as the between-subject factor, was used to determine whether or not participants considered the distracter faces as having the same social status as in the biography procedure. The analysis revealed that the interaction between Distracter Face and Subject Group was significant \([F(11,330)=107.97, M S=68.126, p<.001\), Figure 2b]. Post-hoc tests, performed as before, showed significant differences between distracter faces associated with biographies emphasizing different social status levels (high vs. low). Furthermore, post-hoc tests revealed that the faces associated with biographies belonging to the same social status level did not differ from each other, confirming our manipulation.

\section*{Saccadic eye movement errors}

Practice trials were discarded from the analyses. Saccadic eye movements were defined as correct if landing within \(\pm 2\) degrees of visual angle of the instructed target along the horizontal dimension. Saccadic eye movements landing within \(\pm 2\) degrees of visual angle of the non-instructed target along the horizontal dimension were defined as Gaze Following Errors (GFE) in the incongruent trials (i.e. incongruency between distracter's gaze direction and instructed direction), or as Generic Errors (GE) in the congruent trials (i.e. congruency between distracter's gaze direction and instructed direction). Saccadic eye movements landing outside \(\pm 2\) degrees of visual angle of either the instructed or not-instructed target along the horizontal dimension were defined as Saccades to Nothing Errors (SNE, \(14.5 \%\) of total trials) and were excluded from the analysis.

The first focus of interest was the difference between GFE and GE since it provides a direct and early measure of the automatic tendency to follow the distracting gaze direction. Mean percentages of GFE and GE across subjects were computed for each combination of Distracter Age and Distracter Social Status. These data were entered in a threeway repeated measures ANOVA, with Error Type (GFE and GE), Distracter Age (younger adults, middle-aged adults, and older adults), and Distracter Social Status (high and low) as within-subjects factors. Post-hoc comparisons were performed using the Duncan's test with an alpha level of .05. The analysis revealed that the main effect of Error Type was significant \([F(1,31)=35.59, M S=10385.44, p<.001]\), indicating that participants made more GFE than GE (13.40 \(\%\) vs. \(2.99 \%\) ). This confirmed the automatic tendency of participants to follow the distracter's gaze (for similar results see e.g., Ricciardelli et al., 2002). The analysis also showed a significant main effect of Distracter Age [ \(F(2,62)\) \(=6.01, M S=244.37, p<.005]\) and a significant Error Type \(\times\) Distracter Age interaction \([F(2,62)=12.80, M S=833.07\), \(p<.001]\), indicating that the age of the distracter critically modulated GFE but not GE. Specifically, GFE measured for middle-aged distracters (17.25\%) were higher than GFE measured for both younger ( \(13.55 \%, p<.02\) ) and older distracters ( \(9.39 \%, p<.001\) ), which were also significantly different from each other \((p<.006)\). No other main effects or interactions were significant.

\section*{Saccadic eye movement latencies}

The second focus of interest was the reaction times (RTs) of correct saccadic eye movements, since they could provide an indirect measure of the interference/cueing effect of the distracting gaze. Indeed, although people may be able to suppress the automatic tendency to make saccades in the direction of the distracting gaze, one might expect a higher latency for saccades in the incongruent trials than in the congruent trials. Moreover, if the social status of the distracter exerts an effect on joint attention through a more cognitive/higher-level mechanism than that related to the age of the distracter, as we had hypothesized, then a
different modulation of saccadic eye movements over time should be observed by these two factors. Specifically, the magnitude of the distracter social status effect should increase as the latency of correct saccadic eye movements increases, whereas the magnitude of the distracter age effect should decrease.

To this end, we computed median RT values of correct saccades for the first to the second bin of the individual rank-ordered raw data, separately for each combination of Distracter Age, Distracter Social Status, and Congruency between gaze direction and instructed direction. One subject was excluded from the analysis since the number of his correct saccades was not sufficient to appropriately compute median values of RTs for each combination of the experimental factors. An index of the interference effect of the distracting gaze was then obtained by subtracting the median RT values in the congruent trials from the median RT values in the incongruent trials for each of the Distracter Age and Distracter Social Status conditions, and each participant. These data were entered in a three-way repeated measures ANOVA, with Distracter Age (younger adults, middle-aged adults, and older adults), Distracter Social Status (low and high), and Bin (first and second) as within subjects-factors. Post-hoc comparisons were performed using the Duncan's test with an alpha level of .05 . The ANOVA showed a main effect of Distracter Age \([F(2,60)=\) 8.56, \(M S=10482, p<.001\) ], indicating a higher interference index for both the younger ( 30.15 ms , effect size \(=.56)\) and middle-aged distracters ( 23.94 ms , effect size \(=.44)\) compared to the older distracters \((12.05 \mathrm{~ms}\), effect size \(=.21, p<.001, p<.01\), respectively). In addition, the analysis showed a significant two-way interaction between Distracter Age and Distracter Social Status \([F(2,60)=4.35, M S=5150, p<.02]\), and a significant three-way interaction between Distracter Age, Distracter Social Status and Bin \([F(2,60)=3.64, M S=2441, p<.04]\), Noticeably, as specifically suggested by the three way interaction (Figure 2), the social status of the distracter had an effect on saccadic eye movements with higher latency only. Indeed, post-hoc comparisons showed that, for trials belonging to the second bin, younger distracters with a low social status produced a higher interference index (42.76 ms , effect size \(=.64\) ) than the same distracters with a high social status ( 24.32 ms , effect size \(=.34, p<.02\) ). By contrast, middle-aged distracters produced a higher interference index when they were associated with a high social status ( 33.83 ms , effect size \(=.45\) ) than a low one \((9.93 \mathrm{~ms}\), effect size \(=.13, p<.01)\). No difference between high and low social status was found for older distracters. For trials belonging to the first bin, post-hoc comparisons indicated a higher interference index for both the younger (low status \(=29.10 \mathrm{~ms}\), effect size \(=.68\); high status \(=24.41\) ms , effect size \(=.59\) ) and middle-aged distracters (low status \(=23.74 \mathrm{~ms}\), effect size \(=.52\); high status \(=28.26 \mathrm{~ms}\), effect size \(=.66\) ) compared to older distracters (low status \(=\) 9.61 ms , effect size \(=.21\); high status \(=14.88 \mathrm{~ms}\), effect size \(=.31\) ), independent of social status (all \(p \mathrm{~s}<.05\) ).


Figure 2: Mean index of the interference effect exerted by the distracting gaze (median RT values of correct saccades in the incongruent trials minus median RT values of correct saccades in the congruent trials) as a function of Distracter Age (younger, middle-aged, and older adults), separately for
each Distracter Social (and Bin. Error bars represent the standard errors of means across participants.

Finally, to confirm the finding that the distracter social status modulates joint attention only at a later stage, we focused on the RTs of GFE. Since no effect of the distracter status was found on GFE percentage, one might predict that GFE latencies squarely match those of correct saccades which were placed in the first bin. To this end, we computed the number of GFE with latencies falling in the correct saccades' first bin, separately for each subject. A chi-square test revealed that GFE had latencies which fell in the first bin more frequently ( \(85.3 \%\) ) than would be expected ( \(\chi^{2}=\) \(110.75, d f=30, p<.0001\) ), further supporting the idea that the social status of the distracter exerts a high-level effect on gaze following behaviour.

\section*{Discussion}

In this study, we investigated the impact of face age and social status information on gaze following behaviour in young adults performing an oculomotor task. This was achieved by presenting distracters of different ages whose faces were associated with a fictional biography which could describe the distracter as a high or low status person.

Our results confirmed the automatic nature of gaze following behaviour as the percentage of gaze following errors was higher than the percentage of generic errors (Ricciardelli et al., 2002; 2012). Interestingly, participants made less GFE with older distracters compared to all other distracters. Older distracters also interfered less with the execution of correct saccades, suggesting that older distracting faces are easier to ignore. This result is in line with previous studies that investigated the effect of distracter's age on gaze cueing in young adults (Slessor et al., 2010; Ebner et al. 2010), and reported less distracting effect of averted gaze for elderly faces. It has been proposed that young adults may find it easier to ignore gaze cues from elderly distracters as they are less familiar with their facial features (Deaner et al., 2007). Similar differential results have been observed also for face recognition and processing (for a review see Anastasi \& Rhodes, 2012).
In addition, we found that young adults made more GFE with middle-aged distracters than with younger distracters, indicating a general other-age bias on gaze following
behaviour, rather than an own-age bias. The lack of an ownage bias was confirmed also by the results relative to saccadic eye movement latencies. The occurrence of a super-ordinate categorization (Gaertner \& Dovidio, 2000) of younger and middle-aged distracters may provide a viable explanation of this unpredicted pattern of results. According to Gaertner and Dovidio's (2000) common in-group identity model, when people perceive others as out-group members on the basis of a certain identity cue, and, simultaneously, can form, together with these out-group members, a common super-ordinate super-ordinate group on the basis of another identity cue, the favouritism these people have for their in-group members would be redirected toward outgroup members included within the super-ordinate superordinate group. Therefore, we can surmise that participants of our study classified younger and middle-aged distracters as in-group and out-group members, respectively, on account of face age estimation, and as members of the same super-ordinate group on account of another facial cue estimation, such as facial similarity.

The novel result of our study is that face age, but not social status information, affects GFE. Social status has an effect on saccadic eye movements with higher latency only. Previous studies investigating the role of social status on gaze-mediated covert orienting of attention (Dalmaso et al., 2012) reported that high status individuals produced a stronger gaze cueing effect (Dalmaso et al., 2012); by contrast, our results indicate that in young adults gaze cueing is facilitated (i.e. slower saccadic reaction times) by own-age low-status distracters. The contrast between our results and those of Dalmaso et al.'s (2012) study could be due to differences in the stimuli used. In Dalmaso et al.'s (2012) work, photos depicting only male faces were used as stimuli, while in our study we used both female and male distracters. Indeed, gender is an element from which social status could also be perceived implicitly, and it was established that male faces are perceived to be more dominant (i.e. higher social status) than female faces (Jones et al., 2010). It is possible that the influences on gaze following behaviour of information about social status may also depend on how it is induced, i.e. whether explicitly or implicitly. Alternatively, own-age low status faces could represent a threat to the social identity of high-status groups (Scheepers et al., 2004); as a source of threat, such lowstatus individuals should be monitored continually.

In conclusion, our findings extended previous joint attention studies by indicating that the overt orienting of attention driven by gaze could be differently modulated both by perceptual features and high-order socio-cognitive factors of the seen face. Taken together, our results indicate that a perceptual manipulation of an identity cue (i.e. age), exerts an effect on gaze following parameters in the early stages of saccade programming and execution. By contrast, higher-order manipulation of another identity cue (i.e. social status) affects gaze following parameters at the later stages of saccade programming and execution. This could be due either to the kind of manipulation (perceptual vs. higher-
order) or to the kind of identity cue (age vs. social-status). Future studies are needed to clarify the role of different identity cues (such as age, race, gender, celebrity or political affiliation) and manipulations on the time course and patterns of gaze following parameters, in order to explore the possibility that different identity cues fit into a few broad categories or continua, such as identity valence, which have different effects on overt orienting of attention.

\section*{Acknowledgments}

We thank Natale Vincenzo Maiorana and Roberta Capellini for their help in collecting data. This research was supported by a grant from the Ministero dell'Istruzione, dell'Università e della Ricerca (PRIN project 2008ZN5J5S_004).

\section*{References}

Anastasi, J. S., \& Rhodes, M. G. (2012). The own-age bias in face recognition: A meta-analytic and theoretical review. Psychological Bulletin, 138, 146-174.
Baron-Cohen, S. (1995). Mindblindness: An Essay on Autism and Theory of Mind. MIT Press, Cambridge.
Bonifacci, P., Ricciardelli, P., Lugli, L., \& Pellicano, A. (2008). Emotional attention: Effects of emotion and gaze direction on overt orienting of attention. Cognitive Processing, 9, 127-133.
Cesario, J. F. (2006). Regulatory fit from nonverbal behaviours: How source delivery style influences message effectiveness. Dissertation Abstracts International, 67, 2276.

Ciardo, F., Rossetti, A., Actis-Grosso, R., \& Ricciardelli, P. (2012). How old are you? The effect of face age on gazefollowing behaviour. Perception, 41 (supplement), 129.
Dalmaso, M., Pavan, G., Castelli, L., \& Galfano, G. (2011). Social status gates social attention in humans. Biology Letters, 16, 450-452.
Deaner R. O., Shepherd, S. V., \& Platt, M. L. (2007). Familiarity accentuates gaze cuing in women but not men. Biology Letters, 3, 64-67.
Driver, J., Davis, G., Ricciardelli, P., Kidd, P., Maxwell, E., \& Baron-Cohen, S. (1999). Gaze Perception Triggers Reflexive Visuospatial Orienting. Visual Cognition 6, 509-540.
Ebner, N. C., \& Johnson, M. K. (2010). Age-group differences in interference from young and older emotional faces. Cognition and Emotion, 24, 1095-1116.
Friesen, C., \& Kingstone, A. (1998). The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. Psychonomic Bulletin \& Review, 5, 490-495.
Frischen, A., Bayliss, A. P., \& Tipper, S. P. (2007). Gaze cueing of attention: Visual attention, social cognition, and individual differences. Psychological bulletin, 133, 694724.

Gaertner, S. L., \& Dovidio, J. F. (2000). Reducing intergroup bias: The common ingroup identity model. Philadelphia: Psychology Press.

Greene, D. J., Mooshagian, E., Kaplan, J., Zaidel, E., \& Iacoboni, M. (2009). The neural correlates of social attention: automatic orienting to social and non-social cues. Psychological Research, 73, 499-511.
Jones, B. C., DeBruine, L. M., Main, J. C., Little, A. C., Welling, L. L., Feinberg, D. R. \& Tiddeman, B. P. (2010). Facial cues of dominance modulate the short-term gaze cuing effect in human observers. Proceedings of Royal Society B 277, 617-624.
Innocenti, A., De Stefani, E., Bernardi, N. F., Campione, G. C., \& Gentilucci, M. (2012). Gaze Direction and Request Gesture in Social Interactions. PLoS ONE, 7, e36390.
Liuzza, M. T., Cazzato, V., Vecchione, M., Crostella, F., Caprara, G. V., \& Aglioti, S. M. (2011). Follow my eyes: the gaze of politicians reflexively captures the gaze of ingroup voters. PLoS One, 6, e25117.
Minear, M., \& Park, D. C. (2004). A lifespan database of adult facial stimuli. Behaviour Research Methods, Instruments, \& Computers, 36, 630-633.
Morales, M., Mundy, P., \& Rojas, J. (1998). Following the direction of gaze and language development in 6-montholds. Infant Behaviour and Development, 21, 373-377.
Mundy, P., \& Newell, L. (2007). Attention, joint attention, and social cognition. Current Directions in Psychological Science, 16, 269-274.
Pierno, A. C., Becchio, C., Wall M. B., Smith A. T., Turella, L., \& Castiello, U. (2006). When gaze turns into grasp. Journal of Cognitive Neuroscience, 18, 2130-2137.
Ricciardelli, P., Bricolo, E., Aglioti, S. M., \& Chelazzi, L. (2002). My eyes want to look where your eyes are looking: Exploring the tendency to imitate another individual's gaze. NeuroReport, 13, 2259-2263.
Ricciardelli, P., Carcagno, S., Vallar, G., \& Bricolo, E. (2012). Is gaze following purely reflexive or goal-directed instead? Revisiting the automaticity of orienting attention by gaze cues looking. Experimental Brain Research, 224, 93-106.
Scheepers, D., \& Ellemers, N. (2004). When the pressure is up: The assessment of social identity threat in low and high status groups. Journal of Experimental Social Psychology, 41, 192-200.
Shepherd, S. V., Deaner, R. O., \& Platt, M. L. (2006). Social status gates social attention in monkeys. Current Biology, 16, R119-120.
Shepherd, S. V. (2010). Following gaze: gaze-following behaviour as a window into social cognition. Frontiers in Integrative Neuroscience, 4, 1-11.
Slessor, G., Laird, G., Phillips, L. H., Bull R., \& Filippou, D. (2010). Age-related differences in gaze following: does the age of the face matter? The journals of Gerontology. Series B, Psychological sciences and social sciences, 65, 536-541.
Tipples, J. (2006). Fear and fearfulness potentiate automatic orienting to eye gaze. Cognition \& Emotion, 20, 309-320.
Wiese, H., Schweinberger, S. R., \& Hansen, K. (2008). The age of the beholder: ERP evidence of an own-age bias in face memory. Neuropsychologia, 46, 2973-2985.

\title{
Towards a Statistical Model of Grammaticality
}

\author{
Alexander Clark, Gianluca Giorgolo, and Shalom Lappin \\ firstname.lastname@kcl.ac.uk \\ Department of Philosophy, King's College London
}

\begin{abstract}
The question of whether it is possible to characterise grammatical knowledge in probabilistic terms is central to determining the relationship of linguistic representation to other cognitive domains. We present a statistical model of grammaticality which maps the probabilities of a statistical model for sentences in parts of the British National Corpus (BNC) into grammaticality scores, using various functions of the parameters of the model. We test this approach with a classifier on test sets containing different levels of syntactic infelicity. With appropriate tuning, the classifiers achieve encouraging levels of accuracy. These experiments suggest that it may be possible to characterise grammaticality judgements in probabilistic terms using an enriched language model.
\end{abstract}

Keywords: enriched language models, probability distribution, grammaticality judgements, probabilistic syntax

\section*{Introduction}

The past two decades have seen a lively debate over whether linguistic knowledge is probabilistic or categorically rulebased in nature (see the papers in Bod, Hay, and Jannedy (2003) for some of this discussion). Given the success of probabilistic accounts of learning, representation, and inference across a wide range of cognitive domains, this debate has considerable importance for the way in which knowledge of language is integrated into our general view of human cognition.

On the classical view of syntax developed within linguistic theory over the past sixty years, the grammaticality and the probability of a sentence are entirely distinct properties with no direct relationship. Chomsky (1957) presents the original argument for the irrelevance of probability in determining grammaticality. \({ }^{1}\) This argument depends on the inability of a simple word n-gram model to predict a distinction in probability between a syntactically well-formed but unlikely (semantically anomalous) sentence like Colorless green ideas sleep furiously, and a word salad like Furiously sleep ideas green colorless. Pereira (2000) shows that a smoothed classbased n-gram model trained on a newspaper corpus predicts a significant distinction in probability between the two sentences.

While it is certainly the case that grammaticality cannot be directly reduced to probability, the question of whether there is a significant correlation between the two remains open and interesting. Our general approach is as follows. We train a smoothed class-based trigram model on a filtered subclass of the BNC. We test this model on two corpora. One is divided into original sentences of part of the BNC and their reversed counterparts. The second consists of a subset of orig-

\footnotetext{
\({ }^{1}\) See Fong, Malioutov, Yankama, and Berwick (2013) for a recent discussion of some of the issues involved in identifying grammaticality with probability of occurrence.
}
inal BNC sentences and their permuted variants in which a word in each sentence is randomly exchanged with another word three positions away from it. These distortions constitute syntactic infelicities. The first case involves gross structural ill-formedness similar to the word salad example, while the second introduces subtler, more local mistakes. We score the test corpora using three alternative conditions. Our binary classifiers predict that a string is well-formed (original) or distorted (either reversed or permuted) on the basis of a score derived through normalising its log probability (logprob) value in various different ways. We also test different standard deviations from the distributional norm in setting cut off points for our binary classifiers In our best cases we obtain an accuracy rate of \(98.9 \%\) for the original-reversal test set, and \(79.1 \%\) for the original-permutted test set.

These results suggest that by looking at the internal components of a probability distribution and the stages through which it is computed we can identify additional information that may be used to specify significant correlations between probability and grammaticality. This opens up an interesting set of research questions on the relationship between speakers' knowledge of the probability distribution for a language and their grammaticality judgements.

\section*{Probability and Grammaticality}

As has often been noted, it is not possible to reduce grammaticality directly to probability. First, short ungrammatical sentences generally receive higher probability values than long, complex grammatical sentences containing words with low frequencies. Second, if one specifies a probability value (or even a range of such values) as the minimal threshold for grammaticality, then one is committed to the existence of a finite number of grammatical sentences. The sum of the probabilities of the possible strings of words in a language sum to 1 , and so at most \(1 / \varepsilon\) sentences can have a probability of at least \(\varepsilon\).

On the other hand, probabilistic inference does appear to be pervasive throughout all domains of cognition (Chater, Tenenbaum, and Yuille (2006)). Moreover, language models do seem to play a crucial role in speech recognition and sentence processing. Without them we would not be able to identify speech sounds, and meaningful syntactic and semantic structures in noisy environments. Finally, grammaticality appears to track speakers' acceptability judgements, and these are, in many cases, graded. Probability provides a natural basis for generating such a gradient (Crocker and Keller (2006)).

Our starting point is a language model: a statistical model that defines a probability distribution over sentences.

We construct a log-linear model, parameterised by some vector of parameters \(\Theta=\left\langle\theta_{1}, \ldots, \theta_{k}\right\rangle\). This framework covers a wide range of different models from n-gram models to PCFGs. \({ }^{2}\)

The probabilities defined by this model cannot be used to define a notion of grammaticality for several reasons. First, as the sentences increase in length, the probability of the sentence will always decrease exponentially, for sufficiently long sentences, while we assume that long sentences can be as grammatical as short sentences. Second, one can often substitute a rarer semantically related word for an open class word of the same POS without affecting grammaticality, but the substitution will reduce probability. Figure 1 shows that the log probabilities for sentences that have been reversed or permuted, and are thus generally ungrammatical, overlap completely with the log probabilities of normal sentences (see the next section for details of the experimental protocols). We need to augment our model with an additional component to convert probability into a score that correlates with grammaticality in an interesting way.


Figure 1: Histograms for the distributions of \(\log\) probabilities under the three conditions.

We use statistical properties of the parameters of the model. In order to compute the probability of a sentence with respect to a model we do calculations on the parameters. For a log linear model, this gives a linear function of certain indicator variables; a weighted sum. To compute a score that correlates with grammaticality, we consider other functions, such as a weighted mean, or a minimum over certain scores.

In a trigram model each parameter will then correspond to the \(\log\) conditional probability of one word, given the two

\footnotetext{
\({ }^{2}\) We use smoothing techniques that, in general, can take the model outside the class of log-linear models, but we pass over this technical detail here.
}
preceding words: \(\theta_{w_{i} \mid w_{i-1} w_{i-2}}\). To compute the probability of a sentence we sum the relevant parameters to obtain the log probability. For a sentence \(\left\langle w_{1}, \ldots, w_{n}\right\rangle\) the log probability is
\[
\log P_{\text {TRIGRAM }}\left(\left\langle w_{1}, \ldots, w_{n}\right\rangle\right)=\sum_{i=1}^{n} \theta_{w_{i} \mid w_{i-1} w_{i-2}}
\]

We take the sequence of relevant parameters \(\left\langle\theta_{w_{1} \mid w_{0} w_{-1}}, \ldots, \theta_{w_{n} \mid w_{n-1} w_{n-2}}\right\rangle\), and, rather than summing them, we perform other computations. We consider the average or the minimum of the set of parameters as alternatives for defining values that correspond to grammaticality.

Our most basic score is the mean of this value, the logprob divided by the word length of the sentence:
\[
\text { Meanlogprob }=\frac{1}{n} \log P_{\text {TRIGRAM }}\left(\left\langle w_{1}, \ldots, w_{n}\right\rangle\right)
\]

This eliminates the dependence of the logprob on the length. Our next score divides the logprob of the original trigram model by the logprob with respect to a unigram model.
\[
\text { Normalised }=\frac{\log P_{\text {TRIGRAM }}\left(\left\langle w_{1}, \ldots, w_{n}\right\rangle\right)}{\log P_{\text {UNIGRAM }}\left(\left\langle w_{1}, \ldots, w_{n}\right\rangle\right)}
\]

This removes the variation in logprob caused by rare lexical items. Note that if the unigram model is uniform (if we had equal numbers of each word in the training corpus), then the log of the unigram model would be a multiple of the length, and so it would reduce to the previous value.

Our third score uses the observation that a sentence with one grammatical error in it is ungrammatical. In order to measure grammaticality we look at the minimum of some score over the parts of the sentence. We take the minimum of the ratio of the log trigram probability to \(\log\) unigram probability.
\[
\text { Minimum }=\min _{i}\left[\frac{\log \theta_{w_{i} \mid w_{i-1} w_{i-2}}}{\log \theta_{w_{i}}}\right]
\]

None of these measures will produce a score which is in the range \([0,1]\), though it would be possible to map them into this range. This value will also vary even for grammatical sentences. The scores will be numbers that are distributed in some way. Figure 2 shows the distribution of these scores for the test data. As this score specifies a continuum of values, we are able to accommodate a gradient notion of grammaticality.

Given these three measures we use various standard techniques to see whether new sentences are anomalous or not. For a collection of naturally occurring grammatical sentences we train our models, and then we consider the distribution of these scores. We estimate the mean and standard deviation of the score. We can then judge new sentences as ungrammatical if they are unusually low in score- more than a few standard deviations away from the mean.

Pauls and Klein (2012) apply a related approach to another problem. They use scores based on the logprob values of a language model to discriminate between grammatical and ungrammatical sentences in order to improve the performance of natural language processing systems.

\section*{Experiments and Results}

For our experiments, we use the standard n-gram language model, which is an instance of a Markov model for sequences. To estimate the probability of a sequence of words \(w_{1} \ldots w_{k}\) we use the chain rule of probability, as in (1).
\[
\begin{equation*}
P\left(w_{1} \ldots w_{k}\right)=P\left(w_{1}\right) P\left(w_{2} \mid w_{1}\right) \ldots P\left(w_{k} \mid w_{1} \ldots w_{k-1}\right) \tag{1}
\end{equation*}
\]

The problem with this approach is that we have to estimate the conditional probability of an extremely large number of possible subsequences. Therefore a common method is to reduce the conditional dependencies to a smaller predefined sequence of a given length \(n\), the so called order of a model. Using this assumption we approximate the components in (1) using (2).
\[
\begin{equation*}
P\left(w_{i} \mid w_{1} \ldots w_{i-1}\right) \approx P\left(w_{i} \mid w_{i-n+1} \ldots w_{i-1}\right) \tag{2}
\end{equation*}
\]

The probability assigned to a sequence of words is given by the product in (3)
\[
\begin{equation*}
P\left(w_{1} \ldots w_{k}\right) \approx \prod_{i=1}^{k} P\left(w_{i} \mid w_{i-n+1} \ldots w_{i-1}\right) \tag{3}
\end{equation*}
\]

A common choice for \(n\), that we adopt for our experiments, is three (trigrams).

The standard strategy to estimate the probability of each ngram is maximum likelihood estimation (MLE), which counts the number of times the n -gram appears in a training corpus, and normalizes the count by the sum of the counts of all \(n\) grams that share the same initial subsequence:
\[
\begin{equation*}
P\left(w_{i} \mid w_{i-n+1} \ldots w_{w-1}\right)=\frac{C\left(w_{i-n+1} \ldots w_{i}\right)}{\sum_{w} C\left(w_{i-n+1} \ldots w\right)} \tag{4}
\end{equation*}
\]

To avoid assigning 0 probability to unseen n -grams (a common case, given the huge number of possible n-grams) we use smoothing or discounting, which transfers a small portion of probability mass from seen n -grams to unseen ones. A large number of smoothing techniques have been proposed in the literature (see Chen and Goodman (1999) for a thorough overview). In our experiments we use a form of interpolated smoothing known as Interpolated Kneser-Ney (Goodman (2001)), which has been shown to give consistently good results with different types of metrics.

To reduce the search space of our language model we also employ clustering, which groups together words that occur in similar contexts. In this way we can better estimate the probability of a word following a certain sequence, given the observations we have made of similar words in the same context. Brown, deSouza, Mercer, Pietra, and Lai (1992) introduced the standard technique for using clustering in language models. The general form of a cluster-based language model is given in equation (5), where \(C_{i}\) is the cluster to which word \(w_{i}\) is assigned to.
\[
\begin{equation*}
P\left(w_{i} \mid w_{i-n+1} \ldots w_{i-1}\right)=P\left(w_{i} \mid C_{i}\right) P\left(C_{i} \mid C_{i-n+1} \ldots C_{i-1}\right) \tag{5}
\end{equation*}
\]

The probability \(P\left(w_{i} \mid C_{i}\right)\) is given by the count of occurrences of \(w_{i}\) divided by the count of occurrences of \(C_{i}\), while the other factor of the product can be estimated with a smoothed model like Interpolated Kneser-Ney. Brown et al. (1992) describes a technique for generating the optimal clustering in a corpus, given a parametrically specified number of classes.

We implemented our own procedures for the training and the assessment of \(n\)-gram language models, using Interpolated Kneser-Ney as the smoothing technique. For clustering we applied the improved version of Brown et al. (1992)'s algorithm described in Liang (2005).

Both the training of the language models and the measurement of their performance in the given tasks are performed on portions of the BNC. The BNC is a heterogenous collection of linguistic data. To obtain a more consistent sample of English we first restricted the available texts by excluding transcriptions of spoken language, poetic texts and technical/scientific material. The corpus used for training and the one used for testing were generated from this subset of the BNC by randomly selecting 600 k sentences for training, and 60k for testing. This gave us a training corpus of slightly less than 13 million words, and a testing corpus of approximately 1.3 million words.

To avoid the problem of unknown words in testing, we reconstructed both the training and the testing corpus. We substituted, in both the training and the test corpus, the POS tag for each word which appears less than five times in the training corpus. This insures that the test corpus vocabulary is a subset of the training corpus vocabulary.

Three different types of test corpus (conditions) were generated. The original condition is left intact, and we assume that it contains only grammatical sentences. The permuted condition is generated from the original by randomly swapping two words, separated by two intervening words, in each sentence. The sentences in this corpus are taken to be less grammatical than those in the original condition. Finally, the third test corpus was produced by reversing the order of the words in the original sentences. This reversed condition is considered to be the most syntactically distorted of the three.

We used a simple binary classifier to measure the performance of our language model in predicting the grammaticality of a sentence. After calculating the three scores (Mean log prob, Normalised, and Minimum) in all three conditions we designed two different binary classifiers that assign a label to every sentence in each condition. The first classifier is a simple threshold set to different values for the mean and the standard deviation of the distributions of the alternative normalised scores for the original condition. For each binary comparison the classifier assigns a label to the sentence \(z\) using the following rule:
\[
c_{1}(z)= \begin{cases}\text { original } & \text { if } \operatorname{score}(z) \geq m-S \cdot s  \tag{6}\\ \text { other } & \text { otherwise }\end{cases}
\]
where \(m\) is the mean for the score in the original condition, \(s\) is the standard deviation and \(S\) is a factor by which we move
the threshold away from the mean. The principle of this classifier is that the normalised logprob scores for ungrammatical sentences will be lower than those for grammatical ones, making it possible to distinguish between the two conditions. We adapted this procedure to distinguish between local ungrammaticality (permutation), and more global ungrammaticality (reversed cases).

The second classifier is a simple linear classifier constructed on the basis of the first one. It combines the information from two different scores. This second classifier uses the following general rule:
\[
c_{2}(z)= \begin{cases}\text { original } & \text { if } \operatorname{score}_{2}(z) \geq-\operatorname{score}_{1}(z)+t_{1}+t_{2}  \tag{7}\\ \text { other } & \text { otherwise }\end{cases}
\]
where \(\operatorname{score}_{1}(z)\) is the first of the scores assigned to the sentence, \(\operatorname{score}_{2}(z)\) is the second one, \(t_{1}\) is the best performing threshold for this specific comparison as found in the case of the first type of classifier for the first score, and \(t_{2}\) is the same kind of threshold for the second score. We simply check whether the two scores are above or below the bisector of the second and the fourth quadrant in the space formed by the two scores, and translated by the best thresholds for the same two scores. The intuition here is, again, that grammatical sentences will have consistently better scores than ungrammatical ones.

We performed experiments using both the standard and the cluster-based language models. For the standard case we trained models using words and part-of-speech tags as tokens. In what follows we report only the results for the cluster-based experiments, as these achieved better accuracy. We used 250 clusters. The language model was trained on the training corpus, and the three scores are computed for the sentences in each condition (original, permuted and reversed). Figure 2 summarises the distributions of the three scores for each condition of the cluster-based language model. It is clear that all scores are reasonably good at distinguishing between the original and the reversed conditions, given the small overlap between the distributions. As expected, the overlap between the original and permuted conditions is much higher. It is also interesting to note that the while in the case of Mean log prob and Normalised score the distributions for all the conditions are roughly normal (with some degree of skewing), the Minimum score gives a more irregular distribution, at least for the ungrammatical cases.

On the basis of these distributions we created the first type of classifier. The results for the two comparisons we performed (original/permuted and original/reversed) are summarised in figure 3. The graphs show the accuracy for each score obtained by varying the \(S\) parameter as described in (6). In our experiments we let \(S\) vary in the interval \([0,2.75]\), using a step interval of 0.25 .

In the case of the original/permuted comparison we obtained the best accuracy ( \(77.3 \%\) ) by using the Normalised score and setting the threshold at 0.75 standard deviations to the left of the mean. However the Minimum score seems

Table 1: Linear classifier accuracy
\begin{tabular}{lrr} 
Accuracy & permuted & reversed \\
\hline Mean log prob + Normalised & 71.2 & 97.9 \\
Mean log prob + Minimum & 77.1 & 97.2 \\
Normalised + Minimum & \(\mathbf{7 9 . 1}\) & 98.1 \\
\hline Threshold classifier baseline & 77.3 & \(\mathbf{9 8 . 9}\)
\end{tabular}
to perform better in general for this comparison, obtaining a maximum accuracy of \(77.1 \%\).

Not surprisingly, all three scores perform very well when distinguishing between the original and the reversed version of the sentence, with accuracies above \(95 \%\). The sharp drop in accuracy in the case of the Minimum score that we observe when setting the \(S\) parameter to 2.75 is due to the spike we have in the case of the reversed condition (see rightmost graph in figure 2).

Table 1 reports the accuracy for the linear classifier that combines the results of two threshold classifiers (with the best single classifier scores listed in the bottom row as a baseline comparison). Despite the simplicity of this linear classifier, we observe an improvement in the original/permuted comparison.

\section*{Error analysis}

It is interesting to analyse the cases where our classifiers fail. We looked at the cases that form the tails of the distributions for the Normalised threshold, as it is this score that gives the best general level of classifier accuracy.

The following ten sentences receive the lowest Normalised score according to our language model for the original condition: interview \(\cdot\) Swims \(\cdot / \cdot\) contracts \(\cdot\) then \(\cdot\) TELEPHONE - 75\% • Hotel deal • mimic each item across • Ian ! 90\% These cases are very marginal English sentences. Their presence in the corpus may well be due to transcription error in the BNC, or to the idiosyncratic nature of the text from which they are extracted. However, other cases of false ungrammatical sentences include perfectly acceptable sentences like the following: Amnesty has been given Greetings Magazine's "Best Charity Card of the Year" award .

For permuted sentences, when we analyse the tail of the distribution, we encounter many cases where the permutation produces the same sentence as the original, because the permuted words are identical. In other cases the permutation generates semantically odd, but otherwise well-formed sentences, as in It should be a match of a humdinger. These are the ten permuted sentences that receive the highest Normalised scores (and they are therefore mislabelled as original): He glanced round the bar from the door. He said that he had not been informed of the dissolution of the National Assembly on Jan. 4. • There 's something I hear you to want. - Sometimes, of course, it does not work. • Don't know, I worry why. I I assure you I'm not. - It should be a match of a humdinger. • She put her hand to her brow. • "Yes, I understand ," said Drew quietly. But there was nothing there.

Finally in the case of reversed sentences, we observe that sentences that are assigned extremely high Normalised scores tend to be proper names. Due to their low frequency in the training corpora, proper names are most likely to be replaced by their POS tag in the training and testing phase. Therefore, the language model cannot distinguish the original and the reversed versions of the sequence, given that they appear identical. Again we report here the ten reversed sentences that receive the highest Normalised score. Terrazze Alle . seven - eight - nine • 2 TN. WOKINGHAM , 3 TN. FARNHAM • Debts : MAIDSTONE • Gloucester / BROCKWORTH - FLAUBERT MME • VALLI FRANKIE • PATEL GARGY • BATTERSEA HORSMAN UDO • REUNITE / JAFFE LUCKY

\section*{Discussion and Conclusions}

Clark and Lappin (2011) propose an outline for a stochastic model of indirect negative evidence. In this outline a function maps the probability value of a string, and a set of properties of the string and of the probability distribution over strings of the language, to a threshold value that gives the minimum frequency with which the string must occur in the primary linguistic data in order to be well-formed. The threshold specifies the normalised minimal expectancy of occurrence for a sentence of a certain type (length, lexical class sequence, etc.). This model provides a language learner with a procedure for querying the data to which he/she is exposed in order to determine the extent to which the absence of a string in the data indicates its ungrammaticality.

Here we effectively invert this strategy. We identify a set of structural properties of a string together with parameters for the distribution of logprob-derived scores, in order to define a grammaticality threshold, which we use to classify strings as grammatical or ill-formed. This model offers a stochastic characterisation of grammaticality without reducing grammaticality to probability. It represents a core element of what speakers know about the syntax of their language through a set of parameters in a model whose values correspond to properties of the modified probability distributions that the model generates.

We are not, of course, suggesting that enriched n-gram models are adequate to express the full content of speakers' syntactic knowledge. However, the fact that simple models of the sort that we have used are able to achieve a relatively high degree of accuracy on wide coverage, domain general grammaticality classification tasks suggests that there is an interesting correlation between properties of the probability distribution over the sentences of a language and a speaker's grammaticality judgements.

Should the correlation prove robust it suggests that grammatical knowledge is, to a significant extent, determined by the stochastic patterns of the primary linguistic data to which speakers are exposed. This result will have significant consequences for both the representation of syntactic competence and the nature of the language acquisition process.

In current work we are exploring this correlation further with more sophisticated language models, different distri-
butional parameters and stochastic classifiers, and test data that includes realistic syntactic infelicities. We are evaluating these models against native speakers' acceptability judgements.

\section*{Acknowledgments}

The research described in this paper was done in the framework of the Statistical Models of Grammaticality (SMOG) project at King's College London, funded by grant ES/J022969/1 from the Economic and Social Research Council of the UK. We are grateful to Rens Bod, Stephen Clark, Aarne Ranta, Khalil Sima'an, and Jelle Zuidema. for helpful comments on an earlier draft of this paper. We also thank our PhD students, Jekaterina Denissova and Monika Podsiadlo, for useful discussion of the experimental design of this work, and for logistical support.

\section*{References}

Bod, R., Hay, J., \& Jannedy, S. (2003). Probabilistic linguistics. MIT Press.
Brown, P. F., deSouza, P. V., Mercer, R. L., Pietra, V. J. D., \& Lai, J. C. (1992). Class-based n-gram models of natural language. Computational Linguistics, 18, 467-479.
Chater, N., Tenenbaum, J., \& Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. Trends in Cognitive Sciences, 10(7), 287-291.
Chen, S., \& Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. Computer Speech \& Language, 13(4), 359-393.
Chomsky, N. (1957). Syntactic structures. The Hague: Mouton.
Clark, A., \& Lappin, S. (2011). Linguistic nativism and the poverty of the stimulus. Malden, MA: Wiley-Blackwell.
Crocker, M., \& Keller, F. (2006). Probabilistic grammars as models of gradience in language processing. In G. Fanselow, C. Féry, M. Schlesewsky, \& R. Vogel (Eds.), Gradience in grammar: Generative perspectives (pp. 227245). Oxford University Press.

Fong, S., Malioutov, I., Yankama, B., \& Berwick, R. (2013). Treebank parsing and knowledge of language. In A. Villavicencio, T. Poibeau, A. Korhonen, \& A. Alishahi (Eds.), Cognitive aspects of computational language acquisition (p. 133-172). Springer Berlin Heidelberg.

Goodman, J. (2001). A bit of progress in language modeling. Computer Speech \& Language, 15(4), 403-434.
Liang, P. (2005). Semi-supervised learning for natural language processing. Unpublished master's thesis, Department of Electrical Engineering and Computer Science, MIT.
Pauls, A., \& Klein, D. (2012). Large-scale syntactic language modeling with treelets. In Proceedings of the 50th annual meeting of the association for computational linguistics (pp. 959-968). Jeju, Korea.
Pereira, F. (2000). Formal grammar and information theory: together again? Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences, 358(1769), 1239-1253.


Figure 2: Histograms for the distributions of sentence scores. Each graph shows the distribution of a single score for the three conditions. The \(x\)-axis represents the value of the score and the \(y\)-axis gives a measure of the frequency with which the score is represented in the data. On the left are the scores given by taking the mean (equivalently normalising by length). In the middle are the scores given by normalising with the unigram probability. On the right are the scores using the minimum condition. These scores still overlap significantly, but much less so than the raw logprobs as shown in Figure 1.


Figure 3: Results for the threshold classifier. The two graphs show two comparisons: original/permuted and original/reversed. The \(x\)-axis represents the different values that control the distance from the mean of the threshold, while the y-axis shows the accuracy expressed in percentages.

\title{
Knowledge Helps: Mechanistic Information and Numeric Evidence as Cognitive Levers to Overcome Stasis and Build Public Consensus on Climate Change
}

\author{
Dav Clark (davclark@berkeley.edu) \({ }^{1}\) \\ Michael Andrew Ranney (ranney@berkeley.edu) \({ }^{1,2}\) \\ Jacqueline Felipe (felipe.jacqueline@gmail.com) \({ }^{2}\) \\ \({ }^{1}\) Department of Psychology, 3210 Tolman Hall; Berkeley, CA 94720-1650 USA \\ \({ }^{2}\) Graduate School of Education, 4533 Tolman Hall; Berkeley, CA 94720-1670 USA
}

\begin{abstract}
A clear, growing consensus indicates an urgent need for humans to reduce the burgeoning effects of global climate change ("global warming" or GW). Apt public instruction seems central to achieving critical behavioral changes, but some researchers suggest that U.S. climate attitudes are doomed to cognitive stasis (i.e., that little will be gained by educating the public). Herein are four studies that counter the stasis view. Our laboratory has previously reported findings that (1) virtually no Americans know the basic climate change mechanism, yet it (2) is quickly learned (in a few minutes, e.g., with a 400 -word text), which (3) increases climate change acceptance. Below, Studies 1 and 4 replicate and extend these results to demonstrate (a) efficacy with an online presentation and broader populations and (b) retention up to a month after learning the mechanism. Studies 2-4 explore roles for germane numerical information using estimation with feedback. Study 2 shows that (d) misleading, cherry-picked, statistics can decrease climate change acceptance (and shake metacognition), while Studies 3 and 4 show that (e) surprising scientific information must be presented with care for it to foster beliefs in line with climate science's consensus. In sum, contrary to unnecessarily pessimistic (and correlational) "stasis" arguments, highly germane science information can clearly change the public's understandings and opinions.
\end{abstract}

Keywords: cognitive change, science education, explanation, climate change, global warming, acceptance, mechanism.

\section*{Climate Change as a Behavioral Problem}

Our atmosphere's carbon dioxide \(\left(\mathrm{CO}_{2}\right)\) concentration is higher now than in any of the past 15 million years (World Bank, 2012). Global warming (GW) akin to recent trends last occurred over 17 million years ago, when a \(3-4^{\circ} \mathrm{C}\) gain occurred over 1,500,000 years. Standard models show that continuing our current behavior will yield similar warming in just 100 years. In previous warming periods of this magnitude, widespread extinctions occurred. With imminent warming \(10,000+\) times faster than historical timescales, the biological systems we depend upon (e.g., for food) will clearly be severely impacted (Barnosky, 2009). Nearly all climate researchers have concluded that the problem is urgent and anthropogenic (i.e., essentially \(100 \%\) humancaused). It is thus behavioral, and will be "solved" only by changes in human behavior. The IPCC (Intergovernmental Panel on Climate Change) and Skeptical Science have assembled and disseminated the scientific consensus on GW, but, sadly, the U.S. public is still divided on both GW's existence and its cause (cf. Hoffman, 2011).

A group of climate communication researchers, oddly, suggests that educational ventures would be of little or no
help. Kahan et al. (2012) found (through correlational means) that, for the U.S. (a high per-capita carbon user), direct cognitive approaches (including numeracy and science education) seem to solidify biased viewsreinforcing a kind of cognitive stasis for GW attitudes. This is reminiscent of Lord, Ross, and Lepper (1979), in which people with a strong position tended to polarize further after receiving (not particularly factual) information contrary to their views. Similarly, McCright and Dunlap (2011) highlight data indicating that climate-relevant effects of "education level" are moderated by conservatism or party. (Conservative or "Republican" GW denial was slightly positively related, if at all, with education.) This (also correlational) evidence, they claim, disproves a naïve "knowledge deficit" view-the view that more education can shift the public's beliefs toward the scientific consensus about climate change. However, their own work shows that liberals and conservatives tend to obtain different kinds of information. This split leaves open the possibility that wellconstructed interventions may well induce conservatives to accept the scientific consensus (with little challenge to their core values). Indeed, Lewandowsky, Gignac, and Vaughan (2013) show that offering climate scientists' consensus boosts anthropogenic climate change acceptance.

Our laboratory has provided arguments and many experimental findings that run counter to these "polarization" and cognitive stasis views: For instance, even a small amount of true information can quickly act as a cognitive "lever" to enhance one's understanding and perspective on climate change (Ranney et al., 2012a)—and many other social issues (e.g., abortion and immigration)and even using just a single number/statistic (Garcia de Osuna, Ranney, \& Nelson, 2004; Munnich et al., 2003; Ranney et al., 2008). Below, we offer further experimental results that counter the stasis view for climate education. Notably, we analyze the full spectrum of participants, rather than filtering for those who are already relatively extreme.

Note that new knowledge often facilitates societal shifts and that science "education" has historically driven major social changes-from heliocentrism replacing church doctrine to the acceptance of a tobacco-cancer link in spite of industry obfuscation. (We offer more such germane evidence below.) These data-driven shifts demonstrate how sociologists and social psychologists who hold the stasis view must be incorrect or overly pessimistic. Whether or not they realize it, theorists are haggling over speed, and some
nations learn (e.g., to accept evolution or climate change; Ranney, 2012) faster than others. Of course, learning or acting too slowly can exacerbate existing problems.

We partially agree, though, with those who critique a "knowledge deficit" view of public attitudes (cf. Dickson, 2005). Arbitrary or propaganda-like information need not drive one toward a more empirically supported view. We see the problem, rather, as a wisdom deficit, for which cognitively sophisticated educators can provide the tools that help the public better evaluate the evidence and make choices that match their values. (See Lewandowsky et al., 2012, for a fine discussion of such tools, particularly the correction of misinformation.) We believe that the findings described here will demonstrate that a well considered educational approach is critical for public engagement.

\section*{The GW Mechanism: Extra Greenhouse Effect}

Much of our laboratory's prior research has sought to foster worthwhile, notable conceptual changes with short activities that involve estimations, predictions, or explanations. These activities are followed by small amounts of feedback: Numerically Driven Inferencing studies (NDI; e.g., Ranney et al., 2008) have provided numeric feedback. RTMD (Reinforced Theistic Manifest Destiny) theory, which examines why people in the U.S. are less likely to accept evolution and climate change than are people in peer nations (e.g., Ranney, 2012), has yielded mechanistic interventions.

Ranney et al. (2012a) found that almost no Americans seem to understand the basic mechanism of global warming. Ranney et al. (2012b) includes a 400-word explanation (and experimental stimulus) of the physical-chemical mechanism of the greenhouse effect, here summarized: (1) Earth's surface absorbs (mostly visible) sunlight and then emits infrared light, which greenhouse gases absorb, causing heat energy to leave the atmosphere more slowly than it arrived; (2) as people add more greenhouse gases, the Earth experiences climate change (an added greenhouse effect). In one survey, not a single person out of 270 (mostly public park visitors) could correctly describe (1) and (2). Virtually none of those surveyed could explain a key conceptual piece: the asymmetry of how energy can reach Earth yet then get "trapped" after it arrives (like a "leaky one-wayvalve"), due to the visible-to-infrared energy conversion.

\section*{Mechanism Knowledge is Related to GW Attitudes}

Ranney et al. (2012a) found that the correlation between mechanistic climate change knowledge and attitude toward climate change was robust even when taking into account political party. Mechanistic knowledge correlates with acceptance that global warming is occurring ( \(r=0.22\), \(p=0.0002\) ) and is anthropogenic ( \(r=0.17, p=0.005\) ). Anthropogenic climate change acceptance also predicted financial "willingness to sacrifice" \(\left(\chi^{2}(4)>32, p<0.001\right.\) for each of four items), and one's knowledge score predicted two of these items \(\left(\chi^{2}(1)>3.8, p<0.05\right.\) for both). Further, acceptance of biological evolution (another controversial science topic) was found to predict beliefs and attitudes
toward climate change (as RTMD hypothesizes, and, e.g., Ranney, 2012 found). These findings suggest that the effects of well-chosen aspects of education are both significant and somewhat independent of political affiliation. Indeed, though not reported previously, Ranney et al.'s (2012a) data also showed that evolution acceptance was a significant predictor of climate change acceptance even in a model including the two major political parties \(\left(\chi^{2}(4)=12.3\right.\), \(p<0.02\); N.B., including other parties dramatically reduces quality of fit for any model, likely due to small bin sizes).

\section*{Efficacy in Learning Climate Change's Mechanism}

Ranney, et al. (2012a \& 2012b) also provided two divergent undergraduate samples (from UT-Brownsville, a "HispanicServing Institution," and UC-Berkeley) with the aforementioned 400-word description. Strikingly, this threeminute intervention roughly tripled their mechanistic knowledge on the assessment metric. The intervention also caused both Texas and California undergraduates to increase their climate change acceptance. Contrasting with others' studies noted above, our intervention focused on a fundamental, well-researched knowledge gap, and our assessment focused on acceptance/belief. Such contrasts may explain the difference between observing instructional benefits (as we have) or polarization (as others occasionally have; cf. Lundmark, 2007). We provide further evidence below that such interventions are applicable across broader settings, time-frames, and populations, and that global warming understandings and attitudes are far from static.

\section*{Study 1: A Web-based and Longevity Extension}

Given the replicated demonstrations of significant attitude changes described above, we proceeded to assess whether the mechanism-explanation effects we had obtained were durable or transient. This study extended prior work by delaying the post-test several days. We were also concerned that an "experimental demand" from the classroom setting might have driven our prior results, so we provided the intervention on-line; that is, we assessed whether our materials would elicit significant attitude change even though students participated via their own computers, without experimenter observation. Thus we concurrently explored both the longevity (via delay) and format (on-line) aspects of our phenomenon. We also extended our prompts to incorporate more demographic and introspective queries.

Methods. The instructional materials were those reported in Ranney et al. (2012a \& 2012b; the latter includes the full 400 -word text of our intervention). The empirical differences were that (a) the study was conducted online, via the Qualtrics Inc. (Provo, UT) system, (b) eight items were added to pre- and post-test attitude surveys to add reliability to the related RTMD metrics (specifically, national and religious affinities; these metrics will be reported elsewhere), and (c) five further items were introduced immediately following the instructional material to elicit introspection (about embarrassment, disagreement, etc.).

Undergraduates ( \(\mathrm{N}=80\) ) were recruited via the Research Participation Program (RPP), of the University of California, Berkeley (UCB) psychology department, which allowed us to administer a pre-test to about half of the students (38) between eight and 26 days ( \(\mu=18.5\) days) prior to the study, which may have allayed test-retest effects (although Ranney et al., 2012a, found little evidence for them). Thus, as with Ranney et al. (2012a), some students received the full survey testing "sandwich" while others had no pre-test. A delayed post-test was given to all participants between one and eight days later ( \(\mu=4\) days); this range was used to assess the timecourse of retention in planning subsequent studies. We lack the power to test forgetting over time here (although numerically, we did not find any!).

Results and Discussion. In general and as anticipated, we replicated Ranney et al.'s (2012a) results and extended them by finding that gains were retained over the mean, four-day, delay. (Note: all of this piece's measures use 1-to-9 Likert scales.) Scored knowledge was again linked to self-rated knowledge ( \(r=0.5, p<0.0001\) ) and was similar to that of prior UCB students. Scored knowledge soared from 3.8 (pre-test) to 6.5 (post-test) and 6.3 (delayed post-test); gains were significant ( \(p\) ' \(s<0.0001\), simultaneous comparisons). Stated GW beliefs followed a similar pattern. Mean ratings rose from 6.20 (pre-test) to 6.54 (post-test) and were mostly retained at 6.44 (delayed post-test)—notable gains (again) for a 400 -word text \((t(79)=2.5, p=0.006\) for immediate, and \(t(79)=1.7, p=0.05\) for delayed). The largest post-test gains were found in agreeing with "Human activities are largely responsible for the climate change..." (a 0.25 gain) and certainty that global warming is occurring (a 0.19 gain). The self-rated knowledge mean similarly increased markedly from pre- to post-test ( 4.5 to 5.6 ). This gain's retention, gratifyingly, was also noted on the delayed post-test (5.2; simultaneous comparisons for both the gain and retained gain were again significant; \(p\) 's \(<0.0001\) ). The immediate increase in self-rated knowledge replicates results from Ranney et al. (2012a; the results were not reported then).

In sum, Study 1 extends the finding that well-considered information, even received online, increases anthropogenic GW acceptance and behaviorally relevant attitudes. Further, the conceptual changes that result from reading even 400 words have notable longevity. These effects have been replicated with the general public as well (unpublished data). Computer-based interventions often scale well, enhance reliability, and prove cost-effective; so, given our results, we recommend the online distribution of mechanistic explanations, especially about climate change.

\section*{Altering Beliefs with Factual Numbers}

The aforementioned NDI paradigm has yielded marked attitudinal and conceptual shifts with quite minimalist interventions. NDI and one of its procedures, EPIC (both introduced by Ranney and students), represent a particularly compact, well-specified intervention. EPIC participants (1)
provide an Estimate for each policy-relevant item's quantity, (2) state a preferred target (or monetary allocation) Policy (or Preference) for each quantity, (3) receive true feedback quantities to Incorporate (as new "Information"), and (4) indicate whether their policies have Changed due to the feedback. With just a single wellselected quantity, the EPIC procedure's feedback often shifts one's attitudes. EPIC-spawned conceptual changes are often remarkably durable for such a small intervention (e.g., Ranney et al., 2008), as evidenced by increased estimation accuracy 12 weeks after the procedure (Munnich, Ranney, \& Bachman, 2005). Therefore, we sought to employ NDI interventions in addition to the mechanism intervention from Study 1 and prior studies. Specifically, we presented different participant groups with numerical information that is relevant to global climate change acceptance. We used numbers that were likely to boost acceptance (Studies 3 \& 4), as well as numbers that we thought might erode individuals' acceptance of climate change (Study 2).

\section*{Study 2: Eroding Beliefs with "Evil" True Numbers}

Some organizations publicize out-of-context facts to try to undercut the reality or gravity of human-caused GW. These are usually blatantly cherry picked, such as that Earth slightly cooled by \(0.2^{\circ} \mathrm{F}(.04 \%\) re: absolute-zero terms) from 1940 to 1975 (Jastrow, Nierenberg, \& Seitz, 1991). While surprising, this fact hardly contradicts the ever more obvious warming trend over the last \(125+\) years: one can easily pick endpoints that are oddly high or low in a noisy time series. (The slight decrease is also explained by a planetary motion trend.) Given this rather clear intent to mislead (Oreskes \& Conway, 2010), we (partly tongue-in-cheek) label these numbers "evil." Thus, Study 2's hypotheses are that a few misleading facts can reduce one's (1) climate change acceptance, (2) ratings of knowledge of the issue, and (3) climate-change funding preferences. Of course, lest we erode participants' acceptance of anthropogenic climate change more than fleetingly, we debriefed them right afterward with more complete information-including the mechanism and a large dose of (non-evil) relevant facts.

Methods. The survey and instructional materials were analogous to those used in Ranney et al.'s (2012a) paper-and-pencil second study. The main difference was that the mechanism was replaced with one of two interventions. For one version, part ( \(\mathrm{n}=59\) ) of a UCB college class ( \(\mathrm{N}=104\) ) estimated each of eight items before receiving the feedback values, with an emphasis on maximizing the quantity of feedback numbers given to the student. To this end, this eight-item survey included only a post-test (i.e., no pre-test), and lacked a policy component (thus, it was an EI intervention, lacking " P " or " C "). A more comprehensive engagement containing only two items was administered to the rest of the class ( \(\mathrm{n}=45\) ), and this version included a pretest and extra questions about each item; we asked these students about their surprise level after each feedback value and requested both their climate-change funding Policies
and post-feedback policy Changes regarding/versus various UN (e.g., UNDP millennium) goals.

Results and Discussion. As predicted, preferences for funding GW-related UN goals dropped ( \(\chi^{2}(1)=22, p<0.01\) ) versus all eight non-climate UNDP funding alternatives. (Unfortunately for GW as a social priority, the highest mean pre-test preference for funding climate change initiatives reached only a \(50-50\) split of available funds.) Also as predicted, climate change acceptance significantly dropped, from a 6.5 pre-test mean to a 6.2 post-test mean for the twoitem group \((5.5 \%\) of the available room to drop on the \(9-\) point scale, \(t(42)=-4.3, p<0.001\) ), and significantly to 5.9 for the eight-item group \((11 \%\) of the available room, \(t(88.6)=-2.61, p<0.005)\); note that these shifts were also in the direction of ambivalence (a " 5 " rating), and may reflect confusion rather than disagreement. Our third hypothesis was also supported: self-rated knowledge fell from a mean of 5.0 on the pre-test to 4.5 for the two-item group ( \(12 \%\) of the available room, \(t(44)=-2.5, p<0.01\) ), and plummeted to 2.9 on eight-item survey \((t(87.2)=-5.3, p<0.001)\). This latter drop of 2.1 is \(53 \%\) of the available room to drop on a \(9-\) point scale, which is exceptionally large.

It is clear that even relatively educated members of the public (e.g., undergraduates at a top-tier university) are highly susceptible to misleading, cherry picked facts. Such facts are clearly known to organizations attempting to undermine the overwhelming scientific consensus about climate change. Thus, it seems incumbent upon cognitive science to counter the increasing sophistication with which such organizations distribute misleading information.

\section*{Study 3: "Saintly" Numbers Supporting GW}

Given Study 2's observed efficacy for "evil" numbers and the NDI paradigm's prior successes, this study assessed the utility of numbers that support claims of global warming. Again partly tongue-in-cheek, we call these "saintly" numbers. Given prior NDI studies of similarly "shocking" magnitudes (e.g., Garcia de Osuna, et al., 2004), our hypothesis was that the accurate feedback would increase participants' climate change acceptance, but diminish selfconfidence in their knowledge of the issue.

Methods. Like Study 1, Study 3 was both online and used a UCB-RPP pre-test survey (for a subset of 30 students); however, we increased the delay between pre-test and intervention to a mean of 18 days. We queried the individuals \((\mathrm{N}=60)\) about eight quantities. The eight items also included questions directed at participants' surprise and their reactions to each number. (Monetary preferences were left out of this version because we already observed attitude shifts in the simplified eight-item "evil" intervention.) An added feature of the on-screen intervention is that we could more saliently remind individuals of each of their estimates on the same page on which they incorporated numerical feedback, better ensuring that they contrasted the two. As with Study 1's online survey, a post-test about attitudes and
beliefs was administered both immediately after our intervention and after the 18-day retention interval.

Results and Discussion. Attitudes, acceptance, and beliefs about climate change were stable after this intervention with "saintly" numbers (pre-test: 6.71; post-test: 6.67). This stability was unexpected (but see below for explanations), especially because these items (as with the "evil items) were, as anticipated, able to significantly erode self-rated knowledge ( 5.3 to \(4.0, t(29)=-3.6, p<0.01\) ). This erosion was comparable to that found with the "evil" numbers. These items were also relatively high regarding participant surprise, compared to Study 1's 400-word intervention. The mean surprise rating across Study 3's items was 4.8 , while the mean surprise rating for the 400 words was 2.9. (All ratings above " 1 " indicate some level of surprise.)

One of the most surprising numbers (a 5.2 mean) was the near- \(100 \%\) of active researchers who support anthropogenic climate change's tenets, reflective of the strong relationship between the perceived scientific consensus and climate change acceptance, as Lewandowsky, Gignac, and Vaughan (2013) report. The two numbers most similar to the statistics in the 400 words were similarly surprising, with the rises in atmospheric methane \((+151 \%)\) and \(\mathrm{CO}_{2}(+40 \%)\) yielding respective surprisingness means of 5.9 and 5.1.

Despite these powerful impacts, Study 3 yielded no effect on beliefs or attitudes. This lack of effect is counter to prior NDI studies, in which individuals' preferences and beliefs were often markedly shifted by even a single number. (An experimental silver lining here is the finding that participants will not report greater climate change acceptance by mere dint of experimenter demand!) One possible explanation regards an unintended method change: Participants in prior NDI and RTMD studies were usually told the particular scientific/literature source-both for each statistic that was sought and each true value provided as feedback. Study 3 omitted this, so is possible that participants were less compelled by the authority of this study's statistics, compared to those in Study 2. Another possibility is that these UCB students were near ceiling for acceptance, with a reluctance to admit the disturbing effects of GW. Further, it may be that, as in Study 2, participants were left feeling less knowledgeable-weakening any boost these surprising numbers could have had on climate change acceptance. Finally, perhaps students lacked an appropriate context for integrating this information. The next study illustrates one way to contextualize such feedback statistics.

\section*{Study 4: Consolidating Knowledge-Gain Effects}

Study 4 explores combining (a) the replicated effect of explaining global warming's mechanism and (b) the promising effect of offering representative statistics that support understanding of GW's effects and dangers. Participants were 63 urban San Francisco Bay Area highschool students, who likely better represent the general public than do the prior studies' university undergraduates.

Methods. The students were in three junior-level chemistry classes. Much of the school ( \(40 \%\) ) is on free/reduced lunch (a low-income marker), \(95.1 \%\) is "non-white," and just \(35 \%\) lists English as their primary language. We presented them with (1) a more elaborated mechanistic explanation/minicurriculum, and (2) six key GW statistics. A control group received (1) from above, with (3) six unrelated statistics. We predicted that (i) the mechanistic explanation would yet again yield gains in climate-change understanding and more pro-environmental attitudes, (ii) the key statistics would enhance such effects, and (iii) the effects would remain a month later. Felipe (2012) describes the intervention and results more completely. Everyone received 15 minutes' instruction on climate change's mechanism for each of three days. After estimating each of the six critical climate change quantities, the experimental group ( \(n=33\) ) received the true values as feedback. The control group ( \(\mathrm{n}=30\) ) received six estimation-feedback values that proved equally surprising but which were unrelated to climate (drawn from Ranney et al., 2008). Each student filled out a pre-test, a post-test (three days later; \(\mathrm{N}=63\) ), and a delayed post-test ( 34 days later; \(\mathrm{N}=59\) ). Of each test's many measures, we focus here on scientific mechanistic knowledge, attitudes toward global warming, and Environmental Behavioral Intentions (EBI). We report below only the gist of some of Felipe's (2012) notable findings, yet provide some newer findings about the effects of relevant, surprising numbers on the retention of gains from the mechanistic knowledge curriculum.

Results and Discussion. Pre-test mechanism knowledge was so close to zero that the curriculum hugely increased both groups' GW-mechanism understandings-by \(1,619 \%\), on average (combined \(t(62)=9.31, p<0.0001\) ). Gains in both groups' mean EBI scores were also quite notable \((t(62)=5.91, p<0.0001)\); the effects emphatically replicate our prior three studies' findings that show the importance of mechanistic information in enhancing a person's GW understanding and "pro-environment" attitudes. Even more importantly, the gains were significant 34 days later for both groups (control \(t(27)=3.01\), experimental \(t(28)=5.2\), both \(p\) ' \(<0.002\) ), which seems an especially impressive effect for less than one day's class out of about 170 instructional days. (One might imagine the curriculum's effect if it were extended or reinforced multiple times-or given to the general public.) Interestingly, although the control group greatly gained by learning the mechanism, the experimental group's retention of their mechanistic knowledge was significantly greater than-about twice-that of the control group ( \(t(48.7)=-2.61, p=0.01\); planned comparison after a significant ANOVA interaction term), suggesting that the experimental group's critical statistical information helped reinforce and secure the mechanistic information. Thus, the numbers may have "primed the pump" for learners to more durably encode their new GW mechanism knowledge. The differences show separate utilities for both the mechanistic information and the statistically pertinent information-and the suitability of our brief intervention for high school students, and, likely, the wider public.

Students' acceptance of climate change and concern about its effects were both near ceiling on the pre-test ( 8.3 and 8.1, respectively, thus range-restricted on the 1-9 scales); even so, the experimental group exhibited a significant gain \((t(32)=1.76, \quad \mathrm{p}<0.05)\). Interestingly, the curriculum inadvertently slightly increased students' acceptance that climate change is "just part of a natural cycle" (but to only a modest 3.7 on the 9 -point scale) rather than anthropogenic (which slightly dropped, yet remained at about 7 on the 9 point scale). Upon analysis, this modest, counterproductive result was due to the curriculum focusing on how, prior to humans' influences, Earth's evolution already yielded greenhouse gases and a greenhouse effect. The curriculum failed to communicate carefully that humans have caused an extra greenhouse effect, which represents anthropogenic climate change. This finding highlights the importance of understanding that, while the greenhouse effect in which our species evolved had long kept Earth from being a virtual snowball, the extra greenhouse gases that humans have pumped into the atmosphere during the past \(250+\) years are disturbing a narrow thermal balance. Educators should emphasize that, as with drinking ten gallons of water, a beneficial substance can turn deadly in excess.

\section*{General Discussion}

Our studies have provided an evidential medley that effectively disconfirms the idea that GW-relevant knowledge and attitudes are locked in cognitive stasis. Contrary to those who over-problematize a "knowledge deficit" (or "information deficit") view of global warming communication, we see a "wisdom deficit." Here (and in Ranney et al., 2012a) we have markedly un-problematized any deficit with our interventions' "cognitive levers." In contrast, it is unlikely that offering an ill-structured list of uncompelling facts to an unprepared mind (or thinly veiled rhetoric; cf. Lord, Ross, \& Lepper, 1979) will notably alter beliefs or behaviors-especially for the difficult topic of climate change. Rather, one must be sensitive to specific (mis)understandings that may be relevant to a learner grappling with a domain. Ultimately, we will likely need to engage virtually all people, aiding them in connecting their long-term values to the long-term effects of their behaviors.

Disturbingly, Study 2 showed that climate change acceptance can be readily eroded by misleading, cherrypicked data. To guard against such "evil" misinformation, people need the context to recognize them as the clever propaganda that they are. Such prophylactic interventions may represent promising targets for further educational research initiatives (cf. Lewandowsky et al., 2012).

We are currently studying ways to disseminate the information that we have found to elicit worthwhile cognitive and belief changes. For instance, we are developing on-line instructional materials (e.g., videos) that can widely convey both global warming's mechanism and the statistics that reflect the scientific consensus of climate change-so the public can more fully join that consensus. Even if people forget an offered statistic or a mechanistic
aspect, we hope they will recall being rationally convinced of climate change's reality, danger, and need for action.

We have shown above that on-line survey interventions, brief curricula, and classroom lessons can have a marked and persistent effect on one's knowledge, understanding, beliefs, and attitudes about global warming. Despite arguments to the contrary, some simple cognitivelyinformed interventions may well be fundamental in building humanity's resolve to tackle global climate change.

\section*{Acknowledgments}

We thank Daniel Reinholz, Lloyd Goldwasser, Baljeet Rai, Myles Crain, Sarah Cohen, R. Farjadi, R. McGuane, S. Hu, D. Kazmi, L. Nevo, K. Le, T. Charoenying, A. Galpern, B. Balukjian, the rest of UCB's Reasoning Group, and Jeff Wilson. This research was supported by UCB's Committee on Research and by an Institute of Education Sciences (IES) pre-doctoral training grant (\# R305B090026) to UCB.

\section*{References}

Barnosky, A. D. (2009). Heatstroke: Nature in an age of global warming ( 2 nd ed.). Washington, DC: Island Press.
Dickson, D. (2005). The case for a "deficit model" of science communication. SciDev.net. Retrieved January 19, 2013, from http://www.scidev.net/en/editorials/the-case-for-a-deficit-model-of-science-communic.html
Felipe, J. (2012). Numerical reasoning, knowledge, and environmental behavior regarding climate change. Masters project, Graduate School of Education, University of California, Berkeley.
Garcia de Osuna, J., Ranney, M., \& Nelson, J. (2004). Qualitative and quantitative effects of surprise: (Mis)estimates, rationales, and feedback-induced preference changes while considering abortion. In K. Forbus, D. Gentner, \& T. Regier (Eds.), Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Society (pp. 422-427). Mahwah, NJ: Erlbaum.
Hoffman, A. (2011). The growing climate divide. Nature Climate Change, 1, 195-196.
Jastrow, R., Nierenberg, W., \& Seitz, F. (1991). Global warming: What does the science tell us? Energy, 16(1112), 1331-1345. doi:10.1016/0360-5442(91)90006-8

Kahan, D. M., Peters, E., Wittlin, M., Slovic, P., Ouellette, L. L., Braman, D., \& Mandel, G. (2012). The polarizing impact of science literacy and numeracy on perceived climate change risks. Nature Climate Change, 2, 732-735.
Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., \& Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. Psychological Science in the Public Interest, 13, 106-131.
Lewandowsky, S., Gignac, G., \& Vaughan, S. (2013). The pivotal role of perceived scientific consensus in acceptance of science. Nature Climate Change, 3, 399-404.
Lord C. G., Ross L., Lepper M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. Journal of Personality and Social Psychology, 37, 2098-2109.

Lundmark, C. (2007). The new ecological paradigm revisited: anchoring the NEP scale in environmental ethics. Environmental Education Research, 13, 329-347.
McCright, A. M., \& Dunlap, R. E. (2011). The politicization of climate change and polarization in the American public's views of global warming, 2001-2010. The Sociological Quarterly, 52(2), 155-194.
Munnich, E., Ranney, M., \& Bachman, M. L. N. (2005). The longevities of policy-shifts and memories due to single feedback numbers. In B.G. Bara, L. Barsalou, \& M. Bucciarelli (Eds.) Proceedings of the Twenty-seventh Annual Conference of the Cognitive Science Society (pp. 1553-1558). Mahwah, NJ: Erlbaum.
Munnich, E., Ranney, M., Nelson, J., Garcia de Osuna, J., \& Brazil, N. (2003). Policy shift through NumericallyDriven Inferencing: An EPIC experiment about when base rates matter. In R. Alterman \& D. Kirsh (Eds.), Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society (pp. 834-839). Mahwah, NJ: Erlbaum.
Oreskes, N., \& Conway, E. M. (2010). Merchants of doubt: How a handful of scientists obscured the truth on issues from tobacco smoke to global warming. New York: Bloomsbury Publishing.
Ranney, M.A. (2012). Why don't Americans accept evolution as much as people in peer nations do? A theory (Reinforced Theistic Manifest Destiny) and some pertinent evidence. In K. S. Rosengren, S. Brem, E. Evans, \& G. M. Sinatra (Eds.) Evolution challenges: Integrating research and practice in teaching and learning about evolution (pp. 233-269). Oxford: Oxford University Press.
Ranney, M.A., Clark, D., Reinholz, D., \& Cohen, S. (2012a). Changing global warming beliefs with scientific information: Knowledge, attitudes, and RTMD (Reinforced Theistic Manifest Destiny theory). In N. Miyake, D. Peebles, \& R. P. Cooper (Eds.), Proceedings of the 34th Annual Meeting of the Cognitive Science Society (pp. 2228-2233). Austin, TX: Cognitive Science Society.
Ranney, M.A., Clark, D., Reinholz, D., \& Cohen, S. (2012b). Improving Americans' modest global warming knowledge in the light of RTMD (Reinforced Theistic Manifest Destiny) theory. In J. van Aalst, K. Thompson, M. M. Jacobson, \& P. Reimann (Eds.), The Future of Learning: Proceedings of the Tenth International Conference of the Learning Sciences (pp. 2-481 to 2-482). International Society of the Learning Sciences, Inc.
Ranney, M. A., Rinne, L. F., Yarnall, L., Munnich, E., Miratirx, L., \& Schank, P. (2008). Designing and assessing numeracy training for journalists: Toward improving quantitative reasoning among media consumers. In P A. Kirschner, F. Prins, V. Jonker, \& G. Kanselaar (Eds.), International Perspectives in the Learning Sciences: Proceedings of the 8th International Conference for the Learning Sciences (pp. 2-246 to 2253). International Society of the Learning Sciences, Inc.

World Bank (2012). Turn down the heat: Why a \(4^{\circ} \mathrm{C}\) warmer world must be avoided. Washington, DC: The World Bank.

\title{
Fractal Structure of the Nested Actions in Keeping the Beat
}

\author{
Charles A. Coey (coeyca@mail.uc.edu) \\ Center for Cognition, Action and Perception, Department of Psychology, University of Cincinnati, ML 0376, 4150 Edwards Cl., University of Cincinnati, Cincinnati, OH 45221-0376 USA \\ Justin Hassebrock (hassebja@mail.uc.edu) \\ Center for Cognition, Action and Perception, Department of Psychology, University of Cincinnati, ML 0376, 4150 Edwards Cl., University of Cincinnati, Cincinnati, OH 45221-0376 USA \\ Heidi Kloos (heidi.kloos@uc.edu) \\ Center for Cognition, Action and Perception, Department of Psychology, University of Cincinnati, ML 0376, 4150 Edwards Cl., University of Cincinnati, Cincinnati, OH 45221-0376 USA \\ Michael J. Richardson (richamo@ucmail.uc.edu) \\ Center for Cognition, Action and Perception, Department of Psychology, University of Cincinnati, ML 0376, 4150 Edwards Cl., University of Cincinnati, Cincinnati, OH 45221-0376 USA
}

\begin{abstract}
The current experiments investigated the fractal structure in the nested actions of tapping behavior. The results revealed that task constraints (e.g., tapping to a metronome) alter the fractal structure of a given aspect of the behavior (e.g., intertap interval) and decouple its long-term interactions with other aspects of the behavior (e.g., key-press duration). These results support the idea that fractal structure reflects the dynamical organization of complex systems.
\end{abstract}

Keywords: complex systems, fractal scaling, finger tapping

\section*{Introduction}

There is certainly no shortage of complexity for the student of human mind and behavior. The human system is at once physical, chemical, biological, psychological, and social. We cognitive scientists carefully design experimental tasks and manipulations to gain an empirical purchase on the many forces that shape human behavior. Traditionally, the field has relied on classical techniques of linear statistics. We take different averages or degrees of variability in reaction time to be indicative of the cognitive processes underlying performance in our experimental tasks. Recently, however, researchers have turned to examining more subtle and complex statistical facets of data to understand the processes involved in the organization of human behavior; namely, fractal structure.

Comprehensive review of this statistical property, the available mathematical techniques for its assessment, and the potential implications for theories of cognitive science is not possible in the limited space provided here (see Brown \& Liebovitch, 2010; Delignieres \& Marmelat, 2013; Holden, 2005; Van Orden Kloos, \& Wallot, 2010). Nonetheless, a brief introduction to the topic is warranted. The term "fractal structure" is here being used loosely to refer to patterns of variability in repeated measurements of human behavior. Most traditional, linear statistical
techniques operate on the assumptions that deviations from mean performance will obey a Gaussian distribution and that these "errors" will be uncorrelated with one another. Data displaying fractal structure violates these assumptions. That is, fluctuations in repeated performances exhibit "longterm dependencies" such that errors in early observations are correlated with errors in much later observations. Fractal data obey power-law scaling such that size of a given error is inversely proportional to how often errors of that size occur. Thus, like geometric fractals, these data are said to be "self-similar" and "scale-invariant". Fractal data entail nested patterns of variability wherein small variations in measurement have the same structure as large variations. Such structure in repeated measurements is often referred to as "pink noise", as contrasted against the random variation entailed in "white noise" (Holden, 2005).

In part, these patterns are important to researchers in cognitive science as they have been discovered in a plethora of human data from the simplest of reaction time tasks taking place over the course of minutes (Van Orden, Holden, \& Turvey, 2003) to measurements of self-esteem over a many months (Delignieres, Fortes, Ninot, 2004). More importantly, experimental manipulations of the type typically employed by cognitive scientists have been shown to affect fractal structure. For instance, Kello et al., (2007) demonstrated that reaction times to unpredictable cues were not only slower, but also closer to white noise (i.e., random) variation, than reaction times to predictable cues.

Despite their widespread occurrence, there is not yet a unified account of how these fractal patterns get into the data or what they imply for theories of cognitive science (e.g., Van Orden, Holden, \& Turvey, 2005; Wagenmakers, Farrell, \& Ratcliff, 2005). The current experiments are intended to contribute to the on-going discussion by examining the fractal structure in the nested actions in a tapping task, their dynamical interaction with one another, and the impact of employing different task constraints.

\section*{Experiment 1}

Most statistical techniques used to assess fractal structure require very many observations made under relatively constant task conditions. As such, research revealing these structures in human behavior has typically preferred very simple tasks. One frequently studied task is finger tapping (e.g., Chen, Ding, \& Kelso, 2001; Chen, Repp, \& Patel, 2002; Gilden, Thorton, \& Mallon, 1995; Lemoine, Torre, \& Delignieres, 2006; Madison, 2001; Musha, Katsurai, \& Teramachi, 1985; Ogden \& Collier, 1999; Yamada, 1995).

Generally, studies have found evidence of fractal structure in continuation tapping, wherein participants attempt to keep a steady beat briefly demonstrated to them by a metronome stimulus at the beginning of a trial. In this case, the intervals between taps take on a "persistent" structure (i.e., longer taps tend to be followed longer taps). Interestingly, the fractal structure is different during synchronization tapping, wherein participants synchronize their taps to a constant metronome stimulus. In this case, the intervals between taps take on an "anti-persistent" structure, (i.e., longer taps tend to be followed by shorter taps) whereas the asynchronies between the participant's taps and the metronome show a persistent fractal structure. These findings have been interpreted and modeled as the result of the metronome serving as a corrective feedback mechanism for the maintenance of a given tapping interval (Torre \& Delignieres, 2008).

While these results are reasonably well-understood, to date there have been no investigations of the nested actions comprising finger tapping. That is, most tasks require a behavior that consists of many "sub-actions", all of which may not be measured or examined. In tapping, the task requires striking the key, holding it down for some period of time, releasing the key, and waiting some period of time before striking the key once more. Our first experiment was designed to investigate the fractal structure in these nested actions during continuation tapping, how these nested behaviors might interact with one another across the measured span of behavior, and what differences might be evident during synchronization tapping.

\section*{Method}

\section*{Participants}

Sixteen undergraduate students from the University of Cincinnati participated in the study for partial course credit. All participants were over 18 years of age and right-handed.

\section*{Apparatus}

The participants' tapping behavior was recorded using a USB midi keyboard. The keyboard was connected to a PC computer running Ableton Live (Ableton, Berlin Germany). This software was used to simultaneously record the timeseries of the participants' taps (with a \(\pm 5 \mathrm{~ms}\) error) and present the auditory metronome stimulus to the participant through a pair of headphones.

\section*{Procedure and Design}

After informed consent, participants were instructed that they would complete two trials of tapping behavior while being presented different auditory stimuli. They were then shown how to produce the desired tapping behavior; namely, by resting their right hand on the table and producing taps with their index finger on a key marked with a small piece of tape, being sure to depress and release the key entirely on each tap. Each participant first completed the continuation tapping condition. The stimulus consisted of 10 seconds of a 2 Hz metronome ( 500 ms between beats) followed by 10 minutes of silence. Participants were instructed to synchronize their taps to the metronome for the first 10 seconds, and then to maintain that same beat without the metronome for the remainder of the trial. Each participant then completed the synchronization condition. In this trial, the stimulus simply consisted of 10 minutes and 10 seconds of a 2 Hz metronome. Participants were instructed to synchronize their taps to the metronome for the duration of the trial. At the conclusion of the experiment participants were thanked and debriefed.

\section*{Data Analysis}

The data output by the recording software were collated to yield three different time-series for each trial. The first series contained inter-tap intervals (ITI) where data signified the time elapsed between each tap and the following tap. The second series consisted of key-press durations (KPD) where data signified the time the key was depressed on each tap. The third series consisted of key-release intervals (KRI) where data signified the time between the release of the key of each tap and the following tap. The relationship between these three measures of tapping behavior is depicted in Figure 1. Note that these variables are not independent. For any given tap, determining any two of the variables completely determines the third as well. Thus, we consider this data set to properly consist of only two pieces of information. Nonetheless, we will use all three variables for reasons that will become apparent.


Time

Figure 1: The figure portrays a sequence of three taps and the three measurements collected for each tap.

Prior to fractal analysis, each time-series was subjected to several pre-processing steps to eliminate outliers and linear trends that might otherwise affect the outcome of the test (see Eke et al., 2000; Delignieres et al., 2006). Specifically, individual taps were removed from the data set when either the corresponding ITI was outside the range of \(300-700 \mathrm{~ms}\), or the corresponding KPD was greater than 500 ms . These values were chosen to reflect instances in which the participant failed to either depress or release the key entirely or failed to keep their taps close to the prescribed tempo. When a tap met either of these exclusion criteria, it was removed from each of the three measurement series. Following outlier removal, each time-series was trimmed to 1024 taps as the fractal analysis employed requires series of a length equal to a power of two. Finally, a linear bridge detrending was applied to each series.

The pre-processed series were submitted to a power spectral density (PSD) analysis to assess fractal structure. First, each series is standardized by Z-scoring each value. Then each series is approximated by a set of sinusoids with variable power and frequency by a Fourier transformation. As described above, fractal data obey power-law scaling wherein the size of each deviation is inversely proportional to how often deviations of that size occur. This relationship can be expressed mathematically between the power ( P ) and the frequency (f) of the sinusoids generated by Fourier transformation, where \(\mathrm{P}=1 / \mathrm{f}^{\alpha}\). The "scaling exponent" \((\alpha)\) summarizes the nature of the fractal structure evident in the series with persistent fractal structure indicated by \(\alpha \approx 1\), with random, white noise structure indicated by \(\alpha \approx 0\), and anti-persistent structure indicated by \(\alpha \approx-1\). An estimation of \(\alpha\) can be obtained by plotting power against frequency on double-logarithmic axes, and finding the slope ( S ) of the regression line that best fits this "spectral plot", with \(\alpha=-\mathrm{S}\). In accordance with past research, we estimated \(\alpha\) from only the lowest portion ( \(25 \%\) ) of the power spectrum (Eke et al., 2000; Delignieres et al., 2006).

We also sought to investigate the dynamical interaction of the three measures (ITI, KPD, KRI). To this end, we used cross-correlation analyses. Similar to auto-correlation, cross-correlation computes the correlation between two series across a range of time-lags. The cross-correlation function therefore can capture dependencies between the different tapping variables that exist across several taps.

\section*{Results and Discussion}

Participants generally had no trouble completing the task and there were on average only 5.5 outlier taps per trial. Generally, there were no significant differences in either the means or standard deviations for any of the three variables as a function of experimental condition (all \(p\) 's \(>05\) ). The sole exception was that the standard deviation for ITI was smaller during synchronization ( \(M=26.5 \mathrm{~ms}, S D=4.58\) ) than during continuation tapping ( \(M=31.4 \mathrm{~ms}, S D=8.12\) ), \(t(15)=2.57, p=.021\).

\section*{PSD Analysis}

The change in the fractal structure in ITI across experimental conditions was consistent with the findings of past research (e.g., Chen et al., 2001; Gilden et al., 1995). Specifically, there was a significant decrease in \(\alpha\) from persistent structure during continuation tapping ( \(M=.60\), \(S D=.20\) ) to anti-persistent structure during synchronization tapping ( \(M=-.48, S D=.58\) ), \(t(15)=7.79, p<.001\). Both KPD and KRI showed different patterns of results. There was a small but significant increase in \(\alpha\) for KPD from continuation ( \(M=.71, S D=.23\) ) to synchronization tapping ( \(M=.88, S D=.22\) ), \(t(15)=-2.41, p=.03\). Conversely, there was no difference in \(\alpha\) for KRI between continuation ( \(M=.66, S D=.18\) ) and synchronization tapping ( \(M=.60\), \(S D=.30)\). This pattern of effects is depicted in Figure 2.


Figure 2: Change in \(\alpha\) for ITI, KPD, and KRI from the continuation to synchronization tapping conditions.

Although the observed difference in the fractal structure in ITI is in line with the results of the past tapping research, the effects for KPD and KRI are new findings without established theoretical interpretations. One proposal endorsed by several researchers is that the fractal structure evident in ITI during continuation tapping, and in the asynchronies to the metronome during synchronization tapping, is the empirical signature of the emergent behavior of complex systems (e.g., Chen et al., 2001; Gilden, 2001; Lemoine et al., 2006; Yamada, 1995). Briefly, this account asserts that the structures present in the data are reflective of the dynamical organization of the behavioral system that produced them. The implication is thus that the observed fractal structure does not issue from one particular cognitive or physiological component. Rather, the variation in behavior is the collective result of the interaction of many interdependent processes (Holden, Van Orden, Turvey, 2009). To attempt to extend this account to the results of KPD and KRI, we examined the cross-correlations between the three measures of tapping behavior in hopes of revealing the nature of their dynamical interaction.

\section*{Cross-Correlation Analysis}

The cross-correlation functions for ITI-KRI and for KPDKRI are depicted in Figure 3. As these functions were found to be roughly symmetrical across negative and positive lags, only the positive half of the function is shown here.


Figure 3: Cross-correlations for ITI-KRI and KPD-KRI.
As discussed above, "long-term dependencies" are entailed in fractal variation within a single behavioral measure. The upper panel of Figure 3 suggests that similar long-term dependencies exist between the nested actions involved in continuation tapping behavior. Specifically, the full interval between taps (i.e., ITI) is moderately correlated with the sub-interval (i.e., KRI) out to 15 taps and later. Interestingly, all of this long-term structure is absent during synchronization tapping. This suggests that the constraint of the metronome effectively "decouples" these two dynamics of the tapping behavior. The same basic pattern was evident in the cross-correlation function for ITI and KPD, although it was less pronounced.

In contrast, the cross-correlation function for the two subintervals (KPD and KRI) reveals a fundamentally different pattern across task conditions (lower panel Figure 3). During continuation tapping these variables reveal a moderate negative long-term correlation with one another. Most interestingly, this long-term structure is not damped out by the advent of the metronome in synchronization tapping, but rather grows stronger (i.e., more negative).

It is important to note that the measurement variables analyzed in this experiment are just one window into the processes underlying the tapping behavior. Recall, these
variables are not strictly independent. As such, one might contest that the cross-correlation between KPD and KRI does not reflect the relationship of two separate variables, but simply variation in the times when the key was released. This is essentially correct. As revealed by PSD, and explicated by cross-correlation, the persistent structure in these sub-intervals is unaffected, or is actually stronger, when the metronome constrains the interval between taps (i.e., ITI). Interestingly, this structure in key release times cannot simply be accessed by taking the difference of the key release times (IRI). Submitting IRI to PSD reveals the exact same pattern of effects found for the ITI variable; persistent structure during continuation tapping ( \(\alpha=.55\) ), and slightly anti-persistent structure during synchronization tapping ( \(\alpha=-.33\) ). As such, this variable accesses the same structure in the time between taps as does ITI. Thus, the two independent (sub)behaviors entailed in this task might be best construed as the "tap-to-tap" behavior and the "between-taps" behavior, with our measurement variables being only convenient windows into these dynamics.

\section*{Experiment 2}

Experiment 2 was designed to further investigate the interplay of these nested actions and how task constraints affected their fractal structure. To our knowledge, only one other study has investigated the fractal structure in multiple, nested actions. Kello et al., (2007) conducted a series of reaction time experiments in which they recorded not only the time taken to respond to a stimulus, but also the length of time the participants depressed the key on each response. Taken together, these experiments suggested that reaction times and key contact times were not correlated with one another, and that the fractal structure in reaction times could be affected independently of the structure in key contact times. They did not, however, actually attempt to alter the fractal structure of the key contact times directly. The purpose of Experiment 2 was thus to attempt a manipulation that might constrain the between-taps behavior (i.e., KPD) in our tapping task and thereby investigate the relationship between task constraints and fractal structure generally.

\section*{Method}

\section*{Participants}

Twenty-two undergraduate students from the University of Cincinnati participated in the study for partial course credit. All participants were over 18 years of age and right-handed.

\section*{Procedure and Design}

The design was nearly identical to that of Experiment 1. The primary difference was that half of the participants were instructed not only to synchronize their taps to the metronome during the synchronization condition, but also to attempt to keep the key depressed for the length of the metronome tone. So that the length of the tone would be salient to the participants, the metronome stimulus consisted of alternating 400 ms tones and 400 ms periods of silence.

Prior to this additional manipulation, each participant first completed the continuation tapping condition. In this trial, participants were played the metronome for 10 seconds, and then attempted to maintain the same beat for 8 minutes. Each participant was then given task instructions according to their experimental group and completed the synchronization condition. Participants in the "hold" group both synchronized their taps with the metronome and held the key down for the length of the tone, while participants in the "tap" group simply synchronized with the metronome.

Due to the change in the prescribed tempo of the tapping behavior, the criteria for outlier taps changed. Here, taps were discarded from the data set when either the corresponding ITI was outside the range of \(600-1000 \mathrm{~ms}\), or the corresponding KPD was greater than 800 ms . Also, as this frequency of tapping yielded approximately 600 taps within each trial, the time-series were trimmed to 512 points rather than 1024. The final, pre-processed time-series were submitted to PSD and cross-correlation analyses as before.

\section*{Results and Discussion}

As in Experiment 1, participants had little difficulty with the task and there were on average only 9.8 outlier taps per trial. There were, however, several effects in the linear statistics of the tapping variables. Most importantly, there was a significant interaction effect for mean \(\operatorname{KPD}, F(1,20)=8.84\), \(p=.008\). Mean KPD for the hold group increased strongly from continuation tapping ( \(M \approx 270 \mathrm{~ms}\) ) to synchronization tapping ( \(M \approx 440\) ). In contrast, the tap group KPD only slightly increased from continuation \((M \approx 250)\) to synchronization tapping \((M \approx 300)\). This finding is important in that it indicates that the manipulation between groups was successful in altering their tapping behavior. There were other significant effects in the linear statistics, but as their theoretical import is less germane to the discussion at hand they are not reported.

\section*{PSD Analysis}

As depicted in Figure 4, both groups showed a significant decrease across condition for ITI, and no significant change across condition for KRI. The groups differed, however, in the change in \(\alpha\) for KPD. As in Experiment 1, the tap group showed a (marginally) significant increase in \(\alpha\) from continuation ( \(M=.59, S D=.23\) ) to synchronization tapping ( \(M=.75, S D=.24\) ), \(t(10)=-2.08, p=.064\). Remarkably, this effect was reversed for the hold group, showing a significant decrease from continuation \((M=.93, S D=.33)\) to synchronization tapping ( \(M=.78, S D=.27\) ), \(t(10)=2.25\), \(p=.05\). This effect buttresses the results of Experiment 1 . The fractal structure of KPD changes in the same direction as that of ITI when both of these aspects of tapping are constrained by the metronome (i.e., for the hold group).

\section*{Cross-Correlation Analysis}

The results of the cross-correlation analysis compliment the findings of the PSD analysis. As in Experiment 1, the longterm dependencies between ITI and KPD or KRI evident
during continuation tapping are absent during synchronization tapping. Recall, in Experiment 1 this pattern was reversed for the KPD-KRI cross-correlation. That is, the long-term correlations were stronger during synchronization tapping. This same effect is evident in the cross-correlations for the tap group (upper panel Figure 5). For the hold group, however, this effect is largely absent (lower panel Figure 5). As suggested by the PSD analysis, the KPD of the hold group was constrained by the metronome stimulus. As with ITI, this task constraint appears to have lessened the long-term dependency between these two aspects of the tapping dynamics.


Figure 4: Change in \(\alpha\) for ITI, KPD, and KRI across tapping conditions by experimental group.


Figure 5: KPD-KRI cross-correlation function by experimental group.

\section*{General Discussion}

The current experiments support and extend the previous findings on the fractal structure of finger tapping behavior. Although the measurement variables used in these analyses (i.e., outputs of the MIDI keyboard) might prove only a convenient window into the dynamics of finger tapping behavior, the results do reveal a consistent relationship between the long-term interplay between the different parts of the tapping behavior and how changes in task constraints affected this long-term structure. Specifically, these experiments suggest that when control of any (sub)behavior can be sustained with the aid of task constraints that behavior is effectively decoupled from other parts of the action and shows a reliable shift in its fractal structure.

Though superficially finger tapping may not seem to bear weightily on the issues of interest to the cognitive sciences, these findings do speak to larger theoretical questions about the organization of human mind and behavior. In particular, several researchers have proposed that fractal structure in human behavior reveals the "interaction-dominant" nature of the human system (see Van Orden et al., 2010). That is, these findings suggest that the behavioral in question is not the result of one dominant process (e.g., an internal timer), but instead is organized by many interdependent processes. Whereas more traditional views promote a modular, disembodied impression of the cognitive process, these findings suggest that a behavior as simple as keeping the beat is the product of non-linear interactions across the participant-task system. In short, these ideas invite reconsideration of the nature of the cognitive process and hold promise for addressing the vast complexity inherent in the complete human system.

\section*{Acknowledgments}

We thank Guy Van Orden and Christopher Kello for their kind and insightful contributions to this research.

\section*{References}

Brown, C., \& Liebovitch, L. (2010). Fractal Analysis. London: Sage.
Chen, Y., Ding, M., \& Kelso, J. A. S. (2001). Origins of timing errors in human sensorimotor coordination. Journal of Motor Behavior, 33, 3-8.
Chen, Y., Repp, B. R., \& Patel, A. D. (2002). Spectral decomposition of variability in synchronization and continuation tapping: Comparisons between auditory and visual pacing and feedback conditions. Human Movement Science, 21, 515-532.
Delignieres, D., Fortes, M., \& Ninot, G. (2004). The fractal dynamics of self-esteem and physical self. Nonlinear Dynamics, Psychology, and Life Sciences, 8, 481-510.
Delignieres, D., \& Marmelat, V. (2013). Theoretical and methodological issues in serial correlation analysis. In M. J. Richardson, M. A. Riley, \& K. Shockley (Eds.), Progress in Motor Control: Neural, computational, and Dynamics Approaches. New York: Springer.

Delignieres, D., Ramdani, S., Lemoine, L., Torre, K., Fortes, M., \& Ninot, G. (2006). Fractal analyses for 'short' time series: A re-assessment of classical methods. Journal of Mathematical Psychology, 50, 525-544.
Eke, A., Herman, P., Bassingthwaigthe, J. B., Raymond, G. M., Percival, D. B., Cannon, M., Balla, I., \& Ikrenyi, C. (2000). Physiological time series: Distinguishing fractal noises from motions. European Journal of Physiology, 439, 403-415.
Gilden, D. L. (2001). Cognitive emissions of \(1 / \mathrm{f}\) noise. Psychological Review, 108, 33-56.
Gilden, D. L., Thornton, T., Mallon, M. W. (1995). 1/f noise in human cognition. Science, 267, 1837-1839.
Holden, J. G. (2005). Gauging the fractal dimension of response times from cognitive tasks. In M. A. Riley \& G. C. Van Orden (Eds.), Tutorials in Contemporary Nonlinear Methods for Behavioral Scientists. Retrieved from http:/www.nsf.gov.
Kello, C. T., Beltz, B. C., Holden, J. G., \& Van Orden, G. C. (2007). The emergent coordination of cognitive function. Journal of Experimental Psychology: General, 136, 551-568.
Lemoine, L., Torre, K., \& Delignieres, D., (2006). Testing for the presence of \(1 / \mathrm{f}\) noise in continuation tapping data. Canadian Journal of Experimental Psychology, 60, 247257.

Madison, G. (2001). Variability in isochronous tapping: higher order dependencies as a function of intertap interval. Journal of Experimental Psychology: Human Perception and Performance, 27, 411-422.
Musha, T., Katsurai, K., \& Teramachi, Y. (1985). Fluctuations of human tapping intervals. IEEE Transactions on Biomedical Engineering, 32, 578-582.
Ogden, R. T., \& Collier, G. L., (1999). On detecting and modeling deterministic drift in long run sequences of tapping data. Communications in Statistics - Theory and Methods, 28, 977-987.
Torre, K., \& Delignieres, D. (2008). Unraveling the finding of \(1 / \mathrm{f}\) noise in self-paced and synchronized tapping: A unifying mechanistic model. Biological Cybernetics, 99, 159-170.
Van Orden, G. C., Holden, J. G., \& Turvey, M. T. (2003). Self-organization of cognitive performance. Journal of Experimental Psychology: General, 132, 331-350.
Van Orden, G. C., Holden, J. G., \& Turvey, M. T. (2005). Human cognition and 1/f scaling. Journal of Experimental Psychology: General, 134, 117-123.
Van Orden, G. C., Kloos, H., \& Wallot, S. (2010). Living in the pink: Intentionality, wellbeing, and complexity. In C. Hooker (Ed.), Philosophy of Complex Systems: Handbook of the Philosophy of Science. London: Elsevier.
Wagenmakers, E. J., Farrell, S., \& Ratcliff, R. (2005). Human cognition and a pile of sand: A discussion on serial correlations and self-organized criticality. Journal of Experimental Psychology: General, 134, 108-116.
Yamada, N. (1995). Nature of variability in rhythmical movement. Human Movement Science, 14, 371-384.

\title{
Motor experience interacts with effector information during action prediction
}

\author{
Lincoln J. Colling (lincoln@colling.net.nz) \\ Donders Institute for Brain, Behaviour and Cognition, Radboud University Nijmegen \\ Nijmegen, Netherlands
}

\author{
William F. Thompson (bill.thompson@mq.edu.au) \\ ARC Centre of Excellence in Cognition and its Disorders, Macquarie University \\ Sydney, Australia \\ John Sutton (john.sutton@mq.edu.au) \\ ARC Centre of Excellence in Cognition and its Disorders, Macquarie University Sydney, Australia
}

\begin{abstract}
Recent theory suggests that action prediction relies of a motor emulation mechanism that works by mapping observed actions onto the observer action system so that predictions can be generated using that same predictive mechanisms that underlie action control. This suggests that action prediction may be more accurate when there is a more direct mapping between the stimulus and the observer. We tested this hypothesis by comparing prediction accuracy for two stimulus types. A mannequin stimulus which contained information about the effectors used to produce the action and a point stimulus, which contained identical dynamic information but no effector information. Prediction was more accurate for the mannequin stimulus. However, this effect was dependent on the observer having previous experience performing the observed action. This suggests that experienced and naïve observers might generate predictions in qualitatively difference ways, which may relate to the presence of an internal representation of the action laid down through action performance.
\end{abstract}

Keywords: Joint action; embodied cognition; perceptionaction; action prediction.

\section*{Introduction}

Many types of joint action require two actors to coordinate their actions. Such coordination is especially demanding for joint actions, such as ensemble music and dance performance, where successful completion of the joint action requires precise temporal synchronisation. In these contexts, it is not possible for individuals to observe and then react to the actions of their co-actors because this would introduce disruptive delays. Rather, individuals must anticipate the actions of their co-actors so that they can plan actions that will align with those actions. Because of these time constraints, researchers have emphasised the role of prediction in recent theoretical accounts of joint action coordination (Csibra, 2008; Wilson \& Knoblich, 2005).

Models of predictive mechanisms in motor control, such as forward models and inverse models, can greatly inform our understanding of joint action coordination. Both classes of models are contained within the model of motor control developed by Wolpert and colleagues (e.g., Wolpert, 1997). According to this framework, forward and inverse models are used in tandem to achieve goal-directed behaviour when regular feedback is unreliable because of delays or inaccuracies.

Inverse models act as controllers by transforming a goal state into a series of control commands that are then sent to the controlled system to produce the desired behaviour. In the motor control system, this is implemented by a system that takes the goal state and transforms it into a series of motor commands. Forward models, on the other hand, take the motor commands and transform them into a goal state. The limb also performs a forward mapping from motor commands into a goal state. Therefore, the forward model can be used to predict how the limb is expected to behave. A forward model is particularly useful in motor control where it can be used to bypass delays that occur because feedback must be transmitted from the periphery to centrally located motor control regions. This can compensate for these delays by generating predicted feedback that can be substituted for the delayed feedback.

\section*{Motor involvement in action prediction}

In addition to their role in intrapersonal action prediction, forward models and inverse models are also implicated in the interpersonal action prediction needed for joint action. Csibra (2008) has suggested that during action observation, an inverse model allows observers to reconstruct the motor codes used to produce the observed action. To support this claim, Csibra cites evidence from electrophysiological studies on monkeys as well as neuroimaging studies from humans that show that neurons in motor regions are active not only when actions are produced but also when the same, or similar, actions are passively observed (for a review, see Rizzolatti \& Craighero, 2004). Additionally, Wilson and Knoblich (2005) have proposed that observers are able to construct an internal model of observed actions by mapping the actions onto their own motor systems in a part-by-part, or isomorphic, manner. This internal model acts as a forward model by generating a real-time simulation of the observed action that runs in parallel with incoming sensory information. Information from this model can be substituted for incoming sensory information that reaches the observer through observation. By using internally generated information to drive action planning, delays that result from the processing of external actions can be overcome, and this allows co-actors to plan and coordinate


Figure 1: The zigzag (left) and wave (right) patterns used as stimuli during the recording session.
joint actions in time critical situations. Taken together, forward and inverse models provide an effective mechanism that allows actors to use their own action systems in order to emulate \({ }^{1}\) the actions of others. We refer to this proposal as the emulator hypothesis of action prediction.

A key prediction of the emulator hypothesis is that traces of the observer's motor system should be manifested in the predictions that they generate. The authorship effect provides a means of assessing this. The authorship effect refers to the finding that observers are more accurate at generating predictions about recordings of self-generated actions relative to other-generated actions. More generally, the greater the alignment between the motor dynamics of the observer and the motor dynamics of the agent producing the observed action, the more accurate the predictions generated by the observer (Flach, Knoblich, \& Prinz, 2003; Keller, Knoblich, \& Repp, 2007; Colling, Sutton, \& Thompson, 2010, submitted).

\section*{Motor involvement in event prediction}

In addition to the motor system's role in predicting the actions, evidence from paradigms employing abstract stimuli suggest that the motor system might also be used for sequence prediction in general. For example, findings from fMRI implicate ventral premotor regions in tasks that require participants to generate predictions about abstract sequences (Schubotz \& von Cramon, 2004). Similarly, lesions in premotor regions are associated with deficits in sequence prediction (Schubotz, Sakreida, Tittgemeyer, \& von Cramon, 2004).

Based on these findings, Schubotz (2007) has suggested that motor simulation is a general mechanism for predicting events. In the case of reproducible events-that is, human actions-these events are simulated using the same means that were initially employed to create the event, by using an internal model of the action. However, in the case of event that can't be mapped onto the body Schubotz argues that predictions are generated using an action model of an effector that best matches the general dynamics of the stimulus. Similarly, impoverished action stimuli lacking detail about which effectors were used to produce the action, or actions that are not in the observers repertoire, might be simulated using this more general mechanism. While this might provide a good general description of the stimulus dynamics it may fail to replicate fine-grained details of the stimuli.

\footnotetext{
\({ }^{1}\) Emulate to refers to the process of replicating the functions of a system (e.g., a conspecific's motor system) using different means (e.g., the observer's motor system).
}


Figure 2: Marker positions for recording session.

\section*{Aims of the current study}

The primary aim of the present study is to examine the nature of the internal model that observers use during action prediction. In particular, our aim is to examine whether action prediction is achieved via a general purpose predictive system that, although implemented with the motor system, does not rely on a part-by-part simulation of the observed action. The present study measured prediction accuracy by means of an action synchronisation task similar to that reported in Colling et al. (2010, submitted); however, rather than examining differences in synchronisation accuracy for self-produced and other-produced actions, all participants viewed other-produced actions and we instead varied the properties of the stimulus as well as the relevant motor experience of the observers. Both manipulations were designed to modify the information that participants could access to allow them to map the observed actions onto their own action systems. The motor experience manipulation was designed to provide observers with an internal representation of the action onto which they could map the stimulus, while the stimulus manipulation was designed to modify whether the stimulus could be directly mapped onto the observers' bodies.

To modify the information content in the stimulus, we constructed two sets of stimuli so that the stimulus either contained information about what effectors were employed to produce the action (full information) or only contained the motion information required to perform the synchronisation task, but excluded any information about the effectors used to produce the action (point information).

A manipulation was also designed to examine the role of motor experience on action emulation. This was achieved by dividing the participants into two groups and only providing one of the groups with experience with actually performing the action that they would later observe. Schubotz (2007) has suggested that when observers predict actions that are part of their action repertoire, they emulate the actions using an internal model of that action that has been laid down by the experience of producing the action. Naïve observers, on the other hand, might only employ motor regions that match the general dynamics of the movement. If this is the case, then we can predict that the effect of stimulus content would be modulated by motor experience. In particular, we can predict that naïve observers would not incorporate information about the effectors used to produce the movement into their predictive


Figure 3: Example stimuli from the full information condition (left) and the point information condition (right).
model; thus, the addition of this information should provide no additional benefit on the synchronisation task.

\section*{Methods}

\section*{Participants}

The motor experience group contained of 13 participants (11 females, mean age of 28.1 years). The naïve group contained of 12 participants ( 8 females, mean age of 20.7 year). All participants were right-handed, and all procedures were approved by the Macquarie University Human Subjects Ethics committee.

\section*{Stimuli}

In order to create the stimuli for the test session, five righthanded females (mean age of 24.8 years) performed the movement task while their movements were tracked with motion capture.

The movement task involved tracing out wave and zigzag patterns (see Figure 1) as if drawing them on an imaginary blackboard. The patterns were displayed on two large sheets of cardboard measuring \(0.594 \mathrm{~m} \times 0.841 \mathrm{~m}\). Both patterns contained five upward and five downward movements alternating between long and short. The two patterns differed in terms of the nature of the direction change at the apex of each upward movement. The direction changed sharply for the zigzag pattern, while there was a smooth, flowing direction change for the wave pattern \({ }^{2}\).

Movements were recorded using an 8-camera 3-D passive optical motion capture system (Vicon MX with 4 Vicon MX-F20 and 4 Vicon MX13+ cameras) at a sampling rate of 200 Hz from markers placed on the subject's shoulders, right arm, right hand, and waist (see Figure 2). Raw motion capture data was resampled to 25 Hz and processed with CMotion Visual 3D (C-Motion INC, Rockville MD) to create the test stimuli. For the full information condition, the motion

\footnotetext{
\({ }^{2}\) The difference between the wave and zigzag patterns is of no theoretical interest. Two movement patterns were used only to increase task variety during the stimulus creation phase and ensure participants remained engaged with the task. Statistical analyses confirmed that there were no systematic differences in performance on the test phase as a result of stimulus form (wave, zigzag) and, therefore, this factor was dropped from the analyses reported below.
}


Figure 4: Improvement in synchronisation accuracy for the full information condition relative to the point information condition for the motor experience group and the naïve group. Error bars indicate the \(95 \%\) confidence interval.
capture data was rendered as an animated character consisting of an upper torso, right arm and hand, while the stimuli for the point information condition consisted of only a single point tracking the hand (see Figure 3).

\section*{Procedure}

Participants in the motor experience group undertook a movement session that was identical to the task employed during stimulus creation. Participants performed 3 blocks containing 5 repetitions of each pattern (in random order) with their eyes closed to limit visual experience. The movement session and the test session were on average separated by 15.85 days ( 7 to 27 days).

The task in the test session was to press the response button when the hand of the mannequin, or the marker tracking the hand, reached the apex of each upward movement. Participants were instructed to synchronise the button-press with the display as accurately as possible and were told that this may require them to anticipate when the peak will occur. Each participant performed 4 blocks containing 40 unique stimuli, with equal numbers of full and point stimuli, and equal numbers of wave and zigzag stimuli. Participants in the naïve group were given a brief verbal description of the movement task.

\section*{Results}

Timing error was calculated as the absolute difference between the timing of the peak in the motion capture trajectory and the timing of the button-press. Absolute timing error was used as a dependent measure because it has been shown to provide a good index of accuracy of hitting the target (Spray, 1986). Absolute timing error was analysed by means of a 2 \(\times 4 \times 2\) mixed ANOVA with the within-subjects factors Information Content (full information, point information) and Block (1, 2, 3, and 4), and the between-subjects factor of Experience (motor experience, naïve). The Greenhouse-Geisser procedure was used to correct for violations of sphericity.


Figure 5: (A) Evenly spaced button-presses results in timing errors that vary as a function of peak number. (B) Timing error that does not vary as a function of peak number is a result of the timing of button-presses varying as a function of peak number.

Where appropriate, we report uncorrected \(d f\) s along with the corrected \(p\) value.

There were no systematic differences in synchronisation accuracy related to experimental block, as indicated by the non-significant main effect for Block ( \(F_{3,69}=0.250, p=.861\), \(\varepsilon=.518, \eta_{G}^{2}=.002\) ), and the non-significant interactions for Information Content \(\times\) Block \(\left(F_{3,69}=0.368, p=.777\right.\), \(\left.\eta_{G}^{2}=.001\right)\), Block \(\times\) Experience ( \(F_{3,69}=1.024, p=.352, \varepsilon=\) \(\left..518, \eta_{G}^{2}=.006\right)\) and Information Content \(\times\) Block \(\times\) Experience \(\left(F_{3,69}=0.609, p=.611, \eta_{G}^{2}=.001\right)\). There were also no systematic differences in synchronisation accuracy between the naïve group and the group with motor experience, as indicated by the non-significant main effect of Experience \(\left(F_{1,23}=0.460, p=.504, \eta_{G}^{2}=.016\right)\). Furthermore, there were no systematic differences in synchronisation accuracy between the full information displays and point information displays when the data were collapsed across group and block, as indicated by the non-significant main effect for Information Content ( \(F_{1,23}=8.573, p=.008, \eta_{G}^{2}=.003\) ).

As predicted, the results showed that the effect of information content was modulated by motor experience, as indicated by the significant interaction for Information Content \(\times\) Experience ( \(F_{1,23}=5.413, p=.029, \eta_{G}^{2}=.002\) ). To decompose this interaction, the data were collapsed across block and two paired t-tests were conducted to examine the difference between the two levels of Information Content (full information, point information) for each Experience group. The results of these \(t\)-tests showed that the information content effect was found only for the motor experience group ( \(t_{12}=2.943, p=.012\) ) but not the naïve group \(\left(t_{11}=-0.411\right.\), \(p=.689\) ). This indicates that timing error was significantly higher for the point stimuli relative to full stimuli for the motor experience group ( \(M_{\Delta}=6.855\), 95CI[1.779][11.930]) but not for the naïve group ( \(M_{\Delta}=-1.007\), 95CI[-6.399][4.385]). Therefore, only the motor experience group was able to take


Figure 6: Timing error as a function of peak position for the motor experience group and the naïve group.
advantage of the presence of limb and joint information to enhance synchronisation accuracy (See Figure 4).

A further attempt was made to quantify the difference in task performance between the motor experience group and the naïve group. We examined whether there were differences in task performance that related to whether participants primarily responded to local aspects or global aspects of the stimulus. In the stimuli, the duration of each upward movement alternated from long to short. This irregular pattern leads to local variations in peak timing. Basing responses on global aspects of the stimuli, such as average tempo, would produce a pattern of timing errors that fluctuates from peak to peak. However, by adjusting responses according to the local variations in the stimuli would produce timing errors that are approximately equal for each button-press (see Figure 5). Timing error for each of the final four button-presses (corresponding to each of the final four peaks) was analysed separately for each group by means of a one-way ANOVA with the factor Peak Position (The first peak was dropped from the analysis because the movement leading up to the first peak is neither clearly long nor short).

Analyses showed a significant effect of Peak Position for the naïve group \(\left(F_{3,33}=5.083, p=.031, \varepsilon=.453, \eta_{G}^{2}=.108\right)\), and not for the motor experience group \(\left(F_{3,36}=1.449\right.\), \(p=.254, \varepsilon=.371, \eta_{G}^{2}=.020\) ). This suggests that for the naïve group timing error changed in a low-high-low-high pattern as the trial progressed, while for the motor experience group peak position did not significantly affect timing error. These results are consistent with the naïve group responding to global aspects of the stimuli and the motor experience group responding to local aspects of the stimuli (see Figure \(6)\).

\section*{Discussion}

The primary aim of the present study was to investigate the nature of the action emulation employed during action prediction. In particular, we wanted to investigate whether action prediction relies on observers mapping the stimulus onto their body in a part-by-part manner, or whether they just model
the general dynamics of the action without modelling the specifics of the effectors used to create the stimulus. Furthermore, our aim was to investigate the influence of motor experience on action emulation. Schubotz (2007) has suggested that while abstract stimuli and actions that are not part of the observer's repertoire might be simulated using a general purpose mechanism, actions an observer has experience producing are instead simulated using a model that incorporates the specifics of the effectors used to produce the movement. In order to examine these questions, we varied both the information content of the stimuli and the motor experience of the observers.

\section*{The effect of stimulus information}

In the full information condition, the visual stimulus contained not only the movement information required to perform the task, but also information about the effector used to produce the movement. In the point information condition, the visual stimulus only contained a single moving point. While the point information condition also contained all the movement information required to perform the task, it lacked the additional information about the state of the effectors. As predicted, the results showed enhanced prediction accuracy when observing the full information stimuli.

An alternative explanation for these results is that a difference in the low-level visual features of the stimuli might account for the differences in prediction accuracy. For example, it might be the case that the point information stimulus, which overall contains less visual information, is harder to visually track, and this may manifest as decreased prediction accuracy. However, this could account for the differences in synchronisation accuracy, then this difference should be present in both the experienced and the naïve groups. This was not the case; therefore, the results are not consistent with an explanation based on low-level visual features.

The finding that limb and joint information was able to enhance synchronisation accuracy is also consistent with what is known about the mirror-neuron system, the putative substrate of the action prediction system (Csibra, 2008; Wilson \& Knoblich, 2005). Studies by Buccino and colleagues (e.g., Buccino et al., 2004) have shown that during action observation, regions of the motor cortex are activated in a somatatopic fashion. That is, certain regions show specificity for particular effectors in a manner similar to Penfield's (1954) motor homunculus.

\section*{The effect of motor experience}

A secondary aim of the present study was to examine what influence motor experience would have on prediction accuracy. The results show that motor experience modulated the effect that stimulus type had on prediction performance. In addition, motor experience had an effect on how participants performed the task. In particular, these data show that while the timing error for experienced participants was not affected by the serial position of the peak, the timing error for naïve participants varied according to peak number, and the tim-
ing error for the small peaks was significantly different to the timing error for large peaks.

A pattern of fluctuations would arise if participants kept a relatively steady pace throughout the trial because the spacing of the peaks was not constant throughout the trial, but instead changed according to the height of the peaks. A relatively steady pace for button-presses might occur if participants responded to the global properties of the stimuli, such as the average rhythm (that is, the pace of movement production), or to the general stimulus dynamics. In order to maintain a relatively constant timing error, as seen in the experienced group, participants would need to adjust the timing of each button-press according to the local timing variations in the stimuli that result from the alternating heights of the peaks. This pattern of data, therefore, suggests a global/local bias in stimulus processing that is modulated by motor experience.

The effect of motor experience on processing visual stimuli has recently been noted in several studies. For example, Casile and Giese (2006) have shown that motor training enhances a participant's ability to make a fine-grained visual discrimination of action. In their study, participants were asked to make same/different judgments about gait patterns that they either did or did not have motor experience with. The results showed that participants performed significantly better for trained, or familiar, gait patterns compared with unfamiliar gait patterns, suggesting that in order to make fine-grained visual judgments about the kinematics of an action, observers need to have an internalised model of the action. Similarly, Calvo-Merino, Ehrenberg, Leung, and Haggard (2010) found superior performance on a visual discrimination task of dance moves when those dance moves were part of the observers action repertoire compared with dancers who only had visual experience with the dance moves.

Our findings build on these earlier results, and suggest that observers with motor experience for the observed action are also better able to make fine-grained predictions about the dynamics of a stimulus. This result is consistent with the notion that experienced observers generate predictions about observed actions by employing an internal model of that action that is acquired through motor experience. By mapping the observed action onto their internal model for that action they are better able to capture the fine-grained timing variations in the stimulus because their predictive model more completely captures the constraints specific to the effectors used to produce the action.

\section*{Motor experience modulates stimulus effects}

We have argued that participants with motor experience are more sensitive to the fine-grained timing differences present in an action because they, unlike naïve observers, employ an internalised model of the observed action in order to generate their predictions. This might also help to explain why the influence of limb and joint information was restricted to the group with motor experience. Several neuroimaging studies have shown experience-related differences in motor system activation when observers view actions performed by other
people. For example, Calvo-Merino, Glaser, Grézes, Passingham, and Haggard (2005) asked expert dancers and nondancer controls to view videos of dancers performing in one of two styles (ballet or capoeira). The results showed that activation in motor regions was greater when dancers viewed performers of their own style, suggesting that the motor system is preferentially engaged when observing actions that are familiar. A follow-up study (Calvo-Merino, Grézes, Glaser, Passingham, \& Haggard, 2006), using male and female ballet dancers extended this finding by showing that motor regions were preferentially activated when viewing gender-specific dance moves. As both male and female dancers presumably have equal visual experience with opposite gender dance moves, but different motor experience, this finding suggests that motor engagement with visually presented stimuli is selective for actions for which observers have specific motor familiarity over and above the effects of visual familiarity.

Studies by Schubotz and colleagues (for a review, see Schubotz, 2007) have also implicated premotor regions in prediction of abstract stimuli and in sequence prediction in general. These stimuli cannot be mapped onto the observer's body and, therefore, they might rather be predicted by using a predictive model that exploits the dynamics of an effector that most closely matches the dynamics of the stimulus. Similarly, differences in motor system activation related to motor experience might suggest that inexperienced observers employ general predictive mechanisms, such as those used for sequence prediction, even when the observed action can, at least in principle, be mapped onto their body in an isomorphic, or part-by-part, manner. The addition of limb and joint information was designed to assist the process of mapping the observed action onto the observer's body; however, if, as the neuroimaging data cited above suggests, naïve observers less readily map the observed action onto their body in an isomorphic manner, then providing information to assist this process should provide no additional benefit. This is indeed what was found in the present study.

\section*{References}

Buccino, G., Lui, F., Canessa, N., Patteri, I., Lagravinese, G., Benuzzi, F., et al. (2004). Neural circuits involved in the recognition of actions performed by nonconspecifics: An fMRI study. J Cog Neuro, 16(1), 114-126.
Calvo-Merino, B., Ehrenberg, S., Leung, D., \& Haggard, P. (2010). Experts see it all: configural effects in action observation. Psych Res, 74, 400-406.
Calvo-Merino, B., Glaser, D. E., Grézes, J., Passingham, R. E., \& Haggard, P. (2005). Action observation and acquired motor skills: An fMRI study with expert dancers. Cereb Cortex, 15(8), 1243-1249.
Calvo-Merino, B., Grézes, J., Glaser, D., Passingham, R., \& Haggard, P. (2006). Seeing or doing? influence of visual and motor familiarity in action observation. Curr Bio, 16(19), 1905-1910.
Casile, A., \& Giese, M. A. (2006). Nonvisual motor training
influences biological motion perception. Curr Bio, 16(1), 69-74.
Colling, L. J., Sutton, J., \& Thompson, W. F. (2010). Action synchronisation with biological motion. In W. Christensen, E. Schier, \& S. J (Eds.), ASCS2009: Proceedings of the 9th Conference of the Australasian Society for Cognitive Science (pp. 49-56). Sydney: Macquarie Centre for Cognitive Science.
Colling, L. J., Sutton, J., \& Thompson, W. F. (submitted). Authorship effects in action synchronisation with biological motion.
Csibra, G. (2008). Action mirroring and action understanding: An alternative account. In P. Haggard, Y. Rossetti, \& M. Kawato (Eds.), Sensorimotor foundations of higher cognition (pp. 435-459). New York: Oxford University Press.
Flach, R., Knoblich, G., \& Prinz, W. (2003). Off-line authorship effects in action perception. Brain Cogn, 53(3), 503-513.
Keller, P. E., Knoblich, G., \& Repp, B. H. (2007). Pianists duet better when they play with themselves: On the possible role of action simulation in synchronization. Conscious Cogn, 16(1), 102-111.
Penfield, W. (1954). Mechanisms of voluntary movement. Brain, 77(1), 1-17.
Rizzolatti, G., \& Craighero, L. (2004). The mirror-neuron system. Ann Rev Neuro, 27(1), 169-192.
Schubotz, R. I. (2007). Prediction of external events with our motor system: towards a new framework. TiCS, 11(5), 211-218.
Schubotz, R. I., Sakreida, K., Tittgemeyer, M., \& von Cramon, D. Y. (2004). Motor areas beyond motor performance: Deficits in serial prediction following ventrolateral premotor lesions. Neuropsychology, 18(4), 638-645.
Schubotz, R. I., \& von Cramon, D. Y. (2004). Sequences of abstract nonbiological stimuli share ventral premotor cortex with action observation and imagery. J Neuro, 24(24), 5467-5474.
Spray, J. A. (1986). Absolute error revisited: An accuracy indicator in disguise. J Motor Behav, 18(2), 225-238.
Wilson, M., \& Knoblich, G. (2005). The case for motor involvement in perceiving conspecifics. Psych Bull, 131(3), 460-473.
Wolpert, D. M. (1997). Computational approaches to motor control. TiCS, 1(6), 209-216.

\title{
Models of Human Category Learning: Do They Generalize?
}

\author{
Nolan Conaway (nconawa1@binghamton.edu) \\ Kenneth J. Kurtz (kkurtz@binghamton.edu) \\ Department of Psychology, Binghamton University Binghamton, NY 13905 USA
}

\begin{abstract}
Generalization to new examples is an essential aspect of categorization. However, recent category learning research has not focused on how people generalize their category knowledge. Taking generalization to be a critical basis for evaluating formal models of category learning, we employed a 'minimal case' approach to begin a systematic investigation of generalization. Human participants received supervised training on a two-way artificial classification task based on two dimensions that were each perfect predictors. Learners were then asked to classify new examples sampled from the stimulus space. Most participants based their judgments on one or the other dimension. Varying the relative levels of dimension salience influenced generalization outcomes, but varying category size ( 2,4 , or 8 items) did not. We fit two theoretically distinct similarity-based models (ALCOVE and DIVA) to aggregate learning data and tested on the generalization set. Both models could explain important aspects of human performance, but DIVA produced a superior overall account.
\end{abstract}

Keywords: generalization; categorization; formal models of category learning; similarity; cognitive modeling.

\section*{Introduction}

Categorization is an essential cognitive function categories serve to organize knowledge and, critically, as a basis for extending knowledge to make sense of new experience. A full understanding of human categorization depends on developing models and theories that account for systematic patterns of human learning and generalization performance (for an overview of generalization, see Levering \& Kurtz, 2010).

In classic research, Roger Shepard \((1957,1987)\) put forth the idea of a universal law in which stimulus generalization follows an exponential function of distance in psychological space. This work has had broad implications for theoretical models of categorization. Highly influential reference point models (such as the exemplar view) compute classification in a manner that closely follows Shepard's proposal. Specifically, the class membership of a known item is likely to be generalized to a new item if the two items are highly similar. The key additional design feature needed to account for human classification performance is the inclusion of a selective attention mechanism such that particular dimensions can matter more or less in the computation of similarity. Generalization performance (classification of previously unseen items) has been one of the most important important testing grounds in the debate between exemplar- and prototype-based accounts
(e.g., Homa, 1984; Nosofsky, 1992; see also Medin \& Schaffer, 1978 and the ensuing literature on behavioral experimentation and model-fitting with the 5-4 classification problem).

In a somewhat different approach to studying the generalization of category knowledge, researchers have investigated whether exemplar models can account for rulelike generalization after category learning (Erikson \& Kruschke, 1998, 2002; Nosofsky \& Johansen, 2000). In these studies, participants were asked to classify novel instances after learning an artificial two-way classification based on a unidimensional rule with exceptions. The critical test items were highly similar to the exceptions, but clearly classifiable using the rule. The outcomes of these studies were somewhat mixed and appear to depend on stimulus attributes and also on the structure of the categories that are learned.

The goal of the present research is two-fold: 1) to explore a different approach to investigating the psychology of category generalization; and 2) to use generalization performance as a basis to compare and differentiate models that are highly successful in fitting human learning data. Toward the first goal, our experimental approach is broadly comparable to the psychological studies of generalization discussed above: after a learning phase, participants are asked to classify novel examples. However, our work differs in that we use minimal category learning conditions (small numbers of examples that are readily assigned to two fully coherent classes). Our primary aim is to identify basic, systematic properties of generalization performance.

Regarding the second goal, the field presently offers a small group of formal models of category learning that are general purpose (applicable to any classification problem), that provide explanation at the level of process/mechanism, and that yield good fits to established benchmarks for human category learning. Within the realm of fitting human classification learning performance, there is some sense of having hit the ceiling in terms of differentiating among these models despite their having distinct explanatory elements. Our rationale is that models that do quite well in fitting learning data may diverge in their ability to account for patterns of generalization performance. In particular we are compelled by the prospect of fitting model parameters to the learning data and then holding the models to these values in evaluating ensuing generalization (as discussed below). Toward this end, we evaluate two successful models: a canonical representative of the reference point approach, ALCOVE (Kruschke, 1992) and an updated
version of a competing theoretical alternative, DIVA (Kurtz, 2007).

ALCOVE. ALCOVE is an exemplar based adaptive network model. According to the model, categories are represented by individual exemplars stored in memory. ALCOVE learns to classify by adjusting association weights between exemplar nodes and category nodes, as well as by adjusting a set of attention weights that determine the importance of each stimulus dimension.

DIVA. DIVA offers a more generative than discriminative approach to classification learning and deals in distributed rather that localist internal represenations. Learning to classify examples is accomplished by minimizing reconstructive error along the channels of a divergent autoencoder that is comprised of recoding (input \(\rightarrow\) hidden) weights shared for all categories and separate sets of decoding (hidden \(\rightarrow\) output) weights dedicated to each category. Classification judgments are based on which category channel yields the lowest error, i.e., which channel has been tuned to expect (and successfully reconstruct) a set of features like those of the current item.

DIVA is similarity-based in the sense that the model learns, for each category, how to effectively predict feature values for particular regions in recoding space - when an input item projects into a region that is well handled by a category, the reconstructive error in predicting the features will be low. DIVA does not apply Shepard-like stimulus generalization to categorization - an item is likely to belong to a category because its feature values conform to what a category channel has been optimized to successfully recode and decode, not because it is highly similar to a known member of the category.

Our approach to model comparison. We compare models based on their ability to account for human generalization after category learning. An important advantage of focusing on generalization performance is that we avoid the traditional reliance on post-hoc fits. In all cases, we first fit DIVA and ALCOVE to averaged learning data from each condition in order to find best-fitting parameters across the full set of conditions. This procedure allows us to separate out the parameter fitting process, so that the generalization performance is genuinely a prediction based on a selected model.

We elected to fit ALCOVE using a grid search over its response mapping ( \(\phi\) ), specificity constant (c), association weight learning rate, and attention learning rate parameters. We also fit DIVA using a grid search over the parameters: learning rate, weight range, number of hidden nodes, and a new focusing parameter ( \(\beta\) ) that gives DIVA the ability to account for sensitivity to differences in dimension diagnosticity (Kurtz, 2008).

DIVA's focusing parameter ( \(\beta\) ) allows it to selectively attend to stimulus dimensions based on the disparity in the output activations for that dimension across category channels. DIVA's focusing mechanism differs significantly from selective attention in ALCOVE in that it does not change the encoding of the stimulus or manipulate the representation learned by the model. DIVA's form of focusing is decisional, rather than perceptual or representational in nature, as it operates at the level of the classification response.

\section*{Experiment 1}

This experiment was designed to explore generalization under two conditions: when all stimulus dimensions are diagnostic and equally salient; and when all dimensions are diagnostic, but unequally salient. Figure 1 depicts the two category structures.

Stimulus scaling is an important aspect of our salience manipulation. In order to determine the relation between the stimulus dimensions, we scale the examples in a pairwise similarity study. The similarity study generates a full set of scaled examples, which allows us to manipulate the distance between examples on any dimension. The salience of a dimension can be specified by the distance between the categories on that dimension.

In a pilot study, we explored an extreme case of classification learning in which both stimulus dimensions were diagnostic, but one dimension was much less salient. Participants were generally insensitive to variation in the less salient dimension. In light of these findings, we expected that generalization gradients would show sensitivity given a relatively moderate difference in dimension salience.

Participants and Materials. 108 undergraduates from Binghamton University participated in partial fulfillment of a course requirement. Stimuli were rectangles varying in shading and the distance between two lines within the rectangle. Examples were generated at 8 positions on each dimension ( 8 shading \(* 8\) line spacing \(=64\) examples). The category structures are depicted in Figure 1 along with sample stimuli.

Procedure. Participants were randomly assigned to either the equal salience group or the unequal salience group. In the equal salience condition, the category prototypes were separated by distances of 0.64 and 0.54 on the first and second dimensions (shading and line spacing), respectively. In the unequal salience condition, the category prototypes were separated by a distance of 0.65 and 0.34 on the first and second dimensions. In each condition, there were 4 training examples in each of the two categories.


Figure 1: Top: Four examples of stimuli (taken from the corners of the stimulus space). Bottom left \& center: Category structures with equally and unequally salient dimensions. Bottom right: Test set used for Experiments 1 and 2. Note that all training items are included in the test set. Positions of examples reflect prior scaling.

Each participant completed 32 learning trials. On each trial, a training item was presented on the computer screen and participants were prompted to make a classification decision by clicking one of two buttons (labeled 'Alpha' and 'Beta'). After responding, participants were given corrective feedback on their response. In the test phase, participants classified the 64 examples sampled across the stimulus space (test set depicted in Figure 1). The 8 training items were also presented during the test phase.

Gradient Analysis. In the test phase, participants provide data that yield a generalization gradient of their classification responses. For each participant, we calculated the standard deviation of classification responses at 8 positions on each dimension of the gradient. We then estimated sensitivity to each dimension by calculating the mean of these 8 values. Insensitivity to a dimension is indicated by uniformity of classification responses across that dimension.

Results and Discussion. 24 participants were excluded
from the subsequent analyses for failing to correctly classify 7 out of 8 training items presented during the test phase. The remaining participants were more than \(96 \%\) accurate during the last training block in both conditions.

There were significant individual differences in the generalization data. A k-means analysis revealed three profiles based on the sensitivity estimates described above: these were unidimensional generalization based on either one or the other stimulus dimension (shading or spacing) and multidimensional generalization based on both dimensions. We compared the k-means findings across salience conditions (results are shown in Figure 2).

While a very few participants were sensitive to both dimensions at test, the majority of participants generalized undimensionally. A Fisher's Exact test revealed that the rate of each unidimensional profile differed between salience conditions ( \(p<.001\) ). Participants in the unequal salience condition were more likely to be sensitive to the salient dimension (shading) than participants in the equal salience condition.

We observed a bias towards the line spacing dimension in the equal salience group that is not consistent with the scaling. Interestingly, this may reflect a task difference between pairwise similarity and classification learning that renders participants differentially sensitive to our stimulus dimensions.

The main conclusions we can draw from this study of a 'minimal case' category structure are that: 1) participants tended to generalize according to a single dimension despite an optimal diagonal bound; and 2) dimension salience increased the likelihood of the dimension serving as the basis for generalization.

Modeling Analyses. We tested DIVA and ALCOVE for their ability to account for these generalization findings. Specifically, we sought to determine whether the models could account for: (1) the tendency of learners to generalize based on a single dimension; (2) the substantial degree of selection of each of the two dimension as the focal one by different sets of learners; and (3) the effect of salience on dimensional sensitivity.

Before generating predictions for generalization, we obtained optimal parameter sets by fitting the models to the aggregate learning data (minimizing the sum of squared deviations, SSD, across learning blocks). We then generated predictions for generalization across a range of optimal

Table 1: Parameter values for ALCOVE and DIVA that best fit all conditions of learning performance in Experiments 1 and 2.
\begin{tabular}{|c|c|c|c|c|c|}
\hline \multicolumn{3}{|c|}{ALCOVE} & \multicolumn{3}{|c|}{DIVA} \\
\hline & Experiment 1 \(S S D<.003\) & Experiment 2
\[
S S D<.06
\] & & Experiment 1
\[
S S D<.004
\] & \[
\begin{gathered}
\text { Experiment } 2 \\
S S D<.03
\end{gathered}
\] \\
\hline c (specificity) & 3.4 & 10.5 & number of hidden nodes & 1 & 1 \\
\hline \(\phi\) (response mapping) & 2.8 & 1.45 & \(\beta\) (focusing) & 20 & 80 \\
\hline attention learning & 0.0 & 0.0 & learning rate & 0.14 & 0.18 \\
\hline association learning & 0.1 & 0.3 & initial weight range & +/-0.5 & +/-1.5 \\
\hline
\end{tabular}
parameter sets to gain a full understanding of how the two models performed.


Figure 2: Results of k -means clustering results for Experiment 1. Number of participants shown below each chart.

Both ALCOVE and DIVA provided good fits to the learning data under a range of parameters. The best fitting parameter sets are shown in Table 1. When we tested the models on generalization using these parameters, we found that both models were sensitive to the salience of each dimension, but neither predicted the unexpected bias toward the line spacing dimension that was observed behaviorally.

ALCOVE's attention learning parameter largely governed the model's ability to generalize to a single dimension. ALCOVE produced unidimensional gradients with high levels of attention learning and multidimensional gradients with low levels of attention learning. Given a high attention learning parameter, ALCOVE generalized based on whichever dimension was most salient. We note that ALCOVE lacks any random element such as initial weight values, so the output is deterministic; for this reason, the model does not account for the heavy use of both possible unidimensional rules in the generalization data. Future research will explore generalization using a stochastic version of ALCOVE.

Similar to ALCOVE's use of attention, DIVA's focusing parameter allowed the model to generate either unidimensional or multidimensional gradients. But unlike ALCOVE, DIVA is initialized with random weights on every run. An analysis of results on individual runs revealed that when DIVA's focusing parameter was large and the dimensions were equally salient, the random initial weights sometimes lead to unidimensional generalization based on
either dimension. With larger weight ranges, DIVA produced varied distributions of generalization profiles.

Our analysis of DIVA's generalization also revealed that, with a high focusing parameter, the model is more likely to generalize based on a salient dimension than a less salient dimension. This trend resembles the effect of salience that was observed previously. When the dimensions are equally salient, DIVA tends to produce multidimensional profiles at a greater rate than would be predicted given our behavioral findings.


Figure 3: Category structures for Experiment 2.
These modeling results confirm that generalization provides a promising basis for model evaluation. We found that DIVA and ALCOVE produce generalization gradients that are consistent with the salience of each dimension, and that attentional mechanisms allow similarity-based models to generate unidimensional gradients. Furthermore, a random component can partially explain variability in dimensional selection.

\section*{Experiment 2}

This study was designed to replicate and extend Experiment 1. As in the first study, we manipulated the salience of dimensions by modifying the distance between the two categories. We extend the design by incorporating category size as a between-participants facto (Figure 3 depicts the category structures that were employed). Category size is a potentially interesting factor in our studies because increasing the number of examples in each category also increases variation in representational demands for exemplar models like ALCOVE without altering the solution that the model is required to find. Furthermore, increases in category size should decrease the memorizability of each example (see Homa, 1984 for background on category size effects).

Our primary predictions were that: (1) generalization after learning would reflect sensitivity to a salient
dimension; and (2) shifts in category size would impact the prevalence of integrated, multidimensional generalization.

Table 2: Distance between opposite-category prototypes on each dimension.
\begin{tabular}{|l|l|l|l|l|}
\multicolumn{1}{c}{} & \multicolumn{2}{c}{ Equal Salience } & \multicolumn{2}{c}{ Unequal Salience } \\
\cline { 2 - 5 } \multicolumn{1}{c}{} & Shading & Spacing & Shading & Spacing \\
\hline 2 eg & 0.70 & 0.55 & 0.72 & 0.34 \\
4 eg & 0.67 & 0.55 & 0.69 & 0.34 \\
8 eg & 0.64 & 0.54 & 0.65 & 0.34 \\
\hline
\end{tabular}

Participants and Materials. 228 undergraduates from Binghamton University participated in this experiment toward partially fulfillment of a course requirement. The materials were like those used in Experiment 1.

Procedure. Participants were randomly assigned to one of six conditions ( 2 levels of salience x 3 levels of category size). The category structures are depicted in Figure 3. Participants learned a classification based on two, four, or eight unique examples per category.


Figure 4: Experiment 2 k -means clustering results. Number of participants shown below each chart.

The salience manipulation was similar to that used in Experiment 1 with one departure - we partially re-arranged the members of the second category so that the category prototypes would more evenly spaced apart in the equal salient condition. The distances between prototypes for each condition are shown in Table 2. All other aspects of the procedure are identical to Experiment 1.

Results and Discussion. 56 participants were excluded from subsequent analyses because they made more than one error on training items presented during the test phase. The remaining participants were more than \(94 \%\) accurate during the last training block.

The analysis of the generalization data was conducted as in the first study. Results are displayed in Figure 4. The data do not reveal an effect of category size on generalization. Consequently, our discussion focuses on the salience manipulation across category size conditions.

As in Experiment 1, the majority of participants generalized to a single dimension. A Fishers Exact test (conducted across size groups) reveals a significant effect of salience ( \(p<.01\) ). Participants in the unequal salience group tended to generalize using the salient dimension over the less salient dimension.

We observed the same bias towards the line spacing dimension in the equal salience conditions: our participants were highly sensitive to the line spacing dimension, even when the scaling revealed that the dimensions were equally salient.

Modeling Analyses. We again tested DIVA and ALCOVE on their ability to match human generalization performance. In general, the modeling results for Experiment 2 parallel the results of Experiment 1. Both models found good fits to the aggregate learning data, but neither model predicted the unexpected bias towards the line spacing dimension during generalization. Neither model was affected by our category size manipulation. Parameter information can be found in Table 1.

As in Experiment 1, ALCOVE's attention learning parameter allowed it to account for unidimensional generalization. Given a high attention learning parameter, ALCOVE generalized based on whichever dimension is most salient. But due to the lack of a random component, ALCOVE could not account for the use of either single dimension.

As was the case for attention learning in ALCOVE, DIVA's focusing parameter allowed it to account for unidimensional generalization. Replicating our findings from Experiment 1, we found that when DIVA's focusing parameter was large and the dimensions were equally salient, the random initial weights lead to a distribution of generalization profiles based on either or both dimensions. With larger initial weight ranges, DIVA produced more varied patterns of generalization.

The distributions produced by DIVA reflected the salience of the stimulus dimensions. Specifically, DIVA was more likely to generalize using a salient dimension than a less salient dimension. This trend is similar to the effect of salience that we observed behaviorally. Lastly, as in Experiment 1, DIVA tended to produce more multidimensional profiles when the dimensions were equally salient.

\section*{General Discussion}

Our behavioral results revealed that: (1) category knowledge tends to be generalized based on a single dimension; and (2) the salience of a dimension affects the probability that it is selected. We compared DIVA and ALCOVE on their ability to account for these findings. We learned that these similarity-based models are sensitive to salience differences between dimensions and can use attention to generate unidimensional gradients. We also found that a random component can help predict arbitrary dimension selection: DIVA's initial weights randomly offset the models salience appraisal and allowed it to generalize to a single dimension.

These results help to establish generalization as an important basis for formal model evaluation. By requiring that models account for generalization and learning based on the same parameter fits, we systematically widen the scope of what models are held accountable for explaining. In our work, generalization proved not only to be area where DIVA and ALCOVE made different predictions, but it also provided an opportunity to reduce our reliance on post-hoc fits by searching for parameters using aggregate learning data. In future work, we plan to conduct simulations using a stochastic modification of ALCOVE in order to determine how well the model matches our distributions of human generalization performance. We also plan to conduct new simulations based on fitting the models to individual learning curves rather than aggregate data.

\section*{Acknowledgments}

We would like to thank the members of the Learning and Representation in Cognition (LaRC) Laboratory at Binghamton University.

\section*{References}

Erickson, M. A. \& Kruschke, J. K. (1998). Rules and exemplars in category learning. Journal of Experimental Psychology: General, 127, 107-140.

Erickson, M. A. \& Kruschke, J. K. (2002). Rule-based extrapolation in perceptual categorization. Psychonomic Bulletin \& Review, 9(1), 160-168.
Homa, D. (1984). On the nature of categories. In G. H. Bower (Ed.), Psychology of learning and motivation (Vol. 18, pp. 49-94). New York: Academic Press.
Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. Psychological Review, 99, 22-44.
Kurtz, K. J. (2007). The divergent autoencoder (DIVA) model of category learning. Psychonomic Bulletin \& Review, 14, 560 -576.
Kurtz, K. J. (2008). Advances in modeling human category learning with DIVA. Presented at the 2008 Annual Meeting of the Psychonomic Society, Chicago, IL.
Levering, K., \& Kurtz, K. J. (2010). Generalization in higher-order cognition: Categorization and analogy as bridges to stored knowledge. In M. T. Banich \& D. Caccamise (Eds.), Generalization of knowledge: Multidisciplinary perspectives(pp. 175-196). New York, NY: Psychology Press
Medin, D. L., \& Schaffer, M. M. (1978). Context theory of classification learning. Psychological Review, 85, 207238.

Nosofsky, R. M. (1992). Exemplars, prototypes, and similarity rules. In A. F. Healy, \& S. M. Kosslyn (Eds.), Essays in honor of William K. Estes (pp. 149167). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.
Nosofsky, R. M. \& Johansen, M. K. (2000). Exemplarbased accounts of "multiple-system" phenomena in perceptual categorization. Psychonomic Bulletin \& Review, 7, 375-402.
Shepard, R. N. (1957). Stimulus \& response generalization: A stochastic model relating generalization to distance in psychological space. Psychometrika, 22, 325-345.
Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. Science, 237, 1317-1323.

\title{
Enhancing Robust Learning Through Problem Solving in the Genetics Cognitive Tutor
}

\author{
Albert Corbett (corbett@cmu.edu) \\ Ben MacLaren (maclaren@andrew.cmu.edu) \\ Angela Wagner (awagner@cmu.edu) \\ Human-Computer Interaction Institute, Carnegie Mellon University \\ Pittsburgh, PA 15213 USA \\ Linda Kauffman (lk01@andrew.cmu.edu) \\ Aaron Mitchell (apm1@andrew.cmu.edu) \\ Department of Biological Sciences, Carnegie Mellon University Pittsburgh, PA 15213 USA
}

\author{
Ryan S. J. d. Baker (baker2@exchange.tc.columbia.edu) \\ Department of Human Development, Columbia University Teachers College, New York, NY 10027
}

\begin{abstract}
In this paper, we examine the impact of three learning activities designed to foster more robust learning in a Genetics Cognitive Tutor module on pedigree analysis problem solving, in an experimental study. The three activities are (1) interleaved worked examples with student explanations; (2) enhanced feedback with tutor-provided explanations of problem solving steps; and (3) explicit scaffolding of the reasoning steps in this abductive proc-ess-of-elimination reasoning task. The study included four between-subject conditions, a baseline condition in which students exclusively solved standard problems, and three conditions in which students engaged in one of the new learning activities along with standard problem solving. The scaffolded-reasoning condition was most successful in fostering robust learning, as measured by transfer, retention, and preparation for future learning tests. The enhanced feedback condition, in contrast, yielded the poorest performance on the robust learning measures.
\end{abstract}

Keywords: Education; Problem solving; Robust Learning; Intelligent Tutors.

\section*{Introduction}

Problem solving is an essential learning activity across STEM domains. Successful problem solving results in "robust" knowledge: knowledge that is well-grounded in domain knowledge, and as a consequence, is wellretained by students, transfers more readily to related problem situations and prepares students for more successful future learning (Koedinger, Corbett \& Perfetti, 2012). One of the well-documented risks in problem solving, across STEM domains, is that students can develop superficial knowledge that fails these tests of robust learning. In particular, when students are not well-prepared for problem solving, they can develop problem solving knowledge which focuses on surface elements in problem situations, formal representations, and features of the learning environment itself (Chang, Koedinger \& Lovett, 2003; Chi, Feltovich \& Glaser, 1981; Rittle-Johnson \& Siegler, 1998).

In this paper we examine how to structure problem solving in an intelligent tutoring system to support robust learning in the domain of genetics. Because of its foundational place in the biological sciences, genetics is a large and growing component of high school biology courses, but it is also viewed as one of the hardest topics in biology by both students and instructors, at the secondary and at the post-secondary level (Tsui \& Treagust, 2006). Genetics problem solving is characterized by abductive reasoning. In contrast with deductive hypothesis testing, abductive reasoning starts with a set of observations and reasons backwards to infer properties of the genetic processes that produced the data (e.g., whether a trait is dominant or recessive).

In this paper, we study these issues within a tutor lesson for pedigree analysis in the Genetics Cognitive Tutor (Corbett, Kauffman, MacLaren, Wagner \& Jones, 2010), which has been successfully piloted in both high school and college classrooms. Pedigree analysis relies on a complex reasoning process, which nonetheless lends itself to straightforward natural language description. This study examines whether robust learning is supported by a scaffolded reasoning activity prior to conventional problem solving, or by incorporating explicit explanations during problem solving.

\section*{The Domain: Pedigree Analysis}

Basic pedigree analysis problems pose an interesting challenge both for students and for an intelligent tutoring system. Figure 1 displays a typical pedigree analysis problem, in the Genetics Cognitive Tutor (GCT). This pedigree chart displays four generations in a small family. Females are represented as circles and males as squares. In this family, the founding parents have a daughter affected by a rare genetic trait, as represented by the dark circle. No other family members are affected. The student's task is to determine whether this genetic trait is dominant or recessive, and whether it is X-linked, or transmitted on one of the twenty-two autosomal chromosomes in humans.


Figure 1. The GCT Interface for Pedigree Analysis.
This appears to be a reasonably simple task; the features of the problem representation are readily interpretable and there are only two problem-solving actions, but the task involves complex abductive reasoning. It requires students to employ their knowledge of genetic transmission to reason by process of elimination. For example, the student can eliminate the possibility that this is a dominant trait, because the daughter must inherit the trait from one or both parents, and if the trait were dominant the parent(s) who have the trait allele would be affected.

This is also an atypical and challenging task for an intelligent tutoring system to effectively support. The task involves a complex multi-step reasoning process, but since there are just two solution steps, there are no natural opportunities (that is, no behavioral correlates of intermediate reasoning steps) for the tutor to provide assistance in the form of feedback and advice along the way. Instead, the task is subject to shallow learning, since students can readily memorize conclusions, (e.g., when an affected daughter has unaffected parents, the trait must be autosomal recessive), without any understanding of the underlying genetics that supports the conclusion. Finally, the task is subject to gaming the system (Baker, Corbett, Koedinger \& Wagner, 2004), since each solution step consists of a menu with only three alternatives - "dominant," "recessive," or "cannot be determined," in one case, and "autosomal," "Xlinked," or "cannot be determined in the other case." Unmotivated students can readily click through the menu options to find correct answers.

Summative evaluations of Genetics Cognitive Tutor modules are consistent with these risks; pretestposttest learning gains for basic pedigree analysis are only about half as large as the average gain across all topics (Corbett, et al, 2010).

\section*{Pedigree Analysis Learning Activities}

In this study we developed and evaluated three Cognitive Tutor activities intended to support robust learning in pedigree analysis problem solving. Two activities integrate explicit reasoning explanations into the conventional problem-solving task - worked examples, in which the student explains tutor-generated problem solutions, and enhanced feedback, in which the tutor provides explanations for student problem-solving steps. The third activity, in contrast, explicitly scaffolds the intermediate steps in this abductive process-of-
elimination task and is designed to precede conventional problem solving. As in Cognitive Tutors more generally (Anderson, Corbett, Koedinger \& Pelletier, 1995), in these activities, students receive immediate accuracy feedback on each problem-solving step and can request hints on any problem-solving step.

Interleaved Worked Examples It is well-documented that integrating worked examples with problem solving serves to decrease total learning time and yields improved learning outcomes (Pashler, Bain, Bottge, Graesser, Koedinger, McDaniel \& Metcalfe, 2007; Renkl \& Atkinson, 2003; Sweller \& Cooper, 1985). Recently, several studies have examined the benefits of incorporating worked examples into intelligent tutoring systems (ITSs) for problem solving across a variety of math and science domains, including topics in algebra, geometry, statistics, biology, chemistry and physics (Anthony, 2008; Conati \& VanLehn, 2000; Corbett, MacLaren, Wagner, Kauffman, Mitchell, Baker \& Gowda, 2011; Mclaren, Lim \& Koedinger, 2008; Reed, Corbett, Hoffman, Wagner \& MacLaren, 2013; Salden, Aleven, Schwonke \& Renkl, 2010; Schwonke, Renkl, Krieg, Wittwer, Aleven \& Salden, 2009; Weitz, Salden, Kim \& Heffernan, 2010). In these ITS studies, the chief benefit of incorporating worked examples has been to reduce learning time for a fixed set of activities compared to problem solving alone, but unlike the classic worked-example literature, these ITS studies generally do not find that incorporating worked examples leads to more accurate posttest performance than problem solving alone. The exception is Salden, et. al (2010), which found that adaptively fading examples led to some relative improvement on posttest problem solving. Similarly, the evidence that students learn more deeply when worked examples are integrated into ITSs is mixed at best, although Anthony (2008) and Salden, et al (2010) report better retention of problem solving knowledge and Schwonke, et al (2009) found some evidence of greater conceptual transfer in one of two studies.

Pedigree analysis is a promising domain in which to further explore worked examples, since each step in problem solving depends on a complex, but readily describable reasoning process. Figure 2 displays the worked example interface. Each worked example displays a standard pedigree analysis problem and displays the correct dominance and linkage of the trait directly below the pedigree. These examples also identify a key nuclear family in the pedigree and describe the pattern of affected and unaffected individuals in the family that allows the student to identify the dominance and linkage of the trait. Students select entries in the three menus at the bottom of the screen to explain how to determine the dominance and linkage from the pattern, based on their knowledge of genetics transmission.

Feature Focusing We developed a contrasting activity in which the student generates problem solutions and the tutor provides explanations of the student's correct actions, to directly address two characteristics of basic


Figure. 2. The GCT Interface at the conclusion of a pedigree analysis worked example.
pedigree analysis. The first is that the tell-tale patterns can be hard to identify. For instance, Figure 3 displays two pedigrees, which look similar, each with four affected males, but the trait on the left is autosomal dominant, while the trait on the right is X -linked recessive.


Figure 3. An autosomal dominant pedigree (left) and an X -linked recessive pedigree (right).

The second challenge is that the immediate accuracy feedback generally delivered by Cognitive Tutors (cf. Anderson et al., 1995), is not all that informative in this lesson, since there is a reasonably high probability that the student performed the right action for the wrong reason. Debriefing sessions revealed that students in high school classrooms are aware of the latter risk, and sometimes would like to receive an explanation after selecting a correct menu entry, rather than a hint before.

To address these problems, we developed an enhanced feedback interface displayed in Figure 4. The pedigree is initially displayed entirely in black and without any explanatory text. Following each of the two problem-solving steps, the tutor highlights the relevant pattern in the figure, and provides an explanation. In this example, after the student concluded that the trait is recessive, (1) the relevant pattern was highlighted in green, (2) the conclusion was summarized at the top of the screen in green, and (3) an explanation of the conclusion was displayed in green in the window to the right. After the student concluded that the trait is Xlinked, the relevant pattern remained highlighted in green, the prior dominance conclusion and explanation were grayed out, the linkage conclusion was summarized in green near the figure, and the linkage explanation was presented in green in the adjoining window.


Figure 4. The GCT interface at the end of a pedigree analysis problem with enhanced feedback.

Abductive Reasoning Scaffolds Finally, we developed a problem-solving activity that directly engages students in the reasoning-by-process-of-elimination task. While the other two interventions were integrated with conventional problem solving, this is a separate task that was designed to precede conventional problem solving. Each problem in this task presents the phenotypes of three family members, two parents and a child, as displayed in Figure 5. Immediately to the right, the four possible modes of transmission are listed (autosomal dominant, X-linked dominant, autosomal recessive and X-linked recessive). For each of the four modes, the student enters what the underlying genotype of each of the three family members would have to be, given their respective phenotypes, and under the mode of transmission. (For example, if the trait were autosomal dominant, the two unaffected parents would have to be homozygous recessive, while the affected daughter would have to have a dominant allele.) Then to the far right, the student indicates whether the observed pattern of phenotypes is possible under each mode of transmission, that is, whether the child could inherit its genotype from its parents. (The observed pattern in Figure 5 is impossible for an autosomal dominant trait, since neither parent has a dominant allele to transmit to the daughter.) Finally, at the bottom of the screen the student summarizes which modes of transmission are possible for the observed phenotype pattern.


Figure 5. The GCT Pedigree Analysis Scaffolded Reasoning task at the end of a problem.

This study includes four between-subject conditions and evaluates the success of each of these three interventions in supporting the acquisition of problem-solving skills, and robust learning, compared to standard problem solving.

\section*{Method}

\section*{Participants}

Sixty-four high school students enrolled in high school biology courses were recruited through newspaper ads and classroom handouts to participate in this study for pay. Students were randomly assigned to one of four between-subject treatment groups.

\section*{Procedure}

Students participated in two 2.5 -hour sessions on consecutive days in a CMU computer lab. In Session 1, students:
- viewed an instructional video and read instructional text on basic pedigree analysis;
- completed a conceptual knowledge pretest and a basic problem-solving pretest;
- completed basic pedigree analysis Cognitive Tutor activities, which differed by condition;
- completed a basic problem-solving test and a transfer problem-solving test.

The second session was devoted to an extended preparation for future learning (PFL) activity, as well as a delayed basic problem-solving test. The PFL task was an advanced carrier-probability pedigree analysis task. Each problem in the task displays a large pedigree chart with five or six generations and students calculate the probabilities that various unaffected individuals in the chart carry a single recessive trait allele. Students:
- read instructional text on carrier probabilities pedigree analysis;
- completed an initial PFL paper-and-pencil test
- completed PFL Cognitive Tutor problems;
- completed a second PFL paper-and-pencil posttest;
- completed a delayed basic problem-solving test

\section*{Design}

There were four between-subject conditions in the study, defined by students’ Cognitive Tutor learning activities in the first session.
- Basic Problem Solving (PS): Students completed a set of 78 basic pedigree analysis problems.
- Enhanced Feedback (EF): Students completed the same 78 problems as in the PS group, but completed the first 20 with enhanced feedback.
- Interleaved Worked Examples (WE): Students completed a problem set with 14 interleaved worked
examples and problems to solve, followed by 18 standard problems.
- Scaffolded Abductive Reasoning (SR): Students completed six problems in which the abductive reasoning process was explicitly scaffolded as described above, followed by a set of 18 standard problems.

In Session 2, all students completed the same set of activities focused primarily on the PFL task.

\section*{Tests}

We developed four types of paper-and-pencil tests for the study:
- Problem Solving Tests: Three forms were developed. Each form served as the pretest for \(1 / 3\) of the students in each condition, the session-1 posttest for \(1 / 3\) of the students, and the session-2 delayed test for \(1 / 3\) of the students.
- Conceptual Knowledge Tests: A conceptual knowledge pretest was developed to evaluate students' knowledge of genetic transmission.
- Transfer Tests: A transfer test was developed with two types of problems: one type asked students to solve basic pedigree analysis problems with novel patterns requiring novel reasoning; a second asked students to identify whether family pedigrees were possible or impossible under the four modes of transmission.
- Preparation for Future Learning (PFL): Two forms of a PFL problem-solving test were developed. Each form served as the initial test for \(1 / 2\) of the students in each condition, and as the second test for \(1 / 2\) of the students.

\section*{Results}

Table 1 displays mean accuracy (percent correct) for the tests administered in the study. The conceptual knowledge (CK) and problem solving (PS1) pretests are displayed to the left, followed by the problem solving posttest (PS2) and the problem-solving learning gain from pretest to posttest (PS2-PS1).

The four robust learning tests follow, including the transfer test (TR); the initial PFL test (PFL1), which preceded the session-2 PFL tutor problems; the second PFL test (PFL2), which followed the GCT PFL problems; and finally the delayed basic problem-solving test (PS3). The final column displays students' change in basic problem-solving accuracy over the retention interval (PS3 - PS2).

Table 1: Student test accuracy (percent correct).
\begin{tabular}{cccccccccc}
\hline & & & \multicolumn{4}{c}{ PS } \\
Cond & CK & PS1 & PS2 & Gain & TR & PFL1 & PFL2 & PS3 & \begin{tabular}{c} 
retention \\
change
\end{tabular} \\
& \(\% \mathrm{C}\) & \(\% \mathrm{C}\) & \(\% \mathrm{C}\) & PS2-PS1 & \(\% \mathrm{C}\) & \(\% \mathrm{C}\) & \(\% \mathrm{C}\) & \(\% \mathrm{C}\) & PS3-PS2 \\
\hline SR & 92 & 47 & 47 & 0 & 54 & 36 & 60 & 53 & 6 \\
EF & 86 & 41 & 53 & 12 & 44 & 19 & 40 & 49 & -4 \\
PS & 91 & 43 & 48 & 5 & 47 & 34 & 52 & 48 & 0 \\
WE & 92 & 49 & 56 & 7 & 46 & 31 & 61 & 50 & -6 \\
\hline
\end{tabular}

Average scores on the Conceptual Knowledge pretest (CK) were quite high, averaging about \(90 \%\) correct, indicating that students were very familiar with the transmission genetics underlying pedigree analysis. An ANOVA revealed no significant difference among the four conditions on this pretest, \(\mathrm{F}(3,60)=1.33\), ns.

Average scores on the Problem Solving pretest (PS1) were much lower, averaging \(45 \%\) correct. Again, an ANOVA revealed no significant difference among the four conditions on this pretest, \(\mathrm{F}(3,60)=1.31\), ns.

\section*{Pedigree Analysis Posttest Performance}

We performed an ANOVA on the five paper-andpencil posttest measures of student learning, including the Problem Solving posttest (PS2), and the four robust learning measures: the Transfer test (TR), the Preparation for Future Learning tests (PFL1 \& PFL2) and delayed Problem Solving test (PS3). The main effect of condition is not significant, \(\mathrm{F}(3,60)=1.26\), ns, but the interaction of condition and test type is significant \(F(12,240)=2.25, p<.01\). (The main effect of test type is also significant, \(\mathrm{F}(4,240)=28.59, \mathrm{p}<.001\), but not of particular interest.)

As can be seen in the table, the new enhanced feedback (EF) and scaffolded reasoning (SR) had contrasting impacts. The EF activities yielded the largest prob-lem-solving learning gain, but generally led to the lowest scores on the robust learning tests. In contrast, the new SR activities led to no discernible learning gains from PS1 to PS2, but generally led to the best performance on the robust learning tests.

Basic Problem Solving The new EF condition led to the largest problem-solving learning gains, while the new SR condition led to no discernible learning gains. However, in an ANOVA on the PS gain displayed in Table 1, the effect of condition was not significant, \(\mathrm{F}(3,60)=1.74, \mathrm{p}<.17\).

Robust Learning The SR condition generally outperformed the familiar PS and WE conditions, which in turn outperformed the EF condition on the robust learning measures. The difference is fairly pronounced on the transfer task, and in the retention change scores, where the SR condition is the only condition that displays a small increase in scores over the retention interval. We performed an ANOVA on the transfer test, two PFL tests and the retention change scores, and the effect of condition is significant, \(\mathrm{F}(3,60)=2.80, \mathrm{p}<.05\). The interaction of condition and test measure is not significant.

We performed an ANOVA on each of these four robust learning measures separately and condition was significant only for the retention change measure, \(\mathrm{F}(3,60)=3.41 . \mathrm{p}<.05\), where the SR group is the only one which shows any sign of improving on basic problem solving by virtue of completing the intervening Cognitive Tutor PFL task.

\section*{Tutor Performance}

Session 1 Total Time Table 2 displays the total time that the students in the four conditions spent on session1 GCT pedigree analysis learning activities. The ses-sion-1 tutor activities were designed to hold learning time constant. As can be seen, average time was reasonably constant across conditions, ranging from about 24 to about 27 minutes. We performed an ANOVA on session 1 time on task, and condition was not reliable, \(F(3,60)=0.74\), ns.

Table 2: Student performance on GCT activities in Session 1 and Session 2
\begin{tabular}{c|c|ccc}
\hline & Session 1 & \multicolumn{3}{|c}{ Session 2 (PFL GCT) } \\
Cond & Min. & Min. & \%C & \%hints \\
\hline SR & 27.4 & 25.4 & 58 & 19 \\
EF & 24.9 & 31.2 & 41 & 41 \\
PS & 26.0 & 26.6 & 49 & 32 \\
WE & 23.7 & 26.1 & 55 & 26 \\
\hline
\end{tabular}

Session 2 PFL Tutor Problems Student performance in the session-2 carrier probabilities GCT task provides an additional PFL measure with respect to the four ses-sion-1 learning activities. All students completed the same set of 14 carrier probability problems in the second session. Table 2 displays the average time to complete the problems, student accuracy (the percentage of problem-solving steps on which students' first action was correct), and help requests (the percentage of steps on which a student requested a tutor hint). The students in the SR condition were the most successful in session 2 , responding most accurately, while requiring the least time, and least assistance. In contrast, students in the EF condition performed least successfully on all three measures. In three ANOVAs, the main effect of condition is significant for accuracy, \(\mathrm{F}(3,60)=2.77, \mathrm{p}<.05\), and for hint requests, \(\mathrm{F}(3,60)=3.55, \mathrm{p}<.05\), but not significant for total time, \(F(3,60)=1.86\), ns.

\section*{Summary and Discussion}

Among the three new GCT tasks, the scaffolded reasoning task was the most successful in preparing students for more robust learning in problem solving. The SR combination of a scaffolded reasoning task, in conjunction with a single set of conventional problems, yielded the most robust understanding of pedigree analysis, as measured by transfer, preparation for future learning, and retention of problem-solving skill.

However, design work remains to be done, since the scaffolded reasoning task did not prepare students well for conventional problem solving. Despite their robust learning, students in this condition performed surprisingly poorly on the problem solving posttest, displaying no learning gains.

A more promising design may be to insert the worked example task between the scaffolded reasoning task and conventional unassisted problem solving, to provide students the opportunity to reflect on, and describe how to apply their abductive reasoning skills in
the full problem-solving task. While students in the WE condition did not perform discriminably better than students in the baseline PS condition across the board, there was at least a trend for the WE students to outperform the PS students on the PFL measures.

Finally, the newly designed enhanced-feedback problem solving condition was disappointing. There was a modest and non-significant trend for the EF condition to yield larger learning gains on the problem solving test, but the EF condition led to generally poorer performance on measures of robust learning, especially the PFL test and tutor activities. This may indicate that, to the extent there is a benefit of the enhanced feedback, students are learning to identify the key patterns in pedigrees and to associate them with the corresponding conclusions, but are not developing an understanding of the underlying reasoning. Again, inserting the interleaved worked example activity between the EF task and conventional problem solving might help students build more effectively on any benefits of the EF condition.

\section*{Acknowledgments}

This research was supported by the Institute of Education Sciences award R305A090549 "Promoting Robust Understanding of Genetics with a Cognitive Tutor that Integrates Conceptual Learning with Problem Solving."

\section*{References}

Anderson, J.R., Corbett, A.T., Koedinger, K.R. \& Pelletier, R. (1995). Cognitive tutors: Lessons learned. The Journal of the Learning Sciences, 4, 167-207.
Anthony, L. (2008). Developing handwriting-based intelligent tutors to enhance mathematics learning. Unpub. doctoral dissertation, Carnegie Mellon University.
Baker, R.S., Corbett, A.T., Koedinger, K.R. \& Wagner, A.Z. (2004). Off-task behavior in the cognitive tutor classroom: When Students "Game the System". Proceedings of ACM CHI'2004 Conference on Human Factors in Computing Systems, 383-390.
Chang, N. M., Koedinger, K. R., \& Lovett, M. C. (2003). Learning spurious correlations instead of deeper relations. Proceedings of the 25th Cognitive Science Society, 228-233.
Chi, M.T.H, Feltovich, P.J. \& Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. Cognitive Science, 18, 439-477.
Conati, C. \& VanLehn, K. (2000). Further results from the evaluation of an intelligent tutor to coach selfexplanation. In the Proceedings of the \(5^{\text {th }}\) Internationational Conference on Intelligen Tutoring Systems, 304-313.
Corbett, A., Kauffman, L., MacLaren, B., Wagner, A., \& Jones, E. (2010). A Cognitive Tutor for genetics problem solving: Learning gains and student modeling. Journal of Educational Computing Research, 42, 219-239.
Corbett, A., MacLaren, B., Wagner, A., Kauffman, L., Mitchell, A., Baker, R. \& Gowda, S. (2011). Prepar-
ing students for effective explaining of worked examples in the Genetics Cognitive Tutor. In Proceedings of the Thirty-third Annual Meeting of the Cognitive Science Society, 1476-1481.
Koedinger, K.R., Corbett, A.T. \& Perfetti, C. (2012). The Knowledge-Learning-Instruction (KLI) Framework: Bridging the Science-Practice Chasm to Enhance Robust Student Learning. Cognitive Science, 36, 757-798.
McLaren, B. M., Lim, S.-J., \& Koedinger, K. R. (2008). When is assistance helpful to learning? Results in combining worked examples and intelligent tutoring. In Proceedings of the 9th International Conference on Intelligent Tutoring Systems (pp. 677-680).
Pashler, H., Bain, P., Bottge, B., Graesser, A., Koedinger, K., McDaniel, M., \& Metcalfe, J. (2007). Organizing Instruction and Study to Improve Student Learning. Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.
Reed, S.K., Corbett, A., Hoffman, B., Wagner, A., \& MacLaren, B. (2013). Effect of worked examples and Cognitive Tutor training on constructing equations. Instructional Science, 41, 1-24.
Renkl, A. \& Atkinson, R. K. (2003). Structuring the transition from example study to problem solving in cognitive skills acquisition: A cognitive load perspective. Educational Psychologist, 38, 15-22.
Rittle-Johnson, B. \& Siegler, R.S. (1998). The relation between conceptual and procedural knowledge in learning mathematics: A review. In C. Donlan (Ed.), The development of mathematical skill (pp. 75-110). Hove, UK: Psychology Press.
Salden, R. J. C. M., Aleven, V., Schwonke, R., \& Renkl, A. (2010). The expertise reversal effect and worked examples in tutored problem solving. Instructional Science, 38, 289-307.
Schwonke, R., Renkl, A., Krieg, C., Wittwer, J., Aleven, V., \& Salden, R. (2009). The workedexample effect: Not an artifact of lousy control conditions. Computers in Human Behavior, 25, 258-266.
Sweller, J. \& Cooper, G.A. (1985). The use of worked examples as a substitute for problem solving in learning algebra., Cognition and Instruction, 2, 59-89
Tsui, C-Y. \& Treagust, D.F. (2006). Understanding genetics: Analysis of secondary students' conceptual status. Journal of Research in Science Teaching, 44, 205-235.
Weitz, R., Salden, R. J. C. M., Kim, R. S., \& Heffernan, N. T. (2010). Comparing worked examples and tutored problem solving: Pure vs. mixed approaches. In Proceedings of the Thirty-Second Annual Meeting of the Cognitive Science Society (pp. 2876-2881).


\title{
Dimensions of specificity in musical memory: Evidence from metrical restoration
}

\author{
Sarah C. Creel (creel@cogsci.ucsd.edu) \\ University of California, San Diego, Department of Cognitive Science \\ La Jolla, CA 92093-0515
}

\begin{abstract}
How is musical memory organized? While classic studies of music perception appealed to schematic or symbolic knowledge structures, recent work suggests that listeners form highly-detailed auditory representations of music. Studies of metrical restoration-memory fill-in of the "beat" of a metrically-ambiguous melody-suggest some organizing dimensions in musical memory. However, many potential dimensions remain unexplored. The current study looked for effects of mode (major vs. minor) - a substantial organizing force in Western music-and timbre (what instrument is playing) on metrical restoration. Both mode and timbre can signify particular musical styles. In Experiment 1, listeners showed timbre specificity in metrical restoration, but not mode specificity. However, in Experiment 2, when timbres were extremely unique (one per melody), restoration effects were not observed, suggesting that too much variability leads to diffuse representations which are too weak to support metrical restoration. Implications for the nature of musical memory are discussed.
\end{abstract}

Keywords: perceptual restoration, meter, music perception, metrical restoration

\section*{Introduction}

Recent research suggests that listeners form rich, detailed representations of perceptual information. These details later facilitate recognition (Creel, Aslin, \& Tanenhaus, 2008; Goldinger, 1998; Gjerdingen \& Perrott, 2008; Krumhansl, 2010; Schellenberg et al., 1999) and allow fill-in of ambiguous or absent information (Creel, 2011, 2012; Samuel, 1981). In music particularly, Creel (2011, 2012) has shown that listeners who hear particular metrical (or harmonic) information with a melody will later, upon hearing the melody alone, fill in the missing contextual information (harmony or meter) previously heard with that melody.

These findings are interesting in a number of respects. First, meter is a property previously thought to be largely signal-driven, with listeners extracting metrical regularities via statistical analysis of the signal itself. Creel's work suggests that memory influences meter perception. Second, these findings suggest that similarity-based organization of detailed auditory-temporal memories can support knowledge of distinct genres, such as different musical styles or different languages.

Yet many questions remain. What factors allow listeners to keep particular musical patterns distinct in memorywhat keeps them from bleeding together? Inversely, what factors allow listeners to generalize metrical information?

On first glance, a simple answer to both questions is degree of similarity: listeners generalize to similar musical patterns, and maintain specific representations of less-similar patterns. However, determining equivalent degrees of similarity on varied dimensions is not trivial, as perceivers' use of dimensions can change depending on task and attentional factors (Nosofsky, 1986). That is, we do not know what weights listeners assign to different dimensions in musical memory. Further, some dimensions may be processed integrally, such that their combined effect is not a simple sum of their individual effects. The current study aims to explore the relative strength of various auditorymusical properties on metrical restoration, providing insights into similarity-based organization in musical memory.


Figure 1. First four measures of a melody, in (a) major key with \(6 / 8\) metrical context; (b) minor key with \(3 / 4\) metrical context. Metrical grids indicate perceived emphasis in each meter: large X's denote strong beats, small x's weaker beats,
and .'s indicate the subdivision of each beat. Beat subdivisions are identical in duration in both versions.

\section*{Known influences on metrical restoration}

Previous work in my lab (Creel, 2011, 2012) has examined some factors in memory restoration of meter. In those experiments, as well as the new experiments described here,

I exploit the \(3 / 4-6 / 8\) ambiguity, a musical "ambiguous figure." Certain musical passages with repeating series of 6 sub-beats can be interpreted as being in \(3 / 4\) meter (beats alternating evenly with sub-beats, X . x . x .) or in \(6 / 8\) meter (each beat is followed by two sub-beats: X . . x . . ). Figure 1 shows examples of each meter.

In my experiments, each listener heard a set of 8-12 melodies. Half were presented in a musical context suggesting \(3 / 4\), and half in a context suggesting \(6 / 8\). Melodies were constructed to fit either metrical pattern, allowing a carefully counterbalanced design where, across listeners, each melody was heard in each meter equally often (Table 1). A listener heard each melody multiple times during an exposure phase. Next, all listeners heard each melody without its meter-implying context, followed by probe drumbeats in either \(3 / 4\) or \(6 / 8\). They were asked to rate how well the drumbeats fit with the preceding melody. The question was whether listeners would provide higher ratings for the drumbeats (meters) that matched the contexts that they had previously heard.

Table 1: Example conditions in a metrical restoration experiment. If listeners restore melodies' metrical contexts, then Listener 1 should provide higher probe ratings to the shaded probe trials, and Listener 2 should provide higher probe ratings to the unshaded probe trials.
\begin{tabular}{cccc}
\hline & \multicolumn{2}{c}{ Exposure phase } & Test phase \\
\hline \(\begin{array}{c}\text { Mel- } \\
\text { ody }\end{array}\) & Listener \(\mathbf{1}\) & Listener 2 & All listeners \\
\hline 1 & \(3 / 4\) context & \(6 / 8\) context & \(\begin{array}{c}\text { melody alone melody alone } \\
+3 / 4 \\
\text { melobe }\end{array}\) \\
2 & \(3 / 4\) context & \(6 / 8\) context \(\begin{array}{c}\text { malone } \\
\text { melody alone } \\
+3 / 4 \text { probe }\end{array}\) & \(+6 / 8\) probe
\end{tabular}\(\}\)

In Creel (2011, Experiment 2), listeners heard a set of 8 melodies that were very distinct from each other in terms of timbre, note rate (speed), mode (major, minor, other), and rhythmic patterns. Listeners provided higher ratings for drumbeat probes that matched the contexts they had heard with those specific melodies during the exposure phase. This suggests melody-specific memory for meter.

A second study (Creel, 2012) examined the role of crossmelody similarity on metrical restoration, and so presented melodies with a stronger similarity structure: two timbres, a single note rate, and similar rhythmic patterns across melodies. Listeners associated metrical information with timbre: they showed metrical restoration for a melody played in its original timbre, but not when it was played in the other timbre (as long as timbre and meter patterned consistently across melodies). Further, when melodies were constructed from two different sets of motifs (defined in that study as brief rhythm+contour patterns), listeners showed
metrical restoration for new melodies with those motifs (Experiment 5).

Interestingly, the magnitude of the metrical restoration effects in the 2011 study was much larger than the effect in Creel (2012). Though there were a number of differences between the two sets of studies, one possibility is that the denser similarity structure of the melodies in the second paper led to greater generalization, but, conversely, less individuation. However, it is not clear which one (or more) of the unique properties of the melodies in the 2011 paper generated such strong restoration effects: rate, timbre, mode, rhythmic patterns. Do all dimensions of variation contribute additively to specificity/individuation in memory, or is one particular factor the "smoking gun"?

\section*{Unknown effects on metrical restoration}

As seen in Creel (2011, 2012), timbre and motif content seem to be integral to musical memory. That is, metrical restoration shows timbre specificity and motif specificity. However, numerous dimensions of substantial musictheoretical importance remain untested. First, is there mode specificity? Mode, the particular pitch collection used in a musical piece, may be a signature of musical style: in Western music, the most common modes are major and minor. Other musical styles and cultures are characterized by yet other pitch collections (e.g. Castellano et al., 1984). Mode also contributes to emotional processing: Western listeners associate the major mode with happiness, and the minor mode with sadness (e.g. Hunter et al., 2008).

A second factor not previously examined is rate-the speed at which a musical piece is executed. Do listeners store melodies rate-specifically?

Finally, the role of timbre bears further exploration. Is timbre simply one of many cues that differentiate music in memory? If hearing melodies in two timbres allows listeners to keep metrical patterns distinct, then does hearing melodies in even more timbres create even more distinct metrical representations?

\section*{The current study}

The current study examined influences of mode, rate, and timbre on the metrical restoration effect. Experiment 1 asked whether differences in mode (major or minor), alone or in combination with timbre cues, show specificity effects in metrical restoration. Experiment 2 asked whether maximal differences in timbre ( 1 vs. 12 timbres), and differences in rate, allow even more specificity in memory.

\section*{Experiment 1}

The first experiment compares metrical restoration as a function of timbre-specificity (shown in Creel, 2012) and mode-specificity (not yet explored). We know that listeners show timbre-specific metrical restoration when timbre patterns consistently with meter. Does mode serve a similar function? That is, if mode patterns consistently with meter (e.g. major melodies are always heard in \(3 / 4\), minor melodies in \(6 / 8\) ), will listeners only restore the meter when a
melody is heard in its original mode? Further, do timbre and mode combine additively to provide even more distinct musical memories, and even stronger restoration effects?

\section*{Method}

Participants \(N=107\) participants from the UCSD human participant pool received course credit for participation. Roughly equal numbers of participants took part in Experiments 1a \((n=36), 1 \mathrm{~b}(n=35)\), and \(1 \mathrm{c}(n=36)\).

Stimuli The 18 melodies used here were originally used in Experiments 1-3 of Creel (2012). Melodies were edited slightly to generate clearer metrical contexts. The originals were all composed in major mode. Minor-mode versions were created by lowering the pitch of scale degree 3 , and 6 and 7 in certain contexts, by \(1 / 2\) step, or about \(6 \%\). Melodies were exported from Finale software (MakeMusic, Inc.) in the key of C, played both in a vibraphone timbre and a muted-trumpet timbre. These two timbres were chosen to be highly distinct, based on Iverson and Krumhansl's (1993) perceptual scaling study of timbres.

Design Each participant heard only 12 of the 18 melodies during exposure. Test trials presented all 18 melodies, with manipulations as described below (examples in Table 2). All melodies were presented during the test phase followed by metrical probe drumbeats in \(6 / 8\) and \(3 / 4\).

Experiment la: different timbres. There was a consistent mapping of timbre to meter. For example, a participant might hear six major-mode vibraphone melodies in \(3 / 4\), and six major-mode muted-trumpet melodies in \(6 / 8\). A given participant heard only one mode (major or minor). Test trials presented each melody four times: two probe meters (original meter, other meter) x two timbres (original timbre, other timbre). Mode did not change from training to test.

Experiment 1b: different modes. There was a consistent mapping of mode to meter. For example, a participant might hear six major-mode vibraphone melodies in a \(3 / 4\) metrical context, six minor-mode vibraphone melodies in a \(6 / 8\) metrical context. Thus, the roles of mode and timbre were reversed relative to Experiment 1a. Test trials presented each melody four times: two probe meters (original meter or other meter) x two modes (original mode or other mode). Timbre did not change from training to test.

Experiment 1c: different mode + timbre combinations. There was a consistent mapping of timbre and mode to meter. For example, a participant might hear six majormode vibraphone melodies in \(3 / 4\), and six minor-mode muted-trumpet melodies in \(6 / 8\). Thus, \(3 / 4\) melodies all had the same mode and timbre, while \(6 / 8\) melodies had the other mode and timbre, giving listeners two attributes to link to metrical information. Test trials presented each melody four times: twice for each probe meter, with either the original mode+timbre combination or the opposite mode+timbre combination.

Table 2: Example exposure conditions in Experiment 1.
\begin{tabular}{cccc}
\hline Melody Exp. 1a listener & Exp. 1b listener & Exp.1c listener \\
\hline 1 & major, vib., \(3 / 4\) & major, vib., \(3 / 4\) & major, vib., \(3 / 4\) \\
2 & major, vib., \(3 / 4\) & major, vib., \(3 / 4\) & major, vib., \(3 / 4\) \\
3 & major, tpt., \(6 / 8\) & minor, vib., \(6 / 8\) & minor, tpt., \(6 / 8\) \\
4 & major, tpt., \(6 / 8\) & minor, vib., \(6 / 8\) & minor, tpt., \(6 / 8\) \\
\hline
\end{tabular}

Note. Vib. \(=\) vibraphone; tpt. \(=\) muted trumpet.
Procedure The experiment was run in Matlab using Psychtoolbox3 (Brainard, 1997; Pelli, 1997). Sounds were presented via Sennheiser HD 280 headphones. Before the experiment proper, listeners completed a questionnaire on their academic and performing music experiences. They then went on to an exposure phase, followed by a test phase.

The exposure phase presented each of 12 melodies 6 times each ( 72 trials total). On each trial, listeners were asked to rate, by clicking in a 2 -dimensional grid, the melody's affective quality (sad to happy, on the x-axis) and their subjective judgment of it (like to dislike, on the \(y\) axis). This cover task aimed to keep participants attentive without alerting them to attend specifically to the meter. They were not told that they would later be tested on their knowledge of the melodies.

After exposure, participants were asked to rate drumbeats following each melody. Before beginning the test, they were presented with four example drumbeat probe trials, in order: four bars of Happy Birthday ( \(3 / 4\) meter) followed by "good" drumbeats (in 3/4); Happy Birthday followed by "bad" drumbeats (in 6/8); four bars of Greensleeves ( \(6 / 8\) meter) followed by "good" drumbeats (in 6/8); Greensleeves followed by "bad" drumbeats (in 3/4). They were prompted to consult the experimenter if they had any questions. After this, they proceeded to the test phase.

The test phase presented all 12 melodies that the participant had heard during learning, plus the 6 held-out melodies. Each melody was presented four times: once in the original mode and timbre followed by \(3 / 4\) probe drumbeats ( 4 measures plus a downbeat, or 13 beats); once in original mode and timbre with \(6 / 8\) drumbeats ( 4 measures plus a downbeat, or 9 beats); once in the other mode and/or timbre with \(3 / 4\) drumbeats; and once in the other mode and/or timbre with \(6 / 8\) drumbeats. For each participant, the mode/timbre and meter either matched or mismatched the contexts they had heard at training.

\section*{Results}

All ratings were converted from raw pixel values to a scale ranging from -1 to +1 to allow easier interpretation.

Cover task Participants rated liking and affective content during exposure. Participants rated major melodies happier than minor melodies (1a: between-participants: \(t(33)=4.79\), \(\mathrm{p}<.0001 ; 1 \mathrm{~b}: \mathrm{t}(35)=15.07, \quad \mathrm{p}<.0001 ; 1 \mathrm{c}: \quad \mathrm{t}(35)=11.30\), \(\mathrm{p}<.0001\) ). These ratings differences suggest that participants were attentive during exposure, and further, that they readily distinguished major and minor modes from each other.

\section*{Probe ratings}

Experiment 1a. To determine whether probe ratings differed as a function of prior exposure and instrument match, an analysis of variance (ANOVA) on probe ratings was conducted with Exposure Meter ( \(3 / 4\) or \(6 / 8\) ), Probe Meter ( \(3 / 4\) or \(6 / 8\) ), and Timbre (original, switched) as within-participants factors. Bear in mind that, if there is a metrical restoration effect, then the interaction of Exposure Meter x Probe Meter should be significant. If restoration was timbre-specific-that is, if restoration was stronger when the melody was presented in the original timbre-then there should be a three-way interaction. For ease of interpretation, the restoration effect (Figures 2 and 3) is plotted in this paper as the average difference between the exposed probes (Exposure \(=3 / 4\), Probe \(=3 / 4 ; 6 / 8,6 / 8\) ) and the unexposed probes (Exposure=3/4, Probe=6/8; 6/8, 3/4).

An Exposure Meter x Probe Meter interaction \((F(1,34)=\) 5.30, \(p=.03\) ) verified an overall metrical restoration effect. An interaction of Exposure Meter x Probe Meter x Timbre \((F(1,34)=4.57, p=.04)\) suggested differences in metrical restoration as a function of timbre match. Considering each timbre individually, the Exposure Meter x Probe Meter interaction was only significant for original-timbre trials \((F(1,34)=9.29, p=.004)\), but not for switched-timbre trials \((F(1,34)=.06, p=.80)\). This replicates previous work (Creel, 2012) suggesting that, when meter and timbre covary, listeners do not generalize metrical restoration across a timbre change. Data from "new" melodies (heard for the first time at test) are not discussed due to space restrictions.

Experiment lb. An ANOVA was conducted on probe ratings with Exposure Meter, Probe Meter, and Mode (original or switched) as factors. The ANOVA showed an Exposure Meter x Probe Meter interaction \((F(1,35)=6.84\), \(\mathrm{p}=.01\) ), consistent with metrical restoration. If restoration was mode-specific, there should be a significant Exposure Meter x Probe Meter x Mode interaction. However, this interaction did not approach significance \((F(1,35)=0.00\), \(p=.98\) ), implying that there was no decrement in metrical restoration when a melody was presented in the opposite mode as in training. Bearing this out, the Exposure Meter x Probe Meter interaction was significant for original-mode \((F(1,35)=4.76, p=.04)\) and switched-mode \((F(1,35)=4.66\), \(p=.04\) ) trials individually. This suggests that, in contrast to timbre specificity, listeners do not show mode specificity, but rather mode generality, in meter perception.

Experiment 1c. One might wonder if mode, while not showing specificity effects alone, might augment a timbrespecificity effect. An ANOVA with Exposure Meter, Probe Meter, and Mode+Timbre Combination (original, switched) was conducted on probe ratings. An Exposure Meter x Probe Meter interaction \((F(1,35)=4.44, p=.04)\) suggested an overall metrical restoration effect. The Exposure Meter x Probe Meter x Mode+Timbre Combination interaction did not approach significance \((F(1,35)=0.68, p=.42)\). However, considering each mode+timbre combination individually suggested that the metrical restoration effect was carried by
original mode+timbre trials (significant Exposure Meter x Probe Meter interaction, \(F(1,35)=4.65, p=.04)\), rather than switched mode+timbre trials (not significant; \(F(1,35)=0.46\), \(p=.50\) ). This is numerically consistent with timbre specificity, as in Experiment 1a and Creel (2012). However, it is not at all consistent with stronger specificity effects when both mode and timbre pattern with meter.


Figure 2: Experiment 1, metrical restoration as a function of whether the melody was presented in its original mode and timbre, or in a different mode and/or timbre.

\section*{Discussion}

Experiment 1 replicated previous findings of timbre-specific metrical restoration (Creel, 2012). However, metrical restoration seems to generalize readily across a change in mode. Given the theoretical and affective importance of the major/minor mode distinction, and listeners' demonstrated sensitivity to the affective connotations of mode in the affect-rating cover task, this is somewhat surprising. One possible explanation is that, while harmonic cues to mode in the context were quite strong, cues to mode in the melody alone were weaker. (Note that timbre cues were still available in the melody alone.) If so, stronger harmonic contexts might reveal evidence of mode-specific metrical restoration. Another possibility is that in real musical styles, other factors that covary with mode carry the weight of mode's apparent stylistic impact. For instance, minor-mode melodies might use particular note sequences that are rare in major-mode melodies, and vice versa. Because the melodies used here were changed to minor mode simply by shifting certain pitches down by a small amount, no such stylistic differences were evident here.

\section*{Experiment 2}

While Experiment 1 found no effects of mode on metrical restoration, there were effects of timbre: listeners did not generalize metrical restoration across a change in timbre. One might take this to imply that timbre differences alone could be used to keep different musical styles separate from one another. If this were true, then even more diversity should lead to highly-specific metrical storage. The current experiment addressed this hypothesis.

Listeners were exposed to \(123 / 4\) and \(6 / 8\) melodies either in a single timbre, or in 12 different timbres (one per melody). If more timbres yields more distinct melody representations, then twelve-timbre listeners should show stronger metrical restoration than one-timbre listeners. Crossed with this was a rate-variability manipulation: listeners heard melodies at a single presentation rate, or in three different rates (consistent for a particular melody). If listeners store melodies rate-specifically, then multiple rates should yield more distinct representations, and hence, stronger metrical restoration.

\section*{Method}

Participants \(\mathrm{N}=72\) listeners from the UCSD human participant pool took part in the experiment.

Stimuli Stimuli were 12 major-mode melodies, a subset of those in Experiment 1. The 12 timbres were selected to be discriminable (Iverson \& Krumhansl, 1993) and to span multiple instrument families (percussion, strings, brass, woodwinds) which have timbres similar to each other. For a given participant, each melody had a single rate and timbre.

Design Participants were randomly assigned to one of four combinations of timbre variability ( 1 timbre or 12) and rate ( 1 rate or 3 ). Timbres and rates were counterbalanced such that each timbre, rate, and timbre-rate combination occurred roughly equally across participants and melodies.

Procedure The procedure was identical to that in Experiment 1, except for differences in the stimuli heard.


Figure 3: Experiment 2, metrical restoration effect as a function of timbre and rate variability.

\section*{Results}

Metrical restoration (Figure 3) was assessed in an ANOVA on probe ratings with Exposure Meter and Probe Meter as within-participants factors, and Timbre (one or twelve) and Rate (one or three) as between-subjects factors. As in Experiment 1, an Exposure Meter x Probe Meter interaction \((F(1,68)=14.52, p=.0003)\) indicated a metrical restoration effect overall.

However, none of the higher-level interactions-which would indicate timbre diversity or rate diversity effectswere significant: Exposure Meter x Probe Meter x Timbre \((F(1,68)=2.47, p=.12)\), Exposure Meter x Probe Meter x Rate \((F(1,68)=0.58, p=.45)\), or Exposure Meter x Probe Meter x Timbre x Rate \((F(1,68)=0.34, p=.56)\). Further, the direction of the Timbre interaction effect was numerically opposite that predicted: metrical restoration was more robust for listeners who heard a single timbre than for those who heard 12 different timbres. Individually, metrical restoration was significant only in the two single-timbre conditions (novariability: \(F(1,17)=14.64, p=.001\); rate-variability: \(F(1,17)\) \(=7.25, p=.02\) ). Thus, the strongest evidence for metrical restoration was carried by the low-diversity conditions.

\section*{Discussion}

The results of Experiment 2 were counter to predictions of timbre specificity and rate specificity. There was no effect of rate diversity, which one might think would be closely linked to meter as both properties emerge from musical timing. Further, instead of stronger metrical restoration when each melody had a unique timbre, metrical restoration was numerically smaller-absent-when each melody had a unique timbre. Why wouldn't 12 moredistinct melodies ( 12 unique timbres) generate stronger metrical restoration, due to specificity, than 12 less-distinct (identical-timbre) melodies?

One possible answer is that listeners were not associating timbres themselves with meter, but were associating timbrespecific motifs with meters. That is, listeners were aggregating traces that grouped according to perceptual similarity. Recall that Creel (2012) showed that motif similarity influenced metrical restoration. Suppose that listeners in the current experiment also associated meter with motif-like rhythmic patterns. The melodies in the current study were built from a small set of moderatelyambiguous rhythmic patterns, many of which occurred across multiple melodies. If the same rhythmic patterns were stored separately for separate timbres, then listeners who heard few timbres (Experiment 1; one-timbre condition of Experiment 2) might build up relatively strong motif representations. On the other hand, listeners who heard multiple timbres (the 12-timbre condition of Experiment 2) would store a larger number of timbre-specific motif representations, but these would be weaker because they had been exposed too few times. Thus, the 12-timbre condition may have yielded motif representations that were too weak to generate significant metrical restoration.

\section*{General Discussion}

I began by asking what factors influence metrical restoration. Experiment 1 suggests that timbre may be a stronger influence than mode (major vs. minor), at least in listening situations with limited harmonic information. Further, Experiment 2 suggests that similarity-based grouping of musical memory by timbre may decrease metrical restoration due to too little representational overlap.

Listeners may need many repetitions of motifs in a given timbre (and perhaps in a given rate) for metrical restoration to occur. As shown in Creel (2012), too much motif overlap between two meters blocks metrical restoration. The current study suggests that too little timbre-specific motif overlap in melodic patterns may also thwart metrical restoration. However, I hypothesize that this occurs for two different reasons: too much motif+timbre overlap between meters causes interference, while too little timbre-specific motif overlap within a meter yields representations too weak to generate restoration.

Ongoing work explores the roles of pattern (motif) overlap in metrical restoration. When Experiment 2 presented listeners with multiple rhythmic motifs in diverse timbres-i.e., rhythmic motifs were scattered across melodies and timbres-no metrical restoration was found. A new experiment ( \(N=96\) participants) again uses multiple timbres, but each rhythmic motif is in the same timbre, which should boost representation strength by increasing representational overlap. The magnitude of metrical restoration is .23 , far exceeding (nonsignificant) diversetimbre restoration observed in Experiment 2. This suggests that timbre uniqueness does not inhibit metrical restoration as long as timbres and motifs consistently cooccur. Whether distinct timbres facilitate motif encoding is still unknown.

\section*{The nature of musical memory}

These results speak not only to restoration of metrical information, but also to the organization of musical memory itself. I have argued previously (Creel, 2012) that fine auditory detail in musical memory is not an epiphenomenon, but an organizing force, where similaritybased grouping leads to emergent clusters in memory which shape recognition and processing. Like the "phoneme restoration effect" (Samuel, 1981), metrical restoration and other varieties of musical restoration (harmonic; Creel, 2011) indicate the strength of memory in the processing of auditory events. While the incoming signal itself certainly shapes music processing, memory too is a sizable influence on perception. My studies thus far suggest that musical memory is organized along at least two dimensions: timbre and motif. Future work should assess additional factors in metrical restoration, and also whether metrical restoration and other types of musical restoration-particularly, harmonic restoration-are influenced by the same musical dimensions. If they are not, this might suggest differing attentional weights across dimensions (Nosofsky, 1986) for processing of metrical vs. harmonic information.

An additional question concerns the nature of motifs. Motif structure appears to be a strong organizing force in musical memory, but it is unclear what, functionally, counts as a motif. Creel (2012) defined it as a rhythmic pattern with a particular pitch contour, but many other possibilities are equally consistent with the data: a rhythmic pattern alone; clusters of similar but not identical rhythmic patterns.

Finally, these investigations have implications for other temporal perception phenomena. For instance, Brochard et
al. (2003) showed an ERP signature of listeners' illusory perception of strong-weak alternations in a series of tones of equal amplitude. The research described here suggests that this effect may arise from memory activation of duple meters, the most common meter in Western music. Future studies of restoration effects should continue to reveal how temporal events are represented in the mind.

\section*{References}

Brainard, D. H. (1997). The Psychophysics Toolbox. Spatial Vision, 10, 433-436.
Brochard, R., Abecasis, D., Potter, D., Ragot, R., \& Drake, C. (2003). The "ticktock" of our internal clock: direct brain evidence of subjective accents in isochronous sequences. Psychological Science, 14(4), 362-6.
Castellano, M. A., Bharucha, J. J., \& Krumhansl, C. L. (1984). Tonal hierarchies in the music of North India. Journal of Experimental Psychology: General, 113(3), 394-412.
Creel, S. C. (2011). Specific previous experience affects perception of harmony and meter. Journal of Experimental Psychology: Human Perception and Performance, 37(5), 1512-1526.
Creel, S. C. (2012). Similarity-based restoration of metrical information: Different listening experiences result in different perceptual inferences. Cognitive Psychology, 65(2), 321-51.
Creel, S. C., Aslin, R. N., \& Tanenhaus, M. K. (2008). Heeding the voice of experience: The role of talker variation in lexical access. Cognition, 106, 633-664.
Gjerdingen, R., \& Perrott, D. (2008). Scanning the dial: The rapid recognition of music genres. Journal of New Music Research, 37(2), 93-100.
Goldinger, S. D. (1998). Echoes of echoes?: An episodic theory of lexical access. Psychological Review, 105(2), 251-279.
Hunter, P. G., Schellenberg, E. G., \& Schimmack, U. (2008). Mixed affective responses to music with conflicting cues. Cognition \& Emotion, 22(2), 327-352.
Iverson, P., \& Krumhansl, C. L. (1993). Isolating the dynamic attributes of musical timbre. The Journal of the Acoustical Society of America, 94(5), 2595-603.
Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. Journal of Experimental Psychology: General, 115(1), 39-61.
Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. Spatial Vision, 10, 437 - 442.
Samuel, A. G. (1981). Phonemic restoration: insights from a new methodology. Journal of Experimental Psychology: General, 110(4), 474-94.
Schellenberg, E. G., Iverson, P., \& McKinnon, M. C. (1999). Name that tune: Identifying popular recordings from brief excerpts. Psychonomic Bulletin \& Review, 6, 641-646.

\title{
Reasoning with differing tasks and response formats
}

\author{
Nicole Cruz de Echeverría Loebell (Nicole.Cruz@psychol.uni-giessen.de) \({ }^{1}\) Markus Knauff (Markus.Knauff@ psychol.uni-giessen.de) \({ }^{1}\)
}
\({ }^{1}\) Justus Liebig University Gießen, Experimental Psychology and Cognitive Science, Otto-Behaghel-Str. 10F, Giessen, 35394 Germany

\begin{abstract}
This study investigated the role probabilistic and deductive relations play in the reasoning process. It was predicted that when taking an analytic stance to a problem, it would take longer to evaluate inferences when asked how probable it is that the conclusion is true, than when asked whether the conclusion follows or not from the premises. Contrary to this prediction, people responded faster when the response format was continuous. However, there was no effect of argument type with continuous response format, suggesting people did not assess entailment relations in this condition. Options to address the issue further are discussed.
\end{abstract}

Keywords: deductive/inductive reasoning; dual process theories; task effects; response times.

\section*{Introduction}
"If the animal is a whale, then it must be a mammal"; "If I stay for five more minutes, I shall still catch the train"; "If you exchange these two cables, the telephone will work again". We go about the world constantly making judgments about what might be the case and what consequences we may expect from different situations and actions. Sometimes the reasoning involved occurs rather automatically, at other times it is effortful and time consuming. A lot of it involves conditionals, i. e. statements of the form "if p then q", with " p " and " q " standing for individual propositions such as "you stay 5 more minutes" and "you catch the train".

There is a debate in reasoning research regarding what criteria, or norms for when an inference is correct, people employ when drawing inferences - and, if they employ different criteria, then under what circumstances they reason according to what criterion and in what way the criteria may interact. The two main norms under discussion are a deductive, deterministic one (the conclusion is correct if it follows necessarily from the premises) and a probabilistic one (e. g. the conclusion is correct if its uncertainty is not greater than the sum of the uncertainties of the premises, Adams, 1975). People's answers to reasoning problems are generally sensitive both to the structure of deductive entailment relations involved and to the subjective probability or plausibility of the contents appearing in the relations (i. e. Thompson, 1994; Singmann \& Klauer, 2011).

In some approaches it is argued that people reason using a single norm for argument validity across situations. A major proponent of this position, the theory of mental models (Johnson-Laird \& Byrne, 2002), postulates this to be the deductive norm and accounts for the effect of contextual and probabilistic information on people's inferences by
proposing that people integrate such information in their models of the situation, either by adding or subtracting possibilities considered, or by tagging the models with probabilities (Girotto \& Johnson-Laird, 2004; JohnsonLaird, Legrenzi, Legrenzi, Girotto, \& Caverni, 1999). A further major proponent of the single-criterion position is the probabilistic theory of Oaksford and Chater (Oaksford, Chater, \& Larkin, 2000; Oaksford \& Chater, 2007), which postulates that the effect of contextual and probabilistic information is a consequence of that people generally reason not deductively but probabilistically, in a way that is ecologically rational and that can be modeled using Bayesian theory together with a few further assumptions.

The idea that people use a single norm for argument validity across situations is put into question by a number of findings. Rips (2001) found that when given the same list of arguments which were valid/invalid as well as plausible/implausible, a group of people given deductive instructions endorsed the valid but implausible arguments more often than the invalid but plausible ones. The opposite was the case for a group of people given inductive instructions. Vadeboncoeur and Markovits (1999) found that emphasizing the deductive nature of a task in the instructions led to answers in stronger accordance with such instructions, but that also then the availability of counterexamples to the arguments (making them less plausible even though they were valid) still had an effect. Also the availability of probabilistic information was found to have an effect on people's approach to reasoning problems. For instance, Wolf and Knauff (2008) found that people's strategy of belief revision with conditional inferences was a function of the probability of the conditional when this probability was high or low, but was better explained by the theory of mental models when the probability of the conditional was close to .5 and thus perhaps less informative. A further factor found to influence people's reasoning is the task employed. For example, across several studies the theory of mental models offered a better explanation of reasoning in the conditional inference task, while the probabilistic approach could explain better findings in the truth table task, which is related more directly to the interpretation of conditionals (Geiger \& Oberauer, 2010). Finally, also the response format for otherwise identical tasks, especially whether this is dichotomous or not, has been found to play a role. Oberauer, Geiger, Fischer, and Weidenfeld (2007) found that in the truth table task, the same participants who answered in accordance with a probabilistic interpretation of the
conditional having the three response options "true", "false", and "irrelevant", answered in accordance with a mental model interpretation when the response option "irrelevant" was not available. Further, Markovits and Handley (2005) found that while probability ratings of the arguments of the conditional inference task where uniformly high, proportion of endorsement of the same inferences having binary response format was significantly lower, especially when the inferences where deductively invalid.

Findings like the ones described have led to increasing attempts to find integrative approaches, often in the form of dual-process theories, which assume that people may employ different criteria and ways of thinking under different circumstances. Hereby one process is often described as analytic, under more conscious control, more dependent on working memory resources and more context independent, and the other as heuristic, fast, automatic, context dependent and not much affected by working memory constraints. For instance, Klauer, Beller, \& Hütter (2010) distinguish between a process based on the "logical form" or entailment relations in an argument and one based on content and context information. Sloman (1996, 2002) distinguishes between an associative and a rule based process. The two processes can be related in different ways. For instance, Evans and Over (Evans, 2006; Evans, Handley, Neilens, \& Over, 2010) advocate a defaultinterventionist relation, in which the heuristic process is used as the default, and the analytic process may intervene if there is enough time and the heuristic answer seems insufficient to solve the task. Verschueren, Schaeken, and d'Ydewalle (2005) propose that both processes operate in parallel on a given task, and if the analytic process has enough time and leads to a different result than the heuristic process, it will override the answer arrived at by the heuristic process.

One difficulty with dual-process theories is that they often only explain the effect of deductive validity through the analytic system, while the construction of a representation of the problem to be evaluated can be better attributed to the heuristic system. This puts into question their role as independent forms of solving the same reasoning problem. Also, findings from de Neys (e. g. 2012) suggesting people have not only intuitive heuristics but also logical intuitions, question the idea of an association between the heuristic and the probabilistic on the one hand, and the analytic and the deductive on the other.

The present study aims at investigating further the role of deductive and probabilistic aspects of the reasoning process. Although in general it is plausible that people may approach a task in different ways depending on their goals and constraints of the situation, it is hypothesized that at least some of the findings proposed as evidence for two systems of reasoning may also be explained by making a less strong assumption: through the idea that the reasoning process is a composite one, in which different processes take over different components of the reasoning task, instead of reflecting different approaches to the same task. The two
components considered here are assessment of the probability that a statement is the case (related to the interpretation of the statement) and assessment of what follows from the assumption that a statement is the case. The task of assessing whether something is the case is considered probabilistic: in the context of a conversation, it would be a matter of debate and subject to varying degrees of confidence. In contrast, the task of assessing what follows from the assumption that something is the case is considered (given a deductive task) as deductive and thus in a way deterministic, not probabilistic (something follows or it does not follow from given assumptions). In daily life we are often interested not just in what follows from assuming a certain piece of information, but also in how probable the conclusion itself is: we want to take into account also the uncertainty in the premises and transfer it to the conclusion. However, this is proposed to be a separate task within the reasoning process.

Thus, we hypothesized that, provided people approach a task analytically, it should take longer to answer to the question: "how probable is it that the conclusion from the premises is true?" than to the question: "does the conclusion follow from the premises?" Conversely, if people are given not inferences but only statements to evaluate, it should be faster to answer to the question: "how probable is it that this statement is true?" than to the question: "is this statement true or false?" since the latter case would involve the additional task of setting a threshold - above which one says "yes, it is true" and below which one says it is false - and of comparing the probability of the statement with this threshold. In order to raise the probability that people approach the task analytically, people are often given no time pressure as well as deductive instructions emphasizing the importance of assuming the truth of the premises for the sake of argument. We gave participants no time pressure, but could not emphasize deductive instructions since we wanted to assess the effect of taking into account premise probabilities in addition to entailment relations. We hoped that enough participants would nonetheless take an analytic stance given that in dual-process theories the weight obtained for the parameter representing an analytic approach to the task was often above \(50 \%\) for both binary (Oberauer, 2006) and continuous (Klauer, Beller, \& Hütter, 2010) response formats.

\section*{Method}

\section*{Participants}

Thirty-two students from the University of Giessen took part in the experiment in exchange for payment or course credit. Their mean age was 23.6 years (range: 19-31). They came from different majors, with the exclusion of mathematics, informatics, physics and philosophy. One participant \({ }^{1}\) had taken a course in logic; sixteen had taken at least one course in statistics.

This participant did not show a deterministic response pattern, and her exclusion did not change the pattern of results.

\section*{Design}

The above hypotheses were assessed through a within subject design involving the two main variables task (evaluation of statements or of inferences) and response format (continuous, dichotomous). For statements, a further distinction was made between conditional statements ("if p then \(\mathrm{q}^{\prime}\) ) and the two statements the conditional is composed of ("p" and "q"). For inferences, one could further distinguish inference form. There were four inference forms: "Modus Ponens" (MP: "if p then q", "p", therefore "q"), "Modus Tollens" (MT: "if p then q", "not-q", therefore "not-p"), "Affirmation of the consequent" (AC: "if p then q", "q", therefore "p"), and "Denial of the antecedent" (DA: "if p then q", "not-p", therefore "not q"). Only the first two are deductively valid, because in the other two cases also the negation of the conclusion is compatible with the premises (However, if the conditional is interpreted as a biconditional: "p if and only if q " then all four inferences are deductively valid). The main dependent variable was response latency, but degree of resp. frequency of endorsement was also examined.

\section*{Material and procedure}

Participants viewed either statements or inferences on the computer screen, and were asked to evaluate them on a continuous or dichotomous scale. Statements and inferences were embedded in one of four contexts involving concrete materials but describing arbitrary relations. For example, one such context was the following:

In a workshop in Soko there is a cupboard with blue and yellow drawers for storing the nails and screws. One drawer of the cupboard is opened...
On the next screen appeared the statement or inference to be evaluated, e. g. "If the drawer is blue, then there are nails in it". There were three types of statements: conditionals like the one above ( \(\mathrm{p}->\mathrm{q}\) ), and two statements corresponding to the antecedent ( \(p, ~ e . ~ g . ~ " t h e ~ d r a w e r ~ i s ~ b l u e ") ~ a n d ~ t o ~ t h e ~\) consequent ( \(q\), e. g. "the drawer has nails in it") of the conditional, respectively. There were four kinds of inferences, corresponding to MP, MT, AC and DA. For statements, participants were asked "How probable is it that this statement is true?" with continuous response format (cont), and "Is this statement true or false?" with dichotomous response format (dic). For inferences, the task was to "Consider the statements. How probable is it that the conclusion is true?" with continuous response format and "Assume the statements are true. Does the conclusion follow necessarily from them?" with dichotomous response format. Here we spoke of an evaluation of "the conclusion" and not of a specific statement per se, to make explicit that both response formats involve the evaluation of inferences and not just of statements grouped with other statements.

The continuous response scale was a horizontal line with the endpoints " \(0 \%\) " and " \(100 \%\) " and was divided into 101 points that could be clicked with the mouse. The dichotomous response scale consisted of two adjacent boxes, together as long as the horizontal line of the
continuous response scale, below which stood the words "false" and "true" for statements, and "does not follow" and "follows" for inferences. To the right of each statement and each premise stood a small box filled up to a certain point, representing the probability of the statement (the fuller the box, the more probable the statement). There were four boxes representing the probabilities \(.2, .4, .6\) and .8 . The aim of these boxes was to provide premise probabilities in a non-numeric and yet relative standardized way.

Each of the four contexts was associated with the three statement types, yielding 12 statements for each response format. Further, each context was associated with the four inference types, leading to 16 inferences for each response format. For each participant, one of the four probabilities was randomly assigned to the conditional of one of the four contexts and held constant across the experiment, mimicking the reliability of conditional relations. For each context, the other three probabilities were distributed randomly without replacement across statements, such that e. g. for the context of the workshop, the second premise had a different probability for each of the four inferences. The order of occurrence of the statements and of the inferences was varied randomly for each participant.

Participants were tested individually in two sessions. One session involved evaluation of the 24 statements, the other evaluation of the 32 inferences. The order of sessions was counterbalanced across participants. Within each session, response format was blocked. Instructions at the beginning of each block included familiarization with the response scale and a sample trial. At the end of the second session, all participants worked through 20 trials in which the two response scales were presented alone on the screen (10 times each in random order) and they were to click with the mouse on them as quickly and as randomly as they could. This served to assess differences in response time to the two scales due to processes unrelated to the reasoning task (i. e. motor affordances). This difference was later subtracted from the answers to the reasoning task by centering the values of each participant in each response format around their mean for that response format when presented alone. The experiment was self-paced and lasted about 50 minutes.

\section*{Results and discussion}

The data were analyzed separately for response times and for endorsement ratings as dependent variable. Prior to the analysis of response times, responses faster than 100 ms were eliminated, leading to exclusion of two data points. Elimination of response times outside the interval of the mean plus minus 3 SD for each variable led to no further data exclusions. Since response times have a lower threshold, they do not follow a normal distribution. To compensate for this, the inverse of response times: speed (1/RT), was taken for analysis. This normalizes somewhat the distribution and reduces the impact of outliers while preserving power and ease of interpretation (Whelan, 2008). Measures of speed were then multiplied by 1000 to avoid
working with only very small values (Baayen \& Milin, 2010). Prior to the analysis of endorsement ratings, it was necessary to represent the probability ratings obtained with continuous response format, and the endorsement frequencies obtained with dichotomous response format on the same scale. This was done by transforming mean frequency of the dichotomous items into a percentage value. For example, if a person answered three times yes (coded 1) and one time no (coded 0 ), the mean frequency of acceptance was \((1+1+1+0) / 4=.75=75 \%\) (Markovits \& Handley, 2005). It is thereby important to keep in mind that probability ratings and endorsement frequencies are different measures and may not be directly comparable. Results from such comparisons can be illustrative and useful, but should be interpreted with caution (Singmann \& Klauer, 1010).

Separately for both response speed and endorsement ratings, three ANOVAS were conducted: a general ANOVA across tasks, assessing the effects of task (statements, inferences) and of response format (continuous, dichotomous); an ANOVA for statements assessing the effect of statement type ( \(p->q, p, q\) ) and response format (cont, dic); and an ANOVA for inferences assessing the effect of inference type (MP, MT, AC, DA) and response format. The Greenhouse-Geisser correction of degrees of freedom for lack of sphericity was applied when appropriate. The results are depicted in Figure 1.
in speed due to the scales alone. This analysis (not represented in Figure 1) yielded a main effect of task, \(F(1\), \(31)=134.72, p<.001\), partial \(\eta^{2}=.81\) : answers to statements were faster than to inferences; a main effect of response format, \(F(1,31)=14.96, p=.001\), partial \(\eta^{2}=.33\) : answers were faster when the response format was dichotomous than when it was continuous; and an interaction between task and response format, \(F(1,31)=\) \(6.58, p=.015\), partial \(\eta^{2}=.18\) : the extent to which answers were faster when the response format was dichotomous was greater when evaluating statements than when evaluating inferences. This same ANOVA was then repeated correcting for differences in speed due to the scales alone, i . e. centering the values of each participant in each response format around the participant mean for that response format when presenting the scale alone. This analysis is shown in the upper left panel of Figure 1. It yielded a main effect of task, \(F(1,31)=134.72, p<.001\), partial \(\eta^{2}=.81\) : answers to statements were faster than to inferences; a main effect of response format, \(F(1,31)=15.1, p=.001\), partial \(\eta^{2}=.33\) : answers were faster when the response format was continuous; and an interaction between task and response format, \(F(1,31)=6.58, p=.015\), partial \(\eta^{2}=.18\) : the extent to which answers were faster when the response format was continuous was greater when evaluating inferences than when evaluating statements.

Thus, while in absolute terms it took longer to answer to


Figure 1. The upper panel shows mean speed (adjusted for RT differences between scales when presented alone) of responses for continuous (cont) and dichotomous (dic) response format, across tasks (left column), for statements (middle column) and for inferences (right column). The lower panel shows probability ratings (when response format \(=\) cont \()\) resp. endorsement frequency (when response format \(=\) dic) for the same conditions. Error bars show within subject standard errors (Bakeman \& McArthur, 1996).

For the sake of exposition clarity, only results considered relevant for the hypotheses will be reported in detail. The main hypothesis concerns the effect of task and of response format on response speed. Initially, this analysis was conducted using response speed not adjusted for differences
the continuous than to the dichotomous scale, this relation was reversed when adjusting for differences in response times to each scale when presented alone, such that participants were faster when the response format was continuous. This is in accordance with our hypothesis for
judgments about statements, but contrary to our hypothesis for judgments about inferences.

A possible explanation for why responses where faster with continuous response format both when evaluating statements and when evaluating inferences lies in the lower right panel of Figure 1, depicting endorsement ratings resp. endorsement frequency of the four inferences (MP, MT, AC, DA) as a function of response format. This analysis yielded no effect of response format, \(F(1,31)=1.07, p=\) .31, partial \(\eta^{2}=.03\); an effect of inference type, \(F(3,93)=\) 4.26, \(p=.007\), partial \(\eta^{2}=.12\); and an interaction between inference type and response format, \(F(1,31)=5.77, p=\) .001, partial \(\eta^{2}=.16\). The graphic shows the typically observed pattern of response to the four inferences for dichotomous response format (Bonferroni corrected \(t\)-tests only yielded a significant difference between MP and AC ratings, \(t(31)=4.3, p<.001\) ), whereas there was not a trace of an effect of inference type for continuous response format. Thus, people seem to have taken into account differences in the entailment relations making up the structure of the arguments only when the response format was dichotomous, but not when it was continuous. This finding renders it understandable that people were faster when the response format was continuous.

Finally, it is interesting to note that there was no effect of response format in all three analyses of endorsement ratings resp. endorsement frequency (lower three panels of Figure 1): In the ANOVA across tasks: \(F<1\); In the ANOVA for statements: \(F(1,31)=3.44, p=.07\), partial \(\eta_{2}=.1\); and in the ANOVA for inferences: \(F(1,31)=1.07, p=.31\), partial \(\eta_{2}=.03\).

The absence of an effect of response format in all three analyses speaks against the idea that people build a threshold close to certainty in the condition with binary response format, as had been suggested by Markovits and Handley (2005), who also compared answers in the conditional inference task with binary and continuous response format and found lower levels of inference endorsement when the response format was binary. It rather suggests people endorsed a probabilistic interpretation of the statements throughout: No effect of response format is expected when people judge a statement as true when they judge its probability to be above \(50 \%\) and as false when they judge its probability to be below this value. This is a sensible strategy from a probabilistic perspective because then one's judgments will be right over \(50 \%\) of the time on average. One explanation for the difference between our results and those of Markovits and Handley is that in our experiment one could explicitly see a representation of the statements' probabilities, and this may have made it more likely that they were taken into account as criteria for the judgments. One could assess the issue further using other means of providing probability information, such as through the introduction of a probability learning phase to simulate natural sampling, or through the use of familiar relations for which people can readily build probability estimates.

Although the absence of an effect of inference type for judgments with continuous response format provides a reason for why people's answers were generally faster when the response format was continuous, this absence of an effect is itself surprising and therefore worthy of further consideration. In the study from Markovits and Handley (2005) a similar pattern was observed, with the exception of ratings for MP, which were higher than for the other three inferences. Singman and Klauer (2011), using only a continuous response format, found a more pronounced effect of inference type. No effect of inference type can be expected in the framework of dual-process theories when people take a heuristic stance to the task. A heuristic stance could have been promoted in this experiment through the complexity of the task: In contrast to the two studies above, the relations employed here were arbitrary and each premise was provided with explicit probability information. This may have made it more difficult to explicitly both assess the entailment relations involved in the argument and integrate their probabilities. Thus, one could assess what effect results from simplifying the task, e. g. by providing probability information implicitly by using familiar conditional relations for which people readily build an idea of their probability. This would have the additional benefit that the validity and the soundness of the inferences would converge, ruling out the possibility that people's answers showed no effect of inference type because they were judging not their validity but their soundness, which in the arbitrary relations employed was set to be constant \({ }^{2}\).

The main prediction of this study was that, provided people take an analytic stance to a problem, it would take longer to evaluate inferences when asked how probable it is that the conclusion is true, than when asked whether the conclusion follows from the premises, because integrating probabilities is an additional task to assessing entailment relations. In contrast, we predicted it to take less time to evaluate the truth of a statement when asked how probable it is that the statement is true than when asked whether the statement is true or false, since the latter would involve the additional task of setting a threshold and comparing it with the statements probability. We found that people where generally faster with continuous response format, and that when judging inferences, the entailment relations constituting the structure of the arguments had an effect for dichotomous but not for continuous response format. The results were in accordance with the hypotheses for statement evaluation, but not for inference evaluation. However, they suggest that when evaluating inferences, people did not take an analytic stance to the task when the response format was continuous. One way to promote an analytic stance could be through instructions introducing the task explicitly as one aimed at investigating how analytic reasoning differs from intuitive reasoning, making it important to engage in the former for the sake of the experiment. Such a manipulation was successful in eliciting a heuristic stance in a study from

\footnotetext{
\({ }^{2}\) We thank Momme von Sydow for this helpful suggestion.
}
de Neys and Franssens (2009). One could then assess whether this would make a difference.

\section*{Acknowledgments}

This work was supported by Grant KN 465/11-1 to MK from the Deutsche Forschungsgemeinschaft (DFG) as part of the priority program "New Frameworks of Rationality" (SPP 1516).

\section*{References}

Adams, E. W. (1975). The logic of conditionals: An application of probability to deductive logic. Dordrecht, Holland: Reidel.
Baayem, R. H., \& Milin, P. (2010). Analyzing reaction times. International Journal of Psychological Research, 3(2), 12-28.
Bakeman, R., \& McArthur, D. (1996). Picturing repeated measures: Comments on Luftus, Morrison, and others. Behavior Research Methods, Instruments, \& Computers, 28(4), 584-589.
De Neys, W. (2012). Bias and conflict: A case for logical intuitions. Perspectives on Psychological Science, 7(1), 28-38.
De Neys, W., \& Franssens, S. (2009). Belief inhibition during thinking: Not always winning but at least taking part. Cognition, 113, 45-61.
Evans, J. S. B. T. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. Psychonomic Bulletin \& Review, 13(3), 378-395.
Evans, J. St. B. T., Handley, S. J., Neilens, H., \& Over, D. (2010). The influence of cognitive ability and instructio- nal set on causal conditional inference. The Quarterly Journal of Experimental Psychology, 63, 892-909.
Geiger, S., \& Oberauer, K. (2010). Towards a reconciliation of mental model theory and probabilistic theories of conditionals. In M. Oaksford \& N. Chater (Eds.), Cognition and conditionals: Probability and logic in human thinking, (pp. 289-308). Oxford, UK: Oxford University Press.
Girotto, V., \& Johnson-Laird, P. N. (2004). The probability of conditionals. Psychologia, 47, 207-225.
Johnson-Laird, P. N., \& Byrne, R. M. J. (2002).
Conditionals: A theory of meaning, pragmatics, and inference. Psychological Review, 109, 646-678.
Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M. S., \& Caverni, J. P. (1999). Naive probability: A mental model theory of extensional reasoning. Psychological Review, 106, 62-88.
Klauer, C. K., Beller, S., \& Hütter, M. (2010). Conditional reasoning in context: A dual-source model of probabilistic inference. Journal of Experimental Psychology: Learning, Memory, and Cognition, 36(2), 298-323.
Markovits, H., \& Handley, S. (2005). Is inferential reasoning just probabilistic reasoning in disguise? Memory \& Cognition, 33(7), 1315-1323.

Oaksford, M. \& Chater, N. (2007). Bayesian Rationality:
The probabilistic approach to human reasoning. Oxford, UK: Oxford University Press.
Oaksford, M., \& Chater, N., \& Larkin, J. (2000).
Probabilities and polarity biases in conditional inference. Journal of Experimental Psychology: Learning, Memory, and Cognition, 26, 883-899.
Oberauer, K. (2006). Reasoning with conditionals: A test of formal models of four theories. Cognitive Psychology, 53(3), 238-283.
Oberauer, K., Geiger, S. M., Fischer, K., \& Weidenfeld, A. (2007). Two meanings of "if"? Individual differences in the interpretation of conditionals. The Quarterly Journal of Experimental Psychology, 60(6), 790-819.
Rips, L. J. (2001). Two kinds of reasoning. Psychological Science, 12, 129-134.
Singmann, H., \& Klauer, K. C. (2011). Deductive and inductive conditional inferences: Two modes of reasoning. Thinking \& Reasoning, 17(3), 247-281. Soman, S. A. (1996). The empirical case for two systems of reasoning. Psychological Bulletin, 119(1), 3-22. Sloman, S. A. (2002). Two systems of reasoning. In T. Gilo vich, D. Griffin, \& D. Kahneman. Heuristics and Biases: The Psychology of Intuitive Judgment. Cambridge: Cam bridge University Press.
Thompson, V. A. (1994). Interpretational factors in conditional reasoning. Memory and Cognition, 22, 742758. Whelan, R. (2008). Effective analysis of reaction time data. The Psychological Record, 58, 475-482. Vadeboncoeur, I., \& Markovits, H. (1999). The effect of instructions and information retrieval on accepting the premises in a conditional reasoning task. Thinking and Reasoning, 1999, 5(2), 97-113.
Verschueren, N., Schaeken, W., \& d'Ydewalle, G. (2005). A dual-process specification of causal conditional reasoning. Thinking \& Reasoning, 11, 239-278. Whelan, R. (2008). Effective analysis of reaction time data. The Psychological Record, 58, 475-482.
Wolf, A. \& Knauff, M. (2008). The strategy behind belief revision: A matter of judging probability or the use of mental models? In B. C. Love, Kl McRae, \& V. M. Sloutsky (Eds.), Proceedings of the \(30^{\text {th }}\) Annual Conference of the Cognitive Science Society (pp. 64-70). Austin, TX: Cognitive Science Society.

\title{
Artificial grammar learning of shape-based noun classification
}

\author{
Jennifer Culbertson (jculber4@gmu.edu) \\ Linguistics Program, 4400 University Drive \\ Fairfax, VA 22030 USA
}

\author{
Colin Wilson (colin@cogsci.jhu.edu) \\ Cognitive Science Department, 3400 N. Charles Street \\ Baltimore, MD 21218 USA
}

\begin{abstract}
Systems of noun classification serve to categorize entities based on a set of semantic and/or phonological features. Previous work, for the most part focused on gender-based classes, has suggested that learners acquiring such systems rely primarily on phonological cues, while semantic cues are used only weakly. We show, using an artificial language learning task with adults, that semantic information alone is sufficient to learn a realistic shape-based classification system, challenging the view of phonology bias. Further, our results show that compared to learners exposed to semantically cohesive categories, learners trained on randomly assigned classes are less successful at recalling the category of exposure items. This finding suggests that, contrary to memory-based theories of learning, categories are not necessarily formed by abstraction from memorized exemplars, but can instead be constructed from lower-level properties that category members share.
\end{abstract}

Keywords: classifiers; noun classes; language acquisition; artificial language learning; semantic features

\section*{Introduction}

Systems of noun classification-such as gender, noun class, and classifier systems-distinguish or categorize objects according to salient semantic and/or phonological features. Though such systems may differ in their formal realization, the semantic features on which they are based draw from a common pool that includes physical features (e.g. shape, size), function (e.g. food, tool, habitation), as well as animacy and sociocultural status (Denny, 1976; Dixon, 1986; Lakoff, 1987; Comrie, 1989; Aikhenvald, 2000; Senft, 2000).

For example, in Cantonese, the use of a classifier morpheme is required in constructions involving a numerical or definite noun phrase, as in example (1) below. \({ }^{1}\) The choice of classifier in Cantonese is largely determined by the head noun; for example the classifier \(g o[3]\) is used for people, while the classifier zek[3] is used primarily with animals. Additional classifiers target shape properties like length, dimension, and flexibility.
(1)

> a. sam[1] go[3] jan[4]
> three CL people
> 'three people'
b. sam[1] zek[3] gau[2]
three CL dogs
'three dogs'

\footnotetext{
\({ }^{1}\) Although English does not use them productively, there are nevertheless a number of nouns which can appear with a classifier, e.g. "four strands of hair", "two sheets of paper", "a school of fish".
}

Similarly, in the classifier system of Navajo (Mithun, 1986) nouns are classified according to animacy and shape (among other properties); class marking in this language is found on the verb. Signed languages also commonly have noun classification systems based on shape and other functional properties (Supalla, 1986).

\section*{Acquisition of noun classification systems}

Previous work on the acquisition of systems of noun classification has largely focused on genders and noun classes. Such studies have documented developmental stages including a period of phonological underspecification, and overgeneralization of frequent or default marking, and have highlighted the apparently weak role of semantic (as opposed to phonological or distributional) information (KarmiloffSmith, 1981; Perez-Pereira, 1991; Demuth \& Ellis, 2008; Mariscal, 2009; Gagliardi, 2012). The acquisition of classifier systems, although perhaps less well-studied, indicates a similar developmental trajectory. For example, Tse, Li, and Leung (2007) report that Cantonese-speaking children \((3 ; 0-5 ; 0)\) tend to show early use of classifiers in required contexts but are not adult-like in their choice of classifier until quite late. In particular, children tend to over-use the classifier go3-used for people, but also sometimes referred to as a 'general' classifier (C. Li \& Thompson, 1989)—and to over-generalize other more frequent classifiers. Although P. Li, Huang, and Hsiao (2010) show that Mandarin-speaking children generalize classifiers to novel nouns on the basis of shape features, Tsang and Chambers (2011) argue that adult speakers of Cantonese tend to rely on cues other than the semantic features of the nouns when processing classifiers.

In this paper we investigate the extent to which adult learners can use semantic information alone to acquire category distinctions instantiated in a miniature classifier system. Previous work on artificial language learning suggests that, although the population of most interest may be children within any sensitive period for language acquisition, behavioral patterns exhibited by adults can shed light on both general and language-related learning mechanisms (Wilson, 2006; Culbertson, Smolensky, \& Legendre, 2012; Finley \& Badecker, 2010). The motivation for using an artificial language learning task rather than natural language learning data in this case comes from our hypothesis of why it has been found that phonological cues-even when these are less statistically reliable than semantic properties-are preferentially used by
learners acquiring noun classification systems (Braine, 1987; Frigo \& McDonald, 1998; Gagliardi, 2012). It seems likely that children process a great deal of phonological information about dependencies between nouns and nominal modifiers (such as gender-marked determiners or classifiers) before they acquire the meanings of these elements (Polinsky \& Jackson, 1999). In some sense, then, it is unsurprising that children privilege phonological information at first during language development. Adults may continue to privilege phonological cues, not because they fail to attend to semantics, but simply because their knowledge of noun classes was initially based in phonological processing.

Here, crucially, we use adult English-speakers and construct a miniature language from known objects and their linguistic labels. This removes the problem of acquiring the semantics of nouns and, if our hypothesis is correct, should expose an ability to learn cohesive noun categories on the basis of semantic features alone. While some previous work has suggested that adults can use semantic information to learn classification systems in an artificial language, these studies have exclusively focused on gender-based noun classes (Braine, 1987; Brooks, Braine, Catalano, Brody, \& Sudhalter, 1993). Here we target instead shape-based classifiers, which are likely to be less familiar to English-speaking college students (the population typically targeted).

The system is modeled on Cantonese (sortal) classifiers, in particular those which pick out shape properties of objects. As mentioned above, the particular shape properties indicated by Cantonese classifiers-related to the length, flexibility, and dimensions of objects-are representative of those found in classifier systems typologically (Craig, 1986; Dixon, 1986; Comrie, 1989). Table 1 shows the two Cantonese classifiers, along with the semantic features with which they are associated, on which our system was modeled. The examples provided represent nouns which take the relevant classifier in Cantonese, and are also nouns actually used in the task.

Table 1: Shape-based classifiers tested
\begin{tabular}{lll}
\hline Classifier & Semantic features & Examples \\
\hline zi[4] & rigid, narrow, long & knife, twig, candle \\
jeung[4] & broad, flat, flexible & sheet, card, table \\
\hline
\end{tabular}

\section*{Experiment 1}

In Experiment 1, we tested whether adults could learn and generalize categories of nouns, distinguished by their use of the classifiers in Table 1. We compare learning of a system in which classifier use is conditioned on shape-based semantic properties of nouns to learning a random assignment of nouns to classifier categories. We hypothesized that if learners perceive and make use of semantic information in acquiring noun classification systems, they should succeed in learning the semantically-conditioned language. The
random-assignment condition was used to establish an experimental baseline against which performance in the shapebased condition can be compared, and in particular to assess the role of memory for individual category members in this task. Exemplar-based models of learning argue that category formation begins with a set of memorized exemplars, abstract categories emerging later due to, e.g., computation of featural similarity among exemplars in a given category (Nosofsky, 1986). This predicts that learners exposed to conditioned and random classifier categories should perform equally well when tested on familiar items-in both cases, the set of exemplars presented during exposure should be stored-but should of course differ on their ability to generalize to novel items.

\section*{Participants}

Participants were 20 native English-speaking undergraduates from the Johns Hopkins University. They received a small amount of course credit or extra credit for their participation. No subjects reported difficulties hearing or seeing the stimuli.

\section*{Materials}

The miniature language was comprised of the English numeral words "one" and "two", two nonce classifier morphemes "ka" and "po", and 96 English nouns representing familiar objects. Utterances in the language consisted of a numeral word directly followed by a classifier morpheme, and a noun, as in example (2) below. Utterances were auditorilyusing mac text-to-speech, speaker "Alex"-and orthographically presented and were accompanied by a visual image. The image was a single object for numeral "one" or two of the same objects for numeral "two".
(2) a. one-ka hammer one-CL hammer
b. two-po towel two-CL towel 'two towels'


Figure 1: Example trial

\section*{Design \& Procedure}

Participants were seated in front of a computer, and were instructed that the task was about learning a language similar to English but with two ways of saying the words "one" and "two". They then listened to examples of "one-po", "one-ka" and "two-po", "two-ka". This was followed by 48 familiarization trials, half with objects using the classifier "ka" and half using "po". Half of the trials featured a single object and the other half two objects. On each trial, a visual image appeared with four choices below it, one for each possible numeral-classifier combination followed by the object noun pictured. Participants listened to the auditory stimulus and were required to click the choice which matched what they heard. Figure 1 shows an example trial.

After familiarization, participants took a brief break, and were then instructed that they would see a visual image and four choices below it, as in the familiarization phase, but they would hear no audio. Instead they were required to choose the phrase they thought was most likely to be used in the language. This testing phase was made up of 96 trials, including all the objects seen during familarization, and 48 novel objects. The seen objects were the same as those seen in the familiarization phase, but appeared with the other numeral (e.g. if a participant heard "one-ka hammer" and saw a single hammer during exposure, they saw two hammers at test). No feedback was given.

Participants were randomly assigned to one of two conditions. In the shape condition, the use of "ka" and "po" was conditioned on the semantic properties shown in Table 1 above. The object nouns in each class were a subset of those which actually use the corresponding classifier in Cantonese. As such, although they generally exhibited the relevant properties, there was some amount of variation in the extent to which they did so. For example, the noun "table" takes the classifier jeung[4] in Cantonese even though it does not perfectly exemplify the semantic features of the class.

In the random condition, the use of "ka" and "po" was unconditioned, and nouns were randomly paired with a particular classifier.

\section*{Results}

In analyzing the results of this experiment we were interested in two main questions: (i) Do learners in the shape condition-in which classifier choice is determined by semantic features of nouns-succeed in learning and generalizing the correct categories? (ii) Are the categories learned those which were intended, namely the shape-based categories shown in Table 1? To address the first question, we compared first the performance on seen items across the two conditions. Performance on seen items gives an indication of how well the familiarization set was learned by a given participant. The light colored bars in Figure 2 shows proportion choice of the correct classifier on average for participants in each condition. Analysis of this data using mixedeffects logistic regression (with participants and items as ran-
dom effects) reveals a significant effect of condition ( \(\beta=\) \(1.47, z=5.32, p<0.01\) ), with participants in the shape condition choosing the correct classifier on seen items much more often than those in the random condition ( 0.86 vs .0 .45 ). A significant interaction between condition and number was also found ( \(\beta=-0.29, z=-2.63, p<0.01\) ), indicating the participants in the random condition tended to be less accurate on items with the number "two" compared to "one".

We are also interesting in the extent to which participants in the shape condition could generalize the categorization information they learned during familiarization to novel (unseen) objects at test. As Figure 2 suggests, there was little difference in participants' choice of the correct classifier on seen item, and their choice of the classifier which matched the relevant semantic features on novel nouns. Analysis using mixed-effects logistic regression revealed no significant effect of item familiarity ( \(\beta=0.27, z=1.13, p=0.26\) ). A significant interaction between item familiarity and number was found however ( \(\beta=-0.47, z=-1.98, p<0.05\) ), indicating the participants tended to be less accurate on seen items with the number "one" compared to "two". Note that for participants in the random condition, there is no expected correct classifier for novel items, as the noun categories used in familiarization were random, containing no semantic cues.


Figure 2: Correct choice of classifier for seen and novel nouns in the shape condition, and seen items in the random condition (NB: there is no correct choice for novel nouns in the random condition, since nouns were categorized randomly).

If participants in the shape condition in fact consistently inferred the same set of shape-based categories, we expect to see that their responses on novel test items are highly correlated. On the other hand, participants in the random condition were not expected to infer cohesive categories, and thus we do not expect correlated responses. To assess this, for each pair of participants in the shape condition, we computed the proportion of novel test items that they assigned to the same category. The average agreement proportion for this condi-
tion was high \((0.74, \mathrm{SE}=0.04)\). In contrast, a parallel analysis revealed much lower agreement among participants in the random condition \((0.50, \mathrm{SE}=0.02)\); note that \(50 \%\) agreement would be expected from purely random responding.

\section*{Experiment 2}

In Experiment 2, we sought to replicate our findings in a more diverse population, namely workers on Amazon Mechanical Turk (a service pairing workers with tasks over the internet). This population includes a range of ages and socio-economic backgrounds that may be more representative of the population at large (Mason \& Suri, 2012). In addition, this experiment serves to add to the growing body of linguistic and cognitive research using Mechanical Turk.

\section*{Participants}

Participants were 24 native English-speaking workers recruited through Amazon Mechanical Turk. They received \(\$ 1.00\) for their participation in the study.

\section*{Materials}

The materials were the same as those used in Experiment 1, and participants were again randomly assigned to either the shape condition or the random condition.

\section*{Results}

The results of Experiment 2 replicate the major findings of Experiment 1, as shown in Figure 3. Analysis of this data reveals a significant effect of condition \((\beta=0.91, z=3.77, p<\) 0.01 ), with participants in the shape condition choosing the correct classifier on seen items much more often than those in the random condition ( 0.82 vs. 0.55 ). A significant interaction between condition and number was also found ( \(\beta=0.17, z=2.08, p<0.05\) ), indicating that the participants in the random condition tended to be less accurate on items with the number "one" compared to "two". This interaction is in the opposite direction as what was found in Experiment 1 , suggesting that the effect of number may not be reliable.

In terms of generalization to novel items, participants in the shape condition again show a relatively modest but significant increase in accuracy of classifier choice for seen items in comparison to novel items ( \(\beta=0.38, z=2.08, p<0.05\) ). No other significant effects were observed, again suggesting that differences in performance driven by number in Experiment 1 may not be reliable.

As in Experiment 1, for each pair of participants in a given condition we computed the proportion of novel test items that were assigned to the same category. Average agreement was above chance for the shape condition \((0.65, \mathrm{SE}=0.04)\), but note that this represents a lower level of agreement than that found in Experiment 1 for the same condition. Just as in Experiment 1 , average agreement for the random condition was at the expected chance level \((0.50, \mathrm{SE}=0.02)\).

\section*{Discussion}

In the experiments reported above, we exposed adult Englishspeakers to a miniature artificial noun classification system.


Figure 3: Correct choice of classifier for seen and novel nouns in the shape condition, and seen items in the random condition in Experiment 2 (NB: there is no correct choice for novel nouns in the random condition).

In order to investigate the role of semantic features of nouns in the acquisition of classification systems, we used English words, removing an obstacle present in natural language learning. Child language learners likely go through a stage of development in which phonological but not semantic information is available for the acquisition of noun classification and other grammatical features. The results of our experiments indicate that, when exposed to a realistic classification system (based on two Cantonese sortal classifiers) over known nouns, participants are able to learn the correct categories based on semantic information alone, and can readily generalize this information to new nouns. Learning did not extend to participants exposed to randomly generated noun categories which lacked supporting semantic cues. Our findings were robust in both a population of college students, and among the more diverse population found on Amazon Mechanical Turk-despite a relatively small sample size.

This finding suggests that semantic features of nouns can be quickly used by learners as the basis of a classification system, calling into question the apparently privileged role of phonology cues argued to hold in previous work on this topic (Karmiloff-Smith, 1981; Perez-Pereira, 1991; Tsang \& Chambers, 2011; Gagliardi, 2012). While here we have shown that semantically based noun classification can be learned in the absence of phonological cues, in future work we will ask whether phonological information is nevertheless used preferentially over semantic information when both are simultaneously accessible.

We believe our results are also relevant to understanding the initial stages of category formation. In particular, the dramatic difference in performance for seen items-items which were part of a participant's exposure set-between the two conditions calls into question theories of learning in which
categories are formed by abstraction over a set of stored exemplars (Nosofsky, 1986) (see also (Rouder \& Ratcliff, 2004) for relevant discussion and detailed model comparison). Under such a view, the prediction would be that learners should store the set of exemplars presented during familiarization regardless of whether the particular classifier-noun pairings are random or semantically conditioned. It would then remain unexplained why participants in the random condition fail to use the stored pairings to perform with high accuracy on seen items at test. Our participants succeeded at remembering (or reconstructing) particular examples only when those conformed to a more abstract generalization across items.

\section*{Acknowledgments}

The authors would like to acknowledge Kelly Glazer and Allison Widen for assistance in constructing the experimental materials and running participants.

\section*{References}

Aikhenvald, A. (2000). Classifiers: A typology of noun categorization devices. New York, NY: Oxford University Press.
Braine, M. (1987). What is learned in acquiring word classes: A step toward an acquisition theory. In B. MacWhinney (Ed.), Mechanisms of language acquisition (pp. 65-87). Hillsdale, NJ: Lawrence Erlbaum Associates.
Brooks, P. J., Braine, M. D. S., Catalano, L., Brody, R. E., \& Sudhalter, V. (1993). Acquisition of gender-like noun subclasses in an artificial language: The contribution of phonological markers to learning. Journal of Memory and Language, 32(1), 76-95.
Comrie, B. (1989). Language universals and linguistic typology: Syntax and morphology. Chicago, IL: University of Chicago press.
Craig, C. (1986). Noun classes and categorization: proceedings of a symposium on categorization and noun classification, Eugene, Oregon, october 1983. Philadelphia, PA: John Benjamins.
Culbertson, J., Smolensky, P., \& Legendre, G. (2012). Learning biases predict a word order universal. Cognition, 122, 306-329.
Demuth, K., \& Ellis, D. (2008). Revisiting the acquisition of Sesotho noun class prefixes. Mahwah, NJ: Lawrence Erlbaum.
Denny, J. (1976). What are noun classifiers good for? In Papers from the 12th regional meeting, chicago linguistic society (pp. 122-132).
Dixon, R. (1986). Noun classes and noun classification in typological perspective. Noun classes and categorization, 105-112.
Finley, S., \& Badecker, W. (2010). Linguistic and nonlinguistic influences on learning biases for vowel harmony. In S. Ohlsson \& R. Catrambone (Eds.), Proceedings of the 32nd annual conference of the Cognitive Science Society (p. 706-711). Austin, TX: Cognitive Science Society.

Frigo, L., \& McDonald, J. (1998). Properties of phonological markers that affect the acquisition of gender-like subclasses. Journal of Memory and Language, 39(2), 218245.

Gagliardi, A. (2012). Input and intake in language acquisition. Unpublished doctoral dissertation, PhD dissertation, University of Maryland.
Karmiloff-Smith, A. (1981). A functional approach to child language: A study of determiners and reference (Vol. 24). New York, NY: Cambridge University Press.
Lakoff, G. (1987). Women, fire and dangerous things-what categories reveal about the mind. Chicago, IL: University of Chicago Press.
Li, C., \& Thompson, S. (1989). Mandarin Chinese: A functional reference grammar. Los Angeles, CA: University of California press.
Li, P., Huang, B., \& Hsiao, Y. (2010). Learning that classifiers count: Mandarin-speaking children's acquisition of sortal and mensural classifiers. Journal of East Asian Linguistics, 19(3), 207-230.
Mariscal, S. (2009). Early acquisition of gender agreement in the Spanish noun phrase: starting small. Journal of Child Language, 36(1), 143-171.
Mason, W., \& Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. Behavior Research Methods, 44(1), 1-23.
Mithun, M. (1986). The convergence of noun classification systems. Noun classes and categorization, 379-397.
Nosofsky, R. (1986). Attention, similarity, and the identification-categorization relationship. Journal of Experimental Psychology: General; Journal of Experimental Psychology: General, 115(1), 39.
Perez-Pereira, M. (1991). The acquisition of gender: What Spanish children tell us. Journal of Child Language, 18(03), 571-590.
Polinsky, M., \& Jackson, D. (1999). Noun classes: Language change and learning. In B. A. Fox, D. Jurafsky, \& L. Michaelis (Eds.), Cognition and function in language (pp. 29-50). Stanford, CA: CSLI.
Rouder, J., \& Ratcliff, R. (2004). Comparing categorization models. Journal of Experimental Psychology: General, 133(1), 63.
Senft, G. (2000). Systems of nominal classification. Cambridge: Cambridge University Press.
Supalla, T. (1986). The classifier system in American sign language. In C. Craig (Ed.), Noun classes and categorization (pp. 181-214). Philadelphia, PA: John Benjamins.
Tsang, C., \& Chambers, C. (2011). Appearances aren't everything: Shape classifiers and referential processing in Cantonese. Journal of Experimental Psychology: Learning, Memory, and Cognition, 37(5), 1065.
Tse, S., Li, H., \& Leung, S. (2007). The acquisition of Cantonese classifiers by preschool children in Hong Kong. Journal of child language, 34(3), 495.
Wilson, C. (2006). An experimental and computational study
of velar palatalization. Cognitive Science, 30(5), 945-982.

\title{
Moving from Levels \& Reduction to Dimensions \& Constraints
}

\author{
David Danks (ddanks@cmu.edu) \\ Department of Philosophy, 135 Baker Hall, Carnegie Mellon University Pittsburgh, PA 15213 USA
}

\begin{abstract}
Arguments, claims, and discussions about the "level of description" of a theory are ubiquitous in cognitive science. Such talk is typically expressed more precisely in terms of the granularity of the theory, or in terms of Marr's (1982) three levels (computational, algorithmic, and implementation). I argue that these ways of understanding levels of description are insufficient to capture the range of different types of theoretical commitments that one can have in cognitive science. When we understand these commitments as points in a multi-dimensional space, we find that we must also reconsider our understanding of intertheoretic relations. In particular, we should understand cognitive theories as constraining one another, rather than reducing to one another.
\end{abstract}

Keywords: Level of description; Marr; Philosophy of cognitive science; Reduction; Intertheoretic constraint

\section*{Limitations of Levels}

It is customary within science to talk about our theories as falling at different "levels of description": biology is at a higher level of description than chemistry, which is itself at a higher level than physics. Moreover, talk of levels is not restricted to the relationships between these large-scale domains of science; a sub-symbolic model of causal cognition can be said to be at a lower level of description than some symbolic model of the same cognition or behavior.
"Levels talk" is particularly widespread in the cognitive sciences (as noted by many authors, such as Bechtel, 1994; Bickle, 1998; Marr, 1982). The proliferation of talk about levels is quite unsurprising, given the many different methodologies used to develop theories of human behavior and cognition. At the same time, exactly what is meant by a "level" is often left somewhat vague. Levels of description are sometimes identified with the ontological granularity of a theory, where its level is determined (largely) by its objects. This characterization misses important distinctions, however, such as the difference between a rational analysis that says how one should act, and a process model that describes the cognitive mechanisms generating behavior.

One of the most precise characterization of levels in cognitive science-and certainly the most influential such characterization-was given by Marr (1982), and captured this key distinction. Marr's three levels characterize information-processing devices in general, and processes in the human mind more specifically. The computational level identifies the input and output of the process, as well as constraints on the types of computation done on the input to get the output. The algorithmic level (also called the representation level) specifies an implementation of the computational theory, as well as the representation of the
input and output of the process. Finally, the implementational level describes the physical realization of the representation and the algorithm.

Roughly speaking, the computational level specifies what problem is being (appropriately) solved; the algorithmic level explains how it is solved; and the implementational level gives the details of the physical substrate that does the solving. As a concrete (non-cognitive) example, we can understand a word-processing program as (i) a process for entering, editing, and rendering text documents (the computational level); (ii) a bunch of lines of code that produce the appropriate behavior (the algorithmic level); or (iii) changes of 1 's and 0 's in the internal memory registers of the computer (the implementational level).

As a more cognitive example, consider the problem of learning causal structure from observational data (e.g., Cheng, 1997; Griffiths \& Tenenbaum, 2005). A computational-level model of this problem would characterize the relevant inputs (case-by-case observations or a summary of a sequence of such cases), the output that should result given such input (a representation that can be used for causal inference, decision-making, explanation, etc.), and any relevant cognitive constraints (though in practice, computational-level models rarely incorporate such constraints). An algorithmic-level model would characterize the internal representations and cognitive processes by which we humans happen to solve this challenge. And an implementation-level model would show how the relevant computations are performed in particular brain regions (e.g., frontal cortex as suggested by Fletcher, et al., 2001 or Satpute, et al., 2005).

Marr's three levels were a significant advance in part because they are based on the recognition that the mathematical or computational specification of a cognitive theory significantly underdetermines the commitments that are implied by it. A Bayesian model of causal learning could, for example, be at the computational or algorithmic level, depending on the intended interpretation of the terms in the model. Moreover, these differences in interpretation (and so commitments) can matter: whether some experiment or behavioral measure is a test of a model depends in part on the commitments of that model.

Marr's levels were also intended to help show that there can be distinct models of the same phenomenon that are not competitors. That is, models \(M_{1}\) and \(M_{2}\) can be incompatible (whether mathematically or ontologically) and yet both be correct as long as they are at different levels. For example, Bayesian and associationist models of causal learning are mathematically incompatible-they posit different representations and different learning processes-but can
both be correct if one is at the computational level and the other is at the algorithmic level (Danks, Griffiths, \& Tenenbaum, 2003; Griffiths \& Tenenbaum, 2005).

Unfortunately, Marr's levels suffer from at least two significant flaws. First, and more importantly, they assume that multiple distinct aspects of theoretical commitment must vary together, rather than being able to vary independently (see also McClamrock, 1991). For example, suppose model \(M_{1}\) is a standard computational-level model of human causal learning: it characterizes the relevant inputs and shows which (behavioral) outputs would solve the causal learning task, all while being agnostic about the underlying representations and processes.

Now consider \(M_{2}\) that is mathematically identical to \(M_{1}\), but which claims only that people do generate this (behavioral) output, not that this behavior is how people should solve the causal learning task. That is, \(M_{2}\) is a relatively standard instrumentalist model that characterizes the human behavior without explaining precisely how or why it is generated. \(M_{2}\) is not a computational-level model, as it does not explain why people act as they do (i.e., one of the putative hallmarks of a computational-level model). At the same time, \(M_{2}\) is not an algorithmic-level model, as it does not characterize the underlying representations or cognitive processes. There thus does not appear to be any place to put \(M_{2}\) in the standard three Marr levels.

More generally, Marr's three levels force three different dimensions of variation in theoretical commitment-extent of realism, tightness of approximation, and (importance of) closeness to optimality (all discussed in the next section)to change in lockstep when they can, in practice, vary relatively independently. This observation points towards the second concern about Marr's levels: namely, each of these dimensions has many more than just three levels, as theories can differ (in their commitments) in relatively finegrained ways. Marr's levels are sometimes helpful for providing a quick characterization of the commitments of some theory, if the theory happens to fit one of those templates. But in general, we need a subtler characterization of the types of theoretical commitments we can have for a given cognitive model.

\section*{Dimensions of Variation in Commitments}

In this section, I consider in more detail these three dimensions of variation in one's theoretical commitments. At the end, I show how we can use these dimensions to better understand how Marr's levels force these different dimensions to vary together, though they should be independent in theory (though not always in practice).

\section*{Realist Commitments (or, What Does It Mean to Be a Cognitive Realist?)}

The first dimension is arguably the easiest to understand: the extent of realism about the theory is simply which parts of the theory are supposed to refer to representations or cognitive processes that "really exist" in a standard metaphysical sense. As a simple example, consider a
cognitive model of an individual being asked to add two plus two, and then responding with four. A completely minimal realist commitment for such a model would be to regard it instrumentally: one could commit only to the model offering a correct characterization of the input-output function for human addition. A substantially more realist commitment would claim that there are internal cognitive representations of the numbers ' 2 ' and ' 4 ', as well as some process by which the former representation (perhaps with a copy) is manipulated so as to yield the latter representation. This interpretation presupposes that there is really a representation there (in a sense discussed below) and that there is some process corresponding to addition.

As we see in this example, simply giving the mathematical specification of a cognitive theory is insufficient to determine the realist commitments; those are, in an important sense, outside of the scope of the computational part of the model. At the same time, to fully understand how to interpret a cognitive model, one needs to know what realist commitments to attribute to it. Such specification rarely occurs explicitly for theories in cognitive science (or at least, rarely in journal papers), but is nonetheless an important step. Some information about realist commitments can be conveyed implicitly through the variables in the model, or by asserting that the theory holds at some level of description. "Levels" of description are, however, much too coarse to convey potentially finegrained metaphysical commitments, at least in the sense of stating what things there are held to be in the world.

This dimension of variation is still under-specified, as it is not yet clear which epistemological commitmentscommitments about what we could come to learn or knoware implied by attributing "reality" to cognitive representations or processes. We can usefully understand epistemological commitments in terms of the predictions they license, as prediction is at the core of many epistemic activities, including control, learning, inference, and even parts of explanation.

By looking at constraints on prediction, we see that there are two different types of realist commitments in the cognitive sciences-realism about processes, and about representations. A rough characterization of the distinction between representations and processes suffices for capturing realist commitments: representations are the relatively stable, persistent objects that encode information, and processes are dynamic operations involving those objects that can potentially (but need not) change the state of those objects. That is, representations are whatever encodes information stably over some reasonable timescale, and processes are whatever manipulate that information. This high-level characterization covers most of the standard accounts of cognitive representations and processes; even embodied (e.g., Barsalou, 2008) and dynamic systems (e.g., Port \& van Gelder, 1995) theories of representation (or its apparent absence) fit this general schema, if we focus on the structure of the theory rather than the language used to describe it.

Given this distinction, representation realism implies commitments about the stability of predictions for different types of cognition that use the information encoded in that representation. If the representation "really exists," then the same object is presumably used for (potentially) many purposes, and so predictions in these different contexts should reflect that shared informational basis. For example, realism about the concept 'DoG' implies that behavior in a categorization task involving dogs should be correlated (in various ways) with performance in a feature inference task involving dogs. More generally, representation realism licenses us to use behavior on one task to make predictions about (likely) behavior on different tasks that use the same representations, at least ceteris paribus. Importantly, realism about our cognitive representations does not imply that every one is available for every process; it is certainly possible that we have multiple representational stores, some of which are process-specific. But if the same representation is supposed to be available to multiple processes, then representation realism implies a set of epistemological commitments about correlations or stabilities between predictions about the behaviors that the different cognitive processes generate.

Process realism similarly implies epistemological commitments of inter-prediction correlations and stabilities, but rather for the same task given different inputs, backgrounds, or environmental conditions. That is, if one is committed to the reality of a given cognitive process, then that process should be stable and persistent in its functioning across a range of inputs and conditions. For example, realism about a particular process theory of concept learning implies that this particular process should be active for a variety of inputs that trigger concept learning. Whether I am learning about the concept 'DOG' or the concept 'CAT', the same process should be engaged (since that is the process that is "really there"). Of course, process realism does not imply that every process is triggered for every input or in every condition; rather, process realism is the more minimal claim that there should be correlations and stabilities between the predictions for the different performances of the same task, ceteris paribus.

Critically, the epistemological commitments of process realism and representation realism are separable, at least in the abstract. One could think that the appropriate predictive correlations obtain within a cognitive task but not between them (i.e., process realism without representation realism). For example, performance on a categorization task involving dogs might not imply anything stable for predictions about how people do causal inference about dogs. Alternately, the appropriate stabilities might obtain across tasks for the same information, but not within a task (i.e., representation realism without process realism). For example, there might be correlations between predictions for categorization and feature inference tasks involving dogs, but no stable correlations between the predictions for categorization involving dogs and cats.

One can make realist commitments about only some of the representations or processes in one's theory; process and representation realism are not all-or-nothing affairs. To take a concrete example, consider associative models of contingency (or causal) learning, such as the well-known Rescorla-Wagner model (Rescorla \& Wagner, 1972). At a high level, associative learning models posit that one learns contingencies or correlations (possibly including causal strengths) by updating associative strengths between various factors. Computationally, whenever one observes a new case, the cognitive agent (i) uses some of the observed factors to predict the state of other factors using the appropriate associative strengths, and then (ii) changes associative strengths based on the prediction error.

Most standard interpretations of associative learning models are realist about the associative strengths, but not about the predictions "generated" in step (i) in order to change strengths in step (ii). That is, the former representations "really exist" and are encoded somewhere, but the latter are just a computational device. Similarly, most are realist about the update process that changes the associative strengths, but not about the prediction process that uses some of the associative strengths to predict the states of other factors.

\section*{Degree of Approximation}

A second dimension of variation in the commitments of a cognitive theory is in the intended closeness (to reality) of the theory's approximations. All theories are approximate in some ways, in that they exclude certain factors or possibilities; there is no complete theory that incorporates everything. We can nonetheless distinguish (for a particular theory) different commitments about what is supposed to be captured by that theory. We can think about this dimension as tracking either which factors have been excluded, or the intended scope of the theory.

As a concrete example, suppose one has a model of human addition that predicts that people will respond '93' when asked "what is \(76+17\) ?" A question thus arises when someone responds (erroneously) '83': what does this behavior imply for the theory? One response is to hold that this represents a (partial) falsification of the model, as it made a prediction that was not borne out. A different response is to argue that the behavior is due to some factor that was not included in the model because it falls outside of the intended scope of the model (e.g., a momentary lapse of reason due to distraction). The mathematical or computational specification of a theory does not include what was (deliberately) omitted, but that information is important when deciding how to respond to an apparent mismatch between theory and reality.

This dimension is clearly related to the performance/ competence distinction, but it is also not identical with it. Roughly speaking, a competence theory aims to characterize what people are capable of doing, while a performance theory aims to describe what they actually do. Typically, the former is a theory that aims to explain and predict people's
ideal behavior if they did not face, for example, limits on memory and attention, cognitive processing errors, and other deleterious factors. The latter is supposed to be a theory that accounts for these various factors so as to capture (approximately) actual human behavior in all its messy glory. The mathematical specification of a theory does not entail that it is either a performance or competence theory, and some historical debates in the cognitive sciences occurred precisely because of a misunderstanding about whether (the mathematical specification of) a theory was intended as a competence or performance theory.

The performance vs. competence distinction can be understood as picking out two possible commitments along this dimension of variation (i.e., about the intended scope of a theory). But there are many other intended approximations that one could have in mind, including ones that arise from abstracting away from only some human cognitive limitations and peculiarities, rather than all of them (as in competence theories). The performance vs. competence theory distinction marks an important pair of possible intended commitments of a theory, but fails to capture the full range of possible commitments.

\section*{Importance of Optimality}

The third dimension of variation in a theory's intended commitments is in the putative or claimed optimality of the theory (if any): that is, is the theory additionally claimed to be optimal (or rational), and if so, for what task(s) and relative to what competitors? This additional claim is important because claims about optimality (help to) license so-called "why-explanations." We are often interested not just in how some behavior occurs (i.e., the underlying representations and processes that actually generate it), but also in why that behavior occurs.

Actually tracing the causal history (whether ontogenetic or phylogenetic) of a process or representation can be remarkably difficult, if not impossible. An alternative path to reach a why-explanation is to show that some cognition is optimal relative to competitors, and that there are sufficiently strong pressures on the individual (or lineage) to push the individual to the optimal cognition (and that those pressures actually obtained in these circumstances). If these elements can be shown, then we can conclude that the cognition occurs because it is optimal. This alternative path is a standard way to demonstrate, for example, that some physical trait constitutes an evolutionary adaptation (Rose \& Lauder, 1996).

In practice, many optimality-based "explanations" in the cognitive sciences fail to demonstrate all of the elements; in particular, they frequently fail to show that there are actual "selection pressures" that would suffice to drive an individual towards the optimal cognition, or even to maintain an individual at the optimal cognition. Nonetheless, the intended closeness to optimality (relative to a class of alternatives) of a theory-and so its ability to function in a possible why-explanation-is a critical theoretical commitment about a model that is not implied
simply by its mathematical/computational specification. And clearly, variation in this dimension induces different metaphysical and epistemological commitments, as claims that some theory is optimal imply facts about the causal history of the cognition, and about how the cognition should plausibly change under variations in the environment or learning history.

\section*{Connecting the Dimensions and Marr's Levels}

Marr's levels force these three dimensions of variation to change together, rather than allowing them to vary independently. For example, a theory at the computational level is understood to have a relatively weak set of realist commitments (particularly about processes), significant approximation (since the theory is about how the system should solve a problem, rather than what it actually does), and a fairly strong expectation of optimality. Theories at the implementational level, in contrast, are strongly realist (since they hopefully focus on the underlying biological mechanisms), aim to minimize approximation by incorporating relatively contingent influences, and emphasize causal mechanisms ("how") rather than optimality ("why").

As a result, one must be careful about using Marr's levels to characterize a theory. Use of the terminology can force proponents of a theory into particular commitments that they would prefer to deny, as the levels bundle together commitments that should be kept separate. At the same time, anything that encourages more precise specification of the extra-computational commitments for a theory is a positive. The overall usefulness of Marr's levels principally depends on whether the theory's proponent happens to endorse one of the limited sets of possible commitments that can be expressed in that trichotomy. In many actual cases in cognitive science, however, we have subtler, more finegrained variations in our theoretical commitments.

\section*{From Reduction to Constraint}

Throughout this discussion, I have largely ignored one of the most important uses of levels, whether Marr or otherwise: namely, they provide a framework in which we can understand intertheoretic relationships. That is, we care not only about the commitments of a scientific theory, but also about the ways in which theories are related to one another, and "levels talk" provides an excellent way to understand such relations.

Of course, it is possible that there are no such (interesting) intertheoretic relations in cognitive science, as implied by various claims that psychology is "autonomous" (or other related term) from the underlying neuroscience (e.g., Fodor, 1974, 1997). Proponents of rational analyses often suggest a similar sort of disconnect, as they sometimes hold that the rational analysis says nothing about how the behavior is generated (e.g., Anderson, 1990). There are many theoretical concerns about the autonomy position (see, e.g., the long list in Bickle, 1998). In addition, it is arguably descriptively incorrect: cognitive scientists frequently attend
to the ways in which their theories matter for one another. Regardless of whether it is logically necessary that there be interesting intertheoretic relations, it certainly seems to be contingently true that there are such relations.

The more common way to think about intertheoretic relations in cognitive science is in terms of reduction: roughly, a theory \(H\) at a higher level must (eventually, somehow) "reduce" to a theory \(L\) at a lower level. More precisely, \(H\) reduces to \(L\) when the latter is a finer-grained version of (something approximately equivalent to) the former. There are many different ways of explicating "reduction" with more precision, whether in terms of syntactic equivalence (Nagel, 1961); semantic equivalence (Bickle, 1998); similar causal powers (Schaffner, 1967); replaceability (Churchland, 1985; Hooker, 1981a, 1981b); or even as implementation of a computer program (Danks, 2008). In all of these cases, there is a close connection, or at least sympathy, between talk of "levels" and the focus on reduction as the key intertheoretic relation.

At least two general concerns arise, however, for all of these accounts of "reduction." First, scientific practice (particularly in the cognitive sciences) often does not involve definite, positive, theoretical proposals to serve as the relata of the "reduction" relation. One might claim, for example, that two variables are associated, or that some functional relationship falls in some (perhaps large) family, or that some previously considered theoretical possibility is incorrect (but without any further information about which theoretical possibility actually is right). These different types of theoretical claims can all imply commitments at other levels even if there is no particular broad theory in which they fit (and so no appropriate relata for reduction).

Second, and more importantly, "reduction" is always understood as a between-level relation: \(H\) and \(L\) are theories at different levels about roughly similar phenomena. \({ }^{1}\) Intertheoretic relations arise, however, between theories that do not stand in this type of "hierarchical" arrangement. For example, theories of causal learning and reasoning (e.g., Cheng, 1997; Griffiths \& Tenenbaum, 2005) and theories of "causal" concepts (e.g., Rehder, 2003a, 2003b) investigate different phenomena, and so cannot possibly stand in a reductive relationship in either direction. Nonetheless, these types of theories clearly constrain one another; at the very least, they both depend on representations of causal structure, and so information about one theory can be informative about the other. The focus on "levels of description" or Marr's levels makes it easy to focus on the hierarchically structured theories, but they are not the only ones that constrain one another. Just as we needed a more sophisticated understanding of the dimensions of variation in theoretical commitment, we need a more general account of intertheoretic constraints.

\footnotetext{
\({ }^{1}\) We also sometimes speak of a more general theory "reducing" to a more specific one at the same level in particular conditions (e.g., general relativity reduces to Newtonian dynamics in the limit of \(\left.(v / c)^{2} \rightarrow 0\right)\). Nickles (1973) shows how to keep this type of reduction separate from the type I have been discussing.
}

\section*{Towards an Account of "Constraint"}

At a high level, one cognitive theory \(S\) constrains another theory \(T\) if the extent to which \(S\) has some theoretical virtue \(V\) (e.g., truth, predictive accuracy, explanatory power) is relevant for the extent to which \(T\) has the same theoretical virtue \(V\). More colloquially, \(S\) constrains \(T\) just when, if we care about \(T\) along some dimension, then we should also care about \(S\) along that same dimension (because \(S\) could be informative about \(T\) ). Suppose, for example, that \(T\) reduces to \(S\). Reductions clearly involve constraint in terms of truth: \(S\) and \(T\) plausibly have the same truth-value when \(T\) reduces to \(S\). At the same time, reductions arguably do not always involve constraint in terms of explanatory power: the explanatory powers of the two theories in a reduction can vary relatively independently. Thus, it is important to relativize each application of intertheoretic constraint to a particular theoretical virtue.

To see how a more general notion of "constraint" could be made precise, consider the theoretical virtue of truth. I propose (without argument) that: \(S\) truth-constrains \(T\) if and only if a change in belief in \(S\) from time \(t_{1}\) to time \(t_{2}\) would, for a fully-knowledgeable agent, rationally produce a change in belief in \(T\) from \(t_{1}\) to \(t_{2}\). Note that there is no assumption here that the change in belief in \(S\) is rational; rather, this account of 'constraint' essentially models it as a conditional: "if an individual's belief in \(S\) changes (for whatever reason), then belief in \(T\) should rationally change as well, assuming that she understands the implications of her beliefs."

This proposal clearly includes reduction as a special case constraint: if \(H\) reduces to \(L\) given conditions \(C\), then an increase in belief in \(L \& C\) (alternately, full acceptance of \(L \& C\) ) should rationally lead to an increase in belief in (or full acceptance of) \(H\). For example, if some psychological theory \(P\) reduces to some neuroscientific theory \(N\), then if we come to believe \(N\), then we should also (rationally) believe \(P\). Moreover, in some contexts, a reductive relation can also lead to a downward constraint: if we come to believe \(H\), then that can rule out certain \(L\) s (i.e., any that \(H\) cannot reduce to).

This account of truth-constraint applies much more broadly than just reduction. For example, causal learning theories and theories of causal concepts that use the same representational framework (e.g., causal Bayesian networks) can be understood as mutually supporting: each makes the other more probable. More generally, one regularly finds arguments in cognitive science that are based on converging evidence from disparate domains, measurement methods, or processes. In this model of truth-constraints, the theories in the different domains place symmetric constraints on one another: increases (or decreases) in belief in one theory should rationally lead to increases (or decreases) in belief in others that point in the same direction. That is, the broader intertheoretic relation of "constraint" enables us-in contrast to the more narrowly focused "reduction"-to explicate and justify one of the most common argumentative techniques in cognitive science.

\section*{Conclusions}

The core idea of this paper is that the commitments that we have about our cognitive theories extend far beyond their mathematical or computational specification. Instead, we must be clear about where we are located in a multidimensional space of theoretical commitments. Our degree of realist commitment, permissible degree of approximation, and intended degree of optimality all can vary relatively independently, though they are tightly coupled in the traditional Marr levels.

Moreover, we need a more fine-grained notion of intertheoretic relations to complement this more nuanced picture of theoretic commitments. Cognitive theories sometimes reduce to one another, but more commonly they inform one another only indirectly. I have suggested that a theory of intertheoretic constraints would be most appropriate, but have only sketched how such constraints might look in one particular case. Substantial work remains to be done to characterize the ways that theories can relate to one another, and then to show how these constraints can be used to guide actual practice in cognitive science.

\section*{Acknowledgments}

This work was partially supported by a James S. McDonnell Foundation Scholar Award. An early version of this paper was presented at the 2013 Marshall Weinberg Cognitive Science Symposium.

\section*{References}

Anderson, J. R. (1990). The adaptive character of thought. Hillsdale, NJ: Erlbaum.
Barsalou, L. W. (2008). Grounded cognition. Annual Review of Psychology, 59, 617-645.
Bechtel, W. (1994). Levels of description and explanation in cognitive science. Minds \& Machines, 4, 1-25.
Bickle, J. (1998). Psychoneural reduction: The new wave. Cambridge, MA: The MIT Press.
Cheng, P. W. (1997). From covariation to causation: A causal power theory. Psychological Review, 104, 367405.

Churchland, P. M. (1985). Reduction, qualia, and the direct introspection of brain states. Journal of Philosophy, 82, 122.

Danks, D. (2008). Rational analyses, instrumentalism, and implementations. In N. Chater \& M. Oaksford (Eds.), The probabilistic mind: Prospects for Bayesian cognitive science. Oxford: Oxford University Press.
Danks, D., Griffiths, T. L., \& Tenenbaum, J. B. (2003). Dynamical causal learning. In S. Becker, S. Thrun, \& K. Obermayer (Eds.), Advances in neural information processing systems 15 . Cambridge, MA: MIT Press.
Fletcher, P. C., Anderson, J. M., Shanks, D. R., Honey, R. A. E., Carpenter, T. A., Donovan, T., Papadakis, N. \& Bullmore, E. T. (2001). Responses of human frontal cortex to surprising events are predicted by formal
associative learning theory. Nature Neuroscience, 4, 1043-1048.
Fodor, J. A. (1974). Special sciences: Or the disunity of science as a working hypothesis. Synthese, 28, 97-115.
Fodor, J. A. (1997). Special sciences: Still autonomous after all these years. Nous, 31, 149-163.
Griffiths, T. L., \& Tenenbaum, J. B. (2005). Structure and strength in causal induction. Cognitive Psychology, 51, 334-384.
Hooker, C. A. (1981a). Towards a general theory of reduction, part I: Historical and scientific setting. Dialogue, 20, 38-59.
Hooker, C. A. (1981b). Towards a general theory of reduction, part II: Identity in reduction. Dialogue, 20, 201-236.
Marr, D. (1982). Vision. San Francisco: W.H. Freeman.
McClamrock, R. (1991). Marr's three levels: A reevaluation. Minds \& Machines, 1, 185-196.
Nagel, E. (1961). The structure of science: Problems in the logic of scientific explanation. New York: Harcourt.
Nickles, T. (1973). Two concepts of intertheoretic reduction. The Journal of Philosophy, 70, 181-201.
Port, R. F., \& van Gelder, T. (1995). Mind as motion: Explorations in the dynamics of cognition. Cambridge, MA: The MIT Press.
Rehder, B. (2003a). A causal-model theory of conceptual representation and categorization. Journal of Experimental Psychology: Learning, Memory, and Cognition, 29, 1141-1159.
Rehder, B. (2003b). Categorization as causal reasoning. Cognitive Science, 27, 709-748.
Rescorla, R. A., \& Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black \& W. F. Prokasy (Eds.), Classical conditioning II: Current research and theory. New York: Appleton-CenturyCrofts.
Rose, M. R., \& Lauder, G. V. (Eds.). (1996). Adaptation. San Diego: Academic Press.
Satpute, A. B., Fenker, D. B., Waldmann, M. R., Tabibnia, G., Holyoak, K. J. \& Lieberman, M. D. (2005). An fMRI study of causal judgments. European Journal of Neuroscience, 22, 1233-1238.
Schaffner, K. (1967). Approaches to reduction. Philosophy of Science, 34, 137-147.

\title{
Proactive and Retroactive Interference Effects in Development
}

\author{
Kevin Darby (darby.60@osu.edu) \\ Department of Psychology, 1835 Neil Ave. Columbus, OH 43210 USA
}

\author{
Vladimir M. Sloutsky (sloutsky.1@osu.edu) Department of Psychology, 1835 Neil Ave. Columbus, OH 43210 USA
}

\begin{abstract}
Infants and children are avid learners. This constant aggregation of new knowledge, however, can interfere with past and future learning. Proactive interference (PI) occurs when past learning interferes with new learning, while retroactive interference (RI) is the attenuation of memory for previous learning as a result of new knowledge. Previous work has demonstrated that adults and children display PI and RI effects, but the developmental trajectories of these effects are less clear. The current study developed a new associative learning paradigm to concurrently test PI and RI in preschoolers and adults. Results demonstrated the presence of RI, and these effects were stable across age groups, suggesting that the mechanisms that modulate RI effects may already be mature in these age groups. No PI effects were found in either group, however. This surprising result suggests the role of associative complexity as a possible modulator of PI in these age groups.
\end{abstract}

Keywords: Learning; memory development; proactive interference; retroactive interference.

\section*{Interference effects}

Infants and children are avid learners: they constantly acquire new knowledge. This new knowledge not only expands their sense of the world, but also affects what they already know and what they will learn in the future (Wixted, 2004). Some of these effects are counterintuitive: (1) acquired knowledge may interfere with future learning, the process known as proactive interference (PI), and (2) acquired knowledge may attenuate memory for previously learned information, the process known as retroactive interference (RI). PI and RI effects are particularly important to study in early development because doing so will help determine what factors benefit or detract from the aggregation of early knowledge.

These sources of forgetting may play a role in many early cognitive domains, such as categorization (Mareschal, Quinn, \& French, 2002) and word learning (Levy-Gigi \& Vakil, 2010). Imagine, for example, that a child with bilingual parents learns the word "cat," but is later introduced to the word "gato." Mapping "gato" onto the child's category of cats may be more difficult than learning an entirely new concept in Spanish since the category is already associated with "cat" (PI). Additionally, the mapping between the word "cat" and the category of cats will likely be weakened as a result of learning to associate the category with a second word (RI).

Interference effects have been the focus of a great deal of research. It is clear, for example, that interference occurs in many different learning systems, including connectionist networks (French, 1999; Ratcliff, 1990) and human adults (Bower, Thompson-Schill, \& Tulving, 1994; Wixted, 2004). In adults, RI effects may be modulated by similarity between learning sets as well as mental effort, such that more interference is demonstrated with greater similarity and increased cognitive load (Dewar, Cowan, \& Della Sala, 2007; French, 1999; Wixted, 2004). Additionally, RI seems to be modulated by the engagement of networks in the hippocampal region and surrounding cortices (McClelland, McNaughton, \& O'Reilly, 1995; Wiskott, Rasch, \& Kempermann, 2006). Conversely, PI effects seem to be modulated by executive functions such as attentional control and inhibition of prepotent responses (Baker, Friedman, \& Leslie, 2010; Dick, 2012; Kiesel et al., 2010), and appear to be attenuated by activity in prefrontal regions of the cortex (Badre \& Wagner, 2005).

\section*{Interference in development}

Although the majority of research concerning PI and RI has focused on adults, some evidence suggests that interference effects may also be present early in human development. For example, infants demonstrate RI in a visual recognition task (Turati, 2008) as well as a mobile reinforcement paradigm (Rossi-George \& Rovee-Collier, 1999), and demonstrate PI in visual facial recognition (Tyrrell, Snowman, Beier, \& Blanck, 1990).

Despite the fact that interference occurs across development, the development of the ability to resist each kind of interference is less clear. There is some evidence that RI effects are relatively stable between preschool and school years. Howe (1995) demonstrated that RI effects were similar in preschoolers (approximately 4.5 years) and kindergarteners (approximately 6 years old) in a pairedassociate recall task. Similar findings were reported in 4and 7 -year-olds, using a game-based paradigm (Lee \& Bussey, 2001). It is unclear, however, whether there are developmental differences in RI if a wider age range is considered. In contrast, developmental differences in PI have been reported. Kail (2002) performed a meta-analysis on PI effects in children ages 4-13 years old, as well as an experiment with children in grades 3-6 and undergraduate adults. Both the meta-analysis and experimental results
indicated a decrease in PI effects across these developmental time scales.

The current study was conducted to investigate any differences in PI and RI between preschoolers (5-year-olds) and adults. To do so, we developed a new associative learning paradigm that would be appropriate to measure interference effects in both children and adults (Experiment 1) as well as provide a control for memory decay when specifically measuring RI (Experiment 2). This paradigm has the advantage of testing for both types of interference in a manner that is appropriate for children and adults.

\section*{Experiment 1}

To examine developmental differences in PI and RI, we developed a new associative learning task that allows us to study both types of interference within a single paradigm. In this task pairs of objects were associated with an outcome in three phases. In the first phase, participants learned to predict outcomes based the identities of paired objects; in the second phase, objects were re-paired to stimulate new learning, while in the final phase participants were presented again with the original pairs.

We expected to finds both types of interference in children, whereas the extent to which these effects are present in adults was less clear. Previous research suggests that RI effects are present in adults, to the extent that the learned material is sufficiently similar and cognitively challenging across phases (Dewar et al., 2007; French, 1999; Wixted, 2004). Also, given that cognitive control abilities are substantially more advanced in adults and given that PI effects depend on cognitive control (Baker et al., 2010; Dick, 2012; Kiesel et al., 2010), we expected that PI effects, if found, should be greater in children than in adults.

\section*{Method}

Participants Twenty-six undergraduates at The Ohio State University ( 20 females) and 34 children ( \(m=5.2\) years, SD \(=0.23\) years, 14 females) from the surrounding Columbus community participated in this experiment. Children were tested at local preschools. Adults received course credit and children received stickers for their participation.
Six children did not complete the task due to fatigue ( \(\mathrm{n}=5\) ) or computer error ( \(\mathrm{n}=1\) ). The data from these children were removed from all analyses. Additionally, since the focus of this study was on interference between new and previous learning, we required that participants demonstrate accuracy greater than \(70 \%\) in the initial learning phase of the experiment to be included in the analysis. In this way, we only included participants who demonstrated learning that could induce PI or be subject to RI. This learning criterion resulted in the removal of three adults and ten children. Our final experimental sample, therefore, consisted of 23 adults ( 17 females) and 18 children ( \(\mathrm{m}=5.3\) years, \(\mathrm{SD}=0.27\) years, 8 females).

Stimuli Experimental stimuli consisted of eight objects with common shapes and colors (e.g. blue circle). Each trial
consisted of the presentation of a pair of objects and a visual occluder that resembled a pipe splitting into two ends (see Figure 1). This occluder design was implemented such that an object disappearing behind the occluder could reappear on either side. Crucially, the outcome of the trial (i.e. where the object reappears) depended on the identities of the object pairings.
The object pairings, color of the visual occluder, and color of the background varied by phase: In the first phase, four object pairs were presented along with a white occluder on a dark grey background. In the second phase, objects were repaired and presented with a black occluder on a light grey background. Stimuli in the third phase were identical to those presented in the first phase. The purpose of varying the object pairings was to create interference between learning sets, while contextual information was varied so that new learning would not be too difficult to encode.


Figure 1: An example trial in phase 1
Procedure The task was computer-based, and stimuli were presented using E-Prime. To encourage interest in the task, children were tested using a touch-screen computer (adults were tested with a standard screen). In each trial two objects were presented with a visual occluder as described above. One object was situated directly above the second object, and the relative position of each object in the pair was counter-balanced across trials. The participant was told that one object would move into the occluder and come out on one side, and was asked to predict on which side of the occluder the object would reappear. Responses were made using the left or right arrows on a keyboard (adults) or by touching the relevant area of the touch-screen (children). Immediately after a response was given, the bottommost object would rise and hit the topmost object, which would move directly into the occluder before reappearing on one side approximately one second later. In addition to seeing the outcome of the object movement, participants were given explicit feedback: adults heard a high or low tone corresponding to correct and incorrect responses, respectively, while children were given explicit verbal feedback by the experimenter (e.g. "That's right, it does go
to that side!") in addition to the tone. The side of the object's reappearance was predicted by the object pair. In this way, subjects were able to learn the contingency between object pairs and outcomes.

The identity of the object pairings depended on the phase of the experiment: phase 1 consisted of learning four pairs of objects (such that two pairs reappeared on the left side of the occluder and two on the right). The objects were repaired in phase 2 , such that new learning required subjects to create new associations with the same objects and potential outcomes. The third phase was identical to the first phase, except that order of stimulus presentation varied between phases. Table 1 illustrates the abstract structure of object and outcome pairings in this experiment. For phases \(1-3\), each letter represents an object, while the outcome indicates the side of the object's reappearance from the occluder. Note that each pair in phase 2 includes an object that was associated with the opposite outcome in phase 1.

As noted above, the visual context of these stimuli changed between the phases to facilitate learning and recognition of different learning outcomes in the different phases. Participants were not informed that the context would change between the phases, nor were they told that they would be learning new associations in phase 2 or that they would be relearning the associations from phase 1 in the third phase. Forty trials (10 per pair of objects) were presented to each participant per phase, for a total of 120 trials. Subjects were invited to take short breaks between phases.

Table 1: Abstract object and outcome structure for Experiment 1.
\begin{tabular}{llll}
\hline Phase 1 & Phase 2 & Phase 3 & Outcome \\
\hline A - B & A - F & A - B & Left \\
C - D & C - H & C - D & Left \\
E - F & E - B & E - F & Right \\
G - H & G - D & G - H & Right \\
\hline
\end{tabular}

\section*{Results}

The central question of interest was whether children and adults would demonstrate differing amounts of proactive and retroactive interference effects. To address this question, trials in each phase were divided into 5 blocks (8 trials per phase) to closely examine the learning trajectories of these groups (see Figure 2).

To measure PI we compared the beginning (i.e. first block) of phase 1 to the beginning of phase 2: a decrease in accuracy in the second block would indicate PI. To measure RI we compared the end (i.e. last block) of the first phase to the beginning of phase 3 : since the object pairs were identical in the two phases, a decrease in performance between these blocks would indicate RI. A series of analyses of variance (ANOVAs) and t-tests was used to statistically measure PI and RI.


Figure 2: Accuracy in Experiment 1 by block for each phase in children (top) and adults (bottom), with standard error bars.

Proactive Interference To test PI effects, an ANOVA with block as a within-subject factor and age as a betweensubject factor was performed on the proportion of accurate responses for the first block of the first and second phases (see Figure 3). There were no significant main effects of block or age, and no interaction between these factors ( \(p\) 's>.2). To more directly test PI effects we conducted paired-sample t-tests between the first blocks of phases 1 and 2 separately for children and adults. The difference between blocks was not significant for children, \(t(17)=1.17\), \(p>.2\), or adults, \(t(22)=.49, \mathrm{p}>.6\). These results suggest that PI was not a factor in this experiment in children or adults.

Retroactive Interference To test RI effects, an ANOVA with block as a within-subject factor and age as a betweensubject factor was conducted on accuracy scores in the last block of phase 1 and the first block of phase 3 for children and adults. A significant main effect of block, \(F(1,39)=20.53, p<.001\), indicated that accuracy decreased in
the beginning of phase 3 across age groups. There was also a main effect of age, \(F(1,39)=4.07, p=.05\), suggesting that overall accuracy in these blocks was higher in adults No interaction, however, was found between block and age, \(p>.3\), suggesting that the difference between blocks did not vary as a function of age. The strong main effect of block suggested that RI may be found in individual age groups. Indeed, separate paired-samples t-tests revealed a significant difference in accuracies between the end of phase 1 and the beginning of phase 3 for children, \(t(17)=2.94, p<.01\), \(d=0.69\), as well as adults, \(t(22)=3.55, p<.01, d=0.74\). These results suggest that RI did occur in both children and adults, and that interference did not differ between groups.

One possible explanation for these retroactive interference effects is that subjects simply forgot the relevant associations learned in phase 1 as a result of the time passed between phases 1 and 3 . If this is the case, then the information learned in phase 2 did not interfere with performance in phase 3 but merely served a placeholder for the passage of time. To determine if this was the case, a second experiment was performed to control for memory decay.


Figure 3: PI and RI effects in Experiments 1 and 2. PI effects were calculated as the difference in accuracy between block 1 of phase 1 and block 1 of phase 2; RI effects were calculated as the difference in block 5 of phase

1 and block 1 of phase 3 . Positive values indicate interference; negative values indicate facilitation. \({ }^{*} p<.05\)

\section*{Experiment 2}

One potential interpretation of the RI effects found in Experiment 1 is that participants did not experience interference from learning new associations in phase 2 but simply forgot the associations learned in the first phase due to memory decay. To determine if this was the case, Experiment 2 minimized new learning while retaining the same task structure in the second phase of the task. If the RI effects found in Experiment 1 were due to memory decay, then performance in the beginning of phase 3 should also be
attenuated in this experiment in the absence of new learning. If accuracy has not declined at the start of phase 3, however, we can be confident that the results of Experiment 1 were indeed due to interference and not decay.

\section*{Method}

Participants Twenty-six adult undergraduates (17 females) and 215 -year-old children ( \(\mathrm{m}=5.3\) years, \(\mathrm{SD}=0.21\) years, 13 females) participated. Three children did not complete the experiment due to fatigue ( \(\mathrm{n}=2\) ) or because they were unable to complete the task before the end of the preschool session ( \(n=1\) ). Using the same learning criterion described above, three adults and eight children were further removed from the analysis for failure to demonstrate sufficient learning in the first phase of the task. The final analysis, then, included 23 adults ( 14 females) and 10 children ( \(\mathrm{m}=\) 5.2 years, \(\mathrm{SD}=0.13\) years, 6 females).

Stimuli The stimuli presented in phases 1 and 3 were identical to those in Experiment 1. In the second phase, however, pairs of objects were replaced with horizontally oriented arrows pointing to the left or right side of the screen. This was done so that participants could easily predict the outcome of each trial based on the direction of the arrows. In this way, participants continued performing the same task but with minimal new learning. The occluder and background colors in phase 2 were the same as in Experiment 1.

Procedure The procedure was the same as in Experiment 1: participants were presented with two objects (phases 1 and 3 ) or two arrows (phase 2), and predicted on which side of a visual occluder an object or arrow would reappear.

\section*{Results}

The purpose of this experiment was to determine if the attenuation of performance in Experiment 1 could be explained by memory decay. As such, accuracies in the last block of phase 1 and the first block of phase 3 were compared, as in Experiment 1 (see Figure 4 for the learning curves of each phase for children and adults, and Figure 3 for the differences between the target blocks). An ANOVA with block as a within-subject factor and age as a betweensubject factor revealed a main effect of block that was approaching significance, \(F(1,31)=3.77, p=.06\). However, in contrast to Experiment 1 (where performance dropped in phase 3 compared to phase 1), in this experiment, performance actually improved in phase 3 . There was a significant main effect of age, \(F(1,31)=6.5, p<.05\), indicating that adults' accuracy was higher across blocks, as in Experiment 1. The interaction between block and age was approaching significance, \(F(1,31)=3.77, p=0.6\), possibly reflecting a greater improvement in children's accuracy in the beginning of phase 3 from the end of phase 1. Individual t-tests indicated that the difference between these blocks was not significant in children ( \(p>.1\) ), or adults \((p=1)\). These
findings suggest that simple forgetting cannot explain interference effects observed in Experiment 1.


Figure 4: Accuracy in Experiment 2 by block for each phase in children (top) and adults (bottom), with standard error bars.

\section*{General Discussion}

This study investigated proactive and retroactive interference effects in preschoolers and adults. Results indicated comparable levels of RI in children and adults, but demonstrated no PI in either group. These results support recent claims that RI seems to be a particularly potent source of forgetting in humans (Wixted, 2004).

Additionally, our findings replicate and extend previous demonstrations that RI seems to produce consistent levels of interference across age groups. Specifically, these results provide new evidence that RI effects are stable from the preschool years into adulthood. This consistency may be the result of the early development of the neural systems involved in modulating RI, specifically the hippocampal formation (McClelland et al., 1995; Wiskott et al., 2006). Recent work suggests that the hippocampus and
surrounding areas in the medial temporal lobe have functionally developed by the age of five years (Alvarado, 2000; Bauer, 2008), which is consistent with our findings of stable RI effects following this age.

The results of this study are inconsistent, however, with the previous literature suggesting the presence of PI in children (e.g. Baker et al., 2010) and adults (e.g. Kiesel et al., 2010), as well as its decline with development (Kail, 2002). Why was PI not a factor in this task? One possibility is that the structure of the learned associations was not conducive to this type of interference. Each object in the stimulus set appeared only in a single pair (which was associated with a single outcome), such that each object in a pair was perfectly predictive of a given trial's outcome (see Table 1). As such, it was not necessary to encode an association between the two objects in the pair. This simple structure may have reduced demands on executive function, which might not have been the case if more complex structures were presented. Recall that PI is typically linked to executive functions (Baker et al., 2010; Dick, 2012; Kiesel et al., 2010), which are sub-served by the prefrontal cortex (Badre \& Wagner, 2005). Also recall that the areas of the prefrontal cortex sub-serving executive function mature relatively late (Bunge \& Zelazo, 2006). Therefore, it is quite surprising that PI effects did not transpire in young children. Perhaps the task was too easy to yield such effects (although it was not too easy to yield RI effects).

More broadly, the study of interference effects in development can potentially shed light on a number of important developmental phenomena. Word learning, for example, may depend in large part on associations between sounds and referents (Smith, Jones, Yoshida, \& Colunga, 2003). Recent work has suggested that 12- and 14-month old infants raised in a monolingual or bilingual environment do not differ in their ability to learn simple word-object pairings (Byers-Heinlein, Fennell, \& Werker, 2013). An intriguing possibility is that language background may instead influence the ability to form more complex associations (e.g. between words, referents, and identity of the language).

Many questions await future research. For example, creating more complexity in the structure of associations, such that three-way bindings between object \({ }_{1}\), object \({ }_{2}\), and the outcome are necessary for learning, will help us test the hypothesis that PI is modulated by associative complexity, possibly through increased demands on executive function. Additionally, mapping interference effects in more (particularly younger) age groups will allow us to determine whether interference effects are subject to developmental change and the time scales at which such change occurs. Understanding the mechanisms and developmental time course of these effects will allow us to understand a potentially fundamental aspect of learning and memory and how these processes interact early in life.

\section*{Acknowledgements}

This research was supported by a National Science Foundation (NSF) grant (BCS-0720135) to Vladimir Sloutsky. We thank the children, parents, and preschools that participated in this research.

\section*{References}

Alvarado, M. C. (2000). Revisiting the maturation of medial temporal lobe memory functions in primates. Learning \& Memory, 7(5), 244-256.
Badre, D., \& Wagner, A. D. (2005). Frontal lobe mechanisms that resolve proactive interference. Cerebral Cortex, 15(12), 2003-2012.
Baker, S., Friedman, O., \& Leslie, A. (2010). The opposites task: Using general rules to test cognitive flexibility in preschoolers. Journal of Cognition and Development, 11(2), 240-254.
Bauer, P. J. (2008). Toward a neuro-developmental account of the development of declarative memory.
Developmental Psychobiology, 50(1), 19-31.
Bower, G. H., Thompson-Schill, S., \& Tulving, E. (1994). Reducing retroactive interference: an interference analysis. Journal of Experimental Psychology: Learning, Memory, and Cognition, 20(1), 51-66.
Bunge, S. A., \& Zelazo, P. D. (2006). A brain-based account of the development of rule use in childhood. Current Directions in Psychological Science, 15(3), 118121.

Byers-Heinlein, K., Fennell, C. T., \& Werker, J. F. (2013). The development of associative word learning in monolingual and bilingual infants. Bilingualism: Language and Cognition, 16, 198-205.
Dewar, M. T., Cowan, N., \& Della Sala, S. (2007). Forgetting due to retroactive interference: A fusion of early insights into everyday forgetting and recent research on anterograde amnesia. Cortex, 43(5), 616-634.
Dick, A. S. (2012). Sources of cognitive inflexibility in setshifting tasks: Insights into developmental theories from adult data. Journal of Cognition and Development, 13(1), 82-110.
French, R. (1999). Catastrophic forgetting in connectionist networks. Trends in Cognitive Sciences, 3(4), 128-135.
Howe, M. L. (1995). Interference effects in young children's long-term retention. Developmental Psychology, 31(4), 579-596.
Kail, R. (2002). Developmental change in proactive interference. Child Development, 73(6), 1703-1714.
Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M., \& Koch, I. (2010). Control and interference in task switching--a review. Psychological Bulletin, 136(5), 849-874.
Lee, K., \& Bussey, K. (2001). Children's susceptibility to retroactive interference: The effects of age and degree of learning. Journal of Experimental Child Psychology, 80(4), 372-391.
Levy-Gigi, E., \& Vakil, E. (2010). Developmental differences in the impact of contextual factors on
susceptibility to retroactive interference. Journal of Experimental Child Psychology, 105(1-2), 51-62.
Mareschal, D., Quinn, P. C., \& French, R. M. (2002). Asymmetric interference in 3- to 4-month-olds' sequential category learning. Cognitive Science, 26(3), 377-389.
McClelland, J. L., McNaughton, B. L., \& O’Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. Psychological Review, 102(3), 419-457.
Ratcliff, R. (1990). Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. Psychological Review, 97(2), 285-308.
Rossi-George, a, \& Rovee-Collier, C. (1999). Retroactive interference in 3-month-old infants. Developmental Psychobiology, 35(3), 167-177.
Smith, L. B., Jones, S. S., \& Landau, B. (1992). Count nouns, adjectives, and perceptual properties in children's novel word interpretations. Developmental Psychology, 28(2), 273-286.
Smith, L. B., Jones, S. S., Yoshida, H., \& Colunga, E. (2003). Whose DAM account? Attentional learning explains Booth and Waxman. Cognition, 87(3), 209-213.
Turati, C. (2008). Newborns' memory processes: A study on the effects of retroactive interference and repetition priming. Infancy, 13(6), 557-569.
Tyrrell, D. J., Snowman, L. G., Beier, J. A., \& Blanck, C. M. (1990). Proactive interference in infant recognition memory. Bulletin of the Psychonomic Society, 28(3), 188190.

Wiskott, L., Rasch, M. J., \& Kempermann, G. (2006). A functional hypothesis for adult hippocampal neurogenesis: avoidance of catastrophic interference in the dentate gyrus. Hippocampus, 16(3), 329-343.
Wixted, J. T. (2004). The psychology and neuroscience of forgetting. Annual Review of Psychology, 55, 235-269.

\title{
Self-Regulated Learning with Graphical Overviews: When Spatial Information Detracts from Learning
}

\author{
Sarah Davies (sarah.davies@utah.edu) \\ Kirsten R. Butcher (kirsten.butcher@utah.edu) \\ Corey Stevens (coreybstevens@gmail.com) \\ Department of Educational Psychology, 1705 Campus Center Drive, MBH 327 \\ Salt Lake City, UT 84112 USA
}

\begin{abstract}
Graphical overviews have been studied as a method to improve hypertext learning and digital search. Although previous studies have found learning benefits to graphical overviews of single hypertext, it is unclear if these benefits extend to online learning across multiple (independent) documents. Previous research also has found that graphical overviews facilitate domain focus during online search, but it has not been established whether these benefits are derived from the spatial organization of the graphic or its textual content. This research examined the impact of using graphical overviews organized either spatially (i.e., network view) or textually (i.e., outline view) during self-regulated online learning. Assessments focused on deep understanding of science concepts and the relationships between them. Results indicated that the outline view promoted deeper understanding of science concepts and fewer errors about the relationships between them. Implications are discussed for the design and implementation of instructional materials to support self-regulated learning.
\end{abstract}

Keywords: self-regulated learning; graphical representations; online learning; conceptual browsing; comprehension

\section*{Introduction}

As individual learning tasks increasingly are performed in online environments (Graham \& Metaxas, 2003), there is a strong need to understand how the format of different materials impacts successful self-regulated learning (Pintrich, 2000; Winne, 2001). Self-regulated learning refers to learning situations in which students themselves must organize and manage the learning task (Azevedo \& Cromley, 2004); it can be contrasted with learning in structured environments such as intelligent tutoring systems, where the computer system typically chooses the problems and decides when the student has reached mastery and is ready to move on to new materials (Anderson et al., 1995).

When students work with online learning materials - for example, hypertext documents - the learning task is inherently self-regulated by virtue of non-linear links that allow the learner to choose a unique path through the digital content. Research has found that students have great difficulty in self-regulating their learning with hypermedia, often utilizing ineffective strategies during self-regulated learning tasks (Azevedo et al., 2008). Other research has demonstrated the potential of organizational materials to facilitate more effective self-directed learning in online environments. For example, graphical overviews have been found to facilitate learning when presented before students
work with a hypertext document (Salmerón et al., 2009). However, it is unclear if graphical overviews will have similar facilitative effects in online environments with limited coherence between independent online resources (rather than within a single hypertext document).

There is some evidence that a graphical interface can facilitate learning with varied, independent online resources. Research studying the use of a graphically-organized interface for online browsing showed that it facilitated processing of domain information in a digital library environment when compared to a keyword search interface (Butcher, Bhushan, \& Sumner, 2006). However, it remains unclear if results were driven by the spatial formatting of the graphical interface or its conceptual (textual) content.

This research investigates the effects of a graphical overview (presented as either a text-based outline view or a spatially-organized network view) on students’ selfregulated learning with online digital resources drawn from an educational digital library.

\section*{Self-Regulated Learning with Hypermedia}

When students are asked to self-regulate their learning from hypermedia, they often struggle to organize and process information in ways that support deep understanding (Azevedo et al., 2008). Although successful self-regulated learners engage in strategies such as planning and prior knowledge activation (Azevedo, Guthrie, \& Seibert, 2004), students engaged in self-regulated learning with hypermedia frequently choose to prioritize their reading based upon personal interest or text location (Salmerón, Kintsch, \& Cañas, 2006). Not surprisingly, this failure to attend to conceptual relationships and coherence in the domain can lead students to miss important semantic connections between ideas and to form a more shallow understanding of hypermedia content (Salmerón et al., 2006).

Students may need significant help - especially in activating prior knowledge, organizing knowledge, and processing conceptual relationships - in order to learn effectively with online content. One way to offer this support is to provide the student with useful organizational materials that can be used to guide study and learning. Graphical overviews, which illustrate high-level ideas and the relations between them for a given text or topic, provide one form of organizational materials that has been shown to support learning among students with low prior knowledge (Salmerón et al., 2009).

\section*{Graphical Overviews and Hypertext Learning}

Salmerón and colleagues have examined the impact of graphical overviews as the method of navigation through a hypertext (Salmerón et al., 2009; Salmerón \& Garcia, 2012). These graphical organizations provided students with freedom to choose navigational paths through the hypertext but organized the content that could be viewed across the hypertext using a conceptual overview of the content. Salmerón and Garcia (2012) found that providing young (sixth grade) learners with a graphical overview of a hypertext document improved their knowledge integration during a comprehension task. These results complement earlier findings which showed that providing students with a graphical overview before hypertext study led to increases in comprehension for undergraduate learners (Salmerón, et al., 2009). Salmerón and colleagues have proposed two potential explanations for the observed benefits of graphical overviews: first, graphical overviews may facilitate learning by providing an organizational framework to support online study; second, graphical overviews may facilitate active processing of difficult texts by providing a text macrostructure that frees up additional resources for comprehension processes.

If graphical overviews facilitate learning by providing learners with an organizational framework for domain knowledge, studying their effects within a single hypertext may underestimate their potential benefits. Whereas a single hypertext likely has an overall coherence and topical focus, self-regulated learning in more authentic online environments requires working across independent digital resources that may not be easily integrated. Thus, it is important to consider whether graphical overviews may facilitate learning when students work with multiple online resources (i.e., independent web pages and sites).

\section*{Graphical Overviews and Digital Search}

There is some evidence that graphical overviews change learners' processing when engaged in learning tasks that require work with multiple online resources. Butcher, Bhushan, and Sumner (2006) studied the impact of graphical overviews on students' search and evaluation processes as they attempted to locate useful online resources in an educational digital library. Students used either a graphical representation (a domain overview in the form of a node-link diagram) or a keyword interface to search for relevant digital content. Results showed that using the graphical representation as a search interface increased the depth of domain-relevant processing. Whereas students who navigated digital resources using a keyword interface tended to focus on superficial features of the resources, students navigating the resources with the graphical interface focused on analyzing domain concepts. Changes in the depth of students' processing of digital resources does not provide direct evidence of deeper learning with these resources; however, novice learners engaged in educational search tasks likely are engaged in "search to learn" processes which include iterative rounds of cognitive processing and
interpretation (Marchionini, 2006). Recent research (Butcher et al., 2011) has confirmed the impact of graphical overviews on digital search and evaluation: when graphical representations were used as the basis for preservice teachers' navigation of resources in an educational digital library, students were more likely to identify educationallyuseful online content and to focus on domain-level content when evaluating a web page or site.

\section*{Format and Content of Graphical Overviews}

Although Butcher and colleagues (Butcher et al., 2006; Butcher et al., 2011) have found clear evidence that graphical representations can impact the processes that students use during online search and the overall success of online searches during educational tasks, it remains unclear whether these observed benefits were derived from the spatial format of the graphic (i.e., the spatial organization of the graphical overviews) or its (textual) domain content. Because keyword interfaces may require significant cognitive effort to generate relevant search terms (Marchionini \& White, 2007), it is possible that the benefits of graphical overviews for self-regulated, online learning tasks may be derived from reallocation of cognitive effort from keyword generation to concept analysis. If this were the case, we would expect that removing spatial organization could facilitate even greater benefits by removing processing difficulty associated with examining and understanding spatial information.

If it is largely the textual content of graphical overviews that facilitates learning, more complex spatial formats actually may hurt novice learners. Graphical overviews in the form of a network map (see Figure 1) may depict interrelationships that are too complex for novice learners to understand. Novice learners may be better served by formats that emphasize organizational information in a hierarchical (i.e., linear) manner (see Figure 2). In a comparison of learning from linear and non-linear conceptual overviews, Amadieu and colleagues (2009) found that domain novices reported increased disorientation when learning from a network conceptual overview that depicted important relationships. In contrast, learners reported less disorientation and achieved better recall when learning with a hierarchical conceptual overview. Still, if it is true that graphical overviews promote learning by providing a conceptual framework for domain content, we would expect that a hierarchical graphical overview that removes spatial information would cease to be effective.

The current research extends prior research by examining two forms of graphical overviews during an online learning task: a spatially-organized network view vs. a textuallyorganized (linear) outline view. The use of these two conditions facilitates a direct comparison of whether the spatial format or the domain content of the graphical overviews has the greatest impact on learning outcomes. In addition, this research examines impact within a more authentic online environment, using the graphical overview to facilitate learning across a variety of independent online
resources. Because the network view is designed to demonstrate key conceptual relationships between multiple learning goals, we hypothesized that this graphical overview would facilitate greatest understanding of domain relationships.

\section*{Factual Knowledge vs. Deeper Understanding}

When considering learning outcomes, it is important to recognize that comprehension research has established that different levels of knowledge can be formed during learning (Kintsch, 1998). In this work, we draw upon a well-known, established model of comprehension - ConstructionIntegration (CI) - that distinguishes between three levels of knowledge representation: the surface level, the textbase, and the situation model (Kintsch, 1994). A surface level representation is formed by encoding the specific details of a text (e.g., exact words and sentences). A textbase representation consists of the semantic meaning of a text; thus, a textbase representation drives recall of basic ideas derived from learning materials. The most flexible and durable knowledge representation is the situation model, which is formed when the learner integrates to-be-learned content with prior knowledge. A well-developed situation model drives inference, application, and transfer; as such, students who develop the situation model can be considered to understand materials rather than simply remember them.

The outcome assessments in this research target knowledge at the textbase and situation model levels. As described below, textbase assessments focus on factual knowledge learned during study and recalled during testing. Situation model assessments focus on students' application of learned knowledge, through explanation of concepts and relationships. Errors in student explanations, which may result from superficial reasoning about perceived relationships, also are examined.

\section*{Method}

\section*{Participants}

Twenty-six undergraduate students (8 males, 18 females, \(M\) age \(=23\) ) at a large public university in the western United States participated in this study in partial fulfillment of a class research requirement. One participant was excluded because his major was geology.

\section*{Design}

This study utilized a two-condition, between-subjects experimental design. Participants were randomly assigned to one of the two experimental conditions upon arrival to the study.

\section*{Materials}

Graphical Overviews The graphical overviews in this study were drawn from the Science Literacy Maps published on the National Science Digital Library (NSDL) website. NSDL is a digital library which seeks to provide
access to up-to-date, high-quality, online resources in varied formats that will support education and learning in science, technology, engineering, and mathematics (Zia, 2000). The NSDL Science Literacy Maps are derived from strand maps developed by the American Association for the Advancement of Science (AAAS, 2001); these maps take the form of node-link diagrams. Nodes contain text that describe key learning goals in a topic area. The spatial organization of the nodes and the links between them demonstrate how student knowledge (as evidenced by the learning goals) should progress over time in a given domain.

In the NSDL, the Science Literacy Maps serve as a conceptual browsing interface (Zia, 2000); that is, the maps serve as a graphical search interface. To retrieve relevant digital resources using a conceptual search interface, users select a specific learning goal from the graphical overview (i.e., the Science Literacy Map). Clicking a learning goal brings up a small window that lists the NSDL-catalogued resources relevant to the conceptual information contained in the learning goal; much like a commercial search interface, each listed result provides users a title, a linked URL, and a short description of the resource.

Network Graphical Overview. The network view of the search interface utilizes the standard form of the Science Literacy Maps as found on NSDL.org. Learning goals are represented as nodes and are connected to one another with arrow links (see Figure 1); links between nodes indicate conceptual relationships between the learning goals. The overall spatial organization of the network indicates a more global knowledge organization, showing how learning goals develop over time, across grade levels and subtopics in the domain (see Figure 1).

Outline Graphical Overview. The outline view of the search interface contains the same node content as the network view. That is, all nodes contain the same text describing the same learning goals. However, in this view, the learning goal nodes are listed vertically rather than spatially. Learning goals in the outline view still are grouped by grade level (see Figure 2), but there are no links indicating conceptual relationships and spatial organization has been removed. As in the network view, clicking a learning goal in the outline view will bring up a window showing relevant resources catalogued in the digital library (see Figure 2). The learning goals in the outline view retrieved the same digital resources as in the network view (i.e., both interfaces searched over the same collection of digital resources and used the same algorithms to retrieve content relevant to each learning goal).

Reference Versions of Network and Outline Views Before students used the graphical overview as a search interface to find online digital resources, they were given ten minutes to familiarize themselves with a non-interactive version of the graphic. The non-interactive forms of the graphical overviews utilized the same formatting and content as the interactive (search interface) versions of the graphical overviews as described above (see Figures \(1 \& 2\) ).


Figure 1: The network conceptual search interface is on the left. On the right is its associated noninteractive reference.


Figure 2: The outline conceptual search interface is on the left. On the right is its associated non-interactive reference.

Learning Assessments Learning assessments were administered at the beginning and end of each session. Questions tested participants' factual knowledge of plate tectonics, as well as their understanding of important plate tectonics concepts and relationships between them

Factual Knowledge. Factual knowledge items were designed to capture participants' textbase-level knowledge of plate tectonics. Factual items consisted of generative as well as non-generative (multiple choice and true/false) questions. Generative questions provided participants with images, such as a cross-section of the Earth, and asked them to generate labels for specific components or processes. Participants were asked to generate 13 diagram labels; correct labels received one point and partially-correct labels received half a point, for a total of 13 points. Nongenerative questions tested students on their general knowledge (e.g., the number of Earth's tectonic plates). The non-generative factual assessment consisted of 33 items; participants received one point per correct item, for a total of 33 points.

Conceptual Understanding. Conceptual understanding items were designed to elicit participant explanations about key plate tectonics processes, thereby reflecting participants' situation models. These items asked students to interpret a diagram and explain the plate tectonics processes pictured. Conceptual understanding items were scored using a rubric that categorized explanations from most shallow to most deep, with a maximum of 5 points available per item. See Table 1 for examples of shallow, moderate, and deep answers. There were four conceptual understanding items, for a total of 20 points possible.

Table 1: Conceptual Explanation Examples
\begin{tabular}{|c|l|}
\hline Shallow & \begin{tabular}{l} 
It is showing the movement and direction \\
in which Earth is moving caused by heat.
\end{tabular} \\
\hline Moderate & \begin{tabular}{l} 
The arrows are drawn in a circular pattern \\
because that is how the convection heat \\
current travels beneath the surface.
\end{tabular} \\
\hline Deep & \begin{tabular}{l} 
The rock in the mantle is heated up and \\
due to its then lighter density rises to the \\
surface where it is cooled because it is \\
further away from the core and starts to \\
become more dense and sinks. This \\
process is repeated over and over again \\
and is called convection.
\end{tabular} \\
\hline
\end{tabular}

Relationship Explanations. These items were designed to assess the depth with which students understood conceptual relationships between the learning goals. Relationship explanation items provided students with two distinct learning goals from the graphical overview and asked them to explain the relationship between the learning goals. This assessment presented students with 3 pairs of learning goals at pretest and 6 pairs at posttest. Relationship explanation items were scored as shallow or deep (see Table 2 for examples). Because novice learners often fail to identify and understand important relationships during learning, and because the conditions differed in the explicit portrayal of these relationships, the accuracy of relationship explanations was also examined. Explanations containing incorrect reasoning or mechanisms were marked as containing errors.

Table 2: Relationship Explanation Examples
\begin{tabular}{|c|l|}
\hline Shallow & \begin{tabular}{l} 
They both talk about the movement of the \\
earth and what is causing the earth to move.
\end{tabular} \\
\hline Deep & \begin{tabular}{l} 
Because of heat flow and gravity, we see a \\
pattern of movement within the earth's \\
mantle (convection). The plates ride on the \\
mantle, so this movement translates into the \\
plates interacting with each other.
\end{tabular} \\
\hline
\end{tabular}

\section*{Procedure}

To begin the study session, participants completed a brief survey which gathered demographic information. Next, the pretests were administered to assess participants' prior knowledge of plate tectonics. The learning task included 10minute study of the reference version of the graphical overview (as appropriate to randomly assigned conditions), followed by forty minutes of learning with online digital resources as facilitated by the (condition-appropriate) graphical overview acting as the search interface. During online study, the reference version of the graphical overview was displayed on a second monitor so that participants could refer to it when reading/examining a digital resource.
Following the learning task, posttest assessments were administered.

\section*{Analysis}

As a check of random assignment, factual knowledge at pretest was analyzed using a MANOVA. Posttest learning assessment components also were analyzed using a repeated measures analysis of variance (RMANOVA) and a MANCOVA (see below). Alpha level was set at \(p=.05\) for all analyses.

\section*{Results}

\section*{Prior Knowledge of Plate Tectonics}

A MANOVA for performance on both types of factual knowledge items at pretest did not show an overall condition difference \(\left(F_{(2,22)}=2.50, p=.11\right)\); however, univariate tests indicated that the two conditions did differ significantly on pretest diagram labels \(\left(F_{(1,23)}=5.19, p=\right.\) .03). At pretest, the network overview condition correctly labeled a higher percentage of diagrams ( \(M=.29, S D=.19\) ) than the outline overview condition ( \(M=.15, S D=.10\) ). To control for the variance in learning due to prior knowledge, pretest performance in diagram labels was used as a covariate in a MANCOVA for posttest performance.

\section*{Posttest Performance on Learning Assessments}

Factual Knowledge A RMANOVA was used to examine pre- and posttest performance on the non-generative factual knowledge items. Overall, participants showed a slight but significant learning gain from pre- to posttest ( \(M_{\text {diff }}=.05\); \(\left.F_{(1,23)}=5.30, p=.03\right)\) but there was no significant effect of condition ( \(F<1\) ).

\section*{Conceptual Understanding \& Relationship Explanations} A MANCOVA was used to examine posttest performance on measures of deep comprehension. There was a significant main effect of condition ( \(F_{(3,20)}=4.32, p=.02\) ). Univariate tests showed a main effect of graphical overview condition on conceptual understanding (see Table 3).

Students in the outline graphical overview condition produced conceptual explanations that evidenced deeper understanding of plate tectonics concepts ( \(M=.38, S D=.19\) ) than the network graphical overview condition ( \(M=.34, S D\) \(\left.=.24 ; F_{(1,22)}=9.42, p<.01\right)\). There also was a significant

Table 3: M and (SD) for Assessments of Learning
\begin{tabular}{|lll|}
\hline Assessment Scores (\%) & Network & Outline \\
\hline Factual Knowledge & & \\
Non-generative (pretest) & \(.56(.09)\) & \(.53(.11)\) \\
Non-generative (posttest) & \(.62(.13)\) & \(.57(.06)\) \\
Conceptual Understanding* & \(.34(.24)\) & \(.38(.19)\) \\
& & \\
\begin{tabular}{c} 
Relationship Explanations \\
\% Deep Relationships \\
\% Conceptual Errors*
\end{tabular} & \(.17(.25)\) & \(.24(.25)\) \\
\hline
\end{tabular}
condition difference in the percentage of errors when explaining relationships between plate tectonics concepts \(\left.F_{(1,22)}=8.12, p<.01\right)\). Students in the outline condition generated a smaller percentage of errors ( \(M=.14, S D=.16\) ) than students in the network condition ( \(M=.27, S D=.17\) ).

A non-significant but note-worthy trend was found in the percentage of deep explanations of relationships between concepts \(\left(F_{(1,22)}=3.34, p=.08\right)\). The outline condition produced a higher percentage of deep relationship explanations ( \(M=.24, S D=.25\) ) than the network condition ( \(M=.17, S D=.25\) ).

\section*{Discussion}

After learning from multiple resources online, students in both conditions evidenced a similar increase in factual (textbase level) understanding of plate tectonics concepts. Overall, this is consistent with previous research finding that providing a graphical overview before hypertext study supports textbase comprehension (Salmerón et al., 2009). However, the current results also demonstrate that a spatially-organized graphical representation does not facilitate textbase learning more than a linearly-organized representation. Thus, it may be the textual content of the graphical organizer that facilitates macrostructure processing and leads to learning gains.

Although spatial format does not vary learning outcomes when considering factual (textbase-level) knowledge, it does impact the depth of understanding for important concepts and relationships between them. However, the pattern of results was opposite of hypothesized findings. Current results show that an outline graphical overview provided a learning advantage over a network (spatially-organized) overview: students learning with the outline view produced more deep explanations of science concepts and evidenced fewer erroneous ideas about inter-conceptual relationships.

This is a surprising result, since only the network view visually depicted the conceptual relationships among the learning goals. Indeed, previous studies have hypothesized that a schematic representation of relationships between concepts may provide novice learners with a framework for assimilating knowledge (Salmerón et al., 2009; Butcher et al., 2011). In this study, the spatial depiction of domain relationships compromised deep understanding. Concepts depicted in a network organization resulted in more errors when students explained conceptual relationships; students working with the network view also demonstrated less evidence of deep thinking about concepts. It may be that the graphic illustration of relationships actually precluded students from thinking deeply about the nature of those relationships. By explicitly depicting the conceptual relationships between learning goals, the network view may have caused students to generate fewer of their own inferences or predictions during learning. Alternatively, the network representation of content may have been too complex for novice learners. Previous research has found that students report feeling more disoriented with a network organization than with a more linear representation of
information (Amadieu et al., 2009). Because the network view did not specify the nature of potentially-complex relationships, students may have resorted to more shallow strategies of reasoning, integrating concepts based on superficial, easily-perceivable common features such as shared keywords.

When searching for information online, students typically learn from varied sources (Marchionini, 2006). Creating a deep, flexible understanding of the situation under investigation requires that self-regulated learners be able to synthesize multiple sources of information and integrate their learning with prior knowledge (Butcher \& Kintsch, 2012; Perfetti, Rouet, \& Britt, 1999). By demonstrating potential drawbacks to network-based graphical organizers during online learning, this study contributes an important initial finding to the literature on how to externally support self-regulated learning with multiple online resources. However, more research is needed to understand the specific relationship between the format of graphical overviews and their impact on learning outcomes.

\section*{Acknowledgments}

This material is based upon work supported, in part, by the
National Science Foundation under Grant DUE-0938041. Any opinions, findings, and recommendations or conclusions are those of the authors and do not necessarily reflect the views of the National Science Foundation.

\section*{References}

AAAS (2001). Atlas of Science Literacy, Volume 1. Washington, DC, Project 2061, Association for the Advancement of Science (AAAS) and the National Science Teachers Association.
Amadieu, F., Van Gog, T., Paas, F., Tricot, A., \& Mariné, C. (2009). Effects of prior knowledge and concept-map structure on disorientation, cognitive load, and learning. Learning and Instruction, 19, 376-386.
Anderson, J. R., Corbett, A. T., Koedinger, K. R., \& Pelletier, R. (1995). Cognitive tutors: Lessons learned. Journal of the Learning Sciences, 4(2), 167-207.
Azevedo, R., \& Cromley, J. G. (2004). Does training on self-regulated learning facilitate students' learning with hypermedia? Journal of Educational Psychology, 96, 523-535.
Azevedo, R., Guthrie, J. T., \& Seibert, D. (2004). The role of self-regulated learning in fostering students' conceptual understanding of complex systems with hypermedia. Journal of Educational Computing Research, 30, 87-111.
Azevedo, R., Moos, D. C., Greene, J. A., Winters, F. I., \& Cromley, J. G. (2008). Why is externally-facilitated regulated learning more effective than self-regulated learning with hypermedia? Educational Technology Research \& Development, 56, 45-72.
Butcher, K.R., Bhushan, S., \& Sumner, T. (2006). Multimedia displays for conceptual search processes: Information seeking with strand maps. ACM Multimedia Systems Journal, 11, 236-248.

Butcher, K.R., Davies, S., Crockett, A., Dewald, A., \& Zheng, R. (2011). Do graphical search interfaces support effective search for and evaluation of digital library resources? Proceedings of the Eleventh Annual International ACM/IEEE-CS Joint Conference on Digital Libraries (pp. 315-324). New York, Association for Computing Machinery.
Butcher, K. R., \& Kintsch, W. (2012). Text comprehension and discourse processing. In A.F. Healy \& R.W. Proctor (Eds.), Handbook of Psychology, Volume Four: Experimental Psychology (2nd Ed.). Hoboken, NJ: Wiley.
Graham, L., \& Metaxas, P. T. (2003). Of course it's true; I saw it on the internet! Communications of the ACM, 46(5), 71-75.
Kintsch, W. (1994). Text comprehension, memory, and learning. American Psychologist, 49(4), 294-303.
Kintsch, W. (1998). Comprehension: A paradigm for cognition. Cambridge: Cambridge University Press.
Marchionini, G. (2006). Exploratory search: From finding to understanding. Communications of the ACM, 49(4), 4146.

Marchionini, G., \& White, R. (2007). Find what you need, understand what you find. International Journal of Human-Computer Interaction, 23(3), 205-237.
Perfetti, C. A., Rouet, J.-F., \& Britt, M. A. (1999). Towards a theory of documents representation. In H. van Oostendorp \& S. R. Goldman (Eds.), The construction of mental representations during reading. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
Pintrich, P. R. (2000). The role of goal orientation in selfregulated learning. In M. Boekaerts, P. R. Pintrich, \& M. Zeidner (Eds). Handbook of Self-Regulation. San Diego: Academic Press.
Salmerón, L., Baccino, T., Cañas, J. J., Madrid, R. I., \& Fajardo, I. (2009). Do graphical overviews facilitate or hinder comprehension in hypertext? Computers \& Education, 53, 1308-1319.
Salmerón, L., \& Garcia, V. (2012). Children's reading of printed text and hypertext with navigation overviews: The role of comprehension, sustained attention, and visuospatial abilities. Journal of Educational Computing Research, 47(1), 33-50.
Salmerón, L., Kintsch, W., \& Cañas, J. J. (2006). Reading strategies and prior knowledge in learning with hypertext. Memory \& Cognition, 34, 1157-1171.
Winne, P. H. (2001). Self-regulated learning viewed from models of information processing. In B. Zimmerman \& D. Schunk (Eds.), Self-regulated learning and academic achievement: Theoretical perspectives (2nd Ed). Mahwah, NJ: Erlbaum.
Zia, L. L. (2000). The NSF National Science, Mathematics, Engineering, and Technology Education Digital Library (NSDL) Program: A progress report. D-Lib Magazine, 6: Available
<http://www.dlib.org/dlib/october00/zia/10zia.html\#ref08 \(>\).

\title{
Associative strength and semantic activation in the mental lexicon: evidence from continued word associations
}

\author{
Simon De Deyne (simon.dedeyne@ppw.kuleuven.be) \({ }^{a}\), \({ }^{b}\) \\ Daniel J. Navarro (daniel.navarro@adelaide.edu.au) \({ }^{b}\) Gert Storms (gert.storms@ppw.kuleuven.be) \({ }^{a}\) \\ \({ }^{a}\) University of Leuven, Department of Psychology, Tiensestraat 102, 3000 Leuven, Belgium \\ \({ }^{b}\) University of Adelaide, School of Psychology, 5005 Adelaide, Australia
}

\begin{abstract}
In a word association task, the probability of producing a certain response to a cue is considered to be a direct measure of associative strength between words in the mental lexicon. The common single word association procedure is limited, since the number of words connected to a cue might be underestimated when a single response is asked. The continued association task overcomes this limitation by asking a person to generate multiple associative responses. To test whether continued strengths allow a better approximation of our lexicon, an experiment was conducted in which participants judged the associative strength between words.
Our results show that in contrast to other semantic tasks, continued strength predicts weak to moderate judgments only. Two explanations based on the sampling of information and differential semantic activation of later responses in continued association are proposed. Theoretical implications for semantic activation and methodological implications for derivation of strength are discussed.
Keywords: associative strength, semantic relatedness; word associations.
\end{abstract}

The free word association task has been used extensively to investigate processes and structure in semantic and episodic memory. The task is attractive because it is unconstrained and straightforward, and no apriori restrictions are formulated about what types of relationships between words are deemed relevant. It leads to a rich and varied source of information. Compared to constrained tasks such as feature generation, it tends to provide more thematic relations like DOCTOR NURSE. There is increasing agreement that this thematic information determines much of how natural language concepts are used both in daily life and in language phenomena studied in the lab including semantic priming, metaphor comprehension, categorization and induction (e.g. Lin \& Murphy, 2001; Wisniewski \& Bassok, 1999).

An influential metaphor for the representation of this knowledge presents the mental lexicon as a weighted graph, where the structure of the links between the nodes (words) determines how words relate to each other and get their meaning. Obviously, the value of such a representation hinges on how the words are connected and on what determines the strength of these connections. The key assumption underlying the word association task, is that the number of people that generate a specific response to a cue is an indication of the strength between cue and response. Approximating the relations in the lexicon through word associations explains numerous phenomena: facilitation of word processing in associative priming (Hutchison, 2003), the probabil-
ity of recall in cued-recall tasks (Nelson, Zhang, \& McKinney, 2001), reaction times in lexical decision (De Deyne, Navarro, \& Storms, 2012) and generation frequencies in fluency tasks (Griffiths, Steyvers, \& Tenenbaum, 2007). Moreover, the overlap of the distributions of these strengths for two words indicates how semantically related they are and this is the basis of the success of lexico-semantic models such as Latent Semantic Analysis (LSA, Landauer \& Dumais, 1997) and topic models (Griffiths et al., 2007).

Associative strength is central to how we process the meaning of words, but the traditional way of measuring it, through asking a participant a single word association, is not without limitations. The response frequencies from the single word association task are considered reliable only for the very strong associates, since weaker responses are often missing (Nelson, McEvoy, \& Dennis, 2000). This lack of weak associations is seen as a general drawback of the word association procedure (Aitchison, 2003, p. 101) and has been responsible for questioning the results of previous findings in numerous tasks such as mediated priming (e.g., Chwilla, Kolk, \& Mulder, 2000). Presumably, this reflects dominance effects where for a cue like umbrella a single strong associate such as RAIN accounts for almost all responses (Nelson \& Bajo, 1985). While the exact causes of dominance effects are not well understood, it is obvious that they make the response distributions overly sparse, and bias all kinds of association derived strength measures.

Recently, a large-scale continued word association database was completed involving over 70,000 participants and 3 million responses (De Deyne, Navarro, \& Storms, 2012). In contrast to previous studies, a continued word association task was used in which subjects were presented a short list of stimulus or cue words and asked to give three different responses to each of these cues. The goal of the present study is to investigate how word association frequencies in continued tasks map onto associative strength. If single word associations tend to underestimate or be unreliable for weaker responses, then we would expect that using information encoded in later responses might alleviate this problem. This would support previous findings where semantic relatedness derived from continued association norms results in a better predictor of semantic tasks including pair-wise similarity judgments (De Deyne, Peirsman, \& Storms, 2009), prototypicality judgments (De Deyne, Voorspoels, Verheyen, Navarro,
\& Storms, 2011), and response times in the lexical decision task (De Deyne et al., 2012).

Since continued word association data only became available recently (cfr. De Deyne et al., 2012), few have studied strength derived from multiple responses and how it relates to other measures of associative strength. For instance in the study of Garskof (1965) calculated strength using continued associations to 20 cues and found that a weighted sum depending on the rank of the response correlated higher than a measure of strength that did not take into account response position. In contrast to previous work, this study presents a systematic comparison of measures of strength by looking at the contribution of continued responses alongside that of single responses using a recently proposed task in which participants judge associative strength of word pairs directly (Koriat, 2008; Maki, 2007) and compare it to single associate strength measures.

Sometimes the best way to understand a phenomenon is to take a step back. To aid the interpretation of the pattern of results from the judgment of associative strength task, the second part of this paper describes additional evidence by comparing expected strengths of continued responses with the observed strengths of these responses in the continued task. This analysis allows us to interpret quantitative differences (due to the sampling regime in continued association), and qualitative differences in terms of the types of semantic information activated in later responses.

\section*{Judgment of Associative Strength Experiment}

In a series of experiments on associative strength, Maki (2007) asked subjects to estimate how many people out of 100 would consider two words to be associated. Using a similar judgment of association strength task, our goal was to find out whether continued responses provide a better approximation compared to a single response procedure.

To test this hypothesis we compared various models, starting with a simple one that predicts judgments using the word association counts of the first three response positions ( \(R_{1}\), \(R_{2}, R_{3}\) ). Strength can be forward strength \((F S)\), or the probability that a certain response is generated given a cue or backward strength \((B S)\) : the probability of a certain cue given a specific response. These measures are easily derived by dividing the frequency of a certain response by the total number of responses for that cue.

\section*{Method}

Participants Fifty native Dutch speaking psychology students participated in exchange for course credit.
Stimuli and Materials The stimuli were selected from a set of more than 12,000 Dutch cues that were part of a large scale continued word association database described in De Deyne et al. (2012). Similar to De Deyne et al. (2012) single and multiple response strength were derived from the graph \(G_{1}\) based on the first response \(G_{2}\) based on the secondary and \(G_{3}\) for tertiary responses. These graphs were obtained by con-
verting the bimodal cue by response matrix to a unimodal cue by cue matrix by retaining those responses present in the set of cues. This makes it possible to get estimates of both backward and forward strength since all responses are also present as a cue in such a graph. For each cue, \(F S\) was calculated using only the first response \(\left(G_{1}\right)\) or including the sum of all three responses \(G_{123}\). The cues were determined randomly subject to following conditions. Only responses that were present both in \(G_{1}\) and \(G_{123}\) were considered. The difference in response strength was calculated and responses were selected that differentiated between both graphs.

A total of 80 associated cues and responses were chosen to cover the entire range of forward and backward strength between 0 and 1. All words in the judgment tasks were unique and only Dutch words were admitted that had a word frequency larger than one in the SUBTLEX-NL word frequency norms (Keuleers, Brysbaert, \& New, 2010). Similar as in Maki (2007), 20 unrelated pairs such as RAFT-LION or TASK-SIN were added to the 80 related pairs. Since these do not share any associations, their forward and backward strengths equaled zero. \({ }^{1}\)

Procedure Participants were tested during a collective session in a computer room using an online survey. Similar to Maki (2007) the subjects were familiarized with the word association task. Each participant was asked to give three responses to a set of 15 cues in a task identical to the one described in De Deyne et al. (2012). Upon completion of the word association study they were directed to the instruction page for the judgment of associative strength study and asked to estimate how many out of a hundred persons from Belgium, would give a certain association. An example was shown for a highly related pair (CAPTAIN-BOAT) and a weakly related pair (CAPTAIN-HAT). Finally, they were told to use a sliding scale to indicate their judgments and to consider the entire range of the scale from 0 to 100 . A total of 100 items were presented in a randomized order and had the following format: In a word association task, the word X was presented. How many people out of 100 responded with the word Y? The judgment of associative strength task took about 10 minutes on average to complete.

\section*{Results and Discussion}

The average of all ratings was calculated and the SpearmanBrown formula for split-half reliability was applied on the data from 50 subjects. The result showed that the ratings were highly reliable: \(r_{\text {splithalf }}(100)=.99\).

The judgment of association strength as a function of normed association based on single response strength \(F S_{1}\) is plotted in Figure 1. This Figure shows that weak and moderately strong normed associates are overestimated in the judgments of strength (as indicated by their relative position toward the diagonal), while strong associates tend to be underestimated. This is in line with the previous findings reported

\footnotetext{
\({ }^{1} \mathrm{~A}\) full list of the stimuli is available from http://www .smallworldofwords.com/experiments/
}


Figure 1: Scatter plot for judged and normed \(F S_{1}\) together with regression line and confidence bounds.
by Maki (2007). The relative contribution of different instantiations of associative strength measures based on continued association was investigated through a series of regression analyses where we focused on straightforward predictors that corresponded to interpretable and theoretically interesting aspects of strength. \({ }^{2}\)
Strength in related pairs In a first series of analyses, only the related items are considered as these data have non-zero values for both forward and backward strength measures. The results of the analyses are presented in Table 1. The simplest model predicts judgments of associative strength by the normed strength of the first response \(\left(F S_{1}\right)\). Model 1 accounts already for \(56 \%\) of the variance and found a significant effect of \(F S_{1}, \beta=.75, p<.01\). A second model is one where strength is averaged over all three responses: \(\left.F S_{123}=\left(F S_{1}+F S_{2}+F S_{3}\right) / 3\right)\). This predictor was significant ( \(\beta=.70, p<.01\) ), but the model only captured \(49 \%\) of the variance. In contrast to previous studies (cfr. De Deyne et al., 2011, 2012), the added information from \(R_{2}\) and \(R_{3}\) does not improve the prediction of the judgments of associations strength. Model 3 considers the possibility that judged strength is a function of both forward and backward strength of \(R_{1}\). Both \(F S_{1}(\beta=.70, p<.01)\) and \(B S_{1}(\beta=.21, p<.01)\) were significant predictors and provided the best account of the data so far.

Next, we investigated if \(R_{2}\) and \(R_{3}\) responses provide additional information beyond that captured in \(R_{1}\). Model 4 expands Model 1 by including \(F S_{2}\), resulting in significant effects for \(F S_{1}(\beta=.75, p<.01)\) but not \(F S_{2}(\beta=.11, n s)\) resulting in little extra variance accounted for (see Table 1). Similarly, no effect was found for \(F S_{3}\) in any additional analysis that was not accounted for by either \(F S_{1}\) or \(F S_{2}\). So these will not be discussed further.

\footnotetext{
\({ }^{2}\) To reduce the skew in the count-based strength measures a logtransformation was used.
}

Table 1: Regression models (\#M) for the prediction of judged associative strength. Only significant models are reported and adjusted \(R^{2} \mathrm{~s}\) are used throughout.
\begin{tabular}{rllr}
\hline & & \multicolumn{2}{c}{\begin{tabular}{rlr} 
Related \\
Regression Equation
\end{tabular}} \\
\hline 1 & \(F\)-test & \(F(1,78)=99.8\) & \(69+22 F S_{1}\) \\
2 & \(F(1,78)=75.9\) & \(88+14 F S_{123}\) & .556 \\
3 & \(F(2,77)=58.9\) & \(83+21 F S_{1}+16 B S_{1}\) & .487 \\
4 & \(F(2,77)=51.7\) & \(74+22 F S_{1}+7 F S_{2}\) & .594 \\
5 & \(F(3,76)=52.0\) & \(90+17 F S_{1}+8 B S_{1}+29\) Rel & .562 \\
& & \(\mathbf{2 5 \%}\) Quantile & \\
1 & \(n s\) & - & - \\
2 & \(F(1,37)=6.4\) & \(71+91 F S_{123}\) & .125 \\
3 & \(F(1,37)=4.3\) & \(59+14 F S_{2}\) & .080 \\
4 & \(F(2,36)=8.5\) & \(102+16 F S_{2}+43 B S_{1}\) & .284 \\
5 & \(F(3,35)=11.3\) & \(107+17 F S_{2}+22 B S_{1}+38\) Rel & .448 \\
\hline
\end{tabular}

A final model considered the role of relatedness. It is quite possible that when faced with uncertainty about exact strength, participants use the semantic relatedness between the cue and target to infer how strongly associated they are. Semantic relatedness was calculated as the cosine between the cue and response vector (see De Deyne et al. (2012) for additional details). Intuitively a high cosine indicates many shared associates between two words, while a low cosine indicates few shared associates. Model 5 gave the best fit of the data ( \(R^{2}=.66\) ), with significant effects for both \(F S_{1}\) ( \(\beta=.59, p<.01\) ) and relatedness (Rel, \(\beta=.31, p<.01\) ). \(B S_{1}\) was no longer significant ( \(\beta=.10, n s\) ).
Modeling weak strengths Still, it might be too early to conclude that normed strength from later responses never predicts strength judgments. As can be seen from Figure 1, \(F S_{1}\) at the low end of the scale does not distinguish much of the observed judged data. Possibly, strength derived from later responses results in more stable estimates for those responses that occur less frequent as \(R_{1}\). At this low end of the \(F S_{1}\) scale, participants might make use of richer information, corresponding to information encoded in \(F S_{2}, F S_{3}\), backward strength, or semantic relatedness.

To investigate if the weak strengths are better captured by \(R_{2}\) and \(R_{3}\), a subsection of the data presented was selected by placing a cut-off at the first quartile of \(F S_{1}\), as most of the remaining data were not explained by \(F S_{1}\).

The same models as presented before were now used to predict these data. The results for \(F S_{1}\) in Model 1 confirmed the pattern in Figure 1, as it was unable to predict any data. A significant effect for summed strength \(F S_{123}(\beta=.38, p<\) .05) was found in Model 2, explaining \(13 \%\) of the variance. Since \(F S_{1}\) did not explain the data, a new model consisting of \(F S_{2}\) was tested and found significant ( \(\beta=.32, p<.05\) ). The following models therefore use \(F S_{2}\) rather than \(F S_{1}\). In Model 4, both \(F S_{2}(\beta=.38, p<.01)\) and \(B S_{1}(\beta=.47, p<\) .01) were significant and accounted for \(28 \%\) of the variance.

The final model including relatedness explained most of the variance (45\%) with a significant effect of \(F S_{2}(\beta=.40, p<\) .01 ) and relatedness ( \(\beta=.47, p<.01\) ), but no significance for backward strength \(B S_{1}(\beta=.24, n s)\).

Together, these results support the idea that the judgment of association strength task is sensitive to normed associative strength, and closely replicates the previous findings of (Maki, 2007). However, our main goal was to investigate whether continued responses lead to better approximations of judged strength. Our findings support this hypotheses, but only for weak or moderate strengths. Since no large-scale studies have looked at the effect of continued associations, the next section will go into detail about which mechanisms might cause these results.

\section*{What factors determine the contribution of continued responses?}

A question that arises from the previous findings is why strength measures that include \(R_{2}\) and \(R_{3}\) responses systematically improve the prediction in a variety of semantic tasks such as similarity judgment tasks or lexical decision tasks (De Deyne et al., 2012), but not in the judgment of association task. Can we provide an explanation why they capture no additional information compared to \(R_{1}\) strengths at the high range of the scale?
Sampling without replacement hypothesis. A first explanation is based on the idea that continued responses are biased due to the continued nature of the task. More precisely, participants are not allowed to repeat a response. Especially when a certain \(R_{1}\) association is very dominant, the proportion of participants who did not generate it as \(R_{1}\) but could generate it as \(R_{2}\), will be very low. Summing strengths in these cases might bias strength for such a response. In other words, the strength measures for \(R_{2}\) and \(R_{3}\) do not take into account this sampling without replacement. As consequence of the restriction of sampling without replacement, we expect \(F S_{2}\) to be heavily biased for the strong responses, but at least capture moderate and weak strengths. If sampling without replacement is the main factor governing the observed frequencies for continued responses, then the derived expected strengths for the secondary and tertiary association response should closely agree with the observed strength for \(R_{2}\) and \(R_{3}\).

Given a specific cue with \(N\) different responses one can derive the expected \(R_{2}\) response count for \(x\) from its probability as a first response \(R_{1}\) as follows:
\[
\begin{equation*}
P\left(R_{2}=x\right)=P\left(R_{1}=x\right) \sum_{i=1, i \neq x}^{N} \frac{P\left(R_{1}=i\right)}{1-P\left(R_{1}=i\right)} \tag{1}
\end{equation*}
\]

The same principle holds for the derivation of the joint expected response for \(R_{3}\). For each of the 12,428 cues in \(G_{1}\), the expected \(R_{2}\) strengths were calculated using Equation 1. If differences between expected and observed \(F S_{2}\) strengths are primarily caused by the sampling without re-


Figure 2: Averages and \(S D\) for correlation between expected and observed \(F S_{2}\) (left-hand \(y\)-axis) grouped by entropy \((H)\). A histogram of entropy for each cue with counts was added (right-hand \(y\)-axis).
placement, then the expected and observed values should be similar up to some random noise. For each of the cues, the correlations between expected and observed strength distributions were obtained and had an average correlation of \(r(12428)=.71(S D=.13)\). At this point, it is not clear what determines high or low agreement. A corollary from the strength without replacement explanation is that the degree of bias in \(F S_{2}\) or \(F S_{3}\) will depends on the set-size or heterogeneity of the \(R 1\) response distribution which can be formalized as entropy \(H\) :
\[
\begin{equation*}
H=\sum_{i=1}^{N} p_{i} \log _{2} \frac{1}{p_{i}} \tag{2}
\end{equation*}
\]
where \(N\) is the size of the vocabulary or number of different responses and \(p_{i}\) is the probability for the \(i\) th response. \(H\) increases as the responses become more heterogeneous and equals zero if all responses were identical.

Figure 2 shows the average correlations binned as function of the entropy for the cues. For cues with few responses, the correlation between expected and observed counts is lower. Similarly, the cues with a very heterogeneous response set corresponding to the high entropy words at the right-hand side of Figure 2 also exhibit lower agreement than average entropy cues. A possible explanation of the former effect is due to dominance effects previously observed in cued recall (Nelson \& Bajo, 1985), where a single strong response inhibits the retrieval of other weaker ones. For these low entropy cues we expect higher utility of \(F S_{2}\) or \(F S_{3}\) in the judgment of associative strength assuming that the effect of dominance is removed once the response is generated and additional information becomes accessible. The latter effect could be due to unreliability, where at the high extreme cues elicit only idiosyncratic responses. Little benefit of \(F S_{2}\) can be expected for high entropy cues, since there is no reason to expect very heterogeneous responses to become more coherent in the later responses. For these cues it should be dif-
ficult predicting associative strength whether this strength is based on \(F S_{1}\) or \(F S_{2}\). New pilot studies seem to support this entropy interaction. However, there are number of reasons why sampling restrictions cannot completely explain the observed response distributions for continued responses. First of all, this does not explain why the heterogeneity or entropy increases when more than one response per cue is asked. Second, it might be the case that for different response positions, distinct types of semantic information becomes available.
Time course of Semantic Activation Hypothesis. A possible explanation why some \(R_{2}\) and \(R_{3}\) responses are generated much more (or less) frequent than expected based on \(R_{1}\) when sampling without replacement is taken into account stems from the idea that qualitatively different sources of information are accessed. A first possibility is that the type of response for \(R_{2}\) and \(R_{3}\) is influenced by the previous response beyond previously noted sampling restrictions. Such an order effect is called chaining, and can be illustrated for the cue SWISS, where mOUNTAINS is given more frequently as an \(R_{1}(57 \%)\) than \(R_{2}(16 \%)\), while it is expected \(26 \%\) of cases in \(R_{2}\). Together with the observation that SNOW is given less frequently than expected from its \(R_{1}\) counts, one can assume an associative chain: SWISS \(\rightarrow\) MOUNTAINS \(\rightarrow\) SNOW. The presence of chaining can be quite easily investigated, and previous research suggest this phenomenon is quite rare (De Deyne \& Storms, 2008).

Second, the different time course of automatic and qualitatively different types of semantic information might be a more important factor. Consider for example the cue GORILLA where MONKEY is generated in 72 times as \(R_{1}\). It is expected to occur 21 times as \(R_{2}\) yet occurs only 6 times. At the same time, BIG is generated 18 times as \(R_{2}\), but is expected to occur 6 times at most. Perhaps linguistic or superficial information like superordinate labels precede entity properties as in this example. Both behavioral (Santos, Chaigneau, Simmons, \& Barsalou, 2011) and fMRI studies (Simmons, Hamann, Harenski, Hu, \& Barsalou, 2008) support the idea that gradually deeper semantic information becomes activated. For example, in an experiment by Santos et al. (2011), participants generated about 1.7 responses in a continued time delimited task. In this study, later responses tended to convey a shift from primarily linguistic responses towards taxonomic- and especially thematic- and entity-related responses.

Perhaps a better way to study the time course of semantic activation is based on a comparison between observed and expected response frequencies for continued responses given the response distribution of the first response. Such a comparison is more accurate compared to previous approaches since it is not biased by (lack of) opportunity to generate a previous response in the continued procedure. To investigate what type of information is different in the second and third response of the word associations, we calculated the expected response frequencies for \(R_{2}\) and \(R_{3}\) (cfr. Equation 1) and compared them with the observed response frequencies by subtracting observed from expected \(R_{2}\) and \(R_{3}\). A positive value indicates


Figure 3: Distribution of semantic knowledge for observed responses in \(R_{2}\) and \(R_{3}\) that are either over- or underestimated based on expected \(R_{2}\) and \(R_{3}\) responses.
that the observed response in \(R_{2}\) or \(R_{3}\) is less likely to be generated than expected and this information is underestimated in \(R_{2}\) or \(R_{3}\). A negative value indicates that the response is generated more often than expected and is overestimated in the observed \(R_{2}\) or \(R_{3}\) counts. Since it is practical unfeasible to manually code all possible cue-response pairs only a subset of the data was used. For each of the \(+12,000\) cues the most extreme (one positive, one negative) responses were listed, once for \(R_{2}\) and once for \(R_{3}\). Both sets were sorted and only the 1,000 most negative and 1,000 most positive differences were retained for further analysis.

The relationship between \(2,000 R_{2}\) and \(2,000 R_{3}\) cueresponse pairs was coded as either as entity, introspective, lexical, thematic or taxonomic using similar guidelines as those described in De Deyne and Storms (2008) and Santos et al. (2011). Entity responses encode properties of the cue (e.g., MOON-YELLOW), introspective pairs encode evaluation or affect towards the cue (MOON-PRETTY), lexical attributes encode linguistic properties such as word compound completions, idioms, or rhyme (MOON-walk), thematic information could refer to agents, time and place of an action etc. (MOONASTRONAUT), taxonomic encodes super-,sub- and coordinates, synonyms and antonyms (MOON-PLANET). A detailed discussion of the implications for various types of semantic is beyond the scope of our illustration. For current purposes, we are mainly interested in identifying potential systematicity in qualitative response changes as a function of response position. The results in Figure 3 indicate that this is strongly the case. The largest effect is for taxonomic information which is much less likely to occur \(R_{2}\) and \(R_{3}\) than expected. To a lesser extend, there is also a shift where less lexical responses are generated as \(R_{2}\) or \(R_{3}\). The positive shift shows that entity and thematic responses are generated more frequent than expected for \(R_{2}\) and \(R_{3}\). These findings support the previous conclusions that linguistic information (encoded lexical) precedes conceptual types of information such as entity and thematic information. In contrast to the findings of Santos et al. (2011), our findings also show that taxonomic information is available early in the generation process.

This result also offer a potential qualitative interpretation of the contrast between a lack of effect of \(F S_{2}\) in the judged associative strength task and its significant contribution in similarity judgment and other more semantic tasks if semantic knowledge related to entity features and thematic roles is better encoded in \(R_{2}\) and \(R_{3}\). Clearly, follow-up studies are needed to further evaluate these hypotheses.

\section*{Discussion and Conclusion}

Using a judgment of associative strength task, we investigated the role of normed strength derived continued word associations. In contrast to previous reports where denser representations derived from second and third responses provided better estimates of distributional relatedness and lexical centrality (De Deyne et al., 2012), we found that the contribution of these responses is limited to weak or moderate response strengths. Moreover, in contrast to previous studies, simply summing response frequencies systematically resulted in inferior predictions for judgments of associative strength. Our interpretation of this finding is based on the notion that later responses are likely to underestimate the highest strengths due to sampling without replacement.

When comparing expected strengths under sampling without replacement against the observed strengths, the differences in \(R_{2}\) and \(R_{3}\) are very systematic and point out how semantic activation of types of knowledge changes over time, an issue which has been notoriously difficult to measure using other paradigms including priming. Importantly, using expected response frequencies for continued responses in comparison with actual observed response frequencies might provide a less biased baseline for tracking the time-course of semantic activation through continued association tasks. While different semantic information in continued responses strongly reflects the divergence between expected and observed counts for \(R_{2}\) and \(R_{3}\), it should be noted that other factors might also play a role. Since none of the responses in the association data is stemmed, it is quite likely that some part of the discrepancies will disappear when the data is processed this way. Our findings also result in a number of methodological recommendations as we have shown that ignoring sampling without replacement is problematic for low entropy cues and the use of single or combined strength measure depends on the type of task under consideration (ranging from associative to more semantic in nature).

At a theoretical level, our results challenge the main conclusions about the supposed overestimation bias of weak and moderate associates in judgments of associative strength (Maki, 2007; Koriat, 2008). The previous interpretation rests on the assumption that word association frequencies veridically reflect strength and only a small number of different responses are available (as is the case in single word association). Instead, we propose that this bias might not be due to the judgments themselves but could equally be an artifact of the single association procedure which underestimates low to medium responses.

\section*{Acknowledgments}

This work was supported by a research grant funded by the Research Foundation - Flanders (FWO) to the first author and by the interdisciplinary research project IDO/07/002. Special thanks to Toon Van Borm, Steven Verheyen and Amy Perfors for helpful comments.

\section*{References}

Aitchison, J. (2003). Words in the mind: An introduction to the mental lexicon. Wiley-Blackwell.
Chwilla, D., Kolk, H., \& Mulder, G. (2000). Mediated priming in the lexical decision task: Evidence from event-related potentials and reaction time. Journal of Memory and Language, 42, 314-341.
De Deyne, S., Navarro, D., \& Storms, G. (2012). Better explanations of lexical and semantic cognition using networks derived from continued rather than single word associations. Behavior Research Methods, Advance online publication. 10.3758/s13428-012-0260-7.
De Deyne, S., Peirsman, Y., \& Storms, G. (2009). Sources of semantic similarity. In N. Taatgen \& H. van Rijn (Eds.), Proceedings of the 31th Annual Conference of the Cognitive Science Society (p. 1834-1839). Austin, TX: Cognitive Science Society.

De Deyne, S., \& Storms, G. (2008). Word Associations: Network and Semantic properties. Behavior Research Methods, 40, 213231.

De Deyne, S., Voorspoels, W., Verheyen, S., Navarro, D. J., \& Storms, G. (2011). Graded structure in adjective categories. In L. Carlson, C. Hölscher, \& T. F. Shipley (Eds.), Proceedings of the 33rd Annual Conference of the Cognitive Science Society (p. 1834-1839). Austin, TX: Cognitive Science Society.

Garskof, B. E. (1965). Relation between single word association and continued asociation response hierarchies. Psychological Reports, 16, 307-309.
Griffiths, T. L., Steyvers, M., \& Tenenbaum, J. B. (2007). Topics in semantic representation. Psychological Review, 114, 211-244.
Hutchison, K. (2003). Is semantic priming due to association strength or feature overlap? Psychonomic Bulletin and Review, 10, 785-813.
Keuleers, E., Brysbaert, M., \& New, B. (2010). Subtlex-nl: A new measure for dutch word frequency based on film subtitles. Behavior Research Methods, 42(3), 643-650.
Koriat, A. (2008). Alleviating information of conditional predictions. Organizational Behavior and Human Decision Processes, 106, 61-76.
Landauer, T. K., \& Dumais, S. T. (1997). A solution to Plato's Problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychological Review, 104, 211-240.
Lin, E. L., \& Murphy, G. L. (2001). Thematic relations in adults' concepts. Journal of Experimental Psychology: General, 1, 3-28.
Maki, W. (2007). Judgments of associative memory. Cognitive psychology, 54(4), 319-353.
Nelson, D. L., \& Bajo, M. T. (1985). Prior knowledge and cued recall: Category size and dominance. The American Journal of Psychology, 98, 503-517.
Nelson, D. L., McEvoy, C. L., \& Dennis, S. (2000). What is free association and what does it measure? Memory \& Cognition, 28, 887-899.
Nelson, D. L., Zhang, N., \& McKinney, V. M. (2001). The Ties That Bind What Is Known to the Recognition of What is New. Journal of Experimental Psychology: Learning, Memory, and Cognition, 27, 1147-1159.
Santos, A., Chaigneau, S., Simmons, W., \& Barsalou, L. (2011). Property generation reflects word association and situated simulation. Property generation reflects word association and situated simulation. Language and Cognition,, 3, 83-119.
Simmons, W., Hamann, S., Harenski, C., Hu, X., \& Barsalou, L. (2008). fMRI evidence for word association and situated simulation in conceptual processing. Journal of Physiology - Paris, 102, 106-119.
Wisniewski, E. J., \& Bassok, M. (1999). What Makes a Man Similar to a Tie? Cognitive Psychology, 39, 208-238.

\title{
Language and cognitive load in a dual task environment
}

\author{
Nikolaos Engonopoulos (nikolaos.engonopoulos@ uni-potsdam.de) \\ Potsdam University, Department of Linguistics \\ Karl-Liebknecht-Str. 24-25, 14476 Potsdam, Germany
}

\author{
Asad Sayeed (asayeed@coli.uni-saarland.de) and \\ Vera Demberg (vera@coli.uni-saarland.de) \\ Cluster of Excellence, Saarland University, \\ Campus C7.4, 66123 Saarbrücken, Germany
}

\begin{abstract}
We investigate the effect of linguistic complexity on cognitive load in a dual-task scenario, namely simultaneous driving and language use. To this end, we designed an experiment where participants use a driving simulator while listening to spoken stimuli and answering comprehension questions. Online physiological measures of cognitive load, including the recently established Index of Cognitive Activity, as well as measures of performance in both tasks have been collected with high temporal resolution. The resulting aligned data streams can be used to test a vast array of different hypotheses about the relationship between performance, difficulty, and cognitive load in dual tasks at various levels of temporal resolution and linguistic structure. We present results of the data analysis, including evidence that different linguistic structures may cause measurable changes in cognitive workload on a very fine temporal scale in cases of increased primary task difficulty.
\end{abstract}

Keywords: relative clause; dual task; cognitive load; pupillometry; skin conductance; tracking task; driving; multitasking

\section*{Introduction}

Is there a relationship between psycholinguistic measures of language complexity and quantified cognitive workload in dual-task environments? To answer this question, we experimentally evaluate these measures of language processing in an environment where one task is language-related and the other not. Such language complexity measures have been shown in single-task studies to account for processing difficulty. This work represents a first step in which we investigate the effect of a grammatical structure (German locally ambiguous subject vs. object relative clauses) on a simplified, well-controlled non-linguistic task, a driving task.

Dual tasks are ubiquitous in everyday life, often in situations where attention and performance in the primary task is critical. An example is driving while engaging in dialogue, be it with a passenger, a dialogue-controlled interface, or remotely via mobile phone. Engaging in dialogue generally affects driving performance and safety (Just, Keller, \& Cynkar, 2008; Young, Regan, \& Hammer, 2007).

We manipulated the driving task difficulty and the structural complexity of the linguistic items. We also collected measurements of performance in both tasks and fine-grained physiological indicators of cognitive load, namely skin conductance levels and pupil sizes. We computed values from pupil size for the recent Index of Cognitive Activity (ICA). To our knowledge, this is the first study using the ICA measure in a setting with a language task.

\section*{Background and Related Work}

There is a rich literature on language use while driving a car, largely showing that speaking on the telephone has a negative effect on driving performance (Just et al., 2008; Kubose et al., 2006). Further studies found that this is specific to conversations with remote speakers (independent of whether one uses a hand-held device or free speaking), but that conversations with an in-car passenger are less problematic (Strayer, Drews, \& Johnston, 2003; Drews, Pasupathi, \& Strayer, 2004). It appears that passengers adapt their conversation to the traffic situation, leaving the driver more resources to deal with demands of the driving task when driving becomes difficult (Drews, Pasupathi, \& Strayer, 2008; Crundall, Bains, Chapman, \& Underwood, 2005; Villing, 2009). By contrast, remote conversational partners cannot adapt their speech, so that the driver may reach the point of cognitive overload more easily and thus commit driving errors. However, these lines of research have not taken into account how the fine-grained details of linguistic complexity affect cognitive load and driving task performance.

On the other hand, there is a very rich literature on linguistic processing difficulty in single tasks using brain imaging, ERPs, and reading time studies, as well as a number of dual task experiments generally showing that performance on the linguistic task deteriorates with increased complexity of the other task, see for example King and Just (1991). Finally, multiple models explain the effect of cognitive load in one task on performance in another (Baddeley, 2003; Wickens, 2008; Just, Carpenter, \& Miyake, 2003).

We see, however, unbroken ground in relating the effect of linguistic complexity on a realistic task (e.g., driving) and the size of the interference of linguistic processing with driving performance. This study takes a step in this direction in testing different methods for assessing cognitive load and the effect of one particular linguistic structure-incrementally ambiguous relative clauses-on driving performance in a simplified but controllable and continuous driving task.

\section*{The dual-task experiment}

\section*{The ConTRe task}

Our primary task was a tracking task (Jagacinski \& Flach, 2003) presented as a car driving scenario and called the "Continuous Tracking and Reaction" (ConTRe) task (Mahr, Feld,


Figure 1: A screenshot of the ConTRe steering task.

Moniri, \& Math, 2012). In this task, participants see a simulated 3-D road moving at a constant speed, intended to simulate a moving vehicle. Additionally, two bars of different color appear approximately 20 m in front of the simulated vehicle. The two bars represent the vehicle's position and the target (reference) position. They move laterally across the screen. The reference bar's movement is pseudo-randomly generated by an algorithm, while the "vehicle" bar is controllable by the participant by means of a gaming steering wheel. Participants were instructed to track the reference bar's movements with the controllable bar as closely as possible. To reduce noise in our data, we removed all other elements of the original ConTRe environment (e.g., buildings along the side of the road, and traffic lights), except for the road and the moving bars. A screenshot of the simulated environment can be seen in fig. 1.

This task is a useful abstraction of driving, since it allows a precise and continuous performance measure for steering, essential to driving. We manipulated the difficulty of the ConTRe task by changing the speed of the reference and vehicle bars in order to create a "difficult driving" condition and an "easy driving" condition \({ }^{1}\).

\section*{Language comprehension task}

The spoken comprehension task consists in listening to a sentence containing a relative clause followed by two thematically related 'filler' sentences and a yes/no comprehension question. Questions were related to the relative clause (50\% of the stimuli) or to the filler sentences. All sentences and questions are in German, inspired by Bader and Meng (1999). The stimuli are designed in pairs in such a way that the items in each pair are identical except for the form of the auxiliary of the relative clause (RC), which determines whether it is an object RC (ORC) or a subject RC (SRC). An example of such a relative clause pair is the following:

> Die Lehrerin, die einige Eltern wegen einer solchen Kleinigkeit angerufen [haben / hat], hat nun eine Elternversammlung einberufen.
> "The teacher \({ }_{\text {FEM }}\) [who called some parents / whom

\footnotetext{
\({ }^{1}\) Easy: reference bar maximum speed \(=1 \mathrm{~m} / \mathrm{s}\), controllable bar \(=\) \(2 \mathrm{~m} / \mathrm{s}\). Difficult: reference bar \(=2.5 \mathrm{~m} / \mathrm{s}\), controllable bar \(=4 \mathrm{~m} / \mathrm{s}\).
}
some parents called] because of such a trivial issue, has now called a parents' meeting."

The sentence is locally ambiguous between ORC and SRC until reaching the auxiliary; in previous experiments, increased reaction times in a speeded judgment task (Bader \& Meng, 1999) have been observed when subjects read "haben" (ORC) compared to "hat" (SRC). This is evidence for an interpretive bias toward SRC. All items were synthesized prior to the experiment using MARY TTS (Schröder, Charfuelan, Pammi, \& Türk, 2008) and pauses manipulated so that the critical region duration (hat / haben) is always identical.

\section*{Experimental setup}

Each experiment is divided into 4 recording phases, each lasting about 6 minutes, with short pauses in-between. Each phase is composed of a driving-only phase of 2 minutes followed by a driving-with-language phase of approximately 4 minutes, during which 10 blocks, consisting each of one relative clause, two fillers and one question. Participants answer the question verbally and their response is coded by the experimenter. In the first and the third phase, the driving difficulty is set to "easy", while in the second and fourth phase it is set to "difficult". The order of presented items in the language condition was randomized, and we ensured that each person only saw one condition of each item.

\section*{Measures of cognitive workload}

We have two principal sources of quantified cognitive workload data: physiological and task dependent. Our physiological measures are further divided into two subtypes: pupil area-based (pupillometry) and skin conductance-based, both of which have been widely used in cognitive workload studies, although principally on non-linguistic tasks. Our study is an opportunity to evaluate the relative efficacy of these data sources on linguistic tasks. We also take the opportunity to evaluate a novel form of pupillometric data processing: the Index of Cognitive Activity (ICA). To the best of our knowledge, ours is the first study to investigate the potential of the ICA as a measure of linguistically-induced cognitive load in a dual-task scenario.

Our task-dependent measure is driving performance in our simulated environment, which serves to confirm the "realworld" effect of variations in cognitive workload.

The Index of Cognitive Activity (ICA) Research in pupillometry (Just et al., 2003; Engelhardt, Ferreira, \& Patsenko, 2010; Palinko, Kun, Shyrokov, \& Heeman, 2010) has found that cognition-related changes in pupil size typically amount to a difference of \(20 \%\) relative to the typical pupil size (Laeng, Sirois, \& Gredebäck, 2010). However, light conditions also affect pupil sizes, with brightness-induced changes being much larger than cognitively induced ones (up to \(120 \%\) of typical pupil size).

The Index of Cognitive Activity (ICA; Marshall (2002)) is a patented measure which applies signal processing techniques to filter out slow, large light-induced changes and identify
the occurrence of short, abrupt changes in pupil size, held to be caused by cognitive load. The ICA measure is argued to be robust with respect to changes in light conditions and eye movement. It relates the frequency of rapid small changes in pupil size (also known as pupillary hippus) to cognitive load. The ICA measure has been used for measuring cognitive load in driving simulation tasks (Schwalm, Keinath, \& Zimmer, 2008), simulated driving and visual search (Marshall, 2007), detecting different levels of surgical skill (Richstone et al., 2010), and for measuring linguistically induced cognitive load (Demberg, Kiagia, \& Sayeed, 2013) among other uses. Demberg (2013) provides a more detailed analysis of the ICA measure in the dual task setting presented here.

ICA measurements have been shown to be relatively stable across several commonly used eye tracker models and sample rates ranging from 60 to 300 Hz (Bartels \& Marshall, 2012). We used a head-mounted Eyelink II and sampled at 250 Hz .
Skin conductance response Our second physiological proxy for measuring cognitive load is skin conductance response (SCR), which we calculate from skin conductance level (SCL). Changes in the electrical conductance of the skin are due to activity of the sweat glands, which are in turn controlled by the sympathetic nervous system. Skin conductance amplitude usually changes with respect to its "neutral" (tonic) level in response to unexpected, significant, or aversive stimuli. SCL has been previously used as a measure of cognitive load (Shi, Ruiz, Taib, Choi, \& Chen, 2007). In a dual task experiment with simulated driving and a secondary cognitive task, B. Mehler, Reimer, Coughlin, and Dusek (2009) found that skin conductance levels peaked in cases of mental overload caused by incrementally increasing secondary task difficulty, which was followed by a deterioration in the performance of the primary task. Son and Park (2011) found skin conductance levels along with steering wheel reversals (used as a measure of task performance) to be good input features for an artificial neural network built to predict task difficulty.

We used the Ledalab software (Benedek \& Kaernbach, 2010) to separate our raw skin conductance measurements into an estimate of the tonic component and the phasic component. The software also allows to calculate the number of skin conductance response events. SCR events are the "peaks" of the phasic component of skin conductance; both the number of such events per time unit and the amplitude of the peaks are used in the analysis below.

Driving performance We use performance on the ConTRe task as an additional measure of cognitive load. The task lets us define several measures of task success, including the distance between the reference bar and the controllable bar at each point in time and the speed and acceleration of the controllable bar.

\section*{Results}

We ran our experiment with 24 German native speakers aged 20-34, with the total duration of the recorded samples sum-
ming up to about 12 hours. We performed our data analysis in R using linear mixed effects (LME) modeling with lme4 (Baayen, Davidson, \& Bates, 2008) and mgcv (Wood, 2001).

\section*{Correlation between physiological measures}

The first question we explored was whether our physiological measures are correlated with one another. While there is no significant correlation between the raw skin conductance levels and the ICA, we do find a significant positive correlation between the number of skin conductance events and the ICA (using Spearman's \(\rho\); left ICA: \(\rho=0.06 ; p<0.0001\); right ICA: \(\rho=0.09 ; p<0.0001\) ). One important aspect to keep in mind is also possibly different latencies of the two measures in reaction to a stimulus.

We find a strong correlation between the ICA of the left and right eye (cor \(=0.74 ; p>0.001\), Pearson's product-moment correlation coefficient).

\section*{Response to experimental phases}

Driving performance The next hypothesis we tested was whether our task performance measure in the driving task, i.e., the steering deviation, is sensitive not only to the driving task difficulty, but also to the presence of language. In figure 2 , we have plotted the mean deviation for each of the difficulty settings (easy and difficult driving), with and without the secondary linguistic task. Using linear mixed effects models with a random intercept and random slopes by subject, we found a large significant main effect of driving difficulty (coef \(=0.3 ; t=20.33 ; p<0.001\) ), showing that steering was less accurate when driving was more difficult. We also found a significant positive main effect of whether we are in a language phase (coef \(=-0.05 ; t=-5.00 ; p<0.001\); steering is worse when people are listening to language, see also figure 3 ), as well as a significant interaction between driving difficulty and the language phase, indicating that the effect of language was more burdensome in the difficult driving condition (coef \(=-0.024 ; t=-6.98 ; p<0.001\) ). To confirm whether the effect of language is significant in both driving conditions, we also split the data into two subsets, easy driving and difficult driving, and found that the effect of language was significant in both linear mixed effects models.

This figure illustrates an obvious difference between steering deviation in the easy and difficult driving conditions.


Figure 2: Driving condition/language vs. steering deviation.

Table 1: ICA estimates for the driving plus language phases.
\begin{tabular}{l|rrr|rrr} 
& \multicolumn{4}{|c}{ right ICA } & \multicolumn{3}{c}{ left ICA } \\
& coef & t-value & sign & coef & t value & sign \\
\hline (Intercept) & 0.8116 & 123.40 & \(* * *\) & 0.7965 & 135.63 & \(* * *\) \\
sound playing & 0.0198 & 10.88 & \(* * *\) & 0.0186 & 10.10 & \(* * *\) \\
easy driving & -0.0057 & -2.44 & \(*\) & -0.0004 & -0.21 &
\end{tabular}

Table 2: \# of SCR events reduced during easy driving. (Random slope of driving condition by subject included.)
\begin{tabular}{l|rrc} 
& Estimate & t value & signif. \\
\hline (Intercept) & 0.68626 & 12.550 & \(* * *\) \\
difficulty=easy & -0.06495 & -4.274 & \(* * *\)
\end{tabular}

Pupillometry For the ICA, we find a main effect of driving difficulty in the ICA of the right eye, but not in the left eye (Table 1). Furthermore, we find significantly more blinks during the phases when language was playing. In-depth analysis of the pupillometric data reveals that overall dilation was larger when people were listening to language stimuli, but the number of ICA events was lower (Figure 3). If we look into the language phase, however, the ICA of both eyes went down significantly whenever language wasn't playing (e.g., between stimuli; Table 1: we factored out the effect of blinks or partial blinks on both the pupil area calculations and the ICA). This effect can also be seen in Figure 3, where the 10 ICA spikes in the language region coincide with our 10 blocks of language stimuli.

Skin conductance For skin conductance, we cannot easily compare the easy vs. difficult driving settings, as the skin conductance measuring device was removed between phases, and comparison of absolute values between phases is thus impossible. A measure that can be compared between driving conditions is however the number of skin conductance events. When running a linear mixed effects regression model with this measure as a response variable, we find that more such skin conductance events happened, as expected, in the difficult driving condition, see Table 2.

We do however not find any significant effect of the language vs. no language condition on this measure. Unexpectedly, we find that tonic skin conductance is lower in the driving plus language condition, see Figure 3.

\section*{Cognitive load and language processing difficulty}

To this point, we find that the measures largely behave as expected. Thus we come to our main question: can they detect the effect of fine-grained language complexity? To this end, we analysed the data to see whether we can find a) a correlate for higher processing difficulty in the ambiguous region or right after the disambiguation at hat/haben, and b) whether ORCs lead to less cognitive load than SRCs.

Disambiguating region Detailed analysis of the ambiguous region of the relative clause shows that the Index of Cognitive Activity is high during the ambiguous region of the


Figure 3: Spline plots ( 120 knots; with 0.95 conf intervals) showing the effect of language on physiological measures during an experimental phase ( 2 min driving only followed by 4 min of driving plus language).


Figure 4: SCR during time that stimulus is spoken.
relative clause (during the time span of -2000 msec to 0 msec ), and that the ICA sharply falls right after disambiguation (see Table 4 which shows a significant reduction in ICA of both eyes following disambiguation, encoded as time wrt. onset). These effects hold over and above effects of the steering task, which have been mathematically accounted for by including the task difficulty as a factor in the model. These results indicate that subjects encounter processing difficulty due to the ambiguity. (This is possibly also something they learn during the experiment.)

For skin conductance, we know that effects can be expected 2-4 seconds after the stimulus. Figure 4 shows a significant rise in skin conductance during the five seconds after

Table 3: Mixed effects regression analysis with steering deviation as response variable, for region of 2 s before the onset till 2 s after end of the critical region.
\begin{tabular}{l|lrl} 
& Estimate & t -value & \\
\hline (Intercept) & \(3.562 \mathrm{e}-01\) & 17.07 & \(* * *\) \\
phase time & \(8.459 \mathrm{e}-08\) & 3.44 & \(* * *\) \\
target velocity & \(3.832 \mathrm{e}-01\) & 205.08 & \(* * *\) \\
critical region & \(1.396 \mathrm{e}-02\) & 2.88 & \(* *\) \\
easy driving & \(-2.248 \mathrm{e}-01\) & -64.91 & \(* * *\) \\
target acceleration & \(-2.680 \mathrm{e}-02\) & -5.90 & \(* * *\)
\end{tabular}

Table 4: Mixed effects regression analysis with left and right ICA as response variable, \(100-1800 \mathrm{msec}\) after critical region onset. (Critical region duration: \(0-600 \mathrm{msec}\) )
\begin{tabular}{l|ll|ll} 
& \multicolumn{2}{|c}{ left ICA } & \multicolumn{2}{c}{ right ICA } \\
& Estimate & t-value & Estimate & t-value \\
\hline (Intercept) & 0.7504 & \(35.71 * * *\) & 0.736 & \(37.82 * * *\) \\
subject RC & -0.0354 & -2.12 & \(*\) & \\
phase time & \(-1.16 \times 10^{-7}\) & -2.59 & \(*\) & \\
time wrt. onset & \(-2.78 \times 10^{-5}\) & \(-6.38 * * *\) & \\
steering veloc & 0.0257 & \(5.37 * * *\) & & \\
steering accel & 0.0108 & 2.00 & \(*\) & \\
SRC:phase time & \(1.34 \times 10^{-7}\) & 2.12 & \(*\) & \\
& & & \(4.36 * * *\) \\
& & & & \\
& & &
\end{tabular}
the critical region, which would be consistent with an interpretation that the ambiguity causes higher cognitive load.

But can we see any effect of our linguistic stimuli on the driving performance? We compared steering accuracy at the time of the disambiguating region with steering accuracy during the two seconds before and after, and indeed found that deviation of the controllable bar from the reference bar was significantly larger during the disambiguating region than before or after; see the positive coefficient (Table 3) for the binary variable "critical region".

Subject vs. object relative clauses Finally, we test whether the ICA is sensitive to fine-grained linguistic complexity effects. We isolated the subset of the data which fell within the 1800 msec following the onset of the critical region hat / haben. The duration of this critical region at hat / haben is 650 ms in both conditions, which we imposed by manipulating the duration of the phrase boundary pause during synthesis. On this subset of the data, we built two LME models (one for each eye) with the ICA measure as the response variable and the relative clause type as the fixed effect, while also introducing a random effect per participant.

The results of this analysis are shown in Table 4. We can see that there is a negative effect for the SRC type in both cases, although only the result for the right eye is significant. The interpretation of the coefficient is that SRCs tend to occur with smaller values of ICA than ORCs.

We did not find any significant effects of relative clause condition on skin conductance, overall pupil dilation or steering performance.

Table 5: LME model for answer accuracy.
\begin{tabular}{l|rrc} 
& Estimate & t-value & Sig \\
\hline INTERCEPT & 2.663 & 5.72 & \(* * *\) \\
RC-TYPE (OBJ) & 0.445 & 1.17 & \\
VOICE (PASSIVE) & -1.802 & -3.11 & \(* *\) \\
DRIVINGDIFFICULTY (EASY) & -0.222 & -1.18 &
\end{tabular}

\section*{Performance in the language task}

A last link that we wanted to investigate was the one between performance in the linguistic task (i.e., answer accuracy) and the difficulty of the driving and language tasks. We built a binomial LME model with the answer accuracy as the response factor and driving task difficulty, relative clause type, and the voice (passive vs. active) of the question as fixed effects with a random intercept per participant and subject and a random slope for relative clause type by item \({ }^{2}\). The resulting coefficients are presented in Table 5. While answer accuracy was lower for object relative clauses (74\%) than for subject relative clauses ( \(78 \%\) ), and lower in difficult driving ( \(75 \%\) ) than in easy driving (77\%), these differences did not reach significance. (NB: questions related to relative clauses were only asked after half of the items; i.e., this analysis is based on relatively little data.) The only significant negative effect on answer accuracy was found for passive voice questions, which means that there are significantly more wrong answers to passive voice questions than to active voice ones (this is not unexpected, as it has long been known that passives are more difficult to process than actives (J. Mehler, 1963)).

\section*{Discussion and conclusions}

We designed the tasks in our experiment to require continuous attention. The language task clearly affects performance on the primary steering task: we see the effect of the secondary task in all of our measures. Furthermore, we find effects of linguistic ambiguity and complexity in our measures of cognitive load: during the ambiguous region in our stimuli, we see evidence for higher cognitive load in our pupillometric measure, which is also reflected in a slightly later galvanic skin response. During the disambiguating region, we observe significantly higher steering deviation, which indicates that people are allocating more mental resources to the linguistic task, hence impeding steering performance. We also found evidence for a measurable effect of linguistic complexity in our pupillometric measure ICA: the ICA was significantly higher during the disambiguating region and the following second for the ORC condition compared to SRC. This experiment provides early support for the ICA as a useful measure to assess language-induced cognitive load.

\section*{References}

Baayen, R., Davidson, D., \& Bates, D. (2008). Mixedeffects modeling with crossed random effects for subjects

\footnotetext{
\({ }^{2}\) Best model according to standard comparison with AIC and \(\chi^{2}\).
}
and items. Journal of memory and language, 59(4), 390412.

Baddeley, A. (2003). Working memory: looking back and looking forward. Nature Reviews: Neuroscience, 4.
Bader, M., \& Meng, M. (1999). Subject-object ambiguities in german embedded clauses: An across-the-board comparison. Journal of Psycholinguistic Research, 28(2).
Bartels, M., \& Marshall, S. (2012). Measuring cognitive workload across different eye tracking hardware platforms. In Proceedings of the symposium on eye tracking research and applications (pp. 161-164).
Benedek, M., \& Kaernbach, C. (2010). A continuous measure of phasic electrodermal activity. Journal of neuroscience methods, 190(1-5), 80.
Crundall, D., Bains, M., Chapman, P., \& Underwood, G. (2005). Regulating conversation during driving: A problem for mobile telephones? Transportation Research Part F, 8(3), 197-211.
Demberg, V. (2013). Pupillometry: the index of cognitive activity in a dual-task study. In Proceedings of the 35th annual meeting of the cognitive science society (cogsci-13).
Demberg, V., Kiagia, E., \& Sayeed, A. (2013). Language and cognitive load in a dual task environment. In Proceedings of the 35th annual meeting of the cognitive science society (cogsci-13).
Drews, F. A., Pasupathi, M., \& Strayer, D. L. (2004). Passenger and cell-phone conversations in simulated driving. In Proceedings of the human factors and ergonomics society 48th annual meeting.
Drews, F. A., Pasupathi, M., \& Strayer, D. L. (2008). Passenger and cell phone conversations in simulated driving. Journal of Experimental Psychology: Applied, 14(4), 392.
Engelhardt, P. E., Ferreira, F., \& Patsenko, E. G. (2010). Pupillometry reveals processing load during spoken language comprehension. The Quarterly Journal of Experimental Psychology, 63, 639-645.
Jagacinski, \& Flach. (2003). Control theory for humans: Quantitative approaches to modeling performance. CRC.
Just, M. A., Carpenter, P. A., \& Miyake, A. (2003). Neuroindices of cognitive workload: neuroimaging, pupillometric and event-related potential studies of brain work. Theoretical Issues in Ergonomics Science, 4(1-2), 56-88.
Just, M. A., Keller, T. A., \& Cynkar, J. (2008). A decrease in brain activation associated with driving when listening to someone speak. Brain research, 1205, 70-80.
King, J., \& Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. Journal of Memory and Language, 30(5), 580-602.
Kubose, T. T., Bock, K., Dell, G. S., Garney, S. M., Kramer, A. F., \& Mayhugh, J. (2006). The effects of speech production and speech comprehension on simulated driving performance. Applied cognitive psychology, 20(1), 43-63.
Laeng, B., Sirois, S., \& Gredebäck, G. (2010). Pupillometry - a window to the preconscious? Perspectives on Psychological Science, 7(1), 18-27.

Mahr, A., Feld, M., Moniri, M., \& Math, R. (2012). The ConTRe (Continuous Tracking and Reaction) task: A flexible approach for assessing driver cognitive workload with high sensitivity. In Automotive user interfaces and interactive vehicular applications (pp. 88-91).
Marshall, S. (2002). The Index of Cognitive Activity: measuring cognitive workload. In Proceedings of the 7th conference on human factors and power plants, 2002.
Marshall, S. (2007). Identifying cognitive state from eye metrics. Aviation, space, and environmental medicine, 78(Supplement 1), B165-B175.
Mehler, B., Reimer, B., Coughlin, J., \& Dusek, J. (2009). Impact of incremental increases in cognitive workload on physiological arousal and performance in young adult drivers. Transportation Research Record: Journal of the Transportation Research Board, 2138(-1), 6-12.
Mehler, J. (1963). Some effects of grammatical transformations on the recall of english sentences. Journal of Verbal Learning and Verbal Behavior, 2(4), 346-351.
Palinko, O., Kun, A., Shyrokov, A., \& Heeman, P. (2010). Estimating cognitive load using remote eye tracking in a driving simulator. In ETRA.
Richstone, L., Schwartz, M., Seideman, C., Cadeddu, J., Marshall, S., \& Kavoussi, L. (2010). Eye metrics as an objective assessment of surgical skill. Annals of surgery, 252(1).
Schröder, M., Charfuelan, M., Pammi, S., \& Türk, O. (2008). The MARY TTS entry in the Blizzard Challenge 2008. In Proc. blizzard challenge.
Schwalm, M., Keinath, A., \& Zimmer, H. (2008). Pupillometry as a method for measuring mental workload within a simulated driving task. Human Factors for assistance and automation, 1-13.
Shi, Y., Ruiz, N., Taib, R., Choi, E., \& Chen, F. (2007). Galvanic skin response (gsr) as an index of cognitive load. In CHI '07 extended abstracts on human factors in computing systems (pp. 2651-2656).
Son, J., \& Park, M. (2011). Estimating cognitive load complexity using performance and physiological data in a driving simulator. In AutomotiveUI, 2011, adjunct proceedings.
Strayer, D., Drews, F., \& Johnston, W. (2003). Cell phoneinduced failures of visual attention during simulated driving. Journal of experimental psychology: Applied, 9(1), 23.

Villing, J. (2009). In-vehicle dialogue management - towards distinguishing between different types of workload. In Proceedings of SimPE, fourth workshop on speech in mobile and pervasive environments.
Wickens, C. D. (2008). Multiple resources and mental workload. Human factors, 50(3), 449-455.
Wood, S. (2001). mgcv: GAMs and generalized ridge regression for R. R news, l(2), 20-25.
Young, K., Regan, M., \& Hammer, M. (2007). Driver distraction: A review of the literature. Distracted driving. Sydney, NSW: Australasian College of Road Safety, 379-405.

\title{
Pupillometry: the Index of Cognitive Activity in a dual-task study
}

\author{
Vera Demberg (vera@coli.uni-saarland.de) \\ Cluster of Excellence, Saarland University, Campus C7.4, 66123 Saarbrücken, Germany
}

\begin{abstract}
This paper reports experimental results on the index of cognitive activity (ICA), a recent micro-level measure in pupillometry, which relates processing load to the frequency of rapid small dilations of the pupil. We collected pupil size during a tracking task which was cast in a simulated driving context, as well as for a dual task of simultaneous tracking and language processing. The present results are the first to evaluate the ICA measure on these tasks. We find that the ICA is sensitive both to the simulated driving and the language task, and that it is more responsive to our driving task than overall pupil dilation. Overall, the use of the ICA as opposed to traditional pupillometry seems promising, as our data provide initial evidence that the ICA may be more responsive, and a more fine-grained measure of cognitive load than traditional macro-scale pupil dilation measures.
\end{abstract}

Keywords: Pupillometry, Index of Cognitive Activity, Dual Task, Language, Driving

\section*{Introduction}

The size of the pupil has long been known to reflect arousal (Hess \& Polt, 1960) and cognitive load in a variety of different tasks such as arithmetic problems (Hess \& Polt, 1964), digit recall (Kahneman \& Beatty, 1966), attention (Beatty, 1982) as well as language complexity (Schluroff, 1982; Just \& Carpenter, 1993; Hyönä, Tommola, \& Alaja, 1995; Zellin, Pannekamp, Toepel, \& der Meer, 2011; Frank \& Thompson, 2012), grammatical violations (Gutirrez \& Shapiro, 2010) and context integration effects (Engelhardt, Ferreira, \& Patsenko, 2010). All of these studies have looked at the macrolevel effect of the overall dilation of the pupil as response to a stimulus. Recently, another micro-level measure of pupil dilation has been proposed, called the "Index of Cognitive Activity" or ICA (Marshall, 2000, 2002, 2007), which does not relate processing load to the overall changes in size of the pupil, but instead counts the frequency of rapid small dilation, which are usually discarded as pupillary hippus (Beatty \& Lucero-Wagoner, 2000). The ICA has been argued to be robust to changes in ambient light and eye-movements, and can therefore be hoped to be more reliable and robust than overall pupil dilation. Furthermore, as it does not use the overall dilation of the pupil which can vary as a function of lighting and individual, the frequency of the rapid pupil dilations is argued to be more comparable across tasks and subjects.

If it reliably reflects processing load, the ICA would be a convenient method to assess processing load using an eyetracker, in naturalistic environments, e.g. while driving a car, and could therefore usefully complement the range of experimental paradigms currently used.

To our knowledge, the present paper is the first to test its response to a tracking task, and to analyze properties of the Index of Cognitive Activity such as its response delay to a
stimulus. The application of the method in a realistic scenario (measuring linguistically induced cognitive load during driving) also bears relevance for practical applications.

\section*{The Index of Cognitive Activity}

The Index of Cognitive Activity is a patented measure of cognitive load which has previously only been evaluated on a small range of tasks (Marshall, 2000, 2002, 2007; Schwalm, 2008; Schwalm, Keinath, \& Zimmer, 2008) including digit span tasks, and a simulated driving task. Using the ICA as a measure of processing load is motivated by the finding that pupil size can be affected by two different processes: lighting conditions and cognitive activity. In the overall pupil dilation, these two effects are confounded, even in stable lighting because there is a so-called "light reflex", meaning that the pupil oscillates irregularly and continually. Pupil dilation is controlled by two groups of muscles: circular muscles, which make the pupil contract and radial muscles, which make the pupil dilate. Furthermore, we know that the activation and inhibition patterns are different for reaction to light and reaction to cognitive activity (Marshall, 2000): dilations due to cognitive activity are very short and abrupt, while pupil size changes due to lighting are slower and larger. The ICA therefore tries to disentangle these patterns by performing a wavelet analysis on the pupil dilation record to remove all large oscillations and retain only the very short and rapid events (larger than a specified threshold), which are then attributed to the effect of cognitive activity.

The ICA events (rapid small dilations) per second are counted, divided by the number of expected ICA events per second (30), and the resulting number is then transformed using the hyperbolic tangent function, in order to obtain a number between zero and one \({ }^{1}\). To obtain a continuous measure, blinks are factored out by linear interpolation of adjacent events. When using the EyeTracking.Inc software, an ICA value per second is produced. To obtain finer granularity, we also calculated a per-100-msec ICA value from the ICA events (i.e. the rapid dilation events). Due to the short time span, we could not interpolate for blinks (which take about 100 msecs ) and therefore simply excluded from our analysis time all frames during which a blink or partial blink occurred.

\section*{Background on Pupillometry and the LC-NE area}

It has been observed that pupil dilation is strongly correlated with activity in the locus caeruleus (LC) region of the brain.

\footnotetext{
\({ }^{1}\) The method is patented, and the analysis program has to be licensed from EyeTracking, Inc., San Diego, CA. For details see (Marshall, 2000).
}

LC neurons is bilateral and emits the neuro-transmitter norepinephrine (NE) (Aston-Jones \& Cohen, 2005; Laeng, Sirois, \& Gredebäck, 2012). The LC-NE system is known to be activated by stress and is thought to also have a role in memory retrieval and memory consolidation. The activity of the LCNE system as reflected in pupil dilations can therefore be a valuable method of inspecting cognitive load, and might be particularly useful also in multi-tasking settings.

\section*{Experimental Setup}

We conducted an experiment with 24 subjects, during which participants had to simultaneously perform a tracking task as well as a language comprehension task. We also collected data for the tracking task in a single-task setting. Our tracking task was cast as a simulated driving task ("ConTRe task", (Mahr, Feld, Moniri, \& Math, 2012)). The screen displays a moving road with two periodically moving bars at the horizon. One of the bars moves randomly across the screen ("reference bar"), while the other bar is controlled by the subject with a gaming steering wheel. The task of the participants is to cover the reference bar with their "steering bar", as exactly as possible. Difficulty of the ConTRe task was manipulated by changing the intervals at which the bar moves, as well as the speed at which it moves (the bar then always travels at a constant speed to a randomly determined destination on the horizon), to create an easy and a difficult driving setting \({ }^{2}\).

The linguistic stimuli (loosely based on Bader \& Meng, 1999; see Example (1)) consisted of 40 locally ambiguous subject and object relative clauses in German, where the relative pronoun die is ambiguous between nominative and accusative case. The following NP (einige der Mieter) is also ambiguous between these cases. Accordingly, the relative clause type (subject vs. object relative clause) is ambiguous until the disambiguating verb (hat vs. haben) is encountered.
\[
\begin{align*}
& \text { Die Nachbarin, } \quad \text { [die } \text { dig, nom/acc }^{\text {einige }_{\text {pl, }} \text { nom/acc }}  \tag{1}\\
& \text { der Mieter auf Schadensersatz verklagt hat }{ }_{\text {sg }} / \\
& \text { haben } \left.{ }_{\text {pl }}\right]_{\text {relative clause, traf sich gestern mit Angelika. }}^{\text {"The neighbor, [whom some of the tenants sued }} \\
& \text { for damages / who sued some of the tenants for } \\
& \text { damages }]_{\text {relative clause }} \text {, met Angelika yesterday." }
\end{align*}
\]

The language stimuli were synthesized using the MARY text-to-speech system (Schröder \& Trouvain, 2003). Synthesized stimuli were used to control the exact duration and timing of stimuli and pauses, so that we could more easily align our data for analysis. In particular, we made sure that the disambiguating region (hat / haben) was equally long in both conditions, by manipulating the duration of the pause after hat/haben. Furthermore, using synthesized speech avoids problems with large differences in intonation.

Our experiment was conducted in four phases, between which participants were offered to take a break. Each phase

\footnotetext{
\({ }^{2}\) Driving speed was set to \(40 \mathrm{~km} / \mathrm{h}\) in easy setting, \(70 \mathrm{~km} / \mathrm{h}\) in the difficult driving setting; maximal speed setting for reference bar in easy setting was 1 , and 2.5 in difficult setting; maximal speed setting for steering bar was 2 in easy setting and 4 in difficult setting.
}
included 10 stimuli and 20 fillers, as well as 10 comprehension questions. The order of the stimuli was randomized. We recorded pupil dilations on both eyes using the head-mounted SR EyeLink II eyetracker at 250 Hz .

\section*{Data Analysis and Results}

Methods All analyses reported below were done using the lme4 (Baayen, Davidson, \& Bates, 2008) and mgcv (Wood, 2001) packages in \(R\).

Distribution of the ICA Figure 1 shows the distribution of the ICA calculated per second (top plot) and calculated for a window of 100 ms (bottom plot). While the aggregation is smooth for the 1 s window, there are only few possible distinct events in a 100 msec window (the bumps correspond to 0 events, 1 event up to 5 ICA events). Due to the tanh transformation of the ratio between observed and expected ICA events, the bulk of ICA values lies in a narrow range between 0.7 and 0.95 for the standard per-second aggregation.

The left and right ICA values are strongly correlated with each other (Spearman's rank correlation \(\rho=0.71 ; p<\) 0.0001 ; per-second ICA), but clearly not identical.


Figure 1: Understanding the distribution of ICA values: Density plot for the ICA for different aggregations.

Relationship between the ICA and pupil area Next, we inspect the relationship between the ICA and the overall pupil area. The correlation between these two measures are small (left eye per-second ICA: \(\tau=0.105 ; p<0.0001\); and right eye per-second ICA: \(\tau=0.0146 ; p<0.01 ;\) ). The autocorrelation plot in Figure 2 shows how dynamics of the two measures differ (Figure 2 only shows the left eye but the right eye looks very similar): while the ICA has little autocorrelation in the time-series analysis and changes dynamically, the overall pupil size has a high autocorrelation.

\section*{ICA and the ConTRe Driving Task}

The reference bar moves periodically at a constant speed (ca. every 4 seconds for 1-3 seconds in easy driving and every 2.5 seconds for .5 to 1 seconds in difficult driving). This periodicity can also be seen in the autocorrelation plots shown in


Figure 2: Auto-correlations for the ICA and pupil area.

Figure 4(a). More interestingly, we can also inspect the temporal relationship between the movement of the reference bar and any effect of this in the ICA or the overall pupil area, as shown in Figure 4(b). We can see that there is a time-shifted correlation between the movement of the reference bar and a reaction which we can measure in the ICA, starting at about 700 msec after a movement in the reference bar and peaking at about 1.1 seconds after reference bar movement. This effect is more pronounced in the difficult driving conditions than in the easy driving conditions (these results hold both for the driving only and the driving plus language conditions). As Figure 4(c) shows, there is however almost no discernible effect of the reference bar movement on overall pupil dilation.

These time series analyses are interesting because there was previously no published information on how quickly to expect an effect on the ICA. We however also don't yet know enough about what we actually see in the ICA: is it related to the reference bar stimulus? or maybe rather an effect of the action taken by the participant in the task? In order to shed some light on this question, we also ran an autocorrelation analysis for the ICA and the subject controlled steering bar. As Figure 4(d) shows, the correlation between the ICA and the steering bar is stronger than the correlation between the ICA and the reference bar. As people moved the

            velocity of reference bar
    left ICA ( \(1.3 \mathbf{~ s e c}\) shift)



Figure 3: Spline plot \((\mathrm{k}=10)\) for reference bar velocity and acceleration in the same model fitting the ICA.

(b) Correlation of the right eye ICA with the speed of the reference bar at different time lags; (left eye looks the same).

(c) Almost no time-series correlation can be found between movement of the reference bar and overall pupil size.


Figure 4: Time-series correlations left plots show easy driving, right plots show difficult driving.
steering bar as a reaction to the movement of the reference bar, the latency of the ICA with respect to the steering bar is also much smaller (starting right away and peaking at about 400 msec ). For further analysis, we re-aligned our measurements of the reference bar movement (shift by 1.3 s ) and steering bar movement (shift by 400 msec ) in order to align with the ICA.

Table 1: ICA estimates for the driving plus language phases.
\begin{tabular}{l|rrr|rrr|} 
& \multicolumn{3}{|c|}{ left ICA } & \multicolumn{3}{c|}{ right ICA } \\
& coef & t val & sig & coef & t val & sig \\
\hline (Intercept) & 0.704 & 49.30 & \(* * *\) & 0.730 & 50.49 & \(* * *\) \\
sound file playing & 0.034 & 9.18 & \(* * *\) & 0.033 & 8.99 & \(* * *\) \\
easy driving & -0.008 & -1.01 & & -0.012 & -2.08 & \(*\)
\end{tabular}

This adjusted alignment then allows us to enter these factors in regression and spline models. In a first analysis, we tested whether the ICA is explained only by the speed of the reference bar, or also by its acceleration. Figure 3 shows a spline plot for a model including both reference bar velocity and reference bar acceleration in fitting in turn left and right ICA. The patterns are independent of the driving condition (easy or difficult) and of the presence of language stimuli. We see a roughly linear relationship between reference bar speed and the ICA. The bottom plots of Figure 3 furthermore show a u-shaped correlation between the acceleration of the reference bar and the ICA, indicating that the ICA is larger when the reference bar starts moving or stops moving, and lower when it is not moving or moving at its constant top speed.

\section*{ICA and the language task}

The Effect of Language Figure 5 shows how the left and right ICA and left and right overall pupil sizes evolve during the phases of the experiment, which consist of approximately two minutes of driving followed by four minutes of driving and listening to speech and answering yes-no questions. The speech signal consists of 10 blocks of one item, two fillers and a yes-no question. The blocks are separated by a pause of 2 seconds. It is very interesting to compare the pupil area plots and the ICA plots: pupil area is large at the beginning of a phase, but the pupil contracts soon afterwards. At the beginning of the language phase, pupil dilation increases again, which is what we expected, given the additional load of language processing. Interestingly, this is not the case in the ICA data: The ICA only goes down very little during the drivingonly phase, and is overall lower in the dual-task section than in the single task section.

Another relevant observation is that we can observe 10 clear peaks in the ICA data, corresponding exactly to our 10 items. Such a relationship is not visible in the pupil area data (which also shows some periodicity but without a clear correspondence to stimuli). In a linear mixed effects model including only data from the driving plus language phase with the ICA as a response variable and two predictors (a flag whether a sound file is playing and a flag indicating whether the driving condition was easy or difficult), we find that the ICA is significantly higher when a sound file is being played than when it is not (i.e., between stimuli), see Table 1.

In regressions with pupil area as a response variable, whether the sound file is playing is a significant negative predictor on both the left eye (coef \(=-0.058 ; t=-4.9 ; p>\) \(0.001)\) and the right eye (coef \(=-0.067 ; t=-5.1 ; p>0.001\) ), while the driving difficulty manipulation does not reach significance on either eye.


Figure 5: Spline plot (120 knots) for ICA and overall pupil dilation as a function of the duration of the driving only followed by driving with language task.

Ambiguous Region Next, we would like to see whether the ICA reflects in some way our critical region, i.e. whether we see an effect to the relative clause ambiguity. To this end, we run a spline model showing the development of the ICA during the duration of an item, with three predictor variables: time-shifted steering bar velocity, time-shifted steering bar acceleration and the distance from the critical region. Reference bar velocity does not explain any of the variance in the ICA data once steering bar velocity has been included as a predictor, therefore, our models include only the steering bar data. Figure 6 shows that the ICA is relatively high during the ambiguous region but starts falling right after disambiguation.

Disambiguating Region Note that the two relative clause conditions are collapsed in Figure 6- but can we measure a facilitation in the subject relative clause condition as opposed to the object relative clause? We ran a mixed effects regres-

(b) Overall pupil dilation of left and right eye.

Figure 6: Spline plots with confidence intervals for the ambiguous and critical region. Sentences are aligned for the onset and end of the disambiguating word "hat" / "haben".
sion model with right and left ICA (in turn) as response variables and (time shifted) reference bar velocity and acceleration, (shifted) steering bar velocity and acceleration, relative clause type, phase time (indicating how far into the phase the measurement was taken) and driving difficulty as explanatory variables. We also enter item and subject as random effects, as well as a random slopes for relative clause condition under item and subject.

The mixed effects models shown in Table 2 include data from the time window of 100 msec till 1800 msec after the onset of the critical region. Due to co-articulation, we expect that differences of hat vs. haben should be audible from about 100 msec after the onset, and given our finding of the 1.3 s lag between the reference bar movement and the ICA reaction, the window up to 1.8 s after the onset of the critical region makes sure that we include the relevant part of the data in our model.
\begin{tabular}{l|rrr|rrr} 
& \multicolumn{4}{|c|}{ left ICA } & \multicolumn{3}{c}{ right ICA } \\
& Estimate & t val & sig & Estimate & t val & sig \\
\hline (Intercept) & \(7.247 \mathrm{e}-01\) & 39.24 & \(* * *\) & 0.718417 & 45.54 & \(* * *\) \\
subject RC & \(-3.777 \mathrm{e}-02\) & -2.26 & \(*\) & & & \\
phase time & \(-1.199 \mathrm{e}-07\) & -2.68 & \(* *\) & & & \\
steering velocity & \(2.541 \mathrm{e}-02\) & 11.08 & \(* * *\) & 0.022656 & 10.34 & \(* * *\) \\
steering accel. & \(1.094 \mathrm{e}-02\) & 2.01 & \(*\) & & & \\
SRC:phase time & \(1.411 \mathrm{e}-07\) & 2.23 & \(*\) & & &
\end{tabular}

Table 2: Mixed effects regression analysis with ICA as response variable, for region of \(100-1800 \mathrm{msec}\) after the onset of critical region. (Duration of critical region: \(0-600 \mathrm{msec}\) )

We found again that steering bar velocity is a better predictor of the ICA than reference bar velocity. For the left eye, we find that steering bar velocity, steering bar acceleration, phase time and our critical manipulation, the relative clause type, are significant predictors. In particular, we find that the left eye ICA is significantly lower when the item is a subject relative clause. We also find that the left ICA decrease as a function of when the item is presented within a phase (see also Figure 5). Additionally, we find a significant interaction between phase time and the relative clause condition, which indicates that the difference in ICA between the subject and object relative clauses gets weaker as the experiment proceeds - it is possible that this is a learning effect.

In the right ICA, we see similar tendencies, but, with the exception of steering bar velocity, the predictors fail to significantly improve model fit. It should be noted though that this finding replicates the finding of a language-only study using the same relative clause stimuli, which also found a significant effect of relative clause type on the left eye's ICA but not on the right eye (Demberg, Kiagia, \& Sayeed, 2013).

While we cannot find a significant main effect of relative clause type in a regression with overall pupil size as a response variable, but we do find that the pupil size decreases significantly more quickly in the subject relative clause condition than in the object relative clause condition (this holds for both right eye \((p<0.01)\) and left eye \((p<0.05)\) ).

\section*{Discussion and Conclusions}

In this paper, we have reported our first experimental results on using the index of cognitive activity (ICA) as a measure of cognitive load in a dual-task scenario. Our analysis results show that the ICA and pupil dilation are rather different measures. They have a very low correlation to each other, and also behave differently: the ICA is very dynamic, while pupil dilation changes only slowly. This observation is particularly interesting, as it indicates that the ICA might be used at higher time resolution than overall pupil dilation.

The distribution of the ICA however also shows that there are some limits as to how incrementally it can be used: When calculating the ICA events at 100 msec intervals, the distribution is not smooth, and there is little bandwidth of distinct events (in our data we observe between 0 and 6 such rapid movements per 100 msec window).

The time series analyses reported here furthermore indicate that the ICA is reflecting the ConTRe task steering events,
while no such effect is detectable in pupil area. The autocorrelation analyses also allowed us to understand more about the delay between stimulus and effect in the ICA: there is a lag of about 1.2 s between the movement of the reference bar and an effect in the ICA, and a lag of about 400 msec between the subject's steering action and the ICA. The fact that the correlation of the ICA and the steering bar is larger than the correlation with the reference bar indicates that the ICA might be related to the participants action execution (as opposed to their perception of the steering task). This is also confirmed by mixed effects regression models with the ICA as a response variable and re-aligned steering bar and reference bar velocity as predictors: the reference bar velocity predictor variable does not improve model fit over models which already include steering bar velocity.

Furthermore, we find that the ICA record reflects our secondary task, language comprehension. In a more detailed analysis, we find that the ICA is significantly higher within the dual task condition whenever the language stimulus is not playing, and that the ICA is high during the ambiguous region of our language stimulus and decreases following disambiguation. We also find a significant effect of our language manipulation, showing that the ICA of the left eye is significantly higher in the object relative clause condition than in the subject relative clause condition.

We also compare the ICA measure to traditional overall pupil dilation and find that our primary tracking task is not reflected in pupil dilation. For our language manipulation, the results in overall pupil dilation are consistent with our findings in the ICA: in the subject relative clause condition, the pupil contracts significantly faster than in the object relative clause condition.

\section*{References}

Aston-Jones, G., \& Cohen, J. D. (2005). An integrative theory of locus coeruleus norepinephrine function: adaptive gain and optimal performance. Annual review of neuroscience, 28, 403-450.
Baayen, R. H., Davidson, D. J., \& Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language.
Bader, M., \& Meng, M. (1999). Subject-object ambiguities in german embedded clauses: An across-the-board comparison. Journal of Psycholinguistic Research, 28(2).
Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. Psychological bulletin, 91(2), 276.
Beatty, J., \& Lucero-Wagoner, B. (2000). The pupillary system.
Demberg, V., Kiagia, E., \& Sayeed, A. (2013). The index of cognitive activity as a measure of linguistic processing. In Proceedings of the 35th annual meeting of the cognitive science society (cogsci-13).
Engelhardt, P. E., Ferreira, F., \& Patsenko, E. G. (2010). Pupillometry reveals processing load during spoken lan-
guage comprehension. Quarterly journal of experimental psychology, 63, 639-645.
Frank, S., \& Thompson, R. (2012). Early effects of word surprisal on pupil size during reading. In Proc. 34th annu. conf. cognitive science society (eds n. miyake, d. peebles \& \(r p\) cooper) (pp. 1554-1559).
Gutirrez, R. S., \& Shapiro, L. P. (2010). Measuring the timecourse of sentence processing with pupillometry. In Cuny conference on human sentence processing.
Hess, E., \& Polt, J. (1960). Pupil size as related to interest value of visual stimuli. Science.
Hess, E., \& Polt, J. (1964). Pupil size in relation to mental activity during simple problem-solving. Science.
Hyönä, J., Tommola, J., \& Alaja, A. (1995). Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. The Quarterly Journal of Experimental Psychology, 48(3), 598-612.
Just, M. A., \& Carpenter, P. A. (1993). The intensity dimension of thought: pupillometric indices of sentence processing. Canadian journal of experimental psychology, 47.
Kahneman, D., \& Beatty, J. (1966). Pupil diameter and load on memory. Science.
Laeng, B., Sirois, S., \& Gredebäck, G. (2012). Pupillometry: a window to the preconscious? Perspectives on psychological science, 7(1), 18-27.
Mahr, A., Feld, M., Moniri, M., \& Math, R. (2012). The contre (Continuous Tracking and Reaction) task: A flexible approach for assessing driver cognitive workload with high sensitivity. In Cognitive load and in-vehicle humanmachine interaction workshop at automotiveui.
Marshall, S. (2000). U.s. patent no. 6,090,051.
Marshall, S. (2002). The index of cognitive activity: Measuring cognitive workload. In proc. 7th conference on human factors and power plants (pp. 7-5).
Marshall, S. (2007). Identifying cognitive state from eye metrics. Aviation, space, and environmental medicine, 78(Supplement 1), B165-B175.
Schluroff, M. (1982). Pupil responses to grammatical complexity of sentences. Brain and language, 17(1), 133-145.
Schröder, M., \& Trouvain, J. (2003). The german text-tospeech synthesis system mary: A tool for research, development and teaching. International Journal of Speech Technology, 6(4), 365-377.
Schwalm, M. (2008). Pupillometrie als methode zur erfassung mentaler beanspruchungen im automotiven kontext. Unpublished doctoral dissertation, Universitätsbibliothek.
Schwalm, M., Keinath, A., \& Zimmer, H. (2008). Pupillometry as a method for measuring mental workload within a simulated driving task. Human Factors for assistance and automation, 1-13.
Wood, S. (2001). mgcv: Gams and generalized ridge regression for r . \(R\) news, \(1(2), 20-25\).
Zellin, M., Pannekamp, A., Toepel, U., \& der Meer, E. (2011). In the eye of the listener: Pupil dilation elucidates discourse processing. Int. Journal of Psychophysiology.

\title{
Integration Costs on Auxiliaries? a self-paced reading study using WebExp
}

\author{
Vera Demberg (vera@coli.uni-saarland.de) \\ Cluster of Excellence, Saarland University, \\ Campus C7.4, 66123 Saarbrücken, Germany
}

\begin{abstract}
In their evaluation of the integration cost component of Dependency Locality Theory on the Dundee Corpus, Demberg and Keller (2008) found no significant main effect of DLT integration cost on reading times, but suggested that this might be due to auxiliaries incurring some of the full verb's integration cost and thus facilitating processing of the verb. This hypothesis however, has to date not been tested. The present paper fills this gap by reporting an experiment on subject vs. object relative clauses including auxiliaries, as well as by testing Demberg and Keller's hypothesis directly on the Dundee Corpus.
\end{abstract}

A further contribution of this paper is methodological: we replicate experimental results on the subject vs. object relative clause assymmetry in a self-paced-reading experiment run remotely on the web using WebExp.
Keywords: Dependency Locality Theory, Relative Clause, Auxiliary, WebExp, Dundee Corpus, Self-paced reading, Eyetracking

\section*{Introduction}

Dependency Locality Theory (DLT), proposed by (Gibson, 1998; Gibson \& Pearlmutter, 2000) is a theory of sentence processing which has received quite a lot of attention in the field of psycholinguistics, and has been argued to explain a range of phenomena including including the SRC/ORC processing difficulty asymmetry, difficulty of centre embeddings, cases of processing breakdown, filler-gap dependencies, heavy NP shift and extraposition.
(Demberg \& Keller, 2008) evaluated the integration cost component of DLT on an eye-tracked corpus of newspaper articles (Dundee Corpus; Kennedy and Pynte, 2005), and found that verbs which were preceded by nouns were read more slowly than verbs which were preceded by both auxiliaries and nouns. Demberg and Keller thus hypothesized that integration costs might not be incurred at the main verb (as predicted by DLT), but at the auxiliary, at which it should thus be possible to observe an integration cost effect.

They did however not test whether such an integration cost effect could indeed be detected on the auxiliaries. This paper fills this gap through two studies that test Demberg and Keller's hypothesis: a self-paced reading experiment of matched and controlled subject and object relative clauses containing auxiliaries, as well as a corpus study analysing the auxiliaries across various syntactic constructions in the Dundee Corpus for an integration cost effect.

\section*{Background and Related Work}

\section*{Dependency Locality Theory}

An important component for quantifying processing difficulty in DLT is the so-called "integration cost". Integration cost
(IC) measures the distance between a head its dependent in terms of new discourse referents (DR) that occur inbetween them. Figure 1 shows the dependencies for a subject relative clause (SRC) and an object relative clause (ORC). In the example, discourse referents are marked as either 0 (no new discourse referent) or 1 (new discourse referent). The dependency edge between reporter and the main clause verb admitted is annotated with " +2 " to express that two discourse referents occur between these words (namely, attacked and senator). The interesting case when comparing the subject and object relative clause is the embedded verb attacked: the integration cost is 1 in the SRC case (just the cost of constructing the discourse referent), while the it is 3 in the case of the object relative clause. There is a cost of 1 for constructing the discourse referent, plus a cost of 2 for integrating the relative pronoun who at the trace \({ }^{*} t^{*}\), at which point two discourse referents (senator and attacked) intervene. There are also integrations of senator and attacked and the trace and attacked, but no new discourse referents occur between them, so these integrations are cost-free.


Figure 1: Dependency Locality Theory Integration Cost.

\section*{Auxiliaries in DLT}

Most previous experimental studies on locality effects do not contain auxiliaries. An exception is Experiment 4 from (Warren \& Gibson, 2002), which compares self-paced reading times of an object relative clause with a full or pronominal embedded subject NP vs. a complementizer clause with full or pronominal subject NP, see example (1).
a. Relative clause: The woman who you/the boy had accidentally pushed off the sidewalk got upset and decided to report the incident to the policeman standing nearby.
b. Complement clause: The woman knew that you/the boy had accidentally pushed the girl

\section*{but gave him/you a long lecture anyway.}

Warren and Gibson found that the reading times were longest in the full NP in ORC condition, which is also the one in which highest integration costs are expected. On the auxiliary, the difference between the ORC with full NP and the ORC with pronoun is significant. While this would be in line with an integration effect already on the ORC auxiliary, Warren and Gibson also point out the possibility that the longer reading times on the auxiliary might be a spill-over effect.

The present study seeks to directly investigate whether integration costs can be measured on auxiliaries.

\section*{Result from Demberg and Keller, 2008}

Demberg and Keller's (2008) evaluation of Dependency Locality Theory on the Dundee Corpus showed that there was no general positive correlation between DLT integration costs and reading times. They however looked at integration cost at verbs in more detail, and found that verbs which integrate an auxiliary and a nominal dependent exhibit a reduced estimated reading time compared to verbs that only integrate a nominal dependent, while there seemed to be an overall effect of increased reading time at verbs with more nominal dependents, see Table 1.

Demberg and Keller therefore suggested that the relevant integration cost might not be incurred at the main verb, but at the auxiliary itself, which might integrate nominal dependents and thus incur a non-zero integration cost (DLT assume that auxiliaries are cost-free). When the auxiliary would then

Table 1: First pass durations for verbs (with non-zero integration cost) in the Dundee Corpus: coefficients for verbs grouped by verbal dependents ( \(N>20\) ) and their significance levels for a model fitted on residual reading times (with respect to a model including other predictors known to influence reading times). Abbreviations in the table refer to part of speech tags used by the Penn Treebank annotation: AUX: auxiliary, PRP: personal pronoun, NN: singular or mass noun, NNP: proper noun, singular, RP: particle, MD: modal, NNS: plural noun, RB: adverb, AUXG: auxiliary present participle, TO: preposition to, JJ: adjective, VBP: non-third person singular present verb. Table from Demberg and Keller (2008).
\begin{tabular}{lrcr}
\hline Dependents & Coeff & Signif & N \\
\hline NNP-AUX-AUX & -62.41 & \(* *\) & 21 \\
NNS-AUX-AUX & -35.65 & \(*\) & 57 \\
NNS-MD-AUX & -30.75 & \(* *\) & 110 \\
PRP-AUX-PRP-AUX & -29.72 & \(* * *\) & 184 \\
NN-MD-AUX & -25.35 & \(* *\) & 153 \\
PRP-AUX & -22.64 & \(* * *\) & 700 \\
PRP-AUX-RB & -21.75 & \(*\) & 133 \\
AUXG & -20.26 & \(*\) & 121 \\
NNP-AUX & -19.05 & \(* *\) & 301 \\
TO-PRP & -16.97 & \(* * *\) & 723 \\
NNP & 12.01 & \(* *\) & 1372 \\
NN-RB & 22.26 & \(*\) & 127 \\
NN-NNS & 76.43 & \(* * *\) & 25 \\
PRP-MD-PRP-MD-JJ & 105.4 & \(*\) & 65 \\
\hline
\end{tabular}
be integrated with the main verb, it would facilitate integration (hence the negative coefficient), as the main work of the integration of the nominal dependents has already happened at the auxiliary. They also point out that this explanation is compatible with syntactic theories such as Head-driven Phrase Structure Grammar (Pollard \& Sag, 1994), which assume that auxiliaries inherit the subcategorization frame of the main verb, and that dependents are unified (integrated) into the subcategorization frame at the auxiliary. Demberg and Keller did however not test this hypothesis in their study, so the contribution of this paper is to fill this gap and test both in a controlled experiment and on the Dundee corpus whether the hypothesis that the verb's integration cost can be measured on the auxiliary is true.

\section*{Experiment: Auxiliaries in Relative Clauses}

As a first experiment, we chose to use a strictly controlled experimental setting in which we compare the processing of subject vs. object relative clauses including auxiliaries. The processing difference in subject vs. object relative clauses is well-established: Object relative clauses (as in (2-b)) are more difficult to process than subject relative clauses (1a) (King \& Just, 1991), with increased reading times on the ORC embedded verb as opposed to the SRC embedded verb (Staub, 2010). Dependency Locality Theory (DLT; Gibson, 2000) accounts for this effect in terms of long-distance dependencies, see the explanation of this case in Section on Dependency Locality Theory.

We created 24 subject and object relative clauses with auxiliaries preceding the embedded verb, based on the experimental items from (Staub, 2010), see (2).
a. The mathematician who [aUX had] [v visited] [ NP the chairman] found a solution to the problem.
b. The mathematician who [NP the chairman] [AUX had] [ v visited] found a solution to the problem.

\section*{Data Collection}

We ran a self-paced reading experiment with 126 participants online, using WebExp, www.webexp. info (Keller, 1999), an experimental software that carries out psychological experiments over the internet. Keller, Gunasekharan, Mayo, and Corley (2009) demonstrate that response times collected with WebExp are sufficiently accurate to conduct reaction time experiments over the internet. Experiment 2 from (Keller et al., 2009) replicates results from a lab-based phrase-by-phrase self-paced reading experiment using the WebExp software.

Participants were recruited using Amazon MechanicalTurk, which we used to create HITs linking to the WebExp experiment. In order to encourage participants to complete the whole WebExp experiment, the HIT also contained a field that required participants to fill in a password which was provided on the last screen of the WebExp experiment. We restricted the HITs to workers who were based in the USA, and
have a HIT approval rate \(\geq 80 \%\). In the instructions, we additionally required workers to only participate if they were native speakers of English. We successfully collected data from 126 participants \({ }^{1}\) (approx. 60 per condition). Following recommendations in (Keller et al., 2009), we only allowed workers with a Windows or Linux operating system.
The experiment was programmed as a word-by-word selfpaced reading experiment. Due to a limitation of WebExp, each sentence within a set of sentences to be randomized has to contain the same number of words. In addition to inserting auxiliaries, we therefore edited the items from (Staub, 2010) to conform to this format by adapting the length of the region following the relative clause.

\section*{Mixed-effects modelling}

In order to test whether our manipulation of relative clause type has an effect on the reading times on the auxiliaries, we use linear mixed effects models from the R lme4 package (R. Baayen, 2008; R. H. Baayen, Davidson, \& Bates, 2008). This type of model can be thought of as a generalization of linear regression that allows the inclusion of random factors as well as fixed factors. We treat subjects and items as a random factors, which means that our models contain an intercept term for each subject and each item, representing the individual differences among the subjects and differences between our items. Furthermore, we include random slopes under both subject and item for our predictor (relative clause type), essentially accounting for idiosyncrasies of a participant or item with respect to the predictor, such that only the part of the variance that is common to all participants and all items is attributed to the main effect for our predictor.

We excluded as outliers any reading times shorter than 100 msec or longer than 1000 msec .

\section*{Results}

We found a significant effect of relative clause type on the auxiliary (AUX), the embedded verb (VB) and the determiner (DT), but not on the noun (NN) or the sentence's main verb (VBM), see Figure 2.

Auxiliaries and verbs were read significantly faster in the subject relative clause condition than in the object relative clause condition. We furthermore found that SRC determiners were read more slowly than ORC determiners; a similar effect was found on the noun region in early reading time measures in Staub (2010), see Figure 3. Differences in terms of the location of the effect may be due to differences in selfpaced reading vs. eye-tracking.

The faster reading times on the verb of the subject relative clause are in line with DLT integration cost, while the large and significant effect on the auxiliary seems to support also

\footnotetext{
\({ }^{1}\) There were some problems with WebExp, which sometimes failed to correctly transfer the collected data to our server. This resulted in loss of about \(30 \%\) of our data. This problem had been observed by others and reported earlier; it appears to be independent of the operative system and browser used by workers. Workers whose data failed to transfer were paid normal rates regardless.
}


Figure 2: Coefficients for relative clause condition shown as a bar plot, a negative value at the auxiliary means that the auxiliary was read faster in the subject relative clause condition than in the object relative clause condition.


Figure 3: Experimental results from Staub, 2010 using the same materials without auxiliaries in an eye-tracking study.
an account of an integration cost effect being observable at the auxiliary. There is however a confound in such subject vs. object relative clause stimuli: the regions preceding the auxiliary (or the verb, in other studies) differ in that the SCR auxiliary is preceded by the relative pronoun, while the ORC pronoun is preceded by a noun.

We therefore ran a second analysis which takes into account spill-over effects. We approximate the spill-over effect that a word can cause by calculating with mixed-effects models the effect of one word's log frequency on the following word's reading times \({ }^{2}\). In order to obtain an accurate estimate of typical spill-over effects, we ran this estimation on all words in our experiment, including fillers. used residual reading times for each of the words in our relative clause ma-

\footnotetext{
\({ }^{2}\) While it would be possible to approximate the spill-over effect given not only the previous word but also the word before that, note that for our stimuli, only the previous word can plausibly explain the longer reading times on the ORC auxiliary. (Word aux-2 in the ORC is shorter and more frequent than word aux -2 in the SRC.)
}


Figure 4: When taking spill-over effects into account, the effect on the auxiliary disappeared.
terials as a response variable and re-ran the earlier mixedeffects analysis on these residual reading times. The results then looked rather different: the facilitation effect on SRC auxiliaries, and difficulty effect on SRC determiners disappears when we account for spill-over effect, see Figure 4.

We furthermore find a significant effect of relative clause type on the reading times of the main verb, indicating that the main verb in the subject relative clause condition was read more slowly than the main verb in the object relative clause condition. This effect is puzzling as previous studies on selfpaced reading rather seem to suggest a tendency for longer reading times on the ORC main verb, while reading times in Staub's eye-tracking experiment were virtually identical.

\section*{Discussion}

We find that increased reading times on auxiliaries can be explained in terms of spill-over effects, thus not supporting the hypothesis of (Demberg \& Keller, 2008), who suggested that integration costs might occur at auxiliaries and facilitate integration at the verb. Instead, our findings support the original predictions of Dependency Locality Theory (Gibson, 2000). This result calls for a closer investigation of the Dundee corpus data to directly examine whether any integration cost effect is associated with auxiliaries in that data, see our Experiment 2.

In terms of experimental methodology, we provide evidence for the validity of self-paced reading using WebExp, by replicating the established relative clause asymmetry result on the embedded verb. Initial evidence that timing using WebExp is sufficiently accurate for self-paced reading studies is presented in (Keller et al., 2009). To the best of our knowledge, the present results are the first ones for word-by-word SPR using WebExp, as previous studies used much larger regions.

\section*{Corpus Study: Auxiliaries in Dundee Corpus}

\section*{Data}

For our data analysis, we used the Dundee Corpus (Kennedy \& Pynte, 2005), an English language eye-tracking corpus
based on text from The Independent newspaper. The texts contain about 51,000 words and were read by 10 native speakers of English. The text was presented on a computer screen, five lines at a time at a line length of 80 characters.

Since the corpus data is not syntactically annotated, we parsed the entire corpus with the Stanford parser (Klein \& Manning, 2003; De Marneffe, MacCartney, \& Manning, 2006), which generates both a phrase structure parse and a dependency representation. We calculated DLT integration cost based on the top-ranked dependency output for each sentence. We evaluated our integration cost implementation using a short text that had been hand-annotated with integration cost values (Wu, Bachrach, Cardenas, \& Schuler, 2010). This evaluation gives us an estimate of how well our automatic annotation tool performs, and also enables us to evaluate our new implementation based on the Stanford parser with an older implementation (Demberg \& Keller, 2008) which was based on the MiniPar parser (Lin, 1998), see Table 2.

Table 2: Evaluation on a text of 770 words, manually annotated with integration costs.
\begin{tabular}{l|c|c} 
Parser & \% correct IC & \begin{tabular}{l} 
correlation (Kendall) \\
to manual annotation
\end{tabular} \\
\hline MINIPAR & \(83 \%\) & \(\tau=0.77, p \ll 0.001\) \\
\hline Stanford & \(89 \%\) & \(\tau=0.84, p \ll 0.001\)
\end{tabular}

We then automatically aligned each auxiliary with the automatically determined integration cost calculated for its governing main verb, in order to measure whether any effect of increased integration cost at the verb might be measurable on the auxiliary. In order to decrease noise in the data set, we excluded any cases in which further discourse referents occurred between the auxiliary and the verb, or where no verb could be found within a window of three words after the auxiliary. We furthermore excluded contractions (e.g. we'll).

\section*{Methods}

We analysed the data using linear mixed effects models. Because the corpus data are not as closely controlled as the experimental data from the first study, and because the methodology differs (eye-tracking here vs. self-paced reading in experiment 1), we run mixed-effects models with a larger range of different predictors, including the length of a word in characters WordLength, its log frequency WordFreq, a flag indicating whether the previous word was fixated PrevFix, the frequency of the preceding word to account for spill-over effects PrevFreq, forward and backward transitional probabilities ForwTransProb and BackwTransProb, the word number within the sentence WordNo, the fixation landing position in relation to word length LandPos, the launch distance of the saccade LaunchDist, the surprisal \({ }^{3}\) at the word Surprisal as well as the verb's integration cost IntegCost.

\footnotetext{
\({ }^{3}\) Surprisal was calculated using the Roark parser (Roark, Bachrach, Cardenas, \& Pallier, 2009).
}

As response variables, we use four different measures of reading times: first fixation duration (duration of the first fixation, if any on the first pass through the sentence from left to right), first pass duration (sum of duration of fixations during first pass reading on a word before leaving the word), total duration (sum of the durations of all fixations on a word) and go past times (time spent between the first fixation in first pass reading on a word and first leaving it to the right).

We only analysed auxiliaries which had received at least one fixation. Before fitting LME models, we applied outlier removal: we computed the mean reading time (over all items and participants), and then removed all data points which deviated more than two standard deviations from the mean. Outliers can affect the results of analyses on the Dundee corpus, as (Roland, Mauner, OMeara, \& Yun, 2012) show. Furthermore, this way of trimming the data also reduces the long tail of the reading time distribution, resulting in a distribution that is closer to normal. This left us with 1257 data points.

\section*{Results}

None of our models support the hypothesis that higher integration costs at the verb increase reading times at the auxiliary preceding the verb. In none of the reading time models did the verb's integration cost come out as a significant predictor of reading times on the auxiliary, see for example the best fitting regression model for first pass times in Table 3. When we add the verb's integration cost as a predictor, model fit is not significantly improved ( \(p \approx 0.75\) ), and we even get a negative coefficient ( \(\beta=-0.33, t=-0.293\) ), so there is not much reason to believe at this point that the failure of finding a positive significant effect would simply be due to an insufficient number of data points.

Table 3: Final model for first pass times on auxiliaries, showing that reading times are longer when word length increases, and shorter when the previous word was fixated or was a highly frequent word.
\begin{tabular}{l|rrrc} 
& & & & \\
Predictor & Coeff & Std. Err & t value & Signif \\
\hline (Intercept) & 230.52 & 9.44 & 24.40 & \(* * *\) \\
WordLength & 5.46 & 2.40 & 2.27 & \(*\) \\
PrevFreq & -5.18 & 1.50 & -3.45 & \(* *\) \\
PrevFix & -25.19 & 4.04 & -6.23 & \(* * *\) \\
LandPos & -11.98 & 15.02 & -0.79 & \\
Surprisal & 2.00 & 1.33 & 1.50 & \\
WLen:LandP & -9.02 & 6.13 & -1.47 &
\end{tabular}

To more closely inspect the relationship between DLT integration cost and reading times and understand where the negative coefficient comes from, we ran a generalized additive model with a spline ( \(k=30\) ) for the verb's integration cost as a predictor and the auxiliaries' first pass times as the response variable. As can be seen in Figure 5, there is indeed a negative trend for increasing integration cost, at least for auxiliaries with the most common integration cost values one to
three. (An integration cost value of zero can only occur if the main verb of the sentence is a copula.)

Integration Cost and First Pass Times on Auxiliaries


Figure 5: Spline plot for the verb's integration cost fitting the first pass times on the auxiliary.

\section*{Overall Discussion}

Neither the results from the experiment nor the results from the corpus study support the hypothesis suggested in (Demberg \& Keller, 2008), that integration may already happen at the auxiliary and costs of such an integration would be measurable in reading times. The results from the corpus study are in line with the results from the closely controlled relative clause experiment. While the experiment compared auxiliaries in a specific syntactic construction, the corpus study complements the first experiment in that it includes auxiliaries from many different syntactic constructions. It also shows that the result can be replicated for contextualized reading in more naturalistic conditions.

In both studies, we took care to account for spill-over effects from previous words, and such effects indeed turned out to be important in both studies: in the relative clause study, the interpretation of results changes completely when taking into account spill-over effects, and in the corpus study, the variables capturing overspill effects significantly improve model fit.

Getting back to the fundamental question underlying these studies, these results lead to the following hypotheses about integration cost at auxiliaries:
a) Auxiliaries help integration at the verb in a way which is not directly reflected in their reading times.
b) Auxiliaries do in fact not help integration at the verb, and DLT integration cost, despite showing stable effects in experimental settings, does hence not have much general explanatory power for data of properly contextualized utterances such as those occurring in a corpus.
c) The integration cost estimate using automatic parsing tools is not accurate enough (in particular due to shortcomings in dealing with traces).

The last concern is unlikely to be valid, however, as it only applies to the corpus study, and doesn't explain failure to find an effect in the first experiment. Also, we have taken great care to re-implement integration cost based on a state-of-theart parser, and have used heuristics to account for traces.

Question a) could be addressed experimentally by manipulating constructions with respect to the presence of an auxiliary (the effects of spill-over could be diminuished by using an adverb before the verb such that the word immediately before the verb is always the same).

\section*{Conclusions}

The contribution of this paper is two-fold. We find that increased reading times on auxiliaries in subject vs. object relative clauses can be explained in terms of spill-over effects, thus do not support the hypothesis of (Demberg \& Keller, 2008), who suggested that integration costs might occur at auxiliaries and facilitate integration at the verb. Instead, our findings support the original predictions of Dependency Locality Theory (Gibson, 2000). These results are further supported by a corpus study of auxiliaries from the Dundee Corpus. We can not find any significant effect of the verb's integration cost on reading times at the auxiliary in any of our reading time measures. Given our evidence about the lack of any detectable integration cost effect on auxiliaries, we can no longer explain away the lack of an overall positive effect of integration costs on the verbs from the Dundee corpus as being due to the presence of auxiliaries.

Our second contribution is methodological: the self-paced reading study (Experiment 1) provides evidence for the validity of word-by-word self-paced reading via WebExp, by replicating the established relative clause asymmetry result on the embedded verb. While initial evidence that timing using WebExp is sufficiently accurate for self-paced reading studies is presented in (Keller et al., 2009), the present results are, to the best of our knowledge, the first ones for word-by-word self-paced-reading using WebExp.

\section*{Acknowledgments}

I want to thank Silas Weinbach for his contributions to this work in programming the WebExp and Mechanical Turk experiment, as well as Jesus Calvillo and Asad Sayeed for their contributions to the new integration cost component based on the Stanford parser.

\section*{References}

Baayen, R. (2008). Analyzing linguistic data: A practical introduction to statistics using \(R\). Cambridge Univ Pr.
Baayen, R. H., Davidson, D. J., \& Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language.
De Marneffe, M., MacCartney, B., \& Manning, C. (2006). Generating typed dependency parses from phrase structure parses. In Proceedings of lrec (Vol. 6, pp. 449-454).

Demberg, V., \& Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. Cognition, 109, 193-210.
Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. Cognition, 68, 1-76.
Gibson, E. (2000). Dependency locality theory: A distancedased theory of linguistic complexity. In A. Marantz, Y. Miyashita, \& W. O'Neil (Eds.), Image, language, brain: Papers from the first mind articulation project symposium (pp. 95-126). Cambridge, MA: MIT Press.
Gibson, E., \& Pearlmutter, N. (2000). Distinguishing serial and parallel parsing. Journal of Psycholinguistic Research, 29(2), 231-240.
Keller, F. (1999). Webexp: A java toolbox for web-based psychological experiments. University of Edinburgh, Human Communication Research Centre.
Keller, F., Gunasekharan, S., Mayo, N., \& Corley, M. (2009). Timing accuracy of web experiments: A case study using the webexp software package. Behavior Research Methods, 41(1), 1-12.
Kennedy, A., \& Pynte, J. (2005). Parafoveal-on-foveal effects in normal reading. Vision Research, 45, 153-168.
King, J., \& Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. Journal of Memory and Language, 30, 580-602.
Klein, D., \& Manning, C. (2003). Accurate unlexicalized parsing. In Proceedings of the 41st annual meeting on association for computational linguistics-volume 1 (pp. 423430).

Lin, D. (1998). An information-theoretic definition of similarity. In J. W. Shavlik (Ed.), Proceedings of the 15th international conference on machine learning (pp. 296-304). Madison, WI: Morgan Kaufmann.
Pollard, C., \& Sag, I. A. (1994). Head-driven phrase structure grammar. Chicago: University of Chicago Press.
Roark, B., Bachrach, A., Cardenas, C., \& Pallier, C. (2009, August). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In Proceedings of the 2009 conference on empirical methods in natural language processing (pp. 324-333). Singapore.
Roland, D., Mauner, G., OMeara, C., \& Yun, H. (2012). Discourse expectations and relative clause processing. Journal of Memory and Language.
Staub, A. (2010). Eye movements and processing difficulty in object relative clauses. Cognition, 116, 71-86.
Warren, T., \& Gibson, E. (2002). The influence of referential processing on sentence complexity. Cognition, \(85(1), 79-\) 112.

Wu, S., Bachrach, A., Cardenas, C., \& Schuler, W. (2010, July). Complexity metrics in an incremental right-corner parser. In Proceedings of the 48th annual meeting of the association for computational linguistics (pp. 1189-1198). Uppsala, Sweden.

\title{
Effects of Training on Category Learning
}

\author{
Wei (Sophia) Deng (deng.69@osu.edu) \\ Department of Psychology, The Ohio State University \\ 267 Psychology Building, 1835 Neil Avenue \\ Columbus, OH 43210 USA
}

\author{
Vladimir M. Sloutsky (sloutsky.1@osu.edu) \\ Department of Psychology, The Ohio State University \\ 239 Psychology Building, 1835 Neil Avenue \\ Columbus, OH 43210 USA
}

\begin{abstract}
What information do people extract in the course of category learning? And how does training affect this process? The current study addressed these questions by examining the effects of training on the outcome of category learning in 4to 5 -year-olds and adults. In two experiments, participants were trained on either a classification task or an inference task and then tested with categorization and recognition tasks. The categorical information (i.e., deterministic and probabilistic features) was explicitly given to participants in Experiment 1 but not in Experiment 2. Results with adults replicate previous findings indicating that participants form different representations in the course of classification and inference training (rule-based representation in the former case and similarity-based representations in the latter case). In contrast, regardless of the type of training, young children form similarity-based representations.
\end{abstract}

Keywords: Cognitive Development, Categorization, Learning, Psychology.

\section*{Introduction}

The ability to form categories is an important component of human cognition (see Murphy, 2002, for a review). It has been well established that this ability appears early in development, with young infants capable of forming categories (Eimas \& Quinn, 1994; Oakes, Madole, \& Cohen, 1991). The study of how categories are learned and used can elucidate "a single main theme to cognitive science - the question of how people come to have knowledge" (Murphy, 2002, p. 272).

The relationship between category learning and use can be examined by contrasting two of the fundamental functions of categories - classification and inference (E. Smith, 1994). To test theories of categorization, researchers developed a variety of tasks (see A. Markman \& Ross, 2003, for a review), most of which are based on classification. In a typical classification learning task, participants are presented with stimuli, whose category membership is unknown, and are asked to predict a category each item belongs to. This situation is similar to that of sorting a set of squirrels and hamsters into two distinct groups. Whereas classification involves predicting the category of an item, inference involves predicting a missing feature using information from other features as well as the category. In this case, instead of determining
whether an animal is a squirrel or a hamster, participants predict a value of a given feature (e.g., the type of tail the animal has).

There is evidence that classification and inference learners result in different representations of categories and much of these findings stem from a paradigm developed by Yamauchi and A. Markman (1998). The paradigm is based on the following idea. Imagine two categories A (labeled "A") and B (labeled "B"), each having four binary dimensions (e.g., Size: large vs. small, Color: black vs. white, Shape: square vs. circle, and Texture: smooth vs. rough). The prototype of Category A has all values denoted by " 1 " (i.e., "A", 1, 1, 1, 1) and the prototype of Category \(B\) has all values denoted by " 0 " (i.e., "B", \(0,0,0,0\) ). There are two inter-related generalization tasks - classification and inference. The goal of classification is to infer category membership (and hence the label) on the basis of presented features. For example, participants are presented with all the values for an item (e.g., ?, \(0,1,1,1\) ) and have to predict category label "A" or "B". In contrast, in the inference task participants have to infer a feature on the basis of category label and other presented features. For example, given an item (e.g., "A", 1, ?, 1, 0), participants have to predict the value of the missing feature. It was found that inference learners were more likely than classification learners to infer prototypical features which were correspondingly associated with training items. Multiple studies using this paradigm found that classification learners are sensitive primarily to diagnostic features that distinguish between categories, whereas inference learners are also sensitive to within-category correlations of features, which are not diagnostic but prototypical (Chin-Parker \& Ross, 2002; Chin-Parker \& Ross, 2004; Sakamoto \& Love, 2006; Yamauchi, Love, \& A. Markman, 2002; Yamauchi \& A. Markman, 2000a, 2000b). Furthermore, there is also evidence that adults trained on a classification task attend to the most relevant features (A. L. Anderson et al., 2002). Adults learn to optimize performance in category learning by shifting their attention to different diagnostic features in different situations (Nosofsky, 1984; Rehder \& Hoffman, 2005; Shepard et al., 1961) or learn inattention to newly relevant features (Hoffman \& Rehder, 2010).
The argument that classification learning focuses on the diagnostic features distinguishing categories whereas
inference learning focuses on the prototypical features reflecting within-category information is consistent with the evidence that adults' categorization is often rule-based (Rips, 1989; Allen \& Brooks, 1991). Nosofsky and colleagues (Nosofsky \& Palmeri, 1998; Nosofsky, Palmeri, McKinley, 1994) have proposed a quantitative model of human concept learning that learns to classify objects by forming simple logical rules and remembering occasional exceptions to those rules.

However, there is little agreement on the categorization process in early development. According to knowledgebased approaches, early in development, categorization and inductive generalization is considered to be based on prior categorical knowledge thus to be category based (Gelman \& E. Markman, 1986; Gelman \& Heyman, 1999; Gelman 2004). According to another approach, early categorization is similarity-based (Sloutsky \& Fisher, 2004; Sloutsky, Kloos, \& Fisher, 2007). There is evidence that early generalization is often driven by appearance similarity (Gelman, 1988; Gelman \& E. Markman, 1987; Sloutsky \& Fisher, 2004). In particular, infants are more likely to group items together if the items have overlapping within-group distributions of properties and non-overlapping betweengroup distributions (French, et al., 2004, see also Mareschal \& Quinn, 2001; Mareschal, Quinn, \& French, 2002). Similarly, infants are more likely to generalize non-obvious properties when the two items look alike (Welder \& Graham, 2001) and they are more likely to extend a name to items that have similar shape (E. Markman \& Hutchinson, 1984; L. Smith, et al., 1996). Similar results have been reported with young children, with similarity supporting both categorization of items and induction of non-obvious properties (Gelman, 1988; Gelman \& E. Markman, 1987; Sloutsky \& Fisher, 2004; Sloutsky \& Lo, 1999).

Do children and adults show the same pattern of extracting and processing categorical information in category learning? How does training affect this process? And how do these effects change in the course of development? Does the asymmetry between classification and inference learning found in adults exist in children? Finding such an asymmetry would support the idea that, like adults, children treat classification and inference learning differently and tend to detect and rely on a defining feature to categorize items, whereas a symmetric performance in the classification and inference training would support the idea that children may perform similarity-based categorization and treat two types of category learning equally. The primary goal of this study is to address these questions.

\section*{Overview of the Current Study}

Experiments reported here explored how categories were learned and used under classification and inference training by adults and young children. The basic task consisted of two phases, a training phase and a testing phase. During the training phase, participants had to infer either the category
of a given item (in classification training) or a feature that the item has (in inference training). The testing phase consisted of categorization and recognition tasks and was administered immediately after the training phase. During the testing phase, which was identical for two training conditions, adult and child participants were asked to determine (1) which category the creature was more likely to belong to and (2) whether each picture was old or new. The structures of both training and testing stimuli will be described in the section below.

\section*{Experiment 1}

\section*{Method}

Participants There were 35 adults (16 women) and 21 preschool children ( \(M=56.6\) months, range 53.2-59.5 months; 13 girls) participating in this experiment. In this and the second experiment reported here, adult participants were undergraduate students from the Ohio State University participating for course credit and were tested in a quiet room in our lab on campus. Child participants were recruited from childcare centers, located in middle-class suburbs of Columbus and were tested in a quiet room in their preschool by a female experimenter.
Materials In both experiments reported here, the materials, similar to those used previously by Deng and Sloutsky (2012, 2013), consisted of colorful drawings of artificial creatures that varied in their appearance and in a categoryinclusion rule and that were accompanied by the novel labels "flurp" (Category F) and "jalet" (Category J). For these two categories, we created two prototypes (F0 and J0, respectively) that were distinct in the color and shape of seven of their features: head, body, hands, feet, antennae, tail, and button (see Figure 1). Two categories have a family-resemblance structure and stimuli were derived from the two prototypes by modifying the values of the seven features. The button is the deterministic feature (hereafter "D") and defines the category-inclusion rule: all members of Category F have raindrop-shaped button with the value of 1 whereas all members of Category \(J\) have cross-shaped button with the value of 0 . All the other varying features - the head, body, hands, feet, antennae, and tail - constitute the probabilistic features (hereafter " P ") and reflect the overall similarity among the exemplars.
The training stimuli consisted of High-Match items (i.e., \(\mathrm{P}_{\text {flurp }} \mathrm{D}_{\text {flurp }}\) and \(\mathrm{P}_{\text {jalet }} \mathrm{D}_{\text {jalet }}\) ). All members of \(\mathrm{P}_{\text {flurp }} \mathrm{D}_{\text {flurp }}\) items had four probabilistic features ( P ) consistent with the prototype F0 with the value of 1 and two features consistent with the prototype J0 with the value of 0 . And all of them have the deterministic feature (D) consistent with F0 valued 1. However, all members of \(\mathrm{P}_{\text {jalet }} \mathrm{D}_{\text {jalet }}\) items had four probabilistic features consistent with the prototype J0 with the value of 0 and two features consistent with the prototype F0 with the value of 1 . And all of them have the deterministic feature consistent with J 0 valued 0 .
The testing stimuli consisted of another four sets of items besides the High-Match items. The data analyses reported

Table 1. Example of category structure used in Experiment 1 and Experiment 2.
\begin{tabular}{|c|c|c|c|c|c|c|c|c|c|c|c|c|c|c|c|}
\hline \multicolumn{8}{|c|}{Category F} & \multicolumn{8}{|c|}{Category J} \\
\hline & Head & Body & Hands & Feet & Antenna & Tail & Button & & Head & Body & Hands & Feet & Antenna & Tail & Button \\
\hline F0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & J0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline \(\mathrm{P}_{\text {fiur }} \mathrm{Dffurp}\) & 1 & 1 & 1 & 1 & 0 & 0 & 1 & \(\mathrm{P}_{\text {jalet }} \mathrm{D}_{\text {jalet }}\) & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\
\hline \(\mathrm{P}_{\text {jalet }} \mathrm{D}_{\text {flurp }}\) & 0 & 1 & 0 & 1 & 0 & 0 & 1 & \(\mathrm{P}_{\text {flur }} \mathrm{D}_{\text {jalet }}\) & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\
\hline \(\mathrm{P}_{\text {all-new }} \mathrm{D}_{\text {furp }}\) & N & N & N & N & N & N & 1 & \(\mathrm{P}_{\text {all-new }} \mathrm{D}_{\text {jalet }}\) & N & N & N & N & N & N & 0 \\
\hline
\end{tabular}

Note. The value 1 = any of seven dimensions identical to Category F (flurp, see Figure 1). The value \(0=\) any of seven dimensions identical to Category J (jalet, see Figure 1). The value \(\mathrm{N}=\) new feature which is not presented during training. \(\mathrm{P}=\) probabilistic feature; \(\mathrm{D}=\) deterministic feature. F0 is the prototype of Category F and J0 is the prototype of Category J.


Figure 1. Examples of Stimuli Used in this study.
here only focused on two of them: critical lures (i.e., \(\mathrm{P}_{\text {jalet }} \mathrm{D}_{\text {flurp }}\) and \(\mathrm{P}_{\text {flurp }} \mathrm{D}_{\text {jalet }}\) ), and all-new-P items (i.e., \(\mathrm{P}_{\text {all- }}\) \({ }_{\text {new }} \mathrm{D}_{\text {flurp }}\) and \(\left.\mathrm{P}_{\text {all-new }} \mathrm{D}_{\text {jalet }}\right)\). The High-Match items were used to examine participants' performance of category learning and to assess their recognition accuracy on the old items. Children were above \(91.1 \%\) categorization accuracy on these trials, and adults were above \(95.3 \%\), all above chance ( \(p \mathrm{~s}<.05\) ) and exhibited memory accuracy of \(82.2 \%\) and \(94.5 \%\), respectively. The all-new-P items were catch trials and consisted of six new probabilistic features which were not shown during training. Children and adults exhibited memory accuracy of \(91.0 \%\) and \(98.1 \%\), respectively. The critical lures were Low-Match items: Most of the members of \(P_{\text {jalet }} D_{\text {flurp }}\) items have the \(P\) with the value of 0 but all of them have D valued 1 ; whereas most of the members of \(\mathrm{P}_{\text {flurp }} \mathrm{D}_{\text {jalet }}\) items have P with the value of 1 but all of them have D valued 0 . This set of items was used to assess whether participants relied on overall similarity or category-inclusion rule to categorize new items. Table 1 shows example of category structure with P and D being combined to create three types of stimuli, and Figure 1 shows examples of each kind of stimulus.
Procedure The procedure consisted of two phases, a training phase and a testing phase. During the training phase, participants were given 30 trials (15 trials per category) and they had to infer either the category of a
given item (in classification training) or a feature that the item has (in inference training). Each training trial was accompanied by corrective feedback. The classification and inference training differed in the type of dimensions being predicted. In classification training, participants predicted the category label of a stimulus given information about all other features. In inference training, participants predicted one missing feature of a stimulus given the information about the remaining features as well as the label. The information about P and D was explicitly given to participants before training. They were told that all flurps (or jalets) had raindrop button (or cross button) and most of them had flurps' (or jalets') \(P\) by presenting corresponding probabilistic features one at a time. This information was repeated in the corrective feedback to each response during training. Adult and child participants were randomly assigned to one of the two training conditions. The testing tasks were not mentioned in the training phase of any of the conditions.
Testing phase, including categorization and recognition tasks, was administered immediately after training. During the testing phase, which was identical for two training conditions, adult and child participants were presented with 40 trials of creatures and were asked to determine (1) which category the creature was more likely to belong to and (2) whether each picture was old (i.e., exactly the one
presented during the training phase) or new. The order of the 40 items was randomized across participants. No feedback was provided during the testing phase.

The procedures were identical for both adult and child participants except the way the instructions were presented and the questions were asked. Adult participants read the instructions and questions on the computer screen and pressed the keyboard to make responses, whereas for children instructions as well as questions were presented verbally by a trained experimenter and the experimenter recorded children's responses by pressing the keyboard. The proportion of responses in accordance with the category from which the exemplar was derived (i.e., rulebased responses) was the dependent variable. If classification learners and inference learners process and represent categorical information differently, their performance should be asymmetric between Classification Training and Inference Training conditions. However, if there is no difference between classification and inference training, participants should show symmetric pattern between two training conditions. In addition, if participants rely on the deterministic feature, the proportion of rulebased responses should be high. However, if they rely on multiple probabilistic features, they should make low level of rule-based response.

\section*{Results and Discussion}

All results reported here only focused on performance of the categorization task, specifically on the critical lures (i.e., \(\mathrm{P}_{\text {flurp }} \mathrm{D}_{\text {jalet }}\) and \(\mathrm{P}_{\text {jalet }} \mathrm{D}_{\text {flurp }}\) items). Recall that if participants form a rule-based representation of a category, they should identify the \(\mathrm{P}_{\text {flurp }} \mathrm{D}_{\text {jalet }}\) item as a jalet, whereas if they formed a similarity-based representation, they should identify this item as a flurp.

The main results of Experiment 1 are shown in Figure 2. As shown in the figure, children tended to form similaritybased representations regardless of condition, whereas adults tended to form rule-based representations in the classification condition. These findings were supported by statistical analyses - data in the figure were analyzed with 2 (Training Type: Classification and Inference) by 2 (Age Group: 4-5-year-olds vs. Adults) between-subjects ANOVA. There was a main effect of training type, \(F(1,52)\) \(=5.56, M S E=0.38, p=.022, \eta^{2}=0.097\), and a main effect of age group, \(F(1,52)=24.49, M S E=1.65, p<.001, \eta^{2}=\) 0.320 . Specifically, adults made more rule-based responses in Classification Training than in Inference Training, independent samples \(t(1,31.2)=2.63, p=.013, d=0.92\), with the proportion of rule-based responses above chance in Classification Training, one-sample \(t(1,15)=4.28, p=\) \(.001, d=1.07\), but around chance in Inference Training, one-sample \(t(1,18)=0.94, p=.359\). However, for children, they made comparable rule-based responses in both training conditions ( \(p=.334\) ), with the proportion of rulebased responses significantly below chance in Inference Training, one-sample \(t(1,6)=5.46, p=.002, d=2.06\), and marginally below chance in Classification Training, onesample \(t(1,13)=2.03, p=.064, d=0.54\).


Figure 2. Proportion of rule-based responses by age group and training type in Experiment 1.
The results are consistent with previous evidence (Yamauchi \& A.Markman, 1998; Hoffman \& Rehder, 2010) pointing to the predicted asymmetry between classification and inference training for adults. As predicted, adults tended to process and represent categorical information differently, with classification learners being more likely than inference learners to focus on deterministic feature, which separates two categories. However, there was little evidence that children learned categories differently by classification and inference. The symmetric performance suggested that children treated classification training and inference training equally and, more importantly, unlike adults, children formed similaritybased representation of categories.
One possible limitation of Experiment 1 was that participants were told explicitly about the deterministic and probabilistic features. It is possible that only adults, but not children attended to this information, and as a result, only adults formed rule-based representations. Experiment 2 attempted to eliminate this possibility by not mentioning that there were probabilistic and deterministic features.

\section*{Experiment 2}

\section*{Method}

Participants Twenty-six adults (18 women) and twenty preschool children ( \(M=55.3\) months, range 49.8-60.2 months; 7 girls) participated in this experiment. Two additional adults were texting during experiments and these data were excluded from the analysis.
Materials and procedure The materials were identical to those used in Experiment 1. The overall procedure in Experiment 2 was identical to Experiment 1 except that neither the information of P nor D was given to participants before main experiment or in the feedback (i.e., participants were only given corrective feedback). For the old items (i.e., High-Match items) at test, children were above \(70.1 \%\) categorization accuracy on these trials, and adults were above \(83.9 \%\), all above chance ( \(p\) s \(<.05\) ) and exhibited memory accuracy of \(75.0 \%\) and \(71.1 \%\), respectively. For the all-new-P items, Children and adults exhibited memory accuracy of \(78.2 \%\) and \(82.2 \%\), respectively.


Figure 3. Proportion of rule-based responses by age group and training type in Experiment 2.

\section*{Results and Discussion}

The main results of Experiment 2 are shown in Figure 3. The data were analyzed with 2 (Training Type: Classification and Inference) by 2 (Age Group: 4-5-yearolds vs. Adults) between-subjects ANOVA. The results revealed a significant training type by age group interaction, \(F(1,42)=16.48, M S E=0.64, p=.001, \eta^{2}=\) 0.282 . Independent samples t test indicated that adults made more rule-based responses in Classification Training than in Inference Training, \(t(1,24)=6.14, p=.001, d=\) 2.58, with the proportion of rule-based responses above chance in Classification Training, one-sample \(t(1,9)=4.95\), \(p=.001, d=1.56\), but below chance in Inference Training, one-sample \(t(1,15)=2.30, p=.036, d=0.57\). However, children exhibited comparable proportions of rule-based responses in both training conditions ( \(p=.604\) ), with the proportion of rule-based responses around chance in Classification Training ( \(p=.125\) ) and Inference Training ( \(p\) \(=.140\) ).

The results in Experiment 2 revealed the same pattern as Experiment 1. For adults, there was an asymmetry between classification and inference training; whereas young children's performance in the two training conditions was symmetric, and, regardless of the training condition, they formed similarity-based representations.

\section*{General Discussion}

The reported study examines the effects of training on the outcome of category learning and changes in these effects in the course of development. To achieve this goal, we trained adult and child participants with a category learning task in which participants learned two categories. Each category had a single deterministic feature that differed between the categories and multiple probabilistic features that partially overlapped between categories. Participants who were trained on a classification task were asked to classify items into one of two categories; whereas participants who were trained on an inference task were asked to infer a missing feature of items. Following training, participants were tested on their ability to categorize novel items.

Two major findings stem from the reported results. First, in both reported experiments adults exhibited an asymmetric pattern between classification training and inference training. Their rule-based responses in classification training were consistently higher than those in inference training, which is consistent with previous evidence (Yamauchi \& A. Markman, 1998; Hoffman \& Rehder, 2010) suggesting that adults process and represent categorical information differently. Specifically, classification learners are more likely to focus on deterministic (or rule-based) features than inference learners. However, for young children, the symmetry between classification and inference suggests that they do not treat these two training conditions differently. Second, adults tend to spontaneously detect a defining feature (Experiment 2) and classification learners tend to consistently rely on it to categorize items (Experiment 1 and 2). But there is little evidence that children rely on the deterministic feature in categorization. In contrast to adults, children tend to rely on a pattern of correlated probabilistic features, which reflects the overall similarity.
The results have implications for understanding the mechanisms underlying category learning and how these mechanisms may change in the course of development. Future research will also examine how attention is allocated in category learning by using a combination of eye tracking and behavioral paradigm.

\section*{Acknowledgements}

This research was supported by a National Science Foundation (NSF) grant (BCS-0720135) to Vladimir M. Sloutsky.

\section*{References}

Allen, S. W., \& Brooks, L. R. (1991). Specializing the operation of an explicit rule. Journal of Experimental Psychology: General, 120, 3-19.
Anderson, A. L., Ross, B. H., \& Chin-Parker, S. (2002). A further investigation of category learning by inference. Memory and Cognition, 30, 119-128.
Chin-Parker, S., \& Ross, B. H. (2002). The effect of category learning on sensitivity to within-category correlations. Memory and Cognition, 3, 353-362.
Chin-Parker, S., \& Ross, B. H. (2004). Diagnosticity and prototypicality in category learning: A comparison of inference learning and classification learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 30, 216-226.
Deng, W., \& Sloutsky, V. M. (2012). Carrot-eaters and moving heads: Inductive Inference is better supported by salient features than by category labels. Psychological Science, 23, 178-186.
Deng, W., \& Sloutsky, V. M. (2013). The role of linguistic labels in inductive generalization. Journal of Experimental Child Psychology, 114, 432-455.

Eimas, P. D., \& Quinn, C. (1994). Studies on the formation of perceptually based basic-level categories in young infants. Child Development, 65, 903-917.
French, R.M., Mareschal, D., Mermillod, M., \& Quinn, P.C. (2004). The role of bottom-up processing in perceptual categorization by 3- to 4-month-old infants: Simulations and data. Journal of Experimental Psychology: General, 133, 382-397.
Gelman, S.A. (1988). The development of induction within natural kind and artifact categories. Cognitive Psychology, 20, 65-95.
Gelman, S.A. (2004). Psychological essentialism in children. Trends in Cognitive Sciences, 8, 404-409.
Gelman, S. A., \& Heyman, G. D. (1999). Carrot-eaters and creature-believers: The effects of lexicalization on children's inferences about social categories. Psychological Science, 10, 489-493.
Gelman, S. A., \& Markman, E. (1986). Categories and induction in young children. Cognition, 23, 183-09.
Gelman, S. A., \& Markman, E. M. (1987). Young children's inductions from natural kinds: The role of categories and appearances. Child Development, 58, 1532-1541.
Hoffman, A. B., \& Rehder, B. (2010). The costs of supervised classification: The effect of learning task on conceptual flexibility. Journal of Experimental Psychology: General, 139, 319-340.
Mareschal, D., \& Quinn, P. C. (2001). Categorization in infancy. Trends in Cognitive Sciences, 5, 443-450.
Mareschal, D., Quinn, P.C., \& French, R.M. (2002). Asymmetric interference in 3- to 4-month olds' sequential category learning. Cognitive Science, 26, 377-389.
Markman, E. M., \& Hutchinson, J. E. (1984). Children's sensitivity to constraints on word meaning: Taxonomic versus thematic relations. Cognitive Psychology, 16, 127.

Markman, A. B., \& Ross, B. H. (2003). Category use and category learning. Psychological Bulletin, 129, 592613.

Murphy, G. L. (2002). The Big Book of Concepts, MIT Press.
Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. Journal of Experimental Psychology: Learning, Memory, and Cognition, 10, 104-114.
Nosofsky, R. M., \& Palmeri, T. J. (1998). A rule-plusexception model for classifying objects in continuousdimension spaces. Psychonomic Bulletin \& Review, 5, 345-369.
Nosofsky, R. M., Palmeri, T. J., \& McKinley, S. C. (1994). Rule-plus-exception model of classification learning. Psychological Review, 101, 53-79.
Oakes, L. M., Madole, K. L., \& Cohen, L. B. (1991). Object examining: Habituation and categorization. Cognitive Development, 6, 377-392.

Rehder, B., \& Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. Cognitive Psychology, 51, 1-41.
Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou \& A. Ortony (Eds.), Similarity and analogical reasoning (pp. 21-59). New York: Cambridge University Press.
Sakamoto, Y., \& Love, B. C. (2006). Vancouver, Toronto, Montreal, Austin: Enhanced oddball memory through differentiation, not isolation. Psychonomic Bulletin \& Review, 13, 474-479.
Shapard, R. N., Hovland, C. I., \& Jenkins, H. M. (1961). Learning and memorization of classification. Psychological Monographs, 75(Whole No.517).
Sloutsky, V. M., \& Fisher, A. V. (2004). Induction and categorization in young children: A similarity-based model. Journal of Experimental Psychology: General, 133, 166-188.
Sloutsky, V. M., Kloos, H., \& Fisher, A. V. (2007). When looks are everything: Appearance similarity versus kind information in early induction. Psychological Science, 18, 179-185.
Sloutsky, V. M., \& Lo, Y.-F. (1999). How much does a shared name make things similar? Part 1: Linguistic labels and the development of similarity judgment. Developmental Psychology, 35, 1478-1492.
Smith, E. E. (1994). Concepts and categorization. In E. E. Smith \& D. N. Osherson (Eds.), An invitation to cognitive science (pp. 3-33). Cambridge, MA: MIT Press.
Smith, L. B., Jones, S. S., \& Landau, B. (1996). Naming in young children: A dumb attentional mechanism? Cognition, 60, 143-171.
Welder, A. N., \& Graham, S. A. (2001). The influences of shape similarity and shared labels on infants' inductive inferences about nonobvious object properties. Child Development, 72, 1653-1673.
Yamauchi, T., Love, B. C., \& Markman, A. B. (2002). Learning nonlinearly separable categories by inference and classification. Journal of Experimental Psychology: Learning, Memory, and Cognition, 3, 585-593.
Yamauchi, T., \& Markman, A. B. (1998). Category learning by inference and classification. Journal of Memory and Language, 39, 124-148.
Yamauchi, T., \& Markman, A. B. (2000a). Inference using categories. Journal of Experimental Psychology: Learning, Memory, and Cognition, 26, 776-795.
Yamauchi, T., \& Markman, A. B. (2000b). Learning categories composed of varying instances: The effect of classification, inference, and structural alignment. Memory and Cognition, 28, 64-78.

\title{
Information-Sharing in Three Interacting Minds Solving a Simple Perceptual Task
}

\author{
Michal Denkiewicz (michal.denkiewicz@gmail.com) \\ Department of Psychology, University of Warsaw \\ Stawki 5/7, 00-183 Warsaw, Poland
}

Joanna Rączaszek-Leonardi (raczasze@psych.uw.edu.pl)
Institute of Psychology, Polish Academy of Sciences Jaracza 1, 00-378 Warsaw, Poland

Piotr Migdal (piotr.migdal@icfo.es)
ICFO-Institut de Ciències Fotòniques
08860 Castelldefels (Barcelona), Spain

\author{
Dariusz Plewczynski (darman@icm.edu.pl) \\ Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw \\ Pawińskiego 5a, 02-106 Warsaw, Poland
}

\begin{abstract}
We study three-person groups solving a simple, two alternative forced choice task of perceptual nature. The group members provide individual answers and afterwards discuss and reach a joint decision. Different models of information sharing describe the theoretical relationship between group and individual performance. Experimental data shows that average-performing members can benefit from cooperation, but the groups do not outperform their best members. Results point to voting as the best explanation of the behavior of the groups.
\end{abstract}

Keywords: group decision making; distributed cognition; information sharing

\section*{Introduction}

Whether "two heads are better than one" is no settled question in psychology. Many studies report groups to be less proficient than their most capable members (Corfman \& Kahn, 1995), or that there is no benefit (Heaney, Foster, Gregor, O’Neill, \& Wood, 2010). Groups are often regarded as source of negative influence on individual performance, stemming from conformism (Asch, 1951) or social loafing (Allport, 1924).

There are, however, studies that report benefit from cooperation (Kerr \& Tindale, 2004; Hastie \& Kameda, 2005). In one of such studies Bahrami (B. Bahrami et al., 2010) determined that dyads can outperform their members in a simple two-choice perceptual task, provided that those members have similar individual effectiveness. This group benefit disappeared when communication was forbidden; hence free information sharing was a key factor.

This result is interesting because there is no obvious reason for group benefit to occur in such a simple task. For certain types of tasks, such as concept mastery, concept attainment or learning, group members can pool cognitive resources, or utilize complimentary skill or knowledge. Then a group can provide better solutions in terms of quality, though not necessarily efficiency (Steiner, 1966;

Hill 1982). A notable exception is brainstorming, where participating in a group has a negative impact both on quality and quantity of creative solutions (Taylor, Berry, \& Block, 1958).

If resources cannot be shared, perhaps solutions can be. For so-called "eureka-tasks" a solution can be demonstrated in objective terms, thus (at least in theory) a single participant, who finds the correct answer, can easily persuade other members, leading to a correct group solution. Hence, the chance of group solving a task grows with its size. Examples of such task are Remote Associates Test (Laughlin \& Bitz, 1975), or simply scrambled letters/anagrams.

In Bahrami's study neither of these conditions was met. The participants first performed the perceptual task on their own, without the ability to divide it into parts. Then a group decision was made, based solely on what the participants perceived individually. No reasoning or previous knowledge or was of any use and the only thing, that the participants could communicate was their subjective idea, of what they think the answer was. Still not only did the pairs perform better, than chance (which in this case means the averaged effectiveness of the two participants), but they also outperformed the better of the two members.

Bahrami tested several theoretical models of information sharing in communication, developed in the spirit of signal detection theory, and concluded, that dyad members communicate their own relative confidences, which allows the group benefit to occur.

Later these models were theoretically extended to groups of arbitrary size by Migdał (Migdał, Raczaszek-Leonardi, Denkiewicz \& Plewczynski, 2012). "Aggregation of decisions" was also considered, that is a situation when subgroups of a larger group first reach their decisions, and then try to convince one another.

The topic of this paper is an experimental attempt to verify the applicability of these theoretical models to three-
person groups, and shed light onto how these groups reach their decisions.

\section*{Decision-Making Models}

Consider the following task. On each trial a group consisting of three participants views two sets of stimuli, one after another, for a brief period ( 85 ms ). Each of the two sets consists of six Gabor patches. All patches are identical, with the exception of one (target) patch - which has higher contrast. The task is to determine which screen, the first or the second, contains the target patch. The difference in contrasts between the target patch and non-target patches determines the difficulty of the task. We use the convention, that this difference is positive, when the target is in the second set, and negative, if it's on the first.

Knowing the answers of a given a participant, we can determine the probability, that he or she chooses the second answer as a function of the contrast difference. We model this relationship with a cumulative of the normal distribution, as it provides a good fit to the experimental data (Bahrami et al., 2010).

Again, following the convention we describe this psychometric function with two parameters: slope ( \(s\) ) and bias (b). The slope is a measure of participant's performance and is of primary interest. The bias describes the tendency to give one particular answer - the task is constructed in such a way, that this parameter should be close to 0 . These parameters are related to the standard parameters of a Gaussian curve (mean \(\mu\) and standard deviation \(\sigma\) ) in the following way:
\[
\begin{align*}
& s=\frac{1}{\sqrt{2 \pi} \sigma}  \tag{1}\\
& b=-\mu \tag{2}
\end{align*}
\]

In the same way as individual decision are used to construct curves for participants, group curves can be obtained from group decisions, allowing for a comparison of group and individual slope parameters.

By making different assumptions about communication and decision making process within the group, a theoretical dependency between individual slopes \(s_{1}, s_{2}, s_{3}\) and group slope parameter \(s_{g}\) can be established, in the form of a function \(s_{g}\left(s_{1}, s_{2}, s_{3}\right)\).

If the behavior of the group actually matches the assumptions, then the empirically obtained group slopes should not differ significantly from theoretical predictions. By an information-sharing model we understand a possible set of such assumptions. We consider six such models:

Random Responder (RR) This model assumes that the communication is actually ineffective, and the final decision is randomly chosen from the individual decisions.

Best Decides (BD) It is plausible that the group will simply entrust the decision to the biggest "expert" in the task. That person's decision becomes the group decision. The model assumes that the group initially possesses the knowledge
about who is the best member. This is a somewhat simplistic assumption, yet it is feasible that the best member can be determined in the beginning, based on a small number of trials.

Voting (Vot) The group uses the majority rule to determine the group decision. It requires only the communication of individual decisions.

Weighted Confidence Sharing (WCS) Each member shares his/her own relative confidence. The participants are unable to discern their perceived contrast difference from its reliability (determined by the participant's slope).

Direct Signal Sharing (DSS) Group members communicate both their perceived contrast difference and their confidence separately. This allows for a statistically optimal decision (Sorkin \& Hays, 2001).

Truth Wins (TW) In this model each participant either knows the correct answer, or is aware of not knowing it. In other words a person cannot falsely believe that he knows the answer. This basically means, that if a single group member finds out the correct answer, the group also answers correctly.

Table 1 summarizes the mathematical relationships between group and individual performance, according to the formulas for arbitrary group sizes presented in Migdał et al., 2012.

Table 1: Group slope as a function of individual slopes according to different models.
\begin{tabular}{lc}
\hline Model & \(s_{g}\left(s_{1}, s_{2}, s_{3}\right)\) \\
\hline\(R R\) & \(\frac{s_{1}+s_{2}+s_{3}}{3}\) \\
Vot & \(\frac{s_{1}+s_{2}+s_{3}}{2}\) \\
\(B D\) & \(\max \left\{s_{1}, s_{2}, s_{3}\right\}\) \\
\(W C S\) & \(\frac{s_{1}+s_{2}+s_{3}}{\sqrt{3}}\) \\
\(D S S\) & \(\sqrt{s_{1}^{2}+s_{2}^{2}+s_{3}^{2}}\) \\
\(T W\) & \(s_{1}+s_{2}+s_{3}\) \\
\hline
\end{tabular}

All models, with the exception of the RR model, predict that group performance should exceed average individual performance of that group's members.

In Bahrami's (Bahrami et al. 2010) study of dyads, the WCS model best described the behavior of the participants.

We expected that such behavior should also be seen in three-person groups, that is the groups would outperform their best members, at least in cases of homogenous groups (in terms of performance).

\section*{Experiment}

\section*{Subjects}

Participants were recruited from general population of Warsaw, Poland, using snowball method. There were 15 three-person groups, which gives 45 participants: 30 female and 15 male. All participants were adults (mean age 38, s.d. 15.6) with normal or corrected-to-normal vision. Members of each group knew each other. Written informed consent was obtained from each participant prior to the experiment and each person was rewarded with 25 PLN (approximately 6 EUR) for completing the experiment.

\section*{Experimental setup}

The testing room contained three computer stations, each with a LCD display ( 24 inch, resolution \(=1920 \times 1280\), refresh rate \(=60 \mathrm{~Hz}\) ) and a keyboard. The stations were connected via a local network and arranged in the form of an equilateral triangle, with the displays facing outwards. Each participant saw only his/her own display, but was able to see the fellow group members' faces. To minimize distractions, during the experiment the room was nearly completely dark - the only sources of light were the monitors. However, the lighting was sufficient for the participants to see each other's faces.
Each computer was assigned a color: blue, orange or yellow, for the purpose of identification. The experiment was controlled by custom software based on the PsychoPy framework ("http://www.psychopy.org/"; Peirce, 2008).

\section*{Stimuli}

Each of the two stimuli sets consisted of six Gabor patches evenly distributed around the center of the screen, at a distance of 8 degrees. All patches were vertically oriented and had the following parameters: standard deviation of the Gaussian envelope \(=0.45\) deg. spatial frequency \(=1.5\) cycles/deg. The contrast parameter equaled \(10 \%\) for the non-target patch and \(11.5 \% 13.5 \%, 17 \%\) or \(25 \%\) for the target patch (hence, the contrast difference value was \(1.5 \%\) \(3.5 \%, 7 \%\) and \(15 \%\), respectively). The position of the target patch within its set was chosen randomly each trial.

The background was uniform and gray at all times.

\section*{Task and Procedure}

The experiment consisted of a practice block followed by three experimental blocks. After each experimental block the participants changed their places, moving one seat to the left, so that each participant spent about the same time using each computer. The experimenter was present in the testing
room during the entire experiment, to assure that the procedure was followed.
There were 288 experimental trials - three blocks of 96 trials. The practice sessions consisted of 8 trials. The number of trials with each combination of difficulty level and correct answer combination was equal within blocks. Each trial started with a black fixation cross, placed in the center of the screen, displayed for 500 ms . The two sets of stimuli followed, each visible for 85 ms , separated with a blank screen presented for 1000 ms. Finally, a white question mark appeared, indicating that the participants are to make their individual decisions. The decisions were made by pressing an appropriate key - left or right arrow indicating that the target was in the first or in the second set respectively. The keys were labeled " 1 " and " 2 ". After a button had been pressed the question mark was replaced with a message "Wait for other participants" decisions". It stayed on the screen until all individual decisions were made. So far the participants were not allowed to communicate with each other.
Next, the group decision phase followed. Individual decisions were displayed for 1.5 s , one above the other, each in the color of the respective computer. Then a message appeared asking the participants to discuss the group decision. After the group decision has been agreed upon, a single person was required to input the decision, in the same way as the individual decisions were made. This was always the person to the left of the previous decision maker.
During the group decision phase the participants [] could communicate freely, but were not allowed to leave their seats. The method they could use to arrive at a group decision was not constrained in any way, there was no predefined decision-making scheme (e.g. voting).
After the group decision had been made, feedback was displayed, containing information about the correctness of the group decision and of each individual decision. Feedback was visible for 1.5 s and after it disappeared, the next trial followed immediately.

If the individual decisions were unanimous, the group decision was automatically assumed to be the same. In such case the group decision phase was skipped and feedback was displayed.
The duration of the experimental session depended solely on the pace at which the participants completed the trials, and this was on average 44 minutes.

\section*{Results}

For each participant and each group the slope, \(s\), and bias, \(b\), parameters were estimated by fitting a probit regression model to that individuals (or that groups) decisions. Individual slopes were used to compute the theoretical values of group slope, predicted by each model, according to appropriate formula from Table 1 ( \(\left.s_{\text {model }}\right)\). These values were compared to the values obtained empirically \(\left(s_{\text {group }}\right)\). If the ratio of the empirical and theoretical value, \(s_{\text {group }} / s_{\text {model }}\), was greater than one, it means that the groups outperformed the model, if it was less than one - the groups did not reach
predicted accuracy. A one sample Student's t-test was used to determine the significance of the deviations from theoretical predictions; the quotient was compared with the value of 1 . The results are summarized in Table 2.

Table 2: Experimental results.
\begin{tabular}{llllll} 
model & \(M\) & \(S D\) & \(t(14)\) & \(p\) & \\
\hline RR & 1.42 & 0.32 & 5.16 & \(<.001\) & \(* * *\) \\
BD & 0.93 & 0.24 & -1.11 & .29 & \\
Vot & 0.95 & 0.21 & -0.94 & .36 & \\
WCS & 0.82 & 0.18 & -3.77 & .002 & \(* *\) \\
DSS & 0.71 & 0.15 & -7.19 & \(<.001\) & \(* * *\) \\
TW & 0.47 & 0.11 & -19.23 & \(<.001\) & \(* * *\) \\
\hline \multicolumn{4}{l}{ Note: \({ }^{*} p<.05 .{ }^{* *} p<.01 .{ }^{* * *} p<.001\)}
\end{tabular}

The ratio \(s_{\text {group }} / s_{\text {model }}\) was significantly greater than 1 for the RR model and significantly less than 1 in the case of WCS, DSS and TW models. In other words the actual slope parameters of the groups were significantly greater than those predicted by RR model, and significantly lower, than the predictions of the WCS, DSS, TW models.

In case of some models, namely BD, Vot and TW, a group decision on a given trial can be determined from the individual decisions. This allows for a per-trial comparison of the theoretical decisions and the decisions actually made. A group decision agrees with the BD models if it is the same, as the decision of the best group member (i.e. the one with the highest slope parameter). In case of the Vot model it is the decision made by the majority of participants. Finally, in the TW model it is the correct answer, unless all members were wrong. For each group three values were calculated, indicating the number of group decisions that were consistent with each of these tree models. A paired ttest was performed for each pair of these models. The differences in consistency between the Vot model ( \(M=\) 254.7, \(S D=21.18\) ) and both the BD model ( \(M=218.7, S D\) \(=27.6, \mathrm{p}<.001)\), and TW model, \((M=213.5, S D=19.6, \mathrm{p}\) \(<.001\) ) were significant and positive. The difference in consistency between BD model and TW model was not significant ( \(\mathrm{p}=.77\) ).

\section*{Discussion}

Analysis of individual and group slopes allows us to reject the RR, WCS, DSS and TW models, leaving Vot and BD as plausible explanations of group performance. A trial-by-trial analysis points to Vot model (majority voting) as the best explanation.

The rejection of some models does not mean that the types of behavior they describe did not occur or that they are impossible. It merely shows that they were not dominant in the course of the experiment. Indeed in a small, but not negligible number of trials (about 12\%), the group decisions corresponded neither to BD nor Vot model predictions.

\section*{Conclusions}

Results of the experiment indicate that three-person groups prefer voting as a method of reaching a joint decision, and more advanced communication is rarely employed. Groups outperformed, so to say, their average members, but best members generally made better decisions than the group. This group benefit can be attributed solely to the use of voting as a decision-making scheme.

The failure to outperform the best group members, as it was in the case of pairs in the study by Bahrami et al. (2010), can be explained in many ways. First, the requirement for cognitive resources used for communication and integration of information increase as more group members are added, leading to deterioration of information processing performance. Secondly, in groups of size three, as opposed to dyads, voting becomes possible. Since group members are not directly rewarded for accuracy, and the experimental task is somewhat tedious, employing a simple, relatively good, and socially acceptable method of reaching a group decision seems tempting.

Conformism, or caring for group's coherence, can also play a role, as it can shy away a single correct group member, from confronting the majority decision (and prolonging the decision phase). The social acceptance of voting can afterwards serve as a justification for this group member, if he or she were blamed for not insisting on the correct answer. On the other hand the responsibility for being wrong diminishes, if one is the member of a majority.

It is feasible that the impossibility of automatically resolving a tie in dyads fosters communication and, in turn, increases performance. Adding a third member provides an opportunity to use a simpler and less effective decision making system and, paradoxically, diminish performance. This shows how seemingly simple task and situations can produce non-trivial dependencies.

\section*{Acknowledgements}

The work was supported by EC EuroUnderstanding grant DRUST to JRL, Spanish MINCIN project FIS2008-00784 (TOQATA) and ICFO PhD scholarship to PM, and the Polish Ministry of Education and Science (grants: N301 159735, N518 409238) to DP.

\section*{References}

Allport, F. H. (1924). Response to Social Stimulation in the Group. Social Psychology (pp. 260-291). Boston: Houghton Mifflin Company.
Asch, S. E. (1951). Effects of group pressure on the modification and distortion of judgements. In H . Guetzkow (Ed.), Groups, leadership and men. (pp. 177190). Pittsburgh:PA: Carnegie Press.

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., \& Frith, C. D. (2010). Optimally Interacting Minds. Science, 329(5995), 1081-1085.
Clark, A. (2006). Language, embodiment, and the cognitive niche. Trends in Cognitive Sciences, 10(8), 6-10.

Corfman, K. P., \& Kahn, B. E. (1995). The influence of member heterogeneity on dyad judgment: Are two heads better than one? Marketing Letters, 6(1), 23-32.
Hastie, R., \& Kameda, T. (2005). The robust beauty of majority rules in group decisions. Psychological review, 112(2), 494-508.
Heaney, R., Foster, F. D., Gregor, S., O’Neill, T., \& Wood, R. (2010). Are two heads better than one? An experiment with novice share traders. Australian Journal of Management, 35(2), 119.
Kerr, N. L., \& Tindale, R. S. (2004). Group performance and decision making. Annual review of psychology, 55, 623-55.
Laughlin, P. R., \& Bitz, D. S. (1975). Individual versus dyadic performance on a disjunctive task as a function of initial ability level. Journal of Personality and Social Psychology, 31(3), 487-496.
Migdał, P., Rączaszek-Leonardi, J., Denkiewicz, M. \& Plewczyński, D. (2012). Information-sharing and aggregation models for interacting minds, Journal of Mathematical Psychology, 56, 417-426.
Peirce, J. W. (2008). Generating Stimuli for Neuroscience Using PsychoPy. Frontiers in neuroinformatics, 2 (January), 10.
Sorkin, R., Hays, C., (2001). Signal-detection analysis of group decision making. Psychological Review - New York - 108 (I), 183-203.

Steiner, I. D. (1966). Models for inferring relationships between group size and potential group productivity. Behavioral Science, 11, 273-283.
Taylor, D.W., Berry, P.C., Block, C.H. (1958). Does group participation when using brainstorming facilitate or inhibit creative thinking? Administrative Science Quarterly, 3, 23-47.

\title{
Information Sampling, Conformity and Collective Mistaken Beliefs
}

\author{
Jerker Denrell (Jerker.Denrell@wbs.ac.uk ) \\ University of Warwick, United Kingdom
}

\author{
Gaël Le Mens (Gael.Le-Mens@upf.edu) \\ Universitat Pompeu Fabra, Spain
}

\begin{abstract}
Societies sometimes stick to the status quo instead of switching to superior technologies and institutions. Existing explanations often attribute this to a coordination failure due to payoff externalities: people may know that another alternative is superior but nobody has an incentive to switch unless many others do so. We show that a simple learning argument can provide an alternative explanation. When people learn about the alternatives from their own experiences but tend to adopt the behaviors of others, they will mistakenly learn to believe that a popular alternative is superior to a better, but unpopular alternative. Our model neither assumes that agents engage in motivated cognition nor that they transmit mistaken information to others. Rather, it emphasizes the role of a fundamental asymmetry in access to information about popular versus unpopular alternatives. Our model thus provides a novel, sampling-based, explanation of how conformity in behavior can lead to private acceptance.
\end{abstract}

Keywords: Social Change, Learning, Conformity, Popularity
Explaining why collective mistakes emerge and persist is central to understanding social change, immobility, and differences in welfare (e.g. Elster, 1978; North, 1990). In developing countries, people keep on using poor domestic hygiene practices even though simple changes would save many lives (Curtis, Cairncrosss \& Yonli, 2000), and farmers fail to use fertilizer despite their potential for large increases in productivity (Duflo, Kremer \& Robinson, 2009). In western countries, firms persist in making abundant use of temporary workforce despite the inefficiency of this arrangement (Pfeffer, 1998). How come large groups of people could persist in using inefficient technologies, practices or institutions when better courses of actions are available?

Prior literature has proposed two classes of explanations. The first perspective has demonstrated that such mistakes can occur when payoffs increase with the number of people taking the same action (Arthur, 1989; Elster, 1978; North, 1990). This type of explanation, which relies on network externalities, generally assumes that people know the values of the alternatives. The problem is that nobody has an incentive to switch unless many others do so. The second perspective proposes that collective mistakes can emerge because agents sometimes believe that an alternative that is in fact suboptimal is the best. Rather than emphasizing a coordination failure, such explanations rely on the fact that people may not be aware of the qualities of the alternatives. This perspective assumes that agents use popularity as a signal of quality (Banerjee, 1992; Bikhchandani, Hirshleifer \& Welch, 1992). Explanations that fall in this tradition are thus only valid when agents infer the qualities of the alternatives of the basis of the choices of others.

Here, we show that collective mistakes can still emerge and persist even if payoffs do not depend on the choices of others and people do not use popularity as a signal of quality. Instead of assuming that popularity directly affects an agent's evaluation of the available alternatives, we analyze situations where popularity only affects agents' sampling decisions: In our model, agents are more likely to try the most popular alternative, i.e., the alternative believed by most to be the superior. But agents' quality estimates are solely based on their own experiences with the alternatives. There are many situations where one might expect to see such conformity in behavior (people choose the popular alternative) but not in attitudes (their attitudes depend only on their own experience). For example, people may decide to go along with the majority and select the most popular alternative to avoid being seen as deviant (Cialdini \& Goldstein, 2004), or because of adverse reputation effects to receiving a poor outcome with an unusual alternative (Keynes, 1936; Scharfstein \& Stein, 1990). For example, it is difficult to avoid learning about the research tradition that is dominant in your department.

We show that when agents are more likely to be exposed to popular alternatives, a sampling bias emerges that leads most people to believe that the quality of popular alternatives is superior to that of unpopular alternatives. The intuition behind this result pertains to how people sample information about the available alternatives (Denrell \& Le Mens, 2007; Le Mens \& Denrell, 2011). Consider a medical doctor who has to select one of two possible treatments to cure a patient. Treatment P is popular among other doctors in her reference group and patients whereas treatment R is rarely chosen. Suppose the doctor selects R and the initial outcome is disappointing. While this might be a signal of the poor quality of R , it could also have resulted from other causes. But because patients may not want to continue a treatment with an unpopular drug with disappointing initial results, the doctor is likely to abandon R. In doing so, she will fail to discover that R might be an efficacious treatment. Compare this to what would have happened, had she selected P instead. She might have continued with \(P\), even following disappointing initial results, because patients have heard that P has been efficacious in the past. In doing so, she would have acquired additional information about P and she might have discovered that this treatment was in fact efficacious.

This stylized story illustrates an important asymmetry in opportunities for error correction: an error of underestimation of the efficacy of a popular treatment is less likely to persist than an error of underestimation the efficacy of an unpopular
treatment.
Prior research has identified a number of psychological mechanisms and biases that explain why people might come to prefer what is available or most popular even when it is not the best alternative (Bem, 1972; Festinger, 1957; Heider, 1958). These explanations assume that people adjust their evaluations of the alternatives based on what is seen as a norm. Here, we analyze a simple learning model that demonstrates that it is not necessary to invoke such intra-psychic adjustment. Our model also has novel policy implications, as it suggests that exposing initially skeptical adopters to a new practice may be sufficient to enhance its diffusion.

\section*{Model}

In our model, agents make a sequence of choices among two uncertain alternatives and learn about the qualities of the alternatives from their own experiences. Our model is an extension of standard models of the evolution of coordination (Arthur, 1989). Consider a growing population. In each period, \(t=1,2, \ldots\), one new agent enters the population. Each agent makes a choice, in every period, between two competing alternatives. Everything else equal, agents prefer to select the alternative that is chosen by the largest proportion of others in the previous period. People might care about popularity because of payoff externalities (i.e. when an alternative is more useful if widely spread) or because of adverse reputation effects to receiving a poor outcome with an unusual alternative (Keynes, 1936; Scharfstein \& Stein, 1990). An agent may forego some of the benefits of coordination, however, and choose the less popular alternative, if she believes it is of a higher quality. These assumptions are implemented in the model by the use of the following logistic choice rule. The the probability that Agent \(i\) selects Alt. 1 in period \(t+1\) is:
\[
\begin{equation*}
p C_{1}^{i}\left(\hat{q}_{1, t}^{i}, \hat{q}_{2, t}^{i}, p_{1, t}\right)=\frac{e^{a p_{1, t}+b \hat{q}_{1, t}^{i}}}{e^{a p_{1, t}+b \hat{q}_{1, t}^{i}}+e^{a\left(1-p_{1, t}\right)+b \hat{q}_{2, t}^{i}}} \tag{1}
\end{equation*}
\]
where \(a\) and \(b\) are positive constants, \(p_{k, t}\) is the share of the population choosing Alt. \(k\) in period \(t\), and \(\hat{q}_{k, t}^{i}\) is Agent \(i\) 's estimate of the quality of Alt. \(k\) at the end of period \(t\). The logistic choice rule is often used to model choice under uncertainty (Erev \& Baron, 2005; Sutton \& Barto, 2005) and provides good fit to experimental data on sequential choices (Denrell, 2005).

Whenever Agent \(i\) chooses Alt. \(k\), she can observe some information about the quality of that alternative. The observation, \(z_{k, t}^{i}\), is a random variable drawn from the underlying quality distribution, which is assumed to be normally distributed with expected value \(\mu_{k}\) and common variance \(\sigma^{2}\). We assume a simple updating rule: the revised estimate is a weighted average of the past estimate and the new observation (Busemeyer \& Myung, 1992; Denrell, 2005),
\[
\begin{equation*}
\hat{q}_{k, t}^{i}=(1-\lambda) \hat{q}_{k, t-1}^{i}+\lambda z_{k, t}^{i}, \tag{2}
\end{equation*}
\]
where \(\lambda\) is between 0 and 1 . If \(i\) does not select Alt. \(k\) in period \(t-1\), the estimate remains the same: \(\hat{q}_{k, t}^{i}=\hat{q}_{k, t-1}^{i}\). Agent \(i\) 's
initial estimate, when she enters the population, is a random variable drawn from a normal distribution with mean zero and variance \(\sigma^{2}\).

Note that this formulation assumes that agents only learn from personal experience: an agent only updates her estimate of the quality of the alternative she personally observes. Thus, in this model, agents do not infer the qualities of the alternatives based on the choices of others, nor do they learn from the observations of others.

\section*{Analysis}

It is possible to do a formal analysis of estimates and choices when the number of period becomes large. In the Appendix, we derive an explicit expression for the joint distribution of the asymptotic quality estimates \(\left(\hat{Q}_{1}, \hat{Q}_{2}\right)\) of the two alternatives (eq. 10). We then solve, numerically, for the limiting values of \(p_{1}\), the proportion of agents choosing Alt. 1. Finally, for any given value of \(p_{1}\), we compute, by numerical integration, the probability that Alt. 1 is considered superior.

The predictions of the asymptotic analysis are depicted by the solid lines on the graphs of Fig. 1; the diamonds represent simulated estimates after 500 periods. The left panel shows that when the weight of popularity in the sampling rule (parameter \(a\), see eq. 1 , is small, most agents select the first alternative, which has a higher average quality. If \(a\) is sufficiently large, however, it is possible that most agents select the second, inferior, alternative instead. The right panel shows that, in this case, most agents will also come to believe that the second alternative has a higher quality, i.e., \(P\left(\hat{Q}_{1}<\hat{Q}_{2}\right)>0.5\). In summary, our model implies that, in all cases, most agents will come to believe that the alternative chosen by most has the higher quality, even if such belief is mistaken.

More generally, the expected asymptotic quality estimate of an alternative, for a randomly chosen agent, is an increasing function of the limiting proportion of agents who choose it:
\[
\begin{equation*}
E\left[\hat{Q}_{k}\right]=\mu_{k}-\left(1-p_{k}\right) \frac{b \lambda}{(2-\lambda)} \sigma^{2} \tag{3}
\end{equation*}
\]

This equation demonstrates that the choices of other agents create a systematic externality on the quality estimates of an agent, in spite of the fact that she learns only from her own experience. More precisely, the quality of a rarely chosen alternative is systematically underestimated, and the lower the number of agents who choose it, the more severe the underestimation. The negativity bias also occurs for the popular alternative but is of much lower magnitude. In particular, when the limiting proportion of agents who select the popular alternative is close to 1 , there is almost no bias in the quality estimate of that alternative.

The probability that the process will converge to the inferior alternative, and that this alternative will be believed to have a higher quality, depends crucially on the difference in average qualities and on the variability of the observations. If the mean returns differ substantially, agents will quickly identify the best alternative. For example, suppose \(\mu_{1}=1, \mu_{2}=0\)


Figure 1: Simulated estimates (open diamonds) match the predictions of the asymptotic theory (solid line). The left panel displays the possible limiting values for the proportion of agents who select Alt. 1. These are the roots of the self-consistency equation (eq. 12). The right panel displays the limiting proportion of agents who believe that Alt. 1 is the best. When most agents choose Alt. \(1\left(p_{1}>1 / 2\right)\), most of them also believe that it is the best. And inversely, when most agents choose Alt. 2 ( \(p_{1}<1 / 2\) ), most agents believe that it is the best. Simulation results are averages based on 500 runs of 500 periods each (with \(b=1.5, \mu_{1}=0.25, \mu_{2}=0\) and \(\sigma^{2}=1\) ).
and \(\sigma^{2}=1\). It is then highly unlikely that agents will converge to Alt. 2 and believe it has higher quality. The presence of the variance of the observations in eq. 3 also shows that convergence on the inferior alternative is much more likely if alternatives are difficult to distinguish because observations are noisy. Noisy observations often happen when there is some delay between choices and observations of the corresponding outcomes which makes the association between actions and outcomes difficult. In those settings, the dynamics of experiential learning could be of particular importance.

To understand the intuition underlying these results, note that individuals accumulate biased samples of information about the qualities of the two alternatives. Negative estimates reduce the probability of further sampling, which implies that no further information is available and the quality estimate will not be updated (Denrell, 2005; Fazio, Eiser, \& Shook, 2004). This implies, in turn, that the qualities of the alternatives will tend to be underestimated, as illustrated by eq. 3 .

The magnitude of the information bias, however, is moderated by popularity. To see how, consider extreme cases. Suppose, for example, that Alt. 1 is much more popular than Alt. 2. In this case, decision makers will sample Alt. 1 almost no matter what their quality estimates are. There is thus almost no sampling bias for this alternative, the quality estimate is close to the true quality and there is almost no underestimation tendency for this alternative. When Alt. 1 is much less popular than Alt. 2, sampling will depend more strongly on quality estimates. The quality estimate will thus be subject to the systematic underestimation tendency described above.

It is important to note that our results do not require that the estimates of each individual converge toward a stable value.

In fact, the magnitude of the sampling bias is strongest when \(\lambda=1\). In this case, agents' quality estimates correspond to their last observation of the alternative, and thus the estimates are subject to potentially large changes after each observation. But the population still converges to one of the two alternatives, and the estimates for this alternative become more positive than for the other alternative. More generally, simulations show that people will still tend to evaluate popular alternatives more positively under different assumptions regarding the estimate updating rule, such as when \(\lambda\) declines with the number of observations. For example, when the quality estimate of Alt. \(k\) is the average of all prior observations of that alternative, it can still happen that most people select the inferior alternative and mistakenly believe it to be the alternative of higher quality. But the synchronization of estimates and behavior occurs for values of \(\sigma\) that tend to be larger than when \(\lambda\) is constant. This is not surprising, because in that case, estimates integrate information better (when the environment is stable) than when \(\lambda\) is constant. \({ }^{1}\)

\section*{Coordination and Synchronization of Estimates}

We motivated our assumption that people are more likely to select popular alternatives by referring to settings where people want to conform to the majority. But another reason for wanting to select a popular alternative is the desire to coordinate one's behavior with others (e.g. Hardin, 1968). For example, people might want to go to the same venues as those in the same social group, or they might want to use the same computer platform as others so as to be able to exchange files

\footnotetext{
\({ }^{1}\) This case is very close to Bayesian updating with a prior on \(\mu_{k}\) that has a \(N\left(\mu_{k}, \sigma\right)\) distribution.
}
more easily. It is possible to model such settings as follows: Suppose that there are \(N\) players who choose, in each period, one of \(M\) alternatives. The payoff of Alt. \(k\) follows a Gaussian distribution with mean \(\mu_{k}\) and variance \(\sigma^{2}\). That is, if \(o_{k, t}^{i}\) denotes the payoff Player \(i\) receives from selecting Alt. \(k\) in period \(t\), we have \(o_{k, t}^{i} \sim N\left(\mu_{k}, \sigma\right)\). In addition, there is a coordination bonus: if \(i\) selects Alt. \(k\) in period \(t\) and \(\eta_{k, t}^{-i}\) other players select Alt. \(k\) in that period, Player \(i\) receives a bonus of \(c \eta_{k, t}^{-i}\) where \(c\) is a positive parameter. Suppose there are \(T\) periods and that the goal of each player is to maximize the total payoff she receives over the \(T\) periods.

These assumptions regarding the payoff structure and the goals of the decision makers specify a coordination game (e.g. Gibbons, 1992). In this setup, people have an incentive to select a popular alternative. This formulation defines a setting where people are likely to develop quality estimates consistent with the predictions of our theory without making assumptions about the estimate updating and the choice rules. This suggests a simple way to test our theory in the laboratory: make people play a coordination game such as the one just described, measure if a majority of players believe the alternative chosen by most people to be the superior one (even in cases where it is in fact the inferior alternative) and evaluate if the pattern of estimates can be well explained by an information bias in favor of the popular alternative.

Our assumptions regarding sampling and estimate updating define a heuristic that people can use to play this coordination game. But people can also adopt other heuristics. A widely studied strategy for playing coordination games is the best-reply strategy (e.g. Young, 1998). A decision maker uses a 'best-reply' strategy when she selects the alternative that has the highest subjective expected payoff, assuming that other players will choose the same alternative as they did in the prior period (choice is randomized if more than one alternative has maximal subjective expected payoff). We ran 10,000 simulations of the game, assuming that players use the best-reply strategy, with \(N=10, \mu_{1}=1, \mu_{2}=2, \sigma=3\), \(c=0.1\) and \(\lambda=0.5\). Simulations show that the quality estimates of the players will tend to synchronize with each others, in a fashion similar to what happens in the model we have analyzed earlier in the paper. For example, after 20 periods, the correlation between the quality estimates of Player 1 for Alt. 1 and the sum of the estimates of Players 2 to 10 is 0.64 . Furthermore, the players' choices sometimes coordinate on the inferior alternative (with the above parameters, the likelihood that 6 or more players prefer Alt. 1 is about \(16 \%\) ). This is not surprising because, if \(c\) is high enough, coordination on the inferior alternative is a Nash equilibrium.

\section*{Discussion and Conclusion}

Our model illustrates a novel mechanism that explains why groups may be reluctant to switch to another practice, even when this other practice is superior. Our mechanism complements existing explanations that show how a concern for popularity can lead to lock-in. More precisely, our model
demonstrates how a concern for popularity can also influence evaluations of the qualities of the alternatives even when people learn only from their own private observations. Our theory thus provides a simple mechanism for why public conformity in behavior (a tendency to choose popular alternatives) may lead to private acceptance (the belief that the popular alternative is the best) at the individual level and to collective illusions at the level of the group.

Existing psychological explanations of this synchronization of beliefs with behavior, such as cognitive dissonance theory, or self-perception theory, attribute it to motivated cognition (Bem, 1972; Festinger, 1957; Heider, 1958). Our explanation does not challenge the experimental evidence underlying those explanations. Rather, it suggests a complementary explanation that is likely to be important in realistic settings where popularity affects choices and access to payoff information. A distinctive feature of our analysis is that it assumes that people are good processors of information, but are naïve with respect to the nature of the sample they use to form their quality estimates. This is a standard assumption of the research program on sampling explanations of judgment biases that has received broad empirical support (Fiedler, 2012; Fiedler \& Juslin, 2006; Juslin et al., 2007). We do not claim that people do not engage in motivated cognition. Instead, we point to a fundamental asymmetry regarding the information sample that people have about popular v.s. unpopular alternatives. Because cognition operates on the available sample of information, our sampling explanation operates at a different level of analysis and thus complements explanations that focus on cognitive processes.

How do our results relate to herding models (Banerjee, 1992; Bikhchandani et al., 1992), which also explain collective failures to identify the best alternative? These models also assume that people are good processors of information. But contrary to these models, we do not assume that agents use popularity as a signal of quality. Therefore, our model may be more suitable for contexts where the assumptions of herding models do not apply. In particular, our model fits contexts where people have different tastes or do not believe that others know best, but where their sampling behavior is still influenced by others (Sutton \& Barto, 1998).

Our theory also differs from explanations of collective mistakes that attribute them to a coordination failure due to network externalities (e.g. Elster, 1978; North, 1990). In fact, our theory suggests that, if network externalities affect sampling behavior, the group may not only converge to the inferior alternative but, when this happens, most agents will also come to believe that the inferior alternative is of superior quality. Thus, in a vote about whether to switch to another alternative, most people would favor sticking to the status quo even if it is actually inferior. By contrast, explanations that rely on a coordination failure predict that people will switch if a vote could be organized and switching cost were low. Despite this difference in prediction, our theory complements explanations based on payoff externalities by suggesting that
payoff externalities can have a systematic effect on quality estimates that reinforces the possibility of a lock-in.

More generally, our results point to the systematic effect of unbiased experiences on beliefs. From a policy perspective, this illustrates the potential benefits of exposing initially skeptical adopters to an unpopular practice. Agents may appear to be resistant to unpopular practices not because they are risk averse or conservative, but because their own experiences with the unpopular practice are often skewed towards failures. In this case, inducing agents to try the unpopular practice again might help its acceptance, even when persuasive campaigns are not effective.

\section*{Appendix}

\section*{Preliminaries}

In this appendix, we analyze the asymptotic behavior of the model. Let \(\hat{Q}_{1, t}\) (resp. \(\hat{Q}_{2, t}\) ) be a random variable that refers to the quality estimate for Alt. 1 (resp. Alt. 2 ) of a randomly chosen agent, at the end of period \(t\). Let \(h_{t}\left(\hat{q}_{1, t}, \hat{q}_{2, t}\right)\) be the joint density of the quality estimates in period \(t\). Let \(P_{1, t}\) denote the proportion of the population choosing alternative 1 in period \(t\). Capital letters denote random variables, and corresponding lower case letters denote realizations of the random variables.

The expected proportion of agents choosing Alt. 1 in period \(t+1\) is:
\[
\begin{equation*}
E\left[P_{1, t+1} \mid p_{1, t}\right]=\iint_{\hat{q}_{1, t}, \hat{q}_{2, t}} p C_{1}\left(\hat{q}_{1, t}, \hat{q}_{2, t}, p_{1, t}\right) h_{t}\left(\hat{q}_{1, t}, \hat{q}_{2, t}\right) d \hat{q}_{1, t} d \hat{q}_{2, t}, \tag{4}
\end{equation*}
\]

One of the difficulties in analyzing the asymptotic behavior of equation 4 is that \(P_{1, t}\) is a random variable, which varies from period to period. It is reasonable to suspect, however, that \(P_{1, t}\) will converge to a constant, as \(t \rightarrow \infty\) and the number of agents increases. Let \(p_{1}\) denote this limiting proportion. The intuition is the same as for the law of large numbers: if the number of agents is very large, the proportion choosing Alt. 1 should converge to its expected value.
Another difficulty in solving the above equation is that \(h_{t}\left(\hat{q}_{1, t}, \hat{q}_{2, t}\right)\) will change over time. Nevertheless, there is reason to believe that as \(t \rightarrow \infty, h_{t}\left(\hat{q}_{1, t}, \hat{q}_{2, t}\right)\) will converge to a stationary distribution, denoted \(h\left(\hat{q}_{1}, \hat{q}_{2}\right)\). As \(t \rightarrow \infty\), more agents are added to the system and a larger proportion of agents will have had extensive opportunities to sample the two alternatives. While a new agent, with random estimates, is added to the system in each period, the influence of such agents should become vanishly small over time.

To calculate the equilibrium the model converges to, we adopt both of these simplifying assumptions. That is, we will assume, without a rigorous proof, that \(P_{1, t}\) converges to a constant \(p_{1}\) and that \(h_{t}\left(\hat{q}_{1, t}, \hat{q}_{2, t}\right)\) converges to a stationary distribution \(h\left(\hat{q}_{1}, \hat{q}_{2}\right)\). Under these assumptions, the following self-consistency equation must hold for \(p_{1}\) :
\[
\begin{equation*}
p_{1}=\iint_{\hat{q}_{1}, \hat{q}_{2}} p C_{1}\left(\hat{q}_{1}, \hat{q}_{2}, p_{1}\right) h\left(\hat{q}_{1}, \hat{q}_{2}\right) d \hat{q}_{1} d \hat{q}_{2} . \tag{5}
\end{equation*}
\]

Moreover, for any given value of \(p_{1}\), the stationary joint density of the quality estimates, \(h\left(\hat{q}_{1}, \hat{q}_{2}\right)\), must satisfy
\[
\begin{align*}
h\left(\hat{q}_{1}, \hat{q}_{2}\right)= & \int_{r} h\left(r, \hat{q}_{2}\right) p C_{1}\left(r, \hat{q}_{2}, p_{1}\right) \tau_{1}\left(r, \hat{q}_{1}\right) d r \\
& \quad+\int_{k} h\left(\hat{q}_{1}, k\right) p C_{2}\left(\hat{q}_{1}, k, p_{1}\right) \tau_{2}\left(k, \hat{q}_{2}\right) d k . \tag{6}
\end{align*}
\]

Here, \(p C_{2}\left(\hat{q}_{1}, k, p_{1}\right)\) is the probability that an agent with quality estimates \(\hat{q}_{1}\) and \(k\) will choose Alt. 2 and \(\tau_{k}\left(r, \hat{q}_{k}\right)\) is the probability mass that the quality estimate for Alt. \(k\) transitions from \(r\) to \(\hat{q}_{k}\)
given that the agent samples Alt. \(k\). Because the new estimate equals \(\hat{q}_{k}=(1-\lambda) r+\lambda z_{k}\), where \(z_{k}\) is normally distributed with mean \(\mu_{k}\) and variance \(\sigma^{2}, \tau_{k}\left(r, \hat{q}_{k}\right)\) equals the probability mass that \(z_{k}\) is equal to \(\left(\hat{q}_{k}-(1-\lambda) r\right) / \lambda\).

To explain the above equation (eq. 6), note that the terms to the right add up to the probability that the quality estimates for alternatives 1 and 2 are \(\hat{q}_{1}\) and \(\hat{q}_{2}\), after an agent has sampled one of the alternatives. The first term on the right hand side is the probability that the quality estimates for alternatives 1 and 2 are \(\hat{q}_{1}\) and \(\hat{q}_{2}\), and that the agent sampled Alt. 1 in the previous period. This set of estimates can only emerge, after the agent samples Alt. 1 , if this agent's estimate of the quality of Alt. 2 was equal to \(\hat{q}_{2}\). Similarly, the second term on the right hand side is the probability that the quality estimates for alternatives 1 and 2 are \(\hat{q}_{1}\) and \(\hat{q}_{2}\) and that the agent sampled Alt. 2 in the previous period.

Below, we show how one can solve for \(h\left(\hat{q}_{1}, \hat{q}_{2}\right)\), for any value of \(p_{1}\). Using \(h\left(\hat{q}_{1}, \hat{q}_{2}\right)\), we can then solve for the equilibrium value of \(p_{1}\). Finally, we derive the expected quality estimates in the stationary state.

\section*{The Stationary Distribution of the quality estimates}

For a given value of \(p_{1}\), the quality estimates for a representative agent follow a discrete time markov process, with a general state space \(\mathbb{R} \times \mathbb{R}\). Because there is a positive probability, in any period, that the system could transition from any state to another, the markov process has a unique stationary distribution, which has to satisfy equation 6 . The unique joint density that satisfies this equation is
\[
\begin{equation*}
h\left(\hat{q}_{1}, \hat{q}_{2}\right)=K g_{1}\left(\hat{q}_{1}\right) g_{2}\left(\hat{q}_{2}\right)\left[e^{-a p_{1}-b \hat{q}_{1}}+e^{-a\left(1-p_{1}\right)-b \hat{q}_{2}}\right], \tag{7}
\end{equation*}
\]
where \(K\) is a normalizing constant, i.e.,
\[
\begin{equation*}
1 / K=\iint_{\hat{q}_{1}, \hat{q}_{2}} g_{1}\left(\hat{q}_{1}\right) g_{2}\left(\hat{q}_{2}\right)\left[e^{-a p_{1}-b \hat{q}_{1}}+e^{-a\left(1-p_{1}\right)-b \hat{q}_{2}}\right] d \hat{q}_{1} d \hat{q}_{2} \tag{8}
\end{equation*}
\]
and \(g_{k}(y)=\int_{r} g_{k}(r) \tau_{k}(r, y) d r\), i.e., \(g_{k}(\cdot)\) is the distribution of the random variable the estimate of Alt. \(k\) would converge to if the probability that Alt. \(k\) is selected were equal to 1 in every period. When the quality distribution of Alt. \(k\) is normally distributed with mean \(\mu_{k}\) and variance \(\sigma^{2}\), it can be shown that \(g_{k}(\cdot)\) is a normal density with mean \(\mu_{k}\) and variance \(\sigma^{2} \lambda /(2-\lambda)\).

Using appropriate algebraic manipulations, it can be easily verified that the joint density in eq. 7 satisfies the stability equation 6 . The explicit formula for the normalizing constant is:
\[
\begin{equation*}
1 / K=e^{\frac{b^{2} \sigma^{2} \lambda}{2(2-\lambda)}}\left(e^{-a p_{1}} e^{-b \mu_{1}}+e^{-a\left(1-p_{1}\right)} e^{-b \mu_{2}}\right) \tag{9}
\end{equation*}
\]

The stationary joint density of the estimates is
\[
\begin{align*}
& h\left(\hat{q}_{1}, \hat{q}_{2}\right)= \\
& \frac{\frac{1}{2 \pi \frac{\lambda}{2-\lambda} \sigma^{2}} e^{\frac{-\left(\hat{q}_{1}-\mu_{1}\right)^{2}}{2 \frac{2}{2-\lambda} \sigma^{2}}} e^{\frac{-\left(\hat{q}_{2}-\mu_{2}\right)^{2}}{22-\lambda} \sigma^{2}}\left[e^{-a p_{1}-b \hat{q}_{1}}+e^{-a\left(1-p_{1}\right)-b \hat{q}_{2}}\right]}{e^{\frac{b^{2} \sigma^{2} \lambda}{2(2 \lambda)}}\left(e^{-a p_{1}} e^{-b \mu_{1}}+e^{-a\left(1-p_{1}\right)} e^{-b \mu_{2}}\right)} . \tag{10}
\end{align*}
\]

The probability that an agent will consider Alt. 1 to have a quality higher than Alt. 2 is thus
\[
\begin{equation*}
P\left(\hat{Q}_{1}>\hat{Q}_{2}\right)=\int_{\hat{q}_{1}=-\infty}^{+\infty}\left(\int_{\hat{Q}_{2}=-\infty}^{\hat{q}_{1}} h\left(\hat{q}_{1}, \hat{q}_{2}\right) d \hat{q}_{2}\right) d \hat{q}_{1}, \tag{11}
\end{equation*}
\]
which can be computed by numerical integration.

\section*{The equilibrium value of \(p_{1}\)}

In the stationary state, the probability that Alt. 1 is selected is given by equation 5 . The appropriate substitutions and algebraic manipulations imply
\[
\begin{equation*}
p_{1}=\frac{1}{1+e^{-2 a\left(p_{1}-0.5\right)} e^{-b\left(\mu_{1}-\mu_{2}\right)}} \tag{12}
\end{equation*}
\]

The values that \(p_{1}\) could converge to are the stable roots of the selfconsistency equation (eq. 5), which can be found numerically for any value of \(a\) by using equation 12 . When the value of \(a\) is sufficiently low, such as when \(a=1\), there is only one root and this is above 0.5 . However, if \(a\) is sufficiently large, such as when \(a=3.5\), the equation has three roots, one close to zero, one close to one, and another one, between the two other roots. The intermediary root is unstable, however. That is, the derivative of right-hand-side of the stationarity equation, evaluated at the intermediary root, is higher than 1. As a result, any small disturbance will tend to move the system away from the intermediary root.

\section*{The expected quality estimates}

The marginal distribution of \(\hat{q}_{1}\), given \(p_{1}\), is given by integration of the joint distribution (eq. 10) over \(\hat{q}_{2}\). After simplifications, we get:
\[
\begin{equation*}
E\left[\hat{Q}_{1}\right]=\mu_{1}-\left(1-p_{1}\right) \frac{b \lambda}{(2-\lambda)} \sigma^{2} \tag{13}
\end{equation*}
\]
which is an increasing function of \(p_{1}\). A similar calculation gives \(E\left[\hat{Q}_{2}\right]=\mu_{2}-p_{1} \frac{b \lambda}{(2-\lambda)} \sigma^{2}\), which is a decreasing function of \(p_{1}\). Moreover,
\[
\begin{equation*}
E\left[\hat{Q}_{1}\right]-E\left[\hat{Q}_{2}\right]=\mu_{1}-\mu_{2}+\left(2 p_{1}-1\right) \frac{b \lambda}{(2-\lambda)} \sigma^{2} \tag{14}
\end{equation*}
\]
which is an increasing function of \(p_{1}\). Thus, \(E\left[\hat{Q}_{2}\right]>E\left[\hat{Q}_{1}\right]\) even if \(\mu_{2}<\mu_{1}\) whenever
\[
\begin{equation*}
\left(1-2 p_{1}\right) \frac{b \lambda}{(2-\lambda)} \sigma^{2}>\mu_{1}-\mu_{2} \tag{15}
\end{equation*}
\]

The maximum difference is obtained when \(p_{1}=0\), and
\[
\begin{equation*}
\frac{\mu_{1}-\mu_{2}}{\sigma^{2}}<\frac{b \lambda}{(2-\lambda)} \tag{16}
\end{equation*}
\]

\section*{Acknowledgements}

Gaël Le Mens was supported by Spanish Ministry of Science and Innovation Grant \# ECO2010-17145, and the Barcelona School of Management.

\section*{References}

Arthur, W. B. (1989). Competing Technologies, Increasing Returns, and Lock-In by Historical Events, Economic Journal 97:642-65.
Banerjee, A. V. (1992). A Simple Model of Herd Behavior. Quarterly Journal of Economics 107:797-817.
Bem, D. J. (1972). Self-perception theory. In Advances in experimental social psychology (Academic Press, New York, NY)., Vol 6 pp 1-62.
Bikhchandani, S., Hirshleifer, D. \& Welch, I. (1992). A Theory of fads fashion custom and cultural change as informational cascades. Journal of Political Economy 100:9921026.

Busemeyer, J.R., Myung, I. J. (1992). An adaptive approach to human decision making: Learning theory decision theory and human performance. Journal of Experimental Psychology: General 121:177-194.
Cialdini, R. B., Goldstein, N.J., Social Influence: Conformity and Compliance. Annual Review of Psychology 55:591621.

Curtis. V., Cairncross, S., Yonli, R. (2000). Domestic hygiene and diarrhoea - pinpointing the problem. Tropical Medicine and International Health, 5(1):22-32.
Denrell, J. (2005). Why most people disapprove of me: Experience sampling in impression formation. Psychological Review 112:951-978.
Denrell, J., Le Mens, G. (2007). Interdependent Sampling and Social Influence. Psychological Review 114 (2).: 398422.

Duflo, E., Kremer, M., \& Robinson, J. (2009). Nudging Farmers to Use Fertilizer: Theory and Experimental Evidence from Kenya. NBER Working Paper No. 15131.
Elster J (1978). Logic and Society: Contradictions and Possible Worlds (Wiley, New York, NY).
Erev, I., Barron, G. (2005). On adaptation maximization and reinforcement learning among cognitive strategies. Psychological Review 112:912-931.
Fazio, R. H., Eiser, J. R., \& Shook, N. J. (2004). Attitude Formation Through Exploration: Valence Asymmetries. Journal of Personality and Social Psychology, 87(3), 293-311.
Festinger, L. (1957). A theory of cognitive dissonance (Stanford Univ Press, Palo Alto, CA).
Gibbons, R. (1992). Game theory for applied economists. Princeton University Press.
Hardin, G. (1968). The Tragedy of the Commons. Science, 162(5364), 1243-1248.
Heider, F. (1958). The psychology of interpersonal relations (Wiley, New York, NY).
Keynes, J. M. (1936). General Theory of Employment Interest and Money (Macmillan, London, UK).
Le Mens, G., Denrell, J. (2011). Rational learning and information sampling: On the naïvety assumption in sampling explanations of judgment biases. Psychological Review, 118:379-392.
Munshi, K. (2004). Social learning in a heterogeneous population: technology diffusion in the Indian Green Revolution. Journal of Development Economics, 73:185-213.
North, D. (1990). Institutions, institutional change, and economic performance (Cambridge Univ Press, Cambridge, UK).
Pfeffer, J. (1998). The Human Equation (Harvard Business School Press, Cambridge MA).
Scharfstein, D. S., Stein, J. C. (1990). Herd behavior and investment. The American Economic Review 80(3):465-479.
Sutton, R., Barto, A. G. (1998). Reinforcement learning (The MIT Press, Cambridge, MA).
Young, H. P. (1998). Individual Strategy and Social Structure. (Princeton University Press, Princeton).

\title{
Speed Facilitation In The Absence Of Enhanced Recognition For Target-Aligned But Irrelevant Stimuli Under Cross-modal Presentations
}

\author{
Andrew D. Dewald (adewald@hawaii.edu) \\ Department of Psychology, University of Hawaii at Manoa \\ 2530 Dole Street, Honolulu, HI 96822 \\ Scott Sinnett (ssinnett@hawaii.edu) \\ Department of Psychology, University of Hawaii at Manoa \\ 2530 Dole Street, Honolulu, HI 96822
}

\begin{abstract}
An ignored stimulus is later recognized at enhanced levels if it had previously been aligned with a target from a separate task. This has been demonstrated using both visual and auditory presentations. Here we extend these findings to multisensory conditions. Participants were required to detect immediate repetitions in a sound or picture stream while ignoring superimposed words presented in the opposite modality (either written or spoken, respectively), and then underwent a surprise recognition test for these words. Contrary to the previous unisensory examples (Dewald, Sinnett, \& Doumas, in press; Dewald \& Sinnett, 2012), a significant difference between recognition rates for target-aligned and non-aligned words was not observed. However, a highly significant difference in response latency was observed, with target-aligned words being responded to much more quickly. This finding was robust and observed when the surprise test was presented in either the visual or auditory modalities, as well as across modalities.
\end{abstract}

Key words: Attention, Multimodal Presentation, Response latency, Cross-modal processing.

\section*{Introduction}

Investigations of the relationship between attention and perception have demonstrated significant learning enhancements for certain stimuli in the absence of focused attention (Seitz \& Watanabe, 2003, 2005; Watanabe, Náñez, \& Sasaki, 2001). However, in order to observer these enhancements a number of compulsory prerequisite conditions were required. These included extended exposure rates of unattended stimuli (a random dot motion display) that were presented below threshold (a subset of dots moved coherently and subliminally) and also temporally aligned with a target from an attended secondary task. Under such conditions, enhanced learning performance was observed for the unattended stimuli in later motion discrimination tasks (see, Seitz \& Watanabe, 2003, 2005; Watanabe et al., 2001). Curiously however, when presenting the same type of stimuli (coherent motion) under the same conditions, but at levels that are easily perceptible (i.e., suprathreshold), the aforementioned learning enhancements vanish (Tsushima, Sasaki, \& Watanabe, 2006; Tsushima, Seitz, \& Watanabe 2008). Thus, it appears that the relationship between whether or not learning enhancements occur for irrelevant
stimuli is dependent on whether the initial presentation is sub- or suprathreshold. It is important to note that the investigations that collectively posit this idea have exclusively used random-dot, coherent motion displays (Seitz \& Watanabe, 2003, 2005; Watanabe et al., 2001; Tsushima et al., 2006; Tsushima et al., 2008). A natural ensuing question, therefore, would be whether these findings apply to stimuli that arguably demand a higher level of processing?

Directly addressing this question, Dewald, Sinnett, and Doumas (in press) adapted Seitz and Watanabe's (2003, 2005, see also, Watanabe et al., 2001) motion detection task to include a high-level irrelevant semantic stimulus (words) in an inattentional blindness paradigm (see Rees et al., 1999 for a similar example of the paradigm). Specifically, participants were required to respond to immediate picture repetitions in a stream of serially presented line drawings, while at the same time ignore a simultaneously presented stream of superimposed words. The irrelevant word stream contained a single, unchanging word aligned with the presence of an immediate picture repetition (i.e., targetaligned) as well as seven additional words that were superimposed over the non-repeated pictures (non-aligned; i.e., analogous to exposure frequencies used by Seitz \& Watanabe, 2003). The findings demonstrated that, despite attention being directed away from the task-irrelevant items (i.e., the words), subsequent recognition of these previously irrelevant items was nevertheless enhanced. Critically, this enhancement only occurred for words that had been presented simultaneously with a task-target in the previous task (i.e., target-aligned) when compared to non-aligned irrelevant words.

Similar enhancements for target-aligned stimuli have been observed when measuring recognition performance for irrelevant pictures that had appeared with targets (geometric shapes) in a separate task (e.g., the attentional boost effect; see Swallow \& Jiang, 2010). Collectively, the findings by both Dewald et al. (in press) and Swallow and Jiang seem to paint a different picture than what was described earlier. That is, explicit presentations lead to an enhancement in recognition performance for previously target-aligned items. This is the exact opposite of the inhibited performance observed when explicit motion presentations were used as the irrelevant stimulus (see Tsushima et al., 2006; Tsushima
et al., 2008). Dewald et al. argue that the high saliency of the irrelevant stimuli (i.e., written words rather than a lower level stimulus) likely underpins the difference in findings, assuming that the previous requisite condition of simultaneous presentation is met.

Regardless of the direction of learning effects, the critical component appears to whether the irrelevant stimulus is temporally aligned with a task-relevant target in a previous task (Seitz \& Watanabe, 2003). As these investigations have largely been conducted only in the visual modality, it is important to extend these findings to other sensory modalities in order to determine whether they extend beyond the visual domain. Our recent work (Dewald \& Sinnett, 2012) recently explored this very question by presenting an analogous paradigm using spoken words and sounds (i.e., rather than pictures). A facilitation for targetaligned irrelevant stimuli was observed. Interestingly however, the enhanced performance occurred only when the surprise recognition task was presented in either the same modality as the initial presentation (audition) or across modalities (i.e., audiovisual presentations).

Despite vision being the dominant sense in humans (Chandra, Robinson, \& Sinnett, 2011; Colavita, 1974; Posner, Snyder, \& Davidson, 1980; Sinnett, Spence, \& Soto-Faraco, 2007), it is clear that the human perceptual experience is multisensory in nature. Thus, it is important to explore if the learning effects for irrelevant stimuli within the same sensory modality extend across modalities, as this will further inform how information is processed as a consequence of attentional allocation both within, and across modalities. Generally, performance improves when comparing multisensory to unisensory presentations (see for example Duncan, Martens, \& Ward, 1997; Sinnett et al., 2006; Toro, Soto-Faraco, \& Sinnett, 2005; Wickens, 1984).

The enhanced recognition performance for cross-modal presentations, when compared to unimodal presentations, can be explained by numerous findings that suggest that the capacity of the attentional system is increased if a demanding unisensory task is divided across multiple sensory modalities (i.e., multiple resources theory, see Wickens, 1984). For instance, Sinnett et al. (2006) demonstrated that under multimodal presentations, inattentional blindness for words was ameliorated (i.e., perception improved) when compared with unimodal conditions, regardless of the modality of word presentation (see also Toro et al., 2005 for a similar example involving statistical learning). These findings seem to provide support for an attentional system that is segregated, such that each sensory modality has access to individualized attentional resources (Wickens, 1984, see also Duncan et al., 1997 for an example using the attentional blink)

In the present investigation, we extend unimodal examples of learning enhancements for task irrelevant but target-aligned stimuli, to multimodal presentations. As increased performance has been observed for such presentations (see Duncan et al., 1997; Sinnett et al., 2006), we would expect an overall increase in recognition
performance for both target-aligned and non-aligned items if they are presented in a separate sensory modality from a temporally aligned task-relevant target (e.g., more attentional resources will be available for non-aligned words). Of particular interest is whether or not the comparatively higher scores for target-aligned words will persist under cross-modal presentations. Interestingly, this could possibly jettison the enhancement associated with target-alignment if performance for non-aligned words increases substantially (i.e., a ceiling effect). We presented participants with multisensory visual and auditory streams (adapted from those used in the unimodal conditions in Dewald et al., in press and Dewald \& Sinnett, 2012). This resulted in one of the streams including spoken words with distracting pictures, and the other having written words with distracting sounds. The task was to respond to repetitions in the target stream (i.e., sounds or pictures) and then to subsequently recognize as many words that had been previously presented (i.e., ignored) in the repetition detection task.

The present study also investigates the nature of the surprise recognition task. With the exception of our previous work in the auditory modality (Dewald \& Sinnett, 2012), all research involving this paradigm has presented the recognition task in the visual modality, regardless of whatever modality it was presented in during the repetition detection task. As irrelevant stimuli in the exposure portion of the experiment will be presented in either the auditory or visual sensory modalities, it is necessary to examine if subsequent recognition of these items is affected by whether presentation is in a congruent modality. Our previous work (Dewald \& Sinnett, 2012) did precisely this and systematically manipulated the modality of presentation between exposure and recognition tests. Not surprisingly, when irrelevant items were presented for recognition in the same modality as the exposure (i.e., both visually or both auditorally), learning effects were observed. However, when irrelevant stimuli were presented for recognition in an incongruent modality from their initial exposure, learning enhancements failed to surface for irrelevant items that had been temporally aligned with task-relevant targets (Dewald \& Sinnett, 2012). Lastly, cross-modal presentations lead to the greatest magnitude of enhancement for the previously aligned words in the surprise recognition test. This latter outcome dovetails with previous investigations of attentional allocation across sensory modalities in perceptual and recognition tasks, suggesting that crossmodal presentations generally lead to superior performance when compared to unimodal presentations (Dewald \& Sinnett, 2011; Duncan et al., 1997; Sinnett et al, 2006; Toro et al., 2005). Accordingly, in the present experiment we also presented the surprise recognition tests in the same or different sensory modality, or across modalities. If primary and secondary task modality congruence is a factor as it was in Dewald and Sinnett (2012), then we expect improved results for congruent matchings vs. incongruent matchings between exposure and recognition tasks, and potentially an
additional enhancement for multimodal presentations (simultaneous visual and auditory presentation of the stimulus in the recognition test) given that performance is generally enhanced for multisensory presentations (see Driver \& Spence, 2004). Note, these modality specific enhancements were only seen for target-aligned items.

\section*{Method}

Participants. Seventy-four participants \(\quad(\mathrm{n}=74)\) were recruited from the University of Hawai'i at Manoa in exchange for course credit. A total of 46 participants were assigned to the visual words and sounds condition and a total of 28 participants assigned to the auditory words and pictures condition. The uneven distribution of participants across all conditions was a consequence of convenience sampling. Participants were naïve to the experiment and had normal or corrected to normal vision and hearing. Written informed consent was obtained before participation in the experiment occurred.

Materials. The exact same stimuli and design to create streams were used here as in Dewald et al. (in press, for visual stimuli) and Dewald and Sinnett (2012, for auditory stimuli) except now with multimodal presentations (i.e., pictures presented with spoken words or sounds presented with written words).

Attending to pictures with spoken words. A total of 50 pictures were selected from the Snodgrass and Vanderwart (1980) picture database. Each of the pictures (on average 5 to 10 cm , rotated \(+/-30\) degrees from upright so as to ensure difficulty) was combined with eight one-to-two syllable, high frequency English words (average length of five letters; range 4-6) selected from the MRC psycholinguistic database (Wilson, 1988). The overall average frequency of the eight selected words was 361 per million, ranging between 135 and 782. For the auditory presentation of the words, a native English speaker's voice was recorded reading the list of selected words three times. Three blind listeners chose the best exemplar of each spoken word, with a fourth listener deciding which one was best in the event of a tie. The selected recordings were edited using sound editing software so that all items were the same presentation length ( 350 ms ) and average amplitude.

A stream of 960 picture-spoken word concatenated items was created, with repeated pictures acting as task relevanttargets. The presentation stream was broken into eight blocks of trials (120 each) in which an immediate picture repetition occurred on average one out of every eight trials, equating to 15 task-relevant target repetitions per block, for a total of 120 trials of exposure to a task-relevant target (and specific word, see below). Only eight total words were superimposed over the 960 pictures. Note then that all word types (aligned or non-aligned) were presented in equal proportions (120 times each). This was done to parallel the number of different motions used in Watanabe et al, (2001; see also Seitz \& Watanabe, 2003, 2005), so as to expose the
participants to an unchanging, single, irrelevant word, although also having seven additional irrelevant words all exposed at the same frequency. The same single word was always temporally aligned with the presentation of an immediately repeated picture target. The presentation was pseudorandomized so that on average one out of every eight trials was an immediate picture repetition (and, therefore, the presentation of the same superimposed task-irrelevant target word). Only one superimposed word was aligned with all of the immediately repeated pictures for each participant.

Attending to sounds with written words. The exact same procedure as above was employed but now with sounds, instead of pictures, serving as the task-targets, and visually presented words as the irrelevant stimuli. The sound stimuli were extracted from a database of 100 familiar sounds and were also edited to 350 ms and similar average amplitude (see Sinnett et al., 2006). All other aspects were identical to the previous condition (pictures and spoken words).

Surprise recognition task: For both conditions, a surprise recognition test for the presented words was administered after the completion of the repetition detection task. The test consisted of a total of sixteen words (i.e., half came from the previously presented words, while the other half consisted of foil words that had never been presented before, average frequency of 236 per million with a range of 165-399. The word recognition tasks were randomized and presented by DMDX software (http://www.u.arixona.edu/jforster/dmdx.htm) one at a time, in either the visual or auditory modality, or across modalities. For the visual presentation the words were written in bold, capitalized letters in Arial font at a size of 24 points, and remained on the screen until a response was made. For auditory presentations the words were spoken just as they were in the initial repetition detection task. Cross-modal presentations involved the written word on the screen with the spoken word presented simultaneously.

\section*{Procedure}

Participants were required to attend to the sound (or picture) stream (i.e., they were explicitly instructed to ignore the simultaneously presented, overlaid written/spoken words) and respond to immediate repetitions by pressing the ' \(G\) ' key on the keyboard of the computer. Each item in the sound-word (or picture-word) presentation was presented for 350 ms with a \(150-\mathrm{ms}\) inter-stimulus interval (ISI; silence) for a stimulus onset asynchrony (SOA) of 500 ms . Before the first experimental block, a training block of eight trials was given and repeated until participants were familiar and comfortable with the task. Immediately after the repetition detection task, the surprise word recognition test was administered to all participants (modality type of surprise task dependent on condition). Participants were instructed to press the "B" key if they had heard the word during the repetition detection task or, instead, the "V" key if they had not heard the word before.

\section*{Results}

Target detection accuracy in the repetition detection task. Overall performance accuracy (across all conditions) of immediate target repetition detection revealed that participants were successful at detecting target repetitions in the primary task, ( \(72 \%\) hit rate vs. \(28 \%\) miss rate, \(t(73)=\) 14.67, \(p<.001\) ).

Overall recognition accuracy. Across all conditions, participants were accurate in recognizing the unattended words (both target-aligned and non-aligned) displayed during the repetition detection task at better than chance levels \((86.1 \% S E=1.47, t(73)=17.35, p<.001)\). A threefactor mixed design ANOVA was used to analyze overall (across all conditions) recognition performance for all words. Surprise test modality (auditory, visual, or crossmodal) and exposure modality (visual words vs. auditory words) were between-subjects factors, and target alignment (target-aligned or non-aligned) was a within-subjects factor. There were no main effects for target alignment \((\mathrm{F}(1,68)=\) \(.217, \mathrm{p}=.643\) ), exposure modality in the primary task ( F \((1,68)=2.68, \mathrm{p}=.08\) ), or surprise test modality ( \(\mathrm{F}(2,68\) ) \(=.548, \mathrm{p}=.580\) ). A planed comparison further demonstrated that target-aligned and non-aligned words were recognized at statistically indistinguishable rates, across all conditions (target-aligned: 89.1.0\%, \(S E=3.06\); non-aligned: \(83.8 \%, S E=2.15, t(68)=1.30, p=.195\), Figure 1). Given these null results, no further analyses of recognition performance were conducted.


Figure 1. Recognition percentages pooled across all conditions for Target-aligned words (black bar) and NonAligned (grey bar) words.

Overall recognition speed. To explore if response latency to the words was modulated by target alignment or the modality of presentation of the surprise task, the same threefactor mixed design ANOVA was conducted as above, with surprise test modality (auditory, visual, or cross-modal) and exposure modality (visual words vs. auditory words) as between-subjects factors, and target alignment (target-
aligned or non-aligned) as a within-subjects factor. A main effect of target alignment confirmed that overall, the speed of responding to words was significantly faster for targetaligned words ( \(787.8 \mathrm{~ms}, S E=21.8\) ) when compared to non-aligned words ( \(1378.2 \mathrm{~ms}, S E=80.2\) ) ( \(\mathrm{F}(1,68)=\) \(52.44, \mathrm{p}=.001\) ) (see Figure 2). No main effects were observed for surprise test modality ( \(\mathrm{F}(1,68)=.298, \mathrm{p}=\) .587 ) or exposure modality ( \(\mathrm{F}(2,68)=1.80, \mathrm{p}=.173\) ), nor were any interactions significant except for the three-way interaction ( \(\mathrm{F}(2,68\) ) \(=3.58, \mathrm{p}=.03\) ). To further explore this interaction, further ANOVAs of response speed for each condition were conducted.


Figure 2. Response latencies pooled across all conditions for Target-Aligned (black bar) and Non-Aligned (grey bar) words.

Attending to sounds with written words. A two factor mixed design ANOVA was conducted for response latencies with surprise test modality as a between subjects factor and target-alignment as a within-subjects factor. A main effect of target alignment \((\mathrm{F}(1,43)=34.97, \mathrm{p}<.001)\) was observed, demonstrating that participants responded more quickly to target aligned ( \(812.2 \mathrm{~ms}, S E=29.3\) ) when compared with non-aligned words ( \(1371 \mathrm{~ms}, S E=93.8\) ). There was no main effect for surprise test modality ( \(\mathrm{F}(2,43\) ) \(=.237, \mathrm{p}=.790\) ) nor was there a significant interaction \((\mathrm{F}(2\), 43) \(=1.28, \mathrm{p}=.286\). Planned comparisons also confirmed that when examining response latency in the surprise word recognition task, target-aligned words were responded to significantly faster than non-aligned words in all conditions (Visual Presentation: Target-aligned: \(789.2 \mathrm{~ms}, S E=38.4\) vs. Non-aligned: 1293.7, \(S E=217.4, t(16)=2.92, \mathrm{p}=.02\); Auditory Presentation: Target-aligned: \(717.8 \mathrm{~ms}, S E=63.4\), vs. Non-aligned: \(1519.2 \mathrm{~ms}, S E=38.4, t(11)=5.31, \mathrm{p}=\) .001; Multimodal Presentation: Target-aligned: 901.9 ms , \(S E=46.1\) vs. \(1320 \mathrm{~ms}, S E=72.7,(t(16)=4.86, \mathrm{p}=.001)\). Further confirming the non-significant interaction, there were no significant differences in performance between conditions (all p>.58).

Attending to pictures with spoken words. The same two factor mixed design ANOVA was conducted for response
latencies with surprise test modality as a between subjects factor, and target-alignment as a within-subjects factor. Again, a main effect for target alignment \((\mathrm{F}(1,25)=19.57\), \(\mathrm{p}<.001\) ) was observed, demonstrating once more that participants responded more quickly to target aligned (747.8 \(\mathrm{ms}, S E=30.9\) ) when compared with non-aligned words ( \(1404.5 \mathrm{~ms}, S E=147.2\) ). while no main effect was observed for the modality of the recognition test \((\mathrm{F}(2,25)=1.51, \mathrm{p}=\) .239). The interaction also failed to reach levels of significance \((\mathrm{F}(2,25)=2.81, \mathrm{p}=.079)\). Planned comparisons also confirmed this in each modality presentation in the surprise recognition task (Visual Presentation: Target-aligned: \(659.7 \mathrm{~ms}, S E=67.8\) vs. Nonaligned: 1833.9, \(S E=368.9, t(8)=2.84, \mathrm{p}=.02\); Auditory Presentation: Target-aligned: \(819.8 \mathrm{~ms}, S E=44.5 \mathrm{vs}\). Nonaligned: \(1159.4 \mathrm{~ms}, S E=116.5, t(9)=3.10, \mathrm{p}=.01\); Multimodal Presentation: Target-aligned: \(755.9 \mathrm{~ms}, S E=\) 35.5 vs. \(1247.3 \mathrm{~ms}, S E=200.9, t(8)=2.42, \mathrm{p}=.04)\). Despite the marginal interaction, there were no significant differences in performance between conditions (all p>.05)..

\section*{Discussion}

There are a number of outcomes that necessitate discussion, as the present findings strengthen the understanding of how unattended information is processed when it appears simultaneously with an attended target, especially when considering the multimodal exposures used here. Specifically, the findings exhibit that both presentation types (pictures with auditory words, or sounds with visually presented words) lead to learning effects, exemplified by high recognition rates in the surprise task, despite attention not being directed to the words. This is similar to analogous paradigms using only unimodal visual (Dewald et al, in press) or auditory (Dewald \& Sinnett, 2012) presentations. However, both of these unimodal studies indicated enhanced recognition rates for target-aligned words when compared with non-aligned words. This was not the case with cross-modal presentations, as observed here. That is, although the recognition rates for the unattended stimuli were high, there was no difference between target-aligned and non-aligned items.

The lack of a significant difference in recognition rates based on target alignment is likely due to the cross-modal presentations used here. It is possible that the division of the task could have permitted additional attentional resources to focus on processing all of the words, as shown by the high recognition rates for non-aligned words here (overall 84\%). While it is difficult to statistically compare this rate to our previous studies (already published), it is worth noting that, in analogous but unimodal paradigms, performance for nonaligned words was much lower in either the visual ( \(68 \%\), Dewald et al., in press) or auditory modality (59\%, Dewald \& Sinnett, 2012). Thus, it appears that by presenting the repetition detection task across modalities, additional resources were available that potentially enabled the processing of irrelevant stimuli, resulting in arguably near
ceiling recognition rates for both aligned and non-aligned words. This dovetails well with other research demonstrating enhanced performance under multimodal conditions (Duncan et al., 1997; Sinnett et al., 2006), possibly indicating a segregation of attentional resources across modalities (Wickens, 1984).

Despite the lack of a recognition difference between target-aligned and non-aligned items, the former were responded to significantly faster, regardless of the modality of presentation (target-aligned: 787.8 ms vs. non-aligned: 1378.2 ms ), suggesting alignment did play a role. That is, it is possible that there was improved learning of words that were temporally aligned with a task-relevant target, indicated by response latencies to target-aligned words being faster in all three recognition conditions (visual, auditory, audiovisual). This is an intriguing finding as it indicates a conceivable enhancement for target-aligned material without explicit awareness, as there were no differences in recognition performance. Although, it should be acknowledged that recognition performance might have been at ceiling levels and therefore masked any possible improvement for target-aligned words. Regardless, this finding warrants discussion, as well as further research. Indeed, of the many studies published on this topic (see, Dewald et al, in press, Dewald et al., 2011; Dewald \& Sinnett, 2012; Rees et al, 1999; Sinnett et al, 2006; Swallow \& Jiang, 2010; Tipper \& Driver, 1988) the present experiment is the first to use response latency as a potential measure of enhancement for target-aligned material.

Also of key interest here, is that we did not observe an interaction in performance between target-alignment and the modality of the surprise test, as was observed by Dewald and Sinnett (2012). Across all conditions, regardless of the congruency between presentation and recognition task, there was no significant difference between target-aligned and non-aligned words. Accordingly, this suggests that, at least in the present case, under multimodal presentation, the modality of presentation does not need to match exposure and test conditions. This could be a byproduct of the overall enhanced recognition performance seen after cross-modal presentations. A more systematic approach manipulating presentation (unimodal vs. cross-modal) and surprise test (congruent, incongruent, cross-modal) is required before ruling out that this factor is unnecessary.

Collectively, the present findings provide insight into how irrelevant information is processed when it is presented simultaneously with an attended target across sensory modalities. If certain prerequisite conditions are met, unattended stimuli can be perceived and affect behavior, perhaps even below levels of conscious awareness. Additionally, although a significant difference was not observed here, future research should consider the congruency of modality presentation in both exposure and testing conditions.

\section*{References}

Broadbent, D.E. (1958). Perception and communication. London: Pergamon Press.
Dewald, A.D., Sinnett, S., \& Doumas, L.A.A. (in press). A Window of Perception When Diverting Attention? Enhancing Recognition for Explicitly Presented, Unattended, and Irrelevant Stimuli by Target Alignment. Journal of Experimental Psychology: Human Perception and Performance, 1(39)
Dewald, A.D., \& Sinnett, S. (2012). An inhibited recognition performance for explicitly presented target-aligned irrelevant stimuli in the auditory modality. Proceedings of the Twenty-Ninth Annual Conference of the Cognitive Psychology Society
Dewald, A.D., Sinnett, S., \& Doumas, L.A.A. (2011). Conditions of directed attention inhibit recognition performance for explicitly presented target-aligned irrelevant stimuli. Acta Psychologica, 138, 6067.

Dewald, A.D., Sinnett, S., \& Doumas, L.A.A. (2010). The inhibition and facilitation of stimuli can be modulated by the focus of direct attention. Proceedings of the Twenty-Eight Annual Conference of the Cognitive Psychology Society
Driver, J., \& Spence, C. (2004). Cross-modal spatial attention: Evidence from human performance. In C. Spence \& J. Driver (Eds.), Cross-modal space and cross-modal attention. Oxford, UK: Oxford University Press.
Duncan J, Martens S, Ward R (1997), Restricted attentional capacity within but not between sensory modalities. Nature 387(6635):808-10
Lin, J.Y., Pype, A.D., Murray, \& Boynton, G.M. (2010). Enhanced memory for scenes presented at relevant points in time. PLoS Biol, 8(3), E1000337.
Lupker, S. J. (1984). Semantic priming without association: A second look. Journal of Verbal Learning and Verbal Behavior, 23, 709-733.
Rees, G., Russell, C., Frith, C. D., \& Driver, J. (1999). Inattentional blindness versus inattentional amnesia for fixated but ignored words. Science, 286, 2504-2507.
Seitz, A. R., Kim, R., \& Shams, L. (2006). Sound facilitates visual learning. Current Biology, 16, 1422-1427.
Seitz, A. R. \& Watanabe, T. (2003). Psychophysics: Is subliminal learning really passive? Nature, 422, 36.

Seitz, A. R. \& Watanabe, T. (2005). A unified model for perceptual learning. Trends in Cognitive Science, 9 (7), 329-334.

Seitz, A. R. \& Watanabe, T. (2008). Is task-irrelevant learning really task-irrelevant? PLoS ONE 3(11): e3792
Sinnett, S., Costa, A., \& Soto-Faraco, S. (2006). Manipulating inattentional blindness within and across sensory modalities. Quarterly Journal of experimental Psychology, 59(8), 1425-1442
Snodgrass, J. G., \& Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. Journal of Experimental Psychology: Human Learning and Memory, 6, 1 74-215.
Spence, C., \& Squire, S. (2003). Multisensory integration: Maintaining the perception of synchrony. Current Biology, 13, R519-R521.
Spence, C., \& Driver, J. (1998). Cross-modal links in spatial attention. Pilosophican Transactions of the Royal Society :Biological Sciences, 352(1373), 13191331.

Swallow K. M., \& Jiang, Y. V. (2010). The attentional boost effect: Transient increases in attention to one task enhance performance in a second task. Cognition, 115, 118-132.
Tshushima, Y., Sasaki, Y., \& Watanabe, T. (2006). Greater disruption due to failure of inhibitory control on an ambiguous distractor. Science, 314, 1786-1788.
Tsushima, Y., Seitz, A. R., \& Watanabe, T. (2008). Taskirrelevant learning occurs only when the irrelevant feature is weak. Current Biology, 18(12), 516-517.
Watanabe, T., Náñez,Y., \& Sasak, S. (2001). Perceptual learning without perception. Nature, 413, 844848.

Wickens, C. D. (1984). "Processing resources in attention", in R. Parasuraman \& D.R. Davies (Eds.), Varieties of attention, (pp. 63-102). New York: Academic Press.
Wilson, M. D. (1988). The MRC psycholinguistic database: Machine readable dictionary, version 2. Behavioural Research Methods, Instruments and Computers, 20, 6-11.

\title{
The Great Deceivers: Virtual Agents and Believable Lies
}

\author{
João Dias (joao.dias@gaips.inesc-id.pt) \\ INESC-ID and IST, Universidade Técnica de Lisboa, Portugal
}

\author{
Ruth Aylett (r.s.aylett@hw.ac.uk) \\ Heriot-Watt University, Edinburgh, Scotland
}

\author{
Henrique Reis and Ana Paiva (henrique.reis@ist.utl.pt, ana.paiva@inesc-id.pt) \\ INESC-ID and IST, Universidade Técnica de Lisboa, Portugal
}

\begin{abstract}
This paper proposes a model giving Theory of Mind (ToM) capabilities to artificial agents to allow them to carry out deceptive behaviours. It describes a model supporting an N -level Theory of Mind and reports a study to assess whether equipping agents with a two-level ToM results in them being perceived as more socially intelligent than agents with a singlelevel ToM. A deception game being developed for intercultural training of children, used for this study, is described. Finally, we report results from this study consistent with the hypothesis that a two-level Theory of Mind better supports agents in deceptive behaviour.
\end{abstract}

Keywords: Virtual Agents; Theory of Mind; Deception

\section*{Introduction}

The work reported in this paper arises from the use of synthetic graphical characters interacting in rich virtual worlds. These may be required for interactive drama applications (Mateas \& Stern, 2003), or for story-based education and training applications (Paiva et al., 2004) (Swartout et al., 2006). A key criterion for success is that such agents be believable, that is lead a user, or viewer, to feel that they have an inner life of their own, with goals, motivations and emotions, and are in some sense 'alive' (Bates, 1994). Thus interaction between such characters must display features related to human-human interaction; whether the actions they carry out, their emotional expressions, ability to exhibit empathy, or non-verbal as well as verbal communications. Such features must be contextually appropriate, and in order to achieve this, characters may be driven by an architecture uniting cognitive and affective models, for example using a cognitive appraisal approach (Dias \& Paiva, 2005) (Marsella \& Gratch, 2009).

Computationally implemented cognitive appraisal models are often naive, assuming entirely open behaviour, sometimes referred to as meeting the sincerity condition (Searle, 1976). However, this is unusual in everyday human-human communication where deception often occurs. This may be as simple as masking anger in front of a social superior or fear in front of a child on a dark night (Rosis, Pelachaud, Poggi, Carofiglio, \& Carolis, 2003), (Prendinger \& Ishizuka, 2001), or as complex as deliberately misleading or lying to another person in order to gain an advantage. Deceptive behaviour includes not only the generation of false beliefs in others but also the claiming of desired identities, the exchange of non-existent emotions, and the communication of false preferences or opinions (Wyer \& Epstein, 1996). Thus decep-
tion can be seen as a human-like characteristic that would enhance the believability of synthetic characters portrayed in real world social situations.

A Theory of Mind (ToM) process allows an agent to attribute an artificial mental state to another agent and reason about it. In a single-level ToM, agent A can represent only its belief about what an agent \(B\) is thinking; an agent \(C\) that can not only model what B is thinking but can also model what agent \(B\) thinks about \(C\) has a two-level ToM. In this paper we investigate the hypothesis that an agent with a single-level ToM will be less successful in believable deception then an agent with a two-level ToM. Deception cannot be investigated in abstract but requires a concrete scenario. Our work uses an interactive game played by and with autonomous graphical characters. This is based on the popular game Mafia, or Werewolf, described below, in which deception is fundamental to successful play. The characters are implemented with a cognitive appraisal-based architecture (Dias \& Paiva, 2005) that includes a deliberative mechanism and has been extended to support an N -level ToM mechanism.

\section*{Background and Related Work}

We define a "lie" as a direct communicative act that an agent performs to deceive another agent. We consider deception through verbal mechanisms - speech acts - though deception may also be achieved through non-verbal mechanisms. Deception has been widely studied in AI, though usually with disembodied software agents.

GOLEM (Castelfranchi \& deRosis, 1998) is based on the blocks world of AI planning research. Goals conflict, since agents aim to build different structures from the same available blocks. Agents can achieve goals through their own actions or by asking for "help" from others. Agents have task delegation and adoption preferences and different capabilities, used to plan their actions based on their knowledge of other agents. Deception is instrumental, resulting only from goal conflicts, though it extends to deception about capabilities, goals or personality. However, agents in GOLEM can only produce lies within this limited scope. They cannot for example lie about the requests they have made or plan to make. This would require second order reasoning about the reasoning of other agents, which is not present here.

De Rosis and Carofiglio (deRosis, F; Carofiglio, V; Grassano \& Castelfranchi, 2003) focus on the communicative per-
spective of a deceptive action. In their scenario, a Sender agent tries to convince a Receiver agent that some fact X is not true, where the Sender can lie or use other deceptive strategies. Their system, "Mouth of Truth" implements reasoning models as belief networks (Neapolitan, 1990; Pearl, 1997), where nodes represent belief and probabilities across links to other node represent uncertainty. This allows the Sender to lie not about the belief they want to manipulate, but about one connected to it. Thus uncertainty can be increased for the belief "it rained" if the Sender claims "the floor outside is dry". However, the Sender needs a model of the Receiver's beliefs to be able to do this and so acts as if its own set of beliefs and reasoning rules is replicated in the Receiver. This can then be used to influence the decision making process of the Sender.

The work so far discussed did not ground deception in an explicit model of other agents. Theory of Mind is a term coined by (Premack \& Woodruff, 1978) who define it as the ability to infer the full range of epistemic mental states of others, i.e. beliefs, desires, intentions and knowledge. This is a mechanism that helps to make sense of the behaviour of others in specific contexts and to predict their next action.

Recent work (Harbers \& Meyer, 2009) focuses on a computational implementation of ToM, giving agents the capacity to interact in a believable way with trainees, and to explain their actions and decisions after the training is over. The agents model a trainee's mind and give feedback either through simple action decisions, or by an explanation at the end. Meyer et al. here combined two prominent but conceptually different approaches to the human theory of mind: the Theory-Theory approach (TT) and the Simulation-Theory approach (ST).

In TT, the mental state we attribute to others is not observable, but is knowable through intuition and insight. Implementationally, this is achieved by using inference rules to reason about the beliefs of others. On the other hand, ST claims that each person simulates being another while trying to reason about their epistemic state, using the same structures and processes as those updating their own beliefs and knowledge (Aylett \& Louchart, 2008). Meyer et al. showed that the main difference lay in ease of implementation rather than in outcome, as ST models are better in terms of code re-usability and modularity. Moreover, the TT approach can only deal with BDI (Beliefs Desires Intentions) models (Rao \& Georgeff, 1995) due to a rigid representation of the mental state of other agents in terms of beliefs, limiting it to a specific symbolic representation.

PsychSim (Pynadath \& Marsella, 2005) is a multi-agent based simulation tool for modeling interactions using a decision-theoretic approach. Unlike most such frameworks, where agents select actions maximizing rewards using their own beliefs, PsychSim agents also take into account their beliefs about other agent's beliefs. These recursively- "nested beliefs" may include subjective views of the agent itself. Agents update their beliefs according to the changes in the
world and their subjective interpretations of world dynamics. In particular, messages are implicit ways through which one agent may influence the beliefs of another.

Wagner and Arkin developed algorithms to give an an intelligent robot the ability to deceive (Wagner \& Arkin, 2010). The Deceiver seeks to induce a false belief in another agent, the Target, who is modeled as an action model and utility functions with associated outcomes matrix for a specific situation. This involves performing some action in the environment transmitting a false communication to the Target, so that it will behave in a way benefiting the Deceiver. This modifies the outcome matrix for the Target, the induced outcome matrix. Wagner and Arkin showed that knowledge of the Target affected the success of a deceit attempt. However this work did not explore the implications of different levels of ToM. Although there are systems that implemented a Theory of Mind in agents, and interesting projects on deception, we believe this is the first generic model that combines the two in a way that is flexible enough to be featured in a game. Further, we also show a study that compares different levels of abstraction in the way agents are perceived in terms of lying.

\section*{A Mindreading Agent Model}

Our agent ToM is based on the Mindreading model of (BaronCohen, 1995), and follows the ST approach of Meyer et al.see Figure 1. A central Knowledge Base (KB) stores the agent's beliefs and world knowledge and is the foundation for the agent's behaviour given that its actions are based on its knowledge.


Figure 1: Proposed model for a Mindreading Agent
The ToM has three components, following Baron-Cohen \({ }^{1}\) : the EDD (Eye Direction Detector), SAM (Shared Attention Model), and ToMM (Theory of Mind Mechanism). EDD determines who sees what, while SAM constructs higher level relations between entities (John sees that Luke sees the book). The ToMM represents and stores the mental states of other agents and is used to influence or deceive another agent. However, a deceiving agent must also be able to plan and reason about the consequences of its own actions. Thus our model includes a Deliberation component giving planning capabilities using knowledge from the KB and the ToMM to select the best actions for the agent to perform to meet its current goals.

\footnotetext{
\({ }^{1}\) There is an additional component, the Intentionality Detector but to simplify our model it was not included
}

\section*{Representing Models of Others}

Each Model of Other in the ToMM represents the beliefs of a specific Other the agent knows. A single-level theory of mind allows us to represent an agents's beliefs about another agent's beliefs. However, human adults are able to model more than one level (e.g. beliefs about another's beliefs about another's beliefs). Children start to develop a second level of ToM at around the age of six. Thus agents intended to function believably at the level of older children - as in the Werewolf game used as a study - require a model with more than one level of ToM.

A specific Model of Other contains its own ToMM also containing Models of Others, creating a recursive hierarchical tree-like structure - see Figure 2.


Figure 2: ToMM Hierarchy: 3 agents and 2 levels
Thus three agents, A, B, and C, each with a two-level ToM modeling ability, need six models each. If agents include a three-level ToM, this rises to fourteen models, and with four levels, to thirty models. The more complex the tree structure for the model hierarchy, the more effort is required for each update cycle. We will focus on a two-level ToM, bearing in mind that in the human case, applying more than two levels also causes a substantial overhead. More levels could be used in exchange for a slower reasoning cycle.

The ST approach represents others by simulating ones own processes in that same situation. Hence a ToMM Model of Other corresponds to a simplified version of the Agent Model depicted in Fig. 1, including both data structures and processes. A Model of Other can therefore be updated with a given percept through the same process used to update the agent's own model.

\section*{Updating Models of Others}

When a given percept is received (e.g. a property has changed, or an action was performed), the agent updates its KB and its Models of Others. This is done through the EDD and SAM components.

The EDD determines what entities, objects, and events are perceived by other agents. It first checks whether a target agent is within a certain radius or in the same location as the agent, and if so, asserts that it also receives the percept.However this does not deal with more complex percepts such as a whisper into an ear, where only the specific receiving agent will know what was said. Hence the EDD may also include domain-specific rules about actions with particular restrictions on the perceptual mechanism. A rule specifies information about the action (such as subject, action name,
target, parameters) and associates it with a list of effects. Two main types of effects are used in these rules:
- Global effect - effect of an action assumed to be perceived and shared by everyone (who is close enough). E.g. *:Werewolf(Rob) represents that everyone can perceive Werewolf(Rob).
- Local effect - an effect perceived only by a particular agent. E.g. John:Werewolf \((\) Rob \()\) represents that only John will perceive Werewolf(Rob).

When EDD receives percept \(P\), it determines two lists, perceptionVisibilities and agentVisibilities. The perceptionVisibilities list contains all pairs \(A g: P\), such that agent \(A g\) perceives proposition \(P\), while the agentsVisibilities list contains all pairs of form \(A g: A g\), stating which agents see which other agents. SAM uses this to update Models of Others. It traverses the tree hierarchy, establishing whether a Model \(M\) should perceive \(P\) applying the following test:

\section*{1. Test if Model \(M\) is contained in perceptionVisibilities.}
2. Test if the pair Predecessor \((M): M\) is contained in the agentVisibilities list. Predecessor \((M)\) returns the predecessor of model \(M\) in the tree hierarchy.
3. If both tests are verified, then model \(M\) can perceive \(P\), otherwise the algorithm stops following the remaining subtree and continues the recursive process.

For example, suppose three agents, A, B and C. When A receives a percept \(P\), it will update its own \(K B\) with \(P\), but will also process P in its ToM to update models for B and C . Further, suppose that \(A\) knows that both \(B\) and \(C\) perceived \(P\), and also knows that B does not see C (so it will not see that C perceives P). In this situation A's Model of \(B\) will be updated with P but A's model of B's Model of C (second level) will not be updated.

\section*{Using the ToMM Information}

Agents have two reasoning mechanisms, one forwards (from data to conclusions) using inference rules, and one backwards (from goals to actions needed to achieve them) used to create plans that achieve the agent's goals. An inference rule is a tuple \(\langle R, P, E\rangle\) where \(R\) is the name of the rule, \(P\) (Preconditions) is a list of propositions that need to be verified for the rule to be applied, and \(E\) (Effects) a list of propositions that will be added to or removed from the KB when the rule is applied. Whenever new knowledge is added to the KB, the deliberation component will test the preconditions of the existing Inference rules. If any rule is fired (i.e. its preconditions are verified) the deliberation component will automatically update the KB with the effects in its effects list. If this process adds a new proposition to the KB , the inference process will be repeated until no more changes are verified.

The second mechanism involves goals, plans and actions. A goal is a tuple \(\langle G, P, S\rangle\) where \(G\) is the Goal's name, \(P\) a list of propositions that correspond to the goal's preconditions, and \(S\) a list of propositions that correspond to the goal's
success conditions (i.e. the desired goal state). The deliberation component is constantly checking to see if any goal becomes active by testing its preconditions. Once a goal becomes active, the planner tries to build a plan of actions to achieve the goal's success conditions. The actions used by the planner are defined using a STRIPS-like (Fikes \& Nilsson, 1971) formalism and correspond to a tuple \(<A g, A, P, E>\) where \(A g\) is the agent who performs the action, \(A\) is the action's name, while \(P\) and \(E\) correspond to a list of preconditions and effects. Given the similar representations, Inference Rules can also be used by the planner to build plans of actions; the difference is that when an Inference Rule is selected for execution (when the agent is executing the plan) it is not returned as an action to be performed in the environment. For more details about these mechanisms, please refer to (Aylett, Dias, \& Paiva, 2006).

The first step in making the ToM information available to the deliberation component is to allow the specification of preconditions that are not tested against the agent's own KB but using a particular Model of Other. This is done by specifying explicitly the Model of Other to be tested by representing preconditions as a list of colon separated agents followed by a proposition \(A g_{1}: \ldots: A g_{n}: P\). When the deliberative component finds such a precondition it starts by traversing the tree hierarchy of Models of Others using the list of colon separated agents, and selecting the corresponding Model Of Other. Then the proposition P is tested using the selected Model of Other's KB. As example, \(A: B: \operatorname{Suspects}(A)\) is true if Suspects \((A)\) is true in the Model of \(B\) that is stored in the agent's Model of \(A\) (intuitively representing "I think that A thinks that B suspects him to be the Werewolf"'). If a proposition \(P\) does not specify a Model of Other it will be tested against the agent's own model, in other words, its own KB.

Using preconditions this way allows us to specify goals and inference rules triggered according to beliefs of others. It would be even more useful to model higher-level goals and inference rules, i.e. explicit goals and rules to change the mental states of others. To do so, we use the same mechanism used to specify local and global effects as described previously. An effect is specified as \(A g_{1}: \ldots: A g_{n}: P\), where \(A g_{i}\) is an an agent's name, or the symbol "'*", and represents that only the Models of Others obtained by the list \(A g_{1} \ldots: A g_{n}\) will have the proposition \(P\) added to its KB. The symbol "**" represents that all Models of Others at that particular level will be selected. The planner was extended to be able to handle matching and detection of conflicts between preconditions and local/global effects. In planning terms, a precondition is matched or threatened by a local effect only if their agents lists are compatible and if they refer to the same proposition \(P\). In its simplest version, two agents lists are compatible if they have the same size and the agents are unifiable (the symbol "**" unifies with everything). As examples, the effect \(A: B\) :Suspects \((C)\) matches the precondition \(A: B: \operatorname{Suspects}(C)\), but does not match the precondition \(B: A: \operatorname{Suspects}(C)\), whilst \(A: *: \operatorname{Suspects}(C)\) matches
both \(A: B: \operatorname{Suspects}(C)\) and \(A: D: S u s p e c t s(C)\).
When an inference rule has an effect specified with an agents list (e.g \(A g: P\) ), instead of updating its own KB , the deliberation component will traverse the tree hierarchy in order to update the corresponding Models of Others. Moreover, the ST approach means that the Model of Other corresponds to a version of an Agent Model with its own inference mechanism. When creating a Model of Other, the agent assumes that others will use the same inference rules as its own. Therefore, every update cycle, the inference mechanism will also be executed recursively for each Model of Other. In other words, the agent will simulate other's inference processes, and update the corresponding models. This process is applied even if the effects of the inference rule specify an agent's list. For instance, if the Model of Other of John at level 1, applies an inference rule that results in the effect Rob:Suspects(John), it will update John's Model about Rob's Model at level 2.

Due to its greater complexity, we did not include goal selection/planning, and thus simplified the version of the Agent used as a Model of Other. The agent is therefore not capable of simulating the planning process of others.

\section*{Case Study}

The model above was used to build NPCs that deceive in a system for intercultural training, MIXER (Hall et al., 2011). This is aimed at children aged \(9-11\) and conflict between groups (an in- and out-group scenario) is presented through a social game. Rules act as cultural expectation and if they are varied, conflict will occur. Older children usually define rules before starting to play, but late primary children generally only discover the difference in rules when the conflict occurs, often with game abandonment and shouts of "'it's not fair"" and "'I don't want to play any more". The user acts as an invisible (out-of-game) friend to a character thrust into this situation with the pedagogic aim of showing that the existence of different rules is not the same thing as 'cheating'. MIXER uses variations of the game Werewolf, or Mafia \({ }^{2}\).

A simplified version of the game involves five players, the Villagers, who are divided into two groups, one Werewolf and four potential Victims. Victims have limited information, since they do not know who the Werewolf is (they are 'killed' at night). Characters can be human players or NPCs (Non Playable Characters) running the architecture supporting deception. The goal is to discover who is the Werewolf: the character who is lying.. The Werewolf must lie purposefully: its objective is to remain hidden until no longer outnumbered by Victims. Thus it tries to eliminate Victims while concealing its true identity.

The game has been implemented in turn-based rounds. In each round every character performs the Accuse action in order, naming another character as the Werewolf (see Figure3 ). The Werewolf deceptively accuses one of the victims, knowing they are not in fact the Werewolf. At the end of each turn,

\footnotetext{
\({ }^{2}\) http://en.wikipedia.org/wiki/Mafia_(party_game)
}
the agreed werewolf is excluded from the game and informs the other agents about its true identity. This is used to infer new information about past accusations. The real Werewolf wins if it reaches the last turn alive, when there is only one victim left. At this stage the Werewolf announces its identity. Victims win if they manage to discover who the Werewolf is before the last turn.


Figure 3: An agent performing the Accuse action

The following inference rules allow the victims to reason about past actions, trying to determine possible werewolf suspects:
- I suspect those that were accused by someone I don't suspect
- I stop suspecting someone who accuses a target I suspect
- I suspect those who accused a victim that was eliminated the previous round
- Someone who accuses a target suspects that they are a werewolf
- Someone that is accused will suspect the accuser

\section*{Modeling the Werewolf}

Two versions of the Werewolf agent were implemented. One has a single-level ToM, able to represent what victims believe, but not what victims think it or the other victims believe. The second has a two-level ToM, able to represent what victims think about what it knows and in general, what victims think about the suspicions of others. Both versions also have the inference rules above, used by victims to determine suspects.

The single-level Werewolf has two main strategies compatible with its single-level ToM: eliminate victims that suspect it, and make a victim suspect another victim who has not been accused yet. The second goal corresponds to changing the victim's beliefs, and can be modeled by the success condition \(\left[v_{1}\right]:\) Suspects \(\left(\left[v_{2}\right]\right)\), where \(\left[v_{1}\right]\) is a variable representing a victim and \(\left[v_{2}\right]\) is a variable representing another victim. These variables will be instantiated by the goal activation process, and the agent will then try to make \(\left[v_{1}\right]\) suspect \(\left[v_{2}\right]\).

The two-level Werewolf agent has a strategy commonly used by human players in this game. The agent will "'Lay low"', by avoiding suspicious actions, trying to make victims believe that it thinks the same way they do. This is modeled with the following second level success condition \([v]: S E L F: S u s p e c t s([\) target \(])\), where \([v]\) is a victim, [target \(]\) is another villager that \([v]\) suspects to be the Werewolf, and

SELF represents the Werewolf agent itself. Thus the twolevel ToM Werewolf will accuse villagers that are already being accused by other victims.

\section*{Tests and Evaluation}

Two tests were run comparing these two versions. A first simulation test assessed how well the two types of ToMs performed in the game. In order to test the hypothesis that an agent with a single-level ToM is less successful in believable deception then an agent with a two-level ToM, a second evaluation was conducted with users, assessing their perception of the single-level and two-level ToM Werewolves.

As an autonomous agent architecture is being used, scenarios are unscripted and do not run identically. To avoid different outcomes biasing user responses, a video of a particular run was used for the second test. The simulation test allowed us to select this video.

In the first test, two versions of the system were generated. The first was parameterized so that the Werewolf used the single-level ToM (ToM1 condition), and in the second it used the two-level ToM (ToM2 condition). The victims used a single level ToM in both conditions. With five players, the maximum number of possible rounds is 4 . Both versions ran ten times, from the beginning until the Werewolf was caught or won the game. The number of turns the Werewolf managed in each run was recorded. The video of the best scoring run for each version was used for the second test.

In the TOM1 condition, the Werewolf never lasted four rounds, and so did not win a single game. The ToM2 Werewolf won in two out of ten runs and on average lasted 0.6 more turns than the ToM1 version.

\section*{User Perception of the Lying Agents}

The second test evaluated user-perceptions of the two Werewolf versions' believability. An online questionnaire was used with the two videos selected from the first test. Sixty participants ( \(34 \mathrm{M}, 26 \mathrm{~F}\) ), of which 55 were aged 19-25, were recruited online, and randomly assigned to one of the two versions. They were asked to pay special attention to agents' actions and to try to work out who was lying. They watched the game and then rated affirmations using a Likert scale (ranged from -2 meaning totally disagree, to 2 meaning totally agree) in four sections: (1) affirmations about the game itself; (2) affirmations about all players; (3) the same affirmations as (2) but only for the the liar; (4) affirmations focused on deceptive behaviour. Data was analyzed using a non-parametric Mann-Whitney statistical test to compare conditions ToM1 and ToM2.

Participants perceived the ToM2 condition as more interesting according to A1: "'The game is interesting", ( \(p<\) \(0.05, r=-0.263\) ) and would play this version of the game more "'A2: I would play a game like this"'( \(p<0.05, r=\) -0.292). ToM2 scores were significantly lower ( \(p<0.05\) ) for A3: "'It is easy to win while playing as a Victim"' and significantly higher \((p<0.5)\) for A4: "It is easy to win while playing as a Werewolf"'. We conclude that participants
thought the liar did a more competent job in the ToM2 version.

Answers to A8: "'Players behaved in a predictable way", were also significantly different in the ToM2 condition ( \(p<\) \(0.001,|r|=0.5\) ): player characters were seen as less predictable in ToM1 than ToM2. This is seen as a surrogate for believability given the answers to A10: "'Players are easily deceived"' gave significantly lower values in ToM1 than in ToM2 ( \(p<0.001, r=-0.478\) ), reflecting the more believable performance of the Werewolf in ToM2.

Finally, two additional measures: "how well did the liar play" and its "intelligence" also lead to statistically significant differences between the two conditions ( \(p<0.001\) ) supported by a large effect size \((|r|=0.5)\). We conclude that the liar in ToM2 is perceived as more intelligent than in ToM1. Also statistically significant ( \(p<0.001, r=-0.467\) ) were answers to A15: The liar is affected by others' actions indicating that the Werewolf was seen as more responsive to the play of others in ToM2. Finally, the higher results in ToM2 for A21: The liar managed to deceive the other players ( \(p<0.001, r=-0.524\) ) confirm those for A10 above.

\section*{Conclusions}

This paper advances a model for virtual agents that are able to deceive, embedding a ToM mechanism inspired by work on the human ToM. The model can produce N -level ToM behaviour using a simulation approach, where the agent runs its own mechanisms, reasoning about the beliefs and actions of others as if it was in their shoes. Parametrization allows the number of levels of ToM to be easily varied.

Evaluation was carried out using a social game, MIXER, for intercultural training of children aged 9-11. This game includes one character, the Werewolf, that must lie in order to play successfully. The first test showed that when a Werewolf was given a two-level rather than single-level ToM and played against Villagers with a single-level ToM, the Werewolf's game performance improved. The user testing with 60 subjects showed that participants clearly perceived the ToM2 version Werewolf as better at deceiving the other agents, and, furthermore, saw this as more intelligent behaviour. These results support the hypothesis that an agent with a single-level ToM will be less successful in believable deception then an agent with a two-level ToM.

As future work, it would be interesting to compare different combinations of the scenarios (e.g. one-level werewolf against two-level victims), and to include the simulation of other's planning processes in order to make it posible to reason about other agent's goals and plans. Another interesting extension to this work would be to apply the model to a different type of deception than verbal lies, for example to deceptive display of affective states (Rosis et al., 2003).

\section*{Acknowledgments}

This work was partially supported by the European Community (EC), through the ECUTE project (ICT-5-4.2 257666), and by national funds through Fundação para a Ciência e a Tecnologia (FCT),
under project PEst-OE/EEI/LA0021/2011. The authors are solely responsible for the content of this publication. It does not represent the opinion of the EC or the FCT, which are not responsible for any use that might be made of data appearing therein.

\section*{References}

Aylett, R., Dias, J., \& Paiva, A. (2006). An affectively-driven planner for synthetic characters. In Proc. icaps.
Aylett, R., \& Louchart, S. (2008). If I were you: double appraisal in affective agents. In Aamas-volume 3 (pp. 1233-1236).
Baron-Cohen, S. (1995). Mindblindness: An Essay on Autism and Theory of Mind. MIT Press. Paperback.
Bates, J. (1994). The role of emotion in believable agents. Communications of the ACM, 37, 122-125.
Castelfranchi, R., C; Falcone, \& deRosis, F. (1998). Deceiving in Golem: How to strategically pilfer help. In Autonomous agents: Workshop on deception, fraud and trust in agent societies. Kluwer.
deRosis, F; Carofiglio, V; Grassano, G., \& Castelfranchi, C. (2003). Can computers deliberately deceive? a simulation tool and its application to Turing's imitation game. Computational Intelligence, 19(3), 235-263.
Dias, J. a., \& Paiva, A. (2005). Feeling and reasoning: A computational model for emotional characters. In C. Bento, A. Cardoso, \& G. Dias (Eds.), Progress in artificial intelligence (Vol. 3808, p. 127-140). Springer.

Fikes, R. E., \& Nilsson, N. J. (1971). Strips: A new approach to the application of theorem proving to problem solving. Artificial Intelligence, 2(3-4), 189-208.
Hall, L., Lutfi, S., Nazir, A., Hodgson, J., Hall, M., Ritter, C., et al. (2011). Games based learning for exploring cultural conflict. In Aisb 2011 symposium: Ai \& games, york (pp. 6-7).
Harbers, K. v. d., Maaike; Bosch, \& Meyer, J.-J. (2009). Modeling agents with a theory of mind. In Proc. 2009 ieee/wic/acm int. joint conf. web intelligence and intelligent agent technology - v. 02 (pp. 217-224). Washington, DC, USA: IEEE.
Marsella, S., \& Gratch, J. (2009). Ema: A process model of appraisal dynamics. Cognitive Systems Research, 10(1), 70-90.
Mateas, M., \& Stern, A. (2003). Façade: An experiment in building a fully-realized interactive drama. In Game developers conference, game design track (Vol. 2, p. 82).
Neapolitan, R. E. (1990). Probabilistic reasoning in expert systems: theory and algorithms. New York, NY, USA: John Wiley \& Sons, Inc.
Paiva, A., Dias, J., Sobral, D., Aylett, R., Woods, S., Zoll, C., et al. (2004, July). Caring for agents and agents that care: Building empathic relations with synthetic agents. In Aamas'2004. ACM Press.
Pearl, J. (1997). Probabilistic reasoning in intelligent systems : networks of plausible inference. Morgan Kaufmann. Paperback.
Premack, D., \& Woodruff, G. (1978). Does the chimpanzee have a theory of mind? Behavioral and Brain Sciences, 1(04), 515-526.
Prendinger, H., \& Ishizuka, M. (2001). Social role awareness in animated agents. In Autonomous agents (pp. 270-277).
Pynadath, D., \& Marsella, S. (2005). Psychsim: Modelling theory of mind with decision-theoretic agents. In Ijcai (p. 1181-1186).
Rao, A. S., \& Georgeff, M. P. (1995). Bdi agents: From theory to practice. In Proc 1st int. conf. multiagent systems. San Francisco.
Rosis, F., Pelachaud, C., Poggi, I., Carofiglio, V., \& Carolis, B. (2003). From Greta's mind to her face: modelling the dynamics of affective states in a conversational embodied agent. Int. J. Human-Computer Studies, 59(1), 81-118.
Searle, J. (1976). A classification of illocutionary acts. Language in Society, 5, 1-23.
Swartout, W., Gratch, J., Hill Jr, R., Hovy, E., Marsella, S., Rickel, J., et al. (2006). Toward virtual humans. AI Magazine, 27(2), 96.

Wagner, A., \& Arkin, R. (2010). Acting Deceptively: Providing Robots with the Capacity forDeception. Int. J. Social Robotics, 1-22-22.
Wyer, B. D. D. K. S. K. M., \& Epstein, J. (1996). Lying in everyday life. Journal of Personality and Social Psychology, 70, 979-995.

\title{
One for all, all for one: Agents with social identities
}

\author{
Joana Dimas (joana.dimas@gaips.inesc-id.pt) \\ INESC-ID and Instituto Superior Técnico, Technical University of Lisbon \\ Av. Prof. Cavaco Silva, 2744-016 Porto Salvo, Portugal
}

Phil Lopes (plopes@gaips.inesc-id.pt)
INESC-ID
Av. Prof. Cavaco Silva, 2744-016 Porto Salvo, Portugal

\author{
Rui Prada (rui.prada@gaips.inesc-id.pt) \\ INESC-ID and Instituto Superior Técnico, Technical University of Lisbon \\ Av. Prof. Cavaco Silva, 2744-016 Porto Salvo, Portugal
}

\begin{abstract}
One of the challenges in developing multi-agent systems is the creation of agents able to exhibit human-like behaviours in complex social situations. In order to do so, agents need to be socially aware of their environment and perceive other agents not only as individuals but also as social group members. Following Social Identity and Self-Categorization theories, we developed the Dynamic Identity Model for Agents that provides agents with the ability to adapt their identity and behaviour to the social context. We then implemented it in a social dilemma scenario where different situations were explored.
\end{abstract}

Keywords: Identity; Social Identity; Social Dilemmas; Context-Situated Agents.

\section*{Introduction}

With virtual worlds' increasing complexity, where agents and players are exposed to different scenarios and social contexts, it has become even more important to develop agents whose identity does not remain unaffected, and in turn reacts to its environment in a believable way.

Although some works have been done on agent's identity adaptability, either through the agent's personality (Tan \& Cheng, 2007) or by their culture's background (De Rosis, Pelachaud, \& Poggi, 2004; Mascarenhas, Dias, Afonso, Enz, \& Paiva, 2009), these are adaptations to the player's traits, and does not address the influence of the social context. Moreover, each approach alone did not encompass both individual and social concepts of identity working together and dynamically.

In real life a person's identity is not static and free of influences (Turner, Oakes, Haslam, \& McGarty, 1994; Hogg \& Williams, 2000; Smith \& Mackie, 2000). Instead, several social context factors (Smith \& Mackie, 2000) are known to have an impact on an individual's identity and behaviour, with one of the most studied factors being the presence of in-group or out-group members. In fact, Social Identity (Tajfel, 1972) and Self-Categorization (Turner, Hogg, Oakes, Reicher, \& Wetherell, 1987; Turner et al., 1994) theories explained this process postulating that one's identity can both be personal and social. When in the presence of members of a person's in-group, the individual's behaviour is going to be determined
by its personal identity, and one will relate to others in an interpersonal manner, dependent on his or her personality traits and close personal relationships with others. However, when in the presence of an out-group, a social identity becomes salient, and the perception as group member strengthens, as a person tends to focus his or her perception on the shared features with other in-group members. Consequently, there is a shift of a person's own motives and values from self-interest to group interests (Brewer, 1991). When a social identity emerges, people are more likely to see themselves and others as interchangeable components of a larger social unit rather than unique individuals.

According to (Tajfel, 1972; Turner et al., 1987, 1994), this psychological process of social identification constitutes the basis for in-group cooperation. Because in-group members share the same attributes, they become part of a person's identity and due to this, a person will want to treat all in-group members as he or she would like to be treated. In fact, several studies have already demonstrated that social identity has a positive effect in in-group cooperation and negative effect in out-group cooperation (Goette, Huffman, \& Meier, 2006; McLeish \& Oxoby, 2007), but more specifically that it has an important role in eliciting cooperative behaviour in social dilemmas (Wit \& Wilke, 1992; Kollock, 1998; Weber, Kopelman, \& Messick, 2004). Social dilemmas are, in broad terms, social situations of individual rationality conflict where group interests are at odds with individual ones (Dawes, 1974), and thus making them an interesting application for agents with social identities.

Nonetheless, while some authors have already been modelling the concept of social identity and used it on simulations of crowd behaviour (Fridman \& Kaminka, 2009) or opinion dynamics, such political views (Grier, Skarin, Lubyansky, \& Wolpert, 2008; Lustick, 2002; Salzarulo, 2006), they still did not handle the dynamics of identity, nor have worked on its impact in social dilemmas situations. As such, we developed the Dynamic Identity Model for Agents (Dimas \& Prada, 2013) and implemented it in a social dilemma scenario in order to evaluate it in a game environment.

The paper is organized as follows. Next section we introduce the model, followed by the description of the model's implementation and the platform used to demonstrate the example scenarios we present on the following section. Finally we present some conclusions and future work.

\section*{Dynamic Identity Model for Agents}

The Dynamic Identity Model for Agents (DIMA) aims at providing agents with a dynamic identity that is determined by the social context.

\section*{Agent's Identity}

In DIMA, the agent's identity is not fixed, instead the agent features a sub-set of characteristics that represents the part of the identity that is currently salient on the agent. So in the model each agent has a salient identity that will filter the characteristics that will determine the agent's decision, and also a set of social groups that are known by the agent.
- Salient Identity: representation of the agent's active identity that is going to influence the agent's decision making;
- Social Groups’ Knowledge Base: representation of the agent's known social groups (aggregation of agents that share the same characteristics) and its prototypical characteristics (characteristics that represent the typical agent of that group).
While personal identity is the part of the self-concept defined in terms of idiosyncrasies derived from the intra-group differentiation (Tajfel, 1972), social identity refers to the aspects of a person's self-concept that are derived from the knowledge and feelings about his or her in-group (Tajfel, 1972). As such, the agent is not only going to be able to express its individual identity, but also, for each social group it belongs, the agent will hold a social identity that can be expressed if the situation leads it. In DIMA, an agent's salient identity can have two different levels. It can be social, if an agent's group membership becomes salient trough intergroup differentiation, or it can be personal when no social identity is salient. Thus, the agent's salient identity can be:
- Social: a set of characteristics that the agent shares with the other members of the in-group;
- Personal: a set of characteristics that distinguishes the agent from it's in-group.
In order to represent these two levels, both social and personal identities are defined by:
- Characteristics: representation of the agent's attributes or features that are going to be taken into consideration on the agent's decision making, defined by a name and value.
When an agent's salient identity is personal, the agent's decision will be determined by its personal identity characteristics values, but when the salient identity is social, i.e., then the agent's expressed characteristics' values are going to shift towards the values of the prototypical characteristics of that specific social group.

\section*{Characteristics}

Each characteristic is defined in DIMA by a name and a value:
- Type: a label used to identify the characteristic;
- Value: measurable attribute or feature.

Characteristics can be one of the two types: explicit or implicit. Whereas explicit characteristics can be easily observed and obtained by other agents (e.g. skin or clothes colour, symbols, skills and gender), implicit characteristic are gleaned indirectly by observing the agent's behaviour and expressions, requiring agents with inferring mechanisms. Implicit characteristics can be social values, norms, interests or goals (Hofstede \& Hofstede, 2001; Schwartz, 1992)

\section*{Social Context}

The social context the agent is in will have a great influence on how the agent will perceive itself and others. It will increase the likelihood of the agent behave according to its personal identity or to its social identity, and will also determine which type of identity is going to be salient and influencing the agent's behaviour.

In DIMA, two aspects from the social context are represented:
- Agents Present: agents that share the same space and agents that are not in the space but are referenced in a conversation or by an event.
- Theme: set of characteristics that are relevant in the context, and can be manifested by a place, a talk or an event;

When a specific theme is introduced on the social context, either by a place (e.g. a university), by a topic of a conversation (e.g. a talk about politics), by an event (e.g. travelling outside), or by a task (e.g. cleaning the classroom), the theme will bring out the characteristics that are relevant in that specific social context, and then this set of relevant characteristics is going to be processed by the agent.

It is while looking at each other agents' characteristics that the theme defines as relevant to the current situation, that the agent calculates and perceives if it is in the presence of members with which it shares the same social group (in-group) or not (out-group). If the agent perceives itself as in the presence of only in-group members, its identity is going to be determined by its personal identity. But if the agent is in the presence of out-group members, its identity can be determined by a social identity, according to a formula that we will see next.

\section*{Identity Salience}

Fundamentally, the identity that the agent is going to take in account when processing its decision-making and to generate its behaviour, is going to be determined by the presence or not of the the out-group (Brewer, 1979) but also by several other aspects inherent to the social identity itself. These factors are going to have an impact on the social identity salience
strength, and the more salient a social identity is, more is its influence on the agent's behaviour.

According to Social Identity and Self-Categorization theories (Tajfel, 1972; Turner, 1985; Turner et al., 1987), the salience of a particular social identity is determined by the interaction between how accessible in memory that social identity is to an individual (accessibility), as well as how well it fits the social context (fit) (Turner et al., 1987, 1994). Following (Oakes, 1987), in this model a social identity salience is the product of fit and accessibility (see equation 1 ).
\[
\begin{equation*}
\text { Salience }_{(\text {Social Identity })}=\text { Fit } * \text { Accessibility } \tag{1}
\end{equation*}
\]

Fit between a social identity and the context where the agent is situated is composed by two aspects: comparative fit and normative fit. Comparative fit is defined by the principles of the Meta-Contrast theory (Turner et al., 1987), which states that:
"any collection of people will tend to be categorized into distinct groups to the degree that intra-group differences are perceived as smaller, on average, than inter group differences within the relevant comparative context", p.455, (Turner et al., 1994)
Normative fit refers to the content of that categorization and how well does it match with the characteristics of a social group from the agent's knowledge base.

In order to determine the fit of a social identity with DIMA, first the agent needs to define the social groups present in the context given the actual theme.

All agents present in the social context are going to be clustered into categories, according to the relevant characteristics given by the theme. For this to be possible, all characteristics must have a numeric comparative function which returns the distance between two vectors ranging from 0 to 100 , where 0 means the absence of that characteristic and 100 means that it highly represents the agent.

According to the clustering algorithm results, the agent might perceive as being in the presence of one or more social groups. If the number of clusters is one, that means that the agent is in the presence of one social group. In this case, because of the absence of an out-group the salience of a social identity does not apply, and the agent will use its personal identity. Only in the presence of two or more groups, the agent proceeds in calculating the fit.

In this situation, through normative fit, the agent will be able to determine if it is in the presence of a social group that it already knows and had experience with. So for all social groups in the agent's knowledge base that has those relevant characteristics, the fit is computed by comparing them to all the clusters resulted from the previous clustering process. If no match is found, its because the agent is in the presence of ad-hoc groups (groups who the agent does not have previous knowledge or past experiences with). In those situations the prototypical member, or centroid ( \(C t\) ), of each social group that is going to be used later by the fit is going to be determined by the prototypical member of the present clusters. If
there is actually a match between the social groups found by the clustering algorithm, the agent will use the centroid from the normative social groups that it already knows. The process for computing the value of the normative fit is similar to the comparative fit described bellow.

Calculating the comparative fit of a social identity \(\left(S I_{i}\right)\) is going to be done according to the equation 2 where the distance between the agent's in-group \(\left(S G_{i}\right)\) and any other group \(\left(S G_{o}\right)\) is going to be calculated (inter group differences), and the dispersion of its own social group is measured (intragroup differences). Alfa ( \(\alpha\) ) and Beta \((\beta)\) are weighting values for both distance and dispersion, and since we want to attribute more weight to the distance than to the social group's dispersion, we set the default of \(\alpha\) as 0.8 and \(\beta\) as 0.2 .
\[
\begin{equation*}
\text { ComparativeFit }_{\left(S I_{i}\right)}=\alpha\left(\operatorname{distance}_{\left(S G_{i}, S G_{o}\right)}\right)+\beta\left(1-\operatorname{dispersion}_{\left(S G_{i}\right)}\right) \tag{2}
\end{equation*}
\]

The distance between the agent's group and another group present in the social context is going to be measured by calculating the difference between the out-group centroids \(\left(C t_{\left(S G_{o}\right)}\right)\), that represent the group's prototypical members, and the in-group centroids \(\left(C t_{\left(S G_{i}\right)}\right)\) (see equation 3). If the agent recognizes the groups through the normative fit process then the group's centroids used will be the prototypical members' characteristics from the social groups from the agent's social group's knowledge base, if not, it will be the prototypical members' characteristics of the clusters found trough the clustering algorithm.
\[
\begin{equation*}
\operatorname{distance}_{\left(S G_{i}, S G_{o}\right)}=\frac{\left|C t_{\left(S G_{o}\right)}-C t_{\left(S G_{i}\right)}\right|}{K m d} \tag{3}
\end{equation*}
\]

The dispersion of the agent's social group is measured by calculating the average of absolute differences (MD) of all its members from the prototypical member of the social group (see 4).
\[
\begin{equation*}
\operatorname{dispersion}_{\left(S G_{i}\right)}=\frac{M D_{\left(S G_{i}\right)}}{K m c w} \tag{4}
\end{equation*}
\]

Both distance and dispersion are normalized, using the constants Kmd and Kmcw, where:
- Kmd: is the maximum distance two clusters can hold, and can be calculated according to the equation 5 , where \(N\) is the number of characteristics used for clustering and MAX is the maximum value a characteristic can have;
- Kmcw: is the maximum distance between the centroid member and another member for it to be considered as member of that group. It is a parametrizable value, which is currently set to 50 .
\[
\begin{equation*}
K m d_{S G_{i}, S G_{o}}=\sqrt{N} * M A X \tag{5}
\end{equation*}
\]

Social groups with higher fit are the ones with less clustering dispersion and higher distance from the other social groups.

Accessibility of a particular social group, reflects a person's past experience with that group (Turner et al., 1994). Identities have higher or lower accessibility depending on how accessible is that specific categorization in a person's memory. Identities that have been used more times and displace more emotional valence are more accessible.

The accessibility from new social identities, as the ones from ad-hoc social groups, is implied by the distance between the agent values ( \(<c_{1}, \ldots, c_{n}>_{A g}\) ) and the centroid from its in-group. As such, agents that are closer to the centroid have higher accessibility while agents further from it have lower accessibility, translating this way the connection strength between a agent and that ad-hoc social group (see equation 6).
\[
\begin{equation*}
\operatorname{Acc}(S I)_{t=0}=1-\frac{\left|<c_{1}, \ldots, c_{n}>_{A g}-C t_{\left(S G_{i}\right)}\right|}{K m d} \tag{6}
\end{equation*}
\]

In the presence of normative groups the agent's social identity can have an accessibility value determined by the emotional memory and the easiness of bringing that social identity into the agent's mind (Turner et al., 1994). The emotional valence of a memory is defined by the emotional impact of the actions taken by the agent supported by that identity.

For every time a social identity is salient its accessibility is updated according to the equation 7. The sum of all agent's identities is normalized so when one identity accessibility increases all the others suffer a decay.
\[
\begin{equation*}
\text { Acc }(S I)_{t+1}=\operatorname{Acc}(S I)_{t}+\text { Salience }(S I)_{t} * \text { EmotionalValence }(S I)_{t} \tag{7}
\end{equation*}
\]

The salience of a social identity will be highest if both accessibility and fit are high. The higher a social identity, more impact that will have on the agent's behaviour.

\section*{Implementing DIMA}

For the purposes of experimentation and analysis, DIMA was implemented. The agent behaviour generation system consists of three components: The Characteristic Archetype, The Clustering Algorithm, The Social Identity Calculation.

The Characteristic Archetype consists of an abstract class, which allows the representation of multiple types of characteristics within the system.

In order to calculate the comparative fit and accessibility we used as a clustering algorithm the K-Means algorithm with a few modifications. The clustering algorithm takes into consideration all of the players characteristics values. First it will kick start itself with one K cluster, if there is at least one point who's distance is farther from the distance constraint X , the algorithm will increment K adding one more cluster, forcing the optimization process to restart. The algorithm finishes when the distance constraint heuristic is satisfied.

Figure 1 represents the program pipeline, which starts by assembling a list of the other players known by the agent and their characteristics. Using this list the agent will create a K-Means cluster containing a list with centroids and points. The number of centroids will be a direct representation of the


Figure 1: The Salience Calculation Pipeline.
number of clusters in the agent's K-Means algorithm. The comparative fit and accessibility are then calculated using this K-Means as an input parameter. Finally the salience is obtained through both the comparative fit and the accessibility value.

\section*{Platform}

To explore the above, we used a multi-player game within the Project INVITE \({ }^{1}\) (social Identity and partNership in VIrTual Environments) (Prada et al., 2012) where both humans and virtual agents can participate. The game begins with players stranded on an island due to a plane crash, where an active volcano threatens their lives at any moment. Each player's personal objective is to obtain the largest amount of gold, while at the same time help their campsite members collect wood to build a raft (the team objective) so as to get off the island. The players are faced with the dilemma of either helping their team by collecting wood or gathering gold and thus become rich when saved. If everyone collects mainly gold then the raft will not be built in time and everyone will loose when the volcano erupts. The player who can get off the island with the most gold is the winner.

Although this project aims at exploring the role of social identity and social dilemmas in mixed motive tasks, this platform is fully parametrized and allows the exploration of different scenarios and case studies. Some of the parametrizable variables are: the number of turns until the volcano erupts (end-game condition); number of campsites or teams; number of players per team (that could be a mix of humans and agents); visual characteristics for each player; total wood necessary to finish the raft; number of resources (wood and gold) each player can collect; among others.

\section*{Example Scenarios}

For experimentation purposes a simulation of the game was created. In our scenario the game was limited to 1 turn, 2 teams and 4 players for each of the both campsites, A and B. Players were controlled by virtually intelligent agents. Since different uniform colour has been known to prime differences

\footnotetext{
\({ }^{1}\) http://project-invite.eu/
}
in social group perceptions (Frank, Gilovich, et al., 1988; Peña, Hancock, \& Merola, 2009), it was used a form of differentiation of the two teams, in form of characteristics. In that order players were characterized by a shirt characteristic with the values 0 (red) or 100 (blue), and a campsite characteristic (A and B). Because this work's intentions is the study of the effects of an out-group, each player's K-Means algorithm was limited to a maximum of 2 clusters (i.e. the player's in-group and out-group). Due to the theme of the problem, the campsite will be the most influential characteristic of the clustering algorithm.

The total wood collected by agents is obtained by multiplying the salience value by the total carrying weight, which is 10 . The gold is the difference of the obtained wood value by the total weight. It is expected that agents with a higher salience identity with the campsite (common coloured shirt), will cooperate with more wood.

\section*{Red versus Blue}

In this scenario all campsite A members wear blue shirts while all campsite B members wear red shirts. Because both campsites have members with identical characteristics (i.e. no dispersion) the salience value will be 1 , it's maximum value (see table 1).

Table 1: Red versus Blue Scenario - Campsite A
\begin{tabular}{lccccc}
\hline \hline & Agent 1 & Agent 2 & Agent 3 & Agent4 & Total \\
\hline Colour & Blue & Blue & Blue & Blue & \\
\hline Accessib. & 1 & 1 & 1 & 1 & \\
Comp. Fit & 1 & 1 & 1 & 1 & \\
Salience & 1 & 1 & 1 & 1 & \\
\hline Wood & 10 & 10 & 10 & 10 & 50 \\
Gold & 0 & 0 & 0 & 0 & 0 \\
\hline
\end{tabular}

\section*{Unbalanced Teams}

In this scenario, campsite A has one member wearing a red shirt while the others wear blue shirts, as opposed to campsite B , where one member wears a blue shirt and the others wear red shirts. Although both campsites are similar in their shirt colour distribution, in the perspective of campsite A, the presence of the out-group (campsite B) will be weaker for the red shirt member than for the rest of its members (and vice versa for campsite B). Still, because three of the other members are similar, their salience identity values are strong enough to bias their behaviour to help their team (see table 2).

Table 2: Unbalanced Teams Scenario - Campsite A
\begin{tabular}{lccccc}
\hline \hline & Agent 1 & Agent 2 & Agent 3 & Agent4 & Total \\
\hline Colour & Blue & Blue & Blue & Red & \\
\hline Accessib. & 0.82 & 0.82 & 0.82 & 0.47 & \\
Comp. Fit & 0.68 & 0.68 & 0.68 & 0.68 & \\
Salience & 0.56 & 0.56 & 0.56 & 0.32 & \\
\hline Wood & 6 & 6 & 6 & 3 & 21 \\
Gold & 4 & 4 & 4 & 7 & 19 \\
\hline
\end{tabular}

\section*{Equal Mixed Teams}

In this scenario all campsite A and B members are equally divided between red and blue shirts (i.e. two blue and two red). From the perspective of one campsite, the presence of the out-group will be particularly weak, resulting in a low social identity salience. As such all members are going to behave a little more greedily than in the previous scenarios (see table 3).

Table 3: Equal Mixed Teams Scenario - Campsite A
\begin{tabular}{lccccc}
\hline \hline & Agent 1 & Agent 2 & Agent 3 & Agent4 & Total \\
\hline Colour & Blue & Blue & Red & Red & \\
\hline Accessib. & 0.65 & 0.65 & 0.65 & 0.65 & \\
Comp. Fit & 0.57 & 0.57 & 0.57 & 0.57 & \\
Salience & 0.37 & 0.37 & 0.37 & 0.37 & \\
\hline Wood & 4 & 4 & 4 & 4 & 16 \\
Gold & 6 & 6 & 6 & 6 & 24 \\
\hline
\end{tabular}

\section*{One Team Only}

In this scenario there is only one campsite as such all four agents are in the presence of in-group members. In this situation all agents share the same coloured shirt. Because there is no presence of an out-group, the social identity salience value is 0 , and all members behave accordingly to their personal identity (see table 4).

Table 4: One Team Only Scenario - Campsite A
\begin{tabular}{lccccr}
\hline \hline & Agent 1 & Agent 2 & Agent 3 & Agent4 & Total \\
\hline Colour & Blue & Blue & Blue & Blue & \\
\hline \hline Wood & 0 & 0 & 0 & 0 & 0 \\
Gold & 10 & 10 & 10 & 10 & 40 \\
\hline
\end{tabular}

\section*{Conclusion and Future Work}

Because social identity has a great impact in a wide range of fields and settings such as group formation, cohesiveness, prejudice, conformity, social influence and crowd behaviour (Turner et al. 1994; Hogg, 2003), we believe the study of this other phenomena could also benefit from DIMA.

Running the simulation we found that, as expected, agents whose t-shirt colour matches the majority of their campsite, expressed higher salience identity, cooperating with more wood, while the opposite situation had reverse results. However, in the extreme situations such as Red versus Blue Scenario, or One Team Only Scenario, it was quite evident that agents did not act rationally and presented extreme behaviour (collecting all wood or all gold). In these situations, it looked like they did not care about winning or surviving, respectively, as it would happen in a real situation with humans. As such, for future work, we intent to introduce rational thinking on agent's decision-making in which the influence of social identity salience will work upon. We are also, currently extending DIMA to calculate social identity salience in situations where three or more groups are present, as well as introducing the dynamics of the salience of multiple identities and relations among themselves.

\section*{Acknowledgements}

This work was partially supported by the INVITE project (ref. UTA-Est/MAI/0008/2009) funded by FCT under the UT-Austin/Portugal cooperation agreement and by national funds through FCT Fundação para a Ciência e a Tecnologia, under project PEst-OE/EEI/LA0021/2011, the PIDDAC Program funds.

\section*{References}

Brewer, M. (1979). In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. Psychological Bulletin, 86(2), 307.
Brewer, M. (1991). The social self: On being the same and different at the same time. Personality and Social Psychology Bulletin, 17(5), 475.
Dawes, R. (1974). Formal models of dilemmas in social decision-making. (Tech. Rep.). DTIC Document.
De Rosis, F., Pelachaud, C., \& Poggi, I. (2004). Transcultural believability in embodied agents: a matter of consistent adaptation. Agent Culture: Human-Agent Interaction in a Multicultural World, 75-106.
Dimas, J., \& Prada, R. (2013). You are who you hang out with: Agents with dynamic identity. In Proceedings of the 11th international conference on autonomous agents and multiagent systems (aamas 2013).
Frank, M. G., Gilovich, T., et al. (1988). The dark side of self-and social perception: Black uniforms and aggression in professional sports. Journal of personality and social psychology, 54(1), 74-85.
Fridman, N., \& Kaminka, G. (2009). Comparing human and synthetic group behaviors: A model based on social psychology. In International conference on cognitive modeling (iccm-09).
Goette, L., Huffman, D., \& Meier, S. (2006). The impact of group membership on cooperation and norm enforcement: Evidence using random assignment to real social groups. The American economic review, 96(2), 212-216.
Grier, R., Skarin, B., Lubyansky, A., \& Wolpert, L. (2008). Scipr: A computational model to simulate cultural identities for predicting reactions to events. In Proceedings of the second international conference on computational cultural dynamics (pp. 32-37).
Hofstede, G. H., \& Hofstede, G. (2001). Culture consequences: Comparing values, behaviors, intitutions, and organizations across nations. SAGE Publications, Incorporated.
Hogg, M., \& Williams, K. (2000). From i to we: Social identity and the collective self. Group Dynamics: Theory, Research, and Practice, 4(1), 81.
Kollock, P. (1998). Social dilemmas: The anatomy of cooperation. Annual review of sociology, 183-214.
Lustick, I. (2002). Ps-i: A user-friendly agent-based modeling platform for testing theories of political identity and political stability. Journal of Artificial Societies and Social Simulation, 5(3).

Mascarenhas, S., Dias, J., Afonso, N., Enz, S., \& Paiva, A. (2009). Using rituals to express cultural differences in synthetic characters. In Proceedings of the 8th international conference on autonomous agents and multiagent systemsvolume 1 (pp. 305-312).
McLeish, K., \& Oxoby, R. (2007). Identity, cooperation, and punishment. IZA.
Oakes, P. (1987). The salience of social categories. Rediscovering the social group: A self-categorization theory, 117-141.
Peña, J., Hancock, J. T., \& Merola, N. A. (2009). The priming effects of avatars in virtual settings. Communication Research, 36(6), 838-856.
Prada, R., Raimundo, G., Baptista, M., Dimas, J., Santos, P. A., Martinho, C., et al. (2012). The role of social identity, rationality and anticipation in believable agents. In Proceedings of the 11th international conference on autonomous agents and multiagent systems (aamas 2012).
Salzarulo, L. (2006). A continuous opinion dynamics model based on the principle of meta-contrast. Journal of Artificial Societies and Social Simulation, 9(1), 13.
Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. Advances in experimental social psychology, 25(1), 1-65.
Smith, E., \& Mackie, D. (2000). Social psychology. Psychology Pr.
Tajfel, H. (1972). La catégorisation sociale (social categorization). Introduction à la psychologie sociale, 1, 272302.

Tan, C., \& Cheng, H. (2007). Personality-based adaptation for teamwork in game agents. In Proceedings of the third conference on artificial intelligence and interactive digital entertainment (Vol. 37).
Turner, J. (1985). Social categorization and the self-concept: A social cognitive theory of group behavior. Advances in group processes: Theory and research, 2, 77-122.
Turner, J., Hogg, M., Oakes, P., Reicher, S., \& Wetherell, M. (1987). Rediscovering the social group: A social categorization theory. B. Blackwell, Oxford, UK.
Turner, J., Oakes, P., Haslam, S., \& McGarty, C. (1994). Self and collective: Cognition and social context. Personality and social psychology bulletin, 20, 454-454.
Weber, J., Kopelman, S., \& Messick, D. (2004). A conceptual review of decision making in social dilemmas: Applying a logic of appropriateness. Personality and Social Psychology Review, 8(3), 281.
Wit, A., \& Wilke, H. (1992). The effect of social categorization on cooperation in three types of social dilemmas. Journal of Economic Psychology, 13(1), 135-151.

\title{
Constraining ACT-R Models of Decision Strategies: An Experimental Paradigm
}

\author{
Cvetomir M. Dimov (cvetomir.dimov@unil.ch) and Julian N. Marewski (julian.marewski@ unil.ch) \\ Department of Organizational Behavior, Université de Lausanne, Quartier UNIL-Dorigny, Bâtiment Internef, 1015 Lausanne, Switzerland
}

Lael J. Schooler (schooler@mpib-berlin.mpg.de)
Max Planck Institute for Human Development, Lentzeallee 94
14195 Berlin, Germany

\begin{abstract}
It has been repeatedly debated which strategies people rely on in inference. These debates have been difficult to resolve, partially because hypotheses about the decision processes assumed by these strategies have typically been formulated qualitatively, making it hard to test precise quantitative predictions about response times and other behavioral data. One way to increase the precision of strategies is to implement them in cognitive architectures such as ACT-R. Often, however, a given strategy can be implemented in several ways, with each implementation yielding different behavioral predictions. We present and report a study with an experimental paradigm that can help to identify the correct implementations of classic compensatory and noncompensatory strategies such as the take-the-best and tallying heuristics, and the weighted-linear model.
\end{abstract}

Keywords: Take-the-best, tallying, weighted-linear model, process models, ACT-R

\section*{Introduction}

One important characteristic of well-developed scientific theories is precision. In psychology, theoretical precision can be achieved by complementing verbally formulated theories with formal models. Typically, formal models are specified in terms of mathematical equations or computer code. The goals, level of detail, and level of description of such models vary as a function of the psychological subdiscipline, research questions being asked, or the available technology, to name only a few factors. Computational models have become both increasingly popular and powerful, and have aided cognitive scientists in their endeavor to shed light into the behaviorist's black box. Computer models allow one to specify, on an algorithmic level, the cognitive processes psychological mechanisms are assumed to draw on.

Such process models predict not only what decision a person will make, but also how the information used to make the decision will be processed. The predictions made by these models can thus be tested not only on outcome data (e.g., what item is chosen) but also on process data, including on patterns of information search, response times, or neural activation. Such predictions can eventually differentiate among competing theories that make identical outcome predictions. In particular in the cognitive and decision sciences, describing cognitive processes represents a central goal of theorizing on its own. In fact, the past decades have seen repeated calls to develop process models.

Yet, surprisingly there are relatively few theories of decision making that yield detailed quantitative predictions about process data. Instead, typically qualitative predictions about response times and other process data are tested in experiments. This theoretical and methodological weakness contributes to fuelling important scholarly debates about which decisional processes describe behavior best: simple non-compensatory ones, for which decisions based on some predictors cannot be overturned by others, or complex compensatory integration processes, for which various predictors can neutralize each-other's influence (cf. Bröder \& Schiffer, 2003; Glöckner \& Betsch, 2008; Marewski et al., 2010).

One way to increase the precision of theories of decision making is to implement them in detailed cognitive architectures such as the ACT-R theory of cognition (e.g., Anderson, 2007). ACT-R is a quantitative framework that applies to a broad array of behaviors and tasks, formally integrating theories of memory, perception, action, and other aspects of cognition. ACT-R also allows modeling decision processes. When models of decision making are implemented in ACT-R, quantitative predictions about response time distributions at the millisecond level and other process data can be made and compared to experimental studies. Marewski and Mehlhorn (2011), for instance, implemented several compensatory and noncompensatory decision strategies in ACT-R. In doing so, they modeled for each of the strategies how decisional processes interplay with memory, perceptual, and motor processes, which, in turn, allowed them to quantitatively predict the response time distributions associated with using each strategy in a simple two-alternative forced choice decision task.

While the architectural approach can thus help remedying the aforementioned theoretical and methodological weakness, this approach does not come without its complications. Specifically, often a given strategy can be implemented in numerous different ways in ACT-R (or other cognitive architectures), with each implementation yielding different response time and other process predictions. Part of the problem is that many decision strategies are-in the worst case-only formulated verbally or-in the best case-specified mathematically or algorithmically, without spelling out the strategies' assumptions about lower-level cognitive processes. This specification problem (see Lewandowsky, 1993), namely
how to translate an underspecified theory or strategy into a detailed cognitive model, poses a paramount modeling challenge to the researcher who sets out to find out which implementation is the most adequate one. To illustrate this point, Marewski and Mehlhorn (2011) actually ended up implementing over thirty ACT-R models of similar decision strategies without being able to make strong conclusions about which model most likely represented the correct one.

In this paper, we present and report a study with an experimental paradigm that can help to build and identify the correct implementations of decision strategies. In what we call the train-to-constrain-paradigm, participants are instructed in a detailed step-by-step procedure how to apply specific strategies in a decision task. Since the experimenter thus knows which strategies participants have relied on in the experiment, the resulting response times lend themselves to constraining ACT-R implementations of these strategies. Specifically, as an initial step, here we focus on a variant of that paradigm in which participants are instructed to apply three classic compensatory and non-compensatory strategies, namely the take-the-best (henceforth: TTB) and tallying heuristics, and the weighted-linear model (henceforth: WLM).

The remainder of this paper is structured as follows. First, we will explain in more detail the three decision strategies. Second, we will present the train-to-constrain-paradigm and, in doing so, report a study that we ran using that paradigm. Third, we will report the results of this study, and, fourth, briefly illustrate how these results can be used to build and constrain ACT-R implementations of the three strategies.

\section*{Decision Strategies}

Tallying and WLM have been formulated in different ways (and at times also been given different names); here we use Gigerenzer and Goldstein's (1996) definitions as well as their TTB heuristic. Gigerenzer and Goldstein specified these strategies as models of inductive inference about unknown quantities or future events in simple twoalternative forced choice tasks. In such tasks, a person has to infer which of two alternatives (e.g., cities) has a larger value on a given criterion (e.g., population). One variant of this task that has received considerable attention during the past years is the memory-based decision task illustrated in Figure 1. In this task, a person has to make inferences by relying exclusively on the contents of their memory. The experimental paradigm for identifying correct ACT-R implementations of TTB, tallying, and the WLM that we propose here extends this memory-based task.

Take-the-best. The simple TTB heuristic stands in the tradition of Tversky's (1972) classic elimination by aspects model. TTB bases inferences on the attributes of the alternatives (e.g., whether a city has an airport), which it uses as cues. A cue can have a positive (e.g., a city has an airport, coded as " 1 "), negative (has no airport, coded as "1 "), or an unknown (coded as " 0 ") value. The vector of cue values that define a person's knowledge about a specific
alternative is called the alternative's cue profile. TTB bases inferences on just one good cue. Specifically, TTB orders the cues \(i\) unconditionally according to their validity \(v_{i}\), with \(v_{i}=c_{i} /\left(c_{i}+w_{i}\right), c_{i}\) being the number of correct inferences based on cue \(i\) given that it discriminates between two alternatives (i.e., cue values are \(1 \& 0\), respectively, or \(1 \&-\) 1 , respectively), and \(w_{i}\) the number of incorrect inferences. TTB's rules for searching cues, stopping search, and making a decision can be summarized as follows:

Search: Search through cues in the order of their validity. Stopping: Stop as soon as a cue is found that discriminates between the alternatives.
Decision: Infer that the alternative with the positive cue value has the higher value on the criterion of interest.
As can be seen, TTB is a non-compensatory strategy, which uses solely the first discriminating cue. Translated into a process prediction this implies, for example, that the time it takes to make decisions with TTB should depend on how many cues have been considered before a discriminating cue is found.


Figure 1: Illustration of the memory-based decision task
Tallying. In contrast to TTB and other non-compensatory strategies, many decision models posit that people evaluate alternatives by integrating knowledge about multiple cues. One such heuristic is tallying. This representative of classic unit-weight linear integration models (e.g., Dawes, 1979) simplifies decisions by treating all cues equally. For each alternative, tallying simply counts the cues with positive values and infers that alternatives with the larger number of positive cue values score higher on the criterion of interest. As a consequence, the various cues can neutralize each other's influence on the final decision, thus making tallying a compensatory model. Tallying's search, stopping, and decision rules read as follows:

Search: Search through cues in any order.
Stopping: Stop search after \(m\) out of a total of \(M\) cues (with \(1<m<M\) ) have been accessed.
Decision: Decide for the alternative that is favored by more positive cue values. If the number of positive cue values is the same for both alternatives, guess.

Weighted-linear model. The WLM is similar to tallying in that it integrates all the information available when choosing an alternative. In the WLM, cue values are coded like in TTB. As suggested by its name however, it integrates all cue information by multiplying the cue values by their validities and summing them over for each city, thus
computing the weighted sum of the cues for each city. The WLM's rules can be summarized as follows:

Search: Search through cues in any order.
Stopping: Stop search after \(m\) out of a total of \(M\) cues (with \(1<m<M\) ) have been accessed. Multiply each cue value with its validity and compute the weighted sum of cues for each alternative.
Decision: Decide for the alternative that is favored by the larger weighted sum. If the weighted sum is the same for both alternatives, guess.
The WLM has a long tradition in the cognitive and decision sciences and beyond. For instance, variants of this model have been viewed as optimal rules for preferential choice and are often considered to define rational behavior (cf. Payne, Bettman, \& Johnson, 1993).

\section*{Experimental Paradigm}

The train-to-constrain-paradigm builds on several earlier studies on TTB, tallying, and the WLM (e.g., Bröder \& Gaissmaier, 2007; Bröder \& Schiffer, 2003; Mata, Schooler \& Rieskamp, 2007) and on approaches that teach subjects to rely on specific decision strategies (e.g., Khader et al., 2011; Marewski \& Schooler, 2011).

In our study, we implemented the training portion of our paradigm in a computerized experiment, in which subjects were told that they would participate in a quiz show. In that show, they first learned fictitious facts about how British cities would look like in the future, namely whether these cities would have an international airport, a train station, a university, and/or a premier league soccer team in the year 2100 (such facts are typically judged as useful for inferring city size; cf. Pachur, Bröder, \& Marewski, 2008). In a second step, subjects learned how to employ a strategy that uses these facts as cues to make decisions. During the actual quiz show, they then saw pairs of cities on the computer screen and were instructed to always use the strategy to infer which of the two cities would be larger in the year 2100. Subjects were paid according to the degree to which their decisions agreed with predictions of the respective decision strategy.

Subjects and design. A total of 141 subjects participated in the experiment ( 89 male, \(\mathrm{M}_{\text {age }}=25.3\) ), of which 120 finished it successfully. Subjects were randomly assigned to one of three between-subjects conditions. The conditions differed in terms of the strategy participants learned to use. In the first condition subjects learned TTB, in the other two conditions they learned tallying and the WLM, respectively.

Materials. Sixteen well-known British cities were used as alternatives. These cities correspond to those that most subjects in Pachur et al.'s (2008) pre-study 1 recognized. A pre-study suggested that subjects' familiarity with these cities' names aids them to learn a large number of facts about these cities. Since the degree of familiarity was roughly the same for all cities in both Pachur et al.'s prestudies, no interference effects of familiarity were expected,
and, indeed, also none found. These 16 cities were combined with 8 cue profiles, illustrated in Table 1. In doing so, each of the 8 cue profiles was used twice-albeit with different city names.

Table 1: Cue profiles used
\begin{tabular}{|c|c|c|c|c|c|c|c|c|}
\hline & City1 & City2 & City3 & City4 & City5 & City6 & City7 & City8 \\
\hline Airport & + & + & - & - & + & + & - & - \\
\hline Soccer team & - & - & - & - & - & - & + & + \\
\hline University & - & - & + & + & + & + & + & + \\
\hline Train station & + & - & + & - & + & - & + & - \\
\hline
\end{tabular}

Learning task. The experiment started with a learning task (cf. Bröder \& Schiffer, 2003), in which subjects were taught the 4 cues about the 16 British cities, corresponding to a total of \(4 \times 16=64\) facts. Specifically, during learning, cities and cues were presented repeatedly in a random order until subjects correctly recalled at least 14 of the 16 cities' cue profiles perfectly. Cue profiles were assigned at random to specific cities.

Strategy learning task. After having learned all cues, in each of the three between-subjects conditions, subjects were trained how to use one of three decision strategies. The strategy learning procedure required subjects to go through a stepwise explanation of the decision process assumed by each strategy as well as to apply that strategy correctly on several practice trials that mimic the actual decision task. During practice, subjects received feedback about whether they had applied the strategy correctly, and the strategy was practiced until subjects' decisions concurred to \(100 \%\) with the strategy's predictions. During the strategy learning task, subjects also memorized additional information that is necessary for applying the strategy, such as the cue validities in the case of TTB and WLM. The instructions on how to use each strategy were crafted such that they reflect the strategy descriptions from the literature.

Repetition of learning task. To make sure participants still remembered the 64 facts correctly, one round of the learning task was repeated upon completion of the strategy learning task.

Decision task. In a decision task, 72 pairs of the previously learned British cities were presented (one city on the left side of the computer screen, the other one on the right; see Figure 1). To avoid inducing frequency effects, the pairs were constructed such that each city name appears equally often. Subjects were instructed to always apply the strategy to decide which of the cities will be larger in the year 2100. For each correct application of the strategy, subjects received a bonus payment of 0.5 Euros (0.68 US\$). Each decision inconsistent with the strategy's prediction resulted in a penalty of 0.5 Euros (no feedback was given).

Cue-memory task. In a cue-memory task, subjects had to reproduce the cue values they learned for the cities. The purpose of this task was to collect data about how well subjects remembered the cue values they were taught. This data will be used in future projects to populate the declarative memory of the ACT-R models.

\section*{Experimental Results}

Figure 2 shows the mean of the \(25^{\text {th }}, 50^{\text {th }}\) and \(75^{\text {th }}\) response time percentiles for the three experimental conditions as a function of the number of cues that have to be retrieved from memory prior to finding the most valid discriminating one (henceforth: most valid discriminating cue). Several important observations can be made. First, tallying participants made the fastest decisions. Their response time varied from under 3 s for the \(25^{\text {th }}\) percentile to almost 6 s for the \(75^{\text {th }}\) percentile. This is much faster than previous decision making experiments have reported. For example, Bröder and Gaissmaier (2007) reported mean response times between 6.5 s and 8 s in their first, and between 11 s and 15 s in their second experiment. It should be noted that those experiments did not instruct subjects to rely on specific strategies, but that instead used participants’ decisions to infer, post hoc, by means of strategy classification procedures which strategies subjects have used.


Figure 2: Participants’ aggregate response time percentiles as a function of most valid discriminating cue. Error bars are standard errors of the mean computed across all participants in the respective experimental condition.

Second, the response times of TTB participants fall in the response time range of those reported in these previous experiments. However, this resulted in participants in the TTB condition being slower than tallying participants, which also is a finding that stands in contrast to previous studies, in which post hoc strategy classification procedures were used (e.g., Bröder \& Gaismaier, 2007).

Third, WLM participants are the slowest, which is a result that is consistent with Bröder and Gaissmaier's (2007) earlier studies. Bröder and Gaissmaier reported mean response times between 10s and 11s in their first and between 15 s and 23 s in their second experiment, which fall close to the time range of our participants.

Fourth, as can be seen in Figure 2a, TTB participants' response times increase as a function of most valid discriminating cue. In contrast, Figures 2b and 2c show that for tallying and the WLM the response times do not exhibit such an increase when they are analyzed in the same way as for TTB participants. This result is to a large extent consistent with earlier work: in Bröder and Gaissmaier's (2007) experiments, participants who were inferred to have relied on TTB exhibited strong increases in mean response times as a function of the most valid discriminating cue, while those who were classified as likely users of tallying or the WLM did not exhibit increases that were as strong.

\section*{Implementing Strategies in ACT-R}

In the constraining portion of our paradigm, the observed response times will be used to build and constrain ACT-R implementations of the three decision strategies. Specifically, each individual participant's responses in the memory task can be used to model the contents of that subject's declarative memory after having gone through the training phase. These declarative memory contents can then be used to model the retrieval processes associated with using each of the three decision strategies (cf. Marewski \& Mehlhorn 2011, for this approach). Together with perceptual, motor, and other cognitive processes-all of which can be modeled in ACT-R-these retrieval processes will contribute to the response times predicted by the corresponding ACT-R models of the decision task.

\section*{Overview of ACT-R}

ACT-R describes cognition as a set of modules that interact through a production system. The production system consists of production rules (i.e., if-then rules) whose conditions (i.e., the "if" parts) are matched against the contents of the modules. If a rule's conditions are met, then the production rule can fire and the specified action is carried out. Each module implements different cognitive processes. The declarative module, for instance, enables information storage in and retrieval from memory, the intentional module keeps track of a person's goals, while the imaginal module holds information necessary to perform the current task. A visual module for visual perception and a manual module for motor actions (e.g., typing on a keyboard) simulate interactions with the world. In coordinating the modules, the production rules can only act on information that is available in buffers, which can be thought of as processing bottlenecks, linking the modules' contents to the production rules. For instance, the production rules cannot access all contents of the declarative module, but only these that are currently available in the retrieval buffer. ACT-R distinguishes between a symbolic


Figure 3: Processing stream of the weighted-linear model for the first and last seconds of the decision process. Production rules on the right hand side are stylized representations of the actual ACT-R productions for this model. Note that the model's decision time predictions can vary across different decision trials, for instance, as a function of perceptual and motor processes, or cue activation. Also note that the same production rules fire more than once during the process.
and a subsymbolic system. The symbolic system is composed of the productions rules as well as of the modules and buffers. Access to the information stored in the modules and buffers is determined by the subsymbolic system. This system is cast as a set of equations and determines, for instance, the timing of memory retrieval.


Figure 4: ACT-R predictions of response time percentiles of a tallying and weighted-linear model implementation. Error bars are standard errors of the mean, computed across 30 simulation runs of the ACT-R model.

\section*{Illustrating our ACT-R models}

Figures 4a, 4b and 4c present our preliminary ACT-R models, developed prior to running the experiment as a source of rough predictions of participants' eventual behavior. All of these three models are, perhaps, the most naïve implementations which follow the above mentioned strategy definitions and experimental instructions. In developing these models, no parameters were fitted, but
those from Marewski and Mehlhorn (2011) were used.
All models perform the same task as our experimental subjects: The models "read" two city names off a computer screen, process them, decide for one of them, and enter a response by "pressing" a key. To illustrate this, Figure 3 shows the first and last seconds of an 18 -seconds-long processing stream of our preliminary ACT-R implementation of the WLM. The various decisional, memorial, perceptual and motor processes assumed by the model are coordinated by production rules.

Specifically, by first "reading" the names of both cities, the model tries to retrieve a memory trace of the city names called a chunk. Chunks are facts like "York is a city" or "York has an airport" which model people's familiarity with city names and their cue knowledge about these cities, respectively. For each cue, the model retrieves its validity. If the cue value is positive, the model adds the validity of this cue to the weighted sum of the city, initiating a summation procedure. If the cue value is negative, the model subtracts the validity of the corresponding cue from the weighted sum of that city, initiating a subtraction procedure. Finally, the model compares the total weighted sums of the two cities and chooses the one with the larger total weighted sum by pressing a key. As Figure 4c shows, the predicted response time percentiles of 30 simulation runs of this WLM ACT-R implementation lie close to the \(75^{\text {th }}\) percentile range observed in participants' data (Figure 2c), suggesting that this implementation is not an implausible model, but also that other processes which boost participants' response times, such as memorizing the weighted sum, are present in participants. Our preliminary tallying model (Figure 2b) predicts response times within experimental data, while the TTB model (Figure 2a) is faster. These three models have to be adapted to successfully capture participants' behavior, a more successful example of which is the tallying model presented on Figure 4d, which was built after the experiment. While the former tallying model did not include
memorization of the number of positive cue values of already seen cities, the latter model did, which produced a response time distribution close to participants' response times. Exact modeling of each participant's cue knowledge is the next modeling step to be made. Naturally, after identifying the most promising implementations of all strategies, all models would then have to be tested in new experiments, this way ensuring that they can also account for behavior in tasks for which they were not developed.

\section*{Discussion and Conclusion}

While it goes beyond the scope of this short proceedings paper to present more ACT-R implementations-that is part of a larger research paper-one legitimate question one may raise is what the methodological advantages of our approach over earlier experimental work is. As mentioned above, in earlier studies including Marewski and Mehlhorn's (2011) ACT-R modeling efforts and Bröder and Gaissmaier's (2007) response time analyses for TTB and other heuristics, participants' decisions had to be used to infer, post hoc, by means of strategy-classification and/or other model selection procedures which strategies participants relied upon in an experiment. As a result, the conclusions that could be drawn from analyses of response times crucially hinged on the accuracy of the strategy classification and/or model selection procedure. Our train-to-constrain approach, in contrast, allows identifying the response time patterns associated with a strategy without the need to use potentially inaccurate strategy classification. To illustrate this point, the deviations observed between Bröder and Gaissmaier's and our findings could, besides being a product of differences in the stimuli and materials used, also be a result from the strategy classification method used by these authors. More studies with our paradigm, including experiments that make use of Bröder and Gaissmaier's stimuli and materials, are warranted to decide between these and other competing explanations.

To conclude, response times such as the ones observed in our experimental paradigm can be used to find out which ACT-R implementation best mirrors classic decision strategies used by trained subjects. Once identified, these implementations can, hopefully, be used to model behavior both in previously published studies as well as in new studies in which subjects' decision strategies are unconstrained by training.

\section*{References}

Anderson, J. R. (2007). How can the human mind occur in the physical universe? New York: Oxford University Press.
Bergert, F. B., \& Nosofsky, R. M. (2007). A response-time approach to comparing generalized rational and take-thebest models of decision making. Journal of Experimental Psychology: Learning, Memory, and Cognition, 31, 107-129.
Bröder, A., \& Gaissmaier, W. (2007). Sequential processing of cues in memory-based multi-attribute decisions.

Psychonomic Bulletin \& Review, 14, 895-900.
Bröder, A., \& Schiffer, S. (2003). Take the best versus simultaneous feature matching: Probabilistic inferences from memory and effects of representation format. Journal of Experimental Psychology: General, 132, 277-293.
Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. American Psychologist, 34, 571-582.
Gigerenzer, G., \& Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. Psychological Review, 104, 650-669.
Glöckner A., \& Betsch T. (2008). Modeling option and strategy choices with connectionist networks: Towards an integrative model of automatic and deliberate decision making. Judgment and Decision Making, 3, 215-228.
Lewandowsky, S. (1993). The rewards and hazards of computer simulations. Psychological Science, 4, 236243.

Khader, P.H., Pachur, T., Meier, S., Bien, S., Jost, K., \& Rösler, F. (2011). Memory-based decision-making with heuristics: Evidence for a controlled activation of memory representations. Journal of Cognitive Neuroscience, 23, 3540-3554.
Marewski, J. N., Gaissmaier, W., Schooler, L. J., Goldstein, D., \& Gigerenzer, G. (2010). From recognition to decisions: Extending and testing recognition-based models for multialternative inference. Psychonomic Bulletin \& Review, 3, 287-309.
Marewski, J. N. \& Mehlhorn, K. (2011). Using the ACT-R architecture to specify 39 quantitative process models of decision making. Judgment and Decision Making, 6, 439-519.
Marewski, J. N., \& Schooler, L. J. (2011). Cognitive niches: An ecological model of strategy selection, Psychological Review, 118, 393-437.
Mata, R., Schooler, L. J., \& Rieskamp, J. (2007). The aging decision maker: Cognitive aging and the adaptive selection of decision strategies. Psychology and Aging, 22, 796-810.
Newell, B. R., \& Bröder, A. (2008). Cognitive processes, models and metaphors in decision research. Judgment and Decision Making, 3, 195-204.
Pachur, T., Bröder, A., \& Marewski, J. N. (2008). The recognition heuristic in memory-based inference: Is recognition a non-compensatory cue? Journal of Behavioral Decision Making, 21, 183-210.
Payne, J. W., Bettman, J. R., \& Johnson, E. J. (1993). The adaptive decision maker. New York: Cambridge University Press.
Schulte-Mecklenbeck, M., Kühberger, A. \& Ranyard, R. (Eds.). (2011). A handbook of process tracing methods for decision research: A critical review and user's guide. New York: Taylor \& Francis.
Tversky, A. (1972). Elimination by aspects: A theory of choice. Psychological Review, 79, 281-299.

\title{
Planning ahead through space and time: from neuropsychology to motor control
}

\author{
Mariama C.Dione (mariama.dione@univ-lille3.fr) \\ Laurent Ott (laurent.ott@univ-lille3.fr) \\ Yvonne Delevoye-Turrell (yvonne.delevoye@univ-lille3.fr)
}

\author{
URECA EA 1059, University Lille Nord de France, rue du Barreau, BP 60149 \\ Villeneuve d’Ascq, 59653, France
}

\begin{abstract}
The executive functions have been studied separately in the fields of neuropsychology and of motor control. However, it is not clear whether across fields one is referring to similar cognitive functions. In the present study, we compared the performance scores obtained in a motor spatial-tapping task with those scores obtained in a battery of three neuropsychological tasks which assess respectively the executive functions of updating ( N -back task), inhibiting (GonoGo task) and switching (Letter-number task). Multiple regression analyses revealed significant and specific effects between the motor task and the classical neuropsychological tasks: the timing error measured at slow tempi in the tapping task predicted the scores observed in the updating task only; the spatial error at faster tempi predicted the scores obtained in the switching task only; the contact times at intermediate tempi predicted the scores obtained in the inhibiting task only. Hence, we introduce this easy-to-use non-verbal task as a novel paradigm to assess executive functioning.
\end{abstract}

Keywords: planning; executive functions; motor control; neuropsychology; regression; space; time.

\section*{Introduction}

\section*{Neuropsychological approach of Planning}

Research on the executive functions has historical roots in the study of patients with frontal lobe damage. These patients were the first to demonstrate disruptions to control and organize daily activities independent of any language or memory disorders (Damasio, 1994; Harlow, 1868). The socalled 'frontal' or 'executive' tasks were specifically developed to assess how the main functional deficit encountered by frontal patients, i.e. the planning of daily activities, was affected following brain injury. For example, the Tower of London (Shallice, 1982), a task inspired by a logical reasoning game (the Tower of Hanoi) was proposed to evaluate the ability to inhibit a routine schema - that consisted in producing each move in an isolated and impulsive fashion in response to the true visual configuration of the Tower - in order to define and adopt a cognitive plan to achieve the puzzle in fewer moves (see the model of SAS/GOC, Shallice, 1988). The Wisconsin Card Sorting Test (WCST, Grant \& Berg, 1948), on the other hand, was proposed to target individuals' inability to switch efficiently from an ongoing plan to a novel one. In this task, perseverating errors were measured and used as indicators of cognitive flexibility (Milner, 1963). Finally, working memory tasks, e.g., the n-back task, were developed to
target a person's ability to maintain a cognitive plan active and to update between relevant parts of the plan in function of a given situation. Hence, the executive functions are here directly related to the abilities to inhibit a routine schema in order to adopt and adapt a novel more cognitive plan, to maintain and update parts of a given plan across time and/or to switch flexibly from one plan to another. In the last decade, these cognitive abilities have been referred to in the literature as the inhibiting, the updating, and the switching executive functions (Miyake \& Friedman, 2012; Miyake et al., 2000) and it is common today to use a multiple-test battery to assess the well functioning of the executive functions related to planning abilities.

However, it is the case that the neuropsychological tasks that are classically used to assess the executive functions suffer from a number of validity problems that limits severely the possibility to compare the functions between them. Indeed, the tasks are known to have impurity issues because of the presence of non-executive demands specifically related to the various contents that are used in the tasks (e.g., language, limb displacements, object identification, etc., Burgess, 1997; Phillips, 1997). Second, they present a lack of test-retest reliability, i.e. people can adopt different executive strategies to perform the same task across sessions (Rabbitt, 1997). For example, although the WCST has been designed to reveal a lack of cognitive flexibility, subjects may perform the task by inhibiting certain responses that are no longer appropriate (Miyake et al., 2000). Thus, scores in different sessions may reflect different cognitive strategies, with participants who are sometimes switching between rules and at other times using inhibition to solve the task. The difficulty to characterize the possible relationships between the executive functions is in itself a motivator to consider today a different approach to the evaluation of executive functioning.

\section*{Planning in the Motor Control domain}

A major interest in the field of motor control is to understand how actions are coordinated to enable the execution of complex sequential motor activities, e.g. playing a musical instrument or dancing in rhythm. This question has been particularly studied in the context of sensorimotor synchronization (for a review on SMS, see Repp, 2005). Even if synchronized behaviors require motor coordination both through space and time, research has focused historically on the timing aspects of motor planning and experimental findings have led to the acceptance today
of the existence of two distinct timing modes related to the control and execution of rhythmical action sequences (Robertson et al., 1999; Zelaznik, Spencer, \& Ivry, 2002). The event-based timing mode is primarily involved in tasks that have a clear temporal goal, e.g. trying to keep the beat of a metronome, and is assumed to require an explicit internal representation or memory of the referential temporal interval to produce (Wing \& Kristofferson, 1973). By contrast, emergent timing is assumed to arise implicitly, i.e. from the extraction of temporal regularities emerging naturally from the dynamics of movement control when actions are repeated in smooth oscillatory cycles (Ivry, Spencer, Zelaznik, \& Diedrichsen, 2002; Turvey, 1977). For example I can infer that "I will be late given the speed of my successive footsteps". While finger tapping is the prevailing paradigm to reveal event-based timing, circle drawing has been proposed as the exemplar task to reveal emergent timing (Zelaznik et al., 2002). In terms of statistics, negative lag-1 autocorrelation values are typically observed in fingertapping tasks suggesting that motor responses are controlled through an internal timekeeper (Vorberg \& Wing, 1996; Wing \& Kristofferson, 1973). In contrast, positive or near-to-zero AC-1 values are classically reported in circle drawing tasks suggesting that other mechanisms of control enter into play and enable the implicit emergence of a certain sense of rhythmicity (Lemoine \& Delignières, 2009; Torre \& Delignières, 2008; Zelaznik et al., 2002). Anchoring has been described as a possible mechanism that could explain how timing emerges when movements are continuous rather than discrete. This phenomenon, commonly observed in cyclical movements, consists in a local reduction in spatial and/or temporal errors at a specific location along the trajectory path and is often observed around reversal points, i.e. at points of transition between flexion/extension movements (Beek, Turvey, \& Schmidt, 1992; Roerdink, Ophoff, Peper, \& Beek, 2008). This specific point of transition could be used as a referential to infer timing regularities in the case of continuous movements (Repp \& Steinman, 2010).

In a recent study (Dione \& Delevoye-Turrell, unpublished) the use of a unique task designed as a hybrid of finger tapping and circle drawing was suggested to reveal, assess and compare the two modes of timing between them. In this 'spatial-tapping' task, a picture composed of six discrete visual targets (disposed around a virtual circle) was displayed on a tactile screen. Participants were asked to produce discrete taps on each target, one after the other and to follow the circular trajectory with the arm at the regular pace of a metronome (from 1100 to 300 ms of inter-tone intervals, ITI). The motor actions were assumed to be discrete at slower tempi with the need of cognitive control to maintain long timing intervals through timekeeping at slow tempi; motor actions were proposed to be continuous at faster tempi through the capacity to anticipate only the point of transition between flexion/extension movements when the tempi were too fast. Autocorrelation values were measured up to ten lags to
reveal event-based timing at slow tempi (negative AC-1) and emergent timing at faster tempi (positive AC-6). Furthermore, a detailed spatial analysis was conducted on the spatial endpoint distributions to assess whether the cognitive strategy was turned towards the need to anticipate the point of transition between flexion and extension movement at faster tempi. Spatial ellipses were first measured for each target. The mean area and the angular orientation of each ellipse were then computed. An orientation error was finally calculated as the angular difference between the orientation of the ellipse and the tangent to the circle measured at each target. Performance results revealed first that both the timing ( \(\%\) of \(\mathrm{IRI}_{\text {error }}\) ) and the spatial accuracy (mean spatial area) were perfectly maintained at slowest tempi (from 1100 to 900 ms ). A first small but significant decrease in the performance arose at 800 ms and was maintained until 600 ms . For tempi that were faster than 500 ms the performance was the worst with rather large spatial and timing errors. As predicted for the timing strategies, significant negative autocorrelation values emerged at lag-1 at slow tempi only (from 1100 to 700 ms ) and significant positive AC emerged at lag-6 at faster tempi only (from 500 to 300 ms ). In order to assess whether the actions were controlled or not through an internal timekeeper (event-based timing), the motor delays were measured for each tempo and both lags according to the WK model. Results revealed that the motor delays increased in function of tempi at lag-1 only, suggesting that the timing was event-based at slow tempi only in this spatial-tapping task and that other mechanisms entered into play to explain the correlation factors observed at lag-6. Finally, the spatial analysis confirmed that the endpoint distributions were more oriented in relation to the tangent to the circle at faster tempi, with the emergence of an anchor point at the point of maximal extension in the fastest tempi only.

Overall, these results suggested that in the spatial-tapping task, the timing mode was changed from event-based timing at slow tempi towards emergent timing at faster tempi. Consequently, it is possible to presume that in this motor task the cognitive demands depended on different behavioral strategies in function of the cognitive needs to actively maintain a referential timing interval in working memory (from 1100 to 900 ms ), and to anticipate the spatial point of transition between the two movements of flexion vs. extension involved in the sequence (from 500 to 300 ms ). An intermediate phase was here observed for those tempi between 800 to 600 ms in which the performance was decreased in space and time but in which at the same time the event-based timing strategy was efficiently maintained for \(2 / 3\) of the trials. It is thus possible that in this phase motor inhibition was used in order to avoid a too fast transition from discrete to continuous movement by making an effort to maintain attention on each discrete action in spite of the increase in temporal pressure.

\section*{From neuropsychology to motor control}

In the present study, we suggest that the spatial-tapping task, a task in which several motor actions need to be planned and executed both in space and time could be an interesting method to assess the so-called executive functions. Indeed, the spatial-tapping task requires: (1) working-memory at slow tempi in order to produce discrete tapping actions in reference to memorized timing intervals, (2) the active reduction of a switching cost related to the motor switching between the two biomechanically distinct movements of flexion and extension that compose the movement sequence with the idea that bad switching will lead to poor spatial control of movement trajectory, (3) and finally inhibition of a too fast transition from discrete to continuous actions at intermediate tempi, with finger contact times that are too short to maintain high timing levels of performance.

In order to test this hypothesis performance scores were measured both in a spatial-tapping task and in a battery of neuropsychological tests assessing the three main executive functions. Multiple regression analyses were then computed to explore and reveal specific relationships between the functions.

\section*{Methods}

\section*{Participants}

Twenty-six right-handed students between 18 to 21 years of age and recruited from the University of Lille3 participated voluntarily in the study. All participants received an information letter and provided written informed consent. All participants performed the task with their right hand and reported having normal or corrected-to-normal vision. The protocol received approval of the ethics committee in human sciences of the University of Lille3.

\section*{The spatial-tapping task}

Material \& Stimuli. A picture composed of 6 black targets was displayed on a touch screen Elo Touch 19'' 1915L. The targets ( 10 mm of diameter, distanced of 100 mm ) were placed around a virtual circle of 100 mm of radius. The participants were invited to stand in front of the screen that was placed upon a table and titled at 90 degrees of angle.

Procedure. The subjects' task was to touch each visual target one after the other, starting from the bottom right target, and moving counter-clockwise using the right index finger (fist closed). Participants were instructed to synchronize each pointing action to a series of regular auditory tones (beep duration \(=100 \mathrm{~ms}\) ) that was played through classic computer speakers. Participants were encouraged to maintain their left arm relaxed along the body side. They were clearly instructed of the goal of the task that was to be at best synchronized with the metronome. Each subject performed a total of ten trials. The initial tempo was an inter-tone interval (ITI) of 1200 ms . The temporal
interval was increased by 100 ms between each trial with the fastest one being at an ITI of 300 ms . Participants were required to produce sixty taps for each trial. The total duration of the session was 10 minutes, approximately.

\section*{Performance measures.}

Timing performance. Inter-response intervals (IRIs) were measured as the time interval between the start of two successive taps. Long intervals ( \(>2 *\) ITI) were omitted from all calculations. The \(\mathrm{IRI}_{\text {error }}\) was then computed as the percentage of absolute difference between each IRI and the reference ITI of a given trial. This measure served as an indicator of the magnitude of the timing error.

Spatial performance. The endpoint distributions of the pointing actions were plotted in function of each visual target position. All taps were used (ten data points per ellipse). Through vector calculations, spatial ellipses were then calculated. The mean area of the spatial ellipses was finally measured in \(\mathrm{mm}^{2}\) as an indicator of the magnitude of the spatial error (SE).

Motor Fluency. The Contact time (CT) was defined as the time of finger contact with the touch screen. This measure (in ms ) was used to assess the level of control of the motor response output, with shorter CTs being related to a more fluent gesture.

\section*{Planning indicators.}

Event-based vs. emergent timing. After having suppressed the first six IRIs of each trial, autocorrelation (AC) values were calculated at lag-1 and lag-6 (for details, see Vorberg \& Wing, 1996). These measures served as an indicator of the timing mode that was used to guide the pointing actions with event-based timing being revealed through negative AC-1 values, and emergent timing through positive AC-6 values.

Statistical Analyses. Performance measures and planning indicators were first calculated for individual trials and then averaged across participants. Second, analyses of variances (ANOVA) were conducted with ITI as a repeated measure on measures and indicators. Fisher LSD post hoc tests were used when required and the alpha level was set to 0.05 . The performance measures were then averaged within three phases for each subject according to the moment of change in planning strategy: slow or updating phase, intermediate or inhibiting phase, fast or switching phase.

\section*{Neuropsychological tasks}

Material \& Stimuli. The tasks were all selected from a French version of the TAP computerized battery of tests (Zimmermann \& Fimm, 1994). Participants were seated in front of the computer. The experimenter provided instructions orally. The same instructions were then displayed on the screen. A familiarization trial was performed before each task. One or two response keys were used, in function of task requirements. When only one response key was presented, subjects were asked to respond
with their right hand only. When two response keys were required, participants were asked to respond with their right hand on the right key, and their left hand on the left key. For each task, participants were asked to respond as rapidly as possible while maintaining a low error-rate.

\section*{Tasks procedures.}

Updating \& \(N\)-back task. A series of one-digit numbers were presented one after the other, in the centre of the screen ( 100 items that 15 target items). The subjects' task was to press the response key as fast as possible when the item on the screen was the same as the item presented two times before. Subjects were scored according to the median of their reaction times (RT).

Inhibiting \& Go-noGo task. A straight or a diagonal cross ('+' or ' \(x\) ') appeared briefly in the centre of the screen, for a total sequence length of 40 items. The participant's task was to press a response key as fast as possible for the diagonal cross ('x') only. The target item was present \(50 \%\) of the time. Participants were scored according to the number of false responses.

Switching \& The Letter-number task. A letter and a number were simultaneously presented on the computer screen, for a total sequence of 100 items. Two control conditions and one alternation condition were performed. Two response-keys were used. In the first control or pure block condition, the participants' task was to press the response key that was in the same hemi-field than the letter (for example, if the letter was presented on the left side of the screen, subjects had to press the left response key). In the second control condition, participants had to press the response key that was in the same hemi-field than the number. In the alternation condition, the participants were instructed to alternate a response to a letter, and a response to a number, from trial-to-trial, by pressing the response key that was located in the corresponding hemi-field. Participants were scored according to a switching cost that was measured as the difference in reaction time between the pure blocks (that were pooled together) and the alternation condition.

Statistical Analyses. Performance scores were calculated for each individual in each task and then averaged across participants. Descriptive results (mean, standard, deviation, \(\min \& \max\) values) were then computed. \(\chi 2\) tests were then performed to ensure that performance scores were normally distributed across tasks.

\section*{Multiple regression analyses: ST vs. classical tasks}

Standard multiple regression analyses were conducted to evaluate how well each performance scores ( \(\mathrm{IRI}_{\text {error }}\), CTs, area) obtained in each phase of the spatial-tapping task (regressors) could predict the scores obtained in each of the neuropsychological tasks (dependent variables). The alpha value was set at 0.05 . Following our hypotheses, (1) the timing accuracy in the slow phase of the spatial-tapping task (IRIerror) should require WM abilities (n-back task); (2)
control of the motor response (CTs) as required in the intermediate phase of the spatial-tapping task should be an indicator of inhibition abilities (go-no-go); (3) decrease in the spatial error in the faster phase of the spatial-tapping task (area) should be an indicator of the switching abilities (letter-number task).

\section*{Results}

\section*{The Spatial Tapping task}

\section*{Performance results.}

Timing performances. Results revealed that all participants closely followed the tempo even at fast tempi, with a maximum mean error of \(8 \%\) across the ITI spectrum. As an example, the greatest errors were observed in the fastest tempo of ITI= 300 ms , with IRIs contained between 276 and 324 ms . ANOVA on the \(\mathrm{IRI}_{\text {error }}\) revealed nevertheless that the timing errors were significantly different in function of ITI \((\mathrm{F}(9 ; 225)=7.685 ; \mathrm{p}<0.001)\). Post hoc tests revealed that timing errors were larger at faster tempi, i.e. for \(\mathrm{ITI}=400\) to \(300 \mathrm{~ms}(\) Mean \(=7.6 \% ; \mathrm{SD}=2.8 \%)\) and these values were significantly different from that measured at slower tempi (Mean \(=5.9 \%\); SD \(=1.7 \%\) ).

Spatial performance. ANOVA conducted on the spatial area revealed an increase in the spatial errors at faster tempi \((F(9 ; 225)=83.678 ; p<0.001)\). More specifically, the mean area of the endpoint ellipses were the smallest at slower tempi, i.e., at ITI \(=1100\) to \(800 \mathrm{~ms}\left(\right.\) Mean \(=42.9 \mathrm{~mm}^{2}\); SD \(=\) \(23.7 \mathrm{~mm}^{2}\) ), with all other tempi being characterised by significantly larger errors. Spatial areas were the largest at faster tempi (from \(\mathrm{ITI}=500\) to 300 ms ) and these results were significantly different from all other ITIs (Mean = \(126.8 \mathrm{~mm}^{2} ; \mathrm{SD}=73.96 \mathrm{~mm}^{2}\) ).

Motor fluency. ANOVA conducted on the mean contact times revealed that the CTs were shorter with increasing tempi \((\mathrm{F}(9 ; 225)=31.14 ; \mathrm{p}<0.001)\). Post Hoc tests revealed that the decrease in CT was linear with increasing tempi, with significant differences between the \(\mathrm{n}^{\text {th }}\) trial and the trial \((\mathrm{n}+2)\). Nevertheless, no differences between neighbouring ITIs emerged at the slower tempi, i.e. between ITIs \(=1200\) to 1000 ms for which the largest contact times were measured.

Overall, motor actions were precise in space and time at the slowest tempi ( \(\geq 800 \mathrm{~ms}\) ), less accurate in space but more fluent at intermediate tempi (at 700 and 600 ms of ITI), less precise in both space and time but much more fluent at faster tempi (from 500 to 300 ms of ITI).

\section*{Planning indicators.}

Event-based timing. To note first is the fact that all AC-1 values were negative. Repeated ANOVAs on the AC-1 values showed that these values were significantly different in function of ITI \((\mathrm{F}(9 ; 225)=5.933 ; \mathrm{p}<0.001)\). Post Hoc tests confirmed that the AC-1 values were the largest at slower tempi, i.e. for ITI=1100 to 900 ms of ITI (Mean = \(0.28 ; \mathrm{SD}=0.15\) ) and significantly smaller with increasing tempi, i.e. from \(\mathrm{ITI}=800\) to 400 ms (Mean \(=-0.17, \mathrm{SD}=\) 0.18 ). The AC-1 value was finally the smallest at the fastest

ITI of 300 ms (Mean \(=-0.02, \mathrm{SD}=0.21\) ). These results suggest that the timing mode was event-based for slow tempi ( 1200 to 900 ms ) in this task.

Emergent timing. To note first that all AC-6 values were positive. Repeated ANOVAs on the AC-6 values revealed that these values were significantly different in functions of ITI \((\mathrm{F}(9,225)=8.524 ; \mathrm{p}<0.001)\). Post Hoc tests revealed that the AC-6 values were the smallest at slower tempi, i.e. for \(\mathrm{ITI}=1200\) to \(900 \mathrm{~ms}(\) Mean \(=0.08 ; \mathrm{SD}=0.15)\). At faster tempi (ITI \(=500 \mathrm{~ms}\) to 300 ms ), the positive AC-6 values were the largest (Mean \(=0.31, \mathrm{SD}=0.20\) ) and not significantly different between each other. At intermediate tempi (from 800 to 600 ms ), the AC-6 values were all significantly smaller than at least one of the faster ITIs and larger than at least one of the slower ITI. These results indicate that the timing mode became emergent at fast tempi of 500 to 300 ms of ITI.

Overall, these results suggests that the timing was eventbased at slowest tempi (from 1200 to 900 ms ), emergent at fastest tempi ( 500 to 300 ms ) and in a transition phase at intermediate tempi (from 800 to 600 ms ).

\section*{Neuropsychological tasks}

Performance scores observed in the neuropsychological tasks are presented in table 1. \(\chi^{2}\) tests revealed that performance scores were normally distributed across tasks.

\section*{Multiple Regression Analyses: ST vs. classical tasks}

Beta coefficients and corresponding p-values are presented in table 2 for each regressor in function of each dependent variable. Results confirm our hypotheses: (1) small timing errors in the slow phase of the ST task predict short reaction times in the WM task, (2) short contact times in the ST-task predict larger number of inhibition error in the go-no-go task, (3) smaller are in the fast phase of the ST-task predict smaller switching cost in the letter-number task.

\section*{Discussion}

In this study, we asked whether a simple motor sequencing task could be used to assess the executive functions of planning described in the neuropsychological literature. This motor task was assumed to involve updating in working memory at slowest tempi, inhibiting at intermediate tempi, and switching at faster tempi. This hypothesis was tested by comparing performance scores in the spatial-tapping task to those scores obtained in three neuropsychological tasks selected to target each specific executive function. The findings reported here confirmed our working hypothesis. Indeed, in the slow phase of the spatial tapping task, results suggested that the motor actions were triggered through an internal representation of time intervals, with larger negative auto-correlations at slow tempi (from 900 to 1200 ms ). To perform the task adequately, subjects were required to maintain actively in working memory the target time interval to produce, across the entire duration of the trial.

Table 1: Performance scores in the classic executive tasks
\begin{tabular}{|c|c|c|c|c|}
\cline { 2 - 4 } & \begin{tabular}{c} 
N-back task \\
(UPDATING)
\end{tabular} & \begin{tabular}{c} 
Go-noGo task \\
(INHIBITING)
\end{tabular} & \begin{tabular}{c} 
Letter-number task \\
(SWITCHING)
\end{tabular} & \multirow{2}{*}{\begin{tabular}{c} 
Normal FREQUENCY \\
distribution
\end{tabular}} \\
\cline { 2 - 4 } & Median RT (ms) & False Alarms & Switching Cost (ms) & \\
\hline MEAN & 576 & 1 & 126 & \multirow{3}{*}{} \\
\hline SD & 158 & 1 & 89 & \\
\hline MIN & 368 & 0 & 15 & \\
\hline MAX & 928 & 3 & 371 & 4,08 \\
\hline FREQUENCY \(>(+) 1 \sigma\) & 6 & 0 & 4 & 17,75 \\
\hline FREQUENCY \(\langle(-) 1 \sigma\) & 15 & 23 & 18 & 4,08 \\
\hline\((-) 1 \sigma<\) FQ< \((+) 1 \sigma\) & 5 & 3 & 4 & \\
\hline\(\times 2\) test \((p-\)-value) & 0,467 & 0,051 & 0,996 & \\
\hline
\end{tabular}

Table 2: Multiple regression analyses. Significant results are bolded.
\begin{tabular}{|c|c|c|c|c|c|c|c|}
\hline \multicolumn{2}{|l|}{\multirow[t]{2}{*}{}} & \multicolumn{6}{|c|}{NEUROPSYCHOLOGICAL TASKS} \\
\hline & & \multicolumn{2}{|l|}{N -back (Updating)} & \multicolumn{2}{|l|}{Go-no-go (Inhibiting)} & \multicolumn{2}{|l|}{Letter-number (Switching)} \\
\hline \multicolumn{2}{|l|}{\multirow[t]{2}{*}{SPATIAL TAPPING}} & \multicolumn{2}{|l|}{Median RT (ms)} & \multicolumn{2}{|r|}{False Alarms} & \multicolumn{2}{|l|}{Switching cost (ms)} \\
\hline & & Beta Coeff. & p -value & Beta Coeff. & p -value & Beta Coeff. & p -value \\
\hline \multirow[t]{3}{*}{\begin{tabular}{l}
Slow phase \\
(1200 to 900 \(\mathrm{ms})\)
\end{tabular}} & IRIerror & 0,484 & 0,026 & -0,160 & 0,480 & 0,080 & 0,687 \\
\hline & CT & 0,021 & 0,931 & -0,408 & 0,139 & 0,202 & 0,392 \\
\hline & AREA & 0,066 & 0,804 & 0,434 & 0,144 & 0,389 & 0,134 \\
\hline \multirow{3}{*}{Interm. Phase ( 800 to 600 ms )} & IRIerror & 0,504 & 0,156 & -0,141 & 0,671 & -0,127 & 0,718 \\
\hline & CT & 0,173 & 0,413 & -0,474 & 0,026 & 0,116 & 0,585 \\
\hline & AREA & -0,172 & 0,614 & 0,156 & 0,631 & 0,420 & 0,227 \\
\hline \multirow[b]{3}{*}{Fast phase ( 500 to 300 ms )} & IRIerror & -0,043 & 0,843 & -0,032 & 0,882 & 0,025 & 0,901 \\
\hline & CT & -0,029 & 0,883 & -0,368 & 0,080 & -0,064 & 0,736 \\
\hline & AREA & 0,433 & 0,051 & -0,200 & 0,358 & 0,492 & 0,021 \\
\hline
\end{tabular}

Regression analyses confirmed furthermore the existence of a relationship between these motor results and those performances obtained in a cognitive WM task, suggesting similar WM functions in the cognitive and in the motor domains. In the faster phase of the spatial-tapping task (from 500 to 300 ms ), with actions becoming more circular, it was suggested that the action sequence was divided in two biomechanically distinct movements that compose the sequence, i.e., flexion vs. extension movement patterns. The cognitive goal in this case was then geared towards the need to coordinate smoothly the distinct movements composing the sequence towards a more global trajectory pattern that binds them together, here a circle. In circle drawing, this phenomena is actually measurable through the emergence of an anchor point that is a kinematic reduction in the timing and/or spatial variability at the point of transition between flexion/extension movements (Beek et al., 1992). It has been shown that the anchor point is effectively reduced through explicit anticipatory processes. Indeed, orienting the gaze in advance towards the anchor point significantly reduces the spatial variability observed at this point. In the same vain, flexing and extending the wrist in an anticipatory rhythmic fashion significantly reduces both the spatial and temporal errors respectively at the points of maximum flexion vs. extension related to the movement pathway (Roerdink et al., 2008). In the present results, we observed an anchor point within our circular trajectory at the point of maximal extension (upper left target); point in space at which smaller spatial variability was measured but only in the faster phase of the task, which confirms the role of emergent timing at fast tempi. In reference to the cognitive tests, our results revealed furthermore that the performances at fast tempi in the spatial tapping task were effectively related to smaller switching cost in a classical switching task, suggesting that
switching between two biomechanical movements may be in fact controlled by those similar functions used for cognitive switching. Finally, in the intermediate phase of the spatial-tapping task, it was suggested that motor inhibition of a too fast transition from discrete to continuous actions entered into play to maintain high levels of spatial and timing accuracy in spite of the increase in temporal pressure. This hypothesis was confirmed here with longer contact times in the intermediate phase of the spatialtapping task being significantly related with the ability to inhibit impulsive response in a classical go-no-go task.

In conclusion, we propose the spatial-tapping as a novel paradigm (1) to assess the executive functions in an easy and non-verbal context, (2) to gain a better understanding of the relationships between the distinct executive functions. This approach may be a promising way to reconsider cognitive strategy in broader context and offers a starting point for the study of the functional relationships between motor and cognitive control.

\section*{References}

Beek, P. J., Turvey, M. T., \& Schmidt, R. C. (1992). Autonomous and nonautonomous dynamics of coordinated rhythmic movements. Ecological Psychology, 4(2), 65-95.

Burgess, P. W. (1997). Theory and methodology in executive function research. In P. Rabbitt (Ed.), Methodology of frontal and executive function (pp. 81-116). Hove, UK: Psychology Press.

Damasio, A. R. (1994). Descartes' error: Emotion, rationality and the human brain. Avon, New York: Random House.

Dione, M., \& Delevoye-Turrell, Y. (unpublished). The spatial-tapping task to reveal the coexistence of event-based and emergent timing for the control of rhythmic sequences.

Grant, D. A., \& Berg, E. A. (1948). A behavioural analysis of degree of reinforcement and ease of shifting to new responses in a Weigl-type card-sorting problem. Journal of Experimental Psychology, 38, 404-411.

Harlow, J. M. (1868). Recovery from the passage of an iron bar through the head. Publications of the Massachusetts Medical Society, 2, 327-347.

Ivry, R., Spencer, R., Zelaznik, H., \& Diedrichsen, J. (2002). The Cerebellum and Event Timing. Annals of the New York Academy of Sciences, 978(1), 302-317.

Lemoine, L., \& Delignières, D. (2009). Detrended windowed (lag one) autocorrelation: A new method for distinguishing between event-based and emergent timing. The Quarterly Journal of Experimental Psychology, 62(3), 585-604. doi:10.1080/17470210802131896

Milner, B. (1963). Effects of different brain lesions on card sorting: The role of the frontal lobes. Archives of Neurology, 9(1), 90-100.

Miyake, A., \& Friedman, N. P. (2012). The Nature and Organization of Individual Differences in Executive Functions Four General Conclusions. Current Directions in Psychological Science, 21(1), 8-14.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A.
H., Howerter, A., \& Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "Frontal Lobe" tasks: a latent variable analysis. Cognitive Psychology, 41(1), 49-100.

Phillips, L. H. (1997). Do "frontal tests" measure executive function? Issues of assessment and evidence from fluency tests. In P. Rabbitt (Ed.), Methodology of frontal and executive function (pp. 191-213). Hove, UK: Psychology Press.

Rabbitt, P. (1997). Introduction: Methodologies and models in the study of executive function. In P. Rabbitt (Ed.), Methodology of frontal and executive function (pp. 138). Hove, UK: Psychology Press.

Repp, B. (2005). Sensorimotor synchronization: a review of the tapping literature. Psychonomic Bulletin \& Review, 12(6), 969-992.

Repp, B., \& Steinman, S. (2010). Simultaneous eventbased and emergent timing: synchronization, continuation, and phase correction. Journal of Motor Behavior, 42(2), 111-126.

Robertson, S., Zelaznik, H., Lantero, D., Bojczyk, K., Spencer, R., Doffin, J., \& Schneidt, T. (1999). Correlations for timing consistency among tapping and drawing tasks: evidence against a single timing process for motor control. Journal of experimental psychology. Human perception and performance, 25(5), 1316-1330.

Roerdink, M., Ophoff, E., Peper, C., \& Beek, P. (2008). Visual and musculoskeletal underpinnings of anchoring in rhythmic visuo-motor tracking. Experimental Brain Research., 184(2), 143-156.

Shallice, T. (1982). Specific impairments of planning. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 199-209.

Shallice, T. (1988). From neuropsychology to mental structure. Cambridge, U.K.: Cambridge University Press.

Torre, K., \& Delignières, D. (2008). Distinct ways of timing movements in bimanual coordination tasks: Contribution of serial correlation analysis and implications for modelling. Acta Psychologica, 129(2), 284-296.

Turvey, M. (1977). Preliminaries to a theory of action with reference to vision. Perceiving, acting and knowing, 211-265.

Vorberg, D., \& Wing, A. (1996). Modeling variability and dependence in timing. In Herbert Heuer and Steven W. Keele (Ed.), Handbook of Perception and Action (Vol. Volume 2, pp. 181-262). Academic Press.

Wing, A., \& Kristofferson, A. (1973). The timing of interresponse intervals. Attention, Perception, \& Psychophysics, 13(3), 455-460.

Zelaznik, H., Spencer, R., \& Ivry, R. (2002). Dissociation of explicit and implicit timing in repetitive tapping and drawing movements. Journal of experimental psychology. Human perception and performance, 28(3), 575-588.

Zimmermann, P., \& Fimm, B. (1994). Tests d'évaluation de l'attention (TEA)-Version 2.1. Würselen, Germany: Psytest.

\title{
Spatial congruity effects reveal metaphors, not markedness
}

\author{
Sarah Dolscheid 1,2 \\ (sarah.dolscheid@mpi.nl)
}

\author{
Cleve Graver 4 \\ (gravc243@newschool.edu)
}

\author{
Daniel Casasanto 3,4 \\ (casasand@newschool.edu)
}

\author{
1Max Planck Institute for Psycholinguistics, Nijmegen, NL \({ }_{2}\) International Max Planck Research School for Language Sciences, Nijmegen, NL \({ }_{3}\) Donders Center for Brain, Cognition, and Behaviour, Radboud University, Nijmegen, NL 4Department of Psychology, The New School for Social Research, New York, USA
}

\begin{abstract}
Spatial congruity effects have often been interpreted as evidence for metaphorical thinking, but an alternative markedness-based account challenges this view. In two experiments, we directly compared metaphor and markedness explanations for spatial congruity effects, using musical pitch as a testbed. English speakers who talk about pitch in terms of spatial height were tested in speeded space-pitch compatibility tasks. To determine whether space-pitch congruency effects could be elicited by any marked spatial continuum, participants were asked to classify high- and low-frequency pitches as 'high' and 'low' or as 'front' and 'back' (both pairs of terms constitute cases of marked continuums). We found congruency effects in high/low conditions but not in front/back conditions, indicating that markedness is not sufficient to account for congruity effects (Experiment 1). A second experiment showed that congruency effects were specific to spatial words that cued a vertical schema (tall/short), and that congruity effects were not an artifact of polysemy (e.g., 'high' referring both to space and pitch). Together, these results suggest that congruency effects reveal metaphorical uses of spatial schemas, not markedness effects.
\end{abstract}

Keywords: metaphor, polarity correspondence, markedness, musical pitch, space

\section*{Introduction}

Are high hopes somewhere in the air? Or what about rising prices? And where exactly are you when you are feeling down? Spatial metaphors like these are very common in language. Moreover, according to conceptual metaphor theory, people not only talk in terms of space but they also think metaphorically (i.e. spatially) (Lakoff \& Johnson, 1980). Whereas arguments in favor of this claim were initially based on linguistic data (and thus circular in nature), psychological experiments have now shown that spatial representations importantly contribute to people's understanding of domains like time (Casasanto \& Boroditsky, 2008), social dominance (Schubert, 2005), or valence (Meier \& Robinson, 2004).
Many of these psychological studies base their findings on binary compatibility tasks. In one experiment, for instance, participants were asked to classify dimensions in a metaphoric target domain (i.e., valence: judge the positive or negative valence of a word), while, at the same time aspects of the spatial source domain (i.e, location; up and down) were varied. In line with "GOOD is UP" metaphors, people were faster to evaluate positive words when they
appeared in a high spatial location compared to a low location (and vice versa for negatively valenced words) (Meier \& Robinson, 2004). Similarly, participants made faster judgments about social power when words for powerful people are at the top of a display and powerless people at the bottom (e.g., 'king' above 'slave', rather than vice versa; Schubert, 2005). These "metaphoric congruency effects" (Lakens, 2012), with faster performance for congruent compared to incongruent trials, have been taken as evidence that metaphoric target domains automatically activate congruent spatial information, supporting claims of conceptual metaphor theory (e.g. Meier \& Robinson, 2004; Schubert, 2005).

On an alternative account, however, it has been argued that congruency effects may be better explained as polarity alignment effects, also called markedness effects \({ }^{1}\) (Lakens, 2012). Like many other continuums in language and mind, metaphoric source and target domains (e.g. height or happiness) are considered to be bipolar. That is, they consist of an unmarked or +polar endpoint (e.g. high, happy), and an opposing marked or -polar endpoint (low, sad). Unmarked endpoints (+polar) are commonly defined as the default, evaluatively positive or broader dimension as opposed to the marked (-polar) ones (see e.g., Lehrer, 1985; Proctor \& Cho, 2006; for a critical approach see Haspelmath, 2006). Moreover, there is evidence that polarity differences affect cognitive processing. Participants show faster reaction times for unmarked (+polar) dimensions as compared to marked (-polar) ones (Clark, 1969; Seymour, 1974). Reaction time benefits for congruent metaphoric dimensions (like happy and up) could thus alternatively be explained by an additive processing advantage for + polar endpoints (e.g. happy +polar, up +polar): Across many studies, perceptual and linguistic judgments are faster when the poles of marked continuums are aligned (e.g., 'good' matched with 'up') than when they are misaligned (e.g., 'good' matched with 'down'; Clark, 1969; Lakens, 2012; Proctor \& Cho, 2006). The existence of markedness effects in binary response compatibility tasks raises a question: Does polarity alignment offer an alternative, non-metaphorical explanation for "metaphor congruency effects" like those reported by Meier \& Robinson (2004) and Schubert (2005), which rely on

\footnotetext{
1 Here, the terms "markedness" and "polarity" will be used interchangeably.
}
dimensional compatibility in binary speeded response tasks? And if so, what would this mean for theories of metaphorical mental representation?
Crucially, not all of the evidence for metaphoric thinking comes from (binary) congruency effects. Rather, it has been shown that people's metaphoric representations of domains like time or musical pitch map onto space in a continuous analog fashion (Casasanto, 2010; Dolscheid, Shayan, Majid, \& Casasanto, 2013). English speakers, for instance, who talk about musical pitch in terms of spatial height (high vs. low pitch; see e.g. Stumpf, 2006) also associate higher pitches with higher positions in space in nonlinguistic psychophysical tasks. In one study, participants were asked to reproduce musical pitches while watching lines varying in spatial height. Since lines were presented at multiple positions (i.e., 9 levels of height) in a random order, effects of space on pitch could not be attributed to (binary) polarity. Rather, participants' pitch reproductions were affected by the spatial information in a continuous way; tones accompanied by higher lines were reproduced at a higher frequency on average than the same tones accompanied by lower lines, resulting in a linear influence of height on pitch (Dolscheid et al., 2013). In this study, responses were not speeded, and the metaphor-congruity effects did not rely on the kind of binary stimulus-response compatibility that is believed to give rise to polarity alignment effects (Proctor \& Cho, 2006).
Furthermore, some mappings between space and musical pitch go against markedness. Whereas speakers of many languages (including English) refer to pitch in terms of spatial height, other languages like Farsi or Turkish encode pitch in terms of spatial thickness (Shayan, Ozturk, \& Sicoli, 2011). These thickness-pitch metaphors follow a reversed polarity alignment. Thick (+polar) refers to a low frequency pitch (-polar), whereas thin (-polar) refers to a high frequency pitch (+polar). Since Farsi speakers implicitly represent pitch in terms of thickness (Dolscheid et al., 2013), spatial schemas appear to be more important than polarity alignment.

Although experiments like Dolscheid et al.'s (2013) provide evidence for metaphorical mental representation that cannot be explained by markedness, the role of markedness in binary compatibility tasks remains controversial. Do source-target congruity effects merely show polarity alignment? Or do they reveal metaphoric associations? While metaphors and polarity are often indistinguishable in compatibility tasks (see also Lakens, 2012), we predict that when markedness and metaphor are juxtaposed, congruity effects will support metaphoric thinking, not markedness. What should matter is whether the words that participants have to classify in binary compatibility tasks activate the appropriate spatial schema (e.g., in the case of space-pitch mappings for English speakers, it should be a vertical spatial schema). That is, schema-appropriateness should be necessary, and markedness may not be sufficient to produce congruity effects.

In Experiment 1, we tested compatibility in height-pitch metaphors for 2 pairs of spatial terms, both paradigm cases of marked continuums (Clark, 1973). One pair corresponds to the poles of the correct spatial continuum (high-low), the other to the poles of an incorrect spatial continuum (frontback). High and front both constitute the unmarked or +polar endpoint, whereas low and back represent the marked or -polar endpoint (see e.g., Clark, 1973; Landsberg, 1995). Participants were asked to make binary speeded judgments on high-frequency and low-frequency pitches, classifying pitches either in a polarity-congruent way (e.g. high pitches as high or front), or in a polarityincongruent way (e.g. high pitches as low or back). If polarity alignment drives space-pitch congruity effects, then similar effects should be found when pitch is mapped to any marked linear spatial continuum, regardless of its orientation: High/low and front/back should both produce pitch-congruity effects. Alternatively, if activating a particular spatial schema for pitch is critical (i.e., the schema that is encoded in the participants' language), then high/low should result in a congruency effect, but front/back should not.

\section*{Experiment 1}

\section*{Methods}

Participants Twenty-four English speakers with no reported hearing problems participated for payment ( \(5 \$\) per 30 minutes). Four participants were excluded from analyses for not following instructions (i.e. they responded according to the wrong response mapping throughout at least one condition). They were replaced by a new sample of 4 participants who had not previously participated in the task.

Materials and Procedure Participants were asked to classify tones (one high and one low pitch) as quickly and accurately as possible by pressing buttons on the QWERTY keyboard (Q and P-keys). Stimuli were presented on an Apple iMac using Vision Egg 2.6 (Straw, 2008). Sounds were generated by Audacity software (http://audacity.sourceforge.net/) and comprised two pure tones (frequency: 262 and 440 hertz). Each tone lasted 400 ms . Participants listened to one tone at a time, via sealed headphones. Immediately following the offset of each tone, two response options (e.g., high, low) appeared, one on the bottom left and the other on the bottom right of the screen. Participants were instructed to classify the sound by pressing the button located under the corresponding word (e.g., high or low) as fast and accurately as possible. The left-right locations of the spatial terms varied randomly from trial to trial so that participants could not predict the location of the correct word in advance.

Spatial terms (high-low vs. front-back) were presented in 2 blocks, a high-low block and a front-back block. Within each block, spatial terms were crossed with 2 mappings (congruent, incongruent). The order of blocks was counterbalanced across participants. The order of congruity was counterbalanced within each block. Across blocks,
incongruent and congruent conditions were always presented in alternation. Before each condition, participants received 6 practice trials with feedback. Participants were also given an example illustrating the respective mapping before the practice trials.

Each condition consisted of 24 trials, 96 trials in total. In half of the trials a high pitch was presented, in the other half a low pitch. In the high-low congruent condition, the high pitch had to be classified as high and the low pitch as low. In the high-low incongruent condition, the high pitch had to be categorized as low and the low pitch as high. In the frontback congruent condition the high pitch had to be categorized as front and the low pitch as back (according to patterns of polarity/markedness). In the front-back incongruent condition the low pitch had to be categorized as front and the high pitch as back.

\section*{Results}

All data were analyzed using R (version 2.14.2; http://www.r-project.org/) and the R packages lme4 (Bates \& Maechler, 2009) and languageR (Baayen, 2009; cf. Baayen, 2008). We carried out linear mixed-effects regression models of Space (high-low versus front-back) and Congruency (congruent, incongruent) on accuracy and RTs. Using the principle of backward selection, we started out with a full (conservative) model which took into consideration not only the random intercept but also the random slopes of subject whenever it was appropriate (i.e., when the factor was a within-subject factor). Random intercepts and slopes of items were not included in the analysis due to the small number of items (4 words: high/low, front/back). To interpret the significance, we adopted the criterion that a given cosine was significant if the absolute value of the \(t\)-statistic (or \(z\)-statistic) exceeded 2 (Baayen, 2008).

Accuracy The mean accuracy for all target trials was 92.4\% ( \(S D=8.1\) ). For high/low conditions, accuracy was \(92.4 \%\) ( \(S D=9.7\) ), and for front/back conditions, accuracy was \(92.5 \%(S D=11.3)\). For congruent conditions, accuracy was \(95.9 \%\) ( \(S D=4.5\) ) and for incongruent conditions it was \(88.9 \%(S D=15.5)\). Analyzing accuracy by using a logistic mixed effects model on binary accuracy data yielded no main effects or interaction of Space (high/low, front/back) and Congruency (congruent, incongruent), (Space: \(z=|1.3|\); Congruency: \(z=|0.2|\); Space by Congruency: \(z=|1.2|\) ).

Reaction times Reaction times of the button presses were analyzed by linear mixed effects models. Only correct trials were considered which resulted in the exclusion of \(7 \%\) of the data. Responses greater or less than \(\pm 2\) SDs away from each participant's average RTs were also excluded, which resulted in the removal of \(6 \%\) of the accurate trials.

There was no significant main effect of Space \((t=|0.3|)\). The model yielded a significant main effect of Congruency ( \(t=|3.5|\) ) and a significant interaction of Congruency by Space ( \(t=|3.3|\) ). A linear mixed effect model of Congruency
on reaction times restricted to the level of high/low, yielded a significant effect of Congruency ( \(t=|4.5|\) ), demonstrating a congruity effect of high/low conditions. Restricting the model to the level of front-back yielded no significant effect of Congruency ( \(t=|0.2|\) ) (see Figure 1).


Figure 1: The influence of Space (high-low; front-back) and Congruency (congruent; incongruent) on pitch categorization (plotted in milliseconds).

\section*{Discussion}

In Experiment 1 we find congruency effects for high/low but not for front/back conditions, suggesting that activating the appropriate spatial schema (i.e., spatial height) is what is relevant in such binary response compatibility tasks. Words that activate a different (irrelevant) spatial schema (frontback), however, do not result in a congruity effect. This finding indicates that congruity effects cannot be attributed to markedness (polarity alignment), since 'front' and 'back' also name the unmarked and marked ends of a (sagittally oriented) linear spatial continuum.

\section*{Experiment 2}

Whereas high/low terminology is conventional for pitch in English, front/back is not. Maybe we only find a congruency effect in the case that is lexicalized, but not in the other case (front/back)? This skeptical interpretation would not change the fact that markedness is not sufficient to elicit congruency effects, but it would call into question our claim about activating the right spatial schema. Do we find congruency effects only because participants were using the polysemous words high/low, which can refer to "height" in both space and pitch?
To rule out this alternative explanation, in Experiment 2, we compared congruity effects in two pairs of spatial terms: tall/short and big/small. Neither pair of spatial expressions can be used in conventional English to describe the height of musical pitches (i.e., their frequency). If high/low congruency effects were driven by polysemy, then neither of these pairs of spatial terms should produce a congruity effect. However, if space-pitch congruity effects result from using words that activate a vertical spatial schema, then "tall" and "short" should produce a congruity effect because they are schematically appropriate (even though they are
lexically inappropriate). By contrast, "big" and "small" should not produce any space-pitch congruity effect, because these terms refer to 3 -dimensional size, and should not activate the appropriate 1 -dimensional vertical spatial schema (Dirven \& Taylor, 1988; Taylor, 2002).
In addition to testing whether the height-pitch congruity effect in Experiment 1 depended on the polysemy of "high" and "low," Experiment 2 also provides a second test of the sufficiency of markedness to produce space-pitch congruity effects. "Big" is the unmarked (positive) end and "small" the marked (negative) end of the big-small continuum. Therefore, markedness predicts that judgments should be faster when "big" is matched with "high" than when "small" is matched with "high."

\section*{Methods}

Participants Twenty-four English speakers with no reported hearing problems participated for payment ( \(5 \$\) per 30 minutes). One participant was excluded from analyses for not following instructions (i.e. the participant responded according to the wrong response mapping throughout one condition). He was replaced by a new participant who had not previously participated in the task.

Materials and Procedure The same procedure as in Experiment 1 was used, with the following exceptions. Rather than classifying pitches as high-low or front-back, participants classified them as tall-short for one block and big-small for the other.

In the tall-short congruent condition the high pitch had to be categorized as tall and the low pitch as short. In the tallshort incongruent condition the low pitch had to be categorized as tall and the high pitch as short. In the bigsmall congruent condition the high pitch had to be categorized as big and the low pitch as small (according to patterns of markedness). In the big-small incongruent condition the low pitch had to be categorized as big and the high pitch as small.

\section*{Results}

All data were analyzed using R (version 2.14.2; http://www.r-project.org/) and the R packages lme4 (Bates \& Maechler, 2009) and languageR (Baayen, 2009; cf. Baayen, 2008). We carried out linear mixed-effects regression models of Space (tall-short versus big-small) and Congruency (congruent, incongruent) on accuracy and RTs. Using the principle of backward selection, we again started out with a full (conservative) model which took into consideration not only the random intercept but also the random slopes of subject whenever it was appropriate (i.e., when the factor was a within-subject factor). Random intercepts and slopes of items were not included in the analysis due to the small number of items (4 words: tall/short, big/small). To interpret the significance, we adopted the criterion that a given cosine was significant if the absolute value of the \(t\)-statistic (or \(z\)-statistic) exceeded 2 (Baayen, 2008).

Accuracy The mean accuracy for all target trials was 94.8\% ( \(S D=11.4\) ). For tall/short conditions, accuracy was \(94.6 \%\) ( \(S D=13.1\) ), and for big/small conditions, accuracy was \(94.9 \%(S D=10.1)\). For congruent conditions, accuracy was \(96.2 \%\) ( \(S D=6.7\) ) and for incongruent conditions it was \(93.3 \%(S D=17.1)\). Analyzing accuracy by using a logistic mixed effects model on binary accuracy data yielded no main effects or interaction of Space (tall/short, big/small) and Congruency (congruent, incongruent), (Space: \(z=|1.0|\); Congruency: \(z=|1.0|\); Space by Congruency: \(z=|1.0|\) ).

Reaction times Reaction times of the button presses were analyzed by linear mixed effects models. Only correct trials were considered which resulted in the exclusion of \(4 \%\) of the data. Responses greater or less than \(\pm 2\) SDs away from each participant's average RTs were also excluded, which resulted in the removal of \(4 \%\) of the accurate trials.

There was no significant main effect of Congruency \((t=|0.9|)\). The model yielded a significant main effect of Space ( \(t=|2.3|\) ) and a significant interaction of Congruency by Space \((t=|3.2|)\). A linear mixed effect model of Congruency on reaction times restricted to the level of tall/short, yielded a significant effect of Congruency ( \(t=|3.0|\) ), demonstrating a congruity effect of tall/short conditions. Restricting the model to the level of big-small yielded no significant effect of Congruency ( \(t=|1.5|\) ) (see Figure 2).


Figure 2: The influence of Space (tall-short; big-small) and Congruency (congruent; incongruent) on pitch categorization (plotted in milliseconds).

\section*{Discussion}

In Experiment 2 we find a congruency effect in tall/short but not in big/small conditions. Therefore, congruity effects cannot be attributed to polysemy or markedness. Rather, space-pitch compatibility is based on activating the appropriate spatial schema, which serves as the source domain for English speakers' mental representations of musical pitch.

\section*{General Discussion}

In two experiments, we show binary response-time congruity effects attributable to metaphorical thinking, but not to markedness. Classifying pitches with vertical spatial terms elicited space-pitch congruity effects, but no comparable effects are found when people were asked to classify pitches with terms that name the poles of other marked spatial continuums (front vs. back; big vs. small). Polarity alignment (a.k.a. markedness), therefore, is not sufficient to produce space-pitch congruency effects. Rather, schema-appropriateness is necessary, supporting theories of metaphorical mental representation.
Moreover, congruity effects are not restricted to polysemous words like "high" and "low," which can be used for both space and pitch. Rather, congruity effects can also be found for words like "tall" and "short," which have no musical senses, but which activate a vertical spatial schema: the "active ingredient" in the observed space-pitch congruity effects.
In most cases, the polarities of metaphorical source and target domains are aligned (e.g. Lakens, 2012; Lakoff \& Johnson, 1980). For instance, happy, powerful, good, and high in pitch are all UP (the positive end of this spatial continuum), whereas their antonyms are DOWN (the negative end of the continuum). This relationship between metaphor and markedness makes it hard to determine the cause of many response compatibility effects. However, the polarities of metaphorical source and target domains are not always aligned. Musical pitch provides one domain, in which the marked end of the source domain (space) can be matched to the unmarked end of the target domain (pitch). Farsi speakers, for instance, represent pitch in terms of thickness. In Farsi speakers' language and thought, the unmarked pole of the spatial continuum (thick) is aligned with the marked pole of the pitch continuum (low frequency). Thus, metaphors and markedness can dissociate in Farsi - at least to the extent that markedness can be established in a principled way.

Making psychological predictions on the basis of markedness is problematic because researchers may disagree on how markedness is defined, and even on which end of a given continuum is marked. Whereas Schubert (2005) describes "powerful" as the marked and "powerless" as the unmarked endpoint of the "power" continuum, others have suggested the reverse (e.g. Lakens, 2012). In addition to these inconsistencies, it is not always clear what markedness actually means. By definition, quite a number of attributes like frequency, familiarity, or fluency, seem to be subsumed under the umbrella term markedness (see Haspelmath, 2006). In one experiment, for instance, Lakens (2012) manipulated polarity by adjusting the frequency of the 'marked' endpoint. While usually marked attributes like bad or down (-polar) occur less frequently, this was no longer the case for a group of Laken's participants. Critically, these participants also no longer showed a congruency effect, which was taken as evidence for a polarity account. However, in line with Haspelmath (2006),
it is questionable why one should talk about polarity when actually frequency is driving the effects. Unlike markedness, which is a notoriously ambiguous construct (e.g., Haspelmath, 2006, enumerates 12 distinct usages of this term in cognitive science), metaphors in language are more widely agreed upon. Expressions like "a high soprano" and "a low bass" make clear predictions about the spatial mappings that people should be activating for pitch, and therefore what congruity effects should be found: Linguistic metaphors tell us which end is "up."

Here we find an impact of spatial schemas on sourcetarget congruity as predicted by metaphors in language. Our results suggest an automatic, Stroop-like interference effect of metaphorical associations, converging with other findings of height-pitch congruity effects. In one task, for instance, participants made judgments about musical timbre while spatial height information was varied on a computer screen. Although pitch was irrelevant to the task, people's judgments were affected by the alignment of tonal and spatial height (Evans \& Treisman, 2010), suggesting a highly automatic source-target mapping (see also Rusconi, Kwan, Giordano, Umiltà, \& Butterworth, 2006; for limits of automaticity see Brookshire, Ivry, \& Casasanto, 2010).

Unlike previous experiments, here the spatial source domain was not manipulated physically but rather via linguistic stimuli (i.e., we presented words like high/low; tall/short etc.). This allowed us to directly assess effects of polysemy. In Experiment 2, height-pitch congruity effects could not simply be attributed to lexical overlap (high/low for space and pitch). Rather, we found that words activating a similar vertical schema (tall/short) were sufficient to trigger space-pitch congruity effects even if the words were lexically inappropriate. One could argue, however, that congruity effects in tall/short conditions were still indirectly driven by polysemy. Participants may have activated high/low terminology when classifying pitches, which then in turn led to semantic priming from high to tall, and low to short. However, although we cannot entirely rule out such priming effects, this explanation is unlikely to account for our results, for several reasons. According to Latent Semantic Analysis (LSA; http://lsa.colorado.edu/), 'tall' is more strongly related to 'short' (LSA cosine: .48) than to 'high' (LSA cosine: .31). Moreover, 'short' is about equally strongly related to 'high' (LSA cosine: .30) as to 'low' (LSA cosine: .31). Since activation is expected to spread between the most strongly related items (Collins \& Loftus, 1975), simple spreading activation would have wiped out a tall-short congruity effect rather than producing it. Moreover, although big is more closely related to high than to low (LSA cosine: . 18 versus .12) congruity effects remain absent in big/small conditions. The non-significant big/small effect even points into the opposite direction (see Figure 2), suggesting that semantic priming is unlikely to drive the observed patterns of results. \({ }^{2}\) Thus, while spatial

\footnotetext{
\({ }^{2}\) The trend toward a big-low congruity effect could be driven by underlying associations between size and pitch (e.g., see Evans \& Treisman, 2009) - but not by markendess or semantic priming.
}
terms like high/tall/big may be semantically related and overlap in markedness, we find that activating the appropriate vertical spatial schema is critical for producing space-pitch congruity effects.

\section*{Conclusions}

Metaphor congruency effects have been challenged by a polarity account, claiming that binary response compatibility effects may be better explained by markedness than by metaphorical thinking (Lakens, 2012). Indeed, metaphor and polarity are often hard to distinguish. However, here we show that when polarity and metaphor are juxtaposed, congruity effects support metaphorical thinking, not polarity.

Furthermore, these results show that it is not necessary to use polysemous words to produce source-target congruity effects (i.e., words that can refer to both the metaphorical source and target domains). Words that activate a vertical schema (e.g., tall/short) produce a space-pitch congruity effect despite being lexically inappropriate. Words that activate a different spatial schema (e.g., front/back, \(\mathrm{big} / \mathrm{small})\) do not produce any space-pitch congruity effect, despite naming the poles of other marked spatial continuums.
Together, these results indicate that activating the appropriate spatial schema is the "active ingredient" in space-pitch congruity effects - not polysemy or markedness - supporting theories of metaphorical mental representation.

\section*{Acknowledgments}

We thank Tom Gijssels for assistance in programming and for help with the experimental set-up, and Geoff Brookshire for help with R-analyses. This research was supported by a Travel Grant of Radboud University Nijmegen and an IMPRS fellowship to SD and by a James S. McDonnell Foundation Scholar Award (\#220020236) to DC.

\section*{References}

Baayen, H. R. (2009). LanguageR. R package.
Baayen, R. H. (2008). Analyzing linguistic data: A practical introduction to statistics using \(R\). Cambridge University Press.
Bates, D., \& Maechler, M. (2009). Package "Ime4"(Version 0.999375-32): linear mixed-effects models using S4 classes. Available (April 2011) at http://cran. r-project. org/web/packages/lme4/lme4. pdf.
Brookshire, G., Ivry, R., \& Casasanto, D. (2010). Modulation of motor-meaning congruity effects for valenced words. In S. Ohlsson \& R. Catrambone (Eds.), Proceedings of the 32nd Annual Conference of the Cognitive Science Society (pp. 1940-1945). Austin, TX: Cognitive Science Society.
Casasanto, D. (2010). Space for thinking. In V. Evans \& P. Chilton (Eds.), Language, cognition and space: The state of the art and new directions (pp. 453-478). London: Equinox Publishing.
Casasanto, D., \& Boroditsky, L. (2008). Time in the mind: Using space to think about time. Cognition, 106(2), 579-593.

Clark, H. H. (1969). Linguistic processes in deductive reasoning. Psychological Review, 76(4), 387-404.
Clark, H. H., Carpenter, P. A., \& Just, M. A. (1973). On the meeting of semantics and perception. In W. G. Chase (Ed.), Visual information processing (pp. 311-382). Chicago: Academic Press.
Collins, A. M., \& Loftus, E. F. (1975). A spreading-activation theory of semantic processing. Psychological review, 82(6), 407-428.
Dirven, R., \& Taylor, J. (1988). The Conceptualization of Vertical Space in English: The Case of Tall. In B. RudzkaOstyn (Ed.), Topics in Cognitive Linguistics (pp. 207-229). Amsterdam: Benjamins.
Dolscheid, S., Shayan, S., Majid, A., \& Casasanto, D. (2013). The Thickness of Musical Pitch: Psychophysical evidence for linguistic relativity. Psychological Science, Advance online publication. doi:10.1177/0956797612457374.
Evans, K. K., \& Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features. Journal of vision, 10(1), 6-12.
Haspelmath, M. (2006). Against markedness (and what to replace it with). Journal of Linguistics, 42(1), 25-70.
Lakens, D. (2012). Polarity correspondence in metaphor congruency effects: Structural overlap predicts categorization times for bipolar concepts presented in vertical space. Journal of Experimental Psychology: Learning, Memory, and Cognition, 38(3), 726-736.
Lakoff, G., \& Johnson, M. (1980). Metaphors we live by (Vol. 111). Chicago London: The University of Chicago Press.

Landsberg, M. E. (1995). Syntactic iconicity and linguistic freezes: the human dimension (Vol. 9). De Gruyter Mouton.
Lehrer, A. (1985). Markedness and antonymy. Journal of Linguistics, 21(2), 397-429.
Meier, B. P., \& Robinson, M. D. (2004). Why the Sunny Side Is Up. Associations Between Affect and Vertical Position. Psychological Science, 15(4), 243-247.
Proctor, R. W., \& Cho, Y. S. (2006). Polarity correspondence: A general principle for performance of speeded binary classification tasks. Psychological Bulletin; Psychological Bulletin, 132(3), 416-442.
Rusconi, E., Kwan, B., Giordano, B. L., Umilta, C., \& Butterworth, B. (2006). Spatial representation of pitch height: the SMARC effect. Cognition, 99(2), 113-129.
Schubert, T. W. (2005). Your highness: vertical positions as perceptual symbols of power. Journal of personality and social psychology, 89(1), 1-21.
Seymour, P. H. (1974). Asymmetries in judgments of verticality. Journal of Experimental Psychology, 102(3), 447-455.
Shayan, S., Ozturk, O., \& Sicoli, M. A. (2011). The thickness of pitch: Crossmodal metaphors in Farsi, Turkish, and Zapotec. The Senses and Society, 6(1), 96-105.
Stumpf, C. (2006). Tonpsychologie (Vol. 1). Leipzig: Hirzel.
Taylor, J. R. (2002). Near synonyms as co-extensive categories: "high" and "tall" revisited. Language Sciences, 25(3), 263-284.

\title{
A Two Process Account of the Memory Improvement Due to Choosing
}

\author{
Charles Driver (charles.driver@uqconnect.edu.au) \\ Department of Psychology, University of Queensland, Queensland, Australia, 4072
}

\begin{abstract}
Previous accounts of the mechanism which generates improvement in memory when people freely choose items, compared to other methods of item assignment, conflict and lack integration. I examine facial recognition performance of 40 participants asked to choose either the most or least attractive face from a series of pairs, and find that recognition of chosen faces is greater than recognition of unchosen, while no effect is found for the valence of faces. Considering these results in tandem with prior results and theories, I argue that a two process account of memory improvements due to choice is necessary, with one deliberative process occurring across all options available to choose from, and the other selective process focused on the actual chosen item. I detail the delineation of these processes, and describe and test the current best accounts of each - the multiple-cue hypothesis, and the self-reference effect for memory.
\end{abstract}

Keywords: Self; incidental memory; choice; self-choice effect; multiple cuing; learning and memory;

\section*{Introduction}

Why did you choose that brand at the supermarket, or that politician to vote for? Choices are everywhere, with ramifications from insignificant to major. A form of decision-making, choice more specifically involves a selection from a set of options. Substantial research has addressed the influence of various factors influencing our choices, but comparatively little has examined the direct cognitive impact of actually making a choice. One important cognitive impact is our memory of the choice, both because simply remembering them can be crucial, but also because the memory of previously choosing something has been shown to positively influence us in regards to that choice in future - even when that memory is incorrect. This positive influence occurs during both recall and encoding (Benney \& Henkel, 2006),via an influence on preferences (Bernstein \& Loftus, 2009), further decisions (Chen \& Zhang, 2003), as well as the attributes we associate with items (Benney \& Henkel, 2006; Henkel \& Mather, 2007). A better understanding of the way choices are remembered may thus help us to mitigate some of these arbitrary biases, improve our decision making in general, and inform our broader understanding of both memory and choice.

\section*{Memory improvements due to choice}

Freely choosing items has long been recognized to facilitate memory compared to other methods of item assignment (e.g. random assignment), but current accounts of the mechanisms driving this improvement focus on only
one possible mechanism, and conflict with each other. Early work in the field noted improvements in associative learning performance when participants chose their own learning materials (Perlmuter, Monty, \& Kimble, 1971). It was argued that increased motivation arising from having freely chosen either the stimulus or response increased performance. Later work by Takahashi (1991) examined recall and recognition performance for freely-chosen versus experimenter-assigned options, and found an improvement in memory for freely-chosen items that they termed the selfchoice effect. Takahashi posed a meta-memory interpretation for this effect, suggesting that while choice enhances memory, it does so because participants can select more easily remembered options, and not because freely choosing items increases motivation.

\section*{The multiple-cue hypothesis}

Watanabe (2001) critiqued both Takahashi and Perlmuter's explanations (though further endorsed the selfchoice effect) by demonstrating that the effect was maintained in a constrained choice condition. Watanabe asked participants a series of questions, and provided a number of answers to select from for each, although only one was correct. Participants had to process the various options and make a selection - just as with free choices but only one of the options answered the question appropriately. Because the self-choice effect was maintained when participants went through the process of choosing, yet could not simply choose their favored or more easily remembered options, Watanabe rejected both motivational and meta-memory explanations. Instead, Watanabe proposed a parallel between the self-choice effect, and the generation effect. When participants generate responses to a question, they are generally remembered better than when they are assigned responses (Slamecka \& Graf, 1978). This generation effect also accords with Craik and Lockhart's (1972) levels-of-processing framework, which suggests that the more substantively we think about an item, the better we remember it. The premise for the generation effect is that responding to a question typically involves generating multiple candidate responses, comparing and deliberating between them, then selecting the best response. Watanabe suggests that in order to compare responses, positive and negative attributes of each must be generated. They then argue that while choosing between presented options does not involve the generation of candidate responses themselves, it does involve the generation of attributes for each candidate, and then a
comparison between these generated attributes. They call this the multiple-cue hypothesis, and suggest it is responsible for the improved memory in both generation and self-choice effects; the item attributes that are generated in both cases provide additional cues to facilitate memory of the item. The generative process encourages the encoding of item-specific information (i.e. detailed information about the item), while the comparative process encodes relational information (i.e. how the item connects to other items and concepts in memory).

\section*{The Self-reference effect}

An entirely separate stream of research into choice suggests that rather than improvements in memory being driven by the generative and comparative processes of deliberation, they instead arise because the feeling of freely choosing something causes us to associate it with our selves (Cloutier \& Macrae, 2008). This is suggested as arising because self-reference is encouraged by the sense of personal responsibility in selection. The self-reference effect for memory was proposed by Rogers, Kuiper and Kirker (1977) after many observations that we remember items such as traits, nouns, prose, mental imagery, etc. better when we consider how the items relate to ourselves than to others. Symons and Johnson conducted an extensive meta-analysis of theories and findings for the self-reference effect (1997), and suggest that items encoded in relation to our self construct benefit from the highly developed memory structure (i.e., well organised, richly detailed and massively connected), already in place regarding our self. This self-referential encoding facilitates recognition and recall, as more connections exist to other memory traces. As with the self-choice effect, Symons and Johnson argue that self-reference encourages both item-specific and relational information encoding (i.e., increases encoding of both detail regarding the specific item, and associated connections with other memory traces). Additionally, Symons and Johnson suggest that we use our self-concept as a default source of retrieval cues. This default means that in the absence of other sources of cues, we look to memory traces associated with our self. As such, self-referentially encoded items are also likely to be recalled with reference to the self. Such compatibility of encoding and retrieval conditions - transfer-appropriate-processing (Morris, Bransford, \& Franks, 1977) - leads to improved memory for the items (Wells, Hoffman, \& Enzle, 1984).

When Cloutier and Macrae (2008) extended self-reference to the domain of choices, they examined whether selfreference could improve memory when participants were not explicitly directed to think in terms of the self. They theorized that when someone feels personally responsible for the assignment of an item to their self (i.e. free choice), greater self-reference should be elicited, improving memory for the item. While this is similar to work on the self-choice effect, this study allows a critical delineation between the
act of selecting an item and the process of deliberating between items, by removing the latter. Cloutier and Macrae paired 44 participants, and had each member of the pair take turns to pass slips of paper to the experimenter from either a central bowl (free choice condition), or a participantspecific bowl (assigned condition). The experimenter then read out loud a number printed on the slip, and indicated a positive trait that corresponded to that number from a hidden list - removing any possible influence of deliberation between choices on memory. Even though a choice from the central bowl resulted in just as random and arbitrary a selection as an initial assignment of items to individual bowls, recall and recognition performance of the traits that participants chose themselves via the central bowl was significantly better than both traits chosen by the other participant, and traits when participants chose from the selfspecific bowl. While the explanation that this memory improvement occurs due to increased self-reference needs further examination and testing, this work does show that the mere act of selecting an item can enhance memory for it, so long as a sense of choosing is maintained - even in situations where the choice is arbitrary and no processing of the options is involved.

\section*{Two mechanisms for memory improvement?}

The multiple-cue hypothesis and self-reference accounts pose very different mechanisms for memory improvements due to choosing. Watanabe's (2001) examination of the self-choice effect and multiple-cue hypothesis suggested that deliberation between options drives memory improvements when choosing items, but Cloutier and Macrae's (2008) work on choice and self-reference demonstrated that the sense of freely choosing an item, devoid of deliberation, is sufficient to elicit better memory for it. This latter sense alone could be sufficient to account for the full extent of memory improvement due to choosing, with no need for the multiple-cue hypothesis. However, the effect that the sense of freely choosing has on memory for unchosen items is unknown, whereas Watanabe showed that memory improvements also occurred for deliberated and compared, but unchosen, items. For the sense of freely choosing to sufficiently account for all memory improvements due to choice, it must endow unchosen items with comparable memory improvements to chosen items. That is, chosen and unchosen options that have been deliberated between should exhibit comparable memory performance. This requires a direct comparison between unchosen and chosen item memory performance, which prior works have not pursued (Watanabe compared both to different assignment methods). So, while prior works posit different, but singular, processes for the memory improvement due to choice, a direct comparison between unchosen and chosen can show whether one is sufficient, or that multiple processes are necessary.

\section*{Is self-reference relevant?}

Beyond the question of one or two (or more) processes, is the question of why the sense of freely choosing an item is important for memory - how does it cause the improvement. Cloutier and Macrae (2008) argued that it is the self-referentiality caused by the sense of choosing which drives memory improvements. However, because participants in their studies only chose their own items, and were not asked to choose items from the central bowl for the other participant, it is unclear whether it is the sense of choosing an item to suit oneself, or the mere sense of choosing at all that is driving the effect. Were the latter found to be the case, self-reference would either need to be construed very differently in regards to choice, or a new explanation developed. Cloutier and Macrae acknowledge a further limitation, in that they only provided positive traits in each condition - would the act of selection have improved memory for neutral or negative traits in the same manner as for positive? Items which were deemed less favorable, less in line with personal preferences and desires, should elicit fewer connections in memory with the self. Although far from a complete test of the theory, examining this can enlighten us as to whether the theoretical background of self-reference is a strong basis for this observed effect of freely choosing on memory.

\section*{Outline of the experiment}

To determine whether freely choosing between a pair of options causes differences in memory of unchosen and chosen items, and thus whether one or two processes are necessary to account for memory improvements due to choice, I will examine whether our memory for chosen items is better than our memory for unchosen. Support for this would demonstrate that the sense of freely choosing is insufficient on its own to account for memory improvements due to choosing. Such a finding would then lend support to a two process account, whereby improvement across both chosen and unchosen items arises due to deliberation between the items, while further improvement in memory for only the chosen item arises due to the sense of having personally selected it.

Further to determing whether a one or two process account is necessary, I will also examine the mechanism of the memory improvements due to the sense of freely choosing. I examine this aspect because while I do not debate the existence of the effect, it is very unclear that selfreference is the mechanism. To assess this I will ask whether when we choose options we favor, we will exhibit more improvement in memory than when we choose unfavorable options. Such a finding would lend support to the notion that self-reference is behind the memory improvement associated with the sense of freely choosing.

To examine these questions I will present participants a series of choices between two facial images. Half will be asked to choose the most attractive face, and half the least
attractive. The faces will then be presented one at a time, interspersed with unfamiliar faces, and recognition for the faces will be assessed.

\section*{Method}

\section*{Participants}

Forty undergraduate psychology students (Six male) of mean age 19.57 years \((S D=2.64)\) participated in this study for course credit.

\section*{Materials}

Participants completed a computer-based choice and subsequent memory task. Colour images of female faces were used for the choices offered to participants, the same as in Tangen, Murphy and Thompson (2011). Originally from a Slovakian database of 200 faces (SmartNet IBC, n.d.), the faces were eye-aligned using PsychoMorph (Tiddeman, Burt, \& Perrett, 2001), and clearly identifying extraneous features such as piercings were digitally removed as necessary. An example face pairing is shown in Figure 1.


Figure 1: Example of facial image pairings

\section*{Procedure}

Participants were initially told they would be completing a task in which they had to select faces based on attractiveness, followed by a range of cognitive testing. They were given basic demographic questions followed by a series of 18 choices. Each choice was between a pair of female face images as shown above in Figure 1. The order of pair presentation was randomized for each participant. Half of the participants were asked to "please choose the face you find most attractive", while the other half were asked to choose least attractive. Participants then performed a word unscrambling distracter task for two minutes, after which a surprise memory test commenced. During the memory task, the 18 chosen faces, 18 unchosen faces, as well as 36 new faces, were presented serially in randomized order. Participants were told that they would be presented both old and new faces, but were not informed of the proportion, or of how long the test would last.

\section*{Design}

A 2 (choice type: most attractive, least attractive) \(\times 2\) (memory stimulus: chosen, unchosen) mixed factorial design was used, where choice type was randomly assigned and varied between-participants. Memory stimulus varied within-participants in a randomized order. Recognition was gauged by asking participants "Did you see this face in the earlier sequence of pairs", with a forced-choice response of yes or no. The percentage of recognition responses to each type of memory stimulus was calculated for each participant.

\section*{Results}

Initial examination of the data in Table 1 suggests that chosen faces are recognized more often than unchosen, while both are recognized more often than new faces. Participants appear to be more accurate in their recognition judgment when presented with a new face than either a chosen or unchosen, in that they make the correct judgment \(84 \%\) of the time. However this simply suggests that people are more likely to not recognize a previously seen face than they are to mistakenly recognize a new face.

Table 1: Recognition rates based on choice type.
\begin{tabular}{lccccc}
\hline & \multicolumn{2}{c}{\begin{tabular}{c} 
Choose most \\
Memory \\
attractive \((\mathrm{n}=20)\)
\end{tabular}} & & \multicolumn{2}{c}{\begin{tabular}{c} 
Choose least \\
attractive \((\mathrm{n}=20)\)
\end{tabular}} \\
\cline { 2 - 3 } \cline { 5 - 6 } & M & SD & & M & SD \\
\hline Chosen face & .71 & .14 & & .71 & .12 \\
Unchosen face & .62 & .15 & & .65 & .14 \\
New face & .16 & .8 & & .16 & .9 \\
\hline
\end{tabular}

Recognition rates were assessed with a mixed betweenwithin subjects \(2 \times 2\) (choice type [most, least] \(\times\) memory stimulus [chosen, unchosen]) analysis of variance. There was no main effect of choice type on recognition, \(F(1,38)=\) \(0.12, p=.731\). This lack of main effect of choice type indicates that the type of choice participants were asked to make - choosing either the most or least attractive - did not influence overall recognition rates. There was a significant main effect of stimulus type, \(F(1,38)=12.48, p=.001\). This main effect of stimuli type indicates that previously chosen faces were recognized significantly more often than faces that were shown and participants deliberated over, but left unchosen. No significant interaction between choice type and memory stimuli was observed, \(F(1,38)=0.47, p=\) .490. This absence of interaction indicates that the observed effect of memory stimuli on recognition was not influenced by choice type; the improvement in recognition for chosen compared to unchosen faces was the same whether people were asked to choose the most, or the least, attractive face.

\section*{General Discussion}

When asked to choose between pairs of faces, participants later recognized faces they had chosen more frequently than the faces they did not choose - deliberating between two items and then selecting one led to improved memory for the chosen item compared to the unchosen. This lends support to my proposal that two processes are necessary to account for memory improvements due to choosing. Previous accounts of this memory improvement include the multiple-cue hypothesis (Watanabe, 2001) and the selfreference effect (Cloutier \& Macrae, 2008). My results show that neither explanation alone is sufficient to account for the range of results seen across research in the area, but that in combination they may provide an effective account.

The multiple-cue hypothesis (Watanabe, 2001) suggests that deliberating between items generates additional attributes (which can function as cues) associated with each item, thereby improving recognition and recall of the items. This was based on observations that any items that were part of a selection process (both chosen and unchosen) were better remembered than items that were merely assigned to participants. However, Cloutier and Macrae (2008) showed that even in the absence of deliberation between items, a sense of involvement in the selection process is sufficient to improve memory for items. This sense of involvement could thus have potentially accounted for the full extent of memory gains that people exhibit when choosing between items. However, because personal involvement in the decision is comparable for both chosen and unchosen items, and my results show that participants' recognition of chosen items is greater than unchosen, the sense of involvement is insufficient to explain Watanabe's earlier findings of improved memory for unchosen items compared to items assigned by an experimenter. As such, it appears that at least two processes are necessary to account for memory improvements due to choosing. Some portion of improvement likely occurs as a function of the process of deliberating between options, which improves memory for all items involved in the decision process, while another process related to the act of selecting one option improves memory primarily (or only) for the selected option.

\section*{Multiple-cues and the deliberative process}

Although insufficient to explain the entirety of improvements in memory due to choosing, the multiple-cue hypothesis (Watanabe, 2001) is presently the best explanation for the memory improvement arising from the process of deliberating between items, and is thus responsible for improvements observed for unchosen items involved in a deliberative process compared to unchosen items not involved in a deliberative process. Further understanding of the influence of this deliberative process on memory could be gained by varying the number of options selected between while constraining the time available to choose. I expect that the inclusion of additional
options would reduce the influence of deliberation, while maintaining a similar effect of selection. That is, if choosing between five options instead of two, differences in memory between chosen and unchosen options should be greater, because any quickly rejected options will benefit less from deliberative processes, while chosen options should still benefit similarly from the sense of freely choosing. Note though that it was important in this work to limit the options in choices to two, to avoid confounding an effect of choosing with an effect of rejecting early. This is because if three options are presented, one could be rejected as inappropriate early in the choice process, leading to reduced memory for it, unrelated to the actual act of selection. If this effect were apparent, it would mean that any examination of differences between unchosen and chosen when more than two options are presented is likely to be confounded by attentional differences due to early rejecting. However, now having confirmed the influence of both deliberative and selective processes on memory, choices with more than two options may be able to both further test the two mechanisms, and highlight the relative effects of each.

\section*{Self-reference and the act of selection}

Memory improvements due merely to the act of selection (as distinct from the deliberation between items) have only recently been noted, with the frequently observed but somewhat narrowly specified self-reference effect (Rogers et al., 1977) posed by Cloutier and Macrae (2008) as the only explanation so far. Typically regarded as the phenomenon whereby participants remember items assessed in terms of their relation to their self better than items assessed in relation to others (Symons \& Johnson, 1997), it is generally elicited by explicitly asking participants to think of an item in relation to their self or another. When stepping outside this paradigm, it is difficult to determine a priori the types of items or actions likely to elicit selfreferential processing. Nevertheless, Cloutier and Macrae posed self-referential processing as responsible for participants improved memories of items they felt they chose themselves, compared to memories of items they felt others had chosen. In their case however, rather than varying whether the item was assessed in relation to the self or other, they varied the sense of who was responsible for selecting the item.

To test support for Cloutier and Macrae's proposed selfreference explanation, half of my participants were asked to choose items they found most attractive, while half chose items they found least attractive. Items which more closely represent a participants preferences and desires were expected to be more likely to induce associations with their self (Cloutier \& Macrae, 2008), however, no difference in recognition performance between more and less attractive faces was found, and participants performed no differently when choosing either the most or least attractive faces.

While this result raises questions as to the applicability of self-reference as an explanation for memory improvements due to choice, further research is needed to assess whether this null result is: an artifact of the type of stimuli used; an indication that an items' self-referentiality does not vary based on the strength of accord with personal preferences; or an indication that the observed memory improvement due to the sense of personally choosing is not due to selfreference. These multiple possibilities highlight what I think is a weakness in the application of self-reference based theories - unless directly and overtly introduced, greater self-reference must be assumed in order for self-reference to function as an explanatory mechanism for observed improvements in memory. An alternative approach to this question would be to repeat my experiment, but ask some participants to make choices based on their own preferences, and others to make choices for a friend. This would more closely match the typical approach to eliciting selfreference, such that choosing for a friend should result in less self-reference. Should such an experiment also yield a null result, the predictive usefulness of self-reference as an explanation for the observed effect would seem very questionable.

\section*{Delineating between the two processes}

I have demonstrated that memory improvements due to choosing are likely best described by two distinct processes, as opposed to the singular process accounts described in the literature until now, which I have shown to be inadequate. These two processes are delineated as the initial deliberation between multiple options, or latter act of selecting one specific option. The best accounts of the individual processes at present are the multiple-cue hypothesis (Watanabe, 2001) as driving that portion of improvements due to deliberation, and a self-reference effect (Cloutier \& Macrae, 2008) further improving memory only for those items which are felt to be personally selected. The multiplecue hypothesis was developed in response to a series of works in the area of memory for choices, and allows for relatively clear predictions and testing. Results here neither confirm nor disconfirm the theory, but the further testing I propose involving more choice candidates will go some way towards this. While my results clearly show an effect related only to selection, there are weaknesses with selfreference as an explanation for this effect. My results failed to demonstrate one possible prediction of self-reference, that the valence of items should influence the level of selfreference - yet the weaknesses of self-reference (in current guise) as a theory for memory of choices is seen in that alternative explanations can be given for my observed results which maintain a causal role for self-reference. One such explanation could be that self-reference occurs whenever we judge an item for ourselves, and we do not in fact more closely associate positive items and features with ourselves. My proposal of adhering more closely to the
original self-reference paradigm will be one informative step for this issue.

While I have proposed a number of specific extensions to this line of research, I also suggest that a critical feature of my work - the direct comparison between chosen and unchosen options - is crucial to further investigations related to choice. This distinction allows a clear delineation of the process of deliberating between items, from the act of selecting a single item. That is, the process of deliberating involves cognitive processes that occur across all options, selecting one option rather than another results in differential cognitive effects - direct comparisons between chosen and unchosen options allow these to be examined.

\section*{Towards a reduction of biases}

A final, more general direction this research could fruitfully be extended, is the influence of memories for choosing, and not choosing, on preferences and future decisions. How do the arbitrary biases in memory (Benney \& Henkel, 2006; Henkel \& Mather, 2007; Mather, Shafir, \& Johnson, 2000), and resulting preferences (Bernstein \& Loftus, 2009), that choosing causes vary under different conditions of choice? Could we enhance or minimize this post-choice polarization between options? My results suggest this may be possible: while the choice posed to participants (which face they found more, or less, attractive) varied, the information gained did not - regardless of choice type, we come to know both the face they found most attractive, and the face they found least attractive. Therefore varying the direction of the original choice, manipulating how well the different options and the decision are remembered, may be a very simple way to alter the extent of bias people show towards items after making a choice between them. The earlier suggestion of testing with more than two options could also be explored here. If the multiple-cue hypothesis is adequate, more choices may further reduce our memories for unchosen items, and in turn reduce any biases towards the unchosen options which arise from a memory of not choosing them.

\section*{Acknowledgements}

I thank Jason Tangen for his support across all aspects of the project, and Mike Humphreys and Giulia Paparo for their comments on various portions. The School of Psychology at University of Queensland provided the participant pool.

\section*{References}

Benney, K. S., \& Henkel, L. A. (2006). The role of free choice in memory for past decisions. Memory, 14(8), 1001-1011.
Bernstein, D. M., \& Loftus, E. F. (2009). The consequences of false memories for food preferences and choices. Perspectives on Psychological Science, 4(2), 135-139.

Chen, C. Y., \& Zhang, S. (2003). A dynamic choice process: How choices generate biased memory that influences future choices. In P. A. Keller \& D. W. Rook (Eds.), Advances in consumer research (Vol. 30)
Cloutier, J., \& Macrae, C. N. (2008). The feeling of choosing: Self-involvement and the cognitive status of things past. Consciousness and Cognition, 17(1), 125-135.
Craik, F. I. M., \& Lockhart, R. S. (1972). Levels of processing: A framework for memory research1. Journal of Verbal Learning and Verbal Behavior, 11(6), 671-684.
Henkel, L. A., \& Mather, M. (2007). Memory attributions for choices: How beliefs shape our memories. Journal of Memory and Language, 57(2), 163-176.
Mather, M., Shafir, E., \& Johnson, M. K. (2000). Misremembrance of options past: Source monitoring and choice. Psychological Science, 11(2), 132-138.
Morris, C. D., Bransford, J. D., \& Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. Journal of Verbal Learning and Verbal Behavior, 16(5), 519-533.
Perlmuter, L., Monty, R. A., \& Kimble, G. A. (1971). Effect of choice on paired-associate learning. Journal of Experimental Psychology, 91(1), 47.
Rogers, T. B., Kuiper, N. A., \& Kirker, W. S. (1977). Selfreference and the encoding of personal information. Journal of Personality and Social Psychology, 35(9), 677688.

Slamecka, N. J., \& Graf, P. (1978). The generation effect: Delineation of a phenomenon. Journal of Experimental Psychology: Human Learning and Memory, 4(6), 592.
Symons, C. S., \& Johnson, B. T. (1997). The self-reference effect in memory: A meta-analysis. Psychol Bull, 121(3), 371-394.
Takahashi, M. (1991). The role of choice in memory as a function of age: Support for a metamemory interpretation of the self-choice effect. Psychologia: An International Journal of Psychology in the Orient.
Tangen, J. M., Murphy, S. C., \& Thompson, M. B. (2011). Flashed face distortion effect: Grotesque faces from relative spaces. Perception, 40(5), 628-630.
Tiddeman, B., Burt, M., \& Perrett, D. (2001). Prototyping and transforming facial textures for perception research. Computer Graphics and Applications, IEEE, 21(5), 42-50.
Watanabe, T. (2001). Effects of constrained choice on memory: The extension of the multiple-cue hypothesis to the self-choice effect. Japanese Psychological Research, 43(2), 98-103.
Wells, G. L., Hoffman, C., \& Enzle, M. E. (1984). Selfversus other-referent processing at encoding and retrieval. Personality and Social Psychology Bulletin, 10(4), 574584.

\title{
Motion in Vision and Language: Seeing Visual Motion can influence Processing of Motion Verbs
}

\author{
Carolin Dudschig (carolin.dudschig@uni-tuebingen.de) \\ Department of Psychology, Schleichstr. 4 \\ 72076 Tübingen, Germany
}

Jan Souman (jansouman@hotmail.com)
Department of Psychology, Schleichstr. 4
72076 Tübingen, Germany
Barbara Kaup (barbara.kaup@uni-tuebingen.de)
Department of Psychology, Schleichstr. 4
72076 Tübingen, Germany

\begin{abstract}
In contrast to symbolic models of language understanding, embodied models of language comprehension suggest that language is closely connected with visual and motor processing. In the current study we show that motion words, such as rise or fall, are processed faster if displayed against a background of compatible motion (e.g., upward vs. downward random dot motion with \(60 \%\) motion coherence). However, this interaction between semantic processing and visual processing only occurred if the word and the motion display were presented simultaneously. If the visual motion display was short-lived and occurred 100 or 200 ms after word-onset, no interactions between language and visual motion were found. We suggest that only in situations that do not allow ignoring or strategically suppressing the visual motion display, supra-threshold visual motion can affect language comprehension.
\end{abstract}

Keywords: Language processing; motion verbs; vision; visual motion processing; embodiment; grounding.

\section*{Introduction}

Embodied models of language understanding propose a close connection between language and perceptuomotor processes in the brain (e.g., Barsalou, 1999). Recently, compelling evidence supported the close association between language and other cognitive functions (e.g., Zwaan, Stanfield \& Yaxley, 2002). In the motor domain converging evidence suggests that language facilitates compatible motor actions (e.g., Glenberg \& Kaschak, 2002) and that language comprehension involves cortical motor areas that are also involved in performing the described actions (e.g., Hauk, Jonsrude, \& Pulvermüller, 2004). For example, Glenberg and Kaschak showed that processing sentences such as "Close the drawer" can interfere with motoric responses incompatible with the motion implied in the sentence (e.g., arm movement towards my body). Similar effects have been reported in studies using motion verbs (e.g., rise, climb) or nouns implicitly implying a location (e.g., bird vs. shoe), whereby upward verbs and nouns facilitate upward arm movements (Dudschig, Lachmair, de la Vega, De Filippis, \& Kaup, 2012a;

Lachmair, Dudschig, De Filippis, de la Vega \& Kaup, 2011). In contrast to the effects of language on motor processing, in the perceptual domain there is rather mixed evidence regarding the relation between language and visual processing. In particular, evidence regarding the influence of non-linguistic factors on language processing is rare. This direction of cause is particularly important, as these findings would suggest that mechanisms underlying non-linguistic processes are required and recruited during language processing.

Studies in the visual domain typically investigate the influence of language on perceptual detection or discrimination tasks. For example, it has been shown that words referring to entities with a typical location (e.g., hat vs. shoe) can influence visual target perception in upper or lower screen locations (e.g., Dudschig, Lachmair, de la Vega, De Filippis, \& Kaup, 2012b; Estes, Verges \& Barsalou, 2008). Similar results have been reported for valence words (e.g., Meier \& Robinson, 2004) and religious concepts (e.g., Chasteen, Burdzy \& Pratt, 2010). Additionally, there have been studies demonstrating that visual simulation can also occur during sentence processing and subsequently affect visual discrimination performance (Bergen, Lindsay, Matlock \& Narayanan, 2007). Recently, it has been shown that not only visual discrimination performance but also eye-movements can be affected by words referring to entities in the upper or lower field of vision (Dudschig, Souman, Lachmair, de la Vega, \& Kaup, 2013). More specifically, upward saccades are faster following words referring to entities in the upper visual field (e.g., bird) and in contrast, downward saccades are faster following words referring to entities in the lower visual field (e.g., shoe). Importantly, the relation between language and visual processing was also reported in the other causal direction: Perceiving visual motion patterns can affect language processing. For example, Kaschak, Madden, Therriault, Yaxley, Aveyard, Blanchard and Zwaan (2005) first reported the effects of visual motion perception on language comprehension. In their study, participants viewed visual motion patterns (e.g., upward vs. downward moving
horizontal stripes on a screen) and at the same time had to listen to sentences and perform a sensibility judgment task. The results showed that reading times were slower when the visual motion (e.g., upwards pattern) matched the motion direction implied by the sentence (e.g., "The rocket blasted off'). The authors concluded that language processing demands access to visual processing resources. If these visual processing resources are engaged by the processing of motion patterns, sentence understanding can be impaired.

Interestingly, studies investigating the effect of visual motion percepts on single word comprehension reported opposing results. Meteyard, Zokaei, Bahrami and Vigliocco (2008) analyzed how the understanding of motion verbs (e.g., rise vs. fall) is influenced by activation of motionresponsive visual brain areas. In their study, motion verbs were presented on a screen together with a short-lived ( 200 ms ) visual motion pattern, whereby the visual motion pattern was noisy to a greater or lesser extent. In the nearthreshold condition, the motion display was presented at a coherence level that made it difficult for the participants to detect the motion direction of the motion pattern. In the above-threshold condition, motion coherence was set to a level that clearly allowed classification of the motion direction (upward vs. downward moving pattern). Participants had to perform a lexical decision task. The results showed that near-threshold motion patterns facilitated processing of words implying compatible motions (e.g., rise was faster processed if presented together with a near-threshold upward motion). In the other experiments where the visual motion was set to abovethreshold levels no effect of visual motion perception on language processing was observed. The authors suggested that visual motion activates motion-responsive areas in the brain ( \(\mathrm{MT}+\) ). However, this activation can be suppressed by top-down control mechanisms in the case of abovethreshold motion coherence only. Thus, only in nearthreshold motion patterns the motion information resulted in interactions with semantic language processing. In contrast, in the case of above-threshold visual motion pattern topdown control was recruited and suppressed this visual activation. Importantly, in the study by Meteyard et al. the visual motion patterns were presented very briefly ( 200 ms ) in contrast to 35 sec visual motion percepts in the study of Kaschak et al. (2005). Taken together there is mixed evidence regarding the influence of visual motion perception on language processing. On the one side, abovethreshold and long-lasting visual motion can influence sentences processing (Kaschak et al., 2005), on the other side, only near-threshold visual-motion patterns affected lexical access to single words (Meteyard et al., 2008). Thus, it remains open whether above-threshold visual motion can interact with semantic language processing on a word-level.

In the current study we investigate whether single-word processing can be affected by above-threshold visual motion if visual motion patterns are presented from word onset until response. Such findings would be important for the embodied model of language understanding, as they would
suggest convergence in the empirical evidence in favor of the model, and suggest that both word and sentence processing are influenced similarly by co-occurring visual motion. In order to test this we adapted the visual motion displays used by Meteyard et al. (2008) and created abovethreshold random dot motion displays, that clearly allowed classification of the motion as an upward or downward directed motion. Additionally, we manipulated the stimulus onset asynchrony (SOA) between the word display and the visual motion display. In the 0 ms SOA condition the word and visual motion were displayed simultaneously. In the 100 ms SOA condition first the word was displayed and after 100 ms delay the visual motion pattern appeared. Similarly, in the 200 ms SOA condition, the visual motion pattern followed the word display by 200 ms . Importantly, only in the 0 ms SOA condition word and motion display fully overlapped. Thus, in this condition the simultaneous presentation of word and motion display minimizes the possibility of the participants to ignore the visual motion display. We expected that in conditions were participants were constantly exposed to visual motion during the lexical decision task, visual motion will most strongly influence semantic language processing.

\section*{Method}

\section*{Participants}

Eighteen right-handed psychology students from the University of Tübingen took part in this experiment ( \(M_{\text {age }}=\) \(24.39,16\) female) for monetary reward or course credit.


Figure 1: Trial examples for Go-Trials (word) and NoGo Trials (non-words). Visual motion was either compatible to the motion implied by the verb (top-left display) or incompatible (bottom-left display). Arrows illustrate visual motion direction and were not displayed in the actual experimental setup.

\section*{Stimuli \& Apparatus}

The experiment took place in a sound-attenuated booth. Participants viewed the screen from a 60 cm viewing distance. Experimental procedure was implemented in MATLAB R2010a, Psychtoolbox, 3.0.8.

Words Twenty-four German verbs \({ }^{1}\) referring to upwards motion and 24 verbs referring to a downwards motion were used as experimental stimuli. Upwards and downwards motion verbs did not differ in length ( \(M_{\text {up }}=8.74\) ( \(S D=\) \(\left.1.18), M_{\text {down }}=8.35(S D=2.01), t(44)=.69, p=.49\right)\). Word frequency was retrieved from the Leipziger Wortschatzportal, upwards and downwards motion verbs did not differ in word frequency ( \(M_{\text {up }}=1886.17\) ( \(S D=\) \(3545.31), M_{\text {down }}=1667.70(S D=3134.64), t(44)=.22, p=\) .83. Additionally, 48 pronounceable non-words were constructed. Therefore we used a different set of German verbs and permuted and exchanged various letters.

Visual Motion Patterns Visual motion patterns were adapted from Meteyard et al. (2008) with some adjustments, in order to make the motion clearly visible to the participants. 1000 moving dots were included in each display moving at a speed of \(20^{\circ} / \mathrm{s}\). Dot size was \(0.1^{\circ}\). Dots were presented within an aperture of approximately 15 cm diameter. Figure 1 shows examples of compatible and incompatible visual motion trials.

\section*{Procedure \& Design}

Each experimental trial started with the presentation of a fixation cross in the middle of the screen for the duration of 500 ms (size: 20 pixels). Then, either a word or a non-word replaced the fixation cross. Words were presented in Arial font with a size of \(0.5^{\circ} \times 2.5^{\circ}\) visual angle. In the 0 ms SOA condition, the visual motion pattern was presented together with the word. In the 100 ms and 200 ms SOA conditions, the visual motion pattern followed word onset by 100 or 200 ms , respectively. Words and visual motion were presented until response. Participants had to press the space bar if they decided that the displayed stimulus is a word and withhold response in case of non-word trials. If no response was recorded within 1500 ms the next trial started automatically. The inter-trial-interval was 500 ms . 20 Practice trials were conducted using a separate set of verbal stimuli. The experiment consisted of 576 Go-Trials (word trials) and 576 NoGo-Trials (non-word trials). Each of the 48 words was presented four times in each SOA condition (twice with an upward motion pattern and twice with a downward motion pattern). The experimental design was a within-subject design, with the factors \(\operatorname{SOA}(0,100,200 \mathrm{~ms})\), visual motion

\footnotetext{
\({ }^{1}\) Exemplary German verbs denoting to upwards motion: steigen to rise), erhöhen (to increase), klettern (to climb), wachsen (to grow), hissen (to hoist), erheben (to lift) etc.. Exemplary German verbs denoting to downwards motion: fallen (to fall), sinken (to sink), tauchen (to dive), tropfen (to drip), landen (to land), schütten (to pour), einstürzen (to collapse) etc.
}
(upward, downward) and word direction (upward, downward).

\section*{Results}

All NoGo-Trials and erroneous trials were excluded from analysis. Error exclusion reduced the dataset by 1.40 \%. Additionally, outliers were excluded from reaction time (RT) analysis, with a criterion of 4 SD reducing the dataset by less than \(0.43 \%\). The lexical decision times were analyzed in two repeated measures ANOVAs. In the first ANOVA participant was the random-factor ( \(F_{1}\) : byparticipant analysis) and in an additional ANOVA the stimulus word served as random-factor ( \(F_{2}\) : by-item analysis).

Reaction time results are displayed in Figure 2. There was a main effect of word direction in the by-participant analysis only, \(F_{1}(1,17)=13.06, M S E=834, p<.01, F_{2}(1,46)=1.56\), \(M S E=11520, p=.22\), with responses to downwards word ( 624 ms ) being faster than to upwards words ( 639 ms ). There was no effect of visual motion, \(F_{1}(1,17)=0.12, M S E\) \(=652.8, p=.74, F_{2}(1,46)=0.27, M S E=731.4, p=.60\), nor of \(\mathrm{SOA}, F_{1}(2,34)=0.59, M S E=721.2, p=.56, F_{2}(2,92)=\) 2.53, \(M S E=896.2, p=.09\). There was no interaction between visual motion and SOA, \(F_{1}(2,34)=0.54, M S E=\) \(454.5, p=.59, F_{2}(2,92)=0.64, M S E=901.3, p=.53\). There was no interaction between word direction and SOA, \(F_{1}(2,34)=2.56, M S E=486.4, p=.09, F_{2}(2,92)=0.34\), \(M S E=896.2, p=.71\). There was no interaction between word direction and visual motion direction, \(F_{1}(1,17)=3.15\), \(M S E=573.8, p=.09, F_{2}(1,46)=0.62, M S E=731.4, p=\) .44. Importantly, there was a significant three-wayinteraction between word direction, visual motion and SOA, \(F_{1}(2,34)=3.56, M S E=456.7, p<.05, F_{2}(2,92)=3.03\), \(M S E=901.3, p=.05\). Separate analysis of the SOA conditions showed, that the three way interaction was due to the interaction between word direction and visual motion being significant for the 0 ms SOA condition only, \(F_{1}(1,17)\) \(=8.64, M S E=585, p<.01, F_{2}(1,46)=6.43, M S E=790, p\) \(<.05\) and not for the \(100 \mathrm{~ms} \mathrm{SOA}, F_{1}(1,17)=0.01, M S E=\) \(483.4, p=.94, F_{2}(1,46)=0.96, M S E=839.9, p<.033\) or the \(200 \mathrm{~ms} \mathrm{SOA}, F_{1}(1,17)=0.00, M S E=491.1, p=.97\), \(F_{2}(1,46)=0.04, M S E=904.3, p<.84\). In summary, visual motion direction did interact with lexical processing. However, this was only in trials were word and visual motion display fully overlapped ( 0 ms SOA condition). Posthoc tests showed that this effect was due to faster classification of words referring to upward motion (e.g., rise, climb) if presented on the background of an upward motion in contrast to a downward motion, \(t_{1}(17)=-2.27, p<\) \(.05, t_{2}(23)=-2.40, p<.05\). In contrast word referring to a downward motion (e.g., fall, drip) were faster classified if presented on the background of a downward motion, this was reflected in a trend in the by-subject analysis, \(t_{1}(17)=\) \(1.93, p=.07, t_{2}(23)=1.24, p=.22\).


Figure 2: Reaction time results for the lexical decision task, separately for the three SOA conditions, the word direction and the visual motion direction. Error bars represent confidence intervals for within-subject designs according to Loftus and Masson (1994).

\section*{Discussion}

Converging evidence suggests that language processing is closely related to other cognitive functions and can affect visual and motor processing. Interestingly, some studies also report an effect of motor processes (e.g., Glenberg, Sato, \& Cattaneo, 2008) or visual processing (e.g., Kaschak et al., 2005; Meteyard et al., 2008) on language comprehension, suggesting direct involvement of visual and
motor processes during language understanding. Kaschak et al. (2005) reported that visual motion perception (e.g., downwards motion) interferes with understanding of sentences that imply compatible motions (e.g., "The confetti fell on the parade"). In contrast, Meteyard et al. showed that near-threshold visual motion facilitates lexical access to words that imply compatible motion directions. In the current experiment, we addressed the question whether above-threshold visual motion can affect lexical processing of motion verbs if the participants have no possibility to strategically ignore the visual information. Indeed, our results showed that in conditions where word display and visual motion display occurred simultaneously ( 0 ms SOA ) and persisted throughout the trial, visual motion patterns did interact with lexical processing of the motion verbs. More specifically, we found that upward motion words (e.g., rise) are processed faster if displayed against the background of an upward motion than against downward motion, and the opposite holds for downward motion verbs.

To our knowledge, these findings are the first that show an effect of above-threshold visual motion on single-word processing. But why do we not find interference effects as reported by Kaschak et al. (2005)? First of all, single-word processing might differ regarding the mechanisms how visual processing resources are activated during reading, thus language-vision interactions might occur at different time-points or processing stages. Indeed, previous studies showed that timing can play a crucial role and may determine whether facilitation or interference effects are found (e.g., Boulengner, Roy, Paulignan, Deprez, Jeannerod \& Nazir, 2006). Additionally, in our study we used motion patterns that were very different from Kaschak et al. (moving dot patterns vs. moving bars) and our moving dot patterns were only displayed during each trial. In contrast, Kaschak et al. (2005) displayed motion for as long as 35 s and motion display extended between trials. Moreover, sentences were presented auditory in Kaschak et al.'s study. Thus, differences in task parameters and language material might results in facilitation effects in our study. Indeed, we adapted our visual motion patterns from the study of Meteyard et al. who also reported facilitation effects in case of single-word processing and lexical decision tasks.

This directly leads to the next question: Why do we find a facilitation effect despite using above-threshold motion patterns that can be clearly classified as upward or downward moving motion pattern? In the study of Meteyard et al. these influences of visual motion on language understanding were only observed for near-threshold motion patterns. Previously it has been suggested that the influence of task-irrelevant sub-threshold motion patterns on task performance is stronger than the influence of suprathreshold motion patterns (Tsushima, Sasaki, \& Watanabe, 2006). The authors suggested that sub-threshold motion patterns are processed in the visual cortex similar to suprathreshold motion patterns; however in contrast to suprathreshold motion patterns sub-threshold motion patterns do not automatically result in recruitment of inhibitory control
from the lateral prefrontal cortex (LFPC) in order to inhibit the visual cortex activation (in MT+ ) and thus reduce the influence of the motion percept on responding. The fact that we do find an influence of supra-threshold visual motion patterns on lexical decision task might have several implications. First, as our motion display occurred throughout the whole trial participants might fail to recruit sufficient top-down control mechanisms in order to fully suppress the influence of the visual motion on performance. Additionally, if language processing and visual processing are directly related, small activation in the visual cortex might also be sufficient to influence language processes. Thus, due to top-down control inhibitory control from the LPFC that suppressed visual motion activation, the effects in our study might be rather small. Additionally, in the 100 ms and the 200 ms SOA condition, the LPFC suppression mechanisms on the MT+ activation might be stronger, as it might be easier to suppress the influence of a visual motion display that is delayed in onset to the critical stimulus. Further studies will be required to fully understand the interplay between the language and the visual system and the critical time intervals during language processing, where this interaction occurs.

In summary, our findings have several implications. First, our results suggest that visual motion can also affect language processing if visual motion is presented abovethreshold. Second, these findings pose a challenge to some findings in the motor domain. Typically, in motor tasks participants can see their arm or hand motion. Thus, if participants are instructed to perform a lexical decision task, decision times might be faster in compatible directions, because the visual input from the moving arm or hand will interact with lexical processing. Thus, given that our findings show that word processing interacts with visual motion perception, some findings in the motor domain might also be explained by perception of actual arm or hand movements. In future studies interactions between motor action and language need to be considered carefully, as potentially also being influenced by visual motion perception. In summary, our findings show that language processing and visual processing are closely interrelated. In paradigms, where participants cannot ignore or actively avoid motion perception, language processing can be facilitated by compatible visual motion.

\section*{Acknowledgments}

We thank our student assistants for help with data collection. This project was supported by a Margarete-vonWrangell Fellowship appointed to Carolin Dudschig (European Social Fund and the Ministry Of Science, Research and the Arts Baden-Württemberg) and by the SFB833/B4 project of Barbara Kaup (German Research Foundation).

\section*{References}

Barsalou, L.W. (1999). Perceptions of Perceptual Symbols. Behavioral and Brain Sciences, 22, 637-660.
Bergen. B. K., Lindsay, S., Matlock, T., \& Narayanan, S. (2007). Spatial and linguistic aspects of visual imagery in sentence comprehension. Cognitive Science, 31, 733-764.
Boulenger, V., Roy, A. C., Paulignan, Y., Deprez, V., Jeannerod, M., \& Nazir, T. A. (2006). Cross-talk between language processes and overt motor behavior in the first 200 ms of processing. Journal of Cognitive Neuroscience, 18, 1607-1615.
Chasteen, A.L., Burdzy, D.C., \& Pratt, J. (2010). Thinking of God moves attention. Neuropsychologia, 48, 627-630.
Dudschig, C., Lachmair, M., de la Vega, I., De Filippis, M., \& Kaup, B. (2012). From top to bottom: spatial shifts of attention caused by linguistic stimuli. Cognitive Processing, 13, 151-154.
Dudschig, C., Lachmair, M., de la Vega, I., De Filippis, M., \& Kaup, B. (2012). Do task-irrelevant directionassociated motion verbs affect action planning? Evidence from a Stroop paradigm. Memory \& Cognition, 40, 10811094.

Dudschig, C., Souman, J., Lachmair, M., de la Vega, I., \& Kaup, B. (2013). Reading "Sun" and Looking Up: The Influence of Language on Saccadic Eye Movements in the Vertical Dimension. PloS ONE, 8, e56872.
Estes, Z., Verges, M., \& Barsalou, L. W. (2008). Head up, foot down: Object words orient attention to the object's typical location. Psychological Science, 19, 93-97.
Glenberg, A.M. \& Kaschak, M.P. (2002). Grounding language in action. Psychonomic Bulletin \& Review, 9, 558-565.
Glenberg, A. M., Sato, M., Cattaneo, L., Riggio, L., Palumbo, D., \& Buccino, G. (2008). Processing abstract language modulates motor system activity. The Quarterly Journal of Experimental Psychology, 61, 905-919.
Hauk, O., Johnsrude, I., \& Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. Neuron, 41, 303-307.
Kaschak, M. P., Madden, C. J., Therriault, D. J., Yaxley, R. H., Aveyard, M., Blanchard, A. A., \& Zwaan, R. A. (2005). Perception of motion affects language processing. Cognition, 94, B79-B89.
Lachmair, M., Dudschig, C., De Filippis, M., de la Vega, I., \& Kaup, B. (2011). Root versus roof: automatic activation of location information during word processing. Psychonomic Bulletin \& Review, 18, 1180-1188.
Loftus, G. R., \& Masson, M. E. (1994). Using confidence intervals in within-subject designs. Psychonomic Bulletin \& Review, 1, 476-490.
Meteyard, L., Zokaei, N., Bahrami, B., \& Vigliocco, G. (2008). Visual motion interferes with lexical decision on motion words. Current Biology, 18, R732-R733.
Tsushima, Y., Sasaki, Y., \& Watanabe, T. (2006). Greater disruption due to failure of inhibitory control on an ambiguous distractor. Science, 314, 1786-1788.

Zwaan, R. A., Stanfield, R. A., \& Yaxley, R. H. (2002). Language comprehenders mentally represent the shapes of objects. Psychological Science, 13, 168-171.

\title{
Using Music as a Turn in Conversation in a Lesson
}

\author{
Sam Duffy (s.duffy@eecs.qmul.ac.uk) \\ Patrick G. T. Healey (ph@eecs.qmul.ac.uk) \\ Cognitive Science Research Group \\ School of Electronic Engineering and Computer Science \\ Queen Mary University of London \\ London, E1 4NS, United Kingdom
}

\begin{abstract}
Music is sometimes compared to language as a system of communication, however this comparison is usually at a generic formal, cultural or social level. This paper explores this analogy at the detailed level of interaction: to what extent can musical contributions act as conversational turns? We explore this question through an ethnographic study of music lessons. We describe a new transcription notation designed to capture the interactional details of musical contributions. Using this notation we show that although the ultimate objective of a lesson is development of musical performance, the detailed structure of the musical contributions depends on their interactional organisation. We show that musical contributions display interactional structure at the turn and sub-turn level and are closely integrated with other verbal and non-verbal cues as part of the unfolding conversation.
\end{abstract}

Keywords: music tuition; conversational turn; interaction; repair

\section*{Introduction}

Comparisons of music and speech are well documented, for example Feld and Fox (1994) critically review a broad interdisciplinary collection of anthropological work on the relationship of music to language. Besson, Chobert, and Marie (2011) take a cognitive approach, considering evidence for the bidirectional influence of musical expertise on speech processing and of linguistic expertise on the processing of harmonic sounds. Zatorre, Belin, and Penhune (2002) show that whilst each auditory cortex, in the left and right hemispheres of the brain, has been shown to favour processing of either music or speech, this complimentary specialisation should "be seen as arising from a single underlying principle rather than being unrelated phenomena".

In this paper we look at the relationship between music and speech in social interaction through the study of instrumental music lessons. We describe a new transcription notation designed to capture the interactional details of musical contributions. This is used to investigate the extent to which musical utterances produced during a lesson act as conversational turns. Initially we will summarise some of the characteristics and rules which govern turn taking in conversation. We will then look at the different, context-dependant roles that music can play in different types of interaction. We will use an ethnographic study of clarinet lessons to provide examples of the interplay between music and speech in a pedagogical setting

\section*{What is a Turn?}

In conversation, the participants manage their exchange of units of speech. Sacks, Schegloff, and Jefferson (1974) set out
rules describing the mechanics of how this is achieved, identifying dialogue as a turn organised activity. In order to investigate any turn taking system, be it playing a game of cards or managing a queue, there is a need to define what constitutes a turn. In conversation analysis, the building block for turns is the turn construction unit (TCU), which can be formed from a single word or utterance, a sentence, or a phrase. As one speaker approaches the possible completion of a TCU, another speaker may recognise this as a transition-relevance place (TRP) where they can take the floor. However this is an opportunity rather than an obligation. It may be that the current speaker starts a new TCU and continues with their turn. Since it is possible to predict when turns are heading towards completion, the next speaker can often start their turn without a perceivable gap in the conversation, or even start before the current speaker has finished, causing a brief overlap. Whilst we do not usually talk at the same time as someone else for prolonged periods, brief overlaps like this at transition points are frequent. To investigate these rules in action, a set of notations was proposed for use in conversation analysis transcripts (for example see Figure 1).
```

// The point in the current speaker's speech where overlap
y the other starts
The point where overlap finishes
= latching (no interval between two pieces of talk
() elapsed time in tenths of seconds, used to denote
pauses and silence
Prior syllable is prolonged
empty parenthesis indicate an utterance which could not
be clearly heard
double parenthesis indicate features other than
verbalisation

```

Figure 1: Conversation analysis notation (Sacks et al., 1974)

\section*{Non-verbal interaction}

The gestures which accompany speech are an important scaffold to conversation as well as an integral part of interaction. Bavelas, Chovil, Lawrie, and Wade (1992) explain that conversation is not made up of alternating monologues but is an interactive social system, and show that interactive gestures are essential in maintaining conversation. Cassell and Thorisson (1999) describe the importance of envelope feedback, non-verbal accompaniments to speech such as beat gestures, gaze and head turns. Engle (1998) demonstrate that when gesture and speech are consistent with respect to the underlying referent, they are understood as composite signals rather than separate channels. Clark and Krych (2004) show the im-
portance of visual modality to collaborative work with shared objects, gesturing in relation to an object whilst another is speaking effectively reducing the number of turns required for task completion. Non-verbal interactions can also function as a turn in themselves, the expected response to a verbal turn taking the form of physical activity rather than speech. For example, a request at the family dinner table during a busy multi-party conversation to pass the butter can be actioned without any reference in the continuing conversation (E. A. Schegloff, 2007, pp. 10 Chicken Dinner).

In the same way that gestures act as a scaffold to to speech, non-verbal interaction is vital to the co-ordination of musical sound production. In jazz, a solo may be scored as for a prescribed number of bars with a chord structure, but it may also be open, thrown around the group for others to take a turn or even choose to play against each other, so non-verbal communication with fellow performers is important to manage improvisation. Observing a group of musicians engaged in free improvisation sessions, Healey, Leach, and Bryan-Kinns (2005) found that the musicians used the patterns of body position and orientation, or f-formations (Kendon, 1990) typical of face-to-face conversation to organise the timing of their musical contributions. Moran (2011) observes a group of North Indian musicians whose vocal and bodily responses to musical ideas could be interpreted as comparable to the function of back channelling in everyday conversation. Even in the performance of a predetermined composed score, an ensemble must synchronise entrances and exits, and changes in dynamics and tempo and this is usually achieved by gestures, head and body movement and gaze. In an analysis of co-ordination between members of a string quartet, Davidson and Good (2002) wrote "The nature of the interaction and coordination in conversation are, we believe, analogous to that in small group music-making contexts."

\section*{Turn breakdown and self repair}

An important part of maintaining conversation is dealing with turn-taking errors or rule violations such as a misunderstandings, interruptions, gaps and overlaps. These are frequent in natural dialogue and we manage these breakdowns through repair. Self-repair occurs when the current speaker manages an error within the same turn in which it was made. To do this, the speaker must be able to self-monitor and detect a problem in what they are saying, or see some outward sign of the listener's confusion. Correction is made promptly once the problem has been detected, a neutral holding term such as 'uh' often being used to communicate error detection to the listener and so hold the turn for the self-repair to be made (Levelt, 1983). Self-repair occurs much more frequently than other-initiated repair in natural speech. However even if the other speaker initiates repair, they are much more likely to encourage the original speaker to correct themselves, rather than make a direct correction (E. Schegloff, Jefferson, \& Sacks, 1977).

\section*{The Role of Music in a Pedagogical Context}

There are many different types of interaction involving the production of music. From an ethnomethodological perspective, the social norms which govern how members of a group understand their world and so behave in it (Garfinkel, 1964) are relevant to any interaction, whatever the mode of communication. The context of a musical interaction is therefore important in determining the roles of the participants and the communication content of the music produced. Players in an ensemble rehearsal will use verbal interaction to analyse, discuss and shape their approach to a piece. During a performance, they cannot use the same level of verbal interaction unless they want to share their inner workings with the audience. A soloist is not generally expecting to enter into a dialogue with the audience, however they may engage in extensive non-verbal interaction with fellow performers and the conductor, as part of co-ordinating their performance with an orchestra.

In a pedagogical context, such as an instrumental lesson, both student and tutor produce music but it is subject to immediate scrutiny, their musical utterances being produced with the expectation of immediate feedback. The tutor is not listening to the performance from the perspective of an audience member, but as an expert critic and must be able to immediately verbalise their assessment of the student's performance. In order to prepare a complete piece or movement, they will focus on a small part of it each week, building up the work gradually. Small fragments of music, perhaps only a few bars, are worked on at a time. The student plays them, receives feedback, then plays them again for the tutor to assess if they have incorporated the feedback, and so on in an iterative process.

\section*{Representing Music in a transcript}

It is surprising how few authors looking at interaction in a musical context attempt to represent the musical sounds produced. When they do, one approach is to use musical notation to locate activity on a musical timeline, for example Figure 2 (Holck, 2002). However this presentation is less meaningful for those who do not read music.

In applying the rules of conversation analysis to musical utterances, our notation needs to be comprehensible to those who are used to working with transcripts, whilst being able to capture the interactionally relevant aspects of music production. A system has been devised with two main aims: to make representation of the music understandable for both musicians and non-musicians, and make it possible for a written transcript to convey the full interaction whether utterances are verbal or musical. The starting point was established notation for conversational analysis as shown in Figure 1. This was adapted to produce notation for musical sounds as shown in Figure 3.


Figure 2: Extract from a transcript using musical notation. English translation used with permission Holck (2007, pp 33)
\begin{tabular}{|c|c|}
\hline (1.4) & Long single note and duration \\
\hline - - _ - \({ }^{(2.3)}\) & Short notes in a musical phrase (duration optional) \\
\hline \(\uparrow_{\sim}\) _ _ (1.2) & Rising passage of notes \\
\hline \(\downarrow_{\sim}\) _ _ (2.1) & Falling passage of notes \\
\hline , , , (1.2) & Breath in and duration \\
\hline _//__(1.5) & Interrupted long note \\
\hline [ & Beginning of overlap (verbal or musical) \\
\hline ] & End of overlap \\
\hline \{first octave\} & Additional information for music notation \\
\hline ((smiles)) & Features other than verbalisation e.g.gestures \\
\hline
\end{tabular}

Figure 3: Notation devised to represent musical sounds.

\section*{Applying Musical Notation to Data}

We have observed clarinet lessons with four students and two tutors at two music schools in London which prepare young adults for the study of performance at an undergraduate level. For explanatory purposes we focus on the turn-like organisation of musical contributions in one lesson however, we have found that these phenomena recur across the sample. The student is a male clarinet player studying for his grade eight exam. Throughout, we will indicate the areas of interest which are to be the subject of further work using the broader dataset which has been collected.

\section*{Non-verbal interaction as a scaffold for musical utterances}

The student is being tested on playing scales from memory. The tutor holds a small book containing the syllabus in her hands, at chest height. The student is holding his clarinet in front of his body with both hands and they are facing each other. The tutor has asked him to play a three octave scale. There are eight notes in an octave, the last note and the first note are the same, the last note being an octave higher in pitch. To simplify our transcript, we can represent this as numbers rather than notes \(12345678(1)\), where 8 represents the top of the scale. To play a scale over several octaves you would play seven notes, then start the next octave so a three octave scale could be represented as 1234567123

456712345678 . The instances of octaves are shown separately on the transcript below for clarity. The student (S) is expected to play all three octaves ascending and then descending in one smooth phrase. He commences the ascending scale, flicking his eyes up to the tutor (T) occasionally, but makes an unpleasant squeak at the start of note 7 of the third octave. He elongates this note and then briefly stops playing (Figure 4).
```

T: OK great (0.33) Ah let's have (0.44) F sharp major
((T moves head up from book to look at student))
(0.82) tongued
((S brings clarinet up to mouth))
so by tongued I mean legato [tongued you know
[yeah
((T looks back down to her book))
S: ' ' '(0.82)
=\uparrow1_2_3_4_[5]_6]_7_ {first octave}
[S glances briefly up at tutor]
=\uparrow1_[2]_3_4_5_6_7_ {second octave}
[S glances briefly up at tutor]
=\uparrow 1_2_3_4_5_6_ {third octave}
={squeak} 7

```
\(\qquad\)
``` (0.75)
    pause (0.16)
    =\uparrow [8__] (0.75)
    [S looks up at tutor and raises eyebrows]
S: the[re? ((speaks with clarinet still in mouth))
    [because you want E sharp here
```

Figure 4: The student seeks guidance musically and verbally.

The tutor does not change her position or gaze, or seek to interrupt in this brief pause, and the student continues with the next note in the scale (note 8) again elongating it, looking up at the tutor as he plays and raising his eyebrows. This is a possible parallel with Levelt's word-completion hypothesis (Levelt, 1983) in which it is suggested that speakers have a tendency to complete words after detection of trouble. Musicians may exhibit a similar tendency to finish a musical idea, although this does not always happen, since examples of the student restarting mid-phrase after a mistake have also been found. The circumstances around whether the student finishes a phrase or restarts mid-phrase is a potential area of further investigation. The tutor still does not visibly react. Rather than continue with the scale (which would now be the three octave descent), the student stops and verbally seeks guidance 'there?' with the clarinet still in his mouth (Figure 5). Finally she responds, verbally indicating that the error was with the note that played by not referring to the squeak.

The tutor puts her book down on the music stand, turns away and picks up her clarinet from its resting place on the piano. She then twists her body back to look up and meet the student's gaze as he retries the top of the scale tentatively (Figure 6). They continue to make frequent brief direct eye contact and the tutor nods and encourages him verbally 'yeah' as he starts the descent. He continues down the scale, picking


Figure 5: The student seeks verbal guidance "There?"
up speed to finish with a confident long final note. The tutor continues to hold her clarinet close to her mouth but does not play, letting him finish before giving verbal feedback on the rest of the scale.

```
S: _ _ _ _ `7_8_ {resumes third ascending octave}
S: =
    tentatively}
    [yeah
    [((T nods several times at S))
T holds clarinet in front of her, gaze resting on S
playing, mouthpiece remains several inches below her
mouth
S: +4_3_2_1 {continues third octave, resumes tempo
    and fluidity}
    =\downarrow7_6_5_4_3_2_1{second octave}
    =\downarrow7_6_5_4_3_2_{first octave} =1____{last note}
```

Figure 6: The student continues after encouragement

In this short vignette we see the student's use of gaze to seek feedback whilst playing. When this is unacknowledged, it is escalated to a verbal utterance. When gaze is used to seek a response again but encouragement is received, the student continues to play without seeking feedback verbally. We then see how the student uses tempo to express confidence in the descent, holding the turn long enough to complete the scale. Speeding up talk has been shown to be a way to hold on to a turn (Button, 1993).

## Delayed interruption

We have previously noted that in a music lesson, where a verbal request by the tutor is usually followed by a musical response, it is reasonable to assume that this musical phrase is analogous to a conversational turn, and we should therefore be able to see the characteristics of turn management (Duffy \& Healey, 2012). The tutor will often interrupt the student's playing once a problem has been detected. However rather than stop them as soon as the error has occurred, the end of a musical phrase is preferred by the tutor as a TRP to take the turn, even if they have detected the problem earlier. During the short period of time between detection and interruption, the tutor's non-verbal behaviour reveals their intention to interrupt, such as moving in closer to the music, raising their
arms from their listening pose or picking up a pencil (to write an instruction on the score). If the tutor decides to demonstrate the fragment, they start to pick up their instrument, or bring it closer to a playing position, whilst the student finishes the phrase. Student and tutor will not deliberately play at the same time during this part of the lesson, however we will see later that brief overlap does occur.

We will now examine this in more detail using the notation devised. In this extract, student and tutor have been working on an exam piece together for several weeks (Clarinet Sonata in Eb Major Op. 167 Mvmt IV Molto Allegro by Saint-Saens). It includes passages of ascending and descending scales and arpeggios which are challenging to play fluidly. The tutor suggests that they pick up where they left off the previous week. The student starts by playing the phrase shown in Figure 7 however he plays a wrong note from a different scale which changes the tonality of the passage. The tutor indicates that she has noticed the error by adjusting her gaze and listening position but does not interrupt the student yet. He restarts mid-phrase, from just before the error (Figure 7). This may be for the benefit of the tutor, who has to solve a continuation problem, i.e., how to relate the repair to the original utterance (Levelt, 1983) or for his own benefit in ease of correction. The choice of where to start a musical self-repair is another area worthy of further investigation.


Figure 7: Bars 118-120 Clarinet Sonata in Eb Major Mvmt IV Publisher Durand, 1921. Plate D. \& F. 10,063, Paris.

The tutor moves closer to the score with her pencil and as the student reaches the long note at the end of the phrase, she talks over it and moves the pencil towards the score (Figure 8). We see that rather than interrupt mid-phrase, she lets the student attempt self-repair, only taking over the turn at the TRP presented by the long note at the end of the phrase.
: ${ }^{\uparrow}-----{ }^{\downarrow}--------{ }^{*}$ \{*error, wrong note\} T turns head up from music to the student, moves in from her listening position but does not say anything.
$S$ blows into the clarinet and restarts from just before the error, playing more slowly.
S: [\downarrow- -- - - - `_] [ [\downarrow- - - - ] S: [\downarrow- -- - - - `_] [ [\downarrow- - - - ]
[T pícks`\mp@code{up pencil] [T moves into score with pencil]}     [T pícks`\mp@code{up pencil] [T moves into score with pencil]}
//_(1.37) {final note of phrase}
//_(1.37) {final note of phrase}
T: //ok so make sure that it's A major not E major.
T: //ok so make sure that it's A major not E major.

Figure 8: The tutor delays interruption

## Using 'readiness to play' to signal the intent to interrupt

Here we will see how the tutor uses the position of her clarinet to indicate her 'readiness to play', effectively taking the floor through playing. The student has become stuck on a particular section and the tutor has decided that they will practice it together. She picks up her clarinet and demonstrates the phrase in full at a steady pace, as they both read from the student's score (Figure 9). The tutor bounces her clarinet bell to emphasis the rhythm of the phrase and briefly glances at the student, who is still looking at the score. He starts to nod in time with her clarinet bell, signalling his attention to the demonstration in response to the tutor's checking glance.

```
T plays the phrase in full, at steady pace.
T: }\mp@subsup{\uparrow}{\uparrow}{+
    ((*)
T: Just ((T points to the score))
    //that section (0.2) OK?
S://There?
S brings arm up and points to the score
S: What, just there?=
T: =((nods)) Just do what I've just played
```

Figure 9: The tutor plays the phrase to be worked on in full

The student then tries the phrase himself as directed. He makes several undesirable squeaking sounds in place of the expected notes (Figure10) and as he exhales loudly, the tutor turns towards him.

```
S: ' ' \(\quad(0.6)^{\uparrow}---\{\) squeak \} \{squeak\}
: ((exhales audibiy))
```



Figure 10: The student makes undesirable squeaking sounds

She brings her left hand up to clasp the barrel of her instrument, which she had been holding in her right hand. By moving from holding her clarinet in just one hand, to both hands, the tutor is signalling an intention to interrupt through demonstration. However the student restarts the phrase and the tutor brings her right hand back down to her side. Now she is holding her clarinet in just one hand again, using the change in state of readiness to play to signal that he should keep the turn and continue playing.
This time he manages a substantial part of the phrase without any mistakes however he plays an incorrect note on starting the final scale descent. As he restarts and makes a further mistake, the tutor moves her clarinet back to playing position in three stages, first by bringing her left hand back to the instrument body, then raising the clarinet vertically in her hands so that they are closer the the keys, and then bringing the mouthpiece towards her mouth. Each time she brings the
clarinet closer to the playing position, she is escalating the likelihood of interruption (Figure 11). The student lowers his mouthpiece as soon as he finishes his last note, even though on this final attempt he has played the last part of the phrase correctly. The tutor takes over straight away and plays the end of the phrase that the he has been struggling with.

```
((T moves hands on clarinet closer to keys))
S: &_ _ [_]* {*last two notes are incorrect}
        [T raises clarinet closer to lips]
S: \downarrow_ _ _ **
{**last two notes incorrect, last note distorted}
S: \downarrow_ - - - [___]
        [T places clarinet in mouth]
```

Figure 11: Escalation of visible intention to interrupt

## Self-repair and overlap

Continuing straight on from the last extract we now see the pair negotiate moving into a period of exchanging musical turns. The student takes the floor again in order to try the phrase himself. They now enter a period of alternating turns with brief overlaps, repeating the same phrase again and again, small corrections to the phrasing and rhythm being made by the tutor on each round (Figure 12). The tutor keeps her clarinet in her mouth during the student's turns, resting on her bottom lip, only lifting her top lip to take in breath between playing. The student briefly lowers his clarinet from his mouth during the tutor's turns, indication perhaps that the tutor is going to continue until the phrase is right whilst the student would like to move on.


Figure 12: Student and tutor alternate the same phrase

Then the tutor elongates the phrase, adding in more notes at the beginning and the student copies her, again overlapping with her last note. He manages to play this longer extract, but the tutor repeats it, implying there is still something that needs correction. On his second attempt, the student stumbles after just two notes. He then exhales noisily and continues, restarting twice at mistakes. The tutor keeps her clarinet in her mouth and small head movements indicate that she is preparing to take the floor from the student, however she does let him get to the end of the phrase (Figure 13).

The tutor allows a brief pause before playing the full phrase at a much faster tempo than previously. The student tries to match her but soon stumbles, however the tutor has removed her clarinet from her mouth now and she lets him self repair (Figure 14).


Figure 13: The tutor elongates the phrase

```
T: \
    \uparrow------------------- 
    ((pause (1.33) restarts from error))
S: \downarrow_ _ _ _ _ _ _ _ _ _ _ \uparrow _ _ \downarrow_ _ _
```

Figure 14: The tutor allows the student to self-repair

## A Final Note

It is worth noting that some findings around turn taking mechanisms cannot be so easily applied to the production of music. The phrases played by a clarinetist are defined by the structure of the music - both in terms of notes and phrasing, complete bars, tempo and breathing. Hence some of the recognised ways to manage and hold turns are not available to the player. With a woodwind instrument, holding a turn through an in-breath (Button, 1993) is not always possible since sound production relies on breathing out. It is also difficult to use speed to hold a turn when control of tempo (either consistent speed or dictated rallentando or accelerando) is a goal. Whilst we saw this device used during the part of the lesson devoted to scales, it is less likely to be employed when the student is performing a piece of music for the tutor.

## Acknowledgments

The authors would like to thank the students and tutors who participated in these studies. This work is supported by the Media and Arts Technology programme, EPSRC Doctoral Training Centre EP/G03723X/1.

## References

Bavelas, J., Chovil, N., Lawrie, D., \& Wade, A. (1992). Interactive Gestures. Discourse Processes, 15, 469-489.
Besson, M., Chobert, J., \& Marie, C. (2011, September). Language and Music in the Musician Brain. Language and Linguistics Compass, 5(9), 617-634.
Button, G. (1993). Answers as interactional products: two sequential practices used in job interviews. In P. Drew \& J. Heritage (Eds.), Talk at work: Interaction in institutional settings (pp. 212-234). Cambridge: Cambridge University Press.
Cassell, J., \& Thorisson, K. R. (1999, May). The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. Applied Artificial Intelligence, 13(4-5), 519-538.
Clark, H. H., \& Krych, M. a. (2004, January). Speaking while monitoring addressees for understanding. Journal of Memory and Language, 50(1), 62-81.

Davidson, J. W., \& Good, J. M. M. (2002, October). Social and Musical Co-Ordination between Members of a String Quartet: An Exploratory Study. Psychology of Music, 30(2), 186-201.
Duffy, S., \& Healey, P. G. T. (2012). Spatial Co-ordination in Music Tuition. In N. Miyake, D. Peebles, \& R. P. Cooper (Eds.), Proceedings of the 34th annual conference of the cognitive science society (pp. 1512-1517). Sapporo: Cognitive Science Society.
Engle, R. A. (1998). Not Channels but Composite Signals. In Proceedings of the twentieth annual conference of the cognitive science society (pp. 321-326). University of Wisconsin-Madison: Routledge.
Feld, S., \& Fox, A. a. (1994, October). Music and Language. Annual Review of Anthropology, 23(1), 25-53.
Garfinkel, H. (1964). Studies of the routine grounds of everyday activities. Social Problems, 11(3), 225-250.
Healey, P., Leach, J., \& Bryan-Kinns, N. (2005). Inter-play: Understanding group music improvisation as a form of everyday interaction. In Proceedings of less is more, simple computing in an age of complexity. Cambridge: Microsoft Research.
Holck, U. (2002). 'Kommunikalsk' samspil i musikterapi. Phd., Aalborg University.
Holck, U. (2007). An Ethnographic Descriptive Approach to Video Microanalysis. In T. Wosch \& T. Wigram (Eds.), Microanalysis in music therapy: Methods, techniques and applications for clinicians, researchers, educators and students (chap. 2). London: Jessica Kingsley Publishers.
Kendon, A. (1990). Spatial organization in social encounters: the F-formation system. In Conducting interaction: Patterns of behavior in focused encounters (pp. 209-237). Cambridge: Cambridge Univ Pr.
Levelt, W. J. (1983, July). Monitoring and self-repair in speech. Cognition, 14(1), 41-104.
Moran, N. (2011, May). Music, bodies and relationships: An ethnographic contribution to embodied cognition studies. Psychology of Music.
Sacks, H., Schegloff, E., \& Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. Language, 50(4), 696-735.
Schegloff, E., Jefferson, G., \& Sacks, H. (1977). The Preference for Self-Correction in the Organization of Repair in Conversation. Language, 53(2), 361-382.
Schegloff, E. A. (2007). Sequence Organization in Interaction: Volume 1: A Primer in Conversation Analysis. Cambridge: Cambridge University Press.
Zatorre, R. J., Belin, P., \& Penhune, V. B. (2002, January). Structure and function of auditory cortex: music and speech. Trends in cognitive sciences, 6(1), 37-46.

# Object-based Saliency as a Predictor of Attention in Visual Tasks 

Michal Dziemianko (m.dziemianko@sms.ed.ac.uk) Alasdair Clarke (a.clarke@ed.ac.uk) Frank Keller (keller@inf.ed.ac.uk)<br>Institute for Language, Cognition and Computation<br>School of Informatics, University of Edinburgh<br>10 Crichton Street, Edinburgh EH8 9AB, UK


#### Abstract

The top-down guidance of visual attention is an important factor allowing humans to effectively process incoming visual information. Our understanding of the processes governing attention is not complete, with growing evidence for attention selection based on cognitive relevance. In this paper, we investigate whether models for salient object detection from computer vision can be used to predict attentional shifts in visual tasks. Our results show that the object-based interpretation of saliency provided by these models is a substantially better predictor of fixation locations than traditional pixel-based saliency.


Keywords: eye-tacking; saliency; visual attention.

## Introduction

Virtually every human activity occurs within a visual context and many tasks require visual attention in order of be successfully accomplished (Land \& Hayhoe, 2001). When processing a visual scene, humans have to localize objects, identify them, and establish their spatial relations. The eyemovements involved in this process provide important information about the cognitive processes that unfold during scene comprehension.

A number of models have been proposed to predict eyemovements during scene processing and they can be broadly divided into two categories. The first category consists of bottom-up models that exploit low-level visual features to predict areas likely to be fixated. A number of studies have shown that certain features and their statistical unexpectedness attract human attention (e.g., Bruce \& Tsotsos, 2006). Moreover, low-level features are believed to contribute to the selection of fixated areas, especially when the visual input does not provide useful high-level information (Peters et al., 2005). These experimental results are captured by models that detect salient areas in the visual input and use them to predict attention. The best-known example is the model of Itti et al. (1998), which builds a pixel-based saliency map using color, orientation, and scale filters inspired by neurobiological results.

The second group of models assumes that top-down supervision of attention contributes to the selection of fixation targets (e.g., Torralba et al., 2006). Various types of such supervision have been observed experimentally. Humans show the ability to learn general statistics of the appearance, position, size, spatial arrangement of objects, and their relationships (e.g., Zelinsky, 2008). They also exploit visual memory during scene comprehension tasks (e.g., Shore \& Klein, 2000). Moreover, studies such as those of Chun \& Jiang
(1998) show that participants benefit from learning spatial arrangement of the objects in consecutive searches. Theoretically, such results can be accommodated by the Cognitive Relevance Framework (Henderson et al., 2009), which assumes that attention is allocated to locations that are cognitively relevant for the task performed.

Cognitive relevance predicts that objects should have a privileged status in visual processing, which is in line with experimental evidence suggesting that the allocation of attention is object-based rather than pixel-based. For example, Henderson et al. (2007) argue that saliency does not account for fixated areas in visual search, while Nuthmann \& Henderson (2010) show that the preferred fixation point or landing position is the center of an object: fixations are distributed normally around an object's center of mass, where the spread might be explained by oculomotor errors. Consistent with this, Einhauser et al. (2008) show that the position of objects is a better predictor of fixations than early saliency in tasks such as artistic evaluation, analysis of content, and search.

An alternative view on saliency comes from the computer vision literature, which deals with task of salient object detection: the objects that are perceived by humans as visually most interesting have to be separated from the background. Typically this involves image segmentation and the calculation of visual features in order to select pixels belonging to salient objects. In this context, saliency is a feature of an object, rather than an early pixel-based attractor of attention.

In this paper, we investigate the extent to which methods proposed for salient object detection can be applied to the prediction of fixations. We are not concerned with the prediction of salient image patches, but rather with the selection of objects that are likely to be fixated. This approach allows us to develop computational models of attentional selection based on cognitive relevance defined over objects (Henderson et al., 2007, 2009). We compare the performance of this approach to traditional models which predict fixation locations using pixel-based saliency maps.

## Background

As discussed above, there is experimental evidence for the object-based allocation of attention. Additionally, some objects seem to inherently attract more attention than others, a fact that has been conceptualized using proto-objects: pre-recognition entities that draw attention (Rensink, 2000). Proto-objects have been incorporated into saliency-based models (Walther \& Koch, 2006) and have also been applied


Figure 1: Example of proto-objects extracted from an image using the model of Walther \& Koch (2006). From left to right: original image, saliency map computed according to Itti et al. (1998), proto-object mask. The salient patches, and hence the proto-objects, do not necessarily correspond to the real objects in the scene.
in robotics to create attentional systems for virtual and physical agents (see e.g., Yu et al., 2010). These models perform image segmentation to identify proto-objects: the image is divided into a collection of regions that correspond to areas enclosed by constant, high saliency values. Figure 1 shows an example of such proto-objects extracted from an image using the model of Walther \& Koch (2006).

While Walther and Koch's model is conceptually interesting, its cognitive status is questionable, as there is evidence that it does not predict fixation locations well (Nuthmann \& Henderson, 2010). Alternative models of attention selection based on objects rather than proto-objects have been proposed in computer vision. For example, the work of Liu et al. (2011) focuses on detecting objects annotated by people as salient. These models use machine learning techniques to compute which arrangements of visual features such as center-surround histograms, orientation, scale are perceived as salient. However, in a computer vision context, attentional selection is regarded merely as an engineering task: the aim is to identify areas matching pre-annotated training data, rather than to gain a greater understanding of human behavior.

## Models

We implemented and evaluated three models for salient object detection. Throughout our work we assume that the images are fully annotated with object boundaries, therefore the problem of segmentation and separation of objects from the background does not need to be solved within the models. This assumption makes it possible to evaluate object-based saliency models separately from image segmentation algorithms, which can vary widely in their performance.

## A. Conversion of Standard Saliency

Standard, pixel-based saliency is the baseline against which we evaluate object-based models. The baseline model we use is Torralba et al.'s (2006), which approximates saliency as the probability of the local images feature $L$ in a given location based on the global distribution of these features:

$$
\begin{equation*}
p(L) \propto e^{-\frac{1}{2}\left[(L-\mu)^{T} \Sigma^{-1}(L-\mu)\right]} \tag{1}
\end{equation*}
$$

Here, $\mu$ is the mean vector and $\Sigma$ the covariance matrix of the Gaussian distribution of local features estimated over the
currently processed image. The local features are computed as a set of Steerable pyramid responses computed over three color channels for six orientations and four scales, totaling 72 values at each position.

Based Torralba et al.'s model, we can define a group of models which convert pixel-based saliency values to objectbased salience scores. Such a conversion can be performed by computing functions such as the maximum, mean, median, or mode of the pixels that make up an object. Examples for the use of this method exist in the literature (e.g., Spain \& Perona, 2011), with maximum and mean being common. These models will be referred to as converted in this paper.

## B. Liu et al. Features

Liu et al. (2011) describes a system for salient object detection based on conditional random fields, which simultaneously segments pixels into areas corresponding to objects and computes the pixel's salience. The model is based on three feature channels - contrast, center-surround histograms and spatial color - which are described below. The salience of a pixel is defined to be the a weighted sum of these three feature maps, while the salience of an object is defined as the sum over all pixels within the object's boundary. The full specification of our implementation of Liu's model can be found in Dziemianko (2013). Examples of the feature channels are given in Figure 3.

Multiscale Contrast Contrast is one of the most commonly used features in saliency models and is implemented over a multiscale Gaussian pyramid. In each layer of the pyramid, the contrast at pixel $(x, y)$ is defined to be the mean squared difference of the intensity of pixel at $(x, y)$ and its adjacent neighbors. The multiscale contrast for $I(x, y)$ is then taken to be the sum over the layers of the corresponding pyramid. This has the effect of approximating human receptive field by highlighting high-contrast boundaries while omitting homogeneous regions within objects.

Center-Surround Histograms One of the weaknesses of previous measures of visual salience is that, due to their reliance on high-contrast center-surround features, they tend to emphasis the boundaries of objects while giving very low scores to pixels within an object's boundary (see Figure 2). To tackle this issue, Liu et al. (2011) propose to use region-based features in addition to the center-surrounds described above. These are computed by considering the histogram of colors within an object's bounding box, and comparing it with a surrounding region of equal area (see Figure 2). The $\chi^{2}$ metric is used to measure the distance between histograms and the full details on how these regions are constructed can be found in Liu et al. (2011).
Color Spatial Distribution The last feature used by Liu et al. (2011) is the spatial color distribution, motivated by the observation that salient objects are less likely to contain colors that are distributed widely throughout the image. A simple method for quantifying this is to compute the spatial vari-


Figure 2: An example of high saliency values being assigned to object boundaries due to its reliance on high-contrast features. From left to right: original image, traditional saliency, an object (red) and its surrounding area (green).


Figure 3: Examples of features from Liu et al. (2011). From left to right: original image, multiscale contrast, centersurround histogram, color spatial distribution (image from Liu et al. (2011) with modifications).
ance of color. This involves representing the distribution of colors contained in the image by a Gaussian mixture model. We then carry out soft assignment: for each of these Gaussians, $c$, we then calculate $p(c \mid I(x, y))$, the probability of assigning pixel $I(x, y)$ to the Gaussian $\mathcal{N}\left(\mu_{c}, \Sigma_{c}\right)$. Using these we can calculate the weighted mean and variance for each color component along the horizontal axis:

$$
\begin{align*}
M_{h}(c) & =\frac{1}{|X|_{c}} \sum_{x} p(c \mid I(x, y)) \cdot x  \tag{2}\\
V_{h}(c) & =\frac{1}{|X|_{c}} \sum_{x} p(c \mid I(x, y)) \cdot\left|x-M_{h}(c)\right|^{2} \tag{3}
\end{align*}
$$

where $|X|_{c}=\sum_{x} p\left(c \mid I_{x}\right)$. The vertical spatial variance, $V_{v}$, is computed in the same way, and $V(c)$, the spatial variance of each color component, is then simply defined as:

$$
\begin{equation*}
V(c)=V_{h}(c)+V_{v}(c) \tag{4}
\end{equation*}
$$

Finally, the feature function $f_{s}(x, y)$ is defined as:

$$
\begin{equation*}
f_{s}(x, y) \propto \sum_{c} p(c \mid I(x, y)) \cdot(1-V(c)) \tag{5}
\end{equation*}
$$

## C. Color-component Histograms

In addition to the models described above, we have implemented our own model based on a simplified factored shapes and appearances representation (Eslami \& Williams, 2011). This model shares some characteristics with the spatial color distribution described above, as it assumes that the pixels corresponding to each object have been generated by a number of Gaussians in a feature space (we found Lab-space to be


Figure 4: Examples of scenes used in the visual counting experiment. Targets on the images on the left and in center are man, while for the image on the right it is goggle.
the most effective). However, it performs a comparison of histograms of color cluster assignments within the object and its surrounding area.

In the first phase, the means $\mu$ and covariances $\Sigma$ of these Gaussians are extracted by fitting a Gaussian mixture model (GMM) with $W$ components over all pixels in the image. Similar to Eslami \& Williams (2011), we use $W=15$ Gaussians. At this stage object boundaries and locations are ignored. In the subsequent step, pixels are clustered into $W$ clusters according to the associated GMM components by selecting a component, $\hat{w}$, that maximizes the probability of a pixel being drawn from the Gaussian distribution. The final step of the first phase consists of computing global histograms $H$ of the pixel assignments $\hat{w}$ representing the proportion of pixels belonging to each cluster.

The saliency scores are computed in the second phase. At this stage, the model assumes that the image is fully annotated (i.e., boundaries for each object within the scene are provided). For each object in the scene, we calculate the histogram of pixel assignments over the pixels within the object's boundary. We then define an interestingness value for each object as the Kullback-Leibler (KL) divergence between the local (object) pixel distribution and the global distribution $H$. Intuitively, I represents how different the object is from its surroundings and thus interesting.

## Evaluation

## Method

We evaluate the performance of the models discussed on eyetracking data collected in a visual counting and an object naming task. In the visual counting task, 25 participants were asked to count the number of occurrences of a cued target object, which was either animate (e.g., man) or inanimate (e.g., goggle). The data set consisted of 72 fully objectannotated photo-realistic scenes (both indoor and outdoor), with total of 1809 polygons with mean of $25.12 \pm 11$ and a median of 25 polygons per image, containing zero to three instances of the target object. The data was collected using an Eyelink II head-mounted eye-tracker with a sampling rate of 500 Hz . The images were displayed with a resolution of $1024 \times 768$ pixels, subtending a visual field of approximately $34 \times 30$ degrees. The data set consists of 54,029 fixations. Figure 4 presents examples of scenes used in the experiment.

The object naming dataset (Clarke et al., under revision) contains data collected during an object naming experiment.


Figure 5: Examples of stimuli used in the object naming experiment. Typical responses are: cars, crossing, person for the left, bench, man for the center, and barbecue, charcoal, chimney for the right image.

The stimuli consists 132 fully object-annotated images with a total of 2,858 polygons with mean of $14.2 \pm 5$ and a median of 26 polygons per image. The images were presented to 24 participants after the task was explained using written instructions. Before each trial, participants were asked to fixate a central cross. The image was then displayed for 5000 ms , followed by a beep, after which the participants named objects present in the scene. The image was displayed until the participant finished the trial. Image presentation and apparatus were the same as in the visual counting data set. A total of 2,904 usable trials were collected, resulting in 88,371 fixations. Examples of images used as stimuli are shown in Figure 5.

## Analysis

As well as the models described above, we test two baselines that do not use saliency in any form. The first one weights objects by their Euclidean distance from the center of the image, normalized by object area. This approach is inspired by experimental evidence of center bias in scene viewing (e.g. Tatler, 2007), and will be referred to as center bias.

Secondly, based on the findings of Nuthmann \& Henderson (2010), we also include a baseline that predicts fixations by selecting object centers. In this case, a map is built as a sum of Gaussians centered on the bounding boxes of the object in the image. The covariances of the Gaussians are dependent on object's size, with a factor fitted using 10-fold cross-validation to avoid overfitting the datasets. This baseline is referred to as object overlay.

In the Results and Discussion section below, we show how the different models perform by using receiver operating characteristic (ROC) plots, which indicate the sensitivity (i.e., true positive rate vs. false positive rate) of a classifier as its discrimination threshold varies. Moreover, in order to statistically compare model performance, we calculate the area under the ROC curve (AUC) of each participant. The AUC measures the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. ${ }^{1}$ We submit the AUC means to an ANOVA analysis to compare the performance of the different models pairwise, e.g., saliency against converted (mean). For standard pixelbased saliency, the ROC curve is constructed by thresholding

[^255]| Model | Obj. counting | Obj. naming |
| :--- | :---: | :---: |
| Saliency | 61.66 | 55.87 |
| Object overlay | 63.60 | 59.78 |
| Center bias | 68.02 | 69.17 |
| Converted (max) | 55.27 | 64.66 |
| Converted (mean) | 70.44 | 68.65 |
| Liu et al. 2011 features | 66.67 | 67.42 |
| Color-component hist. | 66.73 | 67.40 |

Table 1: Estimated percentage areas under the ROC curves presented in Figure 6.
the saliency values to select the desired proportion of pixels. The ROC plots for object-based models can not be constructed this method as it would not ensure that entire objects are selected. Instead, an increasing number of objects with the highest saliency values is iteratively selected, and their total area is plotted in the ROC curve. The ROC curves constructed this way are incomplete, representing only selection of up to about $50 \%$ of the image area. Constructing ROC plot for larger selections would result in significant discontinuities due to the fact of all small objects being already selected and essentially only large objects corresponding to surfaces such as floor, sky, or wall being left.

## Results and Discussion

The results are presented in Figure 6. The ROC curves show that selection based on object overlay is better than saliency for thresholds smaller than $40 \%$. Object-based saliency models in turn outperform object overlay. Center bias turns out to be a very competitive baseline, which is only matched by converted (mean).

An analysis of the areas under the ROCs, summarized in Table 1, confirm these observations. The ANOVAs reveal that for both datasets, object position overlay is significantly better than saliency with $F(1,24)=9.27, p<0.005$ for object counting, and $F(1,23)=9,84, p<0.005$ for object naming.

The calculation of area under ROC curve for object-based models is not trivial due to the discontinuity of the plot. We estimated the AUC by interpolating the missing values. ${ }^{2}$ The analysis of the interpolated curves shows that for both datasets, object-based selection is superior to traditional saliency, and to object overlay. These differences are statistically significant, for example converted (mean) is better than saliency with $F(1,24)=165.60, p<0.001$ for counting and $F(1,23)=279.30, p<0.001$ for naming; for color histogram the values are $F(1,24)=34.67, p<0.001$ and $F(1,24)=227.40, p<0.001$ respectively.

The pattern for Converted (max) is more complicated. On the naming data, it is significantly better than saliency $(F(1,24)=132.10, p<0.001)$, but not as good as any of the other methods. On the counting data, it is significantly weaker than standard saliency $(F(1,23)=245.70, p<$

[^256]

Figure 6: Performance of object-based selection of fixation locations on the Visual Count (top) and Object Naming (bottom) datasets. Note that traditional saliency and object-based models cannot be compared directly due to differences in the selection method, see text for details.
0.001 ), operating around chance level. This can be explained by the fact that saliency is sensitive to high contrast edges, usually corresponding to object boundaries. As such, the highest saliency values corresponding to the object might not
fall within the object, but rather belong to its neighbors.
A surprising results is that object-based selection does not outperform selection based on center bias. However, closer investigation of the object rankings based on center bias and

Converted (mean) reveals that the average correlation coefficient between the respective rankings is only 0.50 for the naming and 0.43 for the counting data. This indicates that different sets of objects are selected by the two model for a given threshold, accounting for different subsets of fixations. A combined model would be a promising next step.

## Conclusion

In this paper, we discussed the issue of objectness and its relation to the allocation of visual attention. We demonstrated that it is possible to develop object-based version of saliency. Object-based saliency is not calculated as a value for each of the image pixels (or coordinates), but rather over an area within the boundaries of an object. In this approach, saliency is treated as a feature of an object, similar to other features such as position. This approach is compatible with theories assuming an object-based allocation of attention, such as the Cognitive Relevance Framework (Henderson et al., 2009).

The evaluation we presented used an object counting and an object naming data set. In spite of both of these tasks being object-centric by definition, we believe that our results generalize to other experimental tasks. Such tasks are often either object-centric as well (e.g., visual search), or evidence exists that attentional access is object-based even if the task defined in terms of objects (e.g., in aesthetic judgment or interestingness judgment, see Nuthmann \& Henderson 2010; Einhauser et al. 2008). Indeed it was shown that visual attention is object-based during everyday interaction with the surrounding world (Land et al., 1999). Finally, it has been suggested that free viewing does not mean that viewers look at images without any task constraints, but rather with constraints to which experimenters do not have access (see Tatler et al., 2011, for further discussion).

Even though the intuition that salience is a property of objects has been utilized before, we are not aware of any extensive experimental study aiming to investigate whether objectbased saliency and techniques used to detect salient objects in computer vision can reliably predict human fixations. We showed that the prediction of fixations based on objects and their visual features is not only possible, but superior to standard saliency. However, using the maximum value of saliency within an object was not confirmed as a reliable predictor of whether object is going to be fixated, which is a important result considering the popularity of this feature in previous modeling studies.

## Acknowledgments

The support of the European Research Council under award number 203427 "Synchronous Linguistic and Visual Processing" is gratefully acknowledged.

## References

Bruce, N., \& Tsotsos, J. (2006). Saliency based on information maximization. In Advances in Neural Information Processing Systems 18, (pp. 155-162). Cambridge, MA: MIT Press.

Chun, M., \& Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. Cognitive Psychology, 36, 28-71.
Clarke, A., Coco, M., \& Keller, F. (under revision). The impact of attentional, linguistic and visual features during object naming. In G. Zielinsky, T. Berg, \& M. Pomplun (eds.), Frontiers in Perception Science: Research Topic on Scene Understanding: Behavioural and computational perspectives.
Dziemianko, M. (2013). Modelling Eye Movements and Visual Attention in Synchronous Visual and Linguistic Processing. Ph.D. thesis, University of Edinburgh.
Einhauser, W., Spain, M., \& Perona, P. (2008). Objects predict fixations better than early saliency. Journal of Vision, 8, 1-26.
Eslami, S., \& Williams, C. (2011). Factored shapes and appearances for parts-based object understanding. In Proceedings of the British Machine Vision Conference, (pp. 18.1-18.12). BMVA Press.
Henderson, J., Brockmole, J., \& Castelhano, M. (2007). Visual saliency does not account for eye-movements during visual search in real-world scenes. Eye movements research: insights into mind and brain.
Henderson, J., Malcolm, G., \& Schandl, C. (2009). Searching in the dark: cognitive relevance drives attention in real-world scenes. Psychonomic Bulletin \& Review, 16, 850-856.
Itti, L., Koch, C., \& Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20, 1254-1259.
Land, M., \& Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? Vision Research, 41, 35593565.

Land, M., Mennie, N., \& Rusted, J. (1999). The roles of vision and eye movements in control of activities of daily living. Perception, 28.

Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., \& Shum, H. (2011). Learning to detect salient objects. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33, 353-367.
Nuthmann, A., \& Henderson, J. M. (2010). Object-based attentional selection in scene viewing. Journal of Vision, 20, 1-19.
Peters, R., Iyer, A., Itti, L., \& Koch, C. (2005). Components of bottom-up gaze allocation in natural images. Vision Research, 45, 2397-2416.
Rensink, R. (2000). Seeing, sensing, and scrutinizing. Vision Research, 10-12, 1469-1487.
Shore, D., \& Klein, R. (2000). On the manifestations of memory in visual search. Spatial Vision, 14, 59-75.
Spain, M., \& Perona, P. (2011). Measuring and predicting object importance. International Journal of Computer Vision, 91, 5976.

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. Journal of Vision, 7, 4.
Tatler, B. W., Hayhoe, M., Land, M., \& Ballard, D. (2011). Eye guidance in natural vision: Reinterpreting salience. Journal of Vision, 11.
Torralba, A., Oliva, A., Castelhano, M., \& Henderson, J. (2006). Contextual guidance of eye movements and attention in realworld scenes: The role of global features on object search. Psychological Review, 113, 766-786.
Walther, D., \& Koch, C. (2006). Modelling attention to salient protoobjects. Neural Networks, 19, 1395-1407.
Yu, Y., Mann, G., \& Gosine, R. (2010). An object-based visual attention model for robotic applications. IEEE Transactions on Systems, Man, and Cybernetics: Cybernetics, 5, 1398-1412.
Zelinsky, G. (2008). A theory of eye movements during target acquisition. Psychological Review, 115, 419-433.

# Verbal and Nonverbal Cues Activate Concepts Differently, at Different Times 

Pierce Edmiston (pedmiston@wisc.edu)<br>Department of Psychology, 1202 W. Johnson Street<br>Madison, WI 53706 USA<br>Gary Lupyan (lupyan@wisc.edu)<br>Department of Psychology, 1202 W. Johnson Street<br>Madison, WI 53706 USA


#### Abstract

Although the word "dog" and an unambiguous barking sound may point to the same concept DOG, verbal labels and nonverbal cues appear to activate conceptual information in systematically different ways (Lupyan \& Thompson-Schill, 2012). Here we investigate these differences in more detail. We replicate the finding that labels activate a more prototypical representation than do sounds, and find that sounds activate exemplars consistent with the source of the sound, such that after hearing a barking sound, people are faster to recognize a dog with an open-mouth than a closed mouth, but critically, only when the sound and picture are presented simultaneously. The results are consistent with perceptual cues indexing their source while labels activating a more decontextualized representation of the target category.


Keywords: categorization, concepts, sounds, recognition, cross-modal effects, language

## Introduction

Most concepts are multimodal and can be activated in a variety of ways (Hoffman \& Ralph, 2013). For example, the concept DOG can be activated by seeing a wagging tail, hearing a bark, or petting its furry coat. However, the concept DOG can also be activated by hearing the word 'dog'-without seeing, hearing, or touching an actual dog. This raises the question of how concepts activated by nonverbal sensory cues compare to those activated by verbal category labels.

In the experiments reported here we compare how verbal and nonverbal cues activate representations of purportedly the same concepts. In particular, we focus on visual aspects of familiar animals and artifacts as cued by natural sounds: auditory events with a distinct source (e.g., cat meowing, chainsaw revving), and how these same concepts are activated by verbal labels (words like "cat" and "chainsaw").

The mechanisms underlying recognition of nonverbal sounds and of speech appear to be quite similar. Recognition of both words and natural sounds varies as a function of familiarity, frequency, and context (Ballas, 1993; Stuart \& Jones, 1995). Perception of both natural sounds and speech is influenced by signal ambiguity and noise in similar ways (Aramaki, Marie, Kronland-Martinet, Ystad, \& Besson, 2010; Gygi, Kidd, \& Watson, 2004). Both labels and natural sounds elicit similar N400 event-related potentials-a coarse index of semantic processing (Cummings et al., 2006; Van Petten \& Rheinfelder, 1995)-
even when the identification of the natural sound is incidental to task demands (Orgs, Lange, Dombrowski, \& Heil, 2008). Functional imaging during similar sequential processing tasks reveals largely overlapping cortical areas recruited in processing labels and natural sounds (Dick et al., 2007). Lastly, patterns of naming deficits in patients with aphasia suggest the labeling of everyday objects and the visual recognition of natural sound sources rely on similar cognitive resources (Goll et al., 2010; Saygin, Dick, Wilson, Dronkers, \& Bates, 2003).

The perception of meaningful nonverbal sounds and of words is thus dependent on many of the same properties and activate largely the same semantic networks. Although it may seem that verbal and nonverbal cues are in important respects equivalent, there are critical differences. One such difference is that natural sounds, unlike labels, have a causal relationship with a specific physical source (Ballas, 1993). Recognizing these relationships requires learning, but the relationship between a referent and its natural sound is not arbitrary. We call these relationships "motivated": that is, they are determined by physics (e.g., thunder) or driven by biology (e.g., large dogs-and agitated dogs-have deeper barks). Auditory perceivers are able to exploit such "motivated" relationships and surmise features of a hidden physical source, such as the size of a barking dog (Taylor, Reby, \& McComb, 2008), the shape of resonating plates (Kunkler-Peck \& Turvey, 2000), or the hardness of percussion mallets (Freed, 1990). The perception of these auditory sources is surprisingly accurate, reflecting the lawful relationships between signals and sources in the environment (Fowler, 1990). Importantly, sounds covary lawfully within as well as between categories. For example, a barking sound informs us not only that its source is a dog, but can inform us of the approximate size of the dog.

In contrast, the relationship between labels and their referents is "unmotivated." By this term we do not simply mean that words are arbitrary, i.e., that "dog" refers to dogs by convention (cf. Hockett, 1966), but that there exists a word "dog" that denotes the entire category of dogs rather than a particular type or instance (dachshund, German shepherd, dog-on the left, dog-far away, etc.). In short, barks index specific occurrences of dogs. Even though we can interpret natural sounds at a more categorical level, the surface properties of a specific bark still indexes a particular dog. Verbal labels, on the other hand, abstract over these specifics. When we say "dog" we can leave all
that information unspecified. On this view, labels may activate concepts in a more categorical way. This prediction has been supported by a variety of findings (Lupyan, 2012). For example, Lupyan \& Thompson-Schill (2012) found that label cues resulted in faster visual processing over equally predictive nonverbal cues. This advantage persisted across a number of cue-to-image delay periods and extended to artificially created objects with novel labels and "natural" sounds, suggesting that labels do not activate conceptual representations faster but differently than nonverbal cues. In our view, labels activate representations that emphasize the differences between categories, and thus play a facilitative role in category learning (Lupyan, Rakison, \& McClelland, 2007). These categorical representations enable faster recognition of category-typical objects (Lupyan \& Swingley, 2012), but blur within-category differences reflected in biased exemplar memory (Lupyan, 2008).

However, what is not clear from these previous results is how "unmotivated" and "motivated" cues differ in activating different instances of purportedly the same concept. If "unmotivated" verbal cues activate more categorical representations, then what do "motivated" nonverbal cues activate? Given the inherent causal link between a natural sound and its particular physical source, we predicted that natural sound cues would lead to faster processing of images depicting the production of the auditory cue. The results ended up being more interesting.


Figure 1: Sample stimuli from Experiment 1. Does hearing the sound of a revving chainsaw activate a representation of a chainsaw in action?

## Experiment 1

Hearing a sound characteristic of an animal or artifact may automatically activate particular instances of that category. Consider the kind of chainsaw one might expect upon hearing a chainsaw sound (Fig. 1). Here, we asked whether verbal and nonverbal cues lead to different expectations about subsequent visual information. In Experiment 1 we investigated if label and natural sound cues influence visual processing differently based on the action depicted in target images. In line with previous research, we predicted that when presented a label cue, participants would respond faster to category-typical images. Conversely, we predicted that when presented a natural sound cue, participants would respond faster to sound-matched images.

## Methods

Participants 14 University of Wisconsin-Madison undergraduates participated for course credit.
Materials Auditory cues were spoken labels and natural sounds for 12 target categories of familiar animals and artifacts used in Lupyan \& Thompson-Schill (2012). ${ }^{1}$ Visual images were 4 color photographs for each category: 2 category-typical images and 2 sound-producing images. The images were normed, ensuring unambiguous identification. In addition, participants in a separate image rating study evaluated each picture on one of two dimensions (category typicality and sound match) using a 5-point Likert scale. For category typicality, participants viewed e.g., a dog, and were asked: "How typical is this dog of dogs in general?" For sound match ratings, participants listened to e.g., a bird chirping, saw a picture of a bird, and were asked: "How well does that sound go with this picture?" Each participant performed either category-typicality or sound-matching judgments. As expected, the canonical images were rated higher on category typicality $(M=4.57)$ than on sound match ( $M=3.49$ ), while sound-producing images were rated higher on sound match ( $M=4.37$ ) than on category typicality $(M=4.05)$. These ratings were standardized ( $z$-score) and used as predictors in subsequent analyses.
Procedure Participants completed a category verification task in which an auditory cue-either a spoken category label (e.g., 'cat') or a natural sound (e.g., <meow>)preceded a visual image. Participants determined if each cue-image pair matched on a category level by pressing 'Yes' or 'No' using a labeled gaming controller. For example, if they heard a chainsaw revving or the spoken word "chainsaw" and then saw a picture of a chainsaw, they would press the 'Yes' button. The picture disappeared after each response, and performance feedback was given. Cue type (label, natural sound) and picture exemplar (4 per category) varied randomly within-subjects. There were a total of 576 trials per subject ( $50 \%$ cue-image category match). Each trial began with a 250 msec fixation cross followed by the auditory cue. The target image appeared 1 sec after auditory cue offset. This long delay ensured that participants had ample time to process sounds and labels (see Lupyan \& Thompson-Schill, 2012). The experiment took 30 minutes to complete.

## Results and Discussion ${ }^{2}$

Overall accuracy was high (96\%). Only correct response times (RTs) on matching trials were included. RTs less than 250 msec or greater than 1500 msec were excluded ( $<4 \%$ of correct trials). We fit the data with linear mixed regression (Bates, Maechler, \& Bolker, 2012) to predict response times (RTs) from the interaction between cue type (label, natural sound) and image rating (category-typicality or sound-

[^257]

Figure 2: Significant interaction between cue type and category typicality, but not between cue type and sound match when the target picture lagged auditory cue offset by 1000 msec . Confidence bands denote $\pm 1$ standard error of linear mixed regression point estimates (Mazerolle, 2012). Error bars denote $\pm 1$ standard error of main effect of cue type.
match) with random subject and item effects (target category). As expected (Lupyan \& Thompson-Schill, 2012), responses to label cues ( $M=609 \mathrm{msec}$ ) were reliably faster than responses to natural sound cues ( $M=639 \mathrm{msec}$ ), $F(1,13)=22.03, \quad p<0.001 .^{3}$ The effect of cue type was moderated by category-typicality, $F(1,13)=10.45, p=0.002$ (Fig. 2, left), but not by sound-match, $F(1,13)=0.001$, $p=0.98$ (Fig. 2, right).

To summarize, labels, but not natural sounds, resulted in faster processing of category-typical images, but neither cue resulted in faster processing of sound-matched images. These results replicate previous findings that labels facilitate visual processing more effectively than nonverbal cues (Lupyan \& Thompson-Schill, 2012) and that labels improve recognition of category-typical exemplars (Lupyan \& Swingley, 2012). The results clearly show that labels and natural sounds activate familiar concepts differently and that labels appear to activate a representation that is more categorical/typical. Unexpectedly, natural sounds did not selectively facilitate recognition of pictures that were better matches to the sound-cues. This finding is investigated further in Experiment 2.

[^258]
## Experiment 2

Our second experiment extends the first in two important ways. First, we compiled a more extensive set of stimuli by sampling from the 2-dimensional space of category typical and sound-matched category exemplars (Fig. 3). Second, we varied the cue-to-image delay. We did this because natural sounds, unlike labels, index the animals and objects that produce them. While labels often occur in the absence of the referent (we talk about things not presently in view), sounds are temporally contingent on the presence of the referent. If we hear a bark, chances are a dog is in the vicinity.

In Experiment 2, we investigated if label and natural sound cues influence recognition speed based on the fit between an auditory cue and an image, and on the delay between the cue and the image. In line with the results of Experiment 1, we predicted a label cue would improve processing of category-typical images. We also predicted that a natural sound would improve processing of a fuller set of sound-matched images-that is, where the image depicted an animal or object that was the likely source of the natural sound-and that this effect would be greater when the cue and image were temporally coupled-that is, presented simultaneously.

## Methods

Participants 56 University of Wisconsin-Madison undergraduates participated for course credit.
Materials Auditory cues comprised spoken labels and natural sounds for 10 of the 12 target categories used in Experiment 1 (categories river and toilet were excluded; all sounds edited to 600 msec ). Image ratings (categorytypicality and sound-match) for an augmented set of images were collected via Amazon's Mechanical Turk (mTurk). mTurk workers $(N=42)$ heard either 10 spoken labels or 10 natural sounds to be used in Experiment 2, and were


Figure 3: Sample stimuli from Experiment 2. Categorytypicality was measured independently of sound-match.
presented 8 to 10 pictures for each category with the following instructions: "Please listen to the following audio clip and report how well each image fits with the audio file." Ratings were given on a 5-point Likert scale. From these data, we selected 4 images for each category corresponding to the quadrants depicted in Fig. 3. There was a positive correlation between category-typicality and sound-match (Pearson's $r=0.27$ ). These ratings were standardized ( $z$-score) and used as predictors in subsequent analyses.
Procedure The procedure was the same as in Experiment 1. Cue type (Label, Natural Sound), picture exemplar (4 per category), and image delay (Simultaneous or Delayed 400 msec ) varied randomly within-subject for a total of 427 trials per subject ( $75 \%$ cue-image category match ${ }^{4}$ ). Each trial began with a 250 msec fixation cross. On a random half of the trials, the auditory cue and picture were presented simultaneously; on the remaining trials the picture was presented 400 msec after the offset of the auditory cue. The experiment took 30 minutes to complete.

## Results and Discussion

Overall accuracy was high ( $M=97 \%$ ), except trials in which pictures of scissors were cued by a sound of scissors cutting paper ( $M=91 \%, S D=1.8$ ). Participants also reported difficulties with these trials during debriefing ( 24 out of 56 participants; next most frequent was 5 for bee), and these trials were removed from subsequent analyses ( $<5 \%$ ). ${ }^{5} \mathrm{We}$ excluded trials using the same exclusion criteria as in Experiment 1 ( $<2 \%$ of correct trials removed). Again, we fit the data with linear mixed regression to predict response times from cue type (label, natural sound), delay (simultaneous, delayed), and image rating (category typicality or sound typicality) allowing random subject and item effects (picture category).
Delay and Cue Type We first report how the effect of cue type varied by image delay. As in Experiment 1, responses to label cues were reliably faster than responses to natural sound cues, $F(1,41)=30.14, p<0.0001$. The effect of cue type was moderated by delay, $F(1,41)=6.86, p=0.009$. The RT advantage of labels over natural sounds was greater on simultaneous trials than it was on delayed trials (Fig. 4).
Category Typicality We next report how image ratings of category typicality influenced RTs differently by cue type and by image delay. Category typicality was a reliable

[^259]predictor of RTs, $F(1,41)=10.30, p=0.001$. Importantly, this effect remained constant across both cue types and both image delays. That is, the RT advantage for more categorytypical images over less category-typical images was equivalent for label and natural sound cues, on both simultaneous and delayed trials (Fig. 4, left column). Responses following natural sound cues were predicted by category-typicality of the image during simultaneous and 400 msec delayed trials, an effect not found at the longer delay in Experiment 1.
Sound Match We now report how image ratings of soundmatch influenced response times differently by cue type and


Figure 4: Label and natural sound auditory cues affect response latencies differently by cue-image delay (rows) and by image rating (columns). Confidence bands denote $\pm 1$ standard error of linear mixed regression point estimates (Mazerolle, 2012). Error bars denote $\pm 1$ standard error of main effect of cue type.
by image delay. There was a reliable three-way interaction between sound-match, cue type, and image delay, $F(1,46)=$ 4.67, $p=0.03$. On simultaneous presentation trials, RTs following natural sound cues decreased as the sound-match of the image increased, while RTs following label cues did not vary by sound-match, $t(46)=-3.47, p<0.001$, (Fig. 4, upper right). However, there was no such cue type $\times$ soundmatch interaction at the 400 msec delay, $t(46)=-0.44, p=0.66$ (Fig. 4, lower right). That is, sound-match predicted RTs following natural sounds and not labels when the delay was simultaneous, but not with a 400 msec delay.

To summarize: the image ratings for category-typicality and sound-match correlated with response times based on the cue and the cue-image delay. First, when presented with a spoken label, RTs were predicted by category-typicality of the image, and this effect held across both cue-to-image delay periods. Second, when presented with a natural sound, the sound-match of the image correlated with the response time to that image, but only when the cue-image pair was presented simultaneously. That is, hearing a natural sound improved processing of a particular kind of visual image: a picture depicting an object that could have made the sound at the moment the sound was detected. These results show that the ways in which an auditory cue influences recognition of visual images depends on both the fit of the image to the auditory cue and the time course of the presentation.

## General Discussion

In two experiments we demonstrated that verbal and nonverbal cues systematically differ in how they activate conceptual information, as tested by the speed of visual recognition of category exemplars. Experiment 1 revealed more category-typical exemplars were recognized faster following a spoken label cue but not a natural sound. In addition, Experiment 1 revealed that exemplars that were more sound producing were not recognized faster following either auditory cue. Importantly, responses following natural sound cues did not vary as a function of category-typicality while those following labels did, suggesting that verbal and nonverbal cues are indeed operating on different gradients. Experiment 2 added to these results with a fuller stimulus set and varying image delays. In Experiment 2, but not in Experiment 1, responses following natural sounds did vary with category-typicality. We believe this result to be due to the shorter delays used in Experiment 2 (see Lupyan \& Thompson-Schill, 2012 for differences between labels and natural sounds at longer delays). In Experiment 2, but not in Experiment 1, responses to natural sounds varied as a function of the match between the sound and image, but the relationship was time sensitive. In particular, high soundmatched exemplars were recognized faster following a natural sound only during simultaneous presentation, and sound-match did not predict RTs following verbal cues.

Together, the two experiments reported here highlight the role of multisensory integration as a feature of what we have called "motivated" cues. We associate barking with dogs,
but the bark informs us about the particular dog that made it-a deeper bark is likely to come from a larger dog, and hearing a bark usually temporally coincides with seeing the actual animal. Such contingencies result in audiovisual integration of simultaneous auditory and visual cues that improves detection (Laurienti, Kraft, Maldjian, Burdette, \& Wallace, 2004). For example, Chen \& Spence (2011) reported increased visual detection of masked pictures when presented with a congruent natural sound cue, and that the effectiveness of an auditory cue varied by cue-image delay. The present results support the time sensitivity in cross modal priming of natural sounds and pictures, and measure the strength of this relationship through a "motivated" sound-to-image match.

In contrast, word-to-referent mappings are "unmotivated" (cf. Hockett, 1966). Saying "dog" in a deeper voice does not systematically imply a larger or angrier dog. ${ }^{6}$ So, even though both "dog" and a dog-bark may be unambiguously associated with dogs, the dog-bark indexes a specific dog with a specific size, location, and temperament. The word "dog", while varying systematically with aspects of the speaker (e.g., the lower the pitch, the more likely the speaker is to be male), does not systematically vary with the referent. We can talk about particular dogs, of course, but the word "dog" can and often does remain categorical, abstract.

In addition, these findings establish a heretofore underappreciated relationship between an auditory cue and a sound-matched image in similar cognitive processing tasks. Future attempts to compare semantic and conceptual processing of labels to that of natural sounds may benefit from operationalizing what we have termed the soundmatch between a natural sound and its purported referent (e.g., Saygin, Dick, \& Bates, 2005).

Conclusion We found verbal and nonverbal cues activate different conceptual representations evident in patterns of response latencies to recognize and verify different category exemplars. In a replication of previous findings, verbal cues facilitated recognition of category-typical images. We extended these findings to discern the specifics of conceptual representations activated via natural sound cues: Natural sounds facilitated visual processing of images that fit with the presented sound, but only if the sound and image were presented simultaneously. Critically, these effects were mediated by time, with natural sound cues improving responses to sound-matched images only during simultaneous presentation.

[^260]
## References

Aramaki, M., Marie, C., Kronland-Martinet, R., Ystad, S., \& Besson, M. (2010). Sound categorization and conceptual priming for nonlinguistic and linguistic sounds. Journal of Cognitive Neuroscience, 22(11), 2555-2569.
Ballas, J. A. (1993). Common factors in the identification of an assortment of brief everyday sounds. Journal of Experimental Psychology: Human Perception and Performance, 19(2), 250.
Bates, D., Maechler, M., \& Bolker, B. (2012). lme4: Linear mixed-effects models using S4 classes. R package version 0.999999-0. http://CRAN.R-project.org/package=lme4.

Chen, Y. C., \& Spence, C. (2010). When hearing the bark helps to identify the dog: Semantically-congruent sounds modulate the identification of masked pictures. Cognition, 114(3), 389-404.
Chen, Y. C., \& Spence, C. (2011). Crossmodal semantic priming by naturalistic sounds and spoken words enhances visual sensitivity. JEP: Human Perception and Performance, 37(5), 1554-1568.
Cummings, A., Čeponienė, R., Koyama, A., Saygin, A. P., Townsend, J., \& Dick, F. (2006). Auditory semantic networks for words and natural sounds. Brain Research, 1115(1), 92-107.
Dick, F., Saygin, A. P., Galati, G., Pitzalis, S., Bentrovato, S., D'Amico, S., et al. (2007). What is involved and what is necessary for complex linguistic and nonlinguistic auditory processing: evidence from functional magnetic resonance imaging and lesion data. Journal of Cognitive Neuroscience, 19(5), 799-816.
Fowler, C. A. (1990). Sound-producing sources as objects of perception: Rate normalization and nonspeech perception. The Journal of the Acoustical Society of America, 88, 1236.
Freed, D. J. (1990). Auditory correlates of perceived mallet hardness for a set of recorded percussive sound events. The Journal of the Acoustical Society of America, 87, 311.
Goll, J. C., Crutch, S. J., Loo, J. H. Y., Rohrer, J. D., Frost, C., Bamiou, D. E., \& Warren, J. D. (2010). Non-verbal sound processing in the primary progressive aphasias. Brain, 133(1), 272-285.
Gygi, B., Kidd, G. R., \& Watson, C. S. (2004). Spectraltemporal factors in the identification of environmental sounds. The Journal of the Acoustical Society of America, 115(3), 1252.
Hockett, C. F. (1966). The problem of universals in language. In J. H. Greenberg (Ed.), Universals of Language (2nd. Ed.) (Vol. 2, pp. 1-29). Cambride, MA: The MIT Press.
Hoffman, P., \& Ralph, M. A. L. (2013). Shapes, scents and sounds Quantifying the full multi-sensory basis of conceptual knowledge. Neuropsychologia, 51(1), 14-25.
Kunkler-Peck, A. J., \& Turvey, M. T. (2000). Hearing shape. JEP: Human Perception and Perform, 26(1), 279.
Laurienti, P., Kraft, R., Maldjian, J., Burdette, J., \& Wallace, M. (2004). Semantic congruence is a critical
factor in multisensory behavioral performance. Experimental Brain Research, 158(4).
Lupyan, G. (2008). From chair to "chair": A representational shift account of object labeling effects on memory. Journal of Experimental Psychology: General, 137(2), 348-369. doi:10.1037/0096-3445.137.2.348
Lupyan, G. (2012). Linguistically modulated perception and cognition: the label-feedback hypothesis, 1-13.
Lupyan, G., \& Swingley, D. (2012). Self-directed speech affects visual search performance. The Quarterly Journal of Experimental Psychology, 65(6), 1068-1085.
Lupyan, G., \& Thompson-Schill, S. L. (2012). The evocative power of words: Activation of concepts by verbal and nonverbal means. Journal of Experimental Psychology: General, 141(1), 170-186.
Lupyan, G., Rakison, D. H., \& McClelland, J. L. (2007). Language is not just for talking. Psychological Science, 18(12), 1077-1083.
Mazerolle, M. J. (2012) AICcmodavg: Model selection and multimodel inference based on (Q)AIC(c). R package version 1.26. http://CRAN.Rproject.org/package=AICcmodavg.
Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). Tutorial in Quantitative Methods for Psychology, 4(2), 61-64.
Orgs, G., Lange, K., Dombrowski, J.-H., \& Heil, M. (2008). N400-effects to task-irrelevant environmental sounds: Further evidence for obligatory conceptual processing. Neuroscience Letters, 436(2), 133-137.
Nuckolls, J. (1999). The Case for Sound Symbolism. Annual Review of Anthropology, 28, 252, 225.
Parise, C., \& Pavani, F. (2011). Evidence of sound symbolism in simple vocalizations. Experimental Brain Research, 214(3), 373-380.
Saygin, A. P., Dick, F., \& Bates, E. (2005). An on-line task for contrasting auditory processing in the verbal and nonverbal domains and norms for younger and older adults. Behavior Research Methods, 37(1), 99-110.
Saygin, A. P., Dick, F., Wilson, S. W., Dronkers, N. F., \& Bates, E. (2003). Neural resources for processing language and environmental sounds Evidence from aphasia. Brain, 126(4), 928-945.
Shintel, H., \& Nusbaum, H. C. (2007). The sound of motion in spoken language: Visual information conveyed by acoustic properties of speech. Cognition, 105(3), 681690.

Stuart, G. P., \& Jones, D. M. (1995). Priming the identification of environmental sounds. The Quarterly Journal of Experimental Psychology, 48(3), 741-761.
Taylor, A. M., Reby, D., \& McComb, K. (2008). Human listeners attend to size information in domestic dog growls. The Journal of the Acoustical Society of America, 123(5), 2903.
Van Petten, C., \& Rheinfelder, H. (1995). Conceptual relationships between spoken words and environmental sounds: Event-related brain potential measures. Neuropsychologia, 33(4), 485-508.

# Language and cognitive load in a dual task environment 

Nikolaos Engonopoulos (nikolaos.engonopoulos@ uni-potsdam.de)<br>Potsdam University, Department of Linguistics<br>Karl-Liebknecht-Str. 24-25, 14476 Potsdam, Germany

Asad Sayeed (asayeed@coli.uni-saarland.de) and<br>Vera Demberg (vera@coli.uni-saarland.de)<br>Cluster of Excellence, Saarland University,<br>Campus C7.4, 66123 Saarbrücken, Germany


#### Abstract

We investigate the effect of linguistic complexity on cognitive load in a dual-task scenario, namely simultaneous driving and language use. To this end, we designed an experiment where participants use a driving simulator while listening to spoken stimuli and answering comprehension questions. Online physiological measures of cognitive load, including the recently established Index of Cognitive Activity, as well as measures of performance in both tasks have been collected with high temporal resolution. The resulting aligned data streams can be used to test a vast array of different hypotheses about the relationship between performance, difficulty, and cognitive load in dual tasks at various levels of temporal resolution and linguistic structure. We present results of the data analysis, including evidence that different linguistic structures may cause measurable changes in cognitive workload on a very fine temporal scale in cases of increased primary task difficulty.


Keywords: relative clause; dual task; cognitive load; pupillometry; skin conductance; tracking task; driving; multitasking

## Introduction

Is there a relationship between psycholinguistic measures of language complexity and quantified cognitive workload in dual-task environments? To answer this question, we experimentally evaluate these measures of language processing in an environment where one task is language-related and the other not. Such language complexity measures have been shown in single-task studies to account for processing difficulty. This work represents a first step in which we investigate the effect of a grammatical structure (German locally ambiguous subject vs. object relative clauses) on a simplified, well-controlled non-linguistic task, a driving task.

Dual tasks are ubiquitous in everyday life, often in situations where attention and performance in the primary task is critical. An example is driving while engaging in dialogue, be it with a passenger, a dialogue-controlled interface, or remotely via mobile phone. Engaging in dialogue generally affects driving performance and safety (Just, Keller, \& Cynkar, 2008; Young, Regan, \& Hammer, 2007).

We manipulated the driving task difficulty and the structural complexity of the linguistic items. We also collected measurements of performance in both tasks and fine-grained physiological indicators of cognitive load, namely skin conductance levels and pupil sizes. We computed values from pupil size for the recent Index of Cognitive Activity (ICA). To our knowledge, this is the first study using the ICA measure in a setting with a language task.

## Background and Related Work

There is a rich literature on language use while driving a car, largely showing that speaking on the telephone has a negative effect on driving performance (Just et al., 2008; Kubose et al., 2006). Further studies found that this is specific to conversations with remote speakers (independent of whether one uses a hand-held device or free speaking), but that conversations with an in-car passenger are less problematic (Strayer, Drews, \& Johnston, 2003; Drews, Pasupathi, \& Strayer, 2004). It appears that passengers adapt their conversation to the traffic situation, leaving the driver more resources to deal with demands of the driving task when driving becomes difficult (Drews, Pasupathi, \& Strayer, 2008; Crundall, Bains, Chapman, \& Underwood, 2005; Villing, 2009). By contrast, remote conversational partners cannot adapt their speech, so that the driver may reach the point of cognitive overload more easily and thus commit driving errors. However, these lines of research have not taken into account how the fine-grained details of linguistic complexity affect cognitive load and driving task performance.

On the other hand, there is a very rich literature on linguistic processing difficulty in single tasks using brain imaging, ERPs, and reading time studies, as well as a number of dual task experiments generally showing that performance on the linguistic task deteriorates with increased complexity of the other task, see for example King and Just (1991). Finally, multiple models explain the effect of cognitive load in one task on performance in another (Baddeley, 2003; Wickens, 2008; Just, Carpenter, \& Miyake, 2003).

We see, however, unbroken ground in relating the effect of linguistic complexity on a realistic task (e.g., driving) and the size of the interference of linguistic processing with driving performance. This study takes a step in this direction in testing different methods for assessing cognitive load and the effect of one particular linguistic structure-incrementally ambiguous relative clauses-on driving performance in a simplified but controllable and continuous driving task.

## The dual-task experiment

## The ConTRe task

Our primary task was a tracking task (Jagacinski \& Flach, 2003) presented as a car driving scenario and called the "Continuous Tracking and Reaction" (ConTRe) task (Mahr, Feld,


Figure 1: A screenshot of the ConTRe steering task.

Moniri, \& Math, 2012). In this task, participants see a simulated 3-D road moving at a constant speed, intended to simulate a moving vehicle. Additionally, two bars of different color appear approximately 20 m in front of the simulated vehicle. The two bars represent the vehicle's position and the target (reference) position. They move laterally across the screen. The reference bar's movement is pseudo-randomly generated by an algorithm, while the "vehicle" bar is controllable by the participant by means of a gaming steering wheel. Participants were instructed to track the reference bar's movements with the controllable bar as closely as possible. To reduce noise in our data, we removed all other elements of the original ConTRe environment (e.g., buildings along the side of the road, and traffic lights), except for the road and the moving bars. A screenshot of the simulated environment can be seen in fig. 1.

This task is a useful abstraction of driving, since it allows a precise and continuous performance measure for steering, essential to driving. We manipulated the difficulty of the ConTRe task by changing the speed of the reference and vehicle bars in order to create a "difficult driving" condition and an "easy driving" condition ${ }^{1}$.

## Language comprehension task

The spoken comprehension task consists in listening to a sentence containing a relative clause followed by two thematically related 'filler' sentences and a yes/no comprehension question. Questions were related to the relative clause (50\% of the stimuli) or to the filler sentences. All sentences and questions are in German, inspired by Bader and Meng (1999). The stimuli are designed in pairs in such a way that the items in each pair are identical except for the form of the auxiliary of the relative clause (RC), which determines whether it is an object RC (ORC) or a subject RC (SRC). An example of such a relative clause pair is the following:

> Die Lehrerin, die einige Eltern wegen einer solchen Kleinigkeit angerufen [haben / hat], hat nun eine Elternversammlung einberufen.
> "The teacher ${ }_{\text {FEM }}$ [who called some parents / whom

[^261]some parents called] because of such a trivial issue, has now called a parents' meeting."

The sentence is locally ambiguous between ORC and SRC until reaching the auxiliary; in previous experiments, increased reaction times in a speeded judgment task (Bader \& Meng, 1999) have been observed when subjects read "haben" (ORC) compared to "hat" (SRC). This is evidence for an interpretive bias toward SRC. All items were synthesized prior to the experiment using MARY TTS (Schröder, Charfuelan, Pammi, \& Türk, 2008) and pauses manipulated so that the critical region duration (hat / haben) is always identical.

## Experimental setup

Each experiment is divided into 4 recording phases, each lasting about 6 minutes, with short pauses in-between. Each phase is composed of a driving-only phase of 2 minutes followed by a driving-with-language phase of approximately 4 minutes, during which 10 blocks, consisting each of one relative clause, two fillers and one question. Participants answer the question verbally and their response is coded by the experimenter. In the first and the third phase, the driving difficulty is set to "easy", while in the second and fourth phase it is set to "difficult". The order of presented items in the language condition was randomized, and we ensured that each person only saw one condition of each item.

## Measures of cognitive workload

We have two principal sources of quantified cognitive workload data: physiological and task dependent. Our physiological measures are further divided into two subtypes: pupil area-based (pupillometry) and skin conductance-based, both of which have been widely used in cognitive workload studies, although principally on non-linguistic tasks. Our study is an opportunity to evaluate the relative efficacy of these data sources on linguistic tasks. We also take the opportunity to evaluate a novel form of pupillometric data processing: the Index of Cognitive Activity (ICA). To the best of our knowledge, ours is the first study to investigate the potential of the ICA as a measure of linguistically-induced cognitive load in a dual-task scenario.

Our task-dependent measure is driving performance in our simulated environment, which serves to confirm the "realworld" effect of variations in cognitive workload.

The Index of Cognitive Activity (ICA) Research in pupillometry (Just et al., 2003; Engelhardt, Ferreira, \& Patsenko, 2010; Palinko, Kun, Shyrokov, \& Heeman, 2010) has found that cognition-related changes in pupil size typically amount to a difference of $20 \%$ relative to the typical pupil size (Laeng, Sirois, \& Gredebäck, 2010). However, light conditions also affect pupil sizes, with brightness-induced changes being much larger than cognitively induced ones (up to $120 \%$ of typical pupil size).

The Index of Cognitive Activity (ICA; Marshall (2002)) is a patented measure which applies signal processing techniques to filter out slow, large light-induced changes and identify
the occurrence of short, abrupt changes in pupil size, held to be caused by cognitive load. The ICA measure is argued to be robust with respect to changes in light conditions and eye movement. It relates the frequency of rapid small changes in pupil size (also known as pupillary hippus) to cognitive load. The ICA measure has been used for measuring cognitive load in driving simulation tasks (Schwalm, Keinath, \& Zimmer, 2008), simulated driving and visual search (Marshall, 2007), detecting different levels of surgical skill (Richstone et al., 2010), and for measuring linguistically induced cognitive load (Demberg, Kiagia, \& Sayeed, 2013) among other uses. Demberg (2013) provides a more detailed analysis of the ICA measure in the dual task setting presented here.

ICA measurements have been shown to be relatively stable across several commonly used eye tracker models and sample rates ranging from 60 to 300 Hz (Bartels \& Marshall, 2012). We used a head-mounted Eyelink II and sampled at 250 Hz .
Skin conductance response Our second physiological proxy for measuring cognitive load is skin conductance response (SCR), which we calculate from skin conductance level (SCL). Changes in the electrical conductance of the skin are due to activity of the sweat glands, which are in turn controlled by the sympathetic nervous system. Skin conductance amplitude usually changes with respect to its "neutral" (tonic) level in response to unexpected, significant, or aversive stimuli. SCL has been previously used as a measure of cognitive load (Shi, Ruiz, Taib, Choi, \& Chen, 2007). In a dual task experiment with simulated driving and a secondary cognitive task, B. Mehler, Reimer, Coughlin, and Dusek (2009) found that skin conductance levels peaked in cases of mental overload caused by incrementally increasing secondary task difficulty, which was followed by a deterioration in the performance of the primary task. Son and Park (2011) found skin conductance levels along with steering wheel reversals (used as a measure of task performance) to be good input features for an artificial neural network built to predict task difficulty.

We used the Ledalab software (Benedek \& Kaernbach, 2010) to separate our raw skin conductance measurements into an estimate of the tonic component and the phasic component. The software also allows to calculate the number of skin conductance response events. SCR events are the "peaks" of the phasic component of skin conductance; both the number of such events per time unit and the amplitude of the peaks are used in the analysis below.

Driving performance We use performance on the ConTRe task as an additional measure of cognitive load. The task lets us define several measures of task success, including the distance between the reference bar and the controllable bar at each point in time and the speed and acceleration of the controllable bar.

## Results

We ran our experiment with 24 German native speakers aged 20-34, with the total duration of the recorded samples sum-
ming up to about 12 hours. We performed our data analysis in R using linear mixed effects (LME) modeling with lme4 (Baayen, Davidson, \& Bates, 2008) and mgcv (Wood, 2001).

## Correlation between physiological measures

The first question we explored was whether our physiological measures are correlated with one another. While there is no significant correlation between the raw skin conductance levels and the ICA, we do find a significant positive correlation between the number of skin conductance events and the ICA (using Spearman's $\rho$; left ICA: $\rho=0.06 ; p<0.0001$; right ICA: $\rho=0.09 ; p<0.0001$ ). One important aspect to keep in mind is also possibly different latencies of the two measures in reaction to a stimulus.

We find a strong correlation between the ICA of the left and right eye (cor $=0.74 ; p>0.001$, Pearson's product-moment correlation coefficient).

## Response to experimental phases

Driving performance The next hypothesis we tested was whether our task performance measure in the driving task, i.e., the steering deviation, is sensitive not only to the driving task difficulty, but also to the presence of language. In figure 2 , we have plotted the mean deviation for each of the difficulty settings (easy and difficult driving), with and without the secondary linguistic task. Using linear mixed effects models with a random intercept and random slopes by subject, we found a large significant main effect of driving difficulty (coef $=0.3 ; t=20.33 ; p<0.001$ ), showing that steering was less accurate when driving was more difficult. We also found a significant positive main effect of whether we are in a language phase (coef $=-0.05 ; t=-5.00 ; p<0.001$; steering is worse when people are listening to language, see also figure 3 ), as well as a significant interaction between driving difficulty and the language phase, indicating that the effect of language was more burdensome in the difficult driving condition (coef $=-0.024 ; t=-6.98 ; p<0.001$ ). To confirm whether the effect of language is significant in both driving conditions, we also split the data into two subsets, easy driving and difficult driving, and found that the effect of language was significant in both linear mixed effects models.

This figure illustrates an obvious difference between steering deviation in the easy and difficult driving conditions.


Figure 2: Driving condition/language vs. steering deviation.

Table 1: ICA estimates for the driving plus language phases.

|  | right ICA |  |  |  | left ICA |  |  |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: | :---: |
|  | coef | t-value | sign | coef | t value | sign |  |
| (Intercept) | 0.8116 | 123.40 | $* * *$ | 0.7965 | 135.63 | $* * *$ |  |
| sound playing | 0.0198 | 10.88 | $* * *$ | 0.0186 | 10.10 | $* * *$ |  |
| easy driving | -0.0057 | -2.44 | $*$ | -0.0004 | -0.21 |  |  |

Table 2: \# of SCR events reduced during easy driving. (Random slope of driving condition by subject included.)

|  | Estimate | t value | signif. |
| :--- | ---: | ---: | :---: |
| (Intercept) | 0.68626 | 12.550 | $* * *$ |
| difficulty=easy | -0.06495 | -4.274 | $* * *$ |

Pupillometry For the ICA, we find a main effect of driving difficulty in the ICA of the right eye, but not in the left eye (Table 1). Furthermore, we find significantly more blinks during the phases when language was playing. In-depth analysis of the pupillometric data reveals that overall dilation was larger when people were listening to language stimuli, but the number of ICA events was lower (Figure 3). If we look into the language phase, however, the ICA of both eyes went down significantly whenever language wasn't playing (e.g., between stimuli; Table 1: we factored out the effect of blinks or partial blinks on both the pupil area calculations and the ICA). This effect can also be seen in Figure 3, where the 10 ICA spikes in the language region coincide with our 10 blocks of language stimuli.

Skin conductance For skin conductance, we cannot easily compare the easy vs. difficult driving settings, as the skin conductance measuring device was removed between phases, and comparison of absolute values between phases is thus impossible. A measure that can be compared between driving conditions is however the number of skin conductance events. When running a linear mixed effects regression model with this measure as a response variable, we find that more such skin conductance events happened, as expected, in the difficult driving condition, see Table 2.

We do however not find any significant effect of the language vs. no language condition on this measure. Unexpectedly, we find that tonic skin conductance is lower in the driving plus language condition, see Figure 3.

## Cognitive load and language processing difficulty

To this point, we find that the measures largely behave as expected. Thus we come to our main question: can they detect the effect of fine-grained language complexity? To this end, we analysed the data to see whether we can find a) a correlate for higher processing difficulty in the ambiguous region or right after the disambiguation at hat/haben, and b) whether ORCs lead to less cognitive load than SRCs.

Disambiguating region Detailed analysis of the ambiguous region of the relative clause shows that the Index of Cognitive Activity is high during the ambiguous region of the


Figure 3: Spline plots ( 120 knots; with 0.95 conf intervals) showing the effect of language on physiological measures during an experimental phase ( 2 min driving only followed by 4 min of driving plus language).


Figure 4: SCR during time that stimulus is spoken.
relative clause (during the time span of -2000 msec to 0 msec ), and that the ICA sharply falls right after disambiguation (see Table 4 which shows a significant reduction in ICA of both eyes following disambiguation, encoded as time wrt. onset). These effects hold over and above effects of the steering task, which have been mathematically accounted for by including the task difficulty as a factor in the model. These results indicate that subjects encounter processing difficulty due to the ambiguity. (This is possibly also something they learn during the experiment.)

For skin conductance, we know that effects can be expected 2-4 seconds after the stimulus. Figure 4 shows a significant rise in skin conductance during the five seconds after

Table 3: Mixed effects regression analysis with steering deviation as response variable, for region of 2 s before the onset till 2 s after end of the critical region.

|  | Estimate | t -value |  |
| :--- | :--- | ---: | :--- |
| (Intercept) | $3.562 \mathrm{e}-01$ | 17.07 | $* * *$ |
| phase time | $8.459 \mathrm{e}-08$ | 3.44 | $* * *$ |
| target velocity | $3.832 \mathrm{e}-01$ | 205.08 | $* * *$ |
| critical region | $1.396 \mathrm{e}-02$ | 2.88 | $* *$ |
| easy driving | $-2.248 \mathrm{e}-01$ | -64.91 | $* * *$ |
| target acceleration | $-2.680 \mathrm{e}-02$ | -5.90 | $* * *$ |

Table 4: Mixed effects regression analysis with left and right ICA as response variable, $100-1800 \mathrm{msec}$ after critical region onset. (Critical region duration: $0-600 \mathrm{msec}$ )

|  | left ICA |  | right ICA |  |
| :--- | :--- | :--- | :--- | :--- |
|  | Estimate | t-value | Estimate | t-value |
| (Intercept) | 0.7504 | $35.71 * * *$ | 0.736 | $37.82 * * *$ |
| subject RC | -0.0354 | -2.12 | $*$ |  |
| phase time | $-1.16 \times 10^{-7}$ | -2.59 | $*$ |  |
| time wrt. onset | $-2.78 \times 10^{-5}$ | $-6.38 * * *$ |  |  |
| steering veloc | 0.0257 | $5.37 * * *$ |  |  |
| steering accel | 0.0108 | 2.00 | $*$ |  |
| SRC:phase time | $1.34 \times 10^{-7}$ | 2.12 | $*$ |  |
|  |  |  | $4.36 * * *$ |  |
|  |  |  |  |  |
|  |  |  |  |  |

the critical region, which would be consistent with an interpretation that the ambiguity causes higher cognitive load.

But can we see any effect of our linguistic stimuli on the driving performance? We compared steering accuracy at the time of the disambiguating region with steering accuracy during the two seconds before and after, and indeed found that deviation of the controllable bar from the reference bar was significantly larger during the disambiguating region than before or after; see the positive coefficient (Table 3) for the binary variable "critical region".

Subject vs. object relative clauses Finally, we test whether the ICA is sensitive to fine-grained linguistic complexity effects. We isolated the subset of the data which fell within the 1800 msec following the onset of the critical region hat / haben. The duration of this critical region at hat / haben is 650 ms in both conditions, which we imposed by manipulating the duration of the phrase boundary pause during synthesis. On this subset of the data, we built two LME models (one for each eye) with the ICA measure as the response variable and the relative clause type as the fixed effect, while also introducing a random effect per participant.

The results of this analysis are shown in Table 4. We can see that there is a negative effect for the SRC type in both cases, although only the result for the right eye is significant. The interpretation of the coefficient is that SRCs tend to occur with smaller values of ICA than ORCs.

We did not find any significant effects of relative clause condition on skin conductance, overall pupil dilation or steering performance.

Table 5: LME model for answer accuracy.

|  | Estimate | t-value | Sig |
| :--- | ---: | ---: | :---: |
| INTERCEPT | 2.663 | 5.72 | $* * *$ |
| RC-TYPE (OBJ) | 0.445 | 1.17 |  |
| VOICE (PASSIVE) | -1.802 | -3.11 | $* *$ |
| DRIVINGDIFFICULTY (EASY) | -0.222 | -1.18 |  |

## Performance in the language task

A last link that we wanted to investigate was the one between performance in the linguistic task (i.e., answer accuracy) and the difficulty of the driving and language tasks. We built a binomial LME model with the answer accuracy as the response factor and driving task difficulty, relative clause type, and the voice (passive vs. active) of the question as fixed effects with a random intercept per participant and subject and a random slope for relative clause type by item ${ }^{2}$. The resulting coefficients are presented in Table 5. While answer accuracy was lower for object relative clauses (74\%) than for subject relative clauses ( $78 \%$ ), and lower in difficult driving ( $75 \%$ ) than in easy driving (77\%), these differences did not reach significance. (NB: questions related to relative clauses were only asked after half of the items; i.e., this analysis is based on relatively little data.) The only significant negative effect on answer accuracy was found for passive voice questions, which means that there are significantly more wrong answers to passive voice questions than to active voice ones (this is not unexpected, as it has long been known that passives are more difficult to process than actives (J. Mehler, 1963)).

## Discussion and conclusions

We designed the tasks in our experiment to require continuous attention. The language task clearly affects performance on the primary steering task: we see the effect of the secondary task in all of our measures. Furthermore, we find effects of linguistic ambiguity and complexity in our measures of cognitive load: during the ambiguous region in our stimuli, we see evidence for higher cognitive load in our pupillometric measure, which is also reflected in a slightly later galvanic skin response. During the disambiguating region, we observe significantly higher steering deviation, which indicates that people are allocating more mental resources to the linguistic task, hence impeding steering performance. We also found evidence for a measurable effect of linguistic complexity in our pupillometric measure ICA: the ICA was significantly higher during the disambiguating region and the following second for the ORC condition compared to SRC. This experiment provides early support for the ICA as a useful measure to assess language-induced cognitive load.

## References

Baayen, R., Davidson, D., \& Bates, D. (2008). Mixedeffects modeling with crossed random effects for subjects

[^262]and items. Journal of memory and language, 59(4), 390412.

Baddeley, A. (2003). Working memory: looking back and looking forward. Nature Reviews: Neuroscience, 4.
Bader, M., \& Meng, M. (1999). Subject-object ambiguities in german embedded clauses: An across-the-board comparison. Journal of Psycholinguistic Research, 28(2).
Bartels, M., \& Marshall, S. (2012). Measuring cognitive workload across different eye tracking hardware platforms. In Proceedings of the symposium on eye tracking research and applications (pp. 161-164).
Benedek, M., \& Kaernbach, C. (2010). A continuous measure of phasic electrodermal activity. Journal of neuroscience methods, 190(1-5), 80.
Crundall, D., Bains, M., Chapman, P., \& Underwood, G. (2005). Regulating conversation during driving: A problem for mobile telephones? Transportation Research Part F, 8(3), 197-211.
Demberg, V. (2013). Pupillometry: the index of cognitive activity in a dual-task study. In Proceedings of the 35th annual meeting of the cognitive science society (cogsci-13).
Demberg, V., Kiagia, E., \& Sayeed, A. (2013). Language and cognitive load in a dual task environment. In Proceedings of the 35th annual meeting of the cognitive science society (cogsci-13).
Drews, F. A., Pasupathi, M., \& Strayer, D. L. (2004). Passenger and cell-phone conversations in simulated driving. In Proceedings of the human factors and ergonomics society 48th annual meeting.
Drews, F. A., Pasupathi, M., \& Strayer, D. L. (2008). Passenger and cell phone conversations in simulated driving. Journal of Experimental Psychology: Applied, 14(4), 392.
Engelhardt, P. E., Ferreira, F., \& Patsenko, E. G. (2010). Pupillometry reveals processing load during spoken language comprehension. The Quarterly Journal of Experimental Psychology, 63, 639-645.
Jagacinski, \& Flach. (2003). Control theory for humans: Quantitative approaches to modeling performance. CRC.
Just, M. A., Carpenter, P. A., \& Miyake, A. (2003). Neuroindices of cognitive workload: neuroimaging, pupillometric and event-related potential studies of brain work. Theoretical Issues in Ergonomics Science, 4(1-2), 56-88.
Just, M. A., Keller, T. A., \& Cynkar, J. (2008). A decrease in brain activation associated with driving when listening to someone speak. Brain research, 1205, 70-80.
King, J., \& Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. Journal of Memory and Language, 30(5), 580-602.
Kubose, T. T., Bock, K., Dell, G. S., Garney, S. M., Kramer, A. F., \& Mayhugh, J. (2006). The effects of speech production and speech comprehension on simulated driving performance. Applied cognitive psychology, 20(1), 43-63.
Laeng, B., Sirois, S., \& Gredebäck, G. (2010). Pupillometry - a window to the preconscious? Perspectives on Psychological Science, 7(1), 18-27.

Mahr, A., Feld, M., Moniri, M., \& Math, R. (2012). The ConTRe (Continuous Tracking and Reaction) task: A flexible approach for assessing driver cognitive workload with high sensitivity. In Automotive user interfaces and interactive vehicular applications (pp. 88-91).
Marshall, S. (2002). The Index of Cognitive Activity: measuring cognitive workload. In Proceedings of the 7th conference on human factors and power plants, 2002.
Marshall, S. (2007). Identifying cognitive state from eye metrics. Aviation, space, and environmental medicine, 78(Supplement 1), B165-B175.
Mehler, B., Reimer, B., Coughlin, J., \& Dusek, J. (2009). Impact of incremental increases in cognitive workload on physiological arousal and performance in young adult drivers. Transportation Research Record: Journal of the Transportation Research Board, 2138(-1), 6-12.
Mehler, J. (1963). Some effects of grammatical transformations on the recall of english sentences. Journal of Verbal Learning and Verbal Behavior, 2(4), 346-351.
Palinko, O., Kun, A., Shyrokov, A., \& Heeman, P. (2010). Estimating cognitive load using remote eye tracking in a driving simulator. In ETRA.
Richstone, L., Schwartz, M., Seideman, C., Cadeddu, J., Marshall, S., \& Kavoussi, L. (2010). Eye metrics as an objective assessment of surgical skill. Annals of surgery, 252(1).
Schröder, M., Charfuelan, M., Pammi, S., \& Türk, O. (2008). The MARY TTS entry in the Blizzard Challenge 2008. In Proc. blizzard challenge.
Schwalm, M., Keinath, A., \& Zimmer, H. (2008). Pupillometry as a method for measuring mental workload within a simulated driving task. Human Factors for assistance and automation, 1-13.
Shi, Y., Ruiz, N., Taib, R., Choi, E., \& Chen, F. (2007). Galvanic skin response (gsr) as an index of cognitive load. In CHI '07 extended abstracts on human factors in computing systems (pp. 2651-2656).
Son, J., \& Park, M. (2011). Estimating cognitive load complexity using performance and physiological data in a driving simulator. In AutomotiveUI, 2011, adjunct proceedings.
Strayer, D., Drews, F., \& Johnston, W. (2003). Cell phoneinduced failures of visual attention during simulated driving. Journal of experimental psychology: Applied, 9(1), 23.

Villing, J. (2009). In-vehicle dialogue management - towards distinguishing between different types of workload. In Proceedings of SimPE, fourth workshop on speech in mobile and pervasive environments.
Wickens, C. D. (2008). Multiple resources and mental workload. Human factors, 50(3), 449-455.
Wood, S. (2001). mgcv: GAMs and generalized ridge regression for R. R news, l(2), 20-25.
Young, K., Regan, M., \& Hammer, M. (2007). Driver distraction: A review of the literature. Distracted driving. Sydney, NSW: Australasian College of Road Safety, 379-405.

# Math and Metacognition: Resolving the Paradox 

Shanna Erickson (serickson@ucmerced.edu) Evan Heit (eheit@ucmerced.edu)<br>Cognitive and Information Sciences, University of California, Merced<br>Merced, CA 95343 USA


#### Abstract

Metacognition plays a fundamental role in forming selfjudgments of ability and knowledge. Is metacognition domain and gender specific? Metacognitive judgments and performance were measured across biology, literature, and math content. Undergraduates took three shortened SAT II Subject Tests, and provided estimates of their performance both before and after taking each test. The results were that judgments differed across domain and gender. Overconfidence was evident in all domains, although estimates of ability were more accurate after taking a test. Males tended to be more overconfident, while females were less confident yet more accurately calibrated when estimating ability. Students were over-confident in math, bringing into question the existence of math phobia. Improvement in calibration and gender difference in calibration were most noticeable in math.


Keywords: metacognition, math anxiety, gender differences, mathematics education.

## Introduction

Metacognition, a form of higher-order thinking, plays an important role in cognitive processing. It impacts several areas within cognitive science, such as attention, memory, perception, comprehension, and problem solving (Kitchener, 1983; Metcalfe \& Shimamura, 1994). Metacognition aids intellectual endeavors requiring complex thought processes (Schoenfeld, 1983) and also affects social behavior (Jaccard, Dodge \& Guilamo-Ramos, 2005) and decision making (Cohen, Freeman, \& Thompson, 1998).

Two components of metacognition are of particular importance in education: the ability to monitor what you know, which acts as a basis for predicting retention, and the control processes that are used to enact study activities (Nelson \& Dunlosky, 1991). Students need to use metacognitive control in gauging what they know and deciding what study methods to use (Thiede, Anderson, \& Therriault, 2003; Metcalfe, 2009). This process is constantly changing, as students adapt their behaviors in monitoring a learning goal. Self-regulation is necessary for this process (Kornell \& Bjork, 2007), thus students must select from a variety of strategies, enacting these strategies in goal-directed activities, and monitoring their progress in using these strategies.

Success of metacognition affects students' academic performance (Hattie, Biggs, \& Purdie, 1996; Paris \& Paris, 2001, Coutinho, 2008), as well as their ability to communicate what they know about a particular problem. Being able to communicate their level of understanding to instructors is crucial to the learning process. It guides how classroom and self-study materials are constructed, and can
affect what strategies students are taught for study and examination.

Metacognition has been shown to play a crucial role in gauging one's own knowledge (Sperling, Howard, Stanley \& DuBois, 2004; Schunk \& Ertmer, 2000), including specific academic domains such as reading comprehension (Pressley, 2002), math (Pugalee, 2001), science (Schraw, Crippen, \& Hartley, 2006), and writing (Pugalee, 2001). Any improvements in metacognition would allow students to better judge what they know and how well they will be able to recall information. This holds much promise for improving student academic performance.

Despite the importance of metacognition, people commonly display glaring overconfidence in their selfperception of their own knowledge and various abilities (Kruger \& Dunning, 1999; Dunning, Johnson, Ehrlinger, \& Kruger, 2003). Furthermore, people with lower abilities show an even more exaggerated overconfidence. Students in particular often self-report confidence judgments that are unrelated to their actual performance on assessments (Schraw, 1996). Compounding this is students' inability to allocate study times effectively. Methods of self-guided study often result in non-optimal allocation of study time (Son \& Sethi, 2010). Improved methods are available, but students generally do not employ them, even though it has been shown that it is possible to use metacognitive control. There is potential for optimal study (Son \& Sethi, 2006), but students instead use uninformed metacognitive decisions to structure their study time.

A possible exception to the overconfidence phenomenon is the occurrence of math anxiety (Meece, Wigfield, \& Eccles, 1990; Furner \& Berman, 2003). Math anxiety (or math phobia) is a fear of math that leads to math avoidance or lower math performance (Ashcraft, 2002; Ashcraft \& Krause, 2007) and has been observed in children and adults alike (Wigfield \& Meece, 1988). This sometimes extreme anxiety is harmful in educational and workplace settings (Meece et al., 1990; Furner \& Berman, 2003), undermining national and worldwide priorities to emphasize science, technology, engineering, and math (STEM) achievement. Indeed, a recent national report predicts increased demand for STEM professionals in the US as well as an inadequate supply of prepared graduates (STEMconnector, 2013). Performing math tasks in stressful situations, such as during tests, only compounds math anxiety (Beilock, 2008). Mathphobic attitudes of teachers can also be detrimental to students' math achievement, particularly for female teachers and students (Beilock, 2010).

This fear of math implies that there should be a corresponding underconfidence in self-evaluation of
mathematical ability. The consistent and persisting documentation of widespread math phobia contradicts the finding that people are generally overconfident. How then can we resolve this paradox? We wish to determine if students are as overconfident in math as they are in other academic domains, or if is math an exception to an otherwise global overconfidence.

Past findings indicate that females generally lag behind their male counterparts on standardized test performance in math (Brown \& Josephs, 1999). This is particularly true among high school and college students (Hyde et al., 2006). This gap does appear to have narrowed in recent years (Else-Quest, Hyde, Shibley, Marcia, 2010). However, attitudes toward math between genders still follow differing patterns, and females are more likely to feel intimidated by math than are males (Jakobsson, 2012; Brown \& Josephs, 1999). This lack of confidence often leads to a selffulfilling lag in performance (Brown \& Josephs, 1999; Kiefer \& Sekaquaptewa, 2007) that can lead to gaps in performance between genders.

We wish to explore is if overconfidence generalizes to all domains of academic knowledge and ability, or if it is domain specific. If there exist confidence differences among various academic subjects, this suggests that overconfidence is domain specific and not a general phenomenon that is implied by the findings of Kruger \& Dunning (1999) and Dunning et al. (2003). If overconfidence is a global phenomenon, we would expect to see overconfidence in students' ratings across various academic domains. If metacognition is instead domain specific, we would then expect to find differences in overconfidence among academic domains. In the light of math phobia, we would expect to see underconfidence rather than overconfidence in math tasks, in contrast to other domains.

We also seek to determine if metacognitive ability differs over gender as well, keeping in mind that female students show greater math phobia than males. Finally, we compared metacognitive judgments before and after an intervention, namely taking a test, to determine if students are able to improve their metacognitive judgments. We expected to see improvements, as people could re-evaluate their metacognitive estimates after being exposed to more information in the intervention. This would be consistent with Bayesian accounts of cognition, in that people would be updating their hypothesis of ability based on new observations (Jones \& Love, 2011; Heit \& Erickson, 2011).

## Experiment

We considered test performance, confidence, and calibration in predicted scores. Three comparisons will be highlighted. The first is the comparison among the three different SAT II Subject Tests to assess if overconfidence is a domain specific or general phenomenon. While predictions (estimates before taking an assessment) provide a measure of general confidence within a subject, postdictions (estimates after taking an assessment) provide a more
accurate and comparable measure of metacognitive ability to evaluate knowledge. The use of SAT II Subject Test sample questions gave participants a reference for difficulty level of the assessment before they take it. However, it might have been some years since the participants have taken these, and some participants may have chosen to take a different selection of subject tests than the ones presented in this experiment. Use of retrieval fluency and recognition heuristics would negatively affect metacognition, both for past experience and future performance (Benjamin, Bjork, \& Hirshman, 1998). The use of postdictions brings all participants to a more equitable level of familiarity with the test material before making a judgment of ability. As such, calibration was determined by comparing postdicted estimates of performance with actual scores of performance. If overconfidence is domain specific, we then expect that metacognitive performance would differ among different domains, and that there would be a higher rate of underconfidence within math. If metacognition is domain general, then a similar level of overconfidence should be observed across all three assessments.

The second comparison will be one made between genders. Males were expected to show higher confidence ratings in math than females. Third is the comparison between predictions and postdictions for performance on a task. This allows us to determine if participants improved their metacognitive judgments after completing a task. We expect that postdictions for performance on a task will be more accurately calibrated than predictions for the same task, and results reflected this.

## Method

Participants There were 31 participants in this experiment: 17 female and 14 male. All were UC Merced undergraduates (mean age $=19.03, \mathrm{SD}=0.98$ ) who took the experiment as a form of extra credit in one of their introductory Psychology or Cognitive Science classes.

Tasks and Materials Participants took three tests: a biology, literature, and math test, each consisting of 15 questions. Participants were told before the experiment that they would be taking tests based on SAT II Subject Tests content. Before each test, participants were asked to provide a predicted score (out of 15) for how well they would do. After taking each test, participants provided a postdicted score for how well they thought they performed. They were not told their actual scores on tests.

## Results

Key descriptive results for all participants are shown in Figure 1. The leftmost bar for each category represents average predicted score, the middle bar represents average actual test score, and the rightmost bar represents average postdicted score. Average performance across all tests was $40 \%$. Participants showed general overconfidence in predicted scores before each test. Overconfidence generally persisted in postdicted scores, although drastic reductions in
residual magnitudes show that participants were better able to assess their ability after each test, providing evidence for improvement in metacognitive judgment of ability ( $t=3.30$, $d f=60, p<0.0001$ ). The only test in which participants showed slight underconfidence was biology.

Notably, participants showed high overconfidence in math. The residual for average predicted math score was $35 \%$. This was higher than the residuals for both biology (2.6\%) and literature (19\%).

Results by gender are shown in Figures 2 and 3. Notable was the difference in calibration between genders. Overall, females were more accurate in self-estimates of ability. Differences between their predictions and scores averaged $11 \%$, compared to $29 \%$ for males. Similarly for postdictions, females misestimated their performance by an average of $5 \%$ while males misestimated by an average of $14 \%$. Males were generally overconfident both before and after taking each assessment. Overall, females had lower measures of overconfidence. Within literature and math, females began with overestimates of their ability, but their postdictions were more calibrated. Within biology, females actually started underconfident and became even more so after taking this assessment.

Both genders show little trace of math phobia, as shown by their predominant overestimates of performance. Average prediction and postdiction residuals in math were $27 \%$ (overconfident) and $-5 \%$ (underconfident) for females and were $42 \%$ and $15 \%$ for males. Though participants were generally overconfident with their predictions, they were able to improve their metacognitive judgment accuracy significantly in this domain. Males showed the most marked improvement in calibration in math, and females actually changed their judgments from being overconfident to predominantly underconfident.

In a three-way, predicted versus actual score $\times$ academic subject (biology or literature or math) $\times$ gender (male or female) ANOVA, there was a main effect of gender $F(1,29)$ $=4.48, M S E=13.69, \eta^{2}=0.08, p<0.05$. There was also a significant main effect of predicted ( mean $=59.00$ ) versus actual $($ mean $=40.22)$ score, $F(1,29)=36.61, M S E=10.1$, $\eta^{2}=0.50, p<0.0001$, indicating overconfidence in predictions. There was also a significant main effect of academic subject, $F(2,116)=6.59, M S E=5.33, \eta^{2}=0.08$, $p<0.01$. Notice that scores were lowest overall in math. There was a significant interaction between these two variables, $F(2,116)=16.80, M S E=5.33, \eta^{2}=0.20, p<$ 0.0001 , indicating that degree of overconfidence depended on academic subject. Overconfidence was greatest in math (predicted score $=62.15$, actual score $=27.53$ ). We are careful not to over-interpret the interaction, as actual scores also varied by academic subject. There was also a significant interaction between gender and predicted versus actual score, $F(1,29)=8.19, M S E=10.1, \eta^{2}=0.11, p<$ 0.01 , providing further evidence that overconfidence depended on gender. The remaining main effects and interactions were not significant.


Figure 1: Overall results by test.


Figure 2: Overall results for females.


Figure 3: Overall results for males.
We also conducted a similar analysis on postdicted scores $($ mean $=43.44)$ and actual scores. This ANOVA revealed a main effect of academic subject, $F(2,116)=36.28, M S E=$ 4.85, $\eta^{2}=0.37, p<0.0001$, as well as a main effect of gender $F(1,29)=4.48, M S E=15.49, \eta^{2}=0.09, p<0.05$. There was also a significant interaction between gender and postdicted versus actual scores, $F(1,106.27)=17.55, M S E$ $=6.05, \eta^{2}=0.37, p<0.001$, again showing gender differences in overconfidence. Remaining main effects and interactions were not significant.

Note that although scores were lowest for math, this assessment was not designed to be more difficult than the
other subject tests. In fact, it had the lowest difficulty level. During pilot experiments, test questions were chosen using difficulty ratings provided by College Board. Although we originally chose a variety of easy, medium, and difficult questions for each subject test, performance on this balanced math test was so poor that we substituted easier questions in place of all medium and difficult questions. Thus, the severe overconfidence observed in math is not a result of higher test difficulty level compared to other academic subjects.

Figures 4 and 5 show calibration slopes by domain. The dashed line represents the equation $y=x$ (predicted score $=$ actual score) is used to convey perfect calibration. The closer a line is to this dashed line, the better the calibration. For the predicted scores, there are apparent subject differences, e.g., the slope is highest for math, indicating the highest level of sensitivity to actual performance, and the slope is actually slightly negative for biology. Each domain slope more closely follows the calibration line $y=x$ for postdicted scores, showing that that participants were better able to judge their ability after taking each assessment. Each of models also crosses $y=x$, switching from overconfidence to underconfidence as test performance increases. In addition, correlations between actual scores and residuals calculated from estimated scores $(r=-0.63$ for both predicted and postdicted residuals) reveal that


Figure 4: Overall results for postdicted scores.


Figure 5: Overall results for females.
higher scores are associated with lower residuals. These findings support the previous work by Kruger \& Dunning (1999), Dunning et al. (2003) and Schraw (1996) and show that people with low test scores generally exhibit overconfidence, while people with high test scores are better able to judge their ability. Thus higher performing students tend to be better judges of their ability than are lower performers.

Gender differences were most striking within math. Figure 6 and 7 show calibration models by gender for both predictions and postdictions. Males made predictions with almost no calibration ( $r=0.02$ ), and females were overconfident overall with the predictions. Despite this, both genders were able to make much more accurate postdictions. In fact, these postdiction models were the best of any of the observed estimates of ability in this experiment when compared to other subjects.

## Discussion

The results of this study support past findings that people are generally overconfident in their abilities, although overconfidence does not appear to be exactly the same across domains. This was shown in the varying judgments of ability across academic domains. While both males and females are generally overconfident, females tend to be better calibrated in judging their domain knowledge.


Figure 6: Overall results for females.


Figure 7: Overall results for females.

Postdictions were significantly lower than predictions, showing that people are able to recalibrate their metacognitive judgments towards more accurate judgments after attempting an assessment.

In addressing the paradox of general overconfidence alongside the seeming exception of math phobia, we saw that overconfidence was particularly high in predicted scores for math assessment. This led us to question whether math phobia was present.

Although there was a higher incidence of overconfidence in mathematics, participants showed the greatest beneficial adjustment of metacognitive judgment miscalibration for mathematical ability. All participants were successfully able to recalibrate their estimates towards more accurate judgments of domain knowledge after an assessment.

We have replicated this severe overconfidence in math in other experiments, although gender no longer reached the level of statistical significance. Thus high math overconfidence is not specific to college-level students: In a subsequent experiment $(n=40)$, this result was replicated at a local high school using the same experimental design. In another experiment with college students $(n=46)$, we extended our findings by using the same experiment presented here, but also including Likert scale measures of confidence for each domain, as well as an adapted math Anxiety Rating Scale (MARS) survey (Alexander \& Martray, 1989). Initial results indicate that there does exist math phobia, as we observed MARS ratings similar to other college populations identified as math anxious. We expected higher anxiety ratings to be linked with underestimates of ability. We did not observe this. Instead, math phobia moderated overestimates of ability to be less extreme, although overconfidence still persisted. This is a possible explanation for the coexistence of math phobia and overconfidence in mathematical ability. Further plans include the replication of studies within actual classroom settings in which participants must judge their ability on class assessments.

Does a higher confidence in one subject over another really indicate domain specificity rather than generality? If so, this suggests we may be using different metacognitive methods for different domains such as sciences versus the humanities. Alternatively, we might be using one overarching metacognitive ability that uses different cues and leads to different results across domains.

Our results are relevant for applications in cognitive science, particularly for studying and improving education. We have seen that students are overconfident in math, yet there is evidence that these same students are math phobic. These views pose two strong deterrents for students to seek practice and improvement in math. If students are overconfident in their mathematical abilities and have anxiety about mathematical tasks, they have little incentive to study the subject. This reluctance likely carries over to other science, technology, and engineering subjects that require a significant amount of math background.

We also know that our use of metacognition does not always lead to calibrated self-views of ability. There are optimal models for allocation of study time, but student behaviors do not conform to these (Son \& Sethi, 2006; 2010). Judgments of improvement and learning rate that students use to make time allocation decision are often inaccurate as well (Townsend \& Heit, 2010; 2011). Math phobia has come to be so expected that it has started to influence curriculum design. Already changes have been made in computer science programs to deemphasize math (Tucker, 2001) even though math content is fundamental to this area. The spread of this trend to other science, technology and engineering programs would seriously undermine students' foundational math knowledge. Therefore it will be important to develop techniques that improve students' metacognitive calibration for mathematics and other subjects.

## References

Alexander, L., \& Martray, C. R. (1989). The development of an abbreviated version of the math Anxiety Rating Scale. Measurement and Evaluation in Counseling and Development.
Ashcraft, M. H. (2002). Math anxiety: Personal, educational, and cognitive consequences. Current Directions in Psychological Science, 11(5), 181-185.
Ashcraft, M. H., \& Krause, J. A. (2007). Working memory, math performance, and math anxiety. Psychonomic Bulletin \& Review, 14(2), 243-248.
Beilock, S. L. (2008). Math performance in stressful situations. Current Directions in Psychological Science, 17(5), 339-343.
Beilock, S. L., Gunderson, E. A., Ramirez, G., \& Levine, S. C. (2010). Female teachers' math anxiety affects girls' math achievement. Proceedings of the National Academy of Sciences, 107(5), 1860-1863.
Benjamin, A. S., Bjork, R. A., \& Hirshman, E. (1998). Predicting the future and reconstructing the past: A Bayesian characterization of the utility of subjective fluency. Acta Psychologica, 98(2), 267-290.
Brown, R. P., \& Josephs, R. A. (1999). A burden of proof: Stereotype relevance and gender differences in math performance. Journal of personality and social psychology, 76(2), 246.
Cohen, M. S., Freeman, J. T., \& Thompson, B. (1998). Critical thinking skills in tactical decision making: A model and a training strategy. Making decisions under stress: Implications for individual and team training, 155190.

Else-Quest, N. M., Hyde, J. S., \& Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: a meta-analysis. Psychological Bulletin,136(1), 103.
Coutinho, S. (2008). Self-efficacy, metacognition, and performance. North American Journal of Psychology, 10(1), 165-172.

Dunning, D., Johnson, K., Ehrlinger, J., \& Kruger, J. (2003). Why people fail to recognize their own incompetence. Current Directions in Psychological Science, 12(3), 83-87.
Furner, J. M., \& Berman, B. T. (2003). Review of Research: Math Anxiety: Overcoming a Major Obstacle to the Improvement of Student Math Performance. Childhood Education, 79(3), 170-174.
Hattie, J., Biggs, J., \& Purdie, N. (1996). Effects of learning skills interventions on student learning: A meta-analysis. Review of educational research, 66(2), 99-136.
Heit, E., \& Erickson, S. (2011). In praise of secular Bayesianism. Behavioral and Brain Sciences, 34(04), 202-202.
Hyde, J. S., Fennema, E., Ryan, M., Frost, L. A., \& Hopp, C. (2006). Gender comparisons of mathematics attitudes and affect. Psychology of Women Quarterly, 14(3), 299324.

Jakobsson, N. (2012). Gender and confidence: are women underconfident?.Applied Economics Letters, 19(11), 1057-1059.
Jones, M., \& Love, B. C. (2011). Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. Behavioral and Brain Sciences, 34(04), 169-188.
Kiefer, A. K., \& Sekaquaptewa, D. (2007). Implicit stereotypes and women's math performance: How implicit gender-math stereotypes influence women's susceptibility to stereotype threat. Journal of Experimental Social Psychology,43(5), 825-832.
Kitchener, K. S. (1983). Cognition, metacognition, and epistemic cognition, Human development, 26, 222-232.
Kornell, N., \& Bjork, R. A. (2007). The promise and perils of self-regulated study. Psychonomic Bulletin \& Review, 14(2), 219-224.
Kruger, J., \& Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. Journal of Personality and Social Psychology, 77(6), 1121.
Jaccard, J., Dodge, T., \& Guilamo-Ramos, V. (2005). Metacognition, risk behavior, and risk outcomes: the role of perceived intelligence and perceived knowledge. Health Psychology, 24(2), 161.
Meece, J. L., Wigfield, A., \& Eccles, J. S. (1990). Predictors of math anxiety and its influence on young adolescents' course enrollment intentions and performance in mathematics. Journal of Educational Psychology, 82(1), 60.

Metcalfe, J. (2009). Metacognitive judgments and control of study. Current Directions in Psychological Science, 18(3), 159-163.
Metcalfe, J., \& Shimamura, A. P. (1994). Metacognition: Knowing about knowing. Cambridge, MA: MIT Press/Bradford.
Nelson, T. O., \& Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at
predicting subsequent recall: The "delayed-JOL effect". Psychological Science, 2(4), 267-270.
Paris, S. G., \& Paris, A. H. (2001). Classroom applications of research on self-regulated learning. Educational Psychologist, 36(2), 89-101.
Plake, B. S., \& Parker, C. S. (1982). The development and validation of a revised version of the mathematics anxiety rating scale. Educational and Psychological Measurement, 42(2), 551-557.
Pressley, M. (2002). Metacognition and self-regulated comprehension. What Research Has to Say About Reading Instruction, 3, 291-309.
Pugalee, D. K. (2001). Writing, mathematics, and metacognition: Looking for connections through students' work in mathematical problem solving. School Science and math, 101(5), 236-245.
Schoenfeld, A. H. (1989). Explorations of students' mathematical beliefs and behavior. Journal for Research in Mathematics Education, 20, 338-355.
Schraw, G. (1996). The effect of generalized metacognitive knowledge on test performance and confidence judgments. The Journal of Experimental Education, 65(2), 135-146.
Schraw, G., Crippen, K. J., \& Hartley, K. (2006). Promoting self-regulation in science education: Metacognition as part of a broader perspective on learning. Research in Science Education, 36(1), 111-139.
Son, L. K., \& Sethi, R. (2006). Metacognitive control and optimal learning. Cognitive Science, 30(4), 759-774.
Son, L. K., \& Sethi, R. (2010). Adaptive learning and the allocation of time. Adaptive Behavior, 18(2), 132-140.
Sperling, R. A., Howard, B. C., Staley, R., \& DuBois, N. (2004). Metacognition and self-regulated learning constructs. Educational Research and Evaluation,10(2), 117-139.
Thiede, K. W., Anderson, M., \& Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. Journal of Educational Psychology, 95(1), 66.
Townsend, C., \& Heit, E. (2010). Metacognitive judgments of improvement are uncorrelated with learning rate. Proceedings of the Thirty-Second Annual Conference of the Cognitive Science Society, 1858-1862.
Townsend, C., \& Heit, E. (2011). Framing influences aggregate judgments of learning. Proceedings of the Thirty-Third Annual Conference of the Cognitive Science Society, 756-761.
Tucker, A. B., Kelemen, C. F., \& Bruce, K. B. (2001, February). Our curriculum has become math-phobic! In ACM SIGCSE Bulletin, 33(1), 243-247).
Wigfield, A., \& Meece, J. L. (1988). Math anxiety in elementary and secondary school students. Journal of Educational Psychology, 80(2), 210.

# Candle-Candle-Candle-Candy: Continuous Attraction Toward Previously Seen Phonological Competitors 

Jesse Falke (jfalke@ucmerced.edu)<br>Bodo Winter (bodo@bodowinter.com)<br>Michael J. Spivey (spivey@ucmerced.edu)<br>Cognitive and Information Sciences, 5200 North Lake Road, Merced, CA 95343 USA


#### Abstract

Mouse-tracking provides rich information about temporally sensitive mental processing. In two experiments, we applied this methodology to a phonological cohort task that can be interpreted as a version of the A-not-B task. In the first experiment, participants had to click a word such as "candle" three times in a row on the same side of the computer screen. They then had to click a phonological competitor ("candy") on the other side during the critical trial. This was contrasted with a condition in which the word to be clicked three times in a row was phonologically unrelated to the word at the critical trial. We found that the phonological priming increased attraction toward the competitor. In the second experiment, mouse movements revealed attraction towards the competitor as a function of the number of previous presentations. The results demonstrate that phonological competitors can exert graded influence on motor responses even if the competitors are not simultaneously presented. These results are predicted by and provide evidence for the dynamic field theory of movement preparation and execution. These results can furthermore be interpreted as evidence for continuity underlying the A-not-B task.


Keywords: A-not-B error; deictic pointers; dynamical systems; mouse-tracking

## Introduction

The A-not-B error has been investigated with children for over half of a century (Piaget, 1954). In the standard version of the task, the researcher presents an object to the child and hides it in one location (" A "). When this process is repeated multiple times, the child will often reach for the object in " A " even if it was moved in front of the child's eyes to another location, "B". Eight to ten month old children reliably commit this error (cf., Marcovitch \& Zelazo, 1999).

Smith and Thelen (2003) and Thelen, Schöner, Scheier and Smith (2001) conceptualize the A-not-B task in terms of dynamic field theory (Erlhagen \& Schöner, 2002), where the decision to perform a movement towards either A or B develops in a continuous fashion. The theory views the A-not-B error as the result of changes to an activation field, where both $A$ and $B$ are represented as points on a plane. If the researcher hides a toy under A , the point for A increases in activation and sends inhibitory activation to B . The child reaches for A if a threshold of activation is crossed. Crucially, the memory of this reaching "pre-shapes" the field for the next trial. Over multiple trials, the A region of the field becomes stronger and increasingly exerts inhibitory
influence on the B region of the field, ultimately resulting in the A-not-B error.

Dynamic field theory accurately predicts that if posture is changed between A and B trials (Smith, Thelen, Titzer, \& McLin, 1999), the child does not commit an A-not-B error as often. This follows from the assumption that changes in posture on $A$ and $B$ trials decrease the similarity of preceding memories to the current trials, thus lessening the strength of the pre-shaping of the field.

This account is also compatible with another area of research, deictic pointers (Ballard, Hayhoe, Pook \& Rao, 1997; Chun \& Nakayama, 2000; Richardson \& Spivey, 2000). By associating content and locations with a deictic pointer, a cognitive agent can reference necessary information and use it to aid action without having to build up a detailed model of the world. We will argue that the formation of dynamic fields can be viewed as the formation of deictic pointers.

In this paper, we explore how previously seen stimuli affect perseveration in a mouse-tracking experiment. Mouse-tracking provides a real-time stream of $x, y$ coordinates during movement that has been used to reveal the continuous dynamics underlying a diverse set of cognitive processes, including phonological competition in lexical access (Spivey, Grosjean, \& Knoblich, 2005), the comprehension of sentence negation (Dale \& Duran, 2011), the categorization of typical and atypical objects (Dale, Kehoe, \& Spivey, 2007), and the categorization of faces (Freeman, Ambady, Rule, \& Johnson, 2008), among many others (for a review, see Freeman, Dale, \& Farmer, 2011).

In Spivey et al. (2005), participants saw two objects in opposite corners of the computer screen, e.g. a candy and a candle. They then heard a target word referencing one of the objects, such as "candy", while they executed the movement to click it. When the two objects were phonologically related, the mouse gravitated more toward the competitor object than when they were phonologically unrelated.

Here, we extend the task used by Spivey et al. (2005) to show the graded influence of phonological competitors that are not simultaneously present on the critical trial. Similar to the A-not-B task, we present an object such as "candy" multiple times on one side, and then on the next trial we present "candle" on the opposite side. Dynamic field theory predicts that multiple memory traces of "candy" on one side should exert graded inhibition when seeing the phonologically related "candle" at a different spatial
location. While adults may not commit the full A-not-B error by accidentally clicking on the wrong object, their mouse movements might still reveal continuous attraction towards previously seen competitors.

## Experiment 1

In the first experiment, we presented each participant with phonologically related and phonologically unrelated stimuli. Mouse movements were recorded during each trial to investigate the effects of repeated movements to phonological and non-phonological competitors.

## Methods

Participants Thirty-three undergraduates at UC Merced volunteered to participate to receive partial course credit. All participants were right-handed native speakers of English. 3 stimuli were excluded because of computer lag (0.24 \%).

Stimuli and Procedure The procedure was run using MouseTracker (Freeman \& Ambady, 2010). On each trial, two 200 pixel-wide pictures were presented in the top corners of the screen. Using headphones, we presented a target word for which the corresponding object had to be clicked (e.g., "pickle", "pepper"). Each pair of objects was always phonologically unrelated (e.g., "candle" vs. "lobster"). However, there were 16 critical trials that were preceded either by three trials to the same side without a phonological competitor (lighting-left, pepper-left, speakerleft, candy-right) or three trials to the same side with a phonological competitor (candle-left, candle-left, candleleft, candy-right). These two conditions are called "motor-3" and "phonological-3" respectively. The prime was always the target item. So, for example, a participant might have had to click "candy" three times on the left side and then, on the critical trial, "candle" had to be clicked on the right side (see Fig. 1). Across participants, we balanced the position (left vs. right) in which the target and the primes occurred, and we also balanced which of the competitors occurred as prime, and which as target (i.e., "candy-candy-candycandle" vs. "candle-candle-candle-candy"). In total, there were 8 "motor-3" items and 8 "phonological-3" items (16 critical stimuli)..

There were also 24 filler trials that occurred between critical trials and subsequent priming trials. In the analyses below, we count these filler trials as control trials, as they represent mouse movements toward target objects for which there is no previous prime and no phonological competition. There were thus 88 trials in total. These were preceded by 12 practice trials.
We instructed participants to initialize mouse movements before they heard the sound file. To encourage this, the gain was slowed down to 2 (MouseTracker setting), and the sound file played the target word after a 500 ms delay.


Figure 1: Schematic depiction of experimental trials. Note that on the critical trial, the competitor is not present.

All stimuli were spoken by a native speaker of English. For the prime-3 condition, we recorded three different utterances of the same word by the same speaker to reduce the possibility of selective adaptation effects. Mouse coordinate data was sampled at 60 Hz and was recorded with screen display information, movement durations and final response.

Analyses We inverted the $x$ coordinates of left-going responses so that left- and right-going responses had comparable spatial metrics. We then normalized all responses to have a common origin at ( 0,0 ). Mouse-tracking provides a large set of potential dependent measures. We focused on the Euclidian distance of each measured point from the diagonal line that is defined by the origin in the center of the screen and the corner response box. All analyses we present were time-normalized to 101 time steps per trial.
We analyzed our data in two ways. First, we used the R package lme4 (Bates, Maechler \& Bolker, 2012) to perform growth curve modeling (Mirman, Dixon \& Magnuson, 2008). Time step ( 1 to 101) entered the analysis as a fourth order orthogonal polynomial fixed effect (including lowerorder polynomials), and the crucial effect of interest was the interaction of condition (prime-1, prime-3, control) with time. In the by-subjects-analysis, we included random intercepts for subjects, as well as subject random slopes for time and condition (following Mirman et al., 2008). In the by-items-analysis, we did the same for items. P-values were derived separately for each coefficient based on normal approximated t -values.
Growth curve analysis allows modeling the precise trajectory; however, for comparability with other mouse tracking studies and to get the exact time points of where trajectories differ, we present an alternative analysis following Dale, Kehoe and Spivey (2007), who have shown by means of simulated random trajectories that 8 consecutive t-tests may count as a significant result at $\alpha=$ 0.01 .

With mouse-tracking data, there is always the possibility that average differences between conditions are not due to genuine gradedness in the response, but due to averaging over trials that head straight to a target and trials where participants correct a categorical choice midflight. To assess whether this could explain our results, we computed the bimodality coefficient $b$ (see Freeman \& Ambady, 2010) on the $z$-scored (by subjects and by condition) maximum deviation from the diagonal line and the area under the curve (measures are described in Freeman \& Ambady, 2010: 229). $b$ values over 0.555 are interpreted as evidence for bimodality.

## Results

There were 9 errors in total ( $0.72 \%$ of all trials), all in the control condition. Therefore, there was no indication of a categorical A-not-B error. All subsequent analyses are performed on correct trials only.

Fig. 2 shows the evolution of the Euclidian distances from the diagonal line over time. For the growth curve analysis, likelihood ratio tests revealed a significant difference between the phonological and the pure motor condition by items ( $p=0.025$ ), and a marginally significant difference by subjects ( $p=0.079$ ). In both cases, there were significant interaction effects of condition for the cubic and quadratic components of the model (cubic by subjects, $\mathrm{p}=0.0054$, by items, $p=0.02$; quadratic by subjects, $p=0.024$, by items, $\mathrm{p}=0.021$ ).

The phonological competition condition was significantly different from control by subjects and by items ( $p=0.009$, $\mathrm{p}=0.0065$ ). In both cases, the intercept was higher for phonological competition than for control (by subjects: $\mathrm{p}=0.016$, by items: $\mathrm{p}=0.027$ ), indicating overall larger Euclidian distances for this condition. In the subjects analysis, there were additionally significant effects for interactions between condition and the linear ( $\mathrm{p}=0.003$ ), cubic ( $\mathrm{p}=0.01$ ) and quadratic ( $\mathrm{p}=0.0034$ ) components of the model. The motor priming condition was significantly different from control only by subjects ( $p=0.008$ ) and not by items ( $p=0.175$ ). In both cases there were individual effects for the intercept (higher in motor priming than in control, by subjects: $p=0.01$; by items: $p=0.048$ ), but no effects for higher-order polynomials.

The alternative analysis, following Dale et al. (2007), revealed no consecutive significant differences between phonological and motor priming that passed the 8 t-test criterion. There were 36 significant differences between phonological priming and the control condition in the subjects analysis (time points 59 to 94), and 50 in the items analysis (time points 3 to 25 and 59 to 85). The region that is significant in both analyses is shaded in Fig. 3. There were 29 significant differences ( 73 to 101) between control and motor priming by subjects, as well as 30 by items ( 3 to 32). Interestingly, in this case, these regions were not overlapping.


Figure 2: Euclidian distance as a function of time. Gray area indicates significant differences between phonological competition and control (by subjects and items).

Bimodality analyses revealed no subject with $b>0.555$ for the crucial phonological competition condition, neither for the measure "maximum deviation from the diagonal", nor for the measure "area under the curve". For the motor priming condition, 1 participant had $\mathrm{b}>0.555$ for maximum deviation ( $\sim 3 \%$ ), and 4 participants for area under the curve ( $\sim 12 \%$ ). Again, this shows that the results are fairly unimodal across the board.

## Discussion

The results for the motor-3 and phonological-3 conditions were interesting. In the growth curve analysis, it was surprising that there were significant differences between the two conditions for items, but only marginally significant differences for subjects. This is surprising because the items-based analyses use a smaller sample than the subjects based analyses ( $\mathrm{N}=33$ in the subjects analysis and $\mathrm{N}=16$ in the items analysis). More data will need to be collected to explain this.

Another way to look at the data is to see when and how long the two conditions differed from the control condition in the analysis proposed by Dale and colleagues (2007). While both conditions displayed significant differences from the control by subjects and items, only the phonological priming condition displayed significant differences by both subjects and items simultaneously. In addition, the phonological condition resulted in more total significant differences from the control than the motor priming condition ( 86 vs. 59). If we take the number of significant simultaneous consecutive differences as a measure of strength of the difference (cf., Dale et al., 2007), then the prime-3 condition showed more gravitation away from the diagonal line (towards previously seen competitors) than the prime- 1 condition.

While this experiment has revealed that attraction toward a previously displayed stimulus is modulated by said stimulus being a phonological competitor, there is more to
dynamic field theory. Experiment 2 was designed to further investigate how well dynamic field theory could be applied to these results by varying the number of phonological primes before the critical trial.

## Experiment 2

In this experiment, we modified Experiment 1 slightly. Rather than presenting three motor primes or three phonological primes, the two critical conditions now both used phonological primes. One condition used three primes ("prime-3"), while the other condition used just one prime before the critical trial ("prime-1").

## Methods

Participants Thirty-two undergraduates at UC Merced volunteered to participate and received partial course credit. All participants were right-handed native speakers of English. 9 trials were excluded because of computer lag (0.8\%).

Stimuli and Procedure This experiment had a similar setup to Experiment 1 and made use of the same stimuli. The prime-3 condition was identical to the phonological-3 condition in Experiment 1. There were 18 filler items.

## Results

There were 9 errors in total ( $0.8 \%$ of all trials), 8 in the control condition ("fillers") and 1 in the prime- 3 condition. Crucially, this means that there were no noteworthy differences between the prime-3 and prime-1 error rates, indicating that no categorical A-not-B-like error was committed. Subsequent analyses will be performed on correct trials only.

Fig. 3 shows the evolution of the Euclidian distances from the diagonal line over time. There was a significant interaction between the prime-3 and the prime- 1 condition for the linear component of the growth curve model in the items analysis ( $\mathrm{p}<0.02$ ), and a nearly significant interaction in the subjects analysis ( $p=0.053$ ). This linear component indicates a steeper rise for the trajectory of the prime-3 condition than of the prime-1 condition. However, a likelihood ratio test comparing the full model with the factor "condition" against the model without indicates an only marginally significant overall effect of "prime-3 vs. prime1 " (subjects: $p=0.066$, items: $p=0.09$ ).

There were, however, significant differences between prime-3 and control both by subjects ( $\mathrm{p}=0.0008$ ) and by items ( $p=0.015$ ). This difference seems to stem from the linear component of the model (subjects: $\mathrm{p}=0.0018$, items: $\mathrm{p}=0.026$ ), and for the subjects analysis, there also was a significant difference in intercept ( $\mathrm{p}=0.0049$ ), reflecting overall larger Euclidian distances for the prime-3 condition than for trials without phonological competition and without previous movements towards the competitor. Finally, there was no significant difference of the overall model between prime -1 and the control condition (all p's $>0.1$ ). Comparing this to the effect of the prime-3 condition, this suggests that
the prime-3 trials did in fact deviate more strongly from control trials.

In terms of Dale et al. (2007)'s approach, there were 12 consecutive significant differences between prime-3 and prime-1 (time points 90 to 101) by subjects and none by items. There were 48 consecutive differences for prime-3 versus control (time points 54 to 101) by subjects, and 28 (74 to 101) by items. In contrast, there were 41 consecutive differences for prime-1 versus control (time points 20 to 34, and 64 to 101) by subjects and none by items. The shaded gray area in Fig. 2 shows the portions of prime-3 versus control that are significant in both the subjects and the items analysis.

By-subject bimodality coefficients for maximum deviation of the prime- 1 and prime-3 conditions were all below 0.555 for maximum deviations, indicating that the present results are unlikely due to averaging over bimodal responses. Bimodality coefficients for the area under the curve were above 0.555 for only three participants ( $\sim 9 \%$ ) in the 3-prime condition and for 4 participants (12.5\%) in the prime-1 condition.


Figure 3: Euclidian distance as a function of time. The shaded gray area indicates where prime-3 and control are significantly different from each other by subjects and items.

## General Discussion

In Experiment 1, we found a difference between a purely motor priming condition and a condition that had both a motor priming and a phonological priming aspect. This difference seems to lie within the higher-order polynomials of the curve fit, suggesting that complex details in the shape of the trajectories are of importance in characterizing the difference between the two conditions.. Experiment 2 further established that previous exposure to critical stimuli did affect the trajectory of upcoming trials, and there was indication that the strength of this effect was modulated by the number of previous priming trials.

In both experiments in comparison to the control condition, only the phonological-3 condition produced a
significant deviation in both the subjects and the items analysis simultaneously (shaded area Figs. 2 and 3). This was regardless of the two analysis approaches that we used above.

These results fall straightforwardly out of a dynamic field theory account of how the brain treats memory traces of objects and locations in general (Erlhagen \& Schöner, 2002; Spencer, Barich, Goldberg, \& Perone, 2012), and the A-notB error in particular (Smith \& Thelen, 2003; Thelen et al, 2001). However, in contrast to children, repeatedly clicking on a location ("A") did not lead to a categorical error (there were no significant differences in error rates between the conditions). Instead, there was evidence for a continuous, graded attraction toward the competitor.


Figure 4: Schematic activation patterns of the movement layer in a hypothetical dynamic field simulation of repeated left-moving trials (towards "A").

A similar type of finding comes from Diamond and Kirkham's (2005) adaptation of Zelazo, Frye and Rapus's (1996) Dimensional-Change Card Sort task, typically used with children. After doing multiple trials with one cardsorting rule, Diamond and Kirkham's participants were then told explicitly that the sorting rule has changed for the next block of trials. Children routinely make sorting errors on the first trial of the new block. Adults do not make categorical errors, but they do produce significantly longer reaction times on the first trial with the new sorting rule. Hindy and Spivey (2008) extended this finding by showing that adults also curve their mouse movements significantly toward the old rule's response option.

These results can also readily be interpreted from the theoretical perspective of deictic pointers or visual indices (Ballard et al., 1997) - such that peaks in the dynamic field may function as the pointers. Chun and Nakayama (2000) state that "...memory traces interact with attentional mechanisms to guide eye movements, cognition, and action." In the case of our experiment, these memory traces are built up from preceding trials, similar to the "preshaping" done by the previous trials in the A-not-B task. Fig. 4 provides a visual illustration of these cognitive processes. A peak on one trial builds a memory trace in the field that increases the activation of subsequent trials in that
location ("A"). In addition, inhibitory connections between A and B suppress the activation of the other ("B") location, making these peaks progressively smaller.
The observed difference between the motor priming condition and the phonological priming condition also provides clues as to how semantic tags get associated with deictic pointers. In the motor condition, the only information that is repeatedly associated with the spatial location is the movement. In contrast, the phonological condition had repeated phonological information and visual content in addition to the movement. These richer associations may help account for the greater spatial attraction in these trials.

## Conclusions

In Experiment 1, we showed that presentations of phonological cohort stimuli result in increased spatial attraction toward the competitor's location, even though the cohort is not simultaneously present on the critical trial. There was evidence that this spatial attraction increased due to a genuine effect of phonological competition. In Experiment 2, we modified the conditions to test the dynamic field theory prediction that multiple presentations of similar stimuli result in greater competition. We found evidence that the prime-3 condition resulted in increased spatial attraction toward the competitor in comparison to the prime-1. This can be interpreted as showing that phonological similarity and repeated presentations influence the landscape of the dynamic field. Dynamic field theory as applied to the A-not-B task readily predicts the observed results. For Experiment 1, phonological similarity should influence spatial attraction because the memory traces developed during the repeated trials are stronger than those without phonological similarity. In regards to Experiment 2, repeated presentations should also increase the strength of the memory trace (represented as a pre-shaped field), and cause increased spatial attraction.

Overall, this study and its results add to the literature by providing a indication of how dynamic field theory may be able to account for the data of an A-not-B like task in adults, as well as a novel way of investigating the formation of deictic pointers. The various processes of visual cognition and language in our experiment are spread out in time in such a way that each experimental trial is not independent of previous trials. The landscape of the dynamic field itself is an important theoretical construct for understanding these temporal dynamics. Taken together, these results are powerful support for the value of dynamic field theory modeling and the mouse-tracking experimental methodology. In the future, we intend to conduct additional control experiments within this research program as well as model human data explicitly with dynamic field theory simulations.

## Acknowledgments

We thank Teenie Matlock for making lab space available to us. We thank Robert Räder for helping out with figures.

## References

Ballard, D., Hayhoe, M., Pook, P., \& Rao, R. (1997). Deictic codes for the embodiment of cognition. Behavioral and Brain Sciences, 20, 723-767.
Bates, D., Maechler, M., \& Bolker, B. (2012). lme4: Linear mixed-effects models using S4 classes. R package version 0.999999-0.

Chun, M., \& Nakayama, K. (2000). On the functional role of implicit visual memory for the adaptive deployment of attention across scenes. Visual Cognition, 7(1-3), 65-81.
Dale, R., \& Duran, N. D. (2011). The cognitive dynamics of negated sentence verification. Cognitive Science, 35, 983996.

Dale, R., Kehoe, C.E. \& Spivey, M.J. (2007). Graded motor responses in the time course of categorizing atypical exemplars. Memory \& Cognition, 35, 15-28.
Diamond, A. \& Kirkham, N. (2005). Not quite as grown up as we like to think: Parallels between cognition in childhood and adulthood. Psychological Science, 16, 291297.

Erlhagen, W., \& Schöner, G. (2002). Dynamic field theory of movement preparation. Psychological Review, 109, 545-572.
Freeman, J.B., \& Ambady, N. (2010). MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. Behavior Research Methods, 42, 226-241.
Freeman, J.B., Ambady, N., Rule, N.O., \& Johnson, K.L. (2008). Will a category cue attract you? Motor output reveals dynamic competition across person construal. Journal of Experimental Psychology: General, 137, 673690.

Freeman, J.B., Dale, R., \& Farmer, T.A. (2011). Hand in motion reveals mind in motion. Frontiers in Psychology 2,59. doi:10.3389/fpsyg.2011.00059
Hindy, N. \& Spivey, M. (2008). Motor dynamics of task switching. Proceedings of the 30th Annual Conference of the Cognitive Science Society. (pp. 2474-2479).
Marcovitch, S., \& Zelazo, P. (1999). The a-not-b error: Results from a logistic meta-analysis. Child Development, 70, 1297-1313.
Mirman, D., Dixon, J.A., \& Magnuson, J.S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. Journal of Memory and Language, 59, 475-494.
Piaget, J. (1954). The Construction of Reality in the Child. New York: Basic Books.
Richardson, D. C. \& Spivey, M. J. (2000). Representation, space and Hollywood Squares: Looking at things that aren't there anymore. Cognition, 76, 269-295.
Smith, L.B. \& Thelen, E. (2003) Development as a Dynamic System. Trends in Cognitive Science, 7, 343348.

Smith, L. B., Thelen, E., Titzer, R. \& McLin, D. (1999) Knowing in the Context of Acting: The Task Dynamics of the A-Not-B Error. Psychological Review, 106, 235-260.

Spencer, J. P., Barich, K., Goldberg, J., Perone, S. (2012). Behavioral dynamics and neural grounding of a dynamic field theory of multiple-object tracking. Journal of Integrative Neuroscience, 11, 339-362.
Spivey, M., Grosjean, M., \& Knoblich, G. (2005). Continuous attraction toward phonological competitors Proceedings of the National Academy of Sciences, 102, 10393-10398.
Thelen, E., Schöner, G., Scheier, C. \& Smith, L. B. (2001) The Dynamics of embodiment: A field theory of infant perseverative reaching. Behavioral and Brain Sciences, 24, 1-86.
Zelazo, P., Frye, D., \& Rapus, T. (1996). An age-related dissociation between knowing rules and using them. Cognitive Development, 11, 37-63.

# Communication and Categorization: New Insights into the Relation Between Speech, Labels and Concepts for Infants 

Brock Ferguson (brock@u.northwestern.edu)<br>Sandra R. Waxman (s-waxman@northwestern.edu)<br>Department of Psychology, Northwestern University, 2029 Sheridan Rd.<br>Evanston, IL 60208 USA


#### Abstract

Almost two decades of research has demonstrated that labels facilitate infants' categorization of novel objects. Some interpret this as evidence of an early link between infants' linguistic and conceptual systems. Others suggest that these effects stem exclusively from lower-level processing mechanisms in cross-modal perception, and that words promote categorization only because they are more familiar to infants than non-linguistic acoustic stimuli and therefore easier to process. Here we address these discrepant interpretations using a novel approach. We expose infants to unfamiliar non-linguistic stimuli (sine-wave tone sequences), manipulating the exposure conditions. For 6-month-olds, if the novel acoustic stimuli were embedded within a communicative episode, they subsequently facilitated categorization (Experiment 1), but if they were presented in a non-communicative episode, they had no such effect (Experiment 2). We propose a developmental model that takes infants' burgeoning perceptual and conceptual capacities into account in identifying how communication and words are linked to concepts.


Keywords: language development; words; concepts; categorization; auditory overshadowing; infancy

## Introduction

The nature of word learning has been the focus of a noteworthy debate in recent years. At stake is the relationship between words and concepts: Are words merely associated with objects by infants, as any percept might be associated with another (e.g., Sloutsky \& Fisher, 2012)? Or might even the youngest word learners appreciate words as symbols that refer to concepts (e.g., Waxman \& Gelman, 2009)? Further, if there is a privileged link between words and concepts in infancy, how is it established?

Evidence for this latter position, positing an early and unique link between words and concepts, comes from numerous studies demonstrating that infants integrate domain-specific knowledge about words when they map novel words to objects (Fennell \& Waxman, 2010; Namy \& Waxman, 2000; Woodward \& Hoyne, 1999), generalize words to object concepts (Booth \& Waxman, 2009; Booth, Waxman, \& Huang, 2005), make inferences about hidden properties of named objects (Diesendruck \& Graham, 2010; Gelman \& Heyman, 1999; Graham, Booth, \& Waxman, 2012), and individuate named objects (Dewar \& Xu, 2007; 2009).

There is also evidence for a developmental cascade underlying infants' establishment of a link between words and concepts. Initially, infants appear to hold a broad expectation that words refer to commonalities amongst objects (Waxman, 2003). With development, they refine this broad expectation to link particular types of words (e.g., nouns, adjectives) to particular types of categories (e.g., object categories, property categories) (Booth \& Waxman, 2009). This increasingly precise relation between words and concepts can be observed over the first year in object categorization tasks. Infants hearing human language successfully form categories, but other matched acoustic stimuli (e.g., sine-wave tone sequences) do not (Balaban \& Waxman, 1997; Fulkerson \& Haaf, 2003; Fulkerson \& Waxman, 2007; Waxman \& Markow, 1995). More recent evidence reveals that infants as young as 3- and 4-months (who do not yet segment distinct words from fluent speech) form object categories in the context of human speech, but not in the context of sine-wave tones (Ferry, Hespos \& Waxman, 2010). Thus over the first year, infants' response to words may be a refinement of a broader and earlier response to communicative signals.

Some researchers have argued that the influence of language in these studies reflects cross-modal perceptual processing alone (Robinson \& Sloutsky, 2007; Sloutsky \& Robinson, 2008). Their claims are clear: (1) object categorization tasks with paired acoustic stimuli recruit infants' cross-modal processing abilities, (2) unfamiliar auditory stimuli impede visual processing through "auditory overshadowing", and (3) verbal labels are more familiar to infants than the acoustic stimuli (e.g., tone sequences) to which they are typically compared (Sloutsky \& Robinson, 2008). On this account, words benefit category formation only insofar as they are acoustically familiar.

Here we take a novel empirical approach to tease apart these two accounts. In each experiment, infants participated in a standard object categorization task. But instead of pitting human language against unfamiliar sounds, all infants heard the same unfamiliar sounds: sine-wave tone sequences. Crucially, we introduced infants to these novel sounds in a video before they were presented within an object categorization task. This gave us full control over infants' prior exposure to these novel stimuli, which in turn permits us to ascertain the precise exposure conditions that enable an auditory stimulus to facilitate visual categorization. In Experiment 1, we ask whether embedding
tone sequences in a communicative episode will allow them to subsequently facilitate object categorization. In Experiment 2, we document that this effect cannot be accounted for by appealing to familiarity alone.

## Experiment 1

We introduced 6- and 12-month-old infants to novel acoustic stimuli (sine-wave tone sequences), embedding it in a clearly communicative episode. Next, we presented new tone sequences, this time within the context of the standard object categorization task. We asked whether tone sequences would now (like speech) facilitate categorization. If infants interpreted the novel tone sequences presented in the video as communicative, then tones may now promote categorization in the standard task. However, if infants do not privilege this novel signal with communicative status, or if they resist relating it to object categories, they should not form object categories in the standard task.

We expected the consequences of our manipulation to differ as a function of infants' age. At 6 months, we expected that embedding tones in a rich communicative episode would be sufficient to facilitate categorization but that, by 12 months, infants would require more specific evidence that the signal is referential. This is consistent with evidence that by 12 months, infants distinguish referential from non-referential communicative utterances and only interpret the former as referring to object categories (Fennell \& Waxman, 2010; Hollich, Hirsh-Pasek, \& Golinkoff, 2000; Waxman \& Braun, 2005).

## Methods

Participants Twenty-four healthy, full-term infants participated. Participants included twelve 6-month-olds (6 males, $M=5.94$ months) and twelve 12-month-olds ( 6 males, $M=12.08$ months). Another 13 infants (seven 6-month-olds, six 12 -month-olds) were excluded due to looking for less than $25 \%$ of the familiarization or test phases (8), fussiness (3), or parental interference (2).

Stimuli The design included three phases: exposure, familiarization, and test (see Figure 1). In the exposure phase, infants saw a 2-minute video of two undergraduate women sitting next to each other engaged in a communicative exchange. The "beeper" appeared to produce sine wave tones that had been dubbed over her mouth movements. The "speaker" responded in infantdirected English. Both interlocutors alternated between looking and speaking towards each other and the infant.

In the familiarization phase, infants saw 8 images of members of a single object category (either dinosaurs or fish, counterbalanced). Each image was presented for 20s with 4 s between images. Images were line-drawn and filled with unique solid colours. Each image was paired with a single sine-wave tone sequence, presented at image onset and 10 s post-onset. This sequence ( 2.2 s ), which differed in pitch from the sequences presented in the dialogue, was matched
for pause-length and duration to the labeling phrases used in previous studies (e.g., Ferry et al., 2010).

In the test phase, infants saw two new images in silence for 20s. One image was another member of the familiar category (e.g., another fish), and the other a member of a novel category (e.g., a dinosaur). The left/right position of the novel image was counterbalanced.

Procedure Infants sat on their caregivers' laps approximately 110 cm from the centre of a screen. Auditory stimuli were played through two speakers placed 85 cm apart beneath the screen.

Coding Infants' left-right eye gaze directions were coded frame-by-frame by trained coders blind to the hypotheses. A second coder re-coded the videos to assess reliability (Pearson's $r=.97, p<.0001$ ).

Analyses We analyzed the first 10s of looking to either

| EXPOSURE PHASE |  |
| :---: | :---: |
| Experiment 1 | Experiment 2 |
| Sine wave tones and speech (coordinated with conversation) | Sine wave tones and speech (uncoupled from video) |
| FAMILIARIZATION PHASE (Experiments $1 \& 2$ ) |  |
| 8 images, each paired with the same sine wave tone sequence |  |
| TEST PHASE (Experiments 1 \& 2) |  |
|  |  |

Figure 1: Experimental design of Experiments 1 and 2. Procedure for Exposure, Familiarization and Test Phases, with a sample of representative stimuli.
object in the test phase, as in prior research. (An analysis of the complete test trial yielded the same pattern of results in both experiments.) For each infant, a novelty preference score was calculated as the proportion of looking towards the novel category member. All analyses used arcsin-root transformed proportions.

## Results

As predicted, 6-month-olds $(M=.61, S D=.15)$ had significantly higher novelty preference scores than 12-month-olds $(M=.49, S D=.19), t(22)=1.72$, one-tailed $p$ $<.05$. Planned comparisons to chance showed a clear novelty preference (evidence of categorization) for the 6-month-olds $(t(11)=2.51, p<.03)$ but not the 12-month-olds $(t(11)=-.19, n . s$.$) .$

There were no age differences in looking towards the familiarization images ( $M_{6 \text {-months }}=.43, S D=.15 ; M_{12 \text {-months }}$ $=.53, S D=.13), t(22)=1.71, p=.10$, and no effects of familiarized category, novel object side at test, familiarization looking time, or gender on novelty preference scores (all p's $>.3$ ). All analyses collapsed across these factors.

## Discussion

When infants were introduced to sine-wave tones during a brief communicative episode (dialogue phase), tone sequences then facilitated object categorization (test phase) for 6-month-olds, but not 12-month-olds.

This age difference is striking. We suggest that at both 6and 12 -months, infants flexibly identify the candidate communicative signals in their environment. At 6 months, infants hold a broad expectation linking communicative signals to object categories. But by 12 months, infants recognize the distinct functions of different communicative signals (e.g., speech versus gesture; Martin, Onishi, \& Vouloumanos, 2012; and naming an object versus merely indicating it (e.g., "wow"); Fennell \& Werker, 2003; Namy \& Waxman, 2000). Therefore, at 12 -months, evidence of communicative status alone is insufficient: Infants require more precise evidence that a novel signal is one that refers to objects and object categories.

But could an appeal to signal familiarity alone account for these results? There are two hints that it cannot. First, the particular pattern/pitch of the tones paired with each category member at test were novel (i.e., not presented during the dialogue phase). Second, although 6- and 12-month-olds' exposure to tone sequences was identical, only the 6-month-olds showed evidence of categorization, as we predicted

However, to further tease apart the two accounts, in Experiment 2 we pursue this issue with another group of 6-month-olds.

## Experiment 2

In this experiment, we exposed infants to the very same sine-wave tone sequences (exposure phase) as in Experiment 1, but this time uncoupled them from the
communicative context. During the exposure phase, infants listened to the same auditory signals as in Experiment 1, and saw a video with the same two women, but this time the women cooperated in a joint task in silence. Crucially, infants' exposure to the tones was held constant across both experiments, but in Experiment 2, there was no indication that the tones were part of a communicative interchange. If 6-month-olds' successful categorization in Experiment 1 reflects nothing more than their familiarity with tone sequences, then infants in Experiment 2 should also categorize successfully.

## Methods

Participants Twelve healthy, full-term, 6-month-old infants participated ( $M=5.87$ months). Another 4 infants were tested but excluded due to looking for less than $25 \%$ of the familiarization or test phases.

Stimuli The new exposure video showed two women silently engaged in a cooperative task (mixing ingredients and pouring them, as if making brownies together). They smiled to each other and the infant (as in Experiment 1), but did not communicate verbally. The audio stream included exactly the same "utterances" (tone sequences, English speech) as in Experiment 1, but these were randomly shuffled. (The goal was to remove the prosodic pattern of turn-taking in the spoken utterances that might lead infants to infer that the tone sequences were part of a conversation and therefore communicative). Familiarization and test stimuli were identical to Experiment 1.

## Procedure, Coding \& Analyses Identical to Experiment 1.

## Results

As predicted, 6-month-olds performed differently here than in Experiment $1(t(22)=2.16, p<.05)$. In contrast to Experiment 1, where 6-month-olds averaged a .61 novelty preference at test, those in Experiment 2 performed at the chance level ( $M=.48, t(11)=-0.45$, n.s. $)$.

There were no effects of familiarized category, novel object side at test, familiarization looking time, or gender on novelty preference scores (all $p$ 's > .4). All analyses collapsed across these factors.

## Discussion

These results reveal that mere familiarity with sine-wave tone sequences cannot account for their facilitative effect on object categorization in Experiment 1. Six-month-olds who received the same exposure to these sequences, uncoupled from the communicative episode, show no evidence of categorization.

## General Discussion

In these experiments, we introduce a novel approach for investigating classic questions about the nature of word learning: Are words perceptual features associated with
objects? Or is there a more nuanced link between words and concepts? And, if so, how is it established?

Waxman and colleagues (Balaban \& Waxman, 1997; Booth \& Waxman, 2003; Fulkerson \& Waxman, 2007; Waxman, 2003; Waxman \& Markow, 1995) have long argued for the latter position, and cite evidence that providing a consistent name for distinct members of an object category highlights the commonalities among them and promotes object categorization. On this account, language exerts its influence because infants link language to core conceptual capacities, including object categorization. In contrast, others have suggested that language facilitates categorization only insofar as it is a familiar acoustic stimulus (Robinson \& Sloutsky, 2007a; Sloutsky \& Robinson, 2008). On this account, any adequately familiar stimulus should show facilitative effects: the facilitative effect of an acoustic signal will vary as a function of its familiarity.

In Experiment 1, we asked whether an otherwise inert acoustic stimulus (sine wave tones), introduced within the context of a communicative episode, might facilitate categorization. Six-month-olds showed evidence of categorization, while 12-month-olds did not. In Experiment 2, we asked whether the 6-month-olds' successful categorization could be attributed to their mere exposure to the tone sequences. We provided the same amount of exposure to the sine wave tones, but uncoupled them from the communicative episode. The results were straightforward: infants in Experiment 2 revealed no evidence of categorization in the subsequent task. Stimulus familiarity alone cannot capture these results.

## Auditory overshadowing

Auditory overshadowing is a precise claim about low-level cross-modal processing, and it is relevant to many studies in infant cognitive development including object categorization (Robinson \& Sloutsky, 2007a) and individuation (Robinson \& Sloutsky, 2008). The general processing model invoked is uncontroversial: infants have limited cognitive resources and any stimulus that exhausts these resources will have consequences on subsequent processing.

Thus we do not ask whether auditory overshadowing could, in principle, influence infants' learning (about categories or otherwise), but whether it alone can account for infants' clear patterns of behaviour. The results here join a host of others in demonstrating that in addition to perceptual underpinnings, there are conceptual and socialcommunicative factors that determine whether a paired acoustic stimulus can facilitate object categorization.

Consider, for example, infants' developing knowledge of grammatical categories and its influence on categorization. By 14 months, novel nouns highlight object categories, but adjectives do not (Booth \& Waxman, 2009; Waxman \& Booth, 2001). Adopting an auditory overshadowing interpretation, Sloutsky and colleagues (Robinson \& Sloutsky, 2007a, 2008) argue that nouns are a more familiar
stimulus than adjectives, and thus interfere less with visual processing (Sloutsky \& Fisher, 2012). However, this explanation cannot account for the performance of younger infants ( 9 to 12 months), whose categorization improves when both adjectives and nouns are paired with category exemplars (Waxman \& Booth, 2003; Waxman \& Markow, 1995). Familiarity alone can neither explain this developmental change nor the results of the present studies.

## Communication, cognition, and "natural pedagogy"

Previous claims about the influence of language on categorization have focused primarily on the effect of words presented as labels for object categories (Waxman, 2003). More recent evidence suggests that, for younger infants, human speech more generally can facilitate categorization (Ferry et al., 2010). Three- and 4-month-olds show an increased ability to categorize in the context of human speech despite their inability to reliably segment the speech stream into discrete words (Jusczyk \& Aslin, 1995). The present results go further to suggest that for young infants, speech may be just one of a number of communicative signals that facilitate categorization. Infants in the present studies had no prior exposure to the sine wave tone sequences we presented, and yet merely introducing them as a human communicative signal had a powerful effect on their contribution to infants' subsequent categorization.

Why might communicative signals link to concepts? One recent proposal is that ostensive human communication is "naturally pedagogical" for infants, biasing them to interpret new information as category-relevant and generalizable (Csibra \& Gergely, 2009; Gergely \& Csibra, 2012). A recent study by Yoon, Johnson, and Csibra (2008) demonstrated the effect of communicative signals on cognition in 9-month-olds: in the context of a communicative gesture (pointing), infants most accurately encoded the shape of the object. In a non-communicative (grasping) context, infants most accurately encoded its location. Another study with 9 -month-olds reported object categorization benefits from eye gaze (Wu, Gopnik, Richardson, \& Kirkham, 2010). There is also evidence that communicative object labels enhance object recognition by augmenting core visual processes during encoding (Gliga, Volein, \& Csibra, 2010).

If one posits that the sine wave tone sequences in Experiment 1 were part of an ostensive communicative exchange with the infant (see Csibra, 2010, for a discussion of how infants recognize ostensive signals), the 6-montholds' results align with the theory of natural pedagogy: the presence of the communicative signal facilitated the discovery of category-relevant information.

## Tuning the perceptual and conceptual systems

Natural pedagogy cannot, however, explain the full developmental picture. For example, it cannot account for the results of the 12 -month-olds in Experiment 1. Neither can it explain why, for example, young infants accept gestural labels (like words) to refer to object categories but
older hearing infants do not (Namy \& Waxman, 1998; Suanda \& Namy, 2012), or why young infants map both nouns and adjectives to object categories but older infants are more precise, mapping nouns, and not adjectives, to object categories (Booth \& Waxman, 2009). We therefore suggest a more detailed developmental account. .

Our account builds on a substantial literature suggesting that infants begin life with broad perceptual sensitivities in a variety of social domains (e.g., faces, speech sounds, and hand gestures) but rapidly tune these to make functionally relevant distinctions (Grossmann, Missana, Friederici, \& Ghazanfar, 2012; Palmer, Fais, Golinkoff, \& Werker, 2012; Vouloumanos, Hauser, Werker, \& Martin, 2010; Werker \& Tees, 1984). In language development, this process of perceptual tuning is a critical step, for example, in focusing infants' attention on the signals that are potentially communicative (e.g., human speech) and tuning out those that are not (e.g., non-human primate vocalizations).

We suggest that infants also engage in a process of referential tuning in which they tease apart the particular functions of distinct communicative signals. For example, 12-month-olds expect that human speech, but not noncommunicative vocalizations (e.g., coughing) can refer (Martin et al., 2012). And within human speech, infants gradually distinguish between distinct types of words (nouns and adjectives) and map them accordingly to distinct types of meaning (e.g., to object categories and properties, respectively) (Waxman \& Booth, 2009; Waxman \& Gelman, 2009). In this ongoing, constructive process, infants recruit several knowledge systems (social, linguistic, and conceptual) to infer the intended reference and meaning of communicative signals. When a communicative signal is interpreted as intending to refer to an object category, it can serve to highlight that category and facilitate learning in young infants.

One prediction of the present account is that 12 -montholds' object categorization abilities should benefit from a novel communicative signal if they are given sufficient cues (i.e., beyond mere communicativeness) that the signal is meant to refer to an object or object category. We are currently testing this prediction.

Several other questions remain to be explored. First, this account posits an early expectation that communicative signals in general will relate to meaning in the world. This is consistent with natural pedagogy (Csibra \& Gergely, 2009). However, whether this expectation is innate or acquired prior to 6 months is presently unclear. Second, other studies that explore the influence of social cues in learning do not find a consistent benefit for social cues over non-social cues (e.g., Moore, Angelopoulos, Bennett, 1999; Theuring, Gredebäck, \& Hauf, 2007). Unlike the present experiments, these studies pit social cues against non-social cues in tasks with distractor and target events. Thus their failure to show benefits from social cues may reflect younger infants' limited capacities for inhibitory control and attention deployment. Future research in complex environments can examine this hypothesis and the constraints of learning in
communicative contexts. Finally, it is important to explore the range of conditions under which infants interpret a novel stimulus as communicative.

The present research integrates social, conceptual, and linguistic development for a rich description of infants' early communicative development. We suggest with others (e.g., Noles \& Gelman, 2012; Waxman \& Gelman, 2009) that words are not merely perceptual features that associate with objects, but are communicative symbols, and the products of early perceptual and conceptual tuning.

## Acknowledgements

This research was supported by a Social Sciences and Humanities Research Council of Canada Doctoral Fellowship to B.F., and a National Science Foundation grant to S.R.W. (BCS-0950376).

## References

Balaban, M. T., \& Waxman, S. R. (1997). Do words facilitate object categorization in 9 -month-old infants? Journal of Experimental Child Psychology, 64(1), 3-26.
Booth, A., \& Waxman, S. R. (2003). Mapping words to the world in infancy: Infants' expectations for count nouns and adjectives. Journal of Cognition and Development, 4(3), 357-381.
Booth, A., \& Waxman, S. R. (2009). A horse of a different color: Specifying with precision infants' mappings of novel nouns and adjectives. Child Development, 80(1), 15-22.
Booth, A., Waxman, S. R., \& Huang, Y. T. (2005). Conceptual information permeates word learning in infancy. Developmental Psychology, 41(3), 491-505.
Csibra, G. (2010). Recognizing communicative intentions in infancy. Mind \& Language, 25(2), 141-168.
Csibra, G., \& Gergely, G. (2009). Natural pedagogy. Trends in Cognitive Sciences, 13(4), 148-153.
Dewar, K., \& Xu, F. (2007). Do 9-month-old infants expect distinct words to refer to kinds? Developmental Psychology, 43(5), 1227-1238.
Dewar, K., \& Xu, F. (2009). Do early nouns refer to kinds or distinct shapes? Evidence from 10 -month-old infants. Psychological Science, 20(2), 252-257.
Diesendruck, G., \& Graham, S. A. (2010). Kind matters: A reply to Samuelson and Perone. Cognitive Development, 25, 149-153.
Fennell, C. T., \& Waxman, S. R. (2010). What paradox? Referential cues allow for infant use of phonetic detail in word learning. Child Development, 81(5), 1376-1383.
Fennell, C. T., \& Werker, J. F. (2003). Early word learners' ability to access phonetic detail in well-known words. Language and Speech, 46(2-3), 245-264.
Ferry, A. L., Hespos, S. J., \& Waxman, S. R. (2010). Categorization in 3- and 4-month-old infants: an advantage of words over tones. Child Development, 81(2), 472-479.
Fulkerson, A. L., \& Haaf, R. A. (2003). The influence of labels, non-labeling sounds, and source of auditory input
on 9- and 15-month-olds' object categorization. Infancy, 4(3), 349-369.
Fulkerson, A. L., \& Waxman, S. R. (2007). Words (but not tones) facilitate object categorization: Evidence from 6and 12-month-olds. Cognition, 105(1), 218-228.
Gelman, S. A., \& Heyman, G. D. (1999). Carrot-eaters and creature-believers: The effects of lexicalization on children's inferences about social categories. Psychological Science, 10(6), 489-493.
Gergely, G., \& Csibra, G. (2012). Natural pedagogy. In M. R. Banaji \& S. A. Gelman (Eds.), Navigating the Social World: What Infants, Children, and Other Species Can Teach Us. Oxford University Press.
Gliga, T., Volein, A., \& Csibra, G. (2010). Verbal labels modulate perceptual object processing in 1-year-old children. Journal of Cognitive Neuroscience, 22(12), 2781-2789.
Graham, S. A., Booth, A., \& Waxman, S. R. (2012). Words are not merely features: Only consistently applied nouns guide 4-year-olds' inferences about object categories. Language Learning and Development, 8(2), 136-145.
Grossmann, T., Missana, M., Friederici, A. D., \& Ghazanfar, A. A. (2012). Neural correlates of perceptual narrowing in cross-species face-voice matching. Developmental Science, 15(6), 830-839.
Hollich, G., Hirsh-Pasek, K., \& Golinkoff, R. M. (2000). Breaking the language barrier: An emergentist coalition model for the origins of word learning. Monographs of the Society for Research in Child Development.
Jusczyk, Peter W, \& Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. Cognitive Psychology, 29(1), 1-23.
Martin, A., Onishi, K. H., \& Vouloumanos, A. (2012). Understanding the abstract role of speech in communication at 12 months. Cognition, 123(1), 50-60.
Moore, C., Angelopoulos, M., \& Bennett, P. (1999). Word learning in the context of referential and salience cues. Developmental Psychology, 35(1), 60.
Namy, L. L., \& Waxman, S. R. (1998). Words and gestures: Infants' interpretations of different forms of symbolic reference. Child Development, 69(2), 295-308.
Namy, L. L., \& Waxman, S. R. (2000). Naming and exclaiming: Infants' sensitivity to naming contexts. Journal of Cognition and Development, 1(4), 405-428.
Noles, N. S., \& Gelman, S. A. (2012). Effects of categorical labels on similarity judgments: A critical analysis of similarity-based approaches. Developmental Psychology, 48(3), 890-896.
Palmer, S. B., Fais, L., Golinkoff, R. M., \& Werker, J. F. (2012). Perceptual narrowing of linguistic sign occurs in the 1st year of life. Child Development, 83(2), 543-553.
Robinson, C. W., \& Sloutsky, V. M. (2007). Visual processing speed: Effects of auditory input on visual processing. Developmental Science, 10(6), 734-740.
Robinson, C. W., \& Sloutsky, V. M. (2008). Effects of auditory input in individuation tasks. Developmental Science, 11(6), 869-881.

Sloutsky, V. M., \& Fisher, A. V. (2012). Linguistic labels: Conceptual markers or object features? Journal of Experimental Child Psychology, 111(1), 65-86.
Sloutsky, V., \& Robinson, C. (2008). The role of words and sounds in infants' visual processing: From overshadowing to attentional tuning. Cognitive Science, 32(2), 342-365.
Suanda, S. H., \& Namy, L. L. (2012). Young word learners' interpretations of words and symbolic gestures within the context of ambiguous reference. Child Development, 84(1), 143-153.
Theuring, C., Gredebäck, G., \& Hauf, P. (2007). Object processing during a joint gaze following task. European Journal of Developmental Psychology, 4(1), 65-79.
Vouloumanos, A., Hauser, M. D., Werker, J. F., \& Martin, A. (2010). The tuning of human neonates' preference for speech. Child Development, 81(2), 517-527.
Waxman, S. R. (2003). Links between object categorization and naming: Origins and emergence in human infants. In D. H. Rakison \& L. M. Oakes (Eds.), Early Category and Concept Development: Making Sense of the Blooming, Buzzing Confusion. New York: Oxford University Press.
Waxman, S. R., \& Booth, A. (2001). Seeing pink elephants: Fourteen-month-olds' interpretations of novel nouns and adjectives. Cognitive Psychology, 43(3), 217-242.
Waxman, S. R., \& Booth, A. (2003). The origins and evolution of links between word learning and conceptual organization: New evidence from 11-month-olds. Developmental Science, 6(2), 128-135.
Waxman, S. R., \& Braun, I. (2005). Consistent (but not variable) names as invitations to form object categories: New evidence from 12-month-old infants. Cognition, 95, B59-B68.
Waxman, S. R., \& Gelman, S. A. (2009). Early wordlearning entails reference, not merely associations. Trends in Cognitive Sciences, 13(6), 258-263.
Waxman, S. R., \& Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12- to 13-month-old infants. Cognitive Psychology, 29(3), 257-302.
Werker, J. F., \& Tees, R. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. Infant Behavior and Development, 7(1), 49-63.
Woodward, A. L., \& Hoyne, K. L. (1999). Infants' learning about words and sounds in relation to objects. Child Development, 70(1), 65-77.
Wu, R., Gopnik, A., Richardson, D. C., \& Kirkham, N. Z. (2010). Social cues support learning about objects from statistics in infancy. In S. Ohlsson \& R. Catrambone (Eds.), Proceedings of the 32nd Annual Conference of the Cognitive Science Society (pp. 1228-1233). Austin, TX: Cognitive Science Society.
Yoon, J. M. D., Johnson, M. H., \& Csibra, G. (2008). Communication-induced memory biases in preverbal infants. Proceedings of the National Academy of Sciences, 105(36), 13690-13695.

# Stopping Rule Selection (SRS) Theory Applied to Deferred Decision Making 

Mario Fifić (fificm@gvsu.edu)<br>Department of Psychology, Grand Valley State University<br>Allendale, MI 49501 USA<br>Marcus Buckmann (buckmann@ gmx.net)<br>Max Planck Institute for Human Development, Center for Adaptive Behavior and Cognition, Berlin Germany


#### Abstract

The critical step facing every decision maker is when to stop collecting evidence and proceed with the decision act. This is known as the stopping rule. Over the years, several unconnected explanations have been proposed that suggest nonoptimal approaches can account for some of the observable violations of the optimal stopping rule. The current research proposes a unifying explanation for these violations based on a new stopping rule selection (SRS) theory. The main innovation here is the assumption that a decision maker draws from a large set of different kinds of stopping rules and is not limited to using a single one. The SRS theory hypothesizes that there is a storage area for stopping rules-the so-called decision operative space (DOS)and a retrieval mechanism that is used to select stopping rules from the DOS. The SRS theory has shown good fit to challenging data published in the relevant literature.


Keywords: Stopping rule, deferred decision task, optimal, nonoptimal, decision making.

One of the most important steps of decision making is determining when to stop collecting evidence and proceed with the final decision. This is defined as the stopping rule and it is thought to be an irreplaceable component of almost all cognitive models of decision making.

Take, for example, a patient who is facing a risky medical treatment. The treatment can have a good outcome-that is, the patient will benefit from it-or it can have a bad outcome-that is, the patient will suffer serious side effects. To the patient's surprise, doctors don't have a unanimous opinion on whether the treatment is beneficial or harmful. Thus, the patient decides to ask for several doctors’ opinions. The patient collects either positive opinions (+1) in favor of the risky treatment or negative opinions ( -1 ) against the risky treatment. The total sum of evidence is defined as the critical difference, $d$. But how many opinions should he collect to reduce the risk of making the wrong decision? To help the patient with the decision, his best friend, a statistician, tells him that the number of opinions can be calculated based on the most optimal solution.

## The Optimal Stopping Rule for Evidence Accumulation and Deviations

The determination of the optimal stopping rule in statistical decision making has been examined in great detail by Wald (1947) and from the Bayesian perspective by

Edwards and colleagues (Edwards, 1965). The optimal Bayesian model defines the stopping rule as the minimization of the expected loss, $E(L)$ (De Groot, 1970). The rule prescribed by the optimal model is to continue collecting evidence and to stop only when the expected value of loss is equal to or lower than the expected loss associated with deferring the decision and collecting more evidence.

To calculate the optimal number of doctors the patient should consult, his friend the statistician acquired the conditional distributions of doctors' positive ( + ) opinions given that the treatment can be either beneficial or harmful, $P(+$ opinion | beneficial treatment), $P(+$ opinion | harmful treatment), and also the prior probabilities of beneficial and harmful treatments, $P$ (beneficial treatment) and $P$ (harmful treatment) (e.g., Edwards, 1965; Schechter, 1988). The statistician used all these probabilities to calculate the socalled posterior odds in favor of the hypothesis that the treatment is beneficial given the evidence acquired from $n$ number of doctors, $\Omega\left(\left.\frac{\text { benefical treatment }}{\text { harmful treatment }} \right\rvert\,\right.$ opinion $)$. The posterior odds would indicate the best decision for the finite number of collected doctors' opinions, if the costs and payoffs associated with the risky treatment and the expected diagnostic value of a single opinion are considered. Using mathematical software, the statistician got the number 3 as the optimal stopping rule value for that risky decision. This means that the patient should collect positive and negative doctors' opinion ( +1 s and -1 s ) as long as their cumulative sum $(d)$ is lower than the value of $d=+3$ or higher than the value $d=-3$. The patient should stop evidence collection and make a decision as soon as $d=3$, in which case the patient should accept the risky treatment, or $d=-3$, in which case the patient should reject the risky treatment (e.g., Schechter, 1988).

The relevant literature has revealed that humans do not use the optimal stopping rule. (1) In a deferred decision task in which subjects had the option to defer their decision until they had purchased new information, subjects bought either too little evidence (Phillips \& Edwards, 1966; Pitz, 1968) or too much evidence (Pitz, 1968) compared to the optimal model's predictions. (2) The critical difference value $d$ can change over the course of sampling evidence in a single trial (e.g., Busemeyer \& Rapoport, 1988; Pitz, 1968; Newell, 2005). Subjects tended to make final decisions on smaller critical difference values for larger sets of evidence. To
account for these results, the optimal model should adjust the critical difference value such that it decreases as more evidence is acquired (Pitz, 1968; Viviani, 1979). (3) Subjects frequently terminated evidence collection when the critical difference value was zero ( $d=0$; Pitz, 1968; Pitz, Reinhold, \& Geller, 1969). From the optimal Bayesian viewpoint, this means that decision makers made a final decision even though there was no evidence to support any decision. (4) It has also been shown that human decision makers sometimes stop on a nondiagnostic sequence of evidence (Busemeyer \& Rapoport, 1988). For example, after a series of three positive pieces of evidence the subjects stopped on a negative piece of evidence, $\{+,+,+,-$ $\}$, and made a decision that supported the positive evidence. Note that the last two pieces of evidence were nondiagnostic and stopping on such a pattern of evidence is logically inconsistent with the optimal model.

The optimal approach to decision making has suffered more general criticism. The optimal model can be successfully applied only when a decision maker possesses perfect knowledge of all aspects of a situation. Following Savage (1954) and Binmore (2009), perfect knowledge of an environment is possible if one resides in a so-called small world. Examples of a small world are a controlled laboratory experiment, a lottery, and certain games. In a small world a detailed statistical representation of the environment exists and an optimal model can predict the exact amount of evidence needed to be collected to find the optimal stopping value.

But most decision makers live in a large world. A large world is quite unpredictable and dynamic-it is constantly changing and it is almost impossible to form an exact statistical representation of such an environment. In a large world a decision maker has limited time to make decisions, possesses limited cognitive powers in terms of memory and attention, and usually acts inconsistently (Berg, Biele, \& Gigerenzer, 2008; Gigerenzer, 2008; Schooler \& Hertwig, 2005; Shanteau, 1992; Tversky \& Kahneman, 1974). It is unrealistic to expect that a decision maker living in a large world would be able to employ an optimal model to determine when to stop accumulating evidence. Alternative approaches have been aimed at exploring how to make effective decisions with a limited amount of information and a limited cognitive system.

## Bounded Rationality and Nonoptimal Stopping-Rule Models

According to the bounded rationality approach, making decisions involves simple decision strategies and shortcuts that allow for quick and effortless decisions (e.g., Gigerenzer, 2004). Boundedly rational models require neither exact statistical representation of the environment nor optimization. (For a review of different nonoptimal models for evidence collection see Busemeyer \& Rapoport, 1988; for examples see Fifić, Little, \& Nosofsky, 2010).

Boundedly rational models for determining stopping rules are more suited to real-life decision-making problems and cognitive limitations than is the optimal model. Let us return to our patient example. The patient started to question the optimal value $d=3$ after he learned that the conditional distributions used to estimate the doctors' diagnostic accuracies do not exist for his country. Instead, his friend the statistician used the data from another, much smaller country across the ocean. Not trusting the optimal solution ( $d=3$ ), the patient decided to use another rule. He decided to obtain five doctors' opinions and make his decision based on the majority. This is defined as the fixed-sample-size stopping rule ( $s=5$ in the example). A decision maker determines a fixed amount of evidence to be collected before the collection starts. Our patient may have used a five-opinion stopping rule before-years ago when he bought a car. Alternatively, the patient could rely on another useful cue-a streak of either positive or negative opinions. The patient could stop looking for more opinions after receiving three successive positive or negative doctor opinions ( $r=3$ ) and make a decision accordingly. This is defined as the runs stopping rule (cf. Audley \& Pike, 1965; Estes, 1960). In sports games the runs rule is also known as the hot or cold hand rule (Bar-Eli, Avugos, \& Raab, 2006; Gilovich, Vallone, \& Tversky, 1985; Wilke \& Barrett, 2009). A player who scores a streak of shots in a row is perceived to be "hot" and is a preferred shooter. A player who has a streak of misses is likewise perceived to be "cold."

Although boundedly rational models have been able to explain some observed deviations from the optimal predictions (for details see Busemeyer \& Rapoport, 1988), no single such model has been able to account for them all. Take, for example, the fixed-sample-size stopping rule, which can account for the finding that decision makers sometimes stop on a nondiagnostic sequence of evidence. This rule predicts that the probability of termination should be equal for nondiagnostic sequences of identical length. In contrast, it has been observed that subjects prefer some nondiagnostic sequences over others of the same length (Busemeyer \& Rapoport, 1988). The runs stopping rule can account for the finding that decision makers stop on $d=0$, for example $\{+,+,-,-\}$. To stop on that evidence, the stopping rule value for the negative evidence has to be set on two pieces of negative evidence ( $r=-2$ ). The stopping rule for positive evidence has to be set on a value larger than two pieces of positive evidence (say $r=+3$ ). However, the runs stopping rule has limited explanatory power (Busemeyer \& Rapoport, 1988). For example, it cannot explain stopping when streaks of evidence are missing. In general, more explanatory power is gained by combining several stopping rules (see Pitz et al., 1969) within one framework. We lack a systematic theory to tie together different stopping rules in a single framework for decision making. To remedy this theoretical gap, I propose the stopping rule selection (SRS) theory.

## The SRS Theory

The SRS theory provides the basis for a general approach to decision-making operations. This theory is consistent with the idea of a boundedly rational decision maker who utilizes simple decision rules in real time. In different environments, a decision maker acts adaptively, constantly looking for the best decision strategies, stopping rules, and critical values.

## A formal description of the SRS theory and proposed stopping rules.

The SRS theory aims to provide a unifying framework for the storage and retrieval of multiple stopping rules. It consists of three hypotheses.

Hypothesis 1: Multiple stopping rules. The SRS theory assumes that several different stopping rules can operate concurrently. Decision makers act adaptively to changes in the environment, not only by calibrating different stopping rule values (value criterion) but also by switching between different stopping rules if needed. In real life, multiple stopping rules can be combined in a complex fashion (e.g., Pitz et al., 1969). Take, for example, scoring in tennis: The winner of a tennis game is the player whose score is at least two points higher than the opponent's ( $d \geq 2$ ) and if at least four points have been won so far ( $s \geq 4$ ).

Hypothesis 2: Storage for stopping rules-the decision operative space (DOS). A major component of the SRS theory is a storage place for the stopping rules and their values, which is called the decision operative space (DOS). The DOS can be seen as a variant of an "adaptive toolbox," a collection of domain-specific specialized cognitive mechanisms for decision making built through evolution (Gigerenzer \& Todd, 1999; Payne, Bettman, \& Johnson, 1993; Todd, 1999). Unlike the toolbox concept, the DOS is conceptualized as a structured psychological space. The stopping rules stored in the DOS are sorted on two dimensions: the cognitive effort needed for a certain stopping rule, and the time needed to make a decision using a certain stopping rule (Figure 1A). Depending on the environment, a decision maker can use these two dimensions to estimate which decision tools are the most appropriate to use.

The time scale, on the $x$-axis, is defined as chronological time. The exact expected duration of each stopping rule can be calculated from an analytic expression (e.g., see Feller, 1957, p. 317; also Busemeyer \& Rapoport, 1988; Pitz, 1968; Pitz et al., 1969). Cognitive effort, on the $y$-axis, is defined as the processing complexity of a decision strategy and can be measured by the number of elementary information processes (EIPs, after Payne et al., 1993) engaged in making a decision. As shown in Figure 1A, each point in the DOS represents a stopping rule with a certain stopping value. Stopping values belonging to the same
stopping rule lie on one line: For the runs stopping rule it is $r$, for the critical difference rule, $d$, and for the fixed-samplesize rule, $s$. Overall decision accuracy increases as one chooses as one chooses larger values for the stopping rules. However, the price of improvement is increases in both time and cognitive effort. As depicted in Figure 1, two stopping rules - the critical difference and the fixed-sample-size-are estimated to be of approximately the same complexity. They share the same EIPs, which are counting, differencing, averaging, and memory engagement. They differ on the time needed to complete the operations. The critical difference stopping rule needs more time to finish than the fixed-sample-size rule, for the same critical value. The runs stopping rule uses EIPs that are far simpler than those used by the previous two. To detect runs, a decision maker has only to count evidence, with minimal memory. Although based on simple EIPs, the runs stopping rule requires considerably more waiting time for larger critical values of runs.


Figure 1: (A) The decision operative space (DOS) for three stopping rules. Each point represents a single stopping rule with a stopping value. A straight line connects the same stopping rule with different stopping values. (B) A cast-net retrieval from the DOS. Dotted circles represent three different cast nets.

Hypothesis 3: Retrieval of the stopping rules. A retrieval mechanisms called "cast-net" retrieval is proposed. (cf satisficing approach; Todd, 1997; Todd \& Miller, 1999).

Cast-net retrieval. Selection of stopping rules resembles throwing a cast net and catching fish. A decision maker acts much like a fisherman, casting a net into the operative space. Here, on each throw the catch is a subset of possible stopping rules. To behave adaptively in different environments, decision makers adjust the location in the DOS where the net will be cast, and the size of the net. A decision maker who is not familiar with the environment or encounters much uncertainty in evidence collection may cast a larger net. If familiar with the environment, the decision maker may throw a smaller net. The larger the net is, the more different stopping rules are collected to make a single decision. The SRS theory specifies how several stopping rules could be used simultaneously to make a final decision.

The second property of the cast-net retrieval approach is the double tradeoff. Depending on where stopping rules are retrieved from the DOS, a decision maker may choose to trade off speed and accuracy (cf. Diederich, 2003; Kocher \& Sutter, 2006; Payne et al., 1993) or cognitive effort and accuracy (Payne et al., 1993). Figure 1B shows examples of both tradeoffs. Three cast-net locations are marked by red circles. Moving upward from the lower left circle on the vertical "work harder" path indicates a cognitive effortaccuracy tradeoff, keeping the time value constant. A decision to move vertically in the DOS means choosing to sacrifice frugality of effort to achieve better accuracy. A decision maker works harder to improve overall decision accuracy, as mainly the critical difference stopping rule is sampled. Moving from the lower left circle on the horizontal "take longer" path indicates a speed-accuracy tradeoff, keeping the cognitive effort value constant. A decision to move horizontally means choosing to sacrifice speed to achieve better accuracy. A decision maker takes longer, as mainly the runs stopping rule is sampled. The two tradeoffs can be used to explain adaptive decision making. Under the condition of increased uncertainty, it is expected that a decision maker would increase cognitive effort, and take the "work harder" path. Under time pressure, it is expected that a decision maker would use less time-consuming stopping rules and follow the "take longer" path.

## The SRS Theory: A Walkthrough of the Decision Process

In this section I provide a walkthrough of the decision process behind the SRS theory using the cast-net retrieval approach. The SRS model has two stages. The first stage is characterized by the selection and retrieval of stopping rules and their stopping values. The second is characterized by sequential evidence collection and application of stoppingrule criteria. The process is broken into six steps, three in the first stage and three in the second.

Step 1: Select hypotheses. Depending on the decision problem, a decision maker chooses the choice hypotheses (e.g., Thomas, Dougherty, Sprenger, \& Harbison, 2008). For example, in the patient decision situation described above, the two hypotheses $\mathrm{H}_{1}$ and $\mathrm{H}_{2}$ could be about the risky treatment: $\mathrm{H}_{1}$ : The risky treatment is a beneficial procedure, and $\mathrm{H}_{2}$ : The risky treatment is a harmful procedure. This stage is not under the scrutiny of the SRS model.

Step 2: Cast a net. The plethora of stopping rules and their values presents a challenge for the selection process. To select a subset of the stopping rules and their values, a decision maker throws a cast net into the DOS. To determine the position of the cast net and its span, a decision maker estimates how much time and cognitive effort can be invested in making the decision (on time and cognitive effort dimensions). These position estimates can be influenced by knowledge the decision maker possesses about this particular environment or similar ones. If no knowledge is available then a random starting point can be chosen. For illustration, assume that the following set of rules determines the cast $\{r=1, r=2, s=2, s=3, d=3, d=2\}$.

Step 3: Select a stopping rule. Once the DOS has been reduced by casting a net, several stopping rules and their values are randomly sampled from the net. All stopping rules and their values contained within the net can be retrieved with the same probability, defined by the probability density function $f(x)=\frac{1}{\text { number of retrieved rules }}$. For example, a decision maker could select the following set of stopping rules and their values from the cast net: $\{r=2$, $s=2, d=3\}$. Alternatively the probability of retrieving a certain rule from the cast net can be described by the bivariate normal distribution, $\mathbf{x} \sim(N(\boldsymbol{\mu}, \Sigma))$ (where the bold symbols are vectors), allowing rules that are closer to the center of a net to be retrieved with a higher probability than rules that are caught around the edges of the net.

Step 4: Collect evidence. The second stage starts with evidence accumulation. This step is repeated until a decision is made.

Step 5: Check stopping rule. The SRS model tests whether the evidence accumulated so far meets one of the criteria of the stopping rule selected from the net in Step 3. Assume that the model performs a serial test across three selected stopping rules. If none of the criteria have been met the decision maker looks for more evidence and repeats from Step 4. If any of the stopping value criteria are met, the decision maker stops evidence collection and proceeds with making the final decision.

Step 6: Stop and make a decision according to the hypothesis that was supported by the evidence.

## Face Validity of the SRS Theory: Preliminary Work and Results of Fitting

To establish face validity, I fit the SRS model to challenging data sets published in two separate studies on
determining stopping rules (Busemeyer \& Rapoport, 1988; Pitz, 1968). Our preliminary work showed that the SRS computational model can provide an excellent account of reported human data patterns. It is able to account for between $93 \%$ and $100 \%$ of the variability of Pitz's (1968) data and for about $86 \%$ of observed evidence patterns in Busemeyer and Rapoport's (1988) data. The model has 6 parameters describing the "span" of the cast net. Two parameters for each stopping rule define the range of the stopping rule values captured. As a part of the future exploration is the goal to reduce the number of parameters to only three describing the location and the size of the net.

In addition to showing high fitting accuracy, the SRS model was able to account for all four findings that falsified the optimal approach, described above: (1) People bought too much or too little evidence (Pitz, 1968); (2) the value of the critical difference ( $d$ ) could change over the course of sampling evidence in a single trial (e.g., Busemeyer \& Rapoport, 1988; Pitz, 1968); (3) people terminated evidence collection when the critical difference was zero ( $d=0$; Pitz et al., 1969); and (4) people stopped on nondiagnostic patterns. Regarding the accumulation of evidence, the observed data depart from the optimal model predictions (Table 1): For smaller values of $d$, the subjects collected too much evidence; for larger values of $d$, the subjects collected too little evidence. The SRS model captures this observed data trend as shown in the SRS model-fitting data. Regarding the value of the critical difference ( $d$ ), as can be seen in Table 1, less evidence was needed for larger values of $d$ to terminate evidence collection, compared to the optimal model prediction. This trend is accounted for by the SRS model fit. Regarding the termination of evidence collection when the critical difference was zero ( $d=0$ ), again as seen in Table 1, the SRS model shows that $n>0$ for $d=0$. Finally, regarding stopping on nondiagnostic patterns, the SRS model can also predict the nondiagnostic sequence of evidence (see Table 2). The SRS model fitted the observed patterns $\{1,1,1,0\}$ and $\{0,0,0,1\}$ (see Table 2; remember that 1 stands for positive and 0 for negative evidence). Note that the last two pieces of evidence in each pattern provide the nondiagnostic information for the optimal model.

Table 1: The average number of pieces of evidence ( $n$, shown in the table's cells) collected as a function of critical difference ( $d$ ) for three source reliability values ( $p=.8, .7$, and .6). The observed column shows averaged observed human data (from Pitz, 1968). The SRS column shows the best fit values when the stopping rule selection (SRS) model is fitted to the observed data. The optimal column shows the $n$ values predicted by the optimal model. The $r^{2}$ values are the proportions of explained variability the SRS model can account for.

|  | Source reliability $p=.8$ |  |  |
| :---: | :---: | :---: | :---: |
| $d$ | $r^{2}=1$ |  |  |
|  | Observed | SRS | Optimal |
| $\mathbf{0}$ | 2.73 | 2.71 | 0 |
| $\mathbf{1}$ | 2.75 | 2.8 | 1 |
| $\mathbf{2}$ | 3 | 2.92 | 2.93 |
| $\mathbf{3}$ | 3.67 | 3.59 | 4.71 |
| $\mathbf{4}$ | 5.04 | 5 | 6.41 |


|  | Source reliability $p=.7$ |  |  |
| :---: | :---: | :---: | :---: |
| $d$ | $r^{2}=0.98$ |  |  |
|  | Observed | SRS | Optimal |
| $\mathbf{0}$ | 3.56 | 3.92 | 0 |
| $\mathbf{1}$ | 3.42 | 3.65 | 1 |
| $\mathbf{2}$ | 4.47 | 4.21 | 3.43 |
| $\mathbf{3}$ | 6.07 | 6 | 6.13 |
| $\mathbf{4}$ | 6.64 | 6.53 | 8.86 |


|  | Source reliability $p=.6$ |  |  |
| :---: | :---: | :---: | :---: |
| $d$ | $r^{2}=0.93$ |  |  |
|  | Observed | SRS | Optimal |
| $\mathbf{0}$ | 3.05 | 3.89 | 0 |
| $\mathbf{1}$ | 4.43 | 4.51 | 1 |
| $\mathbf{2}$ | 5.2 | 4.75 | 3.84 |
| $\mathbf{3}$ | 4.74 | 5 | 8.05 |
| $\mathbf{4}$ | 7.12 | 6.86 | 13.37 |

Table 2: The results of the SRS model fit to Busemeyer and Rapoport (1988) data, from the constant cost condition of their Experiment 2. Table shows the matching patterns correctly recognized by the SRS model, as well as the nonmatching patterns. Evidence refers to the observed patterns of evidence prior decision making, where " 1 " and " 0 " stand for positive and negative opinions (recommendations). Response accuracy refers to whether the final decision based on collected evidence was correct. Observed refers to the observed proportion of each pattern. SRS fit refers to the best fitted proportions by the SRS model.

| Evidence | Response <br> accuracy | Observed | SRS fit |
| :--- | :--- | :--- | :--- |
| Observed matched patterns |  |  |  |
| $\{1,1\}$ | Correct | 0.06 | 0.1 |
| $\{0,0\}$ | Correct | 0.07 | 0.1 |
| $\{1,1,1\}$ | Correct | 0.19 | 0.17 |
| $\{0,0,0\}$ | Correct | 0.18 | 0.16 |
| $\{1,0,1,1\}$ | Correct | 0.05 | 0.04 |
| $\{0,1,1,1\}$ | Correct | 0.05 | 0.04 |
| $\{1,1,1,1\}$ | Correct | 0.08 | 0.07 |
| $\{1,1,1,0\}$ | Correct | 0.001 | 0.01 |
| $\{1,1,0,1\}$ | Correct | 0.05 | 0.03 |
| $\{1,1,0,0\}$ | Incorrect | 0.001 | 0.01 |
| $\{1,0,0,0\}$ | Correct | 0.07 | 0.04 |


| Evidence | Response <br> accuracy | Observed | SRS fit |
| :---: | :---: | :---: | :---: |
| $\{0,0,0,0\}$ | Correct | 0.06 | 0.07 |
| $\{0,1,0,0\}$ | Correct | 0.06 | 0.04 |
| $\{0,0,1,0\}$ | Correct | 0.05 | 0.03 |
| $\{0,0,0,1\}$ | Correct | 0.01 | 0.01 |
| Observed nonmatched patterns |  |  |  |
| $\{0,0,1\}$ | Incorrect | 0.002388 | 0 |
| $\{0,1,1\}$ | Correct | 0.009817 | 0 |
| $\{1,0,0\}$ | Correct | 0.002786 | 0 |

## Acknowledgments

This research was supported by the NSF grant (SES1156681) PI: Mario Fifić, Title "Stopping Rule Selection Theory, 2012-2015.

## References

Audley, R. J., \& Pike, A. R. (1965). Some alternative stochastic models of choice. British Journal of Mathematical \& Statistical Psychology, 18, 207-225.
Bar-Eli, M., Avugos, S., \& Raab, M. (2006). Twenty years of "hot hand" research: Review and critique. Psychology of Sport and Exercise, 7, 525-553.
Berg, N., Biele, G., \& Gigerenzer, G. (2008). Consistency versus accuracy of beliefs: Economists surveyed about PSA. [Meeting Abstract]. International Journal of Psychology, 43, 31.
Binmore, K. (2009). Rational decisions. Princeton, NJ: Princeton University Press.
Busemeyer, J. R., \& Rapoport, A. (1988). Psychological models of deferred decision making. Journal of Mathematical Psychology, 32, 91-134.
DeGroot, M. H. (1970). Optimal statistical decisions. New York, NY: McGraw-Hill.
Diederich, A. (2003). MDFT account of decision making under time pressure. Psychonomic Bulletin \& Review, 10, 157-166.
Edwards, W. (1965). Optimal strategies for seeking information: Models for statistics, choice reaction times, and human information processing. Journal of Mathematical Psychology, 2, 312-329.
Estes, W. K. (1960). A random walk model for choice behavior. In K. J. Arrow, S. Karlin, \& P. Suppes (Eds.), Mathematical methods in the social sciences (pp. 265-276). Stanford, CA: Stanford University Press.
Feller, W. (1957). An introduction to probability theory and its applications (2nd ed.). New York, NY: Wiley.
Fific, M., Little, D. R., \& Nosofsky, R. M. (2010). Logical-rule models of classification response times: A synthesis of mentalarchitecture, random-walk, and decision-bound approaches. Psychological Review, 117, 309-348.
Gigerenzer, G. (2004). Fast and frugal heuristics: The tools of bounded rationality. In D. Koehler \& N. Harvey (Eds.), Blackwell handbook of judgment and decision making (pp. 6288). Oxford, UK: Blackwell.

Gigerenzer G. (2007). Gut feelings: The intelligence of the unconscious. New York, NY: Viking.
Gigerenzer, G. (2008). Why heuristics work. Perspectives on Psychological Science, 3, 20-29.
Gigerenzer, G., \& Gaissmaier, W. (2011). Heuristic decision making. Annual Review of Psychology, 62, 451-482.

Gigerenzer, G., \& Todd, P. M. (1999). Fast and frugal heuristics: The adaptive toolbox. In G. Gigerenzer, P. M. Todd, \& the ABC Research Group, Simple heuristics that make us smart (pp. 3-34). New York, NY: Oxford University Press.
Gilovich, T., Vallone, R., \& Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. Cognitive Psychology, 17, 295-314.
Kocher, M. G., \& Sutter, M. (2006). Time is money-Time pressure, incentives, and the quality of decision-making. Journal of Economic Behavior \& Organization, 61, 375-392.
Newell, B. R. (2005). Re-visions of rationality? Trends in Cognitive Sciences, 9, 11-15.
Payne, J. W., Bettman, J. R., \& Johnson, E. J. (1993). The adaptive decision maker. Cambridge, UK: Cambridge University Press.
Phillips, L. D., \& Edwards, W. (1966). Conservatism in a simple probability inference task. Journal of Experimental Psychology, 72, 346-357.
Pitz, G. F. (1968). Information seeking when available information is limited. Journal of Experimental Psychology, 76, 25-34.
Pitz, G. F., Reinhold, H., \& Geller, E. S. (1969). Strategies of information seeking in deferred decision making. Organizational Behavior and Human Performance, 4, 1-19.
Savage, L. J. (1954). The foundations of statistics. New York, NY: Wiley.
Schechter, C. B. (1988). Sequential analysis in a Bayesian model of diastolic blood pressure measurement. Medical Decision Making, 8, 191-196.
Schooler, L. J., \& Hertwig, R. (2005). How forgetting aids heuristic inference. Psychological Review, 112, 610-628.
Shanteau, J. (1992). How much information does an expert use? Is it relevant? Acta Psychologica, 81, 75-86.
Thomas, R. P., Dougherty, M. R., Sprenger, A. M., \& Harbison, J. I. (2008). Diagnostic hypothesis generation and human judgment. Psychological Review, 115, 155-185.
Todd, P. M. (1997). Searching for the next best mate. In R. Conte, R. Hegselmann, \& P. Terna (Eds.), Simulating social phenomena (pp. 419-436). Berlin, Germany: Springer.
Todd, P. M. (1999). Simple inference heuristics versus complex decision machines. Minds and Machines, 9, 461-477.
Todd, P. M., \& Miller, G. F. (1999). From pride and prejudice to persuasion: Satisficing in mate search. In G. Gigerenzer, P. M. Todd, \& the ABC Research Group, Simple Heuristics that make us smart (pp. 287-308). New York, NY: Oxford University Press.
Tversky, A., \& Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185, 1124-1131.
Viviani, P. (1979). Choice reaction times for temporal numerosity. Journal of Experimental Psychology: Human Perception and Performance, 5, 157-167.
Wald, A. (1947). Sequential analysis. New York, NY: Wiley.
Wilke, A., \& Barrett, H. C. (2009). The hot hand phenomenon as a cognitive adaption to clumped resources. Evolution and Human Behavior, 30, 161-169.

# Syntactic priming in language comprehension allows linguistic expectations to converge on the statistics of the input 

Alex B. Fine (afine@bcs.rochester.edu)<br>Department of Brain \& Cognitive Sciences and Department of Linguistics, University of Rochester Rochester, NY 14627 USA<br>T. Florian Jaeger (fjaeger@bcs.rochester.edu)<br>Department of Brain \& Cognitive Sciences and Department of Computer Science, University of Rochester<br>Rochester, NY 14627 USA


#### Abstract

Human language is characterized by variability in that the way in which language is used varies depending, for example, on facts about the identity of the speaker or author, the social context, and surrounding linguistic material. Variability poses formidable challenges to the systems underlying language comprehension, which are known to exploit statistical contingencies in the input to overcome the inherent noisiness of perception; nevertheless, we seem to comprehend language with apparent ease. How is this possible? Here we argue that we are able to comprehend language efficiently in part by continuously adapting to the statistics of novel linguistic situations. We argue further that adaptation specifically allows comprehenders' expectations to converge towards the actual statistics of the linguistic input. Concretely, we show that readers can adjust their linguistic expectations in light of recent experience such that (a) previously difficult structures become easier to process, and, even more strikingly, (b) previously easy to process structures come to incur a processing cost.


Keywords: Sentence processing; experience-based language processing; parsing; reading; learning; adaptation; priming

## Introduction

Human language is variable in the sense that the way in which language is used varies across situations according to, for example, the social context, the surrounding linguistic material, and various facts about the identity of the speaker or author. Variability in this sense pervades our linguistic experience, and has been observed at virtually every level of linguistic representation.

Despite the extent to which language use varies, communication is typically successful. That is, even when faced with novel speakers or accents, we seem to be able to quickly and accurately infer the messages intended by our interlocutors.

Our apparent facility with language is particularly remarkable considering the extent to which linguistic experience has been demonstrated to play a role in language processing. Experience-based accounts of language processing hold that comprehenders generate expectationsabout the probability of observing particular sounds, words, sentence structures, etc.-during online language processing, and that these expectations are informed by and reflect the statistics of previous linguistic experience. By generating expectations that reflect the actual distribution of
events in the environment, comprehenders should, in principle, be able to reduce the average prediction error experienced during online processing, and thus process language efficiently. But if the distribution of words or sentence structures varies according to individual speakers, dialects, etc., then, at first blush, it is no longer clear that generating online linguistic expectations that reflect aggregate statistics over previous experience would be advantageous to the comprehender. How do we comprehend language as well as we do despite variability in the linguistic signal?

Here we present evidence that comprehenders are able to rapidly adapt to or implicitly learn the statistics of novel linguistic situations, focusing specifically on sentence comprehension ("parsing"). We argue that syntactic adaptation allows comprehenders' expectations about the statistics of the input to converge towards the actual statistics, providing an explanation for why experiencebased processing is advantageous despite the variability present in the statistics of the signal. Our experiments build on and attempt to synthesize insights from three lines of research that have till now proceeded largely in parallel: (1) experience-based language processing (e.g., MacDonald, Pearlmutter, \& Seidenberg, 1994), (2) syntactic priming (e.g., Traxler, 2008), and (3) research exploring the link between online processing and implicit learning (e.g., Misyak \& Christiansen, 2012; Wells, Christiansen, Race, Acheson, \& MacDonald, 2009).

To test our hypothesis, we exploit a well-known temporary syntactic ambiguity that provides a window onto comprehenders' expectations, illustrated in (1).

1. The experienced soldiers...
a. ...warned about the dangers before the midnight raid.
b....spoke about the dangers before the midnight raid.
c. ...warned about the dangers conducted the midnight raid.
d. ... who were warned about the dangers conducted the midnight raid.

Verbs like warned give rise to temporary ambiguities since they may occur both as the main verb (MV) of a
sentence ((1a)) or as the verb in a relative clause (RC; (1c)). Sentences (1a) and (1c) can be disambiguated toward the RC reading at conducted, like in (1c). By contrast, (1b) is unambiguously an MV structure because spoke is unambiguously a past tense matrix verb; (1d) is unambiguously an RC because of the relativizer who, which serves as an early disambiguating cue. Sentences like (1c) consistently elicit what are known as ambiguity or gardenpath effects: reading times (RTs) in the disambiguating region (in bold) spike when the ambiguity is resolved towards the relative clause interpretation (1c), compared to unambiguous RCs (1d). No such ambiguity effect is found for ambiguous compared to unambiguous MVs. Experienced-based accounts predict the garden-path effect because verbs like warned are overwhelmingly more likely to occur with MVs than RCs in subjects' previous experience, as evidenced in corpora of written and spoken language.

Given that this frequency difference has a reliable correlate in human behavior, we can take advantage of the MV-RC ambiguity to explore syntactic adaptation. We provide subjects with experience with written language in which the environment-specific syntactic statistics differ sharply from subjects' previous experience with language. If subjects are adapting to the statistics of the input, as we propose, then the manner in which subjects process these structures should change over the course of the experiment. Specifically, if exposed to locally stationary syntactic distributions-i.e., distributions whose parameters remain fixed within the environment-comprehenders' syntactic expectations should converge towards the statistics of the environment. In Experiment 1, we find evidence for rapid, incremental, and cumulative syntactic adaptation: over the course of an experiment where RCs are many times more likely than in subjects' previous experience, the ambiguity effect for these structures continuously decreases until it disappears. Experiment 2 goes a step further. There, we reason that if subjects indeed adapt their expectations to converge towards the statistics of the input, then as subjects come to assign a higher subjective probability to RCs, they should commensurately come to assign a lower subjective probability to MVs. Since MVs by hypothesis compete with RCs for probability mass (in the type of garden path sentences we investigate) it should be possible to make MVs sufficiently unlikely that this structure would actually come to incur a processing cost. This is what we find.

## Experiment 1

In Experiment 1 we ask whether comprehenders can rapidly adjust their syntactic expectations in response to the statistics of a novel linguistic situation (i.e., in response to the statistics of the experiment). We expose subjects to 40 ambiguous and unambiguous RCs, as in (1c) and (1d). Because RCs are infrequent structures, we predict that subjects will display an initially high processing cost for ambiguous relative to unambiguous RCs (i.e., a large ambiguity effect), but as the experiment progresses, and
evidence accumulates that RCs are highly probable within the context of the experiment, we predict that the ambiguity effect should diminish.

## Subjects

80 subjects were recruited via Amazon's Mechanical Turk platform. Only subjects with US IP addresses were allowed to participate. Additionally, instructions clearly indicated that subjects were required to be native speakers of English, and only subjects with at least a $95 \%$ approval rating from previous jobs were included.

## Materials

Critical items were constructed from sentence pairs like (1c) and (1d). Eight different verbs giving rise to the MV/RC ambiguity (watched, washed, taught, served, called, warned, dropped, pushed) were repeated 5 times to yield 40 critical items (only the verbs were repeated; the remainder of the sentences differed between items). Ambiguity was counter-balanced across two experimental lists. In addition, each list contained the same 80 fillers. Filler sentences featured a variety of syntactic structures and, crucially, did not include verbs that give rise to the MV/RC ambiguity (e.g., All the undergraduates in the class had trouble keeping up; The foreign delegates arrived at the embassy surrounded by security guards). Both lists presented stimuli in the same, pseudo-randomized order with 1-3 fillers between each critical item. Two additional lists were created in which the order of items was reversed, yielding a total of 4 orders.

## Procedure

Stimuli were presented in a self-paced moving window display. At the beginning of each trial, the sentence appeared on the screen with all non-space characters replaced by a dash. Subjects pressed the space bar using their dominant hand to view each consecutive word in the sentence. Durations between space bar presses were recorded. At each press of the space bar, the currently viewed word reverted to dashes as the next word was converted to letters. A yes/no comprehension question followed all experimental and filler sentences, with the correct answer to half of all comprehension questions being "yes".

## Results

RTs less than 100 ms or greater than 2000 ms were excluded before computing length-corrected RTs (i.e., RTs with the effect of word length removed) following a procedure similar to the one described in Ferreira and Clifton (1986).

Length-corrected RTs during the disambiguating region (in bold in (1) above) were regressed, using mixed effects regression, onto the full factorial design (i.e., all main effects and interactions) of ambiguity (ambiguous vs. unambiguous) and item order (coded 1-40 and centered). Item order captures the number of RCs observed at a given point in the experiment. Additionally, we included a main
effect of log-transformed stimulus order, which provides an index of how many trials (including both critical items and fillers) have been read at a given point in the experiment. Stimulus order captures the effect of "task adaptation", i.e., general speed-up effects, which can be rather strong in selfpaced reading experiments (all results reported below hold with or without this predictor, and regardless of whether it is log-transformed). For this and all other analyses reported in this paper, we included the maximal random effects structure justified by the data (Jaeger, 2009).

We replicated the significant main effect of ambiguity found in previous studies: RTs in the disambiguating region were greater for ambiguous relative to unambiguous sentences $(\beta=19, \mathrm{p}<.001)$. Also replicating previous work, we found a significant main effect of log stimulus order ( $\beta=-39, \mathrm{p}<.05$ ) and a marginally significant main effect of item order $(\beta=-2, p=.09)$. That is, subjects read stimuli increasingly faster as the experiment progressed, presumably reflecting task adaptation effect (getting used to the self-paced reading paradigm, Fine, Qian, Jaeger, \& Jacobs, 2010) Crucially, there was a significant two-way interaction between ambiguity and item order: the processing cost incurred by ambiguous RCs-the ambiguity effect-significantly diminished as experience with RCs accumulated ( $\beta=-1, \mathrm{p}<.05$ ). In Figure 1, we visualize this interaction by plotting mean length-corrected RTs for ambiguous and unambiguous sentences across four bins of item order, and by plotting the ambiguity effect at all 40 points in the course of the experiment. Both the ambiguity effect and its interaction with item order were observed only in the disambiguating region. The effect of stimulus order was significant or marginally significant in all sentence regions.


Figure 1: Mean length-corrected RTs during the disambiguating region for ambiguous and unambiguous RCs across four bins of item order in Experiment 1, with embedded visualization of the change in ambiguity effect
across the course of the experiment. Error bars give 95\% confidence intervals on the mean.

## Discussion

Experiment 1 demonstrates that comprehenders are capable of rapidly, incrementally, and cumulatively adapting to the statistics of a novel linguistic environment, even after controlling for the effect of practice or task adaptation.

In the Introduction we articulated a conceptualization of syntactic adaptation according to which subjects continuously adjust their expectations such that their expectations about the linguistic environment converge towards the statistics of the linguistic environment. The results of Experiment 1 are compatible with such an interpretation, but do not rule out other plausible ones. For example, it is possible that the results of Experiment 1 are driven by boosts in the base-level activation of the RC structure, but that this happens without specific reference to the statistics of the input (Pickering \& Garrod, 2004), or that adaptation occurs by virtue of episodic memory for the repeatedly encountered structure, which similarly would not need to make reference to the statistics of the environment (Kaschak \& Glenberg, 2004). In Experiment 2, we present a more direct test of the prediction that comprehenders adjust their expectations to converge towards the statistics of the input.

## Experiment 2

Experiment 2 exploits the same temporary ambiguity between MVs and RCs used in Experiment 1. However, unlike in Experiment 1, we expose subjects to both RCs and MVs. As we mentioned above, the ambiguity effect observed for sentences like (1) is driven by large differences in the probabilities of the two structures: upon observing the string The experienced soldiers warned..., subjects have a stronger a priori expectation for an MV interpretation relative to the RC interpretation. In other words, MVs and RCs compete for probability mass: MVs receive a high subjective probability at the expense of RCs. Therefore, if the results of Experiment 1 are driven by convergence towards the statistics of the input, then as subjects come to find RCs more probable, they should also, in turn, find MVs less probable. This effect should be observable in a decreased ambiguity effect for RCs and an increased ambiguity effect for MVs as the experiment progresses.

Experiment 2 employs a between-subject block design to test this prediction. In this experiment, subjects were assigned to one of two groups, which we will call the FillerFirst and the RC-First groups. Subjects were exposed to three blocks of sentences. The composition of the materials in each block, for each group, is shown in Table 1.

Table 1: Summary of the between-subject, block design of Experiment 2.

| Group | Block 1 | Block 2 | Block 3 |
| :---: | :---: | :---: | :---: |
| $\begin{aligned} & \text { RC-First } \\ & (\mathrm{n}=40) \end{aligned}$ | $\begin{aligned} & 16 \quad \mathrm{RCs} \\ & \text { ambiguous) } \end{aligned}$ | $\begin{aligned} & 10 \quad \mathrm{RCs} \quad(5 \\ & \text { ambiguous) } \\ & 20 \text { fillers } \end{aligned}$ | $\begin{aligned} & 10 \quad \text { MVs } \quad(5 \\ & \text { ambiguous })+ \\ & 15 \text { fillers } \end{aligned}$ |
| FillerFirst ( $\mathrm{n}=40$ ) | 16 fillers | $\begin{aligned} & 10 \quad \text { RCs } \quad(5 \\ & \text { ambiguous })+ \\ & 20 \text { fillers } \end{aligned}$ | $\begin{aligned} & 10 \quad \text { MVs } \quad(5 \\ & \text { ambiguous })+ \\ & 15 \text { fillers } \end{aligned}$ |

We conducted Experiment 2 with three specific predictions in mind. We predict (1) that the ambiguity effect for RCs will be diminished from block 1 to block 2 for the RC-first group. This would conceptually replicate Experiment 1. We further predict (2) that the ambiguity effect for RCs during block 2 for the Filler-First group will be greater than that of the RC-first group. If the effects observed in Experiment 1 are due to task adaptation or fatigue, then the ambiguity effect for RCs in Block 2 should be the same for both the Filler-First and the RC-First group. In other words, reading a given number of sentences should have the same effect on reading times regardless of the content of those sentences. Finally, and most crucially, we predict (3) that the ambiguity effect for MVs should increase as experience with RCs increases. If adaptation is a matter of subjects' expectations converging on the statistics of the input, then as the ambiguity effect for RCs decreases, the ambiguity effect for MVs should increase. Thus, we predict a greater ambiguity effect for MVs in block 3 for the RC-First group (where subjects have encountered more RCs by the time they reach block 3) relative to the Filler-First group.

## Subjects

80 subjects were recruited from the University of Rochester community. Informed consent was obtained from all subjects according to the University's scientific research ethics policies. Subjects received $\$ 10$ for their participation.

## Materials

Subjects read a total of 71 sentences over 3 blocks (as outlined in Table 1). RC and MV sentences were created that followed the same template as the critical items from Experiment 1. Two experimental lists were constructed for each group that counter-balanced the conditions (ambiguous vs. unambiguous) for the sentence type (MV or RC) used within each block, totaling four lists. It is important to note that the block structure of the experiment was entirely implicit. From the perspective of the subjects, they simply read 71 sentences without breaks.

## Procedure

The same procedure as in Experiment 1.

## Results

RTs less than 100 ms or greater than 2000 ms were excluded before computing length-corrected RTs, as in Experiment 1. We tested three predictions that follow from the hypothesis that readers adapt to the local statistics of the linguistic environment, enumerated above.

Prediction 1 (does the ambiguity effect in the RC-First group diminish from block 1 to block 2?): We regressed length-corrected RTs during the disambiguating region (underlined in (1)) of sentences read during blocks 1 and 2 in the RC-First group onto ambiguity (ambiguous vs. unambiguous), block (block 1 vs. block 2), and the two-way interaction between these predictors. There was a significant effect of ambiguity ( $\beta=65, \mathrm{p}<.05$ ): ambiguous RCs were read more slowly than unambiguous RCs. There was also a significant main effect of block ( $\beta=-72, \mathrm{p}<.05$ ): subjects read faster during the second block relative to the first block. Finally, the interaction between these two variables, capturing the change in the ambiguity effect from block 1 to block 2, was in the predicted direction and trended towards but did not reach significance $(\beta=18, \mathrm{p}=.2)$. It is likely that the binned comparison of reading times across blocks 1 and 2, combined with fewer observations than in Experiment 1, provides less power than the treatment of RCs as a continuous variable in Experiment 1. To address this, we took data from blocks 1 and 2 for the RC-First group and submitted it to the same analysis reported for Experiment 1. We examined length-corrected RTs during the disambiguating region using the same analysis as in Experiment 1. All critical effects from Experiment 1 were replicated including, importantly, a two-way interaction between ambiguity and item order ( $\beta=2, \mathrm{p}<.05$, after Bonferroni correction for multiple comparisons), replicating Experiment 1.

Prediction 2 (is the ambiguity effect in block 2 greater for the RC-First group than for the Filler-First group?): we regressed length-corrected RTs during the disambiguating region onto group (RC-First vs. Filler-First), ambiguity (ambiguous vs. unambiguous), and the interaction between these two variables. Again, there was a main effect of ambiguity $\mathrm{RCs}(\beta=19, \mathrm{p}<.05)$. There was also a main effect of group: subjects in the RC-First group had overall faster reading times $(\beta=-7, \mathrm{p}<.05)$. Crucially, the two-way interaction between ambiguity and group was marginally significant ( $\beta=-5, \mathrm{p}=.08$ ): the ambiguity effect was smaller in the RC-First group than in the Filler-First group. That is, reading a block of filler sentences does not reduce the processing cost of RCs to the same extent that reading a block of RCs does. This result is shown by the pairs of bars corresponding to block 2 for both groups in Figure 2.

Prediction 3 (is the ambiguity effect for MVs in block 3 greater for subjects who have seen more RCs, i.e. for the RC-First group?): We regressed length-corrected RTs during the disambiguating region of sentences read during block 3 onto ambiguity (ambiguous MV vs. unambiguous MV), group (RC-First vs. Filler-First), and the interaction between these variables. There was a main effect of ambiguity, such that ambiguous MVs were read more
slowly than unambiguous MVs ( $\beta=8, \mathrm{p}<.05$ ). The main effect of group did not reach significance $(\beta=4, \mathrm{p}=.3)$. Crucially, the two-way interaction between ambiguity and group was significant ( $\beta=5, \mathrm{p}<.05$ ): the ambiguity effect for MVs during block 3 was greater for the RC-First group than for the Filler-First group. In other words, subjects who read more RCs subsequently experienced both (1) a reduction in the ambiguity effect for RCs and (2) an increase in the ambiguity effect for MVs. This pattern is visualized in Figure 2 in the right-most pair of bars for each group.


Figure 2: Mean length-corrected RTs during the disambiguating region for ambiguous and unambiguous conditions across all three blocks of Experiment 2. Error bars give $95 \%$ confidence intervals on the mean.

## Discussion

Experiment 2 was designed to further address the hypothesis that comprehenders adjust their syntactic expectations to converge towards the statistics of the input. Specifically, we predicted that, since RCs and MVs compete with each other for probability mass, when subjects come to assign a higher probability to one structure, they should come to assign a lower probability to the other. In Experiment 2, this led to the concrete prediction that a diminished ambiguity effect for RCs should lead to a larger ambiguity effect for MVs, and that this should be greater for the RC-first relative to the Filler-first group. This is what we observed (cf. Figure 2).

## General Discussion

We tested the hypothesis that language comprehenders are able to adapt their syntactic expectations to novel linguistic environments according to the statistics of those environments. In two reading experiments, we provided subjects with experience with distributions of syntactic structures that diverged sharply from their previous experience with English. We predicted that subjects would adapt their expectations (as reflected in changes in RTs)
according to their cumulative recent experience. As predicted, in Experiment 1 subjects came to process a priori infrequent structures that had initially produced longer RTs more quickly when those structures were frequent in the experiment. Experiment 2 replicated this and went a step further: there, subjects not only came to process an a priori infrequent structure more quickly, but also came to process an a priori frequent structure more slowly when it was infrequent in the experiment. Our experiments suggest that readers are capable of adapting to the relative frequencies (/probabilities) of syntactic structures in the current linguistic environment. The results of our experiments have implications for questions concerning the mechanisms underlying language comprehension and for debates about the mechanism underlying syntactic priming. We discuss these in turn.

Previous work on syntactic adaptation has demonstrated that exposure to syntactic structures can have immediate (Traxler, 2008) and cumulative (Kaschak \& Glenberg, 2004) effects on language comprehension, that these effects can be indexed to individual talkers (Kamide, 2012), and that the effects may endure for several days (Wells et al., 2009). Moreover, work on statistical learning has demonstrated a remarkable capacity in children and adults to rapidly extract statistical regularities in novel artificial languages (cf. Gómez \& Gerken, 2000), and has suggested that statistical learning may correlate with language processing in general (Misyak \& Christiansen, 2012). As mentioned in the introduction, however, previous work on experience-based processing, syntactic priming, and statistical learning has all proceeded largely in parallel, and has left open the question of how the immediate effect of experience on language comprehension accumulates over time to give rise to cumulative priming, experience-based processing effects, and environment-specific adaptation. We have attempted to build on all of this work by demonstrating that syntactic adaptation can be profitably construed as the rapid, incremental, and cumulative convergence towards the statistics of a novel linguistic environment. Syntactic adaptation of the kind observed here may therefore offer a route by which the immediate effects of experience ("priming") accumulate to give rise to long-term experience-based processing.

Our results also speak to ongoing debates surrounding the type of mechanism that underlies syntactic priming. Two main views have emerged from previous work. Transient activation accounts hold that priming results from a shortlived boost in the activation of a syntactic representation (Pickering \& Garrod, 2004). By contrast, implicit learning accounts hold that priming is a consequence of an implicit learning mechanism (Chang, Dell, \& Bock, 2006). We believe that implicit learning accounts cover the current results most naturally for at least two reasons. First, subjects in both experiments were sensitive to the cumulative statistics of the environment: the degree to which subjects' expectations for a structure had changed at a given point in the experiment depends on how many times
subjects saw (a) that structure and (b) other structures competing with it for probability mass. To the extent that transient activation accounts do not predict cumulative priming and insofar as learning accounts do (cf. Kaschak, Loney, \& Borreggine, 2006), our results appear to support an implicit learning account. Second, our results provide indirect evidence for error-sensitivity: we observed changes in RTs over the course of both experiments for both RCs and MVs, but changes of a greater magnitude for RCs relative to MVs (see Figure 2): observing a low-probability linguistic event (and therefore one with a relatively large error signal) leads to greater changes in RTs. Errorsensitivity has been argued to be a hallmark of implicit learning (Chang et al., 2006; Fine \& Jaeger, 2013; Jaeger \& Snider, 2013).

Taken together with recent work on adaptation in phonetics and pragmatics (Kurumada, Brown, \& Tanenhaus, 2012; Norris, McQueen, \& Cutler, 2003), our results suggest that adaptation is likely to be a general property of language processing, and a manifestation of a general ability to cope with a dynamic environment.

Finally, our findings demonstrate the fundamental role that experience plays in language processing. Our work suggests that not only is language processing influenced by aggregated prior experience (e.g., MacDonald et al., 1994), but that experience incrementally and rapidly shapes our expectations about the language we speak, thereby allowing us to comprehend language more efficiently.

## Acknowledgments

We gratefully acknowledge Andrew Watts and Narissa Gogos for their assistance in collecting the data reported in this paper. The work presented here was partially funded by an NSF GRF to ABF and NSF IIS-1150028 as well as an Alfred P. Sloan Fellowship to TFJ.

## References

Chang, F., Dell, G. S., \& Bock, K. (2006). Becoming syntactic. Psychological review, 113(2), 234-72.
Ferreira, F., \& Clifton, C. E. (1986). The independence of syntactic processing. Journal of Memory and Language, 25, 348-368.
Fine, A. B., \& Jaeger, T. F. (2013). Evidence for implicit learning in syntactic comprehension. Cognitive Science, 37(3), 578-591.
Fine, A. B., Qian, T., Jaeger, T. F., \& Jacobs, R. A. (2010). Syntactic Adaptation in Language Comprehension. Workshop on Cognitive Modeling and Computational Linguistics (pp. 18-26). Uppsala, Sweden: ACL.
Gómez, R. L., \& Gerken, L. (2000). Infant artificial language learning and language acquisition. Trends in Cognitive Sciences, 4, 178-186.
Jaeger, T. F. (2009). Random effect: Should I stay or should I go? Retrieved from http://hlplab.wordpress.com/2009/05/14/random-effect-structure/

Jaeger, T. F., \& Snider, N. E. (2013). Alignment as a consequence of expectation adaptation : syntactic priming is affected by the prime' s prediction error given both prior and recent experience. Cognition, 127(1), 57-83.
Kamide, Y. (2012). Learning individual talkers' structural preferences. Cognition, 124(1), 66-71.
Kaschak, M. P., \& Glenberg, A. M. (2004). This construction needs learned. Journal of experimental psychology. General, 133(3), 450-67.
Kaschak, M. P., Loney, R. A., \& Borreggine, K. L. (2006). Recent experience affects the strength of structural priming. Cognition, 99, B73-B82.
Kurumada, C., Brown, M., \& Tanenhaus, M. K. (2012). Pragmatic interpretation of speech: It looks like speech adaptation. CogSci 2012. Sapporo, Japan.
MacDonald, M. C., Pearlmutter, N., \& Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. Psych. Review, 101, 676-703.
Misyak, J. B., \& Christiansen, M. H. (2012). Statistical Learning and Language: An Individual Differences Study. Language Learning, 62(1), 302-331.
Norris, D., McQueen, J. M., \& Cutler, A. (2003). Perceptual learning in speech. Cognitive Psychology, 47(2), 204-238.
Pickering, M. J., \& Garrod, S. (2004). Toward a mechanistic psychology of dialogue. Behavioral and Brain Sciences, 27(2), 169-226.
Traxler, M. J. (2008). Lexically independent priming in online sentence comprehension. Psychonomic Bulletin \& Review, 15(1), 149-155.
Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J., \& MacDonald, M. C. (2009). Experience and sentence processing: statistical learning and relative clause comprehension. Cognitive psychology, 58(2), 250-71.

# How high can you jump? <br> Children's Perception of Affordances for Self and Others with Different Abilities 

Justin M. Fine (jmfine@asu.edu)<br>Department of Psychology, Arizona State University<br>Tempe, AZ 85287 USA<br>Heidi Kloos (heidi.kloos@uc.edu)<br>Department of Psychology, University of Cincinnati Cincinnati, OH 45221 USA

Mona M. Jenkins (jenkinmo@mail.uc.edu)<br>Department of Psychology, University of Cincinnati<br>Cincinnati, OH 45221 USA


#### Abstract

The current study investigated whether children and adults can distinguish between actions they are afforded and those afforded to an actor. Participants judged the maximum height they could reach while jumping and they judged the maximum height that the actor could reach while jumping. They did so with and without a weighted backpack, and they did so with and without walking several laps. Results show that before the addition of the weighted backpack, participants rated the actor's abilities as much closer to their own. While wearing the weighted backpack and then walking with it, participants' estimates decreased for themselves, but remained mostly unchanged for the adult.


Keywords: social perception; affordances; agency; embodiment; ecological psychology; development

## Introduction

For individuals to successfully navigate their environment, they must be able to perceive when different actions are possible. How does an individual know whether they can reach a jar from a shelf, step over a barrier, or navigate through traffic without incident? The ability to perceive potential actions is not limited to the individual's actions. Daily activities are filled with social interactions, such as conversational turn-taking (Shockley, Santana, \& Fowler, 2003), helping someone lift an object (Richardson, Marsh, \& Baron, 2007), or detecting whether two people can fit through a doorway (Davis, Riley, Shockley, Cummins-Sebree, 2010). Because people can readily interact and coordinate with other individuals, this suggests that individuals can perceive the actions afforded others and groups of people working together.

In the case of social interaction however, the perceiver doesn't necessarily have a priori information about another person's action capabilities. Two approaches to this problem-the ecological and embodiment perspectivecontend that minimally, perception serves the purpose of guiding action. The ecological approach focuses on the physical and spatial relationship of an observer to the
environment. The embodied approach claims that individuals neurally simulate (Grush, 2004) how they or another might accomplish an action. Both theories' ability to explain social perception in a jumping estimation task was tested in the current study.

## Affordances as the Object of Perception

Several researchers studying how individuals perceive possibilities for action in their environment have narrowed in on Gibson's (1979) concept of affordances. An affordance is meant to capture the relationship of an individual's morphology and action capabilities to the spatial layout of the environment and objects. For an individual to detect an affordance is to perceive an opportunity for action.

Affordance detection is seen in behaviors such as stair climbing (Warren, 1984) or chair sitting (Mark, 1987). Warren (1984) found that individuals selection of the tallest climbable stair is best described by a nearly, invariant ratio of leg-length to stair-riser height. Rather than focusing solely on riser height information, estimates are predicted by a ratio that exists only as a function of perceiver and stair.

Individual's daily routine rarely consist of just solo actions. For example, soccer players must decide whether their teammates are in the correct position to receive a pass. The natural tendency towards such social coordination suggests individuals readily detect what actions other people are afforded; people can accurately report what objects others can reach (Rochat, 1995), lift and move together (Richardson, Marsh, \& Baron, 2007), what chairs another person can sit on (Stoffregen, Gorday, Sheng, \& Flynn, 1999), and how high another person can jump and reach (Ramenzoni, Riley, Shockley, \& Davis, 2008a). This detection ability suggests information is readily available regarding the perceived person and their environment. Stoffregen et al. (1999) found that observers use affordance based information to detect the possible sitting height for other individuals. They asked individuals to watch a video of an actor standing next to a chair and estimate the maximal and preferred sitting height for the actor. As long as the spatial relationship between the
actor and apparatus was preserved, participants could accurately estimate the heights. Estimates were also accurate when participants only saw a kinematic display. The estimates were found to be most accurate when they were scaled by the leg length of the actor in the video, not the participant.

In this case, the estimates are based on scaling the physical morphology of the person to the spatial layout. Studies have also shown that people can perceive the capabilities for others to produce actions that are scaled by biomechanical properties such as jumping to reach an object (Ramenzoni et al., 2008a). In this case, it is less clear what information an observer might use to form a perception about another person's ability.

## Simulations and the Embodied Perceiver

An alternative perspective on social perception and action rests on neurologically driven mechanisms as a basis for behavior. This approach has been brought under the banner of the Common Coding (Prinz, 1997) or Embodied Simulation (Grush, 2004) approach. This approach suggests that social behaviors are explainable by a proposed overlap in how individuals represent perceived and performed actions. In other words, if a person watches an action being performed or plans to produce an action, they simulate the motor program and sensory consequences underlying that action. Simulation behavior is akin to covert imitation behavior (Wilson \& Knoblich, 2005).

The mirror neuron system is thought to underlie such perception and action overlap (Rizzolatti \& Craighero, 2004). The finding that mirror neurons, found in the F5 area of a Macaque monkey's premotor cortex, activate similarly to the viewing and production of an action (e.g., reaching for a glass), provide a mechanism for simulations. The perception of action possibilities in the embodied stance, thus, relies on neural based representations of the observer in the environment.

Behavioral support is found in stimulus-response incompatibility studies and action perception studies. For example, Brass, Bekkering, Wohlschlager, \& Prinz (2000) showed that finger movement reaction times are slower after watching a video of a hand performing the opposite of the instructed movement. They propose this is due to neural interference. Upon seeing the stimulus cue to respond, participants automatically simulate the action they saw; this creates a delayed response due to the overlap between the intended and observed action.

Researchers have suggested that such overlap between perception and action may provide a basis for understanding many social behaviors (Gallese, Keysers, \& Rizzolatti, 2004; Sebanz \& Knoblich, 2009). Knoblich and Jordan (2002) postulate that the mirror neuron system and embodied simulations support the ability to predict potential actions and their outcomes for perceivers and other people. Simulations are derived through a perceiver detecting or representing the actions they are afforded. These simulations are also used to judge the action capabilities of other individuals.

Results supporting a simulation theory of social perception have drawn on behavioral and physiological data. CalvoMerino, Glaser, Grezes, Passingham, and Haggard (2005) found greater activity in cortical regions containing mirror neurons when participants watched videos containing dance movements they were trained to perform. Individuals watching point-light displays are also more sensitive to movements produced by themselves than the movements of other people (Loula, Prasad, Harber, \& Shiffrar, 2005). These findings suggest that an observer's perception of another person's ability is derived from the observer's own capacity for action.

The proposal by Knoblich and Jordan (2002), regarding social action perception, suggests that the perceiver's estimation of other person's capacity to produce actions should be scaled to the perceiver's ability. Interestingly, Ramenzoni et al (2008a) found that putting weights on a participant reduced jump and reach estimates for themselves and an actor even though the actor was not wearing weights. This finding suggests that people may use simulations to estimate others, but use themselves as a frame of reference. It is not clear however, whether such estimates are really based on one's own ability to act per se, or are scaled by another relationship. Ramenzoni, Riley, Shockley, \& Davis (2008b) manipulated observer eye-height in another study as well. They found significant changes in the participants' estimates for themselves and the actor. These findings suggest that eye-height scaled information and embodied simulations both contribute to determining the ability to judge actions for others. Simulation behavior may provide a template for judgments while detection of eye-height or other optically specified information is used to tune those judgments.

## Study Overview

The current study examined whether a person's inherent and manipulated jumping ability affect their judgments of their own and another person's ability equivalently. Specifically, we tested whether individuals' judgments are based solely on their own ability to jump and reach an object or whether estimates are underpinned by simulations tuned by detecting optically specified information. In this case, the detectable information is the eye-height difference between the participant and another person. Thus, we predicted that an observer's estimation accuracy for another person should be related to the difference in eye-height of the perceiver and actor and the similarity of their inherent jumping abilities. If individuals only use simulations to make judgments, reducing observer's abilities should significantly reduce estimates for themselves and the actor. If estimates for the actor remain mostly unchanged, we predict that participants are using simulations tuned by differences in the eye-height between participant and actor.

To test the current predictions, we asked children and adults to estimate the maximum jumping abilities for themselves and an actor. Past studies (Ramenzoni et al., 2008a) have only used adult participants. This population
doesn't discriminate between groups who possess naturally different abilities and potentially different simulation capabilities. Both groups were used under the assumption that children naturally have lower jumping abilities then adults. Thus, they should have inherently different action capabilities to simulate. Participants had never seen the actor walk, jump, or reach for anything, removing any cues regarding the actor's biomechanical abilities, which has been shown to improve individual's judgments of other's (Ramenzoni, Riley, Davis, Shockley, \& Armstrong, 2008c). We manipulated participant's perception of their own jumping abilities and potentially the actors jumping abilities (Ramenzoni et al., 2008a) by increasing their weight. This was accomplished by having participants wear a backpack containing weights. Weighted estimates were provided before and after walking with the backpack.

## Methods

## Participants

Participants were 15 children between 4.5 and 5 years old ( $M$ $=4.9, S D=0.32$ ) and 15 adults between 18 and 24 years old ( $M=21, S D=2.5$ ). Children ranged in weight from 30 to 55 lbs $(M=48.6, S D=7.3)$, in height from 94 to $130 \mathrm{~cm}(M$ $=112, S D=9.3$ ), and in eye-height from 85 to $123 \mathrm{~cm}(M$ $=104, S D=9.6$ ). Adults ranged in weight from 141 to 210 lbs ( $M=159, S D=30.2$ ), in height from 162 to $195 \mathrm{~cm}(M=171$, $S D=7.7)$, and in eye-height from 152 to $185 \mathrm{~cm}(M=162$, $S D=8.6$ ). The actor had a weight of 140 lbs, height of 166 cm , and eye-height of 152 cm . All participants were either undergraduate students at the University of Cincinnati or children of undergraduates.

## Materials

To estimate jumpability, a figurine was suspended by a pulley and rope from the ceiling (see Figure 1). It could be lowered down a wall. Participants stood on a flat surface ( 100 cm x 100 cm ), 7 feet from the suspended object. The actor was positioned one foot to the left of the apparatus, facing the participant. The room was covered in black felt, including the background of where the figurine was suspended. Two adjustable backpacks, one adult-sized and one child-sized, were used to add weight to the participants. The weights used in the bag weighed 15 g each. The amount of weight used per person was approximately $5 \%$ of their body weight ( $\pm 15 \mathrm{~g}$ ). Participants were given help to put on the bags during the experiment.

## Procedure and Design

Participants were asked to play a guessing game. They were instructed to accurately estimate their own and another person's maximum ability to jump for the figurine. Prior to a trial, the figurine was lifted to the ceiling and then lowered down slowly. This was accomplished by an experimenter standing behind the wall and using the pulley system. The instruction was to tell the experimenter to stop lowering the figurine when it was at the reachable height. Participants
were allowed to have the experimenter adjust the apparatus, if the figurine was lowered too much. Estimates were coded using a tape measure drawn onto the wall. The figurine was then lifted to the top of the wall and a new trial started. Participants closed their eyes between each trial, preventing the usage of any spatial cues provided by resetting the apparatus. When estimates were made for the participant's own abilities, the actor was not in the room. When estimates were made for the actor, the actor stood next to the apparatus. A trial started by giving a verbal "go" signal, upon which the participant opened her/his eyes and the figurine was lowered.


Figure 1: Experimental setup. The girl represents the participant making a judgment about herself as well as the actor (the boy in front of the wall). The person behind the wall represents the experimenter lowering the figurine.

The experiment consisted of six types of trials, dependent on whether the participant was making estimates for themselves or the actor, and whether the participant had no weights (no-weights trial), had weights (weights-beforewalking trial), or had walked with weights (weights-afterwalking trial). In the no-weights condition, individuals made estimates from the designated spot. Participants were then given a backpack to wear (pre-weighed to approximately 5\% body weight), but were not allowed to move from the spot. After making their judgments, they were asked to walk 10 circular laps around the room. They then made the two remaining estimates. Two estimates were made for each trial type. The average of the two was used for the dependent variables. After all of the judgments, participants were asked to perform two jumps with the backpack on and two jumps without the backpack. The average across the two jumps of each kind was used to measure actual jumping abilities.

Because we wanted to examine the effects of going from a non-manipulated (no-weights) to a manipulated, but unadjusted (weights before walking), and adjusted (weights after walking) scenario, we did not counterbalance the order of trial type. We did however, counterbalance the order of the Person factor (self vs. other). Combining all of the factors, we utilized a mixed-design of Age Group (Child vs. Adult) x Person (Self vs. Actor) x Trial Type (no-weights, weights-before-walking, and weights-after-walking).

## Results

The following analyses present variables, described below, to analyze the actual jumping abilities, the mean estimated jumping height for the participant and actor (per trial type), and the estimation error (calculated as actual jump-height estimated ${ }_{j u m p}$-height). To examine whether optical information is related to jumpability estimates, the relationship of eyeheight difference to estimation error is also considered.

## Actual Jumping Abilities

First, we analyzed the actual jump height for the child and adult participants in a normal and weighted scenario. This measure was used to determine whether the weight manipulation actually affected jumping ability. Mean jumping height was analyzed with a 2 (Age Group: child, adult) x 2 (Condition: non-weighted, weighted) mixed-design ANOVA. As expected, there was a main effect of the between-group variable of Age Group, $F(1,28)=80.30, p<$ $.05, \eta_{p}^{2}=.74$. Overall, there were differences in the jumping abilities of the children ( $M=160.96, S D=23.66 \mathrm{~cm}$ ) and adult ( $M=227.80, S D=17.98 \mathrm{~cm}$ ) group. The main withingroup effect of Conditions was also significant, $F(1,28)=$ 33.10, $p<.05, \eta_{p}^{2}=.54$. In general, both groups exhibited similar changes in jumping without weights ( $M=195.70, S D$ $=40.88 \mathrm{~cm})$ and with weights ( $M=191.96, S D=39.42 \mathrm{~cm}$ ). The two-way interaction was not significant, suggesting both groups were similar in changes between non-weighted and weighted jumping ability ( $M=7.2, S D=2.34 \mathrm{~cm}$ ). The effect of Condition and lack of interaction reveals that the weights reduced participant's abilities similarly, regardless of Age Group.

## Estimated Jumping Abilities

To determine whether participants jumpability estimates for themselves and the actor were equivalently affected by the weight manipulation, we analyzed the participants' mean jumpability estimates using a 2 (Age Group: child, adult) x 2 (Estimated Person: participant, actor) x 3 (Condition: noweights, weights before walking, and weights after walking) mixed-design ANOVA.

The analysis revealed a significant three-way interaction of Age Group x Estimated Person x Condition, $F(1.69,47.57)=$ 4.94, $p<.05, \eta_{p}^{2}=.74$. Follow-up analyses were performed by splitting the Age Group (child and adult) factor into two separate 2 (Estimated Person: participant, actor) x 3 (Condition: no-weights, weights before walking, and weights after walking) repeated-measures ANOVAS. The results for the adult group yielded a significant interaction between Estimated Person and Condition, $F(1.62,22.71)=27.64, p<$ .05, $\eta_{p}^{2}=.66$. Simple effects compare estimates for the participant versus the actor at each level of Condition revealed no significant effects. The analysis for child participants yielded a two-way interaction between Estimated Person and Condition, $F(1.69,23.76)=32.78, p<.05, \eta_{p}^{2}=$ .70. Simple effects analyses comparing the participant and
actor, across each level of Condition, yielded a significant effect for the weights before walking, $F(1,28)=5.63, p<.05$, $\eta_{p}^{2}=.17$, and after walking condition, $F(1,28)=9.70, p<.05$, $\eta_{p}^{2}=.26$. The mean estimates provided by both age groups for the participant and actor are displayed in Figure 2.

Table 1. Mean jumping estimates and estimation error in cm. All values are rounded to whole integers. Standard deviations are in parentheses.

| Group | Condition | Person | Estimate | Error |
| :--- | :--- | :--- | :--- | :--- |
| Adult | No-weights | Self | $222(7)$ | $8 \quad(13)$ |
|  |  | Actor | $219(9)$ | $16(10)$ |
|  | Before walking | Self | $218(8)$ | $7(14)$ |
|  |  | Actor | $218(9)$ | $17(10)$ |
|  |  | Self | $210(11)$ | $-7(2)$ |
|  |  | Actor | $216(8)$ | $19(10)$ |
| Child | No-weights | Self | $179(13)$ | $-14(17)$ |
|  |  | Actor | $185(11)$ | $45(11)$ |
|  | Before walking | Self | $170(12)$ | $-11(16)$ |
|  |  | Actor | $187(10)$ | $47(10)$ |
|  |  | Self | $157(12)$ | $-12(7)$ |
|  |  | Actor | $188(11)$ | $47(12)$ |

## Estimation Accuracy

The accuracy of estimates were analyzed by examining the mean estimation error (actual $_{\text {jump-height }}-$ estimated $\left._{j u m p-h e i g h t)}\right)$ using a 2 (Age Group: child, adult) x 2 (Estimated Person: participant, actor) x 3 (Condition: no-weights, weights before walking, and weights after walking) mixed-design ANOVA. Analyses revealed a significant three-way interaction of Age Group x Estimated Person x Condition, $F(1.42,39.77)=$ 11.86, $p<.05, \eta_{p}^{2}=.29$.

Follow-up analyses were performed by splitting the age groups into two separate 2 (Estimated Person: participant, actor) x 3 (Condition: no-weights, weights before walking, and weights after walking) repeated-measures ANOVAs. The two-way interaction was significant in the adult group, $F(1.28,18.03)=23.25, p<.05, \eta_{p}^{2}=.62$. Simple effects were used to compare estimation error for the participant versus the actor at each level of Condition. Results yielded a significant difference in the weights before walking condition, $F(1,28)=4.60, p<.05, \eta_{p}^{2}=.31$, and in the weights after walking condition, $F(1.28,18.03)=108.79, p<.05, \eta_{p}^{2}=$ .62. Analyses for the child age group were analyzed similarly. In this case, only the main effect of Estimated Person was significant, $F(1,14)=150.77, p<.05, \eta_{p}^{2}=.92$.

## Accuracy and Eye-Height Scaling

Lastly, we examined whether a relationship between the perceiver's and actor's eye-height explains the accuracy of estimates made for the actor. The focus on actor estimates was chosen because participants provided a consistent level
of accuracy for themselves, but varied in their accuracy for the actor.


Figure 2: Mean jumping estimates by condition, for the participant and the actor. Estimates for the adult (panel A) and child participants (panel B) are shown separately.

Analyses were accomplished using a linear regression to predict the mean estimation error for the actor from the difference in eye-height between participant and actor (eyeheight $_{\text {participant }}$ - eye-height ${ }_{\text {actor }}$ ). Separate regressions were used for each condition (no-weights, weights before walking, and weights after walking). For simplicity, analyses were not split between age group.

The results showed that eye-height difference accounted for a substantial amount of the variance in the no-weights ( $\mathrm{R}^{2}$ $=.70, F(1,28)=63.83, p<.05)$, weights before walking, ( $\mathrm{R}^{2}$ $=.73, F(1,28)=76.75, p<.05)$, and in the weights after walking condition ( $\left.\mathrm{R}^{2}=.65, F(1,28)=52.72, p<.05\right)$. The results for the three separate analyses are displayed in Table 2, and the data is in Figure 3.

Table 2: Results of regression analyses of estimation error for the actor predicted by the eye-height difference between actor and participant.

| Variable | $\beta$ | $\mathrm{SE}(\beta)$ | t | Sig. $(p)$ |
| :--- | :---: | :---: | :---: | :---: |
| No-weights | .834 | .104 | 7.98 | $P<.01$ |
| Before walking | .856 | .098 | 8.76 | $P<.01$ |
| After walking | .808 | .111 | 7.26 | $P<.01$ |

## Discussion

The present study examined what information observer's use when estimating another person's ability to jump for an object.


Figure 3: Scatterplot showing the relationship of the participant's estimation accuracy for the actor with the difference in eye-height.

Like previous findings (Ramenzoni et al., 2008a and 2008b), we anticipated individuals could detect the actor's ability with some accuracy. Some studies (Ramenzoni et al., 2008a) have shown that changing the participant's ability to jump by adding weights alters a perceiver's estimate of their own and an actor's ability, despite not changing the actor's abilities. This has been taken as support that an observer's perceptual judgment is driven by a simulation mechanism. Other studies have shown that estimates of another person's abilities are better described by some physical relationship of the actor (e.g., leg-length) to the environment (Rochat, 1995; Stoffregen et al., 1999) or the physical relationship between two people (Richardson et al., 2007).

Based on the current line of theorizing, we predicted that participant estimates for themselves should decrease when weights were initially added and more so after walking with them on. Additionally, if participants were utilizing eyeheight information, then the estimates for the actor should not decrease significantly. Examining the mean jumpability estimates (Figure 2. and Table 1.), it is clear that participant's estimates for themselves decreased significantly in both age groups, but didn't decrease similarly for the actor. Only with child participants, however, were there significant differences between estimates for themselves and the actor across conditions. The lack of an effect in the adult group is similar findings of Ramenzoni et al. (2008a). The non-effect in the adult group, though, might be due to the eye-height similarity between the adult participants and actor. The children, on average, had a greater eye-height difference to the actor than the adult group.

If perceivers used eye-height information to tune affordance judgments of jumping for the actor, then estimate accuracy should scale with eye-height (Ramenzoni et al., 2008b). Specifically, perceivers with the closest similarity in eye-height to the actor should exhibit the greatest accuracy, assuming they have similar jumping abilities. The regression
analyses of eye-height difference across conditions support this proposal. The $R^{2}$ values demonstrate that a high, and similar, amount of variance is captured by the model across all conditions and age groups. Furthermore, the standardized coefficients are significant and similar across all conditions (Table 2.). Examination of Figure 3 reveals increased accuracy for participants closest in eye-height to the actor. The mean estimation errors (Table 1.) also show that adults were more accurate than children. Interestingly, the scaling relationship shows that as the participant's eye-height decreased away from the actor, there was an increasing tendency to overestimate the actor; as participant eye-height increased away, there was a tendency to underestimate. Together, these findings suggest a potential two part process to perceiving action capabilities for others. Observers can estimate boundaries for another person's abilities by simulating a potential action. Detection of optical information - such as eye-height difference can fine tune these estimates.

## Acknowledgments

The authors thank Anna Silverman for help with data collection.

## References

Brass, M., Bekkering, H., Wohlschläger, A., \& Prinz, W. (2000). Compatibility between observed and executed finger movements: comparing symbolic, spatial, and imitative cues. Brain and cognition, 44(2), 124-143.
Calvo-Merino, B., Glaser, D. E., Grèzes, J., Passingham, R. E., \& Haggard, P. (2005). Action observation and acquired motor skills: an FMRI study with expert dancers. Cerebral cortex, 15(8), 1243-1249.
Davis, T.J., Riley, M.A., Shockley, K., \& Cummins-Sebree, S. (2010). Perceiving affordance for joint actions. Perception, 39(12), 1624-1644.
Gallese, V., Keysers, C., \& Rizzolatti, G. (2004). A unifying view of the basis of social cognition. Trends in Cognitive Sciences, 8 (9), 396-403.
Gibson, J. J. (1979). The ecological approach to visual perception. Lawrence Erlbaum.
Grush, R. (2004). The emulation theory of representation: motor control, imagery, and perception. Behavioral and Brain Sciences, 27, 377-442.
Knoblich, G., \& Jordan, S. (2002). The mirror system and joint action. In M. I. Stamenov \& V. Gallese (Eds.), Mirror neurons and the evolution of brain and language. Amsterdam: John Benjamins.
Loula, F., Prasad, S., Harber, K., \& Shiffrar, M. (2005). Recognizing people from their movement. Journal of Experimental Psychology: Human Perception and Performance, 31 (1), 210.
Mark, L. S. (1987). Eye height-scaled information about affordances: A study of sitting and stair climbing. Journal of Experimental Psychology: Human Perception and Performance, 13(3), 361.

Prinz, W. (1997). Perception and action planning. European journal of cognitive psychology, 9(2), 129-154.
Ramenzoni, V. C., Riley, M. A., Shockley, K., \& Davis, T. (2008a). Carrying the height of the world on your ankles: Encumbering observers reduces estimates of how high an actor can jump. The Quarterly Journal of Experimental Psychology, 61 (10), 1487-1495.
Ramenzoni, V. C., Riley, M. A., Shockley, K., \& Davis, T. (2008b). An information-based approach to action understanding. Cognition, 106 (2), 1059-1070.
Ramenzoni V.C., Riley M.A., Davis T., Shockley, K., \& Armstrong, R. (2008c). Tuning in to another person's action capabilities: perceiving maximal jumping-reach height from walking kinematics. Journal of Experimental Psychology: Human Perception and Performance, 34 (4), 919-928.
Richardson, M. J., Marsh, K. L., \& Baron, R. M. (2007). Judging and actualizing intrapersonal and interpersonal affordances. Journal of Experimental Psychology: Human Perception and Performance, 33 (4), 845-859.
Rizzolatti, G., \& Craighero, L. (2004). The mirror-neuron system. Annu. Rev. Neurosci., 27, 169-192.
Rochat, P. (1995). Perceived reachability for self and for others by 3-to 5-year-old children and adults. Journal of Experimental Child Psychology, 59 (2), 317-333.
Sebanz, N. \& Knoblich, G. (2009). Prediction in joint action: What, when, and where. Topics in Cognitive Science, 1 (2), 353-367.
Shockley, K., Santana, M. V., \& Fowler, C. A. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. Journal of Experimental Psychology: Human Perception and Performance, 29 (2), 326-332
Stoffregen, T. A., Gorday, K. M., Sheng, Y.-Y., \& Flynn, S. B. (1999). Perceiving affordances for another person's actions. Journal of Experimental Psychology: Human Perception and Performances, 25 (1), 120-136.
Warren, W. H. (1984). Perceiving affordances: Visual guidance of stair climbing. Journal of Experimental Psychology: Human Perception and Performance, 10 (5), 683-703.
Wilson, M., \& Knoblich, G. (2005). The case for motor involvement in perceiving conspecifics. Psychological bulletin, 131(3), 460.

# The Effect of Non-Linguistic Patterns on Linguistic Biases 

Sara Finley (sara.finley@waldorf.edu)<br>Department of Psychology, Waldorf College<br>Forest City, IA, 50436, USA


#### Abstract

The present study explores the effects of non-linguistic experiences on biases for linguistic judgments, specifically consonant deletion patterns. When two adjacent consonants come into contact as a result of morphological concatenation, many languages will delete the first consonant (e.g., /bepdok/ becomes /bedok/). Speakers of these languages (as well as English speakers) prefer deletion of the first consonant to the second consonant because the first consonant is perceptually weaker, making it more prone to misrepresentations and modifications. Following exposure to a non-linguistic analogue of consonant deletion in which the second consonant was deleted instead of the first, participants no longer preferred deletion of the first consonant in the metalinguistic judgment task. These results suggest that exposure to non-linguistic materials can interact with linguistic judgments.


Keywords: statistical learning, phonotactics, learning biases, analogy.

## Introduction

One of the major questions in the cognitive science of language is how linguistic and non-linguistic experiences interact to build a productive system of language. For example, children learning language benefit from increased cognitive and social skills in their language capacities, but increased language capacities also help to scaffold cognitive and social growth (Dessalegen \& Landau, 2008). Because language interacts with social, cultural and cognitive aspects of human functioning, it is important to understand how language influences non-linguistic cognition, in addition to how non-linguistic cognition influences language.

The discussions concerning linguistic and non-linguistic interaction have often turned to the question of linguistic relativity, the idea that the specific language one speaks has an effect on how the speaker perceives and interacts with the world (Whorf, 1956). In addition, there is a question of how linguistic knowledge can aid in higher level cognition (Gentner \& Goldin-Meadow, 2003).

The question of how language and thought interact can be addressed not only as whether language affects thought, but also whether non-linguistic information can have an effect on language. While language is a direct way to express one's thoughts, there may be other, more subtle ways in which non-linguistic experience can affect language. These subtle effects could be used to understand the ways in which linguistic knowledge is specific to language (domain
specific) and general to other cognitive processes (domain general). In domain specific views, language is believed to be a key component that to human cognition. The mechanisms that underlie language are separate from other species, and (in the most extreme theories) show no interaction with non-linguistic cognitive functions (Berent, 2012; Pinker \& Jackendoff, 2009). In this view, nonlinguistic cognition should have no influence on linguistic judgments. In a domain general view of language, foundations for the human language capacity arise through social and cultural transmission. The key to linguistic knowledge is an interaction between the need to communicate and the existence of high-level cognitive capacities such as abstract pattern learning and memory (Chater \& Christiansen, 2010). Under this view, nonlinguistic patterns should have a strong influence on linguistic constructs.

One of the strong pieces of evidence for a domain specific approach to language is the idea that there are biases for specific linguistic structures that have no non-linguistic analogues (Berent, Steriade, Lennertz, \& Vaknin, 2007; Culbertson, Smolensky, \& Legendre, 2012; Finley, 2012; Finley \& Badecker, 2008). For example, Finley (2012) found that adult native English speaker show a bias for phonological patterns based on vowel height. Since vowel height is a linguistic construct, it is hard to imagine how such a bias could be influenced by non-linguistic factors.

Other evidence suggests that linguistic patterns may be stored as domain-general rules. Studies of statistical learning for speech segmentation showed similar results to linguistic and non-linguistic materials (Aslin, Saffran, \& Newport, 1997; Saffran, Pollak, Seibel, \& Shkolnik, 2007). In addition, Finley and Christiansen (2011) showed that adult learners can generalize a novel reduplication pattern to from non-linguistic materials to linguistic judgments.

In addition, robust use of analogy in both linguistic and non-linguistic learning tasks (Gentner, 2010) opens the possibility that learners will be able to form connections between non-linguistic patterns and linguistic patterns.

The evidence for both domain general and domain specific learning mechanisms suggests that grammatical principles have many influences. The goal of the present study is to provide experimental evidence that manipulation of nonlinguistic patterns can affect linguistic biases. Specifically, we focus on consonant deletion, a phonological pattern whereby a consonant will delete in the presence of two adjacent consonants.

## Biases in Consonant Deletions

Consonant deletion is a phonological pattern in which one of two adjacent consonants delete (e.g., /depkot/ becomes
/dekot/). In these consonant deletion patterns, there appears to be a cross-linguistic preference to delete the first consonant (Steriade, 2001; Wilson, 2001). In Diola Fogny, when two consonants combine as a result of morphological concatenation, the first consonant deletes (e.g., /let+ku+jaw/ $\rightarrow$ [lekujaw] 'they won't go') (Sapir, 1975; Wilson, 2000, 2001). Wilson (2001) argues that the second consonant is in a perceptually stronger position (onset), while the first consonant is in a perceptually weaker position (coda). If a rule requires deletion of a consonant, speakers will choose to delete the weaker one. A perceptually weak consonant is more likely to be misheard or not heard at all, meaning that over time (diachronically), that consonant may be categorically deleted from the lexical item (Steriade, 2001, 2009).

Finley (2011b) provided evidence that the preference for C 1 deletion over C2 deletion is synchronic and present in speakers of English (a language that does not have regular consonant cluster deletion). In this experiment, monolingual English speaking participants were given a two-alternative forced choice test in which participants chose between two triads. In one triad, the first consonant was deleted (e.g., /bep, dok, bedok/). In another triad, the second consonant was deleted (e.g., /bep, dok, bepok/). Participants were more likely to choose the triads where the first consonant deleted (based on the criteria to choose which triad was more likely to belong to a 'real' language). This result suggests that participants prefer to delete the perceptually weak pattern, despite no exposure to this pattern in the native language.

## Sources of Linguistic Biases

While it is agreed that linguistic biases are prevalent, the sources of such biases are not agreed upon. One possibility is that linguistic biases are derived from pre-existing linguistic knowledge or experience. This knowledge could be innate (Berent, et al., 2007), or inferred indirectly through the course of exposure to other patterns in the language. It is also possible that a bias for a particular linguistic pattern may have roots in domain general cognition (Chater \& Christiansen, 2010). For example, cross-linguistic preferences to avoid changes to the first syllable of a word may result from domain general mechanisms (Beckman, 1998). The first and last items in a list are the most likely to be remembered (referred to as primacy and recency), suggesting that the first and last parts of a word will also be easiest to remember. If beginnings of the words are easier to remember, speakers may avoid altering that part of the word.

One issue with discerning whether non-linguistic cues can influence linguistic biases is that the linguistic and nonlinguistic cues often interact, and the direction of interaction is often unclear. For example, initial syllables may be more likely to be remembered because they are less likely to be altered (and thus have fewer alternative forms to consider). In addition, other linguistic cues such as stress, prominence and volume may play also play in phonological processes,
and these different factors may vary across different languages.

The goal of the present study is to determine whether a non-linguistic analogue of a linguistic pattern can alter a linguistic bias. If the non-linguistic cue can remove (or even reverse) the linguistic bias, it suggests that non-linguistic cues do affect how speakers perceive and interpret language. It is important to note that if a non-linguistic cue can affect a linguistic bias, it in no way implies that all linguistic biases have a non-linguistic basis. However, if a linguistic bias can be influenced by a non-linguistic cue, it opens the possibility that domain general influences affect at least some of the linguistic tendencies found cross-linguistically.

The present experiment makes use of the known linguistic bias in adult English speakers for deletion of C1 in a C1-C2 consonant cluster (Finley, 2011). The present experiment asks whether exposure to a non-linguistic analogue of C2 consonant deletion (as opposed to the preferred C 1 deletion) can reduce or reverse the bias for C 1 deletion in learners.

## Methods

The present study used an artificial language that contained a non-linguistic analogue of a deletion pattern. In Finley (2011a, 2011b), consonant deletion was induced via triads in which two CVC (consonant-vowel-consonant) items were combined to form a CVCVC word (e.g., /bek dof bedof/). We created a non-linguistic analogue using sequences of shapes with various patterns. An analogue for perceptual dis-preference for two consonants in a row was created using visual aesthetics. In the present experiment we treated every consonant as a long rectangle with various patterns, and every vowel as a circle filled with various patterns. The fill patterns were used to create differences between the various circles and rectangles, while maintaining a strong sense of continuity between the shape and size of the circles.

## Participants

All participants were adult native English speakers. Eighteen participants were recruited from Elmhurst College and the surrounding community. Each participant was given a $\$ 10$ gift card for participating. Twelve participants were recruited from the University of Rochester community and paid $\$ 10$ cash for their participation. Twelve control participants were recruited from the University of Rochester community were paid $\$ 5$ cash for their participation. Some participants may have previously participated in an artificial grammar learning experiment, but no participant had been exposed to the stimuli or patterns used in the present experiment. The data for two additional participants could not be used due to malfunctions in the experimental program.

## Design

The experiment was designed to test the ability of adult learners to extend a novel non-linguistic analogue of consonant deletion to a linguistic version of the same task. English speakers have been shown to prefer deletion of the
first consonant of a CC consonant cluster, complying with the general cross-linguistic tendencies (Finley, 2011b; Wilson, 2001). Importantly, the preference shown in English speakers appears without any prior exposure to the pattern. In these previous studies, participants were exposed to triads of shapes presented in the center of the screen for 1000 ms . Participants were told that they would see one of two shapes followed by the combination of the first two shapes (participants were given a practice trial in which all squares and circles were identical, and given a chance to ask questions if necessary).

The patterns of the circles and squares were made in exact analogy to a precious consonant deletion experiment in which two CVC words were combined to form a CVCV word (e.g., /bek/ + /dok/ $\rightarrow$ /bedok/) (Finley, 2011a, 2011b). The visual analogue treated every segment as a separate shape. For example, /a/ was a circle with black 'confetti' squares, and $/ \mathrm{k} /$ was a rectangle filled in with a diagonal brick pattern. Creating stimuli in this manner helped to keep the stimuli as analogous to an experiment that used linguistic materials. It also allowed non-linguistic materials to be balanced similarly to linguistic materials. Examples of the training stimuli can be found in Table 1.

Table 1: Examples of Training Stimuli.


Training consisted of 24 triads repeated five times each in a random order. Immediately following exposure, participants were given a two-alternative forced choice test in two parts. The first part tested knowledge of the non-linguistic pattern, with examples found in Table 2. The second part tested biases towards C1 deletion in a linguistic consonant deletion pattern, with examples in Table 3.

Old Items The first type of test item specifically tested the learner's ability to recognize which of the rectangles was deleted in the exposure items. A participant could respond correctly to these items by remembering the specific items in the exposure set.

New Items The second type of test item used novel shape items. A participant could respond correctly to these items if they extend the pattern seen during exposure to novel items.

Sound Items The third type of test item was designed to assess whether participants who were exposed to the visual deletion pattern would show the same bias towards first-
consonant deletion shown in previous studies. The stimuli were nearly identical to those used in previous studies of consonant deletion (Finley, 2011a, 2011b). Participants were told to select which of the following sets of three sounds was most likely to be from a real language. The sound items were presented in the same manner as the Old and New test items, choosing between deletion of the first consonant (C1) or the second consonant ( C 2 ) of a consonant cluster. Each item in the two-alternative forced-choice task was a tirad: $\mathrm{CVC}_{1}, \mathrm{C}_{2} \mathrm{VC}, \mathrm{CVCVC}$. Participants were told that they would be hearing two sets of three non-words where the third word was a combination of the first two (given /tooth+brush $=$ toothbrush/ as an example), and their job was to select which set of three non-words they preferred.

Table 2: Example Old and New Test Items

|  | Set 1 | Set 2 | Combined Form |
| :---: | :---: | :---: | :---: |
| $\quad 1 \quad$ Old Second Deleted (Correct) |  |  |  |
| First <br> Deleted <br> (Incorrect) |  |  |  |
| New |  |  |  |
| Second <br> Deleted <br> (Correct) |  |  |  |
| First <br> Deleted <br> (Incorrect) |  |  |  |

All stimuli were designed so that the final consonant of the first CVC word was different from the first consonant of the second CVC word. For example, [pik ket] was not a possible pair of words in the experiment because it would be impossible to tell which consonant was deleted. Consonants were drawn from the set $[\mathrm{p}, \mathrm{t}, \mathrm{k}, \mathrm{b}, \mathrm{d}, \mathrm{g}, \mathrm{s}, \mathrm{f}, \mathrm{z}, \mathrm{v}, \mathrm{m}, \mathrm{n}]$, and vowels were drawn from the set $[\mathrm{a}, \mathrm{i}, \mathrm{e}, \mathrm{o}, \mathrm{u}]$ Examples of Sound Items can be found in Table 3.

The Sound stimuli were recorded by an adult female native speaker of English in a sound attenuated booth at $12,000 \mathrm{~Hz}$. Stress was placed on the first syllable using standard English pronunciation, with the exception that no vowels were reduced, meaning though all syllables contained partial stress (as English reduces unstressed syllables). All stimuli items were normalized for intensity (set at 70dB) using Praat (Boersma \& Weenink, 2005).

There were 12 Old Items, 12 New Items and 30 Sound Items (however, a glitch in a group of participants caused
several participants to hear a random set of 20 of the 30 sound items). The Old and New test items were presented together in a random order, before the Sound Items. The items in each test condition were balanced such that half of the items showed deletion of the first consonant/rectangle first, while the other half of the items showed deletion of the second consonant/rectangle first.

Table 3: Sound Item Examples.

|  | CVC 1 | CVC 2 | Combined <br> Form |
| :--- | :--- | :--- | :---: |
| Second Deleted <br> (Non-Linguistic Bias) | div | nup | divup |
| First Deleted <br> (Linguistic Bias) | div | nup | dinup |
| Second Deleted <br> (Non-Linguistic Bias) | kaf | gez | kafez |
| First Deleted <br> (Linguistic Bias) | kaf | gez | kagez |

All phases of the experiment were run in Psyscope X (Cohen, MacWhinney, Flatt, \& Provost, 1993). Participants were given both written and verbal instructions. The entire experiment took approximately 20 minutes.

## Results

Proportion of $\operatorname{Set} 2 / \mathrm{C} 2$ deletion responses for all three different test items are given in Figure 1, with numerical values for means and standard deviations in Table 4.

| Table 4: Means (and Standard Deviations). |  |  |  |
| :---: | :---: | :---: | :--- |
| Condition | Old | New | Sounds |
| Control | 0.53 | 0.47 | 0.37 |
|  | $(0.10)$ | $(0.11)$ | $(0.11)$ |
| Experimental | 0.78 | 0.78 | 0.63 |
| (Old Items | $(0.11)$ | $(0.18)$ | $(0.26)$ |
| Above Chance) |  |  |  |
|  |  |  |  |
| Experimental | 0.46 | 0.47 | 0.44 |
| (Old Items | $(0.059)$ | $(0.14)$ | $(0.21)$ |
| Below Chance) |  |  |  |

We compared the results for the experimental condition to the Control condition by a $2 \times 3$ mixed design ANOVA. We found a significant effect of Training $(F(1,37)=10.89, p=$ 0.002 ), a significant effect of Test $(F(2,74)=7.13, p=$ 0.001 ), and no interaction, $F<1$.

In order to test whether the bias existed in the Controls, and whether the bias was reversed in the Experimental condition, we compared the responses to $50 \%$ chance via one-sample t-tests. The results were significant for the Control condition $t(11)=4.33, p=0.0012$ (in that the Control condition was significantly below chance), but the results were not significant for the Experimental Condition,
$t(26)=1.15, p=0.26$. Because the experiment was concerned with whether exposure to the non-linguistic deletion pattern would change the bias towards C 1 deletion in the consonant test, we compared the responses to the Sound Test Items between the Control and the Experimental Condition. There was a significant difference, $t(37)=2.45, p=0.019$.

Figure 1: Overall Results: Means and Standard Errors.


One possibility for the failure to find a significant difference between the Sound test items and chance (in the one-sample t-test) was that some participants failed to learn the nonlinguistic pattern or remember the items heard in training. One cannot expect the non-linguistic pattern to have any effect on the linguistic pattern without learning the pattern (or at least recognizing the items heard in training). For this reason, we divided participants in the Experimental Condition into two groups: those that scored above $50 \%$ in the Old Items, and those that scored $50 \%$ (chance) or below in the Old Items. Of the 27 participants in the Control Condition, 17 scored above chance in the Old Items, and 10 scored at or below chance. These are presented in Figure 2.

Figure 2: Results with Participants in Experimental Condition: Separated by Response Rate: Means and Standard Errors


The participants who scored at or below chance for Old Items showed results very similar the Control Condition. When compared to the Control Condition via ANOVA, we found no effect of Training ( $F<1$ ), a marginal effect of Test $(F(1,40)=3.00, p=0.061)$ and no interaction $(F(1,40)=$ $1.64, p=0.21)$. When the Sounds Test items were
compared to $50 \%$ chance via one-sample t-tests, there was a marginally significant effect, $t(16)=2.04, p=0.0585$. Of the 17 participants who scored above chance for Old Items, three participants scored below $40 \%$ C2 deletion in the Sound Items. For this minority of participants, exposure to the non-linguistic pattern did not affect the bias. However, the majority of the 17 participants showed C 2 deletion at a rate greater than that of the mean of the Control condition.

## Discussion

The results of the present study demonstrate that linguistic biases can be reduced or altered due to exposure to nonlinguistic material. These results have important implications for cognitive science. First, it suggests that biases found for linguistic patterns are malleable. Different experiences can prime the listener to expect different types of linguistic stimuli, and therefore diminish a pre-existing bias. This means that an innate bias for a particular linguistic structure could be overridden if provided with exposure to the right kinds of data. This may help to create a theory of linguistic biases that can account for the fact that there are exceptions to almost every posited linguistic universal (Evans \& Levinson, 2009).

Second, the results support a theory in which linguistic and non-linguistic data interact. In understanding the domain specificity of language, one must understand what aspects of language interact with non-linguistic cognition, as well as the mechanisms that control this interaction. The results of the present study provide an insight into this question. In the present study, the non-linguistic deletion pattern had a direct analogue to the consonant deletion pattern. This direct analogy allowed participants in the Experimental Condition to interpret the linguistic material differently than participants in the Control condition.

A proposed analysis of the influence of linguistic experiences, non-linguistic experience, and linguistic biases on linguistic biases is presented in Figure 3.

Figure 3: The interactive of linguistic experiences, nonlinguistic experiences and linguistic biases on linguistic judgments.


Linguistic judgments are affected by our linguistic experiences; native English speakers are able to make judgments about English due to their exposure to English. Linguistic judgments are also affected by biases that are independent of language exposure, such as the bias for C 1 deletion over C 2 deletion found in the control condition.

The Experimental condition demonstrated that nonlinguistic experiences can affect linguistic judgments. The non-linguistic experience pushed the participants away from a bias towards C1 deletion. The mechanism proposed in Figure 3 also allows for interaction between linguistic and non-linguistic experiences, as well as an integration between linguistic biases and linguistic experiences. Non-linguistic experiences affect the type of language you are exposed to, and linguistic biases affect the likelihood that you will learn and be exposed to certain types of linguistic materials (Finley, 2012).

The diagram in Figure 3 also allows for individual differences in when non-linguistic experiences will affect linguistic judgments. When non-linguistic experiences and linguistic biases are in conflict (as in the present experiment), biases may trump non-linguistic experiences for some individuals. A small majority of participants in the Experimental condition showed a bias for C1 deletion, despite learning the non-linguistic pattern. This suggests that analogy from non-linguistic to linguistic patterns do not occur for everyone.

Third, the present experiment demonstrates that language and thought interact, and that the direction of interaction can go from non-linguistic patterns to linguistic patterns. The question of language and thought need not extend only to whether language affects thought, but whether nonlinguistic patterns can affect how language is perceived language. The present experiment demonstrates that our non-linguistic experiences can affect how we perceive language.

One question that remains for future research is to understand when non-linguistic patterns may affect linguistic judgments in real-world situations. The present experiment made an arbitrary analogy between consonant deletion and shape deletion. Such direct analogies are rarely found in the real world. Given that patterns in language tend to be abstract and arbitrary, it is difficult to find a nonlinguistic pattern that can be directly linked to language. One possibility may lie within the cognitive and linguistic development of infants and young children. As children learn patterns in their behavior and the behavior of others, they may use those patterns to help learn linguistic patterns. Conversely, children may use their ability to learn patterns to help learn both non-linguistic cognitive skills, as well as linguistic skills. For example, Dessalegen and Landau (2008) demonstrated that children can use labels to solve otherwise difficult non-linguistic tasks. In addition, the robust use of analogy in learning (Gentner, 2010), suggests that learners are capable of analogy from linguistic to nonlinguistic material and vice versa. Future research will work to formalize when and how this analogy occurs.

The results of the present experiment provide further evidence for interaction between linguistic and nonlinguistic patterns. Human learners have a remarkable ability to use analogy to extend a pattern from a nonlinguistic domain to a linguistic domain. Despite the fact that English speakers (as well as speakers of several other
languages) show a bias towards C 1 deletion, this bias was reduced after exposure to a pattern in which the nonlinguistic analogue of C 2 was deleted (as opposed to an analogue of C1).

## Acknowledgments

The author would like to thank Patricia Reeder, Elissa Newport, Emily Kasman, Morgan Smith, Elizabeth Wiemers and Elizabeth Jennings, members of the AslinNewport Lab. In addition, we would like to thank the Elmhurst College Psychology Department, the University of Rochester Brain and Cognitive Science Department, particularly the participants at Elmhurst College. The author assumes all responsibility for any errors. Funding was provided by a grant from Elmhurst College, NIH grant DC00167 to E. Newport, HD37082 to R. Aslin \& E. Newport, and NIH training grant T32DC000035.

## References

Aslin, R. N., Saffran, J. R., \& Newport, E. (1997). Computation of conditional probability statistics by 8-month-old infants. Psychological Science, 9, 321-324.
Beckman, J. N. (1998). Positional faithfulness. Unpublished doctoral dissertation, UMass Amherst.
Berent, I. (2012). The phonological mind. Cambridge: Cambridge University Press.
Berent, I., Steriade, D., Lennertz, T., \& Vaknin, V. (2007). What we know about what we have never heard: Evidence from perceptual illusions. Cognition, 104(3), 591-630.
Boersma, P., \& Weenink. (2005). Praat: Doing phonetics by computer.
Chater, N., \& Christiansen, M. H. (2010). Language acquisition meets language evolution. Cognitive Science, 34, 1131-1157.
Cohen, J. D., MacWhinney, B., Flatt, M., \& Provost, J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. Behavioral Research Methods, Instruments and Computers, 25, 257-271.
Culbertson, J., Smolensky, P., \& Legendre, G. (2012). Learning biases predict a word order universal. Cognition, 122, 306-329.
Dessalegen, B., \& Landau, B. (2008). More than meets the eye: The role of language in binding and maintaining feature conjunctions. Psychological Science, 19(2), 189-195.
Evans, N., \& Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. Behavioral and Brain Sciences, 32(5), 429-492. doi: 10.1017/S0140525X0999094X

Finley, S. (2011a). Generalization to novel consonants in artificial grammar learning. In L. Carlson, C. Hölscher \& T. Shipley (Eds.), Proceedings of the 33rd Annual Cognitive Science Society Annual

Meeting (pp. 3268-3273). Austin, TX: Cognitive Science Society.
Finley, S. (2011b). Perecptual biases in consonant deletion. Proceedings of the XVII Proceedings of the International Congress of Phonetic Sciences (ICPhS XVII).
Finley, S. (2012). Typological asymmetries in round vowel harmony: Support from artificial grammar learning. Language and Cognitive Processes, 27(10), 1550-1562.
Finley, S., \& Badecker, W. (2008). Analytic biases for vowel harmony languages. $W C C F L, 27,168-176$.
Finley, S., \& Christiansen, M. H. (2011). Multimodal transfer of repetition patterns in artificial grammar learning. In L. Carlson, C. Hölscher \& T. Shipley (Eds.), Proceedings of the 33rd Annual Meeting of the Cognitive Science Society (pp. 330-335). Austin, TX: Cognitive Science Society.
Gentner, D. (2010). Bootstapping the mind: Analogical processes and symbol systems. Cognitive Science, 34, 752-775.
Gentner, D., \& Goldin-Meadow. (2003). Whither Whorf. In D. Gentner \& Goldin-Meadow (Eds.), Language in mind: Advances in the study of language and thought (pp. 3-14). Cambridge, MA: MIT Press.
Pinker, S., \& Jackendoff, R. (2009). The components of language: What's specific to language, and what's specific to humans? In M. H. Christiansen, C. Collins \& S. Edelman (Eds.), Language Universals (pp. 126-151). New York: Oxford University Press.
Saffran, J. R., Pollak, S. D., Seibel, R. L., \& Shkolnik, A. (2007). Dog is a dog is a dog: Infant rule learning is not specific to language. Cognition, 105(3), 669680. doi: 10.1016/j.cognition.2006.11.004

Sapir, J. D. (1975). Big and thin: two Diola-Fogny metalinguistic terms. Language in Society, 4, 1-15.
Steriade, D. (2001). Directional asymmetries in place assimilation: A perceptual account. In E. Hume \& K. Johnson (Eds.), The Role of Speech Perception in Phonology (pp. 219-250). London: Academic Press.
Steriade, D. (2009). The Phonology of perceptibility effects: the P -map and its consequences for constraint organization. In K. Hanson \& S. Inkelas (Eds.), The Nature of the Word: Studies in Honor of Paul Kiparsky (pp. 151-180). Cambridge, MA: MIT Press.
Whorf, B. L. (1956). Language, thought and reality. Cambridge, MA: MIT Press.
Wilson, C. (2000). Targeted constraints: An approach to contextual neutralization in Optimality Theory. Johns Hopkins University, Baltimore, MD.
Wilson, C. (2001). Consonant cluster neutralisation and targeted constraints. Phonology, 18, 147-197.

# Using Causal Induction in Humans to Learn and Infer Causality from Video 

Amy Fire (amy.fire@ucla.edu)<br>Song-Chun Zhu (sczhu@stat.ucla.edu)<br>Center for Vision, Cognition, Learning, and Art<br>University of California, Los Angeles<br>Los Angeles, CA 90095 USA


#### Abstract

For both human and machine learners, it is a challenge to make high-level sense of observations by identifying causes, effects, and their connections. Once these connections are learned, the knowledge can be used to infer causes and effects where visual data might be partially hidden or ambiguous. In this paper, we present a Bayesian grammar model for human-perceived causal relationships that is learnable from video. Two experiments investigate high-level causal induction from low-level visual cues. In the first experiment, we show that a computer can apply known heuristics used for causal induction by humans to learn perceptual causal relationships. In the second experiment, we show that our learned model can represent humans' performance in reasoning about hidden effects in video, even when the computer initially misdetects those effects.


Keywords: Perceptual causality; causal induction; statistical models.

## Introduction

A man approaches a closed door. He reaches out to grasp the handle and then stands there. Is it locked? Does he not have the key? He knocks and waits, but the door remains closed. Is there no one on the other side to open it?

Watching these events unfold, humans can readily answer these questions based on their causal knowledge. One way humans can learn causal relationships is through daily observation by internally measuring co-occurrence of events (Griffiths \& Tenenbaum, 2005). Research suggests that humans use a few heuristics to determine whether a cooccurrence is causal, including:

- whether the temporal lag between cause and effect is short, and the cause precedes the effect (Carey, 2009) and
- whether agent actions are responsible for causes (Saxe, Tenenbaum, \& Carey, 2005).

However, learning from daily observation is limited: many actions and effects are hidden. Our prior knowledge about causal relationships between actions and effects allows us to fill in information about the events in the scene.

Some current models represent knowledge with Bayesian networks, e.g., (Griffiths \& Tenenbaum, 2005). These models, however, are disjoint from the low-level visual data that people observe. Instead, models are built using high-level annotations. In reality, agents build knowledge by observing low-level visual data, and models need to be able to deal with uncertainty in observation.

Although Bayesian networks are commonly used to represent causality (Pearl, 2009), grammar models have the expressive power to represent a greater breadth of possibilities than a single instance of a Bayesian network (Griffiths
\& Tenenbaum, 2007). Grammar models allow for multiple configurations and high-level structures, making them more suitable for applications grounded on visual cues; Bayesian networks lack the representative power needed for this.

Grammar models are represented graphically in the AndOr Graph (AOG). In the AOG, Or-nodes represent the multiple alternatives, and And-nodes represent hierarchical decompositions. The AOG naturally lends itself to represent causation where multiple alternative causes can lead to an effect, and each cause is composed of conditions necessary for the effect.

In this paper, we introduce a grammar model for representing causal relationships between actions and object-status changes, the Causal And-Or Graph (C-AOG). We describe methods for learning the model by using co-occurrence to identify potential causal relationships between events and applying the heuristics listed above to those potential relationships. In two experiments, we investigate how the model matches human perceptions of causality. Experiment 1 uses input typical of computer vision detection systems to investigate learning the C-AOG and human perceptions of causality. Experiment 2 demonstrates that the C-AOG models human judgments on imputing hidden variables from video.

## A Grammar Model for Causality

In this section, we introduce the Causal And-Or Graph for causal reasoning, which ties agent actions to fluents.

## Fluents and Actions

Specifically defining those object statuses that vary over time, the term fluents comes from the commonsense-reasoning literature (Mueller, 2006). Relevant here are two kinds of fluents that intentional agents can change: object fluents (e.g., a light can be on or off) and fluents of the mind (e.g., an agent can be thirsty or not thirsty).

The values of these fluents change as a result of agent actions and also trigger rational agents to take action. A lack of change-inducing action (also known as the inertial action) causes the fluent to maintain its value; for example, a door that is closed will remain closed until some action changes it. In this work, fluents are modeled discriminatively.

Actions $\left(A_{i}\right)$ are modeled using the Temporal And-Or Graph (T-AOG), a grammar model for actions (Pei, Jia, \& Zhu, 2011). In the T-AOG, And-nodes group the necessary ways for an action to be performed that allow detection of the action (e.g., object/agent spatial relations, agent poses, scene contexts, and temporal relationships), and Or-nodes provide


Figure 1: A C-AOG for door status as learned in Experiment 1. The value of the top-level fluent is a consequence its children. The fluent transit action nodes indicate the kind of change that occurs in the fluent: step functions for change, flat lines for non-change (or inertial action). Action $a_{0}$ is the inertial action (a lack of state-changing action). Arcs connect children of And-nodes. It should be noted that each photo represents a further set of child And-nodes from the Temporal And-Or Graph (not shown). Thickened lines indicate selections on the Or-nodes that provide a single parse graph.
the alternative methods of performing the action. While hidden Markov models and dynamic Bayesian networks have also been used for action detection from video, the grammar is necessary as it allows representation of high-level structures and multiple configurations.

Our experiments are conducted using a pre-selected set of actions and fluents common to office, hallway, and elevator scenes. Such scenes (and events therein) might be of interest for surveillance, for example.

## The Causal And-Or Graph

The Causal And-Or Graph (C-AOG) is a graphical representation for the grammar of causality. The top levels of one C-AOG learned in Experiment 1 are shown in Figure 1.

In the C-AOG, Or-nodes represent the alternative means of causation (e.g., a monitor, through the computer, can be turned on by someone using a mouse or a keyboard). Arrows point from these causing actions to their fluent effects.

Each And-node is formed from the set of multiple conditions for the action, including its sub-actions. The action nodes in a C-AOG may be inertial actions (resulting in no change); unexplained instances of the fluent are also pooled under the inertial action.

A selection on the Or-nodes is called a parse graph, denoted $p g$ (such as the paths shown by thicker lines in Figure 1). It provides a causal interpretation of each fluent's particular value at a given time, answering "why" the fluent has that particular value.

## Probability on the C-AOG

The probability model over the parse graphs in the C-AOG incorporates the detection probabilities of actions and fluents in a Bayesian manner. In particular, given the video $I$,

$$
\begin{equation*}
\underbrace{P\left(p g_{\mathrm{C}} \mid I\right)}_{\text {posterior }}=\underbrace{P\left(A_{1}, \ldots A_{n} \mid I\right) P\left(\Delta F_{1}, \ldots, \Delta F_{m} \mid I\right)}_{\text {likelihood }} \prod_{v \in V_{\mathrm{C}}^{\mathrm{Or}}} \underbrace{P(w(v))}_{\text {prior }} \tag{1}
\end{equation*}
$$

The likelihood term is the detection probability for the included actions/fluents, and considers actions and fluents independently. $V_{\mathrm{C}}^{\mathrm{Or}}$ is the set of included Or-nodes in the causal explanation, and $w(v)$ returns the selected Or-branch. The prior term gives the switch probability on the Or-nodes for the alternative causes and is learned by maximum likelihood estimation.

## Learning the C-AOG

To learn the C-AOG, potential causal relationships are found by restricting the set of all possible fluent/action interactions with the set of heuristics listed at the beginning. Actions and fluents from all levels of their respective hierarchies are considered.

A joint model is iteratively built up from the initial probability distribution over actions and fluent changes, incorporating a new causal relationship each iteration. In an iteration, the contingency table of each action-fluent pair $\left(A_{i}, \Delta F_{j}\right)$, e.g., Table 1 , is examined. The best causal relationship is determined by maximizing the information gain (IG), which is the Kullback-Leibler divergence (KL) (Kullback \& Leibler, 1951) between the full contingency table of Table 1 and the expected contingency table predicted by the model in the current iteration (similar to work on texture modeling (Zhu, Wu, \& Mumford, 1997)). In particular, in a single iteration, causal relation $c r^{*}$ is added to the model where

$$
\begin{equation*}
c r^{*}=\underset{c r}{\operatorname{argmax}} \mathrm{IG}=\underset{c r}{\operatorname{argmax}} \operatorname{KL}(\mathbf{f}| | \mathbf{h}), \tag{2}
\end{equation*}
$$

$\mathbf{f}=\left(f_{0}, f_{1}, f_{2}, f_{3}\right)$, and $\mathbf{h}$ is the analogous quantity from the current iteration's model. The causal relationships with highest information gains are deemed most significant and are collected into the C-AOG.

Table 1: Contingency table of relative frequencies.

|  | $\Delta F_{j}$ Present | $\Delta F_{j}$ Absent |
| :--- | :--- | :--- |
| $A_{i}$ Present | $f_{0}$ | $f_{1}$ |
| $A_{i}$ Absent | $f_{2}$ | $f_{3}$ |

Our learning method integrates with existing action and fluent detection systems, creating a unified framework for the spatial, temporal, and causal domains. Further, our method is more computationally feasible for large networks of causal connections than Bayesian learning frameworks are (with their prior distributions over graph structures). Traditional causal induction as done by constraint satisfaction (Pearl,


Figure 2: Information gains for causal relations in the order pursued, separated by fluent. Green circles label causes.
2009) or Bayesian formulations (Heckerman, 1995) is intractable to ground on vision sensors. Models such as causal support (Griffiths \& Tenenbaum, 2005) learn a new, larger model each iteration, and the number of possible models grows exponentially. In contrast, the number of computations to learn our model is constant each iteration.

## Experiment 1: Learning Causality

In this experiment, we test the model's ability to learn humanperceived causal relationships. For testing the algorithm, the ground truth is established by linking known causing actions to their fluent effects.

## Video Data Used

To test learning the C-AOG, videos were collected with a Microsoft Kinect, recording the color and depth images simultaneously. The scenes collected include multiple doorways, an elevator, and an office. Figure 1 shows some screenshots of the videos. The entire video collection lasts about 120 minutes, and contains 21 pre-specified action categories. There are 8 to 20 (sometimes simultaneous) instances of each action category.

In this experiment, we first use perfect action and fluent detections to demonstrate learning. We compare these results to those obtained with noisy detections (with varying levels of accuracy), such as would be output from the action and fluent detection system.

## Results and Discussion

Multiple Fluents Figure 2 shows plots of information gains for causal relations in the order pursued, separated by fluent. Causes are added to the model before non-causes with clear cutoffs for the door and light fluents. The cutoff between cause and non-cause is obscure for the computer monitor fluent because the model only acquired partial causal information (the preconditions of power and computer status are hidden).
Noisy Data Randomly flipping action detections leads to the curves shown in Figure 3. As more noise enters the system, the information gained by considering causal relations decreases. While learning works amid noisy scenes (many actions happening simultaneously), clean detections are important.


Figure 3: Information gains for causal relationships in the order pursued for the light fluent.

Hierarchical Action Selection and $\chi^{2}$ Where compound actions (e.g., in the doorway scene, unlocking with a key or entering a code, followed by pushing/pulling the door) are required for the effect, the causing actions may be located within varying levels of the action hierarchy.

For actions hierarchically related to each other in the Temporal AOG, our model incorporates their dependences, minimizing the chance that related actions are selected as causes. Figure 4 shows that Hellinger's $\chi^{2}$ measure (a $\chi^{2}$ that is less sensitive to low expected values in a contingency table (Ferguson, 1996)) fails to identify the correct causes, unable to account for dependence.


Figure 4: Pursuit order for hierarchical causes.

Long Delay, Causal Power, and $\Delta P$ Under the power PC theory (Cheng, 1997), perceptual causality is calculated as:

$$
\begin{equation*}
\text { causal power }=\frac{\Delta P}{P(\text { effect } \mid \text { not cause })} \tag{3}
\end{equation*}
$$



Frame Number (not to scale)
Figure 5: Sample of human judgment key frames.
where $\Delta P$ (Allan, 1980) is given by:

$$
\begin{equation*}
\Delta P=P(\text { effect } \mid \text { cause })-P(\text { effect } \mid \text { not cause }) \tag{4}
\end{equation*}
$$

For an elevator, the only detectable causing action for the door opening is pushing the elevator call button. In this example, our model outperforms causal power as shown in Figure $6 . \Delta P$ performs similarly to causal power.


Figure 6: Pursuit order for the elevator scene.

The failure of causal power and $\Delta P$ originates when an observed event (e.g., walking away) coincidentally always occurs with the true cause (e.g., pushing the elevator call button) and the true cause is not perfectly detected. Both measures favor $100 \%$ correlation, despite how rarely it occurred in the video. The learning method presented here incorporates the frequency that the relationship is observed by examining the full contingency table.

## Further Discussion

Results match exactly with human perceptions of the causal connections between actions and fluent changes, showing that the C-AOG is learnable from co-occurrence and the heuristics listed in the beginning (short temporal lag and agent actions cause fluent changes).

Our results are limited to the action and fluent categories that are pre-specified, despite the fact that many potentially confounding actions were included. Those quantities must be specified in advance so that appropriate detectors can be trained. It is possible, however, that different people would produce different bottom-level actions and fluents.

## Experiment 2: Inference Experiment

In this experiment, our model is validated against humans in the long-term reasoning task of inferring hidden fluent values.

## The Stimuli

Approximately 20 minutes of video data was captured using a Kinect in two scenes: a hallway and an office. Table 2 contains a summary of the fluents contained in the video, as well as the values each fluent can take. While many of these fluents are ordinarily viewable, they are ambiguous in the video (e.g., light status (ambient light may be from a window or a light) or water stream (resolution is not high enough to see it) in Figure 5).

Through a website, volunteer participants $(N=15)$ were shown the test video which paused at preset frames, e.g., those shown in Figure 5. Query points surround either a change in a fluent or a causing action. At each key frame, the participant was asked to assign a total of 100 points to all possible values of each fluent, according to his/her own recognition and reasoning for the events. Assignment of the points corresponded to the subjective probabilities of the fluent values. Each participant was allowed to revise previous judgments with information derived from subsequent frames.

## Reference Estimates

We compare the human responses to predicted fluent values by a baseline random noise model and by the C-AOG.

Baseline Estimate (Random Noise). For a baseline estimate, the hidden fluents were randomly assigned uniformly, without using any detection or causal information (e.g., $50 \%$ for LIGHT ON and $50 \%$ for OFF). The baseline estimate provides a discriminative reference against which we can see how well our model approximates human judgments.
Computer Estimate (The C-AOG). Detectable actions and fluent changes are first extracted from the videos and used

Table 2: List of fluents considered.

```
    Computer: ASLEEP/AWAKE
    Monitor Display: ON/OFF
    Monitor Power: ON/OFF
    Cup: MORE/LESS/SAME
    Water Stream: ON/OFF
    Light: ON/OFF
    Phone: ACTIVE/STANDBY
    Trash Can: MORE/LESS/SAME
    Agent: THIRSTY/SATIATED
    Agent: HAS_TRASH/NOT
```



Figure 7: Sample screenshots for noisy data.
as inputs to the C-AOG model.
The action grammar is pre-specified. Actions are manually segmented, and then poses captured by the Kinect camera are clustered. Temporal parsing transforms the clustered poses into hierarchically-labeled instances from the T-AOG. The maximum probability action detections are used as input.

Fluent changes are detected from the video with the GentleBoost algorithm (Friedman, Hastie, \& Tibshirani, 2000) on features extracted as shown in Figure 7. Non-maximum suppression provides the final detections of fluent changes.

These action and fluent detections (and their probabilities) are then processed with potential causal explanations under the C-AOG (by maximizing the posterior probability of Equation 1). The best-performing consistent causal description over the course of the video is then returned through the Viterbi algorithm (Forney Jr, 1973). Hidden fluents are imputed from this result.

## Results and Discussion

To visualize the results, human, computer, and baseline estimates are reduced to two dimensions using multi-dimensional scaling (MDS) according to the total variation distance between estimates, and plotted in Figure 8.

In the hallway dataset, both fluent and action detections contribute to the causal inference of hidden fluents. The computer performance is very similar to human performance as shown in Figure 8(a). The baseline is far from the cluster of computer and human estimates.

The office dataset only contains detections of actions; all fluents are hidden. The computer's performance is still an improvement over the baseline towards human-level performance, as shown in Figure 8(b).

Misinformation: Correcting Spatio-Temporal Detections In the hallway dataset, multiple changes in the light fluent were detected, yet no causing action was detected, presenting a common situation in vision-detections are usually imperfect. The C-AOG corrects these errors by balancing the maintenance of detections with the consistency of causal explanations. Figure 9 shows typical candidates of the results sorted in order of probability.

The C-AOG result was consistent with human judgments. Humans selected a single value for the light fluent for the duration of the video, but some selected ON while others chose OFF. This reinforces the need to have a probabilistic model


Figure 8: MDS plots of fluent value estimates. Blue dots: human estimates. Red squares: estimates using the C-AOG. Green triangles: baseline estimates. See Further Discussion for notes on the human variability.
capable of maintaining multiple interpretations; the C-AOG result included both solutions.

## Further Discussion

Even though the set of possible fluent values was provided to participants (significantly narrowing their available judgments), the MDS plots show wide variation in human responses. This is due to many factors. First, some participants initialized fluent values differently (e.g., light ON versus OFF in Figure 5), resulting in a large total variation distance. Also, some participants were more cautious than others, recording judgments close to 50/50 where others took an all-or-nothing approach to assigning judgments.


Figure 9: Given action and fluent detections that move the light fluent between ON and OFF without a causing action, the C-AOG prefers this to be explained by incorrect detections of the light fluent. The second most probable class of explanations is that two of the changes had causing actions that were missed by the detection.

As evidenced by the C-AOG's weaker performance, the office dataset was particularly challenging. Action detections were poor and no fluent detections were available to identify conflicts, leaving the system heavily dependent on those incorrect action detections. Despite this disadvantage, the CAOG still provided enough reasoning capability to outperform the baseline. This example underscores the importance of good vision-detection systems.

## Conclusions and Next Steps

In this paper, we have presented a probabilistic graphical grammar model to match human perception of causal relationships between actions and fluent changes, the Causal And-Or Graph (C-AOG).

Experiment 1 showed that the C-AOG of everyday activities can be learned, matching human perceptions of causal relationships. These causal relationships are even learnable amid noise, such as would be present in detection systems. Further, experiment 1 showed that our method models human judgments better than causal power and $\Delta P$.

Experiment 2 showed that the C-AOG can be used as a model of human perception grounded on video to impute values for hidden fluents. This experiment captures the inherent variability of human estimations when confronted with video, and highlights the need for a model that can probabilistically incorporate causality and vision.

One current limitation of the C-AOG is that, if a situation is unexplained, all possible parse graphs are assigned a low probability. In future work, we plan to investigate how adaptive learning can be used to incorporate new instances of fluents into the C-AOG.

## Acknowledgments

This work is supported by ONR MURI grant N00014-10-10933.

## References

Allan, L. (1980). A note on measurement of contingency between two binary variables in judgment tasks. Bulletin of the Psychonomic Society, 15(3), 147-149.
Carey, S. (2009). The origin of concepts. Oxford University Press.

Cheng, P. (1997). From covariation to causation: A causal power theory. Psychological review, 104(2), 367.
Ferguson, T. (1996). A course in large sample theory: Texts in statistical science (Vol. 38). Chapman \& Hall/CRC.
Forney Jr, G. (1973). The viterbi algorithm. Proceedings of the IEEE, 61(3), 268-278.
Friedman, J., Hastie, T., \& Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). The annals of statistics, 28(2), 337-407.
Griffiths, T., \& Tenenbaum, J. (2005). Structure and strength in causal induction. Cognitive Psychology, 51(4), 334384.

Griffiths, T., \& Tenenbaum, J. (2007). Two proposals for causal grammars. Causal learning: Psychology, philosophy, and computation, 323-345.
Heckerman, D. (1995). A bayesian approach to learning causal networks. In Proceedings of the eleventh conference on uncertainty in artificial intelligence (pp. 285-295).
Kullback, S., \& Leibler, R. (1951). On information and sufficiency. The Annals of Mathematical Statistics, 22(1), 7986.

Mueller, E. T. (2006). Commonsense reasoning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.
Pearl, J. (2009). Causality: Models, reasoning and inference (2nd ed.). New York, NY, USA: Cambridge University Press.
Pei, M., Jia, Y., \& Zhu, S.-C. (2011). Parsing video events with goal inference and intent prediction. In Computer vision (iccv), 2011 ieee international conference on (pp. 487-494).
Saxe, R., Tenenbaum, J., \& Carey, S. (2005). Secret agents inferences about hidden causes by 10 -and 12-month-old infants. Psychological Science, 16(12), 995-1001.
Zhu, S.-C., Wu, Y., \& Mumford, D. (1997). Minimax entropy principle and its application to texture modeling. Neural Computation, 9(8), 1627-1660.

# Seeing the forest for the trees predicts accumulation decisions 

## Helen Fischer (helen.fischer@uni-hd.de) ${ }^{1}$ Cleotilde Gonzalez (coty@cmu.edu) ${ }^{2}$

${ }^{1}$ Department of Psychology, Heidelberg University, 69117, Heidelberg, Germany<br>${ }^{2}$ Dynamic Decision Making Laboratory, Department of Social and Decision Sciences, Carnegie Mellon University Pittsburgh, PA 15213 USA


#### Abstract

Stock-flow (SF) systems involving the accumulation of a stock over time are pervasive in many areas of human life. However, people make consistent mistakes when regulating such systems, a phenomenon termed SF failure. We introduce holistic (global) versus analytic (local) processing as a cognitive mechanism underlying the hardly understood SF failure. Using a classic SF problem (department store task), we found that (a) solutions to SF problems were up to four times higher when a global task format highlighting global structure compared to a local task format highlighting local elements was used; (b) a more global processing style is connected to higher solution rates to the SF problem; and (c) procedurally priming participants with more global processing results in higher solution rates to the SF problem. In sum, our results point towards global-local processing as a basic explanation for SF failure.


Keywords: Stock-flow failure; global-local processing; dynamic systems.

Many decisions we make in our daily lives are aimed at keeping a system under control, or in equilibrium. For example, we aim at keeping our weight at a healthy stage, don't each too much, don't eat too little; we aim at keeping our bank accounts under control: buy the things that we need but don't spend too much. These types of systems are called dynamic stock management problems, where a stock (i.e., accumulation) is influenced by decisions to increase the stock (i.e., inflow) or to decrease the stock (i.e., outflow) (Sterman, 2000). Keeping a stock in balance implies that the outflow equals the inflow, that is, the stock does not change when the rate of increase equals the rate of decrease in the stock.
Dynamic stock management is extremely difficult to master even after extended amounts of practice (Diehl \& Sterman, 1995; Paich \& Sterman, 1993). The dominant explanation of these difficulties has been dynamic complexity (Sterman, 2000): the idea that systems that involve multiple decisions and delays between actions and observable outcomes create complex interdependencies that go beyond our cognitive capacity.
In light of these difficulties, more recent studies have pared back dynamic stock management tasks to their fundamental elements - one stock, one inflow and one outflow - and asked for simple 'one-shot' decisions about the system (Booth Sweeney \& Sterman, 2000; Cronin \& Gonzalez, 2007; Gonzalez \& Wong, 2012; Sterman, 2002).

Interestingly, even in these extremely simplified problems, a majority of people performs poorly (Cronin et al., 2009).
This general difficulty in understanding the dynamics of accumulation was termed "Stock-Flow Failure" (SF failure). There is very little research, however, aimed at understanding how people make decisions in these types of systems (Cronin, Gonzalez, \& Sterman, 2009). For example, it was repeatedly found that people have the erroneous tendency to perceive a stock's behavior as directly related to that of its flows (Booth Sweeney \& Sterman, 2000; Cronin et al., 2009; Cronin \& Gonzalez, 2007), a tendency termed correlation heuristic (Cronin et al., 2009). Although the correlation heuristic seems to be robust in SF failures, it remains a re-description of the typically observed behavior rather than an explanation of why the behavior occurs. A goal of the current research is to elucidate some cognitive mechanisms underlying SF failure. Specifically, we introduce Global-Local processing ${ }^{1}$ as a fundamental explanation.
We propose that, to make accurate accumulation decisions, one needs to process information globally and not locally. That is, one needs to see the forest and not the trees. For example, to make a prediction about the amount of money in our bank account at a point of time, we need to see broadly the predicted deposits and withdrawals over the preceding time periods.
Processing styles are content-free ways of perceiving the environment (Tulving \& Schacter, 1990). In global processing, one attends to objects holistically and focuses on the entire Gestalt by "zooming out"; in local processing, one attends to objects elementally and focuses on its details by "zooming in" (Navon, 1977; Schooler, 2002). Just as a global view on a Navon letter (global letter made up of smaller letters, Navon, 1977) means perceiving the whole form and not its component parts, a global view on dynamic systems should mean perceiving the systems' behavioral patterns and not its constituent parts. This should hold not only for complex systems containing a range of interacting variables, but even for "simple" SF systems, because in either case the behavior of the stock depends on the relation between in- and outflow, aggregated over time periods. That is, to regulate dynamic complexity in general, an abstraction process is needed from lower-level

[^263]representations (e.g. about a specific inflow at a specific point in time) to higher-order representations (e.g., about the overall relation between in-and outflow). We expect such a super-ordinate framework to enable the problem-solver not to view each component in isolation, but to view all components as structurally related parts of the system, thus allowing for inferences on the behavior of system as a whole.
To see how a human's tendency to process information globally or locally influences SF-reasoning, we measure individual differences in global-local processing and test participants with two task formats of a commonly used SF problem, the "Department Store" (DS) task (Sterman, 2000; Cronin et al., 2009). We argue that the SF task format that was used previously induced local instead of global processing of the problem and that a representation that induces global processing would lead participants to higher accuracy in their judgments of a stock.
In summary, we investigate the effects of three different aspects of global-local processing on SF reasoning: (a) global vs. local task format of SF problems, (b) individual differences in global vs. local processing and (c) global vs. local perceptual priming.

## Procedural Priming of global-local processing

The tendency to perceive the environment locally versus globally does not only exists as a prior bias in participants, but it can also be triggered, e.g. by instructing participants to focus on the global versus the local letters in the classic Navon-letter task (Navon, 1977; Macrae \& Lewis, 2002). Moreover, global-local processing styles can carry over to subsequent tasks (procedural priming). Such procedural priming must be distinguished from semantic priming in that "the how rather than the what is primed" (Förster \& Dannenberg, 2010, p. 176). As predicted by Schooler's theory of processing shifts (Schooler, Fiore \& Brandimonte, 1997), procedural priming effects can be transferappropriate or transfer-inappropriate. For example, after global procedural priming, participants generated more creative answers such as original uses for a brick than after local procedural priming (Friedman, Fishbach, Förster \& Werth, 2003), thus showing a transfer-appropriate shift.
Procedural priming affects both perceptual and conceptual processing by means of a common attentional mechanism that is used both on the perceptual (e.g. to perceive the global and not the local figure) and the conceptual level (e.g., to select the distant and not the proximal node within the semantic network; Friedman, Fishbach, Förster, Werth, 2003; Förster, 2009). A, say, broadened perceptual scope thus carries over to a broadened conceptual scope, resulting, for example, in more remote associates and higher creativity or the use of broader mental categories (Förster, 2012).

## The department store task

In the DS task, participant are presented with a graph showing the rate of people entering and leaving a department store each minute and over a $30-\mathrm{min}$. interval
(Fig.1). The stock is the accumulation of people in the store over the $30-\mathrm{min}$ interval, the inflow is people entering and outflow people leaving the store. Participants are then asked four questions as shown in the figure. The first two questions test whether participants can read the graph correctly, essentially measuring if they are able to identify the inflow and the outflow. The last two questions test whether participants can infer the stock's behavior based on the behavior of the flows over time.
The main measure of SF failure is the typically low solution rates to questions 3 and 4 (see detailed analyses of different kinds of errors in several publications such as Cronin et al., 2009 and Gonzalez \& Wong, 2012).
The SF failure was also demonstrated using bar charts, tables or texts listing the specific in- and outflows per minute (Cronin \& Gonzalez, 2007; Cronin et al., 2009), for different contents (Brunstein, Gonzalez, \& Kanter, 2010), and also when motivation and learning were induced (Cronin et al., 2009). Thus, so far, SF failure has not only proven to be a highly stable construct, but also the involved cognitive mechanisms remain unclear.
We expect global, as opposed to local, processing to be a beneficial cognitive strategy, however, for two reasons.
First, SF problems (or any problem, for that matter) consist of a set of surface details and an underlying relational structure. SF systems all adhere to the same structure: If the inflow exceeds the outflow, the stock increases and vice versa. Even though the SF structure is simple, "seeing" it is not, but is nevertheless crucial for problem-solving. Because local processing means searching for details, whereas global processing means searching for structural relations (Förster, 2009; Love et al., 1999), we expect global processing to be beneficial for detecting the SF structure and thus for problem solving.
Second, global processing was found to be connected to more superordinate category-use (Förster \& Dannenberg, 2010). Because processing in concrete and narrow categories (e.g., "In minute five, eight people enter, and in minute six, two people enter") represents an erroneous strategy, whereas processing in abstract and broad categories (e.g., "Overall, more people enter than leave") represents a helpful strategy for making inferences about the overall system behavior such as the stock, we expect global processing to be beneficial for problem solving.
The typical SF paradigm may arguably have primed local perceptual processing, however, because local features such as specific numbers of people were highly salient (Fig. 1). Participants might therefore get the impression that exact numbers need to be retrieved and worked with, thus using local processing. In our reasoning, however, it should be beneficial to induce the impression that specific numbers are merely constituent elements, and that the overall figure, the gist of the display needs to be perceived. Since in the Navon-letter-task, it was found that manipulating the relative salience of the local versus the global form triggered local versus global perceptual processing (see Kimchi, 1992, for a review), we expect a task format highlighting surface
elements (local format) to induce local processing and therefore to be detrimental, and a task format highlighting structure (global format) to induce global processing and therefore to be beneficial for SF performance.


Figure 1. Original department store task format as used in Cronin, Gonzalez and Sterman (2009).

In sum, we expect a connection between global-local processing and stock-flow reasoning. Our test of this assumption is threefold:

1. Task format hypothesis: A global task format should enhance solution rates compared to a local task format. We use the same SF problem (department store task) and vary the relative salience of local versus global features.
2. Individual differences hypothesis: Individual more global perceptual processing should be connected to higher solution rates to SF problems. We measure global-local perceptual processing style using a variant of the Navon-letter task, the Kimchi-Palmer figures task (Förster \& Dannenberg, 2010) and correlate participants' mean score with SF problem-solving performance.
3. Priming hypothesis: Priming global perceptual processing should induce a transfer-appropriate, whereas local perceptual processing should induce a transfer-inappropriate shift on subsequent problemsolving. That is, we except induced global perceptual processing to shift to global conceptual processing, resulting in high solution rates in SF problems, and vice versa. We will experimentally induce different perceptual processing styles in participants using the
maps task (Förster, 2005; 2009), and test their effect on solution rates in the department store task.

## Experiment

Participants. A total of 148 participants ( 80 female, 67 male, 1 unknown) with a mean age of 34.9 years ( $S D=12$, range $=18-64$ ) took part in the experiment via Internet. All participants were residents of the US and had completed at least High School, 33\% had a 4 -year college degree in a range of different fields, the largest groups being Business (10\%), Psychology (7\%) and English (3\%).

Materials. A 3(priming: global vs. local vs. control) x 2(task format: global vs. local) mixed design was used, with priming as the between-, and task format as the withinsubjects factor. To procedurally prime participants with a processing style (global vs. local vs. control), the maps task was used (see Förster, 2005; Förster et al., 2009). For each of seven trials, participants were presented with a state map displayed on the screen. In the global condition, participants were instructed to attend to the map in its entity in order to be able to describe its overall shape in one sentence. In the local condition, participants were instructed to attend to only the respective capital marked on the screen in order to be able to describe its location in one sentence. In the control group, participants were instructed to think about an item that characterizes the respective state in order to name it in one sentence. For all three conditions, participants subsequently typed one sentence into an input field while the respective map was still presented on the screen.
To test effects of task formats, the department store task was used in a global and a local format. The local format was very similar to the original format used, thus arguably highlighting local surface features, whereas the global format was designed to highlight global structure of the problem. For both task formats, the original introductory sentence and task display depicting in- and outflows was used (see Fig.1), but the answer options cannot be determined were replaced with 7-point Likert scales assessing subjective confidence: How confident are you in your answer? $0=$ Not confident at all and $7=$ very confident. This was done to assess whether participants have a reliable feeling for correctness as a function of the task format. In the local format, participants answered the following questions:

1. During which minute did the most people enter the store?
2. During which minute did the most people leave the store?
3. During which minute were the most people in the store?
4. During which minute were the fewest people in the store? In the global format, participants answered the following questions:
5. How are the people entering related to the people leaving the store between time periods 1 to 14 ? (a) More people entering than leaving (b) More people leaving than entering (c) Same amount of people entering and leaving.
6. How are the people entering related to the people leaving the store between time periods 14 to 30 ? (a) More people
entering than leaving (b) More people leaving than entering (c) Same amount of people entering and leaving.
7. How would you best describe the accumulation of the number of people in the store between time periods 1 to 14 ? (a) Increasing (b) Decreasing (c) Stable.
8. How would you best describe the accumulation of the number of people in the store between time periods 14 to 30 ? (a) Increasing (b) Decreasing (c) Stable.
To measure individual global-local processing styles, we used the Kimchi-Palmer-figures task (Förster \& Dannenberg, 2010). Participants were presented with triangles and squares that are made up of smaller triangles and squares. Participants indicated for each of 16 trials whether a target figure (e.g., a global square made up of local squares) was more similar to a sample figure that matched its global form or its local form. Display of the figures was counterbalanced with respect to the global (local) match appearing on the left (right). Mean ratings were then conducted for each participant, ranging from 0 (completely local processing style) to 1 (completely global processing style).
Since a bi-directional link exists between good versus bad mood and a global versus local processing style (Gasper \& Clore, 2002), and, in turn, mood is connected to problemsolving (Spering, Wagener \& Funke, 2005), we controlled for mood effects using two 7-point Likert-scales: How do you feel right now? $0=$ Very good and $7=$ very bad; $0=$ Very sad and $7=$ Very happy.
Procedure. Participants were told that they were going to take part in two tasks, one about visual perception, and one about problem solving. Participants first completed one of three randomly assigned between-subjects procedural priming treatments (maps task): global vs. local vs. neutral. Second, participants answered both the local and the global version of the department store task, in random order. Please note that being able to answer the first version correctly (say, the global) was no prerequisite for being able to answer the second version correctly (say, the local). Third, participants completed the Kimchi-Palmer-Figures-task and, as a final set of answers, they answered the mood questions.

## Results

Task format hypothesis. To test whether a global task format improves SF performance relative to a local task format, we compared solutions rates to SF tasks in both formats. For both SF questions 3 and 4, solution rates in the global format were higher than in the local task format (see Table 1). Moreover, mean confidence ratings in the local tasks were not connected to performance in the local tasks, $r$ $=.004$, $p=.48$, but confidence ratings in the global tasks were connected to performance in the global tasks, $r=.37$, $p$ $<.001$.
In sum, the task format hypothesis was confirmed: As expected, mean solution to the SF tasks were higher when a global relative to a local task format was used. Confidence ratings in both format might indicate, moreover, that participants are merely guessing in the local tasks, whereas
they have insight into the problem structure, and therefore a reliable feeling for correctness, in the global tasks.

Table 1. Percentage of participants who answered each of the two SF questions (questions 3 and question 4) correctly as a function of task format (global vs. local).

| Task format | Question 3 | Question 4 | $\chi^{2}$ |
| :--- | :--- | :--- | :---: |
| Local <br> $(\mathrm{n}=148)$ | $20 \%$ | $16 \%$ | $42.3^{* * *}$ |
| Global <br> (n=147) | $57 \%$ | $77 \%$ | $109.3^{* * *}$ |

Note. Local question 3 (4): "During which minute were the most (fewest) people in the store?". Global question 3 (4): "How would you best describe the accumulation of the number of people in the store between time periods 1 to 14 (14 to 30 )?". $* * * p<.001$

Individual differences hypothesis. To test whether globallocal processing styles affect SF-reasoning, processing style was correlated with mean correct solutions to all four SF tasks as a function of priming. After global priming, globallocal processing styles were not connected to mean SF solution, $r(50)=-.05, p=.37$; processing styles were connected to mean SF solutions, however, after local and no priming, $r(99)=.21, p=.02$. A median split was performed to directly compare SF solutions from participants with more global vs. more local processing styles. After no priming (control), and even more so after local priming, participants with a more global processing style performed better than participants with a more local processing style (Table 2).
In sum, the individual differences hypothesis was supported: As long as global priming did not induce a global processing style anyway, participants profited from a preexisting more global processing style when solving SF problems and achieved higher mean solutions than participants with a more local processing style.

Table 2. Mean solution (SD) rates to all four SF tasks as a function of processing style (local vs. global) and priming (local vs. global vs. control).

| Priming | Local <br> processing | Global <br> processing | $t$ |
| :--- | :--- | :--- | :--- |
| Global <br> Priming <br> $(\mathrm{n}=51)$ | $.44(.29)$ | $.43(.28)$ | -.10 |
| Control <br> $(\mathrm{n}=43)$ | $.37(.51)$ | $.52(.27)$ | $-1.85^{*}$ |
| Local Priming <br> $(\mathrm{n}=53)$ | $.31(.26)$ | $.46(.18)$ | $-2.44^{* *}$ |

Note. Local question 3 (4): "During which minute were the most (fewest) people in the store?". Global question 3 (4):
"How would you best describe the accumulation of the number of people in the store between time periods 1 to 14 (14 to 30 )?". ${ }^{*} p<.05,{ }^{* *} p<.01$

Priming hypothesis. To test the effect of priming on solution rates to SF tasks, local and no priming conditions were collapsed, since they did not produce any significant differences in either of the four SF tasks ( $p>.05$ ). To test differences in the number of correct solutions in SF tasks after global priming, two-sample z-tests were conducted. If participants answered the global SF tasks first, global priming had no effect on solution rates, $z=1.04, p=0.15 ; z$ $=0.89, p=0.19 ; z=0.46, p=0.36$ and $z=0.33, p=0.37$ (for the local question 3 and 4 , and the global question 3 and 4 , respectively). However, if participants answered the local questions first, global priming enhanced solution rates in the local SF question compared to local or no priming for question 3 ( $M=24 \%$ vs. $M=12 \%$ ) and 4 ( $M=18 \%$ vs. $M=$ $7 \%$ ), yielding marginal significance of $z=1.44, p=0.07$ and $z=0.33, p=0.06$, respectively. Global priming did not enhance solutions rates to the global tasks 3 ( $M=77 \%$ vs. $M=68 \%)$ and $4(M=88 \%$ vs. $M=87 \%)$, yielding $z=0.8, p$ $=0.21$ and $z=0.33, p=0.23$, respectively.
To control for mood-effects, we compared participants' mean mood in the three priming conditions. Results showed that participants' mood in the global priming condition ( $M=$ 5.8, $S D=1.8$ ) was not different from the local or no priming condition $(M=6.3, S D=1.8), t(149)=-1.58, p=.12$, indicating that the effect of global priming on problemsolving performance was not simply due to mood effects.
In sum, the priming hypothesis was marginally supported: Inducing a global processing style enhanced solutions to the local, but not the global SF questions. An additional priming effect on the local SF tasks did not occur, however, if global SF tasks were answered first, indicating that answering the global SF questions first served as a prime in itself.

## Discussion and Conclusion

The present experiment investigated the cognitive mechanisms underlying SF failure. We proposed globallocal processing as a fundamental, cognitive explanation and tested this notion using three different approaches: Global versus local task formats, individual differences in global-local processing and procedurally priming local versus global processing. Results generally supported our notion of global-local processing to affect SF performance.
In the department store task, participants profited immensely from a global task format highlighting structural relations between the system parts compared to the original local format highlighting features of the constituent parts. Specifically, solution rates in the global format were twice (question 3) or even more than four times as high (question 4) compared to the local format. One might argue, however, that the higher solution rates in the global task format were merely due to the greater amount of information since only the global task format referred to "time periods 1 to 14 " and "time periods 14 to 30 ", respectively. However, the global
task format was specifically designed to unveil the structure of the problem, so that a greater amount of information was inherent in the design of the task format. We would even suspect, moreover, that if questions 3 and 4 left out information about time periods, the global task format would still achieve higher solution rates, simply because questions 1 and 2 already allow participants to detect the problem structure. This, however, is for further research to decide.
As a further result, there was a connection between globallocal processing style and mean solutions to all SF tasks in the control group and the local priming group. Specifically, in line with our expectation, participants using overall global processing were better able to infer the overall behavior of the SF system, as measured by tasks testing an understanding of how the stock reacts to given in- and outflow progressions. Moreover, global processing could be procedurally primed in participants with the map task, resulting in (marginally significant) higher solution rates compared to local processing and the control group. This connection was only present for the local tasks, however, suggesting that participants do not profit from global processing and thus a search for structure, when the task format highlights structure in the first place. In sum, these results point towards global-local processing as a first explanation of the cognitive mechanisms involved in SFreasoning and SF failure.
The present results contribute to an understanding of how people deal with dynamic complexity. Our results merge to the conclusion that less successful participants seem to approach the problem in a piecemeal and concrete manner, whereas successful participants seem to approach the problem in a holistic and abstract manner. Interestingly, locally processing participants tend to stick more closely to the literal information given, whereas globally processing participants tend to go beyond the given information (Friedman \& Förster, 2001). With respect to SF systems, such literal use of information might result in lowest level, categorical representations (e.g., "the inflow is five"), whereas going beyond the information given might result in ordinal (e.g., "in minute 5, the inflow is smaller than the outflow") and increasingly higher-order representations (e.g., "overall, the inflow is bigger than the outflow"). Similarly, fuzzy-trace theory holds that people store two fundamentally different kinds of representations in memory: superficial verbatim representations such as exact numbers and meaning-based gist representations such as the "substance" of information (Reyna, 2012). Based on the present results, it seems plausible to speculate that, after local versus global processing of the task, participants hold fundamentally different representations of the task in memory. This, however, is for future research to decide.
In order to enhance people's ability to deal with SF systems, a range of strategies can be deduced from the present results. For example, it might be helpful to apply the principles of Gestalt psychology for pattern recognition to SF problem displays. With help of the law of good Gestalt,
for example, the constituent elements of SF problems could be grouped to imply global structure and regularity. In a different vein, one could try to enable people to process dynamic problems globally, for example by teaching strategies of abstraction, pattern recognition and induction of higher-order representations. Making use of the finding that a Gestalt view is helpful for dealing with SF problems, one could even try to make computers "see" the patterns in simple SF systems in order to regulate them.
It seems interesting to speculate in how far the benefit of a global, Gestalt view applies to complex systems in general. Systems containing a range of interacting variables can hardly be regulated using analytical strategies, because of limited cognitive capacities of the problem-solver, and because information in real-life is mostly fuzzy in nature. For both reasons, form-generating Gestalt principles could be helpful: They enable the problem-solver both to conceive of the system in its most economic form and to recognize basic similarities and therefore to re-use previously successful regulation strategies. Thus, recognizing patterns in systems might enable one to recognize and use similarities in a noisy world.

## Acknowledgments

This research was supported by a Defense Threat Reduction Agency (DTRA) grant to Cleotilde Gonzalez (HDTRA1-09-1-0053). The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

## References

Booth Sweeney, L. B. \& Sterman, J. D. 2000. Bathtub dynamics: initial results of a systems thinking inventory. System Dynamics Review 16, 249-286.
Brunstein, A., Gonzalez, C. and Kanter, S. (2010), Effects of domain experience in the stock-flow failure. System Dynamics Review, 26(4), 347-354.
Cronin, M., \& Gonzalez, C. (2007). Understanding the building blocks of system dynamics. System Dynamics Review, 23(1), 1-17.
Cronin, A.M., Gonzalez, C., \& Sterman, D. 2009. Why don't well-educated adults understand accumulation? A challenge to researchers, educators, and citizens. Organizational Behavior and Human Decision Process. 108, 116-130.
Diehl, E., \& Sterman, J. D. (1995). Effects of feedback complexity on dynamic decision making. Organizational Behavior and Human Decision Processes, 62(2), 198-215. Förster, J. (2009). Relations between perceptual and conceptual scope: How global versus local processing fits a focus on similarity versus dissimilarity. Journal of Experimental Psychology: General, 138,88-111.
Förster, J. \& Dannenberg, L. (2010). GLOMOsys: A systems account of global versus local processing. Psychological Inquiry, 21, 175-197.
Förster, J., Liberman, N.,\&Friedman, R. S. (2007). Seven principles of goal activation: A systematic approach to
distinguishing goal priming from priming of non-goal constructs. Personality and Social Psychology Review, 11, 211-233.
Friedman, R. S., \& Förster, J. (2001). The effects of promotion and prevention cues on creativity. Journal of Personality and Social Psychology, 81, 1001-1013. Friedman, R., Fishbach, A., Förster, J., \& Werth, L. (2003). Attentional priming effects on creativity. Creativity Research Journal, 15, 277-286.
Gasper, K., \& Clore, G. L. (2002). Attending to the big picture: Mood and global versus local processing of visual information. Psychological Science, 13, 33-39.
Gonzalez, C. \& Wong, H. 2012. Understanding stocks and flows through analogy. System Dynamics Review, 28(1), 327.

Kimchi, R. (1992). Primacy of wholistic processing and global/local paradigm: A critical review. Psychological Bulletin, 112, 24-38.
Love, B. C., Rouder, J. N., \& Wisniewski, E. J. (1999). A structural account of global and local processing. Cognitive Psychology, 38, 291-316.
Macrae, N. \& Lewis, H. 2002. Do I know you? Processing orientation and face recognition. Psychological Science, 13(2), 194-196.
Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. Cognitive Psychology, 9, 353-383.
Paich, Mark, \& John D. Sterman. 1993. Boom, bust, and failures to learn in experimental markets. Management Science, 39(12), 1439-1458.
Reyna. 2012. A new intuitionism: Meaning, memory and development in Fuzzy-trace Theory. Judgment and Decision Making, 7(3), 332-359.
Schooler, J.W., Fiore, S., \& Brandimonte, M. (1997). At a loss from words: Verbal overshadowing of perceptual memories. In D.L. Medin (Ed.), The psychology of learning and motivation: Advances in research and theory, (Vol. 37, pp. 291-340). San Diego, CA: Academic Press.
Schooler, J. W. (2002). Verbalization produces a transfer inappropriate processing shift. Applied Cognitive Psychology, 16, 989-997.
Spering, M., Wagener, D., \& Funke, J. (2005). The role of emotions in complex problem-solving. Cognition and Emotion, 19, 1252-1261.
Sterman, J. D. (2000). Business dynamics: Systems thinking and modeling for a complex world. New York: McGrawHill.
Sterman, J. D. (2002). All models are wrong: Reflections on becoming a systems scientist. System Dynamics Review, 18, 501-531.
Sterman, J.D. \& Booth Sweeney, L. 2002. Cloudy Skies: Assessing public understanding of global warming. System Dynamics Review 18, 207-240.
Tulving, E., \& Schacter, D. L. (1990). Priming and human memory systems. Science, 247, 301-306.

# Grammatical Aspect influences Event Duration Estimations: Evidence from Dutch 

Monique Flecken (m.flecken@donders.ru.nl)<br>Donders Centre for Cognition, Radboud University, PO Box 9104, 6500 HE Nijmegen, the Netherlands

Johannes Gerwien (jo.gerwien@uni-heidelberg.de)<br>Institut für Deutsch als Fremdsprachenphilologie, Heidelberg University Plöck 55, 69117 Heidelberg, Germany


#### Abstract

This study investigates the effect of grammatical aspect marking in Dutch sentences, on speakers' estimations of the duration of highly familiar, everyday events. We first established the 'inherent' or natural duration of different events (Exp. 1). This was then used for the manipulation of aspect (Exp. 2). Participants dragged a slider across the computer screen to estimate the duration of progressive and non-progressive event descriptions. Findings show how the progressive form extends duration estimations for short events, whereas it shortens the perceived duration of inherently medium and long events. We interpret this as psycholinguistic evidence for the function of aspect in Dutch, i.e., giving an 'inside' view of the event and focusing a specific internal time span of the event.


Keywords: grammatical aspect, event representation, Dutch.

## Introduction

Time is an important domain of human experience. For example, most people are able to roughly estimate how long it takes to open a window, to prepare a certain meal or to watch a movie, given normal circumstances. This information about the time course of events is part of world knowledge and our experience with different events and situations. When people talk about events (in finite sentences) the grammar of the language they speak may require them to make specific distinctions which relate to time explicit. They may be required to provide information on whether an event is taking place in the present, or took place in the past (grammatical category of tense). People may also need to specify whether an event has just begun, is in progression, or has reached a state of completion (grammatical category of aspect). However, it is not clear, in what ways world knowledge about temporal features of events and the distinctions provided by the language system interact: how do specific linguistic structures influence the way people represent events? In this study we address this question, and ask specifically how the use of aspectual verb forms in a sentence context affects people's general knowledge about the temporal contours of events, i.e., the duration of events.

## Background

Linguistic theories on grammatical aspect (also viewpoint or verbal aspect) state that the function of progressive aspect is
to modulate the inherent temporal contours of an event, thereby defocusing its boundaries (e.g., Comrie, 1976; Dahl, 2000). Specifically, it expresses a particular perspective on an event in that it is represented as a specific 'ongoing' instance of an event: For example, the semantic difference between 'he passes the ball' and 'he is passing the ball'. The progressive defocuses the boundaries of the event, to give an 'inside' view of a situation and thus 'highlight' its intermediate phases (e.g., Comrie, 1976). It is important to note that event descriptions that mark information regarding tense or aspect ('finite' expressions) do not directly refer to the time span defined by inherent temporal features of an event. With regard to aspect, Klein (1994) for example, distinguishes two temporal layers in language and describes aspect as denoting the relation between the linguistically unspecified time of an event (Time of Situation, TSit), and the specific time span that is being talked about (Topic Time, TT). The function of progressive aspect is to express that this time span (TT) falls within the boundaries of the event (Figure 1). This means that the time span at issue will be viewed as having extended duration (event marked as in progress), but it will be shorter than that of the entire event, as the time span in question does not include the boundary phases. Events not marked for progressive aspect, on the other hand, are unspecific in this regard and can include the entire event ('passes the ball'), thus highlighting a qualitatively different time span compared to events marked for progressive aspect.


Figure 1: Time-relational analysis of progressive aspect (cf. Klein, 1994)

The present study addresses the psycholinguistic reality of the above claims on the function of aspect in a sentence context: how exactly does this grammatical structure influence the way in which events, as expressed by verbal
predicates, are perceived? We focus on potential modulations of the perceived inherent or natural duration of events.

Initial steps in understanding the role of grammatical aspect in event conceptualization have been made from a psycholinguistic perspective in production and comprehension studies. A production study, comparing mono-/bilingual speakers of aspect and non-aspect languages, has looked at event descriptions and patterns in gaze allocation (measured with eye tracking), while subjects were preparing to speak about causative events (which involve an agent acting on an object, e.g., a person knitting a scarf) (Flecken, 2011a). Speakers who used progressive aspect to describe the events (English, Dutch), predominantly allocated gaze to features of the ongoing action (the knitting), and less to the agent of the event (which was the German pattern, non-aspect language). Progressive aspect thus focuses visual attention to ongoing aspects of an event, online, in production.

Comprehension studies have focused on the role of grammatical aspect for our understanding of situations or events, and the relations between different events. Magliano and Schleich (2000), for example, show how grammatical aspect constrains mental models of situations, when connected within a narrative structure. In their comparison of readers' comprehension of sentences marked for progressive and perfective aspect in the past tense, and embedded in a stretch of discourse, they found differences between comprehenders' conceptions of events, despite use of the same lexical information. When reading 'Betty was delivering a baby' versus 'Betty delivered a baby' two different mental representations of the event were formed with consequences for the way in which further contextual information was understood. Using a question-answer paradigm, they explicitly asked whether the critical events were finished or not, at specific points in the story line. In one experiment, they addressed the question whether 'general knowledge' on the duration of events interacted with aspect marking. They included events with a long and a short duration (long duration: ranging from 'watching a movie' to 'writing a novel'; short duration: ranging from 'scratching your nose' to 'packing a suitcase'). Likelihood scores indicated that 'long' events, marked as in progress by means of progressive aspect, were still perceived as ongoing at later sentence positions, in contrast to 'short' events. As they used a rather course measure (yes/no questions), we cannot be sure how exact this difference for aspect marking between short and long events is. Furthermore, the events within each category showed a great range in duration ratings, and included events that may not be familiar to all participants (such as 'giving birth'). A person's lack of experience with a situation or action may result in a less precise mental model of the event. Their findings may be interpreted as showing that the duration of the event described with progressive aspect is interpreted as prolonged, in comparison to the same event described by non-progressive verb forms.

Madden \& Zwaan (2003) also show how verbal aspect constrains speakers' representations of events. In a sentence picture matching task, with pictures showing events at different phases, they found that sentences marked with progressive aspect (in the past tense) elicit an equal amount of choices for pictures showing a completed or an incompleted event. The authors interpret this as showing that speakers can represent different phases of an event as in progression.

Bergen \& Wheeler (2010) also study the effect of aspect on 'mental simulation'. They find that speakers mentally simulate the nucleus of an action, when described in English sentences marked with progressive aspect, in contrast to sentences with perfective aspect.

In, e.g., Anderson et al. (2008) a different methodology was adopted, aiming to get a closer look at online processing of aspectually marked event sentences. They used a mouse tracking paradigm, in which speakers were asked to place a figure on a path, on its way to an endpoint, when listening to sentences describing motion events with and without progressive aspect in the past tense ('was walking to school' versus 'walked to school'). Figures were placed closer to the goal of the motion in the non-aspect condition, indicating that the past progressive focuses attention on internal phases of the past event.

These experiments provide important insights, as they reveal more clearly how aspect influences the processing of event structure. Important questions remain, however: For example, how does progressive aspect modulate event duration estimations for different event types?

In the present study, we take Dutch as our test case, as there is the advantage that this language allows use of sentences describing events in the present tense, both with and without morphological marking of progressive aspect. Production studies on Dutch have shown how progressive aspect is used frequently, but not for all event types. Unlike in English, use is not obligatory in any context (von Stutterheim, Carroll \& Klein, 2009; see for acceptability judgements of progressive and non-progressive event descriptions, Flecken, 2011b). With the investigation of a language other than English, we set out to explore whether the temporal relations described above for progressive aspect (Comrie, 1976; Klein, 1994) apply when Dutch speakers use the progressive aan het construction. In linguistic terms, progressives in different languages will follow the same temporal logic; but do speakers' responses reflect their role so as to modulate their perception of the internal phases of a dynamic situation when estimating event duration? Dutch is a language in which use of progressive aspect is not fully grammaticalized in contrast to English, for example. A comparison with English would have to be carried out on the basis of the same stimuli, however. We thus take first steps in exploring the influence of aspect marking on event duration in Dutch.

## Aims of the present study

In the present study, we draw a distinction between the 'inherent' duration of an event (i.e., the infinite and unspecified time interval or duration of an event, as expressed by bare (infinite) verb phrases, for example 'to write a paper'), which relates to world knowledge about the normal course of an event, and the finite expression of event duration by means of finite sentences or verb phrases, relating to a specific situation ('finite’ event duration). Finite expressions of event duration can include a verbal marker of progressive aspect, or not.

We ask whether speakers of Dutch perceive the duration of an event differently, depending on the specific type of verb form used (progressive or non-progressive) in a sentence context. An example is 'Wij zijn een artikel aan het schrijven' (lit.: we are a paper at-the-write; 'we are writing a paper', progressive verb) versus 'Wij schrijven een artikel' (we write a paper, non-progressive verb form). In Dutch, both instances relate to a specific event, taking place in the here and now.

Dutch speakers estimate the duration of events of different types, described in written sentences, by dragging a slider across a computer screen, using the mouse. Previous studies show how performance on a spatial task may accurately capture speakers' conceptions of temporal dimensions, such as duration (Casasanto \& Boroditsky, 2008). Event sentences will be presented twice, once in a progressive and once in a non-progressive condition. To prevent participants from memorizing the estimated duration of an event, as each is repeated, participants will estimate duration in the absence of a concrete time scale. Sentences will be presented in pairs, which remain the same in both conditions, meaning that the 'pair partner' of an event is thus the main point of reference for duration estimation, rather than an absolute time line. Estimations in minutes/seconds may be more susceptible to memory effects, and may overrule subtle effects of (non-) aspect marking.

We aim to find out how aspect interacts with the 'natural' or inherent duration of events, as judged by speakers on bare verb phrases describing actions and events.

## Experiment 1: 'Inherent' event duration

In experiment 1 native speakers of Dutch were asked to rate all kinds of everyday events and actions described by bare (non-finite) verb phrases (e.g., 'to walk the dog'). Three different samples were asked to rate their familiarity with the events, in how far they are imaginable (to what extent is the event likely to occur in the real world?) and the inherent duration of the events or actions. All ratings were carried out on a five-point scale.

## Method

Participants In total, 30 native speakers of Dutch took part in the experiment, consisting of three parts. They were $(\mathrm{PhD})$ students and postdoctoral researchers at Radboud

University Nijmegen (age range 19-35, balanced for gender).

Materials Stimuli used were written infinite action phrases (bare VPs) relating to everyday actions and events, and described with infinite verb phrases, e.g., 'to peel an apple', 'to open a can', 'to watch a football game'. Sentences were placed in an online questionnaire in a randomized order, and speakers were asked to give online ratings, and specify their age and gender. In total, there were 150 different events/actions.

Procedure Three different samples of 10 native speakers of Dutch took part in three different short experiments, designed as web questionnaires. First of all, the infinite action phrases were rated for familiarity ('how familiar are you with this type of action?') on a scale from 1 (highly unfamiliar) to 5 (highly familiar). Only highly familiar events were selected (ratings of 4 and 5) for Experiment 2. A second sample rated the phrases as to what extent the action was imaginable (rating 1: not imaginable at all, rating 5: highly imaginable). Furthermore, another sample of 10 speakers rated the duration of the infinite action phrases in relation to a 'standardized' event, i.e., to boil pasta, which was specified as lasting for about 7-8 minutes (rating 1 : much shorter than boiling pasta, rating 5: much longer than boiling pasta). This latter rating was conducted to ensure homogeneity of inherent event duration estimations.

## Results

The three rating tasks in Experiment 1 resulted in the selection of 78 different events. All other items were discarded due to a low degree of familiarity, the fact that they were not imaginable, or whether duration ratings showed a high degree of heterogeneity. All in all, 72 items were discarded. The 78 events were divided into three categories of inherent event duration (26 items in each category), on the basis of the duration ratings obtained: short (e.g., 'to turn a key', 'to light a candle'), medium (e.g., 'to set the table'; 'to polish a shoe') and long (e.g., 'to watch a dvd', 'to wash a car'). Items with an average rating of between 1 and 2 were characterized as 'short' events (range of ratings: $1-1.67$ ). Items with an average rating of between 4 and 5 were classified as 'long' events (range of ratings: $4.11-5$ ). Medium events were items with an average rating of between 1.67 and 4.11 .
The 78 items with homogeneously-rated inherent event duration, categorized in three groups (short, medium and long), were used as materials for Experiment 2.

## Experiment 2: 'Finite' event duration

In Experiment 2 we asked native speakers of Dutch to estimate the duration of events, as described in whole sentences, marked with or without progressive aspect.

## Method

Participants In the present study 27 native speakers of Dutch took part, who were all students at Radboud University (age range: 18-32, 16 female, 11 male), and did not have an advanced level of proficiency in a second (or third) language. This was established on the basis of their answers in a language background questionnaire. Students, who reported a stay of over three months in a foreign language country, were excluded from participation.

Materials Stimuli consisted of written sentences describing everyday situations and events. There were in total 78 items, describing 78 different events. Each item was used for a progressive and non-progressive sentence and paired with an item with matching inherent event duration. There were thus 13 pairs in each duration category (short-short pairs, medium-medium pairs, long-long pairs). For the pairings, care was taken to avoid any thematic or semantic relatedness between the two items. Sentences were presented as pairs to provide a kind of reference point for the duration estimations, within each trial. Pairs were always presented in the same aspect (either progressive or non-progressive). The agents of all actions (the subjects of the sentences) were described with two specific names, 'Jan' and 'Paul' in all cases.

Procedure Before subjects came to the lab, they were asked to carry out the same online familiarity rating task as in Experiment 1, dealing with all 78 bare event phrases. Ratings were again made on a scale from 1 to 5 . In the lab, subjects were told that on each trial they would read two sentences describing the situations in which two specific persons, i.e., 'Jan' and 'Paul' were involved right now. They were asked to imagine the situations of both Jan and Paul, and to estimate how long the two agents would be engaged in the activities described. Numbered sentences appeared below each other on a computer screen in a centred position. Within trials, sentences were of approximately the same length, to avoid any visual bias. Lower down, two sliders were presented and subjects were instructed to use the mouse to drag the sliders from left to right, starting with the top one, to estimate duration (Figure 2).


Figure 2: computer screen with sliders dragged slightly to the right (progressive aspect condition, 'long' events: 'John
is telling a fairytale', 'Paul is cleaning the bathroom')

Subjects were instructed that the further they dragged the mouse to the right, the longer they estimated the agent to be engaged in the activity. Furthermore, it was explicitly stated that if they dragged the slider to the right only slightly, this would mean that Jan or Paul are engaged in the activity for a very short time. If they dragged the slider to the utmost right, this would mean that Jan or Paul are performing the activity for a long period of time. The particular part of the slider that was dragged, turned red. Subjects were able to adjust their estimations. After estimating the duration of both sentences, they proceeded to the next trial by clicking a button.

In order to ensure that participants were actually aware of the surface sentence structure, and did not only focus on the bare event characteristics, a question relating to the contents of one of the preceding sentences appeared randomly. Subjects were asked to decide whether they had read that sentence before, by clicking yes or no on a button box. The question sentences were correct half of the time, and the other half contained errors with regard to the type of object described (e.g., for sentence 2 above: Paul is cleaning the kitchen) or the type of aspect used (e.g., for sentence 2: Paul cleans the bathroom). Each sentence pair appeared twice, once in the non-progressive condition, and once in the progressive condition. All trials were pseudo-randomized, so that each repeated item appeared in the second half of the experiment (the second set of 39 trials), to ensure enough distance between repeated items. The occurrence of progressive or non-progressive sentence pairs in the first or second half of the experiment was varied between subjects.

After filling out a sociolinguistic questionnaire, subjects were asked to estimate the precise duration of the different events (described in bare VPs) in minutes (pencil-and-paper test). This was done to double-check, whether the events were rated as belonging to the same duration categories as those established in Experiment 1.

## Results

a) Familiarity ratings All 78 event phrases were rated as familiar (4) to highly familiar (5), replicating the results from experiment 1.
b) Online event duration estimations of sentences For the analysis, we focused on the values of the x-coordinates on the computer screen only, equalling the distance the mouse was dragged towards the right side of the screen. We analyzed our data using mixed effects models ( R , lme4 package). Our goal was to fit a model that would explain the estimations made by the subjects as the result of the impact of various variables, i.e. fixed and random effects. Our fixed effects were 'inherent duration' ('dur') (long, medium, short) and 'aspect' (progressive, non-progressive). The variables were coded as follows: for 'dur', the short event category was coded as the base level ( -1 short, 0 long, 1 medium) and for 'aspect' we coded the non-progressive verb form as the base level ( -1 non progressive form, 1 aan het form). We also aimed at controlling further influences
caused by the experimental design, by taking into account random effects in our model.
The random factors we originally considered were subject, item, and pair. For 'subject' we included by-subject random intercepts, as well as a by-subject random slope, which allowed the predictions for 'inherent duration' to shift by a fixed amount for each subject ${ }^{1}$. With respect to the random factor 'item', two things are important. First, every item (event) belongs to one and only one event duration category. Item is thus a nested random factor. We incorporated this by adding a variable which covered the 'item:dur' interrelationship; this term was also included as a random factor. Second, subjects always rated event pairs and not single events. The pairing of items remained fixed throughout the experiment, for each subject. We thus did not add pair as a separate random factor, as the nested 'item:dur' term would sufficiently capture the variance stemming from random item selection. In general, we follow an approach by Barr et al. (2012) in which the authors argue for a maximal approach, that is, "valid statistical inferences using LMEMs require maximal random-effects structures wherever possible ..."(p.1).

We log-transformed and centred all duration estimations (see Footnote 1 for the formula in R).

Let us turn directly to the fixed effects section in our model (Table 1 below: asp. 1 is aan het condition; dur. 0 is long, dur. 1 is medium event type)

Table 1: Fixed effects in the mixed model

|  | Estimate | Std. Error | t value | p value |
| :--- | :--- | :--- | :--- | :--- |
| Intercept | -0.96918 | 0.10877 | -8.910 |  |
| dur.0 | 1.90106 | 0.12323 | 15.427 | $<.001^{* *}$ |
| dur.1 | 1.00654 | 0.09792 | 10.279 | $<.001^{* *}$ |
| asp.1 | 0.04819 | 0.02364 | 2.039 | $.041^{*}$ |
| dur.0:asp.1 | -0.07118 | 0.03343 | -2.129 | $.033^{*}$ |
| dur.1:asp.1 | -0.06721 | 0.03341 | -2.012 | $.044^{*}$ |

As predicted, for 'inherent duration' ('dur') we find high tvalues (long events $t=15.43$; medium events $\mathrm{t}=10.28$ ), showing that, in contrast to the base level (short events), the two other event types are estimated as significantly longer. There was a significant main effect of 'aspect' $(\mathrm{p}=.041)^{2}$, meaning that short events were estimated as having a longer duration in sentences marked with the aan het form, when compared to the same events described with nonprogressive verbs. Looking at the interaction effects, we find that, compared to our base level, medium and long events marked with the progressive form are estimated as significantly shorter (both p values > .05).

[^264]To exclude the possibility that the above pattern of results is due to the presence of outliers, 32 extreme values (. $008 \%$ ), with a standardized residual at a distance greater than 3 standard deviations from zero, were removed from the data, and the model was refitted.

Table 2: Fixed effects in the mixed model on trimmed data

|  | Estimate | Std. Error | t value | p value |
| :--- | :--- | :--- | :--- | :--- |
| Intercept | -0.97638 | 0.10859 | -8.991 |  |
| dur.0 | 1.93245 | 0.12593 | 15.345 | $<.001^{* *}$ |
| dur.1 | 1.03100 | 0.09883 | 10.432 | $<.001^{* *}$ |
| asp.1 | 0.04781 | 0.02157 | 2.217 | $.027^{*}$ |
| dur.0:asp.1 | -0.06942 | 0.03056 | -2.272 | $.023^{*}$ |
| dur.1:asp.1 | -0.08131 | 0.03052 | -2.665 | $.008^{*}$ |

The trimmed model (Table 2) shows the same significant results for the predictors and their interactions. We conclude that the statistical inferences made in the original model are not confounded by extreme values in the data set.
c) Inherent event duration estimations (bare VPs) Table 3 below displays the average and SD of the duration estimations for the infinite event phrases; these estimations were carried out after the actual experiment. The numbers displayed are duration estimations in minutes.

Table 3: Inherent duration estimations, in minutes

|  | Short | Medium | Long |
| :--- | :--- | :--- | :--- |
| Average | 2.25 | 11.20 | 80.24 |
| SD | 1.52 | 11.27 | 50.52 |
| Lower | 1 | 2 | 14.50 |
| Upper | 6.01 | 58.67 | 206.38 |

The absolute duration estimations support the division into the three categories of inherent event duration, based on Experiment 1.

## General discussion

In Experiment 1, we established three categories of highly familiar, everyday events of different 'inherent' duration (short, medium, long events), on the basis of three rating tasks. In Experiment 2, we used those items and specifically assessed the effect of aspect marking on subjects' duration estimations of the three event types, by means of the 'drag-the-slider-technique'.

First of all, with respect to the different 'inherent' event duration categories, the findings indicate that the method is valid; medium and long events were estimated as lasting significantly longer than short events. The duration estimations made by subjects using a slider on the computer screen, without a fixed time scale, reflect the time spans which are inherently part of the conceptual representation of events, showing that spatial tasks are informative about people's thinking about time (Casasanto \& Boroditsky, 2008).

Second, we find a significant interaction between aspect marking and inherent event duration, suggesting that aspect affects the perceived duration of events described in sentences in a specific way: In Dutch, short events are estimated as having a longer duration when described in the progressive aan het form, whereas medium and long events are estimated as having a shorter duration, when compared to estimations of the same events described by nonprogressive verb forms.

The mechanism underlying the patterns found is explained by a time-relational analysis of aspect: As described above (Klein 1994), progressive aspect marks that the time span being talked about (TT) is placed within the total event time (TSit) whereas unmarked (non-progressive) verb forms are unspecific in this regard. With the progressive, an internal time span is focused and explicitly viewed as 'in progress'. Short events inherently have a short TSit, which can include a transition phase or change in state ('to open a bottle': from 'not open bottle' to 'open bottle'). If language users describe such an event with progressive aspect, the time span at issue is located within the event time (Tsit), and attention is thus directed to the transition phase. Language users experience this as stretching and prolonging the duration of the event in their mental model. For medium/long events, the temporal boundaries (beginning and end) lie further apart (TSit is longer). There are also phases with changes of state with the event 'to repair a bicycle', for example, but it will typically have longer duration. When events are described with progressive verbs, attention is directed to a specific time interval that lies in between the beginning and end of TSit, and, crucially, it does not extend over the entire event. The duration of the event will thus be perceived as shorter, compared to the total time span for the entire event (TSit), as expressed in non-progressive sentences.

World-knowledge about a specific event seems to play a role for the interpretation of aspect - and both layers of duration interact in our subjects' mental models of the events. In general, we provide further evidence that grammatical aspect influences people's representations of events or situations (e.g., Anderson et al., 2008; Madden \& Zwaan, 2003).

## Conclusions

In this study we investigated in how far grammatical aspect has an influence on how people mentally represent the duration of everyday, highly familiar events, described in Dutch sentences. We distinguish between two 'layers' of event duration, which are packaged together in sentences, and which both contribute to the perceived duration of an event. The first 'layer' consists of the 'inherent' duration of an event, which is based on world knowledge. The second layer consists of 'finite' temporal information, expressed by tense and aspect. Our results imply that the inherent duration of events is shared among speakers of a language/culture. This inherent event duration is modulated
by grammatical aspect (aan het in Dutch; previous studies show this for the -ing form in English).
We find psycholinguistic evidence for the function of grammatical aspect in Dutch. By means of progressive aspect, speakers take an 'inside' perspective on an event, by selecting a time interval that falls within the total time period of the event - leading to a complex interaction between aspect marking and the inherent duration of events.

## Acknowledgements

This study was funded by NWO Veni-grant 016.124.107 (to M. Flecken). We thank Mary Carroll for proofreading and commenting on an earlier version of this paper.

## References

Anderson, S., Matlock, T., Fausey, C., \& Spivey, M. (2008). On the path to understanding on-line processing of grammatical aspect. In B. Love, K. McRae, V. Sloutsky (Eds.), Proceedings of the 30th Annual Conference of the Cognitive Science Society (pp. 143-148). Mahwah, NJ: Lawrence Erlbaum.
Barr, D., Levy, R., Scheepers, C. \& Tily, H. (2013). Random effects structure in mixed-effects models: Keep it maximal. Journal of Memory and Language, 68, 255278.

Bergen, B., \& Wheeler, K. (2010). Grammatical aspect and mental simulation. Brain \& Language, 112, 150-158.
Casasanto, D. \& Boroditsky, L. (2008). Time in the Mind: Using space to think about time. Cognition, 106, 579-593
Comrie, B. (1976). Aspect: An introduction to the study of verbal aspect and related problems. Cambridge: CUP.
Dahl, Ø. (Ed.) (2000). Tense and Aspect in the Languages of Europe. Berlin: Mouton de Gruyter.
Flecken, M. (2011a). Event conceptualization by early bilinguals. Insights from linguistic and eye tracking data. Bilingualism: Language \& Cognition, 14 (1), 61-77.
Flecken, M. (2011b). What native speaker judgements tell us about the grammaticalization of a progressive aspectual marker in Dutch. Linguistics, 49 (3), 479-524.
Klein, W. (1994). Time in language. London: Routledge.
Madden, C. \& Zwaan, R. (2003). How does verb aspect constrain event representations? Memory \& Cognition, 31, 663-672.
Magliano, J. \& Schleich, M. (2000). Verb aspect and situation models. Discourse processes, 29 (2), 83-112.
Stutterheim, C. v, Carroll, M. \& Klein, W. (2009). New perspectives in analyzing aspectual distinctions across languages. In W. Klein \& P. Li (Eds.), The expression of time (pp. 195-216). Berlin: Mouton de Gruyter.

# Time (also) flies from left to right... if it is needed! 

Andrea Flumini (andrea.flumini2@unibo.it)<br>Department of Psychology, University of Bologna<br>Viale Berti Pichat 5, 40100 Bologna, Italy<br>Julio Santiago (santiago@ugr.es)<br>Department of Experimental Psychology, University of Granada<br>Campus de Cartuja, 18071 Granada, Spain


#### Abstract

The TIME IS SPACE metaphor consists in the use of a spatial mental time line (either left-right or front-back) to represent time. One of the issues still to be resolved is whether these space-time mappings can be automatically activated independently from the goals of the task. Prior attempts to settle this issue have failed to match adequately the temporally relevant and irrelevant tasks. In the present study we presented Spanish verbs and nonverbs conjugated in past and future forms in both a time judgment and a lexical decision task. Results showed that the left-right space-time mapping is only active when the task requires temporal discrimination, speaking against an automatic activation of the mental time line.


Keywords: time; space; mental time line; automaticity; flexibility; embodied cognition; conceptual metaphor.

## Introduction

As a response to the symbol grounding problem (Harnad, 1990), the Embodied and Grounded Cognition view (e.g., Barsalou, 1999; Gibbs, 2006; Lakoff \& Johnson, 1987) suggests that abstract concepts need to be grounded on concrete domains (i.e., those more directly based on sensory-motor experiences) in order to gain meaning. Under this view, language processing elicits an embodied simulation which is carried on by the very same neural systems used by perception, emotion and action (Barsalou, 2008; Gallese, 2008; Glenberg \& Gallese, 2011; Glenberg et al, 2008). When abstract concepts are referred to, such simulation follows the guide of stored mappings between abstract and concrete concepts. One line of support for this idea comes from empirical studies on the abstract domain of TIME, which seems to be grounded on the concrete domain of SPACE. Response time studies have reported interactions between the processing of the temporal reference of words and sentences and a variety of response mappings: lateralized key presses, forward-backward manual movements, vocal responses (e.g., Boroditsky, 2001; Ouellet et al, 2010b; Santiago et al, 2007; Sell \& Kaschack, 2011; Torralbo et al, 2006; Ulrich \& Maienborn, 2010; Ulrich et al, 2012).
Space-time congruency effects are part of a wide family of studies that manipulate concrete and abstract dimensions in tasks that require elaborating and responding to aspects of the abstract dimension. In this context, modulations due to task-irrelevant concrete dimensions are often found on the processing of the abstract, task-relevant dimension. The resulting
metaphoric congruency effect has been interpreted as the index of the use of underlying concrete representations to organize the abstract dimension, as i.e. in the SNARC effect (Dehaene et al, 1993).
Conceptual Metaphor Theory, which has a longstanding support from linguistics and psychological studies (e.g., Boroditsky, 2000; Clark, 1973; Lakoff \& Johnson, 1980, 1999; Núñez \& Sweetser, 2006; Talmy, 2000), pointed out that our vocabulary about abstract concepts has concrete roots. But one of the most interesting consequences of the empirical findings on conceptual metaphors has been the discovery of the existence of metaphoric mappings not explicitly attested in language (for a review, see Santiago et al 2011). In the last years, the most studied example has been one TIME IS SPACE metaphor, which maps temporal reference onto the left-right horizontal spatial axis. In contrast to the mapping of time onto the frontback axis, which is explicitly attested in many languages (e.g., Sell \& Kaschack, 2011; Torralbo et al, 2006; Ulrich et al, 2012), in his review of cross-linguistic space-time metaphors Radden (2004) observed a total lack of linguistic conventions directly referring in speech to a horizontal left-right time dimension. However, we are all used to conventional associations of time as flowing from left to right (or right to left) along a horizontal axis in written language, graphs, and in many types of graphic devices (e.g., comic strips, calendars, etc.).
The interpretation of conceptual congruency effects as indexes of stable semantic memory mappings has been clearly contradicted by recent experimental results. There is evidence in the literature of different degrees of flexibility/automaticity depending on the abstract dimension studied, the task and materials used, the kind of mappings which are evaluated (Santiago et al, 2011). Nowadays, there seems to be a well-motivated support to the idea that conceptual congruency effects could be of a very contextual nature (e.g., Torralbo et al, 2006; Santiago et al, 2008; Santiago et al, 2012; Lakens et al, 2012).
One of the strongest cases of automatic activation has been observed for the mapping of affective evaluation to front-back responses: participants are faster in responding to positive and negative items by pulling and pushing a lever, respectively (e.g., Chen \& Bargh, 1999). This occurs not only when the decision is based on the valence of the stimuli, but also when performing a lexical decision task (Wentura et al, 2000) and even a stimulus detection task (Chen \& Bargh, 1999), which minimize the taskrelevance of the evaluative dimension. In contrast, space-
time mappings do not seem to be activated so automatically. Recently, two studies extending prior findings with temporal words (e.g., Santiago et al, 2007) to full sentences have tried to address the question of whether it is possible to observe an automatic activation of the mental time-line in an implicit task, investigating both the left-right (Ulrich \& Maienborn, 2010) and the front-back (Ulrich et al, 2012) axes. These studies asked participants to carry out both an explicit temporal judgment task and a sensicality judgment task, observing space-time congruency effects only on the former.
The findings of Ulrich \& Maienborn (2010) on the leftright mental time line left open the possibility that participants did not need to process the temporal reference of sentences in the sensicality judgment task. The nonsensical sentences were constructed by matching an agent and an object that do not fulfill the meaning restrictions of the verb (i.e, as in the past sentence "The fir trees have put on their coat while bathing", or in the future sentence "On next Sunday, the town-hall will marry the pea"). In order to judge whether these sentences are sensible or not, participants might have only assessed whether the action mentioned by the verb can be done by the actor (with the object) on the patient. In other words, whether the arguments fulfill the meaning constraints imposed by the verb. In order to control for this possibility, in their study on the front-back mental time line, Ulrich et al (2012) asked participants in the sensicality task to also perform a time judgment for each sentence at the end of the trial. Again, they failed to observe any interaction between response direction and temporal reference.
Several possibilities are left open by these two studies. A first one is that participants split their judgments into two sequential phases: they first focused on assessing meaning consistency, started response, and then assessed whether the sentence referred to a past or future event (in which case, the effect of temporal reference would be missed by the latency measure). A second, and very interesting possibility is that meaning access at sentence level is less automatic than at the word level, because the meaning of the sequence of words needs to be composed into the overall sentence meaning. Finally, it might be the case that the activation of the front-back time line is not automatic, but we cannot still be certain whether this is also the case for the left-right mental time line (due to the methodological concerns discussed above). A more automatic left-right time line would be consistent with findings of automatic activation of left-right space in tasks that required the processing of ordinal sequences (either learned on the spot or previously known) when the order dimension was completely irrelevant for the task (e.g., Gevers et al, 2004; Previtali et al, 2010), as well as with the well-known SNARC effect in parity tasks (Dehaene et al, 1993). It is clear that the issue of automatic activation of the mental time line is still far from being solved.
In our study we wanted to address simultaneously several of these possibilities. We focused on the processing of time-related single words with left and right responding (thereby testing the activation of the left-right
mental time line) in both time-relevant and time-irrelevant tasks.
To create our materials, we selected isolated Spanish verbs with an intransitive reading (e.g., "dormir" - to sleep). As Spanish is a pronoun drop language, when these verbs are conjugated in past or future tense, they represent a full sentence (e.g., "durmió" means "He slept"). However, their meaning is acquired in a single fixation and through the activation of a single lexical item. So, the chances of a slower, more compositional comprehension strategy are lower. In order to create the nonwords, we modified the set of verbs by changing only one letter in their morphological stem. Therefore, the nonverbs did not pop out as such (e.g., "dormir" was changed to "dorpir"). Moreover, the nonverbs were also inflected in past and future ("durpió"). In this way, we made sure that in order to distinguish the existing from the non-existing verbs, participants had to pay close attention and deeply elaborate the stimulus. We presented these stimuli in a temporal judgment task (decide whether the stimulus refers to the past or the future; Experiment 1) and in a lexical decision task (decide whether the stimulus is a word or a nonword; Experiment 2) with lateralized manual responses. If the left-right space-time mapping can be activated automatically, both experiments should render significant space-time congruency effects. Otherwise, they should arise only in Experiment 1.

## Experiment 1

Experiment 1 used centrally presented Spanish tensed verbs (technically corresponding to full sentences) and nonverbs in an explicit temporal judgment task. Responses were given by means of bimanual lateralized key presses.

## Method

Participants 24 Psychology students from the University of Granada (5 males; age range 19-26 y.; 2 left-handed by self-report) participated for course credit. All participants had normal or corrected-to-normal vision and were naive as to the purposes of the experiment.

Materials We selected 148 Spanish verbs which are intransitive verbs, or at least allow a (very common) intransitive use. Such a kind of verbs was chosen because Spanish is a pro-drop language, so the subject of a verb can be dropped from the sentence. Thus, single conjugated intransitive verbs as used here can stand as full, grammatically correct and sensible sentences. In order to create the nonword set, each verb was modified by changing one letter in its stem, with the constraint of resulting in pronounceable phoneme sequences in Spanish.
The 148 verbs and 148 nonverbs were then conjugated in both the simple past perfect indicative and the simple future indicative (all six possible grammatical persons were more or less equally represented over the set). This resulted in 592 experimental stimuli of four types: past and future verbs, and past and future nonverbs. This total set was randomly divided into four lists of 148 stimuli
each, avoiding item repetition. For example, from the item "faltar/falbar" the following four third person singular tensed versions were created: "faltó" (past verb), "faltará" (future verb), "falbó" (past nonverb), "falbará" (future nonverb), and each of them was randomly assigned to one of the four different lists. Each list was then composed of 37 items of each of the four stimulus types.

Procedure Stimuli were presented centered on a computer screen (Courier New font, 38 points, lower case), black printed on white background. Participants sat at a distance of 60 cm from the computer screen, and placed their left index finger on the Q key and their right index finger on the 9 key of the numerical keyboard in a standard QWERTY keyboard. The distance between response keys was 32 cm . Each trial began with the presentation of a central fixation cross ( 500 ms ) followed by the target verb, that remained on screen until a response was made. Incorrect trials were followed by a 500 ms red uppercase X at the same location of the stimulus. Each incorrect trial was then followed by a 1000 ms blank screen. Correct trials were followed by a 1500 ms blank screen. Participants were instructed to decide whether the presented verb or nonverb referred to either the past or the future.

The experiment was divided into two blocks of 148 trials (separated by a two minutes break) in which the same list of stimuli were responded to using two different mappings of responses (past/future) to keys (left/right). The order of presentation of the two mappings was counterbalanced over participants. We did not control for factors known to affect word recognition times, such as frequency, length, or age of acquisition, because the theoretically interesting effect is the interaction between temporal reference and response hand when participants process the very same list of stimuli using the two possible response-key mappings. Each block was preceded by a short 4 trials training block using different stimuli. The experiment was programmed and run using E-prime 2.0.

Design and Analysis Data were analyzed using a mixed factorial ANOVA with the within-subjects factors Lexical status (word vs. nonword) x Temporal reference (past vs. future) x Key (left vs. right). Counterbalance was included in the design as a between-subjects factor in order to reduce noise, but because it is of no theoretical relevance, its main effect or interactions are not reported here.

## Results

Errors occurred in $6.43 \%$ of trials and were analyzed independently. Reaction times (RTs) exceeding 2 standard deviations from each participant's mean were excluded from the analysis, leading to the removal of an additional $12.01 \%$ of data.

The ANOVA on RTs revealed a significant main effect of the factor Lexical status $(F(1,22)=36.55, M S e=$ 18071.56, $p<.001$ ), due to longer latencies for nonwords ( 1106 ms ) than words ( 989 ms ). This result is completely in line with the psycholinguistic literature about the
lexicality effect (e.g., Kinoshita el al., 2004; Pagliuca et al, 2010) and shows that participants were unable to focus only on the verb inflection in order to give their responses. There was also a Lexical status x Temporal reference interaction $(F(1,44)=13.25, M S e=3451.7, p<.01)$, due to past words being faster than future words ( 982 ms vs. 996 ms , respectively), whereas future nonwords were faster than past nonwords ( 1130 ms vs. 1083 ms , respectively). Finally, and most relevant to current concerns, there was also a significant interaction between Temporal reference and $\operatorname{Key}(F(1,44)=12.03, M S e=$ 16640.96, $p<.01$ ), which showed faster responses to past verbs and nonverbs with the left than with the right hand $(1027 \mathrm{~ms}$ vs. 1085 ms , respectively; Newman-Keuls $p$ $<.05$ ) and to future verbs and nonverbs with the right than with the left hand ( 1004 ms vs. 1075 ms , respectively; Newman-Keuls $p<.05$ ). This interaction was not modulated by Lexical status $(F(1,44)=1.03, M S e=$ 7707.96, $p=.32$; Word: Past verbs - left hand $M=943 \mathrm{~ms}$, Past verbs - right hand $M=1022 \mathrm{~ms}$, Future verbs - left hand $M=1035 \mathrm{~ms}$, Future verbs - right hand $M=958 \mathrm{~ms}$; Nonword: Past verbs - left hand $M=1111 \mathrm{~ms}$, Past verbs right hand $M=1150 \mathrm{~ms}$, Future verbs - left hand $M=$ 1116 ms , Future verbs - right hand $M=1050 \mathrm{~ms}$ ).

The analyses of accuracy revealed only a main effect of Lexical status $(F(1,22)=6.03, M S e=2.43, p<.05)$, which confirmed a greater easiness for participants in processing and responding to words than nonwords (2.1 vs. 2.7 mean errors, respectively).

No other main effects or interactions were significant.


Figure 1. Mean latencies for the factors Temporal reference and Key in Experiment 1.

## Discussion

Experiment 1 found a significant interaction between temporal reference of the stimulus and side of response, taking the form of a standard left-right space-time congruency effect: responses to past sentences were faster with the left hand and responses to future sentences were faster with the right hand, independently of their lexicality. These results replicate and extend prior findings in the literature (e.g., Torralbo el al., 2006; Santiago el al., 2007; Ulrich \& Maienborn, 2010), suggesting, firstly, that a left-to-right mental time-line have been activated in this task, and secondly, that lexical status does not modulate its activation. In other words, the mental time line is used to process the temporal reference of both meaningful and
meaningless words (which are also simple sentences in Spanish).

## Experiment 2

Experiment 2 only differed from the prior experiment in the task instructions: participants were asked to decide whether the stimuli were real Spanish verbs or not. Thus, they carried out a lexical decision task for which temporal reference is irrelevant. The design of the experimental materials made sure that temporal reference information was equally present and salient in both the words and the nonwords.

## Method

Participants 24 Psychology students from the University of Granada (1 male; age range 20-25 y.; 4 left-handed by self-report) participated for course credit. All participants had normal or corrected-to-normal vision and were naive as to the purposes of the experiment.

Materials and procedure Everything was identical to Experiment 1, with the only exception of the instructions: participants decided whether the stimuli were real Spanish verbs or not.

## Design and Analysis

The data were analyzed using a mixed factorial ANOVA with the same factors as in Experiment 1: Lexical status (word vs. nonword) x Temporal reference (past vs. future) x Key (left vs. right) x Counterbalance (not reported further).

## Results

Errors occurred in $5.19 \%$ of trials and were analyzed independently. Reaction times (RTs) exceeding 2 standard deviations from each participant's mean were excluded from the analysis, leading to the removal of an additional $9.56 \%$ of data.
The ANOVA on RTs reported two significant main effects. First, as expected, there was a main effect of Lexical status $(F(1,22)=64.84, M S e=4141.01, p<.001)$ : latencies for nonwords ( 883 ms ) were longer than for words ( 808 ms ) as in Experiment 1. Second, the factor Temporal reference $(F(1,22)=26.22, M S e=2464.2, p$ <.001) was also significant, indicating shorter latencies for past ( 828 ms ) than for future verbs ( 864 ms ). No other main effects or interactions were significant in the RTs analyses. Thus, the lexical decision task failed to replicate the Temporal reference x Key interaction obtained in Experiment $1(F<1$; Past verbs - left hand $M=831 \mathrm{~ms}$, Past verbs - right hand $M=824 \mathrm{~ms}$, Future verbs - left hand $M=866 \mathrm{~ms}$, Future verbs - right hand $M=862 \mathrm{~ms}$ ).
An omnibus ANOVA pooling together both experiments with the between-subjects factor Task (temporal vs. lexical) and the same within-subjects factors mentioned above revealed a significant three-way interaction between Task x Temporal reference x $\operatorname{Key}(F(1,46)=$ 12.59, $M S e=8342.15, p<.001$ ). This confirmed that the two tasks generated different patterns of results.

In the analyses of accuracy there was a main effect of Lexical status $(F(1,22)=10.28, M S e=2.85, p<.01)$, which indicated again that words were easier to process than non-words ( 1.5 vs. 2.3 mean errors, respectively).
No other main effect or interactions were significant.


Figure 2. Mean latencies for the factors Temporal reference and Key in Experiment 2.

## Discussion

The central observation of Experiment 2 was the absence of interaction between left-right responses and temporal reference. This null result occurred in the context of a very clear and sizeable interaction obtained in Experiment 1 using the same stimuli, procedures and participant population. Therefore, it seems that even when specially designed stimuli are used to make sure that temporal reference is processed, the emergence of a congruency effect between left-right space and time is strongly mediated by the context and the goal of the task: the effect can only be found when temporal processing is taskrelevant.

## General Discussion

The present study addressed the question of the automaticity of the activation of the left-right mental timeline. In line with prior findings (e.g., Torralbo et al, 2006; Ulrich \& Maienborn 2010; Ulrich et al, 2012) there is flexibility, not automaticity, in the activation of the mental time-line(s). Short, single words and nonwords especially designed to secure a deep processing generated a strong space-time congruency effect when participants judged their temporal reference, but failed to do it in a lexical decision task. This result agrees well with the conclusions obtained by Ulrich et al (2012) regarding the front-back time line with longer sentences in German, and corroborates those by Ulrich and Maienborn (2010) regarding the left-right time line without some of their potential confoundings.
Present results are also consistent with the view that, all other factors being equal, only the conceptual mappings that are required to carry out the task are set up in working memory (Santiago et al, 2011). It also agrees well with the flexibility observed in the literature on the automaticity of affordance effects (e.g., Borghi et al, 2012; Natraj et al, 2013).

Obviously, present results leave open many future lines of inquiry, and the issue of the automaticity (or lack thereof) of the activation of the mental time line is still not
closed. An important remaining question is whether it is possible to observe the activation of the mental time line in time-irrelevant tasks using different conditions. One interesting possibility has to do with the use of temporal stimuli which have a more direct link to temporal reference, such as dates, months or weekdays. Another possibility is that a more sensitive measure might be able to find the effects (e.g., mouse trajectories). Data are currently being collected about this latter possibility
If the activation of the mental time line remains taskdependent, then there raises the question of why. Other conceptual mappings on the spatial dimension have been shown to be activated automatically at least under certain conditions (e.g., evaluation with approach-avoidance responses, Chen \& Barg, 1999; or number magnitude, Dehaene et al, 1993). Space and time seem to be intrinsically linked from the initial stages in development (Piaget, 1969) and the influence of space on temporal judgments in psychophysics tasks remains until the adult age (Casasanto \& Boroditsky, 2008; Casasanto et al, 2010). Why then participants do not activate the spatial dimension automatically when processing linguistic stimuli with a temporal reference? Future research needs to address this question.
In conclusion, present results corroborate that the leftright space-time congruency effect is strongly mediated by the context and goal of the task, such that it only arises when the task explicitly requires judging temporal reference.

## Acknowledgments

This research has been funded by grants P09-SEJ-4772 (Consejería de Innovación, Ciencia y Empresa, Andalusian Government), PSI2012-32464 (Spanish Ministery of Economy and Competitivity).

## References

Barsalou, L. W. (1999). Perceptual symbol systems. Behavioral and Brain Sciences, 22, 577-660.
Barsalou, L. W. (2008). Grounded cognition. Annual Review of Psychology, 59(1), 617-645.
Borghi, A. M., Flumini, A., Natraj, N., \& Wheaton, L. A. (2012). One hand, two objects: Emergence of affordance in contexts. Brain and Cognition, $80(1)$, 6473.

Boroditsky, L. (2000). Metaphoric structuring: understanding time through spatial metaphors. Cognition, 75(1), 1-28.
Boroditsky, L. (2001). Does language shape thought? English and Mandarin speakers' conceptions of time. Cognitive Psychology, 43, 1-22.
Casasanto, D. J., \& Boroditsky, L. (2008). Time in the mind: using space to think about time. Cognition, 106, 579-93.
Casasanto, D., Fotakopoulou, O., \& Boroditsky, L. (2010). Space and time in the child's mind: Evidence for a cross-dimensional asymmetry. Cognitive Science, 34, 387-405.

Chen, S., \& Bargh, J. A. (1999). Consequences of automatic evaluation: Immediate behavior predispositions to approach or avoid the stimulus. Personality and Social Psychology Bulletin, 25, 215224.

Clark, H. H. (1973). Space, time, semantics, and the child. In: Moore, T. E. (Ed.). Cognitive development and the acquisition of language. New York: Academic Press.
Dehaene, S., Bossini, S., \& Giraux, P. (1993). The mental representation of parity and number magnitude. Journal of Experimental Psychology: General, 122, 371-396.
Gallese, V. (2008). Mirror neurons and the social nature of language: The neural exploitation hypothesis. Social Neuroscience, 3, 317-333.
Gevers, W., Reynvoet, B., \& Fias, W. (2004). The mental representation of ordinal sequences is spatially organized: evidence from days of the week. Cortex, 40, 171-172.
Gibbs, R. W. J. (2006). Embodiment and cognitive science. Cambridge: Cambridge University Press.
Glenberg, A.M., \& Gallese, V., (2011). Action-based language: A theory of language acquisition, comprehension, and production, Cortex, 48, 905-922.
Glenberg, A.M., Sato, M., Cattaneo, L., Riggio, L., Palumbo, D., \& Buccino, G.. (2008) Processing abstract language modulates motor system activity. Quarterly Journal of Experimental Psychology, 61(6), 905-919.
Harnad, S. (1990). The symbol grounding problem. Physica D, 42, 335-346.
Kinoshita, S., Lupker, S.J., \& Rastle, K. (2004). Modulation of regularity and lexicality effects in reading aloud. Memory and Cognition, 32, 1255-1264.
Lakens, D., Semin, G. R., \& Foroni, F. (2012). But for the bad, there would not be good: Grounding valence in brightness through shared relational structures. Journal of Experimental Psychology: General, 141(3), 584-594.
Lakoff, G. \& Johnson, M. (1980). Metaphors we live by. Chicago, IL: The University of Chicago Press.
Lakoff, G. \& Johnson, M. (1999). Philosophy in the flesh: The embodied mind and its challenge to western thought. New York, NY: Basic Books.
Natraj, N., Poole, V., Mizelle, J.C., Flumini, A., Borghi, A.M., \& Wheaton, L. (2013). Context and Hand Posture Modulate the Neural Dynamics of Tool-Object Perception. Neuropsychologia, 51, 506-519.
Núñez, R. E., \& Sweetser, E. (2006). With the future behind them: Convergent evidence from Aymara language and gesture in the cross-linguistic comparison of spatial construals of time. Cognitive Science, 30, 149.

Ouellet, M., Santiago, J., Israeli, Z., \& Gabay, S. (2010b). Is the future the right time? Experimental Psychology, 57, 308-314.
Pagliuca, G., Arduino, L.S., Barca, L., \& Burani, C. (2008). Fully transparent orthography, yet lexical reading aloud: The lexicality effect in Italian. Language and Cognitive Processes, 23, 422-433.
Piaget, J. (1969). The child's conception of time. London: Routledge \& Keagan Paul.

Previtali, P., de Hevia, M.D., \& Girelli, L. (2010). Placing order in space: the SNARC effect in serial learning. Experimental Brain Research, 201, 599-605.
Radden, G. (2004). The metaphor TIME AS SPACE across languages. In: Baumgarten, N., Böttger, C., Motz, M., Probst, J. (Eds.), Uebersetzen, interkulturelle Kommunikation, Spracherwerb und Sprachvermittlung das Leben mit mehreren sprachen: Festschrift fuer Juliane House zum 60. Geburtstag. Zeitschrift für Interkulturellen Fremdsprachenunterricht [Online], 8(2/3), 226-239.
Santiago, J., Lupiáñez, J., Pérez, E., \& Funes, M. J. (2007). Time (also) flies from left to right. Psychonomic Bulletin \& Review, 14, 512-516.
Santiago, J., Ouellet, M., Román, A., \& Valenzuela, J. (2012). Attentional factors in conceptual congruency. Cognitive Science, 36, 1051-1077.
Santiago, J., Román, A., \& Ouellet, M. (2011). Flexible foundations of abstract thought: A review and a theory. In A. Maass \& T. W. Schubert (Eds.), Spatial dimensions of social thought. Berlin: Mouton de Gruyter.
Santiago, J., Román, A., Ouellet, M., Rodríguez, N., \& Pérez-Azor, P. (2008). In hindsight, life flows from left to right. Psychological Research, 74, 59-70.

Sell, A. J., \& Kaschak, M. P. (2011). Processing time shifts affects the execution of motor responses. Brain \& Language, 117, 39-44.
Talmy, L. (2000). Toward a Cognitive Semantics. Volume I: Concept Structuring Systems. Cambridge, MA: MIT Press.
Torralbo, A., Santiago, J., \& Lupiáñez, J. (2006). Flexible conceptual projection of time onto spatial frames of reference. Cognitive Science, 30, 745-757.
Ulrich, R., Eikmeier, V., Vega, I., Ruiz Fernández, S., Alex-Ruf, S., \& Maienborn, C. (2012). With the past behind and the future ahead: Back-to-front representation of past and future sentences. Memory \& Cognition, 40(3), 483-495.
Ulrich, R., \& Maienborn, C. (2010). Left-right coding of past and future in language: The mental timeline during sentence processing. Cognition, 117, 126-138.
Wentura, D., Rothermund, K., \& Bak, P. (2000). Automatic vigilance: The attention-grabbing power of approach- and avoidance-related social information. Journal of Personality and Social Psychology, 78, 1024-1037.

# Surprise! You've Got Some Explaining to Do... 

Meadhbh I. Foster (meadhbh.foster@ucdconnect.ie) Mark T. Keane (mark.keane@ucd.ie)<br>Department of Computer Science \& Informatics, University College Dublin Belfield, Dublin 4, Ireland


#### Abstract

Why are some events more surprising than others? We propose that events that are more difficult to explain are those that are more surprising. The two experiments reported here test the impact of different event outcomes (Outcome-Type) and task demands (Task) on ratings of surprise for simple story scenarios. For the Outcome-Type variable, participants saw outcomes that were either known or less-known surprising outcomes for each scenario. For the Task variable, participants either answered comprehension questions or provided an explanation of the outcome. Outcome-Type reliably affected surprise judgments; known outcomes were rated as less surprising than less-known outcomes. Task also reliably affected surprise judgments; when people provided an explanation it lowered surprise judgments relative to simply answering comprehension questions. Both experiments thus provide evidence on this less-explored explanation aspect of surprise, specifically showing that ease of explanation is a key factor in determining the level of surprise experienced.


Keywords: Surprise; explanation; comprehension, coherence

## Introduction

Life is full of surprises, from bumping into a friend from home while on holidays, to arriving at a surprise party, to opening an amazing birthday gift, or hitting paydirt on that 100-1 racehorse. Surprise has been researched since Darwin's time, perhaps because it involves an interesting mixture of emotion and cognition. Originally, it was conceived of as a "basic emotion" (see Darwin, 1872; Ekman \& Friesan, 1971; Izard, 1977; Plutchik, 1991; Tomkins, 1962), though more recently it has been reappraised as a cognitive state (Kahneman \& Miller, 1986; Maguire, Maguire \& Keane, 2011) because, unlike most emotions, it can either be positively or negatively valenced (Ortony \& Turner, 1990). Although surprise clearly involves an emotional reaction (often accompanied by a startle response) it may also serve a strategic, cognitive goal, as it directs attention to explaining why the surprising event occurred and to learning for the future (see e.g., Maguire et al., 2011; Ranganath \& Rainer, 2003). Accordingly, in Artificial Intelligence (AI), surprise is seen as a candidate mechanism for identifying learning events in agent architectures (Bae \& Young, 2008, 2009; Macedo \& Cardoso, 2001; Macedo, Reisenzein \& Cardoso, 2004).

Imagine that you walk into your house and the walls have changed color from the color they were this morning. If you have no explanation for this turn of events then you would
probably be surprised by this outcome ${ }^{1}$. Many outcomes are surprising, the question is why? Our answer is that outcomes are surprising when they are hard to explain. Specifically, that surprise is a meta-cognitive sense of the amount of explanatory, mental work that was carried out to establish coherence between unfolding events in the world.

To illustrate the point, consider different scenarios for the "re-decoration surprising outcome". If I had left a team of decorators in my house that morning, I would clearly be less surprised by my walls being re-painted, because I had planned for that to occur. If no decorators were contracted, then I would be really surprised at this outcome, because no obvious explanation is forthcoming. However, if my wife and friends have been smirking at me for weeks (the way they do when they throw surprise parties) I would be less surprised because I can explain it as a prank. The experience of surprise will gradually increase across these scenarios as they move from being thoroughly-explainable (contracted decorators) to potentially explainable (smirking friends) to thoroughly-unexplainable (no decorators or smirking) because people have to carry out more explanatory, mental work to establish the coherence of these unfolding events.

In theories of surprise, one group of theorists have focussed on the properties of surprising outcomes, characterising them as low-probability events, disconfirmed expectations or schema-discrepant events (e.g., Meyer, Reisenzein \& Schützwohl, 1997; Reisenzein \& Studtmann, 2007; Schützwohl \& Reisenzein, 1999). Another group of theorists have stressed the importance of (often retrospective) sense-making and the integration of the surprising outcome to make it cohere with previous events (Kahneman \& Miller, 1986; Maguire \& Keane, 2006; Maguire et al., 2011). Theoretically, we are more aligned with the latter than with the former group; the main novelty in our approach being its emphasis on the meta-cognitive, explanatory aspects of the sense-making process. Adopting this meta-cognitive, explanatory approach suggests that experienced surprise may differ (a) for different classes of surprising outcomes (i.e., known versus less-known outcomes) and (b) under different task demands (i.e., being explicitly asked to explain a surprising outcome or not).

[^265]
## Classes of Surprising Outcomes

Viewing surprise from an explanation-perspective, suggests that outcomes may vary in their surprisingness because some are more well-known (directly or vicariously) than others. Intuitively, losing your wallet and losing your belt (that you put on your jeans this morning) are outcomes that could both surprise you during your day. We could call "losing your wallet" a known surprising outcome as it is an experience that people often discuss with one another, suggesting that most people have several "ready-made" explanations for it (see also Schank, 1986); that I left it in a shop, that I dropped it or that I was pickpocketed. In contrast, "losing your belt" is a less-known surprising outcome, suggesting perhaps that there are few or no "ready-made" explanations for $\mathrm{it}^{2}$. We predict that differences in the explanation spaces for these different classes of outcomes will result in different amounts of mental work to make them coherent and, thus, result in different levels of experienced surprise. Traditional probabilistic accounts would recast this known/less-known dimension as some variation of subjective probability, making parallel predictions about levels of surprise. However, obviously, we do not think that subjective probability is the key predictor of behaviour; indeed, in related work where it has been explicitly assessed, it has been shown not to accurately predict levels of surprise (see Maguire et al., 2011, Experiment 1).

## Explanation Task

Viewing surprise as a meta-cognitive effect suggests that if we ask people to explicitly explain the surprising outcome, they will be less surprised than if they receive task demands that are less directed toward explanation (e.g., comprehension questions about the scenario). If people are in "explanation mode" then clearly they should expend less mental effort in explaining the surprising event and hence, other things being equal, should experience less surprise relative to being in some "non-explanation mode". Should such explanation-effects occur, they can probably be explained in some ad hoc fashion by probabilistic accounts; however, we cannot see how a probabilistic account would lead one to perform such a test.

## Experiment 1

To test these predictions, we asked people to make surprise ratings about the outcomes of simple story scenarios describing everyday events. Some outcomes were known surprising outcomes, others were less-known surprising outcomes (see operational definitions in Materials). The task demands were varied by asking participants to either produce the answer to two short comprehension questions about that story or to produce an explanation for why that outcome may have occurred. So, the experiment involved a $2 \times 2$ design with Task (explanation vs. comprehension) as a

[^266]between-subjects variable and Outcome-Type (known vs. less-known) as a within-subjects variable. The questions asked for the comprehension task were very simple, using information clearly and unambiguously presented in the text given to participants (e.g., "Where is [character's name]?").

First, it was predicted that scenarios involving known surprising outcomes would be rated as less surprising than those with the less-known surprising outcomes; as explanations (or partial explanations) for the former would be available for use in making the outcome cohere with the rest of the scenario. Second, it was also predicted that the task demand to find an explanation would result in lower surprise ratings for outcomes, relative to the task demand of answering comprehension questions on the same stories. We made no specific predictions about whether these two variables would interact.

## Method

Participants and Design Forty UCD students (12 male, 28 female) with a mean age of 21.2 years ( $S D=2.07$, range $=$ 19-29) took part voluntarily in this study. Informed consent was obtained prior to the experiment. Participants were randomly assigned to one of two conditions in a 2 (betweensubjects; Task: explanation versus comprehension) x 2 (within-subjects; Outcome-Type: known versus less-known) mixed-measures design.

Materials A material set was created consisting of simple story scenarios with outcomes that were designed to involve known or less-known surprising outcomes (see Table 1). The type of outcome was operationally defined using (a) a pre-test sorting task by an independent group of raters and (b) Latent Semantic Analysis (LSA) scores of coherence.

For the sorting task definition, 20 story scenarios were presented in a pre-test to independent raters $(N=10)$. The raters were assigned to two groups: one group received half the scenarios with a known surprising outcome and the other half of the scenarios with a less-known surprising outcome, and the second group received the opposite. Each rater saw only one outcome for a given scenario. They were asked to determine if a given scenario has an outcome that "falls within the range of reasonable outcomes to the scenario" (i.e., known surprising outcome) or whether it "falls less within the range of reasonable outcomes to the scenario" (i.e., less-known surprising outcome). Of the 20 stories, the raters consistently deemed 9 stories to have separable known and less-known surprising outcomes (Fleiss' kappa showed substantial agreement, $\kappa=.68$, Landis \& Koch, 1977).

For the coherence-score definition, the known and lessknown variants of these 9 stories were scored using LSA. In discourse research (cf., Graesser \& McNamara, 2011), the explanatory coherence of texts is often operationalized by using latent semantic analysis (LSA) scores, where higher LSA scores indicate that the one text is more coherent than another (Landauer \& Dumas, 1996, 1997). For the selected 9 stories used in the experiment, the scenarios with the
known outcomes were scored higher ( $M=.62, S D=.2$ ) than their matched counterparts with less-known outcomes ( $M=$ $.53, S D=.21$ ), a difference that was statistically reliable, $F(1,8)=9.47, p=.015, \eta_{\mathrm{p}}^{2}=.54$.

Four material sets were created. Each of these comprised all nine scenarios, with either four scenarios with known surprising outcomes and five with less-known surprising outcomes, or five scenarios with known surprising outcomes and four with less-known surprising outcomes. As expected, the four material sets used proved to have no effect on subsequent surprise judgments, so these results are not reported in the following analyses ( $p>.12$ )

The order of presentation of these stories was randomised for each participant. Stories were presented on separate pages of a booklet, which began with the appropriate task instructions (explanation or comprehension). Each story was presented on a separate page with the scenario setting on the top of the page, followed by the outcome (known/lessknown), the statement of the task (comprehension or explanation) and a 7 -point scale on which to rate the suprisingness of the outcome (1: not surprised to 7 : very surprised).

Procedure and Scoring Participants were asked to read nine stories and to judge the surprisingness of their outcomes (see Table 1). For the Task variable, the participants in the explanation condition were asked to produce the first explanation they could think of for why the outcome may have occurred, before rating it for surprise; in the comprehension condition the participants were asked to answer two simple comprehension questions about the scenario, before rating it for surprise. For each story, the first question was about the story setting, and the second question was about the outcome.

Table 1: Sample scenario used in Experiment 1.

| Setting | Rebecca is on the beach. <br> She goes for a swim in the water. |  |
| :--- | :---: | :---: |
| Outcome | Known <br> After she dries <br> herself off she <br> notices that her skin <br> has turned red. | Less-known <br> After she dries <br> herself off she <br> notices that her skin <br> has turned turquoise. |

Prior to the experiment, we conducted a pre-test $(N=4)$ to verify that there was no significant difference in the average time taken to produce an explanation compared to that taken to answer the two short comprehension questions; time taken to do one task or the other were not reliably different $(t(2)=-1.41, p=.29$, explanation $M=6.5$ minutes; comprehension $M=7.5$ minutes). Two measures were recorded: (a) the 7 -point scale rating of surprise, and (b) the explanations produced by participants for each scenario in the explanation group. Finally, prior to data analysis one
participant ( $2.5 \%$ of the data) was discarded because they failed to follow the instructions given.

## Results and Discussion

Overall, the results confirmed the predictions that OutcomeType and Task both impact people's perceptions of surprise. The intuition that known outcomes are less surprising than less-known outcomes was confirmed, as was the prediction that instructions to explain the outcome would reduce the overall perception of surprise. So, for example, though both outcomes were deemed to be surprising, the lost-wallet type of scenario was found to be less surprising than the lost-belt type of scenario. No reliable interaction was found between the two variables.

Surprise Judgments A two-way ANOVA confirmed that participants judged stories with known outcomes ( $M=3.92$, $S D=1.18$ ) to be less surprising than those with less-known outcomes ( $M=5.73, S D=0.95$ ), $F(1,37)=128.82, p<$ $.001, \eta_{p}^{2}=.78$, see Figure 1. We maintain that this Outcome-Type effect occurs because known outcomes have associated "ready-made" explanations that are recruited quickly and easily to explain the outcome, lowering surprise ratings. In contrast, stories with less-known outcomes have few "ready-made" explanations to be recruited, so the outcome is harder to explain, resulting in relatively higher surprise ratings.

There was also a significant main effect of Task, $F(1,37)$ $=10.18, p=.003, \eta_{\mathrm{p}}{ }^{2}=.22$, indicating that the explanation group judged the outcomes to be less surprising ( $M=4.40$, $S D=1.03$ ) than the comprehension group ( $M=5.27, S D=$ 0.62 ). This effect occurs because in 'explanation mode' participants find explanations more easily and, hence, for meta-cognitive reasons, their perception of surprise decreases. No interaction between the two variables was found, $F(1,37)=0.00, p=.98, \eta_{\mathrm{p}}^{2}<.001$.


Figure 1: Mean surprise ratings for both levels of Outcome-
Type (known vs. less-known) and Task (explanation vs. comprehension) in Experiment 1

Explanations The explanations provided by the participants in the explanation group provide a key piece of converging
evidence for the view that known outcomes differ from lessknown outcomes. Participants' explanations for each scenario were recorded and classified to identify the most common or dominant explanation for a given scenario. We then carried out a by-materials analysis of the scenarios using the frequency of this dominant explanation as the dependent measure. The ANOVA revealed a main effect of Outcome-Type, in which dominant explanations were found to be more frequently produced to known outcomes ( $M=$ $5.44, S D=1.59)$ than less-known outcomes $(M=4, S D=$ 1.32), $F(1,8)=6.76, p=.03, \eta_{\mathrm{p}}^{2}=.46$. So, participants agree more about the explanations for known outcomes than they do for less-known outcomes, showing that the explanation spaces for these classes of outcomes differ.

## Experiment 2

Our second experiment attempted to replicate the effects found for Outcome-Type and Task, while adding a manipulation to the setting (Setting-Type) designed to elicit counterfactuals, to test another potential aspect of surprise.

Kahneman \& Tversky (1982; Kahneman \& Miller, 1986) proposed that "abnormal events" (our "surprising outcomes") will seem more abnormal if contrasting counterfactual alternatives are highly available; that is, the abnormal event (i.e., losing your wallet) will appear more abnormal if the contrasting counterfactual (i.e., the normal event of "having your wallet") is highly available. Kahneman \& Miller also propose that the availability of the normal event (the counterfactual) can provide an explanation for the abnormal event (the factual one), as people often use the difference between the two events to find an explanation. So, in theory, the elicitation of such counterfactuals could reduce the perceived surprise of an outcome, as it could provide a "quick and dirty" explanation of the surprising outcome. However, this prediction assumes that the counterfactual-inspired explanation is always used (which may not be a given). The literature on counterfactuals (Byrne, 2002; Kahneman \& Tversky, 1982) shows that they tend to be elicited when scenarios mention non-routine events (e.g., if you are told Jack had a car crash when he did not take his usual route home, people naturally draw on the counterfactual scenario of Jack taking his usual route home to find an explanation), though this is not always the case (e.g., Dixon \& Byrne, 2011). So, in this experiment, in addition to the original settings used in Experiment 1 (none), to elicit counterfactuals we changed the setting in the scenarios to stress that the event was either routine (usual) or non-routine (exceptional; see Table 2) for the actor involved.

So, the final design for this experiment manipulated Task (comprehension versus explanation), Outcome-Type (known versus less-known) and Setting-Type (none, usual or exceptional).

## Method

Participants and design Sixty UCD students (27 male, 33 female) with a mean age of 20.95 years ( $S D=4.228$, range
$=18-44)$ took part voluntarily in this study. Informed consent was obtained prior to the experiment. Participants were randomly assigned to one of two conditions in a 2 (between-subjects; Task: comprehension versus explanation) x 2 (within-subjects; Outcome-type: known versus less-known) x 3 (within-subjects; Setting-Type: none, usual, exceptional) mixed-measures design.

Procedure and Scoring As in Experiment 1, participants were asked to read nine stories and to judge the surprisingness of their outcomes. Rather than asking participants how surprised they would be "if this event occurred" (as they were in Experiment 1), they were asked to judge how surprised they would be by the event "if they were the character described". For the Setting-Type variable, the events in the story setting (a) gave no hint as to whether they were routine or not (none), (b) were said to be regular or routine (usual), or (c) said to be non-usual or nonroutine (exceptional). For the Outcome-Type variable, the participants saw either a known or less-known surprising outcome for each story; only one outcome and one setting was seen by each participant for each story (see Table 2 for an example of the materials used). The LSA scores for the three variants of the setting, none, usual and exceptional showed no main effect of this Setting-Type variable ( $p>$ .59).

Table 2: Sample scenario used in Experiment 2

|  | None | Usual | Exceptional |
| :--- | :---: | :---: | :---: |
| Sentence 1 | Lorna is in <br> an ethnic <br> restaurant. | Lorna is in her <br> favourite ethnic <br> restaurant that <br> she has often <br> gone to before. | Lorna is in a <br> new ethnic <br> restaurant that <br> she has never <br> gone to before. |
| Sentence 2 | She has ordered her food and, after a while, the |  |  |
| waiter brings it to her. |  |  |  |

Six material sets were created. Each of these comprised all 9 scenarios, with three variants of each setting type (none, usual, exceptional). Of these, either four scenarios were presented with known surprising outcomes and five with less-known surprising outcomes, or five scenarios with known surprising outcomes and four with less-known surprising outcomes. As expected, the six material sets had no effect on subsequent surprise judgments, so were not included as a variable in the reported analyses ( $p>.5$ ).

The order of presentation of these stories was randomised anew for each participant. Stories were presented sentence by sentence on a desktop computer-screen as participants pressed the spacebar, with each sentence appearing below the preceding one on the screen, until the outcome was presented. At this point, the participants in the explanation condition were instructed to "type in the first explanation you can think of for why this outcome may have occurred:"
and the participants in the comprehension condition saw and answered sequentially two simple comprehension questions about the story. One of these questions was about the information provided in the setting, and the other was about information provided in the outcome. Neither of these questions drew the participants' attention to the SettingType variable, per se. Initially, the participants in this condition saw the first question and, after providing an answer, they pressed the return key, this first question disappeared and the second question appeared. After the explanation/comprehension step, all participants pressed the return key and the question "If you were [character's name], how surprised would you be by this outcome?" On presentation of this question, participants indicated on a 7point scale their surprise judgment (1: not surprised, to 7: very surprised). Three measures were recorded: (a) the 7point rating of surprise, (b) the response time from the time of seeing the outcome sentence to the time in which the surprise judgment was made ${ }^{3}$, and (c) the explanations produced by each participant for each scenario. Finally, prior to data analysis, four participants ( $6.7 \%$ of the data) was discarded for failing to follow the instructions given.

## Results and Discussion

Overall, the results confirmed the predictions that known surprising outcomes and the adoption of an "explanationmode" decreased the perception of surprise; however, there was no strong evidence for a counterfactual effect.

Surprise Judgments A three-way ANOVA confirmed that participants judged known outcomes to be less surprising ( $M=4.51, S D=1.11$ ) than less-known outcomes $(M=6.21$, $S D=.75)$, showing a main effect of Outcome-Type, $F(1,54)$ $=92.46, p<.001, \eta_{\mathrm{p}}^{2}=.63$. There was also a significant main effect of Task, $F(1,54)=4.64, p=.036, \eta_{\mathrm{p}}^{2}=.08$. indicating that participants judged the outcomes of scenarios to be more surprising when they had answering comprehension questions, $(M=5.56, S D=.63)$ as opposed to providing explanations for them $(M=5.09, S D=.85$; see Figure 2). However, there was no main effect of SettingType, $F(2,108)=.002, p=.998, \eta_{\mathrm{p}}^{2}<.001$, no interaction between Outcome-Type and Setting-Type, $F(2,108)=2.78$, $p=.07, \eta_{\mathrm{p}}^{2}=.05$, and no reliable 2-way interactions between the variables (all $F \mathrm{~s}<1$ ).

Explanations Again the frequency with which the most dominant explanation was chosen by the explanation group was calculated for each scenario. A two-way, by-materials ANOVA showed a main effect of Outcome-Type, in which participants were more likely to produce the same dominant explanation for a known surprising outcome ( $M=7.89$, $S D$ $=3.26$ ) than for a less-known outcome ( $M=5.22, S D=$ 2.63), $F(1,8)=6.09, p=.039, \eta_{\mathrm{p}}^{2}=.43$. So, again, participants seem to have a greater degree of shared knowledge in the explanation of known outcomes than they

[^267]do for less-known outcomes, showing that the explanation spaces for these classes of outcomes differ.


Figure 2: Mean surprise ratings for both levels of OutcomeType (known vs. less-known) and Task (explanation vs. comprehension) in Experiment 2

## General Discussion

Overall, the experiments showed that known surprising outcomes are perceived as less surprising than less-known outcomes for the same scenarios, presumably because they are easier to explain. The task of explaining itself was also found to significantly reduce surprise ratings relative to answering comprehension questions in both experiments, again demonstrating how explanation may be the key factor in determining the level of surprise experienced. Finally, the explanations produced by participants were found to be more homogeneous for known outcomes than for lessknown outcomes; that is, there seems to be a shared dominant explanation used to explain known outcomes, that is less present in the case of less-known outcomes. We believe that these results provide converging evidence for an explanation-based account of surprise. Indeed, taken together, the combined effects on surprise found here strongly suggests that surprise may be a metacognitive effect (see Müller \& Stahlberg, 2007; Sanna \& Lundberg, 2012; Touroutoglou \& Efklides, 2010), with perceived surprise reflecting the ease or difficulty of explaining the surprising event.

However, little evidence was found for the counterfactual effect tested for in Experiment 2 (see the Setting-type variable). Both Kahneman \& Miller's Norm Theory (1986) and Teigen \& Keren's Contrast Hypothesis (2003) seem to predict that the ready availability of counterfactuals may influence the degree of surprise experienced; norm theory proposes that counterfactuals are used to explain why the event occurred, while the contrast hypothesis proposes that what was expected to occur (the events these counterfactuals elicit) is contrasted with the outcome to determine the level of surprise. There are several possible reasons for this prediction failure; it could be that our manipulation was not notable enough to elicit counterfactuals (though prior research would suggest
otherwise), or it could be that counterfactuals were generated but not used for explanation, or not considered as good-enough explanations. Of course, it could also be the case that the prediction is just wrong.

The current work also has implications for AI approaches to agent architectures, where it has been proposed that surprise might be used to identify learning events (e.g, Macedo \& Cardoso, 2001; Macedo, Reisenzein \& Cardoso, 2004). This proposal looks like it could be useful, once it is tempered by some consideration of the degree of surprise entailed and the ease of producing an explanation. The current work suggests that both of these aspects of the surprise process can differ considerably and, as such, would deliver very different learning outcomes for an agent.

## References

Bae, B.C., \& Young, R. M. (2008) A use of flashback and foreshadowing for surprise arousal in narrative using a plan-based approach. Lecture Notes in Computer Science, 5334, 156-167.
Bae, B. C., \& Young, R. M. (2009). Suspense? Surprise! or how to generate stories with surprise endings by exploiting the disparity of knowledge between a story's reader and its characters. In Iurgel, I. A., Zagalo, N. \& Petta, P. (Eds.) Proceedings of ICIDS, LNCS 5915. Berlin: Springer.
Byrne, R. M. J. (2002). Mental models and counterfactual thinking. Trends in Cognitive Sciences, 6, 405-445.
Darwin, C. R. (1872). The expression of the emotions in man and animals. London: John Murray.
Dixon, J., \& Byrne, R. M. J. (2011). Counterfactual thinking about exceptional actions. Memory and Cognition, 39(7), 1317-1331.
Ekman, P., \& Friesen, W. (1971). Constants across cultures in the face and emotion. Journal of Personality and Social Psychology, 17(2), 124-9.
Graesser, A. C., \& McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. Topics in Cognitive Science, 3, 371-398.
Izard, C. (1977). Human Emotions. New York: Plenum Press.
Kahneman, D., \& Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. Psychological Review, 93(2), 136-153.
Kahneman, D., \& Tversky, A. (1982). The simulation heuristic. In D. Kahneman, P. Slovic, \& A. Tversky (Eds.), Judgment under uncertainty: Heuristics and biases. New York: Cambridge University Press
Landis, J.R., \& Koch, G.G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159-174.
Landauer, T. K., \& Dumais, S. T. (1996). How come you know so much? From practical problem to theory. In D. Hermann, C. McEvoy, M. Johnson, \& P. Hertel (Eds.), Basic and applied memory: Memory in context. Mahwah, NJ: Erlbaum.

Landauer, T. K., \& Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. Psychological Review, 104, 211-240.
Macedo, L., \& Cardoso, A. (2001). Creativity and surprise. In Proceedings of the AISB'01 Symposium on Artificial Intelligence and Creativity in Arts and Science. United Kingdom: University of York.
Macedo, L., Reisenzein, R. \& Cardoso, A. (2004) Modeling forms of surprise in artificial agents: Empirical and theoretical study of surprise functions. In K. Forbus, D. Gentner \& T. Regier (Eds.), 26th Annual Conference of the Cognitive Science Society. Mahwah, NJ: Erlbaum.
Maguire, R., \& Keane, M. T. (2006). Surprise:
Disconfirmed expectations or representation-fit?
Proceedings of the 28th Annual Conference of the Cognitive Science Society, Hillsdale, NJ: Erlbaum.
Maguire, R., Maguire, P., \& Keane, M. T. (2011). Making sense of surprise: An investigation of the factors influencing surprise judgments. Journal of Experimental Psychology: LMC, 37(1), 176-186.
Meyer, W. U., Reisenzein, R., \& Schützwohl, A. (1997). Toward a process analysis of emotions: The case of surprise. Motivation and Emotion, 21, 251-274.
Müller, P. A., \& Stahlberg, D. (2007). The role of surprise in hindsight bias: A metacognitive model of reduced and reversed hindsight bias. Social Cognition, 25(1), 165-184.
Ortony, A., \& Turner, T. J. (1990). What's basic about basic emotions? Psychological Review, 97(3), 315-331.
Plutchik, R. (1991). The Emotions. New York: University Press of America.
Ranganath, C., \& Rainer, G. (2003). Neural mechanisms for detecting and remembering novel events. Nature Reviews Neuroscience, 4, 193-202.
Reisenzein, R., \& Studtmann, M. (2007). On the expression and experience of surprise: no evidence for facial feedback, but evidence for a reverse self-inference effect. Emotion, 7(3), 601-611.
Sanna, L. J., \& Lundberg, K. B. (2012). The experience of thinking: Metacognitive ease, fluency and context. In P. Briñol \& K. G. DeMarree (Eds.) Social Metacognition, 179-198, UK: Psychology Press.
Schank, R. C. (1986). Explanation patterns. Hillsdale, NJ: Lawrence Erlbaum Associates.
Schützwohl, A., \& Reisenzein, R. (1999). Children's and adult's reactions to a schema-discrepant event: A developmental analysis of surprise. International Journal of Behavioral Development, 23 (1), 37-62.
Teigen, K. H., \& Keren, G. (2003). Surprises: low probabilities or high contrasts? Cognition, 87(2), 55-71.
Tomkins, S. S. (1962). Affect, imagery, consciousness. (Vol. 1). New York: Springer.

Touroutoglou, A., \& Efklides, A. (2010). Cognitive interruption as an object of metacognitive monitoring: Feeling of difficulty and surprise. In A. Efklides \& P. Misailidi (Eds.) Trends and prospects in meta-cognition research. New York: Springer.

# Frequent Frames, Flexible Frames and the Noun-Verb Asymmetry 

Daniel Freudenthal, Julian Pine<br>School of Psychology, University of Liverpool<br>Gary Jones<br>School of Psychology, Nottingham Trent University<br>Fernand Gobet<br>School of Psychology, University of Liverpool


#### Abstract

In this paper we compare several mechanisms for using distributional statistics to derive word class information. We contrast three different ways of computing statistics for independent left and right neighbours with the notion of a frequent frame. We also investigate the role of utterance boundaries as context items and weighting of frequency information in terms of the successful simulation of the noun-verb asymmetry. It is argued that independent contexts can classify items with a higher degree of accuracy than frequent frames, a finding that is more pronounced for larger input sets. Frequent frames classify a larger number of items, but do so with lower accuracy. Utterance boundaries are useful for the development of a noun category, particularly at intermediate levels of frequency sensitivity.


Keywords: Word class derivation, independent contexts, frequent frames.

## Introduction

Several authors have shown that distributional statistics can provide powerful cues for acquiring syntactic categories; words that belong to the same syntactic category tend to be preceded and followed by the same words. Thus, nouns tend to be preceded by determiners and adjectives and followed by verbs. Redington, Chater and Finch (1998), building on work by Chater and Finch (1992), investigate several variants of the same basic principle: for any of a set of target words, a context vector was derived that contained (rank orders of) counts of the 150 most frequent words in the corpus, in positions preceding and following the target words. Redington et al. computed correlations between the context vectors of the target words, which were then used as input to a Hierarchical Cluster Analysis, and concluded that the resulting classes mapped closely onto broad syntactic classes. Redington et al. explore a number of variants of the basic mechanism, but get their best results by using a context of one preceding and one following word, and using a rank order correlation as their distance measure.

An alternative mechanism for acquiring syntactic categories has been proposed by Mintz (2003). Mintz introduces the notion of a frequent 'frame': two lexical items with one word intervening (e.g. He X to). Mintz argues that the (45) most frequent frames in the (English) corpora he analyses show high internal consistency in
terms of the grammatical category of the items that occur in the central position. The notion of a frequent frame is therefore thought to provide a powerful cue that children might employ in the acquisition of syntactic categories. More recent work has confirmed the utility of frequent frames for French (Chemla et al. 2009), but results have been less promising for languages with relatively free word order such as Dutch (Erkelens, 2009) and German (Stumper et al. 2011).

A major difference between the approaches of Redington et al. and Mintz is that the approach described by Redington is inherently graded and frequency sensitive in nature. Thus, in this approach, co-occurrence statistics are collected across all uses of a particular word. Depending on the exact implementation, the approach can also show varying degrees of frequency sensitivity with context vectors containing (rank orders of) word counts. Similarity is then expressed as a correlation-like measure across context vectors, which can be interpreted as a probability of two items being of the same class. This graded context-sensitivity is absent from Mintz's approach. Thus, while Mintz's analysis is restricted to the 45 most frequent frames, it clusters together all items that co-occur in one of these frames. The approach therefore ignores many contexts in which a word may occur, and instead clusters items on the basis of (potentially one) occurrence in specific high frequency contexts.
Typically, mechanisms for extracting grammatical categories are evaluated in terms of accuracy (the extent to which items that are clustered together belong to the same syntactic category) and completeness (the extent to all items within one syntactic category are clustered together). St. Clair et al. (2010), as well as Monaghan (2004), compared frames and independent contexts as used by Redington et al. in the context of connectionist simulations, and found that frames were accurate but resulted in low completeness while independent contexts performed similarly in terms of accuracy but outperformed frames in terms of completeness.
However, while high accuracy is clearly a desirable property of a mechanism that derives syntactic categories (children for instance make very few word class errors), it is less clear if high completeness is desirable, particularly if one is interested in modeling children's early linguistic abilities. Thus, while children ultimately develop
linguistic abilities that suggest the presence of relatively abstract linguistic categories, their early multi-word speech has been characterized as lexically specific. Moreover, completeness is often measured across word classes, when there appear to be developmental discontinuities in children's productive use of grammatical categories. Thus, children appear to be more prepared to produce novel nouns than novel verbs in familiar contexts (Akhtar \& Tomasello, 1997), a finding which has led to the suggestion that children may develop a productive noun category earlier than a productive verb category (Tomasello, 2000).

On the basis of these considerations, it would seem that a mechanism that is plausibly employed by language learning children is one that favours accuracy over completeness and favours the linking of high numbers of nouns over the linking of high numbers of verbs. One factor that might impact on the relative likelihood of linking nouns and verbs is the weighting of frequency information. Nouns for instance have a relatively high likelihood of being preceded by one of a small set of determiners. A second factor that is likely to affect the relative linking of nouns and verbs is the availability of utterance boundaries as framing elements. Nouns have a relatively high likelihood of occurring in utterance final position, and the utterance boundary is thus a potentially powerful cue for a noun category. Freudenthal et al. (2008), in the context of connectionist simulations, provide some evidence in support of this suggestion.

The main aims of this paper then are as follows: 1. To compare frequent frames and measures similar to those used by Redington et al. (1998) in terms of their ability to simulate the word classes apparent in children's early speech, 2. To investigate how different levels of frequency sensitivity as well as the availability of utterance boundaries may impact on these mechanisms.

## Similarity measures used

In the current paper we compare 4 different measures of similarity. We consider the frequent frames approach described by Mintz (2003), as well as 3 different implementations of the independent contexts approach described by Redington et al. (1998). In line with Redington et al., we considered as target words (i.e. words to be classified) the 1000 most frequent words in the corpus. The frequent frames approach closely followed the implementation by Mintz: we considered the target words that co-occurred in the 45 most frequent lexical frames within a corpus. The implementations of independent contexts closely followed the implementation of Redington et al. The context for a given word was encoded as a vector of length 300 consisting of counts of the 150 most frequent words in the corpus in the position directly preceding and following the target word. The actual similarity measures based on these vectors were: 1. Spearman rank order correlation (as used by Redington et al); 2. Cosine similarity based on raw frequency counts;
and 3. Cosine similarity based on the square root of frequency counts. These three different measures differ with respect to the weighting of frequency information, which is highest for cosine similarity based on raw counts and lowest for rank orders. Weighting of frequency information is even lower for frequent frames: this measure only considers whether or not target words cooccur within a given frame, not how often they co-occur.

## Corpora

The analyses were performed on the child-directed speech of the 12 children in the Manchester corpus (Theakston et al., 2001). The child-directed speech in the Manchester corpus is typically in the range of 25,000 to 30,000 utterances per child. Corpora were cleaned up minimally, and only multi-word utterances were analysed. For all corpora the following statistics were collected: for every word in the corpus, counts were collected for the items that preceded and followed it, as well as the frames (A X B) that surrounded them. Frame counts were then tallied across words to determine overall counts for (frequent) frames.

One additional manipulation involved the merging of the corpora for the 12 individual children into one large corpus. This manipulation was included to determine whether the mechanisms under investigation are differentially affected by changes in corpus size.

For the purpose of determining accuracy of derived word pairings, words were assigned their respective Part of Speech (POS) tag as employed on the \%mor coding tier of the corpus. POS tags in the Manchester corpus are relatively detailed and distinguish between main verbs and auxiliaries, as well as nouns, pronouns and proper nouns. The categories employed here are therefore similar to what Mintz terms 'expanded' labelling. Where multiple POS tags were used for one word form (e.g. forms that can be used as either noun or verb) the most frequently used POS tag was assigned. That is: words forms were assigned to the grammatical category in which they were used most frequently.

## Results

There are several ways in which one can evaluate the success of different similarity measures. Redington et al. (1998) performed a cluster analysis, and plotted accuracy and completeness as a function of the similarity level (or number of clusters extracted). While this is informative, it does raise a number of problems in interpreting the outcome. First, it is not immediately obvious at what similarity level one should compare different mechanisms, and second, the clustering process itself can be performed in different ways which have the potential to influence the results in terms of accuracy and completeness. For these reasons we opted to sidestep the clustering process, and perform a more direct evaluation of the similarity scores. This was done by extracting all possible word pairs, and computing the relevant similarity

Table 1: Accuracy and number of classified words for the four different distance measures.

|  | Cos-Sim, Raw Freq |  | Cos-Sim. Sqrt Freq |  | Spearman rank order |  | Frequent Frames |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Acc. | N | Acc. | N | Acc. | N | Acc. | N |
| Anne | .90 | 1555 | .92 | 2367 | .91 | 1314 | .69 | 8688 |
| Aran | .95 | 596 | .94 | 2046 | .92 | 1772 | .66 | 23612 |
| Becky | .83 | 565 | .82 | 1718 | .88 | 1746 | .71 | 5423 |
| Carl | .88 | 576 | .92 | 2113 | .92 | 1894 | .70 | 6191 |
| Dominic | .81 | 409 | .84 | 1418 | .84 | 1313 | .62 | 7904 |
| Gail | .84 | 355 | .89 | 1223 | .91 | 1152 | .59 | 8381 |
| Joel | .76 | 329 | .86 | 1063 | .90 | 1089 | .60 | 7182 |
| John | .76 | 326 | .83 | 1938 | .85 | 2570 | .73 | 9269 |
| Liz | .78 | 340 | .83 | 1171 | .85 | 1546 | .66 | 5354 |
| Nicola | .78 | 356 | .89 | 1115 | .91 | 1272 | .62 | 9949 |
| Ruth | .92 | 276 | .91 | 910 | .89 | 1024 | .61 | 8089 |
| Warren | .79 | 403 | .84 | 1199 | .86 | 1222 | .70 | 14405 |
|  |  |  |  |  |  |  |  |  |
| Average | .84 | 507 | .87 | 1523 | .89 | 1492 | .66 | 9537 |
|  |  |  |  |  |  |  |  |  |
| Merged | .93 | 1559 | .95 | 9561 | .93 | 8636 | .55 | 76676 |

metrics for every word pair. Where a similarity metric exceeded a certain threshold the word pair was considered to belong to the same category.

This procedure obviously raises the question of what threshold should be chosen for the different similarity metrics. Generally, higher values for the threshold will result in higher levels of accuracy and lower numbers of classified items, but these numbers may differ across metrics for a specific value of the threshold. For this reason, we decided to choose a different value for the threshold across metrics such that the resulting accuracy was always relatively high ( $\sim 90 \%$ ) and comparable across the metrics ${ }^{1}$, thus allowing for a meaningful interpretation of differences in completeness. Table 1 shows percentage accuracy as well as number of classified items for the 12 individual children in the Manchester corpus, the average for these children as well as scores for the merged Manchester corpus.

The concept of a threshold for classification is irrelevant for frequent frames (thus making it impossible to peg accuracy at $90 \%$ ), as the notion of a frame entails that two items that co-occur in one of the frequent frames are of the same word class. Table 1 therefore lists accuracy and number of classified items for all word pairs that co-occurred in one of the 45 most frequent frames in

[^268]the relevant corpus. Word pairs that co-occurred in multiple frames were counted only once.

Looking at the individual children and their average in Table 1, it is obvious that there are substantial differences between the different metrics. Frequent frames classify a large number of pairs, but do so at relatively low accuracy. Accuracy levels for frequent frames are lower than reported by Mintz (who reports a type accuracy of .91). This lower accuracy is at least partly caused by the fact that, for the current analyses, words were assigned to their most common category. While such a procedure makes sense for graded measures that collate statistics over different contexts, it may be less appropriate for the frame style analysis. Thus several words can be used as either a noun or a verb (e.g. pull, paint). In the corpora employed here, pull is overwhelmingly used as a verb, while paint is used as a noun more often than a verb (and as a consequence, is considered to belong to the noun class for the current analyses). The frequent frames analysis will classify these items together (resulting in a false alarm) because they co-occur in the frame you $X$ your.

For this reason we performed a second, contextual accuracy analysis on the frames analysis: for every word pair that co-occurred in one of the frequent frames, we considered the actual category of the word within the (most frequent) frame. This analysis resulted in accuracy scores (. 76 on average) that were higher than in the standard analysis, but still lower than those attained by the probabilistic measures.

A comparison of the probabilistic measures also reveals differences. Spearman rank-order and square root cosine similarity classify a similar number of items at similar levels of accuracy. Raw cosine similarity on the other
hand only classifies around a third of the number of items that the other probabilistic measures classify.
Looking at the results for the merged corpus, it becomes apparent that all four measures classify a larger number of items. The three probabilistic measures however, do so with slightly higher accuracy than for the individual children, while the frequent frames measure shows a decrease in accuracy ( $66 \%$ vs. $55 \%$, and $76 \%$ vs. $69 \%$ for the contextual score). Thus, it appears that the merged corpus contains additional information that can be successfully employed by the probabilistic measures but not the frequent frames measure. The increased information in the merged corpus is actually detrimental to the accuracy score for frequent frames. This latter finding appears to be caused by the fact that the frequent frames approach is overly sensitive to the occurrence of 'stray' words within the frequent frames. The fact that a word needs to occur only once within a specific frame to be clustered with all other words within that frame means that infrequent words that are atypical of a particular frame can potentially exert undue influence on overall accuracy scores. This problem becomes more pronounced in larger corpora. Such effects are less of a problem for the probabilistic measures.

## The noun-verb asymmetry

It was argued earlier that children are more willing to use novel nouns in known contexts than they are to use verbs. This finding has been taken as evidence that children develop a productive noun category earlier than they develop a verb category. In this section, we examine to what extent the different metrics show a preference for the clustering of nouns and verbs. This was done by examining the 'hits' from the data in Table 1, and counting the number of noun-noun and verb-verb pairs. The resulting data (proportion of noun-noun pairs relative to noun-noun + verb-verb pairs) are displayed in Table 2.
It is clear from Table 2 that the measures that are most frequency sensitive cluster the highest proportion of nouns. Thus, cosine similarity based on raw frequencies clusters a relatively low number of items but these items consist almost exclusively of nouns. Square Root Cosine similarity and rank order correlation are equally productive in terms of the number of items they classify, with the more frequency sensitive Cosine Similarity linking more nouns. Frequent frames on the other hand overwhelmingly link verbs. It is also apparent from Table 2 that, for frames, there is considerable variation in the proportion of noun-noun pairings: Aran's proportion is highest at $55 \%$, but half the children show a proportion of noun-noun pairings under $5 \%$.

Table 2: Proportion of noun-noun pairings relative to noun-noun plus verb-verb pairings (total N in parentheses), excluding utterance boundaries.

| NV-ratio | Cos- <br> Raw | Cos-Sqrt | Spearman | Frames |
| :---: | :---: | :---: | :---: | :---: |
| Anne | .99 | .93 | .63 | .34 |
|  | $(1387)$ | $(2084)$ | $(1068)$ | $(5836)$ |
| Aran | .99 | .84 | .67 | .55 |
|  | $(560)$ | $(1784)$ | $(1440)$ | $(15222)$ |
| Becky | .99 | .91 | .66 | .03 |
|  | $(461)$ | $(1349)$ | $(1421)$ | $(3685)$ |
| Carl | .99 | .89 | .72 | .01 |
|  | $(507)$ | $(1885)$ | $(1652)$ | $(3964)$ |
| Dominic | .97 | .83 | .48 | .02 |
|  | $(316)$ | $(1040)$ | $(937)$ | $(4499)$ |
| Gail | .96 | .89 | .67 | .02 |
|  | $(284)$ | $(975)$ | $(888)$ | $(4452)$ |
| Joel | .99 | .85 | .63 | .02 |
|  | $(237)$ | $(831)$ | $(853)$ | $(3828)$ |
| John | .99 | .90 | .79 | .40 |
|  | $(246)$ | $(1566)$ | $(2104)$ | $(6378)$ |
| Liz | .97 | .91 | .80 | .10 |
|  | $(250)$ | $.894)$ | $(169)$ | $(3214)$ |
| Nicola | .96 | .74 | .56 | .16 |
|  | $(275)$ | $(907)$ | $(1049)$ | $(6056)$ |
| Ruth | .99 | 633 | .85 | .04 |
|  | $(244)$ | $(746)$ | $(790)$ | $(4645)$ |
| Warren | .98 | .76 | .47 | .52 |
|  | $(311)$ | $(904)$ | $(927)$ | $(9565)$ |
|  |  |  |  |  |
| Average | .98 | .87 | .66 | .28 |
|  | $(432)$ | $(1247)$ | $(1191)$ | $(5945)$ |
|  |  |  |  |  |
| Merged | .99 | .80 | .67 | .57 |
|  | $(1426)$ | $(8577)$ | $(7319)$ | $(63747)$ |

## The role of utterance boundaries

The analyses reported in Table 1 and 2 only considered 'lexical contexts'. That is, only words were considered as context items. The following set of analyses included the beginnings and ends of utterances as context items. Redington et al. (1998) do consider utterance boundaries as context items in one of their analyses (and conclude that they are potentially useful), but the non-parametric nature of their distance metric (rank order correlations) may underestimate the potential utility of utterance boundaries.
Mintz (2003) does not consider utterance boundaries, and it could be argued that there is little reason to consider them. Intuitively, the appealing feature of frames is that (because of their lexical nature) they are highly constraining and hence likely to result in relatively high accuracy. Allowing utterance boundaries in frames limits their constraining nature and may thus reduce accuracy levels. At the same time, however, frames including
utterance boundaries have the potential to capture large numbers of nouns (e.g. The X end\$) and thus might serve to counteract the verb bias apparent in Table 2. Table 3 shows the accuracy scores and number of word pairings for the analysis that include the utterance boundary as a framing element. The proportion of noun-noun pairings relative to noun-noun plus verb-verb pairings are shown in Table 4. For reasons of brevity, Tables 3 and 4 do not present data for the individual children in the Manchester corpus, but only the average and merged data across the 12 children.

Table 3: Proportion correct and number of word pairings including utterance boundaries.

|  | Cos-raw | Cos-Sqrt | Spearman | Frames |
| :---: | :---: | :---: | :---: | :---: |
| Average | .82 | .90 | .90 | .49 |
|  | $(862)$ | $(5739)$ | $(3163)$ | $(81351)$ |
|  |  |  |  |  |
| Merged | .85 | .91 | .92 | .28 |
|  | $(4098)$ | $(23271)$ | $(11271)$ | $(316651)$ |

It is evident from Table 3 that the probabilistic measures deal well with the addition of the utterance boundary as a framing element. Accuracy levels are comparable to those shown in Table 1, while the number of word pairings has increased by a factor 2 to 3 . Inspection of Table 4 furthermore indicates that the increase in completeness is largely the result of increased linking of noun pairs. This is particularly noticeable in the square root cosine similarity model, which links twice as many words as the rank order model. Thus, the average square root cosine similarity model links over 4700 noun-noun pairs, compared to $\sim 1900$ for the rank order model. This difference reflects the greater frequency sensitivity of the cosine model, and suggests that square root of raw frequency represents an optimum level of frequency sensitivity.

Table 4: Number of noun-noun pairings relative to noun-noun plus verb-verb pairings, including utterance boundaries.

| NV-ratio | Cos-raw | Cos-Sqrt | Spearman | Frames |
| :---: | :---: | :---: | :---: | :---: |
| Average | .99 | .95 | .77 | .83 |
|  | $(677)$ | $(4957)$ | $(2501)$ | $(38097)$ |
|  |  |  |  |  |
| Merged | .99 | .90 | .67 | .72 |
|  | $(3256)$ | $(19924)$ | $(9558)$ | $(82924)$ |

## Conclusions

The main aim of this paper was to compare a mechanism for extracting syntactic categories based on independent contexts with Mintz style frequent frames in terms of their accuracy and ability to cluster nouns and verbs. A secondary aim was to investigate the role of frequency sensitivity and availability of utterance boundaries as framing elements.

The analyses presented here suggest that independent contexts result in better predictions than frequent frames. Frequent frames classify a larger number of words, but do so with lower overall accuracy.

Apart from being more accurate, the mechanisms based on independent contexts also cluster more nouns than verbs. This appears to be consistent with the suggestion that children form a productive noun category earlier than they form a verb category. The reverse is true of frequent frames: across the corpora frequent frames overwhelmingly cluster verbs rather than nouns, with noun-noun pairings making up under $5 \%$ of pairings for half the corpora.

When considering large input sets (i.e. the merged Manchester corpus), it becomes obvious that the mechanisms employing independent contexts are able to utilize the additional information contained in larger data sets to classify a larger number of items with similar levels of accuracy. The frequent frames mechanism also classifies a larger number of items when employed on a larger data set, but does so with lower accuracy. This result suggests that one of the strengths of frequent frames - its ability to quickly categorize a relatively large number of items on the basis of limited data - becomes a weakness when faced with larger datasets. Thus, the fact that the approach does not consider the frequency with which items occur in the target frames, results in it being relatively brittle and sensitive to noise and infrequent items in the input.

As an illustration, consider the frame You $X$ to, which is the most frequent frame for the corpus of Carl as well as the merged Manchester corpus. Within Carl's corpus this frame contains 22 unique words, of which 20 (or 91\%) are verbs. Across the Manchester corpus, the same frame contains a total of 89 unique words of which 67 (or $75 \%$ ) are verbs. This increase in non-verbs has a marked impact on the accuracy for the frame which drops from . 74 to .49 . Many of the non-verbs occurring in the frame are legitimate (but infrequent) fillers for the frame (e.g. back, off, down, happy, ready, anything, something, just, how, over, not, one), while others are slightly more exotic: that (from we brought you that to help you), tomorrow (from who's taking you tomorrow to playgroup) to somewhat bizarre: card (from a thank you card to give you).

Since noisiness is an inherent property of the speech signal, which contains frequent repetitions, retracings and restarts, this finding suggests that frequent frames may not be a suitable source of information for category extraction unless combined with some sort of 'clean-up mechanism' or probabilistic element. Such an addition, however, would considerably weaken the great strength of this approach: its ability to rapidly classify items on the basis of little information.

All three mechanisms that computed statistics over independent contexts were able to attain higher accuracy levels, though they classified fewer items. It was argued that the probabilistic nature of these mechanisms allows
them to utilize the additional information in larger corpora without suffering from the brittleness associated with frequent frames.

A similar pattern emerges when including utterance boundaries as context items. For independent contexts, the inclusion of utterance boundaries results in comparable levels of accuracy, coupled with greater levels of completeness. The inclusion of utterance boundaries in frequent frames results in a drop in performance, in particular when considering the merged corpus. Such a finding may not be surprising (and indeed may not be in the 'spirit' of frequent frames), since the inclusion of utterance boundaries leads to the measure being less constrained than the lexical frequent frames that were originally proposed by Mintz. It does, however, provide an additional indication that independent contexts are less brittle and better able to incorporate additional, potentially noisy information.

The increased flexibility of independent contexts is further underscored by the analyses relating to the nounverb asymmetry. Empirical work has suggested that children are more likely to substitute novel nouns in familiar contexts than they are to substitute novel verbs. (Akhtar \& Tomasello, 1997; Tomasello, 2000). If the number of classified nouns versus verbs is an indication of such a tendency, frequent frames would appear to result in levels of verb-richness that are overly high. Thus, when excluding utterance boundaries, noun pairs make up on average only $28 \%$ of noun and verb pairs for frequent frames, compared to approximately $80 \%$ for independent contexts. While the inclusion of utterance boundaries leads to higher levels of noun pairs in the frequent frames approach, the results from Table 2 suggest that this is achieved at the expense of accuracy. For independent contexts, accuracy and noun richness remain high, whilst completeness is improved relative to the condition without utterance boundaries.

Some differences were also apparent within the different implementations of independent contexts. The present paper compared three different measures that differed primarily in terms of the weighting of frequency information. Within the constraints employed (which included a threshold value that results in an average accuracy level of approximately $90 \%$ ), it was apparent that a similarity measure based on raw frequency counts results in relatively low completeness, while completeness for square root cosine similarity and rank order correlations perform at similar levels of completeness, with the square root cosine similarity measure showing more of a noun advantage than the less frequency sensitive rank order correlation. Overall, the inclusion of utterance boundaries leads to higher levels of noun-richness, suggesting it is a useful source of information. The size of this effect of noun-richness however was dependent upon frequency sensitivity: while noun-richness increased for all probabilistic measures, this was most pronounced for the square root cosine
similarity, suggesting that this represents an optimum level of frequency sensitivity.

## Acknowledgements

This research was supported by the Economic and Social Research Council under grant number ES/J011436/1.

## References

Akhtar, N., \& Tomasello, M. (1997). Young children's productivity with word order and verb morphology. Developmental Psychology, 33, 952-965.
Chemla, E., Mintz, T. H., Bernal, S., \& Christophe, A. (2009). Categorizing words using 'frequent frames': What cross-linguistic analyses reveal about distributional acquisition strategies. Developmental Science, 12, 396-406.
Erkelens, M. A. (2009). Learning to categorize verbs and nouns. Unpublished PhD Thesis, Universiteit van Amsterdam, Amsterdam.
Finch, S. P. \& Chater, N. (1992). Bootstrapping syntactic categories. Proceedings of the 14th Annual Conference of the Cognitive Science Society (pp. 820-825).
Freudenthal, D., Pine, J. \& Gobet, F. (2008). On the utility of conjoint and compositional frames and utterance boundaries as predictors of word categories. In B. C. Love, K. McRae, \& V. M. Sloutsky (Eds.), Proceedings of the 30th Annual Conference of the Cognitive Science Society (pp. 1947-1952). Austin, TX: Cognitive Science Society.
MacWhinney, B. (2000). The CHILDES project: Tools for analysing talk ( $3^{\text {rd }}$ Edition). Mahwah, NJ: Erlbaum.
Monaghan, P. \& Christiansen, M. (2004). What information is useful and usable in language acquisition? Proceedings of the $26^{\text {th }}$ Annual Conference of the Cognitive Science Society. Mahwah, NJ: Lawrence Erlbaum.
Mintz, T. (2003). Frequent frames as a cue for grammatical categories in child directed speech. Cognition, 90, 91-117.
Redington, M., Chater, N. \& Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic structures.
St. Clair, M.C., Monaghan, P., \& Christiansen, M.H. (2010). Learning grammatical categories from distributional cues: Flexible frames for language acquisition. Cognition, 116, 341-360.
Stumper, B., Bannard, C., Lieven, E., \& Tomasello, M. (2011). Frequent frames in German child-directed speech: A limited cue to grammatical categories. Cognitive Science, 35, 1190-1205.
Theakston, A.L., Lieven, E.V.M., Pine, J.M. \& Rowland, C.F. (2001). The role of performance limitations in the acquisition of Verb-Argument structure: An alternative account. Journal of Child Language, 28, 127-152.
Tomasello, M. (2000). Do young children have adult syntactic competence? Cognition, 74, 209-253.

# Emotional Speech Processing and Language Knowledge 

Conor I. Frye (cifrye@cogsci.ucsd.edu)<br>Department of Cognitive Science, UC San Diego, 9500 Gilman Dr. M/S 0515<br>La Jolla, CA 92093 USA<br>Sarah C. Creel (creel@cogsci.ucsd.edu)<br>Department of Cognitive Science, UC San Diego, 9500 Gilman Dr. M/S 0515<br>La Jolla, CA 92093 USA


#### Abstract

How does language knowledge affect processing of paralinguistic information-vocal properties that are not directly related to understanding words? This study investigates links between a listener's native language, any other languages they may have experience in, and the ability to identify vocal emotional information in those languages. The study focuses on two particular classes of languages: those with lexical tone, such as Mandarin Chinese, which use pitch properties to distinguish otherwise-identical words; and those without lexical tone, such as English. English listeners and bilingual Mandarin-English listeners listened to sentences and categorized the emotional content of English and Mandarin sentences. Half of the sentences were presented normally; the other half were low-pass filtered to remove all but prosodic cues (pitch and timing). English listeners were better at identifying emotions in English sentences, while bilinguals were equally good at identifying emotions in both languages. This indicates better overall emotion recognition from prosody alone for listeners more familiar with a language. It may point to a connection between tone language experience and augmented paralinguistic processing capabilities.


Keywords: speech perception; paralinguistic perception; voice; language background; individual differences; bilingualism

## Introduction

Spoken language as a medium is not just a symbol system of discrete speech sounds; it is also replete with cues to the talker's identity, region of origin, and emotional state. Although much research has been devoted to understanding how exposure to a language affects speech sound identification (Kuhl, 1994), almost no one has asked how language knowledge affects processing of paralinguistic information-vocal properties that are not directly related to understanding words like speech rate and pitch changes (see Thompson \& Ballkwill, 2006, for an exception). Emotion in the voice is thought to be conveyed by these paralinguistic cues. Though differing languages seem to use similar vocal acoustic cues for the "basic emotions" in non-speech vocalizations such as laughter and crying, it is not clear how readily listeners perceive these emotional cues crosslinguistically when only presented with the auditory signal (Sauter et al., 2010).

## Language-specific recognition of vocal affect

One likely set of cues that listeners use to identify vocal emotion is prosody: pitch and timing information. Happy speech, for instance, typically has more variable pitch and volume, higher overall pitch level, and a faster speaking rate, whereas sad speech sounds exhibit lower average pitch, attenuated loudness and pitch variation, and a slower pace of speech (Morton \& Trehub, 2001). Previous work demonstrates that humans use paralinguistic cues during speech to alert co-communicators to their current emotional state (Kehrein, 2002). However, this ability to attribute certain paralinguistic cues to particular emotional states may not be fully present at birth, but may require learning through lengthy exposure to one's native language.

One indication of the learned nature of paralinguistic processing is that children experience difficulty in identifying vocal emotional cues (Morton \& Trehub, 2001); for instance, 6-year-olds who hear "my mommy gave me a treat" with "sad" emotional prosody will report that the speaker sounded happy, suggesting that they are still learning the mapping between particular speech patterns and emotional states. Further research suggests that these learned aspects may be language-specific (Thompson \& Balkwill, 2006), though those authors do not pinpoint particular cues that may be relevant, nor do they offer a hypothesis as to what level of fluency one needs to access the learned aspects of emotional speech. In a related area, speaker recognition shows some language specificity in infants (Johnson et al., 2011) and adults (Bregman \& Creel, 2012). If encoding of vocal emotional information works similarly to encoding of voices, then good emotional recognition within a language may be dependent on lengthy language experience, and may not generalize to emotion recognition in other languages.

## General ability to recognize vocal emotion

On the other hand, there is evidence that expertise in processing the cues that communicate vocal emotion may generalize widely across domains. This implies that better attention to or encoding of pitch for another purpose or in another domain may lead to better perception of vocal emotion. For example, certain types of languages have been claimed to boost pitch perception abilities: Speakers of tone languages such as Mandarin are better at making relative
pitch distinctions in musical stimuli (Pfordresher \& Brown, 2009). Conversely, musicians are better than non-musicians at brain encoding of linguistic pitch changes (Wong et al, 2007). These studies suggest that facility with pitch processing generalizes even across domains. This implies that good linguistic pitch processing should facilitate pitch processing generally, which would thereby facilitate perception of pitch-related vocal-emotional information, even outside one's native language. The novel prediction for vocal affect detection is that, over and above languagespecific knowledge, tone-language speakers may excel at perceiving vocal-emotional information due to their language background.

## The current study

The current study focuses on two hypotheses about processing of vocal affect. First, the language-specificity hypothesis suggests that listeners are best at identifying vocal affect in their native language, due to lengthy perceptual learning of vocal correlates of emotional states specific only to that language. This also assumes any second language that the speakers are fluent or near-fluent at will also experience this emotional state comprehension. Crucially, a listener who is not a fluent speaker of a language will have difficulty identifying emotion in that language relative to fluent speakers. Second, the tonelanguage benefit hypothesis posits that listeners with a history of speaking tone languages will show good identification of vocal affect even in non-native languages because tone languages generally facilitate listeners' processing of pitch information, and pitch information is one important cue to affect.

To investigate how language background affects emotional speech processing, we asked speakers of Mandarin Chinese (a tonal language) who also spoke English, and speakers of American English (a non-tonal language) to identify emotion in utterances produced in Mandarin and English. Participants' abilities to identify emotions in their own language were compared to their ability to identify emotions in languages unfamiliar to them.

## Methods

## Participants

Thirty-six undergraduates from the University of California, San Diego participated in this study for class credit. Eighteen of the participants were native English speakers who did not speak Mandarin Chinese, and the remaining eighteen were native speakers of Mandarin Chinese who also spoke English fluently as a second language. English was a second language for each of the 18 Mandarin speaking participants, who acquired English at a mean age of 8 years (range: 0-17).

## Stimuli

Eight speakers recorded 96 sentences each in their native language. Four speakers (2 male, 2 female) were native
speakers of English, and the other four (2 male, 2 female) were native speakers of Mandarin Chinese. The speakers' ages ranged from 19 to 26 , with a mean age of 21.75 .

Sentences spanned six different emotional states: anger, disgust, fear, happiness, sadness, and surprise. Sentence semantic content was created to elicit the intended emotion, to make the task of emotional speech production more naturalistic for our speakers. The 16 sentences for each emotion were originally written in English and translated to Mandarin Chinese. Each sentence contained five syllables in the English version. The translation was retranslated separately by all four Mandarin speakers to ensure a good content match with the original English sentences.

Sentences were recorded in a sound-attenuated booth and saved as .wav files. Files were edited so that each sentence had its own sound file. Two types of sound files were created for each sentence. One type of stimulus (Figure 1a) reflected the original recording, complete with naturalistic emotional sentence semantic content. The other (Figure 1b) was low-pass filtered at 500 Hz using Praat software (Boersma \& Weenink, 2011). Low-pass filtering removes high-frequency information including cues to consonants and vowels, while retaining low-frequency information in the speech signal. The result is a muffled, unintelligible sound that preserves fundamental frequency variability, including lexical tone (Mandarin only), prosody, and speech rate. This manipulation allowed for the measurement of prosody recognition without the confound of languagespecific semantic content. Each file was set to an average loudness of 60 decibels.
(a)


Figure 1. Spectrogram of an (a) unfiltered and (2) low-pass filtered angry sentence.

## Procedure

Each participant listened to 384 total sentences, counterbalanced across two conditions so no participant heard the same sentence from the same speaker or in the same language or in the same filter condition twice. Of the 384 sentences, half each were English and Mandarin; crossed with this, half were unfiltered and half were filtered. Participants were asked to identify the emotional state of each sentence as it was presented through Sennheiser HD 280 Pro headphones. The computer monitor displayed the six possible emotions and a number that corresponded to each emotion. Participants pressed the number key that they
thought matched the emotion of the sentence. Perceived emotional responses and reaction times were recorded in Matlab using the Psychtoolbox3 (Brainard, 1997; Pelli, 1997).

## Results

For the current experiment, the language-specificity hypothesis and the tone-language hypothesis make similar predictions regarding the participants due to the fact that our tone language speakers were fluent bilinguals. If listeners showed native language specificity of emotion recognition, then accuracy should be higher for English listeners in English. Mandarin-English bilinguals should perform equally well in both languages-either due to knowledge of both languages, or due to enhanced abilities as tonelanguage speakers. This may vary by degree of language fluency, as assessed by age of English acquisition. Importantly, this pattern should still hold for filtered speech, which crucially does not contain any semantic language information. That is, if listeners are using language-specific acoustic cues to vocal emotion, they should still be more accurate at recognizing emotion in a familiar language even when lexical cues are removed.

To test this hypothesis, we performed a mixed ANOVA on recognition accuracy with Listener Language (English, Mandarin) as a between-participants variable, and Stimulus Language (English, Mandarin) and Intelligibility (unfiltered, filtered) as within-participants variables. The three-way interaction between these variables was significant ( $\mathrm{F}(1,34)=121.70, \mathrm{p}<.0001$ ). This interaction qualified all lower-level effects and interactions.


Figure 2. Accuracy for unfiltered sentences with standard errors. The dotted line represents chance performance

Therefore, we broke the data out into filtered and unfiltered data to better highlight interactions at that level. For unfiltered sentences, with full naturalistic verbal content, there was a significant interaction between Listener Language and Stimulus Language ( $\mathrm{F}(1,34)=325.58$, $\mathrm{p}<.0001$ ), indicating that participants were better able to comprehend emotional affect in languages they spoke highly fluently, which is supportive of our languagespecificity hypothesis. This result is also important as a control for the stimuli used, and demonstrated that the sentences provided ample emotional content clues. When presented with unfiltered speech, English speakers were significantly more accurate with English speech ( $\mathrm{t}(17)=63.706, \mathrm{p}>.0001$ ) whereas Mandarin speakers were equally proficient at identifying emotional affect in both languages $(\mathrm{t}(17)=.891, \mathrm{p}=.385)$ as would be expected from their language background.

Considering the filtered stimuli, which contained only prosodic cues, we again found an interaction of Listener Language and Stimulus Language ( $\mathrm{F}(1,34)=46.278$, $\mathrm{p}<.0001$ ), indicating that even when there was a lack of verbal information, participants were significantly more capable to parse emotional affect when presented with languages they spoke fluently. Considering each listener language group individually, English speakers were significantly more accurate in identifying the intended emotion when given filtered English speech than filtered Mandarin speech $(t(17)=9.949, p<.001)$. Mandarin speakers, however, showed good performance on filtered speech in both languages with no significant differences in accuracy $(t(17)=1.923, p=.0714)$.


Figure 3. Accuracy for filtered (prosody only) sentences with standard errors. The dotted line represents chance performance

Finally, we assessed whether English performance was affected by degree of English fluency by calculating the correlation between bilinguals' English accuracy on filtered speech and their age of English acquisition. This correlation was not significant $(r(16)=-.2176, p=.3856)$, suggesting that somewhat surprisingly, age of second language acquisition was unrelated to ability to process semantic-information-free speech.

## Discussion

Previous work has demonstrated a link between tonal language background and enhancements of abilities in other perceptual domains such as music (Pfordresher \& Brown, 2009), but the link between language background and vocal emotion had been underexplored. The current study explored the relationship between language background and vocal affect identification, focusing on hypotheses of a language-specific benefit and a tone-language benefit in vocal affect identification.

In support of these hypotheses, we showed that listeners are better at discerning emotional content in speech for all languages they have achieved fluency in, even when highfrequency lexical cues are removed. This is important, because the lack of high frequency information removes any clues as to the specific language being presented. This means that all performances on filtered speech represent the participant's processing of the low frequency emotional pitches and tonal changes without the influence of any clues to the actual language. This implies that any responses are based entirely on pitch processing, and any benefits can only come from validation of one of the two hypotheses presented. The data demonstrate, specifically, that English monolinguals identified emotions more accurately in English than in Mandarin, whereas Mandarin-English bilinguals showed equivalent performance in both languages. This is consistent with the language-specificity hypothesis: that listeners are better at discerning cues to emotion in their native language. However, due to the design of the current study, it is also consistent with the tone-language facilitation hypothesis: that tone-language speakers, due to lengthy experience attending to finegrained pitch characteristics of language, have a general advantage at recognizing vocal emotion. That is, Mandarin listeners performed at above-chance levels in both languages because they are tone-language speakers, not only because they also speak English. This would predict that Mandarin listeners would also be superior at emotion recognition in a completely unfamiliar language, a hypothesis we are currently testing. Further study is currently being performed to address these design limitations, and will alleviate the current confound of the tonal language speaking subjects being fluent in the languages of all the presented stimuli

Nevertheless, in the current study, discussion of the implications of both hypotheses is warranted by the data. In regards to the first hypothesis, the language-specific hypothesis, our data are consistent with the expectation that
monolingual listeners perform better in identifying vocal emotion from prosodic cues in their native language. Listeners familiar with two languages (Mandarin and English) performed comparably in both languages. It is possible more subtle effects were also present in the data regarding the bilingual speakers' performances. For instance, if the emotional-speech processing capabilities are affected by the "sensitive period" demonstrated for phonology, then early-exposed speakers of a second language should show native-level abilities in emotional speech processing, and those who acquired their second language later should not. When a Pearson's productmoment correlation was run on the data, however, the correlation suggested no link between a subject's age of acquisition and performance on filtered speech processing. If age of acquisition does not play a correlational role, some other aspect of the subjects' background must be dictating their abilities.

This leaves the possibility that a general benefit for tonelanguage speakers better accounts for the results. The current results, although consistent with a language specific (or familiar-language) benefit, are also consistent with the tone-language benefit hypothesis; speakers of tonal languages performed equally well at identifying vocal affect in both languages presented. However, the data described here cannot distinguish whether bilinguals' good performance in both languages resulted from their tonelanguage background, or simply being fluent in both languages present in the stimuli. Further study will present these subject groups with tonal and non-tonal languages with which they are not familiar.

If there is a general tone-language advantage, Mandarin speakers should outperform English speakers on all unfamiliar languages. This would indicate that a tonal language background affords the speaker with a type of emotional prosody processing training, and that there are aspects of tonal languages that improve the processing capabilities of individuals in emotional speech. This might be true if there are universal pitch characteristics that can be found in all languages, but are very subtle and difficult to pick up on if the listener does not have sufficient training in either that specific language or excellent awareness in pitch perception in general.

There could even be a specific tone-language advantage, such that tone-language speakers would outperform nontone language speakers only for unfamiliar tone languages. This data pattern might hold if tone languages use devices to convey vocal emotion that are similar across a range of tone languages, but different than non-tone languages. We are currently testing these possibilities regarding the tonelanguage advantage with English speakers and MandarinEnglish bilinguals and four languages (English; Mandarin; Dutch [unfamiliar non-tone language]; Vietnamese [unfamiliar tone language]).

The present data demonstrate that there is emotional affect information present even without higher frequency information. It also provides evidence that speakers are
capable of picking up on this information without relying on distinct linguistic information in languages that they are familiar with. The current study provides a tantalizing peek into the emotional affect processing provided by language background, and with further study already in process, moves our understanding of vocal emotional affect processing forward.

## Acknowledgements

CIF was supported by the Center for Research in Language at UC San Diego, and SCC was supported by an NSF CAREER Award (BCS-1057080). We would like to acknowledge the help of undergraduate research assistants Sheryl Soo, Christina Hwang, and Ben Howard, who were instrumental in developing language stimuli and collecting data on participant's language background.

## References

Balkwill, L., \& Thompson, W. (1999). A cross-cultural investigation of the perception of emotion in music: Psychophysical and cultural cues. Music perception, 17(1), 43-64.
Bregman, M. R., \& Creel, S. C. (2012). Learning to recognize unfamiliar voices: the role of language familiarity and music experience. In N. Miyake, D. Peebles, \& R. P. Cooper (Eds.), Proceedings of the 34th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.
Johnson, E. K., Westrek, E., Nazzi, T., \& Cutler, A. (2011). Infant ability to tell voices apart rests on language experience. Developmental Science, 5, 1002-1011.
Kehrein, R. (2002). The prosody of authentic emotions. Proc. Speech Prosody Conf (pp. 423-426).
Kuhl, P. K. (1994). Learning and representation in speech and language. Current Opinion in Neurobiology, 4, 812822.

Ley, R. G., \& Bryden, M. P. (1982). A dissociation of right and left hemispheric effects for recognizing emotional tone and verbal content. Brain and cognition, 1(1), 3-9.
Morton, J. B., \& Trehub, S. E. (2001). Children's Understanding of Emotion in Speech. Child Development, 72(5), 834-843.
Pfordresher, P. Q., \& Brown, S. (2009). Enhanced production and perception of musical pitch in tone language speakers. Attention, Perception, \& Psychophysics, 71(7), 1385-1398. Springer.
Sauter, D. A, Eisner, F., Ekman, P., \& Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. Proceedings of the National Academy of Sciences of the United States of America, 107(7), 2408-12.
Slevc, L. R., \& Miyake, A. (2006). Individual differences in second-language proficiency: does musical ability matter? Psychological Science, 17(8), 675-81.
Thompson, W. F., \& Balkwill, L.-L. (2006). Decoding speech prosody in five languages. Semiotica, 2006(158), 407-424. Walter de Gruyter.

Thompson, W. F., \& Matsunaga, R. I. E. (2004). Recognition of emotion in Japanese, Western, and Hindustani music by Japanese listeners 1, 46(4), 337-349. Thompson, W., Schellenberg, E. G., \& Husain, G. (2006). Perceiving Prosody in Speech. Annals of the New York Academy of Sciences, 999, 530-532.
Thompson, W. F., Schellenberg, E. G., \& Husain, G. (2004). Decoding speech prosody: do music lessons help? Emotion (Washington, D.C.), 4(1), 46-64.

# Cooperation Decreases with Development of Number Sense 

Ellen E. Furlong (ellen.furlong@yale.edu)<br>Department of Psychology, PO Box 208205 New Haven, CT 06511USA

John E. Opfer (opfer.7@osu.edu)
Department of Psychology, 245 Psychology Bldg Columbus, OH 43210 USA


#### Abstract

Cooperation among children can appear haphazard, a finding often attributed to deficient social skills and moral reasoning. Here we took a game theoretical approach to understand development of cooperation, using the prisoner's dilemma to test an alternative source of age-differences in cooperative behavior-how children and adults represent the numerical magnitudes of payoffs for cooperating versus not. We found that as incentives increased solely in numerical magnitude, speed of incentive comparisons decreased and cooperation increased. Further, though children tended to be more cooperative than adults, effect of age on cooperation was moderated by speed of incentive comparison. We conclude that representations of numeric value constrain how economic rewards affect cooperation and that children's greater cooperativeness may be attributed to a poor sense of numerical value.


Keywords: Cooperation; Numerical Cognition; Cognitive Development.

## Introduction

Development of cooperation-how it begins, how it changes over time, and what factors promote it-has invited speculation for at least 350 years. According to Rousseau (1754/2007), cooperation is our birthright, society breeds competition; according to Hobbes (1651/2008), we are naturally competitive, society promotes cooperation. Although scientists champion neither position, nearly all look to the same factors-social constructs-to explain development of cooperation (Miles, Hare \& Tomasello, 2006; Warneken \& Tomasello, 2006; 2007; Warneken, Chen \& Tomasello, 2006). Research on the role of social constructs (i.e., theory of mind, communication, fairness norms, trust, social tolerance) on development of cooperation finds support for both views-development breeds either competition or cooperation, depending on the context (i.e., Damon, 1975; Lane \& Coon, 1972; Piaget, 1932; Warneken \& Tomasello, 2006; 2007; Warneken, Chen \& Tomasello, 2006).

One possible way to explain the role of context on development of cooperation is to consider that cooperation may result, not only from developing social skills, but also from how cooperative incentives are mentally represented (Furlong \& Opfer, 2009). This role for incentive structure has been explored by game theory, which predicts circumstances under which organisms are likely to cooperate and tests these predictions using games such as

Prisoner's Dilemma (PD). Following this approach, we propose a novel and surprising influence on cooperationhow children represent numeric value. In the following sections we: (1) follow Hobbes' lead and provide a game theoretical analysis linking incentive structure to cooperative behavior, (2) explain how developing representations of number affect representation of incentive structures, and (3) show how this analysis accurately predicts Rousseau's claim that cooperation would decrease with age and experience.

## Game Theory Links Incentives to Cooperation

Insight into why cooperation depends heavily on contextual factors comes from game theory, which makes predictions about the incentive structures under which organisms are likely to cooperate. Incentive structures in which small immediate costs of cooperation are offset by large immediate benefits, known as mutualisms, commonly lead to cooperation. Even simple organisms-such as fish and ants-readily engage in cooperation under mutualist incentive structures (Bronstein, 2001; Mesterton-Gibbons \& Dugatkin, 1992; Trivers, 1971).

While cooperative mutualisms occur readily throughout the animal kingdom, reciprocity--in which short-term costs of cooperation are exchanged in expectation of long-term benefits--is relatively scarce. Indeed, in many cases, these exchanges can be explained by simpler mechanisms such as kin selection, where cooperation does not occur in expectation of any future exchange (i.e., Maynard-Smith, 1965; Trivers, 1971; Stevens et al, 2005).

Biologists typically account for high mutualism rates and low reciprocity rates by arguing that mutualism poses relatively few risks (costs are immediate and relatively low and benefits are immediate and relatively high), whereas the additional temporal element of reciprocity makes it fairly risky (costs are immediate and high and future large benefits are tenuous and may never realize; Maynard-Smith, 1965; Stevens et al, 2005; Trivers, 1971). The likelihood of cooperation depends, therefore, on the relation between benefits and costs-in other words, its incentive structure.

How incentive structure can affect cooperative behavior is often examined using the prisoner's dilemma game (Clements \& Stephens, 1995; Noe, 2006; Rapoport \& Chammah, 1965; Valev \& Chater, 2006). The prisoner's dilemma can be conceptualized in this way: Suppose two children, Bonnie and Clyde, have agreed to charge $\$ 3$ per
glass in competing lemonade stands. If Bonnie cooperates and charges $\$ 3$, she'll earn $\$ 3$; however, if she reneges and drops her price, she may be able to sell more lemonade for a cheaper price (say, 2 cups for $\$ 2.50$ each yielding $\$ 5$ ). If both renege, their prices will drop until they sell lemonade at cost-\$1 per cup. If Clyde drops his price, but Bonnie does not, Bonnie will lose her clients to Clyde and earn nothing (Figure 1a).

Generally, if players meet only once, they maximize rewards by defecting; however, if players interact repeatedly, they maximize rewards by cooperating (Axelrod \& Hamilton, 1981; Rapoport \& Chammah, 1965). Sadly, even in iterated dilemmas, people and animals tend to defect (i.e., Baker \& Rachlin, 2002; Dawes \& Thaler, 1998). However, reciprocal dilemmas can elicit mutualistic behavior simply by manipulating incentives-for example, by changing the reward for mutual cooperation from $\$ 3$ to $\$ 6$ and the temptation to defect from $\$ 5$ to $\$ 8$, cooperation rates increase (Figure 1b; Rapoport \& Chammah, 1965; Valev \& Chater, 2006).
(a) A Prisoner's Dilemma
(b) A Mutualism



Figure 1. Payoff matrices characteristic of Prisoner's Dilemma (where cooperation is rare) and Mutulism (where cooperation is common)

## Representation of Incentive Structure Depends on Representation of Numeric Value

Why manipulating incentives results in mutualistic behavior might be explained by how the brain represents numeric quantity. Specifically, as numeric values increase, discriminability decreases; thus, while participants quickly determine that $5>3$, they are slower to determine that $8>6$ (Moyer \& Landauer, 1967; Starkey \& Cooper, 1980).

This numeric size effect fits into a broader literature suggesting non-symbolic numeric quantities may be represented logarithmically: that the brain overestimates differences among small quantities and compresses differences among large quantities (Dehaene, 1997; Nieder \& Miller, 2003). Therefore, the difference between 3 and 5 feels larger than the difference between 6 and 8 .

The suggestion that subjective incentives in the brain may be quite different than objective incentives in the real world
is not new. This is the chief insight of prospect theory (Kahneman \& Tversky, 1979): choices and framing of incentives may affect their subjective value. As Bernoulli (1738/1954) famously observed, "a gain of 2000 ducats is more significant to a pauper than to a rich man though both gain the same amount." The framing of incentives-in this case, the initial endowment-may affect decisions about those incentives.

Although most theories of decision-making rely on prospect theory, in which economic value is subject to size effects (Kahneman \& Tversky, 1979), we argue that numeric value, independent of economic value, can affect cooperative behavior. This hypothesis leads to an interesting implication-namely, converting a reciprocal dilemma into a mutualism may not require manipulating economic values of incentives; rather, it may be accomplished by manipulating numeric values alone.


Figure 2: Payoff matrices used by Furlong \& Opfer (2009).
This surprising hypothesis was recently tested in a series of experiments in which adult participants played one of four prisoner's dilemma games, identical except for incentive structure (Figure 2; Furlong \& Opfer, 2009). As observed in previous studies, subjects in the baseline (\$1) condition showed relatively high rates of defection and low rates of cooperation. When rewards were increased a hundred-fold to $\$ 100$, however, subjects showed the opposite behavior-low rates of defection and high rates of cooperation. This finding could be explained by the standard economic value model-perhaps subjects cooperated more in the $\$ 100$ condition simply because there was more at stake. On the other hand, subjects may have cooperated more simply because 100 is a larger number than 1 . In support of the latter explanation, subjects playing for $100 \phi$, which is economically equivalent to playing for $\$ 1$ and numerically equivalent to playing for $\$ 100$, showed identical behavior to those playing for $\$ 100$. Similarly, subjects in the $1 \phi$ condition behaved identically to those in the $\$ 1$ condition, even though the increase from $1 \phi$ to $\$ 1$
represents a one hundred fold increase in economic value. Thus, cooperative behavior changed in response to numeric value, but not in response to economic value.

These numeric-magnitude effects are consistent with the idea that numbers associated with payoff values are represented logarithmically. That is, the linear model predicts defection whenever the ratio between the reward for mutual cooperation and the temptation to defect is less than 1 , and because $300 \phi / 500 \phi$ equals $\$ 3 / \$ 5$, changing numeric values would not matter. This preservation of ratio information does not obtain if numeric values are scaled logarithmically, as $\ln (300) / \ln (500)$ is approximately 1 (i.e., temptation to defect and cooperate are nearly equal), whereas $\ln (3) / \ln (5)$ is approximately .68 (i.e., temptation to defect is higher than temptation to cooperate).

## Hypothesized Effects of Developing Number Representations on Cooperation

In Furlong and Opfer's (2009) work, big numbers increased cooperation-a finding predicted by the way the mind represents non-symbolic quantities to increase logarithmically with actual value. When representing symbolic quantities, however, important developmental differences emerge (see Opfer \& Siegler, 2012, for review). For young preschoolers, numeric symbols are meaningless stimuli. For example, 2- and 3-year-olds who count flawlessly from $1-10$ have no idea that the number 6 is greater than the number 4 , nor do children of these ages know how many objects to give an adult who asks for 4 or more (Le Corre et al., 2006; Opfer, Thompson, \& Furlong, 2009; Sarnecka \& Carey, 2008). As young children gain experience with the symbols in a given numerical range and associate them with non-verbal quantities in that range, they initially map them to a logarithmically-compressed mental number line (Berteletti et al., 2010; Booth \& Siegler, 2006; Opfer, Thompson, \& Furlong, 2010; Siegler \& Booth, 2004; Siegler \& Opfer, 2003; Thompson \& Opfer, 2010). Over a period that typically lasts $1-3$ years for a given numerical range $(0-10,0-100$, or $0-1000)$, their mapping changes from a logarithmically compressed form to a linear form, in which subjective and objective numerical values increase in a 1:1 fashion. Use of linear magnitude representations occurs earliest for the numerals that are most frequent in the environment, that is the smallest whole numbers, and it gradually is applied to increasingly large numbers.

The logarithmic-to-linear shift in children's representations of symbolic quantities expands children's quantitative thinking profoundly. It improves (1) children's ability to estimate the positions of numbers on number lines (Siegler, Thompson, \& Opfer, 2010), (2) to estimate the measurements of continuous and discrete quantities (Booth \& Siegler, 2006; Laski \& Siegler, 2007; Thompson \& Siegler, 2010), (3) to categorize numbers according to size
(Laski \& Siegler, 2007; Opfer \& Thompson, 2008), (4) to remember numbers that they have encountered (Thompson \& Siegler, 2010), and (5) to estimate and learn the answers to arithmetic problems (Booth \& Siegler, 2006). All of these abilities also have important educational roles, leading to use of linear representations of number being highly correlated with mathematics achievement and a broadly effective target of instructional interventions. Thus, children's representations of symbolic quantities-like those used in the payoff matrices of prisoner's dilemma games-change dramatically with age and experience.

Developmental differences in representations of symbolic magnitudes have important implications for how children and adults are likely to respond to economic incentives. That is, if representations of numeric quantity affect cooperative decisions, adults--who are least likely to use logarithmic representations of symbolic quantity-should show the smallest effect of numeric value on cooperative behavior, whereas young children-who are most likely to use logarithmic representations-should show the largest effect of numeric value on cooperative behavior. This is a somewhat surprising and counter-intuitive prediction: because behavioral variability typically decreases with age, effect sizes generally increase with age. To test this hypothesis, we explored the effects of numeric and unit changes on cooperation in third-grade children, fifth-grade children and adults engaged in a prisoner's dilemma game.

## Method

## Participants

Undergraduate students ( 23 males, 25 females; $\mathrm{M}=19.58$ years of age, $s=1.43$ ), third-grade students ( 19 males, 29 females; $\mathrm{M}=9.33$ years of age, $\mathrm{s}=.33$ ) and fifth-grade students ( 25 males, 23 females; $M=11.06$ years of age, $\mathrm{s}=.43$ ) from largely middle-class schools were randomly assigned to play one of four iterated prisoner's dilemma games (IPDs) identical except for payoff structure (Figure 2). All participants received a sticker (children) or course credit (adults) for participating.

## Design and Procedure

Participants played IPDs against computers using a "Tit-forTat" (TFT) strategy - initially cooperating and thereafter mirroring the participant's behavior on the preceding trial. Participants received no instruction on strategy but were told they were going to play a game called "rock/paper" in which they could earn pretend money (rock was defect and paper cooperate). They were further instructed that the goal was to earn as much money as possible, and that the amount of money they earned depended on how they and the computer played the game. Participants could click on an


Figure 3. Cooperation in adults, 5th graders and 3rd graders in the prisoner's dilemma game.
icon of a piece of paper (cooperate) or a hand in a fist (defect) to make their choice. Once they made their choice the computer's 'choice' was presented as well as a running total of each player's score. Each participant was allowed as much time as they wanted to complete each of 45 trials.

The design was a 2 (unit: dollars or cents) X 2 (number: 1 or 100) factorial design resulting in four games, identical except for payoff structure (Figure 2) - a numerically small dollars condition (\$1), a numerically large cents condition ( $100 \phi$ ), a numerically small cents condition (1申), and a numerically large dollars condition (\$100).

We measured four indices of cooperative behaviorindividual cooperation (total number of trials in which the participant cooperated), mutual cooperation (number of trials in which participant and computer engaged in cooperation together), mutual defection (number of trials in which participant and computer defected together) and forgiveness, a measure of number of trials to cooperate after the computer's first defection.

To ensure children understood the monetary conversion, children were asked, "how many pennies are in a dollar?" Only one child (a third-grader) answered this question incorrectly; his data were excluded from analyses.

Additionally, subjects participated in a computerized number discrimination task in which they were presented with two numbers (i.e., 3 and 5) and asked to press one of two keys to indicate which was the larger as quickly and accurately as possible. Combinations of the numeric values presented to participants in the numeric discrimination task were identical to the prisoner's dilemma task.

## Results and Discussion

First, we explored effects of number, unit and value on cooperative behavior in all three age groups. This analysis is followed by an exploration of the magnitude of the effect of numeric value on cooperation across ages. Finally, we explore the relation between numeric representation in the number comparison task with cooperative behavior in the prisoner's dilemma task.

Two (units: dollars, cents) by two (number: 1, 100) MANOVAs were conducted on the four indices of cooperation. No age groups showed a main effect of unit,
nor did any age group show an interaction of unit with number on their cooperative behavior. Further, no age group showed an effect of economic value ( $1 \phi, \$ 1$ or $\$ 100$ ) on cooperative behavior.

Cooperation in all three groups, however, varied with number (Figure 3; Adults: $F[4,41]=2.66, p=.046$; 5th graders: $F[4,41]=5.09, p=.002$; 3rd graders: $F[4,41]=3.89$, $p=.009$ ). Specifically, numerically greater rewards increased individual cooperation (Adults: $\mathrm{F}[1,44]=10.06, \mathrm{p}=.003$; 5th graders: $\mathrm{F}[1,44]=10.42, \mathrm{p}=.002$; 3rd graders: $\mathrm{F}[1$, $44]=13.49, \mathrm{p}=.001$ ) such that changing rewards from $3 \phi$ to $300 \varnothing$ increased cooperation rates, but an economically identical change from $3 \phi$ to $\$ 3$ did not. The same pattern was evident in rates of mutual cooperation, where numerically large rewards elicited more mutual cooperation than numerically small rewards (Adults: $F[1,44]=7.18$, $p=.01$; 5th graders: $F[1,44]=9.46, p=.004 ; 3$ rd graders: $\mathrm{F}[1$, $44]=8.49, p=.006$ ). Further, numerically large rewards elicited less mutual defection than numerically small ones (Adults: $F[1,44]=9.18, p=.004 ; 5$ th graders: $F[1,44]=6.05$, $p=.02$; 3rd graders: $F[1,44]=9.75, p=.003$ ). While no effect of number was observed for forgiveness in adults and 5th graders, 3rd graders did show an effect of number on forgiveness ( $F[1,44]=5.94, p=.02$ ), requiring fewer trials to 'forgive' their partner for large numeric values than for small numeric values.

A 3 (age: 3rd grade, 5th grade, adult) X 2 (number: 1, 100) MANOVA also revealed main effects of age ( $F[8$, $272]=5.92, p<.001)$ and number $(F[4,135]=8.30, p<.001)$ on cooperation. This effect was observed for individual cooperation $(F[2,138]=9.57, p<.001)$ and mutual defection ( $F[2,138]=19.05, p<.001$ ). Results for mutual cooperation $(F[2,138]=2.18, p=.11)$ and forgiveness $F[2,138]=2.41$, $p=.09)$ trended toward significance. Post-hoc tests revealed 3rd graders had more individual cooperation than both 5th graders and adults ( $p \mathrm{~s}<.01$ ). This pattern held true for mutual defection (5th graders: $\mathrm{p}=.001$; adults: $\mathrm{p}<.001$ ). Fifth graders and adults did not differ from each other on individual cooperation $(p=.82)$ but they trended to differ on mutual defection $(p=.06)$. No differences were found between the groups on mutual cooperation or forgiveness.


Figure 4: 3rd graders showed larger effects of number on individual and mutual cooperation and forgiveness than 5th graders and adults.

We next compared values of Cohen's $d$, a measure of effect size, for each of the four indices of cooperation in each of the three age groups. We expected to find a larger effect of number (a larger value of $d$ ) in third-grade children than in fifth-grade children and adults. This predicted pattern was indeed observed for three of our four measures of cooperation: 3rd graders showed a larger effect of numeric value on individual cooperation ( $d=1.08$ ), mutual cooperation ( $d=1.12$ ) and forgiveness ( $d=0.70$ ) than 5th graders (individual cooperation: $d=0.95$; mutual cooperation: $d=0.90$; forgiveness: $d=0.31$ ) or adults (individual cooperation: $d=0.93$; mutual cooperation: $d=0.78$; forgiveness: $d=0.30$ ). Effect sizes for mutual defection were roughly equal across all three groups (3rd: $d=0.92$; 5th: $d=0.73$; adults: $d=0.88$; Figure 4).

We hypothesized that the age related increases in effect size were due to number representations, and that the 5th graders were already demonstrating adult-like number cognition. Data from individual participants were analysed to determine whether their reaction times in the number discrimination task were best fit using a linear difference between the two comparison numerals (i.e., $5-3$ ) or a logarithmic difference (i.e., $\ln [5]-\ln [3]$ ). This allowed us to classify participants are relying on a more linear or more logarithmic representation.

As predicted, as age increased reliance on a linear representation increased as well ( $\chi 2(2)=4.88, \mathrm{p}=.08) ; 57 \%$ of 3rd graders were best fit by the linear model; $69 \%$ of 5 th graders and $77 \%$ of adults were best fit by the linear model. A 2 (representation type: logarithmic or linear) X 2 (number: 1 or 100) MANOVA on cooperation revealed a significant effect of representation type on cooperation ( $F[4$, 96] $=2.78, p=.03$ ) such that participants best fit by the logarithmic model showed greater individual cooperation ( $F$ $(1,104)=3.41, p=.06)$ and mutual defection $(F$ (1, $104)=7.40, p<.01$ ) than participants best fit by the linear model.

## Conclusion

What is the nature of human cooperation? Are we Rousseauian, naturally cooperative, or are we Hobbesian, naturally competitive? The answer may be Both: we start life cooperative (a la Rousseau), but become competitive with age and experience (a la Hobbes). Our cooperative decisions may be shaped, however, not just by changing social influences, but also by developing numeric representations.

Consistent with this perspective, adults--who represent numbers relatively precisely--showed more individual and mutual cooperation and less mutual defection in response to large numbers, not large economic values. Not only did third- and fifth-grade children also demonstrate this pattern, but age related changes in number representation were associated with changes in cooperation: 3rd graders showed a larger effect of number than the older children and adults in individual cooperation, mutual defection and forgiveness.

Further, numerical representations predicted individual cooperation, mutual cooperation, defection and forgiveness rates; subjects who relied more on logarithmic representations demonstrated higher rates of individual and mutual cooperation, lower rates of mutual defection, and took less time to forgive their partner than subjects who relied more on linear representations. These results suggest that logarithmic representations may make it harder to discriminate incentives, resulting in them being treated more like a cooperative mutualism than a reciprocity.

Our results may be able to shed light on previous findings that children appear Hobbesian or Rousseauian depending on the context (Damon, 1975; Lane \& Coon, 1972; Piaget, 1932; Warneken \& Tomasello, 2006, 2007; Warneken, Chen \& Tomasello, 2006). Perhaps these inconsistencies in cooperation can be explained by how costs and benefits are represented in the minds of children. Children may be more likely to cooperate in tasks in which they perceive the costs to be minimal and/or the benefits large (e.g., holding a door open for a stranger), but may be less likely to cooperate in tasks in which they perceive the costs to be large and the benefits minimal (e.g., providing another child with a reward out of one's own stock).

Thus, while it may not be possible to definitively resolve the Hobbes-Rousseau debate, the combination of game theory and psychology of number may make it possible to predict which circumstances incentive cooperation. Put simply, children may be more Rousseauian when costs and benefits are hard to discriminate, but more Hobbesian when they are easily discriminable.

## References

Axelrod, R. \& Hamilton, W.D. (1981). The evolution of cooperation. Science, 211, 1390-1396.
Baker, F. \& Rachlin, H. (2002). Teaching and learning in a probabilistic prisoner's dilemma. Behavioural Processes, 28, 211-226.

Berteletti, I., Lucangeli, D., Piazza, M., Dehaene, S., \& Zorzi, M. (2010). Numerical Estimation in Preschoolers. Developmental Psychology, 46, 545-551.
Bernoulli, D. (1954). Exposition of a new theory on the measurement of risk. In Louise Sommer (Ed. \& Trans). Econometica, 22, 22-36 (Original work published 1738).
Booth, J., \& Siegler, R. (2006). Developmental and individual differences in pure numerical estimation. Developmental Psychology , 41, 189-201.
Bronstein, J.L. (1994). Our current understanding of mutualism. Quarterly Review of Biology, 69, 31-51.
Clements, K.C. \& Stephens, D.W. (1995). Testing models of non-kin cooperation: Mutualism and the Prisoner's Dilemma. Animal Behaviour, 50, 527-535.
Damon, W. (1975). Early conceptions of positive justice as related to the development of logical operations. Child Development, 46, 301-312.
Dawes, R. \& Thaler, R. (1988). Cooperation. Journal of Economic Perspectives, 2, 187-197.
Dehaene, S. (1997). The Number Sense: How the Mind Creates Mathematics. New York: Oxford University Press.
Furlong, E.E. \& Opfer, J.E. (2009). Cognitive constraints on how economic rewards affect cooperation. Psychological Science, 20, 11-16.
Halberda, J., \& Feigenson, L. (2008). Developmental change in the acuity of the "number sense": The approximate number system in 3-, 4-, 5-, and 6-year-olds and adults. Developmental Psychology, 44, 1457-1465.
Hobbes, T. (2008). Leviathan. In J.C.A. Gaskin (Ed. \& Trans). New York: Oxford World Classics. (Original work published 1651).
Kahneman, D. \& Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, 47, 263291).

Lane, I.M. \& Coon, R.C. (1972). Reward allocation in preschool children. Child Development, 43, 1382-1389.
Le Corre, M., Van de Walle, G., Brannon, E.M., \& Carey, S. (2006). Re-visiting the competence/performance debate in the acquisition of the counting principles. Cognitive Psychology, 52, 130-169.
Maynard-Smith, J. (1965). The evolution of alarm calls. American Naturalist, 99, 59-63.
Meles, A., Hare, B. \& Tomasello, M. (2006). Chimpanzees recruit the best collaborators. Science, 311, 1297-1300.
Mesterton-Gibbons, M. \& Dugatkin, L.A. (1992). Cooperation among unrelated individuals: Evolutionary factors. Quarterly Review of Biology, 67, 267-281.
Moyer, R.S. \& Landauer, T.K., (1967). Time required for judgments of numerical inequality. Nature, 215, 15191520.

Nieder, A. \& Miller, E.K. (2003). Coding of cognitive magnitude: Compressed scaling of numerical information in the primate prefrontal cortex. Neuron, 37, 149-157.
Noe, R. (2006). Cooperation experiments: coordination through communication versus acting apart together. Animal Behaviour, 71, 1-18.

Opfer, J. E., \& Thompson, C. (2008). The trouble with transfer: Insights from microgenetic changes in the representation of numerical magnitude. Child Development, 79, 788-804.
Opfer, J. E., Thompson, C., \& Furlong, E. (2010). Early development of spatial-numeric associations: Evidence from spatial and quantitative performance of preschoolers. Developmental Science, 13, 761-771.
Peters, E., Vastfjall, D., Slovic, P., Mertz, C.K., Mazzocco, K., \& Dickert, S. (2006). Numeracy and decision making. Psychological Science, 17, 407-413.
Piaget, J. (1932). The Moral Judgment of the Child. New York: The Free Press.
Rapoport, A. \& Chammah, A.M. (1965). Prisoner's Dilemma: A Study in Conflict and Cooperation. Ann Arbor: The University of Michigan Press.
Rousseau, J.J. (2007). A Discourse of on the Origin of Inequality and A Discourse on Political Economy. In G.D.H. Cole (Ed. \& Trans). Stillwell, KS: Digireads.com (Original work published 1754).
Siegler, R.S. \& Booth, J.L. (2004). Development of numerical estimation in young children. Child Development, 75, 428-444.
Sielger, R.S. \& Opfer, J.E. (2003). The development of numerical estimation: Evidence for multiple representations of numerical quantity. Psychological Science, 14, 237-243.
Siegler, R.S., Thompson, C.A. \& Opfer, J.E. (2009). The logarithmic-to-linear shift: One learning sequence, many tasks, many time scales. Mind, Brain \& Education, 3, 143-150.
Starkey, P. \& Cooper, R.G. (1980). Perception of numbers by human infants. Science, 210, 1033-1035.
Stevens, J.S., Cushman, F.A. \& Hauser, M.D. (2005). Evolving the psychological mechanisms for cooperation. Annual Review of Ecological Systems, 36, 499-518.
Thompson, C.A. \& Opfer, J.E. (2010). How 15 hundred is like 15 cherries: Effect of progressive alignment on representational changes in numerical cognition. Child Development, 81, 1768-1786.
Thompson, C.A. \& Siegler, R.S. (2010). Linear numerical magnitude representations aid children's memory for numbers. Psychological Science, 21, 1274-1281.
Trivers, R.L. (1971). The evolution of reciprocal altruism. Quarterly Review of Biology, 46, 35-57.
Valev, I. \& Chater, N. (2006). Game relativity: How context influences strategic decision-making. Journal of Experimental Psychology: Learning, Memory, and Cognition, 32, 131-149.
Warneken, F. \& Tomasello, M. (2006). Altruistic helping in human infants and young chimpanzees. Science, 311, 1301-1303.
Warneken, F. \& Tomasello, M. (2007). Helping and cooperation at 14 months of age. Infancy, 11, 271-294.
Warneken, F., Chen, F. \& Tomasello, M. (2006). Cooperative activities in young children and chimpanzees. Child Development, 77, 640-663.

# A Computational Model of Two Cognitive Transitions Underlying Cultural Evolution 

Liane Gabora (liane.gabora@ubc.ca)<br>University of British Columbia<br>Department of Psychology, Okanagan campus, Arts Building, 3333 University Way<br>Kelowna BC, V1V 1V7, CANADA

Wei Wen Chia (cww9989@gmail.com) and Hadi Firouzi (hadi.firouzi@ubc.ca)<br>University of British Columbia<br>Department of Engineering, 5000-2332 Main Mall<br>Vancouver BC,V6T 1Z4, CANADA


#### Abstract

We tested the computational feasibility of the proposal that open-ended cultural evolution was made possible by two cognitive transitions: (1) onset of the capacity to chain thoughts together, followed by (2) onset of contextual focus (CF): the capacity to shift between a divergent mode of thought conducive to 'breaking out of a rut' and a convergent mode of thought conducive to minor modifications. These transitions were simulated in EVOC, an agent-based model of cultural evolution, in which the fitness of agents' actions increases as agents invent ideas for new actions, and imitate the fittest of their neighbors' actions. Both mean fitness and diversity of actions across the society increased with chaining, and even more so with CF, as hypothesized. CF was only effective when the fitness function changed, which supports its hypothesized role in generating and refining ideas.


Keywords: Agent-based model, CF, convergent though, creativity, cultural evolution, divergent thought, dual process, recursive retrieval, stream of thought.

## Introduction

Humans are unique with respect to the ability to generate accumulative, adaptive cultural evolution, a phenomenon referred to as the ratchet effect (Tomasello, Kruger, \& Ratner, 1993). Gaining insight into the origins of the capacity for complex culture is difficult, since all that is left of our prehistoric ancestors are bones and artifacts such as stone tools that resist the passage of time. Although methods for analyzing these remains are becoming increasingly sophisticated, they cannot always distinguish amongst competing theories. Thus, formal models provide valuable reconstructive tools for testing the feasibility of theories concerning the origins of the cognitive mechanisms that have transformed our planet.

Several cognitive mechanisms have been implicated in the ability to evolve culture. One is the capacity to chain thoughts together to generate a sequence of actions or stream of thought (Donald, 1991). Another is contextual focus (hereafter referred to as CF): the capacity to shift between analytic and associative modes of thought (Gabora, 2003). Mathematical models of both have been developed (Gabora \& Aerts, 2009; Gabora \& Kitto, 2012; Veloz et al., 2011). Incorporating chaining into a computational model of cultural evolution increased the fitness and diversity of cultural outputs, as well as the effectiveness of learning (Gabora \& Saberi, 2011). Incorporating CF into a portrait painting computer program
generated artworks that humans preferred over those generated without CF (DiPaola \& Gabora, 2009). However, the portrait painting program did not allow investigation of the effect of CF on the evolution of ideas through cultural interaction. The goal of the work presented here was to understand the relationship between chaining and CF. Specifically, we investigate the feasibility of the hypothesis that RR is broadly useful for improving cultural outputs, while CF is specifically useful for overcoming a new or sudden challenge.

## Early Signs of Human Creativity

The minds of our earliest ancestors, Homo habilis, are referred to as episodic because there is no evidence that their experience deviated from the present moment of concrete sensory perceptions (Donald, 1991). They encoded perceptions of events in memory, but had little voluntary access to them without cues. They were therefore unable to voluntarily shape, modify, or practice skills and actions, and could not invent or refine complex actions, gestures, or vocalizations.

Homo habilis was eventually replaced by Homo erectus, which lived between approximately 1.8 and 0.3 million years ago. This period is considered the beginning of human cultural evolution. Homo erectus exhibited signs of enhanced intelligence, creativity, and adaptability. They made sophisticated task-specific stone hand axes, had complex stable seasonal home bases, and there is evidence of long-distance hunting strategies involving large game, and migration out of Africa (Leakey 1984). It is widely believed that these early signs of creative culture reflect an underlying transition in cognitive or social abilities. The cranial capacity of the Homo erectus brain was approximately $1,000 \mathrm{cc}$, which is about $25 \%$ larger than that of Homo habilis, and at least twice as large as that of living great apes, and $75 \%$ that of modern humans (Aiello, 1996 ).

Some have suggested that these abilities are due to the onset of a theory of mind (Mithen, 1998) or the capacity to imitate (Dugatkin, 2001). However, there is evidence that nonhuman primates also possess theory of mind (Heyes, 1998) and the capacity to imitate (Dugatkin, 2001), yet their cultural complexity do not compare with humans'. Evolutionary psychologists have suggested that our unique abilities were due to the onset of massive modularity (Barkow, Cosmides,
\& Tooby, 1992). However, although the mind exhibits an intermediate degree of functional and anatomical modularity, neuroscience has not revealed vast numbers of hardwired, encapsulated, task-specific modules; indeed, the brain is more subject to environmental influence than was previously believed (Buller, 2005; Byrne, 2000; Wexler, 2006).

Donald (1991) proposed that with the enlarged cranial capacity of Homo erectus, the human mind underwent a transition characterized by a shift from an episodic to a mimetic mode of cognitive functioning, made possible by onset of the capacity to voluntarily retrieve memories independent of environmental cues and chain them into sequences. Donald refers to the cognitive architecture underlying this capacity as a self-triggered recall and rehearsal loop. It enabled information to be processed recursively, and from different perspectives. Voluntary access to memories made it possible to act out ${ }^{1}$ events that occurred in the past or that might occur in the future. Thus not only could the mimetic mind temporarily escape the here and now, but by miming or gesture it could communicate similar escapes to other minds. The capacity to mime thus brought forth what is referred to as a mimetic form of cognition, and allowed for the onset of culture. The selftriggered recall and rehearsal loop also enabled our ancestors to engage in a stream of thought, in which one thought or idea evokes another, and so forth recursively. In this way, attention can be directed away from the external world toward one's internal model of it. Finally, self-triggered recall allowed for voluntary rehearsal and refinement of actions, enabling systematic evaluation and improvement of skills and motor acts.

## An Explosion of Creative Cultural Change

The European archaeological record indicates that an unparalleled cultural transition occurred between 60,000 and 30,000 years ago, at the onset of the Upper Paleolithic. Considering it "evidence of the modern human mind at work," Leakey (1984:93-94) describes this period as "unlike previous eras, when stasis dominated, ... [with] change being measured in millennia rather than hundreds of millennia." Similarly, Mithen (1998) refers to the Upper Paleaolithic as the 'big bang' of human culture, exhibiting more innovation than in the previous six million years of human evolution. It marks the beginnings of traits considered diagnostic of behavioral modernity, including a more organized, strategic, season-specific style of hunting involving specific animals at specific sites, elaborate burial sites indicative of ritual and religion, evidence of dance, magic, and totemism, colonization of Australia, and replacement of Levallois tool technology by blade cores in the Near East. In Europe, complex hearths and many forms of art appeared, including cave paintings of animals, decorated tools and pottery, bone and antler tools with engraved designs, ivory statues of animals and sea shells, and personal decoration such as beads, pendants, and perforated animal teeth, many of which may have indicated social status.

[^269]Whether this period was a genuine revolution culminating in behavioral modernity is hotly debated because claims to this effect are based on the European Palaeolithic record, and largely exclude the African record (McBrearty \& Brooks, 2000). However the dominant view is that modern behavior appeared in Africa between 40,000 and 50,000 years ago, and spread, resulting in displacement of the Neanderthals in Europe (Klein, 1999). From this point on there was only one hominid species: modern Homo sapien, and despite a lack of overall increase in cranial capacity, their prefrontal cortex, and more particularly the orbitofrontal region, increased significantly in size (Dunbar, 1993). in what was most likely a time of major neural reorganization (Klein, 1999). Given that the Middle/Upper Palaeolithic was a period of unprecedented creativity, what kind of cognitive processes were involved?

It is widely believed that a divergent or associative mode of thought predominates during idea generation, while a convergent or analytic mode predominates during the refinement, implementation, and testing of an idea (Finke, Ward, \& Smith, 1992). It has been proposed that the Paleolithic transition reflects fine-tuning of the biochemical mechanisms underlying the capacity to subconsciously shift between these modes, depending on the situation, by varying the specificity of the activated cognitive receptive field (Gabora, 2003; Gabora Kaufman, 2010). This is referred to as contextual fo$\operatorname{cus}^{2}$ (CF) because it requires the ability to focus or defocus attention in response to the context or situation one is in. Defocused attention, by diffusely activating a broad region of memory, is conducive to divergent thought; it enables obscure (but potentially relevant) aspects of the situation to come into play. Focused attention is conducive to convergent thought; memory activation is constrained enough to hone in and perform logical mental operations on the most clearly relevant aspects.

## The Computational Model

We reviewed the evidence for two hypotheses: (1) the earliest signs of culture were due to the onset of the capacity to chain representations together, and (2) the cultural explosion of the Middle-Upper Paleolithic was due to the onset of CF. We investigated these hypotheses using an agent-based model of cultural evolution referred to as "EVOlution of Culture", abbreviated EVOC. EVOC uses neural network based agents that (1) invent new ideas, (2) imitate actions implemented by neighbors, (3) evaluate ideas, and (4) implement successful ideas as actions. EVOC is an elaboration of Meme and Variations, or MAV (Gabora, 1995), the earliest computer program to our knowledge to model not just cultural transmission but cumulative, adaptive, cultural evolution. ${ }^{3}$ It was inspired by the genetic algorithm, a search technique that finds solutions

[^270]to complex problems by generating a 'population' of candidate solutions through processes akin to mutation, selecting the best, and repeating until a satisfactory solution is found (Holland, 1975). The goal behind MAV, and also behind EVOC, was to distil the underlying logic of not biological evolution but cultural evolution. Agents do not evolve in a biological sense-they neither die nor have offspring-but do in a cultural sense, by adaptively modifying each others' ideas for actions. We summarize the architecture of EVOC in sufficient detail to explain our results; for details we refer the reader to previous publications (e.g., Gabora, 1995; Gabora \& Saberi, 2011; Leijnen \& Gabora, 2009).

## Agents

Agents consist of (1) a neural network, which encodes ideas for actions and detects trends in what constitutes a fit action, (2) a 'perceptual system', which carries out the evaluation and imitation of neighbours' actions, and (3) a body, consisting of six body parts which implement actions. The neural network is composed of six input nodes and six corresponding output nodes that represent concepts of body parts (LEFT ARM, RIGHT ARM, LEFT LEG, RIGHT LEG, HEAD, and HIPS), as well as hidden nodes that represent more abstract concepts (LEFT, RIGHT, ARM, LEG, SYMMETRY, OPPOSITE, and MOVEMENT). Input nodes and output nodes are connected to hidden nodes of which they are instances (e.g., LEFT ARM is connected to LEFT.) Activation of any input node activates the MOVEMENT node. Same-direction activation of symmetrical input nodes (e.g., upward motion-of both arms) activates the SYMMETRY node.

## Invention

An idea for a new action is a pattern consisting of six elements that dictate the placement of the six body parts. Agents generate new actions by modifying their initial action or an action that has been invented previously or acquired through imitation. During invention, the pattern of activation on the output nodes is fed back to the input nodes, and invention is biased according to the activations of the SYMMETRY and MOVEMENT hidden nodes. (Were this not the case there would be no benefit to using a neural network.) To invent a new idea, for each node of the idea currently represented on the input layer of the neural network, the agent makes a probabilistic decision as to whether the position of that body part will change, and if it does, the direction of change is stochastically biased according to the learning rate. If the new idea has a higher fitness than the currently implemented idea, the agent learns and implements the action specified by that idea.

## Imitation

The process of finding a neighbour to imitate works through a form of lazy (non-greedy) search. The imitating agent randomly scans its neighbours, and adopts the first action that is fitter than the action it is currently implementing. If it does not find a neighbour that is executing a fitter action than its own current action, it continues to execute the current action.

Table 1: Definition table.

| Term | Definition | Example |
| :--- | :--- | :--- |
| Body <br> Part | Component of agent <br> other than neural net- <br> work. | Left Arm (LA) |
| Sub- <br> action | Set of six components <br> that indicates position of <br> 6 body parts. Each can be <br> in a neutral (0), up (1), or <br> down (-1) position. | RA:-1, LL:1, LA: <br> RL:0, HP:-1; <br> This sub-action <br> is abbreviated <br> $01-110-1$ |
| Action | One or more sequential <br> sub-actions. | $01001-1, \quad-10-1-$ <br> 111 |
| Template | Abstract or prototypical <br> format for a sub-action. <br> Position of a body part <br> can be unspecified (*). | HD:0, LA:*, <br> RA:1, LL:*, <br> RL:1, HP:-1 |

Table 2: Partial set of the templates used in the first fitness function. (The rest are omitted due to lack of space.)

$$
\begin{array}{|l|l|}
\hline T^{1}=\{0, *, *, *, *, *\} & T^{24}=\{1, *, *, 1,1, *\} \\
\hline T^{2}=\{*, 0, *, *, *, *\} & T^{25}=\{1, *, 1, *, 1, *\} \\
\hline T^{3}=\{*, *, 0, *, *, *\} & T^{26}=\{1, *, 1,1, *, *\} \\
\hline
\end{array}
$$

## Evaluation: The Fitness Function

Fitness was evaluated using an adaptation of the Royal Roads fitness function (Forrest \& Mitchell, 1993). Midway through a run the fitness function was changed to test the effectiveness of chaining and CF for adapting to a sudden change in the task constraints or the environment. Definitions of terms used to accomplish this are provided in Table One.

The first fitness function is determined by 45 templates, six of which are shown in Table Two. The second (not shown) is constructed analogously, with different sub-actions. The templates can be thought of as defining the cultural significance of types of sub-actions (such as dance steps). Each template $T^{i}$ consists of six components, one for each body part (i.e., $T^{i}=t_{j}^{i} ; j=1 . .6$ ). Each body part can be in a neutral position (0), up (1), down (-1), or an unspecified position (*). For example, in template $T^{i}=*, 1,-1, *, *, 0$, the left arm is up (LA:1), the right arm is down (RA:-1), the hips are in the neutral position (HP:0), and the positions of other body parts is unspecified (HD:*, LL:*, and RL:*). The templates provide constraints, as well as flexibility with respect to what constitutes a fit action. For example, in an optimally fit action, the head must be in the neutral position (in $T^{1}$ the first component is 0 ) but the positions of other body parts can vary). The optimal sub-actions are $\{0,1,-1,1,-1,1\},\{0,1,-1,1,-1,-1\}$, $\{0,-1,1,-1,1,1\}$, and $\{0,-1,1,-1,1,-1\}$.

Assume that D is a sub-action (i.e., $D=d_{j} ; j=1 . .6$ ) and $T^{i}$ is the $i^{t h}$ template (i.e., $T^{i}=t_{j}^{i} ; j=1 . .6$ ). Thus, $d_{j}$ represents the position of the $j^{\text {th }}$ body part and the value of $d_{j}$ can be either 0 (neutral), 1 (up), or -1 (down). Likewise, the value of
$t_{j}^{i}$ can be $0,1,-1$, or * (unspecified). Accordingly, the fitness of sub-action $D$ is obtained as per Eq. 1 .

$$
\begin{equation*}
F(D)=\sum_{i=1}^{19} \Phi\left(T^{i}, D\right) \times \Omega\left(T^{i}\right) \tag{1}
\end{equation*}
$$

As shown in Eq. 1, fitness is a function of template weight $\left(\Phi\left(T^{i}, D\right)\right)$ and template order $\left(\Omega\left(T^{i}\right)\right)$.
Template Weight $\Phi\left(T^{i}, D\right)$ is a function that determines the weight of sub-action $D$ by comparing it with template $T^{i}$. This weight is set to one if each component of the sub-action (i.e., $d_{j} ; j=1 . .6$ ) either matches the corresponding component of the template (i.e., $t_{j}^{i} ; j=1 . .6$ ) or if the corresponding components of the template is unspecified (i.e., $t_{j}^{i}=*$ ):

$$
\Phi\left(T^{i}, D\right)=\left\{\begin{array}{cc}
1 & \text { if } \forall t_{j}^{i} \in T^{i}: t_{j}^{i}=d_{j} \text { or } *  \tag{2}\\
0 & \text { otherwise }
\end{array}\right.
$$

Template Order $\Omega\left(T^{i}\right)$ computes the order of the template $T^{i}$ by counting the number of components that have a specified value (i.e., $t_{j}^{i} \neq *$ ).

$$
\begin{equation*}
\Omega\left(T^{i}\right)=\sum_{j=1, t_{j}^{i} \neq *}^{6} t_{j}^{i} \tag{3}
\end{equation*}
$$

The fitness functions are difficult to solve because they are rugged; is to have multiple milestones, or fitness peaks, that agents must achieve before reaching the plateau. For example, consider the fitness function given in Table 2. The action $0,0,0,0,0,0$ has a fitness of 6 . An agent may move on from this action to find an actions that fits the third order templates with a fitness of 31 , e.g., $F(D):\{1,1,1,1,1,0\}=$ $3+3+3+3+3+3+3+3+3+3+1=31$.

## Learning

Invention makes use of the ability to learn, and respond adaptively to trends. Knowledge acquired through the evaluation of actions is translated into educated guesses about how to invent fit actions. For example, an agent may learn that symmetrical movement tends to be either beneficial or detrimental, and bias the generation of new actions accordingly.

## A Typical Run

Fitness and diversity of actions are initially low because all agents are initially immobile, implementing the same action, with all body parts in the neutral position. Soon some agent invents an action that has a higher fitness than immobility, and this action gets imitated, so fitness increases. Fitness increases further as other ideas get invented, assessed, implemented as actions, and spread through imitation. The diversity of actions increases due to the proliferation of new ideas, and then decreases as agents hone in on the fittest actions. Thus, over successive rounds of invention and imitation, the agents' actions improve. EVOC thereby models how "descent with modification" occurs in a purely cultural context.

## Method

## Modeling Chaining

The chaining algorithm is illustrated schematically in Figure 1b. Chaining gives agents the opportunity to execute multistep actions. The agent can keep adding a new sub-action to its current action so long as the most recently-added subaction is both novel and successful. A sub-action D is considered novel if at least one of its components is different from that of the previous sub-action. It is considered successful if there exists a template $T^{i}$ such that $\Phi\left(T^{i}, D\right)$ is one.

$$
\operatorname{successful}(D)=\left\{\begin{array}{lc}
\text { true } & \text { if } \exists T^{i}: \Phi\left(T^{i}, D\right)=1  \tag{4}\\
\text { false } & \text { otherwise }
\end{array}\right.
$$

The fitness of an action consisting of more than one subaction is obtained by adding the number of sub-actions to the fitness of the last sub-action in the sequence. For example, if the last sub-action of an action is $D=[0,1,-1,1,-1,1]$ and the number of sub-actions is seven, the fitness of the action is $F(D)+7=14+7=21$. Thus where $c$ is 'with chaining', $w$ is 'without chaining', n is the number of chained sub-actions, the fitness of a chained action, $F_{c}$, is calculated as follows:

$$
\begin{equation*}
F_{c}=F_{w}+n \tag{5}
\end{equation*}
$$

An agent can execute an arbitrarily long action so long as it continues to invent successful new sub-actions. In general, the more sub-actions the fitter the action. This is admittedly a simple algorithm of simulating the capacity for chaining, but we were not interested in the impact of this action per se. The goal here was simply to test hypotheses about how chaining at the individual level affects dynamics at the societal level, by providing agents with a means of implementing multistep actions such that the optimal way of going about one step depends on how one went about the previous step.

## Modeling Contextual Focus

The CF algorithm is illustrated schematically in Figure 1c. In the convergent mode, the current action is only slightly modified to create a new action. In the divergent mode, the current action is substantial modified to create a new action. An agent switches between these modes by modifying its rate of creative change ( RCC ). If the fitness of its current action is low relative to previous actions, RCC increases, causing the agent to shift to a more divergent processing mode conducive to large leaps through the space of possibilities. If action fitness is high relative to that of previous actions, RCC decreases, and the agent shifts to a more convergent mode conducive to minor adjustments. With CF turned off, RCC stays constant throughout the run at $1 / 6$ (i.e., a new action involves change to one of the six body parts). The equation to modify RCC is shown in Eq. 6 where $a$ is a negative value. Since at the start of a run previous fitness is undefined, RCC in this case is a function of the current fitness as per Eq. 7, where $0<b<1$.

$$
\begin{equation*}
\Delta R C C=a\left(F_{\text {new }}-F_{\text {old }}\right) \tag{6}
\end{equation*}
$$



Figure 1: Schematic illustration of (a) neither chaining nor CF, (b) chaining only, and (c) both. Chaining operates within a generation whereas CF operates between generations.

$$
\begin{equation*}
R C C_{\text {initial }}=b^{F_{\text {current }}} \tag{7}
\end{equation*}
$$

In the results shown here $a$ and $b$ were initialized to -0.005 and 0.8 respectively.

## Results

The results of introducing chaining and CF on the mean fitness and diversity (total number of different actions) of actions across all agents in the society are shown in Figures 2 and 3 respectively. All graphs show means of 500 runs. Chaining and CF both significantly increased mean fitness of actions. Without chaining, mean fitness quickly reached a plateau; with chaining it could increase indefinitely. While chaining increased mean fitness throughout the run, CF only increased mean fitness following initial exposure to a new fitness function, i.e., at the beginning of the run, and when the second fitness function was introduced at iteration 50.

Chaining also significantly increased the diversity of actions. Although inspection reveals that there is always convergence on optimal actions, without chained actions, this set is a static (thus mean fitness plateaus) whereas with chained actions the set of optimal actions changes, as increasingly fit actions continue to be found. When agents were first exposed to a fitness function, CF increased both the rate at which new possibilities were generated, and the rate of convergence on the fittest of these, although this effect is more pronounced for the first fitness function than the second. As with fitness, CF exerted no noticeable effect on diversity once the agents had fit actions.


Figure 2: Mean fitness of cultural outputs across the society with both chaining and CF (red line), chaining only (dashed blue line), and neither chaining nor CF (dotted green line).


Figure 3: Diversity of cultural outputs across the society with both chaining and CF (red line), chaining only (dashed blue line), and neither chaining nor CF (dotted green line).

## Discussion

This paper provides valuable insights into the mechanisms underlying the uniquely human capacity for collectively generated, open-ended, adaptive cultural evolution. Our results suggest that once humans became able to sequence thoughts together to generate increasingly complex and refined cultural outputs, they would have found themselves at a significant adaptive advantage. Similarly, our results suggest that once humans became able to employ an exploratory, divergent processing mode when stuck, followed by a shift to a more constrained convergent processing mode to fine-tune their cultural outputs, they would have been capable of generating significantly more valuable cultural outputs. We suggest that a mechanism akin to CF is what makes possible the cumulative creativity exhibited by successful computational models of language evolution (e.g., Kirby, 2001). A potential downfall of processing in a divergent mode is that since effort is devoted to the re-processing of previously learned material, less effort may be devoted to being on the lookout for danger and simply carrying out practical tasks. Since divergent thought carries a high cognitive load, it would not have been useful to think divergently until there was a means to shift back to a convergent mode. Although these results do not prove that onset of the capacity to chain thoughts together into sequences, and to shift between divergent modes of thought through CF, are responsible for our cultural complexity, it shows that they provide a computationally feasible
explanation. We know of no other cognitive mechanisms implicated in the evolution of complex culture for which openended, adaptive cultural change has been demonstrated.

Both chaining and CF were implemented in a simple manner. Future investigations will focus on developing more realistic implementations of chaining and CF. Chaining will use associative recall to reconsider an item from multiple potentially relevant 'perspectives', and the divergent mode of CF will use a sophisticated mathematical model of concepts to facilitate the generation of new concept combinations.

## Acknowledgments

This project was conducted at the Media and Graphics Interdisciplinary Centre (MAGIC), UBC. It was supported by the Natural Sciences and Engineering Research Council of Canada, and Fund for Scientific Research, Flanders, Belgium.

## References

Aiello, L. C. (1996). Hominine pre-adaptations for language and cognition. In P. Mellars \& K. Gibson (Eds.), Modeling the early human mind, Cambridge, UK: McDonald Institute Monographs, 89-99.
Barkow, J. H., Cosmides, L., \& Tooby, J., Eds. (1992). The adapted mind New York: Oxford University Press.
Buller, D. J. (2005). Adapting minds. MIT Press.
Byrne, R., \& Russon, A. (1998). Learning by imitation: A hierarchical approach. Behavioral and Brain Sciences, 2, 667-721.
Cavalli-Sforza, L. L., \& Feldman, M. W. (1981). Cultural transmission and evolution: A quantitative approach. Princeton: Princeton University Press.
DiPaola, S. \& Gabora, L. (2009). Incorporating characteristics of human creativity into an evolutionary art algorithm. Genetic Programming and Evolvable Machines, 10, 97110.

Donald, M. (1991). Origins of the modern mind: Three stages in the evolution of culture and cognition. Cambridge, MA: Harvard University Press.
Dugatkin, L. (2001). Imitation factor: Imitation in animals and the origin of human culture. New York: Free Press.
Dunbar, R. (1993). Coevolution of neocortical size, group size, and language in humans. Behavioral and Brain Sciences, 16, 681-735.
Finke, R. A., Ward, T. B., \& Smith, S. M. (1992). Creative cognition: Theory, research, and applications. Cambridge, MA: MIT Press.
Forrest, S. \& Mitchell, M. (1993). Relative building block fitness and the building block hypothesis. In (L. Whitley, Ed.) Foundations of genetic algorithms. San Mateo: Morgan Kaufman.
Gabora, L. (1995). Meme and Variations: A computer model of cultural evolution. In L. Nadel \& D. Stein (Eds.) 1993 Lectures in Complex Systems. Boston: AddisonWesley.

Gabora, L. (2003). Contextual focus: A cognitive explanation for the cultural transition of the Middle/Upper Paleolithic. Proceedings Annual Conference Cognitive Science Society (pp. 432-437), Boston: Lawrence Erlbaum.
Gabora, L. \& Aerts, D. (2009). A mathematical model of the emergence of an integrated worldview. Journal of Mathematical Psychology, 53, 434-451.
Gabora, L., \& Kaufman, S. (2010). Evolutionary perspectives on creativity. In (J. Kaufman \& R. Sternberg, Eds.) The Cambridge handbook of creativity. Cambridge: Cambridge University Press.
Gabora, L., \& Kitto, K. (2012). Concept combination and the origins of complex cognition. In (L. Swan, Ed.) Origins of mind. Berlin: Springer.
Gabora, L. \& Saberi, M. (2011). How did human creativity arise? An agent-based model of the origin of cumulative open-ended cultural evolution. Proceedings Conference on Cognition \& Creativity (pp. 299-306). New York: ACM.
Higgs, P. (2000). The mimetic transition: a simulation study of the evolution of learning by imitation. Proceedings: Royal Society B: Biological Sciences, 267, 1355-1361.
Heyes, C. M. (1998). Theory of mind in nonhuman primates. Behavioral and Brain Sciences, 211, 104-134.
Hinton, G. E. \& Nowlan, S. J. (1987). How learning can guide evolution. Complex Systems, 1, 495-502.
Holland, J. K. (1975). Adaptation in natural and artificial systems. Ann Arbor, MI: University of Michigan Press.
Hutchins, E. \& Hazelhurst, B. (1991). Learning in the cultural process. In Langton, C., Farmer, D., \& Rasmussen, S. (Eds.) Artificial Life II. Redwood City: Addison-Wesley.
Kirby, S. (2001). Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity. IEEE Transactions on Evolutionary Computation, 5(2), 102-110.
Klein, R. (1999). The human career: Human biological and cultural origins. Chicago: University of Chicago Press.
Leijnen, S., \& Gabora, L. (2009). How creative should creators be to optimize the evolution of ideas? A computational model. Electronic Proceedings in Theoretical Computer Science, 9, 108-119.
Leakey, R. (1984). The origins of humankind. New York: Science Masters Basic Books.
McBrearty, S., \& Brooks, A. S. (2000). The revolution that wasnt: A new interpretation of the origin of modern human behavior. Journal of Human Evolution, 39, 453-563.
Mithen, S. (1998). Creativity in human evolution and prehistory. London, UK: Routledge.
Tomasello, M., Kruger, A., \& Ratner, H. (1993). Cultural learning. Behavioral and Brain Sciences, 16, 495-552.
Veloz, T., Gabora, L., Eyjolfson, M. \& Aerts, D. (2011). A model of the shifting relationship between concepts and contexts in different modes of thought. Proc 5th Intnl Symposium on Quantum Interaction. Heidelberg: Springer.
Wexler, B. (2006). Brain and culture: Neurobiology, ideology and social change. New York: Bradford Books.

# Completion in the Wild: Perception of 3D forms from cross-sections 

## Kristin M. Gagnier (kristin.gagnier@temple.edu)

Spatial Intelligence and Learning Center, Department of Psychology, 1701 N. $13^{\text {th }}$ Street Philadelphia, PA 19122 USA

Thomas F. Shipley (Shipley@temple.edu)<br>Spatial Intelligence and Learning Center, Department of Psychology, 1701 N. $13^{\text {th }}$ Street Philadelphia, PA 19122 USA


#### Abstract

Under conditions where an object is inside another object and only a single face is visible, is there a bias to assume smooth continuation of the surface straight back into the object? To examine the ability to estimate how features progress into a volume, participants viewed 16 pictures of everyday objects (rocks, food, wood) presented with only a single face visible (see Figure 1). Participants reported whether a highlighted region of a picture was present on the surface or extended into the object. If they perceived the region as extending in, they positioned a rod to indicate the angle. Surface responses were rare and instead participants' readily perceived 3D forms from 2D views. Inspection of frequency histograms revealed a systematic bias to estimate the angle of extension in the $80-110^{\circ}$ range. This type of completion process suggests constraints on models of visual completion and has implications for STEM education, in particular, how students deal with ambiguity.


Keywords: perception of 3d volumes, amodal continuation, penetrative thinking.

## Introduction

How does the visual system estimate 3-dimensional (3D) forms of objects embedded in other objects? While this question has received very little empirical attention (Chariker, Naaz, \& Pani, 2011; Hegarty, Keehner, Cohen, Montello, \& Lippa, 2007), we believe it is central to our understanding of amodal completion. Take for example the image shown in Figure 1. There is little doubt that the dark brown region (cinnamon swirl) is 3D and extends into the object, however from this 2-dimensional (2D) view or crosssection, the 3D shape is unknowable (note even if you knew the true 3D shape the answer would still be ambiguous within a mirror reflection because the image could come from either side of a cut). In order to infer the 3D shape of the region, one would have to have a view of the region from another side. Inferring shape from partial information is particularly important for disciplines that rely on 3D visualization (e.g., astronomy, neuroscience, geosciences). However, a common sentiment echoed by geoscientists and noted by Kali \& Orion (1996) for geological stimuli, is that students neglect the ambiguity inherent in a single 2D view and instead are biased to assume that surface boundaries extend perpendicularly into solids.

The goal of this paper was twofold: 1) to examine whether people recognize that the 3D form of an object is unknowable from a single 2D view and 2) to examine whether participants do indeed exhibit a bias to assume smooth continuation of a surface straight back into the solid as suggested by the anecdotal reports of geoscience educators.
To answer this question we showed undergraduate psychology majors pictures of everyday objects such as food, rocks and wood. For each picture, a region was indicated with a colored line, as shown in Figure 1 top. Students indicated if the highlighted region was visible only on the surface or whether it extended into the object. Note that the answer to this question is in fact unknowable. While some of the objects are familiar to viewers (kiwi, bread, etc) and thus the overall shape of the object can be inferred, one cannot know how the cut was made or for unfamiliar objects, whether the indicated region continues into the object or is present only on the surface. To infer the 3D shape one would have to see more than one 2D crosssection of the object. If students saw the indicated region as extending in, they used a rod attached to an inclinometer to indicate the angle at which the region continued into the object (the correct answer for the bread is shown in Figure 1 bottom). We predicted that participants would have a strong sense that the indicated regions were 3D and would exhibit a bias to see the regions as extending back at a $90^{\circ}$ angle relative to the ground surface.


Figure 1: The bread stimulus. Top: Participants indicated if the region highlighted by the red line, was present on the surface or extend into the solid. Bottom: Red line shows the angle at which the swirl extends into the bread. Note participants never saw this view.

## Method

Participants. Participants were 30 Temple University Undergraduates.

Stimuli. The stimuli consisted of 17 color photographs of common objects such as food, wood and rocks. There was one practice image and 16 experimental images. For each picture we selected a specific region of the picture to ask participants about. As shown in Figure 1, this region was indicated with a red line. In each image the area represented a region where a plane might intersect the visible surface. For example in Figure 1, the plane defined by the cinnamon layer between two regions of dough is indicated. Images when presented on the screen were approximately $25 \times 18$ cm.

These categories of images were chosen with two constraints: 1) that we were physically able to slice each object and measure the angle at which each highlighted region extended into the object and 2) that we sample a range of objects that might be familiar to participants.

Our stimuli fell into six broad categories defined by their internal structure: 1) rocks (granite slab), 2) wood (tree ring and a knot), 3) fruits (pineapple, papaya, kiwi), 4) vegetable (onion), 5) animals (fish and beef) and 6) food products that were originally liquid and are now solid (blue cheese, chocolate with almonds and cinnamon bread). These categories were selected because the internal structure ranged from highly structured and constrained by the environment (e.g., wood grain) to relatively unconstrained (e.g., minerals in rock) and thus the orientation is either knowable within a certain range or completely unknowable.

For example, the internal structure of wood is constrained by the environment. As tree structures are generally concentric cylinders, the extension into a slice is a function of the angle of the cut relative to the cylinders. For the fruit stimuli we selected fruits with radial symmetry and thus the internal structure is also structured. An onion although somewhat irregular in shape, has an organized internal structure. For the fish and beef stimuli we asked about how regions of fat extend in. Fat deposits are structured in complex ways by the surrounding muscles and thus organized but not to the same degree as wood, or fruit. The internal composition of rocks can be structured, but the orientation of a mineral's surface relative to the cutting plane is essentially arbitrary.

Apparatus. Stimuli were presented on a 20 -inch Dell monitor. As shown in Figure 2, the monitor was positioned parallel to the ground.


Figure 2: The display used in the experiment. Participants used the black rod to indicate the orientation at which the highlighted region extended into the object.

Procedure. Participants were tested individually. They viewed each picture while standing with their nose over the center of the monitor. Participants were told that we were interested in their opinions of how regions of images continue in 3 -dimensions. They were told that sometimes they would see pictures where they might have a strong sense that a region continued and sometimes they might have a sense that something was present only on the surface. To illustrate these cases, students were shown a picture of a Swiss roll and crayon marks on paper. All students reported seeing the layers of the Swiss roll as extending into the object while the crayon marks were present on the surface.

Participants were shown 16 pictures. For each, their task was to indicate whether the region indicated with the red line was present only on the surface or extended into the object. If they thought it extended into the object, they used a stainless steel rod with an inclinometer (angle measure) attached to indicate the orientation of continuation. Participants placed the edge of the rod on the red line and then moved the rod up and down to indicate the angle. The $0^{\circ}$ was defined relative to the ground plane (i.e. if positioned the rod to indicate straight down, as shown in Figure 2, the angle measurement was $90^{\circ}$ ). After they estimated the angle, they reported their confidence in their response on a 5-point scale. Prior to viewing the 16 pictures participants practiced using the angle measurement device on the image of the Swiss roll. To be sure that there were not differences in the estimates based on the orientation of the picture, after viewing all the pictures and making their responses participants were shown the 16 pictures again but this time the images were rotated 180 degrees. This allowed us to calculate any bias due to their body position relative to the image. Finally, participants were shown the pictures a third time and asked to identify each picture. For any response with a confidence rating of 0 or 1 , we further probed their uncertainty. Participants were asked to select which of the following reasons best described why they were uncertain in their response: 1) they have no idea what it could be, 2) the answer is unknowable 3) there could have been a range of possible angles. Additionally, for the pictures that they
selected surface for, we told them that the region did extend in and asked them if they could make a guess about its orientation using the rod (note these estimates were not included in any analyses). After this, participants completed the Geologic Block Slicing Test (a measure of inferring internal spatial structures from views of multiple sides; Ormand et al, 2011), Spatial Orientation Test (Kozhevnikov \& Hegarty, 2001) and the Mental Rotations Test (Shepard \& Metzler, 1971). Data for these spatial tests are not presented in this paper.

## Results

Although the single 2D view is insufficient to define a 3D shape, participants reported that the answer was unknowable on only $1 \%$ of the trials (12 times out of 960 trials), suggesting that participants do not recognize the need for multiple views to solve the intersection of constraints problem. Consistent with this, participants were confident in their angle estimates. The mean confidence was 3.2 (SD $=1.1$ ) on a 5-point scale. Participants did perceive some of the highlighted regions as being only on the surface, but this was the case on only $26 \%$ of the trials, suggesting that participants tended to perceive the highlighted regions as extending into the object in three dimensions.

In order to calculate the participant's unbiased estimate for each picture, the two estimates were combined by calculating the average of the first estimate and $180^{\circ}$ minus the second estimate. By presenting each picture in two orientations, we could remove any bias that a participant had to orient the rod towards (or away from) their body. For example, consider a case where the estimate for the first view and the second view (when the picture was rotated $180^{\circ}$ ) of a picture was $80^{\circ}$. If the participant were truly responding with an unbiased estimate, this would mean that if the estimate for the first view was $80^{\circ}$, the estimate for the second view should have been $100^{\circ}$. Thus, by subtracting the second estimate from 180, we can avoid any systematic bias (overall participants exhibited an $\sim 4$ degree bias towards their body).

Figure 3 shows the mean angle estimate without bias for each picture along with the $95 \%$ confidence interval for that picture. Inspection of the figure reveals that mean estimates tended to be biased towards $90^{\circ}$. Fifteen of the 16 pictures have mean estimates that are not significantly different from $90^{\circ}$ (the red line denotes $90^{\circ}$ ). The only picture that has a mean estimate significantly greater than $90^{\circ}$ was the "onion" picture.


Figure 3: Mean estimate for each picture. Error bars show $95 \%$ confidence interval for each mean.

Next we examine whether the distribution of responses for each picture was random or whether there was a preferred direction for the distribution. Surface responses were excluded from this analysis. Thus the N for each picture varies based on the number of estimates for that picture. To examine the structure of the distributions we conducted the Rayleigh test (Zar, 1999) using the EZ Rose program (Baas, 2000). The null hypothesis is that the distribution of responses are randomly distributed around a semicircle. If the null is rejected then the distribution of responses has a preferred direction. To examine whether the null is accepted or rejected one compares R (mean vector length) for each picture to the critical value of the test statistic $R_{0.05}$ (see Baas, 2000 equation 10). If $R$ is greater than $\mathrm{R}_{0.05}$ then the distribution of responses has a preferred direction (i.e. is not random). As can be seen in Table 1 the Rayleigh test was rejected for 14 of the 16 pictures. For the "Tree Knot" and Papaya pictures, participants tended to perceive the region as on the surface ( $44 \%$ and $58 \%$ were judged to be "on surface," respectively). Thus the number of estimates was less than the 15 recommend for this test (Baas, 2000). However, an inspection of the frequency distributions (see below) reveals that when estimates were made, they centered around $90^{\circ}$.

## Table 1: Results of the Rayleigh Test.

| Picture | R | $\mathrm{R}_{0.05}$ | $\mathrm{H}_{0}$ |
| :---: | :---: | :---: | :---: |
| Chocolate | 0.91 | 0.38 | rejected |
| Tree Knot | - | - | - |
| Pineapple | 0.97 | 0.43 | rejected |
| Blue Cheese | 0.94 | 0.43 | rejected |
| Salmon | 0.87 | 0.32 | rejected |
| Granite 4 | 0.84 | 0.38 | rejected |
| Wood 1 | 0.97 | 0.45 | rejected |
| Steak 2 | 0.82 | 0.33 | rejected |
| Bread | 0.92 | 0.33 | rejected |
| Kiwi | 0.85 | 0.33 | rejected |
| Steak 1 | 0.80 | 0.36 | rejected |
| Granite 3 | 0.79 | 0.37 | rejected |
| Granite 2 | 0.85 | 0.39 | rejected |
| Granite 1 | 0.85 | 0.43 | rejected |
| Onion | 0.89 | 0.37 | rejected |
| Papaya | - | - | - |

Finally, we examined whether the mean estimates centered around $90^{\circ}$ because some participants estimated the angle at $10^{\circ}$ and others at $170^{\circ}$ and this averaged out to $90^{\circ}$ or whether there was consistency among estimates for all participants. Figure 4 shows the frequency distribution for response for each of the 16 pictures grouped by category. As can be seen in the figure, the distributions are fairly uniform. There are certain pictures, for example the papaya, where participants on averaged perceived the indicated region to be on the surface (thus number of estimates is smaller). However, there were other pictures, like the kiwi and bread where participants agreed the region extended in. Also evident in the distributions is the limited spread. Mean estimates did not encompass the entire $0-180^{\circ}$ spectrum; instead on average they were concentrated around 90.


Wood 2


Papaya


Pineapple



Steak 1


Steak 2




Granite 1


Granite 2


Granite 3


Granite 4


Figure 4: The frequency of surface response and estimates that fall within a $10^{\circ}$ bin from $0-180^{\circ}$ for each picture.

## Discussion

Two conclusions can be drawn from this work. The first is that participants failed to recognize that one cannot know the orientation of a 3D structure from a single 2D view. Instead of recognizing that this situation is ambiguous, participants tended to have consistent intuitions about how regions of pictures extend in 3D into the object. The second conclusion that can be drawn from these findings is that participant's estimates tended to be clustered around $90^{\circ}$, suggesting that estimates were biased to assume that the surface continues straight back into the object. These findings 1) suggest possible constraints on models of amodal completion processes and 2) have implications for STEM education, in particular, challenges that arise when students do not recognize ambiguity. We consider each of these in turn.

In developing our understanding of how the visual system estimates the 3D form of objects, there are important completion phenomena that should constrain computational models. Here we describe a completion phenomenon that to our knowledge has not been recognized -- participants readily perceived 3D forms from 2D views (of both familiar and unfamiliar objects) and the perception is that surface boundaries extend perpendicularly into a solid.

This bias may be informative about the filling out process that occur under conditions where the 3D completion of surfaces can not occur because the edges of an object are not aligned in the 2D projection on the retina (see Tse, 1999). The bias evident in this study may be a product of the way the visual system handles a more common instance of having partial information about an object - when viewed head-on. Normally, if one can only see a single side of an object, it is a result of your current viewpoint (the line-ofsight is perpendicular to the front face). So, the visual system may represent the portions of the occluded object using past experience (i.e. knowledge about cinnamon swirl bread) or some properties of the front face. This process becomes evident when the object (the cinnamon swirl) is surrounded by an opaque region (more bread) and thus other sides are not visible to the observer. The visual system is not flummoxed by this situation but instead rapidly extrapolates from the available single surface to represent an extended 3D structure. Under these conditions, the completion process reflects the assumption that edges on the surface project straight into the object.

Our observations suggest the existence of a visual process that uses available visual information to extend form representations into regions where the form is not visible. Models of visual completion argue that completion processes reflect the system's attempt to construct a representation of the most likely 3D form. An ongoing debate in the literature has examined whether completion processes occur as a result of extrapolation (filling out) or interpolation (filling in). Here, where only one face of a 3D
form is visible, the completion processes must be based on extrapolation - filling out from available visual information (Shipley \& Kellman, 2003) - rather than interpolation between defined regions. Extrapolation occurs in both amodal (Kanizsa, 1979) and modal displays (Shipley \& Kellman, 2003), but previous demonstrations have been extrapolation of planes or edges, not 3D volumes.

In addition to informing our understanding of visual completion processes, these results have important implications for STEM education. Consider a field geology student examining a rock face and trying to make inferences about its 3D structure, or an anatomy student learning where best to make an incision during dissection or surgery. Both of these tasks require inferences about a 3D form from a 2D view. What is critical in both these cases is that an accurate estimate of the orientation of the 3D form requires more information - either in the form of looking at another angle of the rock to see how features penetrate in, or knowing something about the true 3D shape and using that to constrain the estimate of orientation of how the region extends into the volume. There are aspects of the world that might place constraints on the probable internal structure (i.e. grains in wood has a cylindrical structure), however students must recognize the need to seek out additional information in order to make inferences about the 3D structure.

An extensive body of research has examined decisionmaking regarding uncertain events (Kahnamen, \& Tversky 1982), however to our knowledge work in this area has not examined uncertain perceptual situations and their relationships to confusion in the classroom. We believe this is an interesting area to pursue in future research. How best to convey to students that information in an image may appear to determinant, but is in fact ambiguous.

Our aim in this paper is threefold: first, examine 3D completion from 2D views and to bring this type of process to the attention of the research community. By making researchers aware of completion processes in the "wild", we hope to begin a dialogue that may move completion research forward into new domains. Second, to expand the phenomena considered by any model of visual completion. Third, we wish to illustrate the importance of scaffolding student's ability to recognize ambiguity and the necessity to seek out additional information for solving a problem.

## References

Baas, J. (2000) EZ-ROSE a computer program for equal area circular histograms and statistical analysis of two dimensional vectorial data. Comput. Geosci., 26, 153166.

Chariker, J. H., Naaz, F., \& Pani, J. R. (2011). Computerbased learning of neuroanatomy: A longitudinal study of learning, transfer, and retention. Journal of educational psychology, 103(1), 19.

Hegarty, M., Keehner, M., Cohen, C., Montello, D. R., \& Lippa, Y. (2007). The role of spatial cognition in medicine: Applications for selecting and training professionals. In G. Allen (Ed.) Applied Spatial Cognition. Mahwah, NJ: Lawrence Erlbaum Associates.

Kanizsa, G. (1979). Organization in Vision. New York: Praeger.

Kali, Y. \& Orion, N. (1996). Spatial abilities of high-school students in the perception of geologic structures. Journal of Research in Science Teaching, 33, 369-391.

Kozhevnikov, M. \& Hegarty, M. (2001). A dissociation between object-manipulation and perspective-taking spatial abilities. Memory \& Cognition, 29, 745-756.

Ormand, C., Shipley, T. F, Manduca, C, and Tikoff, B (2011). Improving Geoscience Students' Spatial Thinking Skills: Applying Cognitive Science Research in the Classroom. American Geophysical Union Fall meeting (San Francisco, CA).

Shipley, T. F., \& Kellman, P. J. (2003). Boundary completion in illusory contours: Interpolation or extrapolation? Perception, 32, 985-999.

Shepard, R and Metzler. J. "Mental rotation of three dimensional objects." Science 1971. 171(972):701-3.

Tse, P. U. (1999). Volume completion. Cognitive Psychology, 39, 37-68.

Zar, J. H. (1998). Biostatistical analysis (4th ed.). Upper Saddle River, NJ: Prentice Hall.

# Computational exploration of task and attention modulation on holistic processing and left side bias effects in face recognition: the case of face drawing experts. 

Bruno Galmar (brunogal@hku.hk)<br>Janet Hui-wen Hsiao (jhsiao@hku.hk)<br>Department of Psychology, University of Hong Kong<br>Pokfulam Road, Hong Kong SAR


#### Abstract

Drawing artists and non-drawers are like any adult both experts at face recognition. Yet, artists have a richer learning experience with faces: they were trained in rapid sketching of faces. Zhou, Cheng, Zhang and Wong (2011) found that drawing experts showed less holistic processing (HP) for face recognition than non-drawers. Using a computational model of face recognition that did not implement motor processing, we examined whether engagement of local attention and nature of the learning task could account for the reduced HP in drawers without the influence from motor experience. We showed that compared with the non-drawer model that had a global face input (i.e., Hsiao, Shieh \& Cottrell, 2008), a drawer model that incorporated both global face and local facial parts (eyes and mouth) in the input showed reduced HP, suggesting the modulation of local attention engagement. In contrast, the other drawer model that used only global face input but learned to perform an additional face part identification task did not show the reduced HP effect. In addition, both drawer models demonstrated stronger left side (right hemisphere) bias than the non-drawer model. Our data thus suggest that engagement of local attention is sufficient to account for the reduced HP in drawers, and that HP and left side bias effects can be differentially modulated by visual attention or task requirements.


Keywords: Model of face recognition; Holistic processing; Hemispheric lateralization; Visual expertise.

## Introduction

Visual expertise in subordinate-level discrimination has been extensively studied (e.g., Bukach, Gauthier, \& Tarr, 2006), such as our expertise in recognizing individual faces. Several behavioral markers of visual expertise have been identified, including holistic processing (HP), which refers to the phenomenon of viewing faces as a whole instead of various parts (Bukach et al., 2006; although some argue that HP is specific to face recognition; e.g., McKone, Kanwisher, \& Duchaine, 2007). Subsequent studies suggest a correlation between an increase in HP and expertise in subordinate-level individualization, as opposed to expertise in basic-level categorization (e.g., Wong, Palmeri, \& Gauthier (2009)). For example, Wong et al. (2009) trained two participant groups to recognize an artificial object type (Ziggerins) with different training tasks: one group learned to rapidly individualize Ziggerins at the subordinate level, whereas the other group learned rapid sequential categorization at the basic level. The results showed that only the individuation experts showed an increase in HP, even though the two groups had the same amount of
exposure to Ziggerins. This suggests that qualitatively different expertise processing can arise depending on the nature of the training task.

Such a qualitative difference of expertise processing resulting from different learning and training experience has been recently observed for face recognition. Zhou, Cheng, Zhang and Wong (2011) studied two groups: (a) an experimental group was composed of art students who had extensive formal training in sketching and drawing portraits, and (b) a control student group of non-drawers - i.e. who had no prior drawing background or education-. Hence, the two groups had different learning experience in processing faces. Non-drawers would show the typical face expertise any adult is endowed with: being able to recognize at least a thousand of faces. In contrast, art students would have internally assimilated an ordered procedure for rendering faces on a 2D surface (Balas \& Sinha, 2007; Willenbrink \& Willenbrink, 2012), for example: a) sketch the basic head proportion, b) sketch the overall head form and basic lines for features, c) place the brows and lips, and so on. Such a fine-grained procedure relies upon a mix of global and local processing, and featural and configural processing. Art students would not ignore face details which are critical to render a vivid portrait of an individual. Hence, art students are used to scrutinize a face and could be less engaged in HP than non-drawers. This educative guess is supported by eye-tracking studies (Miall \& Tchalenko, 2001; Tchalenko et al. 2003) of eye movements of a skilled artist. Miall and Tchalenko (2001) proposed as an account of the visual encoding of the studied artist Ho: "The capture of visual information detail by detail, rather than in a more holistic manner, is reflected in the way the drawing or painting is built up. Each detail and each element is of intrinsic importance." Using the complete composite paradigm of face recognition, Zhou et al. (2011) found less HP for art students than for non-drawers. Reduced HP with drawing expertise is not an isolate case. Previously, Hsiao and Cottrell (2009) found reduced HP for Chinese readers - who were experts at recognizing Chinese characters - compared with novice Chinese readers. Tso, Au, and Hsiao (2011) further showed that the reduction in HP found in expert Chinese readers depended on their writing rather than reading experience of Chinese characters, since proficient readers who had limited writing experience (i.e. Limitedwriters) showed increased HP as compared with novices, in contrast to the reduced HP observed in Chinese readers who could read and write fluently (i.e., Writers; Tso, 2012).

In the present study, we aimed to examine the underlying mechanism accounting for the results in Zhou et al. (2011) through computational modeling and simulations. Computational modeling is an insightful tool to test ideas on the nature of cognition difficult to test with human subjects (McClelland, 2009). Motor experience, visual attention, and nature of the learning task are all potential factors that may account for drawers' reduced HP in face recognition. These factors may be difficult to disentangle within drawers so that the separate contribution of each to HP is not easily amenable to experimental study. Here, we aimed at testing two simplified models of drawing expertise that did not implement motor processing and to compare them with our previous model of face recognition (i.e., the intermediate convergence model in Hsiao, Shieh, \& Cottrell, 2008), which is to serve as a non-drawer model, in order to examine whether visual attention and nature of the learning task can account for the reduced HP in drawers without the influence from motor experience. Through these two models, we postulated two hypotheses concerning how art students having developed expertise in the task of drawing faces could demonstrate reduced HP in face recognition compared with non-drawers.

The non-drawer model - called base model thereafter shown in Figure 1 is trained to map face images to whole face identity. This global task is intended to reflect ordinary face recognition by non-drawers. The models of drawing expertise are not as purely global as the base model. They embed local processing in addition to the global face identification.

## Rationale behind the first model of drawing expertise

Our first model of drawing expertise shown in Figure 2 is trained to map face, eyes and mouth images to whole face identity. Modeling the encoding of visual information from facial parts such as eyes and mouth to serve the task of whole face identity reflects the engagement by artists in local attention. Using eye-tracking, Tchalenko, DempereMarco, Hu, and Yang (2003) reported that artists do process individually facial parts and even scrutinize faces for informative details: "[...] the experienced painter differed from the novice in his ability to repeatedly target saccades onto a small detail of the model's face, and to lock on to that detail in a steady fixation." Consistently, Zhou et al. (2011) showed that artists had slower response times (RT) compared with non-drawers. This could be because of the additional engagement of local attention on facial parts. The nature of this more local and prolonged visual engagement is translated in the first model of drawing expertise by a larger input layer compared with the base model. A drawing expert may manipulate more encoded visual inputs - as suggested by the expertise literature (Bransford, 2000) - but would still perform the same global identification task than the normal face recognizer. Because of the selective encoding of eyes and mouth in addition of global encoding of the face image, this model reflects engagement of both global and local attention at the encoding stage of visual processing.

## Rationale behind the second model of drawing expertise

Our second model of drawing expertise shown in Figure 3 is trained to map face images to both whole face identity and cluster identities for mouth and eyes. Hence, the rationale is that artists use the same global attentional resources - i.e. the model has the same global input layer as the base model- but artists engage in a more analytical face recognition task. Here, given a face input, the model tries to recognize in addition to face identity, a mouth prototype (a kind of mouth) and a pair of eyes prototype (a kind of eyes). Such partitioning of eyes and mouth in kinds reflects that artists would engage in clustering facial features. This hypothesis is not only sound but also well-grounded. In his Treatise on Painting, the Renaissance genius Leornardo Da Vinci exposes some technical insights on how to develop the skills necessary to a portraitist (Rigaud, 1877). For example, in the section of "How to remember the Form of a Face", Da Vinci mentioned: "If you wish to retain with facility the general look of a face, you must first learn how to draw well several faces, mouths, eyes, noses, chins, [...], all those principal parts which distinguish one man from another." Then, we read: "[...] noses are of ten different sorts: straight, bunched, concave, [...]." In another section entitled "Observations on drawing Portraits", we read: "The uniting of the nose with the brows is in two ways [...]. The forehead has three different forms."

Details on the implementation of these models are given in the next section. We trained the three models to either the same performance level in the whole face identification task or the same amount of epochs, and examined their difference in HP and lateralization. Face processing has been shown to involve right hemisphere (RH) lateralization, as indicated by the left side bias effect: a chimeric face made from two left half faces from the viewer's perspective is usually judged more similar to the original face than one made from two right half faces (Gilbert \& Bakan, 1973). It is commonly assumed that HP is associated with RH lateralization. However, some experimental and computational studies (Hsiao \& Cottrell, 2009; Hsiao \& Cheung, 2011) showed the possibility of increased engagement of RH whereas decreased HP is measured. Another work on Chinese reading expertise (Tso, 2012) revealed a reduced HP for Chinese Writers as compared with Limited-writers; however there was no difference in left side bias between them. Our modeling work is hoped to also shed additional light on this issue.


Figure 1: Base Model

## Modeling Implementation

## Base model for non-drawers

Face recognition by non-drawers is modeled by Hsiao et al.'s (2008) intermediate convergence model of face recognition. This model (Figure 1) incorporated several known observations about visual anatomy and neural computation. Hsiao et al.'s (2008) used Gabor responses over the input images to simulate neural responses of cells in the early visual area, and Principal Component Analysis (PCA) to simulate possible information extraction processes beyond the early visual area. They then used this PCA representation as the input to a two-layer neural network. In addition, they implemented a theory of hemispheric asymmetry in perception, Double Filtering by Frequency theory (DFF, Ivry \& Robertson, 1997) in the model. The theory posits that visual information coming into the brain goes through two frequency-filtering stages. The first stage involves attentional selection of a task-relevant frequency range. At the second stage, the LH amplifies high spatial frequency (HSF) information, while the RH amplifies low spatial frequency (LSF) information. This differential frequency bias in the two hemispheres was implemented in the model by using two sigmoid functions assigning different weights to the Gabor responses in the two hemispheres. In the present implementation, the face input ( $100 \times 135$ pixels) was first filtered with a grid $(6 \times 6)$ of overlapping 2D Gabor filters in quadrature pairs at five scales and eight orientations. The five scales corresponded to 2 to 32 cycles per face (the task-relevant frequency range, depending on the image size. The maximum frequency should not exceed 2 pixels per cycle; the 6th scale, $2^{6}=64$ cycles per image exceeds the maximum frequency of the images, $100 / 2=50$ cycles per image). The resulting Gabor vector representation of the face was split into left and right halves. The perceptual representation of each half was compressed into a 50-element representation. After PCA, each principal component was z-scored to equalize the contribution of each component in the model. The PCA representation was then fed to a feedforward network with one hidden layer of 50 nodes. The number of nodes was determined empirically to allow efficient training of the network of all the three models of the present study. The output layer of the neural network has one output for each of the 53 faces of the testing set. Face images were taken from the CAlifornia Facial Expressions dataset (CAFÉ; Dailey, Cottrell, \& Reilly, 2001). We used two different neutral images for each face to constitute the training and testing sets. The neural network was trained with gradient descent with adaptive learning rate backpropagation from the MATLAB ${ }^{\circledR}$ Neural Network Toolbox (Version 7.0.3). All the networks were trained for both 400 epochs and 150 epochs. 400 epochs is enough for all the models to reach perfect recognition rates on the training sets and near perfect accuracy on testing sets. Training with only 150 epochs offers another viewpoint on the behavior of the three models by decreasing the ceiling effects observed with 400 epochs.

## Implementation of model I of drawing expertise

Our first hypothesis states that drawing experts engage in local attention on specific facial features at the encoding stage in addition to the global encoding process shared with non-drawers. Hence, in addition to the face input, model I includes isolated mouth and isolated eyes as local inputs. We filtered mouth images ( $50 \times 20$ pixels) and eyes images ( $74 \times 18$ pixels) by a bank of Gabor filters of three scales and eight orientations. The three scales corresponded to 2 to 8 cycles per face (The maximum frequency should not exceed 2 pixels per cycle; the 4 th scale, $2^{4}=16$ cycles per image exceeds the maximum frequency of the images, 18/2 $=9$ cycles per image for eyes and $20 / 2=10$ cycles per image for mouth). The size of the filtering grid ( $6 \times 6$ ) was the same for each kind of three - face, mouth and eyes inputs reflecting the engagement of the same resources for processing the global face or anyone of the two local parts. The choice of eyes as a facial feature was motivated by Tchalenko et al.'s (2003) finding that artists primarily focused on eyes. We added also a bottom facial feature: mouth, richly informative for artists. After Gabor filtering, the vector representations of mouth and eyes followed the same scheme of splitting, weighting and compressing as the one for face input. Hence, the neural network of model I was fed with an input layer of length 300 , with 100 PCA values for each of the three inputs. The model I of drawing expertise executes the same classification task as the base model. Hence, the two models have an identical output layer.


Figure 2: Model I of drawing expertise


Figure 3: Model II of drawing expertise

## Implementation of model II of drawing expertise

The second model of a drawing expert in Figure 3 is modified from the base model by adding at the classification stage of the neural network level two tasks. Namely, the model has to map the mouth and the eyes in the face input to respectively a "mouth cluster" and an "eyes cluster". This
second model shares the same input layer with the base model. This means that both models use the same attentional or perceptual resources to encode the input face. However, the expert model is trained with a more analytic task than mere face identification. It has to perform a cluster mapping operation for mouth and eyes. Four eyes and four mouth clusters were defined based on a set of features for eyes and mouth mentioned in textbooks on drawing portraits. This clustering ${ }^{1}$ yielded high recognition rates (> 98\%) for mouth and eyes on both training and testing sets for both training durations.

## Model of the composite task and measure of holistic processing

In human studies, HP is usually assessed through the composite paradigm (Young, Hellawell, \& Hay, 1987). In this paradigm, two stimuli are presented briefly, either sequentially or simultaneously. Participants attend to either the top or bottom halves of the stimuli and judge whether they are the same or different. In congruent trials, the attended and irrelevant halves lead to the same response, whereas in incongruent trials, they lead to different responses. HP is indicated by interference from the irrelevant halves in matching the attended halves; it can be assessed by the performance difference between the congruent and the incongruent trials.

The holistic face processing effect has been accounted for by computational models. For example, Cottrell, Branson, and Calder (2002) trained a computational model to perform a face identification task and an expression judgment task, and showed that the model was able to account for HP effects in both tasks. Richler, Mach, Gauthier, and Palmeri (2007) also used a variant of Cottrell et al.'s (2002) model to account for the HP effect in face recognition. To assess HP in our three models, we applied the method used by Hsiao and Cheung (2011), which was derived from Richler et al. (2007). After training we attenuated the Gabor responses of either the top or bottom half of the images in the test set by multiplying a factor of 0.125 to simulate directing the models' attention to the bottom or top half of the images respectively. For the first model of drawing expertise, for mouth and eyes inputs, only the unattended part was attenuated (eyes are in the top half, mouth is in the bottom half; see Figure 5(a)). The complete composite design was used; it has been shown to be more robust than the partial composite paradigm (Richler, Cheung, \& Gauthier, 2011). We created 4 types of stimulus pairs corresponding to the 4 conditions in Figure 4. Twenty pairs of images in each condition were randomly selected to form the materials (80 pairs in total). We calculated the correlation of the hidden layer representations in each pair as the similarity measure between them.

[^271]

Figure 4: Design of the composite task, with top halves attended.
A threshold was set to be the midpoint between the mean correlation of the "same" stimulus pairs and that of the "different" stimulus pairs. We assumed that the model responded "same" when the correlation of a pair was higher than the threshold, and responded "different" when the correlation was lower than the threshold. The HP effect was indicated by the discrimination performance difference between the congruent and incongruent trials measured by $\mathrm{d}^{\prime}$.

## Measuring hemispheric lateralization effect

The left side (RH) bias was assessed by the accuracy difference between recognizing a left-lateralized stimulus (carrying RH/LSF information) as the original stimulus and recognizing a right-lateralized stimulus (carrying LH/HSF information) as the original one. We defined RH lateralization (RH/LSF preference, Hsiao et al., 2008; Hsiao \& Lam, in press) as the left side bias measured in the biased condition minus that measured in the baseline condition. For the first model of drawing expertise with additional mouth and eyes inputs, lateralized stimuli were also used following the scheme applied to the face input (see Figure 5 (b)).


Figure 5: (a) Illustrative example of a Congruent Same pair for the composite task where bottom half is attenuated. (b) Example of a left-lateralized stimulus for measuring lateralization effects. For (a) and (b), eyes and mouth parts were only used in Model I of drawing expertise.

## Results

## Model I of drawing expertise (Experiment 1)

As shown in Figure 6, the model I of expertise with an input layer completed with mouth and eyes local inputs demonstrated less HP than the base model after either 150 or 400 epochs of training. For the 400 epochs case (the perfect accuracy case on the training set), a directional t-test revealed that model I was statistically significantly less holistic than the base model, $\mathrm{t}(798)=-1.76, \mathrm{p}=.04$, confirming our hypothesis. The mean value of $\Delta \mathrm{d}^{\prime}$ (Congruent d' - Incongruent d') for model I was smaller by a magnitude of 4 than the base model. This could be the result of a stronger ceiling effect. When decreasing the number of training epochs from 400 to $150, \Delta d^{\prime}$ for model I was increased from 0.006 to 0.023 , whereas $\Delta \mathrm{d}^{\prime}$ for the base
model increased from 0.026 to 0.063 . Decreasing the number of epochs did not change the significantly lower amount of HP for model I compared to the base model, $\mathrm{t}(798)=-2.29, \mathrm{p}=.011$. Model I with its increased size of the input layer initially generalized better than the base model. For 150 epochs, model I outperformed the base model ( $98 \%$ versus $91 \%$ recognition rates on the testing sets). However, by 400 epochs, the base model caught up with model I, and both models had equally perfect recognition rates.

Concerning RH lateralization (see Figure 7), a t-test indicated that model I was significantly more subject to a left side bias than the base model, $\mathrm{t}(798)=9, \mathrm{p}<.001$. For 150 epochs, the left side bias was further more accentuated for model I compared with the base model, $\mathrm{t}(798)=16.03$, p <. 001 .


Figure 6: Experiment 1. Holistic Processing


Figure 7: Experiment 1. RH Lateralization
Together the results indicated that our first model of drawing expertise compared with the base model of nondrawers is less holistic as measured by $\Delta \mathrm{d}$ ' and is characterized by a stronger left side (RH) bias effect. This finding of more RH lateralization for the model of drawing expertise was somewhat unexpected: drawers by focusing on parts in addition to global processing could have engaged in more LH/HSF processing than non-drawers. However, the main result here is the replication of Zhou et al. (2011)'s finding of less HP for drawing experts compared with nondrawers.

## Model II of drawing expertise (Experiment 2)

The model II of drawing expertise trained to recognize faces and to map mouths and eyes to respective clusters did
not demonstrate less HP than the base model (see Figure 8). Statistical analysis showed that the expert model was as holistic as the base model for both 400 and 150 epochs, $(\mathrm{t}(798)=-0.38, \mathrm{p}=.35$; $\mathrm{t}(798)=-1.12, \mathrm{p}=.13)$. We expected model II to behave less holistically than the base model but it did not.

Concerning the left side (RH) bias, a t-test showed that model II was significantly more RH lateralized than the base model for both 400 and 150 epochs, $(t(798)=4.56, \mathrm{p}$ $<.001$; $\mathrm{t}(798)=3.17, \mathrm{p}<.001)$. Again, this finding of more RH lateralization for the model of drawing expertise is somewhat unexpected: forcing the model to map eyes and mouth to cluster identities could have favored instead more LH/HSF processing (e.g., Hsiao \& Lam, in press).


Figure 8: Experiment 2. Holistic processing


Figure 9: Experiment 2. RH Lateralization

## Discussion \& Conclusion

Through computational modeling, we explored the nature of drawing expertise and aimed at accounting for Zhou et al. (2011)'s finding of less HP for drawing experts compared to non-drawers. Our first model of drawing expertise relied on engagement of local attention on face parts at the encoding stage in addition to the mere global face encoding in the case of the base model. This model of drawing expertise was successful in accounting for a lesser amount of HP compared with the base model. In the second model of drawing expertise, we kept the input layer of the base model but added to the face identification task, a mapping task of eyes and mouth to cluster identities. This second model was as holistic as the base model. Our modeling idea of an enriched input layer of both local and global information for experts in model I is supported by eye-tracking studies
(Miall \& Tchalenko, 2001; Tchalenko et al. 2003) of artists showing richer and more selective visual encoding by drawing experts compared with non-drawers.

Our findings of the two models of drawing expertise being more RH lateralized than the base model are congruent with the results of Hsiao and Cottrell (2009) on Chinese reading expertise. They found that Chinese character recognition experts have increased RH lateralization but reduced HP compared with novices. Like their results, our finding of increased RH lateralization but reduced HP for the first model of drawing expertise suggests that HP and RH lateralization may be separate processes that do not always go together, depending on the task requirement (Hsiao \& Cheung, 2011). Our finding also provides a testable hypothesis that face drawers may exhibit stronger left side bias in face perception than non-drawers.

Tso (2012) showed that Chinese Writers and Limitedwriters differed in HP but not in left side bias of Chinese characters. Drawers at first sight resemble Chinese Writers in that both achieved expertise through sharpening their motor and visual attention skills by eye-hand coordination while practicing their domain task. Nonetheless, the two groups may also differ in the following way. Chinese Writers were reinforced in a rote motor behavior while learning and copying the sequence of strokes for each character. However, drawers are not only challenged with each face's genuine and instantaneous uniqueness but critically have to render this uniqueness by capturing its gist in the details of the face. Hence, writing Chinese involves more rote motor learning than drawing faces; in contrast, drawers may develop better/finer visual attention skills than Chinese writers. Future work will examine whether our model can also account for Tso's (2012) finding in Chinese Writers and Limited-writers.

Our models of drawing expertise did not embed any motor component to represent motor drawing skills of experts. Hence, we showed that drawing experts and nondrawers could be sufficiently differentiated in terms of the nature (merely global versus both local and global) of attention during visual encoding of faces. We paved a first step in accounting for the nature of drawing expertise. It remains to be investigated what could be the contribution of motor expertise in drawing experts on the amount of HP they engage in.

## Acknowledgments

We are grateful to the Research Grant Council of Hong Kong (project code HKU 745210H to J.H. Hsiao) and the HKU Seed Funding Program for Basic Research (project code 201011159124 to J.H. Hsiao).

## References

Balas, B.J., \& Sinha, P. (2007). Portraits and perception: configural information in creating and recognizing face images. Spatial vision, 21(1-2), 1-2.
Bransford, J. (2000). How people learn: Brain, mind, experience, and school: National Academies Press.

Bukach, C.M., Gauthier, I., \& Tarr, M.J. (2006). Beyond faces and modularity: the power of an expertise framework. Trends in cognitive sciences, 10(4), 159-166.
Cottrell, G.W., Branson, K.M., \& Calder, A.J. (2002). Do expression and identity need separate representations. Paper presented at the Proceedings of the 24th Annual Conference of the Cognitive Science Society.
Dailey, M., Cottrell, GW, \& Reilly, J. (2001). California facial expressions (cafe). unpublished digital images, University of California, San Diego.
Gilbert, C., \& Bakan, P. (1973). Visual asymmetry in perception of faces. Neuropsychologia, 11, 355-362.
Hsiao, J.H., Shieh, D.X., \& Cottrell, G.W. (2008). Convergence of the visual field split: Hemispheric modeling of face and object recognition. Journal of Cognitive Neuroscience, 20(12), 2298-2307.
Hsiao, J. H., \& Cottrell, G. W. (2009). Not All Visual Expertise Is Holistic, but It May Be Leftist The Case of Chinese Character Recognition. Psychological Science, 20(4), 455-463.
Hsiao, J.H., \& Cheung, K.C.F. (2011). Holistic processing and right hemisphere lateralization do not always go together-Evidence from computational modeling. i-Perception, 2(4), 242-242.
Hsiao, J.H., \& Lam, S.M. (In press). The modulation of visual and task characteristics of a writing system on hemispheric lateralization in visual word recognition-A computational exploration. Cognitive Science.
Ivry, R.B., \& Robertson, L.C. (1997). The two sides of perception: MIT Press.
McClelland, J.L. (2009). The place of modeling in cognitive science. Topics in Cognitive Science, 1(1), 11-38.
McKone, E., Kanwisher, N., \& Duchaine, B.C. (2007). Can generic expertise explain special processing for faces? Trends in cognitive sciences, 11(1), 8-15.
Miall, R.C., \& Tchalenko, J. (2001). A painter's eye movements: A study of eye and hand movement during portrait drawing. Leonardo, 34(1), 35-40.
Richler, J.J., Mack, M.L., Gauthier, I., \& Palmeri, T.J. (2007). Distinguishing between perceptual and decisional sources of holism in face processing. Paper presented at the Proceesdings of the Twenty ninth Meeting of the Cognitive Science Society.
Richler, J. J., Cheung, O. S., \& Gauthier, I. (2011). Holistic processing predicts face recognition. Psychological Science, 22(4), 464-471.
Rigaud, J.F. (1877). A Treatise on Painting by Leonardo da Vinci: London: George Bell.
Tchalenko, J., Dempere-Marco, L., Hu, XP, \& Yang, GZ. (2003). Eye movement and voluntary control in portrait drawing. Mind, 2(3), 4.
Tso, R.V.Y., Au, T.K., \& Hsiao, J.H. (2011). The influence of writing experiences on holistic processing in Chinese character recognition. Paper presented at the Proceedings of the 33rd Annual Conference of the Cognitive Science Society.
Tso, Ricky. (2012). Motor experience modulates perceptual representation of objects: the case of Chinese character recognition. (Postgraduate), The University of Hong Kong.
Willenbrink, M., \& Willenbrink, M. (2012). Drawing Portraits for the Absolute Beginner: A Clear \& Easy Guide to Successful Portrait Drawing: North Light Books.
Wong, A.C.N., Palmeri, T.J., \& Gauthier, I. (2009). Conditions for Facelike Expertise With Objects Becoming a Ziggerin Expert-but Which Type? Psychological Science, 20(9), 1108-1117.
Young, A.W., Hellawell, D., \& Hay, D.C. (1987). Configurational information in face perception. Perception, 16(6), 747-759.
Zhou, G., Cheng, Z., Zhang, X., \& Wong, A. C. N. (2011). Smaller holistic processing of faces associated with face drawing experience. Psychonomic bulletin \& review, 1-6.

# A Joint Interference Effect in Picture Naming 

Chiara Gambi (c.gambi@sms.ed.ac.uk)<br>Department of Psychology, 7 George Square Edinburgh, EH8 9JZ U.K.

Joris Van de Cavey (joris.vandecavey@ugent.be)<br>Department of Experimental Psychology, 2 Henri Dunantlaan<br>Gent, 9000 Belgium

Martin J. Pickering (martin.pickering@ed.ac.uk)<br>Department of Psychology, 7 George Square<br>Edinburgh, EH8 9JZ U.K.


#### Abstract

In two experiments we provided evidence for a joint interference effect in picture naming. Participants took longer to name pictures when they believed that their partner concurrently named pictures than when they believed their partner was silent (Experiment 1) or concurrently categorized the pictures as being from the same or from different semantic categories (Experiment 2). However, picture naming latencies were not affected by beliefs about what one's partner said. These findings are consistent with the idea that speakers represent whether another speaker is preparing to speak, but not what they are preparing to say.


Keywords: joint task; co-representation; agent-conflict; language production; picture naming.

In this paper we report results from two experiments that, for the first time, combined a highly constrained language task (picture naming), with a manipulation of the context in which the task is performed (i.e., whether the participant speaks concurrently with her partner or on her own). A similar rationale has been used by researchers who compared solo and joint SR compatibility effects (see Knoblich, Butterfill, \& Sebanz, 2011 for a review), but it has never been applied to picture naming.

A well-known SR compatibility effect is the Simon effect. People are faster responding to "right" stimuli with their right hand and to "left" stimuli with their left hand (congruent trials) than they are responding to "right" stimuli with their left hand and to "left" stimuli with their right hand (incongruent trials). For example, people respond more quickly to the color of a stimulus when the stimulus (e.g., the photograph of a hand) is pointing towards the response hand than when the stimulus is pointing away from the response hand (Sebanz, Knoblich, \& Prinz, 2003).

A similar effect occurs when participants respond only with one hand, but they take turns with another participant who is seated next to them (i.e., they are slower when the pictured hand points towards the other participant than when it points towards themselves). This joint interference effect is interesting because the Simon effect is not observed (or is
reduced) if participants respond with one hand and they perform the task on their own.

The joint Simon effect has been interpreted as evidence that participants represent their partner's potential response and that this representation interferes with their own response on incongruent trials (because the two responses are incompatible, in the same way as a response with one's right hand is incompatible with a response given with one's left hand). We refer to this as the co-representation account of joint interference effects. Interestingly, joint compatibility effects were found when participants sat alone but were led to believe another person performed the task with them. This occurred even when no feedback was available (Atmaca, Sebanz, \& Knoblich, 2011).

The co-representation account has been challenged. Here we are particularly interested in an alternative account put forward by Wenke et al. (2011), the agent-conflict account. According to this account, representing that one's partner is (potentially) about to respond on the current trial interferes with one's own response. However, this occurs because there is a conflict regarding whose turn it is to respond, rather than because of incompatibility between one's own and one's partner's response. In fact, congruent responses should lead to similar amounts of interference as incongruent responses.

Joint interference effects have been almost exclusively investigated in manual tasks (e.g., Simon task, Flanker task, SNARC task), with only two studies using verbal responses (Philipp \& Prinz, 2010; Pickering \& MacLean, 2013) and none looking at picture-naming responses. Importantly, picture-naming responses are subject to varying degrees of congruency. For example, if one participant names the picture of an apple, her partner could either concurrently produce the same word (i.e., apple), or they could concurrently produce an unrelated word (e.g., blouse), or a related word (e.g., banana).

These different degrees of congruency do matter in solo tasks, as shown by several picture-word interference studies. Speakers who name pictures while ignoring distractor words are fastest when the distractor word is the picture's name.

They are slower when the distractor is a different word and slowest when it is a different but semantically related word. The difference in naming latencies between trials with unrelated distractors and trials with related distractors is due to interference between co-activated lexical representations (Levelt, Roelofs, \& Meyer, 1999).

In our study, participants saw pairs of pictures rather than picture-word pairs. When distractor words are replaced by distractor pictures, semantic interference effects generally disappear (Damian \& Bowers, 2003), possibly because distractor picture names are not routinely retrieved or their activation is too weak to out-weight facilitatory effects at the conceptual level. We therefore asked participants to name both pictures in a pair, a task that is subject to semantic interference effects (Aristei, Zwitserlood, \& Abdel Rahman, 2012).We asked whether the time they took to respond might be affected by a representation of their partner's concurrent response.

## Experiment 1

In Experiment 1, a red and a blue picture were simultaneously displayed to two participants seated in different rooms. Before the pictures appeared, an instruction screen showed the names of the two participants accompanied by the words red, blue, or no. Red and blue corresponded to "go" trials: the participant was instructed to name the picture presented in the given color first, and then also name the other picture. No corresponded to "no-go" trials: The participant was instructed to give no response.

We varied the order in which the other participant (the partner) was concurrently naming the pictures (Partner's task), as follows. On trials on which the two participants were assigned the same color, they named the pictures in the same order, therefore producing the same verbal response (SAME condition). On trials on which the two participants were assigned different colors, they named the pictures in reversed order, therefore producing different verbal responses (DIFF condition). Finally, when either of the participants was assigned a "no-go" cue, one participant named the pictures while their partner produced no response (NO condition). See Figure 1 (top) for examples (with apple in blue, blouse in red).

In addition, we introduced a second manipulation, orthogonal to Partner's task. Participants saw either two semantically related (e.g., apple - banana) or two unrelated pictures (e.g., apple - blouse). This served two purposes. The first was to provide a manipulation check. When two semantically related lexical items are activated concurrently (e.g., when speakers are asked to say "apple" and "banana" in close proximity), they interfere with one another (Aristei, et al., 2012). We therefore expected longer latencies when participants named two related than when they named two unrelated pictures (a main effect of semantic relatedness).

Most importantly, we expected Partner's task to affect naming latencies. Specifically, if the co-representation
account can be extended to naming responses, it could be taken to predict that speakers represent the content of their partner's response and activate the corresponding lexical representations.


Figure 1: Sample trial (top) and hypothesized effects according to the three accounts.

Note that, because the speakers always named both pictures, their utterances always contained the same lexical items. However, when the order differed, the picture that the speaker named second was the picture that their partner named first.
Therefore, in the DIFF condition the representation of the partner's response might enhance the activation of the second picture's name. This would in turn result in greater competition between the two pictures' names. Instead, when the order is the same, the first picture's name was the word that one's partner also named first. Therefore, its activation level might be raised and competition with the second picture's name could be reduced. Overall, we should find longer naming latencies in the DIFF condition than in the SAME condition.

This scenario is presented in Figure 1 (panel A). The nodes represent lemmas in Mary's mental lexicon. On the right is a snapshot of the activation level of the nodes apple and blouse just before the onset of the word "apple" when Mary is preparing to utter "apple blouse" (unrelated case), under the different conditions. The degree of activation is indicated by the thickness of the circles. Pointed arrows are excitatory connections, rounded arrows are inhibitory connections.

In addition, the degree of relatedness might also matter (and this was the second purpose of the relatedness manipulation). Specifically, if other-representations are content-specific, the semantic interference effect could be enhanced in the DIFF compared to the SAME condition.

Alternatively, speakers might not represent the content of their partner's response, but they might represent whether their partner responds on the current trial or not (agentconflict account). If so, the relationship between self- and other-representations would not affect processing, and hence naming latencies would be comparable in the SAME and DIFF conditions. For the same reason, there should be no interaction between Relatedness and Partner's task. However, naming latencies should be longer in the SAME and DIFF conditions than in the NO condition. This scenario is presented in Figure 1 (panel B).

Finally, people might not represent other people's responses. Note that our participants could not interact: They named pictures alongside each other, but could not hear each other. Whereas several studies have shown that non-interacting participants display joint interference effects (see above), they all used manual responses. We do not know whether the same would be true for verbal responses, particularly because language is perhaps more tightly linked to communicative situations compared to manual actions. If the Partner's task manipulation has no effect (i.e., no difference between the SAME, DIFF, and NO conditions), we would conclude that another person's utterances are not represented under the conditions tested in our experiment. This scenario is presented in Figure 1 (panel C) as the norepresentation account.

## Method

Participants Twelve pairs of previously unacquainted participants were recruited from the University of Edinburgh student community. All reported to be native English speakers and had no speaking or reading difficulties. They were paid $£ 6$ in return for participation.

Materials Fifty line drawings of everyday objects and animals were paired twice to yield 50 picture-picture pairs ( 25 semantically related, 25 semantically unrelated).

Design and Procedure Partner's task (henceforth, Partner; SAME vs. DIFF vs. NO) and Relatedness (unrelated vs. related) were manipulated within participants and within
items. An item was defined in terms of the first named picture (so apple-blouse and blouse-apple counted as different items). Partner varied on a trial-by-trial basis.

Each participant named a given item once per condition. Pictures were presented into 4 different blocks of 100 trials each. Each block comprised an equal number of trials in each condition for both participants. The order of presentation was pseudo-randomized, separately for each pair and for each block, with the constraint that the same picture never appeared on two consecutive trials. (The order of blocks was counterbalanced across pairs). In addition, we counterbalanced within each block and for each participant the color of the first named picture (blue or red) and the position of the cue (top or bottom half of the screen).

Participants were tested in adjacent soundproof booths. They were seated in front of computer monitors connected to the same machine in the control room (so stimulus presentation was simultaneous). There was a window between the two rooms, but participants could perceive each other only peripherally when facing the monitors.

Upon entering the lab, participants were introduced to one another and taken into the booths. After learning the picture names individually, they were told that they would "work together"; instructions were delivered to both participants at the same time in the control room. Participants then returned to the booths and, after performing 20 practice trials, began the experimental phase. A sample trial is shown in Figure 1 (top). A session lasted about 1 hour.

Recording and Data Analysis An inaudible beep marked stimulus presentation and was recorded together with the participants' responses (on three separate channels), using a multi-channel M-Audio FireWire 1814 device (inMusic, Cumberland, RI, www.m-audio.com) and Adobe Audition (Version 4.0; sampling rate: 48000 Hz ). Beep onsets were automatically tagged using Audacity (Version 1.2.5). Recordings were pre-processed to reduce background noise. Speech onsets were tagged using the Silence finder algorithm in Audacity and manually checked (for lip smacks, etc.). Naming latencies were defined as the time from beep onset to the onset of the participant's response.

The data were analyzed using Generalized Linear mixedeffects models (Bayeen, Davidson, \& Bates, 2008) in R (Version 2.7.2) with a logistic link function for categorical data (Jaeger, 2008). All predictors were contrast-coded. For Partner, we defined two planned contrasts: naming vs. no compared the DIFF and SAME conditions against the NO condition; same vs. different compared the SAME against the DIFF condition.

Fixed and random effects were selected using backward selection. If the model with full random structure did not converge we simplified it by removing higher order terms (first by subjects, then by items). The alpha-level for
likelihood-ratio tests was set to .05 for fixed effects, to .1 for random effects ${ }^{1}$.

Latencies were analyzed only if both pictures were named correctly. Incorrect responses included: naming errors (the wrong name was used), disfluencies, order errors (the name of the second picture was uttered first and vice versa), missing responses. Latencies longer than 3000 or shorter than 300 ms were considered outliers and excluded. Latencies more than 3 standard deviations from the byparticipant mean ( $1.5 \%$ ) were replaced with the cut-off value.

## Results

Accuracy Speakers produced (marginally) fewer incorrect responses when naming related than unrelated pictures ( $\left.\chi^{2}(1)=3.54, p=.06\right)$.

Table 1: \% incorrect in Exp. 1.

|  | DIFF | SAME | NO |
| :--- | :--- | :--- | :--- |
| Unrelated | $7.9 \%$ | $6.8 \%$ | $6.3 \%$ |
| Related | $8.1 \%$ | $5.3 \%$ | $4.9 \%$ |

Table 2: Best fit for accuracy data in Exp. 1.

| Predictor | Estimate | SE | Z |
| :--- | :---: | :---: | ---: |
| Intercept | -3.10 | .18 | -16.97 |
| naming vs. no | .24 | .11 | 2.23 |
| same vs. different | -.23 | .08 | -2.75 |
| related vs. unrelated | -.31 | .15 | -2.05 |
| Random effect | Explained variance estimate |  |  |
| Subjects: intercept | .48 |  |  |
| Items: intercept | .48 |  |  |
| Items: Relatedness | .56 |  |  |

Interestingly, the likelihood of producing an incorrect response was affected by $\operatorname{Partner}\left(\chi^{2}(2)=13.10, \mathrm{p}<.01\right)$ : They produced more incorrect response when their partner was naming than when their partner was silent and also fewer incorrect responses in the SAME than in the DIFF condition (see Table 1 and 2).

Naming latencies Participants took longer to name semantically related than unrelated pictures $\left(\chi^{2}(1)=11.32\right.$, $\mathrm{p}<.001$ ). Crucially, Partner affected naming latencies ( $\chi^{2}(2)$ $=7.80, \mathrm{p}<.05)$ : Latencies were longer when the partner was naming than when he was silent. However, the DIFF and SAME conditions did not differ. Finally, Relatedness and Partner did not interact (see Table 3 and 4).

[^272]Table 3: Mean latencies in Exp. 1.

|  | DIFF | SAME | NO | Tot |
| :--- | ---: | :--- | ---: | :---: |
| Unrelated | 869 | 869 | 855 | 864 |
| Related | 881 | 886 | 872 | 880 |
| Tot | 875 | 877 | 864 |  |
| Semantic |  |  |  |  |
| interference | -12 | -17 | -17 | -16 |

Table 4: Model for naming latencies in Exp. 1.

| Predictor | Estimate | SE | t |
| :--- | :---: | ---: | ---: |
| Intercept | 874 | 24 | 36.72 |
| naming vs. no | 14 | 5 | 2.79 |
| $\quad$same vs. different 1 <br> related vs. | 16 | 4 | .17 |
| unrelated | 5 | 3.36 |  |
| Random effect | Explained variance estimate |  |  |
| Subjects: intercept | 11980 |  |  |
| Items: intercept | 3150 |  |  |

## Discussion

Experiment 1 showed that beliefs about another's task can affect the latency of picture-naming responses, and are thus not consistent with the no-representation account. We take this as evidence that speakers represented that their partner was about to speak. More precisely, our results do not support the co-representation account. Though participants made more errors when their partner prepared an incongruent (DIFF) than a congruent (SAME) response, this pattern was not confirmed by latency data. In addition, while there was a clear semantic interference effect, which replicated previous findings (Aristei, et al., 2012), the effect was no greater in the DIFF ( 12 ms ) than in the SAME condition ( 17 ms ). These results are consistent with the agent-conflict account, as participants took longer to respond when they believed their partner also prepared to respond.

However, we must consider alternative explanations. Note that the slowest conditions (SAME and DIFF) are the ones in which two "go" instructions are displayed on the screen. Participants might be distracted by their partner's instruction more if it is a "go" instruction than if it is a "no-go" instruction, perhaps because "go" instructions are more similar to each other than they are to "no-go" instructions. This might cause interference between memory representations for one's own and the partner's instructions. Participants rarely performed their partner's task by mistake, which seems to suggest that they had little trouble remembering instructions. However, this occurred more often in the DIFF (on $2.3 \%$ of trials speakers named the pictures in their partner's order) than in the NO condition (on $1.2 \%$ of trials speakers gave no response). But more importantly, this explanation cannot account for the fact that latencies were equally long in the SAME as in the DIFF
condition (as in SAME instructions were identical). We return to this issue after Experiment 2.

We conclude that participants experienced interference whenever their partner responded concurrently, because they represented whether it was their partner's turn to respond. But what sort of mechanism could be responsible for this interference effect? The process of "imagining" that one's partner is about to respond might draw away attentional resources from the picture-naming task. If this is the case, "imagining" one's partner performing any task should slow down latencies to the same extent as "imagining" them naming.

However, it is also possible that interference arises because the same mechanisms (i.e., language production mechanisms) are used to represent one's partner naming response and to prepare one's own naming response. If this is the case, we predict less interference when one's partner is preparing a different (non-naming) task than when one's partner is preparing a naming response. Experiment 2 was designed to decide between these alternative explanations.

## Experiment 2

In Experiment 2 we replaced "no-go" trials with a semantic categorization (CAT) task. The SAME and DIFF conditions were exactly the same as in Experiment 1. In the CAT condition, partners were instructed to judge whether the two pictures belonged to the same semantic category or to different semantic categories. They responded by saying "yes" or "no" into the microphone.

Thus, all trials required a response from both participants. If imagining one's partner performing any task was driving the effect we observed in Experiment 1, we should now find no difference between the SAME, DIFF and CAT conditions. Note that both the CAT task and the naming task involve visual processing of the pictures and retrieval of the concepts associated with the depicted entities from memory. In addition, both tasks require articulation of an overt verbal response.

Crucially, however, only the naming task engages language production mechanisms (and specifically the retrieval of the picture's name). Therefore, if the interference effect in Experiment 1 is due to a representation that one's partner is preparing a naming task, we should replicate it in Experiment 2.

## Method

Sixteen new participants from the University of Edinburgh student community were recruited. Materials, design and procedure were as in Experiment 1, except that the CAT condition replaced the NO condition. For the semantic categorization task, participants were told that when they saw the word question (which replaced the word no) next to their name, they were to respond to the following question: "Are the two pictures from the same category?" Data were
analyzed as in Experiment 1; latencies exceeding the 3SDthreshold amounted to $1.7 \%$ of the data.

## Results and Discussion

Categorization Task Participants responded correctly on $94.7 \%$ of the unrelated trials and on $93.6 \%$ of the related trials (a non-significant difference).

Accuracy Speakers produced (marginally) more incorrect naming responses to related than unrelated pictures $\left(\chi^{2}(1)=\right.$ $2.98, \mathrm{p}=.08$ ). More importantly, Partner did not affect the likelihood of producing an incorrect response (see Table 5).

Table 5: \% incorrect in Exp. 2.

|  | DIFF | SAME | CAT |
| :--- | :--- | :--- | :--- |
| Unrelated | $5.6 \%$ | $6.3 \%$ | $6.0 \%$ |
| Related | $7.2 \%$ | $7.1 \%$ | $5.8 \%$ |

Naming latencies Participants took longer to name semantically related than unrelated pictures $(\chi 2(1)=11.04$, $\mathrm{p}<.001$ ). As in Experiment 1, Partner affected latencies $(\chi 2(2)=6.54, p<.05)$ : They were longer when participants believed their partner named pictures than when they believed their partner categorized the pictures. However, the DIFF and SAME conditions did not differ and Relatedness and Partner did not interact (see Table 6 and 7).

Table 6: Mean latencies in Exp. 2.

|  | DIFF | SAME | NO | Tot |
| :--- | ---: | :--- | :--- | :--- |
| Unrelated | 881 | 879 | 874 | 878 |
| Related | 898 | 907 | 885 | 897 |
| Tot | 889 | 893 | 880 |  |
| Semantic |  |  |  |  |
| interference | -17 | -28 | -11 | -19 |

Note that in Experiment 2 two "go" instructions were displayed on every trial, including in the CAT condition; therefore, interference could not have been due to greater interference between memory representations for more similar instructions.
The results of Experiment 2 are not consistent with the co-representation account. As in Experiment 1, naming latencies were very similar in the DIFF and SAME condition. In addition, and unlike in Experiment 1, the likelihood of incorrect responses was very similar in the two conditions (and did not differ significantly from the CAT condition, either). Finally, the semantic interference effect was not larger in the DIFF than in the SAME condition.

Table 7: Model for naming latencies in Exp. 2

| Predictor | Estimate | SE | t |
| :--- | :---: | ---: | ---: |
| Intercept | 884 | 24 | 36.77 |
| naming vs. no | 12 | 5 | 2.47 |
| same vs. different | 3 | 4 | .70 |
| related vs. unrelated | 19 | 5 | 3.48 |
| Random effect | Explained variance estimate |  |  |
| Subjects: intercept | 16490 |  |  |
| Subjects: Size ${ }^{2}$ | 13080 |  |  |
| Items: intercept | 46670 |  |  |
| Items: Relatedness | 4380 |  |  |

Most importantly, we found that naming latencies are longer when speakers believe that their partner is also naming a picture than when they believe their partner is performing a semantic categorization task. Given that the two tasks share all processing stages except lexical retrieval, we conclude that the process of naming pictures is inhibited by the belief that another speaker is concurrently retrieving the pictures' names.

## Conclusion

We showed that people represent their partner's task in a joint picture-naming task. The evidence is not consistent with the co-representation account of joint task effects. Participants did not form content-specific representations of their partner's response. It is possible that this finding is limited to the conditions tested in this study. Interlocutors might form content-specific representations when engaged in a conversation (when they rarely speak at the same time). In addition, the amount of practice and repetition that characterizes picture naming experiments could have masked content-specific effects (perhaps because activation was already at ceiling). Future studies should consider these limitations.

However, our results are consistent with a version of the agent-conflict account, in which interference in naming responses is due (at least partly) to the belief that one's partner is preparing a naming response (as opposed to any response). This is consistent with the idea that people represent others' utterances using some of the mechanisms they use in preparing their own utterances (i.e., language production mechanisms; Pickering \& Garrod, in press).

## Acknowledgments

We would like to thank Eddie Dubourg and Ziggy Campbell. C. Gambi is supported by a University of

Edinburgh scholarship. J. Van de Cavey is supported by an FWO scholarship.

## References

Aristei, S., Zwitserlood, P., \& Abdel Rahman, R. (2012). Picture-induced semantic interference reflects lexical competition during object naming. Frontiers in Psychology, 3, doi: 10.3389/fpsyg.2012.00028
Atmaca, S., Sebanz, N., \& Knoblich, G. (2011). The joint flanker effect: sharing tasks with real and imagined coactors. Experimental Brain Research, 211, 371-385.
Bayeen, R. H., Davidson, D. J., \& Bates, D. (2008). Mixedeffects modeling with crossed random effects for subjects and items. Journal of Memory and Language, 59, 390412.

Damian, M. F., \& Bowers, J. S. (2003). Locus of semantic interference in picture-word interference tasks. Psychonomic Bulletin \& Review, 10, 111-117.
Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. Journal of Memory and Language, 59, 434-446.
Knoblich, G., Butterfill, S., \& Sebanz, N. (2011). Psychological research on joint action: Theory and Data. In B. Ross (Ed.), The psychology of learning and motivation. Burlington: Academic Press.
Levelt, W. J. M., Roelofs, A., \& Meyer, A. S. (1999). A theory of lexical access in speech production. Behavioral and Brain Sciences, 22, 1-75.
Philipp, A. M., \& Prinz, W. (2010). Evidence for a role of the responding agent in the joint compatibility effect. Quarterly Journal of Experimental Psychology, 63, 21592171.

Pickering, M. J., \& Garrod, S. (in press). An integrated theory of language production and comprehension. Behavioral and Brain Sciences.
Pickering, M.J. \& MacLean, J. (2013). Representing others’ words: Just like one's own? Unpublished manuscript.
Sebanz, N., Knoblich, G., \& Prinz, W. (2003). Representing others'actions: Just like one's own? Cognition, 88, B11B21.
Wenke, D., Atmaca, S., Holländer, A., Liepelt, R., Baess, P., \& Prinz, W. (2011). What is shared in joint action? Issues of co-representation, response conflict, and agent identification. Review of Philosophy and Psychology, 2, 147-172.

[^273]
# Semantic Ambiguity Resolution as a Decision Process 

Nicholas Gaylord (nlgaylord@utexas.edu)<br>Colin Bannard (bannard@utexas.edu)<br>Department of Linguistics, University of Texas at Austin<br>305 E. 23rd St. B5100, Austin, TX 78751 USA


#### Abstract

Resolution of the meaning of a semantically ambiguous word requires knowledge about the space of possible meanings of that word, and the selection of a meaning in the light of available evidence and given situational constraints. As such, ambiguity resolution bears many similarities to decision making scenarios more generally. We report on an experiment exploring this analogy by applying some standard manipulations from the decision making literature to a semantic disambiguation task. We explore two particular proposals: (1) that depth of semantic processing can be cast as strategy selection reflecting a risk-sensitive effort-accuracy tradeoff, and (2) that thresholds for inference about meaning in context are situationally flexible and learnable via feedback. One robust property of decision making is people's ability to use feedback in order to adjust responses to maximize payoffs. Participants completed a semantic entailment judgment task in which they received trial-by-trial feedback, and payoff matrices and decision thresholds were manipulated across conditions. We find an effect of risk, with participants employing different comprehension strategies depending on relative gains and losses. We also find that participants were in fact sensitive to varying decision thresholds and accurately adjusted their behavior to match the constraints on what qualified as a true conclusion in different conditions. We take these findings as preliminary evidence that ambiguity resolution in language can be modeled, at least in part, as involving more general decision processes.


Keywords: Speed-Accuracy Tradeoff, risk, decision thresholds, decision making, sentence processing, word meaning, ambiguity

## Introduction

Semantic ambiguity is a widespread phenomenon in natural language, whereby a single word can have more than one interpretation depending on its use. The resolution of semantic ambiguity requires knowledge of the range of possible meanings of an ambiguous word, and the consideration of those possibilities in light of available contextual evidence and given certain situational constraints (such as how strict or precise an interpretation is required). Characterized as such, semantic ambiguity resolution bears many similarities to other scenarios that are studied in research on human decision making. However, the connection between decision making and semantic processing is as yet underexplored.

One robust property of human decision making in other domains is the ability to use feedback in order to adjust responses to maximize benefits (increasing material rewards and/or minimizing cognitive costs). In this paper we look at whether the same behavior might be observed for a semantic disambiguation task. Two particular manipulations were employed, parallel to manipulations in other decision tasks: a) changes to the decision threshold, which separated correct "true" or "false" responses concerning the meaning of a word in context, and $b$ ) changes to the degree of risk (possible material losses) in the decision situation. Such factors
have useful analogues in language understanding. Decision threshold changes are implicated in that different situations call for more (or less) restrictive assumptions as to what can be safely concluded from a potentially ambiguous utterance. Risk is implicated via the potential negative consequences of misinterpretation, which is greater in some cases than others - for example, a failure of interpretation is likely more consequential in a job interview than in a casual conversation.

## Background

Semantic ambiguity has been extensively studied from a variety of perspectives including linguistic theory and psycholinguistics. One important finding from this work is that not all cases of semantic ambiguity are the same - Apresjan (1974) argues that different senses (or uses) of a word can vary in how semantically similar they are, and most psycholinguistic research into the representation of semantic ambiguity arrives at a similar conclusion (Brown, 2008; Frazier \& Rayner, 1990; Klepousniotou, Titone, \& Romero, 2008; Pickering \& Frisson, 2001; Williams, 1992). This position is further supported by various offline judgment tasks (Erk, McCarthy \& Gaylord 2009, To Appear; Gaylord, 2011). In short, there is a growing body of evidence that word meanings are graded - the meanings of individual occurrences of a word can vary quite subtly, and the extent to which word senses apply to a given occurrence varies in a graded fashion as well.

A closely related question is that of how we use the information available in our lexical representations to determine a contextually-appropriate meaning. McElree, Murphy, and Ochoa (2006) and Gaylord, Goldwater, Bannard, and Erk (2012) both investigated the dynamics of this process using a Speed-Accuracy Tradeoff (SAT) design. McElree, Murphy, and Ochoa observed elevated false alarms after short processing delays with stimuli such as Water pistols - are dangerous and Gaylord et al. found the same effect with stimuli such as The dawn broke - Something shattered. In other words, both studies found evidence that when a word is encountered, a context-independent default meaning is activated prior to semantic integration, whether or not it is supported by the occurrence in question. It is likely that these default meanings correspond to those words' most frequent interpretations. There is a current debate as to how information-rich our lexical knowledge must be (cf. Elman 2011) and while evidence is accumulating that our semantic representations provide access to a great deal of richly informative world knowledge, results such as those just discussed also indicate that our knowledge of word meanings contains a more schematic layer that is more efficient to access.

One question that can be raised is why this more schematic level of semantic representation is present despite the fact that it can lead to errors of interpretation. A plausible answer is that sentence comprehension, like other cognitive processes, is subject to economic pressures, and that under many circumstances a shallower processing is sufficient (Barton \& Sanford, 1993; Bever, Sanz, \& Townsend, 1998; Ferreira \& Patson, 2007; Sturt, Sanford, Stewart, \& E Dawydiak, 2004; Swets, Desmet, Jr., \& Ferreira, 2008; Townsend \& Bever, 2001). We hypothesize that these default meanings, as they reflect a word's most likely interpretation, support a shallow semantic processing strategy.

A considerable amount of decision making research addresses the question of strategy selection (Beach \& Mitchell, 1978; Busemeyer, 1993; Gigerenzer, Todd, \& ABC Research Group, 1999; Johnson \& Payne, 1985; Payne, Bettman, \& Johnson, 1988). This work studies how people select more or less effortful decision making strategies in different situations, where increased effort tends to yield increased accuracy. The concept of an effort-accuracy tradeoff is central to strategy selection, and has been seen to be sensitive to risk. Semantic comprehension has been shown to be effortful, and we propose that depth of semantic processing can be cast as a strategy selection problem driven by a risksensitive effort-accuracy tradeoff. We explore this hypothesis through changes across conditions to the payoff matrix dictating potential gains and losses for correct and incorrect responses. We hypothesize that shallow processing strategies (marked by acceptance of default meanings in the absence of contextual support) will be more prevalent under decreased risk, and dispreferred when potential losses are high.

However, parallels with strategy selection are not the only similarity between semantic comprehension and decision making more generally. More generally, ambiguity resolution requires the selection of a possible interpretation of a word in light of available contextual evidence, and given situational constraints on interpretation. As discussed above, meaning-in-context appears to be a very graded phenomenon, and another question is whether people adapt their semantic comprehension behavior to meet situational demands. We explore this question as well by moving the threshold (corresponding to a property of the stimulus) at and above which a response of "true" will be counted as correct in different conditions, and providing trial-by-trial feedback on response accuracy. We hypothesize that participants will use their graded representations of word meanings in order to rapidly learn an optimal decision threshold.

## Experiment

Participants 131 undergraduate psychology students from the University of Texas at Austin completed the experiment in exchange for course credit. Participants received a cash payment of up to $\$ 3.00$ depending on their performance on the task. All participants were native English speakers.

Table 1: Example stimuli, with their associated truth norms. TS = true given context sentence; PS = plausible given context sentence; FS = false (but possible given a different sentence); FV = false given the verb (false regardless of context)

| Context | Probe | Norm | Type |
| :--- | :--- | :--- | :--- |
| The insult burned | Something was mean | 6.65 | TS |
| The insult burned | Something was true | 4.85 | PS |
| The insult burned | Something was warm | 1.30 | FS |
| The insult burned | Something was rolled | 1.30 | FV |
| The log burned | Something was warm | 6.55 | TS |
| The log burned | Something was dangerous | 4.55 | PS |
| The log burned | Something was mean | 1.10 | FS |
| The log burned | Something was fixed | 1.20 | FV |

Table 2: Summary of experimental conditions. Threshold is the truth norm at and above which items were counted as true.

| Condition | Gain/Loss | Threshold |
| :---: | :--- | :--- |
| A | $+5 /-1$ | 3.7 |
| B | $+1 /-5$ | 3.7 |
| C | $+5 /-5$ | 3.7 |
| D | $+5 /-5$ | 2 |
| E | $+5 /-5$ | 6 |

Materials The experiment, 240 trials in length, took the form of a semantic judgment task in which each trial consisted of a context sentence (e.g. The dawn broke) followed by a semantic probe (e.g. Something shattered) to be evaluated as true or false. Each context sentence was paired with a true probe, a plausible but not necessarily true probe, and two false probes: one which would be true under a different meaning of the context verb, and one which was false given any contextually-activated meaning of the verb. The truthfulness of each probe, given its context sentence, was measured via a separate offline norming task and the averages of these ratings established a truthfulness value for each stimulus. Further examples of stimuli are contained in Table 1.
Procedure Participants were told that the experiment would take the form of a game, in which points were gained or lost based on accuracy, and that those points were redeemable for cash at the end of the session. At the start of the experiment, participants were familiarized with the gains and losses associated with correct and incorrect responses, and after each trial they received feedback about their response accuracy (a smiling face for a correct response or a frowning face for an error) and associated gain or loss of points.

A schematic of an experimental trial is shown in Figure 1. The experiment contained 5 conditions (summarized in Table 2), across which we varied risk (via changes to payoff matrices, which were always symmetrical) and decision threshold.

Risk. In condition A participants could gain 5 points for a correct answer, but only lose 1 point for an incorrect answer.


Figure 1: A single trial, depicting a correct "false" response.

Hence the risk associated with giving an errorful response was minimal. In condition B, participants could gain only 1 point for a correct answer but could lose 5 points for an incorrect answer. In condition $C$ participants could gain 5 points for a correct answer and lose 5 points for an incorrect answer. Both conditions B and C are characterized by large potential losses, and are by extension higher in risk than condition A. While the difference between possible gain and loss on a trial is smaller in condition $B$ than in condition $C$, and is in fact exactly the same as in condition A, we know that people tend to evaluate risk relative to a status quo reference point and tend to be loss-averse (Kahneman \& Tversky, 1979). As such, condition B is the highest risk case overall, since possible gains are much smaller than possible losses.

Acceptance Threshold. Across conditions C-E we varied the acceptance threshold while holding risk constant. Acceptance threshold manipulations were accomplished relative to the mean truthfulness ratings that we independently gathered. Very high-rated stimuli were true across conditions, and very low-rated stimuli were consistently false, but stimuli with intermediate truthfulness ratings were counted as true in some conditions but false in others. The threshold for condition C was at median. The threshold for conditions D and E were lower and higher respectively. These thresholds determined the feedback we gave to participants on their responses.

## Results and Discussion

Participants' responses and response times were recorded on each trial. Participant responses of under 150 ms were excluded, as well as the $0.5 \%$ of slowest responses.

Effects of Risk. Risk was manipulated via changes to the study payoff matrix, such that in Condition A possible gains on each trial were large and possible losses small, while in Conditions B and C possible losses were greater. We hypothesized that participants would employ different response strategies in the higher-risk Conditions C (in which uninformed responding would yield a loss relative to the maximum points possible) and $B$ (in which uninformed respond-
ing would be expected to yield a loss relative to the starting point), relative to the low-risk Condition A (in which uninformed responding would be expected to yield a net gain).

We performed a series of multilevel logistic regression models, in which the outcome was the participants' response (true $=1$, false $=0$ ) and in which participant ID was included as a random effect on the model intercept. We first of all examined simple accuracy by looking at whether correct response was a good predictor of actual participant responses across the conditions. As discussed in Wright and London (2009) this is equivalent to a traditional d-prime analysis. A model containing an interaction between correct response and condition was found to give a significantly better fit to the data than a model containing only correct response $\left(\chi^{2}(4)=15.673, p\right.$ $<0.01$ ) and a model containing both terms but no interaction $\left(\chi^{2}(2)=11.172, \mathrm{p}<0.01\right)$. The coefficients revealed the increase in the likelihood of participants responding "true" if the correct response was "true" was significantly greater in both Conditions B and C than it was in condition A.

We next looked in more detail at how participants were making their decisions. In our norming study we obtained graded ratings as to whether the probe sentences were entailed by the context sentences. We assume that participants in our main study were able to utilize intuitions that corresponded to such scales. We first looked at whether our normed truth scale was predictive of response in a series of logistic regression models. A model including the truth norm rating as a predictor gave a better fit to the data than a model including the correct response as sole predictor. A model containing an interaction between truth norm rating and condition was a significantly better fit than a model containing only truth norm $\left(\chi^{2}(4)=17.794, p<0.01\right)$ or one containing both terms but no interaction $\left(\chi^{2}(2)=13.228, \mathrm{p}<0.01\right)$. The coefficients revealed the increase in the likelihood of participants responding "true" as a function of increases in the truth norm was significantly greater in both conditions B and C than in condition A. We next fit separate logistic regression models to the data from each of the conditions and looked at the predictive value of the truth norms in each case. Log Likelihood Ratio Indices (McFadden, 1974) revealed that the truth values had more predictive value in conditions B (0.458) and $\mathrm{C}(0.477)$ than in condition $\mathrm{A}(0.420)$. These data further support the finding that probe truthfulness is a stronger determinant of participant response under increased risk.

We performed a final exploration by defining a simple model based on these truthfulness values and exploring how well it accounts for participants' responses. We assume that an idealized responder would say "true" for a given item with a probability equal to the mean truth rating provided (minus the minimum possible response, 1 ), divided by the difference between the minimum and maximum response (6). We look at the perplexity (an information theoretic measure of how surprised the model is by the data) of such a model when confronted with participant response data. Model perplexity is higher for condition $\mathrm{A}(2.328)$ than it is for conditions

B (2.213) and C (2.110). This indicates that participant responses in the riskier conditions are better described by probe truthfulness than are responses in the less risky conditions.

These analyses suggest that participants are making more sensitive semantic judgements in the riskier conditions. Because semantic comprehension is effortful, one explanation of this is that there is an effort-accuracy tradeoff at work. Participants should be more willing to expend this effort via deeper semantic processing when there is more at stake. A possible consequence of this would be an increase in the time taken to make decisions. This effect is in fact seen, though response times are not elevated across-the-board in the riskier conditions. Rather, stimuli with very high or very low truthfulness values are processed as quickly as in the low-risk condition, and extra time is spent precisely those items that warrant it items with intermediate truthfulness ratings.

Effects of Decision Threshold Placement. We next turn to the effect of changes to the decision threshold. In Conditions C-E, the same stimuli were used but the threshold value at and above which a probe was considered true was varied across conditions. Based on our above argument that word meaning in context is a graded phenomenon whose scales can be used flexibly in making decisions, we hypothesized that participants would adjust their responses to reflect these thresholds. Our norming study showed that people are able to reliably assign graded values as to whether our probe sentence was entailed by our context sentence. We assume that participants in our main experiment will have similar graded evaluations and that they will respond differently depending on our different conditions by inferring an optimal point on their scales at which to accept or reject probes. It is worth reiterating that this is not an arbitrary manipulation - different situations do indeed carry different constraints on meaning-in-context inference. A legal contract, for example, demands a very constrained interpretation of explicitly presented information, while innuendo demands much greater inference.

Acceptance probabilities (across participants) for all stimuli as a function of their truthfulness values are contained in Figure 2, in which it is clearly visible that participants do evaluate stimuli differently between conditions. This is particularly true for stimuli with intermediate (2-6) truthfulness ratings, which are evaluated differently depending on threshold placement. There is much less effect on the acceptance of very high- or very low-rated stimuli. The effect of threshold placement is further supported by the improved fit of a model with an interaction between item truth norm and condition, compared to a model with item truth norm as the only predictor $\left(\chi^{2}(4)=191.69, p<0.001\right)$. Our primary interest, however, is in how rapidly participants learn different decision thresholds. Figure 3 sheds light on this question, showing the changes over trials in the minimum truthfulness ratings for the items that are accepted and the maximum ratings for the items that are rejected for the three conditions. This shows how participants adjust these cutoffs over trials differ-


Figure 2: Acceptance probabilities for stimuli as a function of their truthfulness ratings in Conditions C-E. For each condition, the "true"/"false" threshold is indicated in red.
ently in line with the acceptance thresholds revealed to them via feedback. This is supported by model comparison - a model including a three-way interaction between item truth norm, condition, and trial number gives a significantly better fit ( $\left.\chi^{2}(6)=181.1, \mathrm{p}<0.001\right)$. We take this as evidence that people do dynamically adjust their judgments about meaning-in-context, specifically how broadly or conservatively they interpret meaning, in response to situational constraints.

## General Discussion

We found that participants dynamically adjust their assumptions regarding the conclusions that can be drawn from a given utterance in response to feedback. We also found that they employed different responding strategies depending on risk, and that the difference was not simply due to a speedaccuracy tradeoff. These results suggest that decision making behaviors that have been reported in other non-linguistic do-


Figure 3: Lowess smoothed values for the minimum ratings at which items were accepted (the lower line) and maximum ratings at which items were rejected (the upper line) by trial for our conditions C-E.
mains might reasonably be extended to sentence processing, and in particular to meaning-in-context resolution.

The finding that semantic ambiguity resolution is affected by economic considerations is timely given related developments in the literature. One is the appearance in the sentence processing literature of the so-called "Good Enough" approach (Ferreira \& Patson, 2007), along with other studies (discussed above) that have found people often only engage in shallow syntactic or semantic processing. To the best of our knowledge, the present work is the first to extend this line of inquiry to semantic processing at the lexical level, and the first study to explicitly predict semantic processing depth based on situational characteristics. The connection we make here with the decision making literature suggests further possibilities for studying the effect of situational pressures on language processing. Techniques from this literature, such as trial-by-trial feedback, are being adopted by other research in language processing as well (Lewis, Shvartsman, \& Singh, To Appear) . Another recent development, this time in the theoretical linguistic literature, has been the use of ideas from decision theory and game theory to discuss linguistic communication (Clark, 2012) and particularly pragmatics (Benz, Jager, \& Rooij, 2006). One of the main challenges in extending these accounts is the effective parameterization of utilities. Our findings suggest that standard techniques from the decision making literature might be useful in this regard.

We have argued here that participants' performance in the absence of risk reflects the use of default interpretations of the kind found by McElree et al. (2006) and Gaylord et al. (2012). Depending on the degree of situational risk, people might vary in how readily they will accept an initially-activated default interpretation of a word, presumably because under cer-
tain payoff schemes it is no longer worth the effort of computing a more precise interpretation to avoid a marginal potential loss. A related question to be explored in greater detail in future research is how this readiness to accept default interpretations is affected by the relative strength of the default meaning versus other potentially competing candidate interpretations (Kilgarriff, 2004), and in fact whether it is the case that only one default interpretation is activated.

Participants’ dynamic adjustment to different truthfulness thresholds is equally striking as it shows that in different situations they rapidly learned how conservative or permissive to be regarding the possible conclusions that can be drawn from given information. This is particularly relevant for the study of semantic ambiguity resolution due to the fact that contextualized meaning has long been tied to the set of conclusions that can be drawn from a sentence. While approaches such as that in Chierchia and McConnell-Ginet (2000) are more restrictive in that they characterize sentence meaning through entailments, which are necessarily true, as opposed to here where we also deal with plausible conclusions, the general sentiment of these approaches is nonetheless applicable. Additionally, manipulations of decision threshold such as those employed here may prove useful to the broader study of inference in experimental pragmatics.

An immediate next step is to observe the effects of simultaneously varying both risk and decision threshold. We have already seen that participants rapidly learn to approximate the threshold, and we have seen that participants become more deliberative under higher risk. These facts jointly predict that threshold learning will be both more rapid and more accurate under increased risk. Investigation of these questions is currently underway.

## Acknowledgments

This work was supported in part by a Carlota S. Smith Memorial Research Fellowship to the first author. We also wish to thank Arthur B. Markman for many very important insights as well as the use of his lab, and J. Grant Loomis for his invaluable assistance throughout the data collection process.

## References

Apresjan, J. (1974). Regular polysemy. Linguistics, 142, 5-32.
Barton, S. B., \& Sanford, A. J. (1993). A case study of anomaly detection: Shallow semantic processing and cohesion establishment. Memory \& Cognition, 21, 477-487.
Beach, L. R., \& Mitchell, T. R. (1978). A contingency model for the selection of decision strategies. The Academy of Management Review, 3, 439-449.
Benz, A., Jager, G., \& Rooij, R. van. (2006). Game theory and pragmatics. New York: Palgrave Macmillan.
Bever, T. G., Sanz, M., \& Townsend, D. J. (1998). The emperor's psycholinguistics. Journal of Psycholinguistic Research, 27, 261-284.
Brown, S. (2008). Choosing sense distinctions for WSD: Psycholinguistic evidence. In Proceedings of ACL-08: HLT, short papers (companion volume) (p. 249-252). Association for Computational Linguistics.
Busemeyer, J. (1993). Violations of the speed-accuracy tradeoff relation. In O. Svenson \& A. J. Maule (Eds.), Time pressure and stress in human judgment and decision making. New York: Plenum Press.
Chierchia, G., \& McConnell-Ginet, S. (2000). Meaning and grammar: An introduction to semantics (2nd ed.). MIT Press.
Clark, R. (2012). Meaningful games: Exploring language with game theory. MIT Press.
Elman, J. L. (2011). Lexical knowledge without a lexicon? The Mental Lexicon, 6, 1-33.
Erk, K., McCarthy, D., \& Gaylord, N. (2009). Investigations on word senses and word usages. In Proceedings of $A C L$ 2009.

Erk, K., McCarthy, D., \& Gaylord, N. (To Appear). Measuring word meaning in context.
(To appear in Computational Linguistics)
Ferreira, F., \& Patson, N. D. (2007). The good enough approach to language comprehension. Language and Linguistics Compass, 1, 71-83.
Frazier, L., \& Rayner, K. (1990). Taking on semantic commitements: Processing multiple meanings vs. multiple senses. Journal of Memory and Language, 29, 181-200.
Gaylord, N. (2011). Exploring contextual effects on word meaning via multiple-level similarity judgments. In L. Carlson, C. Holscher, \& T. Shipley (Eds.), Proceedings of the 33rd annual conference of the cognitive science society (pp. 950-955). Austin, TX: Cognitive Science Society.
Gaylord, N., Goldwater, M., Bannard, C., \& Erk, K. (2012). Default verb meanings and verb meaning-in-context: A
speed-accuracy tradeoff study. In Proceedings of architectures and mechanisms for language processing (AMLaP 2012) (p. 218). Riva del Garda, Italy.

Gigerenzer, G., Todd, P. M., \& ABC Research Group the. (1999). Simple heuristics that make us smart. Oxford University Press.
Johnson, E. J., \& Payne, J. W. (1985). Effort and accuracy in choice. Management Science, 31, 395-414.
Kahneman, D., \& Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, 47, 263292.

Kilgarriff, A. (2004). How dominant is the commonest sense of a word? In Sojka, Kopecek, \& Pala (Eds.), Proceedings of Text, Speech, Dialogue. Lecture notes in artificial intelligence (pp. 103-112). Springer.
Klepousniotou, E., Titone, D., \& Romero, C. (2008). Making sense of word senses: The comprehension of polysemy depends on sense overlap. Journal of Experimental Psychology: Learning, Memory, and Cognition, 34, 1534-1543.
Lewis, R. L., Shvartsman, M., \& Singh, S. (To Appear). The adaptive nature of eye-movements in linguistic tasks: How payoff and architecture shape speed-accuracy tradeoffs. Topics in Cognitive Science.
McElree, B., Murphy, G., \& Ochoa, T. (2006). Time course of retrieving conceptual information: A speed-accuracy tradeoff study. Psychonomic Bulletin and Review, 13, 848-853.
McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), Frontiers in econonometrics (pp. 105-142). New York: Academic Press.
Payne, J. W., Bettman, J. R., \& Johnson, E. J. (1988). Adaptive strategy selection in decision making. Journal of Experimental Psychology: Learning, Memory, and Cognition, 14, 534-522.
Pickering, M., \& Frisson, S. (2001). Processing ambiguous verbs: Evidence from eye movements. Journal of Experimental Psychology: Learning, Memory, and Cognition, 27, 556-573.
Sturt, P., Sanford, A. J., Stewart, A., \& E Dawydiak, E. (2004). Linguistic focus and good-enough representations: An application of the change- detection paradigm. Psychonomic Bulletin \& Review, 11, 882-888.
Swets, B., Desmet, T., Jr., C. C., \& Ferreira, F. (2008). Underspecification of syntactic ambiguities: Evidence from self-paced reading. Memory \& Cognition, 36, 201-216.
Townsend, D. J., \& Bever, T. G. (2001). Sentence comprehension: The integration of rules and habits. Cambridge, MA: MIT Press.
Williams, J. (1992). Processing polysemous words in context: Evidence for interrelated meanings. Journal of Psycholinguistic Research, 21, 193-218.
Wright, D., \& London, K. (2009). Multilevel modelling: Beyond the basic applications. British Journal of Mathematical and Statistical Psychology, 62, 439456.

# Individual Differences, Imagery and the Visual Impedance Effect 

Lupita Estefania Gazzo Castañeda (Estefania.Gazzo@psychol.uni-giessen.de)<br>Justus-Liebig University, Experimental Psychology and Cognitive Science, Otto-Behaghel-Str. 10F, 35394 Giessen, Germany<br>Markus Knauff (Markus.Knauff@psychol.uni-giessen.de)<br>Justus-Liebig University, Experimental Psychology and Cognitive Science, Otto-Behaghel-Str. 10F, 35394 Giessen, Germany


#### Abstract

The visual impedance effect describes the fact that unnecessary visual information can impede reasoning (Knauff \& Johnson-Laird, 2002). We explored how this effect is modulated by individual differences in reasoning styles. The main hypothesis of the present work is that the magnitude of the impedance effect depends on the degree to which people use visual mental images during thinking. We conducted two experiments with participants with highly imagistic and highly verbal reasoning strategies. The relational inferences differed in how easily they could be visualized. Our results indicate that (1) verbalizers do not show the visual impedance effect, and (2) that people with a high preference for mental imagery try to imagine even non-visual information visually, always showing the strongest impedance by visualization.


Keywords: Reasoning, individual differences, cognitive styles, visual impedance effect

## Introduction

Many theories have been developed to explain human reasoning (Clark, 1969; Johnson-Laird \& Byrne, 1991; Oaksford \& Chater, 2009; Rips, 1994). A problem common to these theories is that they often exclude the possibility of individual differences in reasoning (Bacon, Handley, \& Newstead, 2003; Ford, 1995). If people are asked how they solve reasoning problems, they usually report different ways of reasoning. While some people report the use of visual imagery (e.g. Egan \& Grimes-Farrow, 1982; Richardson, 1977), others report more language based approaches like rehearsal (Polk \& Newell, 1995), and yet others to think in a more abstract manner (Egan \& Grimes-Farrow, 1982). Based on such observations, Richardson (1977) proposed the differentiation between verbalizers and visualizers. Both were conceptualized as the extremes of a continuum. Visualizers were described as people with high visual imagery, but with poor verbal abilities, and verbalizers were described with the reverse tendencies. Over the years this dichotomy was expanded in accordance with newer neurological findings, and visualizers were divided into object- and spatial- visualizers (e.g. Kozhevnikov, Kosslyn, \& Shepard, 2005). While object visualizers are described as being able to construct vivid, high resolution images, spatial visualizers are described as being especially good in the processing of spatial information (Blazhenkova \& Kozhevnikov, 2009). Such differences in cognitive styles are important because unlike strategies, cognitive styles should be understood as relatively stable and durable
(Blazhenkova \& Kozhevnikov, 2009; Riding \& Cheema, 1991).

The influence of cognitive styles on tasks like anagrams and mental rotation (e.g. Just \& Carpenter, 1985) has already been investigated. Also, individual differences in spatial abilities and mechanical reasoning have been examined (Hegarty \& Sims, 1994). Nevertheless, only few studies analyzed the influence of individual differences on deductive reasoning (e.g. Bacon et al., 2003; Ford, 1995; Sternberg \& Weil, 1980). With the help of verbal protocols, Ford (1995) and Bacon et al. (2003) argued that people resolve syllogisms in two different ways. Some of the participants used a "verbal" strategy and resolved the syllogism via substitution of the terms. Other participants used a "spatial" strategy and resolved the syllogism with the help of schematic drawings which closely resembled Euler circles. However, almost no differences in performance were found (Bacon et al., 2003). Sternberg and Weil (1980) trained their participants to use either visual or rule based strategies in resolving relational inferences. One group of participants received no training. They found an interaction between skill and strategy: the effectiveness of the strategy depended on the verbal or spatial skills of the participant. More importantly, in the same study Sternberg and Weil found that a rule-based strategy lead to the fastest response times. Beyond these initial results, the question of the influence of individual differences based on imagery on other reasoning tasks like relational reasoning problems still remains open. This is surprising, insofar as the role of visual imagery on relational reasoning has been a topic of much controversy (Knauff, 2013).

## The Visual Impedance Effect

For a long time, the role of visual imagery during reasoning was not clear. While some researchers reported imagery as a helpful tool for reasoning (Clement \& Falmagne, 1986; Shaver, Pierson, \& Lang, 1975), others reported opposite results (Johnson-Laird, Byrne, \& Tabossi, 1989; Richardson, 1987; for a detailed review see Knauff 2013; Knauff \& Johnson-Laird, 2002). In search of clarification, Knauff and Johnson-Laird (2002) postulated that these discrepancies are based on a confounding in the items. Many items which are called "visual" are visual as well as spatial. Thus, in order to investigate the role of imagery during reasoning, it is important to disentangle the visual
from the spatial features of a given reasoning problem. By doing this, Knauff and Johnson-Laird (2002) showed that unnecessary visual information is an unnecessary cognitive load in working memory that leads to longer reaction times. They called this effect the visual impedance effect. They adapted their findings to the Theory of Mental Models (Johnson Laird \& Byrne, 1991) and proposed that mental models are spatial and not visual, as other groups propose (e.g. De Soto, London, \& Andel, 1965; Huttenlocher, 1968). This visual impedance effect has been corroborated in experiments with blind persons (Knauff \& May, 2006).

The visual impedance effect and the existence of more or less imagery-based cognitive styles motivated us to assume that people with different abilities in imagery should also perform differently in logical reasoning. We expected that the magnitude of the visual impedance effect would depend on the ability to use imagery during reasoning. Thus, the visual impedance effect should be increased for people with high visual imagery compared to those without a special preference for visual imagery or with a more linguistic cognitive style. Because of their cognitive style, this last group should hardly be affected by unnecessary visual information in tasks and thus show a better performance in relational problems compared to people with high visual imagery, especially in items with unnecessary visual information. To investigate these assumptions we conducted two experiments. In both experiments participants with different preferences for imagery had to solve relational problems. The content of these problems was manipulated in such way that problems were easy or hard to visualize.

## Experiment 1

In the first experiment, we measured the differences in cognitive style with a German version of the VerbalizerVisualizer Questionnaire (VVQ) from Richardson (1977). In a pilot study, we administered a German Version of the VVQ to 120 undergraduate psychology students at the University of Giessen. Using cut-off points at 5 and 12 points we found clearly distinguishable groups of verbalizers and visualizers (the scale ranged from 0 to 15 points).

## Method

Participants 22 participants (18 female, 4 male) from the pilot study participated in the experiment. Half of them were visualizers, the other half were verbalizers. The mean age was 22.8 years $(S D=4.55)$ and they participated for academic credit points.
Materials and Design We created 32 relational inferences. All of them described the same relation (left-right), but the term was either easy (fruits, tools, cutlery or office implements) or hard (nonsense syllables) to visualize. Half of the problems had valid conclusions; the other half had invalid conclusions. Here is an example for an easily visualizable problem with a valid conclusion:

Premise 1: Apple left of Kiwi
Premise 2: Kiwi left of pear
Conclusion: Pear right of apple
The design was a $2 \times 2$ design. The cognitive style of the participants was treated as a between-subjects factor. The ease of visualization for the terms was treated as a withinsubjects factor.
Procedure The experiment took place on a computer in a quiet room, and was programed in Cedrus SuperLab ${ }^{\text {TM }}$. The participants were tested individually. Premises and conclusions were presented on separate slides. The premises were written in black while the conclusion was written in red. The background was white. By pressing the space bar, participants decided when to pass from one premise to the next premise or to the conclusion. The task for the participants was to decide whether the conclusion was valid or not. They gave their decision by pressing one of two keys for "correct" or "false". Between each item, the participants had the opportunity to take a break. Before starting the actual experiment, the participants practiced on four items. To avoid learning effects the terms of these problems were abstract (the letters A, B, C). Dependent measures were premise reading times (not reported here), the mean number of logically correct responses, and the decision times for conclusion-evaluations.

## Results and Discussion

We first analyzed the percentage of correct responses ${ }^{1}$. Examining the problems that were easy to visualize, verbalizers responded correctly to $95.75 \%(S D=5.39)$ of them, while visualizers responded correctly only to $90.91 \%$ ( $S D=11.65$ ) (U-Test, $z=-1.095, p=.273$ ). Examining the problems that were hard to visualize, verbalizers responded correctly to $95.15 \%$ ( $S D=7.94$ ), while visualizers scored 92.73\% ( $S D=9.17$ ) (U-Test, $z=-.643, p=.520$ ). The main effect did not reach statistical significance, which is in accordance with our pervious results (Knauff \& JohnsonLaird, 2002). In the second step, we analyzed the decision times for correct responses. The results are illustrated in figure 1. As expected, visualizers $(M=6212 \mathrm{~ms}, S D=$ 1550) needed more time to resolve problems that were easy to visualize compared to verbalizers ( $M=4917 \mathrm{~ms}, S D=$ 1769). This effect was marginally significant (U-Test, $z=-$ 1.937, $p=.053$ ). For the problems that were hard to visualize, verbalizers ( $M=5700 \mathrm{~ms}, S D=2310$ ) were not significantly faster than visualizers ( $M=6053 \mathrm{~ms}, S D=$ 1703), (U-Test, $z=-.624, p=.533$ ). But contrary to what is implied by the visual impedance effect, decision times for problems that were hard to visualize were no smaller than the ones for problems that were easy to visualize. On the contrary, verbalizers were significantly slower solving problems that were hard to visualize compared to those that were not (Wilcoxon test, $z=-1.956, p=.050$ ). Visualizers showed no difference between both types of problems

[^274](Wilcoxon test, $z=-.533, p=.594)$. So there was something like a visual impedance effect for verbalizers, but not for visualizers. This unexpected result can be explained in two ways. One possible reason might be that nonsense syllables are not only hard to visualize, but also unknown and therefore probably also hard to memorize. This difficulty in memorizing the terms may have led to more cognitive load in working memory and thus to longer decision times compared to items which are known. This would explain the sudden increase in decision time for verbalizers. However, another reason for this unexpected result can be found in the reports many participants made after the experiment. Visualizers in particular, reported visualizing even the nonsense syllables. They reported that the nonsense syllables were visualized as phantasy creatures or names of foreign persons. Obviously, visualizer are so strongly biased towards using visual imagery, that even in tasks where no such visual information is available they transform the given information in such a way that they can use their typical visual thinking style. If so, then visualizers should not show the typical visual impedance effect, but instead of it something like a visual impedance effect on the subjects level. Thus visualizers should always have problems with relational problems, because all problems would be treated as highly visual problems. To test this hypothesis, it is important to use problems with familiar terms which are known, but which are still not easy to visualize. Therefore, in the second study we used the original material from Knauff and Johnson-Laird (2002).


Figure 1: Mean decision times for the conclusion. Error bars represent standard errors.

## Experiment 2

In the second experiment, we decided to measure the differences in cognitive style with the German version of Blazhenkova and Kozhevnikov's (2009) Object-Spatial Imagery and Verbal Questionnaire (OSIVQ). This questionnaire has the benefit that it accounts for the difference between visual imagery and spatial imagery. Thus the OSIVQ makes the same distinctions as the items used by Knauff and Johnson-Laird (2002).

In a pilot study we administered the German Version of the OSIVQ to 148 students at the University of Giessen. We selected our participants on the basis of their scores on the three scales contained within the OSIVQ: the visual scale, the spatial scale, and the verbal scale. We considered a participant as belonging to one of the three cognitive styles, if she or he scored above the sample mean of one scale, but below the sample means of the other scales ${ }^{2}$. Participants with higher deviations were preferred over those with fewer deviations. We selected 13 object visualizers, 6 spatial visualizers and 10 verbalizers. Additionally to these three experimental groups, we also selected a control group, whose scores on the scales did not differ from the sample scale means. The control group consisted of 10 participants.

## Method

Participants All selected participants from the pilot study participated in the experiment. All object visualizers were female ( $n=13$ ), with a mean age of $M=22.54(S D=2.3)$. The group of the spatial visualizers consisted of 3 female and 3 male participants. Their mean age was $M=22.67$ ( $S D$ $=3.14$ ). The group of the verbalizers consisted of 9 female and 1 male participant. Their mean age was $M=22.8$ ( $S D=$ 8.16). Finally, the control group consisted of 7 female and 3 male participants. Their mean age was $M=22.6$ ( $S D=$ 2.59). The participants participated for academic credit points or candies.
Materials and Design We used the relational inferences from Knauff and Johnson-Laird (2002; see table 1). These relations have been evaluated empirically (Knauff \& Johnson-Laird, 2002) and differ in the relative degree to which they can be imagined either visually or spatially. Using these relational terms it is possible to create 32 items. All items consisted of the same terms (dog, cat, ape). Again, half of the problems had valid conclusions; the other half had invalid conclusions. An example for a valid visual problem is:

Premise 1: The dog is cleaner than the cat
Premise 2: The ape is dirtier than the cat
Conclusion: The ape is dirtier than the dog
The design was a $4 \times 4$ design. The cognitive style of the participants was treated as a between-subjects factor. The ease of visualization was treated as a within-subjects factor.

Additionally to the inference task, we also measured spatial, verbal and visual abilities of the participants. The idea was to validate the cognitive style of our participants. A similar procedure was also used by the developers of the OSIVQ (Blazhenkova \& Kozhevnikov, 2009). As a measure for visual abilities we used the Vividness of Visual Imagery Questionnaire (VVIQ: Marks, 1973). It consists of 16 items which examine how easily the participant is able to imagine, visually and vividly, different scenes with open and with closed eyes. As a measure of spatial ability we used a

[^275]mental rotation task, based on the one by Shepard and Metzler (1971). It consisted of the presentation of a target 3D figure, in combination with another similar figure which was either the same in one of six rotated degrees, or a rotated mirror image. The task for the participant was to decide whether both images were the same or not. This task consisted of 48 items. Finally, as a measure of verbal ability we used the subtest "Masselon" from the Berliner Intelligenzstruktur Test (BIS: Jäger, Süß, \& Beauducel, 1997). In this test the participant is confronted with three words (human, feeling, technology) and must then create as many sentences as they can with these three words. For better comparisons with experiment 1, we also administered a German Version of Richardson's (1977) VVQ.
Procedure The experiment always began with the relational inference task, which was programed in SuperLab ${ }^{\text {TM }}$. The procedure for the relational inference task was the same as experiment 1. After completing the relational task, the VVIQ tasks, the rotation task, and the Masselon tasks were presented in a random order. Finally, the participants answered the VVQ and provided written comments on how they believed they solved the tasks. Again, we measured the reading time for each premise (not reported here), the decision time for the conclusion, and whether the task was solved correctly or not.

Table 1: Relations used in Experiment 2, with a description of how easy they were to imagine either as a visual image or a spatial array (adapted from Knauff and Johnson-Laird (2002; p. 368)).

| Relations | Description |
| :---: | :---: |
| Visual |  |
| cleaner-dirtier fatter-thinner | Ease to envisage visually, but hard to envisage spatially |
| Control |  |
| better-worse smarter-dumber | Hard to envisage visually and spatially |
| Visuospatial above-below front-back | Easy to envisage visually and spatially |
| Spatial |  |
| north-south ancestor-descendant | Hard to envisage visually, but easy to envisage spatially |

## Results and Discussion

VVIQ, mental rotation, and verbal abilities The results from the three tasks indicate that our selection of the exponents of the different cognitive styles was successful. The VVIQ was computed in such a way that high scores (max. 80 points) indicated good visual abilities, whereas low scores (min. 16 points) indicated a lack of it. Object visualizers ( $M=51.08, S D=5.61$ ) reached a higher score than verbalizers ( $M=39.75, S D=8.72$ ) on the VVIQ (UTest, $z=-3.072, p=.002$ ). They also reached a higher score than spatial visualizers $(M=40.83, S D=13.45)$ and the control group ( $M=43.70, S D=6.87$ ), but these last two
differences did not reach the adjusted alpha level (U-Test, $z$ $=-1.931, p=.053$ for the comparison with spatial visualizers; U-Test, $z=-2.576, p=.01$ for the comparison with the control group).

Even if there was no significant main effect on the time needed to solve the items in the mental rotation task (Kruskal Wallis, $C h i^{2}=5.440, p=.142$ ), descriptively it was possible to see that, across all items, spatial visualizers ( $M=4.99 \mathrm{~s}, S D=1.88$ ) were faster than verbalizers $(M=$ $7.54 \mathrm{~s}, S D=3.87$ ), than object visualizers ( $M=8.15 \mathrm{~s}, S D=$ 2.85) and the control group ( $M=6.80 \mathrm{~s}, S D=3.41$ ). Poltrock and Brown (1984) proposed that the linear regression slopes of the latencies are an indicator of the rotation speed, in that the smaller the slope, the faster the rotation was performed. As expected, spatial visualizers rotated faster ( $b=29.55, S E=8.39$ ) than verbalizers $(b=$ 72.00, $S E=16.42$ ), object visualizers ( $b=55.79, S E=9.75$ ) and the control group ( $b=48.16, S E=16.77$ ). The groups did not differ in the amount of errors made (Kruskal Wallis, $C h i^{2}=2.641, p=.450$ ).

Our analysis of the Masselon test was based on the amount of written words in valid sentences. Verbalizers ( $M$ $=43.30, S D=11.37$ ) wrote significantly more words than object visualizers $(M=30.54, S D=7.93$; U-Test, $z=-2.86$, $p=.004$ ) and spatial visualizers ( $M=25.17 ; S D=7.11$; UTest, $z=-2.71, p=.007$ ), but they did not differ from the control group ( $M=34.70, S D=12.60$; U-Test, $z=-1.34, p$ $=.182$ ).
The VVQ The comparison of the scores on the VVQ showed that the VVQ is only able to differentiate correctly between verbalizers ( $M=7.20, S D=2.35$ ) and object visualizers ( $M=10.46, S D=2.22$; U-test, $z=-2.875, p=$ .004). Given that the VVQ does not consider spatial visualizers, spatial visualizers ( $M=7.67, S D=3.98$ ) scored similar to verbalizers (U-test, $z=-.174, p=.869$ ) on the VVQ.
Relational inferences Similar to Knauff and Johnson-Laird (2002), we encountered some problems with the spatial relations. On the one hand, several participants reported that they still imagined them in a visual way (e.g. as animals on maps). On the other hand, the spatial terms created "illogical" constellations (e.g. the dog is descendant of the cat), whose difficulty probably confounded the decision times. Therefore, the spatial relational terms were not purely spatial and had to be removed from our analysis. The reported results are based solely on the correct responses to the other three kinds of relational terms.

Based on the results of experiment 1, we assumed that people with a high preference for imagery would not only have difficulties in visual relational problems, but also in non-visual relational problems. Because of their cognitive style, object visualizers would try to imagine even non visual information visually, regardless of how difficult this is, and be impeded by this visualization. To analyze these hypotheses we first considered the percentage of errors made by our participants. As in experiment 1, there were no significant differences between the participants with
different cognitive styles in the percentage of correct answers (Kruskal Wallis, $C h i^{2}=2.060, p=.560$ ). However, there was a significant main effect in the response times when we compared the groups of object-visualizers, spatial visualizers, and verbalizers (Kruskal Wallis, $C h i^{2}=6.855, p$ $=.032$ ). Pairwise comparisons showed that, as expected, object visualizers ( $M=4608 \mathrm{~ms}, S D=2671$ ) took longer to resolve the tasks compared to verbalizers, ( $M=2777 \mathrm{~ms}$, $S D=645$ ). This difference was also significant (U-test, $z=-$ 2.481, $p=.012$ ). Object visualizers were not significantly slower than spatial visualizers ( $M=3857 \mathrm{~ms}, S D=1159$; Utest, $z=-0.614, p=.579$ ) and verbalizers still tended to answer faster than spatial visualizers (U-test, $z=-1.735, p=$ .093). As can be seen in figure 2, neither object visualizers nor verbalizers showed the classical visual impedance effect. In neither group a main effect on decision times for the different kinds of items could be found (Friedman Test, $C h i^{2}=1.077, p=.584$ for object visualizers; $C h i^{2}=.200, p$ $=.905$ for verbalizers). The only group that showed a pattern resembling the classical visual impedance effect were the spatial visualizers. However, because of the small sample size ( $n=6$ ), the main effect did not reach significance (Friedman Test, $C h i^{2}=3.000, p=.223$ ). This trend was not expected and should be investigated in further studies. Nevertheless, the missing visual impedance effect for verbalizers and the long decision times of object visualizers confirm our suppositions derived from experiment 1: while object visualizers do indeed try to visualize even nonvisual information, verbalizers never visualize anything. This leads to a lack of visual impedance effects on the item level, but instead causes visual impedance effects on the subject level.


Figure 2: Mean decision times for the conclusion. Error bars represent standard errors.

## General Discussion

Our findings indicate that, depending on their cognitive style and how easily they are able to use imagery during reasoning, people are influenced in different ways by the imageability of the content of reasoning problems: On the one hand, verbalizers are typically not impeded by visual
characteristics of reasoning problems. They seem to be immune to the visual impedance effect. On the other hand, people who tend to imagine the content of reasoning problems try to visualize even non-visual problems and therefore show a visual impedance effect on all problems, whether the problems are highly visual or not. These results notwithstanding, we are aware of the limitations of our study. In order to strengthen our results, it would be necessary to conduct studies with a greater sample size and to control for gender differences. An additional task for the future is to investigate during which phase of the inference individual differences take effect. Particularly, it remains unclear whether these individual differences play a role only during interpretation and encoding of the reasoning problem premises, or if the individual differences also have an effect on the reasoning process itself. Knauff (2009, 2013) proposes that relational reasoning problems are solved in three steps among which only the first step involves the construction of visual mental images and the other two steps comprise the "real" reasoning processes. However, in this work we did not distinguish between different cognitive styles and so it is still unclear how verbalizers solve such tasks. Do they also create such (irrelevant) initial picturelike representations? One approach that we took previously is to use functional brain imaging to study the neural basis of individual differences in reasoning (Ruff, Knauff, Fangmeier, \& Spreer, 2003). By testing people with different cognitive styles in the scanner it might be possible to see during which phases of the reasoning process these cognitive styles take effect. By doing this, it would also be possible to see to what extent the steps proposed by Knauff (2009, 2013) are generalizable to all cognitive styles. The same could be also done with other individual differences. For example, it might be of interest to investigate whether people with either a holistic or an analytic cognitive style (see Riding \& Cheema, 1991) differ in the construction of mental models.

Another task for the future is to replicate the present results using different formats of presentation. In both studies reported here our items were presented in written form. Considering that verbalizers are often described as having fun and being good at reading and language based tasks (see the relevant items of the VVQ and the OSIVQ), it seems possible that the superior performance of verbalizers resulted not only because they did not use imagery, but also because verbalizers might feel more comfortable with a task presented in their preferred format. Thus, in further studies it is important to present items in other formats, for example acoustically or in an iconic way.

In conclusion, our results support the visual impedance effect. Irrelevant visual details can be a nuisance in reasoning. However, the effect seems to be modulated by the different cognitive styles of individuals. Object visualizers are so profoundly driven by their visual thinking style that they try to visualize almost everything. Thus they show a visual impedance effect even for non-visual reasoning problems. Verbalizers, in contrast, are only
marginally affected by the visual characteristics of reasoning problems. They use more abstract reasoning styles and therefore have no problems with disruptive visual images. We were also able to identify differences between object visualizers and spatial visualizers. Comparing both groups, our findings indicate that the use of spatial representations and processes is the most effective way to solve relational reasoning problems. However, individuals using spatial layout models (Knauff, 2013) seem not to be immune to irrelevant and side-tracking visual details and can therefore be impaired in solving highly visual inference problems. We will continue to explore this effect more thoroughly. A final important corollary of our study is that effects found in general populations (without considering differences in cognitive style) do not necessarily apply to every single person: visual items do not always impede reasoning, they only impede if subjects represent visual features in their mental representation of the task. That is why it is important to incorporate individual differences into theories of reasoning and to highlight such differences in the predictions and assumptions of those theories. Disregarding these differences may lead to unjustified overgeneralizations.

## Acknowledgments

This work was supported by DFG grant KN 465/10-1 to Markus Knauff.

## References

Bacon, A. M., Handley, S. J., \& Newstead, S. E. (2003). Individual differences in strategies for syllogistic reasoning. Thinking and Reasoning, 9, 133-168.
Blazhenkova, O., \& Kozhevnikov, M. (2009). The new object-spatial-verbal cognitive style model: Theory and measurement. Applied Cognitive Psychology, 23, 638663.

Clark, H. H. (1969). Linguistic processes in deductive reasoning. Psychological Review, 76, 387-404.
Clement, C. A., \& Falmagne, R. J. (1986). Logical reasoning, world knowledge, and mental imagery: Interconnections in cognitive processes. Memory \& Cognition, 14, 299-307.
De Soto, L. B., London, M., \& Andel, M. S. (1965). Social reasoning and spatial paralogic. Jorunal of Personality \& Social Psychology, 2, 513-521.
Egan, D. E., \& Grimes-Farrow, D. D. (1982). Differences in mental representations spontaneously adopted for reasoning. Memory \& Cognition, 10, 297-307.
Ford, M. (1995). Two modes of mental representation and problem solution in syllogistic reasoning. Cognition, 54, 1-71.
Hegarty, M., \& Sims, V. K. (1994) Individual differences in mental animation during mechanical reasoning. Memory and. Cognition, 22, 411-430.
Huttenlocher, J. (1968). Constructing spatial images: A strategy in reasoning. Psychological Review, 75, 550-560.

Jäger, A. O., Süß, H. -M. \& Beauducel, A. (1997). Berliner Intelligenzstruktur - Test. Form 4. Göttingen: Hogrefe.
Johnson-Laird, P. N., \& Byrne, R. M. J. (1991). Deduction. Hove, U.K.: Erlbaum.
Johnson-Laird, P. N., Byrne, R. M. J., \& Tabossi, P. (1989). Reasoning by model: The case of multiple quantification. Psychological Review, 96, 658-673.
Just, M. A., \& Carpenter, P. A. (1985). Cognitive coordinate systems: Accounts of mental rotation and individual differences in spatial ability. Psychological Review, 92, 137-172.
Knauff, M. (2009). A neuro-cognitive theory of deductive relational reasoning with mental models and visual images. Spatial Cognition and Computation, 9, 109-137.
Knauff, M. (2013). Space to Reason. Cambridge, MA: MIT Press.
Knauff, M., \& Johnson-Laird, P. N. (2002). Visual imagery can impede reasoning. Memory \& Cognition, 30, 363371.

Knauff, M., \& May, E. (2006). Mental imagery, reasoning, and blindness. Quarterly Journal of Experimental Psychology, 59, 161-177.
Kozhevnikov, M., Kosslyn, S., \& Shephard, J. (2005). Spatial versus object visualizers: A new characterization of visual cognitive style. Memory and Cognition, 33, 710726.

Marks, D. F. (1973). Visual imagery differences in the recall of pictures. British Journal of Psychology, 64, 17-24.
Oaksford, M., \& Chater, N. (2009). Précis of bayesian rationality: The probabilistic approach to human reasoning. Behavioral and Brian Sciences, 32, 69-120.
Polk, T. A., \& Newell, A. (1995). Deduction as verbal reasoning. Psychological Review, 102, 533-566.
Poltrock, S. E., \& Brown, P. (1984). Individual differences in visual spatial ability. Intelligence, 8, 93-138.
Richardson, A. (1977). Verbalizer-Visualizer: a cognitive style dimension. Journal of mental Imagery, 1, 109-126.
Richardson, J. T. E. (1987). The role of mental imagery in models of transitive inference. British Journal of Psychology, 78, 189-203.
Riding, R., \& Cheema, I. (1991). Cognitive styles - an overview and integration. Educational Psychology, 11, 193-215.
Rips, L. J. (1994). The psychology of proof. Cambridge, MA: MIT Press.
Ruff, C. C., Knauff, M., Fangmeier, T., \& Spreer, J. (2003). Reasoning and working memory: Common and distinct neuronal processes. Neuropsychologia, 41, 1241-1253.
Shaver, P., Pierson, L., \& Lang, S. (1975). Converging evidence for the functional significance of imagery in problem solving. Cognition, 3, 359-375.
Shepard, R. N., \& Metzler, J. (1971). Mental rotation of three-dimensional objects. Science, 171, 701-703.
Sternberg, R. J., \& Weil, E. M. (1980). An aptitude X strategy interaction in linear syllogistic reasoning. Journal of Educational Psychology, 72, 226-239.

# Playing for Us: The Influence of Joint Action on Planning in Three-year-olds 

Sarah A. Gerson (s.gerson@donders.ru.nl)<br>Sabine Hunnius (s.hunnius@donders.ru.nl)<br>Harold Bekkering (h.bekkering@donders.ru.nl)<br>Radboud University Nijmegen<br>Donders Institute for Brain, Cognition, and Behaviour, Montessorilaan 3<br>Nijmegen, 6500HE, The Netherlands


#### Abstract

Learning to plan sequences of actions and appropriately adapt our actions during interactions with others are both critical skills upon which much of human society is built. We know that children's joint action and planning skills are both undergoing development during the preschool years, but not much is known about how the joint action context influences young children's planning. In this study, we examined the effect of playing alone or with a joint partner on sequence planning during a problem-solving game in three-year-old children. We found that children were better at planning ahead in the individual than the joint condition of the game despite the joint condition requiring fewer actions on the part of the child. In contrast, children were equally good at problem-solving (i.e., correcting an error) in both conditions. The possible reasons for this difference and directions of future research are discussed.


Keywords: joint action; planning; cognitive development
Working together with others is important across a variety of everyday tasks, ranging from simple, mundane actions to considerably complex plans and action sequences. When we interact with a partner in a work or athletic setting, the complexity of coordinating our actions with another's is quite clear. In contrast, when we perform simple everyday actions such as passing a cup of milk to another person, we likely do this with ease and do not dwell on the coordination with the other or the expectations about the others' action. When acting with another person, planning our own actions requires coordinating our actions with another individual, whether this coordination is conscious and complex or seemingly automatic. Planning our actions when interacting with another is a task that spans many domains and is critical for much of cognitive and social development. Examining the developmental emergence of this skill can shed light on how and when the factors necessary to working with others are integrated.

When performing a task by ourselves, we can create a plan internally and carry out the task without interruption. When jointly acting with another, however, we need to take the other person's actions into account. According to Sebanz and Knoblich (2009), intentional coordination of actions with another requires representing both one's own and one's partner's roles in the task. They suggest that adults engaged in joint actions predict a partner's actions in a joint action task by representing the action of a partner and one's own actions in a functionally equivalent way. In fact,
incorporating a partner's task "affects one's own action planning and performance even when there is no need to take the other's part into account at all" (p. 357). One mechanism thought to underlie the representation and prediction of another's actions is simulation (Gallese \& Goldman, 1998). That is, when one perceives someone else acting in a goal-directed manner, one's own motor system is activated as if one was performing the action oneself (Rizzolatti \& Craighero, 2004). Simulating a partner's action from a first-person perspective can then be used to make predictions about upcoming events (Wilson \& Knoblich, 2005). Additionally, the motor system is preferentially activated for predictions of others' actions within a joint action context (Kourtis, Sebanz, \& Knoblich, 2010). What is simulated and how perceptual information available can be transferred into a motor simulation is still a topic of vivid discussion (see for example Uithol et al., 2011).

The necessity of incorporating another agent's actions in a similar way to one's own actions when interacting with a joint partner suggests that the ability to represent other agents' actions in a similar way to one's own would be a developmental prerequisite for appropriately planning one's actions within a joint context. One piece of behavioral evidence that young children seem to represent others' actions in a similar way to their own actions is that infants' ability to produce particular actions is directly related to their perception, prediction, and motor activation when viewing others perform the same actions (e.g., Cannon et al., 2012; Gerson \& Woodward, in press, van Elk et al., 2008). Meyer and colleagues (2011) found neural evidence that this is especially so in joint action contexts. Greater activation in the motor system was found in three-year-old children watching a joint action partner than when these same children watched someone with whom they were not collaborating. Further, variation in performance on the joint game and in the amount of motor activation observed when the child watched the partner act were related, suggesting that the child's motor system activation was likely related to the integration of their partner's and their own actions.

In addition to a representation of others' actions, the incorporation of others' actions into one's own planning is critical to acting appropriately in joint contexts. In order to address how the presence of others affects planning, research must examine differences in planning one's own actions during individual and joint tasks. A recent study
with adults (Meyer, van der Wel, \& Hunnius, 2013) measured planning of actions that could be performed alone or with another person. It was found that participants learned to initiate actions based on predictions about the subsequent steps in a task after they gained experience acting in the task. This was true in both the individual and joint contexts, suggesting that participants were able to use their experience to predict their own or a partner's actions and plan their actions accordingly.

The research reviewed above indicates that motor activation during the observation and prediction of others' actions is heightened within joint action contexts and that the simulation of others' actions facilitates motor planning in joint contexts. Although motor planning is one important aspect of planning sequences of actions, sequence planning also requires higher-order processes such as future thinking and cognitive control. That is, when performing an initial action that propagates a series of embedded actions, one must plan not only the motor aspects (such as movement, timing, and spatial location) but also consider the consequence of these actions on the future steps in the sequence. Adults are proficient sequence planners, but planning skills are still undergoing development throughout early childhood (Carlson, Moses, \& Claxton, 2004; McCormack \& Atance, 2011). Difficulties in planning and other higher-order cognitive skills have been linked to the relatively prolonged development of the prefrontal cortex (see, for example, Welsh, Pennington, \& Groisser, 1991).

Previous research examining the development of sequence planning within joint action contexts has largely measured children's planning when engaged in a game with a parent or another adult. These studies have found that the development of planning with others is a prolonged process, in that older children (e.g., between 6 and 11 years) often outperformed younger children (e.g., between 3 and 5 years) on planning tasks (e.g., de la Ossa \& Gauvain, 2001; Gauvain, 1992; Gauvain \& Rogoff, 1989). This research, however, focused largely on the role parents played in guiding the joint actions through bids for joint attention, scaffolding of the child's actions, and teaching of strategies or rules. Because parents were involved and influencing children's actions during the joint planning games, measures of the child's planning skills were often measured after the joint task. The lack of planning measurements during joint actions does not take into account whether planning in a joint context adds more cognitive demands to a planning task. In the current study, we explore the planning skills of three-year-old children during a problem-solving task when playing alone or with a partner who acts in a predictable, uniform manner.

We created a game in which the child was required to plan ahead in order to accurately solve a matching game. If he or she did not plan ahead, the child had the chance to correct the error during a subsequent step of the game. All children played this game both alone and in alternating turns with a joint partner, "Kip." The joint partner was a hand puppet introduced during the joint action condition and kip
always acted predictably so that we could assess the influence of a social partner's presence without the social partner's actions directly influencing any of the child's actions. Kip was introduced as separate from the experimenter and the experimenter used a different voice when acting as Kip so that the child did not expect Kip to scaffold his or her actions. We then examined differences between children's accuracy in planning and problemsolving during the individual versus joint conditions. If simulating a person's actions in order to motorically plan one's own actions is the key difference between individual and joint planning, then children's performance during the joint condition should not be hindered. In fact, because children took turns playing with Kip, the joint condition required less motor planning than the individual condition; children only had to place two balls in the correct buckets during each trial instead of four. Therefore, if all planning was carried out through the motor system, children's planning should be better in the joint condition than the individual condition. If, however, other cognitive processes are necessary in order to integrate one's own plans with another person's actions, plans, and goals, then children should perform worse in the joint condition than in the individual condition. That is, if the presence of another actor increases the cognitive demands of higher-order functioning, such as cognitive control, future thinking, and sequence planning, children should perform better in the individual condition than the joint condition.

## Method

## Participants

Thirty-two 37 -month-olds (mean age $=3$ years, 38 days) were included in the final data set for this study ( 15 females, 17 males). All children were recruited from a database of families who volunteered to participate in child studies. An additional 10 children participated but were not included due to equipment malfunction ( $n=2$ ), experimenter error ( $n$ $=2$ ), not completing all trials $(\mathrm{n}=3)$, or lack of learning of the rules of the game or refusal to play with Kip ( $n=3$ ).

## Stimuli and Procedure

Each trial consisted of a set of four balls, four buckets, and a clear, plastic tube that held the balls. There were always two buckets of one color (e.g., green) and two buckets of another color (e.g., yellow). In all but the first training trial, there were two balls of one color (e.g., green), one ball of a second color (e.g., yellow), and one ball that was multicolored (e.g., half green and half yellow). The tube was created to dispense the balls one at a time in a particular order while still allowing participants to see the colors of the upcoming balls (see Figure 1). The multi-colored ball always came out of the tube in the second position, and the three solid-colored balls were pseudorandomly distributed in the first, third, and fourth positions. Except in the demonstration trial, different color combinations (consisting of red, light blue, dark blue, green, and yellow) were used
across trials so as to minimize learning specific rules about colors and to keep the children's attention. In joint play trials, the experimenter wore a hand puppet of a chicken (called "Kip"). The experimenter used a different voice so as to differentiate herself from the puppet.


Figure 1: Example of the game setup. Each trial involved three-solid colored balls, a multicolored ball, and two buckets in each of two colors.

Training Children were taught how the game worked via a set of training trials. First, the experimenter placed a set of four solid-colored balls (brown and black) into matching buckets. This short phase was to teach children that balls had to go into matching buckets. Next, one of the solid balls (the one in the second position) was replaced with a multicolored ball. When the experimenter extracted the multicolored ball, she showed the child that it could go in either the brown or the black bucket. After showing them this, she always left the ball in the inappropriate bucket in terms of meeting the end goal. That is, if there were two brown balls in the tube, the multi-colored ball would be placed in a brown bucket (and vice-versa if there were two black balls). This "mistake" was made in order to show participants the importance of considering the upcoming balls in the tube and to indicate how errors could be corrected. The experimenter then placed a black and brown arrow in front of the bucket to indicate which bucket held the multicolored ball. After the incorrect placement of the multicolored ball, the experimenter would show the child that one of the remaining solid-colored balls no longer had an appropriate bucket in which to be placed. She would talk to the child about how this could possibly be fixed and remind them about the meaning of the arrow and hint about a possible solution: "Do you remember what this arrow means? This means that the multi-colored ball is in this bucket. And where can the multi-colored ball go?" She would then extract the multi-colored ball and place it in the opposite colored bucket. She moved the arrow to the new bucket and then placed the solid-colored ball in the appropriate bucket. After having done this, she would remind the child of how the problem had been solved.

Two training trials followed this demonstration in which the experimenter scaffolded the child throughout the game. These two trials consisted of two different sets of colored balls, randomly assigned. During these trials, the experimenter handed the participant each of the balls and asked him or her to place them in the matching bucket. She
frequently reminded the child that all the balls had to "fit" in the buckets (and pointed to the balls in the tube). If the child struggled, the experiment gave a series of hints. If the child encountered a solid-colored ball that had no matching bucket, the experimenter first gave him or her time to try to solve the problem themselves. Then she gave the participant a series of hints, allowing time for the child to recognize the solution between each hint. As in the demonstration trial, hints increased in detail, ranging from asking what the arrow meant to reminding the child that the multi-colored ball could go in either bucket. If the child still did not respond to the hints, the experimenter moved the mixed ball and demonstrated the solution to the problem. In this way, at the end of the training trials, the experimenter always ensured that the balls were matched with an appropriate bucket at the end of the trial. After these two trials, the experimenter told the child he or she was ready to play without help. Individual or joint play trials then began (counterbalanced between participants).

Individual Play The individual condition consisted of six trials. In each of these trials, the child retrieved each ball from the tube, one at a time, and placed it into a bucket. The experimenter did not participate except to ensure that the child did not retrieve the following ball before placing the one in his or her hand into a bucket. If the child encountered a problem (i.e., a solid-colored ball without a matching bucket), the experimenter did not interfere unless the child looked to the experimenter for help. When the child expressed uncertainty and enquired for help, the experimenter would give the same hints as during the training trial, again giving the child time to solve the problem at each step. After all of the balls were placed in buckets, the experimenter asked the child if they were all correct (regardless of whether or not they were). If the child realized then that there was a problem, the experimenter again only helped (as above) if the child enquired.

Joint Play First, a small hand puppet was introduced to the child. The child was told the name of the puppet (Kip) and that Kip wanted to play with him or her and they could take turns (see Figure 2). The joint play session consisted of nine trials. In the first, fourth, and seventh trial, Kip let the child place the first (and third) ball and Kip placed the second/multi-colored (and fourth) ball. Kip always placed the multi-colored ball in the bucket that allowed all forthcoming balls to be placed correctly. In the other six trials, Kip placed the first and third balls and the child placed the second and fourth balls. This ensured that the number of trials for which the child had to plan (by placing the multi-colored ball correctly) was matched across the individual and joint conditions. If the child incorrectly placed the multi-colored ball and realized this error when later attempting to place a solid-colored ball, the experimenter followed the same procedure as in the individual play trials as far as waiting for the child to enquire in order to give any hints. If Kip had to place the
solid-colored ball that had no matching bucket, she would knock on the full buckets and say "uh oh-this ball can't go in this one" while looking at the empty bucket and would ask for the child's help. If the child did not immediately solve the problem, the experimenter followed the same pattern for giving hints as in other trials.


Figure 2: During the joint action condition, children alternated taking turns with Kip, the hand puppet.

Coding The focal question in this study concerned children's ability to plan where to place the multi-colored ball so that all following balls could fit in matching buckets. For each trial in which the child placed the multi-colored ball (six individual play and six joint play trials), a trained coder judged whether the child placed the multi-colored ball in the correct bucket (for the end goal achievement) before the following ball was retrieved from the tube. This factor will be referred to as planning. The proportion of trials within the individual and joint condition for which the child's planning was correct was calculated and used as a dependent variable. A second question was whether children would correct errors if their initial ball placement was incorrect. For this factor (called problem solving), coders judged whether the child removed the mixed ball and placed it in a correct bucket. If so, the coder noted whether the child carried out this action with or without needing the assistance of hints from the experimenter. The proportion of trials correct after problem solving without hints from the experimenter were calculated for each condition (note: this gave children credit both for initially correct and correctly solved trials without assistance). A second trained coder coded $25 \%$ of the videos and agreed on $99 \%$ of trials.

## Results

As described above, the variable of interest for planning was the proportion of trials for which children were initially correct in their placement of the multi-colored ball and the variable of interest for problem solving was the proportion of trials in which the child had correctly placed all balls (without hints) by the end of the trial. Initially, we conducted a repeated-measures analysis of variance (ANOVA) with Condition (i.e., individual or joint play) and Solution Stage (planning vs. problem-solving) as within subjects factors. The between-subjects counterbalancing factor of Order (i.e., whether the child participated in the
individual or joint condition first) was also included to account for possible learning effects across time. This analysis revealed a main effect of Solution Stage $(F(1,30)=$ 93.33, $p<.001, \mathrm{n}_{\mathrm{p}}^{2}=.76$ ), a Solution Stage X Condition interaction $\left(F(1,30)=5.15, p=.031, \eta_{\mathrm{p}}^{2}=.15\right)$. No other main effects or interactions were found ( $p s>.13, \eta_{p}{ }^{2} \mathrm{~s}<$ .08). The main effect of Solution Stage indicated that the proportion of trials that children successfully planned was significantly lower than their problem solving performance. The interaction suggests that the extent of this difference was affected by condition (individual vs. joint). The lack of main effect or interactions with Order suggests that children who engaged in the joint versus individual task first did not differ from one another in their performance.

In order to follow up on this interaction, we examined pairwise comparisons of estimated marginal means. The difference between individual and joint conditions was significant for planning ( $m d=.11, S E=.048, p=.031$; see Figure 3) in that children were significantly better at planning during the individual than the joint condition. This difference between conditions was not present for problemsolving ( $m d=.001, S E=.035, p=.98$ ). That is, children were equally able to solve the problem in both conditions. Additionally, children performed significantly better during problem-solving than planning within both individual and joint conditions ( $p \mathrm{~s}<.001$ ).

In order to examine planning and problem-solving performance relative to chance levels ( $50 \%$ of trials correct), we conducted one-sample t-tests. In the individual condition, children were better at planning than would be expected by chance ( $M=.61, S E=.028, t(31)=3.95, p<$ .001 , Cohen's $d=1.42$ ). Children were not above chance levels of planning in the joint condition ( $M=.50, S E=.037$, $t(31)=.034, p=.97$, Cohen's $d=.012$ ). When children had the opportunity to correct their errors (i.e., problem solve), they performed at above chance levels in both conditions ( $t \mathrm{~s}$ $>12.3, p \mathrm{~s}<.001$ ).


Figure 3: Children were significantly better at planning in the individual than joint condition ( ${ }^{*} p=.031$ ), but were above chance in problem solving in both conditions.

## Discussion

Children were significantly better at planning their actions appropriately when they played alone than when they took turns playing with a social partner. That is, when playing
alone, they were more likely to take into account the colors of the remaining balls when choosing where to place the mixed ball. When playing with a partner, children's initial placement of the mixed ball was seemingly random (i.e., the placement was correct about half the time [at chance level]). Importantly, this was true despite the fact that children had fewer actions to carry out during the joint condition. In the individual condition, children were responsible for placing all four balls correctly. In the joint condition, however, children only needed to place two of the four balls. The joint partner always played correctly on her trials, so the task of placing half the balls should have, in principle, been easier. The fact that children did not perform as well in this case suggests that something about sharing the task with a partner made it more difficult for the children to plan. That is, motor planning alone was not sufficient for carrying out the task; the demands of sequence planning were made more difficult by the presence of another actor.

In contrast to the difference found in planning, when children encountered a proceeding ball for which there was no matching bucket, they were equally competent at solving this problem regardless of whether they were playing alone or with a partner. The fact that children could and did solve the problem without hints from the experimenter (or Kip) in both conditions suggests that children understood the goal of the task and what actions were necessary in order to achieve this goal. Thus, it was not a lack of understanding of the task that prevented children from planning appropriately during the joint condition. This is impressive given the complexity of the task carried out by the children.

Further, children's planning and problem solving did not change as a function of the order in which they played the individual and joint conditions. This indicates that children did not learn the task over time, regardless of which condition they played first. Additionally, the fact that children who played the joint condition first did not plan more effectively during the individual condition than children who played the individual condition first suggesting that children were not learning how to plan from Kip's turns placing the mixed ball. Given that Kip always placed the mixed ball correctly (on the three trials in which she placed this ball), it was possible that children could have used their partner's correct actions to improve their own planning, but the lack of order effect suggests this was not the case in this study.

An important question to address in future studies is why children were better able to plan during the individual than the joint condition. Several possibilities remain to be examined, including aspects of attention, inhibition, and the social nature of the task.

One possibility is that attention to the future balls to be placed differed when children were playing alone or with Kip. If attention does differ, it suggests that the presence of a partner made it more difficult for children to concentrate on the task at hand and control their attention according to the task goal. Baron (1986) has suggested that the presence of others causes shifts in cognitive processing. This might
be particularly true during early development when attentional control is still developing.

Similarly, children may have struggled to maintain attentional control because of the timing differences between the two task conditions. That is, children could play continuously during the individual condition of the task but were required to pause their own play while their partner acted during the joint condition. It is possible that, it was not simply the presence of the other, but the fact that the child's play was interrupted that made planning more difficult. Whether the break in play led to disrupted attention control or directly to difficulty with planning is unclear, and may be driven by other mechanisms such as inhibitory control or working memory. Ongoing studies in our laboratory aim to address this possibility.

Finally, the mere presence of a social partner, rather than the pauses in play or attention, may have undermined children's planning. Sebanz, Knoblich, and Prinz (2003) suggest that the presence of others influences task performance, regardless of whether one is acting with the other person. They argue, "social facilitation effects are not moderated by the specific actions carried out by others" (p. 12). Instead, they suggest that the presence of another improves performance on simple tasks but impairs performance on more complex tasks. This possibility would be interesting to explore developmentally because of shifts in complexity of particular tasks as children gain both domain-general and domain-specific skills.

The current findings shed light on the difficulties encountered when first attempting to incorporate predictions of a partner's actions with one's own planning. It suggests that planning for two individuals, even when they share a common goal, is more difficult than planning for oneself. The relative complexity of the planning task in this study may have provided the ideal setting in which to examine planning differences across contexts at this age. It is possible that, given a less demanding task (or this task at an older age), children would have performed similarly in both conditions. On the other hand, a more difficult task may have created floor effects in which children would not have performed at above chance levels in either condition. The variability in planning in this study was likely due to an interplay between task difficulty and developmental period. Whether and how individual versus joint planning differs in different developmental periods and at different levels of task complexity should be explored further.

The joint action condition in this study was minimally "joint" in that it involved a turn-taking task in which the social partner always performed correctly. Turn-taking reduced timing and coordination demands common in other joint action tasks. Further, if children learned that the joint partner always acted correctly, he or she could have simply ignored the partner and continued to play without taking him or her into account. The fact that children did perform differently in individual and joint conditions suggests that they likely viewed these conditions differently (but see possibility of timing differences above). Future research
should consider the differential influences of more or less involved interactions with the social partner.

Findings from the current study suggest differences in three-year-old children's planning, but not problem-solving, when they play alone or jointly play with a partner. The mechanisms underlying this difference should be addressed in future research. Given that children of this age have the ability to view a partner as an intentional agent, predict another's actions, and plan their own actions, it seems that the integration of these skills is still undergoing development. How this differs when playing with parents, who may scaffold their actions, or with peers, who are less predictable in their actions, is an interesting avenue of future work. A better understanding of how planning within joint actions develops is important in order to further explore educational consequences, underlying neural mechanisms, and individuals who show a prolonged or atypical developmental pattern.

## Acknowledgments

We would like to extend our gratitude to Angela Khadar, Dirkje van der Aa, Marlene Meyer, Denise Janssen, Celeste Peereboom, Barbarien Bruens, and other members of the Nijmegen Baby Research Center for their help with data collection and thought-provoking discussions. We are especially grateful to the parents and children who participated in this study. This work was funded by a Brain-and-Cognition-Excellence-Grant from NWO (433-09-253).

## References

Baron, R. S. (1986). Distraction-conflict theory: Progress and problems. Advances in Experimental Social Psychology, 19, 1-39. doi: 10.1016/S0065-2601(08)60211-7
Cannon, E. N., Woodward, A. L., Gredebäck, G., von Hofsten, C. \& Turek, C. (2012), Action production influences 12-month-old infants' attention to others' actions. Developmental Science, 15: 35-42. doi: 10.1111/j.1467-7687.2011.01095.x

Carlson, S. M., Moses, L. J., \& Claxton, L. J. (2004). Individual differences in executive functioning and theory of mind: An investigation of inhibitory control and planning ability. Journal of Experimental Child Psychology, 87, 299-319. doi: 10.1016/j.jecp.2004.01.002
De la Ossa, J. L. \& Gauvain, M. (2001). Joint attention by mothers and children while using plans. International Journal of Behavioral Development, 25, 176-183. doi: 10.1080/01650250042000168

Gallese, V., \& Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. Trends in Cognitive Sciences, 2, 493-501. doi: 10.1016/S1364-6613(98)01262-5
Gauvain, M. (1992). Social influences on the development of planning in advance and during action. International Journal of Behavioral Development, 15, 377-398. doi: 10.1177/016502549201500306

Gauvain, M., \& Rogoff, B. (1989). Collaborative problem solving and children's planning skills. Developmental psychology, 25, 139-151. doi: 10.1037//00121649.25.1.139

Gerson, S. A., \& Woodward, A. L. (in press). Learning from their own actions: The unique effect of producing actions on infants' action understanding. Child Development.
Kourtis, D., Sebanz, N., \& Knoblich, G. (2010). Favouritism in the motor system: social interaction modulates action simulation. Biology Letters, 6, 758-761. doi: 10.1098/rsbl.2010.0478
McCormack, T., \& Atance, C. M. (2011). Planning in young children: A review and synthesis. Developmental Review, 31, 1-31. doi: 10.1016/j.dr.2011.02.002
Meyer, M., Hunnius, S., Elk, M. van, Ede, F.L. van, \& Bekkering, H. (2011). Joint action modulates motor system involvement during action observation in 3-yearolds. Experimental Brain Research, 211, 581-592. doi: 10.1007/s00221-011-2658-3

Meyer, M., van der Wel, R. P R. D., \& Hunnius, S. (2013). Higher-order planning for individual and joint object manipulations. Experimental Brain Research, 225, 579588.

Sebanz, N. \& Knoblich, G. (2009). Prediction in joint action: What, when, and where. Topics in Cognitive Science, 1: 353-367. doi: 10.1111/j.17568765.2009.01024.x

Sebanz, N., Knoblich, G., \& Prinz, W. (2003). Representing others' actions: just like one's own?. Cognition, 88, B11B21. doi: 10.1016/S0010-0277(03)00043-X
Rizzolatti, G., \& Craighero, L. (2004). The mirror-neuron system. Annual Review of Neuroscience, 27, 169-192. doi: 10.1146/annurev.neuro.27.070203.144230
Uithol, S., van Rooij, I., Bekkering, H., \& Haselager, P. (2011). Understanding motor resonance. Social Neuroscience, 6, 388-397. doi: 10.1080/17470919. 2011.559129
van Elk, M., Van Schie, H. T., Hunnius, S., Vesper, C., \& Bekkering, H. (2008). You'll never crawl alone: neurophysiological evidence for experience-dependent motor resonance in infancy. Neuroimage, 43, 808-814. doi: 10.1016/j.neuroimage.2008.07.057
Welsh, M. C., Pennington, B. F., \& Groisser, D. B. (1991). A normative-developmental study of executive function: A window on prefrontal function in children. Developmental Neuropsychology, 7, 131-149. doi: 10.1080/8756564910954048

Wilson, M., \& Knoblich, G. (2005). The case for motor involvement in perceiving conspecifics. Psychological Bulletin, 131, 460-473. doi: 10.1037/00332909.131.3.460

# Back on track: Backtracking in counterfactual reasoning 

Tobias Gerstenberg (t.gerstenberg@ucl.ac.uk), Christos Bechlivanidis (c.bechlivanidis@ucl.ac.uk), David A. Lagnado (d.lagnado@ucl.ac.uk)<br>Cognitive, Perceptual and Brain Sciences, University College London, London WC1H 0AP


#### Abstract

Would Dan have died if Bob hadn't shot? In this paper, we show that people's answer depends on whether or not they are asked about what would have caused Bob not to shoot. Something needs to change in order to turn an actual world into a counterfactual world. Previous findings of how people reason about counterfactuals have been mixed: sometimes people appear to backtrack and reevaluate the causes of a counterfactual state (e.g. Rips, 2010). At other times, people appear to treat counterfactuals like interventions that leave the past unchanged (Sloman \& Lagnado, 2005). We experimentally manipulated the order in which participants were asked to consider the consequences of a counterfactual state. The results show that participants are more likely to backtrack when explicitly asked to consider a counterfactual's causes. However, when directly asked about the effects of a counterfactual state, most people don't backtrack.


Keywords: counterfactuals; causality; inference; backtracking.

## Introduction

Counterfactual thoughts play an important part in our everyday lives (see, e.g. Roese, 1997): if we had missed the submission deadline, you wouldn't be reading this paper. If we hadn't embarked on scientific careers, we would have become famous musicians. How do we evaluate the truth of such counterfactual statements? As life does not come with a rewind button, we can never know for sure. Hannes Kürmann, the protagonist in Max Frisch's play Biography: A Game, gets the unique chance to go back in time and play the game of life for a second time. However, despite full awareness of how his unhappy life will unfold and the firm belief that things could have turned out differently, Kürmann cannot bring himself to undo his past (and consequently, his present and future).

Max Frisch's play paints a rather fatalistic picture and suggests that counterfactual thoughts about how our life could have turned out differently are likely to be false. If everything happened as it actually did up until the point of the considered counterfactual, it has to turn out false. At some point, the counterfactual world has to diverge from the actual world in order to ensure the truth of the if-part (or antecedent) of a particular counterfactual statement. At least a change of mind would have been required to transform a scientist's life into that of a rock star.

Often there are a number of ways to realize the truth of a counterfactual's antecedent and the way in which we do so can sometimes have quite dramatic consequences. Consider the following situation: Anne is the commander of a firing squad and blows a whistle to signal to Bob and Chuck that it's time to shoot poor Dan (see Figure 1, cf. Pearl, 2000). Both Bob and Chuck shoot and Dan dies. Let us assume that

(a) What actually happened


Figure 1: If Bob had not shot, would Dan have survived?
the relevant causal relationships are deterministic: whenever Anne gives the signal, Bob and Chuck shoot and they never miss. Furthermore, each of Bob's and Chuck's shots are individually sufficient to bring about Dan's death. What do you think: would Dan have survived if Bob had not shot?

In this paper, we investigate how people evaluate counterfactual statements about simple devices that are structurally equivalent to the scenario just described. We first review theoretical frameworks that yield competing predictions about whether certain counterfactuals are true and then summarize previous empirical work on how people reason counterfactually. In a series of experiments, we test whether or not people spontaneously backtrack by manipulating the order in which participants are asked different counterfactual questions. We find that participants are more likely to backtrack when asked to explicitly consider the cause of the counterfactual's antecedent and suggest that the effect of question order can be explained in terms of a local processing strategy.

## Theories of counterfactual conditionals

Let us illustrate the differences between theories of counterfactuals via the example of the counterfactual conditional "If Bob had not shot then Dan would have survived".

According to Lewis's (1979) account, the counterfactual conditional is true if the counterfactual world in which Bob had not shot $(B=0)$ and Dan would not have died $(D=0)$ is more similar to the actual world than any counterfactual world in which Bob had not shot $(B=0)$ but Dan would have died anyhow ( $D=1$ ). To generate the relevant counterfactual world, we are supposed to imagine a small miracle that transforms $B$ from its original state to the considered counterfactual state and then let the counterfactual world unfold
according to the laws of nature. There are several problems with Lewis's account whereby most of which relate to the underspecified notion of similarity between different worlds (cf. Hiddleston, 2005). While Lewis aims to provide a non-causal account of counterfactuals and reduce causality to counterfactuals, others have argued that this puts the cart before the horse (Hiddleston, 2005; Pearl, 2000).

More recently, theories have been developed that take the notion of causality as primary and evaluate the truth of counterfactuals via reference to explicit causal assumptions that can be represented in causal Bayesian networks (CBN, Hiddleston, 2005; Pearl, 2000). In the spirit of Lewis's (1979) account, these theories evaluate the truth of counterfactuals by referring to similar worlds. However, they differ in how they conceptualize the causal similarity between different possible worlds.

According to Pearl's (2000) pruning theory ${ }^{1}$, the evaluation of a counterfactual involves three steps. First, we update the values of the variables in the causal network based on our observations in the actual world (i.e. $A=1, B=1, C=1$ and $D=1$ ). Second, we change the value of the antecedentvariable (i.e. B) by means of an intervention. Such an intervention results in a mutilated causal network in which all incoming links to the intervened-on variable are removed (see Figure 1b). Third, we evaluate the consequent-variable (i.e. $D)$ based on the variables' values in the mutilated network. Since the intervention in $B$ disconnects all influences of upstream variables, $A$ 's value in the mutilated network remains unchanged. Because Chuck shoots whenever Anne gives the signal (i.e. $C=A$ ) and Dan dies if either Bob or Chuck shot (i.e. $D=\max (B, C)$ ) the counterfactual is false. Dan would have died even if Bob had not shot.

Pearl's (2000) account of dealing with counterfactuals is similar to Lewis's (1979) in that the considered counterfactual world is identical to the actual world up until the point of the antecedent-variable. The antecedent-variable's counterfactual value is realized via an intervention that locally violates the causal relationships of the structure. The resulting counterfactual world is similar to the actual world in that the values of all variables that precede the antecedent-variable (or are causally independent from it) remain unchanged. However, it is dissimilar in that some of the causal relationships that were true about the actual world are not respected in the counterfactual world.

The opposite is true for Hiddleston's (2005) minimalnetwork theory. In this theory, the truth of a counterfactual conditional is evaluated by considering whether it holds in all worlds that are minimally different from the actual world but consistent with its causal laws. Given that the relationships between the actors in our scenario were described as deterministic, there are only two possible worlds that are causally consistent. The actual world (in which the values of all variables are 1) and a counterfactual world in which Anne did not give the signal, neither Bob nor Chuck shot and Dan survived

[^276](i.e. all values are 0). Hence, according to minimal-network theory, the considered counterfactual is true. If Bob had not shot, Dan would have survived (see Figure 1c).

The relevant counterfactual world is dissimilar from the actual world in that all events are different from how they actually were (including events that were temporally prior to the considered counterfactual). However, it is similar in that none of the actual causal relationships have been tampered with.

Note that evaluating the truth of counterfactuals according to minimal-network theory requires us to not only consider the consequences of the antecedent-variable. Bringing about the counterfactual state of the antecedent-variable in a way that is consistent with the causal laws requires us to backtrack and change the values of the antecedent-variable's causes as well. More generally, whereas pruning theory yields that backtracking counterfactuals (e.g. If Bob had not shot then Anne would not have given the signal) are always false, minimal-network theory holds that they can be true (at least in deterministic contexts).

## Psychological studies of counterfactual reasoning

The results of previous studies on how people reason about counterfactuals have been mixed. Sloman and Lagnado (2005) found that people's counterfactual judgments are closely in line with the predictions of pruning theory. In one of their experiments, participants received descriptions of a causal structure identical to the one in the above scenario. In the abstract version of the task, they were informed that $A$ causes $B$ and $C$, and that $B$ and $C$, in turn, each cause $D$. Knowing that $D$ definitely occurred, participants answered the following two counterfactual questions: (a) If $B$ had not occurred, would $D$ still have occurred? (b) If $B$ had not occurred, would $A$ have occurred?

Pruning theory predicts that participants should answer 'yes' to both questions whereas minimal-network theory predicts negative responses. $80 \%$ of the participants answered 'yes' to (a) and $79 \%$ to (b). Responses were similar for scenarios in which the variables and causal relationship were described more concretely (a: $78 \%$, b: $81 \%$; averaged).

However, there has also been empirical support for minimal-network theory (Dehghani, Iliev, \& Kaufmann, 2012; Rips, 2010; Rips \& Edwards, in press). Rips (2010) and Dehghani et al. (2012) focused on backtracking counterfactuals and found that participants' judgments were sensitive to information about the base rates of the antecedentvariable's causes, the way in which these interact (disjunctive vs. conjunctive) and whether the causal links are deterministic or probabilistic. Since pruning theory rules out all backtracking counterfactuals, it cannot account for any of these effects. Recently, Lucas and Kemp (2012) have extended pruning theory to handle backtracking counterfactuals by allowing that variables which are not affected by the counterfactual intervention may take non-actual values.

One might argue that what answers a theory gives to backtracking counterfactuals is not of utmost importance for psychological theorizing. In everyday life, we are normally in-
terested in the effects rather than the causes of counterfactuals. However, as the firing-squad scenario has shown, in some causal structures, whether or not a theory allows for backtracking also affects the truth of non-backtracking counterfactuals. Dan would have survived if Bob had not shot only if we backtrack and change Anne's action.

Rips and Edwards (in press) investigated participants' counterfactual reasoning using abstract devices that were structurally identical to the firing-squad scenario. They varied whether the causal links were described as deterministic or probabilistic (e.g. A's operating always/usually causes $B$ to operate) and whether $B$ and $C$ brought about $D$ in a disjunctive $(D=\max (B, C))$ or conjunctive manner $(D=\min (B, C))$. Furthermore, they manipulated the framing of the counterfactual question between participants. Participants were either asked to consider that a certain component had failed or not operated (e.g. If $B$ had not operated [failed] would $A / C / D$ have operated?). Generally, participants tended to show less backtracking in the failed condition which suggests a local failure in the device than in the not operated condition. Furthermore, there was less backtracking for probabilistic compared to deterministic devices.

We will focus on structures with deterministic causal links for which the predictions between pruning theory and minimal-network theory dissociate strongest. Remember that for the deterministic disjunctive device, Sloman and Lagnado (2005) found that most participants answered positively to the question of whether $A$ (or $D$ ) would have occurred if $B$ had not occurred, Rips and Edwards (in press) found that in their not operated condition, almost all participants answered negatively. In the following, we will explore whether the way in which people mentally process counterfactual questions might account for these divergent findings.

Note that pruning theory and minimal-network theory make different predictions about what states of the system people need to consider when asked whether $D$ would have operated if $B$ had not operated. According to minimalnetwork theory, we first have to backtrack and infer that if $B$ had not operated then $A$ would not have operated. From this it follows that $C$ and $D$ would not have operated. Pruning theory, in contrast, predicts that we can evaluate the truth of the counterfactual without considering the state of $A$ (see Figure 2). Since the counterfactual intervention on $B$ does not affect the state of $C, D$ is predicted to operate even if $B$ had


Figure 2: Hypothesized counterfactual reasoning process in the $D-C-A$ condition.
not operated (because of $C$ ).
This reasoning suggests that the order in which participants are asked to answer different counterfactual questions might influence how likely they show backtracking. In Sloman and Lagnado's (2005) experiment, participants were always asked about $D$ first and then about $A$. In Rips and Edwards's (in press) experiment, participants were asked about $A, C$, and $D$ and free to answer the questions in any order (cf. Figure 3a). Participants indicated their processing order on the response sheet and, generally, answered the questions from left to right (i.e. from $A$ to $B / C$ to $D$ ). In our experiments, we use a computerized task which allows us to manipulate the question order. Based on the discrepancy between Sloman and Lagnado's and Rips and Edwards's findings, we hypothesized that when asked to consider $A$ before $D$, participants will be more likely to show backtracking than when asked about $D$ before $A$. Note that neither pruning theory nor minimal-network theory predict any effects of question order.

## Experiment 1: Replication

We first attempted to replicate Rips and Edwards's (in press) findings in the not operated condition using a computerized interface. Participants ( $N=40$, recruited via Amazon Mechanical Turk) saw eight different devices in randomized order and were asked to answer whether each of the other three components would have operated if $A, B$ or $D$ had not operated (i.e. 8 devices $\times 3$ antecedent components $\times 3$ consequent components $=72$ questions). ${ }^{2}$ The devices differed in whether the causal links were described as deterministic or probabilistic and whether $B$ and $C$ combined disjunctively or conjunctively. The probabilistic devices differed in whether (i) all links were probabilistic, (ii) only the links from $A$ to $B$ and $C$ or (iii) from $B$ and $C$ to $D$ (see Rips \& Edwards, in press, for more details).

The order in which participants were asked about the different antecedent components was counterbalanced ( $A-B-D$ vs. $D-B-A)$. For each antecedent component, participants were free to answer the counterfactual questions for the different consequent components in any order (see Figure 3a). For example, if $B$ was the antecedent component (i.e. if $B$ had not operated) a participant could answer about $A$ (e.g. $A$ would not have operated), $C$ and $D$ in any order. For each counterfactual, participants' response options were to say that the component would have operated, would not have operated or might have operated.

## Results and Discussion

We followed Rips and Edwards's (in press) procedure and coded participants' responses as -1 (does not operate), 0 (might operate) and 1 (operates) in order to run standard statistical analyses. Figure 4 shows a selection of the results. Overall, we closely replicated Rips and Edward's findings with a correlation of $r=.92(R M S E=0.23)$ between the

[^277]

Figure 3: Screenshots of the interface in Experiments 1 and 2 ( $D-C-A$ order condition).
averaged responses to the 72 questions in both experiments. Participants again tended to answer the counterfactual questions from left to right. For example, the average order in which participants indicated to have answered the counterfactual questions when $B$ was the antecedent component was 1.23 for $A, 2.10$ for $C$ and 2.67 for $D$ (the corresponding values in Rips and Edwards's experiment were $A: 1.44, B: 2.21$ and $D: 2.23$ ).

Whereas both pruning theory and minimal-network theory predict the same pattern of responses when $A$ is the antecedent component, their predictions differ when the antecedent components are $B$ or $D$. Minimal-network predicts that the answers to all counterfactual questions are negative. Pruning theory, in contrast, predicts that when $D$ is the antecedent, the answers to all consequent components should be positive. When $B$ is the antecedent component, pruning theory predicts that the answers to both $A$ and $C$ should be positive. For the $D$, the answer is predicted to be negative for the conjunctive and positive for the disjunctive device.

In line with Rips and Edwards's findings and as predicted by minimal-network theory, a majority of participants answered the counterfactual questions negatively. For example, when asked whether $A, C$ and $D$ would have operated if $B$


Figure 4: (i) Mean judgments separated for the disjunctive and conjunctive deterministic device. The labels on the $x$-axis correspond to the consequent-components. Most frequently endorsed structures for the (ii) disjunctive and (iii) conjunctive devices. Note: R\&E = Rips and Edward's (in press) data.
had not operated for the conjunctive device, 24 participants showed backtracking whereas only 8 participants responded in line with pruning theory (see Figure 4b).

## Experiment 2: Order Manipulation

Having replicated Rips and Edwards's finding, Experiment 2 tests the hypothesis that the order in which participants are asked to answer different counterfactual questions influences the degree to which they backtrack. With $B$ as the counterfactual antecedent, we predicted that participants will show more backtracking when asked about $A$ before $D$ and less backtracking when asked about $D$ before $A$.

Between participants ( $N=320$, recruited via Amazon Mechanical Turk), we manipulated the question order ( $A-C-D$ vs. $D-C-A$ ), whether the device was disjunctive or conjunctive as well as whether, in actuality, all or none of the components were operating ( 40 participants per condition). When all components were operating participants were asked to consider the counterfactual that $B$ had not operated (see Figure 5a and b). When none of the components were operating, participants considered that $B$ had operated (see Figure 5c and d).

Our processing hypothesis predicts an interaction between the question order, the type of device and its actual state. The question order is predicted to influence participants' judgments about the counterfactual state of $D$ for (a) the disjunctive device in which everything is actually operating and (d) the conjunctive device in which nothing is operating (see Figure 5 a and d). In these cases, whether $D$ would have been different from actuality depends on whether or not participants backtrack. Accordingly, we predicted that participants in the $D-C-A$ condition are more likely than participants in the $A-$ $C-D$ condition to say that $D$ would have operated for (a) and less likely to say that $D$ would have operated for (d). In contrast, we do not predict an effect of question order for devices (b) and (c). The counterfactual state of $B$ is by itself sufficient to bring about a change in $D$ without the need to consider the states of the other components.

Figure 3 b shows a screenshot of the $D-C-A$ condition. Participants first only saw the text box for $D$. Having answered that question, the response was locked and the next text box


Figure 5: Mean judgments (ii) and frequency of endorsed networks (iii) for different causal devices (i). Note: For each device, the leftmost networks in (iii) are predicted by minimal-network theory and the rightmost networks by pruning theory. The networks in the middle are the most frequently endorsed networks predicted by neither of the two theories.
appeared. All participants just provided answers for a single device.

## Results and Discussion

For ease of interpretation, we analyze the results for devices in which everything is operating initially (Figures 5 a and b) and in which nothing is operating (Figures 5 c and d) separately and focus on participants' answers to component $D$.

For the operating devices $(\mathrm{a}, \mathrm{b})$, there was a significant main effect of structure, $F(1,156)=44.13, p<.001, \eta_{p}{ }^{2}=$ .459 and no main effect of question order ( $p=.097$ ). Participants were more likely to think that $D$ would have operated for the disjunctive ( $M=0.25, S D=0.88$ ) compared to the conjunctive device ( $M=-0.9, S D=0.41$ ).

More interestingly, there was a significant interaction between structure and question order $F(1,156)=9.59, p=$ $.002, \eta_{p}^{2}=.058$. For the disjunctive device, participants in the $D-C-A$ condition were more likely to say that $D$ would have operated ( $M=0.5, S D=0.82$ ) than participants in the $A-C-D$ condition $(M=0, S D=0.88), t(78)=-2.64, p=$ $.01, d=-0.6$. For the conjunctive device, there was no significant difference as a function of question order $(p=.101)$.

The results for the non-operating devices (c, d), closely mirrored the results of the operating devices. Again, there was a significant effect of structure, $F(1,156)=39, p<$ $.001, \eta_{p}{ }^{2}=.302$ and no main effect of question order ( $p=$ .079). Participants were more likely to think that $D$ would have operated for the disjunctive ( $M=0.70, S D=0.68$ ) compared to the conjunctive devices ( $M=-0.29, S D=0.87$ ).

The interaction between structure and question order was significant $F(1,156)=5.26, p=.003, \eta_{p}{ }^{2}=.055$. While there was no significant difference of question order for the
disjunctive device ( $p=.329$ ), in the case of the conjunctive device, participants in the $D-C-A$ condition were less likely to say that $D$ would have operated ( $M=-0.58, S D=0.78$ ) than participants in the $A-C-D$ condition ( $M=0, S D=0.88$ ), $t(78)=3.1, p=.003, d=0.70$.

These results demonstrate that the order in which participants were asked about the different components affected whether they believed that $D$ would have operated. In the $A-C-D$ condition, 36 participants (out of 160 ) answered as predicted by minimal-network theory and 42 as predicted by pruning theory (see Figure 5iii). These numbers shifted towards much less backtracking in the $D-C-A$ condition: only 15 participants answered consistently with minimal-network theory, whereas 68 answered in line with pruning theory.

The results also revealed another interesting pattern: the absolute value of participants' averaged answers about component $A(M=0.15)$ were generally less certain (i.e. closer to $0)$ than their answers about $C(M=0.33)$ and $D(M=0.53)$. The shift towards averaged 0 responses from component $C$ to $A$ in Figures 5a and d for the $D-C-A$ condition is neither predicted by pruning theory nor minimal-network theory. We consider this to be evidence that people process counterfactual questions in a more local fashion rather than simultaneously considering the states of all variables in the system.

For example, when asked whether $D$ would have operated if $B$ had not operated (cf. Figure 5a) most participants in the disjunctive $D-C-A$ condition answer 'yes' to $D$. Having answered positively to $D$ commits participants to saying that $C$ would have operated as well (cf. Figure 2). Otherwise, there is no explanation for why $D$ operates. However, when considering $A$, participants have reached a state of causal inconsistency. Having answered 'yes' to $C$ but knowing that $B$ did not


Figure 6: Hypothesized local counterfactual reasoning process in the $A-C-D$ condition.
operate, they can either resolve this inconsistency by answering 'yes' to $A$ and assuming a fault in $B$. Alternatively, they can answer 'no' to $A$ and assume that $C$ must have operated spontaneously. The same rationale also explains the pattern of results for device (d). For devices (b) and (c), participants’ response to $D$ does not commit them to a particular response for component $C$ - the counterfactual state of $B$ already accounts for the change in $D$.

Participants in the $A-C-D$ condition have to resolve the potential causal inconsistency right at the start (see Figure 6). As the results show, participants are split in how they do so: some backtrack and respond in line with minimal-network theory. Others don't and respond as predicted by pruning theory.

## General Discussion

The capability to think about possible states of the world and reason through what would or could have happened is one of the hallmarks of human cognition. Counterfactual thoughts are of central importance to attributions of responsibility (Lagnado, Gerstenberg, \& Zultan, accepted) and causality (Gerstenberg, Goodman, Lagnado, \& Tenenbaum, 2012). In this paper, the aim was to gain insight into people's counterfactual processing. Based on mixed findings in previous research (Dehghani et al., 2012; Meder, Hagmayer, \& Waldmann, 2009; Rips, 2010; Sloman \& Lagnado, 2005), we investigated whether the order in which participants are asked to reason about the consequences of certain counterfactual states could shed light on these inconsistencies. We first replicated Rips and Edwards's (in press) experiment and then manipulated the order of counterfactual questions in an identical experimental setup.

As hypothesized, participants' answers were more in line with the predictions of minimal-network theory (Hiddleston, 2005) when asked to consider a possible cause of the counterfactual state first. In contrast, when participants were asked to consider the effect of a counterfactual state first, participants showed less backtracking and followed the predictions of pruning theory (Pearl, 2000) more closely. However, the overall pattern of results was not predicted by either theory. We discussed that a more local processing strategy is consis-
tent with this data (cf. Fernbach \& Sloman, 2009, for a similar idea in causal learning). Accordingly, when asked to consider a certain counterfactual, people do not spontaneously think through the implications that this counterfactual state has for the whole system. Rather, participants' responses are indicative of a more local processing strategy that considers only parts of the system. The order in which participants are probed about the counterfactual world hence has a significant effect on what changes they make in order to account for the stipulated counterfactual state. Applied to our initial example, whether Dan is believed to have survived if Bob had not shot depends on whether we are asked to consider Anne first.

While the results of Experiment 1 have shown that participants' responses were closely in line with minimal-network theory, the results in Experiment 2 were more mixed. In future research, we aim to (i) generalize these findings using less abstract stimuli and (ii) investigate more closely what differences between the reported experiments account for participants' tendency to backtrack or not. We speculate that both the explicit contrast between deterministic and probabilistic systems as well as the fact that participants have to think through a great number of different devices, encourages them to endorse a more holistic strategy that favors responses that are causally consistent. However, when not asked explicitly to consider the causes of a counterfactual state, most participants stay on track and don't backtrack.

## Acknowledgments

We thank Lance Rips for providing the data and Brian Edwards for insightful comments. This work was supported by a doctoral grant from the AXA research fund (TG) and an ESRC grant RES-062-33-0004 (DL).

## References

Dehghani, M., Iliev, R., \& Kaufmann, S. (2012). Causal explanation and fact mutability in counterfactual reasoning. Mind \& Language, 27(1), 55-85.
Fernbach, P. M., \& Sloman, S. A. (2009). Causal learning with local computations. Journal of Experimental Psychology: Learning, Memory, and Cognition, 35(3), 678-693.
Gerstenberg, T., Goodman, N. D., Lagnado, D. A., \& Tenenbaum, J. B. (2012). Noisy Newtons: Unifying process and dependency accounts of causal attribution. In Proceedings of the 34th Annual Conference of the Cognitive Science Society.
Hiddleston, E. (2005). A causal theory of counterfactuals. Noûs, 39(4), 632-657.
Lagnado, D. A., Gerstenberg, T., \& Zultan, R. (accepted). Causal responsibility and counterfactuals. Cognitive Science.
Lewis, D. (1979). Counterfactual dependence and time's arrow. Noûs, 13(4), 455-476.
Lucas, C. G., \& Kemp, C. (2012). A unified theory of counterfactuals. In Proceedings of the 34th Annual Conference of the Cognitive Science Society.
Meder, B., Hagmayer, Y., \& Waldmann, M. R. (2009). The role of learning data in causal reasoning about observations and interventions. Memory \& Cognition, 37(3), 249-264.
Pearl, J. (2000). Causality: Models, reasoning and inference. Cambridge University Press.
Rips, L. J. (2010). Two causal theories of counterfactual conditionals. Cognitive Science, 34(2), 175-221.
Rips, L. J., \& Edwards, B. (in press). Inference and explanation in counterfactual reasoning. Cognitive Science.
Roese, N. J. (1997). Counterfactual thinking. Psychological Bulletin, 121(1), 133-148.
Sloman, S. A., \& Lagnado, D. A. (2005). Do we 'do'? Cognitive Science, 29(1), 5-39.

# Musical training and auditory global-local precedence 

Anne T. Gilman (anne.gilman@gmail.com)<br>Juniata College Psychology Department, Good Hall, 1700 Moore Street<br>Huntingdon, PA 16652 USA<br>Elizabeth Ebbets (lizaebbets@gmail.com)<br>Juniata College Psychology Department, Good Hall, 1700 Moore Street<br>Huntingdon, PA 16652 USA


#### Abstract

People demonstrate a consistent tendency to favor holistic or global processing over processing of local details in many perceptual domains; this tendency is called global-local precedence. Formal musical training is associated with qualitative changes in auditory processing, but the number of years required remains unclear, particularly for any perceptual differences between untrained and minimally trained participants. In this study, participants with zero to over ten years' music training identified the direction of pitch changes in three-sound sequences. Only participants with three or more years' training demonstrated a significant global-local precedence. Individuals without musical training consistently identified the local direction of pitch-change sequence elements better than global pitch changes across each sequence. Although musical training was associated both with greater task accuracy and with global-local precedence, improved accuracy did not explain the musically trained participants' preferential processing of global auditory characteristics.


Keywords: Global-local precedence; musical training; temporal processing; auditory perception.

## Perceptual learning and cognition

People's experiences can alter their perceptual processing, even as adults. Perceptual learning thus takes part in core cognitive mechanisms underlying storage of past experiences and their subsequent application to similar but non-identical situations later on. Distinctions between global (or holistic) processing and local (or detail) processing have only recently been examined while taking perceptual learning into account. This connection offers particular interest in the auditory domain, given long-established differences in how trained musicians and musically naïve listeners perceive sounds (Bever \& Chiarello, 1974).

## Global-local precedence

People who become so engrossed in the details of a complex scene or situation that they fail to notice compelling overall patterns are said to be "missing the forest for the trees". This colloquial expression has a parallel in human perception: decades of vision research has documented a bias or precedence towards overall or global characteristics of an image rather than its local details (Navon, 1977). For example, given an image made up of a dozen K's arranged in the shape of an H , viewers are more likely to describe what they see as an H than as a group of K's. When asked to identify either the big letter or the little component letter in a series of many such composite stimuli, viewers are likely to make fewer errors overall when identifying the big rather than the little letters, thus global-local precedence can be identified by greater
overall accuracy for global trials. Viewers in such studies also show a pattern in their errors for incongruent stimuli such as an H made of K's-one could also create a congruent stimulus by arranging K's in the shape of a larger K. Viewers more often mistakenly offer the big letter as an answer on local (little-letter trials) than they err by naming the little letter on global trials, showing an uneven influence-greater for global patterns-of these different processing levels on perception. This uneven influence can also be described as interference by global processing in the local decision. A growing body of evidence supports the overall global-local processing distinction in auditory and other non-visual modalities (Justus \& List, 2005; List \& Justus, 2007, 2010; List, Justus, Robertson, \& Bentin, 2007; Ouimet, Foster, \& Hyde, 2012; Sanders \& Poeppel, 2007).

Cognitive scientists have found ongoing and robust preferential processing of global over local patterns (Love, Rouder, \& Wisniewski, 1999; Ripoll \& Marty, 2005) even while some boundary conditions have been delineated (Navon, 2003). This precedence is considered an outcome of normal maturation rather than one of learning. A large body of evidence, summarized by Poirel and colleagues (2011), suggests that young children shift from a local to a global visual focus around the age of 6 ; Poirel et al. found that lowered greymatter volume corresponded to this shift, again characterizing global precedence as the healthy adult norm. Visual and auditory findings of differences in holistic versus detailed processing bear on issues ranging from autism (Rondan \& Deruelle, 2007) to the acquisition of reading skills (Foxton et al., 2003). For instance, exceptions to global-local precedence for auditory stimuli correspond to autism diagnoses, even though the detail-specific advantages shown by autistic participants were not explained by deficits in global processing as previously thought (Mottron, Peretz, \& Ménard, 2000). In all of the above investigations, global-local precedence is taken as an indicator of healthy and fully-developed perception, and only very recently has this tendency been evaluated with respect to prior musical training in a parallel auditory study (Ouimet et al., 2012).

## Expertise: Musical training

Auditory perception varies qualitatively-not just in accuracy-depending on learning or expertise. Comparing perception of tones presented to the right or left ear has revealed that hemispheric dominance for this task switches
following formal music training (Kellar \& Bever, 1980; Bever \& Chiarello, 1974). Subsequent studies have shown that musical training changes listeners' perception of auditory patterns, both for musical stimuli themselves (Fujioka, Trainor, Ross, Kakigi, \& Pantev, 2004; Foxton, Brown, Chambers, \& Griffiths, 2004) and for the pitch changes that characterize fluent speech (Moreno et al., 2009). Musical training also contributes to changes in numerous other cognitive processes, and better understanding of the precise impact of such training on auditory perception may clarify ongoing questions about music and its contribution to formal measures of intelligence and academic achievement (Schellenberg \& Peretz, 2007).

The one investigation (other than the present study) of the role of prior musical training in auditory global-local precedence found this precedence more strongly demonstrated among less-trained listeners (Ouimet et al., 2012). Accuracy results were contrasted between expert musicians-with at least seven years' formal training-and novices, where the lower-expertise group could have one or two years of prior training, or none at all. This division is consistent with other key precedents in music perception (Warrier \& Zatorre, 2002). However, findings in visual perception, where a mere ten hours' practice with a video game significantly and enduringly changed participants’ skills (Green \& Bavelier, 2003), suggest that the perceptual performance of participants with minimal amounts of musical training should be contrasted with the performance of those with no training at all.

## Evaluating global-local precedence and musical training

The present study reevaluates global auditory bias for possible influences of perceptual expertise, employing the same three-part sweep sequences tested in Sanders and Poeppel's (2007) ERP study to contrast performance of participants with varying levels of musical expertise, including no prior training whatsoever.

## Method

## Participants

Over multiple semesters, 104 undergraduates with normal or corrected-to-normal hearing participated in the experiment, receiving either course participation credit or snack food as compensation.

## Materials

The auditory stimuli selected for this study were created for an ERP study of global versus local auditory processing (Sanders \& Poeppel, 2007). Each stimulus consists of three frequency-modulated octave sweeps or chirps that can go up or down in frequency at the global or local level, as follows. Each chirp lasts only 40 ms , with 190 ms of silence separating the chirps; the three-chirp sequence takes 500 ms total. Each chirp travels smoothly up or down one octave-this direction constitutes the local direction, and all three chirps


Figure 1: Stimulus frequency distributions over time.
of a given sequence go in the same direction. The progression of each sequence is monotonic, with the central pitch of each sweep being higher than the previous one in global upward sequences and lower for downward sequences. The difference in central frequency between the first and the last octave sweeps is .2 octaves. In musical terms, if the central frequency of one chirp were a C note, the central frequency of the next chirp would be a bit over a half step higher or lower that that of the first chirp. See Figure 1 for sample spectrograms, obtained using Praat (Boersma, 2001).

Congruent stimuli (e.g. Figure 1(a)) have changes in frequency which go in the same direction (up or down) at the local and global levels. Incongruent stimuli (e.g. Figure 1(b)) require the participants to accurately distinguish between levels to answer correctly. The stimulus set includes eleven sequences per condition, namely local-up/global-down, local-up/global-up, local-down/global-down, and local-down/global-up, for a total of 44 stimuli. Frequencies used in these stimuli range between a minimum of 0 Hz and a maximum of 3500 Hz . The intensity of each stimulus file measures between 88 and 92 dB , but participants could control playback volume.

## Procedure

Participants were asked to discriminate between upward and downward sweeps (local) and sweep sequences (global) after training involving both passive familiarization and active practice, replicating Sanders \& Poeppel's (2007) behavioral protocol. During familiarization, participants first read a description of the type of sounds they were about to hear, e.g. sounds that go down as a whole, and then heard all of those stimuli (global-down ones in this example) in randomized order. During practice trials, the participants heard a sound stimulus while either "PART" (cueing a local judgement) or "WHOLE" (for a global judgement) was displayed at the center of the screen. They indicated by keypress whether the current stimulus was going up or down, and then they received on-screen feedback showing whether or not their answer was correct. After two rounds of practice, participants then enagaged in four blocks of testing trials, which were identical to the practice trials except that no feedback was provided.

After completing the sound discrimination task, participants indicated whether they had ever studied a musical instrument, including voice. They then had the opportunity to clarify if they had studied music for more than one, more than two, or more than five years.

## Apparatus

Sounds and instructions were presented on two Macintosh iBook G4s equipped Sennheiser HD 202 headphones and running code written by the first author in PEBL (Mueller, 2006). Participants were free to adjust the computer volume to a comfortable level; they pressed keyboard keys ("U" and "D" for "up" and "down") to record their answers.

## Results

An overall global-local precedence in auditory discrimination was demonstrated by our participants, who showed $61.8 \%$ accuracy on global trials versus $51.5 \%$ accuracy on local trials. A $2 \times 2 \times 2$ analysis of variance (ANOVA) comparing the impact of task focus (global or local), stimulus type (congruent or incongruent), and musical training (any or none) showed this difference to be highly significant, $F(1,18296)=$ $55.8, M S E=4262, p<.001$. All analyses were performed using R (R Development Core Team, 2005).

As expected, participants were more successful judging the direction of pitch change for congruent stimuli, demonstrating $66.7 \%$ accuracy on congruent trials and $51.5 \%$ on incongruent trials, $F(1,18296)=451.8, M S E=4262, p<.001$.

To evaluate whether our participants demonstrated any global-local precedence, each person's percent accuracy for global and judgements was compared using the simple difference of their average performance in each condition, $M_{\text {global }}-M_{\text {local }}$. A positive score indicated global-local precedence, while a negative score indicated the opposite. Our participant group as a whole demonstrated an average global-local precedence of .05. This average differed significantly from zero, $t(103)=3.07, p=.0027$. The significant interaction between task focus and sound type, $F(1,18296)=$

| Music Training | Local Advantage | Global Advantage |
| :---: | :---: | :---: |
| None | 15 | 5 |
| Some | 31 | 53 |

Table 1: Number of participants demonstrating each sort of bias according to presence or absence of past musical training
82.7, $M S E=4262, p<.001$, further supported our finding of global-local precedence. A Pearson's correlation between this precedence measure calculated over all trials and the same measure calculated over only incongruent trials showed that these incongruent trials were indeed driving the precedence results, $r=.89, t(102)=20.1, p<.001$. This correlation confirms the presence of the second sign of global-local precedence, where errors on local trials show what can be described as interference from global processing: participants made more errors on local trials, and they did so to a greater extent on incongruent ( $M=0.46$ ) than congruent ( $M=0.67$ ) local trials.

The remaining analyses consistently indicated that musical training is associated with differences in global-local precedence. Those participants who reported any musical training at all showed significantly higher task accuracy (60.2\%) than those with none $(54.7 \%), F(1,18296)=45.6, M S E=$ $4262, p<.001$. Reaction times were lower among musicallytrained participants. These times were not recorded for half of the participants, due to experimenter error; the available times did not show any interactions with other variables. Greater musical training was associated with significantly different global-local precedence, however, $t(31)=-2.5, p=.017$, with an opposite precedence pattern for those with some musical training ( $M=8 \%$ ) as compared to those with none ( $M=-6 \%$ ). This difference clarifies the significant threeway interaction found in the main ANOVA between task focus, sound type, and presence or absence of prior musical training, $F(1,18296)=21.9, M S E=4262, p<.001$.

To evaluate this difference without making assumptions about the distributional characteristics of either the training or the precedence score, we compared this indicator of global versus local advantage to reported prior musical training using a $\chi^{2}$ test (see Table 1). Those with no prior musical background were far more likely to demonstrate better accuracy on local than on global judgements ( 15 out of 20), while participants with some musical training were more likely ( 53 out of 84) to demonstrate global-local precedence. This difference was significant, $\chi^{2}=8.022, d f=1, p=.0046$.

To explore this difference further, participants were grouped according to their reported level of musical expertise. Twenty participants reported having no musical training, 16 reported up to one year of training, 17 up to two years, 2 up to five years, and 49 more than five. For these analyses, responses from the two lone participants with between two and five years of training were combined with the most experienced musicians.

As expected, participants with the most musical training in
our sample showed the greatest overall accuracy in identifying the direction of pitch change in these three-part auditory sequences: those with more than one year of training showed $61 \%$ accuracy; those with up to one year, $57 \%$; and those with none, $55 \%$. All groups performed at above chance levels, $p<.001$.


Figure 2: Participant accuracy in identifying direction of pitch changes. The difference between global and local accuracy shows the opposite pattern among non-trained participants compared to the rest both for congruent and incongruent stimuli. Other than incongruent trials for those with 12 years' training and local incongruent trials for those with none all accuracies differ significantly from chance perfomance (dotted line).

## Discussion

A parallel investigation of global-local precedence in the auditory domain found the opposite results to those presented here, in that their less-trained participants exhibited significant global-local precedence and to a greater extent than their more expert participants (Ouimet et al., 2012). In our study, on the other hand, more expert participants demonstrated global-local precedence, but that precedence only reached statistical significance for those with more than 3 years of training. Those reporting no training at all performed better on local than on global trials, on average. Our study differs from Ouimet et al.'s in multiple ways, however, particularly in the temporal characteristics of the stimuli used and in the grouping of less-expert participants.

## Stimulus differences

Ouimet and colleagues used sequences of tones that correspond to notes on a musical scale (Ouimet et al., 2012, p. 2539), based on hierarchical stimuli developed by Justus and List (Justus \& List, 2005). Although these tone sequences incorporate a hierarchical structure similar to the sequences of
octave sweeps used in our study, there are several important differences.

First, the majority of their sound stimuli lasted much longer. Their three-part sequences of steady tones ranged in duration from 150 ms to 600 ms , so total stimulus durations ranged from 450 ms (comparable to our 500 ms , but with no silence) to 1800 ms . They found only a "negligible" (Ouimet et al., 2012, p. 2539) relation between stimulus duration and accuracy, however, adding support to List and Justus' arguments for the relational invariance of their tone-sequence tasks (List \& Justus, 2010, p.16).

Second, the tone and sweep elements differ sharply in their rate of change of frequency as well as their absolute duration, and the detection of direction of changes in pitch recruits neural resources not required for tone discrimination (Johnsrude, Penhune, \& Zatorre, 2000), possibly adding further difficulty for the sweep task. These stimulus differences, however striking, are unlikely to fully explain our differing results, however, since our participants also showed lower accuracy than that found in prior work with these same octave-sweep sequences (Sanders \& Poeppel, 2007), while still performing at above-chance levels. Some of our error level may arise from having collected data from pool participants at very different points in the semester (Grimm, Markman, \& Maddox, in press). A Pearson's correlation between week of the semester (ranging from 6 to $17, M=11.3$ ) and participant accuracy showed a modest negative correlation, $r=-.25, t(102)=-2.6, p=.0096$. Time of semester showed no interaction with precedence ( $p>.5$ ) or with musical training ( $p>.9$ ); repeating the initial analysis of variance including time of semester did not alter the results obtained.

A third important difference between the two stimulus sets is that in our study, the magnitude of each local change in pitch (one octave) is much greater than each stepped change in the global progression of pitch (one tenth of an octave). In Ouimet et al.'s study (2012), though, global pitch steps are three times the magnitude of local pitch intervals (147 cents, or about $1 / 8$ th of an octave). These stimulus differences complicate our interpretation of which accuracy changes are due to global versus local processing and which are due to discrimination difficulty.

## Participant differences

The other barrier to explaining the opposite precedence findings between the two studies for less-trained participants lies in the mismatch in the two studies' categorizations of musical training levels, a topic of ongoing debate. While Sanders and Poeppel did not measure musical expertise in their ERP study (2007), other assessments of global-local precedence in the auditory domain did record this participant characteristic. In the study presenting the tone-sequence design discussed above, one experiment had more musically trained than untrained participants, while the second had only musically trained participants (Justus \& List, 2005). In one of List and Justus' more recent studies, all participants reported at least 6 years' musical training (List \& Justus, 2007), and
in another, all participants had at least 2 or 3 years, with each participant group reporting a median of 10 years' training (List \& Justus, 2010). Other studies of music perception vary, such as Levitin and colleagues performing fMRI studies on musically-naïve participants who could have up to two years of training (Sridharan, Levitin, \& Menon, 2008; Sridharan, Levitin, Chafe, Berger, \& Menon, 2007) and other studies where participants all had at least five years' training (Vines, Krumhansl, Wanderley, Dalca, \& Levitin, 2011; Vines, Krumhansl, Wanderley, \& Levitin, 2006). Ouimet and colleagues (2012) defined musicians as those reporting 7 or more years (versus non-musicians with less than 3 years) of musical training. If most of their non-musicians had 1-2 years' training, then their finding of global-local precedence in that group comes closer to paralleling our findings. If most of their non-musicians had no musical training at all, the contrast between their and our findings would argue all the more strongly for an examination of the stimulus and task differences. None of the precedents listed above assessed differences between those with little training and those with none. This omission may stem in part from widespread reliance on undergraduates as study participants: finding 20 untrained participants required recruiting over 100. However, since a mere ten hours of video game playing can resolve perceptual deficits previously considered to be genetically driven (Feng, Spence, \& Pratt, 2007), examining constrasts between untrained and minimally trained listeners is essential.

Refining our examination of perceptual acuity associated with small amounts of musical training increases a risk of possible person confounds in the interpretation of our results. Are individuals with better auditory discrimination more likely to have and/or to continue with music instruction? While this risk cannot be ruled out, our results go beyond what would be predicted from person differences alone. If greater accuracy on our task were purely a matter of better auditory acuity, which may indeed influence individuals’ choice to seek or persist in musical training, there would be differences in overall accuracy as found here, but not an interaction with task focus. Another concern, though, might be that novices' greater difficulty with the overall task might disguise a truly universal global-local precedence that can only be detected among those with greater perceptual acuity. This interpretation was not supported by our data, as shown by a follow-up analysis of each participant group. Under the disguise theory, task accuracy should correlate positively with global-local precedence. As shown in Figure 3, all participant groups showed a zero or, in one case, negative correlation between task accuracy and global-local precedence. In simpler terms, better accuracy with this task did not explain the greater global-local precedence found among more-expert listeners.

There are many ways in which our definition of musical training or expertise could and should be made sharper. Error rates in this data set provide some support for concerns about differing performance across the semester within a participant


Figure 3: Non-correlation between task accuracy and globallocal precedence, assessed separately for each participant group according to their prior musical training.
pool. Even with these concerns, though, our results demonstrate a consistent difference according to musical training in an attribute that is more typically presented as a universal trait in healthy populations. Ongoing scientific interest both in the global-local distinction and in cognitive changes associated with musical training suggests that these results may bear on a wide range of investigations and may specifically help future studies of auditory processing avoid a confound of participant expertise.

## Acknowledgments

The authors would like to express our gratitude to Lisa Sanders for sharing stimuli and helpful feedback, and to John Limber for improving an earlier draft of this submission.

## References

Bever, T. G., \& Chiarello, R. J. (1974). Cerebral dominance in musicians and nonmusicians. Science, 185(4150), 537539.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. Glot International, 5, 341-345.
Feng, J., Spence, I., \& Pratt, J. (2007). Playing an action video game reduces gender differences in spatial cognition. Psychological Science, 18(10), 850-855(6).
Foxton, J. M., Brown, A. C., Chambers, S., \& Griffiths, T. D. (2004). Training improves acoustic pattern perception. Current biology, 14, 322-325.
Foxton, J. M., Talcott, J., Witton, C., Brace, H., McIntyre, F., \& Griffiths, T. D. (2003). Reading skills are related to global, but not local, acoustic pattern perception. Nature Neuroscience, 6, 1097-6256.
Fujioka, T., Trainor, L. J., Ross, B., Kakigi, R., \& Pantev, C. (2004). Musical training enhances automatic encoding of melodic contour and interval structure. Journal of cognitive neuroscience, 16, 1010-1021.

Green, C. S., \& Bavelier, D. (2003). Action video games modify visual selective attention. Nature, 423, 534-537.
Grimm, L. R., Markman, A. B., \& Maddox, W. (in press). Semester timing and reward influence performance: End-of-semester syndrome: How situational regulatory fit affects test performance over an academic semester. Basic and Applied Social Psychology.
Johnsrude, I., Penhune, V., \& Zatorre, R. (2000). Functional specificity in right human auditory cortex for perceiving pitch direction. Brain, 123, 155-163.
Justus, T., \& List, A. (2005). Auditory attention to frequency and time: an analogy to visual local-global stimuli. Cognition, 98(1), 31-51.
Kellar, L. A., \& Bever, T. G. (1980). Hemispheric asymmetries in the perception of musical intervals as a function of musical experience and family handedness background. Brain and Language, 10(1), 24-38.
List, A., \& Justus, T. (2007). Auditory priming of frequency and temporal information: Effects of lateralised presentation. Laterality, 12(6), 507-535.
List, A., \& Justus, T. (2010). Relative priming of temporal local-global levels in auditory hierarchical stimuli. Attention, Perception, \& Psychophysics, 72, 193-208.
List, A., Justus, T., Robertson, L. C., \& Bentin, S. (2007). A mismatch negativity study of local-global auditory processing. Brain Research, 1153, 122-133.
Love, B. C., Rouder, J. N., \& Wisniewski, E. J. (1999). A structural account of global and local processing. Cognitive Psychology, 38, 291-316.
Moreno, S., Marques, C., Santos, A., Santos, M., Castro, S. L., \& Besson, M. (2009). Musical training influences linguistic abilities in 8 -year-old children: More evidence for brain plasticity. Cerebral Cortex, 19, 712-723.
Mottron, L., Peretz, I., \& Ménard, E. (2000). Local and global processing of music in high-functioning persons with autism: Beyond cerebral coherence? Journal of Child Psychology and Psychiatry, 41, 1057-1065.
Mueller, S. T. (2006). PEBL: The Psychology Experiment Building Language. (http://pebl.sourceforge.net/)
Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. Cognitive Psychology, 9, 353-383.
Navon, D. (2003). What does a compound letter tell the psychologist's mind? Acta Psychologica, 114, 273-309.
Ouimet, T., Foster, N., \& Hyde, K. (2012). Auditory globallocal processing: Effects of attention and musical experience. Journal of the Acoustical Society of America, 132, 2536-2544.
Poirel, N., Simon, G., Cassotti, M., Leroux, G., Perchey, G., Lanoë, C., et al. (2011). The shift from local to global visual processing in 6-year-old children is associated with grey matter loss. PLoS ONE, 6(6), e20879.
R Development Core Team. (2005). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from
http://www.R-project.org
Ripoll, T., \& Marty, J. (2005). The role of local and global properties in comparison of analogical visual scenes. Psychonomic Bulletin \& Review, 12, 178-184.
Rondan, C., \& Deruelle, C. (2007). Global and configural visual processing in adults with autism and Asperger syndrome. Research in developmental disabilities, 28, 197206.

Sanders, L. D., \& Poeppel, D. (2007). Local and global auditory processing: Behavioral and ERP evidence. Neuropsychologia, 45(6), 1172-1186.
Schellenberg, E. G., \& Peretz, I. (2007). Music, language and cognition: unresolved issues. Trends in Cognitive Sciences, 12, 45-46.
Sridharan, D., Levitin, D. J., Chafe, C. H., Berger, J., \& Menon, V. (2007). Neural dynamics of event segmentation in music: Converging evidence for dissociable ventral and dorsal networks. Neuron, 55, 521-532.
Sridharan, D., Levitin, D. J., \& Menon, V. (2008). A critical role for the right fronto-insular cortex in switching between central-executive and default-mode networks. Proceedings of the National Academy of Sciences of the United States of America, 105, 12569-12574.
Vines, B. W., Krumhansl, C. L., Wanderley, M. M., Dalca, I. M., \& Levitin, D. J. (2011). Music to my eyes: Crossmodal interactions in the perception of emotions in musical performance. Cognition, 118, 157-170.
Vines, B. W., Krumhansl, C. L., Wanderley, M. M., \& Levitin, D. J. (2006). Cross-modal interactions in the perception of musical performance. Cognition, 101, 80-113.
Warrier, C., \& Zatorre, R. (2002). Influence of tonal context and timbral variation on perception of pitch. Perception and Psychophysics, 64, 198-207.

# Visual Strategies in Analogical Reasoning Development: A New Method for Classifying Scanpaths 

Yannick Glady, Jean-Pierre Thibaut, Robert French<br>\{yannick.glady, jean-pierre.thibaut, robert.french\}@u-bourgogne.fr LEAD-CNRS, UMR 5022, University of Burgundy, Pôle AAFE - Esplanade Erasme<br>21065 DIJON. FRANCE


#### Abstract

Development of analogical reasoning is often explained by general maturation of executive functions. A consequence of the involvement of executive functions would be that children and adults differ in the visual strategies they apply when solving analogical problems. Since visual strategies can be studied by means of eye-tracking, we compared the visual scanpaths of children and adults in three different analogical reasoning tasks. This comparison was done by means of a novel technique that combined a recently developed algorithm for computing a "distance" between any pair of scanpaths (Jarodzka, Holmqvist, \& Nyström, 2010), multidimensional scaling (MDS), and a neural network classifier. This analysis clearly showed a difference between adults' and children's visual strategies in solving analogy problems. We focus both on the demonstration that adults and children employ different visual search strategies to solve analogy problems and on the novel technique used to do this. This general technique complements other approaches to eye-movement analysis that rely on local properties of scanpaths, in particular, itemfixation times.


Keywords: Analogical reasoning; development; eye-tracking; strategies.

## Introduction

Analogical reasoning is a ubiquitous process in thinking and reasoning (Gentner \& Smith, 2012; Holyoak, 2012). It can be defined as a comparison of two domains (the source and the target domains) on the basis of their respective relational structure (Gentner, 1983). Studies of analogy making have explored two main explanations for its development, increase of structured knowledge (Gentner \& Rattermann, 1991; Goswami, 1992) and maturation of executive functions (Halford, 1993; Richland, Morrison, \& Holyoak, 2006; Thibaut, French, \& Vezneva, 2010a, 2010b). One important prediction of the executive-function view is that children and adults use different strategies when solving analogy problems. The present study addressed this question by means of a combination of a recently developed algorithm (Jarodzka et al., 2010) for comparing visual scanpaths from an eye-tracker, multi-dimensional scaling (MDS), and a neural net classifier. This technique allowed us to give an affirmative answer to the central question of this paper - namely, whether or not children's analogy strategies are quantifiably different than those of adults.

## Background

Humans rely heavily on vision for virtually every task they do (e.g. categorization, spatial orientation, problem solving, etc.) and it remains a privileged way of acquiring information about the environment. In the case of problem solving, what information is sought and how this search is organized through time to come to a solution for the problem (i.e. visual strategies) may help researchers understand which solving strategies are used. Attention and gaze-fixation are highly correlated, especially for complex stimuli (Deubel \& Schneider, 1996; He \& Kowler, 1992) and the fixation time for a given object is correlated with its informativeness in a scene (Nodine, Carmody, \& Kundel, 1978). This argues in favor of studying eye-movements as indicators of the application of a specific strategy through control of attention.

Eye-tracking data, especially if they involve scanpaths i.e., the complete visual trajectory of a participant's eye movements during the task - are often complex and hard to analyze. For this reason scanpath information is often reduced to static information about the participant's gaze times at specified locations. This simplification, while certainly easier to analyze, generally fails to fully capture the temporal aspects of the data involved in visual strategies. Even when an attempt is made to take into account temporal aspects of the data, it is often difficult to compare two scanpaths because, in general, they differ in length and complexity. Jarodzka et al. (2010) have developed a method that is able to compare any two scanpaths. As the Jarodzka et al. algorithm plays a key role in the analysis that follows, we will describe our variant of this algorithm in some detail below. We combined this scanpath-comparison algorithm with multidimensional scaling and a neural-network classifier to demonstrate that children's analogy-making strategies, as reflected in their visual search patterns across three different problems, are measurably different from those of adults.

We are not the first to use eye-tracking technology to study analogy making, but this type of analysis is, nonetheless, still in its infancy. Eye-tracking techniques were first used by Bethell-Fox, Lohman, \& Snow (1984) to study strategies when reasoning by analogy. They found strategic differences in adults with high or low fluid
intelligence when solving geometric A:B::C:? problems. More recently, Gordon \& Moser (2007) investigated adults' strategies in scene analogy problems. Thibaut, French, Missault, Gérard, \& Glady (2011) also used an eye-tracker to examine infants' gaze locations and item-to-item transitions during an analogy task. However, all of these studies focused on what information was searched for by participants as they attempted to solve the analogy problem.

None of this research compared participants' global scanpaths. In other words, previous eye-tracking studies have focused on local aspects of participants' scanpaths as a means of revealing part of the dynamics of visual search in doing analogy problems. By contrast, in the present study we will use participants' global scanpaths in our attempt to respond to the question of whether children have different visual search strategies than adults when solving visual analogy problems. Woods et al. (2013) showed that the organization of search in visual-attention tasks becomes less variable over the course of development. Because the tasks we used rely on visual attention, we expected children to have more variable scanpaths than adults.

## Experiment

## Methods

## Participants

Subjects were 20 adults ( 14 females, 6 males; mean age $=20 ; 5$ years; $\mathrm{SD}=2.21$; range: 17 to 27), students at the University of Burgundy and naïve to analogical reasoning tasks and 266 -year-olds ( 16 females, 10 males; mean age= 79.5 months; $S D=3.6$; range: 73 to 84 ). For children participating in this experiment, parents' informed consent was required from their parents.

## Materials

Three tasks, each composed of three training trials and four experimental trials, constituted the experiment (see Figure 1). The first task was a scene analogy problem task, the second a standard $\mathrm{A}: \mathrm{B}:: \mathrm{C}:$ ? task and the third an $\mathrm{A}: \mathrm{B}:: \mathrm{C}:$ ? task with the items composing the problems put within a context. Each problem of each task was composed of 7 black and white line drawings.

In the scene analogy problems, the top scene was composed of two elements depicting a binary semantic relation (e.g. a cat chasing a mouse). One of these two elements had an arrow pointing to it. The bottom scene was composed of five drawings: the two elements depicting the same relation as in the top picture (e.g. a boy chasing a girl), a distractor item, and two elements that were consistent with the scene but that had no salient relation with the elements of the relation. These pictures (501x376 pxs) were based on Richland et al., (2006) except for the distractor that was chosen not to be perceptually, only semantically, related to one member of the relation in the bottom picture.

In the standard $\mathrm{A}: \mathrm{B}:: \mathrm{C}:$ ? trials, the $\mathrm{A}, \mathrm{B}, \mathrm{C}$ drawings were presented in the top row along with a black empty square symbolizing the location of the solution. The four remaining pictures (the Target, a Related-to-C Distractor, and two Unrelated Distractors) were presented in a row at the bottom of the screen. The size of each picture was 200 x 195 pxs. The A:B::C:? task within context was constituted of two scenes ( 501 x 376 pxs ). The top picture was composed of two black and white line drawings with a relation between them (e.g. a wolf and meat, with the wolf looking at the meat) with a contextual cue (e.g. a horizontal line for the horizon or the lines of the joining walls and floor for a room). The bottom picture was composed of the five remaining drawings: the C term, the Target, the Related-to-C Distractor and the two Unrelated Distractors. This task differed from the first task in that it was the C term that was


Figure 1. Presentation of the three tasks used for this experiment: a) scene analogy task, b) standard $\mathrm{A}: \mathrm{B}:: \mathrm{C}:$ ? task, c) sceneoriented A:B::C:? task
pointed at with an arrow, and not one of the elements constituting the source relation. It differed from the second task because of the different pictures constituting the problems being grouped in two scenes, but equivalent to the standard $\mathrm{A}: \mathrm{B}:: \mathrm{C}:$ ? task in other respects.

The materials of the last two tasks were based on materials previously used by Thibaut et al. (2011). The four trials of each task were two trials with weak association strengths between A and B, C and T, and C and Dis, and two with strong association strengths in order to equilibrate this factor.

The tasks were displayed on a Tobii T120 eye-tracker device with a $1024 \times 768$ screen resolution.

## Procedure

Appropriate controls were carried out to ensure that the participants knew what the items in each of the problems were and that they understood the instructions. In the first task, they were asked to point to the element in the bottom scene that played the same role as the one which had an arrow pointing to it in the top scene. The two others tasks were administered as in Thibaut et al. (2011). Eye-tracking data was gathered from moment of the initial presentation of the problem to the moment a choice of one of the answers was made. The participant's scanpath for a particular problem consisted of a record of his/her gaze-fixation points taken every 8 ms .

## Data Analysis



Figure 2. Simplification of a scanpath
The goal of this analysis is to compare the sets of children's and adults' scanpaths and to show that there are quantifiable differences in the two. To do this we use a combination of (a variant of) Jarodzka et al.'s (2010) scanpath-comparison algorithm, multidimensional scaling and a neural-net classifier. As the latter two techniques are well known, we will not discuss them at length. However, the Jarodzka et al. algorithm is relatively recent and requires explanation.

## Jarodzka et al. (2010) scanpath-comparison algorithm

The algorithm is designed to determine the similarity of any two scanpaths. It consists of two phases, a simplification
phase and a comparison phase. A scanpath is considered to be made up of a series of "saccade vectors," i.e., a connected series of vectors whose endpoints correspond to coordinates of successive gaze points (Figure 2a). First, the scanpath is simplified by combining into a single vector two consecutive saccade vectors if:
i) their combined length does not exceed 200 pixels in amplitude (i.e., each is very small) and
ii) they are nearly in straight line (i.e., the angle between them is between 2.62 and 3.67 rad ).
In other words if a saccade vector is very small or very linear with respect to its predecessor in the scanpath, the two vectors are combined (Figure 2b).

Once each of the two scanpaths has been simplified, they can be compared. We begin by giving an intuitive explanation of how this is done. Assume, for example, there are two simplified scanpaths, $S_{1}$ and $S_{2}$ made up of 3 and saccade vectors, respectively. In other words, $S_{1}=\left\{u_{1}, u_{2}\right.$, $\left.\mathrm{u}_{3}\right\}$ and $\mathrm{S}_{2}=\left\{\mathrm{v}_{1}, \mathrm{v}_{2}, \mathrm{v}_{3}, \mathrm{v}_{4}\right\}$. Note that these saccade

a) |  | $\boldsymbol{v}_{1}$ | $\boldsymbol{v}_{2}$ | $\boldsymbol{v}_{3}$ | $\boldsymbol{v}_{4}$ |
| :---: | :---: | :---: | :---: | :---: |
| $\boldsymbol{\mu}_{1}$ | $\Delta\left(\mathrm{u}_{1}, \mathrm{v}_{1}\right)$ | $\Delta\left(\mathrm{u}_{1}, \mathrm{v}_{2}\right)$ | $\Delta\left(\mathrm{u}_{1}, \mathrm{v}_{3}\right)$ | $\Delta\left(\mathrm{u}_{1}, \mathrm{v}_{4}\right)$ |
| $\boldsymbol{\mu}_{2}$ | $\Delta\left(\mathrm{u}_{2}, \mathrm{v}_{1}\right)$ | $\Delta\left(\mathrm{u}_{2}, \mathrm{v}_{2}\right)$ | $\Delta\left(\mathrm{u}_{2}, \mathrm{v}_{3}\right)$ | $\Delta\left(\mathrm{u}_{2}, \mathrm{v}_{4}\right)$ |
| $\boldsymbol{\mu}_{3}$ | $\Delta\left(\mathrm{u}_{3}, \mathrm{v}_{1}\right)$ | $\Delta\left(\mathrm{u}_{3}, \mathrm{v}_{2}\right)$ | $\Delta\left(\mathrm{u}_{3}, \mathrm{v}_{3}\right)$ | $\Delta\left(\mathrm{u}_{3}, \mathrm{v}_{4}\right)$ |

b)


Figure 3. Saccade-vector difference table (a): Each of the saccade vectors from each of the two scanpaths are compared based on the chosen metric. (b) The comparison of each pair of stretched scanpaths corresponds to a traverse of the table from the upper-left to the lower-right corner of the saccade-vector difference matrix (the only directions of movement permitted are down, right and diagonally down-and-right). We find the path that produces the lowest total difference value and this value is the similarity measure assigned to $\mathrm{S}_{1}$ and $\mathrm{S}_{2}$
vectors are ordered in time. For example, in $\mathrm{S}_{1}$, the saccade vector $u_{1}$ is followed by $u_{2}$, which is followed by $u_{3}$. To compare $S_{1}$ and $S_{2}$, we need two scanpaths of the same length. To achieve this, we will "stretch" each scanpath by adding immediate repetitions of saccade vectors, so that
they both have the same length. Our goal is to find the two stretched scanpaths, $\mathrm{SS}_{1}$ and $\mathrm{SS}_{2}$ that are as similar as possible with respect to the chosen metric (orientation, length, etc.). This similarity will be the measure of the distance between $S_{1}$ and $S_{2}$.

The easiest way to illustrate this stretching is by means of a saccade-vector difference table for the two scanpaths, $\mathrm{S}_{1}$ and $S_{2}$, defined above.

A saccade-vector difference matrix is first created (Figure 3a). Each of the saccade-vectors making up one of the scanpaths $S_{1}$ is compared to each of the saccade-vectors of the other scanpath $\mathrm{S}_{2}$, according to a metric, generally, vector magnitude or orientation (length in our study). Once this table is constructed, we consider all paths through the table that begin with the comparison of the first saccade vectors in both scanpaths (i.e., cell $(1,1)$ of the table, $\Delta\left(\mathrm{u}_{1}\right.$, $\left.\mathrm{v}_{1}\right)$ ) and end with a comparison of the final saccade vectors in each scanpath (i.e., cell $(3,4)$ of the table, $\left.\Delta\left(u_{3}, v_{4}\right)\right)$ and always move to the right, down, or diagonally down-andright. Three examples of paths through the matrix are illustrated in the right-hand panel of Figure 3. Each path through the table corresponds to the comparison of two specific stretched scanpaths. For example, the uppermost path shown corresponds to a comparison between $\mathrm{SS}_{1}=\left\{\mathrm{u}_{1}\right.$, $\left.\mathrm{u}_{1}, \mathrm{u}_{1}, \mathrm{u}_{2}, \mathrm{u}_{2}, \mathrm{u}_{3}\right\}$ and $\mathrm{SS}_{2}=\left\{\mathrm{v}_{1}, \mathrm{v}_{2}, \mathrm{v}_{3}, \mathrm{v}_{3}, \mathrm{v}_{4}, \mathrm{v}_{4}\right\}$. This path corresponds to the sum of the values in the cells $(1,1),(1,2)$, $(1,3),(2,3),(2,4),(3,4)$ of the saccade-vector difference matrix. When all of these paths through the matrix are considered, the path which has the smallest value (i.e. the smallest cumulative sum of comparisons) is selected. This path corresponds to the two stretched scanpaths that are the most similar. This value, normalized by the number of comparisons done, is the similarity measure assigned to the comparison of scanpaths $\mathrm{S}_{1}$ and $\mathrm{S}_{2}$.

Note that the algorithm as described here differs from Jarodzka et al. (2010) in that it does not rely on the more complex Dijkstra (1959) tree-search algorithm. Instead, we constructed a matrix, cell by cell, with the lowest cumulative sum of comparisons possible for each cell while taking into account the constraints put on the comparisons of the two scanpaths (navigate rightward, downward, or diagonally downward and to the right). In our example, the final distance value between $S_{1}$ and $S_{2}$ is the cumulative sum in $\mathrm{C}(3,4)$ normalized by the number of steps taken through the matrix. This algorithm was computationally less complex for identical results.

The Jarodzka et al. (2010)/MDS/MLP algorithm applied to scanpaths of analogy problems

We only compared the scanpaths from strictly identical problems, but not different trials from the same task. Thus, when we were comparing an adult scanpath and a child's scanpath, the disposition of the items in the problem they were solving was identical.

In this way, for a given set of isomorphic problems (i.e., where all of the items were in identical places on the screen), we computed the differences between all pairs of scanpaths. In other words, if there were $S_{1}$ to $S_{n}$ scanpaths from children and $A_{1}$ to $A_{m}$ scanpaths from adults on the same set of isomorphic problems, we computed the similarity of all pairwise comparisons of scanpaths $S_{i}$ versus $\mathrm{S}_{\mathrm{j}}, \mathrm{S}_{\mathrm{i}}$ versus $\mathrm{A}_{\mathrm{j}}$, and $\mathrm{A}_{\mathrm{i}}$ versus $\mathrm{A}_{\mathrm{j}}$ for all $i$ and $j$.

Once we had calculated the mean differences between scanpaths generated by each participant in each task, we used Multidimensional Scaling to obtain the coordinates on a 2D map that best preserved the distance between scanpaths. As can be seen in Figure 4, for each of the three tasks, the scanpaths clustered according to participant type (Adult or Children). We verified this clustering using a 3layered perceptron (MLP) with a bias node on the input and hidden layers ( 5 hidden units, learning rate $=0.05$, momentum $=0.9$ ) with the coordinates of each scanpath on the MDS map translated into bipolar values and concatenated on input. We used a Leave-One-Out crossvalidation technique to test the robustness of the classification. Leave-One-Out cross-validation is a standard technique in machine learning whereby the classifier (in this case a neural network) is trained on all items but one. Once training is complete, the classifier is tested on the item that had been left out to see whether or not it is classified correctly.

## Results

Using the method of analysis described above, we did a pairwise comparison of all scanpaths generated by adults and children on isomorphic analogy problems. We then conducted a multi-dimensional scaling analysis of this data, which produced the location-map clusters shown in Figure 4. These points are a 2D representation that best reflects the distances between the scanpaths. The crosses correspond to children's scanpaths; the circles correspond to adults' scanpaths.

## Classification of adults' versus children's scanpaths

The Jarodzka et al. (2010) method along with Multidimensional Scaling led to a 2D location map that best represented the relative distances between the set of scanpaths, as calculated by the Jarodzka et al. algorithm (Figure 4). A three-layered feedforward backpropagation network (MLP) with a Leave-One-Out cross-validation method, was used to test the robustness of a classification of the points representing the two groups (i.e. children and adults). For the scene analogy and $\mathrm{A}: \mathrm{B}:: \mathrm{C}:$ ? tasks (Figure 1a and 1b), the network classified $74 \%$ of the participants correctly based on their scanpath ( $70 \%$ of the 20 adults and $78 \%$ of the 23 children for both tasks). For the real-world A:B::C:? task, the network classified $72 \%$ of the subjects correctly ( $65 \%$ of the adults and $78 \%$ of the children). This
was significantly above chance (50\%) for each task (binomial test: $\mathrm{Z}=14.89 ; \mathrm{p}<.001$ for the first and second; $\mathrm{Z}=14.30$; $\mathrm{p}<.001$ for the third). Intuitively, this result can be

b)

c)


Figure 4. Location-map of an MDS analysis of the relative differences among participants for the scene analogy task (a), the standard $\mathrm{A}: \mathrm{B}:: \mathrm{C}:$ ? task (b), and the scene-oriented A:B::C:? task (c).
seen in Figure 3. The adult group tends to be more homogenous than the children as the crosses (children's scanpaths) are more scattered than the circles (adults' scanpaths), and this is reflected in the high degree of accurate classification of the MLP.

## General discussion

The present study addressed the following question in a novel manner: Do children and adults have different visual strategies in analogical reasoning tasks? To answer this, we used an eye-tracking methodology whose data were analyzed by a combination of the Jarodzka et al. (2010) scanpath-comparison algorithm, the transformation of this data into a 2D location map using multidimensional scaling, and, finally, a quantitative adult/child classification by means of a feedforward backpropagation network. The neural-net classification was done by training the network on the scanpath data for all but one participant. Once the network was trained, it was tested on the one scanpath that was left out of the training set. This was done for each participant's scanpath data and the result was scored according to whether the network classified the test scanpath correctly or not. The results obtained with this method agree with previous results from Thibaut et al. 2011 who also showed, by analyzing item gaze times and the number of transitions between items that adults and children differed in their search strategies in the standard $\mathrm{A}: \mathrm{B}:: \mathrm{C}:$ ? analogy task. The present work, using an approach based on individuals' entire scanpaths, also extends this previous work to scene analogy problems and scene-oriented A:B::C:? problems. This scanpath analysis showed, among other things, that children's scanpaths were more variable than those of adults in the three tasks. These differences support the hypothesis of the key role of executive functions in analogy making because the lower variability of adults' scanpaths is indicative of them applying, through control of attention, a previously adopted plan for solving analogy problems (Woods et al., 2013)

The scanpath analysis presented in this paper provides a means of studying various search strategies in analogy making. The technique presented in this paper overcomes thorny problem of comparison of scanpaths of different lengths and allows to take into account the dynamic features of search, which are largely missed in other, more static eye-tracking approaches based on item fixation times. It could also be used, for example, to confirm differences in analogy-making strategies observed in adults in Bethell-Fox et al. (1984) and to classify participants based on their scanpath data (i.e., "elimination strategies" for participants with low fluid intelligence and "constructive matching strategies" for participants with high fluid intelligence). This method is, of course, not limited to studies of analogymaking, and could be used with any other type of problems whose crucial information for its solution could be presented on a screen.

## Conclusion

The method of scanpath analysis presented in this paper provides a new tool to analyze the dynamic aspects of search strategies in a wide variety of experimental contexts. As shown by the results, this method is sensitive to global differences between scanpaths and is useful to discriminate clusters of strategies. In this paper it has been used to show that children's and adults' differ in their variability while solving analogical reasoning problems, suggesting the involvement of executive functions in such tasks. However, to fully understand the causes of these differences, it is inevitable to use local information. Thus, it should be used in combination of other existing methods, in particular, Area-of-Interest (AOI) methods that provide information on what information is sought and how long it is watched (informativeness of stimuli), since this information is not captured by the Jarodzka et al. method. On the other hand, AOI methods give limited information about the dynamic progression of search, something which is captured when full scanpath information is used. In short, the Jarodzka et al. (2010), combined with an MDS analysis and a classifier (backpropagation networks, Support Vector Machines, etc.), provides a potentially far-reaching tool for analyzing participants' dynamic strategies in various problem-solving contexts.

## Acknowledgements

This research has been supported by French ANR Grant 10-BLAN-1908-01 to the second author and a joint ANRESRC grant 10-065 GETPIMA to the last author.

## References

Bethell-Fox, C. E., Lohman, D. F., \& Snow, R. E. (1984). Adaptive reasoning: Componential and eye movement analysis of geometric analogy performance. Intelligence, 8(3), 205-238.
Deubel, H., \& Schneider, W. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. Vision research, 36, 1827-1837.
Dijkstra, E. (1959). A note on two problems in connexion with graphs. Numerische mathematik, 1, 269-271.
Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. Cognitive Science, 7(2), 155-170.
Gentner, D., \& Rattermann, M. J. (1991). Language and the career of similarity. In S. A. Gelman \& J. P. Byrnes (Eds.), Perspectives on Language and Thought: Interrelations in Development (pp. 225-277). New York, NY: Cambridge University Press.

Gentner, D., \& Smith, L. (2012). Analogical reasoning. Encyclopedia of Human Behavior (2nd ed., Vol. 1, pp. 130-136). Elsevier Inc.
Gordon, P. C., \& Moser, S. (2007). Insight into analogies: Evidence from eye movements. Visual Cognition, 15(1), 20-35.
Goswami, U. (1992). Analogical Reasoning in Children. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
Halford, G. S. (1993). Children's Understanding: The Development of Mental Models. Hillsdale, NJ: Erlbaum.
He, P., \& Kowler, E. (1992). The role of saccades in the perception of texture patterns. Vision research, 32(11), 2151-2163.
Holyoak, K. J. (2012). Analogy and relational reasoning. In K.J. Holyoak \& R. G. Morrison (Eds.), The Oxford Handbook of Thinking and Reasoning (pp. 234-259). New York, NY: Oxford University Press.
Jarodzka, H., Holmqvist, K., \& Nyström, M. (2010). A vector-based, multidimensional scanpath similarity measure. ETRA '10: Proceedings of the 2010 Symposium on Eye Tracking Research \& Applications (pp. 211-218). New York, NY.
Nodine, C. E., Carmody, D. P., \& Kundel, H. L. (1978). Searching for Nina. In J. Senders, D. F. Fisher, \& R. Monty (Eds.), Eye movements and the higher psychological functions (pp. 241-258). Hillsdale, NJ: Erlbaum.
Richland, L. E., Morrison, R. G., \& Holyoak, K. J. (2006). Children's development of analogical reasoning: Insights from scene analogy problems. Journal of Experimental Child Psychology, 94(3), 249-273.
Thibaut, J.-P., French, R. M., Missault, A., Gérard, Y., \& Glady, Y. (2011). In the eyes of the beholder: What eyetracking reveals about analogy-making strategies in children and adults. Proceedings of the Thirty-Third Annual Meeting of the Cognitive Science Society (pp. 453-458).
Thibaut, J.-P., French, R. M., \& Vezneva, M. (2010a). Cognitive load and semantic analogies: Searching semantic space. Psychonomic Bulletin \& Review, 17(4), 569-74.
Thibaut, J.-P., French, R. M., \& Vezneva, M. (2010b). The development of analogy making in children: Cognitive load and executive functions. Journal of Experimental Child Psychology, 106(1), 1-19.
Woods, A. J., Göksun, T., Chatterjee, A., Zelonis, S., Mehta, A., \& Smith, S. E. (2013). The development of organized visual search. Acta psychologica, 143(2), 191199.

# A Critical View on Conceptual Blending Theory 

V.V. Glebkin (gleb1514@gmail.com)<br>Gymnasium 1514, 12 Krupskoi Street Moscow, 119311 Russia


#### Abstract

This study addresses conceptual blending theory, originated by J. Fauconnier and M. Turner. The author raises some criticism of the theory's underpinnings and methodology. Particularly, he points at the lack of cultural-historical analysis and the neglect of experimental data as the shortcomings of the theory as stated. It is shown in the paper that the view on blending as an important tool to adapt knowledge to the experience of average people is more correct than its interpretation as a basic instrument for the creation of new knowledge.


Keywords: conceptual blending; criticism; cultural-historical approach; popularization.

## Introduction

Conceptual blending (or conceptual integration) theory is, without a doubt, one of the central conceptual pillars of modern cognitive linguistics, and it has considerable influence on cognitive science in general. According to the authors, J. Fauconnier and M. Turner, conceptual blending is "a great mental capacity that, in its most advanced "doublescope" form, gave our ancestors superiority and, for better and for worse, made us what we are today. We investigate the principles of conceptual blending, its fascinating dynamics, and its crucial role in how we think and live" (Fauconnier, Turner 2002, V; for an extended commentary see: ibid., 389-396). In other words, conceptual blending here is claimed to be the key to the mystery of human evolution and cognition.

At first sight, such a groundbreaking point would be expected to lead to an intensive debate and to meet strong criticism from the less radical researchers. However, in fact, there are no heated debate around conceptual blending theory. We can find a number of papers raising some objections (Gibbs 2000; Harder 2003; Brandt 2005; Oakley, Hougaard 2008, 12; Ferguson, Sanford 2008, 610), but a comprehensive analysis of the theory's underpinnings, methodology, and heuristic potential is a matter of the future. This paper can be considered as a step in that direction.

## Theoretical underpinnings and structure of conceptual blending theory

Although this may be familiar information to some of readers, I will start with a coarse-grained description of the Fauconnier and Turner's attitudes. It allows more relevant understanding of my criticism in the second part of the paper. To be sure, some aspects of conceptual blending theory were touched earlier, but its first systematic description holds, presumably, in Fauconnier, Turner 1994.

Fauconnier, Turner 1996; Fauconnier, Turner 1998; Fauconnier, Turner 2000; Sweetzer 2000; Fauconnier, Turner 2002; Fauconnier, Turner 2008; Fauconnier 2009 should be mentioned as the salient milestones in the theory's development.

The gist of the theory can be formulated as follows: a) the unique feature of human beings is the capacity to create new meanings from existing ones; $b$ ) the main way to implement this capacity is to perform double-scope blending, that is, to build an integrated mental space on the base of a number of input spaces.

A star example illustrating that point is the Buddhist Monk riddle: "A Buddhist monk begins at dawn one day walking up a mountain, reaches the top at sunset, meditates at the top overnight until, at dawn, he begins to walk back to the foot of the mountain, which he reaches at sunset. Make no assumptions about his starting or stopping or about his pace during the trips. Riddle: is there a place on the path that the monk occupies at the same hour of the day on the two separate journeys?" (Fauconnier, Turner 2002, 39).

This riddle has an elegant solution if we imagine the monk strolling up and down on the same day, in other words, combine both walks. In such a blended space the monk is to meet himself and that place is the positive answer to the riddle question. The authors illustrate the solution with the following schema:


Fig. 1. The basic schema of the Buddhist Monk riddle (Fauconnier, Turner 2002, 45)

We can see here the two input spaces (the day of climbing on and that of climbing down), blended space and generic space, containing "what the inputs have in
common: a moving individual and his position, a path linking foot and summit of the mountain, a day of travel, and motion in an unspecified direction" (ibid., 41)).

Another striking example is "The Debate with Kant" ${ }^{1}$. Authors suggest to imagine a contemporary philosopher discussing the issue whether reason is innate capacity when leading a seminar. During that dispute he appeals to Kant as his opponent, namely, states his point, then poses hypothetical objections retrieved from Kant's treatises, then again come up with his own counterarguments, etc. For the audience it looks as face-to-face debate of two modern scholars. For the authors we have here the two input spaces connected with modern philosopher making claims in English and with Kant thinking and writing German. In the blend we find two philosophers speaking English to discuss ultimate philosophical problems. Thus, the blended space emergent structure in some aspects differs from that of input spaces radically reflecting the novel mental (but not ontological) reality.

The next example is so called Regatta. The backstory for it holds such facts: "The clipper ship Northern Light sailed in 1853 from San Francisco to Boston in 76 days, 8 hours. That time was still the fastest on record in 1993, when a mod-catamaran, Great American II, set out on the same course" (ibid., 63). According to authors, "a few days before the catamaran reached Boston, observers were able to say: At this point, Great American II is 4.5 days ahead of Northern Light" (ibid.). This sentence constructs blended space in close similarity with the Buddhist Monk riddle; like two monk trips above, two 140 -yeardistanced voyages are combined into a novel event, in this case, into the boat race. Such time scale compression paves the way for a pictorial perception of two voyages just as "The Debate with Kant" blend provides the audience's emotional engagement in the process of philosophical reasoning.

How wide the field covered by conceptual blending theory is, can be illustrated with two examples below. The first one is complex numbers, the second one is the computer desktop. As is well-known, complex numbers, expressed in the form $a+b i$, can be viewed as points or position vectors in a two-dimensional coordinate system called the complex plane where the real part of a number is represented with the horizontal projection, and the imaginary part with the vertical one. The authors suppose this representation to be the blend, where the first input space holds points in oriented plane with its vector transformations, whereas the second one contains real numbers with operations of addition and multiplication. The generic space in this case holds commutative ring operations on pairs of elements. In that perspective the blended space have a number of new features with regard to input ones (unlike real numbers there is not an order relation between two complex ones; unlike points and vectors complex numbers can be multiplied and divided by each other). So, for

[^278]authors, conceptual blending is an important tool used to create the new knowledge in mathematics (see also Alexander 2011).

A computer desktop, for authors, is the blend of our day-to-day experience space (where we open folders to place or extract documents, throw old folders into a trash can, etc.) and space of formal operations performed in a computer (an abstract language of computer commands which correspond to virtual motions in the blend). In this case, again, the structure of the blend has obvious novelty as compared with input spaces.

Conceptual blending, the authors state, is an important tool to create novel grammatical and lexical constructions in language. Thus, the construction Noun-Phrase Verb Noun-Phrase Prepositional-Phrase, found in a great number of languages to express caused motion, is a conceptual blend of two different actions (e.g., Jack threw the ball into the basket includes three steps: Jack acts on the ball; the ball moves; the ball is in the basket. The blend combines the beginning and the end of the action, omitting the middle part). In some languages, like English, it can be extend to some other classes of action, e.g., Anna sneezed the napkin off the table or The commander let the tank into the compound, etc. (Fauconnier, Turner 1996; cf. Mandelblit 2000).

Such figures of speech as metaphor and metonymy are also blends, according to the authors. They describe highly conventional source-target metaphors as singlescope networks, where the integrated space frame is supplied by only one input space; e.g., Murdoch knocked Iacossa out for companies Murdoch and Iacossa. The integrated space here is a blend of a boxing match and business competition, but the blend topology has no in common with business, it is completely defined by the boxing space frame (Fauconnier, Turner 2002, 126-129).

Meanwhile, only a narrow class of metaphors can be represented by single-scope networks. For instance, this representation is impossible for the expression digging one's own grave (e.g., They dug their own financial grave). In this case the blend inherits the structure from "digging the grave" and "unwitting failure" inputs. However, in the blend the input frames are not simply juxtaposed; the emergent structure is radically different from both of them in some aspects. Thus, in "digging the grave" space people dig the grave not for themselves, but for other people, who have already died. Furthermore, this action here is not a big blunder as opposed to the blend frame. In a similar vein, unwittingly failed (particularly, in the financial sphere) person is unlikely to operate with a real spade as his counterpart in the blend.

This type of integration frame entitled double-scope network characterizes, as mentioned, the unique human capacity distinguished human beings from other species. They date the emergence of this capacity to the epoch about 50,000 years ago and draw the birth of religion, art and language as a conclusion (ibid., 180-187). The gist of their argumentation can be formulated in the form of the following syllogism: a) the general process to provide the
human culture development is the emergence of novel conceptual structures on the base of existing ones; b) in doublescope networks the emergent blend structure has a novel quality with regard to the input spaces; hence, c) doublescope conceptual integration can provide the development of a wide range of cultural forms created by humans.

It is noteworthy that Fauconnier and Turner don't analyze concrete data to argue for this point, and their approach looks a bit like an "ivory tower" theory. At the same time such analysis can be found in the papers of other researchers working within Fauconnier and Turner's paradigm, particularly, in Sweetser 2000. In her work the author describes a hypothetical buffalo hunting ritual where primitive hunters perform a ritual dance in order to provide success in a real hunt in the future. In accord with Sweetser's views, such a ritual is a blend "between (Input 1) the ritual setting and participants and (Input 2) a hunting scene and its participants" (ibid., 319). The blended space holds the new elements which have no counterpart in the input spaces; a buffalo rock painting is transformed here into a real buffalo which is struck in the ritual dance, etc.

A researcher working with blending as a real cognitive process encounters the two general questions: which phases does this process have in on-line regime and what are the criteria to select the elements in the input spaces for the projection into the blend. Fauconnier and Turner try to tackle these issues, although it is hard to say if they are clear about that.

According to authors, "there are three operations involved in constructing the blend: composition, completion, and elaboration" (Fauconnier, Turner 1998, 144). The first stage is characterized by composing the blend from the elements of input spaces; then, the blend is completed with a great range of background conceptual structure; and on the last stage it is developed "through imaginative mental simulation according to principles and logic in the blend" (ibid.). The model described, however, did not rest on any experimental data and it is a problem to check whether it holds water.

With regard to the second issue the authors speak about constitutive and governing principles of conceptual integration. The first ones are connected with the general laws of logic and the rules of language, the second ones are more flexible and not so strictly defined. The authors mark out the overarching goal driving all of the governing principles: Achieve Human Scale, and several subgoals, namely, Compress what is diffuse; Obtain global insight; Strengthen vital relations; Come up with a story; Go from Many to One. Alongside with that, they suggest a number of more concrete principles, such as, Topology Principle ("Other things being equal, set up the blend and the inputs so that useful topology in the inputs and their outer-space relations is reflected by inner-space relations in the blend" (Fauconnier, Turner 2002, 327)) or Integration Principle ("Achieve an integrated blend" (ibid., 328)), etc. Again, the algorithm of applying these principles to particular cases is not transparent; e.g., it is not clear how to find out if topology is useful or not.

Given the overview of conceptual blending theory as completed here, let us move on to the next step.

## Some objections against conceptual blending theory

There are three aspects of conceptual blending theory as stated to be cast in doubt.
A) Contrary to Fauconnier and Turner's interpretation, almost all examples they suggest are connected with the form the conceptual operations are presented in, but not with their essence (the only exception is, perhaps, the quite specific Buddist Monk riddle). The main task of the blend in the examples given is to represent conceptual structure in a convenient, compact, familiar for an average person form; in other words, to provide its popular presentation ${ }^{2}$. The direct link between the popularization and conceptual blending is clear in the popular science literature, where extremely abstract ideas are represented in the form of visible images to be processed by the less educated audience. The S. Hawking's analogy between balloon in which cover tension holds air pressure within it and star where the gravitational interaction between atoms are balanced by the star gas pressure is a classic blend from conceptual blending theory perspective (the input space-1: a balloon, located near the earth's surface; a rubber cover; gas within the balloon; the input space-2: a star; helium and hydrogen atoms within the star; the blend: the star is a balloon with heated air inside, situated among other stars).

Another striking example of such "popular science blend" was suggested by A. Einstein to explain the space curvature in the general theory of relativity. It is the analogy between hypothetical "flat beings" existing in two dimensions and humans living in the three-dimensional space. The "flat beings" are able to perceive the line curvature, but not the space one; the humans can comprehend the curvature of the plane, but the curvature of the space is beyond their comprehension. The blend here holds human beings living inside the plane.

Let us now look from this perspective at Fauconnier and Turner's examples described above. Given "The Debate with Kant" story, the emergence of the blend has no impact on the essence of the problem discussed (in this case, as mentioned, the question whether reason is innate capacity). Such a debate could be held with another philosopher who shared Kant's ideas or, say, in inner dialog of the philosopher with himself. The format described by the authors gives the opportunity to adopt the disputed issue to the audience; it is a "pedagogical trick" transforming an abstract philosophical matter into a kind of

[^279]performance, in other words, coming up with a story.
The "Regatta" example has the same structure. The blend gives here a visible and attractive picture of the events, while touching no ground or even distorting their ground (thus, it is not clear, if the context of the regatta described is relevant to the clipper Northern Light, which goal it pursued during the voyage from San Francisco to Boston, etc.).

The computer desktop example is consistent with the ones examined above. The emergence of the blend helps an average user to work with a computer because it transforms an abstract machine language into the set of objects from his day-to-day experience. Here again we deal with the adaptation of the conceptual structure to the cognitive horizon of the lay observer.

The "complex numbers" case, which is, perhaps, the strongest argument for blending as creation of novel conceptual knowledge, rests on misunderstanding. Complex numbers, as mentioned, are an ordered pair of real numbers which can be represented as a point in oriented plane. However, real numbers can also be represented as a point in oriented line or as a vector which reference point coincides with the origin of coordinates. We can only wonder why the authors address such representation for complex numbers and don't apply it to real ones. A geometric representation of complex numbers is in demand much less than an algebraic or a trigonometric one. The gist of complex numbers has no connection with properties of points in oriented plane or plane vectors. The gist of complex numbers as expansion of real numbers is determined by introducing "imaginary unit" $\boldsymbol{i}(\boldsymbol{i}=\sqrt{ }-1)$ and by a lot of interesting properties connected with that. The analogy with points in plane provides visual image for complex numbers; hence, this case is situated in line with the ones examined above.

The construction Noun-Phrase Verb Noun-Phrase Prepositional-Phrase works, by and large, in the similar direction. Omitting middle links and pointing out only the beginning and the end of the process offer a more visible and dynamic process description which facilitates its perception.

To sum up my contention here, I would like to get to the general point: blending can not provide the emergence of new conceptual knowledge; its function is to adapt existing knowledge to the needs of average people. To consider blending as great mental capacity, which brought about the emergence of various forms of culture, means to put shoes on the wrong foot.
B) The lack of cultural-historical analysis is the next defect of the Fauconnier and Turner's approach. In order to illustrate that let us return to the metaphor digging one's own grave. This expression is likely to appear in different languages in the first decades of the $20^{\text {th }}$ century. According to The Oxford English Dictionary, its earliest example dates back to 1934. Similarly, in Russian language it arises as metaphor in the $1920^{\text {th }}$ alongside with its emergence in direct meaning (e.g., Chasto jertvy prinujdalis' ryt' sebe sami mogilu (Often victims were forced to dig their own grave,
1924)). If it is so, we can suppose that the metaphor digging one's own grave came into being as the comprehension of the new social experience of the first quarter of the $20^{\text {th }}$ century, reflected by literature and language. The picture of people digging their own graves at gun point turned out so vivid and emotionally affecting that it entailed its expansion in other regions. Given this assumption is correct the metaphor structure proves to radically differ from the authors' description. To be sure, we can find here double-scope blending (a man digging his own grave at gun point performs this action by necessity, whereas a man plunging into financial adventure operates by choice; so, it is not fairly correct to speak about direct mapping of the "digging" input space onto the blend), but the other thing is important. Importantly, this or that construction emerges in the here and now, but not in the transcendental reality; in that, its emergence is brought about by socio-cultural shifts, but not by abstract schemas like one suggested by Fauconnier and Turner.
E. Sweetser's interpretation of the buffalo hunting ritual is, perhaps, even a more representative example of distortions influenced by eliminating socio-cultural aspect from the analysis. In order to argue that we need to address the school of cultural-historical psychology and, particularly, the concept complex thinking originated by L. Vygotsky (Vygotsky 1986). The Russian psychologist defined complex as a structure where the bonds between its components are contextual and flexible rather than abstract and fixed. He illustrated his approach with the striking Darwin's example: «A child's use of 'quah' to designate first a duck swimming on the pond, then any liquid, including the milk in his bottle; when he happens to see a coin with an eagle on it, the coin is also called a 'quah', and then any round, coinlike object. This is typical of a chain complex ${ }^{3}$ - each new object included has some attribute in common with another element, but the attributes undergo endless changes» (ibid., 127). When thinking in complexes a child keeps in mind the same objects as an adult (which provides the right communication between them), but his way of operating these objects, and his mental schemas are radically different.

Given this concept as the ground, Vygotsky, among other things, explained the French cultural anthropologist L. Lévy-Bruhl's account of Bororo (the tribe of Northern Brazil) views which sounded counter to Aristotelian logic (Lévy-Bruhl 1978, Lévy-Bruhl 1979). For instance, the Bororo (the tribe in Northern Brazil) boasted that they were red araras (parakeets), which did not merely signify that they would become araras after their death, or that araras metamorphosed the Bororo, but they claimed that they were araras at the current time, which was their actual identity. Levi-Bruhl defined this operation as the law of participation and such way of reasoning as pre-logical thinking.

[^280]For Vygotsky, the Bororo and araras make up a single complex; they are not two discrete entity. A great number of such complexes can be found in primitive and ancient cultures. For instance, magical operations with the name of the enemy were used to damage him, and because of that people endeavored to keep their true name in secret.

The buffalo hunting ritual is the part of the phenomenon described above. A buffalo rock painting and a genuine buffalo are not the two separate objects; they are the elements of the same object, namely, a complex with the topology, rather unusual for the modern people. If it does, to find the blend in this ritual is an obvious mistake.

Reasons of this kind cover a lot of other expressions examined by Fauconnier and Turner (particularly, their analysis of the Grim Reaper metaphor). The lack of the cul-tural-historical component in the analysis leads (at least, in some cases) to failure to reveal the true causes of blend emergence and evolution and to account for the real conceptual structure of the phenomena investigated.
C) It is also to the point to address here a set of experimental researches providing us with some data to estimate the correctness of conceptual blending theory from the psycholinguistic perspective. We will focus on the investigation of counterfactual conditionals as a significant example of such researches (de Vega et al. 2007; de Vega 2008; Ferguson, Sanford 2008; de Vega, Uritta 2011). Fauconnier and Turner take great pains to examine counterfactuals. Such counterfactuals as If Clinton were the Titanic, the iceberg would sink alongside with the Buddhist Monk riddle or "The Debate with Kant" story would be considered as arch examples of conceptual blending. This sentence, for the authors, is the double-scope blend of Titanic and President Clinton mental spaces; in the blend Clinton collides with the iceberg and the iceberg is sinking. In this case blend is figured out as the novel mental space with the unique topology.

A part of experimental data in this field is completely consistent with the interpretation given. Thus, when participants read short stories like Marta switched on the radio and heard the winning lottery numbers. Since she won the lottery prize, the first thing she did was to buy a new Mercedes car and Marta switched on the radio and heard the winning lottery numbers. If she had won the lottery prize, the first thing she would have done was to buy a new Mercedes car and after reading were asked to verify a test probe belonged to the beginning of the story ("heard"), they verified it faster in the counterfactual than in the factual stories, which means, according to the authors, that in counterfactual story the situation model is not updated and the attention of the readers focuses on the initial information. This observation, in turn, argues for the view on counterfactual mental spaces as endowed with special qualities in comparison with factual ones (de Vega et al. 2007; de Vega 2008, 296-297).

The fact that correct comprehension of counterfactuals requires knowledge about both real and counterfactual worlds is also in line with the conjectures of conceptual blending theory (de Vega 2008, 298-299; Ferguson, Sanford 2008, 610; de Vega, Uritta 2011, 962-963).

However, the more precise analysis gives the strong evidence that Fauconnier and Turner's attitudes are too speculative to account for a real time cognitive process. The key question in this context is whether mental spaces are exclusively mental structures, which have no connection with human perception, or they are based on human sensorimotor experience. Although de jure the authors of conceptual blending theory stress the second opportunity, de facto they work with the first one.

In the meantime, the data of the experiments testifies the sensorimotor anchoring of meaning for counterfactual expressions, at least, for special groups of words. Thus, the results of de Vega, Uritta 2011 show that in the process of blend construction counterfactuals apply to the sensorimotor anchors similar to their factual counterparts. Similar point is formulated in Ferguson, Sanford 2008. The authors claim that the processing of a true utterance in the factual context and of a false utterance in the counterfactual one face similar obstacles at the first stage, and the principles of counterfactual space topology are comprehended by the reader only at the second stage of this process.

In other words, the schema of three phases in blend construction (composition, completion, and elaboration) doesn't work, at least, in two aspects. Firstly, the background conceptual structure doesn't complement the blend on the second step; it is kept in mind from the very beginning. Secondly, even after the blend constructed, its elements are perceived at the first phase as objects in the real space with all spectrum of sensorimotor reactions, and only afterwards they are replaced in the counterfactual space.

## Conclusion

I would like to sum up by saying that conceptual blending theory contains a lot of fascinating observations and provocative ideas extending the horizon of our knowledge. At the same time the authors seem to be prone to unreasonable generalizations, and they are not fairly correct in revealing the cognitive meaning of the operation they discovered. The view on blending as an important tool to adapt knowledge to the experience of average people seems to fit the gist of this procedure better than the intention to look at blending as a basic instrument for the creation of new knowledge.

The second weak point of conceptual blending theory is the lack of cultural-historical analysis as well as the absence of experimental data justifying it. It may therefore be interesting in this context to address the demarcation between formal as-if theories and heuristic theories, fitting the process in on-line regime (Gigerenzer, Todd 1999; Hertwig, Hoffrage 2012). Fauconnier and Turner present their theory as heuristic, but, to considerable extent, it looks like as-if theory. The elimination of this contradiction could provide a new impulse for the theory development and strengthen its heuristic potential.

## References

Alexander, J. (2011). Blending in mathematics. Semiotica, 187, 1-48.
Brandt, P.A. (2005). Mental spaces and cognitive semantics: A critical comment. Journal of Pragmatics. 37, 1578-1594. de Vega, M. (2008). Levels of embodied meaning: From pointing to counterfactuals. In de Vega M., Glenberg A., Graesser A. (eds.). Symbols and embodiment: debates on meaning and cognition. Oxford; New York: Oxford University Press, 285-308.
de Vega, M., Uritta M. \& Riffio B. (2007). Canceling updating in the comprehension of counterfactuals embedded in narratives. Memory and Cognition, 35, 1410-1421.
de Vega M. \& Uritta M. (2011). Counterfactual sentences activate embodied meaning: An action-sentence compatibility effect study. Journal of Cognitive Psychology, 23, 8, 962-973.
Fauconnier, G. (2009). Generalized integration networks. In Evans, V., Pourcel S. (eds). New Directions in Cognitive Linguistics. Amsterdam: John Benjamins, 147-160.
Fauconnier, G. \& Turner, M. (1994). Conceptual Projection and Middle Spaces. San Diego: University of California, Department of Cognitive Science Technical Report 9401 (http://papers.ssrn.com/sol3/papers.cfm?abstract_id=129086 2).

Fauconnier, G. \& Turner, M. (1996). Blending as a Central Process of Grammar. In Goldberg A. (ed.). Conceptual Structure, Discourse, and Language. Stanford: CSLI Publications, 113-129.
Fauconnier, G. \& Turner, M. (1998). Conceptual integration networks. Cognitive science. 22, 2, 133-187.
Fauconnier, G. \& Turner, M. (2000). Compression and global insight. Cognitive linguistics. 11, 3/4, 283-304.
Fauconnier, G. \& Turner, M. (2002). The Way We Think. Conceptual Blending and the Mind's Hidden Complexities. N. Y.: Basic Books.

Fauconnier, G. \& Turner, M. (2008). The Origin of Language as a Product of the Evolution of Modern Cognition. In Laks B. (ed.). Origin and evolution of languages: approaches, models, paradigms. L.: Equinox, 133-156.
Ferguson, H. \& Sanford A. (2008). Anomalies in real and counterfactual worlds: An eye-movement investigation. Journal of Memory and Language. 58, 609-626.
Gibbs, R. W., Jr. (2000). Making good psychology out of blending theory. Cognitive linguistics. 11, 3/4, 347-358.
Gigerenzer, G. \& Todd, P. (1999). Fast and Frugal Heuristics: The Adaptive Toolbox. In Gigerenzer G., Todd P. (eds.). Simple heuristics that make us smart. New York: Oxford University Press, 3-36.
Harder, P. (2003). Mental spaces: exactly when do we need them? Cognitive Linguistics. 14, 1, 91-96.
Hertwig, R. \& Hoffrage, U. (2012). Simple Heuristics: the Foundations of Adaptive Social Behavior. In Hertwig R., Hoffrage U.(eds.). Social rationality: simple heuristics in a social context. New York: Oxford University Press, 3-38.

Lévy-Bruhl, L. (1978). Primitive mentality. New York: AMS Press.
Lévy-Bruhl, L. (1979). How natives think. New York: Arno Press.
Mandelblit, N. (2000). The grammatical marking of conceptual integration: From syntax to morphology. Cognitive linguistics. 11, 3/4, 197-251.
Oakley, T. \& Hougaard, A. (2008). Mental spaces and discourse analysis. In Oakley T., Hougaard A. (eds.). Mental spaces in discourse and interaction. Amsterdam; Philadelphia: J. Benjamins Publishing, 1-26.
Sweetser, E. (2000). Blended spaces and performativity. Cognitive linguistics. 11, 3-4, 305-333.
Vygotsky, L. (1986). Thought and language. Cambridge, Mass.: MIT Press.

# Primacy/recency effects in infant categorisation 

Valentina Gliozzi (gliozzi@di.unito.it) ${ }^{1}$<br>Dipartimento di Informatica, Università di Torino<br>Corso Svizzera 185, 10149 Torino, Italy<br>Nadja Althaus (nadja.althaus@psy.ox.ac.uk ) ${ }^{1}$<br>Department of Experimental Psychology, University of Oxford<br>South Parks Road, Oxford OX1 3UD, UK<br>Julien Mayor (julien.mayor@unige.ch)<br>FPSE, University of Geneva<br>40, Boulevard du Pont-d'Arve 1211 Genève 4, Switzerland<br>Kim Plunkett (kim.plunkett@psy.ox.ac.uk )<br>Department of Experimental Psychology, University of Oxford<br>South Parks Road, Oxford OX1 3UD, UK


#### Abstract

We provide evidence that primacy and / or recency effects play a crucial role in infant visual categorization. First, we demonstrate that a connectionist model of infant categorization based on a self-organizing map (Gliozzi, Mayor, Hu, \& Plunkett, 2009) predicts an increased influence of the first and the last stimuli during familiarization on the category boundaries. We then present data from 10-month-old infants which confirm these effects. Future research will allow to discriminate between a primacy or a recency effect.


Keywords:infant categorization, self-organizing maps, connectionist modelling

## Introduction

familiarization/novelty preference paradigms have been widely used in experiments on infant categorization. In these types of experiment, infants are first familiarized with a sequence of stimuli. After the familiarization phase, infants are tested by simultaneously showing them two test stimuli: a within-category test stimulus and an out-of-category test stimulus. After the test phase is completed, category formation is assessed by comparing looking time at the withincategory test stimulus and looking time at the out-of-category test stimulus. Novelty preference - longer looking time at the out-of-category test stimulus than at the within-category stimulus - is taken as an indication for categorization: if looking time is indexed as a measure of surprise, this indicates that the out-of-category test stimulus is less familiar than the withincategory one, and therefore that infants have formed one category over the familiarization stimuli.

The assumption underlying the novelty preference test is that infants form a category representation close to the central tendency of the stimuli. In other words, this representation is equidistant from all the stimuli and represents them equally well, in a process that is unaffected by the order of the stimuli presentation. In this paper, we question this assumption, and argue that the process of category formation is

[^281]more disordered than this, and depends on many familiarization contingencies. In particular, we argue that a primacy or recency effect will affect the category formation process: the number and type of categories formed is modulated by the identity of the first, or last, stimuli presented. Future research will aim at distinguishing the relative roles of primacy and recency effects.
We will first show how the hypothesis of a primacy/recency effect was derived from the analysis of the behavior of a computational model, closely related to the model presented by Gliozzi et al. (2009). The model's predictions have been subsequently tested and validated by testing 10 -month-old infants in Oxford. This manuscript results from of a strong interplay between computational simulations and experimental results.

## Literature and Previous Results

Although it is clear that infants can form categories from visual familiarization stimuli (Younger, 1985; Eimas \& Quinn, 1994; Mareschal \& Quinn, 2001), the way in which familiarization contingencies impact category formation remained elusive until recently (Kovack-Lesh \& Oakes, 2007; P.C.Quinn \& R.S.Bhatt, 2010; Bomba \& Siqueland, 1983; Mather \& Plunkett, 2011) and the nature of the categories formed is yet to be understood.

In a previous experiment, Mather and Plunkett (2011) showed that the order of presentation of the familiarization stimuli can affect categorization. In particular, Mather and Plunkett (2011) compared infant categorization under two familiarization conditions that differ in the order by which the same set of stimuli (those used by Younger (1985)) is presented to infants during familiarization. Examples of familiarization stimuli, as well as of within-category (average), and out-of-category (peripheral) test stimuli can be found in Figure 1. In the high distance condition, infants were familiarized with sequences that maximize the Euclidean distance in feature space between successive stimuli whereas in the low


Figure 1: Example of familiarization sequences in the high distance condition with mild start/end stimuli and of the test stimuli
distance condition the Euclidean distance between successive stimuli is minimized. Mather and Plunkett (2011) found that only infants in the high distance condition successfully exhibited novelty preference at test, indicating that they had formed a category over the familiarization stimuli. Despite seeing the same items, with the only difference being the order of successive stimuli, infants in the low distance condition failed to discriminate between the test stimuli. The authors gave some potential explanations for this finding, ranging from faster habituation in the low distance condition, to the fact that infants in the high distance condition explore a bigger feature space than infants in the low distance condition, hence achieving more robust representations, until the fact that it may be more difficult to discriminate between small changes in feature space in successive stimuli in the low condition than when incremental changes in feature space are larger, as it is the case in the high distance condition. In this paper we provide a further explanation, while trying to gain further insight into the mechanisms underlying category formation with different familiarization contingencies.

Mather and Plunkett (2011)'s results are the starting point of this work. We first reproduce Mather and Plunkett (2011)'s results with a slightly-modified version of the model introduced by Gliozzi et al. (2009). As we will see, the updated model not only captures Mather and Plunkett (2011)'s results but also suggests an interpretation of its behavior which is different from the set of potential explanations provided by Mather and Plunkett (2011). Similarly to Mather and Plunkett (2011), we argue that categorization is affected by the order of presentation of the stimuli. However, in contrast from Mather and Plunkett (2011), we suggest that the largest effect impacting categorization is the identity of the first or last stimulus of the sequence, rather the average Euclidean distance in feature space between successive stimuli. In other words, we argue for a primacy/recency effect. As we will see, the experiments with infants confirm this hypothesis.

## Computational Model

## The model

The model we consider here is an adaptation of the model presented by Gliozzi et al. (2009). The model is a self-organizing map (Kohonen, 1997), which is recognized as a psychologically plausible neural network model (Kohonen, 1993), implementing a biologically plausible approach to human information processing: although our implementation is at a
high level of abstraction, we can be confident that the map architecture and learning algorithms used in the paper can be implemented at a physiological level of information processing. Psychological plausibility is added to our model by the fact that the map can be trained by following the same schedule of infants: by presenting each familiarization stimulus only once (instead of hundreds of times as in standard networks). The model receives visual inputs which are vectors with four dimensions (e.g. $[1,1,5,5]$ ) that represent the stimuli by Younger (1985) used by Mather and Plunkett (2011) (see Figure 1). Each value in the vectors corresponds to one feature in the cartoons presented to infants: length of the neck, length of the legs, the ears' orientation and the size of the tail. The encoding of the stimuli is the same used by Gliozzi et al. (2009), following Mareschal and French (2000). The stimuli can be either "mild", containing feature values close to the overall average (items with feature values 2 and 4 in Figure 1, with mild length legs and neck, etc), or "extreme", containing features further away from the overall average (combinations of values 1 and 5 in Figure 1, with very long or very short legs, very long or very short neck, etc)).

The model, like any self-organizing map, consists of a set of units, spatially organized in regular grids. Each map unit $u$ is associated with a weight vector $W_{u}$ of the same dimension as the input vectors. All weight vectors taken together can be seen as the map's representation of the world. The weight vectors are initialized to small random values. During training, the input vectors are presented to the network. After each presentation of a vector, its best matching unit is identified. This is the unit whose weight vector is closest to the input vector itself (in Euclidean distance). Next, the best matching unit's weights are adjusted to decrease the difference between the associated weight vector and the current input vector, according to the equation

$$
W_{u}(t+1)=W_{u}(t)+a(t)\left(I(t)-W_{u}(t)\right)
$$

where $W_{u}(t+1)$ and $W_{u}(t)$ are the weight vectors associated to unit $u$ at time $t+1$ and $t$ respectively. $I(t)$ is the input vector presented to the network at time $t$. For the best matching unit $u$ and for input $I(t)$, the difference $I(t)-W_{u}(t)$ is called the quantization error (qerr) of the network with respect to $I(t)$. Adjusting the weights can be seen as corresponding to an adaptation of the map's internal representation to accommodate for the new incoming familiarization stimulus. With respect to standard self-organizing maps our model is simpli-
fied and does not have any neighborhood function, due to its limited size. Results extend to a larger version of the model in which there is a (non-shrinking) neighborhood function.

Finally, $a(t)$ is the learning rate at time $t$, defined as $\max \left(0, \min \left(1, \beta * \exp ^{\alpha * \sqrt{q} e r r}\right)\right)($ with $\alpha=4.5, \beta=0.05)$. Results are robust (hold in more than $50 \%$ of the cases) when $\alpha$ ranges from 1 to 10 , and $\beta$ ranges from 0.04 to 0.4 . We have also studied a decreasing variant of the learning rate $a(t)^{\prime}=\max \left(0, \min \left(1, \beta * \exp ^{\alpha * \sqrt{q} e r r}\right)\right) / t$ that allows to replicate results under some parameters' combinations. In the following we restrict our attention to the non-decreasing learning rate. The model's adaptive learning rate has two important properties. The first of these is that it is usually higher than in standard self-organizing maps. This allows the network to be trained in analogy to an infant familiarization procedure: rather than training the network over hundreds of epochs, effectively presenting each stimulus many times, here each stimulus is only presented to the map once. The second property of the learning rate is that it depends on the quantization error: roughly speaking, the more novel the incoming stimulus is, the higher the learning rate will be. The consequence of this is that the learning rate can be considered as a computational counterpart of attention in infants: the adaptive learning rate corresponds to the general finding that infants pay more attention to novel stimuli rather than to familiar ones.

## The model's predictions

In order to replicate Mather and Plunkett (2011)'s results, we have trained our model in the same way in which infants have been familiarized in Mather \& Plunkett's (2011) study: we produced 24 maps per condition (low distance condition versus high distance condition), and each of these was trained with the encoding corresponding to the sequence presented to an infant, with the same schedule used in infant familiarization, i.e. presenting each stimulus exactly once.

After each network was trained, we assessed whether a category had been formed by measuring the network looking time, defined as the quantization error (as in Mareschal and French (2000) and Westermann and Mareschal (2004)). In analogy to the infant experiments, network categorization was assessed during a test phase in which network looking time at the overall average test stimulus was compared to looking time at the peripheral test stimulus: a proportion lower than chance indicates that the stimuli presented during familiarization have been organized in a cluster whose centroid is closer to the overall average test stimulus than to the peripheral one.

For each condition, the average of the ratios for all networks was calculated, and compared to the corresponding ratio calculated by Mather and Plunkett (2011). The model reproduces Mather and Plunkett (2011)'s results with infants: networks familiarized in the high distance condition exhibit a stronger novelty preference for the peripheral test stimulus than those familiarized in the low distance condition.

Although the model successfully reproduces Mather and

Plunkett (2011)'s results, the organization of its internal representation during training suggests an explanation of the results which is different from that provided by Mather and Plunkett (2011). Indeed, the model predicts that the nature of the start and end stimuli impacts categorization more than the Euclidean distance, as suggested by Mather and Plunkett (2011). In particular, novelty preference on test is stronger for maps familiarized with sequences starting and ending with mild values than for those familiarized with sequences starting and ending with extreme values.

In order to understand how Euclidean distance, on the one side, and the nature of start-end stimuli, on the other side, influence the model's behavior, we have conducted simulations in a $2 * 2$ design considering four different conditions. The conditions are obtained by varying the average Euclidean distance between successive stimuli as well as the nature of the start and end stimuli (whether mild or extreme). We thus consider the four possible combinations: low distance \& mild start/end stimuli (Low/Mild); low distance \& extreme start/end stimuli (Low/Extreme); high distance \& mild start/end stimuli (High/Mild); high distance \& extreme start/end stimuli (High/Extreme). In all conditions start and end stimuli are either both mild or both extreme

The model predicts a main effect of start/end stimuli on categorization. For some choices of the learning rate's parameters ( $\alpha$ and $\beta$ ) one obtains an interaction between start/end stimuli and Euclidean distance.

In the following we give an intuitive idea of the model's mechanisms that lead to the prediction. Roughly speaking, the prediction derives from the way in which successive stimuli are organized throughout the training phase: an internal representation (or several internal representations) corresponding to the stimuli experienced is formed and updated run-time, after each stimulus presentation (in line with several other models, as Gliozzi et al. (2009); Gureckis and Love (2004); Westermann and Mareschal (2004)). Depending on the strength of the update of this internal representation after each stimulus presentation (i.e. depending on the value of the learning rate), at the end of the familiarization phase the internal representation is close to the first or last stimulus experienced during familiarization. For our sequences, where start and end type were bound, sequences starting and ending with mild stimuli lead to internal representations of the familiarization stimuli containing mild attributes' values, whereas sequences starting and ending with extreme stimuli lead to internal representations containing extreme attributes' values. For this reason, maps familiarized in the mild condition will find the average test stimulus (that also contains mild values) much more familiar than the peripheral test stimulus, whereas for maps familiarized in the extreme condition the difference will be much less dramatic.

Do infants tested with the same $2 * 2$ design exhibit the same behavior? Can we say that they process the familiarization stimuli in a way similar to the model?

We will see in the next section that infant data reflect the
model's predictions. The question naturally arises on how precisely the model's behavior and infant behavior parallel each other. We address this question by considering looking time throughout familiarization/training. As we will see in the next section, infant looking time decreases throughout the familiarization phase in the low-distance condition while remaining stable in the high distance condition. However, the original model does not exhibit this kind of behavior. In order to achieve this behavior in the model we have to add two elements to the learning mechanism : (i) a form of weight decay: the weights associated to the maps' units that are not involved in training (because they are not selected as the best matching unit) slowly decay towards the initial values, and (ii) a form of habituation: the learning rate decreases if the same unit is the best match over multiple trials. With these two new elements, the network looking time mimics infant looking time also in the familiarization phase.

## Experiments

## Methods

Participants In total, 104 infants (mean age: 310 days; 52 females) took part in this study. An additional 31 infants were excluded due to technical reasons $(\mathrm{N}=12)$ or a failure to reach the looking time criterion ( $\mathrm{N}=19$; criterion: a minimum of 6 trials with looking time data including trials 1 and 8 , total looking time greater than two standard deviations below the mean). Infants were recruited at the maternity ward of the local hospital.
Procedure Infants were seated on the caregiver's lap in front of a large television screen ( $110 \mathrm{~cm} \times 95 \mathrm{~cm}$ ) at a distance of approximately 90 cm . They were presented with eight familiarization trials, followed by four test trials (see Figure 1); all trials were 10 seconds in duration. During the eight familiarization trials, a single familiarization image (subtending ca. 14 degrees visual angle) was displayed either on the left or right hand side of the screen. During the test trials, two images were shown side by side. The first two test trials paired one of the peripheral stimuli with the overall average, with a location switch between the trials, and counterbalancing the position of the average stimulus on Test trial 1 across subjects. Test trials 3 and 4 involved one pairing of the novel stimulus with the average stimulus, and one pairing of the novel stimulus with the peripheral stimulus shown during tests 1 and 2 (order of trials and location of stimuli were counterbalanced). The infant's face was filmed by two cameras mounted above the screen to the left and right. Throughout the procedure, the experimenter monitored infants' gaze from a control room next to the testing booth. Trials were initiated manually by the experimenter after confirming that the infants gaze was directed at the screen, or re-directing the infant's gaze at the screen through verbal communication via microphone (e.g. "Look (baby’s name)! What s next?").


Figure 2: Looking time during familiarization.


Figure 3: Looking time during categorization test trials

## Results

The video streams from left and right cameras were manually scored for infants' gaze direction (left vs. right).

Looking time during familiarization A mixed ANOVA on the looking times for familiarization trials (see Figure 2) with within-subjects factor Block (Block 1: trials 1-4, Block 2: trials 5-8) and between-subjects factors Distance (low, high) and Start/End Stimulus (mild, extreme) revealed a main effect of Block $(\mathrm{F}(1,98)=8.253$, $\mathrm{p}=.005)$ and a Block x Distance interaction $(\mathrm{F}(1,98)=4.072, \mathrm{p}=.046)$. T-Tests confirmed that looking time decreased between Block 1 and 2 in the low-distance conditions, but remained the same in the highdistance conditions.

Categorization: Test trials 1 and 2 In order to assess categorization performance, looking preference scores were obtained for each test trial from each participant by dividing the time spent looking at the average stimulus by the time spent looking at either test stimulus, average or peripheral (see Figure 3 for results). The resulting preference scores from the first test trial were subjected to an ANOVA with factors distance (low vs. high) and start and end stimulus (extreme vs. mild). This revealed a main effect of start/end stimulus $(F(1,92)=6.242, p=.014)$. All other effects remained non-significant (all $F$ s $<.31, p \mathrm{~s}>.57$ ). Follow-up t-tests
showed that infants in the mild start/end stimulus conditions exhibited a preference for the peripheral stimulus on Test trial 1 (Looking proportion for average stimulus 3333: $\mathrm{M}=41.7 \%, \mathrm{SE}=2.9 \%$; $\mathrm{t}(50)=2.882, \mathrm{p}=.006$ ), whereas infants in the extreme start/end stimulus condition exhibited no preference (Looking proportion for stimulus 3333: $\mathrm{M}=51.4 \%$, $\mathrm{SE}=2.4 \%$; $t(44)=.564, p=.576)$. On Test trial 2, the observed pattern of behavior was different. An ANOVA with factors distance and start/end stimulus revealed a significant interaction between distance and start/end stimulus $(\mathrm{F}(1,93)=5.534$, $\mathrm{p}=.021$ ). No other effects were significant (all $F \mathrm{~s}<.75$, $p \mathrm{~s}$ $>$.39). Further analysis of the interaction showed that only infants in the high/extreme condition had a significant preference, again for the peripheral stimulus $(t(23)=2.198, p=.038)$. Preferential looking in all other conditions did not differ from chance ( 0.5 ; all $t \mathrm{~s}<1.2, p \mathrm{~s}>.24$ ).

Novelty preference: Test trials 3 and 4 In order to establish that looking on the first test trials was driven by novelty preference rather than familiarity preference, preference scores were obtained for test trials 3 and 4 by dividing the amount of looking at the novel stimulus by the total looking time for each trial. The preference scores were subjected to an ANOVA with factors Test type (novel vs. average, novel vs. peripheral), Trial order, (novel vs. average first, novel vs. average second), Distance, and Start- and End-stimulus. This revealed a main effect of Trial order $(\mathrm{F}(1,84)=4.895$, $\mathrm{p}=.03$ ). All other effects were non-significant. Follow-up t -tests showed that there was always a significant novelty preference on the first of the two trials (Novel vs. Peripheral: $M=.66, S E=.04, t(50)=4.6, p<.001$; Novel vs. Average: $M=.57, S E=.03 ; t(46)=2.03, p=.048$ ), but on the second test trial infants only exhibited a (marginally) significant novelty preference if they had previously seen the pairing of the novel stimulus and a peripheral stimulus, and were now looking at the average and the novel stimulus ( $M=.57, S E=.04$; $t(46)=2.0, \mathrm{p}=.051)$. Infants who saw the novel stimulus paired with the overall average first did not exhibit a preference on the second novelty preference test trial ( $M=.55$, $S E=.03 ; t(46)=1.6, p=.107)$. These results are consistent with Mather \& Plunkett's (2011) findings.

## Discussion of Experimental Findings

The main effect of start and end stimulus found for Test trial 1 suggests that a recency or primacy effect determines looking on Test trial 1. This is consistent with the model predictions described above. As expected, infants who saw a mild stimulus on familiarization trials 1 and 8 exhibited a preference for the peripheral stimulus on Test trial 1. For these groups, the average stimulus appears particularly familiar when they get to Test trial 1. Infants in the groups with extreme start and end stimuli on the other hand do not exhibit any preference on Test trial 1. This is, empirically, the more surprising result: Younger (1985) reported merely an overall preference for the peripheral stimulus on the equivalent test trial. A conservative interpretation of our data would assume that no category was
formed in the extreme conditions. However, the model's performance indicates that instead of no category being formed the category's central tendency is merely closer to the peripheral stimulus than in the mild condition. This suggests that the null preference we observe is merely due to the fact that with this shifted category representation the average and peripheral test stimuli are equally interesting to the infants.

Test trial 2 is harder to interpret, as the pattern of preferences is very different from Test trial 1. Such order effects are common in familiarization / novelty preference paradigms (for a discussion see Schöner and Thelen (2006)). A likely cause for this is that learning does not stop at the end of familiarization: infants may incorporate both test stimuli presented on Test trial 1 in their category, and this will influence looking preferences on Test trial 2. Further work is required in order to explain the exact patterns observed, but the fact that all four conditions differ on this test trial indicate that Euclidean distance has a secondary impact, i.e. Mather and Plunkett (2011) assumption still holds. Looking times during familiarization imply that Euclidean distance is an important factor for maintaining infants interest during learning. Infants in the high distance conditions maintained looking, whereas infants looking times in the low distance conditions decreased, indicating that they began to habituate. This behavior is consistent with Mather and Plunkett (2011) interpretation of the impact of Euclidean distance on infants' attention.

## General Discussion

Decades of research on early categorization have assumed that categorization patterns were not impacted by the order of presentation of the familiarization stimuli. familiarization sequences were randomised and results averaged over different realisations. Recently, Mather and Plunkett (2011) challenged this view and showed that the order of presentation of the familiarization stimuli had an impact on infant category formation. Reasons for this behavior are yet unclear, which is why we decided to implement a model so as to evaluate the role of the order of presentation of the stimuli on the pattern of categorization.

First, we created a variant of the neural network model introduced by Gliozzi et al. (2009). The model is built with a simple self-organizing map and successfully reproduces Mather and Plunkett (2011)'s results. However, the model proposes an explanation of these results which is different from that provided by Mather and Plunkett (2011). In particular, the model predicts a primacy/recency effect: category formation depends on the nature of the first or last stimuli used in the training sequence.

The model's predictions have been confirmed by data from infants. 104 10-month-old infants were familiarized with sequences in the same four conditions presented to the network. Novelty preference scores on test indicate that responses are mainly driven by primacy/recency effects, whereas the average Euclidean distance influenced looking time during familiarization. This implies that, at odds to common assump-
tions about familiarization, 10-month-old novelty preference responses can be heavily influenced by familiarization stimuli at the start or end of the familiarization sequence, a factor which is often ignored in infant familiarization studies.

Our results are consistent with both primacy and recency effects, and future research will determine whether category formation is more heavily influenced by either primacy or recency.

In conclusion, this paper questions the traditional view underlying the novelty preference procedure suggesting that familiarization stimuli are categorized in an abstract representation of all the stimuli. In this traditional view, the representation formed is independent from familiarization contingencies. On the contrary, our results show that infants are sensitive to the order of presentation of the stimuli and support models that advocate infant category learning as an incremental process by which, on a moment-by-moment basis, infant refine the boundaries of new categories (Gliozzi et al., 2009; Gureckis \& Love, 2004; Westermann \& Mareschal, 2004). In contrast, our results cannot be explained by models in which the infants only establish the category boundaries once they have sampled all familiarization items.

## Acknowledgments

This research was supported by Wellcome Grant 084386/Z/07/Z to Kim Plunkett and the Winkler Career Development Fellowship at St Hugh's College, University of Oxford, to Nadja Althaus. We wish to thank all parents and infants who participated in this study.

## References

Bomba, P. C., \& Siqueland, E. R. (1983). The nature and structure of infant form categories. Journal of Experimental Child Psychology, 35, 294-328.
Eimas, P., \& Quinn, P. (1994). Studies on the formation of perceptually based basic-level categories in young infants. Child Development, 65, 903-917.
Gliozzi, V., Mayor, J., Hu, J.-F., \& Plunkett, K. (2009). Labels as Features (Not Names) for Infant Categorization: A Neurocomputational Approach. Cognitive Science, 33, 709-738.
Gureckis, T., \& Love, B. (2004). Common mechanisms in infant and adult category learning. Infancy, 5(2), 173198.

Kohonen, T. (1993). Physiological interpretation of the self-organizing map algorithm. Neural Networks, 6(7), 895-905.
Kohonen, T. (1997). Self-Organizing Maps, volume 30 of Springer Series in Information Sciences. Springer, Berlin.
Kovack-Lesh, K. A., \& Oakes, L. M. (2007). Hold your horses: How exposure to different items influences infant categorization. Journal of Experimental Child Psychology, 98, 69-93.
Mareschal, D., \& French, R. (2000). Mechanisms of categorization in infancy. Infancy, 1, 59-76.

Mareschal, D., \& Quinn, P. C. (2001, Sep). Categorization in infancy. Trends in Cognitive Sciences, 5, 443-450.
Mather, E., \& Plunkett, K. (2011). Same items, different order: Effects of temporal variability on infant categorization. Cognition, 119, 438-447.
P.C.Quinn, \& R.S.Bhatt. (2010). Learning perceptual organization in infancy:The effect of simultaneous versus sequential variability experience. Perception, 39, 795806.

Schöner, G., \& Thelen, E. (2006). Using dynamic field theory to rethink infant habituation. Psychological Review, 113, 273-298.
Westermann, G., \& Mareschal, D. (2004). From Parts to Wholes: Mechanisms of Development in Infant Visual Object Processing. Infancy, 5(2), 131-151.
Younger, B. (1985). The segregation of items into categories by ten-month-old infants. Child Development, 56, 1574-1583.

# Inventing Prepares Learning Motivationally, but a Worked-out Solution Enhances Learning Outcomes 

Inga Glogger (glogger@psychologie.uni-freiburg.de)<br>Department of Educational and Developmental Psychology, University of Freiburg, Engelbergerstraße 41 D-79085 Freiburg, Germany<br>Julian Kappich (kappich@psychologie.uni-freiburg.de)<br>Department of Educational and Developmental Psychology, University of Freiburg, Engelbergerstraße 41 D-79085 Freiburg, Germany

Rolf Schwonke (schwonke@psychologie.uni-freiburg.de)
Department of Educational and Developmental Psychology, University of Freiburg, Engelbergerstraße 41
D-79085 Freiburg, Germany
Lars Holzäpfel (lars.holzaepfel@ph-freiburg.de)
Institute for Mathematical Education, University of Education, Kunzenweg 21
D-79117 Freiburg, Germany
Matthias Nückles (nueckles@ezw.uni-freiburg.de)
Department for Empirical Instruction and School Research, University of Freiburg, Rempartstraße 11 D-79098 Freiburg, Germany

Alexander Renkl (renkl@psychologie.uni-freiburg.de)
Department of Educational and Developmental Psychology, University of Freiburg, Engelbergerstraße 41 D-79085 Freiburg, Germany


#### Abstract

Solving an open problem as proposed by inventing and productive failure approaches has been shown to prepare learners effectively for subsequent direct instruction even though invented solutions are often suboptimal for the given problems. Inventing can make the learners aware of knowledge gaps (cognitive) and more curious about and interested in the learning contents (motivational effects). However, working on the same problem with a given (optimal) solution helps avoid misconceptions and disorganized knowledge, while providing useful basic knowledge. Therefore, a given solution could be more effective. In an experiment $(N=42)$, we tested to what extent working on an open problem (inventing) versus a solution prepares student teachers for learning strategy evaluation. The inventing group invented criteria to evaluate learning strategies while the worked solution group studied the same problem in a solved, worked-out version. We found differential effects: inventing enhanced knowledge-gap experience, curiosity, and interest. However, studying the worked-out solution enhanced learning outcomes.


Keywords: instruction; invention activities; worked examples; teacher education; learning-strategy assessment.

## Introduction

In order to prepare learners for a new topic and to raise their attention as well as curiosity, teachers often address interesting problems in the beginning before they directly instruct learners about the topic. Similarly, there are experimentally tried-and-tested problem-oriented
approaches (Schmidt, De Volder, De Grave, Moust, \& Patel, 1989) such as inventing problem solutions (Schwartz, Chase, Oppezzo, \& Chin, 2011; Schwartz \& Martin, 2004) or productive failure at initial problems (Kapur, 2010). These approaches aim at preparing learners for subsequent direct instruction (preparation for future learning, Schwartz \& Martin, 2004). However, when preparing a lesson, should a teacher really put such initial problems up for "inventing" or "productive failure" before implementing direct instruction? Is it not more productive to immediately begin with tried-and-tested forms of direct instruction such as example-based learning in order to avoid wasting time when students search for problem solutions that are very hard to find (see Sweller, Kirschner, \& Clark, 2007)?

When starting immediately with methods of direct instruction, a problem might arise: learners often process directly presented information only superficially (Berthold \& Renkl, 2010), leading to little knowledge acquisition and transfer. Problem-oriented introductions such as invention activities can prepare learners to more deeply process directly presented information. For example, Schwartz and Martin (2004) had learners invent formulas describing four different distributions of pitches around a target. Later, the learners were taught the concept of mean deviation. Schwartz and Martin assumed that inventing creates preparedness for future learning by generating "early forms of knowledge" (p. 132). These early forms of knowledge can then be used to easily assimilate further knowledge.

Invention activities can appear problematic because learners might not generate canonical or even false solutions. According to the IKEA effect - the increased valuation of self-made products (Norton, Mochon \& Ariely, 2012) -, these own suboptimal solutions can be valued higher than expert ones. Similarly, research on the continued influence effect (Johnson \& Seifert, 1994) suggests that learners tend to stay with their own suboptimal solution instead of taking up the directly instructed canonical one. However, research on productive failure (e.g., Kapur, 2010, 2012) shows that initial problem-solving activities can be effective even though invented solutions to problems are often suboptimal or even false (see Schmidt et al., 1989, for similar findings). In addition, larger numbers of suboptimal solutions were followed by higher learning outcomes (Kapur, 2012). Difficulties as well as the production of suboptimal solutions can be seen as productive because they cause impasses making the learners realize that certain solutions do not work for all cases. Furthermore, research on impasse-driven learning has shown that instructional explanations are more effective when given in the context of such an impasse (Sánchez, García-Rodicio, \& Acuña, 2009; VanLehn, Siler, Murray, Yamauchi, \& Baggett, 2003). If prior knowledge is not sufficient to solve the inventing task and an impasse is reached, a perceived "vacuum" can help to see more clearly the "information needs" and "knowledge gaps to be filled", which can lead to a better focus on the most relevant contents in a subsequent learning phase.

Besides the more cognitive effects of creating a form of prior knowledge and an experience of knowledge gaps, problem-oriented instruction can influence motivation. Enhancing motivation can foster deep processing, understanding, and transfer (Belenky \& Nokes-Malach, 2012; Entwistle \& Ramsden, 1983; Pintrich, 2000; Pugh \& Bergin, 2006). Schmidt et al. (1989) discussed an epistemic curiosity (i.e., motivation to strive for knowledge) which could explain higher learning outcomes in the problembased condition of their experiment. Interest can be enhanced because "people like to produce things" (Schwartz \& Martin, 2004, p. 171; diSessa, Hammer, Sherin, \& Kolpakowski, 1991; Norman \& Schmidt, 1992). Enhancing learner motivation is argued to be a major advantage of problem-oriented learning in general, but there is little research bearing directly on motivational issues (HmeloSilver, 2004). Most studies do not assess learners' perceived knowledge gaps, either.

Some researchers criticize the postulated effects of such forms of problem-oriented learning (Mayer, 2004). Sweller et al. (2007) as well as Kirschner, Sweller, and Clark (2006) assume that the problem-oriented activities and especially failure within these activities are unproductive. "Not only is unguided instruction normally less effective; there is also evidence that it may have negative results when learners acquire misconceptions or incomplete or disorganized knowledge" (Kirschner et al., 2006, p. 84). They criticize that many studies favoring problem-oriented learning did
not employ an adequate control group. Such a control group would have to engage in the same topic as the experimental group for the same timespan (see Sweller et al., 2007).

Against the background of these different positions on the value of inventing, we designed an experiment with a control group that is adequate and "strong" in the sense of implementing "good" direct instructional procedures, but using the exact same problem as in an inventing group. However, the problem was presented in a worked-out version. Boths group engaged in their problem for the same amount of time. Specifically, we tested the following hypotheses and asked the following research questions:
(1) The inventing activity leads to more experience of knowledge-gaps, more epistemic curiosity, and more interest than working through the worked-out solution. (2) The more knowledge gaps participants perceive, the higher their focus on the most relevant contents in a subsequent learning phase. (3) Does the inventing activity lead to superior learning outcomes when compared to a "strong" control group? (4) Is failure in the inventing activity productive (Kapur, 2010, 2012), that is, (4a) how is the appropriateness of solutions to the invention problem related to learning outcomes? and (4b) how is the number of (different) suboptimal invented solutions related to learning outcomes?

## Method

## Participants and Design

As participants, forty-two German student teachers (sex: 12 female, 30 male; $M_{\text {age }}=22.74, S D=3.44$ ) were randomly assigned to two conditions: "inventing" $(n=21 ; 13$ female, $\left.M_{\text {age }}=22.05 ; S D=2.92\right)$ and "worked solution" $(n=21 ; 17$ female; $M_{\text {age }}=23.43, S D=3.83$ ). As learning domain we used the assessment of learning strategies in learning journals written by high school students. By writing learning journals, for example, as homework after biology or mathematics classes, high school students are encouraged to apply learning strategies. For example, they can develop their own thoughts based on the new learning contents (elaboration strategy). Ideally, they can do this in a detailed way and on their own, for example, "I realized that nothing can grow without mitosis, not even myself! Because (...)." Such an elaboration can be evaluated as high in quality (see Glogger, Schwonke, Holzäpfel, Nückles, \& Renkl, 2012).

The inventing group invented criteria to evaluate the quality of learning strategies applied in learning journals. First, all participants received a short introduction (134 words) about learning journals and the quality of learning strategies in general. The instruction to the subsequent activity (for both groups) read as follows: "On pages B and C, you will find four extracts from learning journals. Each extract shows a variation of the same elaboration strategy (developing own thoughts) in a way a student in a biology class (dealing with the topic mitosis) could have realized it." The extracts looked similar in length, but differed systematically in two quality criteria "detailed elaboration
vs. wordy, but shallow elaboration" and "self-made vs. copied from the lesson". Participants in the inventing group were prompted to contrast the four extracts, rate the quality of each one on a 3-point scale (low, medium, or high), make notes on discerning aspects, and generalize from these aspects to generic evaluation criteria for the learning strategy elaboration. The student teachers had to write down their criteria in a box labeled "my criteria." They were also instructed to check whether or not the final criteria really work to discern all extracts (cf. Roll, Holmes, Day, \& Bonn, 2012). In contrast, participants in the solution condition neither had to rate the extracts nor to invent the evaluation criteria. Instead, they were asked to carefully study the same problem that was worked out by a (fictitious) experienced teacher. That is, the canonical criteria were written in the "my criteria" box, the quality of the four extracts was rated, and short notes about discerning aspects were written down. In summary, the two groups had the exact same work sheets with four extracts. However, the inventing group had to generate a solution to the problem how to evaluate the quality of learning strategies whereas the worked-solution group was given the solution, namely the criteria. That is, the inventing group had to generate core learning principles by contrasting cases, whereas the solution group worked through the contrasted cases with the given principles. The two criteria (principles) were explained in the subsequent learning phase (inter alia), that is, the information on the criteria was redundant for the solution group. Both groups were given the same amount of time ( 15 minutes) for their preparation activity. Participants were compensated with 15 Euros for the average 85 minutes duration of the study.

## Materials

Pretest and Demographic Questionnaire. A web-based pretest assessed participants' topic-specific prior knowledge. Participants received up to five points for the four open questions ( $\alpha=.83$, e.g., "Which learning strategies can students apply by writing a learning journal?"). Two independent raters scored $25 \%$ of the pretests $(I C C=.87)$ and of all following data with open format including posttest. A demographic questionnaire assessed sex, age, number of semesters in teacher education, experience with learning journals, and computer skills.

Process Variables. Questionnaires assessed the participants' experience of knowledge gaps, epistemic curiosity, and interest by items with 6-point rating scales (6: absolutely true). Experience of knowledge gaps was assessed with nine items ( $\alpha=.89$; e.g., "My knowledge was insufficient to complete the task"). Epistemic curiosity was assessed with 10 items $(\alpha=.85)$, based on the "Melbourne Curiosity Inventory - State Form" (Naylor, 1981) and adapted to the present context (e.g., "I feel curious about how to evaluate learning strategies"). Topic-specific interest (Schiefele \& Krapp, 1996) was measured with six items ( $\alpha$ $=.72$; e.g., "Learning how to evaluate learning strategies is entertaining."). In addition, we rated the appropriateness
(quality) of the invented solutions (i.e., in the inventing group only) on a 6-point scale ranging from 1 (not at all appropriate) to 6 (absolutely appropriate, ICC $=.82$ ); and we counted solutions, operationalized by the number of (different) criteria invented by participants.

## Computer-Based Learning Environment and Learning

 Time. After the experimentally varied preparation activity and the questionnaires, participants worked individually in a computer-based learning environment (CBLE). The CBLE explained several sub-categories of elaboration strategies, how they improve comprehension, and how they can be identified in learning journals. A subsequent unit explained the quality criteria of elaboration strategies, using various (new) examples (i.e., not used in the preparation phase). Learners could navigate freely. The focus on the most relevant learning contents was operationalized as the time learners spent in the quality criteria unit. The duration spent in this unit and in the environment as a whole was logged by the software.Posttest. A posttest consisting of seven tasks measured learning outcomes as application of quality criteria on students’ learning strategies $(\alpha=.75)$. Each task consisted of a short extract from a learning journal, representing one sub-category of an elaboration strategy. This sub-strategy was labeled. All extracts were new so that the tasks required transfer (content transfer, Barnett \& Ceci, 2002). Participants were asked to rate the quality of the strategy (low, medium, or high) and to explain their rating by applying the previously learned criteria. Answers were rated on a 6-point scale ranging from 1 (no conceptual understanding) to 6 (very clear conceptual understanding; SOLO taxonomy by Biggs \& Collis, 1982; ICC = .93).

## Procedure

We required participants to work on the web-based pretest four days before the experiment in order to avoid knowledge activation effects. On the day of the experiment, participants first filled out the demographic questionnaire. Next, they worked individually on the task that prepared the following learning phase (inventing vs. worked-out solution) for 15 minutes. Subsequently, questionnaires assessed participants' experience of knowledge gaps, epistemic curiosity, and interest. The participants then worked on the CBLE without time limits ( 20 minutes on average) as we were interested in potential effects of the two conditions on the learning time spent in the environment. After the learning phase, interest was reassessed. Finally, participants worked on the posttest.

## Results

A significance level of .05 was used for all analyses. We used $d$ as an effect-size measure with values between .20 and .50 classified as small, values between .50 and .80 as medium, and values $>.80$ as large (Cohen, 1988). We did not find any significant differences between the groups in prior knowledge (inventing: $M=2.05$ [41 \% correct], $S D=$

Table 1: Means (standard deviations in parentheses) of process variables, and the posttest in the experimental groups, and test statistics.

|  | Knowledge- <br> Epistemic curiosity |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Inventing | $\begin{gathered} 3.80 \\ (0.78) \end{gathered}$ | $\begin{gathered} 4.54 \\ (0.63) \end{gathered}$ | $\begin{gathered} 5.05 \\ (0.58) \end{gathered}$ | $\begin{gathered} 8.02 \\ (2.06) \end{gathered}$ | $\begin{gathered} \hline 3.20 \\ (0.80) \end{gathered}$ |
| Worked solution | $\begin{gathered} 2.75 \\ (1.09) \end{gathered}$ | $\begin{gathered} 4.15 \\ (0.78) \end{gathered}$ | $\begin{gathered} 4.64 \\ (0.52) \end{gathered}$ | $\begin{aligned} & 10.52 \\ & (3.00) \end{aligned}$ | $\begin{gathered} 3.80 \\ (0.84) \end{gathered}$ |
| $t(40)$ | 3.61 | 1.80 | 2.37 | -3.10 | -2.39 |
| $p$ | $<.001{ }^{\text {d }}$ | . $039{ }^{\text {d }}$ | . $012{ }^{\text {d }}$ | . $004{ }^{\text {e }}$ | . $022{ }^{\text {e }}$ |
| $d$ | 1.14 | 0.57 | 0.75 | -0.98 | -0.75 |

Note. All 6-point scales: Knowledge-gap experience, epistemic curiosity, and interest: from 1 (not true at all) to 6 (absolutely true); Posttest: from 1 (no conceptual understanding) to 6 (very clear conceptual understanding). ${ }^{\text {a }}$ Knowledge-gap experience. ${ }^{\text {b }}$ Interest after the learning phase. ${ }^{\text {c }}$ Learning time (in minutes) in the most relevant unit of the CBLE on quality criteria. ${ }^{\text {d }}$ one-tailed. ${ }^{e}$ two-tailed.
1.44, solution: $M=2.37$ [47.4 $\%$ correct], $S D=1.29, t(40)=$ $.75, p=.459$ ) or in demographic variables (sex, age, number of semesters, experience with writing learning journals, and computer skills; all $p$ 's > .05).

Table 1 presents the means and standard deviations of knowledge-gap experience, epistemic curiosity, interest, learning time, and the posttest for the two experimental groups. Directly after the preparation task, the participants of the inventing condition stated higher knowledge-gap experience (large effect), and higher epistemic curiosity (medium effect), than participants of the solution condition, confirming hypothesis 1 . Even after the learning phase, they stated higher interest in learning about the assessment of learning strategies in learning journals (medium effect).

Table 2 shows correlation coefficients between all process variables and dependent variables. Regarding hypothesis 2 , perceived knowledge gaps did not significantly correlate with learning time (simple correlation, Table 2). However, when controlling for condition (and prior knowledge) we found a significant partial correlation, $r(36)_{\text {part }}=.34, p=$ .035, medium effect. Thus, the more knowledge gaps
participants perceived (independent from their condition and prior knowledge), the more time they spent in the most relevant learning unit about quality criteria indeed.

Surprisingly, even though the inventing group experienced more knowledge gaps (correlating with learning time), higher epistemic curiosity, and higher interest, this group did not achieve better learning outcomes (research question 3, see Table 1). Participants of the solution condition even outperformed the inventing condition (medium effect). Controlling for prior knowledge, the effect remained stable, $F(1,38)=4.95, p=.032, d=$ -0.72 . For exploratory purposes, we searched for variables explaining this effect. 'Learning time' is the only variable that correlated significantly with learning outcomes (Table 2 ), even if controlled for condition, $r(38)_{\text {part }}=.55, p<.001$, large effect. Against this background, we analyzed whether the effect of conditions on learning outcomes is mediated by learning time. We tested this mediation effect with a set of related multiple regression equations, following a products-of-coefficients strategy (MacKinnon, 2008). In this approach, there are essentially two assumptions to be met in order to speak of a mediated effect. First, an independent variable must significantly affect a mediating variable (path a). Second, the mediating variable must significantly affect a dependent variable (path b). The significance of the effect can be tested according to Sobel (1982).

The independent variable 'condition' did in fact significantly affect the mediating variable 'learning time' (path a): The worked solution group spent more time learning in the CBLE (quality criteria unit, $b=-1.09, S E=$ $0.41, b^{*}=-.39, p=.010$; controlled for prior knowledge). The learning time predicted learning outcomes significantly (path b: $b=.210, S E=0.0404, b^{*}=.625, p<.001$; controlled for prior knowledge). Learning time significantly mediated the effect of conditions on learning outcomes (Sobel test $=-2.37, p=.018$ [two-tailed], $L C L=-0.418404$, $H C L=-0.04032$ ).

Referring to research question 4 a , the appropriateness of participants' invented solutions $(M=2.8, S D=1.31)$ correlated substantially with learning outcomes, $r(16)_{\text {part }}=$ $.52, p=.028$, (large effect, controlled for prior knowledge. The more a participant failed to invent an appropriate solution to the inventing problem, the more this participant failed in the posttest as well.

Table 2: Intercorrelations of pretest, process variables, and posttest.

|  | Knowledgegap | Epistemic curiosity | Interest pre | Interest post | Learning time ${ }^{\text {a }}$ | Posttest |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Pretest | . 18 | .40* | . 19 | -. 01 | . $30^{+}$ | . 25 |
| Knowledge-gap experience | - | . 52 *** | .33* | . 24 | $.10{ }^{\text {b }}$ | -. 10 |
| Epistemic curiosity |  | - | .61*** | .48** | . 09 | -. 20 |
| Interest pre |  |  | - | . 53 *** | . 01 | -. 11 |
| Interest post |  |  |  | - | -. 13 | $-.29^{+}$ |
| Learning time ${ }^{\text {a }}$ |  |  |  |  | - | .66*** |

Note. $N=44 .{ }^{a}$ Learning time (in minutes) in the most relevant CBLE unit on quality criteria. ${ }^{b}$ Partial correlation differs and is given in the text with condition controlled. ${ }^{+} p<.10 .{ }^{*} p<.05 .{ }^{* *} p<.01 .{ }^{* * *} p<.001$.

Finally, hypothesis 4 b was rejected. The number of suboptimal invented solutions ( $M=4.1, S D=1.52$ ) did not correlate significantly with learning outcomes, $r(16)_{\text {part }}=$ $.28, p=.135$ (one-tailed, controlled for prior knowledge).

## Discussion

In line with the literature on problem-oriented learning (e.g., Hmelo-Silver, 2004), we found that an inventing activity had positive motivational effects (see also Belenky \& Nokes-Malach, 2012). Learners are more curious about and interested in the target learning domain. Learners also become aware of knowledge gaps to be filled. The more knowledge gaps learners perceived, the higher their focus on the most relevant learning contents. The motivational and knowledge-gap effects can be seen as a preparation for learning. However, they did not lead to higher learning outcomes in the inventing group. In contrast to the inventing and productive-failure literature, the worked solution group achieved better learning outcomes, mediated by learning time.

The results are in line with Sweller et al.'s (2007) and Kirschner et al.'s (2006) argument that positive results in problem-oriented learning studies could be an effect of "weak" control conditions or different time-on-task during the experimental variation. The worked-out solution condition resembles a worked-example condition. Possibly, a worked-example effect (Renkl, 2011; Sweller, 2006) had the worked solution group outperform the inventing group, even though self-explanations were not prompted, which is usually sensible in order to exploit the potential of examplebased learning (cf. Chi, Bassok, Lewis, Reimann, \& Glaser, 1989; Renkl, 2011). In the present case, the preparation activity of working through a worked-out solution of a problem obviously prepared learners to learn, possibly by enhancing basic knowledge about quality criteria. This wellorganized basic knowledge might have facilitated working intensively with the instructional explanations and explained strategy examples of different quality in the CBLE. That is, cognitive mechanisms such as deeper elaboration and spontaneous application of the learned concepts to presented examples during the learning phase could account for the results. Findings of a subsequent think-aloud study indicate such mechanisms. The cognitive mechanisms could have predominated motivational mechanisms in this study. Alternatively (or additionally), if the worked-solution functioned as a worked example (a model of a good solution), self-efficacy could have been enhanced. Selfefficacy can enhance effort and persistence (Schunk, 1990) which could explain the short learning time in the inventing group despite enhanced curiosity and interest.

One could interpret a speed-accuracy tradeoff: learning outcomes as well as learning time is higher in worked example. However, in the context of self-regulated learning in a CBLE, where diverse learning paths are provided and working through it can be more or less thorough, enhancing learning time can be advantageous. To put it differently, inventing might have constrained learning time, because
learners did not want to deal thoroughly with the learning contents (see also below). Also note that we did not find any differences between groups in efficacy (learning outcome per minute spent in the quality part, $p=.751$; in the CBLE, $p=.975$ ).
Another evidence for the claim that some "correct" basic knowledge of quality criteria facilitated future learning in the CBLE can be seen in the highly positive correlation between the level of appropriateness of the inventing solutions and the learning outcome. Failure was not productive in the present case. The number of suboptimal invented solutions was not significantly related to learning outcomes. These findings contradict Kapur's $(2010,2012)$ approach of productive failure. Additionally, findings about the continued influence effect (Johnson \& Seifert, 1994) and the IKEA effect (Norton et al., 2012) suggest that the inventing group could have clung to their initial suboptimal solution (quality criteria) even though the canonical criteria were explained and exemplified in the CBLE. Holding on to one's own solution ideas and partly neglecting canonical explanations could be another reason why the inventing group spent less time with learning in the CBLE. If they partly held on to their own solutions, the role of a transition phase, which is a usual part of a productive-failure procedure, could be of major importance: The teacher leads a discussion about students' own (suboptimal) solutions towards the canonical one.

Usually, productive failure and inventing include a collaborative learning setting. Students work on preparatory open problems either in groups (Kapur, 2012; Schmidt et al., 1989; Schwartz \& Martin, 2004) or in pairs (Schwartz et al., 2011; Westermann \& Rummel, 2012). The collaborative setting has not been explicitly discussed as an "active ingredient" of the preparatory activities. However, invention activities might only be effective in collaborative learning settings. This is a point to consider about the present study in which participants worked only individually.
Thus, further studies on problem-oriented learning settings such as inventing should investigate whether collaborative work during preparatory activities is a crucial ingredient of such instructional approaches. It can also be worthwhile to look at the learning processes that follow inventing and inventing-enhanced motivational states (e.g., applying basic knowledge from the preparatory activity, holding on to own solutions) in order to explain why inventing did not enhance learning outcomes when compared to working through a worked-out solution. More generally, it is important to use strong control groups in future research. In doing so, research should not simply investigate which instructional approach is better or worse. Instead, the results of the present study show that future research should achieve a differential analysis of cognitive and motivational effects of the various approaches.

## Acknowledgments

The project "Fostering diagnosis skills of teachers: Identifying and fostering learning strategies through
learning journals" was part of the research program "Bildungsforschung" (educational research) of the BadenWürttemberg Stiftung. We would like to thank our research assistants Corinna Fleischer, Andrea Ohst, Duygu Pektas.

## References

Belenky, D. M., \& Nokes-Malach, T. J. (2012). Motivation and transfer: The role of mastery-approach goals in preparation for future learning. Journal of the Learning Sciences, 21, 399-432.
Berthold, K., \& Renkl, A. (2010). How to foster active processing of explanations in instructional communication. Educational Psychology Review, 22, 2540.

Biggs, J. B., \& Collis, K. F. (1982). Evaluating the quality of learning: The SOLO Taxonomy. New York: Academic Press.
Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., \& Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. Cognitive Science, 13, 145-182.
diSessa, A. A., Hammer, D., Sherin, B. L., \& Kolpakowski, T. (1991). Inventing graphing: Meta-representational expertise in children. Journal of Mathematical Behavior, 10, 117-160.
Entwistle, N. J., \& Ramsden, P. (1983). Understanding student learning. London: Croom Helm.
Glogger, I., Schwonke, R., Holzäpfel, L., Nückles, M., \& Renkl, A. (2012). Learning strategies assessed by journal writing: Prediction of learning outcomes by quantity, quality, and combinations of learning strategies. Journal of Educational Psychology, 104, 452-468.
Hmelo-Silver, C. E. (2004) Problem-based learning: What and how do students learn? Educational Psychology Review, 16, 235-266.
Johnson, H. M., \& Seifert, C. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. Journal of Experimental Psychology: Learning, Memory, and Cognition, 20, 14201436.

Kapur, M. (2010). Productive failure in mathematical problem solving. Instructional Science, 38, 523-550.
Kapur, M. (2012). Productive failure in learning the concept of variance. Instructional Science, 40, 651-672.
Kirschner, P. A., Sweller, J., \& Clark, R. E. (2006). Why minimal guidance during instruction does not work. Educational Psychologist, 41, 75-86.
MacKinnon, D. P. (2008). Introduction into statistical mediation analysis. New York: Erlbaum.
Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? The case for guided methods of instruction. The American Psychologist, 59, 14-19.
Norman, G. R., \& Schmidt, H. G. (1992). The psychological basis of problem-based learning: A review of the evidence. Academic Medicine, 67, 557-565.

Norton, M. I., Mochon, D., \& Ariely, D. (2012). The IKEA effect: When labor leads to love. Journal of Consumer Psychology, 22, 453-460.
Pintrich, N. E. (2000). The role of goal orientation in selfregulated learning. In M. Boekaerts, P. Pintrich \& M. Zeidner (Eds.), Handbook of self-regulation. Orlando, FL: Academic Press.
Pugh, K. J., \& Bergin, D. (2006). Motivational influences on transfer. Educational Psychologist, 41, 147-160.
Renkl, A. (2011). Instruction based on examples. In R. E. Mayer \& P. A. Alexander (Eds.), Handbook of research on learning and instruction. New York, NY: Routledge.
Roll, I., Holmes, N. G., Day, J., \& Bonn, D. (2012). Evaluating metacognitive scaffolding in guided invention activities. Instructional Science, 40, 691-710.
Sánchez, E., García-Rodicio, H., \& Acuña, S. R. (2009). Are instructional explanations more effective in the context of an impasse? Instructional Science, 37, 537563.

Schiefele, U., \& Krapp, A. (1996). Topic interest and free recall of expository text. Learning and Individual Differences, 8, 141-160.
Schmidt, H. G., De Volder, M. L., De Grave, W. S., Moust, J. H. C., \& Patel, V. L. (1989). Exploratory models in the processing of science texts: The role of prior knowledge activation through small-group discussion. Journal of Educational Psychology, 81, 610-619
Schunk, D. H. (1990). Goal setting and self-efficacy during self-regulated learning. Educational Psychologist, 25, 7186.

Schwartz, D. L., Chase, C. C., Oppezzo, M. A., \& Chin, D. B. (2011). Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer. Journal of Educational Psychology, 103. 759-775.
Schwartz, D. L., \& Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. Cognition and Instruction, 22, 129-184.
Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. Sociological Methodology, 13, 290-312.
Sweller, J. (2006). The worked example effect and human cognition. Learning and Instruction, 16, 165-169.
Sweller, J., Kirschner, P., \& Clark, R. (2007). Why minimally guided teaching techniques do not work: A reply to commentaries. Educational Psychologist, 42, 115-121.
Van Lehn, K., Siler, S., Murray, C., Yamauchi, T., \& Baggett, W. B. (2003). Why do only some events cause learning during human tutoring? Cognition and Instruction, 21, 209-249.
Westermann, K., \& Rummel, N. (2012). Delaying instruction: Evidence from a study in a university relearning setting. Instructional Science, 40, 673-689.

# Development of Semantic Knowledge and Its Role in the Development of Category-Based Reasoning 

Karrie E. Godwin (kegodwin@andrew.cmu.edu)<br>Carnegie Mellon University, Department of Psychology, 5000 Forbes Avenue, Pittsburgh, PA 15213 USA<br>Bryan J. Matlen (bmatlen@cmu.edu)<br>Carnegie Mellon University, Department of Psychology, 5000 Forbes Avenue, Pittsburgh, PA 15213 USA

Anna V. Fisher, (fisher49@andrew.cmu.edu)<br>Carnegie Mellon University, Department of Psychology, 5000 Forbes Avenue, Pittsburgh, PA 15213 USA


#### Abstract

Prior research indicates a protracted developmental course in category-based reasoning. One possible explanation for the development of this ability is the gradual reorganization of semantic knowledge. To measure development of semantic knowledge we developed a new paradigm, the Semantic Space task, which uses distance in a two-dimensional space to infer semantic similarities between two objects. Using this paradigm we examined development of semantic knowledge in young children (preschoolers, kindergarteners, and first-grade children) and in adults. We also examined whether conceptual organization as measured by the Semantic Space task is predictive of children's scores on a category-based reasoning task. The findings point to the possibility that development of semantic knowledge plays an important role in the development of category-based reasoning.


Keywords: Semantic Space; Concepts; Cognitive Development

## Introduction

Category-based reasoning is a critical cognitive ability that enables an individual to generalize from the known to the unknown (Hayes, Heit, \& Swendsen, 2010; Proffitt, Coley, \& Medin, 2000). For example, upon learning that a chicken has 39 pairs of chromosomes, one may infer that a dove also has 39 pairs of chromosomes because chickens and doves are the same kind of animal (i.e., birds). Prior research indicates that children's category-based reasoning undergoes a protracted developmental course (e.g., Badger \& Shapiro, 2012; Fisher, Matlen, \& Godwin, 2011; Godwin, Matlen, \& Fisher, in press; Fisher, 2010; Fisher \& Sloutsky, 2005; Sloutsky, Kloos, \& Fisher, 2007), with marked improvements in category-based reasoning apparent between 4 and 6 years of age. However, it is not clear what factors underlie this improvement. One possible explanation for the development of category-based reasoning is representational change: the gradual accretion and reorganization of domain-specific knowledge (Goswami \& Brown, 1989, 1990; Rattermann \& Gentner, 1998).

Representational change has been identified as a factor that fosters cognitive development in a wide array of domains such as numerical development (e.g., Opfer \& Siegler, 2007), problem solving (e.g., Karmiloff-Smith, 1986), and analogical reasoning (e.g., Gentner, Rattermann, Markman, \& Kotovsky, 1995).

One of the few studies to examine the relationship between semantic development and category-based reasoning was conducted by Chi, Hutchinson, and Robin (1989). Chi et al. classified 6 -year-old children as either dinosaur experts or novices, based on their pre-test performance. Subsequently, the children completed an inference task about dinosaurs. The stimuli were digitally modified in order to create novel dinosaurs for both experts and novices. Chi et al. found that children who were classified as domain experts tended to make category-based inferences about the novel dinosaurs (e.g., "he is probably a good swimmer ... cause duckbills are good swimmers", p. 48). In contrast, children who were classified as dinosaur novices tended to make inferences based on a salient attribute (e.g., [the dinosaur] "could walk real fast cause he has giant legs", p. 49). These findings can be taken to suggest that category-based induction is a function of one's domain knowledge (also see Gobbo \& Chi, 1986).

There is also converging evidence suggesting that representational change may play a role in semantic development. First, multidimensional scaling studies have investigated people's ability to classify familiar objects. These studies provide evidence of advancement in classification from initially classifying objects according to more concrete characteristics to utilizing more abstract features when making groupings (e.g., Howard \& Howard, 1977; Saltz, Seller, \& Sigel, 1972). For example, preschoolage children are likely to classify familiar animals on the dimension of size, whereas school-age children are more likely to classify animals along the dimensions of domesticity and predativity.

Second, studies on the development of priming suggest that associative priming (e.g., faster responding to the word 'banana' after 'monkey' is presented) appears early in development whereas semantic priming (e.g., faster responding to the word 'banana' after 'cherry' is presented) develops during the school years (McCauley, Weil, \& Sperber, 1976; Arias-Trejo \& Plunkett, 2011).

Finally, work in cognitive modeling points to a gradual developmental progression in conceptual organization from relatively undifferentiated (e.g., groupings including a penguin, a trout, and an alligator) to more differentiated groupings (Kemp \& Tenenbaum 2008; Rogers \&

McClelland 2004). However, these predictions are yet to be confirmed by empirical studies.

The present study investigates how young children organize knowledge and whether the organizational structure changes over the course of development. In particular, we are interested in examining whether semantic similarity influences how knowledge is organized. In order to examine this issue, we developed a task in which participants are asked to represent the semantic similarity of animal dyads in two-dimensional space. The distance between animal pairs is taken as a measure of how closely the participants represented the concepts. The use of physical distance as an indicator of representational similarity has been successfully used in prior studies (e.g., Goldstone, 1994; Howard \& Howard, 1977).

Unlike multidimensional studies, in which children are free to arrange the items along any desired dimension, we explicitly asked children to put animals of similar kind close together on the game board. Therefore, this paradigm allowed us to examine whether knowledge of semantic similarity changes over the course of development in 4- to 7 -year-old children. We also assessed whether children's semantic organization scores are predictive of their tendency to engage in spontaneous category-based reasoning.

## Method

## Participants

Participants were preschool children ( $N=43$, Mage $=4.32$ years, $S D=0.28$ years), kindergarteners ( $N=22$, Mage $=5.41$ years, $S D=0.30$ years), and first-grade children ( $N=23$, Mage $=6.96$ years, $S D=0.32$ years) attending local preschools and elementary schools. The preschool children were also part of a longitudinal study examining the development of inductive reasoning (see Godwin, Matlen, Fisher, 2012). Adult participants were undergraduate students ( $N=20$, Mage $=20.38$ years $S D=1.22$ years) from a local university who received partial course credit for participation.

## Design \& Procedures

Children were tested individually in a quiet room adjacent to their classroom. Adult participants were tested in a laboratory located on campus. Tasks were administered by hypothesis-blind experimenters.

## Semantic Space Task

This task was designed to assess children's semantic organization. Visual stimuli entailed a game board consisting of a $9 \times 9$ grid (see Figure 1). Two 1" wooden blocks were used as game pieces. The wooden blocks were used as game pieces instead of pictures so that children would use their knowledge about kinds rather than rely on perceptual similarity. A similar approach has been successfully used in prior studies (e.g., Howard \& Howard, 1977).

Linguistic stimuli included 24 pairs of animal names. The stimuli could be classified into one of four categories: (1) semantically-similar dyads (e.g., lamb-sheep), (2) dyads that
share a common setting or habitat (e.g., lamb-horse), (3) unrelated dyads (lamb-swan), and (4) filler items. During the game, the target item was paired with three different test items (i.e., category-choice, setting/habitat match, and unrelated item). Linguistic stimuli is provided in Table 1.


Figure 1: Schematic depiction of the Semantic Space game board. Squares highlighted in red indicate the location of the critical trials. Squares highlighted in yellow mark the location of the filler trials. In the experiment proper, the location of the critical and filler trials was not marked and all squares on the board were white.

Table 1: Linguistic Stimuli for the Semantic Space Task

| Critical Trials |  |  |  |
| :---: | :---: | :---: | :---: |
| Target | Category <br> Choice | Setting/ <br> Habitat | Unrelated |
| Crocodile | Alligator | Fish | Grasshopper |
| Chick | Hen | Goat | Goldfish |
| Lamb | Sheep | Horse | Swan |
| Whale | Dolphin | Octopus | Elephant |
| Monkey | Gorilla | Parrot | Chipmunk |
| Mouse | Rat | Pig | Hippo |
| Filler Pairs |  |  |  |
| 1. Zebra/Turkey; 2. Bear/Snake; 3. Panther/ Turtle; |  |  |  |
| 4. Tiger/Butterfly; 5. Frog/Lion; 6. Giraffe/Seal |  |  |  |

In the Semantic Space task, participants were asked to help Zibbo the zookeeper organize his zoo by placing animals of the same kind close together. At the beginning of the task, the experimenter introduced the game and provided the participants with two examples (Example 1: a bunny and a rabbit were placed on adjacent squares on the game board and the experimenter explained that they should be placed close together because they are the same kind of thing; Example 2: a dog and a shark were placed far apart on the game board and the experimenter explained that they should be placed far apart because they are not the same kind of thing). On each test trial, the experimenter showed the participant where Zibbo put the target animal (e.g., the experimenter placed the first game piece on one of the squares marked in red in Figure 1 and said, "The zookeeper put the mouse here"). Then, the experimenter handed the participant the second game piece and asked him or her to identify where the test item should be placed (e.g., "Where
do you think the hippo should go?"). The participant's response was recorded in order to calculate the distance between the placement of the target and test item. After each trial was administered, the game board was cleared before the experimenter presented the next pair.

Placement of the 18 critical trials (i.e., semanticallysimilar dyads, common habitat/setting dyads, and unrelated items) was pseudo randomized to eight central squares (marked in red in Figure 1). The central squares were utilized for the critical trials in order to equalize the maximum possible distance from the square where the experimenter placed the target. Each of the eight squares was utilized at least twice and no more than three times. The six filler trials were randomly assigned to one of the remaining 24 squares in order to encourage participants to use the entire game board. The animal dyads were presented in one of two pseudo randomized orders. The following stipulations were used when creating the presentation orders: one filler trial was presented after every three critical trials, at least three trials were required in between target repeats, and at least two trials were presented in between semantically-similar dyads. The presentation order was counterbalanced across participants.

Participants' responses were scored in the following way: Raw scores were calculated for each trial by adding the number of squares occupied by the game pieces plus the number of squares between the target and test item (the distance was based on the shortest route between the two game pieces barring diagonal movement). A composite score for non semantically-similar dyads was created by averaging together participants' raw scores for common setting/habitat dyads and unrelated items. A Semantic Space Difference score was calculated by subtracting the average score for semantically-similar dyads from the non semantically-similar composite score. Positive difference scores indicate that participants put semantically-similar dyads closer together and non semantically-similar dyads farther apart. Difference scores approaching zero indicate that participants did not reliably discriminate between semantically-similar dyads and non semantically-similar dyads.

## Category-Based Reasoning Task

The Category-Based Reasoning task is a propertyinduction task in which children are presented with triads of objects and asked to generalize a novel property from the target to one of the test items. Each triad included a target, a category-choice, and a lure (e.g., lamb-sheep-frog). Nine label triads were administered: 3 triads referring to artifacts, 3 triads referring to inanimate natural kinds, and 3 triads referring to animate natural kinds (see Table 2 for the complete list of linguistic stimuli). The 3 animate natural kind triads were also included in the Semantic Space task.

Visual stimuli were presented on the computer and consisted of sets of three identical doors; see Figure 2. The objects remained hidden behind the doors in order to encourage children to rely on the category information
conveyed by the labels. This procedure has been utilized successfully in prior work (e.g., Fisher et al., 2011; Godwin et al., in press).
On each test trial children were told where each object was hiding. Children were told that the target item had a novel property and asked to generalize this property to either the category-choice or the unrelated lure. Concerns regarding the working memory demands of the task are mitigated based on Fisher et al.'s (2011) findings in which children recalled, with high accuracy, which objects were hiding behind each door at the end of each test trial.

All properties were two-syllable blank predicates. Two presentation orders were created: In order 1 all trials were randomized and the presentation order was reversed for order 2. Presentation order was counterbalanced across participants.

Table 2: Category-Based Reasoning Task Linguistic Stimuli

| Target | Category <br> Choice | Lure | Property |
| :---: | :---: | :---: | :---: |
| Artifacts |  |  |  |
| Rug | Carpet | Window | Koski |
| Sofa | Couch | Cup | Creighan |
| Shoe | Boot | Car | Troxel |
| Inanimate |  |  |  |
| Sea | Ocean | Apple | Manchin |
| Hill | Mountain | Flower | Erwin |
| Rock | Stone | Grass | Higa |
| Animate Natural Kinds |  |  |  |
| Alligator | Crocodile | Butterfly | Omat |
| Rat | Mouse | Fish | Lignin |
| Lamb | Sheep | Frog | Matlen |



Figure 2: Schematic depiction of the Category-Based Reasoning task. All instructions were given verbally by the experimenter.

## Picture Identification Task

The picture identification task served to assess children's familiarity with the labels utilized in the Category-Based Reasoning task. The picture identification task is a computer-based task akin to the Peabody Picture Vocabulary Test (Dunn \& Dunn, 1997). Stimuli included 27 labels and 108 pictures. On every trial, children were presented with 4 pictorial response options (the target object and 3 lures). Children were asked to point to the target object. The trials were presented in one of two orders. The presentation order was counterbalanced across participants. The task was administered immediately following the Category-Based Reasoning task.

The Category-Based Reasoning task and the picture identification task were not administered to adults; only preschoolers, kindergartners, and first-grade children completed this portion of the experiment. Additionally, because preschool children were also participating in a related longitudinal study, they participated in the CategoryBased Reasoning task twice, with approximately one week between the two testing sessions. Repeated testing was administered to obtain a more stable estimate of young children's performance. As children's scores on both testing sessions were within $3 \%$ (adjusted means $M$ time $=63 \%$, $M$ time2 $=66 \%$ ), we averaged the scores across the repeated administrations of the task. The analyses reported below are based on these average scores.

## Results

## Semantic Space Task Performance

Preschool children exhibited considerable variability in their performance on the Semantic Space task, with Difference scores ranging from -2.58 to 5.67 , and an average Difference score of 1.37 ( $S D=1.88$ ). Kindergarten children's performance was also highly variable, with Difference scores ranging from -0.92 to 7.00 and an average Difference score of 2.44 ( $S D=2.48$ ). The Difference scores of first-grade children ranged from -0.17 to 8.25 and their average difference score was $4.23(S D=2.08)$. The Difference scores of adult participants ranged from 2.50 to 6.50 and their average Difference score was 4.99 ( $S D=1.13$ ); See Table 3.

Participants' Difference scores were analyzed in a oneway ANOVA with age as the between-subject factor. This analysis revealed a significant effect of age, $F(3$, $104)=20.41, p<0.0001$. This effect was further explored through planned comparisons.

Performance on the Semantic Space task was found to improve with age. In general, preschoolers exhibited greater difficulties discriminating between semantically-similar and non semantically-similar dyads compared to the other age groups. Post-hoc Tukey tests revealed no significant difference between mean Difference scores of preschoolers ( $M=1.37$ ) and kindergartners $(M=2.44), p=0.163$. At the same time, both first graders $(M=4.23)$ and adults $(M=4.99)$ exhibited superior performance compared to preschoolers ( $M=1.37$ ), both $p s<0.0001$. A marked improvement in
performance on the Semantic Space task was observed between kindergarten $(M=2.44)$ and first-grade ( $M=4.23$ ), $p=0.014$. There was no significant difference between the Difference scores of first-graders and adults ( $p=0.583$ ), providing preliminary evidence that semantic differentiation for certain animal categories may begin to reach adult levels by 6 to 7 years of age. Taken together, the pattern of results suggests that the ability to reliably discriminate between semantically-similar and non semantically-similar dyads improves with age; see Figure 3.


Figure 3: Mean difference scores by age group. Semantic Space Difference scores were calculated by subtracting the average score for semantically-similar dyads from the non semantically-similar composite score. Error-bars represent the standard errors of the mean.

Table 3: Semantic Space mean scores by item type and age group.

| Age Group | Mean (SD) |
| :---: | :---: |
| Semantically-Similar Dyads |  |
| Preschool | $4.37(1.35)$ |
| Kindergarten | $4.09(1.63)$ |
| First-Grade | $2.91(0.87)$ |
| Adults | $2.32(0.42)$ |
| Non Semantically-Similar Dyads |  |
| Preschool |  |
| Kindergarten | $5.74(1.51)$ |
| First-Grade | $6.53(1.93)$ |
| Adults | $7.14(1.79)$ |
| Common Setting/Habitat Dyads |  |
| Preschool | $7.31(0.99)$ |
| Kindergarten | $6.83(1.60)$ |
| First-Grade | $6.66(1.87)$ |
| Adults | $6.13(1.14)$ |
| Unrelated Dyads |  |
| Preschool |  |
| Kindergarten | $5.64(1.76)$ |
| First-Grade | $6.61(1.95)$ |
| Adults | $7.62(2.08)$ |

## Picture Identification

The results from the picture identification task indicated that children were familiar with the labels used in the CategoryBased Reasoning task (Preschoolers: $M=.92, S D=.14$, Kindergarteners: $M=.99, S D=.01$, First-Graders: $M=.99$, $S D=.01$ ). As an additional precaution, for the preschool group the Category-Based Reasoning scores were adjusted for their vocabulary knowledge to ensure that children possessed the pre-requisite knowledge to perform categorybased induction. Thus, if a child missed an item on the picture identification task, this trial was removed from this child's Category-Based Reasoning score. This adjustment resulted in the increase of mean Category-Based Reasoning scores in preschoolers from $M=.62$ to $M=.64$. Because the picture identification scores of the other groups of participants were nearly at ceiling, no adjustments to the induction scores were made in the older age groups.

## Category-Based Reasoning Performance

Participants' reasoning scores were submitted to a oneway ANOVA with age as the between-subject factor. This analysis revealed a significant effect of age, $F(2,82)=16.49$, $p<0.0001$. This effect was further explored through planned comparisons.

Category-Based Reasoning performance improved as a function of age; see Figure 4. Posthoc Tukey tests revealed that both kindergarten and first-grade children exhibited superior performance on the Category-Based Reasoning task compared to the preschoolers; all $p s<.0001$. However, there was no significant difference in performance on the Category-Based Reasoning task between the kindergarten children and the first-grade children; $p=.90$.


Figure 4: Proportion of category-based responses by age group. Error-bars represent standard errors of the means. Line indicates chance performance.

The final analysis compared the mean Category-Based Reasoning scores to chance (.50) using single sample t-tests. Participants in all age groups exhibited Category-Based Reasoning performance that was significantly above chance; Preschoolers: $M=.64, S D=.22$; Kindergartners: $M=.87$,
$S D=.18$; First-graders: $M=.89, S D=.14$; all $t s>3.67$, all $p s<0.0001$. The rate of category-based responding in preschool-age children was somewhat higher than in our prior studies ( $M=.54$ across Fisher et al., 2011; Godwin et al., in press, Matlen, Fisher \& Godwin, under review). However, it should be noted that in the present study the sample of preschool children was recruited entirely from a laboratory campus school at a private university and our prior research utilized more diverse community-based samples.

## Is Category-based Reasoning Related to Children's Semantic Space Organization?

A correlational analysis was conducted to examine the potential relationship between children's Category-Based Reasoning performance and their performance on the Semantic Space task. This analysis revealed a significant positive correlation between the Semantic Space Difference scores and Category-Based Reasoning scores when scores were aggregated across preschoolers, kindergarteners, and first-graders, $r=.484, p<0.0001$ (see Figure 5). When separated by age group, there was a significant correlation in the preschool group ( $r=.473 ; p=0.002$ ), a marginally significant correlation in the kindergarten group ( $r=.34$, $p=0.12$ ), and no correlation among first-graders ( $r=.10$, $p=0.66$ ).

It is perhaps not surprising that the magnitude of the correlation between the Semantic Space Difference scores and Category-Based Reasoning scores decreased with age, as children's performance on both tasks improved and variability in performance decreased (e.g., many children achieved ceiling scores on the category-based reasoning task by first-grade). However, it is noteworthy that in preschoolage children, who exhibited a high degree of variability on both tasks, there was a fairly strong relationship between children's category-based reasoning and semantic organization.


Figure 5: A scatterplot of children's Category-Based Reasoning scores and their Difference scores on the Semantic Space task.

## Discussion

The results from the present study point to several novel findings. First, the new paradigm designed to measure development of semantic organization successfully captured increased differentiation among the animal concepts during the preschool and early school years. The gradual increase in the Semantic Space Difference scores in preschoolers, kindergarteners, and first-graders suggests that children increasingly become more sensitive to semantic similarity.

Second, the present findings provide preliminary evidence that individual differences in knowledge organization, as measured by the Semantic Space task, may be related to developmental differences in category-based reasoning.

In conclusion, these findings indicate that children's semantic knowledge undergoes gradual reorganization across development. Additionally, performance on this measure was found to predict preschoolers' inductive generalization performance. This latter finding suggests that the ability to make inductive inferences based on categories may be related to improvements in semantic organization.

## Acknowledgments

We thank Malika Sinha, Laura Pacilio, Alyssa Montanaro, Anna Loiterstein, Like Li, Kayoung Joung, and Rachel Walsh for help collecting data. We thank the children, parents, and teachers for making this work possible. This work was supported by a Graduate Training Grant awarded to Carnegie Mellon University by the Department of Education (R305B090023 and R305B040063).

## References

Arias-Trejo, N. \& Plunkett, K. (2009). Lexical-semantic priming effects during infancy. Philosophical Transactions of the Royal Society Biological Sciences, 364, 3633-3647.
Badger, J. R. \& Shapiro, L. R. (2012). Evidence of a transition from perceptual to category induction in 3- to 9 -year-old children. Journal of Experimental Child Psychology, 113, 131146.

Chi, M. T. H., Hutchinson, J., \& Robin, A. F. (1989). How inferences about novel domain-related concepts can be constrained by structured knowledge. Merrill-Palmer Quarterly, 35: 27-62.
Dunn, L. M., \& Dunn, L. M. (1997). Peabody Picture Vocabulary Test-Third Edition: Manual. Circle Pines, MN: American Guidance Services.
Fisher, A.V. (2010). What's in the name? Or how rocks and stones are different from dogs and puppies. Journal of Experimental Child Psychology, 105, 198-212.
Fisher, A.V., Matlen, B.J., \& Godwin, K.E. (2011). Semantic similarity of labels and inductive generalization: Taking a second look. Cognition, 118, 432-438.
Fisher, A. V. \& Sloutsky, V. M. (2005). When induction meets memory: Evidence for gradual transition from similarity-based to category-based induction. Child Development, 76(3), 583597.

Gentner, D., Rattermann, M. J., Markman, A. B., \& Kotovsky, L. (1995). Two forces in the development of relational similarity.

In T. J. Simon \& G. S. Halford (Eds.), Developing cognitive competence: New approaches to process modeling (pp. 263313). Hillsdale, NJ: LEA.

Gobbo, C., \& Chi, M. T. H. (1986). How knowledge is structured and used by expert and novice children. Cognitive Development, 1: 221-237.
Goswami, U., \& Brown, A.L. (1989). Melting chocolate and melting snowmen: Analogical reasoning and causal relations. Cognition, 35, 69-95.
Goswami, U. \& Brown, A.L. (1990). Higher-order structure and relational reasoning: Contrasting analogical and thematic relations. Cognition, 36, 207-226.
Godwin, K., Fisher, A., \& Matlen, B. (In press). Development of category-based reasoning in 4- to 7 -year-old children: The influence of label co-occurrence and kinship knowledge. Journal of Experimental Child Psychology.
Godwin, K., Matlen, B., \& Fisher, A. (2012). Development of Category-Based Reasoning in Preschool-Age Children: Preliminary Results of a Longitudinal Study. In N. Miyake, D. Peebles, \& R. P. Cooper (Eds.), Proceedings of the 34th Annual Conference of the Cognitive Science Society (pp. 16081613). Austin, TX: Cognitive Science Society.

Goldstone, R. (1994). An efficient method for obtaining similarity data. Behavior Research Methods, Instruments, \& Computers, 26, 381-386.
Hayes, B., Heit, E., \& Swendsen, H. (2010). Inductive reasoning. Wiley Interdisciplinary Reviews: Cognitive Science, 1, 278292.

Howard, D.V. \& Howard, J.H., Jr. (1977). A multidimensional scaling analysis of the development of animal names. Developmental Psychology, 13, 108-113.
Karmiloff-Smith, A. (1984). "Children's problem solving". In M.E.Lamb, A.L.Brown \& B.Rogoff (Eds.) Advances in Developmental Psychology, Vol.III. New Jersey: Erlbaum, 3990.

Kemp, C. \& Tenenbaum, J. B. (2008). Structured models of semantic cognition. Behavioral and Brain Sciences. 31(6), 717-718.
McCauley, C., Weil, C. M., \& Sperber, R. D. (1976). The development of memory structures as reflected by semanticpriming effects. Journal of Experimental Child Psychology, 22, 511-518.
Opfer, J. E., \& Siegler, R. S. (2007). Representational change and children's numerical estimation. Cognitive Psychology, 55, 169-195.
Proffitt, J.B. Coley, J.D., \& Medin, D. (2000). Expertise and category-based induction. Journal of Experimental Psychology: Learning Memory, and Cognition, 26(4), 811828.

Rattermann, M. J., \& Gentner, G. (1998). More evidence for a relational shift in the development of analogy: Children's Performance on a causal-mapping task. Cognitive Development, 13, 453-478.
Rogers, T. T. \& McClelland, J. L. (2004). Semantic Cognition: $A$ Parallel Distributed Processing Approach. Cambridge, MA: MIT Press.
Saltz, E., Seller, E., \& Sigel, I. E., (1972). The development of natural language concepts. Child Development, 43, 1191-1202.
Siegler, R. S., \& Opfer, J. E. (2003). The development of numerical estimation: evidence for multiple representations of numerical quantity. Psychological Science, 14, 237-243.
Sloutsky, V. M., Kloos, H., \& Fisher, A. V. (2007). When looks are everything: Appearance similarity versus kind information in early induction. Psychological Science, 18(2), 179-185.

# Classroom activities and off-task behavior in elementary school children 

Karrie E. Godwin (kegodwin@andrew.cmu.edu)<br>Carnegie Mellon University, Department of Psychology, 5000 Forbes Avenue, Pittsburgh, PA 15213 USA<br>Ma. V. Almeda (mqa2000@tc.columbia.edu)<br>Teachers College Columbia University, Department of Biobehavioral Sciences, 525 W. 120th St., Box 118 New York, NY 10027 USA<br>Megan Petroccia (petrocm@andrew.cmu.edu)<br>Carnegie Mellon University, Department of Psychology, 5000 Forbes Avenue, Pittsburgh, PA 15213 USA<br>Ryan S. Baker (baker2@exchange.tc.columbia.edu )<br>Teachers College Columbia University, Department of Human Development, 525 W. 120th St., Box 118<br>New York, NY 10027 USA<br>Anna V. Fisher (fisher49@andrew.cmu.edu)<br>Carnegie Mellon University, Department of Psychology, 5000 Forbes Avenue, Pittsburgh, PA 15213 USA


#### Abstract

Maintaining focused attention in the classroom is considered an important factor for successful learning. Loss of instructional time due to off-task behavior is recognized as a significant challenge by both researchers and practitioners. However, there has been little research into the factors contributing to off-task behavior. This paper reports results from the first large-scale study investigating how elementary school children allocate their attention in classroom environments and how patterns of attention allocation change as a function of gender, grade level, and instructional format. The findings indicate that instructional format is related to off-task behavior in elementary school students. These findings can begin to form a foundation for development of research-based guidelines for instructional design aimed to optimize focused attention in classroom settings.


Keywords: Off-Task Behavior; Attention

## Introduction

Loss of instructional time due to off-task behavior is a well-established problem in educational settings, recognized both by researchers (e.g., Baker, 2007; Karweit \& Slavin, 1981; Lee et al., 1999) and practitioners (e.g., Lemov, 2010) for over a hundred years (cf. Currie, 1884 as cited in Berliner, 1990). The link between the quality of attention and performance has been demonstrated in the cognitive psychology literature (e.g., Choudhury \& Gorman, 2000; Dixon \& Salley, 2007; DeMarie-Dreblow \& Miller, 1988). It has also been documented that off-task behavior has a negative impact on performance and learning outcomes in school settings (for reviews see Frederick \& Walberg, 1980; Goodman, 1990).

Despite considerable prior research on off-task behavior, designing effective, easy to implement, and scalable interventions to reduce off-task behavior has been challenging. Roberts (2001) suggests that many existing
interventions may be unsuccessful because they do not take into sufficient account the conditions that lead to off-task behavior. The goal of the present study was to expand upon prior research on off-task behavior in elementary school students to begin to elucidate the factors involved in off-task behavior, particularly the factors which are related to classroom activities and thus are malleable.

## Off-task Behavior

There is a variety of reasons why loss of instructional time occurs in schools; these reasons include but are not limited to: weather (e.g., snow days), sudden onset interruptions (e.g., announcements over the loudspeakers), and special events. However, it has been shown that student inattentiveness (i.e., engagement in off-task behavior during instructional time) is the biggest factor that accounts for loss of instructional time (Karweit \& Slavin, 1981). Prior research examining the frequency of off-task behavior has estimated that children spend between $10 \%$ and $50 \%$ of their time off-task in regular education classrooms (Lee et al., 1999; Karweit \& Slavin, 1981). Classrooms employing cognitive tutors report similar results with estimates of offtask behavior constituting $15 \%$ to $25 \%$ of instructional time (e.g., Baker, Corbett, \& Koedinger, 2004; Baker, 2007).

However, there has been limited research examining the factors associated with off-task behavior. Recently researchers have begun to explore the role of classroom design on children's off-task behavior. Godwin and Fisher (2011) found that classroom environments that contained relatively large amounts of visual displays (e.g., charts, posters, manipulatives) elicited more off-task behavior in kindergarten children compared to visual environments that were more streamlined. These design choices were found to hinder children's ability to attend to the content of the lesson and reduced learning outcomes. Related findings were obtained by Barrett et al. (2012). Barrett and
colleagues took a more holistic approach to design and incorporated building factors (e.g., physical space, navigation, furniture scale, etc.), environmental elements (e.g., light, sound, temperature, air quality, etc.), as well as classroom decor (e.g., color, organization, etc.). Barrett et al. found that these design choices (in combination with pupil factors) were related to students' later academic achievement.

Instructional format (e.g., whole-class instruction, small group instruction, etc.) is another important aspect of instructional design. Yet, little is known about the relationship between instructional format and overall rates and types of off-task behavior. The goal of the present study was to examine whether type of instruction is related to incidence of off-task behavior in elementary school students.

## The Present Study

This study examines whether specific instructional strategies are associated with incidence of off-task behavior in elementary school children, both in terms of the overall amount of off-task behavior, and the form which off-task behavior takes. Towards this goal we recorded patterns of attention allocation in elementary school students during a variety of instructional activities (e.g., whole-group instruction, small-group work, etc.).

## Method

## Participants

Twenty-two classrooms participated in the present study. Participating classrooms were selected from 5 local charter schools. Five grade-levels were recruited: Kindergarten through fourth-grade. The distribution across the five gradelevels was as follows: 5 kindergarten classrooms, 4 firstgrade classrooms, 5 second-grade classrooms, 2 third-grade classrooms, and 6 fourth-grade classrooms. The average class size was 21 students ( 10 males, 11 females). However, due to absences the average number of children observed in a single observation session was 18.9 children. The number of children observed per session ranged from 15 to 22 .

## Design and Procedure

Each classroom was observed four times during the second-half of the school year, resulting in a total of 84 observation sessions. Due to time constraints in four of the 22 classrooms only three observation sessions were conducted. The observation sessions were staggered across two time periods (Time 1: February-April 2012, Time 2: May-June 2012) with two observation sessions occurring during each time period. The average delay between observation sessions within a single time period was 3.7 days (the delay ranged from 1 to 14 days). The average delay across time periods was 73.2 days. Each observation session lasted approximately one-hour.

## Operationalization of on- and off-task behavior

For the present study, focused attention was defined as a "state in which attention is directed more or less exclusively
to one target or task" (Ruff \& Rothbart, 1996, p.110). Focused attention was operationalized through visual engagement. If children were directing their eye gaze at the teacher (or classroom assistant), the instructional activity, or toward appropriate instructional materials, the child was classified as on-task. If the child was looking elsewhere, they were classified as off-task. Eye gaze is a common measure of visual attention (for reviews see Henderson \& Ferreira, 2004; Just \& Carpenter, 1976), and it is arguably a reasonable (albeit imperfect) measure of focused attention.

## Coding

All coders were trained in the Baker-Rodrigo Observation Method Protocol (BROMP) for coding behavioral data in field settings (Ocumpaugh, Baker, \& Rodrigo, 2012) using software developed for the android handheld computer. All coders received extensive training consisting of coding videotapes and live observation sessions. Inter-rater reliability was established prior to the study proper. Kappa values ranged from 0.79 to 0.84 . This level of reliability is in line with past classroom research coding off-task behavior, and exceeded the 0.75 threshold to which Fleiss (1981) refers as "excellent" in field settings.

Children were observed using a round-robin coding strategy, in order to reduce the tendency of observers to attend to more salient instances of off-task behavior. The order in which children were observed was determined at the beginning of each session. Each time a child was observed the observation lasted for up to 20 seconds. The first unambiguous behavior observed during the 20 -second period was recorded. Quick glances were considered ambiguous behaviors, and coders were instructed to wait for a clear behavior to occur. If a behavior was noted before 20 seconds elapsed, the coder proceeded to the next child, and a new 20 -second observation period began. Coders observed the children using peripheral vision or side-glances. Peripheral vision was utilized in order to avoid looking directly at the student being observed. This technique makes it less apparent to the child that $\mathrm{s}(\mathrm{he})$ is being observed. This procedure has successfully and reliably captured students' behavior in prior work which assessed student behavior and affect (cf. Baker et al., 2006; Baker et al., 2010; Ocumpaugh et al., 2012).

Coders classified children's behavior as on- or off-task. If the child was looking at the teacher (or classroom assistant), the instructional activity, and/or the relevant instructional materials, they were categorized as on-task. If the child was looking elsewhere, they were categorized as off-task. Contextual clues (i.e., teacher instructions) were also taken into consideration when distinguishing between on- and offtask behavior. For example, if a child was instructed to discuss an idea with a partner, coders would classify conversing with another peer as on-task unless the coders could clearly discern that the conversation was unrelated to the instructional task.

If the child was classified as off-task, the type of off-task behavior was recorded. Six mutually exclusive categories of
off-task behavior were logged: (1) Self-distraction, (2) Peer distraction, (3) Environmental distraction, (4) Supplies, (5) Walking, or (6) Other. Self-distraction entailed engagement with something on the child's own body, such as an article of clothing or an appendage, as well as episodes in which the child would close their eyes. Peer distraction was defined as interacting with or looking at another student(s) when not directed to do so. Environmental distractions include interacting with or looking at any object in the classroom that was not related to the task at hand, while Supplies consists of inappropriately using any object that was part of the assigned task (e.g., playing with a writing utensil). Walking was operationalized as a student physically walking around the classroom when it was not considered appropriate for the task. Other distractions included student behavior that was off-task but did not clearly align with the five aforementioned categories. A seventh category Unknown was also included to capture rare instances in which it was unknown whether the child was on- or off-task, and it was impossible or inappropriate for the observer to relocate in order to obtain a better view of the child. Unknown was also used when students left the classroom for various reasons (e.g., to use the restroom).

Children in each session were treated as a different set of students since it was not possible to link observations across the four sessions. Thus, a total of 1,587 student-session pairs were observed. A student-session pair refers to a specific student observed by a coder within a specific session. The average number of observations per session was 330.13 and the average number of observations per child within a session was 17.58.

## Data Analysis: Variables

Using these data, we attempted to predict within an observation session each student's total on- or off-task behavior as well as the type of behavior the student tended to engage in while off-task. Two categories of predictor variables were considered for incorporation into the models: student characteristics and instructional design. Gender and grade were included as student characteristics. Predictor variables pertaining to instructional design included the proportion of each classroom instructional format and the variable Transitions/Duration of Instructional Format.

Instructional format was included as a predictor variable in order to examine whether certain instructional formats elicit differential amounts of off-task behavior. Six different instructional formats were coded: (1) individual work, (2) small-group or partner work, (3) whole-group instruction at desks, (4) whole-group instruction while sitting on the carpet, (5) dancing, and (6) testing. The proportion of time students spent in each of the aforementioned formats was calculated. The average duration for each instructional format is provided in Table 1.

Transitions were noted every time the teacher paused instruction to change from one activity to another (e.g., transitioning from working on a math problem to listening to a short story). In many cases, transitions coincided with a
change in instructional format (e.g., switching from wholegroup instruction to small-group instruction); however this was not always the case as transitions could occur without a change in instructional format (e.g., with children rotating from one small group activity to another). Transitions were frequently marked by the teacher asking the children to get out new instructional materials (e.g., "Please get out your math binders") or requesting students to change locations (e.g., "Please put your notebooks away and come to the carpet").

Table 1. Time spent in each instructional format

| Average Time Spent (sec) Per Instructional Format |  |
| :---: | :---: |
| Individual Work | 1,424 |
| Small Group | 1,587 |
| Whole-group Instruction at Desks | 1,805 |
| Whole-group Instruction on Carpet | 1,263 |
| Dancing | 141 |
| Testing | 2,530 |

The primary dependent variable was the proportion of ontask behavior of a specific student within a specific session. Additional models were also constructed in order to predict peer off-task behavior and environment-based off-task behavior, as these two types of off-task behavior were common sources of distraction for elementary school children (See Table 2). Environment-based off-task behavior was of particular interest as it is a malleable factor that could theoretically be targeted when designing interventions aimed to mitigate off-task behavior.

## Data Analysis: Approach

We predicted student on-task behavior using a regression tree algorithm (cf. Witten \& Frank, 2005), which sets up a decision tree to predict a numerical value. Binary decisions are made based on specific variables. After several decisions are made, a numerical prediction is given. To determine these specific variables, regression trees find breakpoints within data, where relationships change (mostly) at a certain value of a variable. Regression trees can find more complicated interactions and relationships between variables than is typically possible with linear regression methods, while still remaining more constrained than neural networks or support vector machines-as such, they occupy a moderate position in the trade-off between goodness of fit and flexibility of fit/parsimony. The specific implementation of regression tree used in this paper is REPTree in RapidMiner 5.2 (Mierswa et al., 2006). This relatively rapid algorithm builds a tree using reduced error pruning; an approach designed to produce relatively conservative models (Witten \& Frank, 2005).

Resultant models were evaluated using six-fold student level cross-validation. In this process, students are split randomly into six groups. For each possible combination, a feature is developed using data from five groups of students
before being tested on the sixth "held out" group of students. By cross validating at this level, we increase confidence that features will be accurate for new students.

Within this paper, cross-validation (Efron \& Gong, 1983) is used instead of statistical significance testing for multiple reasons. First, cross-validation assesses how accurate a model is likely to be for new data, rather than assessing the likelihood that a specific data set's results are due to chance. In assessing generalizability, cross-validation has the same goal as the use of information criteria. In fact, the k-fold cross-validation approach used here is thought to be asymptotically equivalent to the Bayesian Information Criterion (BiC) (Shao, 1993). Second, there is not an appropriate statistical significance test for the data used here for two reasons: (1) there is not a well-known statistical significance test for regression trees, and (2) student IDs are not connected across sessions. Testing statistical significance without a student term would result in a bias strongly in the direction of statistical significance; conversely, using a student-session term would result in having an order of magnitude more parameters, biasing strongly against statistical significance.

## Results

Consistent with prior research, children were largely on task: $71 \%$ of children's observed behaviors were on-task. As seen in Table 2, three of the most common types of off-task behavior observed were Peer distractions (45\%), SelfDistractions (18\%), and Environmental distractions (16\%).

Table 2. Descriptive statistics for students' on- and off- task behavior

| Proportion of Observed Behaviors |  |
| :---: | :---: |
| On-task Behavior | $71 \%$ |
| Off-task Behavior | $29 \%$ |
| Proportion of Off-task Behaviors |  |
| Self-Distraction | $18 \%$ |
| Peer Distraction | $45 \%$ |
| Environmental Distractions | $16 \%$ |
| Supply Distractions | $11 \%$ |
| Walking | $3 \%$ |
| Other Distractions | $8 \%$ |
| Descriptive Statistics | Mean (SD) |
| Observations per session | $330.13(63.6)$ |
| Observations per session per child | $17.58(3.7)$ |
| Student/Session pairs observed | 1,587 |

Models predicting on-task behavior were fit based both on instructional design and on limited student demographics (e.g., grade-level and gender). The best overall model predicting on-task behavior was found for the regression tree when student demographics were not included. This model obtained a cross-validated correlation coefficient of $r=0.352$. The cross-validated correlation coefficients for the
"instructional design plus demographic" models were as follows: A regression tree which added gender achieved a cross-validated correlation of 0.322 to the frequency of student on-task behavior and a regression tree model which added grade-level achieved a cross-validated correlation of 0.329. As these "instructional design plus demographics" models achieved lower cross-validated correlation than the simpler model which only considers instructional design, we can infer that including this demographic information does not improve model fit in a generalizable fashion (as mentioned above, this is akin to achieving a better BiC : the additional fit does not compensate for the added model complexity/flexibility of fit). As such, for determining offtask behavior it does not appear to be important whether an elementary school student is a boy or a girl, once instructional design is taken into account. Similarly, gradelevel does not seem to be an important factor, once the influence of grade on instructional design is taken into account.

In this data set, regression trees achieved generally better performance than linear regression. A linear regression model based on instructional design achieved a crossvalidated correlation of 0.221 to the frequency of student on-task behavior. No linear regression model (regardless of the feature set used) performed better than the corresponding regression tree model.

Within instructional design, both the format and the variable Transitions/Duration of Instructional Format were associated with a better model. Removing either of these variable types from the model resulted in worse crossvalidated correlation.

The final regression tree model was rather complex, with 63 leaf nodes (final decision values) and 62 decision nodes. It can be easier to understand some of the key data relationships by considering the cross-validated and regular correlations for single-feature linear regression models. In Table 3, both cross-validated correlations and regular correlations are given. It is worth noting that cross-validated correlations should always be positive (a negative crossvalidated correlation does not imply a negative relationship, but that the relationship reverses direction when applied to different parts of the data; e.g., a negative cross-validated correlation implies that the model is worse than chance). Directionality of the relationship should be inferred from the regular correlation.

As seen in Table 3, the relationship between instructional format and on-task behavior varies as a function of the type of instructional format. Individual work and whole-group instruction at desks were negatively associated with on-task behavior, while small group-work, whole-group instruction while sitting on the carpet, dancing, and testing were positively associated with on-task behavior. It is worth noting that the individual variables may have weak associations, even as reasonable prediction is achieved from a combination of variables. Note, however, that this is not simply a case of an overly-complex model predicting noise; the cross-validated correlation of the overall model is an
indication that the model works on entirely unseen data. The variable Transitions/Duration of Instructional Format was also found to be positively correlated with on-task behavior.

We also generated models to predict peer off-task behavior and environmental off-task behavior, using the same features and modeling methods. The cross-validated correlation of the REPtree model based solely on instructional design was 0.244 for peer off-task behavior and 0.161 for environmental off-task behavior. These correlations did not increase substantially if gender or grade-level were included. With these results, the peer model appears to perform somewhat better than the weak correlation achieved in the environment model, but neither model was as effective as the model predicting the overall amount of on-task behavior.

Table 3. Goodness of single-feature linear regression models at predicting on-task behavior (note that cross-validated correlations are always positive, unless the model performs worse than chance on new data).

| Feature | Direction <br> of <br> relationship | Cross- <br> validated <br> correlations | Correlations |
| :---: | :---: | :---: | :---: |
| Individual <br> work | Negative | .000 | -.018 |
| Small-group <br> work | Positive | .000 | .032 |
| Whole-group <br> instruction at <br> desks | Negative | .114 | -.113 |
| Whole-group <br> instruction <br> carpet | Positive | .110 | .110 |
| Dancing | Positive | .017 | .043 |
| Testing | Positive | .025 | .051 |
| Transitions/ <br> Duration of <br> Inst. Format | Positive | .075 | .075 |
| Gender | Positive | .108 | .109 |
| Grade | Positive | .005 | .039 |

The strongest individual feature correlation (using linear regression and non-cross-validated correlations) for peer off-task behavior is the amount of time spent in wholegroup instruction while sitting on the carpet ( $r=-0.136$ ), followed by the amount of time spent in small-group work ( $r=0.119$ ). Similarly, the strongest individual feature correlation for environmental off-task behavior is smallgroup work ( $r=-0.115$ ). These findings suggest that instructional format does matter for determining specific off-task behaviors. However, the magnitude of correlation for the full model indicates that instructional format determines to a greater degree whether a student will go off-
task, than exactly how they will go off-task. Clearly, the type of instructional format may influence students' choices of how they will go off-task (e.g., a student may be more likely to engage in peer off-task behavior during smallgroup work when another child is in close proximity); however, the exact manifestation of off-task behavior may be influenced by momentary factors (i.e., the most interesting item/person in the classroom at a specific moment).

## Discussion

The present work is the first large-scale study of off-task behavior in elementary school students to investigate the relationship between features of instructional design and incidence of off-task behavior. Specifically, we examined whether type of instructional format (e.g., individual work, small-group work, whole-group work, etc.) and the variable Transitions/Duration of Instructional Format are related to the overall rate of off-task behavior. Our findings indicate that both variables are related to children's engagement in instructional activities. At the same time, children's gender and grade-level (K-4) made only a marginal contribution to off-task behavior once features of the instructional design (e.g., instruction format and variable Transitions/Duration of Instructional Format) were taken into account.

The reported results also indicate that certain types of instructional format are associated with more on-task behavior than others, although further research is required to explicate this finding. There are several possible hypotheses that could be explored. One potential underlying factor is variations in teacher supervision. It is feasible that classroom management is easier for certain instructional formats (e.g., small-group work) than others (e.g., wholegroup instruction at desks). Consequently, instructional formats that are easier for teachers to supervise may result in a reduction in opportunities for students to go off-task.

Secondly, student engagement in the instructional task may also vary across instructional formats. For instance, instructional activities that take place individually or at the students' desks (i.e., whole-group instruction at desks) may be less engaging or motivating than small-group activities which tend to be more socially oriented and include more hands-on learning components. Instructional activities that are more motivational may in turn increase students' on-task behavior. Additionally, instructional duration varies across these formats. Thus, children may be better able to maintain a state of focused attention when instruction consists of small blocks of instructional activities verses instructional activities that occur over a longer duration (cf. Ruff \& Lawson, 1990; Sarid \& Breznitz, 1997). Currently these hypotheses are speculative, and they require additional investigation to determine their viability.

As stated previously, off-task behavior is a significant problem in educational settings because it is thought to impede learning. Optimizing instructional design to promote on-task behavior is a desirable goal; however, there is a paucity of research linking instructional design choices to
attention allocation in classroom settings. The present findings are a first-step in providing empirical evidence to inform instructional design.

## Acknowledgments

We thank Laura Pacilio for her help collecting data. We also thank the children, parents, and teachers who made this project possible. The work reported here was supported by a Graduate Training Grant awarded to Carnegie Mellon University by the Department of Education (R305B090023) and by the Institute of Education Sciences, U.S. Department of Education (R305A110444). The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

## References

Baker, R. S. (2007). Modeling and understanding students' off-task behavior in intelligent tutoring systems. Proceedings of ACM CHI 2007: Computer-Human Interaction, 1059-1068.
Baker, R. S., Corbett, A. T., \& Koedinger, K. R. (2004). Detecting student misuse of intelligent tutoring systems. Proceeding of the $7^{\text {th }}$ International Conference on Intelligent Tutoring Systems, 531-540.
Baker, R.S.J.d., Corbett, A.T., \& Wagner, A.Z. (2006). Human classification of low-fidelity replays of student actions. Proceedings of the Educational Data Mining Workshop at the 8th International Conference on Intelligent Tutoring Systems, 29-36.
Baker, R.S.J.d., D'Mello, S.K., Rodrigo, M.M.T., \& Graesser, A.C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments. International Journal of Human-Computer Studies, 68 (4), 223-241.
Barrett, P., Zhang,Y., Moffat, J., \& Kobbacy, K. (2012). A holistic, multi-level analysis identifying the impact of classroom design on pupils' learning. Building \& Environment, 59, 678-689.
Bloom, B. S. (1976). Human characteristics and school learning. New York: McGraw-Hill.
Carroll, J. (1963). A model of school learning. Teachers College Record, 64, 723-733.
Choudhury, N. \& Gorman, K. S. (2000). The relationship between sustained attention and cognitive performance in 17-24 month old toddlers. Infant and Child Development, 9, 127-146.
Cobb, J.A. (1972). Relationship of discrete classroom behavior to fourth-grade academic achievement. Journal of Educational Psychology, 63, 74-80.
DeMarie-Dreblow, D. \& Miler, P. H. (1988). The development of children's strategies for selective attention: Evidence for a transitional period. Child Development, 59, 1504-1513.
Dixon, W. E. \& Salley, B. J. (2007). "Shh! We're tryin' to concentrate": Attention and environmental distracters in novel word learning. The Journal of Genetic Psychology, 167(4), 393 414.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Pagani, L.S., Feinstein, L., Engel, M., Brooks-Gunn, J., Sexton, H., Duckworth, K., \& Japel, C. (2007). School readiness and later achievement. Developmental Psychology, 43(6), 1428-1446.
Efron, B. \& Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. The American Statistician, 37(1), 36-48.

Frederick, W.C., Walberg, H.J., \& Rasher, S.P. (1979). Time, teacher comments, and achievement in urban high schools. Journal of Educational Research, 73 (2), 63-65.
Godwin, K. \& Fisher, A. (2011). Allocation of Attention in Classroom Environments: Consequences for Learning. In L. Carlson, C. Hölscher, \& T. Shipley (Eds.), Proceedings of the 33 rd Annual Conference of the Cognitive Science Society (28062811). Austin, TX: Cognitive Science Society.

Goodman, L. (1990). Time and learning in the special education classroom. SUNY Press.
Henderson, J., \& Ferreira, F. (2004). In The interface of language, vision, and action: Eye movements and the visual world. Taylor \& Francis, New York.
Hendrickson, L. (1979). The first ' $R$ '. Paper presented at the annual meeting of the American Educational Research Association. San Francisco, CA.
Hough, D. \& Bryde, S. (1996). The effect of full-day kindergarten on student achievement and affect. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.
Just, M., \& Carpenter, P. (1976). Eye fixations and cognitive processes. Cognitive Psychology, 8, 441-480.
Karweit, N. \& Slavin, R. E. (1981). Measurement and modeling choices in studies of time and learning. American Educational Research Journal, 18(2), 157-171.
Lahaderne, H.M. (1968). Attitudinal and intellectual correlates of attention: A study of four sixth-grade classrooms. Journal of Educational Psychology, 59 (5), 320-324.
Lemov, D. (2010). Teach like a champion: 49 techniques that put students on a path to college. San Francisco, CA: Jossey-Bass.
Lee, S. W., Kelly, K. E., \& Nyre, J. E. (1999). Preliminary report on the relation of students' on-task behavior with completion of school work. Psychological Reports, 84, 267-272.
Leinhardt, G. (1977). Program Evaluation: An Empirical Study of Individualized Instruction. American Educational Research Journal, 14, 277-93.
Lloyd, J. W. \& Loper, A. B. (1986). Measurement and evaluation of task-related learning behaviors: Attention to task and metacognition. School Psychology Review, 15(3), 336-345.
Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., \& Euler, T. (2006). YALE: Rapid Prototyping for Complex Data Mining Tasks. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006), 935-940.

Ocumpaugh, J., Baker, R.S.J.d., \& Rodrigo, M.M.T. (2012). Baker-Rodrigo Observation Method Protocol (BROMP) 1.0. Training Manual version 1.0. Technical Report. Manila, Philippines: Ateneo Laboratory for the Learning Sciences.
Pagani, L., Fitzpatrick, C., Archambault, I., \& Janosz, M. (2010). School readiness and later achievement: a French-Canadian replication and extension. Developmental Psychology, 46(5), 984-994.
Roberts, M. (2002). Research in practice: Practical approaches to conducting functional analyses that all educators can use. The Behavior Analyst Today, 3, 83-87.
Ruff, H. \& Rothbart, M. K. (1996). Attention in Early Development. New York: Oxford University Press.
Shao, J. (1993) Linear model selection via cross-validation, Journal of the American Statistical Association 88(422), 486494.

Witten, F., \& Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques. San Francisco: Morgan Kaufmann.

# How pitch accents and focus particles affect the recognition of contextual alternatives 

Nicole Gotzner (nicole.gotzner@hu-berlin.de)<br>Katharina Spalek (katharina.spalek@staff.hu-berlin.de)<br>Department of German Language and Linguistics, Humboldt-Universität zu Berlin, Dorotheenstr. 24, 10117 Berlin, Germany

Isabell Wartenburger (isabell.wartenburger@uni-potsdam.de)<br>Department of Linguistics, Universität Potsdam<br>Karl-Liebknecht-Str. 24-25, 14476 Potsdam, Germany


#### Abstract

Listeners are sensitive to contrastive alternatives in online language comprehension (e.g., Braun \& Tagliapietra, 2010). Such alternatives play a crucial a role in the definition of particles like only which have been found (i) to facilitate recall of contextual alternatives (Spalek, et al., in revision) and (ii) to hamper the rejection of unmentioned alternatives (Gotzner et al., in preparation). The present study investigated the impact of combining a contrastive accent with a particle on memory for contextual alternatives. The results revealed that $\mathrm{L}+\mathrm{H}^{*}$ accents (contrastive) facilitate recognition of a contextual alternative to the accented item compared with $\mathrm{H}^{*}$ accents (non-contrastive). Adding either the particle only or also to the $\mathrm{L}+\mathrm{H}^{*}$ accent slows probe recognition relative to a condition with bare $\mathrm{L}+\mathrm{H}^{*}$ accent. Hence, while contrastive accenting directly increases the salience of alternatives in a listener's mental model, focus particles lead to an initial processing cost.


Keywords: contrastive alternatives, $\mathrm{L}+\mathrm{H}^{*}$ pitch accent, focus particles, recognition memory.

## Introduction

Certain pitch accents convey the information that a statement is contrastive (cf. Pierrehumbert \& Hirschberg, 1990). For example, when uttering the sentence Mary passed the exam with a specific intonation contour on the noun Mary ( $\mathrm{L}+\mathrm{H}^{*}$ according to the ToBI system), a speaker expresses that Mary passed the exam in contrast to other persons in the discourse.
There is evidence that contrastive intonation contours lead to the activation of alternative expressions to the contrastively-stressed elements. For example, Braun and Tagliapietra (2010) found that a sentence produced with an $\mathrm{L}+\mathrm{H}^{*}$ accent on the critical word initiated priming of contrastively-related targets to the critical words in a lexical decision task (e.g., ANTENNA facilitated satellite).

There is also evidence that prosody influences longterm memory for contextual alternatives. Fraundorf, Watson and Benjamin (2010) compared contrastive $\left(\mathrm{L}+\mathrm{H}^{*}\right)$ and non-contrastive ( $\mathrm{H}^{*}$ ) pitch accents in discourses that contained a contrast set with two elements. After exposure to all stimuli, participants had to perform a recognition memory task. The results revealed that the $\mathrm{L}+\mathrm{H}^{*}$ accent increased both the number of hits to correct statements, and the number of correct rejections of the
contrast item, suggesting that contrastive pitch accents enhance memory for the accented element itself as well as for its alternatives. The rejection of unmentioned items (lures), however, was not affected by the pitch accent manipulation.

According to the contrast representation account advocated by Fraundorf et al. (2010) listeners use contrastive accents to encode additional information about items in the contrast set. So, on this account, processing our example sentence Mary passed the exam with an $\mathrm{L}+\mathrm{H}^{*}$ accent would encourage the inference that other students did not pass the exam.

## Focus Particles

Another means of expressing such a contrast is provided by certain lexical items. Placing the focus particle only in front of the utterance (Only) Mary passed the exam results in a similar effect, since the particle expresses that the focused element Mary but not the alternatives lead to a true assertion (cf. König, 1991).

In a previous delayed recall study (Spalek, Gotzner \& Wartenburger, in revision), we explored the impact of focus particles on memory for contextual alternatives. We exposed participants to dialogs that introduced a set of three elements and either contained the particle only, even or no particle (control condition) in the critical sentences. The pitch accent type on the element mentioned in the critical sentences, the focused element, was constant across conditions ( $\mathrm{H}^{*}$ on the accented syllable). The results showed that memory for the alternatives to the focused element improved in the presence of the particles only and even relative to the control condition.

In a subsequent study (Gotzner, Spalek \& Wartenburger, submitted), we tested the effects of focus particles on the recognition of contextual alternatives and rejection of unmentioned alternatives after exposure to an item. We found that participants were slower in rejecting unmentioned alternatives to a focused expression in case the utterances contained the particles only, even or also. This inhibitory effect was present when we enumerated a set of alternatives in the context and when the stimuli only mentioned a semantic category. We concluded that focus particles encourage richer encoding of the alternative set
which leads to better recall of the mentioned alternatives particle when tested at a longer delay (cf. Spalek et al., in revision).

Apart from the fact that focus particles and pitch accents belong to a different formal linguistic level, the two have a different impact on the status of inferences about the alternatives, according to linguistic theory (see for example Krifka, 2007 for an overview). The exhaustivity inference (that the statement does not hold for the alternatives) drawn from a contrastive accent is an implicature which arises through pragmatic reasoning. By definition, implicatures are cancellable inferences and it follows that an utterance with an $\mathrm{L}+\mathrm{H}^{*}$ accent has a reading that does not exclude the alternatives. On the contrary, the exclusion of alternatives is part of the conventional meaning of only, i.e. lexically encoded (cf. Rooth, 1992). In accordance with these theoretical distinctions, experimental work has shown that German participants draw fewer exhaustive inferences from sentences that bear intonational focus than from sentences with only (cf. Onea \& Beaver, 2011). ${ }^{1}$

In contrast to exclusive particles, additive particles like also and even presuppose that a statement holds for at least one of the alternatives and express that the proposition holds for the focused element as well (cf. König, 1991). Yet this linguistic difference did not affect the retrieval of alternatives in the experiments we carried out so far. Therefore, we concluded that the observed effects were driven by the fact that focus particles must refer to a set of alternatives by their semantic definition.

To sum up, there is evidence that focus particles facilitate recall of contextual alternatives (Spalek et al., in revision) and hamper the rejection of unmentioned alternatives when the focused element carries an unmarked accent (Gotzner et al., in preparation). Contrastive pitch accents lead to the activation of contextual alternatives (cf. Braun \& Tagliapietra, 2010) and have been found to facilitate long-term memory for focused elements and contextual alternatives, but not for unmentioned items (cf. Fraundorf et al., 2010).

## The Current Experiment

The aim of the current study was to investigate whether focus particles and contrastive pitch accents rely on the same cognitive mechanism and whether combining the two induces additive effects. That is, if the presence of alternatives is indicated by a contrastive pitch accent in addition to a focus particle, does the accessibility of contextual alternatives increase even further?

In non-tonal languages like English and German focus can be marked prosodically and by specific lexical and syntactic structures (cf. Krifka, 2008) while the different types of focus marking are not mutually exclusive. Following the proposal in Calhoun (2009), we assume

[^282]that by choosing a particular structure, the speaker wishes to make the alternatives particularly salient for the hearer. Hence, it might be that highlighting the alternatives from multiple sources leads to additive effects.

Participants in our experiment were presented with short discourses that mentioned two referents and one of them was mentioned again in the second critical sentence, either pronounced with (a) an unmarked accent $\left(\mathrm{H}^{*}\right)$ or (b) a contrastive one $\left(\mathrm{L}+\mathrm{H}^{*}\right)$. In addition to the $\mathrm{L}+\mathrm{H}^{*}$ accent, condition (c) contained the exclusive particle only and (d) the inclusive particle also. After exposure to the stimuli, participants were asked to recognize the alternative to the noun mentioned in the critical sentences (not the mentioned noun itself).

Additionally, we introduced a delay manipulation that was based on a study by Glenberg, Meyer and Lindem (1987). In their experiment, they found that manipulating the associatedness of a referent and an object affected the mental model listeners constructed from a discourse only after a delay of one filler sentence but not if no filler sentence was presented before test. Since we were not interested in priming effects (cf. Braun \& Tagliapietra, 2010), but in the participants' mental model representation of the contrast set, we administered a similar probe recognition task with a delay manipulation like Glenberg et al. (1987). Accordingly, we hypothesized that the effects only unfold if probe recognition is tested at a delay of one filler sentence.

We predicted that the $\mathrm{L}+\mathrm{H}^{*}$ accent will facilitate recognition of contextual alternatives compared to the $\mathrm{H}^{*}$ accent. Concerning the comparison between the $\mathrm{L}+\mathrm{H}^{*}$ condition and the particle conditions there are three alternative predictions: If accenting and particles rely on the same cognitive mechanism, we should obtain the same magnitude and time-course of the effects. If, however, contrastive accenting and particles are additive, we expect that participants are fastest at recognizing the focus alternative in the two conditions with particles (i.e., (c) and (d)) since they contain a particle in addition to the contrastive accent. Alternatively, it might be that pitch accents are used immediately to encode information about the alternatives (see for example Watson et al., 2008) while such effects take more time to unfold in the case of focus particles. According to this hypothesis, we expect the $\mathrm{L}+\mathrm{H}^{*}$ accent to facilitate the recognition of alternatives after one intervening filler sentence whereas the particles should not cause any positive effects at this point of time.

## Methods

## Participants

A total of 24 native speakers of German ( 15 female and 9 male, mean age 26.1 years, age range $22-30$ ) were recruited from a participant pool at the Institute of Psychology of Humboldt University and were paid 7

Euros in compensation. None of them reported any vision or hearing difficulties.

## Materials

We created 80 discourses that followed the structure of the example presented in (1). The first sentence introduced two referent nouns. The second critical sentence mentioned one of the referents again and described an action. As described above, we introduced a delay variable: On 50 percent of the trials, an additional filler sentence (in brackets in the example) was presented. The filler sentence always consisted of five words and contained a pronoun keeping the referent foregrounded (cf. Glenberg et al., 1987). Across stimuli, the order of mention of the two referents was counterbalanced.
(1) Der Richter und der Zeuge verfolgten die Beweisführung. (Nur/Auch) der RICHTER/der Richter glaubte dem Angeklagten. (Er verkündete das Urteil.)

## Approximate translation:

The judge and the witness followed the argument. (Only/also ${ }^{2}$ ) the JUDGE/the judge believed the defendant. (He announced the verdict.)

The critical sentences were recorded in four versions in each of the conditions $\left(\mathrm{H}^{*}, \mathrm{~L}+\mathrm{H}^{*}\right.$, only and also) by a female research assistant who was trained on focus accentuation and had a middle German accent close to the standard variety of German. After recording, the utterance with the $\mathrm{L}+\mathrm{H}^{*}$ accent (b) was cross-spliced into the two utterances with only (c) and also (d). Thereby, conditions (b), (c) and (d) all contained the $\mathrm{L}+\mathrm{H}^{*}$ accent and all prosodic characteristics of the sentences were held constant. In total, there were 8 experimental conditions: 4 focus conditions crossed with the delay of either 0 or 1 filler sentence.

Acoustic analyses were performed to compare the accented syllable of the $\mathrm{H}^{*}$ (a) and $\mathrm{L}+\mathrm{H}^{*}$ (b) conditions. Figure 1 shows the pitch contour of the accented syllable averaged over all items. Additionally, acoustic analyses were performed to compare the duration, maximum pitch, pitch difference and intensity across accent type condition. Table 1 summarizes the means, standard deviations and results of repeated measures ANOVAs (within item) comparing these acoustic parameters. The analyses revealed that the syllable with $\mathrm{L}+\mathrm{H}^{*}$ accent had a higher pitch excursion, a greater pitch difference, intensity and duration.

The critical trials always probed recognition of the alternative to the noun in subject position, hence requiring

[^283]a yes response ${ }^{3}$. A set of 50 filler items was constructed and 60 items from another experiment were added to counterbalance yes- and no-responses and to prevent subjects from concentrating on the nouns in subject position. Half of the filler items consisted of 2 sentences and half of them consisted of 3 sentences.

Figure 1: Mean pitch contour of the accented syllable across $\mathrm{H}^{*}$ and $\mathrm{L}+\mathrm{H}^{*}$ conditions.


Table 1: Summary of acoustic analyses. Table shows mean values (SE) by accent type and results of F-tests.

| Parameter | H | $\mathrm{L}+\mathrm{H}^{*}$ | $F_{(1,79)}$ | $p$ - <br> value |
| :--- | :--- | :--- | :--- | :--- |
| Duration (s) | 0.17 | 0.23 | 129.1 | 0.001 |
|  | $(0.01)$ | $(0.01)$ |  |  |
| Maximum Pitch | 195.1 | 226.9 | 19.7 | 0.001 |
| (Hz) | $(4.5)$ | $(6.2)$ |  |  |
| Pitch difference | 33.6 | 56.3 | 8.5 | 0.005 |
| (Hz) | $(4.5)$ | $(6.2)$ |  |  |
| Intensity (dB) | 69.1 | 73.3 | 208.0 | 0.001 |
|  | $(1.9)$ | $(2.2)$ |  |  |

Eight experimental lists were created by rotating through the focus ( $\mathrm{H}^{*}, \mathrm{~L}+\mathrm{H}^{*}$, only and also) and delay conditions ( 0 vs. 1 filler sentence) according to a Latin square design. Hence, there were 10 items per condition within a given list. Each list further contained the 110 filler items resulting in a total of 190 items. The lists were pseudo-randomized for each participant so that no more than three filler or test trials were presented in a row and a given focus condition appeared only twice in a row.

[^284]
## Procedure

The experiment started with an instruction displayed on the computer screen. The instructions told the participants that they will be presented with auditory stimuli and that their task is to decide whether a subsequently presented word had appeared in the story or not. They were also asked to respond as accurately and as quickly as possible and to listen to the exact wording. After the instructions were displayed, participants performed four practice trials and were allowed to adjust the sound volume.

Each trial began with the onset of a central fixation cross displayed for 700 ms followed by a discourse that was presented over headphones. Each of the sound files included 2000 ms of silence after the last sentence. On 50 percent of the trials, the audio files contained an additional filler sentence. Probes appeared after the entire audio files, that is, either after the critical sentence (50\%, delay 0 ) or after the filler sentence ( $50 \%$, delay 1 ).

Each probe was administered visually with an offset of 50 ms and the participants indicated whether it had appeared in the discourse by button press. The probe word stayed on the screen until a response was made. If subjects did not respond within a time frame of 4000 ms , the trial counted as a miss. With an offset of 500 ms the next trial was initiated. Every 33 trials, subjects had a short break. In total, there were 6 experimental blocks. The entire experiment lasted about 50 minutes.

## Results

Trials in which subjects responded incorrectly (5.9 \%) or that were more than 2 standard deviations from a participants' mean reaction time ( $4.9 \%$ ) were excluded from the analysis. Accuracy was similar across conditions. The log-RTs were analyzed with a series of mixed models following the procedure described in Baayen (2008). Since we had different predictions for the two delay conditions, we fit two separate models for the two data sets.

Figure 2 shows the mean RTs and standard errors across focus conditions at 0 delay. The final statistical model contained the log-RTs, focus condition and trial as fixed factors, subjects, items as random factors and random slopes for trial. We chose the condition with the $\mathrm{L}+\mathrm{H}^{*}$ accent as reference level, since the two conditions with particles contained an $\mathrm{L}+\mathrm{H}^{*}$ accent as well and since we were interested in whether the particles caused effects in addition to the contrastive accent. The final model had a log likelihood of 222.6 with 828 observations (27 further observations were excluded based on the distribution of residuals and fitted values). Regarding the focus conditions, the model did not reveal any reliable differences across conditions ( $p>.2$; see Table 2 for model details). This is in line with Glenberg et al. (1987) who argued that effects of foregrounding in a mental model should only be observable after a delay of 1 filler sentence.

Figure 2: Mean RT across conditions (delay 0). Error bars indicate standard errors.


Figure 3 displays the mean RTs broken by focus condition at delay 1 . The final mixed effects model for this data set contained the same factors and it had a log likelihood of 206.1 (803 observations, 27 excluded). The analysis revealed that the alternatives were recognized slower in the condition with $\mathrm{H}^{*}$ accent in comparison with the $\mathrm{L}^{+} \mathrm{H}^{*}$ accent ( $t=2.8, S E=.03, p<.01$ ). Hence, the $\mathrm{L}+\mathrm{H}^{*}$ accent facilitated recognition of the alternatives. Compared to the condition with $\mathrm{L}+\mathrm{H}^{*}$ accent, the two particles only and also led to an inhibitory effect ( $\mathrm{L}+\mathrm{H}^{*}$ vs. only: $t=2.0, S E=.03, p<.05 ; \mathrm{L}+\mathrm{H}^{*}$ vs. also: $t=2.5$, $S E=.03, p<.05)$. Thus, adding a focus particle to the contrastive pitch accent led to slower probe recognition times. Table 3 summarizes estimates, confidence intervals and $p$-values extracted by Markov chain Monte Carlo sampling (10000 runs).

Figure 3: Mean RT across conditions (delay 1). Error bars indicate standard errors.


Table 2: Summary of the mixed model (at delay 0).

|  | Estimate | Lower <br> bound | Upper <br> bound | pMCMC |
| :--- | :--- | :--- | :--- | :--- |
| Intercept <br> $\left(\mathrm{L}+\mathrm{H}^{*}\right)$ | 6.47 | 6.37 | 6.56 | 0.0001 |
| $\mathrm{H}^{*}$ | 0.04 | -0.02 | 0.09 | 0.18 |
| only | 0.002 | -0.05 | 0.06 | 0.94 |
| also | 0.01 | -0.04 | 0.07 | 0.72 |
| Trial | -0.002 | -0.003 | -0.0015 | 0.0001 |

Table 3: Summary of the mixed model (at delay 1).

|  | Estimate | Lower <br> bound | Upper <br> bound | pMCMC |
| :--- | :--- | :--- | :--- | :--- |
| Intercept <br> $\left(\mathrm{L}+\mathrm{H}^{*}\right)$ | 6.39 | 6.3 | 6.49 | 0.0001 |
| $\mathrm{H}^{*}$ | 0.08 | 0.03 | 0.14 | 0.006 |
| only | 0.06 | -0.002 | 0.11 | 0.05 |
| also | 0.07 | 0.01 | 0.13 | 0.02 |
| Trial | -0.01 | -0.004 | -0.002 | 0.0001 |

## Discussion

The study reported here extends earlier findings on the role of pitch accents in the retrieval of contextual alternatives in two ways. First, it indicates that the $\mathrm{L}+\mathrm{H}^{*}$ accent not only induces priming or activation of contrastive alternatives (cf. Braun \& Tagliapietra, 2010) but it also benefits recognition memory for contextual alternatives. Second, it shows that such effects already unfold after one intervening filler sentence (in comparison with the long delay introduced by Fraundorf et al., 2010).

The task we used required participants to construct a mental model from the auditory discourses they were presented with and they had to recognize a referent that was an alternative to the element mentioned in the critical sentences. In accordance with Glenberg et al. (1987), the effects we observed only evolved after the inclusion of one filler sentence. This finding also provides evidence that there were no general processing differences across conditions. For example, it might be argued that the conditions with particles contained one more word compared to the other conditions leading to increased processing times. However, we would expect to see the same differences across conditions at 0 delay if this were the right explanation of the effects.

Looking at the recognition times in the $\mathrm{H}^{*}$ and $\mathrm{L}+\mathrm{H}^{*}$ conditions across delays, it becomes obvious that recognition performance was equal at 0 delay. By comparison, the $\mathrm{H}^{*}$ slowed and the $\mathrm{L}+\mathrm{H}^{*}$ facilitated recognition times at delay 1. This pattern of results is in line with a recent study by Husband \& Ferreira (2012). They found that contrastively-stressed as well as neutral prime words primed contrastive targets at an immediate SOA. At a longer SOA, the contrastive associates only
maintained facilitation in the condition with $\mathrm{L}+\mathrm{H}^{*}$ accent. Hence, it seems that initially the activation of contrastive alternatives does not differ across $\mathrm{H}^{*}$ and $\mathrm{L}+\mathrm{H}^{*}$ conditions. At a delay of one filler sentence in our experiment, activation of the alternatives decays in the $\mathrm{H}^{*}$ case (as was also argued by Husband \& Ferreira, 2012) and augments in the $\mathrm{L}+\mathrm{H}^{*}$ case.

This pattern of results confirms the assumption by Watson et al. (2008) that the interpretational domains of the two accent types overlap to some extent in that $\mathrm{H}^{*}$ are compatible with contrastive and non-contrastive referents but that $\mathrm{L}+\mathrm{H}^{*}$ favor contrastive referents. The data are hence in line with the claim that $\mathrm{H}^{*}$ and $\mathrm{L}+\mathrm{H}^{*}$ accent do not necessarily form two discrete categories but that the latter is the more contrastive variant of the former.

The $\mathrm{L}+\mathrm{H}^{*}$ pitch accent made the contrastive alternative more accessible at a delay of one filler sentence. In contrast, adding either an exclusive or inclusive particle led to a processing cost. It is conceivable that the observed pattern is due to a facilitatory effect by the contrastive accent combined with an inhibitory effect of the particles since the conditions with particles contained exactly the same (cross-spliced) recorded utterance.

Such inhibitory effects are consistent with the results of our probe recognition experiments where we found that participants were slower in rejecting unmentioned alternatives to a focused expression in case the utterances contained the particles only, even or also (vs. no particle). This inhibitory effect even showed when we explicitly introduced a set of three elements. Hence, we concluded that the particles led to the activation of further unmentioned alternatives in a listener's mental model slowing the participant's ability to indicate that these elements had not appeared in the story.
Interestingly, Fraundorf et al. (2010) did not find any effects of contrastive pitch accents on the rejection of unmentioned items (so-called lures). Despite the differences between the experiments, it might be that contrastive accents only affect the representation of items in the explicit contrast set while focus particles lead to the retrieval of further alternatives accounting for the inhibitory effects observed here (as well as in Gotzner et al., in preparation).

In conclusion, it seems that focus particles lead to an initial processing cost. The positive effect on the retrieval of contextual alternatives that we have observed in our delayed recall experiments (Spalek et al., in revision) seems to need time to develop. By contrast, pitch accents are used immediately by a listener to encode information about the alternatives (see also Watson et al., 2008 for a similar argument). Yet further research is needed to establish whether focus particles and contrastive pitch accents cause additive effects at longer delays or whether the effects level out. It should also be noted that the different experimental paradigms, probe recognition and delayed recall, impose quite different task demands.

Concerning the theoretical account of the observed effects, the comparison between the different types of focus particles leads to an interesting conclusion. Following the original formulation of the contrast representation account, we should expect the particle only to pattern along with contrastive accents but not the particle also since Fraundorf et al. (2010) speculate that contrastive accents create a representation of what did not happen. However, across all our experiments we found that both exclusive (associated with an exhaustive meaning) and additive particles affected memory for information-structural alternatives to a similar extent. Additive particles like also do not have an exhaustive meaning but they presuppose that a statement holds for at least one alternative. Yet what exclusive and additive particles have in common is that both require a salient set of alternatives.

In light of the findings, we propose a modification of the contrast representation account: We assume that focus-sensitive particles (and contrastive accents) make reference to an alternative set and highlight its relevance for interpretation (see also Krifka, 2007 for a recent definition of focus). This creates a salient representation of the alternatives and makes them more easily retrievable on the long run. In immediate recognition (after one intervening filler sentence), adding a particle to a contrastive accent hampers the participants' ability to match a probe with the alternatives encoded in the mental model, as observed in the present experiment.

The fact that the facilitatory effects of contrastive accenting already unfolded after one intervening sentence whereas the particles led to a processing cost at this point of time could either indicate that the underlying cognitive mechanisms are different in the two domains or simply that prosody is integrated at an earlier point of time. However, further research is required to determine the exact cognitive mechanisms that underlie the observed effects. Taken together with the results of Gotzner et al. (submitted), the inhibitory effects of the particles might indicate that listeners activate further unmentioned items interfering with the recognition of the explicitlyintroduced alternatives. Such effects might be absent in the case of contrastive accents considering the results of Fraundorf et al. (2010).

## Conclusion

We showed that contrastive pitch accents enhance the recognition of contextual alternatives after auditory exposure to an item. The combination of a contrastive pitch accent and a focus particle, in turn, led to an inhibitory effect. We therefore suggest that the facilitatory effects of focus particles on the retrieval of contextual alternatives (cf. Spalek et al., 2013) take time to develop.

## Acknowledgments

This work was supported by the DFG as part of the Collaborative Research Centre SFB 623 "Information

Structure". Isabell Wartenburger is supported by the Stifterverband für deutsche Wissenschaft (Claussen-Simon-Stiftung). We thank Janine Klose and Vinzent Müller for carrying out the experiments and assistance with sound recording. We are also grateful to two anonymous reviewers for helpful comments.

## References

Baayen, R. H. (2008) Analyzing Linguistic Data. A Practical Introduction to Statistics Using R. Cambridge University Press.
Braun, B., \& Tagliapietra, L. (2010). The role of contrastive intonation contours in the retrieval of contextual alternatives. Language and Cognitive Processes, 25, 1024-1043.
Calhoun, S. (2009). What makes a word contrastive? Prosodic, Semantic and Pragmatic Perspectives. In D. Barth-Weingarten, N. Dehé and A. Wichmann (eds.) Where prosody meets pragmatics: Research at the interface (Studies in Pragmatics 8, 53-78), Bingley.
Fraundorf, S.H., Watson, D.G. \& Benjamin, A.S. (2010). Recognition Memory reveals just how CONTRASTIVE contrastive accenting really is. Journal of Memory and Language, 63, 367-386.
Glenberg, A., Meyer, K. \& Lindem, K. (1987). Mental models contribute to foregrounding in text comprehension. Journal of Memory and Language, 26, 69-83.
Gotzner, N., Spalek, K. \& Wartenburger, I. (in preparation). The impact of focus particles on the rejection of contrastive alternatives.
Husband, E.M. and Ferreira, F. (2012). Generating contrastive alternatives: activation and suppression mechanisms. Talk given at the 25 th CUNY Conference on Human Sentence Processing, New York, NY.
König, E. (1991). The Meaning of Focus Particles: A Comparative Perspective. London: Routledge.
Krifka, M. (2007). Basic notions of Information Structure. In Féry, C., Fanselow, G., \& Krifka, M. The notions of information structure. Interdisciplinary Studies on Information Structure. (ISIS), 6. Potsdam: Universitätsverlag Potsdam.
Onea, E. \& Beaver, D. (2011). Hungarian focus is not exhausted. In Ed Cormany, Satoshi Ito \& David Lutz (eds.), Proceedings of semantics and linguistic theory (SALT) 19, 342-359.
Pierrehumbert, J., \& Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In P. Coehn et al. (Eds.), Intentions in communication. Cambridge: MIT Press.
Rooth, M. (1992). A theory of focus interpretation. Natural Language Semantics, 1, 75-116.
Spalek, K., Gotzner, N. \& Wartenburger, I. (in revision). Not only the apples: focus particles influence memory for information-structural alternatives.
Watson, D., Gunlogson, C., \& Tanenhaus, M. (2008). Interpreting pitch accents in on-line comprehension: $\mathrm{H}^{*}$ vs. L+H*. Cognitive Science, 32, 1232-1244.

# The object without qualities: referring with negative properties 

Martijn Goudbeek (m.b.goudbeek@uvt.nl)<br>Tilburg center for Communication and Cognition, School of Humanities Tilburg University, The Netherlands

Inge Haagmans (inge.haagmans@gmail.com)<br>Tilburg center for Communication and Cognition, School of Humanities<br>Tilburg University, The Netherlands

Emiel Krahmer (e.j.krahmer@uvt.nl)<br>Tilburg center for Communication and Cognition, School of Humanities<br>Tilburg University, The Netherlands


#### Abstract

Two experiments investigated the production and comprehension of referring expressions that contain a negative property ("the marker without a cap"). Experiment 1 showed that participants do use negative properties in their object descriptions, but that they were almost always accompanied by other properties, leading to referential overspecification. In experiment 2, participants identified objects based on descriptions that contained negative properties. While participants were faster in identifying objects that were described with preferred properties such as color, response times for objects described with a negative property ("the marker without a cap") and a positive property ("the marker with a cap") did not differ. The results provide behavioral grounds for extending referring expression generation algorithms to include negative properties.


Keywords: Referring expressions, speech production and comprehension, negative properties

## (A sort of) Introduction

Image you are reading an interesting cognitive science manuscript and want to highlight an important passage. On the desk of your colleague (see figure 1) are two markers and since you prefer yellow markers, that is the one you would like her to pass on to you. You could phrase your request like this "Could you pass me the yellow marker, please?". In that case you have produced a referring expression with the property color as a means to distinguish between the two markers. However, a viable alternative to this question is "Could you pass me the marker without the cap, please?". In that case, you have used an negative property to refer to the marker of your choice. This paper investigates to what extent speakers produce referential expressions that contain negative properties and how listeners process these expressions. In doing so, we aim to inform computational models of referring expression generation.

Producing a suitable referring expression can be seen as a problem of choice (Krahmer \& van Deemter, 2012). Which properties does a speaker include in the description when asking for the marker? In addition to color and the absence of a cap, the location ("the marker on the left") and size ("the slightly smaller marker") come to mind as possible distinguishing properties for your marker. How to choose between these properties, when they are all suitable candidates for inclusion in the referring expression? Many current compu-


Figure 1: The two markers on the desk of your colleague.
tational approaches, such as Dale and Reiter's (1995) Incremental Algorithm use a fixed ordering of properties that are serially added to the description until all distractors have been ruled out. Since color is a property that is usually highly preferred (Pechmann, 1989), this is the first property that the Incremental Algorithm would add. Since the inclusion of color in the description uniquely identifies the item of your choice, the algorithm stops and produces "The yellow marker" and because the preference order is a fixed order, negative properties like "without a cap", would never be included in a referring expression.

Standard REG (Referring Expression Generation) algorithms such as the Incremental Algorithm do not consider boolean operations such as negation in the generation process (Krahmer \& van Deemter, 2012). Recently however, several attempts have been made to incorporate negation (and other boolean operators) in REG algorithms, either by extending the incremental algorithm (with a specific focus on referring to sets of objects) with boolean expressions (van Deemter, 2002; van Deemter \& Krahmer, 2006), or by reinterpreting the problem of referring expression generation in terms of description logic or conceptual graphs (Areces et al., 2008; Croitoru \& van Deemter, 2007). All of these approaches are computational in nature and until now the question of if, when, and how human speakers produce referring expressions with negative attributes has not been addressed. The behavioral data presented here can help inform computational approaches to referring expressions by, for example,
making the generated expressions more natural (Viethen \& Dale, 2006; Dale \& Viethen, 2010) or by providing inspiration for the further development of the algorithms. For example, the boolean extensions to the Incremental Algorithm proposed by van Deemter (2002) assumes that negative properties are less preferred than their positive counterparts. Here, we explicitly test this assumption with stimuli that do afford a description with negative properties, but can also be uniquely identified with other, more preferred properties.

While referring expression research is primarily concerned with the production of referring expressions, there is an increasing need to assess how listeners process the descriptions that are generated by REG algorithms (Krahmer \& van Deemter, 2012). The most important criterion for a successful algorithm is whether the expressions generated mimic those of humans. However, humans might not always be good at taking the needs of their listeners into account (Horton \& Keysar, 1996) and the references produced by human speakers might not be the most optimal ones. Thus, if our production experiment shows that speakers do use negative properties in their descriptions, this does not necessary mean that listeners will easily deal with such expressions. By combining a production experiment with a comprehension task, these issues will be addressed.

Two separate experiments will investigate the production and comprehension of referring expressions that contain an negative property. In the production experiment, three research questions are addressed. The first is whether speakers will produce referring expressions with negative properties at all in situations that afford (but not necessitate) the use of a negative property in a description. The second question is whether the number of positive properties necessary for a uniquely identifying description matters. It might be the case that speakers are more likely to use a negative property when the alternative means using a more complex description with, for example, two positive properties. This finding would contrast with the expressions generated by the Incremental Algorithm, which has no backtracking ability to take the length of the resulting referring expressions into account (Dale \& Reiter, 1995; van Deemter, 2002). Finally, the phenomenon of overspecification is addressed. Speakers often produce referential expressions that contain more information than strictly necessary (for example, by referring to the marker with "the yellow marker without a cap"). Speakers have been shown to be more likely to produce overspecified references when they use dispreferred properties such as orientation (Goudbeek \& Krahmer, 2012) and when they refer to target in more complex stimulus arrangements (Koolen et al., in press). Referring expressions with negative properties are arguably more complex and less preferred, leading speakers to overspecify descriptions that contain a negative property.

The comprehension experiment focuses on the processing of negative properties and addresses the question whether identifying objects that are described with negative properties takes more time than identifying objects that are described
with positive properties.

## Experiment 1: Producing referring expressions with negative expressions

In Experiment 1 participants produced descriptions of everyday objects. They could either refer to these objects with one or two positive properties ("the large marker" or "the large yellow marker") or with a negative property ("the marker without a cap" ${ }^{1}$. Additionally, this experiment investigated whether the number of properties necessary in the alternative description influenced the referential choices of the speakers.

## Method

## Participants

Twenty undergraduate students (eleven females) from the participant pool of Tilburg University took part in exchange for partial course credit. They were all native speakers of Dutch and were between 18 and 25 years old.

## Materials

In the production experiment, the stimuli consisted of 96 sets of three objects. The target object was always presented in the middle and was marked with a black rectangle. Of the 96 stimuli, 64 were target objects that needed one of more properties for unique identification and and 32 were typeidentifiable objects. Crucially, of the target objects, 32 could be described with a negative property such as "the marker without a cap" and 32 could be described with a positive property such as "the marker with a cap". The objects that had either positive or negative properties were a marker (with or without a cap), a cup (with or without a handle), a basket (with or without a lid) and a bottle (with or without a cap). See figure 2 for an example of two stimuli.

In addition to these properties, the target objects could alternatively be described with properties such as color or size. These are considered to be preferred properties in REG research (Dale \& Reiter, 1995; Pechmann, 1989) and should thus serve as viable alternatives. To investigate whether the number of preferred properties necessary for identification plays a role in determining whether speakers will use a negative property, there was a condition where one positive property would suffice (e.g., "the orange marker", see figure 2a) and a condition where two positive properties were necessary (e.g., "the large orange marker, see figure 2b). In both conditions, one negative property ("the marker without a cap") would always suffice (see figure 2 ).

In addition, the experiment contained 32 type-identifiable stimuli that could be described by using type only (e.g., "the rabbit", "the strawberry"), leading tot a total of 96 stimuli.

## Procedure

Participants were seated in a sound-attenuated room and were instructed to describe the object in such a way that a naive lis-

[^285]
(a) A target that can be referred to with one positive property ("the orange marker") or a negative property ("the marker without a cap").

(b) A target that can only be referred to with two positive properties ("the large yellow marker") or a negative property ("the marker without a cap").

Figure 2: The target stimuli used in the production experiment
tener would be able to identity the target object from the other two. The target object was always presented in the middle, and marked by a black square (see figure 2). Each stimulus was presented for five seconds, during which the participant's description was recorded. After the presentation and recording, a new set of objects immediately appears on the screen. Stimulus presentation and response recording took place on a PC and was controlled through the open-source package Opensesame (Mathôt et al., 2012). The experiment lasted about ten minutes, after which participants were debriefed and thanked for their cooperation.

## Results

The descriptions of the participants were annotated with respect to which property they used in their description (size, color, pattern, and whether their descriptions contained a negative property or not. We also annotated whether a description contained any redundant properties, to see whether the use of negative properties might cause speakers to overspecify. First, we investigated whether participants used the negative property in their referring expressions at all in the condition that afforded to do so (see the left side of figure 3).

While our prediction is that speakers would use the negative property in their descriptions, the algorithm proposed in van Deemter (2002) would never include negative properties


Figure 3: Proportion of descriptions with a negative property in conditions that afforded the use of negative properties (left) and proportion of descriptions with the corresponding positive property in conditions that afforded their use (right).
in the expressions it generates for these stimuli (since a preferred alternative is available). The results show that in almost half of the cases where the stimulus affords using a negative property, our participants did so $(M=.47, S D=0.31,95 \%$ $\mathrm{CI}=0.33-0.62$ ). A one sample t -test showed that this value did indeed differ significantly from zero $(t)(20)=6.76, p<$ .001 , Cohen's $d=1.52$ ). To investigate the effect of the target type (affording negative properties or not), we compared the proportion of descriptions with negative properties for stimuli that afforded the use of negative properties $(M=.47, S D=$ $0.31,95 \% \mathrm{CI}=0.33-0.62$ ) with descriptions with the corresponding positive properties for stimuli that afforded the use of positive properties $(M=.52, S D=0.30,95 \% \mathrm{CI}=0.38$ 0.66 ). The boxplot in figure 3 shows a large amount of overlap, indicating little difference between using a negative or a positive version of a property. A logistic regression with target type (positive versus negative) as outcome variable and proportion of properties used as predictor confirmed the lack of an effect of target type ( $\beta=0.53, \mathrm{SE}=1.06$, Wald $=0.25$, $\left.p=.62, R_{\text {Nagelkerke }}^{2}=0.01\right)$.

Next, we tested the hypothesis that speakers would be more inclined to use negative properties when the alternative description required two positive properties. Figure 4 shows that speakers indeed produced more descriptions containing negative properties when the alternative contains two positive properties $(M=.57, S D=0.34,95 \% \mathrm{CI}=0.41-0.72$ ) compared to when the alternative contains one positive property ( $M=.38, S D=0.33,95 \% \mathrm{CI}=0.23-0.53$ ). However, a logistic regression analysis with number (one versus two) as predictor and the proportion of descriptions with a negative property as outcome variable only yielded a marginally significant effect ( $\beta=1.72, \mathrm{SE}=1.00$, Wald $=2.98, p<.08$, $\left.R_{\text {Nagelkerke }}^{2}=0.10\right)$.

Finally, we investigated whether referring with a negative


Figure 4: Proportion of referring expressions with a negative property in conditions with alternatives that required one or two positive properties for a uniquely identifying description.


Figure 5: Proportion of overspecification in referring expressions that contained a negative property versus referring expressions that did not contain a negative property.
property causes speakers to overspecify (with a positive property) more than referring with positive properties only. For the subset of stimuli that afforded the use of negative properties, we calculated the proportions of overspecified references (defined as any reference that contains an additional property that would have been sufficent to uniquely identify it) for descriptions with negative and positive properties. Figure 5 clearly shows that when speakers use negative properties, they are more likely to use additional properties ( $M=.78, S D$ $=0.32,95 \% \mathrm{CI}=0.63-0.93)$ than when they do not use negative properties $(M=.123, S D=0.16,95 \% \mathrm{CI}=0.05$ 0.20 ). A logistic regression analysis with referring expression (negative versus positive) as predictor and the proportion of overspecified descriptions as outcome variable showed a significant effect ( $\beta=2.62, \mathrm{SE}=1.75$, Wald $=13.35, p<.001$, $\left.R_{\text {Nagelkerke }}^{2}=0.73\right)$.

## Discussion

The results of this experiment show that speakers certainly do not shy away from using negative properties in their referring expressions. In almost half of the cases participants included a negative property in their description when the stimulus afforded to do so. Importantly, our participants were never forced to use the negative property to identify the target referent: all objects could be uniquely identified by (a combination of) color, size or pattern, or by their type alone. This provides a psycholinguistic motivation for developing ways to generate referring expressions that contain negative attributes. Importantly, the boolean extensions of the Incremental Algorithm described in Areces et al. (2008), van Deemter \& Krahmer (2006), and van Deemter (2002) do not fully do justice to the patterns observed here. For instance, while van Deemter (2002) assumes that negative properties are dispreferred, our speakers produced them just as much as their positive counterparts, even when properties that are considered to be more preferred (such as color and size) were at their disposal.

The comparison between alternatives that contained either one or two positive properties showed, albeit marginally significant, that the more complex the alternative expression becomes, the more likely speakers are to use a negative property in their description. This is difficult to explain for REG algorithms that depend on entering properties from a fixed preference order and that do not take into account the length or the complexity of the resulting referring expression. Interestingly, even though speakers often produce descriptions with negative properties, our analysis also showed that the resulting referring expressions hardly ever contain only these negative properties, but were often overspecified. This is in line with findings from previous studies such as Goudbeek \& Krahmer (2012) and Koolen et al. (in press) that show that speakers are more likely to overspecify when their references include less preferred properties or when visual scenes get more complex. A possible explanation is that speakers could take the processing limits of their listeners into account (Arnold, 2008) and adjust the complexity of their utterances to suit.

## Experiment 2: Understanding referring expressions with negative properties

Experiment 2 investigated the comprehension of referential expressions with negative properties.

## Method

## Participants

Twenty-eight undergraduate students (nineteen females) from the participant pool of Tilburg University took part in exchange for partial course credit. They were all native speakers of Dutch and were between 18 and 25 years old. None of the participants took part in Experiment 1.

## Materials

For Experiment 2, the visual materials used in Experiment 1 were stripped from the black rectangle and were presented in the upper left, middle and upper right corner of the screen (see figure 6). They were complemented with a start box at the bottom of the screen. In addition, we recorded instructions to indicate the target object that contained five different ways to refer to the target. These referential expressions either used a positive property ("click on the marker with a cap"), a negative property ("click on the marker without a cap"), color or size ("click on the yellow marker"), color and size ("click on the large yellow marker"), and type-identifiable stimuli ("click on the strawberry"). All these descriptions were minimally specifying in that they provided sufficient, but not more, information to identify the target object. The instructions were spoken with a neutral intonation by a female speaker of Dutch that was unaware of the goal of the experiment. The position of the target was always either in the left or right upper corner and was counterbalanced across items. Since larger targets are easier to move towards and click on (Fitts, 1954), the size of the objects was counterbalanced as well (e.g., sometimes the instruction was "click on the small marker").

## Procedure

Participants were seated in a sound-attenuated room and were given headphones to listen to the prerecorded instructions. They encountered the setup displayed in Figure 6 and could start the spoken description by moving their mouse pointer over the box labeled "START", after which response recording started. The participants' task was to click as quickly as possible on the object that was being described by the prerecorded referring expression. We used the software package MouseTracker (Freeman \& Ambady, 2010) to present the images and speech stimuli and record the mouse movements and clicks. The experiment lasted about 20 minutes. After the experiment, the participants were debriefed and thanked for their cooperation.

## Results

Table 1 shows the response times of the participants for the five different referring expressions. It should come as no surprise that the participants responded fastest to the typeidentifiable items ("the strawberry"). Furthermore, the descriptions that used one were faster than those that used two two preferred properties. These in turn have faster response times than the conditions with either positive or negative properties, that do not seem to differ much.

These effects were evaluated statistically with a one-way within-subjects analysis of variance with type as a within factor with five levels (positive, negative, one, two, typeidentifiable) and response time as dependent variable. This analysis showed a significant effect of type $(F[4,108]=$ 187.01, $p<0.001, \eta^{2}=0.87$ ). Planned contrast showed that response times to type-identifiable objects $(M=1.6, S D=$


Figure 6: The stimulus presentation in the comprehension experiment. The target in the left upper corner can be described with "the marker without a cap" or "the large yellow marker".

Table 1: Response times, standard deviations and confidence intervals for the five different referring expressions in the comprehension experiment

| Type | RT (s) | SD (s) | 95\% CI |
| :--- | :--- | :--- | :--- |
| Positive property | 2.39 | 0.26 | $2.26-2.46$ |
| Negative property | 2.31 | 0.28 | $2.18-2.40$ |
| Two properties | 1.91 | 0.26 | $1.81-2.01$ |
| One property | 1.72 | 0.24 | $1.63-1.82$ |
| Type identifiable | 1.60 | 0.18 | $1.51-1.66$ |

$0.18)$ were faster than responses to descriptions with one preferred property $(M=1.72, S D=0.26): F[1,27]=38.46, p$ $<0.001, r=.76$, which in turn were faster than responses to descriptions with two preferred properties $(M=1.72, S D$ $=0.26) ; F[1,27]=117.81, p<0.001, r=.90$. These were faster than responses to descriptions with negative ( $M=2.31$ $S D=0.28$ ) or positive properties $(M=2.39, S D=0.26)$ ); $F$ $[1,27]=634.2, p<0.001, r=.98)$, which did not significantly differ from each other $(F[1,27]=2.25, p=0.15, r=$ .27).

## Discussion

In Experiment 2 participants identified objects based on to five types of referring expressions; these either contained a negative property ("the marker without a cap"), its positive counterpart ("the marker with a cap"), one preferred property ("the yellow marker"), two preferred properties ("the large yellow marker") or no properties at all (type-identifiable stimuli such as "the strawberry"). The results showed that listeners' response times closely follow the preferences of speakers. Our participants responded fastest to descriptions where using the targets type was sufficient for unique identification. They took (slightly) longer to respond to descriptions that contained preferred properties such as color or size, and it
took them significantly longer to identify targets that needed two properties to be uniquely identified instead of one. Not surprisingly, the response times were longest for the positive and negative properties that were not as preferred as size or color (having or lacking a cap, a lid, or a handle). Crucially, selecting the appropriate target that was described with a referring expression that contained negative properties did not take longer than selecting a target that was described with a positive property. This is in line with the observation from the production experiment, where speakers used the negative description ("the marker without a cap") as much as the the positive description ("the marker with a cap").

While we carefully controlled for the placement and size of the targets and their properties, the length and complexity of the descriptions was not the same for all descriptions. Descriptions that needed two properties contained more words than descriptions that needed only one property. However, the crucial comparison between descriptions with a positive or negative property differed in only one syllable ("with" versus "without" ${ }^{2}$ ). In addition, we measured response times from the start of the utterance, so our listeners could have already identified the target before the end of the referring expression, but see Arts (2004) for a discussion of measurements onsets in processing referential expressions. Although these issues might be difficult to control (referential expressions containing a negative property are inherently more complex than those with only one property), we do plan to take the length of the utterance into account in future work.

## Conclusion

Taken together, the production and comprehension experiment provide experiential evidence for the use of negative properties in referring expressions. Speakers easily produce expressions such as "the marker without a cap" and listeners are not particularly troubled by them. These findings contribute to our understanding of speech production and perception processes in general, and provide important data for extending the scope of REG algorithms to descriptions containing negative attributes.

## Acknowledgments

Many thanks to Carla van Gruijthuijsen for lending her voice to the stimulus material and to Sandra Augustin for her assistance in conducting the experiments. We gracefully acknowledge the financial support we received from The Netherlands Organization for Scientific Research (NWO, Vici grant 277-70-007 awarded to Emiel Krahmer).

## References

Areces, C., Koller, A., \& Striegnitz, K. (2008). Referring expressions as formulas of description logic. In Proceedings of the fifth international natural language generation con-

[^286]ference (pp. 42-49). Stroudsburg, PA, USA: Association for Computational Linguistics.
Arnold, J. (2008). Reference production: Production-internal and addressee-oriented processes. Language and Cognitive Processes, 23(4), 495-527.
Arts, A. (2004). Overspecification in instructive texts. Tilburg University: Unpublished PhD thesis.
Croitoru, M., \& van Deemter, K. (2007). A conceptual graph approach to the generation of referring expressions. In Proceedings of the 20th international joint conference on artificial intelligence (ijcai) (p. 2456-2461).
Dale, R., \& Reiter, E. (1995). Computational interpretations of the gricean maxims in the generation of referring expressions. Cognitive Science, 19(2), 233-263.
Dale, R., \& Viethen, J. (2010). Empirical methods in natural language generation. In E. Krahmer \& M. Theune (Eds.), (pp. 163-179). Springer.
Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. Journal of Experimental Psychology, 47, 381-391.
Freeman, J. B., \& Ambady, N. (2010). Mousetracker: software for studying real-time mental processing using a computer mouse-tracking method. Behavioral Research Methods, 42(1), 226-241.
Goudbeek, M., \& Krahmer, E. (2012). Alignment in interactive reference production: Content planning, modifier ordering and referential overspecification. Topics in Cognitive Science, 4, 269-289.
Horton, W. S., \& Keysar, B. (1996). When do speakers take into account common ground? Cognition, 59, 91-117.
Koolen, R., Goudbeek, M., \& Krahmer, E. (in press). The effect of scene variation on the redundant use of color in definite reference. Cognitive Science.
Krahmer, E., \& van Deemter, K. (2012, March). Computational generation of referring expressions: A survey. Computational Linguistics, 38(1), 173-218.
Mathôt, S., Schreij, D., \& Theeuwes, J. (2012). Opensesame: An open-source, graphical experiment builder for the social sciences. Behavioral Research Methods, 44, 314324.
Pechmann, T. (1989). Incremental speech production and referential overspecification. Linguistics, 27, 89-110.
van Deemter, K. (2002). Generating referring expressions: Boolean extensions of the incremental algorithm. Сотриtational Linguistics, 28(1), 37-52.
van Deemter, K., \& Krahmer, E. (2006). Graphs and Booleans: On the generation of referring expressions. In H. Bunt \& R. Muskens (Eds.), Computing meaning (pp. 17-53). Dordrecht, The Netherlands: Kluwer.
Viethen, J., \& Dale, R. (2006, July). Algorithms for generating referring expressions: Do they do what people do? In Proceedings of the 4th international conference on natural language generation (pp. 63-70). Sydney, Australia.

# Lay biology in health: How adults conceptualize the benefits of exercise 

Sarah Gripshover<br>sarahjg@stanford.edu<br>Department of Psychology<br>Stanford University

Ellen M. Markman<br>markman@stanford.edu<br>Department of Psychology<br>Stanford University


#### Abstract

We present an approach to health interventions based on the insight that throughout life, individuals construct intuitive theories to predict, explain, and determine how to act on the world. We propose that an intuitive theory-based approach may be useful for teaching adults about the body's biochemical response to exercise. The first step in using this approach is to document the theories that a target population brings to bear on a health domain. We therefore present what is to our knowledge the first investigation of adults' reasoning about exercise. Specifically, we explore how adults explain why exercise confers various benefits, how plausible adults find the physical, emotional, and cognitive benefits of exercise, and how contingent on weight loss adults believe such benefits to be. These findings lay the groundwork for constructing an intervention to motivate adults to exercise more.


Keywords: health, intuitive theories, folkbiology

## Introduction

Throughout life, individuals construct and use coherent belief systems, or intuitive theories, to predict, explain, and determine how to act upon the world (Carey, 1987, 2009; Murphy \& Medin, 1985; Gopnik \& Wellman, 1994; Au, Romo, \& DeWitt, 1999; R. Gelman, Brenneman, Macdonald, \& Roman, 2009; S. Gelman, 2003). This insight has so far remained largely untapped, however, as a resource for health interventions.

We present an approach to health interventions that leverages participants' intuitive theories, builds on them when appropriate, and corrects misconceptions that would interfere with understanding the health message when necessary. This approach goes beyond most knowledge-based interventions that provide facts, "do's and don'ts',' and admonitions without a coherent explanatory framework (see Au, Romo, \& DeWitt, 1999). The intuitive theory based approach takes stock of participants' existing theories in the relevant domain, considers the specific nature of the health message, and determines what conceptual prerequisites need to be in place for participants to understand the message thoroughly, apply it flexibly, and believe it with conviction. The present work takes stock of participants' existing theories regarding exercise.

Gripshover and Markman (in press) used this intuitive theory based approach to teach young children that foods have different nutritional profiles-no one food provides all the nutrients the body needs-and therefore we need a variety of healthy foods. This explanation relies on several concepts, some readily understandable and others opaque. Preschoolers may find invisible, discrete, heterogeneous nutrients inside homogeneous-looking food puzzling because they expect
substances to be continuous (Au, 1994). Because preschoolers understand solutions (Au, Sidle, \& Rollins, 1993; Rosen \& Rozin, 1993), nutrients were explained by analogy to sugar dissolving in water. Nutrient extraction during digestion was explained in mechanical (not biochemical) terms: food enters the body, the stomach breaks it into smaller pieces, extracts the nutrients, and blood carries them throughout the body. This built on children's emerging mechanical understanding of digestion in which food enters and eventually exits the body, but little is known about the intermediate processes of digestion (Teixeira, 2000; Rowlands, 2004). Labels were used to unite perceptually distinct foods into foodgroup categories and convince children that they share important internal properties (S. Gelman \& Markman, 1986). Finally, while children believe that food enables biological processes (Inagaki \& Hatano, 1996), it was additionally explained that different processes require different nutrients. Children learned and generalized this new theory: over half described the role of blood in transporting food when answering open-ended questions about digestion, almost all claimed there were invisible nutrients inside of food, and nearly half justified their hypothetical snack choices in terms of nutrients or variety. Moreover, children even increased their vegetable consumption during snack time even though eating more vegetables was not explicitly mentioned in the intervention.

In the present research, we lay the groundwork for using such an intuitive theory based approach for motivating adults to exercise by taking stock of adults' theories regarding exercise. New findings in exercise research show profound and broad-ranging benefits of exercise, from increasing insulin sensitivity (Borghouts \& Keizer, 2000), to regulating the immune system and reducing inflammation (Woods, Vieira, \& Keylock, 2009), to warding off depression (Dunn, Trivedi, Kampert, Clark, \& Chambliss, 2005; Salmon, 2001) and even decreasing the risk for dementia among the elderly (Kramer, Erickson, \& Colcombe, 2006; Hillman, Weiss, Hagberg, \& Hatfield, 2002; K. Erickson \& Kramer, 2008). Though research is just beginning to delineate the specific mechanisms behind these benefits, the general explanation seems to be that the body sets in motion a cascade of biochemical processes that are specially adapted to sustain and repair body and brain cells and prepare them for the next physical challenge. This repair process explains why exercise improves such a wide range of health indicators, and even causes the creation of new brain cells. Exercise is more than just physical movement: it is in many ways like taking medicine.

We begin by asking what theories adults draw upon to reason about exercise. Very little research exists to guide us
in characterizing these theories. However, we expect nonexpert adults to explain the benefits of exercise by appealing to causal mechanisms that are more or less mechanical in nature; e.g., the heart works harder and thus propels more blood throughout the body; muscles physically grow larger as they become stronger; calories are used up and thus the amount of fat on the body is reduced. Under this view, health benefits would appear to result from biomechanical factors such as weight loss and increased strength of the muscles and heart.

If adults indeed do hold a predominantly biomechanical theory of exercise, we suspect that this theory will be inadequate to explain many of these recently-discovered benefits of exercise satisfactorily. Mental benefits, such as the reduced risk for dementia and improvements in cognitive ability should be especially problematic, because they cross a well-worn ontological boundary between the biological and psychological domains (see, e.g., Carey, 1987; Wellman \& Gelman, 1992, 1998; Inagaki \& Hatano, 1993; Erickson, Keil, \& Lockhart, 2010; Lynch \& Medin, 2006). For example, a normal-weight, or even thin, elderly person who is concerned about memory loss might be more likely to seek out crossword puzzles than physical exercise. An overweight person who embarks on an exercise program and fails to lose a significant amount of weight despite hard work may despair of receiving any health benefits. The present research seeks to document the theories that adults bring to bear on explaining the benefits of exercise as a first step towards creating an intuitive theory based intervention designed to communicate the newfound benefits of exercise to adults.

## Experiment 1

Experiment 1 explores 3 questions: (1) To what extent do adults evoke biomechanical, biochemical, and other causal mechanisms to explain the effects of exercise on the body? (2) How plausible do adults find physical and mental effects of exercise? (3) Do adults view physical benefits of exercise as contingent on weight loss?

## Methods

Participants Participants were 50 adults ( 27 women) from the United States recruited through Amazon's Mechanical Turk (AMT). Ages ranged from 18-25 years $(N=9)$ to $60-$ 69 years $(N=2)$.

Survey A survey was designed using Qualtrics. Question blocks were presented in a fixed order (i.e., plausibility ratings, explanation of benefits, importance of weight loss, and explanation of benefits without weight loss), but items were always randomized within blocks.
Plausibility Ratings Participants were given a list of claims about the effects of exercise and asked to rate their agreement on a scale of $0-100$ with their mouse using a slider. The endpoints of the scale were labeled "totally disagree" and "totally agree," and the midpoint was labeled "neither agree nor disagree." Items included 2 physical effects (improved blood pressure, increased heart rate), 2 emotional
effects (reduced depression risk, improved mood), and 2 cognitive effects (reduced risk for memory loss, improved problem-solving ability). Finally, half of the effects were framed as short-term (immediately after exercise, people experience increased heart rate, improved mood, and improved problem-solving ability) and half as longer-term (regular exercise helps improve blood pressure, reduce depression risk, and reduce risk for memory decline).

Explanation of benefits Participants were shown claims about the benefits of exercise and asked questions framed as, "Research shows that getting regular physical exercise can [claim about benefit]. If you had to guess, why would you say getting plenty of exercise [claim about benefit]?" Participants were asked to assume the claims were true and provide the best explanations they could. Items included 2 physical (reduced blood pressure / blood sugar), 2 emotional (reduced depression risk, reduced stress), and 3 cognitive benefits (improved memory, improved problem-solving ability, and reduced risk for dementia).

Importance of weight loss Seven questions asked whether participants saw value in exercise apart from weight loss. Participants rated how much they agree or disagree with 7 statements such as, "If someone exercises regularly for a long time and never loses any weight, they are probably wasting their time" on a scale of 1-100 using a sliding scale.

## Results and Discussion

Results did not differ by age and gender, so results are collapsed across these categories.

Plausibility Ratings Figure 1 presents mean agreement ratings for each effect of exercise. All ratings were significantly greater than the "neither agree nor disagree" midpoint (all $t$ 's > 3.2), demonstrating that all effects were more plausible than not. Ratings were arcsine transformed to ensure homogeneity of variance and compared to one another using a repeated-measures ANOVA with within-subjects factors of item type (physical, emotional, and cognitive) and time scale (long vs. short-term). Results revealed a main effect of item type $(F[2,48]=50.53, \mathrm{p}<.0001$ and a significant item type by time scale interaction $(F[2,48]=11.28, \mathrm{p}<.0001)$. Posthoc paired $t$-tests showed that emotional effects were less plausible than physical effects (physical $M=90.8 \%$ agreement rating, emotional $M=80.9 \%$ agreement rating, $t(49)=$ $6.09, \mathrm{p}<.0001$ ), and cognitive effects ( $M=66.5 \%$ agreement rating) were less plausible than emotional $(t[49]=6.14$, $p<.0001)$ and physical $(t(49)=9.75, p<.0001)$ effects. No differences were observed between long - and short-term items within physical and emotional effects ( $t$ 's $<1.8$ ), but improving immediate problem-solving ability ( $M=60.1 \%$ agreement rating) was less plausible than reducing risk for dementia ( $M=72.8 \%$ agreement rating).

Explanations of benefits Tables 1-3 present summaries of participants' open-ended explanations. Responses were


Figure 1: Mean plausibility ratings by items. Error bars represent standard error of the mean.
coded into one or more categories.

Physical benefits Explanations for physical benefits included biomechanical mechanisms such as circulation (improved/increased blood flow), "burning off" sugar/calories, clearing debris from blood vessels, strengthening the heart and/or blood vessels, and reducing weight; non-biological explanations (e.g., exercising helps people develop better health-related habits, self-discipline, confidence, etc.); biochemical explanations; and ambiguous biological explanations such as increasing metabolic rate. Table 1 presents the proportion of participants whose explanations featured each category. Biomechanical explanations were by far the most common.

| Items (2) | blood <br> pressure | blood <br> sugar |
| ---: | :--- | :--- |
| burns sugar / calories | 0.00 | 0.52 |
| weight | 0.24 | 0.26 |
| circulation | 0.36 | 0.08 |
| stronger heart | 0.34 | 0.00 |
| clears blood vessels | 0.12 | 0.00 |
| biochemical | 0.00 | 0.12 |
| metabolism | 0.00 | 0.12 |
| non-biological | 0.02 | 0.08 |
| other | 0.34 | 0.30 |
| dk | 0.00 | 0.02 |

Table 1: Proportion of participants appealing to each explanation type for 2 physical benefits of exercise: reducing blood pressure and blood sugar

Emotional benefits Table 2 presents a summary of explanations for the effect of exercise on stress and depression. A handful of biomechanical explanations were observed, such as improved circulation ( $4 \%$ and $2 \%$ for depression and stress) and the release of stress, tension, or negative energy ( $38 \%$ for stress only). However, explanations for stress and depression most commonly appealed to biochemical explanations ( $52 \%$ and $30 \%$ for depression and stress) and non-biological explanations ( $52 \%$ and $30 \%$ for depression and stress).

| Items (2) | depression | stress |
| ---: | ---: | ---: |
| non-biological | 0.52 | 0.32 |
| biochemical all | 0.50 | 0.30 |
| release | 0.00 | 0.38 |
| relax tire out | 0.00 | 0.12 |
| circulation | 0.04 | 0.02 |
| other | 0.12 | 0.12 |
| dk | 0.00 | 0.00 |

Table 2: Proportion of participants appealing to each explanation type for 2 emotional benefits of exercise: preventing depression and reducing stress

Cognitive benefits Table 3 presents a summary of explanations for cognitive benefits of exercise. As with physical benefits, participants appealed predominantly to to biomechanical mechanisms such as circulation (improved blood or oxygen flow to brain and other tissues) and the brain itself getting a "workout" during physical exercise. Participants also claimed that exercise increases feelings of awakeness and improves overall health, and also offered non-biological explanations (e.g., exercise increases social activity, mental discipline, willpower, etc.), and a few biochemical explanations.

| Items (3) | dementia | memory | problem solving |
| ---: | ---: | ---: | ---: |
| circulation | 0.52 | 0.62 | 0.40 |
| nonbio | 0.18 | 0.16 | 0.20 |
| alert clear head | 0.00 | 0.14 | 0.22 |
| physical health | 0.18 | 0.10 | 0.08 |
| exercises brain | 0.14 | 0.12 | 0.06 |
| other | 0.10 | 0.10 | 0.08 |
| biochemical | 0.10 | 0.10 | 0.04 |
| dk | 0.06 | 0.00 | 0.02 |

Table 3: Proportion of participants appealing to each explanation type for 3 cognitive benefits of exercise: preventing dementia, improving memory, and improving problem-solving ability

Biochemical explanations Table 4 presents a summary of biochemical explanations across items. While participants mentioned a wide variety of biochemicals, the most commonly-mentioned was endorphins.

Summary of explanations of benefits Participants evoked primarily biomechanical explanations for physical and cognitive benefits of exercise, and primarily biochemical and nonbiological explanations for emotional benefits. This suggests that although many participants are aware of some biochemical processes related to exercise, they do not view the physical, cognitive, and emotional benefits of exercise as consequences of a single cascade of adaptive biochemical responses. Interestingly, biochemical explanations were overwhelmingly offered for emotional benefits, rather than phys-
$\left.\begin{array}{llllll}\hline \begin{array}{l}\text { Benefit } \\ \text { type }\end{array} & \begin{array}{l}\text { Explanation } \\ \text { type }\end{array} & \text { Item } & \begin{array}{l}\text { conditional } \\ \text { \% of participants }\end{array} \\ \hline \text { physical } & \text { insulin } & \text { blood sugar } & 1.00 & \text { (6 of 6 } \\ \text { emotional } & \text { endorphins } & \begin{array}{l}\text { depression } \\ \text { stress }\end{array} & 0.54 & (14 \text { of 26) }\end{array}\right]$

Table 4: Breakdown of biochemical explanations by benefit type (cognitive, emotional, and physical) and item (preventing depression, reducing stress, etc.). Presented as a conditional proportion of total biochemical explanations given for each item.
ical or cognitive ones, and the vast majority cited endorphins as the causal mechanism. We therefore suspect many of these biochemical explanations may reflect awareness of the wellpublicized "runners' high."

Importance of weight loss Reliability among the 7 questions about the importance of weight loss was high (Cronbach's alpha $=.93$ ) and so a composite importance-of-weight score was created, $M=23.2$ out of $100, s d=14.2$. This score was significantly lower than the "neither agree nor disagree" midpoint of the scale, $t(49)=13.34, p<.0001$, indicating that participants do not view exercise as beneficial solely in the presence of weight loss.

## Biochemical explanations in relation to other measures

To see whether participants who appealed to biochemical causal mechanisms viewed weight loss as less important, we divided participants into two groups: one who never offered biochemical explanations ( $N=18$ ), and one who did ( $N=$ 30). Participants who offered a biochemical explanation had lower importance-of-weight scores $(M=19.1)$ than participants who $\operatorname{did} \operatorname{not}(M=29.5, t[46]=2.56, p=.02)$. We also asked whether participants who offered at least one biochemical explanation rated cognitive and emotional benefits
of exercise as more plausible than participants who offered none, and found that they did not, $t$ 's $<1.3$.

## Discussion

Experiment 1 provided preliminary evidence that adults find the cognitive effects of exercise are less plausible than emotional effects, and emotional effects less plausible than physical effects. Furthermore, we provided preliminary evidence that adults appeal predominantly to biomechanical explanations for explaining the physical and cognitive benefits of exercise, but not the emotional benefits. Instead, adults frequently gave biochemical and non-biological, rather than biomechanical, explanations for the emotional benefits of exercise. However, we suspect these biochemical explanations represent isolated facts, rather than a coherent biochemical theory of the body's response to exercise. Nevertheless, awareness of at least some biochemical effects of exercise in the could be used as leverage for teaching a more coherent biochemical theory.

Finally, although participants were reluctant to endorse statements that blatantly identify weight loss as the primary source of exercise-related health benefits, participants who gave at least one biochemical explanation viewed weight loss as less of a critical factor than participants who identified none. However, it is possible that although adults concede that exercise can offer benefits aside from weight loss, they may still view achieving a healthy weight as important gauge of success if given more subtle questions. Experiment 2 tests this possibility.

## Experiment 2

Experiment 2 explores more closely the importance adults place on maintaining a healthy weight when gauging exercise-related health benefits.

## Methods

Participants Participants were 47 adults ( 23 women) from the United States recruited via AMT. Ages ranged from 18-29 ( $N=23$ ) to 50-59 $(N=4)$.

Survey A survey was designed using Qualtrics. Blocks were presented in a fixed order (i.e., open-ended health behaviors, thinness vs. fitness, weight loss vs. no weight loss), but items within blocks were randomized.

Open-ended health behaviors Participants were asked to list what they believe are the best ways to stay healthy. We were interested in whether participants would be more likely to list maintaining a healthy weight than exercise on its own.

Thinness vs. fitness Participants indicated which of two hypothetical people would be healthier: someone with a healthy weight who rarely exercises and has low physical fitness, or someone who is overweight, exercises frequently, and has high physical fitness.

Exercise vignettes: weight loss vs. no weight loss Participants viewed two vignettes about characters who are overweight and begin fitness programs. Both work out 4-5 times per week for $30-40$ minutes each time for a year. However, one character achieves a weight near the healthy range, and the other character only loses a few pounds and remains overweight. Participants estimated the percent likelihood that each character received a set of physical, cognitive, and emotional benefits. Order was randomized across participants.

## Results and Discussion

Results did not differ by age and gender, so results are collapsed across these categories.

Open-ended health behaviors Table 4 presents behaviors participants identified as important for health. Nearly all participants identified exercise and diet. No participants identified maintaining a healthy weight.

|  | \% of participants |
| ---: | ---: |
| healthy diet | 1.00 |
| exercise | 0.96 |
| weight loss | 0.00 |

Thinness vs. fitness $15 \%$ of participants said that a person who is lean and not physically fit is healthier than someone who is fit and overweight. $23 \%$ of participants said both are equally healthy, and $61 \%$ said that an overweight, fit person is healthier, $\chi^{2}(2)=16.2, p=.0003$.

Exercise vignettes: weight loss vs. no weight loss Reliability across likelihood ratings for physical, cognitive, and emotional benefits was high (alpha $=.93$ for no-weight-loss vignettes and .92 for weight-loss vignettes) mean composite score for each vignette was computed for each participant across items. Participants rated benefits as significantly more likely in the presence of weight loss (no-weight-loss vignettes $M=56 \%$ likely, weight-loss vignettes $M=72 \%$ likely, $t[46]$ $=6.8, p<.0001$ ).

Because the vignettes were identical except for weight loss, it was possible that participants used weight explicitly to differentiate their ratings. We therefore performed a betweensubjects comparison using only each participant's first vignette. Figure 2 presents these ratings. Again, participants rated benefits as more likely when they occurred in the presence of weight loss ( $M=76 \%$ likely) than in the absence of weight loss ( $M=65 \%$ likely), $t(45)=2.4, p=.02$.

This suggests that adults do at least to some extent view weight loss as an indication of progress in an exercise program. This was true for nearly every benefit we presented, including physical and emotional benefits.

## Discussion

Participants identified diet and exercise as the best ways to improve health nearly $100 \%$ of the time, and they recognized


Figure 2: Ratings of percent likelihood that the character in each vignette experienced each benefit. Considering only each participant's first vignette
that a fit, overweight person is likely to be healthier than a normal-weight, unfit person. On the other hand, participants rated nearly all of the benefits of exercise that we presented as less likely to occur for a character who loses very little weight. It is possible that participants found it implausible that a character could truly comply with an exercise program for a year and not lose weight. Exercise physiology research has shown, however, that some individuals do comply rigorously with an exercise program but lose very little weight, and yet still reap health benefits similar to those who lose more weight (King, Hopkins, Caudwell, Stubbs, \& Blundell, 2009). The vignettes furthermore gave identical descriptions of both exercise regimens. Participants still may have assumed the non-weight-loss character was not working out as intensely, however. Further research is needed to determine whether adults view weight loss as a mechanism behind certain health benefits, or as diagnostic of adequate effort.

## General Discussion

These studies to our knowledge provide the first systematic exploration of how adults bring their biological theories to bear on the topic of exercise. Although many participants were aware of some biochemical effects of exercise, we did not find evidence of participants holding coherent theory of exercise triggering a cascade of beneficial chemical responses in the body.

Although adults do not view weight loss as more important than exercise, we found that they do view many benefits as less likely to occur without weight loss. A biochemical theory of exercise may therefore be useful to de-emphasize weight loss as a gauge or mechanism for health benefits.

We also identified one area in which biochemical effects of exercise are already well-known: the emotional benefits of exercise. Widespread knowledge of the "runner's high" can therefore be leveraged to teach a more general biochemical theory of exercise.

Finally, we identified one way in which a biochemical theory of exercise may fail to support a coherent understanding of the wide-ranging benefits of exercise. Adults view cog-
nitive benefits of exercise as less plausible than physical and even emotional benefits. They also explain cognitive benefits primarily by appeal to biomechanical explanations-few participants offered biochemical accounts. This raises the interesting possibility that a lay biochemical theory of exercise does not support a coherent, compelling explanation for the cognitive benefits of physical exercise. This may have implications for motivating older adults to exercise. Maintaining cognitive acuity may be much more important to older adults than more well-known aims of exercise such as building strength and improving physical appearance. At the same time, these cognitive benefits may be the most difficult to understand in light of adults' lay theories. An intervention that explains the cognitive benefits of exercise may therefore be especially useful if targeted towards older adults.

Future research will test whether teaching adults a biochemical theory of exercise increases their conviction about the cognitive benefits of exercise, diminishes focus on weight as an indicator of success, and actually motivate adults to exercise more. The present research documents adults' lay theories of exercise and establishes the potential utility of an intuitive theory based approach to motivating adults to exercise.

## References

Au, T. K. (1994). Developing an Intuitive Understanding of Substance Kinds. Cognitive Psychology, 27(1), 71-111.
Au, T. K., Romo, L. F., \& DeWitt, J. E. (1999). Considering Children's Folkbiology in Health Education. In M. Siegal \& C. C. Peterson (Eds.), Children's understanding of biology and health (pp. 209-234). New York, NY: Cambridge University Press.
Au, T. K., Sidle, A. L., \& Rollins, K. B. (1993). Developing an intuitive understanding of conservation and contamination: Invisible particles as a plausible mechanism. Developmental Psychology, 29(2), 286-299.
Borghouts, L. B., \& Keizer, H. a. (2000). Exercise and insulin sensitivity: a review. International journal of sports medicine, 21(1), 1-12.
Carey, S. (1987). Conceptual Change in Childhood (MIT Press ed.). Cambridge, MA.
Carey, S. (2009). The Origin of Concepts. New York, NY: Oxford University Press.
Dunn, A. L., Trivedi, M. H., Kampert, J. B., Clark, C. G., \& Chambliss, H. O. (2005). Exercise treatment for depression: efficacy and dose response. American journal of preventive medicine, 28(1), 1-8.
Erickson, J., Keil, F. C., \& Lockhart, K. L. (2010). Sensing the coherence of biology in contrast to psychology: young children's use of causal relations to distinguish two foundational domains. Child development, 81(1), 390-409.
Erickson, K., \& Kramer, A. F. (2008). Aerobic Exercise Effects on Cognitive and Neural Plasticity in Older Adults. British journal of sports medicine, 43(1), 22-4.
Gelman, R., Brenneman, K., Macdonald, G., \& Roman, M.
(2009). Preschool Pathways to Science: Facilitating Scientific Ways of Thinking, Talking, Doing, and Understanding. Baltimore, MD: Brookes Publishing Company.
Gelman, S. (2003). The Essential Child: Origins of Essentialism in Everyday Thought. New York: Oxford University Press.
Gelman, S., \& Markman, E. M. (1986). Categories and induction in young children. Cognition, 23(3), 183-209.
Gopnik, A., \& Wellman, H. M. (1994). The Theory Theory. In S. Gelman \& L. A. Hirschfeld (Eds.), Mapping the mind (pp. 257-293). Cambridge: Cambridge University Press.
Gripshover, S., \& Markman, E. M. (in press). Teaching Young Children a Theory of Nutrition: Conceptual Change and Potential for Increased Vegetable Consumption. Psychological Science.
Hillman, C. H., Weiss, E. P., Hagberg, J. M., \& Hatfield, B. D. (2002). The relationship of Age and Cardiovascular Fitness to Cognitive and Motor Processes. Psychophysiology, 39(3), 303-12.
Inagaki, K., \& Hatano, G. (1993). Young Children's Understanding of the Mind-Body Distinction. Child Development, 64(5), 1534.
Inagaki, K., \& Hatano, G. (1996). Young Children's Recognition of Commonalities between Animals and Plants. Child Development, 67(6), 2823-40.
King, N. A., Hopkins, M., Caudwell, P., Stubbs, R. J., \& Blundell, J. E. (2009). Beneficial effects of exercise: shifting the focus from body weight to other markers of health. British Journal of Sports Medicine, 43(12), 924-7.
Kramer, A. F., Erickson, K., \& Colcombe, S. J. (2006). Exercise, Cognition, and the Aging Brain. Journal of Applied Physiology, 101(4), 1237-42.
Lynch, E., \& Medin, D. (2006). Explanatory models of illness: a study of within-culture variation. Cognitive psychology, 53(4), 285-309.
Murphy, G. L., \& Medin, D. L. (1985). The Role of Theories in Conceptual Coherence. Psychological Review, 92(3), 289-316.
Rosen, A. B., \& Rozin, P. (1993). Now you see it, now you don't: The preschool child's conception of invisible particles in the context of dissolving. Developmental Psychology, 29(2), 300-311.
Rowlands, M. (2004). What do children think happens to the food they eat? Journal of Biological Education, 38(4), 167-172.
Salmon, P. (2001). Effects of physical exercise on anxiety, depression, and sensitivity to stress: a unifying theory. Clinical psychology review, 21(1), 33-61.
Teixeira, F. M. (2000). What happens to the food we eat? Children's conceptions of the structure and function of the digestive system. International Journal of Science Education, 22(5), 507-520.
Woods, J. a., Vieira, V. J., \& Keylock, K. T. (2009). Exercise, inflammation, and innate immunity. Immunology and allergy clinics of North America, 29(2), 381-93.

# Bayesian Adaptive Estimation of Psychometric Slope and Threshold with Differential Evolution 

Hairong Gu, Jay I. Myung, Mark A. Pitt, and Zhong-Lin Lu<br>\{Gu.124, Myung.1, Pitt.2, Lu.535\}@Osu.Edu<br>Department of Psychology, Ohio State University<br>1835 Neil Avenue, Columbus, OH 43210 USA


#### Abstract

The adaptive experimentation methodology has been adopted in visual psychophysical modeling in the pursuit of efficiency in experimental time and cost. The standard scheme only optimizes one design in each experimental stage, although simultaneous optimization of multiple designs per stage can be beneficial, but difficult to implement because of a surge in computation. In this study, we incorporated the adaptive experimentation methodology under a Bayesian framework with differential evolution (DE), an algorithm specialized in multi-dimensional optimization problems to explore the multiple-designs-per-stage approach. By taking advantage of parallel computing, DE is computationally fast. The results showed that the multiple-designs-per-stage scheme resulted in a more stable estimation in the early stages of the parameter estimation.


Keywords: Visual psychophysics, Bayesian inference, adaptive estimation, evolutionary computing

## Not All Designs are Equally Informative

Experimental design is a critical step in carrying out effective experiments. Traditionally, the practice of experimental design is guided by heuristic norms, using a one-shot design, chosen at the outset, throughout the course of the experiment. Although this approach may be adequate in some scientific quests, its shortcomings are obvious. First, not all experimental designs are equally informative. The traditional approach does not guarantee that the design, including the number of treatments, the values of treatments, and the number of participants in each treatment, is an optimal choice. A non-optimal design may contribute little to the goal of the experiment. Further, the most informative designs may change as the experiment progresses with more responses being observed. Thus, a one-shot design ignores utilizing what can be learned during the course of an experiment.

Second, the traditional experimental design method typically relies on increasing the number of participants or the number of measurements to increase the power of statistical inference. Obviously, this increases the experimental cost, which would matter for experiments that use expensive technology such as fMRI, or research whose target population is difficult to recruit (children, senior citizens, mentally disordered).

Third, the traditional methods of experimental design center on randomization, reduction of variation, blocking etc., with the purpose of revealing the group or treatment
effects while ignoring the individual variation. However, more and more recognition has been given to the importance of individual differences. For example, in drug development, it is important to know how different people react differently to the same drug to guide the prescription. Thus, experimental designs should not be identical for every participant.

To illustrate how experimental designs can be unequally informative, suppose that a researcher is interested in studying how the rate of detection changes with the brightness of a stimulus. A psychometric function is used to describe the probability $p$ of detecting a stimulus of certain brightness $x$. A simplified example assumes a sigmoid function $p=1 /(1+\exp (-x+t))$, where $x$ is the design variable representing the brightness and $t$ is the parameter, threshold, a characteristic associated with a particular individual, reflected in the shift of the model in the design dimension. Suppose that there are only 5 possible values of $t$. The corresponding predictions are depicted as the five lines in Figure 1. The red line represents a particular subject's true $t$ value and the other four blue lines are from the wrong $t$ values. The researcher conducts an experiment to estimate the threshold value of that subject by presenting two designs with intensity D1 and D2. Visualization of the model suggests that D1 is a good design because the predictions from the five $t$ values are very differentiable so that the observation can be informative of the true $t$ value. On the other hand, D2 would be a bad design because the prediction differences are so small that little information about the exact shift of the true model is given.


Figure 1: A sample psychometric function with 5 possible parameter values (see text) with the true value indicated by the red line and the wrong values by the blue lines. A good design D1 offers the most discriminability, whereas D2 is a bad design for a lack of differentiability in prediction.

## Adaptive Experimentation

In practice, we do not possess full knowledge of the approximate values of good designs because the model can be quite complex and the range of parameters can be much larger. In addition, an experiment usually contains multiple trials, so the best designs at the beginning of an experiment may be different from those at the later trials of the experiment. Therefore, an efficient experimentation should adaptively identify the best design for the current trial based on the responses already collected from the participant. Facing these challenges for a better experimental design regime, a statistical methodology, dubbed adaptive design optimization (ADO, Cavagnaro et al., 2010; Myung et al., 2012) under a Bayesian framework has been developed to meet these needs.

The general framework of ADO is illustrated in Figure 2. The traditional experimentation starts from a particular experimental design, with which data are collected, and it stops at cognitive modeling where data are fit to a proposed model to make statistical inferences. In contrast, in ADO, the inference from cognitive modeling continues to influence the choice the designs for the next experimental stage. To put it in another way, the whole experiment is divided into multiple stages, and in each stage, the design is based on what is learned from the data collected in the previous stages. By doing that, every selected design is the most imminently useful one for the immediate trial. As such, ADO is efficient in a way that it reduces the time, cost of experiments and the number of participants without sacrificing the quality of the statistical inferences.


Figure 2: Schematic illustration of ADO paradigm.
There are other desirable features of ADO that make it more attractive to the traditional experimental methods. It is found that bad designs not only increase the cost of experiments, but also deteriorate the quality of data so as to hurt the final inference. ADO adopts an information theoretic computational algorithm to ensure the quality of the selected designs so that the risk of having bad designs is minimized. Additionally, ADO is able to reveal individual differences in response strategy or characteristics because the designs are tailored based on the subject's responses in each experiment. Classification of participants can also be done after individuals' properties are estimated.

Because of its efficiency and versatility, ADO has found its usage in various disciplines. It has been used for designing electrophysiology experiments in neuroscience (Lewi et al., 2008), drug dosage assignment in clinical drug development (Miller, et al. 2007), etc. In psychology, it has
been implemented in the discrimination of retention models in simulations (Cavagnaro et al., 2010) and human experiments (Cavagnaro et al., 2011).
A promising application of ADO is in psychophysical experiments with potential clinical applications that put high stake on the reliability of the results and usually have tight time restraint on the experiments. In this area, the previous studies have only optimized one design in each experimental stage. The difficulty of exploring a different scheme, multiple designs per stage, lies in a lack of a smarter algorithm and the increase in computation.
In this paper, we explore ways to improve upon the current efficiency of ADO by implementing the multiple-designs-per-stage scheme that is solved with an evolutionary computation algorithm known as differential evolution (DE). In what follows, we begin with a brief introduction of the ADO methodology. We will then review past studies in adaptive experimentation of visual psychophysics, followed by a discussion of the motivation and application of the multiple-designs-per-stage scheme and DE. Finally, we present and discuss results from ADO simulations.

## How ADO Works

In this section, we provide some technical details of ADO. Readers who prefer to skip technicalities may bypass this section. Figure 3 is a schematic illustration of the steps involved in ADO. First, the application of ADO requires that the model should be formulated as a statistical model defined as a parametric family of probability distributions, $p(y \mid \theta, d)$ 's, which specifies the probability of observing an experimental outcome $y$ given a parameter value $\theta$ and a design $d$. As mentioned before, ADO is a circulating process going through design optimization (DO), experiments and cognitive modeling. In each round, the process starts with the assumed or learnt probability distribution of the parameters, the prior distribution $p(\theta)$. Next, in the step of DO, the optimal design $d^{*}$ is selected from a design set $D$ by the principle of maximum utility.
In DO, a utility function $U(d)$ is pre-defined to quantify the usefulness of a design $d \in D$ for the purpose of the experiment. For parameter estimation, the utility $U(d)$ of each design $d$ is the expectation of the local utility $u(d, \theta, y)$ taken over the parameter space and the outcome's sample space, formally written as

$$
\begin{align*}
d^{*} & =\underset{d \in D}{\operatorname{argmax}}(U(d)) \\
& =\underset{d \in D}{\operatorname{argmax}} \iint_{y, \theta} u(d, \theta, y) p(y \mid \theta, d) p(\theta) d y d \theta, \tag{1}
\end{align*}
$$

where $u(d, \theta, y)$ is defined on a set of particular design $d$, parameter value $\theta$ and observation $y$. The goal of parameter estimation is to obtain accurate estimation of the true parameter values with the smallest number of experimental trials. Functionally, an appropriate utility quantifies the usefulness of designs in reducing the variation of the parameter estimates. Or in the language of information theory, a utility amounts to the information gain or the
uncertainty reduction of the unknown parameters after observations are collected. One formulation of utility that directly quantifies the information gain of the parameter $\Theta$ with the observation $Y$ is Mutual Information $I\left(Y_{d} ; \Theta\right)$. According to the property of mutual information, the utility $U(d)$ can be written as

$$
\begin{aligned}
U(d) & =I\left(\mathrm{Y}_{d} ; \Theta\right) \\
& =\iint\left(\log \frac{P\left(\mathrm{y}_{d} \mid \theta\right)}{P\left(y_{d}\right)}\right) p\left(y_{d} \mid \theta\right) p(\theta) d \theta d y_{d}
\end{aligned}
$$

In which $\log \frac{P\left(y_{d} \mid \theta\right)}{P\left(y_{d}\right)}$ corresponds to the local utility $u(d, \theta, y)$ in Equation (1).

Two general methods have been used in ADO to solve the multiple integral problem of Equation (1), grid search and sequential Monte Carlo (SMC). In grid search, the design space is discretized and grids are the fixed designs on the space. To calculate $U(d)$, one way is to discretize the parameter space also and just replace the integral with summation. Or we can draw a large sample of $(\theta, y)$ from the model's prior and sampling distribution, and then calculate Equation (1) by Monte Carlo approximation. On the other hand, in SMC, solving ADO is recasted as a probability density simulation problem. The utility function is extended to a joint distribution with parameters, observations and designs. By adopting Metropolis-Hasting algorithm and simulated annealing procedure, the marginal distribution of $d$ can be obtained. In this paper, we will present a third method, differential evolution (DE) (Storn \& Price, 1997) as an alternative that is specialized in multidimensional optimization problems.

After DO, the optimal design $d_{s}$ for the current stage will be presented to the participant. The responses until the current stage will be used to update the knowledge of the parameters. Mathematically, we calculate the posterior probability distribution of the parameters by Bayes' rule, $P_{g+1}(\theta \mid y, d)=\frac{p(y \mid d, \theta) p_{g}(\theta)}{p(y \mid d)}$. Then the posterior distribution of the parameters of stage $g$ is treated as the prior distribution of the next stage $g+1$. And the ADO process continues.

## Adaptive Estimation of Psychometric Function

In visual psychophysics, a major interest is to study the relationship between the intensity of visual stimuli and their perception. This relationship is usually modeled by a psychometric function with two parameters, threshold and slope. Accurate estimation of the parameter values on individual level not only provides knowledge of the underlying psychophysical process, but also assists in the diagnosis and classification (Lesmes et al., 2010). A major, practical challenge is that a large number of experimental trials is often needed to accurately estimate the parameters with the finding that different design schemes of fixed patterns produce varying accuracy, precision of parameter estimation and model fit (Wichmann \& Hill, 2001).
Addressing this issue, a variety of adaptive experimental methods have been proposed for efficient parameter estimation while the design dimension was restricted to be one. ADO, as a more general optimization algorithm, is able to handle large scale, non-linear models with multiple design variables. Next, within the framework of ADO, the $\Psi$ method (Kontsevish \& Tyler, 1999) was developed that can easily be generalized to incorporate more than one stimulus. It has been applied to such research as diagnosis of visual deficit (Lesmes et al., 2010).

## Multiple-designs-per-stage Scheme

All the methods mentioned above assume that there is just one design to be optimized and one response to be collected in each adaptive estimation stage. It is worthwhile to explore if there is any benefit when more than one design is optimized simultaneously and executed in each stage, by which $d$ in Equation (1) becomes a vector. Intuitively, a multiple-designs-per-stage approach can be beneficial because multiple responses are collected jointly in one stage, and according to the information theory, the joint entropy or information from a set of random variables is more than or equal to the sum of entropy from individual variables. Therefore, we hypothesize that if multiple responses are collected in one stage, the relationship or synergy of the responses can benefit the modeling process more than the case when the responses are collected one by one.


Figure 3: Schematic illustration of the steps involved in adaptive design optimization (ADO).

One computational challenge in the application and implementation of multiple-designs-per-stage scheme is the curse of dimensionality. Most published studies on parameter estimation with psychometric functions used brute-force grid search, which is to fix a certain number of design points on the design space. Because the dimension of the design space increases with the number of designs per stage, the quantity of grids need to enlarge exponentially to keep a certain resolution, which causes a waste of computing resource because most of the grids are far from the best design and not worth being computed in each stage. As such, it begs for a different algorithm that suits multi-dimensional optimization problems in an accurate and efficient way.

## Differential Evolution Search

DE is an evolutionary computation algorithm to optimize nonlinear and non-differentiable continuous functions by keeping track of, iteratively evolving and updating multiple particles. A brief explanation of the algorithm is as follows. To search the global maximum of a $D$ dimensional space, it keeps track of $N P$ D-dimensional vectors $x_{i, G}(i=1,2, \ldots, N P)$, where $N P$ is the number of particles and $G$ the generation index. At the beginning, the vectors can be randomly selected. Then for each target vector $x_{i}$, a mutant vector $v_{i, G+1}$ for the next stage is generated by $v_{i, G+1}=x_{r 1, G}+F \times\left(x_{r 2, G}-x_{r 3, G}\right)$ where $r 1, r 2$ and $r 3$ are randomly chosen integers from 1 to NP except $i$, and $F$ is a constant factor controlling the contribution of the difference of the two randomly chosen vectors. The next step, crossover, creates a trial vector for each target vector with each element either from the mutant vector $v_{i}$ or the target vector $x_{i}$. Then the cost function values of both the target vector $x_{i}$ and the mutant vector $v_{i}$ are computed. If the mutant vector $v_{i}$ yields a smaller cost, the target vector is set to $v_{i}$. Otherwise, the target vector is retained from the last generation. DE is illustrated in Figure 4 with a simple toy example in which DE was used to search the global maximum of a bimodal distribution.


Figure 4: Illustration of DE algorithm searching for the global maximum of a 2-dimensional bimodal distribution. Initially (left), the particles are randomly selected. At $30^{\text {th }}$ generation (right), they converged to the larger mode.
DE is a natural approach to our problem of optimizing multiple designs per stage simultaneously. Because different particles can be processed independently in one stage, DE can benefit from parallel computing.

## GPU-based Parallel Computing

Although ADO retains the quality of the data with fewer trials, the heavy computation of ADO is still an issue to reckon with, especially in real-time experiments. One solution to speed up the computation lies in parallel computing. Traditionally, computer instructions are stored and processed by a central processing unit (CPU), and executed in a serial manner. On the other hand, parallel computing employs multiple cores on a single chip to perform many independent numerical operations simultaneously. Graphic processing units (GPUs) were originally dedicated to processing graphics. However, in recent years, GPUs are being increasingly popular as a general-purpose parallel computing tool in image processing, data mining, and machine learning.

In our previous work, we have implemented GPU computing to accelerate ADO computing. Compared with CPU-based ADO, GPU-based ADO is around 100 times faster, which substantiates the feasibility of using GPU computing to accelerate the computational speed of ADO computing ( $\mathrm{Gu}, 2012$ ). Given that the DE algorithm is intrinsically parallelizable, GPU computing can be beneficial for accelerating the computation.

In the present work, we implemented DE on graphic processing units (GPUs) to speed up the ADO computation.

## Simulations

ADO-based parameter estimation of the psychophysical model in Kontsevich and Tyler (1999) was simulated with artificial data under the assumption that the data are from a stationary process with no variation of lapses or learning. The data-generating model was defined in the following equations

$$
\begin{aligned}
& Y \sim \operatorname{Binomial}(1, \Psi(x)) \\
& \Psi(x)=\Phi(r(x) / \sqrt{2} ; \mu=0, \sigma=1) \\
& r(x)=10^{\wedge}\left(10^{s}(x-t)\right)
\end{aligned}
$$

in which $\quad \Phi(m ; \mu, \sigma)=\frac{1}{\sqrt{2 \pi}} \int_{-\infty}^{m} \exp \left(-\frac{(c-\mu)^{2}}{2 \sigma^{2}}\right) d c ; \quad Y$ represents the experimental observation; $x, t$ and $s$ are the design variable and the parameters, threshold and slope, transformed in log decimal scale. The range of $x, t$ and $s$ are set to be $(0,3),(0,3)$ and $\left(\log _{10} 0.7, \log _{10} 7\right)$, respectively. The prior distributions of $t$ and $s$ are both uniform. In the simulation, the true values for $t$ and $s$ are set to be 1.5 and $\log _{10} 3.5$ or approximately 0.544 .

Multiple designs are optimized at the same time in one stage by DE algorithm. Computationally, DE is used to search for the global maximum of the defined utility function. For a two-alternative forced choice (2AFC) problem, the response $y$ is either 0 or 1 . So the utility function of an $n$-dimensional space can be written as

$$
U(\tilde{d})=\sum_{\theta} \sum_{y_{i}=0,1} u\left(d_{1} \ldots d_{n}, y_{1} \ldots y_{n}, \theta\right) P\left(y_{1} \ldots y_{n} \mid \theta\right) P(\theta)
$$

in which the parameter space $\theta$ is also discretized so that the integral in Equation (1) becomes a summation. The
local utility $u\left(d_{1} \ldots d_{n}, y_{1} \ldots y_{n}, \theta\right)$ is in the form of mutual information $\log \frac{P\left(y_{1} \ldots \mathrm{y}_{n} \mid \theta\right)}{P\left(y_{1} \ldots y_{n}\right)}$.

First the two-designs-per-stage scheme was implemented. Five two-dimensional particles were generated and shown to be enough for the convergence, which was evaluated by the closeness of the particles at the last generation. Until the $50^{\text {th }}$ generation, the 5 particles are identical up to the second decimal number, indicating that 50 generations are enough for DE to locate the maximum of the utility space. The algorithm was coded in parallel computing with a single GPU card, Tesla C2050 by Nvidia, which contains 448 CUDA cores. A third party library in C++, Arrayfire, is called to access the GPU computing function.

One experiment contains a total of 150 stages or 300 trials. To visualize the effect of parameter estimation, the model predictions based on the prior distribution and the posterior distribution at the last stage is shown in Figure 5. On the left, the model prediction is based on the initial uniform distribution of the two parameters. On the right, the prediction is based on the posterior distribution of the $150^{\text {th }}$ stage of the two parameters. Compared to the initial stage, the range of the likely outcome of the model is much narrowed and concentrated, indicating the convergence of the estimation.


Figure 5: The model predictions based on the prior distribution (left) and the posterior distribution at the $150^{\text {th }}$ stage (right). Darker colors indicate high probabilities.

The joint and marginal posterior distributions of threshold and slope at the end of the experiment are shown in Figure 6. Both the posterior distributions tend to converge to the true values of the parameters. Conforming to the previous studies, the estimation of the threshold is more accurate and has less variation in its posterior distribution while the estimation of slope is less stable.

In each stage, one point estimate is computed for each parameter by calculating the mean of the distribution. 100 experiments of 150 stages were run. Let $\theta_{i}$ be the point estimate in each stage, and $\theta_{\text {true }}$ be the true parameter value, each in log decimal scale. Then we can compute the average bias and standard deviation of the estimation in each stage across the 100 experiments by


$$
S D(\theta)=\sqrt{\frac{\sum_{i=1}^{i=I}\left(\theta_{i}-\theta_{\text {true }}\right)^{2}}{I-1}} \cdot 20 d B
$$




Posterior PDF of threshold $t$

Figure 6: The joint and marginal posterior distributions of threshold and slope at the $300^{\text {th }}$ trial.

To compare the two-designs-per-stage scheme with the traditional one-design-per-stage scheme, we ran 100 experiments of 300 stages with one design in each stage and computed the bias and standard deviation of the estimates in each stage. Figure 7(a) shows the comparison between the two different schemes. In the later trials, the two different schemes do not seem to have significant differences. There is no significant bias at the $300^{\text {th }}$ trial for both threshold and slope. The standard deviation of threshold is about 0.2 dB and that of slope is about 1.1 dB . Although the two-designs-per-stage scheme has less fluctuation in the early stages in the bias of threshold, the difference may result from the random effect.
Next, the five-designs-per-stage scheme was implemented. Because the dimension increases, 200 generations are needed for DE to converge. One hundred experiments of 60 stages ( 300 trials in total still) were run and the point estimates were computed for each stage. Figure 7(b) shows the comparison between the five-designs-per-stage and the one-designs-per-stage schemes. We can see that there is much less fluctuation in the bias of threshold for five-designs-per-stage than that of one-design-per-stage at the early trials, which is consistent with the improvement in the two-designs-per-stage scheme. Other than that, there is no obvious difference between the two schemes.

As expected, simply increasing the number of designs in one stage while still keeping the total number of trials constant resulted in improvement in the accuracy of parameter estimation, at least at the early stages. As we hypothesized, the relationship or synergy provided by multiple responses is greater or at least different than the sum of the information from single responses. We expect that such improvement can be more obvious when it is applied to more complex models because in those cases, more trials are needed for simply exploring the model in
the early stages of an experiment. However, we should not expect that the performance continues to improve as the number of designs per stage increases. By the principle of ADO, a good design should be based on solid information conveyed by the participants' responses. A large number of designs per stage may probe into unfruitful regions of the design space. A balance must be sought in deciding how many designs per stage are good for different models.


Figure 7: The comparison of one design per stage with two designs per stage (a) and five designs per stage (b) in the bias and standard deviation of the estimates of threshold and slope.

## Conclusion

In psychophysical studies, many endeavors have been made to bring further efficiency to the process in parameter estimation. One clear direction is in global optimization or multiple steps ahead to improve the current greedy method that only evaluates the design utilities at the next stage. If global optimization provides the ultimate solution, the approach we studied in this paper, multiple designs per stage, is an initial step in this direction. Thus, in this paper, we sought one eclectic choice between the traditional one-shot experimental design at the very beginning of an experiment and the advanced adaptive experimentation with only one design per stage. The results showed that multiple designs per stage can benefit the estimation in the early stages of an experiment. The reason for the benefit is reminiscent of
holistics in Gestalt psychology and the principle in information theory, with the multiple responses offering extra information than the sum of the individual responses.
To realize the optimization of multiple designs in one stage, we integrated the adaptive design optimization framework with an evolutionary computation algorithm, differential evolution, which is specialized in searching a multi-dimensional space for the purpose of optimization. DE can also be naturally applied to models that contain multiple design variables, for which brute-force grid search is usually applied. DE is less computationally demanding than grid search when the design space is large. Other than that, DE can also benefit from parallel computing to accelerate the computation within each experimental stage.
As such, DE-based adaptive design optimization has large potential of applications in the future experiments for parameter estimation.

## References

Cavagnaro, D. R., Myung, J. I., Pitt, M. A., \& Kujala, J. V. (2010). Adaptive design optimization: A mutual information based approach to model discrimination in cognitive science. Neural Computation, 22(4): 887-905.
Cavagnaro, D.R., Pitt, M.A., \& Myung, J.I. (2011). Model discrimination through adaptive experimentation. Psychonomic Bulletin \& Review, 18(1), 204-210.
Gu, H. (2012). Graphic-Processing-Units Based Adaptive Parameter Estimation of a Visual Psychophysical Model. Master thesis submitted to the Department of Psychology of the Ohio State University.
Kontsevich, L. \& Tyler, C. (1999). Bayesian adaptive estimation of psychometric slope and threshold. Vision Research, 39, 2729-2737.
Lesmes, L.A., Lu, Z., Baek, J., \& Albright, T.D. (2010). Bayesian adaptive estimation of the contrast sensitivity function: the quick CSF method. Journal of Vision, 10(3), 17, 1-21.
Lewi, J., Butera, R., \& Paninski, L. (2009). Sequential optimal design of neurophysiology experiments. Neural Computation, 21, 619-687.
Miller, F., Dette, H., \& Guilbaud, O. (2007). Optimal designs for estimating the interesting part of a doseeffect curve. Journal of Biopharmaceutical Statistics, 17, 6.
Myung, J. I., Cavagnaro, D. R. \& Pitt, M. A. (2012). A tutorial on adaptive design optimization. Manuscript submitted for publication.
Storn, R. \& Price, K. (1997). Differential evolution - a simple and efficient heuristic for glabal optimization over continuous spaces. Journal of Global Optimization 11: 341-359.
Wichmann, F. A. \& Hill, N. J (2001). The psychometric function: I. fitting, sampling and goodness of fit. Perception \& Psychophysics, 63(8),1293-1313.

# Spatial meanings for function words? The link between conjunctions and spatial representations 

Ernesto Guerra (ernesto.guerra@uni-bielefeld.de)<br>Cognitive Interaction Technology Excellence Cluster and Department of Linguistics, Bielefeld University, Morgenbreede 39, 33615, Bielefeld, Germany<br>Tyler Marghetis (tmarghet @ucsd.edu)<br>Department of Cognitive Science, University of California San Diego, 9500 Gilman Drive, La Jolla CA 92093-0515, USA<br>Pia Knoeferle (knoeferl@cit-ec.uni-bielefeld.de)<br>Cognitive Interaction Technology Excellence Cluster and Department of Linguistics, Bielefeld University, Morgenbreede 39, 33615, Bielefeld, Germany


#### Abstract

While formal theories of language consider function words to have little semantic content, more recent theoretical work has argued that even function words have meaning. Yet, there is little experimental work on the representations underlying the meaning of function words such as conjunctions. In two offline experiments, we examined whether conjunctions (and, or, but, either...or) are associated in systematic but distinctive ways with spatial information. In Experiment 1, participants drew schematic representations to depict how two abstract conjuncts might be connected by each of the four conjunctions. These drawing were evaluated on three spatial dimensions (distance, containment and size). In Experiment 2, participants evaluated how well schematic sketches (that differed in distance, containment, and size) represented different conjunctions. In both experiments, spatial information was systematically and distinctively associated with conjunctions. Either... or and or conjunctions were reliably associated with the use of large distance and separation via containment of the conjuncts. And, by contrast, was associated with shorter distance between, and no containment of, the conjuncts. Finally, but was associated with differences in size. We discuss implications of these results for the spatial foundation of linguistic meaning, and the link between lexical semantics and logic.


Keywords: Conjunctions, spatial representation, drawing, rating, simulation, embodiment.

## Introduction

Natural language conjunctions such as and and or are used in ways that differ markedly from their logical or "truthtabular" senses. For instance, and often expresses the temporal order of two conjoined events (Bloom et al, 1980). Thus, (1) and (2) mean quite different things:
(1) He ran through the door and slipped on a banana peel.
(2) He slipped on a banana peel and ran through the door.

While (1) and (2) differ only in the order of the conjuncts, this results in a different temporal ordering of the events. Conjunctions can also express causality, counterfactuals, or subordination (see Culicover \& Jackendoff, 1997). This departure from formal logic has long been recognized by linguists of all stripes (e.g., Hoeksema, 1987; Klinedinst \& Rothschild, 2012). But what about those cases where and
and or are used in a sparse discursive context and actually appear to express a simple logical relation? What are the lexical semantics of and and or in their most austere uses?

On a classic formal account, the semantics of these function words is impoverished, contributing to the meaning of an utterance only in virtue of the meaning of the conjoined content words (e.g., Keenan \& Faltz, 1985; cf. Boole, 1854). More recent work, however, has prompted a reconsideration of the semantics of function words, and of the semantic content of grammar more generally. Langacker (2008) has argued that grammar is inseparable from meaning, since it shapes conceptualization in subtle but reliable ways. According to Langacker (1987), conjunctions like and and or prompt the "juxtaposition" of two or more objects or events in a dynamic conceptualization. Moreover, he and others (e.g. Landau \& Jackendoff, 2003; Talmy, 2000) have argued that schematic spatial information may lie at the core of linguistic meaning. Could the "juxtaposition" prompted by conjunctions rely on implicit spatial representations?

This possibility aligns with recent evidence that language comprehension involves the dynamic construction of an embodied mental simulation. In contrast with approaches that posit abstract, symbolic representations (e.g. Landauer \& Dumais, 1997; Markman \& Dietrich, 2000), embodied approaches argue that linguistic meaning is fundamentally tied to perceptual, motor and affective representations (Barsalou, 1999). Understanding "He threw the apple into the air," might involve activating cortical circuits implicated in perceiving the color red (Connell, 2007), perceiving motion (Saygin et al, 2012), or performing the action of throwing (Masson, Bub, \& Warren 2008). To account for how less concrete language is grounded in perception and action, proponents of some embodied approaches to language comprehension have appealed to "metaphorical" representations that map concrete experience to abstract linguistic content (Gibbs, 2006; Gallese \& Lakoff, 2005). For instance, respect can be conceptualized in terms of vertical height-"I look $u p$ to my superiors"-while similarity can be conceptualized in terms of closeness"Our ideas are quite close" (Lakoff \& Johnson, 1999). And, in fact, comprehending language about respect, similarity,
and other abstract concepts appears to involve schematic spatial representations (Guerra \& Knoeferle, 2012; Richardson et al., 2003; but see Bergen et al., 2007). Thus, the meaning of content words-both concrete and abstract-may include schematic spatial information.

But what about function words, such as conjunctions? Could their meaning also involve schematic spatial information, co-opting space to juxtapose conjuncts? There is evidence that grammatical tense, for instance, activates spatial representations. Santiago and colleagues (2007) found that participants were faster to categorize words as referring in the past tense when words were presented on the left (vs. right) side of the screen, but faster for words in the future tense when presented on the right (vs. left)—as if grammatical tense activated a left-to-right mental timeline (see also Torralbo, Santiago \& Lupiáñez, 2006). The "juxtaposition" prompted by conjunctions may also rely on schematic spatial representations, such as containment (see Glenberg, 2010). Comprehension of and, for instance, could involve a spatial grouping of the conjuncts, while or could mark alternatives by separating them spatially via containment. However, there is no clear experimental evidence showing that function words such as conjunctions are indeed related to spatial representations.

In the present two studies, we used two offline tasks to probe spatial representations underlying the meaning of conjunctions. In the first drawing study, participants created schematic sketches of conjunctions; in the second rating study, they rated schematic spatial diagrams on how well they represented different conjunctions. Both drawing and rating tasks have been used to study spatial representations activated by language, but only for concrete and abstract content words (Richardson et al, 2001). If conjunctions also co-opt spatial schemas to keep track of conceptual relations between conjuncts, then we should see a reliable, systematic use of spatial properties like distance, size, or containment to represent different conjunctions.

## Experiment 1: Drawing study

Experiment 1 used a drawing paradigm to examine whether representations of space are used to understand and visually depict the relationships expressed by four conjunctions (and, or, but, and either... or). If spatial representations are co-opted, then participants should systematically use spatial information to differentially represent the relations expressed by conjunctions. Alternatively, if conjunctions relate the meaning of the conjuncts in an abstract or logical fashion, no reliable differences in the use of spatial information should emerge.

## Method

Participants 108 native speakers of German completed the drawing task. They all gave informed consent and received monetary compensation for their participation.

Materials Three German conjunctions (und 'and'; aber, 'but'; and oder 'or') and a German correlative conjunction
(entweder... oder, 'either... or') were presented on a single sheet of paper (Fig. 1). Each conjunction appeared as "Object X conjunction Object Y " at the top of a blank square. Participants could select objects and frames for their drawings (Fig. 2).


Figure 1: Example of the drawings from a single participant.


Figure 2: Objects and frames of different shapes and sizes to be used in the drawings.

Design and procedure In a within-subjects design, with conjunction as a factor ('and', 'but', 'or', 'either... or'), each participant was instructed to make one drawing for each conjunction (see Fig. 1) using the objects in Figure 2. Participants saw two examples for the prepositions with and without. They were told that there were no correct or incorrect answers. The order of the conjunctions (Fig. 1) was counterbalanced.

Analysis We examined the drawings' spatial dimensions of distance, containment and size. Distance was defined as millimeters (mm) between objects' centers; containment codes whether objects were (or weren't) separated by one or more frames; size codes whether the objects had the same or a different size. Normalized distance scores ( $z$-scores) were analyzed with linear mixed effect regression (LMER, lme4 package for R statistical software). Mixed-effects models are suitable for analyzing unbalanced data and capture participants' variation around multiple fixed effects similar to ANOVAs (Quené \& van den Bergh, 2008). Our LMER modeled distance with conjunction as fixed effect, participant as random intercept, and the fixed effect as random slope.

For the analyses with containment and size, we calculated the percentage of representations that used these dimensions (e.g., containment was scored as present when an object was drawn with a frame around it, and size when differentlysized objects were used). A binomial test evaluated whether these percentages differed significantly from chance.

## Results

Distance Figure 3 shows the normalized mean distances between objects by conjunction. A positive deviation from zero (the intercept and grand mean) indicates objects were drawn farther apart than the grand mean object distance; a negative deviation indicates they were closer together. Figure 3 illustrates that while objects were drawn farther apart than average for 'either... or' and 'or', they were drawn closer together for 'and'. Object distance for 'but' did not differ from average. The LMER ${ }^{1}$ model confirmed a main effect of conjunction for distance ( $p<.001$ ).


Figure 3: Normalized mean distances between objects for each conjunction. Error bars represent standard errors.

Post-hoc comparisons (Bonferroni corrected) confirmed shorter between-object distance for 'and' than any of the other conjunctions ( $p \mathbf{s}<.001$ ); objects for 'either...or' were significantly farther apart than those for 'but' ( $p=.01$ ). The difference in distance between 'or' and 'but' did not reach significance ( $p=.24$; uncorrected $p=.037$ ), and 'either...or' and 'or' did not differ ( $p=1$ ).


Figure 4: Participant percentage using containment (upper graph) and size (lower graph) for each of the conjunctions.

Containment \& Size Figure 4 illustrates, for each conjunction, the percentage of participants who used frames around objects and different object sizes in their drawings.

[^287]For 'but', the use of containment did not differ from chance ( $47 \%, p=.6$ ). Instead, differently-sized objects distinguished the conjuncts ( $62 \%, p=.017$ ). For 'either... or', containment ( $61 \%, p=.026$ ), but not size ( $36 \%, p=.008$ ) was used above chance. For 'or', the use of containment did not differ from chance ( $47 \%, p=.6$ ), but differences in size were systematically avoided ( $37 \%, p=.012$ ). Finally, drawings for 'and' avoided the use of containment $(28 \%, p<.001)$ and used size at the level of chance ( $46 \%, p=.4$ ).

## Discussion

As predicted, different conjunctions were reliably associated with particular spatial dimensions. When two objects were conjoined by 'and', they were drawn close together and not separated by frames. By contrast, for 'or' and 'either...or' objects were drawn farther apart and separated by frames. Finally, depictions of 'but' relied on size to contrast the objects, but made no use of containment or distance. These conjunctions, therefore, elicited reliable spatial depictions in the absence of content words or linguistic context.

But do these results reflect spontaneous associations between conjunctions and space, or task-induced strategic reflection? To rule out that participants interpreted all four conjunctions and planned their sketches, perhaps to contrast them, we conducted a self-paced rating study based on the results of Experiment 1. In the rating study, participants only saw one conjunction-schema pair at a time. If sketches served to contrast the conjunctions, then use of space should disappear, or at least be greatly diminished in the rating task when only one pair is rated at a given time.

## Experiment 2: Rating study

Each conjunction ('and', 'or', 'but', ‘either...or') was paired with each of eight spatial schemas, designed to contrast three spatial dimensions: distance, containment, and size (Figure 5). These conjunction-schema pairs were randomly presented, so that participants could not predict the ensuing schema-conjunction pair. If the use of spatial information was not strategic, then ratings of how well a given depiction illustrates the meaning of a conjunction should replicate findings from Experiment 1. Specifically, we predict higher ratings for 'either... or', and 'or' when paired with schemas representing far (vs. close) distance and separated containers (vs. objects-contained). By contrast, ratings for 'and' should be higher with schemas representing close (vs. far) distance and objects-contained (vs. separated containers). Finally, we predict no differences for 'but' on distance- or containment-related schema ratings, but higher than average ratings for size-related schemas.

## Method

Participants A further twenty-four native German speakers completed Experiment 2. They all gave informed consent and received monetary compensation for their participation.

Materials Figure 5 shows the schematic depictions. Seven visual schemas covered the three dimensions analyzed in

Experiment 1 (i.e., distance, containment, size, Fig. 5, A-G); an eighth schema served as a baseline (Fig. 5 H). Each schema was presented on the computer screen with each one of the four conjunctions from Experiment 1.


Figure 5: Depictions for the spatial schemas: far (A); close (B); big (C), small (D), one-container (E); two-containers (F); objects-contained (G); and baseline (H).

Design and procedure A within-subjects design, had schema (eight schemas; Fig. 5) and conjunction ('and', 'or', 'but', 'either...or') as factors. Participants rated each possible pairing on how well a schema depicted a conjunction using a 7 -point scale (1=very bad to 7=very good). Items were presented one at a time on a computer monitor, and participants responded self paced. Experiment Builder v10.6 software (SR Research) recorded responses and randomized trial order.

Analysis For each conjunction, we normalized participants' raw ratings relative to their rating of the baseline schema H by subtracting their baseline rating from their other ratings ${ }^{2}$. Thus, within a conjunction, schemas that were judged more acceptable than baseline received a positive score, but a negative score if they were less acceptable than baseline.

Schema ratings were split into three subsets, based on the three spatial dimensions analyzed in Experiment 1. The distance subset included ratings for far and close schemas; the containment subset included ratings for one-container, two-containers and objects-contained schemas; and the size subset included ratings for big and small schemas. Each set of normalized ratings was then analyzed separately using an LMER model, with schema and conjunction as fixed effects, participant as random intercept, and the main effects and interaction of the fixed effects as random slopes. Planned dependent $t$-tests (Bonferroni corrected) compared ratings for each schema within conjunctions.

## Results

Distance The LMER showed neither main effects of schema nor conjunction ( $p s>.29$ ). However, as predicted, schema and conjunction interacted ( $p=.011$ ), with higher ratings for the far schema for 'and', but the close schema for 'either...or' and 'or' (Fig. 6).

Planned pairwise comparisons assessed the effect for each conjunction. For 'but', ratings did not differ for the far and close schemas ( $p=.92$ ). For both 'either... or' and 'or', by contrast, the far schema received higher ratings ( $p<.001$ and

[^288]$p=.018$, respectively). The pattern reversed for 'and', for which the close schema was reliably preferred ( $p<.001$ ).


Figure 6: Normalized-to-baseline mean rate for far and close schemas for all conjunctions. Error bars represent standard errors.

Containment LMER analyses showed a reliable main effect of conjunction ( $p=.002$ ) but not of schema ( $p=.3$ ); schema and conjunction interacted, as predicted ( $p<.001$, Fig. 7).

Planned pairwise comparisons examined containment preferences for each conjunction. For 'either... or', the twocontainers schema-which maximally separates the two objects-was rated higher than both the objects-contained and the one-container schemas (both $p<.001$ ). Similarly, for 'or', the two-container schema was significantly preferred over the one-container schema ( $p=.002$ ), and was marginally preferred over the objects-contained schema ( $p=.08$; uncorrected $p=.027$ ). By contrast, for 'and', the objects-contained schema-which groups both objects together-received the highest ratings among the containment-related schemas. The one-container schema was significantly disliked, compared to both the objectscontained and the two-container schemas (both $p<.001$, ps for the other comparisons, n.s.).


Figure 7: Normalized-to-baseline mean rate for onecontainer, two-container and objects-contained schemas. Error bars represent standard errors.

Size The LMER showed a main effect of conjunction ( $p=.002$ ). Schemas that highlighted differences in size were rated highly for 'but', nevertheless, dispreferred for all other conjunctions (Fig. 8, other ps n.s.).


Figure 8: Normalized-to-baseline mean rate for big and small schemas for all conjunctions. Error bars represent standard errors.

## Discussion

Experiment 2 confirmed that participants exhibit systematic preferences for spatial representations of conjunctions. Both 'either...or' and 'or' were rated higher with larger distances between objects, while 'and' was rater higher for shorter object distances. These conjunctions were also contrasted by the ratings for containment: the two-container schema was preferred for both 'either...or' and 'or', while the top-rated schema for 'and' contained both objects in a single frame. Finally, schemas that depicted size differences were reliably preferred for 'but', and rejected for all other conjunctions.

## General Discussion

Although previously suggested in the literature (e.g., Glenberg, 2010; Langacker, 2008), until now there was no experimental evidence that space might play a role in the representation of function words such as conjunctions. In two experiments, conjunctions were systematically associated with schematic spatial information, both when participants produced and when they rated spatial representations in the context of conjunctions.

We have framed these results in terms of semantics, and we believe they can shed light on the comprehension of conjunctions in natural language. But they may also tell us something about norms of visual representations or the communicative use of space. Logic and mathematics are rife with spatial diagrams used to represent and reason about logical relations, including and and or diagrams that are strikingly similar to the spatial representations in the current studies (Fig. 9; Guaquinto, 2007). Similarly, Langacker's Cognitive Grammar (2008) relies on spatial diagrams to represent relations between grammar and conceptualization. Sketches and diagrams, after all, are powerful tools for representing abstract concepts (Tversky, 2011).


Figure 9: Venn diagrams use spatial containment to depict logical relations: and (left) and or (right). And is depicted by a compact area, while or involves two separated areas.

If conjunctions are associated with spatial representations, then this may even account for some of the varied senses of and and or that have been discussed in the literature (e.g. Culicover \& Jackendoff, 1997). For instance, since time is also associated with spatial representations (e.g. Santiago et al., 2007), an implicit schematic spatial representation of the conjuncts could perhaps also induce a temporal ordering.

Where does this leave the relation between lexical semantics and logic? Perhaps closer than ever. In their book on the conceptualization of mathematics, Lakoff and Núñez (2000) suggest that "much of what is often called logical inference is in fact spatial inference mapped onto an abstract logic domain" (p.43). If so, then reasoning about logical relations, such as and and or, may rely on "metaphoric" representations of containment and distance (see, e.g., Boot \& Pecher, 2011; Guerra \& Knoeferle, 2012). If both the semantics of conjunctions and formal logic turn out to rely on space, then natural language semantics may be closer to formal logic than recently supposed-if we're willing to accept an appropriately naturalized version of formal logic, and an appropriately embodied version of lexical semantics.

Indeed, a question that remains unaddressed is whether schematic spatial information plays a spontaneous role in the real-time comprehension of conjunctions, when space is not an explicit part of the task. Suggestively, this is the case for content words. Richardson and colleagues (2001) used two offline norming studies to elicit schematic spatial representations associated with both concrete and abstract verbs (e.g. give, respect). They later found that these spatial schemas systematically influenced real-time comprehension of the associated verbs (Richardson et al, 2003), suggesting that the schemas elicited by the offline tasks were active during online language processing. We hypothesize that similar spatial processing may occur during the processing of conjunctions-that is, that the online comprehension of conjunctions may also involve schematic spatial representations of the kind examined here. Such online measures are necessary before we can draw definite conclusions about the semantics of conjunctions.

We do know, however, that conjunctions such as either...or modulate online sentence comprehension (e.g. Frazier, Munn \& Clifton, 2000 for and-coordinations). In a reading study, Staub \& Clifton (2006) examined the effect of the presence or absence of the word either on reading times for the second conjunct of or-coordinated structures (both for noun phrases and independent clauses). They found that the presence of either facilitated the reading of the content that followed the word or. These findings showed that conjunctions (and, either... or) can influence online sentence interpretation. Future studies should investigate whether these online effects extend to influences on spatial processing.

## Conclusion

We have shown that different conjunctions are distinctively associated with spatial dimensions of distance, containment, and size. In both a drawing and a rating task, people
associated 'and' with closeness and containment; 'or' and 'either...or' with distance and separation; and 'but' with contrasting size. Future work will investigate whether these schematic spatial properties are activated during online comprehension, and determine their functional contribution. Nevertheless, the present experiments highlight the use of space to distinguish abstract grammatical relations, suggesting the meaning of different function words can be expressed through distinct visual spatial representations.

## Acknowledgments

This research was funded by the Cognitive Interaction Technology Excellence Cluster (German Research Foundation) and by a PhD scholarship awarded to EG by the Ministry of Education, Government of Chile. We thank the members of the Language \& Cognition Lab (CITEC, University of Bielefeld) for their help with data collection.

## References

Barsalou, L. (1999). Perceptual symbol systems. BBS, 22, 577-660.
Bergen, B, Lindsay, S., Matlock, T., \& Narayanan, S. (2007). Spatial and linguistic aspects of visual imagery in sentence comprehension. Cognitive Science, 31, 733-764.
Bloom, L., Lahey, M., Hood, L., Lifter, K., \& Fiess, K. (1980). Complex sentences: Acquisition of syntactic connectives and the semantic relations they encode. Journal of Child Language, 7, 235-261.
Boole, G. (1854). An investigation of the laws of thought on which are founded the mathematical theories of logic and probabilities. Cambridge, UK: MacMillan and Co.
Boot, I., \& Pecher, D. (2011). Representation of categories: Metaphorical use of the container schema. Experimental Psychology, 58, 162-170.
Connell, L. (2007). Representing object colour in language comprehension. Cognition, 102, 476-485.
Culicover, P., \& Jackendoff, R. (1997). Semantic subordination despite syntactic coordination. Linguistic Inquiry, 28, 195-217.
Gallese, V., \& Lakoff, G. (2005). The brain's concepts: The role of the sensory-motor system in conceptual knowledge. Cognitive Neuropsychology, 22, 455-479.
Gibbs, R. (2006). Metaphor interpretation as embodied simulation. Mind \& Language, 21, 434-458.
Giaquinto, M. (2007). Visual thinking in mathematics: An epistemological study. New York, NY: Oxford University Press.
Glenberg, A. (2010). Embodiment as a unifying perspective for psychology. Wiley Interdisciplinary Reviews: Cognitive Science, 1, 586-596.
Guerra, E., \& Knoeferle, P. (2012). Abstract language comprehension is incrementally modulated by nonreferential spatial information: evidence from eyetracking. Proceedings of the 34th Annual Meeting of the Cognitive Science Society (pp. 1620-1625). Austin, TX: Cognitive Science Society.

Frazier, L., Munn, A., \& Clifton, C. (2000). Processing coordinate structure. JPR, 29, 343-368.
Hoeksema, J. (1987). The semantics of non-boolean "and." Journal of Semantics, 6, 19-40.
Keenan, E., \& Faltz, L. (1985). Boolean Semantics for Natural language. Dordrecht, Holland: D. Reidel.
Klinedinst, N \& Rothschild, D. (2012). Connectives without truth tables. Natural Language Semantics, 20, 1-39.
Lakoff, G., \& Johnson, M. (1999). Philosophy in the flesh: The embodied mind and its challenge to western thought. New York, NY: Basic Books.
Lakoff, G., \& Núñez, R. (2000). Where mathematics comes from: How the embodied mind brings mathematics into being. New York, NY: Basic Books.
Landau, B., \& Jackendoff, R. (1993). "What" and "where" in spatial language and spatial cognition. Behavioral and Brain Sciences, 16, 217-265.
Landauer, T., \& Dumais, S. (1997). A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. Psychological Review, 104, 211-240.
Langacker, R. (1987). Foundations of Cognitive Grammar. Standford, CA: Standford Universtity Press.
Langacker, R. (2008). Cognitive Grammar: A Basic Introduction. New York, NY: Oxford University Press.
Markman, A. \& Dietrich, E. (2000). Extending the classical view of representation. TiCS, 4, 470-475.
Masson, M., Bub, D. \& Warren, C. (2008). Kicking calculators: Contribution of embodied representations to sentence comprehension. JML, 59, 256-265.
Quené, H. \& van den Bergh, H. (2008). Examples of mixedeffects modeling with crossed random effects and with binomial data. JML, 59, 413-425.
Richardson, D., Spivey, M., Barsalou, L., McRae, K. (2003). Spatial representations activated during real-time comprehension of verbs. Cognitive Science, 27, 767-780.
Richardson, D., Spivey, M., Edelman, S. \& Naples, A. (2001). "Language is spatial": Experimental evidence for image schemas of concrete and abstract verbs. Proceedings of the 23 rd Annual Meeting of the Cognitive Science Society. Mawhah, NJ: Erlbaum.
Santiago, J., Lupiáñez, J., Perez, E., \& Fuenes, M.J. (2007) Time (also) flies from left to right. $P B \& R, 14,512-516$.
Saygin, A., McCullough, S., Alac, M., \& Emmorey, K. (2010). Modulation of BOLD response in motion sensitive lateral temporal cortex by real and fictive motion sentences. J Cog Neur, 22, 2480-90.
Staub, A. \& Clifton, C. (2006). Syntactic prediction in language comprehension: Evidence from Either... or. JEP: LMC, 32, 425-436.
Talmy, L. (2000). Toward a cognitive semantics, Volume I: Concept structuring systems. Cambridge, MA: MIT Press
Torralbo, A., Santiago, J., \& Lupianez, J. (2006). Flexible conceptual projection of time onto spatial frames of reference. Cognitive Science, 30, 745-757.
Tversky, B. (2011). Visualizations of thought. Topics in Cognitive Science, 3, 499-535.

# How does social comparison affect regret and relief in children, adolescents and adults? 

Marianne Habib (marianne.habib@ parisdescartes.fr)<br>LaPsyDÉ, CNRS Unit 3521, Paris Descartes University, Sorbonne Paris Cité, France<br>Sorbonne - Labo A. Binet, 45 rue Saint Jacques, 75005 Paris, France

Sylvain Moutier (sylvain.moutier@ parisdescartes.fr)<br>LaPsyDÉ, CNRS Unit 3521, Paris Descartes University, Sorbonne Paris Cité, France<br>Sorbonne - Labo A. Binet, 45 rue Saint Jacques, 75005 Paris, France

Grégoire Borst (gregoire.borst@parisdescartes.fr)
LaPsyDÉ, CNRS Unit 3521, Paris Descartes University, Sorbonne Paris Cité, France
Sorbonne - Labo A. Binet, 45 rue Saint Jacques, 75005 Paris, France
Olivier Houdé (olivier.houde @ parisdescartes.fr)
LaPsyDÉ, CNRS Unit 3521, Paris Descartes University, Sorbonne Paris Cité, France Sorbonne - Labo A. Binet, 45 rue Saint Jacques, 75005 Paris, France
Institut Universitaire de France, 103, Bd Saint-Michel, 75005 Paris, France
Mathieu Cassotti (mathieu.cassotti@parisdescartes.fr)
LaPsyDÉ, CNRS Unit 3521, Paris Descartes University, Sorbonne Paris Cité, France
Sorbonne - Labo A. Binet, 45 rue Saint Jacques, 75005 Paris, France


#### Abstract

Apprehending the development of complex emotions is crucial to understand the development of decision-making. Regret and relief are complex counterfactual emotions, which can arise in private or in social contexts. The aims of the present study were (i) to uncover the development of regret and relief and (ii) to explore the development of a social form of regret and relief in a context of competition. The first experiment provides evidence that the ability to experience regret and relief continues to develop until adolescence, consistent with the implication of the orbitofrontal cortex in their experience. In a context of competition, we observed that adolescents were less able to experience social regret compared to children and adults, whereas their feeling of social relief was reinforced. Besides, adolescents failed to question the appropriateness of their initial decision. This result could provide an explanation for adolescents' enhanced propensity to engage in risky behaviours.


Keywords: Regret; Relief: Counterfactual thinking, Social context; Decision making; Development.

## Introduction

Psychology and neuroscience studies have provided converging evidence that emotions play a crucial role in adaptive decision-making (Loewenstein, Rick \& Cohen, 2008). Thus apprehending the development of basic and complex emotions is crucial to fully understand the development of decision-making. Among these emotions, counterfactually mediated emotions - like regret and relief are related to counterfactual thinking and rely on comparison processes. In a private context, these processes
rely on a comparison between what has happened and what could have happened if the subject had made another choice (Ritov, 1996). The counterfactual comparison has an informative function, as it enables to determine a reference point according to which the obtained outcome will be evaluated. This process can also be motivated by the social context. It will then rely on a comparison between what has happened to the subject and what has happened to another person, like a competitor, who made a different choice (Bault, Coricelli \& Rustichini, 2008).

To date, developmental psychology has mainly focused on the development of counterfactually mediated emotions in young children (Weisberg \& Beck, 2010) showing that the experience of regret develops around 5 years of age, whereas the experience of relief develops around 7 years of age. Recently, the development of regret and relief in adolescence has been investigated in a probabilistic gambling task (Burnett, Bault, Coricelli \& Blakemore, 2010). Participants' emotional ratings revealed that relief, but not regret, develops during adolescence. The lack of development of regret in adolescence is surprising given that increasing feelings of regret and relief are positively correlated with enhanced activity in the orbitofrontal cortex (OFC) that continues to mature until late adolescence (Camille et al., 2004; Gogtay et al., 2004).
A possible explanation of the lack of evidence for the development of regret in adolescence might relate to the nature of the variables used to study this emotion. Previous studies have focused primarily on emotional ratings. Knowing that counterfactually mediated emotions are related to participants' decision and in order to fully
apprehend these emotions, it was necessary to consider the degree to which participants are willing to reconsider their initial choice after experiencing regret (Chua, Gonzalez, Taylor, Welsh, \& Liberzon, 2009).

## Experiment 1 - Private context

Aims and Hypotheses. Thus, the aim of the first experiment was to uncover the development of regret and relief in late childhood, adolescence and adulthood.
In order to do so, participants performed a child friendly gambling task adapted from Camille et al. (2004). We asked participants to choose between two wheels of fortune that differed in the amount of gain and loss expected and the probability of winning. We manipulated the outcome of the wheel of fortune that was not selected by the participants to induce either regret or relief. For each trial, participants rated how they felt about the outcome and their willingness to modify their choice, on a classical likert-type scale.
As the OFC has a fundamental role in the experience of regret, and given the late maturation of this brain area, we expect to observe a progressive development of the emotional experience of counterfactually mediated emotions from childhood to adulthood, in the private context. Besides, the choice rating could be a more sensitive measure in order to study the developmental trajectories and understand the complexity of regret and relief.

## Method

Participants. In this private context, we recruited 53 volunteers: 19 children (mean age $=11.2$ years, $\mathrm{SD}=0.66$ ), 17 adolescents (mean age $=14.5$ years, $\mathrm{SD}=0.40$ ) and 17 university psychology students (mean age $=20.2$ years, SD $=1.48$ ).

Written parental consent was obtained for children and adolescents prior to the assessment session. Participants were tested in accordance with international norms governing the use of human research participants.

Experimental Procedure. Participants performed 36 trials of a child friendly gambling task (cf. fig. 1). For each trial, participants chose between two wheels of fortune, an advantageous wheel (with a positive expected value) and an attractive but disadvantageous wheel (with a negative expected value). Then, two feedbacks, partial and complete feedbacks, were successively provided to participants.

For the partial feedback, the outcome obtained on the selected wheel was displayed on the screen for 4 s (fig. 1.c.). Thus, the partial feedback induced either disappointment (in the case of losses) or elation (in the case of gains). For the complete feedback, participants were informed of the outcome of the alternative wheel for 4 s (fig. 1.f). They could thus compare the obtained outcome to the counterfactual outcome. The complete feedback was designed to induce either regret (when the comparison
between the outcomes was unfavourable to the participant) or relief (in the opposite case).


Figure 1: Experimental design of a trial inducing regret. Two 'wheels of fortune' were displayed on the computer screen (1.a.). After the participant's choice (1.b.), participants were informed of obtained outcome (partial
feedback-1.c.). Then the participant must rate an 'emotional' scale (1.d.) and a 'choice' scale (1.e.). Finally, participants were informed of the complete feedback (1.f.) and had to rate again an 'emotional' scale (1.g.) and a 'choice' scale (1.h.).

After each feedback, participants rated a 7-point 'emotional' scale (fig. 1.d. and 1.g.), ranging from 1 (I am unhappy) to 7 (I am happy). Participants then rated a 7 -point 'choice' scale (fig. 1.e. and 1.f.), ranging from 1 (I wish to modify my choice) to 7 (I do not wish to modify my choice), on which they indicated how much they wished to reconsider their choice.

## Results

Participants' ratings were analysed in four outcome conditions: (a) low loss vs. high loss condition, which should induce minimal relief; (b) low win vs. high win condition, inducing minimal regret; (c) low win vs. high loss condition, inducing maximal relief; and (d) low loss vs. high win condition, inducing maximal regret.
For the 'emotional' and 'choice' ratings analyses, we computed difference scores (see Weisberg \& Beck, 2010). Ratings on the partial feedback were subtracted from ratings on the complete feedback. Thus, the emotional and choice scores ranged between -6 and +6 . We carried out onesample t-tests to determine whether both scores differed from zero. A negative emotional score would suggest that participants experienced regret whereas a positive emotional score would suggest that participants experienced relief. A
negative choice score would suggest that participants wanted to modify their initial choice whereas positive choice score would suggest that participants want to maintain their initial choice.
A 3 (age: children vs. adolescents vs. adults) x 4 (outcomes: maximal regret, minimal regret, maximal relief, minimal relief) mixed-design ANOVA on the 'emotional' scores revealed that these scores differed between the three groups of participants, $\mathrm{F}(2,50)=3.24, \mathrm{p}<.05, \eta_{\mathrm{p}}^{2}=.12$, and between the type of outcomes, $\mathrm{F}(3,50)=57.95, \mathrm{p}<.001, \eta_{\mathrm{p}}{ }^{2}$ $=.54$. The age of the participants affected the emotional scores differently in the four types of outcomes, $\mathrm{F}(6,150)=$ 2.73, $\mathrm{p}<.05, \eta_{\mathrm{p}}^{2}=.10$. In the minimal regret condition, children and adolescents' average emotional scores were lower than adult scores, $t(34)=-1.60, p=.056, d=.72$ for children and $\mathrm{t}(32)=-1.59, \mathrm{p}=.052, \mathrm{~d}=.72$ for adolescents. Similarly, in the maximal regret condition, children and adolescents' emotional scores were lower than adult scores, $\mathrm{t}(34)=-2.41, \mathrm{p}<.01, \mathrm{~d}=.70$ for children and $\mathrm{t}(32)=-3.47$, $\mathrm{p}<.001$, $\mathrm{d}=1.75$ for adolescents (cf. fig. 2). No other differences were significant.
To sum up, all participants experienced regret in the conditions designed to induce this emotion, but children and adolescents' subjective experience of regret was reduced compared to adults. Besides, although all groups reported relief in the condition designed to induce maximal relief, children and adolescents did not experience relief when they lost a small amount of money but avoided losing a higher amount, $\mathrm{t}(18)<1$ and $\mathrm{t}(16)=2,39, \mathrm{p}=.09$ respectively.


Figure 2: 'Emotional' ratings for the private context mean scores (partial feedback ratings subtracted from complete feedback ratings) for regret and relief trials.
Ratings were analysed according to four types of outcome (minimal relief, minimal regret, maximal relief and maximal regret). We compared regret and relief scores to zero (onesample $t$-tests with zero as the test value, Bonferronicorrected, *p < .05, ** p < .005, ***p < .001).

The 3 (age) x 4 (type of outcome) mixed-design ANOVA on the 'choice' scores revealed a main effect of the type of outcome, $\mathrm{F}(3,150)=36.7, \mathrm{p}<.001, \eta_{\mathrm{p}}^{2}=.42$, but no main effect of age, $F(2,50)=2.87, p=.07$. Interestingly, the interaction between age and outcome was significant, F $(6,150)=2.54, \mathrm{p}<.05, \eta_{\mathrm{p}}^{2}=.09$. Planned comparisons of the 'choice scores' revealed that children were less willing to modify their initial choice than were adolescents, $\mathrm{t}(34)=$
$-2.41, \mathrm{p}<.01, \mathrm{~d}=.66$, and adults, $\mathrm{t}(34)=-3.00, \mathrm{p}<.005, \mathrm{~d}$ $=1.23$, in the minimal regret (win-win) condition. Similarly, in the maximal regret (loss-win) condition, children were less willing to modify their initial choice than were adults, $\mathrm{t}(34)=-1.70, \mathrm{p}<.05, \mathrm{~d}=.61$. No other differences were significant.
As expected, adults wished to modify their initial choice in the conditions inducing regret, $\mathrm{t}(16)=-5.43, \mathrm{p}<.001$ in the minimal regret condition and, $\mathrm{t}(16)=-4.55, \mathrm{p}<.001$ in the maximal regret condition. On the contrary, they wished to maintain their initial choice in the conditions inducing relief, $\mathrm{t}(16)=4.23, \mathrm{p}<.005$ in the minimal relief condition and, $\mathrm{t}(16)=4.74 \mathrm{p}<.001$ in the maximal relief condition. As opposed to adults, children did not wish to modify their choice in the minimal regret condition, $\mathrm{t}(18)=1.03, \mathrm{p}>.1$, even if they experienced a significant feeling of regret.

## Discussion

Analyses of the 'emotional' scores revealed developmental differences for both types of counterfactually mediated emotions (regret and relief). All participants experienced regret in the two outcome conditions designed to induce this emotion, but children and adolescents' subjective experience of regret was reduced compared to adults.
Our results are consistent with the ones reported in a recent study (Rafetseder \& Perner, 2012) showing that regret develops progressively from childhood to adulthood and reaches its maximum level in the adult group.
The fact that children experienced relief after a small gain (maximal relief condition) but not after a small loss (minimal relief condition) might suggest that this group has difficulties distinguishing between two outcomes that both lead to a loss and thus might focus more on the loss they obtained rather than on the high loss they avoided (see Weisberg \& Beck, 2012 for similar results).
Moreover, analyses of the 'choice' scores demonstrate that regret affects the participant's willingness to reconsider their initial choice in adults, whereas it does not systematically lead to a reconsideration of the initial choice in children.
Adults wished to modify their initial choice in the conditions inducing regret but wished to maintain their initial choice in the conditions inducing relief. As opposed to adults, children and adolescents expressed no preference about modifying their choice in the condition inducing minimised relief.
Finally, we identified a dissociation in children between the experience of regret and the willingness to reconsider an initial choice, in the minimal regret condition (low win vs. high win condition) specifically. In this condition, the salience of the counterfactual alternative might be reduced as participants have already won on the selected wheel. When the obtained outcome is already good for them, it seems difficult for children to think counterfactually, to take in account the alternative win and then expressing their wish to modify their choice.

## Experiment 2 - Social context

Our first experiment provided evidence that the ability to experience regret and relief continues to develop during late childhood and adolescence.
As little is known about adolescents' sensibility to complex negative emotions, we were particularly interested in the emotional experience of regret in adolescents. It has actually been demonstrated that the anticipation of complex negative emotions - such as regret - can significantly contribute to decrease risky behaviours in adolescents (Conner, Sandberg, McMillan, \& Higgins, 2006; Richard, Van Der Pligt, \& De Vries, 1996). Indeed, studies focusing on the role of anticipated regret in risky decision making have revealed that inciting adolescents to anticipate the regret they could experience after a risky behaviour can significantly decrease the intentions to engage in this behaviour.
In order to apprehend risky decision making in adolescence, neurobiological models have postulated the existence of two distinct brain systems involved in decision making: a cognitive control system - supporting goaldirected decisions through the ability to inhibit impulsive behaviour - and a socio-emotional system - based on the valuation and prediction of potential rewards, that can bias decision (Chein, Albert, O’Brien, Uckert, \& Steinberg, 2011; Sommerville, Jones \& Casey, 2010). These models posit an imbalance between the maturity of adolescents' socio-emotional system and the relative immaturity of the cognitive control system (Sommerville, et al., 2010). Due to this imbalance, adolescents are hypersensitive to rewards, particularly in salient socio-emotional contexts (Chein et al., 2011; Ernst et al., 2005). This kind of context selectively increases adolescents' sensitivity to potential rewards, which could explain why adolescence corresponds to a period of greater risk seeking in everyday life (Chein et al., 2011).

However, to our knowledge, no study has examined the impact of a salient socio-emotional context on adolescents' experience of regret and relief. Yet, in everyday life, adolescents not only experience these emotions alone but also in social contexts - e.g., in school, when they compare their achievements with those of their schoolmates.

Aims and Hypotheses. Thus, the aim of this second experiment was to explore the development of a social form of regret and relief in adolescents, compared to children and adults. Social regret was defined as the negative emotion that one feels when he has missed an opportunity while another person has seized it and social relief as the opposite feeling.
In the socio-emotional context condition, participants were additionally informed that they would be playing against a schoolmate and that their results would be compared to those of the other player throughout the game.
We hypothesized that adolescents should be more influenced by the social context of competition than children and adults (Chein et al., 2011). If adolescents are
hypersensitive to the emotional context, compared to adults and children, they should demonstrate a heightened sensitivity to the gains they obtain and their evaluation of social relief should be biased. Therefore, adolescents should experience an enhanced feeling of social relief. In addition, if the social context also influences negative complex emotions, they should experience a decreased feeling of social regret compared to children and adults. Thus, their willingness to reconsider an initial choice should be attenuated compared to the other groups.

## Method

Participants. In the social context, we recruited 54 volunteers: 18 children (mean age $=11.8$ years, $\mathrm{SD}=0.43$ ), 18 adolescents (mean age $=14.5$ years, $\mathrm{SD}=0.40$ ) and 18 university students (mean age $=20.2$ years, $\mathrm{SD}=1.48$ ).
Written parental consent was obtained for children and adolescents prior to the assessment session. Participants were tested in accordance with international norms governing the use of human research participants.

Experimental Procedure. Participants performed 36 trials of the computerized child friendly gambling task used in experiment 1. The procedure and the stimuli of the gambling task were the same as in experiment 1.
However, in order to induce social regret and social relief, children and adolescents were informed that that they would be playing against a schoolmate and adults were told that they will be playing against another student of the same age and the same institute. We additionally informed participants that all of their choices would be compared to those of their competitor. Thus, the complete feedback informed the participants about the outcome obtained by their competitor, so that they could compare it to their own outcome.

## Results

Participants' ratings were again analysed according to four outcome conditions: (a) low loss vs. high loss condition, which should induce minimal relief; (b) low win vs. high win condition, inducing minimal regret; (c) low win vs. high loss condition, inducing maximal relief; and (d) low loss vs. high win condition, inducing maximal regret. Besides, we computed difference scores for the 'emotional' and 'choice' ratings analyses (cf. experiment 1).
A 3 (age) x 4 (outcome conditions) mixed-design ANOVA conducted on 'emotional' scores revealed a main effect of age, $\mathrm{F}(2,51)=12.64, \mathrm{p}<.001, \eta_{\mathrm{p}}^{2}=.33$, a main effect of outcome condition, $\mathrm{F}(3,153)=51.01, \mathrm{p}<.001, \eta_{\mathrm{p}}{ }^{2}$ $=.50$, and a significant interaction between age and outcome condition, $\mathrm{F}(6,153)=2.36, \mathrm{p}<.05, \eta_{\mathrm{p}}{ }^{2}=.08$. Planned comparisons revealed that in the minimal social relief condition, adolescents' social relief was higher than that expressed by children, $\mathrm{F}(1,51)=12.98$, p $<.001$, $\mathrm{d}=$ 1.19 (cf. fig. 3). No other differences were significant.


Figure 3: 'Emotional' ratings for the social context mean scores on regret and relief conditions. Ratings were analysed according to four types of outcome (minimal relief, minimal regret, maximal relief and maximal regret). We compared regret and relief scores to zero (one-sample ttests with zero as the test value, Bonferroni-corrected, * p < $.05,{ }^{* *} \mathrm{p}<.005,{ }^{* * *} \mathrm{p}<.001$ ).

Notably, adolescents expressed lower social relief in the maximal relief condition compared to the minimal social relief condition, $\mathrm{F}(1,51)=6.18, \mathrm{p}<.05, \mathrm{~d}=0.71$. In the maximal social regret condition, adolescents expressed less regret than children, $\mathrm{t}(51)=2.60, \mathrm{p}<.05, \mathrm{~d}=0.82$, and adults, $\mathrm{t}(51)=3.51, \mathrm{p}<.001, \mathrm{~d}=1.34$. Beside, planned comparisons in this condition also revealed a significant quadratic trend between age and the expression of regret, revealing a U-shaped developmental pattern, $\mathrm{F}(1,51)=$ 12.46, $\mathrm{p}<.001$, and no significant linear trend, $\mathrm{F}<1$. These results suggest that adolescents experience less social regret than children and adults in the maximal social regret condition.
A 3 (age) x 4 (outcomes) mixed-design ANOVA conducted on 'choice' scores revealed that these scores differed between the type of outcome, $F(3,153)=22.21, p<$ $.001, \eta_{\mathrm{p}}^{2}=.30$, but not between the three age groups, $F(2,51)=1.13, p=.33$. Interestingly, the age of the participants affected the choice scores differently in the four outcome conditions, $\mathrm{F}(6,153)=2.67, \mathrm{p}<.05, \eta_{\mathrm{p}}^{2}=.09$.
The willingness to maintain the initial choice was lower in the maximal social relief condition compared to the minimal social relief condition for adolescents and adults, $t(51)=$ $2.12, \mathrm{p}<.05, \mathrm{~d}=0.63$ and $\mathrm{t}(51)=3.51, \mathrm{p}<.005, \mathrm{~d}=0.83$, respectively. Adolescents were less willing to modify their choice in the maximal social regret condition than in the minimal social regret condition, $\mathrm{F}(1,51)=9.52, \mathrm{p}<.01$, $\mathrm{d}=$ 0.67 , and than adults, $\mathrm{F}(1,51)=5.33, \mathrm{p}<.05, \mathrm{~d}=0.83$.

## Discussion

The results of this experiment evidenced that a salient socioemotional context of competition impacts the feeling of regret and relief, specifically in adolescents. Critically, when adolescents obtained an initial negative outcome, their feeling of social relief was reinforced, compared to children and adults. Adolescents were actually far more relieved after obtaining an initial loss (minimal social relief) than after obtaining an initial win (maximal social relief). Given that the only difference between the minimal and maximal social
relief conditions was the presence of an initial loss in the minimal social relief condition, this result may suggest that a salient context of social competition has a direct impact on adolescents' sensitivity to losses, increasing the feeling of relief when the competitor obtained a greater loss.
On the other hand, adolescents' feeling of social regret was considerably attenuated compared to children and adults. Theoretically, the condition of maximal social regret should have the greatest effect on participants' self-esteem. However, even if adolescents are able to experience regret (cf. experiment 1 and Burnett et al., 2010), they did not express a significant feeling of regret in the maximal regret condition. Thus, we argue that this specific lack of social regret in adolescents may be a consequence of a heightened sensitivity to negative outcomes in a social context. Interestingly, this weaker tolerance to losses leads adolescents to down-regulate their feeling of social regret and their willingness to reconsider their choice. As such, when adolescents obtain a negative outcome, they fail to question the appropriateness of their initial decision. This result is in line with studies that revealed a relative deficit in adolescents' ability to tolerate and to learn from negative outcomes compared to adults (Aïte, et al., 2012; Cassotti, Houdé \& Moutier, 2011).
Our findings indicate that adolescents are not only hypersensitive to rewards but also to losses in salient socioemotional contexts by demonstrating that the socioemotional context of competition significantly impacts their feeling of social regret and of social relief after an initial loss. These results are in line with the proposition of an imbalance between the socio-emotional system and the cognitive control system in adolescence (Chein et al., 2011; Sommerville et al., 2010).

## General Discussion

The aims of this paper were to (i) to examine the development of regret and relief from childhood to adulthood and (ii) to explore the development of the ability to experience social regret and social relief in adolescents compared to children and adults in a context of social competition.
The results of the first experiment provide evidence that the ability to experience counterfactually mediated emotions - regret and relief - is attenuated in children and adolescents compared to adults. This result is consistent with the implication of the OFC in the experience of regret and relief (Camille et al., 2004).
Moreover, we observed that regret affects the participant's willingness to reconsider their initial choice in adults, whereas it does not systematically lead to the same reconsideration in children. We actually identified that the experience of regret and the willingness to reconsider an initial choice can be dissociated in children. This result may indicate a developmental dissociation between feeling and doing that has previously been observed among participants of the same age range (Cassotti et al., 2011).

The results from the choice scale reveal the importance of using both 'emotional' and 'choice' ratings to study the precise development of counterfactually mediated emotions such as regret and relief.
The results of the second experiment is the first to evidence that a salient socio-emotional context of competition impacts the feeling of counterfactually mediated emotions - regret and relief - specifically in adolescents. Critically, when adolescents obtained an initial negative outcome, their feeling of social regret was considerably attenuated compared to children and adults, whereas their feeling of social relief was reinforced. The present results suggest that in a social comparison context, adolescents are less able to experience social regret and fail to question the appropriateness of their initial decision, particularly when another person (a competitor) has obtained a higher outcome by choosing differently.

## Conclusion

In conclusion, the present paper evidenced that the ability to experience regret and relief continues to develop in late childhood and adolescence.
This paper is also the first to evidence that a salient socioemotional context of competition can impact the feeling of regret and relief, specifically in adolescents.

Finally, the present results suggest that in a social comparison context, adolescents are less able to experience social regret and fail to question the appropriateness of their initial decision. Thus, this result could provide an explanation for adolescents' enhanced propensity to engage in risky behaviours in everyday life.

## Acknowledgments

We thank the participants of the study and the teacher teams of the Saint Pierre Institute (Brunoy, France) and Guillaume de Normandie Junior High School (Caen, France).

## References

Aïte, A., Cassotti, M., Rossi, S., Poirel, N., Lubin, A., Houdé, O., \& Moutier, S. (2012). Is human decisionmaking under ambiguity guided by loss frequency regardless of the costs? A developmental study using the Soochow Gambling Task. Journal of Experimental Child Psychology, 112(2), 286-294.
Bault, N., Coricelli, G., \& Rustichini, A. (2008). Interdependent Utilities: How Social Ranking Affects Choice Behavior. PLoS ONE, 3(10), e3477.
Burnett, S., Bault, N., Coricelli, G., \& Blakemore, S.-J. (2010). Adolescents' heightened risk-seeking in a probabilistic gambling task. Cognitive Development, 25(2), 183-196.
Camille, N., Coricelli, G., Sallet, J., Pradat-Diehl, P., Duhamel, J.-R., \& Sirigu, A. (2004). The Involvement of the Orbitofrontal Cortex in the Experience of Regret. Science, 304(5674), 1167-1170.

Cassotti, M., Houdé, O., \& Moutier, S. (2011). Developmental changes of win-stay and loss-shift strategies in decision making. Child Neuropsychology, 17(4), 400-411.
Chein, J., Albert, D., O’Brien, L., Uckert, K., \& Steinberg, L. (2011). Peers increase adolescent risk taking by enhancing activity in the brain's reward circuitry. Developmental Science, 14(2), F1-F10.
Chua, H. F., Gonzalez, R., Taylor, S. F., Welsh, R. C., \& Liberzon, I. (2009). Decision-related loss: Regret and disappointment. NeuroImage, 47(4), 2031-2040.
Conner, M., Sandberg, T., McMillan, B., \& Higgins, A. (2006). Role of anticipated regret, intentions and intention stability in adolescent smoking initiation. British Journal of Health Psychology, 11(1), 85-101.
Ernst, M., Nelson, E. E., Jazbec, S., McClure, E. B., Monk, C. S., Leibenluft, E., Blair, J., et al. (2005). Amygdala and nucleus accumbens in responses to receipt and omission of gains in adults and adolescents. NeuroImage, 25(4), 1279-1291.
Gogtay, N., Giedd, J. N., Lusk, L., Hayashi, K. M., Greenstein, D., Vaituzis, A. C., Nugent, T. F., et al. (2004). Dynamic mapping of human cortical development during childhood through early adulthood. Proceedings of the National Academy of Sciences of the United States of America, 101(21), 8174-8179.
Loewenstein, G., Rick, S., \& Cohen, J. D. (2008). Neuroeconomics. Annual Review of Psychology, 59(1), 647-672.
Rafetseder, E., \& Perner, J. (2012). When the alternative would have been better: Counterfactual reasoning and the emergence of regret. Cognition \& Emotion, 26(5), 80019.

Richard, R., Van Der Pligt, J., \& De Vries, N. (1996). Anticipated Regret and Time Perspective: Changing Sexual Risk-taking Behavior. Journal of Behavioral Decision Making, 9(3), 185-199.
Ritov (1996). Organizational Behavior and Human Decision Processes, 66(2), 228-236.
Sommerville, L.H., Jones, R.M., Casey, B.J. (2010). A time of change: Behavioral and neural correlates of adolescent sensitivity to appetitive and aversive environmental cues. Brain and Cognition, 72, 124-133.
Weisberg, D. P., \& Beck, S. R. (2010). Children's thinking about their own and others' regret and relief. Journal of Experimental Child Psychology, 106(2-3), 184-191.
Weisberg, D. P., \& Beck, S. R. (2012). The development of children's regret and relief. Cognition \& Emotion, 26(5), 820-35.

# Unifying Theories of Consciousness, Attention, and Conscious Attention 

Harry Haroutioun Haladjian (h.haladjian@uws.edu.au)<br>School of Social Sciences and Psychology, University of Western Sydney<br>Bankstown Campus, Penrith, NSW 2751 Australia<br>Carlos Montemayor (cmontema@sfsu.edu)<br>Department of Philosophy, San Francisco State University<br>San Francisco, CA 94132 USA


#### Abstract

One of the more challenging research areas in cognitive science is the attempt to understand how the brain supports consciousness. This historically philosophical endeavor is now actively studied in the sciences, with research on visual attention being an especially promising area that can further our understanding of consciousness. A major problem with this cross-disciplinary pursuit, however, is that for philosophers and scientists, the terms consciousness, attention, and conscious attention are ambiguous and used differently even by those within the same academic discipline. The goal of this paper is to begin laying the groundwork for a unified study of consciousness by delineating common terminology for attention and consciousness and by identifying the relationship between the two within the study of conscious attention. This includes categorizing current theories according to a spectrum of theoretical complexity.


Keywords: attention; consciousness; conscious attention; philosophy of mind; perception.

## Introduction

Although the relationship between consciousness and attention has been at the center of recent discussions in cognitive science, the proposals for this relationship are based on assumptions that are problematic. For instance, it is often assumed that the terms "attention" and "consciousness" unambiguously describe specific types of mental phenomena that can be identified experimentally. There is empirical evidence, however, that there are different types of attention, with different neural correlates that cannot be reduced to one another (Parasuraman, 2000), which complicates the attempt to establish a clear relationship between attention and consciousness. Furthermore, while many authors think that it is plausible to define "attention" in terms of several basic types of attention, other theorists think that defining "attention" is hopeless (see Allport, 1993; Johnston \& Dark, 1986).

Similarly, some theorists think that there are at least two types of consciousness (e.g., Block, 1995), and that only one of them is strictly related to the subjective experience of conscious awareness. Other theorists think that the "hard problem" of consciousness (i.e., the study of phenomenal or subjective experience) makes the empirical study of consciousness, unlike the study of perceptual attention, intractable (e.g., Chalmers, 1996; Nagel, 1974). Finally, there are theorists who think that the hard problem of consciousness is just a pseudo-problem, and that
consciousness can and must be studied empirically (e.g., Churchland, 1996; Dennett, 2005). These contrasting opinions and approaches have complicated the study of consciousness in relation to attention, often resulting in a gridlock of concepts between opposing theories. It is possible, however, that many of the current theories on consciousness are not necessarily in opposition, since there may be semantic ambiguities producing these disputes.

Because of the polysemy of the terms "attention" and "consciousness", one should avoid stipulating definitions without first delineating empirical and theoretical constraints that such definitions must satisfy. It is crucial to determine whether the different theoretical perspectives refer to the same types of attention and the same types of consciousness. Based on empirical findings and a theoretical classification of the possible views on this topic, we propose definitions for forms of consciousness, forms of attention, and forms of conscious attention in order to provide a foundation to compare and move forward different theories.

Another goal of this paper is to offer a brief account of recent theories on consciousness, with a focused consideration of how empirical research on attention can provide the grounding for an empirically-driven account of consciousness. One way to do this is by analyzing recent theories on consciousness and attention by categorizing them according to a spectrum of theoretical complexity, starting with the theories that impose the strictest requirements on the interpretation of empirical findings to those that allow the widest range of possible interpretations. For instance, Jesse Prinz (2012) has defended the view that consciousness is just attention. This "strict" view entails that there cannot be any finding about attention that is not a finding about consciousness and vice versa.

Although an identity approach is parsimonious, since it reduces that kinds of cognitive processes associated with consciousness and attention to a single type, it creates the problem of reducing significantly the room for interpretation. For example, the desideratum of empirical adequacy seems to demand more theoretical leniency for the interpretation of research findings that indicate disassociations between attentional processing and conscious awareness. Should it not be possible that some form of attention exists without consciousness, even if consciousness cannot occur without attention? Michael Cohen and colleagues (2012) argue in favor of this possibility. At the extreme opposite of the spectrum, one
finds the view that consciousness and attention can be fully dissociated (i.e., there can be forms of consciousness without attention and vice versa) advocated by Koch and Tsuchiya (2007) and Lamme (2003), among others. This range of possible theoretical complexities provides insight to the approach one can take in studying conscious attention. Due to the diverging views, a meta-analysis is crucial for advancing this field, and future work should focus on such an in-depth analysis. We present a brief overview of the development of the views described above and outline the start of a such meta-analysis.

## Problems of Consciousness

Thomas Nagel (1974) said that the problem of conscious experience is what makes the mind-body problem both interesting and intractable. The problem of how the mind connects with the world would lose its allure, and even become trivial or irrelevant, if one had no idea how solutions to this problem would explain consciousness. Once the theoretical and empirical options to account for consciousness are carefully assessed, however, it becomes clear that they are all problematic. The best way of formulating the intractability of this problem is in terms of what David Chalmers called the "hard problem" of consciousness, that is, why would anything physical have conscious experiences and what is the relationship between physical brain processes and the subjective experience of consciousness? Much has been written about this problem, and there is now widespread consensus that it is not only a difficult philosophical problem, but also one of science's more difficult unsolved puzzles.

## Access vs. Phenomenal Consciousness

Although the problem of consciousness is remarkably intricate, a great deal of progress has been made on the theoretical front. A significant amount of conceptual clarity has been achieved with respect to the question of why functions for cognitive processing may explain some forms of conscious integration required for working memory (what Ned Block, 1995, calls access consciousness), but may not suffice to account for the qualitative aspects of conscious experiences (what Block calls phenomenal consciousness). Access consciousness provides a "workspace" for concepts and multi-sensory information to be accessed for the purposes of reasoning and performing complex actions. These do not necessarily need to reach awareness (i.e., reach a cognitive state where one can report experiencing it). Another theory, by David Rosenthal (2002), proposes that a higher-order thought (HOT) is required for one to be conscious of mental states. These are thoughts about mental states (resulting from sensations or memory retrieval) that allow us to be conscious of them. The relationship between access and phenomenal consciousness (between thoughts and higher-order thoughts in awareness) is one area where attention research may help, for example, by clarifying how thoughts move from access to phenomenal consciousness.

## Self Consciousness

Another source of problems concerning treatments of conscious perception is the role of the self in phenomenal experience. A number of intricate questions originate from this topic. Can one be conscious of something (an emotion, a perceptual representation, etc.) without also being conscious that one is conscious of it? Is the "self" constitutive of every possible experience without itself being experienced? How should we understand consciousness, self-awareness, and the conscious self? One problem with an emphasis on the "self" view is that it seems to demand too much to account for all conscious creatures and because of this reason, it seems to be empirically implausible. Christof Koch (2012), for instance, argues that the self is not necessary to have conscious experiences. He criticizes the mirror test, which infants and most animals fail, as a test for consciousness (although it seems to be a good test for self-awareness). The reasoning is that infants and many animals must have some kind of consciousness (of the phenomenal kind) because they experience pain, feel emotions, etc. They may not have self-consciousness but, the claim is, they do have phenomenal consciousness.

Plausible as this criticism is, however, the relationship between consciousness and self is much more intricate than first appearances suggest. In a passage where Koch is defining the scientific problem of consciousness, he uses two incompatible interpretations of the word "self" when criticizing the conclusion that failure to pass the mirror test indicates the lack of consciousness. One notion of the "self" is the higher-order self that recognizes a particular thought as hers (the recognitional self). Koch seems justified in claiming that the recognitional capacities associated with this kind of self may not be necessary for consciousness. But how to interpret the more primitive "self" that Koch associates with experiences of "flow" (the phenomenal self)? This is a central question that needs to be answered in order to understand the relationship between higher forms of self-awareness and phenomenal consciousness.

## Unconscious Processes

The progress on the experimental front in consciousness research has been dramatic. The situation changed from being one in which the problem was completely ignored (perhaps because it seemed an intractable problem) to one in which substantial resources are spent in research laboratories, producing valuable empirical evidence about the nature of conscious awareness. What paved the way towards this progress was the experimental research on unconscious perception and unconscious cognitive processing. Bernard Baars (1988), for instance, used wellknown unconscious processes (with established research methodologies) to probe the contours of conscious processing. The comparison between the neural correlates of conscious and unconscious processing has already produced crucial insights into the nature of conscious awareness. For example, the thesis that consciousness is the result of a highly integrative process that occurs in a "global
workspace" (Baars, 2002) has been confirmed with neuroscientific evidence (e.g., Dehaene \& Naccache, 2001; Di Lollo, Enns, \& Rensink, 2000).

Another area of advancement is the range of related topics that are studied experimentally with the goal of better understanding consciousness. Experiments on conscious and unconscious perception, binocular rivalry, and mental imagery have expanded our knowledge of perceptual awareness. Additionally, research on the distinction between conscious inclinations for action and the unconscious processes that guide motor control has shown that the processes that reach awareness are indeed just the tip of the cognitive processing iceberg (Rosenbaum, 2002).

## Contents of Consciousness

An important philosophical development that has taken place in the last few years is the incorporation of insights made by psychologists and phenomenologists concerning the content of conscious experience. Susanna Siegel (2006), for example, uses the notion of phenomenal contrasts (a change in how one experiences something) in order to account for the content of conscious vision. This topic in the philosophy of perception concerns our understanding of the difference between conscious perception, illusions, dreams, and hallucinations.

In the history of cognitive psychology, ambiguous images have been considered a paradigmatic case of such contrasts. In the Necker cube (a geometrically ambiguous image that appears to point upward or downward) or the "duck-rabbit" drawing (a semantically ambiguous image that can look like a duck or a rabbit), the stimulus-or perceptual content-does not change but the subject experiences it in one of two alternative ways at a time, and never both at the same time. It is an established finding in vision science that these images alternate at a constant rate, regardless of the intentions of the subject. At first, one interpretation is salient, then it recedes and the other incompatible interpretation becomes the salient one. The subject can also direct her attention, however, and "flip" the interpretations, for example, by focusing on one of the inner corners of the Necker cube (e.g., focusing on the lower inner corner where three edges meet will encourage the ambiguous drawing to be perceived as an upward pointing cube). These attentional contrasts with phenomenological implications show that voluntary and involuntary forms of attention interact with consciously experienced contents in analogous ways.

Other phenomenal changes seem to depend fundamentally on attention, rather than represented content. In discussing the implications of findings on visual attention by Marisa Carrasco and colleagues (e.g., Carrasco, Ling, \& Read, 2004; Carrasco \& Yeshurun, 2009), Block argues that the phenomenal changes in experience, based on changes in attention, are not dependent on either external changes or changes in conceptual aspects of the stimuli (such as semantic ambiguity or expertise). He notes that the quality of these experiences (which he calls "mental paint") feels
"unreal", similar to visual experiences concerning afterimages. Block contends that these findings cannot be explained as illusions because the percept relies on how attention is allocated rather than being a true misrepresentation of the stimuli. The "subjective unreality" of these changes, Block (2010) claims, has not received any empirical investigation. They also remain unaccounted for in a broader theoretical treatment of consciousness. Here is where the study of attention can provide important insights.

## What Is Attention?

Attention research in cognitive psychology is quite active and covers a range of processes-from low-level perceptual systems to high-level cognitive systems. These processes act as "selection" mechanisms to determine what information reaches higher-level cognition, including conscious awareness. In this discussion, we are mainly referring to visual attention, which has the most active research. It is accepted in the scientific community that there are several types of attention comprised of distinct cognitive systems, which have been identified and supported through studies in neuroscience. For example, Posner and Petersen (1990) argued that there are at least three systems that are individually responsible for alerting, orienting, and target detection or executive function (e.g., the top-down processes of visual search). These classifications have held up over years of research, although there is recent evidence for additional attention networks for self-regulation and selfcontrol (Petersen \& Posner, 2012). It is crucial to identify the implications of the various forms of attention on cognition, especially to understand how attention and consciousness are related.

## Bottom-up vs. Top-down Attention

Attention can be stimulus-driven and automatically guided toward important external events that involuntarily catch the focus of attention, or it can be voluntarily guided through willful selection. This distinction is commonly conceived as bottom-up versus top-down processing (see Theeuwes, 2010). That is, attention can be thought of as a process that is exogenous, data-driven, and beyond our control in a cognitively impenetrable manner (bottom-up). This includes pre-attentive mechanisms that are reflexive in nature, such that salient features bias the neural activity for selection into higher processes and can affect behavior without reaching conscious awareness. Alternatively, attention can be described as being endogenous and more deliberate, which biases the competing neural activity in lower-level cognition based on the goals of the current task (top-down). This dichotomy has been challenged recently because there are other forms of attention that do not neatly fall into these categories, such as when learned rewards or habits influence attention (Awh, Belopolsky, \& Theeuwes, 2012). Nevertheless, there are many attentional processes that fall under one of these two descriptions.

## Effortless vs. Effortful Attention

Effortless attention, like bottom-up attention, is thought of as an involuntary, sensory form of attention and does not always reach conscious awareness. These "effortless" processes serve to obtain information from the environment for higher-level representations (which often require more effort to maintain). On the other end of this spectrum is effortful attention, which, like top-down attention, can be described as focused, deliberate, voluntary, or goal-driven and produces the subjective feeling of expending effort. Some complex attentional processes, however, can be so engrossing that they produce the subjective feeling of being involved in a task effortlessly such that one loses a sense of time (Bruya, 2010). It is this latter version of effortless attention that is particularly insightful, which may be related to expertise and is suggestive of how memory systems can interact with attention to influence the perception of effort and time (it is not a straightforward process).

## Varieties of Attention

Beyond the distinctions described above, attention has been characterized under several "varieties". Attention can be feature-based (see Maunsell \& Treue, 2006) and drawn to types of features, generally organized according to specialized regions in the brain that process certain types of sensory information (such as color, motion, or segment orientation). It can also be object-based (Scholl, 2001) and drawn to things in the world that display object-like properties (e.g., cohesion, symmetry). Feature Integration Theory and Object File Theory describe how object-based attention can operate via "object file" representations (Kahneman, Treisman, \& Gibbs, 1992; Treisman \& Gelade, 1980). This is a two stage process that requires the individuation of objects (a bottom-up process) and the identification of the object after a selective attention binds and maintains the features in an object file. This exemplifies the interaction between low-level and high-level forms of attention that makes the study of attention so complex.

Another influential model for attention is the "spotlight model" (Posner, Snyder, \& Davidson, 1980), where attention can be focused on a specific region, or it can be distributed and more diffuse to cover more area (with less detail). This spatial attention operates on empty space or objects in space, and also can quickly determine the "gist" of the information present. Covert attention is a particularly insightful form of attention and refers to the voluntary shift of attention outside the center of one's gaze (Wright \& Ward, 2008). This has been shown through various tasks where a subject views a center of a stimulus display but shifts the focus of attention to the periphery without moving their eyes (or making other physical movements). This type of attention may correspond to the ability to attend to certain thoughts from memory or other mental states that are not immediately linked to sensory information.

Additionally, research on attention has identified peculiar phenomena such as blindsight, inattentional blindness, and the attentional blink. All these describe occasions when
attention fails to perform as expected, particularly because focused attention overlooks targets (e.g., during a search task). The failures may be due to the relevant information not being detected by low-level sensory receptors or could be higher up where it fails to reach awareness. This exemplifies the complexity of the systems that make up the broad term of "attention", which all have the common goal of selecting perceptual information for cognitive processes.

## Conscious Attention

A problem in the empirical study of consciousness is how to identify and explain all the nuances of the theoretical understanding of consciousness at the neural level. For example, even if the experimental evidence confirms that consciousness correlates with a specific pattern of neural activation, what would that finding signify? Could we be able to distinguish access consciousness from phenomenal consciousness? Could it be that the pattern of activation is literally just correlated with consciousness and it neither explains nor identifies what is truly unique about it (i.e., it corresponds to the integration of information but not to the integration mechanism)? Much has been said about this issue, and we will not provide a metaphysical thesis here about the relevance (or lack thereof) of attempts to identify the neural correlates of consciousness. Despite the difficulties underlying the metaphysics of consciousness, we believe that the progress on the experimental front has been substantive. By focusing on the largely unexplored issue of conscious attention, we can outline the general features of an adequate theory of consciousness that would successfully guide future empirical research.

One way of clarifying the relationship between consciousness and attention is by examining the relationship responsible for successful reductions. The spectrum of views that are possible, from most to least restrictive, include: 1) Identity between consciousness and attention, with specific definitions of the kinds of consciousness and attention at stake; 2) Dissociative views of consciousness and attention, where there are several forms of attention without consciousness, but only one form of conscious attention, and attention is a necessary condition for consciousness; 3) Dissociative views that indicate all forms of consciousness are of the same type but that attention is not a necessary condition for consciousness; 4) Dissociative views that indicate there are forms of attention without consciousness but no possible form of consciousness without attention, although there may be many forms of conscious attention; and 5) Full dissociation between consciousness and attention.

The identity thesis for consciousness and attention is the most restrictive of these views and is akin to the reduction of questions about "life" to questions about DNA. According to this view, consciousness just is attention (e.g., Prinz, 2012). There are advantages of this view, but there are also major problems, both theoretical (Koch \& Tsuchiya, 2007) and empirical (Kentridge, 2011). Many of these problems are best understood as possible responses to
two different questions. 1) Are all forms of attention forms of conscious attention? The intuitive response is yes, but the empirical evidence is not clear-cut. 2) The inverse question: are all forms of consciousness forms of conscious attention? Here, things are much trickier and no obvious response seems without problems. The leading intuitions have epistemic or metaphysical flavors, but no leading intuition clearly commands the inquiry. Furthermore, this concerns only theoretical issues-when one looks at the empirical evidence, things are equally tricky. Despite its intuitive strength, the identity thesis is too simplistic to account for such intricate issues as identifying the various systems supporting consciousness and attention-but full dissociation seems to be too strong and so a landscape of options emerges. There seems to be attention without consciousness, for example, as in the case of blindsight. How prevalent are these forms of unconscious attention (i.e., to what extent do they guide cognitive processing)? There may be consciousness without attention and the same consideration about scope is pertinent. Depending on the degree of dissociation, one can envision several possibilities with critical theoretical implications.

What are the possible outcomes? Suppose the degree of dissociation is insignificant. In this case, one could distinguish a few forms of consciousness without attention (or vice versa), but they would be rare cases of little consequence such that one could almost identify consciousness with attention. Yet, even in this case several questions remain. Why would these forms of consciousness without attention (or vice versa) exist? All issues of scope are relevant here. Suppose that all forms of attention are forms of conscious attention but that there are a few cases of consciousness without attention (or attention without consciousness). This possibility would suggest that these forms of consciousness could not be easily integrated with attentional processes, where some forms of consciousness are more resilient to cognitive integration with other processes than others. Or perhaps it is strictly due to there being two fundamental kinds of consciousness.

Suppose, on the contrary, that the degree of dissociation is severe (we focus only on these polar opposites in this paper). Some cases of conscious attention could be associated with what Block calls mental paint, and be highly if not fully dependent upon subjectively unreal attentional contrasts. Other cases of conscious attention could be highly representational and depend on specific mental contents (as attention is generally understood). Finally, other cases of attention could be directed to the conscious self. Of course, there will be many cases in which attention is not accompanied by consciousness (at least phenomenal consciousness) and there will also be cases in which consciousness is not accompanied by attention. The main result would be that consciousness and attention are integrated in some cases, but operate independently from one another. Based on current empirical evidence, however, there is only weak support for consciousness without attention, because there are several types of attention that
must be examined when testing for the presence of consciousness without attention, and studies that claim this dissociation have failed to do so (see Cohen, et al., 2012). Also, this dissociation is unlikely if one accepts the premise that the purpose of attention is to determine what information reaches conscious awareness.

Examining the findings in neuroscience should help clarify the relationship between attention and consciousness. It is accepted that different areas of the brain support different forms of attention. For example, it seems that the cerebellum and other more "primitive" areas of the brain are not necessary for consciousness (Koch, 2012), and yet the cerebellum is crucial for navigation and thus has several areas devoted to attending to features of the environment. Areas associated with emotion, perception, and motivation, which were thought to be deeply related to phenomenal consciousness, are also unnecessary for conscious awareness. So based on the neuroscientific findings, one can make a very plausible case for dissociation. This conclusion has to be evaluated in conjunction with the considerations that led theorists to propose the identity thesis. Furthermore, innovative theories on how consciousness emerges, for example, from recurrent processes in the brain (Lamme, 2006), must be considered in this work.

To advance the understanding of conscious attention, one must provide an integrated account of consciousness and attention based on the latest psychological and neurological findings. By doing so, we can elucidate theoretical distinctions fundamental for an adequate understanding of the conscious mind, such as the distinction between higher forms of self-awareness (e.g., the recognitional self) and more minimal ones (e.g., the phenomenal self). Also, higher forms of attention and consciousness may be associated with the emergence of social interactions within species. That is, as social interactions become more complex (e.g., monogamous mating, social hierarchies, ability to follow gaze), a more sophisticated cognitive system is necessary and this may be correlated to consciousness. Research considering such a social account is also warranted.

## Conclusion

In order to advance the empirical study of consciousness and attention, a concerted effort must be made to unify the two areas in terms of language and goals. Attention research is a promising area for understanding consciousness, especially by clarifying the relationship between consciousness and attention via conscious attention. A main insight from the research findings on attention, which should guide future inquiry, is that attention is mainly concerned with connecting cognitive processing with objects in the external world by processing selective information-it is more analytic and selective in nature than consciousness, which is highly integrative. The past attention research in cognitive psychology, however, presents a challenge for integrated accounts with consciousness, such as the one we pursue here. Most psychologists working on attention had, because of the
intractability of the problem of consciousness mentioned above, either no interest in consciousness or no way to connect their findings with such considerations. Findings on focused attention, divided attention, failures of attention, and other aspects of attention shaped the field without making it explicit how they were compatible with theories of consciousness. Making these connections explicit is another crucial goal for future work that will inform our understanding of conscious attention, and can only emerge from a unified theoretical and conceptual understanding.

## Acknowledgments

This research was partially funded through a Postdoctoral Fellowship from the University of Western Sydney (HHH).

## References

Allport, A. (1993). Attention and control. Have we been asking the wrong questions? A critical review of 25 years. In D. E. Meyer \& S. Kornblum (Eds.), Attention and Performance XIV (pp. 183-218). Cambridge, MA: MIT Press.
Awh, E., Belopolsky, A. V., \& Theeuwes, J. (2012). Topdown versus bottom-up attentional control: A failed theoretical dichotomy. Trends in Cognitive Sciences, 16(8), 437-43.
Baars, B. J. (1988). A cognitive theory of consciousness. Cambridge: Cambridge University Press.
Baars, B. J. (2002). The conscious access hypothesis: Origins and recent evidence. Trends in Cognitive Sciences, 6(1), 47-52.
Block, N. (1995). On a confusion about a function of consciousness. Behavioral and Brain Sciences, 18(2), 227-47.
Block, N. (2010). Attention and mental paint. Philosophical Issues, 20(1), 23-63.
Bruya, B. (2010). Effortless attention: A new perspective in the cognitive science of attention and action. Cambridge, MA: MIT Press.
Carrasco, M., Ling, S., \& Read, S. (2004). Attention alters appearance. Nature Neuroscience, 7(3), 308-13.
Carrasco, M., \& Yeshurun, Y. (2009). Covert attention effects on spatial resolution. Progress in Brain Research, 176, 65-86.
Chalmers, D. J. (1996). The conscious mind: In search of a fundamental theory. New York: Oxford University Press.
Churchland, P. S. (1996). The Hornswoggle problem. Journal of Consciousness Studies, 3(5-6), 402-8.
Cohen, M. A., Cavanagh, P., Chun, M. M., \& Nakayama, K. (2012). The attentional requirements of consciousness. Trends in Cognitive Sciences, 16(8), 411-7.
Dehaene, S., \& Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. Cognition, 79(1-2), 1-37.
Dennett, D. C. (2005). Sweet dreams: Philosophical obstacles to a science of consciousness. Cambridge, MA: MIT Press.

Di Lollo, V., Enns, J. T., \& Rensink, R. A. (2000). Competition for consciousness among visual events: The psychophysics of reentrant visual processes. Journal of Experimental Psychology: General, 129(4), 481-507.
Johnston, W. A., \& Dark, V. J. (1986). Selective attention. Annual Review of Psychology, 37, 43-75.
Kahneman, D., Treisman, A., \& Gibbs, B. J. (1992). The reviewing of object files: Object-specific integration of information. Cognitive Psychology, 24(2), 175-219.
Kentridge, R. W. (2011). Attention without awareness: A brief review. In C. Mole, D. Smithies \& W. Wu (Eds.), Attention: Philosophical and psychological essays (pp. 228-46). Oxford: Oxford University Press.
Koch, C. (2012). Consciousness: Confessions of a romantic reductionist. Cambridge, MA: MIT Press.
Koch, C., \& Tsuchiya, N. (2007). Attention and consciousness: Two distinct brain processes. Trends in Cognitive Sciences, 11(1), 16-22.
Lamme, V. A. (2003). Why visual attention and awareness are different. Trends in Cognitive Sciences, 7(1), 12-8.
Lamme, V. A. (2006). Towards a true neural stance on consciousness. Trends in Cognitive Sciences, 10(11), 494501.

Maunsell, J. H., \& Treue, S. (2006). Feature-based attention in visual cortex. Trends in Neurosciences, 29(6), 317-22.
Nagel, T. (1974). What is it like to be a bat? Philosophical Review, 83(4), 435-50.
Parasuraman, R. (2000). The attentive brain. Cambridge, MA: MIT Press.
Petersen, S. E., \& Posner, M. I. (2012). The attention system of the human brain: 20 years after. Annual Review of Neuroscience, 35, 73-89.
Posner, M. I., \& Petersen, S. E. (1990). The attention system of the human brain. Annual Review of Neuroscience, 13, 25-42.
Posner, M. I., Snyder, C. R., \& Davidson, B. J. (1980). Attention and the detection of signals. Journal of Experimental Psychology, 109(2), 160-74.
Prinz, J. J. (2012). The conscious brain: How attention engenders experience. New York: Oxford University Press.
Rosenbaum, D. M. (2002). Motor control. In H. E. Pashler \& S. Yantis (Eds.), Stevens' handbook of experimental psychology: Sensation and perception (3rd ed., Vol. 1, pp. 315-39). New York: Wiley.
Rosenthal, D. M. (2002). How many kinds of consciousness? Consciousness and Cognition, 11(4), 65365.

Scholl, B. J. (2001). Objects and attention: The state of the art. Cognition, 80(1-2), 1-46.
Siegel, S. (2006). Subject and object in the contents of visual experience. Philosophical Review, 115(3), 355-88.
Theeuwes, J. (2010). Top-down and bottom-up control of visual selection. Acta Psychologica, 135(2), 77-99.
Treisman, A. M., \& Gelade, G. (1980). A feature-integration theory of attention. Cognitive Psychology, 12(1), 97-136.
Wright, R. D., \& Ward, L. M. (2008). Orienting of attention. Oxford: Oxford University Press.

# Discovering Quantification and Number in a Role-Filler Model 

Aaron J. Hamer (ahamer@hawaii.edu)<br>University of Hawaii at Manoa, Linguistics<br>1890 East-West Road, Honolulu, HI 96822 USA

Leonidas A. A. Doumas (leonidas@hawaii.edu)<br>University of Hawaii at Manoa, Psychology<br>2530 Dole Street, Honolulu, HI 96822 USA


#### Abstract

Quantification plays a central role in human reasoning and models thereof, but the discovery and development of quantification remains an open question. We present a theory of how such concepts are learned from experience in the DORA model, a neurally-plausible computational model of relational learning and reasoning (Doumas et al., 2008). The same theory accounts for how concepts of number are acquired in this class of model. We are unaware of any prior model that accounts for the development of both quantification and number from unstructured (e.g., perceptual) input.


Keywords: number; quantification; relational discovery; computational modeling.

## Introduction

Quantification and number are key representational constructs in human cognition. These concepts are foundational in science, mathematics, music, and many other domains of human achievement. Many models of cognition rely on these representational primitives (e.g., any symbolic model that relies on first-order predicate calculus, many Bayesian models such as Piantadosi, Tenenbaum, \& Goodman's (2013) model of quantifier discovery, etc.), but as Carey (2009, p. 456) notes, "There is no proposal I know for a learning mechanism available to non-linguistic creatures that can create representations of objects, number, agency, or causality from perceptual primitives."

These concepts share significant semantic overlap ranging from their function as predicates over sets of objects (Barwise \& Cooper, 1981) to innate, scalar ordering (e.g., one, two, three \& some, many, most; Horn, 1972). Both sets of concepts can be derived from a small set of axioms via set theory (i.e., set membership, identity; Van Heijenoort, 1977). It does not seem unreasonable to consider the problem of their acquisition jointly. While there have been attempts to explain their acquisition in terms of a developmental trajectory from number to quantifiers or vice versa (e.g., Gelman \& Gallistel, 1978; Carey, 2004), we are unaware of any existing model that accounts for the development of representations of both quantification and number from unstructured (i.e., perceptual) input.

## Behavioral Data

## Quantification Facts

Behavioral evidence suggests that there are three broad areas of difficulty with the acquisition of quantification: quantifier spreading, mapping issues, and superlative quantifiers.

Quantifier Spreading Philip and his colleagues (1991a, 1991b) popularized the term quantifier spreading to describe a phenomenon first reported by Inhelder \& Piaget (1964). Children aged six to seven were unable to restrict universal quantifiers to a subset of items present in an array based on a shared feature. When presented with three purple triangles and a purple circle and asked "Are all the triangles purple?" the children would respond in the negative. When asked for an explanation, a typical response was "The circle is purple, too."

Mapping Issues Brooks \& Braine (1996) demonstrated that children have more rigid mappings for the quantifiers all and each than adults. Children preferred a grouped interpretation of all in scenarios such as "All of the roses are in a vase" and a distributed, one-to-one interpretation of each in scenarios such as "Each of the roses is in a vase". They interpreted scenes where roses were distributed over more than one vase as false for the all quantifier and scenes where there was not a one-to-one mapping of roses-to-vases (e.g., more roses than vases, more vases than roses) as false for the each quantifier. Children achieve adult-like performance reasoning about all at around age five but do not reach adult-like performance reasoning about each until age nine.

Superlative Quantifiers Scalar quantifiers can be divided into two types: superlative quantifiers that include their endpoints (e.g., at most three, three or more) and comparative quantifiers that exclude their endpoints (e.g., less than four, more than two). Musolino (2004) showed that five-year-old children performed worse on tasks relying on superlative quantifiers versus comparative quantifiers. Geurts et al. (2010) investigated this phenomenon further and showed that the difficulty of acquiring superlative quantifiers extended to 11 -year-old children. Geurts et al. also showed that superlative quantifiers were more difficult for adults to process (as shown by higher RTs). Hurewitz et al. (2006) found that three-year-olds interpret some as
inclusive of all. This result suggests that some undergoes a transition from a superlative quantifier to a comparative quantifier at some point in development.

## Number Facts

Our discussion of the behavioral data on the acquisition of number will focus on three areas: numerosity and counting, the linear shift, and operational momentum.

Numerosity and Counting Children as young as two-yearsold can subitize, or determine the numerosity of small sets without counting (Gelman \& Gallistel, 1978). However, three-year-olds struggle with the foundations of counting (Grinstead et al., 1997), and have difficulty with cardinality (Wynn 1990, 1992). By three-and-a-half, most children demonstrate exact judgments of numbers up to four and the ability to count to similar magnitudes (Gelman \& Gallistel, 1978; Hurewitz et al., 2006).

The Linear Shift Children initially estimate numerical quantities based on a logarithmic scale before undergoing a shift to using a linear scale at approximately 12 years of age (Siegler \& Opfer, 2003). Logarithmic estimations of quantity are consistent with a perceptual system that obeys the Weber-Fechner law (Fechner, 1860).

Operational Momentum McCrink et al. (2007) showed that adults overestimate sums and underestimate differences, a phenomenon referred to as operational momentum. The pattern of errors fits a Gaussian distribution if magnitudes are represented logarithmically rather than linearly.

## Summary of Behavioral Data

Children struggle with the acquisition of concepts of quantification and number. Some abilities are present early (e.g., subitization at two years) and others develop quickly (e.g., developing counting between ages three and three-and-a-half). Other abilities develop more gradually (e.g., restriction of quantifiers) and some developmental trajectories extend into adolescence (e.g., the linear shift). In some cases earlier points on the developmental trajectory are more compatible with formal logic than the adult norm (e.g., some as a superlative quantifier).

## Developmental Accounts

## Theories of Quantification

Existing accounts of the development of quantification can be grouped into three broad categories: connectionist models, symbolic models, and Bayesian models (e.g., Clark, 1996; Carey, 2004; and Piantadosi, Tenenbaum, \& Goodman, 2013, respectively). Existing connectionist models model the association of externally supplied
symbols such as words with first-order quantifiers. We have not found an account that does not assume pre-existing symbolic representations such as number ${ }^{1}$ (Carey, 2004) or the set theoretic equivalents of number, the existential quantifier, the universal quantifier, or formally equivalent items (i.e., cardinality, non-exhaustion, exhaustion, and membership \& identity, respectively; Piantadosi et al., 2013; Van Heijenoort, 1977).

## Theories of Number

We will examine four classes of models of the acquisition of number: connectionist models, spiking-neuron models, symbolic models, and Bayesian models.

Connectionist Models of Number Existing connectionist models provide an excellent account for the development of subitization via associative learning or summation encoding (e.g., Ahmad, Casey, \& Bale, 2002; Dehaene \& Changeux, 1993; and Verguts \& Fias, 2004). Various models have provided an account for innate ordering via unsupervised competitive recurrent back-propagation networks (e.g., Ahmad et al., 2002) and the association of external symbols with existing representations of number via co-occurrence (Verguts \& Fias, 2004). These models do not address phenomena that occur later in development, nor do they provide an account for the emergence of symbolic representations.

Spiking-Neuron Models of Number These models focus on tying specific abilities or developmental processes to what is known about neuronal behavior. Examples include modeling number as a consequence of gamma oscillations ${ }^{2}$ that predicts subitization behavior that obeys the WeberFechner law (Miller \& Kenyon, 2007) and a tuning function based on neuronal spike trains that accounts for both operational momentum and the linear shift (Prather, 2012).

Symbolic Models of Number Existing symbolic accounts either require "explicit external symbols" (e.g., Carey, 2009) or assume an existing set of quantifier representations (e.g., Gelman \& Gallistel, 1978). While these models account for many developmental phenomena, they openly assume a preexisting cache of symbolic currency to build upon.

Bayesian Models of Number Extant Bayesian models of the acquisition of number share the flaws of Bayesian models of quantification - they assume set theoretic equivalents of number, the existential quantifier, the universal quantifier, or formally equivalent items (i.e., cardinality, non-exhaustion, exhaustion, and membership \& identity, respectively; Piantadosi, Tenenbaum, \& Goodman, 2012; Van Heijenoort, 1977).

[^289]
## Summary of Developmental Accounts

Existing accounts of the development of quantification and number can be grouped into connectionist, symbolic, and Bayesian models. While each class of model has strengths, all existing models fail to account for the development of the symbolic currency such as predicates or set operations that they either map to or build upon. Furthermore, no existing model has accounted for both domains of concepts or all of the key developmental trajectories within a single domain.

## The DORA Model

## Overview

The DORA model is a symbolic connectionist architecture: a computational model using a neural network to store and manipulate structured representations. DORA represents objects and roles in a distributed fashion - that is, as patterns of simultaneous activation over units (analogous to groups of neurons) that represent the semantic features of the item being encoded.

DORA learns structured representations of properties shared between objects by comparing them. Features shared between objects receive input from multiple sources and are isolated via simple Hebbian learning. The resulting representations are comprised of these shared features, are independent of any specific objects, and can be bound to novel objects encountered in the future. ${ }^{3}$ When DORA compares instances of objects searching for another (e.g., a cat searching for a mouse and a sister searching for her brother) it learns representations of searcher (comparing the cat and sister) and sought (comparing the mouse and the brother). When observing a new instance of searching the existing representation of sought can apply (i.e., be bound) to the sought object.

The representations DORA learns are functionally equivalent to single-place predicates that take novel arguments. Although the initial representations that DORA learns contain extraneous features (e.g., the shared features of the cat and sister irrelevant to searcher), comparisons between different instances produce representations that are progressively more refined (i.e., comparing representations searcher learned from different instances causes contextspecific features to wash out).

The DORA model represents multi-place relations by combining sets of these single-place predicates - e.g., after learning representations of searcher and sought they can be combined to form a representation of the multi-place relation searching. If there is anything invariant about a concept (and there must be for us to recognize it), DORA
can learn a structured representation of it.

## Discovery of Quantification and Number

The DORA model learns new representations through a process of iterated comparison of items in the object and role layer, where featural overlaps ${ }^{4}$ are learned as new representations. This process allows for refinement of existing representations by comparing them to other existing representations or new input.

All quantifiers are learned by comparing instances of countable items and extracting numerosity features. There are many accounts of how a connectionist model can acquire basic numerosity features (e.g., Ahmad et al., 2002; Dehaene \& Changeux, 1993; and Verguts \& Fias, 2004); DORA implements a version of the Metric Array Module (Hummel \& Holyoak, 2001) which extracts magnitude features for any metric dimension, such as numerosity or length.


Figure 1: An example of comparing instances of countable items. Note that the featural representation of 3ness is active for higher cardinality sets, at least in quantities where subitization is an effective strategy to extract numerosity features.

Initial comparisons, especially when the arity of compared sets differ, will result in representations of quantifiers such as the all node in Figure 1. Note that the initial representation in this example contains the 3ness node as well. This process allows for the extraction of quantifiers such as all, and through additional experience, quantifiers such as some. The nodes 3ness and all referenced here are purely expository and stand in for the perceptual features that map to these concepts just as the

[^290][^291]nodes for catness and dogness are collapsed representations of the featural invariants present in cats and dogs.


Figure 2: Extraction of the quantity 3.
The same process accounts for the extraction of number representations. As a consequence of this process, concepts that are encountered more frequently (all, one, some) will be learned before concepts that are encountered less frequently (fifteen, at least), and previously learned concepts can be used to bootstrap the learning of future concepts. Eventually, pure conceptual representations of frequently encountered quantifiers and numbers are extracted through repeated comparison.


Figure 3: The resulting representations for frequently encountered quantities.

## Representational Consequences

The representations shown in Figure 3 are pure set representations, suitable for set operations. They can be bound to other relations to create bound sets (solving the quantifier spreading problem, assuming that the cognitive system has developed both these representations and scopelimiting representations and has enough WM to bind them together). There are some other significant consequences of this manner of representation.

Cardinality of the Universal Quantifier All quantifiers are learned through experience; there is never a time when a quantifier is perceived without being predicated over some set. Consequently, the universal quantifier is cardinal. While the cardinality of the universal (and other quantifiers) will change based on the specific context it is experienced or represented in, it will always possess cardinality. This underscores the results from set theory that suggest that numbers and quantifiers are formally equivalent (Van Heijenoort, 1977).

Place-Value Notation Numeral Systems Commonly encountered quantities will be explicitly represented in such a system. It is likely that specific quantifiers for the numbers one through ten exist in such a system. However, it is extremely unlikely that such a system learns a specific representation for quantities such as 347 . However, such representations can be built form the representational currency of lower-order numbers such as three, four, and seven, and a representation for place that takes on features of the base of the numeral system (e.g., 10 for Arabic numerals) and magnitude of the base (e.g., two for the hundreds place), and so on.

## The Way Forward - Count on DORA to Quantify Development

Our theory of quantification and number development handles three major issues not addressed in current models. First, we account for both domains within a single model using a small set of principles (e.g., comparison-based learning, building complex representations from singleplace predicates) and processes. Furthermore, we provide an account for how these symbolic representations are developed and structured as a consequence without drawing from an existing cache of symbolic currency. Finally, our model accounts for a wide variety of developmental trajectories within each domain using the same set of basic parameters and processes, as well as a wide variety of other developmental trajectories.

## Unifying Quantification and Number

One of the core goals of framing the acquisition of quantification and number within the DORA framework is to provide a unified account of their development. Unifying both domains as opposite endpoints of a developmental trajectory has been attempted (e.g., Gelman \& Gallistel, 1978; Carey, 2004) but such attempts fail to account for the intertwined developmental trajectories as they are built on assumptions of mastery within a domain as a foundation on which to build mastery of the other. The most successful Bayesian modeling attempts to account for the development of quantification and number are currently instantiated as separate Bayesian models built on the same set of priors (Piantadosi et al., 2012, 2013). While unifying Bayesian models built on the same set of priors is relatively simple, it remains to be done.

Our account of the development of quantification and number captures key developmental trajectories in both domains as a consequence of comparison-based learning
iterating over previously learned concepts and new experience. The interactions between the developmental trajectories of quantification and number are captured because they arise as a natural consequence of learning both domains at the same time. These interaction effects forced us to deal with both domains simultaneously as modeling either quantification or number learning in isolation failed to account the developmental facts for either domain. DORA cannot model either quantification or number in isolation as successfully as it can account for both together.

## Symbolic Structure Developed, Not Borrowed

Most accounts of cognition fail to explain where the structured symbolic representations they use to solve problems come from. Such structures range from predicates, set operators, and even quantifiers and cardinality. The core function of the DORA model is to extract invariance from unstructured (e.g., perceptual) input via comparison. Using a comparison-based learning mechanism not only explains how such structure arises, but also what this structure looks like. This mechanism creates the representations that many models rely upon.

Bayesian models of development rely on an external source of structured symbols to build a foundation upon. While Bayesian models provide an excellent way to model competency, when modeling development they run into more fundamental issues than failing to account for where the structures they rely on come from. The most successful Bayesian models of the development of quantification and number competency in people (i.e., Piantadosi et al., 2012, 2013) rely on priors that are a superset of the concepts they claim to develop. Put simply, they start with the assumption that people can already count to three and use the quantifiers for existence, some, all, and none. We find it difficult to characterize a model as developmental when it assumes its outputs as priors.

## Modeling Developmental Trajectories

We have provided a brief overview of how DORA learns cardinality and number from experience, but we have not yet laid out how our model handles the developmental trajectories at play.

DORA begins subitizing using the Metric Array Module, a simple, neurally plausible mechanism that could easily be available to two-year-old children. This mechanism outputs magnitude judgments that obey the Weber-Fechner law. Logarithmic judgments of magnitude explain why children treat numbers and analogous quantifiers such as some as superlative quantifiers initially because a point on a logarithmic scale corresponds to a range on a linear scale. As DORA is exposed to many instances of small sets (as children are) it quickly learns to represent small cardinal numbers explicitly. These explicit representations do not rely on logarithmic magnitude features; consequently, children no longer treat these numbers as superlative quantifiers.

Children gain working memory as the prefrontal cortex matures. Quantifier spreading disappears as children are able to marshall the working memory needed to build the complex representations required to simultaneously bind a
quantifier to a scope-limiting representation and match that representation to a particular situation. The representations for cardinal numbers continue to develop throughout childhood as larger and larger numbers become explicitly represented, accounting for the linear shift in early puberty.

We account for all of these developmental facts with a single set of parameters and simple processes. DORA also accounts for over 35 findings surrounding the development of relational thinking (Doumas \& Hummel, 2010; Doumas et al., 2006; Doumas et al., 2008; Sandhofer \& Doumas, 2008), including the relational shift (Rattermann \& Gentner, 1991), the development of relational representations (Smith, 1984), and the development of shape bias (Abecassis et al., 2001).

## Conclusion

Our proposal is a promising account of how concepts of quantifiers and number can be learned from perceptual input. The DORA model's working memory constraints allow a developmental trajectory to be modeled, and make specific predictions about how specific types of quantified reasoning will fail based on working memory demands, such as differing magnitudes of $n$-back tasks. We are exploring these predictions with human participants. Crucially, our model accomplishes these goals using the same parameters and processes that have allowed us to successfully account for more than 35 developmental phenomena in other domains.

## Acknowledgments

The first author would like to thank Daft Punk for the Alive 2007 set.

## References

Abecassis, M., Sera, M. D., Yonas, A., \& Schwade, J. (2001). What's in a shape? Children represent shape variability differently than adults when naming objects. Journal of Experimental Child Psychology, 78, 303-326.
Ahmad, K., Casey, M., \& Bale, T. (2002). Connectionist simulation of quantification skills. Connection Science, 14 (3), 165-201.

Barwise, J., \& Cooper, R. (1981). Generalized quantifiers and natural language. Linguistics and Philosophy, 4, 159219.

Brooks, P. J. \& Braine, M. D. S. (1996). What do children known about the universal quantifiers all and each? Cognition, 60, 235-268.
Carey, S. (2004). Bootstrapping and the origins of concepts. Daedalus, 133(1), 59-68.
Carey, S. (2009). The origin of concepts. New York, NY: Oxford University Press.
Clark, R. (1996). Learning first order quantifier denotations: An essay in semantic learnability. IRCS Technical Report 96-19.
Dehaene, S. \& Changeux, J. (1993). Development of elementary numerical abilities: A neuronal model. Journal of Cognitive Neuroscience, 5(4), 390-407.

Doumas, L. A. A. \& Hummel, J. E. (2010). A computational account of the development of representations underlying object recognition. Cognitive Science, 34, 698-712.
Doumas, L. A. A., Hummel, J. E., \& Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. Psychological Review, 115, 1-43.
Fechner, G. T. (1860). Elemente der Psychophysik. Leipzig: Breitkopf and Hartel.
Gelman, R., \& Gallistel, C. R. (1978). The Child's Understanding of Number. Cambridge, MA: Harvard University Press.
Geurts, B., Katsos, N., Cummins, C., Moons, J., \& Noordman, L. (2010). Scalar quantifiers: Logic, acquisition, and processing. Language and Cognitive Processes, 25(1), 130-148.
Grice, H. P. (1975). Logic and conversation. In P. Cole \& J. Morgan (Eds.), Syntax and semantics: Vol. 3. New York, NY: Academic.
Grice, H. P. (1989). Studies in the way of words. Cambridge, MA: Harvard University Press.
Grinstead, J., MacSwan, J., Curtiss, S., \& Gelman, R. (1997). The independence of language and number. In Greenhill, A., Hughes, M., Littlefield, H., \& Walsh, H. (Eds.), Proceedings of the 22nd Boston University Conference on Language Development, 1. Somerville, MA: Cascadilla.
Horn, L. (1972). On the semantic properties of the logical operators in English. (Unpublished doctoral dissertation). University of California, Los Angeles, CA.
Hummel, J. E., and Holyoak, K. J. (2001). Toward an understanding of analogy within a biological symbol system. In D. Gentner, K. J. Holyoak, \& B. N. Kokinov (Eds.), The Analogical Mind: Perspectives from Cognitive Science. Cambridge, MA: MIT Press.
Hurewitz, F., Papafragou, A., Gleitman, L., \& Gelman, R. (2006). Asymmetries in the acquisition of numbers and quantifiers. Language Learning and Development, 2(2), 77-96.
Inhelder, B., \& Piaget, J. (1964). The Early Growth of Logic in the Child. London: Routledge and Kegan Paul.
Knowlton, B. J., Morrison, R. G., Hummel, J. E., \& Holyoak, K. J. (2012). A neurocomputational system for relational reasoning. Trends in Cognitive Sciences, 16(7), 373-381.
McCrink, K., Dehaene, S., \& Dehaene-Lambertz, G. (2007). Moving along the number line: operational momentum in nonsymbolic arithmetic. Perception \& Psychophysics, 69 (8), 1324-1333.

Miller, J. A., \& Kenyon, G. T. (2007). Extracting numberselective responses from coherent oscillations in a computer model. Neural Computation, 19(7), 1766-1797.
Musolino, J. (2004). The semantics and acquisition of number words: Integrating linguistic and developmental perspectives. Cognition, 93, 1-41.
Philip, W. \& Aurelio, S. (1991). Quantifier spreading: Pilot study of preschooler's "every". In Maxfield, T. \& Plunkett, B. (Eds.), Papers in the Acquisition of WH: Proceedings of the UMass Roundtable, May 1996. Amherst, MA: Graduate Linguistics Student Association.

Philip, W. \& Takahashi, M. (1991). Quantifier spreading in the acquisition of "every". In Maxfield, T. \& Plunkett, B. (Eds.), Papers in the Acquisition of WH: Proceedings of the UMass Roundtable, May 1996. Amherst, MA: Graduate Linguistics Student Association.
Piantadosi, S. T., Tenenbaum, J. B., \& Goodman, N. D. (2012). Bootstrapping in a language of thought: A formal model of numerical concept learning. Cognition, 123, 199-217.
Piantadosi, S. T., Tenenbaum, J. B., \& Goodman, N. D. (2013). Modeling the acquisition of quantifier semantics: a case study in function word learnability. Manuscript submitted for publication.
Prather, R. (2012). Connecting neural coding to number cognition: a computational account. Developmental Science, 15(4), 589-600.
Rattermann, M. J. \& Gentner, D. (1998). More evidence for a relational shift in the development of analogy: Children's performance on a casual-mapping task. Cognitive Development, 13, 453-478.
Sandhofer, C. M. \& Doumas, L. A. A. (2008). Order and presentation effects in learning categories. Journal of Cognition and Development, 9, 194-221.
Siegler, R. S. \& Opfer, J.E. (2003). The development of numerical estimation: Evidence for multiple representations of numerical quantity. Psychological Science, 14, 237-243.
Smith, L. B. (1984). Young children's understanding of attributes and dimensions. Child Development, 55, 363-380.
Van Heijenoort, J. (1977). From Frege to Gödel: a source book in mathematical logic, 1879-1931. Cambridge, MA: Harvard University Press.
Verguts, T., \& Fias, W. (2004). Representation of number in animals and humans: a neural model. Journal of Cognitive Neuroscience, 16, 1493-1504.
Wynn, K. (1990). Children's understanding of counting. Cognition, 36, 155-193.
Wynn, K. (1992). Children's acquisition of number words and the counting system. Cognitive Psychology, 24, 220251.

# Disambiguating NN combinations with left/right stress 

James A. Hampton (hampton@city.ac.uk), Daniel Heussen (daniel@heussen.be), Zarah Argel (zarah_jayne@live.co.uk), Hasina Kanbi (hasina_k91@hotmail.com)<br>Department of Psychology, City University London, Northampton Square, London EC1V OHB, UK


#### Abstract

An interesting subclass of Noun-noun combinations in English can take two meanings depending on whether the first or second word is stressed in speech. A BRICK factory is one that makes bricks, whereas a brick FACTORY is one made of brick. An explanation is offered in terms of a bias for nouns from particular ontological categories to trigger particular semantic interpretations for a combination, together with the proposal that the unstressed noun provides the relation to be used. The explanation is tested in three empirical studies.


Keywords: NN combination, concepts, compounds, meaning, ambiguity, stress

## Introduction

Noun-noun compounds are found in many languages. They are the result of placing two nouns together in order to create a compound noun phrase with a new meaning (Gleitman \& Gleitman, 1970; Levi, 1978). The first nounthe modifier-serves to specialise the meaning of the second noun-the head. To take the example of cheese knife, the head noun knife determines the kind of thing involved-a cheese knife is a knife-while the modifier cheese specifies that it is designed for use with cheese. ${ }^{1}$

Looking at noun-noun compounds in English it is possible to differentiate those that are constructed on the fly to meet the communicative needs of a given moment from those that long ago entered the lexicon. A good example of the former is Downing's (1977) famous apple-juice seat to indicate the seat at a table at which the person drinking apple-juice should sit, whereas examples of the latter are toothpaste or lunch box. As they enter the lexicon, compounds in English may be lexicalized as single orthographic words (e.g., snowman), or optionally hyphenated (e.g., pigeon-hole), while others have two-word spellings (e.g., taxi driver). Lexicalization involves a number of changes in the status of a phrase. In particular, lexicalized compounds are also more likely to lack semantic transparency, and hence to appear in dictionaries. The meaning of such non-transparent phrases must be acquired through hearing them in context, rather

[^292]than being computable from the individual meanings alone. Examples like beer garden, water glass, or wine lake cannot be easily understood without appeal to knowledge external to the meaning of the two individual words.

In addition to the many lexicalized compounds in English, it is also possible to create new forms which can be readily understood. People have little trouble understanding new expressions like camel field, or student gardener, even though they may not have come across these combinations before. An explanation for this productivity was suggested by Gagné and Shoben (1997) in their CARIN model of Noun-Noun (NN) combinations. Earlier, Levi (1978) had proposed that the majority of compounds employ one of about 12 semantic relations, such as USE, LOCATION, MATERIAL COMPOSITION or CAUSE. Gagné and Shoben suggested that each noun in a person's vocabulary may be associated in memory with the relations that it most commonly enters into, as either a modifier or a head. Thus for example mountain as a modifier would normally indicate a location as in mountain goat, or mountain village. On the other hand magazine as a head would normally use an informational "about" relation as in train magazine or psychology magazine. When the two are put together to form mountain magazine, a search is instigated to find a plausible meaning. Since a magazine about mountains strikes most people as more plausible than a magazine in the mountains, in this case the head noun ends up dictating the preferred meaning of the phrase.
In the search for a plausible meaning, two equally plausible meanings can sometimes arise, each based on one of the two nouns and their preferred semantic relations, so that the compound is ambiguous. Kamp and Partee (1995) cite the example of a brick factory, which can either mean a factory that makes bricks or a factory that is made of brick. They also point out that the stress pattern employed when speaking the phrase can disambiguate its meaning. Thus one can compare (1) and (2):
(1) a BRICK factory $=$ a factory that makes bricks
(2) a brick FACTORY = a factory made of brick.

The account offered by Kamp and Partee is that the ambiguity relates to the use of two distinct syntactic forms. They suggest, following Bloomfield, (1933: 228) and Chomsky and Halle's (1968: 15-18) Compound Rule, that left stress is a general signal in English that the phrase is a compound, meaning that its semantics will depend on local
context and the argument structure of the head noun. Note how BEER garden, WATER glass, and WINE lake all take left stress patterns, indicating this compound structure. On the other hand, right stress typically indicates a simpler modifier+head noun phrase. For example a black BIRD is a bird that is black, in contrast to the left-stressed, lexicalized, compound BLACKbird, which in British English refers to the common species of garden bird, Turdus merula.
The meaning of the compound phrase BRICK factory requires an understanding that the concept of factory takes an argument of the kind of thing made in it, with the modifier noun placed into this slot, as in (3).
(3) FACTORY (makes X)

X $=$ \{jam, brick, car, clothing,.. $\}$
On the other hand brick FACTORY is interpreted by treating the noun "brick" as a modifier meaning "made of brick". This account is not entirely satisfactory, since the question of when a NN combination is a "true" compound is hard to make on purely semantic grounds. For example why should the meaning of "factory" not also have an argument equivalent to [made of] in which the noun "brick" could be placed? The occurrence of left stress in compounds is also not as clear or reliable as one might hope (see Bell \& Plag, 2010). While most single orthographic compounds do take left stress (e.g. SUNflower, TOOTHpaste, ICEcream) other highly familiar forms written as two-word phrases take right stress (plum JAM, pumpkin PIE). Nor is it that a given semantic relation such as MADE OF seen in these last two examples is consistently right stressed. Thus in British Received Pronunciation cake (unlike pie) is typically left stressed (GINGER cake, CHOCOLATE cake). It is unclear why the ingredients of jams and pies should be syntactically adjectival while the ingredients of cake should require an argument structure (a point noted by Lees, 1962: 120). Given these difficulties, in this paper we will refer to all NN combinations simply as compounds.

A large corpus-based analysis by Plag (2006) using a variety of models found the assignment of stress in spoken English to be largely unpredictable. The best means of prediction was by using analogy with other similar compounds (e.g. OIL painting, FINGER painting, ACTION painting), suggesting a role for similarity-based generalisation in the assignment of stress (Plag, 2010). Further work by Plag and colleagues has identified evidence for a semantic basis, and for families of semantically similar compounds taking the same stress (Plag et al. 2008). Bell \& Plag (2010) also reported that relative informativeness can direct stress on to the more informative of the two nouns.

The placement of stress in speech generally is clearly a very complex phenomenon (for a review see Cutler et al. 1997). There has been relatively little research on prosody in the psychological literature in relation to the interpretation of compounds. A study by Lynott and Connell (2010) manipulated spoken stress for an arbitrary set of novel NN compounds, and found that dual emphasis differentially speeded the generation of property interpretations (e.g. a zebra mussel as a striped mussel).

However they reported no effects of stress on the frequency with which different interpretations were generated, even though many of the compounds had more than one interpretation.

In this paper we focus on one particular use of stress in the context of the interpretation of compounds. Specifically, we ask how ambiguous compounds such as brick factory are disambiguated with the help of stress. Why should left versus right stress direct interpretation in two different directions? Our proposal is that stress indicates which of the two nouns provides the semantic relation for interpretation of the phrase. In particular we propose that it is the unstressed noun that determines the relation.

Recall the example of brick factory. The claim is that brick as a modifier will typically invoke a [MADE OF] interpretation, as in brick house, brick building, brick wall. On the other hand factory will invoke a [MAKES] relation, as in car factory, hat factory or furniture factory. Placing the stress on the modifier brick thus gives the relation preferred by the head (a factory that makes bricks), whereas stressing the head noun factory gives the relation derived from the modifier (a factory made of brick).

To make the principle operational and testable we needed some means to be able to generate ambiguous compounds where the ambiguity depended on two competing relations, one derived from the modifier and one from the head. Both interpretations needed to be plausible meanings for the written phrase (that is, as read in the absence of auditory stress information). Rather than depend on a frequency analysis of individual words occurring in either position, as Gagné and Shoben (1997) proposed, we adopted a suggestion from Maguire, Wisniewski \& Storms (2010) who proposed, on the basis of a corpus study of semantic patterns in compounding, that preferred semantic relations follow from the general ontological category into which a noun falls, rather than being individual to each noun. Thus brick is a member of the category of compositional materials, along with jam, water, cork, plastic etc. All of these will have a preference as modifiers for a MADE OF relation. Similarly factory belongs to a category of sources or origins of objects, from which the relation MAKES will naturally follow. The idea of interpreting NN compounds by recourse to a superordinate semantic categorization of nouns has had much support, particularly in the domain of automatic processing of natural language (Rosario \& Hearst, 2001), although no agreed semantic taxonomy has yet been developed.

Our strategy in creating a set of ambiguous compounds for testing was therefore first to find individual examples, and then to generate further examples using the same superordinate categories. The results of this process can be seen in Table 1 which shows the analysis of the compounds into general categories, together with examples of the materials used. The classification is necessarily fairly broad and provisional, but it serves to illustrate the analogies between, say, an airplane magazine and a church painting, both of which take either a LOCATION or an ABOUT
relation, and it provides some systematicity to our search for suitable examples. We predicted that with left stress, the compounds will have an interpretation of magazine (or painting) ABOUT an airplane or a church, while with right stress, the location relation will be dominant, yielding a magazine or painting LOCATED IN an airplane or church.

To assess the validity of our analysis, we put our predictions to the test. Experiment 1 and 2 provided participants with the ambiguous compounds spoken with either left or right stress, and asked them to write down their interpretation of the meaning. Experiment 3 provided participants with the spoken phrases, and then asked them to judge the plausibility of a given interpretation which could either match or mismatch the interpretation predicted for the stress pattern. We predicted that stressing the modifier or head would influence both the interpretations generated in Experiment 1 and 2 and the speed and accuracy of judgments in Experiment 3.

## Experiment 1

The purpose of Experiment 1 was to test our proposal that the unstressed word in an ambiguous compound should be the one from which the semantic relation will be derived. Hence a chimpanzee DRAWING should be a drawing done by a chimpanzee (taking the agency from the animate modifier noun), while a CHIMPANZEE drawing will be a drawing of a chimpanzee (taking the "information source about" relation from the head noun).

## Method

Participants Twenty-four students at City University London participated. All were native speakers of English.
Materials Forty ambiguous compounds were recorded spoken by a female voice in Received Pronunciation British English. The compounds were selected from a larger sample through pretesting. Participants in the pre-test read each compound (i.e., no stress information was given) and wrote down an interpretation. Items were then selected where just two alternatives were generated, each by at least $25 \%$ of the participants. For the final selection, on average the more frequent meaning was generated $57 \%$ of the time, and the less frequent meaning $35 \%$ of the time. Each compound was
recorded once with stress on the first word, and once with stress on the second word. To reduce the likelihood that the ambiguity of phrases would be noticed, 12 unambiguous fillers were included, two at the start, and ten more distributed every three to five trials through the rest of the experiment. They were familiar phrases like book bag and oak table.
Procedure Participants were seated in front of a PC, and wore sound insulating headphones, through which the speech samples were played. Instructions were displayed on the screen as follows:

In the present study we are investigating the meaning of so-called noun-noun phrases, phrases consisting of two nouns, such as "park bench". We are interested in your intuitive understanding of these phrases. For each noun-noun phrase, we would like you to write down its meaning in the textbox provided. A short description of the meaning that first comes to mind is sufficient. For example, when you hear "park bench", the first thing that might come to mind might be: A bench in a park.
Once the instructions were understood the trials began. Each trial began with the playing of the recording of a phrase. A button on the screen allowed the participant to repeat the playback if they wished to hear it again. If the replay button was clicked three times, a window appeared displaying the phrase (very occasionally people had trouble hearing the words spoken). Being written, no cue was given as to the stress pattern. A text box was provided on screen into which the participant typed their interpretation of the phrase. A "NEXT" button took them after a short pause to the next screen and a new recorded phrase.

Design Participants were divided into four groups of 6. Two groups had 20 compounds with left stress and 20 with right, while the other two groups had the alternative. In addition two different random orders were used.

## Results

One item was omitted owing to an error in the programming. The results were based on the remaining 39 items. The interpretations entered by participants to each

|  |  | Modifier-based <br> Relation | Head-based <br> Relation | Examples |
| :--- | :--- | :--- | :--- | :--- |
| AGENT/ PATIENT | AGENT | H who is M | H for M | athlete lawyer, celebrity doctor |
| AGENT/ PATIENT | ACTION/EVENT | H by/from M | H for/to M | company award, dolphin strategy |
| AGENT/ PATIENT | INFORMATION SOURCE | H produced by M | H about M | politician novel, chimpanzee drawing |
| LOCATION | AGENT | H comes from M | H done of M | Iceland painter |
| LOCATION | INFORMATION SOURCE | H found in M | H about M | airplane magazine, church painting |
| MATERIAL | INSTRUMENT | H made of M | H for making M | ceramic oven |
| MATERIAL | CONTAINER | H made of M | H contains M | clay bucket, wax pot |
| MATERIAL | INFORMATION SOURCE | H made of M | H about M | chocolate book, paper catalogue |
| MATERIAL | MATERIAL | H made of M | H for M | juice dye, plant poison |
| INSTRUMENT | ACTION/EVENT | H done using M | H done to M | dollar purchase, skateboard damage |

[^293]phrase were collated into a single table, without any information about the original stress pattern that they had heard. The first two authors worked through the table independently to code the interpretations as either (a) the meaning derived from the modifier, (b) the meaning derived from the head, or (c) other or unclassifiable. Initial agreement between the judges was $90 \%$. Disagreements were resolved by each judge reconsidering the disputed interpretations in the knowledge of the other judge's rating. Any remaining disagreements were treated as "other". (There were $13 \%$ of responses that could not be classified.) The predicted effect of stress was borne out in the data across 39 items, with more modifier meanings resulting from right stress (19.8) than from left stress (14.2), and more head meanings resulting from left stress (18.6) than from right stress (15.1). Overall, $57 \%$ of responses were as predicted by stress. The effect was highly reliable across participants, with 21 of 24 following the prediction on average, and none against it ( $\mathrm{p}<.001$, sign test). Across items the effect was less strong statistically, with 25 of 39 compounds following the prediction and 12 against ( $\mathrm{p}=.01$, sign test).

## Experiment 2

The effect in Experiment 1 was relatively small, with stress inducing a bias in interpretation of $57 \%$ versus $43 \%$. Experiment 2 was a replication in which we tested whether a new selection of materials and an improved quality of the sound recordings might show a stronger effect.

## Method

Participants Twenty-four students participated for course credit.

Materials Ambiguous compounds were constructed as in Experiment 1, with 40 compounds, 25 of which were new to this study. (The effect size in Experiment 1 for the 15 items used in both studies was identical to the overall mean effect size for that study, so these were not retained just on the basis of their being "good" items in terms of results.) New recordings were made, under improved recording conditions using a sound-proof studio and high quality microphone and recorder. The speaker had a London accent, more familiar to the student participant pool than was the RP used in Experiment 1. In addition when creating the recordings, to help the speaker produce meaningful stress patterns we used contrastive stress to generate the left versus right stress patterns. The speaker first read out a sentence such as "It's not a CLAY pot it's a ... WAX pot", while pausing before the last two words. In a second recording, the speaker read out the sentence "It's not a wax CANDLE, it's a ... wax POT". All speech except for the final two words of each sentence was then edited out to leave just the final two word phrase for use in the experiment. Hence participants in the experiment proper had no access to the contrastive meaning used in generating the spoken phrases, but just heard each
phrase either with left or with right stress.
Design and Procedure The design and procedure was identical to Experiment 1.

## Results and Discussion

Responses were classified as before. The effect size was considerably increased. For left stress, across the 40 items there were on average 27.8 head meanings and only 11.3 modifier meanings, while with right stress the means were reversed with 14.3 head meanings and 24.0 modifier meanings. The proportion of all responses in line with prediction increased from 57\% in Experiment 1 to $67 \%$ in Experiment 2. Across items, 34 out of 40 (85\%) showed the predicted effect, and only 6 went against the hypothesis. Across participants, 20 (83\%) showed the predicted effect, and only 3 went against. (Both, $\mathrm{p}<.001$ on a sign test).
Experiment 2 strengthened the evidence for our hypothesis. With a new selection of items and improved recording of the stress patterns, the effect size was greatly increased. The relatively weak consistency across items in Experiment 1 can probably be attributed to problems in the original recorded materials. For the set of 15 items used in both experiments effect size correlated across experiments at $.56(\mathrm{p}<.05)$. In the first experiment these items had the same effect size as the remaining items. In Experiment 2 their effect size increased in line with the other new materials ( 8.6 for the 15 retained items, and 7.3 for the 25 new items), supporting the effect of the improved audio recordings.

## Experiment 3

If stress assignment directs interpretation in the way we propose, then it should be easier to judge that a particular given interpretation is plausible for a spoken compound if the interpretation being judged is consistent with the stress pattern used. In Experiment 3, participants heard the same phrases as in Experiment 1 with either left or right stress. They were then immediately given a written interpretation, which was either one of the two plausible meanings, or a new implausible one. When the interpretation was plausible, it could either match that predicted from the assigned stress, or mismatch it (i.e. match the alternative interpretation). We predicted that trials on which a match occurred should lead to faster and more accurate responding.
(Because Experiment 3 was conducted before Experiment 2, the materials and recordings were the same as in Experiment 1).

## Method

Participants Initially 60 students at City University London participated in the study, of which 11 were replaced as they made more than $50 \%$ errors on all trials taken together.

Materials The same 40 spoken word phrases were used as in Experiment 1. Since the programming error for one item
was not detected in time, only 39 of the phrases could be used for the analysis. Fillers were included at the start (the first 5 trials), and throughout the sequence so that not all the phrases were ambiguous. In addition a set of 40 implausible meanings for familiar compound nouns was constructed (e.g., box office "an office about a box", bus seat "a seat that is a bus".)

Procedure Participants heard the spoken phrase over headphones, and then after a delay of 2 seconds saw an interpretation of the phrase on the screen. They had to decide as quickly as possible whether it was a plausible meaning or not. The following instructions appeared on the screen at the start of the experiment:
"In the present study we are investigating the meaning of so-called noun-noun phrases, phrases consisting of two nouns, such as "park bench". We are interested in how long it takes to understand different phrases. You will be presented with a spoken phrase, and shortly after you will see a possible meaning on the screen. If you think the meaning makes sense, then press the ALT GR key (on the right of the space bar). If it doesn't make sense as a meaning for the phrase, then press the ALT key on the left of the space bar. The first five trials are for practice, so feel free to ask if you don't understand what you are supposed to be doing. After that we would like you to proceed, making your responses as fast as you can while not making any errors. The experiment will take 10 to 15 minutes."

Design There were two random orders or presentation and two assignments of stress to each spoken compound. In addition the interpretation offered for judgment could be either the modifier-based or the head-based interpretation.

## Results and Discussion

Mean correct reaction times and error rates were calculated for Matching and Mismatching plausible trials for each participant and each item. Three RTs of over 10 seconds were removed from the analysis entirely, and another 18 reaction times of over 3 standard deviations above the mean for individual participants were truncated. Table 2 shows the results for RT and Error rates. When the interpretation to be judged as meaningful was consistent with the stress assignment, responses were on average 100ms faster and about $10 \%$ more accurate. Two 2-way ANOVA were run for RT and Errors with stress assignment and interpretation as factors, and with participants and items as random effects. (The error distribution for error rates was normal, skew $=-0.1$, matching the assumptions of ANOVA.)

Although RT showed the predicted interaction ( 2338 ms for matching and 2438 ms for mismatching meanings), it failed to reach significance ( $\mathrm{p}=.15$ ). However Errors showed the predicted interaction as significant ( $26 \%$ for matching and $31 \%$ for mismatching trials), with $\mathrm{F}(1,59)=$ $6.0, p=.018$ by subjects, $F(1,38)=4.3, p=.044$ by items.

No main effects were significant. (For filler implausible trials, mean RT was 2300 ms , sd $=771$, and the error rate was $17 \%$.)
While supportive of our hypothesis, the procedure in Experiment 3 is clearly less sensitive to the effects of spoken stress, requiring as it does a "sensicality" judgment. The weak effects may (as in Experiment 1) have reflected some difficulty that some participants may have had in clearly perceiving the spoken phrases. The high error rate may also be owing to this factor. It is also possible that some participants may have focussed on identifying the nonsensical meanings, which would allow them to ignore the spoken stress all together.

|  | MEANING |  |  |  |
| :--- | :--- | :--- | :--- | :--- |
|  | Resp. Time (ms) |  | Error (\%) |  |
| STRESS | Modifier | Head | Modifier | Head |
| Left | 2453 | 2324 | 30.5 | 25.3 |
|  | $(952)$ | $(833)$ | $(18.1)$ | $(17.0)$ |
| Right | 2353 | 2424 | 27.7 | 30.9 |
|  | $(943)$ | $(869)$ | $(16.2)$ | $(17.7)$ |

Table 2. Mean (SD) for response times and errors for Experiment 3

## General Discussion

In these studies we have sought to find evidence for our explanation of how stress assignment can disambiguate the interpretation of NN compounds. We showed in Experiments 1 and 2 that spontaneously generated meanings were influenced in the predicted direction by stress, and this result was supported in Experiment 3 with evidence that online processing of a potential meaning of a phrase was similarly affected by hearing a spoken phrase with the stress on the left or right word, at least in terms of error rates.
The results were typical of psycholinguistic data, in that the main effect of interest was (to various degrees) obscured by other factors and noise in the data. Using a speaker with a London accent, and a procedure for generating the spoken phrases that used contrastive stress ("It's not a CLAY pot it's a WAX pot) produced a marked increase in the size of the predicted effect in the second experiment. It is also very possible that different speakers are differentially responsive to the influence of stress. Both Experiments 1 and 2 found a bimodal distribution across participants. For example in Experiment 2, 16 participants had effects ranging from 43\% to $65 \%$, but the other 8 participants were in the range $87 \%$ to $100 \%$.
The principle explanation that we offer is that the unstressed word in an ambiguous NN compound determines the semantic relation. In light of the role of stress in directing attention, this principle may at first appear paradoxical. One might suppose that attention should be directed towards the noun that is "doing the work". However stress in spoken language is also often used to direct attention to the focus or new information in an utterance (Bell \& Plag, 2010; Bock \& Mazzella, 1983;

Clark \& Haviland, 1977). The difference between "I phoned my mother on FRIDAY" and "I phoned my MOTHER on Friday" is a matter of whether the focus of the utterance is the date of the call or the person called. In the case of our proposed principle, a similar analysis can be made. In an ambiguous compound, one of the nouns provides the background schema from which the general meaning will be derived, and the other provides the highlighted information placed into that schema. Take an example like chimpanzee drawing. Is the issue a matter of what kinds of things chimpanzees get up to, or is it about the kinds of things that get drawn? If the former, then chimpanzee goes unstressed as the new information is that they do drawings. If the latter, then drawing is the unstressed background schema, and the new information is that it is a chimpanzee that is the subject of the drawing. The two interpretations can be (loosely) represented thus;

## (4) A chimpanzee DRAWING

CHIMPANZEE .. [type of creature \{mammal, primate..\}]
.. [activities \{swinging from trees, hooting, DRAWING...\}]
(5) A CHIMPANZEE drawing

DRAWING.. [implements needed \{pencil, paper ...\}] .. [subject \{scene, still life, CHIMPANZEE...\}]
There may therefore be close parallels between the different roles that stress can play within discourse processes and in compound interpretation.

Our result is also consistent with Plag et al.'s (2008) finding that stress assignment is often constant across families of similar compounds, based on the similarity of either head or modifier nouns (see also Plag, 2010). Semantically similar concepts tend to have similar preferred relations, and so enter into similar patterns of stress assignment. For example location modifiers and material modifiers typically take right stress in unambiguous compounds, and there are other cases where given semantic relations are associated with particular stress direction. However, there must be other factors (such as historical accident) at work, as the example given in the introduction of the different stress assignment for pies versus cakes clearly demonstrates. More recently Bell and Plag (2010) have reported that stress can also be predicted from the relative informativeness of the two nouns in a compound, with the more informative being stressed. Our principle fits well with this idea. The unstressed noun sets up a background schema into which the stressed noun is placed as the new information.

The principle that we have described helps to shed some light on at least one aspect of the use of stress patterns in English. It remains to be seen how broadly the principle can now be applied outside of the realm of ambiguous NN compounds.

## References

Bell, M., \& Plag, I. (2010). Informativeness is a determinant of compound stress in English. Unpublished MS from

DFG-Project website, http://www2.uni-siegen.de/~engspra / DFG-Project/Bell-Plag-2010.pdf, viewed March 1, 2011.
Bloomfield, L. (1933). Language. New York: Henry Holt and Co.
Bock, J. K. \& Mazzella, J. R. (1983). Intonational marking of given and new information: Some consequences for comprehension. Memory \& Cognition, 11, 64-76.
Chomsky, N. \& Halle, M. (1968). The sound pattern of English. New York: Harper and Row.
Clark, H. H. \& Haviland, S. E. (1977). Comprehension and the given-new contract. In R.Freedle (Ed.), Discourse production and comprehension (pp. 1-40). Norwood, NJ: Ablex.
Cutler, A., Dahan, D., \& van Donselaar, W. A. (1997). Prosody in the comprehension of spoken language: A literature review. Language and Speech, 40, 141-202.
Downing, P. (1977). On the creation and use of English compound nouns. Language, 53, 810-842.
Gagné, C. L. \& Shoben, E. J. (1997). Influence of thematic relations on the comprehension of Modifier-Noun combinations. Journal of Experimental Psychology: Learning, Memory, and Cognition, 23, 71-87.
Gleitman, L. R. \& Gleitman, H. (1970). Phrase and paraphrase. New York: Academic Press.
Kamp, H. \& Partee, B. (1995). Prototype theory and compositionality. Cognition, 57, 129-191.
Lees, R. D. (1963). The grammar of English nominalizations. Bloomington IN: Indiana University Press.
Levi, J. (1978). The syntax and semantics of complex nominals. New York: Academic Press.
Maguire, P., Wisniewski, E. J., \& Storms, G. (2010). A corpus study of semantic patterns in compounding. Corpus Linguistics and Linguistic Theory, 6, 49-73.
Plag, I. (2006). The variability of compound stress in English: structural, semantic and analogical factors. English Language and Linguistics, 10, 143-172.
Plag, I. (2010). Compound stress assignment by analogy: The constituent family bias. Zeitschrift für Sprachwissenschaft,29, 243-282.
Plag, I., Kunter, G., Lappe, S., \& Braun, M. (2008). The role of semantics, argument structure, and lexicalization in compound stress assignment in English. Language, 84, 760-794.
Rosario, B. \& Hearst, M. (2001). Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP2001). Pittsburgh, PA.

## Acknowledgments

The authors acknowledge Thoslima Khanam and Roberta Pischedda who helped with data collection, Peter Barr for programming, and Sophie Lind who recorded the spoken phrases for Experiment 1. Helpful discussion was provided by Deema Awad, Herbert Clark, Eve Clark, Paul Egré, Christina Gagné, Kate Maclean, Gregory Murphy, Francois Recanati, Thomas Spalding and Gert Storms.

# Improving First-Year Writing Using Argument Diagramming 

Maralee Harrell (mharrell@andrew.cmu.edu)<br>Carnegie Mellon University, Department of Philosophy, 5000 Forbes Ave. Pittsburgh, PA 15213 USA<br>Danielle Wetzel (dfz@andrew.cmu.edu)<br>Carnegie Mellon University, Department of English, 5000 Forbes Ave.<br>Pittsburgh, PA 15213 USA


#### Abstract

There is substantial evidence from many domains that visual representations aid various forms of cognition. We aimed to determine whether learning to construct visual representations of argument structure enhanced the acquisition and development of argumentative writing skills within the context of first-year college writing course. We found a significant effect of the use of argument diagrams, and this effect was stable even when multiple plausible correlates were controlled for. These results suggest that natural-and relatively minor-modifications to standard first-year composition courses could provide substantial increases in student writing ability.


Keywords: argument diagramming; argument mapping; writing; critical thinking; graphic organizers.

## Introduction

The purpose of the First-Year Writing (FW) Program at Carnegie Mellon University is to develop the academic reading and writing skills each student needs to be successful in his or her college career. Each student at CMU must take the course Interpretation and Argument, which is the core of this writing program.

Thus, though not titled 'Critical Thinking,' the FW course taken during the first year is generally one of the student's first introductions to thinking critically at a college level. Among other goals, the specific learning objectives for the FW Program is for students to be able to: (a) analyze a written argument by identifying the conclusion and the premises (both implicit and explicit) and describe how the premises support the conclusion, (b) evaluate a written argument by determining whether the premises do in fact support the conclusion, and whether the premises are reasonable, and (c) write an essay that both analyzes and evaluates one or more arguments.

The over-arching goal for the FW course is to provide foundational reading and writing skills that will enable students to develop advanced literacy in their own disciplines.

Most educators agree that one aspect of "critical thinking" involves the ability to reconstruct, understand and evaluate an argument-cognitive tasks we may describe as 'argument analysis' (see, e.g., Ennis, 1987; Fisher \& Scriven, 1997; Kuhn, 1991). In college, the most common medium through which arguments are analyzed is writing. Interpretation and Argument is a research-based course that
understands that reading and writing are inseparable practices for college-level course work. In the course, students are exposed to a variety of different texts (mostly academic essays) so they can explore a single issue from multiple perspectives and eventually contribute an argument of their own to the discussion. Both the exploration and the contribution rely heavily on argument analysis at various stages.

The first step in this analysis is reading a text for the argument, as opposed to, for example, reading for the plot (as in a novel) or for the facts (as in a textbook). Mandler (1984) provides an overview of research supporting the claim that adults and children as young as 3 -years-old possess "story schemata" that guide understanding when reading or listening to a story. Thus, learning the skill of reading for the argument requires students to develop a new schema, or set of schemata, with which they can interpret the text appropriately.

Schema theory, first introduced by Bartlett $(1932,1958)$ and further developed by Evans (1967), Mandler (1984) and Rumelhart and Ortony (1977), explains cognition as information processing mediated by schemata. A schema is a packet of knowledge containing both data and information about the interconnections among the data. Rumelhart (1980) refers to schemata as the representations of concepts stored in memory, and Sweller (1994) describes schemata as representations of either concepts or problem-solution procedures.

To facilitate the acquisition of new schemas, Sweller (1994) recommends reducing the extraneous cognitive load during the learning process. One common way of reducing extraneous cognitive load is by using graphic organizers (GOs), such as diagrams, to supplement regular reading and instruction. Previous research has shown that students' use of GOs is generally efficacious in producing improvements on a wide range of cognitive tasks - including those generally labelled CT tasks - that are significantly higher than improvements gained by students engaged in reading and regular instruction alone (Horton, et al., 1993; Moore \& Readance, 1984). Thus, we are particularly interested in the efficacy of alternative teaching methods that incorporate GOs to increase argumentative writing performance.

In what might these alternative methods consist? Both Larkin and Simon (1987) and Winn (1991) argue that diagrammatic representations of information can make recognition of important features and drawing inferences
easier than a sentential representation of the same information. Indeed, research on student learning has consistently shown the efficacy of using diagrams to aid text comprehension (Armbruster \& Anderson, 1984; Dansereau, et al.; Novak \& Gowin, 1984; Schwartz \& Rafael, 1985), as well as vocabulary development, postreading activities and writing preparation (Johnson, et al., 1986).

One candidate alternative teaching method, then, is instruction in the use of argument diagrams as an aid to argument comprehension and evaluation (see Figure 1).


Figure 1: Example of a diagram for a simple argument.
If we think of an argument the way that philosophers and logicians do-as a series of statements in which one is the conclusion, and the others are premises supporting this conclusion-then an argument diagram is a visual representation of these statements and the inferential connections between them.

How does argument diagramming develop new schema? The argument diagramming curriculum consists in an online course introducing argument diagramming, followed by inclass and weekly homework assignments on representing the arguments in the course materials in diagrams. The students received oral and written feedback on their diagramming. The students are taught to discriminate between statements (or claims) and other kinds of sentences, as well as the difference between arguments and explanations. The students are also taught to look for words that indicate conclusions (e.g., 'thus' and 'therefore'), premises (e.g., 'because' and 'since'), linked arguments (e.g., 'but' and 'since') and convergent arguments (e.g., lists). All of these types of exercises help students develop an 'argument schema' for reading arguments in a variety of genres.

Recent research on the efficacy of an argument diagramming curriculum on the development of critical thinking skills includes studies on both philosophy students in introductory classes and a mix of undergraduates in critical thinking and informal logic classes. The former studies have shown that instruction that includes the use of argument diagrams to analyze, evaluate and create arguments significantly improves students' critical thinking skills over the course of a semester (Harrell, 2008, 2011, 2012).

The latter studies specifically on computer-supported argument visualization have shown that the use of software specifically designed to help students construct argument diagrams significantly improves critical thinking abilities over the course of a semester (Kirschner, Shum, and Carr 2003; Twardy 2004; van Gelder, Bissett, \& Cumming,
2004). Additionally, research in this area has shown that student's critical thinking about specific topics is improved if students collaborate on argument diagram instruction instead of working alone (Scheuer, McLaren, Harrell, \& Weinberger, 2011). This previous research, however, has all focused on performance on critical thinking skills testsespecially multiple choice tests like the California Critical Think Skills Test-and not on writing tasks.

Even so, we conjectured, that incorporating argument diagramming into our standard curriculum in Interpretation and Argument would help students develop their argumentative writing skills.

Hypothesis: Students who are able to construct argument diagrams and use them during argument analysis tasks will improve in performance on argumentative writing tasks over the course of a semester long composition class significantly more than students in the same class who do not have this ability.
Our first-year writing course was a natural place to study the skills acquisition of our students. We typically teach 2830 sections of this course each semester, with a different instructor for each section. While the general curriculum of the course is set, including the sequence of assignments, each instructor is free to choose the readings for his or her section. The students who take this course are a mix of all majors from each of the seven colleges across the University. This study tests this hypothesis by comparing the pretest and posttest scores of students in Interpretation and Argument who were taught argument diagramming to the scores of those students who were not during the Fall of 2009, and the Spring and Fall of 2010.

## Method

## Participants

Eighty-one students (39 women, 42 men) across 7 sections of Interpretation and Argument were studied. In each semester, each section of the course had a different instructor and the students chose their section. Over the three semesters there were 7 different instructors. The students taught by Instructors 2, 6 and 7 were taught the use of argument diagrams to analyze the arguments in the course readings, while the students in the other sections were taught more traditional methods of analyzing arguments.

## Materials and Procedure

We developed a pretest to be taken at the beginning of the semester, and a companion posttest to be taken at the end. For the next three semesters, students in both the treatment and control groups completed the pretests during the first week of the semester, and the posttest during the last week of the semester. Each test consisted in reading some text and completing two tasks. In Task 1, the student was asked to write an essay analyzing the argument presented by the author in the text. This analysis was to consist in identifying both the content and the structure of the argument. In Task

2, the student was asked to write an essay evaluating the argument presented by the author in the same text. The evaluation was to consist in a claim about the quality of the argument, and reasons to support that claim.

## Results

## Salient Features of Students' Writing

We recognize that text features alone do not constitute "good writing" and that there is no "right way" to read or write a text. We also recognize that privileging some text features over others might ignore other significant features. The features that we chose will help us locate change in demonstrable critical thinking between the pretest and posttest. We analyzed the texts for markers of text development and text coherence. We were interested in seeing to what extent there would be any kind of change in how many different ideas students could generate-about someone else's argument and about their own arguments. Within this category of "development," we identified the following for both Tasks 1 and 2 of the pre- and posttests: the number of different reasons or premises offered for the argument conclusion, and the number of counterarguments considered within the text.

For Task 1, we wanted to determine how much the students were understanding the argument in the text and what statements they would prioritize in their representations of it. For Task 2 only, we also considered whether students provided evidence or elaboration of their reasons. We wanted to distinguish between reasons that were supported with evidence and those that were not. Our concern was instances when students produced a lot of different ideas but failed to support them; we did not want to report "growth" in development without attempting to represent to what extent students were actually supporting their claims.

Because the number of ideas alone does not necessarily equate with good writing, and, in fact, one could argue that too many different ideas within an argument will result in chaos for a reader, we also looked for features that signaled an overall coherence in a written text. Vande Kopple has defined coherence as "prose in which nearly all the sentences have meaningful connections to sentences that appear both before and after them" $(1989,2)$. We also draw upon Enkvist's definition of coherence, "the quality that makes a text conform to a consistent world picture and is therefore summarizable and interpretable" (1990, 49). So, by coherence, we mean those features that enable a reader to make particular kinds of connections within the text. In coding Task 1, we considered the following as coherence markers: logical connections between premises and the argument conclusion, and logical connections between different premises

In coding Task 2, we looked at the following as markers of coherence: logical connections between premises and the argument conclusion, logical connections between different premises, and metacommentary (or "metadiscourse").

Metacommentary is language that writers use, according to Hyland (2003), to compose a text that is clear to a reader.

By providing linguistic "signposts" to readers, writers can create the effect that a text is coherent and holds together in an intentional way. Because these bits of language give clues for making sense of the text, their presence in a text can indicate that a writer is aware of a reader's needs for navigating the text successfully. These bits of language can also show that a writer understands his or her own text in particular ways and can point to a writer's strategic view of his or her writing. We were only interested in the effect that metacommentary has upon the readers-we were not interested in counting the different types. Therefore, coders scored Task 2 holistically for effective use of metacommentary.

## Test Coding

Pretests and posttests were paired by student, and singletest students were excluded from the sample, resulting in 81 pairs of tests. The tests were coded during one extended session, using one set of coders for Task 1, and a different set for Task 2 . Each coder independently coded all pairs of tests in his or her group ( 162 total tests). Each pre-/post-test pair was assigned a unique ID, and the original tests were blinded. To ensure reliability and validity, prior to each coding session, we had an initial coding-calibration session in which we and the coders coded several of the unpaired tests, discussed the codes, and came to a consensus about each code. After this, each coder was given the tests to be coded in a unique random order.

The categories to be coded for Task 1 were: Argument Conclusion, Counter-arguments, Premises, Connections and Errors. "Argument Conclusion" received a code of 1 if the student identified the conclusion of the argument, and a code of 0 if not. "Counter-arguments" received a code that indicated how many counter-arguments the student identified in the text. "Premises" received a code that indicated how many premises the student identified in the text. "Connections" received a code that indicated how many connections between premises or between a premise and the conclusion the student identified in the text. Finally, "Errors" received a code that indicated how many errors the student made; errors identified by the coders were (a) misunderstands counter-argument, (b) missing a major concept, (c) misreading (e.g. overstatement with no qualifiers), (d) misapplied quotation that shows disconnected reading, and (e) other.

The categories to be coded for Task 2 were: Conclusion, Premise, Evidence, Mismatch, Connections, Counterarguments, and Metacommentary. "Conclusion" received a code of 1 if the student stated a thesis, and a code of 0 if not. "Premises" received a code that indicated how many premises the student used in support of the thesis. "Evidence" received a code that indicated how many premises were supported by evidence. "Mismatch" received a code that indicated whether, for each premise, the evidence offered actually supported that premise.
"Connections" received a code that indicated how many connections between premises or between a premise and the conclusion the student identified in the text. "Counterarguments" received a code indicating how many counterarguments the student considered. Finally, "Metacommentary" received a code of 0 if there was no metacommentary, 1 if the metacommentary was present but weak, and 2 if the metacommentary was strong. Then, for each task, the codes from the two coders on these categories were averaged, allowing for a more nuanced scoring of each category than either coder alone could give.

For each task, the primary variables of interest were the individual averages for each category on the pretest and the posttest. In addition, however, the following data was recorded for each student: the student's math, writing and verbal scores on the SAT, the section in which the student was enrolled, the student's final grade in the course, the student's home college, the student's sex, and whether the student had been taught using the AD curriculum.

## Student Characteristics

To determine whether the students in the study differed in any statistically significant characteristic other than being taught AD, we tested how well we could predict students' gains from pre-test to post-test based on the variables we had collected. We performed a regression for Gain using Pretest, Instructor, Gender, Final Grade, College, Math, Verbal, and Writing as regressors. The results indicate that none of the variables besides Pretest and Instructor was a factor in a student's gain. Thus, we are confident that the students in the treatment group were not different in any important aspect from the students in the control group.

## Comparison of Students by AD Instruction

Our hypothesis was that the students in the first-year writing course who received training in Argument Diagramming would gain significantly more in each category on the two tasks than students who did not receive the training. Since the use of argument diagrams was explicitly taught only by Instructors 2, $6 \& 7$, this hypothesis was tested by determining whether the average gain of the students taught by Lecturers $2,6 \& 7$ was significantly different from the average gain of the students taught by Lecturers $1,3,4 \& 5$. The students taught by Lecturers $2,6 \& 7$ are represented in all the tables below by (AD), and the students taught by Lecturers 1, 3, $4 \& 5$ are represented by (No AD). The mean gain for the subpopulations of students in each treatment group is represented given in Figure 2 for Task 1, and in Figure 3 for Task 2.

To determine the predictive value of AD treatment on a student's gain from pretest to posttest, an ANCOVA was conducted for the gains in each category for Task 1 with AD as a factor and the corresponding pretest score as a covariate. So, for example, we conducted an ANCOVA on the Argument Conclusion Gain with AD as a factor and the

Argument Conclusion Pretest as a covariate. The results for Task 1 are given in Table 1.


Figure 2: Comparisons of gains in each category of Task 1 from pretest to posttest for students who were and were not taught argument diagramming.


Figure 3: Comparisons of gains in each category of Task 2 from pretest to posttest for students who were and were not taught argument diagramming.

Table 1: ANCOVA test results for the variable AD for each category on Task 1.

| Category | $F(1,80)$ | $p$ |
| :--- | :---: | :---: |
| Argument Conclusion | 2.47 | 0.120 |
| Counter-arguments | 0.94 | 0.335 |
| Premises | 4.54 | 0.036 |
| Connections | 7.35 | 0.008 |
| Errors | 6.91 | 0.010 |

The effect of AD was statistically significant in each category except Argument Conclusion and Counterarguments for Task 1.

An ANCOVA was also conducted for the gains in each category for Task 2 with AD as a factor and the corresponding pretest score as a covariate. The results for Task 2 are given in Table 2.

Table 2: ANCOVA test results for the variable AD for each category on Task 2.

| Category | $F(1,80)$ | $p$ |
| :--- | :---: | :---: |
| Argument Conclusion | 1.80 | 0.184 |
| Premises | 5.63 | 0.020 |
| Evidence | 6.70 | 0.012 |
| Mismatches | 12.36 | 0.001 |
| Connections | 12.35 | 0.001 |
| Counter-arguments | 5.73 | 0.019 |
| Metacommentary | 10.60 | 0.002 |

The effect of AD was statistically significant in each category except Argument Conclusion for Task 2.

## Discussion

## Findings

The results from Task 1 show that, when reading an argument, students who were taught argument diagramming were significantly more likely than those who were not to identify more of the relevant premises offered that support the author's conclusion, and explain more explicitly how the premises are supposed to work together to support the conclusion. In addition, these students were much less likely to make any errors in their analysis.

The results from Task 2 show that, when evaluating the argument in a text, students who were taught argument diagramming improved significantly more than those who were not in their ability to (a) provide more premises to support their own thesis, (b) offer more evidence in support of each premise (c) have fewer mismatches between premises and evidence, (d) explain more explicitly how the premises are supposed to work together to support the conclusion, (e) offer possible counter-arguments, and (f) provide metacommentary on their response.

Thus, it seems that students who were taught argument diagramming are developing new schema for reading arguments, and learning how to effectively translate this into their own writing. This is reflected most noticeably in the improvement of the metacommentary from pretest to
posttest. We conclude that incorporating argument diagramming into the curriculum of Interpretation and Argument is positively beneficial to realizing several of our course objectives.

## Educational Importance

The primary educational importance of this study is twofold. First, the results indicate that it is possible to significantly improve students' argumentative writing skills over the course of just one semester, even when the course is not only a critical thinking course. Second, these results indicate that a relatively small addition to the curriculum of a first-year writing course can have dramatic benefits for students. The initial instruction in understanding arguments and creating argument diagrams can be given in one or two class-periods (or an online tutorial) and regular, weekly homework assignments can be added to reading, summary and/or reflection assignments. Supplementing one's teaching with argument diagramming does not require a radical reworking of the syllabus, course readings or assignments. This is a great benefit to instructors who may be reluctant to change a curriculum that has been successful.

## Future Work

This study raises as many questions as it answers. While it is clear that the introduction of argument diagramming to the First-Year Writing Program curriculum significantly improves a student's ability to reach several stated course objectives, it would be interesting to explore further the cognitive basis for the effect of argument diagramming. In particular we would like to know what aspects of constructing diagrams help the most in developing new schema.

It would also be interesting to explore whether, once a student learns how to construct argument diagrams, the actual construction of a diagram is important for a particular analysis task. That is, for example, it could be that the new schema is in place, and so the diagrams are no longer needed, or it could be that the construction of a diagram while reading activates the new schema.

We would also like to consider whether there are other skills that we did not measure this time that this addition may help to improve. For example, because our work here did not distinguish between first and second language learners, we cannot speak to whether argument diagramming has more or less of an effect upon second language learners. Additionally, we have anecdotal evidence from several teachers that using argument diagramming during the peer review process was helpful. It would be extremely useful to know whether using argument diagramming in peer review of papers in general makes subsequent drafts better.

Lastly, unlike the relatively solitary activities in which students engage in our FW Courses-like doing homework and writing essays-there are many venues in and out of the classroom in which students may engage in the analysis and evaluation of arguments in a group setting. These may
include anything from classroom discussion of a particular author or topic, to group deliberations about for whom to vote or what public policy to implement. In any of these situations it seems as though it would be advantageous for all members of the group to be able to visually represent the structure of the arguments being considered. We would like to know whether knowing how to construct argument diagrams would aid groups in these situations.

## Acknowledgments

I would like to thank three anonymous reviews for their helpful comments. This study was funded by a joint grant from the Spencer Foundation and the Teagle Foundation, under the title "Systematic Improvement of Undergraduate Education in Research Universities."

## References

Armbruster, B.B., \& Anderson, T.H. (1982). Mapping: Representing informative text graphically. In C.D. Holley \& D.F. Dansereau, (Eds.), Spatial learning strategies. New York: Academic Press.
Bartlett, F. C. (1932). Remembering: A study in experimental and social psychology. Cambridge: Cambridge University Press.
Bartlett, F. C. (1958). Thinking: An experimental and social study. New York: Basic Books.
Dansereau, D.F., Collins, K.W., McDonald, B.A., Holley, C.D., Garland, J., Diekhoff, G., \& Evans, S.H. (1979). Development and evaluation of a learning strategy program. Journal of Educational Psychology, 71, 64-73.
Enkvist, N. E. (1990). Seven Problems in the Study of Coherence and Interpretability. In U. Connor \& A. M. Johns (Eds.), Coherence in writing: Research and pedagogical perspectives. Alexandria, VA: Teachers of English to Speakers of Other Languages, Inc.
Ennis, R. H. (1987). A taxonomy of critical thinking dispositions and abilities. In J. B. Baron \& R. J. Sternberg (Eds.), Teaching thinking skills: Theory and practice. New York: W. H. Freeman \& Co.
Evans, S. H. (1967). A brief statement of schema theory. Psychonomic Science, 8, 87-88.
Fisher, A., \& Scriven, D. (1997). Critical thinking: Its definition and assessment. CA: Edgepress.
Harrell, M. (2012). Assessing the Efficacy of Argument Diagramming to Teach Critical Thinking Skills in Introduction to Philosophy. Inquiry, 27(2), 31-38.
Harrell, M. (2011). Argument Diagramming and Critical Thinking in Introductory Philosophy. Higher Education Research \& Development, 30(3), 371-385.
Harrell, M. (2008). No Computer Program Required: Even Pencil-and-Paper Argument Mapping Improves Critical Thinking Skills. Teaching Philosophy, 31, 351-374.
Horton, P.B., McConney, A.A., Gallo, M., Woods, A.L., Senn, G.J., \& Hamelin, D. (1993). An investigation of the effectiveness of concept mapping as an instructional tool. Science Education, 77(1), 95-111.

Hyland, K. (2003). Second language writing. Cambridge: Cambridge University Press.
Johnson, D.D., Pittleman, S.D., \& Heimlich, J.E. (1986). Semantic mapping. Reading Teacher, 39, 778-783.
Kirschner, P.A., Shum, S.J.B., \& Carr, C.S. (Eds.). (2003). Visualizing argumentation: Software tools for collaborative and educational sense-making. New York: Springer.
Kuhn, D. (1991). The skills of argument. Cambridge: Cambridge University Press.
Larkin, J.H., \& Simon, H.A. (1987). Why a diagram is (sometimes) worth ten thousand words. Cognitive Science, 11, 65-99.
Mandler, J. M. (1984). Stories, scripts, and scenes: Aspects of schema theory. Hillsdale, NJ: Lawrence Erlbaum Associates.
Moore, D. W., \& Readence, J. E. (1984). A quantitative and qualitative review of graphic organizer research. Journal of Educational Research, 78(1), 11-17.
Novak, J.D., \& Gowin, D.B. (1984). Learning how to learn. New York: Cambridge University Press.
Rumelhart, D. E. (1980). Schemata: The building blocks of cognition. In R. J. Shapiro, B. C. Bruce \& W. F. Brewer (Eds.), Theoretical issues in reading comprehension: Perspectives from cognitive psychology, linguistics, artificial intelligence, and education (pp. 33-58). Hillsdale, NJ: Lawrence Erlbaum Associates.
Rumelhart, D. E., \& Ortony, A. (1977). The representation of knowledge in memory. In R. C. Anderson, R. J. Spiro \& W. E. Montague (Eds.), Schooling and the acquisition of knowledge. Hillsdale, NJ: Lawrence Erlbaum Associates.
Scheuer, O., McLaren, B., Harrell, M., \& Weinberger, A. (2011). Will Structuring the Collaboration of Students Improve Their Argumentation? In G. Biswas, S. Bull, J. Kay, \& A. Mitrovic (Eds.), Lecture notes in computer science: Artificial intelligence in education- $15^{\text {th }}$ international conference. Berlin: Springer-Verlag.
Schwartz, R.M., \& Raphael, T.E. (1985). Concept of definition: A key to improving students' vocabulary. Reading Teacher, 39: 198-205.
Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. Learning and Instruction, 4, 295-312.
Twardy, C.R. (2004). Argument maps improve critical thinking. Teaching Philosophy, 27, 95-116.
van Gelder, T., Bissett, M., \& Cumming, G. (2004). Cultivating expertise in informal reasoning. Canadian Journal of Experimental Psychology, 58, 142-152.
Vande Kopple, W. (1989). Clear and coherent prose: A functional approach. Glenview, IL: Scott, Foresman, \& Company.
Winn, W. (1991). Learning from maps and diagrams. Educational Psychology Review, 3: 211-247.

# Why verbalization of facial features increases false positive responses on visuallysimilar distractors: A computational exploration of verbal overshadowing 

Aya Hatano (hatano.aya@c.mbox.nagoya-u.ac.jp), Taiji Ueno (taijiueno7@gmail.com), Shinji Kitagami (kitagami@cc.nagoya-u.ac.jp), and Jun Kawaguchi (kawaguchijun@nagoya-u.jp)<br>Department of Psychology, Graduate School of Environmental Studies, Nagoya University, Furo-cho, Chikusa-ku, Nagoya City, Aichi 4648601, JAPAN


#### Abstract

Verbal overshadowing refers to a phenomenon whereby verbalization of a non-verbal stimulus (e.g., he had slant eyes) impairs subsequent non-verbal recognition accuracy. In order to understand the mechanism by which this phenomenon occurs, we constructed a computational model that was trained to generate an individual-face-specific representation upon input of a noise-filtered retinotopic face (i.e., face recognition). When the model verbalized the facial features before receiving the retinotopic input, the model incorrectly recognized a new face input as one of the different, yet visually-similar, trained items (that is, a false-alarm occurred). In contrast, this recognition error did not occur without prior verbalization. Close inspection of the model revealed that verbalization changed the internal representation such that it lacked the fine-grained information necessary to discriminate visually-similar faces. This supports the view that verbalization causes unavailability/degradation of finegrained non-verbal representations, thus impairing recognition accuracy.


Keywords: verbal overshadowing; face recognition; computational modeling; verbalization

## Introduction

Language is the principal medium for carrying out daily communications. This is still true when communicating our non-verbal experiences, such as recounting a crime scene we have witnessed, or describing the physical appearance of a criminal. Particularly, if we do not have a record of the event such as a picture or video, then conveying an eyewitness memory relies on language. A crucial question in cognitive science, therefore, is the influence of verbalization on non-verbal memory. Many studies have revealed that language has extra-communicative functions, in that it affects such cognitive functions as perception, learning, and memory. For example, in a seminal study by Schooler and Engstler-Schooler (1990), participants watched a video of a bank robbery for 30 seconds and following which half of the participants described the appearance of the bank robber. Subsequently, all of the participants were shown a line-up that consisted of the bank robber's photo and seven distractors. Results revealed that participants who had verbalized the bank robber's appearance were worse at recognizing the target individual than those who had not, a phenomenon known as verbal overshadowing. The procedure of these experiments can be experienced beyond an experimental setting. For example, during criminal investigations, an eyewitness may provide a statement describing the appearance of a criminal and
subsequently identify them from a line-up. In such situations, it is crucial to prevent a false accusation and to examine the credibility of the eyewitness's testimony. Therefore, it is both theoretically and practically important to clarify the mechanism by which verbal overshadowing occurs. For this purpose, we constructed a paralleldistributed processing (PDP) model to simulate the effect of verbalization on subsequent visual recognition.

A closer review of the literature allows us to gain further insight into this phenomenon and therefore to establish a more specific aim for our model. First, although not all of the past studies have split the recognition scores into positive and negative trials, false alarm is sometimes more susceptible to verbalization before recognition than hit rates; that is, participants often inaccurately identify distractors as a target rather than miss a correct target (Meissner, Brigham, \& Kelley, 2001). Furthermore, recognition accuracy in this study was positively correlated with accuracy of the verbal description prior to recognition. Based on these observations, Meissner et al. proposed a recording interference account that assumed verbalization rendered the representations less accurate (compared to visual representations), thus impairing subsequent visual recognition.

Second, Kitagami, Sato, \& Yoshikawa (2002) revealed that verbal overshadowing is also sensitive to the degree of similarity between targets and distractors (manipulated with a morphing technique). Specifically, verbalization impaired subsequent visual recognition only when distractors were highly similar to the target (using a 9-alternative choice task with a "not present" response choice), but the impairment disappeared when similarity was low. It is also worth noting that this manipulation involved a change in the distractors, but not in the target picture itself. We revisited the original data and revealed that accuracy was impaired due to the more frequent choice of a distractor (a false alarm) rather than an incorrect choice of "not present" (a miss). Schooler (2002) explained this result with the transfer inappropriate processing shift hypothesis. This hypothesis assumes that verbalization induces a processing shift from visual to verbal, and that a shift to verbal processing makes finegrained non-verbal information about faces unavailable. This non-verbal information is crucial for discriminating the target from others (see also, Maurer, LeGrand, \& Mondroch, 2002), especially in a high-similarity condition (Kitagami et al., 2002). Although Schooler's hypothesis does not necessarily assume a correlation between recognition accuracy and verbal description accuracy (see also Kitagami
et al., 2002; Fallshore \& Schooler, 1995), this hypothesis and the recording interference hypothesis by Meissner et al. (2001) share two ideas: First, both assume that fine-grained non-verbal information is necessary for face recognition. Second, both expect that a verbal representation which is generated during verbalization lacks such fine-grained information, thus impairing visual recognition.

More recently, Clare and Lewandowsky (2004) introduced an alternative hypothesis, arguing that verbalization shifts the criterion threshold such that participants say "The target is not present in the display" more frequently when in fact the target is present. Although this account can explain a range of existing data, two issues deserve consideration. First, even when a "not present" response was disallowed (that is, responses were forced choice), verbal overshadowing was observed in some studies, especially when elaborative verbalization was encouraged (Fallshore \& Schooler, 1995). Second, the shifting criterion hypothesis cannot explain the fact that false alarm is more susceptible to verbalization than hit rate (Kitagami et al., 2002; Meissner et al., 2001). Thus, as Clare and Lewandowsky also speculated, there may be two mechanisms by which verbalization impairs subsequent visual recognition: One is shifting-criterion (Clare \& Lewandowsky, 2004), and the other is degradation (Meissner et al., 2001) or unavailability (Schooler, 2002) of fine-grained non-verbal representations crucial for face recognition, especially when a distractor is visually confusing. This study focused on the latter possibility, and investigated how the nature of representations changes upon verbalization, and how this affects subsequent visual recognition. Computational modeling is an effective approach for this purpose. An explicitly implemented computational model allows a modeler to directly look at the nature of computations/representations that are underpinning a simulated behavior. The PDP model here was trained for three facial processing tasks: One was to represent the retinotopic input of a face in a non-verbal format (visual encoding/recognition); a second was to activate the correct units for verbal labels upon the same retinotopic input of a face (verbal encoding); the third was to represent a face in a non-verbal format upon verbal inputs (the mental imagery of a face upon verbal cues). After being trained for these tasks, the model was forced to activate some verbal units (i.e., verbalization), and we investigated how this forced activation changed the nature of the computed representation in the model, and how it affected subsequent visual encoding of a retinotopic input.

## Method

## Model Architecture

Figure 1 shows the architecture of the PDP model, built with LENS software (http://tedlab.mit.edu/~dr/Lens/). Three peripheral layers were connected bi-directionally with a single hidden layer. In order to reduce the computational


Figure 1: Architecture of the model (Hinton diagram).
demand in this large model, units between layers were connected sparsely, such that a unit was not connected with others if the external input/target value of that unit was always zero (e.g., a unit in the top-left corner). The bottom layer was named the retinotopic layer, and its activation patterns represented a filtered (Gaussian noise) face stimulus. The left layer was named the verbal layer, and each unit in this layer represented a verbal label for facial features in a localist manner. The right layer was named the visual image layer, and its activation pattern represented a non-filtered (without Gaussian noise) face stimulus. With this architecture and the representations in each layer, we trained the model for the three tasks described below.

## Tasks

Visual encoding of a face from a retinotopic input (visual recognition). In this task, retinotopic face pattern was hardclamped onto the retinotopic layer. Then, the network was trained to activate the non-filtered, unique visual face information of the same person in the visual image layer (individuation or visual recognition - see later).

Verbal encoding of facial features from a retinotopic input (verbalization). In this task, the input was the same as the previous visual recognition task, but the network was trained to activate the correct verbal units for each presented face. For example, if the face had slanted eyes, then the model had to turn on the unit for "slant eyes", and had to turn off the unit for "drooping eyes".

Mental imagery upon verbal cues. In this task, the verbal labels of facial features were presented onto the verbal layer, and the network was trained to activate the visual face information in the visual image layer. As we will explain later, the accuracy for this task never reached 100\% because different faces sometimes shared the same verbal labels (i.e., different targets from the same input pattern).

Recognition. A standard experimental task on human recognition memory employs a N -alternative forced choice task to probe recognition process, particularly when examining verbal overshadowing. The model, however, was not trained for making an explicit N -alternative forced
"choice". Therefore, we should adopt a proper measure to probe the model's recognition. It is one of the most debatable issues in cognitive psychology regarding what process/mechanism is underpinning recognition. Following previous studies (e.g., Plaut \& Behrmann, 2011), we examined whether the model could represent item-specific information (i.e., unique face) as an approximation of recognition process. If the model computes item-specific information of an "old" face in the visual layer from a "new" retinotopic (noisy) input, then it can be considered the model identifies this input as old face by mistake (especially after a verbal label for the old face was activated). In this way, we can at least measure false alarm safely, which is the target of the current study with this procedure.


Figure 2: Four examples of the training patterns. Note that two examples within each half share the same verbal labels, and thus the same pattern activations in the verbal layer. (However, they are different faces with different specific features as shown in the parentheses).

## Representations (face stimuli)

Figure 2 shows examples of the face pictures that we created using montage software (http://www1.mahoroba.ne.jp/~matumoto/nitaroS.html). Sixty-four face pictures were created by combining four types of eyes, four types of nose, and four types of mouth (see the bottom row of Fig. 2 for the possible features) in the following steps. First, we selected two verbal labels for each part of the face - slant eyes, drooping eyes, long nose, button-shaped nose, downturned mouth, and thick lips. Next, we selected two specific types for each verbal label (e.g., slant -eyes [thin] and slant eyes [big] for the label slant eyes, as shown in the right two faces in Fig. 2). In this way, we created four types of eyes, nose, and mouth, resulting in 64 different faces by combining 4 by 4 by 4 . In order to make the model trainable, we did not include other features such as hair. Finally, the size of each picture was $70 * 60$ pixels, and the color information in each pixel was binarized (i.e., black pixel $\rightarrow 1$; white pixel $\rightarrow 0$ ). The resultant 4200-bit vector pattern was used as the target pattern of the visual image layer in the visual recognition tasks (see second row of Fig. 2).

The original bit patterns were transformed into the retinotopic input pattern by smoothing with Gaussian convolution ( $S D=0.5$ ) ( Plaut \& Behrmann, 2011). The original bit patterns were smoothed by Gaussian convolution ( $S D=0.5$ ). In summary, the model had to map a noise-filtered retinotopic input (top row of Fig. 2) into a clearer visual representation (second row of Fig. 2), which is necessary for visual recognition.
The pattern activations in the verbal layer represented the verbal labels in a localist manner (third row of Fig. 2). For example, when presented with a retinotopic pattern of the drooping-eyes, long-nose, and thick-lip face, then the model had to activate the first, second, and third units of the verbal layer (the left two cases of Fig. 2 show these examples). In the mental imagery upon verbal cues task, the same units in the verbal layer were turned on, and the network was trained to activate the visual images in the internal image layer. The accuracy in this task can never be $100 \%$ because sometimes a different target should be generated from the same input pattern (i.e., the same verbal labels). For example, slant eyes (thin) and slant eyes (big) shared the same verbal label, slant eyes. Therefore, the same unit (slant eyes) was turned on for these two cases, but different output patterns (thin or big eyes) should be generated in the visual image layer. This is true to humans: We can imagine various kinds of faces but cannot specify a unique face by simply hearing "slant eyes, long nose, and thick lips". A small amount of Gaussian noise ( $\mathrm{SD}=0.2$ ) was added to the input for the hidden layer to encourage this layer to adopt more polarized outputs.

## Training

Among the 64 face patterns, only 55 patterns were presented during training, and the remaining nine untrained patterns were used to evaluate the network's generalization performance. Furthermore, this allowed us to investigate how differently the model behaved with the trained faces ('old' items) and untrained faces ('new' items), as it was crucial for us to investigate the effect of verbalization before the recognition phase (described later in more detail).

In each trial, units in the corresponding layer (retinotopic or verbal layer) were hard-clamped to their input values, and the network was allowed to cycle 10 times. In each time step, the activation spread to the next layer, gradually being scaled by the values of the interconnecting weights, and then the network would settle into the steady state (an attractor). After 10 cycles of updates, the discrepancy between the output activation pattern generated by the network and the correct target pattern was calculated, and the connection strength was adjusted to reduce the discrepancy. The model was trained with a learning rate of 0.05 , and with a decay parameter set to 0.0000001 . When we evaluated the final performance, we used a strict criterion such that the output was scored correct if the discrepancy was within 0.5 in every unit of the target layer (i.e., the activation was less/more than 0.5 if the target was zero/one for each unit respectively).

Given that young infants recognize their parents easily, it would be natural to assume that visual recognition skills are acquired earlier than an ability to verbalize facial features, or to imagine a face upon verbal cues.Thus, all 55 of the face stimuli were first trained for the visual recognition task. After learning to generate a steady state for more than $50 \%$ of the training items in this task, the other two languagerelated tasks were included in the training schedule.

## Results

## Training tasks

Five independent simulations were run with different random seeds, and we confirmed consistent results across five cases. The training finished after 2837 epochs of training (in each epoch an item appeared once for each task in a random order), at which point the network's performance reached $100 \%$ in both the visual recognition and the verbalization tasks from a retinotopic input for both trained and untrained items (i.e., generalization). The accuracy in the visualization task from verbal labels was $0 \%$ (see above for the reason).

## Visual recognition with/without verbalization

In order to investigate the visual recognition process of a retinotopic input after/without verbalization, we recorded the activation patterns in the visual image layer (right column of Fig. 3) when the network settled on 10 cycles after the retinotopic input presentation (left column of Fig. 3). The upper two rows of Fig. 3 show the pattern activations for a trained ('old') face (drooping eyes [thin], long nose [high], and thick lip [bottom big]) and for a visually-similar, yet untrained ('new') face (drooping eyes [big], long nose [high], and thick lip [bottom big]), respectively. Both retinotopic inputs were correctly mapped onto every unit of the visual image layer. This means that two visually-similar faces were successfully discriminated (see the bigger eyes represented in the second row), unless they were preceded by the verbalization process (100\% accuracy in computation of the individual-specific face information for all the nine untrained items). Next, the middle two rows of Fig. 3 show the activations for the same two items as the upper rows but after verbalization. Specifically, we simulated the following situation: Imagine that the network had encountered the 'old' face shown in top row of Fig. 3 (drooping eyes [thin], long nose [high], and thick lip [bottom big]), and the network had verbalized the correct labels (drooping eyes, long nose, and thick lip). To simulate this situation, the three verbal units for these labels were manually turned 'on' (generating the outputs of 1.0 ) and the network was allowed to cycle 10 times, during which the activations spread into the other layers (it updated its internal status 10 times). After 10 cycles, a retinotopic input for the trained face ('target') and that for the visuallysimilar, yet untrained face ('new') were presented respectively, and the network was allowed to update its status 10 times until each input pattern was mapped onto a


Figure 3: Activation patterns in the visual image layer (right) upon retinotopic inputs (left) for trained 'old' face and for untrained 'new' face in the visual recognition task. Upper two rows: without verbalization. Middle two rows: after verbalization of a similar 'old' face. Bottom row: after verbalization of a dissimilar 'old' face.
steady pattern in the visual image layer (right column). A visual inspection reveals the retinotopic input for the visually-similar 'new' face (drooping eyes [big]) was mapped onto the pattern for the 'old' face (slant eyes [thin]) in the visual image layer (false alarm). Euclid distance of the output pattern from the target "new" face was larger than that from the lure "old" face (i.e., similar to the "old" face pattern). This means that the model actually computed the item-specific information of the "old" face (false alarm). The same analysis was conducted for all the nine "new" items (against its visually-similar "old" face, respectively),

Activations immediately after verbalization of an 'old' face


Similar 'old' face
(drooping eyes, high nose, thick lip)


Dissimilar 'old' face
(slant eyes, button nose, downturned mouth)


Correct recognition
Figure 4: Internal activation patterns in the hidden layer (Hinton diagram: more white units denote higher activation values than black units) at the various kinds of time point (see main text).
and averaged across the five individual simulations. The resultant recognition accuracy was $60 \%$ ( $40 \%$ false alarm), $S D=14.9 \%$, which was significantly lower than $100 \%(t)$ $=5.99, p=.003$ ). This confirms that the example result in Figure 3 is generalized across other patterns. In contrast, the retinotopic input for the 'old' item was mapped onto the correct pattern (drooping eyes [thin]) in the visual image layer, though less weakly than when presented without verbalization (top row). Taken together, these results confirm that false alarm was more susceptible to verbalization than hit rate. Finally, the bottom row of Fig. 3 shows the simulated result in the condition where the distractor was dissimilar to the target. Specifically, the activation patterns were taken from the same untrained 'new' face (drooping eyes [big], long nose [high], and thick lip [bottom big]), but after activations of the irrelevant verbal units (slant eyes, button-shaped nose, and downturned mouth). Thus, the model had encountered a dissimilar person, and verbalized the dissimilar labels before visual recognition of a 'new' item. As a result, the network did not settle into the pattern of the dissimilar target face (slant eyes and downturned mouth), but the represented pattern was more similar to the correct pattern (drooping eyes and thick lip). In other words, the model did not confuse the presented retinotopic input (the 'new' face) with the previously encountered 'old', yet dissimilar face, thus avoiding false alarm in this low-similarity condition (Kitagami et al., 2002).

Finally, in order to understand the mechanism of verbal overshadowing, the pattern activation in the hidden layer was measured at the various kinds of time point (Figure 4). First, the Hinton diagrams at the top row of Figure 4 show
the internal activation patterns when the network successfully discriminated two visually similar 'old' and 'new' faces, respectively (shown in the upper panels of Fig. 3). A visual inspection reveals these two representations are very similar. This concurs with the idea that fine-grained representations are crucial in face recognition (Maurer et al., 2002), without which one would be easily mapped to the other, incorrect, face pattern in the visual image layer (i.e., incorrect recognition).

Next, the left diagram of the middle row of Figure 4 shows the activation pattern immediately after verbalization of 'drooping eyes, high nose, and thick lips'. As a result, this internal representation immediately after verbalization was neither identical with that for visual recognition of the 'old' face (top left) nor that for visual recognition of the 'new' face (top right), concurring with the idea that verbalization generates the representation that lacks finegrained information crucial for face recognition (Maurer et al., 2002). Though lacking such detailed information, it was nonetheless closer to the representation for the 'old' face (top left) than that for the 'new' face (top right). In other words, the model's internal status had already moved towards the pattern for the 'old' face. We will explain later why this representation increased the false alarm of the model when the distractor was similar to the target.

## Discussion

The present computer simulation examined how internal representations changed upon verbalization and how this affected subsequent visual recognition. Without verbalization, the model represented the correct and unique pattern activation for each old face and for a visually-similar new face, respectively, in the visual image layer. This confirms that the model did not confuse two visually-similar retinotopic inputs. On the other hand, the model failed to represent the correct pattern for a new face following the forced activation of verbal units for an 'old', visuallysimilar face (i.e., verbalization). Instead, the represented pattern in the visual layer assimilated to that for the visually-similar 'old' face, suggesting that the model could not differentially recognize the 'new' face from the 'old' face (a false alarm). Importantly, this assimilation was weakened when the preceding verbalization included the features of an 'old', yet dissimilar face. Therefore, these results mirrored Kitagami et al. (2002), who found that participants' false alarm increased upon verbalization when the distractors were similar to the target.

Explicit implementation of a computer model allowed us to directly look at the internal representations to understand why the model behaved in this way. In a normal situation (without verbalization), the model computed very similar, yet unique, internal representations for retinotopic inputs of visually-similar faces. This fact is consistent with the idea that fine-grained representation is necessary for visual recognition of faces (Maurer et al., 2002), especially when discriminating a target from similar distractors. When the model verbalized the facial features, this internal
representation changed such that it was neither identical to that of an 'old' face, nor that of a ''new' face, supporting the argument that verbalization either degrades the finegrained representation (Meissner et al., 2001), or renders it unavailable (Schooler, 2002). Nonetheless, it was closer to the representation for the verbalized 'old' face than to that of a 'new' face. In order to understand why this representation induced a false alarm for a visually-similar face, it is useful to describe the general activity of PDP models here. During training, a PDP network finds a unique attractor state (a unique abstract pattern in the hidden layer) associated with each input pattern. Therefore, generating a correct output is sometimes described as if the internal activity of the hidden layer falls into its unique attractor basin. Though they are unique, similar inputs are associated with similar/neighboring attractor basins (as shown in the top two panels of Fig. 4). Consequently, if the internal representation of the model is degraded for some reason, a similar input can incorrectly drift and fall into the wrong attractor basin, generating an incorrect output. In the current model, verbalization generated the internal representation that lacked fine-grained information crucial for visual recognition (middle-left row of Fig. 4). Though it lacked such information, it was nonetheless closer to the representation of the 'old' face (the top and middle left rows are similar in Fig. 4). In other words, the model's internal status had moved towards the attractor basin for the 'old' face by verbalization. In such a situation, a similar retinotopic input, which would have settled into a unique, yet similar/neighboring attractor basin without verbalization (top-right of Fig. 4), was easily captured by the attractor for the 'old' face. The resultant (captured) internal activation pattern is shown in the left bottom row of Fig. 4, which was more similar to the top-left pattern than the top-right pattern (i.e., incorrect recognition). This is the mechanism by which verbalization impairs subsequent recognition, especially when the distractor is similar to the 'old' face. In contrast, when the dissimilar (inconsistent) labels were verbalized before visual recognition, the hidden layer activation pattern was very different to that for a subsequent 'new' face (i.e., the top and middle right rows are not similar). In this case, the network is not captured by this dissimilar attractor basin, as is shown in the bottom right diagram of Fig. 4.
In summary, as the present study has demonstrated, a computer simulation is a useful tool for investigation of verbal overshadowing. It is difficult to examine verbal overshadowing empirically, given that the standard paradigm involves a single-trial measurement. Therefore, many participants are necessary for detecting a reliable effect, and it is difficult to systematically manipulate a variable as a within-subject factor. In such a situation, it is worthy to implement a computational model in order to understand the mechanism and to provide a theory-driven question that can be empirically testable in human experiments. Of course, any computational modeling should be concerned whether the model's representation/process is the same as the human's, but previous studies have
demonstrated that investigating the internal representation of a model is a useful approach to advance the cognitive theory (e.g., Plaut \& Behrmann 2011). Furthermore, the current model can be extended to other types of perceptual stimuli (not just face). Thus, we expect that the present study would be an important step to clarify the relationship between language and perception in general. Finally, one issue deserves consideration: The current model simulated the increase in false alarm upon verbalization (Kitagami et al., 2002; Meissner et al., 2001), rather than a missed response. Although some studies failed to detect a significant difference between these two measures (Schooler \& Engstler-Schooler, 1990, a recent experimental and computational study (Clare \& Lewandowsky, 2004) suggested that it was actually the increase in "not present" responses, a response type that was not implemented in the current model. A future modeling target would be to understand the mechanism by which verbalization increases both "not present" responses and false alarms, of which the latter particularly occurs when the distractors and targets are similar.

## Acknowledgments

This study was supported by Grant-in-Aid for Scientific Research on Innovative Areas \#23101006 (to S. Kitagami) and Grants-in-Aid for Scientific Research (B) 21330168 (to J. Kawaguchi).

## References

Clare, J., \& Lewandowsky, S. (2004). Verbalizing facial memory: Criterion effects in verbal overshadowing. Journal of Experimental Psychology-Learning Memory and Cognition, 30(4), 739-755
Fallshore, M., \& Schooler, J. W. (1995). Verbal vulnerability of perceptual expertise. Journal of Experimental Psychology: Learning Memory and Cognition, 21, 1608-1623.
Kitagami, S., Sato, W., \& Yoshikawa, S. (2002). The influence of test-set similarity in verbal overshadowing. Applied Cognitive Psychology, 16, 963-972.
Maurer, D., LeGrand, R., \& Mondloch, C. J. (2002). The many faces of configural processing. Trends in Cognitive Science. 6, 255-260
Meissner, C. A., Brigham, J. C., \& Kelley, C. M. (2001). The influence of retrieval processes in verbal overshadowing. Memory \& Cognition, 29, 176-186.
Plaut, D. C. \& Behrmann, M. (2011). Complementary neural representations for faces and words: A computational exploration. Cognitive Neuropsychology, 28, 251-275.
Schooler, J.W. (2002), Verbalization produces a transfer inappropriate processing shift. Applied Cognitive Psychology, 22, 989-997.
Schooler, J. W., \& Engstler-Schooler, T. (1990). Verbal overshadowing of visual memories: some things are better left unsaid. Cognitive Psychology, 22, 36-71.

# The effect of " Trust dynamics " : Perspective taking during collaborative problem solving 

Yugo Hayashi (hayashi.yugo.gp@u.tsukuba.ac.jp)<br>Faculty of Library, Information and Media Science, University of Tsukuba<br>1-2, Kasuga, Tsukuba, Ibaraki, 305-8550, Japan


#### Abstract

We investigated the influence of reflections on self/others' trust within group-based problem solving. We assessed the role of trust dynamics on perspective-taking activities within conflictive groups, extending the experimental framework used by Hayashi (2012) and including conversational agents for controlling participants' interactions related to trust dynamics and perspective taking behavior. Results showed that (1) reflections of self/other trust in conflictive groups may influence trust towards other members, and (2) reflections of trust by members with conflicting perspectives may facilitate trust and perspective taking. This suggests that the level of trust dynamics facilitates trust and can function to manifest perspective taking within cooperative groups.


Keywords: Trust dynamics; perspective taking; collaboration; group dynamics.

## Introduction

Conflict is inevitable in human interactions, even within ingroup activities. We investigated the mediating factor of group dynamics on "perspective-taking activity" within conflictual situations, focusing on how social cognitive factors, such as reflections of " trust dynamics," could facilitate the interaction process. Multiple conversational agents were used to control for group members' interactions, such as perspective-taking and social dynamics.

## Perspective taking and conflicts in groups

Perspective taking is an essential activity for social interaction. It perspective taking requires higher-level cognition and is rife with unintended errors and difficulties (Keysar, Barr, Balin, \& Brauner, 2000; Hayashi \& Miwa, 2009). Perspective taking plays an important role in establishing common ground between adult speakers. Keysar et al. (2000) points out that this activity includes complicated cognitive operations, and people tend to have egocentric biases that hinder such activities. Hayashi, Miwa, and Morita (2006) showed that an individual's egocentric biases constrains perspectivetaking activities during a cooperative task.

Within cooperative group activities (e.g., scientific research groups), perspective taking also plays an important role in discovering new solutions and knowledge. For example, Dunbar (1995) and colleagues investigated the use of inductive reasoning within a scientific research group and proposed the concept of distributed reasoning, where group members achieve their goals by taking charge of different types of inferences. Other studies have also investigated the nature of learners collaboratively working on complex problems while focusing on explaining such activities, role sharing, and reflective interactions (Okada \& Simon, 1997). In
these studies, perspective taking promotes the facilitation of conceptual change for gaining a deeper understanding. However, in such activities, we have to consider the effects of conflicts that occur during interactions. Conflicts occur due to disagreements among group members concerning ideas and opinions about the task being performed (Jehn, Greer, Levine, \& Szulanski, 2008). Overcoming such conflicts not only is important for establishing common ground but also can provide an opportunity for new discoveries. On the basis of these studies, we investigated the effects of a type of social dynamic that is expected to facilitate an "exocentric" perspective for overcoming group conflict.

## Influence of trust dynamics in groups

Organizational and social network studies have shown that the perception of "trust in other member(s)" within the group is important for facilitating group performance (Kramer, 1999; Castellano, Fortunato, \& Loreto, 2009). Within-group perspective taking is an interactive social process that could be mediated by the level of trust and respect within that group. Unfortunately, few studies have investigated how the perception of such dynamics affects cognitive processes during cooperative group activities.

Here, we briefly review what we mean by "trust" and reframe the term for our study. Mayer and Davis (1995) conceptually defines trust as " a willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the truster." Yamagishi and Ymagishi (1994) propose that trust is based on profit and loss to the individual. To reach a solution to the problem, cognitive operations move from the initial unsolved state to the goal state. In group-based problem-solving activities, partners play important roles in providing information to members, and members use that information on the basis of how valuable it is in attaining the goal. The level of belief, such as the trust towards group members, is very important during collaborative group activities. Thus, we define trust as "a belief that the perspective of that person will lead toward a good problem solving outcome."

Trust has several benefits for organizations and their members. When the level of trust is increased, a group is expected to experience superior group processes (e.g., higher levels of cooperation) and higher performance, and vice versa (Dirks, 1999). Parks, Henagar, and Scamahorn (1996) examined how low- and high-trust individuals respond to messages of intent from other participants in a social dilemma task. They
found that low-trusters reacted to a competitive message by decreasing cooperation but were unaffected by a cooperative message. In contrast, high-trusters reacted to the cooperative message by increasing cooperation but were unaffected by the competitive message. In another study, the authors investigated the effects of trust on decision makers' responses to fear during social dilemmas (Dirks \& Ferrin, 2001). They found that high-trusters cooperated more than low-trusters when fear was present but cooperated at the same rate when fear was absent. Klimoski and Karol (1976) investigated the dynamics of interpersonal trust during group creative problem solving with members placed into high-trust, low-trust, or control (no trust) conditions. The high-trust and control groups outperformed the low-trust group on each of 3 tasks based on the number of ideas each group generated.

In sum, shared trust among members enhances individuals’ willingness to engage in various forms of spontaneous sociability, thereby facilitating task performance. However, most past studies have only focused on the direct effect of trust on task performance. We focused on the cognitive process of how reflection of the trust dynamics influences perspectivetaking activities. Figure 1 shows our research framework. Members on the left (with high trust) end up with an exocentric perspective, while members on the right (with low trust) end up with an egocentric perspective.


Figure 1: Research frame work.

## Goal and hypotheses of the current study

The present study investigated the impact of social dynamics on perspective taking activities within conflictive groups. By extending an experimental methodology from Hayashi (2012), we investigated the following two hypotheses.

H1 Reflections of self/other trust in a conflictive group may influence trust towards other members.

H2 Reflections of trust by members with conflicting perspectives may facilitate trust during the perspective-taking process.

## Method

## The rule discovery task: Bistable objects

For the main task, the participants' goals were to count a series of objects presented on a computer display and discover its sequential rule by cooperating with other"participants." Materials for the objects were several sets of random patterns containing several figures on a $6 \times 6$ grid (colored black or white; see Figure 2). In each set, a pattern consisting of combined square blocks was shown against the background of either black or white colors. The background color was controlled to derive, through a Gestalt effect, the change in the problem-solver's perspective (Koffka, 1935).


Figure 2: Experimental stimuli used in the study.
Each set consisted of several "objects" (or patterns) in black or white, each of which consisted of a single block or multiple blocks. As shown in Figure 2, one of the paired objects has a total of 10 " components" (4 black and 6 white). When a participant focuses on white components inside a black background, the white components become the figure, and the six components pop out.

In the present study, a group of four participants collaboratively worked on the problem-solving task through computer terminals connected via a local network. All members exchanged text-based messages to discuss their opinions as to the number of objects and the sequence of the target rule. As shown in Figure 2, four participants saw the objects with either a black background (white perspective) or a white background (black perspective). Two members saw a black and two saw a white background, and each member counted the number objects that were "popping out." For each trial/objects, a square outer-box was shown on the display for one second, which was followed by a stimulus picture presented inside the box frame. Participants were required to send one message per trial to the other members as part of a secondary task (explained in the following section). Participants were told that they had to count the number of objects inside this box frame. The number of white components and black components was controlled, and the total number of components presented to the participants varied between six and twelve. The sequential pattern (Target rule) of the sums of black components and white components was repeatedly presented during each trial (i.e. $6,8,10,12$ or $6,8,10,12$; see

Figure 3). In the initial stage of the task, the local black/white numbers were controlled to be the same. Experiments from a past study showed that participants are fixated to the figure perspective, and participants try to search for the target rule based on his/her perspective (Hayashi et al., 2006; Hayashi \& Miwa, 2009; Hayashi, 2012) . On trial 9, we controlled the trials for differentiation in responses so that participants would report different numbers. Through this manipulation, participants experience conflict.


Figure 3: Example sequence of the objects.
Previous studies show that even though participants confront this conflictual stage, they are biased to their own object/perspective and behave egocentrically (Hayashi et al., 2006). To discover the rule for this task, the participant has to look at the different colored objects and integrate the other two members' perspectives. In the present study, one human participant and three conversational agents acted as human partners. Participants were informed that they would conduct the task with real humans and were told not to chat about anything else related to the task. Two agents were programmed to export messages based on a different perspective (object).

## Group members' behaviors: Conversational agents

In the present study, we set up a text-based chat communication platform by using one server and three clients, including one chat engine and three conversational agents(Figure 4). The system was similar to one used by Hayashi (2012) and developed by Java. On the Server side, a broadcast mechanism was used to distribute all the messages simultaneously. When messages were sent to the Server, they were re-distributed to all Clients and/or Agents.

The conversational computer agents were designed by a typical rule-based system. Based on pre-defined rules, the system can respond to sentences that were inputted by the participants (Figure 4). All three agents were implemented by the rule shown below. All agents autonomously responded to each other's text messages as independent interlocutors.

The conversation agent extracts keywords from the sentences that were distributed by the participant and other agents. The most frequent keywords that are used during this task are related to the (1) number of objects, (2) colors of objects, and (3) rules about the sequence. The agent contains a temporary working memory storage to represent the current status of the input messages from the (1) keywords of the participant, (2) keywords from the agents, and (3) objects


Figure 4: Communication platform.
that were presented on the screen. A rule base, in an "ifthen " format, defined all responses from the agents. When the agent detects keywords of (a) numbers, (b) colors, and (c) the hypothesis, working memory is updated. Then, a patternmatching strategy is used for binding the rules. In the present study, the agents were programmed to respond based on the numbers or colors of the objects that were set for its perspective (Agent A responded based on the black objects, Agent $B$ and $C$ responded based on white objects). The following shows the summarized version of the basic rules utilized in the study:

```
    \(<\) Rules \(>\)
Trials \(<8\)
(agent \(A, B, C\) )
\(->\) numbers \((\) white objects)
Trail \(>8\)
(agent A)
IF : "color(any)" - > colors(white)
ELSE IF : "rules" - > rules
ELSE - > numbers(white)
(agent B,C)
IF : "color(white)"AND "agentBorC" NOTcolor"
\(->\) numbers(black)AND color(black)
ELSE IF : "rules" - > rules
ELSE - > numbers(black)
```

For example, outputs of numbers (black) and colors (black) could be, "Oh, I think there are 6 black objects," "So you see 3 ? I see 3 of them in black, " etc. The server will distribute the messages to each agent in a randomized order with a delay for appropriate turn takings.

## Trust dynamics diagram interface

On each trial, participants were required to work on a secondary task, which was to evaluate and receive results of self/others' evaluations of "trust." This was done to provide reflections regarding trust dynamics of the group members.

Participants were asked to "evaluate each member based on the belief that his/her information will lead to a good outcome for the group." Figure 5 shows an example of the trust dynamics network diagram interface. The participants' representation (node) was located in the center, and the other members (partners A, B, and C) were located above and below. There were four nodes and 12 links for evaluating each member. Each arrow addresses the evaluation towards a particular member. A" + " indicates that the participant evaluated him/her as trustworthy, and a " - " mark means that the participant evaluated the member as untrustworthy. Within the interface, there were three arrows pointing from the participant to Partners A, B, and C. Participants were required to select trust $(+)$ or distrust (-) by using a pull down selection box. Prior to the experiment, all participants completed a training session in order to become familiar with the trust dynamics network diagram interface. During the experiment, participants were required to evaluate other members ( $\mathrm{A}, \mathrm{B}$, and C) during each trial. All members' (agents') outputs were controlled to simulate a pattern of trust dynamics due to each experimental condition.


Figure 5: Trust dynamics network diagram interface.

## Procedure

Seventy-five undergraduate students from a humanities course participated in the experiment (trust condition, $n=$ 27; distrust condition, $n=26$; no-trust condition, $n=22$ ). The experiment took place in a room with 12 laptop computers. Participants performed 32 trials (average time $=35$ minutes). After the task, participants worked on a writing test. In the trust condition, when a conflict occurred on the 9th trial (see Figure 3), the partners/agents were controlled to gradually generate trust towards the members with conflicting perspectives. Conversely, in the distrust condition, there were no such adjustments, and partners/agents kept generating distrust towards members with different perspectives.

Figure 6 shows the simulation of the trust dynamics during the experiment. For trials 1 to 6 , all agents' evaluations were
programmed to select distrust $100 \%$ of the time. On trials 7 to 11 , all agents were programed to generate trust with $80 \%$ variability. After the 12th trial, members with the same perspective ( $\mathrm{C}->\mathrm{B}, \mathrm{B}->\mathrm{C}, \mathrm{A}->$ participant) were programmed to generate trust $100 \%$ of the time. Other's evaluations were weighted to gradually select distrust until the 17th trial (see Figure 6). This was done to manipulate a situation where members evaluated others negatively after the conflict occurred during the main task. In the trust condition, responses were weighted to recover gradually from the 18th trial. However, in the distrust condition, there were no adjustments made. The no-trust condition was the baseline condition, and there was not a trust dynamics diagram provided to the participant.


Figure 6: Results of the agent simulation on trust dynamics.

## Dependent variables

We had two main dependent variables of interest. First, we analyzed the degree of trust towards other members during the following four task stages: (1) Initial Stage (trials 1 to 8), (2) Conflict Stage 1 (trials 9 to 16), (3) Conflict Stage 2 (trials 17 to 24), and (4) Conflict Stage 3 (trials 25 to 32). Evaluations of Participants (agents) A, B, and C were collected based on these stages and coded as follows: (1) trust as 1 point and (2) distrust as -1 point. At each stage, the average score was calculated for each participant and used as a representative value. We then calculated the average of each partner within each trial stage. Next, we analyzed participants' dialog data during the task. This was done to evaluate participants' change in perspective. If the conversation data included a description that referred to the background color, it was counted as "a change in perspective." If a participant's data did not include such information, it was counted as "no perspective change."

## Results

## Evaluations of trust towards other members

Figure 7 shows the trust dynamics results in the trust and distrust conditions. The vertical axis represents the average ratio of trust, and the horizontal axis represents the experimental trials.


Figure 7: Results of the evaluations of the trust to partners.

First, we analyzed the trust condition. We performed a 4 $\times 3$ within-subjects factorial ANOVA with the four evaluation trial-times (Initial Stage vs. Conflict Stage 1 vs. Conflict Stage 2 vs. Conflict Stage 3) and the three partners (Partner (agent) A vs. Partner (agent) B vs. Partner (agent) C) as independent variables. There was a significant interaction between the two factors $(F(6,156)=3.0, p<.01)$. We conducted further analyses to deconstruct this interaction.

First, a simple main effect analysis was conducted on each trial time. In the Initial Stage (trials 1 to 8), Conflict Stage 2 (trials 17 to 24), and Conflict Stage 3 (trials 25 to 32) there were no differences between conditions $(F(2,208)=0.063$, $p=.94),(F(2,208)=2.014, p=.14,(F(2,208)=1.142, p$ $=.32)$ ). However, in Conflict Stage 1 (trials 9 to 16), there were significant differences between conditions $(F(2,202)=$ 8.316, $p<.01$ ). Multiple comparisons using Ryan's method were conducted for Conflict Stage 1. Results indicated that the average score of partner A was higher than partners B and C respectively ( $p<.01 ; p<.01$ ). There were no differences between partners B and $\mathrm{C}(p=.75)$. Next, we examined the distrust condition. We performed the same factorial ANOVA as in the trust condition. There was a significant interaction between the two factors $(F(6,150)=10.110, p<.01)$. Again, we conducted further analyses to deconstruct this interaction.

A simple main effect analysis was conducted on each trial time. In the Initial Stage (trials 1 to 8 ), there were no differences between conditions ( $F(2,200)=0.018, p=.98$ ). However, there were significant differences in Conflict Stage 1, Conflict Stage 2, and Conflict Stage $3(F(2,200)=20.355$, $p<.01, F(2,200)=14.559, p<.01, F(2,200)=21.615$, $p<.01$ ). We conducted multiple comparisons, again, using Ryan's method. For Conflict Stage 1, results indicated that the average score of partner A was higher than that of partners B and C, respectively ( $p<.01 ; p<.01$ ). There were no
differences between partners B and C $(p=.74)$. For Conflict Stage 2, results indicated that the average score of partner A was higher than that of partners B and C , respectively ( $p<$ $.01 ; p<.01$ ). There were no differences between partners B and $\mathrm{C}(p=.26)$. Finally, for Conflict Stage 3, results indicated that the average score of partner A was higher than that of partners B and C, respectively ( $p<.01 ; p<.01$ ). There were no differences between partners B and $\mathrm{C}(p=.53)$.

## Perspective Taking Process

Figure 8 shows the results of the dialogue process of changing perspectives. The vertical axis represents the ratio of perspective change, and the horizontal axis represents the experimental condition. The numerals shown on the cylindrical bars indicate the number of participants in each condition. A chi-square analysis was conducted to check the difference in the number of problem solvers who integrated perspectives during the task. Results indicated that there were significant differences among the conditions ( $\chi^{2}(2)=6.132, p<.05$ ). Next, a multiple comparison was conducted on the two conditions using Fisher's exact test. There were significant differences between the trust condition and the distrust condition and between the trust condition and the no trust condition, respectively ( $p<.01, p<.01$ ). Conversely, there were no differences between the distrust and the no trust condition ( $p$ $<.10$ ). This suggests that a group that interacted with trustbased dynamics were taking into consideration the perspectives of others more frequently than a group that did not.


Figure 8: Results of the perspective taking performance.

## Discussion and Conclusions

We investigated the influence of social cognitive factors, such as trust dynamics during perspective taking activities, in group based problem solving. We investigated how reflections of self/others' trust can facilitate the level of trust toward others and, thus, motivate perspective. Results from the trust dynamics paradigm show that participants felt distrust toward others during the initial stage of the interactions. However, participants gradually constructed trust towards others over time. As the interaction proceeds, trust became stronger. However, when participants perceived members with conflictive perspectives, trust dropped compared to those without a conflicting perspective. This indicates that trust was de-
pendent on whether conflicting information provided by the members was present. Regarding reflective feedback from other members, we found different types of evaluations of trust towards other members (participant $->\mathrm{A}, \mathrm{B}, \mathrm{C}$ ). When participants received a recovery of trust from members with conflictive perspectives (partner B \& C), trust recovered to the same level of trust that was present with the member (partner A) who did not have a conflictive perspective. This shows that reciprocal behavior might have occurred through the interaction process. It is interesting that even if participants experienced conflict from other members, if trust is later presented, he/she will generate feelings of trust. Conversely, if participants do not receive such reflective trust (distrust condition), trust never recovers. These results support H1 and indicate that reflective feedback was influential in building trust towards other members. These results may differ from other types of reflective feedback, such as when you know that others can see everyone's evaluations. This type of feedback might have hindered our results. Furthermore, we found that certain reflections of trust feedback could influence the perspective-taking process. Participants in the trust conditions communicated with others based on conflicting perspectives more so than those in the distrust condition. This supports H2. These results suggest that the level of trust dynamics facilitates trust and can function to manifest perspectivetaking activities within cooperative groups.

The present study provides new implications for cognitive studies on perspective taking during group communication. Specifically, our results capture a trust dynamics factor that could gradually influence the process of trust and exocentric perspectives towards conflictual members. As described earlier, a few studies have investigated the nature of trust in groups and trust's influence on task performance. Some studies investigated the nature of perspective taking in risky dilemma situations. One study Evans and Krueger (2011) observed that in high-risk situations, people heuristically chose distrust; trust increases when costs decrease and benefits increase. However the present study focused on a cooperative situation; our results help elucidate how trust is mediated by problem solving interactions. Some studies (e.g., Klimoski and Karol (1976) ) have investigated the nature of trust on creative problem solving. Unfortunately, there are not many investigations assessing what kinds of trust dynamics are effective in the cognitive process of such activities.

Our present results can contribute to future studies modeling cognitive systems for supporting collaborative activities in groups. Reflections of trust dynamics, and the positive level of trust, are effective in facilitating higher trust towards others, as well as enhanced perspective taking. This model potentially can be used in collaborative learning activities, such as computer-based learning activities with large groups where perspective taking plays an important role. These trust dynamic mechanisms during collaborative learning can motivate learners to work on tasks/materials and be cognizant of the useful contributions of other members.

## References

Castellano, C., Fortunato, S., \& Loreto, V. (2009). Statistical physics of social dynamics. Reviews of mordern physics, 81(2), 591-646.
Dirks, K. (1999). The effects of interpersonal trust on work group performance. Journal of applied psychology, 84, 445-474.
Dirks, K., \& Ferrin, D. (2001). The role of trust in organizational settings. Organization Science, 12(4), 450-467.
Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. In J. R. Sternberg \& E. J. Davidson (Eds.), The nature of insight (p. 365-395). Cambridge, MA, USA: MIT Press.
Evans, A., \& Krueger, J. (2011). Elements of trust: Risk and perspective-taking. Journal of ecperimental Social Psychology.
Hayashi, Y. (2012). The effect of "maverick": a study of group dynamics on breakthrough in collaborative problem solving. In Proceedings of the 34th annual conference of the cognitive science society (p. 444-449).
Hayashi, Y., \& Miwa, K. (2009). Prior experience and communication media in establishing common ground during collaboration. In Proceedings of the 31th annual conference of the cognitive science society (p. 526-531).
Hayashi, Y., Miwa, K., \& Morita, J. (2006). A laboratory study on distributed problem solving by taking different perspectives. In Proceedings of the 28th annual conference of the cognitive science society ( $\mathrm{p} .333-338$ ).
Jehn, K., Greer, L., Levine, S., \& Szulanski, G. (2008). The effects of conflict types, dimensions, and emergent states on group outcomes. Group decisions and negotiations, 17, 465-495.
Keysar, B., Barr, J. D., Balin, A. J., \& Brauner, S. J. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. Psychological Science, 11(1), 32-38.
Klimoski, J. R., \& Karol, L. B. (1976). The impact of trust on creative problem solving groups. journal of applied psychology, 61(5), 630-633.
Koffka, K. (1935). Principles of gestalt psychology. London, UK: Routledge and Kegan Paul.
Kramer, R. (1999). Trust and distrust in organizations. Annual review of psychology, 50, 569-598.
Mayer, R., C, \& Davis, J., H. (1995). An integrative model of organizational trust. Academy of management review, 20(3), 709-734.
Okada, T., \& Simon, H. (1997). Collaborative discovery in a scientific domain. Cognitive Science, 21(2), 109-146.
Parks, D., Henagar, F., \& Scamahorn, D. (1996). Trust and reactions to messages of intent in social dilemmas. Journal of conflict resolution, 40(1), 134-151.
Yamagishi, T., \& Ymagishi, M. (1994). Trust and commitment in the united states and japan. Motivation and Emotion, 18(2), 129-166.

# How listeners respond to speaker's troubles 

Patrick G. T. Healey, Mary Lavelle, Christine Howes, Stuart Battersby, Rose McCabe<br>ph@eecs.qmul.ac.uk<br>Queen Mary University of London,<br>Cognitive Science Research Group,<br>School of Electronic Engineering and Computer Science, London E1 4NS, UK


#### Abstract

Listeners normally provide speakers with simultaneous feedback such as nods, "yeah"s and "mhm"s. These 'backchannels' are important in helping speakers to talk effectively. Two factors are known to influence when a backchannel is produced; if the speaker is looking at the listener or if the speaker is presenting new information. We investigate a third factor: whether the speaker is having trouble speaking i.e. self-repair. If dialogue is an active collaborative process then listener's responses should be especially critical when trouble is encountered. Using data from a corpus of three person dialogues we show that speaker's rate of self-repair is a better predictor of listener responses than speech rate. We also show that listeners respond strongly to speaker troubles independently of whether the speaker is looking at them. We argue that it is the points at which conversation threatens to go off-course that are most significant for coordination. Keywords: Gesture; repair; dialogue


## Introduction

Listening in conversation is not a passive activity. As Goffman (1955) noted, what listeners do while being addressed has important consequences for the way that speakers produce their turns. Goffman distinguished between two general kinds of listener feedback; displays of attention and understanding of what is said and the signalling of interactional functions such as a desire to speak next. Yngve (1970) introduced the term 'backchannel' to describe these uses of simultanous feedback that provide speakers with concurrent information about how their turn is being received.

In a series of experiements examining the effects of listener response behaviours Bavelas and colleagues were able to show that the fluency and effectiveness of a speaker's turns depends directly on the level of feedback they are getting from their addressees (J. B. Bavelas et al., 2000; J. Bavelas et al., 2006). People telling stories to listeners who are engaged in a distractor task speak less fluently and are less compelling than those whose listeners are attending more carefully.

Given the importance of listener responses for successful interaction a key question is what prompts a listener to produce them? Many of the most common backchannel signals, such as nods and smiles, use the visual channel which avoids potential competition with concurrent speech. One common finding in the literature is that addressee responses are reliably correlated with speaker's
gaze. Goodwin (1979) observed that speakers will periodically check whether addressees are attending by looking at them and if they get no response may restart or switch to a new addressee mid-turn. J. B. Bavelas et al. (2002) found that listener responses in their 'close call' story telling task were significantly more likely to occur in a 'gaze window' i.e. when a speaker is looking at a listener than when they are not.

A second common observation in the literature is that backchannels are also associated with the introduction of new information into a dialogue such as the introduction of a new referent or proposal that may warrant some signal of interim acknowledgement or acceptance before the speaker's turn is completed (J. Bavelas et al., 2006; Clark \& Wilkes-Gibbs, 1986; Yngve, 1970). In this case it is the information update that prompts the use of a backchannel to signal understanding 'so far' (Goodwin, 1981).

In this paper we explore the effects of a third factor on listener responses: the degree of difficulty a speaker has in producing their turn. Few conversational turns are produced without some form of online revision or reformulation during their production. Sometimes referred to as disfluencies these self-repairs are indicative of some sort of trouble producing a turn. If conversation is a collaborative process in which each turn is co-produced (Goodwin, 1979; Clark, 1996) then this leads to the hypothesis that the points at which the speaker shows signs of getting into trouble ought to be especially critical for collaborative reponses. This paper tests this hypothesis by investigating the relationship between nodding, speech rate and repair rate in a corpus of three person dialogues.

## Methods

Experimental work on listener backchannel responses has focussed only on dyadic, i.e. two person, interactions. However, natural interactions frequently involve more than two people (Goffman, 1981; Eshghi, 2009). For current purposes three-way interactions also have the practical advantage that they make it possible to compare two kinds of listener depending on who the speaker is looking at while they talk. Given the importance of speaker gaze to the production of backchannels this provides a useful opportunity to compare the responsiveness
of two fully ratified, active participants who differ only in whether they are being looked at while the speaker produces their turn. Note that this differs from the work of Schober \& Clark (1989) who investigated the behaviour of side participants and overhearers whose ability to provide concurrent feedback was restricted.

## Participants

Fifty four participants (30 Male, 24 Female) were recruited to the study through advertising on local community websites. Of those who responded to the advertisement, $40 \%$ participated. Participants within each group had not met prior to the study.

## Procedure

Participants were brought into the laboratory in threes and seated in a triangular formation so that each participant had good visual access to each of the others (see Figure 1). The researcher read aloud a fictional moral dilemma scenario called 'the balloon task' to the seated group. The scenario states that there are four people in a hot air balloon, which is losing height and about to crash into some mountains killing all on board. One person must jump from the ballon to their certain death in order to save the other three. Participants were instructed to debate the reasons for and against each person being saved, and reach mutual agreement about who should jump. The group was provided with an opportunity to ask questions before the researcher left the interaction space and the task began. Interactions ended when participants reached a joint decision. Groups that failed to reach an agreed decision had their interaction terminated at approximately 450 seconds ( 7 minutes 30 seconds).


Figure 1: 2-dimensional image of participants engaged in triadic interaction, wearing the reflective markers

All interactions were recorded in a human interaction laboratory fitted with an optical based Vicon motioncapture system, consisting of 12 infrared cameras and Vicon iQ software. Participants wore a top and a cap with 27 reflective markers attached. Cameras detected the markers at 60 frames per second, resulting in a highly accurate 3D representation of participants' movements over time (see Figures 1 and 2).


Figure 2: The wire frame representation of the interaction in 3-dimensional space

## Data Analysis

For each interaction, speech was transcribed from the 2D video in the annotation tool ELAN (Crasborn \& Sloetjes, 2008). These transcripts, together with the motion capture data were used to produce three measures, speech rate, rate of self repair and rate of nodding for each participant.

Measures of Self-Repair Automatic processing of the transcripts identified, for each turn, the number of words, the number of filled pauses (e.g. er, um) and the number of unfilled pauses, defined as pauses between segments of speech by the same speaker of greater than 200 milliseconds (following e.g. Zellner, 1994, a.o.). Since self-repairs often involve the repetition of words, usually close together, a normalised within-turn repeated words value was calculated, by identifying repeated words in a turn and the distance between them and applying a decay function. Examples of turns including selfrepetition, and their word repeat value are shown below, from a low repeat score in example 1 to a high repeat score in example 3. Repeated words are shown in bold, and their repetition in italics.

[^294](2) Trust me his wife if he's if he's a pilot his wife knows how to do it [1.25]
(3) And and they they said that she said that they emptied the balloon to make it lighter [2.98]

To check validity, this measure was also calculated on a corpus of 52 clinical dialogues which had been handannotated for self-repair (McCabe et al., in preparation). For the 15,191 turns analysed, the within-turn repeated words measure was positively correlated with the handannotated self-repair measure ( $r=0.57, p<0.001$ ) and is therefore used as an index of self-repair. All values were normalised by number of frames in the turn, and mapped to the frame-by-frame motion capture data.

Nodding Head movement was derived from the vertical movement of participants front left head marker. Head nodding was approximated in a two-step process. Firstly, low frequency movements ( 1 Hz and below) and high frequency movements ( 4 Hz and above) were eliminated, in accordance with those described as the parameters of normal head movement in the British Journal of Ophthalmology (Gresty et al., 1976) and fall within the range of ordinary head movement as described by Hadar et al. (1983). Secondly, in line with previous studies (Cerrato \& Svanfeldt, 2006), head nods were identified as vertical movements at a speed $>0.3 \mathrm{~mm} /$ frame, with 7 frames between the top and bottom of the movement.


Figure 3: Indexing dialogue role through speaker head orientation

Speaker Orientation and Recipient role The speech transcript was synchronized with the 3D motion capture data, identifying the identity of the speaker(s) in each frame of interaction. In order to identify the speaker's primary addressee at each point in the dialogue the technique described in Healey \& Battersby (2009)
was used. For each frame of data the speakers' head orientation is calculated using the coordinates of their four head markers. The orientation of the speakers' head is compared to a centre line falling between the speakers' two interacting partners, bisecting the interaction space (Figure 3). Head orientations falling within two degrees of the centre line are excluded. If the speaker's head orientation falls on one side of this line the person on that side is coded as the primary recipient i.e. the person the speaker is primarily orienting to at that point in the dialogue. The other participant is coded, by default, as the secondary recipient. The identity of the speaker (based on hand annotated speech) and the primary and secondary recipients (based on speaker head orientation) is coded for each frame of data. Although in principle head orientation is independent of gaze direction it is nonetheless a reliable indicator of speaker's attention and gaze, especially in multi-party dialogue (Healey \& Battersby, 2009; Jokinen et al., 2010; Loomis et al., 2008).

## Results

Following Boker et al. (2002), windowed crosscorrelations were used to determine the degree of coordination between the head nodding of each participant (i.e. speaker, primary recipient and secondary recipient) and the speaker's speech and repair rates. This method directly compares the rates of speakers' speech and repair at each frame with the head movement of each participant on a lagged frame-by-frame basis within each 30-second window providing: (i) the correlation between speakers' rate of self-repair/speech and participants' nodding, and (ii) the temporal offset at which they occur. Consecutive windows were overlapped to minimize the chance of significant correlations being undetected. Windowed cross-correlation analyses assume local stationarity within each window. Although this may not always be the case, any violations will produce a downward bias of correlation and lag, providing a conservative measure of the magnitude of the effects (as discussed in Boker et al., 2002).

Figure 4 shows the results of the cross-correlation of nodding with speech rate, at lags of up to $\pm 240$ frames (4 seconds). At zero offset speakers nod most, primary recipients nod less and side participants nod least. The Friedman comparisons in Table 1 shows this global pattern of differences between roles is reliable. As the Figure shows, the difference in roles is greatest at zero offset. This is consistent with a pattern in which speakers nod most, primary participants produce some feedback through nods and secondary participants supress their nodding, as indicated by the negative correlation. Since all participants take all roles in this task these effects are only due to differences in who the speaker is looking at.

The cross-correlation of nodding with repair rate, illustrated in Figure 5, shows a different pattern of timing


Figure 4: Cross-Correlation of Speech Rate and Rate of Nodding. Horizontal grey lines indicate the $95 \%$ confidence interval.
and level of responses to repair rate than to speech rate. As Table 1 shows speakers still nod more than primary or side participants in turns that include repairs, however both people in the recipient role at the time the of the repair nod significantly more than they would otherwise. Especially in the 1-3 second offset, i.e. towards the end of the turn involving a repair.

| Pairwise comparison | Friedman's test stat |  |  |  |
| :--- | :--- | ---: | ---: | ---: |
|  |  | Raw | Std | p |
| Repair Rate | Speaker Primary | -0.784 | -6.497 | $<0.001$ |
|  | Speaker Side | 1.042 | 8.634 | $<0.001$ |
|  | Primary Side | 1.825 | 15.131 | $<0.001$ |
| Speech Rate | Speaker Primary | 0.574 | 4.757 | $<0.001$ |
|  | Speaker Side | 1.757 | 14.562 | $<0.001$ |
|  | Primary Side | 1.183 | 9.806 | $<0.001$ |
| Repair vs Speech | Primarer | -0.287 | -2.378 | 0.261 |
|  | Side | 1.071 | 8.875 | $<0.001$ |
|  | 0.428 | 3.550 | 0.006 |  |

Table 1: Non-parametric test results for crosscorrelations by role and speech or repair rate pairwise comparisons

Friedman pairwise comparisons show no reliable difference in speakers nodding as predicted by speech rate or repair rate but both recipient roles show a significantly stronger response to repair rate. Secondary participants, in particular, shift from suppressing their nodding be-


Figure 5: Cross-Correlation of Repair Rate and Rate of Nodding. Horizontal grey lines indicate the $95 \%$ confidence interval.
haviour while the speaker is addressing someone else to a profile much more similar to that of a primary participant especially at offsets of between 1 and 3 seconds.

## Discussion

Despite the fact that all three people involved in the balloon task dialogues are active, ratified participants who are free to respond at any time, the results indicate that there are clear differences in levels of responsiveness depending on who the current speaker is attending to as indexed by theier head orientation. This is consistent with previous work by Goodwin (1979) and J. B. Bavelas et al. (2000); J. Bavelas et al. (2006) who emphasise the importance of speaker gaze in eliciting listener responses.

The results reported here extend existing findings in two ways. Previous experimental work has focussed on the behaviour of listeners in dyadic i.e. two-person dialogues. Here we extend this to three person dialogues. A pragmatic feature of three-person dialgoues is that it becomes harder to judge who is speaking to whom and, as a result, more difficult to co-ordinate the roles of speaker and addressee.

Our results demonstrate that in this context there are concurrent differences in people's levels of responsiveness that depend on whether the speaker is currently oriented to them or to someone else in the conversation; indepen-
dently of what is being said. This replicates findings reported by Healey \& Battersby (2009), for a different corpus, which indicated that listerners who are not oriented to by the speaker, i.e. secondary recipients, are normally less responsive than primary recipients.

Consequently, it is not merely exposure to the content of what is said that determines responsiveness. Interestingly, these results also show for the first time that secondary participants actually suppress non-verbal feedback i.e. their head movements are substantially negatively correlated with the speaker's speech. It appears likely that this is because they are, in a sense, actively displaying their non-recipiency.

Importantly, the present results also suggest the influence of a new factor on overt levels of response; selfrepair or speaker troubles. Although there is no overall effect on Speaker's nodding, Listeners in both the primary and secondary recipient roles respond more strongly to turns in which there is evidence that the speaker is having trouble formulating or articulating their message. This is significant, in part, because selfrepairs are relatively common, occuring in at least a third of turns in natural dialogue even on conservative estimates (Colman \& Healey, 2011). The effect is more marked for secondary recipients who switch from suppressing their responses to producing a profile much closer to that of the primary recipient.

The implication of these differences in patterns of repsonsiveness is that it listener feedback is primarily organised around the successful construction of a turn, not the content of that turn. This strengthens the view that conversation is an active, collaborative process in which people make concerted use of the resources available to them, including speech, gesture and head movements, to produce each turn. However, it also suggests that these resouces are most actively used to help speakers recover from problems in the production of their turn and not, as normally assumed, for acknowledging or 'grounding' new information.

## Acknowledgments

We gratefully acknowledge the support of the Engineering and Physical Sciences Research Council (EPSRC) Doctoral Training Programme (EP/P502683/1).

## References

Bavelas, J., Coates, L., \& Johnson, T. (2006). Listener responses as a collaborative process: The role of gaze. Journal of Communication, 52(3), 566-580.
Bavelas, J. B., Coates, L., \& Johnson, T. (2002). Listener responses as a collaborative process: The role of gaze. Journal of Communication, 52(3), 566-580.
Bavelas, J. B., Coates, L., Johnson, T., et al. (2000). Listeners as co-narrators. Journal of personality and social psychology, 79(6), 941-952.

Boker, S., Rotondo, J., Xu, M., \& King, K. (2002). Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. Psychological Methods, 7(3), 338.
Cerrato, L., \& Svanfeldt, G. (2006). A method for the detection of communicative head nods in expressive speech. Gothenburg papers in theoretical linguistics(92), 153-165.
Clark, H. H. (1996). Using language. Cambridge University Press.
Clark, H. H., \& Wilkes-Gibbs, D. (1986). Referring as a collaborative process. Cognition, 22, 1-39.
Colman, M., \& Healey, P. G. T. (2011). The distribution of repair in dialogue. In Proceedings of the 33rd annual meeting of the cognitive science society (pp. 1563-1568). Boston, MA.
Crasborn, O., \& Sloetjes, H. (2008). Enhanced ELAN functionality for sign language corpora. In 3rd workshop on the representation and processing of sign languages: Construction and exploitation of sign language corpora.
Eshghi, A. (2009). Uncommon Ground: The Distribution of Dialogue Contexts. Unpublished doctoral dissertation, Queen Mary University of London.
Goffman, E. (1955). On face-work: an analysis of ritual elements in social interaction. Psychiatry: Journal for the Study of Interpersonal Processes.
Goffman, E. (1981). Forms of talk. University of Pennsylvania Press.
Goodwin, C. (1979). The interactive construction of a sentence in natural conversation. In G. Psathas (Ed.), Everyday language: Studies in ethnomethodology (pp. 97-121). New York: Irvington Publishers.
Goodwin, C. (1981). Conversational organization: Interaction between speakers and hearers. New York: Academic Press.
Gresty, M., Leech, J., Sanders, M., \& Eggars, H. (1976). A study of head and eye movement in spasmus nutans. British Journal of Ophthalmology, 60(9), 652-654.
Hadar, U., Steiner, T., Grant, E., \& Rose, F. C. (1983). Kinematics of head movements accompanying speech during conversation. Human Movement Science, 2(1), 35-46.
Healey, P. G. T., \& Battersby, S. A. (2009). The interactional geometry of a three-way conversation. In N. Taatgen \& H. van Rijn (Eds.), Proceedings of the thirty-first annual conference of the cognitive science society.
Jokinen, K., Nishida, M., \& Yamamoto, S. (2010). On eye-gaze and turn-taking. In Proceedings of the 2010 workshop on eye gaze in intelligent human machine interaction (pp. 118-123).

Loomis, J., Kelly, J., Pusch, M., Bailenson, J., \& Beall, A. (2008). Psychophysics of perceiving eye-gaze and head direction with peripheral vision: Implications for the dynamics of eye-gaze behavior. Perception, 37(9), 1443-1457.
McCabe, R., Lavelle, M., Bremner, S., Dodwell, D., Healey, P. G. T., Laugharne, R., ... Snell, A. (in preparation). Shared understanding in psychiatristpatient communication: Association with treatment adherence in schizophrenia.
Schober, M. F., \& Clark, H. H. (1989). Understanding by addressees and overhearers. Cognitive Psychology, 21, 211-232.
Yngve, V. H. (1970, April). On getting a word in edgewise. In Papers from the 6th regional meeting of the chicago linguistic society (p. 567-578).
Zellner, B. (1994). Pauses and the temporal structure of speech. In E. Keller (Ed.), Fundamentals of speech synthesis and speech recognition (p. 41-62). Chichester: John Wiley.

# A Computational Model of Systems Memory Reconsolidation 

Peter Helfer (peter.helfer@mail.mcgill.ca)<br>Department of Psychology and Integrated Program in Neuroscience, McGill University, 1205 Penfield Avenue Montreal QC, Canada H3A 1B1<br>Thomas R. Shultz (thomas.shultz@mcgill.ca)<br>Department of Psychology and School of Computer Science, McGill University, 1205 Penfield Avenue<br>Montreal QC, Canada H3A 1B1<br>Oliver Hardt (oliver.hardt @me.com)<br>Department of Psychology, McGill University, 1205 Penfield Avenue Montreal QC, Canada H3A 1B1<br>Karim Nader (karim.nader@mcgill.ca)<br>Department of Psychology, McGill University, 1205 Penfield Avenue<br>Montreal QC, Canada H3A 1B1


#### Abstract

Memory reconsolidation, the re-stabilization of consolidated memories after reactivation-induced destabilization, has received considerable attention in recent years. Nevertheless, the neural processes underlying the phenomenon remain elusive. With the aim of contributing to the development of a theory in this area, we here present a computational model of reconsolidation at the "systems" level. The model is an extension of TraceLink, which has previously been used to account for a range of memory phenomena related to consolidation.


Keywords: Memory reconsolidation, neural network, connectionism.

## Introduction

The phenomenon of memory reconsolidation, the restabilization of consolidated memories after reactivationinduced destabilization, has received considerable attention in recent years with the publication of a series of studies on both animals and human subjects (Nader \& Einarsson, 2010; Nader \& Hardt, 2009). While several computer simulations have modeled consolidation after initial learning, (McClelland, McNaughton, \& O’Reilly, 1995; Murre, 1996), only one model of cellular reconsolidation has been published (Osan, Tort, \& Amaral, 2011), and - to our knowledge - no simulation of systems reconsolidation (Debiec, LeDoux, \& Nader, 2002). In order to fill this gap, we developed an extended version of a previously published computational model of memory consolidation, TraceLink (Murre, 1996), incorporating features that enable it to also account for reconsolidation.
We begin with a brief introduction to the phenomenon of memory consolidation, followed by a description of the TraceLink model. We then discuss the mechanisms believed to underpin systems memory reconsolidation, describe how we implemented them in the model, and, finally, report our simulation results.

## Memory Consolidation

Forgetting and amnesia. The ability to recall acquired memories normally diminishes with time elapsed since learning. Although there is disagreement about the precise shape of the forgetting curve (Anderson \& Tweney, 1997), it is often represented as an exponential so-called Ebbinghaus (1885) forgetting curve, as in Figure 1.


Figure 1: Idealized normal forgetting curve.
In contrast with normal forgetting, memory loss after trauma affects recent memories more than remote ones (McClelland et al., 1995; Scoville \& Milner, 1957; Squire \& Alvarez, 1995), resulting in a curve with the opposite slope, as in Figure 2.


Figure 2: Idealized Ribot gradient.
This graph shows that the ability to recall material learned shortly before onset of amnesia is strongly impaired, whereas older memories are relatively spared. The curve is commonly known as the "Ribot gradient", after the French psychologist Ribot who first postulated it (Ribot, 1882). This temporally graded amnesia gave rise to the idea that a
consolidation process stabilizes newly acquired memories older memories were less affected in amnesia because they had had more time to stabilize.

Types of consolidation. Researchers distinguish between two types of memory consolidation, "systems" consolidation and "synaptic" or "cellular" consolidation (Dudai \& Morris, 2000). Systems consolidation is a process that transitions initially hippocampus-dependent memories to a hippocampus-independent state. In the mammalian brain, the hippocampal formation is involved with the consolidation of "episodic" memories, explicit memories of experienced events. Animal studies as well as human cases of brain damage have shown that memories initially depend on the hippocampus, but gradually become hippocampusindependent. According to the "standard model of systems consolidation" (McClelland et al., 1995; Squire \& Alvarez, 1995), hippocampal memory traces are quickly created but only persist for a limited time, during which they support the more time-consuming construction of neocortical memories. On this view, the temporally graded amnesia observed after hippocampal lesions is due to the fact that older memories have had more time to consolidate in the neocortex, while newer memories are still only weakly represented there (McClelland et al., 1995; Squire \& Alvarez, 1995). This process is called "systems consolidation" because it involves interaction between two brain systems, the hippocampus and the neocortex. In contrast, the so-called "cellular" or "synaptic" consolidation process concerns the stabilization of memories within a single system.

## The TraceLink Model of Memory Consolidation

TraceLink is a connectionist model of systems memory consolidation (Meeter \& Murre, 2005; Murre, 1996). The model has two layers representing hippocampus (HC) and neocortex (NC), respectively. The HC layer has 42 units and the NC layer has 200 units. Each layer is fully connected, i.e. there are independent (asymmetric) connections in both directions between each pair of units, and the two layers are also fully interconnected. Connection weights have values in the range 0.0 to 1.0 . The units have discrete activation levels, either 0.0 (inactive) or 1.0 (active), and a stochastic asigmoid activation function:

$$
\begin{equation*}
P_{j}=\frac{1}{1+e^{-\frac{n e t_{j}}{\text { temp }}}} \tag{1}
\end{equation*}
$$

where $P_{j}$ is the probability that unit $j$ will become (or remain) active, net $_{j}$ is the net input to unit $j$ and temp is a parameter that controls the steepness of the asigmoid function, i.e. the amount of randomness in the model. (For small values of temp, $P_{j}\left(\right.$ net $\left._{i}\right)$ approaches a deterministic step function; for large temp, $P_{j}\left(\right.$ net $\left._{i}\right)$ is close to 0.5 everywhere, i.e. equal probability of becoming active or inactive regardless of net $_{i}$.). A temp value of 0.2 was used in all simulations. The net input $n e t_{j}$ in equation [1] is calculated according to the following formula:

$$
\begin{equation*}
\text { net }_{j}=\sum_{i} w_{i j} a_{i}-\text { inhibition }_{L} \tag{2}
\end{equation*}
$$

where $w_{i j}$ is the weight of the connection from unit $i$ to unit $j$, and $a_{i}$ is the activation level of unit $i$. The term inhibition $n_{L}$ is a layer-specific inhibition quantity that simulates the effect of inhibitory synapses. It is calculated by a feedback algorithm that drives the number of active units in each layer towards a configured equilibrium value, which is also the number of active units in training patterns for the layer. For example, each training pattern for the NC layer has ten active units, and the inhibition mechanism makes the layer preferentially settle into states with that number of active units.

The learning rule is Hebbian with an anti-Hebbian "interference" term that accelerates forgetting of previously learned patterns, especially in the smaller HC layer, where there is more pattern overlap:

$$
\begin{equation*}
w_{i j}(t+1)=w_{i j}(t)+\mu_{T}^{+} a_{i} a_{j}-\mu_{T}^{-}\left(1-a_{i}\right) a_{j} \tag{3}
\end{equation*}
$$

where $w_{i j}(t)$ is the connection weight between units $i$ and $j$ at time $t, a_{i}$ is the activation level of unit $i, \mu_{T}^{+}$is the Hebbian learning rate, and $\mu_{T}^{-}$is the interference or "unlearning" rate. The learning rule strengthens connections between units that are both active, and weakens connections from inactive to active units. Learning rates are specified per "tract" (hence the $T$ subscript). A tract is a set of connections with the same source and destination layers: all the connections from HC units to NC units form one tract, all connections internal to the NC layer form another tract, etc. A tract's learning rates ( $\mu_{T}^{+}$and $\mu_{T}^{-}$) may take on different values during initial acquisition versus consolidation. This simulates the effect of neuromodulation, for example, an increased learning rate in hippocampus in the presence of novel stimuli (Meeter \& Murre, 2005; Murre, 1996).

Initial acquisition. The TraceLink system is trained by presenting a training pattern to both layers ${ }^{1}$ and applying the learning rule to adjust connection weights. The intra-HC and NC-HC tracts have high learning rates and learn patterns well in a single presentation. The intra-NC tract has a much lower learning rate, and as a result a single training cycle only creates a weak trace there.

Recall. To test recall of a training pattern, a subset of the pattern's active NC units (a "cue pattern") are held ("clamped") in the "on" state, and the rest of the units in both layers are randomly set to either the active or inactive state, with equal probability. The whole system is then repeatedly cycled by executing the activation function for all the unclamped units in random order and updating their activation levels accordingly. At the end of each cycle, the inhibition algorithm adjusts the inhibition coefficients of both layers. After a configurable number of such cycles (we

[^295]used 70 in all simulations), the activation pattern into which the system has settled is compared to the original training pattern. Recall accuracy is measured as the percentage of non-cued NC units in the training pattern that have been successfully turned on.

Lesioning. Hippocampal lesion is simulated by simply disconnecting the HC layer (setting all inter-layer connections weights to zero). After initial training, the intact system can normally recall patterns quite well, because the $\mathrm{NC}-\mathrm{HC}$ and $\mathrm{HC}-\mathrm{HC}$ connections provide linkage between the pattern's NC units, but after virtual lesioning recall is poor, because the NC-NC connections are not strong enough to independently enable the system to complete the pattern correctly.

Consolidation. Memory consolidation is simulated by randomly setting each unit's activation level to either 0.0 or 1.0 , letting the system "settle" in the same manner as for recall (but without any cue pattern), and reinforcing whatever state it settles into by applying the learning rule in the NC layer. Because the system is more likely to settle into trained patterns (Hopfield, 1982), this procedure gradually strengthens those patterns in the NC layer. After a pattern has been reinforced in this manner a sufficient number of times, its NC connections become strong enough that the pattern can be recalled even after HC lesioning.

Simulations. In a typical TraceLink simulation, a series of training patterns are presented, one per simulated "day", each followed by a number of consolidation cycles (Meeter \& Murre, 2005; Murre, 1996). Because of interference, especially in the smaller HC layer where patterns overlap more, earlier patterns are gradually overwritten by newer ones. When recall is tested after training a number of patterns, a forgetting curve can be observed: older patterns are recalled less successfully than newer ones. The model is thus able to account for normal forgetting (the idea that interference plays a major role in hippocampal forgetting may be debatable (Hardt, Nader, \& Nadel, 2013)).
While patterns are slowly forgotten in the HC layer, they are gradually strengthened in the NC layer due to consolidation. If the HC layer is "lesioned" after a number of days, the earlier training patterns, which have had more time to consolidate and therefore have a stronger NC representation, are recalled more successfully than the newer ones. The model is thus also able to account for the Ribot gradient observed after hippocampal lesion. See Meeter \& Murre (2005), for more details about the TraceLink model, including accounts of simulations that reproduce a range of human memory phenomena.

## Memory Reconsolidation

It has been shown that reactivating a consolidated memory can return it to a labile state, from which it needs to reconsolidate in order to persist (Nader \& Hardt, 2009). During the period of instability, the so-called "reconsolidation window", memory impairments may be
produced by the same types of intervention that can interfere with initial consolidation, such as lesions and protein synthesis inhibition (Debiec et al., 2002; Nader, Schafe, \& Le Doux, 2000). Some have suggested that that such postreactivation plasticity allows knowledge to be modified when new information is acquired (Hardt, Einarsson, \& Nader, 2010; Lee, 2009). As is the case with memory consolidation, memory reconsolidation has been documented at both the systems and cellular level. The former type, systems reconsolidation, is "the demonstration that reactivation of a remote memory returns the trace to being hippocampus dependent again for a period of time before once again becoming independent of hippocampus" (Debiec et al., 2002).

## Method

Although the physiological events underlying systems memory reconsolidation are not known, researchers have proposed hypothetical mechanisms that could explain the observed phenomena. The present work is a neural-network model of such a hypothesis (Debiec et al., 2002; Hardt et al., 2010; Nadel \& Hardt, 2010; Nader et al., 2000). According to this hypothesis, (1) consolidation renders remote memories hippocampus-independent; (2) reactivation of a consolidated neocortical memory creates a temporary hippocampal trace (or strengthens the existing but decaying trace); (3) the hippocampal trace stimulates the neocortical trace through back-projections; (4) this stimulation has the effect of initially destabilizing the neocortical synapses, making them susceptible to decay and/or modification; (5) continued hippocampal reinforcement prevents decay of (or even strengthens) the neocortical trace while it restabilizes. The model thus provides an explanation for the observed fact that reactivation followed by hippocampal lesion produces amnesia, but neither reactivation nor lesion alone causes memory loss.

## Implementation

In order to model this hypothesis, we implemented a twolayer network along the lines of TraceLink, but with a few additional features: (a) connections have a plasticity attribute; (b) connection weights are subject to time-based decay (Hardt et al., 2013); and (c) the simulation now includes a "reactivation" phase to trigger memory reconsolidation.

Plasticity. The plasticity attribute has a value between 0.0 and 1.0 , representing minimum and maximum plasticity, respectively. Our new learning rule takes plasticity into account:

$$
\begin{equation*}
w_{i j}(t+1)=w_{i j}(t)+p_{i j}\left(\mu_{T}^{+} a_{i} a_{j}-\mu_{T}^{-}\left(1-a_{i}\right) a_{j}\right) \tag{4}
\end{equation*}
$$

where $p_{i j}$ is the plasticity of the connection from unit $i$ to unit $j$. Thus the plasticity affects a connection's sensitivity to training and also its susceptibility to interference.

Connections are created with a $p_{i j}$ value of 1.0 (fully plastic), which subsequently decreases exponentially over simulated time, as expressed by the following formula:

$$
\begin{equation*}
p_{i j}(t+1)=p_{i j}(t) \cdot\left(1-p d r_{T}\right) \tag{5}
\end{equation*}
$$

where $p d r_{T}$ is a plasticity decay rate specific to the tract to which connection $i j$ belongs. In the simulations reported here, the $p d r_{T}$ value was 0.1 for the NC-NC tract, and 0.0 for the other tracts, i.e. plasticity variations in hippocampus were not simulated.

Decay. Connection weights are subject to exponential decay at a rate that is configurable on a per-tract basis. A connection's weight decays by its decay rate modulated by its plasticity, according to the following formula:

$$
\begin{equation*}
w_{i j}(t+1)=w_{i j}(t) \cdot\left(1-p_{i j} w d r_{T}\right) \tag{6}
\end{equation*}
$$

where $w d r_{T}$ is the weight decay rate specified for the tract to which the connection belongs. Thus, as a connection becomes less plastic, it becomes more resistant to decay (Hardt et al., 2013).

Reactivation. In addition to TraceLink's "Acquisition" and "Consolidation" phases, our model has a "Reactivation" phase, during which one or more previously trained patterns are activated, the learning rule [4] is applied, and the plasticity between active units is restored to its maximum value 1.0. Following reactivation, a number of consolidation periods may be executed, as after initial learning.

## Simulations

The following simulations were carried out:
A. Consolidation

1. Train a single pattern.
2. Execute 40 consolidation periods (simulated "days"). At each day, test recall in the intact system and with "lesioned" (deactivated) HC layer.
B. Reactivation/Reconsolidation

Same procedures for training, consolidation and testing as in simulation A , but on day 20 , reactivate the trained pattern, then continue daily consolidation and testing.
C. Reactivation and HC lesion

Same procedure as in simulation B, but on day 21, permanently lesion the HC layer.
The same parameter settings were used in all three simulations, as indicated in Table 1.

An explanatory note about the daily recall tests with intact and "temporarily lesioned" HC : these tests are performed without affecting the continued evolution of the system. No learning or (re)consolidation takes place, and HC is turned back on after testing. The simulation then continues as if the tests had not taken place. Researchers with live subjects, of course, do not have this luxury; in an analogous experiment, they would only be able to get one data point from each subject.

Table 1: Parameter values used in the simulations

| Parameter | Values |  |
| :--- | :--- | :--- |
|  | NC | HC |
| Learning rate during initial acquisition | 0.06 | 0.4 |
| Learning rate during consolidation | 0.02 | 0.0 |
| Learning rate during reactivation | 0.0 | 0.2 |
| Unlearning rate | $75 \%$ of learning |  |
|  | rate |  |$|$| Weight decay rate | 0.1 | 0.1 |
| :--- | :--- | :--- |
| Plasticity decay rate | 0.1 | 0.0 |
| Number of units | 200 | 42 |
| Active units at equilibrium (=pattern size) | 10 | 7 |
| Cue pattern size (units) | 5 | 0 |

The values in the "NC" column apply to the NC layer and intra-NC tract. The values in the "HC" column apply to the HC layer, intra-HC tract and inter-layer tracts.

## Results

## A. Consolidation

Figures 3 a and 3 b show the weight and plasticity of a representative individual connection in the $\mathrm{HC}-\mathrm{HC}$ and NC NC tracts, respectively, during the consolidation simulation. Each of the two monitored connections joined two units that were simultaneously active in the training pattern, i.e. they were connections where significant Hebbian learning took place.


Figure 3: Consolidation. a) Connection weight of a hippocampal connection. b) Weight and plasticity of a neocortical connection. c) Recall performance (averaged results from fifty simulations). Each point on the "lesioned" curve shows the performance with deactivated HC , i.e. as if HC had been lesioned on that day. Vertical bars show standard error.

As expected, HC connections quickly learn the presented pattern, and then decay exponentially. NC connections, on the other hand, quickly become very plastic, but learn only gradually. Around day 17 the HC trace has become too faint for any further consolidation to take place, and the NC trace starts to decay somewhat, but the decay slows down as the plasticity diminishes further and the trace becomes stabilized.
Figure 3c shows the recall performance during the simulation. The upper curve, representing recall in the intact system, shows normal forgetting. The lower curve, recall performance with disabled HC layer, shows a gradient during the consolidation "window", followed by constant performance. These results are similar to those obtained with the original TraceLink model (Meeter \& Murre, 2005); the difference is that forgetting there was purely interference-based, whereas in this simulation it is caused by a combination of interference and decay. (Interference plays a role even though only a single pattern is trained, because the patterns reinforced during (re)consolidation may differ from the trained pattern.)

## B. Reconsolidation

As shown in Figure 4, if the pattern is reactivated on day 20, then (a) the hippocampal trace is rapidly strengthened, (b) the necocortical trace is quickly destabilized and then gradually strengthened and restabilized in a round of reconsolidation, and (c) the recall performance is somewhat improved after the reminder.


Figure 4: Reconsolidation. a) Connection weight of a hippocampal connection. b) Weight and plasticity of a neocortical connection. c) Recall performance (averaged results from fifty simulations).

## C. Reactivation followed by HC lesion

When the HC layer is permanently lesioned after memory reactivation, the results are as illustrated in Figure 5: (a) The
hippocampal trace decays after initial training as in the previous simulation and is boosted by the reactivation


Figure 5: HC lesioning following reactivation. a)
Connection weight of a hippocampal connection. b) Weight and plasticity of a neocortical connection. c) Recall performance (averaged results from fifty simulations). The points on the "intact" curve after day 21 show the performance of the lesioned system.
on day 20. The plot ends at the hippocampal lesion on day 21. (b) The neocortical trace evolves as in experiment B until day 20, the day of reactivation. Following the HC lesion on day 21, instead of being strengthened by reconsolidation, the destabilized NC trace rapidly decays. (c) The recall performance shows rapid onset of amnesia after the hippocampal lesion.

## Discussion

In spite of a growing number of studies on both humans and animals, the neural mechanisms underlying memory reconsolidation are not well understood. The present paper seeks to contribute to the development of a theory by introducing a computational model of reconsolidation.

The key finding in system memory reconsolidation studies is that lesioning after reactivation produces amnesia, whereas neither reactivation alone nor lesioning alone causes memory impairment (Debiec et al., 2002; Nader \& Hardt, 2009). With this in mind, it is interesting to compare Figures 3-5. Figure 3c shows that, once a memory is consolidated in the model, hippocampal lesions without preceding memory reactivation have little effect on it, whereas Figure 5c illustrates that post-reactivation lesions lead to a dramatic drop in recall performance. The cause of this difference is that, after reactivation, the plasticity of the neocortical trace is high, allowing for rapid decay. In Figure 4 c , on the other hand, where hippocampus is left intact after
reactivation, reconsolidation more than compensates for the decay, resulting in moderate strengthening of the memory trace after reactivation.
The neural network model presented here is able to reproduce the empirical results by simulating microprocesses that have been hypothesized to underlie memory reconsolidation - controlled variability in synaptic plasticity and plasticity-dependent synaptic decay rates - and thus demonstrates that these mechanisms in fact can account for the observed effects.
An interesting aspect of this model is that it introduces decay-driven forgetting, in contrast with the TraceLink simulations, where all forgetting was due to interference (Meeter \& Murre, 2005). It is likely that both types of mechanism play important roles in the consolidation and maintenance of memories (Hardt et al., 2013), and we are planning to apply the model to further investigate the relationship between the two. In particular, work in progress includes simulations with multiple training patterns, which will allow us to study the combined effects of decay and even greater interference.
Another direction in which we are planning to extend this work is to apply the model to manifestations of reconsolidation other than amnesia after hippocampal lesions. These include the effects of protein synthesis inhibition (Debiec et al., 2002; Nader et al., 2000) and interference training in the reconsolidation window (Hupbach, Gomez, Hardt, \& Nadel, 2007; Hupbach, Gomez, \& Nadel, 2009; Walker, Brakefield, Hobson, \& Stickgold, 2003).

## Acknowledgments

This research was supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada to TRS. OH and KN were supported by the Canadian Institutes of Health Research and by the Natural Sciences and Engineering Research Council of Canada. OH was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG).

## References

Anderson, R. B., \& Tweney, R. D. (1997). Artifactual power curves in forgetting. Memory \& Cognition, 25(5), 724-730.
Debiec, J., LeDoux, J. E., \& Nader, K. (2002). Cellular and systems reconsolidation in the hippocampus. Neuron, 36(3), 527-538.
Dudai, Y., \& Morris, R. (2000). To consolidate or not to consolidate: what are the questions? In Brain, Perception, Memory. Advances in Cognitive Sciences (Bolhuis J.J. $E d$ ). Oxford University Press.
Ebbinghaus, H. (1885). Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie. Duncker \& Humblot.
Hardt, O., Einarsson, E. Ö., \& Nader, K. (2010). A bridge over troubled water: reconsolidation as a link between
cognitive and neuroscientific memory research traditions. Annual Review of Psychology, 61, 141-167.
Hardt, Oliver, Nader, K., \& Nadel, L. (2013). Decay happens: the role of active forgetting in memory. Trends in Cognitive Sciences.
Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences, 79(8), 2554-2558.
Hupbach, A., Gomez, R., Hardt, O., \& Nadel, L. (2007). Reconsolidation of episodic memories: A subtle reminder triggers integration of new information. Learning \& Memory, 14(1-2), 47-53.
Hupbach, A., Gomez, R., \& Nadel, L. (2009). Episodic memory reconsolidation: updating or source confusion? Memory, 17(5), 502-510.
Lee, J. L. C. (2009). Reconsolidation: maintaining memory relevance. Trends in Neurosciences, 32(8), 413-420.
McClelland, J. L., McNaughton, B. L., \& O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. Psychological Review, 102, 419-457.
Meeter, M., \& Murre, J. M. J. (2005). Tracelink: A model of consolidation and amnesia. Cognitive Neuropsychology, 22(5), 559-587.
Murre, J. M. J. (1996). TraceLink: a model of amnesia and consolidation of memory. Hippocampus, 6(6), 675-684.
Nadel, L., \& Hardt, O. (2010). Update on memory systems and processes. Neuropsychopharmacology, 36(1), 251273.

Nader, K., \& Einarsson, E. O. (2010). Memory reconsolidation: an update. Annals of the New York Academy of Sciences, 1191, 27-41.
Nader, K., \& Hardt, O. (2009). A single standard for memory: the case for reconsolidation. Nature Reviews Neuroscience, 10, 224-234.
Nader, K., Schafe, G. E., \& Le Doux, J. E. (2000). Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. Nature, 406(6797), 722726.

Osan, R., Tort, A. B. L., \& Amaral, O. B. (2011). A Mismatch-Based Model for Memory Reconsolidation and Extinction in Attractor Networks. PLoS ONE, 6(8), e23113.
Ribot, T. A. (1882). Diseases of Memory, an Essay in the Positive Psychology. New York, NY: D. Appleton and Company.
Scoville, W. B., \& Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. Journal of Neurology, Neurosurgery, and Psychiatry, 20(1), 11-21.
Squire, L., \& Alvarez, P. (1995). Retrograde-amnesia and memory consolidation - a neurobiological perspective. Current Opinion in Neurobiology, 5(2), 169-177.
Walker, M. P., Brakefield, T., Hobson, J. A., \& Stickgold, R. (2003). Dissociable stages of human memory consolidation and reconsolidation. Nature, 425(6958), 616-620.

# Expert Blind Spot in Pre-Service and In-Service Mathematics Teachers: Task Design moderates Overestimation of Novices' Performance 

Katharina Hellmann (katharina.hellmann@ezw.uni-freiburg.de)<br>Matthias Nückles (matthias.nueckles@ezw.uni-freiburg.de)<br>University of Freiburg, Department of Educational Science, Instructional and School Research<br>Rempartstraße 11, 79098 Freiburg, Germany


#### Abstract

To act efficiently in the classroom, teachers need to be able to judge the difficulty of problems from a novice's perspective. However, research suggests that experts use their own knowledge as an anchor, adjust estimations for others to their own knowledge and thus underestimate the difficulty that a problem may impose on novices. Similarly, experts should underestimate the benefit for novices of task designs derived from Cognitive Load Theory (CLT), as - following the expertise reversal effect - these should be rather disadvantageous for experts. We investigated pre-service and in-service teachers’ competencies in estimating the difficulty of mathematical tasks for novices. Thirty-four pre-service teachers and thirteen experienced teachers solved tasks that varied in instructional design (optimized for novices following CLT versus non-optimized). Participants solved each task and then estimated how many students of a fictional $9^{\text {th }}$ grade class would be able to solve that task. Solution frequencies were collected from fifty-two $9^{\text {th }}$ grade students. In both expert groups, overestimation was clearly more pronounced for non-optimized than optimized tasks, suggesting an expert blind spot that can be explained in terms of an expertise-reversal effect. The experts failed to adequately take into account the benefits of didactic task variation for novice learners. However, whereas pre-service teachers' overestimations of student performance were large and significant both for non-optimized and optimized tasks, in-service teachers' overestimations were generally small and failed to approach statistical significance. In contrast to preservice teachers, in-service teachers seem to have a better mental model of what a student is able to achieve, thus making better judgments of student performance.


Keywords: expert blind spot; perspective taking; expertise reversal effect

## Theoretical Background

## Expert Blind Spot

Peoples’ judgements of others are very often based on their self-assessment and are therefore cognitively biased (e.g. Tversky \& Kahneman, 1974). In line with this, research in the area of expertise has repeatedly shown that experts tend to misjudge novices' knowledge, achievement, or time on task, amongst others, to a certain degree (e.g. Herppich, Wittwer, Nückles \& Renkl, 2010; Hinds, 1999; Lentz \& de Jong, 2006). This effect has also been found to apply to teachers (e.g. Nathan \& Koedinger, 2000). Teachers, usually referred to as being domain experts in their content area, as they possess a high level of specialized knowledge, are considered to be prone to an expert blind spot (Nathan \&

Petrosino, 2003) when evaluating the difficulty of mathematical problems for students

Following Nickerson’s (1999) anchoring model, teachers may be inclined to use their specialized knowledge as an anchor when assessing the difficulty of problems for students. As a result, they are not able to take the student perspective adequately.


Figure 1: The process of perspective taking through anchoring and adjustment (adapted from Nickerson, 1999)

According to Nickerson (1999, see Figure 1), people tend to build an inaccurate mental model of the potential knowledge of general or specific others. They fail to take into account the specificity or exclusivity of their own knowledge, therefore unconsciously using it as an anchor when estimating other persons’ knowledge. As a result, teachers might underestimate the difficulty that a problem will impose on a student, and overestimate students’ performance.

However, the ability to adequately assess the difficulty of tasks for students is a crucial aspect of teaching expertise. It is necessary for communicating efficiently with students as well as for adapting teaching behaviour in and outside the classroom (e.g. selecting problems for homework, lessons or exams). Teachers should be able to take a novice's perspective and judge task attributes independently of their own perception of difficulty or effortlessness (Helmke, Hosenfeld \& Schrader, 2004).

## Cognitive Load Theory

Cognitive Load Theory (CLT; e.g. Sweller, 2005; Sweller, van Merrienboer \& Paas, 1998) can help to understand how and why experts and novices differ in their perceptions of
task difficulty and how one could deal with these discrepancies.

Working Memory Capacity And Perceived Task Difficulty According to CLT, every learning process is associated with cognitive mental load. The extent, to which a learner experiences this mental load, depends on the degree of the learner's expertise regarding the subject. Experienced learners use already existing knowledge structures, so called schemas. Schemas serve as patterns that help to structure and integrate incoming information (Sweller, 1994). But often, new information is being processed that needs the learner's full working memory capacity. If cognitive schemas do not exist and yet have to be built, working memory, the capacity of which is limited, is loaded to a high extent.
Perceived task difficulty according to CLT should mainly be a function both of the intrinsic mental load imposed on the learner (i.e., the complexity or difficulty of a task) and the amount of extraneous mental load (i.e., the load induced by an ineffective instructional design of the task). Generally, extraneous load has been shown to be an important factor hindering effective learning (e.g. Paas \& van Merrienboer, 1994). Intrinsic load cannot be influenced, as it is inherent to the task itself and can only be moderated by the amount of a learner's prior knowledge. In contrast, extraneous load can and should be reduced. Once working memory is disburdened of extraneous load, more working memory capacity is available for understanding and schema acquisition.

Instructional Techniques Reducing Extraneous Mental Load Novice learners should be provided with learning material designed according to principles derived from CLT. The main principles are: integrated-format (Sweller, 2005), step-by-step-guidance (Kalyuga, Chandler \& Sweller, 2001) or worked examples (Renkl, 2005). Tasks following these design principles substantially reduce the amount of extraneous load imposed on the learner.
An integrated-format in task design as compared to a split-attention design (Sweller, 2005) facilitates learning, as the learner does not have to search and integrate relevant information by himself, before passing on to the solution. With this procedure, information is presented close to each other and allows for an easier processing. A step by step guidance (e.g., Kalyuga, Chandler \& Sweller, 2001) helps the learner to solve a problem without struggling to find all needed solution steps in a correct order. Instead, a processing guideline is given, leaving more working memory capacity available for the understanding of the single steps. Worked examples (e.g., Renkl, 2005), as compared to traditional problem solving techniques, consist of a problem, elaborated solution steps and the solution itself. Again, working memory capacity is free from extraneous load, as no potentially irrelevant trial and error processes are performed. This, once again, results in better schema acquisition and deeper understanding.

Expertise Reversal Effect It is important to emphasize that the effects of the just mentioned CLT principles are only prevailing with regard to novice learners. The positive learning outcome of material that is designed for novice learners may, in contrast, be reversed for experts. The guidance or additional information given by optimized learning material (from now on, the term optimized will be used with regard to learning tasks that are designed in favour of novice learners) can interfere with experts’ advanced cognitive structures and schemas that have already been built. Kalyuga (2007) named this phenomenon expertise reversal effect. He described that an optimized learning tasks is experienced as being more difficult to process and causes a redundancy effect, when presented to expert learners. This results in increased extraneous load and worse performance.
From this follows that the same learning material may cause reversed effects for novice and experienced learners. However, as experts may perceive optimized tasks as being more difficult than non-optimized tasks, they may also be subject to an expert blind spot when assessing the potential difficulty of the tasks for novice learners. This prediction is in line with Nickerson's anchoring and adjustment model (1999). Experts judge optimized learning material as being difficult to solve, use that judgement as an anchor for estimating novices’ performance and thus underestimate novices' performance on these tasks. The opposite is true for non-optimized items, resulting in an overestimation of novices’ performance.

Teachers as domain experts and educators should be knowledgeable of this expertise reversal effect and able to estimate the difficulty of tasks for students as novice learners independently of their own experienced mental load. In the present study we investigated whether this assumption is true for two groups of mathematics experts.

## Research Questions and Predictions

In the present study, we investigated whether pre-service as well as in-service mathematics teachers are subject to an expert blind spot when judging the difficulty of problems for students and whether the two expert groups differ in their estimations. Differences in estimations can be expected due to different levels of teaching experience. The tasks presented to the expert groups varied in instructional design according to CLT, but were comparable in complexity, thus keeping intrinsic cognitive load stable.

Following our theoretical assumptions, both pre-service and in-service teachers should use their expert knowledge as an anchor and underestimate the difficulty of the tasks for novice students in general.

1) Therefore, we predicted that both expert groups would generally overestimate the amount of tasks that novice students would be able to solve correctly (overestimation hypothesis).
An anchoring effect should manifest itself in highly correlated ratings of one's own perceived mental load and estimated task performance of novice learners.
2) Hence, we predicted that the correlation between the experts' self-rated mental load and estimated task performance of novice learners is significantly larger than the correlation between estimated student performance and students' actual performance (anchoring hypothesis).
Following expertise reversal effect, we further expected that the experts would experience less mental load when solving non-optimized than optimized tasks.
3) Consequently, the overestimation of novices' performance should be significantly larger with regard to non-optimized tasks as compared with optimized tasks (expertise reversal hypothesis).

## Method

## Participants

Thirty-four pre-service teachers majoring in mathematics (mean study time being 6.12 semesters, $S D=3.18$ ) and 13 in-service mathematics teachers (mean time of working experience being 12.85 years, $S D=9.13$ ) participated in the study. Two different expert groups were chosen to allow for possible conclusions regarding experience levels. Whereas pre-service teachers usually do not have school teaching experience, the amount of in-service teachers' teaching experience could become evident in their ratings of students performance.

Both expert groups’ estimations were compared to solution frequencies collected from $549^{\text {th }}$ grade high school students (mean age being $14.26, S D=.52$ ). All participants attended the study on a voluntary basis and received financial compensation.

## Study design

We used level of expertise (pre-service teachers vs. inservice teachers vs. novices) and the instructional design of the task (non-optimized vs. optimized mathematical problems) as independent variables. Dependent variables encompassed experts' perceived mental load and performance, their estimations of novices' performance and novices' actual performance on a number of mathematical tasks. Estimations were compared to students’ actual performance

## Instrument and measures

Two mathematics experts created ten tasks on algebra, geometry and trigonometry. To achieve a high level of curriculum validity, contents of the tasks were chosen to meet the requirements expected from pupils on that $9^{\text {th }}$ grade school level (e.g. calculation of area, theorem of Pythagoras, angular sum). Each task was designed in a nonoptimized and optimized version. Tasks without didactic optimization were adapted from mathematics problems currently used in school. Tasks optimization was achieved by using one of the following CLT design principles (the latter being the optimized design): either split-attentionformat vs. integrated-format; or traditional problem solving
vs. step-by-step-guidance; or traditional problem solving vs. worked examples. A task on angular sum, for example, was either designed with help of a diagram and angular degrees being spread over the working sheet making it difficult to match needed information, or presented with a diagram and angular degrees being close to each other (optimized; integrated format). So, whereas each task covered exactly the same mathematical problem (keeping intrinsic cognitive load stable), the design of the task (extraneous cognitive load) varied, allowing for the measurement of differences in mental load and performance due to task design.

Perceived mental load was assessed by the following question adapted from cognitive load literature: "How difficult did you find working on the task?", and measured on a six-point rating-scale ranging from "not at all difficult" to "very difficult".

Teachers' estimations of student performance were collected by using a prototype description of a fictional $9^{\text {th }}$ grade high school class: "Imagine that you are the teacher of a class with 30 students, all having different achievement levels; there are very good, average and very poor students. Now, you want to use the same task that you have just worked on for an exam. How many students of this class will presumably solve the task correctly?"

Participants' task performance was measured as the number of correctly solved tasks (the maximum score being ten). Each of the participants' solutions was rated by two independent mathematics experts as being correctly or falsely solved. When no accordance could be initially found, the two experts discussed their different ratings and agreed on one in a second step.

## Procedure

Each participant received a booklet with ten tasks. Five of the tasks were presented in a non-optimized version and five were presented in an optimized version, balanced within the booklet. Furthermore, each task presented in its nonoptimized version (e.g. angular sum, split-attention-format) had a corresponding item in its optimized version (e.g. analogous angular sum task, integrated-format), placed elsewhere in the booklet. Using this method, repetition effects by having the participants solving the same task twice were avoided, but still estimations based on both task designs were collected.

The participants solved each mathematical problem within a fixed period of time. The time constraint should prevent ceiling effects from occurring. Experts, given unlimited time to solve the tasks, perceive only little to no mental load, as enough working memory capacity is free for solving most tasks correctly, no matter which design is presented. Under these circumstances, an effect of task design on experienced mental load can no longer be detected (Paas, Renkl \& Sweller, 2003).

After having solved each task, participants rated their perceived mental load on a six-point rating scale. Then, they estimated how many students of the fictional $9^{\text {th }}$-grade class would be able to solve the tasks they have just worked on
correctly. After having finished the rating process, participants continued with the next mathematical problem. At the end, demographic data was collected.

## Results

In a first step, we compared pre-service teachers' and inservice teachers' perceived mental load and performance for both item types. For this purpose, each participant's ratings and performance data was aggregated (five for nonoptimized and five for optimized tasks) and then compared with a paired t-test.
In line with CLT, pre-service teachers experienced significantly more mental load when solving optimized ( $M$ $=2.33, S D=0.67$ ) than non-optimized tasks ( $M=2.11, S D$ $=0.58), t(33)=-2.08, p<.05$. However, pre-service teachers did not significantly perform worse on optimized ( $M=75.88 \%, S D=20.17 \%$ ) than on non-optimized tasks ( $M=78.82 \% S D=16.29 \%$ ), $t(33)=0.82$, ns.

In-service teachers experienced no significantly different degree of mental load for optimized ( $M=2.72, S D=0.72$ ) and non-optimized tasks ( $M=2.59, S D=0.94$ ), $t(12)=$ 1.13, ns. Also, performance for optimized ( $M=76.92 \%, S D$ $=17.97 \%$ ) and non-optimized tasks ( $M=73.85 \%, S D=$ 18.94\%) did not differ significantly, $t(12)=-.56$, $n s$.

In a second step, students' solution frequencies were analysed. In line with CLT, the $9^{\text {th }}$-grade students solved more optimized ( $M=51.11 \%, S D=26.51 \%$ ) than nonoptimized tasks ( $M=44.07 \%, S D=24.69 \%), t(53)=2.38, p$ $<.05$ (all performance data are presented in Figure 2).

Finally, performance data between the participant groups were compared in a repeated measures ANOVA. Both preservice $(F(1,86)=44.08, \mathrm{p}<.01)$ and in-service teachers $(F(1,65)=16.69, \mathrm{p}<.01)$ solved significantly more tasks than students did, whereas performance between the expert groups $(F(1,45)=.16, n s)$ did not differ significantly.

## Overestimation Hypothesis

To test the overestimation hypothesis, we computed difference scores. Students' real solution frequencies for each item were subtracted from pre-service and in-service teachers' estimations of how many students would be able to solve this corresponding task correctly. A positive difference score thus indicated an overestimation and a negative score indicated an underestimation. Each participant's difference scores were then aggregated for item type (five scores for non-optimized and five for optimized tasks) and used for further analysis.
As predicted, pre-service teachers overestimated students' performance both on non-optimized tasks, $t(33)=6.29, p<$ .01 , and optimized tasks, $t(33)=2.34, p<.05$ (one-sample t-test). However, in-service teachers’ general overestimation of student performance did not reach statistical significance both for non-optimized $(t(12)=1.21, n s)$ and optimized tasks $(t(12)=-.13, n s)$. Overestimation scores between the expert groups did not differ significantly, $F(1,45)=3.19$, ns. (estimation data for pre-service and in-service teachers are presented in Figure 2).


Figure 2: Pre-service teachers’ (P) and in-service teachers (I) performance and estimation of student performance and students' (S) performance as function of task design (\%)

## Anchoring Hypothesis

To test for an anchoring effect, we computed and compared Fisher z transformed individual correlations. Pre-service and in-service teachers' mental load ratings for each item were correlated with their estimation of student performance for that particular item. Further, the estimation of student performance for each item was correlated with students’ actual performance on that item. This procedure allowed analysing whether experts' estimations were closer to their perceived mental load or to students' actual performance. Experts' perceived mental load (as compared to experts' actual performance on each task) was used for the analysis. Whereas performance on a task cannot be determined immediately by the participants (as it remains unclear whether they solved a task correctly or not), mental load served as adequate and approximate measure of task difficulty. The individual correlations were aggregated (five correlations for non-optimized and five for optimized tasks) and then compared in a repeated measures ANOVA.

As predicted, results showed a significant difference between both correlation types, thus indicating an anchoring effect. Pre-service teachers' estimations of students' performance were significantly more strongly correlated with own perceived mental load than with students' actual performance, $F(1,33)=169.45, \mathrm{p}<.01$. A very similar pattern was found for in-service teachers' correlations, $F(1,12)=35.64, \mathrm{p}<.01$ (correlation coefficients for both expert groups are depicted in Figure 3). It can be concluded that both expert groups used their own perceived mental load as an anchor to estimate the difficulty that the tasks would impose on the students.


Figure 3: The anchoring effect in both expert groups

## Expertise Reversal Hypothesis

As predicted by the expertise reversal hypothesis, both expert groups' overestimations were moderated by the instructional design of the tasks. Pre-service teachers’ overestimation of student performance was significantly larger for non-optimized ( $M=15.88 \%, S D=14.74 \%$ ) than optimized tasks ( $M=7.15 \%, S D=17.81 \%$ ), $t(33)=4.21 ; p$ < .01. Also, in-service teachers' overestimation was significantly larger for non-optimized ( $M=5.63 \%$, $S D=$ 16.82\%) than optimized items ( $M=-.62 \%, S D=16.77 \%$ ), $t$ (12) $=2.54 ; p<.05$ (see Figure 2 for mean scores). Both expert groups seem to have failed to take into account the benefits of didactic optimization of the learning material for novice learners. As was already described in the "Overestimation Hypothesis" section, differences in overestimation scores did not to reach statistical significance for both expert groups.

## Discussion

In the present study, we investigated whether pre-service and in-service mathematics teachers are subject to an expert blind spot, when judging the difficulty of tasks for students. The tasks were designed in accordance with didactic principles derived from CLT, which have differential effects on the learning outcome of experts and novices. Whereas novice learners experience a relief from extraneous mental load when being presented optimized learning material, thus having more working memory capacity available for schema acquisition and therefore performing better on those tasks, the opposite is true for expert learners. These learners, being presented with optimized tasks, experience increased extraneous mental load (due to a redundancy effect) and judge those tasks not only as being more difficult to work on for themselves, but also as being more difficult to solve for novice learners. The reason for this misjudgement lies in an anchoring effect, as experts generally use their own knowledge base and estimations as ground for judging the difficulties that other persons (in this case: novices) may have. To test these assumptions, experts' mental load ratings while working on mathematical tasks and their estimations of novice performance were compared to real solution frequencies obtained from novices.

Results indicate an egocentric bias, as the experts' general estimations for student performance were highly correlated with their own experienced mental load. Especially, the overestimation of students' task performance was significantly larger for non-optimized than optimized items,
indicating an expert blind spot that can be interpreted in terms of an expertise reversal effect. Experts failed to adequately take into account the beneficial or detrimental effects of didactical variation in task design. Rather, they judged both non-optimized and optimized mathematical tasks as being equally difficult for students, which in fact was not the case in our student sample.

However, only pre-service teachers’ general overestimation of student performance was significant, whereas in-service teachers' overestimation failed to reach statistical significance. Relating to Nickerson's (1999) anchoring and adjustment model, in-service teachers seem to have a more accurate mental model of students' knowledge than pre-service teachers do. Teaching experience seems to have had a debiasing effect on an egocentric bias, thus resulting in better judgements of student performance.

Nevertheless, there are certain limitations to the present study. The first one concerns the yet small sample size of inservice teachers compared to pre-service teachers. The results obtained so far should be further consolidated by equalizing sample sizes for both expert groups, thus allowing for a better comparability and generalizability. This would allow for a detailed analysis of the variability in teaching experience between in-service teachers and its effects on the estimations of student performance. Also, though in-service teachers showed no different estimation pattern for both item types than pre-service teachers do, the overall level of overestimation was a different one. With a bigger sample size, this issue could be further investigated and possible influencing variables could be detected.

Another limitation concerns the actual level of expertise in both teacher groups. As presented in the results section, pre-service and in-service teachers solved significantly more tasks correctly that students did. This allows the conclusion that both teacher groups have more specialized knowledge as compared to students and can indeed be called experts. Also, it is not necessarily surprising that pre-service teachers, not yet having gained teaching experience and being presented with learning tasks obtained from school books, do not solve the mathematical tasks in large part. However, it remains unclear why in-service teachers with a high level of specialized knowledge as well as teaching experience only show similar performance rates on the mathematical problems instead of solving almost all of them correctly.

Finally, the present study does not allow for a detailed insight into participants' estimation processes. After having rated each mathematical task, participants had the opportunity to answer an open-format question and give additional information on what they thought made each tasks difficult or easy. This possibility was barely used, thus not allowing for any further insights into participants' cognitive processes while judging the difficulty of the tasks for students.

Future research will address the just mentioned issues and explore ways in which teachers' ability to see the difficulty
of tasks from a student's perspective can be improved and be emphasized already in teacher education. To the authors' knowledge, no research so far has examined experts estimations of novices' performance using instructional design principles derived from CLT. Subsequent studies with different participant groups shall shed more light on anchoring and adjustment processes in experts. Experts’ cognitive processes while solving the mathematical tasks shall be further investigated. Also, longitudinal designs could be conducted in order to analyse effects of intervention programs on teachers' perception of learning material.

## References

Helmke, A., Hosenfeld, I., \& Schrader, F.-W. (2004). Vergleichsarbeiten als Instrument zur Verbesserung der Diagnosekompetenz von Lehrkräften. In R. Arnold \& C. Griese (Eds.), Schulleitung und Schulentwicklung (pp. 119-144). Hohengehren: Schneider.
Herppich, S., Wittwer, J., Nückles, M., \& Renkl, A. (2010). Do tutors' content knowledge and beliefs about learning influence their assessment of tutees' understanding? In S. Ohlsson \& R. Catrambone (Eds.), Proceedings of the 32th Annual Conference of the Cognitive Science Society (pp. 314-319). New York: Erlbaum.
Hinds, P. (1999). The curse of expertise: The effects of expertise and debiasing methods on predictions of novice performance. Journal of Experimental Psychology: Applied, 5, 205-221.
Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. Educational Psychology Review, 19, 509-539.
Kalyuga, S., Chandler, P., \& Sweller, J. (2001). Learner experience and efficiency of instructional guidance. Educational Psychology, 21(1), 5-23.
Lentz, L., \& De Jong, M. (2006). The curse of expertise: Exploring the problem of anticipating readers’ needs. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?rep=rep1\& type=pdf\&doi=10.1.1.136.1732
Nathan, M., \& Koedinger, K. (2000). Teachers' and researchers' beliefs about the development of algebraic reasoning. Journal for Research in Mathematics Education, 31(2), 168-190.
Nathan, M., \& Petrosino, A. (2003). Expert blind spot among preservice teachers. American Educational Research Journal, 40(4), 905-928.
Nickerson, R. S. (1999). How we know - and sometimes misjudge - what others know: Imputing one's own knowledge to others. Psychological Bulletin, 125, 737759.

Paas, F., \& van Merrienboer, J. (1994). Variability of worked examples and transfer of geometrical problemsolving skills: A cognitive-load approach. Journal of Educational Psychology, 86(1), 122-133.

Paas, F., Renkl, A., \& Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. Educational Psychologist, 38(1), 1-4.
Renkl, A. (2005). The worked-out-example principle in multimedia learning. In R. E. Mayer (Ed.), Cambridge Handbook of Multimedia Learning (pp. 229-246). Cambridge: University Press.
Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. Learning and Instruction, 4, 295-312.
Sweller, J. (2005). Implications of cognitive load theory for multimedia learning. In R. E. Mayer (Ed.), Cambridge handbook of multimedia learning (pp. 19-30). New York: Cambridge University Press.
Sweller, J., van Merrienboer, J., \& Paas, F. (1998). Cognitive architecture and instructional design. Educational Psychology Review, 10 (3), 251-196.
Tversky, A., \& Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. Science, 185, 11241131.

# The Recognizability of Authenticity 

Madeleine Henderson (madeleine.henderson@gmail.com) and Liane Gabora (liane.gabora@ubc.ca)<br>Department of Psychology, University of British Columbia<br>Okanagan Campus, 3333 University Way, Kelowna BC, V1V 1V7, CANADA


#### Abstract

The goals of this research were to (1) determine if there is agreement both amongst viewers, and between viewers and the performer, about the extent to which performances are authentic, and (2) ascertain whether or not performers and/or viewers can distinguish between authenticity and skill. An authentic performance is one that is natural or genuine, while an inauthentic performance feels faked, forced, or imitative. Study participants were asked to rate the authenticity and skill level of a series of videotaped performances by dancers and stand-up comedians. Performers also rated their own performances. Authenticity ratings amongst viewers were significantly positively correlated. Ratings between viewers and performers were not significant but all positive. A higher correlation between ratings of both authenticity and skill of performances for viewers than for performers suggests that viewers make less of a distinction between authenticity and skill than performers. The relationship between authenticity and creativity is discussed.


Keywords: authenticity; comedy; creativity; dance; expertise; fake; genuine; individual differences; performance; skill.

## Introduction

With increasing frequency there are calls for research aimed at a synthetic account of how the components of a cognitive system function in synchrony to generate behavior in everyday situations. We propose that the construct of authenticity has an important role to play in such an account. Authenticity refers to the ability to be genuine, to accurately reflect who one really is, and be true to the situation one is in. Writers speak of discovering one's own authentic voice. In theatre research the term 'authentic' is used in discussion of the extent to which a performer gives a performance something personal that goes beyond the script (Lavy, 2005). In the dance community the term 'authentic movement' refers to the strengthening of identity through uninhibited movement of they body in a social context (Goldhahn, 2009). In an area at the intersection of anthropology and tourism research, the term 'authenticity' is used to refer to the extent to which current creative works in a given genre, such as Native American or First Nations art or dance, employ the same tools, techniques, styles, and so forth, as were traditionally used (Daniel, 1996; Maruyama et al., 2008). Thus an authentic performance is one that seems natural, or true to an underlying essence, while an inauthentic performance feels faked, forced, or imitative.

Authenticity is important for many reasons. It feels highly gratifying to both the performer and the observer. It is relevant to many domains of life, including the generation of artistic works and performance (e.g., art, acting, music, and dance), non-artistic performances (e.g., teaching and
newscasts), and everyday social interactions with friends and family. However, despite that performers, viewers, and the general public regularly voice opinions about authenticity, and despite that in the scholarly community authenticity is assumed to be a genuine construct about which viewers and performers are in agreement (e.g., Goldhahn, 2009; Kogan, 2002; Lavy, 2005; McClary, 2007; Nemiro, 1997; Sawyer, 1992; Warja, 1994), we were unable to locate any empirical research that supports this assumption. Indeed we found no empirical evidence for consensus as to which performances are authentic and which are not, either amongst members of an audience, or between a performer and an audience.

## Authenticity and Skill

Audiences without artistic expertise emphasize skill over originality in assessments of visual art, while the reverse is true for audiences with expertise (Hekkert \& van Wieringen, 1990a, 1990b, 1996). This suggests that originality-which might be related to authenticity-can be confused with skill. However, there is evidence that skill and authenticity are distinct constructs (Kogan, 2002). While being skilled in a domain may facilitate authentic performance, it does not guarantee it, nor is it a necessary prerequisite. For example, a dancer may have perfected her craft, and be technically skilled, permitting a wide range of means for selfexpression, but not immerse herself in the work, or simply imitate the instructor, yielding a performance void of authentic style. Conversely, a performer lacking in technical skill may exude personality or detectable "creative release", yielding a performance that comes across as authentic. In short it remains an open question whether viewers confuse a skilled performance with an authentic one.

## Goals of Current Study

Although it would be difficult to pinpoint the potentially myriad factors that contribute to authenticity or a lack of it, it is possible to make headway toward determining whether authenticity is a genuine construct by assessing the extent to which viewers of a performance, and performers themselves, agree in their assessments of authenticity. Thus a first goal of this study was to determine if there is a correlation amongst viewers' assessments of the authenticity of a given performance. A second, related goal was to determine whether there is a correlation between viewers' assessment of the authenticity of a performance and the performer's self-assessment of the authenticity of that performance. We hypothesized that an audience can detect an authentic or inauthentic performance, and that
performances that feel authentic to a performer come across as authentic to an audience, and vice versa.

A third goal was to determine whether authenticity and skill are distinct constructs in the eyes of the performer and/or viewers. Since it is possible to be skilled but perform in an inauthentic manner, or to perform authentically but not be skilled, we hypothesized that both performers and viewers could distinguish between the two constructs.

A final goal was to determine what factors facilitate authentic performance. Previous research on this is inconclusive (e.g., McClary, 2007; Nemiro, 1997; Rhodes, 1999; Sawyer, 1992; Warja, 1994). By asking performers open-ended questions about authenticity we hoped to shed light on this seemingly elusive phenomenon that would pave the way for further studies of the relationship between authenticity and the therapeutic value of creative endeavors.

## The Study

## Participants

Three trained dance performers were recruited from a local dance studio. Dancer A was 25, Dancer B was 29, and Dancer C was 23. Each dancer had between 10 to 12 years of dance experience, and took part in dance at least once per week. Each trained dancer was paid $\$ 30$ for their participation in the study. They met with the experimenter for a video recording session of three hours duration. A four year old child with no formal dance training was also recruited as a dance performer.

Three comedians were also recruited for the study. The first was a 36-year old experienced stand-up comedian with eight years of stand-up comedy experience. He was located from a local directory. The second was a 24 -year old amateur stand-up comedian who had just started doing stand-up comedy one month prior to the study. She was recruited through a psychology of humour class at The University of British Columbia. The third was a 23 -year-old 'social comedian' known to the experimenter. He had no stand-up comedy credentials, but had years of experience being the center of attention for his humour in social situations. None of the comedians were compensated for their participation.

158 University of British Columbia undergraduates were viewers of the performances. 45 were recruited through the SONA system, which enables participation in university research in exchange for credit in a psychology class. 50 students were recruited through psychology of creativity and psychology of humor classes. They were not given incentives to participate. The remaining participants were recruited through online university class message boards, and were also not given incentives to participate. The only exclusion criterion was severe visual impairment, such as blindness. Females accounted for approximately $58.2 \%$ of the sample ( $\mathrm{n}=92$ ) and males accounted for the remaining $41.8 \%(\mathrm{n}=66)$. Most $(83.1 \%, \mathrm{n}=128)$ were between the ages of 17-25, and in a Bachelor of Arts $(64 \%, n=96)$ program.

## Procedure

The experienced dancers were filmed practicing original choreographed dance routines in their dance studio. They were told that the study was about the psychology of movement. They met at the dance studio one hour prior to filming to learn two different modern dance routines. Both routines were choreographed to music and lasted one to two minutes in duration. For one, the music was a quick, highenergy piece, while for the other it was slow and sombre.

After the hour-long practice, each dancer individually performed the fast dance five times. After all dancers had finished, they individually performed the slow dance five times in the same order as the first. Each dance performance was videotaped using a high-definition video camera.

Once the first dancer had completed all her dances, she was directed to a laptop where footage of her routines was uploaded. She was debriefed about the specific reasons for conducting the study, and given a definition of authentic performance. She was then asked to watch her own ten performances in the order in which they were performed, and given a questionnaire with the following items based on each performance:

> Please rate how authentic you felt this performance was based on how you felt you were coming across or how you felt inside during the performance" (Not authentic at all / Somewhat authentic / Neutral or Don't know / Quite authentic / Very authentic)

How would you rate your performance in regards to technical skill?" (Very poor / Poor / Okay / Good / Very good)
This was repeated for the other dancers. All dancers were also asked to fill out an open-ended portion of the questionnaire, which asked the following questions concerning factors that facilitate or hinder authenticity:
(1) Do you feel as though authenticity and technical skill are the same thing or different concepts? Please explain.
(2) Are there particular situations or environments in which you are able to produce your most authentic performance? If so, please tell us about it.
(3) Do you believe that people become more able to find their authentic style with experience?
(4) Was there a known time in your career where you felt that you had made a transition to being more able to express yourself? If yes, describe that transition.

The child dancer was filmed using a high-definition video camera in her home. Filming began when she spontaneously began dancing to upbeat dance music. The camera was not hidden from view and the child was aware she was being filmed. The footage was divided into two video clips of two minutes each. Due to her age, she was not asked to assess how authentic or skilled her performances were, nor to respond to the open-ended questions.

The experienced stand-up comedian was asked to submit between five and ten previously taped performances that he had acquired over his career. We requested that each video clip be under two minutes duration, and that together they
portray a range of authenticity. He was also asked to rate the authenticity and skill of each performance using the same five-point Likert scale administered to the dancers. He submitted six videotaped live performances and his ratings for each.

Video footage was collected from the amateur stand-up comedian without an audience (except for the experimenter). She was asked a series of questions that would potentially promote humourous responses, such as "what was your most embarrassing moment?" or "what is the strangest thing you have seen on campus?" She was also asked to run through segments of her stand-up routine which were narrative in nature, tell funny jokes (not necessarily her own), or make up funny stories and deliver them as though they were real. After approximately 30 minutes of recording, she was asked to look through the footage on a laptop and rate the authenticity and skill of each joke/story segment on the five-point Likert scale. Video footage of the social comedian was collected in the same manner as with the amateur stand-up comedian.

The video clips were loaded onto an online questionnaire using www.surveymonkey.com with the exception of those from Dancer C. Her performances were omitted due to extreme homogeneity in her responses to the Likert items. (Since her performances did not exhibit variation in selfrated authenticity, they were not useful for this study.) Her responses to the qualitative questions were retained.

Viewers were given the following definition of authenticity:

> Authenticity in the performing arts commonly refers to the ability of a performer to perform in such a way that they are able to remain true to who they really are or to the character they are trying to play. Conversely, a performer who is not performing authentically is merely giving a performance that seems artificial or imitated.

Viewers were asked whether they felt they understood the construct of authenticity, and if they did not, further discussion ensued until it was clear to them what authenticity refers to. After each video clip, viewers were required to rate it on the same five-point Likert scales that the performers used. In order to minimize potential order effects, the ordering of the performances was randomly altered every time ten students had completed the survey. Video clips belonging to the same performer were kept together, but the order of the performers and the order of the clips belonging to each performer were randomized.

The students who were recruited from the psychology of creativity and psychology of humor classes were shown the video clips on a projector screen, and they received a paper version of the questionnaire. They were given 30 seconds to rate each performance before the next one commenced. The type of psychology class and the week in which the study was conducted determined the types of performances that were shown. For example, the psychology of creativity class was approached earlier in the study, and was shown the clips of the dancers and the experienced stand-up comedian because these performances were the only ones available at
that time. The psychology of humour class saw only the comedians' performances because dance performances were not relevant to the class content.

## Analysis and Results

The means and standard deviations for the authenticity ratings of the performances are given in Table 1. The highest authenticity ratings were for the dancing child ( $M=$ $4.52)$ and the social comedian $(M=4.05)$.

Table 1: Mean authenticity ratings by viewers and performers for all performances.

| Performer | Viewer ratings <br> $M($ SD $)$ | Performer <br> Self-ratings |
| :--- | :--- | :--- |
| Experienced stand- <br> up comedian | $3.68(1.07)$ | 3.33 |
| Amateur stand-up <br> comedian | $3.19(1.31)$ | 2.50 |
| Social comedian | $4.05(1.08)$ | 4.33 |
| Dancer A | $3.18(1.12)$ | 3.80 |
| Dancer B | $3.27(1.18)$ | 3.30 |
| Dancing child | $4.25(1.01)$ | $\mathrm{N} / \mathrm{A}$ |

## Recognizability of Authenticity

Between-Viewer Ratings To determine whether the viewers agreed as to which performances seemed authentic, the intraclass correlation coefficient $\left(R_{i}\right)$ was calculated. The $R_{i}$ statistic is more appropriate for this study than the widely-used Pearson product moment correlation because the latter ignores the extent to which independent raters agree on any single rating (Cicchetti, 1991). The $R_{i}$ coefficients for the extent of agreement amongst viewers about the authenticity of the performances of each performer are presented in Table 2. All values are statistically significant at the .05 level with the exception of those for Dancer A, and they are all statistically significant at the .01 level with the exception of those for Dancer A and the dancing child.

Table 2: Agreement of authenticity amongst viewers $\left(R_{i}\right)$, and between viewers and performer $(r)$.

| Performer | $\boldsymbol{R}_{i}$ | $r$ |
| :--- | :--- | :--- |
| Experienced stand-up comedian | $.965^{* *}$ | .712 |
| Amateur stand-up comedian | $.890^{* *}$ | .120 |
| Social comedian' | $.858^{* *}$ | .609 |
| Dancer A | .340 | .061 |
| Dancer B | $.879^{* *}$ | .520 |
| Dancing child | $.822^{*}$ | $\mathrm{~N} / \mathrm{A}$ |

${ }^{*} p<.05 ; * * p<.01$

Agreement Between Viewers and Performers. To determine if there was agreement between viewer and performer ratings of authenticity, we merged the multiple viewer ratings to obtain the average composite rating for
each performer. A Pearson product moment correlation was conducted to see if the composite rating is in agreement with the performer's ratings of authenticity. These values are also presented in Table 2.

The highest agreement was between the viewers and the experienced stand-up comedian, followed by the social comedian, Dancer B, the amateur stand-up comedian, and Dancer A. There was considerable variation amongst the performers with respect to the degree to which their assessments of the authenticity of their performances were correlated with the viewers' assessments. While none of the correlations were statistically significant, all were positive. Moreover, significance was based on a small number of performances for each performer. The lack of power from the small $n$ 's indicates that the significance tests were highly prone to type II errors (failure to find a significant difference when one exists). In such situations it may be prudent to focus on the magnitude of the observed effect or relationship instead of the significance tests (Gliner, Leech, \& Morgan, 2002; Serline \& Lapsey, 1993; Wilkinson \& the APA Task Force on Statistical Inference, 1999).

Qualitative Results. To better understand what factors facilitate the expression of authentic creative style we conducted a content analysis of the open-ended questions. There were recurring responses as well as individual differences. Responses to the question, "Is the development of an authentic voice related to experience?", suggest that experience facilitates the development of authentic style, but that this happens differently for different performers. Compare the responses of two dancers:

Experience is what helps one explore his or herself to
discover what authenticity means for them. discover what authenticity means for them.

I have found that by taking a number of different dance styles with a number of different instructors that I have developed (and continue to develop) my own personal style. The more experience that I've gained the more comfortable I've become with myself and my movement and the more ideas that I can "pull out of my hat".

The performers put forward several factors that interfere with the authenticity of their performances: excessive focus on technical perfection, performing in front of large audiences, or audiences that include friends or acquaintances, performing while injured or tired, performing content that is unfamiliar or that does not "lean towards [one's] natural expression", and working with a choreographer that has a different style. The performers also put forward many factors that enhance with the authenticity of their performances. The most commonly cited factor was feeling safe from judgment. Other factors were being in a performing mood, feeling inspired, and teaching choreography. Interestingly, while some performers claimed that having an audience increases the authenticity of their performance, others claimed that it has the opposite effect.

## Distinguishing Authenticity from Skill

Quantitative Results. There was a modest but significant Pearson correlation between mean ratings of authenticity and mean ratings of skill as assessed by viewers. The Pearson correlation for the performers' mean ratings of authenticity and skill was lower but significant. These results are presented in Table 3. Thus although authenticity and skill appear to be related for both performers and viewers, performers made a stronger distinction between them than viewers.

Table 3: Pearson correlation between mean ratings of authenticity and mean ratings of skill $(r)$.

|  | $r$ |
| :--- | :--- |
| Viewers | .641 |
| Performers | .547 |

${ }^{*} p<.001$
Qualitative Results. The qualitative data indicates that the performers unanimously view authenticity and skill as distinct concepts. For example:

Technical skill - is where you learn how to move and hold yourself properly for the desired discipline. Authenticity - is the feeling and expression that you can add to your technical skill to create the "entire picture".

Anybody can master technical skills with enough practice but if you don't have charisma as an artist - or better yet as a stand-up comedian, people won't think you're very funny.

Responses suggested that skill can facilitate authenticity:
Technical skill opened the door of possibilities for me to further express my emotions.

Once I know how to do a proper "plie" and the barre, it is much easier for me to add some expression or feeling because I'm not thinking nearly as much about how the plie should be done and can focus on making it look "pretty."
However, one performer's answers suggested that acquiring skill may interfere with authenticity:

> Sometimes a lot of technical training can make it difficult for the dancer to separate their own authentic style from the teachers. It all comes down to how they have been trained, if their teacher demands uniformity and discourages personal exploration it will be harder. If they have a good teacher who knows how to pull out creativity and massage it, then the experience will benefit their discovery of an authentic style.

The performers claimed that skill can facilitate authentic performance by freeing them from concern with technical details so they could be more fully immersed in the creative process. A preoccupation with skill, however, can prevent a performer from reaching a deeper connection with the task. These qualitative responses, in conjunction with the quantitative results, support the hypothesis that authenticity and skill are related, yet distinct concepts.

## Discussion

The results of this study shed light on the seemingly elusive construct of authenticity. The agreement amongst viewers as to which performances were authentic, a result obtained across a variety of performance types and situations, suggests that authenticity is indeed a real concept as opposed to existing in the eye of the beholder.

The variability in the correlations between authenticity ratings for viewers and performers indicate that when a performance feels authentic to a performer it may or may not come across that way to others. This was addressed by one of the dancers, who noted:

> Some people have very 'quiet' personalities so when they are authentically displaying anger they might be so quiet about it [that] an audience would not see it. Those dancers might be rated 'less authentic' because they are less obvious.

This comment suggests to us that the reason for the low agreements between the amateur stand-up comedian and Dancer A and the viewers is that outward manifestations of their personalities may be subtle for the viewers to detect them. Analyzing how the personality of an artist interacts with the recognizability of authenticity in performance is an interesting direction for future research.

Although the variety of performance types and settings contributed to the generalizability and ecological validity of the findings, caution must taken in drawing conclusions that involve comparisons across performers or performance settings, because differences such as 'in a studio' versus 'at home' could be potential confounds. With this warning, we offer some speculative discussion of between-performer differences. There are several possible explanations for the high agreement amongst viewer authenticity ratings of the experienced comedian. First, over time he may have solidified a strong authentic voice that is readily detectable when present, making an inauthentic performance stand out in contrast. Second, his performances were the only ones that were filmed before he knew he would be rated. Some research indicates that the pressure of knowing one is going to be evaluated can inhibit creative expression (Nemiro, 1997; Rhodes, 1999), so it is possible that the rest of the performers who knew they were going to be evaluated gave performances that were more uniform with respect to authenticity, giving viewers less opportunity to detect differences amongst performances. It would be interesting to investigate whether expertise can entail becoming skilled at faking authenticity, i.e., whether there exist performers for whom expertise is inversely correlated with agreement between performer and viewer authenticity ratings.

The dancing child's high authenticity ratings may reflect in part the stereotype that children are authentic in whatever they do. However, the fact that the social comedian's performance was also rated as highly authentic suggests that these high ratings reflect instead the spontaneity of their performances. While the other performers' performances (though to a lesser extent the amateur comedian) were choreographed or scripted, those of the child and social
comedian were not. This interpretation is consistent with findings that freedom facilitates authenticity (McClary, 2007; Nemiro, 1997; Sawyer, 1992; Rhodes, 1990, Warja, 1994). This explanation is further reinforced by the fact that the experienced stand-up comedian's highest rated performance for authenticity was the only one in which he was forced to improvise (due to verbal feedback from the audience). This points to a weakness of the study. Since most performers knew they would be judged, they may have been less able to release inhibitions and be authentic. Another weakness is that because dancers were confined to rehearsed, choreographed routines, differences between performances of the same dance may have been too subtle for viewers to detect, thus limiting the range of authenticity scores. Future studies will focus on spontaneously improvised performances, which allow authenticity to be expressed through content as well as delivery.

Another direction for future research is to investigate whether there is a difference in the capacity to detect authenticity in live versus videotaped performances. Previous research indicating that there is a constant interaction between a performer and a live audience (Arnold, 1991; Bindeman, 1998; Nemiro, 1997) suggests that viewers may be better able to detect cues or indications of authenticity from a live performance than from a performance on a television or computer screen.

This study of authenticity arose in the context of an interest in what factors affect how the various components of a cognitive system come together to produce overt thought and behavior. It seemed reasonable that an authentic response is one that genuinely reflects the state of one's associative network, including not just one's internal model of the world (including self-understanding) but one's way of seeing and being. We refer to this dynamical structure as a worldview (Gabora, 1999, 2000, 2001). We speculated that: (1) authenticity entails being entirely present and thereby available to detect unresolved questions or issues, and open to change, (2) a lack of authenticity may indicate that elements of one's worldview are repressed or misrepresented because this interferes with the capacity to detect unresolved questions or issues, and be open to change, (3) authentic performance facilitates the process by which one's worldview self-organized into more stable state, while unauthentic performance does not. However, before we could address these issues it was necessary to address the more fundamental question of whether authenticity is a real construct.

The findings reported here, and in particular the key finding that authenticity is recognizable, opens up many questions and perspectives. It led us to speculate that perhaps one is being creative even when not engaged in an overtly creative activity if one responds to a new situation or emotion in a way that authentically reflects how it affects ones' worldview. This is consistent with the honing theory of creativity, according to which creative behavior arises because one's worldview tends to self-organize in response to perturbation to achieve a more stable state, regain
equilibrium, or resolve dissonance (Gabora, 2005; Gabora, Ranjan \& O'Connor, 2012). The creative product or performance is viewed as an external reflection of this internal transformation. This conceptualization of creativity is consistent with the anecdotal evidence obtained in the qualitative portion of this study that authentic performance can be therapeutic. It is also consistent with the notion that personal performance style is the result of inner transformations (Kogan, 2002) and the view that creative performance involves interaction and tension between the creator's conscious and subconscious which impact the creator's identity (Sawyer, 1992).

## Acknowledgments

We are grateful for funding to the second author from the Natural Sciences and Engineering Research Council of Canada and the Concerted Research Program of the Flemish Government of Belgium. We thank the performers, the Creator's Arts Centre, and Celine Dahmen-Formosa for their assistance, and Gordon Rudolph for comments.

## References

Chalnick, A., \& Billman, D. (1988). Unsupervised learning of correlational structure. Proceedings of the Tenth Annual Conference of the Cognitive Science Society (pp. 510-516). Hillsdale, NJ: Lawrence Erlbaum Associates.
Arnold, N. (1991). The manipulation of the audience by the director and actor. In G. D. Wilson (Ed.), Psychology and Performing Arts (pp. 75-81). Amsterdam: Swets \& Zeitlinger.
Bindeman, S. (1998). Echoes of silence: A phenomenological study of the creative process. Creativity Research Journal, 11(1), 69-77.
Cicchetti, D. V. (1991). The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. Behavioral and Brain Sciences, 14, 119186.

Daniel, Y. P. (1996). Tourism dance performances: Authenticity and creativity. Annals of Tourism Research, 23(4), 780-797.
Dudek, S. Z., \& Marchand, P. (1983). Artistic styles and personality in creative painters. Journal of Personality Assessment, 47(2), 139-142.
Gabora, L. (1999). Weaving, bending, patching, mending the fabric of reality: A cognitive science perspective on worldview inconsistency. Foundations of Science, 3(2), 395-428.
Gabora, L. (2000). Conceptual closure. How memories are woven into an interconnected worldview. Annals of the New York Academy of Sciences, 901, 42-53.
Gabora, L. (2005). Creative thought as a non-Darwinian evolutionary process. Journal of Creative Behavior, 39(4), 65-87.
Gabora, L. (2001). Cognitive mechanisms underlying the origin and evolution of culture. Doctoral Thesis, Free University of Brussels, Belgium.

Gabora, L., O’Connor, B., \& Ranjan, A. (2012). The recognizability of individual creative styles within and across domains. Psychology of Aesthetics, Creativity, and the Arts, 6(4), 351-360.
Getzels, J. W., \& Csikszentmihalyo, M. (1976). The creative vision: A longitudinal study of problem finding in art. New York: Wiley-Interscience.
Gliner, J. A., Leech, N. L., \& Morgan, G. A. (2002). Problems with null hypothesis significance testing (NHST): What do the textbooks say? Journal of Experimental Education, 71(1), 83-92.
Goldhahn, E. (2009). Is authentic a meaningful name for the practice of Authentic Movement? American Journal of Dance Therapy, 31, 53-63.
Hekkert, P., \& van Wieringen, P. C. W. (1990a). Complexity and prototypicality as determinants of the appraisal of cubist paintings. British Journal of Psychology, 81, 483-495.
Hekkert, P., \& van Wieringen, P. C. W. (1990b). Experts' appraisal of cubist paintings. In L. Halasz (Ed.), Proceedings of the XIth International Congress on Empirical Aesthetics (pp. 118-121). Budapest: Institute for Psychology of the Hungarian Academy of Sciences.
Hekkert, P., \& van Wieringen, P. C. W. (1998). Assessment of aesthetic quality of artworks by expert observers: An empirical investigation of group decisions. Poetics, 25(5), 281-292.
Kogan, N. (2002). Careers in the performing arts: A psychological perspective. Creativity Research Journal, 14(1), 1-16.
Lavy, J. (2005). Theoretial foundations of Grotowski's total act via negativa and conjunctio oppositorum. The Journal of Religion and Theatre, 4(2), 175-188.
McClary, R. (2007). Healting the psyche through music, myth, and ritual. Psychology of Aesthetics, Creativity, and the Arts, 1, 155-159.
Nemiro, J. (1997). Interpretive artists: A qualitative exploration of the creative process of actors. Creativity Research Journal, 10, 229-240.
Rhodes, C. (1990). Growth from deficiency creativity to being creativity. Creativity Research Journal, 3, 287-299.
Sawyer, K. (1992). Improvisational creativity: An analysis of jazz performance. Creativity Research Journal, 5, 253263.

Serline, R. C., \& Lapsey, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. In G. Keren \& C. Lewis (Eds.), A handbook for data analysis in the behavioral sciences: Methodological issues (pp. 199-228). Hillsdale, DJ: Erlbaum.
Warja, M. (1994). Sounds of music through the spiraling path of individuation: A Jungian approach to music psychotherapy. Music Therapy Perspectives, 12, 75-83.
Wilkinson, L., \& the APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, 54, 594-604.

# Benefits for Processes Cause Decrements in Outcomes: Training Improves Tutors' Interactivity at the Expense of Assessment Accuracy 

Stephanie Herppich (stephanie.herppich@sowi.uni-goettingen.de)<br>University of Göttingen, Educational Institute<br>Waldweg 26, 37073 Göttingen, Germany

Jörg Wittwer (joerg.wittwer@sowi.uni-goettingen.de)
University of Göttingen, Educational Institute
Waldweg 26, 37073 Göttingen, Germany

Matthias Nückles (matthias.nueckles@ezw.uni-freiburg.de)<br>University of Freiburg, Department of Educational Science, Instructional and School Research Rempartstrasse 11, 79098 Freiburg, Germany

Alexander Renkl (renkl@psychologie.uni-freiburg.de)<br>University of Freiburg, Department of Psychology, Developmental and Educational Psychology Engelbergerstrasse 41, 79085 Freiburg, Germany


#### Abstract

Tutoring gives tutors the opportunity to engage in interactive strategies that help them to assess a tutee's understanding. However, tutors without teaching experience often do not engage in interactive strategies and, thus, have difficulty with accurately assessing a tutee's understanding. We conducted an experiment with 39 tutor-tutee dyads to test whether tutors who received training in interactive strategies would become more interactive and more accurate in assessing a tutee's understanding. Results showed that trained tutors provided a more interactive style of tutoring than untrained tutors. However, due to being more interactive, trained tutors produced less accurate assessments than untrained tutors. This suggests that changing the style of tutoring to implement interactive strategies puts a high burden on a tutor's cognitive capacity. Hence, there is obviously little cognitive capacity left that could be used to assess a tutee's understanding. Training methods that automate strategy use might enhance a tutor's assessment accuracy.


Keywords: one-on-one human tutoring; training; tutoring interactions; assessment accuracy

## Introduction

In one-on-one tutoring, tutors have the possibility to engage in interactive tutoring strategies such as asking questions or providing hints. When a tutee responds to a tutor's interactive tutoring strategies, for example, by answering a question, a tutor can learn what a tutee does and does not know (Chi, 2009; Hmelo-Silver \& Barrows, 2006). Thus, in the course of tutoring, a tutor has the opportunity to collect a multitude of information that can be used to summatively assess a tutee's understanding after tutoring session. This summative assessment may also help a tutor to prepare the next tutoring session by choosing material that is suited to a tutee's individual level of understanding (e.g., Chi, Jeong, \& Siler, 2004; Kalyuga, 2007; cf. also the discussion of the
concept of interim assessments for the school context by Perie, Marion, \& Gong, 2009).

However, research has shown that inexperienced tutors, that is, tutors who are not trained in teaching (Chi et al., 2001; Graesser, D'Mello, \& Cade, 2011), often do not engage in interactive tutoring strategies. Instead, they frequently dominate tutoring by providing lengthy explanations (e.g., Chi et al., 2001; Cromley \& Azevedo, 2005). In addition, inexperienced tutors regularly fail to assess a tutee's understanding accurately (Chi et al., 2004; Herppich et al., 2013b).

Against this background, we conducted an experimental study to test whether inexperienced tutors who received training in interactive tutoring strategies would be able to implement an interactive style of tutoring. We were interested in whether a more interactive style of tutoring would benefit a tutor's assessment of a tutee's understanding after tutoring.

## Tutoring Strategies of Experienced and Inexperienced Tutors and Their Influence on Assessment

In contrast to inexperienced tutors, experienced tutors are trained or experienced in teaching (cf. Cromley \& Azevedo, 2005; D'Mello et al., 2010; McArthur, Stasz, \& Zmuidzinas, 1990). Research shows that experienced tutors tend to provide a different style of tutoring than do inexperienced tutors. More specifically, experienced tutors more often engage in interactive tutoring strategies than inexperienced tutors. For example, they frequently scaffold a tutee by providing hints or asking questions (Cade et al., 2008; Chi, Roy, \& Hausmann, 2008; Cromley \& Azevedo, 2005). Scaffolding is a genuinely interactive tutoring strategy because it elicits constructive responses from a tutee (Hmelo-Silver \& Barrows, 2006). In this vein,

Herppich et al. (2013a, 2013b) found that experienced tutors caused tutees to utter more knowledge deficits, that is, incomplete beliefs, incorrect beliefs, or misconceptions, in the course of tutoring than inexperienced tutors. In addition, experienced tutors were more accurate in assessing a tutee's understanding after tutoring than inexperienced tutors. The results suggest that a tutee's uttered knowledge deficits are diagnostically informative because they indicate what a tutee does not know (cf. Chi, et al., 2004; Cromley \& Azevedo, 2005). Thus, tutors might derive information from these knowledge deficits that can be used to assess a tutee's understanding after tutoring.

## Training Inexperienced Tutors

To test whether training inexperienced tutors in interactive tutoring strategies would improve their style of tutoring, we developed a training method that aimed at prompting inexperienced tutors to abstain from giving lengthy explanations and, instead, to engage in more interactive tutoring strategies such as scaffolding (cf. Chi, et al., 2008). As a result of implementing more interactive tutoring strategies in the course of tutoring, tutors were assumed to more intensively engage in collecting diagnostically relevant information that could be used to assess a tutee's understanding after tutoring.

Based on what is known about effective training methods in the domain of learning strategies (Mandl \& Friedrich, 1992), the development of our training method was guided by several principles. First, training methods should inform about the advantages associated with the strategies targeted in the training. Second, training methods should directly convey knowledge about the strategies that need to be trained. Third, training methods should help to practice the targeted strategies (Klauer, 1988; Mandl \& Friedrich, 1992). Research has shown that training methods that are in accordance with these principles are particularly effective (Dignath, Buettner, \& Langfeldt, 2008; Leutner, Leopold, \& Elzen-Rump, 2007).

By now, little attention has been given to training methods that aim at fostering an interactive tutoring style in the service of improving assessment accuracy. However, existing research on training tutors with the aim of enhancing a tutee's learning has well documented that tutors are often able to spontaneously implement the strategies that are targeted in training. Yet, tutors have difficulty with changing their style of tutoring in the long run (King, Staffieri, \& Adelgais, 1998). Moreover, even though tutors are able to change their tutoring strategies, this might not necessarily increase the effectiveness of tutoring (Chi et al., 2001). In their review on tutoring-based instruction, Graesser et al. (2011) summarized research on tutor training in the following way:
...it is difficult to train tutors to adopt particular strategies. They rely on their normal conversational and pedagogical styles.... it is difficult to force the human tutors to adopt changes in their language and discourse,
particularly those levels that are unconscious and involuntary. (p. 422).

## Hypotheses

In this study, we tested the effectiveness of a training method that aimed at helping tutors to implement a more interactive style of tutoring. We addressed the following hypotheses:

1) Trained tutors engage in more interactive tutoring strategies in the course of tutoring than untrained tutors.
2) Trained tutors are more accurate in assessing a tutee's understanding after tutoring than untrained tutors.
3) The more interactive style of tutoring explains why trained tutors are more accurate than untrained tutors in assessing a tutee's understanding after tutoring.

## Method

## Sample and Design

A total of $N=39$ dyads of tutors and tutees participated in the experiment. The topic of tutoring was the human circulatory system. All tutors were university students majoring in biology with a mean age of 22.38 years ( $S D=$ 2.47). Thirty-five tutors were female and 4 tutors were male. Twenty tutors received training in interactive tutoring strategies (= trained tutors), whereas 19 tutors received no training (= untrained tutors). As indicated by a multiplechoice test, all tutors had sufficient knowledge about the human circulatory system. There was no significant difference in knowledge between trained tutors ( $M=8.45$, $S D=2.26)$ and untrained tutors $(M=8.26, S D=1.78), F(1$, 37) $=0.81, p>.05, \eta^{2}<.01$ (small effect). Moreover, trained (mean rank $=18.88$ ) and untrained tutors (mean rank $=21.18$ ) did not differ in their previous experience in providing tutoring, coded as $1=$ no experience, $2=$ sporadic tutoring, $3=$ regular tutoring, $U=167.50, z=-0.69, p>$ $.05, r=-.11$ (small effect). Tutees were seventh-grade students from the middle track of the German school system (i.e., from Realschulen). Of the tutees, 9 were female and 29 were male; one tutee did not indicate gender.

Tutors were randomly assigned to the two experimental conditions (training vs. no training) and tutees were randomly assigned to tutors. The dependent variables in this experiment were the extent to which a tutor elicited knowledge deficits from a tutee in the course of tutoring and the accuracy with which a tutor assessed a tutee's understanding after tutoring.

## Materials

Textbook Passage (Tutees and Tutors) In the tutoring session, the tutor-tutee dyads engaged in a dialogue based on a passage about the human circulatory system. We adapted this passage from the study by Chi et al. (2001). The passage consisted of 59 sentences and each sentence was printed on a separate sheet of paper. The sentences were presented to the tutor and the tutee in a ring binder.

Concepts Test (Tutees and Tutors) We used a shortened version of a test that was employed by Herppich et al. (2013b). This shortened version consisted of 16 multiplechoice items that assessed a tutee's understanding of concepts about the human circulatory system. For example, it included the following item: What is the task of the heart in the human organism? (1) The heart pumps the blood. (2) The heart cleans and filters the blood. (3) The heart supplies the blood with oxygen. (4) Don't know. The items of the original test were adapted from tests developed by Sungur and Tekkaya (2003) and by Michael et al. (2002) or constructed on the basis of the literature on misconceptions of the human circulatory system (e.g., Pelaez et al., 2005). A correct answer indicated a scientifically correct understanding of the concept. Each of the incorrect answers indicated a specific type of incorrect understanding of the concept. Hence, a tutee could achieve a maximum number of 16 points in the concepts test.

To examine the accuracy with which the tutors assessed a tutee's understanding of the human circulatory system after tutoring the tutors were also administered the test.

Training in Interactive Tutoring Strategies (Trained Tutors) The trained tutors received training in interactive tutoring strategies. The training took about 45 minutes and was presented on a computer screen. The training aimed at helping the trained tutors to adopt interactive tutoring strategies that would enable them to elicit knowledge deficits from a tutee. The training consisted of two building blocks. In the first building block, the trained tutors were informed about the problem that tutors often are not interactive and, thus, cannot accurately assess a tutee's understanding (Brown, Campione, \& Day, 1981). Subsequently, the trained tutors were provided with information about three strategies, namely, (1) abstaining from giving lengthy explanations, (2) intensifying question asking, and (3) increasing scaffolding in response to a tutee's contribution (Cade et al., 2008, Chi et al., 2008; Herppich et al., 2013a). To learn about the three strategies, the trained tutors first read an explanatory text and then watched two videos of fictitious tutoring sessions. The first video presented a tutor who failed to engage in interactive tutoring strategies and, thus, to receive information about a tutee's understanding. The second video, in contrast, presented the same tutor who did engage in interactive tutoring strategies, which helped the tutor to receive information about a tutee's understanding (cf. Renkl, 2005). In the second building block, trained the tutors also watched videos that presented positive and negative examples of tutoring strategies. This time, however, the tutoring strategies were not explained to the trained tutors. Instead, the trained tutors were prompted to self-explain what constituted the difference between the positive and negative examples. More specifically, the trained tutors were asked to provide information about the tutoring strategies that they saw in the videos and about the effects of such tutoring strategies for assessing a tutee's understanding (cf. Renkl,
2005). Finally, the trained tutors were required to indicate what they would do in order to change the tutoring strategies that they saw in a negative example. This was done to actively stimulate the application of the to-belearned strategies (cf. Klauer, 1988).

Introductory Text (Untrained Tutors) Instead of receiving training in interactive tutoring strategies, the untrained tutors read a short text. The text provided information about the effectiveness of tutoring and about problems associated with assessing a tutee's understanding. However, the untrained tutors did not receive any instruction on how to solve these problems. Instead, they were asked to provide tutoring in whatever manner they assumed appropriate.

## Procedure

Each tutoring session was divided into three phases: pretest phase, tutoring phase, and posttest phase. On average, a tutoring session lasted about 3 hours.

In the pretest phase, each tutee and each tutor individually read the passage about the human circulatory system. Afterwards, the trained tutors received training and the untrained tutors read the text.

In the tutoring phase, tutor-tutee dyads jointly read the passage about the human circulatory system sentence-bysentence and engaged in a dialogue about each sentence. All tutoring phases were videotaped.

In the posttest phase, the tutees completed the concepts test. The tutors also received the items of the concepts test and were asked to indicate for each item which of the given response options the tutee would choose.

## Codings and Analyses

Elicitation of Knowledge Deficits (Tutors) As an indicator of engaging in interactive tutoring strategies, we coded the knowledge deficits that a tutor elicited from a tutee. To do so, we used a coding scheme adapted from Chi et al. (2004). Every knowledge deficit that a tutee uttered was coded from its beginning to its end (event sampling procedure).

We coded a knowledge deficit whenever a tutor elicited from a tutee an utterance that (1) contradicted a piece of knowledge stated in the textbook passage, that (2) was incomplete, that (3) was vague, that (4) was incorrect and not addressed by the textbook passage, or when the tutee (5) did not utter a certain piece of information at all, that is, the tutee obviously missed this piece of information. In one tutoring session, for example, the tutor asked: "Why does the blood need to go to the lung? What does the lung do?" And the tutee answered: "Yes, um, yes, the lung filters the blood." This answer was coded as utterance of a knowledge deficit because it represents a normatively incorrect understanding. To standardize coding, the coder used a written instruction. For each tutor-tutee dyad, we summed up the number of elicited knowledge deficits.

Summative Assessment (Tutors) To examine the accuracy with which a tutor assessed a tutee's understanding of the human circulatory system after tutoring, we compared a tutee's responses in the concepts test with a tutor's estimations of a tutee's responses in the concepts test. To do so, we made the comparison on an item-by-item basis (cf. Hoge \& Coladarci, 1989). Hence, a tutor could achieve a maximum score of 16 points. Higher scores indicated a higher assessment accuracy.

Mediation Analysis To test our hypotheses, we performed a mediation analysis. We calculated total, direct, and indirect effects in accordance with our hypotheses by applying regression-based path analysis. To test for the statistical significance of an indirect effect, we derived 95\% confidence intervals for indirect effects as well as standard errors for indirect effects via bias-corrected bootstrap (for guidelines, see, e.g., Hayes, 2009, 2012). This approach resolves some methodological problems associated with the Sobel test (Hayes, 2009).

## Results

For all analyses, we used an alpha level of .05 . For directional hypotheses, we used one-tailed tests. In the analyses, trained tutors were coded as 1 and untrained tutors were coded as 0 . As effect size for indirect effects in the mediation analysis, we report $\kappa^{2}$. According to Preacher and Kelley (2011), effects are small when $\kappa^{2}=.01$, medium when $\kappa^{2}=.09$, and large when $\kappa^{2}=.25$. All analyses were performed using SPSS 20.0.0, the PROCESS macro for SPSS introduced in Hayes (2012; to perform the mediation analysis), and AMOS 20.0.0 (to receive standardized path coefficients for the mediation analysis). Table 1 shows the means and standard deviations of the dependent variables.

Table 1: Means and standard deviations (in parentheses) of the experiment's dependent variables

| Variable | Trained | Untrained | All |
| :--- | :---: | :---: | :---: |
|  | Tutors | Tutors | Tutors |
|  | $M(S D)$ | $M(S D)$ | $M(S D)$ |
| Elicited Knowledge | 71.30 | 32.11 | 52.21 |
| Deficits | $(40.46)$ | $(28.63)$ | $(40.01)$ |
| Assessment | 8.05 | 8.21 | 8.13 |
| Accuracy | $(2.54)$ | $(2.30)$ | $(2.40)$ |

## Impact of Training on Implementing Interactive Tutoring Strategies

Our first hypothesis stated that trained tutors would more often engage in interactive tutoring strategies than untrained tutors. Thus, trained tutors should elicit more knowledge deficits from their tutees than untrained tutors. As can be seen in Figure 1, trained tutors elicited more utterances of knowledge deficits from their tutee than did untrained tutors, $R^{2}=.25, F(1,37)=12.08, p<.05,95 \%$ CI $[.26, .74]$. Hence, the trained tutors in fact engaged in more interactive tutoring strategies than the untrained tutors.


Figure 1: Mediation model for the effect of tutor training on assessment accuracy explained by the number of expressed knowledge deficits a tutor elicited from a tutee. Numbers represent standardized path coefficients for direct effects and, in parentheses, the total effect of the independent variable on the dependent variable. ${ }^{*} p<.05$.

## Impact of Training on Summative Assessment

Our second hypothesis stated that trained tutors would more accurately assess a tutee's understanding after tutoring than untrained tutors. However, as the total effect depicted in Figure 1 shows, there was no significant difference in assessment accuracy between trained tutors and untrained tutors, $R^{2}<.01, F(1,37)=0.04, p>.05,95 \%$ CI [-.31, .24]. Hence, if only zero-order relations are taken into account, training tutors to implement interactive tutoring strategies failed to exert an influence on assessment accuracy.

## Interactive Tutoring Strategies as Mediator

Our third hypothesis stated that the higher number of a tutee's elicited knowledge deficits would explain why trained tutors assessed a tutee's understanding after tutoring more accurately than untrained tutors. To statistically test this hypothesis, we computed the indirect effect even though the total effect (i.e., the effect of training on assessment accuracy) was not significant (cf. Hayes, 2009; Shrout \& Bolger, 2002). To test the indirect effect, we constructed a bias corrected $95 \%$ bootstrap confidence interval as well as bootstrap standard errors from 10000 bootstrap samples. We found a significant negative indirect effect indicating that implementing interactive tutoring strategies as a result of receiving training decreased assessment accuracy with a standardized point estimate of $-.27(S E=.10), 95 \%$ CI [-.46, -.12], $\kappa^{2}=.26$ (zero-order correlation between elicited knowledge deficits and assessment accuracy: $r=-.43$, $p<.05$ ). Translated to unstandardized estimates, the number of items correctly estimated by trained tutors was 1.28 points ( $S E=0.54$ ) lower (and not higher) than the number of items correctly estimated by untrained tutors as mediated by the number of elicited knowledge deficits.

## Discussion

This study examined the effectiveness of a training method that aimed at helping tutors to engage in interactive tutoring strategies in the course of tutoring. It was assumed that engaging in interactive tutoring strategies would benefit a tutor's assessment of a tutee's understanding after tutoring.

First, we found that trained tutors in fact showed a more interactive style of tutoring than untrained tutors. Hence, even though the duration of our training was rather short, it was obviously sufficient to help the tutors to implement more interactive tutoring strategies. As a result, tutees tutored by trained tutors more often uttered knowledge deficits than tutees tutored by untrained tutors. This finding is consistent with the results obtained by Herppich et al. (2013a).

Second, however, the trained tutors failed to assess a tutee's understanding more accurately than the untrained tutors. The trained tutors were even less accurate than the untrained tutors. As show by the mediation analysis, this result was explained by the greater extent to which trained tutors engaged in interactive tutoring strategies as a result of receiving training. This effect was probably not observable in the zero-order analysis because the two paths making up the indirect effect were opposite in sign (cf. Hayes, 2009).

An explanation for why trained tutors and untrained tutors did not differ in assessment accuracy, as indicated by the total effect in the mediation analysis, is that the changes in the tutoring strategies due to receiving training might not have been sufficient to produce changes in assessment accuracy. This explanation would be in accordance with the results obtained by Roscoe and Chi (2007), who found that strategies of tutors can only be influenced to a certain extent. Hence, in the context of the present study, the information gained from being more interactive might not have been enough to generate more accurate assessments (cf. Graesser et al., 2011).

However, it still remains an open question as to why the elicitation of knowledge deficits was detrimental for assessing a tutee's understanding after tutoring, as indicated by the indirect effect in the statistical analysis. First, it might be that trained tutors and untrained tutors differed in the types of knowledge deficits they elicited from a tutee. Eliciting a larger number of scientifically incorrect utterances as compared to missing knowledge pieces, for example, might have been more informative for the summative assessment. This is because the incorrect response options in the concepts test were based on common types of incorrect understanding of a concept (e.g., Pelaez et al., 2005). However, the relative number of knowledge deficits elicited per category did not differ significantly between trained tutors and untrained tutors for any of the five categories of knowledge deficits coded.

Second, the detrimental effect of eliciting knowledge deficits on summative assessment might be related to our measure of summative assessment accuracy. During the training, the tutors were repeatedly informed that a tutor should get a picture of a tutee's understanding. As a consequence, the trained tutors might have conceived a tutee's understanding on a more global level than on the level of conceptual understanding. Thus, after having completed the training, being more interactive and receiving more information from the tutees could have drawn the tutors’ attention away from the knowledge they were to
assess in the concepts test. This conjecture could be tested in future research that uses measures of assessment accuracy that are as manageable for tutors as a multiple-choice test on conceptual knowledge but that would tap different levels of a tutee's understanding.

Third, another explanation refers to the fact that the tutors in this study did not possess teaching experience. Hence, the interactive tutoring strategies targeted in the training might have been quite unfamiliar to the tutors. As a result, implementing interactive tutoring strategies during tutoring might have put a fairly high burden on a tutor's cognitive capacity (Feldon, 2007). Thus, there might not have been enough cognitive capacity left to derive information from a tutee's utterances of knowledge deficits as a basis for assessing a tutee's understanding after tutoring.

This interpretation is in accordance with results from research on the acquisition of memory strategies. Often, learners can spontaneously implement a newly learned memory strategy but experience a so-called utilization deficiency (Miller, 1990). That is, implementing the strategy does not immediately improve recall or even hinders it. It is argued that using a newly learned strategy, which is not yet automated, demands most of the cognitive capacity of a learner. Thus, there is little capacity left to spend on processing the material to be recalled (e.g., Miller \& Seier, 1994).

Given this interpretation, it seems to be important to develop training methods that increase the automaticity with which interactive tutoring strategies are executed (Klauer, 1988). When interactive tutoring strategies occur more automatically, there might be more cognitive capacity available that can be used by tutors to assess a tutee's understanding (Feldon, 2007). Future research is encouraged to test whether training methods that target the automaticity of interactive tutoring strategies in fact improve assessment accuracy.

## Acknowledgments

We thank Hannah Bartels, Victoria Denise Claes, Julian Etzel, Sophia Kammer, Rico Krieger, Amelie Krug, Annette Lehmann, Karina Meyer, Bosse Nietsch, Anne-Kristin Rückert, Tatjana Scharping, Eva Wiemers, Lisa Zimmer, and Raoul Zimmermann for their help with many practical aspects of the project. This research was supported by grants from the German Science Foundation DFG (WI 3348/2-1).

## References

Brown, A. L., Campione, J. C., \& Day, J. D. (1981). Learning to learn: On training students to learn from texts. Educational Researcher, 10, 14-21.
Cade, W. L., Copeland, J. L., Person, N. K., \& D'Mello, S. K. (2008). Dialogue modes in expert tutoring. Paper presented at the Ninth International Conference on Intelligent Tutoring Systems, Montreal, Canada.
Chi, M. T. H. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. Topics in Cognitive Science, 1, 73-105.

Chi, M. T. H., Roy, M., \& Hausmann, R. G. M. (2008). Observing tutorial dialogues collaboratively: Insights about human tutoring effectiveness from vicarious learning. Cognitive Science, 32, 301-341.
Chi, M. T. H., Siler, S. A., \& Jeong, H. (2004). Can tutors monitor students' understanding accurately?. Cognition and Instruction, 22, 363-387.
Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T., \& Hausmann, R. G. M. (2001). Learning from human tutoring. Cognitive Science, 25, 471-533.
Cromley, J. G., \& Azevedo, R. (2005). What do reading tutors do? A naturalistic study of more and less experienced tutors in reading. Discourse Processes, 40, 83-113.
Dignath, C., Buettner, G., \& Langfeldt, H. P. (2008). How can primary school students learn self-regulated learning strategies most effectively? A meta-analysis on selfregulation training programmes. Educational Research Review, 3, 101-129.
D’Mello, S., Lehman, B. A., \& Person, N. K. (2010). Expert tutors feedback is immediate, direct, and discriminating. Paper presented at the Twenty-Third International Florida Artificial Intelligence Research Society Conference (FLAIRS 2010), Key West, FL.
Feldon, D. F. (2007). Cognitive load and classroom teaching: The double-edged sword of automaticity. Educational Psychologist, 42, 123-137.
Graesser, A. C., D’Mello, S., \& Cade, W. L. (2011). Instruction based on tutoring. In R. E. Mayer \& P. A. Alexander (Eds.), Handbook of research on learning and instruction. New York: Routledge.
Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. Communication Monographs, 76, 408-420.
Hayes, A. F. (2012). PROCESS: A versatile computational tool for observed variable mediation, moderation, and conditional process modeling [White paper]. Retrieved from http://www.afhayes.com/public/process2012.pdf
Herppich, S., Wittwer, J., Nückles, M., \& Renkl, A. (2013a). Addressing knowledge deficits in tutoring and the role of teaching experience: Benefits for learning and summative assessment. Manuscript submitted for publication.
Herppich, S., Wittwer, J., Nückles, M., \& Renkl, A. (2013b). Does it make a difference? Investigating the assessment accuracy of teacher tutors and student tutors. The Journal of Experimental Education, 81, 242-260.
Hmelo-Silver, C. E., \& Barrows, H. S. (2006). Goals and strategies of a problem-based learning facilitator. Interdisciplinary Journal of Problem-based Learning, 1, 21-39.
Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. Educational Psychology Review, 19, 509-539.
King, A., Staffieri, A., \& Adelgais, A. (1998). Mutual peer tutoring: Effects of structuring tutorial interaction to
scaffold peer learning. Journal of Educational Psychology, 90, 134-152.
Klauer, K. J. (1988). Teaching for learning-to-learn: A critical appraisal with some proposals. Instructional Science, 17, 351-367.
Leutner, D., Leopold, C., \& Den Elzen-Rump, V. (2007). Self-regulated learning with a text-highlighting strategy: A training experiment. Zeitschrift für Psychologie/ Journal of Psychology, 215, 174-182.
Mandl, H., \& Friedrich, H. F. (Eds.). (1992). Lern- und Denkstrategien. Analyse und Intervention [Learning and thinking strategies. Analysis and intervention]. Göttingen: Hogrefe.
McArthur, D., Stasz, C., \& Zmuidzinas, M. (1990). Tutoring techniques in algebra. Cognition and Instruction, 7, 197-224.
Michael, J. A., Wenderoth, M. P., Modell, H. I., Cliff, W., Horwitz, B., McHale, P., ... Whitescarver, S. (2002). Undergraduates' understanding of cardiovascular phenomena. Advances in Physiology Education, 26, 7284.

Miller, P. H. (1990). The development of strategies of selective attention. In D. F. Bjorklund (Ed.), Children's strategies: Contemporary views of cognitive development. Hillsdale: Erlbaum.
Miller, P. H., \& Seier, W. L. (1994). Strategy utilization deficiencies in children: When, where, and why. Advances in Child Development and Behavior, 25, 107156.

Pelaez, N. J., Boyd, D. D., Rojas, J. B., \& Hoover, M. A. (2005). Prevalence of blood circulation misconceptions among prospective elementary teachers. Advances in Physiology Education, 29, 172-181.
Perie, M., Marion, S., \& Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. Educational Measurement: Issues and Practice, 28, 5-13.
Preacher, K. J., \& Kelley, K. (2011). Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. Psychological Methods, 16, 93-115.
Renkl, A. (2005). The worked-out examples principle in multimedia learning. In R. E. Mayer (Ed.), Cambridge handbook of multimedia learning. New York: Cambridge University Press.
Roscoe, R. D., \& Chi, M. T. H. (2007). Understanding tutor learning: Knowledge-building and knowledge-telling in peer tutors' explanations and questions. Review of Educational Research, 77, 534-574.
Shrout, P. E., \& Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. Psychological Methods, 7, 422-445.
Sungur, S., \& Tekkaya, C. (2003). Students' achievement in human circulatory system unit: The effect of reasoning ability and gender. Journal of Science Education and Technology, 12, 59-64.

# Blending and Choosing Within One Mind: Should Judgments Be Based on Exemplars, Rules, or Both? 

Stefan M. Herzog (herzog@mpib-berlin.mpg.de)<br>Center for Adaptive Rationality, Max Planck Institute for Human Development, Lentzeallee 94 14195 Berlin, Germany<br>Bettina von Helversen (bettina.vonhelversen@unibas.ch)<br>Department of Psychology, University of Basel, Missionsstrasse 62a<br>4055 Basel, Switzerland


#### Abstract

Accurate judgments and decisions are crucial for success in many areas of human life. The accuracy of a judgment or decision depends largely on the cognitive process applied. In research on judgment, decision making, and categorization, two kinds of cognitive processes have often been contrasted: exemplar-based processes, which use similarity to previously encountered items to make judgments, decisions, and categorizations, and rule-based processes, which use abstracted cue knowledge. Although most cognitive models of judgment and decision processes assume that people rely on both processes, they differ in whether they assume that one process is selected or that both processes are blended into a single response. The present research takes a functional perspective and investigates what kind of interaction between the two processes leads to accurate responses. Based on crossvalidated simulations in real-world domains, it shows that blending rule- and exemplar-based processes generally leads to better judgments than does choosing between them, suggesting that the default strategy should be a blend of both processes, which is abandoned only when feedback justifies it.


Keywords: accuracy; multiple-cue judgments; decision making; categorization; exemplar models; rules; cognitive models; mixtures of experts; simulation.

## Introduction

Judging quantities, making decisions, and categorizing items are crucial elements of successful human behavior. A vast and diverse literature in cognitive science and judgment and decision making has investigated how people achieve these tasks (e.g., Ashby \& Maddox, 2005; Gigerenzer, Hertwig, \& Pachur, 2011; Kruschke, 2008; Payne, Bettman, \& Johnson, 1993). The many different models and strategies proposed can be broadly classified into two categories with reference to the cognitive processes they assume: exemplarbased processes, which use similarity to previously encountered items to make judgments, decisions, and categorizations, and rule-based processes, which use abstracted cue knowledge (Hahn \& Chater, 1998).

Extensive research has compared the proposed models' ability to describe human behavior. Furthermore, the performance of judgment and decision making strategies in predicting real-world criteria has been thoroughly investigated (e.g., Gigerenzer et al., 2011; Todd, Gigerenzer, \& the ABC Research Group, 2012).

To our knowledge, however, research in cognitive science and judgment and decision making has not previously investigated what kind of interaction between exemplar- and rule-based processes leads to accurate judgments, decisions, and categorizations: relying on just one of the two processes or using both? If both are considered, is it better to choose between them depending on the structure of the task, for instance (Rieskamp \& Otto, 2006), or to blend them into a joint response? This paper presents first answers to these questions.

A functional perspective on the interaction between exemplar- and rule based processes may be useful for at least three reasons. First, examining cognitive models' ability to predict external real-world criteria goes a step further than comparing their ability to describe human behavior in idealized laboratory tasks, by adding a further evaluation criterion. If one class of cognitive models were superior to another in terms of predictive performance, this would make them more attractive as plausible models of human behavior (Chater \& Oaksford, 1999). Second, many cognitive models are inspired by or share similarities with models from research fields interested in predictive performance (such as statistics, artificial intelligence, computer science, and machine learning; see e.g., Jäkel, Schölkopf, \& Wichmann, 2009; Marling, Sqalli, Rissland, Munoz-Avila, \& Aha, 2002), and a functional perspective provides a common ground that serves to re-connect cognitive models with such fields. Third, knowledge of how to profit from the complementary strengths of the two processes could offer prescriptions for improving human judgment, decision making, and categorization by instructing decision makers on when and how to use the two processes.

## Models of Judgment, Decision Making, and Categorization

There are two general approaches to modeling human cognition. First, single general-purpose models have been proposed (e.g., Lee \& Cummins, 2004). For instance, judgment and categorization models assume either only exemplar-based (e.g., Juslin \& Persson, 2002; Kruschke, 1992) or only rule-based processes (e.g., Ashby \& Gott, 1988; Brehmer, 1994). Second, toolbox approaches have been proposed. These assume that people draw on multiple,
different processes to solve the same task (e.g., Gigerenzer \& Selten, 2001). The toolbox approach posits that people adaptively select a tool (i.e., strategy) likely to succeed in the task at hand from a repertoire of strategies: the "toolbox" (Gigerenzer \& Selten, 2001; Payne et al., 1993; Rieskamp \& Otto, 2006; Scheibehenne, Rieskamp, \& Wagenmakers, 2013). Toolbox approaches have gained popularity particularly in decision making (e.g., Gigerenzer \& Selten, 2001; Rieskamp \& Otto, 2006). Yet also in categorization and judgment research, it is frequently assumed that people chose the process that is better suited to solving a task (Ashby, Alfonso-Reese, Turken, \& Waldron, 1998; Juslin, Karlsson, \& Olsson, 2008; Nosofsky, Palmeri, \& McKinley, 1994; von Helversen \& Rieskamp, 2008). For example, COVIS assumes that similarity-based and rulebased processes "race" for an answer, with the faster one determining the response (Ashby et al., 1998).

Although toolbox approaches often assume competition between processes, it is also possible that the processes cooperate. Hybrid or blending models assume that, instead of "choosing" a process for a task, two or more processes are executed simultaneously and their responses are integrated. For instance, the categorization model ATRIUM (Erickson \& Kruschke, 1998) combines both exemplar- and rule-based processes. Inspired by the "mixtures-of-experts" approach from machine learning (Jacobs, Jordan, Nowlan, \& Hinton, 1991), ATRIUM assumes that people have two "experts" in their mind: an exemplar-based and a rule-based one, whose outputs are processed by a gating mechanism. This gating mechanism can "choose" between these modules or "blend" their outputs by averaging their responses. In addition, ATRIUM can learn to rely more strongly on the more successful module (in terms of the probability of choosing or weighted averaging)-either for the whole task or depending on the item presented (i.e., depending on its location in psychological space). Modeling and experimental investigations support ATRIUM's assumption that exemplar- and rule-based processes simultaneously influence how humans categorize (e.g., Erickson \& Kruschke, 1998; Hahn, Prat-Sala, Pothos, \& Brumby, 2010). There is also evidence for such simultaneous influence in the domain of multiple-cue judgments (von Helversen, Herzog, \& Rieskamp, in press).

## Blending and Choosing Within One Mind

The combination of judgments or decisions from different sources is a vibrant topic in research fields such as psychology, judgment and decision making, cognitive science, statistics, artificial intelligence (AI), machine learning, biology, and economics (e.g., Krause, Ruxton, \& Krause, 2010; Kuncheva, 2004; Larrick, Mannes, \& Soll, 2012; Lee, Zhang, \& Shi, 2011; Marling et al., 2002). Combining diverse sources (e.g., forecasts from different experts) generally improves accuracy because different sources often compensate for each other's shortcomings. Depending on the circumstances, either choosing between
("competition") or blending different sources ("cooperation") may lead to better performance.

On the one hand, choosing a specific strategy allows the overall decision process to be adapted to environmental regularities and thus facilitates good performance (e.g., Todd et al., 2012). On the other hand, "blending" (i.e., averaging) different sources can often improve accuracy because errors of different signs cancel each other out. This "wisdom of crowds" phenomenon (Surowiecki, 2004) has recently also been applied to individual minds (e.g., Herzog \& Hertwig, 2009, 2013; Vul \& Pashler, 2008). Combining exemplar- and rule-based processes can be seen as an implicit "crowd within," where the two processes constitute two "experts" in one mind that either compete or cooperate in giving a response. To the extent that exemplar- and rulebased processes complement each other in the errors they commit, combining them may be a successful strategy (Herzog \& von Helversen, 2013).

In the following simulation study, we compare the merits of single purpose models, a competitive toolbox approach, and a cooperative toolbox approach. We focus on exemplarbased and rule-based processes as examples of distinctive cognitive processes because of the prominent distinction between the two in the cognitive literature (Ashby et al., 1998; Hahn \& Chater, 1998; Nosofsky et al., 1994; Persson \& Rieskamp, 2009).

## Different Levels of Interaction: Task or Item

Besides differentiating between choosing (competition) and blending (cooperation) of cognitive processes, we also consider on which level the interaction takes place: the task or item level. In the ecological rationality and adaptive toolbox approach (Todd et al., 2012), it is (implicitly) assumed that the selection of strategies happens on the task level-that is, that all the decisions within the same task are solved using the same strategy (once learning has completed). However, strategy selection (or integration) can also happen on the item level-that is, some items may be better solved by a rule, whereas others require memorization (Nosofsky et al., 1994). To account for this level of interaction, we compared competition and cooperation on the task and the item level.

## Simulation Study: Should Judgments Be Based on Exemplars, Rules or Both?

We investigated the performance of different ways to use exemplar- and rule-based processes in predicting a continuous criterion based on multiple cues. To this end, we conducted cross-validated simulations, informed by ATRIUM's (Erickson \& Kruschke, 1998) cognitive architecture, in five real-world domains. We addressed the following three questions. First, is it better to be equipped with both exemplar- and rule-based processes or is one process enough to achieve accurate judgments? Second, if both processes are used, is it better to choose between them

Table 1: Characteristics of the real-world datasets (adapted from Table 1 in Dana \& Dawes, 2004).
$N=$ number of cases, $k=$ number of cues, $\rho=$ correlation between target variable and predicted values from a multiple linear regression, $\mathbf{v}$ Vector $=$ zero-order correlation between target variable and cues, $\varnothing r_{x i x j}=$ mean correlation among cues.

| Dataset | $N$ | $k$ | $\rho$ | $\mathbf{v}$ Vector | $\otimes r_{x i x j}$ |
| :--- | ---: | ---: | :---: | :---: | :---: | ---: |
| Abalone | 4,177 | 7 | .73 | .63 .58 .56 .56 .54 .50 .42 | .89 |
| NFL | 3,057 | 10 | .54 | .46 .43 .37 .34 .33 .27 .21 .07 .05 .05 | .21 |
| ABC | 955 | 5 | .35 | .32 .20 .06 .04 .02 | .08 |
| NES | 1,910 | 6 | .35 | .26 .17 .15 .15 .13 .12 | .11 |
| WLS | 6,385 | 5 | .20 | .13 .11 .10 .10 .10 | .15 |

(competition) or to blend them (cooperation)? Third, for either choosing between or blending the two processes, is it better to treat all items the same (i.e., integration on the task level) or to treat individual items differently (i.e., integration on the item level)? Item-level integration implies choosing between the processes for each item (in the competitive approach) or weighting the two processes differently for each item when blending (in the cooperative approach).

## Datasets

We analyzed datasets previously used to compare the performance of proper and improper linear models (Dana \& Dawes, 2004). The datasets pertain to five domains: biology, sports, public opinion, political sentiment, and occupational prestige. In all datasets, a continuous target variable was predicted by several cues. For instance, the ABC dataset was derived from a 2002 poll of 955 U.S. households. Respondents' confidence that Osama bin Laden would be captured or killed was predicted by five cues, including the respondent's age, education, gender, and patriotism. See Table 1 for details of the statistical structure.

## Cognitive Models

Exemplar Model To represent an exemplar-based judgment process, we used an exemplar model for multiple-cue judgments (Juslin et al., 2008). The model assumes that judgments are based on the similarity to exemplars stored in memory, where the judgment is an average of the criterion values of the stored exemplars weighted by their similarity to the target item. We used a simplified exemplar model with one single free parameter determining the similarity gradient (see von Helversen \& Rieskamp, 2008).

Rule Model To represent a rule-based process, we used a multiple linear regression model. Such models have been widely used to model human judgment (Brehmer, 1994); they assume that judgments can be understood as the sum of weighted cue values. The model has a free parameter for every cue plus an intercept.

## Simulation Setup

For each simulation run, we randomly drew a learning sample and a test sample. We then fitted the free parameters of the exemplar and the rule model to the learning sampleminimizing the root mean square error (RMSE) between model predictions and criterion values-and used the
estimated parameter values to make predictions for the items in the test sample (for six different strategies described below). We measured estimation accuracy in the test sample using the RMSE between the model's predictions and the criterion values, a commonly used measure of absolute goodness of fit. Seven different sizes of learning samples were used $(20,40,60,80,100,200$, and 500 items) to vary the amount of experience with a domain; all test samples consisted of 250 items. For each dataset and each of the sizes of learning samples, we ran the simulation 1,000 times and averaged the results.

## Using the Rule and Exemplar Models

We tested six strategies for using rule- and exemplar-based processes to make predictions for the test sample.
"Exemplar Model" and "Rule Model" The first two strategies used just one of the two processes exclusively.
"Choosing-Task" and "Choosing-Item" The third and fourth strategy chose either the exemplar or the rule model.

On the task level, "choosing-task" selected in each simulation run the model that was superior in the learning sample and used it for all items in the test sample. To account for differences in model complexity, we used the Bayesian Information Criterion as a selection criterion.

On the item level, "choosing-item" selected in each simulation run and for each item in the test sample the model that was more likely to be superior for this particular test item-based on the performance on similar items in the learning sample. Specifically, for each test item we calculated the RMSE that the exemplar and the rule model had on similar items in the learning sample (i.e., we weighted the RMSE values of each training item using the similarity gradient of the exemplar model). The process with the lower weighted RMSE was then selected and its prediction for this test item was used.
"Blending-Average" and "Blending-Item" The fifth and sixth strategy blended the outputs of the exemplar and the rule model to make a joint prediction.

On the task level, "blending-average" computed for each test item the arithmetic mean of the predictions of the rule and the exemplar model.

On the item level, "blending-item" used in each simulation run and for each item in the test sample a


Figure 1: Cross-validated estimation accuracy (Root Mean Squared Error, RMSE) of six strategies in five domains (for learning samples of different sizes). The upper left panel averages the normalized data across domains; the RMSE values were divided by the largest average RMSE value in each domain. The strategies are explained in the text.
weighted average of both models' predictions-using the same similarity-weighted RMSEs as in "choosing-item." The item-specific weight for the exemplar model was calculated as the proportion of the rule model's weighted RMSE relative to the sum of both models' weighted RMSEs (i.e., the worse the rule model, the larger the weight on the exemplar model).

## Results \& Discussion

Figure 1 shows the generalization performance of the different strategies as a function of the size of the learning sample for the five domains. Because the datasets differed in their range of criterion values, which in turn affected the scale of the RMSE, it was necessary to normalize the RMSEs before aggregating them across datasets. To this end, we divided each RMSE by the largest average RMSE value within the respective domain, so that each RMSE value could be understood as the relative increase in fit. We then constructed a summary learning curve by averaging the normalized RMSEs across the five domains (see Figure 1, upper left panel).

Four results are noteworthy. First, "blending-average" was generally more accurate than either the exemplar or the
rule model; the exemplar model was somewhat better than the averaged predictions of both models only for very small learning samples (i.e., 20 items). Second, "blendingaverage" was generally more accurate than choosing the better model based on its performance in the respective learning sample ("choosing-task"), although choosing was slightly better for very small learning samples (i.e., 20 items) in two of the five datasets. Third, when choosing or blending, it did not pay off to tune one's use of the models to the type of item. Weighting both processes when blending ("blending-item") was less or equally accurate than was giving them equal weights ("blending-average"); similarly, choosing the process depending on the item ("choosing-item") was less or equally accurate than was using the same process for all items ("choosing-task"). Fourth, the differences between strategies decreased as the size of the learning samples increased.

Let us now answer the three questions motivating this simulation. First, in the datasets we investigated, it was generally better to be equipped with both exemplar- and rule-based processes than with just one of the two processes. Second, if both processes were used, it was generally better to blend them than to choose between them. Third, when
choosing between or blending the two processes, it was generally better to treat all items the same (and not to choose or blend, respectively, depending on the type of item; i.e., depending on how much "expertise" the exemplar- and rule-based processes had about a specific part of the psychological space).

## General Discussion

Many cognitive models of judgment, decision-making, and categorization assume that people can use both exemplarand rule-based processes (e.g., Erickson \& Kruschke, 1998). Yet it remained unclear whether using both processes provides a performance advantage over using just one process and, when both processes are available, whether it is better to choose one process depending on the task (i.e., competitive toolbox approach) or to blend their responses (i.e., cooperative toolbox approach). Our simulations in the domain of multiple-cue judgments suggest that combining the two processes (either by choosing between or blending them) leads to better judgments than does relying on just one of them, and that a simple blend (i.e., equal weighting) of both processes leads to accurate judgments. This latter point is consistent with the success of naïve equal weighting strategies (e.g., Dawes, 1979). In another set of simulations, we investigated the combination (i.e., choosing or blending) of exemplar- and rule-based processes in the context of making categorizations (using 38 machine learning benchmark datasets; Herzog \& von Helversen, 2013). Further broadening the scope of the present analysis, we found that blending the outputs of an exemplar- and a rulebased process led to successful categorizations.

Our results resonate with research in AI and machine learning that demonstrates how combining different representations is often beneficial (Kuncheva, 2004; Marling et al., 2002). More specifically, our results suggesting that combining exemplar- and rule-based processes can often increase accuracy in human cognition dovetail nicely with the successful combination of casebased and rule-based reasoning systems in AI (e.g., Marling et al., 2002; Prentzas \& Hatzilygeroudis, 2007).

Besides the general question of whether exemplar- and rule-based processes should be "blended" or "chosen" among, our simulations suggest that it does not pay off to tune one's use of exemplar- and rule-based processes to the type of item one wants to generalize to. This conclusion seems inconsistent with empirical studies suggesting that participants successfully choose between processes in categorization tasks (e.g., Erickson, 2008). Yet these experimental tasks may be unrepresentative of real-world situations. In many experimental studies-especially in categorization research-there is little (or no) doubt about which process is better suited to solving the whole task (or responding to a specific item), and a participant can thus learn to choose between or differentially use the two processes. We speculate that deviating from a simple blending strategy is generally worthwhile only in domains in which one process is clearly superior to the other, both
processes make similar errors, and this statistical structure can be ascertained with enough confidence (see Soll \& Larrick, 2009). However, we would argue that this is typically not the case in real-world domains. It would thus seem prudent that human judges and decision makers, as modeled, for example, by ATRIUM (Erickson \& Kruschke, 1998), start with a simple blend of both processes and deviate from this approach (e.g., by choosing or itemspecific tuning) only when feedback justifies it.

Why is combining exemplar- and rule-based processes so successful in multiple-cue judgment tasks? The use and the performance of exemplar- and rule-based processes in multiple-cue judgment tasks seems to depend on the statistical structure of the task-in particular, the functional relation between cues and criteria (Juslin et al., 2008; von Helversen \& Rieskamp, 2008). If the criterion can be approximated by a linear additive combination of the cues, rule-based processes predominate. In multiplicative tasks, by contrast, exemplar-based processes perform better and are used more frequently. Simulations using artificially created domains (Herzog \& von Helversen, 2013) suggest that the five real-world domains we analyzed in the present simulations represent a mixture of these two kinds of statistical structures (i.e., additive and multiplicative). Consequently, neither of the two processes in isolation was able to capture their statistical structure. To the extent that this result generalizes to decision making and categorization, it suggests one reason why people are equipped with and use both exemplar- and rule-based processes: because only a combination of the two allows people to make successful judgments, decisions, and categorizations in the real world.

## Acknowledgments

We thank the Swiss National Science Foundation for a grant to the first author (100014_129572/1).

## References

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., \& Waldron, E. (1998). A neuropsychological theory of multiple systems in category learning, 105, 442-481.
Ashby, F. G., \& Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. Journal of Experimental Psychology: Learning, Memory, and Cognition, 14, 33-53.
Ashby, F. G., \& Maddox, W. T. (2005). Human category learning. Annual Review of Psychology, 56, 149-178.
Brehmer, B. (1994). The psychology of linear judgement models. Acta Psychologica, 87, 137-154.
Chater, N., \& Oaksford, M. (1999). Ten years of the rational analysis of cognition. Trends in Cognitive Sciences, 3, 57-65.
Dana, J., \& Dawes, R. M. (2004). The superiority of simple alternatives to regression for social science predictions. Journal of Educational and Behavioral Statistics, 29, 317-331.
Dawes, R. M. (1979). The robust beauty of improper linear
models in decision making. American Psychologist, 34, 571-582.
Erickson, M. A. (2008). Executive attention and task switching in category learning: Evidence for stimulusdependent representation. Memory \& Cognition, 36, 749-761.
Erickson, M. A., \& Kruschke, J. K. (1998). Rules and exemplars in category learning. Journal of Experimental Psychology: General, 127, 107-140.
Gigerenzer, G., Hertwig, R., \& Pachur, T. (2011). Heuristics: The foundations of adaptive behavior. Oxford: Oxford University Press.
Gigerenzer, G., \& Selten, R. (Eds.). (2001). Bounded rationality: The adaptive toolbox. Cambridge, MA: MIT Press.
Hahn, U., \& Chater, N. (1998). Similarity and rules: Distinct? Exhaustive? Empirically distinguishable? Cognition, 65, 197-230.
Hahn, U., Prat-Sala, M., Pothos, E. M., \& Brumby, D. P. (2010). Exemplar similarity and rule application. Cognition, 114, 1-18.
Herzog, S. M., \& Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. Psychological Science, 20, 231-237.
Herzog, S. M., \& Hertwig, R. (2013). The crowd-within and the benefits of dialectical bootstrapping: A reply to White and Antonakis (2013). Psychological Science, 24, 117-119.
Herzog, S. M., \& von Helversen, B. (2013). The benefits of combining cognitive processes. Manuscript in preparation.
Jacobs, R. A., Jordan, M., Nowlan, S., \& Hinton, G. (1991). Adaptive mixtures of local experts. Neural Computation, 3, 79-87.
Jäkel, F., Schölkopf, B., \& Wichmann, F. A. (2009). Does cognitive science need kernels? Trends in Cognitive Sciences, 13, 381-388.
Juslin, P., Karlsson, L., \& Olsson, H. (2008). Information integration in multiple cue judgment: A division of labor hypothesis. Cognition, 106, 259-298.
Juslin, P., \& Persson, M. (2002). PROBabilities from Exemplars (PROBEX): A "lazy" algorithm for probabilistic inference from generic knowledge. Cognitive Science, 26, 563-607.
Krause, J., Ruxton, G. D., \& Krause, S. (2010). Swarm intelligence in animals and humans. Trends in Ecology and Evolution, 25, 28-34.
Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. Psychological Review, 99, 22-44.
Kruschke, J. K. (2008). Models of categorization. In R. Sun (Ed.), The Cambridge handbook of computational psychology. New York, NY: Cambridge University Press.
Kuncheva, L. (2004). Combining pattern classifiers: Methods and algorithms. Hoboken, NJ: John Wiley \&

Sons.
Larrick, R. P., Mannes, A. E., \& Soll, J. B. (2012). The social psychology of the wisdom of crowds. In J. I. Krueger (Ed.), Frontiers in social psychology: Social judgment and decision making. New York, NY: Psychology Press.
Lee, M. D., \& Cummins, T. D. R. (2004). Evidence accumulation in decision making: Unifying the "take the best" and the "rational" models. Psychonomic Bulletin \& Review, 11, 343-352.
Lee, M. D., Zhang, S., \& Shi, J. (2011). The wisdom of the crowd playing The Price Is Right. Memory \& Cognition, 39, 914-923.
Marling, C., Sqalli, M., Rissland, E. L., Munoz-Avila, H., \& Aha, D. (2002). Case-based reasoning integrations. $A I$ Magazine, 23, 69-86.
Nosofsky, R. M., Palmeri, T. J., \& McKinley, S. C. (1994). Rule-plus-exception model of classification learning. Psychological Review, 101, 53-79.
Payne, J. W., Bettman, J. R., \& Johnson, E. J. (1993). The adaptive decision maker. Cambridge, UK: Cambridge University Press.
Persson, M., \& Rieskamp, J. (2009). Inferences from memory: Strategy- and exemplar-based judgment models compared. Acta Psychologica, 130, 25-37.
Prentzas, J., \& Hatzilygeroudis, I. (2007). Categorizing approaches combining rule-based and case-based reasoning. Expert Systems, 24, 97-122.
Rieskamp, J., \& Otto, P. E. (2006). SSL: A theory of how people learn to select strategies. Journal of Experimental Psychology: General, 135, 207-236.
Scheibehenne, B., Rieskamp, J., \& Wagenmakers, E.-J. (2013). Testing adaptive toolbox models: A Bayesian hierarchical approach. Psychological Review, 120, 3964.

Soll, J. B., \& Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. Journal of Experimental Psychology: Learning, Memory, and Cognition, 35, 780-805.
Surowiecki, J. (2004). The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations. Garden City, NY: Doubleday.
Todd, P. M., Gigerenzer, G., \& the ABC Research Group. (2012). Ecological rationality: Intelligence in the world. Oxford, UK: Oxford University Press.
von Helversen, B., Herzog, S. M., \& Rieskamp, J. (in press). Haunted by a Doppelgänger: Irrelevant facial similarity affects rule-based judgments. Experimental Psychology.
von Helversen, B., \& Rieskamp, J. (2008). The mapping model: A cognitive theory of quantitative estimation. Journal of Experimental Psychology: General, 137, 73-96.
Vul, E., \& Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. Psychological Science, 19, 645-647.

# Using the letter decision task to examine semantic priming 

Tom Heyman (tom.heyman@ppw.kuleuven.be)<br>Department of Experimental Psychology, 102 Tiensestraat<br>3000 Leuven, Belgium

Simon De Deyne (simon.dedeyne@ppw.kuleuven.be)
Department of Experimental Psychology, 102 Tiensestraat
3000 Leuven, Belgium

Gert Storms (gert.storms@ppw.kuleuven.be)<br>Department of Experimental Psychology, 102 Tiensestraat<br>3000 Leuven, Belgium


#### Abstract

The present research investigates semantic priming with an adapted version of the word fragment completion task. The letter decision task, as we will call it, holds some advantages over the traditionally used lexical decision task in that it eliminates retrospective semantic matching effects, it avoids the need to construct pseudowords, it is more engaging for participants and it enhances semantic processing, which in turn allows for a more fine-grained investigation of semantic activation. The letter decision task requires participants to complete words, from which one letter was omitted like lett_ce (lettuce), as fast as possible. The study found that words are completed faster when the preceding trial comprised a semantically related fragment like tom_to (tomato) than when it comprised an unrelated fragment like guit_r (guitar). Furthermore, the study provides insight in the nature of the priming effect. It demonstrates that priming effects are larger for strongly associated prime-target pairs.


Keywords: Semantic priming; Letter decision task; Associative strength.

## Introduction

Semantic priming is the finding that the processing of targets (e.g., a picture, a word,...) preceded by a semantically related prime (also a picture, a word,...) is enhanced. For instance, the presentation of the word cat facilitates processing of the subsequently presented word dog. One of the debates in the semantic priming literature concerns the source of the priming effect (Hutchison, 2003; Lucas, 2000). The (unresolved) issue revolves around the type of relation between concepts that is necessary for priming to occur. That is to say, words can be associatively related, as evidenced by association norms (De Deyne, Navarro \& Storms, 2012) or because both concepts share certain features. Returning to the cat-dog example, both cats and dogs have four legs, two eyes, are pets, etc. and thus they are related in terms of feature overlap (e.g., McRae \& Boisvert, 1998). Moreover, the strongest associate of cat is dog hence both concepts are also associatively related. Whether priming is driven by word associations or feature overlap (or even something else) is an important question
since it has significant repercussions for theories about the organization of the mental lexicon. Consequently, a lot of research has been devoted to this topic.

The most frequently used paradigms to examine these issues are the lexical decision task, in which participants have to decide whether letter strings form existing words or not, and, to a lesser extent, the pronunciation task, in which participants read aloud words (see the reviews of Hutchison (2003), Lucas (2000) and Neely (1991)). The experimental designs further vary in the degree to which they allow automatic and controlled processes. These latter processes are conscious and strategic and they come into play when the prime-target coupling (e.g., cat-dog) is made explicit (Jones, 2010). This is for instance the case in the standard lexical decision task where participants are required to respond only to the second item of the pair (i.e., the target $d o g$ ) and not to the first (i.e., the prime cat). Strategic effects are volatile and vary over subjects, whereas automatic processes are ubiquitous. Thus, automatic processes are thought to reliably reflect the structure of the mental lexicon (Lucas, 2000). Hence, considerable effort has been put into developing methodologies that prevent controlled processes. One method to reduce strategic effects is the continuous lexical decision task (McNamara \& Altarriba, 1988; Shelton \& Martin, 1992). Here, prime-target pairs are decoupled by asking participants to respond not only to the target but also to the prime.

In the present study, we took a different approach. It was (partly) motivated by the fact that there is little consensus regarding the nature of semantic priming. A possible explanation for the divergent and sometimes unreplicated findings (see Hutchison (2003) and Lucas (2000)) is that the experimental paradigms are not sensitive enough to detect or tease apart subtle effects. The widely used lexical decision task may rely more on superficial processing of words, whereas deeper semantic processing may be necessary to fully uncover the structure of the mental lexicon. Hence, in this study, we used a different method to examine semantic priming. It is an adaptation of the word fragment completion task, a task that has mainly been used in implicit memory studies (i.a., Bassili, Smith \& MacLeod,

1989; Challis \& Brodbeck, 1992; McDermott, 1997; Roediger \& Challis, 1992; Weldon, 1993). There are several variants of the word fragment completion task, but the general idea is that participants are presented with words from which one or more letters are omitted (e.g., r_d or _orn_d_). Participants then are assigned to fill in the gap(s). In some experiments, the dependent variable of interest is the actual answer participants give. Put differently, the question is whether participants complete $\mathrm{r}_{-} \mathrm{d}$ as red or as rod. In other experiments, there is only one correct answer and the crucial dependent variable is the proportion correct responses within a certain time interval or alternatively, the time required to give the correct solution. Concretely, how many participants accurately identify _orn_d_ as tornado and/or what is the average reaction time? In this study, we examined semantic priming using a modification of the latter type. But instead of difficult words with many blank spaces, we opted for relatively simple stimuli with only one blank space. Furthermore, participants were told that the missing letter was always a vowel. The task conceptually resembled a continuous lexical decision task in that participants had to complete both prime and target words (and also unrelated filler items). For instance, on trial $n$ participants got the fragment tom_to (it should be completed as tomato) and on trial $n+1$ they got lett_ce (it should be completed as lettuce). For the sake of clarity, we will therefore coin the term continuous letter decision task to refer to the experimental paradigm in this study. As in a (continuous) lexical decision task, the main dependent variable is reaction time since accuracy will be near perfect. Hence, it is expected that lett_ce is completed faster when it is preceded by a semantically related stimulus like tom_to than when it is preceded by an unrelated stimulus like guit_r (it should be completed as guitar).

We believe that there are some advantages of the continuous letter decision task over the continuous lexical decision task. First of all, in the lexical decision task participants may endorse a retrospective semantic matching strategy. Neely and Keefe (1989) argued that participants might use information about whether the considered letter string is semantically related to the preceding letter string to reduce their response time. Concretely, when there is a semantic relation between two consecutively presented letter strings, the correct answer for the latter letter string is always "word". If there is no such relation, the second letter string is a word or a non-word. In fact, when the proportion of non-words in the experiment is high then the absence of a relation between two consecutive letter strings indicates that the second letter string is more likely to be a non-word. It is possible that participants notice these contingencies, which in turn yields strategic priming effects that are inseparable from (interesting) automatic priming effects. However, the continuous letter decision task introduced here does not suffer from a semantic matching strategy. That is to say, a semantic relation between two words on consecutive trials is not predictive for the correct response to the latter word fragment. The fact that tomato and lettuce are related does
not give information about which vowel is missing in the fragment lett_ce.

A second advantage of the letter decision task with respect to the lexical decision task is that it obviates the need to construct pseudowords. Besides practical convenience, it has also theoretical implications since previous research suggested that the nature of the pseudowords and their similarity to real words modifies priming (Shulman \& Davison, 1977) and also the word frequency effect (Stone \& Van Orden, 1993). Such issues are avoided in the letter decision task.

Thirdly, it is not far-fetched to argue that the letter decision task is more challenging, without becoming burdensome, than the lexical decision task. Although participants may not exactly be filled with joy when performing the experiment, the task is more engaging, which in turn enhances the intrinsic motivation of participants (Deci \& Ryan, 1985).

Finally and perhaps most importantly, the letter decision task presumably involves a deeper semantic processing. In the lexical decision task, shallow processing of letter strings may be sufficient to discriminate words from non-words (Rogers, Lambon Ralph, Hodges \& Patterson, 2004), thereby limiting the facilitatory effect of a related prime. Because the letter decision task is more effortful, a related prime has more potential to exert its influence.

Taken together, it may be fruitful to use the letter decision task to examine semantic priming. Hence, the first goal of the present study was to establish whether a priming effect could be obtained with this task.

A second goal was to examine the nature of the priming effect. Every crucial target like lett_ce (lettuce) was either preceded by a related prime (tom_to, tomato) or an unrelated prime (guit_r, guitar). As is traditionally the case in priming research, one could consider relatedness as a dichotomy (i.e., tomato-lettuce are related whereas guitarlettuce are not). However, one could argue that relatedness is not an all or none matter, but rather that there is variability in the strength with which two words are related (for a similar proposal, see Hutchison, Balota, Cortese \& Watson, 2008). For instance, thunder-lightening has a stronger forward association than tomato-lettuce, meaning that more people give lightning as an association for thunder than lettuce as an association for tomato (based on the large scale Dutch Word Association Database from De Deyne et al., 2012). Thus, one might hypothesize that the priming effect for thunder-lightening is stronger than the effect for tomato-lettuce. The second goal of this study was to examine this prediction.

## Method

## Participants

Participants were 40 first-year psychology students of the University of Leuven ( 7 men, 33 women, mean age 18 years), who participated in return for course credit. All participants were native Dutch speakers.

## Materials

A total of 76 related prime-target pairs like tom_to-lett_ce (tomato-lettuce) were constructed. All stimuli were Dutch word fragments. Primes and targets were always category coordinates. Categories ranged from fruits and music instruments to mammals, tools, professions, etc. Moreover, prime-target pairs had a forward association strength that ranged from $3 \%$ to $30 \%$. These and other measures of association strength were derived from the Dutch Word Association Database (De Deyne et al., 2012). In addition, 76 unrelated filler pairs were constructed.

All word fragments were generated by omitting one vowel from a Dutch noun. Only word fragments that had a unique correct response were used. Of the 76 crucial targets, 16 required an "a" response, 22 an "e" response, 18 an " $i$ " response, 13 an "o" response and 7 a "u" response.

Two lists were created such that a random half of the 76 crucial targets were preceded by their related prime in List A, whereas in List B they were preceded by an unrelated word, and vice versa. The 38 unrelated pairs for each list were constructed by randomly recombining primes and targets, with two limitations. The first is of course that the resulting prime-target pairs were no category coordinates and indeed unrelated, as evidenced by a lack of a forward and backward association between prime and target. Second, a fraction of the related prime-target pairs were response congruent, meaning that the same vowel is missing in both the prime and the target. The unrelated pairs were created in a way that they match in terms of response congruency. When a related pair is response congruent so is the corresponding unrelated pair and the other way around. So for example, there where pa_rd (to be completed as paard, Dutch for horse) was preceded by zebr_ (to be completed as zebra) in List A, it was preceded by t_rwe (to be completed as tarwe, Dutch for wheat) in List B, which was actually the prime for me_l (to be completed as meel, Dutch for flour) in List A. Hence, each list consists of 76 critical prime-target pairs ( 38 related pairs and 38 unrelated pairs) and an additional 76 unrelated filler pairs.

## Procedure

Participants were randomly assigned to one of the two lists. Twenty participants received List A and 20 List B. The task itself was a continuous letter decision task. The continuous nature of the task breaks the 152 pairs down to 304 trials. On each trial, participants were presented with one word fragment. Primes were always shown on odd-numbered trials and targets on even-numbered trials. The order of the pairs within the experiment was random and varied over participants.

On every trial, participants saw a word from which one letter was omitted. They were informed that the missing letter was always a vowel. Participants had to complete the word by pressing either "a", "e", "u", " i ", or "o" on an AZERTY keyboard. The instructions stressed both speed and accuracy. Every word fragment was displayed in the center of the screen and remained present until a response
was made. The inter-trial interval was 500 ms . Before the experimental phase, participants did 20 practice trials. The practice trials were identical to the experimental trials except that 20 new semantically unrelated word fragments were utilized. The experiment was run on a Dell Pentium 4 with a 17.3-inch CRT monitor using Psychopy (Peirce, 2007). It was part of a series of unrelated experiments and took approximately 15 minutes.

## Results

First, the split-half reliability of the response times to the 76 crucial targets was calculated using the Spearman-Brown formula. Split-half correlations for List A and List B separately were obtained for 10,000 different randomizations of the participants. The resulting reliabilities, averaged over the 10,000 randomizations, were .92 for List A and .88 for List B, which is rather high for response times. Note that all analyses were performed only on the 76 crucial target trials.

Erroneously completed targets ( $3.3 \%$ of the data) and targets preceded by an incorrectly completed prime were not included in the analysis ( $5.3 \%$ of the data). Furthermore, responses faster than 250 ms and slower than 4000 ms were removed after which an individual cut-off value for each participant was computed as the mean response time plus 3 standard deviations. Response times exceeding this criterion were also excluded (another $3.9 \%$ of the data was discarded). The exclusion criteria are similar to regular priming studies using the standard lexical decision task, except for the exclusion of target trials following incorrect prime completion. This has to do with the continuous nature of the task: post-error slowing and/or subpar prime processing conceivably obscure target response times and/or priming effects. It should be noted though that the results were qualitatively the same if different exclusion criteria were used.

The log-transformed response times were then fitted using a mixed effects model with a random intercept for participants and items (i.e., the 76 crucial targets). The response times were regressed on 4 predictors: one critical predictor called Relatedness, which is a binary variable indicating whether the target (lett_ce, lettuce) was preceded by a related prime (tom_to, tomato) or an unrelated prime (guit_r, guitar), and three covariates, namely, Contextual Diversity of the target (CD Target ${ }^{1}$, acquired from Keuleers, Brysbaert \& New, 2010), Word Length of the target in number of characters (Length Target) and the logtransformed response time to the prime (RT Prime). To facilitate the interpretation of the effects, CD Target, Length Target and RT Prime were z-transformed. Furthermore, Relatedness was coded such that targets preceded by a related prime served as a baseline. Thus the intercept should be interpreted as the expected response time to a target with

[^296]an average length ( $\approx 6$ characters) and an average contextual diversity ( $\approx 2.4$ ) that was preceded by a related prime with an average response time ( $\approx 1103 \mathrm{~ms}$ ). The analyses were carried out in R (version 2.15.2) ( R development core team, 2011), employing the lme4 package (Bates \& Sarkar, 2007). Markov Chain Monte Carlo p-values (pMCMC) and 95\% highest posterior density intervals (HPD95) were obtained with the pvals.fnc() function of the languageR package, with 10,000 iterations (Baayen, 2008).

The results are summarized in Figure 1, which depicts the $95 \%$ highest posterior density interval for the fixed effects. Note that the HPD95 of the intercept, which ranged from 6.76 to 6.85 , is not presented because it would have distorted the x-axis. Figure 1 shows that all predictors have a HPD95 that excludes zero. Hence, there is a significant priming effect ( $\mathrm{pMCMC}<.001$ ). To grasp the magnitude of the effect, one can derive model predictions based on the point estimates of the fixed effects (i.e., the dots in Figure 1; the estimate of the intercept was 6.8). The expected response time for the average participant and the average target following an average related prime equals 904 ms . This response time increases to 944 ms when the target is preceded by an unrelated prime. In other words, there is a priming effect of 40 ms .

In the previous analysis, Relatedness was a binary predictor. However, a continuous variable is needed to examine whether a stronger relation between word pairs yields a larger priming effect. To this end, five predictors that capture the associative strength between two words were derived from the Dutch Word Association Database (De Deyne et al., 2012). The five predictors are Forward Association Strength (i.e., how often is the target given as an associate to the prime; FS), Backward Association Strength (i.e., how often is the prime given as an associate to the target; BS) and three semantic relatedness measures. Semantic relatedness was calculated by computing the distributional overlap of the vector of association response counts between a pair of words as the cosine between these vectors (S raw). In addition, two variations were included, where (a) the counts were logarithmically transformed (S $\log$ ) or (b) weighted using point-wise mutual information which is often used in semantic vector models (S pmi) (Church \& Hanks, 1989; Turney \& Pantel, 2010). Both related and unrelated prime-target pairs get a score for all five variables. For unrelated pairs, FS and BS values are all zero, but the presence of shared associates results in cosine values for $S$ raw, $S \log$ and $S$ pmi that are often somewhat larger than zero.

A model comparison approach was adopted to assess the merits of these continuous predictors with respect to the binary predictor. In a first step, the same mixed-effects model from the previous analysis was used, but now the binary predictor Relatedness was replaced by one of the five continuous variables. This results in six models of which the fit indices are reported in Table 1. The AIC and BIC scores reported in Table 1 evaluate the goodness of fit against the number of parameters of the model (Akaike, 1974; Schwarz,
1978). Lower values are indicative of a better fit. Since the models compared here are non-nested, AIC and BIC scores were used to assess which model, and thus which predictor, best fits the data. The results show that all continuous measures were better than the binary predictor. The best continuous predictor was $S$ log.

In a second step, we started from the model with S log and added the other continuous variables to investigate whether they can explain the remaining variance. It turned out that only BS was a significant predictor $(\mathrm{pMCMC}=$ $.011)$ besides $S \log (\mathrm{pMCMC}=.006)$.


Figure 1: 95\% highest posterior density intervals of the four regression weights. The dots represent the point estimates of the weights.

Table 1: AIC and BIC scores for the six mixed effects models. Models only differ in the predictor that captures the nature of the prime-target relations (the first column).

| Predictor | AIC | BIC |
| :--- | :--- | :--- |
| S log | 138.8 | 185.9 |
| S raw | 145.1 | 192.2 |
| S pmi | 141.8 | 188.8 |
| FS | 150.8 | 197.9 |
| BS | 140.1 | 187.1 |
| Relatedness (binary) | 152.8 | 199.9 |

## Discussion

The present research proposes a different method, that is, the letter decision task, to examine semantic priming. In this task, participants are shown words from which one letter (i.e., a vowel) is omitted. Participants have to fill in the missing letter as fast as possible. Word fragments were selected such that there was only one correct completion possible, thereby making the task conceptually comparable to the lexical decision task. As argued in the introduction,
there are several advantages over the lexical decision task. Concretely, the letter decision task eliminates retrospective matching effects, it does not require experimenters to construct pseudowords, it is more engaging than the lexical decision task and it involves deeper semantic processing. Crucially, this study shows that the continuous letter decision task can capture semantic priming effects. Hence, the present task is a viable alternative to examine semantic priming in future research. The employed methodology greatly reduces strategic priming effects, although it is theoretically possible that (some) participants engaged in expectancy generation despite the low relatedness proportion ${ }^{2}$. To completely disentangle automatic and strategic processes one might use a standard letter decision task with a short stimulus onset asynchrony. In this paradigm a briefly presented complete prime word is quickly replaced by a to-be-completed target. The short interval prevents expectancy generation (but not retrospective matching in a lexical decision task, see e.g., Shelton and Martin, 1992), while the letter decision task eliminates retrospective matching. In addition, one could manipulate the relatedness proportion in the continuous letter decision task to check whether expectancy generation plays a role. Our lab is currently investigating these issues.

Furthermore, this study provides evidence for the hypothesis that priming effects are greater for strongly related prime-target pairs. Models that regard relatedness as a continuous rather than a binary variable fitted the data better. More specifically, semantic relatedness and backward association strength were shown to predict the response times to the target word fragments the best. Thus, the stronger prime and target words are associated, the faster participants completed the target word. The fact that backward association strength plays a role seems to indicate that the benefit is larger for reciprocally associated primetarget pairs. These findings also highlight the value of the letter decision task. Because this task enhances semantic processing, it allows for a more detailed analysis of semantic activation, which may not be possible with a classic lexical decision task.

The method to assess the merits of continuous predictors over a binary predictor may seem a bit odd. Here, a model comparison approach was used, whereas it might be intuitively compelling to average over participants to obtain a priming effect for each separate item. Indeed, one could look at the average response time of the participants who got the related pair (e.g., tom_to-lett_ce, tomato-lettuce) and subtract it from the average response time of the participants who got the unrelated pair (e.g., guit_r-lett_ce, guitarlettuce) and this for all 76 crucial targets. The resulting 76 priming effects could be regressed on continuous measures like forward association strength, backward association

[^297]strength,... (see Hutchison et al., 2008 for such an approach). However, several researchers have argued against averaging over participants because it inflates type 1 error (Baayen, Davidson \& Bates, 2008; Lorch \& Myers, 1990; Quené \& van den Bergh, 2008). Nevertheless, the results from this study are largely consistent with those from Hutchison and colleagues (2008).

It should be noted that the present research only considers associative strength of prime-target pairs. As described in the introduction, it is debated whether semantic priming is primarily driven by associations between words or by similarity in terms of feature overlap between prime and target. Although this research did not directly address this issue, it does hint at the importance of associations. But we immediately hasten to point out that all related pairs in the experiment were category coordinates, hence there will be considerable feature overlap between related primes and targets as well. Future research incorporating a continuous measure for feature overlap can provide further insight on this matter.

## Acknowledgments

We wish to thank Pieter Moors for his comments and suggestions. Tom Heyman is a research assistant of the Research Foundation-Flanders (FWO-Vlaanderen). Correspondence should be addressed to Tom Heyman, Department of Experimental Psychology, University of Leuven, Tiensestraat 102, 3000 Leuven, Belgium. E-mail: tom.heyman@ppw.kuleuven.be

## References

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19, 716-723.
Baayen, R. H. (2008). Analyzing Linguistic Data. A Practical Introduction to Statistics Using R. Cambridge University Press.
Baayen, R. H., Davidson, D. J., \& Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language, 59, 390-412.
Bassili, J. N., Smith, M. C., \& MacLeod, C. M. (1989). Auditory and visual word-stem completion: Separating data-driven and conceptually driven processes. Quarterly Journal of Experimental Psychology, 41A, 439-453.
Bates, D. M., \& Sarkar, D. (2007). lme4: Linear mixedeffects models using $S 4$ classes. R package version 0.999375-42.

Brysbaert, M., \& New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behavior Research Methods, 41, 977-990.
Challis, B. H., \& Brodbeck, D. R. (1992). Level of processing affects priming in word fragment completion. Journal of Experimental Psychology: Learning, Memory \& Cognition, 18, 595-607.

Church, K., \& Hanks, P. (1989). Word association norms, mutual information, and lexicography. In Proceedings of the 27th Annual Conference of the Association of Computational Linguistics, pp. 76-83, Vancouver, British Columbia.
Deci, E. L., \& Ryan, R. M. (1985). Intrinsic motivation and self-determination in human behavior. New York: Plenum.
De Deyne, S., Navarro, D., \& Storms, G. (2012). Better explanations of lexical and semantic cognition using networks derived from continued rather than single word associations. Behavior Research Methods.
Hutchison, K. A. (2003). Is semantic priming due to association strength or feature overlap? A microanalytic review. Psychonomic Bulletin \& Review, 10, 785-813.
Hutchison, K.A., Balota, D.A., Cortese, M.J., \& Watson, J.M. (2008). Predicting semantic priming at the item level. The Quarterly Journal of Experimental Psychology, 61, 1036-1066.
Jones, L. L. (2010). Pure mediated priming: A retrospective semantic matching model. Journal of Experimental Psychology: Learning, Memory \& Cognition, 36, 135146.

Keuleers, E., Brysbaert, M., \& New, B. (2010). SUBTLEXNL: A new measure for Dutch word frequency based on film subtitles. Behavior Research Methods, 42, 643-650.
Lorch, R. F., \& Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. Journal of Experimental Psychology: Learning, Memory \& Cognition, 16, 149-157.
Lucas, M. (2000). Semantic priming without association: A meta-analytic review. Psychonomic Bulletin \& Review, 7, 618-630.
McDermott, K. B. (1997). Priming on perceptual implicit memory tests can be achieved through presentation of associates. Psychonomic Bulletin \& Review, 4, 582-586.
McNamara, T. P., \& Altarriba, J. (1988). Depth of spreading activation revisited: Semantic mediated priming occurs in lexical decisions. Journal of Memory and Language, 27, 545-559.
McRae, K., \& Boisvert, S. (1998). Automatic semantic similarity priming. Journal of Experimental Psychology: Learning, Memory \& Cognition, 24, 558-572.
Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In D. Besner \& G. W. Humphreys (Eds.), Basic processes in reading: Visual word recognition. Hillsdale, NJ: Erlbaum.
Neely, J. H., \& Keefe, D. E. (1989). Semantic context effects in visual word processing: A hybrid prospective/retrospective processing theory. In G. H. Bower (Ed.), The psychology of learning and motivation: Advances in research and theory. New York: Academic Press.
Peirce, J. (2007). PsychoPy-psychophysics software in Python. Journal of Neuroscience Methods, 162, 8-13.

Quené, H., \& van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. Journal of Memory and Language, 59, 413-425.
R development core team (2011). $R$ : A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, http://www.Rproject.org.
Roediger, H. L., III, \& Challis, B. H. (1992). Effects of exact repetition and conceptual repetition on free recall and primed word-fragment completion. Journal of Experimental Psychology: Learning, Memory \& Cognition, 18, 3-14.
Rogers, T. T., Lambon Ralph, M. A., Hodges, J. R., \& Patterson, K. (2004). Natural selection: The impact of semantic impairment on lexical and object decision. Cognitive Neuropsychology, 21, 331-352.
Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6, 461-464.
Shelton, J. R., \& Martin, R. C. (1992). How semantic is automatic semantic priming? Journal of Experimental Psychology: Learning, Memory \& Cognition, 18, 11911210.

Shulman, H. G., \& Davison, T. C. B. (1977). Control properties of semantic coding in a lexical decision task. Journal of Verbal Learning and Verbal Behavior, 16, 9198.

Stone, G. O., \& Van Orden, G. C. (1993). Strategic control of processing in word recognition. Journal of Experimental Psychology: Human Perception and Performance, 19, 744-774.
Turney, P. D., \& Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. Journal of Artificial Intelligence Research, 37, 141-188.
Weldon, M. S. (1993). The time course of perceptual and conceptual contributions to word fragment completion priming. Journal of Experimental Psychology: Learning, Memory \& Cognition, 19, 1010-1023.

# Topological Similarity of Motor Coordination in Rhythmic Movements 

Shohei Hidaka (shhidaka@jaist.ac.jp) Tsutomu Fujinami (fuji@jaist.ac.jp)<br>Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan


#### Abstract

Recognition of motion is vitally important to any animal. Vision research has proposed a number of algorithms applicable to action recognition. However, unlike successes in early visual perception, the past studies have not yet established the computational theory of action recognition. In the present study, we employ a dynamical systems approach and hypothesize that motions are encoded cognitively as a topological structure abstracted from physical particulars. We investigated whether a common topological nature could be found in a type of rhythmic movement. The topological nature of action dynamics showed a striking similarity, which could not have been identified with other analyses where physical properties were retained. The result suggests that the dynamical perspective serves as a theoretical basis in studying complex human movements.


Keywords: Actions Recognition; Motor Coordination; Dynamical Systems; Invariant Measures

## Recognition of Actions

Recognition of motion is vitally important to any animal. Detection of another animal, whether predator or prey, or a conspecific, and subsequent detailed identification of the other and how it may behave is essential to taking any emergent actions (Johnson, Bolhuis, \& Horn, 1985). Not surprisingly, our visual system is highly specialized to recognize others' actions. How do we recognize bodily movements? Our main focus is that, despite much of the advances, we still miss a parsimonious explanation of "what is an action" or a computational theory of actions. The goal of the present study is, thus, to propose a computationaltheory level description of actions which abstract identities beyond physical particulars. We briefly review psychophysical findings and theoretical works on action recognition.

The past experimental literature has explored the capacity of motion perception using point-light displays (Johansson, 1973) in which the point-lights attached to major joints are only visible in the dark background. The available information is point-wise kinematic motion of multiple body parts. Despite the limited information, people can recognize identity (Troje, Westhoff, \& Lavrov, 2005), gender (Kozlowski \& Cutting, 1977; Troje, 2002), emotions (Pollick et al., 2001; Atkinson; 2009; Hobson \& Lee, 1999), dynamics such as the weight of a lifted object (Bingham, 1987) of actions from point-light displays. Accumulating empirical studies on action perception have suggested that velocity and its higher order derivatives in single or multiple body parts characterize actions: duration of action (Pollick et al., 2001), velocity (DeMeijer, 1989), acceleration (force
or the second order time derivatives) (Chang \& Troje, 2008; 2009), jerk or the third order time derivatives (Cook, Saygin, Swain, \& Blakemore, 2009), and pairwise counter-phase oscillation (Chang \& Troje, 2008; 2009).

Consistent with these empirical findings, most of the theoretical approach employs statistical regularities among motion profiles (Hidaka, 2012). According to a recent review (Troje, 2008), perception of biological motion involves multi-level processing on local and global motion properties. Feature processing consists of four layers from early (low-level) to late (high-level) processing: life detection, structure-from-motion, action recognition, and style recognition. The system detects an autonomous agent, and constructs body structure from its detailed analysis, then is followed by more detailed action analysis.

A couple of models have been proposed for structure-from-motion and action recognition (Giese \& Poggio, 2003; Lange \& Lappe; 2006). In the model of structure-frommotion and action recognition, the model identifies body structure, and subsequently recognizes actions from the pixel-based visualization of point-light displays. In Giese \& Poggio (2003), the model was built based on neurophysiological findings in the visual cortex, and was applied to recognition of action types and action direction in normal, masked, or scrambled point-light displays.

Despite the accumulated empirical evidence and theoretical works, its computational level account attributed to Marr (1982) - a description of function, i.e., set of inputoutput pairs in action recognition - is still missing. The two models above constructed algorithms which recognize a class of actions or properties of actions through processing the features of human bodily movements. However, in general, algorithmic models formalize specific procedures, but their meaning is not often readily apparent. First, a complex model typically loses transparency of mechanism as a cost of generality (For example, multi-layered physiologically-plausible model, Giese \& Poggio, 2003). A drawback of complex models (using nonlinear filters or feature decomposition technique) is that the estimated parameters do not necessarily offer a clear interpretation on which attributes are informative in the recognition processes. Second, such a complex model often outperforms human recognition (Troje, 2002; Davis. \& Geo, 2004; Pollick \& Paterson, 2008) rather than explaining it. It is thus dubious whether those models can explain the action recognition of human beings.

We are interested in the computational level of action recognition rather than in the algorithmic level. We study actions, that is, how our multiple body parts are coordinated
in performing particular tasks. Our human body consists of over two hundreds bones, numerous muscles, and billions of neurons in the central and peripheral nerve systems controlling them with feedback loops. Obviously, making a smooth action requires integrated control across all levels of these interactive systems. Given this complexity in motor control process, it is unrealistic to compute the inverse transformation (as supposed in early vision) from end-point visible actions to its intrinsic motor control patterns. Therefore, instead of such inverse computation, we hypothesize that the goal of action recognition is to compute "dynamical invariances" under smooth transformation. This hypothesis views motor control underlying human movements as a set of dynamical systems, that is, a sequence of interactions between elements involved in controlling movements such as body joints, muscles, neural systems, etc. Our hypothesis can be best understood in the context of the dynamical system perspective on the motor coordination (Shaw et al., 1996; Turvey, 1998; Smith \& Thelen, 2003). The properties retained in dynamical systems for long term can be captured with invariant measures such as attractor dimension or Lyapnov exponent (Kantz \& Schreiber, 1997).

We define a higher dimensional space, i.e., phase space, within which all possible combinations between elements involved in controlling movements can be found. An action is then defined as a trajectory on the space. Trajectories can be projected onto lower dimensional spaces, e.g., actual movements observable from outside. In our study, we collect motion data to reconstruct the dynamical systems by embedding the time series in a higher dimensional space. For graphical examples, Figure 1 illustrates attractors, or the state space which the system may take in the three theoretical dynamical systems, the Hennon map, Rossler system, and Lorenz system (Figure 1 (a-1), (a-2), (a-3)). A univariate time series (as imperfect observation of the system) is shown in Figure 1 (b) for each of these systems. Since the original systems live in two or more dimensions, these univariate time series do not have full information due to missing dimensions. Thus, we need to "reconstruct" the phase space instead of studying the degenerated patterns. By taking the time delay vector (e.g., $\{x(t), x(t+\delta)\}$ ), the topological nature of the phase space is reconstructed (Figure 1c). In Figure 1 (c-1)-(c-3), the time-delay embedding (a map from low to high dimensional space) successfully recovers similar topological structure shown in Figure 1 (a-1)-(a-3) only from the degenerated data Figure 1 (b-1)-(b-3). Although the original phase space is unknown for empirical bodily movements, we expect the intrinsic topological nature can be reconstructed in the same way as the theoretical dynamical systems (Figure 1 (b-4) and (c-4)). See Kantz and Schreiber (1997) for a detailed description of these procedures for nonlinear time series analysis.

To study dynamical invariances, we investigated topological similarities of motor coordination. The rationale for the approach is found in observations such that one can mimic other's behaviors no matter how different their
individual appearances. Topology abstracts over physical particulars such as distance, speed, etc., to extract some dynamical invariances independent of these physical properties. Specifically, we examined the dynamical properties of rhythmic movements for two main reasons. First, rhythmic movements are not just a period but with fluctuating accents, and this is expected to show complexity to some extent neither too simple nor too complex. Second, actions which an actor can maintain continuously and produce a substantial amount of datasets are necessary for characterization of dynamical invariances.


Figure 1: Phase space of (a-1) the Hennon map, (a-2) the Rossler system, (a-3) Lorenz System, and (a-4) the body model and attached markers (filled circles: analyzed, open circles: attached but not reported in this study). (b1-3) A univariate time series from the original phase space in (a1-3) (b-4) An x-axis phase of the Shaker 1 in the expert player in the $60-\mathrm{bpm}$ trial (blue circles) with the estimated noisereduced time series (black dots). (c1-4) The reconstructed phase space from the low dimensional observed time series in (b1-4).

## Chacterizing Complex Rhythmic Actions

The data was originally obtained in order to analyze the levels of expertise in the samba music plays (Yamamoto, Ishikawa, \& Fujinami, 2006; Yamamoto \& Fujinami, 2008). The dataset consists of five players, and each player performed basic samba shaking actions in five different tempos (60, 75, 90, 105, and 120 beats per minute, and each trial lasted 97.4 seconds on average) by being cued with a metronome. While playing, three dimensional motions of 18 markers, attached on body parts and musical instruments, were recorded at 86.1 Hz of sampling rate (Figure 1a-4). As well as the original study, here we aim to find the relationship between dynamical properties among bodily actions. For simplicity, we limited ourselves to analyze a subset of the original datasets, 3190 samples (74.1 seconds long) of four markers attached on the right wrist, right elbow, and two sides of the musical instrument (shaker), having the right shoulder as a reference point (Figure 1a-4). These were the essential parts of the samba actions making sounds directly, and we
expected that dynamic coordination among them would be crucial to characterize the dynamical properties of the samba.

## Preprocess and phase space reconstruction

In the analysis, after down-sampling the original data to 46.05 Hz , the first 250 samples ( 5.81 second long from the beginning of the recording) were excluded as initial setup of the actions, and 3250 samples ( 75.5 second long) of velocities were analyzed for each subject. In order to reduce measurement noise, for each movement of the markers, the local linear projective method was performed after phase space reconstruction of each time series on the 31 dimensional time delay space with 46 msec (i.e., $\{\mathrm{t}, \mathrm{t}+\Delta \mathrm{t}, \mathrm{t}$ $+2 \Delta \mathrm{t}, \ldots, \mathrm{t}+30 \Delta \mathrm{t}\}$ where $\Delta \mathrm{t}=46 \mathrm{msec}$ ) (Takens, 1981). This technique is a nonparametric and unsupervised method which, in principle, reduces observation noise independent of the time series intrinsically generated from a nonlinear dynamical system. Figure 1 (b-4) shows the original data (open circles) and its noise-reduced data (filled circles) after applying the local linear projective method. Due to digitalization in the motion capture system, the original data only takes certain discrete values which may be potential sources of observational noise in the measurement. As the result of the noise reduction, we obtained the 31 dimensional phase space of 3220 points for each coordinate of three dimensional positions of each marker movement in each subject and trial. An example of the reconstructed phase space is shown in Figure 1 (c-4).

## Estimating symbolic dynamics

For each estimated phase space (Figure 1c-4 and Figure 3), a symbolic dynamic is estimated by the symbolic false nearest neighbor method (Buhl \& Kennel, 2005; See also Hidaka \& Yu, 2010). In this algorithm, a symbolic state is assigned to each data point in the given phase space by minimizing the error in the one-to-one correspondence between spatial nearest neighborhood and symbolic nearest neighborhood. After convergence of the iterative minimization procedure, we obtained the series of binary symbolic states for the trajectory in the high dimensional phase space. Each symbolic series of length 5 (e.g., the subsequence " 01100 " as a state) is reported as a state in the present study, but we found the similar results consistently with the symbol length from 5 to 8 . It means that the state transition in each phase space is analyzed by partitioninig into 32 (i.e. $2^{5}$ ) distinct states.

## Results

In order to see the rhythmic properties as phase shifts in repeating actions, we analyzed the temporal profiles of the velocities in the right arm and wrist. Figure 2 shows the histogram of phase differences between body parts with the right shoulder as the reference point. Since the right elbow and wrist are the major body parts playing the shaker, their temporal structure was expected to reflect the rhythmic charateristics. However, not as expected, the five musicians showed quite different distributions in terms of phase shifts among body parts, even for the right wrist and elbow
movement to play the shakers to the same tempo. The peaks found for the elbow and wrist are sharp for musicians A and D, compared with those found in the other musicians. As for A, the peaks of the elbow and wrist come to the same phase, but the peak is less visible for the wrist. As for D , the peaks of the elbow and wrist come later than that of the shoulder. For the other musicians, no obvious feature is found. The frequency uniformly varying over the phase angle shows large fluctuations in arm movements for each musician. The histograms revealed both within-musician fluctuations and individual differences rather than similarity among actions. The results suggest that charaterization of the "same" action (i.e., playing to the samba rhythm) on the levels of physical properties may lead quite different patterns across subjects. Needless to say, changing physical properties such as tempos also directly changes phase differences. The level of physical properties is not sufficient for characterizing actions even if the major parameter of the actions (i.e., tempo) is well controlled.

Next, we analyzed the properties of actions by looking into the dynamical systems underneath body movements. A basic technique to characterize dynamical properties from an empirical time series is phase space reconstruction. A phase space reconstructed by time-delay embedding is visualied as a three dimensional subspace projection (Figure $1 \mathrm{c}-4$ ). The phase space is originally a set of velocity vectors of the four markers including two sides of the shakers, right wrist and elbow. The trajectory on the reconstructed phase space shows an attractor or the state space the system may take. The phase space is 124 dimensional space consisting of 31 time-delay copies of the four dimensional time series. First we analyzed the dimensionality of the attractors as one of invariance for the dynamical system. It is formally measured by correlation dimensions (Kantz \& Schreiber, 1997), and we found the correlation dimensions varying from 1.8 to 2.4 across five musicians and five conditions. These results suggested the state space of the samba rhythm is rather restricted on a low dimensional space.

Since the dimensions of the attractors are lower than three, it allows us to visualize them in the three dimensional space without losing much information. Figure 3 shows the attractors estimated for all five musicans on the five conditions. Visual inspection of the samba attractors grasps the gist of commonalities among the attractors. Consistently across most of the attractors, they share a similar shape of trajectories - a twisted double circle (which may appear different due to a specific visual angle of each attractor). These similar "shapes" of trajectories indicate that the topological nature of the attractors is similar.

In order to quantify the similarity among attractors, we performed analysis based on symbolic dynamics (Buhl \& Kennel, 2005). Symbolic dynamics offers a way to analyze a topological property of a dynamical system by constructing a homeomorphism (map preserving the topological nature) from the original space to a symbolic space. If two state spaces are homeomorphic (topologically identical), we find identity between their symbolic dynamics
as well. Importantly, it is easier to compute similarity between two symbolic spaces than that between two highdimensional phase spaces. This mathematical property is applied in our data analysis. By estimating the symbolic dynamics for each of the samba attractors, we estimated the probability of state space transition on the estimated symbolic space (Figure 4). Each of the top left two-by-two panels shows a probabilistic distribution (stationary distribution) of the symbolic states for each musician and condition. Each symbolic state is defined by a binary symbol series of length 5 (e.g., " 01100 "), and the distribution over its 32 possible states is shown in the figure.

In Figure 4, we show the analysis concerning a pair of musicians and conditions as representative results. Each distribution of symbolic states reflects the topological nature of the underlying dynamical systems - transitivity or connectivity of states in the system. As shown in Figure 4, the distribtions of state spaces are quite similar across subjects ( $\mathrm{R}=0.784$ and $\mathrm{R}=0.962$ ) shown in the panels of Figure 3 and across conditions ( $\mathrm{R}=0.959$ and $\mathrm{R}=0.803$ ) shown in the panels in Figure 2. Across all the pairs of five subjects and five conditions, the average correlation of distiribution is 0.831 . This means that the topological nature of the complex actions as dynamical systems were quite robust across different musicians and different playing tempos.

As a baseline comparison, we performed cross correlation analysis of the physical movements of the shaker (the major axis of the 3D motions with the most variance). By taking a time lag maximizing correlation, the average cross correlation was 0.453 across musicians and trials ( 0.462 across musicians (the same tempo) and 0.478 across tempos (the same musicians)). These results showed the temporal correlation of the shaker movements were not as high as the correlation of the symbolic state space ( 0.831 ) even with the time lag optimally adjusted and the playing tempo being held constant. This result means that the similarity of movements as physical properties cannot explain the topological similarities shown above.


Figure 2: The distribution of phase shift (radian) of right elbow (gray) and right wrist (white).


Figure 3: The reconstructed phase space embedding in three dimensional time delay space in Musician A-E playing at tempo $60,75,90,105$, and 120 BPM. The velocity of the trajectory in the three dimensional space is shown as RGB color code for its visibility.


Figure 4: Stationary distributions of states in each of symbolic dynamics (four top left panels) and their correlations between musicians and conditions (four right side and bottom panels).

The result -- higher similarity between topolgoical properties of the state space -- is quite surprising with consideration to the individual differences on the physical level charateristics (Figure 2) and low cross correlations in the physical movements. The results suggest that the topological properties of the attractors were quite similar across different musicians and tempos, while their physical realizations of the actions differed from person to person.

## Discussion

One of the challenges to the theory of action recognition is formalizing the possible attributes of characteristic actions. In the present study, we hypothesize that an intrinsic topological nature of actions as dynamical systems
characterizes a similarity between actions. This is meant to describe actions on the basis of invariances under nonlinear transformations, rather than the specific features (coordinate systems) the actions have. In other words, this is to abstract the actions from their physical properties. In order to test the hypothesis, we investigated the samba playing action, which is repetitive rhythmic movement. The samba rhythm fluctuates within a certain range, showing a complex accent pattern, even if an auditory cue is given to keep the tempo constant (Figure 2). No common property was found in movements across musicians in physical movements of the right arm.

Subsequently, we analyzed the same data from a perspective of actions as dynamical systems. In the analysis, we define a higher dimensional space in which the action is mapped as a trajectory. We analyzed time-delay vectors, which embed the time series of movements, i.e., lower dimensional data, into higher dimensional space.

The analyses revealed topological similarities in the reconstructed phase space among the musicians and among the different playing conditions. The analyses using the symbolic dynamics quantified the similarities in terms of their topological structures. In sum, these results supported our hypothesis that human actions can be characterized on the basis of invariances as dynamical systems. This invariant nature of the dynamical property can serve as a possible basis for our perception of actions, and offers an explanation of why we perceive them as "the same actions." Interestingly, the patterns revealed in the current analysis (Figure 3) are not just abstract-level depiction, but they also correspond with the introspective view of the samba rhythm (Figure 5: obtained from the most experienced musician A in the post-experiment interview). His drawing represents a general periodic motion with accents at a particular part of the trajectory. The geometric shape of the trajectory closely corresponds with the reconstructed phase space (Figure 3).

Finally, we briefly mention the implications to the two relevant research fields. One is the imitation of bodily movements. It is necessary to map from the body of the self to the body of the other in order to imitate the other's actions (Breazeal \& Scassellati, 2002). One of the traditional approaches to this problem is to compute inverse kinematics (estimating the motor control parameters from the perceived actions) upon which a number of robotics applications have been based (for example, Wolpert, Doya, \& Kawato, 2003). This approach, however, does not offer a sufficient explanation for neonatal imitations (Melzoff \& Moore, 1977) and how actions are identified, because neonate do not have opportunity to learn the cross-modal identity of the action (i.e., visual patterns of the other's action and motor control of the self). In turn, topological similarity of actions revealed in the present study may potentially offer a cross-body identity of actions regardless of their differences in physical particulars.
The present study proposes the dynamical perspective of actions in which it is essential to characterize topological similarities of actions as attractors. It is viewed as a
paradigm shift from cognition as inverse computation for an ill-posed problem to the computation of invariances under smooth transformations.


Figure 5: A drawing by the expert in his introspective explanation of the samba rhythm.

## Acknowledgments

The author is grateful to Nathan T. Nossal for his kind advice for the present manuscript. This study was supported by Artificial Intelligence Research Promotion Foundation, a grant of the NeuroCreative Lab (NPO), Grant-in-Aid for Scientific Research B KAKENHI No. 23300099, and Grant-in-Aid for Exploratory Research 25560297.

## References

Atkinson, A. P. (2009). Impaired recognition of emotions from body movements is associated with elevated motion coherence thresholds in autism spectrum disorders. Neuropsychologia, 47, 3023-3029.
Bingham, G. P. (1987). Kinematic form and scaling: Further investigations on the visual perception of lifted weight. Journal of Experimental Psychology: Human Perception and Performance, 13, 2, 155-177,
Blake, R. \& Shiffrar, M., (2007). Perception of Human Motion. Annual Review of Psychology, 58, 47-73.
Breazeal, C. \& Scassellati, B. (2002). Robots that imitate humans., Trends in Cognitive Sciences, 6 (11), 481-487.
Buhl, M. and Kennel, M. B. (2005). Statistically relaxing to generating partitions for observed time-series data., Phys. Rev. E 71, 046213.
Chang, D. H. F., \& Troje, N. F. (2008). Perception of animacy and direction from local biological motion signals. Journal of Vision, 8, (5):3, 1-10.
Chang, D. H. F., \& Troje, N. F. (2009). Acceleration carries the local inversion effect in biological motion perception. Journal of Vision, 9, (1):19, 1-17
Cook, J., Saygin, A. P., Swain, R., \& Blakemore, S-H., (2009). Reduced sensitivity to minimum-jerk biological motion in autism spectrum conditions. Neuropsychologia, 47, 14, 3275-3278.
DeMeijer, M, (1989). The contribution of general features of body movement to the attribution of emotions. Journal of Nonverbal Behavior, 13, 4, 247-268.

Giese, M. A. \& Poggio, T. (2003). Neural Mechanisms for the recognition of biological movements, Nature Reviews Neuroscience, 4, 179-192.
Hidaka, S. (2012) Identifying Kinematic Cues for Action Style Recognition. In Proceedings of The Thirty Fourth Annual Meeting of Cognitive Science Society, 1679-1684.
Hidaka, S., \& Yu, C. (2010). Spatio-temporal symbolization of multidimensional time series. In ICDM workshops (p. 249-256).
Hobson, R. P. \& Lee, A. (1999). Imitation and Identification in Autism, Journal of Child Psychological Psychiatry, 40, 4, 649-659.
Hubert, B., Wicker, B., Moore, D. G., Monfardini, E., Duverger, H., Fonse'ca, D. Da, Deruelle, C. (2006). Recognition of Emotional and Non-emotional Biological Motion in Individuals with Autistic Spectrum Disorders. Journal of Autism Developmental Disorders, 37, 7, 13861392.

Johanson, G. (1973). Visual perception of biological motion and a model for its analysis. Perception \& Psychophysics, 14, 2, 201-211.
Johnson, M. H., Bolhuis, J. J., \& Horn, G. (1985). Interaction between acquired preferences and developing predispositions during imprinting. Animal Behaviour, 33, 1000-1006.
Kantz, H., \& Schreiber, T. (1997). Nonlinear time series analysis. Cambridge, UK: Cambridge University Press.
Lange, J., \& Lappe, M. (2006). A model of biological motion perception from configural form cues. Journal of Neuroscience, 26, 11, 2894-2906.
Metlzoff, A. N. \& Moore, M. K. (1977). Imitation of Facial and Manual Gestures by Human Neonates, Science, 198, 75-78.
Moore, Hobson, \& Lee (1997). Components of person perception: An investigation with autistic, non-autistic retarded and typically developing children and adolescents., British Journal of Developmental Psychology, 15, 401-423.
Pollick, F. E., Paterson, H. M., Bruderlin, A., Sanford, A. J., (2001). Perceiving affect from arm movement. Cognition, 82, B51-B61.
Pollick, F. E., Lestou, V., Ryu J. Cho, S-B. (2002) Estimating the efficiency of recognizing gender and affect from biological motion., Vision Research, 42, 2345-2355.
Pollick, F. E., Paterson, H., Bruderlin, A. \& Sanford, A. J. (2001) Perceiving affect from arm movement. Cognition, 82, B51-B61.
Pollick F. E., Paterson, E. (2008). Movement style, Movement features, and the recognition of affect from human motion, In Shipley, T. F. \& Zacks, J. M., Understanding Events from Perception to Action, New York: Oxford University Press, 286-307.
Shaw, R. E., Flascher, O. M., \& Mace, W. M. (1996). Dimensions of event perception. Handbook of perception and action, 1, 345-395.

Smith, L. B., \& Thelen, E. (2003). Development as a dynamic system. Trends in cognitive sciences, 7(8), 343348.

Takens, F. (1981). Detecting strange attractors in turbulence., In D. A. Rand and L.-S. Young. Dynamical Systems and Turbulence, Lecture Notes in Mathematics, vol. 898. Springer-Verlag. pp. 366-381.
Troje, N. F. (2002). Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. Journal of Vision, 2, 371-387.
Troje, N. F. (2008). Biological motion perception. In Basbaum, A. et al. (Eds.), The senses: A comprehensive reference (pp. 231-238). Oxford: Elsevier.
Troje, N. F., Westhoff, C., \& Lavrov, M. (2005). Person identification from biological motion: effects of structural and kinematic cues. Perception \& Psychophysics, 67 (4), 667-675.
Turvey, M. T. (1998). Dynamics of effortful touch and interlimb coordination. Journal of Biomechanics, 31(10), 873-882.
Wolpert, D. M., Doya, K., \& Kawato, M. (2003). A unifying computational framework for motor control and social interaction. Phil. Trans. R. Soc. Lond. B, 358, 593602.

Yamamoto, T., \& Fujinami, T. (2008). Hierarchical organization of the coordinative structure of the skill of clay kneading. Human Movement Science, 27(5), 812-822.
Yamamoto, Y., Ishikawa, K., \& Fujinami, T. (2006). Developmental stages of musical skill of samba. Journal of biomechanics, 39, S555.

# Eye Movements Reveal Interplay Between Noun Capitalization and Word Class During Reading 

Sven Hohenstein (sven.hohenstein@uni-potsdam.de)<br>Department of Psychology, University of Potsdam,<br>Karl-Liebknecht-Str. 24/25, 14476 Potsdam, Germany

Reinhold Kliegl (kliegl@uni-potsdam.de)
Department of Psychology, University of Potsdam, Karl-Liebknecht-Str. 24/25, 14476 Potsdam, Germany


#### Abstract

Subjects' eye movements were recorded while they read sentences for comprehension. Sentences were presented with capitalized nouns-in agreement with German spelling rules-or completely in lowercase. Overall reading speed was not influenced by the manipulation of capitalization, but fixation durations were affected by the interplay between capitalization and the word classes of the fixated and the succeeding word. As expected, fixations were shorter for capitalized than lowercase nouns, but unexpectedly they were longer when the upcoming word was also a noun. This modulation was reduced when all words were printed completely in lowercase. We interpret the results as evidence for distributed processing across several words.


Keywords: eye movements; reading; corpus analysis; capitalization; parafoveal processing

## Introduction

The uptake of visual information is critical for the process of reading. Most visual receptors are located in the foveathe central $2^{\circ}$ of the visual field. As distance from the fixation location (and thereby from the fovea) increases, acuity decreases across the parafoveal region. Hence, it is necessary to move our eyes to obtain visual information of the words in a sentence. Foveal information of the currently fixated word is most essential for word processing. There is much evidence for the effect of, among others, orthographic, phonological, and semantic features of the currently fixated, foveal word as well as of the not-yet fixated, parafoveal word on fixation durations. Some of the parafoveal effects are still under dispute (for a recent review see Schotter, Angele, \& Rayner, 2012). The resolution of such disputes is critical for our understanding of eye-movement control during reading, because arguably parafoveal word processing is the primary source of information for guiding the eyes to the next fixation location.

## Capitalization

Although many studies have demonstrated the relevance of low-level orthographic features on eye movements during reading, little is known about the role of capitalization. In languages based on the Roman alphabet, the first words of a sentence and proper nouns are capitalized, that is these words are spelled with an initial capital letter. Additionally, German script has an unusual characteristic: the capitalizat-
ion of all nouns. Therefore, the frequency of capitalized words in German texts is relatively high, making it an ideal language for studying the impact of initial capital letters during natural reading.

Capitalized characters in parafoveal vision may be salient and attract attention to the preview word. Furthermore, German capitalization may reduce the cost of lexical processing. From the first letter alone, readers of German script obtain the word-class information (i.e., whether the next word is a noun vs. non-noun). Deeper lexical (e.g., semantic) processing of the word may start faster than in other languages because of the early availability of word-class information. Preprocessing of parafoveal words appears to be more likely in German than in less transparent orthographies (Laubrock \& Hohenstein, 2012).

There is research demonstrating a beneficial influence of capitalization on reading rate in German (Bock, 1989; Bock, 1990; Bock, Augst, \& Wegner, 1985; Bock, Hagenschneider, \& Schweer, 1989; Gfroerer, Günther, \& Bock, 1989). When uppercase and lowercase letters were used improperly, the reading rate was lower than when the German capitalization rules were observed (Bock et al., 1985). When Dutch subjects were asked to read German texts, they showed the same pattern of reading speed despite the absence of capitalization in Dutch (Bock et al., 1989). These effects were also reflected in eye-movement measures (Gfroerer et al., 1989). Furthermore, German subjects' reading rates for English texts were the same whether the rules of capitalization applied were based on German or English spelling (Bock et al., 1989), indicating that German readers transferred their familiar capitalization rules to text in a different language.

The advantage of capitalization increases with reading skill (Bock, 1990). For $10^{\text {th }}$-grade students, the effects of violations of German capitalization rules are similar to those obtained with adult readers. The same pattern was also present for $7^{\text {th }}$-grade children, but differences were less pronounced. Violation of the capitalization rule had no reliable effect on the reading speed of $3^{\text {rd }}$-grade children. In addition, on the basis of several experiments, Bock (1989) argued that the function of German capitalization rules for reading is independent of word shape, and that they allow differentiating between nouns and non-nouns without analyzing a word's meaning.

There is also evidence for the importance of capitalization in Italian. In Italian, like in most languages based on the Roman alphabet, proper names are spelled with an initial capital letter, whereas common nouns are spelled in lowercase. Peressotti, Cubelli, and Job (2003) demonstrated that reaction times in a lexical-decision task are reduced if proper names are presented capitalized compared to noncapitalized presentation and common nouns. The authors provide an orthographic cue hypothesis that an initial uppercase letter helps pre-activate lexical units corresponding to proper names.

Furthermore, Müsseler, Nisslein, and Koriat (2005) reported an influence of German capitalization rules on the missing-letter effect. Typically, when subjects are asked to underline a certain target letter, the letter is more difficult to detect in function words than in content words. In their study, this missing-letter effect was eliminated when function words within a sentence were capitalized indicating that unfamiliar orthography facilitates the extraction of single letters.

Finally, nouns are best recognized when they are presented with an initial uppercase letter (Jacobs, Nuerk, Graf, Braun, \& Nazir, 2008). In this study, subjects had to type single words that were presented briefly ( 50 ms ) and then followed by a mask. Words were presented with all letters in lowercase, with all letters in uppercase, or with an initial uppercase letter. Accuracy for non-nouns was best when they were presented with an initial uppercase letter or completely in lowercase (both types of presentation are common in German texts due to the capitalization of words at the beginning of a sentence).

Hohenstein and Kliegl (in press; Experiment 2) manipulated capitalization in an eye-movement study of semantic processing of parafoveal words. They asked subjects to read single German sentences for comprehension. Employing the gaze-contingent boundary paradigm (Rayner, 1975), the parafoveal preview for a critical target noun (wool) was semantically either related (silk) or unrelated (soap) to the target and was replaced with the target word during the saccade to the target location. Fixation durations on the target noun were shorter for related compared to unrelated previews, indicating parafoveal semantic information extraction. Most importantly, this effect was not modulated by capitalization: There was no significant difference between sentences presented completely in lowercase and sentences presented in agreement with the German capitalization rules (i.e., with capitalized nouns).

Hohenstein and Kliegl (in press) also analyzed the effect of capitalization on fixation durations on the target word, which was always a noun, and on the pretarget word, which was any of a number of different parts of speech, but never a noun. Whereas the violation of German noun-capitalization was reflected in longer target fixations, the pretarget was fixated more briefly if the sentence was presented completely in lowercase. This inverse effect of capitalization on pretargets and targets could not be explained by either (1) a general reading-speed benefit associated with a presentation
in which nouns are capitalized-a common result of past studies-or (2) a unique effect of capitalization on reading nouns, but not words which are always presented in lowercase. We hypothesize that the obtained effect is due to an interplay of capitalization and the word classes (nouns/nonnouns) of fixated and parafoveal words.

Here we report an eye-movement corpus analysis of German sentences, which were presented either completely in lowercase or with capitalized nouns (following the German spelling rules). The focus will be on the interaction between the word class of the preceding, current, and upcoming word and capitalization.

## Method

This is a reanalysis of the study of Hohenstein and Kliegl, (in press; Experiment 2). We describe the main features of the experiment, judged to be relevant for an appreciation of the present article; for further technical details we refer to the original article.

## Subjects

Thirty-two subjects ( 20 women, 12 men ) participated in the experiment. Their age was between 16 and $39(M=23$, SD $=4.8$ ). All were native speakers of German with normal or corrected-to-normal vision.

## Apparatus

Sentences were displayed on a single line at midscreen height on a 21 -inch monitor. Subjects were seated 24 inches in front of the screen. Sentences were presented in black, boldface, 20-point Courier New font on a white background. Each character was 12 pixels wide- $0.45^{\circ}$ of visual angleat a screen resolution of $1024 \times 768$ pixels. The refresh rate of the monitor was 150 Hz .

Eyes were monitored using an EyeLink II system with a sampling rate of 500 Hz , an instrumental spatial resolution of $0.01^{\circ}$, and an average accuracy of better than $0.5^{\circ}$. Recording was binocular. Heads were positioned on a chin rest to minimize head movements.

## Material

We used the material developed by Hohenstein, Laubrock, and Kliegl (2010). All sentences were constructed around a critical target region and ranged from six to thirteen words. In 100 sentences, the target word was a noun and the preceding word (i.e., the pretarget) was a non-noun. In 24 additional sentences, targets were non-nouns. The sentences did not include any punctuation except the period at the end. Word lengths ranged from two to eighteen characters.

Sentences were presented in two conditions: capitalized and non-capitalized. In the capitalized condition, the sentence's first word and all nouns were spelled with an initial capital letter. In the non-capitalized condition, all words were spelled in lowercase.

## Procedure

Subjects were naive concerning the purpose of the experiment. They were instructed to read single sentences for comprehension. A random sample of one third of the sentences was followed by a three-alternative multiple-choice question that was answered by clicking on one of the response alternatives. Ninety-five percent of all questions were answered correctly, indicating no serious comprehension problems.

At the beginning of the experiment, subjects were instructed that the experiment consisted of two parts. One part of the sentences was presented following the German capitalization rules and the other was presented completely in lowercase. Each part comprised 62 sentences, preceded by six practice sentences, which were not included in the analyses. Subjects were informed about the start of the second part.

When a sentence was initially presented, the preview (related or unrelated) occupied the target location. An invisible boundary located directly after the last letter of the pretarget word was present in each sentence. When either eye crossed the boundary, the preview word on the target position was replaced with the target word. The sentence remained in this final form until the end of the trial. The manipulation of the preview has no relevance for the present study.

## Measures and Selection Criteria

Data from sentences with a blink or loss of measurement was used only until the point in time preceding the first loss and only if the loss occurred after the target region. Saccades were detected with a binocular velocity-based algorithm (Engbert \& Kliegl, 2003; Engbert \& Mergenthaler, 2006). Analyses were based on right-eye fixations.

This first level of screening led to a pool of 39,646 fixations. In a second level of data screening, we excluded the first and last fixations in sentences $(7,830)$ and fixations on the first or last words of sentences $(9,608)$. We used firstpass fixations only (i.e., excluding 5,930 fixations). This second level of screening left us with 24,302 valid withinsentence reading fixations. Our selection procedure is similar to data filtering in a large eye-movement corpus study (Kliegl, Nuthmann, \& Engbert, 2006). We included all nontraining trials. Results were not affected by the exclusion of trials with subsequent comprehension errors.

We computed gaze durations (the sum of all first-pass fixations), first-fixation durations, and single-fixation durations (for a definition of these measures, see Inhoff \& Radach, 1998) for each word. Additional measures included refixation probability and relative landing position (i.e., the position of the first fixation). Furthermore, we calculated reading speed (words per minute) for each sentence.

## Statistical analysis

Inferential statistics for effects on fixation durations are based on linear mixed models (LMMs) specifying subjects and sentences as crossed random factors (Baayen, Davidson, \& Bates, 2008; Kliegl, Masson, \& Richter, 2010). LMMs
are much more resilient to data loss than the classical analysis of variance. Thus, these analyses are very powerful even for datasets with differences in the number of observations between subjects and items. Effects in models with continuous dependent variables were estimated with the lme 4 package (Bates, Mächler, \& Bolker, 2012) in the $R$ environment for statistical computing (version 2.15.2, 64-bit build; R Development Core Team, 2012). LMMs were fitted using the restricted maximum likelihood method.

We specified varying intercepts for both subjects and items. Furthermore, we included varying slopes associated with the effect of capitalization-the only experimental factor in the present study. The full model including variance components for all terms of the experimental design is preferred for statistical analyses (Schielzeth \& Forstmeier, 2009). Coverage probability of confidence intervals associated with fixed effects is better for LMMs including random slopes than for models including intercepts only (Schielzeth \& Forstmeier, 2009; Barr, Levy, Scheepers, Tily, 2013). With the additional inclusion of correlation parameters, models did not longer converge. Hence, these parameters were excluded. In a recent simulation study, Barr et al. (2013) demonstrated that models without random correlations are very similar to full models with respect to coverage probability and power. The authors rank both kinds of models in the first position of desirable model designs.

The primary fixed effects in the analyses were capitalization and the interactions between capitalization and the word classes of the preceding, the current, and the next word. Following the work of Kliegl et al. (2006), we included several additional covariates (word frequency, word length, saccade length, second-order polynomial of relative landing position), which have an influence on fixation durations in reading. The inclusion of these covariates reduced potential confounding for the word class predictors.

Continuous predictors were centered at their mean; relative landing position was centered at .5 ; the factor capitalization (capitalized presentation vs. non-capitalized presentation) entered the analysis as treatment contrast with capitalization as reference category ( 0 vs. 1 ) and the factors for word class (non-noun vs. noun) were specified as sum contrasts ( -0.5 vs. +0.5 ).

We report regression coefficients together with $p$ values based on Markov chain Monte Carlo sampling with 10,000 samples (Baayen et al., 2008). Based on analyses of model residuals, we decided to use the (natural) logarithm of all fixation-duration measures.

## Results

## Global Reading Speed

The analysis of reading speed tested only the effect of capitalization. Mean reading speed was 208 and 207 words per minute in the capitalized and non-capitalized condition, respectively. This difference was not significant ( $\beta=0.85 ; p$ $=.78$ ). The result differs from findings in earlier studies, in
which subjects read reliably slower if capitalization rules were not followed (e.g., Bock et al., 1985).

## Fixation Durations

The analysis of gaze duration ( 20,768 observations), firstfixation $(20,816)$, and single-fixation duration $(17,775)$ revealed the same pattern of effects. We present a detailed analysis of gaze durations only. Table 1 displays the results of an LMM analysis with gaze duration as dependent variable. There were several significant effects associated with word frequency, word length, landing position, and saccade length (see Table 1 for details).

The violation of German capitalization rules did not significantly influence gaze duration as a main effect, but there were several significant interactions associated with it. If the capitalization rules were followed, nouns were fixated more briefly than non-nouns. Interestingly, the word classes of the previous and the next word also had an effect on gaze duration: If the previous word was a noun, fixation time was shorter; if the next word was a noun, gaze duration was longer.

Most importantly, the interaction between capitalization and word class was significant both for the current and the next word. Figure 1 displays these interactions. As expected, a noun was fixated longer if the presentation was non-capitalized and thus with unfamiliar orthography. This trend was reversed with respect to the word class of the word to the right: Fixation duration was reduced if the next word was a noun and presented completely in lowercase. As is apparent from Figure 1, compared to reading under normal German capitalization conditions, fixation durations in sentences presented completely in lowercase are less modulated by properties of the fixated and surrounding text. These interactions were also significant for either first-fixation duration or single-fixation duration as dependent variable.

## Additional Analyses of Reading Behavior

There was a small effect of capitalization on saccade length: Saccades were shorter by 0.157 characters $(p=.003)$ if all words were presented in lowercase compared to the presentation following the German capitalization rules (means values: 6.20 vs. 6.35 ). The interactions between capitalization and words class had no significant effect on saccade length (all $p \mathrm{~s}>.07$ ). Initial landing position, skipping probability, regression probability, and refixation probability were not significantly influenced by capitalization or the interactions between capitalization and word class (all $p$ s $>.18$ ).

## Discussion

In the present study, subjects were asked to read sentences for comprehension. Sentences were presented in one of two capitalization conditions: In the capitalized condition, all nouns were spelled with an initial capital letter (following the German spelling rules); in the non-capitalized condition, all words were presented completely in lowercase.

Table 1: Estimates (regression coefficients) with associated standard errors and $p$-values as well as random-effect variances of a linear mixed-model corpus-analysis with log gaze duration as dependent variable.

| Fixed effects | Dependent variable: log gaze duration |  |  |
| :---: | :---: | :---: | :---: |
|  | Estimate | SE | $p$ |
| (Intercept) | 5.43494 | 0.026 | <. 001 |
| $\log _{10}$ frequency |  |  |  |
| Previous word | -0.02305 | 0.003 | <. 001 |
| Current word | -0.07392 | 0.003 | <. 001 |
| Next word | -0.00005 | <0.001 | . 100 |
| Length ${ }^{-1}$ |  |  |  |
| Previous word | -0.03763 | 0.039 | . 334 |
| Current word | -0.41590 | 0.048 | $<.001$ |
| Next word | 0.01321 | 0.031 | . 667 |
| Relative landing position |  |  |  |
| Linear trend | -6.78905 | 0.400 | <. 001 |
| Quadratic trend | 3.39784 | 0.400 | <. 001 |
| Saccade length |  |  |  |
| Incoming | 0.02457 | 0.001 | <. 001 |
| Outgoing | -0.02955 | 0.001 | <. 001 |
| Capitalization (present vs. absent) | -0.00232 | 0.010 | . 815 |
| Word class <br> (non-noun vs. noun) |  |  |  |
| Previous word | -0.02092 | 0.010 | . 042 |
| Current word | -0.07053 | 0.011 | <. 001 |
| Next word | 0.09283 | 0.009 | <. 001 |
| Capitalization $\times$ word class |  |  |  |
| Previous word | 0.00183 | 0.013 | . 883 |
| Current word | 0.05577 | 0.014 | <. 001 |
| Next word | -0.07686 | 0.013 | $<.001$ |
| Random effects | Variance |  |  |
| Sentences |  |  |  |
| (Intercept) | 0.00154 |  |  |
| Capitalization | 0.00000 |  |  |
| Subjects |  |  |  |
| (Intercept) | 0.02003 |  |  |
| Capitalization | 0.00129 |  |  |
| Residual | 0.13397 |  |  |

Note. All continuous predictors were centred. Frequencies were $\log _{10}$ transformed. For length, the reciprocal was employed. "Current", "previous", and "next" indicate variables associated with the fixated word, the preceding word, and the succeeding word, respectively. The $\times$ symbol indicates an interaction.

Neither fixation times nor reading speed was significantly affected by the violation of capitalization. This result is in contrast to findings in earlier studies (Bock, 1989; 1990; Bock et al., 1985; 1989). We hypothesize the different result is due to changes in communication technology over the last 20 years. The earlier studies were conducted in the 1980s, a time in which text messaging, e-mail, and virtual chat rooms


Figure 1: Gaze duration on the current word as a function of capitalization and word class of the previous, the current, and the next word. Error bars indicate standard errors. All random effects and all fixed effects except the intercept, capitalization, and word class (including interactions) were removed from the data. The removal of effects was applied to each panel separately.
were less widely used than they are today. The use of short text messages with cellphones, in particular, is easier when all words are typed in lowercase. Indeed Schloblinski et al. (2001) found that most German text messages by students do not follow the German capitalization rules. Furthermore, texts in lowercase could be found in e-mails and social networks too (e.g., Schnitzer, 2012). In the present study, participants were young adults. Hence, we suppose our subjects were more proficient in reading text without capitalization than subjects in the earlier studies.

The most important findings are the interactions between capitalization and the word classes of the fixated and the next word. We demonstrated that the effects in a critical target region (Hohenstein \& Kliegl, in press) generalize to reading behavior throughout sentences. If German capitalization rules were followed, nouns were fixated more briefly than non-nouns. Interestingly, the effect was reversed for the upcoming word. We observed an effect of the parafoveal word on the current fixation duration. Parafoveal-on-foveal effects are an indicator of distributed processing of words
during reading (e.g., Kliegl et al., 2006). The case of a word's first character is easy to extract due to the saliency of an uppercase letter in a string of lowercase letters. The direction of the effects could be explained by a segmentation hypothesis: The words preceding nouns are often highly associated with those nouns (e.g., articles, adjectives). Perhaps, when a non-noun is fixated and the next word is a noun, both words are processed during the fixation on the non-noun (the word-group hypothesis; Kliegl, 2007; Radach, 1996). This results in a slowdown if the next word is a noun. Once the noun is fixated, it has already been preprocessed and hence fixation time is reduced.

When all words were presented in lowercase, the effects of word class were significantly reduced. Reading was less modulated by the word class and appeared to be more homogeneous. Although reading speed is comparable in both capitalization conditions, the reading strategy is different because salient orthographic cues are missing when capital letters are not present. This finding is evidence for the importance of parafoveal processing in reading.

In summary, the results reveal an impact of orthographic and visual word features on distributed processing during reading. Furthermore, readers are very flexible in adapting to different reading situations. That said, our results do not generalize to older readers who might encounter problems in the unfamiliar reading situation. Besides age differences, future research should focus on how non-native speakers of German use capitalization cues for their reading strategies.

## Acknowledgments

This research was supported by Deutsche Forschungsgemeinschaft (grants no. KL 955/6-1) as part of Research Group 868 "Computational Modeling of Behavioral, Cognitive, and Neural Dynamics".

## References

Baayen, R. H., Davidson, D. J., \& Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language, 59, 390-412.
Barr, D. J., Levy, R., Scheepers, C., \& Tily, H. J. (2013) Random effects structure in mixed-effects models: Keep it maximal. Journal of Memory and Language, 68, 255278.

Bates, D. M., Mächler, M., \& Bolker, B. (2012). lme4: Linear mixed-effects models using S4 classes (R package version 0.999999-0) [Computer software]. Available at http://CRAN.R-project.org/web/packages/lme4/
Bock, M. (1989). Lesen in Abhängigkeit von der Groß- und Kleinschreibung. [Reading depending on capital and small letters]. Sprache \& Kognition, 8, 133-151.
Bock, M. (1990). Zur Funktion der deutschen Groß-und Kleinschreibung. Einflüsse von Wortform, Muttersprache, Lesealter, Legasthenie und lautem versus leisem Lesen [The function of capital and small letters. Influences of word form, mother tongue, reading age, dyslexia, and loud versus silent reading]. In: C. Stetter (Ed.), Zu einer

Theorie der Orthographie. Interdisziplinäre Aspekte gegenwärtiger Schrift- und Orthographieforschung (pp. 1-33). Tübingen, Germany: Niemeyer.
Bock, M., Augst, G., \& Wegner, I. (1985). Groß oder klein? Zur Funktion des Wortanfangs für den gegenwärtigen Leser. [Capital or small? The function of the beginning of a word for the present-day reader]. Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 17, 191-209.
Bock, M., Hagenscheider, K., \& Schweer, A. (1989). Zur Funktion der Groß-und Kleinschreibung beim Lesen deutscher, englischer und niederländischer Texte [The function of capital and small letters during reading of German, English, and Dutch texts]. In: P. Eisenberg \& H. Günther (Eds.), Schriftsystem und Orthographie (pp. 2355). Tübingen, Germany: Niemeyer.

Engbert, R., \& Kliegl, R. (2003). Microsaccades uncover the orientation of covert attention. Vision Research, 43, 1035-1045.
Engbert, R., \& Mergenthaler, K. (2006). Microsaccades are triggered by low retinal slip. Proceedings of the National Academy of Sciences, 103, 7192-7197.
Gfroerer, S., Günther, H., \& Bock, M. (1989). Augenbewegungen und Substantivgroßschreibung: Eine Pilotstudie [Eye movements and noun capitalization: A pilot study]. In: P. Eisenberg \& H. Günther (Eds.), Schriftsystem und Orthographie (pp. 111-135). Tübingen, Germany: Niemeyer.
Hohenstein, S., \& Kliegl, R. (in press). Semantic preview benefit during reading. Journal of Experimental Psychology: Learning, Memory, and Cognition.
Hohenstein, S., Laubrock, J., \& Kliegl, R. (2010). Semantic preview benefit in eye movements during reading: A parafoveal fast-priming study. Journal of Experimental Psychology: Learning, Memory, and Cognition, 36, 11501170.

Inhoff, A. W., \& Radach, R. (1998). Definition and computation of oculomotor measures in the study of cognitive processes. In: G. Underwood (Ed.), Eye Guidance in Reading and Scene Perception (pp. 29-53). Oxford, UK: Elsevier.
Jacobs, A. M., Nuerk, H. C., Graf, R., Braun, M., \& Nazir, T. A. (2008). The initial capitalization superiority effect in German: Evidence for a perceptual frequency variant of the orthographic cue hypothesis of visual word recognition. Psychological Research, 72, 657-665.
Kliegl, R. (2007). Toward a perceptual-span theory of distributed processing in reading: A reply to Rayner, Pollatsek, Drieghe, Slattery, and Reichle (2007). Journal of Experimental Psychology: General, 136, 530-537.
Kliegl, R., Masson, M. E. J., \& Richter, E. M. (2010). A linear mixed model analysis of masked repetition priming. Visual Cognition, 18, 655-681.
Kliegl, R., Nuthmann, A., \& Engbert, R. (2006). Tracking the mind during reading: The influence of past, present, and future words on fixation durations. Journal of Experimental Psychology: General, 135, 12-35.

Laubrock, J., \& Hohenstein, S. (2012). Orthographic consistency and parafoveal preview benefit: A resourcesharing account of language differences in processing of phonological and semantic codes. Behavioral and Brain Sciences, 35, 292-293.
Müsseler, J, Nisslein, M., \& Koriat, A. (2005). German capitalization of nouns and the detection of letters in continuous text. Canadian Journal of Experimental Psychology, 59, 143-158.
Peressotti, F., Cubelli, R., \& Job, R. (2003). On recognizing proper names: The orthographic cue hypothesis. Cognitive Psychology, 47, 87-116.
R Development Core Team. (2012). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.
Radach, R. (1996). Blickbewegungen beim Lesen [Eye movements during reading]. Münster, Germany: Waxmann.
Rayner, K. (1975). The perceptual span and peripheral cues in reading. Cognitive Psychology, 7, 65-81.
Schielzeth, H., \& Forstmeier, W. (2009). Conclusions beyond support: Overconfident estimates in mixed models. Behavioral Ecology, 20, 416-420.
Schloblinski, P., Fortmann, N., Groß, O., Hogg, F., Horstmann, F., \& Theel, R. (2001). Simsen. Eine Pilotstudie zur sprachlichen und kommunikativen Aspekten in der SMS-Kommunikation [Texting. A pliot study on linguistic and communicative aspects in SMS communication]. Retrieved from http://www.mediensprache.net/networx/networx-22.pdf
Schnitzer, C. V. (2012). Linguistische Aspekte der Kommunikation in den neueren elektronischen Medien: SMS, EMail, Facebook [Linguistic aspects of communication in new electronic media: SMS, e-mail, Facebook]. (Doctoral thesis, Ludwig-Maximilians-Universität, Munich, Germany). Retrieved from http://edoc.ub.unimuenchen.de/14779/
Schotter, E. R., Angele, B., \& Rayner, K. (2012). Parafoveal processing in reading. Attention, Perception, \& Psychophysics, 74, 5-35.

# Here's not looking at you, kid! <br> Unaddressed recipients benefit from co-speech gestures when speech processing suffers 

Judith Holler (judith.holler@mpi.nl) ${ }^{1,2}$<br>Louise Schubotz (louise.schubotz@mpi.nl) ${ }^{1}$ Spencer Kelly (skelly@colgate.edu) ${ }^{3}$<br>Peter Hagoort (peter.hagoort@mpi.nl) ${ }^{1,5}$<br>Manuela Schütze (manuela.schuetze@mpi.nl) ${ }^{1}$<br>Aslı Özyürek (asli.ozyurek@mpi.nl) ${ }^{\text {1,4 }}$

1 Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525XD Nijmegen, The Netherlands<br>2 University of Manchester, School of Psychological Sciences, Coupland Building 1, M13 9PL Manchester, UK 3 Colgate University, Psychology Department, Center for Language and Brain, Oak Drive 13, Hamilton, NY 13346, USA<br>4 Radboud University, Centre for Language Studies, Erasmusplein 1, 6525HT Nijmegen, The Netherlands<br>5 Radboud University, Donders Institute for Brain, Cognition and Behaviour, Montessorilaan 3, 6525 HR<br>Nijmegen, The Netherlands


#### Abstract

In human face-to-face communication, language comprehension is a multi-modal, situated activity. However, little is known about how we combine information from these different modalities, and how perceived communicative intentions, often signaled through visual signals, such as eye gaze, may influence this processing. We address this question by simulating a triadic communication context in which a speaker alternated her gaze between two different recipients. Participants thus viewed speech-only or speech+gesture object-related utterances when being addressed (direct gaze) or unaddressed (averted gaze). Two object images followed each message and participants' task was to choose the object that matched the message. Unaddressed recipients responded significantly slower than addressees for speech-only utterances. However, perceiving the same speech accompanied by gestures sped them up to a level identical to that of addressees. That is, when speech processing suffers due to not being addressed, gesture processing remains intact and enhances the comprehension of a speaker's message.


Keywords: language processing; co-speech iconic gesture; eye gaze; recipient status; communicative intent; multi-party communication.

## Introduction

Human face-to-face communication is a multi-modal activity and often involves multiple participants. Despite this, language comprehension has typically been investigated in uni-modal (i.e., just speech) and solitary (i.e., one passive listener) contexts. The present study investigates language comprehension in the context of two other modalities omnipresent during face-to-face communication, co-speech gesture and eye gaze. Moreover,
it explores the interplay of these modalities during comprehension in a situated, dynamic social context involving multiple interlocutors in different roles.

There is, by now, a plethora of empirical evidence demonstrating that speech and co-speech gestures are semantically integrated during comprehension (e.g., Holle \& Gunter, 2007; Holle, Gunter, Rüschemeyer, Hennenlotter, \& Iacoboni, 2008; Kelly, Özyürek, \& Maris, 2010; Özyürek, Willems, Kita, \& Hagoort, 2007; Willems, Özyürek, \& Hagoort, 2007, 2009). However, only two recent studies have begun to explore to what extent this integration is automatic, and to what extent it is controlled and influenced by the pragmatics of communication, such as the perceived intentional coupling of gesture and speech (e.g., when observing a gesture performed by one person accompanying speech produced by another) (Kelly, Ward, Creigh, \& Bartolotti, 2007; Kelly, Creigh, \& Bartolotti, 2010). The findings suggest that the semantic integration of gesture and speech is indeed sensitive to the intentional coupling of the speech and gesture modalities.

A question that remains is whether this holds when we situate speech and gesture comprehension in a context that is much closer to natural communication, such as in a face-to-face context, where speech and gesture are accompanied by additional nonverbal social cues, such as eye gaze. Due to the saliency of the sclera and the contrast it forms with the iris in the human eye, gaze direction is not only omnipresent but also an extremely powerful social cue in human face-to-face interaction (Senju \& Johnson, 2009). While some studies have investigated speech and gesture comprehension in the presence of eye gaze, they have typically done so without manipulating eye gaze direction as an independent cue (e.g., Green, Straube, Weis, Jansen, Willmes, Konrad, \& Kircher, 2009; Kelly, Kravitz, \&

Hopkins, 2004; Skipper, Goldin-Meadow, Nusbaum, \& Small, 2009; Straube, Green, Jansen, Chatterjee, \& Kircher, 2010; Wu \& Coulson, 2005, 2007).

One exception is a recent study by Holler, Kelly, Hagoort, and Özyürek (2012). Their study involved one speaker alternating her gaze between two recipients, thus rendering one of them addressed and the other unaddressed during each message she communicated. Despite this study involving multi-modal messages consisting of speech and gesture, the study was designed to primarily yield insights into the influence of eye gaze direction on the processing of the gestural component of bi-modal utterances. Thus, while showing that addressed and unaddressed recipients process gestures differently, the findings revealed no effect of eye gaze on the processing of speech. However, as the authors state themselves, this does not necessarily mean that addressed and unaddressed recipients do not differ in how they process speech; one reason being that the paradigm applied in their study required participants to focus attention on the verbal modality to make judgements about the speech they heard. This explicit attentional focus might have masked effects of eye gaze on speech processing that may be revealed in other contexts.

There are some studies that provide us with good reasons to assume that this is indeed the case. For example, Schober and Clark (1989) showed that overhearers process speech less well than addressees in a referential communication task. While this study did not involve a manipulation of eye gaze direction (nor a face-to-face context), it demonstrates that recipient status can have a significant impact on how we process language. This evidence is complemented by more recent studies that did investigate speech processing in the context of gaze. For example, Staudte and Crocker (2012) showed that a robot's eye gaze towards objects in the interlocutors' environment influenced participants' reference resolution, while Knöferle and Kreysa (2012) demonstrated that a person's eye gaze towards objects influences how participants process speech with respect to thematic role assignment and syntax.

Based on this earlier research, we predict that social eye gaze, indicating communicative intent and recipient status in conversation, also influences the processing of speech. Thus, the present study investigates how, in a multiparty setting, different types of recipients (as signaled through a speaker's eye gaze direction) process speech, and speech accompanied by gestures. To do so, we developed a visually focused paradigm that avoids explicit attention to speech to allow us to better observe potential differences in addressed and unaddressed recipients' processing of both uni-modal speech-only and bi-modal speech-gesture utterances.

Like Holler et al. (2012), we implement our task in a situated, triadic communicative setting. However, in our task, participants watched a speaker conveying speech-only or speech + gesture utterances referring to objects (e.g., 'he prefers the laptop'). The gestures accompanying these
utterances in the bi-modal condition were always iconic in nature and depicted a typical feature of the object (such as its function, e.g., a typing gesture). These messages were followed by two object images, one of them having been mentioned in the utterance. The task was simple - speakers were asked to indicate as quickly as possible which of the two images was related to the speaker's preceding message. This paradigm allows us to test, firstly, how different types of recipients process speech when it is the only modality carrying semantic information, and, secondly, how they process semantic messages that are communicated bimodally, via speech and co-speech gesture.

More specifically, we are also interested in seeing whether the findings from our study are in line with the Competing Modalities Hypothesis proposed by Holler et al. (2012). This hypothesis states that unaddressed recipients focus more on gesture than do addressed recipients, since they are processing information from fewer (visual) modalities overall (i.e., no eye gaze, since the speaker's eyes are averted to the other participant). They can therefore devote more cognitive resources to the gestures, and, as a consequence, they process the gesturally depicted meaning more than addressees. In contrast to Holler et al. (2012), whose paradigm was designed to tap primarily into cospeech gesture processing, we here test this hypothesis in a paradigm that allows us to measure the processing of both gesture and speech. That is, if, in the present study, we do observe an effect of recipient status on the processing of speech in a way that is in line with past research (e.g., Schober \& Clark, 1989) - meaning unaddressed recipients process speech less well - then the enhanced processing of co-speech gestures may benefit unaddressed recipients' comprehension of the speaker's message and compensate for some (or even all) of the speech processing disadvantage.

As an alternative, Holler et al. (2012) proposed the Fuzzy Representation Hypothesis. This hypothesis predicts that unaddressed recipients perceive gestures as being less intended for them than for the gazed at recipient. They therefore process gestures less clearly than addressees, and, as a consequence, end up with a fragmented, or fuzzy, representation of the gesturally depicted meaning. If the Fuzzy Representation Hypothesis is true, then we should see no benefitting effect of gestures on the processing of speech. Rather, unaddressed recipients might be slowed down even more when trying to process bi-modal utterances, since not only the speech poses difficulties for them, but also the gestures.

The present study aims to tease apart which of these two hypotheses may best explain how addressed and unaddressed recipients (as indicated by the speaker's eye gaze direction) comprehend multi-modal language in a pragmatically much richer communication context than has been traditionally investigated, that is, in a context that bears somewhat more resemblance to the kind of joint activity that human communication is (Clark, 1996).

## Method

## Participants

32 right-handed, native German speakers (16 female) participated in the experiment (mean age 24.5 yrs ).

## Design

We used a $2 \times 2$ within-participants factorial design, manipulating the gaze direction of the speaker (direct gaze/addressed recipient condition vs. averted gaze/unaddressed recipient condition) as well as the modality of presentation (speech-only vs. speech+gesture).

## Materials and Apparatus

Video clips 160 short sentences of a canonical SVO structure were constructed. Sentences always referred to an object combined with a non-action verb (see below for more detail), e.g. 'he prefers the laptop' ('er bevorzugt den Laptop'). The iconic gestures accompanying the sentences always referred to the object that was mentioned in speech and provided information about its shape, function, or size (see Fig.1, for a gesture depicting the act of typing).

In order to guarantee that the gestures unambiguously referred to the objects mentioned, verbs were carefully selected to be as neutral as possible and were never action verbs. Hence, rather than more commonplace constructions like 'he types on the laptop' where the typing gesture could refer to both 'typing' and 'laptop', verbs like 'prefer' ('bevorzugen'), 'like' ('mögen'), or 'see' ('sehen') were used in the sentences. Our manipulation of both gaze direction and modality of presentation required each sentence to be recorded in four versions: 1. direct gaze (addressed) speech-only, 2. direct gaze (addressed) speech+gesture, 3. averted gaze (unaddressed) speech-only, and 4. averted gaze (unaddressed) speech + gesture (Fig. 1).


Figure 1: Four different versions of the 'laptop' stimulus. $\mathrm{AR}=$ addressed recipient, $\mathrm{UR}=$ unaddressed recipient.

Object pictures We created a total of 320 object pictures. 160 of these were pictures of the objects mentioned in the

160 stimulus sentences (e.g., a picture of a laptop), and an additional 160 pictures were selected to serve as unrelated pictures, such that the 'laptop' would be presented alongside a 'towel', for example (Fig. 2). Object pictures were searched via Google Images and further edited in Adobe Photoshop to have all objects presented in the same quality and size on a white background.

Prior to testing, all 320 pictures were judged by two raters (female native German speakers who did not participate in the main experiment) for their ease of identification.


Figure 2: Example of a pair of object pictures.
Each participant saw each of the 160 video clips in one of the four conditions exactly once, resulting in 160 experimental trials per participant ( 40 trials per condition), plus 24 filler trials, yielding 184 trials overall. To avoid confounding effects of the order in which the pictures were presented on the screen, this order was counterbalanced.

Videos and object pictures were presented on a 15 " laptop screen using Presentation software (http://www.neurobs.com). The audio signal of the videos was presented via high quality Sennheiser headphones.

## Procedure

Participants were tested individually. At the beginning of each testing session, participants were familiarised with the experimental set-up and the course of the experiment, and were seated in front of the experiment laptop where they received their instructions.

Participants were told that they would see a number of pre-recorded video clips of a speaker (in fact a confederate) who, they were told, spontaneously formed short sentences based on line drawings and single words displayed on a screen not displayed in the video shot. They were also told that during the recordings, a second person was present in the room, sitting diagonally across from the speaker. The speaker was supposedly instructed to sometimes address this other (fictitious) participant when producing her utterances (averted gaze condition), and to sometimes address the (actual) participant via a video camera positioned straight across from her (direct gaze condition). Participants were instructed that following each video clip, they would see two pictures of objects on the screen, and that it was their task to indicate via button press which of the two pictures best matched the speaker's message (left button for the left-hand picture, right button for the righthand picture). They were asked to react as quickly and as
accurately as possible. Reaction times of participants' left/right responses were recorded via a button box, as were response accuracies.
In order to ensure that participants were actually watching the video clips and not basing their decision on the spoken part of the message only, they were explicitly asked to look at the screen during the entire course of the experiment. This was further enforced by the presence of a surveillance camera (our checks showed that no participant had looked away), which all participants agreed to be videorecorded with during the experiment.
Before the beginning of the experiment proper, participants completed a total of six practice trials. As in the actual experiment, each trial consisted of a video clip, followed immediately by the two object pictures, which stayed onscreen until the participants pressed a button. After their response, participants saw a fixation cross for a random time interval between 2 and 5 seconds before the next trial started.

## Results

A total of six trials from two participants were excluded from the analysis beforehand because of a technical error. An alpha value of .05 was used throughout our statistical analyses. All p-values reported are two-tailed.
For the analysis of the reaction times ${ }^{1}$, we excluded all incorrect responses, 83 in total ( $=1.62 \%$ of all trials). Also excluded from the analysis were responses more than 2.5 SD above or below each subject's mean reaction time (this resulted in 118 responses being excluded: 40 in the speechonly condition, direct gaze, 31 in the speech-only condition, averted gaze, 23 in the speech + gesture condition, frontal gaze, and 24 in the speech + gesture condition, averted gaze).
Figure 3 shows the reaction time data for the 2 (gaze direction: direct vs. averted) x 2 (modality of presentation: speech-only vs. speech + gesture) repeated measures ANOVA. The results yielded a significant interaction, $\mathrm{F}(1,31)=5.947, \mathrm{p}=.021$. The main effect of modality was not significant, $\mathrm{F}(1,31)=3.431, \mathrm{p}=.074$, and neither was the main effect of gaze, $\mathrm{F}(1,31)=.464, \mathrm{p}=.501$.
In line with our hypotheses, we calculated two a priori contrasts (using paired-samples t -tests), comparing addressed and unaddressed recipients' processing of unimodal speech-only utterances, as well as their processing of the bi-modal speech+gesture utterances. The first comparison showed that unaddressed recipients ( $\mathrm{M}=$ 542 ms ) were significantly slower than addressees ( $\mathrm{M}=$ 530 ms ) at processing speech-only utterances, $\mathrm{t}(1,31)=$ $2.547, \mathrm{p}=.016$. The second comparison, however, showed that unaddressed ( $M=525 \mathrm{~ms}$ ) and addressed $(M=531 \mathrm{~ms})$

[^298]recipients did not differ in their processing of speech + gesture utterances, $\mathrm{t}(1,31)=1.112, \mathrm{p}=.275$.


Figure 3: Addressed recipients' (AR) and unaddressed recipients' (UR) reaction times (ms) in the speech-only and speech + gesture conditions (error bars represent SE ).

## Discussion

This study investigated multi-modal language processing in a situated, socially dynamic communication setting involving multiple parties. The specific question we tried to answer is how different types of recipients, as signaled through a speaker's eye gaze, process speech and speech accompanied by iconic gestures, in a triadic communication scenario. The findings revealed a significant interaction between modality and recipient status. More precisely, they show, first and foremost, that the processing of speech-only utterances is indeed affected by recipient status in our task, since unaddressed recipients were significantly slower in this condition than were addressed recipients. Crucially, addressed and unaddressed recipients did not differ in their processing of speech+gesture utterances. That is, unaddressed recipients significantly benefitted from the information depicted in the gestural modality, allowing them to perform at the same level as addressees when perceiving bi-modal rather than uni-modal utterances.

The findings are thus very much in line with the Competing Modalities Hypothesis (Holler et al., 2012). Unaddressed recipients appear to focus their cognitive resources on the processing of co-speech iconic gestures. At the same time, the findings allow us to further refine this hypothesis; because we found that unaddressed recipients do not process speech more quickly than addressed recipients, the competition effect seems to apply to the visual modalities (gesture and gaze) only. In other words, due to not having to process eye gaze, unaddressed recipients can focus more on gesture and, as a consequence
process this information more. Their increased processing capacity due to the absence of direct gaze does not, however, affect their processing of speech-only utterances.

The reason as to why, in contrast to Holler et al. (2012), we found a numerical but no reliable difference between addressed and unaddressed recipients' processing of bimodal utterances (i.e., unaddressed recipients were slightly faster in the bi-modal condition than addressed recipients were, but not significantly so) is likely to be due to our change in paradigm. As argued in the Introduction, the explicit attentional focus on the verbal modality in Holler et al.'s (2012) study might have masked differences in the processing of speech - an assumption that we were able to corroborate here. In the present study, we purposefully shifted participants' attention towards the visual modality (by asking them to identify pictures) in order to uncover potentially previously masked differences in speech processing, while being aware that this shift in paradigm might, in turn, reduce differences in the processing of visual (i.e., gestural) information between addressed and unaddressed recipients. The present study thus complements that by Holler et al. (2012) nicely. Together, they offer us a more comprehensive insight into how different recipients process uni-modal and bi-modal utterances in the presence of eye gaze.
What remains to be investigated are the exact cognitive mechanisms underlying our Competing Modalities account. Currently, we are unable to determine whether the iconic co-speech gestures benefit unaddressed recipients' processing of speech because they are semantically integrated with the verbal information - thus leading to a richer, unified mental representation of the concept of 'laptop', for example - or whether they lead to a stronger memory trace due to receiving related information from two different input streams (visual and verbal), with this information being associated but stored separately and not as a unified representation (much like a dually-coded representation à la Paivio (1986)). Future studies, preferably involving on-line measures suitable for dipping directly into semantic integration processes, are needed to answer this question.
In conclusion, the present study has brought together three different modalities in a language processing paradigm, and it advances our understanding of how perceived communicative intent, as signaled through a speaker's eye gaze, influences the interplay of these modalities during comprehension in a situated, face-to-facelike (rather than solitary) setting. The findings are striking since we have shown that the ostensive cue of eye gaze has the power to modulate how different recipients process semantic information carried by two concurrent modalities, speech and co-speech gestures. Moreover, we have shown that in situated face-to-face settings involving multiple recipients, the gestural modality can benefit unaddressed recipients - when speech processing suffers, gestures help.

## Acknowledgments

We would like to thank the University of Rostock for their provision of testing space, the Radboud University and MaxNet initiative for financial support, the Neurobiology of Language Department and the Gesture \& Sign Research Group at the Max Planck Institute for Psycholinguistics for helpful feedback during discussions of this study, and the European Commission for funding this research (JH was supported through a Marie Curie Fellowship \#255569, and through ERC Advanced Grant \#269484INTERACT).

## References

Clark, H. (1996). Using language. Cambridge: Cambridge University Press.
Green, A., Straube, B., Weis, S., Jansen, A., Willmes K., Konrad, K., \& Kircher, T. (2009). Neural integration of iconic and unrelated coverbal gestures: a functional MRI study. Human Brain Mapping, 30, 3309-3324.
Holle, H. \& Gunter, T. C. (2007). The role of iconic gestures in speech disambiguation: ERP evidence. Journal of Cognitive Neuroscience, 19, 1175-1192.
Holle, H., Gunter, T. C., Rüschemeyer, S.A., Hennenlotter, A., \& Iacoboni, M. (2008). Neural correlates of the processing of co-speech gestures. NeuroImage, 39, 20102024.

Holler, J., Kelly, S., Hagoort, P., \& Özyürek, A. (2012). When gestures catch the eye: The influence of gaze direction on co-speech gesture comprehension in triadic communication. In N. Miyake, D. Peebles, \& R. P. Cooper (Eds.), Proceedings of the 34th Annual Meeting of the Cognitive Science Society (pp. 467-472). Austin, TX: Cognitive Society.
Kelly, S. D., Creigh, P., \& Bartolotti, J. (2010). Integrating speech and iconic gestures in a Stroop-like task: Evidence for automatic processing. Journal of Cognitive Neuroscience, 22, 683-694.
Kelly, S. D., Kravitz, C., \& Hopkins, M. (2004). Neural correlates of bimodal speech and gesture comprehension. Brain \& Language, 89, 253-260.
Kelly, S. D., Özyürek, A., \& Maris, E. (2010). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. Psychological Science, 21, 260267.

Kelly, S. D., Ward, S., Creigh, P., \& Bartolotti, J. (2007). An intentional stance modulates the integration of gesture and speech during comprehension. Brain and Language, 101, 222-233.
Knöferle, P. \& Kreysa, H. (2012). Can speaker gaze modulate syntactic structuring and thematic role assignment during spoken sentence comprehension? Frontiers in Cognitive Science, 3, 538.
Özyürek, A., Willems, R. M., Kita, S., \& Hagoort, P. (2007). On-line integration of semantic information from speech and gesture: Insights from event-related brain potentials. Journal of Cognitive Neuroscience, 19, 605-
616.

Paivio, A (1986). Mental representations: a dual coding approach. Oxford, UK: Oxford University Press.
Schober, M. F. \& Clark, H. H. (1989). Understanding by addressees and overhearers. Cognitive Psychology, 21, 211-232.
Senju, A. \& Johnson, M. H. (2009). The eye contact effect: Mechanisms and development. Trends in Cognitive Sciences, 13, 127-134.
Skipper, J. I., Goldin-Meadow, S., Nusbaum, H. C., \& Small, S. L. (2009). Gestures orchestrate brain networks for language understanding. Current Biology, 19, 661667.

Staudte, M. \& Crocker, M. (2011). Investigating Joint Attention Mechanisms through Spoken Human-Robot Interaction. Cognition, 120, 268-291.
Straube, B., Green, A., Jansen, A., Chatterjee, A., \& Kircher, T. (2010). Social cues, mentalizing and the neural processing of speech accompanied by gestures. Neuropsychologia, 48, 382-393.
Willems, R. M., Özyürek, A., \& Hagoort, P. (2007). When language meets action: The neural integration of speech and gesture. Cerebral Cortex, 17, 2322-2333.
Willems, R. M., Özyürek, A., \& Hagoort, P. (2009). Differential roles for left inferior frontal and superior temporal cortex in multimodal integration of action and language. NeuroImage, 47, 1992-2004.
Wu, Y.C. \& Coulson, S. (2005). Meaningful gestures: Electrophysiological indices of iconic gesture comprehension. Psychophysiology, 42, 654-667.
Wu, Y. C. \& Coulson, S. (2007). How iconic gestures enhance communication: An ERP study. Brain and Language, 101, 234-245.

# Encountering Multiple Exemplars During Fast Mapping Facilitates Word Learning 

Jessica S. Horst (jessica@sussex.ac.uk)<br>University of Sussex, School of Psychology, Pevensey 1 Building<br>Falmer, Brighton, East Sussex, BN1 9QH, UK<br>Katherine E. Twomey (k.twomey@liverpool.ac.uk)<br>University of Liverpool, Department of Experimental Psychology, Eleanor Rathbone Building, Bedford St South, Liverpool, L69 7WW, UK<br>Samantha L. Ranson (samanthalr@hotmail.com)<br>University of Sussex, School of Psychology, Pevensey 1 Building<br>Falmer, Brighton, East Sussex, BN1 9QH, UK


#### Abstract

Previous research indicates learning words facilitates categorization. In the current study, we investigated whether learning about a category facilitates word learning (retention) by presenting 2 -year-old children with multiple referent selection trials to the same object category. Children either encountered the same exemplar repeatedly or encountered multiple exemplars across trials. All children did very well on the initial task. However, only children who encountered multiple exemplars retained these mappings after a short delay. Overall, these data provide strong evidence that providing children with the opportunity to compare across exemplars during referent selection facilitates retention.


Keywords: word learning; fast mapping; categorization; multiple exemplars

Learning the names for object categories is necessary for children to make sense of their world and to communicate about it effectively. Perhaps it is not surprising, then, that children's early vocabularies are dominated by names for object categories (Samuelson \& Smith, 1999; Waxman, 2003). Previous research has demonstrated a close relationship between vocabulary acquisition and categorization (e.g., Gopnik \& Meltzoff, 1992; Thom \& Sandhofer, 2009). However, although several studies have demonstrated knowing more words facilitates categorization, it remains unclear how experience with object categories may facilitate word learning.

Word learning is a complicated process, involving both fast and slow mapping (McMurray, Horst, \& Samuelson, 2012). The first time a novel name is encountered, the child quickly forms an initial, rough hypothesis of the word's meaning-hence the term fast mapping (Carey, 1978). For example, when presented with a boat, a cup and a novel black-and-white stuffed animal and asked for the penguin, a 2-year-old child can reliably determine that penguin refers to the animal (PENGUIN). However, simply forming this initial mapping does not mean that the child has really learned the word-that is, that the child could recall the name-object association after a delay or in a new context, for example with other novel toys (Horst \& Samuelson, 2008; Riches, Tomasello, \& Conti-Ramsden, 2005, Waxman \& Booth, 2000). Indeed, processing demands
might prevent young children from learning the correct name-object association after only a single exposure (Mather \& Plunkett, 2009).

In contrast to fast mapping, full word learning emerges gradually during a period of slow mapping (Capone \& McGregor, 2005; Carey, 1978; Horst \& Samuelson, 2008). During this phase repeated encounters allow the child to strengthen the name-object association such that it can be recalled after a delay. Importantly, the penguin-PENGUIN association will be strengthened each time the child hears the word penguin and sees the animal in a new situation. For example, a child might see a stuffed penguin at daycare and then later play with a penguin and other animals during bath time at home. Across such situations children learn about the statistical regularity with which the names and their referents co-occur (cross-situational word learning; Munro, Baker, McGregor, Docking \& Arculi, 2012; Smith \& Yu, 2008). Clearly, then, repeated exposures are critical for word learning.

However, children do not only learn names for individual items, but also learn names for object categories. Categories are collections of items which share common features (e.g., Quinn, Eimas \& Rosenkrantz, 1993; Rosch, 1975), but which are still discriminable from each other. In the PENGUIN category, for instance, the majority of members share the common features of black-and-white coloring, two legs and the ability to swim, but the individual members are discriminable. For example, a child can discriminate between a stuffed penguin and a plastic penguin bath toy. Importantly, during early word learning, children not only encounter the same category exemplar repeatedly but may encounter multiple, different exemplars over time.

When children are presented with multiple exemplars across situations, they may compare across items, which induces categorization by helping children to detect both the commonalities and differences between the category members, both of which are critical for categorization (e.g., Kovack-Lesh \& Oakes, 2007; Oakes \& Ribar, 2005). Kovack-Lesh and Oakes (2007) have reported that simply providing the opportunity to compare across exemplars during the transition between trials is enough to help infants form a category they otherwise do not form when presented
with the same items in the same sequence. This is especially important because it demonstrates that young children are able to compare between exemplars across trials.

Namy and Gentner (2002) as well as others (e.g., Casasola, Bhagwat, \& Burke, 2009; Waxman, 2003) have previously argued that applying a common name to multiple exemplars invites children to compare across items and draws their attention to shared commonalities. For example, when two objects are given the same name, children will extend this common name to new objects that share the same perceptual features with the named exemplars (e.g., Samuelson \& Smith, 1999). Importantly, these findings demonstrate that exposing children to multiple, variable exemplars labeled with a common, novel name allows children to detect the similarities between objects and therefore facilitates categorization.

However, it remains unclear whether comparison facilitates children's ability to retain category names because the existing studies on the effect of presenting multiple category exemplars on word learning have focused largely on generalization. For example, in a longitudinal category training study, toddlers who encountered multiple perceptually variable exemplars experienced a significant acceleration in vocabulary growth and were able to generalize novel names to novel exemplars from the same categories, in contrast to children who encountered perceptually similar exemplars (Perry, Samuelson, Malloy, \& Schiffer, 2010).

The current study examines whether providing children with the opportunity to compare across exemplars facilitates their ability to learn and retain names for novel object categories. We tested 2-year-old children because they can complete multiple trials without becoming overly tired and enjoy this particular task. We provided children with multiple fast mapping by mutual exclusivity trials to better understand how encountering multiple exemplars facilitates cross-situational word learning. Further, while previous studies have investigated how encountering multiple exemplars effects children's generalization of novel names, the current study explores the effect on retention. Specifically, children encountered each novel object category across three referent selection trials. Half of the children were repeatedly presented with the same exemplar across trials and half of the children were presented with multiple exemplars across trials. If providing the opportunity to compare across exemplars facilitates crosssituational word learning, then children who fast-mapped multiple exemplars should demonstrate better retention.

## Method

## Participants

Twenty-four children aged 2;6 (13 girls, $M=2 ; 6 S D=43.19$ days; range $=2 ; 4-2 ; 8$ ) with a mean productive vocabulary of 563.75 words $(S D=81.91$ words, range $=391-668$ words) and no family history of colorblindness participated. Children were from predominantly middle class homes. Half of the children were randomly assigned to the single
exemplars condition and the other half were randomly assigned to the multiple exemplars condition. Children's ages and productive vocabularies did not differ between conditions. Data from two additional children were excluded from analyses due to fussiness and experimenter error. Parents were reimbursed for travel expenses and children received a small gift for participating.


Figure 1: Novel stimuli

## Stimuli

Eighteen known objects, chosen because they are highly familiar to 2-year-old children, served as familiar objects: bird, chicken, elephant, fish, giraffe, lion, boat, bus, car, motorcycle, plane, train, block, chair, comb, cup, toy mobile phone and spoon.

Nine novel objects from three categories, chosen because they are not easily named by 2-year-old children, served as the target objects (see Figure 1). Consistent with other studies (e.g., Vlach, Sandhofer, \& Kornell, 2008), the objects in these categories varied in color and texture, but shared the same shape.

The doff category consisted of slightly transparent, plussign shaped tops in yellow, green and red. The cheem category consisted of plastic rods with small balls on one end in blue/orange, orange/blue and yellow/green. The hux category consisted of rubber balloons with elastic strings hanging down in blue/orange, yellow/blue and green. The balloons kept their shape because they had foam balls inside them. All objects were similar in size ( $5 \mathrm{~cm} \times 8 \mathrm{~cm} \times 10 \mathrm{~cm}$ ). Stimuli were presented on a white tray divided into three even sections. A digital kitchen timer was used to time the 5-minute break.

## Procedure and Design

Before the experiment began, the experimenter showed the parent color photographs of the known and novel objects to ensure they were known and novel to the child, respectively (which they were for all children). During the experiment, children were seated in a booster seat at a white table across from the experimenter. Parents sat next to their children and completed a vocabulary checklist and were instructed to
avoid interacting with their children, but to encourage them to respond during the warm-up trials if necessary. None of the children needed parental encouragement after the warmup trials.

Warm-up trials Each session began with three warm-up trials to introduce children to the task. On each trial, children were presented with three randomly selected known objects. First, the experimenter set the tray of objects on the table and silently counted for three seconds to give the child an opportunity to look at the objects (see also, Horst, Scott, \& Pollard, 2010). Then, the experimenter asked the child to select an object by naming it twice (e.g., "Can you find the block? Can you get the block?") before sliding the tray forward. Children were praised heavily for correct responses and corrected if necessary. Between trials the experimenter replaced the tray on her lap and arranged the objects for the next trial out of the child's view.

The same objects were presented on each warm-up trial, but object positions (left, middle, right) were pseudorandomized across trials. Thus, children were asked for a different object in a different position on each trial. These stimuli were later used as known objects during the referent selection trials (see also, Horst \& Samuelson, 2008).

Referent Selection Task. Referent selection trials immediately followed the warm-up trials and proceeded in the same manner except that children were neither praised nor corrected.

Each child was presented with nine sets and saw each set once on a known name referent selection trial and once on a novel name trial for a total of 18 referent selection trials (see Figure 2 for examples). Known name trials were included to ensure that children were mapping the names to the requested targets and not simply mapping novelty to novelty (Horst, Samuelson, Kucker \& McMurray, 2011).


Figure 2: Trials on which a doff exemplar was present

Each set included two familiar objects (e.g., boat and cup) and one novel object (e.g., top). Children in the multiple exemplars condition saw a different novel exemplar in each set. For example, a child might see the green top with the block and lion, the red top with the chair and train and the yellow top with the bus and fish (see Figure 2). Children in the single exemplars condition saw the same exemplar in each set. For example, a child might see the green top with the block and lion, and again the chair and train and once more with the bus and fish. Thus, the only difference between conditions was whether children saw one or three exemplars from each category.

Referent selection trials were presented in three blocks. For example, one child completed all trials with the doff category, then all trials with the cheem category and finally all trials with the hux category. Block order was counterbalanced across participants using a Latin Square design. The order of known and novel trials was pseudorandomized in each block such that the same set (e.g., green top, lion, block) was never presented on two consecutive trials and no more than two trials of either type (i.e., known or novel) were presented sequentially. Object position (left, middle, right) was randomly determined on each trial. Between the referent selection task and the retention task the child remained at the table and colored pictures during a 5minute delay period, which was included to ensure that children's retention was based on long-term memory representations for the novel name-object associations formed during the referent selection phase rather than shortterm maintenance (for a similar argument see, Horst \& Samuelson, 2008).

Retention Task The retention task was the same in both conditions. First, to re-engage children in the task, a new warm-up trial with three different known objects was presented. This was immediately followed by three retention trials, during which children saw three novel exemplars: one from each novel category (top, rod, balloon). The same exemplars were presented on all trials for a given child. In the single exemplars condition, children were presented with the same exemplars encountered earlier. In the multiple exemplars condition, children were presented with one of the three exemplars encountered earlier, with each exemplar (e.g., green top) being presented equally often across children. Object positions were randomized across trials and children were asked for a different novel object in a different position on each trial.

Coding. Children's responses were coded offline from DVD. Responses included touching and picking up objects (see Horst et al., 2011, for a deeper discussion of possible responses on this type of task). A naïve coder coded $20 \%$ of the sessions for reliability. Inter-coder agreement was high, $M=98.08 \%, S D=3.44 \%$ (range $=92.31 \%-100.00 \%$ ). Overall, $90 \%$ of the target words were included in the analyses of children's retention as, like prior studies, only the words that a child correctly fast-mapped at least once
during the referent selection trials were included in subsequent analyses. There was no evidence of group differences in interest/attention during the experiment.

## Results

We first compare children's performance to chance levels and then compare children's performance between conditions. As can be clearly seen in the left panel of Figure 3 , children in both conditions were very accurate at choosing the target object during the initial referent selection task. On known name referent selection trials, 11 children in each condition chose the target on every trial, and one child in each condition chose the target on $8 / 9$ trials. Thus, children's proportion of target choices was the same for both conditions and greater than would be expected by chance (.33), $t(11)=71.73, p<.0001, d=20.60$ (all $p$ s are two-tailed). On novel name referent selection trials, children's proportion of target choices was also greater than expected by chance (.33) both for children in the multiple exemplars condition, $t(11)=6.57, p<.0001, d=2.38$ and for children in the single exemplars condition, $t(11)=4.59$, $p<.001, d=.84$. Again, there was no difference between conditions, $t(22)=.345$, ns. Thus, whether children encountered multiple exemplars or the same exemplars repeatedly during referent selection did not influence children's performance on either known or novel name referent selection trials.

Our main question in this experiment was whether encountering multiple exemplars or the same exemplars repeatedly during referent selection influenced retention. As can be seen in the right panel of Figure 3, only children in the multiple exemplars condition retained more names than expected by chance (.33), $t(11)=5.00, p<.001, d=1.46$. Children in the single exemplars condition failed to retain more words than expected by chance, $t(11)=1.47, n s, d=$ .44. An unpaired $t$-test confirmed that children who encountered multiple exemplars retained more words than children who encountered the same exemplars repeatedly, $t(22)=2.06, p \leq .05, d=.16$.


Figure 3: Children's proportion of correct choices. Dotted line represents chance (.33). Error bars represent one standard error. ${ }^{* * *} p<.0001,{ }^{* *} p<.001,{ }^{*} p \leq .05$.

Table 1: Number of words retained as a function of number of correct referent selection trials. N in parentheses. Exact binomial probabilities, ${ }^{* * *} p<.0001, * * p<.01, * p<.05$.

|  | Correct Number of Trials per Word During Referent Selection |  |  |
| :---: | :---: | :---: | :---: |
|  | One Trial | Two Trials | Three Trials |
| Multiple | 2 (6) | 6 (9)* | 15 (18)*** |
| Exemplars |  |  |  |
| Single | 1 (6) | 3 (8) | 11 (17)** |
| Exemplars |  |  |  |

To further understand how multiple exemplars influence children's ability to retain newly fast-mapped names we also explored retention as a function of number of successful referent selection trials. As can be seen in Table 1 , when children only successfully fast-mapped on one of the three trials, they were unable to retain that name over a 5-minute delay, regardless of whether they saw the same or different exemplars on their two unsuccessful trials. When children successfully fast-mapped twice, they were able to retain that category name if they encountered multiple exemplars but not if they encountered the same exemplar repeatedly. Finally, when children successfully fast-mapped three times, they were able to retain that name whether they had mapped the name to multiple exemplars or to the same exemplar repeatedly. Taken together, these data confirm that multiple exemplars facilitate word learning via fast mapping and that sufficient encounters with the same exemplar can also lead to retention.

## Discussion

The current study explored how providing the opportunity to compare across multiple category exemplars facilitates children's ability to learn and retain names for novel object categories. We presented 2 -year-old children with multiple referent selection trials with the same object category. Children either encountered the same exemplar repeatedly or multiple exemplars across trials. Overall, all children did very well on the initial referent selection task. However, only children who encountered multiple exemplars retained the previously fast-mapped novel names after a delay. Further, these children demonstrated significantly better retention than children who only encountered the same exemplar repeatedly.

Overall, these data demonstrate that experience with multiple exemplars facilitates word learning, specifically retention of fast-mapped names for object categories. Other studies that have explored the relationship between vocabulary and categorization have typically tested children over a long time scale, such as several weeks (Ellis \& Oakes, 2006; Perry et al., 2010). However, the current study reveals that exposing children to an object category, rather than a single category member, facilitates children's ability to learn the name for that category within minutes (see also Kemler Nelson, O'Neil, \& Asher, 2008).

These findings also add to the literature demonstrating that comparison facilitates categorization (e.g., Gentner \& Namy, 1999; Kovack-Lesh \& Oakes, 2007) and that applying a common name to multiple exemplars invites children to compare across items, drawing their attention to shared commonalities (Casasola et al., 2009; Gentner \& Namy, 1999; Namy \& Gentner, 2002; Plunkett, Hu \& Cohen, 2008). It is likely that children also learned from encountering the same exemplars repeatedly, but that this learning was not robust enough to withstand a short delay.

The current study also demonstrates that behavior is the product of nested timescales, consistent with dynamic systems theory (Thelen \& Smith, 1994). Specifically, in the current study, children's ability to retain words emerged as a product of their present (what they were currently seeing), their just previous past (how many exemplars they had just fast-mapped) and their past (their developmental history of learning about names and categories).

Importantly, these data clearly indicate that encountering multiple exemplars led to better novel name retention. We believe that children who encountered multiple exemplars retained words at greater rates because each encounter with a new exemplar invited them to compare the new exemplar to their stored memory representations for that object category, thus enabling them to encode additional information. That is, as each exemplar was encountered children's stored memory representations were updated and elaborated. This explanation is consistent with exemplar theories of categorization (e.g., Murphy, 2002; Medin \& Schaffer, 1978; Nosofsky, 1984), which argue that individual representations are formed each time an exemplar is encountered.

Previous research that has investigated how multiple exemplars influence children's word learning has done so by presenting multiple exemplars at test. After encountering a single exemplar from the target category, children are typically presented with one of two types of test trials. Using referent selection tasks, children are tested with another exemplar from the same category, a completely novel foil and known foils (e.g., Mervis \& Bertrand, 1994; Golinkoff, Hirsh-Pasek, Bailey \& Wenger, 1992). Children are very good at selecting the target as opposed to a completely novel object. However, any delay is minimal (e.g., 1-2 trials later), thus we cannot be sure that we are testing children's long-term memory for new words rather than short-term maintaince.

Using naturalistic play situations, children are tested after a short delay, but they are tested with the same previouslyencountered exemplar, another exemplar from the same category and a novel foil (e.g., Jaswal \& Markman, 2003; Waxman \& Booth, 2000). Children are very good at selecting the previously-encountered exemplar, however, it is not clear that we are testing generalization if the same exemplar is presented again. In addition, these tasks do not control for novelty differences between the test alternatives, which can have a profound effect on children's responses (Horst, et al., 2011).

The current study is different. Specifically, we presented multiple exemplars during referent selection to provide an opportunity to compare across exemplars during fast mapping to investigate the effect of comparison on full word learning. Note, other studies that have tested children's retention for name-object associations learned via referent selection have only included one exemplar for each category (e.g., Horst \& Samuelson, 2008; Kucker \& Samuelson, 2011), although, recently Ankowski, Vlach and Sandhofer (2012) presented multiple exemplars to demonstrate that simultaneous presentation facilitates abstraction and generalization for new category members better than spaced presentation. In addition, the relative novelty of the test alternatives was controlled as each had previously served as a target and each had been encountered the same number of times. Thus, although previous research has tested the strength of children's newly formed nameobject category associations by presenting different exemplars at test, the current study is the first fast mapping study to explore the role of comparison in word learning by manipulating the strength of children's name-object category associations formed during referent selection across encounters with multiple exemplars.

Overall, then, the current study adds to a growing body of evidence that experience with multiple exemplars and within-category variability influences young children's word learning. Importantly, this study demonstrates that categorization can have a profound effect on children's word learning over a short time scale. Thus, the current study is among the first to systematically investigate the interplay between category variability and cross-situational word learning, and as such provides important groundwork for further research in the area, as well as informs our understanding of category learning and cognitive development, more generally.

## Acknowledgments

We thank the parents and children who participated and members of the Word Lab for their administrative assistance. We would also like to thank Emma Axelsson, Karla McGregor, Sammy Perone and Larissa Samuelson for helpful discussions related to this research.

## References

Ankowski, A. A., Vlach, H. A., \& Sandhofer, C. M. (2012). Comparison versus contrast: Task specifics affect category acquisition. Infant and Child Development.
Capone, N. C., \& McGregor, K. K. (2005). The effect of semantic representation on toddlers' word retrieval. Journal of Speech, Language, and Hearing Research, 48(6), 1468-1480.
Carey, S. (1978). The child as word learner. In M. Halle, J. Bresnan \& A. Miller (Eds.), Linguistic Theory and Psychological Reality (pp. 264-293). Cambridge, MA: MIT Press.
Casasola, M., Bhagwat, J., \& Burke, A. S. (2009). Learning to form a spatial category of tight-fit relations: How
experience with a label can give a boost. Developmental Psychology, 45(3), 711-723.
Ellis, A. E., \& Oakes, L. M. (2006). Infants flexibly use different dimensions to categorize objects. Developmental Psychology, 42(6), 1000-1011.
Gentner, D., \& Namy, L. L. (1999). Comparison in the development of categories. Cognitive Development, 14, 487-513.
Golinkoff, R. M., Hirsh-Pasek, K. Bailey, L. M. \& Wenger, N. R. (1992). Young children and adults use lexical principles to learn new nouns. Developmental Psychology (28)1, 99-108.
Gopnik, A., \& Meltzoff, A. (1992). Categorization and naming: Basic-level sorting in eighteen-month-olds and its relation to language. Child Development, 63(5), 10911103.

Horst, J. \& Samuelson, L. (2008). Fast mapping but poor retention by 24-month-old infants. Infancy, 13, 128-157.
Horst, J. S., Samuelson, L. K., Kucker, S. C., \& McMurray, B. (2011). What's new? Children prefer novelty in referent selection. Cognition, 118(2), 234-244.
Horst, J. S., Scott, E. J., \& Pollard, J. A. (2010). The role of competition in word learning via referent selection. Developmental Science, 13(5), 706-713.
Jaswal, V. K. \& Markman, E. M. (2003). The relative strengths of indirect and direct word learning. Developmental Psychology, 29(4), 745-760.
Kemler Nelson, D. G., O'Neil, K. A., \& Asher, Y. M. (2008). A mutually facilitative relationship between learning names and learning concepts in preschool children: The case of artifacts. Journal of Cogntiion and Development, 9(2), 171-193.
Kovack-Lesh, K. A., \& Oakes, L. M. (2007). Hold your horses: How exposure to different items influences infant categorization. Journal of Experimental Child Psychology, 98(2), 69-93.
Oakes, L. M., \& Ribar, R. J. (2005). A comparison of infants' categorization in paired and successive presentation familiarization tasks. Infancy, 7(1), 85-98.
Mather, E., \& Plunkett, K. (2009). Learning words over time: The role of stimulus repetition in mutual exclusivity. Infancy, 14(1), 60-76.
McMurray, B., Horst, J. S., \& Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. Psychological Review, 119(4), 831.
Medin, D. L., \& Schaffer, M. M. (1978). Context theory of classification learning. Psychological Review, 85(3), 207.
Mervis, C. B. \& Bertrand, J. (1994). Acquisition of the Novel Name-Nameless Category (N3C) Principle. Child Development, 65(6), 1646-1662.
Munro, N., Baker, E., McGregor, K., Docking, K., \& Arculi, J. (2012). Why word learning is not fast. Frontiers in Psychology, 3. (41) 1-10.
Murphy, G. L. (2002). The Big Book of Concepts. MIT Press. Cambridge, MA.

Namy, L. L. \& Gentner, D. (2002). Making a silk purse out of two sow's ears: Young children's use of comparison in category learning. Journal of Experimental Psychology-: General, 131(1), 5-15.
Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. Journal of Experimental Psychology: Learning, Memory, and Cognition, 10(1), 104.

Perry, L. K., Samuelson, L. K., Malloy, L. M., \& Schiffer, R. N. (2010). Learn locally, think globally: Exemplar variability supports higher-order generalization and word learning. Psychological Science, 21(12), 1894-1902.
Plunkett, K., Hu, J. F., \& Cohen, L. B. (2008). Labels can override perceptual categories in early infancy. Cognition, 106(2), 665-681.
Quinn, P. C., Eimas, P. D., \& Rosenkrantz, S. L. (1993). Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old Infants. Perception, 22(4), 463-475.
Riches, N. G., Tomasello, M., \& Conti-Ramsden, G. (2005). Verb learning in children with SLI: Frequency and spacing effects. Journal of Speech, Language, and Hearing Research, 48(6), 1397.
Rosch, E. (1975). Cognitive representations of semantic categories. Journal of Experimental Psychology-General, 104(3), 192-233.
Samuelson, L. K., \& Smith, L. B. (1999). Early noun vocabularies: Do ontology, category structure and syntax correspond? Cognition, 73(1), 1-33.
Smith, L. B., \& Yu, C. (2008). Infants rapidly learn wordreferent mappings via cross-situational statistics. Cognition, 106, 1558-1568.
Thelen, E. \& Smith, L. B. (1994). A dynamic systems approach to the development of cognition and action. MIT Press. Cambridge, MA.
Thom, E. E., \& Sandhofer, C. M. (2009). More is more: The relationship between vocabulary size and word extension. Journal of Experimental Child Psychology, 104(4), 466473.

Vlach, H., Sandhofer, C. M., \& Kornell, N. (2008). The spacing effect in children's memory and category induction. Cognition, 109(1), 163-167.
Waxman, S. R. (2003). Links between object categorization and naming: Origins and emergence in human infants. In D. H. Rakison \& L. M. Oakes (Eds.), Early category and concept development: Making sense of the blooming, buzzing confusion. (pp. 213-241). New York: Oxford University Press.
Waxman, S. R. \& Booth, A. E. (2000). Principles that are invoked in the acquisition of words, but not facts. Cognition, 77(2), B33-B43.

# Dilution Effects in Perceptual Information Integration 

Jared M. Hotaling ${ }^{\text {a }}$ (jhotalin@indiana.edu), Andrew L. Cohen ${ }^{\text {b }}$ (acohen@psych.umass.edu), Jerome R. Busemeyer ${ }^{\text {a }}$ (jbusemey@indiana.edu), \& Richard M. Shiffrin ${ }^{\text {a }}$ (shiffrin@indiana.edu)<br>${ }^{\text {a }}$ Department of Psychological \& Brain Sciences, Indiana University<br>1101 E. Tenth St., Bloomington, IN 47408 USA<br>${ }^{\mathrm{b}}$ Department of Psychology, University of Massachusetts<br>Amherst, MA 01003 USA


#### Abstract

In cognitive science there is a paradox: Researchers studying decision making have repeatedly shown that people employ simple and often less than optimal strategies when integrating information from multiple sources. However, researchers working in fields such as categorization, memory, and perception have had great success using optimal models to account for information integration. Is this conflict due to the use of different materials and procedures? We test the hypothesis that stimuli requiring more controlled information integration lead to suboptimal performance, while stimuli that lend themselves to more automatic processing produce more optimal integration. We test for one canonical example of sub-optimal information integration, the dilution effect, using stimuli more commonly found in perception experiments. Dilution was indeed reliable across several conditions. The largest effects occurred in stimuli manipulated so as to discourage automatic processing. We use the Multicomponent Information Accumulation model to explain how stimulus presentation influenced cognitive processing.


Keywords: dilution effect; information integration; models

## Introduction

Information integration is the combining of evidence from multiple sources. Many tasks, from speech comprehension to medical decision making, require such integration. Each source of information on its own provides some evidence, but integrating all information yields the best performance. This article investigates the manner in which information is combined. While the literature provides numerous examples of near-optimal information integration, there are just as many examples where it is far from optimal. The types of stimuli and procedures used often determined the pattern of results. Tasks involving quantitative stimuli, like probability judgments (Tversky \& Kahneman, 1974), seem to implicate heuristic strategies, more than perceptual tasks, like speech comprehension (Oden \& Massaro, 1978). Our research tests the hypothesis that even perceptual information can produce suboptimal integration if it is displayed in a way that discourages automatic processing.

## Suboptimal Integration in Decision Making

Many studies of judgment and decision making suggest that information from multiple sources is integrated via simple heuristics. Sometimes these studies produce behavior approaching optimal decision making (Gigerenzer \& Todd, 1999), but in many other cases, performance is well short of
optimal (Gilovich, Griffin, \& Kahneman, 2002). The conjunctive fallacy, unpacking effects, and the dilution effect are just a few of the many common findings that violate normative models of information integration. These deviations from rational behavior are so numerous that it is now common to assume sub-optimal integration as a starting point for theories of decision making.

## Optimal Integration in Perceptual Domains

In contrast, there are numerous successful applications of optimal or rational models of information integration in domains such as perception (Oden \& Massaro, 1978; Tenenbaum, 1999), categorization (Ashby \& Maddox, 1990, 1992; Nosofsky, 1986), and memory (Anderson, 1991; Shiffrin \& Steyvers, 1997). Researchers in these more perceptual fields begin with an assumption of optimal integration and only later investigate sub-optimal or heuristic-based strategies.

## The Dilution Effect

Although information integration is an object of study by researchers in both decision making and perceptual domains, these fields often seem to operate independently of each other. One reason is the difference in experimental paradigms. Decision making research focuses mainly on linguistic and quantitative stimuli, and is concerned with how individuals use information to form explicit inference or preferences. Perceptual research typically relies on more perceptual stimuli, and concentrates on how the information is produced from external stimulation. Even when words are used as stimuli, as in memory research, the focus of information integration often includes perceptual aspects of the stimuli. The present research aims to bridge this divide through a novel experimental paradigm that combines aspects of each research tradition.

We focus on one example of sub-optimal information integration: the dilution effect. This effect refers to a situation where adding null or weak positive evidence to what is already strong positive evidence reduces the overall belief in a hypothesis. The effect has been replicated in numerous studies (LaBella \& Koehler, 2004; McKenzie, Lee, \& Chen, 2002; Nisbett, Zukier, \& Lemley, 1981; Peters \& Rothbart, 2000), but Shanteau (1975) gives one of the clearest demonstrations of the dilution effect. In his study, an experimenter drew samples of red (R) and white (W)
beads, with replacement, from one of two boxes. Box A was $70 \% \mathrm{~W}$ and $30 \% \mathrm{R}$. Box B was $30 \% \mathrm{~W}$ beads and $70 \% \mathrm{R}$. The participants did not know from which box the beads were drawn. In one condition, the experimenter drew the sequence WWWRWR from one of the boxes. After every two beads, participants estimated the probability that the beads came from Box A. The mean judgments after WW, WWWR, and WWWRWR were $69.3 \%, 64.0$, and 60.6 , respectively. The WW sample provides diagnostic information, information that clearly points to Box A. However the subsequent samples were nondiagnostic; they could have come from either box with equal probability, and should not have changed the estimated likelihood that the entire sequence came from Box A. Yet this non-diagnostic information caused the estimated probability to drop.

## Why Faces?

Although the dilution effect has only been explored using traditional judgment and decision making stimuli, it easily lends itself to perceptual stimuli. We use weak and strong evidence from different parts of a face to investigate the effect. For example, imagine you are asked to identify a face captured on a security camera. The top half of the face is relatively clear, but the bottom half is in shadow and harder to see. The top and bottom halves of the face then lend strong and weak evidence to the decision. The primary goal of this research is to determine whether the information from these sources is combined in an optimal fashion, or sub-optimally as exemplified by the dilution effect.

A benefit of using perceptual stimuli is that issues of interpretation and language understanding do not come into play. For example, the conjunction law is violated less if participants interpret "Linda is a bank teller" to mean that she is a bank teller and not a feminist (Sides, Osherson, Bonini, \& Viale, 2002). People also often misinterpret probabilities, but perform more optimally when information is presented as frequencies (Gigerenzer \& Hoffrage, 1995). The present task employs perceptual stimuli, thereby greatly reducing any undesirable influence of language conventions.

## Testing Models of the Dilution Effect

In addition to testing for the dilution effect, we evaluate three models of information integration. The Simple Bayesian model combines evidence from the two sources of information optimally, according to Bayesian statistical methods, and predicts additive effects. The Averaging model calculates a weighted arithmetic mean of the evidence produced by each source, and always predicts dilution. Finally, we use the Multi-component Information Accumulation model to explain how information is sampled from multiple sources, and accumulates during deliberation. This model accounts for the behavior we observed in our experiment, and provides insight into how stimulus presentation affects information processing.

The goal of our experiment was to replicate the dilution effect using perceptual stimuli and to determine the role of stimulus presentation on performance. In particular, we
tested if images that encouraged more automatic perceptual integration yield reduced dilution effects than images that required more controlled combination of evidence.

In the experiment, participants categorized a test series of faces into two families (Jones or Smith). The test faces were created by morphing together two target faces (representing the patriarch of each family) along a continuum. Different parts of the faces were morphed independently, allowing us to test how individual combined various levels of evidence. In direct analogy to standard work on the dilution effect, the top and bottom halves of a face act as two sources of information. Based on the many studies showing nearoptimal combination of perceptual information it would be natural to expect two halves from the Jones side of the morph continuum to produce even stronger responses in favor of Jones. Alternatively, weak evidence might dilute strong evidence to produce a dilution effect.

To investigate factors controlling the size and reliability of the dilution effect, two manipulations differentially encouraged automatic and controlled integration of information. It is fairly common to distinguish automatic and controlled processing, both in theory and empirical research. Most often automatic processing is assumed to be fast and independent of conscious manipulation, and controlled processing is assumed to be slow and conscious. Automatic processing is usually assumed to be more robust, less prone to large errors, less based on heuristics, and closer to optimal than controlled processing. This line of thinking suggests that the dilution effect is less likely when processing is automatic, and more likely when processing is controlled. We use the automatic/controlled language of Schneider \& Shiffrin (1977) for convenience sake, rather than to make strong claims that information integration is ever entirely automatic or controlled.

In the present experiment we used conditions that manipulated face images so as to bias processing toward or away from automatic processing. In the Together condition the two half faces are shown atop one another, in a normal configuration. Because identification of faces is overlearned, this should promote automatic processing and produce less dilution. That is, weak evidence, when added to strong evidence from the same category, should increase accuracy. In the Split condition the two half faces were separated horizontally. In the Inverted condition the images were displayed upside-down. Because our perceptual systems have rarely needed to recognize split or inverted faces, each half face might be processed separately, with the results later combined using more deliberate strategies. That is, weak evidence should combine less optimally with strong evidence and produce more dilution.

## Method

## Participants

Nineteen students from Indiana University (undergraduate and graduate) were paid $\$ 16$ to participate in this study. All participants had normal or corrected-to-normal vision.

## Stimuli

All of the stimuli used in the experiment were derived from two "target faces" (A and B) selected from the FERET database (Philips, Moon, Rizvi, \& Rauss, 2000). After cropping the image to remove hair and head outline, the faces were warped so that their major facial features aligned. Once the faces are aligned, a morph is essentially a linear combination of the grayscale values of the two faces at each pixel. The cropped areas of the $256 \times 384$ pixels, grayscale images were filled with a sinusoidal grating. Upside-down copies of the two target faces were also made for the Inverted conditions. The four resulting images were used to construct all experimental stimuli.

The experiment began with two short blocks of trials that calibrated morphs levels to the individual. On each trial a half face was presented and participants chose the target that it most closely resembled. The test faces were created by morphing Target $A$ and Target $B$ together along $a$ continuum. Faces favoring $A$ and $B$ were initialized to 94.44\% Target A and 5.56\% Target A, respectively. A staircase algorithm was used to find top and bottom half face morphs for each target and each orientation that produced an intermediate level of accuracy (approximately $72 \%$ ). These morphs became the medium (M) strength half faces, while weak (W) and strong (S) morphs were derived by extrapolation. Weak halves use the morph coefficient halfway between the medium morph and 0.5 . Strong halves used the morph two thirds of the distance between the medium morph and the target.

Having calibrated all morphs levels, test stimuli were created as follows. For each orientation, the W, M, and S top half faces for Target A were crossed with the W, M, and S bottom half faces for Target A. The same procedure was followed for Target B. As a manipulation check, the W and M half faces for A were also paired with the M and W half faces for B, respectively. Whole faces were presented either in a normal configuration (directly above or below the other half face or background) or horizontally split by 60 pixels. The $\mathrm{W}, \mathrm{M}$, and S top and bottom half faces were also presented in isolation with a continuation of the background presented instead of the other half of the face. Pilot testing showed no performance differences between Together and Split half faces, so the latter were omitted. Sample stimuli are shown in Figure 1. This procedure was done separately for upright and inverted faces, yielding 56 test stimuli for each orientation.


Figure 1: Example test faces.

## Procedure

Participants completed two sessions of the experiment on separate days. They were told that they would see a series of faces, each of which belonged to either the Jones or Smith family. They were instructed to use the test face's resemblance to each patriarch to determine the correct family. After several example trials, participants completed two blocks of calibration trials. The first consisted of 72 upright half face trials, interspersed with 48 upright whole face filler trials included to discourage strategies tailored to half faces. Auditory feedback was given after each response, with a high beep for correct and a low beep for incorrect.

After calibration, participants began an integration phase consisting of two blocks of trials in Session 1 and six blocks in Session 2. Each block contained 68 trials. Each test face appeared once per block, with the exceptions of W/W, M/M, and S/S stimuli, which appeared twice. Upright faces appeared in odd numbered blocks. Inverted faces appeared in even numbered blocks.

Each trial began with a test face appearing in one of nine random positions near the middle of the screen. After two seconds the face was masked with one of two scrambled sets of features from the target faces. After 250ms the mask disappeared and the two target faces appeared, one on each side of the screen. Participants chose the family to which the test face belonged. They were then asked, "What is the likelihood that you are correct?", and responded on a 6point scale from $50 \%$ to $100 \%$. A fixed number of points were awarded for each correct choice and the individual with the highest final score received a $\$ 20$ bonus.

## Results

The present analysis focuses on participants' choice proportions, though mean confidence judgments showed a similar pattern of results. We began by removing data from trials in which individuals indicated no confidence in their decision (likelihood judgment of $50 \%$ ), or responded too fast (less than 150 ms ), or too slow (greater than 5 sec ). This procedure removed approximately $12 \%$ trials, across all participants.

Next, we labeled morphs according to the accuracy they produced on half face trials. That is for example, an individual's half face trials determined which Jones top half morphs were strong, medium, and weak. This relabeling proved unnecessary in most cases because accuracy order matched the physical morph order.

A choice response was considered correct if the test face provided stronger evidence for that target than the alternative. For half faces and most whole faces (i.e. those where top and bottom both favored the same target) this was straightforward. On trials where top and bottom halves favored opposite targets the stronger of the two halves indicated the correct response.

Orientation had almost no effect on accuracy, confirming that calibration successfully equated upright and inverted half face morphs strengths. There were also no significant effects of orientation on the dilution effect, so we present
results collapsed across upright and inverted orientation in order to concentrate on evidence level and split. Mean accuracy, collapsed across target, half (top vs. bottom), and orientation is shown in Figure 2. Accuracy tends to increase with evidence strength, providing a coarse check that the stimuli were appropriately calibrated. Accuracy with M/oW faces was below that of even the weak half faces, confirming that these opposite halves were indeed taken as evidence for the alternative category.


Figure 2: Mean accuracy across evidence levels for whole faces (bars) and half faces (lines).

Our primary research question dealt with how people would combine the two halves of a face and when they might show something akin to the dilution effect. To address this question we compared accuracy with each whole face to that with the stronger half alone. Deviation scores were calculated within individuals by subtracting the mean accuracy given the stronger half face from the response (coded as correct or incorrect) given for whole face. A value greater than 0 indicates additive integration, qualitatively in line with the predictions of a simple Bayesian model. A result less than 0 indicates a dilution effect because additional weak positive evidence decreased accuracy. Figure 3 shows mean deviation scores.

A t-test showed mean deviation scores to be significantly below $0, t(6767)=18.61, p<.01$. As expected Split faces produced greater dilution effects than Together faces. A 2 (Orientation) x 2 (Split) repeated measures analysis of variance (ANOVA) confirmed this, with a main effect of Split, $F(1,18)=12.36, M S E=.015, p<.01$. No other effects were significant. Dilution was greatest for W/S faces, where the difference in top and bottom half strengths was largest. Additive effects were largest in the W/W condition, suggesting that some near-optimal information sampling may have occurred.


Figure 3: Mean dilution scores across whole face conditions.


Figure 4: Predicted deviation scores for the McIA model.

## Discussion

These findings are rather surprising given that the dilution effect had not previously been observed in a perceptual context. The bulk of the existing literature suggested that performance would probably resemble that of a nearoptimal integration process, but we found that participants were often less accurate with two pieces of diagnostic evidence than one. Clearly sub-optimal information integration is not limited to the numerical or linguistic stimuli found in traditional judgment and decision making research. Additionally, our results provide insight into how people processed information in the task. As predicted, dilution was greater when automatic perceptual integration of top and bottom halves was made more difficult by splitting the face. Surprisingly, the dilution effect was
present to an equal extent in both Inverted and Upright orientations. Since, inversion was meant to interfere with strategies tailored to upright faces, this may suggest that participants did not treat the stimuli as they would normal faces. In contrast, the split manipulation may have operated at a lower level where splitting disrupted general purpose whole-object automatic processing in either orientation.

## Models of Perceptual Dilution

The obtained pattern of results poses problems for two of the candidate models introduced earlier. The Simple Bayesian model posits that information from top and bottom halves is combined optimally. Consequently, the model predicts additive effects in all conditions, except M/oW, and cannot account for the large dilution effects observed when a W half was paired with a S half.

The Averaging model, on the other hand, assumes that individuals always integrate top and bottom halves of face by taking the average of the evidence produced by each. The model predicts dilution effects whenever top and bottom evidence strengths are unequal, but it predicts deviations scores near 0 for conditions where top and bottom are of the same strength. Thus, the large additive effects in the W/W condition cannot be explained through averaging alone.

As an alternative we propose the Multi-component Information Accumulation model (McIA). This model represents information integration as a process of accumulating evidence to a decision threshold, $\theta$. According to the model, on each trial a participant repeatedly samples information from one of three sources of evidence: the top half, the bottom half, or the whole face. Each sample provides evidence causing preference to move toward one of two decision bounds. These boundaries represent the amount of preference required to make each response. At one moment a sample may favor the Jones response, causing the preference state to take a step toward the Jones boundary. However, the next sample may favor Smith, causing the preference state to step away from the Jones boundary and toward the Smith boundary. In this manner preference evolves as a noisy random walk process until a decision threshold for one response is reached. The model is thus capable of making predictions for both accuracy and response times.

The probability of sampling whole face evidence is a free parameter, $\alpha$, representing the likelihood that the perceptual system would automatically combine the top and bottom halves into a single whole face. Since splitting the halves apart increased the size of the dilution effect, $\alpha$, was estimated separately for Split and Together faces. The probabilities of sampling the top or the bottom half were then each $(1-\alpha) / 2$. The probability of stepping toward the correct decision boundary after a sample is given by the rate parameter, $\delta$. Since the rate of evidence accumulation should vary with stimulus strength, six rate parameters were estimated. These corresponded to W, M, and S morphs for both top and bottom half faces. For half face trials there is only one $\delta$ to sample at each moment. However, on whole
face trials one of three sources of evidence is sampled at each moment. For example, if a stimulus was comprised of a W top and S bottom, the three sources would be $\delta_{\text {Weak } T_{o p} \text {, }}$ $\delta_{\text {Strong }}$ Bottom, and whole face evidence produced by automatically integrating the two halves. If the whole face evidence is sampled, an evidence accumulation rate is calculated as the Bayesian optimal combination of top and bottom rates, assuming independence. This represents the idea that automatic perceptual integration of top and bottom halves produces additional, perhaps configural, evidence for the correct response. The value of $\theta$ proved relatively unimportant for fitting choices, and was arbitrarily set to 10 . In the future we plan to use the McIA model to simultaneously fit choices and response times, which will allow for better estimation of $\theta$.

The best fitting parameters of the McIA model are given in Table 1. Deviation scores based on the model's prediction are shown in Figure 4. The model does a remarkable job of capturing the basic qualitative patterns in the data. It produces dilution effects because deliberation is sometimes driven by the evidence in the weaker half, producing more errors than with the stronger half alone. In the W/S condition this produces very large dilution effects because, for example, $\delta_{\text {Weak Bottom }}$ is much smaller than $\delta_{\text {Strong Top }}$. However, unlike the Averaging model, the McIA model does not always predict dilution. Instead it posits that on some trials the perceptual system automatically combines the top and bottom halves into a configural whole, yielding high accuracy. This explains the additive effects for W/W, as well as the difference between Split and Together conditions. According to the model whole face evidence was sampled 63\% of time for Together faces, but only 37\% of time for Split faces. This supports our hypothesis that separating the top and bottom halves of face encourages more controlled, less optimal strategies.

Table 1: Best Fitting Drift Rate and Attention Parameters of the McIA Model.

| $\delta_{\text {Weak Top }}$ | 0.522 |
| :--- | :--- |
| $\delta_{\text {Medium Top }}$ | 0.549 |
| $\delta_{\text {Strong Top }}$ | 0.586 |
| $\delta_{\text {Weak Bottom }}$ | 0.518 |
| $\delta_{\text {Medium Bottom }}$ | 0.525 |
| $\delta_{\text {Strong Bottom }}$ | 0.594 |
| $\alpha_{\text {Together }}$ | 0.627 |
| $\alpha_{\text {Split }}$ | 0.365 |

## Conclusion

The present results represent a synthesis of two divergent trends in the extant literature. We used the stimuli and procedures of a perceptual categorization study to investigate a central decision making phenomenon. Unlike in many previous studies using perceptual stimuli, we found widespread and reliable sub-optimal integration, in the form of the dilution effect. Informative differences in the size of this effect were also found. The Together condition, which
encouraged automatic face processing, yielded relatively little dilution compared to the Split condition, which encourages more controlled integration.

Note that we do not see processing mode as a binary concept, but rather a continuum between the extremes of fully automatic and fully controlled integration. To the degree that deviation scores were higher for Together conditions than for Split conditions, we posit a greater degree of automatic integration. The McIA model instantiates this idea through a random walk process with three sources of evidence, the top half alone, the bottom half alone, and the whole face, which represents instances where the perceptual system automatically combines the evidence from the two halves. The model explains how processing was modulated by stimulus presentation. Since Together faces were more naturalistic stimuli, participants were able to sample whole face information more often, yielding greater accuracy.

We also found interesting differences in the size of the dilution effect across levels of evidence strength. For conditions where the top and bottom halves were very unequal, significant dilution was observed. The McIA model produces this result by switching attention between top and bottom halves as it repeatedly samples information. Over time, this effectively averages the evidence in each half. In contrast, additive effects were observed in several conditions where top and bottom strengths were equal. The McIA model also predicts this result because averaging the evidence strengths of these two halves (as described above), produces deviation scores near 0 . However, when whole face evidence is sampled, the probability of stepping toward the correct boundary is the Bayesian optimal combination of the two half face $\delta$ values. These whole face samples push accuracy above that of the stronger half alone.

These results pose a serious challenge to the idea that integration of perceptual information is always well described by rational models. The prevalence of dilution effects for even the most natural of stimuli suggests that there is still more work to be done to fully bridge the span between optimal integration in perceptual and sub-optimal integration in judgment and decision making. This work is a first step toward determining the conditions under which sub-optimal information integration is to be expected. In future work we plan to extend this experimental paradigm to investigate other paradoxical phenomena, such as the conjunctive fallacy, the disjunction effect, and availability effects.

## References

Anderson, J. R. (1991). The adaptive nature of human categorization. Psychological Review, 98(3), 409-429.
Ashby, F. G., \& Maddox, W. T. (1990). Integrating information from separable psychological dimensions. Journal of Experimental Psychology: Human Perception and Performance, 16(3), 598-612.
Ashby, F. G., \& Maddox, W. T. (1992). Complex decision rules in categorization: Contrasting novice and
experienced performance. Journal of Experimental Psychology: Human Perception and Performance, 18(1), 50-71.
Gigerenzer, G., \& Hoffrage, U. (1995). How to improve bayesian reasoning without instruction: frequency formats. Psychological Review, 102, 684-704.
Gigerenzer, G., \& Todd, P. M. (1999). Simple heuristics that make us smart. New York: Oxford University Press.
Gilovich, T., Griffin, D., \& Kahneman, D. (Eds.). (2002). Heuristics and biases: The psychology of intuitive judgment. Oxford: Oxford University Press.
LaBella, C., \& Koehler, D. J. (2004). Dilution and confirmation of probability judgments based on nondiagnostic evidence. Memory \& Cognition, 32(7), 1076-1089.
McKenzie, C. R., Lee, S. M., \& Chen, K. K. (2002). When negative evidence increases confidence: change in belief afer hearing two sides of a dispute. Journal of Behavioral Decision Making, 15, 1-18.
Nisbett, R. E., Zukier, H., \& Lemley, R. E. (1981). The dilution effect: nondiagnostic information weakens the implications of diagnostic information. Cognitive Psychology, 13, 248-277.
Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. Journal of Experimental Psychology: General, 115(1), 39-57.
Oden, G. C., \& Massaro, D. W. (1978). Integration of featural information in speech perception. Psychological Review, 85(3), 172-191.
Peters, E., \& Rothbart, M. (2000). Typicality can create, elimate, and reverse the dilution effect. Personality and Social Psychology Bulletin, 26(2), 177-187.
Philips, P. J., Moon, H., Rizvi, S. A., \& Rauss, P. J. (2000). The FERET evaluation methodology for face-recognition algorithms. IEEE transactions on pattern analysis and machine intelligence, 22(10), 1090-1104
Schneider, W., \& Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. Psychological Review, 84(1), 1-66.
Shanteau, J. (1975). Averaging versus multiplying combination rules of inference judgment. Acta Psychologica, 39, 83-89.
Shiffrin, R. M., \& Steyvers, M. (1997). A model for recognition memory: REM--retrieving effectively from memory. Psychonomic Bulletin and Review, 4(2), 145166.

Sides, A., Osherson, D., Bonini, N., \& Viale, R. (2002). On the reality of the conjunction fallacy. Memory and Cognition, 30(2), 191-198.
Tenenbaum, J. B. (1999). Bayesian modeling of human concept learning. In M. S. Kerns, S. A. Solla \& D. A. Cohn (Eds.), Advances in Neural Information Processing Systems (Vol. 11, pp. 59-68). Cambridge, MA: The MIT Press.
Tversky, A., \& Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185, 11241131.

# Cooperation in Prisoner's Dilemma Game: Influence of Players' Social Roles 

Evgenia Hristova (ehristova@cogs.nbu.bg)<br>Maurice Grinberg (mgrinberg@nbu.bg)<br>Iskra Georgieva (issgeorgieva@gmail.com)<br>Milena Borisova (borisova_milena@abv.bg)<br>Department of Cognitive Science and Psychology,<br>New Bulgarian University, 21 Montevideo Street, Sofia 1618, Bulgaria


#### Abstract

The paper aims to extend the findings of a previous study (Grinberg et al., 2012) exploring the impact of social relations on the cooperation in the Prisoner's dilemma game. Relations between players are manipulated by assigning different roles. The roles embodied the four basic types of human relations in line with Fiske's relational models theory (Fiske, 1991): communal sharing, authority ranking, equality matching, and market pricing (players are assigned roles of team mates, chief and subordinate, partners, and opponents, respectively). Cooperation rates, mutual cooperation, mutual defection, and payoffs gained were subsequently analyzed and compared for a series of forty games. As a result we identified that the market-pricing condition is characterized by considerably lower individual and mutual cooperation, higher mutual defection and lower payoff in comparison to the conditions impersonating the remaining three relational types.


Keywords: Prisoner's Dilemma, decision-making, cooperation, social interaction, relational models

## Introduction

## Prisoner's Dilemma Game

Games are formal tools to study social interactions. The Prisoner's dilemma (PD) game is one of the most extensively studied social dilemmas as it is considered to model interactions in many social situations and problems such as overpopulation, pollution, energy savings, participation in a battle, etc. (Dawes, 1980). It is used to study cooperation and conflict in interactions between individuals, groups, and societies (Rapoport \& Chammah, 1965).
In the PD game the players simultaneously choose their moves - to cooperate (C) or to defect (D) - without knowing the choice of the other player. The payoff table for the twoperson PD game is presented in Figure 1. The payoffs of the Prisoner's dilemma game (see Figure 1) satisfy the inequalities $\mathrm{T}>\mathrm{R}>\mathrm{P}>\mathrm{S}$ and $2 \mathrm{R}>\mathrm{T}+\mathrm{S}$. Because of this game structure a dilemma appears, as there is no obvious best move. On one hand, the D choice is dominant for both players - each player gets a larger payoff by choosing D (defection) than by choosing C (cooperation) no matter what the other player chooses. On the other hand, the payoff for mutual defection $(\mathrm{P})$ is lower than the payoff if both players choose their dominated C strategies (payoff $R$ for each player).


|  |  | $\begin{array}{lr} \text { Player } & \text { II } \\ C & D \\ \hline \end{array}$ |  |
| :---: | :---: | :---: | :---: |
|  | C | 3, 3 | 1,4 |
|  | D | 4, 1 | 2, 2 |

Figure 1: Payoff tables for the PD game with standard notation for the payoffs and an example with specific payoff values. In each cell the comma separated payoffs are the Player I's and Player II's payoffs, respectively.

As the PD game is used as a model for describing social dilemmas and studying the phenomenon of cooperation, there is a great interest in the conditions that could promote or diminish cooperation. In formal game theory players are supposed to try to maximize their payoffs in a completely selfish manner (Colman, 2003). From this point of view the dominant strategy in the game is defection (in one-shot or in repeated PD games with a fixed and known number of games). This prediction is in contrast with the behavior of the players observed in laboratory settings or in real life situations.

In human societies, people cooperate all the time and often cooperation is seen as one of the foundations of human civilization (see e.g. Gärdenfors, 2003). Sally (1995) provides a meta-review of the experiments involving PD games published between 1958 and 1995 and shows that in its iterated version (the game is played many times), cooperation choices are made in 20-50 \% of the games (mean $47.4 \%$ ) and even in one-shot games many players cooperate, although much less than in the iterated version.

Several studies have shown how cooperation can emerge from expected utility or anticipatory reinforcement models without any specific relations between the players (see e.g. Grinberg, Hristova, \& Lalev, 2010; and the references there in).

Other theories explain the cooperative behavior in PD games in terms of socially established values and stress the importance of social interaction and relationships. Reputation building theory (Kreps et al., 1982; Andreoni \& Miller, 1993) assumes that the player is building himself a reputation of a cooperative player to build herself the image of trust and thus to provoke cooperation by the other player. Trivers (1972) puts accent on reciprocity as a widespread norm and basis of societies: people reciprocate cooperation
with cooperation. Another influential theory about cooperation in PD game is based on the concept of altruism. It assumes that some players are not strictly self-interested and from an altruistic perspective, cooperation can yield higher payoffs than defection (Cooper et al., 1996).

Although these social theories of cooperation have been proposed to explain cooperative behavior unexpected by normative game theory, it is interesting to consider more general social theories that are more closely related to the game theoretic analysis of social relations. In our opinion such a theory is the relational models theory proposed by Alan Fiske (Fiske, 1991) which is trying to decompose any social interaction to four basic relations and thus seems amenable to game theoretic representation.

Moreover, as the PD game is central in the modeling of social interactions it can be used to explore the existence and limits of the relational social types as posited by relational social models (see e.g. Haslam, 2004). Exploring the potential of games like the PD game as modeling relational types is one of the goals of this paper which is a continuation of a first analysis presented in Grinberg, Hristova, \& Borisova (2012).

## Relational Models Theory

Relational models theory (Fiske, 1992; Fiske \& Haslam, 1996; Rai \& Fiske, 2011) states that there are four basic schemas that are used to build, organize and maintain relationships and interactions among individuals in a society. These models are supposed to be universal and all relations could be described by these models or by combination of them. The four types of relations generate four modes for every aspect of the interactions between people - resource allocation, moral judgments, decisionmaking, etc. These four relation models are the following (Fiske, 1992):

- Communal Sharing - relations in an undifferentiated group of people with equivalent status. Everyone in a community - which could consist of two members or could be very large - has some rights and some duties. The focus is on commonalities and not on distinctions;
- Authority Ranking - implies an ordinal ranking in society and this ranking scheme determines one's relative status. For instance, military hierarchy can be considered a prototype of such relations;
- Equality Matching - relations are based on a model of one-to-one correspondence as in turn-taking, tit-for-tat strategies, etc. The social prototype would be friendship networks, in which reciprocity is a norm which rules the distribution of wealth;
- Market Pricing - based on a model of proportionality in social relations in which people reduce their interaction to some ratios of utility measures. Examples of relations of this type are the ones governed by prices, rational calculations, expected utilities, etc.


## Social Interactions and Cooperation

In formal game theory payoffs, strategies and choices are analyzed independently from any context or meaning. Most experiments for studying PD game employ neutral presentation of the game. I order too be able to control for extraneous variables, game is presented in neutral formulation, choices are labeled as ' A ' and ' B ', or as ' 1 ' and ' 2 ', etc. Participants in the game are usually called 'players', or 'you' and 'the other' and usually are unfamiliar with one another, in most cases also a visual contact between them is avoided. This is done in order to isolate cognitive processing of information and to capture influence of other factors.

However, we think that deeper understanding of decisions in games should consider social relations involved in the interactions. People behave differently in social interactions described as formal games with similar strategies and payoffs, depending on whom they interact with, what is the situation, what are the possible choices.

Sally (2001) states that social interaction is essential and the social dilemmas like PD need to be investigated from such a perspective. In the paper the importance of closeness between players in game strategy building is discussed. According to this account, players change their choices if they perceive the other player as a friend or a stranger.

Other studies focused on the influence of game description, game title, etc. As such labels and description give different context of the interaction, it is expected that they change the behavior of the players.

Some studies explored the influence of the title given to the game. Ellingsen et al. (2012) found more cooperation when the game is labeled 'Community Game' vs. 'Stock Market Game'. Liberman et al. (2004) found a similar effect in the first round cooperation comparing the game titles 'Community Game' and 'Wall Street Game'.

Another study explored the influence of the general interpretation context (Eiser \& Bhavnani, 1974). Participants cooperate more when they are told that the experiments studies 'international negotiation' or 'interpersonal interaction' compared to 'economic bargaining' or neutral description.

Zhong et al. (2007) manipulated several factors - game label, choices labels, outcome labels and found that giving interpretative labels promotes cooperation and this is especially the case when 'trust' and 'cooperation' are used as labels.

However, in all of these studies the influence of the players' roles is not explored. In all of them the players are are labeled neutrally as 'You' and 'Other person' (Ellingsen et al., 2012; Eiser \& Bhavnani, 1974; Zhong et al., 2007) or as 'Player 1' and 'Player 2' (Liberman et al, 2004). As players' labels and roles could also serve to denote social relations, it is worth exploring their influence on cooperative behavior.

## Relational models and PD game

In a previous study (Grinberg et al., 2012), we made a first attempt to apply the Fiske's relational models theory (Fiske, 1991) to playing in PD games. In the experiment (see for details Grinberg et al., 2012), the relational models between players (communal sharing, authority ranking, equality matching, and market pricing) were operationalized by using various ways of distributing the total payoff gained by a dyad of players in a series of Prisoner's dilemma games: each player receives the total payoff (communal sharing), one of the players receives more than the other (authority ranking), each player receives half of the total payoff (equality matching), each player receives a portion of the total payoff proportional to his/hers individual payoffs (market pricing). For these four conditions, the cooperation rates, the mutual cooperation, the mutual defection, and the payoffs gained were analyzed and compared for a series of forty games. The results of Grinberg et al. (2012) showed that the market pricing distribution scheme leads to less cooperation, less mutual cooperation, more mutual defection and less total payoff than in the other three distribution schemes.

This is an interesting result taking into account the fact that in formal game theory, in many experiments, and in many real life situations, the players are perceived as individualistic beings. It is also evidence that the topic deserves further exploration and has motivated the present study.

## Goals of the Study

The goal of the present study is to explore the mapping of the Fiske's relational models theory to Prisoner's dilemma game focusing on the players' roles corresponding to the four relational models (Fiske, 1992) as follows:

- communal sharing - group of people with strong bonds, wherein everyone is equivalent to the other and all resources are common;
- authority ranking - people are ordered hierarchically and the resources are distributed according to the person's rank;
- equality matching - a balanced relationship based on turn-taking, tit-for-tat strategies and equal distribution of the resources;
- market pricing - relations based on proportionality and comparison - it is important 'how the person stands in proportion to others'
We aim to explore what is the influence of the role assigned to the player on a set of game outcomes that characterize the playing of a PD game - cooperation, mutual cooperation, and mutual defection. It is also important to check the influence of the assigned role on the overall payoffs that are received - e.g. what type of model is more beneficial in terms of payoff earned in interactions shaped by the strategic structure of the PD game.

Based on the results obtained in Grinberg et al. (2012), the cooperation rate is expected to be the highest if the players are acting in a communal sharing relation and the
lowest when the players' roles are defined according to the market pricing model. In the latter scenario, we expect a more individualistic behaviour of the players.

## Method

## Stimuli and Procedure

A sequence of 40 Prisoner's dilemma games is used in the current experiment. All of the games had the payoff matrix given in Figure 2.


Figure 2: Payoff table for the PD game used in the experiment.
Participants were tested in pairs. After receiving the appropriate instructions for the corresponding experimental condition, each dyad played 5 training games (whose results were not included in the analysis) followed by 40 games that were further analyzed. On the computer game interface, the cooperation move was labeled ' 1 ' and the defection move was labeled '2'. Matlab 7.6.0 (R2008a) was used for presenting the game and recording the choiches of the players. After each game the subjects got feedback about their own and the other player's choice and payoffs in the current game. They could also constantly monitor their own total payoff; the total payoff of the other player; and the monetary equivalent of their own total payoff.

The instructions for the experiment explained in detail the rules of the game and included several test questions to ensure that participants understood them correctly. There were five instructions that varied only in the description of the players' roles and the corresponding relations between the players in the game.

The experimenters secured that the participants had not visual, verbal and any kind of other contact between them before and during the experiment. Therefore, no player knew who the other player was before the end of the experiment. Subjects were paid real money accordingly to the final payoff in the game. Each session lasted about 20 minutes.

## Experimental Conditions

The players' roles are varied in accordance with the four relational models described above in a between-subjects design. We also added a control condition, exposed to the most common neutral presentation of the PD game. So, there are 5 experimental conditions as a total differing in how players are labeled in the instructions and on the game interface and how the sequence of games is presented (a sentence in the instruction defines the relations between players)

- Team condition - the players are labeled as 'teammates'; instruction: 'You will play a sequence of
games based on team-work between players with your team-mate' (communal sharing relational model);
- Hierarchy condition - the players are labeled as 'chief' and 'subordinate', correspondingly; instruction: 'You will play a sequence of games based on hierarchy between players with your chief/subordinate' (authority ranking relational model).
- Partners condition - the players are labeled as 'partners'; instruction: 'You will play a sequence of games based on equality (parity) between players with your partner' (equality matching relational model);
- Opponents condition - the players are labeled as 'opponents'; instruction: 'You will play a sequence of games based on competition between players with your opponent' (market pricing relational model).
- Players condition - the players are labeled as 'players'; instruction: 'You will play a sequence of games with the other player' (control condition).
The names for players' roles in each experimental condition are used consistently throughout the experimental session - in the instructions and on the game interface.


## Participants

Forty pairs ( 80 participants) took part in the experiment -8 pairs in each experimental condition. Participants were randomly assigned to their experimental condition. In the hierarchy condition, it was randomly determined which player will be in the subordinate role and which player in the chief role.

Data of one dyad was removed because one of the players reported after the game end that he has participated in a similar experiment. Thus, we ended with 7 pairs in the hierarchy condition and data of 78 participants was analyzed ( 46 female, 32 male, mean age 24 years). For this condition, although the players were asymmetrically labeled (chief and subordinate) the results are not significantly different, so they are analyzed together.

## Results

To explore the influence of the players' roles on choices and cooperation in the PD games, the following dependent variables are analyzed: number of cooperative choices for each player; number of games with mutual cooperation in a pair; number of games with mutual defection in a pair. For clarity, in the figures, the results are presented in percentages. However, the analysis is performed using the specified dependent variables.

The average payoff per game (in points) is considered a measure to assess which players' roles led to higher profits.

Each dependent variable is analyzed in ANOVA with players' roles as between-subject factor with 5 levels (team vs. hierarchy vs. partners vs. opponents vs. players).

## Cooperation

The cooperative choices (\%) are presented in Figure 3.The analysis shows a significant influence of the players'
roles on the number of cooperative moves $(\mathrm{F}(4,73)=3.44$, $\mathrm{p}=0.012$ ).

Post-hoc LSD test shows that the cooperation rate in the opponents condition is significantly lower than the cooperation rate in the team condition ( $p=0.003$ ), in the hierarchy condition ( $p=0.003$ ), and in the partners condition ( $p=0.025$ ). All other differences are nonsignificant.


Figure 3: Average percentage of cooperative choices for different players' roles. ('*' means $p<0.05$ ).

This analysis shows that the labels for the players' roles influence the cooperation rate and lead as expected to lower cooperation for players labeled as 'opponents'. In the terminology of Fiske's theory, the market pricing relational model leads to diminished cooperation in comparison to the other three relational models. While this does not seem strange for the team and partner conditions, it is to some extent for the hierarchy condition. For the latter, however, detailed analysis showed that one pair of players cooperated $100 \%$ of games which led to this strange results which is at odds with the results of Grinberg et al. (2012) for the corresponding condition.

## Mutual Cooperation

Average percentage of games in which there is mutual cooperation (both players have chosen to cooperate) is presented in Figure 4.


Figure 4: Average percentage of mutual cooperation in a pair in each distribution condition ('*' means $p<0.05$; '(*)' - marginally significant difference).

The ANOVA does not identify a statistically significant influence of the players' roles on the number of mutual cooperative game outcomes $(F(4,34)=2.03, p=0.112)$. However, the Post-hoc LSD test shows that a difference exists between the opponents and hierarchy condition ( $p=$ 0.013 ). Marginally significant differences are observed between the opponents and partners condition ( $p=0.09$ ), between the opponents and team condition ( $p=0.09$ ), and between control (players) and hierarchy condition ( $p=$ 0.074).

It turns out that mutual cooperation is the lowest ( $\sim 8 \%$ ) in the opponents condition. This result is consistent with the assumption that the competition, distinctive for the money pricing relational model, will induce an individualistic participants' behavior.

Mutual cooperation is also relatively low in the control condition - the condition with neutral description of the players' roles. The interesting result is again in the hierarchy condition for which the hire mutual cooperation is obtained ( $\sim 31 \%$ ) but as discussed earlier it is partially due to one pair of players which cooperated throughout the whole series of games.

## Mutual Defection

The average percentage of games with mutual defection (both players have chosen to defect) is presented in Figure 5. ANOVA does not identify a statistically significant influence of the players' roles on the number of games with mutual defection $(F(4,34)=2.07, p=0.106)$. However, a further conducted Post-hoc LSD test identifies significant difference between the opponents condition and partners condition ( $p=0.009$ ), and marginally significant difference between opponents condition and hierarchy condition ( $p=$ 0.066).


Figure 5: Average percentage of mutual defection in a pair in each distribution condition ('*’ means $p<0.05$; '(*)' marginally significant difference).

Therefore, it can be concluded that when the players are labeled as opponents, mutual defection is a much more typical choice. It should be noted that mutual defection leads to the lowest possible payoff for the pair. Although defection is the dominant strategy for players in one-shot PD games, here the players play 40 games and mutual defection leads to the worst collective payoff - thus the dilemma structure of the game arises as the opposition between individual and collective
rationality. However, it is interesting to note the high mutual defection in the team condition.

## Average Payoff

The payoff analysis was conducted on the basis of the average payoff per game (in points) (see Figure 6). ANOVA shows a significant influence of the distribution type on the payoff $(F(4,73)=2.50, p=0.049)$.


Figure 6: Average payoff per sequence of 40 games for a pair in each distribution condition ('*' means $p<0.05$ ).

Significant differences were established through post-hoc LSD test between the opponents and team condition ( $p=0.006$ ), between the opponents and hierarchy condition ( $p=0.001$ ), between the opponents and partners condition ( $p=0.001$ ).

The payoff for the participants is lowest when the players are opponents (compared to the other three relational models). This is an interesting result especially since the roles of opponents presumably represent the market pricing relational model, which is related to individualistic attitude and profit orientation. However, taking into account that the highest number of games with mutual defection are found in the opponents condition, the result could be explained by the lower payoff that the players get when they both defect.

## Conclusions and Discussion

The presented study aims at further examining the presumable influence of social relations over cooperative and noncooperative behavioral patterns in the Prisoner's dilemma game. Within our experiment subject were assigned different roles that corresponded to the four basic relations, defined by the Relational models theory: communal sharing, authority ranking, equality matching, and market pricing.

The results outline a clear tendency towards lower individual and mutual cooperation, higher mutual defection and lower total payoff when players are directly labeled as 'opponents' (a role model typical for the market pricing relation) in comparison to all other role sets. Simply put whenever participants are led to perceive 1) the other player as their enemy in the game; and 2) the game as a game of open competition, they cooperate less and earn lower payoff both individually and as a pair. This result, though logical and intuitive in nature, questions the actual success potential of a profit-oriented behavioral model within the Prisoner's dilemma game and real life situations reflecting this game.

As it can be concluded competitiveness may not be the best approach towards goal accomplishment whenever a mutual dependency on participants' choices is present regardless of whether we are facing a person who we deem our opponent.

Strikingly similar results were observed in a previous study (Grinberg et al., 2012) examining the effect of the payoff distribution over the cooperation levels in Prisoner's dilemma game. Lower individual and mutual cooperation, higher mutual defection and lower total payoff were observed when the joint profit was divided among players according to their individual contribution - the experimental condition impersonating the market pricing relation. In comparison, in both experiments the conditions reflecting the remaining three relational models are characterized with higher levels of individual and mutual cooperation and payoff plus lower defection rate. What can be concluded as a summary of both studies is that in line with our expectations, the relational model of market pricing, no matter how framed, "awakens" individualistic, egoistic and concurrent behavioral tendencies among subjects resulting in lower level of cooperation within the Prisoner's dilemma game. Moreover, the influence of these tendencies over individuals seems irrespective of the influence of rationality itself. This in its nature supports the idea that human relations may affect our choice of behaviors in a decisive manner irrelevant of our rational awareness.

An interesting area for exploration remains the condition comprising the authority ranking relation. As it can be seen in both studies this relation could lead to high cooperation despite the different roles of the players or the inequality in the payoffs received. Cooperation levels within the condition are more or less the same as the ones observed in the two "cooperative in nature" conditions - communal sharing and equal matching. It can be thus speculated that inequality does not trigger competition to the extent individualism and "self-sufficiency" do. Therefore such a relation of inequality may not be an obstacle for subjects to perceive the game as a game of an in-team dependency and choose a cooperative behavioral pattern.

The studies conducted produce results with broader implication potential. The research on the effect of human relations on the behavior in social dilemmas is fundamental for the understanding of complex phenomena within the field of both decision making in games and real-life situation in economy, politics, military field etc.

## References

Andreoni, J., \& Miller, J. (1993). Rational cooperation in the finitely repeated Prisoner's dilemma: experimental evidence. The Economic Journal, 103, 570-585.
Colman, A. (2003). Cooperation, psychological game theory, and limitations of rationality in social interaction. Behavioral and Brain Sciences, 26, 139-198.
Cooper, R., DeJong, D. V., Forsythe, R., \& Ross, T. W. (1996). Cooperation without reputation: Experimental evidence from Prisoner's dilemma games. Games and Economic Behavior, 12,187-218.

Dawes, R. (1980). Social dilemmas. Annual Review of Psychology, 31, 169-193.
Eiser, J. R., \& Bhavnani, K. K. (1974). The effect of situational meaning on the behaviour of subjects in the prisoner's dilemma game. European Journal of Social Psychology 4 (1): 93-97.
Ellingsen, T., Johannesson, M., Mollerstrom, J., Munkhammar, S. (2012). Social framing effects: Preferences or beliefs? Games and Economic Behavior, 76 (1), pp. 117-130
Fiske, A. P. (1991). Structures of social life: The four elementary forms of human relations: Communal sharing, authority ranking, equality matching, market pricing. New York, NY: Free Press.
Fiske, A. P. (1992). The four elementary forms of sociality: framework for a unified theory of social relations. Psychological Review, 99, 689-723.
Fiske, A. P., \& Haslam, N., (1996). Social cognition is thinking about relationships. Current directions in psychological science, 5, 143-148.
Gärdenfors, P. (2003). How Homo became Sapiens. Oxford University Press.
Grinberg, M., Hristova, E., \& Borisova, M. (2012). Cooperation in Prisoner's Dilemma Game: Influence of Social Relations, Proceedings of CogSci 2012.
Grinberg, M., Hristova E., \& Lalev, E. (2010). Models for cooperative decisions in prisoner's dilemma. In: Nefti, S. \& Gray, J. (Eds.), Advances in Cognitive Systems. IET: London.
Haslam, N. (Ed.) (2004). Relational models theory: a contemporary overview. Mahwah, NJ: Erlbaum.
Kreps, D., Milgrom, P., Roberts, J., \& Wilson, R. (1982). Rational cooperation in the finitely repeated Prisoner's dilemma'. Journal of Economic Theory, 27, 245-252.
Liberman, V., Samuels, S. M., Ross, L. (2004). The Name of the Game: Predictive Power of Reputations Versus Situational Labels in Determining Prisoner's Dilemma Game Moves. Personality and Social Psychology Bulletin, 30, pp. 1175-1185.
Rapoport, A., \& Chammah, A. (1965). Prisoner's dilemma: a study in conflict and cooperation. Univ. of Michigan Press.
Rai, T. S., \& Fiske, A. P. (2011). Moral psychology is relationship regulation: Moral motives for unity, hierarchy, equality, and proportionality. Psychological Review, 117, 57-75.
Sally, D. (1995). Conversation and cooperation in social dilemmas. A meta-analysis of experiments from 1958 to 1992. Rationality and Society, 7, 58-92.

Sally, D. (2001). On sympathy and games. Journal of Economic Behavior \& Organization, 44, 1-30.
Trivers, R. (1972). The evolution of reciprocal altruism. Quarterly Review of Biology, 46, 35-37.
Zhong, C.-B., Loewenstein, J., Murnighan, J. K. (2007). Speaking the Same Language: The Cooperative Effects of Labeling in the Prisoner's Dilemma. Journal of Conflict Resolution, Vol. 51, No. 3, pp. 431-456.

# When does the majority rule? Preschoolers' trust in majority informants varies by task domain 

Jane C. Hu (jane.hu@berkeley.edu)<br>Daphna Buchsbaum (daphnab@berkeley.edu)<br>Thomas L. Griffiths (tom_griffiths@berkeley.edu)<br>Fei Xu (fei_xu@berkeley.edu)<br>Department of Psychology, University of California, Berkeley, Tolman Hall<br>Berkeley, CA 94720 USA


#### Abstract

In order to learn about the world, young children rely on information provided by social partners. Past research has shown children consider a variety of factors when learning from others, including consensus. Corriveau, Fusaro, and Harris (2009) found that in an object labeling task, children trust responses that receive majority support, and they concluded that children prefer members of a majority as social informants. However, it is possible that children prefer majority members only in domains that rely strongly on socially constructed norms, such as object labeling, where non-social information is unavailable. We formalized this prediction using a rational model of learning from testimony across tasks, and compared our model's predictions to children's responses in object labeling and causal learning tasks. We find that in a causal learning task, a domain that relies less on socially constructed norms, children rely more on their personal observations than informant testimony.


Keywords: social learning; Bayesian modeling; social cognition; consensus; testimony; epistemic trust

## Introduction

We humans are inherently social creatures, and throughout our daily interactions, we openly share our thoughts and opinions with one another. The ubiquity of our social sharing and learning is rare among animal species (Tomasello et al., 2005; Warneken \& Tomasello, 2009), and has been cited as an explanation for the robustness of human culture (Boyd, Richerson, \& Henrich, 2011). Listening to others who share their knowledge can save precious time and effort, as learning through experience can be difficult and time-consuming. In listening to others' testimony, we can instantly and effortlessly learn how to prepare a dish, where to hunt, or who to hire to fix the kitchen sink.

Learning from others is especially important for young children, who have a relatively small pool of life experiences to draw on in new situations. However, one potential drawback to social learning is the possibility of receiving incorrect or misleading information. Therefore, it would be advantageous for children to employ mechanisms to evaluate sources' reliability. Previous work has found that children use informants' past accuracy as an indicator of trustworthiness (Sabbagh \& Baldwin, 2001, Birch, Vauthier, \& Bloom, 2008; Koenig, Clement \& Harris, 2004; Pasquini et al., 2004; Corriveau \& Harris, 2009) and selectively imitate others (Gergely, Bekkering, \& Kiraly, 2002; Brugger et al., 2007; Buchsbaum, Gopnik, Griffiths, \&

Shafto, 2011; Schulz, Hooppell, \& Jenkins, 2008). On the other hand, other studies suggest children's social learning is sometimes surprisingly unselective and irrational (Lyons, Young, \& Keil, 2007; McGuigan \& Whiten, 2009).

We can learn not only from reliable individuals, but also from "crowd sourcing" information from a group of people. Adults often turn to others for advice, assuming that opinions held by many must be valid by virtue of their popularity. This intuition echoes the law of large numbers in probability theory: the more individual testimonies, the more likely the collective conclusion of those testimonies is accurate. Corriveau, Fusaro, and Harris (2009) found that three- and four-year-old children view consensus as an indication of reliability; they were more likely to endorse novel object labels that received majority support, and to choose a member of the majority group as an informant about other object labels. The authors concluded that preschoolers prefer information endorsed by the majority, and prefer members of a majority as informants.

The extent to which children prefer members of a majority as informants is still unclear. One possibility is that children prefer majority members as informants in all situations where multiple testimonies are available. In this view, children would indiscriminately weigh information from others as the most valuable source of information, perhaps prizing it above their own observations. Social psychologists have discovered that consensus opinions can override adults' existing opinions (Asch, 1956; Cialdini \& Goldstein, 2004), which can result in internalization of the consensus opinion (Kelman, 1958; Nolan et al., 2008).

However, if children are rational learners, they should not always prefer majority testimonies. Domain demands should affect the weight children place on others' testimony. When learning about domains that are heavily socially constructed (e.g. object labels or tool use conventions), testimony from others should be highly valuable because the relevant knowledge is transmitted through others, and children cannot learn this type of information on their own. By virtue of the social conventions that dictate object labeling, typically only one label is regarded as correct (Markman, 1989). Alternatively, learning about domains that are not socially constructed (e.g. causal relationships, or naïve physics), children should consider not only informants' endorsements but also their own knowledge, because this type of knowledge can be gained through personal observation and is not typically bound by social
convention. This leaves open the possibility that an effect can have multiple causes.
In this paper, we explore how children's endorsement of majority testimony varies as a function of domain type, and compare these empirical results to rational behavior as predicted by a Bayesian model of learning from testimony. Specifically, we compare children's endorsement of majority testimony in an object labeling task versus a causal learning task. We predict that when given two options - one endorsed by a three-person majority, and one endorsed by a single minority informant - children should be more likely to endorse the majority's testimony when learning socially constructed facts (object labeling) than when learning nonsocially constructed facts (causal learning).

## Modeling Testimony Across Task Domains

In order to rationally learn from others' testimony, children must consider several types of information: the testimonies themselves, their own observations, and social and pragmatic cues that can affect the interpretation of others' testimony. The specific cues and information available to children vary depending on domain, leading them to rely more heavily on pure testimony in some domains. Learning object labels is a task that is especially dependent on social conventions. Speakers of a language must implicitly agree that certain words refer to specific objects, concepts, or ideas (Clark, 1988; 1990) and use them accordingly. In contrast, causal knowledge can be gained through nonsocial cues, like personal experience, that also provide reliable information.

A Bayesian ideal observer model is a natural way to formalize our assumptions about the types of evidence available in these different domains, and about the preexisting biases and pragmatic assumptions that learners may bring to linguistic versus causal inferences. Buchsbaum et al. (2012) developed a model of how a rational learner should make causal inferences from both informant testimony and direct observations of causal outcomes. In this model, the learner receives testimony from one or more informants about the causal efficacy of one or more actions, and may also observe the causal outcome of these actions. The learner's goal is to choose a causally effective action. Here, we adapt this model to compare rational inferences from testimony in object labeling and causal tasks.

## Model Details

Our model for causal inference from testimony is very similar to the model presented in Buchsbaum et al. (2012). In this model, learners receive testimony $r_{c, i}$ from informant $i$ about whether they think candidate cause $c$ is effective. Learners can also directly observe the effects $e_{c, j}$ of those causes (with $N_{c}$ being the number of observations of the effect of cause $c$ ) Each cause $c$ has a true underlying causal strength $w_{c}$, where $p\left(w_{c}=\rho\right)=\gamma$ and $p\left(w_{c}=1-\rho\right)=1-\gamma$, where $\rho$ is a relatively high causal strength value, and $\gamma$ is the probability of a cause having high causal strength. The probability of an effect $e$ following $c$ is $w_{c}$. Each informant $i$
has knowledge about the strength of cause $c, k_{c ; i}{ }^{1}$ We assume that $k_{c, i} \in\{0,1\}$, corresponding to two possible states of knowledge of a cause: knowledgeable and naïve. If $k_{c, i}=1$ (informant $i$ knows about the causal strength of $c$ ), then $p\left(r_{c, i}=w_{c} \mid k_{c, i}=1, w_{c}\right)=1-\varepsilon-$ an informant with knowledge of cause $c$ will give correct testimony about the causal strength of $c$ with probability $1-\varepsilon$, where $\varepsilon$ is a small probability of giving incorrect testimony. In this work, we use $\varepsilon=0.01$. On the other hand, if $k_{c, i}=0$ then $p\left(r_{c, i}=w_{c} \mid\right.$ $\left.k_{c, i}=0, w_{c}\right)=p\left(r_{c, i}=w_{c} \mid k_{c, i}=0\right)=0.5-$ the informant will guess uniformly at random between the two possible actions. The probability of informant $i$ being knowledgeable about a particular cause is $p\left(k_{c, i}=1\right)=\tau$.


Figure 1: Dependencies of the variables in our Bayesian ideal observer model.

Finally, we assume $p($ choose $c) \propto p($ effect $\mid c$, obs) children choose causes in proportion to how likely they think they are to produce the effect, given the evidence. We can use this model to compute the probability that the learner should choose to perform a particular action to bring about the effect, using the dependencies defined in our graphical model shown in Figure 1 (for further details, see Buchsbaum et al., 2012). We can use the same model to infer novel object's labels from testimony. In this case, we have unnamed objects instead of causes, and instead of a causal strength, each object has a probability of corresponding to the novel label. However, unlike the causal case, there are no independent observations to incorporate into the model; you cannot "see" whether a label truly names an object.

## Model Predictions

We can examine a simple contrast between object labeling versus causal learning, where in each task, we assume two possibilities per task: i.e. two objects that could be the referent of a novel label, versus two actions to perform on a toy to elicit music. Corriveau, Fusaro, and Harris (2009) showed children three majority informants making one prediction, and a minority informant making an alternate prediction, so in a similar object labeling task, the majority

[^299]a)

b)


Figure 2: Proportion of responses endorsing majority testimony from (a) model predictions and (b) child data.
may each label Object 1 modi once, while the minority informant labels Object 2 modi three times. In a comparable causal task, the majority may all activate a toy using Action 1, while the minority informant activates the same toy three times using Action 2. Here, statements drawing attention to the demonstrated action are treated as testimony that the action is causally effective. As in the graphical model, the effect of an action is independent of the actor. We can use the model defined above to formalize some of the differences between these two tasks, then examine the model's predictions for whether rational learners should endorse majority testimony.

In labeling objects, we know that there exists a pragmatic mutual exclusivity assumption (Markman, 1989). If an informant labels Object 1 as the modi, this strongly implies that they believe that Object 2 is not a modi. In contrast, using one causal action does not necessarily imply that other actions are ineffective. We can capture this difference by having an informant's testimony that Object 1 is the modi implicitly include testimony that Object 2 is not a modi. In contrast, testimony in the causal case about the efficacy of Action 1 is left neutral with respect to the efficacy of Action 2. Instead, we treat testimony about Action 2 as unobserved for this informant (as are any demonstrations of Action 2 they might have performed).

Our remaining modeling assumptions are similar for both the causal and object labeling tasks. From previous work, we know that children assume that causes are relatively rare - most effects can only be brought about in one or two ways (Buchsbaum et., al, 2011; Bonawitz \& Lombrozo, 2012). Similarly, children generally assume that an object has only one basic-level label (Markman, 1989), so if it is a modi, it is probably not also a toma or a blicket. We can represent both of these prior biases by using a small value for $\gamma$, making multiple causes and multiple labels relatively unlikely. We also know that children are biased to assume that causes are deterministic or near-deterministic (Schulz \& Sommerville, 2006), and similarly that if an object is a modi, it is probably a modi every time, rather than occasionally something else. We can represent both of these assumptions using a high value for $\rho$. Finally, we know that children are a priori biased to assume adults are generally knowledgeable and helpful (Taylor, Cartwright, and Bowden, 1991), which can be represented by using a high value of $\tau$.

We can now look at model predictions for the simple object labeling and causal inference tasks described above. We present predictions using the example parameter values $\gamma=0.05 \rho=0.9$ and $\tau=0.8$ in Figure 2a. However, the qualitative differences in model predictions described below are robust to a wide range of parameter values, and in particular hold for any combination of values consistent with our assumptions. Given object-label testimony from a majority of three informants and one minority informant, the model predicts that learners should strongly favor the majority label. This is true not only if we explicitly force the model to consider only hypotheses where exactly one object is a modi (representing a hard mutual exclusivity constraint), but also if we remove this constraint, but continue to hold the softer pragmatic assumption that an informant who calls one object a modi is also saying that the other is not a modi.

In contrast, in the case where three informants activate a toy one way, and the minority informant activates the toy in another, the model predicts that after observing both actions bringing about the effect equally often, learners should be equally likely to choose either action themselves, despite the conflicting testimony. Finally, we examine a case where informants make causal predictions, but do not demonstrate the actions, paralleling the lack of non-testimony evidence in object labeling. In this case, we do not assume that predicting that one action is effective entails that the other action is not. Here, the model again predicts that the learner should endorse the majority's action choice, but only if they believe causes are rare. If they believe that causes are very common, they should continue to be roughly evenly split. Given our assumption that children are biased to believe causes are rare, we predict that they will again endorse the majority's demonstration in this case.

## Experiment 1: Comparing tasks

In this study, we present preschoolers with four informants' conflicting testimony about objects. In the object labeling condition, informants identify the referent of a novel label, and in the causal learning condition, they demonstrate a novel action on the object that results in a song.

## Methods

Participants Participants were 64 preschoolers, 29 male and 27 female (mean age $=4$ years 2 months; range $=36-$ 65 months). Participants were recruited in the San Francisco Bay Area by mail and phone calls or from local preschools
and museums. An additional five children were tested, but were excluded due to fussiness (4) or experimenter error (1).
Materials In the object labeling condition, stimuli were four novel objects. In the causal condition, stimuli were two plush toys, each of which contained a wireless, batterypowered doorbell chime box. The boxes played short melodies when activated by a handheld remote to create the illusion that children's actions were causally efficacious. Pre-recorded video clips of informants' testimonies were shown to children on a 13 " laptop screen.

Procedure Participants were randomly assigned to either the object labeling condition or causal condition. Each participant participated in two test trials of their condition.

In each condition, the experimenter introduced novel objects to participants and explained that they were unknowledgeable about their labels or causes. Participants then watched four video clips of four informants evaluating the objects.

Each clip began with a female informant sitting at a table with the novel objects. She visually inspected them, then picked up one of the toys and called it by the novel label (e.g. modi), or acted on the toy, resulting in the toy playing a short song. In three of the four video clips, the "majority" informants each endorsed one object as a modi or performed one action to elicit music, and the minority informant informant endorsed the other object the modi, or performed an alternate action to elicit music. The minority informant always repeated the novel label or alternate action three times so that each participant heard the label used to refer to each object an equal number of times.

After participants watched the video clips, the experimenter presented the child with the objects from the video clips and asked children to identify the referent of the novel label, or to make the toy play music. Participants' first gestural or vocal response was recorded. Participants in the causal condition were invited to activate toys three times.

Half of the video clips were mirror images of original recordings to control for the location of objects (object labeling condition) and handedness of informants when manipulating toys (causal condition). The trial presented first and identity of the minority informant were also counterbalanced.

## Results and Discussion

Participants were assigned a score $(0,1$, or 2 ) based on the number of trials in which they endorsed the majority informants' testimony (0-2) first responses in the two trials they participated in (see Table 1).

The distribution of scores in the object labeling condition was significantly different from those in the causal condition, $\chi^{2}(1, N=64)=6.72, p<.03$. The proportion of endorsements of majority testimony over the minority informant's was significantly greater in the object labeling trials $(49 / 64)$ than in causal trials $(32 / 64), \chi^{2}(1, N=128)=$ $8.61, p<.003$ (see Figure 2b). These results closely match
our model's predictions. There were no significant differences in responses based on gender or age (younger vs. older than mean age).

Table 1: Participant scores by condition.

|  | Score |  |  |
| :--- | :---: | :---: | :---: |
| Condition | $\mathbf{0}$ | $\mathbf{1}$ | $\mathbf{2}$ |
| Exp 1: Object labeling task | 4 | 7 | 21 |
| Exp 1: Causal task | 13 | 6 | 13 |
| Exp 2: Causal task (no feedback) | 2 | 3 | 12 |

In the causal condition, participants were invited to activate each toy three times. Not all participants made three attempts, but collectively, participants made a total of 179 attempts to activate toys in the causal trials. Ninety-four attempts ( $53 \%$ of total attempts) were actions performed by majority informants, 84 attempts ( $47 \%$ of total attempts) were actions performed by the minority informant, and one attempt was a novel action performed by none of the informants. All participants in this condition attempted at least one action performed by the minority informant.

As predicted by our model, these data show children were more likely to endorse majority testimony when learning socially constructed facts (object labels) than non-socially constructed facts (cause-and-effect relationships).

Though there was no formal coding scheme for children's spontaneous comments during the study sessions, anecdotal evidence suggests that children's intuitions matched our model assumptions about mutual exclusivity. In the causal condition, children's comments suggested they accepted both the majority testimony and the minority informant's testimony ("Both [actions] make it go!"). Furthermore, all children in this condition attempted an action performed by a minority informant at least once in the study, suggesting that children were open to multiple possibilities when learning about cause and effect. In the object labeling condition, however, several children expressed the belief that there was only one correct answer ("That one isn't the modi!" about the minority-endorsed object).

A possible alternative explanation of these results is that children in the causal condition did not use information gained through their observations or informants' testimony; rather, they were simply confused by the task and randomly imitated informants' responses. To rule out this possibility, we designed another causal condition in which we expected children to endorse the majority testimony.

## Experiment 2: The effect of feedback

In the causal condition of Experiment 1, children indiscriminately imitated the majority and minority informant actions, presumably because they were able to rely on their own observations, which suggested both demonstrated actions were equally effective at activating the toy. This second experiment examines how children behave when they do not have their own observations to rely on, but instead only have information from informants. We predict
that when children lack personal observations indicating the efficacy of informants' testimony, they will be more likely to endorse the majority's testimony.

## Participants

Participants were 17 preschoolers, 7 male and 10 female (mean age $=4$ years 4 months; range $=40-62$ months). Participants were recruited in the San Francisco Bay Area by mail and phone calls or from preschools. An additional three children were tested, but excluded due to fussiness.

## Materials and Procedure

The materials and general procedure of Experiment 2 were identical to those used in Experiment 1. The crucial difference between the two experiments was the content of the video clips participants watched. While in Experiment 1, children watched informants in the video clips perform actions that resulted in the toy playing music, in Experiment 2, informants in the video clips only mimed the actions they endorsed, and no music played as a result of miming the actions. In other words, children who viewed the Experiment 2 video clips received no information about the efficacy of the informants' testimony.

The script of the videos also differed from Experiment 1. Unlike the informants in Experiment 1, who did not verbally describe the action they performed, informants in the Experiment 2 video clips explicitly described their endorsed action and its hypothetical causal effect before miming the action, in order to provide context to children about why the action was being mimed: "It plays music if you pull the pink one!"

## Results

As in Experiment 1, Experiment 2 participants were given a score of 0,1 , or 2 based on their first responses (see Table 1). The number of first responses endorsing majority informants' testimony was significantly higher in Experiment 2 than in the Experiment 1 causal condition, $\chi^{2}(1, N=94)=6.84, p<.008$, and was correctly predicted by our model (see Figure 2b).

Participants in Experiment 2 were also invited to attempt to activate each causal toy three times. Participants collectively made 99 attempts to activate the toys. Of those attempts, 64 ( $65 \%$ of total attempts) were actions performed by majority informants, and 35 ( $35 \%$ of total attempts) were actions performed by the minority informant. Unlike in Experiment 1, where all causal condition participants attempted at least one action performed by the minority informant, four of the 14 participants in Experiment 2 attempted only actions endorsed by the majority.

## General Discussion

In this set of studies, we found that children do not indiscriminately endorse majority opinions; rather, their endorsement of majority opinions varies by task domain type and availability of alternate sources of knowledge. In

Experiment 1, children were significantly more likely to endorse majority testimony when learning about socially constructed facts (object labels) than non-socially constructed facts (causal relationships). Experiment 2 found that in the absence of information about the efficacy of informants' actions, children endorse majority testimony. Children's responses were predicted by a Bayesian model, suggesting that children make rational inferences from informants' testimony, and, when available, weigh other sources of information (e.g. personal observations) more heavily than testimony.

Though these results suggest children consider different sources of information in a non-socially-constructed domain, it is unknown whether they would do so in a socially constructed domain. In the causal conditions, the amount of feedback (i.e., hearing the toy play music) children received about actions' effects was easily quantifiable; however, it is less straightforward what would demonstrate positive or negative feedback about informants' endorsements in a object labeling condition. Future studies could explore how to convey feedback in an object labeling condition - perhaps showing successful or unsuccessful communication achieved through using the label - and the effect it would have on children's inferences.

Follow-up studies could also examine the effects of the informants' language in Experiment 2. In designing Experiment 2, we tried to make informants' video demonstrations as natural as possible while maintaining a similar script to Experiment 1, but creating parallel conditions proved difficult. Recall that the informants in Experiment 1 video clips performed their endorsed action without naming their actions. The informants in Experiment 2 narrated their actions and those actions' hypothetical effect ("It plays music if you pull the pink one!") to explain why informants were miming actions and to present possible actions for children to attempt at test. The hypothetical language used could imply to children that informants had prior experience with or knowledge about the toy. This prior knowledge, combined with informants' explicit demonstration, could be interpreted as evidence that informants were acting pedagogically, or upholding a social norm. Future work could examine how much of Experiment 2's effect was driven by children's lack of access to personal observations, versus pedagogical effects or social norm adherence.

Another difference between the conditions is the number of objects used in object labeling and causal conditions. In each of the causal conditions, participants saw the informants perform one of two actions on a single object, but the object labeling condition, participants saw informants call one of two objects by a novel label. A follow-up study to the object labeling condition could feature video clips with informants calling one object by two names, so that procedures of the object labeling and causal conditions would be more parallel.

In Experiment 1, we found that children's endorsement of majority testimony varies by task domain, but the flexibility
with which children incorporate conflicting information has yet to be determined. Accepting majority testimony as universally informative could potentially mislead a learner; individual members of the majority opinion could be mistaken, or the majority opinion as a whole could be flawed (see Esser, 1998 for a review on groupthink). Future work can identify the cues used to identify a reliable or unreliable majority. Children could discount informant testimony for rational reasons - for instance, if an informant is unreliable or unknowledgeable - or for less rational reasons - for instance, bias against out-group members (Kinzler \& Spelke, 2011).

Additionally, the nature of the beliefs underlying children's endorsements has yet to be examined. Children may endorse majority testimony in the moment in order to conform to societal norms, but do not truly believe this testimony correct. In social psychology, this is called compliance. It is also possible that children internalize this new social knowledge and believe it to be true. Future studies could explore whether children are merely complying with social norms in similar object labeling tasks, or whether they internalize the majority's testimony. Children could be asked to teach others the names of objects, or to recall objects' novel labels in sessions hours or days later.

Overall, the similarity between our model's predictions and empirical data suggest that young children are discerning when considering others' testimony; the extent to which they prefer majority testimony is dependent on task domain type. This work also adds to the growing body of literature that suggests children consider information from multiple sources to make rational inferences.

Acknowledgments This work was supported by grant number 0845410 from the National Science Foundation, and two National Science Foundation Graduate Research Fellowships.

## References

Asch, S. E. (1956). Studies of independence and conformity: a minority of one against a unanimous majority. Psychological Monograms, 70(9).

Birch, S. A. J., Vauthier, S. A., \& Bloom, P. (2008). Three-and-four-year-olds spontaneously use others' past performance to guide their learning. Cognition, 107(3), 1018-1034.

Bonawitz, E.B, \& Lombrozo, T. (2012) Occam's Rattle: Children's use of simplicity and probability to constrain inference. Developmental Psychology, 48, 1156-1164.

Boyd, R., Richerson, P. J., \& Henrich, J. (2011). The cultural niche: Why social learning is essential for human adaptation. Proceedings of the National Academies of Sciences, 108(2), 10918-10925.

Brugger, A., Lariviere, L. A., Mumme, D. L., \& Bushnell, E. W. (2007). Doing the right thing: Infants' selection of actions to imitate from observed event sequences. Child Development, 78(3), 806-824.

Buchsbaum, D., Gopnik, A., Griffiths, T. L., \& Shafto, P. (2011). Children's imitation of causal action sequences is influenced by statistical and pedagogical evidence. Cognition, 120(3), 331-340.

Buchsbaum, D., Bridgers, S., Whalen, A., Seiver, E., Griffiths, T. L., \& Gopnik, A. (2012). Do I know that you know what you know?

Modeling testimony in causal inference. Proceedings of the 34th Annual Conference of the Cognitive Science Society.

Cialdini, R. B., \& Goldstein, N. J. (2004). Social Influence: Compliance and Conformity. Annual Review of Psychology, 55, 591621.

Clark, E. V. (1988). On the logic of contrast. Journal of Child Language, 15, 317-335.

Clark, E. V. (1990). On the pragmatics of contrast. Journal of Child Language, 17, 417-431.

Corriveau, K. H., Fusaro, M., \& Harris, P. L. (2009). Going with the flow: Preschoolers prefer nondissenters as informants. Psychological Science, 20(3), 372-377.

Corriveau, K. H. \& Harris, P. L. (2009). Choosing your informant: weighing familiarity and recent accuracy. Developmental Science, 12(3), 426-437.

Esser, J. K. (1998). Alive and well after 25 years: a review of groupthink research. Organizational Behavior and Human Decision Processes, 73(2-3), 116-141.

Gergely, G., Bekkering, H. \& Kiraly, I. (2002). Rational imitation in preverbal infants. Nature, 415, 755.

Kelman, H. C. (1958). Compliance, Identification, and Internalization: Three Processes of Attitude Change. The Journal of Conflict Resolution, 2(1), 51-60.

Kinzler, K.D., \& Spelke, E.S. (2011). Do infants show social preferences for people differing in race? Cognition, 119, 1-9.

Koenig, M., Clement, F., \& Harris, P. (2004). Trust in testimony: Children's use of true and false statements. Psychological Science, 15(10), 694-698.

Lyons, D. E., Young, A. G., \& Keil, F. C. (2007). The hidden structure of overimitation. Proceedings of the National Academy of Sciences, 104(50), 19751.

Markman, E. M. (1989). Categorization and Naming in Children: Problems of Induction. MIT Press.

McGuigan, N., \& Whiten, A. (2009). Emulation and "overemulation" in the social learning of causally opaque versus causally transparent tool use by 23- and 30-month-olds. Journal of Experimental Child Psychology, 104, 367-381.

Nolan, J. M., Schultz, P. W., Cialdini, R. B., Goldstein, N. J., \& Griskevicius, V. (2008). Normative social influence is underdetected. Personality and Social Psychology Bulletin, 34(7), 913-923.

Pasquini, E. S., Corriveau, K. H., Koenig, M., \& Harris, P. L. (2007). Preschoolers monitor the relative accuracy of informants. Developmental Psychology, 43(5), 1215-1226.

Sabbagh, M. A., \& Baldwin, D. A., (2001). Learning words from knowledgeable versus ignorant speakers: Links between preschoolers' theory of mind and semantic development. Child Development, 72(4), 1054-1070.

Schulz, L. E., Hooppell, C., \& Jenkins, A. C. (2008). Judicious imitation: Children differentially imitate deterministically and probabilistically effective actions. Child Development, 79(2), 395-410.

Schulz, L.E., \& Sommerville, J. (2006). God does not play dice: Causal determinism and preschoolers' causal inferences. Child Development, 77(2), 427-442.

Taylor, M., Cartwright, B. S., and Bowden, T. (1991). Perspective taking and theory of mind: Do children predict interpretive diversity as a function of differences in observers' knowledge? Child Development, 62(6), 1334-1351.

Tomasello, M., Carpenter, M., Call, J., Behne, T., \& Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. Behavioral and Brain Sciences, 28.

Warneken, F., \& Tomasello, M. (2009). Helping and cooperation at 14 months of age. Infancy, 11(3), 271-294.

# Visual motion processing and perceptual decision making 

Aziz Hurzook (ahurzook@uwaterloo.ca)<br>Oliver Trujillo (otrujill@uwaterloo.ca) Chris Eliasmith (celiasmith@uwaterloo.ca)<br>Centre for Theoretical Neuroscience, University of Waterloo<br>Waterloo, Ontario, Canada N2L 3G1


#### Abstract

Perceptual decision making is a fundamental cognitive process widely studied in the behavioural sciences (Gold \& Shadlen, 2007; Wang, 2008). We present a novel, biologically plausible model of visual motion processing and perceptual decision making, which is independent of the number of choice categories or alternatives. The implementation is presented in the form of a large-scale spiking neural circuit consisting of three main processes: 1) a velocity filter that uses the principle of oscillator interference to determine the direction and speed of pattern motion using networks of V1 simple cells; 2) a retinotopic representation of motion evidence in the middle temporal area (MT); and 3) competition-less integration of sensory 'evidence' over time by a higher-dimensional attractor network in the lateral intraparietal area (LIP). The mechanisms employed in 1) and 3) are new. We demonstrate the model by reproducing behavioral and neural results from classic perceptual decision making experiments that test the perceived direction of motion of variable coherence dot kinetograms. Specifically, these results capture monkey data from two-alternative forced-choice motion decision tests. We note that without any reconfiguration of the circuit, the implementation can be used to make decisions among a continuum of alternatives.


Keywords: perceptual decision making, continuous decision making, motion processing

## Introduction

An important function of the mammalian brain is the ability to make decisions based on sensory input, and to take action based on these decisions. Organisms are constantly receiving sensory stimuli from their environment, and in order to choose sensible actions, they must sense and accumulate data over time until enough information exists to make a decision.

In this work, we offer two primary contributions in the computational modelling of a classic perceptual decision test. First, we take as our modelling starting point the visual intensity signals falling on the retina, from stimuli like those used in mammalian studies. Second, we show that the structure of the decision task is not relevant to the structure of the percept represented in the association cortex, and propose a novel mechanism to make decisions based on this structure.
A start-to-finish visual motion and perceptual decision circuit. We simulate the essential components of the primate motion perception and decision pathway using biologically plausible techniques at each stage of circuit modelling. From random-dot motion movies we generate burst signals known to occur in LGN (spatiotemporal derivatives of image intensity with noise reduced), the model then extracts velocity (direction and speed) information using a recurrently connected network of V1 simple cells, it then generates maps of optical flow in MT, and finally it integrates this evidence
in LIP using an $n$-dimensional integrator from which the representation of perceived structure emerges, regardless of task structure. Unlike motion energy models and some related proposals (Adelson \& Bergen, 1985; Rust, Mante, Simoncelli, \& Movshon, 2006; Simoncelli \& Heeger, 1998), the velocity selection mechanism we describe shows how recurrently connected spiking neurons can generate the observed spatiotemporal dynamics in V1 simple cells; that is, we show where the phase evolution of separable and inseparable Gabor-like V1 tunings comes from. Also new is our elimination of divisive normalization in the decoding of integrated vector quantities (Simoncelli \& Heeger, 1998), and the use higher dimensional integration in MT. We are not aware of any past spiking neural models that include all of these stages of processing.

Decision making from the temporal integration of structured percepts. Past work employing integrators to explain perceptual decision making assumes that scalar evidence is integrated to a threshold (Wang, 2008). Many separate scalar integrators are proposed to mutually inhibit one another to explain more complex tasks (e.g. deciding between two, four, eight, etc. possible directions of motion). Here, we propose that a single vector integrator can account for any number of directions of motion. The concept of vector addition is simple: when two opposing vectors are added, they cancel; when two similar ones are added, they reinforce. If the vectors are time-dependent, then at any point in the time course of the integration we have the current state of perception (a vector). Thus, 'competition' among alternatives is misleadingthere is no 'race' among 'competing' choice alternatives, as is typical of past models (M. E. Mazurek \& Shadlen, 2003). Moreover, the percept vector is independent of the decision structure. In other words, the number of alternatives (two choices, $n$ choices, a continuum) is irrelevant to the evidence accumulation process. Hence, the DV can be more generally interpreted as the decision radius ('DR', perhaps) of a percept vector evolving through integration in a higher dimensional sphere rather than a point on a line. The percept evolves over time as evidence accumulates, eventually crossing a decision surface ('DS', perhaps, rather than a decision threshold) if enough sensory evidence is accumulated. In the two-alternative forced choice task we use in our simulation, motion signals are integrated in two dimensions $(n=2)$ yet produce a binary decision, without reconfiguration of the circuit.

Our model suggests that the evidence that is accumulating
for perceptual decisions is a task-independent, $n$-dimensional percept structure (a vector) and not simply a task-dependent, one-dimensional category value (or decision variable, 'DV'). Since the percept structure can be interpreted as any timedependent evidence state for any sensory modality, the circuit could provide a more general approach for the analysis of integrate-to-threshold processes. It could thus be applicable to arbitrary decision processes in the brain, of which the motion evidence domain is only one example. In what follows, we provide a summary of the theoretical principles supporting the model, a description of the model itself, experimental details and results.

## Principles of model design

We use the leaky integrate-and-fire (LIF) neuron as our single cell model. The activity of an LIF neuron $a_{i}(J)$ can be thought of as the steady state firing rate of a neuron under a constant current $J$ and is given by

$$
a_{i}(J)=\left[\tau_{r e f}-\tau_{R C} \ln \left(1-\frac{J_{t h}}{J}\right)\right]^{-1}
$$

where $J_{t h}$ is the threshold current of the neuron, $\tau_{r e f}$ is the refractory period for the neuron, and $\tau_{R C}$ is the membrane time constant for the neuron. To reduce computational demands, we focus only on instantaneous firing rates, as opposed to the precise spike time information, using what are known as rate neurons. It has been shown, however, that the same computations can be performed with a slight increase in the number of spiking neurons (Eliasmith \& Anderson, 2003). Neurons in our model are coupled by a model of synaptic dynamics to give rise to biologically realistic dynamics, and hence empirically constrained timing data.

The general modelling techniques we use for building our simulation are collectively called the Neural Engineering Framework (NEF). The NEF is a method for performing large scale computations using any of a variety of simulated neurons (Eliasmith \& Anderson, 2003). The NEF characterizes the encoding of vector values by populations of spiking neurons, and computation of optimal decoders that allow the approximation of linear or nonlinear functions between ensembles of neurons. This allows us to perform arbitrary computations on vector or scalar quantities using simulated neurons. The following paragraphs go on to describe our computational methods and the NEF in more detail.

## Vector representation

Many empirical studies of mammals have found that populations of cortical neurons can encode real-world stimuli (Hebb, 2002). In the NEF, we encode vector-valued stimuli with populations of simulated neurons, or ensembles.

Encoding over neural populations. Each neuron in an ensemble is tuned to receive more ionic current $J$ when responding to a certain stimulus vector $\mathbf{e}_{i}$, known as that neuron's preferred direction vector, and receive less current the further
away the stimulus vector $\mathbf{x}$ is from $\mathbf{e}_{i}$. So given a vector stimulus $\mathbf{x}=\left(x_{1}, x_{2}, \ldots, x_{n}\right)$, we can relate the firing rate of a single neuron in the ensemble $a_{i}$ to the stimulus by

$$
a_{i}(x)=G_{i}[J(x)]=G_{i}\left[\alpha_{i}\left(\mathbf{e}_{i} \cdot \mathbf{x}\right)+J_{i}^{\text {bias }}\right]
$$

where $G_{i}$ is the nonlinear (spiking or non-spiking) function specific to our neuron model, $\alpha_{i}$ is a gain factor, and $J_{i}^{\text {bias }}$ is a background bias current.

Decoding by optimal linear estimation. In addition to being able to encode stimulus values across neural ensembles, we also would like to be able to recover the original stimulus, given an ensemble's firing pattern. Using this method, we can build a representation for arbitrary stimuli with neural ensembles (Eliasmith \& Anderson, 2003). The simplest way to do this is to make the assumption that the stimulus is a linear combination of the neural activities, which turns out to be quite accurate given enough neurons in the representation (Eliasmith \& Anderson, 2003). That is, we assume our stimulus vector $\hat{\mathbf{x}}$ can be represented by

$$
\hat{\mathbf{x}}=\sum_{i=1}^{N} a_{i} \mathbf{d}_{i}
$$

with $N$ being the number of neurons in the ensemble and $\mathbf{d}_{i}$ being a vector of decoding weights for neuron $i$. If we know $\mathbf{x}$, it is possible to find the optimal set of linear decoders $\mathbf{d}$ that minimize the squared error between $\mathbf{x}$ and $\hat{\mathbf{x}}$. This is a common problem in linear algebra, and can be solved as follows:

$$
\begin{aligned}
\mathbf{d} & =\Gamma^{-1} \mathbf{v} \\
\Gamma_{i j} & =\sum_{x} a_{i} a_{j} \\
\mathbf{v}_{j} & =\sum_{x} a_{j} \mathbf{x} .
\end{aligned}
$$

Solving for the optimal linear decoders, $\mathbf{d}$, allows us to recover an estimate of the original stimulus vector given a neural ensemble's activity. As we will see, it also allows us to directly compute the neural connection weights that perform a computation between two or more ensembles.

## Vector transformation

Now that we have defined a way of encoding and decoding stimulus values, we can perform computations between neural ensembles using our encoding and decoding vectors. Suppose we want to have an ensemble $\mathbf{y}$ encode some function of the value another ensemble is encoding, $\mathbf{x}$. i.e. $\mathbf{y}=f(\mathbf{x})$. We simply compute the decoders for $\mathbf{x}$ as above, only substituting $f(\mathbf{x})$ for $\mathbf{x}$ when computing $\mathbf{v}_{j}$. Then in order to encode our desired function, we multiply our new functional decoding weights $\mathbf{d}$ by our encoding weights for population $\mathbf{y}$, yielding a new set of weights between the populations that generate the desired transformation.

$$
\omega_{i j}=\alpha_{j}\left(\mathbf{d}_{i} \cdot \mathbf{e}_{j}\right)
$$

where $\alpha_{j}$ is a gain term associated with neuron $j$. Note that this technique works well for nonlinear functions as well as linear ones, as we are in effect projecting into a higher dimensional space than our representation, effectively turning a nonlinear function into one that is linear in the weight space.

## Population dynamics

The NEF also defines a way of computing functions defined over time, or dynamic functions. Incorporating timedependance is important in understanding and modelling neural responses, since in the real world, neural activity is dependant on time. In general, we describe a linear dynamic function by $d \mathbf{x} / d t \equiv \dot{\mathbf{x}}=A(\mathbf{x})+B(\mathbf{u})$, where $\mathbf{x}$ is the value currently being represented, and $\mathbf{u}$ is an input value from another ensemble.

One useful example of such a function is a twodimensional oscillator, defined by $A=\left(\begin{array}{cc}0 & 1 \\ -1 & 0\end{array}\right)$. To have an ensemble exhibit this behavior, we define a recurrent connection from this population to itself as described in Eliasmith and Anderson (2003). As shown there, it is possible to solve for the connection weights that allow the ensemble to exhibit the desired behavior, allowing for the implementation of arbitrary dynamical systems.

## Visual motion processing and perceptual decision making

The circuit we propose has three main information processing stages: 1) a velocity filter that uses the principle of oscillator interference to determine the direction and speed of pattern motion using networks of V1 simple cells; 2) a retinotopic representation of motion evidence in MT; and 3) competition-less integration of sensory evidence over time by an $n$-dimensional vector integrator in LIP. A schematic circuit diagram is depicted in Figure 1.

## Velocity selection using oscillating networks of V1 simple cells

The extraction of direction of motion employs the oscillator interference (OI) mechanism, depicted in Figure 2. The initial translational motion of an edge in a local region of the visual field is encoded in a burst signal at $t=t_{0}(\phi=0)$ to simulate LGN output. The signal is filtered through an input filter to control the initial phase of the oscillator. The input drives the rotation of the neurally represented state, $\mathbf{x}(t)=(r(t), \phi(t))$, through a progression of Gabor phase angles in the counterclockwise direction, with a rotation period intrinsic to the oscillator. Damping effects cause the neural representation of $\mathbf{x}(t)$ to return quickly to zero without further input. Subsequent input bursts at times $t_{i}$ add vectorially to, and thus interfere with, $\mathbf{x}(t)$. Constructive interference increases $\|\mathbf{x}(t)\|$ while destructive interference decreases it. Thus, if the direction and speed of the edge transiting the input gate of the neural oscillator are sufficiently close to the magnitude and phase of $\mathbf{x}(t)$, a resonance response occurs and $\mathbf{x}(t)$ sustains its magnitude and rotation. High responses from neurons


Figure 1: Unit circuit schematic for perceptual decision circuit. This figure details the circuit associated with each small patch of the visual field indexed by $i$. These units are repeated for each preferred direction, $\theta$. Each cluster of circles shown is a neural ensemble with $N$ LIF neurons. Index $d$ is the dimensionality of the decoded quantity encoded by the ensemble. $T$ is the period of the natural (undamped) frequency of the oscillator. Each MT ensemble pools the activities of several V1 ensembles with the same $\theta$ and $T$; likewise for LIP pooling of MT. The LIP ensemble is an $n$-dimensional integrator whose activity represents the direction of motion vector that emerges as motion evidence accumulates from all directions. In these simulations, $n=2$ as we are testing for the perceived direction of motion in a plane.
tuned to states later in the period indicate strong velocity (direction and speed) correlation for all earlier phase times after $t_{0}$. Summation of the activities of the late-phase neurons from


Figure 2: Velocity selection mechanism based on oscillator interference (OI). The velocity filter is an array of recurrently connected ensembles of direction selective V1 simple cells. The connection weights are determined using the NEF to endow the ensemble with oscillatory phase sensitivity and thus speed selectivity. The system state has components of magnitude and phase, $\mathbf{x}(t)=(r(t), \phi(t))$. The initial (rest) state is $\mathbf{x}(t)=(0,0)$. (1) An initial burst signal from the LGN is triggered by the translational motion of an edge in the receptive field, shown as a bar moving to the right inside the dotted circle, overlapping the input filter. $\mathbf{x}(t)$ begins to increase in magnitude and rotate through the phase angles. (2) Further input bursts at times $t_{1}$ to $t_{4}$ interfere constructively with the system state only if $\mathbf{x}(t) \approx \mathbf{x}\left(t_{i}\right)$. (3) The activity of neurons tuned to phases late in the period will be high only if correlation with visual input is similar earlier in the cycle. The late-phase activities drive an associated direction vector representation in MT. Other V1 oscillators associated with the same patch but tuned to different directions contribute a weight proportional to the component of motion velocity in their preferred direction (bottom, grey arrows).
the oscillator produce a scalar weight of an associated vector represented in a retinotopic field of motion evidence in area MT. This is a generic mechanism that captures motion information from any visual input.

## Motion evidence map in MT

Figure 3 shows time snapshots of sample velocity maps represented in MT. These are depictions of the stimulus motion


Figure 3: Retinotopic velocity maps in MT. Samples of vector read-out (optical flow) maps in MT for a $7 \times 7$ array of receptive fields for times $t=100,120,140 \mathrm{~ms}$ after stimulus input. The response latency was $50-65 \mathrm{~ms}$. Stimulus coherence levels are categorized by column. For all coherence levels, the stimulus produces a distribution of motion responses. The target direction is not obvious from inspection and requires temporal integration.
in the visual field for any number of directions (for clarity we depict eight directions) at the given times. Each point in the $7 \times 7$ array represents the centre of a patch that is the domain of visual signal input to each unit circuit. The scalar output of each V1 oscillator provides the weight of an associated velocity for a given patch in the field. It should be stressed here that no task-dependent categorization of the motion field is imposed.

For complex pattern motion like variable coherence dots, even at high coherence levels ( $50-100 \%$ ), the wide distribution of velocity response maps provides an indication as to why temporal integration is required for the biased direction to emerge.

## Higher dimensional vector integration in LIP

An important contribution of the model is its employment of a higher-dimensional vector integrator. The linear dynamical equation is

$$
\dot{\mathbf{x}}=A \mathbf{x}+B \mathbf{u}(t)
$$

where $A=0, B=I$ (the identity matrix), and $\mathbf{u}(t)$ is the input evidence. Using the NEF we can determine that the recurrent matrix for neurons to implement this dynamical system is

$$
\omega_{i j}=\alpha_{j} \mathbf{d}_{i}(A+I) \mathbf{e}_{j}=\alpha_{j} \mathbf{d}_{i} \mathbf{e}_{j}
$$

where $i$ and $j$ index the same population of neurons. Because the NEF is defined for vector representations, these weights will result in a neural state that represents the integration of information in the dimensionality of $\mathbf{x}$ (in this case $D=2$ ). Multi-dimensional integrators of this sort have been previously employed in neural models of working memory (Singh \& Eliasmith, 2006), but not for decision making.

## Experiment

## Model implementation

The neural system simulation package used to implement the circuit was Nengo, (http://nengo.ca). Table 1 provides the neurophysiological parameters used. A total of $2.9 \times 10^{5}$ spiking LIF neurons were used. The random-dot motion movies were generated using the Psychtoolbox-3 extensions for Matlab ${ }^{\circledR}$ (Kleiner, Brainard, \& Pelli, 2007; Pelli, 1997; Brainard, 1997). The visual input signal was in the preferred directions of the associated V1 oscillators. To simulate thalamic bursting (Butts et al., 2010), temporal derivatives of spatial overlap between the stimuli and oscillator input filter were taken at $2-\mathrm{ms}$ pulse widths.

## Decision test description

We performed a two-alternative, forced-choice, fixed duration test of 1 -second duration, using variable coherence random-dot motion movies for a single patch. The decision threshold value was held fixed and was the only parameter adjusted to fit behavioural data. The length of the percept state vector, when the average success rate of the circuit was $80 \%$, was used as the decision radius (analogous to the decision threshold use by Gold and Shadlen for the same test in monkey trials (Gold \& Shadlen, 2007). The coherence level (motion strength) was lowered progressively, decreasing motion information and stressing the signal-to-noise ratio resolving capability of the circuit. For each coherence level 10 tests were run.

## Results

The model was able to determine direction of motion in the majority of cases down to about 5\% coherency (Figure 5), and showed similar characteristics to data collected from monkeys in (Gold \& Shadlen, 2007). Particularly, as shown in Figure 4, neuron responses in area MT stayed relatively constant over time, with certain neurons showing stronger firing rates when given stronger motion evidence (higher coherency). At the same time, neuron responses in area LIP got stronger over time, particularly when nearing the decision threshold under medium to high coherency. Additionally, as shown in Figure 5, the experimental results relating to the percentage of correct decisions and time taken to make a decision over varying coherency levels were in accordance with experimental data.

## Conclusion

In the TAFC visual decision task we have used to test our model, we have shown the validity the OI velocity selection


Figure 4: Electrophysiology of MT and LIP neurons during the decision task. Recreated from (Gold \& Shadlen, 2007).


Figure 5: Psychometric performance. The circuit can discern motion direction reliably for coherence levels down to $10 \%$, below which it drops to $50 \%$ success (random guess) as motion strength approaches 0 . The disparities in reaction time between our model and the experimental data may be attributable to motor reaction time and other behavioural factors for which we do not account. Monkey data plots recreated from (Gold \& Shadlen, 2007).

|  | Ensemble parameters | Model Value | Biological Value | Reference |
| :--- | :--- | :---: | :---: | :--- |
| V1 | RC constant $\left(\tau_{R C}\right)$ | 20 | $10-20$ | (Shadlen \& Newsome, 1994) |
|  | Post-synaptic constant $\left(\tau_{p s c}\right)$ | 5.0 | $\sim 6.6$ | (Faber \& Korn, 1980) |
|  | Abs refractory period $\left(\tau_{r e f}\right)$ | 2 | $1{ }^{\dagger}$ | (Friedman-Hill, Maldonado, \& Gray, 2000) |
|  | Max firing rate | $100-250$ | $\sim 70-100$ | (Carandini \& Ferster, 2000) |
| $\mathbf{M T}$ | RC constant $\left(\tau_{R C}\right)$ | $10-20$ | (McCormick, Barry W. Connors, \& Prince, 1985) |  |
|  | Post-synaptic constant $\left(\tau_{p s c}\right)$ | 5.0 | $\sim 6.6$ | (Faber \& Korn, 1980) |
|  | Abs refractory period $\left(\tau_{r e f}\right)$ | 5 | $1^{\dagger \dagger}$ | - |
|  | Max firing rate | 100 | $100-200$ | (Felleman \& Kaas, 1984) |
| $\mathbf{L I P}$ | RC constant $\left(\tau_{R C}\right)$ | - | - |  |
|  | Post-synaptic constant $\left(\tau_{p s c}\right)$ | 50 | - | - |
|  | Abs refractory period $\left(\tau_{r e f}\right)$ | 5 | - | - |
|  | Max firing rate | 70 | 70 | (Gold \& Shadlen, 2007) |

Table 1: Neurophysiological parameters used. $\dagger=$ value based on a model estimate. $\dagger \dagger=$ using V1 value. $(-)=$ not available.
mechanism and the effectiveness of integrating a percept vector over time, without any consideration of the number of choice alternatives. The percept vector evolved over time, toward the left or right direction in two dimensions, producing a binary decision. This was due to the nature of the input, the sensory processing and integration mechanisms, and not any imposed task structure. Since the OI mechanism is isometric in the visual plane, identical results would result from forced choice tasks in any direction. We have tested the same model with additional forced-choice options (e.g. 4 and 8), and it performs similarly well (results not shown). Predictably, fewer choice alternatives lead to faster decisions, since the minimum detectable difference in signal level between two alternatives is greater than if that same magnitude were distributed among 8 alternatives.

It is natural for us to consider the percept vector and its temporal integration to a DS in much higher dimensions. The approach we have presented here can likely be applied to higher order sensory or non-sensory decision making that requires integration of evidence over time.

## References

Adelson, E., \& Bergen, J. (1985). Spatiotemporal energy models for the perception of motion. Journal of the Optical Society of America A, 2(2).
Brainard, D. (1997). The Psychophysics Toolbox. Spatial Vision, 10, 433-436.
Butts, D., Desbordes, G., Weng, C., Jin, J., Alonso, J., \& Stanley, G. (2010). The episodic nature of spike trains in the early visual pathway. Journal of neurophysiology, 104(6), 3371-3387.
Carandini, M., \& Ferster, D. (2000). Membrane potential and firing rate in cat primary visual cortex. Journal of Neuroscience, 20(1), 470-484.
Eliasmith, C., \& Anderson, C. H. (2003). Neural engineering: Computation, representation, and dynamics in neurobiological systems. MIT Press.
Faber, D. S., \& Korn, H. (1980). Single-shot channel activation accounts for duration of inhibitory postsynaptic potentials in a central neuron. Science, 208(4444), 612-615.
Felleman, D., \& Kaas, J. (1984). Receptive-field proper-
ties of neurons in middle temporal visual area (MT) of owl monkeys. Journal of Neurophysiology, 52, 488-513.
Friedman-Hill, S., Maldonado, P. E., \& Gray, C. M. (2000). Dynamics of striate cortical activity in the alert macaque: I. Incidence and stimulus-dependence of gamma-band neuronal oscillations. Cerebral Cortex, 10, 1105-1116.
Gold, J. I., \& Shadlen, M. N. (2007). The neural basis of decision making. Annual Review of Neuroscience, 30, 53574.

Hebb, D. O. (2002). The organization of behavior: A neuropsychological theory. In (new ed.). Psychology Press.
Kleiner, M., Brainard, D., \& Pelli, D. (2007). What's new in Psychtoolbox-3? Perception, 36. (ECVP Abstract Supplement)
McCormick, D. A., Barry W. Connors, J. W. L., \& Prince, D. A. (1985). Comparative electrophysiology of pyramidal and sparsely spiny stellate neurons of the neocortex. Journal of Neurophysiology, 54(4).
M. E. Mazurek, J. D., J. D. Roitman, \& Shadlen, M. N. (2003). A role for neural integrators in perceptual decision making. Cerebral Cortex, 3(11).
Pelli, D. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. Spatial Vision, 10, 437-442.
Rust, N. C., Mante, V., Simoncelli, E. P., \& Movshon, J. A. (2006). How MT cells analyze the motion of visual patterns. Nature Neuroscience, 9(11).
Shadlen, M. N., \& Newsome, W. T. (1994). Noise, neural codes and cortical organization. Current Opinion in Neurobiology, 4, 569-579.
Simoncelli, E. P., \& Heeger, D. J. (1998). A model of neuronal responses in area MT. Vision Research, 38(5).
Singh, R., \& Eliasmith, C. (2006). Higher-dimensional neurons explain the tuning and dynamics of working memory cells. The journal of neuroscience, 26(14), 3667-3678.
Wang, X.-J. (2008). Decision making in recurrent neuronal circuits. Neuron, 60.

# What's Up can be Explained by Language Statistics 

Sterling Hutchinson (schtchns@memphis.edu)<br>Department of Psychology/ Institute for Intelligent Systems, University of Memphis<br>365 Innovation Drive, Memphis, TN 38152 USA<br>Max M. Louwerse (maxlouwerse@gmail.com)<br>Department of Psychology/ Institute for Intelligent Systems, University of Memphis<br>365 Innovation Drive, Memphis, TN 38152 USA<br>Tilburg Centre for Cognition and Communication (TiCC), Tilburg University PO Box 90153, 5000 LE, Tilburg, The Netherlands


#### Abstract

Embodied cognition studies have demonstrated that when words found in high physical locations (e.g., bird) are positioned at the top of a screen they are processed faster than when they are positioned at the bottom of the screen. The reverse effect is obtained for words found in low physical locations (e.g., fish). This concept-location facilitation effect has been argued to demonstrate that cognitive processing is fundamentally perceptual in nature. However, questions can be raised with regards to the absolute or relative location of these concept-location words We investigated whether semantic judgments were made with respect to an absolute location on the screen (embodied explanation) or with respect to a relative location in comparison to other words included in the experimental session (statistical linguistic explanation). In a response time experiment we presented participants with physical-location words from existing studies at the top or bottom, top or center, and center or bottom of the screen. For animate words we found a concept location facilitation effect for words presented at the top of the screen, at the center of the screen, and at the bottom of the screen. In addition, however, language statistics explained RTs to center words. Findings indicated that participants made judgments relative to other words on the screen and not relative to their absolute location on the screen, lending support to a statistical linguistic explanation of the findings.


Keywords: concepts; embodied cognition; symbolic cognition; concept-location facilitation; perceptual

## Introduction

Embodied cognition theories state that language is understood through perceptual representations that are grounded in modality-specific somatosensory experience (Barsalou, 1999; Glenberg, 1997; Semin \& Smith, 2008). Words become meaningful only after mentally reenacting external perceptions and experiences associated with that word. Thus, the patterns of neural activity that occur when comprehending a particular word would be similar to those patterns that occur when actually perceiving its referent (Hauk, Johnsrude, \& Pulvermüller, 2004). In other words, according to embodied cognition theories mental representations are couched in the physical and perceptual experiences of the body.

There is a wealth of evidence supporting the embodied cognition account, with evidence showing that when
experimental tasks cue participants to refer to relevant perceptual representations, language processing is facilitated (Glenberg \& Kaschak, 2002; Kaschak et al., 2005; Pecher, van Dantzig, Zwaan, \& Zeelenberg, 2009). For example, Zwaan and Yaxley (2003) demonstrated that when word pairs appeared in their expected physical locations on a computer screen (e.g., ceiling presented at the top of the screen and floor presented at the bottom of the screen), comprehension was faster than when pairs appeared in unexpected physical locations on a computer screen (e.g., floor presented at the top of the screen while ceiling was presented at the bottom of the screen). That is, it is easier to process a word when the expected physical properties of the word match its actual physical properties. Accumulating research like this tends to suggest that individuals rely on perceptual representations, especially in everyday language comprehension.

This embodied cognition account of semantic representations is often contrasted to an amodal (or symbolic) account of cognition, whereby language is represented amodally. A classical symbolic account of language representation argues that semantic information is seated in language and can be derived from relationships that exist between symbols instead of from the mental reenactment of biomechanical and perceptual experiences. In other words, meaning is represented in a linguistic structure within the brain encoded in a formal abstract language, and words are understood from their natural linguistic context instead of from their perceptual features.

Recently, several studies have argued that an extreme symbolic or an extreme embodied cognition account is untenable, and that a more plausible cognitive model includes both perceptual and symbolic processes in language comprehension (Barsalou, Santos, Simmons, \& Wilson, 2009; Louwerse, 2008; 2011a; Paivio, 1986). For instance, Louwerse (2008; 2011a) proposed the Symbol Interdependency Hypothesis. This hypothesis predicts that language encodes the perceptual information we tend to simulate. Consequently, language statistics allows for bootstrapping meaning with only minimal symbol grounding in perceptual experiences. Put differently, according to the idea of symbol interdependency embodied simulations and symbolic relationships are complementary in conceptual processes.

We also know from previous research that language statistics and perceptual simulations explain cognitive processes to different extents under different conditions. For example, linguistic representations are relatively more prominent early during processing whereas complete perceptual representations take longer to generate (Louwerse \& Connell, 2010; Louwerse \& Hutchinson, 2012). Louwerse \& Jeuniaux (2010) found that both task and stimulus influenced whether participants were more likely to rely on linguistic or perceptual information. Thus, findings reporting effects for word pairs attributed to embodied cognition (e.g., Zwaan \& Yaxley, 2003) might likely also be explained by a statistical linguistic account. For example, when participants were asked to make a semantic judgment about word pairs, the statistical linguistic frequency of the word pair best predicted RTs whereas when participants were asked to make an iconic judgment about image pairs, perceptual ratings about the pair better accounted for RTs (Louwerse \& Jeuniaux, 2010). Although both the linguistic and perceptual information about the word pair showed to be relevant in both cognitive tasks, with both verbal and non-verbal stimuli, different types of information were more, or less, important across different conditions.

These studies demonstrate that both language statistics and perceptual simulation must be taken into consideration together. After all, the Symbol Interdependency Hypothesis argues that language encodes perceptual information, making it difficult to disentangle the two variables. That is, effects attributed to statistical linguistic frequencies could also be attributed to perceptual simulation and vice versa. Furthermore, studies demonstrating a language statistics effect use word pairs as stimuli (e.g., Louwerse \& Hutchinson, 2012; Louwerse \& Jeuniaux, 2010; Tse, Kurby, \& Du, 2010).

However, evidence supporting an embodied cognition account also comes from single words, presented in different locations on a computer screen. For example, Šetić and Domijan (2007) presented 'up' and 'down' words one at a time either in an expected physical location or in an unexpected physical location (e.g., butterfly would either appear at the top of the screen (expected location) or at the bottom of the screen (unexpected location)). Participants were asked to determine if the word they saw was something animate (living animal) or something inanimate (non-living entity). As expected, patricipants were faster to process concept-location matches (e.g., butterfly presented at the top of the screen) than conceptlocation mismatches (e.g., butterfly presented at the bottom of the screen). Unlike experiments comparing word pairs, findings for words in isolation, such as those in Šetić and Domijan (2007), are more difficult to also explain with a statistical linguistic account. That is, unigram word frequency does not explain congruency effects, as the set of 'up words' are not all more orless frequent than the set of 'down words'. In fact, when comparing how frequently the 'up words' and 'down words' occurred in a massive corpus of the English language (the Web IT 5-gram corpus; Brants \& Franz, 2006), no difference was obtained between the frequencies of 'up words' and 'down words' inform the

Šetić and Domijan (2007) study, $t(153.37)=0.64, p=.52$. Consequently, the concept-location word results only seem to support an embodied cognition account and are argued to be due to the congruency of the presentation location and the perceptual features of the word: butterfly is processed quickly at the top of the screen because a mental simulation of a butterfly involves perceptual and spatial information about where a butterfly is found in the actual world (above the ground/at the top). This poses a challenge to an account that argues for both linguistic and perceptual simulations factors in conceptual processing, such as proposed by the Symbol Interdependency Hypothesis.
Although it seems straightforward to conclude that these effects must be due to the mental simulation of words, there are alternative explanations. Lakens (2011a; 2011b) argues that such effects might instead be due to polarity correspondence. Proctor and Cho (2006) found that in binary classification tasks, concepts can be processed faster when their polarity matches the response polarity. In other words, when a stimulus and a response are coded as either both positive or both negative, processing is facilitated, e.g., butterfly is processed quickly at the top of the screen because its location is positive (up), as is the response to whether or not it is found in the sky (yes). In order to rule out a polarity correspondence explanation for the results, in a similar experiment, Pecher, van Dantzig, Boot, Zanzolie, and Huber (2010) asked participants to respond to the question Is it usually found in the ocean? or to the question Is it usually found in the sky?. They argued that for a polarity correspondence explanation to be valid, yes responses would be expected to be processed faster at the top of the screen, regardless of the question being asked, and regardless of word meaning. For instance, when being asked if an animal is found in the ocean, one would expect butterfly to be processed faster at the bottom of the screen because it is not found in the ocean, a hypothesis contrary to an embodied cognition explanation and a hypothesis that was not supported. Instead, the results showed just the opposite, i.e., when being asked if the animal is found in the ocean, butterfly was still processed faster at the top of the screen. In a response, Lakens (2011b) still suggested that perhaps butterfly is processed faster at the top of the screen, even when participants are making an ocean judgment because the judgment becomes a relative assessment with down as the default response (as all comparisons are made with reference to the ocean, which is down).
Lakens (2011b) goes further to point out that alternative explanations for data explained solely by perceptual simulations should not be overlooked. In addition, Lakens (2011b) and Louwerse (2011b) both suggest that results from Pecher et al. (2010) might likely also be explained by a statistical linguistic account. That is, although Pecher et al. (2010) concludes that mental simulation accounts for responses in the sky/ocean task, linguistic frequencies do contribute to word meaning and should also be considered. To illustrate, Louwerse (2011b) found that ocean animal names paired with the word ocean occur more frequently than ocean animal names paired with the word sky (and vice versa for sky animal names) and that these frequencies
account for subject RTs. In sum, findings previously attributed to mental simulation accounts can also be explained by a statistical linguistic account, as was also demonstrated in earlier research (Louwerse \& Jeuniaux, 2010). These findings illustrate that task instructions might influence response times because ocean and sky are more or less linguistically associated with the stimuli. In other words, linguistic explanations for these findings should also be explored.

But it remains difficult to offer a linguistic explanation for results when words are presented in isolation. Although task instructions might influence the speeded responses, the frequency of butterfly - sky is only able to account for faster RTs for congruent word categories and tasks while still leaving mental simulations to offer the only explanation for the facilitative effect of the congruency of the presentation location and the perceptual features of the word (as unigram word frequency cannot account for these RTs). Perhaps linguistic information might play a role explaining these concept-location effects for isolated words after all. Although words are presented in isolation on the screen (i.e., one word is presented at a time), it is possible that decisions might be made relative to the other words presented in the other trials of the experiment. Such an explanation would suggest that instead of making judgments relative to the congruency between the concept and the absolute position of the word on the screen (i.e., top of the screen or the bottom of the screen), participants are making judgments relative to the other words in the experiment. That is, participants might show a conceptlocation facilitation effect not because the words are presented on the top and bottom of the screen, but rather because words are asynchronously presented relatively above and below one another throughout the duration of the experiment.

To explore this possibility, in this study we presented participants with isolated words at either the top or bottom (to replicate the original results), top or center, or center or bottom of the screen. According to an embodied cognition account, if responses are faster because word meaning and world location are congruent, we would expect the same high and low words, presented in the center of the screen to show no concept-location facilitation effect because the presentation location is not congruent with the physical and spatial properties of the simulated word. In other words, when butterfly is presented in the center of the screen, processing should not be facilitated.
Alternatively, if decisions are based on the relationship between one word relative to the other words in the experiment (as opposed to being relative to the presentation location of the word; a linguistic explanation), then we might find that high words presented in the center of the screen (concept-location mismatch) will still show a concept-location facilitation effect if low words are presented at the bottom of the screen. That is, when butterfly is presented in the center of the screen, processing will be facilitated if other words in the experiment are 'below' a butterfly. Similarly, we might find that low words presented at the center of the screen would show a concept-location facilitation effect if high words are
presented at the top of the screen. In essence, if conceptlocation facilitation is found when words are presented in relative positions on the screen (i.e., above/below one another) as opposed to absolute positions on the screen (i.e., at the top/bottom of the screen), it might be the case that perceptual simulation (concept-location facilitation effect) is not entirely accounting for RTs but rather, participants are making decisions about words presented in isolation by comparing those words to the group of words included in the experiment.

## Method

## Participants

Eighty-seven undergraduate native English speakers at the University of Memphis participated for extra credit in a Psychology course. Participants were randomly assigned to each of the three conditions (words presented at either a) the top of the screen and the center of the screen, b) the center of the screen and the bottom of the screen, or c) the top of the screen and the bottom of the screen).

## Materials

The experiment consisted of 48 living animal words that could be found in a low spatial location, (such as the ground or ocean, $n=24$ ) or found in the sky (a high spatial location, $n=24$ ). The remaining 48 words consisted of nonliving objects that could also be found in either high ( $n=24$ ) or low ( $n=24$ ) physical locations. Words were extracted from both Pecher et al. (2010) and Šetić and Domijan (2007).

## Procedure

The procedure was almost identical to Pecher et al. (2010) and Šetić and Domijan (2007). Participants were asked if words presented on a $1280 \times 1024$ computer screen were either living or nonliving. This task has the advantage that it does not bias participants to consciously judge the physical location of a word. The center of the screen was positioned at eye level. Similar to Pecher et al. (2010) and Setić and Domijan (2007), each trial began with the presentation of three fixation crosses appearing on the screen for 300 ms . Fixation crosses were presented either at the top, center, or bottom of the screen, depending on where the proceeding word would appear on the screen. This occurred in order to notify participants where the next word would appear.

Words were presented at either the top and the center of the screen, the center and bottom of the screen, or - as in the original Šetić and Domijan (2007) study the top and bottom of the screen, depending upon the between participants condition. Upon presentation of a word, participants indicated whether the word was living or not living by pressing designed counterbalanced keys on the keyboard ( $f$ and $j$ keys). All words were seen once and were counterbalanced for each participant where half the high spatial location words were presented in the upper position (relative to the other presentation location, i.e., top relative to center/bottom or center relative to bottom) and
half in the lower position (i.e., bottom relative to center/top or center relative to top), likewise for the low spatial location words.

If responses were slower than $2,500 \mathrm{~ms}$ a message reading 'TOO SLOW' would appear. Participants were asked to try to be as quick and as accurate as possible in their responses. The next trial began immediately after the subject's response or after the feedback message.

## Results and Discussion

Eleven participants were removed from the analysis because $>40 \%$ of their answers were incorrect. All remaining participants were split evenly between conditions. In all analyses, we used the parameters found in Pecher et al. (2010) for outlier identification and removal. Outliers were identified as those correct responses greater than three standard deviations from the mean per subject per item. Outlier removal (as described above) resulted in a loss of $2.8 \%$ of the data. All error trials were removed, resulting in a loss of an additional $8.7 \%$ of the data.

A mixed-effect regression analysis was conducted on RTs with match/mismatch (match or mismatch between word category (low or high spatial location word) and relative presentation location (relatively high location of top or center or relatively low location of center or bottom)) as a fixed factor and participants and items as random factors (Baayen, Davidson, \& Bates, 2008). The model was fitted using the restricted maximum likelihood estimation (REML) for the continuous variable (RT). F-test denominator degrees of freedom were estimated using the Kenward-Roger's degrees of freedom adjustment to reduce the chances of Type I error (Littell, Stroup, \& Freund, 2002).

In addition to the location presentation manipulation, we investigated the source of the RT differences in this task, linguistic or embodied. An embodied account would be predict a concept-location facilitation effect, whereas a linguistic account would suggest these same effects are driven by language statistics. To further explore if participants were relying on language statistics, we ran analyses using word frequency as a fixed factor to determine if a possible additional explanation for any concept-location facilitation effects may exist. The word frequency factor was calculated as the log frequency of each word being presented obtained using the Web 1T 5gram corpus (Brants \& Franz, 2006).

Unlike Šetić and Domijan (2007), no significant conceptlocation facilitation effect was found for words appearing at the top of the screen, $F(1,2330)=1.46, p=.23$, at the center of the screen, $F(1,1599)=.10, p=.75$, nor at the bottom of the screen, $F(1,2395)=1.76, p=.19$. Just as in Pecher et al., (2010) these findings also fail to replicate the concept-location facilitation effect found in Setić and Domijan (2007). In fact, there was no interaction between location and word category for any of the three word presentation locations and experimental conditions. Pecher et al. (2010) offered the explanation that the concept location facilitation effect is not well understood, with some factors causing facilitation and others causing interference. The linguistic frequency factor did not explain


Figure 1: Average RTs in ms for the words appearing at the top of the screen.


Figure 2: Average RTs in ms for the words appearing at the center of the screen.


Figure 3: Average RTs in ms for the words appearing at the bottom of the screen.
the results either, with no significant main effects for words appearing at the top of the screen, $F(1,2330)=.0001, p=.99$, the center of the screen, $F(1,1599)=.19, p=.66$, nor the bottom of the screen, $F(1,2395)=.11, p=.74$. These current results seem to support neither an embodied cognition account (as there was no concept-location facilitation for the top-bottom condition) nor an alternative linguistic account (as there was no concept-location facilitation for either condition including the center location nor was linguistic frequency significant). In the absence of a replication in both the current study and in Pecher et al. (2010), perhaps the effects reported in Šetić and Domijan (2007) might be attributed to linguistic differences in the Hungarian stimuli. Alternatively, such concept-location facilitation effects might simply be relevant for certain groups of words and not others.

To further explore the results of the current experiment, and the possibility that words are processed relative to the words around them, we analyzed our findings mixed effects model but for animate versus inanimate words. Words that were inanimate again showed no interactions for words appearing at the top of the screen, $F(1,1172)=$. $003, p=.96$ (see Figure 1), the center of the screen, $F(1$, 787) $=.07, p=.80$ (see Figure 2), or the bottom of the screen, $F(1,1072)=.92, p=.34$ (see Figure 3). Linguistic frequency was also not significant for words appearing at the top of the screen, $F(1,1172)=1.53, p=.22$, the center of the screen, $F(1,787)=.62, p=.43$, nor the bottom of the screen, $F(1$, 1072)=.002, $p=.96$.

However, words that were animate did show significant interactions. Words appearing in any given location (top, center, and bottom) were processed faster when that location was relatively the same as the word category. 'Up words' presented in the center were processed faster in the center-bottom condition, whereas 'down words' presented in the center were processed faster in the top-center condition, $F(1,789)=6.10, p<.02$. Figure 2 clearly illustrates RTs for matched and mismatched up and down words presented in the center of the screen, showing that words with a concept-location match are processed faster than words with a concept-location mismatch. Similarly, 'up words' presented in the top of the screen were processed faster in both the top-bottom and top-center conditions, $F(1,1134)=6.80, p<.01$, (see Figure 1). Finally, 'down words' presented in the bottom of the screen were processed faster in both the top-bottom and center-bottom conditions, $F(1,1067)=10.97, p=.001$, (see Figure 3).

In addition, to further explore the impact of linguistic frequency also significantly explained RTs to words presented at the bottom of the screen, $F(1,1067)=5.08, p=$. 02 , but only marginally for words presented in the center of the screen, $F(1,789)=3.22, p=.07$, with no effects for words presented at the top of the screen, $F(1,1134)=2.58$, $p=.10$. These findings seem to be consistent with the idea that decisions are based on the relationship between one word relative to the other words in the experiment, as 'up words' presented relatively above 'down words' still showed a concept-location facilitation effect despite these words being presented in the center of the screen.

In addition, in all conditions, words appearing relatively below other words ( $M=767.45, S D=267.40$ ) were processed significantly slower than words appearing relatively above other words ( $M=889.36, S D=421.41$ ), $t(4926)=15.36, p<.001$. That is, regardless of the absolute location of the word on the screen, where-ever the bottom position was (i.e., center of the screen or bottom of the screen), words presented in that location were processed slower than the same words presented in a relatively higher location. Consider the case of the center presentation location: when words were presented in either the center of the screen or the bottom of the screen, words took longer to process at the bottom and less time to process at the center. However, when those same words were presented in the center or the top, they took longer to process in the center and less time to process at the top. This means that the same words presented in the same location are processed faster or slower simply due to whether other words are appearing above or below them. This at least suggests that comparisons between high and low positions are biased given that the center represents both the relative top and bottom in different conditions.

Finally, to explore whether participants indeed made comparative judgments for words, we assessed whether bigram frequencies were able to account for the response times of center words. As in previous studies (Louwerse, 2008) we operationalized the bigram linguistic frequencies as the $\log$ frequency of a-b (e.g., owl-lizard) or b-a (e.g., lizard-owl) order of word pairs. Because words were presented individually on the screen, pairs were determined by the randomized presentation order. The bigram frequency of each pair was assigned to the second word in the randomly presented pair. The order frequency of all word pairs within 3-5 word grams was obtained using the large Web $1 T$ 5-gram corpus (Brants \& Franz, 2006). A mixed-effect regression analysis was conducted on RTs to center words with the bigram frequency as a fixed factor and participants and items as random factors (Baayen, Davidson, \& Bates, 2008). Bigram frequency was a significant predictor of RTs for center words only, $F(1,906)=3.99, p=.05$. This was true for all center words regardless of experimental condition, implying that participants consider past trials while making judgments about the current word in question, and implying that a linguistic frequencies explain RTs during a conceptlocation facilitation task.

## General Discussion

In three presentation location conditions (top and center, bottom and center, or top and bottom) we failed to replicate a concept-location facilitation effect as found in Šetić and Domijan (2007) for inanimate words. However, when considering animate words, words matched between the relative presentation location and word category resulted in faster RTs than words with a mismatch. This finding suggests that participants make judgments about individual words they see on the screen with respect to other words they see throughout the duration of an experiment. The absolute location of a word on a screen does not seem to impact the concept-location facilitation effect, but rather
the relative location appears to be what is important. This finding suggests that decisions are based on the relationship between one word relative to the other words in the experiment, not only based on the relationship between one word and the embodied physical and spatial properties of that simulated word. In addition, across all three conditions, we found a main effect of location, such that words presented below other words were processed slower. This finding suggested that participants made judgments relative to other words, not only relative to their location on the screen. To further determine whether participants made comparative judgments between words presented asynchronously over the duration of an experiment we also showed that bigram frequencies can predict subject RTs. These findings together indicate that it might be the case that participants are making decisions about words presented in isolation by comparing those words to the group of words included in the experiment, suggesting that findings that are easily attributed to embodied cognition (Pecher et al., 2010; Šetić \& Domijan, 2007) can also be attributed to language statistics.

## References

Baayen, R. H., Davidson, D. J., \& Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language, 59, 390-412.
Barsalou, L. W. (1999). Perceptual symbol systems. Behavioral and Brain Sciences, 22, 577-660.
Barsalou, L. W., Santos, A., Simmons, W. K., \& Wilson, C. D. (2008). Language and simulation in conceptual processing. In M. de Vega, A. M. Glenberg, \& A. C. Graesser (Eds.), Symbols, embodiment, and meaning (pp. 245-283). Oxford, England: Oxford University Press.
Brants, T., \& Franz, A. (2006). Web 1T 5-gram Version 1. Philadelphia, PA: Linguistic Data Consortium.
Glenberg, A. M. (1997). What memory is for. Behavioral and Brain Sciences, 20, 1-55.
Glenberg, A. M., \& Kaschak, M. P. (2002). Grounding language in action. Psychonomic Bulletin \& Review, 9, 558-565.
Hauk, O., Johnsrude, I., \& Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. Neuron, 41, 301-307.
Kaschak, M. P., Madden, C. J., Therriault, D. J., Yaxley, R. H., Aveyard, M. E., Blanchard, A. A., \& Zwaan, R. A. (2005). Perception of motion affects language processing. Cognition, 94, B79-B89.
Lakens, D. (2011a). High skies and oceans deep: Polarity benefits or mental simulation? Frontiers in Psychology, 2, 1-2.
Lakens, D. (2011b). Polarity correspondence in metaphor congruency effects: Structural overlap predicts categorization times for bi-polar concepts presented in vertical space, Journal of Experimental Psychology, 38, 726-736.
Littell, R. C., Stroup, W. W., \& Freund, R. J. (2002). SAS for linear models. Cary, NC: SAS Publishing.

Louwerse, M. M. (2008). Embodied relations are encoded in language. Psychonomic Bulletin \& Review, 15, 838844.

Louwerse, M. M., (2011a). Symbol interdependency in symbolic and embodied cognition. TopiCS in Cognitive Science, 3, 273-302.
Louwerse, M. M. (2011b). Stormy seas and cloudy skies: conceptual processing is (still) linguistic and perceptual. Frontiers in Psychology, 2, 1-4.
Louwerse, M. M., \& Connell, L. (2010). A taste of words: Linguistic context and perceptual simulation predict the modality of words. Cognitive Science, 35, 381-398.
Louwerse, M., \& Hutchinson, S. (2012). Neurological evidence linguistic processes precede perceptual simulation in conceptual processing. Frontiers in psychology, 3: 385.
Louwerse, M. M., \& Jeuniaux, P. (2010). The linguistic and embodied nature of conceptual processing. Cognition, 114, 96-104.
Paivio, A. (1986). Mental representations: A dual coding approach. New York, NY: Oxford University Press.
Pecher, D., van Dantzig, S., Boot, I., Zanolie, K., \& Huber, D. E. (2010). Congruency between word position and meaning is caused by task Induced spatial attention. Frontiers in Psychology, 1, 1-8.
Pecher, D., van Dantzig, S., Zwaan, R. A., \& Zeelenberg, R. (2009). Language comprehenders retain implied shape and orientation of objects. The Quarterly Journal of Experimental Psychology, 62, 1108-1114.
Proctor, R. W., \& Cho, Y. S. (2006). Polarity correspondence: A general principle for performance of speeded binary classification tasks. Psychological Bulletin, 132, 416.
Semin, G. R., \& Smith, E. R. (Eds.). (2008). Embodied grounding: Social, cognitive, affective, and neuroscientific approaches. New York, NY: Cambridge University Press.
Šetić, M., \& Domijan, D. (2007). The influence of vertical spatial orientation on property verification. Language and Cognitive Processes, 22, 297-312.
Tse, C., Kurby, C.A., \& Du, F. (2010). Perceptual simulations and linguistic representations have differential effects on speeded relatedness judgements and recognition memory. The Quarterly Journal of Experimental Psychology, 63, 928-941.
Zwaan, R. A., \& Yaxley, R. H. (2003). Spatial iconicity affects semantic relatedness judgments. Psychonomic Bulletin \& Review, 10, 954-958.

# Influences Beyond Language? A Comparison of Spatial Referencing in Native French Speakers from Four Countries 

Lisa Hüther (lisa.huether@psychologie.uni-freiburg.de)<br>University of Freiburg, Department of Psychology, Engelbergerstrasse 41<br>79085 Freiburg, Germany

Anne Bentz (anne.bentz@psychologie.uni-freiburg.de)
University of Freiburg, Department of Psychology, Engelbergerstrasse 41
79085 Freiburg, Germany

Hans Spada (spada@psychologie.uni-freiburg.de)<br>University of Freiburg, Department of Psychology, Engelbergerstrasse 41<br>79085 Freiburg, Germany

Andrea Bender (bender@psychologie.uni-freiburg.de)
University of Freiburg, Department of Psychology, Engelbergerstrasse 41
79085 Freiburg, Germany
Sieghard Beller (sieghard.beller@uni-paderborn.de)
University of Paderborn, Department of Human Sciences, Warburger Straße 100
33098 Paderborn, Germany


#### Abstract

Research has shown that spatial referencing differs across cultures. Whether "Western" samples, specifically ones speaking the same native language, show the same referencing patterns has not been investigated thus far. Examining spatial referencing behavior across different tasks, we compared samples from four different countries speaking the same language with respect to their application of the intrinsic frame of reference (FoR) and the three variants of the relative FoR. Our findings indicate influences of factors beyond language: While the four French-speaking samples showed an overall preference for the reflection variant of the relative FoR, they differed significantly regarding the extent to which reflection and the intrinsic FoR were applied. Moreover, in all samples, characteristics of the referenced objects, namely whether they were animate or inanimate, influenced FoR use. The order of tasks also had an impact on referencing behavior.


Keywords: space; spatial cognition; frames of reference (FoRs); linguistic relativity; object characteristics; animacy; French.

## The Question of Language's Influence on Cognition

In the past decades, the debate about linguistic relativity (known as the Sapir-Whorf-Hypothesis, e.g. Sapir, 1949; Whorf, 1956), that is, whether language determines cognition, has been revived (e.g. Gumperz \& Levinson, 1996). An influential research area spurring this revival concerns frames of reference in the domain of space.

Languages differ regarding spatial referencing, that is, how they preferentially describe the position of objects in
relation to one another (e.g. Majid, Bowerman, Kita, Haun, \& Levinson, 2004; Mishra, Singh, \& Dasen, 2009). Moreover, it has been shown that across different languages, frames of reference (FoRs) covary in language and cognition (e.g. Danziger, 2011; Haun, Rapold, Janzen, \& Levinson, 2011; Levinson, 2003). However, there is still much debate on how this covariation comes about, specifically whether language determines cognition or vice versa or whether environmental factors influence both language and cognition (see e.g. the debate between Levinson, Kita, Haun, \& Rasch, 2002, and Li \& Gleitman, 2002; and see Haun et al., 2011; Li, Abarbanell, Gleitman, \& Papafragou, 2011).

In their overview of cross-cultural findings, Majid and colleagues (2004) investigated environment (urban vs. rural), habitual action (subsistence patterns) and cognitive styles (individualism vs. collectivism) as possible mediators between FoRs in language and cognition. They found that none of these factors beyond language systematically accounted for differences in non-linguistic FoR use between speakers of different languages.

Commonly, in spatial referencing research, language and culture are treated as one entangled factor (e.g. Burenhult \& Levinson, 2008). As they are closely intertwined (e.g. Kodish, 2003), differential effects of language and culture are arguably difficult to investigate. However, feasible approaches would be to investigate individuals living in the same country but speaking different languages or the other way around: individuals speaking the same language but living in different countries.

Research in this vein indicates that both language and extra-linguistic factors play a role for referencing strategies. Eggleston (2012) compared three samples, namely Spanish speakers from Barcelona and from Nicaragua and a Nicaraguan sample speaking Sumu-Mayangna. While the former two spoke the same language, the latter two lived in the same country. She found that the samples differed with respect to referencing preferences. Shared language was a stronger predictor of spatial referencing behavior than shared environment. Eggleston concludes that the two factors interact. Similarly, indicating influences beyond language, Troadec (2003) found differences in FoRpreferences between two different French speaking samples in Polynesia, in that the absolute FoR was preferred on an island while the relative FoR was preferred in a city. Taken together, these findings indicate that while language and spatial cognition covary, speaking the same language alone does not necessitate identical FoR-preference. Instead, there seem to be differing conventions between communities, at least when the language allows for application of all FoRs.

## Frames of Reference

In times of ever-increasing international cooperation it is important to know possible sources of miscommunication. Implications of research on spatial referencing thus go far beyond research offices, as illustrated by the following example: Task forces from different countries have to come up with strategies of how to enter a building in which terrorists are keeping hostages. If the order "we enter the building from the back left, you guys go in from the front right entrance" are interpreted in different manners, this may have devastating consequences. Knowledge on differences regarding how we describe where things are is one important step in the direction of successful international cooperation.

In order to describe the location of objects in relation to one another, frames of reference (FoRs) are used. They comprise several constituents (cf. Levinson, 2003): a coordinate system (e.g. front, back, left and right), a figure object whose location is to be described and a ground object in relation to which the location of the figure is described (Talmy, 1983).

Three main FoRs have been identified in the literature (Levinson, 2003): absolute, intrinsic and relative FoR. The relative FoR is subdivided into three variants: translation, reflection and rotation. The absolute FoR uses fixed bearings, such as the cardinal directions for the coordinate system. Applying this FoR, a figure object might then be described to be "northeast" of the ground object.

Applying the intrinsic FoR, the coordinate system is centred in the ground object, the figure object's position is thus described from the perspective of the ground object. Hence, this FoR can only be applied when the ground object has intrinsic front, back, left and right sides.

Using a relative FoR, the position of the figure object in relation to the ground object is described from an observer's perspective; the primary coordinate system originates in the
observer's front, back, left and right sides. This primary coordinate system is then projected onto the ground object and transformed into a secondary coordinate system in one of three possible ways: Applying translation, the secondary coordinate system results from a mere shift of the primary system into the ground object. Here, left and right remain oriented as in the observer's primary coordinate system. A figure between the observer and the ground is described to be "behind" the ground object. In the case of reflection, the primary coordinate system is reflected off the ground object. A figure between the observer and the ground is hence described to be "in front of" the ground object, left and right again remain oriented as in the primary coordinate system. In the third variant, rotation, the secondary coordinate system results from rotating the primary system and centring it in the ground as if another observer was facing the observer of the scene. Here, similar to the reflection variant, a figure between the observer and the ground is described to be "in front of" the ground object, however, left and right are also switched. Thus, between the intrinsic and the three relative FoRs, the order "we enter the building from the back left, you go in from the front right entrance" can be interpreted in at least four different ways. However, individuals are mostly unaware of ambiguities in their spatial descriptions (cf. Grabowski \& Miller, 2000).

## The "Western" Bias

Research investigating the link of language and cognition by means of spatial referencing has almost exclusively focused on comparing "Western" (North-American and European) with "Non-Western" (Indigenous) samples. It has been shown that while "Westerners" preferentially use egocentric (relative) referencing strategies, many "Non-Western" cultures use allocentric (absolute) referencing, some even exclusively (e.g. Levinson, 2003). While there is a prevalent implicit assumption that "Westerners" are all the same (e.g. Pederson, 1993), empirical findings comparing referencing behavior within and between Western cultures are scarce. Those studies attempting to do so (e.g. Grabowski \& Miller, 2000; Flaherty \& Richardson, 1996) found that there are differences regarding the application of at least two distinctive FoRs commonly used by speakers of European languages: The intrinsic and the relative FoR. Importantly, the variants of the latter have received very little attention in past research efforts on spatial referencing. In research, the reflection variant is commonly treated to be "the" relative FoR and the only one investigated. However, use of the other two variants has also been reported (e.g. translation in Tongan and Hausa: Bennardo, 2000; Hill, 1982; translation and rotation to some extent in Chinese, Tongan and Farsi speaking samples: Beller, Hüther, Singmann, \& Bender, subm.; Beller, Singmann, Esfandiari, \& Bender, subm.; Bender, Rothe-Wulf, Hüther, \& Beller, 2012). Accounting for the different ways individuals can reference from their own perspective, we found that FoR-preferences of two "Western" populations speaking different languages, namely US-Americans and Germans, differ regarding the
extent to which variants of the relative FoR are applied (Beller et al., subm. [a]; Bender et al., 2012; Hüther, 2010). In the current study, we examined if referencing differences would also occur between Western cultures speaking the same language, namely French.

## Determining the Role of Situational Aspects

The French language allows for the application of all described FoRs. While a preference for the relative FoR is assumed in native speakers of French (e.g. Mishra et al., 2009; Pederson, 1993), there are no empirical investigations of speakers of French regarding the variants of the relative FoR and the intrinsic FoR.

To shed light on the issue of language versus other factors influencing referencing preferences in Western populations, we compared four samples speaking the same language but living in different countries: Belgians, French, Québécois and Swiss. Thus, we kept native language constant while the environment (country) varied between samples. If language alone were the main determinant of referencing preferences, the different groups of French native speakers should not significantly differ in their use of FoRs, irrespective of where they live. However, if FoR-use is a matter of conventions within communities, samples from different countries may differ despite their speaking the same language. Moreover, testing for possible situational factors influencing FoR-use (as suggested by Li \& Gleitman, 2002), we varied whether the referenced objects were animate or inanimate. We assumed a stronger preference for the intrinsic FoR with animate objects. The intrinsic orientation of living beings may be more salient than that of inanimate objects thus making the application of the intrinsic FoR easier. Also, one may be more likely to take, say, a bird's perspective than that of a pencil. We thus assumed influences of the given spatial task at hand, in that intrinsic referencing would occur more often with animate than with inanimate objects.

Considering the three variants of the relative FoR as well as the intrinsic FoR, we set out to answer the following questions: Which referencing preferences can be observed in native speakers of French from France, Switzerland, Belgium, and Canada? Are there inter-individual differences within the countries, indicating variations in FoR-use within communities? Do the samples differ with regard to their preferred FoR, indicating that language is not the only determinant of referencing preferences? And finally, are individuals' referencing choices intra-individually consistent across different situations or do situational influences such as differences in animacy of the to be referenced objects correspond to different FoR-use?

## Method

FoR-use of the four French speaking samples was assessed using an online questionnaire in French. We developed the questionnaire using the Questback software. Within each sample, two versions of the questionnaire were administered in order to control for sequence effects.

Participants A total of 186 students ( 131 female) of the social sciences completed the questionnaire. The Belgian sample consisted of 55 students ( 46 female; mean age 21.8 years, $S D=4.82$ ) from the University of Liège. The French sample consisted of 46 students ( 34 female; mean age 22.5 years, $S D=4.51$ ) attending different universities. The Canadian sample consisted of 57 students ( 29 female; mean age 25.5 years, $S D=6.07$ ) from the University of Montreal. The Swiss sample consisted of 28 students ( 22 female; mean age 24.3 years, $S D=4.27$ ) from the University of Geneva. All participants indicated that French was their native language. Participation was voluntary and was not compensated.
Materials and Procedure All materials were presented in French. After being informed about the procedure and indicating their consent, each participant filled out a questionnaire comprising 40 tasks. Each task contained one of the following eight descriptions:

- The candle [the starfish] is located behind and to the right of the pencil [the bird].
- The candle [the starfish] is located in front and to the left of the pencil [the bird].
- The candle [the starfish] is located behind and to the left of the pencil [the bird].
- The candle [the starfish] is located in front and to the right of the pencil [the bird].
Each of the eight descriptions was presented five times, every time with different pictures as answer options. Out of eight photographs displaying different configurations of animate (starfish and bird) or inanimate (candle and pencil) objects, participants were asked to choose the one that best fitted the description. The photographs differed regarding the direction into which the ground object (bird or pencil) was facing and regarding the position of the figure (starfish or candle). In each task, four of the depicted object configurations corresponded to a distinct FoR (the three relative FoR variants or the intrinsic FoR); choosing one of the other four pictures did not indicate application of one of the FoRs. Animate and inanimate object configurations were presented as two blocks. An example item is shown in Figure 1. Here, choosing b) corresponds to the reflection variant of the relative FoR, c) to the rotation variant, h) corresponds to the translation variant, and selecting d) indicates application of the intrinsic FoR.


Figure 1: Example of an item with configurations of animate objects. The task was to select the photograph corresponding to the description "The starfish is located in front and to the left of the bird."

Table 1: Percentages of FoR choices in the four samples considering object animacy and order of tasks.

|  |  | order |  | items with inanimate objects |  |  |  | items with animate objects |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Country |  |  | ref. | rot. | trans. | int. | n.c.p. | ref. | rot. | trans. | int. | n.c.p. |
| Belgian | ( $\mathrm{N}=30$ ) | A | 83.2 | 1.3 | 5.7 | 7.5 | 2.3 | 71.3 | 1.8 | 4.7 | 19.3 | 2.8 |
|  | ( $\mathrm{N}=25$ ) | B | 46.2 | 1.8 | 6.4 | 43.6 | 2.0 | 42.2 | 2.6 | 4.8 | 44.2 | 6.2 |
|  | ( $N=55$ ) | Total | 64.7 | 1.6 | 6.1 | 25.6 | 2.2 | 56.8 | 2.2 | 4.8 | 31.8 | 4.5 |
| French | ( $\mathrm{N}=21$ ) | A | 65.5 | 5.0 | 3.6 | 21.7 | 4.3 | 54.1 | 4.1 | 4.1 | 30.5 | 7.4 |
|  | ( $\mathrm{N}=25$ ) | B | 38.4 | 2.2 | 3.8 | 49.2 | 6.4 | 38.2 | 6.2 | 3.8 | 44.0 | 7.8 |
|  | ( $N=46$ ) | Total | 52.0 | 3.6 | 3.7 | 35.5 | 5.4 | 46.2 | 5.2 | 4.0 | 37.3 | 7.6 |
| Canadian |  | A | 80.0 | 0.8 | 7.8 | 7.7 | 3.8 | 56.7 | 1.6 | 5.6 | 30.5 | 5.6 |
|  | ( $\mathrm{N}=25$ ) | B | 60.4 | 3.0 | 5.8 | 29.0 | 1.8 | 41.6 | 5.0 | 7.2 | 41.8 | 4.4 |
|  | ( $N=57$ ) | Total | 70.2 | 1.9 | 6.8 | 18.4 | 2.8 | 49.2 | 3.3 | 6.4 | 36.2 | 5.0 |
| Swiss | ( $\mathrm{N}=14$ ) | A | 84.6 | 8.6 | 0.7 | 2.1 | 3.9 | 69.6 | 8.9 | 1.8 | 15.4 | 4.3 |
|  | ( $\mathrm{N}=14$ ) | B | 79.6 | 1.8 | 1.4 | 13.2 | 3.9 | 65.4 | 4.6 | 2.1 | 24.6 | 3.2 |
|  | ( $\mathrm{N}=28$ ) | Total | 82.1 | 5.2 | 1.1 | 7.7 | 3.9 | 67.5 | 6.8 | 2.0 | 20.0 | 3.8 |

Note: Percentages are aggregated across participants and 20 items per block (animate and inanimate); order A: inanimate-animate, order B: animate-inanimate; ref. indicates application of the reflection variant of the relative FoR, rot. $=$ rotation, trans. $=$ translation, int. $=$ intrinsic FoR and the n.c.p. (no clear preference) column denotes the percentages of answers that did not correspond to one of the investigated FoRs (i.e. choices of one of the other four answer options).

Participants were randomly assigned to one of two questionnaire sequences: Approximately half of each sample answered the 20 tasks with inanimate objects first (A), the others started with the 20 animate items (B). After completion of all 40 tasks, demographics were requested and participants were thanked and debriefed.

## Data Analysis and Results

To answer our research questions, we combined descriptive and inferential statistics in analyzing FoR-use in the four samples. As evident in Table 1, albeit to differing extents, in all four investigated French speaking countries the reflection variant of the relative FoR was predominantly used, followed by the intrinsic FoR.

In order to be able to test for significant differences in the samples' FoR-use, we first identified for every participant how often they applied the different FoRs in the two blocks (the possible maximum being 20 per block, indicating application of the same FoR on every item). We then conducted a repeated measures ANOVA with FoR (reflection, rotation, translation, intrinsic, and no clear preference) and block (animate vs. inanimate) as within-subjects-factors and nationality (Belgian, French, Canadian or Swiss) and block order (animate or inanimate items first) as between-subjects-factors. Aside from the expected main
effect of FoRs $\left(F(1.360,242.041)=159.4, p<.001, \mathfrak{y}_{\mathrm{p}}^{2}=\right.$ .472), it revealed a significant interaction of FoRs $x$ nationality $\left(F(4.079,242.041)=2.7, p=.029, \mathrm{y}_{\mathrm{p}}^{2}=.044\right)$. For all analyses, Greenhouse-Geisser corrected values are reported. As apparent in Table 1, the main differences between the samples from the different countries concerned the extent to which the intrinsic FoR and the reflection variant of the relative FoR were applied. Aggregating the data over all items and participants, we found that in the French sample, while the reflection variant of the relative FoR was applied the most ( $48.1 \%$ ), the intrinsic FoR was applied in over a third of the items (37.3\%). In the other three countries, the reflection variant dominated much more clearly over the other referencing options: in the Canadian sample, $60.8 \%$ of the aggregated answers corresponded to the reflection variant ( $26.2 \%$ intrinsic), in Belgium 62.2\% ( $27.3 \%$ intrinsic) and the Swiss sample showed the clearest preference with $74.8 \%$ of the answers corresponding to the reflection variant (only $13.8 \%$ intrinsic).

The repeated measures ANOVA also showed a significant interaction of FoR and animacy $(F(1.411,242.041)=16.7$, $p<.001, \mathrm{y}_{\mathrm{p}}^{2}=.086$ ). As evident from Table 1, this effect was mainly due to overall increased application of the intrinsic FoR with animate objects as opposed to inanimate objects. There was also a significant interaction between

Table 2: Consistent referencing preferences in \% by nationality and object animacy.

|  | items with inanimate objects |  |  |  |  | items with animate objects |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Country | ref. | rot. | trans. | int. | n.c.p. | ref. | rot. | trans. | intr. | n.c.p. |
| Belgium ( $\mathrm{N}=55$ ) | 67.3 | 0 | 3.6 | 23.6 | 5.5 | 54.6 | 0 | 1.8 | 30.9 | 12.7 |
| France $\quad(\mathrm{N}=46)$ | 50.0 | 2.2 | 0 | 36.9 | 10.9 | 41.3 | 0 | 0 | 34.8 | 23.9 |
| Canada ( $\mathrm{N}=57$ ) | 71.9 | 0 | 5.3 | 17.5 | 5.3 | 45.6 | 0 | 1.7 | 28.1 | 24.6 |
| Switzerland ( $\mathrm{N}=28$ ) | 85.7 | 3.6 | 0 | 7.1 | 3.6 | 60.7 | 3.6 | 0 | 14.3 | 21.4 |

Note: Table displays averaged percentages of number of individuals whose FoR-choice was consistent across at least 15 of 20 items per block. N.c.p. (no clear preference) subsumes participants who were not intra-individually consistent.

FoR and order $\left(F(1.360,242.041)=11.9, p<.001, \mathfrak{y}_{\mathrm{p}}{ }^{2}=\right.$ .063): Participants who worked on the animate items first showed a stronger overall preference for the intrinsic FoR (intrinsic: 38.3\%; reflection: 48.9\%) than those who worked on the inanimate items first (intrinsic: 17.3\%; reflection: 70.5 \%).

Taking a closer look at the variance in FoR-choices within the countries, we established for all participants whether they were intra-individually consistent in their FoRuse across tasks. We considered an individual to be consistent, when he/she chose the same FoR in at least 15 of 20 items ( $75 \%$ ). Interestingly, intra-individual consistency was also affected by animacy of the items (cf. Table 2). For inanimate items, the majority of participants behaved intraindividually consistent in terms of their FoR choice (France was the exception with $10.9 \%$ of participants that were not intra-individually consistent). For animate items, FoRchoice across items was less intra-individually consistent (cf. Table 2). There were also inter-individual differences within the countries: In the Belgian, French and Canadian samples a considerable number of participants (consistently) applied either reflection or the intrinsic FoR, whereas the Swiss sample applied reflection very consistently both intraindividually (cf. Table 2), and inter-individually. All samples preferred the reflection variant over the other two variants of the relative FoR.

## Discussion

Considering the three variants of the relative FoR as well as the intrinsic FoR, we found significant differences in referencing behavior between four samples speaking the same native language but living in different countries. French allows for choosing freely between the different FoRs, thus misunderstandings occurring due to this variation in descriptions/interpretations of the same spatial array are possible. While all samples generally preferred reflection over the other variants of the relative FoR, the French used the intrinsic FoR more than the Canadian, Belgian and Swiss samples, the latter almost exclusively applied reflection. Unlike the other investigated countries, France is not officially bilingual. Since second language proficiency influences spatial referencing (cf. Flaherty \& Richardson, 1996), this may account for the French sample differing from the others. The observed general preference for reflection corresponds to our previous findings in German samples. However, indicating situational influences on referencing behavior and thus supporting our hypothesis, the intrinsic FoR was used more frequently with animate than with inanimate items. In previous studies (e.g., Beller et al., subm. [a]) we had found no significant differences with respect to object's animacy. The current study may have facilitated detection of such differences by using photographs, thus making the referencing tasks more lifelike. Another possible explanation inherent in the depicted objects is that the animate objects were bigger and the intrinsic front of the bird may thus have been more salient than that of the pencil. However, this possible effect of the
material would have affected all samples in a similar fashion and hence cannot account for the observed differences between our samples. Presenting the animate and inanimate tasks in blocks may also have had an influence (see Surtees, Apperly, \& Noordzij, 2011). Specifically, task order influenced FoR choice: when animate items came first, the intrinsic FoR was applied more on the following inanimate items as well, and the same effect appeared for the reflection variant of the relative FoR when the order was reversed. This might be explained by some sort of priming effect and/or a tendency to reference consistently (set effects). Regarding consistency, we found that differences within countries were due to inter-individual differences rather than intra-individual differences in referencing. This suggests that while miscommunication regarding the position of objects occurs between people in a given country, individuals tend to keep to their preferred referencing strategy. Note, however, that intra-individual consistency was lower with animate than with inanimate items. Overall, our findings empirically support the generally expected preference for a relative FoR in Western cultures, more specifically, for the reflection variant. However, our data show that the intrinsic FoR is also applied to considerable extents in all investigated populations. Moreover, we found that the extent to which this FoR is applied by French speakers differs both between and within countries. One cannot help but wonder why these possible sources of misunderstandings are not gradually adapted or at least made explicit within language communities, so that the same sentence will not be interpreted in different ways. How does this fit in with Sapir's (1949) famous claim that "We see and hear and otherwise experience very largely as we do because the language habits of our community predispose certain choices of interpretation"? Our findings regarding interindividual differences and differences between countries indicate that interpretation choices may not so strongly be predisposed by the language habits of our community, whether the community is defined by speaking the same language or by living in the same country. We are required to talk about things in space everyday, yet FoRs seem to still be somewhat variably applied, possibly hindering successful communication.

With respect to limitations of the current study, it must be said that our rather homogenous samples of university students majoring in the social sciences do not necessarily warrant generalizability of our findings to the entire population of the investigated countries. On the upside, however, the observed differences cannot be attributed to differences between the samples concerning factors like age or level of education. Moreover, assessing referencing preferences by means of a questionnaire may not adequately represent strategies applied in everyday settings. However, our questionnaire allowed for assessment of the different samples in their usual surroundings with the exact same measure, thus avoiding possible experimenter effects induced by specific dialects for instance. Regarding this train of thought, one might argue that different variants of

French are spoken in the investigated samples, and that differences in referencing behavior may be due to the differences in the linguistic habits of these groups. However, we used the same descriptions across all groups, and, more importantly, our tasks did not require speech production but rather interpretation of a given sentence by choosing one of eight possible depictions. This procedure diminishes possible influences of different dialects on referencing strategies as linguistic input was both minimal and held constant across groups.

In summary, our findings suggest that with respect to the FoRs commonly applied by Westerners, there are differences even between speakers of the same language that seem to be influenced, at least in part, by their living in different countries. In addition, we also found that aspects of the referenced objects had an impact on FoR-preferences. How these differences between speakers of the same language, yet living in different places, come about remains a question for further research.

## References

Bender, A., Rothe-Wulf, A., Hüther, L., \& Beller, S. (2012). Moving forward in space and time: How strong is the conceptual link between spatial and temporal frames of reference? Frontiers in Psychology: Cultural Psychology, 3:486, 1-11.
Beller, S., Hüther, L., Singmann, H., \& Bender, A. (subm. [a]). How relative is the relative frame of reference? Spatial reference in frontal and dorsal arrays across four languages.
Beller, S., Singmann, H., Esfandiari, A., \& Bender, A. (subm. [b]). Frontal and dorsal spatial references in Farsi: Insights from an MPT-analysis.
Bennardo, G. (2000). Language and space in Tonga: The front of the house is where the chief sits! Anthropological Linguistics, 42, 499-544.
Burenhult, N., \& Levinson, S. C. (2008). Language and landscape: A cross-linguistic perspective. Language Sciences, 30, 135-150.
Danzinger, E. (2011). Distinguishing three-dimensional forms from their mirror-images: Whorfian results from users of intrinsic frames of linguistic reference. Language Science, 33(6), 853-867.
Eggleston, A. G. (2012). Spatial reference in SumuMayangna, Nicaraguan Spanish, and Barcelona Spanish. Doctoral dissertation, West Lafayette, Indiana: Purdue University.
Flaherty, M., \& Richardson, W. (1996). Cognition, perception and deixis: the expression of spatial relationships in Spanish and English. Proceedings of the Second International Colloquium on Deixis, Time, Space and Identity. Nancy, France: Centre de Recherche en Informatique.
Grabowski, J., \& Miller, G. A. (2000). Factors affecting the use of dimensional prepositions in German and American English: Object orientation, social context, and
prepositional pattern. Journal of Psycholinguistic Research, 29, 517-553.
Gumperz, J. J., \& Levinson, S. C. (1996). Rethinking Linguistic Relativity. Cambridge University Press.
Haun, D. B. M., Rapold, C. J., Janzen, G., \& Levinson, S. C. (2011). Plasticity of human spatial cognition: Spatial language and cognition covary across cultures. Cognition, 119, 70-80.
Hill, C. A. (1982). Up/down, front/back, left/right. A contrastive study of Hausa and English. In J. Weissenborn \& W. Klein (Eds.), Here and there: Cross-linguistic studies on deixis and demonstration. Amsterdam: Benjamins.
Hüther, L. (2010). Spatial Referencing in The United States of America and Germany: A Cross-Cultural Empirical Study. Diploma Thesis, Albert-Ludwigs-University, Freiburg, Germany.
Kodish, B. I. (2003). What we do with language-What it does with us. ETC: A Review of General Semantics, 60, 383-395.
Levinson, S. C. (2003). Space in Language and Cognition. Cambridge, MA: Cambridge University Press.
Levinson, S. C., Kita, S., Haun, D. B. M., \& Rasch, B. H. (2002). Returning the tables: Language affects spatial reasoning. Cognition, 84, 155-188.
Li, P., \& Gleitman, L. (2002). Turning the tables: Language and spatial reasoning. Cognition 83, 265-294.
Li, P., Abarbanell, L., Gleitman, L., \& Papafragou, A. (2011). Spatial reasoning in Tenejapan Mayans. Cognition, 120, 33-53.
Majid, A., Bowerman, M., Kita, S., Haun, D. B. M. \& Levinson, S. C. (2004). Can Language restructure cognition? The case for space. Trends in Cognitive Sciences, 8, 108-114.
Mishra, R. C., Singh, S., \& Dasen, P.R. (2009). Geocentric dead reckoning in Sanskrit - and Hindi - medium school children. Culture \& Psychology, 15, 386-408.
Pederson, E. (1993). Geographic and manipulable space in two Tamil linguistic systems. In A. Frank \& I. Campari (Eds.), Space information theory. Berlin: Springer.
Sapir, E. (1949). Selected writings. In D. Mandelbaum (Ed.), Culture, language and personality. Berkeley, CA: University of California.
Press.Surtees, A. D. R., Apperly, I. A. \& Noordzij, M. (2011). Sometimes losing your self in space: Children's and adult's spontaneous use of multiple spatial reference frames. Developmental Psychology, 48(1), 185-191.
Talmy, L. (1983): How language structures space. In: H. L. Pick, Jr. \& L. P. Acredolo (eds.): Spatial orientation: Theory, research and application. NY: Plenum.
Troadec, B. (2003). Le développement de la représentation de l'espace à Tahiti: variabilité du côté mer et du côté montagne. Jrnl de la Société des Océanistes, 116, 25-37.
Whorf, B. L. (1956). Language, mind, and reality. In J.B. Carroll (Ed.), Language, thought and reality. Selected writings of Benjamin Lee Whorf. Cambridge, MA: MIT Press.

# Watching Gestures during Learning about Movements with Dynamic Visualization Activates the Human Mirror Neuron System: A fNIRS Study 

Birgit Imhof (b.imhof@iwm-kmrc.de) ${ }^{\text {a }}$<br>Ann-Christine Ehlis (ann-christine.ehlis@med.uni-tuebingen.de) ${ }^{\text {b }}$<br>Florian B. Häußinger (florian.haeussinger@med.uni-tuebingen.de) ${ }^{\text {b }}$<br>Peter Gerjets (p.gerjets@iwm-kmrc.de) ${ }^{\text {a }}$<br>${ }^{a}$ Knowledge Media Research Center, Schleichstrasse 6, 72076 Tuebingen, Germany<br>${ }^{\mathrm{b}}$ Department of Psychiatry \& Psychotherapy, University Hospital Tuebingen, Calwerstr. 14, 72076 Tuebingen, Germany


#### Abstract

This study investigates whether viewing human gestures facilitates learning about non-human biological movements and whether correspondence between gesture and to-belearned movement is superior to non-correspondence. Functional near-infrared-spectroscopy was used to address whether gestures activate the human mirror-neuron-system (hMNS) and whether this activation mediates the facilitation of learning. During learning participants viewed triples of visualizations (animation - gesture video - animation). Results showed that for low-visuospatial-ability learners corresponding gestures led to higher cortical activation in the inferior-frontal cortex (part of the hMNS) and better learning outcomes, whereas for high-visuospatial-ability learners the type of gesture had no influence. Furthermore, results showed that - if presented with non-corresponding gestures - only low-visuospatial-ability learners who activated their inferiorparietal cortex (also part of the hMNS), improve their learning. Thus, activating the hMNS facilitates learning about movements and stimulating the hMNS via gestures seems to be an adequate instructional strategy to enhance learning with dynamic visualizations for low-visuospatial-ability learners.


Keywords: Learning about movements; dynamic visualizations; human mirror-neuron-system; gestures; functional near-infrared-spectroscopy.

## Learning about Movements with Dynamic Visualizations

Many contents in the Natural Sciences as well as in other domains, such as different sport disciplines or scene perception, comprise the understanding of changes in space over time. Dynamic visualizations can easily depict such changes and they may be particularly suited for instructional purposes if these changes do not occur in a discrete or linear way, but rather involve more complex continuous aspects (e.g., acceleration). However, they were not always superior to static visualizations to convey dynamic information (e.g., Imhof et al., 2012). Thus, it is crucial to understand when and for whom dynamic visualizations are beneficial to use them effectively and to exploit their potential for learning. Until now, research on the instructional use of dynamic visualizations has yielded rather heterogeneous results: Not only design factors and individual learner characteristics,
but also context factors, such as, the knowledge domain, task requirements, or additional instructional support, influence the effectiveness of dynamic visualizations (e.g., Höffler \& Leutner, 2007; Lowe, Schnotz, \& Rasch, 2011; Tversky, Morrison, \& Bétrancourt, 2002). These context factors have become a focus of research on dynamic visualizations.

## Learning with Gestures

One idea on how to support learning about movements with dynamic visualizations that is based on the embodied cognition approach and proposed by De Koning and Tabbers (2011) is the active and passive use of gesture. Empirically, Hegarty et al. (2005) showed that gestures are naturally used to express movements of depicted components and thereby also the depicted processes in mental animation problems. Moreover, it has already been shown that the production of gestures during learning is beneficial for acquiring knowledge about different scientific topics and spatial problem solving (e.g., Chu \& Kita, 2011; Cook \& Goldin-Meadow, 2006; Scheiter et al., 2012). However, learners can either produce gestures on their own or they can perceive gestures that are performed by others. In line with the proposal of De Koning and Tabbers (2011), it is also beneficial for learning to perceive gestures that illustrate the depicted contents, for instance, performed by teachers (e.g., Valenzeno, Alibali, \& Klatzky, 2003).

Underlying this gesture watching effect might be the activation of brain areas (i.e., the human mirror-neuronsystem [hMNS]; Fogassi \& Ferrari, 2011; Rizzolatti \& Craighero, 2004) that are typically used to observe, understand and imitate the actions of other persons. In a related line of research, a current hypothesis that has recently received considerable attention (e.g., Ayres et al., 2009; Van Gog et al., 2009) is that the stimulation and involvement of this hMNS might be beneficial for learning about complex continuous aspects with dynamic visualizations. The hMNS is typically activated by human movements, but may be more generally used to also represent other biological or even non-biological movements, if the observer is able to anthropomorphize
these movements (cf. De Koning \& Tabbers, 2011; Engel et al., 2008). Thus, in the domain of learning about biological movements, one effective instructional strategy to activate the hMNS might be to show learners not only the to-belearned movements via dynamic visualizations, but also gestures displaying the to-be-learned dynamics in order to trigger an anthropomorphized encoding. Hence, only showing gestures that map onto the to-be-learned movements should benefit learning about those movements.

This study addresses whether perceiving gestures in addition to dynamic visualizations is also beneficial for learning. Moreover, to investigate the role of the hMNS, the underlying cognitive processes during learning were investigated with neurophysiological methods in this study (i.e., functional near-infrared-spectroscopy [fNIRS]). Until now, to the best of our knowledge, there is no direct test of the assumption that learners' ability to recruit their hMNS during processing dynamic visualizations may influence the effectiveness of the visualizations. Moreover, it still has not been investigated whether hMNS activation can be induced by gesture-based interventions and then transferred to nonhuman movements because of mapping processes. This approach might easily facilitate the understanding of complex dynamic phenomena by implementing embodied visualizations that activate specific brain areas into instructional materials.

## Learners' Visuospatial Ability

Beyond context factors also individual learner characteristics may play a role during learning about biological movements. Because processing continuous changes requires visuospatial ability (cf. Hegarty, 1992), it is likely that learners' visuospatial ability will determine how much the learners profit from visualizations (cf. Hegarty \& Waller, 2005). Often the continuous processes do not occur only in two-dimensional but rather in threedimensional space. Thus not only visual, but also spatial aspects are important. Previous research on visuospatial ability has revealed two important results, namely that (a) learners with higher visuospatial ability outperform learners with lower visuospatial ability during learning with visualizations (see Höffler, 2010, for a meta-analysis) and moreover, there is some evidence, that (b) visuospatial ability may moderate the effectiveness of learning with different visualization formats. Higher visuospatial ability may compensate for "poor" instructions (i.e., in our case unrelated non-corresponding gestures, cf. methods section), whereas learners with lower visuospatial ability suffer from such instructions (cf. ability-as-compensator hypothesis; e.g., Hays, 1996; Hegarty \& Kriz, 2008; Höffler, 2010).

## Research Questions and Hypotheses

This study addressed by using neurophysiological methods (i.e., functional near-infrared-spectroscopy [fNIRS], which is a non-intrusive approach to gather data about cortical activation of humans) the research question whether the hMNS is activated during viewing gestures and whether the
viewing of these gestures is helpful for learning about biological movements because learners map the human movements to the non-human biological movements.

However, maybe solely the circumstance that learners see a human during learning activates the hMNS and is thus sufficient to facilitate learning about biological motions. In other words, it might be also helpful for learners to see gestures that have nothing to do with the to-be-learned content. Thus, this study investigated whether viewing gestures that correspond to the to-be-learned non-human movements facilitate learning about these movements better than unrelated non-corresponding gestures. Additionally, the moderating role of learners' visuospatial ability was addressed. Furthermore, this study tested whether the activation of the MNS mediates the facilitation of learning.

We hypothesize that viewing corresponding gestures facilitates learning more than viewing unrelated noncorresponding gestures. This might be particularly true for low-visuospatial-ability learners, whereas high-visuospatialability learners might not need this type of anthropomorphization to learn about the depicted dynamic processes (cf., ability-as-compensator hypothesis; e.g., Höffler, 2010). Moreover, we hypothesize that learners differ with regard to recruiting the hMNS for processing and that higher hMNS activation is associated with better learning outcomes than lower hMNS activation. This might again be particularly true for low-visuospatial-ability learners, as they do not have available this ability to compensate for such a hMNS actication.

## Methods

## Participants and Design

Forty-five university students ( $M=24.98$ years, $S D=4.57$; 31 females) were asked to learn how to classify different fish according to their movements based on visualizations that illustrated four different movement patterns of fish. For each movement pattern the participants saw three visualizations: Firstly, they saw an animation of the specific movement pattern. Secondly, they saw a video of a person performing gestures with his hands and arms. These gestures either did correspond or did not correspond (i.e., were unrelated) to the fish movement patterns. Therefore, at this point the experimental manipulation with the betweensubjects factor type of gesture took place. Thirdly, the learners saw the initial fish animation again.

An expert regarding fish movements performed the gestures. For the corresponding gestures this expert was instructed to display with his hands and arms representations of the respective movements as clearly as possible (see figure 1 left). For the non-corresponding gestures the expert was instructed to perform gestures with his hands and arms that were unrelated to the fish movement patterns (i.e., waving, circulating the forearms around each other, drumming, and pointing, see figure 1 right).

Each visualization was depicted for 30 s and was followed by pauses of 30 s (black screen) between all
visualizations. The learners were instructed to relax in these pauses. In the pauses, the activations of the brain areas of interest are supposed to decay to the baseline level before the next visualization was displayed.


Figure 1: Learning visualizations in triples: corresponding gestures (left) and non-corresponding gestures (right).

## Materials

Participants had to learn to discriminate four different patterns of fish movements. These movement patterns differ in terms of the body parts that generate propulsion (i.e., the body itself or several fins) and also in the manner of how these body parts move in the three-dimensional space (i.e. different wave-like or paddle-like movements). The four different movement patterns were: 1. undulation of the body; 2 . undulation of the dorsal and anal fins; 3 . oscillation of the dorsal and anal fins (and undulation of the pectoral fins); and 4. oscillation of the pectoral fins. One major challenge in identifying these movement patterns is that fish may deploy other movements in addition (e.g., to navigate), that can easily be confused with movements used for propulsion in another movement pattern.

Animations were rendered based on typical fish performing the four movement patterns. These animations were standardized in terms of the perspective, background, position in the frame, and the swimming direction of the fish. Moreover, in these deliberately designed visualizations, we were able to only show the movements performed for propulsion and omit other irrelevant movements. Beside that, the depicted movements were highly realistic, thus representing the movements of real fish adequately. The movement cycles of the movement patterns were presented in loops in the animations ( 30 s per movement pattern, 25 fps, size: $640 \times 480$ pixels) in the center of the screen.

For each movement pattern, videos of an expert regarding to fish movements were recorded who performed either a corresponding or a non-corresponding gesture. These gestures were presented in the respective conditions in loops in the videos ( 30 s per movement pattern, 25 frames per s , size: $640 \times 480$ pixels) in the center of the screen. The presentation of all visualizations was system-controlled.

## Measures

Learning Outcomes To assess learning outcomes, a movement pattern classification test was administered. This
test comprised 21 dynamic multiple-choice items consisting of underwater videos of real fish performing one of the four to-be-learned movement patterns. To choose for each item the kind of movement pattern that was depicted, learners had to identify the body parts relevant for propulsion and their way of moving. Each item was presented 7 s to the participants and immediately afterwards they had 3 s time to choose the correct answer by pressing a corresponding button. The possible answers were indicated as static screenshots from the learning animations of the four movement patterns. Each item was awarded one point for the correct answer (max. 21 points). The test items were presented in blocks of 30 s so that 3 items were grouped together. Pauses of 30 s (black screen) followed each block.

Learners' Visuospatial Ability Learners' visuospatial ability was assessed with a short version of the paper folding test (PFT, Ekstrom et al., 1976). This test measures the ability to form representations of "object location, movement, spatial relationships, and transformations" (Blazhenkova \& Kozhevnikov, 2009, p. 640) and thus is well suited to cover the domain of fish movements. The short version of the PFT consists of ten multiple-choice items, where participants have to choose the correct answer out of five options. The stimuli are depictions of stepwise folded papers that were punched in the folded state, whereas the answer options depict the punches of various unfolded papers with the punches being either in the correct or incorrect positions. A maximum of three minutes is assigned to work on the items, and each correct answer is worth one point (max. 10 points).

Cortical Activation During viewing the gestures in the learning phase, cortical activation was conducted via fNIRS measurements with an ETG-4000 (Hitachi). As probe set we used a $2 \times 22$ channel array, that was placed over the fronto-temporo-parietal regions centered at the T3-T4 and C3-C4 positions (not exactly terminating on these positions because of the fixed interoptode distances) according to the standard locations of the $10-20$ system. Changes of absorbed near-infrared light were transformed into relative concentration changes of oxygenated $\left(\mathrm{O}_{2} \mathrm{Hb}\right)$ and deoxygenated haemoglobin $(\mathrm{HHb})$. Local increases of $\mathrm{O}_{2} \mathrm{Hb}$ as well as decreases of HHb are indicators of cortical activity (Obrig \& Villringer, 2003).

## Procedure

Participants were tested individually. They first received a printed overview in which they were informed about the procedure on the different parts of the study. Subsequently, they had to answer the PFT and a demographic questionnaire. Subsequently, the fNIRS probe set was placed on the scalp of the participants and adjusted with the help of the experimenter. Then, the learning phase started and the computer-based learning materials were presented. For each of the four to-be-learned movement patterns learners were presented with the triples of visualizations
(fish animation - gesture video - fish animation). In the learning phase the experimental manipulation took place. Learners saw either the corresponding or the noncorresponding gestures. Following the learning phase (12 min ) learners performed a filler task ( 8 min ), in which they listened to music. Subsequently, learners completed the movement classification test ( 8 min ). To answer the test items participants were instructed to put both their forefingers and both their middle fingers on predefined keys. These keys were labeled with screenshots from the corresponding fish animations on the screen. In total, a single experimental session lasted approx. 50 minutes.

## Results

## Learning Outcomes

To analyze learning outcomes we conducted a multiple regression analysis with the categorical predictor type of gesture and the continuous predictor learners' visuospatial ability. We had to exclude four participants because of technical reasons (data loss) resulting in a total number of 41 participants in this analysis. Further, we had to exclude eight test items from the learning outcome measure, because participants answered them with a response rate of more than $95 \%$. The reliability analysis of the remaining 13 test items achieved a good to excellent cronbach's $\alpha$ of 85 .

For learning outcomes the predictors in the regression analysis explained a significant portion of variance ( $p=$ $.01)$. Results showed no effect of type of gesture on learning outcomes ( $p=.41$, ns), whereas there was an effect for learners' visuospatial ability on learning outcomes ( $p=.04$ ). This effect has to be interpreted in terms of the significant interaction between type of gesture and learners' visuospatial ability on learning outcomes ( $p=.04$; figure 2 ).


Figure 2. Interaction between learners' visuospatial ability and type of gesture on learning outcomes.

This interaction was resolved by a simple slopes analysis (cf. Aiken \& West, 1991). It revealed that for participants with high visuospatial ability (defined as one standard deviation above the sample mean) the type of gesture had no influence on learning outcomes ( $p=.34$, ns). As expected, for participants with low visuospatial ability (defined as one standard deviation below the sample mean) corresponding
gestures were better for learning than non-corresponding gestures ( $p=.04$ ). Thus, the corresponding gestures are beneficial for low-visuospatial-ability learners.

## Cortical Activation

To analyze the cortical activation we defined two regions of interest (ROIs) on the left hemisphere for the hMNS among the respective channels. The two ROIs were the left inferiorfrontal cortex (IFC) and the left inferior-parietal cortex (IPC, cf. figure 3). To analyze cortical activation we conducted two multiple regression analyses with the predictors type of gesture and learners' visuospatial ability. We had to exclude additional eight participants from these analyses because the data quality of these participants was too poor resulting in a total number of 33 participants in these analyses. For cortical activation on IPC the predictors in the regression analysis did not explain a significant portion of variance ( $p=.96, \mathrm{~ns}$ ).


Figure 3. Spatial arrangement of the left probeset.
For cortical activation on IFC the predictors in the regression analysis explained a significant portion of variance ( $p<.001$ ). Results showed an effect of type of gesture on IFC activation ( $p<.001$ ) and an effect for learners' visuospatial ability on IFC activation ( $p<.001$ ). These effects have to be interpreted in terms of the significant interaction between type of gesture and learners' visuospatial ability on IFC activation ( $p<.01$; see figure 4).


Figure 4. Effects of type of gesture (G) and learners' visuospatial activities (VSA) on cortical activation (left).

Again a simple slopes analysis was conducted (cf. Aiken \& West, 1991). It revealed that for participants with high visuospatial ability (defined as one standard deviation above the sample mean) the type of gesture had no influence on IFC activation ( $p=.14$, ns). For participants with low visuospatial ability (defined as one standard deviation below the sample mean) corresponding gestures resulted in a higher IFC activation than non-corresponding gestures ( $p<$
.001). Thus, the corresponding gestures helped low-visuospatial-ability learners to activate the hMNS in terms of IFC activation.

## Effects of Cortical Activation on Learning Outcomes

Finally, to address the question whether higher hMNS activation is directly associated with better learning outcomes, we conducted two multiple regression analyses with the three predictors type of gesture, learners' visuospatial ability and cortical activation in terms of IFC activation or IPC activation respectively.

For learning outcomes the predictors in the regression analysis with IFC activation did not explain a significant portion of variance ( $p=.12, \mathrm{~ns}$ ). Interestingly, the predictors in the regression analysis with IPC activation did explain a significant portion of variance for learning outcomes ( $p<$ .01). There was a three-way interaction between the predictors type of gesture, learners' visuospatial ability, and IPC activation on learning outcomes ( $p=.03$; see figure 5 ).


Figure 5. Three-way interaction between type of gesture, learners' visuospatial ability, and IPC activation on learning outcomes.

This triple interaction was resolved by simple slopes analyses (cf. Aiken \& West, 1991). Firstly, this approach revealed that for learners who saw corresponding gestures there was no two-way interaction between participants' visuospatial ability and IPC activation ( $p=.59$, ns). The following simple slopes analyses revealed that IPC activation did not predict learning outcomes for learners who saw corresponding gestures: neither for high-visuospatial-ability learners ( $p=.47$, ns), nor for low-visuospatial-ability learners ( $p=.40$, ns). However, for learners who saw non-corresponding gestures there was an interaction between participants' visuospatial ability and IPC activation $(p<.01)$. We further resolved this two-way interaction between participants' visuospatial ability and IPC activation for learners who saw non-corresponding gestures. The simple slopes analyses revealed that in the group of learners who saw non-corresponding gestures IPC activation negatively predicted learning outcomes for high-visuospatial-ability learners ( $p=.04$ ), whereas for low-visuospatial-ability learners IPC activation positively
predicted learning outcomes $(p=.001)$. Thus, for learners who saw non-corresponding gestures, but have had high visuospatial abilities at their disposal IPC activation is detrimental for learning. However, for learners who did neither have corresponding gestures nor high visuospatial abilities at their disposal, activation of their hMNS in terms of the IPC during processing the unrelated noncorresponding gesture improves their learning.

## Discussion

This study tested whether viewing gestures performed by others is helpful for learning about non-human movements and whether these gestures stimulate anthropomorphization via an activation of the hMNS. The anthropomorphization is stimulated by an external video and is not accomplished by the learners on their own. Our results showed that viewing corresponding gestures activated the hMNS particularly for low-visuospatial-ability learners. These learners achieved the same learning outcomes as high-visuospatial-ability learners. Low-visuospatial-ability learners seem to profit from being demonstrated a connection between non-human biological movements and movements of the human body that correspond to these movements. Thus, learning about biological movements can be facilitated by gesture-based interventions activating parts of the hMNS: Gestures that correspond to the to-be-learned movements and activate the inferior-frontal cortex (IFC). This activation seems to compensate missing viusospatial ability.

Furthermore, our results indicate another way of improving learning about biological movements: When looking at participants who neither have high visuospatial ability, nor received the benefit of viewing corresponding gestures, - namely, the group of low-visuospatial-ability learners who processed non-corresponding gestures - the result pattern was rather heterogeneous: Only participants who activated another part of the hMNS (i.e., the inferiorparietal cortex [IPC]) were able to dramatically improve their learning, whereas participants who did not activate this area achieved only poor results. This indicates that the activation of the inferior-parietal cortex helps participants to learn about biological movements, particularly if they have no access to other facilitating factors. In line with this reasoning, learners who have available two facilitating factors, namely high visuo-spatial abilities and an activation of the IPC, performed worse when they saw noncorresponding gestures. In this case, the two facilitators might compete and interfere with each other resulting in inferior learning outcomes. Nevertheless, higher hMNS activation is associated with better learning outcomes - at least for low-visuospatial-ability learners: for IFC activation it seems that there is a rather stepwise connection in that a certain value has to be reached, whereas for IPC activation it seems that it follows the more activation the better learning.

Stimulating the hMNS by means of gestures seems to be a promising strategy to enhance learning with dynamic visualizations for low-visuospatial-ability learners because this intervention leads to higher activation in their IFC as
part of the hMNS. However, further research needs to replicate these findings with a larger sample size and continue to disentangle the effects of this study. Particularly, our findings have to be replicated with other examples of gestures in different domains, as gestures about fish movements might not be a typical example of gestures. Furthermore, it is very important to investigate how the activation of the IPC can also be fostered by instructions.

Furthermore, gesture-based instructions that support anthropomorphization should be investigated in different instructional domains and settings that involve learning about continuous movements and processes to prove whether they are in general a suitable method to enhance learning about processes with dynamic visualizations.

Further research should also investigate whether effective and less effective dynamic visualizations differ in their ability to activate the MNS, thereby potentially explaining inconsistent results on the effectiveness of dynamic visualizations (e.g., Höffler \& Leutner, 2007; Tversky et al., 2002). The present study is one first step into this field of research and our results suggest that it is important to not only put further effort into designing better dynamic visualizations, but also in providing learners with suitable strategies to adequately process these visualizations.

## References

Aiken, L.S., \& West, S.G. (1991). Multiple regression: Testing and interpreting interactions. Newbury Park: Sage.
Ayres, P., Marcus, N., Chan, C. \& Qian, N. (2009). Learning hand manipulative tasks: When instructional animations are superior to equivalent static representations. Computers in Human Behavior, 25, 348353.

Blazhenkova, O. \& Kozhevnikov, M. (2009). The new object-spatial-verbal cognitive style model: Theory and measurement. Applied Cognitive Psychology, 23, 638663.

Chu, M., \& Kita, S. (2011).The Nature of Gestures' Beneficial Role in Spatial Problem Solving. Journal of Experimental Psychology: General, 140, 102-116.
Cook, S.M., \& Goldin-Meadow, S. (2006). The role of gesture in learning: Do children use their hands to change their minds? Journal of Cognition and Development, 7, 211-232.
De Koning, B.B., \& Tabbers, H.K. (2011). Facilitating understanding of movements in dynamic visualizations: An embodied perspective. Educational Psychology Review, 23, 501-521.
Engel, A., Burke, M., Fiehler, K., Bien, S., \& Rösler, F. (2008). What activates the human mirror neuron system during observation of artificial movements: bottom-up visual features or top-down intentions? Neuropsychologia, 46, 2033-2042.
Ekstrom, R., French, J., Harman, H., \& Dermen, D. (1976). Manual for Kit of Factor-Referenced Cognitive Tests. Princeton: Educational Testing Service.

Fogassi, L., \& Ferrari, P.F. (2011). Mirror systems. Wiley Interdisciplinary Reviews: Cognitive Science, 2, 22-38.
Hegarty, M. (1992). Mental animation: Inferring motion from static diagrams of mechanical systems. Journal of Experimental Psychology: Learning, Memory and Cognition, 18, 1084-1102.
Hegarty, M., \& Kriz, S. (2008). Effects of knowledge and spatial ability on learning from animation. In R. Lowe \& W. Schnotz (Eds.), Learning with animation: Research implications for design. Cambridge, England: Cambridge University Press.
Hegarty, M., Mayer, S., Kriz, S., \& Keehner, M. (2005). The role of gestures in mental animation. Spatial Cognition and Computation, 5, 333-356.
Hegarty, M., \& Waller, D. (2005). Individual differences in spatial ability. In P. Shah, \& A. Miyake (Eds.), Handbook of Visuospatial Thinking. Cambridge University Press.
Höffler, T.N. (2010). Spatial ability: Its influence on learning with visualizations-a meta-analytic review. Educational Psychological Review, 22, 245-269.
Höffler, T.N., \& Leutner, D. (2007). Instructional animation versus static pictures: A meta-analysis. Learning and Instruction, 17, 722-738.
Imhof, B., Scheiter, K., Edelmann, J., \& Gerjets, P. (2012). How temporal and spatial aspects of presenting visualizations affect learning about locomotion patterns. Learning and Instruction, 22, 193-205.
Lowe, R.K., Schnotz, W., \& Rasch, T. (2010). Aligning affordances of graphics with learning task requirements. Applied Cognitive Psychology, 25, 452-459.
Obrig, H., \& Villringer, A. (2003). Beyond the visible Imaging the human brain with light. Journal of Cerebral Blood Flow \& Metabolism, 23, 1-18.
Rizzolatti, G., \& Craighero, L. (2004). The mirror-neuron system. Annual Review of Neuroscience, 27, 169-192.
Scheiter, K., Arndt, J., Imhof, B., \& Ainsworth, S. (2012). Move like a fish: Do gestures aid learning from photographs and videos? In E. de Vries, \& K. Scheiter (Eds.), Proceedings EARLI Special Interest Group Text and Graphics: Staging knowledge and experience: How to take advantage of representational technologies in education and training? Grenoble, France: Université Pierre-Mendès-France.
Tversky, B., Morrison, J., \& Bétrancourt, M. (2002). Animation: Can it facilitate? International Journal of Human-Computer Studies, 57, 247-262.
Valenzeno, L., Alibali, M.W., \& Klatzky, R. (2003). Teachers' gestures facilitate students' learning: A lesson in symmetry. Contemporary Educational Psychology, 28, 187-204.
Van Gog, T., Paas, F., Marcus, N., Ayres, P., \& Sweller, J. (2009). The mirror-neuron system and observational learning: Implications for the effectiveness of dynamic visualizations. Educational Psychology Review, 21, 2130.

# A robustness approach to theory building: A case study of language evolution 

Liz Irvine (elizabeth.irvine@cin.uni-tuebingen.de)<br>Philosophy of Neuroscience, Werner Reichardt Centre for Integrative Neuroscience, Otfried-Müller-Str. 25, 72076 Tübingen, Germany

Seán G. Roberts (sean.roberts@mpi.nl)<br>Language and Cognition Department, Max Planck Instutite for Psycholinguistics Wundtlaan 1, Nijmegen 6525 XD The Netherlands

Simon Kirby (simon@ling.ed.ac.uk)<br>Language Evolution and Computation Research Unit, School of Philosophy, Psychology<br>Language Sciences, University of Edinburgh, Dugald Stewart Building, 3 Charles Street, Edinburgh, EH8 9AD, UK


#### Abstract

Models of cognitive processes often include simplifications, idealisations, and fictionalisations, so how should we learn about cognitive processes from such models? Particularly in cognitive science, when many features of the target system are unknown, it is not always clear which simplifications, idealisations, and so on, are appropriate for a research question, and which are highly misleading. Here we use a case-study from studies of language evolution, and ideas from philosophy of science, to illustrate a robustness approach to learning from models. Robust properties are those that arise across a range of models, simulations and experiments, and can be used to identify key causal structures in the models, and the phenomenon, under investigation. For example, in studies of language evolution, the emergence of compositional structure is a robust property across models, simulations and experiments of cultural transmission, but only under pressures for learnability and expressivity. This arguably illustrates the principles underlying real cases of language evolution. We provide an outline of the robustness approach, including its limitations, and suggest that this methodology can be productively used throughout cognitive science. Perhaps of most importance, it suggests that different modelling frameworks should be used as tools to identify the abstract properties of a system, rather than being definitive expressions of theories. Keywords: Language Evolution; Cultural Evolution; Robustness


## Introduction

A central question in the field of the evolution of language is whether linguistic structure is mainly a product of domainspecific genetic constraints, or of cultural transmission. However, the cultural evolution of language is difficult to study because there is little direct evidence available. Simulations ${ }^{1}$ make it possible to study of the dynamics of cultural evolution, but often include highly simplified mechanisms of learning. Human experiments obviously use a realistic learning mechanism, but present the problem that test subjects already know natural languages, and it is difficult to control for individual differences in learning. While recent work suggests that compositional linguistic structure emerges in iter-

[^300]ated learning contexts under pressures to be learnable and expressive (Kirby, Cornish, \& Smith, 2008), problems with simulations and human experiments potentially make it difficult to justify theoretical claims about the evolution of language.

We explore how the notion of 'robustness' from philosophy of science (e.g. Wimsatt, 1981; Weisberg \& Reisman, 2008) can be used to support claims in language evolution; specifically, we argue that emergent compositional structure is a robust property of iterated learning. We suggest how to extend this methodology to similar claims across cognitive science, that are based on unrealistic models, and little direct empirical evidence.

## Language Evolution

Before discussing the notion of robustness, it is first necessary to introduce work on the iterated learning model (ILM) in language evolution as this will be used as a case-study. The ILM looks at how culturally transmitted systems (e.g. a mapping between linguistic signals and meanings) change by being repeatedly transmitted through a bottleneck (Kirby, 2000). The bottleneck is a restriction of information that could be due to a finite limit on the information to be transferred or a restricted set of meanings to be described. The bottleneck causes the system to change over time, usually towards a more compressible relationship between signals and meanings. This can be interpreted as a pressure on the language to become more 'learnable' by the next generation. In the extreme case, the variation in the system reduces so that there is only one signal. Opposing this is a pressure for expressivity (e.g. a need to distinguish between meanings). A perfectly expressive linguistic system has a different signal for each meaning.

Smith, Kirby, and Brighton (2003) showed that an optimal solution under these two pressures is compositionality: the meaning of a signal is composed of sub-meanings expressed by sub-strings of the signal. This means that there are fewer signal components to be learned than individual meanings, and the signals of unobserved meanings can be re-constructed accurately. The demonstration that cultural transmission can lead to complex linguistic structure contrasts with theories that see linguistic structure as primarily deriving from innate, domain-specific constraints (Chomsky, 1965, see Kirby,

Dowman, \& Griffiths, 2007).
The initial work on the iterated learning model involved computational simulations. Instead of committing to a particular model or simulation framework, a range of computational techniques were used as tools to demonstrate the principles of the iterated learning model. These include grammar induction models (Kirby, 2000), exemplar models (Batali, 2002), neural network models (Kirby \& Hurford, 2002; Swarup \& Gasser, 2008) and self-organising maps (Worgan \& Damper, 2008).

The next step involved translating the ILM into a laboratory experiment. Kirby, Cornish, and Smith (2008) demonstrated that the emergence of compositional structure could be observed in an artificial language which was learned, produced, transmitted and then learned again by human subjects. Participants were exposed to pairings of nonsense words and meanings and asked to memorise them. The meanings were images with structured semantic dimensions: shape (circle, triangle, square), colour (red, blue, black) and movement (horizontal, bouncing, or spiraling motion). Participants were trained on a sub-set of the whole meaning space, but they were then asked to produce a label for every meaning. The labels that were produced became the training data for the next participant. This meant that the language changed as it was transmitted from participant to participant, mirroring the cultural transmission of language.
By this process, the language adapted to two pressures. A bottleneck on transmission was present, because the participants were not trained on all labels. This put a pressure on the language to become more faithfully transmitted. With only this pressure, the number of distinct labels in the language declined. These languages were easy to learn, but not expressive. To counter this, a pressure for expressivity was added by excluding homonyms from the training subset. Under both pressures, the language adapted to become learnable and expressive by becoming compositional. That is, instead of labels being distinct and holistic, sub-parts of each label consistently referred to sub-parts of each meaning. For instance, in one emergent language, all meanings which included the colour blue began with an ' $L$ ' while all meanings that included the spiralling movement ended with 'PILU’ (see Kirby, Cornish, \& Smith, 2008, p. 10684). This meant that the language was both easy to learn and could express all meanings distinctly.

The ILM experiment showed that compositional structure could emerge spontaneously due to the process of cultural transmission. The results mirrored those of the computational simulations, leading to the experiments being thought of as simulations with human participants (see Kirby, Smith, \& Cornish, 2008).

The ILM sparked a lineage of experimental simulations testing different constraints and assumptions of the original simulation, including replacing the exclusion of homonyms with a pressure for communication between two participants (Matthews, Kirby, \& Cornish, 2010; Silvey, Kirby, \& Smith,

2012; Navarro \& Perfors, 2011; Tamariz, Cornish, Roberts, \& Kirby, 2012; Verhoef \& Boer, 2012). Bringing the process full circle, principles elucidated through the human simulations motivated new computational simulations (Smith, Tamariz, Cornish, \& Kirby, 2013). There are differences between these studies, for example the precise distribution of letters in the strings that emerge is not robust across computational and human simulations since human distinctions between vowels and consonants were not built into the computational agents. However, in each case the results were compatible with the theory of structure emerging though repeated transmission through a bottleneck under pressures for learnability and expressivity.

## Problems with abstraction and transparency

Computational models have many advantages: the internal states of individuals are transparent and quantifiable; the exact amount of noise is quantifiable; and exploring the parameter space can be easier than running alternate conditions in human experiments (e.g. Reali \& Griffiths, 2010 model an infinite population). However, the abstractions inherent in computational models can be a weakness because a model's implications rely on the ability to translate between the abstractions and the real world. Computational models of the cultural evolution of language may simplify the representation of linguistic units, learning processes or psychological mechanisms. Simplifying assumptions might be made such as agents sharing an innate, conceptual space for words and meanings (Vogt, 2005) or being able to observe intended meanings (Worgan \& Damper, 2008).

In contrast, laboratory experiments with real humans include artificial languages with concrete analogues to real languages and real learning mechanisms. However, while the learning mechanisms are realistic, they are opaque. It is difficult to deduce the precise mental processes that lead to the emergence of structure. Because of this complexity, it is difficult to maintain absolute experimental control. Furthermore, the participants already have full knowledge of a compositional language. This is a potential confound since the emergent structure may just be a reflection of the participants' existing language rather than being caused by the same process that proto-linguistic humans underwent (e.g. Flynn, 2008; Chomsky, 2011). However, as we shall see below, these problems can be addressed by demonstrating that the outcome is a robust property across different models..

## Aims of models and simulations

One crucial question in this area of research is what simulations are used to learn about, and thus what kind of theoretical inferences can possibly be warranted from them. Simulations of cultural transmission are not intended as instantiations of human language learning, nor the evolution of a real language. However, the ILM is informative about systems that are transmitted through a bottleneck, and that become more structured as a result. The simulations, then, are an example


Figure 1: How simulations, models and lab experiments can contribute to theories of language evolution. Potential problems in the translation from the target phenomenon to models, simulations, and experiments (dotted arrows) are spread over many approaches, while the discovery of robust properties found across all these approaches provide support for the central theory (thick arrows).
of cultural transmission's effect on structure in cultural phenomena, where human language is one of these more general phenomena. Results from simulations of iterated learning can therefore inform the general theory of iterated learning ${ }^{2}$.

In the field of language evolution, simulations of iterated learning are used as thought experiments to show that nativist assumptions are not necessary to account for the emergence of complex linguistic structure. Computational simulations can be more powerful than traditional thought experiments because they must usually be instantiated at a more technical level, and because they allow complex interactions and structures which might be unintuitive (Bedau, 1998).

Di Paolo, Noble, and Bullock (2000) suggest a workable methodology for opaque thought experiments:

1) Exploratory phase: Explore model behaviour, observe patterns
2) Experimental phase: Formulate hypotheses that organise patterns ('explanatory organisation'). Some patterns will be explained by the model dynamics directly. Some patterns will be explained through other observed patterns ('indirect explanation').
3) Explanatory phase: Relate organisation of observations to the theories about natural phenomena, explain the consequences.
[^301]This idea is very similar to recent work in philosophy of science (particularly philosophy of economics), where models and simulations can be seen as 'credible worlds', constructed to explore general theoretical principles (Sugden, 2000, 2009). Recent work on model-based theorising (Weisberg, 2007; Godfrey-Smith, 2006), largely based on examples from population biology, incorporates similar stages of modelling, going through stages of model construction, model analysis (stage 1 and 2) and the (optional) stage of the exploration of how well the model 'fits' the target system, and thus which general principles can be learned from the model (stage 3). Here too, there are questions about how these constructed, and often highly simplified worlds relate to real world systems.

There are two main ways model-world fit can be evaluated, both found in cognitive science. One is to consider the translation from the target system to the model/simulation, for example when models and simulations are used as abstract analogues of a concrete phenomenon (e.g. eye saccades during reading). Here knowledge about the target system is used to construct the model (though of course it may involve simplification, idealisation and so on). Model-world fit is analysed by considering how different the model is from our knowledge of the target system (perhaps it fails to capture important structural or causal features).

The other way to consider model-world fit is to consider the converse; the translation from a model/simulation back to a the target system. Here, a model-simulation is constructed using relevant background knowledge of a family of targets, but not intended to represent any particular target system. Once the principles governing the model/simulation are established, the researcher then looks to see if the model/simulation actually captures any targets in the real world. If any are found, then the researcher infers that the same principles govern the model/simulation and these target systems. It is this kind of translation that is found in Weisberg and Godfrey-Smith's description of model-based science, where justifications must be given for inferences from properties found in a model to properties of a target system. In language evolution, this inference goes from the results of iterated learning paradigms to real cases of language evolution.

In order to support this inference, there are a number of ways that model-target fit can be evaluated. The role that the notion of robustness plays in these evaluations is discussed below, and linked to other cases in cognitive science.

## Model-Target Fit and Robustness

## Robust properties

Robust properties are those that are consistently found across a set of different models, suggesting that it is an 'important' property that derives not from incidental features of the model (e.g. its particular assumptions and simplifications), but from the core structure found across all the models (Levins, 1966; Wimsatt, 1981; Weisberg, 2006; Weisberg \& Reisman, 2008).

As Levins originally put it, "if these models, despite their different assumptions, lead to similar results, we have what we can call a robust theorem that is relatively free of the details of the model. Hence, our truth is the intersection of independent lies." (Levins, 1966, p. 20). As detailed above, in the case of language evolution the emergence of linguistic structure is a robust property of iterated learning systems under the pressures of expressivity and learability.

However, Weisberg and Reisman (2008) identify several (related) kinds of robustness: parameter, structural and representational robustness. First, a model can be robust (i.e. give roughly the same results) for a wide range of parameter settings. Second, a set of crucial causal components that make up the 'core structure' noted above give rise to structural robustness. Finally, representational robustness refers to the range of model descriptions (e.g. programming languages or mathematical formalisms used) for which a model still gives rise to the same results. Of most relevance in the sections below are structural and representational robustness, detailed below.

## Multi-model method

One way that robustness analyses figure in language evolution research is through the use of different types of models. Researchers in the field of cultural evolution see models as tools, not necessarily as reflections of theories (e.g. Cornish, Tamariz, \& Kirby, 2009). As tools, they need not commit the researcher to particular methodological approaches (agentbased or mathematical) nor particular theories of cognition (e.g. humans as Bayesian learners or frequentist learners).

Testing the same ideas across a range of mathematical models and computational simulations, but also across different formalisms or experimental setups within each broad method, is a standard way to explore robust properties. This corresponds to both structural and representational robustness noted above. The same core structures are found across these models, but other variables can differ (e.g. learning algorithm, size of population). That these additional variations do not affect the core finding (emergence of linguistic structure) provides support for the claim that the core structures really are the essential causal components in these models. Further, that these models can be constructed in a range of computational and mathematical frameworks, and still give rise to emergence linguistic structure, means that these results are not related to specific features of these frameworks - they are representationally robust. Both provide support for the claim that linguistic structure is a robust property of ILM models.

However, this method of constructing and comparing a range of computational and mathematical models is not widely found in current practice in cognitive science, where researchers are often wedded to particular frameworks. The field of language evolution may require this kind of approach because its precise object of study is still being identified, so the key variables and relations to consider are not yet obvious. For example, it is difficult to intuit about the abstract properties of a culturally transmitted linguistic system underpinned
by genetic constraints (Christiansen \& Kirby, 2003).

## Human simulations

Another crucial way of exploring the relations between simulations and real systems, used in language evolution, is to replace computer agents in agent-based models with human subjects (Kirby, Cornish, \& Smith, 2008). The inclusion of this much wider range of structural features (such as real biological learning mechanisms) provides a strong test for claims about the core structural features and representational robustness of ILM models. However, that subjects already know natural languages is seen as a strong confounding factor in the interpretation of human simulations. There are two standard responses to this.

Firstly, there are the experimental controls. Compositional structure does not always emerge in these simulations, only when both the pressures of expressivity and learnability are applied. Also, the human participants, far from deliberately introducing familiar linguistic structures, rarely expressed an understanding of the principles behind the experiment, most even not noticing that they were being tested on meanings that they had not been taught (Kirby, Cornish, \& Smith, 2008).
Secondly, there is an argument based on the notion of robustness. Human simulations can be seen as further explorations of structural and representational robustness, that include both actual biological mechanisms (e.g. learning), but also potentially problematic factors (subjects already know natural languages). That linguistic structure still reliably emerges from iterated learning paradigms under pressures for learnability and expressivity, even when significant variables are changed, provides more evidence that the emergence of linguistic structure is a robust property of these models.

## Summary: Learning from simulations in the ILM

Since there is little empirical evidence about the facts of language evolution, the strength of model-target 'fit' may not be most convincingly based on the comparison to the real world, but on the robust properties found under various simplicifications and idealisations of real world target systems. Even if we are not sure of how precisely to represent a target system, the fact that many highly idealised representations of the same system make a similar prediction (e.g. emergence of compositionality), can be sufficient to suggest that this is what is happening in the target system itself. In this case, theoretical claims based on robust properties already have some degree of 'fit' with target systems, purely because of the nature of robustness.

## Robustness in Cognitive Science

The sections above illustrate how to productively use robustness notions in cognitive science. Modelling the same process over different formalisms or frameworks, and over computational, mathematical and human models and simulations, can help identify general principles and core variables and constraints. Each particular model need not have high modelworld fit, but together the emergence of a robust property cuts
through these problems. However, there are of course limits to robustness approaches, and, as with any other methodology, there are no clear cut rules about its application.

First, it is not clear, in general, how many independent lines of evidence (i.e. different models/simulations) one must have in order to identify a 'real' robust property, and the core causal structure that gives rise to it. Yet this may become more clear in specific contexts. In some cases two or three very different models/simulations (e.g. containing very different assumptions) might be sufficient to warrant an inference to the existence of a robust property. Alternatively, a larger group of similar models/simulations that together survey a wide range of alternative assumptions may be required. The identification of surprising convergence across different models will always be context dependent.

Second, robustness analyses can be misleading. One might identify a robust property, and the causal structure that gives rise to it, on the basis of different models that all incorporate the same erroneous assumptions. An inference that this causal structure is also found in the world, and explains some cognitive phenomenon, would therefore be unwarranted. One might also make mistakes in identifying the robust property (perhaps it is more or less specific than found in the models), and what the relevant causal structure is.

Clearly, robustness approaches are defeasible, just as any other methods are. Yet the promotion of the use of a wide range of different frameworks found in robustness-based approaches may minimise the kind of errors identified above, or identify them earlier than approaches that stay within one modelling framework. Further, models will still be held accountable to the usual range of relevant empirical and theoretical work. Therefore robustness analyses should be seen as a additional methodological tool that can help to test and strengthen theoretical claims that are largely made on the basis of models and simulations.

Finally, one might question whether robustness analyses can not only support theoretical claims (as illustrated above), but also show when they are unfounded. In fact, it seems that criticisms of overfitting, highly parameterised models (e.g. Pitt \& Myung, 2002), often based on model comparisons that include controls for model complexity (e.g. Hansen \& Yu, 2001), do just this. Low-parameter models with a stable core of causal components tend to be favoured in cognitive science, which is entirely consistent with a preference for high levels of parameter and structural robustness.

One implication of the use of robustness analyses is that traditional debates about the validity of different modelling approaches may not be constructive. For example, researchers have debated whether 'bottom-up' or 'top-down' approaches are the most productive for researching cognition (Griffiths, Chater, Kemp, Perfors, \& Tenenbaum, 2010; McClelland et al., 2010). With a robustness approach, the question is not about which provides more realistic models or which can provide clearer analytic results, but how they can complement each other's strengths and weaknesses. In this
case, it makes sense for researchers to use both approaches to identify robust properties, and thus converge on mutually supported theories.

## Conclusion

This paper used research in language evolution to illustrate a robustness approach to modelling in cognitive science. It showed how robustness analyses support the identification of linguistic structure as robust property of the processes of cultural transmission, as modelled and simulated across a range of mathematical models, computational simulation frameworks and human experiments.
The robustness approach outlined here strongly contrasts with typical practices in cognitive science, where the aim is often to develop a single model, developed in a specific modelling framework, to account for a narrow range of data. Often, though not always, such practices generate models that lack predictive power and generality, and parameter and structural robustness. With the realisation that such models may have little to do with actual cognitive processes, pressures from new statistical methods of model comparisons are starting to force alternative methods of model construction.

A robustness approach directly promotes the development of models with high parameter, structural and representational robustness. These are often seen as positive features of models, as the output of such models can be traced to the activities of a set of core causal components, not to specific parameter settings, or to artifactual features of the modelling framework used. The use of multiple modelling/simulation frameworks also makes it easier to identify artefacts and core variables in models. Finally, robustness analyses offer a way of providing support for theoretical claims when there is little direct empirical evidence available. In this case, a robustness approach stands as a powerful alternative approach to modelling in cognitive science, and one we recommend highly.

## Acknowledgments

Many thanks to Hannah Cornish for useful discussions.

## References

Batali, J. (2002). The negotiation and acquisition of recursive grammars as a result of competition among exemplars. In T. Briscoe (Ed.), Linguistic evolution through language acquisition: Formal and computational models (chap. 5). Cambridge University Press.
Bedau, M. (1998). Philosophical content and method of artificial life. In T. Bynum \& M. J. H. (Eds.), The digital phoenix: How computers are changing philosophy (p. 135152). Basil Blackwell, Oxford.

Chomsky, N. (1965). Aspects of the theory of syntax. MIT Press (MA).
Chomsky, N. (2011). On the poverty of the stimulus. In Talk given at UCL division of psychology and language sciences, 11/10/2011.

Christiansen, M., \& Kirby, S. (2003). Language evolution: The hardest problem in science? Studies in the evolution of language, 3, 1-15.
Cornish, H., Tamariz, M., \& Kirby, S. (2009). Complex adaptive systems and the origins of adaptive structure: what experiments can tell us. Language Learning, 59, 187-205.
Di Paolo, E. A., Noble, J., \& Bullock, S. (2000). Simulation models as opaque thought experiments. In M. A. Bedau, J. S. McCaskill, N. Packard, \& S. Rasmussen (Eds.), Seventh international conference on artificial life (p.497-506). MIT Press, Cambridge, MA.
Flynn, E. (2008). Investigating children as cultural magnets: do young children transmit redundant information along diffusion chains? Philosophical Transactions of the Royal Society B: Biological Sciences, 363(1509), 3541.
Godfrey-Smith, P. (2006). The strategy of model-based science. Biology and Philosophy, 21(5), 725-740.
Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., \& Tenenbaum, J. (2010). Probabilistic models of cognition: exploring representations and inductive biases. Trends in Cognitive Sciences, 14(8), 357-364.
Hansen, M. H., \& Yu, B. (2001). Model selection and the principle of minimum description length. Journal of the American Statistical Association, 96(454), 746-774.
Kirby, S. (2000). Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners. In C. Knight (Ed.), The evolutionary emergence of language: Social function and the origins of linguistic form (pp. 303-323). Cambridge University Press.
Kirby, S., Cornish, H., \& Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. PNAS, 105(31), 10681-10686.
Kirby, S., Dowman, M., \& Griffiths, T. L. (2007). Innateness and culture in the evolution of language. PNAS, 104(12), 5241-5245.
Kirby, S., \& Hurford, J. (2002). The emergence of linguistic structure: An overview of the iterated learning model. In A. Cangelosi \& D. Parisi (Eds.), Simulating the evolution of language (pp. 121-148). Springer Verlag, London.
Kirby, S., Smith, K., \& Cornish, H. (2008). Language, Learning and Cultural Evolution: How linguistic transmission leads to cumulative adaptation. In R. Cooper \& R. Kempson (Eds.), Language in flux: Dialogue coordination, language variation, change and evolution. London: College Publications.
Levins, R. (1966). The strategy of model building in population biology. American Scientist, 421-431.
Matthews, C., Kirby, S., \& Cornish, H. (2010). The cultural evolution of language in a world of continuous meanings. In A. Smith, M. Schouwstra, B. de Boer, \& K. Smith (Eds.), The evolution of language: Proceedings of evolang 2010.
McClelland, J., Botvinick, M., Noelle, D., Plaut, D., Rogers, T., Seidenberg, M., et al. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cogni-
tion. Trends in Cognitive Sciences, 14(8), 348-356.
Navarro, D. J., \& Perfors, A. F. (2011, Jan). Hypothesis generation, sparse categories, and the positive test strategy. Psychol Rev, 118(1), 120-34.
Pitt, M. A., \& Myung, J. (2002). When a good fit can be bad. Trends in Cognitive Sciences, 6, 421-425.
Reali, F., \& Griffiths, T. (2010). Words as alleles: connecting language evolution with bayesian learners to models of genetic drift. Proceedings of the Royal Society B: Biological Sciences, 277(1680), 429-436.
Silvey, C., Kirby, S., \& Smith, K. (2012). The coevolution of words and meanings. In Proceedings of the european human behaviour and evolution association confrence, 2012.
Smith, K., Kirby, S., \& Brighton, H. (2003). Iterated learning: a framework for the emergence of language. Artificial Life, 9(4), 371-386.
Smith, K., Tamariz, M., Cornish, H., \& Kirby, S. (2013). Language structure is a trade-off between compression and expression. In Proceedings of the 35th annual cognitive science society conference (cogscil3).
Sugden, R. (2000). Credible worlds: the status of theoretical models in economics. Journal of Economic Methodology, $7(1), 1-31$.
Sugden, R. (2009). Credible worlds, capacities and mechanisms. Erkenntnis, 70(1), 3-27.
Swarup, S., \& Gasser, L. (2008). Simple, but not too simple: Learnability vs. functionality in language evolution. In The seventh evolution of language conference.
Tamariz, M., Cornish, H., Roberts, S., \& Kirby, S. (2012). The effect of generation turnover and intelocutor negotiation on linguistic structure. In T. C. Scott-Phillips et al. (Ed.), The Evolution of Language: Proceedings of the 9th International Conference (p. 555-556). World Scientific.
Verhoef, T., \& Boer, B. de. (2012). Holistic or synthetic protolanguage: Evidence from iterated learning of whistled signals. In T. C. Scott-Phillips, M. Tamariz, E. A. Cartmill, \& J. R. Hurford (Eds.), The evolution of language: Proceedings of the 9th international conference (evolang9) (p. 368-375). World Scientific.

Vogt, P. (2005). Meaning development versus predefined meanings in language evolution models. In In packkaelbling (pp. 1154-1159).
Weisberg, M. (2006). Robustness analysis. Philosophy of Science, 73(5), 730-742.
Weisberg, M. (2007). Who is a modeler? The British journal for the philosophy of science, 58(2), 207-233.
Weisberg, M., \& Reisman, K. (2008). The robust volterra principle. Philosophy of science, 75(1), 106-131.
Wimsatt, W. (1981). Robustness, reliability, and overdetermination. Scientific inquiry and the social sciences, 124-163.
Worgan, S. F., \& Damper, R. I. (2008). Removing 'mindreading' from the iterated learning model. In A.D.M. Smith et al. (Ed.), The evolution of language: Proceedings of the 7th international conference (EVOLANG7), Barcelona. World Scientific Publishing Co.

# Context Dependent Utility: Modeling Decision Behavior Across Contexts 

Jonathan Ito (ito@ict.usc.edu)<br>Stacy Marsella (marsella@ict.usc.edu)<br>Institute for Creative Technologies, University of Southern California<br>12015 Waterfront Drive<br>Playa Vista, CA 90094-2536 USA


#### Abstract

One significant challenge in creating accurate models of human decision behavior is accounting for the effect of context. Research shows that seemingly minor changes in the presentation of a decision can lead to drastic shifts in behavior; phenomena collectively referred to as framing effects. Previous work has developed Context Dependent Utility (CDU), a framework integrating Appraisal Theory with decision-theoretic principles. This work extends existing research by presenting a study exploring the behavioral predictions offered by CDU regarding the multidimensional effect of context on decision behavior. The present study finds support for the predictions of CDU regarding the impact of context on decisions: 1) as perceptions of pleasantness increase, decision behavior tends towards riskaversion; 2) as perceptions of goal-congruence increase, decision behavior tends towards risk-aversion; 3) as perceptions of controllability increase, i.e., perceptions that outcomes would have been primarily caused by the decision maker, behavior tends towards risk-seeking.


Keywords: Decision; Appraisal; Context; Framing; Utility;

## Introduction

Descriptive models of human decision behavior seek to accurately describe and predict the decisions people actually make. Creating these models is vital for advancing a more complete understanding of the human decision process and requires addressing the factors that systematically bias the perception and evaluation of decisions.

One significant challenge in creating accurate models of human behavior is accounting for the effect of context on decision behavior. Research has shown that seemingly minor changes in the presentation, or framing, of a decision problem can lead to drastic shifts in behavior; phenomena collectively referred to as framing effects. In a seminal study, now referred to as the Asian Disease Study, Tversky and Kahneman (1981) showed that when outcomes were described, or framed, as gains participants tended to be risk-averse; however, when the same outcomes were framed as losses participants tended to be risk-seeking. Subsequent studies involving domains as diverse as financial planning (Schoorman, Mayer, Douglas, \& Hetrick, 1994), Acquired Immune Deficiency Syndrome (AIDS) (Levin \& Chapman, 1990), Breast Self Examinations (Meyerowitz \& Chaiken, 1987), taxpayer compliance (Liu, Xia, \& Xu, 2011), and judgments of website quality (Hartmann, De Angeli, \& Sutcliffe, 2008) have also demonstrated framing effects to varying degrees. In addition to gain-loss framing, framing can also involve the role of the decision maker (Wagenaar, Keren, \& Lichtenstein, 1988), the salience of outcomes (Van Schie \& Van Der Pligt, 1995),
decision domain (Vartanian, Mandel, \& Duncan, 2011), and perceived need (Mishra \& Fiddick, 2012).

Despite the highly multidimensional nature of context, the prevalence of framing effects in numerous domains, and the profound impact they can have on the decision process, very few decision models explicitly address the multidimensional impact of context on decisions. Existing decision-theoretic approaches which do address framing and context are generally limited by a narrow, one-dimensional view of context. For instance, Prospect Theory (Kahneman \& Tversky, 1979) and Cumulative Prospect Theory (Tversky \& Kahneman, 1992) model the effect of context only to the extent that it applies to outcomes perceived as either gains or losses. Therefore, to address the multidimensional effect of context on decision behavior, previous work has developed Context Dependent Utility (CDU), a framework which seeks to explicitly model the multidimensional impact of context on decision behavior through the integration of Appraisal Theory and decision-theoretic models (Ito \& Marsella, 2011). This work extends previous research by presenting an experimental study exploring the behavioral predictions offered by CDU regarding the multidimensional effect of context on decision behavior. In particular, the results support the behavioral predictions of CDU and suggest that it can dramatically improve the modeling of human decision behavior across distinct contexts.

## Context Dependent Utility

In previous work, Context Dependent Utility (CDU) was developed to explicitly model the multidimensional impact of context on decision behavior (Ito \& Marsella, 2011). The CDU process consists of two primary components: the computational appraisal of the decision situation and an evaluation function aggregating the appraisal information into a real-valued utility.

Appraisal Theory (Lazarus, 1991) is a psychological theory which addresses the process by which emotions arise given the subjective evaluation and interpretation of a situation. Because appraisal theory provides a well-defined framework for the interpretation of features of a situation in terms of their significance, we argue that it provides the means to identify, encode, and integrate contextual information into the decision process. Appraisal as implemented by CDU consists of three distinct evaluations: pleasantness, goal congruence, and control. Each appraisal is defined over individual outcomes as a function of diminishing sensitivity evaluated with respect to some reference point. This follows from the
principle that emotions and appraisals arise primarily from the changes, relative to some reference point, associated with them rather than from any inherent properties of the outcomes themselves (Frijda, 2007). The general appraisal function is shown in (1), in which $0 \leq k \leq 1$, controls the sensitivity of the appraisal.

Pleasantness is implemented as an evaluation of value made with respect to the value of the status quo, $v_{\mathrm{sq}}$, as in (2). Goal congruence is implemented as an evaluation of value made with respect to the value of the aspiration outcome, $v_{\mathrm{ao}}$, as in (3). Control is a measure of the degree to which an outcome will be perceived to have been primarily caused by the decision maker and is implemented as an evaluation of $d e$ cumulative probability, i.e., the total probability of obtaining an outcome at least as preferred as it, made with respect to the probability of the control threshold, $p_{c t}$, as in (4). Note that the decumulative probability representation given in (5) requires that outcomes are in ascending order of value, i.e., $v_{i} \leq v_{i+1}$.

$$
\begin{align*}
\operatorname{appraise}(x, r e f, k) & =\left\{\begin{aligned}
(x-r e f)^{k} \text { if } x-r e f \geq 0 \\
-(r e f-x)^{k} \text { if } x-r e f<0
\end{aligned}\right.  \tag{1}\\
\text { pleas }\left(v_{i}\right) & =\text { appraise }\left(v_{i}, v_{s q}, k_{\text {pleas }}\right)  \tag{2}\\
\operatorname{gc}\left(v_{i}\right) & =\text { appraise }\left(v_{i}, v_{a o}, k_{g c}\right)  \tag{3}\\
\operatorname{ctrl}\left(D_{i}\right) & =\operatorname{appraise}\left(D_{i}, p_{c t}, k_{c t}\right)  \tag{4}\\
D_{i} & =\sum_{j=i}^{n} p_{j} \tag{5}
\end{align*}
$$

The decision evaluation component of CDU is implemented using rank-dependent utility (Quiggin, 1982) as seen in (6). Rank-dependent utility models allow for nonlinear decision weights while maintaining stochastic dominance. The utility function consists of a linearly weighted combination of pleasantness and goal congruence as seen in (7). Outcome weight, $\pi_{i}$, is defined in the standard rank-dependent manner as seen in (8) and (9) in which $a$ and $b$ serve to normalize the weighting function such that $w(0)=0$ and $w(1)=1$.

$$
\begin{align*}
C D U & =\sum_{i=1}^{n} \pi_{i} u\left(v_{i}\right)  \tag{6}\\
u\left(v_{i}\right) & =\left(\beta \text { pleas }\left(v_{i}\right)+(1-\beta) g c\left(v_{i}\right)\right)  \tag{7}\\
\pi_{i} & = \begin{cases}w\left(D_{i}\right)-w\left(D_{i+1}\right) & \text { if } i<n \\
w\left(D_{i}\right) & \text { if } i=n\end{cases}  \tag{8}\\
w\left(D_{i}\right) & =a\left(\operatorname{ctrl}\left(D_{i}\right)\right)+b \tag{9}
\end{align*}
$$

Since appraisals are implemented as functions of diminishing sensitivity with respect to reference points, the underlying utility function becomes increasingly concave as perceptions of pleasantness and goal-congruence increase. Similarly, the underlying weighting function becomes increasingly concave as perceptions of controllability increase. Furthermore, ac-
cording to the principles of the rank-dependent utility formalization employed in CDU, a concave utility function is associated with risk-aversion whereas a concave weighting function is associated with risk-seeking. Therefore, CDU offers the following set of behavioral predictions regarding the effects of pleasantness, goal congruence, and control on the decision process:

Hypothesis 1 As outcomes are perceived as increasingly pleasant, behavior will tend towards risk-aversion

Hypothesis 2 As outcomes are perceived as increasingly goal-congruent, behavior will tend towards risk-aversion

Hypothesis 3 As outcomes are perceived as increasingly controllable, behavior will tend towards risk-seeking

This work presents the results of an experimental framing study designed to test the hypotheses offered by the CDU framework. The study presents participants with a scenario in which they are asked to decide between two competing plans to prevent school dropouts.

## Method

## Participants

For the study, 525 participants from the United States were recruited through Amazon Mechanical Turk.

Each participant received a payment of $\$ 0.40$ for participation. The sample had a self-reported gender distribution of 319 male ( $61 \%$ ) and 206 female ( $39 \%$ ). The median age range was 22 to 34 years with $85 \%$ of participants below 45 years of age. The majority of participants self-identified as white (78\%). Approximately half of participants (50\%) have also completed a 2 year college degree or higher.

Risk propensity, measured using the Subjective Risk Assessment instrument (Dohmen et al., 2005), uses a 7-point scale in which 1 represents being very prepared to take risks, 4 represents being risk-neutral, and 7 represents very unwilling to take risks. The mean of the subjective risk assessment measure for all participants was 3.46 with a median value of 3 and a standard deviation of 1.4 representing a slight overall self-reported tendency towards risk taking.

Additionally, only 42 participants ( $8 \%$ ) self-identified as possessing some real-life expertise involving the prevention of school dropouts.

## Procedure and Design

The study was administered as an anonymous online questionnaire implemented via Qualtrics (Qualtrics Labs Inc., Provo, UT). Before presentation of the decision scenario, demographic information including gender, age, race/ethnicity, and highest level of education was collected along with a measure of subjective risk propensity. Instructions adapted from previous studies administered by Schneider (1992) were then presented to participants regarding the upcoming decision task. In addition, an embedded Instructional Manipulation Check originally developed by Oppenheimer, Meyvis,
and Davidenko (2009) was adapted and employed in the present study to ensure that the instructions were read and to encourage participant attentiveness.

The decision scenario was then presented to the participants. The scenario was based on one originally designed by Fagley, Miller, and Jones (1999) to test standard gainloss framing, but subsequently expanded in the present study to include additional considerations of context. It presented participants with two possible plans to prevent school dropouts: one plan always results in some students dropping out whereas the other plan results in either all students dropping out or no students dropping out. According to decisiontheoretic formalizations of risk, a preference for the plan which always results in the same outcome (some students dropping out) is characterized as a preference towards riskaversion whereas preference for the plan which results in one of two potential outcomes (one good and one bad) is characterized as a preference towards risk-seeking. As in most framing studies, regardless of frame, the underlying values, i.e., numbers of students dropping out or staying in school and outcome occurrence probabilities, associated with the scenario remain unchanged.

The framing of each scenario involved the explicit manipulation of the context associated with the appraisal dimensions of pleasantness, goal congruence, and control. The manipulation of pleasantness was associated with the description of outcomes as gains or losses such that outcomes described as losses were presumed to be relatively unpleasant whereas outcomes described as gains were relatively pleasant. Therefore, in the loss condition, outcomes were described by the number of students that drop out of school; for the gain condition, outcomes were described by the number of students that stay in school; for the neutral condition, outcomes were described using both the number of students that drop out and stay in school.

The manipulation of goal congruence was based on a previous study conducted by Payne, Laughhunn, and Crum (1981) and involved specifying different evaluation criteria to establish what constitutes a successful, or goal-congruent, outcome. In particular, participants were informed that their performance would be evaluated in comparison to the average retention rate, i.e., the percentage of students that stay in school, of other schools in the district. Therefore, in the low retention condition the expected retention rate was $5 \%$ (50 students stay in school or 950 drop out); in the neutral retention condition the retention rate was $40 \%$ ( 400 students stay in school or 600 drop out); and in the high retention condition the expected rate was $75 \%$ ( 750 students stay in school or 250 drop out).

The manipulation of control, derived from research on loci of control (Rotter, 1966), involved depicting the source of uncertainty as either arising from chance events or from the ability of the decision maker. Uncertainty arising from chance events suggests that outcomes are uncontrollable and therefore not caused by the decision maker whereas uncertainty
regarding the ability of the decision maker suggests that outcomes are controllable and will be perceived as having been caused by the decision maker. Therefore, in the chance condition, uncertainty was depicted as arising from the random selection, i.e., lottery, of funding applications; In the ability condition, the source of uncertainty was described as arising from the hypothetical ability of the participant to write a persuasive funding application; and the neutral condition involved a mixture of the two. A full listing of a school dropout scenario in which outcomes are described as gains, retention rate is low, and the source of uncertainty is depicted as arising from one's ability is given in Appendix .

The primary dependent variables in the decision task consisted of both a dichotomous choice between plans, as found in most standard framing studies, and a continuous strength-of-preference response, which included an option indicating indifference, as advocated by Levin, Gaeth, Schreiber, and Lauriola (2002).

After completion of the decision task, participants were asked to directly evaluate the pleasantness, goal congruence, and control of the three potential outcomes, i.e., the sure outcome of the risk-averse alternative and the two potential outcomes of the risky alternative. This served as a manipulation check to ensure that the contextual manipulations employed in the scenario were effective. Participants were then offered the opportunity to express any additional factors that may have influenced their decisions. Finally, participants were asked to indicate whether or not they perceived themselves as possessing any type of expertise relating to school dropouts.

The study was conducted as a $3 \times 3 \times 3$ between-subjects factorial in which both the presentation order of the two dropout prevention plans and the ordering of the two potential outcomes of the risky alternative were balanced. Table 1 illustrates the various factors and the number of participants that were assigned to each combination of factors. To ensure the integrity of the data, responses deemed inconsistent by an automated consistency-verification program were ignored. An inconsistent response was defined as one in which the dichotomous response regarding participant preference between plans does not agree with the associated continuous strength-of-preference measure. For example, a stated dichotomous preference for the sure alternative coupled with a strength-of-preference score favoring the risky alternative was deemed inconsistent.

## Results

As discussed earlier, two distinct but related dependent measures were assessed: a dichotomous choice between alternatives and a continuous strength-of-preference measure. According to the dichotomous measure, 327 participants ( $62 \%$ ) preferred the sure alternative while 198 participants preferred the risky alternative ( $38 \%$ ). Similarly, according to the strength-of-preference measure, in which 1 indicates very strong preference for the sure alternative, 4 indicates indifference, and 7 indicates very strong preference for the risky alternative, mean choice preference was 3.61 with a median

Table 1: Number of Participants per Condition

| Gain/Loss | Retention | Source of Uncertainty |  |  |
| :--- | :--- | ---: | ---: | ---: |
|  | Rate | Chance | Neutral | Ability |
| Loss | Low | 20 | 18 | 23 |
|  | Neutral | 20 | 26 | 18 |
|  | High | 26 | 23 | 22 |
| Neutral | Low | 20 | 18 | 17 |
|  | Neutral | 21 | 21 | 21 |
|  | High | 21 | 22 | 19 |
| Gain | Low | 20 | 19 | 15 |
|  | Neutral | 15 | 17 | 20 |
|  | High | 13 | 16 | 14 |

of 3 and a standard deviation of 1.71. This indicates a slight preference for the sure alternative.

Using the continuous strength-of-preference measure as the dependent variable, a 3-way ANOVA found significant main effects for retention rate $F(2,498)=25.99, p<0.001$, $\eta^{2}=0.09$, and source of uncertainty $F(2,498)=6.53, p<$ $0.01, \eta^{2}=0.02$, while the effect of gain/loss descriptions $F(2,498)=1.84, p=0.16, \eta^{2}=0.01$ approached significance. ${ }^{1}$ Figures 1, 2, and 3 show respectively the main effects of gain/loss description, retention rate, and source of uncertainty on mean preference strength. No significant interactions were detected.


Figure 1: Experiment 1 Effect of Gain/Loss Descriptions on Preference with $95 \%$ Confidence Intervals and Significance Levels ( 0 *** 0.001 ** $0.01 * 0.05$ )

Post-hoc $t$-tests on the effect of gain/loss description on preferences revealed a significant difference between the loss and gain conditions, $t(332.59)=2.29, p<0.05$. For the effect of retention rate on preferences, $t$-tests revealed that the

[^302]

Figure 2: Experiment 1 Effect of Retention Rate on Preference with $95 \%$ Confidence Intervals and Significance Levels ( 0 *** 0.001 ** 0.01 * 0.05 )
difference between the low and high conditions $t(340.55)=$ $-5.93, p<0.001$ and between the neutral and high conditions $t(352.21)=-6.69, p<0.001$ were significant. For the effect of source of uncertainty on preferences, the differences between the chance and the ability conditions $t(328.1)=$ 3.31, $p<0.01$ and between the neutral and ability conditions $t(347.17)=2.39, p<0.05$ were significant.

In addition to the primary dependent variables, direct evaluations of pleasantness, goal congruence, and control over each outcome were assessed to ensure that the contextual manipulations had the intended effect on their associated appraisals. A one-way MANOVA on the effect of gain/loss descriptions on appraisals of pleasantness for each of the three outcomes showed an effect approaching significance $F(6,1042)=1.74, p=0.11$. Similar MANOVAs showed significant effects for retention rate on appraisals of goal congruence $F(6,1042)=18.65, p<0.001$ and source of uncertainty on appraisals of control $F(2,1042)=16.5, p<0.001$ for each of the three outcomes.

## Discussion

The goals of the study were to examine how different aspects of context affect decision behavior and whether shifts in preference in response to contextual changes, i.e., framing effects, are consistent with the predictions offered by CDU.

The experimental results support both the multidimensional effect of context on decision behavior and the predictions offered by CDU with respect to the effects of context. In particular, the present study finds strong support that the contextual dimensions associated with pleasantness, goal congruence, and control do affect decision behavior in the direction predicted by CDU. CDU predicts that when outcomes are described as gains opposed to losses, decision makers will tend to act in a more risk-averse manner to maintain pleasantness. CDU also predicts that when the standards for success, i.e.,


Source of Uncertainty

Figure 3: Experiment 1 Effect of Source of Uncertainty on Preference with $95 \%$ Confidence Intervals and Significance Levels ( 0 *** 0.001 ** $0.01 * 0.05$ )
retention rates, are low, decision makers tend to act in a more risk-averse manner to maintain goal congruency compared to when standards of success are high. Finally, CDU predicts that when the source of outcome uncertainty is depicted as arising from ability, e.g., persuasive writing ability, decision makers tend to act in a more risk-seeking fashion to capitalize on the perceived controllability of the situation compared to when the source of uncertainty is depicted as arising from chance events.

## Conclusion

One significant challenge in creating accurate, descriptive models of human behavior is accounting for the effect of context on decision behavior. Existing approaches at modeling context and its effects on decision behavior, i.e, framing effects, are generally limited by a one-dimensional view of contextual influence and therefore lack the descriptive flexibility to account for a broad range of behavior. Therefore, this work extends previous research (Ito \& Marsella, 2011) on Context Dependent Utility (CDU), a decision framework which seeks to explicitly model the multidimensional impact of context on decision behavior, by presenting experimental data supporting the need for accurate, multidimensional models of contextual influence on decision behavior. In particular, this work presents the results of a study in which participants are asked to choose between two alternative plans (one risky and one risk-averse) to prevent school dropouts. Additionally, to examine the effect of context on overall decision behavior, the context of the decision task is varied between subjects along the following dimensions: whether outcomes are described in terms of gains or losses; the retention rate used to determine the degree to which an outcome is considered successful; and the portrayal of the source of uncertainty in the scenario, i.e., whether outcome variability depends on chance
or ability-based factors.
The present study shows strong support for all of the behavioral predictions offered by CDU regarding the impact of context on decision behavior: 1) as the overall perception of pleasantness increases, which is associated with the description of outcomes as gains rather than losses, decision behavior tends towards risk-aversion; 2) as the overall perception of goal-congruence increases, which is associated with lower retention rates implying lower standards of success, decision behavior tends towards risk-aversion; 3) as the overall perception of controllability, i.e., the perception that outcomes would have been primarily caused by the decision maker, increases, which is associated with depicting the source of uncertainty as arising from one's ability, behavior tends towards risk-seeking. In sum, the present study illustrates the need for models which explicitly represent the multidimensional affect of context on behavior, such as CDU, especially when modeling decision behavior across contextually distinct situations.

## References

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., \& Wagner, G. (2005). Individual risk attitudes: New evidence from a large, representative, experimentally-validated survey. IZA discussion papers, 1730.
Fagley, N., Miller, P., \& Jones, R. (1999). The effect of positive or negative frame on the choices of students in school psychology and educational administration. School Psychology Quarterly, 14(2), 148-162.
Frijda, N. (2007). The laws of emotion. Lawrence Erlbaum Associates.
Hartmann, J., De Angeli, A., \& Sutcliffe, A. (2008). Framing the user experience: information biases on website quality judgement. In Proceedings of the twenty-sixth annual sigchi conference on human factors in computing systems (pp. 855-864).
Ito, J., \& Marsella, S. (2011). Contextually-based utility: An appraisal-based approach at modeling framing and decisions. In Twenty-fifth aaai conference on artificial intelligence (Vol. 2, pp. 60-65). Retrieved from http://ict.usc.edu/ ito/ito2011contextually.pdf

Kahneman, D., \& Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, 47(2), 263291.

Lazarus, R. (1991). Emotion and adaptation. Oxford University Press, USA.
Levin, I., \& Chapman, D. (1990). Risk taking, frame of reference, and characterization of victim groups in aids treatment decisions. Journal of Experimental Social Psychology, 26(5), 421-434.
Levin, I., Gaeth, G., Schreiber, J., \& Lauriola, M. (2002). A new look at framing effects: Distribution of effect sizes, individual differences, and independence of types of ef-
fects* 1. Organizational Behavior and Human Decision Processes, 88(1), 411-429.
Liu, H., Xia, Q., \& Xu, J. (2011). Taxpayer compliance in china: A prospect theory analysis. In Artificial intelligence, management science and electronic commerce (aimsec), 2011 2nd international conference on (pp. 7292-7297).
Meyerowitz, B., \& Chaiken, S. (1987). The effect of message framing on breast self-examination attitudes, intentions, and behavior. Journal of Personality and Social Psychology, 52(3), 500-510.
Mishra, S., \& Fiddick, L. (2012). Beyond gains and losses: The effect of need on risky choice in framed decisions. Journal of personality and social psychology, 102(6), 1136.
Oppenheimer, D., Meyvis, T., \& Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. Journal of Experimental Social Psychology, 45(4), 867-872.
Payne, J., Laughhunn, D., \& Crum, R. (1981). Further tests of aspiration level effects in risky choice behavior. Management Science, 953-958.
Quiggin, J. (1982). A theory of anticipated utility. Journal of Economic Behavior \& Organization, 3(4), 323-343.
Rotter, J. (1966). Generalized expectancies for internal versus external control of reinforcement.
Schneider, S. (1992). Framing and conflict: Aspiration level contingency, the status quo, and current theories of risky choice. Journal of Experimental Psychology: Learning, Memory, and Cognition, 18(5), 1040-1057.
Schoorman, F., Mayer, R., Douglas, C., \& Hetrick, C. (1994). Escalation of commitment and the framing effect: An empirical investigation. Journal of Applied Social Psychology, 24(6), 509-528.
Tversky, A., \& Kahneman, D. (1981). The framing of decisions and the psychology of choice. Science, 211(4481), 453.

Tversky, A., \& Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and uncertainty, 5(4), 297-323.
Van Schie, E., \& Van Der Pligt, J. (1995). Influencing risk preference in decision making: The effects of framing and salience. Organizational Behavior and Human Decision Processes, 63(3), 264-275.
Vartanian, O., Mandel, D., \& Duncan, M. (2011). Money or life: Behavioral and neural context effects on choice under uncertainty. Journal of Neuroscience, Psychology, and Economics, 4(1), 25.
Wagenaar, W., Keren, G., \& Lichtenstein, S. (1988). Islanders and hostages: Deep and surface structures of decision problems* 1. Acta Psychologica, 67(2), 175-189.

## School Dropout Scenario

## Initial Scenario Presentation

Imagine that you have been hired by the school district of a major city to combat the high number of student dropouts. It
is projected that 1000 students in your district will drop out of school during the next year if nothing is done.

## Action Description

Two plans exist to address the student dropout problem. Both plans require similar investments of money, time, and effort from your district. However, only one can be implemented. Based on other districts' experiences with these plans, estimates of the outcomes that can be expected from each plan can be made. Assume for purposes of this decision that these estimates of are accurate and are as follows:

Dropout Prevention Plan A Invest currently available funding in a smaller, relatively affordable dropout prevention program. This plan results in the following outcome:

- 400 of the 1000 students stay in school

Dropout Prevention Plan B Invest currently available funding in a larger dropout prevention program. However, the school district's current funding is insufficient to cover the full cost of this program. Therefore, its success is dependent on obtaining additional funding from the government. Funding approval for your dropout plan depends primarily on your ability to write a persuasive funding application. Historically, $2 / 5$ of your previous applications have been persuasive enough to receive funding. This plan results in one of two possible outcomes:

- $2 / 5$ chance that you are able to write a persuasive funding application. This results in sufficient funding and 1000 of the 1000 students staying in school
- $3 / 5$ chance that you are not able to write a persuasive funding application. This results in insufficient funding and 0 of the 1000 students staying in school


## Evaluation Criteria

The standard used to evaluate your performance will be the average student retention rate (the percentage of students that stay in school) obtained by other school districts in the state. Last year, the average retention rate for the other districts in the state was $\mathbf{5 \%}$. The same rate is expected this year.

If you select a dropout prevention program that ultimately leads to a higher retention rate than those obtained by other districts in the state, you will be evaluated as being successful. Of course, the higher the retention rate is in your district, the more successful you will be evaluated. On the other hand, if you select a dropout prevention plan that ultimately leads to a lower retention rate than the average rate of other districts, your own evaluation will be diminished. Again, the more retention rates are below the average level, the more diminished will be your own evaluation.

Which plan would you adopt?

# A computational exploration on the role of semantic memory in episodic future thinking 

Yuichi Ito (ito.yuichi@nagoya-u.jp), Taiji Ueno (taijiueno7@gmail.com), Shinji Kitagami (kitagami@cc.nagoya-u.ac.jp), and Jun Kawaguchi (kawaguchijun@nagoya-u.jp)<br>Department of Psychology, Graduate School of Environmental Studies, Nagoya University, Furo-cho, Chikusa-ku, Nagoya City, Aichi 4648601, JAPAN


#### Abstract

Episodic future thinking refers to a human cognitive process which generates successive predictions of events that are likely to occur in a cue-specific context in the future. An emerging view is that semantic memory as well as episodic memory contributes to this process, but the exact mechanism remains unclear. We built a computational model that learned to predict the next event upon a presented event (sequence prediction model). After learning the statistical structure in the training sequence, the model was tested for generating successive self-predictions of events triggered by a cue. The generated sequence of events captured some phenomenological features of patients with semantic dementia when the semantic system of the model was damaged. The role of semantics in episodic future thinking and the usefulness of a sequence prediction model are discussed.


Keywords: episodic future thinking; semantics; paralleldistributed processing model; sequence learning

## Introduction

We can project ourselves into the future despite the fact that we have never experienced it. The term episodic future thinking refers to a human ability to envision a plausible future event in a specific time and place (i.e., a specific context) (Atance \& O’Neill, 2001; Schacter, Addis \& Buckner, 2008). Over the last decade, researchers from various fields, including psychology, neuropsychology, and neuroimaging, have investigated episodic future thinking, focusing mainly on the contribution of episodic memory to constructing episodic future thought. More recently, data from patients with semantic dementia have suggested that semantic memory may also play a role (Irish, Addis, Hodges, \& Piguet, 2012). The current study used a computational model to investigate the mechanism by which semantic memory supports episodic future thinking.

## Role of Episodic Memory

The role of episodic memory has been suggested in various studies. For example, some neuroimaging studies have revealed a common neural network involved in the remembering of past, and in imagining future events (Szpunar, Watson, \& McDermott, 2007). These data are consistent with neuropsychological studies with amnesic patients (e.g., hippocampal amnesia or Alzheimer's disease) who showed simultaneous impairments in both remembering past episodes and imagining future events (e.g., Irish et al., 2012). Based on these findings, the
constructive episodic simulation hypothesis was proposed, which assumes that imagining future events requires a system that can retrieve detailed information stored in episodic memory and flexibly recombine them into coherent representations of future events (Schacter, Addis, \& Buckner, 2008). Further support for this idea comes from experimental psychology. For example, both retrieving an episode and imagining a future event are affected by a temporal distance factor in the same manner. Specifically, Addis, Wong, \& Schacter (2008) collected both the past events that participants recalled and the future events they generated, and classified detailed information in these outputs as either internal or external. Internal details are "episodic" information, meaning specific in time and place and related to the central events (i.e., the main event described by the participant). In contrast, external details are not specific in time and place. It was found that, in both recalling of past episodes and thoughts about future episodes, internal details lessened as participants were required to produce farther events from the present in both directions. This means that as episodic future thinking goes farther in terms of temporal distance from the present, the time and place (context) of the generated events deviates from those of the central events (central topic).

## Role of Semantic Memory

More recently, the role of semantic memory in episodic future thinking has also captured attention (Irish et al., 2012). D'Argembeau and Mathy (2011) suggest that construction of episodic future thought typically involves progressive conversion from general to more specific information such that access to general knowledge (semantics) precedes retrieval of time-specific episodic information. In other words, semantic memory provides a "framework" for construction of episodic future event representations, and then episodic information from the past is integrated to form a coherent and elaborated sequence of future events. A key support for this idea came from a study with neurological patients with semantic dementia, characterized by the progressive and insidious loss of conceptual knowledge about objects, facts and the meaning of words, yet preserved non-verbal episodic memory (Irish et al., 2012). Specifically, Irish et al. (2012) found that although their patients were as good at remembering past episodes as controls, their episodic future thoughts lacked internal details. In other words, the sequence of events they generated did not maintain the time and place information (context) that was
cued by an investigator. Note that this was not due to a difference in task difficulty because Alzheimer's disease patients in this study showed simultaneous impairments in both measures. Thus, this dissociative pattern suggests that even if episodic memory is relatively intact, loss of conceptual knowledge has an impact on episodic future thinking.
Motivated by these findings, we employed a paralleldistributed processing (PDP) modelling approach to investigate the mechanism by which semantic memory contributed to episodic future thinking. As we reviewed above, human experiments have provided significant insights, but each has its own limitation: It is relatively difficult to separate the contribution of episodic memory from that of semantic memory in healthy controls. Semantic dementia patients are the best test cases but their verbal outputs are limited such that it is difficult to probe their cognitive processing in detail. In contrast, computational modelling provides an ideal situation where we can directly look at the nature of computation/representations in the model to glean further insights into how semantic memory supports other cognitive processing (e.g., Woollams, Joanisse, \& Patterson, 2009).

## Future Prediction Model

Given there is no computational model for episodic future thinking in the literature, the initial step is to make some simplified assumptions so that the target cognition can be implemented in a computational model. A standard paradigm to probe episodic future thinking involves a presentation of a cue such as time/location/object (e.g., next year's birthday, or $50^{\text {th }}$ birthday, etc.), and a participant successively generates cue-specific predictions on what is likely to happen (e.g., a birthday cake is on a plate in a dining room $\rightarrow$ I blow the candle $\rightarrow$ my friend will pick out the candle $\rightarrow$ the friend will cut the cake $\rightarrow$ the friend will serve me a cake on a plate, etc.). The nature of this generation is successive such that the order of these example sentences cannot be at random. In other words, future thinking includes at least two aspects - computing cuespecific information and successively generating future predictions based on the corresponding previous prediction. Of course, these two aspects are not enough to account for the whole episodic future thinking processing. However, once we assume that episodic future thinking taps at least an ability to generate successive predictions based on the corresponding previous prediction upon a time-/location-/object-specific cue, then there is an existing computational model by Elman (1990) that we can adopt and modify for the current purpose. This model was trained for predicting the next alphabetic letter in an artificial language. Specifically, the model received a 6-bit binary vector, which represented one of the alphabetic letters, and the model was trained for predicting the next 6-bit binary input vector. The presented sequence was not random, but there was a statistical structure regarding what was likely to come next (artificial grammar). The model learned this statistical
structure in the sequence. In later studies, human participants were trained for the same task and were able to use their statistical knowledge after training in order to generate successive predictions about the next letter following their own previous predictions upon a presented cue (Perruchet \& Amorim, 1992). Returning back to the current study, it would be possible to assume that a statistical structure exists even in the event sequence (episode) within the real world. For example, we reasonably guess that the next event would be to blow the candle when a birthday cake is served to the dinner table. Also, we know that someone will cut cake into pieces before biting into a whole cake. Thus, there is some statistical structure in the sequence of events in real world, and our working hypothesis is that the order of successive cue-specific predictions in episodic future thinking should be to some extent constrained by this statistical structure in real world. Once we assume the similarity between the future prediction of the next letter in a given language (Elman, 1990) and the future prediction of the next event in real life, then it is natural to adopt Elman's approach for modelling episodic future thinking (see below in detail). As we admit above, episodic future thinking is a complex cognitive process, but this approach is promising to capture at least the two core characteristics of episodic future thinking mentioned above.

## Method

## Model Architecture, Tasks, and Representations

Figure 1 shows the architecture of the model. Four peripheral layers (input layer, output layer, semantic layer, and recognition layer) were connected bidirectionally through a single hidden layer. The hidden layer and each of the five output layers were connected to themselves. The input layer was sub-divided into five layers, each of which represented one of the five elements of the current event (Figure 1). For example, the first layer represented the context information of the current event. If this context layer was hard-clamped to the binary vector of [100000] , then it meant the current event occurred in Context 1 (e.g., school). The remaining four layers represented the Agent/Action/Object/Instrument of the current event. Thus, if the whole input layer was hard-clamped to the 18 -bit


Figure 1: The architecture of the model (Hinton diagram).

Table 1: Sequence structure of the training set.

| sequence | context |  |  | other information |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | contex <br> t <br> label | pattern | predictabilit y | pattern |  |  |  | predictability |  |
|  |  |  |  | agent | action | object | $\begin{gathered} \text { instrumen } \\ \mathrm{t} \\ \hline \end{gathered}$ | with context | without context |
| event (1) |  |  |  | 010 | 100 | 010 | 100 | 783 |  |
| event (2) | 1 | 100000 | constant | 100 | 010 | 001 | 100 | or |  |
| event (i) |  |  |  | 010 | 001 | 010 | 010 | 100\% |  |
| $\text { event }(i+1)$ | 6 | 000001 | constant | 001 | 010 | 001 | 010 | $33,50,$ | 6 ~ 45\% |
| event (j) |  |  |  | 100 | 100 | 100 | 010 | 100\% |  |
| event $(\mathrm{j}+1)$ |  |  |  | 001 | 001 | 001 | 001 |  |  |
| $\text { event }(\mathrm{j}+2)$ | 3 | 001000 | constant | 010 | 010 | 100 | 100 | 100\% |  |

vector of [(Context) 010000 (Agent) 100 (Action) 001 (Object) 100 (Instrument) 001 ], then the current event was 'In Context 2 (e.g., home), Agent 1 (e.g., John) did Action 3 (e.g., cut) to Object 1 (e.g., cake) with Instrument 3 (e.g., knife)'. The layers in the output side had the same structure, and when presented with the input pattern of the current event, the model was trained to activate the units in the output layer that consisted of the next event (the input 18 -bit vector of the next trial). The sequence structure will be explained later.
Next, the semantic layer consisted of 15 units whose activation patterns represented the 'conceptual knowledge' of the current event (interpretation of the event) in a distributed manner. Following many parallel-distributed processing (PDP) models that incorporated a 'conceptual knowledge' system in their models (Woollams et al., 2009), no attempt was made to design semantic representations that captured the actual meanings of the input pattern (e.g., input words, action, event, etc.). Instead, like past models, artificial semantic representations were created that, nonetheless, captured core characteristics of the meaning of an event. Specifically, we assumed that the meaning of an event would be to some extent related to the action, instrument, and object information of that event (e.g., not an arbitrary mapping). Once we hear these pieces of information, we can guess what happened in that event with some confidence. In contrast, the meaning of an event would be less strongly related to information on who (Agent) did that action. For example, the meaning of cutting an apple with a knife is invariant irrespective of who did that action. Next, the context information also constrains the meaning of the event. We know that certain kinds of events rarely occur in a certain context. For example, passing $a$ ball should not occur in a restaurant. Of course, Agent information would also constrain the meaning of an event (e.g., we might know that John would never eat an apple), but to a lesser extent than context/action/object/instrument
information. Taking these assumptions together, we created the target semantic representations such that the bit-patterns in the context/action/object/instrument input layers were systematically related to part of the target vectors in the semantic layer (i.e., mapping was not completely arbitrary). Then, when presented with the current event pattern in the input layer, the network was trained for generating the correct pattern in the semantic layer in addition to predicting the next event in the output layer. Irish et al. (2012) demonstrated that semantic dementia patients were less accurate than controls for 'knowing (semantic)' nonpersonal events over the past/future 10 years. Thus, we damaged this layer in simulation of the patients' behaviour.

A recognition trial was occasionally inserted during training, in which the network was trained for judging whether the presented event pattern had been experienced before or not. The single unit in the recognition layer served to represent the network's recognition judgment. Specifically, the input layer was hard-clamped to the value of an event representation, and then the network was trained to activate this recognition unit (1.0) if the presented event representation had appeared ('old') before, as part of the main task. In contrast, the recognition unit should be turned off (0.0) if the presented episode representation had never appeared before ('new').

## Sequence Structure of the Training Set

Sequence Structure of Context Information The sequence in the main trial was semi-random. Table 1 shows the structure of the sequence. First, as the left half of Table 1 shows, the context information (i.e., first 6-bit of the 18-bit input vector) was kept constant for several successive events in order to mimic the real world, where we experience successive events in the same context then move to another one. By presenting the first 6-bit information in this way, we can more safely argue that this 6-bit information represents
the context information of an event. Thus, the predictability of the next context information was $100 \%$ in most trials unless it was the boundary of a context-block. After several events, the context information changed into another context semi-randomly (33\%-50\% predictability).

Sequence Structure of Agent/Action/Object/Instrument Information The sequence of the remaining 12-bit information of an event was also semi random. There were 81 possible input patterns, formed by crossing 3 (Agent) by 3 (Action) by 3 (Object) by 3 (Instrument). When the context information was not considered, the predictability of the next event (i.e., next agent/action/object/instrument information) varied from $4 \%$ to $45 \%$ depending on a trial. When the context was considered together, the predictability increased such that it varied from $33 \%$ to $100 \%$ depending on a trial. We implemented the constraint from context information to mimic the real world. For example, it is more difficult to predict what will happen if we see a ball bouncing at a restaurant, but it is less difficult to predict at a park.

Recognition Trials After every nine trials for event prediction (and simultaneous computation of meaning), six trials were inserted to train the model for event recognition. The network received a 18-bit input pattern, and was required to judge whether or not this pattern had been presented before as part of the main task by activating/deactivating the recognition unit. In order not to bias the network's response, 'old' and 'new' trials were evenly distributed (3 trials, each) within each recognition block. The 'old' events were randomly sampled from the main training trials that the network had experienced during event prediction. The 'new' event-set was created in the following steps. First, when we had created the sequence of the main trials, we had ensured that not all the 81 possible input patterns (formed by combining agent, object, action, \& instrument) appeared in every one of the 6 possible contexts. Specifically, in each context, 20-27 possible combination of agent/object/action/instrument information had been randomly sampled and removed from the training set such that these patterns never appeared in that particular context during the main task. These pre-removed patterns served as 'new' events. To be clear, it was possible that these patterns appeared in another context. For example, the network might have received the 18 -bit vector of $[100000,100$, $100,100,100$ (comma denotes the boundary of layers)] but not received that of [010000,100,100,100,10 0 ]. Then, the network would have to activate the recognition unit when presented with the former pattern but would have to deactivate the same unit in the case of the latter. Thus, the network was trained for recognition of a particular event involving a particular context/agent/object/action/instrument. We also ensured that not all the possible 'old' trials and 'new' trials were presented during training, such that we were able to probe the generalization performance of the network to the untrained 'old'/'new' patterns.


Figure 2: Learning curves for event prediction, recognition of trained-items, and recognition of untrained-items.

## Training Parameters

In each trial, 18 units in the input layer were hard-clamped to their input values, and the network was allowed to cycle 10 times. In each time step, the activation spread to the next layer gradually being scaled by the values of the interconnecting weights, and the network settled into the steady state (called as an attractor). After 10 cycles of updates, the discrepancy between the output activation patterns (output event layer and semantic layer) generated by the network and the correct target pattern was calculated, and the connection strength was adjusted to reduce the discrepancy. In recognition trials, only the discrepancy in the recognition unit was considered. A learning rate of 0.01 was set at the beginning of the training. Then, every 10 epochs of training, the learning rate was gradually reduced by 0.001 . A decay parameter was set to 0.0000001 at the beginning and gradually reduced by 0.00000001 as the learning rate was reduced. When we evaluated the network's performances during/after training, we used a strict criterion such that the output was scored correct if the discrepancy was within 0.5 in every unit of the target layer after the $10^{\text {th }}$ cycle (i.e., the activation is less/more than 0.5 if the target is $0.0 / 1,0$. for each unit respectively).

## Results

## Trained Tasks

Figure 2 shows the learning curves for the event prediction task and the recognition task averaged across 10 independent simulations (initiated with different random seeds). The network successfully learned to predict the next event, thus acquiring the statistical structure which existed in the event sequence as well as recognizing the presented event pattern, which was generalized to untrained items. Accuracy for computing the meaning of an event quickly reached $100 \%$ after the training was initiated.

## Episodic Future Thinking

As explained in the introduction, the current model focused to capture at least the ability to compute cue-specific events successively following its own previous event prediction, a core characteristic of episodic future thinking. Thus, we first presented cues (e.g., Context = home, Agent = john, Action $=$ cut, Object $=$ cake, Instrument $=$ knife $)$. Then, once the network generated an output (i.e., prediction of next event), we presented this output vector pattern as the input of the next event, and the network generated the next output (prediction of the next event following its own prediction, see Botvinick \& Plaut, 2004, for the same approach in action learning). This cycle was reiterated 1000 times, and the generated 1000 -event sequence was regarded as an approximation of the network's episodic future thinking. As a result, the network successfully kept the presented context information (Context 1) constant for the first 829 events (average of 10 simulations), but lost this context information after this point.

## Simulation of Semantic Dementia

Following past simulations on semantic cognition, we simulated the episodic future thinking of patients with semantic dementia by removing some of the links between the semantic layer and the hidden layer (e.g., Woollams et al., 2009). Figure 3 shows how long (how many successive events) the network maintained the cued-context information as a function of disease severity (in terms of the proportion of links removed). This 'lesioning' simulation was reiterated 50 times with different links being sampled and removed, and the outcomes were averaged in order to avoid an idiosyncratic result. We found that, as the damage became more severe, the network was increasingly unable to maintain the event sequence of the cued-context (NB., The intact model kept the context for 829 events). Thus, future thinking deviated into another context/topic. Moreover, the proportion of the links removed was negatively correlated with the number of event predictions that maintained the cued-context $[r(17)=.-75, p<.01]$, suggesting that semantics had a causal role in generating a coherent episode in future thinking. Importantly, event recognition accuracy was intact (more than 95\% accurate) after this lesioning. All of these are consistent with the data from semantic dementia patients (Irish et al., 2012).

## Discussion

The current model successfully acquired the statistical structure within the training set, and used this knowledge to generate a context-coherent sequence of events triggered by cues (episodic future thinking). Moreover, when the computation of semantic knowledge was impaired, the model could not generate a context-coherent event sequence, yet preserved its recognition ability of event patterns. Importantly, the number of the events generated in a specific context was negatively correlated with the severity of damage, suggesting the causal role of semantics in episodic


Figure 3: Numbers of successive events in which the network maintained the cued-context information as a function of disease severity.
future thinking (Irish et al., 2012). This is consistent with the idea that the semantic system provides the framework of the event (D'Argembeau \& Mathy, 2011).
How does the semantic system affect the maintenance of context-coherent event sequences? This can be explained in terms of one of the general principles of PDP models. During training, a PDP network finds a unique attractor state (= unique abstract pattern in the hidden layer) associated with each of the input patterns. Once an input value is fed into the model, the activation gradually spreads, and the internal activity of the hidden layer gradually settles onto this unique status, as if it is falling into its unique attractor basin. They are unique, but similar inputs are associated with similar attractor basins. In the current model, the input patterns that share the same context information will fall into similar/neighbouring attractors, thus producing the same context output information to keep a context-coherent episode. However, if the internal representation of the model changes due to an impaired computation at some part of the model, then the network may settle into a wrong attractor basin, generating a wrong output. The diagnostic analysis suggests that this is certainly the case in our model. Specifically, we presented six events in different contexts to the network, and the activation pattern in the hidden layer on which the network settled was measured with/without semantics. Figure 4 shows the similarity structure of these patterns found by a multi-dimensional scaling analysis. With the intact semantic information (filled-markers), the network settles onto the context-specific attractor basins such that the network does not confuse one context with another. However, when the semantic system was damaged (openmarkers), the network's internal status drifted away from its correct attractor, thus generating a different/wrong context representation (e.g., The open-circle is closer to the filleddiamond rather than filled-circle). In other words, semantic representations contribute to "binding" a time-varying event


Figure 4: The similarity structure in the activation patterns of the hidden layer as a function of the input context information and of with/without semantics.
sequence such that it forms a context-coherent episode. One might describe this as a framework within which episodic details are integrated (D'Argembeau \& Mathy, 2011). Interestingly, Schapiro et al. (2013) has recently demonstrated that temporally-close stimuli that form one coherent event are similarly represented (in terms of voxelbased neural patterns) in the inferior/superior anterior temporal lobe and inferior frontal gyrus, both of which are the damaged areas in semantic dementia patients. Damage in this area might disrupt in computation of such similar neural patterns, and bound stimuli might fall apart.
Then, the question is why collapsed semantic knowledge has little effect on episodic recognition accuracy, as was demonstrate in this model as well as in patients with semantic dementia (Irish et al., 2008). This is because recognition of a particular event is both context-specific and agent/action/object/instrument-specific. In other words, it is crucial not to confuse a new event with an old one, even if part of the information contained in that new event is semantically familiar (e.g., you have ever used that instrument before and/or have seen the same action conducted by the same agent, yet in a different context). Therefore, it is possible that event recognition is not influenced by degradation of semantic knowledge (or at least not detected with a standard test).
Admittedly, the ability to generate context-specific event predictions could be simulated if the modules representing schemas or scripts were explicitly built-in by a modeller a priori. However, the model implemented symbolic system must have assumptions about schematic knowledge preliminarily (further discussions, Botvinick \& Plaut, 2004). The present sequential model did not have that symbolic system and developed by learning the statistical structure in the event sequence. This implies that learning sequential structure enables the model to compute schema-like representation (Botvinick \& Plaut, 2004), and can capture the behaviour of semantic dementia patients..
In summary, we have clarified the mechanism by which semantics contribute to episodic future thinking. The sequence prediction model (Elman, 1990) is a useful computational framework that can be extended to an event sequence triggered by a cue such that it successfully
captures the phenomenological and neuropsychological features of episodic future thinking. Certainly, this model does not capture the whole aspects of episodic future thinking, and in this sense, this is a proto-episodic future thinking model. In future work, implementation of essential factors for episodic future thinking is required such as the concepts of "self" or "temporal distance".

## Acknowledgments

This study was supported by Grant-in-Aid for Challenging Exploratory Research (23653224) to J. Kawaguchi, and for JSPS Fellows (24008998) to Y. Ito.

## Reference

Addis, D. R., Wong, A. T., \& Schacter, D. L. (2008). Agerelated changes in the episodic simulation of future events. Psychological science, 19, 33-41.
Atance, C. M., \& O'Neill, D. K. (2001). Episodic future thinking. Trends in cognitive sciences, 5, 533-539.
Botvinick, M., \& Plaut, D. C. (2004). Doing without schema hierarchies: A recurrent connectionist approach to normal and impaired routine sequential action. Psychological Review, 111, 395-429.
D'Argembeau, A., \& Mathy, A. (2011). Tracking the construction of episodic future thoughts. Journal of experimental psychology: General, 140, 258-271.
Elman, J. L. (1990). Finding Structure in Time. Cognitive Science, 14, 179-211.
Irish, M., Addis, D. R., Hodges, J. R., \& Piguet, O. (2012). Considering the role of semantic memory in episodic future thinking: evidence from semantic dementia. Brain, 135, 2178-2191.
Perruchet, P. \& Amorim, M. A. (1992). Conscious knowledge and changes in performance in sequence learning: Evidence against dissociation. Journal of Experimental Psychology: Learning, Memory, and Cognition, 18, 785-800.
Schacter, D. L., Addis, D. R., \& Buckner, R. L. (2007). Remembering the past to imagine the future: the prospective brain. Nature reviews Neuroscience, 8, 657661.

Schacter, D. L., Addis, D. R., \& Buckner, R. L. (2008). Episodic simulation of future events: concepts, data, and applications. Annals of the New York Academy of Sciences, 1124, 39-60.
Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., \& Botvinick, M. M. (2013) Neural representations of events arise from temporal community structure. Nature Neuroscience, 16, 486-492.
Szpunar, K. K., Watson, J. M., \& McDermott, K. B. (2007). Neural substrates of envisioning the future. Proceedings of the National Academy of Sciences of the United States of America, 104, 642-647.
Woollams, A. M., Joanisse, M., \& Patterson, K. (2009). Past-tense generation from form versus meaning: Behavioural data and simulation evidence. Journal of Memory and Language, 61, 55-76.

# The Impact of Problem Space on Reasoning: Solving versus Creating Matrices 

Saskia Jaarsveld (jaarsvel@rhrk.uni-kl.de)<br>University of Kaiserslautern, Center for Cognitive Science, Cognitive and Developmental Psychology, Edwin Schrödinger Strasse, Building 57<br>D-67663 Kaiserslautern, Germany<br>Thomas Lachmann (lachmann@rhrk.uni-kl.de)<br>University of Kaiserslautern, Center for Cognitive Science, Cognitive and Developmental Psychology, Edwin Schrödinger Strasse, Building 57<br>D-67663 Kaiserslautern, Germany

Cees van Leeuwen (cees.vanleeuwen@ppw.kuleuven.be)<br>Catholic University Leuven, Faculty of Psychology and Educational Sciences, Tiensestraat 102<br>B-3000 Leuven, Belgium


#### Abstract

Creative reasoning in ill-defined problem spaces operates differently from classical reasoning in well-defined spaces. To systematically compare the two in an identical knowledge domain, we applied a classical intelligence test: the Standard Progressive Matrices (SPM), in combination with two tests of creativity: the Test for Creative Thinking - Drawing Production (TCT-DP) and the newly developed Creative Reasoning Task (CRT), in which participants are asked to create an SPM-like item, to two age groups ( $N_{l}=511,4-12 \mathrm{y}$ old; $N_{2}=205,6-10 \mathrm{y}$ old). For SPM and CRT the knowledge domain consists of relationships amongst geometrical components in $3 \times 3$ matrices. We developed a typology for scoring the number and complexity of the relationships used in these matrices. For the SPM, we scored frequencies of relationships solved and for CRT those created, and interpreted the scores in terms of differences and similarities between classical and creative reasoning in cognitive development.


Keywords: Cognitive development; intelligence; reasoning; creativity; creative cognition; creative reasoning.

## Classical and Creative Reasoning

In creativity, both convergent and divergent thinking are needed, in order to arrive at a quality formulation (Jaarsveld \& van Leeuwen, 2005). Creative processes often consist of iteratively generating, testing, and selecting intermediate productions, ultimately leading to an integral result. We interpreted this process in terms of the integration of convergent and divergent operations characteristic of creative reasoning (Jaarsveld et al., 2010). Here we will consider the integration of convergent and divergent operations against the alternative possibility that both are used as independent, quasi-additive resources.

Convergent operations are typically associated with classical reasoning. A consequence is that classical and creative reasoning share processing components. Therefore,
if convergent and divergent operations constitute independent resources, test results between classical and creative reasoning will be correlated. Longstanding investigations of intelligence and creativity test scores suggested only a moderate relationship (Wallach \& Kogan, 1965; Kim, 2005; Silvia, 2008). The strength of the relation, however, may be a matter of differences between the knowledge domains of both tests that are unrelated to the differences between reasoning types per se. To illustrate this issue, here we compared a classical intelligence test, the Standard Progressive Matrices (SPM; Raven, 1938/1998), with two creativity tests, one of which, the Creative Reasoning Task (CRT; Jaarsveld et al., 2010; Jaarsveld, Lachmann, \& van Leeuwen, 2012), shares the domain of knowledge with the SPM and the other, the Test for Creative Thinking - Drawing Production (TCT-DP; Urban \& Jellen, 1995) does not (Jaarsveld et al., 2010; 2012).

## Knowledge Domain

In general, classical and creative reasoning tests tend to operate in different knowledge domains. For instance, the SPM, which is considered to measure convergent thinking, operates in the domain of relations among geometrical components contained in a matrix (Figure 1). By contrast, the TCT-DP, which is considered to measure divergent production, operates in the domain of figural associations. Smilansky (1984) introduced a paradigm which we named the Single Knowledge Domain Paradigm. Smilansky asked participants first to solve the SPM and next to create an SPM-like item in a task which we named the Creative Reasoning Task (CRT). Hence, between SPM and CRT cognition operates on the same knowledge domain (Figure 1). Solving a classical reasoning task does not always mean the problem is understood: often a correct solution is accompanied by an incorrect line of verbal reasoning or is obtained without any conceptual understanding (Chi \& VanLehn, 1991; Karmiloff-Smith, 1992; Pine \& Messer, 1999). Such distortions are less likely with ill-defined problems.


Figure 1: An item of the solving test (SPM) and the empty response form of the Creative Reasoning Task (CRT).

## Problem Space

In the literature, classical and creative reasoning are both understood as processes operating in abstract problem spaces (Hayes \& Flowers, 1986; Simon, 1973; Kulkarni \& Simon, 1988; Runco, 2007). A problem space contains all possible states that are accessible from the initial state through iterative application of transition rules, including the ones that bring the problem solver from the initial state to the final solution. Problem spaces in classical reasoning are well-defined; like in a game of chess, no reinterpretation of rules is possible. Problem spaces in creative reasoning are ill-defined, and may allow re-interpretation of rules during the problem solving process. For instance, in rearranging your room you uncover implicit requirements that introduce a set of new transformations and/or eliminate existing ones (Barsalou, 1992) or, when conflicting constraints arise, you introduce new trade-offs (Yamamoto, Nakakoji, \& Takada, 2000).

In our first study we compared reasoning performances in well and ill defined problem spaces, those of the SPM and CRT, respectively (Jaarsveld et al., 2010). For analyzing the performance on both tasks, we developed a scoring method based on a typology of the number and complexity of the relationships in evidence in the $3 \times 3$ matrices that feature in these tests. In a second study (Jaarsveld et al., 2012) we developed for the CRT two sub-scores: Relations, which reflects convergent production in ill defined problem spaces, and Components and Specifications, which reflects divergent production. We compared across grade levels, firstly, the CRT sub-scores of Relations with scores of the SPM and the CRT sub-scores of Components and Specifications with scores on the TCT-DP. Secondly, we analyzed the complexity in matrices solved in the SPM with created in the CRT. This analysis would allow us to observe whether more advanced pupils have a higher developed ability to process complex information (Halford, 1993).

## Method

Participants Children of the first study were from Nursery and Elementary Schools ranging from four to twelve years old ( $N_{l}=511$ ), $52 \%$ girls Mean age per grade in years: Younger Nursery school children ( $M=4.64, N=33$ ), Older Nursery school children ( $M=5.68, N=31$ ), Elementary school Grade 1 ( $M=6.73, N=41$ ), Grade $2(M=7.79, N=$
42), Grade 3 ( $M=8.81, N=59$ ), Grade $4(M=9.80, N=$ 132), Grade $5(M=10.87, N=91)$, Grade $6(M=11.91, N=$ 82). In the second study we only had children from Elementary School ranging from six to ten years old ( $N_{2}=$ 205), $50 \%$ girls. Mean age per grade in years: Grade 1 ( $M=$ $7.06, N=51$ ), Grade $2(M=8.16, N=43)$, Grade 3 ( $M=$ 9.07, $N=51$ ), Grade 4 ( $M=10.05, N=60$ ). Age limits within grades for both studies were not absolute, but the average age increased with 1 year per grade.

Material The SPM is contained in a booklet, which displays one incomplete matrix per page, together with a multiple choice of completion alternatives. Participants had to infer relations between given components and choose the completing figure from among the alternatives given below the matrix (Figure 1). A separate answering sheet is offered, on which individuals mark the number of the alternative they consider to be the proper completion. The CRT asks participants to create an SPM-like item. The instruction was to make the item as difficult as you possibly can such that it will be a hard puzzle for others to solve. On an empty form reflecting the format of the SPM items (Figure 1) participants had to create components and relations, and to draw the completing figure in one of the cells in the lower part of the response form. The TCT-DP asks participants to complete a drawing on a form containing five simple components within a frame and a sixth one outside the frame. The instruction conveyed that one could do nothing wrong and draw as one liked.

Design and procedure Children first performed the solving test (50 minutes). Nursery School children and those up to Grade 3 performed the Coloured Progressive Matrices test (CPM; Raven, 1956/1976) which is designed to assess the cognitive abilities of young children. Older children performed on the SPM. Consecutively, in both studies all were asked to create a matrix in the CRT ( 15 minutes). Finally, the children of the $2^{\text {nd }}$ Study performed the Test of Creative Thinking (10 min). Nursery school children performed the tasks individually; group testing was applied for the classes of the Elementary School.

Analysis The scores of the CPM and SPM equaled the number of items solved correctly. Scores of the CPM were converted to SPM scores according to the scale provided by Raven, Raven, and Court (1998), in order to enable direct comparisons between grade levels. The score of the TCTDP was a summation of grade points (range 0-6) for each of the 14 sub-scores: Continuation, Completion, Connections Made with a Line, Connections Made to Produce a Theme, Figure-based Boundary Transgression, Figure-independent Boundary Transgression, Perspective, Humor and Affectivity, Unconventionality-a: any manipulation of the material; Unconventionality- $b$ : any surrealistic, fictitious and/or abstract elements or drawings; Unconventionality-c: any usage of symbols or signs; Unconventionality-d: unconventional use of given fragments; Speed: drawings
that are made within a certain time limit and show a score above a certain value score extra points. The score of the CRT consisted of the sub-scores Relations and Components and Specifications. Relations scores the logical complexity of relations in complete and incomplete matrices. Relations are typically transformations from one component of the matrix to another. We identified a total of twelve relations. Three for the CPM: Four Identical Components; Continuous Pattern; and Symmetry. One for the SPM: Indication of Mathematical Operation. Three for the CRT: Idiosyncratic and Semantic Coherence; Indication of Form, Texture, Amount or Orientation; and Groups of Three Components. An additional five for the SPM were taken and partly modified from Carpenter, Just, and Shell (1990): Change; Increase and Decrease; Combination; Succession; and Disappear and Remain. We analyzed the matrices of CPM, SPM, and CRT for the relations they contained. Scoring of a relation created in the CRT is done in several steps. First, we listed the relations that apply to the item and for each relation marked the components it covered. Next, for each relation we assigned an index value $i=1,2, \ldots$ to all first appearances of the marked components, starting from the top-left cell of the item, proceeding from left to right through each row from top to bottom. Third, passing through the matrix in the same order as previous, we accumulate a score, in which the first encounter of a component is scored with a value identical to its index; each next time we encounter a component again, we assign the same score as previously, incremented with 1 when it occurs in a row different from where it has previously been encountered, and with another 1 when it occurs in a column in which it has not previously been encountered. The resulting score is the sum total of all values assigned to components of the matrix. The sub-score Components scores the number of different components and the subscore Specifications scores the occurrence of different pictorial specifications (textures and line styles) and transformational specifications (size, orientation, number, and location). These specifications were scored when they did not express a relation. The categories Non Figurative and Figurative indicated matrices which featured components of a geometrical and a figurative character, respectively.

## Results

First, we present results of interrater reliability of the CRT for type and complexity of relations. Next, we present the correlations between test scores reflecting three types of cognitive processes; one which mainly features convergent thinking (CPM and SPM), one which mainly features divergent thinking (TCT-DP), and one in which both types of thinking play a role (CRT). Thirdly, we present to what extent Relations in the items created in the CRT reflected those featured and solved in the CPM and SPM. Finally, we present results of the increasing complexity of matrices created, according to number of relations applied within the matrix.

Interrater reliability Subsets of items of the CRT were scored independently by different raters and interrater reliability was calculated with Cohen's Kappa, $K$, for type of relations and with Pearson correlations, $r$, for the CRT sub-scores. Results ranged from $K=.93$ in the first study ( $n$ = 95), to $K=.94$ in the second study ( $n=69$ ), and from $r=$ .99, $p<.01$ for sub-score Relations, to $r=.91, p<.01$ (both 2-tailed) for sub-score Components and Specifications.

Test Scores Results of test scores between SPM and CRT over all grade levels did not show a significant correlation. They did show a correlation in some grade levels; in the first study in Grades 3 and 6, and in the second study in Grade 1. In the second study, as expected, the CRT sub-score Relations, which according to our theory represents convergent thinking, showed a correlation with the SPM ( $r$ $=.192, p<.01$; partial correlation corrected for TCT-DP: $r$ $=.213, p<.01$ ). The CRT sub-score Components and Specifications, which according to our theory represents divergent thinking, showed a correlation with the TCT-DP ( $r=.147, p<.05$; partial correlation corrected for SPM: .153, $p<.05$ ). Furthermore, scores of SPM and TCT-DP showed a correlation, $r=.225, p<.05$, but, as expected from the assumption that in the CRT the sub-scores represent different thinking abilities, no correlation was found between the CRT sub-scores Relations and Components and Specifications, $r=.016, p=.823$.
Moreover, as expected, there were no correlations between Relations and TCT-DP, and between Components and Specifications and SPM. The latter results hold also for partial correlation analyses. From this we may infer that convergent and divergent thinking play a role in the Creative Reasoning Task and that both can be scored on one end product.

Relations Featured, Solved, and Created in the CPM Condition Frequencies of relations solved and created showed that Young Nursery School children (age in years $M$ $=4.64)$ solved three of the four relations presented but generated a different relation, Relation 1, Idiosyncratic Coherence. Older Nursery School children (age in years $M$ $=5.68$ ) preferred an additional relation, Relation 3, Continuous Pattern. This focus shifts at Grade 1 (age in years $M=6.73$ ) to Relation 2, Four Identical Components and at Grade 2 and 3 (age in years $M=7.79$ and $M=8.81$, respectively) to Continuous Pattern. The dominance of Idiosyncratic Coherence, shows that creative reasoning in the youngest children is dominated by rules that are not deducible logically and clearly arise from an individual interpretation.

Relations Featured, Solved, and Created in the SPM Condition Relation 3, Continuous Pattern is the most frequently created relation by children of the higher grades (Grade 4, age in years $M=9.80$ to Grade $6, M=11.91$ ). Deleting a piece in an overall pattern, whether figurative or non figurative, may be the first abstract relation that plays a
role in generative problem solving. The results of the chisquare tests for independence over the frequencies of relations solved and created in both conditions were significant, p-values smaller than .05 to .001 . In the second study we observed identical results.

Components Results of chi-square tests of Components and Specifications in CPM and SPM condition followed those for Relations. Figurative components were generated by children in both studies although the solving test does not feature these types of components. In the first study the percentage of children who applied figurative components decreased significantly with grade level, $r_{s}=-.671, n=8, p$ <.05, one tailed ( $r_{s}$ Spearman Rank correlation).

Relations in the SPM and created in the CRT as a function of school grade Second study: we observed that only one relation, Combination, showed an increase with grade in both SPM and CRT. Other relations showed either, a decrease in SPM in combination with an increase in CRT, for instance Pattern Completion; an increase in SPM was observed for the relations Change and Succession; a decrease in SPM was observed for the relation Increase and Decrease; a decrease in CRT was observed for the relations Idiosyncratic Coherence, Four Identical Components, and Symmetry. (Spearman Rank correlations of frequencies over grade, $p<.05$ ). For both studies we concluded that SPM and CRT did not show the same trends with grade in the frequency of occurrence of different relations.

Number of Relations Applied per Item Complexity in the SPM matrices as measured by number of relations increased over the series of SPM items, $r_{s}=.900, n=5, p<.05$. We found a corresponding increase in complexity over grade levels in matrices created in the CRT in the first study, $r_{s}=$ $.964, n=8, p<.01$. Components increasingly show variety in number, in size and orientation. In the second study increases in complexity failed to reach significance due to lack of power, $r_{s}=.258, n=4, p=.371$.

## Conclusions and Discussion

We compared across grade levels the performance on the Creative Reasoning Task (CRT), with that on the Progressive Matrices test (CPM and SPM), and the Test of Creative Thinking-Drawing Production (TCT-DP). We used the CRT to measure convergent and divergent thinking, which we consider to play an integrated role in ill defined problem spaces. CRT and SPM operated on the same problem domain; nevertheless, operations used in both tasks differed as a function of the differences in problem spaces. Whereas the SPM uses convergent operations, both divergent and convergent operations are needed for the CRT. The absence of correlations across school grades, therefore, implies that in creative processes as tested by the CRT, convergent and divergent operations do not occur as additive process components, but play an integrated role throughout the process (Jaarsveld \& van Leeuwen, 2005.

Correspondence In addition to contrasts between the tests, similarities were observed in development. Across school grades, we observed increasing complexity in problem solving and problem creation. In the SPM we observed an increase over series of items combining several rules. In the CRT there is a parallel increase in the number of relations applied per item created. Children in more advanced grades also used more components, with an increasingly rich variety of specifications.

Differences Although relations applied in the creating task often featured in the solving task, almost within all grades performance on both tasks was uncorrelated; in the CRT grades were characterized by a preference for specific types of relations. Another difference between both tasks was the absence of concordant increases or decreases over age levels in the application of certain types of relations. Combination was the only one of 12 relations that showed an increase in both tasks. Finally, in creating, figurative components were more persistently preferred, despite the non figurative character of CPM and SPM items. Participants preferred to introduce rules and other elements from their individual episodic/semantic knowledge domains, as opposed to what they encountered in the problem solving task. This difference cannot be understood as a discrepancy in knowledge domain. Creative problem solving, therefore, does not depend entirely upon classical problem solving skills.

Cognitive Development Perspectives Even though the material of the SPM is non figurative, relations created in the CRT tend to be expressed in figurative mode. SingerFreeman and Goswami (2001) observed that three to four year old children understand proportional equivalence, even when the materials (pizza and chocolates) to be matched are not isomorphic. Young children, therefore, do not solve analogy problems on the basis of relational similarities but on the basis of associations (Piaget, Montangero, \& Billeter, 1977). Young children in solving CPM items have the opportunity to learn that matrix components belong together according to certain relationships. They proceeded in the CRT to arrange components according to different, selfdefined relationships.
Whereas children were able to solve most relationships, per grade one type of relation was predominantly applied in the CRT. Zelazo, Frye, and Rapus (1996) observed that knowing a rule in the card sorting task does not imply that it will be used correctly after a new sorting rule has been introduced. These authors observed a change in the ability to switch to a new rule between the age of three and five years old and explained this among others, in terms of the implicitness of rule representation. It could be that representations formed in a well-defined problem space are not understood at a sufficiently explicit level to be carried over to an ill-defined space.
The observed shifts with grade level in rules preferably applied in the CRT seem to correspond to Piaget's
developmental stages. According to Piaget, children between the ages of four and seven years old are in the intuitive thought phase, which is a sub-phase of the preoperational phase. In this phase children develop Conservation (the awareness that altering the state of a substance, does not change it's properties) and Centration (the focusing on one characteristic). In our study these children applied relationships of the types: Idiosyncratic Coherence and Four Identical Components. These are relationships that mostly feature one characteristic.
Piaget considered children between seven and eleven years old to be in the concrete operational phase. In this phase children's ability to think abstractly develops and they learn to understand the concept of reversibility. In our study these children predominantly applied variations of the relationship Continuous Pattern, a relationship which does not require abstract thinking. Deleting a piece from an overall pattern, as application of this rule requires, might be the first abstract operation in this phase.
Although creative productions, therefore, seem to follow Piagetian stages, two observations need to be made: First, despite these overall restrictions, children applied more complex relationships in small frequencies, in the CRT. Second, the contrast between rules used in solving SPM items and those applied in the CRT is not a matter of decalage. Piagetian stages are in evidence in the CRT, but not in the SPM. They are not reflected in classical problem solving but in creative reasoning, which characteristically requires the integration of divergent and convergent reasoning.

Limitations of the CRT The CRT is still in an early stage of development. In its current form, there are several issues that restrict its practical utility. Before the CRT can be administered the SPM has to be completed. This task serves, amongst other things, to make participants acquainted with the particular structure of the matrices problems. Without this phase we would have needed extensive instructions, which renders the task more algorithmic and, therefore, might decrease creative production (Amabile, 1987). To have a solving task precede a generation task is consistent with the general observation that nothing is invented from scratch; creativity implies using old elements in new contexts and seeing relations that no one saw before (Barron, 1981; Boden, 1990; Indurkhya, 1992; Torrance, 1987). Familiarity with the relevant domain and experience with a variety of methods are a prerequisite for generating solutions (Voss \& Post, 1988). For the current study, SPM data were needed anyway. If one is interested only in CRT performance, however, future developments of the test should include a certain number of matrices, specifically constructed to contain the same relations that feature in the SPM. These new matrices, then, are expected to provide participants with an identical solving experience as in the SPM. Moreover, the current CRT asks participants to generate one item only. This was done in order to tap individual abilities at the moment where they had reached
the maximum level of apprehension according to SPM. However, children may not achieve to their maximum abilities in this single item. For this reason, we are currently investigating the effect of including multiple CRT items in the task.

With the Creative Reasoning Task, we were able to answer the question whether individuals who have just solved SPM items in which certain transformations featured, apply these same transformations when they design a new matrix. We found that relations featuring in the solving task differed from those applied in the problem creating task. It was concluded that creative reasoning, as measured by the CRT, does not reflect SPM solving ability, and that both cognitive abilities develop rather independently from each other from Kindergarten to Secondary school.

## References

Amabile, T. M. (1987). The motivation to be creative. In S. G. Isaksen (Ed.), Frontiers of creativity research: Beyond the basics (pp. 223-254). Buffalo, NY: Bearly Limited.
Barron, F. (1981). Originality in relation to personality and intellect. In W. E. Barbe \& J. S. Renzulli (Eds.), Psychology and education of the gifted (pp. 168-178). New York, NY: Irvington Publications.
Barsalou, L.W. (1992). Cognitive psychology: An overview for cognitive scientists. Hillsdale, New Jersey: LEA.
Boden, M. A. (1990). The creative mind: Myths and mechanisms. London: Abacus.
Carpenter, P. A., Just, M. A., \& Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices test. Psychological Review, 97, 404-431.
Chi, M. T. H., \& VanLehn, K. A. (1991). The content of physics self-explanations. Journal of the Learning Sciences, 1, 69-105.
Halford, G. S. (1993). Children's understanding: The development of mental models. Hillsdale, NJ: Lawrence Erlbaum Associates.
Hayes, J. R., \& Flowers, L. S. (1986). Writing research and the writer. American Psychologist, 41, 1106-1113.
Indurkhya, B. (1992). Metaphor and cognition. London: Kluwer Academic Press.
Jaarsveld, S., \& van Leeuwen, C. (2005). Sketches from a design process: Creative cognition inferred from intermediate products. Cognitive Science, 29, 79-101.
Jaarsveld, S., Lachmann, T., Hamel, R., \& van Leeuwen, C. (2010). Solving and creating Raven Progressive Matrices: Reasoning in well and ill defined problem spaces. Creativity Research Journal, 22, 304-319.
Jaarsveld, S., Lachmann, T., \& van Leeuwen, C. (2012). Creative reasoning across developmental levels: Convergence and divergence in problem creation. Intelligence, 40, 172-188.
Karmiloff-Smith, A. (1992). Beyond modularity: A developmental perspective on cognitive science. Cambridge, MA: MIT Press.

Kim, K. H. (2005) Can only intelligent people be creative? Journal of Secondary Gifted Education, 16, 57-66.
Kulkarni, D., \& Simon, H. A. (1988). The processes of scientific discovery: The strategy of experimentation. Cognitive Science, 12, 139-175.
Piaget, J., Montangero, J., \& Billeter, J. (1977). La formation des correlats. In J. Piaget (Ed.), Recherches sur l'abstraction reflechissante I (pp. 115 - 129). Paris: Presse Universitaires de France.
Pine, K., \& Messer, D. (1999). What children do and what children know: Looking beyond success using KarmiloffSmith's RR frame-work. New Ideas in Psychology, 17, 17-30.
Raven, J. C. (1938/1998). Standard Progressive Matrices, Sets $A, B, C, D \& E$. Oxford, Great Britain: Oxford Psychologists Press.
Raven, J. C. (1956/1976). Coloured Progressive Matrices, Sets $A, A B, \& C$. Oxford, Great Britain: Information Press.
Raven, J., Raven, J. C., \& Court, J. H. (1998). Raven manual: Section 3, Standard Progressive Matrices. Oxford, Great Britain: Oxford Psychologists Press.
Runco, M. A. (2007). Creativity, theories and themes: Research, development, and practice. Amsterdam, New York: Elsevier.
Singer-Freeman, K. E., \& Goswami, U. (2001). Does half a pizza equal half a box of chocolates? Proportional matching in an analogy task. Cognitive Development, 16, 811-829.
Silvia, P. J. (2008). Creativity and intelligence revisited: A reanalysis of Wallach and Kogan (1965). Creativity Research Journal, 20, 34-39.
Simon, H. A. (1973). The structure of ill structured problems. Artificial Intelligence, 4, 1012-1021.
Smilansky, J. (1984). Problem solving and the quality of invention: An empirical investigation. Journal of Educational Psychology, 76, 377-386.
Torrance, E. P. (1987). Teaching for creativity. In S. G. Isaksen (Ed.), Frontiers of creativity research: Beyond the basics (pp. 189-215). Buffalo, NY: Bearly.
Urban, K. K., \& Jellen, H. G. (1995). Test of Creative Thinking - Drawing Production. Frankfurt, Germany: Swets Test Services.
Voss, J. F., \& Post, T. A. (1988). On the solving of illstructured problems. In M. T. Chi, R. Glaser, \& M. J. Farr (Eds.), The nature of expertise (pp. 261-286). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
Wallach, M. A., \& Kogan, N. (1965). Modes of thinking in young children: A study of the creativity-intelligence distinction. New York: Holt, Rinehart, \& Winston.
Yamamoto, Y., Nakakoji, K., \& Takada, S. (2000). Hand on representations in two dimensional spaces for early stages of design. Knowledge Based Systems, 13, 357-384.
Zelazo, D. P., Frye, D., \& Rapus, T. (1996). An age-related dissociation between knowing rules and using them. Cognitive Development, 11, 37-63.

## Acknowledgments

We would like to thank Marisete Welter (University of Kaiserslautern, Germany) for providing the scores for the inter rater reliability calculations, and Frits Peters who wrote the software to compute the frequency profiles and the software for the permutation method. Many thanks are also due to the directors, teachers and pupils of the schools who supported our studies.

# Visual Recognition using a Combination of Shape and Color Features 

Sepehr Jalali (tmssj@nus.edu.sg)<br>Cheston Tan (cheston-tan@i2r.a-star.edu.sg)<br>Joo-Hwee Lim (joohwee@i2r.a-star.edu.sg)<br>Jo-Yew Tham (jytham@i2r.a-star.edu.sg)<br>Sim-Heng Ong (eleongsh@nus.edu.sg)<br>Paul James Seekings (mmrl@nus.edu.sg)<br>Elizabeth A. Taylor (tmsohe@ nus.edu.sg)

National University of Singapore, Singapore 119077
Institute for Infocomm Research, A*STAR,Singapore 138632


#### Abstract

We develop and implement a new approach to utilizing color information for object and scene recognition that is inspired by the characteristics of color- and object-selective neurons in the high-level inferotemporal cortex of the primate visual system. In our hierarchical model, we introduce a new dictionary of features representing visual information as quantized color blobs that preserve coarse, relative spatial information. We run this model on several datasets such as Caltech101, Outdoor Scenes and Underwater Images. The combination of our color features with (grayscale) shape features leads to significant increases in performance over shape or color features alone. Using our model, performance is significantly higher than using color naively, i.e. concatenating the channels of various color spaces. This indicates that usage of color information per se is not enough to produce good performance, and that it is specifically our biologically-inspired approach to color that results in significant improvement.


Keywords: Visual recognition; Color; HMAX; Biologically inspired; Visual cortex; Image classification

## Introduction

Many models are inspired by the hierarchical organization of the visual cortex, such as Fukushima (1980) and Riesenhuber and Poggio (1999). Most of these models focus on grayscale information and ignore color information. While the broad use of color information in the primate visual system is well-known, the details are still under active investigation (Conway et al., 2010). Nonetheless, in this paper, we attempt to utilize what is currently known about the use of color to enhance object and scene recognition by computer algorithms. In this paper we utilize the HMAX model (Riesenhuber \& Poggio, 1999), but this approach can be extended to other computational models.

In our experiments, we use the HMAX model (Riesenhuber \& Poggio, 1999) in concatenation with our color model in order to evaluate the use of both shape and color. HMAX is a biologically-inspired model which focuses on the shape processing capabilities of the ventral visual pathway, and has been used to perform classification tasks (Serre, Wolf, Bileschi, Riesenhuber, \& Poggio, 2007).

We focus on extending the model by modelling the highlevel usage of color by incorporating insights from cognitive psychology and neuroscience. The broad intuitive inspiration for our model follows from the fact that colors are recognized categorically just as object classes are, even though color
discrimination and matching is continuous (Palmer, 1999). Interestingly, people of different races (Boynton \& Olson, 1987), as well as chimpanzees (Matuzawa, 1985), organize colors into the same basic color categories, such as red, blue, yellow, green.

More importantly for object and scene recognition, the categorical recognition of color suggests that, if color information is incorporated into object and scene classification, then fine-grained color information (e.g. precisely specified hue) may not be necessary. For example, a beach scene might be recognized from the blue (sky and sea) and brown (sand) regions. It may not be important exactly how blue the sky/sea or how brown the sand grains are. In fact, it may be important to disregard such details in order to perform classification that is tolerant to variations in lighting, and so on.

In addition, the coarse relative spatial position of such color regions may be important. A blue region above a yellowbrown region might suggest a beach scene. If the relative positions are reversed, then the image is probably not a beach scene (or might be an upside-down one). Not only is the detailed spatial information unnecessary, it may be crucial to discard it and only retain coarse spatial information, since the exact spatial relations will depend on factors such as the precise shape of the beach and the camera angle.

Overall, our model can be loosely described as performing object and scene classification by reducing a given image to a "coarse arrangement of categorical color blobs", similar to the idea of spatial aggregation of visual keywords (Lim, 1999), but with realization on the HMAX model. This is different from approaches that utilize color information in a low-level fashion, although the two types of approaches are not mutually exclusive. Crucially, our biologically-inspired approach outperforms the naive use of color, where an image is decomposed into separate color channels that are processed independently until the final classifier stage.

## Related Work

First, we go beyond the intuitive motivation for our approach and review the biological evidence that the primate visual system utilizes color information in a manner that is broadly consistent with our model. Specifically, we review studies of color processing in the high-level visual area of the primate
brain known as infero-temporal cortex (IT), which is commonly associated with invariant object recognition.

In the broadest terms, IT is known to play an important role in color discrimination. A majority of IT neurons are color-selective (Desimone, Schein, Moran, \& Ungerleider, 1985) and two independent studies estimated this proportion to be roughly $70 \%$ (Komatsu, Ideura, Kaji, \& Yamane, 1992; Edwards, Xiao, Keysers, Földiák, \& Perrett, 2003). Contrary to the theory that color processing occurs after more rapid luminance-only processing, no evidence was found that colored images evoke responses that are delayed relative to achromatic images (Edwards et al., 2003). More direct evidence for the role of IT comes from findings that color discrimination is severely disrupted by lesions (Heywood, Shields, \& Cowey, 1988) or cooling (Horel, 1994).

Color-selective neurons in IT are found in clusters, suggesting that they may form a segregated and independent processing network (Conway, Moeller, \& Tsao, 2007). As further evidence of this, one color cluster in IT received projections from a color cluster from another part of IT, suggesting that these clusters of color-processing neurons form reciprocally-connected modules within a distributed network (Banno, Ichinohe, Rockland, \& Komatsu, 2011).

IT neurons are selective for both hue and saturation (Komatsu, 1993). Different cells have different preferred hues, and as a population, the cells' preferred color spans most of the color space (Conway et al., 2007). The colors for which IT neurons are selective for tend to correspond to the basic color names (Komatsu, 1998). Komatsu (1998) proposed that IT has templates corresponding to color categories and may be involved in determining color category by finding the best match over these categories. More recently, the distribution of color-selective neurons found in IT seems to correspond to the three to four most basic colors (Stoughton \& Conway, 2008). The largest peaks align with red, green, and blue, in order of size of peak, with a smaller peak corresponding to yellow. These peaks roughly correspond to colors perceived by humans. Prior to this, neural representation of such unique hues (Hurvich, 1981) had not been found (Valberg, 2001). Note that in the low-level primary visual cortex, the axes defined by cone opponency should more accurately be denoted bluish-red/cyan and lavender/lime opponency (Stoughton \& Conway, 2008), rather than red-green and blue-yellow opponency.

Finally, the region of IT where color-selective neurons are found is coarsely retinotopic (Yasuda, Banno, \& Komatsu, 2010), meaning that spatial information is maintained in a coarse manner, rather than completely discarded or maintained with high fidelity. Overall, these studies are broadly consistent with our proposed "coarse arrangement of categorical color blobs" model of high-level color processing in the primate visual system.

In contrast, most computer vision algorithms utilize color information in a relatively low-level manner. The simplest color extension of a non-color algorithm would be to ap-
ply it independently to the $\mathrm{R}, \mathrm{G}$ and B channels, and then concatenate the features from all 3 channels just before the final classifier stage. Most algorithms are variants of this basic idea, either using some other color space, or fusing the channels before the classifier stage (usually at the dictionary or keyword learning stage). For example, SIFT features can be computed separately for each channel in HSV color space (Bosch, Zisserman, \& Muñoz, 2008), while Brown and Susstrunk (2011) do this for RGB space, along with an NIR (near infra-red) channel. Besides SIFT features, other algorithms use (non-orientation based) histograms in the HSV (Tang, Miller, Singh, \& Abbeel, 2012), Gaussian opponent color (Burghouts \& Geusebroek, 2009), normalized RGB or opponent color spaces (Gevers \& Stokman, 2004). What these algorithms have in common is that in terms of the biology of color vision, they correspond to at most the level of color-opponent cells in the primary visual cortex, the lowest level in the hierarchically-organized visual cortex.

## CQ-HMAX

In this section, we describe our new biologically-inspired model, CQ-HMAX (Color Quantization Hierarchical Max), which uses color information in a hierarchical organization of simple and complex cells. HMAX is a hierarchical model which uses Gabor filters to find simple and complex shapes in the images. Our model has a similar hierarchical structure. However, we use color quantization cores and not Gabor filters, hence our model encodes color information. When combined with HMAX, the overall model includes both color and shape information.

Our color model has a hierarchical structure of simple and complex cells, as can be seen in Fig. 1. We first introduce the model briefly, followed by a more detailed description of each layer. An image pyramid is created in YIQ color space. The $Y$ channel represents luminance information, while the $I$ and $Q$ channels represent chrominance information. The pyramid has 10 scales, with each neighboring scale different by a ratio of $1 /\left(2^{1 / 4}\right)$. In order to evaluate the use of color information in our model, we determined that the YIQ color space produced the best results in comparison with HSV and RGB color spaces. A set of representative values from each color channel is selected as color cores and used to find the best matching unit to each individual pixel value in the pyramid. The $S 1$ layer is created on 10 scales indicating the index of the best matching YIQ core to each pixel in the image pyramids. At the $C 1$ layer, a local max pooling is computed over $\pm 10 \%$ spatial neighborhoods of approximately $6 \times 6$ on $\pm 1$ neighbor scales to find the most frequent color core in each neighborhood. A dictionary of features is sampled randomly from the $C 1$ layer of images. The distance of each dictionary feature to all patches in a neighborhood of that dictionary feature is calculated to create the $S 2$ layer and the best response to each dictionary feature in each image is chosen as the $C 2$ layer to be fed to the SVM layer for classification. We describe each layer in more detail below.


Figure 1: The CQ-HMAX model and the processing of an example beach image.

## $S 1$ Layer and Quantization Cores

The input images are first converted into YIQ color space and a pyramid of 10 scales with a ratio of $2^{1 / 4}$ is created, with the first scale having the shorter side set to 140 pixels, maintaining the aspect ratio of the original image. This image pyramid is then used as the input to the $S 1$ layer. A series of YIQ quantized "color cores" over YIQ channels are created to be used as filters for this layer. We experimented with different numbers of quantization values per color channel, and chose 5 per channel as the optimal number (which results in $5 \times 5 \times 5=125$ cores). In order to choose the optimal cores, 500 images were randomly selected and the color range of these images in YIQ color space was calculated after normalization to the range $[0,1]$. The values of YIQ channel are mostly in the range $[0,1],[0.4,0.7]$ and $[0.4,0.6]$ respectively. These ranges were selected and divided into 5 bins. The quantized values of $Y, I$ and $Q$ after normalization to $[0,1]$ were therefore chosen as follows: $Y=(0,0.25,0.5,0.75,1)$,
$I=(0.4,0.47,0.55,0.63,0.7), Q=(0.4,0.47,0.5,0.53,0.6)$. Using these values results in better classification performance than using the full range $[0,1]$ in each YIQ channel. The outputs at the $S 1$ layer are the index values (i.e. $1,2, \ldots, 125$ ) of the best-matching color core for each element in the image pyramid.

## C1 Layer

The $C 1$ layer provides local invariance to position and scale as it pools nearby $S 1$ units, and as a result, subsamples $S 1$ to reduce the number of units. The $S 1$ pyramid is convolved with a $3 D$ max filter to set the $C 1$ layer size of the bottom of the pyramid to $25 \times 25$ and the highest layer of the pyramid to $5 \times 5$ accordingly. The max is calculated over $\pm 10 \%$ spatial neighborhood on $\pm 1$ neighbor scales in the middle of the pyramid and -2 on the highest level and +2 on the lowest layer of the pyramid (hence it is called a $3 D$ max, as it takes the max over $2 D$ spatial distribution and over $\pm 1$ scale).

This layer provides a model for $V 1$ complex cells. Fig. 1 also shows an example image of $S 1$ and $C 1$ layer. $S 1$ and $C 1$ layers have a distribution of quantization cores from coarse to fine. The higher layers of the $S 1$ pyramid are taken from smaller scales of the images in the input pyramid and respectively the higher levels of $C 1$ layer are computed by taking a $3 D$ max over higher levels of $S 1$ layer. As can be seen in Fig. 1, the higher levels of the pyramid in the $S 1$ and $C 1$ layers represent less detailed information from the image. All levels in the $C 1$ intermediate layer are used for sampling a dictionary of features.

## Dictionary of Features and Distance Table

Once the $C 1$ layer is created, sampling is performed by centering patches of size $4 \times 4$ at random positions and scales using a normalized random number generator function. A distance table is created to store the actual weighted Euclidean distances of the indices from YIQ quantization cores. Since the values of the $Y$ channel are normally distributed between $[0,1]$, but the values of $I$ and $Q$ channels fall in the approximate range of $[-0.6,+0.6]$ and $[-0.5,+0.5]$ respectively, and as in most of the images the actual values of these two latter channels fall between $[-0.1,+0.2]$ and $[-0.1,+0.1]$ (before normalization to $[0,1]$ ) we weighed the distances to have an equal effect in the distance calculation. The distance table weights are calculated as:

$$
\text { DistanceTable }(i, j)=\sqrt{D(1)+\gamma D(2)+\beta D(3)}
$$

$$
\begin{equation*}
\text { Where } D(k)=(Y \operatorname{IQCore}(i, k)-Y \operatorname{IQCore}(j, k))^{2} \tag{1}
\end{equation*}
$$

with $\gamma=3.3$ and $\beta=5$. In Jalali, Lim, Ong, and Tham (2010) and Jalali, Lim, Tham, and Ong (2012) various clustering methods in the creation of the dictionary of features were implemented and it is shown that by use of random sampling in HMAX model, relatively good results can be achieved with a lower computational cost in comparison with clustering of features.

## $S 2$ Layer

Once the dictionary of features and the distance table are created, each entry in the dictionary of features is used as a filter to be convolved on $C 1$ patches of size $4 \times 4$ on the neighbor scales of the dictionary feature in the pyramid. The responses $V(d, p)$ of each dictionary feature, $d$ to all of the neighbor patches of the same size in $\pm 1$ scale and $\pm 10 \%$ in position, $p$ are calculated using a Euclidean distance equation as:

$$
\begin{equation*}
V(d, p)=\exp \left(-\frac{\|d-p\|^{2}}{2 \sigma^{2} \alpha}\right) \tag{2}
\end{equation*}
$$

where $d$ is a feature in the dictionary and $p$ is a patch in the image $C 1$ pyramid. $\sigma$ and $\alpha$ are set to 0.5 and 1 respectively as in Mutch, Knoblich, and Poggio (2010).

## $C 2$ Layer

Once the $S 2$ layer is generated, the maximum values for each patch in the dictionary are taken as the $C 2$ output. This layer
outputs a vector of the same size as the dictionary of features. We chose different sizes for the dictionary of features and in most cases a dictionary of size 10000 was chosen which results in slightly better performances than smaller sizes of about 1000 dimensions.

## Classification Layer

The $C 2$ vectors are classified using a multi-class one-versusrest linear kernel support vector machine. The algorithm used to train the classifier is weighted regularized least-squares after the data is sphered and the mean and variance of each dimension are normalized to zero and one respectively as in Mutch and Lowe (2008).

## Use of HMAX for Encoding Shape Information

For shape information, we used the HMAX model implementation of Mutch and Lowe (2008). In HMAX, the maximum response of the $S 2$ layer is chosen as the $C 2$ layer to be fed to the classifier. An $N$-dimensional vector is calculated as the output of the $C 2$ layer, where each element is the maximum response (everywhere in the image in Serre, Oliva, and Poggio (2007) and in a spatial neighborhood of each dictionary feature in Mutch and Lowe (2008)) over image patches for each dictionary feature where $N$ is the number of features in the dictionary.

Let $V_{i}^{j}$ be the response of the image patch $p_{i}$ to the dictionary feature $d_{j}$ calculated using Eq. 2. The response of the $C 2$ layer is calculated as:

$$
\begin{array}{r}
C 2(j)=\max \left(V_{i}^{j}\right) \text { for } \forall i \in M \\
\text { for } j=1, \ldots, N \tag{3}
\end{array}
$$

where $M$ is the number of valid patches in each image and $N$ is the size of the dictionary of features. This is consistent with the recent HMAX models (Mutch \& Lowe, 2008; Serre, Oliva, \& Poggio, 2007; Jalali et al., 2010; Theriault, Thome, \& Cord, 2011).

## Experimental Results

First we examine the naive use of color by computing various color spaces (RGB, HSV, YIQ) on the Caltech101 dataset (Fei-Fei, Fergus, \& Perona, 2004) and compare the results with grayscale images. The Caltech 101 dataset, includes 101 classes of objects plus a background category. Each class contains between 31 to 800 color images of different sizes. The size of each image is approximately $300 \times 200$ pixels on average. We used 30 randomly chosen images for training from each class and the rest of the images were used in the test phase. We first divide the images into three channels and feed them to the unmodified HMAX (Mutch \& Lowe, 2008) directly and evaluate the classification performance.

As can be seen in Table 1, the use of three different channels and concatenating the $C 2$ vectors of all channels to the SVM provides only marginal improvement. Since the YIQ color space gives the best overall results, we use this color space in our color model. In the rest of this section, we

| Color Component | Caltech101 | Scenes |
| :--- | :--- | :--- |
| Y channel (i.e. gray scale) | 54.65 | 71.48 |
| I channel | 35.20 | 54.62 |
| Q channel | 26.86 | 50.75 |
| YIQ channels concatenated | 55.06 | 72.66 |
| RGB channels concatenated | 26.53 | 73.81 |
| HSV channels concatenated | 31.32 | 73.69 |

Table 1: Results (percentage accuracy) for the naive use of various color channels and color spaces.
evaluated our model on three datasets: Caltech101, Outdoor Scenes and Underwater Images.

## Caltech101 Dataset

The results of using CQ-HMAX on Caltech 101 are shown in Table 2. All experiments are performed 8 times on random splits of training and test sets and the average performance is reported. As can be seen, the use of our color model in this dataset does not outperform HMAX. However, when the $C 2$ features of the color model are concatenated with $C 2$ features of HMAX, the classification results are improved by more than $6 \%$ over HMAX alone. HMAX is a computationally expensive model as Gabor filter responses over different orientations in $S 1$ layer are calculated. However, CQ-HMAX is relatively faster than HMAX as it performs a quantization with 125 cores in the $S 1$ layer instead of Gabor filters.

| Model | Caltech101 | Scenes | UWI |
| :--- | :--- | :--- | :--- |
| HMAX (i.e. shape) | 54.65 | 71.48 | 92.93 |
| CQ-HMAX (i.e. color) | 38.11 | 69.21 | 94.03 |
| CQ-HMAX + HMAX | $\mathbf{6 1 . 0 9}$ | $\mathbf{7 8 . 9 7}$ | $\mathbf{9 6 . 2 3}$ |

Table 2: Results (percentage accuracy) on the Caltech101, Outdoor Scenes and Underwater Images (UWI) datasets.

## Outdoor Scenes Dataset

This dataset contains 8 outdoor scene categories: coast, mountain, forest, open country, street, inside city, tall buildings and highways (Oliva \& Torralba, 2001). There are 2600 color images of size $256 \times 256$ pixels. We used 100 random images per category for training and the rest ( 236 on average) for testing. As can be seen in Table 2, the combination of shape and color significantly improves performance.

## Underwater Images Dataset

We also evaluated CQ-HMAX on the Underwater Images dataset (Jalali, Tan, Lim, Tham, Ong, Seekings, \& Taylor, 2013). This dataset is made of 1664 images of around 740 x 420 pixels from 13 different categories. We used 30 randomly selected images per category for training and the rest for testing. These underwater images contain small objects of various shapes and color against a varied seabed background. The main challenge with these images is in light absorption by the water, and the existence of particles that limit visibility and result in scattering and reflection of light. In this experiment, we created a set of images using both grayscale and color cameras and compared the performance of CQ-HMAX
on color images and HMAX on grayscale images. As seen in Table 2, the classification accuracy increases when color and shape information are combined.

## Conclusions

In this paper, we introduced a new biologically-inspired approach to image classification which uses color in a manner consistent with high-level visual cortex processing by incorporating insights from cognitive psychology and neuroscience. We ran this model on several datasets such as Caltech101, Outdoor Scenes and Underwater Images. The combination of our color features with (grayscale) shape features led to significant increases in performance over shape or color features alone. Using our model, performance is significantly higher than using color naively, i.e. concatenating the channels of various color spaces.

Currently, our model quantizes the YIQ color space into cubed-shaped "color cores" at the $S 1$ layer. Following the work of Shahbaz Khan et al. (2012) and Van De Weijer and Schmid (2006), learning the color values that correspond to semantic color names such as "orange", "brown", could also further improve performance. Alternatively, color cores can be learnt through unsupervised clustering, in which more frequent colors in each dataset are chosen as color cores.

Our model emulates color processing in the high-level IT cortex. Interestingly, the combination of our features with those of Zhang, Barhomi, and Serre (2012) - a biologicallyinspired model that emulates the lower-level cortex - results in classification performance as good (or better) than the state-of-the-art on several benchmark datasets (Jalali, Tan, Lim, Tham, \& Ong, 2013).

## References

Banno, T., Ichinohe, N., Rockland, K. S., \& Komatsu, H. (2011). Reciprocal connectivity of identified colorprocessing modules in the monkey inferior temporal cortex. Cerebral Cortex, 21(6), 1295-310.
Bosch, A., Zisserman, A., \& Muñoz, X. (2008). Scene classification using a hybrid generative/discriminative approach. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(4), 712-27.
Boynton, R. M., \& Olson, C. X. (1987). Locating basic colors in the OSA space. Color Research \& Application, 12(2), 94-105.
Brown, M., \& Susstrunk, S. (2011). Multi-spectral SIFT for scene category recognition. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 177184). IEEE.

Burghouts, G. J., \& Geusebroek, J.-M. (2009). Performance evaluation of local colour invariants. Computer Vision and Image Understanding, 113(1), 48-62.
Conway, B. R., Chatterjee, S., Field, G. D., Horwitz, G. D., Johnson, E. N., Koida, K., et al. (2010). Advances in color science: from retina to behavior. Journal of Neuroscience, 30(45), 14955-63.

Conway, B. R., Moeller, S., \& Tsao, D. Y. (2007). Specialized color modules in macaque extrastriate cortex. Neuron, 56(3), 560-73.
Desimone, R., Schein, S. J., Moran, J., \& Ungerleider, L. G. (1985). Contour, color and shape analysis beyond the striate cortex. Vision Research, 25(3), 441-52.
Edwards, R., Xiao, D., Keysers, C., Földiák, P., \& Perrett, D. (2003). Color sensitivity of cells responsive to complex stimuli in the temporal cortex. Journal of Neurophysiology, 90(2), 1245-56.
Fei-Fei, L., Fergus, R., \& Perona, P. (2004). Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In IEEE CVPR 2004 Workshop on GenerativeModel Based Vision (Vol. 2).
Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics, 36(4), 193-202.
Gevers, T., \& Stokman, H. (2004). Robust histogram construction from color invariants for object recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(1), 113-118.
Heywood, C. A., Shields, C., \& Cowey, A. (1988). The involvement of the temporal lobes in colour discrimination. Experimental Brain Research., 71(2), 437-41.
Horel, J. A. (1994). Retrieval of color and form during suppression of temporal cortex with cold. Behavioural Brain Research, 65(2), 165-72.
Hurvich, L. M. (1981). Color Vision. Sutherland, MA.: Sinauer Associates Inc.
Jalali, S., Lim, J., Ong, S., \& Tham, J. (2010). Dictionary of features in a biologically inspired approach to image classification. In Internationl Conference on Neural Information Processing (ICONIP) (pp. 541-548). Springer.
Jalali, S., Lim, J., Tham, J., \& Ong, S. (2012). Clustering and use of spatial and frequency information in a biologically inspired approach to image classification. In International Joint Conference on Neural Networks (pp. 1-8).
Jalali, S., Tan, C., Lim, J., Tham, J., \& Ong, S. (2013). CQHMAX: A New Biologically Inspired Color Approach to Image Classification. (Manuscript under preparation).
Jalali, S., Tan, C., Lim, J., Tham, J., Ong, S., Seekings, P., et al. (2013). Encoding Co-occurrence of Features in the HMAX Model. In Proceedings of the Annual Conference of the Cognitive Science Society.
Komatsu, H. (1993). Neural coding of color and form in the inferior temporal cortex of the monkey. Biomedical Research, 14, 7-13.
Komatsu, H. (1998). Mechanisms of central color vision. Current Opinion in Neurobiology, 8(4), 503-8.
Komatsu, H., Ideura, Y., Kaji, S., \& Yamane, S. (1992). Color selectivity of neurons in the inferior temporal cortex of the awake macaque monkey. Journal of Neuroscience, 12(2), 408-24.

Lim, J. (1999). Learning visual keywords for content-based retrieval. In International conference on multimedia computing and systems (Vol. 2, pp. 169-173).
Matuzawa, T. (1985). Colour naming and classification in a chimpanzee. Journal of Human Evolution, 14, 283 - 291.
Mutch, J., Knoblich, U., \& Poggio, T. (2010). CNS: a GPUbased framework for simulating cortically-organized networks (Tech. Rep. No. MIT-CSAIL-TR-2010-013 / CBCL286). Cambridge, MA: Massachusetts Institute of Technology.
Mutch, J., \& Lowe, D. (2008). Object class recognition and localization using sparse features with limited receptive fields. International Journal of Computer Vision, 80(1), 45-57.
Oliva, A., \& Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision, 42(3), 145-175.
Palmer, S. E. (1999). Vision science: Photons to phenomenology. Cambridge, MA: MIT Press.
Riesenhuber, M., \& Poggio, T. (1999). Hierarchical models of object recognition in cortex. Nature Neuroscience, 2, 1019-1025.
Serre, T., Oliva, A., \& Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. Proceedings of the National Academy of Sciences, 104(15), 6424.
Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., \& Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 29(3), 411-426.
Shahbaz Khan, F., Anwer, R. M., Weijer, J. Van de, Bagdanov, A. D., Vanrell, M., \& Lopez, A. M. (2012). Color attributes for object detection. In Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 3306-3313).
Stoughton, C. M., \& Conway, B. R. (2008). Neural basis for unique hues. Current Biology, 18(16), R698-9.
Tang, J., Miller, S., Singh, A., \& Abbeel, P. (2012). A textured object recognition pipeline for color and depth image data. In IEEE International Conference on Robotics and Automation (ICRA) (pp. 3467-3474).
Theriault, C., Thome, N., \& Cord, M. (2011). HMAX-S: deep scale representation for biologically inspired image classification. In International Conference on Image Processing (pp. 3-6).
Valberg, A. (2001). Unique hues: an old problem for a new generation. Vision Research, 41(13), 1645-57.
Van De Weijer, J., \& Schmid, C. (2006). Coloring local feature extraction. European Conference on Computer Vision (ECCV), 334-348.
Yasuda, M., Banno, T., \& Komatsu, H. (2010). Color selectivity of neurons in the posterior inferior temporal cortex of the macaque monkey. Cerebral Cortex, 20(7), 1630-46.
Zhang, J., Barhomi, Y., \& Serre, T. (2012). A new biologically inspired color image descriptor. In European Conference on Computer Vision (Vol. 7576, pp. 312-324). Springer.

# Encoding Co-occurrence of Features in the HMAX Model 

Sepehr Jalali (tmssj@nus.edu.sg)<br>Cheston Tan (cheston-tan@i2r.a-star.edu.sg)<br>Joo-Hwee Lim (joohwee@i2r.a-star.edu.sg)<br>Jo-Yew Tham (jytham@i2r.a-star.edu.sg)<br>Sim-Heng Ong (eleongsh@nus.edu.sg)<br>Paul James Seekings (mmrl@nus.edu.sg)<br>Elizabeth A. Taylor (tmsohe@ nus.edu.sg)

National University of Singapore, Singapore 119077
Institute for Infocomm Research, A*STAR, Singapore 138632


#### Abstract

We introduce a method for encoding co-occurrence of features in the HMAX model of visual recognition, and conduct a series of experiments to investigate the contribution of co-occurrence towards better recognition performance. We show that classification accuracy is increased by adding a higher-order layer to the HMAX processing hierarchy, whereby co-occurrence of features is encoded as a new dictionary of features. We show that concatenation of mean pooling, max pooling and co-occurrence information results in better classification results on three datasets (Caltech101, a subset of Caltech256, and TMSI Underwater Images). Overall, we show that incorporating co-occurrence statistics into a biologically-inspired model of visual recognition provides a boost in classification performance above that produced by incorporating occurrence statistics alone.


Keywords: computer vision; HMAX; biologically inspired; co-occurrence statistics; visual cortex; image classification.

## Introduction

Certain categories of visual stimuli can be characterized by the co-occurrence of multiple features. For example, images of cars frequently contain wheels, doors and windows. These co-occurring features do not occur in rigid configurations. Even for a rigid object, 3D rotations can result in inter-feature distances changing when projected as 2D images. However, co-occurring features are generally found close to each other. Using faces as an example, the exact distances between facial features (e.g. eyes, nose, mouth) vary from person to person, but these features are always relatively near to each other.

Can this particular property be exploited to achieve better visual recognition performance? This question cannot be cleanly answered through behavioral experiments unless brain cells encoding co-occurrence can somehow be "turned off"; computational modeling may be a better approach. In this paper, as a proof-of-concept, we modify the biologicallyinspired HMAX model of visual recognition (Riesenhuber \& Poggio, 1999) to encode co-occurrence statistics that are learnt from a training set of images, and we show that recognition performance does indeed improve.

## Background

There is evidence for Max spatial pooling (finding the maximum among a set of inputs from a local spatial region) occurring at multiple levels in the visual system in the primary
visual cortex of cats (Finn \& Ferster, 2007; Lampl, Ferster, Poggio, \& Riesenhuber, 2004), as well as in the higher visual areas of monkeys, such as areas V4 (Gawne \& Martin, 2002) and IT (Sato, 1989). Importantly, however, each of these studies also showed evidence for "Average" pooling occurring, which can be interpreted as encoding the mean occurrence frequency of features.

Beyond just being tuned to the statistics of feature occurrences, there is strong evidence that the primate visual system is also tuned to co-occurrence statistics. This refers to either the joint or conditional probabilities of two (or more) features occurring together within images belonging to a certain object category or across categories. Since a "feature" is not always a precisely defined concept, how can the cooccurrence of two features be distinguished from the occurrence of a single feature that happens to be comprised of two simpler features? To make this distinction unambiguous, experiments were designed such that the elementary features are visually distinct, due to explicit segmentation, due to spatial separation, or from the task context. We term such features, which are the result of sensitivity to co-occurrence, as "cooccurrence features".

In some sense, mid-level features themselves can be considered as co-occurrence features, with their elementary features being simple orientation-sensitive filters (corresponding to orientation-sensitive neurons in the primary visual cortex). Since lines, curves and contours are ubiquitous in images, the presence of a short line segment of a certain orientation strongly predicts that the orientation of a neighboring line segment will be similar. This is particularly so if the relative position of that neighboring line segment is such that the two line segments have the possibility of being collinear.

Our focus here is on high-level features whose elementary features are more complex than simple oriented filters. These high-level features approach the level of semantic object parts or possibly even objects themselves. In the rest of this section, we will review the experimental evidence that the primate visual system develops sensitivity to such high-level co-occurrence features.

In the field known as visual statistical learning (VSL), it has clearly been shown that adult humans develop sensitivity to co-occurrence statistics in images (Fiser \& Aslin, 2001;

Aslin \& Newport, 2012). In a ground-breaking study by Fiser and Aslin (2002) it was shown that 9-month-old infants already developed sensitivity to visual co-occurrence statistics.

There is also an abundance of evidence from monkeys that their visual systems develop sensitivity to co-occurrence statistics. Miyashita (1988) and Sakai and Miyashita (1991), monkeys were trained to recognize pairs of stimuli, in a paradigm known as paired-associate learning. Neurons were found that were sensitive to such trained stimulus pairs, but not other stimulus pairs. The pairings were arbitrary, making the likelihood that such neurons had already possessed such sensitivity vanishingly small. More recently, Hirabayashi and Miyashita (2005) found that populations of IT neurons are sensitive to feature configuration within objects.

Direct evidence for sensitivity to co-occurrence (over and above sensitivity to occurrence) was found by Baker, Behrmann, and Olson (2002). Monkeys were trained to discriminate between objects that were each composed of two distinct parts linked by a line, forming "baton" objects. Compared to untrained objects, selectivity for trained objects was enhanced. This was for both the individual parts, as well as the combined "baton" objects. Crucially, selectivity for the two parts together (i.e. the whole object) was greater than the combined (summed) selectivity for each individual part.

Under what conditions does sensitivity to co-occurrence develop? In human adults, this is an implicit process that develops without awareness of the co-occurrence statistics, using a "cover task" or even through mere exposure (TurkBrowne, Jungé, \& Scholl, 2005; Turk-Browne, Scholl, Chun, \& Johnson, 2009; Aslin \& Newport, 2012). This is also true for human infants (Fiser \& Aslin, 2002; Aslin \& Newport, 2012). In monkeys, most work has been done using active task learning. This is so that the neural selectivity for trained objects can be compared to the control set of untrained objects. Since neural selectivity is enhanced for features that are diagnostic for active task learning (Sigala \& Logothetis, 2002), passive viewing may not be sufficient to produce selectivity that is large enough to be statistically significant when measured from electrode recordings.

How has sensitivity to co-occurrence been measured experimentally? The methods have generally been constrained by the nature of the subjects. Adult human subjects have generally been tested behaviorally, i.e. through their explicit responses (usually simple 'yes/no' tests). More recently, fMRI has been shown to be able to detect co-occurrence sensitivity (Turk-Browne et al., 2009). In human infants, due to their inability to understand or respond explicitly to verbal instruction, experiments have been constrained to using tests for novelty detection that are ubiquitous for infants. In monkeys, due to the ability to conduct invasive experiments that are not possible with humans, scientists have conducted electrophysiological experiments (i.e. using electrodes to record the responses of individual neurons). Such experiments allow for a detailed, "close-up" analysis of the effects of cooccurrence at the level of individual neurons e.g. Baker et al.
(2002); Sakai and Miyashita (1991). However, there are limitations, such as the presence of noise, limited recording time, and the ability to record from at most a few hundred neurons.

Beyond just "being sensitive" to co-occurrence statistics, what are the characteristics of such sensitivity? It is specific to spatial configuration, such as the relative position of the elementary features (Hirabayashi \& Miyashita, 2005). In addition, this sensitivity is reflected not in strength of neural responses per se, but rather in the selectivity for co-occurring features relative to non-co-occurring features (Baker et al., 2002).

One special case of sensitivity to co-occurrence of features is that of faces. The elementary features are semantic face parts such as the eyes, nose and mouth. It is very well-established that humans and monkeys are sensitive to the combination and relative configuration of face parts. Specifically, any change to the normal configuration of the face leads to reduced neural responses and poorer recognition accuracy. One manifestation of this is the Face Inversion Effect (FIE), whereby inverted faces are much more poorly recognized than upright faces (Yin, 1969). Faces with the parts in scrambled configurations are also poorly recognized. Furthermore, the sensitivity to co-occurrence seems to be unavoidable. In what is known as the Composite Face Effect, people are sensitive to the bottom halves of faces, even when they are explicitly instructed to ignore them during a discrimination task (Young, Hellawell, \& Hay, 1987).

Generally, such sensitivity requires normal visual experience during infancy in order to develop (Le Grand, Mondloch, Maurer, \& Brent, 2004). It also develops quickly, reaching adults levels (at least qualitatively) by age 4 (Heering, Houthuys, \& Rossion, 2007); this is consistent with the notion that passive exposure is sufficient for co-occurrence sensitivity to develop (see above). Evidence for sensitivity to co-occurrence for face parts has also been found at the level of single neurons. Freiwald, Tsao, and Livingstone (2009) found that in one of the brain regions that respond selectivity to faces, neurons on average responded to combinations of two to three face parts, rather than individual parts. Cooccurrences have been studied in a series of experiments such as Edelman, Yang, Hiles, and Intrator (2002).

Use of co-occurrences of features for creating more complex features in Fidler, Boben, and Leonardis (2008) shows an improvement in classification accuracy, and bag-of-features approaches show improvements in classification results using frequency of patches in the images in (Fei-Fei \& Perona, 2005). Co-occurrence information can be used to find partpart and part-whole relations of features of different receptive field sizes. If a feature is occurring too often in a class (and not likewise in other classes), it is more likely to be a discriminant feature in that class and if two features are cooccurring in a class often in a neighborhood, they may be part of a more complex feature and can have a part-part relationship and they might be more related to the object rather than the background (unless the background is also repetitive, e.g.
sky in airplane images). Also, if there exist features of different sizes and they co-occur in the same position on different scales they are likely to have a part-whole relationship.

## HMAX Model

The HMAX model (Riesenhuber \& Poggio, 1999) simulates the feed-forward path of the visual cortex. This model is used to find a good trade-off between invariance and selectivity. $S 1$ cells provide selectivity by responding to oriented filters and $C 1$ cells provide invariance by pooling over neighboring scales and positions. We use the HMAX model presented in Mutch and Lowe (2008) in the first three layers (S1, C1 and $S 2$ ). Here we have a brief review on this model and show our modifications to it.

In this implementation, an image is fed into the structure and 10 different scales of the image are created as inputs to $S 1$ layer. Gabor filters in 12 orientations are created as $S 1$ layer filters:

$$
\begin{equation*}
G(x, y)=\exp \left(-\frac{\left(X^{2}+\gamma^{2} Y^{2}\right)}{2 \sigma^{2}}\right) \cos \left(\frac{2 \pi}{\lambda} X\right) \tag{1}
\end{equation*}
$$

where $X=x \cos \theta-y \sin \theta$ and $Y=x \sin \theta+y \cos \theta$. The values of $x$ and $y$ vary between -5 and 5 , and $\theta$ varies between 0 and $\pi$. The parameters $\gamma$ (aspect ratio), $\sigma$ (effective width), and $\lambda$ (wavelength) are all taken from Serre, Wolf, and Poggio (2005) and are set to $0.3,4.5$, and 5.6 respectively.

A fixed size of Gabor filters is implemented on different scales of the images where the smaller edge of the biggest image is set to 140 pixels while maintaining the aspect ratio (the image pyramid of 10 scales created each layer by a factor of $2^{1 / 4}$ smaller than the last using bicubic interpolation). The response of a patch of pixels $X$ to a particular $S 1$ filter $G$ is given by:

$$
\begin{equation*}
R(x, y)=\left|\frac{\sum X_{i} G_{i}}{\sqrt{\sum X_{i}^{2}}}\right| \tag{2}
\end{equation*}
$$

These outputs are sent to the $C 1$ layer, which performs a local $3 D$ max operation on both scale $( \pm 1)$ and position $(3 \times 3$ neighborhood) of the filter responses. The output of this layer is a pyramid consisted of between 500-2000 different patches of size $4 \times 4,8 \times 8,12 \times 12$ and $16 \times 16$ in 8 scales depending on the size of the input image. In this level one or two samples are randomly sampled from each training image (from random scales and positions) and a dictionary of features of size 4096 is created. This dictionary is then made sparse by selecting the highest response from each orientation and setting the rest to 0 .

The response of a patch of $C 1$ units $X$ to a particular $S 2$ feature/prototype $P$ (a dictionary feature), of size $n \times n$, is given by a Gaussian radial basis function:

$$
\begin{equation*}
R(X, P)=\exp \left(-\frac{\|X-P\|^{2}}{\sigma^{2}}\right) \tag{3}
\end{equation*}
$$



Figure 1: In HMAX, the max on the columns is taken as the response for creating $C 2$ output vector. In contrast, histogram approaches based on SIFT methods use the frequency of feature occurrence, i.e. the normalized sum of the max values on the rows.

The values of $R$ are stored as $S 2$ layer. The distance of each sample from each training image with each entry on the dictionary is calculated and a local max is taken in $C 2$ layer in $\pm 1$ scale and $\pm 10 \%$ spatial neighborhood (despite a global max in Serre et al. (Serre et al., 2005)). These $C 2$ features are sent to the SVM for training. For testing images the same hierarchical procedure is repeated. In (Mutch \& Lowe, 2008) sparse prototypes are calculated and the maximum response from all directions for each window is taken and SVM normals method (Mladenić, Brank, Grobelnik, \& Milic-Frayling, 2004) is used to select the features with higher weights. In this approach, SVM is run a few times, and each time features with lower weights are dropped. In this HMAX implementation, once $S 2$ features are calculated, the $C 2$ layer is calculated as:

$$
\begin{array}{r}
C 2(n)=\max \left(V_{k}^{n}\right) \text { for } \forall k \in M \\
\text { for } n=1, \ldots, N \tag{4}
\end{array}
$$

As can be seen in Figure 1 in conventional HMAX approaches, the max on the columns is taken as the value for $C 2$ either in a local neighborhood of each feature or globally. Since taking the max in a local neighborhood (in $\pm 1$ scale and $\pm 10 \%$ spatial neighborhood) is shown to improve the performance by about 5\% in Caltech101 dataset in Mutch and Lowe (2008), in our experiments we also use a local neighborhood for calculating the responses. We also eliminate the local inhibition in $S 2$ level proposed in Mutch and Lowe (2008) as it increased the performance. Once a feature belongs to the first or last scale in the pyramid, we extend the neighborhood to two neighboring scales. Same method is used for features which fall in the borders of each scale, and $+20 \%$ or $-20 \%$ of their neighborhood is used for comparisons.

If we take the sum of the values on rows in Figure 1 and normalize them, these are "HMean" features, which are also biologically-inspired, and significantly improve classification results when concatenated with HMAX features (Jalali, Lim,

Tham, \& Ong, 2012). HMean is equivalent to the feature occurrence frequency in "bag-of-features" methods.

## Encoding Co-occurrence of Features

For each class, we first find the value and index of the mostfrequently occurring features (MOF). The next step is to encode the co-occurrence of these features as can be seen in Figure 1. For every class, we calculate the co-occurrence of the most frequent features and store it as a $S 3$ dictionary feature. Hence a new dictionary of features is added to the model which is composed of \#MOF $\times$ \#MOF entries for each class, where \#MOF was set as 20. In this dictionary of features, the value of each dictionary feature is calculated as:

$$
\begin{equation*}
C 3(i, j)=C 2(i) C 2(j) \exp \left(-\frac{\left\|S_{i}-S_{j}\right\|^{2}}{\sigma^{2}}\right) \tag{5}
\end{equation*}
$$

where $S_{n}$ represents the spatial position of the $C 2$ feature and $\sigma=0.5$.

This dictionary encodes the value of co-occurrence of every pair of features selected for each class. Hence we will have $N N$ dictionaries where $N N$ stands for the number of categories in the classification task. These dictionaries are concatenated to create the $C 2$ dictionary of features. In the training and test phases, the respective feature to each dictionary feature is found (the most similar feature in every image) and the similarity of the values in dictionary of features are calculated for every image. This results in a \#MOF $\times$ \#MOF $\times$ NN feature as the $C 3$ feature and it is concatenated to $C 2$ feature vector and sent to the classifier for classification. The extended model for encoding the co-occurrence of features is shown in Figure 2.

## Experimental Results

We evaluated our co-occurrence model on the Caltech101 dataset (Fei-Fei, Fergus, \& Perona, 2004). The model was trained on 30 images per category (standard for this dataset; see Mutch and Lowe (2008)), and tested on all the other images. We also used the Caltech 256 dataset (Griffin, Holub, \& Perona, 2007), because it allows for more images per category than Caltech101. In particular, we considered only the 14 (out of 256) categories which had 200 or more images. We trained the model on 150 images (so that there would be at least 50 images for testing), and tested on the rest. We also examined classification accuracy as a function of number of training images for Caltech256. This was motivated by the concern that co-occurrence features could require more data for reliable co-occurrence statistics to be extracted, before the advantage of co-occurrence could be properly manifested.

We also evaluated the performance of our model on a new dataset consisting of images of underwater targets. The main challenge with underwater images is the existence of particles that limit the visibility in unclear waters and results in scattering, reflection and absorption of light, and the differential absorption of light of different wavelengths by water itself. This dataset consists of 1664 images (roughly $740 \times 420$ pixels in


Figure 2: Diagram of model processing hierarchy.
size) from 13 categories. Example images from this dataset are shown in Figure 3. We used 30 images per category for training, and the rest for testing.

Results are shown in Table 1. For all images, only intensity (luminance) information was used. All results were derived using 8 random train/test splits. For all three datasets, the combination of HMAX and co-occurrence features gave better results (classification accuracy) than either type of feature alone (Caltech101: $59.3 \%$ vs. $54.7 \%$ vs. $57.7 \%$; Caltech 256 : $64.4 \%$ vs. $60.2 \%$ vs. $48.6 \%$; Underwater Images: $98.7 \%$ vs. $92.9 \%$ vs. $92.2 \%$ ). Since co-occurrence features were derived from the co-occurrence of HMean features, we also compared which of these two feature types (co-occurrence vs. HMean) gave better results when combined with HMAX. Again, for all three datasets, combining co-occurrence features with HMAX produced better results than combining HMean with HMAX (Caltech101: $59.3 \%$ vs. $58.9 \%$; Caltech256: $64.4 \%$ vs. $61.3 \%$; Underwater Images: $98.7 \%$ vs. $98.3 \%$ ). Furthermore, for all datasets, the combination of all three feature types was better than just HMAX and HMean together (Caltech101: $60.1 \%$ vs. $58.9 \%$; Caltech 256 : $64.1 \%$ vs. $61.3 \%$; Underwater Images: $99.0 \%$ vs. $98.3 \%$ ).

We also examined the effect of disregarding spatial distance (i.e. the exponential in Eq. 5). As seen in Table 1, for all datasets, results were better when spatial distance was taken into account (Caltech101: 57.7\% vs. 55.1\%; Caltech256: $48.6 \%$ vs. $44.2 \%$; Underwater Images: $92.2 \%$ vs. $83.3 \%$ ).


Figure 3: Examples from TMSI Underwater Images dataset.

Table 1: Classification performance on the Caltech101, Caltech256 (subset - see text for details), and TMSI Underwater Images datasets.

| Method | Caltech101 | Caltech256 <br> (subset) | Underwater <br> Images |
| :--- | :--- | :--- | :--- |
| HMAX | 54.7 | 60.2 | 92.9 |
| Co-occurrence <br> (no distance) | 55.1 | 44.2 | 83.3 |
| Co-occurrence | 57.7 | 48.6 | 92.2 |
| HMAX + <br> Co-occurrence | 59.3 | 64.4 | 98.7 |
| HMAX + HMean | 58.9 | 61.3 | 98.3 |
| HMAX + HMean <br> + Co-occurrence | 60.1 | 64.1 | 99.0 |

In order to evaluate the effect of number of training images for the creation of co-occurrence features, we trained the model with varying numbers of training images per category. As shown in Figure 4, the performance boost when adding co-occurrence features was greatest when using 150 training images. However, for fewer than 150 training images, the boost from adding co-occurrence features is unreliable. Nonetheless, looking at just HMAX alone, performance seems to asymptote at 150 training images, but for the combination of HMAX and co-occurrence features, performance seems to increase roughly linearly with the number of training images. While empirically, co-occurrence may help performance in all datasets, similar analyses (i.e. performance boost as a function of number of training images) for the other 2 datasets may not be meaningful, since the maximum number of training images is only 30 per category.


Figure 4: Classification accuracy on Caltech256 as a function of number of training images.

## Discussion

In this paper, we showed that combining co-occurrence features with regular HMAX features leads to better classification performance than using either feature type alone. Furthermore, adding co-occurrence features to HMAX increases performance more than adding occurrence features. The three types of features encode different information, and therefore the combination of all three feature types gave the best overall performance. For co-occurrence, the spatial distance between the two co-occurring features also contributes to better performance. In this work, we focused solely on HMAX. However, in future work, our co-occurrence method can be applied to other vision algorithms.

In preliminary experiments not reported here, we experimented with creating co-occurrence features from HMAX features (rather than HMean features, as done in this paper). However, this resulted in either a drop in performance or no change. This will be investigated further in future work.

Fig. 4 suggests that the performance boost from using cooccurrence may be limited by the number of training images. More detailed investigation is limited by the relatively small number of images per category in these datasets. Further investigation may require utilizing or creating larger datasets.

Another prospect for further improvement is to encode cooccurrence of more than two features. However, besides possibly requiring even more training data than two-feature co-occurrence, there may be diminishing returns for such "higher-order" co-occurrences. This is because relatively fewer classes will have the underlying visual structure that will benefit from encoding such co-occurrences.

In this paper, the choice of features for encoding cooccurrence was based on their frequency. Choosing discriminative (rather than frequent) features for co-occurrence encoding may be a more direct approach to maximizing classification performance. To choose discriminative features, one approach is to train the SVM several times and remove fea-
tures with low weights, as in Mutch and Lowe (2008), or to simply use features with mean response values that differ the most between different classes.

## References

Aslin, R. N., \& Newport, E. L. (2012). Statistical Learning: From Acquiring Specific Items to Forming General Rules. Current Directions in Psychological Science, 21(3), 170176.

Baker, C. I., Behrmann, M., \& Olson, C. R. (2002). Impact of learning on representation of parts and wholes in monkey inferotemporal cortex. Nature Neuroscience, 5(11), 12106.

Edelman, S., Yang, H., Hiles, B., \& Intrator, N. (2002). Probabilistic principles in unsupervised learning of visual structure: human data and a model. Advances in Neural Information Processing Systems, 1, 19-26.
Fei-Fei, L., Fergus, R., \& Perona, P. (2004). Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. IEEE. CVPR 2004. In Workshop on generativemodel based vision (Vol. 2).
Fei-Fei, L., \& Perona, P. (2005). A bayesian hierarchical model for learning natural scene categories. In IEEE Conference on Computer Vision and Pattern Recognition (Vol. 2, pp. 524-531).
Fidler, S., Boben, M., \& Leonardis, A. (2008). Similaritybased cross-layered hierarchical representation for object categorization. In IEEE Conference on Computer Vision and Pattern Recognition (pp. 1-8).
Finn, I. M., \& Ferster, D. (2007). Computational diversity in complex cells of cat primary visual cortex. Journal of Neuroscience, 27(36), 9638-48.
Fiser, J., \& Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. Psychological Science, 12(6), 499-504.
Fiser, J., \& Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. Proceedings of the National Academy of Sciences of the United States of America, 99(24), 15822-6.
Freiwald, W. A., Tsao, D. Y., \& Livingstone, M. S. (2009). A face feature space in the macaque temporal lobe. Nature Neuroscience, 12(9), 1187-96.
Gawne, T. J., \& Martin, J. M. (2002). Responses of Primate Visual Cortical V4 Neurons to Simultaneously Presented Stimuli. Journal of Neurophysiology, 88(3), 1128-1135.
Griffin, G., Holub, A., \& Perona, P. (2007). Caltech-256 object category dataset (Tech. Rep. No. 7694). California Institute of Technology.
Heering, A. de, Houthuys, S., \& Rossion, B. (2007). Holistic face processing is mature at 4 years of age: evidence from the composite face effect. Journal of Experimental Child Psychology, 96(1), 57-70.
Hirabayashi, T., \& Miyashita, Y. (2005). Dynamically modulated spike correlation in monkey inferior temporal cortex
depending on the feature configuration within a whole object. Journal of Neuroscience, 25(44), 10299-307.
Jalali, S., Lim, J., Tham, J., \& Ong, S. (2012). Clustering and use of spatial and frequency information in a biologically inspired approach to image classification. In International Joint Conference on Neural Networks (pp. 1-8).
Lampl, I., Ferster, D., Poggio, T., \& Riesenhuber, M. (2004). Intracellular measurements of spatial integration and the MAX operation in complex cells of the cat primary visual cortex. Journal of Neurophysiology, 92(5), 2704-13.
Le Grand, R., Mondloch, C. J., Maurer, D., \& Brent, H. P. (2004). Impairment in holistic face processing following early visual deprivation. Psychological Science, 15(11), 762-8.
Miyashita, Y. (1988). Neuronal correlate of visual associative long-term memory in the primate temporal cortex. Nature, 335, 817-20.
Mladenić, D., Brank, J., Grobelnik, M., \& Milic-Frayling, N. (2004). Feature selection using linear classifier weights: interaction with classification models. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 234-241).
Mutch, J., \& Lowe, D. (2008). Object class recognition and localization using sparse features with limited receptive fields. International Journal of Computer Vision, 80(1), 45-57.
Riesenhuber, M., \& Poggio, T. (1999). Hierarchical models of object recognition in cortex. Nature Neuroscience, 2, 1019-1025.
Sakai, K., \& Miyashita, Y. (1991). Neural organization for the long-term memory of paired associates. Nature, 354, 152-5.
Sato, T. (1989). Interactions of visual stimuli in the receptive fields of inferior temporal neurons in awake macaques. Experimental Brain Research, 77(1), 23-30.
Serre, T., Wolf, L., \& Poggio, T. (2005). Object recognition with features inspired by visual cortex. In IEEE Conference on Computer Vision and Pattern Recognition (Vol. 2, pp. 994-1000).
Sigala, N., \& Logothetis, N. K. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. Nature, 415, 318-20.
Turk-Browne, N. B., Jungé, J., \& Scholl, B. J. (2005). The automaticity of visual statistical learning. Journal of Experimental Psychology: General, 134(4), 552-64.
Turk-Browne, N. B., Scholl, B. J., Chun, M. M., \& Johnson, M. K. (2009). Neural evidence of statistical learning: efficient detection of visual regularities without awareness. Journal of Cognitive Neuroscience, 21(10), 1934-45.
Yin, R. K. (1969). Looking at upside-down faces. Journal of Experimental Psychology, 81(1), 141-145.
Young, A. W., Hellawell, D., \& Hay, D. C. (1987). Configurational information in face perception. Perception, 16(6), 747-59.

# The Assumption of Class-Conditional Independence in Category Learning 

Jana Jarecki (jarecki@mpib-berlin.mpg.de)*<br>Björn Meder (meder@mpib-berlin.mpg.de)*<br>Jonathan D. Nelson (nelson@mpib-berlin.mpg.de)*<br>*Center for Adaptive Cognition and Behavior (ABC), Max Planck Institute for Human Development, Lentzeallee 94 14195 Berlin, Germany


#### Abstract

This paper investigates the role of the assumption of classconditional independence of object features in human classification learning. This assumption holds that object feature values are statistically independent of each other, given knowledge of the object's true category. Treating features as classconditionally independent can in many situations substantially facilitate learning and categorization even if the assumption is not perfectly true. Using optimal experimental design principles, we designed a task to test whether people have this default assumption when learning to categorize. Results provide some supporting evidence, although the data are mixed. What is clear is that classification behavior adapts to the structure of the environment: a category structure that is unlearnable under the assumption of class-conditional independence is learned by all participants.


Keywords: Multiple-cue classification learning; classconditional independence; naïve Bayes; causal Markov condition

## Introduction

Categorization is fundamental for cognition. Grouping together objects or events helps us to efficiently encode environmental patterns, make inferences about unobserved properties of novel instances, and make decisions. Without categorization we could not see the woods for the trees.

Despite the ease with which we form categories and use them to make inferences or judgments, from a computational perspective categorization is a challenging problem. For instance, different diseases can cause similar symptoms, entailing that diagnostic inferences are often only probabilistic. Patients may have new symptom combinations and still require a diagnosis. Depending on the specific assumptions the physician makes about the relationship between the diseases and symptoms, a physician could justifiably make very different inferences about the diseases.

In the present paper, we investigate the role of the possible assumption of class-conditional independence of features in category learning. Class-conditional independence holds if the features of the category members are statistically independent given the true class. This assumption can facilitate classification and learning of category structures. The concept of class-conditional independence underlies the naïve Bayes classifier in machine learning (Domingos \& Pazzani, 1997), and is also a key assumption in some psychological classification models (e.g., Fried \& Holyoak, 1984; Anderson, 1991). It is related to ideas of channel separability in sensory perception (Movellan \& McClelland, 2001). Similar ideas are found
in Reichenbach's (1956) common-cause principle in the philosophy of science and in causal modeling (Spirtes, Glymour, \& Scheines, 1993; Pearl, 2000).

Both the philosophical and psychological literature make claims about the normative bases of the assumption of class-conditional-independence of features. Our focus here is not on the general normativity or nonnormativity of that assumption, but on whether the assumption of class-conditional independence may (perhaps tacitly) underlie people's inferences in learning and multiple-cue categorization tasks. We think of this assumption as one of many possible default (heuristic or meta-heuristic) assumptions that, if close enough to an environment's actual structure, may facilitate learning and inferences.

## The Psychology of Conditional Independence

Some psychological models of categorization incorporate assumptions of class-conditional independence, such as the category density model (Fried \& Holyoak, 1984) or Anderson's (1991) rational model of categorization. Both models treat features of instances as class-conditionally independent to make inferences about category membership or unobserved item properties.

Other research has focused more directly on the role of conditional independence assumptions in human reasoning. For instance, a key assumption in many formal causal modeling approaches (e.g., Pearl, 2000; Spirtes et al., 1993) is the so-called causal Markov condition, which assumes that a variable in a causal network is independent of all other variables (except for its causal descendants), conditional on its direct causes. As this assumption facilitates probabilistic inferences across complex causal networks it was suggested that people's causal inferences could also comply with this conditional independence assumption.

Von Sydow, Meder, and Hagmayer (2009) investigated reasoning about causal chains and found that subjects' inferences indicated a use of conditional independence assumptions, even if the learning data suggested otherwise. ${ }^{1}$ Other research, however, found violations of the causal Markov condition (Rehder \& Burnett, 2005). Asked to infer the prob-

[^303]ability for one effect when knowing the common cause of several effects, people's judgments were influenced by the status of the other effects rather than treating all effects as independent of each other given the cause. One explanation for this "nonindependence effect" (Rehder \& Burnett, 2005) is that it might be due to subjective explanations that disable all causal links between the cause and effects at once (Walsh \& Sloman, 2007). Other researchers have argued that these Markov violations do not indicate flawed human reasoning, but reflect the use of abstract causal knowledge that is sensitive to contextual information (Mayrhofer, Hagmayer, \& Waldmann, 2010).

## Research Questions

Should the assumption of class-conditional feature independence be used in classification learning? Do people use that assumption to guide learning about the structure of a novel environment? We extend previous research fourfold: (1) We use optimal experimental design principles (Myung \& Pitt, 2009; Nelson, 2005) to explicitly address the assumption in classification, (2) we are interested in categorization learning as opposed to causal reasoning, (3) we investigate how people's experience with a new environment shapes their classification behavior, whereas many previous studies have measured explicit numerical probability judgments. (4) We use an experience-based research paradigm, whereas previous studies used numerical (Rehder \& Burnett, 2005) or verbal (Mayrhofer et al., 2010) formats. Personal experience of events has been shown to result in different behavior and learning than word- or number-based presentation of probabilities (Hertwig, Barron, Weber, \& Erev, 2004; Nelson, McKenzie, Cottrell, \& Sejnowski, 2010). Before describing the task we designed, let us turn to the normative question of class-conditional independence in classification.

## Class-Conditional Independence in Classification

Categorization entails assigning an object to a class. Let $F$ denote an object consisting of a vector of feature values $\boldsymbol{f}$, and let $C$ denote a random variable whose values are the possible classes $c_{1}, \ldots, c_{n}$. The posterior probability of the class given the observed feature values, $P$ (class $\mid$ features), can be inferred using Bayes' rule:

$$
\begin{equation*}
P(C=c \mid F=\boldsymbol{f})=\frac{P(F=\boldsymbol{f} \mid C=c) P(C=c)}{P(F=\boldsymbol{f})} \tag{1}
\end{equation*}
$$

where $P(F=f \mid C=c)$ denotes the likelihood of feature value vector $\boldsymbol{f}$ given class $c, P(C=c)$ is the prior probability of the class, and $P(F=\boldsymbol{f})$ is the occurrence probability of the feature configuration. An important question is how we estimate the relevant probabilities to infer the posterior probability. Estimating the classes' prior probabilities, $P(C=c)$, from the data is relatively straightforward. However, estimating the likelihood of the features given the class, $P(F=f \mid C=c)$, is more complicated, as the number of probabilities grows exponentially with the number of features (the curse of di-
mensionality). One way to sidestep the problem is to assume that features are class-conditionally independent.

## Class-Conditional Independence

If class-conditional independence holds the individual features within a class are statistically independent (e.g., Domingos \& Pazzani, 1997). This means that the probability of a feature configuration given a class can be factorized such that:

$$
\begin{equation*}
P(F=\boldsymbol{f} \mid C=c)=\prod_{j=1}^{J} P\left(F_{j}=\boldsymbol{f}_{j} \mid C=c\right) \tag{2}
\end{equation*}
$$

where $P(F=f \mid C=c)$ denotes the likelihood of the feature configuration given the class, $P\left(F_{j}=f_{j} \mid C=c\right)$ is the marginal likelihood of the $j^{\text {th }}$ feature value given the class, and $j=1, \ldots, J$ indexes the different features. Thus, according to the assumption of class-conditional independence, the likelihood of each feature value combination can be estimated from the likelihoods of the individual feature values.
Advantages The key advantage of assuming that features are class-conditionally independent is that it reduces the curse of dimensionality. For example, for 10 binary features there are $2^{10}$ possible feature configurations. That means, we have to estimate 1024 likelihoods of feature configurations for each class. Assuming class-conditional independence reduces the number of required likelihoods from 1024 to 8.

Another benefit is that class-conditional independence allows inferences about new feature configurations. Even if a particular combination of feature values has not been observed yet, assuming class-conditional independence allows inference of the likelihood of the feature configuration from the marginal likelihoods of the individual feature values, thereby enabling computing the posterior class probabilities.

Robustness While class-conditional independence may rarely exactly hold in real-world environments, violations of this assumption do not necessarily impair performance. For instance, a widely used classifier in machine learning is the naïve Bayes model, which treats features as classconditionally independent and computes the posterior class probabilities accordingly. Both simulation studies and analytic results demonstrate the robustness of this model under a variety of conditions (Domingos \& Pazzani, 1997). For instance, if the optimality criterion is classification accuracy (error minimization, i.e., a zero-one loss function), then even if the derived posterior probabilities do not exactly correspond to the true posterior, as long as the correct category receives the highest posterior probability, classification error will be minimized.

Summary Treating features as class-conditionally independent in a classification task can be helpful, as it simplifies the problem of parameter estimation and violations of classconditional independence do not necessarily entail a loss in classification accuracy. On the other hand, assuming classconditional independence also puts constraints on the types
of classification problems that can be solved. For instance, treating features as class-conditionally independent can make it impossible to solve certain classification problems, such as nonlinearly-separable category structures (Domingos \& Pazzani, 1997).

From a psychological perspective, however, presuming class-conditional independence might be a plausible default assumption in category learning. If features are (approximately) class-conditionally independent, this facilitates learning and inference substantially. We designed an experiment to investigate whether people initially presume class-conditional independence, and if people change their beliefs and classification behavior when class-conditional independence does not hold in the environment.

## Experiment

Our goal was to examine whether people use classconditional independence as a default assumption in category learning when the true environmental probabilities are not known yet, that is, early in learning. In order to test this question, we designed a learning environment in which classification decisions would be strongly different if the learner presumes class-conditional feature independence, rather than basing classification decisions solely on the previous instances with the exact same configuration of feature values.

## Method

Participants Thirty subjects ( $M_{\text {age }}=23, S D=3.3$ years, $70 \%$ females) participated in a computer-based experiment in exchange for 12 Euro.
Task Participants' task was to learn classify objects with three binary features into one out of two categories. As stimuli we used simulated biological "plankton" specimens differing in three binary features ("eye", "tail", and "claw", shown in the left image in Figure 1). The classes were labelled as "Species A" vs. "Species B". The assignment of the actual physical features and their values to the underlying probabilities, as well as the class labels, were randomized across participants.
Procedure We used a trial-by-trial supervised multiplecue probabilistic category learning paradigm (e.g., Knowlton, Squire, \& Gluck, 1994; Meder \& Nelson, 2012; Nelson et al., 2010; Rehder \& Hoffman, 2005). After introducing the task and familiarizing subjects with the three features, on each trial a plankton exemplar with a specific feature value combination was randomly drawn according to the true environmental probabilities (see below) and displayed on the screen. After participants made a classification decision, feedback on the true class was given and the next trial started. Learning continued until criterion performance was achieved. Criterion performance was defined as both (1) an overall classification accuracy of $98 \%$ over the last 100 trials, and (2) accurate classification of the last five instances of every individual configuration of features.

Environment Using optimal experimental design (OED) principles (Myung \& Pitt, 2009; Nelson, 2005) we conducted simulations to find environmental probabilities that best differentiate between a learner that assumes class-conditional independence and a learner that makes predictions based only on previous instances of the same feature configuration. The possible environmental probabilities for our task consisted of the following parameters: (i) the base rate of Species A (determining the Species B base rate), (ii) the likelihoods of each of the eight possible feature value combinations given Species A and (iii) the corresponding values for Species B. The parameter values were obtained via optimization, using genetic algorithms to search for desirable environments which had frequent configurations of features with large absolute discrepancies between the actual posterior probability of Species A, and the posterior probability presumed based on the class-conditional independence assumption. Formally, the genetic algorithm optimized the following fitness function:

$$
\begin{equation*}
\sum_{i=1}^{I}\left[P_{\text {true }}\left(C=c \mid F=f_{i}\right)-P_{\text {cci }}\left(C=c \mid F=f_{i}\right)\right]^{2} \times P\left(F=f_{i}\right)^{2} \tag{3}
\end{equation*}
$$

where $i$ indexes all possible feature value combinations and the subscripts true vs. cci indicate the posteriors calculated according to the true vs. class-conditionally independent parameters.

The obtained environment is summarized in Figure 1. The environment contains five out of eight possible feature combinations (henceforth denoted as $111,000,100,010,001$ ); the remaining three combinations $(011,101,110)$ do not occur. The figure illustrates the category base rates, the likelihoods of the feature configurations given the two classes, as well as the marginal likelihoods of the features, which provide the basis for inferring posterior probabilities according to the class-conditional independence assumption. Note that although nothing in the optimization prescribed finding a deterministic environment, in fact the posterior probabilities of Category A are one or zero, for each of the feature configurations that occurs.

In this environment, assuming class-conditional independence leads to classification decisions that systematically deviate from decisions based on the true environmental probabilities. Table 1 summarizes the feature configurations, their probability of occurrence, the posterior probabilities according to the true environmental probabilities, and the posterior probabilities derived assuming class-conditional independence. For four out of the five feature configurations, the classification decision derived assuming class-conditional independence conflicts with the actual class membership (indicated by $\neq$ in Table 1).

Consider feature configuration 111. This item always belongs to Species $A$ in the true environment. If features are treated as class-conditionally independent, it belongs to Species A with probability 0.91 . The small difference between the actual probability of 1.00 and 0.91 should not


Figure 1: Task environment. a) Stimuli and base rates of classes. b) Joint likelihoods of true environment. c) Marginal likelihoods used assuming class-conditional independence
change the learner's classification decision for this stimulus. This, however, is not true for the other items. For instance, according to the true environment, item 000 belongs to Species A with probability 1 , but assuming class-conditional independence entails that it belongs to Species B with probability 0.67 . Thus, a learner assuming class-conditional independence would believe that on average about $67 \%$ of the 000 items belong to Species B, despite experiencing that it always belongs to Species A. The same divergence holds for the other three configurations $(100,010,001)$ : whereas all of those items actually belong to category B , treating features as class-conditionally independent entails that the probability for category A is higher (0.58).

Table 1: True environment vs. assuming classconditional independence (cci).

| Features | $\begin{array}{c}P \text { (features) } \\ \text { true env }\end{array}$ | $P$ (class $\mid$ features) |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| true env |  |  |  |  |$)$

The strongest discrepancy is for the 000 configuration, which is the second-most-frequent configuration, occurring with probability .28. Note that a hypothetical learner (even with perfect memory) who assumes class-conditional independence of features, and is unable to give up this assumption, will never learn the true statistical structure of this environment, even after completing a quadrillion learning trials.

Achieving criterion performance would also be impossible if learners looked at one feature only (at 1 xx , or x 1 x , or xx 1 and ignoring the x ). Considering single features, participants should think any feature configuration belongs to Species A
with probability 0.78 . This holds for attending solely to any of the three features.

## Hypotheses

If participants make no (not even tacit) assumptions of classconditional feature independence, and learn each item separately, then items could be learned in order of their frequency of occurrence (a frequency-of-configuration hypothesis). If participants approach the task by assuming features to be class-conditionally independent, classification decisions should systematically deviate from ones derived from the true environmental probabilities, especially early in learning (a posterior-discrepancy hypothesis).

Both hypotheses predict the fewest errors for item 111, the most frequent feature configuration and the one for which the class-conditional independence posterior is closest to accurate. For the four critical items, the difference in posterior probability is the largest for item 000. The posterior-discrepancy hypothesis predicts the most errors for item 000 , and thus that the ordering of errors should be $111<100 \approx 010 \approx 001<000$. However, the frequency-ofconfiguration hypothesis predicts that the ordering of classification errors should be $111<000<100 \approx 010 \approx 001$.

Key empirical questions are therefore whether there are any systematic differences in learning rate for the individual items, whether the early learning data suggest a presumption of class-conditional independence, and if so, whether the occurrence frequency of an item or the degree to which classconditional independence fails on it determine learning.

## Results and Discussion

All participants reached criterion performance, i.e. learned the category structure (in a mean number of 391 trials, $\mathrm{SD}=155, \mathrm{Md}=348$, range 210 to 808 trials). To reach criterion performance, participants needed to classify each individual feature configuration correctly five times in a row. To investigate whether there was a difference in learning speed for the different feature configurations, we calculated the number of times each item needed to be observed before reaching this criterion (Table 2). We will first consider learning time and then error rates.

Table 2: Number of trials an item needed to be seen to correctly classify it five times in a row.

| Features |  |  | Trials |  |  |  |
| :--- | :--- | :--- | :--- | :--- | ---: | ---: | ---: |
|  |  |  | mean | $S D(S E)$ | median |  |
| 1 | 1 | 1 |  | 10.4 | $10.7(1.9)$ | 7.0 |
| 1 | 0 | 0 |  | 11.4 | $8.0(2.1)$ | 7.5 |
| 0 | 1 | 0 |  | 11.5 | $7.7(2.1)$ | 9.0 |
| 0 | 0 | 1 |  | 11.5 | $7.0(2.1)$ | 9.0 |
| 0 | 0 | 0 |  | 15.8 | $11.5(2.9)$ | 13.5 |

In our data most subjects learned item 111 before item 000 ( 22 out of 30 , binomial $p<.02$ ), which is consistent with both hypotheses. Did learning time follow
the frequency-of-configuration hypothesis, or the posteriordiscrepancy hypothesis? The posterior-discrepancy hypothesis predicts an ordering of $111<100 \approx 010 \approx 001<000$, whereas the item-frequency hypothesis's ordering prediction is $111<000<100 \approx 010 \approx 001$. The critical difference in predictions is between the learning time for items 100,010 , and 001 and item 000 . The frequency hypothesis predicts that item 000 will be learned faster, whereas the posterior discrepancy hypothesis predicts that items 100, 010, and 001 will be learned first. Here, our results strongly support the posterior discrepancy hypothesis, and contradict the item frequency hypothesis. Items 100, 010 and 001 were learned more quickly by more people than item 000 , despite item 000 's greater frequency (item 001 faster: 21 out of 30 , binomial $p<.05$; item 010 faster: 20 out of 30 , binomial $p<.1$; item 100 faster: 21 out of 30 , binomial $p<.05$ ). Moreover, there was a nonsignificant trend for items 100,010 , and 001 to take longer than item 111; consistent with the posterior discrepancy hypothesis but not the configuration frequency hypothesis.


Figure 2: Percentage incorrect classifications for the first 50 trials each item was encountered.

The error rates throughout early learning are summarized in Figure 2. This figure corroborates the analysis of the number of learning trials required for each stimulus configuration: item 000 was clearly the most difficult to learn. As this feature configuration is the one for which the difference in posterior probability is largest when assuming class-conditional independence versus using the full true environmental probabilities, this finding is consistent with the idea that people treat features as being class-conditionally independent early in learning. However, items 100, 010 and 001 were much closer to (or even indistinguishable from) item 111, consistent with the above analysis in Table 2.

## General Discussion

The present paper examined the role of the assumption of class-conditional independence of features in category learning. While different types of conditional independence assumptions play an important role in various scientific debates and computational models of cognition, little is known about their descriptive validity in the context of classification learning with multiple cues. Our goal was to empirically investigate whether people initially (early in learning) treat features as class-conditionally independent. The present results partially support the idea that people initially treat features as class-conditionally independent and make classification decisions accordingly. We think of the results as tentative because some aspects of the data are not perfectly clear.

Our focus in the present study was on participants' behavior early in learning, when evidence about the category structure and environmental probabilities is limited. This approach is similar to the studies of Smith and Minda (1998), who investigated possible transitions in categorization strategies and stimulus encoding over the course of learning.Their finding was that late in learning exemplar models (e.g., Medin \& Schaffer, 1978) accounted best for subjects' behavior, but that this was not the case early in learning (in which a prototype model seemed to better account for human performance, see below). This is also a possible explanation for the finding that despite strongly violating class-conditional independence, the environment in our experiment was clearly learnable. Participants could have initially treated features as class-conditionally independent and computed posteriors accordingly and later shifted to an exemplar-based strategy to minimize classification error.

A key methodological aspect of our study was to use optimal experimental design principles to find environments that would allows us to directly test whether people use classconditional independence as a default assumption in categorization. Interestingly, the optimizations told us that the best environment to differentiate between a learner that assumes class-conditional independence and a learner that makes predictions based only on previous instances of the same feature configuration was deterministic. The crucial aspect of this environment, however, is not that it is deterministic, but that it entails a nonlinearly separable category structure. Since the class-conditional independence model induces a linear decision bound (Domingos \& Pazzani, 1997), it could not achieve criterion performance in this particular task environment.

This, in turn, relates our study to earlier research in psychology, which investigated whether linearly separable categories are easier to learn than nonlinearly separable ones (e.g., Medin \& Schaffer, 1978; Medin \& Schwanenflugel, 1981). This research focused on two types of categorization models, exemplar- and prototype-models, both of which assume that categorization decisions are derived from similarity comparisons (either to specific exemplars stored in memory or to prototypes of categories). By contrast, we investigated category learning and human subjects' initial assump-
tions from the perspective of probabilistic inference (see also Anderson, 1991; Fried \& Holyoak, 1984), a conceptually different view. Nevertheless, there are some interesting connections between our work and these earlier (similarity-based) models. For instance, assuming class-conditional independence entails that not all information (about feature configurations and corresponding class probability) is encoded during learning, but only marginalized conditional likelihoods and category base rates. In this respect the class-conditional independence model is similar to prototype models, which encode parametric information of central tendencies (e.g., mean or mode of feature values) that form the prototype (e.g., Smith \& Minda, 1998).

Importantly, these accounts assume that information is stored separately for each feature and the to-be-classified item is compared to the prototypes separately on each feature dimension individually. Conversely, a learner who makes no assumptions about the structure of the relations between classes and features and directly tracks the true environmental probabilities is conceptually more similar to exemplar models of category learning. The difference is that prototype models, like our independence model, do not need to store each individual instance that is experienced.

In sum, the current paper adds to the debate about the role of conditional independence assumptions for computational models of cognition. The task environment identified based on optimal experimental design principles allowed us to directly examine the descriptive validity of this assumption in category learning. Here, we do find evidence consistent with its use.

## Acknowledgments

This research was supported by grants NE 1713/1 to JDN and ME 3717/2 to BM, from the Deutsche Forschungsgemeinschaft (DFG) as part of the priority program "New Frameworks of Rationality" (SPP 1516). We would like to thank Gregor Caregnato for data collection and Laura Martignon, Michael Waldmann, Ralf Mayrhofer, and the reviewers for their helpful comments. We also thank Jorge Rey and Sheila O'Connell (University of Florida, FMELe) for allowing us to base our artificial plankton stimuli on their copepod photographs.

## References

Anderson, J. R. (1991). The adaptive nature of human categorization. Psychological Review, 3, 409-429.
Domingos, P. \& Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning, 29, 103-130.
Fried, L. S. \& Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 2, 234-257.
Hertwig, R., Barron, G., Weber, E., \& Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. Psychological Science, 15, 534-539.

Knowlton, B., Squire, L., \& Gluck, M. (1994). Probabilistic classification learning in amnesia. Learning \& Memory, 1 , 106-120.
Mayrhofer, R., Hagmayer, Y., \& Waldmann, M. (2010). Agents and causes: A Bayesian error attribution model of causal reasoning. In Proceedings of the 32nd Annual Conference of the Cognitive Science Society.
Meder, B. \& Nelson, J. D. (2012). Information search with situation-specific reward functions. Judgment and Decision Making, 7, 119-148.
Medin, D. L. \& Schaffer, M. M. (1978). Context theory of classification learning. Psychological Review, 3, 207-238.
Medin, D. L. \& Schwanenflugel, P. (1981). Linear separability in classification learning. Journal of Experimental Psychology: Human Learning and Memory, 355-368.
Movellan, J. R. \& McClelland, J. L. (2001). The MortonMassaro law of information integration: Implications for models of perception. Psychological Review, 108, 113148.

Myung, J. \& Pitt, M. (2009). Optimal experimental design for model discrimination. Psychological Review, 116, 832840.

Nelson, J. D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. Psychological Review, 112, 979-999.
Nelson, J. D., McKenzie, C. R. M., Cottrell, G. W., \& Sejnowski, T. J. (2010). Experience matters. Psychological Science, 21, 960-969.
Pearl, J. (2000). Causality. Models, Reasoning and Inference. New York: Cambridge University Press.
Rehder, B. \& Burnett, R. (2005). Feature inference and the causal structure of categories. Cognitive Psychology, 50, 264-314.
Rehder, B. \& Hoffman, A. (2005). Eyetracking and selective attention in category learning. Cognitive Psychology, 51, 1-41.
Reichenbach, H. (1956). The Direction of Time. Berkeley: University of California Press.
Smith, J. D. \& Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 24, 14111436.

Spirtes, P., Glymour, C., \& Scheines, R. (1993). Causation, Prediction, and Search. New York: Springer.
Von Sydow, M., Meder, B., \& Hagmayer, Y. (2009). A transitivity heuristic of probabilistic causal reasoning. In Proceedings of the 31st Annual Conference of the Cognitive Science Society.
Walsh, C. R. \& Sloman, S. A. (2007). Updating beliefs with causal models: Violations of screening off. In J. R. A. M. A. Gluck \& S. M. Kosslyn (Eds.), Memory and Mind: A Festschrift for Gordon H. Bower.

# Similarity, causality and argumentation 

William Jiménez-Leal (w.jimenezleal@uniandes.edu.co)<br>Christian Gaviria (cgaviria@uniandes.edu.co)<br>Departamento de Psicología, Universidad de los Andes. Cra. $1 \mathrm{~N}^{\circ}$ 18A- 12, Edificio Franco, Bogotá, 111711. Colombia


#### Abstract

Similarity is a notion that is widely used both in cognitive science and in argumentation theory. These research programs have, however, developed in large part separately and in consequence rely on disparate notions of similarity. Only recently there has been a proposal for specifying how similarity actually plays a role in judging slippery slope arguments. We present here further theoretical discussion and empirical evidence in order to show how similarity can play a role in slippery slope arguments and in argumentation in general. In the experiment presented here, we manipulated the availability of causal information, and showed that people are sensitive to it when judging arguments' strength. We conclude that similarity between causal properties of the elements presented in arguments is crucial for arguments' strength assessments.


Keywords: Argumentation, similarity, causality, analogical reasoning.

## Introduction

The degree of conviction that an argument generates depends on many elements. The effectiveness of some arguments seems to depend on the perceived similarity between the elements presented in the premises and the conclusions that might follow. For example, such is the case of the argument based on precedent, where the similarity between past events and the one under discussion is such as to warrant following the same course of action as with the precedent (Walton, 2010). Similarly, some arguments fail because the relation of similarity between premises and conclusion is weak. The fallacy of false analogy (Tindale, 2007) is one example, where there is a comparison between situations based on superficial similarities that do not support the conclusion. Walton, Reed \& Macagno (2008) recognize that judgements of similarity between a class and an exemplar are key for the quality of arguments from verbal classification (from definition, vagueness, arbitrariness) (See also Macagno, 2009; Walton, 2009). The notion of similarity is thus central to explaining why people deem some arguments good or bad, and it is taken as a primitive element for explaining how people evaluate arguments.

Similarly, the typologies of arguments put forward by perspectives like the dialectical (Walton, 2010) and the pragma-dialectical (van Eemeren, Houtlosser, \& Snoeck, 2007), rely on identification of similarity. In the pragmadialectical perspective, for example, one of the three main types of arguments is the 'argumentation based on comparison" (van Eemeren, et al., 2007), where the argument and the standpoint argued for refer to different things but share a predicate. In the example "It is not at all
necessary to give James a 10 dollar allowance, because his brother always got 10 dollars a week", the similarity between James and his brother regarding the money needed, is the justification that allows one to proceed from premise to conclusion (Hitchcock \& Wagenmans, 2011). In fact, the questions proposed to identify this type of argumentation scheme presuppose the notion of similarity (e.g. "Are there enough relevant similarities in the things that are compared?")

Similarity thus plays a dual role in argumentation: not only is it proposed that similarity judgments are performed by people engaged in argumentation, but it is also suggested that argumentation schemes are to be identified by questions that imply similarity judgements. That is, similarity plays a role both in explaining what people do, and also as a tool that the argumentation scholar needs to identify arguments and evaluate its correctness.

Even though there has been vigorous research on the role of similarity in several psychological processes (Goldstone \& Son, 2005), and despite argumentation research consistently using this construct as a tool to characterise several argumentation schemes (Walton et al, 2008), little work has been done to integrate the findings of cognitive science into our understanding of how people reason with arguments. In what follows, we will briefly examine the most common notions of similarity currently in use in cognitive science and consider the only work we are aware of that explicitly makes use of this idea to explain argument strength (Corner, Hahn \& Oaksford, 2011). This will lead us to consider causality as one of the key ideas that is missing when using similarity as an explanatory principle. We will then present some empirical evidence to support our claims.

## Similarity and cognitive science

It is difficult to overstate the importance of similarity as an explanatory tool in cognitive science. From categorisation to analogy, similarity judgements are advanced to explain very diverse phenomena. Links between rules and similarity as well as the very need of appealing to similarity in explaining cognition have been widely discussed (Sloman \& Rips, 1998; Goldstone, Day \& Son, 2010). It is more or less accepted that alternative ways of conceiving similarity capture different intuitions about our use of this notion, and that all have different weaknesses and strengths.

The multiplicity of contexts in which it is possible to use the notion of similarity is consistent with the diversity of ways in which people judge that objects are alike. One can distinguish three main models to conceive similarity: geometric models, featural models and alignment based
models. ${ }^{1}$ While none of these models can possibly capture the flexibility of similarity, they offer important insights into how similarity can be possibly used in the context of argumentation.

Geometric models are based on multidimensional scaling of similarity and dissimilarity judgments provided by participants. People judge how alike two objects are, and their ratings are used to generate a set of points organised in a metric space. The similarity of two objects is an inverse function of the distance between points that represent the objects. The distances measured depend on the number of dimensions inferred (Goldstone el at., 2010). Certainly, a geometrical representation of similarity seems to be at play in the case of induction of blank properties, as proposed by the similarity-coverage model of induction (Osherson et al. 1990). Knowing that "bears require Biotin for haemoglobin synthesis" makes one more likely to believe that wolves require that substance when compared to whales. In this case, the induction is possibly supported by a similarity judgement along the dimensions of "animal with fur", and "lives in the woods".

Notice that this conception of similarity relies on the idea of objects represented by dimensions, which can be adequately captured by classification tasks. The problem with this idea is that it makes geometric models too heavily committed to the assumptions of minimality, symmetry and the triangle inequality, as pointed out by Tversky (1977). These assumptions make the model psychologically implausible for some similarity judgments (e.g. asymmetric judgments like "Korea is more similar to China than China is to Korea"). Featural models capitalise on some problems of geometric models to advance a notion of similarity based on weighted feature-matching. Here the objects are characterised as a set of features, and resemblance is established by some linear combination of shared and distinctive features, with their respective weights. Featural models have found success particularly in explaining categorisation tasks (Verguts et al, 2004). Typically, these tasks involve a learning phase where participants are presented several exemplars that belong to an artificial category (e.g. Flowers whose colour, number of petals and size can vary). Participants are then tested with new exemplars, whose characteristics may match the ones presented in the study phase.

Both geometric and featural representations of similarity have traditionally been used to analyse tasks with relatively unstructured inputs. The link between the inferred dimensions or features had traditionally been overlooked. In response to this problem, and inspired by research in analogy and metaphor, Gentner and Markman (1995) proposed the idea of having situations as the input of the comparison process, and thus starting with complex inputs. Similarity between objects is in this case derived from the

[^304]role the object fulfils in the scene. This principle guides the selection of characteristics relevant for the comparison process. Alignment-based models assume that similarity comparisons involve a mechanism of structure-mapping, called structural alignment, that seeks maximal structurally consistent matches. When maximizing these matches, there is a set of matched characteristics and two sets of differences, alignable and non-alignable. The latter are key to establishing similarity and explaining the effects of asymmetry and minimality identified by Tversky.

While these approaches to similarity have met different degrees of success in explaining phenomena like categorisation and metaphor, little has been done to specify the particular approach at play when turning to the idea of similarity in the context of explaining argumentation. Thus we now consider the extant literature about argumentation.

## Similarity in argumentation

To our knowledge, there are only two explicit proposals for using similarity as an explanatory tool in argumentation. The first one is attributed to Walton (2010, 2012), who presents an analysis showing how arguments from precedent are based on arguments from analogy and classification. The second one is attributed to Corner, Hahn and Oaksford (2011), who, in the framework of the Bayesian approach to argumentation (Hahn \& Oaksford, 2007), claim that the mechanism underlying the slippery slope arguments (SSAs) consists of a category boundary re-appraisal process, which in turn depends on the perceived similarity between an exemplar and a category. We now consider each one in turn.

In law, arguments from precedents involve applying an earlier decision to a later case deemed to be the same. Of course, the issue at stake here is when two cases can be considered the "same" in light of the precedent. By the same token, in arguments from analogy a decision is suggested because the case is similar to another one, where "how similar" is the critical question. Walton (2010) is interested in finding an objective way "to identify, analyse and evaluate arguments from analogy" (p. 217), and proposes that arguments from precedent are a special case of arguments from analogy, which in turn are cases of arguments by classification or definition. Given Walton's interest in legal reasoning, the inputs of the process are "cases", complex situations that afford comparisons at multiple levels. Comparisons are only possible if there is a "plausible story" that connects the cases being compared.

The mechanism proposed to establish similarity is an abstract structure called a "story scheme" (Bex, 2009), which is a template that contains a connected sequence of events or actions represented by variables, so that different stories can be represented as instances of it (Walton, 2012). Once the right story scheme is selected, it is possible to establish if the case argued for is an instance of the story scheme. For example, the argument that selling unhealthy food is analogous to selling a malfunctioning car, and thus the same controls should be implemented for food, is only possible in a story scheme that can encompass both food
and cars. In other words, to judge the quality of the argument, it is necessary to decide whether the coverage of the story scheme is to be extended to include the new case.

The methodology proposed by Walton is, ultimately, a combination of story schemes and the argumentation scheme from analogy. As such, it is a tool used by the argumentation analyst and it is not intended to have psychological reality. However, its use does imply a functioning cognitive system able to comprehend similarity. As such, the judgements of similarity from story schemes implicitly rely on an alignment-based approach.

The second case is the work of Corner et al. (2011), whose goal is to provide the objective basis for judging SSAs. SSAs are arguments where a proposal is put forward but its consequences are thought to be undesirable, so that if the proposal is allowed, the undesirable consequence will unavoidably follow (e.g. "if freedom of speech is refused to extremist groups, then there will be censorship to any kind of political expression") (Volokh, 2003). Corner et al. present evidence on how the strength of slippery slope arguments is related to the perceived similarity of the elements present in the premises of the argument. They propose that the mechanism underlying the judged strength of slippery slope arguments is the assessment of similarity between the exemplars presented in their premise and conclusion.

More specifically, Corner et al. claim that when assessing a SSA there is a process of category boundary re-appraisal (Corner et al., 2011), and thus the exemplar under discussion can be rightly considered within the scope of the category discussed. Consequently, people are more willing to accept arguments of this form when the similarity between the elements presented in the premises is high, and otherwise consider the argument fallacious. How good an SSA is, depends on extending the category boundaries to include the case under discussion. For example, the acceptability of the argument "If voluntary euthanasia is legalised, then in the future there will be more cases of 'medical murder" (Corner et al, 2011, p. 133), depends on being able to redraw the limits of the category 'medical murder' to include 'euthanasia'. Their claims are based on extensive findings from work on exemplars' effects on categorisation (Nosofsky, 1986).

This idea is certainly a step forward in integrating research in cognitive science and argumentation. However, the generality of Corner et al's proposal is lessened when considering the materials used in their experiments.

In their experiments $2,2 \mathrm{a}$ and 3 , they use numerically defined exemplars. They describe a situation where there is a discussion about the inclusion of a new territory in the category of places of outstanding beauty (PONB). Participants were presented with cases of areas that were either declared PONB or not, including the number of animal species in each place, as the decisive criterion. For example, they were told that location A (114 species) and location B (149 species) were not considered eligible as PONB, whereas location C (224 species) and D (259
species) were. In the testing phase, which corresponds with the SSA, a new pair of exemplars was presented in terms of two conditions: similar (194 and 179 species) and dissimilar ( 218 vs. 179 species). If the mechanism of SSA is an instance of category boundary re-appraisal, it would be expected that (1) an argument involving a comparison of items should be better evaluated when they are similar; and (2) that the results of a categorisation task would support this prediction. They did in fact find a good match between categorisation decisions and the strength of SSAs.

The similarity of the cases considered in Corner et al (2011) depends on numerical thresholds (e.g. number of species in a natural park necessary to declare it a PONB (exp. 2); number of years of imprisonment for knife/gun crime (exp. 3)), given by the fact of dealing with numerically defined categories with only one dimension. As such, their proposal suggests at least one question; namely, will the link between similarity and SSAs show up in cases where the similarity metric depends on more than one dimension (or no dimensions at all: features, stories, etc.)? The next logical step is then to examine the functioning of the hypothesized mechanism in the cases suggested.

The common theme in the work of Walton and Corner et al. is the idea of a more basic mechanism at work when dealing with arguments: Walton, at the level of the scholar of identifying and analysing argumentation schemes; Corner et al, at the cognitive level of individuals faced with arguments. We believe the latter is a particularly promising avenue of research since it relies on the accumulated knowledge of cognitive science about similarity and promotes the integration of cognitive science and argumentation theory (Hornikx \& Hahn, 2012). However, as has been acknowledged, the evidence presented by Corner et al. is limited to cases where similarity judgments operate within a dimension, suffering, in consequence, from one of the main criticisms put forward against geometric approaches to similarity, that is, overlooking the connection between the judged dimension and other aspects of the objects under consideration.

## Causality and similarity

We believe there is a complementary way of conceiving similarity in the context of argumentation that comes from the literature on causal categorisation. This literature offers a way of dealing with the dichotomy between dimensions and features, and also accounts for the fact that features are usually correlated in exemplars.

There is ample evidence of causal effects on similarity assessments in the categorisation literature (Rehder, 2003). According to the causal model of categorisation, the observed correlation between exemplars' features are understood as evidence of an underlying mechanism at work, resulting in those features (Rehder \& Burnett, 2005). Whereas Rehder (2006) considers causality and similarity as two independent sources of information, it is possible to interpret categorisation as cases that depend on similarity judgments inspired by causality. Similarity is not a fixed
notion, as noted above, and proposing that causal information determines our perception of similarity amounts to saying that features and dimensions that enter comparisons are governed by a more general principle. In fact, causal-based models can actually provide a way to solve the apparent opposition of models based in either features or dimensions. Kemp, Shafto, Berke and Tenenbaum (2008) propose a causal model that integrates both kinds of knowledge, relations between objects and relation between features. The evidence accumulated recently in favour of the causal models of categorisation and induction gives good grounds for suggesting that the similarity assessment at work in the case of argumentation depends on the perceived causal similarity, in at least some relevant cases.

This would lead one to consider causal-based similarity judgements as the mechanism underlying some argumentation schemes, which can be characterised by the inclusion of a new exemplar under the scope of the category. Some of these argument types have been suggested above: analogy, precedent, classification, definition and slippery slope. Similarly, some forms of the SSA could be considered special cases of causal similaritybased argumentation, where how slippery a slope is, depends on the causal links shared by the exemplars presented in the premises of the argument, as the evidence of their features provide. ${ }^{2}$

Here we do not commit to a particular model of causal reasoning, only to the idea underlying causal-based models of categorisation. However, our proposal has the general appeal of using the logic of weighted feature-matching, where the weights are assigned following a psychological principle, viz. causal representation. In consequence, the strength of the arguments that depend on this mechanism can be predicted by establishing what the particular causal mechanism at work is.

## The present experiment

As a first attempt to test this idea, we designed an experiment where the presence of the causal efficient feature was manipulated as well as the overall similarity (number of matched features), in the context of a slippery slope argument. This is a 3 (number of matched features) X 2 (presence/absence of a causally relevant feature), between-subjects design. We expect to see a main effect of the causally relevant feature, regardless of the overall similarity indicated by simple feature matching. In consequence, arguments based on causally matched information will be judged stronger.

[^305]
## Experiment

## Participants

132 university students ( 77 female) with ages between 18 and $34(\mathrm{~m}=21.25, \mathrm{sd}=3.36)$ took part in this study. The students came from several different undergraduate programs. Participants were randomly allocated to one of the six possible combinations, with 65 and 67 participants in the causal condition and non-causal condition, respectively, and 41,42 and 39 for each one of the groups defined by the number of matched features.

## Materials and procedure

Participants were tested in groups at the end of one of their classes. Each participant was provided with a fourpage booklet containing, in the first page, some general instructions, in the following two pages the main task, and in the last page participants were requested to provide basic demographic information.

The main task involved making judgments relative to two scenarios (drugs and fertilizers). The first part of each scenario described a situation where a government agency had to decide whether to allow the use of a new substance (drug/fertilizer). The second part of the scenario included a table comparing the features of a banned substance and the corresponding characteristics of the new substance under consideration. Each table had four items, where the number of matched characteristics $(1,2$ or 3$)$ and the presence of the key causal feature (matched or not) were manipulated. For example, table 1 shows the information presented in the fertilizer scenario, with two matched features in the noncausal condition. Polenoy is the currently banned fertilizer, and Soilex the fertilizer the government is considering whether to allow. In this case, the "high concentration of nitric acid" was the key causal feature.

Causal features were selected from ratings provided by an independent group of 20 participants who selected what characteristic they considered more important for banning/allowing fertilizers and potentially addictive substances. Agreement on the most causal feature for the fertilizer scenario was $100 \%$, and in the drugs scenario was $80 \%$.

Table 1: Information presented in a sample item

|  | POLENOY | SOILEX |
| :--- | :---: | :---: |
| Doses lower than 50kg per hectare | YES | YES |
| Highly soluble in water | YES | NO |
| Delivered with sprinklers | YES | NO |
| High concentration of nitric acid | YES | YES |

An argument was then presented claiming that the new substance should not be allowed, because allowing it would inevitably lead to removing the ban on the former substance too ("If we allow Soilex now, we are going to have to allow Polenoy. In consequence, we should not allow Soilex"). In both scenarios the arguments were uttered by fictitious characters in positions of power. Participants were asked to rate how convincing each argument was on a 10-point scale. All of the participants rated both scenarios, with the presentation order counterbalanced. Finally, participants rated how negative/positive (on a scale of 1 to 10) they
considered the consequence stated in the conclusion of each argument to be, as a measure of the perceived utility.

## Results

Results of each scenario were submitted to a 2 (presence/absence of a causally relevant feature) X 3 (number of matched items: 1, or 3) between subjects Anova. Results for each scenario are considered separately because their respective utility ratings differed (paired $t(131)=4.56$, $\mathrm{p}<0.001$ ), even though there was no significant difference between their acceptance ratings (paired $t(131)=0.6, p=.5$ )
For both items, there was a main effect of causal information. People rated the argument with the matched relevant causal feature as more convincing than the case without the matched feature ( 5.8 vs. 3.7 for the drugs scenario and 5.3 vs. 4 for the fertilizer scenario)(see table 2). The differences are statistically significant in both cases, F $(2,128)=19.95, \mathrm{p}<0.01, \eta^{2}=0.03 ; \mathrm{F}(2,128)=3.92, \mathrm{p}<0.05$, $\eta^{2}=0.18$.

Table 2: Summary of argument strength ratings by causal information and number of matched items

| Scenario | Causal Info | \# of matched features | Mean (s.d) |
| :---: | :---: | :---: | :---: |
| Drugs | Yes | 1 | 5.29 (3.15) |
|  |  | 2 | 6.14 (2.41) |
|  |  | 3 | 6.15 (3.26) |
|  |  | Total | 5.76 (2.91) |
|  | No | 1 | 3.03 (2.56) |
|  |  | 2 | 5.17 (2.71) |
|  |  | 3 | 3.40 (1.64) |
|  |  | Total | 3.69 (2.51) |
| Fertilizers | Yes | 1 | 4.91 (3.11) |
|  |  | 2 | 5.72 (2.61) |
|  |  | 3 | 5.35 (1.98) |
|  |  | Total | 5.24 (2.69) |
|  | No | 1 | 4.50 (2.87) |
|  |  | 2 | 5.76 (3.10) |
|  |  | 3 | 2.15 (1.82) |
|  |  | Total | 4.14 (2.99) |

There was also a main effect of the number of matched features $\left(F(2,126)=3.99, p<0.05, \eta^{2}=0.05 ; F(2,126)=3.58\right.$, $\mathrm{p}<0.01, \eta^{2}=0.05$, for fertilizer and drugs, respectively). The degree of persuasion that an argument exerted changed with the number of matched features for both scenarios. Post hoc tests (Tukey) revealed that the mean acceptance rating was lower when having only one matched feature (the causal characteristic), compared to two or three matches. However, there is no consistent pattern of differences across scenarios when having more than one matched feature.

Interaction was significant for the fertilizer scenario $(\mathrm{F}(2$, $128)=3.62, \mathrm{p}<0.05, \eta^{2}=0.02$ ) but not for drugs item $(\mathrm{F}=1.1)$. In the fertilizers scenario, the difference between the causal and non-causal condition was larger when having only one feature matched. Maybe in this scenario having a single feature matched was more salient.

Utility ratings were significantly higher for fertilizer than for drugs ( 4.5 vs. 37 , paired $\mathrm{t}(131)=4.56, \mathrm{p}<0.01$ ), which means people were more in agreement with fertilizer use. However, the utility ratings were not correlated with acceptance ratings in either scenario ( 0.04 and 0.002 ) and did not differ as a function of the inclusion of causal information or matched items ( F 's $<1$ for all anovas).

In short, both items were rated in the predicted direction, with the more convincing arguments being those that have a causally relevant matched feature to the sample item. However, the number of items does not have a clear effect. Increasing the number of matches is not linearly associated with higher argument acceptance ratings, but adding a matched feature does have an effect. The present experiment does not support a firm conclusion about this aspect.

## Discussion

People are sensitive to causal information in the assessment of argument strength in SSA. In the case of the scenarios used in this experiment, people recognise the causal feature (e.g. concentration of nitric acid) as the key characteristic that produces the undesirable consequence and thus determines the acceptability of the SSA. This experiment adds support to Corner et al's proposal of category reappraisal as the mechanism at work in SSA, and also sets it in the larger context of the use of causal information for categorisation. A potential problem with the interpretation of the data presented is that is not possible to discard that the other features presented were also interpreted as causal. This would explain why adding a matched feature was associated to higher acceptance ratings. Even in this case, it would still stand that matching the most causally efficient feature is related with a significant increase in the acceptability of the argument.

## Summary and conclusions

Similarity clearly plays a role in argument evaluation. Here we have presented evidence of a particular way in which this can occur. The experiment presented shows that people are sensitive to causal information when judging how similar a new exemplar is to a known class. This finding complements Corner et al's (2011) work, by further specifying the mechanism at play, beyond the case of a geometrical notion of similarity. SSAs often imply an evaluation of how "inevitable" an undesirable consequence is once the proposed action has been effected, and in that sense, causal knowledge linking the elements in the argument is especially relevant.

Taking into account research in cognitive science on similarity has several benefits. First, it makes it possible to take further steps in the integration of dissimilar perspectives in the study of argumentation (Hornikx \& Hahn, 2012). The different ways in which different evaluations of similarity can play a role in argumentation are still unexplored. Second, it can help predict cases where arguments may be considered bad or fallacious, by providing an understanding of similarity ratings of the
elements under discussion. The reverse is also true: by examining the way people assess arguments, it might be possible to examine the conceptual representation of the world, and the causal structure implied. The actual scope of the causal similarity mechanism is a matter of empirical research. For example, the perceived strength of ad hominem arguments, such as the ad hitlerum (Harris et al, 2012), where adopting policy X would lead to the adoption of other undesirable policies, might depend on the similarity of the causally relevant links that connect policy X with other undesirable policies. Third, the study of causal similarity from a cognitive perspective can potentially provide a unifying theme to the study of the argument typologies proposed in informal logic. The dialectical (Walton, 2010) and pragma dialectical (van Eemeren, Houtlosser, \& Snoeck, 2007) approaches propose typologies that, useful as they are for the study of argumentation, might conceal important unifying psychologically themes in the evaluation of arguments. One of them, as suggested in this paper, is the use of categorical causal information.

There are several other questions that can be explored using judged causal similarity as a framework. For example, it is clear that complex situations require the rapid evaluation for alignable matches and mismatches (Gentner \& Markman, 1995). Are alignable differences more important when they refer to causal characteristics? The way similarity is assessed, in the absence of specific characteristics to be matched (cf. geometrical models), will probably have a differential impact on argument strength, when compared to cases where the exemplars are fully specified by a set of characteristics. A causal-based model of categorisation offers a wealth of hypotheses to be investigated.

## Acknowledgments

William Jiménez-Leal was funded by the FAPA grant at Universidad de los Andes.

## References

Bex, F. (2009). "Analysing Stories Using Schemes." In H. Kaptein, H. Prakken and B. Verheij (Eds.),Legal Evidence and Proof: Statistics, Stories, Logic. Farnham: Ashgate.
Corner, A., Hahn, U., \& Oaksford, M. (2011). The psychological mechanism of the slippery slope argument. Journal of Memory and Language, 64, 2, 133-152.
van Eemeren, F. H., Houtlosser, P., \& Snoeck, H. A. F. (2007). Argumentative indicators in discourse: A pragma-dialectical study. Dordrecht, Netherlands: Springer.
Gentner, D., \& Markman, A. B. (1995) . Similarity is like analogy : Structural alignment in comparison. En C. Cacciari (Ed.), Similarity in language, thought and perception. Brussels: BREPOLS.
Goldstone, R. L., Day, S., \& Son, J. Y. (2010). Comparison. In B. Glatzeder, V. Goel, \& A. von Müller (Eds.) On Thinking: Volume II, towards a theory of thinking. Heidelberg, Germany: Springer Verlag GmbH.

Hahn, U., \& Oaksford, M. (2007). The Rationality of Informal Argumentation: A Bayesian Approach to Reasoning Fallacies. Psychological Review, 114, 3, 704-732.
Harris, A. J. L., Hsu, A. S., \& Madsen, J. K. (2012). Because Hitler did it! Quantitative tests of Bayesian argumentation using ad hominem. Thinking and Reasoning, 18 (3), 311343.

Hitchcock, D. \& Wagenmans, J. (2011) The pragmadialectical account of argument schemes. In van Eemeren, F. H., Feteris, E. T., Garssen, B., \& Snoeck, H. F. (Eds.) Keeping in touch with pragma-dialectics: In honor of Frans H. van Eemeren. Amsterdam: John Benjamins Pub. Co.

Hornikx, J., \& Hahn, U. (2012). Reasoning and argumentation: Towards an integrated psychology of argumentation. Thinking and Reasoning, 18 (3), 225-243
Kemp, C., Shafto, P., Berke, A. \& Tenenbaum, J. B. (2007). Combining causal and similarity-based reasoning. Advances in Neural Information Processing Systems. 19.
Macagno, F, (2009) Reasoning from Classification and Definition. Argumentation, Vol. 23, pp. 81-107,
Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. Journal of Experimental Psychology. General, 115 (1), 39-61
Osherson, D. N, Smith, E., Wilkie, O., López, A., Shafir, E. (1990). Category-Based Induction. Psychological Review, 97 (2), 185-200.

Rehder, B. (2003). Categorization as causal reasoning. Cognitive Science, 27, 5, 709-748.
Rehder, B. (2006). When similarity and causality compete in category-based property generalization. Memory \& Cognition, 34, 1, 3-16.
Rehder, B. (2009). Causal-based property generalization. Cognitive Science, 33, 3, 301-344.
Rehder, B \& Burnett, R. (2005) Feature inference and the causal structure of categories. Cognitive Science, 50: 264314, 2005.
Sloman, S. A., \& Rips, L. J. (1998). Similarity and symbols in human thinking. Cambridge, Mass: MIT Press.
Tindale, C. W. (2007). Fallacies and argument appraisal. Cambridge: Cambridge University Press.
Tversky, A. (1977) Features of Similarity. Psychological Review, 84(4), Jul 1977, 327-352.
Verguts, T., Ameel, E., \& Storms, G. (2004). Measures of similarity in models of categorization. Memory \& Cognition, 32, 3, 379-389.
Volokh, E. (2003). The Mechanisms of the Slippery Slope. Harvard Law Review, 116, 1026-1138.
Walton, D. N. (1992). Slippery slope arguments. Oxford England: Clarendon Press.
Walton, D.N.(2008) Arguing from Definition to Verbal Classification: The Case of Redefining 'Planet' to Exclude Pluto. Informal Logic, Vol. 28, No. 2 (2008), pp. 129-154.
Walton, D. N. (2010). Similarity, precedent and argument from analogy. Artificial Intelligence and Law, 18 (3), 217-246.
Walton, D.N. (2012) Story Similarity in Arguments from Analogy. Informal Logic, 32(2), 199-218.
Walton, D. N., Reed, C., \& Macagno, F. (2008). Argumentation schemes. Cambridge: Cambridge University Press.

# Good Decisions, Good Causes: Optimality as a Constraint on Attribution of Causal Responsibility 

Samuel G. B. Johnson (samuel.johnson@yale.edu)<br>Department of Psychology, Yale University<br>2 Hillhouse Ave., New Haven, CT 06520 USA

Lance J. Rips (rips@northwestern.edu)<br>Department of Psychology, Northwestern University<br>2029 Sheridan Road, Evanston, IL 60208 USA


#### Abstract

How do we assign causal responsibility for others' decisions? The present experiments examine the possibility that an optimality constraint is used in these attributions, with agents considered less responsible for outcomes when the decisions that led to those outcomes were suboptimal. Our first two experiments investigate scenarios in which agents are choosing among multiple options, varying the efficacy of the forsaken alternatives to examine the role of optimality in attributing responsibility. Experiment 3 tests whether optimality considerations also play a role in attribution of causality more generally. Taken together, these studies indicate that optimality constraints are used in lay decision theory and in causal judgment.


Keywords: Causal attribution; decision-making; theory of mind; responsibility; lay decision theory.

## Introduction

Many of the decisions we make on a daily basis are thoroughly mediocre. This conclusion has been the joint product of the philosophical discipline of Decision Theory, which aims to characterize the decisions we ought to make given our knowledge and priorities (e.g., Jeffrey, 1965), and the psychological discipline of Judgment and Decision-Making (JDM), which aims to characterize actual decision-making behavior (e.g., Tversky \& Kahneman, 1974). JDM research has shown that normative decision theory largely fails as a descriptive theory of human decision-making, documenting a plethora of ways in which our actual decision-making does not live up to normative standards.

Less is known, however, about how people conceptualize and evaluate the decisions of others-a research question one might term lay decision theory. The present research begins to examine this question, investigating how people assign causal responsibility to agents for the outcomes of their decisions.

In these experiments, we consider situations in which an agent made a decision that led to an outcome with probability $\mathrm{P}_{\mathrm{ACT}}$ (always $50 \%$ ), but could have made an alternative decision that would have led to that outcome with probability $\mathrm{P}_{\mathrm{ALT}}$ (which was varied between $10 \%$ and $90 \%$ across conditions). For example:

Angie has a shrub, and wants the shrub's flowers to turn red. She is considering two brands of fertilizer to apply:
If she applies Formula PTY, there is a $50 \%$ chance that the flowers will turn red.
If she applies Formula NRW, there is a $10 \%$ chance that the flowers will turn red.
Angie chooses Formula PTY, and the flowers turn red. Thus, Angie's actual choice of Formula PTY led to the outcome with probability $\mathrm{P}_{\mathrm{ACT}}=50 \%$, and her alternative choice of Formula NRW led to the outcome with $\mathrm{P}_{\text {ALT }}=$ $10 \%$. Because $\mathrm{P}_{\text {ACT }}>\mathrm{P}_{\text {ALT }}$, Angie's choice was optimal. However, if Formula NRW had led to the outcome with $\mathrm{P}_{\mathrm{ALT}}=90 \%$, then $\mathrm{P}_{\mathrm{ALT}}>\mathrm{P}_{\mathrm{ACT}}$, and Angie's choice would have been suboptimal. Finally, if $\mathrm{P}_{\text {ALT }}$ had been $50 \%$, then $\mathrm{P}_{\mathrm{ACT}}=\mathrm{P}_{\mathrm{ALT}}$, and there would have been no uniquely optimal decision.

There are at least three possible predictions one could make about how judgments of Angie's responsibility for the outcome would vary as a function of the counterfactual alternatives that Angie forsook.

First, according to the Counterfactual Stability view, counterfactual alternatives are irrelevant to computing causal responsibility in situations like Angie's. This seems plausible, since $\mathrm{P}_{\mathrm{ACT}}$ is fixed across all conditions at $50 \%$, and the likelihood of goal completion does not depend on the efficacy of the alternative. Moreover, it is stipulated in the vignette that the agent achieves her goal, eliminating uncertainty about the outcome. This result would be obtained if people are permissive of suboptimal decision-making when computing causal responsibility. In actual decision-making practice, after all, computational limitations prevent us from analyzing every possible course of action, so we often settle for an option that is satisfactory even if not optimal (Simon, 1956).

Second, according to the Difference-Making view, judgments of responsibility are a linear function of the difference made by the actual choice, relative to the alternative choice. On this view, responsibility judgments would be proportional to $\left[\mathrm{P}_{\mathrm{ACT}}-\mathrm{P}_{\mathrm{ALT}}\right]$, known as $\Delta \mathrm{P}$ in the causal learning literature. On this view, one is most responsible for an outcome when the quality of the alternative options is very low, because the choice made a
large difference, while one is viewed as less responsible as the size of this gap decreases. For suboptimal choices (i.e., when $\Delta \mathrm{P}<0$ ), one's responsibility could further decrease (or be viewed as preventive) as the forsaken alternatives become increasingly efficacious and $\Delta \mathrm{P}$ becomes increasingly negative. Though prior research on causal attribution (e.g., Cheng \& Novick, 1992; Spellman, 1997) does not directly predict this result, this pattern would be most consistent with those previous findings.

Finally, according to the Optimality view, judgments of responsibility would be higher for optimal decisions than for suboptimal decisions, without consideration for how much better that decision is, compared to its alternatives. As we know humans to be satisficers and heuristic decision-makers, it may seem unlikely on the surface that we should require optimal behavior from others when assigning causal responsibility. However, theoretical considerations make this view seem less far-fetched.

Dennett (1987; see also Davidson, 1967) proposed that mental state inferences can often be accomplished by invoking a well-formedness rule called the Principle of Rationality. Just as we can solve the equation ' $X+Y=Z$ ' if given the values of two of the three variables, so can we make inferences about agents' actions, goals, and situational constraints by using the principle that agents act rationally to satisfy their goals, given situational constraints. These inputs to the rationality 'formula' can be either states of the world (when reasoning teleologically), or mental states (i.e., beliefs, desires, and intentions, when reasoning mentalistically). In either case, the Principle of Rationality produces a unique prediction for one element given the other two, just as the facts of arithmetic yield a unique solution for ' $2+Y=5$ '. Actions conforming to the Principle of Rationality are optimal, relative to the agent's goals and situational constraints.

Previous research has shown that both adults and infants often make inferences afforded by the Principle of Rationality. In a series of experiments using a violation-of-expectation paradigm, Csibra et al. (1999) presented young infants with simple visual displays of geometric figures. The infants successfully used teleological constraints to predict these figures' actions from their goals and situational constraints. More recently, Baker, Saxe, and Tenenbaum (2009) developed a computational model to capture adults' inferences about goals from a display of the agent's movements in a simple maze.

The Principle of Rationality could lead people to discount the causal efficacy of suboptimal decisionmakers in two ways. First, because actions are assumed to follow the Principle of Rationality, apparently suboptimal actions are often reinterpreted as optimal actions under different assumptions-for example, that the agent was acting under a different goal, or that the agent's beliefs were incomplete or erroneous (Baker et al., 2009; Buchsbaum et al., 2011). Although the action is optimal under such a reinterpretation, the assumptions made about the agent (such as ignorance) to rationalize the action may
undermine the agent's perceived causal role in producing the outcome. Second, Csibra et al. (1999) have suggested that conformity to the Principle of Rationality is used as a principle for determining which entities are treated as agents, that is, as subject to folk-psychological principles. A decreased belief in the decision-maker's status as an agent could lead to a decreased attribution of causation. Indeed, such reasoning may not be restricted only to human agents. Kelemen and Rosset (2009) found that even adults apply teleological principles 'promiscuously' to inanimate objects. If people use efficiency cues to classify entities as agents, they may assign greater causal responsibility for objects that fulfill their causal affordances in the most efficient manner.
If the Optimality view is correct, responsibility judgments would be higher when $\mathrm{P}_{\mathrm{ACT}}>\mathrm{P}_{\text {ALT }}$ (an optimal choice) than when $\mathrm{P}_{\mathrm{ACT}}<\mathrm{P}_{\text {ALT }}$ (a suboptimal choice). However, the size of $\left[\mathrm{P}_{\mathrm{ACT}}-\mathrm{P}_{\text {ALt }}\right]$ would not affect judgments, since optimality is a qualitative property of a choice and does not depend on the magnitude of this difference. This view does not make a specific prediction about judgments when $P_{A C T}=P_{\text {ALT }}$, because there is no uniquely optimal choice in such situations.

Three experiments distinguished among these possibilities. First, we examined whether an agent's perceived responsibility (Experiment 1A) or causal contribution (Experiment 1B) for the outcome of a decision depends on the efficacy of an alternative, forsaken option. Second, we replicated and extended these findings using situations in which agents decide among three options (Experiment 2A) or in which the base rate of the outcome is specified (Experiment 2B), ruling out an alternative interpretation of Experiment 1. Finally, we explored the possibility that optimality constraints are used in assessing causation more generally, even for inanimate causes (Experiment 3).

## Experiments 1A and 1B

In Experiment 1, we examined whether the predictions of the Counterfactual Stability, Difference-Making, or Optimality view best capture judgments about vignettes such as those presented in the introduction. Additionally, to assess the consistency of these effects across measures, we included questions both about responsibility (in Experiment 1A) and about causation (in Experiment 1B).

## Method

Participants Fifty participants ( $56 \%$ female) were recruited from Amazon Mechanical Turk to participate in Experiment 1 A , and a different group of 50 participants ( $44 \%$ female) to participate in Experiment 1B.
Materials and Procedure Participants read five vignettes similar to the text given above, with five different cover stories. In these vignettes, $\mathrm{P}_{\mathrm{ACT}}$ was fixed at $50 \%$, while $P_{\text {ALT }}$ was varied across cover stories (at $10 \%, 30 \%, 50 \%$, $70 \%$, and $90 \%$ ) using a Latin square. The corresponding values of $\Delta \mathrm{P}\left(\mathrm{P}_{\text {ACT }}-\mathrm{P}_{\text {ALT }}\right)$ are thus $40 \%, 20 \%, 0 \%$,
$-20 \%$, and $-40 \%$, respectively. Participants were asked to rate their agreement with either a responsibility statement (e.g., "Angie is responsible for the flowers turning red") in Experiment 1A, or with a causal statement ("Angie caused the flowers to turn red") in Experiment 1B, on an 11-point scale (0: 'disagree'; 5: 'neither agree nor disagree'; 10: 'agree'). Manipulation check questions were included in this and all subsequent studies to monitor comprehension of the vignettes; however, these questions are not discussed further because no participants were eliminated from the analysis for these or any subsequent experiments.

## Results

As shown in Table 1, judgments of responsibility and causation were higher when $\mathrm{P}_{\mathrm{ACT}}>\mathrm{P}_{\mathrm{ALT}}$ than when $\mathrm{P}_{\mathrm{ACT}}$ $<\mathrm{P}_{\mathrm{ALT}}$, while judgments were intermediate when $\mathrm{P}_{\mathrm{ACT}}=$ $P_{\text {ALT }}$. Yet, the magnitude of the difference between $\mathrm{P}_{\mathrm{ACT}}$ and $P_{\text {ALT }}$ had no effect on judgments beyond the direction of the difference, consistent with the Optimality view.

This pattern was confirmed with a mixed-model ANOVA on judgments, with $\mathrm{P}_{\text {ALT }}(10 \%, 30 \%, 50 \%, 70 \%$, or $90 \%$ ) as a within-subjects factor, and Experiment (responsibility question or causal question) as a betweensubjects factor. This revealed a significant main effect of $\mathrm{P}_{\mathrm{ALT}}, F(4,392)=13.97, M S E=3.34, p<.001, \eta_{\mathrm{p}}{ }^{2}=.13$, and a main effect of Experiment, $F(1,98)=10.54, M S E=$ 15.41, $p=.002, \eta_{\mathrm{p}}^{2}=.10$, with responsibility ratings higher overall than causal ratings ( $M=6.09, S D=1.67$ vs. $M=4.95, S D=1.83$ ). This main effect may have occurred because the word 'cause' triggered a deterministic causal concept at odds with the probabilistic character of the decision problem. There was no interaction between $\mathrm{P}_{\mathrm{ALT}}$ and Experiment, $F(4,392)=$ $0.87, M S E=3.34, p=.48, \eta_{\mathrm{p}}^{2}<.01$, indicating no reliable difference in the effect of $\mathrm{P}_{\mathrm{ALT}}$ between experiments.

To explore the main effect of $\mathrm{P}_{\mathrm{ALT}}$, pairwise planned comparisons were conducted on adjacent $\mathrm{P}_{\text {ALT }}$ conditions (means from the combined experiments are presented in the bottom row of Table 1). The $10 \%$ and $30 \%$ conditions did not differ, $t(99)=1.09, S E M=0.21, p=.28, d=0.11$, nor did the $70 \%$ and $90 \%$ conditions, $t(99)=-0.21$, SEM $=0.24, p=.84, d=-0.02$. However, the $50 \%$ condition differed significantly from both the $30 \%$ condition, $t(99)$ $=-3.13, S E M=0.25, p=.002, d=-0.31$, and the $70 \%$ condition, $t(99)=2.22, S E M=0.21, p=.029, d=0.22$.

Table 1: Results of Experiment 1 (SDs in parentheses).

| $\mathbf{P}_{\text {AlT }}$ | $\mathbf{1 0 \%}$ | $\mathbf{3 0 \%}$ | $\mathbf{5 0 \%}$ | $\mathbf{7 0 \%}$ | $\mathbf{9 0 \%}$ |
| :--- | :--- | :--- | :--- | :--- | :--- |
| Exp. 1A | 6.74 | 6.60 | 5.82 | 5.68 | 5.60 |
|  | $(2.24)$ | $(2.10)$ | $(2.35)$ | $(2.34)$ | $(2.49)$ |
| Exp. 1B | 5.96 | 5.64 | 4.84 | 4.06 | 4.24 |
|  | $(2.34)$ | $(2.29)$ | $(2.47)$ | $(2.73)$ | $(2.58)$ |
| Mean | 6.35 | 6.12 | 5.33 | 4.87 | 4.92 |
|  | $(2.31)$ | $(2.24)$ | $(2.45)$ | $(2.66)$ | $(2.61)$ |

## Discussion

These results are most consistent with the Optimality view. Although there was an effect of $\mathrm{P}_{\text {ALT }}$ on responsibility ratings, this occurred only because judgments were dependent on the sign of $\Delta \mathrm{P}$ : Judgments were higher when $\mathrm{P}_{\mathrm{ACT}}>\mathrm{P}_{\text {ALT }}$ than when $\mathrm{P}_{\mathrm{ACT}}=\mathrm{P}_{\mathrm{ALT}}$, and higher when $\mathrm{P}_{\mathrm{ACT}}=\mathrm{P}_{\mathrm{ALT}}$ than when $\mathrm{P}_{\mathrm{ACT}}<\mathrm{P}_{\mathrm{ALT}}$, but the magnitude of $\Delta \mathrm{P}$ did not affect judgments. This finding cannot be explained by either the Counterfactual Stability view, according to which responsibility judgments would be invariant over different values of $\mathrm{P}_{\text {ALT }}$, or by the Difference-Making view, according to which responsibility judgments would be proportional to $\Delta \mathrm{P}\left(\mathrm{P}_{\mathrm{ACT}}-\mathrm{P}_{\mathrm{ALT}}\right)$. Moreover, the lack of magnitudedependence held for both attributions of responsibility and of causation, indicating that these results are not due to idiosyncratic properties of either phrasing.

Although the Counterfactual Stability and DifferenceMaking views cannot explain the results of Experiment 1, these results do not uniquely entail the Optimality view, because participants could have made these responses on the basis of whether $\Delta \mathrm{P}>0$. In more complex decision problems, it is possible for multiple options to have positive $\Delta \mathrm{P}$, yet for only one option to be uniquely optimal. Experiment 2 investigated whether people would still be sensitive to optimality in more complex scenarios.

## Experiments 2A and 2B

To examine whether the response pattern in Experiment 1 was based on optimality or simply on whether $\Delta \mathrm{P}>0$, Experiment 2 used vignettes in which agents faced three choices, of which $(A)$ had a low probability of leading to the goal, $(B)$ had a moderate probability of leading to the goal, and $(C)$ had the highest probability of leading to the goal. Thus, $C$ is the optimal choice, but both $B$ and $C$ have positive $\Delta \mathrm{P}$ relative to $A$. If judgments are based on optimality, then an agent choosing $C$ should be rated more responsible than an agent choosing $B$, since $C$ is optimal but $A$ and $B$ are not. However, if people are merely sensitive to $\Delta \mathrm{P}$ being positive, they should rate agents choosing $B$ and $C$ equally highly, since $\Delta \mathrm{P}>0$ for both.
Experiment 2 employed two different framings of the "least optimal" alternative. In Experiment 2A, all three options were described as alternative choices with varying probabilities of success. In Experiment 2B, the "least optimal" alternative was described as a base rate-the probability of the goal occurring in the absence of any action. This second phrasing makes the fact that $\Delta \mathrm{P}>0$ for both of the other alternatives more salient, providing a stronger test against the Optimality hypothesis.

## Method

Participants One hundred participants (52\% female) were recruited from Amazon Mechanical Turk to participate in Experiment 2A, and a different group of 100 participants ( $49 \%$ female) participated in Experiment 2B.

Each experiment was conducted as part of a session that included additional experiments not reported here; the order of the experiments was counterbalanced.
Materials and Procedure In Experiment 2A, participants read two vignettes from Experiment 1, modified to read:

Angie has a shrub, and wants the shrub's flowers to turn red. She is thinking about applying a fertilizer, and has three options:
If she applies Formula LPN, there is a $10 \%$ chance that the flowers will turn red.
If she applies Formula PTY, there is a $50 \%$ chance that the flowers will turn red.
If she applies Formula NRW, there is a [30/70]\% chance that the flowers will turn red.
Angie chooses Formula PTY, and the flowers turn red.
For Experiment 2B, the phrase "if she applies Formula LPN" was replaced by the phrase "if she applies nothing," to make the $10 \%$ base rate more salient. Whether Formula NRW had a $30 \%$ or $70 \%$ chance of leading to the goal ( $\mathrm{P}_{\mathrm{ALT}}$ ) was manipulated within-subjects (in the former case, the actual choice was optimal, while in the latter case, the actual choice was suboptimal), with the assignment of $\mathrm{P}_{\mathrm{ALT}}$ to vignette counterbalanced. Participants rated the agent's responsibility for the outcome on the same 11-point scale as Experiment 1A.

## Results and Discussion

As shown in Figure 1, agents were viewed as less responsible when their choice was suboptimal, whether the "least optimal" option was described as an alternative (Experiment 2A) or as a base rate (Experiment 2B). This occurred even though $\Delta \mathrm{P}$ was positive for both choices.

An ANOVA was conducted on responsibility judgments, with $\mathrm{P}_{\text {ALT }}(30 \%$ or $70 \%$ ) as a within-subjects factor and Experiment ( 2 A or 2 B ) as a between-subjects factor. There was a main effect of $\mathrm{P}_{\mathrm{ALT}}, F(1,198)=18.57$, $M S E=1.91, p<.001, \eta_{\mathrm{p}}{ }^{2}=.09$, with responsibility rated higher when $\mathrm{P}_{\mathrm{ALT}}=30 \%(M=6.83, S D=2.01)$ than when $\mathrm{P}_{\mathrm{ALT}}=70 \%(M=6.24, S D=2.14)$. Thus, responsibility ratings were higher for optimal decisions.

There was also a main effect of Experiment, $F(1,198)=$ 12.57, $M S E=6.37, p<.001, \eta_{\mathrm{p}}{ }^{2}=.06$, with judgments higher in Experiment 2B $(M=6.98, S D=1.53)$ than in Experiment 2A $(M=6.09, S D=2.01)$. This may have occurred because Experiment 2B described the least effective option as an omission rather than as an action, creating a qualitative difference among the options. However, there was no interaction between Experiment and $\mathrm{P}_{\mathrm{ALT}}, F(1,198)=0.30, M S E=1.91, p=.59, \eta_{\mathrm{p}}^{2}<.01$.

Although the Difference-Making account would make the same predictions as the Optimality account for these cases, the results of Experiment 2 show that the responses in Experiment 1 are unlikely to have been based merely on considerations of whether $\Delta \mathrm{P}>0$, since this was the case for both the optimal and the suboptimal case in Experiment 2. The Optimality view is the only account that is consistent with the results of both experiments.


Figure 1: Mean responsibility judgments (Experiment 2) and causal judgments (Experiment 3) on 11-point scales.

## Experiment 3

One potential explanation for the results of Experiments 1 and 2 is that participants were reinterpreting apparently suboptimal actions as guided by a different set of assumptions about the agent's knowledge or goals (e.g., Baker et al., 2009; Buchsbaum et al., 2011). Angie's objectively suboptimal decision to use Formula PTY may have led participants to view her as ignorant of the choice situation, or as having some other goal in mind other than making the flowers turn red, and participants may have accordingly downgraded her responsibility.

Another possibility, however, is that these results reflect principles used to designate entities as subject to our folk-psychological theorizing in the first place. Csibra et al. (1999) suggested that the Principle of Rationality is used for this purpose; indeed, efficiency may even be detected at the perceptual level (Gao \& Scholl, 2011). If individuals failing to behave optimally are not conceptualized as agents to the same extent as those behaving optimally, this could lead people to discount their role in causally producing the outcome.

If our earlier effects were obtained at least in part because optimality is used as a principle for designating agents, optimality considerations might be used more generally in causal reasoning, outside the social realm, because people may reason about efficient causes as though they were endowed with agent-like properties. This prediction, while counterintuitive, is bolstered by findings of 'promiscuous' teleological reasoning (Kelemen \& Rosset, 2009), with children and even adults under time pressure treating natural kinds as though they were artifacts endowed with purposes. Experiment 3 examined this prediction by testing whether people treat event types as more causal when they lead optimally to their effect, relative to other possible causes.

## Method

Participants One hundred participants ( $44 \%$ female) were recruited from Amazon Mechanical Turk to participate in Experiment 3. This experiment was
conducted as part of a session that included an additional experiment not reported here; the order of the experiments was counterbalanced.
Materials and Procedure Participants read two vignettes adapted from Experiment 2 B so that they no longer referred to a choice made by a human agent, but instead to the probability of an effect occurring given two different (non-human) causes:

There is a certain shrub that has flowers which sometimes turn red. There are two brands of fertilizer: Formula PTY and Formula NRW.
When nothing is applied, there is a $10 \%$ chance that the flowers turn red.
When Formula PTY is applied, there is a $50 \%$ chance that the flowers turn red.
When Formula NRW is applied, there is a [30/70]\% chance that the flowers turn red.
Each participant was then asked to what extent they agreed with the statement that "Formula PTY causes the flowers to turn red" on the same 11-point scale used in previous experiments. In both conditions, participants judged the strength of a cause with $50 \%$ efficacy (i.e., $\mathrm{P}_{\mathrm{ACT}}=50 \%$ ), while the efficacy of the other cause ( $\mathrm{P}_{\mathrm{ALT}}=$ $30 \%$ or $\mathrm{P}_{\mathrm{ALT}}=70 \%$ ) was manipulated within-subjects and counterbalanced with vignette.

## Results and Discussion

As shown in Figure 1, Formula PTY was judged more causal when it was optimal than when it was suboptimal, even though it always had a $50 \%$ chance of leading to the effect. A paired-sample $t$-test revealed that causal ratings were higher in the $\mathrm{P}_{\mathrm{ALT}}=30 \%$ condition $(M=6.58, S D=$ 1.67) than in the $\mathrm{P}_{\mathrm{ALT}}=70 \%$ condition $(M=6.00, S D=$ 1.72), $t(99)=4.50, S E M=0.13, p<.001, d=0.45$.

This result suggests that the mechanism underlying the optimality effect in lay decision theory is not specific to human agents, but can be extended 'promiscuously' to other entities, with people discounting a cause's efficacy in the face of a superior alternative cause.

One concern about this result might be potential scale or contrast effects. For example, suppose that participants implicitly judge each cause in each vignette (including the alternative cause that was not asked about), and always give the 'best' cause for each vignette the same rating. Then, a participant in the optimal condition might have assigned the better (actual) cause a rating of ' 7 ' and implicitly assigned the worse (alternative) cause a rating of ' 5 ', and in the suboptimal condition implicitly assigned the better (alternative) cause a rating of ' 7 ' and the worse (actual) cause a rating of ' 5 ', leading to our effect. Similarly, a contrast effect could have occurred if the psychological weight of $\mathrm{P}_{\mathrm{ACT}}$ differed between conditions. $\mathrm{P}_{\mathrm{ACT}}(50 \%)$ could have felt like a larger magnitude when compared to $\mathrm{P}_{\mathrm{ALT}}=30 \%$ than to $\mathrm{P}_{\mathrm{ALT}}=70 \%$, leading to higher ratings in the optimal condition.

Although these possibilities can only be ruled out definitively with future study, our within-subjects design
renders these explanations unlikely. The vignettes were read and judged consecutively, which both calls attention to the consistency of $\mathrm{P}_{\mathrm{ACT}}$ across conditions, and creates pressure to give identical responses across conditions. Nonetheless, converging evidence from other tasks will be of use in ruling out these possibilities more directly.

The present result should be distinguished from the superficially similar phenomena of discounting (e.g., Khemlani \& Oppenheimer, 2011) and cue competition (e.g., Waldmann \& Holyoak, 1992). Discounting occurs when one has a prior causal schema in which two causes (e.g., Formulas PTY and NRW) are each sufficient for an effect (the flowers turning red). If one cause (Formula NRW) is known to occur on some particular occasion, this makes the other cause (Formula PTY) less likely to be present on that occasion, because the known presence of Formula NRW "explains away" the effect and removes any reason to posit Formula PTY. Thus, the discounting phenomenon involves prior knowledge of causal types influencing subsequent inferences about causal tokens.

In the related phenomenon of cue competition, tokenlevel observational data affect subsequent formation of type-causal schemas. For example, in backward blocking, two candidate causes are first paired with the effect (i.e., Formulas PTY and NRW are both applied for several trials on which the flowers turn red), then one of the candidates alone is paired with the effect (i.e., only Formula PTY is applied for several trials on which the flowers turn red). Observing that the alternative cue (Formula PTY) produces the effect by itself reduces the belief that Formula NRW causes the effect in general.

However, the logics underlying these phenomena do not apply to our experimental situation. The input to the discounting process is type-causal schemas, and the output token-causal inferences; the input to cue competition is token-causal observations, and the output type-causal schemas. In our task, in contrast, participants made type-causal judgments from knowledge about statistical relationships at the type level. Thus, the present phenomenon is conceptually distinct.

Although Experiment 3 suggests that the optimality effect in lay decision theory occurs at least in part because conformity to the Principle of Rationality is used to designate entities as subject to folk-psychological principles, this does not preclude the possibility that some participants in Experiments 1 and 2 were additionally reinterpreting the agents' actions as optimal under a different set of assumptions. Nonetheless, Experiment 3 shows that re-interpretation cannot be a full explanation. Indeed, the effect in Experiment 3, which cannot not be explained in terms of re-interpretation, was of similar magnitude to that in Experiment 2.

## General Discussion

The present studies examined whether optimality is used as a cue for assigning causal responsibility. In Experiment 1, agents were judged more responsible for, and more
causal in, producing an outcome when their decision was the optimal choice for obtaining the outcome, but the magnitude of the difference between the efficacy of the optimal and suboptimal choices did not affect judgments. Experiment 2 showed that perceived responsibility is greater when a decision is optimal than when suboptimal, even when the suboptimal option is superior to a worst option or to the base rate of the outcome. Finally, Experiment 3 demonstrated an optimality effect in reasoning about causation for inanimate causes, suggesting that the optimality effect occurs at least in part because entities acting optimally are more likely to be designated as agents subject to our folk psychology.
Our results suggest several potentially promising avenues for future research. We are seldom confronted in real life with decisions for which we know the exact probabilities, more often entertaining a range of probabilities as potentially valid (Levi, 1985). A more ecologically valid test of our optimality hypothesis would specify realistic decision alternatives for which participants had a range of prior beliefs about the efficacy for achieving an outcome, rather than a single probability value. We chose to instead specify the probabilities so as to maximize experimental control. However, replicating the current results with more naturalistic stimuli would both enhance the generality of our findings and allow for exploration of boundary conditions.
Little appears to be known concerning folk beliefs about decision-making, what we term lay decision theory. In addition to shedding light on our theory of mind abilities, understanding the principles of lay decision theory may have practical implications for behavioral game theory, in which people must model others' behavior in order to make their own decisions. The present research addresses only a small fraction of the questions that might be asked: for example, how these beliefs are used in explaining and predicting behavior, whether (and when) people conceptualize decisions in terms of mental states or as states of the world (i.e., with mentalistic or teleological representations), how people conceptualize more complex decision problems in which multiple goals must be balanced against one another, and whether optimality constraints are applied equally to our own behavior as to the behavior of others. We are currently conducting research to probe these questions.

## Conclusion

Rationality constraints are commonplace heuristics for making inferences about human actions. The present research shows that such constraints also play a role in the evaluation of decisions, affecting how causal responsibility is assigned for an outcome.
As suboptimal decision-makers ourselves, it may appear hypocritical for us to hold others less responsible for the outcomes of their decisions when they decide suboptimally. Yet, we may have little choice-our rationality is, after all, bounded.

## Acknowledgments

This research was partially supported by funds awarded to the first author by the Yale University Department of Psychology. We thank Fabrizio Cariani, Winston Chang, Angie Johnston, Doug Medin, Emily Morson, Axel Mueller, Eyal Sagi, and Laurie Santos for their insightful suggestions, and the city of Nashville, TN for its inspirational, sunny November weather.

## References

Baker, C. L., Saxe, R., \& Tenenbaum, J. B. (2009). Action understanding as inverse planning. Cognition, 113(3), 329-349.
Buchsbaum, D., Gopnik, A., Griffiths, T. L., \& Shafto, P. (2011). Children's imitation of causal action sequences is influenced by statistical and pedagogical evidence. Cognition, 120(3), 331-340.
Cheng, P. W., \& Novick, L. R. (1992). Covariation in natural causal induction. Psychological Review, 99(2), 365-382.
Csibra, G., Gergely, G., Bíró, S., Koós, O., \& Brockbank, M. (1999). Goal attribution without agency cues: The perception of 'pure reason' in infancy. Cognition, 72(3), 237-267.
Davidson, D. (1967). Truth and meaning. Synthese, 17(3), 304-323.
Dennett, D. C. (1987). The intentional stance. Cambridge, MA: MIT Press.
Gao, T., \& Scholl, B. J. (2011). Chasing vs. stalking: Interrupting the perception of animacy. Journal of Experimental Psychology: Human Perception and Performance, 37(3), 669-684.
Jeffrey, R. C. (1965). The logic of decision. New York: McGraw-Hill.
Kelemen, D., \& Rosset, E. (2009). The human function compunction: Teleological explanation in adults. Cognition, 111(1), 138-143.
Khemlani, S. S., Oppenheimer, D. M. (2011). When one model casts doubt on another: A levels-of-analysis approach to causal discounting. Psychological Bulletin, 137(2), 195-210.
Levi, I. (1985). Imprecision and indeterminacy in probability judgment. Philosophy of Science, 52(3), 390-409.
Simon, H. A. (1956). Rational choice and the structure of the environment. Psychological Review, 63(2), 129138.

Spellman, B. A. (1997). Crediting causality. Journal of Experimental Psychology: General, 126(4), 323-348.
Tversky, A., \& Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185(4157), 1124-1131.
Waldmann, M. R., \& Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. Journal of Experimental Psychology: General, 121(2), 222-236.

# Learning complementary action with differences in goal knowledge 

Jeremy Karnowski (jkarnows@cogsci.ucsd.edu)<br>Department of Cognitive Science, 9500 Gilman Drive<br>La Jolla, CA 92093-0515 USA

Edwin Hutchins (ehutchins@cogsci.ucsd.edu)<br>Department of Cognitive Science, 9500 Gilman Drive<br>La Jolla, CA 92093-0515 USA


#### Abstract

Humans, as a cooperative species, need to coordinate in order to achieve goals that are beyond the ability of one individual. Modeling the emergence of coordination can provide ways to understand how successful joint action is established. In this paper, we investigate the problem of two agents coordinating to move an object to one agent's target location through complementary action. We formalize the problem using a decisiontheoretic framework called Decentralized Partially Observable Markov Decision Processes (Dec-POMDPs). We utilize multiagent Q-learning as a heuristic to obtain reasonable solutions to our problem and investigate how different agent architectures, which represent hypotheses about agent abilities and internal representations, affect the convergence of the learning process. Our results show, in this problem, that agents using external signals or internal representations will not only eventually perform better than those that are coordinating in physical space alone but also outperform agents that have independent knowledge of the goal. We then employ information theoretic measures to quantify the restructuring of information flow over the learning process. We find that the external environment state varies in its informativeness about agents' actions depending on the agents' architecture. Finally, we discuss how these results, and the modeling technique in general, can address questions regarding the origins of communication.


Keywords: Dec-POMDPs; multi-agent Q-learning; Behavioral Info-Dynamics; mutual information

## Introduction

The moment we move from a study of individual cognition to a detailed analysis of the social realm, we have committed ourselves to the investigation of a different type of system. There is no centralized controller; this system is inherently decentralized. The questions we ask, however, may be similar. Just as we wish to study how a individual decision maker adapts its behavior in a task environment, we can investigate the ways in which multiple, possibly non-identical, decision makers reorganize their internal world and their external interactions to form a new functional system that solves a problem which cannot be addressed by one individual alone (Hutchins, 1995).

One important problem that cooperative agents face is how to coordinate their movements to arrive at a goal known only to one of the agents. This problem was addressed in Hazlehurst and Hutchins (1998), where the authors constructed an algorithm that allowed for a set of agents to converge on similar form-meaning mappings which also related to their movements within a given environment. This setup, like many modeling studies that focus on issues of hidden goals of other agents, has a strong predilection towards imita-
tive learning. Not all learning and reorganization in a multiagent system is imitative, however, and another focus of modeling should be on complementary action learning (Hutchins \& Johnson, 2009). It has been shown elsewhere that agents can learn to coordinate in complementary ways without sharing information about each other (Sen, Sekaran, Hale, et al., 1994), but this presumes an environment where there is only one destination and both agents know its identity. By combining aspects from these two studies, we can investigate scenarios in which agents must collaboratively, through complementary action, arrive at a goal location known to only one agent.

While it is typically intractable to find the optimal solution to many multi-agent coordination problems, these problems are particularly important because their inherent challenges highlight several important features of social interaction and group dynamics that need to be studied:

1. Non-stationary World: Agents are constantly adapting to the statistics of their environment, including other agents. Since other agents do not have a fixed method of interacting with the world a priori, the world is inherently nonstationary (Buşoniu, Babuška, \& Schutter, 2008).
2. Non-independent Sampling: An agent's own actions affect its incoming sensory information and this in turn affects the regularities it can extract from the world (Lungarella \& Sporns, 2005). Motor activity and sensory information obtained from the environment are interdependent; the way we move in the world shapes our understanding of it and these patterns of data have structure.
3. Distribution of Knowledge: Not all agents in the world have access to the same information or capabilities. The social realm is comprised of more than just a set of identical individual problem solvers (Hutchins, 1995).

Another prominent research direction in studying multiagent systems is determining "(h)ow to develop... problem solving protocols (information flow) that enable agents to share results and knowledge in a timely, effective manner" (Sen, 1997). It is important to understand how a group of individual agents reorganizes in functional ways that alter the flow of information; we need to understand "what information goes where and in what form" (Hutchins, 1995) and how these pathways change. This situation is complicated by the
fact that researchers in Cognitive Science hold different assumptions about the internal organization and external behavior of agents, which specifies the model elements, and this constrains the possible ways to reconfigure information flow. This situation can be rectified, however, by utilizing a common formalism for comparing and contrasting the consequences of different sets of assumptions.

In this paper, we utilize a formal framework, Decentralized Partially Observable Markov Decision Processes (DecPOMDPs), to place our problem of interest into a larger set of multi-agent coordination problems in order to investigate coordination problems when agents have access to different amounts of information (Karnowski, accepted). We then discuss how several assumptions about agent architecture map into specific changes in the problem structure, demonstrating how we can vary our hypotheses by altering the components of the Dec-POMDP. Through the use of multi-agent Qlearning, we can demonstrate the speed with which agents reorganize themselves into stable patterns of behavior that allow them to coordinate their actions and achieve a joint goal. This reorganization brings differences in performance, however, based on the assumptions made about agent capabilities. We utilize mutual information to measure the changes in statistical dependencies among streams of information and to show how agents' behaviors respond to environmental regularities. We conclude by discussing how one problem formulation may provide insights into the study of the evolution of communication and future directions in this area.

## Methods

## Decentralized Partially Observable Markov Decision Processes (Dec-POMDPs)

Dec-POMDPs (D. Bernstein, Zilberstein, \& Immerman, 2000) are a way to formalize multi-agent coordination problems. They provide a common structure that aids in the discussion of related problems and the development of solution techniques. While there exist other frameworks that tackle problems of agent coordination and problem solving (Dec-POMDP-COM, MTDP, and COM-MTDP with perfect recall), many of them have been shown to be formally equivalent (Seuken \& Zilberstein, 2008). The reason for the variety is that the frameworks emphasize different features. For instance, while Dec-POMDPs and Dec-POMDP-COMs (DecPOMDPs with communication) (Goldman, Allen, \& Zilberstein, 2007) are formally equivalent, the former tends to focus on bodily coordination in physical space and the latter with problems that also involve symbolic coordination. In addition to communication, frameworks often contain assumptions about the representational capacities of their agents, providing agents with, for example, the ability to model the goals or actions of other agents (Claus \& Boutilier, 1998). Providing a language for researchers in Cognitive Science to systematize problems in cooperative multi-agent interactions and make explicit their assumptions about individual architecture will allow for a thorough comparison of current models and
the exploration of regions between models with different assumptions.

Formally, a Dec-POMDP can be defined by a tuple $\langle\{A g\}, S,\{A\}, P,\{\Omega\}, O, R\rangle$, where $\{A g\}=\{1,2, \ldots, n\}$ is the set of agents, S is the possible states of the world, $\{A\}=$ $\left\{A_{1}\right\} \times\left\{A_{2}\right\} \times \ldots \times\left\{A_{n}\right\}$ is the set of joint actions (with $a=\left(a_{1}, a_{2}, \ldots, a_{n}\right)$ being a joint action and action $a_{i}$ is the action of agent i ), P is the transition function (with $P\left(s^{\prime} \mid s, a\right)$ being the transition to state s' given current state $s$ and joint action a), $\{\Omega\}$ is the set of possible observations, O is the matrix that defines the probability of seeing observation o given state s , and $R=R\left(s, a, s^{\prime}\right)$ is the reward for taking the joint action a in state s and transitioning to state s '. The goal of solving a Dec-POMDP is to find a joint policy $\pi=\left\{\pi_{1}, \pi_{2}, \ldots, \pi_{n}\right\}$ (where each $\pi_{i}$ is a local policy of one agent that maps an observation of a state to an action, i.e. $\pi_{i}: S \rightarrow A_{i}$ ) such that the group minimizes some cost function over time (similarly, it can maximize a reward function).

## Multi-agent Q-learning

Dec-POMDPs are a useful abstraction which allows for a common language when speaking about coordination problems. These problems, are typically difficult to solve (D. Bernstein et al., 2000), but solution algorithms are a current research trend (Spaan \& Oliehoek, 2008). Another way to address these problems is to use on-line adaptive heuristic algorithms that provide good approximate solutions, such as Q-learning (CJC, 1989), as they stochastically approximate off-line learning of optimal policies. In this paper, we use the Q-learning algorithm in a multi-agent context (Buşoniu et al., 2008). Within each agent, state-action pairs are strengthened depending on the outcome of the chosen action. For instance, if an agent transitions to state $s^{\prime}$ after performing action $a$ while in state $s$, an agent will receive a reinforcement $R$ and update the value of that state-action pair $(s, a)$ :

$$
\begin{equation*}
Q(s, a) \leftarrow(1-\alpha) Q(s, a)+\alpha\left(R+\max _{a^{\prime} \in A} Q\left(s^{\prime}, a^{\prime}\right)\right) \tag{1}
\end{equation*}
$$

Other parameters relate to the learning algorithm itself. The learning rate, $\alpha$, determines the degree to which the current state is updated given new experience, and the discount factor, $\gamma$, specifies how influential future states and actions are to the current state. In this experiment, actions were chosen in a greedy manner.

## Behavioral Info-Dynamics

Consider an isolated animal collective $X$ consisting of $n$ freely moving animals. Temporal data is collected on each animal's behavior generating a unique time series. Given a collection of sensorimotor time series data from a set of animals, we can measure statistical dependencies during different behavioral patterns. Tononi, Sporns, and Edelman (1994) (and later Tononi, Edelman, and Sporns (1998)) introduced a set of appropriately defined information-theoretic measures to capture the statistical properties of a system with $n$ components. While their methods were originally designed to study
neural systems, more recent work has adapted these measures to study sensorimotor coordination in embodied agents by collecting sensor and motor time series data (Lungarella, Pegors, Bulwinkle, \& Sporns, 2005). We utilize a Python implementation of these measures (available at https:// github.com/OpenCV-at-DCog-HCI/BID) to further extend these measures to study the behavior of a system of agents. In this paper, we focus only on the mutual information between pairs of time series. Depending on their interaction with the world, solitary agents and collections of agents exploit different statistical dependencies among streams of information. We can show these changes by measuring mutual information (Sporns, Karnowski, \& Lungarella, 2006; Di Prodi, Porr, \& Wörgötter, 2010).

Entropy defines the uncertainty inherent in a time series, or the average amount of information present. For instance, if knowing the state of the system at a given point in time will give you a lot of information about the time series as a whole, then this will contribute to a lower entropy. This could happen if that state is highly unlikely, and thus is more informative. If every state, however, is equally likely, then knowing the state at one point in time gives no information about the time series as a whole and entropy is maximal.

$$
\begin{equation*}
H(X)=-\sum_{j=1}^{n} p\left(x_{j}\right) \log \left(p\left(x_{j}\right)\right) \tag{2}
\end{equation*}
$$

Mutual information measures the dependence between two distributions (and in our case, time series). It is defined as the Kullback-Leibler distance ( $D_{K L}$ ) between the joint distribution $p\left(X_{1}, X_{2}\right)$ and the independent distribution $p\left(X_{1}\right) p\left(X_{2}\right)$. Mutual information is also defined as the sum of the entropies of the individual parts with the joint entropy subtracted out.

$$
\begin{align*}
& M I\left(X_{1}, X_{2}\right)=D_{K L}\left[p\left(X_{1}, X_{2}\right) \| p\left(X_{1}\right) p\left(X_{2}\right)\right]= \\
& H\left(X_{1}\right)+H\left(X_{2}\right)-H\left(X_{1}, X_{2}\right) \tag{3}
\end{align*}
$$

Any dependence between the two time series will increase the mutual information between them. For instance, if the state of one agent provides a lot of information about the state of another agent, this will result in higher mutual information. If the agents are completely independent, then this predictive power is lost, and mutual information will be zero.

## Problem and Experimental setup

To explore how two agents could coordinate via complementary actions to arrive at a hidden goal, we created an extension of the 'block pushing problem' (Matarić, 1996; Sen et al., 1994) where two agents are tasked to move from a start location to the goal, which is one of two possible locations, and follow as closely as possible a path P between the two. At every timestep, Agent i uses a force $\vec{F}_{i}$, where $0 \leq\left|\vec{F}_{i}\right| \leq F_{\max }$ on the block at an angle $\theta_{i}$, where $0<\theta_{i}<\pi$, which results in the block being offset by $\left|\vec{F}_{i}\right| \cos (\theta)$ in the $x$ direction and $\left|\vec{F}_{i}\right| \sin (\theta)$ in the $y$ direction. The new position of the block is calculated by vector addition of the displacement created
by the two agents. The new coordinates are then assigned to the correct discrete bin. The location of the block is used as feedback for the agents, depending on which scenario is being considered.

In our problem, $\{A g\}$ is a set of two agents, $S$ is the xcoordinate in a $20 \times 20$ grid world, the actions are a vectoraddition of individual agent actions that combine force and angle ( $0.2 \leq\left|\vec{F}_{i}\right| \leq 2.0$ ) in 0.2 increments and $15 \leq \theta_{i} \leq 165$ in 15 degree increments), $P$ is deterministic (the probabilities of moving to the next state given a joint action is 1 and the rest are zero), the set of observations is always the current x-coordinate in the grid world but more information is added depending on the scenario (for the agent with the goal, the current goal is also added to the observation), $\Omega$ is deterministic (the probabilities of an agent perceiving a particular observation given a state is 1 and the rest are zero), and the feedback depends on the scenario.

The first goal of our study was to establish a baseline. We implemented the scenario as found in (Sen et al., 1994):
0. Agent 2 also knows goal (Full Information): Both agents receive an observation of their x-coordinate and the goal. Their feedback is a function of their distance from the goal path $P$.

Even though there are two possible paths, there is only one goal for each trial, and therefore our agents acted in similar manner and replicated the results obtained by Sen et al. (1994). We then set out to construct a situation where there is a disparity in the amount of information accessible to each agent. In our 'base case', we consider the impact of removing information about the goal from Agent 2 and only allowing Agent 1 to have this knowledge. From here, our models were motivated by research agendas within Cognitive Science. Given different assumptions of agent architecture, we alter the Dec-POMDP in specific ways:

1. Agent 1 knows the goal but Agent 2 does not ('Base Case'): Agent 1 remains identical to previous results, but the observation Agent 2 receives does not contain information about the goal. The feedback for Agent 2 is a function of the distance from the closest path (i.e. when there is no information about the goal, the closest path is the best)
2. Agent 2 tracks probability of goal ('Theory of Mind'): Giving an agent the ability to represent the goal of another agent and make inferences about that goal given data is one way to conceptualize Theory of Mind. In this situation, Agent 2 begins a trial with the prior belief that either goal is the possible target. At each time step, the state of the world is a sample with which Agent 2 updates its belief of the current goal via Bayes rule. The probability of this sample is the probability that the x -coordinate is sampled from a Gaussian distribution with the x-coordinate of the goal being the mean and a standard deviation of 2.5 (Altering this distribution is future work). The probability space was discretized into 10 bins. The feedback for Agent 2 is
an weighted average (given current belief) of the feedback for both paths.
3. Agent 1 can make sounds ('Communication'): Agent 1 produces either a 0 or 1 which becomes part of the state which Agent 2 will experience on the next time step. The feedback for Agent 2 is a function of the closest path.
4. Agent 1 can make sounds and Agent 2 tracks probability of goal ('Theory of Mind' and 'Communication'): This is a combination of the previous two alterations. The feedback for Agent 2 is the weighted average of the feedback for both paths.

The feedback in each of these cases is determined by a function of the distance from the desired path, $f(\delta x)=$ $K * a^{-\delta x}$, similar to the original setup in Sen et al. (1994). This provides a high value for being on the path and an exponentially decreasing value further away from the desired path. Starting out the learning process with high values for state-action pairs and providing feedback after every trial was another feature in Sen et al. (1994) that allowed the agents to explore the available actions (alternatively, one could set the values in the beginning to be zero, but receiving feedback after just one trial would bias the agent to take the same path every trial). Also, any updates to state-action pairs could not be larger than the original high value (in our case, this was set to 100).

At the beginning of every trial, the two agents start at $(x, y)=(10,0)$ and the goal is randomly chosen from two options: $(3,20)$ or $(17,20)$. They make individual actions which combine into a joint action as outlined above. If the agents move the object outside of the $20 \times 20$ grid world, then the trial ends. Similarly, if the agents arrive at the goal state, the trials ceases. In the rare chance that agents would take more than 100 timesteps, the trial would also stop (forcing the angles to not allow agents to travel parallel to the x-axis helps alleviate this problem). An additional feature incorporated into the world dynamics was an automatic movement forward if the agents did not move forward enough on a trial. This was added to ensure agents did not remain still and allowed for better convergence.

## Results

In our experiments, agents always began with equally valuable state-action pairs and this caused their actions to be selected randomly. Over many trials, as agents adjust the values of different actions within each state, their behaviors begin to become patterned. Practices reduce the entropy of the shared environment, which leads to better policies and to a decrease in the average distance from the goal path. One would suspect, however, that performance would be best when there is complete information for both agents and that scenarios in which one agent has partial and incomplete information, the resulting joint actions would lead to poorer performance. This is not what we find, as shown in Figure 1. Having the ability to produce and utilize sounds allows agents, over time,
to perform better than those with complete information. Having the ability to represent and make inferences about the goals of another agent provides even more improvement in joint coordination.


Figure 1: The average distance of the actual path from the goal path given different agent assumptions ( $\alpha=0.01, \gamma=$ $0.9)$. Each experiment had 5000 trials and the data has been averaged over 100 experiments. Other learning rates ( $\alpha \in$ $\{0.1,0.2,0.3\}$ ) resulted in the similar patterns of performance with different rates of convergence.

We can determine how the two agents functionally reorganized themselves based on the levels of statistical dependence between different data streams. Mutual information provides a way to measure how predictable one data stream is from another. As we can see in Figure 2, both the scenario in which Agent 1 and Agent 2 have full knowledge of the goal and the 'base case', where Agent 2 does not know the goal, there is an increase in the mutual information between the x-coordinate and the angle of Agent 2 but this mutual informativeness plateaus. In the scenarios where there is Theory of Mind, Agent 2 is receiving a wealth of information about the goal through its current location but not necessarily needing to rely on any connection between its angle action choices and its location, which would have forced it to be more precise in its actions. In the scenarios with sound, there is a lot of extra structure in the shared environment that becomes highly predictive of the x-coordinate and therefore in the actions of Agent 2, including the angle. Another situation was created in which Agent 1 produced a sound but the state also included another random noise (to take away the special nature of the sound but not its ability to be manipulated). While the graph does not show the full increase of MI, other simulations showed this had the same trend as the case with communication, just over a longer period of time. This makes sense if agents were learning to utilize structure, but randomness was slowing this process down.

We did not find that the forces with which agents pushed the box had any predictive power for other data streams. When there was an increase in mutual information, it ap-


Figure 2: The mutual information between the x-coordinate and the angle of Agent 2.
peared to be due to the high predictability of angle and $x$ coordinate. As the world dynamics forced agents ahead one step if they did not apply enough force, it may have been the case that this affected the importance of force as a predictive element. This is probably not the case, however, as the agents in our model (and those in Sen et al. (1994)) only observe the x-coordinates, which would in turn dampen some of the informativeness of force in agent action choices.

## Discussion

In this paper, we have discussed the benefits of utilizing a common theoretical framework for addressing cooperative multi-agent problems in Cognitive Science and demonstrated how changes to framework elements can encapsulate various hypotheses about agent actions and internal representational capacities. We have designed a new multi-agent problem, focusing on understanding the acquisition of complementary actions in a goal-directed task where there is an information disparity. We used Q-learning, an algorithm commonly used in modeling single agent decision making, in a multi-agent setting to investigate how agent hypotheses affect the convergence of the learning process. And finally, we used mutual information to quantify how informative one data stream, the
x-coordinate, is about another data stream, the angle chosen by Agent 2 and charted the changes in this informativeness over time.

The results for this particular problem formulation provide a partial ranking of models based on performance. There are, however, a couple of caveats. First, while our simulated agents chose their actions in a greedy manner, different results might be obtained through other action selection methods, such as using a Boltzmann action selection mechanism. Second, Dec-POMDPs are typically used when there is some uncertainty in state transitions (due to modeling motor noise) or observations (due to sensory noise or partial view of the world). While this problem does not utilize this feature, future work manipulating these parameters may change the success of models with different assumptions about agent architecture.

This work highlights several of the open problems in the study of the emergence of communication, as it simultaneously investigates the origin of signaling channels, the sources of representation in signals, and the roles of social interaction in learned communication systems (Lyon, Nehaniv, \& Cangelosi, 2006).

Future work related to this particular example will strive to explore how agents could learn to discover that one information stream is informative about another, a hallmark of communication. As a starting point, for instance, we are particularly interested in the case where the agents have an ability to put structure into the shared environment through sounds. In this case, it could be that the agent with the goal is able to create noises, which allows the second agent to adjust its policy given this external structure. This in turn forces more regular behavior to which the speaking agent can then adjust. Originally, the noise was not functionally related to the current state; in the beginning, sounds just happened. As engagement proceeds, that noise ends up carrying information, and at that moment, the sounds would become a signaling channel.

This process, however, hasn't held any commitments to the content of that signaling channel. It may turn out that the speaking agent, through features of the algorithm, converges on highly rewarding action-sound pairings and the second agent only need adjust its behavior accordingly. In either case, we suspect that putting structure out into the world may create stable regularities with which agents could take advantage and eventually internalize (Vygotsky, 1978). Agent interactions themselves would be the determining factor behind the sources of representations in the signals they employ. In problems similar to ours, it is often the case that multi-agent Q-learning fails, precisely because neither agent experiences a stationary environment (Claus \& Boutilier, 1998). Placing stationary-creating behavior at the center of new algorithms is also possible future work.

Here we have shown that we can operationalize several assumptions in Cognitive Science and discover what structure and organization emerge from these hypotheses. In the present examples, however, agents are endowed with cer-
tain abilities a priori. We would really like to explore the conditions under which language-like abilities and Theory of Mind-like processes could emerge from ongoing interactions between autonomous agents. Additional future work will look at the space between these hypotheses and how various learning algorithms could take agents from a lack of abilities to a state where additional mental abilities have emerged through agent interactions.

## Acknowledgments

The authors would like to thank Chris Johnson, Ben Bergen, and Ben Cipollini for helpful discussions. Jeremy Karnowski is a Jacobs Fellow and is the recipient of a CARTA Graduate Fellowship in Anthropogeny.

## References

Bernstein, D., Zilberstein, S., \& Immerman, N. (2000). The complexity of decentralized control of markov decision processes. In Proceedings of the sixteenth conference on uncertainty in artificial intelligence (pp. 32-37).
Bernstein, D. S., Givan, R., Immerman, N., \& Zilberstein, S. (2002). The complexity of decentralized control of markov decision processes. Mathematics of Operations Research, 27(4), 819-840.
Buşoniu, L., Babuška, R., \& Schutter, B. De. (2008, March). A comprehensive survey of multi-agent reinforcement learning. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 38(2), 156-172.
CJC, H. (1989). Learning from delayed rewards. Cambridge University, Cambridge, England, Doctoral thesis.
Claus, C., \& Boutilier, C. (1998). The dynamics of reinforcement learning in cooperative multiagent systems. In In proceedings of national conference on artificial intelligence (aaai-98 (pp. 746-752).
Di Prodi, P., Porr, B., \& Wörgötter, F. (2010). A novel information measure for predictive learning in a social system setting. From Animals to Animats 11, 511-522.
Goldman, C., Allen, M., \& Zilberstein, S. (2007). Learning to communicate in a decentralized environment. $A u$ tonomous Agents and Multi-Agent Systems, 15(1), 4790.

Hazlehurst, B., \& Hutchins, E. (1998). The emergence of propositions from the co-ordination of talk and action in a shared world. Language and Cognitive Processes, 13(2-3), 373-424.
Hutchins, E. (1995). Cognition in the wild. MIT Press.
Hutchins, E., \& Johnson, C. (2009). Modeling the emergence of language as an embodied collective cognitive activity. Topics in Cognitive Science, 1(3), 523-546.
Karnowski, J. (accepted). Modeling collaborative coordination requires anthropological insights. Topics in Cognitive Science.
Lungarella, M., Pegors, T., Bulwinkle, D., \& Sporns, O. (2005). Methods for quantifying the informational
structure of sensory and motor data. Neuroinformatics, 3(3), 243-262.
Lungarella, M., \& Sporns, O. (2005). Information selfstructuring: Key principle for learning and development. Proceedings The 4nd International Conference on Development and Learning 2005, 25-30.
Lyon, C., Nehaniv, C., \& Cangelosi, A. (2006). Emergence of communication and language. Springer.
Matarić, M. (1996). Learning in multi-robot systems. Adaption and Learning in Multi-Agent Systems, 152-163.
Sen, S. (1997). Multiagent systems: Milestones and new horizons. Trends in Cognitive Sciences, 1(9), 334-340.
Sen, S., Sekaran, M., Hale, J., et al. (1994). Learning to coordinate without sharing information. In Proceedings of the national conference on artificial intelligence ( pp . 426-426).
Seuken, S., \& Zilberstein, S. (2008). Formal models and algorithms for decentralized decision making under uncertainty. Autonomous Agents and Multi-Agent Systems, 17(2), 190-250.
Spaan, M., \& Oliehoek, F. (2008). The multiagent decision process toolbox: Software for decision-theoretic planning in multiagent-systems.
Sporns, O., Karnowski, J., \& Lungarella, M. (2006). Mapping causal relations in sensorimotor networks. In Proc. of the 5th int. conf. on development and learning.
Tononi, G., Edelman, G., \& Sporns, O. (1998). Complexity and coherency: integrating information in the brain. Trends in cognitive sciences, 2(12), 474-484.
Tononi, G., Sporns, O., \& Edelman, G. (1994). A measure for brain complexity: relating functional segregation and integration in the nervous system. Proceedings of the National Academy of Sciences of the United States of America, 91(11), 5033-5037.
Vygotsky, L. (1978). Mind in society. Harvard University Press.

# Representational Shifts Towards the Prototype in Memory for Hue 

Laura Jane Kelly (lkelly@ucmerced.edu)<br>Evan Heit (eheit@ucmerced.edu)<br>Cognitive and Information Sciences, University of California, Merced<br>5200 N. Lake Road, Merced, CA 95343 USA


#### Abstract

Representational shifts in memory have been a recent topic of interest and debate (Blanco \& Gureckis 2012; Lupyan, 2008; Richler, Gauthier \& Palmeri, 2011; Richler, Palmeri \& Gauthier, 2012). Whether there are true systematic biases in memory due to a stimulus being labeled has been proposed and contested. The fundamental proposal that representations shift toward the prototype has not previously been demonstrated. In the present experiment, participants judged colored silhouettes by color category or by preference, then were asked to remember the hue of the original silhouette among five narrowly distinct options. By using the single dimension of hue, we are able to show prototypical representational shifts in memory for colored silhouettes after a few minutes. We did not observe a difference between color labeled and preference judged silhouettes, refuting the claim that labeling is the source of prototypical representational shifts.


Keywords: Concepts and Categories; Representation; Perception; Memory; Labels; Color

## Introduction

A representational shift is a spatial metaphor for a systematic difference between the representation as measured and the original stimulus that inspired the representation. The representation is said to have shifted from the original sensory input somewhere along its cognitive path before being measured. The representational shift hypothesis (Lupyan, 2008) suggests that when explicit labels are used while perceiving an object, the encoded representation shifts from what it would be without explicit labeling to a more prototypical representation.

The original experiments (Lupyan, 2008), described in more detail below, did not directly address the main predictions of the representational shift hypothesis: the existence of a systematic shift or the direction of that shift. The analysis relied on inferring a representational shift from a pattern of non-directional forgetting. The data only directly indicated worse memory for labeled objects. The existence of a systematic shift and the direction of the potential shift is simply not shown by the data collected.

In the experiment presented here, we use color to test memory for recently presented silhouettes of animals and objects. In a paradigm similar to Lupyan (2008), the objects are labeled or not labeled followed by a surprise memory task. Instead of a yes-no recognition task, we present an array of 5 hue variations for the participants to choose among. By switching from multidimensional objects to unidimensional hue, and testing fine variations of hue
memory rather than course-grained recognition memory, we can see whether there are systematic shifts of hue memory and their direction within the dimension of hue. The pattern of false alarms is used to look for representational shifts.

Why do these representational shifts matter? We already know there are top-down influences on memory (e.g., Heit, 1997). Two particularly dramatic examples of the imperfection of memory are eyewitness testimony (Wells \& Olson, 2003) and flashbulb memories (Schmolck, Buffalo, \& Squire, 2000). Small differences between the experience of an object or color and its representation in memory may seem minor in comparison to changing the race of a shooter based on stereotypes or radically rewriting how you heard about a defining national moment over the course of a few years. However, taking into account how pervasive these small differences would be, representational shifts could have extensive effects on how meaning is build and supported. Prototypes can be conceived of as resulting from the build up of exemplars over time (Nosofsky, 1986; Palmeri \& Nosofsky, 2001). As a ramification of the representational shift hypothesis, exemplars in memory would not accurately reflect experience alone but would also have a bias related to when the example of the category was experienced. The stronger the category, the stronger the pull of the prototype, and as a result, the more new exemplars are biased towards categories as they already stand.

## The Representational Shift Hypothesis

Lupyan (2008) proposed that there would be a difference in encoding and retrieving memories of objects if the categorization was explicit rather than only implicit at the time of encoding. Previous work had shown better category learning with labels than without labels (Lupyan, Rakison \& McClelland, 2007), suggesting that there are differences in how categories with labels are activated. The difference in activation of the category could influence the encoding of exemplars of the category. Specifically, the representational shift hypothesis asserts that concurrent labels activate a more prototypical representation of the category than would the perceived exemplars activate alone. The stored encoding is hypothesized to be a mixture of the exemplar activation and the label activation; the final representation has shifted toward the prototype as a result of the interaction of perception and semantic memory.
The original work (Lupyan, 2008) used an experimental paradigm consisting of a presentation of chairs and lamps followed by a surprise recognition test including both the old objects and matched new object lures that were very
similar to the old objects. There were two conditions, category judgments and preference judgments. During the initial presentation of the objects, each object was either judged to be a chair or lamp, or was judged to be liked or disliked. The results were then analyzed in terms of both hits and false alarms. Lower hit rates for the categorically judged objects than for the preference judged objects were taken to indicate a distorted or shifted memory. High false alarm rates, on the other hand, would have been taken to indicate overall poor memory. The results showed the predicted lower hit rate without a higher false alarm rate in the category judgment condition but not in the preference judgment condition. This pattern of forgetting was taken to be evidence for the representational shift account.

## Challenges to a Representational Shift Account

Not all researchers accept the representational shift hypothesis. An alternative hypothesis to explain the low hit rates (Lupyan, 2008) is depth of processing. A depth of processing account predicts both high false alarm rates and low hit rates for the category judgment condition because only minimal exemplar specific information would be encoded for categorical judgments but more detailed information about each exemplar would be processed and encoded for preference judgments. Follow-up experiments explored this idea using additional judgment conditions of location (Richler, Gauthier \& Palmeri, 2011) and of orientation (Blanco \& Gureckis, 2012). They tested whether a preference judgment simply forces more fine-grained processing of an object than a category judgment does, leading to more detailed encodings. The location and orientation judgments were not expected to require as much processing of the actual item as preference judgments. According to the representational shift hypothesis, if labeling forces a more typical encoding than non-labeling, then a category judgment should result in the lower hit rates without a change in false alarms while preference, location and orientation judgments should all have higher false alarm rates. The labeling conditions did not uniformly create lower hit rates than non-labeling conditions. The hit rates were only lower in comparison to the preference condition supporting depth of processing.

Richler, Palmeri, and Gauthier (2012) tested the representational shift hypothesis, or labeling effect, against a paradoxical production effect (MacLeod, Gopie, Hourihan, Neary \& Ozubko, 2010) which is characterized by more distinct and accurate memory for vocally named words and objects. Explicit category responses by button push, by silent labeling, and by verbal labeling were contrasted with preference judgments using the same surprise recognition memory test paradigm as Lupyan (2008). Richler et al. found varying levels of memory strength. Verbal labeling was remembered most accurately, then preference, followed by silent naming, with button press categorization being the least strong. The pattern of forgetting along with the pattern of accurate memory was used to support the distinctiveness
of processing account which emphasizes the uniqueness of features over depth of processing

In each of these follow-up experiments, the pattern of forgetting was used to support a depth of processing account.

## But What About the Shifts?

There is now one set of experiments in favor of the representational shift hypothesis and three sets opposed. However, none of these experiments truly get at the main prediction of the representational shift hypothesis: a stimulus processed with overt categorization will undergo a shift where the encoded representation in memory will be more prototypical than the original stimulus. The literature accepts an effect that is only implied, arguing about its cognitive mechanism rather than its existence. Rather than looking at the absence of recognition and trying to infer what processes could result in poor recognition memory, we can instead look at patterns of recognition to see if the encoding has shifted and by how much which should lead more straightforwardly to possible underlying processes. One way to go about looking at the shifts is to move the categorization and subsequent shifts onto a single dimension, in this case the hue dimension, rather than trying to infer shifts in chair-versus-lamp space, with multiple unknown dimensions.

## Hue Perception and Memory

In psychological color research, it has been found that there is a near universal progression of basic color names (e.g., Berlin \& Kay, 1969) and optimized focal or prototypical shades of color within a named category (e.g., Regier \& Kay, 2009). This has led to further research looking at categorical perception where colors are labeled differently based on their relation to the focal colors and the color boundaries of their language's color categories (e.g., Kay \& Kempton, 1984). An implication of this literature is that colors are not always experienced the same even if they are objectively the same wavelength of light. A speaker of Russian will see two shades as more different from each other when they cross the boundary between the Russian light blue and dark blue basic categories than an English speaker who would categorize them both as blue (Winawer et al., 2007). Color space has a distorted topography of similarity and difference based on the categories applied to it.

As implied by the term categorical perception, it is tempting to conclude that there is always an issue of categorization at play with color perception. We routinely perceive colors to be different than they actually are as in these color perception tasks. However, in the psychophysics literature, hue is highly memorable. Hue is one of three dimensions along with lightness and saturation that are used to describe color. Under some circumstances, participants can quite accurately reproduce the hue that they have just seen (Pérez-Carpinell, Baldoví, de Fez, and Castro, 1998). These results contrast with the psychological literature


Figure 1: Example colored silhouettes. The silhouettes are representative of the two color categories: red and green, as well as the two animacy categories: living and non-living.

Focused on distortion, instead suggesting that we do not always influence the colors we see with top-down knowledge but truly record hues as they objectively are in the world.

Color is a domain where memory has proven accurate to experience and is a domain where categories affect perception. The representational shift hypothesis can be tested in a domain where memory has shown to be accurate to experience without labels. By adding labels alongside hue perception we can see if representations of the colors do indeed shift.

## Experiment

The present experiment was designed to test the predictions of the representational shift hypothesis: There is a systematic prototypical shift of memory for overtly labeled objects, and preference-judged objects are not subject to the same shift. Preference-judged objects could lack a representational shift or at the least demonstrate a less strongly prototypical shift.

## Method

Participants 39 participants were recruited through UC Merced's participant pool. All participants were monolingual or early (by age 10) bilingual English speakers. Participants had normal or corrected to normal vision. They also had normal color vision, tested using the CITY color vision test (City University, 2002) at the conclusion of the experimental session.

Stimuli 40 colored silhouettes were created in Adobe Photoshop for the study phase of the experiment. All colors


Figure 2: A representation of CIE L*C*h color space. The variations in hue are calculated in degrees on the plane of lightness and the radius of saturation. (http://nyman.netsolution.ch/IT8FujiProvia.htm)
were calculated in device independent CIE $L^{*} a^{*} b^{*}$ color space and converted to RGB device dependent color space via unique monitor profiles created by a X-rite il Display Pro color calibrator to ensure color constancy across testing stations. 20 silhouettes were living things such as a giraffe and a butterfly, and 20 silhouettes were non-living objects such as a pan and an airplane (see Figure 1). 8 main hues, 4 reds and 4 greens were selected. The colors had the lightness and saturation values of their category's focal color (Sturges \& Whitfield, 1995). Each object was randomly assigned a hue of red and a hue of green. The silhouettes were randomly assigned to 4 groups of 10 objects, 5 living and 5 non-living. The groups were then assigned a color category between subjects (i.e. Participant 1 saw groups $1 \& 2$ in their red hues and groups $3 \& 4$ in their green hues while Participant 2 saw groups $1 \& 3$ in red and groups $2 \& 4$ in green, etc.) The semi-random creation of colored silhouettes preserves color and animacy balances while counterbalancing the color/shape pairings across participants.

For the testing phase of the experiment, four variations of each of the 8 main hues were calculated in CIE L*C*h color space (Figure 2) with a distance of $4^{\circ}$ along the hue dimension from its adjacent hue. A test scale for each silhouette of two steps more typical, one step more typical, the original, one step less typical, and two steps less typical was created.

## Procedure

The experiment consisted of two main parts: a judgment phase and a memory test. Each participant encountered 80 judgment trials followed by 40 memory trials. For the judgment phase, the participants had been instructed to


## Hue Variant Chosen by Judgment Condition



Figure 4: Average percent of each hue typicality chosen by participants. $20 \%$ chosen would be expected if participants were choosing at random.
remember the silhouettes as they would show up more than once but were not explicitly told of a memory test.

Initially, participants were instructed in the two types of judgments: color (" 1 " for red or " 2 " for green) and preference (" 3 " for like or " 4 " for dislike). Tags were placed above the keys on the keyboard to remind participants of the mappings mid-task. The trials were presented in alternating judgment blocks with 10 trials per block. Each silhouette was judged twice by each participant, both within the same judgment type (e.g., if the giraffe was judged for color the first time it appeared, the giraffe would be again judged for color the second time it appeared). The silhouettes were judged in each judgment condition between participants.
Each trial consisted of a fixation cross ( 1500 ms ), the silhouette to be judged ( 300 ms ), a question mark eliciting the judgment for that block ( 700 ms ), followed by a blank screen ( 1000 ms ).

After the judgment phase, participants were then tested for hue memory. The memory trials consisted of a circular array of the 5 hue variants of a particular silhouette (Figure 3). The array had the hues in graded clockwise order with the most typical hue rotated to a random position by trial resulting in a consistent appearance of selecting from a gradient of hues but avoiding position effects that would be present in a line. Each of the 5 positions had a location label 1 through 5 that participants entered on the keyboard to make their selection. There was no time limit imposed on the memory test responses with an intertrial interval of 1500 ms .

## Results

Not all participants proved equally skilled or motivated to complete this task. The following criteria needed to be met in order to include a participant's data in the analysis: at least $80 \%$ accuracy for color judgments (4 participants did
not meet this criteria), and responded using at least 3 of the 5 memory test positions ( 6 participants did not meet this criteria). As a result, 29 participants were included in the analyses.

As can be seen in Figure 4, there was a systematic shift of responses in the hue memory test. As described above, the position of the hues was randomized, disambiguating position responses from typicality. The judged stimuli selected were both clockwise and counterclockwise to the focal color in color space disambiguated hue space direction and the direction of typicality. Only when taking into account the relative direction in hue space of the focal color does the systematic bias emerge from the data.

There are three main questions to be addressed in the data analysis to test the representational shift hypothesis: Is there a shift of memory for color judged silhouettes? Is there a shift for preference judged silhouettes? Is there a difference between the two shifts should they both be observed?

To test whether there is a shift for items in the color judgment condition, a one sample t-test with a comparison value of 3 was performed on the average typicality value of the hue chosen at memory test. Responses lower than 3 are more typical of the color category, a response of 3 is true to the original color viewed, and response values higher than 3 are atypical of the color category. The mean response typicality for color judged silhouettes is 2.70 . The mean is significantly lower than 3 , indicating that the shift is in the prototypical direction, $\mathrm{t}(28)=-3.93, \mathrm{p}<.001$. In the color judgment condition, there is an observed representational shift is the prototypical direction as predicted by the representational shift hypothesis.

To test whether there is a representational shift for items in the preference condition, another one sample t-test with a comparison value of 3 was preformed. Again the mean response typicality is less than 3 at 2.56 indicating a prototypical representational shift, $\mathrm{t}(28)=-5.82, \mathrm{p}<.001$. The
representational shift hypothesis does not explicitly reject such a shift, but it does predict that the shift due to labeling should be stronger.

Finally, the mean is significantly greater for the color condition (2.70) than the preference condition (2.59), paired $\mathrm{t}(28)=2.05, \mathrm{p}<05$. The direction of the difference is actually the opposite of the prediction made by the representational shift hypothesis. There is more of a prototypical shift for preference-judged stimuli than color judged stimuli. Even if the two conditions being statistically different were a matter of chance, it is quite unlikely that they are truly different in the opposite direction.

## Discussion

As designed, this paradigm allows us to directly view representational shifts within hue color space. Participants were exposed to colored silhouettes then tested on 5 variations of the original silhouette. Participants on average chose hues that were slightly more prototypical of the basic color category than the hue that they had originally seen. This is the prototypical representational shift predicted by the representational shift hypothesis. However, the type of judgment made on the silhouette did not make a difference in whether a prototypical representational shift occurred. In contrast to the predictions of the representational shift account, participants chose a more prototypical hue in even larger proportions when they had made preference judgments about a silhouette rather than labeling the color category.

## Depth of Processing

In previous experiments there have been measurable differences between conditions (Blanco \& Gureckis, 2012; Lupyan, 2008; Richler et al. 2011, 2012) that were interpreted as superior memory for preference-judged items. The restriction to a single dimension for variation limited the ways in which preference-judged items could be uniquely encoded into memory to benefit recognition. For the dimension tested, the depth of processing predictions could have been turned around with color being more deeply processed with category judgments than with preference judgments. While making preference judgments, participants were not limited to opining on the color. It is plausible that they paid more attention to whether they liked giraffes and pans than the particular hue. Thus, a depth of processing account of the present results appears most plausible.

## Implications of Representational Shifts

Earlier we discussed a potential implication of representational shifts: a pervasive influence of past experience and existing categories on new representations. These representations then become part of the categories that proceed to influence representational shifts in future experiences. Rather than categories being an accumulation of raw experience, these distortions in how new exemplars
are encoded support existing category structures and discourage new categorization schemes from developing in well-categorized domains. As categorization schemes mature, the representational shifts would cascade, potentially reaching the extreme of the characteristically discrete looking categorical perception effect (Harnad, 1987; but see Huette \& McMurray, 2010). These implications hold for the representational shift effect regardless of the mechanism behind the shifts.

## Online Influence of Categories

Categories, regardless of overt labeling, affect memory. In the context of this experiment, hue memory was not as accurate as it has been reported in other research (e.g., Pérez-Carpinell et al., 1998). Given that categorization of objects is fairly automatic (Grill-Spector \& Kanwisher, 2005), having category judgments in the course of the experiment could have created a 'category-relevant' context where regardless of explicit categorization responses, categories were activated for all stimuli. The implication is that categories alter memory formation online rather than being a permanent perceptual bias consistent with recent research (e.g. Landau, Dessalegn, \& Goldberg, 2010). A non-category relevant context with added overt labels is a possible scenario where labeling may have an effect on representational shifts.

## Time Course

Our previous research (Kelly \& Heit, 2012) has shown that atypical representational shifts-shifts away from the prototype - are found with immediate (half second and five second delay) recognition tests. At some point between five seconds and a few minutes representations go from being distorted atypically to being distorted typically. Representations are not veridical to the stimulus initially or subsequently.
It's possible that these shifts are symptomatic of competing needs of working memory and long term memory. Working memory could privilege differentiating information in case details are important in the moment. Detailed information that will go unused would deplete resources unnecessarily when being encoded into long term memory, making relying on the category general information to supply a complete representation upon recall more advantageous. Variables that change depth of processing could be indicative of how likely specific differentiating information is to be needed in the future. These explanations for the mobility of representational shifts are purely speculative and need to be researched further.

## Push or Pull?

Lupyan (2008) hypothesized the mechanism of blended representations between exemplar and category prototype as the cause of prototypical representational shifts. Another proposed mechanism to account for this bias is boundary truncation (Huttenlocher, Hedges, Lourenco, Crawford, \&

Corrigan, 2007). Rather than having a pull toward the prototype via prototype activation, there is a push toward the prototype by disregarding extreme information. The shift into the category would be consistent with both theories and needs to be explored further.

## Conclusion

Representational shifts exist in memory for hue. There are prototypical shifts in memory for colored silhouettes encountered minutes before test. These shifts can be seen and measured on a unidimensional testing ground. Conceptual space has long been known to be contorted in color (Kay \& Kempton, 1984) and beyond (Goldstone, 1992). Our results support the idea that the creation and maintenance of these contortions could be due to representational movement.

## References

Berlin, B., \& Kay, P. (1969). Basic Color Terms: Their Universality and Evolution. Berkeley: University of California Press.
Blanco, N. and Gureckis, T.M. (2012). Does category labeling lead to forgetting? Cognitive Processing.
City University (2002). A web-based colour vision test. http://www.city.ac.uk/avrc/colourtest.html.
Goldstone, R. L. (1998). Perceptual learning. Annual review of psychology, 49(1), 585-612.
Grill-Spector, K., \& Kanwisher, N. (2005). Visual recognition: As soon as you know it is there, you know what it is. Psychological Science, 16, 151-160.
Harnad, S. R. (Ed.). (1990). Categorical perception: The groundwork of cognition. Cambridge, UK: Cambridge University Press.
Heit, E. (1997). Knowledge and concept learning. In K. Lamberts \& D. Shanks (Eds.), Knowledge, Concepts, and Categories, 7-41. London: Psychology Press.
Huette, S. \& McMurray, B. (2010). Continuous dynamics of color categorization. Psychonomic Bulletin \& Review, 17(3), 348-354.
Huttenlocher, J., Hedges, L. V., Lourenco, S. F., Crawford, L. E., \& Corrigan, B. (2007). Estimating stimuli from contrasting categories: Truncation due to boundaries. Journal of Experimental Psychology: General, 136, 502519.

Kay, P., \& Kempton, W. (1984). What is the Sapir-Whorf hypothesis?. American Anthropologist, 86(1), 65-79.
Kelly, L.J. \& Heit, E. (2012, November). Representational Shifts in Memory for Hue. Poster presented at the $53^{\text {rd }}$ Annual Meeting of the Psychonomic Society, Minneapolis, Minnesota.
Landau, B., Dessalegn, B., \& Goldberg, A.M. (2010). Language and space: Momentary interactions. In P. Chilton and V. Evans (Eds.), Language, cognition and space: The state of the art and new directions. Advances in Cognitive Linguistics Series (Ed, V. Evans, B. Bergen, J. Zinken). London: Equinox Publishing.

Lupyan, G. (2008). From chair to "chair:" A representational shift account of object labeling effects on memory. Journal of Experimental Psychology: General, 137(2), 348-369.
Lupyan, G., Rakison, D.H., \& McClelland, J.L. (2007). Language is not just for talking: Labels facilitate learning of novel categories. Psychological Science, 18(12), 10771083.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. Journal of Experimental Psychology: General, 115(1), 39.
Palmeri, T. J., \& Nosofsky, R. M., (2001). Central tendencies, extreme points, and prototype enhancement effects in ill-defined perceptual categorization. The Quarterly Journal of Experimental Psychology, 54, 197235.

Pérez-Carpinell, J., Baldoví, R., de Fez, M. D. and Castro, J. (1998). Color memory matching: Time effect and other factors. Color Research and Application, 23, 234-247.
Regier, T. \& Kay, P. (2009). Language, thought, and color: Whorf was half right. Trends in Cognitive Science, 13, 439-446.
Richler, J.J., Gauthier, I., \& Palmeri, T.J. (2011). Automaticity of basic-level categorization accounts for labeling effects in visual recognition memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 37, 1579-1587.
Schmolck, H., Buffalo, E. A., \& Squire, L. R. (2000). Memory distortions develop over time: Recollections of the OJ Simpson trial verdict after 15 and 32 months. Psychological Science, 11(1), 39-45.
Sturges, J. \& Whitfield, T.W.A. (1995). Locating basic colours in the munsell space. Color Research and Application, 20, 364-376.
Wells, G. L., \& Olson, E. A. (2003). Eyewitness testimony. Annual Review of Psychology, 54(1), 277-295.
Winawer, J., Witthoft, N., Frank, M.C., Wu, L., Wade, A.R., \& Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. Proceedings of the National Academy Sciences 104, 7780-7785.

# Adaptive Foraging: Effects of Resource Conditions on Search Paths in a Web-Based Foraging Game 

Bryan Elvis Kerster (bkerster@ucmerced.edu)<br>Christopher T. Kello (ckello@ucmerced.edu)<br>Cognitive and Information Science, 5200 N. Lake Road<br>Merced, CA 95343 USA<br>Theo Rhodes (theorhodes@gmail.edu)<br>Cognitive Science and Psychology, 7060 Route 104<br>Oswego, NY 13126 USA

Ralph Jerry Bien-Aime (rbien-aime@ucmerced.edu)<br>Cognitive and Information Science, 5200 N. Lake Road Merced, CA 95343 USA


#### Abstract

Foraging is a search process common to mobile organisms, and foraging paths commonly exhibit statistical patterns akin to Lévy walks. There may be common factors and benefits underlying these patterns, but investigations are hindered by difficulty in assessing and manipulating search environments and task conditions. In the present study, a simple foraging game was developed to isolate and manipulate two factors hypothesized to make Lévy walks adaptive search strategies-sparsity, and spatial clustering of targets in the search environment. Players navigated a fuel-limited ship over a 2D grid to find as many targets as possible, rendered as asteroids in outer space. Over 1800 participants were recruited to play using Amazon's Mechanical Turk, in order to widely sample the parameter space defined by degrees of target sparsity and clustering. Observed search paths resembled Lévy walks with memory, and those of high performers were found to vary adaptively with clustering, but not sparsity. Results indicate that Lévy-like walks can emerge from search strategies and algorithms adapted to environments with clustered resources.


Keywords: Foraging; Lévy walks; crowdsourcing; adaptive search.

## Introduction

Perhaps the most ancient kind of search function in biological organisms, in terms of evolutionary history, is foraging-moving about one's environment in search of resources like food, locales like shelter, or other organisms like mates. Studies of animal foraging have found that foraging paths tend to resemble Lévy walks (Viswanathan et al., 1996). Paths are clustered such that most path segments are relatively short, but they are interspersed with longer segments, occasionally much longer. Paths resemble Lévy walks in that distributions of path lengths follow an inverse power law, $\mathrm{P}(l) \sim 1 / l^{\alpha}$, where $\alpha \sim 2$. Lévy-like paths are observed for foragers from bacteria (Berg, 1993) to humans (Rhee, Shin, Hong, Lee, \& Kim, 2011).

Lévy foraging paths can be modeled simply as random walks with path lengths sampled from a power law, although path directions may be correlated over time (Viswanathan et al., 2001), and mechanisms of navigation
are left unspecified. Despite their simplicity, Lévy walks have proven influential because they suggest that search benefits are conferred by power law path lengths, at least under certain foraging conditions (Viswanathan \& Buldyrev, 1999). In particular, when $\alpha \sim 2$ and targets are sparsely and randomly distributed, Lévy walks maximize the rate of finding targets compared with Gaussiandistributed random walks.

The potential benefits of Lévy walks recently have led cognitive scientists to investigate whether they occur in perceptual, memory, and decision-making search tasks. First were Rhodes and Turvey (2007), who investigated Lévy walks in a classic category recall paradigm (Bousfield \& Sedgewick, 1944). Participants recalled as many animals as they could from long-term memory, for twenty minutes. Inter-response intervals were used as indirect measures of memory "path lengths", and they were found to be best fit by inverse power law functions with exponents near two. Then, Rhodes, Kello, and Kerster (2011) found that saccade lengths in visual foraging tasks also followed a heavy-tailed distribution resembling efficient Lévy walks, although the lognormal function provided the best fit to data. A lognormal can be viewed as a constrained power law (Stephen \& Mirman, 2010), which should be expected when search is constrained to a relatively small space (a computer monitor). Most recently, Radicchi and Baronchelli (2012) found search intervals to be Lévy-like when buyers searched the bid space in online auctions, and observed exponents were shown to maximize economic gains.

These and other similar studies raise the question of what mechanisms and factors give rise to Lévy-like search paths across so many different species and foraging conditions. Theoretical analyses suggest that sparsity of targets is a factor, but it is prohibitively difficult to test this hypothesis in natural foraging conditions, including visual and memory foraging of natural scenes and categories. Also, most theoretical analyses have assumed randomly distributed targets (Viswanathan \& Buldyrev, 1999), but food and other resources may instead tend to be clustered in nature, as is the case with plankton distributions, for instance (Mackas \& Boyd, 1979). In terms of mechanism, Lévy-like foraging
may be intrinsic to both biological and cognitive search functions, in which case Lévy walks would occur regardless of search conditions. Alternatively, Lévy walks may emerge as a result of interactions between search processes and their environments. These interactions may unfold over the course of minutes and even faster timescales.

In the present experiment, we examined the roles of sparsity and clustering in a web-based video game designed to mimic canonical foraging. We used a video game because it allowed us to know and manipulate search conditions. We made the game web-based to collect data from large numbers of participants on Amazon's Mechanical Turk. Recent studies show that Turk yields data comparable to university participant pools (Germine et al., 2012; Snow, O’Connor, Jurafsky, \& Ng, 2008), and we confirmed this in a pilot study run through the UC Merced participant pool (not reported). Turk allowed us to robustly sample the parameter space created by factorial manipulation of sparsity and clustering. Turk also allowed us to collect enough data to compare players who find greater versus fewer numbers of targets, i.e. high versus low performers.

Comparisons of high versus low scorers are critical because they test whether adaptive search is associated with closer-to-optimal (i.e. higher scoring) performance. That is, do high performers adapt their search strategies, as measured by path length distributions, to changes in sparsity and clustering? Do high scoring foraging paths more closely resemble Lévy flights with the theoretically optimal exponent of two? Addressing these questions will provide evidence on 1) whether Lévy-like foraging paths can emerge from searcher-environment interactions on timescales no longer than minutes, and 2) whether adapting paths to sparsity or clustering is associated with better performance.

## Methods

The foraging game was framed as a task of exploring outer space to find resources on asteroids (see Figure 1, and http://cogmech.ucmerced.edu/downloads.html to play). Participants used a mouse (or functionally equivalent device) to move a spaceship over a $1280 \times 1024$ grid of space. Movement was controlled at two scales, zoomed in and zoomed out. When zoomed out, the entire space was visible at once, and participants clicked on a location to "fly" the ship to that spot (shown by animation). Participants pressed the space bar to zoom in 15X at a given location, at which point they again could navigate the ship via point-and-click. Hubble images were used as background to help engage players by giving the sense of outer space, and to provide environmental cues that are, in general, ubiquitous to natural search conditions. These cues may encourage use of memory in navigation (e.g. Vinson, 1999), which shall be discussed later.

Asteroids were visible only when zoomed in, and resources were harvested by moving to them such that collision occurred between the asteroid and ship graphics. A set amount of fuel was provided for each play of the game,
and fuel usage (shown by a fuel bar) was a linear function of distance traveled, plus a small constant for each zoom in/out. The amount of fuel provided was determined based on pilot work to allow for about 5 minutes per play, and to enable players to find some but not all asteroids. Each successful harvest was indicated by sight and sound, and asteroids could only be harvested once (the un/harvested status of asteroids was not displayed). Each harvest added one point to the score (no fuel was added), and play continued until all fuel was expended.


Figure 1: Example game space shown zoomed out (above), along with the corresponding zoomed in view (below).
Current score is shown in upper left corner, and fuel bar with remaining fuel in red is shown in upper right corner.

The game was designed to mimic foraging as exemplified by aquatic birds hunting for fish, or the eyes scanning a scene to gather visual information. Relatively short movements are made during resource acquisition (while in the water and close to the surface, or during fixations), interspersed with longer-scale movements when no resources are acquired (while flying high above the water, or during saccades when visual information uptake is attenuated; Ross, Morrone, Goldberg, \& Burr, 2001). Foraging costs in natural searches (e.g. risk and energy expenditure) were lumped into the fuel cost of travel, although time costs were also a factor, given the natural tendency to minimize time spent foraging. The game was coded in Flash so it could be distributed via the web, with game data collected on a local server. We
used Amazon’s Mechanical Turk to find people willing to play the game twice (plus a 1 minute practice session to learn) for 75 cents in compensation. Pilot work indicated that the availability and quality of Turk workers fell off precipitously for tasks lasting more than 10-15 minutes. A step-by-step demo, along with instructions in English, was presented at the beginning of play, and each play was set to last about 5 minutes (assuming no breaks). Two plays were required for each paid work session.


Figure 2: Example of 150 asteroids clustered at 0.05, 0.15, 0.25 , and 0.5 , left to right, top to bottom.

The number of asteroids per play was set at four different levels: 25, 50, 100, and 150. Pilot work indicated that 25 asteroids meant that players occasionally found only a few of them (or even none), and 150 meant that players found asteroids nearly every time the zoomed in. Clustering of asteroids was manipulated at four different levels of a probabilistic parameter: $0.05,0.15,0.25$, and 0.5 . This parameter controlled the probability of dividing asteroids evenly (0.5) or entirely to one side (0.0) in an algorithm that divided a given set of asteroids recursively into alternating horizontal and vertical splits of a given 2D space. Asteroids were placed when only one remained in a given recursively split section of the space (placed at random in the section), and/or when the space could be split no further (see Figure 2 for example asteroid distributions, and Figure 3 for pseudocode). This algorithm created clusters whose sizes followed a nested scaling relation to varying degrees, consistent with findings of scaling law clustering of natural resources (Humphries et al., 2010; Mackas \& Boyd, 1979). The algorithm also created asteroid distributions that were independent of Hubble image backgrounds, and participants were informed of this independence during the demo/ instruction period.

The full $4 \times 4$ factorial of sparsity and clustering levels was tested. Each participant played twice in only one of the 16 possible conditions, chosen at random at the start of each Turk session. The demo and instructions included an
example asteroid distribution for the condition the participant was in, to help them formulate an informed foraging strategy. Players were also encouraged to achieve the best score possible by maintaining a high score board, and allowing high scorers to enter their initials for display to other players.

```
function distributeResources(rectangle, prob_split)
{
    // stop when no stars or no space left
    if (rectangle.stars_remaining < 1) return
    else if (rectangle.size < 1 pixel) {
        place remaining stars at pixel
        return
    }
    else if (rectangle.stars_remaining == 1) {
        place star randomly in rectangle
        return
    }
    // split rect in half, alternate between vert and horiz
    (rectangle1,rectangle2) = splitRectangle(rectangle,alternate)
    // randomize bias for placing stars in each half
    if (random_prob() < 0.5) prob_split = 1 - prob_split
    for each star {
        if (random_prob() < prob_split) place star in rectangle1
        else place star in rectangle2
    }
    distributeResources(rectangle1, prob_split)
    distributeResources(rectangle2, prob_split)
}
```

Figure 3: Pseudocode for asteroid distribution algorithm.
Note that a smaller parameter leads to increased clustering.

## Results

A total of 1,825 game sessions were administered on Turk. Participants who did not produce more than 80 zoomin actions per play were excluded from analysis (603 participants). Pilot work indicated that participants who simply expended fuel to complete the task, rather than endeavored to find asteroids, were revealed by low numbers of zoom-in actions. Of the remaining 1,222 participants, 393 played in two or more Turk sessions. Analyses with and without these repeats indicated no qualitative change in results, so both were included in the reported results. Analyses combine zoomed in and zoomed out path lengths.

Visual inspection of zoomed out flight paths revealed directional movements that ranged in their temporal correlations, which express a very simple memory (i.e. effect of history) in search paths. Two example paths at the two ends of this range are shown in Figure 4. Paths that consisted of highly regular directional movements were seen as "sweep" strategies designed to systematically cover the space in left-right, top-down, spiral, and other search patterns. Other paths consisted of apparently haphazard directional movements, akin to random walks. In the middle were mixtures of the two, plus directional movements that followed irregular contours of Hubble images (despite instructions that distributions were independent of images).

To minimize effects of practice and learning, only the second of two plays per Turk session was analyzed. Performance was measured as the proportion of available targets found, and plays were divided into three categories of performance, for each of the 16 game conditions: Top 20, middle 20, and bottom 20 scores. More than 60 Turk
sessions were randomly assigned to each game condition, so any additional plays were excluded from reported analyses. The constant of 20 plays per cell simplified statistical analyses, and excluding intermediate performances helped to further distinguish our three categories.


Figure 4: Two example flight paths of more versus less directionally correlated movements (top versus bottom). Red and green lines indicate zoomed out and zoomed in movements, respectively. Blue dots indicate clicked locations, and yellow dots indicate points of harvest. White boxes indicate areas of zoom in.

All results are graphed and analyzed as a function of sparsity, clustering, and performance category. A three-way analysis of variance was conducted for each dependent measure, but we report only main effects and two-way interactions relevant to our research questions and hypotheses. First, we examined score as function of sparsity, clustering, and performance levels (Figure 5). The main effect of performance category is itself based on score, and is so large throughout our analyses that reporting its reliability was unnecessary.

As for the other two main effects, sparsity was not reliable, $\mathrm{F}(3,18)=2.1, \mathrm{p}>0.1$, but clustering was, $\mathrm{F}(3,18)=$ 38.42, $\mathrm{p}<.05$. The interaction of performance level with sparsity was also not reliable, $F(6,18)=0.84, \mathrm{p}>0.5$, but it did interact with clustering, $\mathrm{F}(6,18)=55.41, \mathrm{p}<.05$. Visual inspection shows that scores improved with clustering for high performers, but the opposite effect occurred for low
performers. These results show that foragers adapted to clustering but not sparsity, and low performers appeared to adapt counterproductive strategies in terms of score. Thus we have initial evidence that high performers took advantage of the spatial correlations in clustering, suggesting that foraging paths adapted based on interactions between search processes and game conditions.


Figure 5: Proportional score as a function of sparsity, clustering, and performance category.

However, one might argue for an alternate interpretation of the data. It may be that each forager chooses a strategy a priori without regard to conditions, and effects of clustering merely show that strategies matter more for greater clustering. We tested this possibility by examining the change in score from first to second play. We found that score increased over time for high performers ( $+19 \%$ ), but decreased for low performers ( $-14 \%$ ), $t(638)=17.7, \mathrm{p}<.01$. This difference suggests that strategies changed over the course of play, for better or worse, indicating that strategies were indicative of interactions between search processes and game conditions.


Figure 6: Path Length as a function of sparsity, clustering, and performance category.

Next we examine mean path length, shown in Figure 6. High performers had shorter path lengths overall, reflecting the fact that shorter path lengths allowed for greater coverage of the space by reducing zoom costs. This main
effect would be expected to diminish, and possibly even reverse, if fuel costs were greater per zoom.

There were no main effects of sparsity on mean path length, $\mathrm{F}(3,18)=2.19, \mathrm{p}>.1$, or clustering, $\mathrm{F}(3,18)=1.7$, p $>$.2. Once again, the interaction of performance category with sparsity was not reliable, $\mathrm{F}(6,18)=1.66, \mathrm{p}>.15$, but it was with clustering, $\mathrm{F}(6,18)=4.61, \mathrm{p}>.05$. Visual inspection shows that path lengths for high performers increased with clustering, whereas they decreased for low performers. The increase for high performers presumably reflects the increased need for larger jumps as clusters became sparser. Again, low performers appeared to adjust strategies as well, but in counterproductive ways.

Next we examine whether foraging paths resembled Lévy walks, in the sense that path length distributions were power law distributed with estimated exponents near two. We used multi-model inference (Symonds \& Moussalli, 2010) to test which of four different functions provided the best fit to the distribution of path lengths for each participant (mean of 217 path segments per participant): Normal, exponential, lognormal, and Pareto. Only the latter two are heavy-tailed and Lévy-like, and the method uses Akaike's information criterion (AIC) to find the function with the shortest information-theoretic distance to the data.

The lognormal function provided the best fit for $68 \%$ of the participants, with the remaining trials roughly evenly split between normal and exponential fits (Pareto never provided the best fit). As mentioned earlier, the lognormal is akin to a constrained power law, and the foraging game constrained movements in terms of a limited amount of space. An example distribution from one participant is plotted in Figure 7 in logarithmic coordinates, which is representative of the majority of participants. The constrained, normal-like portion of the distribution is seen as a slight hump on the left side, and the power law-like tail is seen as a negatively sloped line on the right.


Figure 7: An representative path length distribution for one participant, plotted in logarithmic coordinates

We further examined the tails of path length distributions in two ways. First, the fitted lognormal functions have $\mu$ and $\sigma$ parameters, where the latter roughly corresponds with the heaviness of the tail. Heavier tails indicate more Lévy-like distributions. Best-fitting $\sigma$ 's are shown in Figure 8 for all
participants, including those whose data were better fit by normal or exponential distributions.

The overall pattern of results was similar to those reported earlier. There was no main effect of sparsity, $\mathrm{F}(3,18)=1.17$, $p>.35$, but clustering was again reliable, $\mathrm{F}(3,18)=10.52$, p $<.05$. Visual inspection reveals the possibility of an effect in the high/mid performers which was supported by a reliable interaction of performance category with clustering, $F(6,18)=12.29, p<.05$. Once again, there was no reliable interaction with sparsity, $\mathrm{F}(6,18)=1.77, \mathrm{p}>.15$. Visual inspection shows that the tails of path length distributions were heavier overall for low performers, but they became heavier with greater clustering only for high performers.

To gauge whether distributions were becoming more similar to the theoretical power law exponent of two, we fit regression lines (see Figure 7) to the right half of distributions in logarithmic coordinates, and results are shown in Figure 9. None of the main effects (excluding performance category) or interactions were reliable, but slopes were generally in the neighborhood of the theoretical optimum of -2 (negative of the optimal exponent). Moreover, slopes for high performers were closest in their approach towards -2 with greater clustering.


Figure 8: Lognormal $\sigma$ as a function of sparsity, clustering, and performance category.

To summarize, search paths generally resembled efficient Lévy walks as predicted, at least to some degree. The majority of path length distributions were heavy-tailed, and tails resembled truncated power laws with exponents near two. Distributions most closely resembled Lévy walks for high performers in the most clustered resource conditions. Foraging paths were not like random walks, in that path directions were never drawn purely at random. Instead, path directions tended to be correlated over time. This tendency can be quantified simply by computing the proportion of times that next steps went in the same direction, within some threshold. Angular changes were between 0 and 180 degrees (collapsing left versus right turns) and divided evenly into 45 bins. The proportion of movements falling into the smallest angular bin was $25 \%$, about ten times greater than chance. Thus foraging paths had memory in that the direction of each step was sensitive to previous steps.


Figure 9: Regression slopes as a function of sparsity, clustering, and performance category.

## Discussion

In the present study, analyses of data from a web-based foraging game showed that Lévy-like search paths emerge from search processes that change depending on the clustering of resources. This was true for both high and low performers, although only high performers changed their search processes adaptively. The lack of a sparsity effect was conspicuous because prior theoretical analyses suggest that the benefits of Lévy-like search paths are most prevalent when items are rare to be found. However, prior analyses focused on random Lévy walks, whereas observed foraging paths clearly had memory. Memory sometimes manifested as "sweep" searches, but more generally, directions of next steps depended on previous steps.

Our results suggest that the prevalence of Lévy-like walks in both animal and cognitive searches can be better modeled by processes with memory that attend to target and task conditions, as opposed to random walk processes. The observed effects of clustering indicate that search processes take advantage of spatial correlations in resource distributions when they exist and are known or learned. A simple approach to modeling an effect of spatial correlations is for search processes to follow a gradient of resource density. These and related modeling ideas (e.g. Ferreira, Raposo, Viswanathan, \& da Luz, 2012) are potentially interesting topics for future research.

## Acknowledgments

This work was supported by a grant from the National Science Foundation, BCS 1031903 (PI Kello). The authors thank members of the Cognitive and Information Sciences graduate group for helpful comments and feedback.

## References

Berg, H. C. (1993). Random walks in biology. Princeton University Press.
Bousfield, W. A., \& Sedgewick, C. H. W. (1944). An analysis of sequences of restricted associative responses. Journal of General Psychology.
Ferreira, A. S., Raposo, E. P., Viswanathan, G. M., \& da
Luz, M. G. E. (2012). The influence of the environment
on Lévy random search efficiency: Fractality and memory effects. Physica A: Statistical Mechanics and its Applications, 391(11), 3234-3246.
Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., \& Wilmer, J. B. (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. Psychonomic bulletin \& review.
Humphries, N. E., Queiroz, N., Dyer, J. R. M., Pade, N. G., Musyl, M. K., Schaefer, K. M., ... Sims, D. W. (2010). Environmental context explains Lévy and Brownian movement patterns of marine predators. Nature, 465(7301), 1066-9.
Mackas, D. L., \& Boyd, C. M. (1979). Spectral analysis of zooplankton spatial heterogeneity. Science, 204(4388), 62.

Radicchi, F., \& Baronchelli, A. (2012). Evolution of optimal Lévy-flight strategies in human mental searches. Physical Review E, 85(6), 1-8.
Rhee, I., Shin, M., Hong, S., Lee, K., \& Kim, S. (2011). On the levy-walk nature of human mobility. /ACM Transactions on, 19(3), 630-643.
Rhodes, T., \& Turvey, M. (2007). Human memory retrieval as Lévy foraging. Physica A: Statistical Mechanics and its Applications, 385(1), 255-260.
Ross, J., Morrone, M., Goldberg, M., \& Burr, D. (2001). Changes in visual perception at the time of saccades. Rev. Neurosci, 24(2), 113-21.
Snow, R., O’Connor, B., Jurafsky, D., \& Ng, A. (2008). Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks. ... Methods in Natural Language ..., (October), 254-263.
Stephen, D. G., \& Mirman, D. (2010). Interactions dominate the dynamics of visual cognition. Cognition, 115(1), 154165.

Symonds, M. R. E., \& Moussalli, A. (2010). A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike’s information criterion. Behavioral Ecology and Sociobiology, 65(1), 13-21.
Vinson, N. (1999). Design guidelines for landmarks to support navigation in virtual environments. Proceedings of the SIGCHI conference on Human ....
Viswanathan, G., Afanasyev, V., Buldyrev, S., Murphy, E., Prince, P., \& Stanley, H. E. (1996). Lévy flight search patterns of wandering albatrosses. Nature, 381(6581), 413-415.
Viswanathan, G., Afanasyev, V., Buldyrev, S. V., Havlin, S., Luz, M. G. E. da, Raposo, E. P., \& Stanley, H. E. (2001). Lévy flights search patterns of biological organisms. Physica A: Statistical Mechanics and its Applications, 295, 85-88.
Viswanathan, G., \& Buldyrev, S. V. (1999). Optimizing the success of random searches. Nature, 401(6756), 911.

# Neural-network Modelling of Bayesian Learning and Inference 

Milad Kharratzadeh (milad.kharratzadeh@mail.mcgill.ca)<br>Department of Electrical and Computer Engineering, McGill University, 3480 University Street<br>Montreal, QC H3A 2A7 Canada

Thomas R. Shultz (thomas.shultz@mcgill.ca)<br>Department of Psychology and School of Computer Science, McGill University, 1205 Penfield Avenue Montreal, QC H3A 1B1 Canada


#### Abstract

We propose a modular neural-network structure for implementing the Bayesian framework for learning and inference. Our design has three main components, two for computing the priors and likelihoods based on observations and one for applying Bayes' rule. Through comprehensive simulations we show that our proposed model succeeds in implementing Bayesian learning and inference. We also provide a novel explanation of base-rate neglect, the most well-documented deviation from Bayes' rule, by modelling it as a weight decay mechanism which increases entropy. Keywords: Neural-network; Bayes’ rule; Bayesian learning and inference; base-rate neglect; weight decay; entropy


## Introduction

Bayesian models are becoming prominent across a wide range of problems in cognitive science including inductive learning (Tenenbaum, Kemp, \& Shafto, 2006), language acquisition (Chater \& Manning, 2006), and vision (Yuille \& Kersten, 2006). While these Bayesian ideas provide computation level models, it is beneficial, and sometimes necessary, to appeal to some implementation-level (biological) models to explain human behaviour. Connectionist approaches provide a neural-based model of cognitive processes.

There is growing evidence in neuroscience supporting the relevance of Bayesian models on a neural level (Doya, Ishii, Pouget, \& Rao, 2007). Many perceptual and sensorimotor tasks that are learned and performed by the central nervous system can be described in a Bayesian framework (Konrad \& Wolpert, 2004). Neural computations, as simple as summing up the firing rates, can be seen as analogous to a Bayesian inference process, with population activity patterns encoding posterior distributions (Pouget, Dayan, \& Zemel, 2003).

In theoretical terms, connectionist models show promising prospects in implementing computational-level Bayesian ideas. Under certain assumptions, and inspired by the nature of neural activation functions, neural units can compute posterior probability values (McClelland, 1998). So-called Boltzmann machines, a type of stochastic recurrent neural network, were also suggested to implement Bayesian learning (Ackley, Hinton, \& Sejnowski, 1985). Later work introduced generative networks called restrictive Boltzmann machines that can make bidirectional inferences based on what they learned (Hinton \& Osindero, 2006).

All these connections between Bayesian and neuralnetwork models motivate further exploration of the relation between the two. In this paper, we propose a complete, modular neural-network structure implementing Bayesian learn-
ing and inference in a general form. We do this by using three main modules, two responsible for computing priors and likelihoods based on observations, and one responsible for applying Bayes rule and computing the posteriors. We show that our model is able to successfully implement Bayesian learning and inference and replicate analytical results with high precision in a brain-like fashion which could later be used to gain intuition into how brains implement Bayesian reasoning. Our work also provides a framework to study the deviations from optimal Bayesian reasoning which result from base-rate neglect (Kahneman \& Tversky, 1996; Eddy, 1982).

Our work is novel in its precise and complete implementation of a Bayesian framework in a modular, brain-like fashion. The proposed network takes observations as inputs and computes the posterior probabilities (i.e., updated beliefs). Moreover, using a fast, constructive learning algorithm (sibling-descendent cascade-correlation) for the network provides the advantage of a self-organizing learning and inference method which is similar to humans' developmental, autonomous inference and learning (Shultz \& Fahlman, 2010). Another novelty of this work is the modelling of base-rate neglect as a weight decay mechanism.

The idea of neural implementation of Bayesian phenomena was suggested before. Shi and Griffiths (2009) introduced a scheme for implementing importance sampling with radial basis function (RBF) networks. They assume that the unit activation functions are of radial basis type. Hence, in their model, a single RBF neuron measures likelihood, resulting in a straightforward implementation of importance sampling. Also, Griffiths et al. (2012) used a linear network to approximate the generalization performance of a probabilistic model. Their linear networks produce different solutions from Bayes on structures other than those based on Wishart priors. Our work differs in two major ways. First, we assume that the characteristics of the likelihood/priors and the network's structure are unknown a priori and learn them for a wide range of likelihood and prior distributions through a constructive training phase. Also, these functions are learned precisely (i.e., no approximation) using a population of neurons with simple though realistic activation functions (sigmoid) rather than linear or complex ones (RBF).

## The Basics of Bayesian Learning and Inference

The Bayesian framework addresses the problem of updating beliefs and making inferences based on observed data. As-
sume that we have a set of mutually exclusive and exhaustive hypotheses, $\mathcal{H}=\left\{h_{1}, \ldots, h_{N}\right\}$, and want to infer which of these hypotheses best explains observed data. In this setting, we denote the degree of belief in different hypotheses by probabilities. Bayesian inference is based on a simple formula known as Bayes' rule. This rule specifies how posterior probabilities (of a hypothesis being true given the observed data) can be computed using the product of data likelihood and prior probabilities:

$$
\begin{equation*}
P\left(h_{i} \mid d\right)=\frac{P\left(d \mid h_{i}\right) P\left(h_{i}\right)}{P(d)}=\frac{P\left(d \mid h_{i}\right) P\left(h_{i}\right)}{\sum_{i=1}^{N} P\left(d \mid h_{i}\right) P\left(h_{i}\right)} . \tag{1}
\end{equation*}
$$

Priors, $P\left(h_{i}\right)$, represent how much we believe in a hypothesis before observing data. Likelihoods, $P\left(d \mid h_{i}\right)$, denote the probability with which we would expect to observe these data if a hypothesis were true. The denominator, known as marginal probability of data, is a normalizing sum which ensures that the posteriors for all hypotheses sum to one.

The Bayesian framework is generative. This means that observed data are generated by an underlying mechanism or hypothesis. Then, the role of inference is to evaluate different hypotheses and choose the one which is the most likely mechanism responsible for generating the data (i.e., the one with the highest posterior probability). These generative processes can be specified by probabilistic models. Here, we describe likelihoods and priors respectively with probability distributions and mass functions as their generative mechanisms.

## Proposed Connectionist Model

We construct a modular neural network implementing Bayesian learning and inference. The algorithm we use to build our artificial neural modules is a variant of the cascadecorrelation (CC) method called sibling-descendant cascadecorrelation (SDCC) which is a constructive method for learning multi-layer artificial neural networks (Baluja \& Fahlman, 1994). CC offers two major advantages over standard backpropagation (BP) methods. First, it constructs the network in an autonomous fashion (i.e., a user does not have to design the topology of the network). Second, its greedy learning mechanism can be orders of magnitude faster than the standard BP algorithm. In addition to these, SDCC has another important benefit; due to its design, it can often reduce the depth of the network drastically (Baluja \& Fahlman, 1994).

We build our model in a modular fashion. Module 1 (shown in Fig. 1) implements Bayes' rule. For this module, we assume that there are two hypotheses to compare; extending it for any finite number of hypotheses is straightforward. There are three inputs (the prior and the two likelihoods) and one output (the posterior), and the module learns and implement (1). We run this module once for each hypothesis.

When data are observed in consecutive rounds, posteriors at one round are taken as priors for the next round; this is how the beliefs are updated in light of new observed data (rational inductive inference). Therefore, by connecting the output of module $1, P\left(h_{1} \mid d\right)$, to the input corresponding to the prior, $P\left(h_{1}\right)$, we can model this aspect of human cognition.


Figure 1: Module 1 computes the posterior based on the likelihoods and prior and according to Bayes' rule.

Module 1 assumes that the values of prior and likelihood probabilities are given and then applies Bayes' rule using them as input. Module 2 (shown in Fig. 2) is responsible for computing the likelihoods. It takes observation(s) as input(s). Due to the generative nature of the Bayesian framework, we describe likelihoods as probability distributions. The role of module 2 is to learn these distributions as the underlying mechanisms generating data. However, this should be done without any implicit or explicit knowledge about the specifications of the distribution and solely based on the observed training data.


Figure 2: Module 2 computes the likelihoods based on the observed data.

We denote the generative process (probability distribution) as a function of the observed data and hypothesis, $f(d, h)$. For example, if the likelihood distribution for hypothesis $h$ is a Gaussian with average $h$ and standard deviation 1, then:

$$
\begin{equation*}
f(d, h)=\frac{1}{2 \pi} e^{\frac{(d-h)^{2}}{2}} \tag{2}
\end{equation*}
$$

Finally, module 3 (shown in Fig. 3) computes the hypotheses' priors by learning their generative discrete distribution function. Note that the input to this module can be chosen from a finite number of possible hypotheses, $\mathcal{H}=$ $\left\{h_{1}, \ldots, h_{N}\right\}$, and hence the generative distribution is discrete. It takes the hypotheses as input and gives their prior probability as output. We represent the generative mechanism of priors as a probability mass function denoted by $g(h)$. For instance, this function can be of the following form:

$$
\begin{equation*}
g(h)=\alpha \cdot e^{\frac{h^{2}}{2}} \tag{3}
\end{equation*}
$$

where $\alpha$ is chosen so that the sum of priors equals 1 .


Figure 3: Module 3 computes the prior probabilities.

If we have $N$ hypotheses, to run a complete Bayesian learning and inference, we learn modules 1 and 3 one time and use them $N$ times (for respectively computing the posterior and prior of each hypothesis). However, since the likelihood distributions might be different for different hypotheses, we learn $N$ different units of module 2 and use each of them once.

## Simulation Results

## Setup

We test each module individually. Using SDCC, we first train each module by presenting input(s)-output pairs as training patterns. Then, we test the generalization ability of the learned module by utilizing a set of input(s)-output testing patterns. The accuracy of the module's outputs is examined by comparing them with the correct outputs presented in the testing set. In all our modules, the hidden units have sigmoid activation functions and the output units have linear activation functions.

## Module 1

Module 1 takes the two likelihoods, $P\left(d \mid h_{1}\right)$ and $P\left(d \mid h_{2}\right)$, and the prior, $P\left(h_{1}\right)$ as inputs and gives the posterior, $P\left(h_{1} \mid d\right)$, as the output. The training set is all the triplets starting from ( $0.1,0.1,0.1$ ) and going up to $(0.9,0.9,0.9)$ with steps of size 0.1 paired with appropriate output derived from (1). The testing set is all the triplets starting from $(0.05,0.05,0.05)$ and going up to $(0.95,0.95,0.95)$ with steps of size 0.1 paired with correct outputs. Thus, we have 1000 training and 1000 testing points and these two sets have no overlap.

Because of the random nature of SDCC, we get a different network with different structure every time we run the algorithm on the training set. Across 50 learned networks for module 1, on average, each network had 13.2 hidden units and 3.4 hidden layers, and the training took 970 epochs.

We compare the outputs of module 1 with the actual results of Bayes' rule by plotting them against each other in a scatter plot. In order to check the generalization accuracy of the built network we use the testing set data (which are not used in training the network) in our analysis. The results in Fig. 4 show that there is a high correlation between the outputs of the network and the true values. Also the slope and $y$-intercept of the fitted line are respectively near 1 and 0 . In sum, the learned module 1 produces highly precise outputs and we can conclude that it implements the Bayes' rule successfully.

## Module 2

Module 2 computes the likelihood given the observed data. This module learns the distribution generating data solely based on the training set presented to it during the training phase. It has no prior information about the form or characteristics of this distribution. Different hypotheses can have different likelihood distributions, hence we run one module 2 for each hypothesis. Through several experiments, we show that module 2 can learn a variety of likelihood distributions.


Figure 4: Outputs of module 1 plotted against true values.

We start with the Gaussian distribution. In this case, the likelihood is given as in equation (2). The observed data, $d$, is the input and its likelihood is the output. Given the hypothesis $h$, we have a Gaussian likelihood with mean $h$, and standard deviation 1. For example, $h$ could be 0 or 1 in the case where we have two hypotheses. Each of these distributions can be learned using SDCC. We define our training set to be the collection of equidistant points from -5 to 5 with steps of size 0.01 and the testing set to be equidistant points from -4.975 to 4.975 , with the same step size, all paired with appropriate outputs derived from (2). Across 50 learned networks, on average, module 2 had 1.8 hidden layers and 8.2 hidden units.

To check the accuracy of the module's outputs, we plot them against the actual Gaussian values in a scatter plot shown in Fig. 5. The high correlation and the equation of the fitted line show that this module succeeds in learning the Gaussian distribution. To further assess the performance of module 2, we plot the probability distribution function generated by it alongside the actual Gaussian distribution in Fig. 6a. We observe that the two curves are very close.


Figure 5: Outputs of module 2 plotted against true values of a Normal distribution. For the Normal, there are two $x$ values for every y value; hence, there are two lines of dots.


Figure 6: Outputs of module 2 compared with the actual values for three sample distributions.

The replication of the original likelihood distribution by our model is of special importance, because it is done without any explicit or implicit information about the actual distribution. The only information available to the learner is the probability values of the points in the training set. Based on that and by generalization, module 2 learns the actual distribution which generates the data. This capacity is not limited only to the Gaussian distribution. In our simulations, we observe that a wide range of distribution functions can be learned by this module. In Fig. 6, we show that for a couple of sample probability distributions module 2 produces results very close to the actual distribution functions.

## Module 3

Module 3 computes hypotheses' priors. Although their inputs and outputs are of different nature, modules 2 and 3 are functionally the same. They both compute the output based on a probability distribution over the input. Therefore, like module 2 , module 3 is capable of learning the underlying structure of its inputs - the possible hypotheses. The only difference between modules 2 and 3 is that the likelihood distributions (in module 2) are continuous while prior distributions (in module 3) are discrete. We analyse module 3 in more detail while discussing base-rate neglect in the next section.

## Base-rate Neglect as Weight Decay

In contemporary cognitive science, rationality in learning and inference is frequently defined and measured in terms of conformity to Bayes' rule. However, this appears to conflict with the Nobel-prize-winning work showing that people are somewhat poor Bayesians due to biases such as base-rate neglect, representativeness heuristic, and confusing the direction of conditional probabilities (Kahneman \& Tversky, 1996). Even experienced medical professionals deviate from optimal Bayesian inference and make major errors in their probabilistic reasoning (Eddy, 1982). More recently, Prime and Shultz showed that base rates (i.e., priors) are not entirely ignored but just de-emphasized (Prime \& Shultz, 2011).

We first show how base-rate neglect can be interpreted in the Bayesian framework. Then, as an important contribution
of this work, we show that this neglect can be modelled as weight decay in our proposed neural network. This explanation is particularly of interest because it is neurologically plausible and in accordance with theories explaining memory loss and decline in some other cognitive functions as a result of synaptic decay over time (Hardt, Nader, \& Nadel, in press).

Base-rate neglect is an error in computing the posterior probability of a hypothesis without taking full account of the priors. In the Bayesian framework, in the extreme case of entirely ignoring the priors, Bayes' rule in (1) becomes:

$$
\begin{equation*}
P\left(h_{i} \mid d\right)=\frac{P\left(d \mid h_{i}\right)}{\sum_{i=1}^{N} P\left(d \mid h_{i}\right)} \tag{4}
\end{equation*}
$$

Looking at this equation from a different perspective, we can assume that in the original Bayes' rule, all the hypotheses had equal priors and these priors were cancelled out to give equation (4). Therefore, in the Bayesian framework, base-rate neglect is translated into assuming equal priors (i.e., equiprobable hypotheses). This means that the more the original priors (base rates) are averaged out and approach the uniform distribution, the more they are neglected in Bayesian inference. We can explain this more abstractly by using the notion of entropy defined in information theory as a measure of uncertainty. Given a discrete random variable $X=\left\{h_{1}, \ldots, h_{N}\right\}$ with probability mass function $P(\cdot)$, its entropy is defined as:

$$
\begin{equation*}
H(X)=-\sum_{i=1}^{N} P\left(h_{i}\right) \log _{2} P\left(h_{i}\right) \tag{5}
\end{equation*}
$$

Entropy quantifies the expected value of information contained in a distribution. It is easy to show that a uniform distribution has the maximum entropy (equal to $\log _{2} N$ ) among all discrete distributions over the set $\left\{h_{1}, \ldots, h_{N}\right\}$. In sum, we can conclude that in the Bayesian framework, base-rate neglect is equivalent to ignoring the priors in the form of averaging them out to get a uniform distribution, or equivalently, increasing their entropy.

We show that weight decay in our proposed neural network produces the same results as just described, namely approaching a uniform distribution and increasing entropy. We then conclude that we can model base-rate neglect in the Bayesian framework by a weight decay mechanism in our brain-like

(a) The priors' initially Gaussian distribution approaches uniform as time passes and weights decay more (decay rate $=0.2$ ).

(b) The entropy of prior distributions increases as time passes and weights decay more.

Figure 7: Effects of connection weight decay.
network implementing the posterior inference. We take priors as the states of a learning and inference system. As weights decay, the system moves towards more stable states and thus the entropy increases. In special cases, some priors are updated so often that the effects of decay are overcome and these priors stay strong. On the other hand, likelihoods are new evidence and thus, they are not subject to much decay.

In module 3, an SDCC network learns the hypotheses' priors. Assume that we have $N$ hypotheses, $\left\{h_{1}, \ldots, h_{N}\right\}$, with probability mass function given by equation (3). We present the results for this specific mass function, but the results are similar for other discrete distributions. Our training set is the collection of 401 equidistant points from -10 to 10 (with steps of size 0.05 ), and our testing set is the collection of 399 equidistant points from -9.975 to 9.975 , all paired with the correct outputs derived from (3). We choose these sets such that there is no overlap between the testing and training sets in order to measure the generalization abilities of the module. Note that we do not specify the form of the probability mass function in any way and the network learns it by generalization from the training input. Also note that in the case we consider here, $N=399$ as we define our hypotheses to be the collection of points in the testing set.

After learning, the network's weights decay exponentially over time steps as follows:

$$
\begin{equation*}
W_{i}(t+1)=(1-r) \cdot W_{i}(t)=(1-r)^{t} \cdot W_{i}(1) \tag{6}
\end{equation*}
$$

where $W_{i}(t)$ is the weight of connection $i$ at time $t$ and $r$ is the decay rate. Clearly, as the connection weights of the learned network decay, the output will change. Fig. 7 demonstrates the effects of weight decay on the output of module 3. In Fig. 7a, for $r=0.2$, we observe that as time passes, and hence as the weights continue to decay, the distribution of the hypotheses approaches a uniform distribution. We consider a discrete case where we have a finite number of hypotheses, and therefore Fig. 7 represents the probability mass function where the sum of all probabilities must be equal to 1 . Note that in the 15th time step, the distribution is almost uniform; thus, the value of the probability is $1 / N=1 / 399$ (this small value should not be mistaken with zero in Fig. 7(a)).

There is a literature on using weight decay to improve the ability of neural networks to generalize (Krogh \& Hertz, 1992). Krogh and Hertz showed that a weight decay can improve generalization by suppressing any irrelevant components of the weight vector. This effect is evident in Fig. 7a, as the bumps (overfitting) get smoothed out as time passes.

In Fig. 7b, entropies are plotted as a function of time for four values of decay rate. In all four cases, entropy increases with time until it converges to its maximum which corresponds to uniform prior distribution. In our case, the maximum entropy is $\log _{2} N=\log _{2} 399=8.64$.

In conclusion, we show that the proposed neural network model contributes to the resolution of the discrepancy between demonstrated Bayesian successes and failures by modelling base-rate neglect as weight decay in a connectionist network implementing Bayesian inference. This is done by showing that as weights decay, the priors' probability mass function approaches a uniform distribution and its entropy increases. Consequently, the prior terms eventually cancel out from Bayes' rule, resulting in the neglect of base rates.

## Discussion

We propose a modular neural-network structure to implement Bayesian learning and inference. Through simulations, we show that the proposed three modules, responsible for computing Bayes' rule, likelihoods, and priors, succeed in learning their assigned task. Employing a weight-decay mechanism, we provide a novel explanation of base-rate neglect, the most well-documented deviation from Bayes' rule. We show that weight decay increases the entropy of priors in a Bayesian system. In nature, this is very similar to the second law of thermodynamics which states that the entropy of isolated systems never decreases and that they evolve towards equilibrium - the state of maximum entropy. Our model of base-rate neglect predicts that older, less continuously supported (i.e., more isolated) priors would be subject to more decay, and consequently more neglect.

Our model is a first step towards implementing the Bayesian framework with neural networks. As such, it still
has several limitations. For instance, our current model is only capable of handling a finite number of hypotheses. When the number of hypotheses gets uncountably infinite, the current structure will be infeasible due to large numbers of outputs that must be remembered.

Our networks learn to approximate Bayesian functions with the output being a function value. However, this assumption might be unrealistic and does not explain how the brain innately represents probabilities. It is more realistic to approximate such functions from input instances occurring at various frequencies. Although we do not address this here, our further experiments show that SDCC networks can do this, which is effectively probability matching. With that, our modular neural network, similar to human brain, makes sense of data by representing probability distributions and applying Bayes' rule to find the best explanation for any given data.

In this paper, we apply our model to base-rate neglect. However, there are many other subtle and potentially difficult Bayesian phenomena, such as hierarchical Bayesian structures and causal networks, to consider. Also, our work does not address the origin of the Bayesian competencies; all we show is that neural networks can implement Bayesian inference and learning. The origin of theses capacities could be in evolution, learning, development or some combination. Finally, we only use SDCC as the learning method and do not try other neural-network approaches.

There is plenty of scope for future research to address the issues just discussed. For instance, we can illuminate the issue regarding the origin of the Bayesian competencies of our model by agent-based simulations of evolution of Bayesian inference and learning. Preliminary results in the context of social learning strategies have shown that evolution favours Bayesian learning, based on passing posteriors, over imitation and environment sampling (Montrey \& Shultz, 2010).

With no doubt, Bayesian models provide powerful analytical tools to rigorously study deep questions of human cognition that have not been previously subject to formal analysis. These Bayesian ideas, providing computation-level models, are becoming prominent across a wide range of problems in cognitive science. On the other hand, connectionist models offer an implementation-level framework for modelling mental phenomena in a more biologically plausible fashion. We present this work in the spirit of theoretical unification and enhancement of these two approaches. We are not advocating replacement of one approach in favour of the other.

## Acknowledgements

This research was supported in part by McGill Engineering Doctoral Award to MK and a Discovery grant to TRS from the Natural Sciences and Engineering Research Council of Canada. Mark Coates and Deniz Üstebay contributed thoughtful comments on an earlier draft.

## References

Ackley, H., Hinton, E., \& Sejnowski, J. (1985). A learning algorithm for Boltzmann machines. Cog. Sci., 147-169.

Baluja, S., \& Fahlman, S. E. (1994). Reducing network depth in the cascade-correlation learning architecture (Tech. Rep.). Carnegie Mellon University, School of Computer Science.
Chater, N., \& Manning, C. D. (2006). Probabilistic models of language processing and acquisition. Trends in Cognitive Sciences, 10(7), 335-344.
Doya, K., Ishii, S., Pouget, A., \& Rao, R. (2007). Bayesian brain : probabilistic approaches to neural coding. Cambridge, MA : MIT Press.
Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: problems and opportunities. In D. Kahneman, P. Slovic, \& A. Tversky (Eds.), Judgment under uncertainty: Heuristics and biases. Cambridge Univ. Press.
Griffiths, T. L., Austerweil, J. L., \& Berthiaume, V. G. (2012). Comparing the inductive biases of simple neural networks and bayesian models. In Proc. the 34th Annual Conf. of the Cog. Sci. Society.
Hardt, O., Nader, K., \& Nadel, L. (in press). Decay happens: the role of active forgetting in memory. Trends in Cog. Sci..
Hinton, G. E., \& Osindero, S. (2006). A fast learning algorithm for deep belief nets. Neural Computation, 18, 2006.
Kahneman, D., \& Tversky, A. (1996). On the reality of cognitive illusions. Psychological Review, 103, 582-591.
Konrad, K., \& Wolpert, D. (2004). Bayesian integration in sensorimotor learning. Nature, 427, 244-247.
Krogh, A., \& Hertz, J. A. (1992). A simple weight decay can improve generalization. Advances in neural information processing systems, 4, 950-957.
McClelland, J. L. (1998). Connectionist models and Bayesian inference. In M. Oaksford \& N. Chater (Eds.), Rational models of cognition. Oxford, UK: Oxford Univ. Press.
Montrey, M., \& Shultz, R. (2010). Evolution of social learning strategies. , 95-100.
Pouget, A., Dayan, P., \& Zemel, R. (2003). Inference and computation with population codes. Annual Review in Neuroscience, 26, 381-410.
Prime, H., \& Shultz, T. R. (2011). Explicit Bayesian reasoning with frequencies, probabilities, and surprisals. Proc. of 33rd Annual Conf. Cog. Sci. Society.
Shi, L., \& Griffiths, T. (2009). Neural implementation of hierarchical Bayesian inference by importance sampling. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, \& A. Culotta (Eds.), Advances in neural information processing systems 22 (pp. 1669-1677).
Shultz, T. R., \& Fahlman, S. E. (2010). Cascade correlation. In C. Sammut \& G. I. Webb (Eds.), Encyclopedia of machine learning (p. 139-147). Heidelberg, Germany: Springer-Verlag.
Tenenbaum, J. B., Kemp, C., \& Shafto, P. (2006). Theorybased Bayesian models of inductive learning and reasoning. Trends in Cognitive Sciences, 10(7), 309-318.
Yuille, A., \& Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? Trends in Cog. Sci., 10(7), 301-308.

# Uncertainty can increase explanatory credibility 

Sangeet Khemlani ${ }^{1}$, Daniel Gartenberg ${ }^{2}$, Kun Hee Park ${ }^{2}$, and J. Gregory Trafton ${ }^{1}$<br>${ }^{1}$ US Naval Research Laboratory, Washington, DC 20375 USA<br>${ }^{2}$ George Mason University, Fairfax, VA 22030 USA


#### Abstract

In daily conversations, what information do people use to assess their conversational partner's explanations? We explore how a metacognitive cue, in particular the partner's confidence or uncertainty, can modulate the credibility of an explanation. Two experiments showed that explanations are accepted more often when delivered by an uncertain conversational partner. Participants in Experiment 1 demonstrated the general effect by interacting with a pseudoautonomous robotic confederate. Experiment 2 used the same methodology to show that the effect was applicable to explanatory reasoning and not other sorts of inferences. Results are consistent with an account in which reasoners use relative confidence as a metacognitive cue to infer their conversational partner's depth of processing.


Keywords: explanations, confidence, uncertainty, collaborative reasoning, human-robot interaction

## Introduction

What makes an explanation believable? Researchers have recently discovered several conceptual and structural properties that distinguish credible explanations (for reviews, see Keil, 2006; Lombrozo, 2006). Good explanations are often relevant and informative (Grice, 1975; Wilson \& Sperber, 2004). Likewise, people appear to prefer explanations that are simple (Chater, 1996; Lagnado, 1994; Lombrozo, 2007; but cf. Johnson-Laird, Girotto, \& Legrenzi, 2004), and in situations of uncertainty, they appear to prefer explanations that have narrow latent scope, i.e., those that account for only observed phenomena (Khemlani, Sussman, \& Oppenheimer, 2011). These preferences show that properties intrinsic to the explanation itself can cause individuals to judge the explanation to be better, more likely, more plausible, and more credible.

However, individuals also rate explanations by appealing to extrinsic information, e.g., information about the context in which the explanation was provided rather than the material content described by the explanation. Extrinsic information is particularly important when reasoners have to evaluate another individual's explanations. In those situations, factors such as the individual's motivation, mood, and confidence can affect the believability of his or her explanation. In this paper, we focus on how confidence can modulate an explanation's credibility. We first describe confidence as a metacognitive signal, and then explain how confidence can affect the believability of an explanation. Two studies show that when an agent appears uncertain, individuals accept the agent's explanations more often. We discuss the phenomenon in light of intuitive and analytic reasoning systems.

## Confidence and explanatory credibility

Subjective confidence is among the most widely investigated metacognitive signals (Dunlosky \& Metcalfe, 2009). In many cognitive tasks it is correlated with accuracy, though people are often systematically overconfident about their performance (Lichtenstein, Fischhoff, \& Phillips, 1982; Lindley, 1982; McClelland \& Bolger, 1994). Much of the research on subjective confidence addresses how individuals integrate cues from their task performance or else their declarative knowledge to assess their confidence in a particular decision of theirs. Confidence is often construed as a signal predictive of translating judgments to actions (Dunning, 2007; Tversky \& Koehler, 1994), and researchers have accordingly proposed many models of how that signal is constructed (Albert \& Sponsler, 1989; Erev, Wallsten, \& Budescu, 1994; Ferrell \& McGoey, 1980; Gigerenzer, Hoffrage, \& Kleinbölting, 1991; Griffin \& Tversky, 1992; Juslin, 1994; Koriat, 2012; May, 1986; Pfeifer, 1994; Wallsten \& Gonzáles-Vallejo).

In daily interactions with others, people frequently provide cues to their own level of confidence for their conversational partners to interpret, and they use their partner's cues to interpret the content of their partner's statements. Despite the prevalent use of confidence signals in modulating informational content, little work has established how individuals integrate cues to a partner's confidence or lack thereof into their own decision-making, and few if any of the aforementioned models of subjective confidence can explain how confidence is assessed in others. Suppose, for example, that you ask a friend what she thinks of a new restaurant that has opened up in her neighborhood. If she says, "It's good!" her intonation may provide a cue to a high level of confidence in her response. Alternatively, if she hesitates and says, "It's...good..." then you may negate the material content of her response and prefer instead to explain her lack of confidence as indicative of her disapproval.

In the present investigation, we examined how individuals incorporate their partners' levels of confidence when they assess their partner's explanations of a confusing scenario. Reasoners could modulate their acceptance in their partner's explanation in one of two ways. An intuitive prediction is that people should accept an explanation more often when the explanation is delivered by a confident partner than an uncertain partner. People who exhibit this behavior should infer, implicitly or explicitly, that the partner's confidence is proportional to the explanation's credibility. Preliminary support for this prediction comes from recent studies on so called "powerless language", which show that statements that include hedging phrases such as "sort of", "kind of",
and "probably" are rated more negatively compared to nonhedged statements (Blankenship \& Holtgraves, 2005; Durik, Britt, Reynolds, \& Storey, 2008; Liu \& Fox Tree, 2012). Hedges may provide a cue to a low level of confidence, and therefore cause people to attenuate their belief in the statement.

Alternatively, if people prefer explanations when they are delivered by an uncertain partner, then it may be because the partner's uncertainty provides pragmatic cues to the strength of the explanation. For example, an uncertain expressional cue such as a furrowed brow may suggest that the partner was engaged in more analytical thinking (Alter, Oppenheimer, Epley, \& Eyre, 2007), and an analytical response may be preferred to an intuitive one.

In what follows, we report two experiments that tested whether confidence or uncertainty affects explanatory credibility. In both studies, participants engaged in a dyadic interaction with a pseudo-autonomous humanoid robot. The robot allowed us to impose stringent controls on the verbal and expressional cues that participants received.

## Experiment 1

Experiment 1 tested whether an explanation was more or less acceptable if it came from a confident or an uncertain confederate. To generate systematic social interactions, the experiment called on participants to engage in a dyadic interaction with a pseudo-autonomous robotic confederate, a humanoid mobile, dexterous, social (MDS) robot (Breazeal et al., 2008). Participants were told that they were interacting with the robot through a web-based chat interface (see Figures 1 and 2). Participants' task was to read a problem to the robot, listen to the robot's response, and then decide whether they agreed, did not understand, or disagreed with the robot. If they did not understand, or else if they disagreed with the robot, they verbally explained their reason for not accepting the robot's response, and their verbal protocols were recorded. All of the robot's responses were pre-recorded, and we manipulated whether the robot delivered its responses using cues of confidence or uncertainty.

## Method

Participants. 38 native-English speaking undergraduates from George Mason University participated in exchange for partial course credit. None of the participants had received any training in logic.

Procedure. Participants engaged in a dyadic interaction with a pseudo-autonomous robotic confederate. Before they began the study, they were shown a video of humans engaged in natural language dialogue with an MDS robot (Hiatt et al., 2011). Participants were told that they would interact with the robotic confederate online, but that the confederate had only limited abilities to comprehend natural language, and that the confederate would be unable to respond to unrelated questions. In actuality, all of the robot's responses were pre-recorded. Participants were
instructed to use a chat interface to read problems to the confederate and listen to the confederate's responses. The interface was written in Objective C for an iPad tablet computer.

The experiment began when the confederate introduced itself as "Lucas", an MDS robot, and waited for the participant to initiate the study by reading the first problem. Figure 1 shows a schematic of the interface. Participants first read a description of a problem to the confederate (Figure 1a); when they finished, they pressed a button and listened to the confederate's response (Figure 1b); when the robot finished speaking, the participants indicated whether they agreed with, did not understand, or disagreed with the robot's response (Figure 1c); finally, if they disagreed or did not understand the robot, they were given an opportunity to explain their disagreement verbally (Figure 1d), and they moved on to the next problem.

a.

b.

c.

d.

Figure 1. A schematic diagram of the chat interface used for the pseudo-interaction in Experiments 1 and 2.

Design and materials. Problems consisted of a conditional generalization (1), a categorical statement (2), and an inferential prompt, e.g.,

1. If James does regular aerobic exercises then he strengthens his heart.
2. But, James did not strengthen his heart.
3. What, if anything, follows?

The problems invite both explanatory (e.g., "James had a congenital heart defect") and deductive (e.g., "James did not do regular exercises") responses. However, people tend to elicit explanations for such problems (Lee \& Johnson-Laird, 2006). In the present study, participants listened to and evaluated the confederate's explanation of ten separate problems, which were drawn from five different domains: biology, economics, mechanics, psychology, and natural phenomena (see the Appendix for the full set of materials). Explanations were adapted from reasoners' most frequently generated spontaneous explanations in studies that used similar materials (Khemlani \& Johnson-Laird, 2012). For each explanation, the robotic confederate delivered its response using a verbal cue and an expressional cue to its level of confidence. Half of the participants received confident verbal and expressional cues, and the remaining received uncertain cues. The explanations in both conditions were delivered with the same intonation. Figure 2 provides examples of the verbal and expressional cues. The
materials were balanced for their length across both conditions.

Post-experimental questionnaire. Participants who perceive their interaction with the robot as staged may respond differently than those who believe the interaction is real. To examine this factor, participants completed a postexperimental questionnaire after they finished the experiment proper. The questionnaire assessed whether the participants had believed (erroneously) that they were interacting with an autonomous robot, or whether they believed (accurately) that the interaction was staged. In our analyses, we present data from the most direct question they answered, which was as follows:
"Did Lucas's responses seem natural?

1. No, his responses usually looked like pre-recorded videos.
2. I'm not sure.
3. Yes, he usually responded like a human would."

After participants answered the questionnaire, they were debriefed that the interaction was staged.

## Results and discussion

Figure 3 shows the percentage of agreement for the explanations as a function of the confederate's confidence. Surprisingly, participants accepted explanations more often when the confederate was uncertain ( $75 \%$ agreement) than when it was confident ( $63 \%$ agreement; one-tailed MannWhitney test, $\mathrm{z}=1.75, \mathrm{p}=.04$, Cliff's $\delta=.33$ ). In both conditions, participants accepted explanations signifi-


Figure 2. The interface used in Experiments 1 and 2 (a). The robotic confederate was either confident (b) or uncertain (c) for the duration of the study. Confident expressional cues included wide open eyes, raised eyebrows, and a straight mouth orientation. Furthermore, the confident confederate preceded its responses with confident verbal cues, e.g., "Oh, I've got it!" or "That's easy." Uncertain expressional cues included narrow eyes, half-cocked eyebrows (a furrowed brow analog), and a slanted mouth orientation. Uncertain verbal cues included expressions such as, "Hmm, that's a tough one" and "Huh, I don't know for sure."
cantly more often than chance (Wilcoxon tests, zs $>2.25$, ps $<.02$ ). Their agreement varied across the different types of materials (Friedman analysis of variance, $\chi^{2}=49.9, \mathrm{p}<$ .0001 ). Across the study, $45 \%$ of the participants responded that they believed the interaction was pre-recorded.

To assess whether the effect of uncertainty on explanatory credibility was robust across the different materials, we fit the data to a generalized mixed-effects model (Baayen, Davidson, \& Bates, 2008) with a binomial error distribution and a logit link function using the lme 4 package (Bates, Maechler, \& Bolker, 2012) in R (R Core Team, 2012). The model took into account a single fixed effect, i.e., the confederate's confidence, as well as three additional random effects: the participant variance, the problem variance, and whether or not the participant believed that the interaction was pre-recorded. The model yielded a significant main effect of confidence $(b=.77, \mathrm{SE}=.37, \mathrm{p}=.04)$. The results suggest that the effect held whether or not the participants believed that the interaction was staged.


Figure 3.
Agreement percentages for explanations as a function of whether those explanations were delivered by a confident or an uncertain confederate. $95 \%$ confidence intervals shown.

Experiment 1 tested whether reasoners would accept explanations more or less often when given by an uncertain confederate compared to a confident confederate. However, the study did not establish whether the effect is unique to explanatory reasoning. It may be the case that the effect is widespread, and that it is applicable to any sort of inference, not just to the evaluation of explanations. To test the boundary conditions of the effect, participants in Experiment 2 evaluated both explanations and deductions.

## Experiment 2

Experiment 2 sought to replicate the effect of uncertainty on explanatory credibility, as well as to test whether it applied to any sort of inference, or whether it was localized, in part, to explanatory reasoning. The study was similar to the previous one, with one exception: the robotic confederate in the present study provided two types of responses, either an explanation or else a deduction. Recall that the problems used in the previous study, e.g.,

If James does regular aerobic exercises then he strengthens his heart.
But, James did not strengthen his heart.
What, if anything, follows?
invite two different sorts of reasoning strategies. One could construct an explanation that goes beyond the information in the premises (Khemlani \& Johnson-Laird, 2011). Or else one could make a modus tollens deduction, which is a logical deduction that takes the following abstract form. If $A$ then B. Not B. Therefore, not $A$. The inference is valid, i.e., the conclusion is true whenever the premises are true, but it is difficult for naïve reasoners. Thus, in the present study, the robotic confederate's responses concerned either an explanation or else a modus tollens deduction. Half of the participants interacted with a confident confederate and the other half interacted with an uncertain one. If the effect of uncertainty on credibility applies to any sort of response, then there should not be an interaction between the type of inference and the confederate's confidence. In contrast, if the effect is unique to explanatory reasoning, then there should be no difference between participants' evaluations of confident and uncertain deductions, but there should be a difference in their evaluations of explanations.

## Method

Participants, design, and procedure. 45 native Englishspeaking participants were recruited though the same participant pool as in Experiment 1. None of them had received training in formal logic. They solved ten reasoning problems by engaging in a web-based chat interaction with a pseudo-autonomous robotic confederate (see Figures 1 and 2), and they were taught to use the interface using the same procedure as in the previous study. Their task was to read each problem aloud to the confederate, listen to the confederate's response, and then judge whether they agreed, did not understand, or disagreed with the response. On half of the problems, the confederate would produce an explanation, and on the other half, it would produce a deduction (see Appendix). Twenty-three participants interacted with a confederate that produced confident responses and the remaining interacted with one that produced uncertain responses. After completing the last problem, participants filled out the same post-experimental questionnaire that was described for Experiment 1.

## Results and discussion

Figure 4 presents the percentage of agreement to deductions and explanations as a function of whether the response was delivered by a confident or an uncertain confederate. Participants agreed with deductions almost at ceiling ( $87 \%$ ) and accepted them reliably more often than they accepted explanations ( $63 \%$; Wilcoxon test, $\mathrm{z}=3.8$, p $<.0001$, Cliff's $\delta=.55$ ). Likewise, they accepted uncertain responses more often than confident responses ( $81 \% \mathrm{vs}$. $71 \%$; Mann-Whitney test, $\mathrm{z}=2.47, \mathrm{p}=.01$, Cliff's $\delta=.43$ ). However, the main effect of confidence was driven entirely


Figure 4. Agreement percentages for deductions and explanations as a function of whether they were delivered by a confident or an uncertain confederate. $95 \%$ confidence intervals shown.
by the difference between confident and uncertain explanations, and the data yielded a significant interaction between the type of inference and the confederate's confidence (Mann-Whitney test, $\mathrm{z}=1.95, \mathrm{p}=.05$, Cliff's $\delta$ $=.48$ ). The results suggest that the effect of uncertainty on credibility applies to explanations and not deductions. As in the previous study, agreement varied as a function of the contents of the problems (Friedman analysis of variance, $\chi^{2}$ $=43.49, \mathrm{p}<.0001$ ), and $58 \%$ of the participants reported that they believed the interaction was pre-recorded.

To assess whether the effect and the relevant interaction were both reliable across the different materials, we fitted the data to another generalized mixed-effects model. The model took into account two fixed effects, i.e., the confederate's confidence and the inference type, and the three pertinent random effects, i.e., the participant variance, the problem variance, and whether or not the participant believed that the interaction was pre-recorded. The model yielded a significant main effect of the type of inference (b $=-2.07, \mathrm{SE}=.36, \mathrm{p}<.0001$ ), however it yielded no main effect of confidence $(\mathrm{b}=.05, \mathrm{SE}=.42, \mathrm{p}=.90)$. Instead, it yielded a significant interaction between the type of inference and the confederate's confidence ( $\mathrm{b}=1.07$, $\mathrm{SE}=$ $.54, \mathrm{p}=.045$ ). As in Experiment 1, the analysis shows that the effect held in spite of any variance from the different materials or the perception that the interaction was staged.

## General Discussion

We used a novel experimental methodology to study how reasoners incorporate metacognitive information to judge one another's explanations. In two experiments, reasoners interacted with a robotic agent that appeared to deliver its responses in a confident or else an uncertain demeanor. One might expect that people should agree with confident explanations more often. Yet Experiment 1 showed that participants accepted explanations more often when they came from an uncertain confederate compared to a confident one. Experiment 2 tested whether the effect held
more generally for deductions, but it found instead that it was limited to explanations.

Why do reasoners accept explanations more often when they come from an uncertain source? The results are counterintuitive, particularly since confidence is correlated with informational accuracy. Indeed, at first blush, the results of our studies conflict with recent findings on hedging behavior and powerless language (Blankenship \& Holtgraves, 2005; Durik, Britt, Reynolds, \& Storey, 2008; Liu \& Fox Tree, 2012). However, we hypothesize that one reason for a speaker to produce uncertain expressions, gestures, and verbal cues is to signal to a listener that the speaker is engaged in deeper analytic processing, and furthermore, that the speaker is considering alternative possibilities. This proposal accounts for why the effect is manifest for explanations but not modus tollens deductions: explanations require reasoners to think about multiple possibilities and to go beyond the information presented in the premises, whereas modus tollens deductions do not. If our hypothesis is true, then we should find a similar effect of uncertainty on credibility for deductions that require reasoners to consider multiple possibilities compared to those that do not.
The present data reveal a robust credibility effect for human-robot interactions, and critics are justified in wondering whether the effect will still hold in dyadic human-human interactions (but cf. Moon \& Nass, 1996, for evidence that people treat interactive computers as though they were human). Similar studies with human confederates are feasible, but the human-robot interaction paradigm we employed has several advantages to traditional studies with human confederates. First, robotic confederates can be programmed to yield very precise expressional and gestural cues that are consistent for all participants in the study, while even the best human confederates are susceptible to irregular behaviors. Second, robotic confederates can be programmed to implement complex experimental designs and counterbalancing schemes. For example, the software in Experiment 2 was written so that exactly half of the robot's responses were explanations. Despite these advantages, however, future studies should examine the credibility effect in, albeit less controlled, human studies. One promising methodological compromise is to run pseudo-dyadic interaction studies over the Internet (Summerville \& Chartier, 2012).

The results we present have psychological implications, as well as implications for robotics researchers. A major goal for the interdisciplinary community of human-robot interaction research is to develop social robots that humans trust (Fong, Thorpe, \& Baur, 2001; Goodrich \& Schultz, 2007; Steinfield et al., 2006). The credibility effect we show implies that humans are likely to take into account metacognitive signals (and their robotic analogs) in assessing information from autonomous systems. Research on the modulatory effects of confidence on higher order reasoning is of multidisciplinary relevance, and can be applied to developing broader theories of confidence
monitoring in humans as well as more natural and trustworthy autonomous robots.

## Acknowledgements

This research was funded by a National Research Council Research Associateship awarded to SK and ONR Grant \#s N0001412WX30002 and N0001411WX20516 awarded to JGT. We are also grateful to Bill Adams, Len Breslow, Magda Bugajska, Tony Harrison, Laura Hiatt, Phil JohnsonLaird, Ed Lawson, Eric Martinson, Malcolm McCurry, Lilia Moshkina, Frank Tamborello, Alan Schulz, and the ARCH Lab at George Mason University for their helpful comments and their assistance in data collection.

## References

Albert, J.M., \& Sponsler, G.C. (1989). Subjective probability calibration: A mathematical model. Journal of Mathematical Psychology, 33, 289-308.
Bates, D., Maechler, M., Bolker, B. (2012). Ime4: Linear mixed-effects models using S 4 classes. Retrieved fromhttp://CRAN.Rproject.org/package=lme4.
Blankenship, K. L., \& Holtgraves, T. (2005). The role of different markers of linguistic powerlessness in persuasion. Journal of Language and Social Psychology, 24, 3-24.
Breazeal, C., Siegel, M, Berlin, M., Gray, J., Grupen, R., Deegan, P, Weber, J., Narendran, K., \& McBean, J. (2008). Mobile, dexterous, social robots for mobile manipulation and human-robot interaction. In Proceedings of International Conference on Computer Graphics and Interactive Techniques. New York, NY: ACM.
Chater, N. (1996). Reconciling simplicity and likelihood principles in perceptual organization. Psychological Review, 103, 566-581.
Dunlosky, J., \& Metcalfe, J., (2009). Metacognition. Thousand Oaks, CA: US Sage Publications, Inc.
Durik, A., Britt, M.A., Reynolds, R., \& Storey, J. (2008). The effects of hedges in persuasive arguments: A nuanced analysis of language. Journal of Language and Social Psychology, 30, 341-349.
Erev, I., Wallsten, T. S., \& Budescu, D. V. (1994). Simultaneous overconfidence and conservatism in judgment: implications for research and practice. Psychological Review, 101, 519-27.
Ferrell, W.R., \& McGoey, P.J. (1980). A model of calibration for subjective probabilities. Organizational Behavior and Human Performance, 26, 32-53.
Fong, T., Thorpe, C., \& Baur, C. (2001). Collaboration, dialogue, and human-robot interaction. In Proceedings of the $10^{\text {th }}$ International Symposium of Robotics Research.
Gigerenzer, G., Hoffrage, U., \& Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. Psychological Review, 98, 506-528.
Goodrich, M., \& Schultz, A.C. (2007). Human-robot interaction: A survey. Foundations and Trends in Human-Computer Interaction, 1.
Grice, H. P. (1975). Logic and conversation. In P. Cole and J. Morgan (Eds.) Syntax and semantics, pp. 41-58. New York: Academic.
Griffin, D., \& Tversky, A. (1992). The weighting of evidence and the determinants of confidence. Cognitive Psychology, 24, 411-435.
Hiatt, L.M., Harrison, A.M., Lawson, W.E., Martinson, E., \& Trafton, J.G. (2011). Robotic secrets revealed, Episode 002 [Video file]. Retrieved from http://goo.gl/ru28e.
Johnson-Laird, P.N., Girotto, V., \& Legrenzi, P. (2004). Reasoning from inconsistency to consistency. Psychological Review, 111.
Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. Organizational Behavior and Human Decision Processes, 57.
Keil, F. C. (2006). Explanation and understanding. Annual Review of Psychology, 57, 225-254.

Khemlani, S., \& Johnson-Laird, P.N. (2011). The need to explain. Quarterly Journal of Experimental Psychology, 64, 2276-88.
Khemlani, S., \& Johnson-Laird, P. N. (2012). Hidden conflicts: Explanations make inconsistencies harder to detect. Acta Psychologica, 139, 486-491.
Khemlani, S., Sussman, A., \& Oppenheimer, D. (2011). Harry Potter and the sorcerer's scope: Scope biases in explanatory reasoning. Memory \& Cognition, 39, 527-535.
Koriat, A. (2012). The self-consistency model of subjective confidence. Psychological Review, 119, 80-113.
Lagnado, D. (1994). The psychology of explanation: A Bayesian approach. Masters Thesis. Schools of Psychology and Computer Science, University of Birmingham.
Lee, N.Y.L., \& Johnson-Laird, P.N. (2006). Are there cross-cultural differences in reasoning? In D.S. Macnamara \& J.G. Trafton (Eds.), Proceedings of the 28th Annual Conference of the Cognitive Science Society (pp. 459-464). Austin, TX: Cognitive Science Society.
Lichtenstein, S., Fischhoff, B., Phillips, L.D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, A. Tversky (Eds.), Judgment under uncertainty: Heuristics and biases. Cambridge, UK: Cambridge University Press.
Liu, K., \& Fox Tree, J. E. (2012). Hedges enhance memory but inhibit retelling. Psychonomic Bulletin \& Review, 19, 892-898.
Lombrozo, T. (2006). The structure and function of explanations. Trends in Cognitive Sciences, 10, 464-470.

Lombrozo, T. (2007). Simplicity and probability in causal explanations. Cognitive Psychology, 55, 232-257.
McClelland, A.G., \& Bolger, F. (1994). The calibration of subjective probabilities: theories and models 1980-1994. In G. Wright and P. Ayton (Eds.), Subjective Probability. Chichester: Wiley.
Moon, C., \& Nass, C. (1996). How "real" are computer personalities? Communication Research, 23, 651-674.
Pfeifer, P.E. (1994). Are we overconfident in the belief that probability forecasters are overconfident? Organizational Behavior and Human Decision Processes, 58, 203-213.
R Core Team (2012). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/
Steinfeld, A., Fong, T., Kaber, D., Lewis, M., Scholtz, J., Schultz, A., \& Goodrich, M. (2006). Common metrics for human-robot interaction, in: Proceedings of the 1st ACM/IEEE International Conference on Human-Robot Interactions, HRI'06.
Summerville, A., \& Chartier, C.R. (2012). Pseudo-dyadic "interaction" on Amazon's Mechanical Turk. Behavioral Research Methods, online publication.
Wallsten, T.S., \& Gonzáles-Vallejo, C. (1994). Statement verification: A stochastic model of judgment and response. Psychological Review, 101, 490-504.
Wilson, D., \& Sperber, D. (2004). Relevance theory. In G. Ward and L. Horn (Eds.) Handbook of pragmatics, pp. 607-32. Oxford, UK: Blackwell Science.

Appendix. The problems used in Experiment 1 and Experiment 2, which consisted of a conditional generalization (column 1) and a categorical statement (column 2).

| Premises (spoken by the participant to the confederate) |  | Responses (spoken to the participant by the confederate) |  |
| :---: | :---: | :---: | :---: |
| Conditional generalization | Categorical | Explanation (Experiments 1 and 2) | Deduction (Experiment 2) |
| If a person is bitten by a viper then he will die | However, a man named Matthew did not die | Matthew received an antidote | Matthew was not bitten by a viper |
| If James does regular aerobic exercises then he strengthens his heart | But, James did not strengthen his heart | James had a congenital heart defect | James did not do regular aerobic exercises |
| If a car's engine is tuned in a special way then its fuel consumption goes down | However, one car's fuel consumption did not go down | The car had engine problems that increased consumption | The car's engine was not tuned in the special way |
| If the aperture on a camera is narrowed, then less light falls on the film | But in one instance, less light did not fall on the film | It was completely dark, so there was no light at all | The aperture on the camera was not narrowed |
| If a person pulls the trigger on a pistol, then the pistol fires | However, it turned out that the pistol did not fire | The safety had not been taken off the pistol | Nobody pulled the trigger |
| If a substance such as butter is heated then it melts | However, one piece of butter did not melt | The heat was too low to melt the butter | The piece of butter was not heated |
| If Chemical A and Chemical B come into contact with one another then there will be an explosion | But there was no explosion | There was not enough of either of the substances | The two substances did not come into contact with one another |
| If a person receives a heavy blow to the head then that person forgets some preceding events | However, Pat did not forget any preceding events | Pat was wearing a helmet at the time | Pat did not receive a heavy blow to the head |
| If people make too much noise at a party then the neighbors complain | But the neighbors did not complain | The neighbors were away on summer vacation | People did not make too much noise at the party |
| If the banks cut interest rates then the GDP increases | But the GDP did not increase | Cutting rates is not enough in an economic decline | The banks did not cut interest rates |

# Mental simulation and the construction of informal algorithms 

Sangeet Khemlani ${ }^{1}$ and Phil Johnson-Laird ${ }^{2}$<br>sunny.khemlani.ctr@nrl.navy.mil, phil@princeton.edu<br>${ }^{1}$ Naval Research Laboratory, Washington, DC 20375 USA<br>${ }^{2}$ Princeton University, Princeton NJ 08540 USA


#### Abstract

We describe two studies that show that when individuals who are not programmers create algorithms, they rely on mental simulations. Our studies concerned a railway domain in which carriages are rearranged - a simple environment but equivalent in computational power to a Turing machine. Participants successfully solved rearrangement problems (Experiment 1), and created algorithms to solve them (Experiment 2) and their performance corroborated the use of simulation. The participants tended to use loops and to prefer while-loops even though they are of greater computational power than for-loops. Their ability to create algorithms for abstract problems improved when they first had to create algorithms for more concrete problems. We devised a computer program that creates its own algorithms for rearrangement problems. It generates Lisp functions that operate on lists and creates descriptions of them in everyday language. The complexity of the resulting algorithms predicts participants' difficulty in devising them.


Keywords: algorithms, computer programming, creativity, deduction, problem-solving, reasoning

## Introduction

A long controversy about human thinking is whether it depends on logic (Rips, 1994), probabilities (Oakford \& Chater, 2007; Tenenbaum \& Griffiths, 2001), or mental simulations (Craik, 1943; Johnson-Laird, 1983; Hegarty, 2004). Many inferences such as syllogistic deductions can be explained by mechanisms that depend on any of the three approaches (see Khemlani \& Johnson-Laird, 2012, for a review). Indeed, few inferential tasks unequivocally depend on one approach. Computer programming may be such a task: it is readily explained by appealing to mental simulation (Bornat, Dehnadi, \& Simon, 2008; Caspersen, Bennedsen, \& Larsen, 2007; Kurland \& Pea, 1985). To debug faulty code, programmers have to mentally simulate the algorithm to discover the situations in which the computer failed to produce the expected output. It is less apparent how mental rules of logic or probabilities could be used develop algorithms. Logic can be used to deduce the consequences of a program, but the creation of a program goes beyond logic (cf. Gulwani, 2010; Kitzelmann, Schmidt, Mühlpfordt, \& Wysotzki, 2002). Probabilities hardly enter into the process, because computer programs are deterministic, and the language of the probability calculus is ill equipped to operate over the structures of programs. Mental simulation is therefore an appropriate framework with which to characterize the ability to create algorithms, and researchers can benefit from studying the simulations programmers use to solve tasks (Holt, BoehmDavis, \& Schultz, 1987).

Expert programming depends on more than just mental simulation, however. Programmers often have specialized knowledge of programming languages, of relevant software platforms and tools, and about computer science in general (Boehm-Davis \& Ross, 1992). For that reason, many studies have tested the ability of novice programmers to write computer programs (see, e.g., Anderson, Pirolli, \& Farrell, 1988). Few have investigated how those without any background in programming try to create algorithms. Miller (1974) pioneered such studies. He examined the way college students unfamiliar with computers wrote instructions for others to follow, and found that they tended not to use loops in their instructions, even though they could understand them (Miller, 1981). More recently, Pane and his colleagues carried out a study in which they presented nonprogrammers with static descriptions of an agent moving in a popular video game, PacMan, and the participants had to summarize how agents moved in general. They again preferred not to make use of loops, but when they did, they appeared to rely on while-loops (Pane et al., 2001).

Despite these results, there exists no psychological theory of how non-programmers construct algorithms. To develop such a theory, and to study algorithmic creativity in nonprogrammers, we designed a novel problem-solving task environment in which reasoners have to sort the order of a list in various ways. We introduce the environment below, and then explain how individuals build kinematic mental models to construct algorithms for their solutions. We then describe two experiments that show that reasoners intuitively understand the environment (Experiment 1) and that they can mentally create algorithms for the problems in the environment (Experiment 2).

## Rearrangement problems and the railway environment

We studied how individuals who have never learned computer programming create algorithms in everyday language. For problems that they readily understood, we used the railway environment shown in Figure 1. The environment consists of a railway track and a siding. It is an analog of a finite-state device with two stacks - the left track (a) holds the input and also acts as a stack, the siding (b) acts as another stack, and the right track (c) holds the final output. Participants' task is to move the cars from the left track to the right track into a specific order. Cars can move only from the siding to and from the left track, and from left track to right track. Multiple cars can be moved at once, i.e., any move of a selected car applies to all cars in front of it. For example, in Figure 1, if you moved the E car


Figure 1. The railway domain with an example of an initial configuration in which a set of cars is on the left side (a) of the track, the siding (b) can hold one or more cars while other cars are moved to the right side of the track (c).
to the right track, then the F car would move along in front of it. To restrict the environment to a single stack, cars could move from the siding only to the output on the right track. In summary, only three sorts of move are possible in the railway environment:

R: one or more cars moved from left track to right track.
S : one or more cars moved from left track to siding.
L : one or more cars moved from the siding to left track.
One constraint is that cars can be neither removed nor added to trains in our rearrangement problems - if they could be, then the railway environment would be equivalent to a universal Turing machine power.

Experiment 1 below investigated all 24 possible rearrangements of four cars, and examined whether the participants perseverated, i.e., made one or more unnecessary moves. They can use a simple variant of "means-ends" analysis in which they work backwards from the required goal, invoking operations relevant to reducing the difference between the current state and the goal (e.g., Newell \& Simon, 1972; Newell, 1990). For rearrangement problems, they need only envisage each successive car in the goal. Suppose, for instance, they have to re-arrange the order ABCD into ACBD . The starting state is: $\mathrm{ABCD}[$ ], where the square brackets denote the contents of the siding, which is empty at the start. Their immediate goal is to get D to the far end of the right track: [ ] . . . D. So, they move D from left to right track: ABC [ ]D. The next partial goal is to get B to the right track, and so they need to move C out of the way onto the siding: $\mathrm{AB}[\mathrm{C}] \mathrm{D}$. Now, they can move B to the right: $\mathrm{A}[\mathrm{C}] \mathrm{BD}$. They move C off the stack: $\mathrm{AC}[\mathrm{BD}$. The next move is intriguing. They should move both A and C together from left to right track. But, if reasoners perseverate, they may move only C to the right track. Their solution won't be minimal, because they then have to make a separate move of A to right track.

We investigated how reasoners solve single instances of such problems, but our primary goal was to understand the processes and representations non-programmers use to create algorithms. In the following section, we explain how kinematic mental models can be used to construct algorithms, and illustrate the predictions that the modelbased theory makes.

## A model-based theory of algorithmic creativity

How do naïve individuals create informal algorithms? We hypothesize that individuals simulate solutions to problems, where a simulation consists of a sequence of kinematic mental models representing states of the world, real or imaginary, and the sequence itself represents a logical or temporal order of the states (Johnson-Laird, 1983, Ch. 15). Reasoners use such simulations to carry out three separate steps to create an algorithm: 1) they solve at least two different instances of a rearrangement problem using a kinematic sequence of moves; 2) they scan the kinematic sequences to abduce a pattern; 3) they translate the pattern into a verbal description. We address the three steps in turn.

Step 1: Problem-solving as simulation. The first step is to solve two different instances of a rearrangement problem. Otherwise, re-arrangements are ambiguous. At any point in the simulation, only a single move is made, and so to reverse, say, four carriages, reasoners can begin by envisaging the transformation from the start state:

$$
\mathrm{ABCDEF}[] \rightarrow[\mathrm{]} \ldots \mathrm{~A}
$$

This partial goal calls for a move of five cars onto the siding, $\mathrm{A}[\mathrm{BCDEF}]$, so A can be moved to right track, [BCDEF]A. The next partial goal is to get B to right track, and so it should be moved to left track, $\mathrm{B}[\mathrm{CDEF}] \mathrm{A}$, and over to right track, [CDEF]BA. A repeated loop of these two operations moves each car in turn off the siding and to right track, and solves the problem.

Two variables should affect performance in the solution of rearrangement problems: the number of moves and the number of their operands. Obviously, the greater the number of moves, the more difficult a problem should be - the only sort of theory that would not make this prediction would be one that made no appeal to simulation. A more subtle prediction concerns the number of operands. In a reversal problem, such as the one above, each move after the first has an operand of one car. We can contrast this case with the solution of a palindrome problem, such as:

## ABCCBA[ ] $\rightarrow$ [ ]AABBCC

There are three cars, BCC , on the left that match the goal, but they are blocked, and so to solve the problem, the blocking cars are moved onto the siding: $\mathrm{ABCC}[\mathrm{BA}]$. The three cars on the left are moved to the right: $\mathrm{A}[\mathrm{BA}] \mathrm{BCC}$. One car on the siding matches the goal, and so it is moved to the left: $\mathrm{AB}[\mathrm{A}] \mathrm{BCC}$. Two cars on the left match the goal, and so they are moved to the right: $[\mathrm{A}] \mathrm{ABBCC}$. The car on the stack matches the goal, and so it is moved to the left and then over to the right, and the problem is solved. Its minimal solution required a total of 10 cars to be moved in 6 moves. This solution has a mean number of operands per move greater than that for the reversal problems, and so the theory predicts that the palindrome problems should be more difficult than reversal problems of the same number of
moves. And individuals may make an unnecessary move in their solution of the problem, i.e., they may fail to solve the problem parsimoniously. Number of operands has a family resemblance to "relational complexity", which concerns the number of arguments in a relation, and which affects problem difficulty (Halford, Wilson, \& Phillips, 1998). However, the number of operands concerns, not the number of arguments of an operator, but whether the value of a single argument is one or more entities.

Step 2: Pattern abstraction and abduction. The second step in creating an algorithm is to recover the structure of the solutions - the loop they contain, and any operations before or after it. Consider the moves to reverse trains of four and five cars, respectively:

## (S3 R1 L1 R1 L1 R1 L1 R1) (S4 R1 L1 R1 L1 R1 L1 R1 L1 R1)

where 'S3' means move three cars from left track to the Siding, 'R1' means move one car from left track to Right track, and 'L1' means move one car from the siding to Left track. The loop of operations is (R1 L1). But, how many times should it be iterated? There are two ways to find the answer. The simpler is to observe the conditions in the simulation when the loop ceases, respectively:

```
D[ ]CBA
E[ ]DCBA
```

In both cases, the siding is empty, and so this condition determines that a while-loop should continue until the siding is empty. The alternative answer depends on computing the number of times that a for-loop should be executed, and it calls for the solution of a pair of simultaneous linear equations to obtain the values of $a$ and $b$ in:

$$
\text { number-of-iterations }=\mathrm{a} * \text { train-length }+\mathrm{b} .
$$

Step 3: Conversion to natural language. The third and final step is to map the structure of the solution into a description. A general algorithm for reversing the order of cars applies to trains of any length. Hence, it needs to describe a loop of moves. When reasoners convert the algorithm to a natural language description, their responses should yield the condition in which the loop stops (an indication that they've constructed while-loop) or else reflect the number of times for which the loop should be executed (an indication that they've constructed a for-loop). The solution of simultaneous equations calls for more than just simulation, whereas the halting conditions of a loop can be observed in a simulation, and so the theory predicts that correct responses should tend to use while-loops more often than for-loops.

We have implemented all three steps in a computer program that discovers and outputs algorithms to solve any re-arrangement problem that depends on a single loop. It
outputs a for-loop, a while-loop, and a translation of the while-loop into informal English (see Appendix). Each of these algorithms solves any instance of the relevant class of rearrangements.

Experiment 1 tested whether solutions to rearrangements depend on the number of moves and the number of operands. Experiment 2 tested whether reasoners use simulation to construct algorithms, and therefore formulate while loops, and whether the theory predicts the relative difficulty of different sorts of problem.

## Experiment 1

Experiment 1 tested the effects of number of moves and number of operands on the solution of simple rearrangement problems in the railway environment. The problems were simple and called for the rearrangement of only four cars. Hence, our interest was in whether the participants could solve the problems without making redundant moves. The participants had to solve all the 24 possible rearrangements of trains containing four cars. Their minimal solutions call for various numbers of moves ( $1,4,5,6,7$, or 8 ), and as a consequence the theory predicts an increasing trend in redundant moves for these problems. The total numbers of operands in minimal solutions was $(4,6,8,10$, or 12 ), and as a consequence there should be an increasing trend in redundant moves. Because these two variables are only partially correlated, we were able to examine their effects independently (see Table 1 below).

## Method

Participants. Twenty undergraduate students at Princeton University served as participants, and none had had any prior training in logic or computer science.

Design and procedure. Participants acted as their own controls and carried out all 24 problems, which were presented in a different random order to each of them. When they had completed the experiment, they carried out two of the problems again, but they had to think aloud as they did so. They were tested individually, and carried out the experiment on a PC running LispWorks 4.4. They interacted with the system using the mouse and the keyboard of the computer. They were shown a three-minute instructional video that guided them through the elements of the railway environment, and that presented the instructions. The key instruction stated that they should try to solve each problem with as few moves as possible.

## Results and discussion

Non-programmers were able to solve rearrangement problems with ease: they produced very few incorrect solutions. Table 1 presents the participants' mean numbers of moves to solve the problems depending on the minimum number of moves and the total number of operands. We dropped the two extreme problems from the statistical

|  | Total number of operands <br> \# of moves in <br> a minimal |  |  |  |  | (cars) <br> moved in minimal solutions |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| solution | $\mathbf{4}$ | $\mathbf{6}$ | $\mathbf{8}$ | $\mathbf{1 0}$ | $\mathbf{1 2}$ | Mean \# of <br> actual moves |  |  |
| 1 | 1.0 |  |  |  |  | 1.0 |  |  |
| 4 |  | 4.3 | 4.7 | 4.6 |  | 4.5 |  |  |
| 5 |  | 5.5 | 5.2 |  |  | 5.4 |  |  |
| 6 |  |  | 6.5 | 6.6 |  | 6.6 |  |  |
| 7 |  |  | 8.9 |  | 8.9 | 8.6 |  |  |
| 8 | 1.0 | 4.9 | 6.5 | 6.9 | 8.6 | 8.4 |  |  |
| Mean \# of |  |  |  |  |  |  |  |  |
| actual moves | 1.0 |  |  |  |  |  |  |  |

Table 1. The mean numbers of moves in Experiment 1 in rearrangement problems as a function of the total number of moves in their minimal solutions and the total number of operands (cars) to be moved.
analysis so that they would not bias the results, i.e., the problem that required only one move to solution, and the problem that had a total of 12 operands. Given that the participants solved the problems, it is hardly surprising that the mean number of the participants' moves increased with the minimal number of moves required to solve a problem (Page's trend test, $\mathrm{L}=1809.5, \mathrm{z}=8.47, \mathrm{p}<.0001$ ). But, the results also showed that their mean number of moves also increased with the number of operands (Page's trend test, L $=276, \mathrm{z}=5.69, \mathrm{p}<.0001$ ). In other words, the participants tended to fail to find minimal solutions, and as the mean number of operands increased so the number of their moves increased, independently of the total number of moves in a minimal solution. (For brevity, we spare readers the latency results, but their patterns corroborated both of these effects.) There was a reliable tendency for the participants to make redundant moves. Every participant made at least one redundant move (Binomial, $p=.5^{20}$ ).

In summary, the experiment shows that naive individuals can solve simple rearrangements. It corroborated the prediction that the number of moves affected the difficulty of the problem, and thereby supported simulation-based accounts. Likewise, it corroborated the prediction unique to the model-based theory that the number of operands should affect the difficulty of a problem. The following experiment tested whether non-programmers could formulate general solutions for rearrangement problems.

## Experiment 2

In Experiment 2, the participants had to formulate algorithms to solve three sorts of rearrangement: reversals, such as ABCDEFGH becomes HGFEDCBA; palindromes, such as ABCDDCBA becomes AABBCCDD; and parity sorts, such as ABCDEFGH becomes ACEGBDFH. Participants had to construct the algorithms in their mind's eye with no access to the railway environment. They were familiar with the environment, because they had just solved five practice problems on it, but these problems were simple rearrangements that differed from the problems in the experiment proper. They were then shown the inputs and outputs for each of the problems, and they had to write down algorithms for solving them. They did so for fixed-
length problems in which trains of eight cars had to be rearranged, and indefinite-length problems in which trains of any number of cars had to be rearranged. The fixedlength problems should be easier than indefinite-length problems, because only the former can be solved without loops. Likewise, complexity and number of operands predict a trend in difficulty over the three sorts of general rearrangements: reversals should be easier than palindromes, which in turn should be easier than parity sorts. The latter should be the hardest to solve because they call for an extra operation in their algorithm (see the Appendix).

## Method

Design and materials. The participants acted as their own controls and carried out six problems: the three sorts of rearrangement as both fixed-length problems of eight cars and indefinite-length problems of any number of cars. The session began with five practice problems akin to those in Experiment 1, which the participants merely had to solve by interacting with the railway system. These problems were unrelated to the experimental problems: each of them had a train of 6 cars, and a solution depending on 8 moves. The experiment proper followed, and the participants' task was to type out a procedure that would solve each problem, but they were not allowed to interact with the railway environment. They carried out two blocks of trials, one of the definite problems and one of the indefinite problems, presented in a counterbalanced order to two groups of participants. The order of the three sorts of rearrangement was randomized for each participant within the blocks. For the indefinite-length problems, the participants were told that a car containing an ellipsis stood in place for any number of cars that had the same pattern.

Participants and procedure. Twenty students from the same population as before took part in the experiment. They watched an instructional video and were told how to interpret the car containing an ellipsis. They then solved the practice problems using the same procedure as before. In the experiment proper, the participants were told to write a description of a procedure for solving each of the experimental problems as efficiently as possible. They were free to use their own words in any way that they wanted, but they no longer were allowed to manipulate the cars in the railway environment.

## Results and discussion

Two independent raters scored the correctness of the algorithms and whether they contained a while-loop, a for loop, or no loop whatsoever (see Appendix for examples of correct responses). Inter-rater reliability was high for judgments of correctness (Cohen's $\kappa=.82$ ) and the sorts of loops that participants devised $(\kappa=.73)$. A third independent rater resolved the disagreements. Performance with the fixed-length problems was at ceiling ( $90 \%$ correct)
a.
b.


Figure 2. The percentages of correct algorithms (panel a) and the response times in s (panel b) for the indefinite-length problems as a function of the sort of rearrangement, and whether they occurred in the first or second block of trials.
and much better than the indefinite-length problems (52\% correct; Wilcoxon test, $\mathrm{z}=3.5, \mathrm{p}=.0004$; Cliff's $\delta=.64$ ). Figure 2 accordingly shows only the performance for the indefinite-length problems, and the Appendix provides examples of participants' correct algorithms. The three sorts of rearrangement yielded the predicted trend in accuracy and in the time to respond (see Appendix; Page's trend tests, zs $>3.08$, ps $<.002$ ). Likewise, the participants used many more while-loops ( $74 \%$ of correct solutions) than for-loops ( $26 \%$ of correct solutions) for indefinite-length problems. The use of while-loops correlated with accuracy ( $\mathrm{r}=.32$, p $<.0005$ ), whereas the use of for-loops did not ( $\mathrm{r}=.14$, $\mathrm{p}=$ .10). The differences in ability were striking: the best participant created a correct algorithm for every problem, whereas the worst did so for only a third of the fixed-length problems and for none of the indefinite-length problems.

## General Discussion

The ability to create algorithms might seem to be a case of competence in pure mathematics with little relation to everyday life. Problems in rearranging cars in toy trains may similarly seem remote from the exigencies of daily life. However, algorithmic thinking is regularly called for, e.g., in laying place settings on a table, in determining kinship relations, in following a recipe or a set of instructions. Other sorts of algorithmic thinking are needed to determine the consequences of knitting patterns, instructions for kits, maintenance manuals, and, above all, algorithms in computer programs.

Algorithmic thinking is easier when you can manipulate an external environment and solve a problem using only partial means-ends analysis, i.e., you can use the railway environment and solve a rearrangement of the cars in a train, one car at a time (Experiment 1). But suppose that your task is to devise an algorithm for the general problem of sorting cars in this way - so that cars in odd-numbered positions precede cars in even-numbered positions. The algorithm for this task is not obvious. According to the present theory, the way that you carry it out is to make another simulation so that you can figure out what is going on. You should then
notice that there is a loop of two operations (move one car to the right, and then one car onto the siding) that has to be repeated while more than two cars remain on the left track. It follows that while-loops should occur more often than for-loops in putative algorithms, because it is easier to envisage halting conditions for while-loops from simulations than to use them to compute the number of iterations for a for-loop. The difficulty of the task also depends on the Kolmogorov complexity of the program, as indexed in the number of its instructions (in Lisp or in everyday language), and on the number of operands (Experiment 2).

Computer scientists often complain about the lack of any valid test of the likely ability of naive individuals as computer programmers (e.g., Bornat, Dehnadi, \& Simon, 2008). The rearrangement problems in our experiments may provide the basis for such a test. At the very least, we now know that individuals differ reliably in their ability both to solve problems in the railway domain (Experiment 1), and to formulate informal algorithms for their solutions (Experiment 2). The question remains as to whether such tasks are reliable predictors of ability. Mathematicians, logicians, and computer programmers, learn to reason about the repeated loops of operations that are needed in recursive functions. Previous studies have examined how novice programmers cope with such reasoning in trying to specify algorithms in a programming language (see, e.g., Anderson \& Jeffries, 1985). Our studies have shown that naive individuals with no training in computer programming are able to make simulation-based deductions, to solve rearrangement problems, and even to abduce informal algorithms for their general solution.

The evidence we have reported corroborated the theory based on mental models. To the best of our knowledge, no other theory of naïve algorithmic creativity exists. But, a theory could be developed in principle from an axiomatization of the domain in first-order logic (see, e.g., McCarthy \& Hayes, 1969; McCarthy, 1986; Rips, 1994). A typical axiom would capture the effects of a move, e.g.:

For any $\mathrm{x}, \mathrm{y}$, if x is a car $\& \mathrm{y}$ is a train $\& \mathrm{z}$ is a train $\& y$ is on right track $\& z$ is on left track \& $x$ is at the front of $y \& R 1$ is carried out then $x$ is at back of $\mathrm{z} \& \operatorname{not}(\mathrm{x}$ is at front of y ).

No one has proposed such an account, and so it is not yet possible to pit it against the model-based theory. But, we cannot rule it out, and remark only that the approach runs into difficulties. Our participants' think-aloud protocols raise problems for it, because they report moving cars around in a mental simulation of the railway environment. Likewise, their reliance on simulations predicts their use of while-loops in algorithms, because simulations yield the halting conditions for while-loops more readily than the number of iterations for for-loops. These results seem difficult, if not impossible, to explain without recourse to the use of mental simulations.

## Acknowledgements

This research was supported by a National Science Foundation Graduate Research Fellowship to SSK and by NSF Grant No. SES 0844851 to PJL to study deductive and probabilistic reasoning. We are grateful to Monica Bucciarelli, Sam Glucksberg, Adele Goldberg, Geoffrey Goodwin, Louis Lee, David Lobina, Max Lotstein, Robert Mackiewicz, Paula Rubio, and Carlos Santamaria, for their helpful comments and criticisms.

## References

Anderson, J.R., \& Jeffries, R. (1985). Novice Lisp Errors: Undetected losses of information from working memory. Human-Computer Interaction, 1, 107-131.
Anderson, J. R., Pirolli, P., \& Farrell, R. (1988). Learning to program recursive functions. In M. Chi, R. Glaser, \& M. Farr (Eds.), The nature of expertise (pp. 153-183). Hillsdale, NJ: Erlbaum.
Boehm-Davis, D., \& Ross, L. (1992). Program design methodologies and the software development process. International Journal of Man-Machine Studies, 36, 1-19.
Bornat, R., Dehnadi, S., \& Simon (2008). Mental models, consistency and programming aptitude. Proceedings of the Tenth conference on Australasian Computing Education Conference, 10, 53-61.
Caspersen, M.E., Bennedsen, J., \& Larsen, K.D. (2007). Mental models and programming aptitude. ACM SIGCSE Bulletin, 39.
Corballis, M. (2011). The recursive mind. Princeton: Princeton University Press.
Craik, K. (1943). The Nature of Explanation. Cambridge, UK: Cambridge University Press.
Gulwani, S. (2010). Dimensions in program synthesis. In Proceedings of the 12th International ACM SIGPLAN Conference. Hagenberg, Austria.
Halford, G.S., Wilson, W.H., \& Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for
comparative, developmental, and cognitive psychology. Behavioral and Brain Sciences, 21, 803-865.
Hegarty, M. (2004). Mechanical reasoning as mental simulation. Trends in Cognitive Sciences, 8, 280-285.
Holt, R.W., Boehm-Davis, D., \& Schultz, A. (1987). Mental representations of programs for students and professional programmers. In Empirical Studies of Programmers: Second Workshop (pp. 33-46). Ablex Publishing Corp.
Johnson-Laird, P.N. (1983). Mental models. Cambridge: Cambridge University Press.
Khemlani, S., \& Johnson-Laird, P.N. (2012). Theories of the syllogism: A meta-analysis. Psychological Bulletin, 138.
Kitzelmann, E., Schmidt, U., Mühlpfordt, M., \& Wysotzki, F. (2002). Inductive synthesis of functional programs. In Calmet, J., Benhamou, B., et al. (Eds.) Artificial Intelligence, Automated Reasoning, and Symbolic Computation. New York: Springer.
Kurland, D. M., \& Pea, R. D. (1985). Children's mental models of recursive LOGO programs. Journal of Educational Computing Research, 1, 235-244.
Miller, L. (1974). Programming by non-programmers. International Journal of Man-Machine Studies, 6, 237-260.
Miller, L. (1981). National language programming: Styles, strategies, and contrasts. IBM Systems Journal, 20, 184-215.
Newell, A. (1990). Unified theories of cognition. Cambridge, MA: Harvard University Press.
Newell, A., \& Simon, H.A. (1972). Human problem solving. Englewood Cliffs, NJ: Prentice-Hall.
Oaksford, M., \& Chater, N. (2007). Bayesian rationality. Oxford: Oxford University Press.
Pane, J.F., Ratanamahatana, C.A., \& Myers, B.A. (2001). Studying the language and structure in non-programmers' solutions to programming problems. International Journal of Human-Computer Studies, 54, 237-264.
Rips, L.J. (1994). The psychology of proof. Cambridge, MA: MIT Press.
Tenenbaum, J. B., \& Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. Behavioral and Brain Sciences, 24.

Appendix. Natural language solutions (as outputted by the computer program for abducing them) to three sorts of general problem: reversals, palindromes, and parity sorts, and examples of correct algorithms created by participants; and the percentage of participants' algorithms that correctly solved the given problems in Experiment 2.

| Problem | Automatically generated algorithms | Examples of correct algorithms | \% Correct |
| :---: | :---: | :---: | :---: |
| Reversal | 1 Move one less than the cars to siding. 2 While there are > zero cars on siding 3 ...move one car to right track 4 ...move one car to left track. <br> 5 Move one car to right track. | " $i$ 'll move everything in the side track. then i'll move each letter back onto the left track and then to the right track." (Participant 14) | 90\% |
| Palindrome | 1 Move one less than half the cars to siding. 2 While there are > two cars on left track 3 ...move two cars to right track 4 ...move one car to left track. <br> 5 Move two cars to right track | "step1: cut the train into half, move the right half to siding <br> step2: for both half trains on the left and siding track, move a pair of carts of the same letter to the right. Continue doing so until all the carts are on the right track." (Participant 1) | 68\% |
| Parity sort | 1 While there are > two cars on left track <br> 2 ...move one car to right track <br> 3 ...move one car to siding. <br> 4 Move one car to right track. <br> 5 Move one less than half the cars to left track <br> 6 Move half the cars to right track | "Move the rightmost car to the right track, and move the next car to the side track. Continue alternating between right track and side track until the left track is empty. Then move all cars from the side track to the left track, and then to the right track." (Participant 7) | 55\% |

# Channels of multimodal communication: Relative contributions to discourse understanding 

Andrej A. Kibrik (aakibrik@gmail.com)<br>Institute of Linguistics of the Russian Academy of Sciences and Lomonosov Moscow State University B. Kislovskij per. 1, Institute of Linguistics RAN, Moscow 125009 Russia<br>Natalia B. Molchanova (natascha.molchanova@gmail.com)<br>BearingPoint<br>B. Ordynka 40-4, Moscow 119017 Russia

Keywords: discourse; communication channels; understanding; multimodality

## 1. Introduction: Communication channels

The mainstream view of linguistic form, characteristic of modern linguistics, can be formulated as follows: language consists of hierarchically organized segmental units, such as phonemes, morphemes, words, phrases, and sentences. Mainstream linguistics thus equates linguistic form with verbal form, that is, the segmental vocal material. However, as we all know, apart from sound, there are other channels (or components) of communication, in the first place through vision. The visual channel is what is sometimes named with the cover term body language, including gesture, mimic, gaze, posture, etc. (see e.g. McNeill, 1992; Kendon, 1994; Goldin-Meadow, 1999; Krejdlin, 2002; Butovskaja, 2004; Andersen, 2007; Burgoon et al., 2011).

Furthermore, the vocal material is not exhausted by verbal elements. There is also prosody, that is, non-verbal (= nonsegmental) aspects to sound, including intonation, tempo, pausing, loudness, discourse accents, tonal registers, etc. (see e.g. Cruttenden, 1986; Kodzasov, 2009).
An unbiased view should probably be the following: all of these components must be taken into account in a realistic model of communication. For example, imagine that you are staying in a hotel room with thin walls and can hear people next door talking. You cannot hear words (the verbal component) but you can hear prosody, and you get something about the conversation, for example you may know that the people are quarreling. On the other hand, prosody-free talk, as sometimes heard from TV simultaneous interpreters on the Euronews channel, is unnatural and hinders comprehension. In this study we address the question of the relative contribution of the various communication channels or components to the overall comprehension of spoken discourse.

## 2. Views on the importance of various communication channels

The traditional approach of mainstream linguistics has been to consider the verbal channel so central that prosody and the visual channel have often been downgraded as "paralinguistics". Many contemporary textbooks in linguistics barely mention prosody and do not mention gesture and body language at all (see e.g. Hall, 2005).

The other extreme is represented by the view common in applied psychology that words matter less than prosody and especially than body language. It is very often that the following figures are quoted, going back to Mehrabian (1971): body language conveys $55 \%$ of information, prosody conveys $38 \%$ of information, and the verbal component only $7 \%$ of information ${ }^{1}$, see e.g. http://jobsearch.about.com/od/interviewsnetworking/a/nonv erbalcomm.htm. According to this view, "words may be what men use when all else fails" (Krejdlin, 2002: 6).

Most likely, the truth lies between these two extremes. All of the communication channels must be valuable and none can be negligible. This kind of balanced approach is characteristic of the modern multimodal paradigm (see e.g. Granström et al. eds., 2002; Norris, 2004; Ventola et al. eds., 2004; Bengio \& Bourlard eds., 2005; Royce \& Bowcher, 2007; Jewitt ed., 2011). According to Kress (2002), "A multimodal approach assumes that the message is 'spread across' all the modes of communication. If this is so, then each mode is a partial bearer of the overall meaning of the message." To use a quotation from the computational domain, "within biology, experimental psychology, and cognitive neuroscience, a separate rapidly growing literature has clarified that multisensory perception and integration cannot be predicted by studying the senses in isolation" (Cohen \& Oviatt, 2006). Kibrik (2010) described the research program of multimodal linguistics, taking into account all of the communication channels in an integrated approach.

Taking up the challenge of Mehrabian (1971), in this study we try to numerically estimate the contribution of each communication channel into the overall process of message understanding. (Cf. two early psychological studies Walker, 1977 and Hollandsworth et al., 1979, arriving at rather opposite conclusions, and also Cutica \& Bucciarelli, 2006.)

[^306]
## 3. Experimental design

The experimental design, first developed by Andrej Kibrik in 2006 at the Deparment of Theoretical and Applied Linguistics, Philological Faculty, Moscow State University (see Kibrik \& El'bert, 2008, Kibrik, 2010), consists of several elements. For the purposes of this study, we differentiate between three communication channels, or components, including two vocal channels, the verbal and the prosodic ones, and the visual channel comprising all elements of body language; see Figure 1.


Figure 1: Three communication channels.
If we take a sample of natural discourse, we can isolate three communication channels. For example, if we have a recording of communication, a video without sound is equivalent to the visual channel alone. We also need to isolate the verbal channel and the prosodic channels; specific technical ways of how that can be made possible are explained in sections 4 and 5 below. Assuming that the three channels have been isolated, we can produce eight $\left(2^{3}=8\right)$ variants of the sample discourse and present them to separate groups of experimental participants. These eight variants include three in which only one channel is represented, three in which two channels are represented, one with the three channels (the original material), and the null variant in which nothing has been shown to participants. We will thus need eight groups of participants, each presented with one of the eight kinds of experimental material.
The null variant of the experimental discourse and the corresponding group is necessary in order to evaluate which part of the overall content can be inferred on the basis of background knowledge and common sense.

At the next stage the degree of the participants' understanding of the discourse can be assessed with the help of a questionnaire, and such assessment may be used as an estimate of a communication channel's contribution to the overall discourse understanding.

## 4. Experiment A: movie-based material

The first line of studies in this paradigm was implemented in a series of experiments by Ekaterina Èl'bert, particularly in her diploma thesis (2007), and further reinterpreted and refined in Kibrik and El'bert (2008). In this line of studies the decision was made to use an excerpt of a movie as experimental discourse. Specifically, the Russian TV serial "Tajny sledstvija" ("Mysteries of the investigation") was
used. The experimental excerpt ran for 3 minutes and 20 seconds, and it was preceded by a 8 minutes context excerpt, starting from the beginning of a series. The experimental excerpt fully consisted of a conversation, to ensure that we are testing the understanding of discourse rather than of the film in general.

The two vocal channels were separated from each other through the following procedures. The verbal channel was presented in the written mode, by means of temporally aligned running subtitles. The prosodic channel was obtained from the original sound by superimposing a filter creating the "behind a wall" effect. Figure 2 illustrates a snapshot from the experimental type "visual plus verbal", in other words, video plus running subtitles.


Figure 2: Frame from the experimental material "visual plus verbal".

99 participants took part in the study, divided into eight groups, each group comprising 10 to 17 persons. All eight groups watched the identical context excerpt. As for the experimental excerpt, each of the eight groups had access to different material. The null group did not see anything apart from the context excerpt, three groups only had access to one communication channel of the experimental excerpt (either verbal or prosodic or visual), the other three groups to two communication channels (verbal+prosodic = original sound; verbal+visual = video and subtitles, see Figure 2; prosodic+visual $=$ video and filtered sound), and the eighth group watched the original version of the experimental excerpt.

The context and the experimental excerpts were shown to the whole group of participants on a large screen. Each participant was instructed to attend the context and the experimental excerpt and then answer a set of questions concerned with the experimental excerpt alone. The questionnaire was constructed in accordance with the received principles of test tasks (Panchenko, 2000). There were 23 multiple-choice questions in the questionnaire; a participant was supposed to choose only one answer out of four listed variants. Here is an example of a question, along with the offered answers (translated from the Russian original):

What Tamara Stepanovna offers Masha before the beginning of the conversation:
a. to take off her coat
b. to have a cup of tea
c. to have a seat
d. to have a drink

One of the available answers (in this particular case, c) was correct, two were plausible but wrong ( $\mathrm{a}, \mathrm{b}$ ) and one implausible (d); the latter was aimed at filtering out incompetent participants.

## Results of Experiment A

Percentage of correct answers was used as a way to assess a participant's degree of discourse understanding. The summarized results are shown in a diagram in Figure 3.


Figure 3: Degrees of discourse understanding in Experiment A.

We see from the second, third, and fourth columns in Figure 3 that each individual communication channel is substantially informative: The verbal channel is leading in this respect ( $72.4 \%$ correct answers), but the two other (prosodic: $51.5 \%$, visual: $61.7 \%$ ) also significantly (MannWhitney test, $\mathrm{p}<0.05$ ) prevail over the null condition (leftmost column, $38.3 \%$ ). The hierarchy of the individual channels turns out verbal>visual>prosodic (significant according to Kruskal-Wallis test $(\mathrm{H}(2,69)=24.2$, $\mathrm{p}<0.01)$ ). In spite of the prevalence of the verbal channel, the difference in the contributions of individual channels is not dramatic, and second, the degree of understanding in the "verbal alone" condition is significantly (Mann-Whitney test, $\mathrm{p}<0.05$ ) lower than in the original material condition (three channels in conjunction, the rightmost column in Figure 3, 87.4\%).
Another conclusion from the results of Experiment A concerns the comparison of the three groups that had access to two communication channels; see columns fifth to seventh from the left. There is a very noticeable (but not reaching the level of statistical significance) dip in the condition "visual+prosodic" (51.6\%), compared to two other pairwise combinations (verbal+prosodic: 70.7\%; verbal+visual: $77.8 \%$ ). Apparently, that dip means that language users have difficulties integrating information from the visual and prosodic channels, in the absence of
verbal material. In a natural setting, this condition can be compared to observing communication via a glass that is penetrable for prosody but blocks the verbal material. Most likely, the dip in the "visual+prosodic" condition is due to the unusual character of such situations in real life, as well as to the participants' inability to integrate information from the visual and prosodic channels in the absence of verbal material.

## 5. Experiment B: conversation-based material

At the following stage of the project, we modified and/or improved a number of the methodological decisions made in Experiment A, including the kind of stimulus material, the technical methods of isolating the prosodic channel and the verbal channel, the questionnaire, and the interviewing procedure. The below description of Experiment B is organized as follows. Each of the mentioned methodological decisions made in Experiment A is assessed, and a modification/improvement realized in Experiment $B$ is presented.

Several problems of the movie-based stimulus material, used in Experiment A, were detected, including the following. First, the plot of the movie in certain instances facilitated guessing by the experiment participants. Second, it was not possible to exclude the familiarity of the movie to some of the participants. Third, the quasi-natural behavior of the actors could affect the results. Fourth, all speakers were of the same gender (women) which made it difficult for the participants to distinguish between voices, especially in the "prosodic alone" condition.

The solution realized in Experiment B was to employ a recording of natural dialogue between two speakers. In order to make the dialogue structured and predictable, a guessing game "Little garages" was recorded. One of the speakers, a woman, was laying a number of toothpicks on the table and was asking the guesser, a man: "How many little garages?" The guesser was trying to figure out how to provide a correct answer, which was difficult (because the intended amount of little garages was in fact the number of the the first speaker's fingers kept on the table at the moment). The guessing process lasted for 19 minutes, out of which the stimulus material of 5 minutes and 55 seconds was produced. The stimulus material consisted of a dialogue between the two speakers, culminating in the guesser's ultimate success. A frame from the guessing game recording appears in Figure 4.

The acoustic filter used in Experiment A produced the material in the "prosodic alone" condition that was excessively noisy. The solution used in Experiment B was to radically decrease the signal at all frequencies except for the speaker's average F0 frequency. This led to a more satisfactory "behind the wall" effect.


Figure 4: Frame from the recording of the guessing game.
The main problem associated with the "verbal alone" condition in Experiment A was that the subtitles operated in the visual, rather than the vocal, mode. This had created a substantial deviation from the situation of spoken discourse, also leading to the undesired interaction and/or competition between the written verbal material and the visually perceived video material. In addition, some participants experienced difficulties in following the subtitles appearing and disappearing at the same pace as spoken words in the original material. The solution introduced in Experiment B was to produce an artificial spoken prosody-free signal. Both speakers participating in the recording were requested to individually pronounce each word that occurred in their conversation. All thus elicited words were then glued together in the right order, thus providing prosody-free discourse, devoid of intonation, reduction, differences in tempo, etc.

As far as the questionnaire is concerned, the imperfection of Experiment A is seen through the insufficient gap between the results of the null group and the original material group: $38.3 \%$ vs. $87.4 \%$. These numbers indicate that the participants were able to reconstruct the correct answer quite often and, on the other hand, even the full original material did not provide reliable access to a correct answer. In order to improve the questionnaire, a testing stage was introduced in Experiment B, in which trivial questions were identified (high null group results), as well as unfortunate questions (low original material group results). Trivial and unfortunate questions were filtered out, and the number of questions was reduced from 30 to 17 . The improved results in the two contrastive groups turned out $24.7 \%$ and $91.2 \%$ of correct answers, see below.

The interviewing procedure was improved in Experiment B. In Experiment A the participants were of various and uncontrolled age and life experience. The presence of multiple participants in the room could have led to undesirable and uncontrolled interference. Finally, the need for a large room, loud speakers, and a big screen is an unnecessary technical complication to the procedure. In Experiment B the participants were controlled for age,
geographical origin, and social status: only students of Moscow origin were recruited, which provided a homogeneous sample. They were also balanced in terms of gender. The experiment was implemented in a remote fashion: the stimulus material was posted on youtube.com, and the questionnaire at Googledocs. The guidelines closely directed the participants' sequence of actions, from one experimental part to another and from one group of questions to another, so there are reasons to believe that the procedures were very similar in all participants. All participants worked in comparable, independent, and comfortable conditions, and there was no need for technical excessiveness such as a big screen and loud speakers. 92 participants altogether took part in the experiment, out of which 20 were employed at the testing stage and 72 at the main stage (from 10 to 15 in each experimental group).

## Results of Experiment B

The quantitative results of Experiment B are shown in Figure 5.


Figure 5: Degrees of discourse understanding in Experiment B.

The main findings of Experiment B are similar to those obtained in Experiment A. All three communication channels, taken in isolation (columns two to four from the left) are substantially and comparably informative: they lead to $58.8 \%, 45.6 \%$, and $48.8 \%$ of correct answers, compare that to the $24.7 \%$ in the null group. The hierarchy of informativeness is again verbal>visual>prosodic. The conditions with two channels available (columns five to seven from the left) demonstrate the following results: $73.5 \%, 88.2 \%$, and $52.4 \%$. Compared to Experiment A, we here get a much cleaner picture as concerns the better participants' performance in the two channels conditions as contrasted with the one channel conditions. Finally, we see again a dramatic dip in the "visual+prosodic" condition: the second last column counting from the left.

## 6. Discussion

The main conclusion of Experiment B is the following: in spite of the substantial differences in the methodology from Experiment A, the results are remarkably similar. With minor differences the overall picture in Figures 3 and 5 is
very similar. This makes us believe that our conclusions about the relative contributions of various communication channels to the overall discourse understanding are fairly robust.
Now, the picture in Figure 5 is cleaner and crisper in two respects: the more obvious advantage of the two channel conditions over the one channel conditions and the better contrast between the null group and the original material group.

In order to provide a response to Mehrabian's (1971) famous (or infamous) numbers, the following method can be applied. Suppose the three communication channels are independent (this is a strong assumption, but it is necessary for calculating the relative contributions of the channels). We can sum up all percentages in the one-channel conditions and then normalize them to $100 \%$. Let us perform this operation on the results of both experiments, looking at the numbers in columns two to four from the left in Figures 3 and 5 (percentages are rounded to 1 per cent). The outcome of this procedure is shown in Table 1.

Table 1: Normalized contributions of the three communication channels.

|  |  | Experiment A | Experiment B |
| :--- | :--- | :--- | :--- |
| Summed percentages |  | $72+52+62=186$ | $59+46+49=154$ |
| Normalized <br> contributions | Verbal | $72 \%: 1.86 \approx 39 \%$ | $59 \%: 1.54 \approx 38 \%$ |
|  | Prosodic | $52 \%: 1.86 \approx 28 \%$ | $46 \%: 1.54 \approx 30 \%$ |
|  | Visual | $62 \%: 1.86 \approx 33 \%$ | $49 \%: 1.54 \approx 32 \%$ |

Once again, we see the striking similarity in the results of the two experiments: the numerically evaluated contributions of the three channels never differ from each other by more than $2 \%$. So the contributions of the channels are stable irrespective of the specifics of methodology.
Also, the gender differences between the participants were explored in Experiment B. Two particularly interesting results were obtained for the conditions "verbal alone" and "visual+prosodic"; they are shown in Table 2.

Table 2: Performance of men and women in two conditions in Experiment B (percentages of correct answers indicated)

| Condition | Men | Women | Advantage |
| :--- | :--- | :--- | :--- |
| Verbal alone | 59.1 | 69.9 | Women: +10.7 |
| Visual + prosodic | 66.1 | 51.6 | Men: +14.5 |

As is clear from Table 2, in the condition "verbal alone" the women have demonstrated a striking advantage, providing correct answers much more frequently than the men. In contrast, the men demonstrated a strong advantage in the condition "visual+prosodic" that, as was discussed above, corresponds to an unusual situation and generally creates a difficulty in comparison with other two-channel conditions. These results conform to certain generalizations about gender intelligence, such as the women's better performance in verbal tasks and men's better performance in novel situations (see e.g. Bendas, 2006).

## 7. Conclusions

This study is the first linguistically-informed demonstration of the importance of several communication channels for understanding natural discourse. The following conclusions can be drawn from the reported study.

First, all communication channels are highly significant in encoding content and understanding of discourse. Therefore, the attitude common in mainstream linguistics, according to which linguistic communication is performed mostly by the verbal component, whereas other channels are negligible, is incorrect.

Second, among the communication channels the verbal channel is the leading one. Therefore, the viewpoint popular in applied psychology, according to which the contribution of the verbal component is negligible, is erroneous as well.

Third, the specific normalized contributions of the verbal, prosodic, and visual channels are in the vicinity of $38 \%$, $30 \%$, and $32 \%$, respectively.

Fourth, participants have difficulties integrating the information from the visual and prosodic channels, in the absence of the verbal channel. This suggests that in normal communication the verbal channel plays the role of an anchor to which the information from other channels is attached.

Fifth, men and women perform differently in the conditions of isolated communication channels, women having advantage in the "verbal alone" condition and men having advantage in the novel and unusual "visual+prosodic" condition.

As was pointed out in section 5 , many questions from the original questionnaire were filtered out for certain substantial reasons, which has reduced the number of questions from 30 to 17 . In combination with the large number of conditions (eight), this has led to the fact that the quantitative tendencies observed in Experiment B do not quite reach the level of statistical significance. In April 2013 we collected additional data, bringing the number of subjects in each group to at least fifteen (total=132). We expect that, when the statistical analysis is completed, full significance of the results will be attained, as well as a formal comparison of the results of the two experiments.

A number of methodological issues remain for further research. In particular, we would like to pinpoint two of those. First, we are planning to experiment with monologic discourse addressed to public audience, such as presentations of travel agents in front of a group of people. This would complement the already attained results from our studies of dialogic communication. Second, we will keep working on refining the methods allowing to isolate the verbal channel. Both of the so far employed methods have their shortcomings, the subtitles switching from the auditory to the visual modality and the prosody-free talk being the unnatural kind of input. We will keep searching for additional methods helping to present the "verbal alone" condition in a more ecologically valid way.

A major problem in the studies of human communication and discourse is associated with the fact that different
disciplinary traditions and paradigms address communication from different angles, not consulting each other's results. Linguists usually only pay attention to the verbal component, while non-verbal communication is mostly explored by social psychologists. In this study we propose an approach that is hopefully relevant for each of the fields studying human communication and bridging the gap between them.
We would like to conclude with a quotation from Ron Scollon (2006): "Any use of language is inescapably multimodal".

## Acknowledgements

We are grateful to Olga Fedorova, Anna Laurinavičiute, and Andriy Myachykov who provided useful and important input during the preparation of this paper. The study was conducted with financial support from grant \#11-04-00153 of the Russian Foundation for the Humanities.

## References

Andersen, P. (2007). Nonverbal Communication: Forms and Functions. Long Grove: Waveland Press.
Bendas, T. V. (2006). Gendernaja psixologija [Gender psychology]. Saint-Petersburg: Piter.
Bengio, S., \& Bourlard, H. (Eds.) (2005). Machine learning for multimodal interaction. Berlin: Springer.
Burgoon, J. K., Guerrero, L. K., \& Floyd, K. (2011). Nonverbal communication. Boston: Allyn \& Bacon.
Butovskaja, M. L. (2004). Jazyk tela: priroda i kul’tura [Body language: nature and culture]. Moscow: Nauchnyj mir.
Cohen, P. R., \& Oviatt, S. L. (2006). Multimodal interaction with computers. In K. Brown (Ed.), Encyclopedia of language and linguistics, 2nd ed. Oxford: Elsevier.
Cruttenden, A. (1986). Intonation. Cambridge: Cambridge University Press.
Cutica, I., \& Bucciarelli, M. (2006). Why gestures matter in learning from a discourse. In B. M. Velichkovsky, T. V. Chernigovskaya, Yu. I. Alexandrov, \& D. N. Akhapkin (Eds.), The Second Biennial Conference on Cognitive Science. Vol. 1. Saint-Petersburg: Saint-Petersburg State University, Philological Faculty, 40-41.
Èl'bert, E. M. (2007). Vklad verbal'nogo, prosodicheskogo i vizual'nogo kanalov $v$ ponimanie ustnogo diskursa [Contribution of the verbal, prosodic, and visual channels in the understanding of spoken discourse]. Diploma thesis. Moscow State University.
Goldin-Meadow, S. (1999). The role of gesture in communication and thinking. Trends in Cognitive Science, 3, 419-429.
Granström, B., House, D., \& Karlsson, I. (Eds.) (2002). Multimodality in Language and Speech Systems. Norwell: Kluwer.
Hall, C. J. (2005). An Introduction to Language and Linguistics: Breaking the Language Spell. London: Continuum.

Hollandsworth, J. G. Jr., Kazelskis, R., Stevens, J. \& Dressel, M. E. (1979). Relative contributions of verbal, articulative, and nonverbal communication to employment decisions in the job interview setting. Personnel Psychology, 32, Issue 2, 359-367.
Jewitt, C. (Ed.) (2011). The Routledge Handbook of Multimodal Analysis. London: Routledge.
Kendon, A. (1994). Do gestures communicate? A review. Research on Language and Social Interaction, 27, 175200.

Kibrik, A. A., \& Èl'bert, E. M. (2008). Understanding spoken discourse: the contribution of three information channels. In Yu.I.Alexandrov et al. (Eds.), Materials of the 3rd International Conference on Cognitive Science, 82-84. Moscow: IP RAN.
Kibrik, A. A. (2010). Multimodal'naja lingvistika [Multimodal linguistics]. In Yu. I. Alexandrov \& V. D. Solovyev (Eds.), Kognitivnye issledovanija, issue 4. Moscow: IP RAN.
Kodzasov, S. V. (2009). Issledovanija v oblasti russkoj prosodii [Studies in the field of Russian prosody]. Moscow: Jazyki slavjanskix kul'tur.
Krejdlin, G. E. (2002). Neverbal'naja semiotika [Nonverbal semiotics]. Moscow: NLO.
Kress, G. (2002). The multimodal landscape of communication. Medien Journal, 4, 4-19.
McNeill, D. (1992). Hand and mind: what gestures reveal about thought. Chicago: University of Chicago Press.
Mehrabian, A. (1971). Silent messages. Belmont, CA: Wadsworth Publishing Company.
Norris, S. (2004). Analyzing multimodal interaction: A methodological framework. London: Routledge.
Panchenko, A. A. (2000). Razrabotka testov [Developing tests]. Khabarovsk.
Royce, T. D., \& Bowcher, W. L. (2007). New directions in the analysis of multimodal discourse. Mahwah: Lawrence Erlbaum.
Scollon, R. (2006). Multimodality and the language of politics. In K. Brown (Ed.), Encyclopedia of language and linguistics, 2nd ed. Oxford: Elsevier.
Ventola E., Charles, C., \& Kaltenbacher, M. (Eds.) (2004). Perspectives on multimodality. Amsterdam: Benjamins.
Walker, M. B. (1977). The relative importance of verbal and nonverbal cues in the expression of confidence. Australian Journal of Psychology, 29, Issue 1, 45-57.

# Quantifying the Coherence of Pedestrian Groups 

Adam W. Kiefer (adam_kiefer@brown.edu)<br>Department of Cognitive, Linguistic \& Psychological Sciences<br>Brown University<br>190 Thayer Street<br>Providence, RI 02912 USA

Stéphane Bonneaud (stephane_bonneaud@brown.edu)<br>Department of Cognitive, Linguistic \& Psychological Sciences<br>Brown University<br>190 Thayer Street<br>Providence, RI 02912 USA

Kevin Rio (kevin_rio@brown.edu)<br>Department of Cognitive, Linguistic \& Psychological Sciences<br>Brown University<br>190 Thayer Street<br>Providence, RI 02912 USA

William H. Warren (bill_warren@brown.edu)<br>Department of Cognitive, Linguistic \& Psychological Sciences<br>Brown University<br>190 Thayer Street<br>Providence, RI 02912 USA


#### Abstract

Coherent collective behavior emerges from local interactions between individuals that generate group dynamics. An outstanding question is how to quantify group coherence in order to understand the nature of these dynamics. We investigate this problem in the context of a small group of pedestrians instructed simply to walk to a goal. To measure the degree of coordination in a group, we employed principal components analysis to estimate dimensional compression, and cross-recurrence quantification analysis to estimate the coupling strength between individuals. The results indicate lower-dimensional behavior and more stable coupling in real groups compared to reshuffled virtual groups. These findings demonstrate spontaneous local coordination in pedestrian groups that gives rise to coherent collective behavior, and offer an approach for investigating group dynamics in more complex contexts.


Keywords: group locomotion; group coordination; crossrecurrence quantification; principal components analysis

## Introduction

Group dynamics arise from local interactions between individuals that are governed by a multi-level set of processes. At the most basic level, these interactions depend on a coupling between individuals based on perceptual information, which may further depend on higher-order cognitive and social constraints. To understand the emergence of collective behavior, it is necessary to begin by characterizing both the local informational coupling and the global group behavior. Such an approach requires a complementary set of analysis tools to quantify observable
properties, such as the degree and stability of coordination, at both the individual and group levels.

In the context of locomotion, we focus on the coupling between individual pedestrians that yields the formation of a coherent crowd. A recent dynamical model of locomotor behavior (Fajen \& Warren, 2003, 2007; Warren \& Fajen, 2008) has characterized both individual behavior and pedestrian interactions, including coordination in leaderfollower and side-by-side dyads (Rio \& Warren, 2011; Page \& Warren, 2012), and may be generalized to coordination in groups (Rio, Bonneaud \& Warren, 2012). Here we investigate measures of the degree of coordination in small groups, or group coherence.

Relevant behavioral variables to index the locomotor trajectory of an agent include (1) the agent's direction of travel, or heading ( $\phi$ ) and (2) the agent's speed (s). Each of these variables can be considered a degree of freedom (DoF) of pedestrian locomotion, and thus the DoF of a group of N pedestrians can be operationally defined as a system consisting of $\mathrm{N} \times 2$ DoF (i.e., $\phi$ and $\mathbf{s}$ ).

It has been proposed that behavioral coordination between two agents arises from the coupling of DoF via shared information variables (Riley, Richardson, Shockley \& Ramenzoni, 2011). Shared information between agents allows the DoF to directly regulate one another. This permits the characterization of interpersonal coordination in terms of the reduction of DoF, or dimensional compression, due to the behavioral reorganization of the newly assembled system. In the context of pedestrian interactions, a follower controls their speed by nulling change in the leader's visual
angle, and a pedestrian walking beside a neighbor controls their speed by nulling change in the neighbor's visual direction. Thus, visual information that serves to couple DoFs (i.e. $\phi$ and $\mathbf{s}$ ) gives rise to pedestrian coordination and, ultimately, coherent crowds (Bonneaud \& Warren, 2012; Moussaïd, Helbing \& Theraulaz, 2011; Ondřej, Pettré, Olivier, \& Donikian, 2010; Rio, Bonneaud \& Warren, 2012).

We aim to advance the analysis of collective behavior by developing methods to quantify the degree of coordination among pedestrians in groups. We focus on both the basic coordination mechanism - the local coupling between pairs of neighbors - and the global characteristic of group coherence. The problem then becomes how to quantify coherence as a measure of collective behavior. To that end, we must identify analysis tools that can be used to characterize coordination at multiple scales of measurement.

Principal Components Analysis (PCA) is one way to quantify the overall dimensional compression of an observed system (Riley et al., 2011). An advantage of PCA is that it can take all of the DoF, or variables of interest, in a given system and identify new collective variables (the principal components), based on relations within highdimensional datasets. It also indexes the load magnitude of the original variables of interest on the identified principal components, and this can help uncover how the behavioral variables are coupled together in the organized system. These characteristics make PCA an important tool for revealing global properties of a system.

However, PCA is a linear analysis and cannot measure the local coordination between agents. That question requires an analysis tool that quantifies patterns of coordination between two behavioral variables. Cross-recurrence quantification (CRQ) is better suited for this purpose. CRQ is a nonlinear analysis that indexes repeating patterns in a pair of time series (Shockley, Butwill, Zbilut, \& Webber, 2002; Webber \& Zbilut, 1994), and has already demonstrated its utility in interpersonal coordination (e.g., Ramenzoni, Riley, Shockley \& Baker, 2012; Richardson, Dale \& Shockley, 2008). In particular, when analyzing side-by-side walking, Page and Warren (2012) found CRQ to output a reliable measure of the coupling strength, or degree of coordination, between the walking speed (s) of two pedestrians as their behavior evolved over time. In contrast to PCA, CRQ is limited to a pairwise analysis of time series, and thus provides a measure of coupling strength in a dyad rather than the overall coordination of the group. Taken together, PCA and CRQ allow us to characterize coordination and coherence at a local (i.e., dyad) and more global (i.e., group) level of behavior.

To study group coherence, we began with observations of a simple and highly controlled locomotor task: four pedestrians walking to a common goal. While quantitative measures of crowd dynamics should apply to more complex scenarios (see Moussaïd et al., 2012), we believed this approach would reveal essential coordination dynamics as a first pass to understanding crowd behavior. In the present
experiment, we instructed groups of four participants to walk toward one of three goals; the group's initial density was varied on each trial (see Figure 1). As described above, we analyzed time series of two behavioral variables for each participant: the heading direction ( $\phi$ ) and speed (s). This resulted in a total of eight DoF for the four-agent system. We hypothesized that the groups would exhibit dimensional compression in all conditions, compared to virtual groups we constructed by randomly sampling the same participants from different trials (see Method section). We also expected a greater reduction in DoF as density increased, due to larger changes in visual angle and visual direction at smaller distances, as well as to spatial constraints on walking. With regard to CRQ, we hypothesized that the coupling strength would be greater in all conditions compared to virtual groups, and that the leader-follower pairs would exhibit stronger coupling than the side-by-side pairs, as observed in our previous studies of two pedestrians (Rio \& Warren, 2011; Page \& Warren, 2012).

## Participants

Five groups of four participants ( $\mathrm{N}=20$; M age $23.57 \pm 0.93$ years; 12 females, 8 male) from Brown University and the greater region were compensated $\$ 15$ for their participation. Participants had no history of cognitive deficits, lower extremity injury, or neuromuscular disorders that would inhibit normal locomotor activity. The experiment was approved by the Brown University Institutional Review Board and adhered to guidelines for the ethical treatment of participants.

## Materials and Apparatus

The experiment took place in a $12 \times 14 \mathrm{~m}$ open room. The head position of each participant was tracked with a MicroTrax inertial tracker affixed atop a lightweight bicycle helmet on the head. Each tracker communicated with an IS900 ultrasonic overhead grid tracking system (InterSense, Billerica MA, USA) and provided 6 DoF position ( 4 mm RMS error) and orientation ( $0.1^{\circ}$ RMS error) data at 60 Hz . Three cardboard goal poles (approximately 2 m tall and 0.5 m in diameter) were placed at an initial distance of 8 m and angular offsets of $12.53^{\circ}$ to the left (pole 1 ), $0^{\circ}$ (pole 2), and $12.53^{\circ}$ to the right (pole 3) of the midpoint of the front two participants (see Figure 1). Colored tape was used to mark four possible starting positions in a square configuration, with initial spacing of $0.5,1.0,1.5$, or 2.5 m on a side.

## Design \& Procedure

Each group completed eight trials in each of 12 conditions, four densities ( $0.5,1.0,1.5,2.5 \mathrm{~m}$ spacing) crossed with three goal positions (left, straight, right; see Figure 1). This resulted in a total of 96 trials, presented in a random order, in each experimental session. Goal position was changed only to vary the task between trials, and thus was not included as a factor in the statistical analyses.

At the beginning of each trial the four participants were randomly assigned to the four positions in the square
configuration: (1) front right, (2) front left, (3) back right, or (4) back left (Figure 1). Once they were standing in the correct location, an experimenter gave a "go" signal and the group began to walk straight ahead. As the last participant crossed a line 1 m in front of the initial positions of the front participants, the experimenter gave a verbal command to walk to goal 1,2 , or 3 . The only instruction given to the participants was to continue walking to the specified goal at a comfortable pace without stopping. Participants were not told to stay together as a group or to maintain the initial configuration. Each trial lasted approximately 6 to 8 s .


Figure 1: The four possible starting positions for each of the four possible starting densities. From this view, the participants would walk from left to right. Note the dotted line 1 m from the midpoint between the front two participants that represents when the experimenter "goal" command was given. The heading and speed variables ( $\phi$ and $\mathbf{s}$, respectively) under each agent indicate the eight DoF of the system (i.e., the eight variables analyzed in the present experiment). $\mathrm{FR}=$ Front Right; FL = Front Left; BR = Back Right; BL = Back Left.

## Data Reduction and Analysis

The tracking system recorded the head position ( $x$ - and $z$ coordinates) of each participant at a sampling rate of 60 Hz . The raw (unfiltered) position data were used to compute the participant's speed (s) and heading ( $\phi$ ) from the displacement between successive samples, according to the following equations:

$$
\begin{align*}
& \boldsymbol{s}_{i}=\frac{\left(\left(x_{i}-x_{i-1}\right)^{2}+\left(z_{i}-z_{i-1}\right)^{2}\right)^{.5}}{\Delta t}  \tag{1}\\
& \phi_{i}=\tan ^{-1}\left(\frac{x_{i}-x_{i-1}}{z_{i}-z_{i-1}}\right) \tag{2}
\end{align*}
$$

where $x_{i}$ and $z_{i}$ are the head position on the $i$ th frame, in room coordinates. The $\phi$ and $\mathbf{s}$ time series were used for all subsequent analyses.

Virtual Group Construction For each real group trial, a paired virtual group trial was constructed by randomly selecting a time series from the same participants in the same group and condition, but from different trials. Thus all task constraints were matched, except that the four
participants in the virtual group were not perceptually coupled with each other. The virtual groups were created to ensure that any results that indicated significant coordination between participants were due to the perceptual coupling, not the task constraints (e.g., the common goal, the simultaneous goal command, or similar preferred walking speeds). After random selection of the four time series, they were temporally aligned based on the time the goal command was given by the experimenter. To equate their length (a requirement of both PCA and CRQ analysis), a time series was then potentially cropped at the beginning and/or end. This resulted in four time series of equal length that were aligned by the goal command.

Principal Components Analysis (PCA) PCA identifies linear relationships within multi-dimensional datasets and then maps the original data into a newly defined space. The principal components (i.e., axes of space) represent the dataset's primary dimensions of variation, but do not necessarily map directly onto the original dimensions of the actual measurement. The end result is a representation of potentially new, important variables that best account for the variance within the observed system.

In the context of the present experiment, eight variables of interest representative of the 8 DoFs of the observed system (i.e., $\phi$ and $\mathbf{s}$ for each participant) were submitted to a single PCA. The data were normalized using a $z$-score transform prior to analysis. PCA was performed in Matlab using the princomp function and the results were examined in a similar fashion to Ramenzoni et al. (2012).

First, the number of components that together account for $90 \%$ or more of the variance in the data set was determined. To investigate dimensional compression in the real vs. virtual group, a $4 \times 2$ mixed-model analysis of variance (ANOVA) was conducted on number of components, with initial density as a within-subjects factor and group (real vs. virtual) as a between-subjects factor, averaged across goal position.

Next, the amount of variance accounted for by the first principal component (PC) in the real vs. virtual group was compared using an identical mixed-model ANOVA. The analysis was limited to the first PC because the subsequent components were dependent on the first PC. Greater variance accounted for by the first PC in the real group indicates dimensional compression, and thus greater coherence, in the visually coupled system.

Finally, the mean correlation coefficient ( $r$ ) for the loading of each behavioral variable on the first PC was examined to investigate which of the eight variables were most influential in characterizing the group's behavior. The $r$ values were transformed using a Fisher's z' transform and submitted to a $4 \times 8 \times 2$ mixed-model ANOVA with initial density and agent position as within-subjects factors, and group as a between-subjects factor, again averaged across goal position. The aim of this analysis was to examine whether the speed or heading of an agent in a particular
position more strongly influenced the group's behavior and whether this influence depended on density.

Cross-recurrence Quantification (CRQ) A nonlinear, two-dimensional cross-recurrence quantification (CRQ) analysis was used to quantify the time-correlated activity between the heading time series of each dyad in the group, and, separately, the speed time series of each dyad (see Figure 2 for the analysis steps). A CRQ analysis is conducted by first embedding the pair of normalized time series in a multidimensional, time-delayed phase space (see Marwan, Romano, Thiel \& Kurths, 2007; Ramenzoni et al., 2012; Richardson, Schmidt, \& Kay, 2007; Shockley et al., 2002; Webber \& Zbilut, 1994). Because not all variables that make up the behavior in a dynamical system are necessarily knowable a priori, phase space reconstruction allows for the behavior of these potentially "hidden" variables in the dynamical system to be evaluated via their interaction with, or influence on, the known variable (in this case the $\mathbf{s}$ time series). Thus, the structure of the reconstructed phase space can reveal the underlying dynamics of the dynamical system as a whole. Specifically, the "neighborliness" of points within some tolerance or radius in phase space can indicate recurrent points in the two time series. These represent states in one time series that closely correspond to previous or future states in the other time series, and can illustrate behavioral patterns of coordination in the observed system. The recurrent points are identified and represented in a cross-recurrence plot (see Figure 2, bottom), from which a suite of measures can then be computed to quantify these patterns (see Shockley et al., 2002 and Marwan et al., 2007 for a review of analysis procedures).

In the present experiment, only cross-maxline (CML) was computed and analyzed: specifically the longest diagonal line of consecutive recurrent points on a cross-recurrence plot. This provides a measure of the longest time interval that the speed (or heading) of two participants was coupled during a given trial. CML is known to be sensitive to the temporal stability of coordination between two time series, associated with the coupling strength between agents (Richardson et al., 2007). A previous CRQ analysis of speed with two pedestrians revealed stronger coupling between leader-follower pairs than side-by-side pairs (Page \& Warren, 2012). The parameters used for CRQ were as follows: embedding dimension $=5$; delay $=3$ data points; radius within which points are counted as recurrent $=1.0 \%$ of the actual distance separating points in reconstructed phase space.

## Results

## PCA

Number of Components The number of components required to account for $90 \%$ of the variance was significantly reduced in real groups $(M=3.71 \pm 0.12)$ compared to virtual groups $(M=5.76 \pm 0.07), F(1,8)=$
233.22, $p<.001$ (see Figure 3, top). Thus, the visual coupling between agents reduced the DoF of the group significantly more than the external task constraints, indicative of emergent global coherence. Surprisingly, there was no effect of initial density ( $p>.05$ ), implying that group coherence at low densities was comparable to that at high densities.


Figure 2: A schematic of the steps in the CRQ analysis. For each trial, the speed time series of the FR agent (top left) and BR agent (top right) are unfolded separately into a shared reconstructed phase space via time-delayed copies of each measured time series, denoted as $\mathbf{s}_{\mathrm{FR}, \mathrm{BR}}(\mathrm{A})$. Recurrent points within a given radius (B) and strings of recurrent points (C) are identified with respect to each point in phase space and represented in a cross-recurrence plot (bottom) with each axis representative of the $\mathbf{s}_{\mathrm{FR}}$ and $\mathbf{s}_{\mathrm{BR}}$ time series at each time step. Each pixel indicates a recurrent point, and the diagonal line structures indicate the length of a string of recurrent points, or the co-evolution of the two time series at different time delays. The longest diagonal line, cross-maxline (CML), was computed for each dyad in the group.

PC 1 The first principal component accounted for significantly more variance in real groups ( $M=52.43 \% \pm$ 0.79 ) than in virtual groups ( $M=39.74 \% \pm 0.45$ ), $F(1,8)=$ 190.42, $p<.001$. This result confirms dimensional compression in group behavior, due to the visual coupling. There was, again, no effect of initial density ( $p>.05$ ) on the variance accounted for by PC 1.

Contribution of Variables to PC 1 The composition of the first principal component was further examined to determine the relative contributions of each behavioral variable, by computing the loading ( $r$ ) of each variable on PC1 (see Figure 3, bottom). A significant agent position $\times$ group
interaction was observed for $r, F(7,56)=408.03, p<.001$. Follow-up $t$-tests (Bonferroni corrected $p \leq .01$ ) indicated that the $\mathbf{s}$ variable was more strongly correlated with PC 1 in the real groups than in the virtual groups (all $p<.001$ ), whereas the $\phi$ variable was not (all $p>.01$ ), for all four agent positions. Within the real groups, the $\mathbf{s}$ variable was more strongly correlated with PC1 than the $\phi$ variable ( $p<$ .001 ), whereas $\mathbf{s}$ and $\phi$ did not significantly differ in the virtual groups (all $p>.01$ ), for all agent positions. Thus, the behavior of real groups was more coherent than that of virtual groups, primarily due to the coordination of walking speed; thanks to the presence of a common goal, heading was independently aligned in both groups.


Figure 3: The amount of variance accounted for by each component beginning with PC1 (top), and the loading ( $r$ ) of behavioral variables onto PC 1 (bottom). Black bars = real groups, white bars $=$ virtual groups, significant differences $(p$ $<.001)$ indicated with Duncan Grouping.

## CRQ

The results of PCA indicated the importance of speed (more than heading) as a variable of interest in the current dataset. Accordingly, the CRQ analyses focused on the speed time series for all six dyads on each trial. Representative crossrecurrence plots for a real and virtual dyad appear in Figure 4. A significant main effect of group was observed for cross-maxline length (CML), $F(1,8)=34.83, p<.001$. Specifically, the real group exhibited a mean CML ( $M=$ $49.93 \pm 0.03$ ) more than twice as long as the virtual group ( $M=20.73 \pm 0.02$ ), irrespective of dyad, goal position, or initial density. Surprisingly, this implies that the coupling is equally stable at high and low densities, and for leaderfollower and side-by-side dyads.


Figure 4: Sample cross-recurrence plots for speed time series from a real (top) and a virtual (bottom) leader-follower dyad. Note the diagonal lines visible in the cross-recurrence plot for the real dyad, indicative of a temporally stable speed coupling between agents.

## Discussion

The present experiment attempted to measure the degree of coherence in pedestrian groups, based on analyses of two behavioral variables, heading $(\phi)$ and speed (s), during goaldirected locomotion. We expanded the analysis from interpersonal coordination to the behavior of small groups, as a path to understanding collective crowd dynamics.

The PCA found that visually coupled pedestrian groups exhibited significant dimensional compression across all experimental conditions, compared to virtual groups. The results indicate that the task constraints (e.g. common goal, simultaneous command, preferred walking speed) accounted for a reduction of approximately 2.2 DoF (from 8 to 5.8 ) in the virtual groups. However, the visual coupling produced a further reduction of approximately 2.1 DoF (from 5.8 to 3.7). This is indicative of a functional reorganization of DoF thanks to the informational coupling of behavioral variables, yielding the emergence of coherent collective behavior.

In addition, PC 1 analysis offers preliminary evidence of a new collective variable that accounts for group coherence in the present case. The loading of behavioral variables on PC1 suggests that agent speed is a primary contributor to the new group dynamics. However, the comparatively weak contribution of the behavioral variable of heading direction is likely due to the external constraint of a common goal in this particular task. Taken together, these findings support the reduction of DoF in interpersonal coordination proposed by Riley et al. (2011; Ramenzoni et al., 2012).

The CRQ analysis provided more specific results about the coupling strength between particular dyads in the group.

The speed variable exhibited a significantly more stable coupling in real groups than virtual groups, with no differences between leader-follower, side-by-side, and diagonal dyads. Taken together, the PCA and CRQ results indicate that the reduction in group DoF in the present task is due in large part to the coordination of speed at the dyad level, resulting from the visual coupling between neighbors.

While the overall results supported the hypothesis of group coordination via local coupling, the analyses diverged from our expectations in two important respects. First, we anticipated that the degree of coordination would increase as group density increased, but we did not observe an effect of density. It is possible that the range of densities tested was too small to observe an effect, or that the task constraints combined with a short walking distance limited the degree of variability in individual trajectories. But consistent with this finding, we previously observed that speed coordination in pairs of pedestrians is also independent of distance over 1-3m (Rio \& Warren, 2011). Second, we were surprised that coupling strength did not differ among dyads, given we had previously observed greater speed coordination between leader-follower than side-by-side pairs. Again, it is possible the task constraints may have limited the variability in individual behavior. In subsequent experiments, we are measuring pedestrian groups over longer distances without a common goal or timing signal.

The present work is a starting point for understanding collective behavior in pedestrian groups. We have begun by focusing on the local coupling between agents, on the hypothesis that this generic coordination mechanism will scale up from small groups to large crowds and perhaps to swarms across species. It is likely that other cognitive and social variables also constrain this coupling. For example, cognitive processes such as decision-making and motivation, and social factors such as group membership, dominance relations, and social communication, may influence the selection of goals, neighbors, speeds, and control laws and shape group dynamics. The present experiment suggests an approach to quantifying multi-agent coordination in many of these contexts. Future work will continue to scale up these analyses to larger groups in various pedestrian scenarios, with the aim of understanding the emergence of collective behavior and global patterns in large crowds.

## Acknowledgments

This research is funded by NIH 5R01 EY010923-25. The authors would also like to thank Henry Harrison.

## References

Bonneaud, S. \& Warren, W. H. (2012) A behavioral dynamics approach to modeling realistic pedestrian behavior. $6^{\text {th }}$ International Conference on Pedestrian and Evacuation Dynamics, Zurich, Switzerland.
Fajen, B. R. \& Warren, W. H. (2003). Behavioral dynamics of steering, obstacle avoidance, and route selection.

Journal of Experimental Psychology: Human Perception and Performance, 29, 343-362.
Fajen, B. R. \& Warren, W. H. (2007). Behavioral dynamics of intercepting a moving target. Experimental Brain Research, 180, 303-319
Marwan, N., Romano, M. C., Thiel, M., \& Kurths, J. (2007). Recurrence plots for the analysis of complex systems. Physics Reports, 438, 237-329.
Moussaïd, M., Guillot, E. G., Moreau, M., Fehrenback, J., Chabiron, O., et al. (2012). Traffic instabilities in selforganized pedestrian crowds. PLOS Computational Biology, 3, e1002442.
DOI: 10.1371/journal.pcbi. 1002442
Moussaïd, M., Helbing, D. \& Theraulaz, G. (2011). How simple rules determine pedestrian behavior and crowd disasters. Proceedings of the National Academy of Sciences of the United States of America, 17, 6884-6888.
Ondřej, J., Pettré, J., Olivier, A-H., \& Donikian, S. (2010. A synthetic-vision based steering approach for crowd simulation. ACM Transaction on Graphics 4, 123. DOI: 10.1145/1778765.1778860.
Page, Z. \& Warren, W. (2012). Visual control of speed in side-by-side walking. Journal of Vision, 12, 188.
Ramenzoni, V. C., Riley, M. A., Shockley, K. \& Baker, A. A. (2012). Interpersonal and intrapersonal coordinative modes for joint and single task performance. Human Movement Science, 31, 1253-1267.
Richardson, M. R., Schmidt, R. C., \& Kay, B. A. (2007). Distinguishing the noise and attractor strength of coordinated limb movements using recurrence analysis. Biological Cybernetics, 96, 59-78.
Riley, M. A., Richardson, M. J., Shockley, K. \& Ramenzoni, V. (2011). Interpersonal synergies. Frontiers in Psychology, 2, 38.
Rio, K., Bonneaud, S. \& Warren, W. H. (2012). Speed coordination in pedestrian groups: Linking individual locomotion with crowd behavior. Journal of Vision, 12, 190.

Rio, K. W. \& Warren, W. H. (2011). A speed control law for pedestrian following based on visual angle. Journal of Vision, 11, 899.
Richardson, D., Dale, R., \& Shockley, K. (2008). Synchrony and swing in conversation: Coordination, temporal dynamics, and communication. In G. Knoblich (Ed.), Embodied Communication (pp. 75-94). New York City, NY: Oxford University Press.
Shockley, K., Butwill, M., Zbilut, J. P., \& Webber, C. L. Jr. (2002). Cross recurrence quantification of coupled oscillators. Physics Letters A, 305, 59-69.
Warren, W. H. \& Fajen, B. R. (2008) Behavioral dynamics of visually-guided locomotion. In A. Fuchs \& V. Jirsa (Eds.), Coordination: Neural, behavioral, and social dynamics. Heidelberg: Springer.
Webber, C. L. Jr., \& Zbilut, J. P. (1994). Dynamical assessment of physiological systems and states using recurrence plot strategies. Journal of Applied Physiology, 76, 965-973.

# Naturalistic Word-Concept Pair Learning With Semantic Spaces 

Brent Kievit-Kylar ${ }^{1}$, George Kachergis ${ }^{\mathbf{3}}$, and Michael N. Jones ${ }^{1,2}$<br>\{bkievitk, gkacherg, jonesmn\}@indiana.edu<br>Cognitive Science Program<br>${ }^{2}$ Department of Psychological \& Brain Sciences<br>Bloomington, IN 47405 USA<br>${ }^{3}$ Psychology Department, Leiden University<br>Leiden 2311 EZ Netherlands


#### Abstract

We describe a model designed to learn word-concept pairings using a combination of semantic space models. We compare various semantic space models to each other as well as to extant word-learning models in the literature and find that not only do semantic space models require fewer underlying assumptions, they perform at least on par with existing associative models. We also demonstrate that semantic space models correctly predict different word-concept pairings from existing models and can be combined with existing models to perform better than either model can individually.


Keywords: Childes; natural language processing; semantic space models; associative learning.

## Introduction

While the task of word-concept matching may seem trivial to an adult, imagine the task from the perspective of a young child. A child hears a series of vocalizations, which are then parsed into word units, and must perceive instances of objects in their environment through visual inspection. From that information the child must determine a set of objects or concepts that are present. These two tasks are challenging enough, but the child must then find a way to correlate the words that he/she has heard with the objects in the immediate environment. Extracting the correct mappings from the myriad possible ones is complicated by things such as the potential absence of matches between object and word (e.g., an object is mentioned that is not present), and the fact that not all words refer to objects (e.g., verbs and function words).

Several techniques have been proposed to help simplify the word-concept acquisition problem, the majority of which require the child to have prebuilt assumptions (e.g., a novel word must map to a novel object), or to perform complex Bayesian logic calculations (Frank et al. 2007). In this paper, we explore a new class of model that is based on the rapidly growing field of semantic space models. In particular, we generalize Kintsch's (2001) predication algorithm to the problem of word-concept learning in semantic spaces.

Kintsch's (2001) algorithm simulates the process of matching based on shared neighbors in a semantic space. The result of our adaptation is a model that learns to map words and objects to semantic clusters, greatly simplifying the problem of word-object mapping. Rather than casting the problem as one of learning associations between independent words and independent objects, a semantic
space approach can take advantage of the fact that similar words carry mutually reinforcing information about each other's object referents. In addition, the similarity between a noun and semantically related verb or adjective contains information about the noun's referent. Bounce may often be used when a ball is present in the environment, even in absence of the noun ball. The semantic similarity between the words bounce and ball may be used as an indirect cue to the mapping between the noun and object.

We next provide background on the problem, data, and existing models of word-concept learning. Then we turn to a summary of a variety of semantic space models used, and a general purpose technique for creating word-concept learners from semantic models adapted from Kintsch's (2001) algorithm. Finally, we test these models on a labeled fragment of the CHILDES corpus and explore the benefits of combining different semantic models into hybrids.

## Child Learning Models

While there are a number of existing word-concept mapping models from the child learning literature, we will focus on two recent models that have both been applied to the objecttagged corpus data that we use (described below).
The data used for training in these simulations are from an annotated version of the Rollins section of the CHILDES corpus (MacWhinney, 2000) used by Frank et al (2007). The entire corpus takes place over approximately ten minutes of talk taken from a caregiver to a child. Each sentence is annotated with the objects that are visible to the child when that sentence was being spoken. Thus the corpus consists of entries in the form $\left\{\mathrm{W}_{0}, \mathrm{~W}_{1}, \ldots, \mathrm{~W}_{\mathrm{n}}, \mathrm{C}_{0}, \mathrm{C}_{1}, \ldots\right.$ $\left.\mathrm{C}_{\mathrm{m}}\right\}$ where $\mathrm{W}_{\mathrm{i}}$ is a word token and $\mathrm{C}_{\mathrm{i}}$ is a concept token (each represented by a string). A unique identifier was used to differentiate concepts from words, as both are represented in the dataset by similar strings of characters. In this paper, we will use the angle brackets as delimiters, such that "dog" represents the word dog, and " $<\operatorname{dog}>$ " represents the concept or object dog.

## Frank et al.'s (2007) Bayesian Framework

Frank et al. (2007) propose a Bayesian model to jointly learn word-concept mappings, as well as which objects a speaker intends to speak about in a situation. Using a model similar to Latent Dirichlet Allocation (LDA) used by Topic models, they assume that words are generated from the lexicon according to what objects are present and are likely
to be talked about (i.e., the intention of the speaker). This model is a computational-level model that does not specify learning mechanisms, but rather specifies how to calculate the likelihood of a particular lexicon, given all of the situations that one has observed. The inferred lexicon is simply a collection of word-object pairings, and tends to be small because the prior favors smaller lexicons. The model handles nonreferential words: if a given word appears with many objects only a few times, these mappings will likely not be added to the lexicon. The inferred lexicon will mostly be comprised of the highest co-occurring word-object pairs; there is no explicit penalization for linking words to multiple objects, nor a word to multiple objects. There is no learning of associations among words, nor among objects. Frank et al. demonstrate impressive performance from this model on subsequent testing of word-object pairings.

The semantic space approach we propose differs theoretically from the Frank et al. (2007) model in at least two ways. Firstly, while the Frank et al. approach attempts to calculate an underlying generator that maps from concepts to words through the lexicon, the semantic space approach is more passive, projecting words and concepts onto points in psychological space. Secondly, the semantic space approach attempts to learn the relations between words, including the relations between concepts. This added structure allows a semantic space model to bootstrap additional partial information from indirect relationships.

## Kachergis, Yu, \& Shiffrin (2012) Associative Model

Kachergis et al. (2012) introduced an incremental model that learns word-object associations. Competing attentional biases for familiarity (i.e., already-strong associations) and for stimuli with uncertain associates (i.e., high entropy) allow this model to exhibit mutual exclusivity and other word-learning principles, as well as associative learning effects such as blocking and highlighting (Kachergis, 2012).

The model stores knowledge in $M$, a word-object association matrix that grows during training. Cell $M_{w, o}$ is the strength of association between word $w$ and object $o$. Before the first trial, $M$ has no information: each cell is set to $1 / \mathrm{m}$. Association strengths decay, and on each new trial a fixed amount of associative weight, $\chi$, is distributed among the associations between words and objects, and added to the strengths. The rule for distributing $\chi$ (i.e., attention) balances a preference for attending to unknown stimuli with a preference for strengthening already-strong associations. When a word and referent are repeated, extra attention (i.e., $\chi$ ) is given to this pair-a prior knowledge bias. Pairs of stimuli with no or weak associates also attract attention, whereas pairings between uncertain objects and known words, or vice versa, do not attract much attention. Stimulus uncertainty is captured using entropy $(H)$, a measure that is 0 when the outcome of a variable is certain (e.g., a word appears with only one object), and maximal $\left(\log _{2} n\right)$ when all of the $n$ possible object (or word) associations are equally likely (e.g., for a novel stimulus, or one that appears with all stimuli equally). In the model, on each trial the entropy of
each word and object is calculated from the normalized row (column) vector of associations for that word (object), $p\left(M_{w,}\right)$, as follows:

$$
H\left(M_{w, \cdot}\right)=-\sum_{i=1}^{n} p\left(M_{w, i}\right) \cdot \log \left(p\left(M_{w, i}\right)\right)
$$

The update rule for adjusting and allocating strengths for the stimuli presented on a trial is:

$$
M_{w, o}=\alpha M_{w, o}+\frac{X \cdot e^{\lambda \cdot(H(w)+H(o))} \cdot M_{w, o}}{\sum_{w \in S} \sum_{o \in S} e^{\lambda \cdot(H(w)+H(o))} \cdot M_{w, o}}
$$

In this equation, $\alpha$ is a parameter governing forgetting, $\chi$ is the attention weight being distributed, and $\lambda$ is a scaling parameter governing differential weighting of uncertainty and prior knowledge (familiarity). As $\lambda$ increases, the weight of uncertainty (i.e., the exponentiated entropy term, which includes both the word and object's association entropies) increases relative to familiarity. The denominator normalizes the numerator so that exactly $\chi$ associative weight is distributed among the potential associations on the trial. For stimuli not on a trial, only forgetting operates. This model aims to capture the process of learning simple wordconcept associations using basic cues a learner may have.

## Semantic Space Models

Semantic space models have seen a great amount of both attention and success in the literature over the past decade. There are a variety of semantic space models currently in the literature, but all are fundamentally based on the assumption that the contexts in which a word occurs may be used to infer its meaning, commonly projected into a highdimensional psychological space. Words that frequently cooccur in contexts together, or that frequently occur in similar contexts, become more proximal in semantic space. We explore a variety of semantic space model representations of the CHILDES data here, all using the same mapping mechanism adapted from Kintsch's (2001) algorithm. We next very briefly describe each representation model used in our comparison.

## BEAGLE

The BEAGLE model (Jones \& Mewhort, 2007) uses holographic vector manipulation to represent word similarities. In BEAGLE, each new word encountered is assigned an environmental vector with elements generated independently from a Gaussian distribution, and a lexical vector of the same length but initialized to zeros. When encountering a sentence, the environmental vector of each word is added to the lexical vector of each word it co-occurs with. Similarity is measured using cosine similarities between words' lexical vectors.

## ESA

Explicit semantic analysis (Gabrilovich \& Markovitch, 2007) was designed for use with the Wikipedia corpus. It uses a centroid-based classifier that correlates given input
text to a weighted list of concepts associated with each target word.

## FDTRI

Fixed Duration Temporal Random Indexing, introduced by Jurgens and Stevens (2009), attempts to bypass the computational difficulty inherent in singular value decomposition through the use of random projections onto lower dimensional space. Similar to BEAGLE, each word has an environmental vector, although FDTRI vectors are generated to be sparse. Rather than producing a word-bymeaning matrix, FDTRI incorporates time in a word-by-meaning-by-time tensor. The additional temporal information could be useful in word-concept pairing, if the sequential information given by the caregiver is relevant to object detection. For example, a caregiver may be more likely to start with the label and then continue with a description of the object.

## HAL

The Hyperspace Analogue to Language (HAL; Lund \& Burgess, 1996) uses a fixed size window that is slid along the corpus. A matrix is built which is an accumulation of pairs of words that co-occur within any given window of text. Order information is partially preserved through the use of "occurring before," and "occurring after" cooccurrence matrices.

## LSA

Latent Semantic Analysis (LSA; Landauer \& Dumais, 1997) operates by applying singular value decomposition to a word-by-context frequency matrix, reducing the matrix from high dimensionality (documents) to lower dimensional space (latent semantic components). The premise is that the reduction removes irrelevant features of the word usage, yielding a semantic abstraction in the resulting space.

## ISA

Baroni, Lenci and Onnis (2007) developed incremental semantic analysis (ISA) to analyze children's speech data. ISA is based on random indexing models with a few variations. First, updating occurrence information includes both the signature (or environmental vector) as well as information on the learned history of the other word. This allows ISA to capture higher order relations. Second, word frequency discounts are updated online as the model learns for information about the distribution of words in the world.

## PMI

Pointwise Mutual Information (PMI; Church \& Hanks, 1989) is a basic information theoretic metric that looks at the probability of two words occurring together relative to the probability of each word occurring individually. This provides a first order word co-occurrence metric, and has demonstrated remarkable effectiveness at explaining human
semantic data without resorting to complex inference mechanisms.

## Word-Concept Models

We developed a generalized technique to transform any semantic space model into a word-concept learning model. Word-concept models are divided into a learning phase and a prediction phase. In the learning phase, the model is applied to sentences composed of words. The generalized modification for a word-concept model is to simply concatenate the word tokens with the concept tokens into a single concept/label sensory episode. In the prediction phase, we attempt to assign an object token to each word token.
All semantic space models have the ability to determine similarity between any word pair, thus prediction can be as simple as finding the object with the maximum similarity to the given target word:

$$
\begin{equation*}
\operatorname{maxarg}_{i}\left(\operatorname{sim}\left(w_{\text {targ }}, o_{i}\right)\right) \tag{1}
\end{equation*}
$$

While this performs reasonably, it is possible to improve on this technique by adding a second step. Building from Kintsch's (2001) predication algorithm, we first activate the neighbor set of N most similar words to our target word:

$$
\begin{equation*}
\mathrm{NSet}_{\text {targ }}=\operatorname{maxarg}_{i}\left(\operatorname{sim}\left(w_{\text {targ }}, w_{i}\right)\right) \tag{2}
\end{equation*}
$$

Then for every object, we calculate its activation, $A c t_{i}$, as the similarity between that object and every one of the top N word matches, weighted by the similarity of that word to the target word:

$$
\begin{equation*}
A c t_{i}=\square_{j \varepsilon N S e t} \operatorname{sim}\left(w_{t a r g}, w_{j}\right)^{p} * \operatorname{sim}\left(o_{j}, w_{j}\right) \tag{3}
\end{equation*}
$$

The mechanism provides a match not only to the target word but also to the target's region of semantic space. This is particularly important because there are always more words than concepts. Using Kintsch's (2001) predication allows non-nouns to influence the outcome of the similarity measurement through their similarity to the nouns. Thus, if the word "red" is strongly associated with the word "apple" in the discussion and "red" is also associated with the concept <apple>, then "red" can be used to discover the underlying link between "apple" and <apple>. This mapping can be done implicitly, without knowledge of the part of speech as long as the target words that are to be matched to objects are known.

## Experiment

Each of the above models was trained on the CHILDES corpus and the results were compared to the gold standard model in Frank et al. (2007), as well as to a baseline model that simply counts which words and objects co-occur. There are many different ways to evaluate model performance, and
there does not seem to be agreement in the field about the correct measure to use. To remain comparable to Frank et al (2007), we examined the best F-score (the harmonic mean of precision and recall) achieved by each model. We also looked at the overall proportion of pairings matching the golden standard if all 37 words are assigned a concept meaning. We do this in order to determine what mappings the system makes if forced, although we note that this random slice of parent-child interaction may not be enough to disambiguate all of the mappings.

We also explore hybrid models, asking which two models contribute the most complementary (non-redundant) information. This exercise may hint at what combinations of mechanisms are most important for learning in the natural language environment.

## Results

One popular way of visualizing the results of word-concept learning is through a confusion matrix as shown in Figure 1. The confusion matrix shows the similarity between wordconcept pairings as gradients from black to white (with lighter being a higher association) filling each grid cells. According to the gold standard, each word is associated with exactly one object (except for "bird" which can refer to $<$ duck $>$ or $<$ bird $>$ ). The cells outlined in red indicate the correct word-object pairings according to the gold standard.

In Figure 1a and 1b, a winner-takes-all filter has been applied for each word. Thus, the object that has the highest association has been assigned the similarity of 1 , and all others have been assigned a similarity of 0 . The values returned by the semantic space models cannot be directly interpreted as probabilities for pairing selections. Hence, only relative similarity measures are used here. In Figure 1c, we display the gradient similarity ratings after having been scaled to a power of five (thus exaggerating the differences between predictions).


Figure 1: Confusion matrix results.


Figure 2: Word space. For each model by word, a black square indicates if model correctly identified that word.

It is important to understand which word-object pairs each individual model gets correct or incorrect. It is also important to see which word/object pairs are overall more or less likely to be found by the model. Figure 2 shows for each model, and each word, the probability of accurately identifying the target word. Black squares indicate word-bymodel pairs in which model correctly identifies the object associated with that word. There is also a calculated average correctness for each word (rightmost column) and for each model (bottom row). Some of the semantic space models are non-deterministic (BEAGLE, FDTRI and ISA). For these models, 100 runs were computed and a correct identification granted when more than half of the runs correctly identified that pairing. No partial credit was given in any form for coming close to correctly matching each word. Parameter values were selected to be those optimized (relative to expected overall matches) for the individual model.
The F-scores are important indicators to help understand how well each model has correctly inferred the word-object pairings. To calculate the F-scores for each model, we first computed a similarity matrix for every word-object pair. For non-deterministic models, the similarities were averaged across 100 similarity runs for the given model. Next, maximum values were determined within each word to select an object. Pairings that were correct were labeled true, and pairings labeled that were incorrect, were labeled false. These similarity measures were then ordered based on strength (both correct and incorrect measures). For each N, precision and recall figures were then calculated for each of the top N word pairings.

This results in the receiver operator curves shown in Figure 3, and Table 1 shows the maximum F score values taken from the ROC curve chart for the top 5 models.


Figure 3: ROC curve for all models.
Table 1. Top F Scores

| Model | Hybrid | BEAGLE | Frank | ESA | COOC |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Best F | .83 | .55 | .54 | .54 | .53 |

Using the Word 2 Word language visualization tool (Kievit-Kylar \& Jones, 2012) we can visualize all of the word/object pairings as a graph. In Figure 4, words are nodes and each similarity measure is an edge. In the visualization below, we see all concepts lined up across the top with each word referring to them shown below. The green connections indicate the strongest similarities observed by the system. Ideally, all lines would link to the word directly above.


Figure 4: Word network visualization of BEAGLE model solution.

Because different models seem to make different mistakes, we also explored how hybrid models might be able to exploit these differences. Each model $M$ is able to assign some similarity measure to and word-object pair $\mathrm{M}_{\text {sim }}\left(\mathrm{W}_{\mathrm{x}}, \mathrm{O}_{\mathrm{y}}\right)$. We considered hybrid models of the following form: $\mathrm{M}_{\mathrm{A}, \mathrm{Bsim}}\left(\mathrm{W}_{\mathrm{x}}, \mathrm{O}_{\mathrm{y}}\right)=\mathrm{M}_{\text {Asim }}\left(\mathrm{W}_{\mathrm{x}}, \mathrm{O}_{\mathrm{y}}\right) * \mathrm{c}+\mathrm{M}_{\mathrm{Bsim}}\left(\mathrm{W}_{\mathrm{x}}, \mathrm{O}_{\mathrm{y}}\right)$. Each pair of models was optimized, relative to the average number of matches with the golden standard, for the constant c .

Figure 5 shows the correct number of matches for each optimum pairing of models. A heat map of colors has been added to indicate highest (red) to lowest (purple) values. The optimum model is a co-occurrence by BEAGLE hybrid with the later having a weight three times greater than the former. This hybrid model results in 30 correct mappings on the gold standard.


Figure 6: Word network visualization of optimum hybrid.


Figure 7: Thresholded confusion matrix for best hybrid.

|  | beagleE | beagleF | esaE | esaF | fdtriE | fdtriF | hale | halF | isaE | isaF | IsaE | IsaF | Kachergis | cooc |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| beagleGold | 24 | 27 | 21 | 19 | 12 | 19 | 19 | 15 | 8 | 19 | 10 | 19 | 20 | 30 |
| beagleE |  | 21 | 18 | 19 | 12 | 11 | 14 | 20 | 8 | 10 | 10 | 21 | 17 | 24 |
| beagleF |  |  | 18 | 19 | 12 | 18 | 14 | 23 | 8 | 10 | 10 | 23 | 22 | 22 |
| esaE |  |  |  | 19 | 20 | 18 | 20 | 19 | 17 | 18 | 21 | 20 | 23 | 21 |
| esaF |  |  |  |  | 22 | 19 | 22 | 22 | 19 | 19 | 20 | 19 | 24 | 21 |
| fdtriE |  |  |  |  |  | 15 | 15 | 15 | 12 | 16 | 13 | 12 | 19 | 29 |
| fdtriF |  |  |  |  |  |  | 15 | 19 | 10 | 15 | 15 | 11 | 15 | 24 |
| hale |  |  |  |  |  |  |  | 16 | 13 | 15 | 10 | 14 | 16 | 28 |
| half |  |  |  |  |  |  |  |  | 14 | 15 | 11 | 14 | 15 | 28 |
| isaE |  |  |  |  |  |  |  |  |  | 12 | 11 | 8 | 15 | 18 |
| isaF |  |  |  |  |  |  |  |  |  |  | 11 | 11 | 11 | 27 |
| IsaE |  |  |  |  |  |  |  |  |  |  |  | 10 | 17 | 23 |
| IsaF |  |  |  |  |  |  |  |  |  |  |  |  | 15 | 22 |
| Kachergis |  |  |  |  |  |  |  |  |  |  |  |  |  | 20 |

Figure 5: Hybrid Pairings.

## Conclusions

The semantic space approach to word-concept learning is a fruitful endeavor with potential to better understand how humans make use of mechanisms and mutually reinforcing information sources across learning. The best pure semantic space models was able to predict the gold standard to a higher degree of accuracy than existing models while still conforming to known semantic and processing constraints. The adaptation of Kintsch's (2001) mechanism for predication allows semantic models to consider not only the semantic similarity between a word and object, but to also consider mutual information from the semantic neighborhoods. This procedure provided a benefit to each of the semantic space models tested.

Hybrid models also provide interesting insight into the word/concept-learning problem. The optimum hybrid model merged the co-occurrence model with BEAGLE. This optimal fusion makes intuitive sense, as the co-occurrence model provides first-order co-occurrence information that can be best supplemented by the higher-order co-occurrence information inherent in the semantic space models. The performance of the hybrid model suggests that infants may be capitalizing on both raw co-occurrence information and an emerging ability for higher-order semantic abstraction. Knowledge of which words are similar to each other from linguistic experience may be used to bootstrap word-object mappings across learning.

## Acknowledgments

This research was supported by grants from Google Research and NSF BCS-1056744 to MNJ.

## References

Baroni, M., Lenci, A., and Onnis, L. (2007). Isa meets lara: A fully incremental word space model for cognitively plausible simulations of semantic learning. In Proceedings of the 45th Meeting of the Association for Computational Linguistics. 49-56.

Church, K.W., Hanks, P. (1989). Word Association Norms, Mutual Information and Lexicography. In: Proceedings of the $27^{\text {th }}$ Annual Conference of the Association of Computational Linguistics, 76-83.
Frank, M. C., Goodman, N. D., \& Tenenbaum, J. B. (2007). A Bayesian framework for cross-situational wordlearning. Advances in neural information processing systems, 20, 1212-1222.
Gabrilovich, E. and S. Markovitch. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. Proceedings of IJCAI, 1606-1611.
Jones, M. N., \& Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. Psychological Review, 114, 1-37.
Jurgens D., Stevens K. (2009). Event detection in blogs using temporal random indexing. In Proceedings of the Workshop on Events in Emerging Text Types, 9-16.
Kachergis, G. (2012). Learning Nouns with DomainGeneral Associative Learning Mechanisms. Proceedings of the 34th Annual Conference of the Cognitive Science Society 533-538.
Kachergis, G., Yu, C., \& Shiffrin, R. M. (2012). An Associative Model of Adaptive Inference for Learning Word-Referent Mappings. Psychonomic Bulletin \& Review, 19(2), 317-324.
Kievit-Kylar, B., \& Jones, M. N. (2012). Visualizing multiple word similarity measures. Behavior Research Methods, 44, 656-674.
Kintsch, W. (2001). Predication. Cognitive Science, 25, 173-202.
Landauer, T.K., Dumais, S.T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104(2), 211-240.
Lund, K., \& Burgess, C. (1996). Producing highdimensional semantic spaces from lexical co-occurrence. Behavior Research Methods, Instruments, and Computers, 28, 203-208.
MacWhinney, B. (2000). The CHILDES Project: Tools for Analyzing Talk. Lawrence Erlbaum.

# Selection of Linker Type in Emphatic Reduplication: Speaker's Intuition meets Corpus Statistics 

Özkan Kılıç (ozkan.kilic@gmail.com)<br>Cem Bozşahin (bozsahin@metu.edu.tr)<br>Department of Cognitive Science, Graduate School of Informatics<br>Middle East Technical University (METU), Ankara, Turkey


#### Abstract

Turkish Emphatic Reduplication (TER) occurs in adjectives and adverbs to accentuate their meanings. The current experimental study to investigate the selection of the linker type in TER indicated that responses from the participants correlate with some lexical statistics. The result relies on the statistics from a corpus with approximately 2 million Turkish words, which we use in lieu of lexical statistics. The frequency order of linker choice reported by the participants was exactly the opposite of the order of frequency of words in which the linker and the first consonant co-occur consecutively in the corpus. Such a direct link to lexicon was unexpected. We suggest that TER, an apparently phonological operation, depends on lexical access for selecting the appropriate linker whose cooccurrence with the initial consonant of the reduplicated word is infrequent. Our results relate morphology and lexicon in more ways than the blocking phenomena, and suggest that TER may be morpholexical.


Keywords: Morphology; emphatic reduplication; lexicon; lexical frequency.

## Introduction

Turkish is generally considered to be a language with subject-object-verb word order, and it is morphologically agglutinating, with considerably involved morphology. The Turkish word structure depends heavily on suffixes, which are transparently stacked on the leftmost stem like "beads on a string" (see Kornfilt, 1997; Lewis, 2000; Göksel \& Kerslake, 2005 for a detailed review of Turkish). However, not all Turkish morphology is suffixal, and it is not always the case that we can enumerate the allomorphies clearly in non-suffixal morphology and phonology, with some good variation among speakers, which is not the case for suffixal morphology. Perhaps not surprisingly, these cases involve items which might appear to look like prefixes. We suggest a way to relate these differences in the processing of bound items to lexical statistics, as evidence for another case of interaction between morphology and lexicon.

Turkish Emphatic Reduplication, henceforth TER, is one such resource showing greater speaker variation. It is a derivational process which intensifies the meaning of adjectives and some adverbs. Phonologically it involves the duplication of the initial ( $C$ ) $V$ of the base, and addition of a prefix-like item as a linker to the root, which is a consonant from the set $\{p, s, m, r\}$. (Demircan, 1987; Oztaner 1996; Wedel, 1999; Yu, 1999; Kelepir, 2000; Kim, 2007; Dhillon, 2009). All words beginning with a vowel are infixed with $-p$ as the linker. In some cases the $(C) V+$ linker "prefix" is also followed by an additional "infix" from the set $\{-A,-I l,-A m\}$ as in some of the examples below.

| (1) ka-s-katı | ç1-r-1l-çıplak |
| :--- | ---: |
| RED solid | RED naked |
| 'hard as a rock' | 'totally naked' |
| pa-r-am-parça | dü-p-e-düz |
| RED torn | RED plain |
| 'completely torn apart' | 'utterly' |

## Related Work

Inkelas \& Zoll (2005) use cophonology to explain emphatic reduplication. A cophonology is a morphological function associated with particular morphological constructions to model morphologically conditioned phonology. Cophonologies receive words or morphemes as input, and perform some operations such as constraint ranking, truncation, and velar deletion on the input to be sent to the phonological interface (Inkelas \& Orgun, 1995; Inkelas, 2000). Truncation and addition act on the word beyaz 'white' to produce bem-. Then, the mother node links the subconstituent daughters to the input and shifts stress to the truncated one to form bémbeyaz 'snow white'. Demircan (1987), and later Wedel (1999, 2000), examined TER as a phonological operation and summarized the linker selection constraints as follows:
(2) (i) The linker from the set $\{p, s, m, r\}$ cannot be identical with the initial consonant $\left(C_{1}\right)$ of the base: pembe 'pink' $\rightarrow$ *peppembe, although $p \in\{p, s, m, r\}$. Perpembe is possible but not likely (see below).
(ii) The linker cannot be identical to the second consonant $\left(C_{2}\right)$ of the base: pembe $\rightarrow$ *pempembelpespembe, although $m \in\{p, s, m, r\}$.
(iii) The phonetic features \{coronal, sonorant, labial, continuant $\}$ of the linker cannot be identical with those of the second segment of the base. The linker with the most contrasting features is selected for perceptual salience.
(iv) The linker is selected in a way that it can establish an optimization or balance among the features contributing to the featural contrast with respect to base.

Some examples in Turkish seem to be orthogonal to these constraints. For example, in addition to çı-r-ll-çıplak, which is the commonly assumed reduplicated form of çıplak as in (1), çı-s-çıplak, çı-r-çıplak, çı-m-çıplak and çı-p-çıplak do occur in the Web.

The constraints in (2i-iv) can be violated when phonological productivity is put to the test in relation to morphology and the lexicon. We asked 50 participants to reduplicate pırasa 'leek', and along with the expected pımpırasa and pıspırasa, we also received pırpırasa and pıppırasa, in more or less equal distribution.

In this paper we report our experimental study which investigates the selection of the linker type, and its relation to morphology and the lexicon. The results indicate that responses from participants show disagreements with the literature about familiar words which are known to be non-TER targets. It also suggests that TER may be morpholexical, which is quite contrary to current literature.

The database for our statistics is a corpus of approximately 2 million Turkish words. We consider the database to be a rough approximate of a native speaker's lexical statistics. We asked the participants to emphatically reduplicate some words from this corpus. We asked them to play a "what-if" game to reduplicate the words that they would normally resist to reduplicate, such as masa 'table'. Among 50 participants and 31 target words, only one target, the only one that begins with a vowel, showed unanimous agreement, all others showing varying degrees of agreement. We present our method and the experiments toward understanding these results.

## TER vs. Reduplication

Kim (2007) points out that the most productive "prefixation" in Turkish is observed in reduplication. According to Göksel \& Kerslake (2005), Turkish duplication can be observed in three ways: M-reduplication, doubling, and TER, as explained above. We show below that, unlike TER, the first two kind are phonological-syntactic operations, therefore not operating on morphological properties but phonologicalsyntactic ones. In this regard they are expected to be less susceptible to lexical statistics.

## M-reduplication

If a word or compound to be m-reduplicated starts with a vowel, the original word is prefixed with $m$-, and then duplicated as shown in (3a). If it starts with a consonant other than $m$-, the consonant is replaced with $m$-, and the new form is duplicated as shown in (3b). In case the word or the compound starts with $m$-, it is followed by the word falan 'like, so and so'. M-reduplications can occur in all syntactic positions.
(3) a. $[\text { Çocuklar }]_{N P}\left[[\text { akıcı makıcı }]_{A D V}[\text { konuşmazlar }]_{V}\right]_{V P}$ child-PLU fluent M-DUP speak-NEG-AOR-3PL lit. 'Children do not speak fluently (and the like)'
b. $[\text { Çocuklar mocuklar }]_{N P}\left[[\text { akıcı }]_{A D V}[\text { konuşmazlar }]_{V}\right]_{V P}$ child-PLU M-DUP fluent speak-NEG-AOR-3PL lit. 'Children (and the like) do not speak fluently'

We note that the results of this process are two independent words, both phonologically and syntactically. We can, for example, choose the duplicated form in (3b) as the target of a construction:
(4) a. Çocuklar mocuklar akıcı konuşmazlar. childPLU M-DUP fluently speak-NEG-AOR-3PL 'Children (and the like) do not speak fluently'
b. Mocuklar hiç konuşmazlar. M-DUP never speak-NEG-AOR-3PL lit. 'The likes do not speak at all.'

## Doubling

Doubling occurs in two ways: simple doubling, and doubling in lexical formations. In simple doubling, the word is repeated, as in (5). Depending on the syntactic category of the targeted lexeme, it can produce adverbials, adjectivals and measure terms (Göksel \& Kerslake 2005).

## (5) tek tek zaman zaman <br> one DUP time DUP <br> 'one by one' 'time to time'

Some additional morphemes, such as the plural suffix and the question particle ( QP ), are attached to the sister constituents as in (6a) and (6b), or one of the constituents undergoes phonetic changes, as in (6c), for doubling in lexical formations.
(6) a. güzel-ler güzel-i bir kız beautiful-PLU beautiful-POSS a girl 'a very beautiful girl'
b. güzel mi güzel bir kız
beautiful QP beautiful a girl
'a very beautiful girl'
c. ufak tefek bir kutu
little $\phi_{i}$ (little) a box
'a tiny box'
Among these alternatives, the last one seems closest to a morphological-lexical operation. $\phi_{i}$ in (6c) stands for cophonology (Orgun, 1996; Orgun, 1999; Inkelas \& Zoll, 2005), which is the morphological function associated with particular morphological constructions to model morphologically conditioned phonology. The basic idea is shown in Figure 1 .


Figure 1: Template for cophonologies.

Inkelas \& Zoll (2005) employ cophonologies in their Morphological Doubling Theory (MDT), and stress that the theory is morphologically motivated because it makes use of roots, morphs and affixes, rather than mora, coda or foot. The model works in a binary manner, in which there are two inputs called daughter nodes, and the output in the tree's root is called the mother node. In MDT, the reduplicant and base are both generated by morphology as part of a construction that also embodies semantic and phonological generalizations concerning the output of reduplication (Inkelas, 2005).

However, Göksel \& Kerslake (2005:101) classify lexical formations as idiomatic expressions, and our intuitions are consistent with this observation. Take for example the following from the same page in their book:

```
(7) a. konu komşu
    LEX-FORM neighbor
    'neighbors', lit. 'neighbor-neighbor'
b. süklüm püklüm
crestfallen LEX-FORM
'in a crestfallen manner'
```

Firstly, they are not always right-headed or left-headed, which seems a bit unusual for a purportedly morphological operation. Secondly, as Göksel \& Kerslake (2005) point out, the copy may or may not exist independently. We add to this an additional observation that the doubled word is indeed a phonological word, not a suffix, prefix, and not necessarily a lexeme.

It is not clear to us whether examples such as ( 6 c ) should be made part of an experiment on lexical statistics versus speaker production of bound elements, because it is not clear whether something other than multi-word word formation is involved here, although we are quite certain that independent words should be left out, i.e. examples such as ( $6 \mathrm{a}-\mathrm{b}$ ). We chose to take idiomatic lexical formations ( 6 c ) outside the scope of our experiments too, and concentrate only on TER.

In Morphological Doubling Theory, MDT, the features given above can be ranked by the cophonologies to determine the linker in the emphatically reduplicated form of TER. Yu (1999) argued that the allomorphy in Turkish reduplication could be accounted for by positing morphotactic constraints, which spell out the form of each of the allomorphs that dominate certain phonotactic constraints. The ultimate selection of the appropriate morph depends on the harmonic satisfaction of the lower-ranked phonotactic constraints on the linker.

Demircan (1987) analyzed 121 emphatically reduplicated adjectives, and concluded that the frequency of reduplicated adjectives showed the ranking $-p>-m>-s>-r$, in which $-p$ is the most selected linker, and $-r$ the least. Of all the adjectives, $46 \%$ are reduplicated with $-p, 29 \%$ with $-m, 18 \%$ with $-s$, and $7 \%$ with $-r$. In another study, $\operatorname{Wedel}(1999,2000)$ concluded that TER with the linker $-r$ might be lexicalized.

In contrast with these findings, and with the constraints summarized in (2), we noted some frequently occurring "exceptions" in §1, for example, among others, çı-r-ll-çıplak,
which is the expected reduplicated form of çıplak, but also çı-s-çıplak, çı-m-çıplak, çı-r-çıplak and çı-p-çıplak.

There seems to be no easy generalization across the speakers about how they emphatically reduplicate a novel word. Auspiciously, there are tendencies depending on the linker type, with respect to lexical co-occurrence frequencies, which we explain below.

## Method and Findings

In order to thoroughly understand the linker selection choices available to native speakers of Turkish, a questionnaire consisting of 31 nonadjectival words composed of Turkish nouns and verbs was prepared. The word list was given to 25 male and 25 female participants, all university graduates, average age 34.20, and all native speakers of Turkish. Nonadjectives were deliberately selected to guarantee that the participants would be very unlikely to have applied TER to the words in the list before. The participants were asked the following questions: if the words were to be treated as adjectives, how would they emphatically reduplicate them? They were allowed to give single word answers. We also asked whether they had ever reduplicated the words, how they knew how to reduplicate the words, and the average time in seconds it took them to reduplicate each word. Because our plans did not involve reaction time experiments, we considered the last piece of information as the perceived difficulty and/or effort on behalf of the speaker. About two-thirds of the subjects took the experiment in our presence, and our timing seems to concur what they reported.

We have been told that they had never reduplicated any of the words before. For their own explanation of self performance, the participants responded that they reduplicated the words 'intuitively', and that each word required about $5 \mathrm{sec}-$ onds or less for reduplication. All the participants used $-p$, $-m,-s$, and $-r$ for the linker position, but none used $-A,-I l$ or $-A m$ as an additional infix. This result is shown in more detail in Table 1.

Looking closer at the results, most productions seem to satisfy the constraints previously reported (Demircan, 1987; Wedel, 1999; Wedel, 2000), but, certainly, some formations such as böpböcek, firfirın, mammasal and kemkemir, violate these constraints. Unlike the study by Demircan (1987), the order of the linker type frequencies in this study is $-p>-s>-$ $m>-r$. Moreover, the reduplicated forms with $r$-linkers seem to disconfirm Wedel's $(1999 ; 2000)$ conclusion that $r$-forms might be lexicalized. They seem to be just less frequently used.

Explanations for these findings might lie in lexical statistics, in particular n-grams of graphemes, which we use in lieu of phonemes because of lack of speech data. For this end we used corpora to approximate a lexical statistic. When we examined the METU-Sabancı Turkish Treebank (Atalay et al., 2003), we found that there are 43,571 roots, of which 5,533 are distinct. The linker order found in the current study is exactly opposite of the frequency of words with roots that end

Table 1: TER results for the nonadjectives.

| Word | $p$-linker | $m$-linker | $s$-linker | $r$-linker |
| :---: | :---: | :---: | :---: | :---: |
| bıçak | 11 | 10 | 28 | 1 |
|  | (bıpbıçak) | (bımbıçak) | (bısbıçak) | (bırbıçak) |
| böcek | 13 | 8 | 25 | 4 |
| cevap | 30 | 0 | 16 | 4 |
| cami | 28 | 8 | 14 | 0 |
| çorba | 29 | 0 | 16 | 5 |
| dilek | 32 | 11 | 7 | 0 |
| davet | 26 | 3 | 21 | 0 |
| duvar | 27 | 7 | 16 | 0 |
| ĕ̆len | 50 | 0 | 0 | 0 |
| firin | 17 | 2 | 23 | 8 |
| felek | 12 | 3 | 30 | 5 |
| getir | 32 | 0 | 18 | 0 |
| götür | 37 | 0 | 13 | 0 |
| hüzün | 43 | 2 | 5 | 0 |
| jilet | 36 | 6 | 8 | 0 |
| kıble | 18 | 2 | 30 | 0 |
| kemir | 23 | 8 | 14 | 5 |
| leğen | 36 | 3 | 11 | 0 |
| laf | 26 | 7 | 17 | 0 |
| masal | 14 | 5 | 29 | 2 |
| nizam | 32 | 8 | 10 | 0 |
| pırasa | 14 | 9 | 17 | 10 |
| resim | 38 | 4 | 8 | 0 |
| surat | 43 | 6 | 0 | 1 |
| seçim | 32 | 9 | 0 | 9 |
| şerit | 26 | 19 | 0 | 5 |
| tutkal | 37 | 5 | 8 | 0 |
| tekerlek | 24 | 7 | 18 | 1 |
| vazo | 29 | 2 | 19 | 0 |
| yutkun | 38 | 3 | 9 | 0 |
| zarf | 40 | 10 | 0 | 0 |
| (\%) | 57\% | 11\% | 28\% | 4\% |

with the grapheme found in the linker; see Table 2, the second column. The third column repeats the same count in a much larger corpus, the BOUN corpus of 490 million words (Sak, Güngör, \& Saraçlar, 2011). The database has 45,035 distinct stems, whose frequency of ending with $p / \mathrm{m} / \mathrm{s} / r$ is reported in the third column. They have the same rank as the other corpus. We take the rank as an adequate relative measure of the lexical choice of endings in Turkish.

Next we consider the co-occurrence of the linker type with other consonants that are likely to be at the initial segment. For a word without TER beginning with the sequence $C_{1} V_{1} C_{2} \ldots$, the consonant co-occurrence on the boundary of the "prefix" and the base will be one taken from the set $\left\{p C_{1}\right.$, $\left.m C_{1}, s C_{1}, r C_{1}\right\}$. Prefixation is a morphological operation, and we wanted to see if early lexical access can be contrasted with early morphological processing by frequency.

One hypothesis is that the linker may be selected so that the

Table 2: Number of distinct stems/roots terminating with the same grapheme as a linker type, in two large corpora.

| Root ending | METU-Sabanc1 | BOUN |
| :--- | :---: | :---: |
| $p$ | 100 | 639 |
| $s$ | 128 | 780 |
| $m$ | 281 | 1,620 |
| $r$ | 470 | 3,523 |

first segment of the reduplicated word has less resemblance to an existing root. One way to check this effect is to see if the consonant co-occurrence is minimized for the linker type- $C_{1}$ pairs. In order to test this hypothesis, the statistics from the METU Turkish Corpus (Say et al., 2002) is studied. Table 3 shows the number of distinct words containing the consonant co-occurrences composed of one linker and the initial grapheme of the nonadjectival word from the corpus. For example, 46 words in the corpus have $p b$ as substring. Similarly, 482, 101 and 633 words have $p m, p s$ and $p r$ as substrings, respectively.

Table 3: Linker-consonant co-occurrences in the corpus.

| Consonant | Linkers |  |  |  |
| :---: | :---: | :---: | :---: | :---: |
|  | $p-$ | $m-$ | $s-$ | $r-$ |
| $b$ | 46 | 482 | 101 | 633 |
|  | $(p b)$ | $(p m)$ | $(p s)$ | $(p r)$ |
| $c$ | 44 | 435 | 136 | 705 |
| $c$ | 112 | 13 | 48 | 602 |
| $d$ | 106 | 1599 | 148 | 9958 |
| $f$ | 25 | 28 | 66 | 191 |
| $g$ | 11 | 92 | 54 | 1519 |
| $h$ | 189 | 114 | 168 | 200 |
| $j$ | 7 | 1 | 7 | 90 |
| $k$ | 275 | 134 | 845 | 2575 |
| $l$ | 1799 | 3171 | 1655 | 8005 |
| $m$ | 340 | 257 | 519 | 5156 |
| $n$ | 90 | 82 | 140 | 559 |
| $p$ | 100 | 447 | 404 | 499 |
| $r$ | 952 | 201 | 139 | 277 |
| $s$ | 529 | 926 | 612 | 3119 |
| $s$ | 10 | 90 | 10 | 624 |
| $t$ | 820 | 109 | 4338 | 3321 |
| $v$ | 25 | 25 | 61 | 61 |
| $y$ | 132 | 122 | 719 | 346 |
| $z$ | 36 | 161 | 16 | 195 |
| $(\%)$ | $6 \%$ | $16 \%$ | $17 \%$ | $61 \%$ |

The order of the linker type selection frequencies reported by the participants is exactly the opposite of the order of the frequency of the words in which the linker and the first consonant $\left(C_{1}\right)$ co-occurred in the corpus. To exemplify: if the participants' choices occurred in the order $p C_{1}>m C_{1}>$ $s C_{1}>r C_{1}$ (where $x C_{1}$ indicates frequency of co-occurrence of $x$ and $C_{1}$, in this sequence), then the linker type and conso-
nant co-occurrences in the corpus independent of reduplication are in the order $p C_{1}<m C_{1}<s C_{1}<r C_{1}$. This seems to be true for all cases that we have tried. For example, the following are produced by the participants in varying frequencies for the word masa 'table': mammasa, masmasa, mapmasa, marmasa. When we look at the co-occurrence of $\mathrm{mm}, \mathrm{sm}$, $p m, r m$ in the corpus, their frequency rank is the opposite of the ranking of the four alternatives by the participants. Such a direct relation to lexical frequency would be surprising if phonological and morphological contrasts were the sole bases of ranking as suggested by the constraints listed in (2). The same considerations apply to cophonology.

The Mann-Whitney non-parametric test was employed to compare the distributions of the two sets given in Table 1 and Table 3. It shows that the two sets are significantly different ( $U=693, p<.05, r=.79$ ). For the participants' responses, the frequency of answers significantly and negatively correlated with linker types, ranging from $-p$ to $-r$ (Pearson's $r(124)=-.64, p<.01)$. On the other hand, the frequency of consonant collocations in the corpus significantly and positively correlate with linker types, ranging from $-p$ to $-r$ (Pearson's $r(124)=.34, p<.01)$.

## Discussion and Conclusion

The study of Turkish Emphatic Reduplication (TER) on unexpected targets, e.g. on nonadjectival words, and the 'intuitive' responses on part of the participants which violate the phonological feature constraints, show that although TER is a morphologically conditioned phonological process, it seems to depend on the knowledge of distributions such as consonant co-occurrence statistics and root ending statistics. The participants tended to dissimilate the linker type they chose from the known consonant co-occurrences and root endings of their language.

Besides the phonological constraints, selecting an appropriate linker so that the first segment of the reduplicated word has less resemblance to an existing root-word is additionally effective in the process of reduplication. This dissimilation tendency can be observed in Table 1. For example, dar-davet, dur-duvar, göm-götür, gör-götür, ger-getir, gem-getir, hürhüzün, var-vazo and zar-zarf were not produced because dar 'tight', dur 'stop', göm 'bury', gör 'see', ger 'stretch', gem 'curb', hür 'free', var 'exist' and zar 'die' are already existing stems in Turkish. Thus, selecting a linker which has a frequent (admittedly orthographic) representation in the corpus would seem to steer the speaker to considering as if there was a root instead of a prefix. This we think might point out more ways to look at morphology-lexicon relation, rather than just the "blocking" kind such as went/*goed and git/*gittir/götür (leave/cause to leave/take away). Also, considering the fact that participants were able to respond within 5 seconds, it seems to us that the speakers are putting the co-occurrence frequencies in their language to online use.

The findings underline that cophonologies, i.e. morphological functions (Inkelas \& Zoll, 2005), might require prior
knowledge of known words' frequencies to select the appropriate linker after the truncation of the base in light of the phonological constraints. To be able to employ these statistics, TER as a process needs access to a speaker's lexicon.

We therefore suggest that Turkish emphatic reduplication, an apparently phonological operation, depends on global lexical knowledge for selecting an appropriate linker whose cooccurrence with the initial consonant of the reduplicated word is infrequent. (Yavas, 1980 was first to point out the lexical source of the linker type.)

We argue further that there are sufficient reasons to take emphatic reduplication as morpholexical, rather than phonological or cophonological. First, the ranking of TER elicitations from non-TER targets conforming to TER's base form paradigm is not consistent. This suggests that something other than phonological ranking is also at work.

The lexical constraints on TER seems to be more than the stem's part of speech and lexical statistics. The process is very productive when we can entertain a $((e, t),(e, t))$ reading for TER, from a $((e, t),(e, t))$ base, be it adjective or adverb, to $((e, t),(e, t))$ result. This is a semantic constraint. Yemek 'food' is not the right type (currently it seems to be $(e, t)$ ), and we have *yepyemek/*yemyemek/*yesyemek/*yeryemek, although the first syllable of the base would be a phonologically legitimate input to TER if the semantic type could be satisfied: yeşil 'green' $\rightarrow$ yemyeşil 'all green'. Place names cannot be of the right semantic type either, and similarly fail to undergo TER if not forced: Mordoğan $\rightarrow$ *mosmordoğan.

The semantics of reduplication works on aspectual properties or intensive aspects, depending on the morphological and lexical property of the $((e, t),(e, t))$ base: çabuk 'quick' $\rightarrow$ çarçabuk 'in haste', which is aspectual and derivational (cf. syntactic/phonological reduplication çabuk çabuk 'hurriedly'), and mor 'purple' $\rightarrow$ mosmor 'deep purple', which is intensive. Note also the case of yeşil $\rightarrow$ yemyessil above, which seems to be intensive in some other way.

Connectives and postpositions are the hardest targets for TER, presumably because of their semantic type in addition to morphology: ama 'but' $\rightarrow$ *apama, göre 'according to' $\rightarrow$ *göpgöre. Other semantically potential targets for TER, e.g. VPs, are in fact $(e, t)$, and as such they behave as expected: иуи 'sleep' $\rightarrow$ *upиyu, düşün 'think' $\rightarrow$ *düpdüşün. Additional linkers $\{-A,-I l,-A m\}$ are never used by our participants, which suggests that such forms in (1) are probably lexicalized.

It is clear that the reduplicated "prefix", the linker types $\{p$, $s, m, r\}$, or the "infix" from $\{-A,-I l,-A m\}$ are not morphological objects. They are not affixes or morphemes. It seems also clear that the process is not purely lexical or phonological. Its "allomorphy" is open-ended; there seems to be no discernible TER morpheme, or a purely morphophonological process.

We point out that (i) the process is codetermined by morphology and the lexicon, (ii) its semantics depend on lexical properties, and (iii) it cannot be repeated: masmavi $\rightarrow$ *masmasmavi and apaçık $\rightarrow$ *apapaçık. Therefore it is most likely
a morpholexical rule, with subsequent phonological effects, rather than causes, as exponence.

## References

Atalay, N., Oflazer, K., \& Say, B. (2003). The annotation process in the Turkish Treebank. In Proceedings of the EACL workshop on linguistically interpreted corpora (pp. 13-14). Budapest, Hungary.
Demircan, O. (1987). Emphatic reduplication in Turkish. In H. E. Boeschoten \& L. T. Verhoeven (Eds.), Studies on modern Turkish: Proceedings of the 3rd conference on Turkish linguistics (pp. 24-41). Tilburg: Tilburg University Press.
Dhillon, R. (2009). Turkish emphatic reduplication: Balancing productive and lexicalized forms. GLS, 71, 3-20.
Göksel, A., \& Kerslake, C. (2005). Turkish: A comprehensive grammar. London: Routledge.
Güngör, T. (2003). Lexical and morphological statistics for Turkish. In Proceedings of TAINN (pp. 409-412).
Inkelas, S. (2000). Phonotactic blocking through structural immunity. In B. Stiebels \& D. Wunderlich (Eds.), Lexicon in focus (pp. 7-40). Berlin: Akademia Verlag.
Inkelas, S. (2005). Morphological doubling theory: Evidence for morphological doubling in reduplication. In B. Hurch (Ed.), Studies on reduplication (pp. 63-86). Berling: Mouton de Gruyter.
Inkelas, S., \& Orgun, C. O. (1995). Level ordering and economy in the lexical phonology of Turkish. Language, 71, 763-793.
Inkelas, S., \& Zoll, C. (2005). Reduplication: Doubling in morphology. Cambridge: Cambridge University Press.
Kelepir, M. (2000). To be or not to be faithful. In A. Göksel \& C. Kerslake (Eds.), Studies on Turkish and Turkic languages. Harrassowitz Verlag.
Kılıç, O., \& Bozşahin, C. (2012). Semi-supervised morpheme segmentation without morphological analysis. In Proceedings of the workshop on language resources and technologies for Turkic languages, LREC. İstanbul, Turkey: ELRA.
Kim, H. (2007). The full-to-partial reduction in Korean and Turkish reduplication. Linguistic Research, 26, 121-148.
Kornfiltl, J. (1997). Turkish. London and New York: Routledge.
Lewis, G. (2000). Turkish grammar. Oxford: Oxford University Press. (2nd ed)
Orgun, C. O. (1999). Morphological doubling theory: Evidence for morphological doubling in reduplication. In B. Hermans \& M. von Oostendorp (Eds.), The derivational residue in phonological optimality theory (pp. 247-267). Amsterdam: John Benjamins.
Oztaner, S. M. (1996). A word grammar of Turkish with morphophonemic rules. Master thesis, Department of Computer Engineering, Middle East Technical University.
Sak, H., Güngör, T., \& Saraçlar, M. (2011). Resources for Turkish morphological processing. Language Resources
and Evaluation, 45(2), 249-261.
Say, B., Zeyrek, D., Oflazer, K., \& Ozge, U. (2002). Development of a corpus and a treebank for present-day written Turkish. In K. Imer \& G. Dogan (Eds.), Proceedings of the eleventh international conference of Turkish linguistics (pp. 183-192). Famagusta, Northern Cyprus: Eastern Mediterranean University Press.
Wedel, A. (1999). Turkish emphatic reduplication. Master thesis, Department of Linguistics, University of California, Santa Cruz.
Wedel, A. (2000). Perceptual distinctiveness in Turkish emphatic reduplication. In Proceedingsof the 19th west coast conference on formal linguistics (pp. 546-559).
Yavas, M. (1980). Borrowing and its implications for Turkish phonology. Unpublished doctoral dissertation, Univ. of Kansas.
Yu, A. (1999). Dissimilation and allomorphy: The case of Turkish emphatic reduplication. Master thesis, Department of Linguistics, University of California, Berkeley.

# Connecting Learning Goals and Component Cognitive Skills in Digital Games 

Stephen S. Killingsworth (s.killingsworth@vanderbilt.edu)<br>Department of Teaching and Learning, Box 230, 230 Appleton Place<br>Nashville, TN 37203 USA

Douglas Clark (doug.clark@vanderbilt.edu)<br>Department of Teaching and Learning, Box 230, 230 Appleton Place<br>Nashville, TN 37203 USA


#### Abstract

Growing bodies of research have investigated how digital games might be used as pedagogical tools and separately, how playing commercial games influences basic cognitive capacities or skills. The goal of the present research is to draw from these separate lines of research to ask how changes in basic cognitive capacities and formal learning gains may be related. The present study employed a game in which a ship moves through different environments using forces. The game teaches the basic relationships between objects and forces in Newton's Laws of Motion. Students played one of two versions of the game. The predictive version encouraged planning and reflection, by allowing students unlimited time to place forces along a path. In the real-time version, forces immediately affected the player when selected. The results suggest that learning was equivalent across the versions, but changes in attentional capacities may differentially contribute to learning between versions.


Keywords: Education; Psychology; Learning; Classroom studies; Experimental research with children; Digital games

## Introduction

Video games have been present in mainstream culture for decades, but have recently become a popular topic for research. One branch of research on video games in Cognitive and Social Psychology, has investigated the impact of recreational game play on basic cognition and behavior. A second branch of research has investigated the impact of games specifically designed to teach concepts within a discipline. Though these divisions do not cover all the relevant work, they do account for a majority of publications on videogames. In the present work, we investigate how games can train concepts and basic cognitive capacities. Beyond this, we begin to address the complex question of how cognitive skill training and discipline-specific learning may each contribute to learning gains on an assessment of students' basic understandings of Newton's Laws of Motion.

Much of the recent research on videogames in Cognitive Psychology has been connected to the somewhat surprising finding that some of commercial action video games may actually train basic cognitive capacities of players (e.g. Dye, Green, \& Bavelier, 2009; Feng, Spence, \& Pratt, 2007). One particularly interesting finding is that games may train networks that control three basic aspects of visual attention (Dye et al., 2009). There have been some concerns about the conclusions drawn in these studies (Boot, Blakely, \&

Simons, 2011). However, the possibility of a positive impact of games that may otherwise have negative social effects (e.g., Carnagey, Anderson, \& Bushman, 2007) has been a compelling topic for research.

Other research on videogames for learning has focused on learning discipline specific content knowledge, skills, processes, attitudes, and engagement (e.g., NRC, 2010). This education-focused work spans several fields and is often referred to as research on "serious games" or "games for learning" although there are multiple other names as well. Again, this work has typically focused on how games produce learning gains in a particular discipline or skill.

The present project differs from most prior Cognitive Psychology and Education-focused work, but is designed to benefit from the approaches of both of those areas of research. The present work uses a conceptually-integrated game (Clark \& Martinez-Garza, 2012) under development called EGAME in which the target concepts are integrated directly into gameplay mechanics, rather than being presented through separate activities. The basic prototype of the game involved in this study (see Figure 1) was designed to promote an accurate intuitive understanding of Newton's Laws. The game provides puzzle-like scenarios in which players use a limited palette of forces to move a ship to a target. Unlike in many popular games, movement in this game is controlled by combining unidirectional forces of varying magnitudes and durations. Furthermore, the game models realistic motion and is sensitive to the constraints of the environment (e.g., the presence or absence of friction).

Two versions of our game prototype were used in this study. The first, predictive, version of the game was designed to encourage planning and reflection. In this version, students dragged forces from a palette onto a level map. The students would then "run" the simulation to observe the results of their choices. This design minimized competition between cognitive resources necessary to select forces and the resources available to observe and evaluate the effects of choices. The placement play phase involved selecting locations for forces, looking at the palette, and dragging icons with the mouse. The observation play phase involved watching the ship respond to forces placed on the map (and optionally stopping the simulation).

The real-time control version of the game combined placement and observation. Students had unlimited time to look at a level and plan before selecting a force, but as each force icon was clicked, the ship moved accordingly. In this
version, students made force selections as the ship moved and had to time actions appropriately. Our expectation was that this design imposed greater cognitive demands on students. For example, the real-time game encouraged more strategies such as memorizing available forces and preparing actions before beginning a level. Moreover, the real-time version required continuous monitoring of the position of the ship and continuous shifting of attention between the force palette and the game map [see Droll \& Hayhoe (2007) for how attention and working memory may be coordinated in related contexts]. Due to the presumably greater load imposed by the real-time game, we hypothesize greater learning gains for the predictive game version than for the real-time game version (hypothesis 1).

Neither version of game was truly an "action game," like those that have been shown to train cognitive capacities in other studies, but our manipulation of game versions allowed us to isolate certain features of typical action games. More specifically, as in typical action games, the real-time game type encouraged monitoring multiple regions of the screen and timing actions with onscreen motion. Thus, the primary differences between the game types are in terms of how players must distribute attention and select relevant information. Therefore, in investigating differences in capacities that might be trained by the two game types, we focused on changes in scores on the attention network test (ANT) across players in each version of the game. Based on brain imaging and behavioral evidence, the ANT is reported to measure attentional capacities in terms of three distinct network components: (1) an executive component, related to inhibiting irrelevant information, (2) an orienting component, related to shifting the focus of attention to particular spatial locations, and (3) an alerting component, related to preparing to process upcoming information (see Dye et al., 2009 and Rueda et al., 2004). In research by Dye et al. (2009), the authors find that frequent action game players had larger scores on the executive and orienting components of the ANT and had faster baseline RTs with equivalent accuracy. Given these findings and the similarities between the real-time game and typical action games, we hypothesize that changes in orienting and executive attention networks (and baseline RT) after gameplay will be larger for the real-time game group (hypothesis 2).

In addition to measuring changes in attention networks, we investigated the relationship between gains in basic cognitive capacities, gains on our formal assessment, and measures of motivation. At the most basic level, we predict that motivation will support learning and that we will observe a positive correlation between motivation and physics learning gains for both game types (hypothesis 3). We also predict that network scores on the ANT pretest and will be more strongly positively correlated with learning gains on the physics test for the real-time game (hypothesis 4). This hypothesis is based on the premise that the real-time game imposes greater attentional demands and thus, students with a greater initial capacities might learn
more more than others fort that game. Though we do not have a specific prediction for how changes in basic cognitive capacities will relate to changes in physics understanding across versions, we also predict that changes in ANT network scores may have different relationships to learning gains across the two game versions (hypothesis 5). Our final hypothesis, following Dye and colleagues (2009) is that students that more frequently played action video games will have higher initial orienting and executive scores on the ANT (hypothesis 6).


Figure 1: Screenshot of EGAME level.

## Method

## Subjects

143 middle school students ( 70 female and 73 male) in the Southeastern United states participated in this study. The school served a racially diverse, primarily middle-class population. Students participated together during their normal $8^{\text {th }}$ grade science class for approximately 3 hours of game play and 1 hour of pre-post assessments spread across one week. The sample consisted of students from 6 classes under the same teacher. Data was only used from students who completed the assent form. All analyses only included students that completed the measures reflected in those analyses.

## Equipment

Students used MacBook Air computers to play the game. The game and cognitive tests were designed using Adobe Flash. The prototype versions of the game used in this study as well as current versions of the game can be viewed at www.surgeuniverse.com.

## Assessments and Questionnaires

Physics Understanding Students completed pre-and posttests consisting of 12 questions based on the Force Concept Inventory (Hestenes et al., 1992). Questions covered the following basic concepts relevant to understanding Newton's Laws: vector combination and diagonal motion (vectors); the relationship between velocity, acceleration, and position (acceleration); the influence of friction on
motion (friction); the influence of mass on motion (mass); and the influence of gravity on motion (gravity).

Attention Networks (ANT) We administered an adapted child-friendly version of the ANT developed by Rueda et al. (2004). The ANT evaluates the efficiency of three distinct attentional networks (executive, orienting, and alerting). In the pre- and post-test, 144 critical trials were presented in a fixed random order. On each trial (after a 1500 ms ITI), a fixation cross was presented ( 400 to 1600 ms ). Following this, one of four cue types was presented ( 150 ms ). Cues were gray circles occupying approximately the same area as the target $\left(1.7^{\circ}\right)$. Cue conditions were: no cue, a central cue (at fixation), a double cue (at possible target locations), or a spatial cue (at the upcoming target location). After a 450 ms delay, the target stimulus was presented either $1.9^{\circ}$ above or below the prior fixation location. The target was a spherical furry character used in game tutorials. The target was presented alone (neutral trials) or flanked by distractors ( 2 to the left and 2 to the right). Students responded to what direction the target was facing. On incongruent trials, distractors faced the opposite direction of the target. On congruent trials, all characters faced the same direction. The critical stimuli were presented for up to 1500 ms . Feedback was provided in the following forms at fixation: correct response: "+10 pts", incorrect response: "oops", and delayed response: "too slow".

Mental Rotation Students completed a mental rotation task adapted from Widenbauer \& Jansen-Osmann (2008). The task required students to decide whether two images were identical or mirrored. Because numerous students misunderstood the instructions and for the sake of brevity, data from this task are not discussed further.

Motivation and Engagement (QCM and GEQ) The game engagement questionnaire (GEQ) is a measure developed by Brockmyer and colleagues (2009). The questionnaire yields a single composite score of engagement in terms of: presence, flow, absorption, and immersion. Each item had three choices: "no", "sort of", and "yes". We adapted this questionnaire to refer to our game. For more details on the GEQ, see Brockmyer et al. (2009).

The QCM is a measure of achievement motivation. The QCM differentiates the following factors: anxiety, challenge, interest, and probability of success. We used a modified version of the short form of the QCM (Freund et al., 2011). Specifically, we replaced "task" with "game" in all questions and removed one concerning item: "I am afraid I will make a fool out of myself".

Gaming Experience Survey Following Dye et al. (2009), we asked students to list the 10 games they had played the most frequently in the past 12 months. Using this, students were classified as action game players or not.

## Design and Procedure

The study used a pretest-intervention-posttest design. Students were seated at lab tables mostly in pairs, though some students were alone. Students were pseudo-randomly assigned to one of two game versions (predictive or realtime) in each class. 79 students played the real-time game and 64 played the predictive game. Assignment was not random because students were allowed to sit in their typical seats and pairs of students seated together were placed in the same condition. This prevented students from seeing the alternate game version and allowed them to consult one another if they chose. All students worked individually. Before playing the game, students completed three separate tasks that were integrated with the game content: the physics pre-test (adapted from the FCI), the ANT, and the mental rotation task. After the pre-tests, students played the their version of the game. The content of the game levels roughly corresponded with one or more of the aforementioned categories of questions on the FCI-based test.

Students played the game for approximately three days of class time and completed different numbers of levels in this period according to their abilities. Several simple tutorials were included and two questions were included within the first 10 levels of the game to help students connect the material in the game to Newton's Laws.

Students completed the questionnaire on current motivation (QCM) after playing the first level of the game and the game engagement questionnaire (GEQ) after playing approximately 38 levels. Students were asked to stop playing after approximately 20 minutes on the third day. After playing, students first completed the FCI, ANT, and the mental rotation post-tests, then completed the gaming experience survey and provided feedback about the game.

## Results

## Initial Equivalence of Student Groups

The distribution of students classified as action gamers on the gaming experience survey did not significantly differ by across the game version groups (predictive vs. real-time). Furthermore, game type groups did not significantly differ prior to treatment in terms of any subscales on the physics understanding test or the ANT (i.e., Alerting, Orienting, or Executive scores).

## Measures of Motivation and Engagement

A univariate ANOVA was conducted with GEQ scores as the dependent variable and game version as a betweensubjects variable. There were no significant differences between student engagement ratings across the predictive ( $M=45.25, S D=8.09$ ) and real-time $(M=44.27, S D=$ 5.96) game versions, $F(1,124)=.62, p=.43, \eta_{p}^{2}=.01$. Additionally, separate univariate ANOVAs were conducted with QCM components as dependent variables and version as a between-subjects variable. The univariate ANOVA for

QCM "probability of success" showed a significant difference between versions, $F(1,139)=46.57, p<.0001$, $\eta_{p}^{2}=.25$. Students in the real-time game had significantly higher estimates $(M=5.00, S D=1.08)$ than students in the predictive game $(M=3.45, S D=1.62)$. Thus, it appears that students in the real-time game version may have had higher achievement motivation to start. To note, the QCM challenge component was dropped from covariate analyses due to a large correlation with the interest component, $r(139)=.61, p<.0001$.

## Student Gains and Version Comparisons

Physics Understanding A repeated-measures MANCOVA was conducted with test administration (pre vs. post) as a within-subjects factor. Game version (predictive vs. realtime) was included as between-subjects factors. Each question type (vectors, acceleration, friction, mass, and gravity) was entered as a separate dependent measure. The multivariate analysis showed that overall learning gains were non-significant from pre- to post-test, $F(5,137)=$ 2.27, $p=.05, \eta_{p}^{2}=.08$. Separate univariate ANOVAs for each question type were examined with the same factors as above. These tests showed that only the vectors question type showed small but significant learning gains, $F(1,141)$ $=6.55, p=.01, \eta_{p}^{2}=.04$, from pre- $(M=.22, S D=.26)$ to post-test ( $M=.29, S D=.29$ ). No interactions with game version were significant for any of these tests, so our hypothesis of an overall advantage for the predictive game was not supported.

Attention Networks (Baseline and Network Scores) A repeated-measures ANOVA was used to evaluate baseline RT (neutral trials) between test administration times. Game version was included as a between-subjects factor. This test did show a significant effect of test administration, $F(1$, $100)=36.27, p<.0001, \eta_{p}^{2}=.27$, with faster post- $(M=$ $542, S D=91)$ than pre-test $(M=582, S D=90)$ RTs. The interaction between test administration and version was not significant, $F(1,100)=.20, p=.66, \eta_{p}^{2}=.002$. A similar ANOVA with baseline accuracy showed no significant effects.

Following this analysis, individual repeated measures ANOVAs were used to compare gains in network scores. We calculated network scores from difference scores among median RTs following Rueda et al. (2004): Executive score = incongruent - congruent trials, orienting score $=$ spatial single cue trials, alerting score $=$ double - no cue trials. Test administration was included as a within-subjects factor and game version was included as a between-subjects factor. The ANOVA for alerting scores showed that scores significantly increased, $F(1,100)=27.48, p<.0001, \eta_{p}^{2}=$ .22 , from pre- $(M=5.96, S D=36.52)$ to post-test $(M=$ 44.46, $S D=58.71$ ) administration. The ANOVA for orienting scores showed that scores significantly increased, $F(1,100)=100.88, p<.0001, \eta_{p}^{2}=.50$, from pre- $(M=-$
33.75, $S D=46.54$ ) to post-test $(M=18.53, S D=43.08)$ administration. Additionally, there was a significant interaction between test administration and game version for orienting scores, $F(1,100)=7.46, p=.007, \eta_{p}^{2}=.07$. This interaction reflected that the differences (post-pre) in orienting scores were larger for the predictive version ( $M_{\text {Diff }}$ $\left.=74, S D_{D i f f}=55\right)$ than the real-time $\left(M_{D i f f}=42, S D_{D i f f}=60\right)$ version.

Finally, the ANOVA for executive scores showed that scores significantly increased, $F(1,100)=139.40, p<$ $.0001, \eta_{p}^{2}=.58$, from pre- $(M=32.95, S D=36.22)$ to posttest ( $M=96.82, S D=55.20$ ) administration. These findings do not support our second hypothesis. In fact, the only difference between the two game versions we observed was in the opposite of the predicted direction (with larger gains in orienting scores for the predictive game).

Attention Networks (Omnibus ANOVA, RTs and Accuracy) To compare the specific effects of cues and flankers, separate repeated-measures ANOVAs for RTs and for accuracy were conducted. However, given these analyses are not of primary importance to our research questions, the details of these analyses are not reported here. We note three important results from these analyses, however. First, no interactions involving game version were significant. Second, main effects for accuracy ANOVAs were similar to those for RT ANOVAs. Finally, we observed spatial cues reducing congruency effects, which has been observed other ANT studies. Furthermore, this effect was greater for the pre-test.

## Covariate Analyses of Student Gains

Attention Networks (Gaming Experience) A univariate ANOVA was conducted to evaluate baseline RT differences on the ANT pre-test between recreational action game players and others. Recreational action game playing was included as a random factor. Action game playing did not influence baseline RT in this comparison, $F(1,81)=.01, p$ $=.92, \eta_{p}^{2}<.001$. Following this, separate univariate ANOVAs were conducted for each ANT network score. None of the network scores were significantly different for action game players: alerting scores: $F(1,81)=.03, p=.87$, $\eta_{p}^{2}<.001$; orienting scores: $F(1,81)=3.17, p=.08, \eta_{p}^{2}=$ .04 ; and executive scores, $F(1,81)=.01, p=.93, \eta_{p}^{2}<$ .0001. Orienting was the only component to approach significance [action game players $(M=-27, S D=37)$, non action game players ( $M=-45, S D=49$ )]. Overall our results did not corroborate those of Dye and colleagues. However, we did observe a marginally larger pre-test orienting score for action game players.

Physics Understanding with Covariates First, to determine how baseline measures of attention influenced learning gains, separate repeated-measures MANCOVAs were conducted for each game version. Test administration (pre vs. post) was included as a within-subjects factor. ANT
pre-test network scores (alerting, orienting, and executive) were included as covariates. Each question type (vectors, acceleration, friction, mass, and gravity) was included as a separate measure. Neither multivariate nor univariate tests showed any significant effects of the covariates for either game version. Thus, our fourth hypothesis, that ANT pretest scores will be more closely correlated with learning gains for the real-time game was not supported.

Following the above analyses with ANT pre-test scores, a similar analysis was conducted including difference scores between the ANT pre- and post-tests, aggregate GEQ scores, and QCM component scores (probability of success, anxiety, and interest). First, separate repeated-measures MANCOVAs were conducted for each game version. For the real-time game, there was a significant interaction between test administration and ANT orienting score in the multivariate test, $F(5,45)=3.31, p=.01, \eta_{p}^{2}=.27$. None of the other effects for the real-time game were significant in the multivariate test.

Because we were interested in the specific effects for each question type, univariate tests were explored as well. For the real-time game, the interaction between test administration and ANT orienting gains was significant for the vectors question type, $F(1,49)=6.09, p=.02, \eta_{p}^{2}=.11$, and for the friction question type, $F(1,49)=6.38, p=.02$, $\eta_{p}^{2}=.12$. The interaction between test administration and ANT executive gains was significant for the vectors question type, $F(1,49)=5.17, p=.03, \eta_{p}^{2}=.10$, and for the friction question type, $F(1,49)=4.87, p=.03, \eta_{p}^{2}=.09$. Partial correlations with difference scores controlling for other covariates showed that gains on vectors and friction questions increased with smaller ANT orienting, $r(47)=-$ $.33, p=.02$, and executive gains, $r(47)=-.31, p=.03$.

For the predictive game, no effects were significant in the multivariate test. In unvariate tests for the predictive game, interactions with test administration were significant for the mass question type with the QCM anxiety, $F(1,29)=4.98$, $p=.03, \eta_{p}^{2}=.15$, and GEQ score, $F(1,29)=4.85, p=.04$, $\eta_{p}^{2}=.14$. Gains on the mass question increased with increasing QCM anxiety, $r(27)=.38, p=.03$ and GEQ engagement, $r(27)=.38, p=.04$. Similarly, gains on the gravity question increased with increasing QCM interest scores, $F(1,29)=5.83, p=.02, \eta_{p}^{2}=.17$. These findings partly support our fourth hypothesis that increased motivation would support greater physics learning gains, however this was limited to the predictive game.

Finally, the predictive game showed a significant interaction between test administration and ANT executive gains for the friction question type, $F(1,29)=4.39, p=$ $.045, \eta_{p}^{2}=.13$. Gains on the friction question increased with smaller executive gains, $r(27)=-.36, p<.05$. Together, the differences in correlations between learning gains and ANT gains for the real-time and predictive games support our fifth hypothesis.

## Discussion and Conclusions

Overall, few of our initial hypotheses were supported: Players did not demonstrate better learning with the predictive than with the real-time game (hypothesis 1), changes in attention network scores were not greater for the real-time game (hypothesis 2), scores on the ANT pre-test did not predict learning gains for either version (hypothesis 4), and action videogame players did not have higher initial network scores (hypothesis 6). However, we did observe that motivation was correlated with learning gains, at least for the predictive game (hypothesis 3 ), and we did find that changes in ANT network scores had different relationships to learning gains across the two game versions (hypothesis 5). The remainder of this section is devoted to discussing specific findings of interest.

Perhaps the most interesting finding from the above analyses is that there was a robust negative correlation between participants' orienting/executive ANT gains and physics understanding gains in the real-time game. ANT scores increased from pre- to post-test for both game version groups, suggesting that students may have had more attentional resources available to distribute attention after playing either game version (see Dye et al., 2009). However, the greater the ANT gains, the smaller were the learning gains observed in the real-time game. One interpretation of these findings is that learning gains for the real-time game were greater for those students that gained less in terms of available attentional resources though realtime game play. There may be competition for resources between learning to spread attention quickly and widely in the real-time game and resources for extracting disciplinespecific content from the game.

Another notable finding is that there were no overall differences in learning between the real-time and the predictive game. Despite the additional load presumably imposed by the real-time game, learning was equivalent. Several possible explanations will be explored in future work. Students might simply replay levels more often in the real-time game, so that load limitations are overcome. Additionally, the real-time game may have certain advantages over the predictive game. One possible advantage is that students are not required to anticipate or visualize the results of cumulative force applications to form a coherent plan - students implement plans piecemeal, as needed. Each decision can be made relative to the current direction of motion and about how each force will alter the current trajectory. Furthermore, in the real-time version, students get immediate feedback about whether each action undertaken results in an expected outcome.

For the predictive game, multivariate tests showed a somewhat greater influence of motivation and engagement, such that greater motivation/engagement was correlated with larger learning gains. One possibility is that performance on the predictive game was influenced by motivation due to the time gap between planning, execution/observation, and revision. If students failed to use what they observed to inform a subsequent placement phase,
then they may have adopted something more like a trial-and-error approach at each placement phase. However, the real-time game delivered just-in-time feedback on choices, which may have facilitated identifying incorrect actions even with lower motivation.

Another interesting finding was the interaction between game version and test administration for ANT orienting scores. Considering orienting/executive scores were both larger for game players in the Dye study, we might expect to see larger orienting scores for the real-time game because the rapid responses required game are more similar to those required in action games. However, we observed the opposite (greater orienting score changes for the predictive game). In the predictive game, (1) there were additional visual landmarks (forces placed by the student) to monitor as the ship approached and (2) attention could be devoted exclusively to orienting to relevant landmarks in the observation phase (as forces were not being selected). Such differences may account for gains in ANT orienting scores. Interestingly, these gains in orienting scores did not correlate with learning gains for the predictive game. This could strengthen the claim that orienting gains were obtained from improving monitoring of relevant landmarks during motion, which one would not expect to influence physics learning.

A final point involves the comparison of individuals classified as action game players to other students. Dye and colleagues (2009) showed that action game players had higher scores on orienting and executive ANT components and faster baseline RTs (but with equal accuracy). In contrast to these prior findings, we found only a marginal relationship between prior gaming experience and ANT orienting scores. These differences may be due to differences in the form of the ANT administered. Another difference that may have contributed was our testing the ANT in a classroom whereas Dye et al. tested in the home. Despite these differences, it is worth noting that our participants showed increased scores from pre- to post-test in the direction expected from Dye et al.'s results for recreational playing. Thus, it does seem that playing our game may induce changes in attentional networks.

One limitation of this study is that there was no baseline condition with which to compare the game version treatments. Therefore, gains on the physics assessment and in the components of the ANT could result from a testing effect. Preliminary results do indicate that EGAME produces larger physics learning gains than a control game with adult participants. Future research will need to address this issue.

A second limitation of this study involves the prototype nature of the versions of the EGAME game at the heart of this study. EGAME is being continually improved based on these and other findings, but the game is still a work in progress. Our assumption (underscored here) has been that without scaffolding, formal learning gains will be minimal. Two future plans involve (1) introducing feedback based on the game play and (2) incorporating dialog interactions to
support explicit articulation through self-explanation and directed questioning. The results of the current study are an important step toward integrating basic research on cognition and learning with applied research informing the design of digital games for learning.

## Acknowledgments

This material is based upon work supported by the NSF under Grant No. 1119290.

## References

Boot, W. R., Blakely, D. P., \& Simons, D. J. (2009). Do action video games improve perception and cognition. Frontiers in Psychology, 2(226), 1-6.
Brockmyer, J. H., Fox, C. M., Curtiss, K. A., McBroom, E., Buckhart, K. M., \& Pidruzny, J. N. (2009). The development of the game engagement questionnaire: A measure of engagement in video game-playing. Journal of Experimental Social Psychology, 45, 624-634.
Carnagey, N. L., Anderson, C. A., \& Bushman, B. J. (2007). The effect of video game violence on physiological desensitization to real-life violence. Journal of Experimental Social Psychology, 43, 489- 496.
Clark, D. B., \& Martinez-Garza, M. (2012). Prediction and explanation as design mechanics in conceptuallyintegrated digital games to help players articulate the tacit understandings they build through gameplay. Steinkuhler et al. (Eds.), Games, Learning, And Society (pp. 279-305). Cambridge: Cambridge University Press.
Droll, J. A., \& Hayhoe, M. M. (2007). Trade-offs between gaze and working memory use. JEP: Human Perception and Performance, 33(6), 1352-1365.
Dye, M. W. G., Green, C. S., \& Bavelier, D. (2009). The development of attention skills in action video game players. Neuropsychologia, 47, 1780-1789.
Feng, J., Spence, I., \& Pratt, J. (2007). Playing an action video game reduces gender differences in spatial cognition. Psychological Science, 18(10), 850-855.
Freund, P. A., Kuhn, J.-T., \& Holling, H. (2011). Measuring current achievement motivation with the QCM: Short form development and investigation of measurement invariance. Personality and Individual Differences, 51(5), 629-634.
Hestenes, D., Wells, M., \& Swackhamer, G. (1992). Force Concept Inventory. The Physics Teacher, 30, 141-158.
NRC (Honey, M. A., \& Hilton, M. Eds.) (2010). Learning Science Through Computer Games and Simulations. National Research Council. Washington, DC: National Academy Press.
Rueda, M. R., Fan, J., McCandliss, B. D., Halparin, J. D., Gruber, D. B., Lercari, L. P., Posner, M. I. (2004). Development of attentional networks in childhood. Neuropsychologia, 42(8), 1029-1040.
Wiedenbauer, G. \& Jansen-Osmann, P. (2008). Manual training of mental rotation in children. Learning and Instruction, 18(1), 30-41.

# Two for one? Transfer of conceptual content in bilingual number word learning 

Katherine Kimura<br>kkimura@ucsd.edu<br>Department of Psychology<br>University of California, San Diego

Katie Wagner<br>kgwagner@ucsd.edu<br>Department of Psychology<br>University of California, San Diego

David Barner<br>barner@ucsd.edu<br>Department of Psychology<br>University of California, San Diego


#### Abstract

Bilingual speakers are confronted with a unique challenge when learning language as they must learn to express the same concept in two separate languages. Here, we examine whether learning number words in one language (i.e., L1) facilitates the acquisition of analogous number words in a second language (i.e., L2) or whether extensive experience and familiarity with numbers within the second language is required to learn words in L2. To do so, we tested 68 bilinguals speakers between the ages of 2 and 4 years and show that conceptual knowledge of numbers in L1 reliably predicted children's conceptual knowledge of numbers in L2, suggesting that knowledge transferred from one language to the other. The effect, however, was limited to two developmental transitions: one-knower to two-knower and subset knower to CP-knower. Familiarity with L2 numbers as well as age were also significant predictors of children's conceptual understanding of numbers.


Keywords: bilingualism; conceptual transfer; word learning; number words.

## Introduction

When children learn language, they are confronted with the problem of discovering how words encode conceptual content and thus encode their experience of the world. Although children eventually overcome this challenge and learn to associate specific words with specific concepts, this process is slow and often involves making difficult inductive inferences regarding the meanings of words (Quine, 1960). In these cases where slow, inductive inferences are required it is often unclear whether children's difficulty lies with forming the concept to be referenced (see Carey, 2009) merely mapping the correct linguistic symbol to the correct concept. This distinction between conceptual and linguistic development is difficult to disentangle because linguistic experience is almost always correlated with other factors that influence conceptual development including biological maturation and non-linguistic experience. Although there are some striking examples where language can be isolated from these other developmental factors, for example international adoptees as well as late learners of sign language, these cases may have limited generalizability due to severe linguistic delay or a sharp disruption of first language learning.

In contrast, bilingual children sometimes have limited knowledge of their second language (i.e., L2) while still having an intact first language experience (i.e., L1). This separation of conceptual development and L2 linguistic development can provide a unique test case for exploring how linguistic competence and conceptual development are
related, while avoiding the challenges introduced by latelearners of sign language and international adoptees.

More specifically, bilingual speakers allows us to test whether conceptual learning accomplished first in one linguistic medium might facilitate the acquisition of corresponding content in a second language by eliminating several steps in the second language acquisition, thus resulting in a faster second language acquisition rate relative to the first language acquisition. In cases where children must acquire concepts before mapping language to those concepts, L2 acquisition of those words should be faster than L1 acquisition because these concepts can transfer from L1 to L2. In contrast, when children merely require increased exposure to the language in order to map a word to a pre-existing concept, no L2 facilitation would be expected because this process is necessarily languagespecific.

In the present study we explored this idea by investigating the acquisition of number words (e.g., one, two, three) - a central test case in the study of conceptual change (see Carey, 2009). To do so, we tested children learning two languages and asked whether learning number word meanings in one language (i.e., L1) facilitates the acquisition of analogous number words in a second language (i.e., L2).

Early in acquisition, children as young as two years learn to recite a partial count list in a serial order (e.g., one, two, three, four, five, etc.), pointing at objects as they do so (see Gelman \& Gallistel, 1978; Frye, Braisby, Lowe, Maroudas, \& Nicholls, 1989; Fuson, 1988). Despite this seemingly procedural understanding of the relationship between counting and cardinality, children at this stage in development typically have little to no understanding of how counting represents number (i.e., how the last number of the count list represents the exact cardinality of the set) nor have they acquired the meanings of any of the number words (i.e., that numbers refer to specific quantities of a set). Soon, however, children begin to acquire an exact meaning for the number one, reliably giving one object when asked for one and more than one when asked for a contrasting number. After six to nine months as a 'oneknower,' children learn the meaning of two, becoming a 'two-knower' and, following this sequential pattern, learn the meanings of three and four (Wynn, 1990, 1992). During these early stages of number word learning, these children who are classified as one-, two-, three-, and four-knowers have meanings for only a subset of their number words (i.e., one, two, three, and four) and are thus collectively referred to as 'subset knowers.' Eventually, twelve to eighteen
months after children first acquire the concept of one, they discover the cardinal principle that governs counting and recognize that the counting procedure can be used to label the cardinality of sets, at which point they are considered Cardinal Principle knowers or 'CP-knowers' (for evidence and discussion regarding these stages, see Le Corre \& Carey, 2007; Lee \& Sarnecka, 2011; Piantadosi, Goodman, \& Tenenbaum, 2012; Sarnecka \& Carey, 2008; Wynn, 1990, 1992; for discussion of what these children actually know, see Davidson, Eng, \& Barner, 2012).

By the time children acquire the concept of cardinality, they have experienced at least three important qualitative shifts in their conceptual understanding of numbers that differentiates non-knowers from one-knowers, one-knowers from two-knowers, and subset knowers from CP-knowers (for review, see Carey, 2009). First, in order to learn the meaning of one, children must have already acquired their first linguistic representations of an exact cardinality. During this stage, children begin to recognize that number words represent specific numerosities, for example that one represents precisely one item rather than an undefined quantity or an amount defined by contrasting a number, such as not one. Second, when children acquire the meaning of two in languages that mark the singular-plural distinction, they experience a fundamental shift in their understanding of quantity that differentiates one-knowers from twoknowers. Unlike the concept of one, which corresponds with the singular marker ' $a$,' the specific concept of two is not marked by morphology in English, French, and Spanish (the languages targeted in this study) as the plural morphology can refer to sets of any size two or greater. This suggest that once children acquire the concept of two, they may undergo a conceptual leap as they must acquire this new concept of duality. Third, when children become CP-knowers, they learn of the unique relationship between counting and cardinality, specifically that the counting procedure assigns number words to sets. That is, children understand that the last number recited in the count list refers to the specific cardinality of the set.

The idea that each stage involves significant conceptual change predicts that, once such changes have occurred in one language (i.e., the PNL), subsequent learning of words that encode identical concepts in a second language (i.e., the SNL) should be substantially easier, at least to the degree that acquisition in the PNL is delayed by the process of constructing the relevant content. For those children who understand the unique relationship between counting and cardinality, learning may also be facilitated by recognizing that the two count lists server similar functions in the respective languages, thus allowing the formation of an analogical mapping between the two (for a discussion of analogical mapping, see Gentner, 1983; 2003; Gentner \& Markman 1997).

Although the construction of conceptual content predicts that children's understanding of numbers may transfer from one language to another, it is equally possible that acquiring concepts in one language is independent of acquiring
identical concepts in a second language. For example, children may learn the meaning of one as a function of the frequency of associations between the word 'one' and sets of one, a process that is irrespective of children's acquisition of 'uno' in Spanish or 'un' in French. That is, knowledge may be acquired as a result of exposure to the number word and, therefore, may be represented and stored in the language in which the concept was originally acquired and fail to automatically transfer to a second language.

Previous studies of mathematical competence in bilinguals find little evidence of transfer across languages. For example, bilingual speakers exhibit a strong preference for one language over another when performing arithmetic, sometimes preferring the language of original instruction despite being a dominant speaker of another language (Dehaene, 1997; Dehaene, Spelke, Pinel, Stanescu, \& Tsivkin, 1999; Spelke \& Tsivkin, 2001). When asked to perform simple mathematical computations in their second language, bilinguals not only perform calculations more slowly but also do so with lower accuracy (Kolers, 1968; Marsh \& Maki, 1976; McClain \& Huang, 1982). However, these studies tell us little about how earlier processes of bilingual learning and representational transfer take place. This is because these studies focus on mathematical operations, which may depend on a different, broader, set of representational resources, including memorized procedures and facts that may be uniquely dependent on linguistic encoding (for discussion, see Dehaene, 1997). Thus, although relevant to understanding the bilingual representation of number, these studies do not directly address whether early transfer of numerical concepts is possible in bilingual learners, and thus whether the foundations of arithmetic learning can be shared across languages. Here, within the context of number word acquisition, we explored this issue by testing children who were second language learners and assessing their ability to successfully denote the cardinalities of number words in each language.

In the present study, we tested two populations of bilingual 2- to 4-year-olds: French-English speakers and Spanish-English speakers. Children participated in two tasks in each language. First, they completed a Give-a-Number task, which assessed their comprehension of number words, and second they completed a counting task, which assessed their familiarity with the count list in each language thus acting as a proxy for their relative exposure to numbers. Both of these measures allowed us to ask whether, when controlling for counting ability and, thus, familiarity with numbers, knowledge of number words in the L2 was predicted by knowledge of number words in the L1. More specifically, we asked whether there was evidence of conceptual transfer in subset knowers, CP-knowers, or both. Thus, we tested whether transfer is mediated by earlier acquisition of exact cardinal meanings, like one, two, three, and four, and whether it can be mediated by learning how the counting procedure works when children become CPknowers.

## Method

## Participants

Sixty-eight bilingual learners of either English and French or English and Spanish from the San Diego metropolitan area participated. In the French-English (FE) sample, 23 children (13 male) between the ages of 2;11 and 5;0 ( $M=$ $3 ; 9, S D=0 ; 6$ ) participated. These children were primarily recruited from a French language preschool where instruction is conducted exclusively in French. In the Spanish-English (SE) sample, 45 children (22 male) between the ages of $2 ; 2$ and $5 ; 0(M=4 ; 2, S D=0 ; 9)$ participated after being recruited from either a Spanish immersion preschool or a departmental database. Participants were from predominately Non-Hispanic Caucasian or from Hispanic middle-class families and were contacted either through letters distributed by teachers at local preschools or by phone using a departmental recruitment database. An additional 15 children participated but were excluded for completing the tasks in a language other than the one being tested, for example, speaking in Spanish when the tasks were conducted in English ( $\mathrm{N}=2$ ), for being trilingual speakers $(\mathrm{N}=2)$ and for failure to complete the counting task in at least one language ( $\mathrm{N}=$ 11).

As reported by the caregivers, 5 of the FE children were primarily French speakers, 14 of the SE children were primarily Spanish speakers, and 43 of the FE $(\mathrm{N}=14)$ and SE ( $\mathrm{N}=29$ ) children were primarily English speakers. Two additional children were listed as having both Spanish and English as their primary language. For the remaining four children, no primarily language was reported.

## Procedures

Testing sessions lasted approximately 20 minutes and consisted of two tasks: a Give-a-Number task followed directly by a counting task. Both tasks were administered once in English and once in either French or Spanish, such that each child completed both tasks in one language before completing identical tasks in his or her second language. The order in which the languages were tested was randomized across children. As an additional measure of a child's fluency in both languages, we initially asked caregivers to complete the Language Development Survey in English and either French or Spanish (Rescorla, 1989). However, because many parents were unable to complete the survey in both languages (e.g., parents were monolingual), we discontinued its use and do not report the data here.

Give-a-Number Task This task was adapted from Wynn (1992) using the non-titration method developed by Sarnecka and Carey (2008) and was used to assess children's comprehension of number words in each language. The experimenter began by presenting the child with a red paper plate and ten plastic fish and inviting the child to play a game with her toys. For each trial, the
experimenter asked the child to place a quantity of the fish inside the red circle, omitting singular and plural markings by asking, for example, "Can you put $N$ in the red circle? Put $N$ in the red circle and tell me when you're all done." Once the child responded, the experimenter then asked, "Is that $N$ ? Can you count and make sure?" and encouraged the child to count in the language tested. If the child recognized an error, the experimenter allowed the child to change his or her response. Following the completion of each trial, the objects were returned to their original positions and the next trial was administered until all were completed.

Participants completed up to twenty-one trials, consisting of three trials for each of the seven numbers tested (i.e., 1, 2, $3,4,5,8$, and 10 ). The order was quasi-randomized such that each number was tested once before any number was repeated, thus resulting in three sets of seven numbers. Children were defined as an $N$-knower (e.g., three-knower) if they correctly provided $N$ (e.g., 3 fish) on at least two out of the three trials that $N$ was requested and, of those times that the child provided $N$, did so in response to a request for $N$ on at least two-thirds of all trials. If children responded correctly on two out of the three trials for each number tested, then they were classified as CP-knowers.

Counting Task After administering the Give-a-Number task, the experimenter asked the child, "Can you count as high as you can?" If the child failed to respond or indicated that he or she did not know how to count, the experimenter provided the first number of the count list (e.g., one) with rising intonation in an attempt to clarify the instructions and encourage the child to continue counting. In the event that the child failed to respond after the prompt, the experimenter reassured the child and ended the task.

After the task, the experimenter recorded the highest number recited, noting any errors such as omission (e.g., "...13, 14, 16") and cyclical repetition (e.g., "...8, $9,10,1$, 2 "). The child's highest number was defined as the largest number counted to before error. For example, fourteen was the highest number recorded for a child who omitted fifteen, whereas ten was the highest number recorded for a child who cyclically repeated the first ten numbers. In cases where children failed to accurately count at the onset of the task yet recited a string of numbers (e.g., " $6,7,8 \ldots$.."), the highest number was recorded as zero. In contrast, children who refused to count were excluded from the analysis.

For each child, the language with the highest number recorded (e.g., fourteen) was coded as his or her Primary Number Language (i.e., PNL), while the language with the lowest number recorded (e.g., diez) was coded as the child's Secondary Number Language (i.e., SNL). For example, a child who counted to fourteen in English and diez in Spanish was coded as having English as her PNL and Spanish as her SNL. In cases where the highest number recited was matched in both languages (e.g., ten and diez), PNL was defined as the child's primary language as reported by the parent $(\mathrm{N}=1)$ or, when the parent indicated no preference for either language, was instead coded as

English ( $\mathrm{N}=1$ ). The highest number children counted to without error ranged from 1 to 100 in PNL and from 0 to 39 in SNL. Except in the one case noted earlier, parental report was not used to determine a child's primary language for numbers. This is because children frequently encounter number words in formal classroom settings where instruction is often conducted in a language that is not spoken by the parent.

## Results

Figures 1 and 2 reveal the mean performance on the counting task in each language at each knower level separated by FE speakers and SE speakers. As expected, children's familiarity with the count list as reflected in the highest number recited increased as their comprehension of number words progressed. Preliminary analyses revealed no significant difference in performance between FE and SE children. As a result, all analyses are collapsed across languages.


Figure 1: The mean performance on the counting task by knower level for French-English speakers.


Figure 2: The mean performance on the counting task by knower level for Spanish-English speakers.

To examine the relationship between children's comprehension of numbers, children's familiarity and exposure to number words, and children's general maturation, we conducted Spearman's correlations between knower level, highest count, and age. Not surprising, SNL knower level was significantly correlated with PNL knower level, $\rho=0.81, p<0.01$, SNL counting, $\rho=0.72, p<0.01$, and age, $\rho=0.67=, p<0.01$, indicating that children's understanding of numbers deepened as a function of their familiarity with the count list and their general cognitive development.

To isolate the individual effects of familiarity with the count list on the one hand and comprehension of numbers on the other hand, we conducted a logistic regression to predict children's knower level in SNL using SNL counting, PNL knower level, and age as predictors. The full model, when compared against a constant only model, significantly predicted SNL knower level, indicating that the predictors as a set reliably differentiated children's knower level in SNL, $r^{2}(U)=0.45, \chi^{2}(6)=85.8, p<0.01$, with a misclassification rate of 0.27 . A likelihood ratio test further revealed a main effect of PNL knower level, $\chi^{2}(4)=24.07, p$ $<0.01$, suggesting that children's comprehension of numbers in PNL significantly predicted children's comprehension of numbers in SNL, perhaps through conceptual transfer (see Figure 3). However, the effect was restricted to two developmental transitions, as reflected in the parameter estimates: a transition from a 1 -knower to a 2 knower, $\beta=2.91, \beta(\mathrm{SE})=1.13 ; \chi^{2}(1)=5.20, p=0.02$, and a transition from a subset knower to a CP-knower, $\beta=2.41$, $\beta(\mathrm{SE})=0.91 ; \chi^{2}(1)=3.51, p=0.06$. There were also effects of SNL counting, $\chi^{2}(1)=12.05, p<0.001$ and age, $\chi^{2}(1)=$ $3.90, p=0.05$.


Figure 3: The percentage of children at each SNL knower level by PNL knower level.

## Discussion

We asked how transfer of number concepts from a first language and familiarity with numbers in a second language may facilitate the acquisition of number words in the second language. We found that familiarity with number words
facilitated the acquisition of number concepts and that when children acquired the meaning of number words in their primary language that knowledge transferred to their secondary language during two stages of conceptual development. Transfer was seen when children became twoknowers and when they became CP-knowers.

One possible mechanism explaining the effect of transfer at the level of CP-knowers is a process of analogical reasoning that results in the mapping of analogous concepts across languages. As CP-knowers, children not only learn the meanings of numbers greater than four but also recognize the unique relationship between counting and cardinality. More specifically, CP-knowers learn that counting can be used to label the cardinality of sets and eventually that each successive number is one greater than the preceding one (i.e., $\mathrm{N}+1$ ). Importantly, however, by recognizing that number words belong to a class that forms a structured list, children may infer that the lists in each of their languages operate according to the same principles. As a result, children who are CP-knowers may transfer their knowledge of counting from one list to the other through this process of analogical reasoning.

Another possible mechanism is that number word learning, in general, is a process of conceptual change in which new concepts, such as one, two, and three, are constructed. According to Carey (2009), prior to learning small number words, children cannot represent exact cardinalities via language (see also Le Corre \& Carey, 2007; Sarnecka \& Gelman, 2004; Wynn, 1990, 1992). Although infants can keep track of small numbers of individual objects (Feigenson \& Carey, 2005), and can represent the approximate cardinality of large sets (e.g., Xu \& Spelke, 2003), they may not be able to represent the precise numerosity of sets as a property distinct from the individuals themselves. On this view, number word learning is hard, in part, because it involves creating new conceptual resources. Consequently, once these resources have been built, learning the same meanings in a second language should be substantially easier - i.e., there should be "conceptual transfer." In this case, language transfer might also be observed in bilingual speakers during each of the subset stages of number word acquisition.

In contrast, we failed to find evidence of any transfer from the primary to the secondary number language when children become 1-knowers and 3-knowers. This supports an alternative view that children learn the meanings of these number words (three and one) as a function of their exposure to the number words in their secondary number language, indicating that they must employ a languagespecific mapping between words and meanings. That is, children may consistently hear the word "uno" in association with sets of one and, as a result, form direct mappings between the specific word, "uno," and the quantity, one. Without knowledge of the count list structure that CP-knowers are privy to, bilingual subset knowers are unable to draw comparisons between these numbers. Consequently in these two cases, children gained little
advantage in their SNL number knowledge by graduating to the next level of n-knower in PNL.

While transfer failed to occur at most subset knower levels, there was evidence that the transition from being a one-knower to a two-knower did transfer across languages despite these children's lack of knowledge of the count list structure. This particular transition may mark a significant, conceptual milestone in number word learning that can be transferred across languages. Whereas English, French, and Spanish use singular-plural morphology to mark the exact quantity of one (e.g., "a"), none of these languages use morphology to mark the exact cardinality of two (e.g. a dual marker like that used in Slovenian, Corbett, 2000). For this reason, there may be a conceptual barrier that children have to pass before they are able to map number words two or greater onto their corresponding quantities. Once this barrier is passed in one language, children are able to learn the words two and three in both languages, given that they have sufficient familiarity with number words in both languages. After passing this barrier, the transition from two-knower to three-knower is not transferred across languages, unlike the previous transition from one-knower to two-knower.

The singular-plural morphology in English, French and Spanish also explains why no transfer was seen when children become 1-knowers. The singular-plural morphology may facilitate a concept of the exact cardinality of one before children begin the number acquisition process (Barner, Libenson, Cheung, \& Takasaki, 2009).

In conclusion, we found that learning number words in one language facilitates the acquisition of the analogous number words in a second language at particular points in the number word acquisition process that are characterized by conceptual milestones. Although we suggest that relatively simple concepts, like counting, transfer across languages, it remains uncertain to what extent this occurs. Future studies should explore the generalizations and limitations of conceptual transfer by testing more advanced numerical abilities like estimation.

## Acknowledgments

We thank the members of the Language and Development Lab for scheduling participants, collecting data, and coding responses. This work was supported by the James S. McDonnell Foundation Scholar Award to DB.

## References

Carey, S. (2009). The Origin of Concepts. New York: Oxford University Press.
Corbett, G. G. (2000). Number. Cambridge: Cambridge University Press.
Barner, D., Libenson, A., Cheung, P., \& Takasaki, M. (2009). Cross-linguistic relations between quantifiers and numerals in language acquisition: Evidence from Japanese. Journal of Experimental Child Psychology 103, 421-440.

Davidson, K., Eng, K., \& Barner, D. (2012). Does learning to count involve a semantic induction? Cognition, 123, 162-173.
Dehaene, S. (1997). The Number Sense: How the Mind Creates Mathematics. Oxford University Press: New York.
Dehaene, S., Spelke, E., Pinel, P., Stanescu, R., \& Tsivkin, S. (1999). Sources of mathematical thinking: Behavioral and brain-imaging evidence. Science, 284, 970-974.
Feigenson, L., \& Carey, S. (2005). On the limits of infants' quantification of small object arrays. Cognition, 97, 295313.

Frye, D. Braisby, N. Love, J. Maroudas, C. Nicholls, J. (1989). Young children's understanding of counting and cardinality. Child Development, 60, 1158-1171.
Fuson, K. (1988). Children's counting and concepts of number. New York: Springer-Verlag.
Gelman, R. \& Gallistel, C. R. (1978). The child's understanding of number. Cambridge, Mass: Harvard University Press.
Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. Cognitive Science, 7, 155-170.
Gentner, D. (2003). Why we're so smart. In D. Gentner and S. Goldin-Meadows (Eds.), Language in mind: Advances in the study of language and thought (pp. 195-235). Cambridge, MA: MIT Press.
Gentner, D., \& Markman, A. B. (1997). Structure mapping in analogy and similarity. American Psychologist, 52, 4556.

Kolers, P. (1968). Bilinguals and information processing. Scientific American, 218, 78-86.
Le Corre, M. \& Carey, S. (2007). One, two, three, four, nothing more: An investigation of the conceptual sources of the counting principles. Cognition, 105, 395-438.
Lee, M. D. \& Sarnecka, B.W. (2011). Number-knower levels in young children: Insights from a Bayesian model. Cognition, 120, 391-402.
Marsh, L. G., \& Maki, R. H. (1976). Efficiency of arithmetic operations in bilinguals as a function of language. Memory and Cognition, 4, 459-464.
McClain, L., \& Huang, J. Y. S. (1982). Speed of simple arithmetic in bilinguals. Memory and Cognition, 10, 591596.

Piantadosi, S., Tenenbaum, J. \& Goodman, N. (2012) Bootstrapping in a language of thought: a formal model of numerical concept learning. Cognition, 123, 199-217.
Quine W. V. (1960). Word and Object. The Mit Press.
Rescorla, L. (1989). The Language Development Survey: A screening tool for delayed language in toddlers. Journal of Speech and Hear Disorders, 54, 587-599.
Sarnecka, B. W. \& Carey, S. (2008) How counting represents number: What children must learn and when they learn it. Cognition, 108, 662-674.
Sarnecka, B.W. \& Gelman, S.A. (2004). Six does not just mean a lot: Preschoolers see number words as specific. Cognition, 92, 329-352.

Spelke, E. S., \& Tsivkin, S. (2001). Language and number: A bilingual training study. Cognition, 78, 45-88.
Wynn, K (1990) Children's understanding of counting. Cognition, 36, 155-193.
Wynn, K. (1992). Children's acquisition of the number words and the counting system. Cognitive Psychology, 24, 220-251.
Xu, F., \& Spelke, E. S. (2000). Large number discrimination in 6-month-old infants. Cognition, 74, B1-B11.

# The Effects of Personality in a Social Context 

Kirsty Kitto (kirsty.kitto@qut.edu.au)<br>Information Systems School, Queensland University of Technology<br>2 George Street, Brisbane, 4000, Australia.

Fabio Boschetti (fabio.boschetti@csiro.au)<br>Marine Research, Commonwealth Scientific and Industrial Research Organisation (CSIRO)<br>School of Earth and Geographical Sciences, The University of Western Australia<br>Private Bag 5, Wembley, 6913, Australia


#### Abstract

The contextuality of changing attitudes makes them extremely difficult to model. This paper scales up Quantum Decision Theory (QDT) to a social setting, using it to model the manner in which social contexts can interact with the process of low elaboration attitude change. The elements of this extended theory are presented, along with a proof of concept computational implementation in a low dimensional subspace. This model suggests that a society's understanding of social issues will settle down into a static or frozen configuration unless that society consists of a range of individuals with varying personality types and norms. Keywords: contextual models; quantum decision theory; attitude change; agent based modelling


## Modelling Attitude Change in a Social Context

The ability to model and predict human responses to changing social conditions is fast becoming highly desirable in a world facing a number of global challenges. This social behaviour is frequently driven by their internally held attitudes of the individuals in a society (Ajzen, 2005; Fazio \& Petty, 2008). For example, privately held attitudes play a critical role in people's personal choices about their health, education, social groups, and housing, as well as the importance they attribute to national issues such as the environment, immigration and state security. However, attitudes are highly contextual, and this makes them extremely difficult to model formally. People's attitudes are not static immutable objects, but change in response to persuasion (Seiter \& Gass, 2010), and the attempt to maintain cognitive consistency (Cooper, 2007). We often express different attitudes and opinions in accordance with the social scenario we find ourselves in (Bond \& Smith, 1996), and it is frequently the case that an explicitly expressed attitude is quite different from an internally held one (Greenwald \& Banaji, 1995).

The Elaboration Likelihood Model (ELM) (Petty \& Cacioppo, 1986); and the Heuristic-Systematic Model (HSM) (Chaiken, 1987) are the two traditional models of attitude change, but both depend upon a number of poorly defined variables, which led Mosler, Schwarz, Ammann, and Gutscher (2001) to create a computational model of attitude change in order force a more accurate specification of the largely heuristic ELM. In essence, both models posit that some processes of attitude change require relatively high amounts of mental effort, resulting from situations where individuals are motivated to pay attention to a message, or have the cognitive capacities to consider it carefully. In these high
effort or high elaboration processes, people's attitudes will be determined by an effortful examination of all relevant information, and so changing them will expend high amounts of cognitive energy. In contrast, other low effort or low elaboration processes of persuasion require relatively little mental consideration by the persuadee, resulting in attitudes determined by factors like emotions, 'gut feeling', liking, and reference to authority.

There are few analytical models capable of describing the dynamics of low elaboration attitude change. While high elaboration processes are more logical and considered, hence frequently following processes similar to first order logic, low elaboration processes are more difficult to control, and are frequently more open to subtle social influences. While it must be acknowledged that involuntary factors such as disgust can play a very important role in low elaboration attitude change (Griskevicius et al., 2013; Rozin, Haidt, \& McCauley, 2008), these responses are themselves often mandated by previous social conditioning. It is very difficult to separate low elaboration attitude change from the social context (both current and historical) in which it occurs.

Furthermore, the underlying personalities of individuals in a society can reveal stark differences in how they will respond to their social context. For example, the Asch conformity experiments (Bond \& Smith, 1996), while not directly applying to attitude change, revealed stark differences in the conformity of subjects when responding to a group of confederates who had been instructed to lie about a perceptual task. While a control group of subjects who performed the same task alone revealed an error rate of less than $1 \%, 75 \%$ of the experimental group of subjects gave an incorrect answer to at least one perceptual task. These incorrect responses often matched those of the lying confederate group. Interestingly, by performing post task interviews, Asch established that there was a wide range of individual responses to these tasks. Some individuals reacted confidently to their individual perceptions, whereas others became more withdrawn and hesitant. Some yielded easily to the group decisions, even to the point of actively believing that the group answer was the correct one. This suggests that the underlying personality of the subjects was a key factor affecting their likelihood of conforming with the group, or truly reporting their differing perceptual observations.

In this paper, we shall introduce a dynamical model of
low elaboration attitude change, showing how it is possible to mathematically represent the manner in which the social context of an agent can affect their expressed attitudes. The model uses a cognitive state to represent an attitude, but is non-deterministic, with the probability of an agent acting taken to depend not just on this state, but also on: (1) the social context in which an agent finds themselves; and, (2) their underlying personality. A simple computational implementation will be discussed, and the way in which agent personalities affect individual attitude changes, and in turn affect the dynamics of the society as a whole will be explored.

## Modelling Decisions in a Social Context

Our model takes Quantum Decision Theory (QDT) (Busemeyer, Pothos, Franco, \& Trueblood, 2011; Busemeyer \& Bruza, 2012) as its starting point, due to its implicit capacity to represent the effect of context upon a decision. QDT has been shown capable of providing a unified explanation for many of the so called 'violations' of rational decision theory that are exhibited by individual humans, and so offers a promising new approach to the modelling of human decision making in context. A recent work by the authors (Kitto \& Boschetti, 2013) has introduced a social extension of the basic QDT model. It proposes a mechanism by which a society of agents self-organises into a set of ideologies representing their combined, and often contradictory, attitudes towards a social issue. This section will briefly introduce that model, but full details can be found in the longer paper.

## The Basic QDT

We shall begin with a consideration of an agent $A$, called Alice, who is deciding whether or not to 'act' in response to a given social issue. Recognising that $A$ 's decision is likely to depend upon their social context, we shall represent her cognitive state as a vector $|A\rangle$ in a vector space, ${ }^{1}$ the structure of which will depend upon the nature of the issue under consideration. If $A$ has decided to act on this issue, then we shall denote this state of action as the vector $|1\rangle$, to represent a situation where it is true that she has chosen to act (in contrast to a state of inaction which we denote as $|0\rangle$ ).

These decisions only make sense with respect to a particular social context, and the probability of $A$ acting could change with a new social setting. However, the quantum formalism can easily incorporate this contextuality due to its vectorial representation of the state $|A\rangle$. Thus, QDT represents the cognitive state of Alice, defined with respect to the

[^307]

Figure 1: An agent attempts to decide whether or not to act. (a) Their probability of action is proportional to the length squared of the projection of their state onto the axes labelled $\left|0_{p}\right\rangle$ (no action) and $\left|1_{p}\right\rangle$ (action); (b) The changing context of a decision. The probability of the agent acting changes between the two depicted contexts, which can immediately be seen by the different lengths of the projections from the state $|A\rangle$ onto the two different 'act' axes $\left|1_{p}\right\rangle$ and $\left|1_{q}\right\rangle$.
context $p$ as

$$
\begin{equation*}
|A\rangle=a_{0}\left|0_{p}\right\rangle+a_{1}\left|1_{p}\right\rangle, \text { where }\left|a_{0}\right|^{2}+\left|a_{1}\right|^{2}=1 \tag{1}
\end{equation*}
$$

a situation that is illustrated in Figure 1(a). Pythagoras theorem is used to extract the probabilities of $A$ acting (or not) in this context, with the probability of action given by $\left|a_{1}\right|^{2}$ and that of inaction similarly given by $\left|a_{0}\right|^{2}$. Thus, the projection of the state $|A\rangle$ onto the current context decides the probabilities of action for this model (Isham, 1995).

With reference to Figure 1(a), we see that in the context $p$ Alice is genuinely undecided. The cognitive state $|A\rangle$ represents an agent who has yet to decide how to act within some context, in contrast to the more standard modelling scenario where the agent has decided how to act, but we as modellers do not know what that decision is. Thus, the probabilities that arise in this model are fundamentally different from those of the more standard Kolmogorovian approaches (both Bayesian and frequentist), and this difference can have a profound effect with a change in context.

This can be seen with a consideration of figure 1(b), which is an elaboration of figure 1(a), and represents the changing probabilities of action that arise in the case of two different contexts, $p$ and $q$. With reference to figure $1(\mathrm{~b})$ we can quickly see that while our agent is highly likely to act in context $q$, this is not the case in context $p$, where $A$ is much less likely to act (since by examination of the figure we can see that while $\left|a_{0}\right|>\left|a_{1}\right|$ in context $p,\left|b_{1}\right|>\left|b_{0}\right|$ in context $q$ ).

## Social Framings of an Issue

This simple model can be naturally extended across a set of multiple agents which we shall call a society $\{|A\rangle,|B\rangle,|C\rangle \ldots\}$, all of whom are considering an issue, where each individual agent $X$ is described with a cognitive state $|X\rangle$ which is expected to change in time.

We assume that agents can make decisions to act within one of two contexts, which we denote as local, and global.

This is taken to represent the manner in which, while we frequently make internal or private decisions (as represented by a local frame), we must sometimes cast our choices within a societal domain (as represented by a global frame) when for example, we are required to vote in a general election. The local frames of the individuals in a society might be similar to a global understanding, or they might differ substantially, depending upon the agent and how they think about the world. Local frames might arise from a wide range of both external and internal factors, such as the socioeconomic status of an agent, their educational background, race etc. and so are likely to be highly complex, and multidependent variables. As a first approximation we shall model them as another basis in the two dimensional vector space already introduced for the states and global frame. This allows us to anticipate that global frames will result from an aggregation function applied to the local frames of every agent who somehow identifies with that ideology. At this point in time, we define identification by performing a distance measure; the global frame that most closely aligns with the local frame of the agent is the one to which the agent is deemed to belong. However, we note that this identification is not intrinsic to the theoretical model per se, rather it is expected to evolve as the model is applied to different social scenarios, and extended into a higher dimensional state space than the early 2D implementation discussed below. We currently use clustering for the definition of global frames via aggregation, but we anticipate that there are many potential methods for defining global frames, and that different ones will prove necessary for different issues (List, 2012)

Kitto and Boschetti (2013) claimed that this framework provides an opportunity to model low elaboration processes of attitude change nontrivially, due to its explicit recognition of the context in which an agent makes a decision. The QDT approach allows for the probability of an agent acting to vary over the full range $(0,1)$ in response to the range of angles that can be taken by the cognitive state of the agent within the Hilbert space that represents the issue currently under consideration. Thus, in order to evaluate Alice's probability of acting, we must take both her current cognitive state $|A\rangle$, and her current social context $p$ (as represented by a global or local frame) into account.

We postulate that an agent who has made a decision is likely to feel a certain amount of cognitive dissonance (Cooper, 2007) as their internal cognitive state will not be aligned with their decision (unless their cognitive state was already aligned with the relevant frame from which they are currently considering an issue). This means that they will feel a certain amount of psychological discomfort, which will drive them to alter their view of the world to fit with their decision within the context that it was made. They can do this in the current model by adjusting either their cognitive state, or their local framing of the issue, to more accurately reflect their decision. However, the literature suggests that some people are more comfortable with cognitive dissonance than others; their personalities will therefore play a
key role in how this adjustment occurs. For example, some agents will feel far less comfortable with uncertainty than others, and so be more affected by dissonance (Sorrentino \& Roney, 2000; Sorrentino \& Hewitt, 1984). In order to model these intuitions, we note that an agent whose cognitive state lies close to the axes representing their current frame will be more certain about their likely future actions than one whose cognitive state lies between those axes (i.e. has the cognitive state forms a $45^{\circ}$ angle between choosing to act and choosing not to act in the frame $p$ ). This leads us to introduce a measure of the uncertainty that an agent experiences about their likely future decisions, using binary entropy $H_{b}(P) \equiv-P \log _{2} P-(1-P) \log _{2}(1-P)$, which is a function taking its minimum values at $P=0$ and $P=1$, and its maximum at $P=1 / 2$. Here, the probability $P$ is defined with reference to the probability of the agent acting (or not) within the given context. Referring to Figure 1(a), we can rewrite the binary entropy of our agent within the context $p$ using a set of geometric variables

$$
\begin{equation*}
H_{b}(P(\theta))=-\left|a_{1}\right|^{2} \cdot \log _{2}\left(\left|a_{1}\right|^{2}\right)-\left|a_{0}\right|^{2} \cdot \log _{2}\left(\left|a_{0}\right|^{2}\right) \tag{2}
\end{equation*}
$$

where $\theta$ is the angle between the $\left|1_{p}\right\rangle$ basis state and the state of the agent $|A\rangle$. This entropy measure is then used in a model of the two different drives for cognitive consistency that we hypothesise are experienced by an agent making a decision in a social context:

1. A desire for internal cognitive consistency. This drives agents to align their cognitive state with the local frame within which they are currently considering an issue.
2. A desire to 'fit in' with the society and its current norms. This desire is expressed by a pull of agent's local frame towards the current global frame (or ideology) to which they belong, which serves to reframe their understanding of the issue.

Defining $\Theta$ as the angle between the agent's current state $|A\rangle$ and the decision to act in the global frame to which they currently belong we introduce a function which measures the uncertainty of the agent $A$ with respect to both frames:

$$
\begin{equation*}
H(|A\rangle, \theta, \Theta)=w_{i}(A) H_{b}(P(\theta))+w_{s}(A) H_{b}(P(\Theta)) \tag{3}
\end{equation*}
$$

where the weights $w_{i}(A)$ and $w_{s}(A)$ refer to agent $A$ 's need for internal consistency and social conformity respectively. These weights can be set to range over a population of agents, indicating a rough parameterisation of a society's social make-up. This measure can naturally be extended to consider the uncertainty of the whole society of $N$ agents:

$$
\begin{align*}
H_{b_{\text {Tot }}} & =\sum_{i=1}^{N} H\left(|i\rangle, \theta_{i}, \Theta_{i}\right)  \tag{4}\\
& =\sum_{i=1}^{N}\left[w_{i}(i) H_{b}(P(|i\rangle, \theta))+w_{s}(i) H_{b}(P(|i\rangle, \Theta))\right] \tag{5}
\end{align*}
$$

which should decrease as the agents achieve cognitive consistency and so settle into a set of stable ideologies, or global attitudes about the world.

## Time Evolution

The weights $w_{i}(A)$ and $w_{s}(A)$ can be considered as personality variables, and they will affect each agent's future actions, in addition to their current cognitive comfort (as is represented by (3)). At present, we update agent states and local frames slightly differently according to the frame in which the decision was initially made.

Local Decisions If the decision was in the local frame, then only the cognitive state of the agent is updated (within the local frame). Thus, an agent who has chosen to act within a certain framing of a problem will shift their state towards the decision ('yes' or 'no') that they made in that context. The size of this shift is defined as dependent upon two factors: (1) the personality profile of the agent (given in this case as $w_{i}$, as it represents the desire of an agent to align their cognitive state with their local frame); (2) the angle $\theta$. Writing $\theta_{0}$ for the angle between the agent's state and the $\left|0_{p}\right\rangle$ axis, and $\theta_{1}$ for the angle between their state and the $\left|1_{p}\right\rangle$ axis, the new angle between the agent's state and the frame will become:
if $A$ decides $\left\{\begin{array}{l}\text { to act: } \theta_{1}\left(|A\rangle_{t+1}, w(A)\right)=\theta_{1}\left(\left|A_{t}\right\rangle\right) \times w(A) \\ \text { not to act: } \theta_{0}\left(|A\rangle_{t+1}, w(A)\right)=\theta_{0}\left(\left|A_{t}\right\rangle\right) \times w(A)\end{array}\right.$
where $w(A)$ depends upon the comfort of $A$ with holding an attitude that is dissonant from their decision. Thus, for this update process $w(A)=w_{i}(A)$. Agents who decide to act will thus experience a rotation of their cognitive state by a certain distance dependent upon their personality towards the $\left|1_{p}\right\rangle$ axis (recall that $\theta$ is the distance between the $\left|1_{p}\right\rangle$ axis and the current state of the agent $|A\rangle$ ), and agents who decide not to act will experience a rotation of their cognitive state in the opposite direction.

Global Decisions If the decision was made in the global frame, then both the cognitive state of the agent and their local frame are updated (with reference to their global frame). Thus, in addition to the update of the cognitive state that is represented in equation (6), the local frame of the agent will shift towards the global axis that represents the decision made by the agent. The amount by which the local frame shifts is given by an equivalent version of equation (6), thus the new angle between the local frame and the global frame is given by (6), but with $w(A)=w_{s}(A)$.

## Implementation

A proof of concept model has been implemented in MATLAB, which allows for an investigation of the timewise behaviour of this new agent based modelling paradigm. Space does not permit a full explanation of this implementation, however, we direct the interested reader towards the actual MATLAB script ${ }^{2}$ which implements the basic pseudocode shown in Figure 2.

While the model that we have presented is admittedly very simple, it does exhibit a number of key features which one

[^308]```
Number of global frames \(=G\)
Number of agents = N
For \(i=1\)..N
    Assign coherence \& consistency variables
    If RandomPersonality \(=0\) then
        conformity \(=0.5\) and consistency \(=0.5\)
    If RandomPersonality \(=1\) then
        consistency \& conformity range over [0-1]
    Assign cognitive states \& local frames randomly
For each timestep
    Find the position of the global frames (use k-means)
    For each agent
        Calculate which global frame the agent belongs to
        Probabilistically choose to act or not in one frame
        If acting in local frame then update cognitive state
        If acting in global frame then update cognitive
        state and local frame
        Calculate entropy of the agent
        Calculate total entropy of system
```

Figure 2: Pseudocode for the computational implementation.
could reasonably expect should be found in an agent based model of attitude change. For example, Kitto and Boschetti (2013) describes the manner in which a population selforganises into a set of ideologies, which evolve and update in time. As predicted, the entropy (4) has a tendency to decrease in time. It is also possible to guide the behaviour of the population, through shifting a global frame, and to then watch the system reorganise into a new semi-stable configuration. In this paper we shall instead focus upon one key feature that has not yet been described, namely, the importance of personality in driving the attitude changes of a society of individuals.

## The Importance of a Personality Spread

Two different seeding strategies have been utilised to initialise the consistency and conformity parameters ( $w_{i}(A)$ and $w_{s}(A)$ ) for each agent within the computational model. A random distribution is possible, where each agent is seeded with parameters that randomly range from 0 to 1 , or alternatively all agents can be seeded with a fixed personality distribution. This allows for an investigation of the effect that varying personality spreads can have upon a population.

Random Personality When agents are seeded with a random personality mix the time evolution of the system is predictably at its most erratic. While the entropy of the system has a tendency to decrease throughout a run, the agents tend not to find a stable configuration, and the system remains in a state of flux and change; states, local, and global frames can all move throughout a run.

Figure 3 shows a set of shots from a typical run for this scenario, along with the entropy plot as it gradually decreases through time, subject to some stochastic variance as agents realign their local frames. Two global frames were specified, and their location at each timestep found using a k -means style algorithm. Agent's cognitive states are represented using black lines, global frames by the large dots above the cog-
nitive states, and local $|1\rangle$ frames as small black spots.


Figure 3: A typical run of a system initialised with agents of random personality spread. Note that the entropy of the system has a tendency to decrease in time, but that it never fully minimises or stablises.

Figure 4 shows a collection of entropy plots for two, three, and four global frames, all initialised with a random mix of personality parameters. Note that in all cases the entropy decreases, but that the system shows more erratic behaviour when more global points of view are available for the agents to align with. The limited nature of the current computational implementation (which has only been performed for two dimensions) means that arbitrarily adding more frames to what is a very small space does not result in realistic behaviour, however, work is currently in progress to extend this model to a higher dimensional state space, and this would allow for the interaction of far more social contexts to be investigated.

Fixed Personality In contrast, when the personality mix of the agents is fixed at $w_{i}(A)=0.5, w_{s}(A)=0.5$ the system


Figure 4: A collection of entropy plots for differing numbers of global frames, initialised with a random mix of personality.
exhibits a far more stable time evolution pattern, and becomes fixed in a static configuration around timestep 25. Figure 5 shows a typical run for this scenario. Note that the entropy minimises very early during a run, as the agents settle into a stable scenario that does not need to re-adjust. All agents can find a state and local frame that minimises (3), and the system rapidly settles down. This dynamics is also evident for for higher numbers of global frames.


Figure 5: A typical run of a system seeded with a population of fixed personality type. (In this case $w_{i}(A)=0.5$ and $w_{s}(A)=0.5$.) The system quickly stabilises into a configuration where all agents are of one, or the other, state of mind. This behaviour is observed for all fixed personality profiles.

## Evolution Requires Consistency and Cohesion

This brief discussion highlights the need for a society to contain a range of personality types. A society of individuals who all have the same personality mix quickly becomes static in this model, it settles down into a scenario where the attitudes of the agents, and their framing of those attitudes, do not change in time. This situation becomes even more dramatic when the society is seeded with individuals who have nonzero values only for conformity or for consistency. In both of these scenarios the model does not evolve at all, it remains in the same state as the one that it was initialised in.
This behaviour plausibly reflects the behaviour of societies in general. Difference of opinion and a varying response to the social context are both key and essential features of a society, and yet such behaviour does not tend to be well captured by current modelling technology. Thus, the contextualised apparatus of QDT offers an interesting new perspective on the modelling of social behaviour that we feel holds promise for future extension and expansion to a more realistic set of scenarios.

## Conclusions

We begin our conclusion with something of a caveat. The model presented here does not utilise the standard complex Hilbert space of quantum theory, nor even the standard Schrödinger based time evolution equation of that model. Indeed, while the dynamics of equation (6) are unitary (Isham, 1995), they are not a part of the standard formalism of quantum theory. We propose that QDT is the first of a class of contextual models of human cognition, but do not expect that a straightforward application of the quantum formalism will suffice to model every contextually dependent cognitive system.
The geometric nature of the model presented here provides a dramatic departure from more standard state based modelling methodologies. In particular, the interaction between the cognitive state of the agent $(|A\rangle)$ and of the basis in which they choose to make their decision (as represented by the basis $\{|0\rangle,|1\rangle\}$ ) means that in a different social context, the agent is highly likely to make a different decision as to how to act. Thus, in adopting a framework inspired by QDT, a very new approach to the treatment of context has been obtained. Furthermore, as the model presented here is developed, we anticipate that it will become necessary to progress to a complex space in order to represent the full range of personality variables and their associated cognitive states. In particular, the interference effects that are apparent in QDT, are not implemented in the current simple form of this model. In summary, it is the contextuality of human decision making in a social context that is captured by this model, but more cognitive effects are likely to be possible within this framework.
However, this initial step is important. Uncertainty dominates in scenarios where contextuality arises, but it is a cognitive effect apparent in the minds of the agents themselves, not in that of the modeller (Payne, Bettman, \& Schkade, 1999), and this is not well captured by our current probabilistic approaches. We have shown one viable approach towards capturing contextual social effects, based upon QDT. A proof of concept computational model was discussed, and a set of varying personalities was shown to be essential for the dynamical evolution of the model. Thus, a way forwards presents, and future work will seek to develop this exciting new approach.

## Acknowledgments

KK was supported by an Australian Research Council Post Doctoral Fellowship (DP1094974) during her completion of this work. The authors would like to acknowledge many helpful discussions with Peter Bruza and Iain Walker.

## References

Ajzen, I. (2005). Attitudes, personality and behaviour (2nd ed.). Open University Press.
Bond, R., \& Smith, P. B. (1996). Culture and Conformity: A Meta-Analysis of Studies Using Asch’s (1952b, 1956)

Line Judgment Task. Psychological Bulletin, 119, 111137.

Busemeyer, J., \& Bruza, P. (2012). Quantum models of cognition and decision. Cambridge University Press.
Busemeyer, J., Pothos, E., Franco, R., \& Trueblood, J. (2011). A Quantum Theoretical Explanation for Probability Judgment Errors. Psychological Review, 118(2), 193218.

Chaiken, S. M. (1987). The heuristic model of persuasion. In P. Zanna, J. M. Olson, \& C. P. Herman (Eds.), Social influence: The Ontario symposium (Vol. 5, p. 3-39). Hillsdale, NJ: Erlbaum.
Cooper, J. (2007). Cognitive dissonance: 50 years of a classic theory. Sage.
Fazio, R., \& Petty, R. (Eds.). (2008). Attitudes: Their structure, function, and consequences. Psychology Press.
Greenwald, A. G., \& Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. Psychological Review, 102, 4-27.
Griskevicius, V., Ackerman, J. M., Cantú, S. M., Delton, A. W., Robertson, T. E., Simpson, J. A., et al. (2013). When the economy falters, do people spend or save? responses to resource scarcity depend on childhood environments. Psychological Science, 24(2), 197-205.
Isham, C. J. (1995). Lectures on Quantum Theory. London: Imperial College Press.
Kitto, K., \& Boschetti, F. (2013). Attitudes, ideologies and self-organisation: Information load minimisation in multiagent decision making. Advances in Complex Systems.
List, C. (2012). The theory of judgment aggregation: An introductory review. Synthese, 187(1), 179-207.
Mosler, H.-J., Schwarz, K., Ammann, F., \& Gutscher, H. (2001). Computer Simulation as a Method of Further Developing a Theory: Simulating the Elaboration Likelihood Model. Personality and Social Psychology Review, 5, 201215.

Payne, J. W., Bettman, J. R., \& Schkade, D. A. (1999). Measuring constructed preferences: Towards a building code. Journal of Risk and Uncertainty, 19(1), 243-270.
Petty, R. E., \& Cacioppo, J. T. (1986). Communication and Persuasion: Central and Peripheral Routes to Attitude Change. New York: Springer-Verlag.
Rozin, P., Haidt, J., \& McCauley, C. R. (2008). Disgust. In M. Lewis, J. M. Haviland-Jones, \& L. F. Barrett (Eds.), Handbook of emotions (3rd ed., p. 757-776). New York: Guilford Press.
Seiter, R. H., \& Gass, J. S. (2010). Persuasion, social influence, and compliance gaining (4th ed.). Boston: Allyn \& Bacon.
Sorrentino, R., \& Roney, C. (2000). Uncertain mind: Individual differences in facing the unknown. Psychology Press.
Sorrentino, R. M., \& Hewitt, E. C. (1984). The uncertaintyreducing properties of achievement tasks revisited. Journal of Personality and Social Psychology, 47(4), 884-899.

## Comprehension cueing strategies in elderly: a window into cognitive decline?


#### Abstract

Language abilities gradually decline as we age, but the mechanisms of this decline are not well understood. The present study investigated comprehension of subject vs. object who and which direct questions (DQs), embedded questions (EQs) and relative clauses (RCs) in 39 cognitively healthy native speakers of Spanish. The elderly participants (n $=21$ ) were further classified according to their scores on a general cognitive test, Montreal Cognitive Assessment (MoCA), into a group with low MoCA scores, LM ( $\mathrm{n}=10$ ), and a group with normal MoCA scores, NM ( $\mathrm{n}=11$ ). A mixed-model, repeated-measures analysis of variance (ANOVA) showed that the elderly participants achieved significantly worse accuracy and speed than the young participants (Y) in all tasks. Accuracy was significantly lower and reaction times significantly longer in the LM group compared to the NM group in DQs and RCs. Accuracy in comprehension of EQs was also worse in LM compared to NM, with no significant difference in RTs between the two groups. The results are explained within the competition model and reliance on a language-specific cueing strategy. Reliance on cueing strategies in sentence comprehension may be an effective indicator of cognitive decline associated with aging.


Keywords: comprehension; wh-dependencies; aging.

## Introduction

Cognitive aging is typically associated with a decline in speed of processing and deterioration of memory and attention (Salthouse, 2009). Language abilities also gradually decline as we age, which is reflected in decreased vocabulary, smaller mean number of clauses per utterance, simplified syntactic structure of produced sentences, reliance on optimization strategies when choosing referring expressions as well as difficulty in comprehension of complex sentences (Kemper, Thompson \& Marquis, 2001; Grossman, Cooke, De Vita, Chen, Moore et al., 2002; Hendriks, Englert, Wubs \& Hoeks, 2008). Older adults’ language comprehension decline appears to be due not to sensory, but cognitive demands of spoken language, with complex syntax slowing down the comprehension even when sentence understanding is accurate (Tun, Benichov \& Wingfield, 2010). Research on English has shown that comprehension of structures that require a syntactic operation of movement and involve a longer gap between a moved element and its trace $(t)$, such as object relative clauses (e.g., The cat ${ }_{i}$ that the dog chased $t_{i}$ is black), is impaired in elderly adults, while comprehension of subject relative clauses, in which this gap is smaller (e.g., The cat ${ }_{i}$ that $t_{i}$ chased the dog is black), is spared (e.g., Zurif, Swinney, Prather, Wingfield \& Brownell, 1995; StineMorrow, Ryan \& Leonard, 2000). One explanation of this finding is that the object relative clauses require allocation of more working memory (WM) resources than subject
relative clauses, and WM limitation is one of key features of cognitive aging (Zurif et al., 1995; Caplan \& Waters, 1999; Stine-Morrow et al., 2000; Grossman, Cooke, De Vita, Alsop, Detre et al., 2002).

Furthermore, neuroimaging research has shown that when processing complex sentences, healthy seniors compared to young participants show reduced activation in the core language areas (e.g., inferior frontal regions), while showing additional activation of some areas that are not considered the "core" sentence processing network as well as difference in the coherence of connectivity of the involved brain areas (Peelle, Troiani, Wingfield, \& Grossman, 2010; Tyler, Shafto, Randall, Wright, Marslen-Wilson et al., 2010). Activation of the brain regions that are not typically involved in language processing has been interpreted as an indicator of compensatory processes (Grossman et al., 2002; Wingfield \& Grossman, 2006; Tyler et al., 2010).

Better understanding of the earliest changes in typical cognitive aging is also an important step towards better understanding of the Alzheimer's disease (AD) continuum. Structural and metabolic changes in AD brain occur long before cognitive symptoms become apparent (Dubois et al., 2007, 2010; Sperling et al., 2011). Crucially, even small metabolic and structural alterations in the brain may affect the dynamics enabling cognitive function (Buckner, Snyder, Shannon, LaRossa, Sachs, et al., 2005). Thus, it is important to understand the brain's ability to engage alternate networks and rely on cognitive strategies compensating for a deteriorating cognitive function.

One goal of the present study was to determine whether elderly native speakers of Spanish rely on compensatory strategies in sentence comprehension. We chose to study comprehension of wh-structures (i.e., structures formed by wh-words, such as what, who, which, etc.): direct and embedded questions introduced by interrogative pronouns qué ("what, which") and quién ("who") and relative clauses introduced by que. Like in English, the distance between a moved element and its gap is longer in object than in subject $w h$-structures, as shown in (1-2):
(1) ¿Quién ${ }_{i} t_{i}$ comió una naranja?
(2) ¿A quién mordió $_{j}$ el perrito $t_{j} t_{i}$ ?

However, in Spanish preposition a marks object whquestions and therefore it could serve as a processing cue. Since it appears before the moved $w h$-word, it signals an object structure, allowing the parser to assign a temporary thematic role before encountering the gap. Thus, reliance on this cue would facilitate comprehension of object structures, resulting in their good comprehension, even though they are syntactically more difficult than subject structures and require more WM resources.

## 2. Present Study

### 2.1 Participants

We tested 39 neurologically healthy native speakers of Spanish, of which 21 were older and 18 were young persons. There was a statistically significant difference in age between the groups $(t(27)=28.457, p<0.05)$ and years of education ( $t$ (30) = 6.76, $p<0.05$ ), but not in gender distribution $\left(\chi^{2}(1)=1.857, p=0.17\right)$.

The group of elderly was divided into two subgroups, based on their MoCA scores: since the scores lower than 26 indicate mild cognitive impairment (MCI) (Chertkow, Massoud, Nasreddine, Belleville, Joanette, et al., 2008), we used a cut-off score of 26 to dichotomize the elderly participants into a Normal MoCA scores group (NM) ( $\geq 26$ ) and a Low MoCA group (LM) (<26). Comparing the age means of the latter two groups revealed that the LM group ( $73.8 \pm 6.25$ ) was significantly older than the NM group ( $66.45 \pm 5.14$ ): $t(19)=2.95, p=0.008$ ). The two groups did not differ significantly in years of education $(t)=6.14, p$ $=0.54)$ or in gender distribution $\left(\chi^{2}(1)=0.064, p=0.8\right)$.

All participants were healthy, with no history of stroke, neurological disorders, alcohol/drug abuse, or other conditions that could affect cognition. They all reported normal hearing, and normal/ corrected to normal vision. All participants were recruited through the Ingema Foundation. Participants' characteristics are summarized in Table 1.

### 2.2 Evaluative measures

In addition to a test of global cognition (MoCA), we administered the Month Ordering Test to assess verbal WM (VWM). This test assesses storage and manipulation of material with semantic content, i.e., names of the months in calendar, which makes it highly relevant for studies of sentence comprehension (Almor et al., 2001; Goral, ClarkCotton, Spiro, Obler, Verkuilen et al., 2011). The months are given in a non-canonical order and participants' task is to repeat them canonically. There are 20 strings of months in total, distributed across 5 levels, with 4 strings at each level, containing a different number of months to order. Each correctly ordered string is scored as one point. Thus, the total possible score is 20 . Participants' scores on evaluative measures are summarized together with their demographic characteristics in Table 1.

Table 1: Participants characteristics.

|  | Elderly |  | Young |
| :--- | :--- | :--- | :--- |
|  | $\mathbf{L M}(\mathbf{n}=\mathbf{1 0})$ | NM $(\mathbf{n}=\mathbf{1 1})$ | $\mathbf{Y}(\mathbf{n}=\mathbf{1 8})$ |
| Age | $73.8 \pm 6.2$ | $66.45 \pm 5.1$ | $24.6 \pm 2.6$ |
| Age range | $65-85$ | $60-78$ | $20-30$ |
| Gender $(\mathrm{m} / \mathrm{f})$ | $4 / 6$ | $5 / 6$ | $4 / 14$ |
| Education $(\mathrm{y})$ | $10.56 \pm 5$ | $11.25 \pm 3$ | $17.44 \pm 1.9$ |
| MoCA | $21.9 \pm 3.0$ | $27.42 \pm 2$ | $28.83 \pm 1.2$ |
| VWM | $10.78 \pm 2.4$ | $14.67 \pm 1.8$ | $15.17 \pm 2.1$ |

There were three experiments in the study. Experiment 1 tested comprehension of who and which NP direct questions (DQs) extracted from a subject vs. object position in a sentence. It contained 40 sentences: 20 who DQ (ten subject and ten object questions) and 20 which NP DQs (ten subject and ten objects questions). Each question was preceded by a declarative sentence describing a situation from everyday life, such as: Pablo is eating apples and Juan is eating oranges. Thus, for a subject position, a who question would be: Who is eating oranges? And a which-NP question would be: Which boy is eating oranges? The sentences were presented auditorily, and possible answers-Pablo, Juanappeared in a written form, on the left and right side of the computer screen, respectively. The participants indicated their responses by pressing the left or right arrow on the keyboard, depending on whether the correct answer was on the left or on the right side of the screen.

Experiment 2 tested comprehension of embedded questions (EQs). There were 80 EQs: 40 who and 40 which $N P$ questions, with 20 subject and 20 object questions within each group. Half of the questions ( $n=40$ ) contained one prepositional phrase (PP) and the other half contained two PPs. EQs were tested in a verification paradigm: participants were required to listen to a sentence, followed by a verification statement, and decide whether the statement was correct or incorrect relative to the sentence. The participants indicated their answers by pressing the left vs. right arrow on the keyboard, depending on whether "Correct" and "Incorrect" appeared on the left or right side of the computer screen.

In Experiment 3, we tested comprehension of relative clauses (RCs). There were 10 subject and 10 object RCs introduced by que. The tested structure was preceded by a simple sentence providing a context. As in Experiments 1 and 2, participants were required to indicate their answers by pressing the left or right keyboard arrow.

Sentence stimuli for each experiment were first randomized in Excel and then recorded in Audacity (http://audacity.sourceforge.net/). Prerecorded sentences were imported in the DMDX (www. http://www.u.arizona.edu/~kforster/dmdx) and presented auditorily over a PC computer and a set of speakerphones.

### 2.4 Procedures

Participants were instructed to respond to a question as fast and as accurately as possible. The next sentence was initiated by the subject's response. The left and right arrow responses for correct answers were counter-balanced across conditions in each experiment. There was a time window of $5,000 \mathrm{msec}$ for answers. If the participant did not respond within that time, the answer options disappeared from the screen, and a fixation cross appeared, indicating that a new auditory stimuli was about to appear. A failure to respond within $5,000 \mathrm{msec}$ was scored as an error. There was a $30-$ second break after every 20 sentences. Feedback showing whether the answers were correct or incorrect was given on

### 2.3 Experimental measures

the computer screen only during the practice trials. There was no feedback during the actual testing.

Each session began with the experimenter describing the study, and the participant reading and signing the informed consent. After that, demographic details were collected and precise instructions on how to execute the experimental tasks were given. This was followed by the participant's taking 8 practice trials. After a satisfactory performance on the practice trials, the participants were tested on the experimental measures. Finally, MoCA and the Month Ordering Test to assess verbal WM (the VWM test henceforth) were administered.

All the materials were administered in the same order to each participant, except for the experimental stimuli, which were administered as two different randomizations, which were introduced to allow controlling for the effects of stimulus ordering. Testing was carried out in a quiet room at Ingema laboratory facilities in San Sebastián. It was conducted individually with each participant and completed in a single session, which lasted approximately 1 hour and 10 minutes. The study was conducted in accordance with the Declaration of Helsinki and was approved by the local Ethics' Committee.

## 3. Results

### 3.1 Evaluative Measures

There were statistically significant differences between the elderly group overall and the young participants on MoCA $(t(25)=4.431, p<0.0005)$ and VWM test $(t(36.5)=2.714$, $p=0.01$ ), indicating better performance of the younger compared to the older participants. Within the group of elderly participants, the NM group outperformed the LM group on both tests - MoCA: $t(19)=5.42, p=0.001$, and VWM: $t(19)=3.792, p=0.001$. There was a significant positive correlation between years of education and MoCA ( $\mathrm{r}=0.597, \mathrm{~N}=39, p<0.01$ ), and between MoCA and VWM scores ( $\mathrm{r}=0.611, \mathrm{~N}=39, p<0.01$ ). There was a significant negative correlation between age and MoCA scores ( $\mathrm{r}=-0.607, \mathrm{~N}=39, p<0.01$ ), and between age and VWM scores ( $\mathrm{r}=-0.495, \mathrm{~N}=39, p<0.001$ ). All tests were two-tailed.

### 3.2 Experimental Measures

Accuracy and RTs of understanding wh-dependencies were analyzed in a mixed-model, repeated-measures analysis of variance (ANOVA), with a between-subject factor comparing groups (LM, NM, Y) and within-subject factors comparing the extraction site in a sentence (subject/object) and the type of wh-word (who/which).
Experiment 1: Direct Questions. The accuracy analyses showed that the main effect of group was significant, i.e., there were statistically significant differences in the participants' overall sentence comprehension between the groups $(F(2,36)=30.421, p=0.001)$. The results of a posthoc Tukey test showed significant differences in
comprehension between the Y group and the LM group ( $p<$ 0.005 ), and between the NM group and the LM group ( $p<$ 0.005 ). In both comparisons, the LM group had lower accuracy. The difference between the Y and the NM groups was not significant ( $p=0.81$ ). The main effect of extraction site (subject/object) was significant $(F(1,36)=4.564, p<$ 0.04 ), reflecting better comprehension of object structures, whereas the main effect of $w h$-word was not significant $(F(1,36)=0.187, p=0.668)$.

The analysis of RTs also showed that the main effects of the group $(F(2,36)=37.844, p<0.001)$ and extraction site were significant $(F(1,36)=4.479, p=0.041)$, and so was the effect of the two-way interaction between the extraction site and group $(F(2,36)=3.593, p=0.038)$. Tukey test showed significant differences between the Y group and the NM group ( $p<0.005$ ), between the Y and LM groups ( $p<$ 0.005 ), and between the NM and the LM groups ( $p<$ 0.001 ), with the LM group reacting slower in both cases.

Experiment 2: Embedded Questions. Comprehension of EQs did not show a significant effect of extraction site (subject/object) $(F(1,36)=3.517, p=0.69)$. However, the main effect of wh-word (who/which) was significant $(F(1,36)=5.623, p=0.023)$, and so was the interaction between the extraction site and type of wh-word $(F(1,36)=$ 5.001, $p=0.032$ ). The type of $w h$-word also interacted with PP $(F(1,36)=5.454, p=0.025)$. There were significant differences in the participants' overall sentence comprehension $(F(2,36)=61.990, p<0.001)$. The results of Tukey test showed significant differences for every pair of groups (LM vs. Y: $p<0.005$; LM vs. NM: $p<0.005$; NM vs. Y: $p<0.001$ ), where the Y group was the most accurate, while the LM group was the least accurate. Since the lowest scores were achieved on tasks in Experiment 2, percent correct responses across all conditions are given in Table 2 as another view into the data.

Table 2: Percent correct responses on Experiment 2.

|  | Who | Who | Who | Who | Whi | Whi | Whi | Whi |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | S1PP | O1PP | S2PP | O2PP | S1PP | 01PP | S2PP | O2PP |
| LM | 60 | 51 | 57 | 58 | 56 | 65 | 55 | 45 |
| NM | 69 | 62 | 87 | 70 | 88 | 83 | 85 | 78 |
| Y | 94 | 93 | 97 | 92 | 98 | 94 | 98 | 94 |

Since questions in all experiments required a choice between two possibilities, $50 \%$ correct represented chance performance on all tasks, and scores between 26-75\% were considered to be within the range of chance. A score of 75\% or better was considered better than chance performance, while a score of $25 \%$ or below was taken to indicate a systematic reversal in the interpretation of a particular construction (there were no such scores in our data).

The analysis of the RT data has shown that the main effects of $w h$-word and extraction site were not significant, but the interaction between these two factors was significant $(F(1,36)=11.263, p=0.002)$. There was a significant effect of PP $(F(1,36)=22.369, p<0.001)$, and it interacted with
the group $(F(2,36)=4.139, p=0.024)$. A three-way interaction between PP, group, and $w h$-word was also significant $(F(2,36)=3.315, p=0.048)$. RTs differed significantly among the groups $(F(2,36)=14.049, p<$ 0.001 ), and the post-hoc Tukey test showed that the Y group was faster than the LM group ( $p=0.001$ ) and the NM group ( $p=0.031$ ). The difference in RTs between the latter two groups was also significant ( $p=0.05$ ).

Thus, the results of Experiment 2 indicate that embedded questions containing additional phrases such as PPs are in general difficult to process for cognitively healthy older adults, in particular to those with mildly affected general cognition and VWM.
Experiment 3: Relative Clauses. There were no significant within-subject effects in participants' comprehension of RCs. Neither the main effect of extraction site was significant $(F(1,36)=1.651, p=0.2)$ nor its interaction with the group $(F(1,36)=0.744, p=0.42)$. There were significant differences in the participants' overall comprehension of RC among the groups $(F(2,36)=9.662$, $p<0.001$ ). The results of Tukey test showed that there were significant differences in the comprehension between the Y and the LM groups ( $p=0.001$ ), and between the NM group and the LM group ( $p=0.024$ ). In both cases the LM group had lower accuracy. RTs differed significantly among the groups $(F(2,36)=26.784, p<0.001)$, and the Tukey test showed significant differences between every pair of groups: the Y group was the fastest, while the LM group was the slowest one.

### 3.3 Summary of results

Overall, lower accuracy and longer reaction times in comprehension of DQs and RCs were found in LM compared to NM participants. Comprehension of EQs was also worse in the LM group compared to the NM group (accuracy), but this was not associated with significant differences in RTs between the two groups. The Y group showed significantly better comprehension accuracy and speed in all tasks. Adding one or two PPs to the whstructures in EQs pushed the comprehension of the LM group to the chance level on all EQs, as well as comprehension of the NM group of 3 out of 4 types of who EQs. Note that adding the PPs only extended the length of sentences, without adding new layers of structure. Thus, extra processing load, even if imposed only linearly and not hierarchically, leads to a difficulty in comprehension of whstructures in healthy elderly adults. This finding supports the notion that excessive processing demands may turn the cueing strategy ineffective.

## 4. Discussion

The fact that the LM group turned out to be significantly older than the NM group may reflect dynamics of language deterioration associated with aging. While the results of evaluative measures showed that age affected both MoCA and VWM scores (the higher the age, the worse the results),
education also affected the scores, with the more years of education being associated with the better scores. However, lack of a statistically significant difference in years of education between the LM and NM groups indicates that the differences in results of cognitive tests between these two groups cannot be explained in terms of a general difference in years of education. Our results generally agree with previous findings on more accuracy errors and longer RTs in syntactic processing in elderly native speakers of English (Obler et al., 1991).

### 4.1 Subject vs. object

An interesting finding of the present study is better comprehension of object than subject who DQs. Given that the distance between the moved wh-word and its trace is longer in object questions (2) than in subject questions (1), we would expect object structures to be more demanding for processing. According to distance-based accounts, the shorter distance between the trace and its gap poses less burden on WM in subject- than in object-wh-questions, which explains the data from English discussed in the Introduction. However, our finding that object DQs were better comprehended than subject DQs is not in line with such accounts.
The idea that processing in structurally different languages reflects the structural differences among languages and that in different languages different types of information may serve as cues in sentence processing is the backbone of the competition model (Bates \& MacWhinney, 1987; MacWhinney, 1987). According to this model, the language processor chooses which information to attend to in determining sentence meaning based on specific characteristics of cues. For instance, the preposition $a$ in Spanish is not a highly available cue, because it appears only with animate direct objects. It is not a highly reliable cue, because it can convey several different meanings. However, it is not costly to process, and despite its weak cue validity, it can guide sentence comprehension: the preposition $a$ has "an extremely high contrast validity... Among normal speakers, in fact, it is the most overriding cue in determining semantic role" (Benedet et al., 1998, p. 332).

Thus, the finding that object DQs were comprehended better than subject DQs reflects a strategy based on syntactic cueing: object DQs in our experiments are introduced by a PP (a qué), beginning with the preposition $a$, which signals the grammatical role of object and the thematic role of Patient. Therefore, it is possible for the processor to rely on $a$ in correctly predicting the grammatical function and temporarily assign a thematic role to the initial constituent in a sentence such as (2) before encountering the gap. Once it encounters the gap, the temporarily assigned thematic role is confirmed or disconfirmed. Our data show that this information was utilized by the LM group in the comprehension of direct object who questions. This strategy reduces the processing demands on WM and facilitates comprehension when WM resources are reduced. However,
if the processing demands are too high, as in examples of EQs extended with additional PPs, the cueing strategy is not effective.
Another indicator that comprehension of EQs in two groups of elderly was at chance due to processing limitations is related to the finding that their comprehension of RCs was accurate. This finding shows that in older Spanish speakers this cuing strategy can be effective in syntactically more difficult conditions (e.g., object RCs and DQs), but it may not be effective when processing load is too high, regardless of syntactic complexity (extended EQs). Since the strategy of reliance on the preposition $a$ is language-specific, it is not available to speakers of English, and therefore the patterns of comprehension of object structures differ in the elderly speakers of these two languages.

### 4.2 Who vs. which

Wh-word-order in Spanish requires that a wh-word occupies a sentence- or clause-initial position, prohibiting preverbal subjects (Jaeggli, 1982; Goodall, 2004). Wh-words in multiple wh-questions, however, can switch between subject and object positions, as shown in (4-7):
(4) ¿Quién compro qué?
"Who bought what?"
(5) ¿Qué compro quién?
"What did who buy?"
(6) Juan sabe qué dijo quién.
"Juan knows what who said."
(7) Juan sabe quién dijo qué.
"Juan knows who said what." (Jaeggli, 1982, p.156).
Since quién and qué can switch their positions in a sentence, it appears that they do not obey the Superiority requirements (Chomsky, 1973; Pesetsky, 1987). This further means that there are no syntactic differences between quién and qué, and thus no syntactic reason to expect differences in their processing.
There are, however, differences between quién and qué at the discourse level: quién "who" is non-referential and non-discourse-linked, while qué "which" is referential and discourse-linked (D-linked). Some researchers argue that this difference affects processing (Hickok \& Avrutin, 1995): D-linked expressions are easier to comprehend, because they refer to a set of objects that is already known to the hearer. By contrast, who/what refers to an unlimited set of objects with which the hearer is not familiar, which makes them more difficult to process. Other researchers, however, pointed out that it is precisely their D-linked nature that makes which expressions more difficult, because they require processing and integration of information at two levels - syntax and discourse (Avrutin, 2000). Our data support the former view, showing the effect of $w h$-word in EQs, i.e., better comprehension of which questions. However, additional processing load of 1 or 2 PPs cancelled out the cuing strategy based on structural sentential features,
which facilitated comprehension of direct object who questions.
In conclusion, the main finding of the present study is a decline in comprehension of $w h$-structures and reliance on cueing as a compensatory mechanism in sentence comprehension in older native-Spanish-speaking adults. This strategy is effective in syntactically demanding conditions, when WM demands are not too high. However, excessive WM load prevents the use of the strategy. Thus, our results agree with previous findings in suggesting that it is not syntax per se, but limitation of WM resources which are necessary for processing that is affected in aging.
Further research needs to address questions pertaining to language-memory interface in older people with lower scores on tests of global cognition, such as MoCA, whose language, although appears to be normal, shows signs of decline when tested more carefully. Studying the compensatory mechanisms and strategies employed in language processing in such individuals may help us to understand better the transition from healthy aging to mild cognitive impairment and the AD continuum.

## References

Almor, A., MacDonald, M. C., Kempler, D., Andersen, E. S., \& Tyler, L. K. (2001). Comprehension of long distance number agreement in probable Alzheimer's disease. Language \& Cognitive Processes, 16(1), 35-63.
Avrutin, S. (2000). Comprehension of Discourse-Linked and Non-Discourse-Linked Questions by Children and Broca's Aphasics. In: Grodzinsky, Y., Saphiro, L. \& Swinney, D. (Eds.), Language and the Brain (pp. 295313). San Diego: Academic Press.

Bates, E., \& MacWhinney, B. (1987). Competition, variation, and language learning. In B. MacWhinney (Ed.), Mechanisms of Language Acquisition (pp. 157193). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.
Benedet, M., Christiansen, J. \& Goodglass, H. (1998). A cross-linguistic study of grammatical morphology in Spanish and English speaking agrammatic patients. Cortex, 34, 309-336.
Buckner, R., Snyder, A.Z., Shannon, B.J., LaRossa, G., Sachs, R. et al. (2005). Molecular, structural, and functional characterization of Alzheimer's disease: evidence for a relationship between default activity, amyloid, and memory. Journal of Neuroscience, 25, 770917.

Caplan, D. \& Waters, G. (1999). Verbal Working Memory and Sentence comprehension. Behavioural and Brain Sciences, 22, 77-126.
Chertkow, H., Massoud, F., Nasreddine, Z., Belleville, S., Joanette, Y., Bocti, C., Drolet, V., Kirk, J., Freedman, M., Bergman, H. (2008). Diagnosis and treatment of dementia: Mild cognitive impairment and cognitive impairment without dementia. CMAJ, 178(10), 12731285.

Chomsky, N. (1973). Conditions on Transformations. In: Anderson, S. \& Kiparsky, P. (Eds.), A Festschrift for Morris Halle, Holt, Rinehart, and Winston, New York.
Dubois, B., Feldman, H.H., Jacova, C., Dekosky, S.T., Cummings, J.L., et al.( 2007). Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDSADRDA criteria. Lancet Neurology, 6(8), 734-746.
Dubois, B., Feldman, H.H., Jacova, C., Cummings, J.L., Dekosky, S.T., et al. (2010). Revising the definition of Alzheimer's disease: a new lexicon. Lancet Neurology, 9(11), 1118-27.
Goodal, G. (2004). On the Syntax and Processing of Whquestions in Spanish. In: Schmeiser, B., Chand, V., Kelleher, A. \& Rodriguez, A. (Eds.), WCCFL 23 Proceedings (pp. 101-114). Somerville, MA: Cascadilla Press.
Goral, M., Clark-Cotton, M., Spiro, A., Obler, L.K., Verkuilen, J. et al. (2011). The Contribution of Set Switching and Working Memory to Sentence Processing in Older Adults. Exp Aging Research, 375(5), 516-538.
Grossman, M., Cooke, A., De Vita, C., Alsop, D., Detre, J., Chen, W., \& Gee, J. (2002). Age-Related Changes in Working Memory during Sentence Comprehension: An fMRI Study. NeuroImage, 15, 302-317.
Grossman, M., Cooke, A., De Vita, C., Chen, W., Moore, P., Detre, J., Alsop, D. \& Gee, J. (2002). Sentence Processing Strategies in Healthy Seniors with Poor Comprehension: An fMRI Study. Brain and Language, 80, 296-313.
Grossman M, White-Devine T. (1998). Sentence Comprehension in Alzheimer's disease. Brain and Language, 62, 186-201.
Grossman, M., Murray, R., Koenig, P., Ash, S., Cross, K., Moore, P. et al. (2007). Verb acquisition and representation in Alzheimer's disease. Neuropsychologia, 45, 2508-2518.
Hendriks, P., Englert, C., Wubs, E. \& Hoeks, J. (2008). Age Differences in Adults; Use of Referring Expressions. J Log Lan Inf
Hickok, G. \& Avrutin, S. (1996). Comprehension of WhQuestions in Two Broca's Aphasics. Brain and Language, 52, 314-327.
Jaeggli, O. (1982). Topics in Romance Syntax. Foris Publications. Dordrecht - Holland/Cinnaminson - U.S.A.
Kempe, V. \& MacWhinney, B. (1999). Processing of Morphological and Semantic Cues in Russian and German. Language and Cognitive Processes, 14(2), 129171.

Kemper, S., Marquis, J. \& Thompson, M. (2001). Longitudinal Change in Language Production: Effects of Aging and Dementia on Grammatical Complexity and Propositional Content. Psychology and Aging, 16(4), 600614.

Kempler, D., Almor, A. \& MacDonald, M.C. (1998). Teasing Apart the Contribution of Memory and language Impairment in Alzheimer's Disease: An Online Study of

Sentence Comprehension. American Journal of SpeechLanguage Pathology, 7, 61-67.
MacWhinney, B. (1987). The Competition Model. In: MacWhinney, B. (Ed.), Mechanisms of Language Acquisition (pp. 249-308). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.
Obler, L.K., Fein, D., Nicholas, M. \& Albert,M.L. (1991). Auditory comprehension and aging: Decline in syntactic processing. Applied psycholinguistics 12, 433-452.
Peelle, J.E., Troiani, V., Wingfield, A. \& Grossman, M. (2010). Neural Processing during Older Adults' Comprehension of Spoken Sentences: Age Differences in Resource Allocation and Connectivity. Cerebral Cortex, 20, 773-782.
Pesetsky, D. (1987). Wh-in situ. Movement and unselective Binding. Reuland, E. \& ter Meulen, A. (Eds.), The Representations of (In)definiteness (pp. 98-129). Cambridge, Mass: MIT Press.
Salthouse, T. A. (2009). When does age-related cognitive decline begin? Neurobiology of Aging, 30, 507-514.
Sperling, R.A., Aisen, P.S., Beckett, L.A., Bennett, D.A., Craft, S. et al. (2011). Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging and the Alzheimer's Association workgroup. Alzheimer's \& Dementia, 7(3), 280-292.
Stine-Morrow, E.A.I., Ryan, S. \& Leonard, S. (2000). Age Differences in On-Line Syntactic Processing. Experimental Aging Research, 26, 315-322.
Tun, P.A., Benichov, J. \& Wingfield, A. (2010). Response latencies in Auditory Sentence Comprehension: Effects of Linguistic versus Perceptual Challenge. Psychological Aging, 25 (3), 730-735.
Tyler, L.K., Shafto, M.A., Randall, B., Wright, P., MarslenWilson, W.D. et al. (2010). Preserving Syntactic Processing across the Adult Life Span: The Modulation of the Frontotemporal Language System in the Context of Age-Related Atrophy. Cerebral Cortex, 20, 352-364.
Wingfield, A. \& Grossman, M. (2006). Language and the Aging Brain: Patterns of Neural Compensation Revealed by Functional Brain Imaging. Journal of Neurophysiology, 96, 2830-2839.
Zurif, E., Swinney, D., Prather, P., Wingfield, A. \& Brownell, H. (1995). The Allocation of Memory Resources during Sentence Comprehension: Evidence from the Elderly. Journal of Psycholinguistic Research, 24, 165-182.

# Returning the Ticket - Mental Time Travel Reconsidered 

Markus Kneer<br>Institut Jean Nicod<br>kneer@ens.fr


#### Abstract

Mental Time Travel (MTT) is, roughly, an individual's capacity to project herself into the past or future by remembering or imagining first-personal experiences respectively. MTT is further presumed to have a distinct, concrete though dispersed neural correlate, and hence describes a neuro-cognitive phenomenon.

Opening with a brief sketch of the development and current state of the art, the essay pursues three central aims: Firstly, it constitutes a plea for more conceptual rigour on the cognitive side of the fence, so as to ensure that meaningful lessons can be drawn from neurological enquiry about it. Secondly, a partial conceptual qualification of the necessary requirements of MTT as traditionally conceived is proposed, as they seem vague, uninformative and arbitrary. Finally, a revision of MTT is attempted, which aspires to include a variety of mental states so far not associated with MTT. MTT, as it is currently defined and investigated, I will argue, stands too heavily in the genealogical debt of research into episodic memory, and suffers from an astonishing neglect of considerations pertaining to imagination.


## 1. What is Mental Time Travel?

The fact that certain types of imagination activate the same brain zones as episodic memory provoked the hypothesis that there is a single neuro-cognitive system which enables human beings to engage in mental time travel (MTT). Mental time travel is, roughly, an individual's capacity to project herself into the past or future by remembering or imagining first-personal experiences respectively. As a neuro-cognitive phenomenon, MTT is presumed to have a distinct, though dispersed, neural correlate.

Evidence from various disciplines is consistent with the MTT hypothesis. Studies in ontogenetics have confirmed that episodic memory and prospection (mental time travel into the future) emerge in parallel in children aged around three to four. Furthermore, episode specific details decrease with age both for generated past and future events. ${ }^{1}$

Lesion studies show that ventromedial frontal damage leads to loss of episodic memory and prospection, while leaving large parts of the cognitive apparatus in tact. ${ }^{2}$ Moreover, patients with hippocampal amnesia are unable to generate everyday imaginary experiences. ${ }^{3}$

[^309]Neuroimaging draws a similar picture. According to a variety of studies, episodic states about future and past have a common underlying cerebral base: When talking freely about past or future events, PET and fMRI scans revealed shared activity in regions including the ventromedial prefrontal cortex, and parts of the medial temporal lobe. ${ }^{4}$

I'll open with a brief sketch of the leading account of mental time travel. ${ }^{5}$ I will then proceed to argue that this account is misconceived for two fundamental reasons: (1) Its genealogical debt to episodic memory and autonoetic consciousness as well as the shallow conception of imagination in play give rise to an ad hoc and unnecessarily constrained account of episodic states. (2) The necessary capacities for MTT as traditionally conceived are unfounded, their formulation is conceptually vague and uninformative. This will severely obstruct empirical research into the ontogenetic and neurological foundations of the phenomenon.

## 2. Foundations of the Traditional Account

MTT has developed out of the psychological study of memory, and in particular Tulving's (1972) landmark distinction between semantic and episodic memory. The former takes propositional form since it is factual. I

[^310]can e.g. recall that Paris is the capital of France. Episodic memory, by contrast, refers to an individual's engaging in an episode of past personal experience, e.g. when I remember what my first arrival in Paris was like. It is characteristically accompanied by a particular feeling of "warmth and intimacy" (W. James, 1890), in other words it is phenomenologically rich. Later on, autonoetic - "self-knowing" - consciousness became the distinguishing mark of episodic memory. Tulving, who coined the term "autonoetic consciousness", defines it as "the kind of consciousness that mediates an individual's awareness of his or her existence and identity in subjective time extending from the personal past through the present to the personal future" (1985: 1). In response to neuropsychological findings, the restriction of episodic states soon dropped away; episodic states were henceforth considered to encompass mental time travel both into the past and future.

My first criticism concerns the scope of the leading account: MTT is restricted to episodic states exclusively concerned with an individual's personal past, present and future. Employing a detailed typology of imaginative states I will draw up a rival account which construes the phenomenon in question more broadly. I will argue from the following two hypotheses:

## Common Kind Hypothesis:

There exists a basic mental state, called "episodic state", in which we undergo phenomenologically rich experiences from a first person perspective. Such states are "quasi-perceptual" in so far as they resemble perceptions, and draw heavily on past perceptual and proprioceptive intake, but are not direct representations of reality. Episodic states comprise of episodic memory - quasiperceptions of the past, and participatory imagination - quasi-perceptions of hypothetical and potentially future episodes.

## Common Capacity Hypothesis:

Episodic states supervene on a single type of brain state; they can be characterized by a differentiated neuronal correlate and are the product of a particular neurosystem of the brain.

## 3. Problems of Scope and a Rival Account

MTT, as it is traditionally conceived, is construed unnecessarily narrow, since it insists on (i) a clearly defined temporal component, which (ii) involves an explicit awareness of a narrative self and concerns (iii) episodes which are explicitly personal/autobiographical - in the sense that the subject involved must be the thinker's empirical self, and the scenarios must be true past or probable future experiences. Both constraints can be directly derived from Tulving's characterization of MTT taking place in "subjective time extending from the personal past through the present to the personal future" and countless other passages. They are equally present in Suddendorf \& Corballis $(1997,2007)$.

The main reason for the narrow construction of MTT seems to lie in the genealogy of the term, coming from episodic memory and hence focusing heavily both on a temporal component and some sort of autobiographical element. Autonoetic awareness as the central property of episodic states has further helped to foster such a questionable conception. The problem extends into the experimental paradigms: Not only is rather few research done on future MTT (i.e. the imagination component as narrowly conceived), but furthermore hardly any experiments included imagination not structured autobiographically or temporally in the relevant way.

Mental voyage, as I propose it, explicitly includes three types of episodic imagination which are not encompassed in MTT: (i) Episodes about what would have happened to myself if I had acted differently in the past. (ii) Imaginations involving my empirical self in scenarios which do not have a temporal specification whatsoever. (iii) Imaginations not involving my empirical self, but another self, or a general self - i.e. episodes concerning what it is like for Jack to win an Oscar, or for someone to win the lottery respectively. ${ }^{6}$

[^311]
## 4. Central Capacities for MTT

## Methodological Considerations

Mental time travel is generally conceived as a neuro-cognitive or brain/mind system ${ }^{7}$, that is, as a mental capacity which has a real, singular, though dispersed neural correlate. Crucially, the description of its functions and properties take place in two conceptual spaces: an abstract vocabulary pertaining to mind as elaborated in psychology and philosophy and the vocabulary referring to concrete phenomena of the brain as employed by neuroscience. The mental vocabulary is dominant due to our still very limited understanding of the brain. Hypotheses concerning the functioning of the brain as well as experiential paradigms are largely formulated in mental terms, which gives rise to a variety of complications: (i) The conceptions of the mind and its capacities are manifold, so a choice has to be taken which (ii) is likely to leave its mark on the formulation of the respective hypothesis, experiments, and hence the empirical "findings". (iii) A neuro-cognitive hybrid vocabulary facilitates conceptual confusion if (for instance) a brain phenomenon is "associated" with a mental phenomenon which gives rise to different interpretations in distinct conceptual frameworks of the mind.

Given the complications arising from two distinct types of interacting vocabularies, empirical underdetermination, and the sensitivity of the subtraction method in hemodynamic techniques (PET and fMRI), two things should be clear: Adequate enquiry into such neuro-cognitive phenomena can only succeed if its constituents are (i) conceived of in minimal, rather than complex units and (ii) defined as rigorously as possible on the cognitive side so as not to obstruct and confuse enquiry on the neurological side. The literature on MTT does not adhere to these criteria. The central capacities of MTT are both theoretically ad hoc and so ill defined that while scholars take themselves to be in conceptual agreement they frequently operate with - and do empirical research based on - radically different concepts. We will now turn both to the misspecification of

[^312]the central capacities of MTT and the latent conceptual anarchy.

## Autonoetic Consciousness

The central criterion of episodic memory and states of mental time travel more generally, is autonoetic awareness (Wheeler, Stuss \& Tulving, 1997; Tulving 2002; Suddendorf \& Corballis 1997, 2007). Though a relatively recent, and entirely technical concept, astonishingly there are four conceptions of autonoetic consciousness which stand in rivalry.

Autonoetic consciousness understood as the distinguishing mark of mental states which have a particular "feel" to them is roughly equivalent with Block's (1995) "phenomenal consciousness". Autonoetic awareness in this sense is usually cashed out by reference to the feeling of "warmth and intimacy" (W. James, 1890), or the "subjectivity" of such states in comparison to, for example, "objective" semantic memory. A second account puts the stress on (narrative) selfawareness, reasonably enough, since autonetic consciousness has frequently been characterized as "self-knowing" consciousness (Tulving, 1985), and it is due to autonoesis that an individual is supposedly able to project his self into the past and future. A third account focuses on the kind of (phenomenal) feature which allows us to distinguish, for instance, an episodic memory from an imagination or a daydream. Episodic states seem to come pack and parcel with a certain phenomenological feature pertaining to time, which allows us not to confound past, present and future episodic states; in the case of memory we witness, for instance, a "feeling of pastness". Finally, there are various hybrid accounts which include some or all of the mentioned features. Unsurprisingly, with autonoesis - "the hallmark of episodic memory" (Tulving) being such a promiscuous concept, the episodic/semantic memory distinction is also drawn in all sorts of ways.

The three mentioned aspects, phenomenality, narrative self-awareness, and subjective temporal indexing are neither inconsistent nor necessarily unrelated, but nonetheless distinct. Problematically, however, different authors seem to work with different
definitions. ${ }^{8}$ There is, furthermore, no apparent reason for lumping them together especially inexplicitly, and in varying constellations. In fact, it seems favourable to keep them separate so as to curb the spreading conceptual confusion, and in particular so as to control - as far as possible - separately for each in experiments.

## Consciousness

How best to reorganize what is left of autonoetic consciousness? Block (1995) distinguishes four concepts of consciousness: Self-consciousness constitutes the possession and competent mastery of the concept of the self; monitoring consciousness is the metacognitive process of one's realizing to be in a certain state (e.g. to know that one believes X). Access-consciousness is the property of a representation which "is broadcast for free use in reasoning and for direct "rational" control of action (including reporting)". Finally, and for our purposes probably most importantly, there is phenomenal consciousness, which is notoriously hard to define. Under phenomenal consciousness we understand the experiential properties of a conscious state, or what it "is like" to be in that state. ${ }^{9}$ Block highlights that phenomenal and access consciousness are conceptually distinct, though admits that they might contingently always appear in parallel in human subjects.

My proposal is to abolish the concept of autonoetic awareness since it is vague and unnecessarily lumps together all sorts of phenomena which are best kept separate. Phenomenal consciousness (or phenomenality) is most salient in perceptual experiences, however it also characterizes episodes of remembering and sensory imagination, though the phenomenal properties are not as pronounced. It is probably phenomenality, intimately related to the first-person perspective, which determines the distinctions between episodic

[^313]and semantic memory and between sensory and propositional imagination in the first place.

## Subjective Time

As mentioned before, time (like colour or smell, for instance) also seems phenomenologically salient, in particular, duration ("an instant seeming an eternity") and the subjective temporal location ("it being early"; "something being a long time ago"). It is a feature of episodic states that they frequently present themselves as a particular episodic state (e.g. memories through an attached feeling of "pastness"). However, as pointed out above, various types of participatory imagination do not have an explicit temporal component, and are hence not located in what Tulving calls "subjective time" (Tulving, 2002, 2005; a feature reproduced in virtually all articles on MTT). Furthermore, though there frequently seems a phenomenally salient indicator present in episodic states, it can easily be wrong (as for example in implanted or false memory, possibly in states of déjà-vu experiences if one wants to count them amongst episodic states). ${ }^{10}$

In short, the temporal component should be abolished, as there is reason to include episodic states without them into the common kind; the fact that temporal features do not form part of memory traces (Friedman, 1993) leaves us with the perfectly viable option of their being "attributed" by an extra capacity, if at all.

## Involvement of the Self

Self-awareness to a rather high extent is considered a fundamental prerequisite of MTT by all leading accounts. ${ }^{11}$ For Tulving

[^314]and colleagues, the relevant degree of selfawareness required for MTT corresponds to Stuss \& Benson's (1986) third functional level of the frontal lobes ${ }^{12}$. The latter provides "the ability to introspect on one's own thoughts and to realize the relation of self to one's social environment" (1997:334), and is further deemed "intimately related" to autonoetic consciousness.

Mitchell's (1994) threefold account of the self is employed by Suddendorf \& Corballis with questionable success to illuminate the role of the self in MTT. In Mitchell's framework, the second type of self - a "self as built on kinaesthetic-visual matching" allows us to engage in pretense, planning, imaginative experience and fantasy. Suddendorf \& Corballis deem this insufficient for MTT, which apparently calls for a self "built on symbols, language and artefacts" so as to allow the individual to understand social norms, and to dissociate self and other. It rests obscure, however, how pretense, imagination and fantasy (i.e. second-level activities) are possible without the capacity to dissociate from one's present states (only available on the third level) and hence whether Mitchell's account is not rather ad hoc. ${ }^{13}$

Ignoring the specificities of the two proposed accounts of the self, let us note that they both share two essential features: Firstly, they require the mastery and competent application of an explicit concept of the self, and secondly, they pertain to what is frequently called a "narrative" account of the self.

The narrative self contrasts with the "minimal" self, i.e. immediate awareness of oneself as the subject of experience. A minimal self is conceived to be little more than "a bare locus of consciousness, void of personality" (G. Strawson, 1999: 493), aspects of continuity over time are neither included in the definition nor deemed necessary. By contrast, an

[^315]adequate account of the latter - i.e. of how to explain an awareness of oneself from past to present and future, both in terms of one's individual experience and testimony of others' - is the centrepiece of "narrative" conceptions of the self (cf. Gallagher, 2000 for a recent review). Both minimal and narrative self demand the competent mastery of a selfconcept. However, even simpler conceptions are thinkable, amounting to no more than a point of consciousness to which action and experience is relativized.

## The Self Reconsidered

Though self-awareness in the general sense is undoubtedly an important aspect of MTT, it is difficult to say something insightful about it. Coherent ad hoc stories about the involvement of a narrative self in MTT can be told to abundance within the framework of our mental vocabulary; however, it rests entirely in the dark of what should constitute (and how to test for) the involvement of narrative self-consciousness on the level of the brain. ${ }^{14}$

Rather than working with the proposed, extremely complex notions of self-awareness as necessary capacities for MTT, it seems preferable to focus on minimal, separable ingredients. Episodic thought is not fundamentally characterised by its relation to the subject's self, but rather by being a particular mode of thought - it is essentially first-personal, but not essentially autobiographical as often presumed. Furthermore, episodic thought is susceptible to a certain measure of control: I can bring out more vividly, or shift attention to, different aspects of a past event; in imagination the freedom of control is even more pronounced.

On top of narrativity, we might be inclined to question mastery of a self-concept as a necessary requirement of MTT. Perry (1998) for instance, contrasts agent-relative knowledge with self-attached knowledge. The former takes place "from the perspective of a particular agent who does not need to have an idea of self, or a notion of himself' (BBB) but who is nonetheless capable of placing

[^316]himself within his environment. One way to make this process explicit are utterance involving demonstratives, e.g. when he thinks/says "There is an apple". In selfattached knowledge, by contrast, the agent disposes of a self-notion and expresses his knowledge by means of the (pure) indexical "l", e.g. "I see an apple". Both involve the firstperson perspective, essential for episodic states, but only the latter type of subject can make this explicit to herself.

Interestingly, the first kind of knowledge, and the very basic type of self involved would satisfy a variety of episodic states: If I engage in a phenomenally rich first-person episode with the content "There was an apple", or follow the instruction "Imagine an apple", I do not necessarily need the first-person indexical; an appropriately imagined content of the kind "There is a juicy red apple" fulfils the demand no less than "I see a juicy red apple in front of the eye of my mind". However, whenever the person who remembers or imagines is to take himself as the object of his episodic state, he or she is required to have an explicit notion of himself, otherwise he couldn't attribute any experiences or properties to himself. What this confirms, once again, is that the capacity to engage in episodic states is - at least in principle - independent of the capacity to engage in autobiographical episodic states (though due to the contingent set-up of the brain this might not actually be so).

A variety of insights follow from this: Firstly, we have another argument why autonoesis i.e. "self-knowing consciousness" - might not be a fundamental requirement of MTT in so far as there exists a kind of basic episodic state which does not depend on the mental time traveller having even a very basic notion of himself. Secondly, and relatedly, the firstperson perspective might be entirely severed form any type of complex narrative self in so far as there is a variety of possibilities to "fill" it: Apart from one's own self, it could be another's or a general self. Thirdly, it might be hypothesized that MTT is intimately related to the mastery of demonstratives (and as concerns autobiographical episodic states, also indexicals) and temporal concepts, or their respective non-conceptual counterparts. So an important open question is to what
extend the non-conceptual and pre-linguistic resources (e.g. of small children) suffice for MTT. ${ }^{15}$

In short, there is no need to presume MTT in the need of an awareness of a complex narrative self, whose involvement will most likely prove impossible (or very difficult) to test for. Furthermore, it is not clear why complex introspective abilities, metarepresentation and awareness of one's social environment etc. should be necessary for basic episodic thought. As has been argued, the only true requirements for basic MTT are the involvement of the first-person perspective and a given measure of control of the episodic content. Non-autobiographical episodic states (There was an elephant or This will be wet and cold etc.) might not even demand an explicit notion of the self. Finally, at least in so far as mental voyage is concerned, it will be extremely misleading to stress the requirement of (narrative) selfawareness, both in so far as certain types of participatory imagination are neither relativized to "subjective time" (i.e. are not locatable within a subject's personal narrative) or to the subject's own self at all. They might simply be about what it would be like to win the Nobel Prize or how it would feel to have one's neighbour's problems. Rather than operating with various ad-hoc accounts of the narrative self, it seems advisable to take the first-personal mode and control (or sense of agency) as the basic requirements for MTT, and to devise abilities allowing us to judge to what extent an explicit notion of the self, and conceptual abilities are required as well.

## Bibliography

Addis, D. R., A. T. Wong, et al. (2007). "Remembering the past and imagining the future: Common and distinct neural substrates during event construction and elaboration." Neuropsychologia 45(7): 1363-1377.

[^317]Addis, D. R., A. T. Wong, et al. (2008). "Age-Related Changes in the Episodic Simulation of Future Events." Psychological Science 19(1): 33-41.

Atance, C. M. and D. K. O'Neill (2001). "Episodic future thinking." Trends in Cognitive Sciences 5(12): 533-539.

Bechtel, W. (2002). "Decomposing the Mind-Brain: A LongTerm Pursuit." Brain and Mind 3(2): 229-242.

Bechtel, W. (Forthcoming in Skipper et al. (eds.): Philosophy and the Life Sciences: A Reader. MIT Press. ). "The Epistemology of Evidence in Cognitive Neuroscience."

Bermúdez, J. L. (1998). The Paradox of Self-Consciousness, MIT Press.

Bischof-Köhler, D. (2000). Kinder auf Zeitreise Theory of Mind, Zeitverständnis und Handlungsorganisation. Bern Göttingen [u.a.], Huber.

Block, N. (1995). "On a confusion about a function of consciousness." Behavioral and Brain Sciences 18(2): 227287.

Botzung, A., E. Denkova, et al. (2008). "Experiencing past and future personal events: Functional neuroimaging evidence on the neural bases of mental time travel." Brain and Cognition 66(2): 202-212.

Bub, D. N. (2000). "Methodological issues confronting PET and fMRI studies of cognitive function." Cognitive neuropsychology(Print) 17(5): 467-484.

Buckner, R. L. and D. C. Carroll (2007). "Self-projection and the brain." Trends in Cognitive Sciences 11(2): 49-57.

Buckner, R. L. and W. Koutstaal (1998). "Functional neuroimaging studies of encoding, priming, and explicit memory retrieval." Proceedings of the National Academy of Sciences 95(3): 891.

Buckner, R. L., M. E. Raichle, et al. (1996). "Functional Anatomic Studies of Memory Retrieval for Auditory Words and Visual Pictures." Journal of Neuroscience 16(19): 6219.

Cabeza, R., F. Dolcos, et al. (2002). "Similarities and Differences in the Neural Correlates of Episodic Memory Retrieval and Working Memory." Neuroimage 16(2): 317-330.

Cabeza, R. and L. Nyberg (2000). "Imaging Cognition II: An Empirical Review of 275 PET and fMRI Studies." Journal of Cognitive Neuroscience 12(1): 1-47.

Cermak, L. S. and F. I. M. Craik (1979). Levels of processing in human memory, Lawrence Erlbaum Associates New York: Distributed by Halsted Press Division, Wiley, Hillsdale, NJ.

Clayton, N. S., T. J. Bussey, et al. (2003). "Can animals recall the past and plan for the future?" Nat Rev Neurosci 4(8): 68591.

Clayton, N. S., D. P. Griffiths, et al. (2000). "Declarative and episodic-like memory in animals: Personal musings of a Scrub Jay." The evolution of cognition: 273-288.

D'Argembeau, A., C. Comblain, et al. (2003). "Phenomenal characteristics of autobiographical memories for positive, negative, and neutral events." Applied Cognitive Psychology 17(3): 281-294.

D'Argembeau, A., S. Raffard, et al. (2008). "Remembering the past and imagining the future in schizophrenia." Journal of Abnormal Psychology 117(1): 247-251.

D'Argembeau, A. and M. Van der Linden (2006). "Individual differences in the phenomenology of mental time travel: The effect of vivid visual imagery and emotion regulation strategies." Consciousness and Cognition 15(2): 342-350.

Dennett, D. C. (1991). Consciousness explained, Little, Brown and Co Boston.

Fink, G. R., H. J. Markowitsch, et al. (1996). "Cerebral Representation of One"s Own Past: Neural Networks Involved in Autobiographical Memory." Journal of Neuroscience 16(13): 4275.

Friedman, W. J. (1993). "Memory for the time of past events." Psychological bulletin 113(1): 44-66.

Friston, K. J., C. J. Price, et al. (1996). "The Trouble with Cognitive Subtraction." Neuroimage 4(2): 97-104.

Gallagher, S. (2000). "Philosophical conceptions of the self: implications for cognitive science." Trends in Cognitive Sciences 4(1): 14-21.

Gardiner, J. M. (2001). "Episodic Memory and Autonoetic Consciousness: A First-Person Approach." Philosophical Transactions: Biological Sciences 356(1413): 1351-1361.

Gerrans, P. K., J. (MS). "Moral Meta-Cognition."
Gerrans, P. S., D. (MS). "Implicit Mental Time Travel."
Hassabis, D., D. Kumaran, et al. (2007). "Patients with hippocampal amnesia cannot imagine new experiences." Proceedings of the National Academy of Sciences 104(5): 1726.

Hassabis, D. and E. A. Maguire (2007). "Deconstructing episodic memory with construction." Trends in Cognitive Sciences 11(7): 299-306.

James, W. (1890/2007). The Principles of Psychology, Cosimo Classics.

Jennings, J. M., A. R. McIntosh, et al. (1997). "Cognitive Subtractions May Not Add Up: The Interaction between Semantic Processing and Response Mode." Neuroimage 5(3): 229-239.

Klein, S. B., J. Loftus, et al. (2002). "Memory and temporal experience: The effects of episodic memory loss on an amnesic patient's ability to remember the past and imagine the future." Social Cognition 20(5): 353-379.

Levine, B. (2004). "Autobiographical memory and the self in time: Brain lesion effects, functional neuroanatomy, and lifespan development." Brain and Cognition 55(1): 54-68.

Levine, B., E. Svoboda, et al. (2002). "Aging and autobiographical memory: dissociating episodic from semantic retrieval." Psychol Aging 17(4): 677-89.

Loftus, E. F. (1995). "The reality of repressed memories." Annual Progress in Child Psychiatry and Child Development, 1994.

Loftus, E. F. and D. Davis (2006). "Recovered memories." Annual Review of Clinical Psychology 2: 469-498.

Macphail, E. M. (2000). "The search for a mental Rubicon." The Evolution of Cognition, MIT Press, Cambridge.

Nagel, T. (1974). "What is it like to be a bat." Philosophical Review 83(4): 435-450.

Neisser, U. (1988). "Five kinds of self-knowledge."
Philosophical Psychology 1(1): 35-59.
Nelson, K. (2001). "Language and the self: From the "experiencing l"to the "continuing me"." The self in time: Developmental perspectives: 15-33.

Nyberg, L., R. Cabeza, et al. (1996). "PET studies of encoding and retrieval: The HERA model." Psychonomic Bulletin and Review 3(2): 135-148.

Okuda, J., T. Fujii, et al. (2003). "Thinking of the future and past: the roles of the frontal pole and the medial temporal lobes." Neuroimage 19(4): 1369-1380.

Perry, J. (1998). "Myself and I." Philosophie in Synthetischer Absicht: 83-103.

Povinelli, D. J. (2001). "The self: Elevated in consciousness and extended in time." The self in time: Developmental perspectives: 75-95.

Price, C. J., C. J. Moore, et al. (1997). "Subtractions, conjunctions, and interactions in experimental design of activation studies." Human Brain Mapping 5(4): 264-272.

Schacter, D. L. and D. R. Addis (2007). "On the constructive episodic simulation of past and future events." Behavioral and Brain Sciences 30(03): 331-332.

Squire, L. R., J. G. Ojemann, et al. (1992). "Activation of the hippocampus in normal humans: a functional anatomical study of memory." Proc Natl Acad Sci US A 89(5): 1837-41.

Steinvorth, S., S. Corkin, et al. (2006). "Ecphory of autobiographical memories: An fMRI study of recent and remote memory retrieval." Neuroimage 30(1): 285-298.

Strawson, G. (1999). "The Self and the SESMET." Journal of Consciousness Studies 6(4): 99-135.

Suddendorf, T. and J. Busby (2003). "Like it or not? The mental time travel debate: Reply to Clayton et al." Trends in Cognitive Sciences 7(10): 437-438.

Suddendorf, T. and J. Busby (2003). "Mental time travel in animals?" Trends in Cognitive Sciences 7(9): 391-396.

Suddendorf, T. and J. Busby (2005). "Making decisions with the future in mind: Developmental and comparative identification of mental time travel." Learning and Motivation 36(2): 110-125.

Suddendorf, T. and M. C. Corballis (1997). "Mental time travel and the evolution of the human mind." Genetic, Social, and General Psychology Monographs 123(2): 133-167.

Szpunar, K. K., J. M. Watson, et al. (2007). "Neural substrates of envisioning the future." Proceedings of the National Academy of Sciences 104(2): 642.

Tulving, E. (1972). "Episodic and semantic memory." Organization of memory: 381-403.

Tulving, E. (1985). "Memory and consciousness." Canadian Psychology-Psychologie Canadienne 26: 1-12.

Tulving, E. (1993). "What Is Episodic Memory?" Current Directions in Psychological Science 2(3): 67-70.

Tulving, E. (2002). "Episodic memory: From mind to brain." Annual Review of Psychology 53(1): 1-25.

Tulving, E., C. A. Hayman, et al. (1991). "Long-lasting perceptual priming and semantic learning in amnesia: a case experiment." J Exp Psychol Learn Mem Cogn 17(4): 595-617.

Zentall, T. R. (2006). "Mental time travel in animals: A challenging question." Behavioural Processes 72(2): 173-183.

# The time-course of processing discourse connectives 

Judith Köhne (judith.koehne@uni-bamberg.de)<br>Otto-Friedrich-Universität Bamberg<br>Markusplatz 3, 96045 Bamberg

Vera Demberg (vera@coli.uni-saarland.de)<br>Cluster of Excellence, Saarland University,<br>Campus C7.4, 66123 Saarbrücken, Germany


#### Abstract

While there is some evidence that causal discourse relations are processed incrementally, the time-course of comprehending concessive discourse markers (e.g., nevertheless) has hardly been investigated. Given that concessives are often defined as negative causals, there may be similarities between the processing of concessives and negations (e.g., a delay). This paper investigates the time-course of processing causal versus concessive discourse markers in German within both a visual-world experiment and a reading experiment. We find that while concessive discourse markers can be processed rapidly if the context is constraining enough, there is a delay compared to causal contexts.


Keywords: Discourse connectives; prediction; concessives; incrementality; eye-tracking; visual world

## Processing Discourse Relations

A large number of experiments reveals that language comprehension is generally incremental and even predictive (Marslen-Wilson, 1973; Tanenhaus \& Trueswell, 1995). However, there is also evidence that incoming information across the sentence level does not always immediately update local predictions and affect global interpretation (Sanford \& Garrod, 1998). It is therefore an interesting question which information from the discourse the comprehender considers, and how strongly and fast it affects comprehension and active predictions. One way to investigate this issue is to focus on the time-course of discourse connectors. Experimental evidence suggests that discourse connectors such as because, therefore, and however facilitate coherence building and hence comprehension: Millis \& Just (1994) found that when sentences were connected by discourse markers (because and although), people were able to more successfully answer comprehension questions and to more quickly read the second sentence.

We however know much less about the time-course of processing discourse connectors. While some have argued that people only integrate them at the end of a sentence (Millis \& Just, 1994), other experiments indicate that this integration happens much earlier (Traxler et al., 1997).

The experiments described in this paper examine the timecourse of integrating causal connectors (e.g., therefore) versus concessive connectors (e.g., however). These are particularly interesting to compare, because concessives have sometimes be referred to as "negative causals" (König \& Siemund, 2000). That means that the processing of concessives (compared to causals) may resemble the processing negation. In
particular, our experiments aim to answer the following questions:

- Are causal and concessive connectives processed incrementally (possibly eliciting predictions), or with a delay?
- Do concessives elicit an active search for alternatives (as has been shown for negation, Kaup et al., 2006)?
- Regarding global interpretation, are concessive discourse relations integrated as smoothly as causal discourse relations or do they cause processing difficulties (resembling negation; Carpenter \& Just, 1975; Kaup et al., 2006)?


## Background

## The Time-Course of Processing Connectors

Few studies have investigated the time-course of processing causals and concessives. In short discourses of two clauses, Millis \& Just (1994) observed longer wrap-up times at the end of the second clause when a (causal or concessive) discourse connector was present, as compared to the same sentences without a discourse connector. Millis and Just hypothesized that a representation of the second clause was constructed without taking into account the first clause, and only later integrated with the first clause.

Millis and Just's "Connective Integration Model" of late integration of discourse connectors and earlier parts of the discourse was however refuted, at least for causal connectors. Traxler, Bybee, \& Pickering (1997), for instance, found evidence for an early integration of because and the preceding discourse: When comparing processing of causal and diagnostic sentences, the greater difficulty in diagnostics occurred well before the end of the second clause. This indicates that processing of the second clause was affected early on by its relation to the preceding context. Further evidence for incremental processing of causal discourse relations comes from an ERP study without explicit connectives: Kuperberg et al. (2011) found that causally-related sentences were easier to process than sentences which were not standing in a causal relationship, revealing that causal coherence can influence the earliest stages of semantically processing incoming words.

These findings bring up a number of questions with respect to the exact time-course of processing connectives. Specifically, it is an open question in how far, and how quickly, people generate predictions taking into account discourse connectives, and how concessive connectives might differ from causals in this respect.

An interesting theory regarding predictions elicited by causal and concessive connectors comes from Murray (1995): In a series of studies, he found a greater beneficiary effect of the presence of contrastive and concessive (i.e., adversative) discourse connectors as opposed to causal and additive connectors. Murray concludes that adversative connectives create stronger expectations for the upcoming sentence than causal or additive connectors (Murray, 1995, p. 120). Murray does however not control for the ambiguity of discourse connectors, which means that his hypotheses may be taken with a grain of salt. Moreover, his findings could alternatively be accounted for by a causality-by-default account (Kuperberg et al., 2011; Sanders, 2005): The cause for the low facilitation for causal connectors may be that similar expectations are generated in the absence of any connector.

## Concessives as Negative Causals

To date, there is very little research on the time-course of processing concessives. An early study by Townsend (1983) reveals that concessives are processed more slowly than causals and that recall is worse for concessives than for causals. These findings may suggest an interesting relationship between causals and concessives. In fact, concessives are sometimes referred to as "negative causals" (König \& Siemund, 2000). Experimental studies support that causals and concessives establish the same type of relation, but are different in polarity (Louwerse, 2001; Sanders et al., 1992).

A delay of processing concessives on the one hand and defining concessives as negative causals on the other hand seems to be in line with a frequently supported theory of negation processing: Many experiments point to an account where there is a general delay in processing negation (e.g., Carpenter \& Just, 1975).

Kaup et al. (2006), for instance, found in a self-paced reading study combined with a picture naming task that when processing contradictory predicates (e.g., The door is not open / closed), ,
people are first mentally simulating the positive state (open door) and only later the positive state is negated. That means people only later searched for alternatives and mentally closed the door (see also Lüdtke et al., 2008). Ferguson et al. (2008) examined the time-course of processing negation in discourse using eye-tracking in reading and ERP. Interestingly, they also found that counterfactual negated discourse information was not used incrementally but had a delayed effect on comprehension.

Other studies, on the contrary, reveal that a delay can be attenuated or completely removed when the negation is expected or pragmatically licensed (Nieuwland \& Kuperberg, 2008; Dale \& Duran, 2011). Staab (2007), for instance, found in a series of ERP studies that negation in discourse context was processed fast. More than that, if readers were forced to process slowly and deeply, negation was even used as a cue to rapidly anticipate how the sentence continues. These very different results suggest that the time-course of processing negation may be influenced by a number of factors such
as the kind of negated information and the discourse context. It is an interesting question how processing negative causals may enrich this picture.

## Experiment 1: Visual World Study

## Methods

Participants We tested 36 participants, 4 of which had to be excluded due to eye-tracking problems. Data of 32 participants ( 8 male, average age 26) was analyzed.
Design, Materials \& Procedure We constructed 20 items, each consisting of three spoken sentences in German, and a static scene (see Example (1) and Figure 1).

Marc denkt über einen kleinen [Snack nach. Er hat gerade Lust, etwas $]_{\text {topic }}[$ Süßes / Salziges zu essen] category . [Daher / Dennoch holt er sich] ${ }_{\text {connector }}$ [aus der Küche] ${ }_{\text {extended }}$ [die appetitliche / den appetitlichen] ${ }_{\text {pretarget }}$ [Waffel / Kuchen / Brezel / Käse] ${ }_{\text {target }}$.
Marc fancies a [snack. He feels like having something $_{\text {topic }}[\text { sweet }]_{\text {category. }} \quad[$ Therefore / Nevertheless, he gets] $]_{\text {connector }}$ [from the kitchen] $]_{\text {extended }}$ $\left[\text { the }{ }_{[f e m] /[\text { masc }]} \text { delicious }{ }_{[\text {fem }] /[\text { masc }]}\right]_{\text {pretarget }}$ [waffle / cake / pretzel / cheese $]_{\text {target }}$.


Figure 1: Stimulus for visual world experiment.
The first sentence introduces a situation or topic, such as food in Marc denkt über einen Snack nach. ("Marc fancies a Snack"). The second sentence always identifies a category (e.g., sweet things), matching two of the depicted objects (waffle and cake). Two other objects in the scene belong to another category (the counter category, salty things: cheese and pretzel). The third sentence begins either with a causal (Daher/Dennoch) or a concessive (Deswegen/Trotzdem) connector (2-level within-participant factor), followed by subject and verb (holt er sich, "he gets"; connector region). This region precedes another phrase (aus der Küche, "from the kitchen"; extended connector region), the gender-marked pretarget noun region (e.g., die appetitliche), and the target noun (causal: Waffel, concessive: Brezel). Target nouns are always congruent with the preceding discourse. Visuals worlds include the four objects belonging to the category and the counter category and two distractor objects (here, cup and wire whisk), embedded in a simple scene (here, kitchen ).

Category given in Sentence 2, gender of target noun, and condition (i.e., causal/concessive) were fully counterbalanced, resulting in 8 lists. Every participant was assigned to one of the lists and saw each of the 20 items in one version only. 40 filler discourse-scene pairs were included, following the same general pattern but using a range of discourse relations and markers (e.g., später, "later"), making the target noun unpredictable. All items and half of the fillers were followed by a comprehension question about the target noun but referring to it by its category rather than its name (Holt Marc sich etwas Süßes?, "Does Marc get something sweet?"), which participants answered by button press (YES/NO). Half of the questions' correct answer was "yes", the other half "no". Order of presentation was pseudo-randomized with at least one filler in between two items. Participants were tested individually and their eye-movements were tracked. Their task was to look and listen carefully enough to reply to the comprehension questions. The experiment lasted about 30 minutes.
Predictions When the category (e.g., sweet) is mentioned, fast and incremental processing predicts participants to look more often at the two objects matching this category (waffle and cake) in both conditions.

For the third sentence, predictions for causal and concessive sentences differ: In the causal condition, people are predicted to keep looking at the category objects until the casemarked pretarget region. During the pretarget region then, fast and incremental processing predicts more looks towards the gender-congruent object, and finally, when the target is mentioned, more looks to the target. In the concessive condition, however, hypothesizing that the concessive connector is processed eagerly and incrementally predicts participants to change from looking to the category objects to looking to the two counter-category (salty) objects (pretzel and cheese), as soon as the scope of the concessive connector is clear.

In particular, the scope could be inferred and a search for alternatives could be initiated after the subject and verb following the connector (connector region). The hypothesis that the concessive connector is processed fast and incrementally, also predicts participants to start looking more frequently at the final target object during the gender-marked pretarget region. A late integration account, or a simple lexical priming account would not predict this pattern but that participants keep looking at the category objects (sweet things) until they hear the target word.

## Data Analyses \& Results

For eye-movement analyses, we compared inspections to the four areas of interest (AOIs): target (e.g., waffle), category competitor (sharing category with target, e.g., cake), gender competitor (sharing gender with target, e.g. pretzel), and unrelated competitor (sharing neither category nor gender with target, e.g., cheese). Four time regions were of interest: category region, connector region, extended connector region, and pretarget region. Eye-movements were analyzed using logistic regressions, entering the data into linear mixed effect
models with logit-link function (from the lme4 package in R; Bates, 2005). AOI and Condition (causal/concessive) were used as a Fixed Factors and Participant and Item as random factors. Main effects were tested based on model comparison using a $\chi$ test (Baayen et al., 2008). Random slopes for Participant and Item were evaluated based on model comparison as well and included when they improved the model fit. For contrasts between levels (AOIs), we report Wald-z values and p-values as well as coefficients ( $\beta$ ) and standard errors (SE).

Analyses reveal that in the category region, participants inspected the two objects matching this category (causal: target + category competitor; concessive: gender competitor + unrelated competitor) significantly more frequently than the counter-category objects, independent of condition (effect AOI: $\chi(1)=49.26, p<.001$; no effect condition: $\chi(1)=1.99$, $p=.16$; no interaction: $\chi(1)=1.88, p=.17)$. In the connector region, there was an effect of AOI $(\chi(1)=7.78, p<.01)$, no effect of condition $(\chi(1)=0.26, p=.61)$, but, importantly, an interaction $(\chi(1)=4.17, p<.05)$ : In the causal condition, the category objects were still looked at significantly more often than the counter-category objects $(\chi(1)=$ $11.38, p<.001$ ); in the concessive condition, however, participants inspected the two counter-category objects just as much as the category objects $(\chi(1)=0.30, p=.58)$. As illustrated in Figure 2, this is due to them first looking more at the category objects, but gradually starting to look more at the counter category objects, as the scope of the concessive becomes clear. In the extended connector region then, we find significantly more looks to the objects of the countercategory objects in the concessive condition $(\chi(1)=15.19$, $p<.001$ ) as well as still significantly more looks to the category objects in the causal condition $(\chi(1)=18.64, p<.001)$. That means that looking at the target category (i.e., category in causal and counter-category in concessives) was independent of condition in this region (effect of AOI: $\chi(1)=33.65$, $p<.001$, no effect of condition: $\chi(1)=0.76, p=.38$, and no interaction: $\chi(1)=0.11, p=.74)$. This reveals that the concessive marker was immediately interpreted, and that people engaged in an active search for alternatives. In the pretarget region (when shifted 200 ms$)^{1}$, the target was looked at more frequently than all other objects in both conditions (effect AOI: $\chi(3)=20.42, p<.001$, no effect condition: $\chi(1)$ $=0.01, p=.92$, no interaction: $\chi(3)=0.87, p=.87$; effect AOI causal: $\chi(1)=63.16, p<.001$; effect AOI concessive: $\chi(3)=12.62, p<.01)$. In the causal condition, the differences to gender competitor ( $\beta=-1.11, S E=0.19, z=-5.92$, $p<.001$ ) and unrelated competitor ( $\beta=-1.26, S E=0.19$, $z=-6.53, p<.001$ ) are significant and the difference between target and category competitor is marginally significant ( $\beta=-0.32, S E=0.17, z=-1.87, p=.06$ ).

In the concessive condition, the difference between target and category competitor, on the contrary, fails to reach signif-

[^318]icance $(\beta=-0.20, S E=0.17, z=-1.18, p=.24)$, whereas the target was looked at significantly more often than gender competitor $(\beta=-10.89, S E=0.18, z=-4.88, p<.001)$ and unrelated competitor $(\beta=-1.07, S E=0.19, z=-5.65$, $p<$.001).

Accuracies and Reaction Times for comprehension questions were analyzed the same way as eye-movements, except that we used linear regressions rather than logistic regressions for response times. While response times did not differ across conditions ( $\chi(1)=0.44, p=.51$ ), accuracy was significantly lower in the concessive condition (78\%) than in the causal condition ( $84 \% ; \chi(5)=11.17, p<.05$ ). More detailed analyses reveal that this difference was driven by the lower answer accuracy for those questions in the concessive condition in for which the correct answer was "yes".


Figure 2: Results for causal (top) and concessive (bottom) conditions.

## Discussion

These results clearly reveal that both causal and concessive discourse markers were integrated rapidly into on-line comprehension and that processing the concessive led to a search for alternatives.

In the causal condition, processing was rapid and stable enough to combine with grammar information to predictively identify the target referent. In the concessive condition, there is a similar tendency but it did not reach significance. This
may mean that processing concessives is more difficult and does not allow people to rapidly take gender marking into account. The result that looks to the target category exceed looks to the other objects later in the concessive condition than the causal condition may reflect slower processing in the concessive condition. However, since, in the causal condition, the objects belonging to the target category were already looked at most before the connector region, the finding cannot be clearly interpreted.

The finding that accuracy of question answering was worse in the concessive than the causal condition (when the correct answer was "yes') might suggest that processing in the concessive condition was shallower, causing a late cognitive burden for global interpretation. An alternative possibility is that suppressing the category directly mentioned in the second sentence (e.g., sweet) in combination with having to categorize the target (e.g., pretzel - salty) might be difficult (as in sweet... however.. pretzel - ... something salty?)

Experiment 2 evaluates whether our finding that discourse markers can be integrated rapidly, shaping predictions about upcoming words, can be replicated in a reading experiment.

## Experiment 2: Reading Study

## Methods

Participants We tested 30 participants, 6 of which had to be excluded due to eye-tracking problems. Data of 24 participants ( 5 male, average age 24) was analyzed.

Design, Materials \& Procedure Items for Experiment 2 consist of 24 three-sentences discourses, following a similar logic as the ones of Experiment 1. However, rather than reducing the set of possible predictions by providing a picture, a more strongly constraining first sentence introducing two scenarios is employed. The second sentence makes one of these two options more salient. The third sentence begins with either a causal or a concessive marker, followed by a region which determines the focus of the concessive, a pretarget region which contains case-marking, and the target noun region (see Example 2). The target noun is not used in the preceding context. Half of the sentences are congruent (e.g., head and ears cold - therefore - hat), and half incongruent (e.g., neck cold - however - scarf), resulting in a 2(causal/concessive) $\times 2$ (congruent/incongruent) withinparticipant design. All sentences are grammatically correct.

Lotte braucht für den Winter noch Kleidungsstücke um Kopf und Hals zu wärmen. An Kopf und Ohren friert sie besonders. Daher / Dennoch guckt sie als allererstes nach [einer schön warmen / einem schön warmen] pretarget [Mütze/Schal, die/der nicht zu bunt aussieht $]_{\text {target }}$.
Lotte needs clothes to keep her head and neck warm for the winter. Her head and ears feel particularly cold. Therefore / However, she first of all looks for [a nicely warm] gender-marked pretarget [hat / scarf that does not look too colorful $]_{\text {target }}$.

The 24 items were intermixed with 48 filler discourses, which followed the same pattern as the items but using a range of non-causal/concessive discourse markers (e.g., later, in particular). All items and half of the fillers were followed by yes/no-comprehension questions, asking about the target noun without referring to it by name (Schaut Lotte als erstes nach einem Kleidungsstück für den Kopf?, "Does Lotte first of all look for clothing for the head?"), answerable by button press. Half of the questions' correct answer was "yes", the other half's "no". We created 8 lists, according to the numbers of versions per item: 2(salience second sentence) $\times 2$ (causal/concessive) $\times 2$ (match/mismatch). Participants saw only one version of each item.

Discourses were presented on the center of the screen, divided into two parts: The first screen showed the first and second sentence and the second screen contained the target sentence. The question, if present, followed the discourse on a third screen. Reading was self-paced, controlled by button press. The order of presentation was pseudo-randomized with at least one filler in between two items. Participants were tested individually and their eyes were tracked. They were asked to read carefully to be able to correctly answer comprehension questions. The experiment lasted approximately 30 minutes.

Predictions Given our results from the first experiment, we hypothesized that people would be able to eagerly integrate discourse context and the the discourse connector to predict the target noun. This predicts a mismatch effect (as expressed in longer reading times: first pass durations, regressions, and total reading time) in the pretarget region when the grammatical gender of determiner and adjective does not match the grammatical gender of the predicted target noun. This mismatch effect is moreover predicted to continue in the target region.

## Data Analyses, Results \& Discussion

First pass durations, regression durations, and total reading times in the pretarget (determiner and adjective) and target (noun and final phrase) region were analyzed using linear regressions (see Experiment 1). Trials with track loss in more than one of all regions (Sentence 1, Sentence 2, discourse maker region, pretarget region, target region, question) and with reading times smaller than 50 ms were excluded from analyses.

For the causal condition, we found a consistent tendency for people to read more slowly in mismatching than matching sentences in both the pretarget and the target regions in all measures; none of these trends, however, reached significance. In the concessive condition, similar but weaker tendencies were found, but only for first pass reading times. While this could mean that discourse information cannot be integrated fast enough to give rise to prediction neither in causals nor in concessives, we considered the possibility that some of our items were not clear or constraining enough to enable readers to anticipate the target noun.

To still get an idea about the indicated difference between causal and non-causal contexts (i.e., using the causal condition as a baseline), we excluded those items for which no mismatch effect was observable in total reading times in the causal condition in the target region (i.e., when reading times were not higher for mismatches than matches, not even any time after the target noun was encountered). Based on the remaining 19 items, we found significantly longer reading times for mismatches in the pretarget region of causal sentences for all measurements (first pass $\chi(1)=5.38, p<.05$; total time: $\chi(1)=7.27, p<.01$; regression: $\chi(1)=4.99, p<$ .05).

However, even for these 19 predictable discourses, when the discourse relation was concessive, there was only a marginal effect of mismatch for first pass durations in the pretarget region $(\chi(1)=3.43, p=.06)$ but no further effects for the pretarget region (total time: $\chi(1)=.02, p=.89$; regression: $\chi(1)=0.07, p=.80$ ) and not even in the target region (first pass $\chi(1)=1.12, p=.29$; total time: $\chi(1)=0.02, p=$ .88; regression: $\chi(1)=0.15, p=.70)$.

For comprehension-question accuracy, there was no effect of condition (causal: $80 \%$, concessive: $82 \% ; \chi(1)=$ $1.40, p=.24)$, a marginal effect of mismatch $(\chi(1)=$ 3.57, $p=.06$ ) and a significant interaction $(\chi(1)=4.01, p<$ .05): Accuracy was significantly higher for match than mismatch only for causals $(\chi(1)=5.81, p<.05)$ but not for concessives $(\chi(1)=1.17, p=.68)$. For Reaction Times, there was no effect of match $(\chi(1)=0, p=1)$ nor condition $(\chi(1)=0, p=1)$, and no interaction $(\chi(1)=0.38, p=.54)$.

Experiment 2 therefore indicates that, given that the discourse is really clear and constraining, in causally related sentences, readers are generally able to make predictions based on quickly integrating discourse context. In other words, if the target noun was predictable, then it was predicted rapidly (i.e., in the pretarget region). Global interpretation, as well, was influenced by the congruency of the discourse, as indicated by results from question answering accuracy. For concessives, on the contrary, there is no consistent evidence, that either of this was the case.

## General Discussion

While results from Experiments 1 and 2 are not fully in line with one another, it is likely that prediction is easier with a constraining visual scene which is co-present during the entire discourse than with a linguistic context which is only read once and needs to be remembered and re-accessed. Moreover, while there was a prediction effect for concessives in the extended connector region in Experiment 1, prediction was also slower (or more slowly stable) in the concessive than the causal condition (no significant effect in the concessive case in the pretarget region). Possibly, processing concessives was simply more slowly than processing causals in both experiments.

Another possible explanation however is that the scope of the concessive is more ambiguous than the scope of causals
in experimental items of Experiment 2:
causal Timmy wants to do $A$ and B. A is more important.
Therefore .-
concessive Timmy wants to do $A$ and $B$. $A$ is more important. However .

In the causal case, the causal connector can only refer to the previous sentence "A is important", hence only $A$ is a sensible continuation. In the concessive case, the concessive marker might take scope either over the second sentence, in which case the prediction would be, as anticipated, However B. But, it is also possible for the concessive to take scope over both initial sentences, leading to a prediction However $C$, that is, Tommy goes on to do something entirely different. In that case, the space of possible predictions is wide open and cannot be expected to cause a gender mismatch effect.

For the visual world experiment, this difference in scope between causals and concessives is not an issue as the visual scene is very explicit. An interesting aspect about the hypothesis that concessive markers give rise to less specific predictions than causal markers is that it stands in apparent contrast to the hypothesis by Murray (1995) discussed earlier : Murray suggests that concessives are highly constraining while causal connectives are moderately constraining and leave more open hypotheses. We believe that Murray's and our hypotheses are not necessarily contradictory, however: While concessive markers may be less ambiguous with respect to the discourse relation they are marking (see also Asr \& Demberg, 2012a,b), they may at the same time be more ambiguous with respect to the scope of their argument.

Results from Experiments 1 and 2 are generally in line with studies revealing immediate interpretation of discourse markers (e.g., Traxler et al., 1997). However, our data also supports that negating a discourse relation (i.e., via adversative markers) may cause a delay in processing, at least when a directly mentioned state of affairs needs to be rejected and its opposite needs to be both mentally accessed and found (on a scene or in memory). That means that concessive discourse markers are a type of negation that can cause processing difficulties. Moreover, Experiment 1 supports that negation can give rise to a search for alternatives (Kaup et al., 2006).

## Conclusions

We investigated the time-course of processing marked causal and concessive discourse relations within two experiments. Results from a visual world experiment (Exp. 1) provide clear evidence that, at least in this highly constraining scenario, both causals and concessives can be processed incrementally and give rise to predictions. Concessives, specifically, elicit an active search for alternatives. A reading experiment (Exp. 2) confirms this finding for causals but not for concessives. Results of both experiments indicate that concessives may be more difficult to process than causals, causing a delay. In Experiment 2 , difficulties with concessives may also be due to ambiguity in scope.

## References

Asr, F. T., \& Demberg, V. (2012a). Implicitness of discourse relations. In Proceedings of coling. Mumbai, India.
Asr, F. T., \& Demberg, V. (2012b). Measuring the strength of linguistic cues for discourse relations. In Proceedings of the workshop on advances in discourse analysis and its computational aspects (adaca). Mumbai, India.
Baayen, R., Davidson, D., \& Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language, 59, 390-412.
Bates, D. (2005). Fitting linear mixed models in r. $R$ News, 5, 27-30.
Carpenter, P., \& Just, M. (1975). Sentence comprehension: A psycholinguistic processing model of verification. Psychological review, 82(1), 45.
Dale, R., \& Duran, N. (2011). The cognitive dynamics of negated sentence verification. Cognitive science, 35(5), 983-996.
Ferguson, H., Sanford, A., \& Leuthold, H. (2008). Eye-movements and erps reveal the time course of processing negation and remitting counterfactual worlds. Brain research, 1236, 113-125.
Kaup, B., Lüdtke, J., \& Zwaan, R. (2006). Processing negated sentences with contradictory predicates: Is a door that is not open mentally closed? Journal of Pragmatics, 38(7), 1033-1050.
König, E., \& Siemund, P. (2000). Causal and concessive clauses: Formal and semantic relations. TOPICS IN ENGLISH LINGUISTICS, 33, 341-360.
Kuperberg, G., Paczynski, M., \& Ditman, T. (2011). Establishing causal coherence across sentences: An ERP study. Journal of Cognitive Neuroscience, 23(5), 1230-1246.
Louwerse, M. (2001). An analytic and cognitive parametrization of coherence relations. Cognitive Linguistics, 12(3), 291-316.
Lüdtke, J., Friedrich, C., De Filippis, M., \& Kaup, B. (2008). Eventrelated potential correlates of negation in a sentence-picture verification paradigm. Journal of cognitive neuroscience, 20(8), 1355-1370.
Marslen-Wilson, W. (1973). Linguistic structure and speech shadowing at very short latencies. Nature, 244, 522-523.
Millis, K., \& Just, M. (1994). The influence of connectives on sentence comprehension. Journal of Memory and Language; Journal of Memory and Language.
Murray, J. (1995). Logical connectives and local coherence. Sources of Coherence in Reading, 107-125.
Nieuwland, M., \& Kuperberg, G. (2008). When the truth is not too hard to handle an event-related potential study on the pragmatics of negation. Psychological Science, 19(12), 1213-1218.
Sanders, T. (2005). Coherence, causality and cognitive complexity in discourse. In Proceedings/actes sem-05, first international symposium on the exploration and modelling of meaning.
Sanders, T., Spooren, W., \& Noordman, L. (1992). Toward a taxonomy of coherence relations. Discourse processes, 15(1), 1-35.
Sanford, A., \& Garrod, S. (1998). Incrementality in discourse understanding. The construction of mental representations during reading, 1.
Staab, J. (2007). Negation in context: Electrophysiological and behavioral investigations of negation effects in discourse processing. ProQuest.
Tanenhaus, M. K., \& Trueswell, J. (1995). Sentence comprehension. In J. Miller \& P. Eimas (Eds.), Handbook of perception and cognition, vol 11: Speech, language, and communication (2nd ed., pp. 217-262). San Diego, CA: Academic Press.
Townsend, D. (1983). Thematic processing in sentences and texts. Cognition, 13(2), 223-261.
Traxler, M., Bybee, M., \& Pickering, M. (1997). Influence of connectives on language comprehension: eye tracking evidence for incremental interpretation. The Quarterly Journal of Experimental Psychology: Section A, 50(3), 481-497.
Traxler, M., Sanford, A., Aked, J., \& Moxey, L. (1997). Processing causal and diagnostic statements in discourse. Journal of Experimental Psychology: Learning, Memory, and Cognition, 23(1).

# The role of morphology in spelling: Long-term effects of training 

Kendall L.D. Kolne (kendall.kolne@mail.mcgill.ca)<br>Katherine Hill (katherine.hill@mail.mcgill.ca)<br>Laura M. Gonnerman (laura.gonnerman@mcgill.ca)<br>School of Communication Science and Disorders, McGill University, 1266 Pine Ave W<br>Montreal, QC, H3G 1A8 CAN


#### Abstract

We compared the effectiveness of two spelling interventions: one focused on morphological structure and one emphasizing word meanings,, on spelling acquisition in French speaking children in $3^{\text {td }}$ and $5^{\text {th }}$ grades. The morphology intervention led to significantly greater improvement in spelling than the vocabulary intervention, especially for children in grade 5 . To compare the long-term effects of the two interventions, we tested the children's spelling ability six-months after the conclusion of the intervention program. Results show that both grades maintain an increase in spelling accuracy compared to their pre-intervention performance. Additionally, the children in grade 5 who received morphological instruction retained more spelling knowledge than those who received the vocabulary instruction. These results suggest that teaching children about the structure of complex words supports their spelling ability in the long-term, providing evidence for the importance of morphological knowledge in literacy development.


Keywords: morphology; spelling; literacy development; vocabulary; intervention; French

## Introduction

Learning to spell is a critical aspect of literacy development, yet research has typically focused on the development of reading skills. Understanding the process of learning to spell has become particularly important in Quebec, where a widespread decline in children's spelling ability has become apparent (Jalbert, 2007). Contributing to this decline is the difficult nature of French spelling. French has a one-tomany mapping of sounds-to-orthography, so the same sound may be written in a number of different ways. Additionally, silent letters are common in written French, so children must learn to spell parts of words for which there is no overt pronunciation to guide them. These features of written French make learning to spell in this language a complex task.

Recent evidence suggests that literacy instruction focused on morphological knowledge, or on the ability to recognize and process sub-lexical units in language (e.g., recognizing that the word reheatable is made up of three sub-parts, the prefix re-, the stem heat, and the suffix -able) may assist children's spelling development. In fact, children who have greater metalinguistic awareness of morphological structure are better able to spell words correctly (e.g., Deacon, Kirby, \& Casselman-Bell, 2009; Sénéchal, 2000) and teaching children explicitly about the morphological relationships
between words improves their reading and writing skills (see Bowers, Kirby, \& Deacon, 2010, for a review).

While morphological awareness training may be a beneficial teaching method for fostering literacy development, there are a number of important issues to be resolved to ensure that children receive the most effective instruction. Firstly, most of this evidence is derived from studies of English-speaking children, and little is known about the contribution of morphological skills to writing ability in French (cf. Sénéchal, 2000; Sénéchal, Basque, \& Leclaire, 2006; Pacton \& Deacon, 2008). French has a richer morphological system than English, so it is likely that morphology may have an even more influential role in learning to spell in French. Intervention studies with French-speaking children are needed to test this hypothesis.

Additionally, children as young as two to three years demonstrate knowledge of morphology (Berko, 1958; Clark, 1993, Gonnerman, 2007), but it is not clear when this knowledge begins to influence spelling ability. Some researchers have argued that morphological knowledge has an early influence as children begin to develop literacy skills (e.g., Deacon \& Kirby, 2004), while others report that the influence of morphological knowledge on spelling ability does not have a large impact until later in development (e.g., Carlisle, 1995; Kirby et al., 2012; Singson, Mahony, \& Mann, 2000). To provide the most effective instruction to children, it is crucial to understand the most appropriate stage of development to introduce morphological training.

Typically, instruction of morphological structure also involves discussion of word meaning, because morphologically related words share similar form and meaning. Previous studies have yet to investigate the distinction between morphological and vocabulary instruction (e.g., St-Pierre \& Dubé. 2012), thus the relative contribution of morphology versus semantics to improving spelling ability is unknown. To disambiguate the potential benefit of morphological knowledge from the benefits of word meaning instruction, it is necessary to isolate the teaching of morphological structure and compare its effects on spelling outcomes to that of vocabulary training.

Finally, it is important to find out whether the benefits of a morphological intervention program can be maintained across time, and whether the knowledge will transfer to new words not taught in the intervention. Carlisle (2010) conducted a review of instructional programs using morphological awareness training to improve literacy outcomes, and reported that the majority of these studies fail
to report the long-term maintenance of the effects, or the transfer of learning to new words. It is critical to evaluate both the maintenance and transfer of learning to ensure that a morphological intervention provides children with longlasting abilities beyond the context of the intervention.

We have conducted an intervention study to investigate the role of morphological training for improving spelling in Quebec French. In a previous study, we analyzed and reported the results immediately after the conclusion of the intervention. The focus of the present study is to examine the long-term effects of the intervention, as measured at a follow-up session six months after the conclusion of the intervention. We compared the long-term effects of morphological instruction for $3^{\text {rd }}$ graders and $5^{\text {th }}$ graders, explicitly contrasting its relative contribution to spelling ability with that of vocabulary instruction. Thus, our research question is two-fold:

1. Is there a difference in relative long-term intervention effectiveness by grade? That is, will a morphology intervention improve long-term spelling performance of children in grade 3 versus 5 ?
2. Is there a difference in long-term intervention effectiveness by instruction method? That is, will a morphology intervention lead to great long-term spelling improvement than a vocabulary intervention?

In the sections that follow, we describe the intervention that was conducted, as well as the spelling outcomes following the intervention for children in grades 3 and 5 To address our research questions, we present data from a sixmonth follow-up test evaluating the long-term effectiveness of the morphology and vocabulary training for improving spelling performance.

Overall, we expect that the children will experience some degree of forgetting, such that their spelling accuracy at the six-month follow-up will be lower than at post-intervention; however we expect that the children will retain some of the spelling knowledge from the intervention, so their spelling scores at the six-month follow-up will be higher than at the pre-intervention. Moreover, we predict that the greater benefit observed for the morphology intervention will be maintained in the long-term.

## The Present Study

We developed an intervention to target the spelling of a set of morphologically complex words, with emphasis on either morphology or vocabulary instruction. The present study aims to assess the long-term outcomes of our spelling intervention. Six months after the intervention ended, we went back to the school and administered the same spelling test to the children who had participated in the intervention. The children's performance on this test at the six-month follow-up will be compared to their performance on the test
as measured before the intervention as well as immediately after the intervention.

## Methods

## Participants

Eighty-four children were recruited from one elementary school in the greater Montreal area and took part in the intervention. Children from two Grade 3 and two Grade 5 classes within the school participated. The primary language of instruction in this school is French. 36 children from Grade 3 participated ( 23 girls and 13 boys), as well as 48 children from Grade 5 ( 27 girls and 21 boys).

Children were randomly assigned to one of the two treatment groups, based on their general spelling abilities prior to their participation in the intervention study. General spelling ability was assessed using a modified version of the Test Ortho3 from the Batterie d'Évaluation du Langage Écrit et de ses troubles (BELEC) (Mousty, Leybaert, Alegria, Content, \& Morais, 1994). Children in each intervention group were also matched on language background (monolingual Francophone, or multilingual), and gender, with approximately equal ratios of boys to girls in each treatment group.

## The intervention

Children in grade 3 and grade 5 took part in the intervention. The children were divided into two groups, one which received instruction explicitly focused on the morphological structure of the words to be learned (Morphology group), the other receiving instruction focused on the meanings of the words (Vocabulary group). For example, the Morphology group was taught that there are two parts to the word finlandais, namely the stem finland and the suffix -ais, while the Vocabulary group was taught that the word finlandais describes something or someone that comes from the country, Finland. The children were taught to spell an identical set of 30 words, with only the emphasis of instruction differing across intervention groups. The intervention was given during 10 weekly sessions, each lasting one hour.

Ten suffixes were taught in the intervention. The suffixes were relatively frequent and productive in Quebec French, such that they are preferentially used to form new words. Three words were chosen containing each of the 10 suffixes, creating the list of 30 words that were taught in the intervention. These words were relatively infrequent, so it would be unlikely that the children in grade 3 or 5 would already know these words.

The 30 words were distributed across the 10 intervention sessions, with three words taught per session. In each session, the children in the Morphology group were taught the three words with the same suffix. For the Vocabulary group, words with the same suffix were distributed across the 10 sessions, such that the words with the same suffix were never taught in the same session. For example, in the first session, the Morphology group was taught finlandais,
japonais, and camerounais, whereas the Vocabulary group was taught ogresse, huileux, and galanterie. Thus, each group was taught the same words, just in different sessions.

## Materials for assessing intervention effectiveness

We developed a test to determine the effectiveness of the intervention on children's spelling ability. This test was administered before (pre-intervention), immediately after (post-intervention), and six months after the intervention concluded (six-month follow-up). We designed this spelling test to measure specific outcomes from our intervention. The test assessed the spelling of complex and simple words, and required children to generalize stems and suffixes taught in the intervention to new words not taught in the intervention. The items on the test were either the exact complex word taught in the intervention (i.e., a taught stem and a taught suffix), a taught or an untaught stem without a suffix, or a combination of a taught/untaught stem and suffix in a complex word (i.e., a taught stem with a new suffix, or a new stem with a taught suffix).

## Procedure

All students took the spelling test in the classroom at the same time. The instructor read each sentence once, repeating the missing words as many times as necessary for all students to fill in the missing word. The instructor was a female native speaker of Quebec French.

## Results and Discussion

We assessed the effects of our intervention immediately following the conclusion of the intervention program, analyzing the changes in spelling performance from pre- to post-intervention. Before we report the results of the sixmonth follow up, the pre- to post- test analyses will be summarized. As the focus of the present study is the longterm spelling outcomes, only statistics including the sixmonth follow-up scores will be reported in this paper.

There were 15 children who participated in the original intervention who were absent from the six-month follow-up session. These children were excluded from the following analyses. Additionally, 3 children were absent from either the pre- or post-intervention assessment, and these children were also excluded from the following analyses.

The children's performance on the spelling test was scored based on whether the whole words were spelled correctly, and also whether the stems and suffixes of complex words were spelled correctly. Accordingly, each complex word received three scores, one for the whole word, one for the stem, and one for the suffix. Mean percent correct scores on the whole words, stems, and suffixes were calculated for the following analyses.

## Question \#1: Is there a difference in relative longterm intervention effectiveness by grade?

Pre- to post- intervention summary We compared the changes in spelling accuracy over all the items on the
spelling test, from pre- to post-intervention, for grade 3 and 5 students. The results of this analysis revealed that children in both grades improved their spelling from pre- to postintervention, with children in grade 5 scoring higher overall than those in grade 3. However, the children in grade 3 showed a greater differential between pre- and postintervention than those in grade 5, indicating that the children in grade 3 were aided more by the intervention, irrespective of the type of instruction.

To test whether these differences remained six months after the intervention, we calculated mean percent correct at each test time. These mean scores for grades 3 and 5 are displayed in Figure 1. We entered the whole word accuracy scores on all of the spelling test items into a $2 \times 3$ ANOVA with the factors Grade (grade 3 or grade 5) and Test Time (pre-intervention, post-intervention, or six month post) to assess the long-term effects of the intervention for each grade. The main effect of Grade was significant, $F(1,64)=$ $16.98 p<.001$, indicating that the children in grade 5 scored significantly higher than the children in grade 3 . The main effect of Test Time was also significant, $F(2,128)=174.92$, $p<.001$, as was the interaction of Grade and Test Time, $F(2,128)=6.73, p=.002$, indicating significant differences between the spelling performance of grade 3 and 5 children across the three testing sessions.


Figure 1: Overall mean percent correct on the spelling test, for grade 3 and grade 5 at pre-intervention, post-intervention and the six-month follow-up.

Post-intervention to six-month follow-up To specifically examine the potential differences in the long-term effects of the intervention for grade 3 and 5 children, a planned comparison of the whole word accuracy scores for all items, with the factors Grade (grade 3 or grade 5) and Test Time (post-intervention or six month post) was conducted. The results show that the grade 5 children had significantly higher spelling scores than the grade 3 students from postto six month post-intervention, $F(1,64)=11.55, p<.001$. Collapsing across both grades, scores were significantly higher at post-intervention than at the six month follow-up, $F(1,64)=12.01, p<.001$, indicating that the children had forgotten some of the spelling knowledge they gained from the intervention six months later. Interestingly, the
interaction of Grade and Test Time was not significant, $F(1$, $64)=.35, p=.55$, indicating no difference between grade 3 and grade 5 in the amount of spelling knowledge that was forgotten. In fact, there was only a small, albeit significant, decrease in spelling ability six months after the intervention, approximately $5 \%$ in each grade.

Pre-intervention to six-month follow up To ensure that six months later the children retained much of the spelling knowledge they originally gained from the intervention, we conducted a planned comparison of the whole word spelling accuracy scores of all items, with the factors Grade (grade 3 or grade 5) and Test Time (pre-intervention or six month post-intervention). Once again there was a significant main effect of Grade, $F(1,64)=17.57, p<.001$, such that the children in Grade 5 scored higher than those in Grade 3. The main effect of Test Time was significant, $F(1,64)=193.01$, as was the interaction between Grade and Test Time, $F(1,64)=10.85, p=.002$. These results indicate that children in both grades maintained their spelling improvement, scoring higher at the six-month follow-up than at pre-intervention. Moreover, the children in grade 3 improved more from pre-intervention to the six-month follow-up than the children in grade 5. Thus, the children display long-term learning, having retained a large amount of the spelling knowledge that they gained from the intervention six months later.

## Question \#2: Is there a difference in long-term intervention effectiveness by instruction method?

Pre- to post- intervention summary Given the differences between grades in intervention effectiveness, we analyzed pre- to post- intervention differences between the Morphology and Vocabulary group for each grade
separately. In general, children in both instructional groups increased from pre- to post- intervention, indicating that both types of instruction effectively improved children's spelling ability for both $3^{\text {rd }}$ and $5^{\text {th }}$ graders. Looking more closely at the accuracy for stems and suffixes of the test items, differential effects according to intervention group emerged, with the Morphology group showing a larger increase in spelling accuracy than the Vocabulary group.

The results immediately following the intervention suggest that the instruction focusing on the morphological structure of words provides an advantage to children over an intervention that focuses on word meanings. Specifically, children who have had morphological-based training were able to generalize the knowledge they gained in the intervention to be able to correctly spell morphologically related words that had not been taught directly. While the Morphology group showed differential improvements over the Vocabulary group in both grades, the morphological intervention provided the strongest benefit for children in grade 5.

To determine whether the advantage of a morphological intervention over a vocabulary intervention for learning to spell was maintained after a period of no instruction, we compared the changes in spelling accuracy of the two intervention groups from immediately after the intervention to the six-month follow-up assessment. Additionally, we compared the long-term effects of the morphology and vocabulary instruction for $3^{\text {rd }}$ and $5^{\text {th }}$ grade separately, to determine the developmental stage for which the spelling intervention is most effective. Each grade was thus examined separately in the following analyses.. The mean percent correct on the complex words, stems and suffixes for both intervention groups are displayed in Table 1 for Grade 5, and in Table 2 for Grade 3.

Table 1. Grade 5 mean percent correct on complex words, stems and suffixes at post-intervention and six-month follow-up.

|  | Morphology Group |  |  |  |  | Vocabulary Group |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Postintervention |  | Six-month follow-up |  | Mean <br> Difference | Postintervention |  | Six-month follow-up |  | Mean Difference |
|  | M | $S D$ | M | $S D$ |  | M | $S D$ | M | $S D$ |  |
| Complex Words | 83.33 | 17.25 | 74.31 | 17.40 | -9.02 | 86.84 | 15.29 | 68.42 | 21.40 | -18.42 |
| Stems | 86.96 | 9.66 | 78.99 | 13.52 | -7.97 | 80.78 | 15.76 | 75.06 | 16.00 | -5.72 |
| Suffixes | 91.67 | 7.11 | 88.19 | 8.27 | -3.48 | 93.42 | 6.41 | 83.55 | 13.54 | -9.87 |

Table 2. Grade 3 mean percent correct on complex words, stems and suffixes at post-intervention and six-month follow-up.

|  | Morphology Group |  |  |  |  | Vocabulary Group |  |  |  | Mean Difference |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Postintervention |  | Six-month follow-up |  | Mean Difference | Postintervention |  | Six-month follow-up |  |  |
|  | M | $S D$ | M | $S D$ |  | M | $S D$ | M | $S D$ |  |
| Complex <br> Words | 81.25 | 16.08 | 63.39 | 21.63 | -17.86 | 78.33 | 21.89 | 65.00 | 16.50 | -13.33 |
| Stems | 72.98 | 15.27 | 67.70 | 13.82 | -5.28 | 73.91 | 18.00 | 68.12 | 19.28 | -5.79 |
| Suffixes | 88.39 | 7.70 | 81.70 | 13.97 | -6.69 | 79.58 | 19.11 | 76.67 | 16.61 | -2.91 |

Performance on complex words We first looked at the long-term changes in whole word spelling accuracy of the complex words that were taught in the intervention. The whole word scores for the complex taught words were entered into a separate ANOVA for each grade, with the factors Intervention Group (morphology or vocabulary) and Test Time (post-intervention or six-month postintervention). Looking first at the results for grade 5, the main effect of Test Time was significant, with children scoring higher at the post-test session, than the pretest session, $F(1,35)=21.98, p<.001$. The main effect of Group was not significant, $F(1,35)=.05, p=.81$, nor was the interaction of Test Time and Group,, $F(1,35)=2.52, p=$ .12. Thus, both groups display some forgetting of how to spell the complex words that were taught in the intervention, but this change is not differential based on the intervention group.

For the $3^{\text {rd }}$ graders, the main effect of Test Time was once again significant, $F(1,27)=20.68, p<.001$, while the main effect of Group was not significant $F(1,27)=, p=.92$. Unlike the pattern observed in the $5^{\text {th }}$ grade, the interaction of Group and Test Time was not significant, $F(1,27)=.44$, $p=.51$. For children in grade 3 , after six months both groups showed a similar decrease in spelling accuracy for the complex words taught in the intervention.

Performance on stems To assess the long-term effects of instruction on the spelling of taught stems, mean percent correct scores for taught stems were entered into an ANOVA with the factors Test Time (post-intervention or six-month-post intervention) and Group (morphology or vocabulary), for each grade separately. The results for the $5^{\text {th }}$ grade children showed a significant main effect of Test Time, $F(1,35)=12.70, p=.001$, but not a significant main effect of Group, $F(1.35)=1.44, p=.24$, nor an interaction between Test Time and Group $F(1,35)=.35, p=.56$.

Similarly, in the $3^{\text {rd }}$ grade, the main effect of Test Time was significant, $F(1,27)=9.68, p=.004$, while the main effect of Group and the interaction of Test Time and Group were not, $F(1,27)=.01, p=.91, F(1,27)=.02, p=.89$, respectively. For both Grade 3 and Grade 5, performance on the taught stems decreased somewhat for both the morphology and vocabulary groups, but this small decrease was the same across both groups. Thus, the initial learning based on the intervention resulted in approximately 21-31 percent increases in spelling of the stems, and after 6 months, both groups still showed significant improvements in spelling, only dropping 1 to 6 percent in their scores.

Performance on suffixes We compared the long-term effects of the two intervention types on the spelling of suffixes taught in the intervention. For each grade, the mean percent correct scores for taught suffixes were entered into separate ANOVAs, with the factors Test Time (postintervention or six-month post-intervention) and Group (morphology or vocabulary). For grade 5, the main effect of Test Time was significant, $F(1,35)=18.22, p<.001$, while
the main effect of Group was not, $F(1,35)=.30, p=.56$. Interestingly, the interaction of Group and Test Time was significant $F(1,35)=4.08, p=.05$, revealing that six months after the intervention, the morphology group showed greater retention for the spelling of taught suffixes. This finding suggests that for children in grade 5 , instruction focused on morphological structure is more beneficial in the long-term for learning to spell morphologically complex words than instruction focused on word meaning.

The analysis for grade 3 children showed that the main effect of Test Time was marginally significant, $F(1,27)=$ $4.00, p=.06$, and that the main effect of Group was not, $F(1,27)=1.86, p=.18$. In contrast to Grade 5, the interaction of Test Time and Grade was not significant for Grade $3, F(1,27)=.64, p=.43$. There is a slight decrease in the spelling of taught stems at the six-month follow-up for both intervention groups, and this decrease is not different by intervention received. Given the differing pattern of results for performance in the spelling of taught suffixes, with the $5^{\text {th }}$ graders in the morphology group showing greater retention, the morphology-based instruction seems to provide an advantage over a vocabulary-based instruction for learning to spell at later stages of literacy development.

## General Discussion

The present study evaluated the long-term effectiveness of a morphology-based intervention for elementary school-aged French-speaking children. The intervention contrasted the effects of a training program focused on the morphological structure of words, with one that concentrated only on word meaning. While other intervention studies have confounded morphology and vocabulary instruction (see Bowers, Kirby, \& Deacon, 2010, for a review), our study design allowed us to disambiguate the relative benefits of morphology and vocabulary instruction for spelling outcomes. Additionally, by conducting the intervention with children in $3^{\text {rd }}$ and $5^{\text {th }}$ grade, we could assess the effects of morphological instruction at different stages of literacy development.

While both interventions led to significant spelling improvements from pre- to post-intervention, the Morphology group displayed significantly greater improvement in their ability to generalize their spelling knowledge beyond the words that were taught in the intervention. The differential benefit in favour of the morphology group was particularly pronounced for the children in grade 5 . Overall, the results suggest that teaching children about morphological structure successfully improves spelling accuracy more than instruction based on word meaning does.

In addition, in the results reported here, we demonstrate the long term learning effects of the morphology intervention by re-examining the children after a six-month delay. We found that for both the morphology and vocabulary groups, the improvement in spelling accuracy remains six months later, as the children spell significantly better at the six-month follow-up than at pre-test. These effects hold for children in both grades 3 and 5 . The
children do display some forgetting at the six-month mark, with scores significantly decreasing from post-intervention to six-month follow-up, however, the decreases were very small (approximately one to six percent) and there were no differences in the amount of forgetting between grades. This finding suggests that, regardless of instruction type, children benefit from our spelling intervention.

Importantly, when examining the differential effects of instruction type, we found a significant, long-term advantage for grade 5 children in the Morphology group over children in the Vocabulary group. At the six-month follow-up, those who received morphology instruction showed greater retention of spelling knowledge than those who received the vocabulary instruction. Our intervention study and the subsequent follow-up suggest that morphological training provides sustained improvement to children's spelling accuracy in French, greater than instruction on word meaning, particularly for older elementary school-aged children.

## Conclusion

Findings from our follow-up study provide support for an advantageous role of morphology instruction for spelling outcomes in Quebec French. Explicitly teaching children about the components of complex words helps them to spell stems and suffixes better, and to generalize their knowledge beyond the words taught in the intervention. For older children, these effects are maintained well after instruction is finished, indicating that morphology instruction would be a useful tool for dealing with the spelling difficulties observed in Quebec. While we did not see the same differential long-term benefit of morphology training in the younger children, our findings indicate that both types of intervention were very beneficial in the long-term. As such, an intervention combing instruction of morphological structure and vocabulary knowledge may be especially helpful for these children.

## Acknowledgments

This research was supported by a grant awarded to the third author from the Fonds de recherche du Québec Société et culture (FRQSC). We would also like to thank Eve Bergeron and Rebecca Amzallag for their help with testing. Finally, a huge debt of gratitude goes to the school director, teachers, and students for their enthusiastic participation in the intervention study.

## References

Berko, J. (1958). The child's learning of English morphology, Word, 14, 150-177.
Bowers, P.N., Kirby, J.R., \& Deacon, S.H. (2010). The effects of morphological instruction on literacy skills: A systematic review of the literature. Review of Educational Research, 80, 144-179.

Carlisle, J.F. (2010). Effects of instruction in morphological awareness on literacy achievement: An integrative review. Reading Research Quarterly, 45, 464-487.
Carlisle, J.F., \& Stone, C.A. (2005). Exploring the role of morphemes in word reading. Reading Research Quarterly, 40, 428-449.
Clark (1993). The lexicon in acquisition. Cambridge: Cambridge University Press.
Deacon, S.H., \& Kirby, J.R. (2004). Morphological awareness: Just "more phonological"? The roles of morphological and phonological awareness in reading development. Applied Psycholinguistics, 25, 223-238.
Gonnerman, L.M. (2007). Children's recognition of novel derived words. In H. Caunt-Nulton, S. Kulatiake, \& I. Woo (Eds.). Proceedings of the $31^{\text {st }}$ Annual Boston University
Jalbert, P (2007). L'épreuve obligatoire d'écriture de la fin du troisième cycle du primaire en français, langue d'enseignement: Comparaison des résultats de 2000 et 2005. Ministère de l'Éducation, du Loisir et du Sport http://www.mels.gouv.qc.ca/lancement/TablePilotage_Pro gFormation/Ecriture3eCyclePri mFLE.pdf.
Kirby, J.R., Deacon, S.H., Bowers, P.N., Izenberg, L., Wade-Woolley, L., \& Parrila, R. (2012). Children's morphological awareness and reading ability, Reading and Writing, 25, 389-410.
Pacton, S., \& Deacon, S. H. (2008). The timing and mechanisms of children's use of morphological information in spelling: A review of evidence from English and French. Cognitive Development, 23, 339-359.
Sénéchal, M. (2000). Morphological effects in children's spelling of French words. Canadian Journal of Experimental Psychology, 54, 76-86.
Sénéchal, M., Basque, M.T., \& Leclaire, T. (2006). Morphological knowledge as revealed in children's spelling accuracy and reports of spelling strategies. Journal of Experimental Child Psychology, 95, 231-254.
Singson, M., Mahony, D., \& Mann, V. (2000). The relation between reading ability and morphological skills: Evidence from derivational suffixes. Reading and Writing, 12, 219- 252.
St-Pierre, M-C., \& Dubé, J-F. (2012). Morphological awareness intervention: Does severity play a role in spelling/reading improvement? Aula Abierta, 40, 15-22.

# How big is the BFG? The impact of redundant size adjectives on size perception 

Emiel Krahmer (E.J.Krahmer@uvt.nl)<br>Tilburg center for Cognition and Communication (TiCC), School of Humanities, Tilburg University, The Netherlands

Marret K. Noordewier (M.K.Noordewier@fsw.leidenuniv.nl)<br>Faculty of Social and Behavioral Sciences, Social and Organizational Psychology, Leiden University, The Netherlands

Martijn Goudbeek (M.B.Goudbeek@uvt.nI)<br>Tilburg center for Cognition and Communication (TiCC), School of Humanities, Tilburg University, The Netherlands

Ruud Koolen (R.M.F.Koolen@uvt.nl)<br>Tilburg center for Cognition and Communication (TiCC), School of Humanities, Tilburg University, The Netherlands


#### Abstract

In two experiments we study how redundant size modifiers influence the perceived size of objects. We show that when objects are referred to with overspecified descriptions (for example, using a description like "the large red chair" in a situation where all chairs are equally large but different in color), participants subsequently estimate the object to be larger than when objects are referred to using minimally distinguishing descriptions (e.g., "the red chair"). In Experiment 1, we show this effect with adult language users and different kinds of size modifiers. In Experiment 2, the same effect is shown for children of two different age groups (7- and 10-year olds), and for different kinds of visual size contrasts. Interestingly, we observe an inversely proportional relation between the age of our child participants and the difference in size estimates for minimal and overspecified descriptions, suggesting that language users gradually become better at avoiding false pragmatic inferences from redundant adjectives as they grow older.


Keywords: Reference, Overspecification, Language Development, Conversational Implicature

## Introduction

Arguably, referring to a giant as "big" is somewhat excessive, certainly when there is only one giant in sight. But would calling a giant "big" and "friendly" (as Roald Dahl, 1982, does in his well-known children's novel The $B F G$; short for Big Friendly Giant), nevertheless have an impact on the perceived size (or friendliness) of said giant? And would this effect be the same for younger children as for older ones or even for adults? These are essentially the questions we address in this paper.

## Background

Speakers frequently produce definite descriptions such as "the big friendly giant", "the red chair" and "the large ball",
since they allow them to link their utterances to the physical world surrounding them. One central problem that a speaker has to solve when planning such a referring expression is to decide which properties to include in the reference. A chair can be red, but also large, plush, modern, with or without cushions and armrests, cheap or expensive, etc. So which properties to select? A successful reference includes sufficiently many properties to allow the addressee to determine which chair the speaker has in mind, but not too many, as Dale and Reiter (1995) propose in their computational interpretation of Grice's (1975) maxims for reference production.

One prima facie plausible option would be to opt for the smallest set of properties that distinguish the target object from the other objects in the context (Dale 1989). However, it has been repeatedly found that this is not necessarily what speakers do (e.g., Olson, 1970; Pechmann, 1989; Belke \& Meyer, 2002; Engelhardt et al., 2006; Koolen et al., 2011). In many cases, speakers produce overspecified descriptions, which contain one or more redundant modifiers. For example, they produce a description such as "the large red chair", in a situation where "the red chair" would have been sufficient to single out the target. A number of speakerinternal factors have been shown to influence the likelihood of speakers producing an overspecified description, ranging from the pressures of incremental speech production (speakers may start producing a referring expression before scanning of a visual scene is complete; Pechmann, 1989) to scene complexity (more overspecification in complex visual scenes; Koolen et al., 2012) and conversational setting (more overspecification when misunderstandings are costly; Arts et al., 2011), suggesting that overspecification does not have a single distinct cause.

However, in this paper we focus on the impact of overspecification on language understanding, and here the picture is less clear. Some researchers have suggested that
overspecification can help addressees with identifying an intended target (e.g., Arts et al., 2011; Paraboni et al., 2007), while others argued that they slow down identification (e.g., Engelhardt et al., 2006; 2011). Importantly, all these studies focus only on object identification. In this paper, we argue that overspecification in referring expressions may also have other important side effects for addressees.

A prominent view in language understanding is based on the assumption that "utterances convey only relevant information" (Frank \& Goodman, 2012). This assumption can be traced back at least to the work of Grice (1975), who postulates among other things that speakers should not make their contribution more informative than is required (this is half of his well-known Maxim of Quantity). A speaker that violates ("flouts") this maxim, by providing more information than needed, thereby triggers a conversational implicature, suggesting to the listener that the additional material is meaningful after all. Imagine, for instance, that a speaker tells you to "sit by the newly-painted table" (Dale \& Reiter, 1995), while there is only one table in the room. In that case, you may think the modifier "newly-painted" is redundant (since it is more informative than required for the purpose of identification), and this might cause you to infer the conversational implicature, intended by the speaker, that it is best not to rest your arms on this table in order to keep your clothes unstained.

However, one can also think of situations where an addressee may reason that the redundant information is somehow relevant, even when the speaker did not intend it in this way (Grice would call this a false conversational implicature). After all, as we argued above, speakers may overspecify for a variety of reasons. Our first hypothesis therefore is that if an object is described redundantly as "large" or "small" this will influence how the size of this object is perceived. Redundantly referring to a target object as large (small) may cause people to perceive or remember the target as larger (smaller) than when such a redundant size modifier is not included in a description. Even though we focus on redundant size adjectives here, we conjectured that other redundant adjectives (e.g., referring to color) could have similar effects (a possibility we discuss below).

Moreover, it has been argued that children are more likely to derive false conversational implicatures than adults (Siegal \& Surian, 2004). On the one hand, we know from earlier research that children have a general tendency to regularly produce underspecified or ambiguous referring expressions until they are about seven years old (Deutsch \& Pechmann, 1982; Matthews, Lieven, \& Tomasello, 2007), and before that age only marginally benefit from redundant information in target identification (Sonnenschein, 1982; Ackerman, Szymansi \& Silver, 1990; Davies \& Katsos, 2010). Indeed, one could argue that it requires relatively sophisticated pragmatic reasoning to understand the implications of redundant information in descriptions. Therefore our second hypothesis is that children are more susceptible to redundant size modifiers than adults when making size estimates. Given that earlier work suggests that
the relevant pragmatic reasoning is under development until children are about 7 years old, we test this both with child participants of on average 7 years (Group 3 in the Dutch elementary school system) and 10 years old (Group 6) in the experiment.

## The current studies

We test these two hypotheses in two experiments (one with adults, one with children of two age groups), which rely on the same basic idea: participants hear descriptions referring to objects in a visual scene. Descriptions can either be minimally specified or overspecified (containing a redundant size modifier). After participants have processed a description, the objects in the visual scene disappear from view, and participants are asked to indicate how large they think the target object (which no longer is visible) was.

## Experiment 1: Adults

## Method

Participants Participants were 68 undergraduate students from Tilburg University (49 female) who participated for partial course credits. Their mean age was 21.5 years ( $\mathrm{SD}=$ 2.4). All were native speakers of Dutch, the language of the experiment.


Figure 1: Example scene consisting of two same size chairs, with different colors (may not be visible in a black and white print). In the minimal condition, the target would be referred to as "the red chair", while in the overspecified condition it would be "the large red chair".

Materials Stimuli were created using pictures of furniture items from the Object Databank, created by Michael Tarr and colleagues, and often used in research on reference (e.g., van Deemter, Gatt, van der Sluis \& Power, 2012). Three furniture items were selected for the current experiment (chair, couch, desk), and manipulated for color (either red or blue) and size (either large or small). Each stimulus consisted of two objects, one target (the object being referred to) and one distractor. There were 12 different targets ( 3 object types x 2 colors x 2 sizes), and the left-right position of the target with respect to the distractor was counterbalanced. In the critical stimuli, the distractor was always of the same type and size as the target and only differed in color (see Figure 1 for an example). Each target was referred to once with a minimal description (e.g., "the red chair") and once with an overspecified description containing a redundant size adjective (e.g., "the large red chair" or "the small red chair", depending on the size of the target). This created a total of $12 \times 2=24$ critical trials. The
experiment also contained 24 filler trials, in which the visual objects consisted of different types, sizes and colors.

Procedure During the experiment, individual participants were seated in front of a computer screen, on which pairs of objects were presented as in Figure 1, together with a prerecorded spoken description (e.g., "the [large] red chair"), produced by a female speaker with neutral intonation (i.e., with nuclear stress on the noun and no pitch accent on the adjective) and presented to participants over headphones. After a fixed interval, both objects disappeared from the screen and a horizontal slider appeared, together with the question "How large was the $\qquad$ ?", where the gap in this question was filled by the type of the target (e.g., "chair") The slider had the shape of an elongated, isosceles triangle with the tip ("small") on the left- and the base ("large") on the right-hand side (see Figure 2). Upon appearance, the slider handle was positioned in the middle; the handle had to be moved before the participant could proceed to the next stimulus. For analysis, the position after being set by the participant was mapped to a score between 0 and 100 (with higher number indicating larger size estimates).

In addition to the size question, participants were also asked to indicate the color of each object referred to on a one-dimensional saturation scale, ranging from lighter to darker, again with the handle initially positioned in the middle, on the assumption that a redundant mention of color (like "red") would cause participants to perceive an object as "redder" than when color was not mentioned in a description. However, no reliable effects of redundant color adjectives were found, and we will not describe the results of this measure further. In the general discussion we do return to this issue.

Experiment 1 had a within-participants design: all participants produced a size estimate for all targets. Stimuli were presented in a random order. Before the actual experiment started, a three trial training session (with a fan as target object type) was presented, to make participants familiar with the experimental set-up. After the training session there was no further interaction between participants and experimenter.


Figure 2: Slider used in Experiment 1 for size estimates.
Statistical analysis To test for significance, we conducted a $2 \times 2$ repeated measures Analysis of Variance (ANOVA), with Size (levels: large, small) and Description (levels: minimal, overspecified) as independent variables and average size estimate as the dependent variable.

## Results and discussion

Figure 3 summarizes the results. First of all, a main effect of Size was found, $F(1,67)=461.94, p<.001, \eta^{2}=.87$. Large targets were estimated to be larger $(M=60.45,95 \% \mathrm{CI}=$ (57.83, 63.07)) than small ones $(M=23.17,95 \% \mathrm{CI}=(20$, 26.34)). This serves as a manipulation check and indicates that the slider worked exactly as intended. In addition, a main effect of Description was found, $F(1,67)=5.30, p<$ $.05, \eta^{2}=.07$. Targets that were referred to using an overspecified description were estimated to be larger than targets that were referred to using minimal descriptions. Importantly, this main effect was qualified by an interaction between Size and Description, $F(1,67)=15.16, p<.001, \eta^{2}$ $=.18$. This interaction can be explained by inspection of Figure 3: large targets that are referred to redundantly are estimated to be larger $(M=62.81,95 \% \mathrm{CI}=(59.85,65.76)$ than ones that are referred to minimally ( $M=58.1,95 \%$ CI $=(55.34,60.85))$, while small targets that are referred to redundantly are estimated to be smaller ( $M=22.10,95 \% \mathrm{CI}$ $=(18.88,25.33))$ than ones that are referred to minimally $(M=24.24,95 \% \mathrm{CI}=(20.86,27.62))$.


Figure 3: Mean size estimates in millimeters (range 1-100) for large and small objects for minimally specified descriptions (e.g., "the red chair") or overspecified ones ("the large/small red chair").

Experiment 1 clearly showed that adults are sensitive to redundant size modifiers in distinguishing descriptions. In Experiment 2, we conduct a comparable experiment with child language users in two different age categories, to see whether younger and older children are similarly sensitive to redundant modifiers. In addition, in this experiment we also vary the visual size of the target, to see whether size differences between target and distractor influence any effects of redundant size modifiers.

## Experiment 2: Children

## Method

Participants Sixty normally developing children were included in the study, in two age groups: 30 younger children ( 13 girls, 17 boys), with an average age of 7.1 years (range: 6.6-8.3), all in Group 3 of the Dutch elementary
school system; and 30 older ones ( 15 girls, 15 boys), with an average age of 10.2 years (range: 8.7-11.5), all in Group 6 . The children participating in this experiment were native speakers of Dutch, the language of the experiment. All children came from the Mgr. Zwijsenschool in Kerkdriel (Gelderland, The Netherlands). Parental consent for participation in the experiment was obtained for all children prior to the experiment.

Materials Pictures of eight different photorealistic children's toys (football, teddy bear, train, slide, rubber duck, doll, boat, spin top) were used as targets in this experiment (see Figure 4 for two representative examples). Each target was presented together with one distractor toy. Four different conditions were created for each of the eight targets by varying the size of the distractor (which could either be depicted as large as or smaller than the target) and by varying the reference to the target (which could either include a redundant size modifier or not), resulting in $8 \times 4$ $=32$ critical trials. Note that in this experiment each target could uniquely be identified by its type ("the football"), so including a size adjective always resulted in an overspecified description. In addition, eight control trials were included, one for each target type, in which the target was combined with a smaller object of the same type (e.g., a small football), so that the size adjective in a description such as "the large football" was informative and not redundant. This allowed us to check whether any differences in size-estimates for non-redundant adjectives between age groups could be observed. In all 40 stimuli the left-right position of the target with respect to the distractor was counterbalanced.


Figure 4: Examples of visual stimuli used in Experiment 2, with two different children's toys. Again the target could either be referred to in a minimal way ("the ball") or an overspecified one ("the large ball").

Procedure Children performed the experiment individually, and were seated in front of a computer monitor in a quiet room in the school building. The procedure for younger and older children was exactly the same and went as follows: after a brief training session, in which the magnitude estimation scale was practiced, stimulus presentation started with a pair of toys presented on a white background for 4 seconds. After this a white screen appeared for 8 seconds, during which children were asked to answer a pre-recorded question "How large was $\qquad$ ?" The gap in this question was filled by a description of the target, which could either be minimal ("the football") or overspecified ("the large football"). For the audio recordings, a male adult speaker
was used, who realized each question with a neutral intonation. Since the experiment was conducted in Dutch, this implies that the nuclear stress always occurred on the noun and the adjective was produced without a pitch accent.

Children were asked to indicate their size estimate on a magnitude estimation scale of 100 millimeters (consisting of a horizontal line without units of length added), with on the left-hand side a picture of the target reduced by a factor of 1.5 , and on the right hand side the same picture enlarged by a factor of 1.5 , in such a way that the real value was exactly in between (remember that the target figure was not visible to the child during the size estimation phase of the experiment, so children could not directly map the perceived size onto the scale). Children could indicate the estimates on paper, using a booklet that was positioned in front of them. After completing one trial, the next pair of toys appeared on the screen. The entire experiment lasted approximately 10 minutes. After the experiment, size estimates were manually measured in millimeters, with higher numbers indicating larger sizes. Measurements were done blind for condition.

Design and analyses The experiment had a $2 \times 2 \times 2$ mixed design, with Description (minimal, overspecified) and Size (target and distractor equally large, target larger than distractor) as within-participant factors, Age group (younger, older) as a between-participant factor and average size estimate as the dependent variable. Tests for significance were conducted using a repeated measures ANOVA.


Figure 5: Mean size estimates in millimeters (range 1-100) of younger (avg. 7.1 years) and older (avg. 10.2 years) children for targets that were minimally specified (e.g., "the ball") or overspecified ("the large ball").

## Results and discussion

The results showed that Size had a significant impact on children's size estimates. If the distractor was smaller than the target, children perceived the target as larger ( $M=47.3$, $95 \% \mathrm{CI}=(42.39,52.22)$ ) than when both target and distractor had the same size $(M=44.2,95 \% \mathrm{CI}=(39,6$, 49.6)), $F(1,58)=5.22, \mathrm{p}<.05, \eta^{2}=.08$. In other words, size
estimates are relative, even though the target always had the same size. Crucially, we also found a significant effect of Description on size estimates. If a target was referred to with a redundant size modifier, children perceived it as larger $(M=50.5,95 \% \mathrm{CI}=(44.99,56.01))$ than when the reference did not include such a modifier ( $M=41.43$, $95 \%$ $\mathrm{CI}=(36.98,45.88)), F(1,58)=47.03, \mathrm{p}<.001, \eta^{2}=.45$. This effect was independent of whether the target was visually larger than the distractor (no significant interaction between Description and Size was found).

Interestingly, we did find a significant interaction between Description and Age, revealing that older children are less sensitive to redundant size modifiers than younger ones, $F(1,58)=5.02, \mathrm{p}<.05, \eta^{2}=.08$. This interaction is illustrated in Figure 5, showing that when the target is minimally referred to ("the football") the size estimates of younger children $(M=41.7,95 \% \mathrm{CI}=(35.4,48.0))$ were almost the same as those of older ones ( $M=41.2,95 \% \mathrm{CI}=$ ( $34.86,47.45$ )), while overspecified descriptions ("the large football") caused younger children to make larger estimates $(M=53.74,95 \% \mathrm{CI}=(45.94,61.54))$ than older ones $(M=$ $47.26,95 \% \mathrm{CI}=(39.48,55.06))$. When we conducted a separate analysis over the eight additional items which were referred to using non-redundant modifiers, no significant age differences in size estimates were found either, suggesting that it is indeed only redundant size modifiers for which younger children are more sensitive.

No further significant main effects or interactions were found.


Figure 6: Average difference in size estimate in millimeters for redundant and minimal description ("the large $X$ " minus "the $X$ ") as a function of age (comparing younger children, older children and adults).

## General discussion

In two experiments we have shown that when target objects are referred to using descriptions containing redundant size modifiers (e.g., "the large red chair" in a situation where two chairs are equally large but different in color), participants subsequently estimate the object to be larger
than when objects are referred to using minimally distinguishing descriptions (e.g., "the red chair"). In Experiment 1, we showed this effect with adult language users and with two different size modifiers ("large" and "small"). In Experiment 2, the same effect is shown for children of two different age groups (7- and 10-year olds), but this time with descriptions for object of different types (e.g., a ball and a teddy bear). Interestingly, this effect was found to be independent of whether the target was actually larger than the distractor or not; in both cases, a redundant size modifier had a comparable effect.

Even though the two experiments were slightly different in the way they were conducted (e.g., furniture targets and digital size estimates in Experiment 1 versus children's toys and size estimates on paper in Experiment 2), the essential idea was the same: participants had to process object description which were either overspecified or minimal, and after the objects had disappeared from view, they were asked to estimate the perceived size of the target object that had just been referred to on a scale from 1 to 100 . Therefore, it is interesting to plot the difference in size estimates for overspecified and minimal description as a function of age, as is done in Figure 6. Inspection of this figure reveals a clear trend. We already saw in Experiment 2 that 7-year olds were more sensitive to redundant information than 10 -year olds, but the pattern for the 10 years olds seems comparable to that of the adults in Experiment 1. A univariate ANOVA confirms this: overall, there is a significant effect of age on the difference in size estimates, $F(2,125)=5.91, p<.01, \eta^{2}=.09$, but pairwise comparisons using the Bonferroni method showed that only the younger children differ significantly from the other two age groups. This appears to be consistent with the earlier work (cited in the introduction) showing that children younger than 8 still regularly produce referring expressions that may be underspecified and do not benefit from overspecified descriptions in target identification.

In this study we concentrated on redundant size modifiers, and the question naturally arises whether different kinds of adjectives could have similar effects. Our experiences with color adjectives in Experiment 1 suggest that this may not be the case. In particular, hearing a redundant description of a target as "the large red chair" (when both chairs are red) did not cause participants' to perceive the target as 'redder' than when hearing a minimal description ("the large chair"). Potentially, this could be due to the one-dimensional saturation slider that was used (after all, colors differ along multiple dimensions, also including lightness and hue, and it is not entirely clear which corresponds to, say, 'redness'). Alternatively, it could be due to the fact that color is an absolute property, while size is a relative one (e.g., Rips \& Turnbull, 1980). An adjective like "large" implies a comparison (an object is only large compared to another object), which may explain why participants are more likely to modify a size than a color estimate. We leave this as an issue for future research.

Our current results suggest that including redundant modifiers in a referring expression can be used strategically. If a speaker subtly wants to emphasize a property of a target, she can just mention it in a distinguishing description, irrespective of whether it rules out any distractors. In fact, mentioning any redundant property may help in attracting attention to a particular target. Koolen, Krahmer and Swerts (2012), in a study with children from two age groups similar to the ones under study here, found that when children were offered a choice between two identical looking sweets, they opted significantly more often for the one which was referred to in a redundant way ("this red sweet") than for the one that was minimally described ("this sweet"), and even thought the former would taste better than the latter. Interestingly, this effect was found to be stronger for younger than for older children, confirming that as they grow older, and their pragmatic skills increase, children are less likely to be influenced by overspecification.

## Conclusion

Wrapping up, we can state that calling a giant both "big" and "friendly" will make him seem larger than merely calling him "friendly", although the size of this effect is presumably inversely proportional to the age of the addressee.

## Acknowledgments

Many thanks to Ellen Evers for her help with setting up Experiment 1, and Maud Broekmeulen for her help with conducting Experiment 2. We received financial support from The Netherlands Organization for Scientific Research, (NWO, Vici grant 277-70-007), which is gratefully acknowledged.

## References

Ackermann, B., Szymanski, J. \& Silver, D. (1990). Children's use of common ground in interpreting ambiguous referential utterances. Developmental Psychology 26, 234-245.
Arts, A., Maes, A., Noordman, L. \& Jansen, C. (2011). Overspecification facilitates object identification. Journal of Pragmatics, 43, 361-374.
Belke, E., \& Meyer, A. (2002). Tracking the time course of multidimensional stimulus discrimination: Analysis of viewing patterns and processing times during "same""different" decisions. European Journal of Cognitive Psychology, 14, 237-266.
Dahl, R. (1982). The BFG. London: Jonathan Cape.
Dale, R. (1989). Cooking up referring expressions. In: Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 6875.

Dale, R., \& Reiter, E. (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. Cognitive Science, 18, 233-263.

Davies, C., \& Katsos, N. (2010). Over-informative children: Production/comprehension asymmetry or tolerance to pragmatic violations? Lingua, 120, 1956-1972.
van Deemter, K., Gatt, A., van der Sluis, I. \& Power, R. (2012). Generation of referring expressions: assessing the Incremental Algorithm. Cognitive Science, 36 (5), 799836.
van Deemter, K., Gatt, A., van Gompel, R., \& Krahmer, E. (2012). Towards a computational psycholinguistics of reference production. Topics in Cognitive Science, 4(2), 166-183.
Deutsch, W. \& Pechmann, T. (1982). Social interaction and the development of definite descriptions. Cognition, 11, 159-184.
Engelhardt, P., Bailey, K. \& Ferreira, F. (2006). Do speakers and listeners observe the Gricean maxim of quantity? Journal of Memory and Language, 54, 554-573.
Engelhardt, P.E., Demiral, S.B., \& Ferreira, F. (2011). Over-specified referential expressions impair comprehension: An ERP study. Brain and Cognition, 77, 304-314.
Frank, M., \& Goodman, N. (2012). Predicting Pragmatic Reasoning in Language Games. Science, 336, 998.
Grice, H. (1975). Logic and conversation. In: Cole, P. \& Morgan, J. L. (Eds.), Speech Acts. Academic Press, New York, 41-58.
Koolen, R., Gatt, A., Goudbeek, M., \& Krahmer, E. (2011). Factors causing overspecification in definite descriptions. Journal of Pragmatics, 43, 3231-3250.
Koolen, R., Goudbeek, M., \& Krahmer, E. (2012). The effect of scene variation on the redundant use of color in definite reference. Cognitive Science, to appear.
Koolen, R., Krahmer, E., \& Swerts, M. (2012). Developmental changes in children's processing of redundant information in definite object descriptions, submitted.
Matthews, D., Lieven, E. \& Tomasello, M. (2007). How toddlers and preschoolers learn to uniquely identify objects for others: a training study. Child Development, 78, 1744-1759.
Olson, D.R. 1970. Language and thought: Aspects of a cognitive theory of semantics. Psychological Review, 77, 257-273.
Paraboni, I., van Deemter, K. \& Masthoff, J. (2007). Making referents easy to identify. Computational Linguistics, 33, 229-254.
Pechmann, T. (1989). Incremental speech production and referential overspecification. Linguistics, 27, 89-110.
Rips, L. \& Turnbull, W. (1980). How big is big? Relative and absolute properties in memory. Cognition, 8, 145174.

Siegal M. \& Surian, L. (2004). Conceptual development and conversational understanding. Trends in Cognitive Sciences, 8, 534-538.
Sonnenschein, S. (1982). The effects of redundant communication on listeners - when more is less. Child Development, 53, 717-729.

# The case of cognitive ecology for cognitive processes in everyday life situations 

Mattias Kristiansson<br>Department of Computer and Information Science, Linköping University<br>SE-58183 Linköping, Sweden


#### Abstract

Cognitive ecology is a term that has been used in environments that are more tightly coupled and purposespecific than environments of everyday life. In this paper I consider cases from a cognitive ethnography of older adults. These cases show the analytical use of understanding the diachronic and synchronic cognitive ecology in which cognitive processes of everyday life occur. Specifically I discuss how the social and physical ecology and changes in these can shape goals, the use of cognitive artifacts and the use of other cognitive resources in agent environments that are not as purpose created and not as tightly coupled as environments of previous studies in this field.


Keywords: Cognitive ecology; distributed cognition; everyday life; older adults

## Introduction

This paper elaborates on the notion of a cognitive ecology applied to the domain of older adults coping with cognitive problems and situations in everyday life. Examples from a cognitive ethnography of older adults will be analyzed. The reason for doing this is to shed some light on what a cognitive ecology can be in a social and physical environment that is not as tightly coupled or information dense as the cases where the concept of cognitive ecology have been used previously. By focusing on older adults I hope to understand how circumstances in everyday life can constrain, shape and alter the use of certain cognitive strategies that assist and therefore become important for the understanding of the cognitive process. This analysis will have a specific focus on ecology, as contrasted to the idea that the agent actively shapes the cognitive process. I believe this is important because it allows us to understand the role of the active agent more firmly in an environment that often is not as tightly structured, with a specific goal or purpose as a navigational bridge (Hutchins, 1995) or an early modern theatre (Tribble, 2011). First I turn to the concept of cognitive ecology and then I briefly turn to the tension between the idea that individuals contribute to the cognitive process and the idea that cognitive process are shaped by the circumstances. Finally several aspects of cognitive ecologies through the light of examples from the conducted cognitive ethnography are discussed.

## Cognitive ecology

The understanding of cognition in relationship to environmental factors has now been a prominent undertaking in cognitive science for a while. The term "cognitive ecology" is now occasionally used to describe
the study of cognition in context, emphasizing the general notion that cognition is something taking place and developing in an ecology that constrains, alters and forms cognitive processes (Hutchins, 2010a; Tribble \& Sutton, 2011; Tribble, 2011). Tribble (2011, p.151) held that the idea of distributed cognition and the approach of cognitive ecology are basically the same. I will not here assess this statement, but in this paper I view it as a continuum between what can be seen as a distributed cognitive process and what formed, constrained, or altered this process. In this paper I want to focus on the latter aspect. "Cognitive ecology" has mostly been used in the field of animal cognition where the focus is on how the ecology shapes intracranial cognitive process (Dukas \& Ratcliffe, 2009). In this paper I use the term cognitive ecology to explain something that also shapes processes that incorporates both intracranial and extracranial resources.

Hutchins (2010a) notes that cognitive ecology both can be viewed from a synchronic perspective (that is functional relationships in the present), and a diachronic perspective (that is cognition as development of cognitive ecologies). Much of the research into this field has focused on what goes on in the present without saying much about the developmental aspect of the cognitive process or mediated action (Sutton, Harris, Keil, \& Barnier, 2010; Wertsch, 1998). This difference is important because what can constrain the use of resources is not always found in the present, "on the spot" (Hutchins, 2010b). We live and are shaped by cultural practices that to some extent determine the ways we "do things". Clark (2008) also emphasizes the understanding of interaction between different systems and specifically the continuous reciprocal causation between these systems. A key foundation for these related principles is the understanding of how the processes of constraining, altering and forming cognitive processes occur. For instance, why does someone use a particular artifact in a certain way? The answer can be found outside the individual and the specifics of the artifact.

Cognitive ecology suggests a unit of analysis that focuses on "units defined in terms of dynamic patterns of correlation across elements" (Hutchins, 2010a, p.705). What the correct unit of analysis should be to explain cognitive phenomenon is therefore not given before we have some understanding of the synchronic and diachronic ecology of where the phenomenon takes place.

Tribble (2011) used an ecological approach when she studied and historically analyzed theatre practices in the English renaissance theatre. The objective for her analysis was to explain the impressive performance of individuals
performing up to six different plays in a week with irregular and limited practicing time. One explanation of such a fact could easily be that these people had amazing memory abilities. But she proposed a larger dynamical model of this memory performance where it is important to understand the ecological differences that exist between theatre practices today and theatre practice back then. Back then "much preparation was individual, facilitated by the individual parts containing only the character's lines and his cues." (p.14). Much of this success was also facilitated by the ecological niche of the physical and social environment, where for instance parts were written to suit less experienced actors.

In her analysis she introduces the term "cognitive thrift", which is a principle that suggests that in a highly cognitively demanding environment, such as the theatre in this era, "every incentive would have been to minimize any additional cognitive burdens" (Tribble, 2011, p.32). In her conclusion she notes that cognitive ecologies "place more or less weight in internal mechanisms, on central control, or on particular forms of cognitive artifacts and social systems" (p.153). It can be argued that doing comparative studies of cognitive ecologies allow cognitive scientists to understand the relative contribution of different parts of a system to uphold reasonable performance.

In the case of healthy older adults coping with everyday life, it is not as easy to say that this is a highly cognitively demanding environment. Older adults cope well with everyday life in comparison with their performance in labsettings, and one suggested explanation of this is that older adults seldom need to perform at their cognitive maximum in their normal life (Salthouse, 2012). How something as a cognitive ecology works in a setting where performance in a specific way is not often as demanding is to my knowledge rather unexplored. A term such as "cognitive thrift" might not apply in this context. This is because when we talk about cognitive ecology and distributed cognition we often do so in the domain of so called cognitively rich environments where a slight change in the ecology can profoundly shape the process and the performance.

The question that follows is what a cognitively rich environment is? In the case of early modern theatre or on a navigational bridge it can be interpreted as a measure of how much information that flow across various media, which directly relates to the problem at hand. How densely does the information flow across the various (tightly coupled) media to solve problems in everyday life is for me still an open question. Neither can we easily say that the environments and processes that take place within these environments of everyday life are task specific since these environments often have multiple purposes. How a cognitive ecology can form, constrain and alter cognitive processes and to some extent predict (according to some measure) successful performance in the lives of older adults would not necessarily be based on the same principles as in the highly demanding environments.

## The ecology as opposed to the active individual

As noted above, one point made by Tribble (2011) is that even though memory demands were high in the early modern theatre practices, much of this taxing work was not solely placed on individual cognitive abilities. Much of this pressure was left to various aspects of the overall physical and social system of the theatre in work. The tension between the idea of an active individual and a shaping society or environment has been around for some time in various scientific fields (Wertsch, 1998). In cognitive science many have argued that too much emphasis has been put on the individual, placing to many cognitive abilities simply inside the skull as default (c.f. Hutchins, 1995). Wertsch (1998) argues that this is a question without an obvious answer since answers to this question are often not based on empirical grounds. In this paper I use one side of this dichotomy, the circumstances that shape, as an analytical tool to understand important aspects of the process.

Wertsch (1998), by adopting the "pentad" proposed by Burke (1969), uses a further elaborate analytical tool in his focus on mediated action as the unit of analysis. The pentad consists of act, scene, agent, agency, purpose, or in Wertsch:s words "What? Where? Who? How? and Why?" (p.13). The point is not that these are true reflections of reality; it is rather that they are tools for the interpretation of reality. Focusing on mediated action can be understood as emphasizing certain parts of the pentad and de-emphasizing other parts. The scene is for instance not included much in an analysis of mediated actions (Wertsch, 1998). But on the other hand Wertsch argues that focusing on mediated action allows us to be in the middle of an individual and collective/distributed perspective. The agent and her mediational tools (see cognitive tools) are irreducible to each other in terms of the action. In my examples below I will use the idea of a scene as something that realizes and in a true sense constrains and alters the cognitive process in certain directions and therefore also sometimes alters the mediated action. The scene is here part of the cognitive ecology that Tribble uses in her analysis.

Wertsch (1998) also focuses on the fact that mediational tools have often been developed for other purposes than the reasons they are used for in the present. Therefore he emphasizes investigation of both consumption of mediational means and production of mediational means. Regarding the production of mediational tools he notes that tools are often borrowed from other sociocultural contexts and that the processes of what he calls a spin-off of actions with certain mediational tools are not always developed from a clear purpose of an inventor.

Even if we in this description find the notion of an agent that borrows and produce mediational tools, the idea of the pentad suggests that we can analyze what is not physically part of the agent and the tool and say something about how the agent and her mediational tools became orchestrated in an action or in a distributed cognitive process.

The role of the individual can theoretically be pictured as a continuum from a top-down driven agent to a bottomdriven agent to factors that stand outside the role of the agent but still support the cognitive process. Clark (2006) talks about ecological control as something we do when we do not micromanage every point in the process but still search for opportunities. When we do not micromanage the process, much of what constrain the process is outside the individual's scope of control. Certain processes have been developed to suit certain ecologies. To illustrate this I now turn to cases from a conducted cognitive ethnography of elderly people.

## Ethnography of everyday life

Previous research in cognitive aging suggests that older adults actually do have an active role in their compensatory practices for declining memory abilities (c. f. Frias, Dixon, \& Bäckman, 2003). Through self-reports older adults often report that they adopt external memory aids and cooperation with social others (c. f. Cavanaugh, Grady, \& Perlmutter, 1983; Frias et al., 2003). With such premises, even though they are based on self-reports, it is worth asking to what extent such a wide unit of analysis as cognitive ecology is applicable and at all important in these kinds of less problem-centered environments.

The material referred to below was collected as part of a cognitive ethnography during the summer of 2010. The scenery of this is in the home healthcare system where I worked as a healthcare assistant. Within this context I conducted interviews outside the role as an assistant, and observation in the role as an assistant. Most of the participants had normal cognitive functioning for their age and some had diagnosed memory declines. The specifics for each case are provided with the examples (but for more information see Dahlbäck, Kristiansson, and Stjernberg, 2013).

The following sections are categorized according to conclusions I can draw from the specific examples presented, together with the overall material collected in relationship to earlier theorizing in cognitive ecology (Hutchins, 2010a; Tribble, 2011) and earlier presented ideas of Wertsch (1998).

First I consider the general case that environmental factors together can enact certain cognitive processes. Second I relate cognitive ecologies of everyday life in relationship to the use of cognitive artifacts. Third I discuss the social nature of everyday life and how these social circumstances form cognitive processes. Fourth I note that the ecology can form the goal of cognitive processes and last I discuss how we can understand diachronic processes by understanding how ecologies shape cognitive processes.

## Environmental factors enact the nature of cognitive processes

A is 91 years old with a normal cognitive decline for his age. He has problems with hearing and particularly seeing.
(All examples in this paper are verbatim translations from Swedish from my original field notes.)
"He tells me that he goes to the grocery store almost every day: "there is always something you need and there is also a seating arrangement where there is always someone you know from previous work places". [...] When I ask him if he writes shopping lists he says that he doesn't and that he remembers everything in his head. He pictures how he usually goes through the important places in his home before he goes to the store, checking whether something is missing." (Excerpt from A)

His troubles with seeing were apparent at other times during this interview. The case notes how processes of remembering can be (a) constrained, in this case by his seeing impairment, (b) altered, by the fact that he lives rather close by the grocery store and (c) motivated by the fact that going to the store (almost) every day also has a social incentive. If his cognitive ecology would have been different in terms of social network, physical surrounding and limitations, his processes of remembering could have been distributed differently.

A is also aided by his routine of going through the usual places in his home where things often are missing. By doing this he provides himself with a mental anchor for remembering what was missing at the particular places at home. Partly because he more likely can recall what he found missing, but also because certain places constrain what he can possibly need. In a sense he has invented the method of loci himself. This together with the fact that he goes to the store almost every day makes his process of purchasing groceries a resilient one (c. f. Hollnagel, Woods, \& Leveson, 2006). His strategy works rather well in this specific ecology.

Doing ethnography in the context of the home healthcare services creates a special kind of cognitive ecology. This is because the ethnographer is in many cases part of the cognitive process. Since distributed cognition emphasizes the social aspect of cognition being participatory shapes the phenomenon that we try to investigate. Consider the next entry.
"A large part to achieve the smoothest possible performance is to know by heart what routines apply to what person. Of course there is a general routine of logging into the system, saying hi etc. [...] But to do it as smoothly as possible you need to know what the home environment looks like and the viewpoint of this service from the perspective of the person. Where should the socks hang? Where is the medicine locker? How do you prepare a sandwich in the correct way? (Excerpt from B)

This case is also a about the order of doing things. The smoothest possible performance is about coordinating work and to together remember what to do where, in what order, when. "She didn't recognize me because it was the first time I was there. She started pointing at the medicine locker and asked if I had the keys. [...] When I told her that I had the keys she rose and walked to the other room, apparently to
let me take off the socks before taking the medicine". (Excerpt from B)

Notice how the information of me having the keys initiates a more complex routine where the medicine is not the first goal of the routine. Even if this entry is also about how the work environment taxes the cognitive processes of the worker it also highlights an interesting cognitive ecology where it is important that all actors have a somewhat similar picture of how the activity should develop. If this shared picture is not the case this is indeed a cognitively demanding situation for both parts. But here the ecology of the home healthcare services provides with some structure regarding the predetermined goals of the assistance that the assistant know and can use to adapt to the circumstances of the visit. The receiver also adapts and initiates question, but overall the situation in the case above is cognitively taxing since the ecology is not as when the more experienced workers arrive.

In the case of B (as perhaps opposed to A ) the practice between a home healthcare assistant and the healthcare receiver is a rather predefined practice with certain goals, which have been established over several iterations of the service across several assistants. The receiver has adopted a general routine that works in the ecology of this service, a kind of a "cognitive thrift". The routinized coordinated practice is in this sense more equal to how the cognitive ecology is shaping the cognitive processes on a navigational bridge or in a theatre.

## How the cognitive ecology can shape the roles of artifacts

The case of $C$ below shows something similar to the case of the home healthcare setting above, but this case also shows how the role of a cognitive artifact, in this case a shopping list, are given an unspecified or a degraded role when used in a new cognitive ecology. This entry is from the first time she receives shopping assistance due to a physical problem.
"C uses a shopping list for the shopping session. She makes it clear that it is important for her that she remembers paracetamol as she has none at home and is in some pain. She constantly consults the shopping list to remind herself where to go. In the end, we cannot find paracetamol. I am not used to this supermarket, so I am of no help. She stops and asks a worker, who tells her that it is to be found after the check-out. She wants me to remind her if she forgets. After the check-out she has indeed forgotten, so I remind her." (Excerpt from C)

From the perspective of Wertsch (1998) the mediated remembering through this shopping list has been transformed to a mediated remembering both through the shopping list and a social other. Using the perspective of a cognitive ecology makes it possible to predict that we need to view this activity from different units of analysis. Information is propagated mostly between the subject and her shopping list and also between the subject and her assistant; but also to a lesser degree between the assistant
and C:s shopping list. The idea that artifacts exist with a functional relationship to their ecology has also been noted by Garbis (2002) that studied a tightly coupled cooperative process management setting. Remember that the task described above as defined through the home healthcare service was not about remembering things, it was about the physical challenge grocery shopping entails. But nevertheless the activity provides a certain kind of cognitive ecology that provides certain kinds of resources, that in this case inevitably creates a kind of process.

It is possible to view this process from two perspectives. One is that the individual must be active in this process, choosing resources and utilizing the resources sufficiently to perform reasonable good. The other, as noted, is to emphasize the circumstances that give rise to the role and utilization of resources.
" $D$ has memory problems and cannot always remember whether the home healthcare personnel have been on their visit to her, so she keeps the used time and day-specific plastic medicine envelope on her kitchen table after it has been used as a way of helping her to assure herself that they have been there that day. For this visit, she comes running after me as I am about to throw away the plastic envelope in the bin." (Excerpt from D)

The case is that this envelope has information so that it works as memory trace of previous activities. In this case it is worth noting, despite her memory problems, the active role of D to achieve good performance (Dahlbäck et al., 2013). She takes a cognitive artifact developed for one purpose and uses it in a different context for a different purpose (compare Wertsch, 1998).

But again this is also in a relationship to the cognitive ecology and how it realizes the use of an artifact in an efficient way. The experienced and the inexperienced home healthcare worker create different cognitive ecologies taxing mental resources of parts of the system differently. Under normal circumstances this cognitive system is a rather stable one. On the other hand part of the normal cognitive ecology is that there are different agents in operation creating a normal variation in the system. Another thing about the ecology in this case that is worth noting is that she comes running when she hears me throwing away the plastic envelope. She lives in a relatively small apartment and can therefore not be too far away from the action taking place. In this case she hears a sound from the bin that usually is not there. The artifacts in the cases of C and D have certain existences due to ecological factors.

## The social happenstance

I have already talked about the understanding of the cognitive process in the home healthcare services as a special kind of ethnography since the ethnographer is literally part of the cognitive process. In cases when people have a pronounced cognitive decline, that idea is not very strange since they are in the home healthcare system for that reason. But most of the time (at least in this specific unit) people are not in the home healthcare system because of
cognitive problems, but for a variety of physical problems. The point is that much of what in specific situations has formed the cognitive process is not always part of the cognitive process, but can be considered part of the cognitive ecology.

Consider the case of E, who each morning calls a few of her sisters to simply update the status of their lives. The fact that this happens each day tells us that they are rather good at keeping track of each other. The routine gives an arena for distributed processes of remembering. The fact that they are calling each other each morning and having a social environment that allows for such communication is part of their cognitive ecology. This is a kind of cognitive arena since it likely shapes the nature of the communication, which in turn shapes the distributed processes as they arise.

In a similar way we could view the case of $C$. In an interview with A I asked him in relation to his seeing problems what happens when they re-arrange in the grocery store, whereupon he quickly answers that "there is always someone that you can ask about the location of things". In the future detailed studies of situations where the social arena works as a resource can be of importance if we want to understand how individuals utilize this arena sufficiently.

## The smoothest possible performance

Part of understanding the ecology of cognitive phenomenon is to understand what a reasonable performance is for the particular subject or group. In the English Renaissance theatre it was the "smoothest possible performance" and not necessarily perfect recall (Tribble, 2011). In the case of $A$ above it is possible that the smoothest possible performance is not to perfectly recall all the groceries needed each day. For A, depending on the importance of the grocery, forgetting to buy something one day includes a new possibility to remember to purchase it the next day. A process-oriented view on memory deemphasizes the product of what to remember. To understand how humans remember we need to look on the process of remembering (c. f. Dixon, 1999). One important aspect of this is that this is from the perspective of the scientist conducting her research. The product from the subjective perspective in real-life settings can in a very true sense be the most important aspect. Consider the case of E .
" $E$ has an appointment at the podiatrist. She has a note from the podiatrist which she has posted on her fridge. She has turned the note around and written the date again, though bigger this time. She has also noted this in her calendar, located on the kitchen table. This calendar is always located on the kitchen table. For some unknown reason, the dates have gotten mixed up, and the wrong date has been transferred to her calendar." (Excerpt from E)

This is an interesting example since it shows the usage of different external sources for the same information. We can note that remembering appointments can be considered a highly important task to perform perfectly on since it is maneuvered with so many different resources. The smoothest possible performance is in this case, as in contrast
to the grocery shopping, perfect recall. The point here, similar to what Tribble (2011) noted, is that the social and cultural environment to some extent determines what "the smoothest possible performance" is. This is also an example of how the understanding of the mediated action gives insight to the nature of the purpose of the action (Wertsch, 1998).

Consider once again the case of A . What would happen if A would have a longer way to the grocery store? An apparent consequence would be that A:s physical limitations would be strained and therefore he may decide to not go to the store at all. But imagine that he still would manage despite the larger physical demand, and perhaps decide to walk there every second day. Would not his cognitive processes be composed in a different way? It would at least change what the smoothest possible performance would be. If one had a long road to the grocery store one would not as likely want to forget to buy something. Further, if the goal of the activity changes, the process will likely also change. Perhaps his loci-inspired remembering would be backed up with a shopping list despite his problems with seeing. This is of course an imagined world, but not an unlikely world.

## Cues of a diachronic process

As previously mentioned, an understudied part of cognitive ecologies is the diachronic perspective; that is how cognitive processes develop over time. The discussion of B shows how the study of the ecology can give insight into how cognitive processes develop. If one aspect of the ecology changes the process may also do so. Consider the next case of E :
"E demonstrates how to clean spoons discoloured by tea with the help of baking powder. E stands by the sink while the daughter and I are sitting by the kitchen table. [...] The daughter notices that $A$ uses the wet spoon in the powder container: "you can't do that, it will ferment". E answers quickly, and suggests that it will not ferment and will not be used for baking: "yes I can, because it is old baking powder". When E returns the container to the cupboard, the daughter remarks that she shouldn't place it next to the active baking powder. E rebels against her daughter's suggestion and places it next to the active container. She stops for a moment and lifts it a couple of times and says that she will anyway pay attention to and remember by the weight that it is the right one. The daughter remarks that at some point the containers will be of the same weight and they will be indistinguishable. E adds that anyway she always tests if the powder is active before baking."

We can view this from the perspective of cultural practice. E has learned to clean spoons in this way and (at least she claims) incorporated that knowledge into another activity, baking. She does not need an external memory aid to find the correct active baking powder. She instead remembers the practice of testing the powder, which appears to work instead of the daughter's suggested strategy. Understanding the practices of a group can predict development of cognitive strategies over time. Certain
ecologies of practices shape certain needs to develop certain kinds of cognitive processes. In the case of $E$ a cognitive process is used that can resist the process of cognitive aging rather efficiently.

Previously I have also discussed examples where we can predict that changes in cognitive ecologies such as suddenly being part of the home healthcare system constrain and alter the cognitive processes. Changing the social circumstances in general such as giving more structure or expectation of certain activities changes the cognitive processes involved, such as the use of an artifact for some purpose. The case of D previously is likely the clearest example of this, where the role of the envelope as an artifact is given two functions to serve different components of the activity. Understanding changes and ecological factors in people's lives can help us understand the diachronic aspect of cognitive processes and possibly also predict cognitive performance in new ecologies.

## Concluding remarks

Still there is a need to understand what the differences are between kinds of ecologies and what principles are at work in the shaping processes of cognitive processes in everyday life. It is still rather unexplored how distributed cognitive processes work in a less clear purpose-driven environment, in a less information dense environment, that are guided by a more or less clear socially structured environment, and that are more or less demanding for individuals. The importance of having a cognitive process adapted to the cognitive ecology can be different between these environments. It can be that the role of the individual is far more important in everyday life situations, and that the role of the ecology is more subtle in these environments since it is an ecology that do not shape, alter or constrain the cognitive process as much as other ecologies.

I have used the case of older adults to show that cognitive ecology is a useful term to understand cognitive processes in everyday life. But I do not believe that the points made throughout the second half of the paper are specific to older adults. Older adults are an interesting group for many reasons. One is that they have been investigated quite thoroughly in lab-settings and that we know that the prediction of lab-performance to the settings of their everyday life is low. Cognitive ecology is one way to understand this. But we all live in an everyday life where the social network and the length to the grocery store to some extent shape our cognitive processes. This paper shows that the ecology can shape how much effort we put into our cognitive processes, how we distribute them and how we create routines for them in everyday life environments.

## Acknowledgements

I thank Nils Dahlbäck, Fredrik Stjernberg and two anonymous reviewers for helpful feedback on a previous version of this paper.

## References

Burke, K. (1969). A grammar of motives. Berkeley: University of California Press.
Cavanaugh, J. C., Grady, J. G., \& Perlmutter, M. (1983). Forgetting and use of memory aids in 20 to 70 year olds everyday life. International Journal of Aging and Human Development, 17(2), 113-122.
Clark, A. (2006). Soft selves and ecological control. In D. Spurrett, D. Ross, H. Kincaid, \& L. Stephens (Eds.), Distributed Cognition and the Will (pp. 101-122). Cambridge, MA: MIT Press.
Clark, A. (2008). Supersizing the mind: embodiment, action, and cognitive extension. New York: Oxford University Press.
Dahlbäck, N., Kristiansson, M., \& Stjernberg, F. (2013). Distributed remembering through active structuring of activities and environments. Review of Philosophy and Psychology, 4(1), 153-165.
Dixon, R. A. (1999). Exploring cognition in interactive situations: the aging of $\mathrm{N}+1$ minds. In T. M. Hess \& F. Blanchard-Fields (Eds.), Social Cognition and Aging (pp. 267-290). San Diego and London: Academic Press.
Dukas, R., \& Ratcliffe, J. M. (2009). Cognitive Ecology II. Chicago: University of Chicago Press.
Frias, C. M. De, Dixon, R. A., \& Bäckman, L. (2003). Use of memory compensation strategies is related to psychosocial and health indicators. The journals of gerontology. Series B, Psychological sciences and social sciences, 58(1), 12-22.
Garbis, C. (2002). The Cognitive Use of Artifacts in Cooperative Process Management: Rescue Management and Underground Line Control. Linköping Studies in Arts and Science.
Hollnagel, E., Woods, D. D., \& Leveson, N. (2006). Resilience Engineering: Concepts and Precepts. Farnham, Surrey: Ashgate Pub Co.
Hutchins, E. (1995). Cognition in the Wild. Cambridge, MA: MIT Press.
Hutchins, E. (2010a). Cognitive ecology. Topics in Cognitive Science, 2(4), 705-715.
Hutchins, E. (2010b). Enculturating the supersized mind. Philosophical Studies, 152(3), 437-446.
Salthouse, T. (2012). Consequences of age-related cognitive declines. Annual review of psychology, 63, 201-26.
Sutton, J., Harris, C. B., Keil, P. G., \& Barnier, A. J. (2010). The psychology of memory, extended cognition, and socially distributed remembering. Phenomenology and the Cognitive Sciences, 9(4), 521-560.
Tribble, E. B. (2011). Cognition in the Globe: Attention and Memory in Shakespeare's Theatre. New York: Palgrave Macmillan.
Tribble, E. B., \& Sutton, J. (2011). Cognitive ecology as a framework for shakespearean studies. Shakespeare Studies, 39, 94-104.
Wertsch, J. V. (1998). Mind As Action. Oxford: Oxford University Press.

# What Aspects of Cyber Cruelty are judged most distressing? An Adaptive Conjoint Study with Two Independent Samples 

Christina Kuhlmann (s1chkuhl@uni-trier.de)<br>Stephanie Pieschl (pieschl@uni-muenster.de)<br>Torsten Porsch (t.porsch@uni-muenster.de)<br>Department of Psychology, Westfälische Wilhelms-Universität Münster, Fliednerstraße 21, 48149 Münster, Germany


#### Abstract

Cyberbullying is defined as bullying via electronic means including the defining characteristics of repetition over time, intent to harm, and power imbalance. However, this normative top-down definition is discussed controversially. We argue that the term "cyberbullying" and the associated defining criteria might constrict our focus artificially. Therefore, we investigate bottom-up which aspects of cyber cruelty contribute to victims' distress in an adaptive conjoint design with two independent samples (sample 1: $n=131$; sample 2: $n=82$ ). Six potentially relevant factors were investigated, each with multiple attributes: number of incidents, perpetrator status, perpetrator motive, and type, medium, and publicity of cyber incident. Contrary to the definition of cyberbullying, number of incidents, publicity, and type of cyber cruelty emerged as most important factors. These results allow us to further map the cognitive representation of cyber cruelty and are practically relevant for the definition and measurement of cyberbullying.


Keywords: cyberbullying; electronic communication; emotional distress; cognitive representation; conjoint analysis.

## Theoretical Background

Cyberbullying - namely bullying via electronic means - is a prevalent problem among today's youth with mostly negative consequences (Tokunaga, 2010). In order to adequately research this phenomenon and to ultimately design effective prevention and intervention measures, a precise conceptualization of this construct is paramount (Ybarra, Boyd, Korchmaros, \& Oppenheim, 2012).

However, one area of controversy is the literal connotations of the composite term "bullying". Today most scientists agree that bullying denotes an "aggressive, intentional act or behavior that is carried out by a group or an individual repeatedly and over time against a victim who cannot easily defend him or herself" (Olweus, 1993; cited in Smith, Mahdavi, Carvalho, Fisher, Russell, \& Tippett, 2008, p. 376). Thus, repetition, power imbalance, and intent to harm are considered the key defining characteristics of bullying. But research shows that the understanding of "bullying" differs between historical eras, cultures or age groups (Smith \& Monks, 2008). For example, in cultural comparisons, one of the biggest challenges is finding translations of "bullying" with equivalent meaning. Most often, terms vary in breadth and cognitive connotations; "the social construction of meaning and its cultural and temporal variability become apparent" (ibid., p. 110).

With the advent of electronic communication and the first reported cases of online cruelty, the term "cyberbullying" was coined to refer to this new phenomenon. The definition of conventional bullying was transferred to cyberspace, and cyberbullying was defined as "an aggressive, intentional act carried out by a group or individual, using electronic forms of contact, repeatedly and over time against a victim who cannot easily defend him or herself" (Smith et al., 2008, p. 376). However, this theory-based top-down definition of cyberbullying has been discussed controversially ever since (e.g., Dooley, Pyzalski, \& Cross, 2009; Grigg, 2010; Menesini \& Nocentini, 2009; Pieschl et al., in press).

Recent empirical investigations about the connotations and cognitive representations of the term "cyberbullying" offer the possibility to shed further light onto this issue from a data-driven, bottom-up perspective. Results from a multidimensional scaling analysis with 2,257 adolescents from six European countries (Menesini et al., 2012) show that the most important dimension of cyberbullying is characterized by the imbalance of power and the second most important dimension is characterized by intentionality. When adolescents classify a scenario as cyberbullying, they seem to mainly consider the presence of these criteria. Focus-group interviews of 70 Italian, Spanish and German adolescents (Nocentini, Calmaestra, Schultze-Krumbholz, Scheithauer, Ortega, \& Menesini, 2010) on the other hand show that in some cases, subjects consider the publicity of an incident as a substitute of the criterion of repetition. Further, they consider victims' perceived level of distress more important than an existing imbalance of power and view victims' interpretation of an incident more critical than an existing intent to harm. These results seem to imply that the cyber-victims' experience is more important than the adherence to normative criteria. Adolescents from another focus group study go even one step further; they consider the term cyberbullying "vague, inadequate and restricted" (Grigg, 2010, p. 151) because of the broad and varied set of negative incidents that can happen on the internet but that are not covered by this term.

We argue that these investigations about subjects' interpretation of the term "cyberbullying" can only show one side of the coin: Subjects evaluate the normative criteria of cyberbullying. But generations of students have been taught the definition of "bullying" in school. Therefore, it is not surprising that they consider incidents as "cyberbullying" that are consistent with this learned
definition. Thus, the term "cyberbullying" might artificially constrict researchers' and practitioners' focus. Many hurtful online experiences do not fall into this narrow definition.

Therefore, we advocate a complementary route of investigation to also shed light onto the other side of the coin: We explore which aspects of cyber incidents are evaluated as most distressing and use these as cognitive criteria underlying a more inclusive definition of cyber cruelty. This approach is consistent with adolescents' views (Grigg, 2010; Nocentini et al., 2010). It is also consistent with the diagnosis of psychological disorders; only those disorders are considered that cause clinically significant distress or impairment in specific areas of functioning (American Psychiatric Association, 2000).
We assume that not only defining criteria of "cyberbullying" are relevant to the experience of distress but also cyber-specific factors (for an overview see Table 1). More specifically, the criterion of repetition can be captured in a straightforward way by investigating the impact of different number of incidents. Power imbalance, on the other hand, can have many facets, such as age, competence or intelligence; in this study we consider the social status of the perpetrator in terms of perceived popularity (Pieschl et al., in press) as well as anonymity (Dooley et al., 2009; Menesini et al., 2012) (perpetrator status). Besides intent to harm and the related motives of feeling superior and whish for appreciation (Olweus, 1996), we also consider that perpetrators might not be aware of the consequences of their behavior and instigate seemingly cruel incidents for fun (Twyman, Saylor, Taylor, \& Comeaux, 2010) or that they might seek retaliation (Vandebosch \& Van Cleemput, 2008) (perpetrator motive). For these criteria of (cyber-)bullying, we predict that those incidents including repetition, power imbalance and intent to harm will be more distressing than other incidents.

As first cyber-specific factor, we consider selected types of cyberbullying and cyber cruelty (Pieschl et al., in press; Willard, 2007): harassment (insults or threats), denigration (spreading rumors), outing (revelation of secrets), impersonation (passing off as someone else) and exclusion (from online groups and activities). The second cyberspecific factor is the medium. Because recently, hardware and software applications merge, we will consider the representational code as most relevant dimension; we predict that pictorial incidents will be more distressing than written / verbal ones (Pieschl et al., in press; Smith et al., 2008). Our third cyber-specific factor is the publicity of the incident; we predict that public incidents are more distressing than semi-public and private ones (Nocentini et al., 2010).

In the context of cyberbullying, distress has mainly been investigated on an emotional level, for example as feeling upset or stressed (Ortega, Elipe, Mora-Merchán, Calmaestra, \& Vega, 2009). But it also incorporates cognitive facets such as helpless cognitions, for example thoughts like "My situation is hopeless" (Pieschl et al., in press). Furthermore, previous research shows that
participants with a history of victimization consistently report higher levels of distress when confronted with bullying and cyberbullying scenarios (Bauman \& Newman, 2013). Thus, we also predict that previous cyber experience is relevant for the level of cognitive-emotional distress.

## Hypotheses

In this study we investigate two main hypotheses in an adaptive conjoint design: (1) Not all factors are equally important for the experience of distress as a result of cyber cruelty; the utility values of these factors differ significantly. More specifically, we predict that not all defining characteristics of (cyber-)bullying (number of incidents, perpetrator status, and perpetrator motive) are judged more important than cyber-specific factors (type, medium, and publicity of cyber incident). (2) The part-worth utility values of the attributes of each factor differ significantly (all attributes are given in Table 1). We predict more distress associated with more frequent cyber incidents (number of incidents), with popular perpetrators (perpetrator status indicating power imbalance) who have an intent to harm (perpetrator motive), and pictorial (medium) and public (publicity) cyber incidents. For type of cyber incident, this is an explorative research question. Furthermore, we explore two research questions about between-subject differences: (3) The results regarding (1) and (2) differ significantly according to previous cyber-experience; previous cybervictims report the highest level of distress, significantly more than previous cyber-perpetrators. (4) The results regarding (1) and (2) can be validated in two independent samples; in both samples we predict similar results.

## Method

## Samples

Sample 1 consists of 133 high school students. Data from 2 students had to be excluded because of missing data, thus the final sample size is $n=131$. These 43 boys ( $32.8 \%$ ) and 88 girls ( $67.2 \%$ ) are on average $17.47(S D=1.01)$ years old and spend on average 2.43 hours ( $S D=1.50$ ) on weekdays and 3.66 hours ( $S D=2.72$ ) on weekend days on the internet. Sample 2 consists of 91 young adults. Data from 9 young adults had to be excluded because of missing data, thus the final sample size is $n=82$. These 18 young men ( $22.0 \%$ ) and 64 young women ( $78.0 \%$ ) are on average 20.29 ( $\mathrm{SD}=1.14$ ) years old and spend on average 2.90 hours ( $S D=1.75$ ) on weekdays and 3.85 hours ( $S D=2.69$ ) on weekend days on the internet.

## Material

Adaptive Conjoint Analysis: Distress Measure An adaptive conjoint analysis (ACA; Gustafsson, Herrmann, \& Huber, 2007) was presented by the online survey system Unipark (© Questback). In this part of the study, participants had to imagine that they were cyberbullied and they had to rate their level of distress associated with multiple fictitious incidents that were described by
combinations of attributes. The relevant factors and attributes can be seen in Table 1. Because of the high number of factors and attributes in this study, we used a fractional factorial design. Yet the maximum gain of information was reached by using the adaptive conjoint design. The system automatically arranges the scenarios based on subjects' previous judgments by choosing those attributes whose comparison provides the maximum of new information.

Table 1: Factors (defining characteristics and cyberspecific) and corresponding attributes investigated in the adaptive conjoint analysis.

| Defining Characteristics | Attributes |
| :--- | :--- |
| number of incidents | once <br> $2-3$ times per month <br> weekly <br> multiple times per week |
| perpetrator status | anonymous <br> popular <br> unpopular |
| perpetrator motive | intent to harm <br> feeling superior <br> appreciation by others <br> retaliation <br> fun |
| Cyber-Specific Factors | Attributes |
| medium | written / verbal <br> pictorial |
| publicity | private <br> semi-public <br> public |
| type of cyber incident | harassment <br> denigration <br> outing <br> impersonation <br> exclusion |

In the preference for levels phase of the ACA, participants rated the level of distress associated with each attribute on a 6 -point scale ( $1=$ not upsetting $-6=$ very upsetting). In the attribute importance phase, the most and least distressing attributes of each factor were contrasted and participants had to judge on a 4-point scale if these were "equally upsetting" or "one is more upsetting than the other". In the phase of paired-comparison trade-off questions, we presented two fictitious cyber incidents, each consisting of 2 or 3 attributes of different factors. On a 5 -point scale with one situation located at each end, subjects had to decide which situation was more distressing. In the final calibrating concepts phase, participants had to rate the level of distress associated with cyber incidents consisting of a combination of 4 attributes of different factors on a scale from 0 (not at all distressing) to 100 (very distressing).

Based on participants' judgments, Unipark (© Questback) automatically computes part-worth utilities for each
attribute and utility values for each factor. High values indicate that a specific factor (or attribute) is relatively important for participants' judged distress while low values indicate relative unimportance.

Cyber Experiences Questionnaire We adapted the cyberbullying questionnaire of Riebel, Jäger and Fisher (2009) to include the following five of Willard's (2007) categories: harassment, denigration, impersonation, outing, and exclusion. Students were asked how often these incidents happened to them via cell phone or the internet (cyber-victim) and how often they instigated such incidents themselves (cyber-perpetrator) in the last two months. All answers were given on 5 -point scales with the categories "never", "once", "2-3 times per month", "weekly", and "multiple times per week". Cyber involvement was diagnosed if students gave at least once a different answer than "never" (cyber-victim and cyber-perpetrator). Participants who reported both cyber-victim and cyberperpetrator experiences were classified as cyber-perpetratorvictims. Note that we do not refer to these experiences as cyberbullying because some of the conventional criteria for bullying are not fulfilled, for example repetition over time.

## Procedure

Sample 1 was recruited at the Open Day of the Westfälische Wilhelms-Universität Münster, Germany. High school students visiting the Department of Psychology volunteered and their data was collected in group sessions in a computer lab. Sample 2 was recruited from a database of adult volunteers maintained by the same department of psychology; participants were sent the link to the online survey and answered at will. All participants answered the same electronic survey presented by Unipark (© Questback). It consists of demographic questions, the adaptive conjoint analysis, the cyber experience questionnaire and further questions ${ }^{1}$.

## Results

## Descriptive Results

Sample 1 and sample 2 differ significantly in age $(t(211)=-18.92, p<.001)$ and average internet use on weekdays $(t(211)=-2.11, p=.036)$. On average, sample 2 participants are older and spend more time on the internet on weekdays. Sample 1 and sample 2 also differ significantly in their cyber experience ( $X^{2}(3)=13.82$, $p=.003$ ). In sample 1 , only $37.4 \%$ of students were not involved in cyber incidents in the last two months, $22.9 \%$ were classified as cyber-victims, $15.3 \%$ as cyberperpetrators, and $24.5 \%$ as cyber-perpetrator-victims. In

[^319]sample 2, a total of $61.0 \%$ young adults were not involved in cyber incidents in the last two months, $18.3 \%$ were classified as cyber-victims, $12.2 \%$ as cyber-perpetrators, and $8.5 \%$ as cyber-perpetrator-victims. Therefore, the variables "sample" and "cyber-experience" will be included as between-subject factors in all subsequent analyses.

## Hypothesis 1: Factor Differences

To investigate hypothesis 1, we computed a repeatedmeasure ANOVA with the utility values of the six factors as repeated-measure dependent variable and cyber-experience and samples as between-subject factors. The results show a significant main effect of the repeated-measure factor $(F[4.5,921.7]=19.67, p<.001)$, but no significant differences between groups with different cyber-experiences $(F[3,205]=1.73, \quad p>.05)$ or between samples ( $F[1,205$ ] $=3.45, p>.05$; see Figure 1), and no significant interactions between these factors $(F[13.5,921.7]=0.5$, $p>.05$ ).


Figure 1: Utility values of the six factors extracted from the conjoint analysis; these values indicate distress.

As can be seen in Figure 1, publicity, number, and type of cyber incidents were assigned the highest utility values and these did not differ significantly from one another (publicity vs. number: $F[1,205]=0.75, p>.05$; number vs. type: $F[1,205]=0.96, p>.05)$. Perpetrator motive was judged significantly less important (type vs. motive: $F[1,205=12.42, p=.001)$, followed by perpetrator status (motive vs. status: $F[1,205]=8.72, p=.004$ ). Medium of incident was judged least important but did not differ significantly from perpetrator status $(F[1,205]=3.00$, $p>.05)$.

## Hypothesis 2: Attribute Differences

To investigate hypothesis 2 , we computed repeated-measure ANOVAs for the attributes of each factor separately. In each ANOVA the part-worth utility values of all attributes of one factor constitute the repeated-measure dependent variable, while cyber-experience and samples constitute the between-subject factors. We report only significant effects ordered by factor.

In all ANOVAs we found main effects of the repeatedmeasure variable: More frequent incidents were judged more distressing (number: $\quad F[2.1,432.4]=183.11$, $p<.001$ ), popular perpetrators were judged more distressing than anonymous ones, followed by unpopular ones (status: $F[1.9,389.7]=50.79, p<.001$ ), the intent to harm was judged more distressing than fun, followed by retaliation, appreciation by others, and feeling superior (motive: $F[3.6,731.7]=35.64, p<.001)$. Furthermore, pictorial incidents were judged more distressing than written / verbal ones (medium: $F[1,205]=123.00, p<.001$ ), more public incidents were judged most distressing, followed by semipublic ones and private ones (publicity: $F[1.5,306.7]=203.20, p<.001$ ), and outing was judged most distressing, followed by harassment, denigration, exclusion, and impersonation (type: $F[3.8,774.6]=7.51$, $p<.001$ ). To give one more specific example (number; see Figure 2): incidents "multiple times per week" were judged significantly more distressing than "weekly" incidents $(F[1,205]=11.90, p<.001)$ which were in turn judged significantly more distressing than incidents "2-3 times per month" ( $F[1,205]=72.36, \quad p<.001)$ and those were judged significantly more distressing than a single incident ("once": $F[1,205]=157.84, p<.001$ ).


Figure 2: Part worth utilities of the attributes for the factor number extracted from the conjoint analysis; these values indicate distress.

In all ANOVAs we also found main effects of the betweensubject factor cyber-experience. More specifically for number of incidents $(F[3,205]=3.21, p=.024)$, perpetrator status $(F[3,205]=3.21, p=.024)$, perpetrator motive $(F[3,205]=3.21, p=.024)$, and for medium $(F[3,205]=3.21, p=.024)$, publicity $(F[3,205]=3.21$, $p=.024)$, and type of cyber incident ( $F[3,205]=3.21$, $p=.024$ ). In all cases cyber-victims judged most attributes significantly more distressing than cyber-perpetrators. Additionally, we found significant interactions between the repeated-measure factor and cyber-experience for perpetrator motive ( $F$ [10.7, 731.7] $=1.87, p=.041$ ) and publicity $(F[4.48,306.7]=2.7, p=.025)$. To give one more specific example (number; see Figure 2): across all
frequencies of cyber incidents, cyber-victims judged level of distress was significantly higher than that of cyberperpetrators ( $M D=.11, p=.012$ ). Furthermore, the judged level of distress of non-involved participants did not differ significantly from that of cyber-victims ( $M D=.03, p>.05$ ) and the judged level of distress of cyber- perpetrator-victims did not differ significantly from that of cyber-perpetrators ( $M D=.01, p>.05$ ).

We found no significant main effects of the two samples in any of the ANOVAs.

## Discussion of Results

Our first hypothesis was confirmed: These results show that there are aspects of cyber cruelty that are perceived as significantly more distressing than others. As predicted, not all defining characteristics of cyberbullying were judged more important than cyber-specific factors. More specifically, while number of incidents was considered among the most important factors, status and motive of the perpetrator - indicative of power imbalance and intent to harm respectively - are considered significantly less important. On the other hand, the cyber-specific factors type of cyber incident and publicity are among the most important factors. Medium of cyber incident emerged as the least important factor.

Our second hypothesis was also confirmed: These results indicate that, for each factor, some of the associated attributes are perceived significantly more distressing than others. As predicted, more distress was associated with more frequent incidents (number), with popular bullies rather than with unpopular bullies (status), with intent to harm (motive), with pictorial rather than with written incidents (medium), and with more publicity. Furthermore, results regarding the perpetrator status indicate that anonymous perpetrators are perceived more distressing than unpopular ones but less distressing than popular ones. Additional results regarding perpetrator motives indicate that all other motives but intent to harm were judged significantly less distressing, more specifically retaliation, fun, appreciation by others and feeling superior. Finally, our explorative research question regarding types of cyber incidents indicates that outing was considered most distressing, followed (in order of descending importance) by harassment, denigration, exclusion, and impersonation. Consequently, the most distressing case of cyber cruelty would be the following one: several public incidents of outing per week, by a popular bully in form of pictures or videos, where the perpetrator wants to harm the victim.

Our third and fourth hypotheses were also (mostly) confirmed: We found no significant effects of the betweensubject factor cyber-experience regarding hypothesis one, but in all analyses regarding hypothesis two the results confirm our predictions (hypothesis 3): For all investigated factors, cyber-victims (and often non-involved participants) reported significantly more distress across all attributes than cyber-perpetrators (and often cyber-perpetrator-victims)
(main effects); further interactions indicate that these differences disappear for very distressing attributes.
Furthermore, the pattern of results in sample 1 and sample 2 did not differ significantly in any of our analyses, pointing to the validity of our findings (hypothesis 4).

On a theoretical level, these findings underline, on the one hand, that defining characteristics of conventional bullying are indeed relevant to the experience of cyberbullying. Repeated incidents with intent to harm and power imbalance are perceived more distressing than other incidents. On the other hand, these results also show that cyber-specific factors are just as or even more important for victims' experience of cyber cruelty. Especially the type of incident and its publicity seem to be important. Therefore, the experience of cyber cruelty seems not (only) to be determined by the (artificial) boundaries of a normative, theory-driven, top-down definition of cyberbullying. Rather, subjects' cognitive representation of such incidents (datadriven bottom-up approach) shows that all proximal factors that concern the content of the incident and thus also the cyber-victim directly - namely number of incidents, type of incident, and the incident's publicity - are judged more important for the experience of distress than more distal factors regarding the perpetrator or medium - namely status and motive of the perpetrator or medium.

## Limitations and Implications

The advantage of using the innovative approach of adaptive conjoint analysis to assess implicit judgments unfortunately goes hand in hand with a possible loss in external validity. Since the attributes needed to be suitable for every potential combination of attributes, they had to be expressed on a rather abstract level. Therefore, imagining concrete cyber incidents might have been complicated and the imagined situations might have been quite idiosyncratic. Additionally, we do not know if the results can be generalized since the experience of cyber cruelty presumably also depends on further personal and contextual factors, for example on previous cyber-experience as shown in this study. But we do not know if the effects of cyber-experience are due to the fact that previous cyber-victims are better able to take the victim perspective or if they point to a cumulative vulnerability as a result of cyber-victimization. Further research is needed. However, the fact that we could replicate our results with two independent samples points to the validity of our findings.

Despite these limitations these results have further theoretical and practical implications that are not only highly relevant for psychology, but might also have implications for other cognitive science disciplines such as linguistics or philosophy. For example, the question of how technical terms such as "cyberbullying" are conceptualized and might constrain human cognition clearly lies at the intersection of psychology and linguistics. On a more concrete theoretical level these results contribute to the controversial discussion within psychology about the definition of cyberbullying: We suggest that cyber-victims'
level of distress should be taken into account for the definition and diagnosis of cyberbullying. Such definitions and diagnosis criteria should not only be based on normative, theory-driven, top-down considerations but also on subjects' cognitive representations of cyber cruelty (datadriven bottom-up approach). We would like to propose the affected subjects' level of distress as potential defining criterion. However, because of the widely accepted criteria of "bullying", another more inclusive term for all kinds of cyber cruelty might be more useful. Further research regarding this issue is needed. However, we can still draw some practical conclusions: For example, the distress associated with outing indicates that adolescents need to be advised of the dangers of sharing private information online. Additionally, the distress associated with publicity indicates that education about data protection and privacy settings could also contribute to the prevention of cyber cruelty.

## Acknowledgments

The data was collected as part of the first author's Bachelor thesis requirement.

## References

American Psychiatric Association. (2000). Diagnostic and statistical manual of mental disorders (4th ed., text revision). Washington, DC: Author.
Bauman, S., \& Newman, M. L. (2013). Testing assumptions about cyberbullying: Perceived distress associated with acts of conventional and cyber bullying. Psychology of Violence, 3(1), 27-38.
Dooley, J. J., Pyżalski, J., \& Cross, D. (2009). Cyberbullying versus face-to-face bullying: A theoretical and conceptual review. Zeitschrift Für Psychologie/Journal Of Psychology, 217(4), 182-188.
Grigg, D. (2010). Cyber-aggression: Definition and concept of cyberbullying. Australian Journal Of Guidance And Counselling, 20(2), 143-156.
Gustafsson, A., Herrmann, A. \& Huber, F. (2007). Conjoint measurement: Methods and applications. Berlin Springer.
Menesini, E., \& Nocentini, A. (2009). Cyberbullying definition and measurement: Some critical considerations. Zeitschrift Für Psychologie/Journal Of Psychology, 217(4), 230-232.
Menesini, E., Nocentini, A., Palladino, B., Frisén, A., Berne, S., Ortega-Ruiz, R., \& ... Smith, P. K. (2012). Cyberbullying definition among adolescents: A comparison across six European countries. Cyberpsychology, Behavior, And Social Networking, 15(9), 455-462.
Nocentini, A., Calmaestra, J., Schultze-Krumbholz, A., Scheithauer, H., Ortega, R., \& Menesini, E. (2010). Cyberbullying: Labels, behaviours and definition in three European countries. Australian Journal Of Guidance And Counselling, 20(2), 129-142.
Olweus, D. (1996). Gewalt in der Schule. Was Lehrer und Eltern wissen sollten - und tun können.[Bullying at
school: What we know and what we can do]. Verlag Hans Huber. 2. Aufl., Bern.
Ortega, R., Elipe, P., Mora-Merchán, J. A., Calmaestra, J., \& Vega, E. (2009). The emotional impact on victims of traditional bullying and cyberbullying: A study of Spanish adolescents. Zeitschrift Für Psychologie/Journal Of Psychology, 217(4), 197-204.
Pieschl, S., Porsch, T., Kahl, T., \& Klockenbusch, R. (in press). Relevant dimensions of cyberbullying - results from two experimental studies. Journal of Applied Developmental Psychology.
Raskauskas, J., \& Stoltz, A. D. (2007). Involvement in traditional and electronic bullying among adolescents. Developmental Psychology, 43(3), 564-575.
Riebel, J., Jäger, R. S. \& Fischer, U. C. (2009). Cyberbullying in Germany - An exploration of prevalence, overlapping with real life bullying and coping strategies. Psychology Science, 51(3), 298-314.
Smith, P. K., \& Monks, C. P. (2008). Concepts of bullying: Developmental and cultural aspects. International Journal Of Adolescent Medicine And Health, 20(2), 101-112.
Smith, P.K., Mahdavi, J., Carvalho, M., Fisher, S., Russell, S., \& Tippett, N. (2008). Cyberbullying: Its nature and impact in secondary school pupils. Journal Of Child Psychology And Psychiatry, 49(4), 376-385.
Tokunaga, R. S. (2010). Following you home from school: A critical review and synthesis of research on cyberbullying victimization. Computers In Human Behavior, 26(3), 277-287.
Twyman, K., Saylor, C., Taylor, L. A., \& Comeaux, C. (2010). Comparing children and adolescents engaged in cyberbullying to matched peers. Cyberpsychology, Behavior, and Social Networking, 13(2), 195-199.
Vandebosch, H., \& van Cleemput, K. (2008). Defining Cyberbullying: A Qualitative Research into the Perceptions of Youngsters. CyberPsychology \& Behavior, 11(4), 499-503.
Willard, N. E. (2007). Cyberbullying and cyberthreats: Responding to the challenge of online social aggression, threats, and distress. Champaign, Ill: Research Press.
Ybarra, M. L., Boyd, D., Korchmaros, J. D., \& Oppenheim, J. (2012). Defining and measuring cyberbullying within the larger context of bullying victimization. Journal of Adolescent Health, 51(1), 53-58.

# Hierarchical Control and Sense of Agency: Differential Effects of Control on Implicit and Explicit measures of Agency 

Devpriya Kumar (devpriyak@cbcs.ac.in)<br>Centre of Behavioural and Cognitive Sciences, University of Allahabad<br>Allahabad 211002, India<br>Narayanan Srinivasan (nsrini@cbcs.ac.in)<br>Centre of Behavioural and Cognitive Sciences, University of Allahabad<br>Allahabad 211002, India


#### Abstract

While recent studies show dissociation between the implicit and explicit aspects of 'sense of agency', the mechanisms underlying these different aspects of agency are not yet clearly understood. We argue that the control achieved at different levels of hierarchy is important for different aspects of agency. In the current study, we investigate how changes in control at the perceptual-motor level and at goal level influence implicit and explicit measures of sense of agency. In a given trial, participants were first required to aim at a target in a noisy environment and then shoot at the target. After certain interval, a circle flashed at the location where participant aimed while pressing the trigger. Participants estimated the interval between action and presentation of the circle that acted as a measure of intentional binding, an implicit measure of agency and also rated an explicit sense of authorship. The results suggest that different aspects of agency and dissociation between implicit and explicit aspects of agency are mediated by control achieved at various levels.


Keywords: Sense of agency; event-control approach; intentional binding; control; hierarchical system; interval estimation.

## Introduction

With every action that we perform, we not only influence our environment but also modify our conscious mental state of being the agent of the action. This feeling is known as sense of agency (Pacherie, 2011). While elusive, this sense of agency (SoA) is central to our conscious experience and has recently gained popularity among philosophers as well as scientists (see Gallagher, 2006, for a review and insight on related issues).
Sense of agency is a complex, multifaceted, phenomenon (Pacherie, 2011). In general two aspects of SoA have been emphasized and studied in detail. Firstly, sense of Agency as first order experience, in which agency is generally linked to the intentional aspect of task. Here an action is considered to be self-generated when the effect of an action matches the intention of the participant (Moore, Lagando, Deal, \& Haggard, 2009; Farrer \& Frith, 2002), or when agency is linked to bodily movement as in the famous rubber hand illusion (Farrer, Frank, Georgieff, Frith, Decety, \& Jeannerod, 2003). This aspect of agency is also called pre-reflective or implicit sense of agency. Secondly, sense of agency as reflective attribution (or sense of
authorship), in which participant is asked to report his/her subjective sense of belief in causing an action (Ebert \& Wegner, 2010; Haggard \& Moore, 2010). Recent models of 'self' (Synofzik, Vosgerau, \& Newen, 2008) take into account these two aspects of agency.

An important measure of agency that has gained prominence in the last decade is intentional binding. Intentional binding refers to the finding that participants perceive the self-generated action and its effect to be temporally closer to each other (Haggard, Clark, \& Kalogeras, 2002b). The concept of intentional binding has been linked in the literature strongly to the sense of agency, that is, the experience of agency is greater when intentional binding is stronger. A recent review (Moore \& Obhi, 2012) suggests that intentional binding has been associated with implicit measures of agency like efference, sensory feedback, causal feedback, and intentionality. Haggard and Clark (2003) have suggested that when motor cortex is stimulated to produce a movement similar to a voluntary movement, intentional binding is not affected, suggesting that intentional binding does not depend on the sensory signals produced during movement, but rather it depends on the efference copy generated during action planning.

Desantis and colleagues (Desantis, Cedric, \& Waszack, 2011) showed that, when participant believes that he/she has control over the environment, intentional binding is stronger. In the original study showing intentional binding, (Haggard, Clark, \& Kalogeras, 2002b) the outcome of a participant generated action was manipulated in terms of intention (intended or unintended). Results indicate that intentional binding is stronger for the intended effect compared to the unintended effect. These and many more studies indicate a strong link between the implicit measures of agency and intentional binding, suggesting that intentional binding can be used as a reliable measure of implicit sense of agency.

In addition, researchers have investigated the relationship between intentional binding and explicit sense of agency or reflective sense of agency (Moore \& Obhi, 2012). In one such study (Ebert \& Wegner, 2010), participants were presented with a picture, which can either move congruent or incongruent to the direction participant moved the joystick. The delay between the participant's movement of the joystick and the movement of the object on the screen
was manipulated at three levels $(100 \mathrm{~ms}, 400 \mathrm{~ms}$, and 700 ms ). At the end of the trial participants were asked to perform a interval estimation task (to measure intentional binding) and give rating of authorship (explicit measure of agency). Authors reported dissociation between the two measures, with the congruency between action and effect, having a greater effect on intentional binding compared to the explicit measure of sense of authorship. The mechanisms underlying these two different measures and aspects of SoA are still not fully understood necessitating a study to understand the mechanisms involved in SoA. The current study investigates the mechanisms involved in determining the implicit and explicit measures of sense of agency and the way in which these two measures might be related to each other.
We argue that the concept of control exercised by the participant over perception-action events can provide us a basic framework to understand both explicit and implicit sense of agency. Recent studies have shown that control might play a crucial role in influencing sense of agency (Desantis, Cedric, \& Waszack, 2011; Moore, Lagando, Deal, \& Haggard, 2009; Jordan, 2003; Kumar \& Srinivasan, 2012; Kumar \& Srinivasan, under review). Studies based on event-control approach (Jordan, 2003; Kumar \& Srinivasan, 2012) suggest that all our interactions with the environment (which are in form of perceptual-action couplings) are constrained by multiple hierarchical control loops extending across organism and his environment (see Jordan, 2003 for details). Sense of agency, according to this framework is determined by the highest level of control loop at which participant is able to exercise control.
We used a modification of the paradigm used by Ebert and Wegner (Ebert \& Wegner, 2010). In the current experiment, the participant had to aim and shoot at a noisy target with the help of joystick and the noise in the environment was manipulated. By changing the amount of noise, we manipulated the perceptual-motor control that the participant can exercise. After the first task, the scenario was made static and a circle flashed at the location where subject aimed during the first task. The duration interval between the time when subject presses the trigger and when the circle is flashed was manipulated. Participant is later asked to estimate this interval and give a confidence rating for authorship of action. Estimated interval acted as a measure of intentional binding and confidence rating measured participant's subjective sense of authorship.
According to the event-control approach, sense of self depends upon the highest level at which control is exercised. In the current paradigm, control can be exercised at two levels; firstly, at the perceptual-motor level, that is the joystick level control and secondly, at the goal level, that is, whether or not participant is able to correctly aim at the target. We hypothesized that when participant misses the target, sense of agency would increase as a function of perceptual-motor control. When the participant hits the target, SoA would be independent of perceptual-motor control.

## Method

## Participants

Thirteen volunteers from University of Allahabad participated in the Experiment.

## Stimuli and Apparatus

Stimuli consisted of eleven natural scenes (resolution $3648 \times 2736$ ) from a custom database. Every scene contained a target region in the form of three concentric circles, placed randomly somewhere in the scene. Experiment was conducted on a 14 " monitor at a resolution of $800 \times 600$, with input from keyboard and joystick. The experiment was designed using MATLAB 2010b and psychophysics toolbox 3.

## Procedure

Participants were instructed that the experiment consists of two phases, practice phase and the main experiment. They were also told that they have to perform time interval estimation and were instructed about what millisecond stands for and an approximate idea of the concept (see Ebert and Wegner, 2010 for more details).

## Practice Session

In the practice session, a fixation cross was presented on the screen. Participants were instructed that they have to press a trigger to initiate trial and they can press the trigger when they feel like. After the trigger was pressed, the fixation cross on the screen turned blue in color indicating that the trigger has been pressed. After a random interval (out of $0 \mathrm{~ms}, 100 \mathrm{~ms}, 200 \mathrm{~ms}, \ldots, 900 \mathrm{~ms}$ ), a blue circle was flashed on screen. Participants were asked to estimate the time interval between trigger press and the circle flashing on the screen. Response was made using a ten point scale ( 0,100 , $200, \ldots, 900$ ). At the end of every trial, participant was given feedback about his/her estimate. The practice session served two purposes. Firstly, it helped improve interval estimation ability and also its assessment. Secondly, it made participants believe that the interval is manipulated at ten levels in the main session too. A total number of 200 practice trials were given with 20 trials for each of the ten intervals. Data from the practice session was used to perform preliminary analysis.

## Main Session

In the main session, for a particular trial, participants were instructed that they have to aim at a target, by moving the joystick and press the trigger, within 15 seconds. To manipulate the amount of control, a random movement was added to the scene. To decrease the amount of control that participant can exercise, amount of random movement was increased. This control varied from trial to trial. We manipulated control at three levels (low control, medium control, and full control). At a random interval after the participant pressed trigger, a blue circle was presented at the
location (always at the centre of the screen) where participant aimed while pressing the trigger. The SOA between the trigger press and the presentation of blue circle was manipulated at three levels $(100 \mathrm{~ms}, 400 \mathrm{~ms}$, and 700 ms ).

The circle remained on the screen for 500 ms , after which the participant was asked to report the interval between trigger press and appearance of circle, on a ten point scale similar to the practice session. In the main session participant was not given feedback regarding the interval estimated. This was followed by a second question, in which participant had to report the sense of authorship, on a seven point scale (similar to the questions used by Ebert \& Wegner, 2010). There were a total of 216 trials in the experiment, with 24 trials in each condition. We recorded estimated interval, rating of authorship, and whether or not the participant hit on the target in each trial. In the main session participants were not given feedback regarding the estimated interval.

## Results

## Preliminary Analysis

Data from the practice session suggest that participants in general are able to correctly estimate the time interval. Similar to Ebert and Wagner (2011), we calculated the mean correlation between actual time and estimated time (mean $r$ $=0.683$ ) that was significantly greater than zero $t(12)=$ $12.3489, p<.01$. Data from one participant that was beyond two standard deviations from the mean $(r=0.2112)$ was removed from further analysis. In the main experimental session, the outcome (target hit/miss) was not controlled or counterbalanced across SOA (given that this is completely dependent on the performance of the participant in a given trial). Hence, to remove bias due to the unbalanced aspect of target hit/miss, we performed a correlation between target accuracy and SOA. The correlation between accuracy and SOA was not significant (mean $r=-0.0129$ ) indicating a lack of relationship between them.

## Interval estimation task

Repeated measures ANOVA with SOA and control as factors on the estimated interval showed an expected significant main effect of SOA, $F(2,22)=23.46, p<.01$ indicating that participants' estimates increased as SOA increased. The effect of control as well the interaction between control and SOA was not significant. We categorized data further into two categories: (1) when participants hit the target and (2) when participants missed the target. For each category, we performed a two-way repeated measures ANOVA across three levels of SOA and three levels of control.
When participants were successful in hitting the target, there was a main effect of $\operatorname{SOA}, F(2,22)=27.17, p<.001$. Estimated interval for $100 \mathrm{~ms}, t(11)=7.76, p<.01$ and 400 ms SOA, $t(11)=4.21, p<.05$, was significantly different from that for 700 ms SOA. The main effect for control ( $p=$ 0.98 ) and interaction between control and SOA ( $p=0.26$ )
was not significant. When participants were not successful, there was a main effect of $\operatorname{SOA}, F(2,22)=34.01, p<.01$, with mean rating for 100 ms significantly different from rating for $400 \mathrm{~ms}, t(11)=7.55, p<.01$, and rating for 400 ms significantly less than rating for $700 \mathrm{~ms}, t(11)=8.6, p<.01$. The main effect of control was significant, $F(2,22)=6.86$, $p<.01$. Paired $t$-tests between different control conditions suggested a decrease in estimated interval with increase in control, with close to significance difference between, low control and medium control, $t(11)=1.619, p=.057$, and a significant difference between medium control condition and full control condition, $t(11)=3.14, p<.01$. The interaction between SOA and control was not significant ( $p$ $=0.7$ ).

(a)

(b)

Figure 1: Estimated intervals as a function of control and SOA (a) when goal was accomplished and (b) when goal was not accomplished

We used the interval estimation task to assess intentional binding between self-triggered event (cause) and a second perceptual event (effect). Results suggest that intentional binding is greater (estimated interval is less) as the amount of control increases, that is when higher level goal is not achieved. When higher level goal is achieved, intentional binding (interval estimate) is not influenced by the amount of control.

## Self-reported control

When subjects were successful in hitting the target, there was a main effect of control, $F(2,22)=35.57, p<.01$. There was an increase in self-reported control as amount of control was increased, from low to medium, $t(11)=3.91, p$ $<0.05$ and from medium to full, $t(11)=12.35, p<.05$. The
main effect of SOA ( $p=0.28$ ) and the interaction effect ( $p=$ 0.26 ) was not significant.

When subjects were not successful, there was a main effect of control, $F(2,22)=5.62, p<.05$. Post-hoc comparisons show a significant difference between low control and medium control conditions, $t(11)=4.5, p<.05$ as well as low control and high control conditions, $t(11)=$ $6.3, p<.01$. The main effect of SOA $(p=0.88)$ and the interaction ( $p=0.4$ ) was once again not significant.


Figure 2: Rating for the sense of Authorship as a function of control and SOA (a) when target is achieved and (b) when target is not achieved

Self-reported control was a measure of subjects' sense of control and authorship (Ebert \& Wegner, 2010). The results suggest that participant's sense of authorship depends on the amount of control and is independent of success or failure in achieving target.

## Relationship between the Two Measures

Results show dissociation between the sense of agency and sense of authorship with respect to various control levels. To further analyze how these dependent measures are related, we performed a correlational analysis between the measure of intentional binding (interval estimate) and authorship (self-reported control) for the two levels of control. When the target goal is not achieved, there was a significant positive correlation between the two measures ( $r$ $(108)=0.2, p<.05)$. When target goal is achieved, the
correlation between the two measures is not significant ( $r$ $(108)=0.13, p=.16)$ indicating differences between the two measures. Sense of authorship seems to depend on the amount of control that a participant exercises at the perceptual-motor level, whereas intentional binding depends on both lower joystick control level as well as the higher goal level control.

## Control \& SOA as Predictors

To further explore how control and SOA can be used to explain the differences in intentional binding as a function of goal, we performed two simultaneous multiple linear regressions for the estimated interval, treating target hit/miss as a dichotomous variable, with control and SOA as independent factors and intentional binding as the dependent measure.

In target miss condition, control ( $\beta_{1}=-92.3, t=-2.5$ ) and SOA ( $\beta_{2}=0.48, t=15.6$ ) were significant (adjusted- $\mathrm{R}^{2}=$ $0.19, F(2,1035)=124, p<.01)$. In target hit condition, SOA ( $\beta_{1}=0.49, t=18.65, p<.05$ ) was significant, but the $\beta$ value for Control ( $\beta_{2}=4.5, t=0.21, p=0.82$ ) was not significant (adjusted- $\left.R^{2}=0.24, F(2,1094)=174, p<.01\right)$. The analysis suggests that estimated interval decreased as the amount of control increases for the trials in which participants missed the target but control is not a significant predictor when the target goal was achieved.

## Discussion

In the current experiment, we investigated the role of control at multiple hierarchical levels in determining the sense of agency (both implicit, via measuring intentional binding and explicit, via rating on sense of authorship). There are a few important results that can be inferred from the data. Firstly, the high correlation between estimated and actual interval supports the idea of using interval estimation task as a valid measure of intentional binding (Ebert \& Wegner, 2010). Secondly, consistent with the findings from that study, our findings show dissociation between the two measures of agency. Thirdly, our study provides support to the hierarchical event-control framework in understanding self and sense of agency (Kumar \& Srinivasan, under review; 2012; Jordan, 2003).

## Control and Intentional Binding

The results for the measure of intentional binding support the hypothesis that sense of agency depends hierarchically on the amount of control at various levels. The results support to the framework provided by event-control approach (Jordan, 2003; Kumar \& Srinivasan, 2012) suggesting that control might play a key role in determining sense of agency. The results of the study are consistent with findings by Berberian and colleagues (Berberian, Sarrazin, Blaye \& Haggard, 2012), who showed the presence of intentional binding in a complex task and a decrease in intentional binding as a function of automaticity in control. A major difference between our study and that by Berberian et al. (2012) study was that although both studies
manipulated control at different levels, the manipulation of control in the two studies is different in nature. In their study, (Berberian et al., 2012), manipulated control along a single dimension (i.e. automaticity level). However, in the current study, control is varied at two different levels (at goal level, and at perceptual-motor level). A second difference is that, in their study, the authors report a strong correlation between the two measures of agency and we find dissociation between the two measures when subjects are able to achieve control at the goal level (hit the target). Both the studies, along with others (Jordan, 2003; Desantis, Cedric, \& Waszack, 2011; Kumar \& Srinivasan, under review) provide evidence that control is correlated to the amount of intentional binding and plays a key role in determining sense of agency.

## Control and Sense of Authorship

For the sense of authorship, our hypothesis was not completely supported by the results. Participants did show an increase in rating with control for the sense of authorship, when participants missed the target, thus supporting first part of our hypothesis. But, this increase in sense of authorship with control was also present when participants accurately hit the target indicating that the explicit measure of sense of agency is independent of control when participants hit the target. In combination, these results suggest dissociation between the intentional binding and sense of authorship. This dissociation has also been found in earlier studies (Ebert \& Wegner, 2010), but the underlying mechanisms are not yet clear (Haggard \& Moore, 2010; (Moore \& Obhi, 2012).

## Underlying Mechanism

Haggard and Moore (2010), commenting on the Ebert and Wegner (2010) study raised certain issues that remain unanswered from the study. Firstly, whether the exact mechanism of consistency is retrospective or prospective in nature? In the current study, control is predictive in nature, the control at perceptual-motor level was based on the prior expectation of participant when they moved joystick to aim at the target. At the goal level, participant's expectation of the outcome occurred prior to the event (as with congruency in the case of Ebert and Wegner. But, unlike congruency (Ebert \& Wegner, 2010), the effect at the goal level occurred immediately before (or at the time) they pressed the trigger. Hence, goal level control can also be assumed to be predictive in nature. This suggests that the mechanism linking control and intentional binding is influenced by predictive processes.
A second issue was the exact causal nature of the link between intentional binding and sense of authorship. Our results suggest that intentional binding is sensitive to the hierarchical levels of event-control. In comparison, the sense of authorship seems to be less sensitive to the eventcontrol hierarchy. Hence, we would like to suggest that intentional binding and sense of authorship are not causally linked to each other, but are rather mediated by amount of
control at different levels that can be exercised by participants.

## Conclusions

We have shown that the theory of event control provides a successful framework to understand sense of agency. We suggest that both implicit and explicit aspects of sense of agency are mediated by hierarchical levels of control, but differently. The dissociation between implicit and explicit aspects of agency can be attributed to a difference in the way hierarchical nested control at multiple levels mediate the different aspects of agency. We have also confirmed that interval estimation task can be used to successfully measure intentional binding.

If it is actually the case that these nested control loops mediate agency, what exactly causes these control loops to mediate various aspects of agency in a different fashion? Possible answers might lie in the nature of control and the potential perception-action interactions between the organism and the environment that are dependent on the control. The study provides a pathway to understanding differences in sense of agency and further experiments would enable to naturalize and understand self.

## References

Berberian, B. S., Blaye, P., \& Haggard, P. (2012). Automation technology and sense of control: A window on human agency. PLoS One, 3, e34075.
Desantis, A., Cedric, R., \& Waszack, F. (2011). On the infuence of causal beliefs on feeling of agency. Consciousness and Cognition, 20, 1211-1220.
Ebert, J. P., \& Wegner, D. M. (2010). Time warp: Authorship shapes the perceived timing of actions and events. Consciousness and Cognition, 19, 481-489.
Farrer, C., \& Frith, C. (2002). Experiencing oneself vs another person as being the cause of an action: the neural correlates of the experience of agency. Neuroimage, 15, 596-603.
Farrer, C., Frank, N., Georgieff, N., Frith, C. D., Decety, J., \& Jeannerod, M. (2003). Modulating the experience of agency: a positron emission tomography study. Neuroimage, 18(2), 324-333.
Haggard, P., \& Clark, A. (2003). Intentional action: Conscious experience and neural prediction. Consciousness and Cognition, 12(4), 695-707.
Haggard, P., \& Moore, J. (2010). Intentional binding and higher order agency experience. Consciousness and Cognition, 19(1), 490-491.
Haggard, P., Clark, S., \& Kalogeras, J. (2002b). Voluntary action and conscious awareness. Nature Neuroscience, 5(4), 382-385.
Humphreys, G., \& Buehne, M. (2009). Magnitude estimation reveals temporal binding. Journal of Experimental Psychology, 35(5), 1542-1549.
Jordan, S. J. (2003). Emergence of self and other in perception and action: An event-control approach. Consciousness and Cognition, 4(1), 633-646.

Kumar, D., \& Srinivasan, N. (2012). Hierarchical eventcontrol and subjective experience of agency. Frontiers in Psychology, 3:410.
Kumar, D., \& Srinivasan, N. (under review). Naturalizing self: A hierarchical event-control approach.
Moore, J. W., \& Obhi, S. S. (2012). Intentional binding and sense of agency: A review. Consciousness and Cognition, 21, 1211-1220.
Moore, J. W., Lagando, D., Deal, D. C., \& Haggard, P. (2009). Feeling of control: Contingency determines experience of self. Cognition, 110, 279-283.
Pacherie, E. (2011). Self-Agency. In S. Gallagher, Oxford Handbook of the Self (pp. 442-464), Oxford: Oxford University Press.
Synofzik, M., Vosgerau, G., \& Newen, A. (2008). Beyond the comparator model: a multifactorial two-step account of agency. Consciousness and Cognition, 17(1), 219-239.
Shrager, J., \& Langley, P. (Eds.) (1990). Computational Models of Scientific Discovery and Theory Formation. San Mateo, CA: Morgan Kaufmann.

# Methods for Classifying Errors on the Raven's Standard Progressive Matrices Test 

Maithilee Kunda (mkunda@gatech.edu) ${ }^{1}$<br>Isabelle Soulières (soulieres.isabelle@uqam.ca) ${ }^{2}$ Agata Rozga (agata@gatech.edu) ${ }^{1}$ Ashok K. Goel (goel@cc.gatech.edu) ${ }^{1}$<br>${ }^{1}$ School of Interactive Computing, Georgia Tech, 85 Fifth Street NW, Atlanta, GA 30308 USA<br>${ }^{2}$ Département de Psychologie, Université du Québec à Montréal C.P. 8888 succursale Centre-ville, Montréal (Québec) H3C 3P8 Canada


#### Abstract

Although many psychometric tests, like Raven's Progressive Matrices, are commonly evaluated according to total score, additional variables can lend insight into the underlying cognitive processes. We examine conceptual errors on the Raven's Standard Progressive Matrices (SPM) test. We present a complete classification of error types on the SPM using a two-kind coding scheme, yielding $\geq 95 \%$ inter-rater reliability. We also examine how to extract error data from a computational model, and we present a method for measuring errors through systematic ablation to create a "population" of models whose performance can be examined as a group. We present a preliminary analysis of error patterns on the SPM from typically developing individuals, individuals diagnosed with autism, and a computational model called ASTI. We discuss what the error patterns suggest regarding cognition on the SPM and routes towards improving the ASTI model.


Keywords: ablation experiments; computational modeling; error patterns; mental imagery; psychometrics; Raven's Progressive Matrices; visual representations.

## Introduction

Raven's Progressive Matrices (RPM) is a widely used series of intelligence tests that consist of multiple choice visual analogy problems, as in Fig. 1. Each problem contains a matrix of geometric figures with one figure missing; the correct missing figure that completes the matrix pattern must be selected from a set of answer choices.

Performance is generally measured in terms of overall score, i.e. number correct, which can then be used as an index into normative test data to determine an IQ score or percentile ranking for that individual. While total score is certainly an important variable, serving as a coarse measure of an individual's overall ability, there are alternative dimensions of performance that may provide a finer-grained view of an individual's cognitive processing:

1) Per-item accuracy, e.g. differential item functioning, takes into account potential variation even when individuals may obtain the same total score (Facon \& Nuchadee, 2010; Lynn, Alik, \& Irwing, 2004; Van Herwegen, Farran, and Annaz, 2011).
2) Reaction time can be used to understand the stages of processing in solving a single item (Bethell-Fox, Lohman, \& Snow, 1984) or to compare performance
across individuals or groups (Soulières et al., 2009).
3) Patterns of errors-for a problem answered incorrectly, which of the given distracters is selected?-have been studied as a window into cognitive strategy (Bromley, 1953; Gunn \& Jarrold, 2004; Miller \& Raven, 1939; Van Herwegen, Farran, and Annaz, 2011; Vodegel Matzen et al., 1994).
All of these dimensions represent measurable aspects of the "output" of a human cognitive system taking the RPM test. The "input" to such a system, in addition to the test itself, can be conceptualized as the set of cognitive functions drawn upon while solving the test. Unlike the output measures, it is difficult to directly measure cognitive functioning. Some studies have used eye-tracking as a measure of visual attention (Bethell-Fox, et al., 1984; Carpenter, Just, \& Shell, 1990), and some have used verbal reporting protocols (Carpenter et al., 1990) though verbal report may bias the cognitive strategies used by participants (DeShon, Chan, \& Weissbein, 1995).

Another way to elucidate these invisible cognitive mechanisms is to construct computational models of various aspects of RPM problem solving and then inspect these models in relation to human behavioral data. Aspects of RPM (or RPM-like) problem solving that have been investigated using computational models include:


Figure 1: Example of an RPM-like problem.

1) Knowledge representation, i.e. visual versus verbal representations of problem content (Hunt, 1974; Kunda, Goel, \& McGreggor, 2013; McGreggor, Kunda, \& Goel, 2011).
2) Goal-subgoal maintenance (Carpenter et al., 1990).
3) Problem-solving process, i.e. constructive matching (mentally constructing the answer and then selecting an answer choice) versus response elimination (inspecting each answer choice to find the best fit) (Bethell-Fox et al., 1984; Lovett \& Forbus, 2012).
4) Answer selection process in terms of confidence (McGreggor \& Goel, 2012) or probability (Little, Lewandowsky, \& Griffiths, 2012).
However, in the extant literature on computational models of the RPM, many models tend to focus on only one measure of output performance: total score. We believe it is not only valuable but critical that models examine the other dimensions of "output" that we have mentioned, in order to investigate how models relate to human cognition at increasingly fine-grained levels of resolution.

In this paper, we focus on one such "output" measureerror patterns-and one computational model-the ASTI model, described in detail in a previous publication (Kunda et al., 2013). We first present an operationalization of error patterns on the Raven's Standard Progressive Matrices (SPM) test, in the form of a two-kind classification of conceptual error types. Then, we briefly summarize the algorithms and performance of the ASTI model. Finally, we present a method for analyzing the errors made by a computational model, and we give preliminary results based on a comparison of the errors made by the ASTI model against human error data from typically developing individuals and individuals diagnosed with autism, along with an evaluation of what these differences in error patterns can tell us about cognitive processing on the RPM.

## Types of Conceptual Errors on the SPM

One way to examine errors on an RPM test is to look at which distracter is chosen in comparison to those most frequently chosen (Thissen, 1976; van der Ven \& Ellis, 2000). However, many studies have shown that errors can also be classified according to conceptual type, which may provide additional insight into what it means when a certain error is made (Forbes, 1964; Horner \& Nailling, 1980).

However, there is currently one significant barrier to the widespread analysis of error patterns on the SPM test; while the published manuals for two of the RPM tests, the Colored Progressive Matrices (CPM) and the Advanced Progressive Matrices (APM), include taxonomies of conceptual error types, the manual for the Standard Progressive Matrices (SPM) does not (Raven, Raven, \& Court, 2003). Vodegel Matzen et al. (1994) attempted to use the APM error type classifications on a portion of the SPM, but inter-coder reliability reached only about $70 \%$. The authors concluded that classification of SPM distracters seemed "problematic" in that there did not seem to be a systematic methodology used for constructing distracters.

The taxonomies given in the CPM and APM manuals (Raven et al., 2003), although having different labels, seem to represent the same four notions of error types. We now present a synthesized description of these four error types which, along with criteria used to classify a particular distracter, are also summarized in Table 1.

1) Incomplete correlate (IC) errors occur when the chosen distracter is almost, but not quite, correct. For example, some IC distracters have the correct shape but the wrong texture, as exemplified by distracter \#1 in Fig. 1. These kinds of errors are made when a test-taker more or less "gets" the problem, in terms of identifying the relevant matrix relationships, but then fails to fully account for all of the details when selecting an answer.
2) Repetition (R) errors occur when the chosen distracter copies a matrix entry adjacent to the blank space, as shown by distracters \#3 and \#8 in Fig. 1. The choice of an R distracter may represent perseveration or fixation on the matrix entries, in which an answer is selected via perceptual matching between the answer choices and the matrix entries closest to the blank space.
3) Difference (D) errors occur when the chosen distracter is qualitatively different in appearance from the other choices. D distracters include completely blank entries, as exemplified by distracter \#2 in Fig. 1, as well as those that have extraneous or complex shapes not found in the matrix. A D distracter might be chosen because it visually "pops" from among the other choices.
4) Wrong principle (WP) errors occur when the chosen distracter is a copy or composition of elements from various matrix entries, as exemplified by distracters \#4 and \#6 in Fig. 1. A WP distracter might be chosen if the test-taker fails to educe the correct relationship from the matrix and combines the entries according to some other rule or relationship to produce an answer.

## Two-Kind Taxonomy and Coding Results

The main difficulty we observed in coding SPM distracters is that the same distracter often seems to fall under multiple categories, e.g. it might represent a repetition as well as an incomplete correlate; this difficulty was shared by VodegelMatzen, et al. (1994). From this observation, we realized that the four error types listed above actually represent two orthogonal classifications of distracters:

Kind I: Relationship of distracter to matrix entries: Repetition, difference, and wrong principle errors all have to do with how a distracter is related to information in the matrix and in the other answer choices, without any regard to the content of the correct answer choice. In particular, errors of the first kind assume the participant is attending to irrelevant or erroneous aspects of the problem, and that they are not able to discover even a partial solution.

Kind II: Relationship of distracter to correct answer: Incomplete correlate errors have to do with how a particular distracter is related to the correct answer choice. These errors assume the participant correctly guesses some part of the solution but does not quite attain the correct answer.

Table 1: Criteria for classifying distracters on the SPM.

| Error type | $\#$ | Criteria |
| ---: | ---: | :--- |
| Kind I: | 1 | Repetition of matrix entry to left of blank space |
|  | 2 | Repetition of matrix entry above blank space |
|  | 3 | Repetition of matrix entry to top-left of blank space |
|  | 4 | Filled completely white or black |
| Kind I: | 5 | Union of matrix entries or aspects of them, so that union has more components than any single matrix entry |
| Difference | 6 | Maximizes some feature value or makes it more complex |
|  | 7 | Differs qualitatively from matrix and other answers, or contains information not found anywhere in matrix |
| Kind I: | 8 | Copy of matrix entry not adjacent to blank space |
| Wrong | 9 | Rotation/reflection of matrix entry |
| Principle | 10 | Other transformations or combinations of matrix entries or aspects of them, including negative images |
|  | 11 | Negative (color-inversion) of correct answer |
|  | 12 | Change only in fill, texture, or style |
| Kind II: | 13 | Rotation/reflection of correct answer |
| Incomplete | 14 | Change only in spatial layout of elements |
| Correlate | 15 | Change only in size or scale, in either or both dimensions (allowing for feature-wise scaling) |
|  | 16 | Change only in number of discrete elements (allowing for slight changes in layout) |
|  | 17 | Incomplete, with missing element or portion |

Using this two-kind taxonomy, two raters independently coded all 432 distracters on the $\mathrm{SPM}^{1}$ in two separate passes, first for Kind I and then for Kind II. Kind I classification used a copy of the test booklet in which no answers had been marked, and raters assigned every distracter to one of categories \#1-10 in Table 1. Kind II classification used another test booklet copy in which the correct answers had been marked and the matrix portions of each problem had been cut off, so only the answer choices were visible; raters assigned each distracter to one of categories \#11-17 in Table 1, or left it uncategorized.

Initial agreement between the two raters was $82 \%$ for Kind I errors and (coincidentally) $82 \%$ for Kind II errors. Kappa coefficients were calculated to test for independence between raters. The kappa values were 0.79 for Kind I errors and 0.67 for Kind II errors.

Discrepancies were resolved during a negotiation phase between the two raters. Each discrepancy was discussed, and each rater presented a rationale for the classification. It was found that there were several systematic discrepancies easily resolved by making the coding criteria more specific. For example, Criterion \#5 in Table 1 was modified to specify that this type of distracter had to have more elements in it than any entry in the matrix, which was not originally part of the criterion. Table 1 shows the final coding criteria, after these changes had been incorporated.

After the negotiation phase, rater agreement was recalculated. Post-negotiation agreement was $95 \%$ for Kind I errors and $98 \%$ for Kind II errors. Remaining differences were resolved by the primary rater based on consideration of the conceptual error type intended to be captured.

Fig. 2 shows the overall proportions of error types across all distracters of the SPM. Interestingly, there is roughly the same proportion of incomplete correlate distracters as

[^320]correct answers, and all remaining distracters are divided nearly evenly among the three remaining error types.


Figure 2: Proportions of each error type on the SPM.

## The ASTI Model

In previous work (Kunda et al., 2013), we presented a computational model of problem solving on the RPM, the Affine and Set Transformation Induction (ASTI) model. This model was constructed in order to investigate problem solving on the RPM using visual mental representations. All extant computational RPM models had previously relied on propositional forms of representation (e.g. Carpenter et al., 1990), despite a breadth of evidence from human studies suggesting that problem solving can proceed using either visual or verbal forms of representation (see Kunda et al., 2013, for a summary of these studies).

The ASTI model also has implications for a recent study of RPM performance in individuals diagnosed with autism, which found that these individuals seemed to use predominantly visual strategies (Soulières et al., 2009), in line with other empirical evidence showing a visual cognitive bias in autism (Kunda \& Goel, 2011).

The ASTI model uses purely visual representations in the form of pixel-based images along with affine and set transformations designed to emulate the types of operations observed in studies of human mental imagery. The model uses a constructive matching approach; first, it examines different subsets of the matrix entries (each an individual image), under each of these transforms to induce a "best-fit" overall transform. Then, the ASTI model applies this bestfit transformation to the remaining matrix entries to generate a predicted answer image. Finally, this predicted answer is compared to each answer choice to select the best match.

```
Initialization
        Read matrix entries into list of i mages M
        Read answer choices into list of i mages A
    For any two i mages a and b, define a
    similarity metric S(a, b) }->z\in{[0,1
4 Define set of base transforms T
5 Define set of analogies lo }->\mp@subsup{|}{1}{\prime}\mathrm{ , where lo
    contains i mage sequences representing
    complete row, column, or di agonal lines in
    the matrix, and for each io \inIO, I I has
    the corresponding i mages i_ representing
    the parallel partial line in the matrix
Transformation Induction
    For each i mage sequence io Elo, induce the
    best.fit composite transform tc:
        For each base transform t \in T:
            Apply t to the first i mage(s) in io
            to produce i mage it
            Search all possible translation
            offsets (x, y) between io and it to
            find the best match, as calculated
            by S(iol (x,y), it)
5 Select the'best.fit base transform
            tb as per S, as calculated above
        tc is then a composition of ts and the
        translation offset ( }x,y\mathrm{ )
    Obtain a final transform tf by selecting
    that tc which produces the best average
    fit, across each subset of parallel io flo
Candidate Prediction and Answer Selection
1 Choose i mage sequence io that results in
    the best-fit tf, according to s as
    calculated in the previous step
2 Apply tf to corresponding partial i mage
    sequence i_ El_ to produce candidate
    answer i mage i c
3 For each answer choice ia }\inA, comput
    similarity S(ic, ia)
4 \text { Select the best-fit answer choice iA as}
    per S, as calculated above
```

Figure 3: Algorithm used by the ASTI model.

## Obtaining Error Data from the ASTI Model

The current version of the ASTI model correctly answers 50 out of 60 problems on the SPM. One difficulty with high performing computational models such as ASTI is that it is not immediately clear how errors made by the model might be analyzed in a meaningful way, as error data can only be collected on 10 of the 60 problems.

We use a method for obtaining error data from a computational RPM model through model ablation (Cohen \& Howe 1988). The ASTI model uses affine transforms (rectilinear rotations and reflections), as well as addition, subtraction, and pair-wise image composition (union,
intersection, etc.); the model also inspects the matrix according to rows, columns, and diagonals. By removing access to subsets of these mechanisms, we can observe the errors made by general classes of ASTI configurations.

Table 2 lists mechanisms used for $2 \times 2$ matrices (found in Sets A and B of the SPM) and $3 \times 3$ matrices (found in Sets C through E of the SPM). Ablating combinations of these mechanisms yields 96 different model configurations, whose total scores range from 15 to 50 correct.

Table 2: Mechanisms for Ablation in the ASTI Model

| Type | Image sets | Transforms |
| ---: | :--- | :--- |
| $\mathbf{2 \times 2}$ | 1. Rows | 1. Identity <br> matrices |
|  | 2. Columns | 2. Rotation/reflection <br> 3. Addition/subtraction |
| $3 \times 3$ | 1. Rows | 1. Identity |
|  | 2. Columns | 2. Rotation/reflection |
|  | 3. Diagonals | 3. Addition/subtraction <br> 4. Composition |

## Analysis of Error Patterns

Using the new classification of error types on the SPM that we described above, we conducted an analysis to compare the error patterns of typically developing individuals, individuals diagnosed with autism, and the ASTI model.

Human data were obtained from previous studies done at the Hôpital Rivière-des-Prairies in Montreal, Canada. Participants diagnosed with autism received a best-estimate multidisciplinary diagnosis after evaluation with standard diagnostic instruments, the ADOS and ADI-R (Lord et al., 1999; Rutter et al, 2003).

Using a cutoff of 17 years, participants were grouped into children and adults. Data included answer choices given for each SPM problem, including a few instances in which no answer was given. (One participant in the autism group was excluded from analyses, as he had selected answer choice " 1 " for more than half of the problems.)

Table 2 summarizes total SPM score, age, and Wechsler full-scale IQ information for these groups. While total SPM scores between TD and AUT groups are not significantly different, the ASTI SPM scores are significantly lower. This introduces a potential confound, if error types are dependent on overall ability. To address this issue, we conducted an analysis using three subgroups (TD children, AUT children, and the ASTI model) individually matched on total SPM score. Table 3 gives data on these subgroups.

We looked at the proportions of each error type that were made on the entire SPM test, averaged across participants in each group. Fig. 3 presents the results of these comparisons for the score-matched subgroups. Results for the full groups of children and adults were similar, and so we present detailed results of this first analysis only.

There is no significant difference in overall error distributions between the TD and AUT groups, $\chi^{2}(N=826)$
$=1.89, p=0.60$, whereas the error distribution from the ASTI model differs significantly from each of the human groups, $\chi^{2}(N=826)=91.62, p<0.001$ for TD, and $\chi^{2}(N=$ $826)=98.69, p<0.001$ for AUT.

A one-way ANOVA was used to test for differences in error proportions among the three groups. Proportions differed significantly for repetition, $F(2,111)=6.20, p=$ 0.003 , and difference errors, $F=32.03, p<0.001$, but did not differ significantly for incomplete correlate, $F=0.14, p$ $=0.87$, or wrong principle errors $F=1.61, p=0.20$.

Table 2: Demographic data for full participant groups. Values as shown as: mean (standard deviation).

|  | Children |  | Adults |  | Model |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  | TD | AUT | TD | AUT | ASTI |
| N | 54 | 108 | 52 | 44 | 96 |
| SPM | 42.61 | 37.43 | 50.69 | 48.43 | 32.57 |
| score | $(9.79)$ | $(12.17)$ | $(5.38)$ | $(9.64)$ | $(9.74)$ |
| Age in | 11.96 | 11.02 | 22.98 | 26.80 | $\mathrm{n} / \mathrm{a}$ |
| years | $(3.40)$ | $(2.99)$ | $(4.28)$ | $(6.72)$ |  |
| IQ | 109.82 | 84.38 | 106.91 | 97.61 | $\mathrm{n} / \mathrm{a}$ |

Note: IQ data was not available for all participants.
Table 3: Demographic data for score-matched subgroups.

|  | Children |  | Model |
| :---: | :---: | :---: | :---: |
|  | TD | AUT | ASTI |
| N | 38 | 38 | 38 |
| SPM | 38.26 | 38.26 | 38.29 |
| score | $(8.07)$ | $(8.09)$ | $(8.07)$ |
| Age in | 11.11 | 10.76 | $\mathrm{n} / \mathrm{a}$ |
| years | $(3.30)$ | $(2.71)$ |  |
| IQ | 106.08 <br> $(9.08)$ | 88.83 <br> $(18.79)$ | $\mathrm{n} / \mathrm{a}$ |

Note: IQ data was not available for all participants.


Figure 3: Proportions of each error type made on the SPM by typically developing (TD) individuals, individuals diagnosed with autism (AUT), and the ASTI model.
(Error bars represent one standard deviation.)

## Discussion

We discuss results from two perspectives. First, what does this analysis tell us about the error patterns shown by the TD versus AUT groups? Second, what does this analysis tell us about the error patterns shown by the ASTI model?

First, we see that the distribution of conceptual errors made on the SPM does not seem to differ significantly between the TD and AUT groups. Following a prior study suggesting that individuals with autism tend to use visual strategies to solve these kinds of problems (Soulières et al., 2009), one interpretation may be that looking at error types of this kind does not by itself indicate potential differences in problem solving modality (i.e. visual/verbal). However, as TD individuals most likely use a combination of visual as well as verbal strategies on the SPM, another, currently unexplored, hypothesis is that differences in error types may only surface for problems solved verbally by the TD group and visually by the autism group. If this is the case, then detecting such differences would require a finer-grained analysis of error types on various subsets of SPM problems instead of across the entire test as a whole.
To address the latter question, comparisons of errors between human participants and the ASTI model show agreement on two types of errors (incomplete correlate and wrong principle) and discrepancies on the other two types (repetition and difference). Looking at these differences in error patterns lends valuable insight into how specific aspects of the ASTI model affect its overall behavior and simultaneously suggests concrete avenues for improving the cognitive fidelity of the ASTI model.

First, with regard to the relative increase in repetition errors, the ASTI model predicts answers based on the matrix entries adjacent to the blank space. Thus, it is likely that its prediction is visually similar to an adjacent matrix entry, leading to an error of repetition. While humans do often make repetition errors, they also likely draw upon more aspects of the matrix when selecting an answer, which the ASTI model could also be modified to do.

Second, regarding the relative scarcity of difference errors made by the ASTI model, recall that these errors are made according to how a particular distracter might seem different or more complex than the other answer choices. Making difference errors thus should only affect test-takers using a response elimination strategy, i.e. looking at the answer choices as a set at the start of or during problem-solving. Test-takers using a constructive matching strategy already have an answer in mind before moving to inspect the answer choices, and if this answer is constructed by examining and combining matrix entries, it would likely be similar to these entries and thus not be likely to lead to a difference error.

Difference errors may thus be considered a result of testtakers fixating on the visual salience of one particular answer choice over another. The ASTI model currently does not contain mechanisms to detect salience or perform response elimination; the addition of these mechanisms will improve the fidelity with which problem-solving strategies used by the ASTI model mirror those of humans.

## Conclusion

The main motivation for this work stems from the view that conceptual types of errors made on the Raven's tests can serve as an important additional measure of behavioral performance, above and beyond total score. To this end, this paper makes two primary contributions.

The first major contribution is the new classification of error types on the SPM using a two-kind approach that yielded $\geq 95 \%$ inter-rater reliability. This classification should have considerable utility for further studies of human or machine SPM performance, and it adds a significant new component of information for the RPM family of tests, as both the CPM and APM tests already have such error classifications, but the SPM previously did not. One area of future work is to examine the error patterns made by humans on different subsets of test problems, instead of across the test as a whole, to achieve a finer-grained analysis of what kinds of errors people make on certain problems.

The second major contribution is the methodology presented for measuring the conceptual errors made by a computational model on the RPM. Looking at the errors made by the ASTI model has led us to propose two modifications to improve its cognitive fidelity: first, the model should consider additional aspects of the matrix when generating answer predictions, in addition to just the adjacent entries, and the model should be able to adopt a response elimination strategy and also be susceptible to the visual salience of particular answer choices.

Neither of these observations would have been possible by looking at total score alone, or even at the pattern of correct vs. incorrect answers. Future work on test-taking by humans and computational models should continue to look at multiple performance measures, beyond just total score, to fully understand performance and cognitive implications.

## Acknowledgments

We are grateful to the US National Science Foundation for its support through NSF (RI) Grant \#1116541, "Addressing visual analogy problems on the Raven's intelligence test," and through a GRFP fellowship, and to the US Office of Naval Research for support through an NDSEG fellowship. We also thank Isabelle Simard, Patricia Jelenic, Bryan Wiltgen, and Keith McGreggor for their assistance.

## References

Bethell-Fox, C., Lohman, D., \& Snow, R. (1984). Adaptive reasoning: Componential and eye movement analysis of geometric analogy performance. Intelligence, 8, 205-238.
Bromley, D. (1953). Primitive forms of response to the Matrices test. Journal of Mental Sciences, 99, 374-393.
Carpenter, P., Just, M., \& Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the RPM test. Psychol. Rev., 97, 404-431.
Cohen, P., \& Howe, A. (1988). How evaluation guides AI research. AI Magazine, 9 (4), 35-43.
Dawson, M., Soulieres, I., Gernsbacher, M., \& Mottron, L.
(2007). The level and nature of autistic intelligence. Psychological Science, 18 (8), 657-662.
DeShon, R. P, Chan, D., \& Weissbein, D. A. (1995). Verbal overshadowing effects on Raven's Advanced Progressive Matrices: Evidence for multidimensional performance determinants. Intelligence, 21 (2), 135-155.
Facon, B., \& Nuchadee, M. (2010). An item analysis of Raven's CPM among participants with Down syndrome. Research in Developmental Disabilities, 31, 243-249.
Forbes, A. (1964). An item analysis of the advanced matrices. British J. Educational Psychology, 34, 223-236.
Gunn, D., \& Jarrold, C. (2004). Raven's matrices performance in Down syndrome: Evidence of unusual errors. Res. Developmental Disabilities, 25 (5), 443-457.
Horner, J. \& Nailling, K. (1980). Raven's Colored Progressive Matrices: Interpreting results through analysis of problem-type and error-type. Clinical Aphasiology Conference, Minneapolis, MN. 226-239.
Kunda, M., \& Goel, A. K. (2011). Thinking in Pictures as a cognitive account of autism. Journal of Autism and Developmental Disorders. 41 (9): 1157-1177.
Kunda, M., McGreggor, K., \& Goel, A. K. (2013). A computational model for solving problems from the RPM intelligence test using iconic visual representations. Cognitive Systems Research, 22-23, 47-66.
Lord, C., Rutter, M., DiLavore, P., \& Risi, S. (1999). The ADOS-G. Los Angeles, CA: Western Psychol. Services.
Lovett, A., \& Forbus, K. (2012). Modeling multiple strategies for solving geometric analogy problems. Proc. $34^{\text {th }}$ Annual Conference of the Cognitive Science Society.
McGreggor, K., \& Goel, A. K. (2012). Fractal analogies for general intelligence. Artificial General Intelligence, LNAI 7716, 177-188.
Miller, F., \& Raven, J. (1939). The influence of positional factors on the choice of answers to perceptual intelligence tests. British Journal of Medical Psychology, 18, 35-39.
Raven, J., Raven, J. C., \& Court, J. H. (2003). Manual for Raven's Progressive Matrices and Vocabulary Scales. San Antonio, TX: Harcourt Assessment.
Rutter, M., Le Couteur, A., \& Lord, C. (2003). ADI-R. Los Angeles, CA: Western Psychological Services.
Soulières, I., Dawson, M., Samson, F., Barbeau, E., Sahyoun, C., et al. (2009). Enhanced visual processing contributes to matrix reasoning in autism. Human Brain Mapping, 30 (12), 4082-4107.
Thissen, D. M. (1976). Information in wrong responses to the Raven's Progressive Matrices. Journal of Educational Measurement, 13 (3), 201-214.
van der Ven, A.H.G.S., \& Ellis, J.L. (2000). A Rasch analysis of Raven's standard progressive matrices. Personality and Individual Differences, 29 (1), 45-64.
Van Herwegen, J., Farran, E., \& Annaz, D. (2011). Item and error analysis on Raven's CPM in Williams Syndrome. Research in Developmental Disabilities, 32, 93-99.
Vodegel-Matzen, L., van der Molen, M., \& Dudink, A. (1994). Error analysis of Raven test performance. Personality and Individual Differences, 16 (3), 433-445.

# Experts' Explanations Engage Novices in Deep-Processing 

Andreas Lachner (andreas.lachner@ezw.uni-freiburg.de) Matthias Nückles (matthias.nueckles@ezw.uni-freiburg.de)

Department for Educational Science, University of Freiburg

Rempartstrasse 11, 79098 Freiburg, Germany


#### Abstract

Experts and intermediates fundamentally differ in the ways they explain subject matter to novices. Experts provide less details but in a highly coherent format, whereas intermediates provide many additional details but in a format with low coherency. In a recent study, we found that experts' explanations enabled novices to acquire more transferable and flexible knowledge as opposed to explanations by intermediates mainly due to the higher coherence of experts' explanations. In order to investigate more directly how experts' and intermediates' explanations differently triggered novices' processing of the explanations, we conducted a think-aloud study. Results indicated that novices learning with an expert's explanation processed the explanations deeper than novices with intermediates’ explanations. In line with this, deep processing was significantly related to novices' transfer. Thus, expertise can be regarded as an essential prerequisite for generating effective instructional explanations that engage novices to process the subject matter deeply and to generate transferable knowledge.


Keywords: expertise; processing; instructional explanations; transfer.

## Introduction

Experts and intermediate students fundamentally differ in the ways they explain subject matter to novices. Figure 1 shows two propositional representations, one intermediate student's explanation and one expert's explanation, taken from a study by Lachner, Gurlitt and Nückles (2012). These propositional representations about bacterial endocarditis, an inflammation of the heart valves, structurally differ in several important respects.


Figure 1: Graphical representation of an intermediate's explanation and an expert's explanation.

The typical expert's explanation consisted of only a few, mainly advanced, clinical concepts $(N=24)$. Beyond that, the expert's explanation was very coherent, as she related all
explanatory concepts to each other, resulting in a single very coherent chunk of knowledge. In marked contrast, the typical intermediate’s explanation provided many concepts ( $N=52$ ). Although there was one interrelated chunk about pathophysiological processes of bacterial endocarditis, the intermediate was less likely to relate basic pathophysiological concepts with advanced concepts, which resulted in many fragmented knowledge blocks $(N=8)$.

These two explanations nicely illustrate well-known differences between experts and intermediates. For instance, research on categorization shows that experts tend to organize their knowledge around abstract principles, which allows them to integrate their knowledge in a more coherent manner, whereas novices organize their knowledge around superficial features, which results in less coherent knowledge structures (Chi, Feltovich, \& Glaser, 1981; Rottman, Gentner, \& Goldwater, 2012).

In the same vein, in the medical domain, it has been shown that experts subsume basic medical concepts under advanced concepts, which results in very condensed schemata, whereas intermediates rather rely on detailed knowledge, as they have not yet acquired these advanced clinical principles. This subsumption process is also known as knowledge encapsulation (Boshuizen \& Schmidt, 1992, Rikers, Schmidt, \& Boshuizen, 2000; 2002).

Lachner et al. (2012) found that these effects for coherence and knowledge encapsulation also hold true for instructional explanations, specifically explanations written for novice medical students. Compared to intermediates, medical experts wrote more coherent and equally more encapsulated explanations, meaning that they omitted more details in their explanations. However, both intermediate students and experts used the same amount of advanced concepts. Apparently, experts adapted their choice of words, but not the way they would structure an explanation.

## Learning from Instructional Explanations

As explanations by experts and intermediates fundamentally differed on the level of coherence and encapsulation, explanations by experts and intermediates might also affect student learning differently. For instance, Hinds, Patterson and Pfeffer (2001) investigated how the instructor's domain expertise affected novices' learning in the domain of electrical engineering. More specifically, they examined how novices studying an intermediate's explanation differed from novices studying an expert's explanation with regard to their performance on near transfer and far transfer tasks. Results indicated an interaction effect. Although novices
with an intermediate's explanations outperformed novices with an expert's explanation on near transfer tasks, there was a clear benefit for novices with experts' explanations on far transfer tasks. In a related study, Boekhout, van Gog, van de Wiel, Gerards-Last, and Geraets (2010) showed that worked examples constructed by experts led to larger benefits for novices in transfer tasks than worked examples constructed by intermediates. However, with regard to the acquisition of factual knowledge, novices with experts’ worked examples did not differ from novices learning from intermediates' worked examples.

## Beneficial Features of Instructional Explanations

Bridging findings from expertise research and tutoring research, Lachner and Nückles (2013) investigated which expertise-related textual features of explanations accounted for the better transfer of novices learning with experts' explanations. Specifically, they examined how coherence and encapsulation of the instructors' explanations, as coherence and encapsulation were selective indicators for expertise (Lachner et al., 2012; Rikers et al., 2002; Rottman et al., 2012), affected novices’ learning outcomes. Similarly to Hinds et al. and Boekhout et al., Lachner and Nückles found that novices studying with experts' explanations significantly outperformed novices with intermediates' explanations on transfer tasks. At the same time, in line with Boekhout et al., they did not find a significant difference between experts’ and intermediates’ explanations regarding novices' factual knowledge.

More importantly, Lachner and Nückles (2013) conducted a mediation analysis to investigate whether encapsulation, as measured by the omission of detailed knowledge, or coherence, as measured by the number of isolated fragments (see Figure 1), accounted for novices’ transfer. Results indicated that the degree of encapsulation had no effect on novices' transfer, whereas explanatory coherence clearly mediated the effect of instructors' expertise on novices' transfer. Therefore, the authors could show that it was the coherence of experts' explanations that enabled novices to transfer their acquired knowledge to other medical tasks.

Nevertheless, although the Lachner et al. study suggests that explanatory coherence fostered novices' transfer, they did not examine which learning processes were provoked by experts' versus intermediates' explanations that could explain the transfer effect.

## Processing of Instructional Explanations

Bransford and Schwarz (1999) proclaimed that for flexible transfer, learning with "understanding" is necessary. Studies by Gilabert, Martinez and Vidal-Abarca (2005) and Linderholm et al. (2001) support this view, as they found that the high coherence of texts fostered students' active processing of the text material. As coherent explanations highlighted important causal relations between concepts, the coherence of explanations probably served as a valuable scaffold to engage students’ processing.

Text processing can be regarded as the construction and integration of multiple independent representations of a text (Kintsch, 1988; Kintsch, 2004). First, learners construct a text base which contains the essential meaning of the text, mainly by translating the text into propositions, or in other words, by paraphrasing the text and by bridging inferences to connect information within the text (Kintsch, 1988). Second, learners construct a situation model by doing selfexplanations to fill coherence gaps with their prior knowledge. Kintsch (2004) argued especially processing activities, that aim to enrich the situation model, are needed to develop a deep understanding.

In the study by Lachner and Nückles (2013), the mediating variable between instructors' expertise and novices' transfer was explanatory coherence. In the study described here, we examined whether the effect of expertise on novices' learning can be explained by novices' processing activities. In line with text comprehension research, (Gilabert et al., 2005; Linderholm et al., 2001), we assumed that experts' coherent explanations may better promote novices' deep processing and novices' acquisition of flexible knowledge compared to intermediates' less coherent explanations.

## Research Questions and Hypotheses

To investigate novices’ processing activities while learning with experts' and intermediates' explanations, we conducted a think-aloud study. The aims of our study were twofold. First, we wanted to replicate the findings by Lachner and Nückles (2013) that experts' explanations were better suited to foster novices' transfer compared to intermediates' explanations. Second, as we were interested in novices' processing activities, we examined how novices processed explanations by intermediates and experts using a thinkaloud procedure.

## Learning Hypotheses

In line with previous research (Boekhout et al., 2010; Hinds et al., 2001; Lachner \& Nückles, 2013), we hypothesized that novices would benefit more from experts' explanations as opposed to intermediates' explanations in transfer tasks. Experts' coherent explanations would better enable novices to construct an appropriate situation model of bacterial endocarditis and thus enable them to transfer their knowledge of bacterial endocarditis to other tasks (Kintsch, 2004).

For novices' factual knowledge gain, we refrained from making clear predictions, as Boekhout et al. (2010) and Lachner and Nückles (2013) did not find any significant differences between explanations by experts and intermediates. As coherence mainly accounted for the construction of a rich situation model and not for the generation of an adequate text base (Gilabert et al., 2005; Kintsch, 2004), novices with intermediates' explanations could perform just as well in a factual knowledge test as those with experts' explanations.

## Processing Hypotheses

As suggested by research on text comprehension (Gilabert et al., 2005; Linderholm et al., 2001), we assumed that experts' coherent explanations would encourage novices’ deep processing compared to intermediates’ explanations. Therefore, we expected that novices with experts' explanations would outperform novices with intermediates' explanations with regard to the proportion of bridging inferences and self-explanations, whereas intermediates' explanations would trigger novices' paraphrasing. For negative monitoring, we refrained from making clear predictions, as the fewer details in experts' explanations could trigger novices’ monitoring, as well as the lack of coherence of intermediates' explanations.

Beyond that, as Kintsch (2004) suggested, we assumed that novices' transfer was significantly related to novices’ deep processing.

## Method

## Participants

Sixty-eight novices from the University of Freiburg, Germany participated in the study. They were recruited from medicine ( 45 students) and biology programs (23 students). 70.60 \% were female; their mean age was 20.25 ( $S D=1.87$ ). Participants were on average in their first semester ( $S D=1.24$ ) and had not yet attended any courses in cardiology. Participants were financially compensated with 10 Euro for their participation.

## Design

Novices were randomly assigned to one of four explanations about bacterial endocarditis, an infection of the heart valves (two experts’ explanations and two intermediates' explanations). We used a pretest-posttest design with type of explanations, that is, experts' explanations or intermediates' explanations, as independent variables. There were two classes of dependent variables: We analyzed novices’ learning outcomes with both a factual knowledge test that measured novices' knowledge about central concepts and interdependencies of bacterial endocarditis, and with a transfer test that required the participants to apply their acquired knowledge of bacterial endocarditis to other medical phenomena. Additionally, we collected novices' processing activities by means of think-aloud-protocols (i.e. paraphrasing, bridging inferences, self-explanations, and negative monitoring) while studying the explanations.

## Materials

## Case description

We provided the participants with a general case description of a fictitious patient suffering from bacterial endocarditis. It included central findings of laboratory data, and descriptions of symptoms. The case description had been used in previous classical studies on the nature of expertise, as bacterial endocarditis can be regarded as a prototypical
heart disease that requires deep-level knowledge about embolisms, the structure of the heart, and the circulatory system (Boshuizen \& Schmidt, 1992).

## Explanations

We selected two experts' explanations and two intermediates’ explanations from a recent study by Lachner et al. (2012). We selected the explanations according to their degree of coherence, which can be regarded as the number of fragments in the explanation (see Figure 1). We chose the explanations of the two experts with the smallest number of fragments (Expert A: 1 fragment; Expert B: 3 fragments) and those of the two intermediates with the highest number of fragments (Intermediate A: 8 fragments; Intermediate B: 8 fragments).
In the Lachner et al. study, this structural feature of explanations mediated the effect of the instructors' expertise with regard to novices' transfer. The experts in that study were cardiologists who had at least 15 years of working experience. Intermediates were medical students in their fifth year of studying. The explanations were 157 words (SD $=36.03$ ) long on average. The explanations pointed out the biomedical processes and causes of bacterial endocarditis, and how the symptoms mentioned in the case description could be related to the underlying biomedical processes.

## Factual knowledge test

A factual knowledge test was used as pre- and posttest and measured novices' conceptual understanding of bacterial endocarditis. It consisted of nine multiple choice items with four answer possibilities and one correct solution (e.g. "What is the reason for the diastolic in cases of endocarditis?"). Participants received one point for each correct answer, yielding a total possible score of nine points.

## Transfer test

To measure novices' transfer, we constructed two complex questions that required novices to transfer their acquired knowledge of bacterial endocarditis to other complex medical phenomena ("Why can endocarditis result in a cardiogenic shock?", "Can endocarditis be the cause of a stroke?"). First, participants’ written answers to these questions were segmented into individual statements and then compared to reference answers constructed by a medical expert. A scorer who was blind to the participants’ treatment condition used a strict manual in which participants received 0.5 points for each unit of the reference answer. For each task, participants could obtain 4.5 points, which resulted in a maximum score of 9 points for both answers.

## Procedure

Participants were tested individually in a quiet room. They were randomly assigned to one of the four explanations. An experimental session lasted 60 minutes. During the experimental session, participants were not allowed to proceed before being signaled by the experimenter (exact time on task). First, participants answered the pretest (10 minutes). Then, in the learning phase, they received the case description and one of the randomly assigned explanations
( 25 minutes). Participants were instructed to think aloud while they studied the explanation. If participants did not think aloud for more than 5 seconds, the experimenter prompted them to continue talking. In the post-test phase, participants answered the factual knowledge test (10 minutes) and accomplished the two transfer tasks (15 minutes).

## Analyses and coding

For the analyses of novices' learning processes, their think aloud protocols were transcribed and segmented into idea units. Based on Chi (2000), each idea unit was categorized as paraphrasing, bridging inferences, self-explanation, and negative monitoring (see Table 1). Thirty percent of the protocols were co-rated by a second rater. In assigning verbalizations to categories, inter-rater agreement was very good ( $\kappa=.88$ ). Thus, only one rater coded the rest of the protocols.

Table 1: Categories to rate the think-aloud protocols.

| Category | Description |
| :--- | :--- |
| Paraphrase | Novice simply restated or <br> paraphrased a text segment from <br> the explanation. |
| Bridging inferences | Novice relates different text <br> passages of the explanation to <br> better understand relations <br> between sentences. |
| Negative | Novice connects new information <br> with prior knowledge by self- <br> explaining. Indicators are the <br> generation of examples or making <br> predictions. |
| monitoring | Novice utters his /her non- <br> comprehension |

## Results

We used an alpha level of .05 for all statistical analyses. As an effect size measure, we used partial $\eta^{2}$ qualifying values $<.06$ as small effect, values in the range between .06 and .14 as medium effect, and values > . 14 as large effect (see Cohen, 1988).

A series of ANOVAs and $\chi^{2}$ tests revealed no significant differences between the experimental conditions concerning age, $F(1,66)=1.22, p=.27$; gender, $\chi^{2}(1)=2.50, p=.11$; study programs, $\chi^{2}(1)=.59, p=.44$; prior knowledge, $F(1$, 66 ) $=1.16, p=.29$, and the number of processing activities, $F(1,66)=.84, p=.36$.

## Learning Hypotheses

Table 2 provides an overview of the means and standard deviations for the factual knowledge and the transfer test. To investigate differences in factual knowledge between novices who learned with an intermediate's explanation and novices learning with an expert's explanation, we performed
an ANCOVA with type of explanation as a fixed factor, novices' posttest scores as dependent variable and novices' prior knowledge as a covariate. There was no significant difference for type of explanation regarding novices' factual knowledge, $F(1,65)=1.90, p=.17, \eta^{2}=.03$. Thus, we could replicate the results from Lachner and Nückles (2013) that novices benefited from intermediates' and experts’ explanations to a similar extent.

With regard to the transfer hypothesis, we found that novices learning with an expert's explanation significantly outperformed novices learning with an intermediate's explanation on the transfer tasks, $F(1,65)=15.56, p=.00$, $\eta^{2}=.19$. Thus, as in the study by Lachner and Nückles, experts' explanations better supported novices in solving transfer tasks as opposed to intermediates' explanations.

Table 2: Means and standard deviations (in parentheses) for the learning outcome measures.

|  | Novices with <br> Intermediates’ <br> Explanations | Novices with <br> Experts’ <br> Explanations |
| :--- | :---: | :---: |
| Dependent variable | $3.29(1.34)$ | $3.71(1.78)$ |
| Factual knowledge | $4.18(1.73)$ | $4.76(1.30)$ |
| Transfer | $3.63(1.78)$ | $5.44(1.84)$ |

## Processing Hypotheses

Table 3 shows the mean proportions and standard deviations of novices' processing activities. With regard to our processing hypothesis, we conducted a MANCOVA with type of processing activities (paraphrase, bridging inferences, self-explanations and negative monitoring) as dependent variables, type of explanation as independent variable and novices’ prior knowledge as covariate. The MANCOVA revealed a significant effect for type of explanation, $F(3,63)=3.25, p=.03, \eta^{2}=.13$. Separate ANCOVAS showed that this effect was specifically due to the differences in the proportions of paraphrasing and selfexplanations (see Table 3).

Table 3: Mean proportions and standard deviations (in parentheses) of novices’ processing activities

|  | Explanations by |  | $F^{a}$ | $p$ | $\eta^{2}$ |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  | Intermediate | Expert |  |  |  |
| Paraphrases | . 51 (.24) | . 33 (.22) | 9.65 | . 00 | . 13 |
| Bridging inferences | . 22 (.12) | . 24 (.17) | 0.25 | . 62 | . 00 |
| Self- <br> Explanations | . 11 (.11) | . 20 (.22) | 4.79 | . 03 | . 07 |
| Negative monitoring | . 16 (.15) | . 22 (.21) | 2.17 | . 15 | . 03 |
| ${ }^{\text {a }} d f=1,65$ |  |  |  |  |  |

As expected, novices learning with intermediates' explanations used more shallow processing strategies directed at the construction of the text base (i.e. paraphrasing a text segment) as compared with novices’ learning with experts' explanations. In contrast, regarding the proportion of self-explanations, novices learning with experts' explanations used more deep-level processing strategies (i.e. self-explanations) as opposed to novices learning with intermediates' explanations.

However, there was no difference for type of explanation regarding the proportion of bridging inferences. Thus, novices used the same amount of bridging inferences to establish coherence within their text base regardless of which type of explanation they received. Additionally, we did not find any significant differences for type of explanation regarding negative monitoring, which suggests that intermediates' and experts' explanations entailed comprehension problems to a similar extent.

To test whether novices’ transfer was associated with the proportion of novices' deep processing of the explanations, we computed a Pearson's correlation. To obtain a single score for deep-processing of the explanations, we computed the proportion of deep processing learning activities (i.e. bridging inferences and self-explanations) that aimed at constructing a sufficient situation model for each participant. This was appropriate because the different values were significantly inter-correlated, $r(68)=.53, p=$ .00. Novices' deep processing activities were significantly correlated with novices' performance on transfer tasks, $r(68)=.30, p=.01$. Evidently, novices’ deep processing led novices to better transfer their knowledge to other tasks.

## Discussion

The main goal of the present study was to investigate how experts' and intermediates' explanations differently affected novices' processing and novices' learning outcomes.

For novices' performance on transfer tasks, we could replicate findings of previous studies (Boekhout et al., 2010; Hinds et al., 2001; Lachner \& Nückles, 2013) that experts’ explanations more effectively enabled novices to transfer their knowledge acquired from the explanations to other related medical phenomena. Similar to findings by Boekhout et al. and Lachner and Nückles, we did not find any significant differences between experts' and intermediates' explanations regarding novices' factual knowledge gain. Apparently, intermediates' and experts' explanations were comparably appropriate to establish a solid text base. However, it must be noted that the average factual knowledge gain was rather low (see Table 2), which can be mainly attributed to the brief text length of our instructional explanations.

With regard to novices' processing of the explanations, we can conclude that experts' explanations engaged novices in a deeper processing of the explanations as opposed to explanations generated by intermediates. Novices with experts' explanations made significantly more selfexplanations and less paraphrasing compared to novices
with intermediates’ explanations. However, in contrast to our assumptions, we did not find any differences for the proportions of bridging inferences and negative monitoring. Apparently, experts' omissions in their explanations and the lack of coherence in intermediates' explanations may have balanced each other out and therefore resulted in a trade-off in the novices' bridging inferences and negative monitoring.

Beyond that, we could show that novices' performance on transfer tasks was significantly related to novices’ deep processing. Apparently, intermediates' less coherent explanations triggered shallow processing activities that solely aimed at the construction of a solid text base. In contrast, experts’ explanations mainly triggered novices’ deep processing, which resulted in the construction of a better situation model and a better performance on the transfer test. As intermediates primarily relied on shallow processing, they probably constructed a less coherent and therefore less effective situation model that resulted in a lower performance on the transfer tasks (Kintsch, 2004). However, there was only a moderate correlation between novices' transfer and deep-processing activities. However, think-aloud protocols are less reliable to measure unconscious comprehension processes (Graesser et al., 1997). Therefore, in subsequent studies, behavioral measures should be included as a complementary measure to tap implicit processes of comprehension more directly (Holmqvist et al., 2011; Kaakinen \& Hyona, 2005).

What are the broader theoretical implications of our research? First, although research on the expert-blind spot (Hinds, 1999; Nathan \& Koedinger, 2000) suggests that experts sometimes have difficulties in taking a novice's perspective, their instructional explanations nevertheless effectively support novice students in acquiring deep and flexible knowledge due to the superior coherence of their explanations. Compared to intermediates, experts produce explanations that highlight central principles of the subject matter in a very coherent manner. This supports novices in processing the explanations deeply in order to establish coherent and flexible representations of the subject domain.

Second, we could show that the effect of coherence on novices' deep-processing and on novices' transfer performance also holds true in more naturalistic settings, such as in giving explanations. In our study, we used real instructional explanations by experts and intermediates, instead of constructing highly coherent vs. low-coherent explanations (e.g. Ainsworth \& Burcham, 2007; Gilabert et al., 2005; Linderholm et al., 2001). Despite the promising results of our study, there are also some limitations and open questions. One limitation of this experiment is the use of only one phenomenon of cardiology, namely bacterial endocarditis, which possibly restricts the generalizability of our experiment. However, bacterial endocarditis can be regarded as a classic disease, which requires fundamental knowledge about the circulatory system, the structure of the heart, and embolisms. In a similar vein, future studies should investigate whether the effect of the higher
coherence of experts' explanations on novices' processing and transfer also holds true for other subject-domains.

Overall, the present study shows that experts' explanations are an effective means to foster novices' deep processing of complex contents. Due to their highcoherence, experts' explanations prompt novices to process the explanations deeply by focusing on central principles, which results in more flexible knowledge structures and subsequently in a better transfer of knowledge to other tasks. In doing so, experts' explanations can be considered as a valuable scaffold for engaging novices in deep processing and in a meaningful construction of knowledge.

## References

Ainsworth, S., \& Burcham, S. (2007). The impact of text coherence on learning by self-explanation. Learning and Instruction, 17(3), 286-303.
Boekhout, P., van Gog, T., van de Wiel, M. W., GerardsLast, D., \& Geraets, J. (2010). Example-based learning: effects of model expertise in relation to student expertise. The British Journal of Educational Psychology, 80(4), 557-566.
Boshuizen, H. P., \& Schmidt, H. G. (1992). On the role of biomedical knowledge in clinical reasoning by experts, intermediates and novices. Cognitive Science, 16(2), 153184.

Bransford, J. D., \& Schwartz, D. L. (1999). Rethinking transfer: a simple proposal with multiple implications. Review of Research in Education, 24, 61-100.
Chi, M.T.H. (2000). Self-explaining expository texts: the dual processes of generating inferences and repairing mental models. In R. Glaser (Ed.), Advances in Instructional Psychology. Hillsdale, NJ: Lawrence Erlbaum Associates.
Chi, M. T. H., Feltovich, P. J., \& Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. Cognitive Science, 5(2), 121-152.
Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.
Graesser, A. C., Millis, K. K., \& Zwaan, R. A. (1997). Discourse comprehension. Annual Review of Psychology, 48(1), 163-189.
Gilabert, R., Martinez, G., \& Vidal-Abarca, E. (2005). Some good texts are always better: text revision to foster inferences of readers with high and low prior background knowledge. Learning and Instruction, 15(1), 45-68.
Hinds, P. J. (1999). The curse of expertise: The effects of expertise and debiasing methods on predictions of novice performance. Journal of Experimental Psychology: Applied, 5(2), 205-221.
Hinds, P. J., Patterson, M., \& Pfeffer, J. (2001). Bothered by abstraction: the effect of expertise on knowledge transfer and subsequent novice performance. The Journal of Applied Psychology, 86(6), 1232-1243.
Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., \& Van de Weijer, J. (2011). Eye tracking: A
comprehensive guide to methods and measures. OUP Oxford.
Kaakinen, J. K., \& Hyona, J. (2005). Perspective effects on expository text comprehension: evidence from thinkaloud protocols, eyetracking, and recall. Discourse Processes, 40(3), 239-257.
Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. Psychological Review, 95(2), 163-182.
Kintsch, W. (2004). The construction-integration model of text comprehension and its implications for instruction. In R. B. Ruddell \& N. J. Unrau (Eds.), Theoretical models and processes of reading. Newark, DE: International Reading Association.
Lachner, A., Gurlitt, J., \& Nückles M. (2012). A graphoriented approach to measuring expertise - detecting structural differences between experts and intermediates. In N. Miyake, D. Peebles, \& R. P. Cooper (Eds.), Proceedings of the 34th Annual Conference of the Cognitive Science Society (653-658). Sapporo, Japan: Cognitive Science Society.
Lachner, A., Gurlitt, J., \& Nückles M. (2013). Why experts' explanations foster novices' transfer - Coherence matters. Manuscript submitted for publication.
Linderholm, T., Everson, M. G., van den Broek, P., Mischinski, M., Crittenden, A., \& Samuels, J.(2001). Effects of causal text revision on more or less-skilled readers' comprehension of easy and difficult texts. Cognition \& Instruction, 18, 525-556.
Nathan, M. J., \& Koedinger, K. R. (2000). Teachers' and researchers' beliefs about the development of algebraic reasoning. Journal for Research in Mathematics Education 31,168-190.
Rikers, R. M. J. P., Schmidt, H. G., \& Boshuizen, H. P. A. (2000). Knowledge encapsulation and the intermediate effect. Contemporary Educational Psychology, 25(2), 150-166.
Rikers, R. M. J. P., Schmidt, H. G., \& Boshuizen, H. P. A. (2002). On the constraints of encapsulated knowledge: Clinical case representations by medical experts and subexperts. Cognition and Instruction, 20(1), 27-45.
Rottman, B. M., Gentner, D., \& Goldwater, M. B. (2012). Causal systems categories: Differences in novice and expert categorization of causal phenomena. Cognitive Science, 36(5), 919-932.

# Do risk-averse people lie less? A comparison of risk-taking behavior in deceptive and non-deceptive scenarios 

Tei Laine (lainet @ihpc.a-star.edu.sg)<br>Kayo Sakamoto (sakamotok@ihpc.a-star.edu.sg)<br>Institute of High Performance Computing, A*STAR, 1 Fusionopolis Way, \#16-16 Connexis North, Singapore 138632<br>Tomi Silander (silander@comp.nus.edu.sg)<br>School of Computing, National University of Singapore, 21 Lower Kent Ridge Road, Singapore 119077


#### Abstract

We studied decision making in situations in which there is a monetary incentive to take risk, and in which the risk taking option sometimes involves deception. We conducted a within participant experiment in which we compared risk taking in deception conditions to pure (non-deceptive) gambles with equivalent risks and outcomes. We confirmed the fourfold pattern of risk attitudes in both conditions. We found that participants chose fewer risky options when the risky option was associated with deception, but that those who deceived more in the deception condition also took more risks in the gamble condition. We conclude that people who tend to take risks in gambles, also take them when it involves deception, although to a lesser extent.


Keywords: Decision making; risk attitudes; deception; incentives.

## Introduction

Despite being a fundamental construct in many economic theories, individual risk attitude does not exhibit the construct stability generally associated with personality traits. Many studies (e.g. Berg, Dickhaut, \& McCabe, 2005; Holt \& Laury, 2002; Isaac \& James, 2000) have shown that inconsistency in people's risk taking depends for instance on the nature (hypothetical vs. real) and magnitude of outcomes, the task and the situation they are facing (e.g., lottery vs. auction vs. game show), and the risk elicitation method (e.g., questionnaires vs. laboratory experiments).

Just as people are willing to take risk in certain situations and not in others, there are also situations in which people are willing to deceive and others in which they are not. We studied the interplay of these two tendencies in situations where deception was risky but the decision maker could choose not to deceive and not to take risk. In these situations there was, by design, no trade-off between risk-taking and deception.

Sakamoto, Laine, and Farber (forthcoming) found that perceived detection risk is one of the factors that determines whether people decide to deceive or not, and that it is evaluated differently in gain- and loss-facing situations. In the current study we sought to find out whether people's decisions to deceive are driven by their attitudes towards dishonest behavior or towards risk taking, and whether their non-deceptive risky choices can predict their decisions to deceive. Particularly, we were interested in if people's risk taking behavior in the deceptive domain also follows the four-fold pattern of risk attitudes (Tversky \& Kahneman, 1992), namely that on average people are risk averse when facing high probability gains and low probability losses (e.g., buying insurance) and
risk seeking when facing low probability gains (e.g., buying a lottery ticket) and high probability losses.

In the extensive decision making literature, few studies have addressed the four-fold pattern directly (Harbaugh, Krause, \& Vesterlund, 2009). Those that have, have not found convincing evidence in support of the pattern, supposedly due to methodological issues related to e.g. the elicitation method, usage of complicated or hypothetical prospects, or presence of low and high probability prospects in the gain domain only.

Often the number of participants and the number of decisions per participant have also been relatively small. In order to test the four-fold pattern of risk attitudes we designed a within participant study, and collected data from substantial number of individuals. We compared risk-taking decisions in two hypothetical situations, one of which was an abstract gamble and the other a real-life decision situation presumably familiar to many of our participants, namely filing a tax return. Using several analysis and modeling techniques we found support for the four-fold pattern of risk attitudes, as well as differences in risk-seeking vs. risk-avoiding behaviors in these two conditions. Despite differences in risk taking between participants who decided to deceive and those who did not, we found consistency in individual decisions across the two conditions.

## Risk attitudes when facing gain vs. loss

An abundance of empirical evidence has shown that people weight losses and gains asymmetrically, so that a loss is generally considered worse than an equivalent gain (Kahneman \& Tversky, 1979). This gain versus loss trade-off has been termed loss aversion. The principle of loss aversion is controversial since, as Gal (2006) points out, it is used in an ad hoc manner to explain a number of phenomena involving losses and gains, such as the sunk-cost effect, the endowment effect and status-quo bias (Harinck, Dijk, Beest, \& Mersmann, 2007); while at the same time, these phenomena have been presented as evidence for the existence of loss aversion. Tversky and Kahneman (1991) notwithstanding, it remains unclear how strong the experimental evidence is for loss aversion (Bateman, Kahneman, Munro, Starmer, \& Sugden, 2005; Novemsky \& Kahneman, 2005).

In general people prefer avoiding losses to making gains, and when facing gains they exhibit risk aversion, which is considered a fundamental element in theories of human decision making under risk (Holt \& Laury, 2002). Its true nature
is not well understood, for instance how its existence depends on the size of the risky outcomes (Holt \& Laury, 2002), as laboratory experiments usually use relatively low monetary incentives. The observed pattern of human risky behavior is more complex than briefly described above, and it relates not only to the magnitudes of gains and losses but also to their probabilities.

Markowitz (1952) proposed a value function, defined over gains and losses, that underweights small gains and small losses relative to large gains and large losses. This implies risk-seeking behavior for small gains and risk-avoiding behavior for large gains, whereas the reverse is true for losses (Haisley, Mostafa, \& Loewenstein, 2008). Kahneman and Tversky (1979) explain the equivalent risk preference pattern - dubbed the four-fold pattern of risk attitudes - with a probability weighting function that overweights low probabilities and underweights moderate and high probabilities.

## Lie aversion

We are interested in whether and how the risk preference changes when risk is associated with deceptive behavior that results in a better outcome than risk-avoiding behavior. Pure lie aversion would mean that the cost of lying is derived from the act of lying only. Even if there is evidence that lie aversion exists, it is not always pure, but is relative to the circumstances or linked to the consequences of lying (especially in repeated interactions where reputation is at stake) or beliefs about the outcomes and expectations of others (Erat \& Gneezy, 2012; Gneezy, 2005; Hurkens \& Kartik, 2009; López-Pérez \& Spiegelman, in press; Lundquist, Ellingsen, Gribbe, \& Johannesson, 2009).

López-Pérez and Spiegelman (in press) devise an experiment to isolate pure lie aversion. To rule out altruistic or guilt-avoidance motivations for truth-telling, none of their treatments induced loss for the receiver, but instead involved a slight increase in the sender's payoff if she decided to lie. Even if the majority of the participants lied, the number of participants (about one third) who never did was the same in each treatment. This led the authors to conclude that pure lie aversion does exist. Lundquist, Ellingsen, and Johannesson (2009) also find evidence for lie aversion, with the effect increasing with the size of the lie (people prefer not to stretch the truth too much), and with free communication as opposed to predefined messages.

Erat and Gneezy (2012) as well find strong evidence for lie aversion, but also convincing evidence that people are willing to lie, even at their own cost, if it significantly helps the other person, and even more so if their own payoff increases without increasing the other person's costs. Gneezy (2005) also finds that people tend to lie if there is no cost associated with lying itself, and if the lying benefits themselves without hurting others.

Hurkens and Kartvik (2009) argue that people can be categorized roughly into two kinds: to those who would never lie and to those who will always lie if the benefit from lying
exceeds the benefit from telling the truth. Gibson, Tanner and Wagner's results (2012) reject this static type-based model, but they argue that significant within and among individual heterogeneity exists in lie aversion and willingness to engage in deceptive behavior, providing evidence that intrinsic preferences are non-separable form economic incentives. In summary, people are sensitive to the outcomes attainable by lying, and aversion to lying cannot be explained solely by the negative (guilt) feeling from the act of lying itself, but rather must take some account of what can be achieved (benefits) or avoided (costs) by lying.

## Experiment

We designed an experiment to study how incentives, i.e, monetary gain vs. no-gain and monetary loss vs. no-loss, and the associated risks affect people's propensity to choose a deceptive risky option, in conditions where the risky option is associated with a better outcome (if successful) than the sure option. We compared risky choices in the deception condition to the pure gamble condition in which there was no deception. As the deception condition we chose a real-life scenario of filling in an annual income tax return. The risk in these scenarios was defined as the probability that the tax return would be audited and the information found to be in error.

Mainly supportive but also mixed evidence exists about the effectiveness of audit probability and fines as a deterrent for tax evasion. Maciejovsky, Schwarzenberger, and Kirchler (2012) review several studies that found a positive effect of audit probability on tax compliance, and also studies that failed to find any support for tax fines as effective deterrents for tax evasion. The authors suggest affect as a determinant of tax behavior, but other sources have also been suggested, such as trust, fairness, and social norms (Maciejovsky et al., 2012). We are not aware of any studies that have linked deception aversion to tax behavior or that have focused primarily on willingness to deceive in taxes across conditions of variable risk and outcome size.

## Method

Participants Using Amazon Mechanical Turk (MTurk, http://www.mturk.com/) we recruited 690 participants to complete an online questionnaire in Qualtrics software (http://www.qualtrics.com). After discarding data from participants who failed the attention check question we had 672 participants ( 362 women, 308 men, 2 unknown; median age 29 years, age range: 18-73 years). All participants were native English speakers, aged 18 or above, and residing in the US. Each participant received USD 1.00 for their participation.

Material We prepared 18 questions in two conditions that asked for choices between a sure and a risky option. The difference between the conditions was that one of them used simple monetary gambles, whereas the other used real-life scenarios of filling the tax return (deception scenario). In the deception condition risky outcome was always associated
with the deceptive option, and the sure outcome with the honest option. We chose this design because of the inherent risky nature of deception: there is always a chance, however minimal, that the deception is detected, leading to an adverse outcome. In other words, examples in which risk taking is associated with a sure outcome would have been, in our opinion, so artificial - especially the cases with sure losses - that we expected them to bias our participants' decision behavior.

We chose the tax return as our deception scenario since it has the extremely valuable feature of being usable symmetrically for the gain and loss domains, depending on whether the taxpayer is facing additional taxes or a tax refund. ${ }^{1}$

Four types of scenarios were prepared in both conditions: high probability gain, high probability loss, low probability gain, and low probability loss, so that for each outcome level (gain and loss) we created both a low probability and a high probability scenario. These types were designed to test the four-fold pattern of risk attitudes, and the value of the sure option was chosen to match the expected value of the risky option. For each type we created four outcome and probability variations, as shown in Table 1. ${ }^{2}$

Table 1: The money to gain or lose, and their probabilities.

| $p$ | 0.20 | 0.05 | 0.05 | 0.10 | 0.80 | 0.95 | 0.95 | 0.90 |
| :---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| Risky option \$ | 20 | 100 | 1000 | 5000 | 20 | 100 | 1000 | 5000 |
| Sure option \$ | 4 | 5 | 50 | 500 | 16 | 95 | 950 | 4500 |

Example scenarios in the gain condition are shown in Figure 1. The wording in the corresponding loss scenarios was "have to pay", instead of "get back" (deception), and "lose" instead of "win" (gamble). The choice options were presented horizontally next to each other, and their order was randomized.

Procedure After giving their informed consent the participants were asked to make their choices in 18 deception scenarios and 18 gamble scenarios (the order of sets was selected randomly for each participant). All participants answered all 36 questions, so the deception vs. gamble manipulation was within participant. After finishing the choice questionnaire they filled in optional background information, such as age, gender, and education. The questionnaire ended with a debriefing. It took them about 20 minutes to finish the whole experiment.

[^321]
#### Abstract

Deception scenario John is doing his yearly taxes. He has to answer a question about how many dependents he has. John is aware that only children under 19 years of age qualify as dependents for tax purposes; his son recently turned 19 , and so does not qualify. If he tells the truth, indicating he has zero dependents, he will get back $\$ 950$. Alternatively he could say he has one dependent, in which case he would instead get back $\$ 1000$. He knows that such mistakes are fairly common, and that he can relatively easily find an excuse for his "clerical error" if his tax return is audited. If in the audit his answer is found to be in error, however, he would get back $\$ 0$. He also knows that $5 \%$ of such errors are caught. If you were John, would you say [] You have zero dependents and get back $\$ 950$ for sure. [] You have one dependent, in which case you have a $95 \%$ chance of getting back $\$ 1000$, but also a $5 \%$ chance of getting back $\$ 0$.

Gamble scenario Imagine that you face the following two alternatives and you must choose one of them. Which one would you choose? [] Win \$950 for sure. [] $95 \%$ chance to win $\$ 1000, \$ 0$ otherwise.


Figure 1: Example questions.

## Results

We started by plotting the relative frequencies of risk takers in gains and losses for each question separately (Figure 2). When facing losses, clearly far more participants took risks in gamble scenarios than in deception scenarios across all probabilities of losing, and even more so for the higher probabilities. In gains, the risk taking varied with the amount to be gained in both conditions, but more so for gambles in which more participants took risk when facing small rather than large gains with low probabilities, whereas not much difference was seen between scenarios with high probabilities.

First we wanted to ensure that the concepts of gain/loss and gamble/deception really explain the variance in the data. Even if the overall frequency of risk taking is only $28.4 \%$, two-way repeated-measures ANOVA (2 scenario conditions $\times 2$ outcome conditions) found significant main effects both for gamble $(M=6.51)^{3}$ vs. deception ( $M=2.82$ ), $\mathrm{F}(1,671)=666.03, \mathrm{p}<0.01$, and for gain $(\mathrm{M}=3.19)$ vs. loss $(\mathrm{M}=5.78), \mathrm{F}(1,671)=651.43, \mathrm{p}<0.01$, as well as an interaction, $\mathrm{F}(1,672)=420.11, \mathrm{p}<0.01$. Pairwise t-tests with Bonferroni adjustment showed that all pairwise differences between these four conditions (i.e., deception gain, $\mathrm{M}=1.19$ (Risk taking score $\in[0,8]$ ); deception loss, $\mathrm{M}=1.64$; gamble gain, $\mathrm{M}=2.00$; and gamble loss, $\mathrm{M}=4.50$ ) were significantly different, $\mathrm{p}<0.001$.

Risk seeking and risk aversion To see if the data supports the concepts of risk seeking and risk aversion, we applied factor analysis that uses a multidimensional item response model

[^322]

Figure 2: Relative frequencies of risk takers; the larger the icon (G/D), the larger the associated outcome in that question.
for binary data. We chose solutions with the smallest number of factors that showed a meaningful structure: a one-factor solution for deception and a two-factor solution for gambles. The Promax rotated factor loadings are shown in Table 2.

For gambles the factor loadings clearly align with the fourfold pattern of risk attitudes, if we interpret the Factor 1 as "risk seeking" and Factor 2 as "risk aversion." Even though not shown in Table 2, a two-factor solution for deception also showed a similar pattern to the gamble data. It thus seems that our participants were driven by the risk attitude in both gamble and deception conditions.

However, as shown in Table 2, just one factor was enough to explain all responses in the deception condition, whereas no such meaningful pattern was detected in the gamble data. It thus seems that in the gamble condition our participants were driven by risk attitudes, but in the deception condition these attitudes were joined by other considerations.

On another note, unlike what would be suggested by the four-fold pattern of risk attitudes, the participants tended to choose the sure option in more than half of the low probability gain questions $(\mathrm{t}(671)=18.30, \mathrm{p}<0.01)$. More specifically, their risk taking in this condition was heavily affected by the value of the gamble, so that they were much more willing to trade off a sure $\$ 4$ for unsure $\$ 20$ than a sure $\$ 500$ for unsure $\$ 5000$, so they gambled in the former case but chose the sure option in the latter. This clearly contradicts what

Table 2: Promax rotated factor loadings.

|  | Question |  | Gamble |  |  |
| :--- | :--- | ---: | ---: | ---: | ---: |
| Deception |  |  |  |  |  |
| $p$ | Outcome | $\$$ | F 1 | F 2 | F 1 |
| High | Gain | 1000 | 0.174 | $\mathbf{0 . 6 6 0}$ | $\mathbf{0 . 6 9 8}$ |
|  |  | 20 | 0.001 | $\mathbf{0 . 4 0 6}$ | $\mathbf{0 . 7 7 3}$ |
|  |  | 100 | 0.221 | $\mathbf{0 . 5 4 6}$ | $\mathbf{0 . 6 8 9}$ |
|  |  | 5000 | 0.185 | $\mathbf{0 . 6 1 4}$ | $\mathbf{0 . 6 6 4}$ |
| High | Loss | 1000 | $\mathbf{- 0 . 8 6 4}$ | -0.026 | $\mathbf{0 . 8 6 7}$ |
|  |  | 20 | $\mathbf{- 0 . 7 0 0}$ | 0.050 | $\mathbf{0 . 9 2 3}$ |
|  |  | 100 | $\mathbf{- 0 . 8 4 0}$ | 0.063 | $\mathbf{0 . 9 1 7}$ |
|  |  | 5000 | $\mathbf{- 0 . 7 9 7}$ | -0.021 | $\mathbf{0 . 7 8 2}$ |
| Low | Gain | 1000 | $\mathbf{- 0 . 3 5 2}$ | -0.130 | $\mathbf{0 . 8 1 6}$ |
|  |  | 20 | $\mathbf{- 0 . 3 9 1}$ | -0.065 | $\mathbf{0 . 9 0 0}$ |
|  |  | 100 | $\mathbf{- 0 . 4 1 7}$ | -0.152 | $\mathbf{0 . 8 3 9}$ |
|  |  | 5000 | $\mathbf{- 0 . 3 5 3}$ | 0.245 | $\mathbf{0 . 7 1 7}$ |
| Low | Loss | 1000 | -0.065 | $\mathbf{0 . 6 8 6}$ | $\mathbf{0 . 6 5 1}$ |
|  |  | 20 | -0.260 | $\mathbf{0 . 4 8 3}$ | $\mathbf{- 0 . 8 2 0}$ |
|  |  | 100 | -0.091 | $\mathbf{0 . 6 8 2}$ | $\mathbf{0 . 7 0 0}$ |
|  |  | 5000 | -0.180 | $\mathbf{0 . 6 7 0}$ | $\mathbf{0 . 6 6 0}$ |

Items in bold denote the largest factor loading for the question.
$p=$ probability
one would expect from the four-fold pattern of risk attitudes, which predicts risk-seeking in all of these cases. We hypothesize that this may be an effect of our participant population, but more rigorous analysis of the difference the amount of money makes in risky and deceptive choices is a subject for future studies.
Risk aversion or deception aversion We have already established that the participants chose less risky options in the deception condition than in the pure gambling condition. However, the interesting question is whether their gambling behavior and deception behavior are statistically related.

We initially run $\chi^{2}$-analyses to find out if our participants' risk taking in the gamble scenarios and in the corresponding deception scenarios were correlated. For all questions the responses in these two conditions were not independent; the results were very or extremely significant for all questions except high probability loss with risky outcome of $\$ 20$ (for which it was still significant).

We also conducted another factor analysis for the whole dataset, i.e. gamble and deception responses combined. For the same reason as before, we chose a four-factor model. The Promax rotated factor loadings are shown in Table 3. For each question we chose the factor on which it loaded the strongest. We see Factor 1 appearing as a "deception aversion" factor, and Factor 3 as "risk aversion." Compared to the two-factor solution for gambles, in this solution the "risk seeking" factor is split into two, Factors 2 and 4. Also some questions in deception condition, which are expected to load the strongest on "deception aversion" load strongly on "risk aversion" factor.

We then considered the determinants of deceptive behav-

Table 3: Promax rotated factor loadings. Question

| D ? | $p$ | OC | \$ | F1 | F2 | F3 | F4 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Yes | High | Gain | 1000 | 0.355 | 0.180 | -0.458 | -0.108 |
|  |  |  | 20 | 0.539 | 0.193 | -0.344 | -0.088 |
|  |  |  | 100 | 0.328 | 0.138 | -0.555 | 0.063 |
|  |  |  | 5000 | 0.324 | 0.314 | -0.497 | -0.012 |
| Yes | High | Loss | 1000 | 0.970 | -0.168 | 0.205 | -0.037 |
|  |  |  | 20 | 0.956 | -0.043 | 0.034 | 0.042 |
|  |  |  | 100 | 0.981 | -0.100 | 0.090 | 0.061 |
|  |  |  | 5000 | 0.907 | -0.223 | 0.228 | 0.048 |
| Yes | Low | Gain | 1000 | 0.750 | -0.158 | 0.027 | -0.124 |
|  |  |  | 20 | 0.867 | 0.119 | 0.029 | 0.002 |
|  |  |  | 100 | 0.806 | -0.053 | 0.076 | -0.151 |
|  |  |  | 5000 | 0.649 | -0.112 | 0.016 | -0.102 |
| Yes | Low | Loss | 1000 | 0.435 | 0.051 | -0.475 | 0.158 |
|  |  |  | 20 | 0.562 | 0.080 | -0.429 | -0.050 |
|  |  |  | 100 | 0.400 | 0.205 | -0.411 | 0.072 |
|  |  |  | 5000 | 0.528 | 0.098 | -0.322 | 0.296 |
| No | High | Gain | 1000 | -0.140 | 0.117 | -0.844 | -0.188 |
|  |  |  | 20 | -0.235 | 0.042 | -0.601 | -0.363 |
|  |  |  | 100 | -0.170 | 0.203 | -0.693 | -0.136 |
|  |  |  | 5000 | -0.083 | 0.138 | -0.758 | -0.212 |
| No | High | Loss | 1000 | 0.027 | -0.875 | -0.038 | -0.240 |
|  |  |  | 20 | 0.135 | -0.741 | -0.059 | -0.115 |
|  |  |  | 100 | 0.079 | -0.838 | -0.083 | -0.193 |
|  |  |  | 5000 | 0.180 | -0.787 | 0.016 | -0.104 |
| No | Low | Gain | 1000 | 0.041 | -0.104 | 0.066 | -0.652 |
|  |  |  | 20 | 0.043 | -0.259 | -0.042 | -0.691 |
|  |  |  | 100 | -0.011 | -0.260 | 0.099 | -0.752 |
|  |  |  | 5000 | -0.020 | -0.136 | -0.229 | -0.518 |
| No | Low | Loss | 1000 | -0.105 | -0.254 | -0.730 | 0.319 |
|  |  |  | 20 | -0.135 | -0.361 | -0.584 | -0.011 |
|  |  |  | 100 | -0.140 | -0.289 | -0.745 | 0.257 |
|  |  |  | 5000 | -0.130 | -0.360 | -0.740 | 0.253 |

D?=deception condition(Yes)/gamble condition(No); OC=Outcome. $p=$ probability
ior by correlating factor scores of Factor 2 in the gamble data (which we interpreted as "pure" risk aversion) to the factor scores of the four factors in the combined data. The correlations are $0.46(\mathrm{p}<0.01), 0.04(\mathrm{p}=$ n.s.), $0.75(\mathrm{p}<0.01)$, and -0.11 ( $\mathrm{p}<0.01$ ) for combined factors $1,2,3$, and 4 , respectively. Note that the correlation to the combined Factor 1 is relatively high even though most questions highly associated with that factor are not highly associated with Factor 2 in gambles.

The factor pattern, together with the correlation of factor scores, suggests that deception aversion does not fully explain the reluctance to choose the risky option in the deception scenarios, but rather that something profound in the risk attitudes also plays a role.

As many as 279 (out of 672) participants never chose the
risky option in the deception condition, i.e. they never deceived. We wanted to see how these participants behaved in the gamble condition. The relative risk taking frequencies of non-deceptive and deceptive participants in the gamble questions are plotted in Figure 3 for gains and losses. Compared to the participants who never deceived, the participants who chose at least one risky option in the deception condition also more often chose a risky option in all gamble conditions.


Figure 3: Relative frequencies of risk takers; the larger the icon (H/D), the larger the associated outcome in that question.

To quantify the statistical significance of this observation, we conducted a mixed 2 (never deceived vs. sometimes deceived) $\times 4$ (four-fold pattern) ANOVA that showed a significant main effect of those who never deceived ( $\mathrm{M}=5.735$ ) (Score $\in[0,16]$ ) vs. those who deceived at least once $(\mathrm{M}=7.056), \mathrm{F}(1,672)=36.83, \mathrm{p}<0.001$. Non-deceiving participants also appeared to be risk averse in gambles.

The factor structure for non-deceiving participants' gamble responses follows the structure for the rest of the participants, shown in Table 2. As the above ANOVA shows, they took significantly fewer risks overall, and more specifically they chose significantly fewer risky options in low probability gain questions ( $\mathrm{M}=1.072$ ) (Score $\in[0,4]$ ), compared to the others $(\mathrm{M}=1.310), \mathrm{F}(1,614)=7.6308(\mathrm{p}<0.01)$. There is something particular in these participants and their responses, which is also reflected in the factor solution (see Table 3): the low probability gain questions are explained by their own factor, separate from the other supposedly risk seeking questions.

Zooming down to the individual question level, a 1 -sided proportion test showed a very significant reduction in risk taking for this participant group in all high probability gain questions ( $\mathrm{p}<0.001$ ), and also a significant reduction in all low probability loss questions ( $\mathrm{p}<0.01$ ). High probability loss and low probability gain groups both contained questions for which the test results were not significant. None of the questions in these groups showed statistically very significant ( $\mathrm{p}<0.01$ ) results.

In summary, honest participants chose less risky options in the gamble conditions in which one would expect risk aversion. Compared to the rest of the participants, the honest participants were more risk averse overall, and especially so in conditions in which one would expect risk avoidance.

## Discussion

Our central finding is that the participants who always chose the honest option in deception scenarios also displayed a distinctive pattern of behavior in non-deceptive gamble scenarios. They took fewer risks than normal in conditions that normally elicit risky responses, and were also more risk averse than other participants in conditions that normally elicit risk aversion. From this we conclude that it may not be pure lie aversion that determines the likelihood of risky deceptive behavior even in seemingly perfectly lie-averse people, but rather that risk attitude also plays a role. In other words, these "honest" participants may still be driven in part by an unusually strong aversion to risk, rather than purely by aversion to deception itself. Our results also support a (less surprising) generalization about risk-takers, namely that people who tend to take risks under normal conditions also tend to take them in deception context. Together, these results can be summed up as indicating that there is some within participant consistency in risk taking across conditions that do or do not involve deception. In other words, deception aversion cannot be regarded as a "pure" factor, and does not totally overrun risk seeking in the deception domain. However, the nature of our experimental design does not allow a fully conclusive distinction between these two forces. Future studies are planned to address this issue.

## Acknowledgments

We are grateful to Ilya Farber for the feedback and suggestions related to this work. We would also like to thank Wan Kum Seong for creating the Qualtrics survey for us, and managing the data collection.

## References

Bateman, I., Kahneman, D., Munro, A., Starmer, C., \& Sugden, R. (2005). Testing competing models of loss aversion: An adversarial collaboration. Journal of Public Ecnomics, 89, 1561-1580.
Berg, J., Dickhaut, J., \& McCabe, K. (2005). Risk preference instability across institutions: A dilemma. Proceedings of the National Academy of Sciences, 102(11), 4209-4214.

Erat, S., \& Gneezy, U. (2012). White lies. Management Science, 58, 723-733.
Gal, D. (2006). A psychological law of inertia and the illusion of loss aversion. Judgment and Decision Making, l(1), 2332.

Gibson, R., Tanner, C., \& Wagner, A. F. (2012, April). Preferences for truthfulness: Heterogeneity among and within individuals. Switzerland. (Working Paper Series: National Centre of Competence in Research)
Gneezy, U. (2005). Deception: The role of consequences. The American Economic Review, 95(1), 384-394.
Haisley, E., Mostafa, R., \& Loewenstein, G. (2008). Myopic risk-seeking: The impact of narrow decision bracketing on lottery play. Journal of Risk and Uncertainty, 37, 57-75.
Harbaugh, W. T., Krause, K., \& Vesterlund, L. (2009). The fourfold pattern of risk attitudes in choice and pricing task. The Economic Journal, 120(545), 595611.
Harinck, F., Dijk, E. V., Beest, I. V., \& Mersmann, P. (2007). When gains loom larger than losses. Psychological Science, 18(12), 1099-1105.
Holt, C. A., \& Laury, S. K. (2002). Risk aversion and incentive effects. American Economic Review, 92(5), 1644-55.
Hurkens, S., \& Kartik, N. (2009). Would I lie to you? On social preferences and lying aversion. Experimental Economics, 12, 180-192.
Isaac, R. M., \& James, D. (2000). Just who are you calling risk averse? Journal of Risk and Uncertainty, 20(2), 177187.

Kahneman, D., \& Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, 47, 263291.

López-Pérez, R., \& Spiegelman, E. (in press). Why do people tell the truth? Experimental evidence for pure lie aversion. Experimental Economics.
Lundquist, T., Ellingsen, T., Gribbe, E., \& Johannesson, M. (2009). The aversion to lying. Journal of Economic Behavior \& Organization, 70, 81-92.
Maciejovsky, B., Schwarzenberger, H., \& Kirchler, E. (2012). Rationality versus emotion: The case of tax ethics and compliance. Journal of Business Ethics, 109, 339-350.
Markowitz, H. (1952). The utility of wealth. The Journal of Political Economy, 60(2), 151-158.
Novemsky, N., \& Kahneman, D. (2005). The boundaries of loss aversion. Journal of Marketing Research, 42, 119128.

Sakamoto, K., Laine, T., \& Farber, I. (forthcoming). Deciding whether to deceive: Determinants of the choice between deceptive and honest communication. Journal of Economic Behavior \& Organization.
Tversky, A., \& Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. The Quarterly Journal of Economics, 106(4), 1039-1061.
Tversky, A., \& Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and Uncertainty, 5(4), 297-323.

# Maximizing learning from collaborative activities 

Rachel J. Lam (rachel.lam@asu.edu)<br>Arizona State University<br>Teachers College, 300 E. Lemon St.<br>Tempe, AZ 85281


#### Abstract

Utilizing a Preparation for Future Learning paradigm and the Interactive-Constructive-Active-Passive framework, this study examined how two different kinds of cognitively engaging activities prepared students to learn from collaborating. Findings show that preparing prior to collaborating improved learning, but a difference was not detected in the type of preparation. In addition, differences in learning outcomes were only present in measures of deep knowledge. Analyses used a multilevel method targeted to dyadic data. Discussion addresses designing collaborative classroom activities that are effective and efficient for deep learning, as well as the importance of aligning assessments to depth of learning.


Keywords: collaborative learning; preparation for future learning; cognitive engagement; classroom learning.

## Introduction

Collaborative learning has become a common instructional strategy in a variety of educational settings because of its potential to boost student learning. Through peer discussion, students can receive immediate feedback, ask questions, generate explanations, challenge each other, jointly construct understanding, and elaborate on each other's ideas, which are all behaviors that have been shown to improve learning outcomes in both the classroom and laboratory. However, despite the extensive research that has been conducted on collaborative learning, the literature is still unclear as to what factors lead to the best learning outcomes, in particular, for deep understanding of concepts. Thus, this work aimed to investigate two factors that may improve deep knowledge, in particular, in a conceptual (as opposed to a problem-solving) domain: (a) individually engaging in the learning material prior to collaborating and (b) "constructively" engaging, where students are generating (constructing) new knowledge beyond the learning material.

There are mixed results as to how collaboration affects student learning (Barron, 2003; Craig, Chi, \& VanLehn, 2009). In general, students do not always take advantage of the benefits collaboration affords, thus, researchers have searched for ways to help students collaborate more effectively. Methods such as training students in collaboration skills (Hausmann, 2006; Uesaka \& Manalo, 2011), providing structured guidance to students while interacting (Coleman, 1998; Walker, Rummel, \& Koedinger, 2011), and designing collaborative learning environments that elicit meaningful discussion (Engle \& Conant, 2002; Kapur \& Bielaczyc, 2012) have been found to improve learning from collaborating. However, there are also challenges and limitations to these methods.

One limitation to training students in specific skills before collaborating is that they often fail to retain those skills after time (Webb, Nemer, \& Ing, 2006). The challenge of structured guidance during collaboration is that too much can constrain creativity and flexible discussion, which can hinder learning (Cohen, 1994). Therefore, one question that remains is, does the effort and time that it takes to train or guide students in collaborative behaviors really pay off? Work that has investigated the design of collaborative activities to naturally elicit effective dialoguing addresses this challenge, showing that open-ended and flexible tasks can enrich discussion (Janssen, Erkens, Kirshner, \& Kanselaar, 2010; Van Boxtel, Van der Linden, \& Kanselaar, 2000). However, this only occurs when students have sufficient prior knowledge (Nokes-Malach, Meade, \& Morrow, 2012). Thus, a collaborative learning method that avoids the time and effort needed to train students in particular skills or structure their instance-by-instance dialogic behaviors, while providing the opportunities for students to acquire adequate prior knowledge is investigated in the current study.

## Cognitive theoretical models

Two cognitive theoretical models supported the design of the collaborative activities in this study. The Interactive-Constructive-Active-Passive (ICAP) framework and the Preparation for Future Learning (PFL) paradigm are described below.

## The ICAP framework

The ICAP framework differentiates student engagement in learning tasks by categorizing students' overt behaviors as Interactive, Constructive, Active, or Passive, and is founded on theoretical assumptions about how those behaviors link to different cognitive processes (Chi, 2009; Menekse, Stump, Krause, \& Chi, 2012). An Interactive behavior might be debating or extending a partner's idea and the cognitive process underlying Interactive engagement would be co-creating knowledge. Inventing a rule, self-explaining, or creating a concept map would be Constructive, the underlying cognitive process being creating new knowledge. Active behaviors include highlighting a textbook chapter or copying solutions steps from the board, and correspond to assimilating knowledge. Listening or watching would be considered Passive, corresponding to the process of storing knowledge. The ICAP hypothesis makes the prediction that Interactive activities will produce better learning outcomes than Constructive activities, which are better than Active
activities, which are all better than Passive activities: $\mathrm{I}>\mathrm{C}>\mathrm{A}>\mathrm{P}$. There is empirical support for the ICAP hypothesis, although the Interactive category carries several caveats (Menekse et al., 2012). One is that engagement should only be considered Interactive when both individuals in a dialogue are engaging constructively. This does not always occur (literature on the process of collaboration in learning settings attests to this claim). Thus, this current study will address the question of how learning is affected by interacting on a Constructively designed task or an Actively designed task.

## The PFL paradigm

This paradigm takes into account how earlier learning experiences can shape future learning, under the perspective that prior learning can activate a mental model to either facilitate or hinder the learning of a new concept (Schwartz, Sears, \& Chang, 2007). Although the PFL paradigm was introduced in the literature over two decades ago (Schwartz \& Bransford, 1998), more recent work has used this model to investigate learning outcomes in a variety of domains (Chin et al., 2010, in elementary school science; Gadgil \& Nokes-Malach, 2012, in cognitive psychology; Schwartz, Chase, Oppezzo, \& Chin, in press, in physics). This work has shown that invention-type tasks better prepare students to learn from a lecture (Schwartz \& Martin, 2004). In other words, tasks that are set up to cognitively engage students in a "constructive" way, by causing students to generate new knowledge (Chi, 2009), are those that best prepare students to learn in a future task. The majority of the work that has investigated the PFL paradigm uses some form of didactic instruction (i.e. lecture) as the future task, thus, little is known about the effects other forms of instruction as future tasks, such as collaboration. The current study utilizes the PFL model to structure collaborative learning activities for students, however, the future activity is peer discussion (instead of a lecture) and students individually (rather than collaboratively) engage in the preparation task.

## Measures of learning and mental models

In light of using the two aforementioned cognitive perspectives as the basis for this study, the measures of learning outcomes should be viewed as representing student mental models of the concepts being tested. Mental models can be assessed through externalizations such as selfgenerated concept maps, matrices, drawings, and freewriting (Janssen et al., 2010; Schwartz, 1995; Van Amelsvoort, Andriessen, \& Kanselaar, 2007). Multiplechoice or $\mathrm{T} / \mathrm{F}$ tests are often used to measure student learning with regard to accuracy or correctness of knowledge, however, these are not necessarily appropriate to fully assess a mental model (Bransford \& Schwartz, 1999; Schwartz et al., 2007). A more complete picture of student knowledge can be captured by combining these types of assessments. With respect to measuring depth of knowledge, shallow knowledge can be equated to the "surface features" of a mental model, while deep knowledge
lies in the "structure" of the model (Chi \& VanLehn, 2012). Surface features can be facets such as labels and definitions, physical characteristics, or other plain facts. Structural knowledge is much more complex, representing the relationships between the features of a concept and/or the process by which a concept occurs or functions. Thus, the current work used student-generated written responses to assess deep, structural-based learning, while T/F pre- and posttests were used to assess shallow, surface-feature learning.

## Method

The study used a $2 \times 2$ experimental design examining Preparation (No Prep and Prep) and Type of Task (Active and Constructive). The two dependent variables were shallow learning and deep learning. In order to preserve both internal validity and ecological validity, the study was conducted as a classroom study across four introductory psychology classes with equal representation of the four conditions in each classroom. The students participated in the study as a part of their "regular" classroom activity for the weekly topic of "concepts of memory."

## Participants

Ninety students from four Psych 101 courses at a large community college in a Southwestern city in the United States participated in this study. The mean age of students was 21 years and the sample represented an ethnically diverse population (46\% Hispanic, 37\% Caucasian, 10\% African American, and 7\% Asian, Native American, or Middle Eastern). Fifty six percent of the students were female, $44 \%$ were male.

## Materials

Regarding the topic of interest, prior research attests to the difficulty that students have in deeply understanding the differences between a variety of concepts of memory, in particular, for encoding- and schema-based concepts (Schwartz \& Bransford, 1998). Thus, all learning activity materials and assessments were based on Schwartz and Bransford's (1998) materials. These materials were the only form of instruction to students for the topic. Students received no other instructional material (lecture, textbook readings, etc.) prior to the study and, therefore, were assumed to have limited prior knowledge of the concepts.

The study used the following materials: (1) pretest and demographic survey, (2) four versions of learning materials based on condition, (3) posttest, and (4) scoring rubrics.
(1) The pretest consisted of T/F questions that were very slightly modified from Schwartz and Bransford's (1998) verification measure, which was used in several studies on concepts of memory.
(2) The materials used during the learning phase were equivalent in domain content, however, the specific task instructions varied according to the ICAP cognitive engagement definitions and whether or not the condition
included a preparation period. In Prep conditions, students were given a portion of the class time to individually work on the task prior to engaging with a partner, while students in the No Prep conditions worked with a partner for the entirety of the learning phase. Active tasks asked students to work within the existing learning materials (i.e. they did not have to generate inferences beyond the materials to complete the tasks), while the Constructive tasks required students to invent concepts. To provide an example, the Constructive task required students to answer questions such as, "Why do people remember certain kinds of information, but not other kinds?" after studying a memory experiment and its results. They had to generate ideas about the process of memory. The Active version of the task, on the other hand, instructed students to study a list of memory terms and their descriptions. They then applied the terms to the same memory experiment included in the Constructive version by writing the term next to the appropriate result of the experiment. These students had to "search and select," but did not necessarily have to generate any new knowledge. Since the Active tasks took much less time to complete (as shown in a prior pilot study of this work), they included a secondary memory experiment task that was identical in structure to the first, but with a different cover story. This was to control for time-on-task, which was equalized across the four conditions.
(3) The posttest included the same $T / F$ questions that were used in the pretest. To avoid a "testing effect" (i.e. learning solely attributed to the recognition of identical test questions at a later testing phase), the ordering of the questions was changed and there were four to five days inbetween the tests. (See work by Bjork and Storm, 2011, for details regarding the conditions under which testing influences learning.) Student gain scores from pre- to posttest served as the measure of shallow learning.

Two additional tasks were included on the posttest to obtain a measure of deep learning. These were "prediction" tasks, where students had to study novel experiments on memory (i.e. they did not appear in the learning materials) and synthesize their recently learned knowledge in order to apply it to new experimental conditions, generate new inferences about how memory works, predict the results of the experiments, and provide evidence of their reasoning for predictions. Students freely wrote their responses to a set of sub-questions that all corresponded to a basic question of, "Based on what you now know about memory, how do you think the results of these experiments will turn out?"

Because these types of prediction tasks are likely deeply cognitively engaging, there was concern that including any on the pretest might influence students to engage differently in the learning activity tasks. In particular, the Active conditions may have become contaminated if students were primed in a pretest task to think more deeply about the concepts. Thus, the pretest only included the shallow T/F questions. Although this prevented obtaining any measure of deep knowledge prior to the learning phase, this was of less concern since it was highly unlikely that students had
prior deep knowledge of memory concepts. As already mentioned, they not did have previous instruction on the topic in their classes and in addition, they produced low shallow knowledge scores at pretest ( $M=50.8 \%, S D=21.6$ ). Thus, rather than a gain score, the deep learning measure used only the posttest prediction task scores.
(4) Scoring rubrics were developed in order to quantify students' responses to these prediction tasks. Responses were coded by how well they represented any of the following eight concepts: elaboration, schema, gist, serial position effect, generation effect, obstacle recall, interference, and encoding failure. These concepts may have been explicitly learned in the Active conditions, through the "search and select" tasks, or may have been implicitly learned in the Constructive conditions, through the "invention of concepts" tasks. A code of "other" was used for responses that represented novel ideas about memory (i.e. ideas that were not taught through the activities). This coding translated to a score ranging from 0-3 points, based on a holistic-style rubric. A higher score indicated knowledge of a broader range of concepts, representing a more complete mental model of memory. A score was also given for the quality of students' reasoning supporting the relationship between their predictions and the concepts, also ranging from $0-3$ points. This score indicated knowledge of the relationships between the concepts and their applications to novel settings, thus, representing a better structured mental model. A total score of 0-6 was possible. Two raters scored a randomly selected $20 \%$ of the data and intraclass correlation was used to assess inter-rater reliability, $\operatorname{ICC}(2,1)=.76, \mathrm{p}<.001$. One rater scored the remaining tests.

## Procedure

The study took place over the course of a week. On the first day, students took the pretest and filled out the demographic survey. Students were given 15 minutes to complete the pretest.

Students completed the learning activity phase during the next class. They were randomly assigned to one of the four conditions: (a) No Prep-Active, (b) No Prep-Constructive, (c) Prep-Active, and (d) Prep-Constructive. For No Prep conditions, students were randomly assigned to a partner and told to follow the instructions on their packets. They were encouraged to share ideas, try come to agreement before writing down an answer, and not to worry about writing right or wrong answers. Instructions varied depending on whether students were completing the Active or Constructive version (described in the Materials section), but all students were told to try to contribute equally to the discussion. For Prep conditions, students first completed an individual packet. They were told not to worry about right or wrong answers and to do their best. They were informed that they would use this packet to work with a partner. After the individual work (ranging from 15-20 minutes), students were randomly paired and spent the remaining class period doing their collaborative packet (10-15 minutes). They were told to share their ideas, try to contribute equally to
discussion, and come to agreement before writing down their answers. At the end of class, all materials were collected (each pair turned in a jointly completed collaborative packet). Students spent 30-35 minutes on the learning task in all conditions.

The posttest was given in the following class and was completed individually. Students spent 35-50 minutes on the posttest. Any students who finished before 30 minutes passed were asked to go over their answers one more time.

## Results

To avoid violation of the assumption of independence of subjects (which traditional ANOVA assumes) (Kenny, Kashy, \& Cook, 2006), a dyadic multilevel model was used for all analyses. Figure 1 illustrates the model. The analytic technique was a linear mixed model with the Restricted Maximum Likelihood (REML) method, appropriate to cope with dependency between partners within dyads.


Figure 1: Multilevel dyadic design.

## Shallow learning

Analysis of the pre- and posttests compared learning gains across conditions. "Normalized change" calculations were used to adjust learning gains by accounting for influences of pretest scores, yielding a more sensitive measure of gains (Marx \& Cummings, 2007). When post>pre, the following formula was used: post-pre/1-pre. When post<pre, a different formula was used: post-pre/pre. Although students gained in all conditions, there was no reliable difference between conditions. Table 1 summarizes these results.

Table 1: Shallow learning mean scores

| Condition | $n$ | Pretest $\%$ | Posttest $\%$ | Adj. <br> Gain |
| :--- | :--- | :--- | :--- | :--- |
| No Prep-Active | 14 | 53.6 | 72.6 | .43 |
| No Prep- <br> $\quad$ Constructive | 18 | 49.1 | 61.1 | .21 |
| Prep-Active | 15 | 46.7 | 63.3 | .28 |
| Prep- <br> $\quad$Constructive <br> Total | 19 | 53.5 | 71.9 | .40 |

Note: Due to incompletion of the T/F questions at either pre- or posttest, the total sample was reduced from 90 to 66 students.

These results are not surprising because even the "lowest" condition (No Prep-Active) constitutes an effective teaching strategy in a number of ways. Students were provided terms and definitions, the opportunity to apply those to real-world examples, and the benefit of engaging in discussion. Because these pre- and posttests were used to assess the knowledge of the surface features of memory concepts, students were expected to gain in all conditions. The differences between conditions were only hypothesized for deep learning, attesting to the sensitivity of the manipulation of the conditions. The deep learning results are below.

## Deep learning

Ninety students completed the prediction task portion of the posttest. The prediction task posttest scores were reliably different across conditions. There was a main effect of Preparation $F(1,41.1)=5.79, p<.03$, but no effect of Type of Task, nor an interaction effect. Students who prepared in the task individually in either type of task before collaborating showed evidence of deeper learning. See Figure 2.


Figure 2: Prediction task results.
This result was not expected since prior work supports the notion that "constructively" engaging activities should produce improved learning above "actively" engaging activities. As shown in Figure 2, there is virtually no difference between the Active and Constructive conditions when students individually prepared prior to collaborating. One interpretation of these results is that the inclusion of preparation prior to discussion in a collaborative activity boosts learning such that it overrides any effects of type of task. It is possible that the inclusion of an individual preparation period increases the likelihood that students will engage constructively in a dialogue, regardless of whether the task itself requires generation of new knowledge. In other words, the preparation may have spontaneously impelled students to engage constructively even in Active tasks, thus, further exploratory analyses were conducted to check this.

## Preparation facilitates constructive engagement

To assess differences between how the No Prep-Active and Prep-Active students engaged in the tasks, the collaborative activity worksheets were examined. They were scored by student effort, rather than in correctness of responses, since these were never intended to measure learning. Support for such a strategy can be found in work on dynamic assessments (Bransford \& Schwartz, 1999; Schwartz et al., 2007), which measure readiness to learn, rather than learning outcomes. Thus, as related to the PFL paradigm, these learning tasks can be viewed as readiness tasks that prepared students to engage in collaboration, with the posttest prediction task measuring learning. It is possible that students from the Prep-Active condition developed an enhanced readiness for learning in discussion, accounting for the improved performance on the prediction tasks.

Each dyad that completed at least $94 \%$ ( 15 of 16 items) of the activity worksheet received an effort score of two; those that completed at least $75 \%$ received an effort score of one; zero points were given to any dyads that completed under $69 \%$ of the activity (only four dyads). Since amount of work completed is not a thorough indication of how deeply engaged students were in the activities, the number of times students within a pair disagreed was also taken into account. The worksheets included a line for each item that asked students if they agreed on the answer, and if not, they were instructed to explain their disagreements. (Work on argumentation shows that students benefit from talking through disagreements, Asterhan \& Schwarz, 2009.) Analysis of discourse could have provided a better measure of engagement, however, that was beyond the scope of this paper. Thus, activity effort and average number of disagreements per pair were used to measure engagement.

Results showed that dyads in the Prep-Active condition produced a higher activity score ( $M=1.71$ ) compared to those in the No Prep-Active condition ( $M=1.43$ ), and had a slightly higher average number of disagreements (.55 compared to .45 , respectively). Although none of these differences were significant, put together they provide some support that preparation may influence students to engage constructively in an activity, even when the activity in and of itself does not require such engagement.

## Discussion and future work

This study tested the effects of preparation and type of task on shallow and deep learning in a collaborative activity. Students engaged in either an Actively or Constructively designed task, and either worked individually during part of the learning phase, then collaborated (Prep), or worked jointly the entire time (No Prep). (Recall that time-on-task and domain content of the learning materials were the same across all conditions.) Results showed that preparation improved deep learning outcomes, but no difference was detected for type of task. The main effect of preparation on outcomes extends the PFL paradigm, showing that peer discussion can serve as a beneficial future learning task (i.e. the future task need not be lecture). Considering the learning
opportunities that peer discussion offers as compared to didactic forms of instruction, this is an important finding towards design of classroom activities, especially with regard to deep learning. Although this study cannot inform on the comparison between collaborative learning and didactic instruction as future learning tasks of a PFL model, it supports the need for more work in this area.

The ICAP framework was not necessarily supported as an effective tool for designing learning activities since, overall, there were no differences in type of task on learning. However, the ICAP hypothesis predicts outcomes based on student engagement, not on task instructions. Thus, the exploratory analysis showed that students in the Prep-Active condition might have engaged constructively, justifying a null effect. Additionally, one might argue that because all four conditions included collaborative activities, the level of engagement for all conditions was actually Interactive. What is of interest here is that there then should have been an overall null result, however, that did not occur in the deep learning outcomes. Chi (2009) discusses the idea that working in pairs does not automatically make engagement Interactive, and that to be truly Interactive, both students must at minimum be engaging constructively. Thus, with regard to design of learning tasks, one way to better ensure that students engage Interactively in collaborative tasks is to include an individual preparation task prior to discussion. Future work is examining discourse data from a sampling of pairs from this study to further inform on how discourse processes related to learning, within in the contexts of the ICAP framework and PFL paradigm.

This study draws concern toward prior work that has not used analytic techniques that account for dependency between partners within dyads. Future work in areas of collaborative learning should utilize dyadic or multilevel models to analyze data that includes individual student outcomes (such as individually completed posttests).

Regarding the learning assessments, this study shows the usefulness of distinguishing between deep and shallow learning, and that different kinds of measures are needed to evaluate learning of varying depths. By using a mental model perspective to understand outcomes, one can see that a measure of "surface feature" knowledge would have shown no effects across conditions. An appropriate measure of deep "structural" knowledge was needed to tease apart how learning was affected by the collaborative tasks.

To conclude, it appears that one way to maximize the benefits of collaboration on deep learning is to include a preparation task, which allows students to develop a readiness for learning in future discussion. Preparation also may elicit spontaneous constructive engagement in future discussion, and it may not be necessary to otherwise design collaborative activities to specifically engage constructive behaviors. In addition, students who prepared only spent half the amount of the time collaborating. Thus, using a PFL paradigm to structure collaborative activity is also efficient, in that students can make the most effective use of their time engaging in discussion.

## Acknowledgments

This work was partly funded by a GRSP grant from the Graduate \& Professional Student Association at Arizona State University. Thanks to Julie Morrison for feedback on materials, Ruth Wylie and Kasia Muldner for their mentorship, Micki Chi for her early influence on this work, and Kathy Nakagawa for her support and encouragement.

## References

Asterhan, C.S.C. \& Schwarz, B.B. (2009). Argumentation and explanation in conceptual change: Indications from protocol analyses of peer-to-peer dialog. Cognitive Science, 33, 374-400.
Barron, B. (2003). When smart groups fail. Journal of the Learning Sciences, 12(3), 307-359.
Bjork, E.L. \& Storm, B.C. (2011). Retrieval experience as a modifier of future encoding: Another test effect. Journal of Experimental Psychology: Learning, Memory, and Cognition, 37(5), 1113-1124.
Bransford, J.D. \& Schwartz, D.L. (1999). Rethinking transfer: A simple proposal with multiple implications. Review of Research in Education, 24, 61-100.
Chi, M.T.H. (2009). Active-Constructive-Interactive: A conceptual framework of differentiating learning activities. Topics in Cognitive Science, 1(1), 73-105.
Chi, M.T.H. \& VanLehn, K. (2012). Seeing deep structure from the interactions of surface features. Educational Psychologist, 47(3), 177-188.
Chin, D.B., Dohmen, I.M., Cheng, B.H., Oppezzo, M.A., Chase, C.C., \& Schwartz, D.L. (2010). Preparing students for future learning with Teachable Agents. Educational Technology Research and Development, 58, 649-669.
Cohen, E.G. (1994). Restructuring the classroom: Conditions for productive small groups. Review of Educational Research, 64(1), 1-35.
Coleman, E.B. (1998). Using explanatory knowledge during collaborative problem solving in science. Journal of the Learning Sciences, 7(3\&4), 387-427.
Craig, S.D., Chi, M.T.H. and VanLehn, K. (2009). Improving classroom learning by collaboratively observing human tutoring videos while problem solving. Journal of Educational Psychology, 101(4), 779-789.
Engle, R.A. \& Conant, F.R. (2002). Guiding principles for fostering productive disciplinary engagement: Explaining an emergent argument in a community of learners classroom. Cognition and Instruction, 20(4), 399-483.
Gadgil, S. \& Nokes-Malach, T.J. (2012). Overcoming collaborative inhibition through error correction: A classroom experiment. Applied Cognitive Psychology, DOI: 10.1002/acp. 1843
Hausmann, R.G.M. (2006). Why do elaborative dialogs lead to effective problem solving and deep learning? Proceedings of the $28^{\text {th }}$ Annual Meeting of the Cognitive Science Society (pp. 1465-1469). Vancouver, B.C.: Sheridan Printing.

Janssen, J., Erkens, G., Kirschner, P.A., \& Kanselaar, G. (2010). Effects of representational guidance during computer-supported collaborative learning. Instructional Science, 38(1), 59-88.
Kapur, M. \& Bielaczyc, K. (2012). Designing for productive failure. Journal of the Learning Sciences, 21(1), 45-83.
Kenny, D.A., Kashy, D.A, \& Cook, W.L. (2006). Analysis of Dyadic Data. New York: Guilford.
Marx, J.D. \& Cummings, K. (2007). Normalized change. American Journal of Physics, 75(1), 87-91.
Menekse, M., Stump, G.S., Krause, S., \& Chi, M.T.H. (in press). Differentiated overt learning activities for effective instruction in an engineering classroom. Journal of Engineering Education.
Nokes-Malach, T.J., Meade, M.L., \& Morrow, D.G. (2012). The effect of expertise on collaborative problem solving. Thinking \& Reasoning, 18(1), 32-58.
Schwartz, D.L. (1995). The emergence of abstract representations in dyad problem solving. The Journal of the Learning Sciences, 4(3), 321-354.
Schwartz, D.L. \& Bransford, J.D. (1998). A time for telling. Cognition and Instruction, 16(4), 475-522.
Schwartz, D.L., Chase, C.C., Oppezzo, M.A., \& Chin, D.B. (in press). Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer. Journal of Educational Psychology.
Schwartz, D.L. \& Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. Cognition and Instruction, 22(2), 129-184.
Schwartz, D.L., Sears, D., \& Chang, J. (2007). Reconsidering prior knowledge. Carnegie Symposium on Cognition, 319-344. Mahwah, NJ: Lawrence Erlbaum.
Uesaka, Y. \& Manalo, E. (2011). The effects of peer communication with diagrams on students' math word problem solving processes and outcomes. Proceedings of the $33^{\text {rd }}$ Annual Meeting of the Cognitive Science Society (pp. 312-317). Austin, TX: Cognitive Science Society.
Van Amelsvoort, M., Andriessen, J., \& Kanselaar, G. (2007). Representational tools in computer-supported collaborative argumentation-based learning: How dyads work with constructed and inspected argumentative diagrams. Journal of the Learning Sciences, 16(4), 485521.

Van Boxtel, C., Van der Linden, J., \& Kanselaar, G. (2000). Collaborative learning tasks and the elaboration of conceptual knowledge. Learning and Instruction, 10, 311330.

Walker, E., Rummel, N., \& Koedinger, K.R., (2011). Designing automated adaptive support to improve student helping behaviors in a peer tutoring activity. ComputerSupported Collaborative Learning, 6, 279-306.
Webb, N.M., Nemer, K.M., \& Ing, M. (2006). Small-group reflections: Parallels between teacher discourse and student behavior in peer-directed groups. Journal of the Learning Sciences, 15(1), 63-119.

# Measuring the separate effects of practice and fatigue on eye movements during visual search 

Sophie N. Lanthier (snlanthier@gmail.ca)<br>Department of Psychology, 2136 West Mall,<br>Vancouver, BC, V6T 1Z4, Canada

Evan F. Risko (Evan.F.Risko@gmail.ca)
Department of Psychology, 202 Psychology Building
Memphis, TN, 38152

Daniel Smilek (dsmilek@uwaterloo.ca)<br>Department of Psychology, 200 University Ave West, Waterloo, ON, N2L 3G1, Canada

Alan Kingstone (alan.kingstone@ubc.ca)<br>Department of Psychology, 2136 West Mall, Vancouver, BC, V6T 1Z4, Canada


#### Abstract

Two experiments were conducted to examine how time-ontask (i.e., practice and fatigue) influences eye movements during visual search. In Experiment 1, we examined how practice influences eye movements during an extended visual search task. Results replicate the findings that over the course of a visual search task, performance improves and fixation duration increases. Yet changes in fixation duration did not correlate with changes in search performance. In Experiment 2, we examined how fatigue influences eye movements during an extended visual search task. To manipulate fatigue, participants either did or did not receive breaks. Those who did not receive breaks replicated the findings in Experiment 1. Critically, participants who did receive breaks showed no increase in fixation duration over the course of the visual search task. These results indicate that the increase in fixation duration with time-on-task reflects fatigue, and that this measure of fatigue can be derived independent of measures of performance improvements, such as shorter response times.


Keywords: Visual Search, Attention, Practice, Fatigue, Eye Movements.

Previous research (Horowitz, Cade, Wolfe \& Czeisler, 2003; Wolfe et al, 2007) has shown that after long passages of time, critical items are missed during visual search. Given that many studies have shown that fatigue negatively impacts the ability to allocate attention across a broad spectrum of tasks (Casagrande, Violani, Curcio \& Bertini, 1997; Dawson \& Reid, 1997; Drumer \& Dinges, 2005; Fairclough \& Graham, 1999; Lyznicki, Doege, Davis \& Williams, 1998; Marcus \& Loughlin, 1996; Williamson \& Feyer, 2000), it is reasonable to think that the fatigue which arises over time while engaging on a task contributes to search performance failures over time. Fatigue can be manifested in a number of different ways. For example, fatigue is considered to result from working (e.g., Winwood, Winefield, \& Lushington, 2006), mental stress (Baumeister, 2002), psychopathology (Berrios, 1990), boredom (Wyatt \&

Langdon, 1937), disease (Whitehead, 2009), and lack of sleep (Durmer, \& Dinges, 2005). In the present context, we refer to the fatigue that is associated with time-on-task (e.g., Neri, et al., 2002; Stern, et al., 1993; Wilkinson, 1961). Pinpointing the role of fatigue on search performance is inherently challenging, given that a) the negative effect of fatigue overlaps with the positive effect of practice on a task, and b) the same behavioural measures (e.g., response time, RT) are used to assess both the positive and negative effects of time on task. Further, several studies have found that RT measures are actually insensitive to fatigue (Baulk, Reyner \& Horne, 2001; Gillberg, Kecklund \& Akerstedt, 1996; Milosevic, 1997). Utilizing the fact that fatigue appears to influence oculomotor control (Bocca \& Denise, 2006; De Gennaro, Ferrara, Urbani \& Bertini, 2000; Galley 1989; Galley \& Galley, 1998; Hoffman, 1946; Luckiesh \& Moss, 1937; Morris \& Miller 1996; Saito, 1992; Schleicher, Galley, Briest \& Galley, 2008; Sirevaag \& Stern 2000; Stern, Boyer, \& Schroeder, 1994; Summala, Häkkänen, Mikkola, \& Sinkkonen, 1999), and could lead to an increase in fixation duration, the present paper investigated the intriguing possibility that fixation duration could provide an index of fatigue over time that is separable from the beneficial effects of practice on a task.

To achieve this aim, in two experiments the eye movements and performance of individuals were monitored as they performed a visual search task. Experiment 1 confirmed that changes in fixation duration are not correlated with the improvements in search that come from practice. Experiment 2 manipulated levels of fatigue, and showed that as fatigue levels increased, so did fixation duration, and that this could be separated and measured reliably from the positive effects that practice has on search performance. Collectively these data provide evidence that eye movements can be used as an indicator of fatigue that is
independent of other performance changes, in particular, the positive performance changes that accompany practice.

## Experiment 1

Participants performed a standard attentionally demanding visual search task, whereby search time increases with set size, and indicating that a target is absent takes longer than reporting a target's presence (Wolfe, 1998). An SR Research EyeLink 1000 desktop mount eye tracking system measured fixation duration.

## Method

Participants Twelve undergraduates (8 female) received course credit for participating.

Design A 2 (Target presence: present vs. absent) x 2 (Set size: 7 vs. 14 items) x 12 (Block: 1-12) within design was used.

Stimuli The stimulus displays were presented on a 24 -inch monitor set at a resolution of 1920 by 1200, with participants seated 80 cm from the screen. The visual search display consisted of target and distractor letters in an imaginary $6 \times 6$ grid, with cell-centres separated by 170 pixels horizontally and 128 pixels vertically. The only letters used as targets were " $E$ ", " $K$ ", " $P$ ", and " $Z$ ". Target letters never appeared as distractors in the visual search displays. All letters were presented in 36 -point Lucida Console font, measuring approximately 0.6 cm horizontally and 1.2 cm vertically, subtending 0.6 degrees of visual angle.

Procedure Each participant received 600 trials which were divided into 12 blocks of 50 trials. Eye tracking calibration was conducted before each block. Additionally, drift correction was conducted every 10 trials. Each target letter appeared in $25 \%$ of the target present displays. The displays were randomized for each participant. The visual search task consisted of the search for one of the target letters amongst either 6 or 13 distractor letters. Trials began with a 500 ms display indicating the target to be searched for on that trial, followed by a blank screen for 200 ms and then the search display. The participants were instructed to search for the target letter and press the right button on a gamepad if the target was present and the left button if the target was absent. The display remained on the screen until a response was made, at which point a blank screen was presented for 200 ms . The experiment took approximately 40 minutes.

## Results \& Discussion

Participants were encouraged to familiarize themselves with the task for the first 12 trials of the study, so those trials were removed from analysis. If the pupil was undetectable at any point during a trial (this includes blinks) then the trial was removed ( $8.4 \%$ of the trials). An error analysis was conducted on the remaining trials on which errors occurred
(4.5\%). The remaining analyses were conducted on the correct responses only. Each measure was analyzed using an Analysis of variance (ANOVA) with Target Presence (present vs. absent), Set Size (7 vs. 14 items), and Block (1 to 12 ) as within participant factors. We report all significant main effects and interactions. Because the focus of the study relates directly to time-on-task, we focus on the main effect of block and its interactions with other variables.

Performance Change in RT is plotted as a function of block (i.e., practice) in Figure 1. For RT, there was a main effect of Target Presence, $\mathrm{F}(1,11)=84.29$, MSE $=809646.70$, p .001, whereby responses were shorter on target present trials than target absent trials. There was a main effect of Set Size, $F(1,11)=190.71$, MSE $=303531.15, \mathrm{p}<.001$, with responses shorter on set size 7 trials than set size 14 trials. In addition, there was a Target Presence by Set Size interaction, $\mathrm{F}(1,11)=103.96, \mathrm{MSE}=147848.77, \mathrm{p}<.001$, indicating that the effect of target presence was greater on set size 14 trials than set size 7 trials. Critically, there was a main effect of Block, $\mathrm{F}(11,121)=5.18$, MSE $=31649.60$, p $<.001$, reflecting that responses became faster as block increased. No other main effects or interactions were significant. For errors, there was a main effect of Target Presence, F(1, 11) $=28.02$, MSE $=0.01, \mathrm{p}<.001$, whereby fewer errors were made on target absent than target present trials. No other main effects or interactions were significant.


Figure 1: Change in RT as a function of Block relative to the first Block in Experiment 1.

Eye Movements Change in fixation duration is plotted as a function of block (time-on-task) in Figure 2. Measurement of fixation duration revealed a main effect of Target Presence, $F(1,11)=40.60, M S E=1403.60, p<.001$, indicating that fixation durations were shorter on target absent than target present trials. There was a main effect of Set Size, $F(1,11)=8.40$, MSE $=350.20$, p $<.02$, with fixation durations being shorter on set size 14 trials than on set size 7 trials. Note that the average RT on trials without
targets and with 14 items was approximately 2-3 seconds, and our understanding of fatigue does not operate on this time scale. In other words, fatigue gradually increases throughout a task and, as such, is not measured over the course of a single trial. Critically, there was a main effect of Block, $F(11,110)=3.10, M S E=261.40, p<.002$, indicating that participants' average fixation duration increased as block increased.


Figure 2: Change in Fixation Duration as a function of Block relative to the first Block in Experiment 1.

Relation Between Performance and Eye Movements Finally, we calculated the change in both fixation duration and RT as a function of block for each participant. We then correlated these measures to examine the relationship between changes in RT and fixation duration over time. There was no correlation between the change in RT and the change in fixation duration as a function time on task, $\mathrm{r}(12)$ $=0.37, p=0.234$, demonstrating that the positive effect of practice on visual search is not driving the change in fixation duration ${ }^{1}$. Our working hypothesis is that this increase in fixation duration reflects an increase in fatigue a factor that increases with time-on-task, but is qualitatively distinct from those factors that benefit from time-on-task and lead to a performance improvement. Experiment 2 put this interpretation to a direct test.

## Experiment 2

As in Experiment 1, participants performed a standard attentionally demanding visual search task while their eye movements were monitored. Critically, half of the subjects received a three-minute break between each block of trials and the other half of the subjects did not. Fatigue is associated with time on task (Wilkinson, 1961; 1963; 1965).

[^323]As time-on-task increases the effect of fatigue on performance also increases. Providing subjects with a break reduces the amount of fatigue experienced throughout the task (Neri, et al., 2002; Stern, et al., 1993). If changes in fixation duration are related to fatigue, then we would not expect fixation duration to increase as a function of block in the Break group. However, in the No Break group, we would expect fixation duration to increase as a function of block as it did in Experiment 1.

## Method

Participants Thirty-two undergraduate students (14 female) received course credit for participating.

Design A 2 (Break condition: breaks vs. no breaks) x 2 (Target presence: present vs. absent) x 2 (Set size: 7 items vs. 14 items) x 12 (Block: 1-12) mixed design was used. The break condition was manipulated between participants; target presence, set size and block were manipulated within participants.

Stimuli and Procedure The stimuli and procedure were the same as those used in Experiment 1, except that each participant now received 576 trials which were divided into 12 blocks of 48 trials. The drift correction was now conducted every 12 trials. In addition, half of the participants now received breaks and half did not. Participants in the break condition received eleven 3 minute breaks, one at the end of each block of trials. Participants in the no break condition did not receive breaks throughout the task. The experiment was approximately 40 minutes in the no break condition and 75 minutes in the break condition.

## Results \& Discussion

Analysis was preceded by the same trial removal procedure used in Experiment 1 resulting in the removal of the first 12 trials and trials on which the pupil was undetectable (9.5\%) or errors occurred (3.1\%). Each measure was analyzed using a mixed ANOVA with Breaks (No Breaks vs. Breaks) as a between participant factor and Target Presence (present vs. absent), Set Size (7 vs. 14 items), and Block (1 to 12) as within participant factors.

Performance Change in RTs for the break and no break conditions are plotted as a function of block in Figure 3. For RT, there was a main effect of Target Presence, $F(1,30)=$ 219.84, MSE $=836634.27, \mathrm{p}<.001$, and Set Size, $\mathrm{F}(1,30)$ $=447.56$, MSE $=367483.96, \mathrm{p}<.001$. In addition, there was a Target Presence by Set Size interaction, F(1, 30) = 255.93, $\mathrm{MSE}=128222.02, \mathrm{p}<.001$, which indicates that search was more efficient when the target was present than when it was absent. Critically, there was a main effect of Block $\mathrm{F}(11,330)=12.94$, MSE $=78443.64, \mathrm{p}<.001$, whereby search became more efficient, that is, the time to determine whether a target was present or absent, became faster as block increased. This finding replicated the practice effect on performance observed in Experiment 1. There was
also a Block by Set Size interaction $F(11,330)=4.88$, MSE $=18439.20, \mathrm{p}<.001$, which indicates that search became more efficient over time. No other main effects or interactions were significant. For errors, there was a main effect of Target Presence, $\mathrm{F}(1,30)=43.16$, MSE $=166.03$, $\mathrm{p}<.001$, and Set Size, $\mathrm{F}(1,30)=16.14$, MSE $=29.59$, $\mathrm{p}<$ .001. In addition, there was a Target Presence by Set Size interaction, $\mathrm{F}(1,30)=14.15$, MSE $=30.30, \mathrm{p}<.001$, indicating that the effect of target presence was greater on set size 14 trials than set size 7 trials. No other main effects or interactions were significant.


## Block

Figure 3: Change in RT as a function of Block relative to the first Block in Experiment 2.

Eye Movements Change in fixation duration is plotted as a function of block (time-on-task) in Figure 4. Measurement of fixation duration revealed a main effect of Target Presence, $\mathrm{F}(1,30)=132.73$, MSE $=1739.87, \mathrm{p}<.001$, and a main effect of Break, $\mathrm{F}(1,31)=2.88$, MSE $=14847.77$, p $<0.01$, such that fixation durations were shorter in the no break condition than in the break condition. There was a Target Presence by Set Size interaction, $\mathrm{F}(1,30)=28.84$, MSE $=197.40, \mathrm{p}<.001$, and a Block by Target Presence interaction, $\mathrm{F}(11,330)=4.09$, $\mathrm{MSE}=161.66, \mathrm{p}<.001$, Most critically, there was a main effect of Block, F(11, 330) $=2.41, \mathrm{MSE}=289.61, \mathrm{p}<0.007$, whereby participants' average fixation duration increased as block increased and a Block by Break interaction, $\mathrm{F}(11,330)=1.87$, MSE $=$ 289.61, p < 0.042. No other main effects or interactions were significant.

To further assess the interaction of Block with Break, we calculated the slope relating Fixation Duration to Block. The effect of block was larger in the no break (1.14) condition than in the break condition (0.08), $t(30)=2.57, p$ < .015. The slopes were significantly different from zero in the no break condition, $t(15)=3.43, p<.005$, but not in the break condition, $t(15)=.32, p=.752$. This interaction confirms our hypothesis that the increases in fixation duration observed in Experiment 1, and in the No Break
group in Experiment 2, reflect increasing fatigue as a function of increased time-on-task ${ }^{2}$.


Figure 4: Change in Fixation Duration as a function of Block relative to the first Block in Experiment 2.

Relation Between Performance and Eye Movements in Experiment 1 and 2 Our argument that the increase in fixation duration reflects an increase in fatigue over the course of the task was based in part on the absence of a correlation between the changes in fixation duration and RT as a function of time observed in Experiment 1. Experiment 2 provides strong and convergent support for this argument by dissociating fixation duration and RT (i.e., a Block x Break interaction in fixation duration, and the lack of this interaction in RT). Finally, to increase the power of the original analysis in Experiment 1 we pooled the data from Experiment 1 and the no break condition of Experiment 2. There was a significant correlation, $\mathrm{r}(28)=0.386, p<0.04$, demonstrating that improvements in RT over time are related to decreases in fixation duration over time in individual subjects. As such, the positive effect of practice on visual search (i.e., shorter RTs) is not driving the increase in fixation duration, as individuals who improved the least or became slower over time also showed the greatest increase in fixation duration over the course of the task. This finding supports the notion that an increase in fixation duration is related to an increase in fatigue that results from an increase in time on task.

[^324]
## General Discussion

Many everyday tasks require looking for objects over extended periods of time. For example, security officers may be required to conduct searches for the larger part of the workday. Given that performing even the most basic task for an extended period of time can lead to fatigue, and ultimately performance failures (Dawson \& Reid, 1997; Drumer \& Dinges, 2005; Williamson \& Feyer, 2000), the present study provides a measurement tool for assessing this change in state even when it is co-occurring with the positive effects of practice.

There are a number of reasons that fatigue might influence fixation duration. The increase in fixation duration may reflect an oculomotor disengage deficit. Bocca and Denise (2006) demonstrated that the effect of fatigue on saccadic latency was more pronounced when the fixation remained on the screen than when the fixation was removed prior to appearance of an eye movement target (see also Versace et al., 2006). This pattern of results has typically been interpreted as reflecting a difficulty in disengaging the oculomotor system from fixation (e.g., Kingstone \& Klein, 1993). Further, there is evidence that damage to the parietal lobe leads to problems disengaging attention (Olk, Hildebrandt \& Kingstone, 2009) and fatigue is known to disrupt parietal lobe function. Thus, fatigue may influence the ability to initiate an eye movement.

Another possibility is that fatigue influences the decision to move the eyes from one location to another. In the present context, the decision to move from one location to another is dependent on the decision that the target is not at the currently attended location. The ability to make this decision may be disrupted by fatigue induced as time-ontask increases. For example, a number of studies have demonstrated a selective impairment of frontal lobe function when an individual is fatigued (e.g., Harrison and Horne, 2000; Jones \& Harrison, 2001). The frontal lobe has also been associated with goal maintenance (e.g., Miller \& Cohen, 2001). Thus, activation of the goal "find the letter X" may decrease when individuals become fatigued as time-on-task increases. This decreased goal activation may lead to lapses resulting in longer fixation durations. Lapses have long been associated with fatigue and are thought to arise because of transient disruptions in cognitive control (e.g., Lim \& Dinges, 2008).

One important goal of future research will be to tease apart which of these mechanisms, or combination of mechanisms, drive the oculomotor mechanisms (e.g., fixation duration) that are sensitive to the effects of fatigue in search, even when it overlaps with the positive effects of practice on a task. This is a finding of potentially great importance because it means that one could use fixation duration as a means to detect the presence of fatigue well before its inevitable negative effects begin to override the positive benefits of time on task. For instance, security guards at airports might be performing their examinations of luggage x-rays quickly and effectively, but measures of fixation duration could determine that the searcher is
growing fatigued and a break would be well advised. Similarly, it is clear that driving requires one to constantly be searching the visual world and yet the dangers of fatigue are no less profound. In principle measures of fixation duration could be obtained noninvasively while one is driving, and when reliable increases in fixation duration are detected, a driver could be encouraged to take a small restbreak at the next opportunity, with the potential that lives on the roadways could be saved. In sum, while the ability to sustain attention in both fatigued and non-fatigued individuals has attracted a great deal of research recently (e.g., Lim and Dinges, 2008), the relations between indices of sustained attention and spatial attention (i.e., eye movements) have received little consideration. This is surprising given the importance of vision to the performance of everyday tasks. The present work represents an initial and significant step forward in understanding how oculomotor measures of sustained attention vis-a-vis fixation duration can be used to detect the presence of observer fatigue independent of the beneficial effects that time-on-task can have on visual search performance.

## Acknowledgments

This work was supported by a Natural Sciences and Engineering Research Council (NSERC) Scholarship to SNL, a Killam Fellowship to EFR, NSERC operating grants to DS and AK. Address correspondence to Sophie N. Lanthier, Psychology Department, The University of British Columbia, 2136 West Mall, Vancouver, British Columbia, V6T 1Z4, Canada. Email: snlanthier@gmail.ca. Telephone: 604-822-0069. Fax: 604-822-6923.

## References

Baulk, S., Reyner, L. \& Horne, J. (2001). Driver sleepiness - evaluation of reaction time measurement as a secondary task. Sleep, 24(6), 695-698.
Baumeister, R. F. (2002). Ego Depletion and Self-Control Failure: An Energy Model of the Self's Executive Function. Self and Identity, 1, 129-136.
Berrios, G. E. (1990). Feelings of fatigue and psychopathology: a conceptual history. Comprehensive psychiatry, 31(2), 140-151.
Bocca, M. L. \& Denise, P. (2006). Total sleep deprivation effect on disengagement of spatial attention as assessed by saccadic eye movements. Clinical Neurophysiology, 117(4), 894-899.
Casagrande, M., Violani, C., Curcio, G. \& Bertini, M. (1997). Assessing vigilance through a brief pencil and paper letter cancellation task (LCT): effects of one night of sleep deprivation and of the time of day. Ergonomics, 40(60), 613-630.
Dawson., D. \& Reid, K. (1997). Fatigue, alcohol and performance impairment. Nature, 388(6639), 235.
De Gennaro, L., Ferrara, M., Urbani, L. \& Bertini, M. (2000). Oculomotor impairment after 1 night of total sleep deprivation: a dissociation between measures of
speed and accuracy. Clinical Neurophysiology, 111(10), 1771-1778.
Durmer, J. S. \& Dinges, D. F. (2005). Neurocognitive consequences of sleep deprivation. Seminars in Neurology, 25(1), 117-129.
Fairclough, S. H. \& Graham, R. (1999). Impairment of driving performance caused by sleep deprivation or alcohol: a comparative study. Human Factors, 41(1), 118-128
Galley, N. (1989). Saccadic eye movement velocity as an indicator of (de)activation. A review and some speculations. Journal of Psychophysiology, 3, 229-244.
Galley, N., \& Galley, L. (1998). Mental fatigue: an inquiry using saccadic and blink parameters. Journal of Psychophysiology, 12, 101.
Gillberg, M., Kecklund, G., \& Akerstedt, T. (1996). Sleepiness and performance of professional drivers in a truck simulator - comparison between day and night driving. Journal of Sleep Research, 5(1), 12-15.
Harrison, Y. \& Horne, J. A. (1996). Occurrence of ""microsleeps' during daytime sleep onset in normal subjects. Electroencephalography and Clinical Neurophysiology, 98(5), 411-416.
Hoffman, A. C. (1946). Eye movements during prolonged reading. Journal of Experimental Psychology, 36(2), 95118.

Horowitz, T. S., Cade, B. E., Wolfe, J. M. \& Czeisler, C. A. (2003). Searching night and day: A dissociation of effects of circadian phase and time awake on visual selective attention and vigilance. Psychological Science, 14(6), 549-557.
Jones, K. \& Harrison, Y. (2001). Frontal lobe function, sleep loss and fragmented sleep. Sleep Medicine Reviews, 5(6), 463-475.
Kingstone, A. \& Klein, R. M. (1993). Visual offsets facilitate saccadic latency: Does pre-disengagement of visuo-spatial attention mediate this gap effect? Journal of Experimental Psychology: Human Perception \& Performance, 19, 1251-1256.
Lim, J. \& Dinges, D. F. (2008). Sleep deprivation and vigilant attention. Annals of the New York Academy of Sciences, 1129, 305-322.
Luckiesh, M. \& Moss, F. K. The science of seeing. New York: Van Nostrand.
Lyznicki, J. M., Doege, T. C., Davis, R. M. \& Williams, M. A. (1998). Sleepiness, driving, and motor vehicle crashes. Council on Scientific Affairs, American Medical Association. The Journal of the American Medical Association, 279(23), 1908-1913.
Marcus, C. L. \& Loughlin, G. M. (1996) Effect of sleep deprivation on driving safety in housestaff. Sleep, 19(10), 763-766.
Miller, E. K. \& Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. Annual Review Neuroscience, 24, 167-202.
Milosevic, S. (1997). Drivers' fatigue studies. Ergonomics, 40(3), 381-389.

Morris, T. L. \& Miller, J. C. (1996). Electrooculographic and performance indices of fatigue during simulated flight. Biological Psychology, 42(3), 343-360.
Neri, D. F., Oyung, R. L., Colletti, L. M., Mallis, M. M., Tam, P. Y. \& Dinges D. F. (2002). Controlled breaks as a fatigue countermeasure on the flight deck. Aviation, Space and Environment Medicine, 73(7), 654 - 664.
Olk, B., Hildebrandt, H. \& Kingstone, A. (2009). Involuntary but not voluntary orienting contributes to a disengage deficit in visual neglect. Cortex, 46, 11491164.

Saito, S. (1992). Does fatigue exist in a quantitative measure of eye movements? Ergonomics, 35(5), 607-615.
Schleicher, R., Galley, N., Briest, S. \& Galley, L. (2008). Blinks and saccades as indicators of fatigue in sleepiness warnings: looking tired? Ergonomics, 51(7), 982 - 1010.
Sirevaag, E. J. \& Stern, J. A. (2000). Ocular measures of fatigue and cognitive factors. In Backs, R. W. \& Boucsein, W (Eds.), Engineering Psychophysiology, (pp. 269-288). Mahwah, NJ: Lawrence Erlbaum Associates.
Stern, J. A., Boyer, D., \& Schroeder, D. (1994). Blink rate: a possible measure of fatigue. Human Factors, 36(2), 285-297.
Stern, J.A., Boyer, D., Schroeder, D., Touchstone, M. \& Stoliarov, N. (1993) Blinks, saccades and fixation pauses during vigilance task performance (Reports DOT/FAA/AM-84). (Washington DC: Federal Aviation Administration, Office of Aviation Medicine).
Summala, H., Häkkänen, H., Mikkola, T. \& Sinkkonen, J. (1999). Task effects on fatigue in overnight driving. Ergonomics, 42(6), 798-806.
Versace, F., Cavallero, C., De Min Tona, G., Mozzato, M. \& Stegagno, L. (2006). Effects of sleep reduction on spatial attention. Biological Psychology, 71(3), 248-255.
Williamson, A. M. \& Feyer, A. M. (2002). Moderate sleep deprivation produces impairments in cognitive and motor performance equivalent to legally prescribed levels of alcohol intoxication. Occupational and Environmental Medicine, 57, 649-655.
Whitehead, L. (2009). The measurement of fatigue in chronic illness: a systematic review of unidimensional and multidimensional fatigue measures. Journal of Pain and Symptom Management, 37(1), 107.
Winwood, P. C., Winefield, A. H., \& Lushington, K. (2006). Work-related fatigue and recovery: the contribution of age, domestic responsibilities and shiftwork. Journal of Advanced Nursing, 56(4), 438-449.
Wolfe, J. M., Horowitz, T. S., Van Wert, M. J., Kenner, N. M., Place, S. S. \& Kibbi, N. (2007). Low target prevalence is a stubborn source of errors in visual search tasks. Journal of experimental psychology: General, 136(4), 623-635.
Wolfe, J.M. (1998). What Can 1,000,000 Trials Tell Us About Visual Search? Psychological Science, 9(1), 33-39.
Wyatt, S., \& Langdon, J. N. (1937). Fatigue and Boredom in Repetitive Work. Industrial Health Research Board Report. Medical Research Council, 77.

# Applying the dynamics of post-synaptic potentials to individual units in simulation of temporally extended ERP reading data 

Sarah Laszlo (slaszlo@binghamton.edu)<br>Department of Psychology and Program in Linguistics, 4400 Vestal Parkway East<br>Binghamton, NY 13903 USA<br>Blair C. Armstrong (b.armstrong@bcbl.edu)<br>Basque Center on Cognition, Brain and Language, Paseo Mikeletegi 69, $2^{\text {nd }}$ Floor<br>20009 DONOSTIA SPAIN


#### Abstract

In prior work, we have demonstrated that attention to the neural implementation of cognitive function is critical in creating models capable of simulating the physiological traces of those functions (e.g., Event-Related Potentials; ERPs). Here, we extend our Parallel Distributed Processing (PDP) model of ERP data elicited during the reading of single word forms to the simplest more temporally extended phenomenon: the ERP repetition effect. Simulations demonstrate that reproducing the dynamics of the ERP repetition effect can be accomplished by imposing the temporal envelope of post-synaptic potentials on individual units in the model.


Keywords: Parallel Distributed Processing; EventRelated Potentials; N400; Visual Word Recognition; Neural Computation

## Introduction

When PDP models were first introduced in the 1980s, part of the reason for their popularity was that they allowed the simulation of cognitive function with a computational architecture that was thematically similar to that employed by real neurons. In particular, the activation of a computational unit in a PDP model is determined by weighted summation of excitatory and inhibitory input-- similar to the manner in which the potential of a neuron is determined. However, especially in the domain of word reading, the neural metaphor introduced in the 1980s has made relatively little progress since that time. Instead of focusing on improving the neural metaphor, work has largely focused on increasing the number and sophistication of cognitive tasks that can be reproduced (e.g., Harm \& Seidenberg, 2004; Perry, Ziegler, \& Zorzi, 2007).

This situation is unfortunate for several reasons, two of which are particularly relevant to the present research. First, the incorporation of neural constraints in PDP models, in domains besides reading, has inspired significant theoretical progress. As a representative example, consider the manner in which models implementing the details of impaired dopaminergic gating in schizophrenia have been important in outlining a unified account of the
widespread cognitive impairments observed in that dysfunction (e.g., Braver, Barch, \& Cohen, 1999). As we attempt to demonstrate here, similar improvements in understanding could potentially be made in the domain of visual word recognition through models implementing relevant features of neural computation.

Second, though there is substantial disagreement between modeling groups about fundamental theoretical constructs (e.g., distributed versus local representation, importance of learned behavior, importance of computational homogeneity; see Seidenberg \& Plaut, 2006, for review), there is surprising agreement from many adherents of PDP models, dual-route models, and even Bayesian models, that improvement could be made to models of visual word recognition (and cognitive models more generally) by incorporating more neural constraint (Harm \& Seidenberg, 2004; Perry, et al., 2007; Griffiths, Chater, Kemp, Perfors, \& Tenenbaum, 2010). This agreement comes at a time when there exists a similar agreement that greater computational specificity is required in theories introduced to unify a voluminous ERP reading literature (e.g., Barber \& Kutas, 2007; Van Berkum, 2008; Laszlo \& Federmeier, 2011).

## The ERP Model

The ERP Model (Laszlo \& Plaut, 2012) improves contact between computational models of visual word recognition and the neural implementation of cognitive function in two principle ways. First, the ERP model's fundamental purpose is to simulate ERP waveforms, which are direct measurements of the activity of cortical neurons. This departs from traditional reading models, which instead focus on simulation of behavioral data. In particular, the ERP model simulates key effects on the N400 ERP component. The N400 is thought to represent the obligatory access of semantics in response to the presentation of an orthographic word form (for review, see Kutas \& Federmeier, 2011). This process has been explicitly couched in computational terms concordant with the PDP framework, such as


Figure 1: [Left] Architecture of the ERP model. INH stands for "inhibitory". [Right] Temporal dynamics of excitatory and inhibitory units.
parallelism and distributed representation (Laszlo \& Federmeier, 2011). The ERP model has demonstrated that PDP architecture can produce the critical effects on the N400 that led to its being considered the product of PDP architecture in the first place, such as a lack of sensitivity to lexicality as compared with a much larger effect of orthographic neighhorhood size (Laszlo \& Plaut, 2012).

Second, we have demonstrated that successful simulation of N 400 component effects requires implementation of an important constraining characteristic of neural computation: the separation of excitation and inhibition (Laszlo \& Plaut, 2012). In the ERP model, individual units have excitatory or inhibitory connections, never both. Further, inhibitory connections in the model are range-restricted, in that inhibitory connections are present only within a level of representation, never between, just as inhibitory neural projections are typically restricted to within a cortical area (this implementation is thematically similar to that in the TRACE model). Between-level connections in the ERP model are always excitatory. In addition to being range-restricted, inhibitory units in the ERP model are out-numbered by excitatory units: only one inhibitory unit is present at each level of representation. Finally, in the cortex, some populations of inhibitory units respond more quickly than others to input. In the model, this differential time course is simulated on the inhibitory units by means of the multi-linear "elbow" activation function, which produces unit activations that approximate the sum of "fast" and "slow" inhibitory
sub-populations. Figure 1 displays the architecture of the ERP model and the activation dynamics for excitatory and inhibitory units. Outside of the neural constraints just described, the ERP model is a typical PDP model that follows in the tradition of PDP word recognition models that have preceded it (most recently Harm \& Seidenberg, 2004). That is, its task is to associate a distributed pattern of orthographic input with a distributed pattern of semantic output, through non-linear (sigmoidal) transformation over several banks of hidden units. It accomplishes this task by acquiring connection weights over a training period of supervised learning with the back-propagation through time algorithm.

## ERP Repetition Effects

The ERP model successfully simulates important component effects elicited when participants read an unconnected list of text. This type of reading material, of course, does not resemble realistic reading material in numerous respects. Most importantly for the current research, realistic text comprehension pervasively relies on context for interpretation of individual word forms. Thus, to extend the ERP model's relevance to the processes involved in reading more realistic material, it is important to extend its sensitivity to context. The simplest type of context, and a type that produces robust modulations of the N 400 , is the immediate repetition of a word form (e.g., DOG DOG). This simple form of context requires that the processing of word, in a minimal fashion, be dependent on what has come before it, and is thus a reasonable first step in making the bridge between simulating the response to isolated items and simulating the response to items embedded in context.
Figure 2 displays canonical ERPs elicited when words (DOG), acronyms (DVD), pseudowords (GORK), and illegal strings of letters (XFQ) are repeated. Repetition effects on the N400 are characterized by a positivity in response to a $2^{\text {nd }}$ presentation, regardless of item type. The classic explanation of N400 repetition effects is that when an item is repeated in a short period of time ( $\sim 10$ seconds), its semantic features are still somewhat active from the prior presentation. Consequently, fewer-- unspecified-resources need be devoted to activating the same features a second time, resulting in a reduced N 400 . This interpretation has been essentially unchallenged since its formation (Rugg, 1985), but, as we will see, the model will suggest a subtly different account.
ERP repetition effects are prevalent enough in not only the reading literature, but also the memory and perception literatures, that their mechanics have been considered in computational models before (Huber,


Figure 2: Grand averaged ERPs elicited in response to first and second presentations of words, acronyms, pseudowords, and illegal strings, over the middle parietal electrode. The classic N400 repetition effect-reduced N400s for repeated items - is boxed. Note: negative is plotted upwards by convention.

Tian, Curran, O'Reilly, \& Woroch, 2008). This work, however, focused on early (i.e., pre-N400) repetition effects. An implemented computational account of N400 repetition effects, in contrast, is to our knowledge not present in the literature, and is a goal of the present simulations.

## Unit Fatigue, Post-Synaptic Potentials, and the Alpha Function

In the model, N400 activity is linked to mean activation in the semantic level of representation. Thus, in order to effect a simulated reduced N 400 in response to a repeated item, less activity must occur in semantics in the model when an item is repeated than when it is presented for the first time. In particular, specific units must become less active in response to an input when they have recently been active than when they have not; in other words, individual units must have the capacity to become selectively fatigued. Importantly, this fatigue must occur at the level of individual units-- not across the entire semantic level of representation-because units that have NOT recently been active must be free to activate to their maximum level (e.g., when a new item is presented instead of a repetition).

Thus, the desired dynamic for individual units in the model in the context of item repetition is one where an initial activation peak (in response to the first item in a pair) is followed by a subsequent decline in activation. Interestingly, this dynamic profile is similar to that of post-synaptic potentials (PSPs), as simulated in neural computation with the alpha function:

$$
\begin{equation*}
V=\alpha t e^{\frac{-t}{T}} \tag{1}
\end{equation*}
$$

Where $V$ is a measure of membrane potential, $\alpha$ is a scaling parameter that determines the maximum value of $V, t$ is the number of time steps since the unit became active, and $T$ is a free parameter that determines the time step at which $V$ peaks (see David, Kiebel,

Harrison, Mattout, Kilner, \& Friston, 2006). Figure 1 displays the shape of the alpha function.
Thus, in neural computation, PSPs are simulated with a function that resembles that desired for simulation of repetition effects. This is especially interesting in light of the fact that the source of the ERP signal is cortical post-synaptic potentials. Independent observations about 1) the dynamics of the function needed to implement repetition effects and 2) the source of ERPs thus converge to suggest a method for simulating ERP repetition effects: constraint of unit activation in the model with the alpha function.

As inhibitory units in the model are already constrained with the elbow function, to allow them to simulate the response of fast and slow inhibitory populations, we confine application of the alpha function to excitatory units. We aimed to determine whether imposing this profile would enable the model to simulate ERP repetition effects.

## Simulations

The architecture of the model is displayed in Figure 1, and is identical to that used in Laszlo \& Plaut (2012), with the exception that, now, excitatory unit activation is constrained by the alpha function. To understand how this is accomplished, think of the value of the alpha function at a particular time step as a scaling parameter. In simulations, the parameter $\alpha$ (see Equation 1) was set such that the permitted values of $V$ fell in $[0,1]$. Thus, when a unit activation is multiplied by $V$, that multiplication results in that unit's activation being scaled by $V$. When the alpha function is in its peak state, at $t=T, V$ is 1 , so multiplying unit activation by $V$ does not change the original unit activation. However, when the alpha function is in its fatigued state, when $t>T, V<1$, such that multiplying unit activations by $V$ reduces those activations, effecting unit fatigue.


Figure 3: Simulated ERPs elicited in response to repeated and non-repeated presentations of words, acronyms, pseudowords, and illegal strings. The dashed $y$-axis indicates stimulus onset. All units in the model data are arbitrary. In the simulated ERPs, as in the real ERPs, all item types produce reduced semantic activation when an item is repeated as compared to when it is not.

In the cortex, of course, not all neurons generate PSPs in response to all inputs. Thus, some neurons become fatigued in response to particular inputs, and some do not. In order to implement fatigue that mirrors the cortical situation, units in the model progress along the alpha function at different rates. Specifically, $t$ for purposes of calculating $V$ is not simply the total number of time steps that have elapsed since the presentation of the input. Instead, $V$ is calculated separately for each unit. In these by-unit calculations, $t$ is incremented not with every time step in the model, but only when a unit's activation on the prior time step exceeded a threshold. This threshold is a fixed parameter in the model. The result of this method for determining $t$ is that only units that respond to a particular input become fatigued. Units that do not respond to a particular input do not become activated above threshold, and therefore do not become fatigued.

## Training

Weights in the model were initialized to small, random values. The orthographic autoencoder was then trained via back-propagation through time for 20000 epochs to reproduce orthographic inputs on an identical output layer. Then, with the weights in the autoencoder and all inhibitory weights fixed, the remainder of the network was trained for 15000 epochs to associate input orthographies with output semantics. Each training pattern was presented for 16 time steps. Training items consisted of 62 words and 15 acronyms. Importantly, the entire network's activation was reset to its initial level after each item during training, meaning that each input during training was isolated from others. Thus, the model received no training on repeated items. The model's output dynamics in response to repeated items must therefore be an emergent characteristic of its architecture-- newly implemented to simulate PSPs-when extended to these novel input scenarios, not
simply the result of training it on the desired response to repetitions.

## Testing

The trained network was presented with input pairs either of the form AA (repetitions) or AB (nonrepetitions). Each item of the pair was presented for 16 time steps, with a single time step of blank input between each item of the pair. In testing, the network was not re-initialized between items in a pair (but was re-initialized between pairs). In non-repetitions, the B item was always of the same lexical type as the A item (i.e., words were followed by words, etc.).

In addition to trained items, the network was tested on repetitions and non-repetitions of pseudowords (85) and illegal strings (279)-- these comprised all possible nonwords in the model's orthography. The nonwords provide a particularly hard test for the model, since they were not presented to the model during training. When presented with nonword pairs, in order to, correctly, produce reduced activation on repetition but not nonrepetition trials, the model must produce dynamics it has never been trained on in response to items it has never been exposed to.

## ERPs

Target ERPs for simulation were drawn from the singleitem ERP corpus (for details, see Laszlo \& Federmeier, 2011). Briefly, it includes responses from participants who passively read an unconnected list including 75 each of words, pseudowords, acronyms, and illegal strings-- all of which repeated once-- while EEG was recorded. Figure 2 displays the target phenomenon for simulation: N400 amplitude is reduced on second presentation for all item types.

## Results <br> ERPs



Figure 4: Simulated ERPs elicited in response to repeated and non-repeated items in a model in which the alpha function is not applied. Simulated waveforms are essentially identical across presentations in these simulations, which is why only a single wave trace is visible in the figure: the second trace is directly beneath the first. Unlike the ERPs of the alpha function model, ERPs from this simulation do not display repetition effects for any item type.

Grand-averaged ERPs were computed over the middle parietal electrode site for each item type (words, pseudowords, acronyms, and illegal strings) on each presentation (first and second). N400 peak latency was measured from $250-450 \mathrm{~ms}$; N400 mean amplitude was then measured according to the full width at half max (FWHM) of that peak. This resulted in quantification of N400 mean amplitude over the $350-450 \mathrm{~ms}$ window. Using FWHM to determine the window of measurement allows for better consistency in measurements taken from real and simulated ERPs, as temporal units in the simulated ERPs are arbitrary (i.e., have no meaningful counterpart in milliseconds), but nevertheless have a peak and a FWHM of that peak.

The impact of repetition was assessed by analyzing the mean amplitude data for each item type using linear mixed effect regression, with item as a random factor and item type as a fixed factor. Markov Chain Monte Carlo sampling was used to generate $p$-values. These analyses replicated the standard finding: N400 mean amplitudes were reduced for all item types (all $p s<$ 0.0003 ).

## Simulations

Simulated ERPs were generated by averaging semantic activation for each time step in the model for the second item in each item pair; the time series of those averages across time steps is the simulated ERP. Figure 3 presents simulated ERPs for first and second presentations of each item type. As is evident from the Figure, simulated ERP amplitudes were reduced for each item type. Simulated N400 (sN400) peak latency was measured as simply the latency of the most positive peak in the simulated ERPs; since N400 activity is linked to mean semantic activation in the model, the peak of mean semantic activation in the model is transparently the peak of the sN400. Mean amplitude of the sN400 was then measured according to the FWHM of that peak, in analogy with measurement of the N400. Analysis identical to that described for the
human ERPs revealed a substantial sN400 amplitude reduction for all item types (all $p<0.005$ ).

To assess the degree to which the alpha function was responsible for the simulated repetition effects, we conducted a second simulation in which the only modification was the removal of the alpha function (essentially, this model was a replication of Laszlo \& Plaut, 2012). In what follows, we will refer to this simulation as the No-Alpha simulation, and the original simulation as the Alpha simulation. Figure 4 displays results of the No-Alpha simulation. As is evident in the Figure, the No-Alpha model did not exhibit a sN400 repetition effect, in contrast with both the empirical data and the Alpha simulation. Numerically, the difference between first and second presentation sN400 mean amplitude was not different than 0 to 5 degrees of decimal precision for any item type.

## Discussion

Our goal was to extend the original ERP model from being insensitive to context to being sensitive to the minimal context of whether an item has been repeated. We aimed to achieve this by improving the neural realism of the model. This improvement took the form of imposing the fatigue dynamic of PSPs on individual units in the model. The choice of this particular dynamic was motivated both by the empirical need to identify a fatiguing dynamic as well as the observation that the source of the ERP signal is cortical PSPs. Results indicated that, even when presented with a situation never encountered in training (item pairs) and items never encountered in training (pseudowords, illegal strings), a variant of the ERP model implementing unit fatigue reproduced the standard pattern observed in ERP studies: namely, that repeated orthographic items elicit reduced N400s. Importantly, reduced sN 400 s in response to repetition were not obtained in a version of the model without unit fatigue.
These results support the general conclusion that improving the neural realism of PDP models is a
strategy that can greatly extend the type of phenomena such models are able to explain. More importantly, however, this data provides a potential explicit mechanistic explanation of ERP repetition effects that subtly differs from that typically offered in the literature. As already discussed, the classic explanation of N400 repetition effects is that, when an item is first encountered, it invokes access of its associated semantics (or, in the case of nonwords, the semantics of visually similar items). Then, when the same item is repeated, there is less lexical-semantic processing required to re-activate the pre-activated semantics, resulting in a reduced N400 (see Rugg, 1985).

The source of N400 repetition effects in the model, in contrast, is not pre-activation of semantic features-- as is visible in Figure 3, network activity drops back almost to zero between items in a pair, before the onset of the simulated N400. Instead, semantic activity is reduced due to the fatigue of individual semantic units. While the traditional view of N 400 repetition effects is based on unspecified principles of cognitive resource, the simulations suggest a view based on explicit mechanistic principles of the underlying neural system.

More exploration-- both empirical and computational-- of fatigue as an explanation of repetition effects is clearly needed: for example, it has been demonstrated in the ERP literature that additional repetitions of word forms (i.e., third, fourth, or more presentations) do not further diminish the N400 response (Young \& Rugg, 2007), and it is not clear that the ERP model would exhibit this pattern. Similarly, in the present simulations words were considered a monolithic group, but it is well known that N400 repetition effects are strongly influenced by lexical factors such as word frequency (e.g., Young \& Rugg, 2007), and it is again not clear that the ERP model would respond similarly. Thus, although the current work suggests an interesting alternative explanation of N400 repetition effects, based on realistic neural mechanisms and processing dynamics, clearly there is significant additional work to be done to explore this explanation further. The explicit simulation implemented here is hoped to provide a foundation for this future work.

## Acknowledgments

The authors acknowledge M. Monk and the members of Binghamton University Modeling Meeting-especially K. Kurtz-for insightful discussion. This research was supported by the Research Foundation of the State University of New York and the Basque Center for Brain, Language, and Cognition.

## References

Barber, H.A., \& Kutas, M. (2007). Interplay between computational models and cognitive electrophysiology in visual word recognition. Brain Research Reviews, 53, 98-123.
Braver. T.S., Barch, D.M., \& Cohen, J.D. (1999). Cognition and Control in Schizophrenia: A Computational Model of Dopamine and Prefrontal Function. Biological Psychiatry, 46, 312-328.
David, O., Kiebel, S.J., Harrison, L.M., Mattout, J., Kilner, J.M., \& Friston, K.J. (2006). Dynamic Causal Modeling of Evoked Responses in EEG and MEG. Neuroimage, 30, 1255-72.
Griffiths, T.L., Chater, N., Kemp, C., Perfors, A., \& Tenenbaum, J.B. (2010). Probabilistic models of cognition: exploring representations and inductive biases. Trends in Cognitive Sciences, 14, 357-364.
Harm. M.W., \& Seidenberg, M.S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. Psychological Review, 111, 662-720.
Huber, D.E., Tian, X., Curran, T., O’Reilly, R.C., \& Woroch, B. (2008). The Dynamics of Integration and Separation: ERP, MEG, and Neural Network Studies of Immediate Repetition Effects. Journal of Experimental Psychology: Human Perception and Performance, 34, 1389-1416.
Kutas, M. and Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). Annual Review of Psychology, 62, 621-647.
Laszlo, S., \& Federmeier, K.D. (2011). The N400 as a snapshot of interactive processing: evidence from regression analyses of orthographic neighbor and lexical associate effects. Psychophysiology, 48, 176-186.
Laszlo, S., \& Plaut, D.C. (2012). A neurally plausible parallel distributed processing model of event-related potential word reading data. Brain and Language, 120, 271-281.
Perry, C., Ziegler, J.C., \& Zorzi, M. (2007). Nested incremental modeling in the development of computational theories: The CDP + model of reading aloud. Psychological Review, 114, 273-315.
Rugg, M.D. (1985). The effects of semantic priming and word repetition on event-related potentials. Psychophysiology, 22, 642-647.
Seidenberg, M.S., \& Plaut, D.C. (2006). Progress in understanding word reading: Data fitting versus theory building. In S. Andrews (Ed.), From Inkmarks to Ideas: Current Issues in Lexical Processing (pp. 25-49). Psychology Press: Hove, UK.
Van Berkum, J.J. A. (2008). Understanding Sentences in Context: What Brain Waves Can Tell Us. Current Directions in Psychological Science, 17, 376-380.
Young, M. P., \& Rugg, M.D. (2007). Word frequency and multiple repetition as determinants of the modulation of event-related potentials in a semantic classification task. Psychophysiology, 29, 664-676.

# Movement Correlation as a Nonverbal Cue in the Judgment of Affiliation during Social Interaction 

Nida Latif (n.latif@queensu.ca) ${ }^{1}$<br>Adriano V. Barbosa (adriano.vilela@gmail.com) ${ }^{3}$<br>Eric Vatikiotis-Bateson (evb@mail.ubc.ca) ${ }^{3}$<br>Monica S. Castelhano (monica.castelhano@queensu.ca) ${ }^{1}$ Kevin G. Munhall (kevin.munhall@queensu.ca) ${ }^{1,2}$<br>${ }^{1}$ Department of Psychology, Queen's University<br>Kingston, ON K7L 3N6 Canada<br>${ }^{2}$ Department of Otolaryngology, Queen's University<br>Kingston, ON K7L 3N6 Canada<br>${ }^{3}$ Department of Linguistics, University of British Columbia<br>Vancouver, BC V6T 1Z4 Canada


#### Abstract

It has been demonstrated that brief exposure to behavioral information is sufficient for making accurate social judgments. Movement coordination during social interaction, is one potential cue. Although coordination between individuals has been identified, our ability to perceive it when making judgments regarding affiliation (friends vs. strangers) is unknown. In the present studies, we investigated how correlated movement contributes to observers' accuracy when judging affiliation. Using correlation map analysis to quantify coordination, we showed that individuals familiar with each other correlated their movements more frequently. Observers were able to use coordination as a cue, but only when the information presented was restricted to movement related to speech (i.e. while only viewing faces). These results suggest that observed movement coordination is influenced by speech-related movements. We suggest that social perception is multi-faceted and cues may be prioritized differentially based on availability.


Keywords: social perception; social interaction \& conversation; movement correlation

## Introduction

Humans are constantly immersed in social interaction and conversation. It is not surprising that we have mechanisms to facilitate these interactions, both while engaging in and observing them. One mechanism is our ability to create a rich representation of social cues from very brief exposures as short as a few seconds. These short exposures or "thin slices" of behavioral and linguistic information are sufficient for making remarkably accurate judgments regarding social situations. Research has demonstrated great accuracy in judgments in a variety of domains including personality, social status and mental states (Ambady \& Rosenthal, 1993). However, little known about how specific cues contribute to our accuracy in social perception.

Understanding how we use available social cues is important to human behavior because monitoring others' intentions and actions is a prerequisite for modulating and guiding our own behavior, interactions, and relationship formation (Foulsham et al., 2010). In addition, using social cues is often impaired in many social and psychological disorders such as autism spectrum disorder and
schizophrenia leading to difficulty in successful interaction (Klin et al., 2002).

Previous research has looked at specific motion patterns that occur during interpersonal communication. Studies have demonstrated that individuals unintentionally synchronize and coordinate their movements and converge in linguistic properties during conversation (Richardson, Dale \& Shockley, 2008; Richardson \& Dale, 2005; Chartrand \& van Baaren, 2009; Pardo, 2006). This has been shown with different attributes of conversation such as facial expression, postures and accents (Capella \& Planalp, 1981; McHugo, Lanzetta, Sullivan, Masters \& Englis, 1985). Individuals even unintentionally coordinate their movements without visual information from their partner (Shockley et al., 2003). Coordination without visual information suggests that convergence in behavior can be directly influenced by vocal information exchanged during conversation. Further, studies examining social-cognitive variables in convergence have shown through subjective observation that individuals with good rapport coordinate their movements (Grahe \& Bernieri, 1999). Also, friends converge more in linguistic properties than strangers (Dunne $\& \mathrm{Ng}, 2002$ ). Coordination may occur because of inherent biological and behavioral rhythms as well as a coupling of conversation-engaged individuals' mental representation of their perceptions of each other (Richardson \& Dale, 2005; Meltzoff \& Prinz, 2002).

Although the presence of convergence in nonverbal and linguistic properties has been examined, our ability to perceive this convergence has not been investigated. In particular, the contribution of correlated movements has not been examined objectively in the perception of affiliation (i.e. whether individuals engaged in conversation are friends or strangers). The current studies use movement and coordination quantification methods to examine convergence between interacting individuals. These methods allowed us to investigate whether the amount of coordination differs as a function of affiliation and whether it contributes to the accuracy of affiliation judgments made by an external observer.

## Experiment 1: Movement Analysis

This experiment investigated if there was an observable variation in coordination between individuals as a result of known affiliation differences.

## Methods

Participants Sixty-two undergraduates (Mean Age $=21.2$, 36 females) from Queen's University were recruited in pairs to engage in video-recorded conversation. Thirty-one dyads were either recruited as friend pairs or were experimentally paired. Conversations from 15 same gender friend pairs (10 female), 12 same gender stranger pairs ( 6 female) and two mixed friend and stranger pairs were video recorded.

Stimulus Collection Stimuli were collected by video recording, unstructured conversation between two participants, using a single camera aimed to capture both individuals ${ }^{1}$. Individuals sat on fixed chairs and were left to converse without an experimenter present for approximately 10 minutes.

Stimulus Analysis An algorithm computing spatiotemporal coordination was used on the video clips. In these experiments, only the visual information was examined using this algorithm. The algorithm developed by Barbosa et al. (2012) first computes optical flow using a standard image processing technique where velocities of brightness patterns in an image are calculated within a region of pixels and summed to give a global value for a particular cluster of pixels (Horn \& Shunck, 1981). Then, a correlation analysis is used to compute instantaneous correlation between movement signals within a specified region of interest.

Using the optical flow analysis, the Barbosa et al. (2012) algorithm computes total motion in an identified region of interest by summing the optical flow in that region. Regions of interest were drawn around each individual engaged in conversation, for a gross estimate of their total body motion. Correlation Map Analysis (CMA) was then used to quantify the coordination between the two speakers' movements. A key characteristic of CMA is that it computes the correlation between a pair of signals as a function of both time and the lag between the signals. This not only allows us to characterize the correlation throughout the duration of the signals, but also to capture correlations between events that are not perfectly aligned in time. Therefore, CMA is able to capture coordination between signals, where events in the signals are related to each other but do not necessarily happen at exactly the same time; rather, they fluctuate around some specific lag between the signals as the signals evolve through time. Capturing coordination at a time lag allows for alternating behavior, such as that seen in social interaction, to be captured. This kind of mechanism is ideal for biological rhythms that are rarely synchronous and allows for convergence in social

[^325]interaction to be quantified (Winfree, 1980, Barbosa et al, 2012).

Correlation Data Analysis Average distributions of correlation were created for each of the friends and strangers groups. For statistical comparison, a resampling nonparametric technique was used to create null distributions. The correlations for friends and strangers were compared to a null distribution. The distributions were simplified to look only at the positive lags $(0-+0.5 \mathrm{~s})$. Because of the rhythmical structure of conversation, the positive and negative lags tend to be redundant and we included only the positive lags in all analyses. This resulted in distributions looking at 16 lags in total, including 0 lag (i.e. completely synchronous correlation; Frame rate $=30 \mathrm{fps} ; 0.5 \mathrm{~s}=15$ frames plus one 0 lag frame).

Motion Magnitude Analysis To control for the magnitude of motion when looking at the correlation, distributions of all motion magnitudes from the friends and strangers pairs were generated. The two distributions were compared to determine differences in total motion.

## Results

Correlation Analysis The probability distributions of each of the friend and stranger correlations at each lag were compared with a null distribution representing the correlations computed between the motions of all subjects who were not actually in a conversation together. The means for the correlations at lags closest to synchronous were significantly different $(\mathrm{p}<0.05)$ from the null distribution for both friends and strangers. These results are displayed in Figure 1 where the first half of both the friend and stranger distributions (first 8 lags/frames) displays higher mean correlations than observed for the random pairings. Correlations significantly different at time-points closest to synchronous indicates that individuals engaged in conversation are highly sensitive to their partner's movements and coordination occurs within moments of the movement first being initiated. Thus, the data indicate that engaging in a face-to-face conversation produces correlations greatly exceeding what would be produced by chance pairings of motion signals.

To determine how correlation differed based on affiliation, the friends and strangers distributions were subtracted from each other to create a difference distribution. This was compared with two null difference distributions: One looked at correlation differences between randomly paired individual's movements and randomly assigned affiliation categories and the second distribution looked at correlation differences between real pairs that actually conversed but who were arbitrarily categorized as a particular affiliation for this analysis. Both the overall difference as well as the mean difference per lag was significantly different when compared to both null distributions where friends had more correlated events than
strangers. ( $\mathrm{p}<0.05$ ) (See Figure 3). Although comparisons to both these null distributions is interesting in that they suggest that friends' and strangers' conversation contain content unique to their affiliation categorization, the comparison to the null containing real pairs is more informative. The real-pair null contains motion that can be attributed to conversational motion in general as opposed to random motion. These results support our hypothesis that affiliation results in correlational variation, and that friends correlate more frequently than strangers.


Figure 2. (Top to Bottom) Distributions of average correlation for the null, friend and strangers. Lighter colors indicate larger proportion of events occurring at correlation values plotted along the $x$-axis. Lag counts (in frames) indicate 16 temporal points between 0 and 0.5 s where average correlation was computed. Here, both friends and strangers had correlation value greater than the null but only significant in the first few frames.

Motion Magnitude Analysis A distribution of magnitudes of motion for the friends and strangers was created to determine any differences in motion present within the groups. Results indicated that friend pairs contained more motion (Mean=0.88 pixels/frame, $\mathrm{SE}=5.79 \mathrm{e}-04$ ) than stranger pairs (Mean=0.84 pixels/frame, $\mathrm{SE}=5.38 \mathrm{e}-04$ ). This analysis was carried in preparation for Experiment 2. We wanted to be able to control the amount of motion presented to observers making perceptual judgments so that judgments about accuracy were not being affected by motion differences.

Results from Experiment 1 demonstrated in a quantified manner that correlation was an inherent part of conversational movement and that friends coordinated more


Figure 3. (Top to Bottom) Three-dimensional average correlation difference distributions for friends-strangers, random-pair subtractions and real-pair random subtractions. Lighter colors indicate higher correlation differences along the x -axis with height indicating frequency of events. Lag counts (in frames) indicate 16 temporal points between 0 and 0.5 s where average correlation was computed. Greater positive peaks indicate more correlated events for friends in comparison to strangers.
than strangers. This supported our hypothesis and previous studies that indicated that familiarity and good rapport resulted in linguistic and behavioral coordination. The correlation data from this study were used to identify stimuli for a perceptual judgment task in Experiment 2.

## Experiment 2: Perceptual Judgment

The previous experiment demonstrated that affiliation led to differences in movement correlation. This experiment investigated whether observers were attuned to the correlational differences and if perception of correlation was influencing accuracy of affiliation judgments. In this study, we varied amount of correlation while controlling for the amount of motion in the perception of thin-slices of conversation

## Methods

Stimuli The analysis of motion magnitudes from Experiment 1 was used to control the amount of motion so that differences in perception of clips could be attributed to correlation differences rather than confounded by motion
differences. Clips were selected from a window that was centered at half a standard deviation around the mean of the sum of the motion distributions for friends and strangers. Clips and their corresponding correlation values were extracted if they were contained within a 5 s continuous ${ }^{2}$ clip that contained average motion from within our defined thresholds. Correlations were re-computed using the same procedure as Experiment 1 for only those clips that were controlled for motion magnitude to ensure that the friend/stranger correlation results were true for our perception stimuli.

All possible clips were sorted from lowest to highest average correlation. Six of the lowest and six of the highest correlated clips were selected for each of the friends and strangers groups ( $\mathrm{n}=24 \mathrm{clips}$ ) such that each conversing pair was only presented once. The final clips contained ten same gender friend pairs, ten same gender stranger pairs and two mixed-gender pairs for each group.

Procedure Twenty undergraduates (Mean age $=20.8,16$ females) from Queen's University participated for monetary compensation. A within-subjects design was used where all participants viewed both high and low correlated friend and stranger clips. The 24 five-second clips were presented and participants were asked to perform a social judgment rating using a Likert scale. On a scale of 1-7, participants indicated whether the two individuals engaged in conversation had just met (1) or were friends (7). Following the experiment, each participant was asked to record the kind of information they used to make their judgments.

## Results

Correlation Analysis Average correlations and correlation difference patterns seen in Experiment 1 were also observed in the 24 stimulus clips, confirming that the restricted magnitude of motion was not influencing correlations.

Social Perception Accuracy The accuracy of affiliation perception was determined by computing the average score for all videos presented as a function of their correlation and affiliation. Figure 3 displays the perceptual rating results. A factorial ANOVA was performed and results indicated that although participants could accurately discriminate between friends and strangers $(\mathrm{F}(1,20)=4.28, \mathrm{p}=0.05)$, correlation did not seem to affect perceptual judgments. In addition, scores were analyzed using an extreme groups analysis (Preacher, Rucker, MacCallum \& Nicewander, 2005) where neutral responses were eliminated and a factorial ANOVA was

[^326]

Figure 3. Average score for affiliation rating. A greater score represents a preference towards a judgment of 'friends' and a lower score represents preference towards a 'strangers' rating.
performed. Results demonstrated the same effect as the simple analysis.

The perceptual judgment results in this experiment demonstrated that observers could clearly make affiliation judgments. However, our results showed no evidence that degree of correlation between pairs influenced perceptual judgments. Subjective responses of reported cues used by participants indicated that subtle movements such as those of the hands and the mouth were given precedence. A high correlation between speech and face/head motion related to speech has been demonstrated in previous studies (Barbosa et al, 2008); perhaps these smaller gestures of speech were not contributing to our correlation measures as much as larger body motions. Experiment 3 was conducted to investigate whether the correlational structure of smaller speech related movements, such as face/head motion, might better account for the perceptual data.

## Experiment 3: Selected Perceptual Judgment

In this experiment, we looked at whether an observer's ability to make perceptual judgments of affiliation altered when the social information presented was restricted to more subtle, speech related correlations within the face/head region.

## Methods

Stimuli All video recordings from Experiment 1 were cropped at participants' shoulders to include only head and facial movement. Correlation and motion analysis was performed in the same manner as Experiment 1. Motion thresholds and clip selection criteria were created using the same procedure as Experiment 2. Twenty-four clips were selected to include 12 same gender friends and 12 same gender strangers. ${ }^{3}$ These included six of the highest and six of the lowest correlated clips in each category.

[^327]Procedure Twenty undergraduates from Queen's University (Mean Age=20.1, 16 females) participated in this experiment. The perceptual judgment task used the same procedure as that used in Experiment 2.

## Results

Correlation Analysis Overall, average correlations and correlation differences reflected the same pattern as Experiments 1 and 2 confirming that movement correlation was influenced by affiliation, independent of amount of motion.

Social Perception Accuracy Analysis of affiliation perception accuracy was performed as in Experiment 2. Average score for all videos as a function of correlation and affiliation was computed and a factorial ANOVA was performed. Results showed that there was a significant effect of correlation indicating observers provided a higher proportion of 'friends' responses for highly correlated clips $(\mathrm{F}(1,20)=7.78, \mathrm{p}=0.01)$. There was no significant effect for affiliation indicating that scores were not dependent on the true affiliation between conversing individuals. Results are presented in Figure 4. Observers were more likely to misperceive high-correlation pairs as friends and lowcorrelation individuals as strangers. These results demonstrated that movement correlation was a significantly influencing cue when perceiving subtle, speech-related movement.


Figure 5. Average score for affiliation rating when viewing only faces. A greater score represents a preference towards a judgment of 'friends' and a lower score represents preference towards a 'strangers' rating.

## General Discussion

In these studies, we were interested in examining how affiliation between two individuals resulted in variation in movement and coordination. Additionally, we were interested in determining whether this variation influenced the accuracy of observers making rapid judgments regarding that affiliation. Based on previous studies on convergence in linguistic and behavioral properties, we predicted that familiarity would result in greater coordination which would influence judgments of affiliation by external observers.

The results of these studies demonstrated that there was indeed variation in motion and coordination resulting from affiliation. In general, movement correlation was present during social interaction, regardless of affiliation, although higher correlations were present for friend pairs. This was supported by studies that suggest that we interactively align our representations of conversation content (Garrod \& Pickering, 2004). But, is this correlation used as information when making social judgments?

The observation that participants were not sensitive to the correlation differences presented in Experiment 2 suggests that other cues in the full body stimuli such as static postural cues as well as motion cues might have influenced the way participants attributed affiliation. The correlational structure of the larger body movements used to select stimuli clearly was not the major determinant of participant responses. Observers reported that they prioritized more subtle movements related to speech when producing their judgments. Experiment 3 tested this by restricting visible motion to the head and face area to minimize the contribution of other possible cues and showed that coordination was a determinant of affiliation judgments. Observing a clear decrease in accuracy by eliciting use of coordination cues indicates that integration of many cues, including movement correlation, contributes to our remarkable ability to make accurate social judgments.

These studies provided us with two important conclusions regarding movement coordination in social perception: 1) Perception of unintentional coordination observed during social interaction in previous studies is directly influenced by speech-related movement and 2) Multiple factors contribute to social perception however, observers can use coordination as their basis of their affiliation judgments.

Previous studies have shown individuals engaged in conversation become mutually entrained in their movements and this coordination persists even when individuals are interacting verbally without visual input from the other individual. Even when facing an individual with whom participants were not conversing, coordinated movements persisted with the direct conversation partner (Shockley et al., 2003). These findings can be explained by the fact that speech-related movement of the head and face is directly correlated with the auditory signal of speech (Barbosa et al., 2008). Our studies have demonstrated that coordination occurred between individuals actively involved in the conversation but also that third-party observers were sensitive to the speech-related correlation between talkers. The complete explanation for participants' performance in Experiment 2 and 3 warrants further research. We know that full body information provides additional information contributing to greater accuracy but the exact nature of this information has yet to be identified.

Human communication provides a rich information set for making judgments and many cues can contribute to perceptual decisions. The perceptual strength of cues will vary with the context and the observer's history with a
judgment and the manner in which multiple social cues are integrated is still unknown. As in the study of the general visual world (Gibson, 1968), we need to identify potential sources of information in the social world.

Future work examining where observers visually fixate when making affiliation judgments is necessary since the distribution of attention will be a window into the perceptual cues observers use. Information about the allocation of attention may explain individual differences in performance as well as accuracy differences in different social contexts.

These studies aimed to investigate social perception in a less arbitrary and more objective manner. We used motoric correlation to address how judgments of affiliation could be affected by nonverbal factors. We demonstrated that affiliation influenced coordination of movement during social interaction. Further, we showed that observers used this correlation information as a cue, at the expense of accuracy, when making judgments but only when the rich social information set present in human communication was restricted. This broad area of research will continue informing us about our sensitivity to information used for the successful social interactions we encounter everyday.

## Acknowledgments

This research was supported by funding from the Natural Sciences and Engineering Research Council of Canada, the Brian R. Shelton Fellowship and Queen's University.

## References

Ambady, N. \& Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. Journal of Personality and Social Psychology, 64, 431-441.
Barbosa, A. V., Dechaine, R.-M., Vatikiotis-Bateson, E., \& Yehia, H. C. (2012). Quantifying time-varying coordination of multimodal speech signals using correlation map analysis. The Journal of the Acoustical Society of America, 131(3), 2162-2172.
Barbosa, A., Yehia , H. \& Vatikiotis-Bateson, E. (2008). Algorithm for computing spatiotemporal coordination. In Proceedings of the International Conference on Auditory-Visual Speech Processing: Moreton Island, Australia
Capella, J. \& Planalp, S. (1981). Talk and silence sequences in informal conversations III: Interspeaker influence. Human Communication Research, 7(2), 117-132.
Chartrand, T.L. \& van Baaren, R. (2009). Human mimicry. Advances in Experimental Psychology, 42, 219-274.
Dunne, M. \& Ng, S (1994). Simultaneous speech in small group conversation: All-together-now and One-at-atime? Journal of Language and Social Psychology, 13(45), 45-69.
Foulsham, T. \& Underwood, G. (2008). What can saliency models predict about eye movements? Spatial and
sequential aspects of fixations during encoding and recognition. Journal of Vision, 8(2), 1-17.
Gibson, J.J. (1968). What gives rise to the perception of motion? Psychological Review, 75(5), 335-346.
Grahe, J.E. \& Bernieri, J. (1999). The importance of nonverbal cues in judging rapport. Journal of Nonverbal Behavior, 23(4), 253-269.
Horn, B.K.P. \& Schunck, B.G. (1981). Determining optical flow. Artifical Intelligence, 17, 185-203.
Klin, A., Jones, W., Schultz, R., Volkmar, F., \& Cohen, D. (2002). Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. Archives of General Psychiatry, 59, 809-816.
McHugo, G.J, Lanzetta, J.T., Sullivan, D.G., Masters, R.D., \& Englis, B.G. (1985). Emotional reactions to a political leader's expressive displays. Journal of Personality and Social Psychology, 49(6), 1513-1529.
Meltzoff A.N. \& Prinz, W. (2002) The Imitative Mind: Development, Evolution and Brain Bases, Cambridge, United Kingdom: Oxford University Press.
Pardo, J.S. (2006). On phonetic convergence during conversational interaction. Journal of the Acoustical Society of America, 119, 2382-2393.
Richardson, M. Dale, R. \& Shockley, K. (2008). Synchrony and swing in conversation: coordination, temporal dynamics and communication. In I. Wachsmuth, M. Lenzen, G. Knoblich (eds), Embodied Communication in Humans and Machines. (pp. 75-93). Oxford, United Kingdom: Oxford University Press.
Richardson, D.C \& Dale, R. (2005). Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. Cognitive Science, 29, 1045-1060.
Shockley, K. Santana, M.V. \& Fowler, C.V. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. Journal of Experimental Psychology: Human Perception and Performance, 29, 326-332.
Winfree, A.T. (1980). The geometry of biological time. Springer: New York.

# Overcoming Screen Inferiority in Text Learning 

Tirza Lauterman (Tirzal@tx.technion.ac.il)<br>Faculty of Industrial Engineering and Management, Technion - Israel Institute of Technology<br>Technion City, Haifa, 3200003 Israel<br>Rakefet Ackerman (Ackerman@ie.technion.ac.il)<br>Faculty of Industrial Engineering and Management, Technion - Israel Institute of Technology<br>Technion City, Haifa, 3200003 Israel


#### Abstract

Metacognitive monitoring that accompanies a learning task reflects the predicted achievements at test during and at the end of studying the materials. Monitoring reliability is strongly associated with the quality of study regulation and with ultimate performance at test, because it is by this subjective assessment that people decide whether and how to invest more time. Previous studies that compared learning texts on screen to learning from printed texts found that screen learners performed worse and were overconfident about their success. The present research examined two methods for overcoming screen inferiority in these respects. Gaining experience with the study-test task with six different texts allowed improvement. Writing keywords after a delay from learning already eliminated screen inferiority from the first studied texts. In both methods, predictions of performance did not reflect changes in test scores. The two methods clearly affected screen and paper learners differently. This study outlines directions for overcoming screen inferiority, but also calls attention to the effects of context on cognitive and metacognitive processes, beyond the mere interaction between the person and the task content.


Keywords: Reading comprehension; e-learning; humancomputer interaction; metacognitive monitoring; overconfidence.

## Introduction

Learning from texts is a central task in many daily situations. Models of self-regulated learning (Dunlosky \& Hertzog, 1998; Nelson \& Narens, 1990) suggest that reliable subjective assessment of knowledge, or metacognitive monitoring, is essential for effective regulation of learning (Metcalfe \& Finn, 2008; Thiede, Anderson, \& Therriault, 2003). Worryingly, the typical finding in metacognitive studies is that monitoring accuracy regarding comprehension of texts is quite poor (see Maki, 1998). Research suggests that learners use heuristic cues to assess their knowledge (Koriat, 1997). Low monitoring accuracy might be a result of using non-predictive cues. In the case of text learning, such cues may be ease of processing (Dunlosky \& Rawson, 2005) or domain familiarity (Glenberg, Sanocki, Epstein, \& Morris, 1987).

Kintsch (1998) proposed a model of representation levels. According to this model, reading comprehension is constructed from three levels of text representation: words
and signs, sentences, and inference level. It can be derived from this theory that when high-order comprehension is tested, prediction of performance should be more accurate when it relies on the highest representation level of the text. Indeed, studies that demonstrated improvements in monitoring accuracy in text learning often used methods for increasing in-depth processing of the studied materials. In particular, Thiede and his colleagues used writing keywords or writing a summary of the text after a delay (Anderson \& Thiede, 2008; Thiede, Dunlosky, Griffin, \& Wiley, 2005). In another study they made sure to instill appropriate test expectancy for directing participants to the level of processing required for the test (Thiede, Wiley, \& Griffin, 2011). Monitoring reliability is measured in the literature in two respects, resolution and calibration. Resolution is the extent to which predictions of performance at test discriminate between better and lesser known items studied. Calibration is the gap between the predicted performance and actual score at test, and reflects the extent of over- or under-confidence. The above mentioned methods had benefits for performance at test and for resolution. Calibration was not the focus of the mentioned studies that examined the effects of in-depth processing, but is the focus of the present study, as detailed below.

Nowadays, text learning in computerized environments is widespread in numerous domains. For example, reading in depth is required for lawyers using computerized repositories of forensic precedents and for higher education candidates when they face the reading sections in online screening exams such as the Graduate Management Admission Test (GMAT). Thus, it is worthwhile considering whether performance and monitoring accuracy are affected by the reading media of screen versus paper.

Previous studies indicated that people process data more shallowly in computerized environments than they do when studying from print (e.g., Liu, 2005; Morineau, Blanche, Tobin, \& Guéguen, 2005). Ackerman and Goldsmith (2011) addressed these questions by comparing learning texts on screen to learning the same texts from paper, and took the metacognitive processes into account. They found that screen learners performed worse and were overconfident about their success. Overall, people tend to prefer reading texts in depth from print rather than from computerized environments, including modern e-books (Jamali, Nicholas,
\& Rowlands, 2009; Olsen, Kleivset, \& Langseth, 2013; Woody, Daniel, \& Baker, 2010). So a question is raised whether the observed screen inferiority depends on the reluctance of the participants regarding studying texts on screen. Indeed, the results of Ackerman and Goldsmith (2011) were obtained from students who strongly prefer print over computerized learning. However, Ackerman and Lauterman (2012) recently found similar outcomes among engineering students, but only under mild time pressure. Importantly, these students are used to reading from screen and have only a moderate preference for print.

As explained above, overconfidence reflects a calibration bias. This aspect was neglected in studies that attempted to improve monitoring reliability by increasing depth of processing. The present study examined whether methods found effective for improving resolution are also effective for reducing overconfidence. However, notably, most of the previous improvements in monitoring accuracy were achieved in computerized conduction of the experiments (e. g., Anderson \& Thiede, 2008). The present study examined whether such methods are particularly effective on screen, where processing is hypothesized to be shallower even for people experienced in reading from screen. This hypothesis is important in two respects. First, it may point to practical directions for reducing screen inferiority. Second, it has theoretical significance in pointing out that the extent of improvement depends on study context, beyond variables related to the learners and/or to the task content.

## Experiment

The first method we used for reducing screen inferiority relative to paper learning was gaining experience with the task. Multiple study-test cycles were used for providing the participants with appropriate test expectancy for allowing adjustment of their processing level to the requirements and improving the correspondence between the cues used for monitoring and the gained knowledge (Thiede et al., 2011).

The participants of the first group worked on six texts, all on screen or all on paper. The present sample was drawn from the same population used by Ackerman and Lauterman (2012). Following on from them, the participants learned each text under mild time pressure, predicted their performance at test, and answered multiple-choice test questions before moving to the next text.

For the second group, we attempted to direct the participants to a high level of text representation. We did it by asking them to write keywords for each text. It was found effective by Thiede et al. (2005) for improvement of monitoring resolution, but only when there was a delay between text learning and keywords writing. This group studied two texts consecutively. They then wrote keywords, predicted their success at test, and were tested on each of the two texts by their study order. Because of the delay and the study of two texts in a row, test performance for the whole second group was expected to be lower than for the first group that was tested on each test immediately after studying it. The question is whether the delayed keyword
writing reduces screen inferiority because it helps participants who naturally process the information more shallowly on screen, to process it more deeply and therefore eliminate screen inferiority.

## Method

Participants. Eighty undergraduate students from the Faculty of Industrial Engineering at the Technion with no learning disabilities participated in the study. Mean age was 25.8 years old and $48 \%$ were women.

Materials. The six texts, 1000-1200 words (2-4 pages) each, dealt with various topics (e.g., the advantages of coalbased power compared to other energy sources; adult initiation ceremonies in various cultures). An additional, shorter text ( 200 words) was used for familiarizing the participants with the procedure. The texts were taken from web sites intended for reading on screen. Each text formed the basis for a multiple-choice test including five questions testing memory of details and five questions testing higherorder comprehension.

Procedure. The experiment was administered in groups of up to eight participants in a small computer lab. Each group was randomly assigned to read from screen or from paper and for the immediate-test or the delayed keywords conditions. The procedure for the immediate-test group was identical to that used by Ackerman and Lauterman (2012) and was the same for screen and for paper. The participants read each text for seven minutes and were directed to study it for a multiple-choice test. Immediately after reading they provided their predictions of performance (POPs) on two scales (25-100\%), one for memory for details and one for higher-order comprehension, and then answered the test questions. The mean of the two ratings was used for the analyses. This procedure was repeated six times.

For the delayed keywords condition, the participants read two texts consecutively. After reading both, they wrote four keywords for the first text, filled in their POPs, and took the test for the first text. The same procedure (keywords, POPs, test) was done then for the second text. This procedure was repeated for two more text pairs, which were not included in the present analyses. The entire procedure was explained to the participants in advance and the order of the texts was counterbalanced across participants.

## Results

We started our analysis by examining whether the first two texts of the immediate-test group replicate the screen inferiority in performance and overconfidence found by Ackerman and Lauterman (2012) under the same conditions. Figure 1 panel A presents the results. A two-way Analysis of Variance (ANOVA) with Measure (POP vs. test score) $\times$ Medium (screen vs. paper) revealed a main effect of the measure, $F(1,38)=54.64, M S E=101.80, p<.0001$, suggesting a general overconfidence. There was also a


Figure 1: Predictions of performance (POP) and test scores for the first and the last two texts studied for an immediate test are presented in panel A and panel B, respectively. Panel C presents the results for the two texts for which the test took place after a delay and after providing keywords. The error bars represent the standard error of the means.
significant interactive effect, $F(1,38)=12.83, M S E=$ $101.80, p=.001$. As can be seen in the figure, test scores were lower on screen than on paper, $t(38)=2.76, p<.01$, while POP showed the opposite direction, though insignificantly, $t(38)=1.69, p<.10$. Overconfidence was measured as the mean gap between POPs and test scores. The opposite direction of changes - lower test scores and higher POPs on screen - yielded a higher overconfidence level than on paper, $t(38)=3.58, p=.001$. These findings replicate the findings of Ackerman and Lauterman (2012) and form the starting point for our attempts to reduce screen inferiority.

In comparison to the first two texts, a similar ANOVA on the last two texts of the immediate-test group showed only the main effect of the measure, $F(1,38)=11.79, M S E=$ 121.91, $p=.001$, which reflected general overconfidence. There was no interactive effect, $F<1$. A three-way ANOVA of Pair Order (first vs. last) $\times$ Measure (POP vs. test score) $\times$ Medium (screen vs. paper) revealed a triple interactive effect, $F(1,38)=9.42, M S E=68.73, p<.005$. Test scores improved on screen, $t(38)=3.87, p=.001$, but not on paper, $t<1$, and there were no differences in the POPs, both $t \mathrm{~s}<1.2$. Thus, by gaining experience with the task, screen learners improved their test scores, but did not acknowledge this improvement. The outcome was a reduction in their overconfidence, $t(38)=4.08, p=.001$.

The first two texts of the delayed-keywords group also showed a significant overconfidence, $F(1,76)=89.57, M S E$ $=132.80, p<.0001$, but resulted in an elimination of screen inferiority relative to paper, with no interactive effect of measure and media, $F<1$. The triple interaction when comparing the two conditions was significant here as well, $F(1,76)=7.53, M S E=132.80, p<.01$. In this case, the difference stemmed from a near significant reduction in performance after the delay on paper only, $t(39)=1.86, p=$ .07. Screen learners, in contrast, scored similarly in immediate tests without keywords as after a delay but with writing keywords. As in the immediate-test, POPs did not mirror the performance changes found on paper. Thus, the delayed keywords procedure eliminated screen inferiority
relative to paper learning in both performance and overconfidence.

## Discussion

In light of previous findings of screen inferiority relative to paper learning in both performance and overconfidence (Ackerman \& Goldsmith, 2011; Ackerman \& Lauterman, 2012), the present study examined whether students learning from texts presented on screen benefit from using methods that were found in previous studies to contribute to the resolution of metacognitive judgments. As expected, media differences in both performance and overconfidence were eliminated. One group eliminated the media effect by gaining experience with the task and the other group eliminated it by writing keywords and being tested after a delay.

Although predicted, the findings of differences between the media in the effects of the two methods on performance are striking. In the group that gained experience with the task, performance improved for screen learners only. In the group that provided keywords and was tested after a delay, performance was not lower relative to immediate testing for screen learners only. We interpret these findings to suggest that participants who studied on paper spontaneously engaged in effective in-depth learning. Thus, the two methods did not change the effectiveness of their processing. This made experience with the task unnecessary. The keywords provided upon delay also could not increase depth of processing, and thus the delayed test took its toll. For the screen learners, in contrast, spontaneous learning was less effective, so experience with the task led them to improve learning regulation. The delayed keywords led them to overcome the toll of the delayed test. Clearly, this explanation is speculative and requires further research; however, it accords the particular effective regulation found by Ackerman and Lauterman (2012) on paper only with the same population.

Another striking finding is the mismatch between changes in performance and POPs. In all cases, the POPs were
almost constant, while performance was affected by the manipulations and the media used. Thus, overconfidence differences stemmed almost solely from differences in performance. These findings correspond to the wellestablished literature, which suggested that metacognitive judgments are more affected by the materials' internal characteristics than by the external conditions in which the task is performed. For example, while people take into account the a-priori difficulty of paired associates (e.g., related vs. unrelated word pairs), they do not sufficiently appreciate the benefit of repeated memorization of the same list of items (Koriat, Sheffer, \& Ma'ayan, 2002). Similarly, when guided to engage in imagery for elaborated processing of paired associates, although performance improved, it was not appreciated in recall predictions (Rabinowitz, Ackerman, Craik, \& Hinchley, 1982). However, in contrast to this low sensitivity of the metacognitive judgments to knowledge variations, in the previous studies with the same materials (Ackerman \& Goldsmith, 2011; Ackerman \& Lauterman, 2012) POPs did show sensitivity to time conditions, freely allocated, time pressure, and unexpected interruption of the learning. This sensitivity to time conditions, exhibited some correspondence with performance, in particular when studying on paper. The comparison between the previous studies and the present one highlights the dissociation found here between POPs and performance at the tests. The screen participants in the present study did not acknowledge knowledge improvement, even when it was pronounced (last two texts of the immediate test condition). The present line of research examined media and time frames. It will be interesting for future studies to further examine these factors and others that affect POPs' sensitivity to changes in performance.

To sum up, the consistent screen inferiority in performance and overconfidence can be overcome by simple methods, such as experience with task and guidance for in-depth processing, to the extent of being as good as learning on paper. The findings have clear implications. First, software designers and policy makers in numerous contexts should take into account the differences between the media in the quality of monitoring and regulation of learning. Second, the principle of improving the reliability of the cues used for monitoring, which guided us in choosing the methods for improvement, should be taken into account when designing training towards using computerized environments that involve extensive textual sections. However, the observed media differences in the effectiveness of the methods should draw attention to the fact that some methods reported in the literature were examined only on one medium, either screen or paper. From the theoretical perspective, the media effects draw attention to the effects of the context on learning regulation and outcomes, beyond the interaction between a person, with his or her given learning skills, and the study materials (see also Morineau et al., 2005).

## References

Ackerman, R., \& Goldsmith, M. (2011). Metacognitive regulation of text learning: On screen versus on paper. Journal of Experimental Psychology: Applied, 17(1), 1832.

Ackerman, R., \& Lauterman, T. (2012). Taking reading comprehension exams on screen or on paper? A metacognitive analysis of learning texts under time pressure. Computers in Human Behavior, 28, 1816-1828.
Anderson, M. C. M., \& Thiede, K. W. (2008). Why do delayed summaries improve metacomprehension accuracy? Acta Psychologica, 128, 110-118.
Dunlosky, J., \& Hertzog, C. (1998). Training programs to improve learning in later adulthood: Helping older adults educate themselves. In D. J. Hacker (Ed.), Metacognition in educational theory and practice. Mahwah, NJ: Lawrence Erlbaum Associates Inc Publishers.
Dunlosky, J., \& Rawson, K. A. (2005). Why Does Rereading Improve Metacomprehension Accuracy? Evaluating the Levels-of-Disruption Hypothesis for the Rereading Effect. Discourse Processes, 40(1), 37-55.
Glenberg, A. M., Sanocki, T., Epstein, W., \& Morris, C. (1987). Enhancing calibration of comprehension. Journal of Experimental Psychology: General, 116(2), 119-136.
Jamali, H. R., Nicholas, D., \& Rowlands, I. (2009). Scholarly e-books: the views of 16,000 academics: Results from the JISC National E-Book Observatory. Aslib Proceedings: New Information Perspectives, 61, 33-47.
Kintsch, W. (1998). Comprehension: a Paradigm for Cognition. New York: Cambridge.
Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. Journal of Experimental Psychology: General, 126, 349-370.
Koriat, A., Sheffer, L., \& Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. Journal of Experimental Psychology-General, 131(2), 147-162.
Liu, Z. (2005). Reading behavior in the digital environment. Journal of Documentation, 61(6), 700-712.
Maki, R. H. (1998). Test predictions over text material. In D. J. Hacker (Ed.), Metacognition in educational theory and practice. The educational psychology series. Mahwah, NJ, US: Lawrence Erlbaum Associates, Publishers.
Metcalfe, J., \& Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. Psychonomic Bulletin \& Review, 15(1), 174-179.
Morineau, T., Blanche, C., Tobin, L., \& Guéguen, N. (2005). The emergence of the contextual role of the ebook in cognitive processes through an ecological and functional analysis. International Journal of HumanComputer Studies, 62(3), 329-348.
Nelson, T. O., \& Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. Bower
(Ed.), The Psychology of learning and motivation: Advances in research and theory (Vol. 26). San Diego, CA: Academic Press.
Olsen, A. N., Kleivset, B., \& Langseth, H. (2013). E-Book readers in higher education student reading preferences and other data from surveys at the University of Agder. SAGE Open, 3(2).
Rabinowitz, J. C., Ackerman, B. P., Craik, F. I., \& Hinchley, J. L. (1982). Aging and metamemory: The roles of relatedness and imagery. J Gerontol, 37(6), 688695.

Thiede, K. W., Anderson, M. C. M., \& Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. Journal of Educational Psychology, 95(1), 66-73.
Thiede, K. W., Dunlosky, J., Griffin, T. D., \& Wiley, J. (2005). Understanding the delayed-keyword effect on metacomprehension accuracy. Journal of Experimental Psychology-Learning Memory and Cognition, 31(6), 1267-1280.
Thiede, K. W., Wiley, J., \& Griffin, T. D. (2011). Test expectancy affects metacomprehension accuracy. British Journal of Educational Psychology, 81(2), 264-273.
Woody, W. D., Daniel, D. B., \& Baker, C. A. (2010). Ebooks or textbooks: Students prefer textbooks. Computers \& Education, 55(3), 945-948.

# Effects of Explicit Abstract Knowledge and Simple Associations in Sequence Learning 

Jessica C. Lee (jlee4128@uni.sydney.edu.au)<br>Evan J. Livesey (evan.livesey@sydney.edu.au)

School of Psychology, University of Sydney
Sydney, NSW 2006 Australia


#### Abstract

This study examined the effect of an explicit relational rule on sequence learning in a 3 -choice serial reaction time task. Simple probabilistic contingencies between pairs of response cues were used in such a way that the sequence of cues moved predominantly in one direction (i.e. either clockwise or counterclockwise). Performance on cued and miscued responses was compared for a group given a hint about the abstract rule describing the relationship between the response cues, and a group given no information about this relationship. Experiment 1A demonstrated that XYZ and XYX subsequences showed performance differences when the location of the target on each trial was random. Experiment 1B showed that giving participants the explicit hint affected XYZ subsequences more than XYX subsequences. Implications for sequence learning and, specifically, the interaction between rule and instance learning are discussed.


Keywords: serial reaction time; awareness; motor learning; volitional control; sequence learning; rule vs. instance learning

## Introduction

Humans are remarkably capable at learning about sequential material, such as the underlying sequence of locations of a target in a serial reaction time (SRT) task (Nissen \& Bullemer, 1987). A typical paradigm involves participants pressing keys corresponding to the location of a target as it appears in positions on a computer screen. Speed and accuracy of responses are emphasized. Unbeknownst to participants, some or all of the positions can be predicted using deterministic or probabilistic rules. Participants exposed to this structured material generally show a reduction in reaction times ( RTs ) for predicted locations, relative to unpredicted (random) locations, suggesting that they have learned about the underlying sequence. The SRT task is an example of an implicit learning paradigm that appears to show robust learning in the absence of verbalizable knowledge (e.g. Destrebecqz \& Cleeremans, 2001; Willingham, Nissen \& Bullemer, 1989) and intention to learn (Jiménez, Méndez, \& Cleeremans, 1996; Jones \& McLaren, 2009). Since implicit learning is, according to some definitions, unconscious (Reber, 1993), this classification implies that learning is independent of explicit knowledge, and should not be susceptible to cognitive influences (Lewicki, 1986).

However, the implicit status of sequence learning has been challenged by later studies showing that sequence knowledge is reportable when appropriate tests are used (e.g. Jimenez, Mendez, \& Cleeremans, 1996; Perruchet \& Amorim, 1992). The results from several studies (e.g.

Dominey, Lelekov, Ventre-Dominey, \& Jeannerod, 1998; Jimenez, Vaquero, \& Lupianez, 2006; Jones \& McLaren, 2009) suggest that while learning is generally automatic (in that it does not require an intention to learn and occurs under a variety of learning conditions), giving participants knowledge about the sequence or instructions to search for an underlying rule can change what is learned, implying that learning is under some degree of volitional control. Conflicting results and ongoing disagreement about appropriate methodology has meant that no firm conclusions can be made about the status of implicit learning and what learning mechanisms it embodies (Shanks \& St. John, 1994).

A corollary of the implicit/explicit distinction made by several researchers is between the learning of rules and instances. Explicit learning is assumed by some to involve knowledge in the form of symbolic propositions that are apt for describing abstract relations between events (Mitchell, De Houwer, \& Lovibond, 2009). In contrast, implicit learning can be seen as the accumulation of statistical information in an incremental fashion, and many have argued is better suited for learning the surface structure or physical properties of events in a sequence rather than abstract relations (McLaren, Green, \& Mackintosh, 1994; Perruchet \& Amorim, 1992). Conceptualizing implicit learning in this way allows it to be explained using the same simple associative learning mechanisms that have been postulated to explain animal learning (McLaren, Green, \& Mackintosh, 1994). In support of this conceptualization, models such as Elman's (1990) Serial Recurrent Network (SRN) have been quite successful in modeling human performance in the SRT task using associative mechanisms (e.g. Cleeremans \& McClelland, 1991). The SRN captures statistical regularities in the pattern of responses and allows predictions to be made based on a limited temporal context of responses. However, models such as these presuppose that explicit representation of sequence knowledge is limited, and assume rule learning to be a separate, higherorder process. While it seems obvious that humans are capable of rule learning and hypothesis testing, the question of interest is whether these explicit processes have any place in sequence learning.

Where abstract relations between events can be described in ways that do not depend on the physical properties of those events, the content of rule learning often differs from that of instance learning (e.g. Natal, McLaren, \& Livesey, 2013; Livesey \& McLaren, 2009; Shanks \& Darby, 1998;). It is possible to derive evidence for both rule and instance
learning in an SRT task with appropriately constructed sequences. For example, Dominey et al. (1998) found that participants under implicit and explicit learning conditions were able to learn about surface contingencies in an SRT task, but only those in the explicit condition were able to learn about, and transfer their knowledge of, the underlying abstract rule. This study shows that what is learned in an SRT task depends on the learning conditions imposed, and while rule learning requires appropriate learning conditions (sufficient cognitive resources, or in this case, explicit instructions to search for a rule), instance learning can occur automatically.

In Dominey et al.'s (1998) study, evidence of a dissociation between the learning of abstract and surface structure was sought by testing transfer to sequences containing different surface features but the same abstract rule. An alternative approach used by Jones and McLaren (2009) is to allow participants to make a prediction about the target's location before the onset of the target, with the assumption that some sequences will benefit from an intentional search for sequences more than others. In a twochoice (X,Y) SRT task, Jones and McLaren found that participants given incidental learning conditions showed the strongest evidence of learning for subsequences containing an alternation (e.g. YYX, YXY) and the least evidence of learning for subsequences consisting of runs of the same response (XXX). However, when participants were presented with two cue positions and given instructions to predict what would happen on the next trial, this pattern of results was reversed, with the best learning occurring for the more salient XXX subsequences. This study shows that the effect of explicit instructions can differentially affect certain subsequences. Giving participants the intention to learn or alluding to the existence of sequences does not entail that all subsequences will benefit from these manipulations. However, the fact that learning in the SRT task is even affected by these manipulations suggests that sequence learning cannot be considered implicit in the traditional sense.

The translation of abstract rules to the performance of a concrete action is not necessarily a straightforward task, even when the rule is explicitly identified. Many researchers have assumed that it requires intentional mental effort (Gick \& Holyoak, 1983; Gomez, 1997; Shanks \& St. John, 1994). Although the rules used in the studies by Dominey et al. (1998) and Jones and McLaren (2009) could be explained verbally and symbolically, they were relatively complex. Once known, implementation of the rule involved retention of at least two items in working memory in order to use the abstract relationship to determine the next response.

In contrast, in this study a relatively simple rule was used; the sequence of response cues moved clockwise most of the time for some participants, and moved counterclockwise most of the time for others. This rule can be applied purely on the basis of the preceding response but was still abstract in the sense that it involved a relationship between at least two events and can be applied flexibly to any of the
response cues in the sequence. Although the rule was probabilistic in nature and not necessarily obvious to participants performing the task, it was easy to describe and (we assumed) easy to implement once recognized explicitly. Thus the primary aim of current study was to explore the degree to which learning in an SRT task could be affected by explicit knowledge of an abstract rule describing the probabilistic contingencies in the task.

## Experiment 1

Experiment 1 used a simplified version of the SRT task with three response locations (A, B and C, see Figure 1) and probabilistic contingencies to minimize hypothesis-testing strategies and the development of explicit sequence knowledge during training (see Jiménez \& Méndez, 2001). The target never appeared in the same location twice in a row, meaning that each set of three consecutive responses in the sequence contained either three unique responses (XYZ, e.g. $\mathrm{ACB}, \mathrm{ABC}$ ) or a repetition of one response as the first and third in the set (XYX, e.g. ABA, ACA).


Figure 1. Illustration of the target positions (A, B, C) and an example of the contingencies arranged between them (in this case, resulting in a predominantly clockwise direction of motion). Curved, bold lines indicate cued trials ( $p=.75$ ) and dotted straight lines indicate miscued trials ( $\mathrm{p}=.25$ ).

Experiment 1A sought to establish performance differences on these two subsequences using a single control group that performed the SRT task with a randomly generated sequence. Experiment 1 B compared learning between two groups: a group given a hint before the experiment about the nature of the contingencies embedded within the target locations (hint group), and a group who did not receive a hint (no hint group). The contingencies were arranged such that most of the time, the target appeared to be moving in one direction (clockwise or counterclockwise), with the direction of motion randomly chosen for each participant. If the target was moving clockwise, for example, there was a .75 probability that the next location would be the next clockwise position (cued trials), and a . 25 probability that the next location would be the next
counterclockwise direction (miscued trials, and vice versa for counterclockwise, see Figure 1).

Since the contingencies were probabilistic, an explicit hint about the direction of motion would still mean that the location of the target on any given trial could not be predicted with complete accuracy. We expected that both groups in Experiment 1B would learn the sequence, but the group given the hint would show both better overall learning (a larger cueing effect) and higher levels of awareness in subsequent tests. Furthermore, we hypothesized that even with a simple abstract rule its application on each trial would not be straightforward. Since responses are made rapidly, there is little time to prepare for the next response based on the direction of motion. Thus, although the directional rule is applicable to every response, we expected that the effectiveness of the hint when applied to specific instances would be greater on the more salient XYZ subsequences than the XYX subsequences because the presence of a consistent direction of motion for several responses would facilitate the use of the rule.

## Method

Participants and Apparatus In Experiment 1A, fifteen University of Sydney staff and students participated. In Experiment 1B, forty-six University of Sydney first year Psychology students participated in exchange for course credit. The experiment was programmed using PsychToolbox for Matlab (Brainard, 1997; Pelli, 1997) and run on Apple Mac Mini desktop computers connected to 17 inch CRT monitors, refreshed at a rate of 85 Hz . Participants made responses using a standard Apple keyboard and mouse. Testing was conducted in individual cubicles in groups of up to six.

Procedure For both experiments, participants performed a SRT task where they were asked to respond to a series of targets appearing in one of three positions (on the left, top and right) on the computer screen by pressing corresponding arrow keys. Training in both experiments was presented to participants in one continuous block of 720 responses. In Experiment 1 A , the position of the target was completely random, such that no learning could occur. In Experiment 1B, the sequence of locations followed a probabilistic rule such that the sequence of response locations usually moved in a clockwise or counterclockwise direction around the screen, and each trial was either cued ( $75 \%$ of the time) or miscued ( $25 \%$ of the time, see Table 1). The direction of motion was randomly chosen for each participant.

Participants in Experiment 1B were allocated to either a hint or no hint group. The hint group received written instructions at the beginning of the experiment that stated that the target moved in either a clockwise or counterclockwise direction most of the time, and that their task was to work out which direction it went. The no hint group did not receive any explicit instructions about the possibility of an underlying sequence, nor did the control group in Experiment 1A. All participants were informed that
they would have to respond as quickly and as accurately as possible to the targets appearing around the screen.

After the training phase, both groups in Experiment 1B were told that there was a pattern in the sequence of locations in the training phase, and they would now be asked questions about the sequence. Participants completed a recognition test and a prediction test, in counterbalanced order. Participants in Experiment 1A did not complete any awareness tests, as there were no contingencies to assess.

Table 1. Probability of occurrence for each triplet type and conditional probability of the last response cue (cue t) in each triplet given the preceding сие (сие $t-1$ ).

|  | $p($ triplet $)$ | $p($ cue $t \mid$ cue $t-1)$ |
| :--- | :---: | :---: |
| XYZ - Cued | 0.5625 | 0.75 |
| XYZ - Miscued | 0.0625 | 0.25 |
| XYX - Cued | 0.1875 | 0.75 |
| XYX - Miscued | 0.1875 | 0.25 |

Recognition Test On each trial in the recognition test, participants were presented with two sequences in which they had to respond to the target in the same way as in training. One of the two sequences was the same sequence they saw in training, and the other sequence was the opposite (the direction of motion was reversed). Each sequence contained 12 response cues and participants completed 10 trials. After responding to the two sequences, participants were asked to press a key to indicate which of the two sequences they thought was most similar to their training sequence.

Prediction Test The prediction test simply presented participants with the target in one of the three positions and asked which of the remaining two positions they would predict the next position to be. This test consisted of 3 trials (one for each of the target positions).

## Results

All of the following RT analyses refer to response times on correct trials only, excluding responses $>1$ second.

Experiment 1A Participants in Experiment 1A took on average 317 ms to respond on XYZ trials (with $98 \%$ accuracy), and 363 ms to respond on XYX trials (with $94 \%$ accuracy). Thus participants were both faster, $F(1,14)=$ 119.91, $p<.001$, and more accurate, $F(1,14)=21.81, p<.001$, on XYZ trials, relative to XYX trials. These performance differences indicated that the repetition of a recentlyperformed response on XYX trials interfered with fast and accurate responding, or conversely that performing all three responses without repetition facilitated responding. This effect was not based on differential contingencies (after each response, the remaining two cues were equally likely) and is similar to alternation effects found in other choice response tasks, which are most likely not based on sequence learning effects (e.g. see Barrett \& Livesey, 2010). In any
case, performance differences provided further impetus for examining the XYX and XYZ subsequences separately.

Experiment 1B: Training All trials were classified as being either cued or miscued, and the cueing effect taken as the difference between the cued and miscued trials.

Figure 2 shows the mean RTs for cued and miscued trials in Experiment 1B for both the hint $(n=23)$ and no hint ( $n=23$ ) groups, and the mean RTs for the control group ( $n=15$ ) in Experiment 1A, across the 4 training quarters. It is evident that in both the hint and no hint groups, participants were slower to make a response on miscued trials, and faster to make a response on cued trials, relative to what would be expected without any contingencies (the control group). An ANOVA with group (hint x no hint) as a between-subjects factor, and cueing (cued x miscued) and quarter (1-4) as within-subjects factor revealed an overall cueing effect, $F(1,44)=344.15, p<.001$, and a marginal interaction with group, $F(1,44)=3.99, p=.05$, with the hint group exhibiting a larger cueing effect overall. There was also a significant linear trend in the cueing effect, $F(1,44)=74.17, p<.001$, which did not interact with group, $F<1$, indicating that the cueing effect increased during training.


Figure 2. Mean reaction times for the hint and no hint groups in Experiment 1B, and mean reaction times for the control group in Experiment 1A.

To examine whether the effect of the hint differed between the two subsequences, a repeated measures ANOVA with cueing (cued $x$ miscued), subsequence type (XYZ x XYX) and quarter (1-4) as within-group factors and group (hint x no hint) as a between-subjects factor was performed. As hypothesized, a significant 4-way interaction was found $F(3,132)=3.37, p=.02$. The cueing effect for both
subsequence types and for both groups is shown in Figure 3. It is evident that while both the hint and no hint groups obtain similar cueing effects for the XYX subsequences across training, the hint group's cueing effect increased sharply in the $3^{\text {rd }}$ quarter. This may be because it took participants in the hint group some time to translate the hint given at the start of the SRT task into confident knowledge about the direction of motion, and therefore for this knowledge to affect their performance.

There was a significant group difference for XYZ cueing, $F(1,44)=7.29, p=.001$, but not for XYX cueing, $\mathrm{F}<1$. Thus it appears that the effect of the hint increased cueing for XYZ subsequences but not for XYX subsequences, relative to the no hint group.


Figure 3. Cuing effect by training quarter in Experiment 1B, for both groups and subsequence types, showing a significant difference between hint and no hint groups for XYZ cuing, but not for XYX cuing.

Recognition Test The hint group showed a recognition score (61.1\%) that was statistically above chance, $F(1,21)=8.56, p=.008$, while the no hint group did not (55.2\%, $F<1$ ).

Prediction Test Mirroring the recognition test, the hint group showed a level of performance (65.4\%) that was statistically higher than chance, $F(1,21)=8.06, p=.01$, and the no hint group did not $(59.4 \%, F(1,21)=1.58, p=.22)$.

Cueing and Awareness To examine the relationship between awareness and cuing, each participant's recognition score was correlated with their cuing effect for XYZ and XYX sequences separately (Figure 4). There was a significant correlation between recognition and cuing for the

XYZ subsequences in the hint group only, $r(23)=.45, p=.03$. Comparing the correlation coefficients between groups, there was a stronger relationship between awareness and cuing in the hint group for XYZ subsequences ( $r(22)=.45$ ) than the no hint group $(r(22)=.2), z=2.2, p=.028$.


Figure 4. Scatterplots of the cuing effect (in seconds) as a function of recognition accuracy for each subsequence type
(XYX and XYZ) and each group (Hint and No Hint) in Experiment 1B.

## General Discussion

In this study, we observed a robust sequence learning effect using probabilistic contingencies arranged between 3 target locations in both a group given an explicit hint about an abstract rule underlying the contingencies and a no hint
group who performed the task as usual. The hint group exhibited a marginally larger cuing effect overall and produced above-chance results on both a recognition and prediction test.

Experiment 1A demonstrated that participants were both faster and more accurate to respond on XYZ subsequences (consisting of 3 unique responses) than on XYX subsequences (when the response on the current trial is the same as the response 2 trials back) when there were no contingencies present. Closer inspection of the different subsequences indicated that the benefit of the hint group over the control group in Experiment 1B was only evident on subsequences that did not contain a repetition (XYZ), and that participants in both groups learned about subsequences with a repetition (XYX) equally well. While the hint group were able to produce results on the awareness tests at a level greater than chance, there was a significant correlation between cueing and recognition on XYZ subsequences only.

The results from this study suggest that the relationship between explicit abstract knowledge and performance in sequence learning tasks is complex. The SRT task utilized in this experiment shows a strong dissociation between two subsequence types - participants given an explicit hint about the underlying contingencies could use this knowledge to produce a larger cueing effect (relative to those given no hint) on XYZ subsequences, and the amount of cueing was related to how well those in the hint group performed in the recognition test. On the other hand, whether or not participants received the hint did not make any difference to the magnitude of the cueing effect on XYX trials, and the cueing effect was not correlated with recognition performance.

These results indicate that while the hint may have been successful in helping participants to discover the abstract rule amongst the contingencies, this knowledge could only be applied to XYZ trials and not XYX trials. One potential reason for this may be that certain subsequences are learned better in intentional learning conditions because they are more salient (see Jones \& McLaren, 2009). The repetition of two clockwise or counterclockwise positions in the XYZ subsequences may be particularly salient if participants are searching for the correct direction of the target. This is in line with our initial prediction that the abstract information provided by the hint would be easier to implement (and thus produce a greater facilitatory effect) on trials where the direction of motion was consistent for several cues. An alternative reason may be simply that XYX subsequences are harder to respond to in general, and therefore applying explicit knowledge on these trials may also be more difficult. According to this explanation, whatever property of the XYX subsequences that produced the slower reaction time and lower accuracy in Experiment 1A might also be responsible for interfering with the expression of any knowledge of the sequence that participants had acquired.

While Experiment 1B demonstrated that giving participants explicit knowledge did affect their SRT
performance, it is also obvious that this knowledge is not necessary to display a cueing effect. In fact, very robust cueing effects were evident in the no hint group, along with poor performance (not differing significantly from chance) on awareness tests. However, the fact that giving participants an explicit hint about the underlying sequence affected performance suggests that there is some degree of volitional control in sequence learning, and that learning is not impervious to cognitive influences such as intention to learn. However, whether or not explicit knowledge about an abstract rule can be expressed in sequence learning seems to be dependent on the properties of the subsequences to be learned.

In summary, this experiment demonstrates that sequence learning does not appear to be independent of explicit knowledge of the abstract relations between the cues in the sequence. However, participants are not able to apply their knowledge equally to all subsequences even when, in principle, the abstract relation applies to all instances. These results are most consistent with an explanation in which simple associations between events are learned and expressed relatively automatically, but explicit symbolic knowledge has a strong influence on performance only when specific conditions permit its use.

## Acknowledgments

This research was supported by a University of Sydney Bridging Support Grant to EJL. The authors would like to thank Lauren Shone and Dorothy Kwok for helpful comments on an earlier version of this manuscript.

## References

Barrett, L. C. \& Livesey, E. J. (2010). Dissociations between expectancy and performance in simple and twochoice reaction-time tasks: A test of associative and nonassociative explanations. Journal of Experimental Psychology: Learning, Memory and Cognition, 36, 864877.

Brainard, D. H. (1997). The Psychophysics Toolbox. Spatial Vision, 10, 433-436.
Cleeremans, A., \& McClelland, J. L. (1991). Learning the structure of event sequences. Journal of Experimental Psychology: General, 120, 235-253.
Destrebecqz, A. \& Cleeremans, A. (2001). Can sequence learning be implicit? New evidence with the process dissociation procedure. Psychonomic Bulletin and Review, 8, 343-350.
Dominey, P. F., Lelekov, T., Ventre-Dominey, J., \& Jeannerod, M. (1998). Dissociable processes for learning the surface structure and abstract structure of sensorimotor sequences. Journal of Cognitive Neuroscience, 10, 734-751.
Elman, J. L. (1990). Finding structure in time. Cognitive Science, 14, 179-211.
Gick, M. L. \& Holyoak, K. J. (1983). Schema induction and analogical transfer. Cognitive Psychology, 15, 1-38.

Gomez, R. L. (1997). Transfer and complexity in artificial grammar learning. Cognitive Psychology, 33, 154-207.
Jiménez, L. \& Méndez, C. (2001). Implicit sequence learning with competing explicit cues. The Quarterly Journal of Experimental Psychology, 54A, 345-369.
Jiménez, L., Méndez, C., \& Cleeremans, A. (1996). Comparing direct and indirect measures of sequence learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 22, 948-969.
Jones, F. W., \& McLaren, I. P. L. (2009). Human sequence learning under incidental and intentional conditions. Journal of Experimental Psychology: Animal Behavior Processes, 35, 538-553.
Lewicki, P. (1986). Nonconscious social information processing. New York: Academic Press.
Livesey, E. J., \& McLaren, I. P. L. (2009). Discrimination and generalization along a simple dimension: Peak shift and rule-governed responding. Journal of Experimental Psychology: Animal Behavior Processes, 35, 554-565.
McLaren, I. P. L., Green, R.E. A., \& Mackintosh, N. J. (1994). Animal learning and the implicit/explicit distinction. In Implicit and explicit learning of languages (ed. N. C. Ellis), pp. 313-332. London: Academic.
Mitchell, C. J., De Houwer, J, \& Lovibond, P. F. (2009). The propositional nature of human associative learning. Behavioral and Brain Sciences, 32, 183-246.
Natal, S. D. C., McLaren, I. P. L., \& Livesey, E. J. (2013). Generalization of feature- and rule-based learning in the categorization of dimensional stimuli: Evidence for dual processes under cognitive control. Journal of Experimental Psychology: Animal Behavior Processes, 39, 140-151.
Nissen, M. J., \& Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. Cognitive Psychology, 19, 1-32.
Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. Spatial Vision, 10, 437-442.
Perruchet, P., \& Amorim, M. A. (1992). Conscious knowledge and changes in performance in sequence learning: Evidence against dissociation. Journal of Experimental Psychology: Learning, Memory, and Cognition, 18, 785-800.
Reber, A. S. (1993). Implicit learning and tacit knowledge. Oxford, England: Oxford University Press.
Shanks, D. R. \& Darby, R. J. (1998). Feature- and rulebased generalization in human associative learning. Journal of Experimental Psychology: Animal Behavior Processes, 24, 405-415.
Shanks, D. R. \& St. John, M. F. (1994). Characteristics of dissociable human learning systems. Behavioral \& Brain Sciences, 17, 367-447.
Willingham, D. B., Nissen, M. J., \& Bullemer, P. (1989). On the development of procedural knowledge. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15, 1047-1060.

# Learning and representation of causative motor actions: a neural network model and its use in an embodied theory of language 

Jeremy Lee-Hand (leehandjeremy@gmail.com) and Alistair Knott (alik@cs.otago.ac.nz) Department of Computer Science, University of Otago, New Zealand


#### Abstract

In this paper we present a neural network model of motor learning structured around circuits which associate motor actions with their sensory effects, as proposed by Hommel et al. (2001). The network implements a novel model of causative actions, which bring about specified distal movements in manipulated target objects (e.g. bending a lever). It also serves as the basis for a novel embodied account of the syntax of causative sentences such as John bent the lever.


Keywords: motor control, neural networks, embodied models of language

## Effect-based action representations in neuroscience and language

A common idea in models of action representation is that an agent's actions (also known as motor programs) are encoded in a way which makes reference to the sensory effects they bring about. This idea has a long history, but in recent research it is most strongly associated with Prinz's (1997) theory of 'common coding' and Hommel et al.'s theory of 'event codes' (Hommel et al., 2001). The key idea uniting these models is that motor programs are not defined purely within the motor domain: their neural representation includes a specification of the sensory effects they bring about, in one or more sensory modalities. This position can be supported both on theoretical grounds and through experiments; we will give brief examples of each kind of argument.

Theoretically, a strong argument for this view of action representation comes from considerations about how actions are learned. It is uncontroversial that an agent's repertoire of motor programs is learned through some kind of reinforcement. A reinforcing signal is a sensory signal. When an agent executes a motor program and generates a rewarding signal, an association is made between the sensory signal and this particular program. After a certain amount of training, if all goes well, the sensory signal will become associated with a range of related motor movements, which bring it about in different ways or under different circumstances, perhaps in ways which are parameterised or organised by features of the sensory stimulus. At this point, if the agent activates the sensory signal, this will bring about one of these movements, and result in reward. But equally importantly, the group of motor movements associated with the sensory signal can now be thought of as comprising an action category, in virtue of their shared ability to evoke the stimulus. Categories are defined around central concepts or prototypes, and in this case the unifying concept is a sensory one. For this reason, it makes sense to talk about action categories as being defined by the sensory effects they bring about.

Experimentally, the idea that actions are defined by their effects has been suported in several ways. For instance, there have been many studies exploring variations on the wellknown stimulus-response compatibility effect (Simon, 1969). A good example is a study by Hommel (1993). Here subjects had to respond to an auditory stimulus by pressing a button, either with the left or right hand. The tone of the auditory stimulus indicated which button the subject should press. But as a distracting factor, the stimulus was also presented either on the left or the right. The classical stimulus-response compatibility effect is that subjects are slower to respond if the spatial location of the stimulus is incompatible with the hand which must respond. In Hommel's experiment, button presses generated a reafferent visual stimulus whose location could be decoupled from the location of the hand pressing the button, to explore whether the compatibility effect operates in the domain of motor movements or that of their sensory consequences. Button presses consistently produced a visual stimulus: illumination of a light. In one condition the light appeared on the same side as the hand (e.g. left button presses illuminated a light on the left), while in another it appeared on the opposite side (e.g. left button presses illuminated a light on the right). Hommel found that the stimulus-response compatibility effect depended on compatibility with the perceptual effects of button-presses, rather than on the hand which was used. This shows that the way subjects encode actions does make some reference to their sensory consequencesat least enough to interfere with stimulus-response mappings. Effect-based representations of motor actions are also supported by several studies of the neural representation of actions; see for instance Umiltà et al. (2008); Matsumoto et al. (2003).

Another interesting piece of evidence for effect-based action representations comes from a completely different area of cognitive science: linguistics. The evidence comes from a phenomenon called the 'causative alternation'. This is found in many languages, but we will illustrate with English. Consider the following two sentences:
(1) John bent the lever.
(2) The lever bent.

As these show, the verb bend can be used in a transitive sentence, where the lever appears as the object (Example 1) or in an intransitive sentence, where it appears as the subject (e.g. Example 2). On the face of it, a syntactician would have to assume two different senses of the word bend: one which describes the lever as an agent and one as a patient. But this is counterintuitive, since what happens to the lever is the same
'Underlying' syntactic structure
John CAUSED the lever bent
'Surface' syntactic structure
John bent the lever $\square$

Figure 1: Derivation of John bent the lever by movement from an underlying syntactic structure


Figure 2: (a) The hand/arm (b) detail of a single finger pad
in each case. A way to avoid assuming this implausible ambiguity in the verb bend is to argue that the transitive sentence John bent the lever really means John caused [the lever to bend]. This analysis can be neatly expressed in syntactic theories which posit that sentences have an 'underlying' syntactic structure which is distinct from their surface form: an idea associated with Chomskan accounts of syntax (see e.g. Chomsky, 1995). In a Chomskyan framework, we can argue that the underlying structure of Example 1 is John caused [the lever bent], as shown on the left of Figure 1. At this level of analysis, 'the lever' is the subject of bend, just as it is in Example 2. In a Chomskyan model, the surface structure of Example 1 is produced by moving the lower verb bent into the position of the higher verb caused, as shown on the right of Figure 1.

In this paper we have two aims. We will first introduce a computational model of the learning and control of causative actions, which implements a particular take on Hommel et al.'s conception of event codes. The model has several interesting features as an account of action representation, which we will briefly discuss. But our other main aim is to juxtapose an account of processing in the motor control network with the syntactic analysis of causative verbs just sketched above. We will argue that the network may provide a framework that allows the syntactic analysis to be expressed in terms of neural mechanisms.

## A platform for learning and control of simulated actions

Our computational model was implemented in a software environment for simulating hand/arm actions called GraspProject (Lee-Hand et al., 2012; for details see Neumegen, 2013). The environment is built on top of the JMonkey games engine, which uses the Bullet physics engine to define objects made up of linked rigid bodies, and OpenGL to render graphical views of these. GraspProject provides a simple model of the hand and arm, with three degrees of freedom in the arm


Figure 3: Architecture of the motor control network
(two at the shoulder and one at the elbow) and one in the hand controlling grip aperture (see Figure 2a). It also provides a fairly rich model of the touch sensors in the fingers. Finger pads are modelled as deformable grids of rigid bodies connected by springs. (A single finger pad in light contact with a solid surface is shown in Figure 2b.) Information about light touches is provided by collision detectors on each pad, and information about stronger touches which deform the surface of the skin is read from the joint angles between adjacent pads.

## Architecture of the motor control network

Our model of the motor system is a neural network for learning hand actions directed at target objects. It provides a simple model of some aspects of infant motor development.

The general architecture of the network is shown in Figure 3. It consists of three sub-networks arranged in sequence. These are assumed to be trained at three successive developmental stages, by reward signals of different degrees of complexity. In this scheme, the system is initially rewarded by very simple sensory signals, which train a simple motor circuit, but as learning takes place in this circuit, more complex reward signals become available, which in turn train higherorder circuits. The first two networks are described in detail in Lee-Hand et al. (2012), and their interaction with the third network is described in Lee-Hand (2013).

The first network to be trained is called the reach network (see the lower part of Figure 3). This network learns a function which maps a visual representation of a target object onto
a goal motor state of the hand and arm. (The visual representation has two components, one relating to the position of the target, the other to its shape. The former representation maps to a goal arm state; the latter to a goal hand state.) During training, the agent visually attends to objects in its perispace, and executes hand/arm actions at random. Sometimes these actions result in its hand touching the target, evoking a touch signal (the simple touch signal). This signal is intrinsically rewarding (as in Oztop et al., 2004). The touch signal has two functions. First, it allows a proprioceptive representation of the agent's current motor state to be copied into the medium holding its goal motor state (see the upper arrow leading from the simple tactile signal). Second, it allows the reach network to be trained, so that the current visual representation of the target object is associated with this newly specified goal motor state, and similar presentations of the target in the future will automatically elicit an appropriate motor goal.

This simple circuit implements a particular version of Hommel et al.'s model of event codes. Learning in the circuit creates what can be thought of as a single simple action category, associated with the sensory representation of a touch to the hand: a network which maps visual stimuli onto motor goals which will bring about this representation. Motor goals in the circuit are associated with sensory stimuli in three ways. Any representation in the motor goal medium is implicitly associated with one particular reward stimulus (a simple touch sensation). Specific motor goals are associated axiomatically with specific motor states (sensed proprioceptively) when the reward stimulus is evoked. And specific motor goals are also associated through learning with arbitrary sensory stimuli (in this case visual), which carry information about the motor states associated with reward signals. Again this happens at the time the reward stimulus is evoked. The key devices in the circuit are reward-gated learning and copy operations. These devices will be replicated in the other two networks.

The reach network generates a motor goal-but of course there must also be a mechanism which achieves this goal. At the first developmental stage, we assume this mechanism is a simple feedback motor controller. This device takes the current motor state and the goal motor state and generates a motor signal proportional to the difference between them, in a direction which reduces this difference. (The controller is not shown in Figure 3.) A feedback controller does not need to be trained; it can be assumed to be present at birth. (We use a PID controller; see e.g. Araki, 2006). However, mature motor control involves a mixture of feedback control and feedforward control (see e.g. Kawato et al., 1987). Feedforward control exploits learning about the properties of the agent's motor system to optimise action trajectories. If we think of the feedforward controller in sufficiently general terms, we can say that it is through learning in this controller that an agent can acquire a repertoire of different action categories. Different actions (like grabbing or punching or slapping) have different characteristic trajectories of the
hand and fingers; the feedforward control system somehow learns about the distinct effects of particular trajectories and creates action categories associated with each. However, it is not clear how different trajectories are represented in the biological motor control system. There is good evidence that agents do not compute detailed trajectories in advance; these are only generated 'on the fly', as an action is actually underway (see e.g. Cisek, 2005). Our network implements a particular idea about how trajectories are represented. We assume that the agent evoking a goal motor state can generate learned perturbations of this goal state as an action is under way, which deviate the hand from the normal course it would take under simple feedback control. For instance, to generate a trajectory bringing the hand onto the target from above, the goal state could be temporarily perturbed to a point above the target, so the hand initially moves higher than it would normally do. This idea is discussed in more detail and evaluated in Lee-Hand et al. (2012). This kind of learning takes place in the second network in our model, the motor program network (see the middle of Figure 3).

The motor program network learns to map a goal motor state onto a perturbed goal motor state, which is applied at the start of a reach action and removed when the hand is at a specified distance from the target. Learning in this network begins when the reach network reliably generates actions that lead to reward signals. During training, random perturbations are applied to the goal motor state produced by the reach network. From time to time, these perturbations result in richer tactile reward signals than those used to train the reach network. There are several different signals, which result from particular perturbations. Some perturbations result in a grasp or near-grasp, which generates a characteristic rich tactile stimulus. Others result in a slap movement, which generates another, different, tactile stimulus. (These rich stimuli are almost never generated through pure feedback control, because they result from special trajectories.) When a rich tactile stimulus is generated, copy and learning operations take place in the motor program circuit which are analogous to those in the reach circuit. First, the tactile stimulus is copied to an area holding 'motor programs'. Second, the motor program network is trained to map the current goal motor state, plus the currently active motor program, onto the perturbation which resulted in the reward. After this learning, activating a specific motor program will generate an action with a characteristic trajectory. We envisage motor programmes competing with one another, with the winner being selected.

Note that the motor program network must execute in parallel with the simple reach network. It basically modulates the behaviour of the simple network, in a manner reminiscent of Brooks' (1991) subsumption architecture. In order to execute a motor program, it is important that the whole motor program circuit is enabled, or turned on. Accordingly, while different motor programs provide different input to the motor program network, they also uniformly generate a control signal to enable the network they provide input to.

The final network to be trained is the causative action network (see the top of Figure 3). Our assumption here is that there is a higher level of motor control where sensory reward signals are generated within a perceptual module whose primary function is to classify actions observed occurring in the external world. There is a well-studied perceptual module of this kind in the brain, implemented in a pathway from sensory cortices (in particular visual cortex) through the superior temporal sulcus (STS) and inferior parietal cortex to the premotor cortex (see e.g. Keysers and Perrett, 2004). When an agent allocates attention to an external object, representations in this pathway encode the actions of this object in various ways. Canonically, the action recognition pathway is active when an agent is passively observing the external world. But consider what happens when the agent is attending to an external object as a target, while directing a hand action towards it. Any actions evoked in the action recognition pathway in this scenario are potentially actions brought about by the hand action. We propose that during action execution, action signals evoked in the action recognition pathway function as reward signals, which train the causative action network to bring about particular distal actions in the world.

Training in this higher-level motor circuit again proceeds by random generation of perturbations to the goal motor state delivered by the reach network. In this circuit, sequences of perturbations are applied, to generate still more complex trajectories. (This is depicted in the diagram by a recurrent input, though in our implementation we 'unroll' this recurrence and generate exactly two perturbations.) Some of these sequences cause particular patterns of movement in the target object, which are interpreted as external actions by the action recognition system. Activation of an action representation in the action recognition system when performing an action on a target object is hard-wired to generate a reward signal. This signal has two effects. First, the observed action is copied to a specialised motor medium: specifically, a medium in which action plans are held. Second, the causative action network is trained to map the basic goal motor state delivered by the reach network onto the sequence of perturbations which led to reward. Note that the network also takes representations in the action planning system as input. After training, the causative action network can take a simple goal motor state, plus an action representation in the action planning system, and generate a sequence of perturbations which will lead to observation of a specific action on the attended target. And different patterns in the action planning system will lead to different observed actions.

This network enables a rich repertoire of actions to be learned. It preserves Hommel et al.'s idea that action representations are organised around their perceptual effects. But since the action recognition network generates rich, highlevel percptual signals, a correspondingly rich set of motor programs can be established. At the same time, the basic mechanisms through which learning happens are the same as in much simpler motor learning systems.

Part of the design of the causative action circuit is that 'cause' is motor program in its own right, which competes within the motor program selection system against regular motor programs like 'grasp' and 'slap'. One important difference is that the 'cause' action enables the causative actions network rather than the motor program network, but other than that it counts as a regular motor program. This raises some important questions about how causative actions are planned and executed. When an agent decides to perform a causative action, presumably he has some particular caused action in mind. But at the time of planning, this caused action is in the future: minimally, the agent must bring his hand into contact with the target object before he can cause it to move in any way. Moreover, there is hardly ever a clear way of decomposing a causative action into a simple reach action and a subsequent manipulation. In order to cause a particular action in a target object, the trajectory of the hand towards the object must typically be biased from the very start: for instance, to cause an object to squash, the hand must approach the target from a particular direction, and with particular force. So the movements which bring about the caused action must be initiated some time before the action is perceived.

Our way of addressing this issue in the network is to activate the motor correlates of perceived actions in a medium holding planned actions, rather than in the medium of regular motor programs like 'grasp' and 'slap'. An underlying assumption in our model is that an agent brings about actions through planned sequences of sensory or motor operations (for details see Knott, 2012). We also assume that planned sequences are selected as wholes, and that the component actions in a planned action sequence are active in parallel in the working memory medium where actions are planned. (This assumption is well supported by single-cell recordings in monkeys; see e.g. Averbeck et al., 2002). When the causative actions network is exploring causative actions, it will activate the 'cause' motor programme experimentally, and choose a random sequence of perturbations. In some cases, this results some time later in activation of an action in the action recognition system: say 'squash'. This observed action activates a corresponding planned action; additionally the sequenceplanning system will learn that the sequence 'cause', 'squash' is a good one to execute in the current context, so that when a similar context occurs in the future, it will activate this planned sequence. Now consider what happens when the planned sequence is executed. The agent first executes the motor programme 'cause'. This enables the causative action network, which generates a sequence of perturbations. Crucially, the causative action network also takes input from the planning medium in which the caused action ('squash') is active as part of the planned sequence. So as soon as it is initiated, the network is configured to generate the perturbation sequence which led to the caused action, even before this action actually occurs.

The key mechanism enabling causative actions to be executed is one which activates a sensory representation (the
squash action) as a goal some time before it is evoked as a sensory stimulus. Note that something very similar happens in the other networks; for instance in the reach network the actual motor state where the touch sensation occurs is activated as a goal motor state. In the simple network this activation is possible because visual perception provides information about reward-associated motor states. In the higher-level causative actions network, the advance notification of reward comes from the working memory system which stores prepared actions. But the effect is much the same.

## Experiments learning causative actions

We built two objects in the simulation environment which could undergo specialised kinds of action. One comprised two horizontal planes connected by a spring; this object could be 'squashed' by pushing down on it. The other was a lever which could pivot around a joint; this object could be bent. Our action recognition system consulted the states defined in the physics engine directly; we did not attempt to simulate the action recognition module (though we did simulate the visual inputs to the reach network in a simple way). Each training trial involved the presentation of a single object (either the bendable target, the squashable target or a rigid object) in one of several possible positions. This led to activity in the nework and a motor action. If the rigid object was presented, 'grasp' was activated in the motor program circuit, so that this circuit could be trained; otherwise the activated program was always 'cause', the operation activating the causative action network. This network always generated two perturbations in sequence. At the start of training, perturbations were annealed with noise; this was gradually reduced as the network learned. If the action recognition system detected an action $A$, it generated a reward, and the sensed action activated a corresponding unit in the action planning system, resulting in the sequence ['cause', A]. This sequence plan was also associated with the visual target shape, so subsequent presentation of this target would activate the plan. As training progressed, the system learned perturbations which brought about particular perceived actions, and also learned to map visual target representations onto appropriate sequence plans. Details of the network's training and testing are given in LeeHand (2013).

## A syntactic analysis of causative actions, and a sensorimotor interpretation

It is interesting to compare the structures used by our network to learn and generate causative actions with the syntactic structure of sentences reporting causative actions. As discussed at the start, a Chomskyan analysis posits an underlying level of syntactic representation at which the sentence John bent the lever contains an explicit 'cause' action as its main verb, which takes as its argument a nested action ('The lever bent'). The network model also contains an explicit 'cause' operation (the 'cause' motor programme), and a nested action (the action delivered by the action recognition system).

An emerging school of thought in cognitive science is that concrete sentences get their meanings by evoking embodied representations in the sensorimotor system (see e.g. Glenberg and Kaschak, 2002; Barsalou, 2008). The network model of causative actions presented above may support an interesting 'embodied' account of the semantics of causative transitives like John bent the lever. We will conclude by discussing this possibility.

The underlying structure of sentences in Chomsky's (1995) Minimalist model is called 'logical form' (LF). The LF of John bent the lever is shown in Figure 4. LF represen-


Figure 4: LF of John bent the lever, including head-raising operation
tations have a right-branching recursive structure: the units of recursion, called X-bar schemas (XPs), are indicated with boxes in the figure. (Details of the higher two XPs are omitted for simplicity.) Knott (2012) uses Minimalist LF structures to express a strong claim about the embodied nature of sentence meanings. In his proposal, the LF of any sentence reporting a concrete episode in the world can be interpreted as a description of the sensorimotor processes through which this episode is experienced. Knott assumes Ballard et al.'s (1997) account of sensorimotor processing, which posits that this processing is organised into well-defined sequences of attentional or motor operations called 'deictic operations'. The key idea in Knott's proposal is that the LF structure of a concrete sentence describes a sequence of deictic operationsspecifically, that each XP in the structure describes a single operation. The sensorimotor denotations of XPs are shown in red in Figure 4.

This general proposal makes a specific prediction about the sequential structure of a causative action. As shown in Figure 4 , the XP introducing CAUSE immediately dominates the XP presenting the nested action. Knott's proposal thus predicts that a causative action involves two stages: activation of a 'cause' operation, followed by experience of the 'bend' action. And indeed, execution of causative actions in our network model has this sequential character. So the recursive structure of LF has the right general form.

However, an additional neat correspondence can be drawn
between the structure of LF and the structure of the sensorimotor routine. As discussed above, the hand movement initiated by the 'cause' motor program must follow different trajectories to the target to achieve different causal effects on it. In the network model we catered for this by having the causative actions network take input from a medium representing a casual effect as it is planned, rather than as it is later observed. We assumed that this medium holds the entire sequence of actions being executed by the agent, as a sustained signal, so as soon as the causative network is engaged, its output is influenced by the planned action to be brought about. Now note that in the LF structure in Figure 4, the lower verb bent raises to the position of the higher verb CAUSE (as shown by the arrow) so that it can be pronounced in this higher position, giving the surface form John bent the cup. We propose to explain the Minimalist device of verb raising in sensorimotor terms by assuming in general that surface verbs describe motor actions as they are planned rather than as they actually occur. The reason why the verb bent can be pronounced at the higher verb position is that it denotes a signal in the planning medium, which is tonically active through the whole executed routine. In fact, this account of verb raising follows naturally from a wider sensorimotor account of verb raising which was proposed by Knott (2012) and is implemented in a neural network model of sentence generation described in Takac et al. (2012). But in the current context, the key point is that the sensorimotor interpretation of LF explains both the structural relationship between the CAUSE verb and its complement VP and the extended syntactic domain of the nested verb bend at LF which allows it to appear in the position of the CAUSE verb in surface structure.

## Summary

In this paper we presented a neural network model of the learning and execution of causative actions. The model embodies a particular take on Hommel et al.'s proposal that actions are defined in terms of their effects: a basic circuit implementing this principle is replicated in three different components of the network, at different levels of abstraction. At the same time, the network provides the basis for an interesting account of the syntactic structure of causative sentencesspecifically, of the relation between a cause predicate and the action which this predicate introduces.

## References

Araki, M. (2006). PID control. In H. Unbehauen, editor, Control Systems, Robotics and Automation Vol. II.
Averbeck, B., Chafee, M., Crowe, D., and Georgopoulos, A. (2002). Parallel processing of serial movements in prefrontal cortex. PNAS, 99(20), 13172-13177.
Ballard, D., Hayhoe, M., Pook, P., and Rao, R. (1997). Deictic codes for the embodiment of cognition. Behavioral and Brain Sciences, 20(4), 723-767.
Barsalou, L. (2008). Grounded cognition. Annual Review of Psychology, 59, 617-645.

Brooks, R. (1991). Intelligence without representation. Artificial Intelligence, 47, 139-159.
Chomsky, N. (1995). The Minimalist Program. MIT Press, Cambridge, MA.
Cisek, P. and Kalaska, J. (2005). Neural correlates of reaching decisions in dorsal premotor cortex: Specification of multiple direction choices and final selection of action. Neuron, 45, 801-814.
Glenberg, A. and Kaschak, P. (2002). Grounding language in action. Psychonomic Bulletin and Review, 9(3), 558-565.
Hommel, B. (1993). Inverting the Simon effect by intention. Psychological Research, 55, 270-279.
Hommel, B., Müsseler, J., Aschersleben, G., and Prinz, W. (2001). The theory of event coding (TEC): A framework for perception and action learning. Behavioral and Brain Sciences, 24, 849-878.
Kawato, M., Furawaka, K., and Suzuki, R. (1987). A hierarchical neural network model for the control and learning of voluntary movements. Biological Cybernetics, 56, 1-17.
Keysers, C. and Perrett, D. (2004). Demystifying social cognition: A Hebbian perspective. Trends in Cognitive Sciences, 8(11), 501-507.
Knott, A. (2012). Sensorimotor Cognition and Natural Language Syntax. MIT Press, Cambridge, MA.
Lee-Hand, J. (2013). A neural network model of causative actions. MSc thesis, Dept of Computer Science, University of Otago.
Lee-Hand, J., Neumegen, T., and Knott, A. (2012). Representing reach-to-grasp trajectories using perturbed goal motor states. In Proceedings of the Pacific Rim Conference on Artificial Intelligence (PRICAI), pages 250-261.
Matsumoto, K., Suzuki, W., and Tanaka, K. (2003). Neuronal correlates of goal-based motor selection in the prefrontal cortex. Science, 301, 229-232.
Neumegen, T. (2013). A computational platform for simulating reach-to-grasp actions: modelling physics, touch receptors and motor control mechanisms. MSc thesis, Dept of Computer Science, University of Otago.
Oztop, E., Bradley, N., and Arbib, M. (2004). Infant grasp learning: a computational model. Experimental Brain Research, 158, 480-503.
Prinz, W. (1997). Perception and action planning. European Journal of Cognitive Psychology, 9(2), 129-154.
Simon, J. (1969). Reactions towards the source of stimulation. Journal of Experimental Psychology, 81, 174-176.
Takac, M., Benuskova, L., and Knott, A. (2012). Mapping sensorimotor sequences to word sequences: A connectionist model of language acquisition and sentence generation. Cognition, 125, 288-308.
Umiltà, M., Escola, L., Intskirveli, I., Grammont, F., Rochat, M., Caruana, F., Jezzini, A., Gallese, V., and Rizzolatti, G. (2008). When pliers become fingers in the monkey motor system. PNAS, 105(6), 2209-2213.

# Coordinate-free Thinking 

Joop Leo (joop.leo@phil.uu.nl)<br>Department of Philosophy, Utrecht University, P.O. Box 80126<br>3508 TC Utrecht, The Netherlands


#### Abstract

Thinking outside the box is a nice metaphor, but we are so used to putting things in boxes that it may seem essential to our way of thinking. This idea is enforced by the standard view of relations that says that relata always come in a certain order. There is, however, an alternative view on relations, the antipositionalist view, according to which the constituents of the complexes of a relation neither come in a certain order, nor do they occupy positions. Instead, a relation is conceived as a network of interrelated complexes. We show that this view facilitates a coordinate-free way of thinking and that it may thus have a heuristic value.


Keywords: Antipositionalism; relational complex; relation; substitution.

## Introduction

Consider the following state 'out there':


According to the standard view on relations, we have two distinct complexes here: the cat's being on top of the mat and the mat's being under the cat. So, in the first complex the cat comes first and in the second complex the mat comes first. However, in the state itself, there is no such order. This makes the standard view weak.

A better view on relations is the positionalist view. According to this view, each relation comes with a number of positions to which objects can be assigned. The state above can be expressed in a neutral way by assigning the cat to the position Top and the mat to the position Bottom. But for symmetric relations like the adjacency relation, we would get two indistinguishable complexes for a single state. ${ }^{1}$ This makes also this view objectionable.

Kit Fine developed a radically different view on relations, the so-called antipositionalist view (Fine, 2000). In this view,

[^328]the objects of a complex neither come in a certain order, nor are objects assigned to positions. Instead, a relation consist of a network of the complexes interrelated by substitutions. We may illustrate this as follows.


If we substitute in the complex of the cat on the mat a dog for the cat and a table for the mat, we get the complex of the dog on the table. Furthermore, substituting in the last complex a caterpillar for the dog and a mushroom for the table gives the same result as substituting in the original complex the caterpillar for the cat and the mushroom for the mat.

Since the antipositionalist view rejects order and positions as fundamental, it is in fact a coordinate-free view on relations.

A detailed comparison of mathematical models for the different views on relations provide strong support for the claim that antipositionalism is the superior view on relations (Leo, 2008, 2010, 2013b). Nevertheless, the view does not seem to be in line with our ordinary way of thinking and of expressing ourselves. Our natural languages are linear and in many languages we exploit the linear order to our advantage to express factual situations concisely. We say, for example, "the cat is on the mat" and we know that by mentioning the cat first that he or she is on top. We may, however, be misled into thinking that the order is essential for the underlying relation.

This misconception is reinforced by standard logic, which also imposes an order that is not present in the represented states 'out there' in reality. In fact, standard logic functions like a distorting mirror. It does not faithfully represent reality. However, of an impeccable logic we expect that it can represent reality in a very pure and natural way. Only, such a logic does not yet exist.

In this paper a sketch will be given of a new logic of relations that matches well with a coordinate-free way of thinking. Furthermore some ideas will be given for utilization of the new logic.

## A coordinate-free view of the world

Suppose you had to make clear that a block $a$ is on top of another block $b$, but that you were not allowed to use expressions in which the order $a$ and $b$ are mentioned is relevant for
the meaning, and also that you are not allowed to use positions like Top and Bottom. In other words, you would have to give a coordinate-free account of the state.

Then what you could do is point at a situation where $a$ is indeed on top of $b$. Now there is a big chance that someone else will not directly know what kind of relationship you mean between $a$ and $b$. It might, for example, be thought that you wanted to express that $a$ and $b$ are close to each other. But it would help if you would point at a lot of other situations and in case there is a vertical placement between two objects you would say "This is another state of the same relation that can be obtained from the original state by substituting this object for $a$ and that object for $b$, ," and for dissimilar situations you would say "This state is not of the same relation."

The relation could in this way be grasped without using a specific order for the objects and also without using positions. In this setting, we understand the relation by explicitly using the notion of substitution. A very interesting question is how much of the world we could understand in terms of substitutions.

I consider substitution as a primitive kind of operation. We have, for example, a clear understanding of what it means to substitute Romeo for Adam in Adam's loving Eve. What is less clear, however, is how important the notion of substitution is for our understanding of the world.

When you think about Adam's loving Eve and of Romeo's loving Juliet, you probably do not think explicitly in terms of substitutions. Rather, you think in terms of the love relation applied in a certain way to the persons involved. This is in line with the standard view and the positionalist view on relations, which considers a relation as something with 'holes' in which things can be put.

As I argued before, the standard view and the positionalist view on relations are wrong. ${ }^{2}$ But this does not mean that the views are also harmful. In Leo (2010), an explicit justification was given for using positional representations. Moreover, positional representations are very practical.

Nevertheless, realizing that on a fundamental level there are no positions in relations is a liberating thought. It opens the door to explore new ways of learning and of looking at the world.

Because substitution may be an essential ingredient for any truly intelligent system, the insights are likely also relevant for the field of artificial intelligence. One of the challenges will be to investigate whether and how a general notion of substitution can be implemented in artificial systems.

## Developing a logic of relations

The motivation for developing a logic of relations is that we like to have a formal framework that captures the essence of 'real' relations. In this section, a short description is given of a new logic in which we can reason about relations in a coordinate-free way. Although the technical details are kept

[^329]to a minimum, some basic knowledge of predicate logic will be assumed.

Let us start with the love relation. In predicate logic this relation is represented by a binary predicate symbol, say $L$, and the state of Adam's loving Eve is represented by a formula like $L(a, e)$. The order of the arguments $a$ and $e$ play a role for the interpretation. But it makes no sense to say that in the state itself Adam comes first. The key question is: How can we get rid of the order?

## The main idea

As in predicate logic, the new logic also has terms and formulas, where the terms represent entities in the world (or in our domain of discourse) and the formulas represent assertions about these entities. For example, a term may represent Adam and a formula may represent the assertion that Adam loves Eve.

To prevent problems with the order, we will in our new logic not accept terms $F\left(x_{1}, \ldots, x_{n}\right)$ and formulas $P\left(x_{1}, \ldots, x_{n}\right)$ for any $n \geq 2$. At the same time we do not want to loose anything of the expressive power of predicate logic. Fortunately this is possible.

The main idea for a new logic of relations is to use terms to represent not only individuals, but also all kinds of complex entities 'out there' and to build formulas from the terms only with equality and normal logical connectives and quantifiers. So, we will get rid of predicates altogether-with the exception of equality.

The terms may, for example, not only represent persons like Adam and Eve, but also complexes like Adam's loving Eve and they may even represent substitutions like the one from Adam's loving Eve to Romeo's loving Juliet. Formally, for these entities the terms might look like:

$$
a, e, L_{a e}, s
$$

In addition, we have terms like $\operatorname{src}(s)$, representing the source of the substitution $s$, i.e. the complex of Adam's loving Eve, $\operatorname{tgt}(s)$, representing its target, i.e. Romeo's loving Juliet, and $s(r)$, representing the substitution of Adam by Romeo in the original complex.

As we said, the formulas represent assertions about the terms. To express in a formula that Adam loves Eve we could say

$$
L_{a e}=L_{a e}
$$

This might look as something that is trivially true, but this is not the case since we do not assume that all terms have an interpretation. It is similar to what we have in natural languages with non-referring terms like 'Vulcan', 'ether', 'Santa Claus', and ' 5 loves Eve', and in arithmetic with terms like ' $\frac{1}{0}$ '. The way equations are interpreted in our logic guarantees that if $t=t^{\prime}$, then both $t$ and $t^{\prime}$ have an interpretation. In other words, for an equality assertion to be true, the existence of the interpretation of the terms is required.

Because we need existence assertions so often, we abbreviate the formula above as

$$
\mathrm{E}!L_{a e}
$$

Now if we would like to express that Eve loves Adam as well, then this might require a much larger formula involving src, tgt, $=$, logical connectives and quantifiers. For this reason, it is convenient to introduce abbreviations like:

$$
E!L_{a e}[a \mapsto e, e \mapsto a]
$$

Here $L_{a e}[a \mapsto e, e \mapsto a]$ stands for the result of substituting $e$ for $a$ and $a$ for $e$ in the complex $L_{a e}$.

What I presented here is only an example. It is just to give an impression of the logic of relations. In Leo (2013a) a more detailed and formal description of the syntax and semantics of the logic is given.

## Axiomatization

For the logic of relations, a rather straightforward axiomatization can be given. Here, we do not give the more general axioms, but only axioms that say something specific about complexes and substitutions.

The logic has two constants:
src, tgt

The constants will be interpreted as partial functions that give the source and the target of a substitution.

For convenience, we give a few definitions:

$$
\begin{array}{ll}
\mathrm{E}!t: & t=t \\
t \simeq t^{\prime}: & \mathrm{E}!t \vee \mathrm{E}!t^{\prime} \rightarrow t=t^{\prime} \\
t \text { is a complex : } & \exists s(t=\operatorname{src}(s)) \\
t \text { in } t^{\prime}: & \exists s\left(\operatorname{src}(s)=t^{\prime} \wedge \mathrm{E}!s(t)\right)
\end{array}
$$

The first formula, E! $t$, says that (the interpretation of) the term $t$ is defined. The formula $t \simeq t^{\prime}$ expresses weak equality between the terms $t$ and $t^{\prime}$, i.e. if either $t$ or $t^{\prime}$ is defined, then so is the other and their contents are the same. The formula ' $t$ is a complex' is true if and only if $t$ is the source of a substitution. And the last formula, $t$ in $t^{\prime}$, expresses that $t$ is an object that belongs to $t^{\prime}$, which is the case if $t^{\prime}$ is the source of a substitution for which $s(t)$ is defined.

The axioms are as follows :

## Source and target axioms

Any substitution has a source and a target:

$$
\forall s(\mathrm{E}!\operatorname{src}(s) \leftrightarrow \mathrm{E}!\operatorname{tgt}(s))
$$

The target of a substitution is a complex as well:

$$
\forall s(\mathrm{E}!\operatorname{tgt}(s) \rightarrow \operatorname{tgt}(s) \text { is a complex })
$$

## Constituents axiom

For any complex, all substitutions are defined for the same objects:

$$
\forall x \forall s(x \text { in } \operatorname{src}(s) \rightarrow \mathrm{E}!s(x))
$$

## Extensionality of substitutions axiom

A substitution is uniquely determined by what objects are substituted for the constituents of a complex:

$$
\forall s, s^{\prime}\left(\operatorname{src}(s)=\operatorname{src}\left(s^{\prime}\right) \wedge \forall u\left(s(u) \simeq s^{\prime}(u)\right) \rightarrow s=s^{\prime}\right)
$$

## Identity of substitutions axiom

Substituting for each object of a complex the same object results in the same complex:

$$
\begin{aligned}
& \forall x(x \text { is a complex } \rightarrow \exists s(\operatorname{src}(s)=x \wedge \operatorname{tgt}(s)=x \wedge \\
& \forall u(u \text { in } x \rightarrow s(u)=u))
\end{aligned}
$$

## Composition of substitutions axiom

Substitutions can be composed like partial functions:

$$
\begin{aligned}
\forall s, s^{\prime}(\operatorname{tgt}(s) & =\operatorname{src}\left(s^{\prime}\right) \rightarrow \exists s^{\prime \prime}\left(\operatorname{src}\left(s^{\prime \prime}\right)=\operatorname{src}(s) \wedge\right. \\
& \left.\left.\operatorname{tgt}\left(s^{\prime \prime}\right)=\operatorname{tgt}\left(s^{\prime}\right) \wedge \forall u\left(s^{\prime \prime}(u) \simeq s^{\prime}(s(u))\right)\right)\right)
\end{aligned}
$$

Furthermore, to make deductions we use modus ponens as our single rule of inference:

$$
\text { from } \varphi \text { and } \varphi \rightarrow \psi, \text { infer } \psi
$$

With these axioms and this rule of inference we have a powerful system to reason about all kinds of relational structures in a natural way.

## Advantages of the logic of relations

With this design of the new logic we get the same expressive power as that of first-order predicate logic. In addition, it has some significant advantages compared to predicate logic:

1. We got rid of the artificial order of the arguments of a relation. The logic allows us to express relations in a neutral way. This applies not only to everyday relations like the love relation, but to mathematical relations as well. We speak of the less-than relation and the greater-than relation, but it would be more natural and correct to say that there is just a single strict ordering relation. In the logic of relations this single relation can be formulated in a convincing way.
2. The logic of relations seems conceptually simpler than predicate logic. The formulas only make use of terms, logical connectives, quantifiers and equality. Predicate symbols do not occur at all-except equality. I consider this a significant advantage. The logic allows us to talk in a purely 'logical' way about the things that are 'out there'.
3. Substitution-the basic operation of the new logic-is a very intuitive notion. In my view it is more elementary than assuming that arguments of a relation come in a specific order or that relations have fixed positions to which arguments can be assigned.
4. In the logic of relations, certain complex properties of objects can be expressed very concisely. For example, that objects $a$ and $b$ have exactly the same relations can be expressed as

$$
\forall x(x \text { is a complex } \rightarrow \mathrm{E}!x[a \mapsto b, b \mapsto a])
$$

In predicate logic we need in some cases an infinite number of formulas for this.
5. In principle, the logic of relations allow complexes to have any number of objects, including infinitely many. This is not the case for normal predicate logic.

What might seem a disadvantage of the logic of relations is that the formulas can be relatively long. However, using abbreviations like $t\left[u_{1} \mapsto v_{1}, \ldots, u_{n} \mapsto v_{n}\right]$ may solve this problem. We might even go a step further and let certain formulas look exactly like formulas of ordinary predicate logic, for example by writing $L(x, y)$ for $E!L_{a e}[a \mapsto x, e \mapsto y]$.

In conclusion, the logic of relations has the potential to represent reality more adequately than predicate logic.

## Learning relations

The new logic of relations may influence the way we look at the world. But how do we normally 'learn' relations? Is it by substitution, by abstraction, by positional representations or via processes with a completely different logic? And what is the role of language in learning relations?

A psychological investigation of the way we learn relations would be extremely useful. It might, however, not be easy to develop experiments to determine how we learn all kinds of relations. It will require experts in cognitive psychology to design appropriate tests and experiments.

And there are also more questions to be asked, for example, how small children and animals learn relations. Answers to these questions might deepen our insight in fundamental aspects of the way we understand and represent the world. In addition, they might suggest new learning programs.

Some theoretical research has been done in this field. For example, in Tomlinson and Love (2006) a model of relational learning has been developed. But as far as I know, the role that the notion of substitution may play in learning relations has never been explicitly considered. A more explicit investigation could fit in nicely with the research field of the way we reason. Interesting research in this more general field has been done in Johnson-Laird (1983); Evans, Newstead, and Byrne (1993); Gentner and Smith (2012). In particular, work on analogical reasoning seems to be most relevant.

A related question is what is the best way for artificial intelligence systems to learn relations. There is encouraging re-
search with respect to developing algorithms for learning relations (Richards \& Mooney, 1992; Heyer, Läuter, Quasthoff, Wittig, \& Wolff, 2001; Katrenko \& Adriaans, 2007). In some cases the goal is to build systems that 'discover' relations, and in other cases to find instances for which a given relation holds.

As we saw, the basic operation of the new logic is substitution. If we will be able to implement a general notion of substitution in an AI system, then it might perhaps be possible for such a system to learn a variety of relations by feeding it large sets of samples. As a result we might get sophisticated systems that are able to discover all kinds of subtle regularities in the world.

## Impact of a coordinate-free view

The introduction of a coordinate system in the 17 th century by René Descartes marked a major step forward for mathematics and physics. The idea is quite simple: to each point in the plane a pair of numbers is assigned. This made it possible to describe geometric shapes by algebraic equations. However, in the twentieth century coordinate-free treatments of certain geometric topics turned out to be simpler and more elegant. In particular this is the case for vector analysis and differential geometry.

The choice of a particular coordinate system often turns out to be irrelevant. In physics, this may have an underlying reason; around 1910, Albert Einstein introduced the principle of general covariance, according to which the basic laws of physics remain invariant under changes in frames of reference. From this, one should not immediately conclude that a completely coordinate-free formulation of the laws is always possible. It would also be misleading to call the coordinates used in physics artefacts.

For relations the situation is different. According to the antipositionalist view-the superior view on relations-a relation has on a fundamental level no positions and no order for the relata. The view is genuinely coordinate-free. We do not have the problem of having to deal with irrelevant details like choosing an order, since such details are simply lacking.

This observation may not immediately have an effect on how we perceive the world around us. However, I expect an impact in the longer term:

1. Becoming more acquainted with the new logic of relations may make us more familiar with the idea that relations around us are indeed networks of interrelated complexes. This may help us to 'discover' new relations and instances of relations in an easier way.
2. An interesting application of the new logic presented in this paper will be the development of a philosophically driven alternative for set theory-the standard foundation of mathematics. There is a substantial need for this, since we do not live in a world of sets, but in a world of things with relations between them. Although almost everything
can be coded in set theory, the coding is in some cases quite artificial. ${ }^{3}$
3. The new logic may be useful for developing a new programming language in which complexes and substitutions play a central role. Programs written in such complexoriented programming languages may have a simpler internal structure. This would be of great interest.
4. Finally, I can imagine that some day an effect of this research may be found in websites, shops, airports, and cities: innovations in the design of such places may be inspired by what is for us the most logical way to relate things.

It will be obvious that to accomplish the goals mentioned above, more innovative and interdisciplinary research needs to be done-in particular by computer scientists, linguists, logicians, philosophers, and psychologists. But the results so far are promising.

## Acknowledgments

This research was supported by a VENI grant from the Netherlands Organization for Scientific Research (NWO). I thank Niels van Miltenburg, Vincent van Oostrom, and Albert Visser for their comments on the paper.

## References

Evans, J., Newstead, S., \& Byrne, R. (1993). Human reasoning: The psychology of deduction. Hove: Lawrence Erlbaum Associates Ltd.
Fine, K. (2000). Neutral relations. The Philosophical Review, 109, 1-33.
Gentner, D., \& Smith, L. (2012). Analogical reasoning. In V. S. Ramachandran (Ed.), Encyclopedia of human behavior (2nd ed.) (pp. 130-136). Oxford: Elsevier.
Heyer, G., Läuter, M., Quasthoff, U., Wittig, T., \& Wolff, C. (2001). Learning relations using collocations. In Proceedings of the ijcai workshop on ontology learning.
Johnson-Laird, P. (1983). Mental models. Cambridge: Cambridge University Press.
Katrenko, S., \& Adriaans, P. (2007). Learning relations from biomedical corpora using dependency tree levels. In K. Tuyls (Ed.), Lecture notes in computer science (pp. 6180). Berlin: Springer.

Leo, J. (2008). Modeling relations. Journal of Philosophical Logic, 37, 353-385.
Leo, J. (2010). The logical structure of relations. Doctoral dissertation, Department of Philosophy, Utrecht University.
Leo, J. (2013a). A logic of natural relations (Logic Group Preprint Series No. 307). Utrecht: Department of Philosophy, Utrecht University.

[^330]Leo, J. (2013b). Relational complexes. Journal of Philosophical Logic, 42, 357-390.
Richards, B., \& Mooney, R. (1992). Learning relations by pathfinding. In Proceedings of the tenth national conference on artificial intelligence (pp. 50-55). San Jose, CA.
Tomlinson, M., \& Love, B. (2006). Learning abstract relations through analogy to concrete exemplars. In Proceedings of the cognitive science society (pp. 2269-2274). Vancouver, Canada.

# Categorization and Abstract Similarity in Chess 

Pablo León-Villagrá (pleonvil@uos.de) Frank Jäkel (fjaekel@uos.de)<br>Institute of Cognitive Science, University of Osnabrück<br>49069 Osnabrück, Germany


#### Abstract

Chess experts remember meaningful chess positions better than novices (de Groot, 1978; Chase \& Simon, 1973). This can be explained with a larger number of chunks in experts' long-term memory (Gobet \& Simon, 1998). These chunks are mainly based on visual representations-that is, pieces on squares. However, a recent experiment highlighted that experts prefer to group chess positions by abstract similarities that cannot be explained purely visually (Linhares \& Brum, 2007). Based on these data it was claimed that chess expertise, in addition to chunks, crucially relies on abstraction and analogies. These data and the conclusions were heavily criticized because the instructions strongly biased the participants to group positions in a certain way (Bilalić \& Gobet, 2009). Here, we successfully replicated this experiment with less explicit instructions. In addition, we collected category labels for the groupings that allowed us to explore the abstract principles that participants used.


Keywords: Analogy, Categorization, Chess, Expertise, Pattern Recognition, Representations, Similarity

## Introduction

After a match strong chess players often comment that aspects of their game were similar to well-known classical games. For example, after his win against Aronian in January 2013 world champion Viswanathan Anand stated at the press conference: "It was the same concept [...], Rubinstein's version was even Rook takes c3 and Rook to h3, but essentially [it was] the same idea [...]." Or take another example, Rosentalis commented on one of his games: "When playing Qa3 the game Smyslov-Reshevsky came to my mind" (Rowson, 2001). The left panel of Figure 1 shows the position in Rosentalis' game which made him remember the position from Smyslov-Reshevsky (right panel). The two positions share no obvious visual similarity and differ considerably with regard to the pieces and their arrangement on the board. Nevertheless, Rosentalis perceived both positions to share some crucial aspects and based on this similarity he considered the move Qa3 (which allowed an exchange of queens). How do chess players represent chess positions and what kind of similarity do expert chess players perceive in positions that are visually very different?

The classical conception of expertise in chess is based on the idea that finding the right move is a process of recognition and association (Gobet \& Simon, 1998). There are convincing data indicating that experienced chess players have access to a large database of stored patterns, called chunks, and these chunks are associated with plausible plans and ideas (de Groot, 1978; Chase


Figure 1: The left panel shows the game RosentalisAppel before white played Qa3. The right panel shows the game Smyslov-Reshevsky, World Championship 1948. Rosentalis commented that his game was similar to Smyslov-Reshevsky although there are few obvious visual similarities (Rowson, 2001).
\& Simon, 1973). A chunk in chess is, hence, defined as a unit of information in long-term memory containing a meaningful grouping of pieces on squares, plus associated moves and ideas. Each chunk consists of up to five pieces and the size of the stored chunks is positively correlated with skill. Furthermore, under the assumption that experts and novices can both retain $7 \pm 2$ chunks in short-term memory, more skilled players can make better use of their short-term memory because they have the right chunks available. Hence, differences in skill are, to a great part, based on differences in the number and the size of the chunks stored in long-term memory. In order to accommodate various findings that were inconsistent with the original chunking theory the concept of a chunk was later expanded to more complex structures, so-called templates. Templates are formed if positions reoccur frequently and in addition to the template core (which is a classical chunk) can contain free variables (Gobet \& Simon, 1996, 1998). Even though this notion expands classical chunking, in actual implementations of the theory templates are still accessed via discrimination nets and thus patterns of specific pieces on squares are fundamental for recognition.

In the anecdotal examples mentioned above strong players did not seem to rely on purely visual information, such as pieces on squares, to retrieve relevant information from memory. Linhares (2008) has argued forcefully that chess research should not focus too strongly on the visual aspects of the game. According to him, although visual similarities between stored chunks and
presented positions surely play a role in chess expertise, abstract-level relations and analogies are more important for understanding expert performance (Linhares \& Brum, 2007). Chess experts excel because their representations are on a high level of abstraction and these representations are not explainable by chunking alone.

In other areas of expertise it is well established that experts rely on abstract representations while novices concentrate on superficial aspects. Chi, Feltovich, and Glaser (1981) asked experts and novices to sort physics problems into groups. They found that the novices grouped problems based on superficial similarities (problems with pulleys vs problems with springs) whereas experts grouped problems based on physical, non-obvious principles (conservation of energy vs conservation of momentum). Inspired by this work, Linhares and Brum (2007) constructed a set of chess positions that could be grouped either by superficial, visual similarity or based on abstract principles. In their experiment they showed subjects 20 positions that formed 10 pairs based on 10 abstract ideas, like "material gain due to a double attack" or "endgame with bishops of the same color." These pairs are fairly natural as they consist of wellestablished categories in chess. Importantly, they constructed the material in a way that there were also 5 obvious pairings based on purely visual similarity. That is, the pieces and their respective positions on the board were extremely similar. But these 5 pairs were strategically or tactically very different situations due to small, but crucial, differences. Figure 2 shows an example of positions that can be grouped either visually or abstractly.

Linhares and Brum (2007) then asked chess players of varying strengths, from relatively strong masters to unrated amateurs, to pair their 20 chess positions based on strategical similarity. The expert players almost exclusively grouped the positions into abstract pairs while the novices only matched about half of the abstract pairs. Almost no visual pairs were chosen by the experts whereas novices often paired by visual similarity. Linhares and Brum concluded that multiple levels of encoding of chess positions exist, from surface representations of concrete piece relations to abstract semantic or conceptual representations consisting of abstract roles of pieces. Expert chess players perceive positions as global semantic arrangements and associate them with future developments and plans. Therefore, what differentiates experts and novices is the level of abstraction at which positions can be represented.

Bilalić and Gobet (2009) reproduced the study by Linhares and Brum introducing a condition in which participants were not asked to pair positions based on abstract similarity but on visual similarity. They did this because in the original study the instructions explicitly encouraged grouping by abstract similarity and discour-


Figure 2: Four of the positions presented in the experiment by Linhares and Brum (2007). While positions 14 and 15 and positions 17 and 18 are visually almost identical they differ considerably in their abstract essence. On the other hand, positions 14 and 17 and positions 15 and 18 are very similar on an abstract level. Positions 14 and 17 are examples of endgames with opposite-colored bishops in which no side can make progress as the opponents pawns are fixed on the wrong color. In contrast, positions 15 and 18 consist of endgames with bishops of the same color and are easily winning for white because the black pawns are attackable. For example, in position 18 white can immediately win the pawn on d6 and proceed to win the game.
aged grouping by visual similarity. They argued that this is a big methodological flaw of the original study and wanted to demonstrate that the explicit instructions simply biased the subjects to respond as the experimenter wished. In the abstract-similarity condition Bilalić and Gobet could replicate the original resultsexperts paired more than twice as many abstract pairs as novices. In the visual-similarity condition, in which players were instructed to group positions together based on visual similarity, both groups matched an equally low number of abstract pairs and a high number of visual pairs. The point that Bilalić and Gobet wanted to make with this experiment is that experts will group the material in any way they are instructed to. Hence, the strong conclusions that Linhares and Brum drew from their data are not warranted. Bilalić and Gobet argued
that the original experiment did not show any evidence that analogical reasoning or abstract similarity play an important role in chess expertise. According to them, experts did not group abstract pairs because they thought that was the natural thing to do but because they were explicitly instructed to do so.

Linhares and Brum (2009) responded to this criticism by stating that experts can of course behave like novices if told to do so. But there is an important asymmetry: "Novices cannot behave as experts." (Linhares \& Brum, 2009, p. 750) The original experiment was meant to demonstrate that experts can group the stimuli by abstract similarity whereas the novices have to rely on superficial similarity, just like in the study by Chi et al. (1981). We agree with this observation but still think that Bilalić and Gobet had a valid point. The original experiment just shows that expert chess players can group by abstract similarity if told to do so, but this in itself does not show that noticing abstract similarities and making analogies is as crucial for chess expertise as Linhares and Brum claim. As participants were explicitly discouraged to pair by visual similarity we don't know whether expert players considered visual similarity relevant at all-however unlikely this may seem. Visual similarity, in any case, might still play the dominant role in memory retrieval during a game (despite the anecdotal evidence mentioned in the introduction). We think that even if the link between real-world expert performance and subjects' pairings in the experiment by Linhares and Brum is unclear, it is still interesting to try and directly probe experts and novices for their intuitions about the similarity of positions. This was also the first step in the work of Chi et al. (1981). But similarity is a difficult notion. Unless the "respects for similarity" (Medin, Goldstone, \& Gentner, 1993) are precisely specified neither subject nor experimenter can be sure about what is meant by "similarity." The pairing experiment was cleverly designed to compare abstract and visual similarity against each other and thus the material probably biased the subjects to focus on these two aspects. In an actual chess game there might be even more respects for similarity than the two (abstract vs visual) that Linhares and Brum had in mind for their pairs. However, even in their experiment participants might perceive several notions of similarity in conflict with each other-but because of the clear instructions they use the one that was intended by the experimenters.

The present paper tried to replicate Linhares' and Brum's study once again. In our experiment less explicit instructions about the nature of pairs were provided. Therefore, every participant was able to pair the positions based on his or her individual, intuitive understanding of similarity in chess. We think that this is the "missing condition" in the debate. As we didn't give the subjects any obvious instructions on what we meant
by similarity, potentially there could be other notions of similarity that participants might consider relevant or in conflict with the abstract similarities that Linhares and Brum had constructed. Visual surface similarity could be one of them - but not the only one. Hence, if participants, even under our fuzzy instructions, grouped the stimuli mainly as predicted by Linhares and Brum this would be somewhat stronger evidence for their claims. But if the experts spontaneously grouped by visual similarity or in any other way this would show that Bilalic's and Gobet's methodological concerns are indeed important. In addition, we decided to go beyond Linhares' and Brum's original study by also asking participants to provide a category label for each pairing. In the study by Chi et al. (1981) the category labels proved to be very helpful for understanding the difference in the representations of experts and novices. These category labels will allow us to see more directly what the participants deemed relevant for the task.

## Methods

The design was based to a large extent on the original experiment and the same set of stimuli was used (Linhares \& Brum, 2007). In contrast to the original experiment, the present study changed the instruction given to the participants and permitted a more differentiated evaluation of the positions in the orientation phase.

## Participants

32 participants were recruited at local chess clubs, of which two participants aborted the experiment due to fatigue. The remaining 30 participants were all at least familiar with the rules of chess and basic strategies. The participants' skill is reflected in their DWZ rating. The DWZ rating system is an adaptation of the Elo rating system for the German national chess federation. Like the Elo scale the DWZ rating allows a fine differentiation of skill based on the players performance at chess tournaments. The mean DWZ rating was 1395.5 ( $\mathrm{SD}=$ $750.3, \min =0, \max =2461$ ), 5 players had no official rating. On average each player had performed about 2-3 hours of chess-practice per week in the last year. The mean age was 32.7 years ( $\mathrm{SD}=15.5$, min $=12$, max $=74$ ). Only two participants were female and both female participants were unrated. The participants were divided into an expert and a novice group according to the mean playing strength of all registered German players $(\mathrm{M}=1518)$. The expert group was composed of 16 participants with a mean DWZ of $1945.9(\mathrm{SD}=268.2$, $\min =1569, \max =2461$ ). The novice group had 14 members, with a mean DWZ of 766.4 ( $\mathrm{SD}=611.3$, min $=0, \max =1490)$. Splitting the participants in this way is not ideal since both groups now contain players that are around the German average. This will blur the differences between novices and experts, making it harder to find an effect if there is one. However, this
grouping is consistent with the grouping that was used in previous studies and allows for a better comparison with these studies (Bilalić \& Gobet, 2009; Linhares \& Brum, 2007). In addition we also calculated and report correlations here.

## Procedure

At the beginning of the experiment verbal instructions explaining the basic procedure of the survey were provided. The participants received a form of consent and after signing it proceeded to the first part of the experiment. Each participant received a questionnaire that asked for gender, age, and chess rating (DWZ). The final question asked for the amount of chess practice the subject had performed per week in the last year (Scale: 0-1 hours per week, 1-2 hours per week, 2-3 hours per week, 3-4 hours per week, more than 4 hours per week).

Familiarization Phase After these general questions the main experiment started. Participants received a short instruction on how to perform the survey. The twenty chess positions were presented in random order. The task consisted in giving an evaluation of the position and checking the particular response ("White has a winning advantage", "White has a minor advantage", "The position is equal", "Black has a winning advantage", "Black has a minor advantage", "No evaluation is possible"). For each position white was to move and unlimited time was granted to perform the task. Nevertheless, the participants were instructed to perform like in an over-the-board chess game and to use a reasonable time investment per position. After evaluating a position participants had to write down the next move for white. The familiarization phase served the purpose of activating, for each stimulus, the representations that would also be relevant in an actual game.
Pairing Phase The task in the second part of the experiment was to group the twenty chess positions from the familiarization phase into pairs. Participants received overview-sheets in which the twenty positions were displayed. The sheets were placed in front of the participants so that they could inspect all position simultaneously. There were three types of overview-sheets with different random arrangements of the positions. Positions were labeled with successive numbers (from 1 to 20) and each participant received one type of randomized overview-sheet. Instructions were to pair together positions which "intuitively seemed similar." Additionally, participants should perform "as if they were thinking about similar positions in a chess game." It can be argued that this additional instruction biased participants more than necessary. However, we decided to include it so that subjects understand that they should use the representations that they would use in an actual game.

Participants had to fill the position pairs into a designated table. Several of the expert players asked for
clarification of the instructions, indicating that our instruction was indeed vague, as intended. They asked in what regard they were to interpret the similarity and were told that this was up to them and they should follow their intuitions.

Labels and Features The final part of the survey consisted of ten sheets with questions about the chosen pairs. Participants were instructed to use the overviewsheet and their response table as an aid. First of all, participants were asked to name a headline or topic for each chosen pair. After that, participants had to name attributes or features which made them select the pair. Finally, they had to rate the similarity of both positions and the prototypicality for each of the two positions for the chosen headline. The ratings are not easily interpretable as the labels were highly idiosyncratic and we usually did not have enough identical pairings for a statistical analysis. We still performed an exploratory and qualitative analysis (which had unclear results) but omit it here due to space limitations.

## Stimulus Material

The twenty chess positions in the present study were the same as in the original study (Linhares \& Brum, 2007). The positions were constructed in a way that 10 abstract pairs (i.e., all 20 positions) could be selected. At the same time, five control pairs (i.e., 10 positions out of 20) consisted of visually almost identical positions. The positions with a high visual similarity were very different on an abstract level, while the abstract pairs had no similarity on a visual level. An example is shown in Figure 2. All positions displayed a middlegame or endgame position and in all positions white was to move. Most positions were relatively easy to solve.

## Results

## Pairings and Evaluations

The two groups differed significantly in the number of abstract and control pairs that they chose. The expert group chose a significantly higher number of abstract pairs ( $\mathrm{M}=5.5, \mathrm{SE}=.62$ ) than the novice group ( M $=3.1, \mathrm{SE}=.54, t(28)=-2.9, p=.007)$. The novice group matched significantly more visual pairs ( $\mathrm{M}=3.4$, $\mathrm{SE}=.37)$ than the expert group $(\mathrm{M}=1.7, \mathrm{SE}=.38$, $t(28)=3.1, p=.004)$. As mentioned above, grouping participants in this way is not ideal, therefore we also calculated the correlation between DWZ rating and the choices, excluding the five unrated participants. DWZ rating correlated positively with the number of abstract pairs, $r=.52, p=.007$ and negatively with the number of control pairs, $r=-.49, p=.014$ (Figure 3). Not surprisingly, there was also a positive relationship between rating and correct evaluation of the position, $r=.77$, $p<10^{-5}$.


Figure 3: Number of chosen abstract pairs (left) and visual pairs (right) as a function of skill.

## Category Labels

In order to compare the labels given by each participant to the labels of the other participants, the first author came up with a hierarchical classification scheme for the labels based on the abstract categories given in the original design of the material (Linhares \& Brum, 2007). He then assigned the participants' labels to these classes. Some additional classes were added later based on labels that were given by several participants, for example "last pairing(s)" (the participant could not match all positions and some pairs had to be chosen randomly from the remaining material). One other unexpected label occurred relatively frequently: Situations described as deadlocked, stuck, or impenetrable were summarized in one label ("deadlocked"). This classification resulted in the hierarchical system of labels shown in Table 1.

The most common label for the novice group was "visual similarity" ( 34 out of 140 labels), followed by "last pairing" (12). This is in stark contrast to the experts who chose these labels rarely $(5,1)$. The most common labels for the expert group were "checkmate in one" (15 out of 160 labels), "endgame with opposite colored bishops" (13), "pawn endgame" (13), "passed pawn in the pawn endgame" (13) and "endgame with bishops of the same color" (12). Those labels were chosen much less frequently by the novices.

Overall, novices chose a higher number of pairs based on visual characteristics and often gave purely visual descriptions. In addition, novices often chose general levels of description ("check", "endgames" or "bishop endgame") while most experts used more specific labels (e.g., "passed pawn in the pawn endgame", "bishop endgame with bishops of the same color"). Finally, various unexpected labels were given which in many cases could not be classified using the descriptions (or more general instances of these descriptions) of the categories by Linhares and Brum (17 idiosyncratic labels in the novice group and 11 in the expert group).

## Label-based Reanalysis of Pairing

The material consisted of pairs of visually very similar positions that could also be interpreted as clear and well-established instances of abstract categories. Even

1. Visual Similarity [Novices: 34, Experts: 5]

## 2. Tactics

(a) Material Gain [N: 1, E: 3]
i. Double Attack [N: 7, E: 11]
ii. Discovered Attack [N: 2, E: 1]
(b) Check [N: 6, E: 1]
i. Checkmate [N: 4, E: 9]

- Checkmate in One [N: 5, E: 15]
- Discovered Checkmate [N: 1, E: 6]
- Smothered Checkmate [N: 2, E: 10]

3. Endgames [N: 3, E: 0]
(a) Pawn Endgame [N: 7, E: 13]
i. Pawn Chain [N: 2, E: 6]
ii. Passed Pawn [N: 2, E: 13]
iii. Opposition [N: 6, E: 10]
(b) Bishop Endgame $[\mathrm{N}: 8, \mathrm{E}: 6]$
i. Bishops of the same color $[\mathrm{N}: 1, \mathrm{E}: 12]$
ii. Bishops of different color [ $\mathrm{N}: 1, \mathrm{E}: 13$ ]
4. Other Labels
(a) Last Pairing [ $\mathrm{N}: 12, \mathrm{E}: 1]$
(b) No Label [N: 0, E: 2]
(c) Incomprehensible Label [N: 5, E: 1]
(d) Deadlocked [N: 7, E: 9]
(e) Advantageous Positions [N: 4, E: 2]
(f) Drawish Positions [N: 3, E: 0]
(g) Idiosyncratic Labels [ $\mathrm{N}: 17$, E: 11]

Table 1: Hierarchical classification scheme used for participants' category labels.
though the results in preceding studies (Linhares \& Brum, 2007; Bilalić \& Gobet, 2009) were very clear, in our study several unexpected pairings could be observed. Analysis of the labels given by the participants for particular pairings showed that unexpected abstract relations existed in the material. Therefore, several pairs originally designed as visual pairs allowed for a plausible and consistent abstract classification. One of the most striking examples of such an underlying unexpected abstract similarity was the pair consisting of positions 3 and 6 (not shown). Although designed as a visual pair, this pair was perceived as abstract by two of the strongest participants in this study, stating that both positions share similarity with the abstract concept of a fortress. Also, some abstract pairs were chosen for the wrong reasons. For example, if a participant chose an abstract pair but gave the label "last pairing without similarity" it is very improbable that the abstract pair had been chosen
based on abstract similarity.
The labeling task we used in this study allows us to reconsider for each pair whether it should be considered as an abstract pairing or not, independent of the intended pairings by Linhares and Brum. In total, about half of the labels for abstract pairs in the novice group did not contain any sort of abstract information and most explicitly stated that they were not selected due to abstract similarity. A reevaluation of the pairings chosen by the subjects based on their labels resulted in a slightly higher number of abstract pairs for the expert group but did not change the general difference considerably (M expert abstract $=6.3, \mathrm{SE}=.67, \mathrm{M}$ novice abstract $=1.8, \mathrm{SE}=$ .55). The reanalysis did increase the correlation between rating and number of abstract pairs ( $r=.66, p=.0003$ ) considerably, while it did not change the correlation between rating and the number of visual pairs $(r=-.49$, $p=.01$ ).

## Discussion

The present study replicated the results obtained by Linhares and Brum (2007) and Bilalić and Gobet (2009). On average experienced chess players chose considerably more abstract pairs than the novice group. On the other hand, novice players selected about twice as many visual pairs than the expert group.

As in our experiment no explicit instruction about the nature of expected pairs was given, the number of chosen abstract pairs was smaller than in the previous studies. The results obtained in this study weaken the methodological concerns raised by Bilalić and Gobet (2009). Experts did not simply do what they were instructed to do, but in our experiment freely chose to pair positions based on abstract similarity. Even though the material contained a considerable amount of visually almost identical positions, these possible pairings were not interpreted as relevant for the task by the experts.

As we have the category labels for the pairings as well, we have very direct evidence for what the participants deemed relevant for each pairing. The labels clearly show that novices very often spontaneously grouped by visual similarity whereas experts did not and preferred wellestablished abstract chess categories that, probably, simply weren't available to many of the novices. Furthermore, even though novices sometimes also used abstract categories, we could see that there was a tendency for experts to use better differentiated categories. This is a common finding in the study of expertise (Johnson \& Mervis, 1997).

The present study showed that the categorization behavior observed in Linhares' and Brum's study was not simply based on the instruction but was a genuine, asymmetric characteristic of expert players. However, we still agree with the conclusion of Bilalić and Gobet that "it may well be that analogy is central to expert cognition
[...]. This cannot, however, be demonstrated by asking experts to look for analogy in problems." (Bilalić \& Gobet, 2009, p. 746). Future research should therefore follow more closely the example of Chi et al. (1981). While the stimuli that Linhares and Brum (2007) used were cleverly designed to contrast abstract and visual similarity, they may have biased participants to consider mostly these two aspects. The next step should be to have subjects group a representative selection of stimuli to avoid this bias. In addition, it remains to be demonstrated beyond anecdotal evidence that abstract similarity is important for actual chess playing. One way forward could be to look for abstract categories and analogical processes in think-aloud protocols that were collected during actual games.

## References

Bilalić, M., \& Gobet, F. (2009). They do what they are told to do: The influence of instruction on (chess) expert perception - Commentary on Linhares and Brum (2007). Cognitive Science, 33(5), 743-747.

Chase, W., \& Simon, H. (1973). Perception in chess. Cognitive Psychology, 4(1), 55-81.
Chi, M., Feltovich, P., \& Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. Cognitive Science, 5(2), 121-152.
de Groot, A. (1978). Thought and choice in chess. Mouton De Gruyter.
Gobet, F., \& Simon, H. (1996). Templates in chess memory: A mechanism for recalling several boards. Cognitive Psychology, 31 (1), 1-40.
Gobet, F., \& Simon, H. (1998). Expert chess memory: Revisiting the chunking hypothesis. Memory, 6(3), 225-255.
Johnson, K., \& Mervis, C. (1997). Effects of varying levels of expertise on the basic level of categorization. Journal of Experimental Psychology: General, 126(3), 248.

Linhares, A. (2008). The emergence of choice: Decision-making and strategic thinking through analogies (Tech. Rep.). Available from http://cogprints.org/6615/
Linhares, A., \& Brum, P. (2007). Understanding our understanding of strategic scenarios: What role do chunks play? Cognitive Science, 31(6), 989-1007.
Linhares, A., \& Brum, P. (2009). How can experts see the invisible? Reply to Bilalić and Gobet. Cognitive Science, 33(5), 748-751.
Medin, D., Goldstone, R., \& Gentner, D. (1993). Respects for similarity. Psychological Review, 100(2), 254.

Rowson, J. (2001). The seven deadly chess sins. Gambit Publications.

# Prior Linguistic Knowledge Influences Implicit Language Learning 

Janny H.C. Leung (jannyleung@hku.hk)<br>School of English, The University of Hong Kong, 735 Run Run Shaw Tower, Centennial Campus, The University of Hong Kong, Hong Kong

John N. Williams (jnw12@cam.ac.uk)
Department of Theoretical and Applied Linguistics, University of Cambridge 9 West Road, Cambridge CB3 9DP, United Kingdom


#### Abstract

We report three experiments that explore the effect of prior linguistic knowledge on implicit language learning. Native speakers of English and native speakers of Cantonese participated in implicit learning (IL) experiments that involved different learning materials. In Experiment 1, both participant groups showed evidence of learning a mapping between articles and noun animacy. In Experiment 2, neither group showed learning of the mapping between articles and a linguistically anomalous concept (the number of capital letters in an English word or the number of strokes in a Chinese character). In Experiment 3, the Chinese group, but not the English group, showed evidence of learning a mapping between articles and a concept derived from the Chinese classifier system. It was concluded that first language knowledge affected implicit language learning, and that IL, at least when natural language learning is concerned, is not a completely unconstrained domain-general mechanism.


Keywords: implicit learning; form-meaning connections; vocabulary learning; second language acquisition; cross-linguistic influence

## Introduction

Traditional implicit learning (IL) (as contrasted with explicit learning, EL) research has sought to minimise the effect of prior knowledge, either by using artificial grammars (e.g., Reber, 1967) or artificial event sequences (e.g., serial reaction time experiments; e.g., Nissen \& Bullemer, 1987). Perruchet and Pacton (2006: 237) suggest that even arbitrary materials may not be "neutral" enough, for they may interact with related situational knowledge. It is typically assumed that domain general mechanisms underlie IL, but in many real life situations the learner may bring relevant domainspecific prior knowledge or dispositions, so the question arises as to whether, or how, these impact upon the IL process.

Natural language learning is a case in point - some believe that linguistic universals constrain both first and second language acquisition (Chomsky, 2006; Hawkins, 2004) and that in both cases the theoretically interesting learning processes operate at the implicit level. Even those who dispute the nature of such linguistic universals would accept that second language acquisition (SLA) is heavily influenced by first language (L1) knowledge, or L1-based processing strategies (MacWhinney, 2008; Ellis \& Sagarra, 2011). Second language (L2) learners approach SLA with existing linguistic knowledge and habits they have gathered from their first language acquisition experience. Cross-linguistic influence is well documented in the SLA literature, much of which is concerned with identifying the ways in which elements from one language get incorporated into another, accounting for errors, contrastive analysis, and interaction of transfer effects with other factors. Ellis (1994/2001: 300) argues that no theory of SLA "can be considered complete" if it ignores the learner's prior linguistic knowledge. In a similar vein, if IL is posited as an underlying mechanism of language acquisition, one must also consider whether and how the influences of prior linguistic experience on learning take place implicitly. In the SLA literature, cross-linguistic influences are sometimes thought of in terms of hypothesis testing and learner strategies (Corder, 1981; Tomasello \& Herron, 1989), implying a certain degree of intention and awareness in the process. Although it is difficult to imagine that such influences involve only explicit processes, there does not seem to be empirical effort to demonstrate such influences operating at the implicit level during learning. Moreover, cross-linguistic influence is found to be subject to general constraints such as language proficiency, sociolinguistic factors, markedness, prototypicality, language distance and psychotypology, and developmental factors (Ellis, 1994/2001). The interaction of such constraints with domain general learning mechanisms begs for research.

Our earlier work has begun to show that learning processes supporting implicit language learning
effects are not completely unconstrained. Leung and Williams (2012) used an efficiency measure (Seger 1994) to measure learning of a semi-artificial article system comprising the pseudowords gi, ul, ro, ne. Participants were told that the use of the articles depended on the distance of the object described by the accompanying noun ( $g i$ and $u l$ were used with an accompanying noun which referred to a near object; ro and ne were used if it referred to a far object). Unbeknownst to them, article use also depended on noun animacy ( gi and ro were used with animate objects and $u l$ and ne with inanimate objects). Objects were pictorially presented, along with an audio presentation of the corresponding noun phrase. The task for the participants was to respond as quickly as possible whether the noun referred to a living or a non-living thing ${ }^{1}$. Learning was measured by an increase in reaction time when the hidden regularity was violated (Violation trials) compared with reaction time in grammatical trials (Control trials). After performing 272 training trials in the task, native speakers of English who claimed to be unaware of the hidden regularity (as revealed in a standardised verbal report) nevertheless slowed down significantly in animacy decisions when the correlation between article use and noun animacy was reversed. In contrast, in a second experiment article use correlated with the relative size of objects rather than their animacy. This time there was no significant change in relative size decision times when the mapping between article use and relative size of the objects described by the nouns was reversed. The findings suggest that some meanings are more amenable to the IL of form-meaning connections than others. One possible explanation for this is the availability of grammatical processes and representations based on participants' existing linguistic knowledge, but to probe into this issue, comparisons between participants with different first languages have to be made.

The present study aims to consolidate and extend our earlier work by further exploring potential first language influences on IL of a semi-artificial grammatical system. We report below three experiments involving different learning materials and two language groups (native Chinese and native English speakers).

[^331]
## Experiment 1

## Objective

To test whether animacy, a conceptually salient feature, may be implicitly mapped onto articles by native speakers of English and Chinese.

## Participants

Thirty native speakers of English from the University of Cambridge and 27 native speakers of Cantonese Chinese from the University of Hong Kong, with a varied second language background. All Hong Kong participants spoke English as an L2, and some also knew other second/foreign languages.

## Materials and Procedure

All experimental materials were digitalized and presented with E-Prime software.

Participants were told that they would be introduced to a miniature article system from a language not known to them (Table 1). They were told that the articles ${ }^{2}$ were used to encode the distance between the speaker and the object ( $g i$ and ro for near objects and $u l$ and $n e$ for far objects). Therefore gi dog may be read as 'the near dog', ro table as 'the near table', ul mouse as 'the far mouse', and ne car as 'the far car'. Participants could spend as much time as they needed to remember the mapping between the articles and the distance system. Participants were however not told that the use of these articles also depended on the animacy of the accompanying noun ( $g i$ and $u l$ for animate objects and ro and ne for inanimate objects).

Table 1 The miniature article system in Experiment 1

| Miniature article <br> system |  | Participants not told |  |
| :---: | :---: | :---: | :---: |
|  | animate | inanimate |  |
| Participants <br> were told | near | gi | ro |
|  | far | ul | ne |

A total of 176 animate and inanimate nouns were used for the experiment, each appearing twice (once with each possible article). In the training phase, participants were exposed to visually presented noun phrases (article and noun combinations). While the same article system was used for all participants, the nouns were presented in each participant group's first

[^332]language（i．e．，English or Chinese；see table 2）．For instance，an English subject may see gi fox and a
Chinese subject may see $g i$ 狐狸（＇fox＇）on the screen． The task for the participants was first to make a decision about the animacy of the object（living or non－living，M or C key）described by the noun．The noun phrase then disappeared and was replaced by the prompt＂N／F＂．Participants had to indicate the distance meaning of the article（ $\mathrm{M}=$ near， $\mathrm{K}=\mathrm{far}$ ）． Both decisions had to be made as quickly as possible and reaction time was recorded．Response buttons were configured such that the near／far buttons were in a logical arrangement－one above the other，which also helped reduce interference with the other decision which is in a horizontal arrangement；see figure 1 for visualisation．Reaction time for the animacy decision was measured from the onset of the noun phrase；reaction time for the near／far decision was measured from the onset of the N／F（near／far） prompt，which appeared immediately after the animacy response．Feedback was provided；if participants gave a wrong response，the display did not change．Eight practice trials were provided before the experiment started．A total of 204 grammatical trials were presented in the training phase，with equal numbers of trials presented with each article．These trials were divided into four blocks，although no division between blocks was apparent to the participants．The first block consisted of 84 trials， which were made up of 84 nouns used in combination with an equal number of one of the two grammatically possible articles for each noun．Block 2 contained 60 trials，comprising 60 new nouns， again in combination with an equal number of appropriate articles．In Block 3， 28 of these nouns were repeated with a correct article of opposite distance from Block 2 （e．g．if gi pig had occurred in Block 2，then ul pig occurred in Block 3）．Within Blocks 1 to 3 the trials were divided into fixed groups of four，with each article occurring once．For each participant the order of trials within groups was randomised as was the order of groups．This procedure meant that no more than two successive trials would involve the same article．In Block 4 the remaining 32 nouns from Block 2 were repeated． Half of them occurred with an article of different distance from Block 2，but correct animacy（e．g．，if $u l$ parrot had occurred in Block 2 then gi parrot occurred in Block 4）．These were Control trials．The other half of the Block 4 trials occurred with the article of opposite distance and animacy（e．g．if ro tent had occurred in Block 2，then ul tent occurred in Block 4）．These were Violation trials．Control and Violation trials were randomly intermixed，and the nouns were rotated around conditions across subjects，
resulting in two presentation lists．Note that although nouns were repeated from Block 2 to Blocks 3 and 4， no article－noun combination was repeated throughout the experiment．After Block 4 participants filled out a questionnaire which probed awareness of the relevance of animacy to article usage，awareness of violations，and at what point in the experiment awareness of the regularity developed．Participants who reported no awareness were then encouraged to guess what factors（apart from near／far）determined when the articles were used．The whole experiment took about 45 minutes to complete．

Table 2．Sample animate and inanimate nouns used in Experiment 1

| $\mathrm{gi} / \mathrm{ul}$ |  | ro／ne |  |
| :---: | :---: | :---: | :---: |
| English <br> version | Chinese <br> version | English <br> version | Chinese <br> version |
| fox | 狐狸 | piano | 鋼琴 |
| buffalo | 水牛 | microscope | 㪚微鏡 |
| gorilla | 大猩猩 | telescope | 望遠鏡 |
| seal | 海豹 | kettle | 水壺 |



Figure 1．Configuration of response buttons on millisecond accurate keyboard／response box．

## Results

In the post－experiment debriefing，two levels of unawareness were assessed：a failure to report knowledge of the hidden regularity，and a failure to suggest any relevance of the target concept to article use when prompted to guess．Data from participants who failed to report knowledge are reported below； data from participants who failed to guess were found to show a pattern consistent with the summarized findings and are not reported here．

Outlying response times，at cutoff limits at $\pm 2.5$ standard deviations from each subject＇s mean in the Control and Violation blocks respectively were removed．In addition，data were excluded from any participant whose mean response time for the first decision over the two critical conditions was more than 2.5 standard deviations from the group mean．

Considerable variability in response times was observed，possibly because participants varied in how they distributed processing time over the two decisions and so inordinately slow participants were excluded on the basis that they may have been approaching the task in a different way from the majority．

Evidence of learning is based on the difference in reaction time across Control and Violation trials，as tested by a paired－sample t－test．Only data for the first（in this case animacy）decision are reported（no effects in any experiment were obtained for the second，near／far，decision）．

Twenty native English and 20 native Chinese did not report relevant knowledge of the regularity in response to the first debriefing question（awareness rates $33 \%$ and $26 \%$ respectively）．One Chinese participant was excluded due to slow overall animacy decision response times．Response times to make animacy decisions were significantly slower for Violation than Control trials even for participants who showed no awareness of the relevance of animacy，or related concepts，to article usage（see fig． 2，＇English animacy＇and＇Chinese animacy＇）， indicating that the animacy rule has been learned implicitly．The combination of alphabetical articles with characters is orthographically odd for the Chinese，but this did not seem to have affected their learning．


Figure 2．Reaction times（RT）in milliseconds（ms） for unaware participants in all three experiments， ＊ $\mathrm{p}<0.05,{ }^{* *} \mathrm{p}<0.01$ for difference between control and violation conditions

The question that follows is whether anyone with any language background will learn any regularity．A
linguistically unnatural learning target is adopted in the following experiment to investigate this．

## Experiment 2

## Objective

To test whether a linguistically anomalous，form－ based，distinction may be implicitly mapped onto articles by native speakers of English and Chinese．

## Participants

Twenty three native speakers of English from the University of Cambridge and 29 native speakers of Chinese from the University of Hong Kong，with a varied second language background，participated in the study．

## Materials and Procedure

The experiment shares a similar design with Experiment 1．The animacy system was now replaced by a non－semantic distinction for each participant group．In the English version，the hidden regularity was that the choice of article depended on whether the word had one capital letter or two．In the Chinese version，the hidden regularly was whether the first character in a two－character noun has more strokes than the second and vice versa，with a strokes difference between the characters being big enough that no counting would be necessary（see Table 3 below）．

Table 3．Sample nouns used in Experiment 2

| gi／ro |  | ul／ne |  |
| :---: | :---: | :---: | :---: |
| English <br> version <br> （case） | Chinese <br> version <br> （stroke） | English <br> version <br> （case） | Chinese <br> version <br> （stroke） |
| foX | 天鵝 | goRillA | 獅子 |
| piAno | 月餅 | TelesCope | 剪刀 |
| buffaLo | 牙醫 | sEAl | 學生 |
| miCroscope | 天橋 | kEttlE | 瀑布 |

The design and procedure were the same as Experiment 1 except that here participants had to indicate，as their first decision，whether the noun contained one or two capital letters（ M and Z keys respectively）or whether there were more strokes in the second or first character（ M and Z keys respectively）．The second decision indicated whether the article meant＇near＇$(\mathrm{M})$ or＇far＇$(\mathrm{K})$ ，as before．

## Results

Twenty native English and 28 native Chinese did not report relevant knowledge of the regularity in response to the first debriefing question（awareness rates $13 \%$ and $3 \%$ respectively）．One slow English unaware participant was excluded．

In both language groups，the reaction times across Control and Violation trials for unaware participants were not significantly different（see fig．2），indicating that implicit language learning in this domain does not occur for a linguistically unnatural，and non－ semantic，distinction．The next experiment examines whether learning occurs for any linguistically natural semantic distinction，regardless of whether it is reflected in the first language．

## Experiment 3

## Objective

To test whether a concept derived from the Chinese classifier system，which is not grammaticised in English，may be implicitly mapped onto articles by native speakers of English and Chinese．

## Participants

Twenty seven native speakers of English from the University of Cambridge and 32 native speakers of Chinese from the University of Hong Kong，with a varied second language background，participated in the study．

## Materials and Procedure

The learning target in this experiment is derived from a shape distinction in the Chinese classifier system． The Chinese classifier 張（zoengl in Cantonese）is generally used with thin flat objects（e．g．，sheets of paper，photos，blankets），and the counter 條（tiu4 in Cantonese）is generally used with long thin objects （e．g．，rivers，straws，ties）${ }^{3}$ ．Both classifiers frequently occur in daily usage．

The same design as the above experiments was adopted，except that the long／flat distinction became the hidden regularity governing article use in this experiment．The same article system was used with the two participant groups；nouns were presented in the first language of the participants．For example， one may find items such as gi shoelace and ro envelope in the English version，and their equivalents in the Chinese version（e．g．，gi 鞋帶 and ro 信封）．

[^333]The design and procedure were the same as in previous experiments except that here participants had to indicate，as their first decision，whether the noun referred to an object that was long（M key）or flat（Z key）．The second decision indicated whether the article meant＇near＇（M）or＇far＇（K），as before．

## Results

Twenty four native English and 26 native Chinese did not report relevant knowledge of the regularity in response to the first debriefing question（awareness rates $11 \%$ and $19 \%$ respectively）．One slow English and two slow Chinese unaware participants were excluded．A significant slowdown in Violation trials when compared with the Control trials was obtain among the unaware Chinese participants but not the unaware English participants（see fig．2），suggesting that implicit language learning is sensitive to prior linguistic knowledge．

## Discussion

One might imagine that in this paradigm participants merely learn associations between articles and patterns of keystrokes，e．g．，that＇gi＇is associated with the sequence M （living）－ M （near），or＇ne＇with Z（non－living）－K（far）．Control trials respect these patterns，but Violation trials break them，e．g．＇gi＇ would occur with the sequence Z （non－living）－ M （near）．However，if this were the case，then the nature of the categorization being performed on the noun should have made no difference whatsoever．The fact that it did suggests that the learning effects were due to learning associations between the articles and conceptual categories．

It is important to note that in all of the experiments the relevant＇hidden＇conceptual distinction had to be attended and computed in order to perform the task． But this did not guarantee that the association with the articles would be learned．Some equate statistical learning with IL（e．g．，Conway \＆Christiansen，2006）， but statistical computations should not be sensitive to the nature of the data（Perruchet \＆Pacton，2006）．

The finding that IL effects are sensitive to the nature of the concepts involved points to an interaction between the domain general learning mechanism and linguistic knowledge，which，according to many linguists，is domain specific．Semantic IL in natural language is constrained by the availability of conceptual distinctions to grammaticisation，which varies cross－linguistically．

Where no evidence of learning was obtained，it is possible that measurable learning would develop over
time. This is to say that linguistically unnatural or unfamiliar semantic categories may not be unlearnable. However, the variable amount of learning obtained after equivalent exposure shows that implicit language learning is sensitive to prior linguistic knowledge. The present study thus extends the existing literature and is congruent with studies which show that cross cultural processing biases also apply to unconscious knowledge (such as a preference for local versus global perspectives in Kiyokawa et al., 2012), and can help explain SLA studies which find L1 semantic structures in L2 processing (e.g., Jiang 2004).

We provide evidence that cross-linguistic influences may take place implicitly, and caution against a presumption that L1 transfer is based on hypothesis testing or learner strategy. It remains unclear to what degree such influences take place implicitly, or explicitly, or both implicitly and explicitly, in different SLA settings, and it seems likely that individual differences exist. A better understanding of the mechanism underlying cross-linguistic influences has obvious theoretical and pedagogical implications. Theoretically it sheds light on debates in language acquisition on domain specificity and linguistic universals; pedagogically it informs teaching/learning methodologies that aim to promote or discourage different kinds of influences.

But the kind of cross-linguistic influence that we have demonstrated may have gone beyond simple transfer. In our experiments, English participants were only sensitive to animacy - a fundamental conceptual distinction, even though it is only subtly marked in English, and there is no article-noun agreement. Chinese participants were sensitive to all semantic (but not non-semantic) distinctions, presumably through experience of their classifier system, but generalizing to novel distinctions.

Although many assume that first language acquisition is essentially implicit, it is only recently that research has shown possibilities of adult SLA taking place implicitly. Apart from further exploring the mechanisms of crosslinguistic influence, which are pertinent to SLA research, future research into implicit language learning may also inform IL research as to its nature and constraints.

## Acknowledgments

This project (code RES-000-22-3030) has been generously supported by a joint research scheme under Research Grants Council, Hong Kong, and Economic and Social Research Council, United Kingdom.

## References

Conway, C. M. and Christiansen, M. H. (2006) Statistical Learning Within and Between Modalities: Pitting Abstract Against StimulusSpecific Representations. Psychological Science, 17:10, 905-912.
Chomsky, N. (2006) Language and Mind. Cambridge University Press.
Corder, P. (1981) Error analysis and interlanguage. Oxford: Oxford University Press.
Ellis, N. C. \& Sagarra, N. (2011) Learned attention in adult language acquisition: A replication and generalization study and meta-analysis. Studies in Second Language Acquisition, 33: 4, 589-624.
Ellis, R. (1994/2001) The Study of Second Language Acquisition. Oxford, U.K.: Oxford University Press.
Hawkins, John. A. (2004) Efficiency and complexity in grammars. Oxford, U.K.: Oxford University Press.
Jiang, N. Semantic Transfer and Its Implications for Vocabulary Teaching in a Second Language. The Modern Language Journal, 88:3, 416-432.
Kiyokawa, S., Dienes, Z., Tanaka, D., Yamada, A. and Crowe, L. (2012) Cross cultural differences in unconscious knowledge. Cognition, 124: 16-24.
Leung, J. H. C. and Williams, J. N. (2012) Constraints on Implicit Learning of Grammatical Form-Meaning Connections. Language Learning, 62: 2, 634-662.
MacWhinney, B. (2008) A Unified Model. In N. Ellis \& P. Robinson (Eds.) Handbook of cognitive linguistics and second language acquisition. Lawrence Erlbaum Press.
Nissen, M.J., \& Bullemer, P. (1987) Attentional requirements of learning: Evidence from performance measures. Cognitive Psychology, 19:1, 1-32.
Perruchet, P. and Pacton, S. (2006) Implicit learning and statistical learning: one phenomenon, two approaches. Trends in Cognitive Science, 10:5, 233-238.
Reber, A. S. (1967). Implicit learning of artificial grammars. Journal of Verbal Learning and Verbal Behavior, 6, 855-863.
Tomasello, M. and Herron, C. (1989) Feedback for Language Transfer Errors. Studies in Second Language Acquisition, 11, 386-395.
Williams, J. N. (2005). Learning without awareness. Studies in Second Language Acquisition, 27, 269304.

## Reading Motor Intentions

Lewkowicz Daniel (daniel.lewkowicz@etu.univ-lille3.fr)<br>URECA laboratory, Univ Lille Nord de France, F-59000 Lille, France

Delevoye-Turrell Yvonne (yvonne.delevoye@univ-lille3.fr)<br>URECA laboratory, Univ Lille Nord de France, F-59000 Lille, France<br>MESHS, USR 3185, F-59000 Lille, France


#### Abstract

Some evidence in very recent psychological studies have demonstrated that motor simulation ability is crucial for the correct understanding of social intentions. The present study was conducted first to confirm that the nature of the motor intention leads to early modulations of movement kinematics. Then, we tested whether humans could read an agent's intention when observing the very first element of a complex action sequence. Results revealed early variations in movement kinematics and further showed that human agents can use these deviants to distinguish above chance level between three different social actions. Similar performance levels were found using an artificial classifier (Neural Network) and this procedure demonstrated furthermore that decisions could be taken on the basis of information contained in the first 500 ms of movement kinematics. Taken together these results confirm the importance of motor simulation for adapted social interaction, and suggest how robotic adaptive controllers may use as input low-level motor information (e.g. kinematics) to afford biologically inspired social behaviors.


Keywords: Classifier; kinematics; sequences; motor control; intentionality: social interaction; internal models; prediction; motor planning; biological movement.

## Introduction

In everyday activities, the grasping of an object might be performed with different prior intentions: e.g. touch, move, throw or pass. Ansuini et al. (2008) have measured the prior-to-contact grasping kinematics for reach-to-grasp movements performed toward a bottle filled with water. By comparing hand shaping across tasks involving different subsequent actions - pour the water into a container; throw the bottle; move the bottle from one spatial location to another - the authors demonstrated how the prior intention in grasping the object strongly affected the positioning of the fingers during the reaching and the contact phase of the action (Ansuini, Giosa, Turella, Altoè, \& Castiello, 2008). In another series of studies, Becchio and collaborators investigated the effects of social context on reach-to-grasp actions. They found initial adjustments reflecting specific planning strategies (Becchio, Sartori, Bulgheroni, \& Castiello, 2008a) as well as online adjustments (Sartori, Becchio, Bulgheroni, \& Castiello, 2009) when performing under social context (see : Becchio et al., 2010 for a review).

More recently, researchers have gone one step further to suggest that not only end-point constraints and social
contexts affect movement kinematics, but that these deviants may be used to read motor intention. For example, when observing actions performed under social context or not, Castiello and collaborators demonstrated that humans can successfully use kinematic cues of reach-to-grasp movements to predict the final goal of the action (Sartori, Becchio, \& Castiello, 2011). Similar results were also found using point-light displays of simple reach to grasp movements (Manera, Becchio, Cavallo, Sartori, \& Castiello, 2011). However, in these studies, the classification rates were obtained under a forced two-choice paradigm, and for the most subtle differences (cooperative vs individual preferred speed or competitive vs fast speed) the classification rates were very small (near 50\%).

In the present work, we wanted to study the capacity of humans to read motor intention in a sequence of 2 motor elements. One novelty of this study is that the sequences were performed entirely during an interactive situation with a con-specific, without any interruption or verbal instruction between the sequences. As such, we recorded sequential actions during an ecologically inspired task (Jungle Speed), a simple face-to-face game using a unique manipulated object. Our main focus was to compare human and artificial categorization performances for three different sequential actions that took part during the game. To test the hypothesis that kinematics alone is sufficient to read social intention, we fed the artificial classifier with movement kinematics only.

Confronting Jacob \& Jeannerod's (2005) reading motor intention hypothesis, we hypothesized that human agents are able to read motor intention through the simple observation of arm kinematics of the first element of a 2-sequence action. This is possible due to the fact that arm kinematics of the reach to grasp movements reveal specific deviants in function of goal intention from an ideal optimized trajectory. Finally, if motor simulation is sufficient, then an artificial neural network should be able to learn from the deviants and predict as well as humans, the motor intention of an observed agent. In the following section, we first describe the methods we used to make the observation videos (Part A), which were then played to human agents (Part B) and used as input parameters to an artificial neural network (Part C).

## Creating Stimuli

Two adults participated in the study, one experimenter and the other as subject. Both participants were right handed
as verified with the Handedness questionnaire (Oldfield, 1971). They had no prior knowledge of the experiment and provided informed consent before participating in the experimental session that lasted approximately 90 minutes. The subjects' movements were recorded using a (1) a video camera (Sony Handycam) and (2) 4 Oqus infrared cameras (Qualysis). 5 infrared reflective markers were placed on the thumb (tip), index (base and tip) and the wrist (scaphoïd and pisiform). Cameras were calibrated before each session, allowing the system to reach a standard deviation smaller than 0.2 mm , with a 200 Hz sampling rate


Figure 1. Schematic representation of the experimental setup.
The game. The object that was to be manipulated by the subject was a wooden dowel of width 2 cm and height of 4 cm that was placed precisely 20 cm in front of the starting position ('pick' position). The subject started each trial by pinching index and thumb together at the starting position (see Figure 1). Each trial, the subject's task was to reach and grasp the dowel between thumb and index finger in order to move it from the initial position to one of the three 'place' positions during an adapted version of the jungle-speed game. The game consisted in 4 blocks of approximately 40 rounds. First, the subject's task was to pick and place the dowel at the 'Play' position in order to set the initial condition of the game. Then, at the 'go' signal (high pitch) both participants reached for the dowel as quickly as possible. The competitive move was not recorded and is not part of this study, although, this was indeed intentionally omitted during instructions given to the subject. During competitive move, the first who have grasped the dowel, won the round and scored a point. The second phase consisted in a rewarding phase. The dowel was first always repositioned at pick position and the subject wait at starting position for the next audio tone (low pitch). During the rewarding phase just after the competitive move, the subject has to reach to grasp the dowel and place it either at the 'You' position if the experimenter scored during competitive move, or at the 'Me' Position if the participant scored during competitive move. The game went on until one of the two players reached 20 points. Thus, we recorded twice as much 'Play' moves than 'You' or 'Me' moves during the game. Nonetheless, after 5 rounds of training to
set up the game rules, no other verbal instructions were given during the blocks. The three different positions ('Play'; 'Me'; 'You') where the dowel had to be placed were delimited by visual marks directly placed on the tabletop.

The recordings. The best 16 recordings of each category ('Play', 'Me' and 'You') were extracted using VirtualDub and kept for future use as stimuli. Each sequence was delimited with a 1 -second pre-trial, i.e., before the initial movement onset, and was cut exactly one frame before the index finger contacted the object. Movies were compressed with FFdshow codec (MJPEG) at 50 frames per seconds with a screen resolution of $720 \times 576$. Video clips were coupled with the recordings of the arm kinematics using 4 Oqus infrared Cameras (Qualisys). Infrared reflective markers were placed on the index (base and tip), the thumb (tip), the wrist (scaphoïd and pisiform) of each participant, as well as on the object. Cameras were calibrated before each recording session, allowing the system to reach a standard deviation smaller than 0.2 mm for all three absolute positions at a 200 Hz sampling frequency. Care was taken as to provide no contextual information within the video clips (torso, gaze, face expression), i.e., only the hand and the target object were fully in view. Velocity profiles are presented in Figure 2 and show that play, me or you sequences show deviations during both first and second motor element (amplitude and the width of the bell-shaped curves, first and second peaks of velocity, time position for local minima).


Figure 2. Mean velocity profiles for the three categories of sequences. Each bell-shape curve corresponds to a motor element. The first is the reach to grasp element, and the second bell-shape curve is placing element. The local minima are used to segment the two motor elements and compute movement times.

## Human Categorization Performance

The short video clips were presented to a panel of human subjects to test whether human agents were able to predict
the goal of a sequential action when shown only the first sub-element of a sequence, i.e., the reach movement.

Participants, Apparatus and Software. Twenty-six young adults (mean age: $21.82 \pm 2.76$ years, range $=18-29$ years) participated in the study. All subjects were right handed (Oldfield, 1971) and had no prior knowledge of the experiment. Subjects provided informed consent before participating in the experimental session that lasted approximately 45 minutes. Participants were seated comfortably facing a table in a dark and silent room. For each trial, participants started by placing their hand on response keys that were delimited by tape placed directly on the keyboard keys. Stimuli were presented on a laptop computer with MATLAB software (Mathworks) with the PsychToolbox environment.

Experimental Procedure. The participants' task was to answer on the keypad after each video clip presentation whether the social intention of the sequence was 'let's Play' (5 key), 'for Me' (2 key) or 'for You' (8 key). A 1-second blank screen was displayed in between two trials. Participants were instructed to give their answers as fast and as accurately as possible. They were obliged to provide an answer within a 4-second time window otherwise the trial was cancelled and presented at the end of the block. A feedback message was given to tell participants if their response was too slow. Each block consisted in the random presentation of a series of 48 stimuli, i.e., 16 different video clips for each of the three categories (Play; Me; You). After a 5-minute pause, the next block of 48 video clips was presented.

Dependent variables and statistical analyses. The number of correct responses (correct prediction of the ongoing action) and the response times were calculated for each category. The dependent measures were submitted to a repeated-measure ANOVA with Block and Category (Play; Me; You) as within factors. The participants' scores for each category were compared using a Chi-square between the observed scores and the random distribution between categories corrected by total the number of answer of that category. In other term, because the total amount of answer is not exactly the same between categories, we consider the guessing base-rate of each category separately. The alpha level of significance was set to 0.05 .

Response times. Results showed no effect of Block on response time, $\mathrm{F}(2,50)=1.401, \mathrm{p}=.256$, indicating that participants answered with a similar response time in the first ( $\mathrm{M}=878$, $\mathrm{SD}=382 \mathrm{~ms}$ ), the second $(\mathrm{M}=$ $848, S D=315 \mathrm{~ms})$, and in the third block $(\mathrm{M}=944, \mathrm{SD}=$ 316 ms ). No effects of Category were found on response time, $\mathrm{F}(2,50)=2.621, \mathrm{p}=.083$, indicating that participants answered within the same delay both for 'Play' ( $\mathrm{M}=900$, SD $=294 \mathrm{~ms})$, 'Me' $(\mathrm{M}=866, \mathrm{SD}=294 \mathrm{~ms})$, and 'You' categories ( $\mathrm{M}=905$, $\mathrm{SD}=300 \mathrm{~ms}$ ).

Number of correct responses. There was an absence of Block effect on classification performances, $\mathrm{F}(2,50)=$ $0.102, p=.903$. However, a main effect of Category was obtained, $\mathrm{F}(2,50)=16.022, \mathrm{p}<.001, \eta_{\mathrm{p}}^{2}=.39$. Post-hoc Scheffé analyses further showed that participants were more accurate for trials in the 'Me' category ( $\mathrm{M}=57.53$, $\mathrm{SD}=$ 13.02 \%) than in the 'You' ( $\mathrm{M}=40.87, \mathrm{SD}=12.12 \%)$ and in the 'Play' category ( $\mathrm{M}=47.27, \mathrm{SD}=13.04 \%$ ). More importantly, Chi-squared tests showed highly significant difference between observed frequencies and random guessing baselines for the 'Me' (guess rate $=36.98$, $\mathrm{p}<.001$ ), 'Play' (guess rate $=36.12, \mathrm{p}<.001$ ) or 'You' (guess rate $=26.90, \mathrm{p}<.001$ ). These results confirmed that performance was significantly greater than chance level.

## Categorization with Artificial Neural Networks

In the following section, we present the simple feedforward neural network that was developed to demonstrate the possibility to categorize on the only basis of motor kinematics.

Architecture and Learning procedure. A simple classification Neural Network was constructed with N neurons (1-23 neurons) as inputs, 3 hidden layers and 3 output neurons (one for each category). The N size is the sub-selection of the total movement duration. Activation functions for the output layers were symmetrical and sigmoid, between -1 and 1 .

With this NN, the instantaneous velocity in 3D was calculated between the recorded positions of the wrist for two subsequent frames. A threshold of $20 \mathrm{~mm} . \mathrm{s}^{-1}$ was then determined to compute the reaction time (RT) delay between the start of the recording and the actual beginning of the movement. Second, a sampling parameter was used to compute the average velocity across 10 frames. Third, the mean velocity values were converted from $\mathrm{mm} . \mathrm{s}^{-1}$ to $\mathrm{m} . \mathrm{s}^{-1}$ in order to get data within an overall range of 0 to 1 . Finally, a training set (25\%) and a test set (75\%) were randomly picked from the 144 different kinematic recordings. 20 different networks were trained to obtain the classification performance for every specific target time widow (i.e. time window for kinematic recognition). The mean response and variance across the 20 networks are described in the result section as the NN success rate (this value is always lower than the best performing network). By varying the amount of data fed as input, we computed the classification performance from multiple time windows. The learning procedure was a back-propagation algorithm using the FANN library (Nissen, 2003). Target error (to stop the learning) was set to MeanStandardError < 0.001 with a maximum number of epochs set to 10000 , and 300 iterations between each test (evaluation of target global error.

Classification results in function of time. Results revealed that the simple artificial classifier was able to converge in most cases. The classifier succeeded in
discriminating between categories for input sizes above 9, i.e. with a time window of 450 ms . For the input size of 9 , single sample $t$-tests confirmed that all categories were above chance level, $\mathrm{p}<.001$ : 'Play' ( $\mathrm{M}=55.70 \mathrm{SD}=8.08$ \%); 'Me' category ( $\mathrm{M}=56.70 \mathrm{SD}=4.16 \%$ ), and 'You' category ( $\mathrm{M}=50.33 \mathrm{SD}=5.63 \%$ ). Figure 3 presents the detailed results for 12 different input sizes, between 50 ms to 1150 ms with a step of 100 ms . From the input size of 5 ( 250 ms ) to 9 ( 450 ms ), only 2 categories were successfully recognized while the other remained below chance level; Below 250 ms , only one category was correctly classified. The crucial point to note here is nevertheless the fact that by 450 ms all categories are classified above chance level; a point in time that occurs before the end of the first subelement of movement sequence confirming the capacity of a simple network to predict motor intention by the use of lowlevel kinematics early on during motor execution.

Artificial Neural Network (Classifier)


Figure 3. Results obtained with the ANN. Note that with an input size of 450 ms , most of the networks classify the movements with a higher rate than chance level and before the end of the first motor element (vertical grey bar).

## Discussion

In the present contribution, we report experimental data confirming that motor intention can be read through the simple observation of movement kinematics. More specifically, we first showed that the three different motor intentions that were used in a simplified version of the Jungle Speed game (Asmodee eds.) modified the kinematics of the first (reach) sub-element of the sequential action. Second, human agents were able to classify rapidly ( $<1$ s) and above chance level ( $>40 \%$ ), the trial category when observing a video-clip of the reaching movement only of the sequence. Third, a classic feedforward neural network was also able to categorize motor intention through the use of low-level kinematic information of, once again, the reaching sub-element only. In the following section, we discuss these findings in more detail and describe how this work can help
advance the development of future cybernetic systems that will afford true human-robot interactivity.

Kinematics reflecting motor intention. In the abundant literature of manipulating actions, the effects of end-point constraints on the early parts of movement kinematics have been investigated extensively in experimental psychology. In individualistic situations, multiple sources have been reported to modify and shape hand trajectory in two-element sequences such as second-target distance (Gentilucci, Negrotti, \& Gangitano, 1997), end-target orientation (Haggard, 1998; Hesse \& Deubel, 2010) or second-action type (Armbrüster \& Spijkers, 2006; Marteniuk, MacKenzie, Jeannerod, Athenes, \& Dugas, 1987). In social interactive manipulative tasks, final-goals have also been reported as having an effect on reach-to-grasp kinematics such as giving vs. placing an object (Becchio et al., 2008a), cooperative vs. competitive actions (Becchio, Sartori, Bulgheroni, \& Castiello, 2008b; Georgiou, Becchio, Glover, \& Castiello, 2007), absence vs. presence of social request (Ferri, Campione, Dalla Volta, Gianelli, \& Gentilucci, 2011), verbal communicative vs. non-communicative intentions (Sartori, Becchio, Bara, \& Castiello, 2009). The kinematic effects reported here are consistent with this literature and suggest that when planning a sequential action with multiple sub-elements, the requirements of the endpoint element are back-propagated to constrain the way the very first element of the sequence will be planned and performed. Thus, it is possible to suggest that low-level motor components may contain early indices that reflect the end-point motor intention of an agent.

Reading intentions. In the present study, the first part of each movement was identical, i.e., the agents initiated their move with their hand placed on the starting pad of the playing area, and reached for and grasped the wooden-peg that was always at the same position on the table. However, the second part of the move was specific and directly related to the game intention: lift the wooden peg to take it ('Me' category), to give it ('You' category) or to place it on the table ('Play' category). Thus, any kinematic deviants observed on the first part of the sequence may be related to the social intention of the second part. By measuring two basic motor parameters (peak velocity and movement duration), we showed that it was possible to dissociate the three types of social interaction categories (Figure 2). We then tested the fact that human observers could use these deviants to classify observed actions above chance level. The video clips were created in order to show the first subelement only, without any contextual cues; care was also taken to cut the end of the reaching action, one frame before object contact, in order to avoid providing any cues on movement direction of the second part of the sequence. Our findings demonstrate that classification is possible and that in certain cases, the participants' performance can be extremely precise (up to $67 \%$ of correct classification for the best of participants). But how is this possible?

An alternative low-level hypothesis. It is nevertheless possible that the understanding of motor intention is based on more low-level cue readings. As suggested by the work of Perrett and al. (Perrett et al., 1989), the visual system definitely contribute to action recognition and the performance showed by humans could be interpreted as the resolution of an "inverse" problem (goal attribution) with a simple bayesian inference about which goal explain the best which action (Csibra, 2008). Indeed, despite a total absence of contextual cues within the video clips (body, head, eyes), we demonstrated in the present study that participants were able to read motor intention significantly above chance level. Hence, the subjects' responses could be guided by the slight deviances from the optimal strategy (i.e. to grasp without any subsequent action) in the low-level motor kinematics. This confirms recent results presented by Stapel et al. (2012) who showed that in absence of contextual cues, kinematics could be a key source of information to predict intentions of ongoing actions. To go further in this low-level hypothesis, we conducted a second work for which we used a very simple artificial neural network classifier and we showed that this simple classifier performed as well as our human subjects in categorizing the three different socialintended video-clips. Further studies, namely brain imaging, are needed in order to determine whether the good performance reached in our human individuals was due to direct coding of the low-level kinematic parameters or whether the kinematics deviants are simple by products and that even for simple actions, human performances engage in a cognitive motor simulation to read motor intention (see e.g. Kilner, Friston, \& Frith, 2007; Kilner \& Frith, 2008).

It is to note that correct classification of the three social categories was far from being perfect, reaching in the best of cases $60 \%$ of correct identification. Hence, kinematics can be used for predicting ongoing actions but cannot be the only source, used by human agents to judge motor intention. It has been shown that during natural sequential task (i.e. preparing a sandwich), eye movements are stereotyped and predictive (Hayhoe, Shrivastava, Mruczek, \& Pelz, 2003). During the task, the eye precedes the hand movements in systematic way ensuring a good coding of object position for accurate planning of arm (Johansson, Westling, Bäckström, \& Flanagan, 2001). This coordination between eye and hand movements during manipulative tasks have extensively been tested in experimental psychology and have demonstrated that e.g., eye movement onset is always faster than hand movement onset, and the peak velocity of both eye and hand movements are strongly correlated, suggesting that they possess a coupled function. It is thus possible that using both gaze position and the hand movements kinematics, an observer would be able to increase the efficiency of intention reading (see also : (Bekkering \& Neggers, 2002).

## Perspectives for interactive and social robotics.

The application of our work would be to develop
robots that afford true interaction, i.e., being able (1) to read motor intention in human kinematics in order to adapt but also (2) to move with biological realistic kinematics, in order allow others to understand the intention of the robot. Following the data presented here, we hypothesize that a humanoid robot could become interactive if it moved following the laws of biological movement with action sequences that integrate back propagation of terminal intention. Such a phenomenon would provide the means for human agents to read intentionality and thus, gain in understanding the goal of the robot's movements. Furthermore, including social deviants in the motor kinematics within early steps of motor sequences would also allow safe interaction with large industrial robots by affording humans the possibility of anticipating false moves in joint actions that share similar work spaces.

Implementing robots with the architecture necessary to "afford intentionality" would need to integrate the different brain regions that are known to play a role in motor planning and motor-sensory predictive mapping. De Rengervé et collaborators (de Rengervé, Hirel, Andry, Quoy, \& Gaussier, 2011) have recently reported on such an architecture, which included amongst other areas, the cerebellum and the basal ganglia. Tested on both software and hardware, this neural architecture has demonstrated its efficiency on data collected in a hydraulic robotic arm. With a series of imitation trials, this system demonstrated the capacity to learn how to perform sequential actions that respected biological laws, i.e., to perform movements with kinematics that mirror those performed by human agents. As such, this robot arm has demonstrated increased interactivity with human agents affording augmented interaction both in time and in space (none published results). Ongoing studies are now being conducted to assess whether this interactivity is associated to an increase in the capacity of human collaborators to read the robot's intention.

## Conclusion

We have here described experimental findings in humans demonstrating that it is possible to read motor intention through the simple observation of kinematic deviants. Classification capacities were significantly above chance level and provided human subjects the means to dissociate between three different socially oriented actions. We argue in the present study that reading intentionality may not depend on a high-level cognitive function as suggested in the psychological literature. Internal simulations may not be systematically required and understanding other intentions may in certain cases relate to a direct coding of those kinematic deviants that back propagate from end-point to early on during sequence execution; this direct coding would emerge through years of joint-action experiences, during interactions with adult con-specifics. As a first step to support this hypothesis, we report in the present study simple neural networks that were able, after learning the meaning of kinematic deviants, to classify the three categories of actions to the same degree of accuracy than
our human participants. These preliminary results stresses the importance of further developing the optimal theories of motor control in order to include the effects on sequential actions such as, back propagation phenomena of social context.

## Acknowledgments

The work reported here was financially supported by a grant from the French Research Agency ANR-2009-CORD-014INTERACT. The work provided by DL was also financed by the region Nord-Pas-de-Calais (France).

## References

Ansuini, C., Giosa, L., Turella, L., Altoè, G., \& Castiello, U. (2008). An object for an action, the same object for other actions: effects on hand shaping. Experimental Brain Research, 185(1), 111-119.
Armbrüster, C., \& Spijkers, W. (2006). Movement planning in prehension: do intended actions influence the initial reach and grasp movement? Motor control, 10(4), 311-329.
Becchio, C., Sartori, L., Bulgheroni, M., \& Castiello, U. (2008a). The case of Dr. Jekyll and Mr. Hyde: A kinematic study on social intention. Consciousness and Cognition, 17(3), 557-564.
Becchio, C., Sartori, L., Bulgheroni, M., \& Castiello, U. (2008b). Both your intention and mine are reflected in the kinematics of my reach-to-grasp movement. Cognition, 106(2), 894-912.
Becchio, C., Sartori, L., \& Castiello, U. (2010). Toward You The Social Side of Actions. Current Directions in Psychological Science, 19(3), 183188.

Bekkering, H., \& Neggers, S. F. W. (2002). Visual search is modulated by action intentions. Psychological Science, 13(4), 370-374.
Csibra, G. (2008). Action mirroring and action understanding: An alternative account. Sensorymotor Foundations of Higher Cognition. Attention and Performance XXII, 435-459.
De Rengervé, A., Hirel, J., Andry, P., Quoy, M., \& Gaussier, P. (2011). On-line learning and planning in a pick-and-place task demonstrated through body manipulation. In 2011 IEEE International Conference on Development and Learning (ICDL) (Vol. 2, p. 1-6).
Ferri, F., Campione, G. C., Dalla Volta, R., Gianelli, C., \& Gentilucci, M. (2011). Social Requests and Social Affordances: How They Affect the Kinematics of Motor Sequences during Interactions between Conspecifics. PLoS ONE, 6(1), e15855.
Gentilucci, M., Negrotti, A., \& Gangitano, M. (1997). Planning an action. Experimental Brain Research, 115(1), 116-128.
Georgiou, I., Becchio, C., Glover, S., \& Castiello, U. (2007). Different action patterns for cooperative
and competitive behaviour. Cognition, 102(3), 415 -433.
Haggard, P. (1998). Planning of action sequences. Acta Psychologica, 99(2), 201-215.
Hayhoe, M., Shrivastava, A., Mruczek, R., \& Pelz, J. (2003). Visual memory and motor planning in a natural task. Journal of Vision, 3(1).
Hesse, C., \& Deubel, H. (2010). Advance Planning in Sequential Pick-and-Place Tasks. Journal of Neurophysiology, 104(1), 508-516.
Johansson, R. S., Westling, G., Bäckström, A., \& Flanagan, J. R. (2001). Eye-hand coordination in object manipulation. the Journal of Neuroscience, 21(17), 6917-6932.
Kilner, J. M., Friston, K. J., \& Frith, C. D. (2007). Predictive coding: an account of the mirror neuron system. Cognitive Processing, 8(3), 159-166.
Kilner, J. M., \& Frith, C. D. (2008). Action observation: inferring intentions without mirror neurons. Current Biology, 18(1), R32-R33.
Manera, V., Becchio, C., Cavallo, A., Sartori, L., \& Castiello, U. (2011). Cooperation or competition? Discriminating between social intentions by observing prehensile movements. Experimental Brain Research, 211(3-4), 547-556.
Marteniuk, R. G., MacKenzie, C. L., Jeannerod, M., Athenes, S., \& Dugas, C. (1987). Constraints on human arm movement trajectories. Canadian Journal of Psychology, 41(3), 365-378.
Nissen, S. (2003). Implementation of a fast artificial neural network library (fann). Report, Department of Computer Science University of Copenhagen (DIKU), 31.
Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia, 9(1), 97-113.
Perrett, D. I., Harries, M. H., Bevan, R., Thomas, S., Benson, P. J., Mistlin, A. J., ... Ortega, J. E. (1989). Frameworks of analysis for the neural representation of animate objects and actions. Journal of Experimental Biology, 146(1), 87-113.
Sartori, L., Becchio, C., Bara, B. G., \& Castiello, U. (2009). Does the intention to communicate affect action kinematics? Consciousness and Cognition, 18(3), 766-772.
Sartori, L., Becchio, C., Bulgheroni, M., \& Castiello, U. (2009). Modulation of the action control system by social intention: unexpected social requests override preplanned action. Journal of Experimental Psychology. Human Perception and Performance, 35(5), 1490-1500. doi:10.1037/a0015777
Sartori, L., Becchio, C., \& Castiello, U. (2011). Cues to intention: The role of movement information. Cognition, 119(2), 242-252.

# The Influence of Collective Opinion on True-False Judgment and Information-Sharing Decision 

Huaye Li (hli21@stevens.edu)<br>Yasuaki Sakamoto (ysakamot@stevens.edu)<br>Howe School of Technology Management, Stevens Institute of Technology<br>Hoboken, NJ 07030 USA


#### Abstract

The purpose of the current work is to examine when and how knowing collective opinion influences people's judgments and decisions in social media environments. In particular, the present work focuses on people's true-false judgment of statements found on websites and the likelihood of sharing these statements. The results from Experiment 1 revealed that, for false statements, collective opinion had little influence on people's true-false judgments, but, for true and debatable statements, their judgments were biased toward collective opinion. The results from Experiment 2 indicated that the likelihood of sharing the true, debatable, and false statements followed the collective opinion, and that people were less likely to share false statements than debatable or true ones without collective opinion. These findings extend past work on social influence and advance understanding of how people make judgments and decisions in social media websites.


Keywords: Collective opinion, true-false judgment, information-sharing decision, social media, social influence

## Introduction

Social media technologies, such as Twitter and Facebook, have become part of many people's everyday lives. Using these technologies, people not only acquire new information but also generate content and influence trends. Given the growing use and participatory nature of social media, better understanding of how people behave in such an environment is essential.

The objective of the current paper is to contribute to this need by reporting the results from two experiments that examine how people make true-false judgments on statements found on websites and how they make decisions about whether or not to share these statements in a social media environment. In the current work, sharing of a statement means passing of the statement to others. In particular, the work presented here focuses on how social influence plays a role in people's true-false judgments and sharing decisions in a social media context.

## Collective Opinion

One of the main functions of social media is to share opinions with others and collectively make decisions (e.g., Glushko et al., 2008). Collective opinion, such as how many people have liked or shared a message, is part of social media technologies. In Twitter, an example of collective opinion is the number of re-tweets, or the number of people
who re-tweeted a particular message. Re-tweeting is a kind of sharing, in which an original tweet, a brief message of 140 characters or less, is broadcast to the re-tweeter's followers through a simple clicking of a re-tweet button. The re-tweeted messages become available to the public as well, and the number of re-tweets associated with a tweet signifies the popularity of the tweet. The more ret-tweets, the more popular. Facebook uses 'like' to indicate the popularity of photos, stories, communities, and so on. Many review websites allow users to indicate their opinion about the usefulness of a particular review.

Despite the abundance of collective opinion in social media websites, when and how it affects people's judgments and decisions in such an environment is not well understood. However, there are classic studies on social influence in face-to-face environments, and more recent work on social influence in online environments, whose findings indicate that people use collective opinion to make their judgments and decisions in various situations.

## Social Influence

Past research in face-to-face environments has shown that people have a strong motivation to compare their opinions with others (Festinger, 1954), and they often adopt the decisions of others (e.g., Cialdini \& Goldstein, 2004; Deutsch \& Gerard, 1995; Gureckis \& Goldstone, 2006) due to their desire to make correct responses under uncertainty (Sherif, 1935) or their desire to be liked by others (Asch, 1955). By relying on others' opinions, individuals can learn and entertain solutions that they would not have even considered otherwise (Bandura, 1965).

More recent work has shown that knowing other's decisions also influences people's decisions in online environments. In an online market experiment, whereas good music was always downloaded by many and bad music was always unpopular, the popularities of the pieces in between varied depending on whether or not people knew the number of downloads the pieces had (Salganik, Dodds, \& Watts, 2006). In another set of online experiments, subjects liked the same online news stories more when the stories had many existing supporters than when the stories had only a few supporters (Sakamoto et al., 2009). Subjects even switched their preferences when the experimenter flipped the assumed numbers of previous supporters (Sakamoto, 2010; Salganik \& Watts, 2008). These past studies on social influence suggest that people's liking and rating can follow collective opinion in social media
environments. The current research extends this past work on social influence to true-false judgments and sharing decisions in a social media environment.

## Hypotheses

Social influence may be especially strong in social media environments, as people experience a lot of information whose factual accuracy is unclear, and they rely on collective opinion in an attempt to reduce uncertainty. During crises, for example, uncertainty is high and people are under pressure mentally. People generate a large amount of information in an attempt to make sense of the situation, and they readily share this information without verifying its factual accuracy, resulting in the dissemination of false rumors (e.g., Allport \& Postman, 1947; DiFonzo \& Bordia, 2007). Consequently, a large amount of debatable information appears during responses to disasters.

In relation to debatable information, one focus of the current work is to examine how collective opinion influences people's judgment about whether a statement is true, false, or debatable. The statements used in the current research were related to health advice. A statement is true when the advice in the statement is supported by clear evidence according to health professionals. A statement is false when it contains information identified as incorrect by health professionals. A statement is debatable when health professionals cannot verify its factual accuracy because evidence is mixed or missing. The first hypothesis is that the true-false judgment of debatable statements will be most prone to social influence because their factual accuracy is unclear; people will adopt collective opinion in an attempt to make a correct decision when they encounter debatable statements. In contrast, people's perceptions of true and false statements are relatively strong and hard to change, just as good and poor pieces of music were immune to social influence in Salganik et al.'s (2006) research.

Another focus of the present work is to examine how collective opinion influences people's intention to share a statement differently depending on whether the statement is true, debatable, or false. Because it is unlikely that people strongly feel that they should or should not share the healthrelated statements used in the present work, they will rely on collective opinion to make their decision. The second hypothesis is that social influence takes place for true, debatable, and false statements; people's decision will follow the collective opinion. If this is the case, there will be a positive social influence, in which increasing the value of collective opinion increases the likelihood of sharing.

An alternative account is that people want to share information that others have not shared. There is evidence in consumer research that some people want to be unique and differentiate themselves from others (Berger \& Heath 2007; Snyder \& Fromkin 1980; Tian, Bearden, \& Hunter 2001). According to this view, in deciding whether or not to share a statement in social media, people will go against the information about the number of people who already shared the statement. This leads to a negative social influence.

It is unclear whether positive or negative social influence takes place in information sharing decisions. In the information transmission literature, researchers focus mainly on how factors such as valence and source credibility relate to the spread of information (e.g., Fragale \& Heath 2004; Ha \& Ahn 2011; Heath 1996; Rene et al. 2012). Those who study social dimensions tend to examine how social network structures, including the number of followers and position in the social network, affect information diffusion (e.g., Cha et al. 2010; Huberman et al. 2009; Kwak et al. 2010; Xin et al. 2012) and how to maximize the spread of influence through online social networks and the extent to which one could predict online popularity (Kempe, Kleinberg and Tardos, 2003; Kim, Kim \& Cho, 2011). Social influences on sharing decisions are not well studied.

Another important question the current work can address is whether people are more likely to share true, debatable, or false information, without knowing collective opinion. Although this question does not involve the main topic of social influence, the answer will be useful. Here, we borrow an idea from the rumor psychology literature. The third hypothesis is that debatable statements result in a higher likelihood of sharing than false and true statements because debatable statements will induce informational ambiguity and anxiety by being disputable. Ambiguity and anxiety are proposed to be strong predictors of rumor spread (Anthony 1973; Shibutani 1966).

## Experiment 1

The main purpose of Experiment 1 is to examine how collective opinion influences people's judgment about whether a statement is true, false, or debatable.

## Method

Participants In return for a nominal fee, 227 workers of Amazon's Mechanical Turk (https://www.mturk.com) completed the experiment. The mean age was 36 . Using Mechanical Turk, a few research groups have replicated classic psychological phenomena and have shown that researchers can collect high-quality data (e.g., Buhrmester, Kwang, \& Gosling, 2011; Mason \& Suri, 2011; Paolacci, Chandler, \& Ipeirotis, 2010). We followed their recommendations.
Materials From Discovery, Food Networks and National Institute of Health (NIH), 120 statements about health advice were selected with two constrains: each statement was clearly identified by health professionals as true, debatable, or false, and the information carried by each statement was familiar to most people. Of 120 statements, 40 were true, 40 were debatable, and 40 were false.
Design and Procedure Subjects were instructed to read a health-related statement online, and to rate the extent to which the statement was true using a 7 -point scale, where 1 was definitely false and 7 was definitely true. A response
around the middle of the scale indicated debatable. The 120 statements were presented sequentially in a random order.

There were three conditions.

1. Control: Statements were presented with no social information. Fifty unique subjects rated each statement. Figure 1 shows an example of the actual screen used in the control condition.
2. Real: Each statement was presented with real collective opinion, which was the mode of the 50 ratings from the control condition. We used the mode because it preserved extreme ratings. With mean and median, the ratings tended to go toward the middle of the scale. Figure 2 shows an example.
3. Invented: Each statement was presented with invented collective opinion. We transformed the observed mode as follows: $1 \rightarrow$ (became) $7,2 \rightarrow 6,3 \rightarrow 5,4 \rightarrow 7$ if the mean was smaller than $4,4 \rightarrow 1$ if the mean was larger than $4,5 \rightarrow 3,6 \rightarrow 2$, and $7 \rightarrow 1$. Figure 3 shows an example.
```
Background: People are bombarded with more informaton than they can process in social media, such as Twitter and Facebook. Although people can find a lot of useful information in social media, they also encounter a significant amount of unverified information. Please help us improve social media environments. Your responses are valuable!
Question A: Does the statement below contain the word fiber? Yes No
Oatmeal contains soluble fiber, which rectuces your low.density lipoprotein
LDL), the "bad" cholesterol.
Question B: Imagine that you are reading the statement above in social media, such as Twitter and Facebook. To what extent do you think this statement is true or false? 1234567
definitely false \(\circ \circ \rho \circ \circ \circ \circ\) definitely true
***If you want to know more about the above statement, please email us (intuitive.analytic@gmail.com)***
```

If this is your first HIT from this batch, please complete the following:

Thank you for your participation!
Figure 1. Example of Experiment 1's control condition

```
Question A: Does the statement below contain the word fiber? Yes No \begin{tabular}{|l|}
\hline \(\begin{array}{l}\text { Oatmeal contains soluble fiber, which reduces your low-density lipoprotein } \\
\text { (LDL), the "bad" cholesterol. }\end{array}\) \\
\hline
\end{tabular}
Question B: Imagine that you are reading the statement above in social media, such as Twitter and Facebook. To what extent do you think this statement is true or false?
```

```
True-false rating by majory of others like you: }
```

True-false rating by majory of others like you: }
1234567
definitely false 0 0 0 0 0 definitely true

```

Figure 2 Example of Experiment 1's real condition
Question A: Does the statement below contain the word fiber? Yes No

Question B: Imagine that you are reading the statement above in social media, such as Twitter and Facebook. To what extent do you think this statement is true or false?
True-false rating by majory of others like you: 2
            1234567
    definitely false \(\circ \circ \circ \circ \circ \circ \circ\) definitely true

Figure 3. Example of Experiment 1's invented condition

Table 1 shows how the collective opinion observed in the control condition was used to generate the collective opinion shown to subjects in the real and invented conditions.

Table 1. Collective opinion
\begin{tabular}{ccc}
\hline \begin{tabular}{c} 
Observed in the \\
control condition
\end{tabular} & \begin{tabular}{c} 
Used in the real \\
condition
\end{tabular} & \begin{tabular}{c} 
Used in the \\
invented condition
\end{tabular} \\
\hline 1 & 1 & 7 \\
2 & 2 & 6 \\
3 & 3 & 5 \\
4 & 4 & 1 or 7 \\
5 & 5 & 3 \\
6 & 6 & 2 \\
7 & 7 & 1 \\
\hline
\end{tabular}

\section*{Results and Discussion}

All participants were included in the analyses. The main interest of Experiment 1 was whether and how collective opinion might influence people's true-false judgments. Figure 4 shows the overall pattern of the results.

A 3 (statement: true, false, debatable) by 3 (condition: control, real, invented) analysis of variance (ANOVA), with true-false ratings as a dependent measure, revealed a significant main effect of statement, \(F(2,117)=33.95, p<\) .001. Collapsing across condition, the true-false ratings for true (5.05), debatable (4.72), and false statements (3.62) differed significantly. There was also a significant main effect of condition, \(F(2,396)=26.65, p<.001\). Collapsing across different types of statements, the true-false ratings in the control (4.97), real (5.53), and fake conditions (4.64) conditions differed significantly, indicating that there might be some sort of social influence. The statement by condition interaction was significant, \(F(2,234)=11.28, p<.001\), indicating that the pattern of social influence differed depending on the statement type.

Given the significant interaction between statement and condition, we further analyzed social influence within each statement type using a one-way ANOVA with condition as independent variable and true-false ratings as a dependent measure. Within true and debatable statements, the control, real, and invented conditions differed significantly in the true-false ratings, \(F(2,117)=18.29, p<.001\), and \(F(2,117)\) \(=5.601, p<.001\), respectively. However, within false statements, there was no significant difference across the three conditions, \(F<1\).

Experiment 1 tested the first hypothesis that the true-false judgment of debatable statements would be most prone to social influence. This hypothesis was partially supported. Although, as predicted, people adopted the collective opinion when making true-false judgments for debatable statements, the judgments of true statements resulted in the same pattern. As predicted, there was little social influence for the true-false judgments for false statements. As figure 4 shows, for true and debatable statements, the ratings increased in the real condition and decreased in the inverted
condition relative to the ratings in the control condition, indicating a positive social influence. The true-false judgments followed the collective opinion for true and debatable statements.


Figure 4. Results of Experiment 1 are shown. Error bars represent \(95 \%\) confidence intervals.

\section*{Experiment 2}

The focus of Experiment 1 was on the effect of collective opinion on people's true-false judgments. The main purpose of Experiment 2 is to examine how collective opinion affects people's likelihood of sharing information in a social media environment. The results from Experiment 2 will also address the question of whether people are more likely to share true, debatable, or false information, without knowing collective opinion.

\section*{Method}

Participants In Experiment 2, 220 workers of Amazon's Mechanical Turk completed the experiment for a nominal fee. Their mean age was 28 .
Materials The same as Experiment 1.
Design and Procedure The same as Experiment 1 except that, in Experiment 2, the question asked how likely it is that subjects will share the information, and the collective opinion indicated information regarding collective likelihood. Figures 5, 6, and 7 show an example of the screen presented to the subjects in Experiment 2's control, real, or invented condition, respectively.

Background: People are bombarded with more information than they can process in social media, such as Twitter and Facebook. Although people can find a lot of useful information in social media, they also encounter a significant amount of unverified information. Please help us improve social media environments. Your responses are valuable!
Question A: Does the statement below contain the word fiber? Yes \({ }_{\circ}\) No
Oatmeal contains soluble fiber, which reduces your low-
density lipoprotein (LDL), the "bad" cholesterol.
Question B: Imagine that you are reading the statement above in social media, such as Twitter and Facebook. How likely is it that you will share this information?
```

***If you want to know more about the above statement, please email us (intuitive.analytic@gmail.com)***
If this is your first HIT from this batth, please complete the following:
IN
Thank you for your participation:

```

Figure 5. Example of Experiment 2's control condition

Question A: Does the statement below contain the word fiber? Yes \({ }^{\text {No }}\)
Oatmeal contains soluble fiber, which reduces your low-
density lipoprotein (LDL), the "bad" cholesterol.
Question B: Imagine that you are reading the statement above in social media, such as Twitter and Facebook. How likely is it that you will share this information?
Likelihood of sharing by majory of others like you: 6.5
not at all

Figure 6. Example of Experiment 2's real condition

Question A: Does the statement below contain the word fiber? Yes No
Oatmeal contains soluble fiber, which reduces your low-
density lipoprotein (LDL), the "bad" cholesterol.
Question B: Imagine that you are reading the statement above in social media, such as Twitter and Facebook. How likely is it that you will share this information? Likelihood of sharing by majory of others like you: 3


Figure 7. Example of Experiment 2's invented condition

\section*{Results and Discussion}

All participants were included in the analyses. The main interest of Experiment 2 was whether and how collective opinion might influence people's sharing decisions. Figure 8 shows the overall pattern of the results in Experiment 2.

A 3 (statement: true, false, debatable) by 3 (condition: control, real, invented) ANOVA, with people's likelihood of sharing as a dependent measure revealed a significant main effect of statement, \(F(2,117)=12.99, p<.001\). Collapsing across condition, the likelihood of sharing true (4.27), debatable (4.39), and false statements (3.96) differed significantly. There was also a significant main effect of condition, \(F(2,396)=10.55, p<.001\). Collapsing across statement, the likelihood of sharing in the control (4.29), real (4.46), and invented conditions (3.87) differed significantly, indicating the presence of social influence. Moreover, there was a significant statement by condition interaction, \(F(2,234)=11.16, p<.000\). The pattern of social influence differed depending on the type of statement.

We analyzed social influence within each type of statement using a one-way ANOVA with condition as independent variable and the likelihood of sharing as a dependent measure. Within true, debatable, and false statements, the control, real, and invented conditions differed significantly in the likelihood of sharing, \(F(2,156)\) \(=13.92, p<.001\), and \(F(2,156)=10.58, p<.001\), and \(F(2\), \(156)=4.399, p=0.005\), respectively.

The results from Experiment 2 show that collective opinion affects people's likelihood of sharing information, and support the hypothesis that a positive social influence takes place for true, debatable, and false statements. In Experiment 2, people's intention to share information followed the collective opinion. In addition, Experiment 2's results partially supported the third hypothesis that debatable statements would result in a higher likelihood of sharing than false and true statements. Although people were more likely to share debatable statements than false statements in the control condition, the likelihood of sharing debatable and true statements did not differ significantly, as can be seen in Figure 8.


Figure 8. Results of Experiment 2 are shown. Error bars represent \(95 \%\) confidence intervals.

\section*{General Discussion and Implication}

In the current paper, we reported results from two experiments that examined how collective opinion might influence people's true-false judgments and information sharing decisions. In Experiment 1, we found that, for false statements, collective opinion had little influence on people's true-false judgments, but, for true and debatable statements, their judgments were biased toward collective opinion. In Experiment 2, we learned that the likelihood of sharing the true, debatable, and false statements followed the collective opinion, and that people were less likely to
share false statements than debatable or true ones without collective opinion. The current results reveal that whether or not people adopt collective opinion in social media contexts depends on the type of judgment they make and the type of information they evaluate.

In the real social media environments, collective opinion is updated constantly. Future research may examine several iterations of the current experiments, in which collective opinion is updated after each run based on the ratings of the previous run. By doing so, we can study the evolution of people's ratings and collective opinion. Do the ratings converge or diverge after several iterations? The ratings might diverge when using mode as collective opinion as in the current work, but they might converge when collective opinion takes the form of median or mean.

Another characteristic of social media is that there are diverse kinds of information. The focus in the current work was information related to health advice. Future work should extend the current findings to other kinds of statements. We are currently examining the role collective opinion plays in sharing information related to natural disasters, such as Hurricane Sandy in 2012 and the Great East Japan Earthquake in 2011.

Finally, although we used rating scales in the current work, people's information sharing decisions in social media environments are binary. For example, there is a retweet button in Twitter. One extension of the current work might be to measure actual behavior by creating a 'share' button that sends a message to an associated email account when pressed. We can create a button and a Gmail account for each condition in an experiment, and ask subjects to click the button if they want to share the message. Although the measure of intent using a rating scale can provide us information about the strength of intent, it may or may not translate to actual behavior. When stimuli are tweets, one can examine whether or not there is a positive correlation between the likelihood of sharing and the actual re-tweeting number in Twitter. On the other hand, clicking of a share button cannot capture information about how much people want to share.

In conclusion, better understanding of how people make judgment and decision in social media contexts is important. People use social media technologies to share information everyday, even during responses to disasters. More research along this line can help the development of a set of recommendations for enhancing people's social media literacy and for improving the design of socialcomputational systems to improve the quality of information in social media, and more generally, to increase the productivity and wellbeing of our society. We hope that the present work can stimulate further investigation of social influence in social media environments.

\section*{Acknowledgments}

This material is based upon the work supported by the National Science Foundation under Grant No. IIS-1138658.

\section*{References}

Allport, G.W., \& Postman, L. (1947). The Psychology of Rumor. Henry Holt and Company, New York, New York.
Anthony, S. (1973). Anxiety and Rumor. The Journal of Social Psychology, 89, 91-98.
Bandari, R., Asur, S., \& Huberman, B. A. (2012). The Pulse of News in Social Media: Forecasting Popularity. Arxiv preprint arXiv:1202.0332.
Berger, J., \& Heath, C. (2007). Where consumers diverge from others: Identity-signaling and product domains. Journal of Consumer Research, 34, 121-134.
Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D., Marlow, C., Settle, J. E., \& Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. Nature, 489, 295-298.
Buhrmester, M., Kwang, T., \& Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? Perspectives on Psychological Science.
Cha, M., Haddadi, H., Benevenuto, F., \& Gummadi, K.P. (2010). Measuring user influence in Twitter: The million follower fallacy. In Int'l AAAI Conference on Weblogs and Social Media (ICWSM).
Cialdini, R. B., \& Goldstein, N. J. (2004). Social influence: Compliance and conformity. Annual Review of Psychology, 55, 591-621.
DiFonzo, N., \& Bordia, P. (2007). Rumor psychology: Social and organizational approaches. American Psychological Association, Washington.
Festinger, L. (1954). A theory of social comparison processes. Human Relations, 7, 117-140.
Fragale, A., \& Heath, C. (2004). Evolving informational credentials: The (mis)attribution of believable facts to credible sources. Personality and Social Psychology Bulletin, 30, 226-236.
Glushko, R. J., Maglio, P. P., Matlock, T., \& Barsalou, L. W. (2008). Categorization in the wild. Trends in Cognitive Sciences, 12, 129-135.
Gureckis, T. M., \& Goldstone, R. L. (2006). Thinking in groups. In S. Harnad \& I. Dror (Eds.), Distributed cognition: Special issue of pragmatics \& cognition, 14 (pp. 293-311). Amsterdam, The Netherlands: John Benjamins.
Ha, S., \& Ahn, J. (2011). Why are you sharing others' Tweets?: The impact of argument quality and source credibility on information sharing behavior. ICIS 2011 Proceedings.
Heath, C. (1996). Do people prefer to pass along good news or bad news? Valence and relevance of news as a predictor of transmission propensity. Organizational Behavior and Human Decision Processes, 68, 79-94.

Huberman, B., Romero, D., \& Wu, F. (2009). Social networks that matter: Twitter under the microscope. First Monday, 14, 1-5.
Kempe. D, Kleinberg. M. J, \& Tardos. E, (2003). Maximizing the spread of influence through a social network, KDD, pp. 137-146, ACM.
Kim. S, Kim. S, \& Cho. H, (2011). Predicting the virtual temperature of web-blog articles as a measurement tool for online popularity, in IEEE 11th International Conference on Computer and Information Technology (CIT), 31, pp. 449-454.
Kwak, H., Lee, C., Park, H., \& Moon, S. (2010). What is Twitter, a social network or a news media? Proceedings of the 19th international conference on World Wide Web WWW'10.
Lerman. K., \& Hogg. T, (2010). Using a model of social dynamics to predict popularity of news. WWW, 621-630, ACM.
Mason, W., \& Suri, S. (2011). A Guide to Behavioral Experiments on Mechanical Turk. Behavior Research Methods.
Paolacci, G., Chandler, J., \& Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. Judgment and Decision Making, 5, 411-419.
Rene, P., Antonios, G., \& Frank, S. (2012). Emotional divergence influences information spreading in Twitter. AAAI 2012.
Sakamoto, Y. (2010). The impact of collective opinion on online judgment. Annual Meeting of the Cognitive Science Society.
Sakamoto, Y. (2012). Following trendsetters: Collective decisions in online social networks. HICSS 2012.
Sakamoto, Y., Ma, J., \& Nickerson, J. V. (2009). 2377 people like this article: The influence of other's decisions on yours. In N. Taatgen, H. van Rijn, L. Schomaker, and J. Nerbonne (Eds.), Proceedings of the 31st Annual Meeting of the Cognitive Science Society.
Sherif, M. (1935). A study of some social factors in perception. Archives of Psychology, 27, 1-60.
Shibutani, T. (1966). Improvised news: A sociological study of rumor. Indianapolis, IN: Bobbs-Merrill.
Snyder, C., \& Fromkin, H. (1980). Uniqueness: The human pursuit of difference. New York: Plenum.
Soll, J. B., \& Larrick, R. P. (2009). Strategies for revising judgment. Journal of Experimental Psychology: Learning, Memory and Cognition, 35, 780-805.
Tian, K., Bearden, W., \& Hunter, G. (2001). Consumers' need for uniqueness: Scale development and validation. Journal of Consumer Research, 28, 50-66.
Xin, S., Ying D., \& Jerome B. (2012). Multiple spreaders affect the indirect influence on Twitter. WWW 2012 Companion, Lyon, France.

\title{
Different Strategies in Solving Series Completion Inductive Reasoning Problems: An fMRI and Computational Study
}

\author{
Peipeng Liang, Xiuqin Jia (ppliang1979@gmail.com, xiuqin.jia@gmail.com) \\ Department of Radiology, Xuanwu Hospital, Capital Medical University, Beijing, China \\ Brain Key Lab of Magnetic Resonance Imaging and Brain Informatics, Beijing, China \\ Niels A. Taatgen (niels@ai.rug.nl) \\ Department of Artificial Intelligence, University of Groningen, Groningen, The Netherlands \\ Ning Zhong (zhong@maebashi-it.ac.jp) \\ Department of Life Science and Informatics, Maebashi Institute of Technology, Maebashi 371-0816, Japan \\ Kuncheng Li (lkc1955@gmail.com) \\ Department of Radiology, Xuanwu Hospital, Capital Medical University, Beijing, China \\ Brain Key Lab of Magnetic Resonance Imaging and Brain Informatics, Beijing, China
}

\begin{abstract}
Neural correlate of human inductive reasoning process is still unclear. Number series and letter series completion are two typical inductive reasoning tasks, and with common core component of rule induction. Previous studies have demonstrated that different strategies are adopted in number series and letter series completion tasks even the underlying rules are identical. In the present study, we examined cortical activation as a function of two different reasoning strategies for solving series completion tasks. The retrieval strategy, used in number series completion tasks, involves direct retrieving of arithmetic knowledge to get the relations between items. The procedural strategy, used in letter series completion tasks, requires counting a certain number of times to detect the relations linking two items. The two strategies require essentially the equivalent cognitive processes, but have different working memory demands (the procedural strategy incurs greater demands). The procedural strategy produced significant greater activity in areas involved in memory retrieval (dorsolateral prefrontal cortex, DLPFC) and mental representation/maintenance (posterior parietal cortex, PPC). An ACT-R model of the tasks successfully predicted behavioral performance and BOLD responses in DLPFC and PPC. The present findings support a general-purpose dualprocess theory of inductive reasoning regarding the cognitive architecture.
\end{abstract}

Keywords: Number series; Letter series; Inductive reasoning; Adpative control of thought-rational (ACT-R)

\section*{Introduction}

Inductive reasoning, defined as inferring a general rule or relation from specific elements, is traditionally considered as one of the most important constitutes of human intelligence (Spearman, 1923). Several studies were performed to investigate the neural underpinnings of human inductive reasoning using different types of tasks, including sentential (e.g., House cats have 32 teeth; Lions have 32 teeth; therefore, all felines have 32 teeth. Was the given conclusion plausible given the premises? Goel et al., 1997; 2004), figural (e.g., infer the rule underlying the figural
stimuli consisting of novel animals. Goel et al., 2000) and numerical (e.g., number series completion tasks. Liang et al., 2007; Yang et al., 2009; Zhong et al., 2011; Jia et al., 2011) tasks. The recruitment of fronto-parietal regions together with their left lateralization are convergent reported in most of these studies, however, the detailed activation patterns are modulated by the heterogeneousness of the experimental task. Neural correlate of human inductive reasoning process is still unclear, and needs more experimental studies.

The series completion problem, including letter series (e.g., c e g ?) and number series (e.g., 357 ?), is a kind of typical inductive reasoning task (Pelligrino, 1985; Thurstone, 1938; Thorndike et al., 1986) and always used in general fluid intelligence (Gf) test (Cattel, 1963; Hayslip et al., 1995; Johnston et al., 2010; Redick et al., 2012). However, number series and letter series completion tasks are solved differently. Evidences demonstrated that different strategies (by definition, a strategy is a goal-directed procedure under the deliberate control of the participant (Rosenberg-Lee and Anderson, 2009)) were employed in number series and letter series completion tasks, in which each item in one has a same-rule counterpart in the other (Quereshi, 2001), and number series tasks are easier and more familiar than letter series tasks (Quereshi \& Seitz, 1993; Quereshi \& Smith, 1998). This has been confirmed by a recent pilot study with post-test oral report in our group, in which subjects are required to solve the two kinds of series completion tasks comprising items based on identical rules. A "retrieval" strategy is used in solving number series tasks, in which the relation between two adjacent items can be directly attained by retrieving the corresponding arithmetic fact from long-term memory (e.g., \(1113: 11+2=13\), thus, the rule is +2 ). As to letter series tasks, a "procedural" strategy is adopted, in which participants require to step-wise count the corresponding adjacent letter of an item in order to find the relation linking two items (e.g., k n: k ... l ... m ... n; thus, the rule is +3 ). The investigation of strategies using imaging has the potential to enrich our understanding the
neural substrates of inductive reasoning, in terms of locus, level, and duration of activity.

The retrieval and procedural strategy involve performing equivalent cognitive processes (retrieving of declarative memory to detect the relation between the adjacent two items, internal representation and maintenance, and response output), thus engage the same brain areas. However, the working memory demands differ between the two strategies. In the retrieval strategy (e.g., 11 13), participants directly get the relation (e.g., +2 ) between the adjacent two items by retrieving once the corresponding arithmetic fact (e.g., \(11+2=13\) ); While in the procedural strategy (e.g., k m), participants require to perform twice retrieval step by step (e.g., l is next to k , and m is next to l ) and twice internal maintaining/updating a counter, and then the rule can be determined according to the counter. In this way, the procedural strategy incurs many more working memory demands than the retrieval strategy. Thus, the two strategies should differentially engage brain areas that are sensitive to working memory load. It is predicted that the two strategies can be differentiated by the extent of activity within the same brain areas, including the left prefrontal cortex, recruited in memory retrieval, and posterior parietal cortex, involved in mental representation (see a summarization in Anderson et al., 2008).

The goal of the current study was to employ computational cognitive modeling to make specific predictions about the strategy differences. Specifically, based on the aforementioned specification, we expect to demonstrate that the strategy difference can be distinguished neurally by differential engagement of the same brain areas. To make our predictions more precise (in terms of the timing and level of activity), we plan to build computational models of the experimental tasks using the adaptive control of thought-rational (ACT-R) cognitive architecture (Anderson, 2007). ACT-R proposes that cognitive process is the result of the independent activity of distinct modules which are coordinated by a central production system. The ACT-R models automatically generate predictions for activity in the ACT-R modules, which we could then compare with activity in the brain areas of interest. For the tasks in the present study, the differential engagement was mainly due to the differences in retrieval and maintenance demands between the two strategies. Two modules were thus of particular interest in this study: the retrieval module, which is responsible for the retrieval of declarative memories and linked to the lateral inferior prefrontal cortex (LIPFC) and the imaginal module, which is responsible for the encoding and maintenance of internal representation of the problems and linked to the posterior parietal cortex (PPC) (Anderson, 2007; Qin et al., 2004). As the procedural strategy incurs more demands of retrieval and maintenance, thus, we expected LIPFC and PPC have a greater response to the procedural strategy (used in letter series tasks) than the retrieval strategy (used in number series tasks). Together, we will test the two competitive theories empirically and computationally.

\section*{Materials and Methods}

\section*{Subjects}

Twenty-three paid healthy undergraduate and postgraduate students (11 females; \(24.1 \pm 3.7\) years old) participated in the experiment. Writhen informed consent was obtained from each participant and this study was approved by the Ethics committee of Xuanwu Hospital, Capital Medical University.

\section*{Stimuli and experimental design}

Four kinds of tasks were organized into a \(2 \times 2\) factorial design (Table 1). The first factor was Content, consisting of two levels, number-related (24 inductions and 24 baselines) and letter-related ( 24 inductions and 24 baselines) tasks. The second factor was Task in which the first level was the induction condition consisting of series completion tasks (24 number series inductions and 24 letter series inductions) and the second level was the baseline condition (24 number judgment baselines and 24 letter judgment baselines). This yielded four types of tasks: number series induction (NumIR), letter series induction (LetIR), number judgment baseline (Is10) and letter judgment baseline (IsJ). In particular, interferential tasks, which are identical in pattern to inductions but without common rules (e.g., "1 38 " or "a c f"), were included into NumIR and LetIR tasks based on a pilot study. There were twelve interferential tasks within 60 induction tasks in the current study, with six in NumIR tasks and another six in LetIR tasks.

For all kinds of tasks, there were three sequentially presented numbers or letters (e.g., "1 35 " or "a c e"). Number-related tasks and letter-related tasks were matched for magnitudes and operations. All the letters involved were lowercase and within the range of a-z. Correspondingly, all the numbers involved were within the range of 1-26. Half of NumIR and LetIR tasks was forward (e.g., "1 35 " or "a c e"; the rule is: +2 ) while another half was backward (e.g., "13 9 5 " or "m i e"; the rule is: -4). In the answer options of NumIR and LetIR tasks, the distances between the correct and the false answer were less than 3. Half of Is10 and IsJ tasks were with the answer of "yes" while another half were with the answer of "no".

Table 1: Examples of experimental tasks.
\begin{tabular}{llllc}
\hline & Task & \multicolumn{3}{c}{ Options } \\
NumIR & 111315 & A. 18 & B. 17 & B \\
LetIR & t s r & A. q & B. p & A \\
Is10 & 142310 & A. yes & B. no & A \\
IsJ & n w j & A. no & B. yes & B \\
\hline
\end{tabular}

\section*{Stimuli Presentation}

Stimuli from all conditions were organized into three sessions and presented randomly in an event related design. The order of sessions was counterbalanced among subjects. The beginning of a trial was signaled by a cue of the task type ("Finding a rule" for NumIR and LetIR tasks, "Is there
' 10 '" for Is10 tasks and "Is there ' J '" for IsJ tasks) for 2 s. The numbers/letters then appeared on the screen with the first number/letter appearing at 2 s , the second at 2.6 s , and the last at 3.4 s . After the appearance of the third number/letter, subjects were instructed to press the left or right button (counterbalanced among subjects) when the answer was acquired. Three numbers/letters would remain on the screen for at most 6.6 s (since the presentation of the third number/letter), or the button-pressing response within 6.6 s would stop it. Subsequently, two answer options were displayed and subjects were asked to choose the correct answer by pressing the corresponding buttons (left button for "A"). Subjects were instructed to respond as accurately and quickly as possible and move to the next trial if the stimuli advanced before they could respond. The interstimulus interval (ISI) was 8 s and quick responses would leave more rest time within the trial. Thus, reaction times (RT) were recorded based on the first button response and accuracies were acquired based on the second button response.

\section*{MRI data acquisition}

Scanning was performed on a 3.0 Tesla MRI system (Siemens Trio Tim; Siemens Medical System, Erlangen, Germany) with a 12 -channel phased array head coil (see details in Liang et al., (2007)). This acquisition sequence generated 364 volumes for each session. The scanner was synchronized with the presentation of every trial.

\section*{Data Analysis}

\section*{Data preprocessing}

Data were analyzed using SPM5 software (http://www.fil.ion.ucl.ac.uk) and the preprocessing steps are identical to Jia et al. (2011).

\section*{FMRI analysis}

Only correct response trials were included in data analysis (wrong response trials and interferential trials were not included). The epoch of interest is the duration between the presentations of the first number/letter to the first button response. The BOLD signal was modeled using the canonical HRF with time derivative and the RT as duration trial by trial.

For confirmatory analysis, region of interest (ROI) analyses focused on two predefined functional regions in ACT-R, LIPFC (TAL., -43, 23, 24) and PPC (TAL., -23, 63, 40) (Anderson, 2007). The statistical results were based on the beta-values (of general linear model implemented in SPM5) averaged within the two ROIs.

For exploratory analysis, we are primarily interested in the conjunction analysis to explore activations common to both strategies [(NumIR-Is10) and (LetRI-IsJ)]. The Task by Content interaction comparisons ([(NumIR-Is10)-(LetIRIsJ)] and [(LetIR-IsJ)-(NumIR-Is10)]) were also executed to reveal the areas specific to each strategy. The activation
reported survived a voxel-level intensity threshold of p < 0.05 corrected for multiple comparisons.

\section*{Results}

\section*{Behavioral performance}

We carried out analyses of variance for two factors: Content (number vs. letter) and Task (induction vs. baseline) on both RT and accuracy. Behavioral scores indicated that subjects performed the task in the expected manner. The main effects of Content and Task were significant for RT and accuracy: Response to induction tasks was significantly longer ( \(\mathrm{F}(1,22\) ) \(=70.10, \mathrm{p}<0.001\) ) and less accurate \((\mathrm{F}(1,22)=61.56, \mathrm{p}<\) 0.001 ) than that of baseline tasks; Response to letter tasks was significantly longer \((\mathrm{F}(1,22)=68.38, \mathrm{p}<0.001)\) and less accurate \((\mathrm{F}(1,22)=69.94, \mathrm{p}<0.001)\) than that of number tasks. The interaction effect between Task and Content was significant for RT \((\mathrm{F}(1,22)=69.24, \mathrm{p}<0.001)\) and accuracy \((\mathrm{F}(1,22)=66.75, \mathrm{p}<0.001)\) : The difference between LetIR and IsJ (i.e., LetIR-IsJ) (RT: \(\mathrm{F}(1,22)=30.59\), \(p<0.001\); accuracy: \(F(1,22)=73.81, p<0.001)\) than the difference between NumIR and Is10 (i.e., NumIR-Is10) (RT: \(F(1,22)=6.39, p<0.05\); accuracy: \(F(1,22)=2.89, p=0.10)\). This indicated that the significant interaction effect is primarily driven by the strategy difference, i.e., the difference between LetIR and NumIR.

\section*{Imaging results}

Confirmatory analysis We performed repeated measures Task by Content for the two ROIs. The two ROIs showed the identical patterns. The main effect of Content \((F(1,22)=\) 35.46, \(p<0.001\) for LIPFC and \(F(1,22)=14.23, p<0.001\) for PPC), Task \((F(1,22)=67.10, p<0.001\) for LIPFC and \(F(1,22)=35.45, p<0.001\) for PPC) and the Task by Content interaction effect \((F(1,22)=29.64, p<0.001\) for LIPFC and \(F(1,22)=4.60, p<0.001\) for PPC) were all significant. The post-hoc multiple comparisons showed that the interaction effect was driven by the strategy difference, i.e., the difference between letter series reasoning tasks and number series reasoning tasks \((F(1,22)=46.58, p<0.001\) for LIPFC and \(F(1,22)=13.51, p<0.001\) for PPC). There were no significant difference between the two kinds of baselines (i.e., IsJ versus Is10; \(F(1,22)=3.98, p=0.06\) for LIPFC and \(F(1,22)=1.85, p=0.19\) for PPC).
Exploratory analysis Conjunction analysis indicated that both strategies showed common activation in the left dorsolateral prefrontal cortex (DLPFC, BA 46, 9) and bilateral superior parietal lobe (SPL, BA 7) extending to left inferior parietal lobe (IPL, BA 40) (Figure 1). Regions specific to LetIR [(LetIR-IsJ)-(NumIR-Is10)] were found in the left middle frontal gyrus (BA 9, 46), bilateral superior frontal gyrus extending into anterior cingulate cortex/medial frontal gyrus (BA 6, 8, 9, 32), inferior frontal gyrus/insula (BA 13, 45, 47), and bilateral superior parietal cortex (BA 7). No significant activations were found specific to NumIR [(NumIR-Is10) - (LetIR-IsJ)].


Figure 1: Activations common to both strategies.

\section*{The model}

Having now reviewed the result of the experiment, we come to the question whether we can understand them in the frame work of the ACT-R theory. The model we constructed primarily depends on the visual module to perceive the stimuli, the manual module to respond, the retrieval module to retrieve a fact from memory, and the imaginal module to encode and update its stored representation. Figure 2 presents the sequences of activity in the four modules, for a " +2 " NumIR problem or a "backward next" relation LetIR problem solved by each of the strategies. Specifically, the underlying productions driving these modules are not shown. As the fronto-parietal network plays an important role in inductive reasoning, the retrieval and imaginal module are of special interest in the present study, in the following, we would illustrate the prediction of BOLD responses in the two modules.

The fit of the predictions of the model to the RT and ACC data are presented in Figure 3. The parameters were estimated to fit the behavioral data: a factor that scaled the time to retrieve a declarative memory fact ( 0.3 sec ) and the time to modify the contents of the imaginal module (i.e., 0.2 sec ). This leads to a predicted effect size ( 605.42 ms for NumIR, 2116.17 ms for LetIR, 551.25 ms for baseline conditions) that is highly similar to the observed effect size ( 586.4 ms for NumIR, 2221.4 ms for LetIR, 530.4 ms for Is10, and 606.1 ms for IsJ). Only the deviation of data is presented because the model makes identical predictions. It can be seen that the model produced a reasonable fit to the behavioral data.

In order to generate BOLD predictions, we convolved the module activity with a gamma function. As is typical (e.g., Boyton et al., 1996; Glover, 1999), if the module is engaged, it will produce a BOLD response \(t\) time units later according to the function:
\[
H(t)=m(t / s)^{\alpha} e^{-(t / s)}
\]

The parameter \(m\) is the magnitude parameter and determines the height of the function; the parameter \(s\) is the scale parameter and determines the time scale, and the parameter is the shape parameter and determines the narrowness of the function. The cumulative BOLD responses in a particular module is the sum of the individual BOLD responses driven by a module's activities. This can be modeled by convolving the hemodynamic response, \(H(t)\), with a demand
function, \(D(\mathrm{t})\), which has a value of 1 when the module associated with that region is active, and 0 otherwise:
\[
B(\mathrm{t})=\int_{0}^{t} D(x) H(t-x) d x
\]

Once the timings of buffers actions are all set, we can predict the BOLD functions by estimating the magnitude parameter \(m\), the exponent \(\alpha\), and the latency scale \(s\) for each brain region. The estimates of these parameters and measurement of the quality of the prediction are given in Table 2. Figure 4 displays percent change of BOLD response (relative to the baseline defined by the average BOLD response of the first two scans and the last two scans), along with the prediction of the ACT-R model to be presented.

Table 2: ACT-R parameters and the BOLD predictions.
\begin{tabular}{lcc}
\hline & Imaginal & Retrieval \\
\hline Exponent \((\boldsymbol{\alpha})\) & 3 & 3 \\
Scale \((s)\) & 2.2 & 1.8 \\
Magnitude & & \\
\(\quad M \Gamma(\alpha+1)^{*}\) & 5.6 & 5.5 \\
Correlation \((r)\) & 0.94 & 0.92 \\
\hline
\end{tabular}


Figure 2: A schematic representation of the ACT-R model's solution to the number- and letter-series problem of "11 1315 " and "t s r", solved by retrieval strategy and procedural strategy, respectively.

\section*{Discussion}

The primary purpose of the current study was to investigate the neural correlates of the two cognitive strategies in series completion inductive reasoning tasks. Both confirmatory and exploratory analysis indicated that DLPFC (L.>R.) and PPC (bilaterally) were commonly recruited in the two kinds of tasks, but with different extent. Additionally, the exploratory analyses identified additional regions in dorsomedial prefrontal cortex (DMPFC) and ventrolateral prefrontal cortex (VLPFC) was active for the procedural strategy but not for the retrieval one. These results suggest that some aspects of the behavioral signatures of strategies may be recovered from imaging data. We constructed the computational model to simulate participants' behavior. The results showed that there was reasonable fitness between the model prediction and the empirical data.


Figure 3: Data (red solid lines) and model fits (black dashed lines) for NumIR, LetIR, Is10, and IsJ.

In the current study, the significant co-activation of the left DLPFC and bilateral PPC is identified to be strategyindependent. These results replicated a recent fMRI study of number series completion task in our group (Jia, et al., 2011), and were consistent with previous studies of sentential inductive reasoning task (Goel and Dolan 2004) and figural inductive reasoning tasks (Induction minus Perceptual baseline; Goeland Dolan, 2000). This implies that these two regions may be the key regions involved in inductive reasoning. Moreover, the left DLPFC is more specific to sentential induction, as contrast to deduction (Goel and Dolan, 2004). The activity of the DLPFC is right lateralized in Goel et al. (2000), but is left lateralized in the other studies (Jia, et al., 2011; Goel and Dolan, 2004) and the current study, which may be ascribed to the different kinds of experimental materials (figure versus number, sentences).

As to the functional role of DLPFC and PPC, different interpretations were proposed. In the domain of inductive reasoning, the left DLPFC was related to the use of world
knowledge in the generation and evaluation of hypotheses (Goel and Dolan, 2004), rule identification and extrapolations (Jia, et al., 2011), and the fronto-parietal network was more specific to rule identification (Jia, et al., 2011). (Note: In Goel and Dolan (2000), the authors are interested in bilateral hippocampus, which is specifically associated with rule inference, and right lateral orbital prefrontal cortex, which is associated with the task by difficulty interaction. The former is interpreted in terms of semantic encoding of novel stimuli, and the latter in terms of hypothesis selection. While the DLPFC (R.>L.) and bilateral PPC, identified in Rule Induction minus Perceptual Baseline, is not explained of their functional roles.) It is not surprise of these different explanations, as different domain and context were situated in.


Figure 4: The BOLD signals obtained for the prefrontal and parietal regions for NumIR and LetIR condition and the prediction of the ACT-R model of the task.

In this study, with the help of computational modeling, we made a more general account. The DLPFC and PPC regions identified in the exploratory analysis, are largely overlapped the regions involved in the ACT-R LIPFC and ACT-R PPC. Therefore, consistent with the explanation in ACT-R theory, it was inferred that the left DLPFC is associated with memory retrieval of semantic information/knowledge and PPC is associated with mental representation of problem states (Anderson, 2007; Danker \& Anderson, 2007). In the current study, although the procedural strategy and the retrieval strategy commonly recruited the left DLPFC and bilateral PPC, the strategy variation can be distinguished by the extent of cortical activity within these regions. The procedural strategy incurs more working memory demands (both the retrieval and maintenance demands) than the retrieval strategy. The ACTR model fitted well with the experimental behavioral and imaging data, which demonstrated the validity of this general account.

In summary, by coupling this empirical work with computational modeling, we have deepened our understanding of what constitute the strategy variation (the retrieval strategy versus the procedural strategy) and quantify their predictions. These results support the generalpurpose dual-process theory of inductive reasoning.

\section*{Acknowledgments}

This work was supported by the Natural Science Foundation of China (No. 61105118), Beijing Nova Program (No. Z12111000250000).

ACT-R code can be accessed by logging in ccmureasoninglab@gmail.com (userid: ccmureasoninglab; password: peipengliang).

\section*{References}

Anderson JR. (2007). How can human mind occur in the physical universe ? New York : Oxford University Press.
Anderson JR, Carter CS, Fincham JM, Qin YL, Ravizza SM, Rosenberg-Lee M. (2008). Using fMRI to test models of complex cognition. Cognitive Science, 32, 1323-1348.
Barbey AK, Barsalou LW. (2009). Models of reasoning and problem solving. In Squire L, editor, Encyclopeida of Neuroscience. Oxford: Academic Press. 8, 35-43.
Boyton GM, Engel SA, Glover GH, and Heeger DJ. (1996). Linear systems analysis of functional magnetic resonance imaging in human V1. Journal of Neuroscience, 16, 42074221.

Chein JM, Moore AB, Conway ARA. (2011). Domaingeneral mechanisms of complex working memory span. Neuroimage, 54, 550-559.
Danker JF and Anderson JR. (2007). The roles of prefrontal and posterior parietal cortex in algebra problem solving : A case of using cognitive modeling to inform neuroimaging data. Neuroimage, 35, 1365-1377.
Feeney A, Shafto P, Dunning D. (2007). Who is susceptible to conjunction fallacies in category-based induction? Psychonomic Bulletin \& Review, 14(5), 884-889.
Glover GH. (1999). Deconvolution of impulse response in event-related BOLD fMRI. Neuroimage, 9, 416-429.
Goel V, Dolan R. (2000). Anatomical segregation of component processes in an inductive inference task. Journal of Cognitive Neuroscience, 12, 1-10.
Goel V, Dolan R. (2004). Differential involvement of left prefrontal cortex in inductive and deductive reasoning. Cognition, 93, B109-B121.
Goel V, Gold B, Kapur S, Houle S. (1997). The seats of reason: A localization study of deductive \& inductive reasoning using PET ( \(\mathrm{O}^{15}\) ) blood flow technique. NeuroReport, 8, 1305-1310.
Jia XQ, Liang PP, Lu J, Yang YH, Zhong N, Li KC. (2011). Common and dissociable neural correlates associated with component processes of inductive reasoning. Neuroimage, 56, 2292-2299.
Johnston A, Gradisar M, Dohnt H, Billows M, Mccappin S. (2010). Adolescent sleep and fluid intelligence performance. Sleep and Biological Rhythmes, 8, 180-186.

Liang PP, Zhong N, Lu SF, Liu JM, Yao YY, Li KC, Yang YH. (2007). The neural mechanism of human numerical inductive reasoning process: A combined ERP and fMRI study. LNAI 4845, Springer, 223-243.
Pelligrino JW. (1985). Inductive reasoning ability. In Sternberg RJ (Eds.), Human abilities: An informationprocessing approach. New York: Freeman.
Prado J, Chadha A, Booth JR. (2011). The brain network for deductive reasoning: a quantitative meta-analysis of 28 neuroimaging studies. J Cogn Neurosci., 23(11), 34833497.

Qin YL, Cameron SC, Silk EM, Stenger VA, Fissell K, Goode A, Anderson JR. (2004). The change of the brain activation patterns as children learn algebra equation solving. PNAS, 101(15), 5686-5691.
Quereshi MY. (2001). Factorial similarity of letter and number series comprising identical rules. Current Psychology: Developmetal, Learning, Personality, Social Spring, 20(3), 278-286.
Quereshi MY, Seitz R. (1993). Identical rules do not make letter and number series equivalent. Intelligence, 12, 110119.

Quereshi MY, Smith H. (1998). Reasoning ability in older adults measured through letter and number series. Current Psychology, 17, 20-27.
Redick TS, Shipstead Z, Harrison TL, Hicks KL, Fried DE, Hambrick DZ, Kane MJ, Engle RW. (In press). No evidence of intelligence improvement after working memory training: A randomized, placebo controlled study. J. Exp. Psychol. Gen.

Rosenberg-Lee M, and Anderson JR. (2009). Neural correlates of arithmetic calculation strategies. Cognitive, Affective, \& Behavioral Neurosicence, 9(3), 270-285.
Spearman C. (1923). The nature of inteUigence and the principles of cognition. London: Macmillan.
Thorndike RL, Hagen EP, Sattler JM. (1986). The StanfordBinet Intelligence Scale: Fourth Edition. Guide for administering and scoring. Chicago: Riverside Publishing Company.
Thurstone LL. (1938). Primary mental abilities. Chicago: University of Chicago Press.
Yang YH, Liang PP, Lu SF, Li KC, Zhong N. (2009). The role of the DLPFC in inductive reasoning of MCI patients and normal agings: An fMRI study. Sci China Life Sci., 52(8), 789-795.
Zhong N, Liang PP, Qin YL, Lu SF, Yang YH, Li KC. (1983). Neural substrates of data-driven scientific discovery: An fMRI study during performance of number series completion task. Sci China Life Sci., 54(5), 466473.

\title{
Inductive and deductive reasoning in Obsessive-Compulsive Disorder
}

\author{
Janice H. Liew (j.liew@psy.unsw.edu.au) \\ School of Psychology, University of New South Wales \\ Sydney, 2052, Australia \\ Brett K. Hayes (b.hayes@unsw.edu.au) \\ School of Psychology, University of New South Wales \\ Sydney, 2052, Australia
}

\begin{abstract}
This study examined the hypothesis that individuals with obsessive-compulsive disorder (OCD) show a selective deficit in inductive reasoning but are unimpaired in their ability to make deductive inferences. 100 participants from an analog sample made inductive or deductive inferences about arguments that differed according to causal consistency and validity. They also completed a task examining sensitivity to the implications of diverse evidence in induction. Participants who were high or low on obsessivecompulsive symptoms showed similar patterns of induction based on causal knowledge and similar patterns of deduction. However, those with the highest level of OCD symptoms showed less of a preference for diverse evidence when evaluating inductive arguments, compared to those with the lowest level of symptoms. This difference was found across both OCD-relevant and OCDneutral items, and persisted when the effects of group differences in general ability were controlled. These results indicate that both inductive reasoning based on background knowledge and deductive reasoning are intact in individuals with high OCD-traits but the use of inductive heuristics such as evidence diversity is impaired.
\end{abstract}

Keywords: inductive reasoning, deductive reasoning, psychopathology, cognitive neuropsychiatry

\section*{Introduction}

Obsessive-compulsive disorder (OCD) is characterized by the experience of unwanted, repetitive intrusions in the form of thoughts, impulses or images. These obsessions are often accompanied by compulsions (repetitive behaviors or mental acts) that represent attempts to reduce or neutralize the marked distress that the obsessions cause.

A variety of biological, cognitive and social factors affect the onset and maintenance of OCD symptoms (Riggs \& Foa, 1993). Recently, a number of theoretical accounts have suggested that deficits in reasoning contribute to OCD symptomatology. Some of these accounts suggest that people with OCD have difficulty in reasoning about uncertain or probabilistic information (O’Connor, 2002; Pélissier \& O'Connor, 2002); that is, they show a deficit in some forms of inductive reasoning. These accounts suggest that this is a global deficit, such that people with OCD show poorer inductive reasoning compared to controls across a range of stimulus materials and content domains. Moreover, this impairment in inductive reasoning is thought to be found together with a relatively intact ability to reason deductively. Unlike induction, deduction involves the evaluation of arguments that are certain to be either valid or
invalid on the basis of logical rules (Heit, Rotello, \& Hayes, 2012).

A review of the empirical evidence however, reveals only mixed support for this "impaired induction but spared deduction" account of OCD. In support of this account, Pélissier and O'Connor (2002) found that individuals with OCD had more difficulty than controls in drawing plausible probabilistic conclusions from a set of verbal statements about everyday situations. This impairment in inductive reasoning was found together with apparently intact deductive reasoning, as measured by performance on the Wason Selection Task and ability to discriminate between valid and invalid verbal syllogisms. Moreover, this pattern was found with stimulus materials that had little connection with the content of OCD patients' obsessions.

Other work however, has suggested that the reasoning deficit in OCD extends to deduction as well as induction. For example, Simpson, Cove, Fineberg, Msetfi, and Ball (2007) found that people with OCD were poorer than controls at discriminating between logically valid and logically invalid syllogisms with OCD-neutral content.

This mixed pattern of evidence reflects, at least in part, a general problem with the methods used in previous attempts to examine inductive and deductive reasoning in people with OCD. These studies have made little attempt to match tasks that ostensibly assess inductive and deductive reasoning on dimensions such as overall task difficulty, stimulus content and task familiarity. Hence, differences in performance between nominally inductive and deductive tasks may actually reflect task-specific characteristics rather than in the cognitive processes that underlie inductive and deductive reasoning.

A major aim of the current studies was to re-examine inductive and deductive reasoning in those with OCDrelated traits and controls, using a method that addressed this major limitation of previous work. The general approach is patterned after Rips (2001) and Heit and Rotello (2010) who asked university undergraduates to evaluate a set of verbal arguments that varied in both logical validity and inductive plausibility. Crucially, different groups evaluated the set of arguments on the basis of logical necessity (deduction condition) or the overall plausibility (induction condition) of the conclusions.

Another important aim was to carry out a more exhaustive examination of possible deficits in inductive reasoning in those with OCD-related traits. A review of
research on induction in people without OCD (typically university undergraduates) suggests that such reasoning is influenced by at least two distinct factors (cf. see Hayes, Heit, \& Swendsen, 2010 for a detailed review). On the one hand, people often evaluate inductive arguments using their prior knowledge of causal or taxonomic relations between argument premises and conclusions (Medin, Coley, Storms, \& Hayes, 2003). For example, Rips (2001) found that participants were more likely to accept conclusions in inductive arguments when these were consistent with background causal knowledge (e.g. see the top right cell of Table 1), than when they were causally inconsistent (e.g. see the bottom right cell of Table 2), even though neither conclusion is logically entailed by the premise.

A second factor influencing induction is the use of general heuristics for assessing probabilistic evidence (Heit, Hahn, \& Feeney, 2005; Osherson, Smith, Wilkie, López, \& Shafir, 1990). Such heuristics include the sample size (or "monotonicity") principle in which the strength or plausibility of an inductive conclusion tends to increase with the number of instances of positive evidence observed (Osherson et al., 1990). Another important heuristic is premise diversity. All things being equal, more diverse evidence (e.g., cows and mice have property \(X\) ) is usually seen as a stronger basis for inductive generalizations (e.g., mammals have property \(X\) ) than less diverse evidence (e.g., cows and horses have property \(X\) ). Although there is some debate about the normativity of this principle (e.g., Lo, Sides, Rozelle, \& Osherson, 2002), a large body of evidence shows that most reasoners use this heuristic when evaluating inductive evidence (see Heit et al., 2005 for a review).

Previous work on inductive reasoning deficits in OCD has blurred this distinction, with some researchers examining induction based on background knowledge (e.g., Pélissier \& O'Connor, 2002), while others have examined the use of domain-general heuristics in probabilistic and inductive reasoning (e.g., Fear \& Healy, 1997). We sought to clarify the nature of inductive deficits in people with OCD by assessing those with low and high levels of OCDrelated traits on each of these two types of inductive tasks.

\section*{The Current Study}

To re-examine inductive and deductive reasoning in individuals high and individuals low on OCD symptoms, we administered two tasks. The first examined inductive and deductive reasoning using a common set of stimulus materials. Following Rips (2001), participants were asked to judge either the inductive strength or deductive validity of four types of arguments. Table 1 illustrates this design with arguments that vary in logical validity and consistency with causal knowledge. Crucially, different groups were instructed to evaluate the same set of arguments on the basis of either deductive validity or inductive plausibility.

Previous work with undergraduates (e.g., Heit \& Rotello, 2010; Rips, 2001) has found that this instructional manipulation interacts with argument type. In particular, Rips (2001) found that under deduction conditions, binary
judgments of argument strength were primarily affected by validity, regardless of causal consistency. In contrast, those given induction instructions were highly sensitive to causal consistency. In this condition, causally-consistent but logically invalid arguments (e.g., arguments like those in the top right cell of Table 1) were judged to have similar argument strength to logically valid arguments. According to Rips (2001), this pattern shows that people use qualitatively different criteria for evaluating arguments when doing induction and deduction.

If those with OCD-symptomology exhibit spared deductive reasoning but impaired inductive reasoning, then they should show a different pattern of performance on the Rips induction-deduction task. Specifically, they may show sensitivity to logical validity under deduction instructions but may not show the same sensitivity to causal consistency as controls, when given induction instructions.

Table 1: Examples of the argument types used in the Rips induction - deduction task. Participants were asked to evaluate the conclusion (below the line in italics) assuming the premises (above the line in normal font) to be true.
\begin{tabular}{|c|c|c|}
\hline \multirow[b]{2}{*}{Causal status} & \multicolumn{2}{|c|}{Validity} \\
\hline & Valid & Invalid \\
\hline \multirow[t]{2}{*}{Causally Consistent} & If Jill rolls in the mud, & Jill rolls in the mud. \\
\hline & Jill rolls in the mud. & \\
\hline \multirow{4}{*}{Causally Inconsistent} & Jill gets dirty. & Jill gets dirty. \\
\hline & If Jill rolls in the mud, & Jill rolls in the mud. \\
\hline & \begin{tabular}{l}
Jill gets clean. \\
Jill rolls in the mud.
\end{tabular} & \\
\hline & Jill gets clean. & Jill gets clean. \\
\hline
\end{tabular}

Note that the Rips task examines induction based on background causal knowledge. With the aim of providing a more comprehensive examination of possible inductive deficits in OCD, we also administered an inductive reasoning task which tested sensitivity to the diversity heuristic. In the diversity task, participants were asked to make judgments about which of two pairs of premises would provide better evidence for a more general inductive conclusion (see Table 2 for an example). One premise pair (non-diverse set) contained two very similar premises, while another (diverse set) contained premises that were less similar (but still nested within the conclusion category).

Those with no OCD-symptomology were expected to show a robust preference for the diverse set (cf. Heit et al., 2005; Osherson et al., 1990). However, if inductive reasoning is impaired in people with OCD, then we would expect to see less evidence of the diversity heuristic in those with OCD symptoms. Indirect support for this prediction comes from the finding that relative to controls, individuals with OCD often make repeated observations of the same or similar items before making a probability judgment (e.g., Fear \& Healy, 1997; Volans, 1976). As shown in Table 2, the prediction about differences between diversity-based reasoning in those low or high in OCD traits was examined
using "OCD-neutral" arguments as well as arguments with content relevant to common obsessions (OCD-relevant).

Table 2: Examples of the four argument types used in the Diversity task. The premises are given in normal font above the line and are assumed to be true. Conclusions are given in italics below the line.
\begin{tabular}{lllll}
\hline \begin{tabular}{c} 
Premise \\
Sets
\end{tabular} & OCD-Neutral & \begin{tabular}{c} 
Content \\
OCD-Relevant
\end{tabular} \\
\hline Diverse & \begin{tabular}{l} 
All cows have an \\
ileal vein
\end{tabular} & \begin{tabular}{l} 
All gold coins are \\
contaminated by the \\
bacteria hemonasella coli
\end{tabular} \\
& \begin{tabular}{l} 
All mice have an \\
ileal vein
\end{tabular} & \begin{tabular}{l} 
All dollar bills are \\
contaminated by the \\
bacteria hemonasella coli
\end{tabular} \\
& \begin{tabular}{lll} 
All mammals have \\
an ileal vein
\end{tabular} & \begin{tabular}{l} 
All forms of money are \\
contaminated by the \\
bacteria hemonasella coli
\end{tabular} \\
\hline
\end{tabular}

It is important to note that unlike many previous studies (e.g., Pélissier \& O'Connor, 2002; Simpson et al., 2007), we did not test patients who had received a formal diagnosis of OCD. Instead we employed an "analog-sample" of undergraduates who showed relatively low or high levels of OCD symptomology as measured by a widely used selfreport screening questionnaire. This approach is justifiable given that non-treatment seeking individuals who score highly on self-report measures of obsessive-compulsive symptoms often do meet diagnostic criteria for OCD (Burns, Formea, Keortge, \& Sternberger, 1995).

\section*{Method}

Participants. One-hundred undergraduate students who spoke English as their primary language participated for course credit.

These participants were all assessed using the Obsessive-Compulsive Inventory-Revised (Foa et al., 2002). This is an 18 -item self-report measure that assesses subjective experience of OCD symptoms in the past month. Item ratings are made on a 5 -point Likert scale ( \(0=\) not at all distressing, \(4=\) extremely distressing). The OCI-R has been shown to reliably distinguish between individuals with OCD and non-OCD controls, and to have high internal consistency and test-retest reliability (Foa et al. 2002). Individuals high on obsessive-compulsive symptoms were defined as those scoring equal to or greater than 21 on the

OCI-R ( \(n=44, M=29.95, S D=7.19\) ), which is consistent with the cut-off used for distinguishing between nonanxious controls and those with OCD (Foa et al., 2002). Of these participants, the majority were female \((n=30)\) and the mean age was 18.75 years ( \(S D=1.64\) ). Low-OCD individuals were defined as those scoring less than 21 on the OCI-R ( \(n=56, M=10.85, S D=5.42\) ); the majority of these participants were female ( \(n=30\) ) and the mean age was 20.46 years ( \(S D=6.67\) ).

\section*{Design and Procedure.}

All participants were tested individually in the UNSW Cognition and Reasoning lab. All were administered a Rips induction-deduction task and a premise diversity task, with order of task presentation counterbalanced across participants. After completion of the reasoning tasks, all participants also completed a test of general ability (the twosubtest short-form of the Wechsler Abbreviated Scale of Intelligence) (WASI; Wechsler, 1999), and the OCI-R. The general ability test was included so that possible group differences in reasoning performance could be differentiated from group differences in overall cognitive ability.

Rips induction-deduction task. The Rips inductiondeduction task consisted of 16 arguments that varied factorially in logical validity (valid vs. invalid) or causal consistency (consistent vs. inconsistent) (see Table 1 for examples), such that there were four argument types. The valid items were based on either the inference form modus ponens (If \(p\) then \(q, p\) therefore \(q\) ) (as in the Table 1 example) or conjunctive syllogism (not ( \(p\) and \(q\) ), p therefore not \(q\) ), such that valid items followed an acceptable logical structure but invalid items did not. Item content for 12 of the arguments was taken from Rips (2001), and the remaining four arguments were generated by the researchers.

Instructions for evaluating these arguments (deduction vs. induction) were manipulated between subjects, with approximately equal numbers allocated to each condition. Those in the induction condition were told that strong arguments are those for which "assuming the information above the line is true, this makes the sentence below the line plausible", whilst those in the deduction condition were told that valid arguments are those for which "assuming the information above the line is true, this necessarily makes the sentence below the line true". They were instructed to examine each argument and make a binary judgment about the conclusion ("strong" or "weak" in the induction condition; "valid" or "invalid" in the deduction condition). Arguments were presented one at a time on a computer screen in random order, and responses were made via onscreen buttons. There was no time limit on responding.

Premise Diversity task. This consisted of 30 items, each made up of two pairs of premises (one diverse, one nondiverse) and a general conclusion (see Table 2 for examples). Assignment of premises to the diverse or nondiverse set was based on pre-test ratings of the similarity
between the categories mentioned in the premises (e.g., cows and horses) by an independent group of participants who took no part in the main experiment. This pre-test established that diverse premise pairs ( \(M=4.52, S D=1.31\) ) were reliably perceived as less similar to one another than the non-diverse pairs ( \(M=7.94, S D=.54\) ), \(t(10)=11.75\), \(p<.01\). Premise and conclusion categories (e.g., cows, horses, mammals; gold coins, silver coins, money) were selected so that they would be familiar to participants, but the properties attached to each (e.g., "have an ileal vein", "are contaminated by hemonasella coli") were unfamiliar (cf. Osherson et al., 1990). For each item, participants had to choose which of the premise pairs provided stronger evidence for the conclusion, as illustrated in Table 2.

Half the diversity items were content-neutral (related to animals), whilst the other half were OCD-relevant (containing emotional content related to common OCDrelated concerns, such as washing and checking). The leftright positioning of diverse and non-diverse premises was randomized, as was item order. The full set of items is available from the authors.

\section*{Results}

Preliminary Analyses. A one-way ANOVA analysis of general ability scores between OCD-high and OCD-low groups indicated that the groups did not differ in estimates of general ability, \(F(1,98)=3.36, p=.07\). However, general ability estimates for those scoring in the highest quartile on the OCI-R ( \(M=105.37, S D=11.62\) ) were lower than those in the lowest quartile \((M=112.4, S D=7.55), F(1,98)=\) 298.04, \(p<.01\). As there were general ability differences in the lowest and highest quartiles of OCD symptoms, general ability was controlled in all analyses.

Rips Induction-deduction task - Proportion of positive responses. The proportion of trials in which OCD-high and OCD-low participants judged an argument as "valid" in the deduction condition or "strong" in the induction condition is given in Figure 1. As Figure 1 indicates, the proportion of positive responses across item types is clearly affected by argument consistency, \(F(1,98)=261.94, p<.01\), and validity, \(F(1,98)=223.33, p<.01\).

The key question however, is whether responses to valid-inconsistent (V-I) and invalid-consistent (Inv-C) items differ as a function of instruction and OCD-group status. Crucially, as in Rips (2001), we found a crossover interaction between the relative likelihood of making a positive response to V-I and Inv-C and the instruction manipulation, \(F(1,96)=10.82, p<.01\). Figure 1 shows that under deduction instructions, there was a higher rate of positive responding to V-I items than to Inv-C items, but that this pattern reverses under induction instructions. This suggests that people applied qualitatively different criteria to evaluating argument strength in the induction and deduction conditions. Notably, as is clear from Figure 2, this effect was found in both OCD-low and OCD-high groups (i.e. there was no significant group x item x instruction
interaction, \(p=.595\) ). All of these results remained robust when group comparisons were restricted to the highest and lowest quartile groups on the OCI-R (OCD-low, \(\mathrm{n}=24\); OCD-high, \(\mathrm{n}=27\) ). These results challenge the view that induction involving the use of background causal knowledge is selectively impaired in OCD.


Figure 1. Proportion of positive responses ('strong' or 'valid') for each item type, by instruction condition and group. V-C = Valid + causally consistent items; V-I = Valid + causally inconsistent items; Inv-C = Invalid + causally consistent items; Inv-I = Invalid + causally inconsistent items.

A further analysis of the deduction data was carried out by calculating an "interaction index", which measures the influence of causal consistency on positive responding for valid and invalid problems, whilst correcting for response bias (see Dube, Rotello, \& Heit, 2010). The interaction index was calculated using the formula;
\[
\begin{equation*}
\text { Interaction index }=\left(\mathrm{H}_{I}-\mathrm{F}_{I}\right)-\left(\mathrm{H}_{C}-\mathrm{F}_{C}\right) \tag{1}
\end{equation*}
\]

H denotes the rate of hits (responding "valid" to a logically valid item). F denotes false alarms (responding "valid" to a logically invalid item), and \(C\) and \(I\) denote causally consistent and causally inconsistent arguments respectively. The index is scaled such that a positive index suggests that people find it easier to discriminate between valid and invalid items with unbelievable conclusions. An interaction index score was calculated for each participant and mean scores were compared between the OCD-low ( \(M=0.39, S D\) \(=.05)\) and OCD-high ( \(M=0.34, S D=.07\) ) groups. Consistent with previous work, the interaction index calculated for the OCD-low group was positive (Dube et al., 2010), as was the index calculated for the OCD-high group. The interaction index scores did not differ between these groups, \(F(1,95)=.38, p=.54\). In other words, there were no OCD-group differences in the impact of causal consistency on judgments of logical validity. Again it appears that OCD-low and OCD-high show similar patterns of reasoning based on background causal knowledge.

Proportion of Diverse Pairs Chosen. Overall, both OCDhigh and OCD-low groups were more likely to choose the diverse premise pairs as providing stronger evidence for inductive generalization than would be expected by chance, (OCD-low, \(t(55)=7.98, p<.01\); OCD-high, \(t(42)=4.05\), \(p<.01)\). The relative preference for diverse pairs in OCDhigh and OCD-low groups was compared. There was no effect of OCD status for overall proportion of diverse pairs chosen, \(F(1,106)=1.35, p=.25\). Participants showed a reliable diversity effect (i.e. selection of diverse pairs above chance) for both OCD-neutral, \(t(99)=7.27, p<.01\), and OCD-relevant items, \(t(99)=9.24, p<.01\).

We again reanalyzed these data restricting group comparisons to those individuals showing the most extreme scores on the OCI-R (i.e. the lowest and highest quartiles). As can be seen in Figure 2, individuals exhibiting the highest OCD symptoms shows reduced preference for diverse evidence in induction than those with low levels of OCD symptoms, \(F(1,51)=7.41, p<.01, d=1.09\). Individuals with high scores on the OCI-R were less likely to show a preference for diverse evidence, regardless of whether item content was neutral, \(F(1,51)=5.95, p<.05\), \(d=1.06\), or emotionally relevant, \(F(1,51)=7.89, p<.01\), \(d=1.01\). Moreover, this difference persisted when group differences in general ability were controlled by using individual scores on the ability test as a covariate. \({ }^{1}\)


Figure 2. Proportion of Diverse Pairs Chosen by Lowest and Highest Quartiles on the OCI-R.

Overall, these data suggest that non-clinical adults with the highest levels of OCD symptoms were less likely to make use of the diversity heuristic in inductive reasoning than those who show low levels of symptomatology.

\footnotetext{
\({ }^{1}\) Moreover, linear regression analyses showed that scores on the OCI-R explained a significant amount of variance in the proportion of diverse pairs chosen overall after the common variance explained by general ability had been controlled (i.e. when OCI-R scores were entered into the equation after general ability), \(\mathrm{R}^{2}=\) \(.33, F(1,97)=5.95, p<.01\), and for both neutral items, \(\mathrm{R}^{2}=.35\), \(F(1,97)=6.77, p<.01\), and OCD relevant items, \(\mathrm{R}^{2}=.29, F(1,97)\) \(=4.42, p<.05\).
}

\section*{General Discussion}

Previous work (e.g. Pélissier \& O’Connor, 2002) has suggested that people with OCD show a selective deficit in inductive reasoning but unimpaired ability to reason deductively. This study tested this hypothesis in two ways. First, we compared the inductive and deductive performance of those with low or high levels of OCD-related traits using a common stimulus set for both tasks. Second, we examined the performance of these two groups on two types of inductive problems; one based on the use of background knowledge to determine inductive validity and another examining the inductive heuristic of evidence diversity.

Overall there was mixed support for the hypothesis of a selective inductive deficit in people with OCD-related traits. Results from the Rips induction-deduction task replicated the main findings of other comparisons of inductive and deductive reasoning in nonclinical populations (e.g, Heit \& Rotello, 2010; Rips, 2001). Induction and deduction instructions led participants to evaluate arguments in qualitatively different ways. Evaluations of inductive strength were based on consistency with prior knowledge. Evaluations of deductive validity were evaluated according to logical necessity. Crucially, there were no differences between OCD-low and OCD-high groups in patterns of inductive and deductive reasoning. These data provide little support for a selective deficit in inductive reasoning based on background knowledge in people with OCD.

An important finding however is that those who showed the highest level of OCD symptomatology exhibited an atypical pattern of induction based on the diversity heuristic. Those in the highest OCD symptom quartile were less likely to see diverse premise pairs as a stronger basis for inductive generalization than those in the lowest quartile. This difference persisted when the effects of general ability were factored out. This suggests that although inductive reasoning based on consistency with background knowledge may be intact in people with high-OCD symptoms, this group does show an impaired understanding of the implications of evidence diversity. Moreover, this appears to be a global impairment, affecting inductive reasoning about both OCD-related and OCD-neutral items.

Further work is needed to identify the specific source of this inductive impairment. It is notable that although sensitivity to evidence diversity is robust in nonclinical groups (Heit et al., 2005), there are some cases where this heuristic interacts with other factors, such as property knowledge. When diverse premises share a highly specific or idiosyncratic property, inductive generalizations based on diverse premises may actually be weaker than those based on non-diverse premises (Feeney \& Heit, 2011). For example, Medin et al., (2003) found that people were less likely to generalize a property shared by camels and desert rats to other mammals, than a property shared by camels and rhinos, even though the first set of premises was rated as more diverse. It seems unlikely however, that this type of mechanism could explain the weakening of the diversity effect in people with OCD-symptoms. If this effect was
driven by the high OCD group inferring more specific or idiosyncratic relations between diverse premise pairs, one could reasonably expect that this effect would be stronger in items with OCD-relevant content. However, the weakening of the diversity effect in people with OCD symptoms was found across both OCD-relevant and neutral items.

A more likely explanation of reduced sensitivity to diversity in the OCD-high group relates to preservative tendencies observed in other studies of probabilistic and inductive reasoning in OCD patients (e.g. Fear \& Healy, 1997; Volans, 1976). Such studies have found that when asked to evaluate evidence for an uncertain conclusion, people with OCD-related traits often repeatedly choose to examine similar or redundant types of evidence.

Overall, we found some evidence for impaired inductive reasoning in people with OCD-related traits, but only when a general inductive heuristic was involved. By contrast, high-OCD individuals did not differ from controls in induction based on background knowledge or in deductive reasoning. Clearly, given the analog nature of our samples, we must be cautious in generalizing the deficit in the use of the diversity heuristic to clinical populations. However, given that the level of OCD symptomology is likely to be more severe in those seeking or undergoing treatment, it seems reasonable to speculate that such individuals will also show impairment in diversity-based inductive reasoning.

This study is one of the first to apply contemporary methods and theories of induction and deduction to examine reasoning deficits related to a specific clinical condition. Our view is that a careful examination of patterns of spared and impaired reasoning in such groups can contribute to the understanding of reasoning in both clinical and non-clinical populations (cf. Caramazza \& Coltheart, 2006). For example, the finding that OCD-related symptoms are associated with impairments in some forms of induction but not others suggests that more than one underlying cognitive process drives inductive reasoning.

\section*{References}

Burns, G. L., Formea, G. M., Keortge, S., \& Sternberger, L. G. (1995). The utilization of nonpatient samples in the study of obsessive compulsive disorder. Behaviour Research and Therapy, 33, 133-144.
Caramazza, A., \& Coltheart, M. (2006). Cognitive Neuropsychology twenty years on. Cognitive Neuropsychology, 23, 3-12.
Dube, C., Rotello, C. M., \& Heit, E. (2010). Assessing belief bias with ROCs: It's a response bias effect. Psychological Review, 117, 831-863.
Fear, C. F., \& Healy, D. (1997). Probabilistic reasoning in obsessive-compulsive and delusional disorders. Psychological Medicine: A Journal of Research in Psychiatry and the Allied Sciences, 27, 199-208.
Feeney, A., \& Heit, E. (2011). Properties of the diversity effect in category-based inductive reasoning. Thinking \& Reasoning, 17, 156-181.

Foa, E. B., Huppert, J. D., Leiberg, S., Langner, R., Kichic, R., Hajcak, G., \& Salkovskis, P. M. (2002). The Obsessive-Complusive Inventory: Development and validation of a short version. Psychological Assessment, 14, 485-495.
Hayes, B., Heit, E., \& Swendsen, H. (2010). Inductive reasoning. Wiley Interdisciplinary Reviews: Cognitive Science, 1, 278-292.
Heit, E., Hahn, U., \& Feeney, A. (2005). Defending diversity. In W. Ahn, B. C. Goldstone, A. B. Love \& P. Wolff (Eds.), Categorization inside and outside the lab: Festscrift in honour of Douglas L. Medin (pp. 87-100). Washington, DC: American Psychological Association.
Heit, E., \& Rotello, C. M. (2010). Relations between inductive reasoning and deductive reasoning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 36, 805-812.
Heit, E., Rotello, C. M., \& Hayes, B. K. (2012). Relations between memory and reasoning. Psychology of learning and motivation, 57, 57-101.
Lo, Y., Sides, A., Rozelle, J., \& Osherson, D. (2002). Evidential diversity and premise probability in young children's inductive judgement. Cognitive Science, 26, 181-206.
Medin, D. L., Coley, J. D., Storms, G., \& Hayes, B. K. (2003). A relevance theory of induction. Psychonomic Bulletin \& Review, 10, 517-532.
O'Connor, K. (2002). Intrusions and inferences in obsessivecompulsive disorders. Clinical Psychology and Psychotherapy, 9, 38-46.
Osherson, D. N., Smith, E. E., Wilkie, O., López, A., \& Shafir, E. (1990). Category-based induction. Psychological Review, 97, 185-200.
Pélissier, M.-C., \& O'Connor, K. P. (2002). Deductive and inductive reasoning in obsessive-compulsive disorder. British Journal of Clinical Psychology, 41, 15-27.
Riggs, D. S., \& Foa, E. B. (1993). Obsessive compulsive disorder. In D. H. Barlow (Ed.), Clinical handbook of psychological disorders (2 \({ }^{\text {nd }}\) ed., pp. 189-239). New York: Guilford Press.
Rips, L. J. (2001). Two kinds of reasoning. Psychological Science, 12, 129-134.
Simpson, J., Cove, J., Fineberg, N., Msetfi, R. M., \& Ball, L. J. (2007). Reasoning in people with obsessivecompulsive disorder. British Journal of Clinical Psychology, 46, 397-411.
Volans, P. J. (1976). Styles of decision-making and probability appraisal in selected obsessional and phobic patients. British Journal of Social \& Clinical Psychology, 15, 305-317.
Wechsler, D. (1999). Wechsler Abbreviated Scale of Intelligence (WASI). San Antonio, TX: Harcourt Assessment.

\title{
The Influence of the Attention Set on Exogenous Orienting
}

\author{
Ahnate Lim (ahnate@hawaii.edu) \\ Department of Psychology, University of Hawaii at Manoa \\ 2530 Dole Street, Honolulu, HI 96822 USA \\ Scott Sinnett (ssinnett@hawaii.edu) \\ Department of Psychology, University of Hawaii at Manoa \\ 2530 Dole Street, Honolulu, HI 96822 USA
}

\begin{abstract}
The processing of incoming sensory information relies on interacting mechanisms of sustained attention (the ability to focus attention and ignore irrelevant stimuli) and attentional capture (the ability of certain stimuli to reflexively attract one's attention). Being able to precisely predict what can capture attention when it is engaged in a demanding task is important both for understanding the nature of attention as a cognitive system and also for practical applications. While evidence indicates that exogenous capture, a mechanism previously understood to be automatic, can be eliminated while concurrently performing a demanding task, we reframe this phenomena within the theoretical framework of the 'attention set' (Most et al., 2005). Consequently, the specific prediction that cuing effects should reappear when dimensions of the cue overlap with those in the attention set (i.e., elements of the demanding task) was empirically tested and confirmed. Suggestions for further theoretical refinement and empirical testing are discussed.
\end{abstract}

Keywords: Theories of attention; attention set; exogenous cuing; orienting; attentional capture; perceptual load

\section*{Introduction}

As an information processing mechanism, one of the characteristic dichotomies of attention is how it must have the capacity to be both focused and distractible at the same time. The ability to ignore irrelevant stimuli and closely attend to a specific task at hand is fundamental to goal directed behavior. Conversely, the ability to be distracted by potentially dangerous events or to be drawn towards relevant information outside the current task or area of focus can be crucial for avoiding harm and responding effectively to the environment. In fact, neurological evidence has demonstrated that these dissociable mechanisms are underpinned by distinct and interactive neural networks (Corbetta, Patel, \& Shulman, 2008; Corbetta \& Shulman, 2002).

Many of the attentional mechanisms that we use on a daily basis can be characterized by the way in which they enable goal-directed and top-down control of behavior. Indeed, top-down attentional control has been observed to play a role in visual search (Wolfe, 2007), endogenous (participant directed) orienting of attention to spatial locations (Posner, 1980), and even feature integration (Treisman \& Gelade, 1980), to name but a few. Although
the environment may contain stimuli that actively compete for and capture attention, what ultimately becomes selected for subsequent processing can be influenced by "top-down signals" that filter for behaviorally relevant objects (Desimone \& Duncan, 1995).

Considering the importance of responding effectively to changes in the environment, the attentional system allows for stimuli to 'reflexively' capture attention, in a 'bottomup' environmentally triggered fashion. It has been shown that this aspect of attention can be dependent on the particular nature of the stimuli and environmental circumstances at hand. Such factors may include the role of stimulus saliency (Jonides \& Yantis, 1988) and relevancy to behavioral goals (Yantis \& Egeth, 1999), for example.

Of direct interest to the current study, recent research has suggested that the reflexive orienting of attention can, at times, be interrupted when an observer is undergoing a difficult and demanding task (Santangelo, Belardinelli, \& Spence, 2007). In other words, where such exogenous, or stimulus-driven, mechanisms were previously thought to be automatic (Müller \& Rabbitt, 1989), more recent evidence has suggested that these effects may be eliminated in a state of focused attention. For example, several recent studies have demonstrated that requiring participants to perform a concurrent demanding task can effectively eliminate the ability of exogenous cues to capture attention (Santangelo et al., 2007; Santangelo, Finoia, Raffone, Olivetti Belardinelli, \& Spence, 2008; Santangelo \& Spence, 2008; Theeuwes, 1991; Yantis \& Jonides, 1990). Two of these studies, for instance, employed central-arrows as \(100 \%\) predictive cues in a target detection task, while also deploying abrupt visual onsets as exogenous cues (Theeuwes, 1991; Yantis \& Jonides, 1990), and found that the abrupt visual onsets had no effect on performance. Yet a different study by Van der Lubbe and Postma (2005) used more eccentric (peripheral) exogenous cues and obtained evidence to the contrary, where effects of the abrupt visual onsets were observed even when attention was engaged.

Thus, there appears to be evidence indicating that under some circumstances exogenous cuing effects remain, while under others these effects are eliminated. While these experiments often employ a demanding central task, the question remains as to why opposing findings have been observed. Importantly, the answer to this question should
provide insight as to how the sustained attentional system interacts with the attentional capture system.

Although it is inherently difficult to formulate theories of attention that are both broad in scope (encompassing several classes of phenomena) while concurrently possessing predictive power for detailed behavioral outcomes, there are frameworks that could provide initial scaffolding towards such comprehensive theories. One such general framework for combining aspects of both inattentional blindness (i.e., an indirect measure of sustained focus) and attentional capture has been proposed by Most, Scholl, Clifford, and Simons (2005). Central to their theoretical framework, is the idea of an 'attention set' that is synonymous with the current task at hand or state of mind. The authors postulate that this 'attention set' should be the most influential factor in determining what captures attention. Incidentally, the idea that the current frame of mind determines how attention is allocated has also been proposed by Neisser's construct of the perceptual cycle (1976). While Most and colleagues' formulation provides an explanatory construct for both sustained attention and attentional capture, their emphasis on the attention set can be used to infer precise predictions. Specifically, Most et al. (2005, p. 218) proposed that:
> "Although some stimulus properties (e.g., uniqueness) can affect noticing, to a larger extent the unexpected objects that people consciously see depend on the ways in which they 'tune' their attention for processing of specific types of stimuli-that is, on the attentional set that they adopt."

Consequently, this leads to the prediction that irrelevant events that are within the same attentional set should be capable of capturing attention (i.e., irrelevant events that are similar to the targets used in a separate and attended to task), whereas events that fall outside of the attention set should go unnoticed (e.g., a gorilla walking amidst a group of people passing a basketball while counting passes, see for example Simons \& Chabris, 1999).

Most and colleagues’ (2005; 2001) predictions regarding the influence of the attention set were supported by a series of empirical studies centered around a paradigm in which participants counted the number of bounces of a subset of items moving within a display. Crucially, an unexpected object entered the display after several trials and detection rates for these objects were used as a measure of attentional capture. In this way, Most et al. were able to manipulate the composition of the attention set (the items moving and bouncing within the display), and observe the subsequent effects on attentional capture. Of critical importance to their theory, the findings suggest that the capture of awareness is influenced both by top-down and bottom-up interactions, where the most influential factor is ultimately the attention set adopted (although certain bottom-up factors such as stimulus salience can increase the chance that objects will be noticed). In general, when unexpected items possessed features that overlapped with those in the attentional set, participants consistently noticed them, whereas when the items were outside the attention set, participants rarely
noticed them. Bearing in mind that Most and colleagues' (2005; 2001) experiments were adaptations of an inattentional blindness paradigm where participants were tested on their awareness and processing of an unexpected event, the question remains as to whether the same predictions would generalize to a different task setting where attention is focused on a central area (rather than across the experimental display), and attentional capture is measured through exogenous cuing rather than conscious detection.

To recall a related example that was mentioned earlier in more detail, Santangelo and colleagues devised a paradigm involving both a demanding central task and an exogenously cued target detection task, and found that exogenous orienting does not capture attention in a mandatory fashion when undergoing a demanding central task (see Santangelo et al., 2007; Santangelo \& Spence, 2007, 2008). That is, when one's attention is engaged in performing a perceptually or attentionally demanding task, the automatic effects of exogenous cues seemingly disappear. This finding is especially important considering that previous accounts of exogenous cuing suggest that peripheral cues automatically capture attention (Jonides, 1981; Müller \& Rabbitt, 1989; Van der Lubbe \& Postma, 2005).

While it is possible that the elimination of the cuing effect could be related to an increase in perceptual load and a concomitant reduction in available attentional resources, as suggested by Santangelo et al. (2007), Most et al.'s (2005) theoretical framework could equally predict the same result. That is, Most et al. would predict that the elimination of the cuing effect would be related to the fact that the peripheral cues were not contained in the 'attention set' (i.e., the cue was not a part of, nor was it related to, anything in the central task). This was precisely the case in the paradigm used by Santagelo and colleagues (2007; 2008). Specifically, participants were required to detect a number amongst a rapid serial presentation of letters and numbers, while the peripheral cue was a geometric shape (i.e., not a letter or number). Adopting Most et al.'s logic, the peripheral cue was task irrelevant and not related to anything in the attention set (letters or numbers), therefore it is not surprising that it failed to capture attention. Accordingly, one can predict that if the irrelevant peripheral cues were to be manipulated such that they overlapped with the current attention set (i.e., the peripheral cues and central targets come from the same category or share the same features), they should successfully capture attention despite being completely irrelevant to the task at hand. In the present study, our goal was to investigate the relationship between central processing and the peripheral capture of attention (exogenous orienting), and how this relationship is mediated by the attention set.

Using a within subjects design, participants performed a difficult central task requiring them to detect numbers that were presented within a stream of rapidly presented letters. On a subset of trials participants responded to the location of a peripherally presented target (above or below) that was
orthogonally cued. Critically, we presented different types of peripheral cues to each participant, such that the cue was either of the same content as the central task or different. Note that the cue itself was completely irrelevant to the task and in theory would be outside of the attention set if it did not share any stimulus characteristics with items in the attention set (i.e., the central task in this case). If Most et al.'s (2005) prediction holds, exogenous cuing effects should be eliminated when peripheral cues are different from stimuli in the central stream. However, if peripheral cues are related to (or were even subsets of) the central task, then an exogenous cuing effect (i.e., attentional capture) should emerge.

\section*{Methods}

\section*{Participants}

Twenty-three participants (mean age \(=22 \pm 4 ; 13\) females) were recruited from undergraduate courses at the University of Hawaii at Manoa, and offered course credit for their participation. All participants were naïve as to the purpose of the experiment and had normal or corrected to normal vision. Ethical approval was obtained from the University's Committee on Human Subjects.

\section*{Stimuli}

All stimuli were presented on a 20 ", iMac using Bootcamp and DMDX software (Forster \& Forster, 2003). Observers sat approximately 60 cm from the display. Stimuli in the central rapid serial visual presentation (RSVP) stream was constructed from randomly chosen non-repeated letters (11 selected from set of 17: B, C, D, E, F, J, K, L, M, N, P, R, S, \(\mathrm{T}, \mathrm{Y}, \mathrm{X}, \mathrm{Z}\) ), each presented for 100 ms with a 16.7 ms interstimulus interval (ISI). For digit detection trials, numbers were selected from a set of six: \(2,3,4,5,6,9\). Visual targets were black circles (subtending \(2^{\circ}\) ) and cues were either black rectangles \(\left(2.5^{\circ} \times 1.7^{\circ}\right)\) or numbers of comparable size (i.e., outside or in the attention set of the primary task, respectively; see Figures 1 and 2). Aside from the use of number cues on half of the trials, all stimuli, presentation times, and counterbalancing were constructed to be similar to the unimodal visual condition used in Santangelo et al.'s experiment (2007).

\section*{Procedure}

All participants were presented with written instructions for the task on the computer screen. Next they were presented with practice trials and given accuracy and reaction time feedback after the end of each trial. The participants had the option of repeating the instructions, repeating the practice trials, or continuing with the experiment. The experimenter also monitored participants during the practice trials to ensure their understanding of the task.

For the actual task, participants were required to monitor the RSVP stream presented in the center of the display, and to respond to the occurrence of a numerical digit. A digit occurred on the majority of trials ( \(67 \%\) ). On the remaining
trials (33\%) the digit was not presented and instead, participants responded to the location of a spatial target that could have occurred in one of the four corners of the screen. A peripheral cue was presented on all trials, but was irrelevant to either task. The cue could have validly predicted the side of the spatial target or not (note, a spatial target was not present on digit trials). Responses were made using one of three keys following detection of either 1) a number, 2) an upward spatial target, or 3) a downward spatial target.


Figure 1: Schematic representation of the task. See text for details.


Figure 2: The two different cue types used in the task.
Each trial began with a fixation cross ( 1000 ms ) followed by the RSVP stream of 11 items. On digit detection trials, the numbers randomly occurred in either the third, sixth, or ninth position in the stream (see also, Santangelo et al., 2007). A spatial cue was also presented on each trial (for 100 ms , identical to item duration), occurring in the third or sixth position on either the right or left side of the display equiprobably. When spatial targets occurred a number was not presented in the stream, and the spatial target appeared two positions after the cue ( \(5^{\text {th }}\) or \(8^{\text {th }}\) position). The two
types of cues, rectangles or numbers, also occurred equiprobably (see Figure 2). Each experimental session consisted of 196 randomized trials, 132 of which were the digit detection task, and 64 of which were target detection (Santangelo et al., 2007). Cue combinations and trial repetitions were counterbalanced. Participants were instructed to respond as soon as targets were detected.

\section*{Results}

Mean reaction times (RTs) and error rates were analyzed using three repeated measures ANOVAs (analysis of variance): one for the overall experiment and two separate ANOVAs for the digit and spatial target detection conditions. Assumptions of sphericity were tested on all analyses, with Huyn-Feldt corrections being applied to \(p\) values where appropriate.


Figure 3: Mean error rates across tasks and cue types. Error bars indicate standard error values.

The first ANOVA was performed on the RT data with factors of task type (digit or target) and cue type (rectangle or number). There was no main effect of task type, \(F(1,22)\) \(=1.2, p=.3\), indicating that there were no overall differences in RTs across digit ( \(M=594 \mathrm{~ms}\) ) and target ( \(M\) \(=556 \mathrm{~ms}\) ) detection tasks. There was, however, a main effect of cue type, \(F(1,22)=8.4, p=.008\), indicating that, overall, RTs were slower when number cues ( \(M=589 \mathrm{~ms}\) ) occurred compared to rectangle cues \((M=563 \mathrm{~ms})\). There was no interaction between task and cue types, \(F(1,22)<1\), \(n s\), indicating no differences in RT patterns across the two tasks. In examining the error data, there was a main effect of task type, \(F(1,22)=36.2, p<.001\), with lower error rates for the digit task ( \(2 \%\) ) compared to the target task ( \(15 \%\) ). Error rates were also higher on trials with number cues ( \(10 \%\) ) than on those with rectangle cues ( \(7 \%\) ), \(F(1,22\) ) = \(4.5, p=.045\), indicating that on average the task was more difficult when number cues were present. Notably, the analysis revealed a marginally significant interaction between task and cue types, \(F(1,22)=3.5, p=.07\), indicating that number cues tended to be more distracting
than rectangle cues ( \(18 \%\) vs \(13 \%\), respectively) during spatial target detection, but not during digit detection ( \(3 \%\) vs \(2 \%\), see Figure 3).

A second three way ANOVA performed on the digit detection condition with factors of digit position (3), cue position (2), and cue type (2) revealed that participants detected the digits significantly faster when they were presented in the ninth ( \(M=455 \mathrm{~ms}\) ) position than when presented in the sixth ( \(M=584 \mathrm{~ms}\) ) or third ( \(M=736 \mathrm{~ms}\) ) positions respectively, \(F(2,44)=25.6, p<.001\). Reaction times were also faster when cues were presented in the third position \((M=574 \mathrm{~ms})\) than in the sixth position \((M=613\) \(\mathrm{ms}), F(1,22)=13.0, p<.01\). There was also a significant interaction between digit position and cue position, \(F(2,44)\) \(=5.4, p=.008\), suggesting that performance was worse when the cue occurred at the same time as the digit. Although there was no main effect in mean RTs between trials with number cues compared to trials with rectangle cues \(F(1,22)<1, n s\), there was a significant three-way interaction, \(F(2,44)=4.1, p=.04\), indicating a different pattern of RTs between rectangle and number cues. Specifically, when the target and cue both occur in the third position the number cue adversely affected performance whereas the rectangle cue did not (see Figure 4). No significant differences were found in the error rate data.


Figure 4: Interaction between target position, cue position in temporal stream, and cue type. Graph A shows trials with rectangle cues, whereas Graph \(B\) shows those with number cues.

The third, and most important, two way ANOVA was performed on the spatial target detection condition with factors of cue validity (2) and cue type (2). While there were no main effects of cue validity on reaction times, \(F(1,22)=\) \(3.0, p=.1\), there was a main effect of cue type, \(F(1,22)=\) \(11.0, p=.003\), where reaction times were slower when the target preceding cues were numbers \((M=570 \mathrm{~ms})\) compared to when they were rectangles \((M=543 \mathrm{~ms})\). Paramount to this study, the interaction between cue validity and cue type was also significant, \(F(1,22)=4.7, p=.04\), indicating the presence of cuing effects for number cues on the one hand ( 554 ms for valid cues, and 587 ms for invalid cues), and the lack of cuing effects for rectangle cues on the other hand ( 539 ms for valid cues, and 547 ms for invalid cues, see Figure 5). No significant differences were found in the error rates across cue type or validity.


Figure 5: Interaction of cuing effects within the spatial target detection condition. Cue validity: \(0=\) invalid cue, and \(1=\) valid cue.

\section*{Discussion}

The main purpose of this experiment was to test Most et al.'s (2005) theoretical framework on attention. To this end, our findings unequivocally support the prediction that irrelevant items that fall within the attentional set are capable of capturing attention while irrelevant items that fall outside of the attention set do not. That is, as was predicted, peripheral cues that were in the same category, or had overlapping features with the central task (numbers) had a cuing effect ( 33 ms ) on spatial target detection, while items outside the category (rectangles) did not (i.e., the cuing effect was eliminated). Not only was a cueing effect observed for peripheral number distractors, but this type of cue also led to a general increase in RTs for spatial trials. This indicates that despite being irrelevant to the task, the number cues were nevertheless processed, and served as more effective distractors when compared with the rectangles.

Although valid number cues effectively captured attention, they did not facilitate overall faster reaction times. That is, the mean reaction time for trials with rectangle cues was in fact faster than for number cues, despite the lack of a
cuing effect within this condition. This may possibly be due to the interaction of the number cues with the requirements of the central task. That is, any potential facilitating effects on performance of the valid number cues were probably offset by the overlap and interference with the digit detection task. The effects of this interference were also observed in the higher error rates for trials with number cues when compared to rectangle cues for spatial target detection (Figure 3).

Reaction times on the digit detection trials also point towards greater interference from the number cues. The interaction indicates that performance was worse on trials when the cue occurred at the same time as the digit, with more interference occurring from number cues when presented in the third frame. It is worth noting here that the lack of clearly distinguished differences between the effects of the rectangle and number cues on digit detection may be due to a more general distracting effect of the number cues. That is, the number cue may induce a distracting effect that generalizes beyond those particular trials to even cause the rectangle cues to become more distracting than they naturally would be. An effective way to test this theory would be to have participants also perform a task that consisted only of rectangle cues, and to then compare the pattern of results.

Aside from providing support for the theoretical position that the most influential factor for attentional capture is the 'attention set'-or the current items in focus-our findings also lend support to the notion that attention focuses on objects and features in addition to spatial location (e.g., Duncan, 1984; Egly, Driver, \& Rafal, 1994; for a review, see Scholl, 2001). In refining our understanding of the attention set, it becomes imperative to more precisely define the attention set itself, for the reason that when one is engaged in a task, there are usually multiple objects or different classes of events to attend to. For example, in this experiment, we defined the attention set as being the digit detection task, due to the fact this occurred the majority ( \(67 \%\) ) of the time. The most important object in the central stream was the number, and accordingly the identity of peripheral cues was manipulated to be numbers on half of the trials. \({ }^{1}\) The question remains however, as to what role the letters in the letter stream do play in the attention set.

Despite the fact that the letters within the RSVP stream are of a different category than the number targets, they are nevertheless processed by virtue of proximity to the number targets (both temporally and spatially) and the fact that the participants must monitor the stream in order to accurately detect the number amongst letters. We speculate that the letters should also be in the attention set, but is this assumption warranted? It is possible that the letters themselves may undergo some form on inhibition, and therefore a cue that is a letter might not capture attention. Even if this assumption were to be warranted, what would

\footnotetext{
\({ }^{1}\) Note that on each trial the number cues were different than the numbers presented in the central RSVP task, thereby avoiding any potential confounds.
}
be the precise role of letters in the attention set? Would the letters be afforded equal roles to the numbers, or would their roles be lesser, perhaps even of an inhibitory nature?

Thus it is clear that although Most and colleagues' (2005) framework provides a constructive foundation to build upon, further theoretical refinement and specification through experimentation is needed. Given that many aspects of attention appear to operate in context dependent manners, exploring these contexts within the unifying framework of the attention set may prove to be an informative approach for understanding the mechanisms of attention.

\section*{References}

Corbetta, M., Patel, G., \& Shulman, G. L. (2008). The reorienting system of the human brain: from environment to theory of mind. Neuron, 58(3), 306-324.
Corbetta, M., \& Shulman, G. L. (2002). Control of goaldirected and stimulus-driven attention in the brain. Nature Reviews Neuroscience, 3(3), 201-215.
Desimone, R., \& Duncan, J. (1995). Neural mechanisms of selective visual attention. Annual review of neuroscience, 18(1), 193-222.
Duncan, J. (1984). Selective attention and the organization of visual information. Journal of Experimental Psychology: General, 113(4), 501-517.
Egly, R., Driver, J., \& Rafal, R. D. (1994). Shifting visual attention between objects and locations: evidence from normal and parietal lesion subjects. Journal of Experimental Psychology: General, 123(2), 161.
Forster, K., \& Forster, J. (2003). DMDX: A Windows display program with millisecond accuracy. Behavior Research Methods, Instruments, \& Computers, 35(1), 116.

Jonides, J. (1981). Voluntary versus automatic control over the mind's eye's movement. In J. Long \& A. Baddeley (Eds.), Attention and Performance \(I X\) (pp. 187-203). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.
Jonides, J., \& Yantis, S. (1988). Uniqueness of abrupt visual onset in capturing attention. Attention, Perception, \& Psychophysics, 43(4), 346-354.
Most, S. B., Scholl, B. J., Clifford, E. R., \& Simons, D. J. (2005). What you see is what you set: sustained inattentional blindness and the capture of awareness. Psychological Review, 112(1), 217.
Most, S. B., Simons, D. J., Scholl, B. J., Jimenez, R., Clifford, E., \& Chabris, C. F. (2001). How not to be seen: The contribution of similarity and selective ignoring to sustained inattentional blindness. Psychological Science, 12(1), 9-17.
Müller, H. J., \& Rabbitt, P. M. (1989). Reflexive and voluntary orienting of visual attention: time course of activation and resistance to interruption. Journal of Experimental Psychology: Human Perception and Performance, 15(2), 315.
Neisser, U. (1976). Cognition and reality: Principles and implications of cognitive psychology. San Francisco: W. H. Freeman.

Posner, M. I. (1980). Orienting of attention. Q J Exp Psychol, 32(1), 3-25.
Santangelo, V., Belardinelli, M., \& Spence, C. (2007). The suppression of reflexive visual and auditory orienting when attention is otherwise engaged. Journal of Experimental Psychology: Human Perception and Performance, 33(1), 137-148.
Santangelo, V., Finoia, P., Raffone, A., Olivetti Belardinelli, M., \& Spence, C. (2008). Perceptual load affects exogenous spatial orienting while working memory load does not. Experimental brain research, 184(3), 371-382.
Santangelo, V., \& Spence, C. (2007). Multisensory cues capture spatial attention regardless of perceptual load. Journal of Experimental Psychology: Human Perception and Performance, 33(6), 1311-1321.
Santangelo, V., \& Spence, C. (2008). Is the exogenous orienting of spatial attention truly automatic? Evidence from unimodal and multisensory studies. Consciousness and cognition, 17(3), 989-1015.
Scholl, B. J. (2001). Objects and attention: The state of the art. Cognition, 80(1-2), 1-46.
Simons, D. J., \& Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. Perception-London, 28(9), 1059-1074.
Theeuwes, J. (1991). Exogenous and endogenous control of attention: The effect of visual onsets and offsets. Attention, Perception, \& Psychophysics, 49(1), 83-90.
Treisman, A. M., \& Gelade, G. (1980). A feature-integration theory of attention. Cognitive psychology, 12(1), 97-136.
Van der Lubbe, R. H. J., \& Postma, A. (2005). Interruption from irrelevant auditory and visual onsets even when attention is in a focused state. Experimental brain research, 164(4), 464-471.
Wolfe, J. M. (2007). Guided Search 4.0: Current progress with a model of visual search. In W. Gray (Ed.), Integrated Models of Cognitive Systems. NY: Oxford.
Yantis, S., \& Egeth, H. E. (1999). On the distinction between visual salience and stimulus-driven attentional capture. Journal of Experimental Psychology: Human Perception and Performance, 25(3), 661.
Yantis, S., \& Jonides, J. (1990). Abrupt visual onsets and selective attention: voluntary versus automatic allocation. Journal of Experimental Psychology: Human Perception and Performance, 16(1), 121.

\title{
Modeling the Relational Shift in Melodic Processing of Young Children
}

\author{
Ahnate Lim (ahnate@hawaii.edu) \\ Department of Psychology, University of Hawaii at Manoa \\ 2530 Dole Street, Honolulu, HI 96822 USA
}

\author{
Leonidas A. A. Doumas (leonidas@hawaii.edu) \\ Department of Psychology, University of Hawaii at Manoa 2530 Dole Street, Honolulu, HI 96822 USA
}

\author{
Scott Sinnett (ssinnett@hawaii.edu) \\ Department of Psychology, University of Hawaii at Manoa \\ 2530 Dole Street, Honolulu, HI 96822 USA
}

\begin{abstract}
As a ubiquitous trend in the cognitive development of children, the 'relational shift' accounts for a change in preference for absolute percepts towards a preference for relational percepts, and is observed across a wide variety of domains. Extensive evidence indicates that this prepotency for relational processing is also observed in how children process melodies. When recalling melodies, younger children typically recall more absolute pitch properties than older children, while the exact opposite occurs in older children. Using DORA (Discovery Of Relations by Analogy; Doumas et al., 2008), a domain-general symbolic-connectionist model of relation learning, we simulated the relational shift in melodic perception of children age 3-6 years based on an experiment by Sergeant and Roche (1973). DORA's performance matched the children's well, suggesting common developmental and perceptual mechanisms between the relational shift in melodic processing and the shift seen across other domains.
\end{abstract}

Keywords: Melodic perception; relation learning; development; relational shift; absolute pitch; computational modeling; DORA.

\section*{Introduction}

One of the fundamental cross-domain trends in human development is characterized by a qualitative transformation, or shift, in how children process information. Evidence from developmental psychology overwhelmingly indicates that while children initially attend to, recall, and reason about absolute perceptual properties, around the age of 4-6 they begin to rely on structured relational properties (Allport, 1924; Gentner \& Rattermann, 1991; Halford, 2005; Pollack, 1969; Vernon, 1940). This shift has been observed in areas such as language (Gentner, 1988), spatial tasks (Case \& Khanna, 1981; DeLoache, Sugarman, \& Brown, 1985), number comprehension (Gelman \& Gallistel, 1978; Michie, 1985), and visual shape perception (Abecassis, Sera, Yonas, \& Schwade, 2001), to name but a few. This phenomenon has been termed the 'relational shift,' as the characteristic trend is towards greater reliance on relational attributes as children mature.

Consistent with the developmental trajectory for the relational shift in other domains, in the domain of music
children also develop from initially processing more absolute aspects of melodies to processing more relational aspects as they grow older. In an especially revealing study, Sergeant and Roche (1973) trained three groups of children from the age of three to six to reproduce melodies from invariant recordings. When the children were required to recall the melodies one week later, the younger group reproduced the absolute pitches more accurately than the older group, while the older group reproduced the relational aspects (melodic shape, interval sizes, and tonality) more accurately than the younger group. This perceptual shift and exchange in proficiency levels between the recall of absolute and relational musical aspects in younger and older children has been replicated in many other studies as well (Saffran, 2003; Saffran \& Griepentrog, 2001; Sergeant, 1969; Sergeant \& Roche, 1973; Stalinski \& Schellenberg, 2010; Takeuchi \& Hulse, 1993).

Given the prevalence of the relational shift across multiple domains, it is reasonable to assert that any comprehensive theory or model of cognitive development must necessarily account for this phenomenon. One of the models of higher cognition that has successfully been used to account for the relational shift in development is DORA (Discovery Of Relations by Analogy; Doumas, Hummel, \& Sandhofer, 2008). DORA has been used to simulate the relational shift in visual shape perception (Doumas \& Hummel, 2010), analogical problems (Doumas, Morrison, \& Richland, 2009; Morrison, Doumas, \& Richland, 2011), categorical reasoning, spatial reasoning, general relational reasoning, and progressive alignment (Doumas et al., 2008).

In this study, we aim to understand how the relational shift in melodic processing occurs in children. We hypothesize that the same processes that cause the relational shift in other domains are also responsible for the shift in the domain of melodic processing. Specifically, we propose that as children learn about the world, they increasing rely on relational invariants in the environment. This reliance is itself a direct result of the cognitive processes that allow for relational invariants to be detected in the first place. That is, equipped with a cognitive architecture that allows for intersection discovery of shared properties, the natural trend over time (i.e., repeated exposure) is to preferentially
perceive the world in terms of these regularly occurring relational invariants (Doumas \& Hummel, 2010; Doumas et al., 2008).

This raises the question as to how relational invariants are discovered in the first place. Our proposal for this mechanism of discovery is instantiated in DORA's symbolic-connectionist architecture, and has been used to account for how melodic perception occurs in infants (Lim, Doumas, \& Sinnett, 2012). Consequently, providing an account for the relational shift in melodic processing may also help to shed light on other issues. For instance, the argument could be made against the existence of a musical relational shift by citing evidence of infants' ability to detect relational properties from melodies (Plantinga \& Trainor, 2005; Stalinski \& Schellenberg, 2012; Trehub, Bull, \& Thorpe, 1984). That is, given that the relational shift indicates that younger children preferentially process melodies based on absolute percepts (i.e., absolute pitch), would evidence of infants ability to process melodies based on relative percepts (i.e., relative pitch) not be contradictory? Since DORA has been used to simulate the latter phenomenon (Lim et al., 2012), by using DORA to simulate the former phenomenon (i.e., the relational shift in musical processing), we hope to provide an answer to this question as well. \({ }^{1}\)

\section*{The LISA/DORA models}

LISA (Learning and Inference with Schemas and Analogies; Hummel \& Holyoak, 1997, 2003) is a symbolicconnectionist model of analogy and relational reasoning. DORA is an extension of LISA that learns structured (i.e., symbolic) representations of relations from unstructured inputs. That is, DORA provides an account of how the structured relational representations LISA uses to perform relational reasoning can be learned from examples. At present, DORA accounts for over 30 phenomena from the literature on relational learning, and cognitive development, and as it learns representations of relations it develops into LISA and can simulate the additional \(40+\) phenomena in relational thinking for which LISA accounts for (e.g., Doumas et al., 2008). In the following, we provide a very brief description of the LISA/DORA models (for full details, see Hummel \& Holyoak, 1997, 2003; Doumas et al., 2008)

LISAese Representations In LISA (and DORA after it has gone through learning), relational structures are represented by a hierarchy of distributed and localist codes (see Figure 1). At the bottom, "semantic" units (small circles in Figure 1) represent the features of objects and roles in a distributed fashion. At the next level, these distributed representations are connected to localist units (POs) representing individual

\footnotetext{
1 Due to spatial constraints, we provide only summary information on melodic and relational processing here, for more background on melodic processing, including details about absolute and relative pitch and the other features used within these simulations, see Lim et al. (2012).
}
predicates (or roles) and objects (triangles and larger circles in Figure 1). Localist role-binding units (RBs; rectangles in Figure 1) link object and role units into specific role-filler bindings. At the top of the hierarchy, localist P units (ovals in Figure 1) link RBs into whole relational propositions.


Figure 1: LISA/DORA representation of the proposition, chase (dog, cat).

Relational structures (or propositions) are divided into two mutually exclusive sets: a driver and recipient(s). In LISA/DORA, the sequence of firing events is controlled by the driver. Specifically, one (or at most three) proposition(s) in the driver become(s) active (i.e., enter working memory). When a proposition enters working memory, role-filler bindings must be represented dynamically on the units that maintain role-filler independence (i.e., POs and semantic units) to allow for reusability of units and preservation of similarity across different bindings (Hummel \& Holyoak, 1997). In LISA, binding information is carried by synchrony of firing (with roles firing simultaneously with their fillers). In DORA, binding information is carried by systematic asynchrony of firing, with bound role-filler pairs firing in direct sequence (for details, see Doumas et al., 2008). \({ }^{2}\)
Relational Learning In broadest strokes, DORA learns structured representations by comparing objects to isolate their shared properties and to represent these shared properties as explicit structures. More specifically, DORA starts with simple feature-vector representations of objects (i.e., a node connected to set of features describing that object; large and small circles from Figure 1). When DORA compares one object to another, corresponding elements (i.e., shared features) of the two representations fire simultaneously (Figure 2a). Any semantic features common to both objects receive twice as much input and thus become roughly twice as active as features connected to one but not the other (Figure 2b). By recruiting a new PO unit and learning connections between that unit and active semantics via Hebbian learning (wherein the strength of connections is a function of the units' activation), DORA learns stronger connections between the new PO unit and more active

\footnotetext{
\({ }^{2}\) Asynchrony-based binding allows role and filler to be coded by the same pool of semantic units, which allows DORA to learn representations of relations from representations of objects (Doumas et al., 2008).
}
semantic units (Figure 2c). The new PO thus becomes an explicit representation of the featural overlap of the compared objects and can act as a single place predicate, taking other object representations as arguments to form role-filler pairs (Figure 2d; see also Doumas et al., 2008). Applied iteratively, this process allows DORA to learn structured explicit single-place predicate representations of any properties that compared objects may share. Comparison also allows DORA to learn representations of multi-place relations by linking sets of constituent role-filler pairs into relational structures (i.e., to learn the chases relation by linking together representations of the roles chaser and chased; see Doumas et al., 2008 for details). Moreover, when DORA has learned representations of whole relational structures, this intersection discovery algorithm allows it to learn schemas by comparing sets of structural relations to one another. For example, if after DORA has learned about a dog chasing a cat (chase (dog, cat)) and a boy chasing a girl (chase (boy, girl)), it can compare these and learn a representation coding for the intersection of the two chase relations and their arguments, or chase (generic-object1, generic-object2).


Figure 2: The process through which DORA learns a singleplace predicate representation of "higher" from two musical notes.

Mapping For the purpose of analogical mapping, LISA/DORA learns mapping connections between units coactive of the same type in the driver and recipient (e.g., between PO units in the driver and PO units in the recipient). These connections grow whenever corresponding units in the driver and recipient are active simultaneously. They permit LISA to learn the correspondences between matching structures in separate analogs. They also permit correspondences learned early in mapping to influence the correspondences learned later.

\section*{Methods}

In this section we describe the Sergeant and Roche (1973) study, followed by the details and outcomes of DORA's simulations.

\section*{Task Description}

In Sergeant and Roche's (1973) cross-sectional study, children were divided into three groups: one group with children between 3 to 4 years, one with children of 5 years, and one with children of 6 years. All groups received the same training and testing procedures. They were trained to vocally reproduce three melodies from an invariant recording in six training sessions spread out over three weeks. All children were given the exact same melodies at each training session. Each melody lasted for 8 or 16 bars.

One week after training, the children were then asked to vocally recall the melodies, which were tape recorded and scored by two independent judges on perceptual dimension (pitch accuracy), and conceptual dimensions (melodic shape, intervals, and tonality).

\section*{Simulations}

In the first simulation, we simulated the development of representations of individual relations that could define auditory sequence from experience with the world. In the second simulation, we used the representations DORA learned during the first simulation to simulate the behavior of Sergeant and Roche's (1973) subjects. Crucially, these two simulations were interleaved, which allowed us to test
Simulation Part 1 In the first simulation we tested DORA's ability to learn relational concepts from examples. This simulation proceeded like several simulations of relation learning from our previous work (e.g., Doumas \& Hummel, 2005, 2010; Doumas et al., 2008; Doumas et al., 2009). We started DORA with representations of 100 objects (represented as PO units) attached to random sets of features (chosen from a pool of 100). We then defined 4 relations (those that could be used to describe a melodic sequence, e.g., contour (higher/lower), and interval (long-interval, short-interval, medium-interval)).

Each relation transformation consisted of two roles each with three semantic features (e.g., for the higher relation, both the roles above and below were each defined by three specific semantic units). Each of the 100 objects was attached to the features of between 2 and 4 relational roles chosen at random such that if an object was part of a relation, it was attached to the features of one of the roles, chosen at random. For example, objectl might be attached to the features for above (one role of the relation higher) and start-long-interval (the agent role of the relation longinterval). We presented DORA with sets of objects selected at random, and allowed it to compare the objects and learn from the results (as per DORA's relation learning algorithm). As DORA learned new representations it would also use these representations to make subsequent comparisons. For instance, if DORA learned an explicit
representation of the property above by comparing two objects both attached to the features of above, it could use this new representation for future comparisons. On each trial we selected between 2 and 6 representations and let DORA compare them and learn from the results (i.e., perform predication, and relation learning routines). We assume that this act of inspection and comparison is similar to what happens when children encounter objects in the worldwhere objects are part of several relations-and learns from these experiences.

We ran 600 learning trials and measured the quality of the representations DORA had learned after each 100 trials. Quality was calculated as the mean of connection weights to relevant features (i.e., those defining a specific transformation or role of a transformation) divided by the mean of all other connection weights +1 ( 1 was added to the mean of all other connection weights to normalize the quality measure to between 0 and 1). A higher quality denoted stronger connections to the semantics defining a specific transformation relative to all other connections (i.e., a more pristine representation of the transformation). Figure 3 indicates the quality of the representations DORA learned at each level of iteration. Early in learning, DORA's representations are 'dirty' in that it's representations of relations and their roles are also highly connected to extraneous features specific to the instances from which the representations are learned. These representations are consequently very context dependent. As learning progresses however, DORA's representations become progressively more refined. By the end of learning, DORA has learned representations of relations and their roles that are context-independent, connected strongly to only the features specific to the particular relational roles defining the relation and very weakly connected to context features. Thus, in time DORA can use these representations to reason about instances regardless of context, like older children and adults (see, e.g., Doumas et al., 2008).

For the analysis herein, the 'quality' of DORA's representations (how relationally clean or context dependent they are) is considered an analogous measurement to the vocal reproductions of the children in Sergeant and Roche's (1973) study. That is, more pristine representations in DORA would be analogous to children reproducing melodies with more conceptual (relational) dimensions, whereas dirty presentations in DORA would be analogous to children reproducing melodies with more perceptual (absolute) dimensions.

Simulation Part 2 During the second simulation we simulated Sergeant and Roche's (1973) training and test conditions. We created a 20 note melody represented as 20 PO units attached to features indicating absolute frequency (between f 1 and f 24 ), the note's place in the sequence (120), two semantics describing whether the note is higher (above) or lower (below) the previous note in the sequence, two semantics describing the relative interval from the previous note (high-, medium-, low-interval), a semantic
describing the absolute interval from the previous note, and four random features (from a pool of 100). The features represent the properties that infants, children, and adults are capable of representing about melody (Thorpe \& Trehub, 1989; Trehub et al., 1984). Importantly, all of the frequency direction (higher/lower) and frequency interval, both absolute and relative, can be generated from raw frequency values (i.e. sensory input) using a simple comparator circuit described in Doumas et al. (2008) and Hummel and Biederman (1992).


Figure 3: The quality of DORA's representations as a function of learning iterations.

During training, we presented DORA with the note sequence and allowed it to fire each two note sequence in the melody (e.g., notes 1 and 2, then notes 2 and 3). During each two note firing sequence DORA attempted to retrieve relations from LTM describing the sequence (these representations were the same as those DORA had learned during part one of the simulation; see below for details). If DORA successfully retrieved a relation from LTM, DORA predicated the respective roles of the relation about the notes in the sequence. For example, if a two note sequence caused DORA to recall the higher \((x, y)\) relation from LTM (consisting of the roles above \((x)\) and below \((y)\) ), DORA would link the above PO to the note that was higher with an RB unit, and the below PO to the note that was lower. This process reflects our assumption that children and adults attempt to understand melodies using representations at their disposal. After DORA has attempted to classify the 2 note sequences in the melody, DORA stores the resulting representation in LTM.

Importantly, to simulate 4,5, and 6-year olds, we used representations that DORA had learned during the first part of the simulation in DORA's LTM. Specifically, to simulate the representations of 4 year olds, we used the representations that DORA had learned after 200 training iterations, to simulate 5 year olds we used the representations DORA had learned after 400 iterations, and to simulate 6 year-olds we used the representations DORA had learned after 600 iterations. At each age we also included distractor predicates describing extraneous properties (e.g., loudness, timbre, etc.) in LTM. For every relevant relation in DORA's LTM (i.e., relations describing
higher and relative interval) we also included 2 irrelevant relations. Our addition of distractor relations in LTM instantiates our assumption that children learn about multiple relations at the same time during development.

We trained DORA in this manner six times (reflecting the six training sessions from the Sergeant \& Roche (1973) study). After the second training session, and after each subsequent training session, DORA compared the representation it had learned during training to the representation it had learned during the previous training session and learned a new representation (or a schema) using it's learning algorithm.

To simulate the testing phase from Sergeant and Roche's (1973) study, we examined the representation of the melody DORA has learned after the six training sessions. Four-yearold DORA's relational representations were quite dirty and tied to the semantics of the objects from which they had been learned. DORA, consequently, had difficulty retrieving these representations from memory given the melody as a context cue. As a result, the representation of the melody that DORA stores is essentially the melody itself, without much (if any) explicit relational information predicated about it. As DORA get's older (i.e., has its LTM populated with representations produces by more extensive learning during simulation part 1 ), it becomes more likely to retrieve and thus predicate relations about the two note sequences in the melody during training. More precisely, 4 -year old DORA retrieved predicates about only \(18 \%\) of the 2 note sequences it thought about, 5 year-old DORA retrieved predicates about \(63 \%\) of the two note sequences it thought about, and 6 year-old DORA about \(91 \%\) of the instances it though about. Importantly, the predicates in DORA's LTM that it could retrieve varied in their refinement across ages (as described above). We used the representations that DORA had learned after the six training session as a proxy for what it would recall as melody production during the test session of the Sergeant and Roche study. We evaluated these final representations for the presence of relational properties with the assumption that increases in relational properties indicate increased reliance and accuracy on the conceptual dimensions of melodic shape and relative interval. Just as the children in Sergeant and Roche (1973), early in development DORA's ratio of relational/categorical features to absolute features was low, but as DORA learned the ratio increased strongly. At age 4, the ratio value was 0.85 . This value increased to 1.1 at age 5 and 1.6 at age 6 . This progression very closely mirrors the change in reliance on absolute to relational properties observed in children.

\section*{Discussion}

The purpose of this study was to 1 ) test our hypothesis that a common mechanism could potentially underly both the relational shift in melodic perception and the relational shift observed in other domains, and to 2) instantiate this mechanism within a computational model. To this end, our hypotheses were supported by DORA's simulations, which matched the behavioral data from children in Sergeant and

Roche's (1973) study. To our awareness this is the first time the relational shift in melodic processing has been modeled using 1) a neurally plausible architecture, 2) a domaingeneral model of cognition, and 3) the first run of simulations without any parameter fittings.

Consequently, DORA's success in simulating both the relational shift in children's melodic processing in this study, and in simulating infants' ability to detect relational properties of melodies (Lim et al., 2012), provides insights into a misunderstood (what we view as nonexistent) contradiction. Specifically, the argument has been made (e.g., Stalinski \& Schellenberg, 2012) that the evidence for a relational shift in melodic processing may be contradicted by findings that infants can process relational properties of melodies (for a review, see Trehub, 2001). We argue that these two findings are not contradictory, as evidence of a relational shift does not indicate that younger children cannot process relations, only that they show a preference for absolute pitch percepts. As they grow older this preference shifts towards relational melodic features (Takeuchi \& Hulse, 1993). Our theory posits that detection of relational features in melodies by infants (and humans of all ages for that matter) is facilitated by the temporal nature of melodies (each note in the melody sequentially occurs over time), and the corresponding temporality through which our brain encodes and recalls each note (i.e., binding through asynchrony). \({ }^{3}\)

We propose that cognitive systems (e.g., DORA) that use temporality as a binding mechanism between the individual units (notes) of a perceptual group (melodies), is inherently equipped to detect relational invariants within the group (Lim et al., 2012). Through development, learning (i.e., repeated exposure to the environment) occurs and the system inevitably detects more relational invariants, develops cleaner representations that are closer to these invariants (Simulation 1), and learns that this type of information is valuable and predictive. As a result, the system comes to prefer these types of percepts, as observed in the relational shift and predicted by DORA.

It has been proposed that the ability to detect relational properties in melodies may have a common ontogenetic origin as the ability to process vocal speech patterns, where our ability to detect relational melodic features may be a by product of our ability to detect invariants in speech (Terhardt, 1974). Additionally, our model lends support to the notion that absolute (i.e., perfect) pitch-the ability to recount specific note names or frequencies of auditory stimuli-may be a common ability in all humans that is robust in early childhood and subsequently diminishes through development, specifically as the relational shift occurs (Sergeant, 1969; for a review on aboslute pitch, see Takeuchi \& Hulse, 1993). We speculate that the high correlation of musical training during childhood with

\footnotetext{
\({ }^{3}\) Although we propose for time as a binding mechanism, we agnostically acknowledge that other mechanisms could serve a similar function. For a detailed algorithmic level account of DORA's mechanisms, see Doumas et al., 2008.
}
absolute pitch abilities (that subsequently endures into adulthood) may be due to increased exposure to pitch relevant stimuli as young children, and hope to examine such questions in future research.

\section*{References}

Abecassis, M., Sera, M. D., Yonas, A., \& Schwade, J. (2001). Whats in a shape? children represent shape variability differently than adults when naming objects. Journal of Experimental Child Psychology, 78(3), 213239.

Allport, G. W. (1924). Eidetic imagery. British Journal of Psychology. General Section, 15(2), 99-120.
Case, R., \& Khanna, F. (1981). The missing links: Stages in children's progression from sensorimotor to logical thought. New Directions for Child and Adolescent Development, 1981(12), 21-32.
DeLoache, J. S., Sugarman, S., \& Brown, A. L. (1985). The development of error correction strategies in young children's manipulative play. Child Development, 928939.

Doumas, L. A. A., \& Hummel, J. E. (2005). A symbolicconnectionist model of relation discovery. Paper presented at the 23 rd Annual Conference of the Cognitive Science Society.
Doumas, L. A. A., \& Hummel, J. E. (2010). A computational account of the development of the generalization of shape information. Cognitive science, 34(4), 698-712.
Doumas, L. A. A., Hummel, J. E., \& Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. Psychological Review, \(115(1), 1\).
Doumas, L. A. A., Morrison, R. G., \& Richland, L. E. (2009). The Development of Analogy: Working Memory in Relational Learning and Mapping. Paper presented at the 31st Annual Conference of the Cognitive Science Society, Amsterdam, The Netherlands.
Gelman, R., \& Gallistel, C. (1978). The child's understanding of number. Cambridge, MA: Harvard University Press.
Gentner, D. (1988). Metaphor as structure mapping: The relational shift. Child Development, 47-59.
Gentner, D., \& Rattermann, M. J. (1991). Language and the career of similarity. In S. A. Gelman \& J. P. Byrnes (Eds.), Perspectives on language and thought: Interrelations in development (pp. 225-277). London: Cambridge University Press.
Halford, G. S. (2005). Development of thinking. In K. J. Holyoak \& R. G. Morrison (Eds.), The Cambridge handbook of thinking and reasoning (pp. 529-558). New York: Cambridge University Press.
Hummel, J. E., \& Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. Psychological Review, 99(3), 480.
Hummel, J. E., \& Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. Psychological Review, 104(3), 427.

Hummel, J. E., \& Holyoak, K. J. (2003). A symbolicconnectionist theory of relational inference and generalization. Psychological Review, 110(2), 220.
Lim, A., Doumas, L. A. A., \& Sinnett, S. (2012). Modeling Melodic Perception as Relational Learning Using a Symbolic-Connectionist Architecture (DORA). Paper presented at the 34th Annual Conference of the Cognitive Science Society, Sapporo, Japan.
Michie, S. (1985). Development of absolute and relative concepts of number in preschool children. Developmental Psychology, 21(2), 247.
Morrison, R. G., Doumas, L. A. A., \& Richland, L. E. (2011). A computational account of children's analogical reasoning: balancing inhibitory control in working memory and relational representation. Developmental Science, 14(3), 516-529.
Plantinga, J., \& Trainor, L. J. (2005). Memory for melody: Infants use a relative pitch code. Cognition, 98(1), 1-11.
Pollack, R. (1969). Some implications of ontogenetic changes in perception. In Elkind \& Flavell (Eds.), Studies in cognitive development (pp. 365-407): O.U.P.
Saffran, J. R. (2003). Absolute pitch in infancy and adulthood: The role of tonal structure. Developmental Science, 6(1), 35-43.
Saffran, J. R., \& Griepentrog, G. J. (2001). Absolute pitch in infant auditory learning: evidence for developmental reorganization. Developmental Psychology, 37(1), 74.
Sergeant, D. (1969). Experimental investigation of absolute pitch. Journal of Research in Music Education, 17(1), 135-143.
Sergeant, D., \& Roche, S. (1973). Perceptual shifts in the auditory information processing of young children. Psychology of Music, 1(2), 39-48.
Stalinski, S. M., \& Schellenberg, E. G. (2010). Shifting perceptions: Developmental changes in judgments of melodic similarity. Developmental Psychology, 46(6), 1799.

Stalinski, S. M., \& Schellenberg, E. G. (2012). Music Cognition: A Developmental Perspective. Topics in Cognitive Science.
Takeuchi, A. H., \& Hulse, S. H. (1993). Absolute pitch. Psychological bulletin, 113(2), 345.
Terhardt, E. (1974). Pitch, consonance, and harmony. The Journal of the Acoustical Society of America, 55, 1061.
Thorpe, L. A., \& Trehub, S. E. (1989). Duration illusion and auditory grouping in infancy. Developmental Psychology, 25(1), 122.
Trehub, S. E., Bull, D., \& Thorpe, L. A. (1984). Infants' perception of melodies: The role of melodic contour. Child Development, 55(3), 821-830.
Vernon, M. D. (1940). The relation of cognition and phantasy in children. British Journal of Psychology. General Section, 31(1), 1-21.

\title{
Effects of Response and Presentation Format on Measures of Approximate Number System Acuity
}

\author{
Marcus Lindskog (marcus.lindskog@psyk.uu.se), Anders Winman (anders.winman@psyk.uu.se), Peter Juslin (peter.juslin@psyk.uu.se) \\ Department of Psychology, Uppsala University, P. O Box 1225, Uppsala, 75142 Sweden
}

\begin{abstract}
Human adults, infants, and non-human animals are believed to be equipped with an Approximate Number System (ANS) supporting non-symbolic representations of numerical magnitudes. Recent research has questioned both the validity and reliability of tasks intended to measure acuity in the ANS. Issues with validity and reliability might be due to differences in methodology. In the present study, we compare four tasks designed to measure ANS acuity, using a within-subjects design. The tasks are compared with respect to response and presentation format effects previously studied in the psychophysics literature, but largely ignored in the ANS literature. We find a presentation format effect and show that when non-symbolic numerical stimuli are presented sequentially the magnitude of the second stimulus is overestimated. Further, the results indicate that people's sensitivity to differentiate between non-symbolic numerosities is dependent on response format. The implications of the results to measures of ANS acuity are discussed.
\end{abstract}

Keywords: Approximate number system, response format, presentation format, validity, reliability

\section*{Introduction}

Imagine walking in the countryside. As you approach a large field you spot two flocks of sheep, one with only white sheep and one with only black sheep, and amuse yourself by making a snapshot judgment of whether there are more white than black sheep. Later in your walk, you encounter another two flocks of sheep. This time the two flocks emerge from a tunnel, one flock after the other, separated by some short time interval. Once again, you test your judgment skills by deciding which of the two flocks that is the more numerous.

Human adults, infants and non-human animals have a common ability to represent numerical magnitudes, such as the number of sheep, without using symbols (Feigenson, Dehaene, \& Spelke, 2004). The core cognitive system supporting this ability, the Approximate Number System (ANS), represents magnitudes in an approximate fashion with representations becoming increasingly imprecise as numerosity increases (Dehaene, 2009; but see, Brannon, Wusthoff, Gallistel, \& Gibbon, 2001).

The accuracy with which the ANS can represent numerical magnitudes, often referred to as the acuity of the ANS, is conceptualized as the smallest change in numerosity that can be reliably detected and is often quantified by a Weber fraction (w). Acuity in the ANS progresses (i.e., \(w\) decreases) developmentally from
childhood to adolescence (Halberda \& Feigenson, 2008; Halberda, Ly, Wilmer, Naiman, \& Germine, 2012) but even among adults there is considerable individual variability (e.g., Halberda \& Feigenson, 2008; Halberda et al., 2012; Tokita \& Ishiguchi, 2010).

Studies using brain imaging have identified a neurological basis for the ANS in the intraparietal sulcus (IPS) on the lateral surface of the parietal lobe (Castelli, Glaser, \& Butterworth, 2006). Within the IPS, specialized neurons (numerons) sensitive to numerosity have been identified in macaque monkeys (Nieder, Freedman, \& Miller, 2002). The IPS, however, is not only activated by non-numerical stimuli. In humans, it is also activated when they observe numbers in different modalities and when they perform simple arithmetic tasks (Piazza, Izard, Pinel, Le Bihan, \& Dehaene, 2004). That the IPS is activated for a wide variety of numerical stimuli suggests a relationship between ANS acuity and achievement in formal mathematics. Such a relationship has been documented with children, even when controlling for a large number of cognitive abilities, (Halberda, Mazzocco, \& Feigenson, 2008; Inglis, Attridge, Batchelor, \& Gilmore, 2011) but results from studies on adults are mixed (Gebuis \& van der Smagt, 2011; Inglis et al., 2011; Price, Palmer, Battista, \& Ansari, 2012).

At least part of the mixed results might be attributed to differences in methodology. Recently the reliability and validity of some of the most commonly used measures of ANS acuity have been challenged (Gebuis \& Van der Smagt, 2011; Gilmore, Attridge, \& Inglis, 2011; Lindskog, Winman, Juslin, \& Poom, 2013; Price et al., 2012) and studies indicate that while some formats show reasonable reliability and validity others are neither reliable nor valid (Lindskog et al., 2013). The differences in reliability and validity between different tasks that measure ANS acuity highlight the question of whether task features influences performance. Put differently, will you be better at deciding which of the two flocks of sheep that is the more numerous when you see them coming out of the tunnel, one flock at a time, or when you see both flocks at the same time on the field? The question of what factors that influence the reliability of ANS acuity measures is important also because reliability sets an upper limit on correlations between ANS acuity and other cognitive abilities. The present study addresses this question by comparing four ANS acuity tasks, in a within-subjects design, that use different presentation and response formats.

\section*{Response and Presentation Formats}

While ANS acuity tasks use response and presentation formats that have been studied within the psychophysics literature (e.g. Macmillan \& Creelman, 2005) little or no attention has been given to how the choice of format influences the measures of ANS acuity per se. The typical ANS acuity tasks present participants with two arrays of non-symbolic stimuli. Most often the stimuli are dots (e.g., Halberda et al., 2008) but other types of stimuli, for example arrays of squares, have also been used (e.g., Maloney, Risko, Preston, Ansari, \& Fugelsang, 2010). After being presented with the two arrays participants' ability to differentiate between the numerosities of the two arrays is tested using one of two response formats. With a comparison format, similar to the two-alternative forced choice (2AFC) procedure used in psychophysics (Macmillan \& Creelman, 2005), participants indicate which of the two arrays that is the more numerous. For example, Halberda et al. (2008) presented participants with two arrays of dots, one array of yellow dots and one of blue dots, and asked participants to indicate whether blue or yellow was the more numerous color. In contrast, with a discrimination format, similar to a same-different procedure used in classification tasks in psychophysics (Macmillan \& Creelman, 2005), participants are to respond whether the two arrays have the same amount of dots or if the amounts of dots in the two arrays differ. The distinction between the two formats is relevant because even though psychophysicists have long known that the discrimination format is notoriously more difficult than the comparison format (Macmillan \& Creelman, 2005) this difference has not been acknowledged and investigated within the ANSliterature. Therefore, the present study compares the two response formats directly in a within-subjects design.

In addition to the discrimination-comparison distinction, tasks that measure ANS acuity can be classified with respect to how the stimulus is presented temporally. With a simultaneous presentation format, both arrays of stimuli are presented at the same time. For example, in the study by Halberda et al. (2008) the arrays of blue and yellow dots were spatially intermixed and presented on a monitor at the same time. In contrast to the simultaneous presentation format, several studies have employed a sequential presentation format (e.g., Gilmore et al., 2011) where the two arrays are presented one at a time, separated temporally by a short interstimulus interval (ISI). Two reasons make the presentation format distinction important. First, previous research has indicated that while tasks using a simultaneous presentation format exhibit a reasonably good validity, tasks with a sequential presentation format do not (Lindskog et al., 2013). Second, the introduction of an ISI in 2AFC tasks has been shown to introduce a bias, the time-order-error (TOE), where the second stimulus is commonly judged larger more often than the first (Hellström, 1985; Macmillan \& Creelman, 2005). Consequently, a sequence presenting
the two arrays as; Less numerous \(\rightarrow\) More numerous, would be correctly reported more often than the opposite sequence; More numerous \(\rightarrow\) Less numerous. Whether a TOE exists or not in ANS acuity tasks is an empirical and potentially important question. In the present study, we compare the two presentation formats and investigate if a TOE is present in ANS acuity tasks when using the sequential presentation format together with the comparison response format.

\section*{The Present Study}

Experiment 1 was designed to investigate the effects of presentation format and response format on performance in non-symbolic number differentiation tasks (i.e., ANS acuity tasks). To foreshadow the results; the experiment documented a TOE with a sequential presentation format and a comparison response format. Experiment 2 was designed to investigate the origin of the TOE by the use of direct estimates of non-symbolic numerosities. Experiment 1 also documented an effect of response format with better performance in the comparison than the discrimination format. We designed Experiment 3 to investigate if this effect was due to features of the ANS or due to a difference in sensitivity related to task features.

\section*{Experiment 1}

In Experiment 1, participants performed four tasks designed to measure ANS acuity. The tasks were modeled from those used in previous research on ANS acuity. The experiment was designed to compare response formats and presentation formats in general and more specifically to investigate if two classical phenomena documented in the psychophysics literature, the time-order-error and the comparison/discrimination difference, were present in tasks measuring ANS acuity.

\section*{Method}

Participants. Participants (10 Male, 20 Female) were undergraduate students from Uppsala University with a mean age of 26.1 years ( \(S D=6.6\) years). They received a movie ticket or course credits for their participation.

Materials and procedure. Participants carried out a set of four tasks, described in detail below and illustrated in Figure 1. The order of tasks was counterbalanced using a Latin square. In none of the tasks did participants receive feedback on their performance.

Parallel comparison. The parallel comparison task was based on Halberda et al. (2008). On each of the 200 trials, participants saw spatially intermixed blue and yellow dots on a monitor. Exposure time ( 200 ms ) was too short for the dots to be serially counted. We used five ratios between the numerosity of the two arrays of dots (1:2, 3:4, 5:6, 7:8, \(9: 10\) ) with the total number of dots varying between 11 and 30. One fifth of the trials consisted of each ratio. Half of the


Figure 1: Illustration of the parallel (Panel A) and sequential (Panel B) presentation formats together with the comparison (Panel A) and discrimination (Panel B) response format.
trials had blue and half had yellow as the more numerous set. The dots varied randomly in size. To counteract the use of perceptual cues dot arrays were matched for total area on half of the trials and for average dot-size on the other half of the trials. The participants judged which set was more numerous by pressing a color-coded keyboard button.

Sequential comparison. The sequential comparison task used the same stimuli as the parallel comparison task. Here, however, the stimuli were presented sequentially and separated by a 300 ms interstimulus interval. The order of color, and whether the first or second array was the more numerous, was counterbalanced over trials.

Parallel discrimination. The parallel discrimination task presented the stimuli in the same way as the parallel comparison task. Stimuli for half of the trials were created as in the comparison tasks with the same ratios between the numerosity of the two sets of dots and the same total number of dots. For the second half of the trials, both sets of dots (i.e. the blue and yellow set) had the same number of dots. Using the same numerosities as when the two sets differed in the number of dots resulted in the total number of dots varying between 10 and 32. In addition, while the comparison tasks required participants to determine whether the blue or the yellow set of dots was the more numerate, the parallel discrimination task asked participants to determine if the two sets of dots had the same or different amount of dots.

Sequential discrimination. The sequential discrimination task used the same presentation format as the sequential comparison task and the same response format and stimuli as the parallel discrimination task.

\section*{Results and Discussion}

Because the discrimination tasks do not easily allow for the modeling of an individual weber fraction, and because previous research (Lindskog et al., 2013) indicates that proportion correct is just as reliable and valid as \(w\), we used proportion correct as a measure of performance in all of the four tasks.

We compared performance in the four tasks by entering proportion correct as dependent variable into a 2 x 2 repeated measures ANOVA with presentation format
(parallel/sequential) and response format (comparison/discrimination) as within-subjects independent variables. This analysis showed a significant main effect of presentation format \((F(1,29)=55.6, p<.001)\) with better performance with the sequential format ( \(M=.73, S E M=\) \(.009)\) than with the parallel format ( \(M=.67, S E M=.008\) ). There was also a significant main effect of response format \((F(1,29)=546.1, p<.001)\) with higher proportion correct with the comparison ( \(M=.81, S E M=.01\) ) than with the discrimination ( \(M=.59\), SEM \(=.008\) ) format. The two-way presentation format by response format interaction did not reach significance ( \(F<1\) ).There were no effects of the order of the ANS tasks \((F<1)\).

In the two tasks using a sequential presentation format, the dot arrays are presented in one of two orders, either the larger or the smaller array came first. To investigate if this ordering influenced performance we entered proportion correct as dependent variable into a \(2 x 2\) repeated measures ANOVA with response format (comparison/discrimination) and array-size order (larger-smaller/smaller-larger) as within-subjects independent variables. The significant interaction \((F(1,29)=19.2, p<.001)\), illustrated in Figure 2, show that while the array-size order does not influence performance with the discrimination format there is a significant and substantial difference between the two orders with the comparison format.


Figure 2: Proportion correct as a function of response format and array order. Vertical bars denote 95 \% - confidence intervals.

In the comparison format the order which presents the smaller array first leads to significantly higher proportion correct ( \(M=.92, S E M=.012\) ) than the order which presents the larger array first ( \(M=.75, S E M=.022\) ). The analysis thus suggested that there might be a TOE present.

A TOE may occur either because the first stimulus is underestimated, because the second stimulus is overestimated or because it is psychologically easier to detect an increase in numerosity rather than a decrease. We designed Experiment 2 to distinguish between these three possibilities and to investigate the origin of the TOE when using non-symbolic numerosities as stimuli.

The response format effect might emerge for at least two different and independent reasons. First, it might be a feature of the ANS that it is adapted to detect the direction of a difference. For example, the ANS might have developed to determine that bush A contains more berries than bush \(B\), rather than to just determine that there is a difference in the amount of berries on the two bushes. Second, it might be that participants' sensitivity is higher with the comparison format than with the discrimination format as suggested by signal detection theory (Macmillan \& Creelman, 2005). We designed Experiment 3 to distinguish between these two possibilities.

\section*{Experiment 2}

Experiment 2 investigated the origin of the TOE observed in Experiment 1. Participants made direct estimates of the number of displayed dots in a task closely matching the sequential tasks from Experiment 1.

\section*{Method}

Participants. Twenty undergraduate students took part in the study, 12 females and 8 males. Average age was 24.8 ( \(S D=5.49\) ). Participants received a cinema voucher or course credits for their participation.

Stimulus and procedure. Stimuli were three numbers of dots ( 8,11 , and 14) that were presented in temporal sequence in stimulus pairs (e.g. \(8-11,14-8\) etc.) in a randomized order in a fully crossed design (two presentation positions (first/second) by three numerosities (8/11/14)). The dots were either blue or yellow. The sequence of colors was always the same for each participant, but randomized between subjects. Each stimulus pair was presented 9 times. Intermixed with these stimulus pairs each numerosity also occurred in isolation as a control. Together this made up 96 trials per participant. The numerosities were presented for 200 ms , with a blank interstimulus interval of 300 ms . Half of the trials were controlled for average dot-size, half for cumulative area. The task consisted of directly estimating the number of dots. This was done by entering a single number (for control stimuli) or two numbers with the keyboard. The input box was color coded, and always occurred in a left/right fashion corresponding to first/second
position. Participants were told that if they altogether had missed a presentation of stimuli, they could indicate this by entering an error code.

\section*{Results and Discussion}

Stimuli for which participants indicated that they had missed the presentation, as well as outlier responses ( \(|z|>3\) ) were excluded from the analysis. These data made up \(2.3 \%\) of the responses. There were no effects of color sequence order or stimulus type (size/area controlled).

Figure 3 shows judgments for control stimuli that appeared in isolated presentations. As can be seen in the figure, ratings were quite sensitive to the number of dots \((F(2,38)=38.0, p<.001\), one-way repeated measures ANOVA), but with a slight overestimation (the actual number is depicted in the dotted line in the figure).

Figure 4 shows the data of the two presentation positions and different numerosities. As can be seen in the figure, ratings are higher in the second presentation position.


Figure 3: Mean judgments of the three numerosities when presented separately (dotted line shows actual numerosity). Vertical bars denote 95 \% - confidence intervals.


Figure 4: Mean judgments of the three numerosities for each presentation position (first /second). Vertical bars denote 95 \% - confidence intervals.

A two-way ANOVA with numerosity (8/11/14) and stimulus presentation position (first/second) as independent within-subjects variables and judged numerosity as dependent variable shows that this presentation position effect is statistically significant \((F(1,19)=34.5, p<.001)\). The interaction was not significant \((F(2,38)=1.8, p=.17)\)

A one-way ANOVA with condition (control/1st presentation/2nd presentation) as independent variable shows a significant effect on absolute error \((F(2,38)=11.4\), \(p<.001\) ). Error was lowest in the control condition ( \(M=\) \(1.93, S E M=.26\) ), higher in the first presentation position ( \(M\) \(=2.16, S E M=.27\) ) and highest in the second presentation position ( \(M=2.55\), \(S E M=.34\) ). A Scheffé's post hoc test revealed that the error in the second presentation position was statistically significant from the two other conditions, which did not differ significantly from each other. The means of the absolute difference between participants' estimates and the three numerosities 8,11 , and 14 (i.e. the absolute error) were \(1.42,2.16\), and 2.28 respectively. This increase in absolute error was statistically significant \((F(2,38)=6.8, p=.00292)\).

The results of Experiment 2 show that when two numerosities occur in a brief temporal sequence, separated by a short interval, the second numerosity is rated as more numerous than the first, and with a larger error. There is no clear indication of interference in the reversed temporal direction, presenting a second numerosity does apparently not have a deteriorating effect on the judgment of the first numerosity. The results support the one of the proposed explanations for the TOE found in Experiment 1; Participants' better performance with the smaller \(\rightarrow\) larger presentation order than with the larger \(\rightarrow\) smaller order is due to the inflation in experienced numerosity of the second stimulus. This effect leads to participants correctly identifying this order, but will hinder performance on the larger \(\rightarrow\) smaller sequence.

\section*{Experiment 3}

In Experiment 3, we added an extra response alternative to the comparison format. In addition to answering whether the blue or yellow array was the more numerous, participants could also respond that they had the same numerosity. If a feature of the ANS is that it is adapted to detect direction (i.e. to detect that \(A>B\) rather than just that \(A \neq B\) ) we expected better performance with this new response format than with a discrimination format as a result of increased performance on trials with different amount of dots. However, if the effect from Experiment 1 could be attributed to a change in sensitivity, the opposite was expected because adding the extra response alternative would make the task harder.

\section*{Method}

Participants. The participants from Experiment 2 participated in Experiment 3.

Materials and procedure. Half of the participants carried out the two comparison tasks described in Experiment 1. The other half carried out the same tasks but with an alteration to the response format. The alteration combines the response formats of the comparison and discrimination tasks in Experiment 1. In the original discrimination tasks, participants could respond same or different while the comparison format had blue and yellow as response alternatives. In the modified task, participants had three response options: blue, yellow, and same. All other features of the task were identical to the comparison tasks of Experiment 1. The order of tasks was counterbalanced.

\section*{Results and Discussion}

We compared performance in the four tasks by entering proportion correct as dependent variable into a \(2 \times 2\) split-plot ANOVA with presentation format (parallel/sequential) as within-subjects independent variable and response format (same-different/yellow-same-blue) as between-subjects independent variable. Both the main effect of presentation format \((F(1,18)=36.3, p<.001)\) and the main effect of response format \((F(1,18)=20.4, p<.001)\) were significant while the interaction was not \((F<1)\). Participants in the same-different condition performed better ( \(M=.61\), SEM \(=\) .012) than did those in the yellow-same-blue condition ( \(M=\) \(.51, S E M=.012\) ). Further, and replicating the results from Experiment 1, performance was better in the sequential ( \(M\) \(=.60, S E M=.013\) ) than in the parallel presentation format ( \(M=.52, S E M=.011\) ).

These results show that the response format difference from Experiment 1 was eliminated, and even reversed, when an extra response alternative was added to the comparison format. This indicates that the discrimination format is more difficult than the comparison format and that the difference seen in Experiment 1 could be accounted for by a difference in sensitivity. However, even though the results lend support for a sensitivity explanation it does not exclude the possibility that the ANS is adapted not only to detect differences but also to detect the direction of a difference. This should be a question for future research to examine in more detail.

\section*{General Discussion}

Recently, a large body of research has investigated the ANS and its relationship to mathematical achievement. This research has used several different tasks to measure ANS acuity. The present study extends previous research by investigating response and presentation format effects on performance in ANS acuity tasks and shows that comparisons between tasks might not always be straightforward.

In Experiment 1, we found three effects with potentially important implications. First, the sequential presentation format yielded approximately \(8 \%\) (. 72 vs. .67) better performance than the parallel format. In ANS experiments
were w, which is modeled on proportion correct (e.g., Halberda et al., 2008), rather than proportion correct is used as a performance measure this corresponds to a \(30-45 \%\) difference in \(w\) for a typical participant ( \(w=[.15-.20]\) ). Thus, changing the presentation format can give rise to a substantial difference in estimated \(w\).

Second, in the sequential comparison task the order of stimulus was found to affect performance, similar to a TOE. Experiment 2 showed that the effect was due to an overestimation of the second stimulus compared to the control stimulus while no such bias could be found for the first stimulus. While it remains for future research to determine why the second stimulus is overestimated, one possibility could be residual activation in the IPS from the first stimulus. The effect has implications for measurements of ANS acuity. First, it will be necessary for future research using a sequential presentation format to counterbalance the order of stimulus for each ratio. Second, counterbalancing the order of stimulus might not be sufficient if numerosities are used that give rise to asymmetric differences in proportion correct. That is if the gain in one presentation order is larger/smaller than the loss in the opposite order. It remains for future research to investigate such asymmetries.

Finally, performance with a comparison format was significantly better than with a discrimination format. We proposed two possible explanations for this effect, that the ANS is adapted to detect direction or a difference in sensitivity, and showed in Experiment 3 that the latter was supported. This suggests that research on ANS acuity might benefit from, in addition to \(w\) and proportion correct, introducing a measure of sensitivity as a performance measure. The pattern of results, however, does not exclude the possibility of the ANS being a system adapted to detect the direction of a difference. This possibility should be an intriguing question for future research.

\section*{Acknowledgments}

This research was sponsored by the Swedish Research Council. We thank Håkan Nilsson, Philip Millroth, Leo Poom, and Mona Guath for valuable comments on earlier versions of the manuscript.

\section*{References}

Brannon, E. M., Wusthoff, C. J., Gallistel, C. R., \& Gibbon, J. (2001). Numerical subtraction in the pigeon: Evidence for a linear subjective number scale. Psychological Science, 12, 238-243.
Castelli, F., Glaser, D. E., \& Butterworth, B. (2006). Discrete and analogue quantity processing in the parietal lobe: A functional MRI study. Proceedings of the National Academy of Sciences of the United States of America, 103, 4693-4698.
Dehaene, S. (2009). Origins of Mathematical Intuitions. Annals of the New York Academy of Sciences, 1156, 232259.

Feigenson, L., Dehaene, S., \& Spelke, E. (2004). Core systems of number. Trends in Cognitive Sciences, 8, 307314.

Gebuis, T., \& Van der Smagt, M. J. (2011). False approximations of the Approximate Number System? PLoS ONE, 6, e25405.
Gilmore, C., Attridge, N., \& Inglis, M. (2011). Measuring the approximate number system. The Quarterly Journal of Experimental Psychology, 64, 2099-2109.
Halberda, J., \& Feigenson, L. (2008). Developmental change in the acuity of the "number sense": The approximate number system in 3-, 4-, 5-, and 6-year-olds and adults. Developmental Psychology, 44, 1457-1465.
Halberda, J., Ly, R., Wilmer, J. B., Naiman, D. Q., \& Germine, L. (2012). Number sense across the lifespan as revealed by a massive Internet-based sample. Proceedings of the National Academy of Sciences, 109, 11116-11120.
Halberda, J., Mazzocco, M. M. M., \& Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. Nature, 455, 665-668.
Hellström, Å. (1985). The time-order error and its relatives: Mirrors of cognitive processes in comparing. Psychological Bulletin, 97, 35-61.
Inglis, M., Attridge, N., Batchelor, S., \& Gilmore, C. (2011). Non-verbal number acuity correlates with symbolic mathematics achievement: But only in children. Psychonomic Bulletin \& Review, 18, 1222-1229.
Lindskog, M., Winman, A., Juslin, P. \& Poom, L. (2013). Measuring approximate number system acuity reliably. In preparation.
Macmillan, N. A., \& Creelman, C. D. (2005). Detection theory: a user's guide (2nd ed.). Mahwah, N.J: Lawrence Erlbaum Associates.
Maloney, E. A., Risko, E. F., Preston, F., Ansari, D., \& Fugelsang, J. (2010). Challenging the reliability and validity of cognitive measures: The case of the numerical distance effect. Acta Psychologica, 134, 154-161.
Nieder, A., Freedman, D. J., \& Miller, E. K. (2002). Representation of the quantity of visual items in the primate prefrontal cortex. Science, 297, 1708-1711.
Piazza, M., Izard, V., Pinel, P., Le Bihan, D., \& Dehaene, S. (2004). Tuning curves for approximate numerosity in the human intraparietal sulcus. Neuron, 44, 547-555.
Price, G. R., Palmer, D., Battista, C., \& Ansari, D. (2012). Nonsymbolic numerical magnitude comparison: Reliability and validity of different task variants and outcome measures, and their relationship to arithmetic achievement in adults. Acta Psychologica, 140, 50-57.
Tokita, M., \& Ishiguchi, A. (2010). How might the discrepancy in the effects of perceptual variables on numerosity judgment be reconciled? Attention, Perception \& Psychophysics, 72, 1839-1853.

\title{
Effects of Text Titles and the Timing of Keywording Tasks on Metacognitive Monitoring
}

\author{
Marie Lippmann (lippmann@psychologie.tu-dresden.de) \\ Psychology of Learning and Instruction, Technische Universität Dresden, Zellescher Weg 17 \\ 01062 Dresden, Germany \\ Susanne Narciss (susanne.narciss@tu-dresden.de) \\ Psychology of Learning and Instruction, Technische Universität Dresden, Zellescher Weg 17 \\ 01062 Dresden, Germany
}

Neil H. Schwartz (nschwartz@csuchico.edu)
Department of Psychology, 400 West First Street
Chico, CA 95929, USA

\author{
Robert W. Danielson (robert.danielson@usc.edu) \\ Rossier School of Education, 3470 Trousdale Parkway \\ Los Angeles, CA 90089, USA
}

\begin{abstract}
Successful learning from text takes place when the cognitive demands of the learning task - i.e. the comprehension and retention of text material - and the metacognitive demands of the learning task - i.e. the accurate assessment of one's own learning process-are met. The present study was designed to investigate text titles - a factor known to affect cognitive learning processes- as well as the timing of keywording tasks - a factor known to affect metacognitive processes - and their effects on metacognitive monitoring and learning outcomes. The results of the study showed that both factors affected learning on the cognitive as well as the metacognitive level.
\end{abstract}

Keywords: Text-based learning; metacognition.

\section*{Effects of Titles on Learning and Metacognitive Monitoring}

Text titles, a common feature of written text, affect cognitive learning processes and outcomes by: (a) providing a context for an upcoming text (Ausubel, 1968), (b) activating relevant prior knowledge (Ausubel, 1968), and (c) guiding a reader's attention towards certain information in the text (Lorch \& Lorch, 1996). Titles also serve as retrieval cues for previously learned text information (Sadoski, Goetz \& Rodriguez, 2000) and foster the recall of text information that is related to the title (Ritchey, Schuster \& Allen, 2008). While the cognitive effects of titles are well-investigated, it is interesting to consider their potential influence on metacognitive monitoring, as well.

Metacognitive monitoring takes place when learners evaluate their own learning process with respect to a learning goal (Butler \& Winne, 1995; Dinsmore, Alexander \& Loughlin, 2008). In other words, learners engage in metacognitive monitoring whenever they judge their current state of learning relative to a desired state of learning (Thiede \& Dunlosky, 1999). The quality of metacognitive
judgments is influenced by the cues that learners use to make their judgments. According to the Cue-UtilizationFramework (Koriat, 1997), metacognitive judgments are inferential in nature. '[Learners] do not monitor directly the strength of the memory trace of the [to-be-judged information], but use a variety of cues that are generally predictive of subsequent test performance’ (Koriat, 1997, p.2).

The authors believe that titles may function as such cues whenever they are used to prompt learners' metacognitive judgments. Considering that a title related to a text provides a stronger link to relevant information from the text than an unrelated title, related titles should serve as more valid cues for metacognitive monitoring than unrelated titles.

\section*{Effects of Immediate vs. Delayed Keywording on Metacognitive Monitoring}

Aside from cues that arise from the text material, such as titles, learning tasks provide further cues for metacognitive monitoring (Thiede, Anderson \& Therriault, 2003). Keywording tasks are a type of learning task in which learners summarize a previously studied text using a set of keywords. The timing of when learners generate their keywords affects the quality of their metacognitive judgments with respect to recall test performance at a later point in time. Learners who generate keywords immediately after reading a text are less accurate in their metacognitive monitoring than learners who generate keywords after a delay (Thiede, Anderson \& Therriault, 2003; Thiede, Dunlosky, Griffin \&Wiley, 2005).

Current research relates these findings to Activation Theories of Text Understanding (Britton \& Guelgoez, 1991)-theories that describe a spreading activation during reading. More text information is available shortly after reading a text than after a delay, when text information has
decayed in memory. That means learners who generate keywords immediately after reading a text experience a high ease of recall in every keywording task. The high ease of recall in each keywording task makes it hard for learners to distinguish between well and less-well learned texts. It is hard for learners to make that distinction because the performance on an immediate keywording task is not a valid indicator of performance on tests that occur at a later point in time when text information has decayed in memory.

Learners who generate keywords after a delay do not experience the same ease of recall in every keywording task, because the learners need to access text information that has been subject to memory decay to a much larger extent at the time of the keywording task. Hence, the learners may be able to generate only a few keywords for a text that they do not recall well, while they may generate more keywords for a text they recall better. Since delayed keywording requires learners to access text information that has been subject to memory decay to a larger extent than immediate keywording, delayed keywording is a more valid indicator of recall test performance with regard to tests taken at a later point in time. Hence, learners who generate keywords after a delay provide more accurate metacognitive judgments than learners who generate keywords immediately.

While most of the current research has focused on the effects of the timing of keywording tasks on relative monitoring accuracy - i.e. the ability of learners to distinguish between well-learned and less well-learned text, the present study aims to investigate the effects of the timing of keywording tasks on monitoring bias - i.e. the extent to which learners over- or underestimate how much they have learned from a text.

\section*{Purpose of the Present Study and Hypotheses}

In order to develop learning materials that foster successful learning from text, learning materials should be constructed so that they foster learning on the cognitive, as well as on the metacognitive level. Thus, the present study was designed to investigate how related vs. unrelated text titles, and immediate vs. delayed keywording, affect metacognitive monitoring and learning outcomes.

The dependent measures of the study were comprised of a set of cognitive and metacognitive measures, namely a) performance on a keywording task as measured by the number of keywords correct, b) metacognitive judgments of learning for each text as measured by ratings on a 6-point Likert scale, and c) recall test performance as measured by the number of idea units recalled correctly in a free-recall essay task. Monitoring bias (d) was calculated by relating learners' metacognitive judgments to their recall test performance using the Self-Criterion-Residual-Strategy (Paulhus \& John, 1998).

With regard to the objectives of the present study, the authors aimed to investigate the following hypotheses:

\section*{Hypotheses: Effects of Titles}
1. Titles affect learning outcomes - titles related to a text serve as more valid retrieval cues than unrelated titles. Thus, the authors expect higher recall test performance for texts with related titles in both, the keywording task (Hypothesis 1.1) and an essay task (Hypothesis 1.2).
2. Titles affect metacognitive monitoring - titles related to a text serve as more valid cues for metacognitive monitoring than unrelated titles. Thus, the authors expect unrelated titles to evoke a stronger monitoring bias than titles related to the text.

\section*{Hypotheses: Effects of Immediate vs. Delayed Keywording}
3. The timing of keywording tasks affects learning outcomes in the keywording task (Hypothesis 3.1), but not in the essay task (Hypothesis 3.2). Learners who generate keywords immediately after reading a text have access to text information that is presumed to still be rather active in their memory. They experience a high ease of recall in every keywording task and are able to generate many correct keywords. Learners who generate keywords after a delay need to access information in their memory that has been subject to decay to a much larger extent. They do not experience the same ease of recall as learners in the immediate keywording group and, thus, are expected to generate a smaller number of correct keywords. The authors do not expect to find the same effect in the essay task, because, in the essay task, learners in both keywording conditions have to rely on text information that has been subject to memory decay to the same extent (i.e. about the same amount of time has passed in between reading and essay writing).
4. The timing of the keywording task affects metacognitive monitoring - learners who generate keywords immediately after reading a text experience a high ease of recall in every keywording task, which may cue them to overestimate their ability to retrieve the same text information at a later point in time, when memory activation for text information has decayed. Thus, the authors expect learners in the immediate keywording group to show a stronger overestimation bias than learners in the delayed keywording group.

\section*{Hypotheses: Interactive Effects of Titles and Immediate vs. Delayed Keywording}
5. Titles and the timing of keywording tasks interact with the learning outcomes in the keywording task (Hypothesis 5.1.), but not with the essay task (Hypothesis 5.2). While learners who generate keywords after a delay rely on titles as retrieval cues, learners in the immediate keywording group do not, because the text they just read is presumed to still be rather active within memory. Thus, the authors expect learners in the immediate keywording group to generate
more correct keywords for texts with unrelated titles than learners in the delayed keywording group. The authors do not expect to find the same effect in the essay task, in which learners in both keywording groups have to access text information that has been subject to memory decay to the same extent.
6. Titles and the timing of keywording tasks interact with metacognitive monitoring - learners who generate keywords immediately are expected to overestimate how much they learned from texts with unrelated titles. Learners who generate keywords after a delay are expected to show less of an overestimation bias with regard to texts with unrelated titles.

\section*{Methods}

Participants. 213 undergraduate students of an American university - 56 males and 157 females - participated in the study. Participant's ages ranged from \(18-57\) years ( \(\mathrm{M}=\) 22.2).

Design. The study follows a 2-Keywording (Immediate vs. Delayed) x 3-Title (Related/Close vs. Related/Distant vs. Unrelated) - design with repeated measures on the factor 'Titles'. The order of topic and title appearance was balanced within a Latin Square.
Materials. The study was conducted online. Materials were comprised of 6 expository texts derived from online databases and modified to suit the purpose of the study. Each expository text consisted of 2 distinct subtopics of an overall related theme. Themes varied for each text and were chosen from topics which are neither part of the standard US high school curriculum, nor part of the standard undergraduate curriculum at the university from which participants were recruited. The text concerning the overall theme of 'Art', for example, was comprised of the subtopics 'Expressionist Painting' and 'Dualism in Art'. To control for confounding effects between a topic and its position in the text, the order of topic appearance was counterbalanced within a Latin Square, so that every participant experienced every title condition twice throughout the study. In order to control for confounding effects of text position, the order of text appearance was also balanced within the Latin Square. Each subtopic in a text consisted of 30 idea units. Idea Units were defined as "single, meaningful piece[s] of information conveyed by the passage, whether [they] consisted of a word, a definition, or a phrase in the passage" (Meyer, 1975). The subtopics were balanced for word count (range: 190-284 words) and readability (Flesh-Kincaid readability score; range: 11-13). The readability range was chosen to match the target participant group of undergraduate university students. Each text was accompanied by one of three titles - a title that was related to the first subtopic in the text (Related/Close), a title that was related to the second subtopic in the text (Related/Distant), or a title that was unrelated to either of the subtopics in the text (Unrelated). While the authors had explicit hypotheses on the effects of related versus unrelated titles on metacognitive monitoring and learning outcomes, the distinction between

Related/Close and Related/Distant titles was made in order to detect whether the position of the related information in the text would have distinct effects on metacognitive monitoring and learning outcomes.

Study procedure. Participants were randomly assigned to the immediate or delayed keywording condition. Each participant read 6 texts and was instructed to learn as much from them as possible. Each text was presented for 2.5 minutes. Participants were asked to generate a maximum of 6 keywords prompted by the title, in order to capture the main gist of each text. The immediate-keywording group generated keywords immediately after reading each text. The delayed-keywording group generated keywords only after reading all 6 texts. After reading and keywording, participants provided a metacognitive judgment of learning on a 6 -point Likert scale \((1=\) learned very little to \(6=\) learned very much). Then, the text titles were presented one at a time, and participants were asked to write essays about what they remembered from the text. The time limit for each essay was 3 minutes. Reading and writing times were controlled in order to encourage participants to engage in each task thoroughly. Reading and writing times were allocated according to data derived from a pilot study conducted prior to the actual investigation.

\section*{Results}

Keywording task. Keywords were scored using a 4-category scoring rubric. Keywords could be correct, incorrect, missing, or they could be derived from prior knowledge, rather than from the text. Only correct keywords were included in the keyword analyses. The results of the keywording task were analyzed using a 2-keyword (Immediate vs. Delayed) x 3-title (Related/Close vs. Related/Distant vs. Unrelated) Analysis of Variance (ANOVA) with repeated measures on the factor 'titles' and Bonferroni correction for multiple testing.

The ANOVA revealed a significant main effect for the timing of keywording tasks \([\mathrm{F}(1,211)=132.64\); MSerror \(=\) 2.71; \(\mathrm{p}<0.01\); partial \(\eta 2=0.39\) (large effect)]. Learners who generated keywords immediately after reading a text were able to generate more correct keywords ( \(\mathrm{M}=3.97\); SD \(=0.09)\) than learners who generated keywords after a delay ( \(\mathrm{M}=2.47\); \(\mathrm{SD}=0.09\) ).

The ANOVA also revealed a significant main effect for titles \([\mathrm{F}(2,422)=20.86\); MSerror \(=1.19 ; \mathrm{p}<0.01\); partial \(\eta 2=0.09\) (moderate effect)]. Learners generated more correct keywords when the title was related to the text ( M ~ 3.41; \(\mathrm{SD}=0.09\) ), than when the title was unrelated ( \(\mathrm{M}=\) 2.82; \(\mathrm{SD}=0.09\) ). The number of correctly generated keywords did not differ significantly depending on whether the title-related information was stated first in the text \(\left(\mathrm{M}_{\mathrm{RC}}\right.\) \(\left.=3.41 ; \mathrm{SD}_{\mathrm{RC}}=0.09\right)\), or second \(\left(\mathrm{M}_{\mathrm{RD}}=3.42 ; \mathrm{SD}_{\mathrm{RD}}=0.09\right)\). In other words, learners generated more correct keywords as long as the title was related to the text, no matter in which position the related information appeared.

The main effects of keywording and title conditions were further qualified by a significant two-way interaction
between the timing of keywording tasks and titles [ F (2, 422) \(=11.95\); MSerror \(=1.19 ; p<0.01\); partial \(\eta^{2}=0.05\) (small effect)]. Learners who generated keywords immediately after reading a text had no problem generating keywords for texts with unrelated titles ( \(\mathrm{M}=3.87\); \(\mathrm{SD}=\) 0.13 ). Learners who generated keywords after a delay, on the other hand, generated a smaller number of correct keywords for texts with unrelated titles ( \(M=1.78\); \(\mathrm{SD}=\) \(0.13)\).

Metacognitive judgments of learning. Metacognitive judgments of learning were analyzed using a 2 -keyword (Immediate vs. Delayed) x 3-title (Related/Close vs. Related/Distant vs. Unrelated) ANOVA with repeated measures on the factor 'titles' and Bonferroni correction for multiple testing. The results of the ANOVA revealed a significant main effect for keywording conditions [F (1, 211) \(=7.47\); MSerror \(=1.66 ; \mathrm{p}<0.01\); partial \(\eta^{2}=0.03\) (small effect)]. Learners who generated keywords immediately after reading a text provided higher judgments of learning ( \(\mathrm{M}=3.14\); \(\mathrm{SD}=0.07\) ) than learners who generated keywords after a delay ( \(M=2.86\); \(S D=0.07\) ).
The ANOVA also revealed a significant main effect for titles [F \((2,422)=39.62\); MSerror \(=0.68 ;\) p \(<0.01\); partial \(\eta 2=0.16\) (moderate effect)]. Learners provided higher judgments of learning for texts with related titles ( \(\mathrm{M} \sim 3.2\); SD \(=0.07\) ) than for texts with unrelated titles ( \(M=2.60\); SD \(=0.07\) ). The judgment magnitude did not vary significantly depending on whether the related information was stated close to the title \(\left(\mathrm{M}_{\mathrm{RC}}=3.25\right.\); \(\left.\mathrm{SD}_{\mathrm{RD}}=0.07\right)\), or distant from it \(\left(\mathrm{M}_{\mathrm{RD}}=3.16 ; \mathrm{SD}_{\mathrm{RD}}=0.07\right)\).

Essay task performance. Essays were scored for idea units using a 5 -category scoring rubric. Recalled idea units could be correct, incorrect, partially correct to \(50 \%\) or \(25 \%\), or correct, but derived from prior knowledge rather than from the text. Only partially or fully recalled idea units derived from the texts were included in the essay analysis. The results of the essay task were analyzed with a 2 -keyword (Immediate vs. Delayed) x 3-title (Related/Close vs. Related/Distant vs. Unrelated) ANOVA with repeated measures on the factor 'titles' and Bonferroni correction for multiple testing.

The ANOVA revealed a significant main effect for titles \(\left[\mathrm{F}(2,422)=21.14 ;\right.\) MSerror \(=3.54 ; \mathrm{p}<0.01\); partial \(\eta^{2}=\) 0.09 (moderate effect)]. Learners recalled more idea units from text with related titles ( \(\mathrm{M} \sim 4.9\); \(\mathrm{SD} \sim 0.21\) ) than from texts with unrelated titles ( \(\mathrm{M}=3.89\); \(\mathrm{SD}=0.21\) ). The number of idea units recalled did not vary significantly depending on whether title-related information was stated first in the text \(\left(\mathrm{M}_{\mathrm{RC}}=4.98 ; \mathrm{SD}_{\mathrm{RC}}=0.22\right)\), or second \(\left(\mathrm{M}_{\mathrm{RD}}\right.\) \(=4.84 ; \mathrm{SD}_{\mathrm{RD}}=0.21\) ).

Metacognitive monitoring bias. Metacognitive monitoring bias was computed using the Self-Criterion-ResidualStrategy (SCR-Strategy: Paulhus \& John, 1998). For SCRAnalyses, self-reports (i.e. metacognitive judgments of learning) are regressed on an external criterion (i.e. essay task performance). The standardized residuals are used as indices for monitoring bias (i.e. the extent to which an
individual's monitoring accuracy differs from the average monitoring accuracy observed in the participant sample). The closer the standardized residual is to 0 , the more accurate the learner. Standardized residuals with negative values indicate underestimation, while standardized residuals above 0 indicate overestimation.
In a first step, the metacognitive judgment of learning for each text was regressed on the learner's essay task performance on that text and the standardized residuals from these simple regressions were saved. In a second step, the mean standardized residual for each title condition was computed for each participant. In a third step, a 2-keyword (Immediate vs. Delayed) x 3-title (Related/Close vs. Related/Distant vs. Unrelated) ANOVA with repeated measures on the factor 'titles' and Bonferroni correction for multiple testing was computed on the mean standardized residuals for each title condition.
The ANOVA revealed a significant main effect for keywording conditions \([F(1,211)=5.72\); MSerror \(=0.95\); \(p=0.02\); partial \(\eta^{2}=0.03\) (small effect); see figure 1]. In general, learners showed virtually the same monitoring bias ( \(\mathrm{M}=|0.09|\); SD \(\sim 0.05\) ) in both keywording groups - except that learners who generated keywords immediately tended to overestimate how much they had learned ( \(\mathrm{M}=0.09\); SD \(=0.05\) ), while learners who generated keywords after a delay tended to underestimate how much they had learned ( \(\mathrm{M}=-0.09 ; \mathrm{SD}=0.06\) ).
The ANOVA also revealed a significant main effect for titles [F \((2,422)=29.12\); MSerror \(=0.39 ; \mathrm{p}<0.01\); partial \(\eta^{2}=0.12\) (moderate effect); see figure 1]. Learners tended to overestimate how much they had learned when texts were related to the title, while it did not matter whether the titlerelated information appeared first in the text \(\left(\mathrm{M}_{\mathrm{RC}}=0.16\right.\); \(\left.\mathrm{SD}_{\mathrm{RC}}=0.05\right)\), or second \(\left(\mathrm{M}_{\mathrm{RD}}=0.11 ; \mathrm{SD}_{\mathrm{RD}}=0.05\right)\). When texts were unrelated to the title, learners tended to underestimate themselves instead \(\left(\mathrm{M}_{\mathrm{UR}}=-0.26 ; \mathrm{SD}_{\mathrm{UR}}=\right.\) \(0.05)\).


Figure 1: Metacognitive monitoring bias at timing of keywording task (Immediate vs. Delayed) x titles (Related/Close vs. Related/Distant vs. Unrelated).

\section*{Discussion}

\section*{Effects of Titles and Qualifying Interactions}

Hypothesis 1: The authors expected titles to influence learning outcomes in both - the keywording task (Hypothesis 1.1) and the essay task (Hypothesis 1.2). The results of the study are in line with the hypotheses - learners generated more correct keywords for texts with related titles, than for texts with unrelated titles. Learners also recalled more idea units from texts with related titles, than from texts with unrelated titles. These findings support the idea that related titles provide a stronger link to relevant text information than unrelated titles, and, thus, serve as more valid retrieval cues for recalling text information than unrelated titles.

It is important to note that the timing of the keywording task influenced how strongly learners relied on titles when generating keywords (Hypotheses 5.1). While learners in the immediate keywording group were able to generate almost as many correct keywords for texts with unrelated titles (M \(=3.87\); \(\mathrm{SD}=0.13\) ) as for texts with related titles \((\mathrm{M}=4.01\); \(\mathrm{SD}=0.13\) ), learners in the delayed keywording group generated less correct keywords for texts with unrelated titles \((M=1.77 ; S D=0.13)\) than for texts with related titles ( \(\mathrm{M}=2.81 ; \mathrm{SD}=0.12\) ). This finding is in line with the assumptions of Activation Theories of Text Understanding (Britton \& Guelgoez, 1991) suggesting a spread of activation during reading. Learners who generated keywords immediately after reading a text were able to access text information that was presumably still active within memory. That is, the learners did not have to rely on the title as a retrieval cue to the same extent as learners in the delayed keywording group. The delayed keywording learners needed to access text information from memory that had decayed to a much larger extent at the time of their keywording task. This interaction was not observable in the essay task (Hypothesis 5.2), because for the essay task, learners in both keywording groups had to access text information in their memory that had been subject to decay. About the same amount of time had passed in between reading and essay writing in both keywording groups. Thus, learners in the immediate keywording group could not rely on information that was presumably active within memory for the essay task, but needed to access information that had decayed.

Hypothesis 2: The authors expected titles to affect metacognitive monitoring. The authors specifically hypothesized that related titles would serve as more valid cues for making metacognitive judgments than unrelated titles, resulting in a smaller monitoring bias for texts with related, than for texts with unrelated titles. The results of the study showed that monitoring bias was indeed influenced by the title conditions. Learners tended to overestimate how much they had learned from texts with related titles, while they tended to underestimate how much they had learned from texts with unrelated titles. The strength of the monitoring bias differed between title conditions in the way
the authors hypothesized - while learners just slightly overestimated how much they had learned from texts with related titles (Mean standardized residual \(=0.14\) ), they underestimated how much they had learned from texts with unrelated titles to a much larger extent (Mean standardized residual \(=-0.26\) ). This finding supports the idea that related titles serve as more valid cues for making metacognitive judgments than unrelated titles.

\section*{Effects of Immediate vs. Delayed Keywording}

Hypothesis 3: The authors expected the timing of the keywording task to influence the number of correctly generated keywords (Hypothesis 3.1), but not the number of correctly recalled idea units in the essay task (Hypothesis 3.2.) The results of the study provided evidence for these hypotheses. Learners in the immediate keywording group generated more correct keywords ( \(\mathrm{M}=3.97\); \(\mathrm{SD}=0.09\) ) than learners in the delayed keywording group ( \(\mathrm{M}=2.47\); \(\mathrm{SD}=0.09\) ). This finding again supports the assumptions of Activation Theories of Text Understanding (Britton \& Guelgoez, 1991). Learners can easily access text information shortly after reading, while it is harder to access text information after a delay when memory activation has decayed. This effect was not observable in the essay task anymore, because for the essay task, learners in both keywording groups had to access text information in their memory that had been subject to decay to the same extent, i.e. about the same amount of time had passed in between reading and essay writing in both keywording groups.
Hypothesis 4: The authors expected the timing of the keywording task to influence metacognitive monitoring. The authors specifically hypothesized that learners in the immediate keywording group would show a stronger overestimation bias than learners in the delayed keywording group. The results of this study support this hypothesis. While learners in the immediate keywording group showed a slight overestimation bias ( \(M=0.09\); \(\mathrm{SD}=0.05\) ), learners in the delayed keywording group showed a slight underestimation bias ( \(M=-0.09\); \(S D=0.06\) ). It is important to note that this result was influenced by the effects of the unrelated title condition, although the authors could not detect the hypothesized interaction (Hypothesis 6). That is, the general underestimation bias evoked by unrelated titles decreased the mean monitoring bias in both keywording conditions.
In order to detect whether the direction of monitoring bias was overall affected by the unrelated title condition, the authors removed the effects of the unrelated title condition from the analysis by conducting a separate analysis for texts with related titles only. That is, the authors compared the mean standardized residual for texts with related titles in the immediate keywording condition \((\mathrm{M}=0.2 ; \mathrm{SD}=0.08)\) to the mean standardized residual for texts with related titles in the delayed condition ( \(\mathrm{M}=0.07\); \(\mathrm{SD}=0.08\) ). The t -test revealed a significant difference between the keywording groups ( \(\mathrm{t}(211\) ) \(=-11.86 ; \mathrm{p}<0.01\) ). While learners in both keywording groups tended to generally overestimate how
much they learned from texts with related titles, as indicated by mean standardized residuals above 0 , learners who generated keywords immediately showed a significantly stronger overestimation bias than learners who generated keywords after a delay. This finding is in line with former research investigating the delayed keywording effect (Thiede, Anderson \& Therriault, 2003). The authors believe that this effect is due to the high ease of recall that learners experience in the immediate keywording task, as indicated by the large amount of keywords generated correctly. The ability to generate a large number of keywords may cue learners to believe that they have learned the text information well and that they will be able to recall it at a later point in time, as indicated by higher judgments of learning in the immediate keywording group. Yet, a learner's performance on an immediate keywording task is not a valid indicator of performance in the essay task, which takes place at a later point in time when text information has been subject to memory decay. Thus, learners who generate keywords immediately tend to show a strong overestimation bias due to the ease of recall they experience in their keywording task. Learners who generate keywords after a delay, on the other hand, need to access text information that has already been subject to memory decay to a larger extent and that is a much better indicator of performance in the essay task, which takes place after an even larger delay. That means that learners in the delayed keywording group do not experience an ease of recall that could cue them to overestimate themselves to the same extent as learners in the immediate keywording group, resulting in more accurate metacognitive monitoring in the delayed keywording group, as compared to the immediate keywording group.

\section*{Conclusions and Outlook}

The present study contributes to the current literature in three ways. First, the authors could show that titles do not only affect cognitive learning processes and learning outcomes, but also metacognitive monitoring, with related titles functioning as more valid cues for making metacognitive judgments than unrelated titles. Second, the results of the study showed that delayed keywording does not only foster relative monitoring accuracy (i.e. the ability to distinguish between well and less-well learned texts), but also prevents overestimation bias (i.e. the tendency to overestimate oneself) to a larger extent than immediate keywording. Third, the results of the study showed that titles and the timing of keywording tasks interact with regard to certain learning tasks. This finding raises the question of how closely cognitive and metacognitive processes are related - a question that may be very interesting to investigate in further studies.

\section*{References}

Ausubel, D. (1968). Educational psychology: A cognitive view. New York: Holt, Rinehart, and Winston.

Britton, B. K., \& Gülgöz, S. (1991). Using Kintsch’s computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. Jounal of Educational Psychology, 83, 329345.

Butler, D. L. \& Winne, P. H. (1995). Feedback and selfregulated learning: A theoretical synthesis. Review of Educational Research, 65, 245-281.
Dinsmore, D.L., Alexander, P.A. \& Loughlin, S.M. (2008). Focusing the conceptual lens on metacognition, selfregulation, and self-regulated learning. Educational Psychological Review, 20, 391-409.
Koriat, A. (1997). Monitoring one’s own knowledge during study: A cue-utilization approach to judgments of learning. Journal of Experimental Psychology: General, 126, \(349-370\).
Lorch, R. F., Jr. \& Lorch, E. P. (1996). Effects of organizational signals on free recall of expository text. Journal of Educational Psychology, 88, 38-48.
Meyer, B. J. F. (1975). The organization of prose and its effects on memory. New York: American Elsevier Publishing.
Paulhus, D.L. \& John, O.P.(1998). Egoistic and moralistic biases in self-perception - The interplay of self-deceptive styles with basic traits and Motives. Journal of Personality, 66 (6).
Ritchey, K., Schuster, J. \& Allen, J. (2008). How the relationship between text and headings influences readers‘ memory. Contemporary Educational Psychology, 33, 859874.

Thiede, K. W., Anderson, M. C. M. \& Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. Journal of Educational Psychology, 95 (1), 66- 73.
Thiede, K. W., \& Dunlosky, J. (1999). Toward a general model of selfregulated study: An analysis of selection of items for study and selfpaced study time. Journal of Experimental Psychology: Learning, Memory, and Cognition, 25, 1024-1037.
Thiede, K.W., Dunlosky, J., Griffin, T.D. \&Wiley, J. (2005). Understanding the Delayed-Keyword Effect on Metacomprehension Accuracy. Journal of Experimental Psychology, Learning, Memory, and Cognition, 31 (6). Sadoski, M., Goetz, E.T. \& Rodriguez, M. (2000). Engaging Texts: Effects of concreteness on comprehensibility, interest, and recall in four text types. Journal of Educational Psychology, 92 (1), 85 - 95.

\section*{Acknowledgements}

The present study was funded by the German Academic Exchange Service (DAAD), as well as the Gesellschaft von Freunden und Förderern der TU Dresden (GFF). The authors thank David Philhour (CSU Chico) for statistical support and insight.

\title{
The relationship between blocking and inference in causal learning.
}

\author{
Evan J. Livesey (evan.livesey@sydney.edu.au) \\ Jessica C. Lee (jlee4128@uni.sydney.edu.au) \\ Lauren T. Shone (lsho0771@uni.sydney.edu.au) \\ School of Psychology, University of Sydney \\ Sydney, NSW 2006 Australia
}

\begin{abstract}
The blocking effect in causal learning, once taken as a hallmark of associative learning, has recently been explained in terms of an explicit deductive reasoning process. Yet when the conditions necessary for deduction are removed, a small blocking effect is often still present. We examined the relationship between blocking and participants' performance on analytical thinking and probabilistic reasoning measures. Inferential processes predict blocking or an absence of blocking in this situation, depending on the observer's consideration of conditional probabilities. Although Bayesian inference predicts blocking, most individuals are not inclined to use this form of probabilistic reasoning explicitly, an observation we confirmed using a logical problem with similar properties to the relationships present in the blocking effect. Furthermore, participants who showed the greatest capacity for analytical reflection were less likely to show a blocking effect, suggesting that blocking in causal learning is the product of an intuitive and unreflective thought process.
\end{abstract}

Keywords: Blocking; causal learning; inferential reasoning; associative learning; Bayesian inference.

\section*{Introduction}

Many theories of causal learning assume that when individuals make judgments about the relationship between putative causes and their effects, some form of inferential reasoning is involved. However, theories differ substantially in how they place inferential reasoning amongst other contributing mechanisms. Some authors have argued that all causal judgments are necessarily the product of explicit inferential processes based on consciously mediated propositions about the relationships between events (Mitchell, De Houwer and Lovibond, 2009). Others assume that in making causal judgments about a cue, relatively automatic memory retrieval mechanisms based on associative learning play a much greater role, bringing to mind the events that were previously paired with that cue. According to this account, inferential thoughts of an analytical nature - for instance based on formal logic and reasoning - play a smaller role, in some cases perhaps only when strongly encouraged.

\section*{Blocking in causal learning}

The blocking effect has become an important test bed for these arguments. In a typical blocking experiment, one cue (A) is presented and is reliably followed by a particular outcome. In a second stage, \(A\) is presented with another cue (B) and this compound of two cues is followed by the
outcome. B is never presented by itself and its relationship with the outcome is thus ambiguous. When asked to give a rating of the extent to which each of a number of cues causes the outcome, participants often given a lower rating for B than for control cues ( C and D ) that were also presented in compound and followed by the outcome but were never presented on their own.

The cues and outcomes are often presented within a hypothetical scenario. For instance, in the allergist task, the participant assumes the role of a doctor trying to determine the cause of a patient's allergic reactions. The participant might observe that when the patient eats Fish they suffer from a reaction \((\mathrm{A}+)\), and later when the patient eats Fish and Rice \((\mathrm{AB}+)\), they suffer from the same allergic reaction. The patient might also suffer from an allergic reaction after eating Mushrooms and Pasta (CD+), but does not suffer a reaction after eating various other foods (e.g. E-). After learning to predict what will happen after certain meals, through a process of trial and error, the participant must then make an explicit judgment about the extent to which a food or foods cause the allergic reaction, or the likelihood that a reaction will occur given that certain foods have been consumed.

The blocking effect is well documented in causal learning experiments using the allergist task and other similar scenarios. Its presence was originally taken as evidence that a similar associative learning process was responsible for causal learning and conditioning in humans and other animals because blocking in classical conditioning is widely replicated and well explained by associative learning theories (Dickinson, Shanks, and Evenden, 1984). Several other prominent theoretical approaches to causal reasoning also provide explanations of blocking (e.g. Cheng, 1997; Griffiths, Sobel, Tenenbaum, Gopnik, 2011; Waldmann, 2000). Whether based on associations or statistical computation, many theories of causal learning share an assumption that causal judgments partly reflect an implicit sensitivity to the contingencies between observed events. This sensitivity allows the observer to make judgments about causation with little deliberate mental effort, even when the causal relationships between cues and outcomes are ambiguous and must be inferred indirectly, as in the case of blocking (e.g. see Sternberg and McClelland, 2011).

\section*{Blocking and inferential reasoning}

Recently, several authors have argued for an explanation of blocking that relies only on inferential reasoning based
upon a relatively simple set of propositions (De Houwer, Beckers, \& Glautier, 2002; Lovibond, Been, Mitchell, Bouton, \& Frohardt, 2003). Proponents of this account point out that there are circumstances under which the observer can logically deduce that the blocked cue (B) is not a cause of the outcome. For instance, this position is reached if one assumes that the effects of the patient's allergies are additive and that a more severe reaction could be observed if it were present. Holding these assumptions, if one does not observe an increase in the severity of the outcome when B is eaten at the same time as the allergenic food A , then one can deduce that B does not contribute to the allergic reaction. For example, if eating Fish causes an allergic reaction of severity 5 (on a fictitious allergy scale with a maximum of 10) and eating Fish and Rice also causes an allergic reaction of severity 5 , then Rice has not made the reaction worse and thus probably isn't a cause of the reaction itself. Consistent with this inferential reasoning hypothesis, Lovibond et al. (2003; see also De Houwer et al., 2002; Livesey \& Boakes, 2004) observed that pretraining and explicit instructions that encourage this outcome additivity assumption enhance the blocking effect.

Lovibond et al. (2003) also argued that if the observer assumes that the effects of the causal cues do not add to create a larger effect then this deduction is no longer valid and therefore there should be no blocking observed. This "nonadditive" assumption is encouraged by explicitly showing that the addition of two causes does not result in a stronger outcome than one cause on its own. According to this argument, participants with an assumption that the outcome is nonadditive should identify that they cannot be certain of the causal status of B, any more than the control cues \(C\) and \(D\), and thus give each of these cues an equivalent causal rating that reflects that uncertainty.

In practice, a statistically robust blocking effect is often observed even after explicit nonadditive pretraining, albeit one that is numerically smaller than after additive pretraining (e.g. Lovibond et al., 2003; Mitchell, Lovibond, Minard, \& Lavis, 2006). The presence of this persistent blocking effect has been viewed by some as a problem for the inferential reasoning account of causal learning because blocking after nonadditive pretraining is not the result that a participant would generate when applying inferential reasoning in a rational way (Lovibond et al., 2003).

Yet it is worth noting that, at least from the perspective of classical probability theory, this blocking effect is entirely rational. For both the blocking and control cases, the problem involves determining the probability of the hypothesis that a certain cue, \(X\), is a reliable cause of the outcome, \(\mathrm{p}(\mathrm{X}+)\). Relevant information is gained from observing that X in compound with another cue does cause the outcome \((\mathrm{XY}+\) ). Thus the problem becomes one of calculating the conditional probability that X is a cause of the outcome given the observation that the compound XY causes the outcome, \(\mathrm{p}\left(\mathrm{X}^{+} \mid \mathrm{XY}+\right)\). We can use Bayes' theorem to calculate this conditional probability as follows:
\[
p(X+\mid X Y+)=\frac{p(X Y+\mid X+) \times p(X+)}{p(X Y+)}
\]
where
\[
p(X Y+)=p(X+)+p(Y+)-p(X+) p(Y+)
\]

In the case of the blocked cue, \(B\), we can assume that participants are already certain that \(A\) causes the outcome the first time they experience \(\mathrm{AB}+\) trials, i.e. \(\mathrm{p}(\mathrm{A}+)=1\). In the case of the control cue, D , there is equal uncertainty about it and cue \(C\), and thus \(p(C+)=p(D+)\). In the absence of any further information, these unconditional probabilities, as well as \(p(B+)\), are assumed to be equal to the base rate (the probability that the outcome will occur on any given trial or for any given cue). If we assume that, when a nonadditive outcome follows a compound of two cues, the outcome is independently caused by at least one of the cues, then \(\mathrm{p}(\mathrm{XY}+\mid \mathrm{X}+)=1\). This is a reasonable assumption unless it is explicitly shown to be false, as in the case of patterning discriminations (Harris \& Livesey, 2008; Livesey, Thorwart, \& Harris, 2011). The predicted blocking effect derived from these assumptions is a function of the base rate probability, as shown in Figure 1. As the base rate approaches zero, \(p(B+)\) approaches zero and \(p(D+)\) approaches 0.5 . As the base rate approaches \(1, \mathrm{p}(\mathrm{B}+)\) and \(\mathrm{p}(\mathrm{D}+)\) both approach 1 . Importantly, for every base rate between 0 and \(1, \mathrm{p}(\mathrm{B}+)\) is less than \(\mathrm{p}(\mathrm{D}+)\). Most causal learning experiments (including this study) present equal numbers of outcome and no outcome trial types, meaning that the base rate is around 0.5 . This means that a modest blocking effect is predicted, is can be seen in Figure 1.


Figure 1. Probability that a test cue (B or D) causes the outcome as a function of the baseline probability that any given cue causes the outcome. Values were calculated by applying Bayes' theorem to the propositions that can be derived from a typical blocking design involving the "blocked" cue B and the control cue D (see parentheses).

The solution can also be derived without the previous equations using a series of simple inferential steps. For ease
of illustration, let us also assume that the probability of any cue shown in the experiment causing the outcome is 0.5 . Given these assumptions, for any given compound of two cues A and B, there are four equally likely possibilities; i) A and \(B\) are both causal, ii) A only is causal, iii) B only is causal, or iv) neither A nor B is causal. In the case of the blocking cue, we know that A leads to the outcome, which allows us to rule out two of these possibilities (iii and iv), leaving possibility (i) in which B is causal, and possibility (ii) in which \(B\) is not causal. Thus the probability the \(B\) causes the outcome is 0.5 . In the case of the control cues, we observe only that the compound causes the outcome, which allows us to rule out only possibilities (iv) that neither cue causes the outcome. The remaining three possibilities are still equally likely, and D causes the outcome in two of these three possibilities. Thus the probability that D causes the outcome is 0.67 (likewise for C ).

\section*{Inferences and probabilistic reasoning}

Although it may seem surprising to some that blocking under these circumstances is completely logical, the temptation to conclude that blocking is the result of an explicit rational inference based on classical probability theory needs to be tempered by an equally striking observation. In a host of similar situations, most participants are very unlikely to apply this form of reasoning. The rationale applied above to blocking shares formal qualities with other problems involving conditional probabilities, which most normal adults find extremely difficult (e.g. BarHillel \& Falk, 1982). A prominent example is the MontyHall dilemma (see Burns \& Wieth, 2004), in which participants are so resistant to the solution derived from conditional probabilities that the problem is often referred to as a cognitive illusion. Thus, even though the blocking effect under nonadditive assumptions could be described as being rational, one should question whether participants are capable and inclined to explicitly use the inferential process that is necessary to arrive at the judgment in a rational and logical fashion.

If participants do use explicit reasoning processes akin to Bayesian inference, and the nonadditive blocking effect is a consequence of this reasoning, then the participants who show the greatest inclination to engage critically in inferential reasoning will be the most likely to give ratings in line with the blocking effect. Alternatively, Lovibond et al. (2003) assume that the most prevalent rational inference will be one in which the blocked and control cues are treated as being equally ambiguous, and thus no difference in their causal ratings should be observed. If this assumption is correct then those participants who are most likely to engage in that rational inference will be the least likely to produce a blocking effect in their judgments of causality. This hypothesis also implies that the blocking effect that has previously been observed after nonadditive pretraining is the result of a non-rational process such as a failure to retrieve the outcome associated with the blocked cue (Mitchell et al., 2006).

The current study sought to assess exactly what types of reasoned inference participants were inclined to use in this situation and how the inferential skills of individual participants were related to the blocking effect.

\section*{Blocking and critical thinking}

To test the relationship between inferential thinking and blocking, we coupled a typical blocking task with a test of cognitive reflection developed by Frederick (2005). The test presents three mathematical problems, each of which can be solved with minimal calculation. The problems were specifically designed to provoke an intuitive answer that is incorrect. Deriving the correct answer requires a modest amount of self-reflection and analytical thought in order to reject the first number that comes to mind and to then apply the inferences that are appropriate for the logic of the question at hand. Frederick's (2005) analysis of this cognitive reflection test (CRT) over multiple samples of young American adults revealed that a substantial proportion scored 0 out of 3 on the test, revealing a strong tendency to accept and report the intuitive foil answer for each question. CRT performance is associated with general cognitive ability (Frederick, 2005). However, some studies have shown that performance on the test is influenced by the conditions under which the information is presented; for instance when the questions are more difficult to read they are more likely to be answered correctly (Alter, Oppenheimer, Epley, \& Eyre, 2007). This suggests that participants' propensity to engage in critical reflection of the questions fluctuates and can be manipulated. For this study, the CRT was administered immediately after participants had finished making the causal judgments and thus, we assumed would assess their engagement in critical reflection around the time when the key measures of blocking were taken.

Participants were also given an additional problem designed to have similar logical properties to the contingencies in the blocking effect, in particular the presence of relevant conditional probabilities. Participants were instructed to "Imagine you are playing a game where, on every turn, a player tosses two normal everyday coins - a 50 -cent coin and a \(\$ 1\) coin - in the air. The coins are not biased: they are equally likely to show heads or tails. If either of the coins lands heads up, the player wins the round." They were then given two scenarios and asked to provide a probability for each:
1) "It is your turn next and you toss the coins. The \(\$ 1\) coin shows heads but the 50 -cent coin falls out of sight. What is the probability that the 50 -cent coin is showing heads?"
2) "Your turn to toss the coins comes around again. This time, when you toss the coins, both coins fall out of sight. The other players in the game say (honestly) that you have won but you cannot see the coins. What is the probability that the 50 -cent coin is showing heads?"

The answer to the first of these questions is relatively straightforward. Because the \(\$ 1\) coin lands heads, the fact that the participant has won has no bearing on the
probability that the 50 -cent coin is showing heads. Thus the correct answer is 0.5 . The answer to the second question is more difficult because the information indicating that the participant has won is important for the probability that either one of the coins has landed heads. The correct answer is 0.67 because two of the three equally probable circumstances that could lead to the participant winning involve the 50 -cent coin landing heads. We anticipated that most participants would say that the probability in this instance was also 0.5 . This result would be consistent with the logical inference that Lovibond et al. (2003) assume is most likely to occur in a blocking experiment with nonadditive outcomes.

Of most importance in this experiment was the relationship between CRT performance and blocking, and specifically whether blocking was found to be larger or smaller in those individuals that showed greater capacity for cognitive reflection. The coin-toss problem was added to further assess how participants engaged in inferences about similar uncertain events. If, as expected, many participants conclude that the uncertain events in each part of the cointoss problem are equally likely, then it shows a tendency to use the inferential reasoning described by Lovibond et al. (2003). On the other hand, if participants tend to give the correct answer then it suggests they are very capable of using conditional probabilities in this context and may do so to make explicit inferences in causal learning that would produce a blocking effect.

Table 1: Design of the current Experiment.
\begin{tabular}{|c|c|c|c|}
\hline Pretrain & Train 1 & Train 2 & Test \\
\hline & A+ & AB+ & B \\
\hline W- & & CD+ & C, D \\
\hline X+ & E+ & & \\
\hline Y+ & G- & F+ & E, F, EF \\
\hline Z- & GH- & GH- & EM, FM \\
\hline WZ- & IJ+ & KL- & H \\
\hline XY+ & L- & L- & L \\
\hline
\end{tabular}

Note: Letters A-M and W-Z denote randomly allocated foods used as predictive cues. These cues were followed by either no allergic reaction (-) or an allergic reaction (+). Trials above the dotted line in Train 1, Train 2 and Test comprise the blocking contingencies.

\section*{Method}

Participants. Forty-four introductory psychology students at the University of Sydney participated in the experiment in partial fulfillment of course requirements ( 32 female, mean age \(=18.9\) years).
Apparatus and Stimuli. Participants were tested in individual cubicles in a quiet laboratory. The causal learning experiment was programmed using the Psychophysics toolbox for Matlab and was presented using Apple Mac Mini computers attached to 17 inch displays. Experimental stimuli included images of a banana, apple, fish, lemon, cheese, milk, coffee, eggs, garlic, bread, pasta, peanuts,
avocado, meat, mushrooms, olive oil, strawberries, peas, and rice accompanied by written labels. The allocation of foods to cue (A, B, etc.) was randomized for each participant. The CRT and coin-toss problems were administered in paper and pencil format, with each test presented on a single side of A4 paper, printed clearly in 14 point Times New Roman font.
Procedure. Participants were asked to assume the role of a doctor whose task was to ascertain which foods were causing the allergic reactions of a fictitious patient, Mr X. Participants were given general instructions about the scenario and the procedure, as well as explicit instructions about the nonadditive nature of the outcome. The latter was reinforced by presenting a pretraining phase in which two cues ( X and Y ) had demonstrably nonadditive effects. Here trials with \(\mathrm{X}, \mathrm{Y}\) and the compound XY were presented, each with followed by an identical allergic reaction. The presentation of the reaction outcome was the same throughout the experiment and was always accompanied by a fictitious severity index showing the same level of severity for all allergic reactions.

For each of the Pretrain, Train 1 and Train 2 phases shown in Table 1, each of the trial types was presented 8 times in a randomized order. On each trial, either one or two foods were presented and participants predicted what outcome ("no allergic reaction" or "ALLERGIC REACTION") occurred by clicking either option. When an outcome was selected the options disappeared and were replaced with feedback about the actual outcome.

In the Test phase, participants were presented with a cue (or cues) and asked to make several judgments. First they were asked to judge "What is the probability that this food (these foods) will cause Mr X to have an allergic reaction?" and were required to make a rating on a linear analogue scale ranging from 0 to 1 with 0.1 increments marked along the scale. They were also asked to rate "How confident are you that your first rating is correct?" and "How severe will the reaction most likely be?" on additional linear analogue scales. The order of presentation of trials within the test phase was randomized, with each trial type presented only once. The critical cues in this phase for assessing blocking were cues B, C and D.

On completion of the allergist task, participants were given the CRT and conditional probability coin-toss problem in paper and pencil form. Participants were told to take as much time as they needed to finish these questions. Two versions of the coin-toss problem were used (counterbalanced between participants), one with the "neither coin visible" question first, the other with the " \(\$ 1\) coin visible" question first. Above the response line for each question, participants were reminded that "A. If EITHER of the coins shows heads, you win the round" and "B. You know that you have won this round."

\section*{Results}

Learning during the pretraining and training phases of the causal judgment task was generally very rapid. In the final
block of pretraining, phase 1 and phase 2 training, mean accuracy exceeded 0.95 for every cue-outcome contingency. All participants performed well above chance. Statistical analyses focused on the critical test data only. All analyses were performed with an alpha level of 0.05 .

Test Ratings. Of greatest importance was the probability rating for cue \(\mathrm{B}(\mathrm{M}=0.52)\) compared to the mean probability rating for C and \(\mathrm{D}(\mathrm{M}=0.65)\). The difference between these ratings was statistically significant, \(\mathrm{t}(43)=\) \(2.83, \mathrm{p}=.007\) ), indicating a reliable blocking effect overall.

CRT scores. Performance on the CRT was generally poor. Participants made on average just 0.70 correct responses out of a maximum of 3 . The vast majority ( \(84.1 \%\) ) of errors resulted from reporting of the intuitive foil answers to each item (for further details, see Frederick, 2005). Table 2 shows the number of participants who scored \(0-3\) on the CRT test, and the mean blocking score for participants with each score.

Table 2: Frequency of CRT scores and blocking score.
\begin{tabular}{|c|c|c|c|}
\hline CRT score & N & \multicolumn{2}{|c|}{Blocking} \\
\hline /3 & Participants & Mean & SEM \\
\hline 0 & 26 & . 207 & . 059 \\
\hline 1 & 7 & . 019 & . 115 \\
\hline 2 & 9 & . 005 & . 086 \\
\hline 3 & 2 & -. 001 & 0 \\
\hline total & 44 & . 13 & . 047 \\
\hline
\end{tabular}

Note: "Blocking" refers to the difference in probability rating given for the control cues \(\mathrm{C} / \mathrm{D}\) and the target cue B .

Of greatest interest was whether the number correct was related to blocking (as indicated by the difference in probability ratings for \(B\) and \(C / D\) ). The correlation between CRT score and blocking was negative and significant, \(\mathrm{r}=-\) \(0.304, \mathrm{p}=.045\). As can be seen in Table 2, this was mainly due to a large blocking effect in those that scored 0 on the CRT, with little variance in blocking scores amongst participants with CRT scores of 1 to 3 . Participants who scored 0 on the CRT showed significantly more blocking than those who scored more than \(0, \mathrm{t}(42)=2.27, \mathrm{p}=.028\). This is illustrated in Figure 2.

Coin-toss problem. Of the 44 participants, 29 responded 0.5 for the answer to both questions. Just two participants gave the correct responses, answering 0.67 for the scenario where neither coin is visible and 0.5 for the scenario where one coin is visibly showing heads (both scored 2 out of 3 on the CRT task and both exhibited a blocking effect). The remaining 13 participants did not systematically assign a higher probability to the "neither coin visible" scenario (M \(=.52)\) than to the "one coin visible" scenario \((\mathrm{M}=.54)\). As can be seen in Figure 3, participants who gave the same response to both questions (i.e. \(0.5 / 0.5\) ) produced equivalent blocking scores to those that produced different answers to the coin-toss problem, \(\mathrm{t}(42)=0.15, \mathrm{p}=.88\).


Figure 2. Mean causal probability judgments for the blocked cue B and control cues (mean of C and D), as a function of CRT performance. Left: Mean ratings for participants who failed to correctly answer a single question on the Cognitive Reflection Test. Right: Mean ratings for participants who scored at least 1 on the CRT. Error bars show SEM of the difference between B and \(\mathrm{C} / \mathrm{D}\) ratings.


Figure 3. Mean causal probability judgments for the blocked cue B and control cues C and D, as a function of answers to the coin-toss problem. Left: Mean ratings for participants who answered 0.5 for both items. Right: Mean ratings for participants who gave other answers (including two who gave the correct answers). Error bars show SEM of the difference between \(B\) and \(C / D\) ratings.

\section*{Discussion}

Overall, participants showed a modest but statistically reliable blocking effect. This observation is typical of many studies in causal learning, including several that involve non-additive pretraining to discourage participants from deducing that cue B is not causal (e.g. Lovibond et al., 2003; Mitchell et al., 2006).

More importantly, the size of the blocking effect was significantly related to participants' CRT performance. In particular, participants who scored zero on this test showed a substantial blocking effect whereas those that answered at least one of the three questions correctly gave comparable judgments for cue B and the control cues. The participants that scored zero on the CRT demonstrated the weakest ability to reflect critically on the questions in order to reject the most obvious answer and derive the correct one. These results are consistent with Lovibond et al.'s (2003) assertion that participants who reason carefully about the cues in a
blocking task involving nonadditive outcomes will judge the blocked and control cues to be equally likely to cause the outcome rather than adopting a Bayesian inference that appropriately accounts for conditional probabilities and is actually best aligned with the blocking effect itself.

Thinking about cause and effect under uncertainty is a difficult task and people do not readily adopt the approach typified by classical probability theory. The final probability question we used in this study is an example - with formal qualities similar to the blocking contingencies - where only two participants out of 44 gave the correct answer. Most (29 out of 44) gave the same answer, \(\mathrm{p}=0.5\), to both problems, suggesting that they assumed the status of the unseen coin was unaffected by information about the outcome (i.e. winning the round) in both of the examples. This logic is very similar to Lovibond et al.'s (2003) argument about reduced blocking with a nonadditive outcome. They argued that participants will conclude that no information is known about cue B and, likewise, no information is known about either of the cues C or D and, therefore, all three should be given the same rating. However, unlike the coin toss scenario, the conservative logic expressed in this inference was not as prevalent in the causal ratings for the cues \(B, C\) and D. Furthermore, participants who gave the \(0.5 / 0.5\) response to the coin-toss problem, and thus should not show blocking based on Lovibond et al.'s inference, were just as likely to show a blocking effect in their causal ratings as those who gave different answers to the coin-toss problem. These results suggest that, although the logic described by Lovibond et al. is prevalent in decisions involving uncertain causal relationships, the application of the inference is not necessarily consistent across different scenarios.

This result is correlational and should be interpreted cautiously. Blocking may arise from other forms of explicit inference, such as deductive reasoning, which is encouraged by additive outcome assumptions (Lovibond et al., 2003). Thus the key relationship observed in this study should only arise if the assumptions that participants bring into the experiment are tightly constrained to prevent deduction.

Conclusion. Although the blocking effect is arguably rational, even when assuming that the outcome is nonadditive, it is nonetheless associated with an uncritical mode of causal judgment. Only the minority of participants displaying some critical analytical ability on the CRT gave equivalent ratings to the blocked and control cues, consistent with the type of inferential reasoning outlined by Lovibond et al. (2003). The results are consistent with an account of causal learning that assumes that judgments are based on both explicit inferences and some form of associative learning or other automatic psychological operation that approximates Bayesian inference.

\section*{Acknowledgments}

This research was supported by a University of Sydney Bridging Support Grant to EJL.

\section*{References}

Alter, A. L., Oppenheimer, D. M., Epley, N., \& Eyre, R. N. (2007). Overcoming Intuition: Metacognitive Difficulty Activates Analytic Reasoning. Journal of Experimental Psychology: General, 136, 569-576.
Bar-Hillel, M., \& Falk, R. (1982). Some teasers concerning conditional probabilities. Cognition, 11, 109-122.
Burns, B.D. \& Wieth, M. (2004). The Collider Principle in Causal Reasoning: Why the Monty Hall Dilemma is so Hard. Journal of Experimental Psychology: General, 133, 434-449.
Cheng, P. W. (1997). From Covariation to Causation: A Causal Power Theory. Psychological Review, 104, 367405.

Dickinson,A., Shanks, D., \& Evenden, J. (1984). Judgement of act-outcome contingency: The role of selective attribution. Quarterly Journal of Experimental Psychology, 36A, 29-50.
De Houwer, J., Beckers, T., \& Glautier, S. (2002). Outcome and cue properties modulate blocking. Quarterly Journal of Experimental Psychology, 55A, 965-985.
Frederick, S. (2005). Cognitive Reflection and Decision Making. Journal of Economic Perspectives, 19, 25-42.
Griffiths, T. L., Sobel, D. M., Tenenbaum, J. B., Gopnik, A. (2011). Bayes and Blickets: Effects of Knowledge on Causal Induction in Children and Adults. Cognitive Science, 35, 1407-1455.
Harris, J.A. \& Livesey, E.J. (2008). Comparing Patterning and Biconditional Discriminations in Humans. Journal of Experimental Psychology: Animal Behavior Processes, 34, 144-154.
Livesey, E. J. \& Boakes, R. A. (2004). Outcome additivity, elemental processing and blocking in human causality judgements. Quarterly Journal of Experimental Psychology, 57B, 361-379.
Livesey, E. J., Thorwart, A., \& Harris, J. A. (2011). Comparing positive and negative patterning in human learning. Quarterly Journal of Experimental Psychology, 64, 2316-2333.
Lovibond P. F., Been S.-L., Mitchell C. J., Bouton M. E., Frohardt R. (2003). Forward and backward blocking of causal judgment is enhanced by additivity of effect magnitude. Memory \& Cognition 31, 133-42.
Mitchell C. J., De Houwer J., Lovibond P. F. (2009). The propositional nature of human associative learning. Behavioral and Brain Science, 32, 183-246.
Mitchell C.J., Lovibond P.F., Minard E, Lavis Y. (2006). Forward blocking in human learning sometimes reflects the failure to encode a cue-outcome relationship. Quarterly Journal of Experimental Psychology 59, 830844.

Sternberg, D. A. (2012). Two Mechanisms of Human Contingency Learning. Psychological Science, 23, 59-68.
Waldmann, M. R. (2000). Competition Among Causes But Not Effects in Predictive and Diagnostic Learning. Journal of Experimental Psychology: Learning, Memory and Cognition, 26, 53-76.

\title{
Modeling Continuous Representations in Visual Working Memory
}

\author{
Johannes Lohmann (johannes.lohmann@uni-tuebingen.de) \\ Martin V. Butz (martin.butz@uni-tuebingen.de) \\ Cognitive Modeling, Department of Computer Science, Department of Psychology, Sand 14, Tübingen, 72076, Germany
}

\begin{abstract}
Visual working memory (VWM) is a crucial part of our cognitive system. Currently there is an active debate how the apparent limitations of VWM should be described. Limited-slot and flexible-resource theories are discussed, but so far the temporal dynamics of representations stored in VWM are not fully understood. In this paper we present data that supports the notion of dynamic VWM contents with changing precision. To account for these observations in a qualitative way, we propose a neural network that is able to account for emerging capacity limits as well as for changes in the precision of stored information.
\end{abstract}

Keywords: VWM; Visual Attention; Neural Network

\section*{Introduction}

Successful interaction with our environment requires the storage and maintenance of task-relevant information. Planning of behavior as well as evaluation of the outcome would not be possible without this ability. The fact that we are able to perform these tasks indicates that there is some kind of interface between our environment and cognition. This interface is provided by the working memory system. Due to its importance in linking cognition to the external world, working memory has been studied extensively for decades.

An apparent limitation of working memory and especially visual working memory (VWM) with respect to the amount of preserved information has been observed throughout the years (see Miller, 1956 for an early review). Until today it is still in discussion how these limitations might be characterized. While limited-slot theories (Zhang \& Luck, 2008) state that the maximum number of stored representations is limited and cannot be increased by decreasing the precision of individual representations, flexible-resource theories (Bays \& Hussain, 2008) assume a mnemonic resource that can be used to either store a large number of low-precision representations or a small number of high-precision representations. The differentiation between both theories is fundamental as the alternatives suggest rather different bases of cognition.

There is empirical evidence for both alternatives (see Fukuda, Awh, \& Vogel, 2010 for a recent review), even if resource models seem more plausible from a modeling perspective (Berg, Shin, Chou, George, \& Ma, 2012). Recent studies investigated the neural basis for capacity limits, but again the results are mixed. While Anderson, Vogel, and Awh (2011) found behavioral as well as electrophysiological evidence for a limited number of slots in humans, Buschman, Siegel, Roy, and Miller (2011) reported mixed evidence in rhesus monkeys. The results imply the existence of discrete slots, containing some kind of resource that determines the precision of the stored representations.

To sum up, it remains unclear if capacity limits of VWM can be described in terms of slots or resources, but the empirical evidence indicates that it is unlikely that a strict resource or slot model will be the consensus - some kind of mixed model seems most probable. Interestingly, in contrast to the limitations with respect to the number of stored items, the temporal dynamics of VWM contents are far less investigated. Also in this respect the results are inconclusive. Zhang and Luck (2009) reported clear evidence for abrupt loss of information, whereas Salmela, Lahde, and Saarinen (2012) found evidence for the gradual loss of information. The results of Zhang and Luck (2009) are more in line with a slot model. The gradual decay observed by Salmela et al. (2012) is more in line with a resource model, in which the precision of the representations degrades over time. To investigate the temporal dynamics of information maintained in VWM we applied the same experimental paradigm as Zhang and Luck (2009). Furthermore, we developed a neural network model that can qualitatively account for the observations.

In the next section we describe the experimental setup. Next we report the obtained results. After this we give an outline of the neural model. A short discussion concludes the paper.

\section*{Experimental Setup}

Since the influential study of Luck and Vogel (1997), change detection paradigms have become the standard approach in VWM research. As it was highlighted by Brady, Konkle, and Alvarez (2011), however, change detection paradigms do not allow to assess the precision of the representation that underlies the response of the participants. Zhang and Luck (2009) proposed a paradigm that allows to obtain a measure for the precision of the stored representations.

Participants had to remember Fourier descriptors which varied continuously in their phase (see Fig. 1 for the trial sequence and the stimuli). After the presentation of the stimuli a response screen appeared after a variable interstimulus interval (ISI). The position of one of the presented shapes was highlighted and participants had to indicate the presented stimuli on a "shape-wheel" containing the whole shape dimension. If the critical shape was in working memory, participants should report a shape close to the original shape, the response distribution should be bell-shaped and its deviation would be a measure of the precision of the representation. Without a representation participants should guess. Guessing should follow a uniform distribution. Together these processes result in a mixed distribution, consisting of a bellshaped distribution reflecting the precision (referred to as \(\sigma\)


Figure 1: Sequence of a single trial. After the initial presentation of the fixation cross, three, four, or five Fourier descriptors were displayed on a invisible circle for 200 ms . After this a variable interstimulus interval (ISI) of 500, 2500, or 5000 ms followed. At the end of the trial a "shape-wheel" appeared. One of the stimuli positions was highlighted and participants had to indicate the identity of the descriptor on the wheel.


Figure 2: Normalized histogram of exemplary artificially generated data and fitted von Mises distribution.
in the following) and a vertical offset, indicating the probability that participants had no representation of the critical shape (referred to as \(\mathrm{p}_{\mathrm{F}}\) in the following).

We were interested in the effects of different ISIs on the precision and the amount of stored information, i.e. the variation of \(\sigma\) and \(\mathrm{p}_{\mathrm{F}}\). A slot-model would predict a constant \(\sigma\) over time and furthermore that \(\sigma\) would not be affected by the number of stimuli. Once the available slots are filled \(\mathrm{p}_{\mathrm{F}}\) should strongly increase. A resource model would predict an increasing \(\sigma\) with increasing numbers of items, while \(\mathrm{p}_{\mathrm{F}}\) should not change.

\section*{Participants}

16 healthy students ( 11 males) of the cognitive sciences participated in our experiment, their age ranged between 22 and 33 years (mean age 23.125). All participants reported normal or corrected-to-normal vision and received course credit in exchange for their participation. All participants provided informed consent.

\section*{Apparatus}

Stimulus presentation and collection of responses were performed by an IBM-compatible computer with a \(22-\mathrm{in}\). dis-
play. The stimuli were displayed at a resolution of 1680 by 1050 pixels. The experiment was implemented in C++.

\section*{Stimuli}

The stimulus set consisted of three, four, or five Fourier descriptors. The phases were chosen pseudo-randomly. The minimal phase difference was \(30^{\circ}\). Each descriptor subtended \(1.6^{\circ} \times 1.6^{\circ}\) degrees of visual angle (viewing distance \(\approx 70\) \(\mathrm{cm})\). The descriptors were randomly arranged on an invisible circle with a radius of \(4.7^{\circ}\) of visual angle. This arrangement was restricted to six possible locations, each spaced by \(60^{\circ}\). To obtain the responses a wheel containing 30 shapes evenly distributed over the phase space was presented as well as a cue indicating the critical descriptor (see Fig. 1). The shape wheel was centered on the screen with a radius of \(10.2^{\circ}\) of visual angle.

\section*{Procedure}

Each trial started with a fixation cross which lasted for 2000 ms . After this three to five Fourier descriptors were presented for 1000 ms . This was followed by a blank screen which lasted for 1000,2500 , or 5000 ms . Then the response screen was presented and one of the locations was highlighted. The participants responded by clicking on the "shape-wheel". We collected the absolute angle as well as the angular distance to the cued descriptor. Trials lasted until the participants confirmed their response by pressing a key. All \(3 \times 3\) combinations of number of stimuli and ISI were repeated 30 times, totaling 270 trials. The whole procedure took about 60 min utes.

\section*{Results}

We applied a quantitative model (Zhang \& Luck, 2008) to the data to estimate the probability that a cued descriptor was present in memory \(\left(1-\mathrm{p}_{\mathrm{F}}\right)\) as well as the precision \((\sigma)\) of the representation. We first describe this model, then we report the obtained results.

\section*{Data Analysis}

According to the model participants have a noisy representation of the crucial descriptor in some trials, in the remaining trials participants are assumed to guess. The noisy representation can be described in terms of a von Mises distribution \({ }^{1}\), whereas guessing is modeled as a uniform random process. The resulting mixed distribution is displayed in Fig. 2 and can be described as
\[
\begin{equation*}
p\left(x \mid \mu, \kappa, \mathrm{p}_{\mathrm{F}}\right)=\left(1-\mathrm{p}_{\mathrm{F}}\right) \frac{e^{\kappa \cos (x-\mu)}}{2 \pi I_{0}(\kappa)}+\frac{1}{2 \pi} \mathrm{p}_{\mathrm{F}}, \tag{1}
\end{equation*}
\]
where \(\mu\) is the mean of the von Mises distribution, \(\kappa\) denotes the density of the distribution (this can be considered as the inverse of the deviation), \(I_{0}\) is the modified Bessel function of 0 th order, and \(\mathrm{p}_{\mathrm{F}}\) is the guessing probability.

\footnotetext{
\({ }^{1}\) Due to the circular nature of the response dimension a Gaussian distribution is not feasible.
}

Later on we will report the deviation of the distribution instead of \(\kappa\), which can be obtained by
\[
\begin{equation*}
\sigma=\sqrt{1-\frac{I_{1}(\kappa)}{I_{0}(\kappa)}} \tag{2}
\end{equation*}
\]
where \(I_{1}\) is the modified Bessel function of 1 st order.
Since the phase of the relevant descriptor varied from trial to trial, the data analysis was based on the angular distance between the response and the phase of the descriptor. Hence the mean of the von Mises distributions should equal 0 . As noted above, \(\sigma\) reflects the precision of the representation, larger values of \(\sigma\) indicate a lower quality of the representation. The probability \(\mathrm{p}_{\mathrm{F}}\) is an indicator for the amount of preserved information. Higher values of \(\mathrm{p}_{\mathrm{F}}\) indicate less information to be preserved.

The parameters cannot be directly inferred from the observed data. Therefore we fitted the mixed distribution via maximum likelihood estimation.

\section*{Estimated Parameters}

We concentrated on the estimated values of \(\mathrm{p}_{\mathrm{F}}\) and \(\sigma\) of the mixed model. We estimated \(\mu, \kappa\), and \(\mathrm{p}_{\mathrm{F}}\) separately for each participant and each level of the varied factors. The results have to be treated with caution, since the data basis for these fits was quite small, consisting of only 90 samples per fit. For three participants the likelihood values remained comparatively small, indicating that the applied model was not well suited for their data. Hence, the respective data sets were not entered in the analysis.

Fig. 3 shows the obtained results. The estimates with respect to the number of stimuli are plotted on the left panel, whereas the estimates for the different ISIs are plotted on the right panel. Error bars indicate the standard error of the mean.

For a quantitative analysis of the differences of the estimates we performed paired t-Tests. Significant differences on a \(5 \%\) level are indicated by an asterisk. With respect to the number of stimuli, \(\sigma\) increased significantly. For \(\mathrm{p}_{\mathrm{F}}\) the estimate was significantly higher in case of four items compared to three items. With respect to the ISI, significant differences were only observed for \(\sigma\). None of the six \(\mu\) parameters differed significantly from zero.

The precision of preserved information decreases with the number of items. The amount of preserved information seems less effected, since there was no significant difference between the \(\mathrm{p}_{\mathrm{F}}\) estimates for three and five items. The precision also seems to change over time, whereas the amount of stored information remains constant. Compared to the effect of the number of stimuli the effect of the ISI is much weaker. The fact that we observed significant changes in the precision of stored information is in conflict with the results reported by Zhang and Luck (2009), but fits the observations of Salmela et al. (2012).


Figure 4: Overview of the model architecture. Input units accumulate spikes from sensory neurons. Memory units receive activation from the input units. Lateral inhibition as well as self-recurrent excitation determine the overall activations of the memory units, which is proportional to the precision of the respective representation. See text for details about the parameters.

\section*{Neural Model}

So far there are only few models that can account for the temporal dynamics of VWM contents. One of these models is the time-based resource-sharing (TBRS) model (Barrouillet, Bernardin, \& Camos, 2004, see Oberauer \& Lewandowsky, 2011 for an implementation). TBRS assumes an interplay between temporal decay and a refreshment process to account for dynamic changes in the quality of VWM representations. The encoding stage is neglected however. Another, more neural model is the dynamic field theory (DFT, Johnson, Spencer, \& Schöner, 2009), where the dynamic interactions between excitatory and inhibitory layers of neurons are applied to model dynamic changes in VWM content. On the one hand, DFT has a lot of desirable features, for instance capacity limits emerge naturally from the model properties. On the other hand, DFT has a lot of degrees of freedom, rendering direct fitting to observed data rather difficult. Furthermore, the encoding stage is not specified.

We propose a model that can account for encoding of stimuli, as well as for the maintenance of stored representations. We want to achieve this with less degrees of freedom than DFT but still with a neural model. The proposed model is a combination of the theory of visual attention (TVA, Bundesen, 1990) and the short term memory network proposed by Usher and Cohen (1999). Since this network models discrete states, we extended it with the single trace fragility theory (Wickelgren, 1974) to model continuous changes in


Figure 3: Parameter estimates with respect to the number of stimuli (left panel) and ISI (right panel). Significant differences in the estimates are indicated with an asterisk. Deviation \((\sigma)\) increases with number of stimuli as well as with ISI. For the failure probability \(\mathrm{p}_{\mathrm{F}}\) effects were observed for the number of stimuli only.
the precision of stored representations. The input layer of the model accumulates spikes from sensory neurons. The spiking rate of the sensory neurons is based on TVA. Activation from the input layer is forwarded to memory units. The overall activations of the memory units depend on lateral inhibtion and self-recurrent excitation. The activity of the memory units varies continuously and is proportional to the precision of the respective representation. Temporal decay is modeled via the leakage of input activation in the absence of sensory input. Initially, the binding between memory units and the input layer is fragile. The longer a representation resides in VWM, the stronger the binding and the weaker the leakage. Fig. 4 gives an overview of the model architecture. The different components of the model are described in the next paragraphs. After this we give a short example of the performance of the model.

\section*{Encoding of Stimuli}

The encoding stage is modeled via TVA. TVA is a quantitative model of visual encoding that is well suited to account for the selection and categorization of visual stimuli. TVA models visual attention by integrating bottom-up as well as top-down processes.

TVA assumes a competition between different categorizations for incorporation in VWM. This competition can be quantified via a race model, where the rate parameter \(\mathrm{v}(x, i)\) determines the time needed for a categorization of the type "item \(x\) belongs to category \(i\) " to be finished. The rate parameter depends on the task as well as sensory parameters (see Bundesen, 1990 for details).

We assume a fixed number of sensory neurons that spike during the presentation of stimuli with a rate equal to \(v(x, i)\) (cf. Bundesen, Habekost, \& Kyllingsbæk, 2005). Spikes are accumulated in a separate input layer, which forwards activation to the memory layer that is described in the next paragraph.

\section*{Maintenance of Representations}

We model VWM in terms of a dynamic neural network (Usher \& Cohen, 1999). The activation of each unit is affected by three processes. First, activation is stabilized by self-recurrent excitatory feedback. Second, each unit is inhibited by its neighbors. Third, a unit might receive sensory input that increases its activation (see Fig. 4). We modeled this input in terms of accumulated spikes emitted from the sensory neurons described in the previous paragraph. The dynamics of a single unit in the memory layer can be described via the following differential equation:
\[
\begin{equation*}
\frac{d \kappa_{i}}{d t}=-\kappa_{i}+\alpha F\left(\kappa_{i}\right)-\beta \sum_{j \neq i}^{N} F\left(\kappa_{j}\right)+I_{i}+\text { noise } \tag{3}
\end{equation*}
\]
where \(\kappa_{i}\) indicates the activation of unit \(i, \alpha\) is the selfrecurrent excitatory weight, \(\beta\) is the weight of the lateral inhibition, \(I_{i}\) denotes the current sensory input supporting unit \(i\), noise indicates a uniform noise term, and \(F(\kappa)\) is the activation-function (in this case a linear one).

Without sensory input Eq. 3 can be used to model decay of activations to a baseline, given the noise is small enough to prevent random fluctuations. The system has a lot of interesting emergent properties despite its simple structure. It is possible to model serial position effects, capacity limits (with proper choices for \(\alpha\) and \(\beta\) ), and the development of stable states.

We assume the activation of the memory units to be proportional to the precision of the according representations, which is quantified by the \(\kappa\) parameter of the von Mises distribution in the mixed model. We assume one memory unit per stimulus. If there are more stimuli, lateral inhibition is increased, resulting in an overall reduced activity. The reduced precision observed for higher numbers of stimuli (see Fig. 3, lower panel) emerges naturally. In its original formulation the described network can be used to model discrete states of mem-
ory units, either the activation is above the baseline or not, the transitions are non-linear. Since our data indicates continuous decay of the precision over time, we extended the original model with the single trace fragility theory (Wickelgren, 1974). This extension is described in the next paragraph.

\section*{Temporal Decay}

As it can be inferred from Eq. 3, the activation of a memory unit depends on the sensory input. Since Buschman et al. (2011) found that neural activation first decays in early areas, we modeled continuous changes in the precision by a decay of the accumulated sensory input \({ }^{2}\). Since older memory traces are more resistant than younger ones (Jost's second law), we assumed this decay to slow down over time.

The single trace fragility theory provides a formal framework for these assumptions, by introducing the concept of fragility, which quantifies the susceptibility of stored information for temporal decay. In our model, fragility refers to the binding between input and memory units (see Fig. 4), which modulates the leakage of the input units. Applied to the input term in Eq. 3, decay after display offset can be described by:
\[
\begin{equation*}
\frac{d I_{i}}{d t}=-\xi f I_{i} \tag{4}
\end{equation*}
\]
where \(\xi\) is the decay rate and \(f\) denotes the fragility, which is reduced over time:
\[
\begin{equation*}
\frac{d f}{d t}=-\psi f^{2} \tag{5}
\end{equation*}
\]
here \(\psi\) denotes the decay rate of the fragility. We assume that fragility starts to decline when a categorization is encoded, hence early categorizations are less prone to temporal decay. Furthermore, we assume that the temporal decay starts after display offset, when sensory information is no longer available to maintain the activity of the memory units. This mechanism concludes the model specification.

\section*{Modeling Continuous Representations}

To illustrate the functionality of the system, let us assume a simple display containing three stimuli, lets further assume that only one feature dimension of these stimuli is task relevant (e.g. shape) and that the stimuli do not differ with respect to visibility (see Fig. 4, center). In this case the spiking rate of all sensory neurons is equal during the presentation of the stimuli. This process is displayed in the upper panel of Fig. 5. Encoding is a random process, hence the encoding times \(\left(t_{0}\right)\) differ. After the offset of the display \((t>D)\) the decay mechanism described in Eq. 4 begins to operate. Over time the decay attenuates, since the fragility is reduced (see Eq. 5).

Lateral inhibition between stored representations is visible around time-step 30; the encoding of an additional representation reduces the activation of the previously stored ones.

\footnotetext{
\({ }^{2}\) Please note that Buschman et al. (2011) interpreted this finding in terms of an encoding failure instead of a temporal decay.
}


Figure 5: Example of the dynamics produced by the proposed model. Dynamics of the input units are displayed in the upper panel, the overall activation of the memory units is displayed in the lower panel. Parameter D in the upper panel refers to the presentation duration, while \(t_{0}\) denotes the encoding time of the according representation. Activation of the memory units is assumed to be proportional to the precision of the representation, the decreasing activation reflects the increasing deviation in Fig. 3 (right panel).

After the offset of the stimuli (at time-step 50, vertical dashed line), decay starts. For highly active representations a nearly exponential decay occurs, which attenuates over time, whereas the decay of less active, or older representations can be better described by a power function. As can be concluded from the Fig. 5, it is possible that representations are completely lost (i.e. the according activation falls below zero). The resting level of the system varies, depending on the strength of the lateral inhibition, that is the number of simultaneously active units.

The model so far specifies activations over time; to fit the model to the data in the experiments, we need to convert these activations into probability distributions for memory recollection. We assume the activation of a memory unit to be proportional to the precision of the according representation. Hence, the activations are a measure of the \(\kappa\) parameter of the von Mises distribution. The overall response distribution can be modeled as a mixed distribution, with \(\kappa\) 's equal to the activations multiplied with a constant scaling factor (values below zero are considered to be zero). Since a von Mises distribution with \(\kappa=0\) becomes a uniform distribution, \(\mathrm{p}_{\mathrm{F}}\) can be modeled as well.

Our model is able to account for the encoding of visual stimuli, as well as for changes in the precision of the stored representations. The observed reduction in precision (see

Fig. 3, left panel) due to more stimuli can be accounted for by TVA and lateral inhibition. The change in precision over time is captured by the network dynamics described in Eq. 3 and the temporal decay described in Eq. 4. The apparently high number of parameters is greatly reduced since the TVA parameter \(v\) can be set to one (like in the example). This leaves \(\alpha, \beta, \xi, \psi\), and the number of sensory neurons per input unit as free parameters. The different components of the model are modular, for instance it is quite simple to model the memory dynamics without Eq. 4. To evaluate the relevance of the different components, we will perform Bayesian model comparisons, once a broader data basis has been acquired.

\section*{Conclusions}

In this paper we presented data on the temporal changes in the precision of VWM contents. The results indicate that the precision of the representations is strongly affected by the number of stimuli. Precision decreased over time, whereas the amount of preserved information was rather stable (no significant increase of \(\mathrm{p}_{\mathrm{F}}\) ). We proposed a neural network model, which models the encoding as well as the maintenance of information. So far the key findings can be replicated qualitatively, since the overall precision decreases with increasing memory loads as well as over time.

The different approaches to account for VWM properties reflect different paradigms of understanding cognition. While limited-slot models with static precision adhere to the computer metaphor of cognition, resource models are more in line with the dynamic system approach to cognition. The fact that VWM capacity is limited is not questioned, however. Given the findings of Buschman et al. (2011), who observed discrete slots with flexible resources per hemisphere and a possible gating mechanism in the lateral intraparietal cortex, our model offers an implementation of this neural mechanism. As it was noted above, capacity limits emerge naturally in the network due to the interplay of excitation and lateral inhibition, while the individual activations implement a continuous process, modeling the precision of individual representations over time. Thus, the approach integrates slot and resource model perspectives. Further evaluation is necessary to assess the quantitative fit of the model. This requires a broader data basis, involving larger numbers of stimuli to assess the ability of our model to account for higher memory loads, which is usually better accounted for by slot models (Rouder, Morey, Morey, \& Cowan, 2011).

\section*{References}

Anderson, D., Vogel, E. K., \& Awh, E. (2011). Precision in visual working memory reaches a stable plateau when individual item limits are exceeded. The Journal of Neuroscience, 31, 1128-1138.
Barrouillet, P., Bernardin, S., \& Camos, V. (2004). Time constraints and resource sharing in adults' working memory spans. Journal of Experimental Psychology: General, 133, 83-100.

Bays, P. M., \& Hussain, M. (2008). Dynamic shifts of limited working memory resources in human vision. Science, 321, 851-854.
Berg, R. van den, Shin, H., Chou, W., George, R., \& Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. PNAS, 109, 87808785.

Brady, T., Konkle, T., \& Alvarez, G. (2011). A review of visual memory capacity: Beyond individual items and toward structured representations. Journal of Vision, 11, 1-34.
Bundesen, C. (1990). A theory of visual attention. Psychological Review, 97, 523-547.
Bundesen, C., Habekost, T., \& Kyllingsbæk, S. (2005). A neural theory of visual attention: Bridging cognition and neurophysiology. Psychological Review, 112, 291-328.
Buschman, T., Siegel, M., Roy, J., \& Miller, E. (2011). Neural substrates of cognitive capacity limitations. PNAS, 108, 11252-11255.
Fukuda, K., Awh, E., \& Vogel, E. K. (2010). Discrete capacity limits in visual working memory. Current Opinion in Neurobiology, 20, 177-182.
Johnson, J. S., Spencer, J. P., \& Schöner, G. (2009). A layered neural architecture for the consolidation, maintenance, and updating of representations in visual working memory. Brain Research, 1299, 17-32.
Luck, S. J., \& Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. Nature, 390, 279-281.
Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review, 63, 81-97.
Oberauer, K., \& Lewandowsky, S. (2011). Modeling working memory: a computational implementation of the TimeBased Resource-Sharing theory. Psychonomic Bulletin \& Review, 18, 10-45.
Rouder, J. N., Morey, R. D., Morey, C. C., \& Cowan, N. (2011). How to measure working memory capacity in the change detection paradigm. Psychonomic Bulletin \& Review, 18(2), 324-330.
Salmela, V., Lahde, M., \& Saarinen, J. (2012). Visual working memory for amplitude-modulated shapes. Journal of Vision, 12, 1-9.
Usher, M., \& Cohen, J. D. (1999). Short term memory and selection processes in a frontal-lobe model. In D. Heinke, G. W. Humphries, \& A. Olsen (Eds.), Connectionist models in cognitive neuroscience (pp. 78-91). London: SpringerVerlag.
Wickelgren, W. A. (1974). Single-trace fragility theory of memory dynamics. Memory \& Cognition, 2, 775-780.
Zhang, W., \& Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. Nature, 453, 233-235.
Zhang, W., \& Luck, S. J. (2009). Sudden death and gradual decay in visual working memory. Psychological Science, 20, 423-428.

\title{
Studying sign processes in the emergence of communication
}

\author{
Angelo Loula (angelocl@ecomp.uefs.br) \\ Cognitive and Intelligent Systems Lab (LASIC), State University of Feira de Santana (UEFS) Feira de Santana, BA, Brazil \\ Ricardo Gudwin (gudwin@dca.fee.unicamp.br) \\ Department of Computer Engineering and Industrial Automation, School of Electrical and Computer Engineering, University of Campinas (UNICAMP), Campinas, SP, Brazil \\ João Queiroz (queirozj@pq.cnpq.br) \\ Institute of Arts and Design, Federal University of Juiz de Fora (UFJF) Juiz de Fora, MG, Brazil
}

\begin{abstract}
Communication depends on the production and interpretation of representations, but the study of representational processes underlying communication finds little discussion in computational experiments. Here we present an experiment on the emergence of both interpretation and production of multiple representations, with multiple referents, where referential processes can be tracked. Results show the dynamics of semiotic processes during the evolution of artificial creatures and the emergence of a variety of semiotic processes, such as sign production, sign interpretation, and sign-object-interpretant relations.
\end{abstract}

Keywords: Sign; Communication; Semiotics; Artificial Intelligence; Computer Simulation.

\section*{Introduction}

Computational modeling and simulation of the emergence of semiotic processes, such as language and communication, ,has been consolidating as an important methodology and has a growing community of researchers involves in scientific and technological issues related with such phenomena (Wagner et al., 2003; Nolfi \& Mirolli, 2010; Noble et al., 2010). This computational approach have been dedicated to building environments and creatures through which it is possible to simulate the minimum conditions to observe the evolution and emergence of semiotic behaviors. As the main form of interaction between agents, in most of these synthetic experiments, communication has, particularly, been a significant research subject. Primarily, it depends on the production of representations (by an utterer) and the interpretation of them (by an interpreter). Despite the fact that representation processes are in the foundations of communication, little discussion about such processes can be found, such as, the emergence of fundamental types of representations and their referential relations.

We have previously simulated the emergence of interpretation of two different types of representations (symbols and indexes) in communicative interactions (Loula et al., 2010), and have also studied further the cognitive conditions to the emergence of such interpretation processes (Loula et al, 2011). However, the experiments previously done focused only on the emergence of interpretation, with fixed production of a single representation with only one referent. Here we propose to evaluate representation
processes in the emergence of both interpretation and production of multiple representations, with multiple referents. To do so, we use a similar scenario of resource collecting, but apply a neural network model as the cognitive architecture for creatures, that can become utterers and interpreters. The experiment involves empirical constraints from studies of animal communication and also theoretical constraints from Peircean pragmatic theory of signs.

In the next section, we review related work on simulation of the emergence of communication. Next, we briefly describe the theoretical principles and biological motivations. We then describe our experiment on the emergence of production and interpretational processes in communication events. Finally, we outline our results and conclusions and point out perspectives on the study of the emergence of sign processes types.

\section*{Related work}

The emergence of communication in computer simulations is the topic of many works, but discussions on the underlying representational processes find little space in such literature. We review two important ones to reveal this.

Floreano et al (2007) studied the evolutionary conditions that could allow the emergence of a reliable communication system, following biological motivations of animal communication. Groups of robots controlled by neural networks were evolved to adapt to a foraging task. Robots could use a ring of blue light as a signal, as the authors call, to each other. Even though the ring of light was used to cooperate, nothing was discussed on what it represented to the robots or what type of representation could be involved.

In an experiment with artificial creatures in a grid world, Cangelosi (2001) simulated the emergence of communication systems to name edible and poisonous mushrooms. He proposed the emergence of different modalities of representations in this experiment on the evolution of communication. In typifying communication systems, Cangelosi (2001) distinguished between signals, which have direct relation with world entities, and symbols, which in addition are related to other symbols. The simulated creatures were controlled by a neural network which were both evolved and trained in various tasks, and,
at the end, a shared communication system emerged, involving signals and symbols, according to Cangelosi. But he did not described how these signals and symbols were interpreted by the creatures, i.e. if a heard signal was first mapped to a mushroom as its referent, and then to an action, or if it was mapped to an action, with a referent being associated with it.

Other works have also studied the emergence of communication in artificial agents (see Nolfi and Mirolli, 2010, Wagner et al. 2003). Nevertheless, we have not found works that have studied the emergence of different types of interpretations processes and referential processes.

\section*{Theoretical and Empirical Constraints}

Synthetic experiments are heavily influenced by theoretical principles and biological motivations, and such background should be an essential part of any synthetic experiment (Parisi, 2001). To model the emergence of communication processes based on different types of representation, it is certainly important to look at theoretical models and principles, and also look for biological motivations, and avoid arbitrary or naïve assumptions about the underlying processes.

In computational simulation works dealing with the emergence of communication, there is always something that is communicated from an agent to another one, and that is given various names: signal, symbol, sound, word, expression, or utterance. In most of these works, that which is communicated also seems to have representation capabilities. We have used the term representation in the first section, to emphasize this and also to apply a more familiar word for the artificial intelligence community. Nevertheless, we will now use the expression 'sign', as a technical term in a theoretical background.

A sign is defined, following Peirce (1958), as something that refers to something else, an object (which the sign represents in some aspect) and produces an effect (interpretant) in the interpreter. A sign is also defined by Peirce in relation to communication, as something that mediates between an utterer and an interpreter (Peirce, 1967, MS318), with the sign originating in the first and determining its interpretant in the last (Peirce, 1967, MS11). It is important to notice that a sign is only regarded as a sign if and when it is interpreted, so a sign communicated by an utterer is only a sign to the interpreter, but not to the utterer himself, unless it interprets the produced sign himself.

Sign processes show a remarkable variety. A basic typology (and the most fundamental one), proposed by Peirce (1958), differentiates between iconic, indexical, and symbolic processes. Icons are signs that stand for their objects by a similarity or resemblance, no matter if they show any spatio-temporal physical correlation with an existent object. In this case, a sign refers to an object in virtue of a certain quality which is shared between them. Indexes are signs which refer to their objects due to a direct physical connection between them. Since (in this case) the sign should be determined by the object (e.g. by means of a
causal relationship) both must exist as actual events. Spatiotemporal co-variation is the most characteristic property of indexical processes. Symbols are signs that are related to their object through a determinative relation of law, rule or convention. A symbol becomes a sign of some object merely or mainly by the fact that it is used and understood as such by the interpreter, who establishes this connection.

Communication is a process that occurs among natural systems and therefore we followed biological motivations on building our synthetic experiment. Animals communicate in various situations, from courtship and dominance to predator warning and food calls. Following Peirce's definition of sign classes, many animals can actually be capable of communicating by means of diverse types of signs (Ribeiro et al., 2007).

To further explore the mechanisms behind communication, a minimum brain model can be useful to understand what cognitive resources might be available and process underlining certain behaviors. Queiroz and Ribeiro (2002) described a minimum vertebrate brain for vervet monkeys' predator warning vocalization behavior. It was modeled as being composed by three major representational domains: the sensory, the associative and the motor domains. Different first-order sensory representational domains (RD1s) receive unimodal stimuli, which are then associated in a second-order multi-modal representation domain (RD2) so as to elicit symbolic responses to alarmcalls by a first-order motor representation domain (RD1m).

In order to model the emergence of indexical and symbolic interpretation competences, we must specify the requirements for each and how to recognize them. Indexical interpretation is a direct interpretation of signs, such that the interpreter is guided by the sign to recognize its object as something spatio-temporally connected to it. In the minimum brain model, this corresponds to an individual capable of connecting RD1s to RD1m without the need for RD2. But a symbolic interpretation undergoes the mediation of the RD2 to connect the sign to its object, in such a way that a habit (either inborn or acquired) must be present to establish this association. Thus, in symbolic interpretation,


Figure 1: Cognitive architectures for representations interpretations. Left: Type 1 architecture - RD1s are connected directly to RD1m. Right: Type 2 architecture data from visual RD1s and auditory RD1s can be associated in RD2 before connecting to RD1m.

RD2 is needed once it is the only domain able to establish connections between different representation domains. To evaluate what conditions might elicit each response type indexical or symbolic -, we implemented these two possible cognitive processing paths as mutually exclusive paths: either the creature responds to auditory events indexically, with direct motor actions (Type 1 architecture), or the creatures responds to auditory events symbolically, associating them with a visual stimulus and responding as if that was seen (Type 2 architecture) (see figure 1). For an external observer, who only watches the information available to the creature and its motor responses, it may not be possible to see changes in the interpretation process. But the underlying mechanisms behind each sign process are qualitatively different.

\section*{The experiment}

The scenario to test the conditions for the emergence of semiotic processes is inspired by food foraging behavior of animals. Two types of resources can be found in the environment, with positive and negative values. Resources are differentiated by perceptual features and creatures can produce the two types of signs, which are also perceptually different. Here all creatures are potential utterers and interpreters. An evolutionary process is applied to allow creatures' adaptation to the task of collecting resources, which involves action selection, sensorial categorization, and sign production and interpretation.

The environment is a 50 by 10 grid and the resource (alternating positive and negative in each trial) is placed in one position. Five creatures are placed randomly at each trial, and they are capable of seeing resources within 2 positions distance and hear signs within 25 positions. Each perception of a resource or a sign corresponds to a sequence of bits. A resource is a 4 bit sequence, with positive ones starting with 01 bits and negative ones, with 10 bits. Signs produced by creatures have 3 bits, and can start with bits 01 (called sign 1) or with bits 10 (called sign 2). The other bits are randomly generated.

Creatures can execute a limited set of motor actions: move forward, turn left, turn right, stand still, positive visual taxis, negative visual taxis, positive auditory taxis, negative auditory taxis, protect; and vocal actions: produce sign 1, produce sign 2. Taxis actions are directional motor responses to sensorial input, with positive taxis guiding to the sensorial input position and negative taxis guiding away from it. When a sign is produced by a creature it can be heard by other creatures in the next instant.

Creatures are controlled by a feed-forward neural network with three layers (figure 2), with weights between -2.0 and +2.0 with 0.1 intervals. In type 1 cognitive architecture, auditory middle layer is connected to the output layer, but in type 2 architecture auditory middle layer is connected to the visual middle layer, defining an associative memory between auditory activations and visual activations.

To allow better analysis of neural network activation patterns and to augment the descriptive power of assessing


Figure 2: Neural networks used by creatures. Top: Type 1 architecture with direct connection between auditory middle layer and output layers. Bottom: Type 2
architecture with auditory middle layer connected to visual middle layer.
neurons activations, we applied a winner-takes-all (WTA) mechanism to the each middle layer and each output layer. In WTA, only the neuron with the highest positive activation (calculated as the sum of products of inputs and weights) is going to have an output of 1.0 . The other neurons within the layer will have a null activation. If no neuron has a positive activation, then all will have null output. Applying WTA activation, both visual middle layer and auditory middle layer perform a localist categorization, with each neuron responding for a given pattern from sensorial input. In the motor output layer and in the vocal output layer, the use of WTA allows only one motor action and one vocal action, at most, to be executed.

The neural networks of all creatures have initially a type 1 architecture with random weights. Evolution allows the creatures to adapt to the task of collecting positive resources and avoiding negative resources. Creatures that stand on the same position as the positive resource gain 10 resource units per instant. When a negative resource is placed on the environment, any creature standing in the same position loses 100 units. Besides, each creature, in any position, not executing a protection action when a negative resource is present, loses 10 units. However, execution of the protection action costs 5 units per instant.

Since this experiment involves the co-evolution of sign production and interpretation, to allow a stable communication system to emerge, selection is done at group level of identical individuals, following conclusions from Floreano et al (2007). The performance of each individual in the proposed task corresponds to the number of resource units collect by their group, composed of cloned individuals (same weights and architecture type).
The population of 500 individuals is divided in 100 groups of 5 clones. Each group is evaluated for 4 to 8 trials, half with positive resource and half with negative one. After groups are evaluated, one individual from the 20 best groups are selected for the next generation. These individuals will be included in the next generation along with 80 new individuals, and each of these 100 individuals will be cloned to form 100 groups of 5 clones again. To generate the 80 new individuals, the 20 best ones go through recombination and mutation. During recombination, if architecture types from parents are not the same, type 1 prevails. Mutation can change each weight to a new value also change architecture type, with \(1 \%\) chance of going from a type 2 to a type 1 and \(0.05 \%\) of changing from a type 1 to a type 2 .

During the evolutionary process, we observed groups performance in resource collecting task, the auditory middle layer connection (cognitive architecture type), and sign production, along with sign-object relations, sensorial categorization and motor responses.

\section*{Results}

To evaluate the emergence of sign interpretation and sign production, we simulated the experiment in an initial configuration. There are two cycles. In the first one, 300 generations, creatures did not have the auditory sensor and could not vocalize, and creatures were evaluated for 4 trials, in which they were near the resource. In cycle 2, after 300 generations, creatures were evaluated in 8 trials, 4 without communication, just as before, and 4 with communication, when one creature was placed 1 position away from the resource, but the other 4 were placed far away.

First simulation. During cycle 1, in a few generations, creatures adapted to the task of resource collecting, with the best groups obtaining around 2000 resource units. Creatures quickly acquired a response of protection action for a negative resource seen and a visual taxis response to a positive resource. Visual input categorization was also gradually adjusted, and ended up with creatures categorizing positive resources in neuron number 3 (see neuron numbers in figure 2), negative resources in neuron number 2 and null visual inputs in neuron number 1 .

At the start of the second cycle, resource collection performance dropped with the extra trials when most creatures are placed far away from the resource. Observing vocal responses to resources in the start, groups were split between those which vocalized sign 1 , sign 2 or no sign. Motor responses to signs, however, were not appropriate yet. The first adaptive use of communication appeared later on, with creatures vocalizing sign 1 to positive resources
and responding to it with auditory taxis, with sign 1 categorized in neuron 1 from auditory middle layer. In a further generation, this sign production behavior was expanded; a group vocalized sign 1 to both types of resources. This helped all creatures see the type of resource present in the environment, since they went after the sign utterer, saw the resource, and chose the correct motor response, but with a certain delay because of the distance they had to move. This response was a typical indexical interpretation of the sign that drew interpreters' attention to the utterer, searching for referents spatial-temporally related to the sign.

Later on, one group started a distinguished vocalization for each type of resource, with sign 1 being used for negative resource and sign 2 for positive one. Sign 1 was categorized in neuron 2 in auditory middle layer and had a protection action response. Sign 2 was categorized in neuron 1 with an auditory taxis response. This response dominated all groups after some generations and was the final behavior of all best groups. We could notice that type 1 architecture was dominant during all cycle 2 , and therefore indexical interpretation was the only interpretation processes during this simulation.

Second Simulation. A type 2 architecture, with symbolic interpretation, allows the establishment of multi-modal associations and could make auditory signs connect to visual signs. We have previously proposed that symbolic interpretation can act as a cognitive shortcut to an earlier acquired competence, when a cognitive trait is hard to be acquired (Loula et al., 2010). If the interpreter already had appropriate motor responses to visual inputs and this sensormotor coordination had a high acquisition cost, symbolic interpretation could connect auditory signs to visual signs and reuse the previous acquired competence.

In this second experiment, we modified the scenario of the first simulation by making it harder to learn motor coordination. The WTA activation in the output layer is changed so that the highest active neuron only stays active if its activation (sum of products of neuron inputs and weights) is 1.0 higher than the second highest active neuron. Since weights are between -2.0 and 2.0 , this severely limits possible connection weights.

We executed the simulation with this new configuration. Results are shown in figure 5, with the number of collected units and the auditory middle layer connection.
The initial resource collecting performance was much worse than in the first simulation, because most groups remained non-functional with the modified activation of output layer. It took much longer for creatures to respond appropriately to visual inputs. At the end of cycle 1 , the number of collected resources by the best groups was similar to the first simulation. At this point, creatures categorized positive resources in neuron number 1 of the visual middle layer, negative resources in neuron number 3, and null visual inputs in neuron 2 .

When cycle 2 started and creatures gained auditory sensor and vocal actions, all of them had the auditory middle layer


Figure 4: Evaluation of foraging task and auditory middle layer connection along the generations for the second simulation.


Figure 5: Activation of auditory and visual middle layers for each sign for the second simulation.
connected to the output layer. Different from the last simulation, almost none of the groups vocalized a sign. In the first generations of this second cycle, some groups already had the auditory middle layer changed to connect to the visual layer, but still without vocalizing. After some generations, the first group, to actually produce and interpret signs, vocalized sign 1 when a negative resource was seen, and when this sign was heard, it was categorized by neuron 1 in the auditory middle layer, which in turn activated neuron 3 in the visual middle layer, since creatures in this group had a type 2 architecture. Neuron 3 in visual layer is used to categorize seen negative resources, so an associative
memory is established. From then, other groups appear with different strategies, therefore, around generation 400, there was an intense competition between alternative strategies, with different architecture types and different categorizations, as can be seen in figures 4 and 5 .

During this dispute, a new strategy appeared with a group of creatures using distinct signs for each resource type: sign 2 for positive resources and sign 1 for negative ones. However, when either signs were heard, they were categorized in the same neuron (number 1) in the auditory middle layer and then connected to neuron number 1 in the visual middle layer, used to categorize seen positive
resources. Even though, diverse signs were produced, sign categorization was indistinct and the same motor response of visual taxis was produced. This was still an adaptive response, since even if a negative resource was present, it guided creatures close to the resource which, when seen, made creatures have a protection action. Note that this is different from what occurred in the first simulation, because in this case there was an interpretation (misleading, however) where any sign always refers to a positive resource, before the creature actually saw the resource.

This undistinguished categorization of signs, however, was not the best adaptive behavior still, since there was a delay in the protection action choice. In a later generation, starting from this strategy, a group categorized signs differently, with sign 2 being categorized in neuron 1 of the auditory middle layer, but sign 1 categorized in neuron 2 which is then associated with neuron 3 in the visual middle layer, used for seen negative resources, ending with an immediate protection action. This becomes the dominant communication strategy, with specific signs being produced by an utterer for each resource type and when an interpreter hears each signs. It internally associates signs with a referent and responds with an adaptive action, without the need of seeing the referent in the environment, so symbolbased communication emerged.

\section*{Discussion and Conclusion}

In previous works, we have simulated the emergence of interpretation processes based on symbolic and indexical signs. Here, besides interpretation processes, we were able to observe the emergence of sign production, sign-object and sign-object-interpretant relations, involving multiple signs, objects and interpretants.

In the first simulation, motor actions could be easily coordinated with sensorial input, and therefore an adaptive behavior evolved into a direct response to the communicated signs. This response defined an indexical interpretation and had two distinct moments for actual interpretation: first, interpreter's attention is directed to the utterer, then, when interpreter gets close to the utterer, it can visualize the object and thus the sign is spatiotemporally associated with the object. At the end of first simulation, this indexical interpretation occurs for sign 2, which is associated with a positive resource, as the interpreter creature performs auditory taxis and finds the utterer creature and the positive resource. In case of sign 1, we cannot say that the creatures interpret it as referring to a negative resource, because upon hearing sign 2 , the creatures perform a protection action and stop moving without seeing the negative resource.

Increasing cost of cognitive traits acquisition, we imposed a restriction on neural network's output layer activation in the second simulation. In this case, symbolic interpretation of signs was the adaptive response, confirming our hypothesis that symbolic processes can act as a cognitive shortcut mapping auditory signs to visual categories and reusing visual module mapping to motor actions.

In the second simulation, interpretation of each sign became distinguished with creatures associating each type of sign with a different visual category. Symbolic interpretation of signs emerged as a cognitive shortcut to already established visual-motor connections.

An important step for our experiment was to apply a neural network with a localist activation as the cognitive architecture for artificial creatures. This way, we were able to analyze middle layer categorization of sensorial inputs and what each category activates in the subsequent layer. For type 2 architecture, it was possible to determine the referent of each sign evaluating neural network activation. This model, therefore, allows an objective verification of what signs represent for interpreters in an experiment on the emergence of communication.

Additional investigations on semiotic processes and cognitive processes must be done to complement studies on the emergence of communication and language. We expect that a careful analysis of underlying semiotic processes will bring forth novel findings and stimulate new discussions.

\section*{References}

Cangelosi, A. (2001). Evolution of communication and language using signals, symbols, and words. IEEE Transactions on Evolutionary Computation, 5(2), 93-101. Floreano, D., Mitri, S., Magnenat, S., \& Keller, L. (2007). Evolutionary conditions for the emergence of communication in robots. Current Biology, 17, 514-519.
Loula, A., Gudwin, R. \& Queiroz, J. (2011) Cognitive conditions to the emergence of sign interpretation in artificial creatures. In Proceedings of the 11th European Conference of Artificial Life, ECAL'11, 2011, France. (p. 497-504)
Loula, A., Gudwin, R., \& Queiroz, J. (2010) On the emergence of indexical and symbolic interpretation in artificial creatures, or What is this I hear? In Fellermann, H., et al., editors, Artificial Life XII, pages 862--868. MIT Press.
Nolfi, S. and Mirolli, M., Eds. (2010). Evolution of Communication and Language in Embodied Agents. Berlin: Springer Verlag.
Parisi D. (2001). Simulazioni - la realtà rifatta nel computer. Il Mulino : Bologna.
Peirce, C.S. (1958) Collected Papers of Charles Sanders Peirce. Harvard University Press, Cam-bridge, Mass.
Peirce, C.S. (1967) Annotated catalogue of the papers of Charles S. Peirce. Amherst: University of Massachusetts.
Queiroz, J. \& Ribeiro, S. (2002). The biological substrate of icons, indexes, and symbols in animal communication: A neurosemiotic analysis of vervet monkey alarm calls. In The Peirce Seminar Papers 5. New York: Berghahn.
Ribeiro S, Loula A, Araújo I, Gudwin R, Queiroz J (2007) Symbols are not uniquely human. Biosystems 90:263-272.
Wagner, K., Reggia, J., Uriagereka, J., \& Wilkinson, G. (2003) Progress in the simulation of emergent communication and language. Adaptive Behavior, 11(1), 37-69.

\title{
Spatial Correspondence Parameters at the Basis of Transfer of Learning in Social Contexts
}

\author{
Luisa Lugli (l.lugli@unibo.it) \\ Department of Communication and Economics, University of Modena and Reggio Emilia, via Allegri 9 \\ 42121 Reggio Emilia, Italy \\ Cristina Iani (cristina.iani@unimore.it) \\ Department of Communication and Economics, University of Modena and Reggio Emilia, via Allegri 9 42121 Reggio Emilia, Italy \\ Nadia Milanese (nadia.milanese@gmail.com) \\ Department of Communication and Economics, University of Modena and Reggio Emilia, via Allegri 9 42121 Reggio Emilia, Italy \\ Sandro Rubichi (sandro.rubichi@unimore.it) \\ Department of Communication and Economics, University of Modena and Reggio Emilia, via Allegri 9 42121 Reggio Emilia, Italy
}

\begin{abstract}
Recent works indicated that performing a joint spatial compatibility task with an incompatible stimulus-response mapping affects subsequent joint Simon task performance, eliminating the social Simon effect (social transfer of learning effect or SToL effect). Crucially, the SToL effect was not tuned to the specific identity of the co-actor, and depended on the overlap between the spatial relations of the practice and transfer tasks. Starting from these findings, this study aimed at investigating which spatial relations between stimulus ( S ), response (R) or participant \((\mathrm{P})\) positions are relevant for the SToL effect to occur. Two experiments were run in which the participant-response associations were incompatible (participants were required to respond with crossed arms), whereas the stimulus-response and stimulus-participant associations were manipulated. We found that learning derived from the practice task did not transfer to the subsequent task when stimulus-response associations were spatially incompatible and stimulus-participant associations were compatible (Experiment 1). However, a SToL effect was evident when stimulus-participant associations were spatially incompatible and stimulus-response associations were compatible (Experiment 2), hence suggesting that the spatial relation between stimulus and participant positions is crucial for the SToL effect to occur.
\end{abstract}

Keywords: social cognition; joint performance; spatial compatibility; social transfer-of-learning

\section*{Introduction}

Learning involves the acquisition and modification of new or existing knowledge through the application of which humans may be able to perform new tasks. This knowledge is shaped by the experience humans could acquire alone or in a social context (e.g., Vygotsky, 1978).

As regards individual performance, there is evidence that knowledge acquired in a task (i.e., practice task) can be transferred to and affects the way a subsequent task (i.e., transfer task) is performed. In the transfer-of-learning (ToL) paradigm, developed by Proctor e Lu (1999, see also Iani, Rubichi, Gherri, \& Nicoletti, 2009) participants are required to perform a spatial compatibility task with an incompatible stimulus-response (S-R) mapping (i.e., they are instructed to press a right key when a left stimulus is presented and a left key when a right stimulus is presented), followed by a Simon task in which stimulus location is irrelevant and responses have to be emitted on the basis of a non spatial stimulus feature (e.g., color). When the Simon task is performed alone, performance is more efficient when stimuli and responses spatially correspond (corresponding trials) than when they do not correspond (non-corresponding trials). Thus, if participants are instructed to press a right key to a red stimulus and a left key to a green stimulus, their reaction times (RTs) will be shorter and accuracy higher if a red stimulus appears on the left compared to when it appears on the right. The influence of the irrelevant spatial stimulus feature on performance is known as the Simon effect (Simon \& Rudell, 1967; Rubichi \& Nicoletti, 2006; Rubichi, et al., 1997; Rubichi, et al., 2004; for reviews, see Proctor \& Vu, 2006; Rubichi, et al., 2006).
It has been demonstrated that performance on the Simon task could be modulated, that is the Simon effect is reduced, eliminated or reversed, after practicing with a spatially incompatible mapping (e.g., Iani et al., 2009; Proctor \& Lu, 1999; Tagliabue, et al., 2000). This is thought to occur because the non-corresponding stimulus-response associations acquired during the transfer task remain active and influence performance in the subsequent Simon task. Hence, the fact that after an incompatible practice the Simon
effect is modulated indicates that performance depends not only on the goals of the task that is currently being performed, but also on immediate prior experience.
Sebanz and colleagues (Sebanz, Bekkering, \& Knoblich, 2006; Sebanz, Knoblich, \& Prinz, 2003) have shown that the Simon effect occurs even when the Simon task is shared between two agents with each one responding only to one stimulus color (from now on, social Simon effect). In the social variant of the Simon task, one participant has to press the left key in response to green stimuli and the other participant has to press the right key in response to red stimuli, so that each participant is performing a go/no-go task. The observation of a social Simon effect provides evidence that, although each participant is responsible for only half of the task and hence for only one response alternative, they tend to represent the co-actor's task and to integrate self and other's task into a common representation (see also Ferraro et al., 2012). Starting from these evidences, two studies (Milanese, Iani, \& Rubichi, 2010; Milanese, Iani, Sebanz, \& Rubichi, 2011) investigated, by means of the social transfer of learning (SToL) paradigm, whether and to what extend specific contextual determinants influence the way knowledge acquired in a given task could be transferred to a subsequent one. Milanese et al. (2010) modified the transfer of learning paradigm used in individual context. In their modified paradigm (from now on, the SToL paradigm), two participants performed together the spatial compatibility task (practice task) and the Simon task (transfer task) one after the other. They found that individually and jointly acquired stimulus-response associations acquired in the practice task remained active and transferred to the joint Simon task leading to an elimination of the joint Simon effect, whereas jointly acquired stimulus-response associations did not transfer to individual task performance. In other words, transfer-oflearning effects were maximal only when both practice and transfer tasks took place in a social setting, suggesting that what was transferred was not only what was specifically practiced, but also aspects of the interactive context in which learning took place.
Milanese et al. (2011) further investigated the elements of the contexts that needed to remain constant for transfer between a jointly performed practice task and a subsequent joint transfer task to occur, that is the identity of the co-actor and the spatial relation between the two co-actors. Results showed that a spatially incompatible practice performed jointly with another person influenced performance on a subsequent joint Simon task even if the co-actor's identity changed (Experiment 1), whereas when participant's position changed from the practice to the transfer task (that is, participant sitting on the left in the practice session sat on the right in the transfer session, the opposite was true for the other participant), the social-transfer-of-learning effect did not occur (Experiment 2). To sum up: the SToL effect was not tuned to the specific identity of the co-actor, and depended on the overlap between the spatial relations of the practice and transfer tasks.

Starting from these results, one might wonder which specific spatial relations are really necessary in order to obtain the modulation of performance on the subsequent joint Simon task. We know that in the solo condition, the non-corresponding link between stimulus and response positions is crucial. What does it happen when the joint task requires a further spatial determinant that is the participant's position? In other words, what does it happen if participants are required to take into consideration both the position of the response-key and the position of their body? To this aim, we performed two experiments, using the SToL paradigm, in which we manipulated the spatial relation between the stimulus, the response and the participant. For sake of clarity, the position of the response-key referred to the right/left button location on the keyboard, and the participant's position referred to the left/right displacement of the participant's body with respect to the center of the table. In both experiments the participant-response associations were incompatible (participants were required to respond with crossed arms), whereas the stimulusresponse and stimulus-participant associations were manipulated. In the practice session of Experiment 1, stimulus-response associations were spatially incompatible, while stimulus-participant ones were spatially compatible. We achieved this by requiring participants to respond with their arms crossed to the stimulus which was contralateral with respect to the position of the response-key (i.e., for instance, participants sitting on the left responded by pressing the right key to the left stimulus). In the practice session of Experiment 2, stimulus-participant associations were spatially incompatible, while stimulus-response ones were spatially compatible. In this experiment, participants were required to respond with their arms crossed to the stimulus that was contralateral with respect to their sitting position (i.e., for instance, participants sitting on the left responded by pressing the right key in response to the right stimulus). These manipulations will allow us to define which incompatible association is crucial for the SToL effect to occur.

\section*{Experiment 1}

\section*{Method}

Participants Sixteen students (1 male; 4 left-handed; age range: 19-26 years) of the University of Modena and Reggio Emilia took part in Experiment 1 for partial fulfillment of course credit. They reported normal or corrected-to-normal vision and were naïve as to purpose of the study. Once selected, they were randomly paired.
Apparatus and stimuli As in Milanese et al. (2010, 2011), stimuli in the spatial compatibility task were white solid squares ( 4.5 X 4.5 cm ), whereas stimuli in the Simon task were red or green solid squares ( \(4.5 \times 4.5 \mathrm{~cm}\) ). All stimuli were presented on a black screen, 9.5 cm to the left or to the right of a central fixation cross ( 1 X 1 cm ). Stimulus presentation was controlled by an IBM computer. In both tasks, responses were executed by pressing the "z" or "-" keys of a standard Italian keyboard with the left or right
index finger, respectively. In the spatial compatibility task participants' hands were crossed (the participant sitting on the left pressed the right key with his/her left hand; the participant sitting on the right pressed the left key with his/her right hand).Viewing distance was about 60 cm .
Procedure The experiment consisted of two consecutive sessions: a practice session and a transfer session. Participants first performed a joint spatial compatibility task (practice session) with an incompatible mapping (stimulus positions were mapped incompatibly to response-key positions, that is participants were required to respond to the controlateral stimulus with respect to the response-key position). Each participant was instructed to respond to only one of the two stimulus locations by pressing the contralateral key (by crossing their arms) and refraining from responding when a stimulus appeared in the alternative position. Hence, half of the participants responded to left stimuli by pressing a right key, whereas the other half responded to right stimuli by pressing a left key (see Fig. 1). After a 5 -min rest, participants were administered a joint Simon task (transfer session), in which the red and green stimuli were always location-irrelevant trials. Participants were instructed to respond to only one stimulus color by pressing the key at their disposal. For half of the participants, instructions required to press the right key in response to red stimuli and the left key in response to green stimuli, whereas for the other half instructions required to respond with the opposite stimulus-response mapping. Each participant kept the same position in both practice and transfer tasks, but changed the response position from the practice to the transfer task. That is, for instance, the participant sitting on the left and responding with the right key in the practice session sat on the left and responded with the left key in the transfer session (see Fig. 1).
In both tasks, a trial began with the presentation of the fixation cross at the center of a black background. After 1000 ms the stimulus appeared to the left or to the right of the fixation. In the spatial compatibility task, the stimulus remained visible for 600 ms , and the maximum time allowed for a response was 1200 ms . In the Simon task, the stimulus remained visible for 800 ms and the maximum time allowed for a response was 1000 ms . The inter-trialinterval was 1 s , and it was initiated immediately after the response was made.
The spatial compatibility task was composed of 12 training trials and 300 experimental trials divided into 3 blocks. The Simon task consisted of 12 training trials and 160 experimental trials divided into two blocks of 80 trials each. For both tasks, instructions stressed both speed and accuracy of performance.


Figure 1: Schematic representation of the experimental conditions used in the two experiments. In the practice session the participant sitting on the left (A) was required to press the right key in response to the left stimulus (participant-response and stimulus-response associations were spatially incompatible and the stimulus-participant association was compatible, Experiment 1) or to the right stimulus (participant-response and stimulus-participant associations were spatially incompatible and the stimulusresponse association was compatible, Experiment 2). In both experiments, each participant kept the same sitting position in both practice and transfer tasks, while the position of the response-key changed.

\section*{Results and discussions}

Since our predictions concern performance in the joint Simon task, for the current and the following experiment we report only the data for the Simon task (transfer session). Correct reaction times (RTs) were submitted to a repeatedmeasures Analysis of Variance (ANOVA) with Correspondence (corresponding vs. non-corresponding) as within-subject factor.

Responses in corresponding trials ( 328 ms ) were faster than responses in non-corresponding trials ( 350 ms ), \(F(1,15)=48.18, p<.001\), see Fig. 2. The significant 22 -ms Simon effect indicates that the joint Simon task was not influenced by prior joint performance on the spatial compatibility task \({ }^{1}\). Thus, practice on a spatial compatibility

\footnotetext{
\({ }^{1}\) In order to understand whether the \(22-\mathrm{ms}\) social Simon effect found in Experiment 1 was influenced by prior practice, we compared the data of this experiment with the data of the baseline condition of Milanese et al. (2010)'s Experiment 1 in which a 14ms social Simon effect was evident. Correct RTs for the two conditions were submitted to an ANOVA with Correspondence (corresponding vs. non-corresponding trials) as within-subject factor and Condition (baseline vs. Experiment 1) as between-
}
task with an incompatible association between the participant and the location of the response-key is not sufficient to produce SToL. These results suggested that stimulus-participant associations may play a crucial role in the occurrence of the STol effect.


Figure 2: Means reaction times (ms) for the transfer session of Experiments 1 and 2 as a function of stimulus-response correspondence.
subjects factor. The Correspondence x Condition interaction did not reach significance, \(F(1,30)=2.91, p=.10\), indicating that the effects found in the two experiments did not differ. Based on this result, we can safely conclude that in our Experiment 1 there was no evidence of SToL.

\section*{Experiment 2}

\section*{Method}

Participants Sixteen new right-handed students of the University of Modena and Reggio Emilia (all female; age range 19-20 years), selected as in the previous experiment, took part in Experiment 2.
Apparatus and stimuli and procedure Apparatus and stimuli were the same as in Experiment 1, whereas the procedure varied as follows. Participants performed the joint spatial compatibility task with a different incompatible mapping: stimulus positions were mapped incompatibly to participants seating position, that is participants were required to respond to controlateral stimuli with respect to their seating position. Each participant was instructed to respond to only one of the two stimulus locations by pressing the contralateral key (by crossing their arms) and refraining from responding when a stimulus appeared in the alternative position. Each participants kept the same position in both practice and transfer tasks, while the position of the response-key changed. That is, for instance, the participant sitting on the left and responding with the right key in the practice session sat on the left and responded with the left key in the transfer session (see Fig. \(1)\).

\section*{Results and discussion}

Correct RTs were submitted to an ANOVA with Correspondence as within-subject factor. RTs did not differ between corresponding ( 316 ms ) and non-corresponding trials ( 319 ms ), \(F<1\) (see Fig. 2). The lack of a significant Simon effect is indicative of SToL. This result can be taken as an indication that it is the spatial association between the stimulus and the participant acquired during the practice task that is crucial for the occurrence of the SToL effect, while stimulus-response associations are irrelevant.

\section*{Additional analysis}

In order to compare the Simon effect found in the two experiments, we ran an ANOVA with Correspondence (corresponding vs. non-corresponding trials) as withinsubject factor and Experiment (Experiment 1 vs. Experiment 2) as between-subjects factor. Responses in corresponding trials ( 322 ms ) were faster than responses in non-corresponding trials \((334 \mathrm{~ms}), F(1,30)=25.64, p<.001\). The main effect of Experiment was nearly significant, \(F(1,30)=3.64, p=.07\), showing that responses were slower in Experiment \(1(339 \mathrm{~ms})\) than in Experiment \(2(317 \mathrm{~ms})\). The Correspondence x Experiment interaction was significant, \(F(1,30)=15.23, p<.001\), indicating that the \(22-\mathrm{ms}\) found in Experiment 1 differed from the \(3-\mathrm{ms}\) (non-significant) effect found in Experiment 2.

\section*{Discussion}

It is well known that in the ToL paradigm, when the practice task is performed in a solo condition, what is acquired and transferred to the subsequent Simon task is an association between stimulus and response-key positions (Proctor \& Lu, 1999; see also see also Iani, et al., 2009).
The results of previous studies (e.g., Milanese et al., 2010, 2011) seem to suggest that in social settings a crucial factor for the occurrence of transfer-of-leaning effects may be the type of relation between the participant and the stimulus positions acquired during practice rather than the relation between stimulus and response-key positions. The present study was aimed at assessing the relative contribution of the spatial relations between stimulus and response-key, between stimulus and participant or rather between participant and response-key positions in the occurrence of the SToL effect. More specifically, the participant-response associations were always incompatible (participants were required to respond with crossed arms), whereas stimulusresponse and stimulus-participant associations were manipulated. In this way we were able to investigate independently whether crucial for the SToL effect to occur is the incompatible association between stimulus-response positions (Experiment 1) or between stimulus-participant positions (Experiment 2).
We found a SToL effect when participant-response and stimulus-participant associations were spatially incompatible and stimulus-response associations were compatible (Experiment 2). No evidence of SToL was found when participant-response and stimulus-response associations were spatially incompatible and stimulusparticipant association were compatible (Experiment 1).
The present findings suggest that the incompatible association between the positions of the stimulus and of the participant may be crucial for the emergence of the SToL effect. It would seem, thus, that in a joint setting, where participants are (implicitly) required to take into account the presence of another person, the participant's position acquires greater relevance than in a solo setting. These results point to an intriguing and debated issue remained open so far: do correspondence effects emerging in joint setting depend on the relationship not only between stimuli and responses but also between stimuli and responding agents?
Recently, Philipp and Prinz (2010; see also Liepelt, et al., 2010) proposed that the joint compatibility effect may rely not only on the stimulus-response spatial correspondence (as is known to be crucial for the standard Simon effect to occur), but also on social correspondence, that is the one between stimulus and responding agents. According to these authors, when the Simon task is shared between two acting individuals, space may be used as an indication of whose turn it is. This would mean that a stimulus appearing on the left does not bring to the automatic activation of the left response, but rather is perceived as a stimulus signaling that the person sitting on the left is in charge of responding.

Starting from this account, our study investigated whether social correspondence may play a crucial role also in the SToL effect. It should, however, be noted that in Philipp and Prinz's study, the positions of the participant and of the response-key always corresponded and hence the correspondence between stimulus and response-key position could not be distinguished from the correspondence between the stimulus and the responding agent. In the current study, we separated the positions of the response-key and of the participant, as in the practice task participants were required to respond with crossed arms. In this way we were able to investigate independently whether crucial for the SToL effect to occur is the incompatible association between stimulus and response-key positions (Experiment 1) or between stimulus and participant positions (Experiment 2). Based on the findings of the current study, the observation of a null joint Simon effect in the transfer task could be the result of the acquisition and subsequent transfer of the incompatible link between the stimulus and participant positions. In other words, as the present results suggest, the SToL effect in these studies may rely not only on the spatial association correspondence between stimulus and response, differently from the individual condition (ToL effect), but also, and crucially, on the incompatible link between the stimulus and participant positions. These results underline the importance of both spatial and social features. Indeed, it seems that acting in a social context increases the importance of the participants' position with respect to the stimulus.
To conclude, in this study, we used the social transfer of learning paradigm to identify which elements of the context need to remain constant for social transfer-of-learning effects to occur. This issue is not trivial, because sometimes the practice context (i.e., the context in which we implicitly acquire new knowledge) and the transfer context (i.e., the subsequent context in which we utilize the acquired knowledge) are not identical and may differ in several aspects. The results of this study are particularly relevant since they provide insights on the way we represent another's task (and how we integrate the other agent information about action with our information, see Knoblich \& Jordan, 2003) in particular joint action situations.

\section*{Acknowledgments}

This work was supported by MIUR PRIN (project 2008ZN5J5S).

\section*{References}

Ferraro, L., Iani, C., Mariani, M., Nicoletti, R., Gallese, V., Rubichi, S. (2012). Look What I Am Doing: Does Observational Learning Take Place in Evocative TaskSharing Situations? PLoS ONE, 7(8): e43311. doi:10.1371/journal.pone.0043311.
Iani, C., Rubichi, S., Gherri, E., \& Nicoletti, R. (2009). Cooccurrence of sequential and practice effects in the Simon
task: Evidence for two independent mechanisms affecting response selection. Memory \& Cognition, 37, 358-367.
Knoblich, G., \& Jordan, J.S. (2003). Action Coordination in Groups and Individuals: Learning Anticipatory Control. Journal of Experimental Psychology: Learning, Memory and Cognition, 29(5), 1006-1016.
Liepelt, R., Wenke, D., Fischer, R., \& Prinz, W. (2010). Trial-to-trialsequential dependencies in a social and nonsocial Simon task. Psychological Research, 75, 366-375.
Milanese, N., Iani, C., \& Rubichi, S. (2010) Shared learning shapes human performance: transfer effects in task sharing. Cognition, 116(1), 15-22.
Milanese, N., Iani, C, Sebanz, N., \& Rubichi, S. (2011) Contextual determinants of the social-transfer-of-learning effect. Experimental Brain Research, 211, 415-422.
Philipp, A.M., \& Prinz, W. (2010).Evidence for the role of the responding agent in the joint compatibility effect. Quarterly Journal of Experimental Psychology, 63, 21592171.

Proctor, R. W., \& Lu, C.-H. (1999). Processing irrelevant information: Practice and transfer effects in choicereaction tasks. Memory \& Cognition, 27, 63-77.
Proctor, R. W., \& Vu, K.-P. L. (2006). Stimulus-response compatibility principles: Data, theory, and application. Boca Raton, FL: Taylor \& Francis.
Rubichi, S., \& Nicoletti, R. (2006).Simon effect and handedness: Evidence for a dominant-hand attentional bias in spatial coding. Perception \& Psychophysics, 68(7), 1059-1069.
Rubichi, S., Nicoletti, R., Iani, C., \& Umiltà, C.A. (1997). The Simon effect occurs relative to the direction of an attention shift. Journal of Experimental Psychology: Human Perception and Performance, 23, 1353-1364.
Rubichi, S., Nicoletti, R., Pelosi, A., \& Umiltà, C. A. (2004). Right-left prevalence effect with horizontal and vertical effectors. Perception \& Psychophysics, 2, 255263.

Rubichi, S., Vu, K., Nicoletti, R., \& Proctor, R. (2006). Two-dimensional spatial coding. Psychonomic Bulletin \& Review, 13, 201-216.
Sebanz, N., Bekkering, H., \& Knoblich, G. (2006). Joint action: Bodies and minds moving together. Trends in Cognitive Sciences, 10, 70-76.
Sebanz, N., Knoblich, G., \& Prinz, W. (2003). Representing others' actions: just like one's own? Cognition, 88, 11-21.
Simon, J. R., \& Rudell, A. P. (1967). Auditory S-R compatibility: The effect of an irrelevant cue on information processing. Journal of Applied Psychology, 51, 300-304.
Tagliabue, M., Zorzi, M., Umiltà, C., \& Bassignani, C. (2000). The role of LTM links and STM links in the Simon effect. Journal of Experimental Psychology: Human Perception and Performance, 26, 648-670.
Vygotsky, L. S. (1978). Mind and society: The development of higher mental processes. Cambridge, MA: Harvard University Press.

\section*{Body movements affect Counting}

\author{
Luisa Lugli (l.lugli@unibo.it) \\ Department of Philosophy and Communication Studies, University of Bologna,Via A. Gardino 23, 40122 Bologna, Italy \\ Department of Communication and Economics, University of Modena and Reggio Emilia, via Allegri 9, 42121 Reggio Emilia, Italy \\ \section*{Giulia Baroni (giulia.baroni4@unibo.it)} \\ Department of Philosophy and Communication Studies, University of Bologna,Via A. Gardino 23, 40122 Bologna, Italy \\ Filomena Anelli (filomena.anelli@unibo.it) \\ Department of Philosophy and Communication Studies, University of Bologna,Via A. Gardino 23, 40122 Bologna, Italy Department of Education Sciences, University of Bologna,Via Filippo Re 6, 40126 Bologna, Italy \\ \section*{Anna M. Borghi (annamaria.borghi@unibo.it)} \\ Department of Psychology, University of Bologna, Viale Berti Pichat 5, 40127 Bologna, Italy \\ Institute of Cognitive Sciences and Technologies, National Research Council, Via S. Martino della Battaglia, 4400185 Roma, Italy \\ Roberto Nicoletti (roberto.nicoletti@unibo.it) \\ Department of Philosophy and Communication Studies, University of Bologna,Via A. Gardino 23, 40122 Bologna, Italy
}

\begin{abstract}
Aim of the present study is to investigate whether and to what extent movements performed with the whole body can influence calculation processes. Participants were asked to perform additions or subtractions while executing an ascending or descending movement in a passive (i.e., by taking the elevator) or active (i.e., by taking the stairs) mode. Results revealed a congruency effect between the type of calculation made and the direction of the movement performed, but only when participants experienced it through a passive mode. Our data are in line with studies providing evidence of a strict link between numerical and spatial representations, and between motor actions and number magnitude processing (motor-to-semantic effect). Implications of the results for the embodied and grounded nature of numerical cognition will be considered and discussed.
\end{abstract}

Keywords: numerical cognition; body movement; embodied cognition

\section*{Introduction}

One of the main challenges of embodied and grounded cognition views (e.g., Barsalou, 2008) is to account for the representation of abstract concepts, such as numbers. The present study focuses on numerical cognition, an important area in which this challenge can be addressed. The mental representation of quantity (i.e., number magnitude) is the main focus of an increasing number of researches. In a well known study, Deahene and colleagues (Dehaene, Bossini, \& Giraux, 1993) demonstrated that numbers ranging from 0 to 4 and from 6 to 9 facilitate left and right responses, respectively, even if number magnitude was a taskirrelevant feature. These results led the authors to claim that number magnitude is visuo-spatially represented along an horizontal mental number line (MNL) with small numbers on the left and large numbers on the right of a continuum (for a review, see Hubbard, Pinel, Piazza, \& Dehaene, 2005). In a recent study, Holmes and Lourenco (2012) investigated the strength of this Spatial Numerical Association of Response Codes (SNARC effect), on both the horizontal (i.e., small left; large - right) and vertical (i.e., small - bottom; large - up) MNL and found that the horizontal organization is rather dominant. Their data also showed, though, that the vertical organization depends on how number are conceptualized. In other words, the vertical organization emerged only when numbers are conceptualized as magnitudes that elicit an orientation (e.g., \(1^{\text {st }}\) floor from surface, \(2^{\text {nd }}\) floor from surface, etc.).

To our knowledge, all the studies so far were aimed at investigating whether and how the numerical magnitude, that is the representation of the number as a small or large quantity, impacts two processes. First, recent studies analyzed the influence of the numerical magnitude on the spatial representation of the horizontal/vertical axis, that is left/low and right/high spatial position (Fisher, Castel, Dodd, \& Pratt, 2003; Nicholls, Loftus, \& Gevers, 2008; Pecher \& Boot,
2011). Second, several researches focused on the impact of the magnitude on the action related processes investigating the bi-directional interaction between the magnitude code and the action plans of grasping objects with precision or power grip (Andres, Ostry, Nicol, \& Paus, 2008; Badets, Andres, Di Luca, \& Pesenti, 2007; Badets, Bouquet, Ric, \& Pesenti, 2012; Badets \& Pesenti, 2010, 2011; Chiou, Chang, Tzeng, \& Wu, 2009; Ranzini et al., 2012).

Starting from this evidence, the current study aims at investigating whether and to what extent processes leading to numerical magnitudes, such as arithmetical calculations of addition and subtraction, can be influenced by real upward and downward movements experienced with the whole body. In a study by Knops, Viarouge, and Dehaene (2009) participants tended to select the numerosity displayed in the upper right location for additions, and in the upper left location for subtractions (Space-Operation Association of Responses: SOAR). Differently from Knops and colleagues, our task required participants to: (a) keep adding or subtracting the same quantity from a starting number in a given period of time; (b) repeat the result of each calculation aloud, so that participants had to focus on the on-line and progressive calculation process; (c) experience an ascending or descending movement with the whole body in a passive (i.e., taking the elevator) or active (i.e., taking the stairs) mode.

In a recent research, Hartmann, Grabherr, and Mast (2011) demonstrated that the passive displacement of the body was sufficient to influence the magnitude of self-generated numbers. In their Experiment 1, participants were positioned on a motion platform and were asked to generate numbers at random while the platform was moving (leftward, rightward, downward, upward, forward, and backward) or when it was stationary. Results indicated a bias for small numbers, which were generated more easily during leftward and downward motions as compared to rightward and upward motions, respectively. Differently from Hartman et al. (2011), we asked participants to experience the movements through an active and a passive mode, and to make additions and subtractions rather than generating random numbers.

To summarize, our experimental procedure is new with respect to current literature in two aspects: first, we focused on the processes leading to a numerical magnitude, rather than focusing on the number magnitude per se. Second, we asked participants to experience real movements engaging their whole body through passive and active modes. We hypothesized a congruency effect between the direction of the experienced ascending or descending movement and the spatial orientation inferred by the type of calculation made, that is additions-upward orientation vs. subtractions-downward orientation. More precisely, we predicted that participants would be facilitated when asked to make additions while experiencing an upward movement and subtractions while experiencing
a downward movement (congruent conditions) as compared to the opposite instructions (incongruent conditions). Finally, we also assessed whether this congruency effect was modulated by the passive or active mode through which the movements were experienced. It is crucial to say that the type of movements and the sense of their direction were different in these two modes. In fact, when taking the elevator participants passively perceived themselves moving in a given direction. Conversely, when the stairs were taken, participants performed an overt and real motor action with a full physical body involvement. Moreover, the sense of the movement was fast and clearly vertical when using the elevator, while it was more progressive and less vertical when using the stairs, since that the awareness of going up or down changed progressively step by step. Hence, these two modes can modulate results differently: if the congruency effect requires an active motor process, we hypothesize to find the effect only in the stairs mode. Conversely, if the fast and vertical passive displacement of the body is sufficient to obtain the congruency effect, we expect the effect even when participants take the elevator.

\section*{Method}

Participants Twenty-eight undergraduate students from the University of Bologna ( 15 females, age range 19-24 years) took part in the experiment in exchange of 5 Euros. The majority of them had a background in humanities. All were naïve as to the purpose of the experiment and gave informed consent.

Apparatus and Stimuli Participants were asked to add or subtract 3 to a starting number (e.g., 342) for 22 seconds and to say the result of each calculation aloud (e.g., \(345,348,351\) or \(339,336,333\) and so on, for additions and subtractions, respectively, until 22 seconds were elapsed). In order not to make the calculation process too easy, the starting numbers: a) were composed by three digits; b) started with two different digits (i.e., 3 or 5 , such as 378 or 516).

Procedure The task required participants to make the additions or subtractions while taking the elevator or taking the stairs. In order not to mix these modes, the task was divided in two blocks, whose order was balanced between subjects. In one block, participants performed the calculations while taking the elevator, whereas in the other block calculations were performed while taking the stairs. Within each block, four trials were performed, resulting from the combination of the two types of calculation, additions and subtractions, and movement, ascending and descending. We designed each block in order to make additions and subtractions always alternate: an addition was always followed by a subtraction and vice versa. At the beginning of each trial, the experimenter explicitly informed the participant about the type of calculation that had to be performed and about the type of movement that was going to be experienced. For each trial, the experimenter spoke the starting number aloud
and then a go signal followed. Immediately after the go signal, the passive or active movement began and, at the same time, participants were required to repeat the starting number and then to keep saying the result of each calculation aloud for 22 seconds consecutively, until the experimenter gave the stop signal. If a calculation error was made, the trial was stopped and a new trial began with a different starting number. No feedback was given during the task and the importance of accuracy over speed was stressed at the beginning of each trial. The experimenter was present during the whole experiment. For the passive mode, she went up/down using the elevator together with the participant. For the active mode, she walked close to the participant while going up/down the stairs and asked the participant to keep her pace throughout the whole movement. In other words, each participant and the experimenter went up/down the stairs together so that the number of steps taken was held constant across participants. Overall, the experiment lasted about 15/20 minutes.

\section*{Results}

We considered the correct number of calculations made within the 22 seconds time window as our dependent variable. Since we predicted a congruency effect between the direction of the movement and the type of calculation, we divided the trials in congruent (ascending movements - additions; descending movements - subtractions) and incongruent (ascending movements - subtractions; descending movements additions), and then we averaged the number of calculations separately for each group of pairings.

The correct calculations were entered into a repeatedmeasures ANOVA with Congruency (congruent vs. incongruent) and Mode (elevator-passive movement mode vs. stairs-active movement mode) as withinsubjects factors. The Congruency \([F(1,27)=10.20\), \(\left.\operatorname{MSE}=7.51, n^{2}{ }_{p}=0.27, p<.01\right]\) and Mode \([F(1,27)=\) \(\left.32.50, \mathrm{MSE}=3.96, n_{p}^{2}=0.55, p<.001\right]\) factors were significant. We found a higher number of calculations: a) for congruent pairings ( \(\mathrm{M}=11.2\) ) with respect to incongruent ones ( \(\mathrm{M}=10.7\) ); b) when the movement was experienced with a passive \((\mathrm{M}=12)\) with respect to an active ( \(\mathrm{M}=9.8\) ) mode. Crucially, the Congruency x Mode interaction was significant \([F(1,27)=5.16\), MSE \(\left.=1.36, n_{p}^{2}=0.16, p=.03\right]\). Fisher's LSD posthoc test showed that, in the passive mode, participants performed a higher number of calculations for congruent pairings than for incongruent ones ( \(\mathrm{Ms}=\) 12.5 vs. 11.5 , respectively, \(p<.01\) ). Conversely, this pattern failed to emerge for the active mode ( \(\mathrm{Ms}=9.9\) vs. 9.8, for congruent and incongruent pairings, \(p=\) 0.95 , see Figure 1).


Figure 1: Number of calculations for congruent and incongruent pairings performed through a passive (i.e., elevator, leftmost panel) or active (i.e., stairs, rightmost panel) mode. Bars are standard error of the mean.

\section*{Discussion}

We investigated whether calculation processes, such as additions and subtractions, are influenced by real movements experienced with the whole body. Our findings showed a facilitation, in terms of correct number of calculations made, for the congruent condition, that is when additions and subtractions were performed while experiencing an ascending and descending movement, respectively, with respect to the opposite mapping. This result is in line with recent behavioral findings showing the influence of the motor process over the semantic one (i.e., motor-to-semantic effect, see Badets et al., 2012; Badets \& Pesenti, 2010, 2011; Ranzini et al., 2012).

The fact that we did not find the congruency effect for the active mode could be due to different factors. First, climbing the stairs required an overt movement and a full physical body involvement. Second, the sense of the movement direction was more progressive and less vertical, with respect to when the elevator was used, since the awareness of going up or down changed step by step. Hence, the lack of a congruency effect could be due to the higher amount of resources required by the dual-task of climbing the stairs and counting at the same time, as also suggested by the lower amount of calculations yielded for this mode compared to the passive one. In addition, the fact that the movement was probably perceived as faster and more vertical when taking the elevator, could have yielded a significant effect for this mode only.

In a recent study Hartmann and colleagues (2011, Experiment 1; see also Hartmann, Farkas, \& Mast, 2012) found that the number generation process was influenced by experiencing a passive whole body movement. Our results are in line with these findings, but we also obtained a congruency effect between the body movements and the calculation processes, instead of a given set of numbers. Moreover, participants in our study also experienced real movements with their whole body instead of passive movements while seated in a chair. We claim, indeed, that experiencing
ascending and descending movements with the whole body can influence the processes responsible for numbers representations as magnitude with an upward and downward orientation, that is addition and subtraction calculations.

Our results can have intriguing implications for the embodied and grounded cognition view, which claims a close link between perception and action, due to the influence of both our sensory-motor system and previous experiences on the cognitive processes (e.g., Barsalou, 2008). Of particular relevance to our work is the debate concerning the claim that both concrete and abstract concepts, such as numbers, are grounded in perception-action systems (e.g., Borghi \& Pecher, 2011; Fischer, 2012; Gianelli, Ranzini, Marzocchi, Rettore Micheli, \& Borghi, 2012; Pecher \& Boot, 2011). So far few studies demonstrated the grounding of abstract concepts on sensory-motor experiences (for reviews, see Borghi \& Cimatti, 2012; Pecher, Boot, \& van Dantzig, 2011) and thus our study, by focusing on an interesting example of abstract concepts such as numbers, can be relevant for this issue. In fact, the influence of whole body movements on numbers representation can be interpreted as a proof that also number processing can be embodied. Interestingly, our findings are in line with a recent proposal, advanced by Fischer and Brugger (2011) on the origin of the Spatial-numerical associations (SNAs), that recognizes the grounded and embodied nature of numerical cognition, which would origin from finger counting.

To summarize, our study contributes to highlight the close link and interaction between our everyday activities, as movements in real-life situations, and higher-order cognitive processes, as spatial representation and number processing.

\section*{Acknowledgments}

Many thanks to the members of the EMCO group (www.emco.unibo.it) and HuCog group (http://www.hucog.unimore.it/site/home.html) for helpful comments and discussions. We would also like to thank Prof. Roberto Bolzani for a significant help in data analysis and results discussion.

This work was supported by grant (PRIN 2008) from MIUR (Ministero Italiano dell'Istruzione, dell'Università e della Ricerca) and by the European Community, FP7 project ROSSI (www.rossiproject.eu), Emergence of Communication in Robots through Sensorimotor and Social Interaction.

\section*{References}

Andres, M., Ostry, D.J., Nicol, F., \& Paus, T. (2008). Time course of number magnitude interference during grasping. Cortex, 44, 414-419.

Badets, A., Andres, M., Di Luca, S., \& Pesenti, M. (2007). Number magnitude potentiates action
judgment. Experimental Brain Research, 180, 525534.

Badets, A., Bouquet, C.A., Ric, F., \& Pesenti, M. (2012). Number generation bias after action observation. Experimental Brain Research, 221, 4349.

Badets, A., \& Pesenti, M. (2010). Creating number semantics through finger movement perception. Cognition, 115, 46-53.

Badets, A., \& Pesenti, M. (2011). Finger-number interaction. An ideomotor account. Experimental Psychology, 58, 287-292.

Barsalou, L.W. (2008). Grounded cognition. Annual Review of Psychology, 59, 617-645.

Borghi, A.M., \& Cimatti, F. (2012). Words are not just words: the social acquisition of abstract words. RIFL - ISSN: 2036-6728. doi: . 10.4396/20120303.

Borghi, A.M., \& Pecher, D. (2011). Introduction to the special topic Embodied and Grounded Cognition. Frontiers in Psychology, 2, 187. doi: 10.3389/fpsyg.2011.00187.

Chiou, R.Y.-C., Chang, E.C., Tzeng, O.J.-L., \& Wu, D.H. (2009). The common magnitude code underlying numerical and size processing for action but not for perception. Experimental Brain Research, 194, 553-562.

Dehaene, S., Bossini, S., \& Giraux, P. (1993). The mental representation of parity and number magnitude. Journal of Experimental Psychology: General, 122, 371-396.

Fischer, M.H. (2012). A hierarchical view of grounded, embodied, and situated numerical cognition. Cognitive Processing, 13, 161-164.

Fischer, M.H., \& Brugger, P. (2011). When digits help digits: Spatial-numerical associations point to finger counting as prime example of embodied cognition. Frontiers in Psychology, 2, 260. doi: 10.3389/fpsyg.2011.00260.

Fischer, M.H., Castel, A.D., Dodd, M.D., \& Pratt, J. (2003). Perceiving numbers causes spatial shifts of attention. Nature Neuroscience, 6, 555-556.

Gianelli, C., Ranzini, M., Marzocchi M., Rettore Micheli L., \& Borghi, A.M. (2012). Influence of numerical magnitudes on the free choice of an object position. Cognitive Processing, 13,1, 185-188.

Hartman, M., Grabherr, L., \& Mast, F.W. (2011). Moving along the mental number line: Interactions between whole-body motion and numerical cognition. Journal of Experimental Psychology: Human Perception and Performance, 38(6), 141627.

Hartmann, M., Farkas, R, \& Mast, F.W. (2012). Selfmotion perception influences number processing: Evidence for a parity task. Cognitive Processing, 13, 189-192.

Holmes, K.J., \& Lourenco, S.F. (2012). Orienting numbers in mental space: Horizontal organization trumps vertical. The Quarterly Journal of Experimental Psychology, 65, 1044-1051.

Hubbard, E.M., Piazza, M., Pinel, P., \& Dehaene, S. (2005). Interactions between number and space in the parietal cortex. Nature Reviews Neuroscience, 6, 435-448.

Knops, A., Viarouge, A., \& Dehaene, S. (2009). Dynamic representations underlying symbolic and nonsymbolic calculation: Evidence from the operational momentum effect. Attention, Perception, \& Psychophysics, 71(4), 803-821.

Nicholls, M.E.R., Loftus, A.M., \& Gevers, W. (2008). Look, no hands: A perceptual task shows that number magnitude induces shift of attention. Psychonomic Bullettin \& Review, 15, 413-418.

Pecher, D., \& Boot, I. (2011). Numbers in space: Differences between concrete and abstract situations. Frontiers in Psychology, 2, 121. doi: 10.3389/fpsyg.2011.00121.

Pecher D., Boot I., \& van Dantzig S. (2011) Abstract concepts: Sensory-motor grounding, metaphors, and beyond. In: Ross B, editor. The Psychology of Learning and Motivation, Burlington: Academic Press. vol. 54, pp. 217-248.

Ranzini, M., Lugli, L., Anelli, F., Carbone, R., Nicoletti, R., \& Borghi, A.M.(2012). Graspable objects shape number processing. Frontiers in Human Neuroscience, 5, 147.

\title{
Won't You Think of the Children?: Traits Predicting Intergenerational Preferences
}

\author{
Christian C. Luhmann (christian.luhmann@stonybrook.edu) \\ Department of Psychology, Stony Brook University \\ Stony Brook, NY 11790 USA \\ Sarah S. Pak (sarah.pak@ princeton.edu) \\ Department of Psychology, Princeton University \\ Princeton, NY 08540 USA
}

\begin{abstract}
Individuals are frequently asked to make sacrifices in an attempt to produce benefits for future generations. Such decisions are referred to as intergenerational dilemmas. Previous research on intergenerational dilemmas has shown that situational manipulation of factors such as the delay between sacrifice and benefits and the perceived similarity with future others modulate intergenerational preferences. However, it is unclear whether there are traits that predict intergenerational preferences across a variety of dilemmas. Individual differences were quantified using econometric measures of delay discounting and social discounting. Results indicated that individual differences on these measures accounted for a significant portion of the variance observed in a broad measure of intergenerational preferences.
\end{abstract}

Keywords: intergenerational choice, delay discounting, social discounting

It is increasingly clear that many of our everyday actions not only have immediate consequences, but also have consequences for those in future generations. Decisions about such actions are somewhat peculiar in that a thorough evaluation requires considering the interests of individuals that do not yet exist. These decisions become even more complicated when they require short-term sacrifices on the part of the present generation in order to achieve benefits for (or to avoid harming) future generations. Decisions about such tradeoffs have been referred to as intergenerational dilemmas (e.g., Gardiner, 2006).

Intergenerational dilemmas are frequently encountered in the context of policy-making and involve everything from global warming and overfishing to more mundane decisions about infrastructure investments. Occasionally, policy makers place will act on behalf of future generations. For example, Norway's gasoline prices are among the world's highest at \(\$ 10.12\) per gallon, resulting from taxes on fossil fuels designed to reduce global warming (Randall, 2012; Romero, 2005). However, due to a lack of political will, intergenerational dilemmas are often resolved in ways that favor the current generation.

Intergenerational dilemmas are an example of a larger class of social dilemmas in which the interests of the decision maker are at odds with the interests of others. Despite the pervasive assumption of self-interested motives by economists (see Frederick, Loewenstein, \& O’Donoghue,

2002 for a review of this literature), altruistic behavior has been observed in a range of contexts and across a variety of species (Piliavin \& Charng, 1990). The dominant models suggest that altruism is highly dependent on reciprocity. That is, decision makers may act in order to achieve benefits for others, even at personal cost, if the beneficiaries may later return the favor. In the context of intergenerational dilemmas, however, reciprocity plays no obvious role. That is, decision makers in the present generation have no reason to behave altruistically towards future generations because future generations have neither an opportunity to reciprocate beneficial actions nor the means to retaliate for detrimental actions. In fact, according to traditional models of altruism, it is in the present generation's best interest to make decisions that ignore the welfare of future generations.

\section*{An Economic Perspective}

When contemplating intergenerational dilemmas at the policy level, economists are typically employed to produce cost-benefit analyses that are used to guide policy-making. When a policy's consequences (either costs or benefits) extend over long time periods, these analyses are forced to specify exactly how current and future welfare are balanced. The strategy typically taken is to take the interests of future generations into consideration, but to a lesser extent than the interests of the current generation. The degree to which future consequences impact intergenerational choices is controlled by what is known as the social discount rate (Moore, Boardman, Vining, Weimer, \& Greenberg, 2004). Assuming a social discount rate of ten percent, immediate consequences are considered to be twice as important as identical consequences that will occur in seven years and ten times greater than consequences that will occur in 14 years. For example, imagine a proposal to fix aging sewer systems. If the proposed legislation would cost \(\$ 100\) million dollars immediately, but would avert a potentially costly failure estimated to occur in 30 years, then the ultimate cost of the failure would have to exceed \(\$ 2\) billion in order to justify the immediate expenditure.

These social discount rates can lead to potentially undesirable conclusions. For example, the ten percent discount rate suggests that the welfare of the current generation's grandchildren (two generations or 50 years from now) will be valued at less than one percent of the
welfare of the current generation. Thus, these standard discount rates suggest that we ought to essentially disregard the welfare of future generations and instead act to maximize welfare over a short temporal horizon.

However, policymakers and even economists themselves often disagree about what social discount rates are appropriate. For example, the Stern Review (2007), a comprehensive report assessing the costs of climate change, was criticized by some economists for employing a social discount rate that was too low (e.g., Beckerman \& Hepburn, 2007), whereas others have argued for a lower discount rate (e.g., a discount rate of zero, Cline, 2008).

In contrast, Dasgupta (2007) noted that one could be similarly dissatisfied with the selection of another parameter in Stern's model, what Stern calls eta ( \(\eta\) ). Eta is an ethical parameter that reflects people's attitude about disparities in welfare both between individuals within the current population and disparities in welfare between current and future populations. Dasgupta (2007) argues that society ought to have more egalitarian attitudes than implied by the value of eta Stern selected.

\section*{A Psychological Perspective}

Psychologists have also investigated factors related to intergenerational preferences (though to a far more limited degree). Wade-Benzoni (2008), for example, has compiled a set of factors that appear to modulate intergenerational preferences. These factors include the delay and uncertainty associated with future consequences, affinity towards future generations, and the behavior of past generations. Both individually and taken together, these factors appear to predict, to varying degrees, one's intergenerational choices.

One intuitive influence on intergenerational preferences is the delay between the current generation's behavior and the associated consequences. This mirrors the economic idea of a social discount rate reviewed above. That is, consequences expected to occur only after long delays are discounted more (i.e., exert less of an influence on intergenerational choices) than those expected to occur after shorter delays. For example, Wade-Benzoni (2008) found that participants who were told that future generations would begin to reap the benefits of a proposed gas tax in the relatively near future were willing to bear significantly higher gas taxes than those who believed the benefits were more temporally delayed.

Preferences in intergenerational dilemmas have also been found to depend on the affinity between decision-makers and the recipients of future benefits. For example, WadeBenzoni (2008) asked office staffers to distribute a sum of money between themselves and a future subject in the study. Results indicated that participants left significantly larger sums of money if they believed they were leaving money for a fellow staff member (high affinity) than when they believed they were leaving money for a stranger (low affinity). This is in line with past findings (e.g., Hoffman,

McCabe, \& Smith, 1996) that social distance acts to attenuate generosity toward others.

Lastly, intergenerational preferences appear to be influenced by the behavior of past generations. As described above, intergenerational generosity (or greed) cannot generally be reciprocated. However, there is recent evidence that individuals unable to reciprocate may "pay forward" past acts on unrelated third parties (Gray, Ward, \& Norton, in press). In the case of intergenerational dilemmas, this would suggest that individuals might attempt to "reciprocate" the actions of previous generations, but to do so with future generations. That is, if previous generations have sacrificed on our behalf, then perhaps we may be more willing to do so on behalf of future generations. Consistent with this suggestion, Wade-Benzoni (2002) found that intergenerational precedents can exert a strong influence on intergenerational choices, but only when individuals believed that previous generations were willing to make sacrifices. Apparently, previous generations' generosity serves as a model for the current generation in a way that previous generations' selfishness does not.
This previous work has pointed to several major factors that influence intergenerational preferences. However, these previous studies have focused on the manipulation of situational factors. For example, beliefs about the consequences of overfishing predicted were related to willingness to accept fishing quotas (Wade-Benzoni, 2008). Furthermore, because factors such as delay, uncertainty, and precedent should vary from one intergeneration dilemma to another, one would also expect people's intergenerational preferences to vary from one dilemma to another as well.
The goal of the current study is to explore how intergenerational preferences may be predicted by decisionrelated traits that are relatively stable across decision making contexts. That is, the current study concerns our ability to predict individual differences in intergenerational preferences. Because of our focus on individual differences, we evaluate decision-relevant traits using measures that are both quantitatively rigorous and that generalize across a variety of contexts.
Specifically, we evaluate decision makers' preferences regarding delay and social distance because past work has found situational manipulation of these factors to modulate intergenerational preferences. To evaluate preferences about delay, we use a standard delay discounting task (Kirby \& Marakovic, 1996). Delay discounting refers to the tendency for immediate rewards to be preferred over delayed rewards and for the value of rewards to decline with increasing delay. To evaluate social preferences, we employ a recently developed measure of social discounting (Rachlin, 2002). Similar to delay discounting, work on social discounting has found that rewards to the self are preferred over rewards given to others and that the subjective value of others' rewards declines as social distance increases.

\section*{Method}

\section*{Participants}

Sixty-three Stony Brook University undergraduate students participated for partial course credit.

\section*{Dependent Measures}

The intergenerational decision-making task consisted of four scenarios. Specifically, we adapted items involving the topics of overharvesting fisheries and a gasoline tax (based on materials from Wade-Benzoni, 2008). Two additional scenarios were developed specifically for the current study, one involving an increase in tuition and one involving an increase in rent. These items were included to increase the relevance of the intergenerational dilemmas to our undergraduate participants. Each scenario embodied the same basic set of features characteristic of a standard intergenerational dilemma. That is, the scenarios each described an immediate, costly sacrifice and stated that the benefits of this sacrifice would only be enjoyed by other individuals (but not by the participant) and that the benefits would only arrive at some point in the future. As in WadeBenzoni's (2008) study, each scenario involved reading a brief passage that provided factual information about the issue, including short-term costs and future benefits, and included a graphical visualization of the relationship between the magnitude of the short-term sacrifice and the corresponding benefits to future generations. The graph did not include numbers of any sort and was not intended to be thoroughly informative. Instead, it was intended to simply illustrate the idea that greater present sacrifice would yield greater future benefit. Participants were then asked to indicate whether they would agree with a series of proposed sacrifices. For example, participants were asked if they would agree to pay an additional \(\$ 0.20\) tax, raising the price of a gallon of gasoline to \(\$ 3.20\).

The delay discounting task was adapted from a previous study by Kirby and Marakovic (1996). On each trial of this task, participants chose between a smaller reward, which was available immediately (i.e., "tonight"), and a larger reward, which was only available after some delay. For example, one item asked participants to select between \$30 dollars tonight and \(\$ 85\) to be delivered in 70 days. Each of these items is associated with a discount rate that represents how patient a decision maker would need to be in order to be indifferent between the immediate and delayed rewards. For example, in the preceding example, indifference would be associated with a discount rate of exactly .008571 . The task consisted of 27 items. Twenty-one of these items were identical to those used by Kirby and Marakovic (1996), capable of detecting discount rates from 0.0007 to 0.25 . In our experience, we have found that undergraduates' preferences fall toward the impatient end of this range. To ensure that we did not artificially exclude particularly impatient participants, we amended the original 21 items with 6 additional items that extended the range of measurable discounting rates from 0.0007 to 1.0 , a change
we have adopted in previous investigations (Luhmann, in press).

The social discounting task was adapted from Jones and Rachlin (2006). Participants were first asked to imagine 100 people, ranging from one's closest friend or relative (i.e., person \#1) to a mere acquaintance (person \#100). On each trial, participants were asked to choose between a reward for themselves and a reward for someone on their list of 100 people. For example, one item asked participant to choose either \(\$ 30\) dollars for themselves and \(\$ 85\) for Person \#70. The specific quantities were identical to those used in the delay discounting task. That is, the reward magnitudes were identical and the delays (e.g., 70 days) were converted into social distances (e.g., person \#70).

\section*{Procedure}

Before the experiment began, instructions were read to participants. Participants were told that they would complete a number of measures on their preferences on a variety of topics. The instructions further emphasized that there were no "correct answers". The order for both the scenarios within the intergenerational decision task and the three measures themselves were counterbalanced across participants. The entire procedure took approximately 30 minutes.

\section*{Results}

Each participant's delay discount rate was estimated as the discount rate most consistent with her choices. For example, if a participant chose the larger reward for all items representing discount rates equal to and smaller than .01 but chose the smaller reward for all items equal to and larger than .02 , her discount rate would be estimated as the geometric mean of the discount rates associated with the items on each side of this "switch point" (.014 in this case). If more than one discount rate was found to be equally consistent with a set of choices, the geometric mean of the consistent estimates was taken to be the participants' discount rate (for further details, see Kirby \& Marakovic, 1996). The procedure for estimating of social discount rates was identical. Because discount rates are highly skewed, they were transformed by taking their natural log before being submitted to the statistical analyses described below.

To quantify participants' intergenerational preferences, we first estimated the maximum sacrifice that each participant would accept in each scenario. This maximum was estimated using a procedure similar to that used to estimate the discount rates. For example, if a participant agreed to all the sacrifices equal to and smaller than \$300 but rejected all sacrifices equal to and larger than \(\$ 400\), her maximum willingness was estimated as the mean of these two "cross-over" quantities (e.g. \$350). If more than one estimate was found to be equally consistent with a set of choices, the mean of the most consistent estimates was taken. These estimates were then normalized by computing z-scores. This yielded a total of six z-scores for each participant, one for each scenario. Finally, each
\begin{tabular}{ccccc}
\hline Variable & Coefficient & SE & \(\boldsymbol{t}\) & \(\boldsymbol{p}\) \\
\hline Intercept & -1.59 & 0.574 & -2.77 & 0.007 \\
Delay Discounting & -0.79 & 0.351 & -2.24 & 0.028 \\
Social Discounting & -1.06 & 0.404 & -2.61 & 0.011 \\
Delay * Social Discounting & -0.54 & 0.246 & -2.18 & 0.033 \\
\hline
\end{tabular}

Note: Overall \(R^{2}=0.1264\) ( \(\mathrm{p}<.005\) )
participant's six z-scores were averaged. These averages represent participants' intergenerational preferences: their average, relative willingness to sacrifice on behalf of future generations.

We constructed a multiple regression model with the social and delay discount rates acting as predictor variables and the intergenerational preference measure acting as the outcome variable (Table 1). Results demonstrated that this model accounted for a significant proportion of the variance in intergenerational choices. Turning to the individual factors (see Figure 1), results indicate that participants' delay discount rates significantly predicted their intergenerational choices, with lower delay discount rates (i.e., greater patience) predicting greater willingness to sacrifice on behalf of future generations. Social discount rates were also a significant predictor of intergenerational preferences, with lower social discount rates (i.e., greater generosity) predicting greater willingness to sacrifice on behalf of future generations. Finally, results indicate that the interaction between social discounting and delay discounting was also a significant predictor of intergenerational preferences. Specifically, the direction of this relationship suggests that delay and social discounting combined super-additively to predict intergenerational choices. That is, a decision maker who was both patient and generous was even more willing to sacrifice on behalf of future generations than would have been expected by her individual delay and social discount rates.

\section*{Discussion}

The goal of the current study was to investigate intergenerational preferences; the willingness to make sacrifices on behalf of future generations. Whereas prior studies have focused on situational factors that influence intergenerational decision making, we have instead investigated how individuals' decision-related traits might predict their intergenerational preferences. Our results suggest that both delay discount rates (i.e., patience) and
social discount rates (i.e., generosity) were significant predictors of intergenerational preferences. Individuals displaying greater patience when choosing between personal rewards were also significantly more willing to make sacrifices for the benefit of future generations. Similarly, individuals who made more generous choices, more frequently preferring rewards to others at personal cost, were also more inclined to make intergenerational sacrifices. Finally, we also found that individuals who displayed both greater generosity and greater patience were even more willing to sacrifice on behalf of future generations than would be expected given these individual traits.

The finding that individuals' generosity predicts their intergenerational preferences is consistent with previous research. For example, Jones and Rachlin (2009) have demonstrated that social discount rates (but not delay discount rates) predict altruistic behavior in public goods games, which is a multi-player version of the classic Prisoner's Dilemma game (Axelrod, 1984). Those with low social discount rates (i.e., high generosity) have been found to be more cooperative than those with high social discounting rates. Jones and Rachlin (2009) suggest that when assessing the tradeoff between personal rewards and rewards to others, the latter is necessarily discounted according to the social distance between the decision maker themselves and the others. Wade-Benzoni (2008) has also reported that individuals are more willing to make intergenerational sacrifices on behalf of similar others than dissimilar others. Wade-Benzoni (2008) refers to this dimension as affinity, but it is roughly equivalent to social distance, particularly as it has been conceived by Trope and colleagues (Trope \& Liberman, 2011).

The finding that individuals’ patience predicts their intergenerational preferences is somewhat more curious. Of course, individuals' distaste for delayed payoffs is a robust finding (Soman et al., 2005). Indeed, delay has been found to systematically devalue rewards. Wade-Benzoni (2008) has reported a related finding in which intergenerational


Figure 1 - Partial residual plots illustrating the relationship between intergenerational preferences and delay discounting (A), social discounting (B), and the interaction between delay and social discounting (C). Discount rates have been log-transformed.
preferences were found to be more generous when the benefit to future generations was described as occurring sooner rather than later. Given that intergenerational benefits only arrive in the distant future, it may seem reasonable that patience should be associated with greater intergenerational discounting. However, delay discount rates are typically assumed to describe attitudes toward the delay associated with personal rewards, not the rewards of others. Given that intergenerational tradeoffs are between the current self and future others, it is not immediately obvious why one's evaluation of one's own future rewards is particularly relevant. However, if one is attempting to assess the magnitude of the intergenerational benefits, it may not be possible to perform this evaluation without one's own intertemporal attitudes influencing the valuation. Alternatively, some researchers have suggested that intertemporal attitudes may reflect one's beliefs about the uncertainty present in the environment (Bixter \& Luhmann, 2012). That is, it is permissible, and even advisable, to be impatient if it is believed that future rewards are unlikely to be delivered as promised. Under this view, patience is not about one's unique attitudes toward personal rewards, but about the risk associated with waiting; risk that everyone faces. Consistent with this account, Wade-Benzoni (2008) reported that intergenerational preferences were more selfish when the future benefits were associated with greater uncertainty.

The interaction between patience and generosity is interesting and may be a natural extension of the reasoning outlined above. Because intergenerational sacrifices are made so as to bring about benefits that are both temporally and socially distant, it makes sense that these two factors might jointly influence intergenerational preferences. Indeed, the way in which discounting is typically formulated suggests that rewards are reduced by a discount factor that combines both the magnitude of the dimension (e.g., delay) and the decision maker's attitude toward that dimension (e.g., patience). When rewards are discounted along more than one dimension, these discount factors are combined multiplicatively (e.g., Ho, Mobini, Chiang, Bradshaw, \& Szabadi, 1999), which naturally predicts an interaction and more specifically suggests that patience and generosity should combine super-additively.

It is also interesting to note that our results found patience and generosity to be independent because prior work (Jones \& Rachlin, 2009) has found these traits to be significantly correlated. The correlations reported by Jones and Rachlin were not overwhelming ( \(r=.25-.28\) ), so it is possible that we did not have sufficient power to detect this relationship. However, the predictive power each factor provided in our multiple regression analysis suggests that this may not be a plausible explanation. It is perhaps even more surprising that we failed to find any overlap between patience and generosity because our study assessed these dimensions using nearly identical tasks (e.g., identical rewards and distances). If participants were not paying close attention to the materials and simply making choices based on the
numbers presented on each trial, their choices should have been identical. This suggests that the independence of delay and social discounting is even stronger evidence for separable traits.

The current results suggest implications for policymaking. Specifically, the current study suggests that social discount rates should reflect a variety of decision-related psychological attitudes. For example, one cannot simply assume that the delay between immediate costs and future benefits captures the entirety of the current generation's attitudes toward future generations. Echoing the concerns of Dasgupta (2007), more general attitudes about equality, fairness, and generosity appear to be just as powerful in determining individuals' intergenerational preferences. Though this may complicate the calculation of social discount rates, attitudes about social equality are arguably easier to contemplate because they can be evaluated intragenerationally.

The notion that intergenerational preferences are a unique blend of intertemporal and social preferences also has implications for those seeking to encourage intergenerational sacrifice. For example, policymakers can make efforts to deemphasize the delay until intergenerational benefits will be achieved (thereby mollifying impatience) and deemphasize the social differences between current and future generations (encouraging greater generosity). Indeed, because of the interactivity between these factors, policymakers accomplishing both of these goals simultaneously would be expected to get an extra "boost" of selflessness toward future generations.

\section*{References}

Axelrod, R. (1984). The Evolution of Cooperation. New York: Basic Books.
Beckerman, W., \& Hepburn, C. (2007). Ethics of the discount rate in the Stern review on the economics of climate change. World Economics, 8(1), 187-201.
Bixter, M. T. \& Luhmann, C. C. (2012). Evidence for implicit risk: Delay information primes the processing of uncertainty information. Manuscript submitted for publication.
Cline, W. (2008, January). Comments on the Stern Review. Remarks at the Roundtable Discussion on the Stern Review of the Economics of Climate Change and Its Critics, American Economic Association, Annual Meeting, New Orleans. Retrieved from: http://www.iie.com/publications/papers/paper.cfm?Resear chID=874
Dasgupta, P. (2007). Commentary: The Stern review's economics of climate change. National Institute of Economic Review, 199(4), 4-7.
Frederick, S., Loewenstein, G., \& O’Donoghue, T. (2002). Time discounting and time preference: A critical review. Journal of Economic Literature, 40(2), 351-401.

Gardiner, S. M. (2006). A perfect moral storm: Climate change, intergenerational ethics and the problem of moral corruption. Environmental Values, 15, 397-413.
Gray, K., Ward, A. F., Norton, W. I. (in press). Paying it forward: Generalized reciprocity and the limits of generosity, Journal of Experimental Psychology: General.
Ho, M. Y., Mobini, S., Chiang, T. J., Bradshaw, C. M., \& Szabadi, E. (1999). Theory and method in the quantitative analysis of 'impulsive choice' behaviour: Implications for psychopharmacology. Psychopharmacology, 146, 362372.

Hoffman, E., McCabe, K., \& Smith, V. L. (1996). Social distance and other-regarding behavior in dictator games. The American Economic Review, 86, 653-660.
Jones, B.A., \& Rachlin, H. (2009). Delay, probability, and social discounting in a public goods game. Journal of The Experimental Analytical Behavior, 91(1), 61-73.
Jones, B. A. \& Rachlin, H. (2006). Social discounting. Psychological Science, 17(4), 283-286.
Kirby, K. N., \& Marakovic, N. N. (1996). Delaydiscounting probabilistic rewards: Rates decrease as amounts increase. Psychonomic Bulletin \& Review, 3(1), 100-104.
Luhmann, C. C. (in press). Discounting of Delayed Rewards is Not Hyperbolic, Journal of Experimental Psychology: Learning, Memory, \& Cognition.
Moore, M. A., Boardman, A. E., Vining, A. R., Weimer, D. L., \& Greenberg, D. H. (2004). "Just give me a number!" Practical values for the social discount rate. Journal of Policy Analysis and Management, 23(4), 789-812.
Piliavin, J. A. \& Charng, H. W. Altruism: A review of recent theory and research. Annual Review of Sociology, 16(1), 27-65.
Rachlin, H. (2002). Altruism and selfishness. Behavioral Brain Science, 25(2), 239-250.
Rachlin, H., \& Jones, B.A. (2007). Social and delay discounting. Journal of Behavioral Decision Making, 20(1), 1-15.
Randall, T. (2012, August 13). Highest \& cheapest gas prices by country. Bloomberg News. Retrieved from http://www.bloomberg.com/slideshow/2012-08-13/highest-cheapest-gas-prices-by-country.html.
Romero, S. (2005, April 30). The \$6.66-a-Gallon Solution. The New York Times. Retrieved from http://www.nytimes.com/2005/04/30/business/worldbusin ess/30norway.html
Stern, N. (2007). The Economics of Climate Change: The Stern Review. Cambridge: Cambridge University Press.
Soman, D. G., Ainslie, Frederick, S., Li., X., Lynch, J., Moreau, P., Mitchell, A., Read, D., Sawyer, A., Trope, Y., Wertenbroch, K., \& Zauberman, G. (2005). The psychology of intertemporal discounting: Why are distant events valued differently from proximal ones? Marketing Letters, 16(3, 4), 347-360.

Trope, Y., \& Liberman, N. (2011). Construal-level theory of psychological distance. Psychological Review, 117(2), 440-463.
Wade-Benzoni, K.A. (2002). A golden rule over time: Reciprocity in intergenerational allocation decisions. The Academy of Management Journal, 45(5),1011-1028
Wade-Benzoni, K. A. (2008). Maple trees and weeping willows: The role of time, uncertainty, and affinity in intergenerational decisions. Negotiation and Conflict Management Research, 1(3), 220-245.
Wade-Benzoni, K.A., Hernandez, M., Medvec, V., \& Messick, D. (2008). In fairness to future generations: The role of egocentrism, uncertainty, power, and stewardship in judgments of intergenerational allocations. Journal of Experimental Social Psychology, 44(1), 233-245.
Wade-Benzoni, K.A., \& Tost, L. P. (2009). The egoism and altruism of intergenerational behavior. Personality and Social Psychology Review, 13(3), 165-193.

\title{
Tell Us Your Story: Investigating the Linguistic Features of Trauma Narrative
}

\author{
Jeremy A. Luno (jaluno@memphis.edu) \\ Department of Psychology / Institute for Intelligent Systems, University of Memphis 400 Innovation Drive, Memphis, TN 38152 USA
}
J. Gayle Beck (jgbeck@memphis.edu)

Department of Psychology / University of Memphis
202 Psychology Building, Memphis, TN 38152 USA

\author{
Max Louwerse (maxlouwerse@gmail.com) \\ Department of Psychology / Institute for Intelligent Systems, University of Memphis 400 Innovation Drive, Memphis, TN 38152 USA \\ Tilburg Centre for Cognition and Communication (TiCC), Tilburg University \\ PO Box 90153, 5000 LE, Tilburg, The Netherlands
}

\begin{abstract}
Linguistic features can predict several aspects of human behavior. Little is known, however, about whether syntactic, semantic and structural language features can also predict psychological disorders such as posttraumatic stress disorder (PTSD). The current study investigated whether the linguistic properties in trauma narratives written by survivors of a Motor Vehicle Accident (MVA), change as function of the intensity of PTSD symptoms. A short form diagnostic tool known as the Posttraumatic Stress Disorder Checklist (PCL) was used to determine the severity of participant PTSD symptomatology. Using a text capture paradigm participants were then asked to write a neutral narrative or a narrative that described their traumatic event. PCL scores were compared to linguistic variables from eight different computational linguistic algorithms. Results from this study suggested that the relative intensity of PTSD symptomatology affects syntactic, semantic, and structural aspects of the narrative.
\end{abstract}

Keywords: Posttraumatic Stress Disorder; PTSD; Trauma Narrative; Linguistic Features.

\section*{Introduction}

Language patterns can be good predictors of relations in the world. For instance, language statistics can predict the modality of a word (Louwerse \& Connell, 2011), the iconic relationship of words (Louwerse, 2008), social networks (Hutchinson, Datla, \& Louwerse, 2011), and even geographical locations of cities (Louwerse \& Benesh, 2012; Louwerse \& Zwaan, 2009). Language patterns have also shown to be predictors of aspects of human behavior. For instance, linguistic features predict fraudulent events (Louwerse, Lin, \& Semin, 2010), predict an individual's personality type (Gill, Nowson, \& Oberlander, 2009), whether they are lying (Hancock, 2004), and even to what extent they visit their doctor's office (Campbell \& Pennebaker, 2003).

Despite the predictive utility of language, there is very little computational linguistic research that investigates whether language features predict psychological symptoms such as those associated with Posttraumatic Stress Disorder
(PTSD). Considering the ability of linguistic patterns to demonstrate mental processes and behaviors, and the concise delineations of PTSD symptomatology, the question can be raised whether linguistic patterns in written narratives from trauma survivors can be related to the severity of the trauma. More specifically, can language use reflect PTSD symptoms and their relative intensity? This research question was investigated in the current study.

PTSD is an anxiety disorder diagnosed to persons having "experienced, witnessed, or having been confronted with events that involve potential death, serious injury, or a threat to the physical integrity of oneself or others" (DSM-IV, 2000, p. 467). PTSD patients will persistently re-experience the event, while simultaneously avoiding thoughts, and/or environmental reminders of the event, with some or all of these symptoms lasting for longer than one month. Specific symptoms include re-experiencing the traumatic event, avoiding thoughts of the event, mental/emotional numbing, as well as hyper-arousal. These symptoms are further delineated to include flashbacks, nightmares, sleep difficulties, and irritability. These symptoms are clustered into three overarching categories, "Re-experiencing Symptoms", "Avoidance Symptoms", and "Hyperarousal Symptoms (DSM-IV, 2000), even though not all symptoms are necessary for a PTSD diagnosis.

In recent years the prevalence of PTSD has increased. Studies indicate that \(20 \%\) of women and \(9 \%\) of men will develop PTSD, while \(6.8 \%\) of those diagnosed will live with the disorder indefinitely (Kessler et al., 1995). This increase is due not only to an awareness of the disorder in the clinical community, but also to the development of diagnostic tools such as the PTSD Checklist (PCL). The PCL is a 17 -item self-report measure that monitors trauma symptomatology much like the 30 -item structured interview, Clinician Administered PTSD Scale (CAPS). Despite the CAPS being the longstanding method to PTSD diagnosis, current research validates that the PCL correlates highly with the CAPS measure as well as with its diagnostic efficiency (Blanchard, Jones-Alexander, Buckley, \& Fomeris, 1996).

For instance, one study demonstrated the reliability of the PCL even within highly specific civilian populations (e.g. college students, Motor Vehicle Accident survivors) (Elhai, Gray, Docherty, Kashdan, \& Kose, 2007). These direct questions administered in both the CAPS and PCL have proven to be a valid way of determining PTSD (Blanchard et al., 1996). However, it would be desirable to have an alternative, perhaps less direct, measure that reveals PTSD symptoms identified by the CAPS and PCL. This would allow for patients to be prescreened. Given the evidence from computational linguistic measures predicting human behavior, it might be the case that the language used by PTSD patients predicts the severity of the disorder.

There are indications that language use in PTSD patients might be indicative of the disorder. For instance, one main issue impeding recovery from the disorder is that sufferers have difficulty mentally integrating the event into their current cognitive schemas (Dalgleish, 2004). Traumatic experience is mentally represented by two constructs, situationally accessible memories (SAMs), and verbally accessible memories (VAMs), but only VAMs can be deliberately retrieved; SAMs are activated by situation dependent reminders of the event (Brewin et al., 1996). Ironically, even though VAMs are memory representations that are readily accessible, PTSD patients will still report confusion of the details, as well as difficulty in forming coherent accounts of the traumatic event (Ehlers, Ehring, \& Kleim, 2012). The language of a PTSD patient might reflect this confusion, or reflect their lack of clarity in recalling the details of the traumatic event.

Language use of trauma survivors has been studied before (i.e. therapeutic measures and interventions) (Sloan et al., 2012). One such example comes from work conducted by Tausczik and Pennebaker (2010). In their study writing samples collected in temporal units (e.g. over an extended time period). Some of these samples were related to a participant's traumatic experience, while others were nonemotional in nature. Tausczik and Pennebaker not only demonstrated the health benefits of narrative production by noticing a decrease in doctor visits for those writing about the traumatic experiences, but they were also able to identify word categories related to depression (e.g. pronouns), seeing similar patterns in health improvements as a function of increases or decreases in the usage of these categories.

Similarly, Smyth, Stone, Hurewitz, and Kaell (1999) investigated the effectiveness of PTSD treatment as a result of narrative production rather than fragmented discourse (i.e., discourse without narrative structure) In their study narrative production showed to alter a trauma survivor's tendency to avoid thoughts, as well as aid their recovery. The implication here is that the mental integration of traumatic events can be facilitated by means of narrative production. Smyth et al. (1999) also showed that writing leads to a reduction in symptoms of patients with chronic illness. The Smyth et al. \((1999 ; 2001)\) findings support those of Tausczik and Pennebaker (2010) by isolating the type of
writing task that produces the greatest benefit; specifically, narrative production vs. fragmented and/or controlled discourse.

Often, the analyses of trauma language include global characteristics. For instance, Mansfield et al. (2010) described the foci of their analysis with terms such as complexity, personal growth, and resolution, which indicate positive change in thinking when reflecting on a difficult event, the ability to see different perspectives and outcomes, or the reconciliation with difficult life experiences. Similarly, Tuval-Mashiach et al. (2004) used terms such as coherence, self-evaluation, and meaning. Clearly, these concepts are not problematic in themselves - interrater reliability avoids that the abstract terms might lead to ambiguity - yet a computational operationalization of such concepts is difficult.

Rather than focusing on the global characteristics of trauma language, our analyses of trauma narratives utilized word-level linguistic models that not only offer a large spectrum of linguistic dimensions, but also provide theoretical grounding in the organization of their categories. Based on the theoretical frameworks these models provide, it is possible to isolate constructs similar to those of the aforementioned studies, while simultaneously revealing constructs previously unconsidered?

Because little computational linguistic work has been done on the analysis of trauma narratives, computational linguistic algorithms to analyze the data covered a wide range of dimensions, including syntactic and semantic algorithms. These algorithms can generally be classified into general structural (e.g., word count), syntactic (e.g., connectives) and semantic (e.g., word choice) dimensions of language, whereby some used a bag-of-word approach (e.g. LIWC), whereas others used a probability approach (MRC), whereas yet others relied on the computation of different factors (e.g., type-token ratio). Eight different algorithms were used, categorized in Figure 1.

Within the syntactic dimension, two algorithms were used, one focusing on general linguistic features, the other on interclausal relationships. For general linguistic features, we used 67 features from the Biber model (1988). These features primarily operate at the word level (e.g., parts-ofspeech) and can be categorized as tense and aspect markers, place and time adverbials, pronouns and pro-verbs, questions, nominal forms, passives, stative forms, subordination features, prepositional phrases, adjectives and adverbs, lexical specificity, lexical classes, modals, specialized verb classes, reduced forms, dispreferred structures, and co-ordinations and negations.

Specific interclausal relationships were captured using Louwerse's (2002) parameterization, including positive additive, (e.g. also, moreover), negative additive (e.g. however, but), positive temporal (e.g. after, before), negative temporal (e.g. until), and causal (e.g. because, so) connectives.

The semantic dimension can be broken down in three subdivisions: psycholinguistic ratings, conceptual


Figure 1. Linguistic Category Distinctions
relations, and comprehensive classifications. Psycholinguistic ratings are computed using the MRC Psycholinguistic Database (Coltheart, 1981), to get ratings on the familiarity, concreteness, imagability and meaningfulness of words.

The conceptual dimension covers three categories: interpersonal, social and emotional language. Interpersonal language use is captured by the linguistic category model (LCM) (Semin \& Fiedler, 1991). The model consists of a classification of interpersonal (transitive) verbs that are used to describe actions or psychological states and adjectives that are employed to characterize persons. This classification gives insight into the meanings of verbs and adjectives that people use when they communicate about actors and their social events. The model makes a distinction between five different categories of interpersonal terms

Social language features were captured by Pennebaker et al.'s (2007) Linguistic Inquiry and Word Count (LIWC). LIWC consists of 63 syntactic (e.g., pronouns) and semantic word categories (e.g., death, family) that focus on semantic aspects of discourse, namely aspects of discourse related to social phenomenon.

Emotional words are captured by the classification proposed by Johnson-Laird and Oatley (1989). These words are classified into two classes, broadly basic emotions (anger, fear, disgust, happiness etc.) and complex emotions (guilt, pity, tenderness etc.). The basic emotions indicate no cognitive load hence they are also called raw emotions, whereas the complex emotions indicate cognitive load.

For the comprehensive category WordNet (Miller \& Fellbaum, 1998) was used, consisting of 150,000 words in 44 base types, including 25 primitive groups for nouns (e.g. time, location, person, etc.), 15 for verbs (e.g. communication, cognition, etc.), 3 groups of adjectives, and 1 group of adverbs. For the structural dimension, discourse features such as type/token ratio and word count were used.

The current study aims to utilize these linguistic categories to analyze written narratives produced by participants who experienced an MVA. The intent is to complement the aforementioned studies on the basis of trauma narrative language, but to include measures that have yet to be utilized in these studies; namely the categories featured here, but also participant scores on the PCL. The PCL scores obtained in our sample population will act as a dependent variable from which the associated linguistic features will be compared. Due to PCL scores resting on a continuum, it is possible that the usage of certain linguistic categories will increase and/or decrease as a function of these scores.

Following the analyses of texts captured in this study, the linguistic patterns found will be used as predictor variables to analyze texts collected in studies conducted by Shipherd and Beck \((1999,2005)\). Two data sets from their studies will be analyzed; texts collected from MVA survivors, as well as texts collected from survivors of sexual trauma. In both data sets, there are samples from trauma survivors suffering from PTSD, as well as individuals exposed to trauma though not suffering from PTSD. All of the subjects included in their studies were evaluated using the CAPS or PCL criteria. This allows for a parallel analysis of all of their texts as well as those collected here. The intent is to not only support the reliability of the predictor variables, but also to confirm or deny the possibility that these variables will dependently fluctuate within the range of PCL scores obtained, as we have demonstrated here.

\section*{Experiment}

The current pilot study utilized linguistic category frequencies to analyze written narratives produced by participants who experienced an MVA. These frequencies were then compared with different levels of participant PTSD symptomatology as measured by the PCL.

Table 1: Six featured linguistic categories.
\begin{tabular}{ll}
\hline \multicolumn{1}{c}{ Category } & \multicolumn{1}{c}{ Examples } \\
\hline Determiners (Biber Model) & these, those, few, many, every, any, much, a, an, the, some, all \\
Death (LIWC Model) & died, dead, dying, fatal, alive, grief, mortal, demise, decease(d) \\
Function (LIWC Model) & definitely, ahead, might, nearly, across, enough, among, been \\
Causal Negation (Connectives Model) & although, nevertheless, unless, provided that \\
Punctuation (Biber Model) & \(. ‘, " / ?!()-\) \\
Word Count & (count of words per document) \\
\hline
\end{tabular}

\section*{Methods}

\section*{Participants}

Forty-three undergraduate students from the University of Memphis ( 31 females) participated in the study for course credit. Participants were prescreened to assess PCL scores and to determine whether they had experienced an MVA. All participants in this study had at some time experienced an MVA, though not all participants suffered from PTSD.

\section*{Procedure}

A \(2 \times 2\) design was employed, both counterbalanced and randomized, where half of the participants first wrote about their MVA, while the others completed the neutral text first. Each condition included both tasks. Both text capture tasks were ten minutes in length. Each task, regardless of order, was partitioned by a ten minute cognitive distractor task to minimize carryover effects. The cognitive distractor task utilized in this experiment was a number-based Sudoku puzzle.

\section*{Results and Discussion}

PCL scores ranged from 17-65 \((M=25.07, S D=9.38)\). Five participants from the study \((10 \%\) of the sample population) had PCL scores high enough to suggest a PTSD diagnosis (composite score of 35 of higher), falling within the range of earlier reported estimates of PTSD prevalence (Kessler et al., 1995).

Mixed effects regression analyses were conducted on the normalized frequencies of the linguistic variables with PCL scores and text type (trauma or neutral) as fixed factors, and condition (neutral and trauma narrative) as a random factor (Baayen, Davidson, \& Bates, 2008). The model was fitted using the restricted maximum likelihood estimation (REML) for the continuous variable (PCL scores). F-test denominator degrees of freedom were estimated using the Kenward-Roger's degrees of freedom adjustment to reduce the chances of Type I error (Littell, Stroup, \& Freund, 2002).

Results suggested a significant relationship between PCL scores and six linguistic variables, four categories from the syntactic dimension, and one from both the semantic and structural dimensions. (See Table 1).

PCL scores were related to the use of punctuation, \(F(1\), 81.09) \(=13.474, p<.001\), such that when PCL scores increased the more punctuations were found. Word Count and PCL score were related, \(F(1,81.029)=4.467, p<.05\), with high PCL scores yielding longer texts. The relevance of these categories might be explained by the fact that the patient is unable to lock in the event's specifics, instead using more words in an attempt to accurately describe the event. The presence of these linguistic units could be a byproduct of suppression, in that they are avoiding the acknowledgement of the traumatic event's specifics.

Another variable related to PCL scores were the frequency of determiners (Biber), \(F(1,82.004)=8.597, p<\) . 01 , with higher PCL scores yielding fewer determiners. Determiners, in written and spoken discourse, are used to add discrete specification to the information being conveyed (Argamon et al., 2003; Biber et al 1998; Mulac \& Lundell 1994). It makes sense that determiners in texts written by a PTSD sufferer would find less use, as the disorder affects an individual's ability to concretize specifics from the event.

Negative causal connectives (e.g., although, nevertheless) showed a positive relation with PCL scores, \(F(1,81.136)=\) \(4.74, p<.05\). The increase of negative causal connectives as a function of higher PCL scores might be explained both by suppression and avoidance. Negative causal connectives imply a causal relation that is negated, perhaps to create a distance to the events described, or reflecting an uncertainty in the claims made in the preceding clause.

In addition, the LIWC semantic category "death" was related to PCL scores, \(F(1,81.127)=7.113, p<.01\), likely explained by the inherent nature of traumatic events regardless of whether the trauma experienced was psychological or physical in nature. Also from LIWC, the category "function" yielded a positive relation with PCL scores, \(F(1,82)=6.911, p=.01\) (e.g. as PCL scores increased, the use of "function" words increased).

Interestingly, emotions categories from the LIWC and Johnson-Laird and Oatley Emotions models did not reach significance in relationship to PCL scores. However, "Negative Emotions" from the LIWC model reached significance in comparison of text types, \(F(1,81.151)=\) 4.955, \(p<.05\), as well as "Basic Emotions" from the Johnson-Laird and Oatley Emotion model, \(F(1,81)=9.742\), \(p=.002\). The findings here imply an emotional foundation
in texts written about the PTSD sufferer's recall of the traumatic event.

\section*{General Discussion}

Previous research has shown that linguistic features can predict a multitude of aspects of human behavior. Whether linguistic features might also be indicative of psychological disorders is however less clear. We investigated whether the linguistic properties in trauma narratives written by survivors of a Motor Vehicle Accident (MVA), change as a function of the intensity of posttraumatic stress disorder (PTSD) symptoms. The severity of participant PTSD symptomatology was compared to linguistic variables from eight different computational algorithms.

It may not be surprising from the quantity of variables used in our analysis that some would demonstrate the hypothesized relationship between PCL scores and language use. However, the results featured here can be grounded in theoretical constructs that are in line with what would be expected from the clinical psychological literature. Just as well, since not all of the associated PTSD symptoms are required for a PTSD diagnosis, it is possible that not all of the symptom clusters would emerge in the texts collected. Of the three main PTSD symptom clusters, the variables revealed in our analyses align most with that of the "Avoidance" distinction. The Avoidance cluster of the PTSD symptom inventory identifies a PTSD sufferer's propensity to avoid thoughts and reminders of the event. This cluster includes the thought suppression behavioral phenomenon often associated with PTSD as well (DSM-IV, 2000).

For example, Chung and Pennebaker (2007) have highlighted the ability of function words to reveal psychological states. Function words are subjectively personal, and have a multifaceted utility in the personalization of discourse. Due to the personal nature of traumatic experience it is notable that function word usage would fluctuate as a result of the severity of the traumatic event. The difficulty in mentally integrating the traumatic experience could be one reason why higher PCL scores would reflect a decreased usage of this category. The PTSD patient is unable to work the experience into their current cognitive schemas. The lack of personalization in the narrative reflects the individual's inability to map the experience into long-term memory.

The deficiency of the determiners category can be explained through "Avoidance" symptoms. Determiners, in written and spoken discourse, are used to add discrete specification to the information being conveyed (Argamon et al., 2003, Mulac \& Lundell 1994; Biber et al 1998). It makes sense that determiners in texts written by a PTSD sufferer would find less use, as the disorder affects an individual's ability to concretize specifics from the event.

An increase in negative causal connectives aligns with the "Avoidance" distinction as well, as though an individual suffering from PTSD second-guesses their statements, demonstrating uncertainty as a product of cognitive
distancing. The word count and punctuation could be explained here as well, as the participant is unable to lock in the specifics, and instead uses more words in an attempt to accurately describe the event. The presence of these linguistic units could be a byproduct of suppression, in that they are avoiding the acknowledgement of the traumatic event's specifics. Or just as well, as suggested earlier, PTSD survivors are unable to recall these specifics. From these relationships there is evidence that a person suffering from PTSD trauma can cognitively distance themselves from the details of the event. And while it does not map well in the "Avoidance" cluster, the revealed presence of words from the semantic category "Death" can be explained by the inherent nature of traumatic events regardless of whether the trauma experienced was psychological or physical in nature.

The emotion categories that were shown to differ across text types can be reasoned to explain a PTSD patient's propensity for emotional numbing. Despite the existence of emotional numbing, emotions are still attached to the experience. In writing about the traumatic experience these emotions are faced and thus resurface. This might explain the relative intensity of emotion word usage when describing the traumatic event and the lack of presence in narratives irrelevant to this experience.

Even though the computational linguistic variables show a relationship with PTSD measures, it is clearly not the case that a direct relationship can be assumed, nor should the findings here be seen as an attempt to replace existing clinical psychology measures of diagnosis. At the same time, the current study is encouraging enough to pursue further analysis that might provide a first filter to identify those at risk of PTSD. The categories revealed here are promising, as they align with crucial aspects of the fragmented nature of a PTSD sufferer's recall of their traumatic experience. As well, from the categories discovered here, it is reasonable to presume these same categories and patterns will appear in the analysis of texts collected during the Shipherd and Beck (1999, 2005) studies.

\section*{References}

American Psychiatric Association. (2000) Diagnostic and statistical manual of mental disorders (Revised \(4^{\text {th }}\) ed.) Washington, DC: Author.
Argamon, S.; Koppel, M.; Fine, J.; and Shimoni, A. 2003. Gender, genre, and writing style in formal written texts. Text 23(3).
Biber, D. (1988). Variation Across Speech and Writing. Cambridge: Cambridge University Press.
Blanchard, Edward B. Jones-Alexander, Jacqueline Buckley, Todd C. Forneris, Catherine A. ; Behaviour Research and Therapy, Vol 34(8), Aug, 1996. pp. 669673.

Campbell, R. S., \& Pennebaker, J. W. (2003). The Secret Life of Pronouns Flexibility in Writing Style and Physical Health. Psychological Science, 14(1), 60-65.

Chung, C. K., \& Pennebaker, J. W. (2007). The psychological function of function words. Social communication: Frontiers of social psychology, 343-359.
Coltheart, M. (1981). The MRC Psycholinguistic database. The Quarterly Journal of Experimental Psychology Section A , 33 (4), 497-505.
Dalgleish, T. T. (2004). What might not have been: An investigation of the nature of counterfactual thinking in survivors of trauma. Psychological Medicine, 34(7), 1215-1225.
Ehlers, A., Ehring, T., \& Kleim, B. (2012). Information Processing in Posttraumatic Stress Disorder. In J. Beck, \& D. Sloan (Eds), The Oxford Handbook of Traumatic Stress Disorders (pp. 191-218). New York, NY: Oxford University Press.
Elhai, J. D., Gray, M. J., Docherty, A. R., Kashdan, T. B., \& Kose, S. (2007). Structural validity of the posttraumatic Stress disorder checklist among college students with A trauma history. Journal of Interpersonal Violence , 22 (11), 1471-1478.

Gill, A. J., Nowson, S., \& Oberlander, J. (2009). What are they blogging about? Personality, topic and motivation in blogs. In E. Adar, M. Hurst, T. Finin, N. S. Glance, N. Nicolov, B. L. Tseng (Eds.): Proceedings of the Third International Conference on Weblogs and Social Media, ICWSM 2009, San Jose, California, USA, The AAAI Press 2009.
Hancock, Jeffrey T. Curry, Lauren E. Goorha, Saurabh Woodworth, Michael ; Discourse Processes, Vol 45(1), Jan-Feb, 2008. pp. 1-23.
Hutchinson, S., Datla, V., \& Louwerse, M. M. (2012). Social networks are encoded in language. In N. Miyake, D. Peebles, \& R. P. Cooper (Eds.), Proceedings of the 34th Annual Conference of the Cognitive Science Society (pp. 491-496). Austin, TX: Cognitive Science Society.
Johnson-Laird, P. N., \& Oatley, K. (1989). The Language of Emotions: An Analysis of a Semantic Field. Cognition and Emotion, 3 (2), 81-123.
Kessler, R. C., Sonnega, A., Bromet, E., Hughes, M., \& Nelson, C. B. (1995). Posttraumatic stress disorder in the National Comorbidity Survey. Archives of general psychiatry, 52(12), 1048.
Louwerse, M. (2002). An analytic and cognitive parameterization of coherence relations. Cognitive Linguistics, 291-315.
Louwerse, M.M. (2008). Embodied representations are encoded in language. Psychonomic Bulletin and Review, 15, 838-844.
Louwerse, M.M. \& Benesh, N. (2012). Representing spatial structure through maps and language: Lord of the Rings encodes the spatial structure of Middle Earth. Cognitive Science.
Louwerse, M.M. \& Connell, L. (2011). A taste of words: Linguistic context and perceptual simulation predict the modality of words. Cognitive Science , 35, 381-398.

Louwerse, M.M., Lin, K., Drescher, A., \& Semin, G. (2010). Linguistic cues predict fraudulent events in a corporate social network. In S. Ohlsson \& R. Catrambone (Eds.), Proceedings of the 32nd Annual Conference of the Cognitive Science Society (pp. 961-966). Austin, TX: Cognitive Science Society.
Louwerse, M.M. \& Zwaan, R.A. (2009). Language encodes geographical information. Cognitive Science, 33, 51-73.
Mansfield, C. D., McLean, K. C., \& Lilgendahl, J. P. (2010). Narrating traumas and transgressions: Links between narrative processing, wisdom, and wellbeing. Narrative Inquiry, 20(2), 246-273.
Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the ACM, 38(11), 39-41.
Mulac, A. \& T. L. Lundell, (1994). Effects of gender-linked language differences in adults' written discourse: Multivariate tests of language effects, Language \& Communication 14(3).
Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., \& Booth, R. J. (2007). The development and psychometric properties of LIWC2007. Austin, TX, LIWC. Net.
Semin, G. R., \& Fiedler, K. (1989). The Cognitive Functions of Linguistic Categories in Describing Persons: Social Cognition and Language. Journal of Personality and Social Psychology, 54 (4), 558-568.
Semin, G. R., \& Fiedler, K. (1991). The Linguistic Category Model, its Bases, Applications and Range. European Review of Social Psychology, 2.
Shipherd, J. C., \& Beck, J. G. (1999). The effects of suppressing trauma-related thoughts on women with raperelated posttraumatic stress disorder. Behavior Research and Therapy, 37, 99-112.
Shipherd, J. C., \& Beck, J. G. (2005). The Role of Thought Suppression in Posstraumatic Stress Disorder. Behavior Therapy, 36, 277-287.
Sloan, Denise M. Marx, Brian P. Bovin, Michelle J. Feinstein, Brian A. Gallagher, Matthew W.; Behaviour Research and Therapy, Vol 50(10), Oct, 2012. pp. 627635.

Smyth, J. M., Stone, A. A., Hurewitz, A., \& Kaell, A. (1999). Effects of writing about stressful experiences on symptom reduction in patients with asthma or rheumatoid arthritis. JAMA: the Journal of the American Medical Association, 281(14), 1304-1309.
Smyth, J., True, N., \& Souto, J. (2001). Effects of Writing about Traumatic Experiences: The Necessity for Narrative Structuring. Journal of Social and Clinical Psychology, 20 (2), 161-172.
Tausczik, Y. R., \& Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods . Journal of Language and Social Psychology, 29 (1), 24-54.
Tuval-Mashiach, R., Freedman, S., Bargai, N., Boker, R., Hadar, H., \& Shalev, A. Y. (2004). Coping with trauma: Narrative and cognitive perspectives. Psychiatry: Interpersonal and Biological Processes, 67(3), 280-293.

\title{
Language Dominance Modulates Cross-language Lexical Interaction in Late Immersed Bilinguals
}

\author{
Barbara C. Malt (barbara.malt @lehigh.edu) \\ Department of Psychology, Lehigh University, Bethlehem, PA 18015, USA \\ Ping Li(pul8@psu.edu) \\ Department of Psychology, Penn State University, University Park, PA 16802, USA \\ Eef Ameel (eef.ameel@psy.kuleuven.be) \\ Department of Psychology, University of Leuven, 3000 Leuven, Belgium \\ Aneta Pavlenko (apavlenk @temple.edu) \\ Department of Teaching and Learning, Temple University, Philadelphia, PA, 19122, USA \\ Huichun Zhu (huichun.zhu@gmail.com) \\ Department of Psychology, Lehigh University, Bethlehem, PA 18015, USA
}

\begin{abstract}
Languages differ in the way they package elements of the world into words, which poses a challenge for bilinguals. We examined word use patterns for common household objects for late-immersed Chinese-English bilinguals to investigate how the bilingual lexical network develops when the first language is fully mature at the time of second-language immersion. We found changes to both first- and secondlanguage word use with increased English dominance, indicating continued plasticity and mutual influence.
\end{abstract}

Keywords: bilingualism; word learning; word use; lexicon; categorization.

\section*{Introduction}

Second-language learning research traditionally examined transfer from the first language (L1) to the second (L2) assuming a stable L1. Separately, language attrition research examined changes to L1 in the face of L2 dominance. Only recently has it been appreciated that L1 and L2 may exert mutual influences, and that performance in each may best be understood by studying their interplay across conditions of learning and use (Schmid \& Köpke, 2007).

Most inquiry from this new perspective has focused on phonology and morpho-syntax. These domains are considered to engage procedural memory and potentially be affected by phenomena such as critical periods for learning. In contrast, the lexicon is considered to be stored in declarative memory, with performance subject to standard memory parameters such as frequency of retrieval (e.g., Ullman, 2004). But appropriate use of words depends on much more than retrieval of word forms. Languages differ in the way they package elements of the world into words. For instance, in English, upholstered seats for one person receive the same name as hard wooden seats for one person (chair), whereas in Mandarin they receive the same name as upholstered seats for several (safa). Even cognates show
differing patterns. In Spanish, a Coke bottle is botella but a baby bottle is mamadera, and a tennis ball is not a bola but a pelota.

These subtle differences can be thought of in terms of a lexical network in which the conceptual level of representation includes features, instances, and associations rather than unitary concept nodes. Associated word forms of the two languages can have different patterns of connection to elements of the conceptual layer (Ameel, Storms, Malt, \& Sloman, 2005; Pavlenko, 2009; Van Hell \& de Groot, 1998). Given this model, influences of one language on the other can be conceived as changes to the weights on connections between word forms and elements of the conceptual layer. When a new L2 word form is taught as, or implicitly assumed to be, a translation equivalent of an L1 word, the network will set initial weights to match those of the L1 word. With experience, the connection weights might be adjusted to more closely match those of native L2 speakers. However, cross-connections between words might cause adjustments to the L1 connection weights as well, shifting them away from those of native L1 speakers.

The theoretical issues raised about bi-directional influence in phonology and morpho-syntax (e.g., Schmid, 2011; Köpke, Schmid, Keijzer, \& Dostert, 2007) are closely echoed for lexical knowledge when framed in these terms. A potential key variable on the degree of mutual influence is age of acquisition. The network's weight configurations may stabilize after mastery of L1 and become resistant to change. If L2 is introduced after the network has stabilized, and if the L2 connection patterns initially reflect those of L1, two consequences may result. First is difficulty adjusting the L2 weights toward the L1 standard. Second is that the L1 will be protected from an influence of L2 -because weights for the L1 words are resistant to change, and also because the L2 weights will diverge little from L1 weights and so have little potential to influence L1 weights. This situation resembles a critical period effect with regard
to acquisition of the L2. Because it also entails protection of the L1 from change, though, it may better be framed in terms of entrenchment of the network. The effects of L1 lexical entrenchment have begun to been tested in connectionist models. Zhao and Li (2010) simulated early versus late bilingual learners and found significant differences between them with regard to the organization of word classes.

A different possibility is that the network may not exhibit stabilization that is resistant to change after the initial L1 learning. In that case, with sufficient input at any time of exposure, the L2 connection weights may be gradually shaped to a close approximation of those of native speakers. This requires that the network be never fully committed to the weight configurations even later in learning. The degree to which a network should be flexible versus committed poses a classic 'stability-plasticity' dilemma in computational modeling (see Li, Farkas, \& MacWhinney, 2004). Under this scenario, because of cross-connections between the L1 and L2 lexicons, the more the L2 weights diverge from initial L1 settings, the greater the impact on L1 usage may be. Conversely, predominant use of the L1 may leave L1 patterns largely intact. This possibility is compatible with suggestions that continued use of L1 protects it against attrition. However, because the more one language is used, the less the other must be, it also implies the trade-off that greater preservation of the L1 patterns will entail lesser progression in the L2.

Yet a third possibility is that if the acquired L1 pattern does not resist change, the network will, under the influence of L2 input, arrive at a configuration for both languages that is a compromise between L1 and L2 patterns. In this case, patterns of usage may not fully match those of monolingual speakers of either language. Ameel et al. (2005) found a convergence of this sort for Belgian early bilinguals who grew up with both French and Dutch. It may be less likely to be found for late bilinguals, having one well-established language before substantial exposure to the second. Conversely, the patterns of the two languages may be functionally separable for late bilinguals, allowing mastery of native -like patterns for both given sufficient exposure to each. This outcome may be most likely under conditions where the languages are dissimilar overall and/or in terms of naming patterns within a domain, yielding weaker crossconnections as L2 learning takes place.

We focus here on late L2 learners to examine plasticity of the network after L1 is well-established. In related work, Malt \& Sloman (2003) found that, for immersed L2 users of English, elements of non-native usage patterns for concrete nouns in the L2 persisted for many years despite evolution toward more native-like usage. On the other hand, Pavlenko and Malt (2011) found evidence for some L2 impact on L1 word use in Russian immigrants to the U.S. who continued to speak Russian at home. These bilinguals treated several L1 Russian terms for drinking vessels as if they were more equivalent to English terms than did largely monolingual speakers in Russia. The L2 influence on L1 was greatest for
those who came to the U.S. in early childhood, but some influence was seen even for those who arrived after age 18. These results point toward limited but continued plasticity of the network. Malt and Sloman's study did not isolate late learners, however, and neither study looked at performance in L1 and L2 in relation to each other, or at outcomes as a function of language dominance or attainment. Malt and Sloman's participants came from many L1 backgrounds, creating variable L1-L2 similarity, whereas Pavlenko and Malt's Russian-English language pairing can be considered to involve languages of intermediate similarity.

The current study investigated naming patterns for common household objects by native speakers of Chinese attending school in the U.S. They named the objects in both English and Chinese, in separate sessions. Participants had arrived no earlier than age 15, thus having a mature L1 at the time of immersion. They varied in the extent to which they had become more English-dominant. Their two languages are dissimilar on many dimensions from syntax to writing systems, and the naming patterns in this domain are dissimilar. We asked three questions that will shed light on fundamental aspects of how bilingual patterns of word use develop for dissimilar languages under conditions of late L2 immersion. These are:
(1) What constrains learning of subtle aspects of \(\mathbf{L} \mathbf{2}\) word use patterns given mature L1 knowledge at time of immersion? Can L2 usage evolve toward native-like patterns as a function of experience, even in light of a mature L1, or will entrenched L1 knowledge defeat reshaping of L2 word usage patterns?
(2) How stable are \(\mathbf{L} 1\) usage patterns when L 2 immersion occurs after L1 is mature? In particular, are highly entrenched L1 patterns immune to an L2 influence, or does the impact vary depending on L2 experience?
(3) If progress in L2 mastery is observed and linked to the extent of L2 dominance, does it have a negative relation to the preservation of the native L1 patterns, or do they vary independently?

\section*{Method}

Sixty-two Lehigh University students, native speakers of Mandarin, participated. All used English on a daily basis. Average age of immersion in English was 21, with a minimum of 15 . Mean self-rated proficiency for English was 4.94 on a 7 -point scale; for Chinese, 6.92. Twenty-five largely monolingual speakers of Mandarin resident in China and 28 largely monolingual speakers of English resident in the U.S. served as comparison groups.

Stimuli for assessing naming patterns consisted of 67 pictures of objects for preparing and serving foods and 73 pictures of objects for holding and dispensing products such as health and beauty aids, cleaners, and foods (see Ameel et al., 2005). For brevity we call the first the dishes set and the second the bottles set, but each contained many objects with other names, as reported below.

Figure 1. Sample pictures from the dishes set.


Figure 2. Sample pictures from the bottles set.


Each set of pictures was presented on a web page. Instructions indicated (in English or Mandarin, depending on test session) that for each picture, they should give whatever name seemed best or most natural, and that their response could be one word or more than one. The photos followed, with each accompanied by a response box into which participants typed their choice of name.

Monolingual speakers of English and of Mandarin each participated in only one experimental session in which they viewed the web pages (with order balanced across participants) and typed in their responses to the pictures.

Bilinguals participated in three sessions. In the first, conducted in English, they filled out an extensive language history questionnaire (in English) that asked for information such as age of exposure to English, years of formal instruction, age of immersion, years of immersion, and other aspects of language experience and usage. They then completed an English word/non-word discrimination task as one measure of English proficiency. The second session was also conducted in English, always by a native speaker, and participants' responses were in English. Participants first completed the naming task for the two stimulus sets (with order balanced across participants). Additional measures of proficiency and current language accessibility were then taken including a speeded picture-naming task and a verbal fluency task in which they were asked to list all the exemplars they could to each of three prompts (Clothing, Transportation, and Food) in 60 seconds each. Last, they told the story depicted in a wordless picture book to the experimenter. The final session was conducted in Mandarin by a native speaker of Mandarin and took place at least one week after the second. The same tasks were completed in the same order, with responses in Mandarin.

\section*{Results}

Monolingual naming patterns. We first tabulated the names produced by monolingual speakers of each language to determine the most common ("dominant") name for each picture. Those names, along with how many objects of the set had each listed name as dominant, are presented in Tables 1 and 2. The tables show that the lexical categories of the two languages do not have a simple relation for either stimulus set. For both, each language has one broad term that covers \(1 / 3\) or more of the objects, but these terms do not correspond closely to one another: Objects labeled by a single term in one language are distributed across several in the other. This is also true for most of the other terms that cover multiple objects of the set. These complex relations pose a challenge for the L2 learner. The neatest mapping across languages is the close correspondence of Mandarin bei to the combined English cup, mug, and glass. However, in this case, an L1 speaker of Mandarin must still learn to segment a broader category into several narrower ones.

For the dishes stimulus set, the number of terms that are dominant for at least one object is similar between the two languages, with 9 for English and 8 for Mandarin. For the bottles set, however, English has 13 compared to Mandarin's 5. The greater number of discriminations, along with absence of a clean mapping between any major terms, may make acquiring English naming patterns for the bottles set more challenging for Mandarin-English bilinguals. At the same time, if dissimilarity decreases cross-connections to the L1, it may exert less influence on the L1.

Table 1a: Distribution of names across the 67 pictures of the dishes set, grouped by English.
\begin{tabular}{ll}
\hline English & Mandarin \\
\hline 27 bowl & \begin{tabular}{l}
19 wan, 3 pen, 1 pan, 1 die, 1 bei, 1 yao, 1 \\
\\
12 mug
\end{tabular} \\
yan hui gang \\
9 cup & 9 bei \\
8 dish & 3 pan, 3 yan hui gang, 2 pen \\
6 plate & 4 pan, 2 die \\
2 glass & 2 bei \\
1 pot & 1 guo \\
1 jar & 1 wan \\
1 tray & 1 pan \\
\hline
\end{tabular}

Table 1b: Distribution of names across the 67 pictures of the dishes set, grouped by Mandarin
\begin{tabular}{ll}
\hline Mandarin & English \\
\hline 24 bei & 12 mug, 9 cup, 2 glass, 1 bowl \\
20 wan & 19 bowl, 1 jar \\
9 pan & 4 plate, 3 dish, 1 tray, 1 bowl \\
5 pen & 3 bowl, 2 dish \\
4 yan hui gang & 3 dish, 1 bowl \\
3 die & 2 plate, 1 bowl \\
1 guo & 1 pot \\
1 yao & 1 bowl \\
\hline
\end{tabular}

Table 2a: Distribution of names across the 73 pictures of the bottles set, grouped by English.
\begin{tabular}{ll}
\hline English & Mandarin \\
\hline 37 bottle & 33 ping, 3 tong, 1 he \\
7 can & 3 ping, 2 tong, 1 guan, 1 he \\
6 container & 3 he, 2 ping, 1 tong \\
5 box & 5 he \\
4 jar & 2 ping, 1 guan, 1 he \\
4 tube & 4 guan \\
3 stick & 2 ping, 1 guan \\
2 case & 2 he \\
1 basket & 1 he \\
1 canister & 1 he \\
1 carton & 1 he \\
1 grinder & 1 ping \\
1 shaker & 1 ping \\
\hline
\end{tabular}

Table 2b: Distribution of names across the 73 pictures of the bottles set, grouped by Mandarin.
\begin{tabular}{ll}
\hline Mandarin & English \\
\hline 44 ping & 33 bottle, 3 can, 2 stick, 2 jar, 2 \\
container, 1 grinder, 1 shaker \\
16 he & 5 box, 3 container, 2 case, 1 bottle, \\
& \begin{tabular}{l}
1 carton, 1 jar, 1 basket, 1 can, 1 \\
canister
\end{tabular} \\
7 guan \(^{1}\) & 4 tube, 1 jar, 1 stick \\
6 tong \(^{3}\) bottle, 2 can, 1 container \\
\hline
\end{tabular}

What constrains L2 learning given mature L1 knowledge at time of immersion? Can L2 usage evolve toward native-like patterns as a function of \(L 2\) language experience, or will entrenched L1 knowledge defeat reshaping of word usage patterns? Naming performance of each participant for each stimulus set was assessed using a measure of individual agreement with monolingual name choice across all pictures of the set. For each object, the bilingual was credited for the name produced for it proportional to the number of monolingual English speakers who produced that name. For instance, if a given object was called bottle by \(80 \%\) of monolingual speakers, jar by \(10 \%\), container by \(5 \%\), and \(j u g\) by \(5 \%\), then a bilingual who called it bottle received a score of .8 , one who called it jar received a score of .1 and so on. A 0 was assigned for responses not produced by any monolingual speaker. An individual's scores for the 67 dish pictures and 73 bottles pictures were each averaged to produce a summary value for each person for each set. As a baseline for comparison, we also calculated the mean level of agreement for individual monolingual speakers of English with their own monolingual group for each stimulus set.

\footnotetext{
\({ }^{1}\) Because responses were typed in pinyin, we cannot distinguish guan with tone 3 from guan with tone 4 , but for our stimuli, most or all are likely to be guan4.
}

To evaluate whether language experience - in particular, the dominance of one language over the other in current usage - influences match to the monolingual pattern, bilinguals were divided into two groups according to the extent to which English had become dominant for them. To do so, all the individual language performance measures other than naming responses were correlated with one another and with responses to the various language history questions. The relative number of items produced to the category prompt Clothing in English vs. Mandarin correlated significantly with more other measures (20 out of 36) than any other performance measure and was selected as the basis for grouping. To the extent that the bilinguals can retrieve more English than Mandarin words for items of clothing in 60 seconds, their English can be assumed to be more highly activated than their Mandarin.

The distribution of number of English minus Mandarin clothing items produced by each participant was examined for a break point. Participants assigned to the Higher English Dominance group ( \(\mathrm{n}=27\) ) had a mean value of 0.15 (s.d. 3.22), indicating that on average they produced about equal numbers in Chinese and English. Those assigned to the Lower English Dominance group ( \(\mathrm{n}=35\) ) had a mean value of -9.0 (s.d. 3.26), indicating that on average they produced 9 more in Chinese than English. Correspondingly, mean self-rated English proficiency for the Higher English Dominance group was 5.14 and for the Lower, 4.76. For Chinese self-ratings, it was 6.87 and 6.95 respectively. Table 3 presents the mean individual agreement scores of each speaker group to the monolingual English group for each stimulus set.

Table 3: Mean agreement scores of monolinguals and bilinguals to the monolingual English group.
\begin{tabular}{llll}
\hline & & Monolingual \\
English
\end{tabular} \begin{tabular}{l} 
Higher \\
English \\
Dom. \\
Bilinguals
\end{tabular}\(\quad\)\begin{tabular}{l} 
Lower \\
English \\
Dom. \\
Bilinguals
\end{tabular}

An ANOVA with speaker group as a between-subjects factor showed a significant main effect of speaker group for both stimulus sets: F \((2,87)=46.16, p<.0001\) for dishes; \(\mathrm{F}(2,86)=8.14, p<.001\) for bottles. Post hoc comparisons (LSD) showed that bilinguals differed significantly from monolinguals for both stimulus sets ( \(p<.001\) ). The effect of extent of English dominance differed by stimulus set, though. For dishes, the bilingual groups differed from each other ( \(p<.001\) ), but for bottles, they did not. Thus, the ability to progress toward an L2 native-like naming pattern differs by semantic domain.

For dishes, both bilingual groups differed from the monolinguals by greatly over-using cup and plate and under-using mug and dish. Higher English dominance bilinguals added to the dominant term list one word
dominant in monolingual usage for only one object (pot) and one dominant for two (glass), as well as a more important term, mug, dominant for 12 objects for the monolinguals. However, the bulk of their progress does not appear to be due to addition of these vocabulary words. Removing from the data the three stimuli that had monolingual dominant names of pot or glass leaves the scores virtually unchanged (. 50 vs. .45). Furthermore, about 40\% of lower English dominant bilinguals did produce mug, even though it was not dominant for any object for them. Looking only at the scores of those who did produce mug in each group, the mean score for lower English dominance was . 47 and for higher English dominance was .52, maintaining the difference between groups. It appears that progress in matching monolingual patterns is largely due to more appropriate use of terms known to both groups.

For bottles, both bilingual groups differed from the monolinguals by greatly over-using bottle and, to a lesser extent, box, and by under-using container, jar, and several minor terms (dominant for monolinguals for only one to three objects of the set). As already noted, though, there was no sign of greater shaping of the word-object connections toward native-like with higher English dominance.

In short, these late bilinguals speaking two dissimilar languages do show movement toward overcoming entrenched L1 patterns as a function of language experience. However, they do so only for one of the domains, a point to which we will return.

How stable are L1 usage patterns when L2 immersion occurs after L1 is mature? In particular, are highly entrenched L1 patterns immune to an L2 influence, or does the impact vary depending on \(L 2\) experience? Performance on Mandarin naming was scored in the same way as for English naming. Table 4 presents the mean individual agreement scores of each speaker group to the monolingual Mandarin group for each stimulus set.

Table 4. Mean agreement scores of monolinguals and bilinguals to the monolingual Mandarin group.
\begin{tabular}{llll}
\hline & & Higher & Lower \\
& Monolingual \\
Mandarin & \begin{tabular}{l} 
English \\
Dom. \\
Bilinguals
\end{tabular} & \begin{tabular}{l} 
English \\
Dom. \\
Bilinguals
\end{tabular} \\
\hline Dishes & \(.85(.07)\) & \(.69(.03)\) & \(.68(.04)\) \\
Bottles & \(.86(.06)\) & \(.63(.09)\) & \(.68(.06)\) \\
\hline
\end{tabular}

An ANOVA with speaker group as a between-subjects factor showed a significant main effect of speaker group for both stimulus sets: \(\mathrm{F}(2,84)=100.94, p<.0001\) for dishes; \(\mathrm{F}(2,82)=73.44, p<.001\) for bottles. Post hoc comparisons (LSD) showed that bilinguals differed significantly from monolinguals for both stimulus sets ( \(p s<.0001\) ). This indicates that entrenched L1 patterns are not immune from an L2 influence, even under late immersion for bilinguals speaking dissimilar languages. Changes appear to be largely
due to over-extension of ping and guan and under-extension of he and tong. The latter two may have particularly unclear relations to any English words (see Tables 2a and b).The effect of extent of English dominance again differed by stimulus set, though. For bottles, the bilingual groups differed from each other ( \(p<.01\) ); those with higher English dominance were further from the monolingual standard. For dishes, the groups did not differ. Whereas progression toward the \(\mathbf{L} 2\) standard with greater English dominance was shown for dishes, greater loss of agreement with the L1 standard appears here for bottles.

If progress in \(L 2\) mastery is observed and linked to the extent of \(L 2\) dominance, does it have a negative relation to the preservation of the native \(L 1\) patterns, or do they vary independently? The data already presented suggest that they must vary independently, since bilinguals showed differential progress toward the L2 standard only for dishes and differential movement away from the L1 standard only for bottles. To further address this question, we correlated individual bilinguals' mean scores for English and Mandarin performance. For dishes, there was no relation between the two ( \(r=.10\), n.s.). For bottles, there was a small and marginally significant positive relation rather than a negative one ( \(r=.21, p=.06\) ). Overall, then, it appears that progress in one language does not mandate a declining performance in the other across the board with respect to naming patterns.

\section*{Discussion}

We initially outlined several possibilities for how the bilingual lexical network might develop under conditions of late L2 learning. The data argue against the idea that the network stabilizes at an L1 configuration that both protects it against L1 change and prevents progress in L2 acquisition. The data also argue against a reciprocal relation where shifts toward the naming pattern of one language inevitably result in shifts away from the other. At the same time, there was no evidence for the full separation of the two language learning experiences (whereby there could be preservation of the L1 while also progressing toward L2). The current data are most compatible with the situation found in Ameel et al. (2005)'s data for simultaneous French-Dutch bilinguals in Belgium: The network adjusts weights for both languages such that convergence results, and the word usage patterns for each language are more similar for bilinguals than they are for two monolinguals of the corresponding languages.

This outcome is more surprising in the current context, given that the two languages were acquired asynchronously and are dissimilar on many dimensions, as well as having divergent naming patterns with no cognates that might promote incorrect assumptions of word-to-word equivalences. In light of the naming strategies adopted by the bilinguals in each language, though, it may be less surprising. For both L1 and L2, the trend was to over-extend the words that are prominent in the domain (covering a large
numbers of objects for monolinguals) and under-extend words used for smaller subsets of objects. We cannot ensure that our stimulus sets exactly match the distribution of objects in the real world, but we sampled widely and it is likely that, if anything, we somewhat over-represented less common object types. It is probably inevitable that bilinguals receive less of the input needed to maintain (for L1) or establish (for L2) appropriate connection weights to object types for the infrequent words, and will use these less. Some mutual influence may then be exerted across the word-object inputs more commonly encountered, such that the major categories come to resemble each other more.

The remaining critical question is why the two domains showed different outcomes for the effect of increased English dominance. The observed outcomes most likely do derive from the different L1-L2 relations in the two domains, as alluded to earlier. The agreement scores of the monolinguals show that Mandarin speakers use their dominant terms for both domains with a high degree of consistency. English speakers used theirs with much lower consistency, and the full sets of response shows many more uncommon terms (e.g., cylinder, dispenser, vial, tub, tin; platter, saucer, trough, Tupperware) used sporadically. The lower English consistency is especially pronounced for the bottles set. In addition, as noted earlier, this set lacks any terms having a neat mapping to the Chinese terms, whereas the dishes set at least has a fairly clean correspondence of cup, mug, and glass jointly to Chinese bei. For dishes, bilinguals may be able to make progress in the distinctions among cup, mug, and glass without reshaping their use of bei. For bottles, bilinguals are more likely to struggle to acquire the native-like distinctions without success because the input is so highly variable. Nevertheless, the more they tilt toward becoming English dominant, the less they are reinforcing their Chinese usage patterns, and those weaker word-object connection weights may further diminish.

The current discussion has been framed in terms compatible with connectionist modeling. The network perspective provides a framing that links theoretical issues for the lexicon with those for phonology and morpho-syntax and highlights questions about bi-directional influence on patterns of word usage. Implementation is an important next step toward understanding the dynamics of lexical crosslanguage influence. Modeling stands to yield significant insights into the competition and representation of multiple languages in the bilingual mind (see Li, 2013 for a recent discussion). In naturalistic or experimental settings it is often difficult to bring learning variables under tight control, but these variables can be parametrically manipulated in a computational model. For instance, characteristics of the naming patterns, amounts of input in each language, proficiency or dominance in L1 or L2, and temporal characteristics of the input (blocked by language, as in a complete switch to L2, or intermixed as for immersed bilinguals who maintain contact with an L1 community) can be manipulated to make further predictions about what effects might emerge under what circumstances.

Conversely, the behavioral data as reported here help inform the nature of the models to be developed.

\section*{Acknowledgments}

This work was supported by a grant from the National Science Foundation (\#BCS-1057855) to Barbara Malt and Ping Li. Eef Ameel is supported on a postdoctoral fellowship from the Fund for Scientific Research-Flanders. We thank Shijing Jiang, Zinan Jiang, Rachel Jobe, Gwendolyn Johnson, Lauren O'Brien, Nikolas Rivera, Aislinn Rowan, Shenchi Tian, and Rachel Williams for assistance with data collection and analysis, and Ben Zinszer for helpful discussion.

\section*{References}

Ameel, E., Storms, G., Malt, B.C., and Sloman, S. A. (2005). How bilinguals solve the naming problem. Journal of Memory and Language, 53, 60-80.
Schmid, M. S. \& Köpke, B. (2007). Bilingualism and attrition. In B. Köpke, M. S. Schmid, M. Keijzer, \& S. Dostert (Eds.). Language attrition: Theoretical perspectives. Amsterdam: John Benjamins.
Köpke, B., Schmid, M.S., Keijzer, M., \& Dostert, S. (Eds.). Language attrition: Theoretical perspectives. Amsterdam: John Benjamins.
Li, P. (2013). Computational modeling of bilingualism: How can models tell us more about the bilingual mind? Bilingualism: Language and Cognition, 16, 1-6.
Li, P., Farkas, I., \& MacWhinney (2004). Early lexical acquisition in a self-organizing neural network. Neural Networks, 17, 1345-1362.
Malt, B. C. and Sloman, S. A. (2003). Linguistic diversity and object naming by non-native speakers of English. Bilingualism: Language and Cognition, 6, 47-67.
Pavlenko, A. (2009) Conceptual representation in the bilingual lexicon and second language vocabulary learning. In Pavlenko, A. (Ed.) The bilingual mental lexicon: Interdisciplinary approaches. Clevedon, UK: Multilingual Matters.
Pavlenko, A. \& Malt, B.C. (2011). Kitchen Russian: Crosslinguistic differences and first-language object naming by Russian-English bilinguals. Bilingualism: Language and Cognition, 14, 19-45.
Schmid, M. (2011). Language attrition. Cambridge: Cambridge University Press.
Ullman, M. T. (2004). Contributions of neural memory circuits to language: The declarative/procedural model. Cognition, 92, 231-270.
Van Hell, J. G., \& De Groot, A. M. B. (1998). Conceptual representation in bilingual memory: Effects of concreteness and cognate status in word association. Bilingualism: Language and Cognition, 1, 193-211.
Zhao, X., \& Li, P. (2010). Bilingual lexical interactions in an unsupervised neural network model. International Journal of Bilingual Education and Bilingualism, 13, 505-524.

\title{
Do Language Structure or Language Proficiency Affect Critical Evaluation?
}

\author{
Emmanuel Manalo (emmanuel.manalo@gmail.com) \\ Kyoko Watanabe (watanabekyoko@aoni.waseda.jp) Chris Sheppard (chris@waseda.jp) \\ Faculty of Science and Engineering, Waseda University \\ 3-4-1 Ohkubo, Shinjuku-ku, Tokyo 169-8555, JAPAN
}

\begin{abstract}
This study examined whether language structure or language proficiency might influence students' use of evaluative language in written reports, and whether instruction might improve students' use of evaluative language. Reports in Japanese and in English written by second year Japanese university students, who had received instruction in academic discourse pertaining to critical evaluation, were analyzed for use of evaluative statements. This revealed no disadvantage for use of the Japanese language, which is considered as having a more indirect structure that may make critical evaluation more difficult. English proficiency test scores, however, were found to correlate with production of evaluative statements in English, but not in Japanese, suggesting that inadequate second language proficiency could limit critical evaluation use. The second year students' use of evaluative statements was also found higher than their first year counterparts' (who had not yet received instruction), suggesting that such instruction is beneficial for skills development in both languages.
\end{abstract}

Keywords: critical evaluation; critical thinking; language structure; second language proficiency; cognitive cost

\section*{Introduction}

The cultivation of students' abilities to critically evaluate the soundness of knowledge claims and arguments is considered as one of the most important objectives of education (e.g., Glassner, Weinstock, \& Neuman, 2005) and, with the proliferation of unvetted available information through the Internet and other forms of media in modern societies, the ability to determine credibility has become crucial (e.g., Thomm \& Bromme, 2011). Developing students' critical thinking skills (the broader set of skills to which critical evaluation belongs) is, however, not without its challenges (e.g., Halpern, 1998). There are various factors that have been claimed to affect people's use of critical thinking, including some culture-related factors. Asian students, in particular, have often been portrayed as lacking in critical thinking skills compared to Western students (e.g., Atkinson, 1997; Fox, 1994), and many tertiary instructors have been found to subscribe to such a view (e.g., Lee \& Carrasquillo, 2006; Robertson, Line, Jones, \& Thomas, 2000).

One explanation that has been put forward for the apparent differences in critical thinking skills manifested by students from different cultural groups concerns the structure of their native language. This explanation posits that, due to their structure, some languages may present constraints in the ease with which certain thinking skills can
be carried out or expressed. This explanation is sometimes referred to as the "Sapir-Whorf hypothesis" (see Au, 1983; Hockett, 1954), which suggests that languages differ in the relative ease with which they can be used to convey certain ideas. An example of a claim of this kind is Bloom's (1981) proposal that counterfactual thinking (i.e., thinking about what might have been, contrary to facts) may be more difficult in Chinese compared to English.

More recent observations of linguistic differences, such as "indirectness" being a feature more prevalent in some languages, particularly Asian languages (e.g., Kong, 2005), would appear to support the notion that language structure could affect the ease with which certain modes of thinking could be undertaken or expressed. In a study by Itakura and Tsui (2011), for example, evidence was found that book reviewers use different strategies to convey critical evaluation when writing in Japanese compared to English. For example, in Japanese, criticism is usually indirectly conveyed and is frequently preceded by an apology.

\section*{Language Structure or Language Proficiency?}

Previous studies, however, had not clarified whether language structures could actually impose constraints in what users of the language can do. Although the earliermentioned study by Bloom (1981) claimed to have found evidence for this where counterfactual thinking in the Chinese language is concerned, subsequent investigations failed to replicate or support Bloom's results (Au, 1983). Thus it remains unclear whether, for example, the structure of a language like Japanese would make it relatively more difficult to undertake tasks like critical evaluation (cf. Itakura and Tsui's, 2011, findings), and hence make a person appear less competent in his or her critical thinking skills.

Concerning international students who have been reported as appearing less competent in critical thinking skills (cf. Lee \& Carrasquillo, 2006; Robertson et al., 2000), there is another possible explanation that other authors have previously suggested (e.g., Floyd, 2011; Lun, Fischer, \& Ward, 2010; Paton, 2005) but which had not been adequately tested. This explanation hinges on the fact that many international students have to use a second language (L2), like English, in their host environment. It suggests that, if a person is not so proficient in a language, he or she would generally manifest lower competence in carrying out tasks when using that language. Tasks that are likely to get affected include cognitive tasks like critical thinking.

This possible influence of language proficiency on critical thinking skills application can be explained in terms of cognitive cost (i.e., the mental resources cost associated with executing tasks). Language processing entails the use of cognitive resources in working memory (Baddeley, 1986, 1998), and lower proficiency in a language would require the use of more resources (i.e., the cognitive cost would be higher). The application of critical thinking skills would likewise require the use of working memory resources. The resources available in working memory, however, are limited (Baddeley, 1986, 1998) and, if a considerable amount of those resources has already been expended on utilizing a language in which proficiency is low, there may not be adequate resources remaining for the satisfactory execution of critical thinking.

The negative impact of the higher cognitive cost entailed in using a language in which proficiency is low, on the execution of other cognitive tasks, has been demonstrated in previous research. Takano and Noda \((1993,1995)\) showed that the use of a foreign language detrimentally affects performance in concurrently undertaken non-linguistic tasks like arithmetic calculation and mental imagery, and Manalo and Uesaka (2012) reported evidence indicating that students' lower proficiency in an L2 limits their ability to use diagrams when explaining information in that L2. Where critical thinking is concerned, both Lun et al. (2010) and Floyd (2011) reported indications that lower proficiency levels in English could detrimentally affect Asian students' performance in critical thinking tests administered in English. However, neither of those studies used appropriate, objective measures of L2 proficiency to reliably confirm the connection between L2 proficiency and critical thinking skills performance.

\section*{Overview of the Present Study}

The main purpose of the present study was to examine the possible influences of language structure, and proficiency in L2, on students' manifestation of critical thinking in their writing. The study was not intended to be a comprehensive test of the language structure hypothesis: it examined only whether, in the written work of Japanese university students, there might be observable differences in the presence of critical thinking qualities, depending on the language being used, Japanese or English. Critical thinking was operationalized as students' use of evaluative statements. Such use was chosen for investigation because it comprises a salient expression of critical evaluation, which in turn is central to the notion of critical thinking application (cf. Fisher \& Scriven's, 1997, p. 21, definition of critical thinking as "skilled and active interpretation and evaluation of observations and communications, information and argumentation" - italics added).

In the present study, Japanese was deemed an appropriate language to examine because, like a number of other Asian languages, it employs patterns of expression that make it more indirect and inductive compared to English (e.g., Itakura \& Tsui, 2011; Scollon \& Wong-Scollon, 1991).

Evaluation, however, requires precision and directness in conveying judgments about the quality or value of the subject being referred to. Thus, structural features of the Japanese language could make the production of evaluative language relatively more difficult. If so, it should be possible to detect lower rates of evaluative language use in the students' written work in Japanese compared to English.

As this study was focusing on students' written work in both L1 and L2, it was equally important to consider whether using an L2 may detrimentally affect students' critical evaluation performance. Thus, possible relationships between students' TOEIC test scores (Test of English for International Communication, a norm-referenced test of English listening comprehension and reading skills, widely used as a measure of students' English language proficiency levels in Japan; http://www.ets.org/toeic) and their production of evaluative statements were investigated. The question here was whether L2 proficiency would manifest as a limiting factor because lower proficiency entails higher cognitive cost when using the L2, leaving insufficient resources in working memory for critical evaluation. If this explanation is supported, a relationship should be found between the students' TOEIC scores and their evaluative statements production in the L2, but not in the L1. A relationship in the L1 would suggest that general language or intellectual abilities - rather than L2 proficiency - affect critical evaluation performance. The reason is that language abilities, and intellectual abilities and performance, are generally considered as being related (e.g., Ackerman, 1986; Neisser et al., 1996). Thus, a student with higher language and intellectual abilities could be expected to score higher in the TOEIC test, and evidence better performance in tasks like critical evaluation - in both their L1 and L2.

The research conducted comprised two related studies. In Study 1, evaluative statements that second year Japanese university students produced in Japanese (their L1) and in English (their L2) were examined. These students had received instruction on academic discourse. Thus, they were not naïve as to the requirements of expressing evaluative language, and any differences in the writing they produced in L1 and L2 could be attributed to either the inherent structure of the language they were using or their proficiency in using that language (particularly the L2).

In Study 2, the same writing task was given to first year students who had received little instruction on academic discourse, and nothing explicit on the production of evaluative language. The purpose of this second study was to find out if the characteristics of L1 and L2 written work produced by the first year students, compared to their second year counterparts, differed - and hence, whether the additional instruction that had been received by the more advanced second year students might have made a difference.

\section*{Study 1}

The first study was carried out to test the hypothesis that students' production of evaluative statements in Japanese
and in English would differ. A second hypothesis was also tested: that, if L2 proficiency is a limiting factor in students' critical evaluation performance, their TOEIC scores would be related to their evaluative statements production in L2, but not in L1. Lower use of evaluative statements in the students' L2 work should also be observable if this L2 proficiency hypothesis applies.

\section*{Method}

Participants The participants were 111 Japanese university students in their second year of study in science and engineering disciplines. For these students, Japanese is L1 and English is L2. These students were taking a compulsory English communication skills development course that covers oral and written academic discourse in task-based discussion and research development. The students came from four different classes in that course.

The students were required to sit the TOEIC test at regular intervals during their period of enrolment, and their scores on that test were available to their course teachers.

Materials and Procedure As part of the communication skills course, the students were provided class instruction, textbook explanations and examples (Anthony, Rose, \& Sheppard, 2010), and practice in the use of language appropriate for critical evaluation, including ranking and debating different reasons and other forms of alternatives (e.g., clearly stating the premises, and then drawing conclusions). These were all provided in English.

For the purposes of the present investigation, the students were additionally provided with a single page Japanese translation of the part of the textbook dealing with how to make valid arguments. They were also supplied brief (one page) written examples (one in English and one in Japanese) of how alternative reasons could be ranked according to judgments about their relative importance. The example texts conveyed someone's opinion about the most important reason for learning the English language, among four possible reasons. The texts provided examples of evaluative statements and provision of support for claims, although those were not labeled or overtly identified in any way in the texts. The equivalence and appropriate use of language in the English and Japanese versions were checked by several bilingual teachers of the course. Although all materials provided in the course are usually in English, the Japanese versions were supplied in this case to avoid possible disadvantage to the students' production of evaluative language in Japanese (i.e., without the Japanese versions, it could be argued that the students might have simply been unfamiliar with the equivalent Japanese expressions for critical evaluation).

During two 90 -minute class sessions of the communication skills course, the students were introduced to the Titanic and Space Shuttle Challenger disasters, including four basic causes that have been proposed for the occurrence of each of those disasters. During the class sessions, the students participated in guided exercises to
explore and discuss the disasters and their corresponding possible causes.

For homework, the students were asked to write two brief reports to explain what they considered to be the most important cause of each of the disasters. To avoid any possible misunderstandings about the requirements of the homework task, written instructions were provided in Japanese. The students were randomly assigned to write one report in English and the other in Japanese (i.e., if they were asked to write the Titanic report in English, they had to write the Challenger report in Japanese, and vice versa).

Analyses The following were counted and scored in the analysis of the students written work:
a) Number of sentences [Total];
b) Number of evaluative sentences (i.e., sentences where some evaluation of the relative value of the topic is made) [Evaluative];
c) Number of evaluative sentences specifically about the causes of the disaster (i.e., sentences where some evaluation is made about the relative importance of the causes given for the occurrence of the disaster) [Causes];
d) Number of evaluative sentences that are supported by reason or evidence of some kind [Supported].
Operational criteria were drawn up for determining what data counted under each of these categories. For example, where "evaluative sentences" were concerned, the following were required: the sentence must explicitly say something about the worth or value of the subject, and that worth or value must be in comparison to something else. Conditional statements that explicitly convey a relative evaluation of the subject were counted. The following examples, in contrast, did not count: the use of simple adjectives or adverbs to describe something, prescriptive statements not explicitly expressing a relative evaluation or judgment, and conditional statements in general.

Inter-rater reliability was checked by asking an independent coder to score a randomly selected sample of \(25 \%\) of the data. Reliability coefficients obtained (Cronbach's alphas) were deemed to be satisfactory (e.g., .922 and .940 in English and .960 and .963 in Japanese for the "Evaluative" and "Causes" scores, respectively).

Analyses of variance were conducted to compare the students' scores in each of the categories noted above in English and in Japanese. Correlational analyses were carried out to examine possible relationships with the students' most recent TOEIC test scores.

\section*{Results}

Table 1 shows the means, and standard deviations (in brackets), obtained under each category for the students' written work in English and in Japanese.

No significant effects were found due to the task (i.e., the Titanic compared to the Challenger reports). The analysis however revealed significant effects due to language in the total number of sentences written [Total], \(F(1,110)=11.51\),
\(p=.001, \eta_{\mathrm{p}}{ }^{2}=.095\); the number of evaluative sentences [Evaluative], \(F(1,110)=4.85, p=.030, \eta_{\mathrm{p}}{ }^{2}=.042\); the number of evaluative sentences about causes [Causes], \(F(1\), \(110)=5.00, p=.027, \eta_{\mathrm{p}}^{2}=.044\); and the number of evaluative sentences with support [Supported], \(F(1,110)=\) \(9.61, p=.002, \eta_{\mathrm{p}}^{2}=.080\). These results indicate that the students wrote more sentences in English compared to Japanese, but they wrote more evaluative sentences, evaluative sentences about causes, and evaluative sentences with support in Japanese compared to English.

Table 1: Mean report scores according to language used
\begin{tabular}{lcccc}
\hline & Total & Evaluat. & Causes & Support. \\
\hline English & 20.35 & 3.48 & 3.38 & 2.08 \\
& \((5.58)\) & \((1.77)\) & \((1.77)\) & \((1.42)\) \\
\hline Japanese & 18.72 & 3.76 & 3.67 & 2.49 \\
& \((5.91)\) & \((1.80)\) & \((1.80)\) & \((1.38)\) \\
\hline
\end{tabular}

Because the total number of sentences that the students wrote in English and in Japanese differed, the proportions (i.e., Evaluative, Causes, and Supported sentences as proportions of Total) were also calculated and compared according to the language used. The comparisons revealed significant differences in each case: for Evaluative, \(F(1\), \(110)=20.17, p<.001, \eta_{\mathrm{p}}^{2}=.155\); for Causes, \(F(1,110)=\) 20.29, \(p<.001, \eta_{\mathrm{p}}^{2}=.156\); and for Supported, \(F(1,110)=\) 24.90, \(p<.001, \eta_{\mathrm{p}}^{2}=.185\). These results, depicted in Figure 1, indicate that the proportions of Evaluative, Causes, and Supported sentences were higher in the reports that the students wrote in Japanese compared to those they wrote in English.

The results of the correlational analysis are shown in Table 2. In the students' written work in English, TOEIC scores correlated significantly with all categories of scores obtained. However, in Japanese, TOEIC scores significantly correlated only with Total and Supported sentences.

\section*{Discussion}

Differences were found in both actual numbers and proportions of evaluative sentences that the students produced in English and in Japanese. The direction of the differences, however, was opposite to the language structure-based prediction: higher proportions of evaluative sentences were found in Japanese instead of English. This result suggests that the students were better at producing evaluative language in their L1. The significant correlation found between students' TOEIC scores and their English writing scores, and the lack of significant correlations in Japanese where the Evaluative and Causes sentences were concerned, suggest that the students' English/L2 proficiency accounts for at least part of that difference.

The significant correlations between TOEIC scores and the numbers of Total and Supported sentences that were also present in Japanese suggest that general language skills (which is also correlated with TOEIC scores) may affect
students' productivity and use of evidence in writing, irrespective of the language being used.

As noted earlier, the student participants in this first study had already received instruction in academic discourse that includes the use of evaluative language. Therefore, an important next question to address was, "To what extent had that instruction affected the relative production of evaluative language in English and in Japanese?" - which was pursued in the second study.


Figure 1: Evaluative, evaluative about causes, and evaluative supported sentences, as proportions of the total number of sentences written, in English and in Japanese.

Table 2: Correlation coefficients between students' TOEIC scores and categories of their report scores, according to the language used (effect sizes shown in brackets)
\begin{tabular}{lcccc}
\hline & Total & Evaluat. & Causes & Support. \\
\hline English & \(.22^{*}\) & \(.22^{*}\) & \(.22^{*}\) & \(.23^{*}\) \\
& \((.047)\) & \((.049)\) & \((.048)\) & \((.051)\) \\
\hline Japanese & \(.27^{* *}\) & .18 & .15 & \(.25^{* *}\) \\
& \((.075)\) & \((.032)\) & \((.023)\) & \((.062)\) \\
\hline\(* p<.05{ }^{* *}\) &
\end{tabular}
* \(p<.05\). ** \(p<.01\).

\section*{Study 2}

The purpose of the second study was to examine whether first year students manifest lower use of the target evaluative language compared to the second year students, and whether any such differences might be consistent across English and Japanese.

\section*{Method}

Participants The participants were 44 Japanese university students who were in their first year of studies in the same science and engineering faculty as the students in Study 1. The students came from two classes of a compulsory first year English communication skills course which deals with various aspects of oral and written academic discourse, but nothing explicit about evaluative language (which is not covered until the second year course).

Materials, Procedure, and Analysis For one of their homework assignments, the students were given brief
reading materials (in English and in Japanese) about the Titanic and Space Shuttle Challenger disasters, including the proposed causes of those disasters. These materials were drawn from the textbook used in the second year course. The Japanese translations were provided to these first year participants to ensure that their subsequent writing performance would not have been compromised by possible difficulties in understanding the English versions. The content of those materials were not covered in class.

The homework task that the students had to do was the same as that given to the second year students: to produce two brief reports to explain what they considered to be the most important cause of each of the disasters, after reading the materials provided. Like the second year students, they were randomly assigned to write one report in English and the other in Japanese. Also, like the second year students, they were provided with the one-page examples (one in English and one in Japanese) of how alternative reasons (for learning the English language) could be ranked according to judgments about their relative importance. The crucial difference was that the first year students were not provided class instruction and exercises on the use of academic discourse specifically pertaining to evaluative language.

The written reports that the students produced were analyzed and scored in the same manner described in the first study. The first and second year students' data were then compared.

\section*{Results}

Analyses of variance revealed significant effects due to year of enrolment (first year compared to second year) in the students' scores for: Total, \(F(1,153)=23.37, p<.001, \eta_{\mathrm{p}}{ }^{2}\) \(=.133\); Evaluative, \(F(1,153)=27.79, p<.001, \eta_{\mathrm{p}}^{2}=.154\); for Causes, \(F(1,153)=27.15, p<.001, \eta_{\mathrm{p}}^{2}=.151\); and for Supported, \(F(1,153)=6.99, p=.009, \eta_{\mathrm{p}}{ }^{2}=.044\). Significant language effects were also found for Total, \(F(1\), \(153)=26.84, p<.001, \eta_{\mathrm{p}}^{2}=.149\); Causes, \(F(1,153)=4.31\), \(p=.04, \eta_{\mathrm{p}}^{2}=.027\); and Supported, \(F(1,153)=14.03, p<\) \(.001, \eta_{\mathrm{p}}{ }^{2}=.084\). No significant interaction effects between language and year were found; nor were any significant effects found due to the task (Titanic versus Challenger).

These results indicate that, compared to the second year students, the first year students wrote fewer sentences in total for their reports. They also produced fewer evaluative statements (evaluative sentences, evaluative sentences about causes, evaluative sentences that are supported). These differences in the students' production of evaluative language are depicted in Figure 2. Significant language differences were found in the total number of sentences, number of evaluative sentences about causes, and number of evaluative sentences with support that the students wrote: in each case, the students produced more in Japanese compared to English.

\section*{Discussion}

The results of Study 2 showed that the second year students wrote more sentences in their reports, and produced more of
the target evaluative language, compared to the first year students. This finding suggests that instruction on appropriate language to use - which had been provided to the second year students - can improve students' abilities in manifesting critical evaluation in their written work. Although as noted the instruction was provided almost entirely in English, the significant language effects found were all in favor of the Japanese language, which suggests that there is transfer across the languages in skills acquisition. In other words, skills taught and learned in English also produce improvements in the production of evaluative language in Japanese.


Figure 2: Mean numbers of evaluative, evaluative about causes, and evaluative supported sentences produced in English and Japanese reports by the first year (English 1, Japanese 1) and second year (English 2, Japanese 2) students.

\section*{General Discussion}

The findings of this study provide clear evidence that, at least for Japanese students, using the Japanese language (their L1) presents no disadvantage compared to English (their L2) in the production of evaluative language (i.e., the Japanese language structure is not a limiting factor). How Japanese students' evaluative language use might compare to that of students whose first language is structured differently (e.g., native English speakers responding to the same tasks), or students who are fully bilingual in Japanese and English, would need to be examined in future research. However, in the present study, there appeared to be no obvious deficits in evaluative language production in Japanese among the second year students who had received instruction in the necessary academic discourse.

There is evidence in the present study, however, that language proficiency can be a limiting factor in the production of evaluative language. The significant correlations between the students' TOEIC scores and their production of evaluative sentences in English (their L2) but not in Japanese (their L1) - indicate that performance varied with L2 proficiency. This provides useful evidence to corroborate previously made claims (e.g., Floyd, 2011; Lun
et al., 2010; Paton, 2005) that some of the shortcomings in critical thinking skills manifested by international students can be attributed to their having to use an L2 in which they may not be as proficient compared to their native speaker counterparts.

The finding about L2 proficiency being a potential limiting factor in students' use of the target critical evaluation language suggests that, to address the perceived deficiencies in Asian and other foreign students' critical thinking skills, educational strategies that would improve their proficiencies in English (or whatever language is used in the host country) would be helpful.

The findings of this study also show that appropriate classroom instruction promotes university students' development of skills in critical evaluation. The second year students evidenced similar writing profiles to those of first year students; however, having received instructions in academic discourse relevant to critical evaluation, they also produced more of the target evaluative language. They did this in both languages, L1 and L2, even though academic discourse instruction was primarily provided in the L2 suggesting some transfer of skills across languages.

\section*{Acknowledgments}

This research was supported by a grant-in-aid (A11550900) received from the Japan Society for the Promotion of Science. The authors would like to thank Takashi Kusumi for his support.

\section*{References}

Ackerman, P. L. (1986). Individual differences in information processing: An investigation of intellectual abilities and task performance during practice. Intelligence, 10, 101-139.
Anthony, L., Rose, R., \& Sheppard, C. (2010). Concept building and discussion: Foundations. Tokyo: DTP Publishing.
Atkinson, D. (1997). A critical approach to critical thinking in TESOL. TESOL Quarterly, 31, 9-37.
Au, T. K. F. (1983). Chinese and English counterfactuals: The Sapir-Whorf hypothesis revisited. Cognition, 15 , 155-187.
Baddeley, A. D. (1986). Working memory. Oxford: Oxford University Press.
Baddeley, A. D. (1998). Human memory. Boston: Allyn \& Bacon.
Bloom, A. (1981). The linguistic shaping of thought. Hillsdale, N J: Erlbaum.
Fisher, A., \& Scriven, M. (1997). Critical thinking: Its definition and assessment. Norwich, UK: Centre for Research in Critical Thinking.
Floyd, C. B. (2011). Critical thinking in a second language. Higher Education Research and Development, 30, 289 302.

Fox, H. (1994). Listening to the world. Urbana, IL: National Council of Teachers of English.

Glassner, A., Weinstock, M., \& Neuman, Y. (2005). Pupils’ evaluation and generation of evidence and explanation in argumentation. British Journal of Educational Psychology, 75, 105-118.
Halpern, D. F. (1998). Teaching critical thinking for transfer across domains. American Psychologist, 53, 449-455.
Hockett, C. E. (1954). Chinese versus English: An exploration of the Whorfian thesis. In H. Hoijer (Ed.), Language in Culture. Chicago: University of Chicago Press.
Itakura, H., \& Tsui, A. B. M. (2011). Evaluation in academic discourse: Managing criticism in Japanese and English book reviews. Journal of Pragmatics, 43, 13661379.

Kong, K. C. C. (2005). Linguistic resources as evaluators in English and Chinese research articles. Multilingua 24, 275-308.
Lee, K. S., \& Carrasquillo, A. (2006). Korean college students in United States: Perceptions of professors and students. College Student Journal, 40, 442-456.
Lun, V. M. C., Fischer, R., \& Ward, C. (2010). Exploring cultural differences in critical thinking: Is it about my thinking style or the language I speak? Learning and Individual Differences, 20, 604-616.
Manalo, E., \& Uesaka, Y. (2012). Elucidating the mechanism of spontaneous diagram use in explanations: How cognitive processing of text and diagrammatic representations is influenced by individual and taskrelated factors. In P. Cox, P. Rodgers, \& B. Plimmer (Eds), Diagrams 2012, Lecture Notes in Artificial Intelligence (LNAI) 7352 (pp. 35-50). Berlin Heidelberg: Springer-Verlag.
Neisser, U., Boodoo, G., Bouchard, T. J. Jr., Boykin, A. W., Brody, N., Ceci, S. J., ... Urbina, S. (1996). Intelligence: Knowns and unknowns. American Psychologist, 51, 77101.

Paton, M. (2005). Is critical analysis foreign to Chinese students? In E. Manalo, \& G. Wong-Toi (Eds.), Communication skills in university education: The international dimension (pp. 1-11). Auckland: Pearson Education.
Robertson, M., Line, M., Jones, S., \& Thomas, S. (2000). International students, learning environments and perceptions: A case study using the Delphi technique. Higher Education Research \& Development, 19, 89-102.
Scollon, R., \& Wong-Scollon, S. (1991). Topic confusion in English-Asian discourse. World Englishes 10, 113-125.
Takano, Y., \& Noda, A. (1993). A temporary decline of thinking ability during foreign language processing. Journal of Cross-cultural Psychology, 24, 445-462.
Takano, Y., \& Noda, A. (1995). Interlanguage dissimilarity enhances the decline of thinking ability during foreign language processing. Language Learning, 45, 657-681.
Thomm, E., \& Bromme, R. (2011). "It should at least seem scientific!" Textual features of "scientificness" and their impact on lay assessments of online information. Science Education, 96, 187-211.

\title{
Moral choices: The influence of the "Do not play God" principle
}

\author{
Amelia Gangemi (gangemia@unime.it) \\ Department of Cognitive Science, Via Concezione, 6 Messina, 98121- Italy \\ Francesco Mancini (mancini@apc.it) \\ Scuola di Psicoterapia Cognitiva, Viale Castro Pretorio, 116, Rome, 00185 - Italy
}

\begin{abstract}
A wide literature demonstrates that people prefer harm caused by omissions over equal or lesser harm caused by actions. This omission bias has been explained referring to several principles, such as causality or responsibility. A convincing research view has been suggested by Sunstein (2005): harmful acts are generally worse than harmful omissions when moral intuitions reflect the "Do not play God" principle: inactions interfere less with the "natural order." In two preliminary studies, we examine the influence of the "Do not play God" principle on individuals moral preferences, using the switch version of the trolley problem. Study 1 demonstrates that our participants' justifications for their inaction choice explicitly refer to the intention of not interfering with the "natural order". Study 2 demonstrates that the presence of stimuli influencing a reduction of protagonist's decisional autonomy (e.g. an authority) activates the "Do not play God" principle, leading them to prefer inaction.
\end{abstract}

Keywords: Omission bias; Moral choices; Trolley dilemma; "Do not play God" principle.

\section*{Introduction}

It's quite common the intuition that it is worse for a doctor to kill a patient with a deadly disease then let him die by abstaining from any kind of medical intervention. Consequentialist philosophers argue that these cases should be considered equivalently (Singer, 1979). In a number of well-controlled experiments Baron and colleagues have shown instead that people consider harmful acts worse than harmful omissions with otherwise identical, predictable outcomes (i.e., omission bias). For example, Spranca, Minsk and Baron (1991) showed that people find it worse when somebody who wants to harm a person offers this person a food item with an allergenic ingredient than when she passively watches the person who does not know about the ingredient taking this item himself. Ritov and Baron (1990) used also vaccination to illustrate the bias: many people consider the risk of harm from vaccination as more serious than the risk from omitting vaccination. This bias seems to affect real vaccination decisions (Asch et al., 1994; Meszaros et al., 1996), and it has been replicated in several situations (e.g. Royzman \& Baron, 2002; Baron \& Leshner, 2000).

Wroe and Salkovskis (2000) explain this kind of choices arguing that most people regard themselves as more responsible for what they actively do than what they fail to do, and that this omission bias occurs due to perceived differences in causality and differing degrees of responsibility. In particular, according to these authors participants' judgments about the immorality of commission depend on several factors that ordinarily distinguish omission and commission, such as physical movements in commission.

Sunstein (2005) suggests instead that harmful acts are generally worse than harmful omissions when individuals' moral intuitions reflect an overgeneralization of the "Do not play God" principle: omissions or inactions interfere less with the "natural order." Omission generally carries less negative moral weight than commission, intervening less in individual's destiny.

Let's go back to the doctor who decides to let a patient die by refraining from any kind of medical intervention. According to the common sense, this case is considered less morally negative than the case in which the same doctor gives the patient a medication that quickly kills him. But the former case (to suspend any medical treatment, i.e. the omission) does not imply less responsibility or physical movements than the second decision (to give a deadly medication, i.e. the commission): the doctor could ask to move the patient from the emergency room, or could proceed by removing tubes or catheters, or finally by communicating the nurses his instructions and supervising how they follow them. In other words, contrary to Wroe and Salkovskis (2000), the harmful omission could not necessarily imply less responsibility or physical movements than actions. From this point of view there are no differences between action and omission.

So what does differentiate them? According to Sunstein's hypothesis, the omission choice has less impact on the "natural order" (in the example the patient's destiny), not violating the "do not play God" principle, and this would make the difference. It's like people say: nobody can claim the right to decide over the life and death of someone, even if s/he explicitly asks to die. S/he can only follow and adapt to the events.

Moreover, Haidt and Baron (1999) showed that the differences between harmful actions and omissions disappear (i.e. the omission has the same moral weight of
action) with people that are in roles that make them responsible. They have an equal responsibility to prevent harm through both action and omission, like for example a captain of a ship, who is equally responsible for both the acts and omissions that lead to similar harm for the passengers. That is, the higher is the social role, the higher is the decisional autonomy and the right to intervene on the natural order (i.e. responsibility to protect) and thus the less is the weight of the "do not play God" principle.

So far, no empirical studies have investigated whether the "do not play God" principle influences individuals choices when faced with problems like moral dilemmas.

The present study aimed to examine whether individuals tend to prefer harm caused by omissions over equal or lesser harm caused by acts, on the basis of their moral intuition based on the "Do not play God" principle. To this aim, we used the well-known switch version of the trolley problem. In its original form, the problem asks people to suppose that a runaway trolley is headed for five people, who will be killed if the trolley continues on its course. The question is whether one would throw a switch that would move the trolley onto another set of tracks, killing one person rather than five. This moral dilemma requires participants to choose one of two undesirable courses of action (both involving loss of life). The action option requires subjects to act, thereby causing the death of one person (but indirectly saving the lives of others). According to Sunstein's hypothesis, it allows modification of the "natural order" in the attempt to minimize the number of victims. The omission option involves no action, and the failure to act results in the deaths of five people. But omission does not modify the "natural order" and respects the "Do not play God" principle.

In line to the consequentialist point of view, in this dilemma, people should prefer the action option: it involves a lower number of victims. According to Suntein, people would prefer the omission option, consistent with the "Do not play God" principle.

To test this hypothesis two different studies were carried out. In the first, we wanted to verify whether subjects preferring inaction would actually tend to justify it by referring to the "Do not play God" principle, while those preferring action would tend to justify it by referring to the consequentialist idea of minimizing suffering and victims. According to Cushman and colleagues (Cushman, Young, \& Hauser, 2006), we assume that the principles used in judgments are articulated in justifications. In the second study, we wanted to verify whether the preference for the omissions would enhance with stimuli leading to a reduction of protagonist's decisional autonomy (e.g. an authority).

When faced with the original trolley problem, most subjects (80-90\%) prefer action (see Greene, Sommerville, Nystrom, Darley, \& Cohen, 2001; Greene, Nystrom, Engell, Darley, \& Cohen, 2004; Greene, Cushman, Stewart, Lowenberg, Nystrom, \& Cohen, 2009). In order to avoid this sort of ceiling effect, which could interfere with the results of the second experiment, we used a version of the
problem with a modified proportion of victims, five vs. three instead of the original five vs. one.

In a preliminary study of the first study we tested this modified version with a group of 54 volunteers, undergraduate students from the University of Rome, with a mean age of 20.2 (ranging from 18-32). All participants were given four moral dilemmas (see below) with the new proportion of victims. Each dilemma required participants to indicate which of two courses of action they would take if confronted with such dilemmas in real life (Greene \& Haidt, 2002; Greene et al., 2004). Participants were asked to respond to each dilemma by marking "yes" (action) or "no" (inaction). The total number of inaction choices made by each participant was the dependent variable. With this modified trolley problem, there were about \(50 \%\) action choices in all dilemmas.

\section*{Study 1}

In this study we wanted to verify whether subjects preferring inaction would tend to justify it by referring to the "Do not play God" principle (e.g. "Who am I to decide who lives and who dies?'"), while those preferring action would tend to justify it by referring to the consequentialist attempt to minimize suffering (e.g. "it's better that three people die instead of five"). Four moral dilemmas were shown to a group of participants. For each dilemma participants were asked to justify their choice. Two judges codified all justifications into two categories: deontological and consequentialist.

\section*{Method}

\section*{Participants}

Participants were 69 undergraduate and postgraduate students recruited by advertisements at the University of Rome (Italy) ( 45 females and 24 males). Their ages ranged from 18 to 45 with a mean age of 23.8 . All of the participants were thus volunteers and provided informed consent.

\section*{Materials and Procedure}

After completing a demographic questionnaire, participants received seven brief scenarios comprised of 6-8 sentences each. Four scenarios concerned moral dilemmas, each requiring participants to indicate which of two courses of action they would take if confronted with such dilemmas (Greene \& Haidt, 2002; Greene, et al., 2004). In one (action) alternative, the participant acted, thereby killing three human beings, but saving the lives of five others. In the second (inaction) alternative the participant did not act, and therefore did not kill three human beings. However, in this second alternative, the participant's failure to act resulted in more deaths (five) than in the first alternative. The other three scenarios required participants to choose between action and inaction. These control scenarios did not involve moral dilemmas. They did not present any victims or harm, but were included as filler items. The order of the seven dilemmas was randomized. Following is an example of the two kinds of dilemmas, moral and control, presented in the study (translated from Italian):

Moral Dilemma

You are near a Ferris wheel. It does not work. Just under the wheel, there are five tourists. Suddenly, the wheel starts turning and soon a cabin will kill them. There is no way to warn them and they cannot escape in any way. The only way to save the five tourists is to pull a lever that can change the rotation of the wheel. Unfortunately, there are three people on the other side that would be killed. Should you pull the lever?

\section*{Control dilemma}

You have just sent an e-mail order for three books that you need for your studies (they are by your favorite writer), when a colleague suggests that you buy the same books and two more (five books in total) at a discount. The order cancellation procedure requires too much time. Should you proceed with the cancellation procedure?

The text of all seven scenarios is available on the web at www.apc.it. After having responded to each scenario by marking "yes" (action) or "no" (inaction), participants were asked to justify their choice in their own words. Of 276 justifications, 18 were removed from the analyses because participants provided a nonsensical response or a judgment that made it clear they had misunderstood a scenario. Two colleagues of the authors who were blind to the hypotheses being tested coded a total of 258 justifications. Justifications were coded into two exclusive categories:

Deontological: justification refers to the importance of not substituting God, not interfering with a destiny already determined, or not taking the responsibility of deciding for others.

Consequentialist: justification refers to the importance of saving the greatest number of lives.

\section*{Results}

According to our hypotheses, we found that almost all participants preferring inaction (96\%) justified it by referring to the "Do not play God" principle (e.g. "I cannot decide who lives and who dies") \((\chi 2(1, N=69)=213.6 ; p\) \(<0.001\) ), while most of those preferring action (86\%) justified it by referring to the importance of minimizing suffering (e.g. "it’s better that three people die instead of five") \((\chi 2(1, N=69)=133.9\), see Table 1\()\).

Table 1. Proportions (and frequencies) of justifications given by participants for their action/inaction choices.
\begin{tabular}{ccc}
\hline & \multicolumn{2}{c}{ Choice } \\
\hline Justification & Action & Inaction \\
Deontological & \(14(38)\) & \(96(242)\) \\
Consequentialist & \(86(226)\) & \(4(10)\) \\
\hline
\end{tabular}

Table 2 provides the overall agreement between coders for the two categories of justifications, along with Cohen's kappa, a statistic of interobserver reliability for which values between .60 and .70 are considered fair, from .70 to .75 are considered good, and above .75 are considered excellent (Fleiss, 1981). The overall agreement for the four moral dilemmas, .83 , was quite high.

Table 2. Agreement between coders for each moral dilemma for the two categories of justifications, along with Cohen's kappa.
\begin{tabular}{cc}
\hline & \begin{tabular}{c} 
Inter-Observer \\
Reliability
\end{tabular} \\
\hline Dilemma & kappa \\
1 & .84 \\
2 & .71 \\
3 & .89 \\
4 & .90 \\
\hline
\end{tabular}

Overall, these results show that participants tend to prefer omission in order to respect the moral principle of "Do not play God."

\section*{Study 2}

With the first study we have demonstrated that individual preferences for omissions in moral dilemmas are actually influenced by the moral goal of respecting the "Do not play God" principle. But these results raise the question whether participants' justifications of their moral choice reflect their actual reasons for deciding. It is possible that participants simply report a justification that corresponds to their decision, but it is not clear whether the justification preceded and causally influenced their decision (e.g. Haidt, 2001), although Cushman and colleagues (Cushman, Young, \& Hauser, 2006) state that the principles used in judgments are well articulated and reflected in justifications. For these reasons a second study is carried out.

With this second experiment we wanted to further verify whether the preference for the omissions in problems like the trolley dilemma is influenced by the goal of respecting the "natural order", activated by a restriction of one's decisional autonomy.

To this aim, we compared three versions of the trolley dilemma to isolate the effects of 1.authority presence, and 2.closeness, on moral judgments concerning harmful actions. The original version of the trolley dilemma was used as control condition (neutral problem, cf. study 1). We expected that in the "authority" problem participants would choose the inaction options more than participants in the "closeness" and neutral problems. The presence of an authority would indeed limit the decisional autonomy of the protagonist, leading participants to prefer the omission.

\section*{Method}

Subjects
Participants were 105 undergraduate and postgraduate students recruited by advertisements at the University of Rome (Italy) ( 70 females and 35 males). Their ages ranged from 18 to 51 with a mean age of 24.5. All of the participants were thus volunteers and provided informed consent.

\section*{Materials, and procedure}

Subjects responded to one of three versions of the trolley dilemma, in a between-subject design. In each condition, participants received the seven scenarios used in the earlier
experiment (4 moral dilemmas, 3 control dilemmas), in which information about the presence of the "authority" and the "closeness" of the protagonist to the victims were systematically varied. In the "authority" condition ( \(n=45\) ), the moral scenarios presented an authority close to the protagonist (e.g. a policeman, a judge). In the closeness condition ( \(\mathrm{n}=30\) ), in all the moral scenarios the protagonist was close to the potential victims. In the neutral condition ( \(n=30\) ), participants were given the original version of the trolley dilemmas.

In each condition, the order of the seven dilemmas was randomized. Following is an example of the two kinds of moral dilemmas, "authority" and "closeness", presented in the study (translated from Italian). Each version started with the same stem but ends differently:

Start of the dilemma:
You are near a Ferris wheel. It does not work. Just under the wheel, there are five tourists. Suddenly, the wheel starts turning and soon a cabin will kill them. There is no way to warn them and they cannot escape in any way. The only way to save the five tourists is to pull a lever that can change the rotation of the wheel. Unfortunately, there are three people on the other side that would be killed.

The "authority" script continues as follows:
You are in the cabin and close to the lever. You know that the cabin is under video surveillance and that cameras are connected to the police and the security service. Should you pull the lever?

The "closeness" script continues as follows:
You are in the cabin very next to the five tourists and you can see clearly their faces from there. Should you pull the lever?

The text of all seven scenarios is available on the web at www.apc.it. In all conditions, as in the earlier study, each dilemma required participants to indicate which of two courses of action they would take if confronted with such dilemmas (Greene \& Haidt, 2002; Greene, et al., 2004). The total number of inaction choices made by each participant was the dependent variable.

\section*{Results}

As expected, the proportion of scenarios for which participants chose inaction was significantly greater in the "authority" condition ( \(\mathrm{F}(2,102\) ) \(=9.55, \mathrm{p}<0.001, \mathrm{M}=2.47\), \(\mathrm{SD}=1.15\) ), than in the other two experimental conditions ("closeness": \(\mathrm{M}=1.46, \mathrm{SD}=1.3, \mathrm{t}(73)=3.48, \mathrm{p}<.002\); neutral: \(\mathrm{M}=1.43 ; \mathrm{SD}=1.88, \mathrm{t}(73)=3.9, \mathrm{p}<.001)\). No differences were found between the "closeness" and neutral conditions, \(\mathrm{t}(58)=0.1, n . s\).

This result demonstrates that participants' preferences for the inaction depend on the goal of reducing or limiting one's own decisional autonomy, according to the not play God moral principle.

\section*{General Discussion}

The two studies present evidence that moral judgments of harmful acts and omissions are affected by the degree of their interference with the "natural order". According to Sunstein (2005), harmful acts are worse than harmful
omissions because individuals' moral intuitions reflect an overgeneralization of the "Do not play God" principle. In this perspective, omissions or inactions interfere less with the "natural order." Omission generally carries less negative moral weight than commission, since it interferes less with individual's destiny.

In particular, in two studies we demonstrated that the "Do not play God" principle influences individuals' moral preferences when faced with problems like the trolley dilemma, traditionally used in moral psychology to study how people reason when choosing between two morally unacceptable courses of action. In particular, in the first study we demonstrated that participants preferring omission justified this choice according to the "Do not play God" deontological principle. The second experiment demonstrated that the presence of an "authority" lead people to limit their decisional autonomy, thus preferring the inaction, that is what interferes less with a given order ("do not play God" principle).

Our findings may contribute to the explanation of the omission bias, which is defined as the tendency to judge harmful actions as worse or as less moral than equally harmful omissions (inactions) (e.g. Baron \& Ritov, 2004; Spranca, Minsk, \& Baron, 2003).

Several experiments have found that across a variety of moral dilemmas, subjects' judgments about the permissibility of harming an individual align with some principles, which usually distinguish between action and inaction, such as harm intended as the means to an end is worse than harm foreseen as the side effect of a pursuit, or that harm involving physical contact with the victim is worse than harm involving no physical contact.

Sunstein (2005) suggests that harmful acts are generally worse than harmful omissions because according to the "Do not play God" principle, they interferes less with the "natural order." Overall, results from our two preliminary experiments confirm this hypothesis. In Study 1, our participants' justifications for their inaction explicitly refer to the intention of not interfering with destiny. In Study 2, our participants are affected in their moral choice by the presence of an authority, which induce them to choose the inaction, the option that does not modify "the natural order".

The omission bias may thus be better considered as a part of a deontological theory that people tend to approve (Sinnott-Armstrong, Young \& Cushman, 2010; Waldmann, Nagel, \& Wiegmann, 2012), and actively influences both commonsense morality and law, including constitutional law, by treating harmful omissions as morally unproblematic or categorically different from harmful actions (see for example the current debate on euthanasia).
The current studies are to be considered as preliminary studies on this topic. They also present several limitations that call for further investigation. First, participants of both our studies were predominantly female, young, and influenced by Catholic culture. It is possible that our results may not generalize to a broader population. Thus, they may not work well for others that differ in culture and religion.

Further experiments could test whether our results will be replicated for individuals of different in cultures and religions.

\section*{References}

Asch, D., Baron, J., Hershey, J. C., Kunreuther, H., Meszaros, J., Ritov, I., \& Spranca, M. (1994). Determinants of resistance to pertussis vaccination. Medical Decision Making, 14, 118-123.
Baron, J., \& Leshner, S. (2000). How serious are expressions of protected values. Journal of Experimental Psychology: Applied, 6, 183-194.
Baron, J., \& Ritov, I. (2004). Omission bias, individual differences, and normality. Journal of Organizational Behavior and Human Decision Processes, 94, 74-85.
Cushman, F., Young, L., \& Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. Psychological Science, 17, 1082-1089.
Greene, J. D., \& Haidt, J. (2002). How (and where) does moral judgment work? Trends in Cognitive Sciences, 6, 517-523.
Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., \& Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. Neuron, 44, 389400.

Greene, J., Sommerville, R. B., Nystrom, L. E., Darley, J. M., \& Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. Science, 293, 2105-2108.
Greene, J.D., Cushman, F.A., Stewart, L.E., Lowenberg, K., Nystrom, L.E., \& Cohen, J.D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. Cognition, 111, 364-371.
Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment Psychological Review, 108, 814-834.
Haidt, J., \& Baron, J. (1996). Social roles and the moral judgment of acts and omissions. European Journal of Social Psychology, 26, 201-218.
Meszaros, J. R., Asch, D. A., Baron, J., Hershey, J. C., Kunreuther, H., \& Schwartz-Buzaglo, J. (1996). Cognitive processes and the decisions of some parents to forego pertussis vaccination for their children. Journal of Clinical Epidemiology, 49, 697-703.
Royzman, E. B., \& Baron, J. (2002). The preference for indirect harm. Social Justice Research, 15, 165-184.
Singer, P. (1979). Practical ethics. Cambridge, England: Cambridge University Press.
Sinnott-Armstrong, W., Young, L., \& Cushman, F. (2010). Moral intuitions. In J. M. Doris \& Th e Moral Psychology Research Group (Eds.), The moral psychology handbook. Oxford, England: Oxford University Press.
Spranca, M., Minsk, E., \& Baron, J. (1991). Omission and commission in judgment and choice. Journal of Experimental Social Psychology, 27, 76-105.

Spranca, M., Minsk, E., \& Baron, J. (2003). Omission and commission in judgment and choice (p. 31). Philadelphia, PA: University of Pennsylvania.
Sunstein, C. (2005). Moral heuristics. Behavioral and Brain Sciences, 28, 531-573.
Waldmann, M. R., Nagel, J., \& Wiegmann, A. (2012). Moral judgment. In K. J. Holyoak \& R. G. Morrison (Eds.), The Oxford Handbook of Thinking and Reasoning (pp. 364-389). New York: Oxford University Press.
Wroe, A. L., \& Salkovskis, P. M. (2000). Causing harm and allowing harm: a study of beliefs in obsessional problems. Behaviour Research and Therapy, 38, 1141-1162.

\title{
Towards agents with human-like decisions under uncertainty \({ }^{1}\)
}

\author{
Nuno Marques (nunocm@gmail.com) \\ Francisco Melo (fmelo@inesc-id.pt) \\ Samuel Mascarenhas (samuel.fm@gmail.com) \\ João Dias (joao.dias@gaips.inesc-id.pt) \\ Rui Prada (rui.prada@gaips.inesc-id.pt) \\ Ana Paiva (ana.paiva@inesc-id.pt) \\ INESC-ID, Instituto Superior Técnico \\ Av. Prof. Cavaco Silva, TagusPark \\ 2780-990 Porto Salvo, Portugal
}

\begin{abstract}
Creating autonomous virtual agents capable of exhibiting human-like behaviour under uncertainty is becoming increasingly relevant, for instance in multi-agent based simulations (MABS), used to validate social theories, and also as intelligent characters in virtual training environments (VTEs). The agents in these systems should not act optimally; instead, they should display intrinsic human limitations and make judgement errors. We propose a Belief-Desire-Intention (BDI) based model which allows for the emergence of uncertainty related biases during the agent's deliberation process. To achieve it, a probability of success is calculated from the agent's beliefs and attributed to each available intention. These probabilities are then combined with the intention's utility using Prospect Theory, a widely validated descriptive model of human decision. We also distinguish risk from ambiguity, and allow for individual variability in attitudes towards these two types of uncertainty through the specification of indices. In a travelling scenario, we demonstrate how distinct, more realistic agent behaviours can be obtained by applying the proposed model.
\end{abstract}

Keywords: Intelligent agents; Decision making; Cognitive biases

\section*{Introduction}

Uncertainty is a natural part of our world. No one can claim to know everything, no one can predict the future. We deal with uncertainty on our everyday lives and our behaviour is constantly influenced by it, even if we do not always realize it. However, in the context of virtual agents, uncertainty has usually been seen as a problem that the agent must overcome (eg. planning Peot \& Smith, 1992), and thus most existing systems are aimed at achieving optimal agent behaviour under these conditions.

Our approach is different, in which we acknowledge the often sub-optimal, even "irrational" behaviour of humans when confronted with uncertain situations. These decision biases and judgement errors have been extensively studied and are supported by a wealth of empirical evidence (eg. Kahneman \& Tversky, 1979; Camerer \& Ho, 1994). We propose an agent model based on the classical Belief-Desire-Intention (BDI) paradigm, which seeks to integrate in the agent's deliberation process these deviations from rational behaviour.

\footnotetext{
\({ }^{1}\) This work was partially supported by the Portuguese Fundação para a Ciência e Tecnologia under project PEstOE/EEI/LA0021/2011. (INESC-ID multiannual funding) through the PIDDAC Program funds. Additionally, it was funded by the National Project SEMIRA (ERA-Compl/0002/2009), and European Project eCute (ICT-5-4.2 257666) projects.
}

Agents with the aforementioned characteristics can be specially useful for Multi-Agent Based Simulations (MABS) (Davidsson, 2001). In these systems, human behaviour is modelled at the individual (agent) level, and the resulting structure is analysed after it emerges from the agent interactions. Typically, MABS have been used to validate social theories (eg. Davidsson, 2002). The inclusion of uncertainty is of special importance in market simulations, as it strongly impacts the decisions of the agent (Arthur, 1991). From socio-cultural research, the Uncertainty Avoidance dimension of human cultures, identified by Hofstede (Hofstede, 2001), is another example where these agents could be used in the context of MABS. Our solution is also relevant for use in serious games, particularly virtual training environments. As these simulation often focus on social and communication aspects (eg. Johnson \& Valente, 2009; Kim et al., 2009), it is increasingly important to embed the virtual characters with human-like behaviour.

This paper is organized as follows. We start by giving a possible definition of uncertainty and describing Prospect Theory, and follow with work related to ours. Then we present the model, and demonstrate it using an example scenario. Finally we discuss future improvements.

\section*{Background}

In tackling the effects of uncertainty, one should first have an accurate definition of the term. However, this is not an easy task because different research fields or problem approaches use it with different meanings.

One important step is distinguishing uncertainty from the closely related concept of risk. In a decision context, the later refers to choices involving known chances (eg. a spin of a roulette wheel). However, uncertainty arises in a decisions involving personal opinions (eg. betting on what football team will win a game). Moreover, uncertainty has distinct facets (Smithson, 2008): epistemic randomness or risk uncertainty is the subjective counterpart of risk, and is usually represented by subjective probabilities; ambiguity, which results from overlapping beliefs (i.e, strong reasons to believe and not believe) or uncertainty about probabilities (second order uncertainty); and vagueness, reflected by fuzzy statements (eg. "John is tall" - what does "tall" mean?).

The topic of how humans choose (or should choose) under
uncertainty has been extensively studied over the last centuries. Decision making theories which seek to predict the optimal choice, such as the classical Expected Utility theory (EU), are called normative. However, people do not generally obey the axioms of normative theories (some examples of violations are described in the following section). Given our goal of achieving human-like behaviour, we focus on theories seeking to describe how humans actually act. Within these, decision behaviour has been observed to differ when the subject is offered a description of available choices (decisions from description paradigm), versus when he can learn by direct experimentation (decisions from experience paradigm), as shown by Hertwig, Barron, Weber, \& Erev, 2004. As we will see, the solution proposed in this paper assumes that the agent learns by asking and not by experimentation, and therefore we restrict ourselves to the former category.

\section*{Prospect theory}

The most validated descriptive theory of human decision is called Prospect Theory (PT) (Kahneman \& Tversky, 1979; Tversky \& Kahneman, 1992). Some of the decision biases it addresses are:

Framing Effects: there is evidence that the framing of options (in terms of gains or losses) significantly impacts the choices people make (Tversky \& Kahneman, 1986);

Nonlinear preferences: the idea that a risky prospect is linear in outcome probabilities has been proved false, most prominently by the Allais paradox (Allais, 1979);

Source dependence: as demonstrated by the Ellsberg paradox, people's decisions depend not only on the degree of uncertainty but also on its source; this phenomena has been explained both from an ambiguity aversion (people dislike ambiguity, Ellsberg, 1961) and from a competence hypothesis perspective (people prefer a bet on their area of competence when compared to equivalent bet based on objective probabilities, Heath \& Tversky, 1991);

Fourfold pattern of risk: empirical studies indicate that people are generally risk averse for high probability gains and low probability losses, and risk seeking for low probability gains and high probability losses (Tversky \& Kahneman, 1992).

These biases are accounted by PT by assuming a framing phase prior to the actual evaluation; a value function ( \(v\) ) which distort utilities; and a weighting function \((\pi)\) which distorts probabilities. Our focus is on the biases created by these functions, and how to integrate them in the BDI model, as the modelling of framing effects has already been explored in a context similar to ours (Ito \& Marsella, 2011). In PT, the valuation attributed to a prospect (i.e, a gamble) \(f\), which has \(n\) possible outcomes \(x_{i}, i=1 \ldots n\), each with utility \(U_{i}\) and probability \(p_{i}\), is given by:
\[
V(f)=\sum_{i} v\left(U_{i}\right) \pi\left(p_{i}\right)
\]

The choice for the specific value and weighting functions is arbitrary, as long as they obey certain properties. The value function should reflect the effects of diminishing sensitivity (variations in utility are less perceived the further they are from the reference point), and thus be concave for gains and convex for losses (S-shape). Furthermore, it should be steeper for losses than for gains, reflecting the phenomena of loss aversion.

The weighting function transforms a probability, and should also reflect the effects of diminishing sensitivity. However, in this case there are two boundaries ( \(p=0\) and \(p=1\) ), and thus the resulting function is inverse \(S\)-shaped. The curvature of these functions reflect the individual propensity to decision biases, which is usually accounted for by assuming parametrized functional forms.

Our integration of PT in the BDI model, as shown later, is restricted to choices involving prospects with at most two outcomes. Therefore, both Prospect Theory and its more recent development, Cumulative Prospect Theory (CPT) (Tversky \& Kahneman, 1992), coincide. We are presenting the original formulation of the theory. It is also important to note that, although PT is originally based on studies where probabilities were objectively stated, and thus related to decisions under risk, its fundamental properties were also verified in decisions under risk uncertainty (Tversky \& Fox, 1995).

\section*{Related Work}

In Pezzulo's et. al. proposal, measures of ignorance (what the agent does not know), contradiction and uncertainty (difference between opposing beliefs) are computed by the agent, and used in the decision process using custom rules (Pezzulo, Lorini, \& Calvi, 2004).

FAtiMA-PSI (Dias \& Paiva, 2005; Mascarenhas, Dias, Prada, \& Paiva, 2010) is an architecture geared towards the creation of believable virtual characters. It has a strong focus on emotional aspects and human motivations. It already represents some forms of uncertainty, as stochastic action outcomes and estimations of goal success based on past observations. This architecture also provides several parameters which allow an author to define agents with different personalities. However, it does not model unreliable perceptions - the environment is considered completely observable, and the decision process is based on EU theory.

The graded BDI model and abstract architecture (g-BDI, Casali, Godo, \& Sierra, 2009) extends the classical BDI model by allowing uncertainty to be represented in the agent's mental attitudes. g-BDI permits not only uncertain (graded) beliefs, but also desires and intentions. Graded desires correspond to degrees of preference (or rejection) over states of the world, and graded intentions represent the preference over specific ways (plans) to achieve desires. The formalization of the g-BDI allow different contexts to operate each in its own logic. Thus, the belief context (BC), for example, can use probability measures to represent uncertainty.

The Contextually-Based Utility (CBU) model (Ito \&

Marsella, 2011) combines principles of cognitive appraisal theories with decision theoretic notions, with the main purpose of capturing framing effects with greater accuracy. For each possible goal outcome, a contextual utility value is calculated using two salient features: pleasantness, the outcome's intrinsic attractiveness or unattractiveness; and congruence, how much achieving the goal contributes to the agent's expectations. A decision weight is also computed using the outcomes' probabilities. These three measures are transformed by an S-shaped function which models the effects of diminishing sensitivity in relation to a variable reference point, and are then linearly combined to obtain a goal's final valuation.

The work presented above share with ours the purpose of achieving human-like agent behaviours in decisions under uncertainty. However, almost none of them apply a validated descriptive decision theory. The exception is CBU, which is not an agent model in itself. Thus, our approach differs in what we consider fundamental requisites of our solution: 1) capturing widely validated findings on human decision behaviour; and 2 ) being a generally applicable agent model.

\section*{Model}

We chose the BDI model because its folk psychology roots are consistent with our goal of modelling human-like behaviours, and also due to its flexibility and wide application. An overview of the proposed model is shown in figure 1. In the present section, each component is explained in detail.


Figure 1: Model overview

\section*{Uncertainty representation}

This component deals on how to represent uncertainty in the agent beliefs. We used bayesian probability, as it is the most developed model of uncertainty representation. We assume a world defined by a set of crisp (true/false) propositions, represented by upper-case letters, for example \(\mathrm{O}=\) "Hotel is open". Exhaustive and mutually exclusive subsets of propositions are called variables. A probability distribution over the values of each variable forms the agent's belief state, i.e, his opinions on the actual value of propositions. Thus, if \(\Theta=\left\{A_{1}, \ldots, A_{n}\right\}\) is a variable, then \(\sum_{i} P\left(A_{i}\right)=1\), where \(P\left(A_{i}\right)\) denotes the subjective probability of \(A_{i}\) being true.

In order to focus on the behavioural consequences of uncertainty, we made two simplifying assumptions: 1) all variables are conditionally independent; and 2) the agent can only be uncertain about static propositions (propositions whose value never change, because no actions exist with such effects). We expect to address these limitations in future work.

Belief revision The agent can change its opinions by making either a direct observation or by asking questions to other agents. In the latter case, the degree at which the agent believes what he is told depends on the evidence's credibility. We represent the credibility of an evidence provided by source \(i\), asserting a proposition \(A\), as \(\operatorname{cr}\left(\varepsilon_{i}^{A}\right) \in[0,1[\). The value \(c r\) is calculated based on the history of previously received answers from the same source. We follow the method proposed in (Pearl, 1988), which allows Bayes Rule to be applied to uncertain evidence. Assuming independent sources, the agent's beliefs are updated using the formula below:
\[
P\left(H \mid \varepsilon_{i}^{A}\right)= \begin{cases}\left(\operatorname{cr}\left(\varepsilon_{i}^{A}\right)+\frac{1-c r\left(\varepsilon_{i}^{A}\right)}{n}\right) \cdot \frac{P(H)}{P\left(\varepsilon_{i}^{A}\right)} & \text { ifH }=A \\ \frac{1-c r\left(\varepsilon_{i}^{A}\right)}{n} \cdot \frac{P(H)}{P\left(\varepsilon_{i}^{A}\right)} & \text { otherwise }\end{cases}
\]
where \(P(H)\) is the prior probability of value \(H\) in a variable. When new evidence comes which asserts \(A\), the above formula increases the belief in \(A\) while decreasing the belief in the other values of the same variable, such that they still sum to one. Note that a direct observation corresponds to \(\operatorname{cr}\left(\varepsilon_{i}^{A}\right)=1\), and thus always leads to absolute certainty on a variable's value.

\section*{Solutions and doubts}

This section serves as a bridge between the preceding (uncertainty representation) and following (decision process) components, by demonstrating how the agent's beliefs are integrated in the decision process. In the BDI model, the general behaviour of an agent is guided by the active intention - a commitment to achieve a goal and a plan to do it. We call each plan generated by the planning process a solution. We assume that each solution only contains indispensable actions (as usual in classical planning), and therefore if a single action fails, the solution also fails.

The execution of an action, in turn, is dependent on its Action Pre Conditions (APC) validity. APCs are world propositions or their negation. Thus, success in achieving a goal is ultimately dependent on the validity of APCs and, if among the APCs some correspond to uncertain propositions, they possibly invalidate the entire solution and make the goal impossible to attain. Uncertain APCs are what we call doubts, and they are the basic components from which uncertainty related biases will arise (see Figure 2). To distance ourselves from the problem of planning in uncertain environments, during planning doubts are considered valid and thus ignored.

\section*{Decision process}

Within the deliberative process of the agent, our model deals with the evaluation of competing solutions, which essentially


Figure 2: An example of a solution containing two doubts.
correspond to the intentions of the agent - possible paths to commit to and pursue. Prospect Theory (PT), the psychology based model described before, is used during this process. This theory has been shown to accurately predict the choice behaviour of humans in many real life situations, and as such it nicely fits our goal - achieving actual, and not optimal behaviour. The prospects (i.e, the solutions) only have two possible outcomes: success and failure.
Value function We assume a utility of 0 for failing, and a utility of success given by the sum of utility of each individual action within in the solution. Using the status quo as the reference point \((v(0)=0)\), the above definition implies that a solution is always a positive prospect. We applied Tversky and Kahneman two branch power function, but because we only have positive outcomes we can restrict to the positive part (with \(\alpha=0.88\) as estimated in Tversky \& Kahneman, 1992):
\[
v=U^{\alpha}, U \geq 0
\]

Weighting function The weighting function \(\pi(p)\) transforms the beliefs the agent holds about states of the world into the decision weights he actually utilises when making a decision. The probability assigned to the successful outcome corresponds to no invalid doubts at all (as a single doubt failure invalidates the whole solution). Thus, variables being independent, \(p\) is obtained by multiplying the probabilities of all solution doubts being valid. The probability \(p\) is then transformed by the weighting function. We use the function proposed by Wu \& Gonzalez, 1999:
\[
\pi=\frac{\delta p^{\gamma}}{\delta p^{\gamma}+(1-p)^{\gamma}}
\]

In the above function, the two parameters reflect two distinct cognitive biases: \(\gamma\) (estimated as \(\gamma=0.44\) ) represents the impact of diminishing sensitivity to probability variations through the function curvature; \(\delta\) (estimated as \(\delta=0.77\) ) represents the attractiveness to gambles through the function elevation.

Individual variability towards risk uncertainty is allowed via the indirect specification of the \(\delta\) parameter. Staying within the parameter values estimated by \(\mathrm{Wu} \&\) Gonzalez, 1999, we propose to substitute \(\delta\) with \(\delta^{R}\), such that:
\[
\delta^{R}=\delta+(0.5-R)
\]
where \(R \in[0,1]\) is a risk aversion index available to the scenario author. A value of \(R\) close to 1 results in a risk seeking attitude, as the agent will generally overweight probability values; on the other hand, a value close to 0 corresponds to a risk averse attitude, as the probabilities are underweighted.

The resulting functions are shown in figure 3.


Figure 3: Weighting function at different levels of risk aversion \((R)\).

\section*{Ambiguity}

Although PT captures risk biases, ambiguity related biases are not considered. Before discussing them, we present how is ambiguity quantified in the proposed model.

Research in psychology and neuroscience proposes information entropy of possible meanings one attributes to a situation (or the world) as a mathematical quantification of ambiguity (Takahashi, Oono, Radford, \& Others, 2007; Hirsh, Mar, \& Peterson, 2012). Information entropy, also known as Shannon Entropy, measures the amount of "disorder", or uncertainty, associated with a random variable. If \(\theta=\) \(\left\{A_{1}, \ldots, A_{n}\right\}\) is a variable, it's entropy is given by (we use \(e\) as the base of the logarithm):
\[
H(\theta)=-\sum_{i}^{n} p\left(A_{i}\right) \log p\left(A_{i}\right)
\]

In our model, the variables correspond the agent's belief state. The ambiguity being experienced by the agent, however, depends only on the goal under pursuit, i.e, the solution under execution. Thus, it arises from solution doubt variables:
\[
\Delta(S)=\sum_{i} H\left(\theta_{i}\right), \theta_{i} \in\{\operatorname{doubts}(S)\}
\]

Where \(\Delta(S)\) represents the ambiguity of solution \(S\). High levels of ambiguity are caused by ignorance about the actual value of doubts - when the probability distributions are "flat", and/or there are many doubts. Lower levels of ambiguity, on the other hand, are characterized by strong beliefs (probabilities close to either 1 or 0 ) and/or a fewer number of doubts. Note that the amount of ambiguity is independent on what variable values are preferred, and thus distinct from risk.

Existing research seem to indicate that people are averse to this type of ambiguity, i.e, they tend to choose lower entropy
over higher entropy options (Takahashi et al., 2007). In order to represent this effect, we first reduce it to the \([0,1[\) interval:
\[
\Delta^{*}=1-e^{-k \Delta}
\]

Where \(k=0.2\) is a normalizing factor. Then, we introduce ambiguity in the \(\delta^{R}\) parameter defined before, such that:
\[
\delta^{\Delta}=\delta^{R}\left(1-A \Delta^{*}\right)
\]

Where \(A \in[0,1]\) is an ambiguity aversion index. Therefore, values of \(A\) close to 1 reinforce the aversion to solutions with high ambiguity.

\section*{Example scenario}

In this section we demonstrate how the proposed model gives origin to different behaviours, by focusing on preference reversals caused by distinct attitudes towards uncertainty. The scenario being presented is deliberately simple, in order to facilitate the presentation.

Consider an author who wishes to create an uncertain travel scenario. In addition to the agents introduced further below, it contains two other entities: the city of Agra (location of the TajMahal) and the region of Ladakh (home to several beautiful Buddhist monasteries), which will be the available travel destinations. When an agent learns about them (eg. at initialization time), the associated properties are stored in the agent's memory as variables. By default, properties are certain, but uncertainty about specific entities or specific properties may be specified through the agent's configuration file.

In the presented test case, all agents will share the exact same knowledge: they are completely ignorant about finding accommodation in Ladakh, but are almost certain to find it in Agra. This is represented by two different probability distributions over the variables originated by the property accAvailable of Agra and Ladakh, as shown in table 1. The

Table 1: Agent's uncertain knowledge
\begin{tabular}{|lcl|}
\hline \multicolumn{3}{|c|}{ Agra(accAvailable) } \\
\hline Values & \multicolumn{2}{c|}{ Belief } \\
True & 0.9 & Entropy \\
False & 0.3 & 0.3251 \\
\hline \hline \multicolumn{3}{|c|}{ Ladakh(accAvailable) } \\
\hline Values & \multicolumn{2}{c|}{ Belief } \\
True & 0.5 & Entropy \\
False & 0.5 & 0.6931 \\
\hline
\end{tabular}
atomic actions available to the agents in order to achieve their Travel() goal are specified in table 2, which in this case differ only in their utility, as the doubts are equivalent (having a place to sleep). The created solutions are shown in table 3.
\begin{tabular}{lll} 
Actions & \begin{tabular}{c} 
Table 2: Actions \\
Utility
\end{tabular} \\
Doubts
\end{tabular}

In the first case, we show the impact of distinct attitudes towards risk. The solutions are evaluated by Peter, an extremely
\begin{tabular}{lllll} 
& \multirow{2}{*}{ Table 3: Solutions } & & \\
Solutions & Util & Prob & Amb \\
\hline \(\boldsymbol{S}_{1}:\) TravelTo(Agra) \(\rightarrow\) Visit(Agra) & 6 & 0.9 & 0.3251 \\
\(\boldsymbol{S}_{2}: \operatorname{TravelTo(Ladakh)~} \rightarrow\) Visit(Ladakh) & 10 & 0.5 & 0.6931
\end{tabular}
risk averse and ambiguity tolerant agent ( \(R=1 ; A=0\) ), and Sara, who is risk seeking and also ambiguity tolerant \((R=0\); \(A=0\) ). We start with Peter's evaluation:
\[
\begin{gathered}
V\left(S_{1}\right)=v(6) * \pi(0.9)=6^{0.88} *\left(\frac{1.27 * 0.9^{0.44}}{1.27 * 0.9^{0.44}+(1-0.9)^{0.44}}\right)=\mathbf{2 . 0 1} \\
V\left(S_{2}\right)=v(10) * \pi(0.5)=10^{0.88} *\left(\frac{1.27 * 0.5^{0.44}}{1.27 * 0.5^{0.44}+(1-0.5)^{0.44}}\right)=1.61
\end{gathered}
\]

And Sara's evaluation is shown below:
\[
\begin{gathered}
V\left(S_{1}\right)=v(6) * \pi(0.9)=6^{0.88} *\left(\frac{0.27 * 0.9^{0.44}}{0.27 * 0.9^{0.44}+(1-0.9)^{0.44}}\right)=3.72 \\
V\left(S_{2}\right)=v(10) * \pi(0.5)=10^{0.88} *\left(\frac{0.27 * 0.9^{0.44}}{0.27 * 0.5^{0.44}+(1-0.5)^{0.44}}\right)=\mathbf{4 . 2 4}
\end{gathered}
\]

As we can see, Peter avoids the higher utility solution due to its greater risk. On the other hand, Sara is not concerned, and visits Ladakh.

In the second test case, we demonstrate the effects of ambiguity. We use two other agents, both average risk averse. However, John is ambiguity tolerant \((R=0.5 ; A=0)\) while Laura is ambiguity averse ( \(R=0.5 ; A=1\) ). John evaluates the solutions similarly as shown before, resulting in the valuations: \(V\left(S_{1}\right)=3.24\) and \(V\left(S_{2}\right)=\mathbf{3 . 3 0}\), and therefore goes to Ladakh. In Laura's decision, the parameter \(\delta^{A}\) includes the effects of ambiguity as shown below:
\[
\begin{gathered}
\Delta_{1}^{*}=1-e^{-0.2 * 0.3251}=0.0630 \\
\Delta_{2}^{*}=1-e^{-0.2 * 0.6931}=0.1294 \\
\delta_{1}^{\Delta}=0.77 *(1-1.0 * 0.0630)=0.7215 \\
\delta_{2}^{\Delta}=0.77 *(1-1.0 * 0.1294)=0.6703 \\
V\left(S_{1}\right)=v(6) * \pi(0.9)=6^{0.88} *\left(\frac{0.7215 * 0.9^{0.44}}{0.7215 * 0.9^{0.44}+(1-0.9)^{0.44}}\right)=3.17 \\
V\left(S_{2}\right)=v(10) * \pi(0.5)=10^{0.88} *\left(\frac{0.6703 * 0.5^{0.44}}{0.6703 * 0.5^{0.44}+(1-0.5)^{0.44}}\right)=3.04
\end{gathered}
\]

For Laura, the ambiguity associated with travelling to Ladakh is too great to overcome the potential reward, and she prefers to go to Agra. As it can be seen, the effects of ambiguity are independent of risk, given that the agent's only differ in their ambiguity attitude. It should be noted that, although we fixed the uncertainty levels and varied the agent parametrizations, an equivalent preference reversal can be achieved by varying the ambiguity levels while maintaining risk and the indices.

\section*{Conclusion}

In this paper, we proposed an agent model which combines the flexibility and generality of the BDI paradigm with a widely validated descriptive decision model, in order to capture decisions biases in the agent's deliberative process. The two test cases presented in the previous section showed how different parametrizations in agent's attitudes towards uncertainty have distinct behavioural consequences. A rational agent using EU, on the other hand, would always select the
same solution. Furthermore, by considering risk and ambiguity as separate constructs, we expect to better capture human decisions under uncertainty.

Although the modelled risk biases are well grounded on empirical evidence, the effects of ambiguity require further validation. In particular, the magnitude at which entropy causes an aversion effect in a decision context should be further studied, so that a more accurate expression can be used in the calculation of \(\delta^{A}\). We also expect to address the present limitations by possibly integrating a Bayesian Belief Network (Pearl, 1988), allowing for conditional dependences between variables, and using a planner capable of dealing with partially observable domains (eg. Peot \& Smith, 1992), in order to drop the static propositions assumption. Finally, we consider accounting for framing effects as an important step in achieving more realistic decision behaviours, and the ideas behind the CBU model (Ito \& Marsella, 2011) are certainty worth exploring.

\section*{References}

Allais, M. (1979). The so-called allais paradox and rational decisions under uncertainty. Springer.
Arthur, W. B. (1991). Designing economic agents that act like human agents: A behavioral approach to bounded rationality. The American Economic Review, 81(2), 353359.

Camerer, C. F., \& Ho, T. H. (1994). Violations of the betweenness axiom and nonlinearity in probability. Journal of risk and uncertainty, 8(2), 167-196.
Casali, A., Godo, L., \& Sierra, C. (2009). g-bdi: A graded intensional agent model for practical reasoning. Modeling Decisions for Artificial Intelligence, 5-20.
Davidsson, P. (2001). Multi agent based simulation: beyond social simulation. Multi-Agent-Based Simulation, 141155.

Davidsson, P. (2002). Agent based social simulation: A computer science view. Journal of artificial societies and social simulation, 5(1).
Dias, J., \& Paiva, A. (2005). Feeling and reasoning: A computational model for emotional characters. Progress in Artificial Intelligence, 127-140.
Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. The Quarterly Journal of Economics, 643-669.
Heath, C., \& Tversky, A. (1991). Preference and belief: Ambiguity and competence in choice under uncertainty. Journal of risk and uncertainty, 28(1), 5-28.
Hertwig, R., Barron, G., Weber, E. U., \& Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. Psychological Science, 15(8), 534-539.
Hirsh, J. B., Mar, R. a., \& Peterson, J. B. (2012, April). Psychological entropy: A framework for understanding uncertainty-related anxiety. Psychological review, 119(2), 304-20.
Hofstede, G. (2001). Culture's consequences: Comparing values, behaviors, institutions, and organizations across
nations. Sage Publications, Inc.
Ito, J., \& Marsella, S. (2011). Contextually-based utility: An appraisal-based approach at modeling framing and decisions. In Proceedings of the twenty-fifth \(\{A A A I\}\) conference on artificial intelligence, august 7-11, 2011, san francisco, ca, united states (Vol. 2, pp. 60-65).
Johnson, W. L., \& Valente, A. (2009). Tactical Language and Culture Training Systems: using AI to teach foreign languages and cultures. AI Magazine, 30(2), 72.
Kahneman, D., \& Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica: Journal of the Econometric Society, 47(2), 263-292.
Kim, J., Hill, R., Jr, Durlach, P., Lane, H., Forbell, E., Core, M., ... Hart, J. (2009). BiLAT: A game-based environment for practicing negotiation in a cultural context. International Journal of Artificial Intelligence in Education, 19(3), 289-308.
Mascarenhas, S., Dias, J., Prada, R., \& Paiva, A. (2010). A dimensional model for cultural behaviour in virtual agents. Applied Artificial Intelligence, 24(6), 552-574.
Pearl, J. (1988). Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann.
Peot, M., \& Smith, D. (1992). Conditional nonlinear planning. ... conference on Artificial intelligence planning ..., 189-197.
Pezzulo, G., Lorini, E., \& Calvi, G. (2004). How do I Know how much I don't Know? A cognitive approach about Uncertainty and Ignorance. Proceedings of COGSCI ....
Smithson, M. (2008). Psychology ambivalent view of uncertainty. Uncertainty and risk: Multidisciplinary perspectives, 205-218.
Takahashi, T., Oono, H., Radford, M. H. B., \& Others. (2007, January). Comparison of probabilistic choice models in humans. Behavioral and Brain Functions, 3(1), 20.
Tversky, A., \& Fox, C. R. (1995). Weighing risk and uncertainty. Psychological review, 102(2), 269.
Tversky, A., \& Kahneman, D. (1986). Rational choice and the framing of decisions. Journal of business, 251-278.
Tversky, A., \& Kahneman, D. (1992, October). Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and uncertainty, 5(4), 297-323.
Wu, G., \& Gonzalez, R. (1999). Nonlinear decision weights in choice under uncertainty. Management Science, 74-85.

\title{
A Computational Framework for Attentional 3D Object Detection
}

\author{
Germán Martín García and Simone Frintrop \\ \{martin, frintrop\}@iai.uni-bonn. de \\ Institute of Computer Science III, Universität Bonn, 53117 Bonn, Germany
}

\begin{abstract}
We present a computational framework for the detection of unknown objects in a 3D environment. It is based on a visual attention system that detects proto-objects which are improved by iterative segmentation steps. At the same time a 3D scene model is built from measurements of a depth camera. The detected proto-objects are projected into the 3D scene, resulting in 3D object models which are incrementally updated. Finally, environment- and object-based inhibition of return enables to withdraw the attention from one object and switch to the next. We show that the system works well in cluttered natural scenes and can find and segment objects without prior knowledge.
\end{abstract}

\section*{INTRODUCTION}

Object detection is one of the tasks which are easy to solve for humans but hard for machines. Especially unsupervised object detection, i.e., finding all objects in a scene without previous learning, is largely unsolved in machine vision. \({ }^{1}\) However, a system that is able to localize unknown objects in unknown environments is tremendously useful for robotics. For example, a future robot that shall assist in a household must be able to operate autonomously in a new house and is permanently faced with new, unknown objects. Since humans are able to solve such tasks easily, a promising approach for technical systems is to mimic the human visual system. \({ }^{2}\)

In humans as in machines, one of the challenges is to deal with the huge amount of perceptual input. Despite the parallelity of the brain, its capacity is not sufficient to deal with all sensory data in detail and a selection has to take place. Neisser (1967) was the first who proposed a twostage processing of perception that solves this task: first, a pre-attentive process selects regions of interest in parallel, and, second, an attentive process investigates these regions sequentially in more detail. This view has since then widely spread and many psychological theories and models build upon this dichotomy (e.g. Treisman \& Gelade, 1980; Wolfe, 1994). Rensink (2000) has further developed this idea with his coherence theory of attention. It states that the pre-attentive processing determines structures, which he calls proto-objects, that describe the local scene structure of a spatially limited region. After that, focused attention selects a small number of proto-objects which form a coherence field representing a specific object.

Here, we present a computational framework that follows Rensink's idea of proto-objects as pre-processing step for object detection. Our approach generates proto-objects with a

\footnotetext{
\({ }^{1}\) The winner of the latest Semantic Robot Vision Challenge (http://www.semantic-robot-vision-challenge.org) was only able to detect 13 out of 20 objects (Meger et al., 2010), although in this challenge, the target objects were known in advance.
\({ }^{2}\) However, note that our intention is to obtain an improved technical system rather than to mimic the HVS as closely as possible.
}
bottom-up visual attention system (Klein \& Frintrop, 2012) and improves their shape by iterative segmentation steps. In contrast to other attention models, we operate on 3D data from a depth camera and are thus able to obtain 3D object models in space, which are incrementally updated by integrating new perceptual data.

In computational systems based on bottom-up visual attention, the focus of attention is directed to the most salient region in the scene. In order to scan the whole scene, this requires a way to withdraw attention from that region and switch to the next. In human vision, this is performed by inhibition of return mechanisms (IOR) that inhibit the currently attended region (Tipper et al., 1994).

In most computational systems, IOR is implemented by zeroing values in the saliency map (Itti et al., 1998). This is sufficient in static images, but when acting in a 3D world, the correspondence between spatial locations and image regions is required. This affects also the IOR mechanism, since when the perspective of the observer changes or objects are moving, inhibition has to move with them, preventing attention to re-visit the objects directly. This motivates the use of a 3D map that grounds the perceptions in space and enables to maintain a coherent IOR representation over space and time. Corresponding to human vision (Tipper et al., 1994), our IOR mechanism is both object- and environment-based.

The contributions of this paper are threefold. First, instead of operating on 2D images, we perform attention-based object detection on 3D data; this enables us to situate the attention system in a 3D environment, resulting in a coherent representation of objects over time. Secondly, it allows for performing not only an environment-based but also an objectbased inhibition of return mechanism that operates in space and time. Finally, the use of salient blobs instead of only fixation points for initializing the segmentation process lets us bound the amount of perceptual data to be processed.

\section*{Related Work}

Many computational attention systems have been built during the last two decades, first for the purpose of mimicking and understanding the human visual system (survey in Heinke \& Humphreys, 2004), and second to improve technical systems in terms of speed and quality (survey in Frintrop et al., 2010). The general structure of attention systems is based on psychological models such as the Feature Integration Theory (Treisman \& Gelade, 1980) and states that features are computed in parallel before they are fused to a saliency map.

One component of attention systems is the inhibition of return mechanism. While IOR is simple on static images, image sequences introduce the challenge of establishing correspon-


Figure 1: System Overview. The RGB-D camera provides color and depth streams that are processed to obtain proto-objects and a 3D representation of the scene. Here, one proto-object is fixated (1), segmented (2), and projected to the 3D scene (3). The inhibition (5) did not yet take place.
dences between objects over time. In this context, Backer et al. (2001) perform object-centered IOR. However, their approach operates on simple artificially rendered scenes instead of real world data and on 2D images instead of 3D data as we do. Additionally, we combine object-centered and environment-centered IOR to enable both types of inhibition.

Walther and Koch (2006) use an attention system to obtain saliency maps and generate proto-objects inside this map by thresholding. Unsupervised object detection was also tackled by Kootstra and Kragic (2011) who produce saliency maps with a symmetry-based attention system. They use the most salient points as hypothetical centers of objects; these are then provided as seeds to the segmentation process. The figural goodness of the segmentations is evaluated by Gestalt principles. In a robotics context, Meger et al. (2010) search for objects with the mobile robot "Curious George". The robot used a peripheral vision system to identify object candidates with help of a visual attention module. Then, close-up views of these candidates were recorded with a foveal vision system and investigated by a recognition module to identify the object.

\section*{General Structure}

A general overview of the system is depicted in Figure 1. We acquire data with a depth camera that provides color as well as depth information, and is moved around the scene to obtain different viewpoints. The color and the depth information are investigated in two separate processing streams. The color stream determines proto-objects with help of a bottom-up visual attention system (Fig. 1, top), while the depth stream generates a 3D map of the scene (Fig. 1, bottom). The two streams are combined by projecting the proto-objects into the 3D scene. This results in 3D object models that are incrementally updated when new camera frames are available.

The system operates in two behaviors: the saccade behav-
ior and the fixate behavior. When the system starts, it first finds the most salient proto-object (1. in Fig. 1), which is then attended for several frames (fixate behavior), allowing other modules to improve the shape of the attended proto-object by segmentation (2.) and project it to the 3D scene (3.). After fixating an object for a while, the saccade behavior takes over to determine the next focus of attention. This is enabled by object-based and environment-based inhibition of return mechanisms (4.), that inhibit the region of the segmented object \(O\) and the surrounding region \(A\). To maintain a coherent inhibition of return representation, even when moving the camera, the inhibition values are stored within the 3D map data. From its 3D representation, the data can be projected to produce a 2D IOR map (5.), that is used for inhibiting protoobjects in the saliency map. When the attended object is inhibited, a saccade to the next salient proto-object is generated.

\section*{Proto-Object Detection}

We perform object detection in two steps: first, we detect proto-objects in each frame with a visual attention system and second, the extend of the proto-objects is improved by a segmentation step.

\section*{Attention System: Generation of Proto-Objects}

The first step of object detection is the generation of protoobjects with a visual attention system that mimics the preattentive processing stage of the human visual system. Such systems usually investigate several feature channels such as color and orientation in parallel and finally fuse the resulting conspicuities in a single saliency map (Frintrop et al., 2010). The peaks in the saliency map can be interpreted as proto-objects (e.g. Walther \& Koch, 2006). While in human attention, top-down factors also play an important role, such information is not always available in robotics. Therefore, we compute here only the bottom-up attention.


Figure 2: Top left to bottom right: original RGB image; its corresponding saliency map \(S M\); saliency map after adaptive thresholding \(S M^{\prime}\); the \(S M^{\prime \prime}\) map after the final thresholding.

In this work, we use the CoDi system to compute saliency maps (Klein \& Frintrop, 2012). The structure follows the standard architecture of Itti et al. (1998), consisting of intensity, color, and orientation feature channels which belong to the most important features in the human visual system (Wolfe \& Horowitz, 2004). In contrast to other saliency systems, the center-surround contrast is computed with respect to feature distributions; these are approximated by Normal distributions and their distance is quickly computed by the \(W_{2}\) distance (Wasserstein metric based on the Euclidean norm).

To allow the detection of arbitrarily sized salient regions, we perform the computations on 8 different scales. The color channel consists of a red-green and a blue-yellow channel, following the opponent-process theory of human color vision (Hurvich \& Jameson, 1957). The orientation channel computes center surround differences of Gabor filters of four different orientations: \(0^{\circ}, 45^{\circ}, 90^{\circ}, 135^{\circ}\). The saliency map \(S M\) is the result of fusing the color and orientation channels.

To generate the image blobs that correspond to protoobjects, two thresholding operations are performed: first an adaptive thresholding using a Gaussian kernel \({ }^{3}\)
\[
S M^{\prime}(x, y)= \begin{cases}S M(x, y) & : S M(x, y)>T(x, y)  \tag{1}\\ 0 & : \text { otherwise }\end{cases}
\]
where \(T(x, y)\) is the weighted mean of the neighborhood of \((x, y)\). Finally, a binary thresholding is performed on \(S M^{\prime}\) at a percentage of the global maximum saliency value MAX:
\[
S M^{\prime \prime}(x, y)= \begin{cases}S M^{\prime}(x, y) & : S M^{\prime}(x, y)>0.3 \times M A X  \tag{2}\\ 0 & : \text { otherwise }\end{cases}
\]

Fig. 2 shows the saliency map \(S M\) and the thresholded maps \(S M^{\prime}\) and \(S M^{\prime \prime}\) for an example image. On \(S M^{\prime \prime}\) we find the connected components (proto-objects) and compute their average saliency \(\overline{\text { sal }}\). This method provides us with salient blobs instead of only fixation points which determines the center of

\footnotetext{
\({ }^{3}\) We use the adaptiveThreshold function of the OpenCV library: http://opencv.org/
}
fixation as well as the size of the region to use for further investigation. Too small or too big blobs are discarded. If information for the inhibition of objects is already available in terms of a 2D IOR map \(I\) (see below), it is used to inhibit already visited regions. This is done by computing the overlap \(o\) between each blob and \(I\). Finally, the proto-object with the highest value \(\overline{s a l} *(1-o)\) is attended.

Thus, the computational attention system fulfills its two main purposes: first, it directs attention to a region of interest and, second, it bounds the amount of perceptual data to be processed afterwards while ignoring the rest.

\section*{Improving Proto-Objects by Segmentation}

After finding proto-objects, we improve their shape by a segmentation step that bundles parts of the image data. This has a similar effect as grouping mechanisms in human perception that facilitate figure-ground segregation (Wagemans et al., 2012). Such segmentation steps are likely to exist at all levels of human visual processing (Scholl, 2001).

Here, we use the approved GrabCut segmentation (Rother et al., 2004) that was originally proposed for segmenting objects in images with help of user interaction. It takes a rectangle as input, as well as an initialization of pixels with their likelihoods of being object or background. Segmentation is based on the color similarity of neighboring pixels, thus regarding two of the most important factors of perceptual grouping (similarity and proximity). GrabCut performs foreground/background segmentation by iteratively minimizing an energy function. The energy function measures how different each pixel is from the foreground/background model to which it is assigned, as well as from its direct neighbors. It penalizes pixels different from the foreground model to be labeled as foreground as well as labeling pixels as foreground when all its neighbors are background.

The rectangle required for initialization is determined automatically with help of the proto-objects and the information about already detected objects. The pixels of the currently attended proto-object are merged with the information of this object from previous frames (if available). This information can be gathered from the 3D scene representation raycasted to a 2D object map that will be explained later on (cf. Fig. 1). Now, the smallest rectangle \(r\) containing all merged pixels is determined (cf. Fig. 4, top), as well as a rectangle \(r^{\prime}\), obtained by expanding \(r\) 's dimensions by \(10 \%\).

For initializing segmentation, GrabCut requires four possible pixel likelihood values: FG (foreground), BG (background), \(P R_{\_} F G\) (probably foreground) and \(P R_{-} B G\) (probably background). These are obtained by defining three intervals between 0 and the saliency maximum \(\max\) in \(R\) :
\[
L(x, y)= \begin{cases}F G & : S M^{\prime \prime}(x, y) \in\left[v_{3}, \max \right],(x, y) \in R  \tag{3}\\ P R_{-} F G & : S M^{\prime \prime}(x, y) \in\left[v_{1}, v_{3}\right],(x, y) \in R \\ P R_{\_} B G & : S M^{\prime \prime}(x, y) \in\left[0, v_{1}\right],(x, y) \in R \\ B G & :(x, y) \in R^{\prime} \backslash R,\end{cases}
\]
where \(R\) and \(R^{\prime}\) are the sets of pixels contained in rectangles \(r\)


Figure 3: Top: a book as example object. Middle: initialization of GrabCut, the grayscale values correspond to the four possible likelihoods \(F G\) (white), \(P R_{-} F G\) (light gray), \(P R_{\perp} B G\) (dark gray), and \(B G\) (black). Bottom: the segmentation result.
and \(r^{\prime}\) respectively, and \(v_{i}=i \cdot \frac{\max }{4}\) defines each of the interval limits. The likelihoods are corrected by incorporating the information about the current and all other objects. This is done by setting the pixels that correspond to the current object in the 2D object map as \(P R \_F G\), and the ones corresponding to other objects as \(B G\). An example of the initialization values is displayed in Fig. 3. Five iterations of GrabCut produce a binary object mask \(O\) for the attended blob.

\section*{Creating a 3D Scene Map}

While the color image was used to detect proto-objects, the depth data is used to build a 3D map of the scene. This is done with the KinectFusion algorithm \({ }^{4}\) (Newcombe et al., 2011), which builds a 3D map of the environment by integrating multiple range scans from a moving depth camera such as Kinect. It performs two processes in parallel, namely, tracking of the pose of the camera, and registration of the depth scans into a complete scene representation. The result is a 3D scene map consisting of voxels (cf. Fig. 5, right).

To represent the scene at time \(k\), a global truncated signed distance function (TSDF) \(S_{k}(p) \rightarrow\left[F_{k}(p), W_{k}(p)\right]\) is computed by integrating the depth measurements, where \(p \in \mathbb{R}^{3}\) is a point in space, \(F_{k}(p)\) the TSDF value and \(W_{k}(p)\) a weight. The function is discretized in a voxel grid; its zero crossings are points that lie on surfaces. Thus, from the voxel grid, a point cloud can be rendered by choosing the voxels containing zero TSDF values.

\section*{Extended 3D Scene Map}

Our system stores all object information in a 3D structure. It is an extended version of the voxel grid defined in the previous section. For convenience, we will refer to the new voxel grid as \(S_{k}[c]\), where voxel \(c=(x, y, z), x, y, z \in[1 . . V o l]\) and Vol is the number of cells into which the grid is discretized. We extend the \(S_{k}\) function to
\[
\begin{equation*}
S_{k}[c] \rightarrow\left\{F_{k}[c], W_{k}[c], L_{k}[c], L W_{k}[c], I_{k}[c], I W_{k}[c]\right\} \tag{4}
\end{equation*}
\]
where \(F_{k}[c]\) and \(W_{k}[c]\) are the values defined before, \(L_{k}[c], L W_{k}[c]\) are variables that contain object label information, and \(I_{k}[c], I W_{k}[c]\) are IOR related and will be explained

\footnotetext{
\({ }^{4}\) We use the open source implementation available in the Point Cloud Library (http://pointclouds.org/)
}
later on. The 3D information from the voxel grid can at any time be projected to produce a 2 D image containing IOR or object label information (details follow). \({ }^{5}\)

\section*{Generating 3D Object Models}

Now, the 3D object models are created and updated using the binary object mask \(O\) from the segmentation stage. Let us denote the function that maps pixels in the image to voxels in the grid as map : \(p \in \mathbb{Z}^{2}, T \in \mathbb{R}^{4}, D \in \mathbb{Z}^{m \times n} \rightarrow c \in \mathbb{Z}^{3}\), where \(p\) is a pixel, \(T\) the camera pose, and \(D\) a depth image with dimensions \(m \times n\). The pixels in the object mask are mapped to their corresponding voxels in the grid:
\[
\begin{equation*}
\operatorname{map}\left(O, T_{g, k}, D_{k}\right) \rightarrow O^{\prime}=\left\{c: c \in \mathbb{Z}^{3}\right\} \tag{5}
\end{equation*}
\]
where \(g\) is the global frame of reference.
Now it has to be decided which label to assign to the voxels in \(O^{\prime}\). There are two mechanisms corresponding to the fixate and saccade behaviors of the system. During the fixate behavior, the label of the currently attended object is used. When the saccade behavior selects a new focus of attention, it performs as follows. On the set of voxels \(O^{\prime}\) corresponding to the new proto-object, we extract the current labels \(>0\) : \(L a b=\left\{L_{k}[c]: L_{k}[c]>0, c \in O^{\prime}\right\}\). We find the most frequently occurring label \(l\) in \(L a b\). If less than \(5 \%\) of the voxels are labeled, we assign \(l\) a new value corresponding to a newly detected object. The value of \(l\) is now used to update the voxels contained in \(O^{\prime}\). This simple scheme lets us integrate the overlapping segmentations of different views of the same objects in the 3D map.

To be flexible against wrong segmentations or overlapping objects, weights are assigned to the labels. Every time the same label is assigned to a voxel, its label weight \(L W_{k}\) is incremented. If a voxel is updated with a different label, the weight is decremented. Eventually it could reach 0, resulting in an unlabeled voxel. This mechanism lets us incrementally build the object representations with a certain tolerance to failure; furthermore, by thresholding the label weight we can specify the degree of confidence in our object representations that we want for rendering the labeled point cloud. In our experiments, we used \(L W_{k}=5\), meaning that a voxel has to be assigned to a specific object at least 5 times to be considered for this object.

\section*{3D IOR Map}

After fixating an object for several frames, the object must be inhibited to enable the next saccade. To allow a coherent IOR over time, we store the inhibition values within the 3D voxel grid: \(I_{k}[c]\) is a binary flag denoting whether that voxel shall be inhibited and \(I W_{k}[c]\) is a weight that determines how long the effect shall take place. Having IOR information in 3D coordinates lets us generate 2D IOR maps \(I_{k}\) from the required camera poses throughout the sequence.

\footnotetext{
\({ }^{5}\) In (Newcombe et al., 2011), the TSDF function is raycasted, given a camera pose, to generate a depth map prediction. Using this method in our extended TSDF function means we can generate 2D IOR or object label maps for every new pose of the camera.
}


Figure 4: Table Top sequence at different points in time (columns). From top to bottom: (i) image of the scene with currently attended object (blue rectangle); (ii) the saliency map and the segmented part from the currently attended object; (iii) inhibition of return maps; white: object-based IOR, gray: environment-based IOR; (iv) the 3D scene map including detected objects

According to human vision, we use two types of IOR mechanisms: environment-based and object-based IOR (Tipper et al., 1994). The latter comes intuitively from the segmented object mask \(O\). The environment-based IOR is initialized by the regions close to the object but not on the object, i.e., from a so called attended mask \(A=R^{\prime} \backslash O\). The two masks are mapped as in the previous section to obtain their respective voxel sets \(O^{\prime}\) and \(A^{\prime}\). For every voxel \(c\) in \(O^{\prime}\) and \(A^{\prime}\), its weight \(I W_{k}[c]\) is incremented. When it reaches a certain threshold, the IOR flag \(I_{k}[c]\) is activated. The weight of all not considered voxels is decremented. If a weight eventually reaches 0 , the IOR flag is reset to 0 as well.

\section*{Evaluation}

To evaluate our system we recorded two video sequences in an office environment with an RGB-D camera that provides depth as well as color information. The first sequence shows a setting of objects on a table top (cf. Fig. 4). The complexity of this setting corresponds to the complexity of scenes in current state of the art benchmarks and papers on unsupervised object detection in machine vision (cf. Meger et al., 2010; Kootstra \& Kragic, 2011). However, the real world can be much more complex. Therefore, we recorded a second sequence, that shows a very cluttered setting (Fig. 5). Both settings were recorded turning the camera so that the scene was observed from different viewpoints (cf. Fig. 1). \({ }^{6}\)

Fig. 4 illustrates several steps of our approach at different time points. First, the book was attended (fixate behavior).

\footnotetext{
\({ }^{6}\) Videos of the complete sequences as well as the resulting 3D representations can be found at http://vimeo.com/cogbonn/
}

After fixating it for several frames, the region is inhibited (3rd row) and the attention switches to the next proto-object (saccade behavior). This proto-object consists of two real objects (cup and tea box) since these objects are overlapping from this point of view and have similar saliency. The procedure continues, until all objects on the table have been detected.

For the second sequence, we present for space reasons only the resulting 3D map with detected objects (Fig. 5, right). Here, the approach finds 19 objects after 438 frames ( \(\sim 13\) sec ). More objects could be found by longer observing the sequence, but some would be missed, e.g., due to high similarity to the background, and no current computer vision system would be able to find all objects without pre-knowledge in such a complex setting. Note that several of the "objects" still have proto-object characteristics, meaning that they show parts of objects (handle of dishwashing brush (6), bottom of coffee machine (18)) or clusters of objects (tea boxes (11)). Such semantic ambiguities could only be resolved by a recognition system that investigates the attended regions in more detail, or by a robot that interacts with objects and decides on objectness depending on the connectivity of object parts.

To evaluate our system quantitatively, we measure how precisely the detected objects were segmented. For this, the points in the 3D map corresponding to objects were manually labeled to serve as ground truth. We generally denote the ground truth of each object as \(G\), and the 3D points of the object detected by our system as \(S\). We measure the precision \(p\) and recall \(r\) of the detected objects with respect to the ground truth as \(p=(S \cap G) / S\), and \(r=(S \cap G) / G\). The values are shown in Tab. 1 and Fig. 5. It can be seen that the

\begin{tabular}{|l||l|l|l|l|l|l|l|l|l|l|}
\hline object & \(\mathbf{1}\) & \(\mathbf{2}\) & \(\mathbf{3}\) & \(\mathbf{4}\) & \(\mathbf{5}\) & \(\mathbf{6}\) & \(\mathbf{7}\) & \(\mathbf{8}\) & \(\mathbf{9}\) & \(\mathbf{1 0}\) \\
\hline \hline precision & 93 & 69 & 92 & 99 & 62 & 52 & 90 & 60 & 100 & 99 \\
\hline recall & 40 & 43 & 28 & 40 & 61 & 28 & 36 & 36 & 21 & 37 \\
\hline \hline object & \(\mathbf{1 1}\) & \(\mathbf{1 2}\) & \(\mathbf{1 3}\) & \(\mathbf{1 4}\) & \(\mathbf{1 5}\) & \(\mathbf{1 6}\) & \(\mathbf{1 7}\) & \(\mathbf{1 8}\) & \(\mathbf{1 9}\) & \\
\hline \hline precision & 23 & 90 & 83 & 98 & 91 & 99 & 100 & 89 & 100 & \\
\hline recall & 47 & 40 & 35 & 39 & 31 & 30 & 8 & 1 & 3 & \\
\hline
\end{tabular}

Figure 5: Coffee Machine sequence. Left: color image. Right: 3D scene map with detected objects (numbers denote labels). Bottom: precision/recall values in \%
\begin{tabular}{|l||l|l|l|l|l|l|}
\hline object & Book & Cup & Cereals Box & Car & Sponge & Pot \\
\hline \hline precision & 99 & 55 & 98 & 99 & 97 & 94 \\
\hline recall & 64 & 62 & 53 & 54 & 56 & 9 \\
\hline
\end{tabular}

Table 1: Table Top sequence: precision/recall values in \% (cf. Fig. 4).
precision values are mostly very good (more than \(90 \%\) for 17 out of 25 objects), that means that only few voxels were accidentally assigned to an object. A bad value usually indicates that a cluster of objects was detected and compared with separate objects in the ground truth (e.g. objects 5 and 11). The recall values are lower, meaning that often not all of the voxels that belong to an object were detected. In the future, this can be improved by additional post-processing steps based on grouping mechanisms for figure-ground segregation.

\section*{Conclusion}

We have presented a flexible framework for the detection of unknown objects in a 3D scene. Unlike other approaches, the system uses depth values additionally to a color image of a scene and is thus able to generate 3D object models that are incrementally updated when new information is available. All perceptual data is spatially grounded and thus consistent over different viewpoints. The results show that the algorithm is able to detect many objects in scenes with high clutter, without using any prior knowledge about the type of objects.

Applying attention mechanisms in space and time introduces new challenges, for example the question of how and when to switch attention between salient regions. We introduced an environment- and object-based inhibition of return mechanism that addresses this problem by using the information from the 3D environment and object models for inhibition.

\section*{References}

Backer, G., Mertsching, B., \& Bollmann, M. (2001). Dataand model-driven gaze control for an active-vision system. IEEE Trans. on PAMI, 23(12).
Frintrop, S., Rome, E., \& Christensen, H. I. (2010). Computational visual attention systems and their cognitive foundations: A survey. ACM Trans. on Applied Perception, 7(1).
Heinke, D., \& Humphreys, G. W. (2004). Computational models of visual selective attention. A review. In Connectionist models in psychology. Psychology Press.
Hurvich, L., \& Jameson, D. (1957). An opponent-process theory of color vision. Psychological review, 64(6).
Itti, L., Koch, C., \& Niebur, E. (1998, Nov). A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on PAMI, 20(11).
Klein, D. A., \& Frintrop, S. (2012). Salient pattern detection using W2 on multivariate normal distributions. In Proc. of DAGM-OAGM. Springer.
Kootstra, G., \& Kragic, D. (2011). Fast and bottom-up object detection, segmentation, and evaluation using Gestalt principles. In IEEE Int'l Conf. on Robotics and Automation.
Meger, D., Muja, M., Helmer, S., Gupta, A., Gamroth, C., Hoffman, T., et al. (2010). Curious george: An integrated visual search platform. In Canadian conference on computer and robot vision.
Neisser, U. (1967). Cognitive psychology. New York: Appleton-Century-Crofts.
Newcombe, R. A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A. J., et al. (2011). KinectFusion: Realtime dense surface mapping and tracking. In Proc. of IEEE Int'l Symposium on Mixed and Augmented Reality.
Rensink, R. A. (2000). The dynamic representation of scenes. Visual Cognition, 7, 17-42.
Rother, C., Kolmogorov, V., \& Blake, A. (2004). GrabCut: Interactive foreground extraction using iterated graph cuts. ACM Trans. Graph., 23, 309-314.
Scholl, B. J. (2001). Objects and attention: the state of the art. Cognition, 80, 1-46.
Tipper, S. P., Weaver, B., Jerreat, L. M., \& Burak, A. L. (1994). Object-based and environment-based inhibition of return of visual attention. J. of Experimental Psychology.
Treisman, A. M., \& Gelade, G. (1980). A feature integration theory of attention. Cognitive Psychology, 12, 97-136.
Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., et al. (2012). A century of Gestalt psychology in visual perception: I. perceptual grouping and figure-ground organization. Psychological Bulletin.
Walther, D., \& Koch, C. (2006). Modeling attention to salient proto-objects. Neural Networks, 19(9), 1395-1407.
Wolfe, J. M. (1994). Guided search 2.0: A revised model of visual search. Psychonomic Bulletin and Review, l(2).
Wolfe, J. M., \& Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? Nature Reviews Neuroscience, 5, 1-7.

\title{
Effect of Positive and Negative Instances on Rule Discovery: Investigation Using Eye Tracking
}

\author{
Miki Matsumuro (muro@cog.human.nagoya-u.ac.jp) \\ Kazuhisa Miwa (miwa@is.nagoya-u.ac.jp) \\ Graduate School of Information Science, Nagoya University, Fro-cho, Chikusa-ku, Nagoya, Japan
}

\begin{abstract}
This study investigates how participants reject an initial rule when they face positive and negative instances of an initial rule. Using eye movement data, we analyzed a perspective that indicated the type of rules that participants consider. The results show that, only in negative instances, participants considered rules from the perspective that they used for finding and confirming the initial rule. We concluded that, when participants face negative instances, they tried to change the initial rule peripherally to explain them. This appeared in the form of a longer consideration from an initial perspective in negative instances.
\end{abstract}

Keywords: Rule discovery; eye tracking; anomalous data; attentional learning; perspective

\section*{Introduction}

Finding regularities is one of the most important activities not only in science but also in many aspects of daily life. People can find regularities in their daily experiences or observed data and use them for predictions and decision making. However, if a rule that they find at first (i.e., an initial rule) is incorrect, they need to reject it on the basis of their observed data and start considering a new one. We focus on such a situation, specifically one in which people face negative evidence to the initial rule. In particular, we investigate the period in which both positive and negative instances appear.

The process of finding and rejecting the initial rule consists of the following three phases. The first phase is the "initial phase" in which participants find and use the initial rule. In this phase, participants always observe positive instances. Our study focuses on the second phase called the "transition phase." In this phase, participants face both positive and negative instances. The third phase, the "post-transition phase," follows the transition phase, and it includes the time in which participants try to find a rule that is valid for all instances. Almost all instances in this phase are negative to the initial rule; hence, the initial rule is absolutely rejected.

\section*{Perspective and Eye Tracking}

Using a relatively easy rule discovery task, such as Wason's 2-4-6 task, many researchers have shown that disconfirmation is effective to reject the initial rule (e.g., Wason, 1960). They obtained these results from analyses based on the best guess regarding the rule collected by verbal or written reports. However, this method does not reveal thoughts before participants reveal their best guess. This period contains many "unexpressed thoughts" that cannot be reflected in the best guess, such as thoughts that are abandoned before they become obvious rules or that do not yet reach obvious rules. Compared with other phases, it is difficult to make one's thoughts into
the form of a rule in the transition phase; therefore, more unexpressed thoughts appear during this phase. Although unexpressed thoughts are crucial information regarding the process in the transition phase, a traditional analysis of verbal reports of the rule has difficulty in interpreting them because they do not take the form of a rule.

We use the following two new approaches to investigate unexpressed thoughts: the analysis of "perspective" and "eye tracking." The perspective constrains the type of rules that people consider. There are an incredible number of available rules; therefore, participants cannot simultaneously consider all rules. By deciding upon the perspective, participants consider only rules matching their perspective, which makes them able to consider an adequate number of rules. The perspective includes the function or role of rules and the factor used for rules. For example, Klahr and Dunbar (1988) represented hypotheses as frames and clustered them on the basis of the role of frames in scientific reasoning. Participants in their experiment considered the hypotheses in a cluster that had an identical role in their mind. The information provided by the analysis of the perspective about the unexpressed thoughts in the transition phase is more useful than that provided by the analysis of the reported best guess.

Lien and Lin (2011) investigated the effects of perspective on rule discovery performance using Wason's 2-4-6 task. They found that participants who could discover the rule changed their perspective more frequently than those who could not. However, they analyzed only the reported rules, which have a problem as mentioned above. Haverty, Koedinger, Klahr, and Alibali (2000) showed that the crucial factor for rule discovery was the ability to judge which type of rules participants should pursue, that is, which perspective they should use. The protocol analysis used in their study has been used in many preceding studies on rule discovery. Although it provides a detailed process of rule discovery, it still has two problems for the purpose of this study.

The first problem is that participants mention only their conscious thoughts. The verbal protocol cannot always capture unexpressed thoughts because participants are not always conscious of them. The second problem is that protocol data is coarse-grained on the time scale. In general, the speed of thinking is faster and more fluent than that of verbalization. In the transition phase, short thoughts that come and go in the mind appear frequently. Such short thoughts are difficult to put into words. For these reasons, we use eye tracking, which continuously gives more fine-grained data. Recently, eye movement data has been utilized in many studies, such as
category learning and insight problem solving, and has produced good results (e.g., Knoblich, Ohlsson, \& Raney, 2001; Rehder \& Hoffman, 2005).

\section*{Hypotheses}

We define a perspective needed to find an initial rule as an "initial perspective." The hypotheses in this study are as follows. First, in the initial phase, participants mainly consider rules from the initial perspective. The preceding studies on category learning showed that once participants learned a rule, they focused only on the relevant features and ignored irrelevant features (e.g., Rehder \& Hoffman, 2005). Similarly, once participants found a valid rule (i.e., an initial rule), they focused only on the relevant perspective (i.e., the initial perspective) and ignored irrelevant perspectives.

The transition phase has the following two hypotheses: the "shift hypothesis" and "retain hypothesis." In the shift hypothesis, we predict that participants will shift their focus from the initial perspective to other perspectives as soon as they face only one negative instance. According to Popper, a hypothesis can be normatively falsified by just one negative instance. Using Wason's 2-4-6 task, many studies have shown that one disconfirmation is enough to reject the initial rule. If the shift hypothesis is supported, participants will consider rules from the initial perspective to be the same as or shorter than those from other perspectives in the transition phase. The second retain hypothesis suggests that focus on the initial perspective will be retained. In this case, participants will continue to consider rules from the initial perspective rather than those from other perspectives in the transition phase.

There are two possible explanations for retaining consideration from the initial perspective. First, participants use the initial perspective for peripheral changes of the initial rule in order to adapt it to negative instances; for example, to add an additional or exceptional rule. Chinn and Brewer (1993) showed that participants were likely to hold onto their theories by reinterpreting data, using the peripheral theory change, and so on. In addition, Dunbar (1995) demonstrated that participants did not often change their hypotheses even if they faced inconsistent evidence. The second explanation is based on attentional learning. The attentional learning model (e.g., EXIT; Kruschke, 2003) explains eye movement data in category learning according to the following process. First, participants learn the relevant features that should be focused on and irrelevant features that should be ignored. Positive feedback to their learned rule reinforces their focus on the learned features. On the other hand, if they face negative feedback, they shift their focus from the learned relevant features to other features. On the basis of this model, the participants in this study would also shift their focus from the learned initial perspective to other perspectives when they face negative instances. In contrast, when they face positive instances, they again focus on the initial perspective; as a result, consideration from the initial perspective is observed.

We can judge which explanation is accepted from how par-
ticipants consider rules in each negative and positive instance. If the "peripheral change" explanation is accepted, participants consider rules from the initial perspective when they face negative rather than positive instances. The initial rule is not valid in negative instances; therefore, participants need to consider modified rules from the initial perspective to improve the initial rule. On the other hand, when participants observe positive instances in which the initial rule is valid, they do not need much consideration from the initial perspective. If the "attentional learning" explanation is accepted, participants consider rules from the initial perspective when they face positive rather than negative instances. Observing positive instances reinforces the focus on the initial perspective because the initial rule is valid in these instances. On the other hand, negative instances shift participant's focus from the initial perspective to other perspectives.

\section*{Task}

We created a new rule discovery task in which a fixation area and perspective correspond. Figure 1 shows an example screenshot of the task display corresponding to one instance. One instance consists of three panels (arrow, compass, and number) and eight letters (a-h) in the center of the display. These eight letters were arranged in a circle with letter \(a\) taking the 12 o'clock position. One of these letters was displayed in each instance (we call it the "target letter"). Participants were asked to find a rule determining which panel related to the target letter and how the objects in the related panel determined it through observations of some instances. We told them in advance that only one of the three panels was related to the target letter and that the other two panels had no relationship.

We gave a different function to each of the three panels. In the number panel, two numbers from 1 to 4 were displayed. Each number indicated the order of each letter, such as 1 was \(a\) and 2 was \(b\). In the arrow panel, two circular arrows from \(45,90,135\), and \(180^{\circ}\) were displayed. The angle of each arrow indicated the degree of a rotation from the starting letter


Figure 1: Example screenshot of the task display used in this study.
to certain letter. For instance, if the participant set \(a\) as a starting point, a \(45^{\circ}\) arrow indicated \(b\) and a \(90^{\circ}\) arrow indicated \(c\). In the compass panel, two directions from north, east, south, and west were displayed in Chinese characters. Each direction of the compass panel indicated the direction on a map, that is, north was the upper one \((a)\) and south was the lower one (e). The objects within each panel were allowed to be combined. Each panel with a different function corresponded to a different perspective; therefore, the observation times of each panel corresponded to the amount of investigation from each perspective.

We prepared two rules that determined the relationship between the target letter and panel. One was the "initial rule" discovered in the initial phase. In this rule, the target letter is the letter moved from \(a\) by a sum of the angles of two arrows on the arrow panel. For example, when the arrows are at 135 and \(180^{\circ}\) (Figure 1), the target letter \(h\) is identified by moving from \(a\) (at 12 O'clock) by the sum of the two arrows' angles \(\left(135+180=315^{\circ}\right)\). The other rule is the "target rule" that is valid throughout the task. In this rule, the target letter is the letter in the opposite (replace north with south) position to the position of the letter indicated by a combination of two directions on the compass panel. For example, with west and south directions (Figure 1), the target letter \(h\) 's position was indicated by combining two directions (west and south \(=\) south-west) and then replacing south with north (north-west). The initial perspective needed to find the initial rule was the arrow perspective, and the target perspective needed to find the target rule was the compass perspective.

The task consisted of eleven blocks. A block consisted of five observations of an instance and four tests. In an observation, participants considered rules by observing such an instance, as exemplified in Figure 1. By clicking on the NEXT button or after a certain number of seconds, the next instance was presented. Participants were not permitted to return to prior instances. After observing five instances, participants announced the rule which they found and then started the test. In the test display, three panels and the buttons of each of the eight letters were presented. Participants predicted the target letter and selected the button of the letter. The next problem was presented by clicking on the button or after 15 sec . Four problems were conducted in total.

We separated eleven blocks into the following three phases: the first four blocks were the initial phase, the next three blocks were the transition phase, and the final four blocks were the post-transition phase. In the initial phase, only positive instances that supported both the initial and target rules were presented. All participants were expected to find the initial rule and focus on the initial perspective. In the transition phase, both the positive and negative instances to the initial rule are presented, whereas the target rule is valid in all these instances. In the post-transition phase, all instances were positive to the target rule and negative to the initial rule. In this phase, we expected that participants would try to find the target rule. All participants completed eleven blocks, even
if they found the target rule.

\section*{Predictions}

On the basis of our hypotheses, we made the following predictions. First, in the initial phase, the observation time of the arrow panel corresponding to the initial perspective would be longer than those of the other panels. In the transition phase, we expected different result for each hypothesis. If the shift hypothesis was approved, the difference of the observation time between the three panels would disappear. If the retain hypothesis was approved, the observation time of the arrow panel would be longer than those of the other panels. Furthermore, when the retain hypothesis was supported, there were two predictions based on the two explanations for observations of the difference. The prediction derived from the peripheral change explanation was that the tendency of the longer observation time of the arrow panel is more significant in negative instances than that in positive instances. On the other hand, if the attentional learning explanation is accepted, the tendency of the longer observation time of the arrow panel is more significant in positive instances than that in negative instances.

\section*{Experiment 1}

Hereafter, when we mentioned the positive and negative instances, they were defined according to the initial rule. The primary purpose of Experiment 1 was to confirm that the initial rule was rejected progressively in the transition phase, as expected. For this purpose, we compared two types of conditions: one "divided" condition in which the transition phase was removed, and two "mixed" conditions in which the transition phase was conducted as explained in the task section. The second purpose was to investigate whether the more negative instances the participants were given, the faster they rejected the initial rule in the transition phase. For this purpose, we manipulated the ratio of negative instances in the transition phase in the two mixed conditions.

In the divided condition, to remove the transition phase, the post-transition phase followed immediately after the initial phase; namely, after the initial phase ended, all presented instances switched into negative. This manipulation was predicted to make participants reject the initial rule drastically after the initial phase because there was no transition phase in which both positive and negative instances were presented. On the other hand, in the two mixed conditions, the initial rule was rejected progressively in the transition phase. The two mixed conditions consisted of a "mixed-increase" condition in which the ratio of negative instances increased in the transition phase and a "mixed-few" condition in which one negative instance appeared in a block. If the ratio of negative instances affects the timing of rejecting the initial rule, the rejection in the mixed-few condition would occur slower than that in the mixed-increase condition.

\section*{Method}

Participants Ninety undergraduates participated in Experiment 1. Each participant was assigned to one of the three conditions.

Task and Procedure We manipulated the amount of negative instances in blocks 5, 6, and 7. In the divided condition, only negative instances were presented after the second trial of block 5. The number of negative instances increased from one to three in the mixed-increase condition and was always one at each block in the mixed-few condition. Before the second trial in block 5, all instances were positive in all conditions. The maximum observation time of each instance was 35 sec .

The rule discovery task was conducted one at a time in a classroom. Each participant engaged in the task individually at his or her own pace using a computer terminal. Before starting the task, participants were instructed about the task and learned the functions of each panel through practice. Participants used a keyboard to describe the rule they found.

\section*{Results}

Seven participants who did not follow the instructions were excluded from analyses. We categorized the reported rules in each block into the following three categories: the initial rule, the target rule, and others or none. In Experiment 1, we focused on the rejection of the initial rule; therefore, we analyzed the discoverer of the initial rule who described only the initial rule in blocks 3 and 4. Twenty-four non-discoverers were excluded from analyses. The data of 24 participants in the divided condition, 12 participants in the mixed-increase condition, and 22 participants in the mixed-few condition were used for analyses.

Figure 2 shows the ratio of participants in each condition whose reported rule was categorized in the initial rule in each block. As expected, in the divided condition, a drastic decline was observed in block 5 . Only two participants ( \(9.09 \%\) ) stated the initial rule in block 5 and none stated it in blocks 6 and 7. The rejection of the initial rule was slower in the two mixed conditions than that in the divided condition. The ratio of participants who stated the initial rule in block 5 was high and nearly the same in the mixed-increase condition ( \(91.66 \%\) ) and mixed-few condition ( \(86.36 \%\) ). In block 6, this ratio was


Figure 2: Ratio of participants in each condition whose reported rule was categorized in the initial rule in each block.
moderately lower in the mixed-increase condition (16.66\%) than that in the mixed-few condition \((31.81 \%)\). More than \(15 \%\) of participants kept the initial rule in block 7 in both conditions.

The same result was observed in the tests. The use of the initial rule to identify the target letter decreased more slowly in the two mixed conditions than that in the divided condition. The ratio of participants who stated the target rule started increasing in block 5 and reached around \(50 \%\) in block 11 . There was no difference in the three conditions.

\section*{Discussion}

In the results of Experiment 1, we observed a drastic decline in the number of participants who identified the initial rule only in the divided condition. This means that, when only negative instances were presented after a certain point, the initial rule was rejected without taking any time to consider it. In contrast, we observed progressive declines in the two mixed conditions. These results support our prediction that, in the transition phase in which both positive and negative instances were presented, participants would reject the initial rule progressively, considering whether they should reject it. Participants in the mixed-few condition tended to reject the initial rule more slowly than those in the mixed-increase condition. This means that the more negative instances participants were given, the faster they rejected the initial rule. The results of Experiment 1 show the validity of our task.

\section*{Experiment 2}

The purpose of Experiment 2 was to investigate how participants considered rules from each perspective in the transition phase using the task used in Experiment 1. Using eye movement data, we tested the two hypotheses: the shift and retain hypotheses.

\section*{Method}

Participants Twenty undergraduates participated in Experiment 2.

Task and Apparatus We conducted the rule discovery task two times, identified as task 1 and task 2. Task 1 consisted of three blocks and had no rule, and the combinations of the objects on each panel and the target letter in the task were randomly created. The purpose of task 1 was to confirm that there was no inherent difference between the observation times of each panel. Task 2 had the same format as the mixed-increase condition in Experiment 1. The maximum observation time of each instance was 40 sec .

Participants' eye movements were recorded using a Tobii T60 eye tracker at 60 Hz . We presented the task display on a 17 -in monitor with a resolution of \(1280 \quad 1024\) pixels. Participants were seated approximately 60 cm away from the monitor. The size of visual angle on the panels was approximately \(7.57^{\circ} \quad 5.05^{\circ}\).

Procedure Experiment 2 was conducted individually. First, participants were instructed about the task and learned the


Figure 3: Ratio of participants who announced each rule.
functions of each panel through practice. Task 1 was conducted before task 2. Participants were asked to find the rule but were not instructed that there was no rule. The procedure for each task in Experiment 2 was the same as that in Experiment 1, except that the best guess of the rule was announced verbally. Before starting each block, a calibration for recording eye movement was performed. All eye movements throughout each task were recorded.

\section*{Results}

Four participants who did not discover the initial rule were excluded from analyses. Two participants whose eye movement data were not recorded more than \(50 \%\) of the time and one participant whose fixations on the lower panel were not recorded correctly were excluded from analyses.
Reported Rules Figure 3 shows the ratio of participants who announced the initial or target rule in each block. As observed in Experiment 1, the ratio of participants who announced the initial rule declined progressively from blocks 5 to 7. The ratio of participants who announced the target rule was also similar to that of Experiment 1: this number started increasing in block 5 before reaching \(46.15 \%\) in block 11 .
Eye Movement We calculated each participant's observation time of each panel. Then, we calculated the mean observation time of each panel in task 1, the initial phase, and the transition phase to investigate two hypotheses. We did not use the data in block 1 in which not all participants announced the initial rule for calculating the mean observation time in the initial phase. First, we conducted a one-way within subject ANOVA (panel: arrow, compass, and number) for the observation time in task 1 to confirm that there was no coherent difference between three panels. A significant main effect of the panel was not observed \((F(2,24)<1)\). This result means that the different contents of each panel or functions of rules did not affect observation time.

To investigate which hypothesis, the shift or retain hypothesis, was supported, we conducted a 2 (phase: initial and transition) 3 (panel: arrow, compass, and number) within subject ANOVA for mean observation times (Figure 4). The main effect of the phase factor \((F(1,12)=\) \(17.351, p=.001\) ) and the panel factor reached significance \((F(2,24)=13.174, p<.001)\). The interaction of the two factors was not significant \((F(2,24)<1)\). To confirm the difference in the observation times in each phase, we conducted a


Figure 4: Mean observation time of each panel in each phase of the task 2 (bars show standard errors).
planed one-way within subject ANOVA (panel: arrow, compass, and number) for the observation times of each phase. In both the initial and transition phases, a significant main effect of the panel factor was observed (initial: \(F(2,24)=\) 23.367, \(p<.001\); transition: \(F(2,24)=5.528, p=.011)\). We observed the same tendency in both phases. The observation time of the arrow panel was significantly longer than those of the compass panel (initial: \(p<.001\); transition: \(p=.059\) ) and the number panel (initial: \(p=.002\); transition: \(p=.045\) ). There was no significant difference between the observation times of the compass and number panels (both phases: \(p s=1.000\) ). These results support the retain hypothesis that participants focused on the initial perspective; therefore, we continued the analyses to examine two explanations for the retain hypothesis.

We calculated the observation times for each positive and negative instance in all instances from the second trial in block 5 to the fifth trial in block 7, i.e., the transition phase (Figure 5). We conducted a 2 (instance: positive and negative) 3 (panel: arrow, compass, and number) within subject ANOVA for the mean observation time. As a result, the interaction between the instance and panel factors reached significance \((F(2,24)=5.489, p=.011)\). The significant simple main effect of the panel factor was not observed in positive instances \((F(2,24)=1.723, p=.200)\). In contrast, in negative instances, the simple main effect of the panel factor was significant \((F(2,24)=7.123, p=.007)\). The observation time of the arrow panel was significantly longer than those of the compass panel marginally ( \(p=.067\) ) and the num-


Figure 5: Mean observation time of each panel for each positive and negative instance in the transition phase (bars show standard errors).
ber panel ( \(p=.016\) ). There was no significant difference between the observation times of the compass and number panels ( \(p=1.000\) ). The main effect of the instance factor \((F(1,12)=57.298, p<.001)\) and the panel factor reached significance \((F(2,24)=6.139, p=.007)\). The result that the participants considered rules from the initial perspective in negative instances rather than positive instances supports the peripheral change explanation.

\section*{Discussion}

In Experiment 2, we tested two hypotheses using eye movement data. In the initial phase, the results were consistent with our prediction: the observation time of the arrow panel was longer than those of the other panels. This means that the investigation from the initial perspective was conducted more often than that from other perspectives. This tendency was retained in the transition phase. This supports the retain hypothesis that focus on the initial perspective is retained if participants face negative instances. The analysis comparing the observation times in each positive and negative instance reveals that, although participants considered the rules to the same extent from all perspectives in positive instances, they considered rules from the initial perspective longer than those from other perspectives in negative instances. These results suggest that retained focus on the initial perspective is because of an attempt to improve the initial rule from the initial perspective when negative instances are given.

\section*{General Discussion}

In this study, we investigated the process of finding and rejecting an initial rule using eye movement data. We analyzed a perspective that indicated what types of rules participants considered. Our experiments yielded the following results: (1) In the initial phase in which only positive instances were given, participants focused on the initial perspective rather than other perspectives. (2) In the transition phase in which both positive and negative instances were given, participants retained a tendency to focus on the initial perspective. (3) The tendency to focus on the initial perspective was observed only when participants faced negative instances. On the basis of these results, we accept the retain hypothesis and the peripheral change explanation.

Participants' consideration of each rule from each perspective is out of the range of our study. We consider some possibilities about what participants were thinking from the initial perspective on the basis of the results of our experiments and the Chinn and Brewer's classifications (1993; Mason, 2001). The first possibility is that participants did not accept negative instances as valid, which corresponds to two of Chinn and Brewer's classifications: ignoring and rejecting. We can deny this possibility because we, as experimenters, guaranteed the validity of the instances. The second possibility is that participants did not try to explain negative instances, which corresponds with the categories of excluding and abeyance. If this possibility is true, the observation times of negative instances would be short; however, they were not. Furthermore, the
participants' task was to explain all instances. The third possibility is that they reinterpreted negative instances, which is Chinn and Brewer's category of reinterpreting. This possibility is rejected because interpretations other than those we provided regarding how to use the objects on each panel were not allowed.

The last two Chinn and Brewer's classifications are peripheral change and theory change, which, in this study, correspond to a peripheral change of the initial rule and a consideration of a whole new rule, respectively. For the following three reasons, it is most likely that participants tried to change the initial rule peripherally. First, in both experiments, many participants announced the initial rule in the transition phase. Second, the observation times of negative instances were longer than those of positive instances. Finally, there is normatively no benefit to focus on the initial perspective if participants consider a whole new rule. When participants kept the initial rule in their minds, they did not require much investigation from the initial perspective in positive instances. Consequently, a difference in the observation times from each perspective was not observed.

\section*{References}

Chinn, C. A., \& Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. Review of Educational Research, 63, 1-49.
Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. In R. J. Sternberg \& J. E. Davidson (Eds.), The nature of insight. Cambridge MA: MIT Press.
Haverty, L. A., Koedinger, K. R., Klahr, D., \& Alibali, M. W. (2000). Solving inductive reasoning problems in mathematics: not-so-trivial pursuit. Cognitive Science, 24, 249298.

Klahr, D., \& Dunbar, K. (1988). Dual space search during scientific reasoning. Cognitive Science, 12, 1-48.
Knoblich, G., Ohlsson, S., \& Raney, G. E. (2001). An eye movement study of insight problem solving. Memory \& Cognition, 29, 1000-1009.
Kruschke, J. K. (2003). Attention in learning. Current Directions in Psychological Science, 12, 171-175.
Lien, Y.-W., \& Lin, W.-L. (2011). From falsification to generating an alternative hypothesis: Exploring the role of the new-perspective hypothesis in successful 2-4-6 task performance. Thinking \& Reasoning, 17, 105-136.
Mason, L. (2001). Responses to anomalous data on controversial topics and theory change. Learning and Instruction, 11, 453-483.
Rehder, B., \& Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. Cognitive Psychology, 51, 1-41.
Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. Quarterly Journal of Experimental Psychology, 12, 129-140.

\title{
Automatic generation of naturalistic child-adult interaction data
}

\author{
Yevgen Matusevych (Y.Matusevych@uvt.nl) \\ Department of Culture Studies, Tilburg University PO Box 90153, 5000 LE Tilburg, the Netherlands
}

\author{
Afra Alishahi (A.Alishahi@uvt.nl) \\ Department of Communication and Information Sciences, Tilburg University PO Box 90153, 5000 LE Tilburg, the Netherlands
}

\author{
Paul Vogt (P.A.Vogt@uvt.nl) \\ Department of Communication and Information Sciences, Tilburg University PO Box 90153, 5000 LE Tilburg, the Netherlands
}

\begin{abstract}
The input to a cognitively plausible model of language acquisition must have the same information components and statistical properties as the child-directed speech. There are collections of child-directed utterances (e.g., CHILDES), but a realistic representation of their visual and semantic context is not available. We propose three quantitative measures for analyzing the statistical properties of a manually annotated sample of child-adult interaction videos, and compare these against the scene representations automatically generated from the same child-directed utterances, showing that these two datasets are significantly different. To address this problem, we propose an interaction-based framework for generating utterances and scenes based on the co-occurrence frequencies collected from the annotated videos, and show that the resulting interactionbased dataset is comparable to naturalistic data. We use an existing model of cross-situational word learning as a case study for comparing different datasets, and show that only interaction-based data preserve the learning task complexity.
\end{abstract}

Keywords: Child language acquisition; computational modeling; child-directed speech; cross-situational word learning.

\section*{Introduction}

A usage-based approach to language claims that natural languages in all their complexity can be learned merely from the input (or usage) data that is available to human learners. Computational modeling has been extensively used as a methodology for supporting this view: using a dataset that is statistically similar to child-directed input, a computational model can show that certain linguistic representations are learnable without domain-specific prior knowledge. Therefore, the input to a cognitively plausible model of language acquisition must have the same information components and statistical properties as the natural child-directed speech (CDS). A careful analysis and reconstruction of such data is a prerequisite of developing a model.

Recent decades have seen a significant growth in the variety and quantity of data collections for studying language. One major resource in this domain is CHILDES (MacWhinney, 2000), a collection of corpora containing recorded interactions of adults with children of different age and language groups. The interaction transcriptions have been used in several models of grammar induction from a large text corpus (e.g., Clark, 2001). The problem arises when a learning task demands perceptual and linguistic input. This
might be due to the nature of the process under study (e.g., learning the meaning of words) or the theoretical framework on which the model is based (e.g., construction grammar). In such cases, each utterance must be paired with a representation of its visual context. Many of the databases in CHILDES contain video recordings of the interaction sessions, but these recordings are mostly not annotated and hard to use without preprocessing or manual coding. Some models in fact use a small set of manually annotated videos as input (e.g., Yu \& Ballard, 2007; Frank, Tenenbaum, \& Fernald, 2013), but this approach is limited in quantity and scalability.

A common strategy for dealing with this challenge is to use artificially generated input: each sentence is constructed by randomly sampling from a presumed distribution over a list of words; the visual context is similarly built by sampling from a set of symbols which represent concepts or objects (e.g., Siskind, 1996; Niyogi, 2002). To make the data more naturalistic, some models select sentences from the transcriptions of actual child-adult interactions, and build the accompanying scene artificially by assuming a semantic representation for each word in the sentence and combining them (Fazly, Alishahi, \& Stevenson, 2010).

Generating the visual context automatically based on childdirected utterances (Utterance-Based Data, or UBD) eliminates the quantity concern (since manual annotation of the surrounding scene is not needed). However, the generated context is different from what the child observes in important ways. In a natural interaction scenario between a child and an adult, the surrounding scene is rather consistent or changes minimally (although the attention of the participants might move from one set of perceivable objects or actions to another). In contrast, in the automatic scene generation approach the utterance determines the scene, so the visual context can change drastically from one sentence to the next. A disproportional variation in visual context or scene can affect language learning; for example, context diversity has been shown to facilitate cross-situational word learning (Kachergis, Yu, \& Shiffrin, 2009). Moreover, a UBD approach guarantees that the relevant meanings for all the words in an utterance are included in the constructed scene. Artificial noise can be added by post-processing data and ran-
domly removing some meaning elements from the scene, but the noise ratio can still be unrealistically low.

UBD also differs from actual exchanges between children and their caregivers in that it lacks any interaction-based features. Crucially, utterances and actions directed at the learner at each point in time are independent of the learner's reaction to previous input data. In reality, the content of adult's utterance often depends on what the child just did or said (Kishimoto, Shizawa, Yasuda, Hinobayashi, \& Minami, 2007; Chouinard \& Clark, 2003). Interaction is suggested to be an essential mechanism of language development (e.g., MacWhinney, 2010).
In this paper, we investigate the characteristics of the visual context in a sample of child-adult interaction sessions, and compare them to those in an automatically generated one. We show that in every measure, the two contexts are considerably different, and argue that these differences might have implications for modeling child language learning. We propose a hybrid approach for generating an input corpus of utterance-scene pairs, where co-occurrence frequencies collected from a sample of manually annotated videos are used for generating utterances and visual contexts. Our framework not only takes the usage frequencies of the words and objects into account, but also includes interaction-based features such as dependence of adult's utterance on child's recent behavior. Finally, as a case study, we use an existing model of word learning (Fazly et al., 2010) to compare the complexity of the learning task using UBD vs. Interaction-Based Data (IBD, generated by our proposed framework). Our results show that using UBD for word learning unrealistically simplifies the learning task. Using IBD, in contrast, yields results that are closely comparable to the ones based on manually annotated scenes from videos of child-adult interaction.

\section*{Analyzing Utterance-based Input}

We analyze the cognitive plausibility of UBD by comparing its characteristics to a carefully annotated set of video recordings of child-adult interactions. The details of this data set are described below.

\section*{Data set}

As part of a larger project to study cross-cultural aspects of child-language acquisition (CASA MILA; Vogt and Mastin, 2013a, 2013b), three 13-month-old children from the Netherlands were recorded on video. The videos were recorded at the children's homes and involved interactions with one of the parents. The parents were instructed to continue their daily routines and ignore the recordings.

For each child, we selected an interaction session in a toy playing setting. The video fragments were 8 min 37 sec , 8 min 50 sec and 10 min long. We excluded some short episodes from the analysis, namely those (1) where the child or the adult was not captured properly by the camera, and (2) where the other parent was present, and the child's attention was focused on him/her. From the videos we extracted
adults' and children's gaze directions, actions, objects or participants that the actions were directed at, and utterances. Using this data, we constructed a corpus of child-directed utterances, each paired with a representation of the accompanying scene.

Scene representation There is no easy way to determine which elements a child perceives as potential referents at a certain moment of time. In fact, any object, action or event from the natural environment can be occasionally referred to in speech. However, studies suggest that children use certain mechanisms and constraints such as referential and salience cues to focus on relevant aspects of the scene (e.g., Behrend, 1990; Moore, Angelopoulos, \& Bennett, 1999). In particular, Yu, Smith, Shen, Pereira, and Smith (2009) show that objects in child's and parent's hands dominate the child's visual field.

In coding the interaction context in the video recordings, we consider two different interpretations for a scene:
active: all the objects that either participant (or both of them) is acting on or looking at during an utterance, in addition to the actions that (s)he performs (a similar approach was used by Frank, Goodman, and Tenenbaum (2008)).
all: the full set of visible objects, the action(s) performed during an utterance and the participants.

In addition, a third dataset was automatically generated:
UBD: Fazly et al. (2010) construct a scene by putting together the semantic symbols that correspond to the words in the accompanying utterance. Referential uncertainty is simulated by merging the representations of two consecutive scenes, and pairing them with only one of the utterances. They include noise into the data by removing the semantic symbol of one word from the scene for \(20 \%\) of the input items. Since we wanted to compare our results to those of Fazly et al. (2010), we applied the exact same approach to the child-directed utterances that we extracted from the CASA MILA recordings.

\section*{Measures}

To compare the datasets described above, we use three measures: scene stability, noise, and referential certainty.
Scene stability As mentioned before, the stability of the visual scene is one of the main points of deviation between natural interaction settings and the artificially generated input. We measure scene stability as the overlap between every pair of consecutive scenes. Since in both cases (the produced scenes in UBD and the annotated ones in our data set) a scene is represented as a set of symbols, we define the overlap between each two sets as the cardinality of their intersection divided by the cardinality of their union:
\[
\operatorname{overlap}\left(S_{i}, S_{i+1}\right)=\frac{\left|S_{i} \cap S_{i+1}\right|}{\left|S_{i} \cup S_{i+1}\right|}
\]

Noise We count a word's usage (or token) in an utterance as noisy if its semantic symbol is not included in the scene representation for that utterance. The total number of noisy words in an utterance, then, is calculated as
\[
\operatorname{noise}\left(U_{i}\right)=\frac{\left|U_{i}\right|-\left|U_{i} \cap S_{i}\right|}{\left|U_{i}\right|}
\]
where \(S_{i}\) is the current scene, and \(U_{i}\) is the (correct) meaning representation of the current utterance. To avoid making arbitrary decisions about the meaning of abstract or function words, we limit our analysis of noise to objects and actions.

Referential certainty We define the referential certainty for a scene as
\[
\operatorname{certainty}\left(S_{i}\right)=\frac{\left|U_{i} \cap S_{i}\right|}{\left|S_{i}\right|}
\]

Conceptually referential certainty shows what portion of a scene is referred to in the respective utterance. Note that this measure is the opposite of the more commonly used referential uncertainty, but it avoids the problem of having zero denominators in case the meaning representation of the utterance does not overlap with the scene.

\section*{Results}

We calculated the above measures for three datasets: childdirected utterances extracted from CASA MILA and paired with two interpretations of the accompanying visual scene (i.e. active and all), or with UBD-style automatically generated scene representations. Results are shown in Table 1.

Table 1: Plausibility measures for three datasets
\begin{tabular}{|l|c|c|c|}
\hline & all & active & UBD \\
\hline Scene stability & 0.916 & 0.436 & 0.112 \\
\hline Noise & 0.414 & 0.426 & 0.099 \\
\hline Referential certainty & 0.019 & 0.112 & 0.602 \\
\hline
\end{tabular}

The average values provided in the table inform us that the all condition differs substantially from the other two in terms of scene stability ( \(\mu=0.916\) vs. 0.436 and 0.112 ) and referential certainty ( \(\mu=0.019\) vs. 0.112 and 0.602 ). For this reason, and taking into account the fact that the standard deviation values for the all condition are rather small as compared to the respective means ( \(\sigma_{\text {stability }}=0.065 ; \sigma_{\text {certainty }}=0.032\) ), we eliminate this condition from the analysis. \({ }^{1}\) To compare the other two conditions, we ran the Mann-Whitney \(U\)-test for each of the three measures. We found significant differences between the annotated data (active condition) and the UBD in terms of all three measures: scene stability ( \(M d n=0.400 \mathrm{vs}\). 0.059 ; \(U=5230, n_{\text {active }}=274, n_{U B D}=133, p<.001, r=-.583\) ), noise ( \(M d n=0.400 \mathrm{vs}\). \(0.000 ; U=8927, n_{\text {active }}=278, n_{U B D}=139, p<.001, r=-.466\) ) and referential certainty \(\left(M d n=0.000\right.\) vs. \(0.571 ; U=3910, n_{\text {active }}=278, n_{U B D}=139, p<\) \(.001, r=-.690\) ). This demonstrates that UBD may be an easier input for the learner than the natural data.

\footnotetext{
\({ }^{1}\) The noise values for the active and the all conditions are almost equal, since the way we interpret a scene has little impact on the amount of noise in utterances. Due to this fact, for noise we also use only the active condition in the further analysis.
}

\section*{An interaction-based framework for input generation}

We propose an interaction-based framework for generating input data which resembles the verbal and non-verbal exchanges between a child and a caregiver. Our model is inspired by the language game model used to study the evolution of language (Steels, 1996; Vogt \& Haasdijk, 2010). In this model, agents communicate with each other through verbal and non-verbal behavior. Language game interactions involve a context, and agents communicate about items in this context, potentially learning associations between words and items.

We simulate the input generation process as a series of interactive sessions between two agents, Adult and Child. Each session starts with constructing a visual context (i.e., a collection of objects), followed by a sequence of exchanges between the two agents, until one of them leaves or terminates the session. In each turn, Adult performs an action (AdAct) while producing an utterance (AdUttr), to which Child responds by performing another action (ChAct) and producing an utterance (ChUttr, implemented as presence or absence of a verbal reaction). \({ }^{2}\) The main algorithm can be described as follows:
```

for }s\leftarrow1\mathrm{ to number of interaction sessions do
t\leftarrow0;
Context }\leftarrow\mathrm{ setupContext(s);
repeat
t\leftarrowt+1;
Situation}t\leftarrow < initialize(Context)
Situation}t\leftarrow < updateAdult(AdAct tr1 , AdUttr r-1 );
Situation}t\leftarrow \leftarrow updateChild(\mp@subsup{ChAct }{t-1}{},\mp@subsup{\mathrm{ ChUttr}}{t-1}{})
(AdAct },\mathrm{ AdUttr
Situation}\mp@subsup{}{t}{}\leftarrow\mathrm{ updateAdult(AdAct
(ChAct}t,\mp@subsup{ChUttr}{t}{})\leftarrow\operatorname{childTurn(Situation}\mp@subsup{|}{t}{})
until ChAct = 'leave' or AdAct }=\mathrm{ 'leave';
end

```

Each of the main steps in the algorithm are explained in more detail below.

Visual context From the sample data we extracted all the objects that were directly used by adults or children in their interactions. In each computational simulation, we randomly selected a fixed number of objects from the list and added them to the context. Since the size of the visual context may depend on the interaction domain (e.g., toy playing, book reading, etc.), we added it as a parameter to our framework.
Actions and action types We compiled two lists of actions, one for each agent. Actions might take arguments that can be an object type or the agents themselves (e.g., take toy or touch child). In order to base our computational model on more general behavioral patterns rather than on occasional events, we classified agents' actions into six types, based on

\footnotetext{
\({ }^{2}\) Since children in our sample video recordings were too young to talk, we did not gather enough statistical information about their produced utterances. However, the main concern of our framework is to create realistic child-directed input, and the child-produced data is an outcome of the learning model.
}
the factors that motivate them. These action types are listed in Table 2.

Table 2: Action types and their motivating factors
\begin{tabular}{|c|c|c|}
\hline Action type & Motivating factor & Example \\
\hline Continuation & Same person's previous action & \begin{tabular}{l}
Adult \(_{t}\) : [move bag] \\
Adult \(_{t+1}\) : [move box]
\end{tabular} \\
\hline Reaction & Other person's previous action & \begin{tabular}{l}
Child \(_{t}\) : [put ball] \\
Adult \(_{t+1}\) : [take ball]
\end{tabular} \\
\hline Result & Same person's prev. utterance & \begin{tabular}{l}
Adult \(_{t}\) : Bumba first \\
Adult \(_{t+1}\) : [take Bumba]
\end{tabular} \\
\hline Reaction to utterance & Other person's prev. utterance & Adult \(_{t}\) : The tree Child \(_{t+1}\) : [take toy tree] \\
\hline Initiating & None & Adult \({ }_{t}\) : sit down] \\
\hline
\end{tabular}

Utterances and utterance types We compiled a list of utterances produced by adults. Some of these contain placeholders which, depending on the context, can be filled with the labels for the respective actions and their arguments. Similar to actions, we recognized six utterance types based on their motivating factor, as listed in Table 3.

Table 3: Utterance types and their motivating factors
\begin{tabular}{|c|c|c|}
\hline Utterance type & Motivating factor & Example \\
\hline Accompanying & Same person's current action & Adult \(_{t}\) : [show ball] Adult \(t_{t}\) :This is a ball \\
\hline Continuation & Same person's prev. utterance & Adult \(t_{t}\) : Dad the ball? Adult \(t_{t+1}\) : Can dad the ball? \\
\hline Reaction & Other person's previous action & \begin{tabular}{l}
Child \(_{t}\) : [stand up] \\
Adult \(_{t+1}\) : Gonna walk?
\end{tabular} \\
\hline Answer & Other person's prev. utterance & Child \(_{t}\) : babbling Adult \(t_{t+1}\) : Yeah, Bumba \\
\hline Unknown & None & Child \(_{t}\) : babbling \\
\hline
\end{tabular}

Producing actions and utterances At each step \(t\) during a session, the actions and utterances produced by the agents are sampled from the frequency distributions collected from the annotated videos, each conditioned on the current situation. A situation includes all the relevant parameters, including the current and previous utterances and actions of both agents, the action arguments, and the visual context. Thanks to these parameters, the agents do not produce completely random actions and utterances, and the interaction process appears to be logical. (For more details on the estimated probabilities for each variable, see Matusevych (2012).)

Each turn in a session consists of the following steps:
1. The current situation \(\left(\right.\) Situation \(\left._{t}\right)\) is set to include the visual context (Context), the previous actions (AdAct \(t_{t-1}\) and ChAct \(_{t-1}\) ) and utterances \(\left(\right.\) AdUttr \(_{t-1}\) and ChUttr \(\left._{t-1}\right)\)
2. Adult's next action is generated:
(a) An action type for Adult ( AdActType \(_{t}\) ) is randomly selected, conditioned on Situation \({ }_{t}\)
(b) An action for Adult \(\left(A d A c t_{t}\right)\) is randomly selected, conditioned on AdActType \({ }_{t}\) and Situation \(_{t}\)
(c) Arguments for the action are randomly selected, conditioned on \(A d A c t_{t}\)
(d) Situation \(_{t}\) is updated to include \(A d A c t_{t}\) and its arguments
3. Adult's next utterance is generated:
(a) An utterance type for Adult (AdUttrType \({ }_{t}\) ) is randomly selected, conditioned on Situation \({ }_{t}\)
(b) An utterance for Adult \(\left(A d U t t r_{t}\right)\) is randomly selected, conditioned on AdUttrType \(t_{t}\) and Situation \({ }_{t}\)
(c) Situation \(n_{t}\) is updated to include \(A d U t t r_{t}\)
4. Child's next action and utterance ( \(C h A c t_{t}\) and \({\left.C h U t t r_{t}\right) ~ a r e ~}_{\text {ar }}\) generated in the same way as Adult's.

\section*{A sample interaction session}

We illustrate the interaction process using the following example (see Table 4).

Table 4: A fragment of a generated interaction
\begin{tabular}{|c|c|c|c|}
\hline \multicolumn{4}{|c|}{ Context: puzzle, piece-clown, bin, ball, piece-frog } \\
\hline Turn & Agent & Action & Utterance \\
\hline 1. & Adult & play puzzle & - \\
1. & Child & play piece-clown & babbling \\
2. & Adult & point puzzle & It fits here. \\
2. & Child & touch bin & babbling \\
3. & Adult & play puzzle & Yes? \\
\hline
\end{tabular}

The example can be interpreted as following. Adult starts the interaction by playing with a puzzle toy without saying anything. Child plays with a clown-shaped puzzle piece and babbles. Adult points at the puzzle saying It fits here. However, Child's attention is distracted by the bin, which he touches. He continues babbling. Adult continues playing with the puzzle toy, asking Yes?. His question can be interpreted either as a support for his previous utterance or as an attempt to clarify the child's utterance. The interaction goes on in this manner until one of the agents leaves.

\section*{Comparing IBD and UBD}

We used the interaction-based framework for generating a dataset. While in UBD scenes were constructed from utterances, in IBD each scene included salient elements, namely, the objects that the agents had in their hands, the agents' most recent actions and their arguments-in a manner similar to the active condition in the Analyzing Utterance-based Input section above. Using the same three measures-scene stability, noise, and referential certainty-we compare IBD to UBD and manually annotated CASA MILA data (both conditions).


Figure 1: Plausibility measures for four datasets

As can be seen in the charts (Figure 1), for each of the three measures the input data generated by our framework is much closer to the manually annotated data from the interaction videos than UBD. Again, for the reasons specified in the Analysis section, we did not use the all condition in the further analysis. For the other three conditions, the KruskalWallis \(H\)-test showed the significant difference in terms of stability \((H(2)=213.822, p<.001)\), noise \((H(2)=95.725, p<.001)\) and certainty \((H(2)=289.410, p<.001)\). To examine the pairwise differences between the three groups, we used Mann-Whitney tests, taking into account Bonferroni correction (which resulted in .025 level of significance). The difference between active and IBD was not significant in terms of noise ( \(M d n=0.400\) vs. \(0.333 ; U=37897.5, n_{\text {active }}=278, n_{\text {IBD }}=278, p>.025\) ), and significant with only small effect size in terms of scene stability \(\left(M d n=0.400\right.\) vs. \(\left.0.500 ; U=30012, n_{\text {active }}=274, n_{I B D}=274, p<.001, r=-.174\right)\) and certainty \(\left(M d n=0.000\right.\) vs. \(0.000 ; U=35002.5, n_{\text {active }}=278, n_{I B D}=278, p<\) \(.025, r=-.100\) ). However, the difference between UBD and IBD was significant with a large effect size for each measure: scene stability ( \(M d n=0.059\) vs. \(0.500 ; U=2586, n_{U B D}=133, n_{I B D}=\) \(274, p<.001, r=-.699\) ), noise ( \(M d n=0.000\) vs. \(0.333 ; U=10008.5, n_{U B D}=\) \(139, n_{I B D}=278, p<.001, r=-.426\) ) and certainty ( \(M d n=0.571\) vs. 0.000 ; \(\left.U=2451.5, n_{U B D}=139, n_{\text {IBD }}=278, p<.001, r=-.760\right)\). These results confirm that data generated by the proposed framework is more suitable for training and evaluating cognitive models than UBD. We further investigate this claim by using these different data sets in an existing model of word learning.

\section*{Case study: learning word meaning}

We used the cross-situational word learning model of Fazly et al. (2010) as a case study for our proposed input generation framework. Our goal is to show that the complexity of the learning task depends on the properties of the input data, and less realistically generated input can considerably simplify the task.

\section*{Description of the model}

The model of Fazly et al. (2010) incrementally learns the meaning of each word (e.g., play) as a probability distribution over all the possible meaning components, each represented as a unique symbol (e.g., Play, Ball). At each moment in time, the model receives a new input item, consisting of an utterance and its (ambiguous) semantic representation, which is an unordered set of symbols. The model uses its previous knowledge of word meanings to align each word in the current utterance with the most likely symbols in the current scene representation. It then uses these alignments to update the meaning of each word by accumulating such cross-situational evidence over time.

\section*{Model performance on different types of input}

We compared the performance of the word learning model on four different data sets:
1. the manually annotated portion of CASA MILA (active);
2. UBD generated from the same data set (UBD-CASA MILA);
3. original UBD used by Fazly et al. (2010) and generated from the Manchester corpus in the CHILDES database (UBD-Manchester);
4. IBD generated by our framework as a result of simulations with 19 objects in the environment (which was the average context size in the analyzed CASA MILA dataset). \({ }^{3}\)
For measuring the learning success at each moment, we used effective ratio calculated as the number of words that the learner has acquired at that time, divided by the number of words that she heard so far. The growth of the effective ratio over time is presented in Figure 2. Note that the size of \(U B D\) (CASA MILA) set is two times smaller than that of the original CASA MILA dataset, because only every other natural utterance could be included into UBD. For UBD (Manchester) and \(I B D\) the graphs show values averaged over 10 word learning simulations.


Figure 2: Overall model performance on four different datasets for 300 (left) and 6500 (right) utterances

The graph on the left shows the effective ratio over the course of 300 input items, which is slightly more than the size of the CASA MILA dataset. It is clear from these results that the performance of the word learning model is very similar when it is trained on data collected from CASA MILA and on data generated by our framework (IBD). In contrast, the model performs much better when it is trained on any of the UBD sets. This difference again suggests that UBD is not representative of what young language learners have access to, and a more realistic approach to data generation must be applied. The graph on the right shows the same measure over the course of 6500 utterances (the size of UBD-Manchester). The same pattern can be seen: there is a considerably large gap between the learning curves in UBD and IBD cases. It is also clear that in the latter case, the size of the input to the learner does not have to be constrained by the amount of data available in an existing collection.

\section*{Conclusion and discussion}

We manually annotated a small dataset of video recordings of child-adult interactions and collected various types of co-

\footnotetext{
\({ }^{3}\) Since one of the main parameters of the framework was the context size, we also investigated whether the learning process would vary with the number of objects in the environment, but our manipulations did not result in changing the overall learning pattern in terms of effective ratio.
}
occurrence frequencies of utterances, utterance types, accompanying actions and action types, action arguments and participants, and other objects available in the visual context. Using three quantitative measures, we compared the characteristics of these utterances and their surrounding scenes with the product of the most realistic existing approach to automatic generation of scene representations (Fazly et al., 2010). Our analyses show significant differences between the two datasets, and using an existing model of word learning as a case study further demonstrates that automatically generated utterance-based data simplifies the learning task to an unrealistic scale. However, manual annotation as an alternative approach (e.g., Yu \& Ballard, 2007; Frank et al., 2013) is not scalable due to the limited quantity of the data available. The hybrid approach that we propose eliminates these problems: we present an input generation framework which can produce an infinite stream of child-adult interaction data containing both linguistic and visual information, whose statistical properties are closely comparable to those of manually annotated data.

Any data annotation or generation scheme inevitably incorporates assumptions about important components and information cues in language learning, which can be seen as built-in biases brought to the learning task. However, computational models need data and will benefit from any attempt to make this data more naturalistic.

An extension of the proposed framework can potentially provide certain interaction features such as the participants' focus of visual attention and head movement. Such extra features can allow computational models to systematically investigate the impact of interaction factors in language learning.

The dataset that we analyzed was limited in size and the interaction domain (toy playing). We add a parameter to our framework to account for potential variation in the size of the visual context. But humans' linguistic behavior (e.g., the structural and pragmatic characteristics of utterances) may also depend on the domain to some extent (e.g., Choi, 2000). Therefore, a larger and more diverse collection of interaction videos will provide a more realistic base for estimating the input generation probabilities in our framework. The larger CASA MILA corpus of interaction data that is currently under development (Vogt \& Mastin, 2013a) is one suitable candidate for such expansion.

\section*{References}

Behrend, D. A. (1990). Constraints and development: A reply to Nelson (1988). Cog. Development, 5(3), 313-330.
Choi, S. (2000). Caregiver input in English and Korean: Use of nouns and verbs in book-reading and toy-play contexts. J. of Child Lang., 27(1), 69-96.

Chouinard, M. M., \& Clark, E. V. (2003). Adult reformulations of child errors as negative evidence. J. of Child Lang., 30(3), 637-669.
Clark, A. (2001). Unsupervised induction of stochastic
context-free grammars with distributional clustering. In Proc. of CoNLL'2000 (pp. 105-112).
Fazly, A., Alishahi, A., \& Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. Cognitive Science, 34(6), 1017-1063.
Frank, M., Goodman, N., \& Tenenbaum, J. (2008). A bayesian framework for cross-situational word-learning. In Proc. of NIPS'2008 (pp. 457-464).
Frank, M., Tenenbaum, J. B., \& Fernald, A. (2013). Social and discourse contributions to the determination of reference in cross-situational word learning. Language, Learning and Development.
Kachergis, G., Yu, C., \& Shiffrin, R. M. (2009). Frequency and contextual diversity effects in cross-situational word learning. In Proc. of CogSci'2009 (pp. 2220-2225).
Kishimoto, T., Shizawa, Y., Yasuda, J., Hinobayashi, T., \& Minami, T. (2007). Do pointing gestures by infants provoke comments from adults? Infant Behavior and Development, 30(4), 562-567.
MacWhinney, B. (2000). The CHILDES project: Tools for analyzing talk (3rd ed.). Lawrence Erlbaum Associates.
MacWhinney, B. (2010). Computational models of child language learning: An introduction. J. of Child Language, 37(3), 477-485.
Matusevych, Y. (2012). Modeling child-adult interaction: A computational study of word learning in context Master's thesis, Tilburg University, the Netherlands.
Moore, C., Angelopoulos, M., \& Bennett, P. (1999). Word learning in the context of referential and salience cues. Developmental Psychology, 35(1), 60-68.
Niyogi, S. (2002). Bayesian learning at the syntax-semantics interface. In Proc. of CogSci’2002 (pp. 697-702).
Siskind, J. M. (1996). A computational study of crosssituational techniques for learning word-to-meaning mappings. Cognition, 61(1-2), 39-91.
Steels, L. (1996). Emergent adaptive lexicons. In From Animals to Animats 4: Proc. of the Fourth International Conference on Simulation of Adaptive Behavior (pp. 562-567).
Vogt, P., \& Haasdijk, E. (2010). Modeling social learning of language and skills. Artificial Life, 16(4), 289-309.
Vogt, P., \& Mastin, J. D. (2013a). Anchoring social symbol grounding in children's interactions. Künstliche Intelligenz, 27, 145-151.
Vogt, P., \& Mastin, J. D. (2013b). Rural and urban differences in language socialization and vocabulary development in Mozambique. Proc. of CogSci'2013.
Yu, C., \& Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. Neurocomputing, 70(13-15), 2149-2165.
Yu, C., Smith, L. B., Shen, H., Pereira, A. F., \& Smith, T. (2009). Active information selection: visual attention through the hands. IEEE Transactions on Autonomous Mental Development, 1(2), 141-151.

\title{
Coordinating turning while walking and talking
}

\author{
Eric Mayor (eric.mayor@unine.ch) \\ Institute of Work and Organizational Psychology, University of Neuchâtel \\ Emile-Argand 11, 2000 Neuchâtel, Switzerland
}

\author{
Adrian Bangerter (adrian.bangerter@unine.ch) \\ Institute of Work and Organizational Psychology, University of Neuchâtel \\ Emile-Argand 11, 2000 Neuchâtel, Switzerland
}

\begin{abstract}
Few studies have investigated multitasking in joint actions, especially two joint actions performed by two people together and coordinated via multimodal communication. We investigate the case of two people walking and talking together, a common combination of joint actions. In an experiment, pairs talked together in four varying conditions of mobility. A narrator told a story to a partner. They did this while either standing immobile, walking along a straight-line itinerary, or walking along a complex itinerary featuring several turns. They also completed a walking task along a complex itinerary without having to tell a story. One person (the navigator) was also entrusted with a map of the itinerary. We analyzed how participants coordinated turning while telling a story. Narrators relied more on verbal means to signal turning, and were more distracted during the turn, leading to more repetition of story-related content.
\end{abstract}

Keywords: conversation, coordination, walking, multimodal communication, joint action, collaboration, multitasking.

\section*{Multimodal Coordination of Concurrent Joint Actions}

Multitasking, or the concurrent performance of two different tasks, is common in everyday life. An important question concerns the effect of multitasking on task performance. Research on multitasking has revealed much about the basic cognitive processes involved, showing that sharing processing resources (attention, working memory, and executive control) between multiple tasks can impair performance (Salvucci \& Taatgen, 2011). Much of this research, however, has focused on multitasking behavior of individuals engaged in solitary tasks. Some research focuses on situations where people coordinate concurrent joint actions (e.g., either an individual action and a joint action or two joint actions). For example, Fussell, Kiesler, Setlock, Scupelli, and Weisband (2004) investigated how people coordinated two projects, each one with a different partner, face-to-face and via instant messaging. But few studies investigate the role of dialogue in coordinating concurrent joint actions. This is a significant oversight, because dialogue (which involves both verbal and nonverbal acts, i.e., multimodal communication; Stivers \& Sidnell, 2005) is the commonest means of coordinating joint action (Clark, 1996).

Investigations into the role of multimodal dialogue in coordinating multiple joint actions can significantly expand cognitive science research on multitasking. Recognizing
how processing resources are distributed among multiple individuals and coordinated over multiple communicative modalities challenges existing cognitive theories on multitasking. As we will see, investigating such phenomena requires theories about coordinating meaning and identities in interaction.

What coordination problems arise when people perform concurrent joint actions, and how do people use multimodal communication to solve these problems? An initial investigation of this issue was proposed by Chevalley and Bangerter (2010). They used Clark’s (1996) theory of language use to propose a model of how people suspend a joint action they are doing together in response to an interruption, and how they reinstate those actions after the interruption. Participants have to coordinate on at least three aspects in switching from one joint action to another. First, when reinstating a joint action after a suspension, participants have to update their common ground (Clark, 1996) about the state of the action. They do this by talking together about where they were in the action. Second, they have to attend to their partners' face needs. According to politeness theory (Brown \& Levinson, 1987), people have a right to positive consideration by others (positive face) as well as a right to act freely without being unduly imposed on (negative face). Suspending a joint action and making one's partner wait while one does something else constitutes a threat to negative face. To mitigate this threat, participants engage in politeness like warning about an interruption, asking permission to suspend, minimizing its duration (just \(a \mathrm{sec}\) ) or apologizing. Third, coordinating responses to an interruption raises the question of a division of labor among interaction partners. For example, only one participant in a joint action may be the target of an interruption, leaving the other participant free to keep the current state of the action in memory. The non-interrupted participant may then play a crucial role in reconstructing the state of the action once the interruption is over. Indeed, asymmetries in conversational roles or access to privileged knowledge affect the way partners coordinate suspending and reinstating joint actions (Bangerter, Chevalley, \& Derouwaux, 2010).

Here we pursue this line of inquiry but focus on the case where two people accomplish two joint actions concurrently with each other (rather than suspending one joint action for a longer period of time in order to engage in another one possibly involving another person). In such a case, conflicts between resources used for one task but required for another
may arise. Multimodal communication is a potential way of circumventing this "bottleneck", because communicating about one joint action in a different modality (e.g., via gestures) might leave the primary modality (e.g., talk) undisturbed. Another way of circumventing the bottleneck is to distribute task components among different individuals (Hutchins, 1995). In doing so, participants in multiple concurrent joint actions minimize overall collaborative effort (Clark, 1996). More generally, in coordinating multiple concurrent joint actions, participants respond to two fundamental imperatives of conversation (Enfield, 2006). An informational imperative requires participants to coordinate joint understanding of both actions (e.g., where they are in a narrative, when they are going to turn a corner), and an affiliational imperative requires them to manage each other's identities and commitments to the joint action (e.g., not interrupting a speaker at an interesting point in a story). We apply theories of conversation as joint action to explain processes occurring in multimodal coordination of concurrent joint actions.

\section*{Walking While Talking}

We report initial findings from an experimental study investigating how two people coordinate two concurrent joint actions, namely talking together while walking together. We chose walking and talking because it is a commonly occurring combination in everyday life. Many everyday conversations take place in situations of mobility. For example, hospital personnel spend substantial amounts of time engaging in various activities while walking (Bardram \& Bossen, 2005).
Talking together is a common joint action that is coordinated through a variety of channels, including speech, paraverbal information, gaze, gesture, body posture and so on. Depending on the type of conversation, participants may occupy different roles that constrain their relevant contributions. Of course, talking together has been largely studied in various disciplines (Sacks, Schegloff, \& Jefferson, 1974, Clark, 1996), but comparatively little is known about how conversation is coordinated with other, non-linguistic joint actions.

Walking together is also a common joint action where partners must coordinate walking speed and posture in order to position themselves abreast of each other. Synchronizing gait requires coordination via tactile (hand-holding) or visual signals (Zivotofsky \& Hausdorff, 2007). In some cases, when walking constitutes a means of locomotion to a particular place known to only one of the partners, roles may also emerge (i.e., one person using a map). Indeed, even transitory forms of collective mobility like crossing a street as a group when the traffic light for pedestrians is green require coordination (Relieu, 2008).

Walking normally requires few cognitive resources, and people are typically able to walk and do something else at the same time. But there are measurable decrements in task performance in such cases. For example, older adults are less able to memorize while walking (Lindenberger,

Marsiske \& Baltes, 2000). Also, adults who answer questions while walking are less fluent than while stationary (Kemper, Herman, \& Lian, 2003) Another study (Yatani \& Truong, 2009) found that users of handheld devices are more effective when standing than when walking. These studies fall short of studying true joint actions because they do not investigate interactive conversation. However, they are relevant for understanding walking performed in conjunction with other actions, and suggest that the small decrements in performance could be easily increased by making walking more difficult (e.g., by having participants navigate a complex itinerary using a map rather than just walking a predetermined path). Thus, walking constitutes a convenient and malleable candidate task to investigate in conjunction with talking.

\section*{Our Experiment}

In our experiment, pairs of participants were videotaped while talking together in four within-subjects conditions of varying mobility (the Task variable) designed to instantiate different combinations of concurrent demands related to walking and talking (Table 1). The talking task involved one person (the narrator) telling a story to the other (the partner). Participants kept these roles for the duration of the study. In the talk-only condition, pairs were standing immobile while the narrator told the story. In the talk-andwalk condition, they walked together along a straight-line itinerary which was indicated on a map while the narrator told the story. In the talk-and-navigate condition, they walked together along a complex itinerary (i.e., featuring five turns) which was indicated on a map while the narrator told the story. In the navigate-only condition, they walked together along a similarly complex itinerary (i.e., also featuring five turns) which was indicated on a map but could talk about whatever they wanted, thus creating a situation where navigation is clearly prioritized.

Table 1. Demands of talking and walking instantiated in four within-subjects conditions.
\begin{tabular}{|c|c|c|}
\hline & Talking Demands & Walking Demands \\
\hline Talk Only & High & None \\
\hline Talk and Walk & High & Low \\
\hline \begin{tabular}{c} 
Talk and \\
Navigate
\end{tabular} & High & High \\
\hline Navigate Only & Low & High \\
\hline
\end{tabular}

In addition, either the narrator or the partner was entrusted with the responsibility of making sure the pair followed the itinerary correctly. The person responsible (hereafter the navigator) was given the map. This constituted a betweensubjects variable.

Thus, the design of the experiment was a 4 (Task, withinsubjects) X 2 (Navigator, between-subjects) design. In such a setting, it is possible to investigate many interesting questions. For example, the coordination of story-telling involves the narrator regularly seeking a back-channel
response from the partner. This is often done via gaze (Bavelas, Coates, \& Johnson, 2002). If the partner is distracted and thus kept from producing back-channel responses, the quality of the story suffers (Bavelas, Coates, \& Johnson, 2000). However, when walking and talking, gaze may not be as freely available for this purpose as when people are talking without moving. The effect of walking on gaze allocation and therefore on story-telling coordination via back-channels can be investigated by comparing the talk-only condition with the other conditions. Other comparisons are possible, for example comparing the talk-and-navigate condition with the navigate-only condition allows investigating to what extent talking may interfere with a navigational task, with navigational performance being measured by changes in walking speed (e.g., slowing down or stopping) or by errors (e.g., wrong turns).
In this paper, our analysis focuses on how participants coordinate turning to the left or to the right according to the itinerary while talking. Turning while talking is a good example of how an acute coordination demand may emerge from one joint action, thereby jeopardizing coordination of the other joint action. In our experiment, the responsibility for navigating was often implicitly entrusted to the navigator, who was the only participant who had easy visual access to the map. Thus, turning was typically coordinated via some kind of signal from navigators to the other participant. There are several ways to do this. Navigators might tell other participants to turn, for example by uttering we're going to turn to the right. Or they might point in the direction of the turn. They might also swivel their gaze in the direction of the turn, or nudge or push their partner, or use a combination of several signals. Some pairs even managed to turn without any visible or audible coordination signals (albeit quite rarely). How might participants decide to coordinate a turn? When narrators are navigators, they have the floor, because they are responsible for telling the story. Thus, it seems easier for them to signal the turn via verbal means. On the other hand, when partners are navigators, they must interrupt the narrator and gain the floor if they want to signal the turn verbally. This is a potential threat to the narrator's face (Bangerter, Chevalley, \& Derouwaux, 2010). If, as predicted by joint action theories of conversation, participants deal with this problem by distributing collaborative effort across modalities and by a distribution of labor, we would expect partners as navigators to rely relatively less on verbal means to signal turning than narrators as navigators.
To test this possibility, we investigated the effect of the Task and Navigator variables on the coordination of turning. For each of the five turns in the talk-and-navigate condition for each pair, we coded what kind of verbal or nonverbal means they used to coordinate the turn. We compared this data with the verbal and nonverbal means used to coordinate turning in the navigate-only condition. Because there are no narrator and partner roles in the navigate-only condition, it serves as a baseline for comparison with the effect of roles in the talk-and-navigate condition.

We also investigated the effect of the Task and Navigator variables on the coordination of storytelling. When narrators are navigators, they may be more distracted when they have to both communicate about the turn and keep track of the story they are telling. This might make participants more likely to lose track of the story, and thus more likely that some utterance relative to the story will have to be repeated after the turn as a means of reconstructing the story line (Chevalley \& Bangerter, 2010).

\section*{Method}

\section*{Participants}

Eighty people ( 46 women and 34 men) participated in 40 pairs. Pairs were composed irrespective of gender. Participants were native French speakers and did not know each other before the study.

\section*{Procedure}

We video-recorded each pair in one static and three mobile conditions. In all conditions, participants were also equipped with audio recorders and tie-clip microphones. In the talkonly condition, participants were filmed with a hand-held video camera from a distance of several meters. In the three mobile conditions, participants walked abreast. They were filmed frontally with a device consisting of either a GoPro Hero2 camera or a Contour HD camera attached to a perch that was held by the experimenter who walked about 1.5 m behind the pair. The perch extended over the heads of the pair (see Figure1). It was just above and the out of their field of vision when they looked ahead. The experimenter calibrated his walking speed to the participants' in order to maintain the camera at a constant distance from them. The perch also featured a supplementary backup audio recorder attached above the participants' heads. In this way, the setup allowed frontal mobile videotaping of the participants from above their heads to below their knees (Figure 2).


Figure 1: Setup of portable videocamera perch.
Twenty ordered combinations of the four conditions were randomly computed and randomly assigned to pairs in each
between-subjects condition (the same combinations were used in both conditions). Pairs performed the tasks in the order thereby defined.
In the walking conditions, participants followed an itinerary using a map, responsibility for the navigation being randomly assigned to the narrator or partner before the experiment. Participants were asked to navigate from a starting point to a precisely marked end point. Thus, even straight-line itineraries required some monitoring on the part of the navigator to avoid undershooting or overshooting the end point. All itineraries had a total length of approximately 400 meters. Recordings took place outdoors in a quiet urban area.


Figure 2: Still pictures of two pairs (in both cases, the narrator is the navigator and is on the left). Bottom picture:

The narrator is initiating a turn by gesturing.

\section*{Data preparation}

Video was synchronized with the sound of the two audio recorders (on a separate track) and a file was produced per condition for each group. A video clip of each turn was prepared. Clips started approximately 15 seconds before the initiation of the turn and lasted 30 seconds.
Based on a viewing of each clip, a detailed qualitative description of how each pair coordinated each turn was written by the first author. The description featured a sequential list of the circumstances of the turn, as well as any visible or audible behavior dedicated to coordinating the turn, including specifications of which participant was on the inside of the turn, descriptions of gestures (e.g.,
pointing), verbatim transcription of any utterances or the direction of gaze.

A research assistant then coded each description on the following variables:
- Who produced a signal (narrator or partner).
- When it was produced (before, during or after the turn)
- The signal produced (look at map, look at other participant, look in the direction of the turn, look elsewhere, point in the direction of the turn, point on the map, other gesture, give directions verbally, request help, agree)
- Repetitions of previous story content

Interrater agreement was assessed by having two coders independently code 25 turns. Cohen's kappa indicated excellent agreement (all kappas > .90).
The individual turn-coordination signals were grouped together to compute frequencies with which three types of signals were produced: gaze, gesture and utterance. The number of repetitions per turn was also computed.

\section*{Results}

Pairs took the same amount of time to complete the task in all four conditions, Wilks' lambda \(=.930, F(3,37)=.922\), \(p=.44,(M=297.5 \mathrm{~s}, S D=65.7 \mathrm{~s})\).

Because Task is a within-subjects variable, we performed repeated-measures analyses with the frequencies (by turn) of gaze, gesture, utterance and repetition as dependent variables. Because turns are nested within groups and the dependent variables are count data, we ran mixed model Poisson regressions in R 3.0. These analyses take into account the random effects of pairs. The independent variables were Navigator role and Task, which were entered in that order in the models, prior to the interaction term. Independent variables were dummy coded (0 vs 1 ). Categories coded 0 were Partner for the Navigator variable and Talk-and-navigate for the Task variable. The models were fitted by the Laplace approximation. Table 2 shows the means for each dependent variable as a function of Task and Navigator role. In what follows, \(b\) coefficients for each effect represent natural-log-transformed values.

Table 2: Mean frequencies (standard deviations) of gaze, gestures, utterances, and repetitions by Task and Navigator role per turn.
\begin{tabular}{|l|c|r|r|c|}
\hline & \multicolumn{2}{|c|}{ Talk-and-navigate } & \multicolumn{2}{c|}{ Navigate-only } \\
\hline & Narrator & Partner & Narrator & Partner \\
\hline Gaze & \(4.50(2.52)\) & \(4.36(2.74)\) & \(3.92(2.57)\) & \(4.13(2.29)\) \\
\hline Gesture & \(1.34(2.36)\) & \(0.87(1.03)\) & \(0.81(1.30)\) & \(0.69(1.12)\) \\
\hline Utterance & \(1.63(2.92)\) & \(0.58(1.24)\) & \(1.56(2.34)\) & \(1.36(2.26)\) \\
\hline Repetition & \(0.15(0.38)\) & \(0.03(0.17)\) & \(0.10(0.30)\) & \(0.12(0.41)\) \\
\hline
\end{tabular}

Gaze is used frequently in coordinating turning. While gaze shifts might be primarily produced by participants to steer their own individual walking trajectory, they might also attract the attention of the other participant and thus serve as an unintended cue that a turn is imminent. Pairs
gazed marginally less in the navigate-only condition than in the talk-and-navigate condition ( \(b=-.12, S E=0.07, p=\) 0.07 ). Navigator role was not a significant predictor of gaze. It is worth noting that this model does not fit the data significantly better than a null model (deviance \(=4, d f=3\), \(n s\) ). (Differences from the null models are significant for all other dependent variables.)
Gestures were used regularly, albeit less often than gaze. Pairs gestured marginally less when the partner was responsible for the itinerary than when the narrator was ( \(b=\) \(-0.36, S E=0.20, p=0.07\) ). In the navigate-only condition, pairs gestured less than in the talk-and-navigate condition ( \(b\) \(=-0.48, S E=0.14, p=0.0007\) ). As expected, pairs in the navigate-only condition used less utterances to coordinate turning than did pairs in the talk-and-navigate condition ( \(b=\) \(-1.06, S E=0.29, p=0.0003\) ). The interaction of task and navigator was also significant ( \(b=.86, S E=0.19, p<\) 0.0001): In the talk-and-navigate condition, pairs discussed the navigation task more when the narrator was responsible for navigation than when the partner was. On the contrary, in the navigate-only condition, pairs discussed the navigation task equally often, irrespective of navigator role.


Figure 4: Example of a progressive breakdown in the story following a missed turn.

Utterances related to turning included directions but also expressions of uncertainty, like I just need to look or I think
that's it, as well as occasional requests for assistance, which sometimes could completely override the narrative activity. In one exceptional case (depicted in Figure 4), the narrator progressively realizes she is lost, first interrupting her story by saying I don't understand where to go anymore while pointing vaguely in the direction of the turn. She then looks at her partner and laughs, and then asks her can you help \(m e\), while showing the map to her partner. All the while, the pair is walking straight ahead without slowing down. Subsequent to the frames shown in Figure 4, the pair will slow down and come to a complete stop while the partner explains to the narrator where to go. Only once they have corrected their trajectory will the narrator resume her story. This example illustrates the complex interplay of the multimodal signals produced (verbal utterances, gaze, pointing, and showing the map). It also illustrates a momentary but complete breakdown in one task (talking) when coordination requirements of the other task (walking) briefly overwhelm participants' available resources.

Repetitions of story-related content were infrequent. When they did occur, it was mostly the last utterance before the turn that was repeated immediately after the turn was complete. Nonetheless, repetitions of story-related utterances were less frequent when the partner was responsible for navigation than when the narrator was ( \(b=-\) \(1.60, S E=0.71, p=0.02\) ). There was also an interaction ( \(b\) \(=1.76, S E=0.81, p=0.03\) ). In the talk-and- navigate condition, pairs repeated story content more when the narrator was responsible for navigation than when the partner was. This was not the case in the navigate-only condition, possibly because no participant had an assigned role regarding the discussion (usually participants engaged in small talk while navigating in this condition, each contributing to the discussion).

\section*{Discussion}

Talking together while walking together constitutes a complex set of concurrent joint activities. Using the example of turning, we have shown how the division of labor among pairs affects the coordination of the turn. Narrators used more verbal utterances to signal a turn than partners. This finding converges with those of Chevalley and Bangerter (2010) and Bangerter, Chevalley, and Derouwaux (2010), who found that it was more effortful for listeners to suspend a conversation than for speakers. In refraining from interrupting speakers, listeners also deployed more politeness, suggesting they were trying to mitigate the face threat of interrupting the speaker. In the present case, partners may have preferred to accomplish some signals via gesture, in order to avoid interrupting the narrator's story.

We also found that narrators repeated story-related utterances after a turn more often when they were navigators than when they were not, suggested that they were distracted by the double responsibility of narrating and signaling the turn. It may also be the case that this finding is related to the previous finding that narrators use more verbal
means. Given that they have the floor, narrators may find it comparatively easier to interrupt their story to signal the turn. But in doing so, they may potentially interfere more with their own recall of where they were in the story than if they would use gestural means to signal the turn.

Our findings confirm that, in coordinating concurrent joint actions, participants need to manage common ground, pay attention to face wants of their partners, and that they may accomplish these constraints via a division of labor and using multimodal communication. Thus, coordinating concurrent joint actions expands the phenomenon of multitasking into the realm of conversational interaction and requires consideration of social as well as cognitive processes.

\section*{Acknowledgments}

Research was supported by the Swiss National Science Foundation (Grant no. 100014_138195/1 to Adrian Bangerter). We thank Besarta Domaqi, Rosalie Muriset, Phillip Strahm and Lucie Voillat for assistance in conducting the study.

\section*{References}

Bangerter, A., Chevalley, E., \& Derouwaux, S. (2010). Managing third-party interruptions in conversations: Effects of duration and conversational role. Journal of Language and Social Psychology, 29, 235-244.
Bardram, J., \& Bossen, C. (2005). Mobility work: The spatial dimension of collaboration at a hospital. Computer Supported Cooperative Work, 14, 131-160.
Bavelas, J. B., Coates, L., \& Johnson, T. (2000). Listeners as co-narrators. Journal of Personality and Social Psychology, 79, 941-952.
Bavelas, J. B., Coates, L., \& Johnson, T. (2002). Listener responses as a collaborative process: The role of gaze. Journal of Communication, 52, 566-580.
Brown, P., \& Levinson, S. (1987). Politeness: Some universals in language use. Cambridge: Cambridge University Press.
Chevalley, E., \& Bangerter, A. (2010). Suspending and reinstating joint activities with dialogue. Discourse Processes, 47, 263-291.
Clark, H. H. (1996). Using language. Cambridge: Cambridge University Press.
Enfield, N. J. (2006). Social consequences of common ground. In N. J. Enfield \& S. C. Levinson (eds.), Roots of human sociality: Culture, cognition, and human interaction (pp. 399-430). Oxford: Berg.
Fussell, S. R., Kiesler, S., Setlock, L. D., Scupelli, P., \& Weisband, S. (2004). Effects of Instant Messaging on the management of multiple projects. CHI 2004 (pp. 191198). NY: ACM Press.

Hutchins, E. (1995). Cognition in the wild. Cambridge, MA: MIT Press.
Kemper, S., Herman, R, \& Lian, C. (2003). The cost of doing two things at once for young and older adults:
talking while walking, finger tapping, and ignoring speech or noise. Psychology and Aging, 18, 181-192.
Lindenberger, U., Marsiske, M., \& Baltes, P. B. (2000). Memorizing while walking: Increase in dual-task costs from young adulthood to old age. Psychology and Aging, 15, 417-436.
Relieu, M. (2008). Mobilité et coordination sociale. Note sur l'analyse ethnométhodologique des déplacements en situation naturelle. Actes des 8èmes Journées Francophones Extraction et Gestion des Connaissances, Sophia Antipolis.
Sacks, H., Schegloff, E. A. \& Jefferson, G. (1974). A simplest systematics for the organisation of turn-taking for conversation. Language, 50, 696-735.
Salvucci, D., \& Taatgen, N. (2011). The multitasking mind. Oxford: Oxford University Press.
Stivers, T., \& Sidnell, J. (2005). Introduction: Multimodal interaction. Semiotica, 156, 1-20.
Yatani, N., \& Truong, K. (2009). An evaluation of stylusbased text entry methods on handheld devices studied in different user mobility states. Pervasive and Mobile Computing, 5, 469-508.
Zivotofsky, A. Z., \& Hausdorff, J. M. (2007). The sensory feedback mechanisms enabling couples to walk synchronously: an initial investigation. Journal of NeuroEngineering and Rehabilitation, 4, 28.

\title{
Insight Follows Incubation In The Compound Remote Associates Task
}

\author{
Sean W. Mc Carthy (smccartesol@gmail.com) \\ Department of Psychology \\ Loyola University Chicago \\ 1032 W Sheridan Rd \\ Chicago, IL 60626 USA \\ John M. Malony (jmalony@luc.edu) \\ Department of Psychology \\ Loyola University Chicago \\ 1032 W Sheridan Rd \\ Chicago, IL 60626 USA \\ Robert G. Morrison (rmorrison@luc.edu) \\ Department of Psychology, Neuroscience Institute \\ Loyola University Chicago \\ 1032 W Sheridan Rd \\ Chicago, IL 60626 USA
}

\begin{abstract}
The phenomenon of insight is frequently characterized by the experience of a sudden and certain solution. Anecdotal accounts suggest insight frequently occurs after the problem solver has taken some time away from the problem (i.e., incubation). Here we used Compound Remote Associates problems to examine how incubation affects the subjective experience of insight at different levels of problem fixation. We hypothesized that incubation would elicit a mind-set change resulting in improved problem solving performance regardless of the initial level of fixation. Second, we predicted to the extent that insight reflects a person's assessment of mind-set change, the experience of insight would be more likely after incubation. Results were consistent with these predictions. These findings suggest that the role of incubation in producing insight may have more to do with changing mind-set than forgetting information that fixates problem solvers.
\end{abstract}

Keywords: creativity; fixation; incubation; insight; problem solving

\section*{Introduction}

People frequently describe solving problems with either an analytic, step-by-step process, or a comparatively unconscious process resulting in unexpected answers (Boden, 1994; Morrison, in press; van Steenburgh et al., 2012). In the latter situation people show little ability to predict their sudden insight (Metcaffe, 1986), yet have great confidence in the solution that seemingly came from unconscious processing (Simonton, 2012; Smith \& Ward, 2012). This experience often follows a time away from the problem, also known as incubation (van Steenburgh, et al., 2012).

Insight has been studied using a variety of different approaches. Beginning with the Gestalt psychologists, researchers attempted to create problems where the experience of insight was more likely (e.g., Duncker's (1945) Candle Problem; Katona's (1940) Matchstick Arithmetic Problems; Mednick's (1962) Remote Associates Problems). Using these types of problems researchers have examined the experience of insight, for instance by asking participants to monitor their problem solving progress in situ (Melcalfe, 1986) or asking participants to report whether they experienced insight upon problem completion (Bowden \& Jung-Beeman, 2003a). The latter approach allows researchers to make post hoc comparisons between problems solved with and without insight on a problem-by-problem basis for each participant.
Alternatively, some studies have examined how problem-solving context could facilitate insight solutions (e.g., Barid et al., 2012; Kounios et al., 2008; Smith \& Blankenship, 1991; Storm, 2010, 2011; Wallas, 1926). For instance, Smith and Blankenship (1989) argued that incubation allows problem solvers to forget (or perhaps inhibit) mental representations resulting in fixation and thereby achieve an insight solution.

\section*{The role of incubation in promoting insight}

Building on an earlier study by Smith and Blankenship (1989), Kohn and Smith (2009) asked participants to solve remote associates problems (Mednick, 1962) in which participants must discover a single word that is a remote associate of three different words. Prior to attempting to solve each problem participants completed an initial task
(a) Unrelated - Direct Conditions

(b) Blocking - Incubate
\begin{tabular}{|c|c|c|c|}
\hline Two-Word Task & \multicolumn{3}{|r|}{Compound Remote Associates Problem} \\
\hline 20 Seconds & First Epoch 20 Seconds & Digit Monitoring Task 40 Seconds & Second Epoch 10 Seconds \\
\hline Blocking: Cross Rain Out & \multicolumn{2}{|l|}{Cross Rain Tie \(\Rightarrow 2-8-3-9-5 \Rightarrow\)} & Cross Rain Tie \\
\hline  & \multicolumn{2}{|l|}{\(\int \begin{gathered}\text { Moving to Digit } \\ \text { Span contingent } \\ \text { on not solving in } \\ \text { first epoch. }\end{gathered}\)} & \[
\sqrt{ } \sqrt{6}
\] \\
\hline \begin{tabular}{l}
Cross Out \\
Rain Out
\end{tabular} & Crossbow Rainbow Bowtie & \[
\begin{gathered}
2 \\
(3-9,9-5)
\end{gathered}
\] & Crossbow Rainbow Bowtie \\
\hline
\end{tabular}

Figure 1: (a) Unrelated - Direct and (b) Blocking - Incubate example trials. In Unrelated Compound Remote Associates (CRA) trials the preceding Two Word Phrase Task (TWPT) problem has no words in common with the CRA problem while in Blocking CRA trials the preceding TWPT problem contains two of the CRA problem words which pair with a third word that is not the correct answer for the CRA problem, thereby increasing CRA problem fixation. In Direct CRA trials participants have two contiguous epochs to try to solve the CRA problem, while in Incubate CRA trials the two epochs are separated by a 40 s incubation period in which participants perform the Digit Monitoring Task (DMT).
designed to manipulate the level of fixation experienced while trying to solve the remote associates problems. Participants briefly tried to solve each remote associates problem and then were given either a second continuous solution period or a brief 30s incubation period during which they performed a working-memory distractor task. Kohn and Smith found a trend towards participants showing an improvement in performance for problems on which they were initially more fixated and received an incubation period. Using a different type of insight problem, Baird and colleagues (2012) also found a benefit for incubation; however, the greatest benefit was found not from a difficult distractor task or simple rest during incubation, but rather from a task designed to promote mind-wandering. This latter result suggests that the benefit of incubation may not be to help participants overcome fixation, but rather, to promote the appropriate cognitive processing conducive to insight. Likewise, sleep studies by Cai and colleagues (2009) demonstrated that implicit priming of answers to unsolved Remote Associates Problems helped participants solve the problems after REM sleep compared to non-REM sleep or an equivalent rest period. This result suggests that time alone is insufficient for incubation effects, but rather solutions involving insight require a change in the underlying cognitive processing used for problem solving. However, none of these studies specifically asked participants whether they had experienced insight while solving the problem.
Bowden and Jung-Beeman (2003a) developed a subjective measure of insight for use with Compound Remote Associates problems (CRA; Bowden \& JungBeeman, 2003b) variants of Mednick's (1962) classic Remote Associates Task problems. Specifically Bowden and Jung-Beeman (2003a) asked participants after they had
solved a CRA problem to report via a numeric scale, how much insight they had experienced. Jung-Beeman and colleagues (Bowden \& Jung-Beeman, 2003b; JungBeeman et al., 2004; Kounios et al, 2006, 2008) have used various versions of this methodology to perform post hoc sorting of problems based on the participant's subjective experience. Using this methodology along with various neuroimaging methods they found evidence that right anterior superior temporal gyrus, a brain area associated with semantic integration, was specifically engaged just prior to CRA problems that participants reported solving with insight (Jung-Beeman, et al., 2004). Importantly, they also found evidence for neural activity indicative of visual gating just prior to the right temporal activity suggesting that a part of solving with insight might involve inhibiting the external world in favor of subconscious processing. Likewise, Kounios and colleagues (2008) identified this same neural signature before participants had initially seen problems they subsequently reported solving with insight, suggesting that the visual gating was likely indicative of a different problem solving strategy (Kounios et al., 2008).

\section*{Current Study}

The purpose of this study was twofold. First we wanted to explore whether taking time away from a problem (i.e., incubation) contributes to the subjective experience of insight. Second, to investigate whether incubation specifically helps participants overcome fixation, we adapted Kohn and Smith's (2009) paradigm for use with CRA problems (Bowden and Jung-Beeman, 2003b) and a subjective measure of insight. Specifically, we used Kohn \& Smith's two-word task to manipulate the degree of fixation prior to attempting to solve a CRA problem. We then manipulated incubation by either giving participants a second immediate opportunity to solve the problem, or


Figure 2: In Epoch 1, there was a reliable effect of blocking on CRA resolution rates demonstrating the effectiveness of the TWPT problem fixation manipulation. In Epoch 2, there was a reliable effect of incubation, with no interaction with initial TWPT induced fixation. Error bars represent \(\pm 1\) SEM.
instead gave them a period of incubation where they performed a working-memory distractor task. Whenever participants solved CRA problems, they were asked to report whether they experienced insight or not. This procedure thus allows us to evaluate whether insight is more likely after incubation and whether insight solutions were likely the result of release from problem fixation.

We hypothesized that incubation with a mild workingmemory distractor would elicit a mind-set change resulting in improved CRA problem solving performance regardless of the initial level of fixation. Second, we predicted that if the experience of insight reflects a person's assessment of mind-set change they would report greater insight on successfully solved problems after incubation than if they simply continued working on the problems without an incubation period.

\section*{Method}

\section*{Participants}

Eighty undergraduate students ( 60 female) from Loyola University Chicago participated in the experiment. Participants gave informed consent to take part in the study. The Loyola University Chicago Institutional Review Board approved all recruitment methods and procedures.

\section*{Task Descriptions}

Three tasks implemented in e-Prime 2.0 were used in this experiment. The primary task consisted of Compound Remote Associate problems (CRA; Bowden \& JungBeeman, 2003b). Each CRA problem consists of three unrelated words that can each be paired with a fourth target word that is a remote associate of each of the cue words to
make three compound word pairs (see Figure 1 for an example problem).
After the methods of Kohn and Smith (2009), we manipulated CRA problem fixation through use of a preceding Two-Word Phrase Task (TWPT) problem corresponding to each CRA problem. This task required participants to combine three presented words, two of which were from the corresponding CRA problem, into two two-word phrases (see Figure 1). This was intended to create a strong association for two of the CRA words to a word that was not the correct CRA answer, and thereby induce CRA problem fixation. We used the corresponding TWPT problem before the CRA problem in the Blocking condition (see Figure 1b), while we used an unrelated TWPT problem created for a different CRA problem in the Unrelated condition (see Figure 1a).
Lastly, we used a Digit-Monitoring Task (DMT; Kohn \& Smith, 2009) as the distractor task during incubation. In the DMT participants saw a series of digits from 1 to 9 presented one digit each second for 40 s. Participants were to track the total number of times that two odd digits were presented in a row and report that at the end of the incubation period.

\section*{Testing Procedure}

Forty-four CRA problems were rotated between four counterbalanced conditions (i.e., Unrelated/Direct, Unrelated/Incubate, Blocking/Direct, Blocking/Incubate; see Figure 1 for a schematic of two of the conditions). Each trial began with a TWPT problem for 20s followed by a CRA problem. On Direct trials if the participant did not solve the CRA problem in 20s (Epoch 1) they were given 10 additional seconds to solve the problem (Epoch 2). On Incubate trials that they did not solve in 20 s they performed the DMT for 40 s and then were given an additional 10s to solve the CRA problem. To encourage


Figure 3: In the first epoch, reports of insight were significantly higher in the unrelated condition suggesting that overcoming fixation was not responsible for the experience of insight. In the second epoch reports of insight were greater following incubation suggesting that the incubation task helped participants to elicit a mind-set change resulting in an insight solution. Error bars represent \(\pm 1\) SEM.
participants to form links between the TWPT and the CRA problems we used 6 additional CRA problems in the Helping condition. In these problems the correct answer for the CRA problem was given as the third word in the TWPT problem.

The definition of insight given to subjects was taken from Jung-Beeman et al. (2004). Briefly, the feeling of insight was described as a sudden experience in which a fully formed answer came to mind all at once. Upon solving a CRA, subjects were asked if they experienced insight. The subjects responded verbally with either yes or no.

\section*{Results}

Due to the CRAs being divided into a first 20s epoch and a second 10s epoch, accuracy was calculated using resolution rates (Kohn \& Smith, 2009). For the first epoch the Resolution rate was simply equal to the proportion solved correctly. For the second epoch we corrected for the number of problems solved in the first epoch and used the number of problems attempted during the second epoch as the denominator in the proportion correct calculation.

Resolution rate in the first epoch was impacted by fixation with participants solving more problems when they were preceded by a TWPT that did not result in greater fixation (see Figure 2; \(t(79)=5.6, p<.001\) ). Next we evaluated whether performance in Epoch 2 was impacted by incubation and whether this interacted with our fixation manipulation. A two-way within subjects ANOVA yielded a main effect of incubation (see Figure 2; \(F(1,79)=11.5, p=.001, \square_{\mathrm{p}}^{2}=.13\) ), but no main effect of fixation \(F(1,79)=.48, p=.5, \square_{\mathrm{p}}^{2}=.006\) ) and no interaction \(\left.F(1,79)=.73, p=.4, \square_{\mathrm{p}}^{2}=.009\right)\). Following the analysis of Kohn and Smith (2009) we also performed planned comparisons to look at the effect of incubation on Blocking and Unrelated trials independently. As in Kohn
and Smith's study participants showed a reliable difference in CRA resolution rate with respect to incubation in the Blocking condition \(\left(F(1,79)=12.0, \mathrm{p}=.001, \square_{\mathrm{p}}{ }^{2}=.13\right)\). However, unlike Kohn and Smith we found a trend towards a difference for the unrelated condition as well \(\left(F(1,79)=2.8, \mathrm{p}=.10, \square_{\mathrm{p}}^{2}=.03\right)\), consistent with our failure to find a reliable interaction between incubation and fixation. Thus, overall our results suggest that incubation aided in CRA problem solving regardless of the level of fixation as manipulated by the TWPT.
Overall, \(62 \%\) of all correct answers were answered with insight and \(38 \%\) were answered without insight. In an effort to measure participant's subjective experience of insight within each condition, we calculated an insight score for each participant by subtracting total number of correct non-insight answers from their total number of correct insight answers and dividing by the resolution rate.
Insight score in the first epoch was impacted by fixation with participants reporting greater insight on solution when they had less fixation as manipulated by the Two-Word Task (see Figure 3; \(t(79)=2.6, p=.012\) ). Next we evaluated whether the experience of insight in Epoch 2 was impacted by incubation and whether this interacted with our fixation manipulation. A two-way within subjects ANOVA yielded a main effect of incubation (see Figure 3; \(F(1,79)=9.0, p=.004, \square_{\mathrm{p}}^{2}=.10\) ), but no main effect of fixation \(F(1,79)=2.3, p=.14, \square_{\mathrm{p}}^{2}=.03\) ) and no interaction \(\left.F(1,79)=.78, p=.4, \square_{\mathrm{p}}^{2}=.01\right)\). Our results suggest that incubation increased the experience of insight, just as it aided solution performance. Likewise, the experience of insight does not appear to be majorly impacted by the initial degree of problem fixation.

\section*{Discussion}

Using a similar incubation and fixation paradigm with different remote associates problems, Kohn and Smith
(2009) reported that incubation led to higher resolution rates when participants were subjected to a task intended to cause problem fixation. They suggested that this improvement was due to distraction during incubation helping participants overcome problem fixation by forgetting wrong associations. In our study, we found that in spite of a strong initial fixation effect, incubation helped participants solve problems regardless of the level of fixation. In addition, participants experienced greater insight when they successfully solved problems after incubation regardless of fixation compared to when they successfully solved problems in a continuous period (Direct condition). Our results suggests that incubation does contribute to the experience of insight, and that incubation likely contributes to insight problem solving in ways other than just through forgetting fixation.

Recently Baird and colleagues (2012) presented evidence suggesting that what people do during incubation affects how likely they are to solve insight problems. Importantly, they found that a more demanding task resulted in less improvement than a less demanding task that encouraged mind wandering. Likewise, Cai and colleagues (2009) found that when participants experienced REM sleep during a Remote Associates Task incubation period they were more likely to benefit from an implicit semantic clue prior to incubation than if they had non-REM sleep or simply rested during incubation. Like our results, these findings suggest that something more than just forgetting must occur during incubation to facilitate insight.

One possible role for incubation may be to shift the mood of the participant. In our study when participants solved CRA problems during the first epoch prior to incubation they reported less insight when they had previously solved a TWPT problem intended to create CRA problem fixation than when they solved an unrelated TWPT problem (see Figure 3 Epoch 1). It is possible that the frustration resulting from fixation may encourage a negative mood. Several previous studies have suggested that participants are more likely to solve insight problems when they are in a positive mood (e.g., Isen, Daubman, \& Nowicki, 1987; Subramaniam et al., 2009). Subramaniam and colleagues showed that when people were high in selfreported positive affect prior to testing they were more likely to solve CRA problems and report insight. Van Steenburgh and colleagues (2012) have speculated that this effect of positive affect may be due to the ability of positive affect to encourage a broadening of attention (see also Rowe, Hirsch, \& Anderson, 2007). A broad attentional focus has long been known to be associated with creative behavior (e.g., Ansburg \& Hill, 2003; Mendelsohn \& Griswold, 1966). While it seems unlikely that performing the DMT incubation task in the present study would likely elicit a positive mood it is possible that the shift away from being stuck on a problem may result in at least a less negative mood perhaps resulting in a broader attentional mindset.

While our findings do support the idea that incubation can contribute to a change in mindset that aids in solving problems with insight, much remains to elucidate the precise nature of cognitive change that occurs during incubation.

\section*{Acknowledgments}

The authors thank Stephanie Hare, Leonidas Skiadoupolis, Kelly Brandstadt, and Krishna Bharani for assistance in collecting data and for excellent technical support and Mark Beeman and Marcia Grabowecky for helpful discussions. The American Federation of Aging Research/Rosalinde and Arthur Gilbert Foundation, the Illinois Department of Public Health, and the Loyola University Chicago Deans of Arts and Sciences and the Graduate School provided generous support.

\section*{References}

Ansburg, P. I., \& Hill, K. (2003). Creative and analytic thinkers differ in their use of attentional resources. Personality and Individual Differences, 34, 1141-1152.
Baird, B., Smallwood, J., Mrazek, M. D., Kam, J. W., Franklin, M. S., \& Schooler, J. W. (2012). Inspired by distraction: Mind wandering facilitates creative incubation. Psychological Science, 23(10), 1117-1122.
Bowden, E. M., \& Jung-Beeman, M. (2003). Aha! insight experience correlates with solution activation in the right hemisphere. Psychonomic Bulletin \& Review, 10(3), 730-737.
Bowden, E. M., \& Jung-Beeman, M. (2003). Normative data for 144 compound remote associate problems. Behavior Research Methods, Instruments, \& Computers: A Journal of the Psychonomic Society, Inc, 35(4), 634639.

Bowden, E. M., Jung-Beeman, M., Fleck, J., \& Kounios, J. (2005). New approaches to demystifying insight. Trends in Cognitive Sciences, 9(7), 322-328.
Cai, D. J., Mednick, S. A., Harrison, E. M., Kanady, J. C., \& Mednick, S. C. (2009). REM, not incubation, improves creativity by priming associative networks. Proceedings of the National Academy of Sciences of the United States of America, 106(25), 10130-10134.
Duncker, K., \& Lees, L. S. (1945). On problem-solving. Psychological Monographs, 58(5), i-113.
Isen, A. M., Daubman, K. A., \& Nowicki, G. P. (1987). Positive affect facilitates creative problem solving. Journal of Personality and Social Psychology, 32, 11121131.

Jung-Beeman, M., Bowden, E. M., Haberman, J., Frymiare, J. L., Arambel-Liu, S., Greenblatt, R., . . . Kounios, J. (2004). Neural activity when people solve verbal problems with insight. PLoS Biology, 2(4), e97.
Katona, G. (1940). Organizing and memorizing. New York: Columbia University Press.
Kohn, N., \& Smith, S. M. (2009). Partly versus completely out of your mind: Effects of incubation and distraction
on resolving fixation. Journal of Creative Behavior, 43(2), 102-118.
Kounios, J., Fleck, J. I., Green, D. L., Payne, L., Stevenson, J. L., Bowden, E. M., \& Jung-Beeman, M. (2008). The origins of insight in resting-state brain activity. Neuropsychologia, 46(1), 281-291.
Kounios, J., Frymiare, J. L., Bowden, E. M., Fleck, J. I., Subramaniam, K., Parrish, T. B., \& Jung-Beeman, M. (2006). The prepared mind: Neural activity prior to problem presentation predicts subsequent solution by sudden insight. Psychological Science, 17(10), 882-890.
Kounios, J., Frymiare, J. L., Bowden, E. M., Fleck, J. I., Subramaniam, K., Parrish, T. B., \& Jung-Beeman, M. (2006). The prepared mind: Neural activity prior to problem presentation predicts subsequent solution by sudden insight. Psychological Science, 17(10), 882-890.
Mednick, S. A. (1962). The associative basis of the creative process. Psychological Review, 69, 220-232.
Mendelsohn, G. A., \& Griswold, B. B. (1966). Assessed creative potential, vocabulary level, and sex as predictors of the use of incidental cues in verbal problem solving. Journal of Personality and Socia Psychology, 4, 421423.

Metcalfe, J. (1986). Premonitions of insight predict impending error. Journal of Experimental Psychology: Learning, Memory, and Cognition, 12(4), 623-634.
Morrison, R. G. (in press). Problem solving. In M. Aminoff \& R. Daroff (Eds.), The encyclopedia of the neurological sciences ( \(2^{\text {nd }}\) Edition). San Diego: Academic Press. http://www.canlab.org/uploads/2/8/1/3/2813551/morriso n_problemsolving_2ed_ens.pdf
Rowe, G., Hirsch, J. B., \& Anderson, A. K. (2007). Positive affect increases the breadth of selective attention. Proceedings of the National Academy of Sciences USA, 104, 383-388.
Simonton, D.K. (2012). Genius. In K. J. Holyoak \& R. G. Morrison (Eds.), Oxford handbook of thinking and reasoning (pp. 492-509). New York: Oxford University Press.
Smith, S. M., \& Blankenship, S. E. (1991). Incubation and the persistence of fixation in problem solving. The American Journal of Psychology, , 61-87.
Smith, S.M., \& Ward, T. (2012). Cognition and the creation of ideas. In K. J. Holyoak \& R. G. Morrison (Eds.), Oxford handbook of thinking and reasoning (pp. 456-474). New York: Oxford University Press.
Storm, B. C., \& Angello, G. (2010). Overcoming fixation. creative problem solving and retrieval-induced forgetting. Psychological Science, 21(9), 1263-1265.
Storm, B. C., Angello, G., \& Bjork, E. L. (2011). Thinking can cause forgetting: Memory dynamics in creative problem solving. Journal of Experimental Psychology.Learning, Memory, and Cognition, 37(5), 1287-1293.
Subramaniam, K., Kounios, J., Parrish, T. B., \& JungBeeman, M. (2009). A brain mechanism for facilitation
of insight by positive affect. Journal of Cognitive Neuroscience, 21(3), 415-432.
van Steenburgh, J.J., Fleck, J.I., Beeman, M., \& Kounios, J. (2012). Insight. In K. J. Holyoak \& R. G. Morrison (Eds.), Oxford handbook of thinking and reasoning (pp. 475-491). New York: Oxford University Press.
Wallas, G. (1926). The art of thought. New York: Harcourt Brace.

\title{
Modeling a reaction time variant of the Perruchet effect in humans
}

\author{
Amy McAndrew (am375@exeter.ac.uk) Fayme Yeates Frederick Verbruggen Ian P.L. McLaren \\ School of Psychology, College of Life and Environmental Sciences, University of Exeter, UK.
}

\begin{abstract}
This paper presents a reaction time (RT) experiment that follows on from the work of Perruchet, Cleeremans, and Destrebeceqz (2006), investigating the extent to which reaction times (RTs) are governed by the conscious expectancy of a particular response. In this experiment, participants were presented with a single stimulus (which we will call the conditioned stimulus; CS) followed by one of two outcomes (which we will call unconditioned stimuli; USs); to which participants had to make an appropriate instrumental response. On every trial we recorded the time taken to make this response and participants were asked to rate their expectancy that one of the USs (US1) was going to occur. We found that the expectancy rating for US1 correlated negatively with RT on US1 trials. Over successive runs of reinforcement, when participants rated US1 as less likely to occur they were slower to respond to US1 (lower ratings, higher RTs). When we calculated the expectancy for US2 as the complement of that for US1 expectancy, expectancy of US2 correlated positively with RTs. Thus, across runs of reinforcement, participants responded more quickly to US2 when considering US2 less likely (low rating, low RT). We argue that the requirement to make a conscious expectancy rating results in participants attending more to US1 occurrences than those of US2. This results in a qualitatively different relationship between conscious expectancy and automatic responses that cannot be reconciled by a single processing system account. A dual processing system explanation of learning is proposed to explain these results. In support of this position, we successfully modeled our US2 RT data using a modified version of the Augmented simple recurrent network (Yeates, Jones, Wills, McLaren, \& McLaren, 2013).
\end{abstract}

Keywords: Perruchet effect; Modeling; Dual processing systems; AugSRN; Associative learning

\section*{Introduction}

Recently, there has been a lively debate on the extent to which learning is governed by a single processing system or dual processing systems (e.g. McLaren, Green, \& Mackintosh, 1994). A single processing system view advocates one conscious reasoning process (e.g. Lovibond, \& Shanks, 2002). From this viewpoint, conditioned responding (CR) obtained in an instrumental conditioning paradigm is driven by contingency knowledge that develops during the course of conditioning between a conditioned stimulus (CS) and unconditioned stimulus (US). Within a dual processing system framework, associative automatic processes can be responsible for the CR without explicit contingency knowledge. Based on this account, an associative link forms between a representation of the CS and representation of the US. Presentation of the CS activates the link between the CS and US, which activates the US representation, which then produces a CR.

One of the most convincing sources of evidence (Mitchell, De Houwer, \& Lovibond, 2009; Shanks, \& St John, 1994) for dual processing systems is the Perruchet effect (Perruchet, 1985). In the reaction time (RT) version of this experiment employed by Perruchet, Cleeremans, and Destrebeceqz (2006), participants hear an auditory tone (the CS ) on every trial. Half the time the CS is followed by a visual US to which participants have to make a keypress response. On the other half of the trials there is no US and participants are not required to make a response. Participants make an online expectancy rating on every trial regarding the extent to which they think the US is going to appear on that trial.

Across successive CS-US (reinforced) trials, expectancy ratings that the US will occur decrease. However, after experiencing runs of nonreinforced, CS-noUS, trials participants' ratings indicate they think it more likely that the US will occur; and thus, that a response is more likely to be required. This is consistent with the gambler's fallacy phenomenon (Burns, \& Corpus, 2004). In contrast, the CR (the instrumental response to the US measured by RT) gets faster (improves) with successive reinforcement. This means consecutive CS-US trials result in shorter RTs, whereas runs involving an absence of the US result in slower responding. This pattern of responding is hard to reconcile with the gambler's fallacy, as participants become quicker to respond to the US at the same time as their expectancy of the US (and thus their expectancy that they are required to make a response) decreases. An associative account can, however, explain the change in RT with reinforcement history, as over successively reinforced trials the associative link between the CS representation and the US representation becomes stronger, leading to faster RTs. This link is extinguished and weakened by the absence of the US on the CS-noUS trials, leading to slower RTs. Thus, a dual processing systems account is required to explain both the conscious processes underlying expectancy along with the RT pattern that captures our automatic, associative learning about CS-US relationships (McLaren, Green, \& Mackintosh, 1994).

The experiment presented here aims to further investigate the effects observed in a RT version of the Perruchet paradigm, and to provide support for a dual processing systems account of learning. To build on the original experiments, we presented participants with two USs in order to obtain RT data on every trial and to keep the demands of each trial consistent. We were therefore able to take a measure of CR for the two USs separately and compare these to expectancy of each US. If RT and expectancy of the US are found to follow different trends this would imply that a single processing system
explanation of learning would be unable to explain the results and that a dual processing systems account would be more appropriate. If our assumptions regarding the nature of the processes underlying RT performance are correct, we should be able to model these associatively. Therefore, to assess this claim, we used a model of human learning (the revised augmented simple recurrent network: RASRN; Yeates, Jones, Wills, McLaren, \& McLaren, 2013) in an attempt to simulate the instrumental responding of participants in this experiment.

\section*{Method}

\section*{Participants}

64 University of Exeter students ( 13 men and 51 women) were recruited for course credit to participate in this experiment. Their ages ranged from 18 to 49 years, with a mean age of 21 .

\section*{Design and Stimuli}

The CS was visually presented to participants as a brown cylinder ( \(11 \times 7 \mathrm{~cm}\) ) in the centre of a white screen. The words "Peanut Butter" and "Brown Sugar" were the two USs that followed the presentation of the CS. Both USs were presented to and counterbalanced across each participant as US1 and US2. Each of the USs was presented half the time after the CS, forming a partial reinforcement schedule where the occurrence of each US was equally likely.

In a typical Perruchet design, we are interested in runs of reinforced and non-reinforced trials, therefore a repeatedmeasures factor of run length (the number of a given trial type that occur consecutively in a row) was constructed. There were 8 levels of this factor; \(-4,-3,-2,-1,+1,+2,+3\), and +4 . When analyzing the sequence of trials given to each participant in this experiment, we can examine repetitions of the same US (D, different trials) or repetitions of the opposing US (S, same trials) as equivalents of these positive and negative runs of trials, respectively. A CR measurement is taken on the trial after the run itself, thus when considering US1 trials, a +2 trial would have involved two consecutive CS-US1 trials prior to this, whereas a -3 trial would have been preceded by a run of three CS-US2 trials (see Table 1 for an example of how runs are labeled within the sequence).

Table 1. An example of a sequence of CS-US pairings and the corresponding run lengths of these trials. These are labeled both in terms of classic Perruchet positive and negative runs; and in terms of same (S) and different (D) runs. Trial type indicates whether US1 or US2 is paired with the CS (which occurs on each trial).
\begin{tabular}{cccccccc}
\hline \begin{tabular}{c} 
Trial \\
type
\end{tabular} & US1 & US1 & US2 & US2 & US2 & US1 & US2 \\
\hline \begin{tabular}{c} 
Run \\
length
\end{tabular} & & +1 & -2 & +1 & +2 & -3 & -1 \\
\cline { 2 - 8 } & & S1 & D2 & S1 & S2 & D3 & D1 \\
\hline
\end{tabular}

We aimed to compile sequences of US1 and US2 trials that involved these same (S/positive) and different (D/negative) runs from 1 to 4 , following a binominal
distribution as shown in Table 2. However, the original Perruchet experiments only comprised of one CS and one US, while the current experiment involves two USs. As each run has to end in the opposite trial type (e.g. a US1 run would have to end in a US2 trial), two 'different' runs of length five are included in each block. These are a requirement for the sequence, are counterbalanced across the US type across blocks and excluded from the analysis; and so are not discussed further.

Table 2. The binomial distribution of run lengths.
\begin{tabular}{ccccccccc}
\hline Run & -4 & -3 & -2 & -1 & +1 & +2 & +3 & +4 \\
\cline { 2 - 9 } \begin{tabular}{c} 
length
\end{tabular} & D4 & D3 & D2 & D1 & S1 & S2 & S3 & S4 \\
\hline \begin{tabular}{c} 
Number \\
of runs
\end{tabular} & 2 & 4 & 8 & 16 & 16 & 8 & 4 & 2 \\
\hline
\end{tabular}

In this experiment, each participant experienced two blocks of 57 trials, which comprised of unique, randomized sequences of run lengths. These sequences were constructed using MatLab. We measured both expectancy and RT as our dependent variables and compared them across run length for both USs separately.

\section*{Procedure}

A cover story was given to participants, who were told they were playing the role of a doctor seeing a number of patients with both diabetes and a nut allergy. Participants were exposed to the CS for 5 seconds on each trial and were told that this brown cylinder could represent either peanut butter or brown sugar. During this time, participants had to make a rating on a scale of 1 to 9 regarding the extent they thought this trial would be a US1 trial. For half of the participants, peanut butter was US1; for the other half, brown sugar was US1. If US1 was peanut butter, they were told that a rating of 1 would indicate: "I definitely do not think the patient will need adrenaline"; up to a rating of 9: "I definitely think the patient will need adrenaline". Adrenaline was replaced by insulin when brown sugar was US1. Participants were told that half the time "peanut butter" would appear after the CS and on the other half of trials "brown sugar" would appear. One of these stimuli (the US) was then presented immediately after the CS. Participants were instructed to respond as quickly as possible to the stimuli to administer adrenaline to "peanut butter" and insulin to "brown sugar" with left Ctrl and left Alt keys (counterbalanced) to avoid anaphylactic shock or hyperglycemia, respectively. The US remained onscreen until a response was made, followed by a variable ITI of 2 to 5 seconds before the next trial commenced. Participants were allowed a short break between the two blocks to allow them to rest.

\section*{Results}

Both RT and expectancy data were collected using MatLab and PsychToolbox (Brainard, 1997). RTs for US1 and US2 were recorded on each trial in milliseconds (ms). Any RTs over 1 second were excluded from the analyses. The mean RT for each run length for US1 and US2 can be seen in Fig. 1 top panel. In terms of expectancy, participants
were required to make ratings based on the extent they thought US1 was going to occur. Therefore, we divided the data into average expectancy for US1 on US1 trials and average expectancy for US1 on US2 trials for each participant on each run length, see Fig. 1 bottom panel.

Reaction times


\section*{Expectancy of US1}


Figure 1. The top panel displays the RT data for US1 and US2 across run length. The bottom panel displays the expectancy for US1 on US1 and US2 trials across run length.

A two-way repeated-measures analysis of variance (ANOVA) was run on the RT data using the factors US (US1 versus US2) and run length ( \(-4,-3,-2,-1,+1,+2,+3\), \(+4)\). A significant interaction between US and run length was found, \(F(7,238)=2.58, M S E=0.025, p=.029\), as well as a significant linear trend interaction, \(F(1,34)=8.84, M S E\) \(=0.085, p=.005\). This indicates that there is a significant difference in US1 and US2 RTs across run length. From Fig. 1 top panel, it can be seen that US1 RTs appear to increase after a run of US1 trials (i.e. RT increases when run length increases), whilst US2 RTs decrease after a run of US2 trials (i.e. RT decreases when run length increases).

One-way repeated-measures ANOVAs were then used to analyze the US1 and US2 RT data separately. There is a highly significant main effect of run length for the US2 RTs, \(F(7,336)=6.21, M S E=0.07, p<.001\). There was also a significant linear trend decreasing from -4 to +4 across run length, \(F(1,48)=16.86, M S E=0.27, p<.001\). With regards to US1 RTs, however, the numerically increasing linear trend from -4 to +4 was not significant.

A two-way repeated-measures ANOVA was also run on the US1 expectancy data, again with the factors US and run length. A significant interaction between US and run length was found, \(F(7,371)=3.39, M S E=22.42, p=.017\), as well
as a significant linear trend interaction, \(F(1,53)=4.43, M S E\) \(=48.92, p=.040\). This indicates expectancy of US1 on US1 differs significantly from expectancy of US1 on US2 trials across run length. From Fig. 1 bottom panel, it appears that expectancy for US1 on US1 trials decreases across run length whilst expectancy of US1 on US2 trials increases across run length.

One-way repeated-measures ANOVAs were then used to analyze expectancy on US1 and US2 trials separately. There is a significant main effect of expectancy of US1 on US2 trials across run length, \(F(7,399)=2.51, M S E=9.78, p=\) .041, and a significant linear trend increasing from -4 to +4 , \(F(1,57)=5.38, M S E=33.78, p=.024\). With regards to expectancy of US1 on US1 trials, a marginally significant main effect of run length was found, \(F(7,392)=2.44, M S E\) \(=11.26, p=.051\). However, the decreasing numerical linear trend was not reliable.

\section*{Discussion}

Regarding the expectancy measure (Fig. 1, bottom panel), we should make it clear from the start that both lines on the graph reflect US1 expectancy, however we have split this by whether the rating was taken on a US1 or US2 trial. Expectancy for US1 on US1 and US2 trials can be explained by the gambler's fallacy phenomenon (Burns \& Corpus, 2004). Expectancy of US1 after a run of US1 trials numerically decreases, while expectancy of US1 after a run of US2 trials increases. Thus, after a run of US1 trials the participant thinks US2 is more likely to occur, so expectancy of US1 declines; but after a run of US2 trials the participant now believes it is US1s turn, so expectancy of US1 increases.

Within the RT data, participants' responses to US1 numerically increased as a function of run length. This indicates participants were faster to respond after successive CS-US2 trials, and therefore were slower after successive CS-US1 trials. We found a negative correlation between US1 expectancy and US1 RTs, \(r=-.871, n=8, p=.005\). Thus, after a run of CS-US1 trials participants made lower ratings that US1 would occur and were slower to make US1 responses. Therefore it would appear that a propositional explanation would be sufficient to explain this result, by simply claiming expectancy directly influenced RT.

Turning to US2, we propose that, logically, if a participant is expecting one US to happen then they are not expecting the other, so if a participant is expecting a US1 trial to occur then that implies they are not expecting a US2 trial. This would suggest expectancy of the two USs is complementary such that, if expectancy of US1 is the highest possible rating (9), then expectancy of US2 should be the lowest possible rating (1). We can assume that these sum \((9+1=10)\) and thus calculate expectancy for US2 as equal to 10 minus US1 expectancy. Based on this assumption, we can predict participants' expectancy of US2 on US2 trials as being the complement of their expectancy of US1 on US2 trials, see Fig. 2. If this supposition is true, then we have shown expectancy of US2 on US2 trials decreases as a function of run length. Therefore, higher ratings of US2 are made if participants have experienced a
run of US1 trials, and vice versa. This pattern of responding can be attributed to the propositional, gambler's fallacy phenomenon discussed previously.

\section*{Expectancy on US2 trials}


Figure 2. This graph displays expectancy of US1 on US2 trials and the hypothetical expectacny of US2 on US2 trials.

In order to verify if our inference regarding expectancy of US2 on US2 trials was correct, 32 of our participants carried out a further two experimental blocks to those described in the earlier method section. In these blocks, two (identical) cylinders were presented (successively) and participants had to make an expectancy rating to each. One cylinder required the participants to make a "peanut butter" rating, the other a "brown sugar" rating. Participants then had to make the appropriate RT response as in the previous blocks. Comparing participants expectancy of US1 on US1 trials and their expectancy of US2 on US1 trials, there was a highly significant negative correlation, \(r=-.969, n=8, p<\) .001. This shows that on US1 trials, if participants were for example, expecting a US1 trial they were not expecting a US2, and vice versa. Additionally, comparing expectancy of US1 on US2 trials and expectancy of US2 on US2 trials, there was also a highly significant negative correlation, \(r=-\) \(.944, n=8, p<.001\). This also shows that on US2 trials, if participants were expecting a US1 they were not expecting a US2. Therefore, our earlier assumption receives considerable empirical support from this check.

Given that expectancy of US2 on US2 trials decreases as a function of run length, interestingly we found that US2 RTs also decreased as a function of run length (see Fig. 1). Participants were faster to respond to US2 on a run of CSUS2 trials, even though their expectancy that US2 would occur had decreased. We have therefore demonstrated a positive correlation between expectancy of US2 on US2 trials and US2 RTs, \(r=.833, n=8, p=.010\). For example, after a run of CS-US2 trials, participants rate that a US1 trial is more likely (and therefore a US2 is less likely), yet are faster to respond to US2. It is consequently hard to reconcile this expectancy with the RT result if we take the position that a single propositional explanation could explain our data. We would argue that associative, link-based processes are required to explain the RTs for US2. One version of this would be that when a person experiences the CS followed by US2, a link is set up between the two representations of these stimuli. After a run of CS-US2 trials this would strengthen the link between these stimuli, resulting in a stronger CR (i.e. a faster key press response) to US2.

However, after a run of CS-US1 trials, the link between CS and US1 strengthens, but the link between the CS and US2 weakens (extinction). Hence, the more consecutive CS-US1 trials there are, the weaker the CR to US2 (i.e. the slower the RT). The results for US2 are in agreement with previous Perruchet RT experiments, in which a single propositional process cannot explain both the expectancy and RT data.

In one experiment we have shown two different results, one where expectancy and RT appear positively correlated, and another where they are negatively correlated. We have, as a consequence, proposed a dual processing systems explanation of the US2 result. We would now like to pursue this further, by speculating how associative and propositional processes could produce both results. We hypothesize that the difference between the two effects (for US1 and US2) lies in where participants' attention is focused. As participants are directed to focus on one US (US1), to which they are making expectancy ratings, this effectively manipulates the expression of both propositional and associative processing systems for that US. We assume that because participants are attending to US1, they spend less time thinking about US2 and this would suggest conscious reasoning processes are more focused on the processing of US1 than US2. If US2 is not being consciously processed (to the same extent) then changes in US2 performance in the experiment might be driven by an alternative processing system. By reducing attention to US2, we believe we have created an environment conducive to associative learning. In contrast, a large amount of cognitive resource is being directed to processing US1, and perhaps this has led conscious processes to play a larger role in RT performance for this outcome, and inhibited the expression of associative processes in this case.

\section*{Modeling}

To explore how associative processes might be driving instrumental responses to US2, we chose to simulate this experiment using an established model of associative learning. We chose the augmented SRN (Cleeremans, \& McClelland, 1991; as revised by Yeates, Jones, Wills, McLaren, \& McLaren, 2013), which is particularly wellsuited to this task as the simple recurrent network (SRN; Elman, 1990) was devised to account for learning that is observed across sequences of trials. Our aim was to ascertain the extent to which learning is driven by the development of associative strength between the CS and US2, or whether the sequential structure of the experiment (runs of US1 and US2) is what drives this result.

The model (see Fig. 3) is a connectionist network that feeds input activation to a hidden layer, which in turn feeds activation forward into an output layer, each employing the logistic activation function (Rumelhart, Hinton, \& Williams, 1985). The activation of the hidden layer is copied back into a set of context units on each trial, which are then fed into the hidden layer as input on the next trial. This recurrent loop provides the model with a memory of the hidden layer's representation of the last trial. Learning occurs through back-propagation of error correction, comparing output activation to expected responses. Connection weight
changes to represent both short- and long-term learning are calculated using fast and slow learning rates, respectively. Fast weights have a higher learning rate but decay more rapidly, and were introduced to the model by Cleeremans and McClelland (1991) to account for the short term priming effects evident in their data. The slow weights reflect more permanent learning that takes a longer time to develop due to the lower learning rate.

\section*{Revised Augmented SRN}


Figure 3. Architecture of the revised version of the Augmented SRN by Yeates et al. (2013)

The model in this simulation involved two output units to represent Ctrl and Alt keypress responses to US1 and US2. As well as the context units (copy of the previous trial hidden unit activation) there were five additional input units. These followed revisions to the SRN by Yeates et al. (2013, see for further discussion) and included both a representation of the previous response made (two units, one for US1 and one for US2) as well as a representation of the on-screen stimuli on the current trial (one CS unit and two US units, one to represent each of US1 and US2). The free parameters of the model were: 20 hidden units with the learning rates set at 0.4 and 0.533 for slow and fast learning rates, respectively (based on Jones, \& McLaren, 2009).

The model was run 64 times with random initial weights of between -0.5 and 0.5 to give the same \(n\) of networks as participants in the experiment. Each of these simulations used binary input activations representing the exact occurrence of the CS and USs taken from the unique sequences that each of the 64 participants were given. Mean square error (MSE) was calculated as an index of responding to the US on each trial from the squared difference between output activations and the expected activations for the two possible responses ( 0.1 and 0.9 for incorrect and correct response, respectively). Trials were analyzed according to run length and US, like the variables of interest used in the behavioral experiment.

We analyzed the MSE for each US using one-way repeated measure ANOVAs and thus examined the modeling data in the same fashion as the behavioral data. There was a main effect of run length in both US1, \(F(7,406)\) \(=1339.80, M S E=0.67, p<.001\), and US2, \(F(7,441)=\) \(1546.46, M S E=0.67, p<.001\). Thus, for both US1 and US2 MSE differed according to run lengths. Furthermore, we found that there was a highly significant linear contrast on run length for both USs, \(F(1,58)=2633.43, M S E=4.44, p\)
\(<.001\), and \(F(1,63)=2908.722, M S E=4.14, p<.001\), for US1 and US2 respectively. This is seen quite clearly in Fig. 4, which shows a decreasing linear trend for both USs (which do not differ significantly) across run length. It can also be seen from the graph the two functions of MSE lie almost entirely on top of one another. Thus, responding to both of these USs is extremely similar, both demonstrating a reduction in error as run length increases.

\section*{Modelling data}


Figure 4. Graph of the mean square error (MSE) of the model

When comparing the modeling data to the human data we are using MSE as an approximation to RTs, as this is what we consider to capture the automatic, associative relationship between CS and US. We can see that human RT responding to US2 has the same, decreasing function across increasing reinforcement as is produced by the AugSRN. This is supported by a significant positive correlation between run length on RT and MSE results for US2, \(r=\) \(.895, n=8, p=.003\). Clearly then, the Augmented SRN is a good model of human performance on US2 in our experiment, but a poor one for US1.

Further investigation, however, reveals that the basis for performance may not be the conventional associative explanation offered for the Perruchet effect. There is no doubt that transient fluctuations in the strength of CS-US associations could explain the results observed for US2. But, the Augmented SRN can also learn about the sequence of events that take place, rather than just in terms of CS-US associations; and with the parameters given in Yeates et al., (2013) it could be that the pattern shown in Fig. 4 is based on this type of learning, rather than CS-US learning. This can be investigated by running the same simulation, but with the CS unit permanently set to zero so that no change in CS-US associations is possible. When we did this, the same function emerged, see Fig. 5. Thus, we would appear to have evidence suggesting that transient changes in CS-US associations might not be the basis of the function shown in Fig. 4. This result is reminiscent of that reported by Mitchell, Wardle, Lovibond, Weidemann and Chang (2010), who were able to get a Perruchet type effect in an RT experiment without any CSs. We have essentially the same result in our simulation, using a model that is well known for its ability to generate sequential effects.

But if sequential effects are the correct explanation of our modeling result, the removal of all the input units from the model (leaving only the hidden and output layers) should
abolish this effect, as there would be nothing left in the model that could produce sequential effects (no input or copy-back from the hidden layer). However, when we did this, we found the same decreasing function in MSE as seen before in our previous simulations (see Fig. 5). This demonstrates that sequential effects are not necessarily driving our result, but rather that the associative fluctuations between the hidden and output units are.

\section*{Further modelling}


Figure 5. Graph of the MSE for the further modelling
At the beginning of these last simulations the hidden units have activation values of 0.5 (corresponding to zero input). Therefore, after a reinforced trial the link between any hidden unit and the output unit will be strengthened. Consequently, if another reinforced trial follows the previous one this link is again strengthened leading the model to produce \(a\) smaller MSE. In contrast, \(a\) nonreinforced trial weakens this link, and the MSE increases. Therefore, there is an associative explanation for the Perruchet effect that emerges from this model, just not the classic explanation as it is usually cited. It is worth emphasizing that it is an associative explanation that applies here, and not one based on conscious, cognitive expectancy of the US. The pattern seen for US1 in our empirical data follows that generated by the expectancy ratings given by our participants and is quite different from both the pattern seen for US2 and the pattern generated by our model simulating an explanation in terms of CS-US associations, sequential effects, or hidden to output layer connections. The correlation between human RTs and modeling data for US1 is negative and non-significant across run length, \(r=-\) \(.562, n=8, p=.148\). Thus, an associative explanation will not fit these data, and a more cognitive model is required.

\section*{General discussion}

This paper presents behavioral and modeling data based on a new RT variant of the classic RT Perruchet paradigm. In our behavioral experiment we produced a Perruchet-type effect whereby expectancy of US2 decreased as a function of run length while RT responses to US2 decreased. We have rejected a single processing system explanation of learning in favor of a dual processing systems argument to explain this result. The propositional, gambler's fallacy heuristic (Burns \& Corpus, 2004) explains why expectancy of US2 decreased as the run of CS-US2 trials increased, as participants are deciding that it is less likely another US2 trial will happen if they have experienced a run of US2
trials. However, within the RT data, after a run of CS-US2 trials participants are faster to respond to US2 despite low expectancy that US2 will occur. This seems paradoxical when considered from a single systems view, but an associative account can explain the RT result, in terms of fluctuating hidden-output unit associations, sequential effects or CS-US associations. Our feeling is that it would be possible to parameterise the Augmented SRN to produce the US2 pattern of results on the basis of any of these potential mechanisms, though it would appear that in our current simulations the effect is mainly carried by fluctuating hidden-output associations. Note, however, that in Fig. 5 the pattern is more pronounced when the input to the model is enabled (suggesting that sequential effects can contribute), and we have run other simulations that show that the presence or absence of a CS representation can also strengthen or weaken this effect indicating that CS-US associations can also be effective in this model. More research will be needed to determine which of these mechanisms is the correct explanation for our data.

In conclusion, the evidence for dissociable propositional and associative processes provided by Perruchet type RT experiments is perhaps stronger than we thought. Explaining these effects with reference to a single propositional system, however, is likely to prove a difficult challenge for theorists of that persuasion.

\section*{ACKNOWLEDGEMENTS}

This research was funded by an ESRC studentship to Amy McAndrew and an ESRC grant to Ian McLaren.

\section*{REFERENCES}

Brainard, D.H. (1997). The Psychophysics Toolbox, Spatial Vision, 10, 433-6. Burns, B.D., \& Corpus, B. (2004). Randomness and inductions from streaks: "Gambler's fallacy" vs. "Hot hand". Psych. Bulletin and Review, 11, 179-184. Cleeremans, A., \& McClelland, J.L. (1991). Learning the structure of event sequences. Journal of Experimental Psychology: General, 120, 235-253.
Elman, J.L. (1990). Finding structure in time. Cognitive Science, 14, 179-211.
Jones, F.W., \& McLaren, I.P.L. (2009). Human sequence learning under incidental and intentional conditions. Journal of Experimental Psychology: Animal Behavior Processes, 35, 538-553.
Lovibond P.F., \& Shanks, D.R. (2002). The role of awareness in Pavlovian conditioning: Empirical evidence and theoretical implications. Journal of Experimental Psychology: Animal Behavior Processes, 28, 3-26.
McLaren, I.P.L., Green, R.E.A., \& Mackintosh, N.J. (1994). Animal learning and the implicit/explicit distinction. In N.C. Ellis (Ed.) Implicit and explicit learning of languages (pp. 313-332). New York, NY: Academic Press.
Mitchell, C.J., De Houwer, J., \& Lovibond, P.F. (2009). The propositional nature of human associative learning. Behav. and Brain Sciences, 32, 183-246.
Mitchell, C.J., Wardle, S.G., Lovibond, P.F., Weidemann, G., \& Chang, B.P. (2010). Do reaction times in the Perruchet effect reflect variations in the strength of an associative link? Journal of Experimental Psychology: Learning Memory and Cognition. 36, 567-72.
Perruchet, P. (1985). A pitfall for the expectancy theory of human eyelid conditioning. Pavlovian Journal of Biological Sciences, 20, 163-170.
Perruchet, P., Cleeremans, A., \& Destrebeceqz, A. (2006). Dissociating the effects of autonomic activation and explicit expectancy on reaction times in a simple associative learning task. Journal of Experimental Psychology: Learning, Memory, and Cognition, 32, 955-965.
Rumelhart, D.E., Hinton, G.E., \& Williams, R.J. (1985). Learning internal representation by error propagation. In D. E. Rumelhart \& J. L. McClelland (Eds.), Parallel Distributed Processing: Explorations in the Microstructre of Cognition. (Vol. 1: Foundations.). Cambridge, MA: Bradford Book/MIT Press.
Shanks, D.R., \& St. John, M.F. (1994). Characteristics of dissociable human learning systems. Behavioural and Brain Sciences, 17, 367-447.
Yeates, F., Jones, F.W., Wills, A.J., McLaren, R.P., \& McLaren, I.P.L. (2013). Human sequence learning under incidental conditions. Journal of Experimental Psychology: Animal Behavior Processes, 39(2), 166-173.

\title{
Can US sensitization account for the electrodermal variant of the Perruchet effect?
}

\author{
Amy McAndrew (am375@exeter.ac.uk) \({ }^{\text {a }} \quad\) Gabrielle Weidemann \({ }^{\text {b }} \quad\) Ian P.L. McLaren \({ }^{\text {a }}\) \\ \({ }^{\text {a }}\) School of Psychology, College of Life and Environmental Sciences, University of Exeter, UK. \\ \({ }^{\mathrm{b}}\) School of Social Sciences and Psychology, University of Western Sydney, Australia
}

\begin{abstract}
During experiments employing Perruchet's (1985) paradigm there are runs of reinforced (CS-US) trials and non-reinforced (CS-noUS) trials. Conditioned responding (CR) is measured, for example, using eyeblink responses (Perruchet, 1985), reaction times (Perruchet, Cleeremans, \& Destrebeceqz, 2006), or changes in skin conductance (SCR; McAndrew, Jones, McLaren, \& McLaren, 2012), as well as an online measure of expectancy for the unconditioned stimulus (US). A double dissociation between CR and conscious expectancy of the US is typically found, whereby expectancy of the US decreases while the CR increases across runs of successively reinforced trials. A gambler's fallacy explanation can be offered for the expectancy data, whereas an associative explanation can be used to explain variations in the CR (consistent with the dual processing theory of McLaren, Green, \& Mackintosh, 1994). However, skeptics of this effect have proposed nonassociative explanations of the CR data seen in these experiments. They note that every CS-US pairing is confounded by the presence of the US. Therefore, it is possible that US sensitization, the phenomenon whereby repeated US presentations leads to stronger unconditioned responding to the US, could produce the increasing CR pattern with successive reinforcements (Weidemann, Tangen, Lovibond, \& Mitchell, 2009). Two experiments are presented investigating whether US sensitization can explain the recently published electrodermal version of the Perruchet effect.
\end{abstract}

Keywords: Perruchet effect; US sensitization; Dual processing systems

\section*{Introduction}

The Perruchet effect (Perruchet, 1985, Perruchet, Cleeremans, \& Destrebeceqz, 2006) is often cited as one of the most convincing pieces of evidence of a dissociation between explicit, conscious, propositional processes and implicit, automatic, associative processes (e.g. Mitchell, De Houwer, \& Lovibond, 2009). McAndrew, Jones, McLaren and McLaren (2012) ran an electrodermal variation of the classic Perruchet experiment in which participants saw a visual conditioned stimulus (CS), a brown cylinder, which was partially reinforced by an electric shock (the unconditioned stimulus, US). The participants made online expectancy ratings on every trial as to whether they thought the US was going to occur. Changes in their autonomic skin conductance response (SCR) were also measured as an index of conditioned responding (CR).

We found that the SCR increased over successive reinforcements, while expectancy of the US decreased across the same sequence of trials. This mirrored the original findings of Perruchet and colleagues in both the eyeblink (Perruchet, 1985) and reaction time (RT; Perruchet, Cleeremans, \& Destrebeceqz, 2006) variants of this paradigm. The gambler's fallacy heuristic (Burns \&

Corpus, 2004), an explicit, propositional phenomenon, explained the expectancy data, implying that as participants experienced successive runs of reinforced (CS-US) trials, they were less likely to rate the subsequent trial as being paired with a US. Conversely, when participants experienced runs of successively non-reinforced (CS-noUS) trials they rated it as increasingly likely they would experience a US. However, this explanation did not apply to the SCR data, but an associative account did (e.g. McLaren, Forrest and McLaren, 2012). SCR increased over successive reinforcements, meaning that the CR was strongest when the participants had experienced a run of CS-US trials. In associative terms, during this type of Pavlovian conditioning, the link between the representation of the CS and the representation of the US was strengthened by the successively reinforced trials, producing a larger CR. However, after a run of CS-noUS trials, the link between the representations of the CS and the US was weakened by extinction, causing smaller changes in SCR and therefore a weaker CR. Hence, these results are consistent with a dual processing systems account of learning, with an explicit propositional system generating the expectancy data and an associative system the changes in SCR.

The Perruchet effect is one of the most compelling examples of dual processing systems due to the simultaneous measurement of CR and expectancy. Previous research demonstrating dissociations between these two variables has often involved subliminal presentations of the CS (e.g. Balderston, \& Helmstetter, 2010), but this research is often criticized about whether the presented stimuli are truly subliminal (e.g. Mitchel, De Houwer, \& Lovibond, 2009). Alternatively, researchers have attempted to use postconditioning questionnaires to assess contingency awareness, however it has been argued this type of measure could be subject to interference or forgetting influencing the reliability and validity of the awareness measure (Lovibond, \& Shanks, 2002). The Perruchet paradigm however, overcomes these problems and is a more convincing demonstration of a double dissociation.

However, the dual processing system account of the Perruchet effect depends critically on the assumption that the linear trend in CRs is the result of associative learning. Alternately the pattern of CRs could be accounted for by US sensitization; this effect refers to the increase in unconditioned responding (UR) seen when there is repeated exposure to a US (Weidemann, Tangen, Lovibond, \& Mitchell, 2009). In the Perruchet experiments, every CS-US pairing unavoidably involves presentation of the US. Therefore, it may not be the pairing of the two stimuli (CS and US) strengthening or weakening the associative link between the stimuli that is causing fluctuations in the SCR.

Instead, it could simply be exposure to the US driving this effect. If this were true, this would undermine the Perruchet effect as evidence for dual processing systems.

Research investigating US sensitization as an explanation of CR in the eyeblink and RT variants of the Perruchet effect has produced mixed results; with US sensitization failing to account for data from the eyeblink paradigm (e.g. Weidemann et al., 2009), but Mitchell et al. (2010) finding US sensitization a plausible explanation of the RT data despite Barrett and Livesey (2010) disagreeing. Given the inconsistency on this point, we felt that further investigation was important to determine whether US sensitization could account for the variations in the CR observed in the Perruchet effect. If it were found that US sensitization could adequately account for these results, a single, explicit, nonassociative processing explanation of the results would be sufficient.

In particular, it was important to try and determine whether US sensitization could explain the results reported in our 2012 paper. A lot of past research within the electrodermal domain finds a strong positive correlation between CR and conscious contingency knowledge, for example, if participants fail to develop CS-US contingency knowledge they often fail to show any CR (Lovibond \& Shanks, 2002). The implication is that to see a CR using electrodermal procedures, participants must explicitly expect the shock to happen when they are presented with a CS. This view is in stark contrast to our earlier findings. We hypothesized that we were able to dissociate the CR and expectancy of the US because of the nature of the Perruchet task. In our experiment there was a partial reinforcement schedule, half the trials were followed by a shock and half were not, and the participants were made explicitly aware of this from the beginning of the experiment. The participants were therefore put into a state of uncertainty from the start, as they were unable to accurately predict when the shocks were going to happen. Consequently, given that the participants were unable to use their rational, inferential processes to determine when the shocks were going to happen, this provides a context in which some reliance on alternative processing systems, which could be implicit or associative, might be expected. There is some evidence in the electrodermal domain to support this hypothesis. One example is Knight, Nguyen, and Bandettini (2003), who presented participants with tone stimuli, one continually reinforced by white noise (CS+) and another never paired with white noise (CS-). They varied US predictability by presenting the CSs above and below the perceptual threshold and found that even in the absence of any conscious ability to discriminate between the stimuli, there was still evidence of higher CRs to the CS+ than the CS-. Additionally, evidence of an implicit/explicit learning distinction can be found within the neuroimaging literature. Different brain structures appear to be involved in different aspects of learning to the extent that one can differentiate brain regions involved in conscious and unconscious learning (e.g. Tabbert, Stark, Kirsch, \& Vaitl, 2006).

Our aim here is to establish whether associative processes govern the CR in our experiments, by checking whether US sensitization can account for our findings. If we can rule out this explanation of our results, then we can add our experiment to the others that show that SCR and conscious expectancy can dissociate.

\section*{Experiment 1}

One of the simplest ways to investigate whether US sensitization governs CR in our Perruchet experiment was to simply remove the CSs. In this way participants would only experience noCS-US trials and noCS-noUS trials. If the same increasing patterns in SCR were found as in the original experiment, this would imply that the result was not dependent on CS-US pairings, as there are no CSs presented in the experiment. Under these circumstances we could conclude that US sensitization would be driving responding. However, if SCR fails to increase over successive US presentations, this would tend to suggest that a US sensitization account could not explain the electrodermal variant of the Perruchet effect.

\section*{Method}

\section*{Participants}

24 University of Exeter students participated in this experiment, 16 women and 8 men; ages ranging from 18-35 (average, 21 years). All were paid \(£ 10\).

\section*{Stimuli}

The US was a 500 ms electrical impulse administered with a PowerLab 26T generator using stainless steel electrodes attached to the left proximal and medial phalanges of the index finger. Participants set their own shock level between 5 and 20 mA where they deemed the shock to be "definitely uncomfortable but not painful".

Throughout the experiment there was a black cross (5 x 5 cm onscreen) in the centre of a white screen. The cross was used to fixate participants' attention.

Skin conductance was measured using LabChart software via MLT116F GSR electrodes attached to the medial phalanges on the left third and fourth fingers. Online explicit expectancy of the US was recorded using a Contour Shuttle Xpress device. Roughly every five seconds participants were required to make an expectancy rating about the extent they thought the shock would happen at that moment in time. The device had 5 buttons and fit nicely into participants' hand whereby 1 button corresponded to 1 finger. The different expectancy values were: 1 "There will definitely not be a shock", 2 "There might not be a shock", 3 "Not sure either way", 4 "There may be a shock", and 5 "There will definitely be a shock". A continuously available key explained which buttons represented which expectancy ratings.

\section*{Design}

There were two repeated-measures factors in this experiment. The first was run length, i.e. the number of trials of the same type in a row; there were six different run
lengths: \(-3,-2,-1,+1,+2\), and +3 . The run length measure is taken on the trial after the run itself. For example, \(a+1\) run length SCR is taken on the trial after a participant has previously experienced a US trial that itself was preceded by a noUS trial. A +2 measurement is taken on the trial following two consecutive US trials. Whereas, a -1 run length measurement is taken if the participant has just experienced exactly one noUS trial, and a - 2 measurement is taken when a participant experiences two noUS trials in a row. A switch between a positive and negative run length measurement occurs when the trial type just experienced switches from a US to noUS trial and vice versa. The other factor used in the design and analysis is the presence or absence of the US on the trials in the run, i.e. shock ( + ) or no shock (-).

The trial sequences used in this experiment were matched to the sequences used in the original McAndrew, Jones, McLaren and McLaren (2012) experiment, using the same trial structure and run distributions, see Table 1. In the McAndrew et al. experiment, on shock ( + ) trials, a 500 ms US was administered after 4500 ms of the CS being on screen, whereas on no shock (-) trials no US occurred. SCR recordings were taken on every trial, during the five seconds prior to CS onset (PreCS), five seconds while the CS was on screen and five seconds after the CS (PostCS). The intertrial interval (ITI) was randomly varied between 30 and 40 seconds on each trial in order to stop participants timing the onset of the CS. Long ITIs were required to allow the SCR recording to reach baseline after the previous US. This experiment, in keeping with the procedures used in the original experiment, was split into two blocks to allow recalibration to the shock to reduce habituation. Overall there were 46 trials, 23 per block. An extra trial was added at the end of each block, the \(23^{\text {rd }}\) trial, to take measurements from the last experimental trial. As there were no CSs in this experiment, on each trial, a hypothetical 5 second "CS" period was measured (where the CS would have occurred), and 5 seconds before this as a "PreCS" measure was taken. On shock trials, the shock would occur in the last 500 ms of the "CS" period.

Table 1. Trial types by frequency of occurrence.
\begin{tabular}{lcccccc}
\hline \multicolumn{1}{c}{ Variable } & & noUS (-) & & \multicolumn{2}{c}{ US (+) } \\
\hline Run length & 3 & 2 & 1 & 1 & 2 & 3 \\
Frequency & 2 & 4 & 8 & 8 & 4 & 2 \\
\hline
\end{tabular}

\section*{Procedure}

The participants were told they would receive shocks randomly throughout the experiment without any warning they were going to occur. The participants were asked to rate their expectancy that the shock would occur at that moment in time roughly every 5 seconds throughout the duration of the experiment. Expectancy ratings were made using the Shuttle Xpress device, on the scale 1 (definitely no shock) to 5 (definitely shock). Otherwise they were asked to remain still to avoid motion artifacts in the SCR.

\section*{Results}

The SCR data was recorded in micro-Siemens in LabChart and exported to Excel. For each trial the mean SCR was taken for the hypothetical "PreCS" and "CS" periods. The data was standardised using a log transformation to reduce the variability between participants. The change in SCR prior to US onset (as a consequence of preceding runs), was calculated using the formula "CS-PreCS". Mean CR for each run length was then calculated for each participant and across participants. For the expectancy data, the rating made closest to the hypothetical CS period was used as the participants' expectancy of the US on that trial. A mean expectancy rating for each run length was calculated for each participant and then across participants. Additionally, as in the 2012 experiment, the SCR and expectancy data were collapsed to form levels 1, 2 and 3. Level 1 averages run lengths +1 and -3 , level 2 run length +2 and -2 , level 3 run lengths +3 and 1. This was done to treat the data in a similar fashion to the 2012 experiment to enable direct comparisons to be made.

Two-way repeated measures analyses of variance (ANOVA) were carried out separately on the SCR and expectancy data. With regards to SCR, there was no significant linear trend over level, \(F(2,46)=0.26, M S E=\) \(0.004, p=.774\), see Fig. 1, a Bayesian analysis (Dienes, 2011) confirmed that we have strong evidence for the null hypothesis, rejecting US sensitization as an explanation of our effect, Bayes factor \(=0.32\). There was a significant effect of US presence, with a higher mean SCR after US absent trials \((-0.02)\) than US present trials \((-0.04\), see Fig. 2), \(F(1,23)=5.43, M S E=0.003, p=.029\). The interaction between level and US presence was not significant ( \(p>.05\) ).

There was a significantly increasing linear trend over level in the expectancy data, \(F(2,46)=5.04, M S E=0.257, p\) \(=.014\), see Fig. 1. There was also a significant effect of US presence with a higher mean expectancy rating for US absent trials (3.68) than US present trials (3.17, see Fig. 3), \(F(1,23)=18.41, M S E=0.496, p<.001\). In addition, there was a significant linear interaction between level and US presence, \(F(1,23)=14.77, M S E=0.213, p=.001\), reflecting the fact that measures taken after US present ( + ) trials increase as a function of level whereas those taken after US absent (-) trials slightly decrease.

\section*{Experiment 1 Results}


Figure 1. Graph depicting changes in SCR and expectancy as functions of level.

\section*{Discussion}

This experiment aimed to investigate the extent to which US sensitization is a plausible explanation of the SCR result observed in our original 2012 experiment. In that earlier experiment participants experienced exactly the same sequences of shocks as used here (NB. a different sequence for each participant), but the shocks followed a CS (a picture of a brown cylinder). This CS also occurred on noshock trials, so that it had a \(50 \%\) rate of reinforcement. McAndrew et al. (2012) found that autonomic SCR increased significantly with level, whilst explicit expectancy decreased significantly with level.

In the current experiment participants experienced runs of USs and noUSs in the absence of this CS. Hence, there was no associative structure to drive changes in SCR, and now SCR is essentially flat across level. This indicates that US sensitization is not occurring in this experiment, as successive shocks are not leading to an increase in SCR as would be expected if this were happening. This suggests that US sensitization is not responsible for the SCR pattern seen in the 2012 experiment, making the case that associative processes are responsible somewhat stronger.

Supporting this analysis, a significant effect of US presence was found in the SCR data, with higher SCRs observed on measurements taken after noUS (-) trials. Therefore, there were bigger changes in SCR just after nonreinforced runs (when shocks were not occurring; see Fig. 2). We conjecture that the SCR may be subject to habituation rather than sensitization using our procedures, because exposure to shocks appears to be causing smaller rather than larger SCR fluctuations. Alternatively, it could be that any learning to the temporal cues is being expressed more in the PreCS period than the CS period, explaining the negative difference scores.

\section*{Changes in SCR}


Figure 2. Graph depiciting SCR as a function of run length.
Regarding expectancy, the data is more complicated. A significantly increasing effect of level was found as well as a significant US presence effect such that expectancy of shock was higher after US absent (-) trials. These two findings at first seem paradoxical, with the latter suggesting participants gave higher expectancy ratings if there had not been any shocks, while the former implies the opposite (see Fig. 3). The increase in expectancy with level is entirely driven by the US present \((+)\) trials, and could simply reflect use of another heuristic, the "hot hand" effect (Burns \&

Corpus, 2004). Here we speculate that participants are simply tracking runs of shocks, and once they have had two in a row decide that the run is likely to continue. Regarding the US presence effect, participants gave higher ratings of shock after US absent trials where no USs occurred. This implies participants expected shocks more when there had not been one recently. We speculate that because participants knew that in this experiment, the only thing that would happen was that intermittent shocks would occur, as time elapsed ratings for a shock occurring would tend to increase as they knew that eventually a shock had to happen. We recorded the expectancy ratings participants made over each trial in 5 second bins (we had to exclude one bin as this varied from 5 to 15 seconds due to recording issues caused by a variable ITI). Supporting our speculation, expectancy significantly increases as time elapses between trials, \(F(1,23)=75.41, M S E=11.99, p<.001\), i.e. once a participant is shocked ratings of another shock occurring are lower just after this, but as time elapses the rating increases. In some sense this is a temporal equivalent of the gambler's fallacy heuristic, but this time it is an entirely rational reflection of the sequences experienced in this experiment.

By combining these two effects we can explain the pattern of results shown in Fig. 3. The gradual increase in expectancy with time since the last shock sets the overall trend, and tracking of runs of USs explains the increasing trend for positive runs superimposed on this overall effect.

This pattern of results conflicts with our 2012 expectancy finding that expectancy of shock decreased as a function of level. Comparing both experiments, we have two very distinct patterns of responding, which suggests we have fundamentally changed the paradigm and the demand characteristics from our original experiment, leading participants to approach the task differently. Given that both SCR and expectancy show an overall decline across run length in this experiment we could even claim that our data are consistent with the conscious expectancy-driven account of SCR often found in the literature on electrodermal conditioning. What this pattern of effects cannot do, however, is explain the quite different pattern found by McAndrew et al. (2012).

\section*{Expectancy rating}


Figure 3. Graph depiciting expectancy as a function of run length.

We hypothesise that the crucial factor causing the difference between the results of this experiment and the 2012 experiment is the absence of the CS. In this experiment participants only experience one stimulus (the US) as opposed to two interacting stimuli (the CS and the US). But can we be sure that sensitization is not a factor in our experiment? We see two possible issues that need to be investigated. The first stems from the possibility that any change in SCR consequent on experience of shock is being expressed in the PreCS period because of timing issues. If sensitization were occurring, but manifesting in the PreCS window, then this would have the effect of driving our SCR measure down. In the McAndrew et al. (2012) experiment the CS could be used to eliminate this timing issue, and so the sensitisation would now manifest in the CS period and drive the measure up. Weidemann et al. (2009) have also proposed that the expression of US sensitization is dependent on a weakly conditioned discrete cue being present. Given this, we cannot establish that US sensitization is not driving the SCR in our 2012 experiment as matters stand. This is addressed in Experiment 2.

\section*{Experiment 2}

This experiment uses a discrete cue to provide the correct context for the expression of US sensitization (Weidemann et al., 2009). Therefore, there are CS-US, CS-noUS, noCSUS and noCS-noUS trials in this experiment. The US sequences used were mapped to those in the previous experiments except that we added strategically placed CSs, one per run length per block. These CSs were placed on these specific runs to avoid the build up of associative strength, and alternated in terms of being reinforced or not. SCR was measured on these trials, and, due to the absence of any associative structure during the preceding run, if an increase in SCR across run length (and level) is found, US sensitization would explain this. However, if we fail to find an increasing pattern an alternative explanation for our 2012 results must be sought, perhaps an associative one.

\section*{Method}

\section*{Participants}

24 people participated in this experiment, all University of Exeter undergraduate students, 15 women and 9 men, ages ranging from 18 to 24 years old (average, 19 years). All participants were paid \(£ 10\).

\section*{Stimuli}

The same stimuli were used in this experiment as in the previous one. However, on the trials where a CS was presented, a brown cylinder ( \(19 \times 13 \mathrm{~cm}\) onscreen) appeared for 5 seconds (the same as used in McAndrew et al., 2012). Participants were asked to make explicit expectancy ratings just as they were in the first experiment, every 5 seconds.

\section*{Design}

The sequences were the same as those used in Experiment 1, however, a CS was added to 6 trials per block, one on each of the \(-3,-2,-1,+1,+2\) and +3 runs. There is only one
+3 and -3 run per block so these always had a CS. A CS was then randomly allocated to a \(+2,-2,+1\), and -1 run. Three additional trials were inserted at the start of each block, CSUS, CS-noUS, CS-US, in order to create the weakly conditioned discrete cue. Thus, overall there were 52 trials.

\section*{Procedure}

Participants were told that sometimes they would see a brown cylinder come on the screen. Half the time it would be followed by a shock and half the time not, but sometimes they would receive a shock when the cylinder was not there. Other procedures were as in the previous experiment.

\section*{Results}

The SCR data was treated in much the same way as in Experiment 1, with regards to data collection and log transformation. A measure of the CR was taken for each trial on which the CS was present using the formula, CSPreCS. A mean SCR measure was then recorded for each run length and averaged across participants. With regards to the expectancy data, the expectancy rating made during the actual CS period was taken as participants' expectancy of the US on that trial. Again, a mean rating for each run length was calculated for each participant and then across participants. Both data sets were collapsed to form the variable level (see Fig. 4).

\section*{Experiment 2 Results}


Figure 4. Graph depicting changes in SCR and expectancy as functions of level.

Two-way repeated measures ANOVAs were run on the SCR and expectancy data separately. In the SCR data, there was no significant linear trend over level, \(F(1,23)=1.82\), \(M S E=.023, p=.190\), Bayes factor \(=0.31\), so we have strong evidence for the null hypothesis leading us to reject US sensitization as an account for this result. Nor was there a significant difference between the US present and US absent trials or any interaction ( \(p>.05\) ). With regards to the expectancy data, there was a significantly increasing linear effect across level, \(F(1,23)=5.52, M S E=2.19, p=.028\). Additionally, there was a significant effect of US presence, with a higher mean expectancy rating for US absent (3.69) compared to US present (2.90) trials, \(F(1,23)=39.35, M S E\) \(=22.56\), \(p<.001\), see Fig 5. However there was no interaction between these two variables ( \(p>.05\) ).

\section*{Discussion}

Weidemann et al. (2009) proposed that in order to see the effects of US sensitization, the experimental context had to incorporate a discrete CS. Therefore, Experiment 2 aimed to do this to keep the context of the experiment similar to that of the original 2012 experiment. The CSs were strategically placed on one of each US run length, to measure UR without the build up of associative strength. In some sense our manipulation has been successful, as now the SCR changes recorded are all positive, as was the case in our 2012 experiment.

Analysis of the SCR data shows we have found another case where SCR is flat across run length and level. There is no sign of an increasing trend, which would be expected if presentation of the US is sensitizing participants. This null result strengthens the case against the nonassociative US sensitisation account as an explanation of the SCR result seen in the original Perruchet experiment. In the absence of any evidence to the contrary, we propose that associative processes are driving performance in the original 2012 paradigm, consistent with a dual processing system account of learning (McLaren, Green, \& Mackintosh, 1994).

\section*{Further Results}


Figure 5. Graph depicting changes in SCR and expectancy as functions of run length.

With regards to the expectancy data, despite our changes to the paradigm, we have obtained the same pattern as we found in Experiment 1. There is a significant linear increase in expectancy across level, yet an overall decrease in expectancy from US absent to US present trials. The explanation we proposed for Experiment 1 can also account for this result. Despite CSs being present in this experiment, participants are still instructed that shocks will happen intermittently throughout the experiment regardless of the presence/absence of the CS. Therefore, as time elapses participants give higher ratings for shock, which then decrease once a shock has occurred. Once again we speculate that they track runs of USs, and this effect is superimposed on the overall trend due to elapsed time.

\section*{Conclusion}

In two experiments we investigated whether nonassociative US sensitization could explain the original result found in McAndrew et al.'s (2012) experiment. In
both cases there was no evidence for sensitization to the US, and the pattern of results was different to that obtained in the Perruchet paradigm. We conclude that a dual processing system explanation (e.g. McLaren et al., 1994), appealing to explicit, propositional processes to explain the expectancy data as opposed to associative, autonomic accounts of the SCR data, is still the most convincing account of McAndrew et al.'s (2012) results.

\section*{ACKNOWLEDGEMENTS}

This research was funded by a studentship from the ESRC.

\section*{REFERENCES}

Balderston, N.L., \& Helmstetter, F.J. (2010). Conditioning with masked stimuli affects the timecourse of skin conductance responses. Behavioral Neuroscience, 124(4), 478-489.
Barrett, L.C., \& Livesey, E.J. (2010). Dissociations between expectancy and performance in simple two-choice reaction-time tasks: A test of associative and nonassociative explanations. Journal of Experimental Psychology: Learning, Memory, and Cognition, 36(4), 864-877.
Burns, B.D., \& Corpus, B. (2004). Randomness and inductions from streaks: "Gambler's fallacy" versus "Hot hand". Psychonomic Bulletin and Review, 11, 179-184.
Dienes, Z. (2011). Bayesian versus Orthodox statistics: Which side are you on? Perspectives on Psychological Sciences, 6(3), 274-290.
Knight, D.C., Nguyen, H.T., \& Bandettini, P.A. (2003). Expression of conditional fear with and without awareness. PNAS, 100(25), 1528015283.

Lovibond P.F., \& Shanks, D.R. (2002). The role of awareness in Pavlovian conditioning: Empirical evidence and theoretical implications. Journal of Experimetnal Psychology: Animal Behavior Processes, 28(1), 3-26.
McAndrew, A., Jones, F., McLaren, R., \& McLaren, I.P.L. (2012). Dissociating expectancy of shock and changes in skin conductance: An investigation of the Perruchet effect using an electrodermal paradigm. Journal of Experimental Psychology: Animal Behavior Processes, 38(2), 203-208.
McLaren, I.P.L., Forrest, C.L., and McLaren, R.P. (2012). Elemental representation and configural mappings: Combining elemental and configural theories of associative learning. Learning and Behavior, 40, 320-33.
McLaren, I.P.L., Green, R.E.A., \& Mackintosh, N.J. (1994). Animal learning and the implicit/explicit distinction. In N.C. Ellis (Ed.) Implicit and explicit learning of languages (pp. 313-332). New York, NY: Academic Press.
Mitchell, C.J., De Houwer, J., \& Lovibond, P.F. (2009). The propositional nature of human associative learning. Behavioral and Brain Sciences, 32, 183-246.
Mitchell, C.J., Wardle, S.G., Lovibond, P.F., Weidemann, G., \& Chang, B.P.I. (2010). Do reaction times in the Perruchet effect reflect variations in the strength of an associative link? Journal of Experimental Psychology: Learning, Memory and Cognition, 36(2), 567-572.
Perruchet, P. (1985). A pitfall for the expectancy theory of human eyelid conditioning. Pavlovian Journal of Biological Sciences, 20, 163-170.
Perruceht, P., Cleeremans, A., \& Destrebeceqz, A. (2006). Dissociating the effects of autonomic activation and explicit expectancy on reaction times in a simple associative learning task. Journal of Experimental Psychology: Learning, Memory, and Cognition, 32, 955-965.
Tabbert, K., Stark, R., Kirsch, P., \& Vaitl, D. (2006). Dissociation of neural responses and skin conductance reactions during fear conditioning with and without awareness of stimulus contingencies. Neuroimage, 32(2), 761.
Weidemann, G., Tangen, J.M., Lovibond, P.F., \& Mitchell, C.J. (2009). Is Perruchet's dissociation between eyeblink conditioned responding and outcome expectancy evidence for two learning systems? Journal of Experimental Psychology: Animal Behavior Processes, 35, 169-176.

\title{
Reasoning about diamonds, gravity and mental states: The cognitive costs of theory of mind
}

\author{
Ben Meijering \({ }^{1}\) (b.meijering@rug.nl), Hedderik van Rijn \({ }^{2}\), Niels A. Taatgen \({ }^{1}\), and Rineke Verbrugge \({ }^{1}\) \\ \({ }^{1}\) Institute of Artificial Intelligence, PO Box 407 \\ 9700 AK Groningen, The Netherlands \\ \({ }^{2}\) Department of Psychology, Grote Kruisstraat 2/1 \\ NL-9712 TS Groningen
}

\begin{abstract}
Theory of mind (ToM) is required when reasoning about mental states such as knowledge, beliefs, desires, and intentions. Many complex reasoning tasks require domaingeneral cognitive resources such as planning, resistance to interference, and working memory. In this paper we present a study of the additional cognitive costs of reasoning about mental states. We presented participants with sequential games in which they have to reason about another player. In the so-called player condition, the other player is reasoning about the participant, whereas in the so-called balance condition, the other player is reasoning about a balance scale. Both types of games require the same comparisons, but only differ in the required depth of ToM reasoning. Games in the player condition require one additional switch between perspectives. The results show that participants make different types of mistakes in the player condition as compared to the balance condition. This finding implies a different reasoning process when reasoning about mental states. The results also show faster decreasing reaction times in the balance condition than in the player condition. Based on these findings, we argue that reasoning about mental states requires unique cognitive resources.
\end{abstract}

Keywords: Theory of mind; perspective taking; decision making; sequential games; social cognition.

\section*{Introduction}

In many social interactions we reason about one another. If, for example, our decisions or outcomes depend on someone else's actions, we try to predict what the other will do. Predicting the other's actions requires an understanding of how behaviors are caused by mental states such as beliefs, desires, goals, et cetera. Such an understanding is often referred to as theory of mind (Premack \& Woodruff, 1978).

A theory of mind, or ToM, is starting to develop around the age of three to four years (Wellman, Cross, \& Watson, 2001; Wimmer \& Perner, 1983). However, younger infants already are susceptible to others' mental states (Onishi \& Baillargeon, 2005). One possible explanation is that they are able to read others' behavior, but cannot yet explicitly reason about the underlying mental states. Only after many interactions, reading many distinct behaviors, do children start to develop a theory of how behaviors generally correspond with beliefs, desires, intentions, et cetera.

So far, we have introduced ToM as being a theory (Gopnik \& Wellman, 1992). However, we do not want to exclude another definition of ToM that considers it to be an ability or skill to reason about mental states of oneself and
others (Apperly, 2011; Leslie, Friedman, \& German, 2004; Van Rij, Van Rijn, \& Hendriks, 2010; Wimmer \& Perner, 1983). In fact, a theory alone would not suffice when reasoning about others' mental states. Such reasoning is an entire process of generating many possible mental state interpretations (Baker, Saxe, \& Tenenbaum, 2009), and ToM reasoning might be qualitatively different from other kinds of reasoning.

Some studies have shown similar but uncorrelated developmental trends in ToM tasks and non-mental tasks that require similar representations (Arslan, Hohenberger, \& Verbrugge, 2012; Flobbe, Verbrugge, Hendriks, \& Krämer, 2008). For example, a relative clause in the sentence "The goat that pushes the cat" requires a similar representation as the complement clause in "Alice knows that Bob is writing", but only the complement clause requires a mental state representation. As children become older, they get better at understanding both types of sentences. However, their performance does not correlate when the factor age is controlled for. These findings show that ToM tasks might consume unique cognitive resources. It is important to note, however, that these tasks might have differed with respect to other factors, besides the aspect of mental representations.

Some studies show similar performance in ToM tasks, on the one hand, and equivalent but non-mental control tasks, on the other. In the false-belief or Sally-Anne task, for example, children have to attribute a false belief about an object's current location to Sally (Wellman et al., 2001; Wimmer \& Perner, 1983). Sally stores an object at location A, but the object is moved from location A to location B while Sally is away. Therefore, Sally still thinks that the object is at location A. To pass this task, children should acknowledge that Sally falsely believes that the object is still at location A. The false-sign task is a similar but nonmental counterpart of the false-belief task. An object is first stored at location A, indicated by an arrow. Next, the object is moved from location A to location B, but the arrow still points at location A. The false sign in this task is the arrow pointing at location A, which is similar to Sally's false belief. Children's accuracy in both tasks is similar, and their performance correlates, even after correcting for age (Apperly, 2011; Leekam, Perner, Healey, \& Sewell, 2008; Sabbagh, Xu, Carlson, Moses, \& Kang, 2006). This finding implies that mental state reasoning might not qualitatively differ from other kinds of reasoning.


Figure 1. Examples of two-player Marble Drop games. A white marble is about to drop, and its path can be manipulated by turning the orange and blue trapdoors. In these example games, participants have to obtain as many orange diamonds as possible and they control the orange trapdoors. The other player has to obtain as many blue diamonds as possible and controls the blue trapdoor. In game \(a\), the optimal decision for a participant is to let the white marble drop into the topmost bin, thereby obtaining 3 orange marbles. The 4 orange diamonds in the bottom-left bin are not obtainable, as the other (blue) player's optimal decision is to let the white marble drop into the middle bin: The other player knows that the optimal (orange) decision at the bottom trapdoors is to go left, yielding a suboptimal outcome of 1 blue diamond for Player 2. Games \(a\) and \(c\) are second-order games, because participants (as Player 1) have to reason about the other player (i.e., Player 2) who in turn has to reason about Player 1. The games in \(b\) and \(d\) are first-order counterparts of the games in \(a\) and c, respectively. They require the same comparisons, as the outcome of the balance scale is congruent with Player 1's last correct / rational decision: Both depend only on Player 1's diamonds in the bottom two bins. However, the games with the balance require one fewer switch between Player 1 and Player 2 perspectives.

Similar accuracy of responses in ToM tasks and their nonmental counterparts, however, does not necessarily imply a similar reasoning process. Moreover, differences might manifest themselves elsewhere, for example, in the reaction times. If, for example, both tasks require overlapping cognitive functions but ToM tasks require additional cognitive processing, the response patterns might not differ as much as the associated response times. Moreover, differences in accuracy might not manifest themselves until the tasks become more complex and exhaust cognitive resources.

Given these mixed findings, the question remains whether reasoning about mental states requires additional cognitive resources. Complex reasoning tasks consume cognitive resources, because oftentimes they require integration of information in the overall reasoning process. Integration of information and reasoning require executive functions such as planning, set shifting, resistance to interference, and working memory. It is not yet obvious why these executive functions alone would not suffice to reason about mental states.

In this study we investigate whether reasoning about mental states consumes unique cognitive resources. Participants are presented so-called Marble Drop games (Figure 1) in which they have to reason about another player. Marble Drop games have a recursive structure because the best possible, or optimal, decision at the first trapdoor depends on the other player's decision at the
second trapdoor, which in turn depends on the outcome at the third trapdoor (Meijering, Van Rijn, Taatgen, \& Verbrugge, 2011). The crucial factor in this experiment is whether the outcome at the third trapdoor is determined by Player 1's decision (player condition) or by the physics of a balance scale (balance condition). Both conditions require the same comparisons, but games in the player condition require one additional switch between player perspectives: Player 1 has to reason about what Player 2 thinks that Player 1 will do at the final trapdoor. If reasoning about mental states requires additional cognitive resources, games in the player condition would be more difficult than games in the balance condition.

\section*{Method}

Participants are always assigned to the role of Player 1, and in both conditions they need to take the perspective of Player 2 to predict the outcome at the second trapdoor. This perspective taking requires ToM. As explained previously, the decision at the second trapdoor depends on the outcome at the third trapdoor. If the participants (i.e., Player 1) control that trapdoor, they need to switch perspective again. They need to re-take their own perspective from within Player 2's perspective. This requires second-order ToM. In the balance scale condition, participants do not have to switch perspective again, and thus need first-order ToM at most. They still need to make the same comparisons, as the outcome of the balance scale depends on Player 1's payoffs
and this outcome is congruent with Player 1's goal to maximize his or her payoffs.

If ToM requires unique cognitive resources, we expect that participants respond faster in the balance condition than in the player condition, because the balance condition requires one switch less between Player 1 and Player 2 perspectives than the player condition. We also expect better performance in the balance condition, because Marble Drop games in which Player 1 controls the third trapdoor might appear to be less deterministic. The assumption, here, is that it is easier to attribute knowledge of physics to Player 2 than to attribute to Player 2 epistemic reasoning about Player 1, as epistemic reasoning involves testing of multiple possible Player 2 perspectives.

\section*{Participants}

Forty-two first-year Psychology students (30 female) participated in exchange for course credit. The average age was 21 years, ranging from 18 to 25 . Each participant reported normal or corrected-to-normal visual acuity.

\section*{Stimuli}

Of all possible payoff structures, only those that are diagnostic of second-order ToM reasoning were included in the experiment. A game is diagnostic of second-order ToM reasoning if it requires a participant to reason about each decision point to arrive at the optimal decision. An example of a non-diagnostic payoff structure is one in which Player 1's first payoff, in the topmost bin, is the maximum payoff in that game. In that case, Player 1 would not need to reason about the second and third decision points. The payoff structures are listed in a table, which can be found at http://www.ai.rug.nl/~meijering/marble_drop.html.

\section*{Design}

The experimental design consists of two between-subjects conditions: balance condition versus player condition. In the player condition, participants are presented with the original second-order ToM games (Meijering, Van Rijn, Taatgen, \& Verbrugge, 2012). In the balance condition, participants play the games with the same payoff structures, but the third decision point is replaced by a balance scale. Importantly, the games in both conditions are equivalent, as they require the same comparisons between payoffs. In each game, the outcome of the balance is the same as the last correct / rational decision of Player 1, because both only depend on the number of Player 1 diamonds in the bottom two bins (see Figure 1).

\section*{Procedure}

After giving informed consent, participants were seated in front of a 24 -inch iMac. They were randomly assigned to the balance scale condition or the player condition. The participants were instructed that their goal was to obtain as many diamonds as possible of their target color, either blue or orange, which was counterbalanced between participants. They were also instructed that Player 2's (i.e., the
computer's) goal was to obtain as many marbles as possible of the other color.

The experimental procedure is the same in both ToM conditions. Participants are presented 62 unique games. At the start of each game, participants have to decide whether to stop the game, by letting the white marble drop into the top bin, or to continue the game, by letting the white marble drop onto Player 2's trapdoor. The game stops if Player 2 decides to let the white marble drop into the middle bin. If Player 2 decides to let the white marble drop onto the third trapdoor, participants in the player condition have to decide whether to stop the game in the bottom-left or bottom-right bin. In the balance condition, the physics of the balance scale determine whether the marble drops into the bottomleft bin or the bottom-right bin. Importantly, the balance scale is set in motion as soon as the white marble drops onto it. Otherwise, Player 2 would not have to reason about the balance scale. Each game is fully animated. See Figure 1 for some example games.

After each game, participants receive feedback that mentions Player 1's outcome. If, for example, the marble drops into a bin that contains two diamonds for Player 1, the feedback mentions: "You get 2 ".

To familiarize participants with the rules of Marble Drop games, participants are presented additional feedback during the first 12 games. Feedback explicitly mentions whether the outcome is the highest attainable Player 1 payoff. In case a participant obtains 3 diamonds and could not have obtained more, feedback is: "Correct. You get 3. The highest possible payoff!". In case a participant obtains 3 diamonds, but could have obtained 4, feedback is: "Incorrect. You get 3. You could have obtained 4".

\section*{Results \& Discussion}

The data consist of 62 unique Marble Drop games (i.e., payoff structures) for each participant. In the statistical analyses, the games are blocked to accommodate non-linear and differential learning rates: The first 12 'training' games comprise the first block, and the remaining 50 games are split into 5 subsequent blocks of 10 games each. The graphs show means and standard errors, which are represented by error bars.

The data are analyzed by means of linear mixed-effects models (Baayen, 2008; Baayen, Davidson, \& Bates, 2008) to accommodate random sources of variation due to sampling of participants and items (i.e., payoff structures). Specifically, each model allows for by-participant and byitem adjustments of the intercept. For each analysis that we report below, we first constructed a full factorial model with all main and interaction effects. Based on likelihood ratio comparisons, we removed main and interaction effects for as long as the corresponding parameters were not justified. If a comparison preferred a simplified model, we report the log-likelihood statistics. The correctness of responses is analyzed by means of logistic linear mixed-effects models, as correctness of responses is a binary variable (incorrect vs. correct).

\section*{Mean Proportion Correct}

The proportion of correct responses in each block is averaged across participants and depicted in Figure 2. The figure does not show great differences between performance in the balance scale and player conditions.

A full-factorial model with main effects and an interaction effect of Condition and Block did not fit the data better than an additive model, \(\chi^{2}(5)=6.08, n s\). The parameters of the additive model are discussed below.

There is a significant effect of Block, \(\beta=1.37, z=10.89\), \(p<.001\). As can be seen in Figure 2, performance increases over the course of playing many Marble Drop games.


Figure 2: Mean proportion of correct responses per block; depicted separately for participants in the balance condition (light gray) and the player condition (dark gray).

There is no effect of Condition, as can be seen in Figure 2. In contrast to our hypothesis, the probability of making a correct decision does not differ between the balance scale and player conditions. An analysis of the types of errors (next section: Types of Errors), however, shows differential errors between the balance scale and player conditions.

\section*{Types of Errors}

The errors that participants made were categorized according to game type, as an overall analysis might not be sensitive enough to differentiate between the balance scale and player conditions. Two types of games were distinguished on the basis of Player 2's (programmed) decision, which is either stop the game or continue.

There is no main effect of Player 2 decision, \(\beta=-.08, z=\) \(-.575, n s\), which means that the difficulty of a game does not depend on Player 2's decision. This finding implies that there is no reason to believe that there are particular subsets of hard(er) payoff structures among the selected payoff structures.

There is a significant interaction effect between the factors Condition and Player 2 response (see Figure 3), \(\beta=\) \(.65, z=3.349\), and \(p<0.001\). In the balance scale condition, the probability of making a correct decision does not differ between games in which Player 2's decision is to stop, on the one hand, and games in which Player 2's decision is to continue, on the other hand. In the player condition, in contrast, there is a difference. One possible explanation is that participants in the player condition expect Player 2 to continue in most games, and this expectation pays off in games in which Player 2 actually decides to continue. In each game, Player 2 has a greater payoff in one of the last two end states than in the earlier end state, and participants might assign too great a probability to Player 2 going for that payoff. Participants in the balance condition, in contrast, might estimate those probabilities more accurately (i.e., lower), because games with a balance scale can be considered more deterministic.


Player 2 decision
Figure 3: Mean proportion of correct responses across participants, depicted separately for the balance scale and player conditions, and Player 2's decision.

\section*{Reaction Times}

There are differences in the types of errors between participants in the balance scale and player conditions, but what about the reaction time data? RTs are analyzed to find out whether a switch between perspectives comes with a time-cost. The RTs are log-transformed as reaction times are skewed to the right. Figure 4 shows the average log-RT across participants.
Figure 4 shows differential learning rates between participants in the balance scale and player conditions, especially in the first half of the experiment, in blocks 1 to 3 . In the second half, blocks 4 to 6 , the learning rates do not seem to differ that much. To specifically accommodate for differential learning rates, the factor Block was reparameterized as a new factor Half, with levels 1 and 2 , and
a new factor Block with levels 1, 2, and 3 within each level of Half. The results of the full factorial LME with main and interaction effects of Condition, Half, and Block are discussed below.

The main effects of Half and Block (with linear contrast) are significant, \(\beta=-.22, t=-7.82, p<.001\), and \(\beta=-.18, t\) \(=-5.37, p<.001\), respectively. From the first to the second half of the experiment, and within each half, the RTs decrease linearly. The interaction between Half and Block is also significant, \(\beta=.15, t=3.19, p=.0015\). The decrease in RTs is stronger in the first half of the experiment than in the second half.


Figure 4: Average log-RT across participants plotted against block, separately for the balance scale and player conditions.

The interaction between Condition and Block is significant, \(\beta=.17, t=-3.43, p<.001\). The decrease in RTs in the first half of the experiment is less strong in player condition than in balance scale condition. This finding is partly congruent with the hypothesis that RTs are shortest in the balance scale condition because it requires fewer switches between perspectives than the player condition. There is, however, no main effect of Condition, \(\beta=.14, t=\) \(1.2, n s\). Thus, on average, the RTs do not differ between the balance scale condition and the player condition. However, participants in the balance scale condition do become faster towards the end of the first half of the experiment, whereas participants in the player condition do not become faster. A possible explanation is that participants in the balance scale condition are quicker over the course of playing multiple games to attribute an understanding of gravity to Player 2. In contrast, participants in the player condition need to play more games and test multiple Player 2 perspectives.

The interaction between Condition, Half, and Block is also significant, \(\beta=-.14, t=-2.83, p<.005\). As can be seen in Figure 4, the differential learning rates in the first half of
the experiment disappear in the second half of the experiment, where the RT trends do not differ that much between the balance scale condition and the player condition.

In sum, there is an interaction effect of Condition and Block on the RTs, and this effect is mainly present in the first half of the experiment. There, the RTs decrease more in the balance scale condition than in the player condition. This interaction effect, between Condition and Block, seems to disappear in the second half of the experiment. A possible explanation for the latter finding is that, initially, participants in the balance scale condition settle more quickly on the correct Player 2 perspective than participants in the player condition, who test multiple Player 2 perspectives across multiple games.

\section*{General Conclusions}

In this study we investigated whether ToM requires additional cognitive resources. We presented two types of games that required the same comparisons but differed with respect to the required depth of ToM reasoning: Games in the player condition required second-order ToM , as participants had to reason about a Player 2 that, in turn, reasoned about them; Games in the balance scale condition required first-order ToM, as participants had to reason about a Player 2 that reasoned about a balance scale. Our results show different errors between these conditions, which implies that the reasoning was not the same in the balance scale and player conditions. Moreover, the reaction time trends differed. The learning rate was faster for participants in the balance scale condition than for participants in the player condition. A faster learning rate in the balance condition is congruent with our hypothesis that it is easier to play against a Player 2 that reasons about gravity than playing against a Player 2 that reasons about mental states.

We assumed that games with a balance scale are easier to play because they appear to be more deterministic than games in which Player 1 has the last decision. This assumption is congruent with the RT data: Longer RTs in the player condition could be the cause of participants' testing of multiple possible Player 2 perspectives. Games in the balance scale condition, in contrast, require testing of fewer possible Player 2 perspectives, yielding shorter decision times.

Besides a faster learning rate in the balance condition, we expected a greater proportion of correct decisions. However, the probability of making a correct decision does not differ between the balance scale (i.e., first-order ToM) condition and the player (i.e., second-order ToM) condition. One possible explanation is that knowledge about gravity is not automatically attributed to Player 2. We expected that participants in the balance condition would automatically 'see' how Player 2's decision depends on the outcome of the balance, as young children have already mastered many balance scale configurations (Van Rijn, Van Someren, \& Van der Maas, 2003). However, attributing an
understanding of gravity to Player 2 might be less of an automatic process than reasoning about gravity oneself.

Based on our findings, we conclude that participants do need ToM in Marble Drop games. Sequential games such as Marble Drop can be critiqued for not requiring ToM: If Player 2's strategy is known, the optimal (Player 1) decision can be determined without reasoning about Player 2's reasoning about Player 1's last possible decision. Applying backward induction, an algorithm based on sequential payoff comparisons, would yield the optimal decision. However, Meijering et al.'s (Meijering et al., 2012) eye tracking study shows that participants use more complicated and diverse reasoning strategies, not only backward induction. Moreover, backward induction would not be able to account for different types of mistakes and differential reaction times in the two conditions, as backward induction always works the same, irrespective of condition. Our findings provide strong support for the idea that sequential games are not just a decision-making problem but also evoke reasoning about mental states and thus require ToM.

In fact, it seems that sequential games are a particularly good paradigm to test reasoning about mental states, as they require active application of ToM. If Player 2's strategy is not yet known, participants need to actively find the correct Player 2 perspective. In any given game, multiple Player 2 perspectives might apply, but only that of a rational Player 2 is consistent with Player 2's actual decisions across all games. Active application of ToM is required to test multiple perspectives and find that of a rational Player 2.

To conclude, our findings are congruent with findings from fMRI studies showing that mental state reasoning employs brain regions that differ from the regions involved in cognitive control (Apperly, 2011; Saxe, Schulz, \& Jiang, 2006). Our findings suggest that perspective taking requires additional cognitive resources, as opposed to just greater cognitive control, as one additional switch between perspectives induces not only longer reaction times but also qualitatively different decisions.

\section*{References}

Apperly, I. A. (2011). Mindreaders: The cognitive basis of "theory of mind." Hove, UK: Psychology Press.
Arslan, B., Hohenberger, A., \& Verbrugge, R. (2012). The development of second-order social cognition and its relation with complex language understanding and memory. In N. Miyake, D. Peebles, \& R. P. Cooper (Eds.), Proceedings of the 34th Annual Conference of the Cognitive Science Society (pp. 1290-1295). Austin, TX: Cognitive Science Society.
Baayen, R. H. (2008). Analyzing Linguistic Data: A Practical Introduction to Statistics Using R. New York, USA: Cambridge University Press.
Baayen, R. H., Davidson, D. J., \& Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language, 59(4), 390-412.
Baker, C., Saxe, R., \& Tenenbaum, J. B. (2009). Action
understanding as inverse planning. Cognition, 113(3), 329-349.
Flobbe, L., Verbrugge, R., Hendriks, P., \& Krämer, I. (2008). Children's application of theory of mind in reasoning and language. Journal of Logic, Language and Information, 17(4), 417-442.
Gopnik, A., \& Wellman, H. M. (1992). Why the child's theory of mind really is a theory. Mind and Language, 7(1-2), 145-171.
Leekam, S., Perner, J., Healey, L., \& Sewell, C. (2008). False signs and the non-specificity of theory of mind: Evidence that preschoolers have general difficulties in understanding representations. British Journal of Developmental Psychology, 26(4), 485-497.
Leslie, A. M., Friedman, O., \& German, T. P. (2004). Core mechanisms in "theory of mind." Trends in Cognitive Sciences, 8(12), 528-533.
Meijering, B., Van Rijn, H., Taatgen, N. A., \& Verbrugge, R. (2012). What eye movements can tell about theory of mind in a strategic game. PloS One, 7(9). doi:10.1371/journal.pone. 0045961
Meijering, B., Van Rijn, H., Taatgen, N., \& Verbrugge, R. (2011). I do know what you think I think: Second-order theory of mind in strategic games is not that difficult. In L. Carslon, C. Hoelscher, \& T. F. Shipley (Eds.), Proceedings of the 33rd Annual Conference of the Cognitive Science Society (pp. 2486-2491). Austin, TX: Cognitive Science Society.
Onishi, K. H., \& Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? Science, 308(5719), 255-258.
Premack, D., \& Woodruff, G. (1978). Does the chimpanzee have a theory of mind? Behavioral and Brain Sciences, l(4), 515-526.
Sabbagh, M. A., Xu, F., Carlson, S. M., Moses, L. J., \& Kang, L. (2006). The development of executive functioning and theory of mind: A comparison of Chinese and U.S. preschoolers. Psychological Science, 17(1), 74-81.
Saxe, R., Schulz, L. E., \& Jiang, Y. V. (2006). Reading minds versus following rules: Dissociating theory of mind and executive control in the brain. Social Neuroscience, 1(3-4), 284-298.
Van Rij, J., Van Rijn, H., \& Hendriks, P. (2010). Cognitive architectures and language acquisition: A case study in pronoun comprehension. Journal of Child Language. Journal of Child Language, 37(3), 731-766.
Van Rijn, H., Van Someren, M., \& Van der Maas, H. L. J. (2003). Modeling developmental transitions on the balance scale task. Cognitive Science, 27(2), 227-257.
Wellman, H. M., Cross, D., \& Watson, J. (2001). Metaanalysis of theory-of-mind development: The truth about false belief. Child Development, 72(3), 655-684.
Wimmer, H., \& Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition, 13(1), 103-128.

\title{
Modeling the Development of Determiner Productivity in Children's Early Speech
}

\author{
Stephan Meylan \\ smeylan@berkeley.edu \\ Department of Psychology \\ University of California, Berkeley
}

\author{
Michael C. Frank \\ mcfrank@stanford.edu \\ Department of Psychology \\ Stanford University
}

\author{
Roger Levy \\ rlevy@ucsd.edu \\ Department of Linguistics \\ University of California, San Diego
}

\begin{abstract}
The English definite and indefinite articles (also known as \(d e\) terminers) are a useful index of early morphosyntactic productivity in children's speech, and give evidence about children's representation of syntactic abstractions. Previous work (i.e. Pine \& Lieven, 1997) used a measure of productivity that shows a strong sensitivity to sample size and does not account for the relationship between adult input and children's learning. In this paper, we develop a more robust metric by employing a hierarchical Bayesian model to characterize the degree of generalization implicit in observed determiner usage. By inferring parameters for a generative model over longitudinal corpora, we measure the trajectory of grammatical category abstraction. Our results are consistent with the hypothesis that child learners exhibit adult-like patterns of generalization quite early in the acquisition of determiners.
\end{abstract}

Keywords: grammatical productivity; development; syntax; morphosyntax; modeling

\section*{Introduction}

How do children begin to use the rich combinatorial structure of language to express novel thoughts? Nativist accounts propose an innate specification of syntactic categories that allow the child learner to exploit regularities in language structure from birth (Valian, 1986). Constructivist theories, on the other hand, contend that abstract categorical knowledge is built up over time as the child learner generalizes from specific usages to form broader combinatorial rules (Tomasello, 2003). The indefinite determiner "a" and the definite determiner "the"-the shortest and most frequent words in the English language-are a locus of interest for both theoretical viewpoints. Because they are both frequent and obligatory, determiners are an early index of morphosyntactic \({ }^{1}\) productivity that can be observed cross-linguistically.

A context-free grammar production rule (Figure 1) captures the intuition that a noun phrase can be created by choosing a determiner from the abstract DET category and a noun from the abstract N category. This noun phrase in turn combines with other phrase structures, like a verb phrase or preposition, to form higher-order structures. For most singular nouns, a grammatical NP can be formed using either determiner. Furthermore, hearing a novel word with one determiner suggests that use with the the other is also likely grammatical. Hearing someone introduce "a blickmoo" for the first time, you would not hesitate to request "the blickmoo" even if you had never heard that sequence of words before.

\footnotetext{
\({ }^{1}\) In English, determiners and nouns are separate words by linguistic criteria (e.g. an adjective may come between a determiner and a noun). Many other languages use determiners that are morphologically integrated with the noun (see Kramsky, 1972 for an overview).
}


Figure 1: Many noun phrases can be created by combining a word from the abstract categories determiner and noun \((\mathrm{NP} \rightarrow \mathrm{DET}+\mathrm{N})\). It is an open question whether children's early representations are organized around these abstractions.

When do children share that same judgment? Valian (1986) showed that children between \(2 ; 0\) and \(2 ; 6\) demonstrate a variety of productive syntactic categories, including determiners. Using a distributional analysis of children's speech, she found that determiners were used in a fashion consistent with an adult-like grammar. Determiners were never used as the sole content of an utterance, never appeared in a sentencefinal position, and were always sequenced correctly with respect to adjectives and nouns in noun phrases.

Pine \& Lieven (1997) challenged Valian's assertion of adult-like grammatical productivity in children's speech by citing an apparent limit to productivity in determiner use. As a quantitative metric, Pine and Lieven presented the overlap measure for determiners: the number of nouns used with both determiners (in some sample), divided by the number of nouns used with either (in the same sample). For 11 children from \(1 ; 0\) to \(3 ; 0\) this proportion ranged from 0 to .23 , which Pine and Lieven interpreted as being extremely low for a speaker with productive determiner syntax. Rather than making full use of the combinatorial productivity of nouns and determiners, on this metric children thus seemed to be very conservative in their productions and to show a strong tendency to use nouns with only a single determiner. Pine and Lieven interpreted this finding as supporting item-based theories of learning in which there is only gradual generalization from individual instances to abstractions like DET.

Valian et al. (2009) objected that Pine \& Lieven (1997) failed to take noun frequency into account in considering determiner use. Because the overlap measure is necessarily 0 for all nouns that appear only once, Valian and colleagues argued that the overlap measure, especially when calculated over small datasets, under-represents productivity. Highly frequent nouns were much more likely to be used with both determiners: more than \(80 \%\) of nouns used at least 6 times were used with both "a" and "the." Our own analyses of the Providence corpus (Demuth \& McCullough, 2009) confirm


Figure 2: Determiner overlap (proportion of nouns used with both "a" and "the") increases as a function of the number of tokens in a speech sample, as seen from CHILDES files for 4 children from the Providence corpus. Dashed and solid lines show loess smoothers for the child and parent respectively.
this issue: overlap is deeply confounded with sample size. Sample size is the best predictor of both child and parent overlap, regardless of age (Figure 2).

Yang (2010) supplemented this argument by showing that-regardless of sample size-the overlap measure is necessarily low because of the Zipfian distribution of noun frequencies. The Zipfian frequency distribution of nouns results in a long tail of words seen only once, so if overlap is calculated as the proportion of nouns seen with both determiners, it will necessarily be low. Yang additionally observed that nouns vary in their determiner preference (e.g., "the bathroom" is more frequent than "a bathroom", but "a bath" is more frequent than "the bath"), unlike the simplest probabilistic instantiation of a productive context-free rule scheme as in Figure 1, where the probabilities of Determiner \(\rightarrow\) "the" and Determiner \(\rightarrow\) "a" would be independent of the noun's identity (Booth, 1969).

But while the overlap statistic is flawed, there is currently no replacement that directly measures the productivity of children's determiner use. Hence, in the current study, we develop a novel method for quantifying determiner productivity. We use a hierarchical Bayesian model to estimate adults' and children's determiner productivity (metric model) and then develop a variant that estimates the linkage between adult input and child generalization (linking model). In each model, one key parameter can be interpreted as a graded metric of productivity robust to variation in sample size and noun frequency distribution. Bayesian inference gives us the posterior distribution of this parameter given child and adult caregiver production data, allowing us to quantify determiner productivity and examine its developmental timecourse.


Figure 3: Interpretation of the \(v\) parameter, a concise metric of grammatical productivity. At low values of \(v\), little or no information is shared between nouns. At higher \(v\) values, nouns exhibit more consistent usage as a class, indicating the existence of a productive \(\mathrm{DET}+\mathrm{N}\) rule.

\section*{Metric Model}

We model the use of each determiner with a noun as a draw from a binomial distribution (a single weighted coin flip). The use of "the" is heads, and the use of "a," tails. The idiosyncratic determiner preference for each noun can thus be thought of as a coin's weighting, ranging from zero (a noun used only with "a") to one (a noun used only with "the"). We model variability in noun-specific determiner preferences by assuming some distribution underlying these preferences; specifically, we assume that each noun's preference is drawn from a beta distribution with mean \(\mu_{0}\) (the underlying "average" preference across all nouns) and scale \(v\), giving us a hierarchical beta-binomial model (Gelman et al., 2004). \({ }^{2}\)

The scale parameter \(v\) in our model plays a central role in quantifying cross-noun variability and thus gives us a continuous space in which to quantify learner productivity (Figure 3). At one end of the range, when \(v=0\), we have an extreme "island" learner for whom every noun is produced with only one determiner or the other. At the other end of the spectrum, as \(v\) approaches infinity, we have an extreme over-generalizer who has identical determiner preference for all nouns. The \(v\) parameter thus establishes a continuum on which we can place constructivist and nativist hypotheses.

By estimating values of \(\mu\) and \(v\) for individual children over the course of their development, we can examine how these parameters change, potentially reflecting developmental changes in productivity. Here we use the metric model to compare mother and child productivity for the six children in the Providence corpus (Demuth \& McCullough, 2009).

\section*{Model Details}

A full graphical model representation of the linking model is shown on the left side of Figure 4 . We assume that data \(d\)

\footnotetext{
\({ }^{2}\) Many readers may be more familiar with the more common parameterization of the beta distribution in terms of shape parameters \(\alpha=\mu \nu\) and \(\beta=(1-\mu) \nu\).
}

\(\nu\) : Degree of dispersion in determiner preference
\(\nu\) : Degree of dispersion in determiner preference
\(\alpha\) : Uninformative priors over \(\nu\) and \(\mu_{0}\)
\(\alpha\) : Uninformative priors over \(\nu\) and \(\mu_{0}\)
\(\eta\) : Noise parameter downweighting observed data
\(\eta\) : Noise parameter downweighting observed data
\(\alpha_{\eta}\) : Uniformative prior over \(\eta\)
\(\alpha_{\eta}\) : Uniformative prior over \(\eta\)
\(\mu_{i}\) : Idiosyncratic determiner preferences for each of M noun types
\(\mu_{i}\) : Idiosyncratic determiner preferences for each of M noun types
\(d_{i}\) : Observed determiner-noun pairs (N pairs)
\(d_{i}\) : Observed determiner-noun pairs (N pairs)
Metric Model
Metric Model
\(\mu_{i} \sim \operatorname{Beta}\left(\mu_{0} \nu,\left(1-\mu_{0}\right) \nu\right)\)
\(\mu_{i} \sim \operatorname{Beta}\left(\mu_{0} \nu,\left(1-\mu_{0}\right) \nu\right)\)
\(d_{i} \sim \operatorname{Bern}\left(\mu_{i}\right)\)
\(d_{i} \sim \operatorname{Bern}\left(\mu_{i}\right)\)
Linking Model
Linking Model
\(\mu_{i} \sim \operatorname{Beta}\left(\mu_{0} \nu+\eta r_{i}^{A},\left(1-\mu_{0}\right) \nu+\eta\left(N_{i}^{A}-r_{i}^{A}\right)\right)\)
\(\mu_{i} \sim \operatorname{Beta}\left(\mu_{0} \nu+\eta r_{i}^{A},\left(1-\mu_{0}\right) \nu+\eta\left(N_{i}^{A}-r_{i}^{A}\right)\right)\)
    \(N_{i}^{A}\) : total number of instances of the ith noun with "a" or "the"
    \(N_{i}^{A}\) : total number of instances of the ith noun with "a" or "the"
    \(r_{i}^{A}\) : no. of successes (observances of the) for the ith noun
    \(r_{i}^{A}\) : no. of successes (observances of the) for the ith noun
    \(d_{i}^{C} \sim \operatorname{Bern}\left(\mu_{i}^{C}\right)\)
    \(d_{i}^{C} \sim \operatorname{Bern}\left(\mu_{i}^{C}\right)\)
    \(d_{i}^{A} \sim \operatorname{Bern}\left(\mu_{i}^{A}\right)\)
    \(d_{i}^{A} \sim \operatorname{Bern}\left(\mu_{i}^{A}\right)\)

Figure 4: Graphical representations of the metric and linking model. Shaded nodes indicate observed data (determiner-noun productions) or uninformative priors set by the researcher.
(individual determiner observations) are generated as draws from a binomial with parameter \(\mu_{i}\) for each of \(M\) noun types. These parameters are in turn drawn from a beta distribution with parameters \(\mu_{0}\) and \(\nu\). The \(\mu_{0}\) parameter describes the overall mean determiner preference, and the \(v\) parameterthe central target of inference-describes the degree to which individual noun preferences vary around the overall average \(\mu_{0}\). We complete the model via an uninformative prior distribution over \(\mu_{0}\) and \(v\).

Given a sample of determiner-noun pairings, we can use Bayesian inference to produce full posteriors over \(\mu\) and \(v\). In practice, we perform inference using Gibbs sampling via the JAGS package (Plummer, 2003); grid-sampling of posterior distributions and trace plots confirmed good convergence properties (see also the Appendix).

\section*{Corpus Selection and Extraction}

The Providence corpus (Demuth et al., 2006) consists of longitudinal in-home recordings from six children from New England and contains a relatively high density sample from the onset of single words at about \(1 ; 3\) to \(3 ; 0\). Utterances from each child and their mother were extracted from CHILDESformatted transcripts (MacWhinney, 2000) and augmented with a machine-generated syntax tier in CLAN (Sagae et al., 2010). Using these syntactic trees, we automatically extracted modifiers associated with each noun, as well as their part of speech. For the model input, noun uses were subset to those with a definite or indefinite determiner, yielding \(5-15 \times 10^{3}\) age-referenced DET+N tokens for the mothers and \(1.5-5 \times 10^{3}\) for the children.

For each mother and child, we performed a sliding-window analysis, examining successively older subsections of the corpus. On the basis of artificial corpus simulations (see Appendix), a window size of 1024 tokens was selected. On a linear sequence of tokens, each new window contained 10 new tokens from the full dataset and omitted the earliest 10.

This method yielded on average 150 measures of determiner productivity for each speaker. Additionally, an overlap measure was calculated for each 1024 token window according to the procedure described in Pine \& Lieven (1997).

\section*{Results and Discussion}

An item-based learning theory predicts a developmental increase in children's generalization across nouns (as measured by \(v\) ) as individual item-based constructions give way to a general production rule. In contrast, a theory positing full morphosyntactic productivity predicts no major difference in generalization over development; instead, children and parents will show the same level of productivity from early on.

Our sliding window analysis reveals no clear developmental trend in children's productivity (Figure 5), consistent with the early productivity account. For both the adult and the child, individual conversational bouts show high variance, but \(v\) values for the children are as higher or higher than those in the speech of their mothers, and children exhibit adult-like peaks of noun groupedness from the beginning of production. Nevertheless, for several of the children (e.g. William, Ethan, Violet), it is clear that the amount of data is not sufficient to allow the temporal granularity for a strong test.

Although it gives similar results to the overlap statistic, the \(v\) parameter in our model is preferable. While the overlap measure is confounded by sample size (see above), additional data only improves our estimate of \(v\). Posterior inference gives an explicit representation of the model's uncertainty in a data set, making it readily apparent when the sample size is too small to estimate model parameters.

This property of the model allows us to note that the variability in the estimates of productivity for adults and children seem to be quite reliable. In both cases, there is substantial variability that is not explained by the child's age. We hypothesize that this variability is due to the changing conversational and discourse dynamics between recordings in the


Figure 5: Sliding window analysis results. The metric model shows no clear developmental trend in children's productivity, nor a major difference in productivity between children and their mothers. On the left, black points and vertical gray bars represent the mean of the posterior and the \(95 \%\) highest posterior density interval on \(v\); horizontal gray bars show the temporal extent of the window used in the model at each point.
corpus, leading to the introduction by chance of many nouns with similar or dissimilar determiner preferences in context. Denser data will be needed, however, to test this hypothesis more fully.

\section*{Linking Model}

Although the metric model's results are suggestive of productivity from the earliest ages of children's determiner production, several aspects of the metric model limit the strength of the conclusions we can draw from it. First, the model fails to control for differences in the distribution of nouns for which determiners are produced by the speaker. For example, if children's determiner-noun productions disproportionately involve high-frequency nouns compared with adult productions, and if higher-frequency nouns tend to have more balanced determiner preference, it would inflate the metric model's estimate of children's productivity.

Additionally, an advocate of the island-learner position could justly point out that a child might produce relatively equal numbers of both determiners for a given noun Ywhich disfavors low values of \(v\)-not due to generalization but because the child has learned both "a Y" and "the Y" as islands from the input. Our linking model remedies these shortcomings by explicitly linking the determiner preference for child productions of a given noun to the experience the child has had with that noun in input from the caregiver. In the linking model, \(v\) more directly represents the strength of a child's generalization across nouns: as \(v\) approaches zero, we have a true island learner whose productions for a given noun reflect only experience with that noun from adult input;
as \(v\) approaches infinity, we have a true overgeneralizer for whom noun-specific variabilty in determiner frequencies in input are completely ignored. While it allows for a more nuanced picture of the relationship between a child's input and his or her productions, the linking model does not allow us to compare measures of adult and child productivity directly; in this sense it is complementary to the metric model.

\section*{Model Details}

The generative structure for the linking model is given on the right side of Figure 4. As before, we assume a hierarchical beta-binomial model linking different noun-specific determiner preferences together into a general determiner preference with mean \(\mu_{0}\) and scale \(v\). Here, however, adult determiner productions \(d^{A}\) for a given noun in the child's input contribute explicitly to the child's determiner preference \(\mu\) for that noun. We formalize the effect of the input on the child's determiner preference by assuming that the child acts as an ideal observer. Adult input for a given noun serves as binomial count observations, which the child combines with its beta-prior pseudocounts to yield Bayesian inference on the posterior distribution over the determiner preference for that noun. \({ }^{3}\) We allow adult input to be downweighted by a "for-

\footnotetext{
\({ }^{3}\) Note, however, that while the linking model contains an idealobserver component, it is not an ideal-observer model in its totality. Most critically, \(\mu_{0}\) and \(v\) are not learned by the child from adult data, but rather reflect the relationship between adult input and the child's productions. In principle, the child's productions can even be highly discrepant from the adult input, if \(v\) is large and \(\mu_{0}\) does not match the overall distribution of adult determiner use. Conversely, if the posterior on \(\mu_{0}\) is a close match to adult determiner use, it suggests
}


Figure 6: Linking model results for inference on \(v\) (left) and simulated vs. empirical overlap measure (right). Black points and vertical gray bars represent the mean of the posterior and the \(95 \%\) highest posterior density interval on \(v\), respectively. Horizontal gray bars express the temporal extent of the window used to fit the model at each point.
getting" or "noise" parameter \(\eta\), motivated not only theoretically from the consideration that a child is unlikely to be able to store and learn with perfect fidelity from every determinernoun production in its input, but also empirically: without it, it is hard for even an extreme island learner version of our model to reproduce a pattern sometimes seen in our dataset, where the determiner distribution for a given noun will be relatively balanced for the adult but highly skewed for the child.

\section*{Dataset, Results and Discussion}

We used the same window size (1024 tokens) as for the metric model for a sliding window analysis using the linking model, but used all parent data up to and including the period of child usages for each window. Results of the linking model indicate that children generalize beyond the input that they receive (Figure 6, left), though there is some evidence of variation across children in generalization strength: the weakest generalizer, Alex, shows a \(v\) around 0.6 , and the strongest generalizer, William at the latest stage in our dataset, shows a \(v\) around 2.3 Posterior means for \(\eta\) varied between 0.071 and 0.599 , with substantial variation between children; posterior means for \(\mu_{0}\) varied between 0.145 and 0.717 . As with the metric model, we observed no evidence for a developmental trend from lesser to greater generalization: while some children (Ethan, Violet, William) seem to show a trend toward increasing \(v\) over time, other children (Naima, Alex) show no directional trend, and one child (Lily) has a decreasing trend.

Finally, although we have argued that the overlap measure is not useful for quantifying productivity across sample sizes, we can use it as a goodness-of-fit metric for our model within a sample. We do this by using the adult data and the joint posterior of the fitted model for each window to generated

\footnotetext{
that the child is indeed generalizing from adult productions across nouns in his or her production behavior.
}
simulated determiner productions for the specific noun distribution in that window, and comparing the overlap measure for the simulated data with the empirical overlap measure in that window. For nearly all windows of all children, empirical overlap falls within the range of simulated overlaps, validating the model's overall fit to the data (Figure 6, right).

\section*{General Discussion}

We constructed two models to quantify the productivity in children's early determiner usage and to compare this to that of their mothers. These models instantiated a statistical tradeoff between memorization of the observed data ("island learning") and extreme generalization. Results from both models suggested that the children in our sample were neither extreme generalizers nor extreme island learners. Contra the constructivist hypothesis, neither model provided clear evidence for developmental change in children's generalization behavior over time, and by the summary measure of productivity furnished by the metric model their speech was not quantitatively distinguished from that of their parents. Yet contra the full-productivity nativist hypothesis, there is clear evidence for item-specific combinatorial preferences between determiners and nouns ( \(v\) values are relatively low in the metric model; compare Figures 3 and 5) and that children are at least somewhat sensitive to the specifics of adult input ( \(v\) values are low in the linking model).

Nevertheless, while the current results are consistent with early productivity, our modeling work leaves unaddressed a number of issues that both preclude a conclusive judgment in this debate thus far and also point the way towards future work. As we alluded to when introducing the linking model, it is difficult to rule out the possibility that apparently "productive" determiner behavior for a given noun may reflect the child's having learned both determiners with that noun as is-
lands. This difficulty is compounded by the fact that though the Providence corpus is extensive, it still records only a small fraction of the total adult input each child in the corpus has received. In the future this difficulty may be addressed by more complete datasets; additionally, our model could be extended by allowing imputation of unrecorded adult data, which would allow our uncertainty regarding the content of this input to be incorporated into inferences about productivity in child behavior.

A second challenge is that an advocate of the fullgeneralization position could reasonably object that nounspecific determiner preferences in child productions that mirror adult input may be driven by other factors to which both adults and children are sensitive in determiner production, such as referential context (e.g., Maratsos, 1979; KarmiloffSmith, 1981). Our model could be extended to account for these effects by conditioning determiner probabilities not only on noun identity but also on other contextual factors recoverable from corpus data; this move might allow a richer investigation of the developmental trajectory of how these aspects of the knowledge underlying fully proficient determiner deteminer use are learned and used in naturalistic production.

The gold standard for demonstrating the existence of productive knowledge of determiner syntax would, of course, be the combination of a novel noun with determiners that the child has not yet heard used with that noun. Regardless of the outcome of such a study, however, we believe that our probabilistic, data-driven approach would retain potential to advance our understanding of how linguistic knowledge develops. The modeling framework presented here provides an alternative to the extreme positions of all memorization or all generalization embodied by constructivist and nativist viewpoints. Although our model contained many simplifying assumptions, including not only those mentioned above but also the restriction to two determiners, it has given initial traction in measuring how experience from local episodes may lead to global generalizations. For the problem of determiner productivity, the simplifying assumptions can be relaxed one by one; and the general architecture can be applied to study a broad range of phenomena beyond the development of determiners, such as the emergence of plural markings and other morphological generalizations. We hope that exploring the space of models that combine the best features of both islandand generalizing-learner accounts may lead to new insights into the emergence of productive language.

\section*{Acknowledgments}

Thanks to the members of the Language and Cognition Lab at Stanford and the Concepts and Cognition Lab at UC Berkeley for valuable discussion. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1106400.

\section*{References}

Booth, T. L. (1969). Probabilistic representation of formal languages. In IEEE conference record of the 1969 tenth annual sym-
posium on switching and automata theory (pp. 74-81).
Demuth, K., Culbertson, J., \& Alter, J. (2006). Word-minimality, epenthesis, and coda licensing in the acquisition of english (Vol. 49).
Demuth, K., \& McCullough, E. (2009). The prosodic (re)organization of children's early english articles. J. of Child Language, 36, 173-200.
Frank, M. C., Tenenbaum, J. B., \& Gibson, E. (2013). Learning and long-term retention of large-scale artificial languages. PLoS ONE, 8, e52500.
Gelman, A., Carlin, J., Stern, H., \& Rubin, D. (2004). Bayesian data analysis. Chapman \& Hall/CRC.
Karmiloff-Smith, A. (1981). A functional approach to child language: A study of determiners and reference (Vol. 24). Cambridge University Press.
Kramsky, J. (1972). The article and the concept of definiteness in language. Paris: Mouton.
MacWhinney, B. (2000). The childes project: Tools for analyzing talk. Lawrence Erlbaum Associates.
Maratsos, M. (1979). Learning how and when to use pronouns and determiners. Language acquisition. Cambridge, 225-240.
Pine, J. M., \& Lieven, E. (1997). Slot and frame patterns and the development of the determiner category. Applied Psycholinguistics, 18, 123-138.
Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling.
Sagae, K., Davis, E., Lavie, A., Macwhinney, B., \& Wintner, S. (2010). Morphosyntactic annotation of childes transcripts. J. of Child Language, 37, 705-729.
Tomasello, M. (2003). Constructing a language: A usage-based theory of language acquisition. Cambridge, MA: Harvard University Press.
Valian, V. (1986). Syntactic categories in the speech of young children. Developmental Psychology, 2, 562-579.
Valian, V., Solt, S., \& Stewart, J. (2009). Abstract categories or limited-scope formulae? the case of children's determiners. J. of Child Language, 36, 743-778.
Yang, C. (2010). Who's afraid of George Kingsley Zipf? (Unpublished manuscript)

\section*{Appendix: Model Validation with Artificial Corpora}

To test the validity of our Gibbs sampling procedure procedure and establish the minimum number of \(\mathrm{DET}+\mathrm{N}\) samples necessary to parameterize the model, we tested the metric model on artificial noun and determiner counts generated according to known statistical properties. We varied \(\mu\) was from .1 to .9 in increments of .1 , and \(v\) at \(.05, .1, .5,1,5,10\), and 50. We additionally varied the number of tokens from \(2^{0}(1)\) to \(2^{24}\left(1.6 \times 10^{7}\right)\) (the upper limit corresponding to the order of magnitude of tokens heard by a child; Frank et al., 2013), with token distributions generated from both uniform and Zipfian word frequency distributions.

As in the main simulations, we estimated posteriors for the parameters \(\mu\) and \(\nu\) and compared with the known \(\mu\) and \(v\) used to generate the input data. MCMC chains here and in the main simulations consisted of 1000 samples after a burn-in of 1000 adaptive samples and 1000 updates, with no thinning. We employed Gelman diagnostics as well as manual inspection of traces to check for sufficient burn-in time and mixing. Grid sampling confirmed that likelihoods were sufficiently peaked to constrain parameter estimates and were consistent with posteriors produced inferred with MCMC. Measures of the reliability of inference (mean and standard deviation in the difference from the true value) helped establish a minimum window size for sliding window analyses to correspond with error less than some fixed value \(\varepsilon\).

\title{
Spatial Reasoning in Native Speakers of Russian and German
}

\author{
Maria Mikheeva (maria.micheewa@mail.ru) \\ Kazan (Volga Region) Federal University, Department of Psychology, Kremlyovskaya-Str. 18, 420008 Kazan, Russian Federation
}

\author{
Leandra Bucher (leandra.bucher@psychol.uni-giessen.de) \\ Justus Liebig University, Experimental Psychology and Cognitive Science, Otto-Behaghel-Str. 10F, 35394 Giessen, Germany
}

\author{
Jelica Nejasmic (jelica.nejasmic@ psychol.uni-giessen.de) \\ Justus Liebig University, Experimental Psychology and Cognitive Science, Otto-Behaghel-Str. 10F, 35394 Giessen, Germany
}

\author{
Markus Knauff (markus.knauff @ psychol.uni-giessen.de) \\ Justus Liebig University, Experimental Psychology and Cognitive Science, Otto-Behaghel-Str. 10F, 35394 Giessen, Germany
}

\begin{abstract}
The relationship between reasoning and language has been frequently studied. Here we explore principles of spatial reasoning in Germans and Russians. We compared the performance of Russians in three different settings to the performance of Germans. The task was to construct layouts of wooden blocks according to verbal instructions, describing the relations of these blocks. Subsequently pieces of new information, introduced as incontrovertible facts and partly contradicting the initial descriptions, were given. Participants re-arranged the blocks to take into account the new facts. Recent research conducted with Germans has shown that although alternatives are logical equivalent - there are preferences for certain solutions. The question was whether Russians show the same or different preferences. Our results suggest that construction and revision of spatial models follow similar principles. However, we observed differences between the groups regarding the flexibility to apply a principle based on the order of words in a sentence.
\end{abstract}

Keywords: Spatial reasoning; Relational reasoning, Crosscultural similarities; Language; Russia; Belief revision

\section*{Introduction}

Misunderstandings happen so often between people from different countries, triggering the important question: how is language connected to our mental representation of the world? What role does it play in reasoning? Answers suggested to that question are provided by an important theory in that area: the Sapir-Whorf hypothesis affirms that language influences thought (Zvegintsev, 1960; Levinson, Kita, Haun, \& Rasch, 2002; Levinson \& Meira, 2003). On the other hand, there is the view that the mind is organized in a modular way with separate modules dedicated to certain abilities (e.g. Tsimmerling, 2000; Nowak, Komarova, \& Niyogi, 2001; Kulikov, 2012). A further important question, to some extent related to language, is: how do mental representations differ across different cultures? What role do cultural backgrounds play in reasoning? While some studies suggest cultural dissimilarities (e.g. Oyserman \& Lee,
2008), other studies show that there are common cognitive principles used by reasoners of different cultures. Crosscultural similarities have, for instance, been shown in topological reasoning (Knauff \& Ragni, 2011; Knauff, 2013).

The present study is concerned with spatial relational reasoning and the influence of language and culture. We briefly analyze relevant work on linguistic influence on thinking, comparative topology of German and Russian, and spatial relational reasoning. We then present an experiment, designed to investigate the construction and revision of spatial models.

\section*{Construction and Revision of Spatial Mental Models}

Imagine you need to find the house № 28 in a street, unfamiliar to you. You have received the following description of the precise location by friend A , informing you that:
(1) "There is a hotel to the right of a café.", and
(2) "The house № 28 is to the left of the café."

The description allows for one (determinate) model to construct. In order to construct the model, spatial information is inserted successively. Based on the information given by statement (1), the model
(3) "Café - Hotel"
is initiated and extended by (2) "House № 28 ", resulting into the model:
(4) House № 28 - Café - Hotel

A lot of studies have explored factors that influence reasoners when they construct models, among them the order of objects as inserted into the model, and other order effects (e.g. Payne, 1993; Ehrlich \& Johnson-Laird, 1982; Payne \& Baguley, 2006; Bucher, Krumnack, Nejasmic, \& Knauff, 2011; Krumnack, Bucher, Nejasmic, \& Knauff, 2011; Nejasmic, Krumnack, Bucher, \& Knauff, 2011).

Imagine you find out a little later that the information uttered by friend A is unreliable. Friend B - who lives in the street in question - informs you that as a fact:
(5) "The house № 28 is to the right of the hotel."

The more reliable and incontrovertible information partially contradicts friend A's description needs to be taken into account. The following alternatives are possible:
(6) Café - Hotel - House № 28
(7) Hotel - House № 28 - Café

Both variations of the initial model are logically equivalent. Nevertheless, when confronted with ambiguous relational information, human reasoners frequently prefer one alternative over the other (Jahn, Knauff, \& JohnsonLaird, 2007; Krumnack, Bucher, Nejasmic, \& Knauff, 2010; Bucher et al., 2011; Krumnack, Bucher, Nejasmic, Nebel, \& Knauff, 2011; Bucher \& Nejasmic, 2012; Knauff, Bucher, Krumnack, \& Nejasmic, 2013).

\section*{Preferred model revision}

The process of model revision with verbal descriptions, using binary relations \(\mathrm{r}(\mathrm{X}, \mathrm{Y})\) as facts has been shown to rely on the following principle: the functional distinction of X as the "to-be- located object" (LO) in contrast to Y as the "reference object" (RO) specifies the location of the LO relative to the known location of the RO (e.g. Huttenlocher \& Strauss, 1968; Miller \& Johnson-Laird, 1976; Talmy, 1983; Landau \& Jackendoff, 1993). For the revision of horizontal linear arrangements, the following finding concerning reasoners' preferences is characteristic
Initial arrangement \(\quad\) A B C
Counterfact \(\quad \mathrm{C}\) is left of A,
with \(C\) as the relation's \(L O\)
with \(C\) as the relation's \(L O\)
Preferred revision: C A B
Note that the logical equivalent (non-preferred) alternative for revising the initial model by relocating the counterfact's RO (here: A) would results in the revised model: B C A. We refer to the preferred principle as the LO-principle (compared to the RO-principle).

\section*{Linguistic Influence on Thought and Comparative Typology of German and Russian Languages}

The Sapir-Whorf hypothesis postulates that language determines thought or at least that linguistic categories influence thought and certain kinds of non-linguistic behavior (Zvegintsev, 1960). Li and Gleitman (2002) investigated influences of individual languages (e.g. English and Dutch) on spatial reasoning. Chatterjee (2011) studied language as a form of mental representation of space. With the current study, we investigate influences of different languages (Russian vs. German) on spatial mental representations. Moreover, to dissociate between influences that result from linguistic aspects on the one and cultural aspects on the other hand, the study took place in two cultural settings (in Germany and in Russia).

First, we briefly explain the structures of both languages in terms of comparative typology. Both languages are from the Indo-European family. German belongs to the West Germanic family, Russian is a Slavic language. They are inflexional languages (from Lat. Flectivus «flexible»). The term refers to a language, where word-building with inflexions dominates. Inflexions are morphemes which can have much significance; e.g. the article „die" (as in "die Katze", "the cat") in German, indicates the gender (feminine), the case (nominative), and the number (singular). Russian is even more inflexional compared to German. Inflexional languages can be synthetic or analytic. The German language is between the synthetic and analytic languages, it has some characteristics of both language types. In a synthetic language, a word contains all the grammar, e.g., by inflexional endings, prefixes, suffixes. An analytic language is a language which reproduces grammatical relationships syntactically. Accordingly, it uses only unbound morphemes, and only separate words like articles etc. (Anokhina \& Kostrova, 2006). For instance: "хорошая новость" and "eine gute Neuigkeit" ("good news") - in Russian, the ending ,"as", and in German the (additionally needed) article „eine" indicates: feminine, nominative, singular. There are some more differences between German and Russian which we do not want to explain here in detail. What is relevant for the current study is the flexibility in word order in the two languages. In German, the possibility of ordering words within a sentence in a certain way is much more limited compared to Russian. Of course, the "freedom" of word order in Russian is not unlimited and also regulated by semantic and stylistic factors (as in German) (Anokhina \& Kostrova, 2006). An example is given in table 1 .

Table 1. Word order in German and Russians languages
\begin{tabular}{lll}
\hline Russian & German & Exact meaning \\
\hline Я люблю тебя & Ich liebe Dich & I love you \\
\hline Тебя я люблю & \begin{tabular}{l} 
Word order \\
not possible
\end{tabular} & \begin{tabular}{l} 
I love you, exactly \\
you, not another \\
person
\end{tabular} \\
\hline Я тебя люблю & \begin{tabular}{l} 
Word order \\
not possible
\end{tabular} & \begin{tabular}{l} 
I love you, exactly \\
you, not another \\
person
\end{tabular} \\
\hline Люблю я тебя & \begin{tabular}{l} 
Word order \\
not possible*
\end{tabular} & \begin{tabular}{l} 
I love you (with an \\
even stronger \\
significance in the \\
sense that I can do \\
nothing about it)
\end{tabular} \\
\hline Люблю тебя я & \begin{tabular}{l} 
Word order \\
not possible**
\end{tabular} & \begin{tabular}{l} 
You are loved by me, \\
not someone else
\end{tabular} \\
\hline Тебя люблю я & Dich liebe ich & \begin{tabular}{l} 
I love you, exactly \\
you, not another \\
person
\end{tabular} \\
\hline
\end{tabular}
*the word order would be possible in a German question ("Liebe ich Dich?"); **the word order would be possible in a German passive sentence ("Du wirst von mir geliebt.")

Analogously, the Russian spatial language is also more flexible than the German language. For example, dynamic local relations in German indicate source locations ("where from?") and directions of motion ("where to?"), while in Russian such relations have a triple function, they indicate: location ("where"), source location, and direction (Khoruzhaya, 2007).
To summarize, there are major differences between German and Russian. Previous research on spatial relational reasoning suggests that reasoners have strong preferences which are often based on linguistic cues that are connected to the sentence structure. The main finding in a range of experiments on the variation of spatial models (Bucher et al., 2011; Krumnack et al., 2011; Bucher \& Nejasmic, 2012; Knauff, et al., 2013) is that the variation is preferably done by the relocation of objects that are perceived as more flexible compared to other objects. These objects are usually the so called to-be-located objects (LO) of a relational statement as compared to reference objects (RO) which are perceived as more stationary. We refer to this preference as the LO-principle.
The question is whether cognitive principles such as the LO-principle are used independently from linguistic or cultural aspects. Concerning the language aspect, it would be plausible if Russians, i.e. native speakers of a language that is by nature very flexible concerning the word order and sentence structure are accordingly more flexible in the application of such cognitive principles. In this specific case they might use the RO-principle more frequently as an alternative solution in the reasoning task. Culture is another important aspect to look at when we look at similarities of cognitive principles. There are many definitions of culture. Oyserman and Lee (2008, p. 311) say that "culture matters to the extent that individuals living in different societies are likely to have differing experiences". Criado (2009, p. 295) explains that culture is "a set of shared values, beliefs, expectations, customs, jargon, and rituals". What seems to be indisputable is that a cultural environment can have an impact on the way an individual thinks.
In order to explore both, language and culture influence, we conducted the same experiment with native speakers of German and of Russian as participants. Three different samples of Russian participants were tested in two different cultural environments:
1. The first sample was tested in German, in Germany
2. The second sample was tested in Russian in Germany
3. The third sample was tested in Russian in Russia

The purpose was to control for both, the language and the cultural setting.

\section*{Experiment: Construction and Revision of Block Arrangements}

\section*{Method}

The first part of the experiment can be referred to as "construction phase". The task was to physically construct layouts of wooden blocks according to a verbal instruction,
describing the relations of these blocks. The second part can be titled "revision phase". Once a layout was constructed, a piece of new information, introduced as an incontrovertible fact contradicted a part of the initial description. The task was to re-arrange the blocks such that it cohered with the "fact".
Participants Altogether, we tested 76 volunteers who performed in the task either in German or in Russian in Germany or in Russia. All participants gave informed consent to participation. Participants were tested individually. Each participant was tested only once.

Language abilities were assessed by self-report. Russian participants tested in Germany rated their German language abilities as "very good", and reported to be capable of writing and speaking fluently. They were fluent in Russian as their mother tongue and in German as a second language, and have been living and were educated in Germany for a considerable time. Russian participants tested in Russia reported to be not familiar with the German language while German participants reported to be not familiar with the Russian language.

The sample of Germans tested in German in Germany consisted of 11 ( 5 male; age: \(M=24.91 ; S D=2.95\) ) native speakers of German, all students from the University of Giessen. None of them has ever studied Russian.

The sample of Russians tested in German in Germany consisted of 19 ( 3 male; age: \(M=24.05\); \(S D=4.18\) ) native speakers of Russian.
The sample of Russians tested in Russian in Germany consisted of 20 ( 3 male; age: \(M=25.45\); \(S D=5.26\) ) native speakers of Russian.

The sample of Russians tested in Russian in Russia consisted of 26 ( 1 male; age: \(M=20.35\); \(S D=0.63\) ) native speakers of Russian. They were all students from the Federal University of Kazan (among them 19 students of psychology). None of them has ever studied German or has been to Germany.
Materials, Procedure, and Design 32 items were presented, each consisting of two premises and an inconsistent fact. The task was presented on a 19 "-computer screen, using Microsoft PowerPoint (Version 2007) running in the windows environment XP on a standard personal computer. PowerPoint slides were presented by the experimenter in a sequentially and individually adapted manner according to participants' performance.
In all items, the two premises and the contradictory fact (presented in red) had the surface structure as follows: first object - relation (either "left of" or "right of") - second object.

\section*{Example: "Yellow left of red"}

Participants were provided with wooden square blocks (size: \(2.5 \times 2.5 \times 2.5 \mathrm{~cm}\) ), red, green, yellow, and blue colored on a plate in front of them and instructed to construct and subsequently revise their block layouts.
The construction phase: participants were instructed to pick up the colored blocks, one at a time using one hand, and arrange them according to the information provided by
the premises into a linear one-dimensional order. The premises informed about the determinate order of the blocks with the blocks represented by the respective colors (red, green, yellow, and blue).

Example:
1st premise: "Blue right of red"
2nd premise: "Green right of blue"
Spatial arrangement: red - blue - green
The location of the third object was counter-balanced across all problems. In a recent experiment, very similar to the one reported here, and presented in German to German participants, word order has been shown to be crucial for the physical construction of spatial models (Bucher et al., 2011, Experiment 2). Here, accordingly, based on the description of the \(1^{\text {st }}\) premise, two possible construction orders were possible:
1. Starting on the left side and continue to the right, e.g. (consider the 1st premise from the above example) putting down the red block first and placing the blue block to the red one's right side.
2. Starting on the right side and continue to the left, e.g. putting down the blue block first and placing the red block to the blue one's left side
The resulting orders are describable as \(1-2-3\) and \(2-1-\) 3 , with the numbers indicating the order by which objects had been put down; e.g. red first - blue second - green third (order \(1-2-3\) ) and red second - blue first - green third (order 2-1-3). The question was whether there would be order effects when constructing the arrangements in the Russian samples similar to those found in Germans.
The revision phase: subsequently after participants had constructed the order of the three colored blocks, they were asked to revise their order according to the inconsistent fact.

Example-fact: "Green left of red"
Participants were free with the revision of their initially constructed arrangements. After each trial, the wooden blocks were put back onto the plate by the experimenter. Four practice trials (neither recorded nor analyzed) preceded the experimental trials. Performance was recorded on a video tape by the experimenter and analyzed after the experimental session (Bucher et al., 2011).
In previous experiments, using a very similar experimental set-up, presented in German to German participants, the finding was that of a clear preference (e.g. \(89.52 \%, S D=\) 11.30; see Bucher et al., 2011, experiment 2) of LO relocations as compared to RO relocations. The question was whether participants of the Russian samples would apply the LO-principle similarly to the German participants.

\section*{Results and discussion}

Construction: Mean percentage rate of correctly constructed orders was \(97 \%\) ( \(S D=4.46\) ). There was no difference in number of mistakes in construction between the samples ( \(p>0.30\) ). We analyzed the order of objects put down during the construction in every sample, running Wilcoxon's tests for each sample, separately (Siegal \& Castellan, 1988). In the German sample the 1-2-3-order was
used more frequently compared to the 2-1-3-order ( \(z=-\) 2.01; \(p<0.05\) ). The same applies to the remaining samples: Russians tested in German in Germany ( \(z=-3.93\); \(p<0.01\) ), Russians tested in Russian in Germany ( \(z=-3.97 ; p<0.01\) ), and Russians tested in Russian in Russia ( \(z=-4.49 ; p<\) 0.01 ). Our results provide evidence that the same principle (putting down blocks in word order 1-2-3 rather than in the order 2-1-3) was applied similarly by participants of all four samples. Thus, we can conclude that the cognitive principle is not affected by linguistic aspects. Russians tested in Germany as well as Russians tested in Russia used the same principle, suggesting cross-cultural similarities.
We continued our analysis by comparing the magnitude of the preference applied by Germans and Russians in the different settings. The Kruskall-Wallis test revealed a difference between the samples ( \(\mathrm{p}<0.05\) ). Pair wise comparisons, using Wilcoxon's tests revealed that the group of Germans differed from all Russian samples (all \(p s>.05\) ).
Despite the overall similarities found across all samples, we found that Russians were more strictly in the application of this principle. The Germans performed more flexible in the task, using the alternative word order (2-1-3) more frequently. Figure 1 depicts the result graphically. We continue to discuss this difference in the General Discussion of this paper.

Revision: Mean percentage rate of correctly revised models was \(99.14 \% ~(S D=1.74)\). There were no differences in the amount of mistakes between the samples ( \(p>0.40\) ). Erroneous problems were excluded from further analyses. We ran Wilcoxon's tests for each sample, separately. That was to analyze which principle the revision followed. The tests indicated that in the German sample LOs were relocated more frequently then ROs \((z=-2.95 ; p<0.01)\). The same principle was applied by reasoners in the other samples: Russians tested in German in Germany ( \(z=-3.83 ; p\) < 0.01), Russians tested in Russian in Germany ( \(z=-3.96 ; p\) \(<0.01\) ), and Russians tested in Russian in Russia ( \(z=-4.56\); \(p<0.01\) ). Again, our results suggest similarities across both language groups. There was a clear preference for LO relocations across all samples (figure 2 depicts the results graphically). The principle was equally used by Russian native speakers, who were tested in Russian and in German. This suggests that linguistic aspects were not modulating the effect. Russians tested in Germany as well as


Figure 1. The figure depicts the difference between Germans and Russians during the construction of block arrangements.
The word order effect was more pronounced in Russians than in Germans. Error bars indicate standard errors.


Figure 2. For revision, the LO principle (relocation of the to-be-located object, LO as opposed to the reference object, RO) was preferably applied by participants of all samples. Error bars indicate standard errors.

Russians tested in Russia used the same principle, suggesting similarity across the cultures. We continued our analysis by comparing the magnitude of the preference applied by Germans and Russians in the different settings. The Kruskall-Wallis test revealed that there are no differences between the samples ( \(p>0.30\) ). Unlike the word order effect, the LO effect was equally strong across all groups. This finding further corroborates the assumption of a cross-cultural and cross-linguistc cognitive principle.

\section*{General Discussion}

The present study investigates aspects of spatial relational reasoning in reasoners from Germany and Russia. We explored principles applied for the construction and the revision of spatial models in four types of samples. A sample of German native speakers who were tested in German in Germany, a sample of Russian native speakers, tested under the same conditions, a sample of Russian native speakers tested in their mother tongue but in the German cultural environment, and a sample of native speakers of Russian who were naive to the German language, tested in their native Russian cultural environment.
The study was motivated by recent findings that principles applied by German reasoners in spatial relational reasoning tasks were based on linguistic cues. One study (Bucher et al., 2011) suggests that during the physical construction of spatial (block) arrangements, the word order plays a role in guiding the construction process while for the revision of these models the asymmetry of LOs and ROs of relational statements provide the crucial cues for reasoners (Bucher \& Nejasmic, 2012; Knauff et al., 2013). Here as a novelty, a similar task was presented to native speakers of Russian. We were concerned with the dissociation of linguistic and cultural aspects. In order to dissociate these aspects to a certain degree, we splitted the Russian group into three subsamples, allowing a rough distinction of cultural from linguistic influences. The results indicate cross-cultural similarities for both cognitive principles applied during construction and revision of spatial models. Across all samples (German and Russians), the construction followed the word-order-principle. This effect has been previously shown in Germans (Bucher et al., 2011) and could be repetitively shown here. The revision was found to be
guided by the LO-principle. This principle had been repeatedly shown in German reasoners and here - for the first time - in Russians. The effect was comparable in magnitude across all samples. We conclude that both principles reflect similar mechanisms.

However, Germans used the alternative principle (starting construction with the second object) more often compared to the Russian samples and performed thus construction processes more flexible. Please, note that with the current experiment, it is not distinguishable whether reasoners used the first object mentioned in the premise or the LO as a starting point for their construction, because the first object in a statement was identical with the LO. However, Bucher et al. (2011) argue that different cognitive principles are applied during construction (first vs. second object as starting object) and revision (relocation of LO vs. RO), respectively. The authors also provide empirical evidence for their view. The results of the current study show that German native speakers are more flexible when applying the word order principle, compared to Russians, while the LO-principle is applied equally robustly by both groups. This can be taken as corroborating evidence that the cognitive mechanisms underlying construction and revision are distinguishable, however comparable for Germans and Russians.
Nevertheless, we found differences in the manifestations of the word-order effect between the groups. The effect was stronger in Russians compared to Germans. This indicates that Russians used the alternative word order less frequently than Germans did when they constructed their models. When we bear in mind that the Russian language allows for many variations of word orders in a sentence, the result might look counterintuitive, at the first glance. However, we must note that speakers of Russian already make many decisions when they construct a sentence. Maybe, it is the compensation for this "liberty" in the canonical word order of the Russian language which we find reflected by the high adherence to the word order principle. Also important in the present context is that although we might have found crosscultural similarities between Germans and Russians, as well as cross-linguistic principles that were applied during the construction phase, there is an alternative interpretation of the results. While (as in previous experiments) German reasoners might have based the construction preferably on the word order, it is possible that Russian reasoners applied the LO principle, i.e. put the LO as first object on the table. With the present study we cannot rule out this alternative interpretation but we are currently running experiments designed to look at this problem.

\section*{Acknowledgments}

This study was supported by a grant of the Deutsche Forschungsgemeinschaft, (KN 465/6-2). It has been carried out in the framework of the partnership between Kazan (Volga Region) Federal University and Justus-Liebig University Giessen. We thank Anja Gatzsche for assistance with the experiment.

\section*{References}

Anokhina, S. P. \& Kostrova, О. А. (2006). Сравнительная типология немецкого и русского языков: Учебное пособие. СамГПУ Press.
Bucher L., Krumnack A., Nejasmic J., \& Knauff, M. (2011). Cognitive processes underlying spatial belief revision. In L. Carlson, C. Hölscher, T. Shipley (Eds.), Proceedings of the 33rd Annual Conference of the Cognitive Science Society (pp. 3477-3482). Austin, TX: Cognitive Science Society.
Bucher, L. \& Nejasmic, J. (2012). Relocating multiple objects during belief revision. In C. Stachniss, K. Schill, \& D. Uttal (Eds.), Lecture Notes in Artificial Intelligence: Spatial Cognition (pp. 476-491). Berlin, Germany: Springer.
Chatterjee, A. (2011). Language and space: Some interactions. Trends in Cognitive Sciences, 2, 55-61.
Criado, R. (2002) Taking scripts as a model of lesson organisation for the integration of culture and language in ELT. Revista Alicantina de Estudios Ingleses 22, 295-314.
Ehrlich, K. \& Johnson-Laird, P. N. (1982) Spatial descriptions and referential continuity. Journal of Verbal Learning and Verbal Behavior, 21, 296-306.
Huttenlocher, J. \& Strauss, S. (1968). Comprehension and a statement's relation to the situation it describes. Journal of Verbal Learning \& Verbal Behavior, 7, 300-304.
Jahn, G., Knauff, M., \& Johnson-Laird, P. N. (2007). Preferred mental models in reasoning about spatial relations. Memory and Cognition, 35, 2075-2078.
Khoruzhaya, ㄴ. (2007) Семантические функиии пространственных и временных предлогов в немеиком и русском языках. Kuban State University Press.
Knauff, M. (2013). Space to Reason. Cambridge, MA: MIT Press.
Knauff, M. \& Ragni, M. (2011). Cross-Cultural Preferences in Spatial. Reasoning. Journal of Cognition and Culture, 11, 1-21.
Knauff, M., Bucher, L., Krumnack, A., \& Nejasmic, J. (2013). Spatial belief revision. Journal of Cognitive Psychology, 1, 1-10.
Krumnack, A., Bucher, L., Nejasmic, J., \& Knauff, M. (2010). Spatial reasoning as verbal reasoning. In S. Ohlsson \& R. Catrambone (Eds.), Proceedings of the 32nd Annual Conference of the Cognitive Science Society (pp. 1002 - 1007). Austin, TX: Cognitive Science Society.
Krumnack, A., Bucher, L., Nejasmic, J., \& Knauff, M. (2011). Efficiency and minimal change in spatial belief revision. In L. Carlson, C. Hölscher, \& T. Shipley (Eds.), Proceedings of the 33rd Annual Conference of the Cognitive Science Society (pp. 2270-2275). Austin, TX: Cognitive Science Society.
Krumnack, A., Bucher, L., Nejasmic, J., Nebel, B., \& Knauff, M. (2011). A model for relational reasoning as verbal reasoning. Cognitive Systems Research, 11, 377392.

Kulikov, D. (2012). Гипотеза лингвистической относительности Сепира-Уорфа и её методологическая оценка в работах М.К.Петрова. Политическая концептология, 2, 189-199.
Landau, B. \& Jackendoff, R. (1993) "What" and "where" in spatial language and spatial cognition. Behavioral and Brain Sciences, 16, 217-265.
Levinson, S. C. \& Meira, S. (2003). 'Natural concepts' in the spatial topological domain. Language 79, 485-516.
Levinson, S., Kita, S., Haun, D., \& Rasch, B. (2002). Returning the tables: Language affects spatial reasoning. Cognition, 84, 155-188.
Li, P. \& Gleitman, L. (2002). Language and spatial reasoning. Cognition, 3, \(265-294\).
Miller, G. A. \& Johnson-Laird, P. N. (1976). Language and Perception. Cambridge MA: Havard University Press.
Nejasmic, J., Krumnack, A., Bucher, L., \& Knauff, M. (2011). Cognitive processes underlying the continuity effect in spatial reasoning In L. Carlson, C. Holscher, \& T. Shipley (Eds.), Proceedings of the 33rd Annual Conference of the Cognitive Science Society (pp. 11271132). Austin, TX: Cognitive Science Society.

Nowak, M., Komarova, N., \& Niyogi, P. (2001). Evolution of universal grammar. Science, 291, 114-118.
Oyserman, D. \& Lee, S. (2008). Does culture influence what and how we think? Effects of priming individualism and collectivism. Psychological Bulletin, 2, 311-342.
Payne, S. J. (1993). Memory for mental models of spatial descriptions: An episodic construction trace account. Memory \& Cognition, 21, 591-603.
Payne, S. \& Baguley, T. (2006). Memory for the process of constructing an integrated mental model. Memory \& Cognition, 34, 817-825.
Siegal, S. \& Castellan, N. J. (1988). Nonparametric tests for the behavioral sciences. McGraw-Hill College
Talmy, L. (1983). How language structures space. In H. L. Pick \& L. P. Acredolo (Eds.), Spatial Orientation: Theory, Research and Application (pp. 225-282). London, UK: Plenum.
Tsimmerlimg, A. (2000). Американская лингвистика сегодняшнего дня глазами отечественных языковедов. Вопросы языкознания, 2, 1-20.
Zvegintsev, V. A. (1960). Теоретико - лингвистические предпосылки гипотезы Сепира - Уорфа. Новое в лингвистике, 1, 111-134.

\title{
Problem Solving Between Action-Selection and Action-Completion in a Simple Domain
}

Gareth E. Miles (gareth.miles@southwales.ac.uk)
School of Psychology, University of South Wales, Trefforest, CF37 1DL, UK

\author{
Stephen J. Payne (s.j.payne@bath.ac.uk) \\ Department of Computer Science, University of Bath, Bath, BA2 7AY
}

\begin{abstract}
Experiment 1 demonstrates that problem solving knowledge can be applied while a move is in progress in certain Tower of London (ToL) problems. A two-stage move process is often delayed in the second stage when participants have been misled by similarity to a previous problem. We suggest this is indicative of misgivings about the chosen move caused by on-going analysis of the move that is being made. Experiment 2 swapped the stages of the two-stage process and again reported more hesitancy in the second stage when participants had been misled. We conclude that it is desirable for models of problem solving to evolve so that they can apply the same learned problem solving knowledge both before a move is selected and while the move is being made. We then describe a model of ToL problem solving that fulfills these criteria and has been computationally-implemented within an embodied cognitive architecture.
\end{abstract}

Keywords: Problem Solving, Embodied Cognition, Production Systems

\section*{Introduction}

The current paper tries to distinguish between two accounts of how a move (or action) is selected and executed during solving a problem. The first account holds that knowledge about how to solve a problem is used to decide the move and then this move is simply executed. A description of the problem acts as input to the problem solving process (here we term this a situation-only decision process). On this basis an action is selected and then executed. The second account suggests that problem solving occurs after the move has been selected as well as beforehand. In this account problem solving also occurs after an action is selected, immediately prior to and during the execution of said action. The input to this problem solving processes is not only a description of the problem but also a description of the intended action (we term this a situation-action decision process).

Problem solving after a move has been selected is rare in current psychological theories, for example this does not typically occur in ACT-r models (see Anderson, 2007). While theories of problem solving based solely on situationonly problem solving have the advantage of simplicity, they lack the power to use existing problem solving knowledge to evaluate ongoing or imminent actions. Theories featuring situation-action problem solving must necessarily feature some situation-only problem solving in order to derive an action for consideration. However, this situation-only problem solving does not necessarily have to be complex indeed an algorithm that simply picks an action at random
that hasn't been considered before might be sufficient. Some recent theories, particularly those exploring embodied problem solving, favour a combination of simple situationonly problem solving prior to action selection followed by more complex situation-action problem solving afterward (e.g. Miles, 2009, 2011; Schuboltz, 2007), the latter is often based on the mental simulation of the results of the action.

The differences between the two accounts are important because each implies a different form for knowledge about solving problems. Situation-action problem solving representations potentially could replace much of the need for situation-only problem solving representations. A key element of situation-action problem solving knowledge is that it typically either encourages or discourages an already selected action. By contrast situation-only problem solving knowledge is concerned with suggesting an action. Once situation-action problem solving knowledge is added to a theory of problem solving then there is less emphasis on the need to select the correct action first. An unsuitable action can be selected then rejected using situation-action knowledge. Indeed, a situation-action account suggests that people often cycle through a series of possible actions at any given stage of solving a problem, allowing situation-action knowledge to confirm or deny each action.

Hence the two accounts of problem solving are fundamentally different. The traditional model suggests a single decision point, followed by action. The situationaction position suggests a series of tentative decisions regarding possible actions followed by evidence gathering while that action is held in mind. It is notable that the second position is much more temporally scalable than the first. For example it is awkward to model speed-accuracytradeoffs in the first style of problem solving, while the ability to vary the amount of time spent gathering evidence is implicit in the second proposal.

\section*{Empirical Support for Situation-Action Knowledge}

Currently there is only indirect empirical support for the existence of situation-action knowledge. A necessary precondition of this knowledge is the ability to represent an action without necessarily executing that action.

Neuroscience has provided evidence of a representational role for parts of the brain associated with action, for example the premotor cortex (e.g. Decety et al., 1994). The existence of mirror neurons that are activated both by performing an action and observing an action (Gallese, Fadiga, Fogasse, \& Rizzolatti, 1996) suggests that the motor areas of the brain are involved in thinking as well
as doing. Over the last 15 years a large body of work has pointed to the conclusion that motor areas of the brain are used for representational roles as well as for executing actions (see Barsalou, 2009, p. 1285, for a brief review).

If representations in the motor areas of the brain are available to other areas of the brain, then logically these other areas will be able to make use of these representations when deciding what to do. It is exactly this logic that supports the existence of situation-action knowledge. Simply the problem solving parts of the brain are aware of the situation, they are also aware of the action that has been represented in the premotor cortex and related areas. It makes sense for the problem solving areas of the brain to make use of this knowledge regarding the conjunction of situation and action to either spur the motor areas into that action or pull them back from the brink of making an error.

The paper begins by reporting two Experiments that looked for evidence of continuing problem solving in the final stages of move selection in the Tower of London (variant) problem. We then show how these data can be computationally modeled using situation-action knowledge.

\section*{Experiment 1}

The paradigm used in both Experiments reported here works by biasing the selection of a first move in a given Tower of London problem. The bias occurs because participants have earlier solved a problem that is either superficially similar or the same as the target problem. In the repeat condition the bias supports the correct first move (of two possible moves), while in the false-analogy condition the bias supports the incorrect first move (again of two possible moves). These conditions are contrasted with problems where there is no bias.

\section*{Method}

Design: The Experiment was presented in two blocks, firstly of 3-disk problems then of 4-disk problems. Each block featured two training problems, a one minute pause, then two target problems. Each target problem had an inverse version, which although superficially similar has a different optimal first move. In each block one of the target problems was the same as one of the training problems (repeat condition) and one target problem was the inverse of the other training problem (false-analogy condition). Comparisons were planned between these conditions and the unbiased performance on the final training problem (novel condition).

Participants: Twenty-eight undergraduates participated in Experiment 1, each received either 30 minutes course credit or \(£ 2\).

Materials and apparatus: The ToL problems were presented on a desktop computer. Participants responded using a mouse. To move a disk the participants had to click on the disk they wanted to move and then click on the peg they wished to move it to. At the top of the screen the goal state was shown while the current state was interacted with in the main area of the screen. Disks were shown in
different colours and each was labeled with a different letter.

Procedure: At the beginning of each of the two blocks participants were first presented with an orientation task. The orientation task required six moves from a flat start state, with no goal displayed. Instructions were then displayed on the screen for a minimum of 30 seconds.

For 10 seconds prior to all training and target problems, the goal state for the problem was presented, on top of, and obscuring, the initial configuration for the problem. During this period a miniature representation of the start state was shown near the top of the screen, but all the disks in this representation were blocked grey, preventing participants from beginning a solution to the problem. This part of the procedure was designed to act as a cue to the related training problem. This display was then removed revealing the interface.

During the training phase problems the participant was only allowed the number of moves in the optimal solution to a problem. Once they had made this number of moves (and the goal state was not reached) a panel appeared (for 3 seconds) obscuring the problem, with "Try Again!" displayed in large letters. The problem was then reset to its original start state. This restriction was critical in ensuring all participants learned the same correct solution for each training problem (each problem had only one optimal solution path).

Timed lockouts were used between problems and between blocks. Participants were locked out for 30 seconds between all consecutive problems. The pause between training and target phases was one minute. There was also a minute lockout between blocks.

Prior to each target problem in a block a hint was given during the last 15 seconds of the lockout time. In the 3-disk block it was phrased as follows "The next problem will be the same as one you have already done." while in the 4-disk block it was "Note: You will have already solved the next problem". This change of phrasing was designed to increase the salience of the hint in the 4-disk block.

\section*{Results}

All latency data were log transformed for analysis; the raw data are summarised in Table 1. Comparisons were made between i) false-analogy condition and the repeat condition, ii) false-analogy condition and the final training problem (novel condition). The later comparison is subject to order effects, but the order effects (one would expect improved performance with practice) run counter to the predicted effects of condition (false-analogy < novel).

\section*{The 3-disk problems}

No significant effects of condition were found in measures of 3-disk problem solving. Most participants were still learning the basic methods needed to solve the ToL during this block, and this may have disrupted performance on the experimental conditions.

Table 1: Number of optimal first moves (from 28), first click latency (secs), and second click latency (secs) in Experiment 1
\(\left.\begin{array}{lccl}\hline \text { Condition } & \begin{array}{c}\text { No. of } \\
\text { Optimal 1 }\end{array} \\
\text { moves }\end{array} \quad \begin{array}{ccc}1^{\text {st }} \text { Click } \\
\text { latency (SD) }\end{array} \quad \begin{array}{c}2^{\text {nd }} \text { Click } \\
\text { Latency (SD) }\end{array}\right]\)\begin{tabular}{lcll} 
Repeat & 28 & \(5.18(1.66)\) & \(1.33(1.57)\) \\
Novel & 26 & \(6.22(1.66)\) & \(1.39(1.54)\) \\
False Analog. & 12 & \(7.35(1.83)\) & \(1.94(1.84)\) \\
\hline
\end{tabular}

\section*{The 4-disk problems}

Significantly more optimal moves were made under the repeat condition than the false-analogy condition, 28/28 vs. \(12 / 28, \mathrm{p}<.001\). A within-participant T-test found no significant differences between the false-analogy and repeat conditions on measures of the time taken to initiate the first move ( \(\mathrm{t}<1\) ). However participants took less time to complete a move in the repeat condition than they did in the false-analogy condition, \(\mathrm{t}(27)=5.86, \mathrm{p}<.001\).

In the novel condition participants succeeding 26 times on 28 first attempts at the final training problem, this compares to 12 times from 28 attempts of false-analogy target problems, \(\mathrm{p}<.001\). There was no significant difference between the novel condition and the false-analogy condition selection on time taken to initiate the first move by clicking a disk \((\mathrm{t}<1)\) but it took longer for participants to click on the location the disk was to go to in the false-analogy condition, \(\mathrm{t}(27)=2.99, \mathrm{p}<.01\).

Of the 28 participants, two made an error in the 'novel' condition. The remaining 26 were split into those that made an error on the subsequent false-analogy condition and those that did not; groups error \((\mathrm{N}=14)\) and correct \((\mathrm{N}=12)\). These data were analysed in \(2 \times 2\) mixed design ANOVA on disk destination click latency, with condition (false-analogy Vs. novel) as a within participant factor and error group (error Vs. correct) as a between participant factor. There were no interactions with, or main effects of group (all F \(<1\) ). This analysis suggests that hesitancy over the move being made was present both in those who did make the correct move those who didn't.

These results are consistent with the hypothesis that participants were engaged in problem solving during the final stages of completing the move. The hesitancy seen in the final stages of the move in the false-analogy condition is best explained as second thoughts about a move that has previously been decided upon. Despite these second thoughts, the original move is at least sometimes completed, but sometimes an alternative move is chosen (as indicated by the lack of differences in hesitancy between those who made errors and those that made the correct move in the false-analogy condition). This suggests that problem solving knowledge is being used after the first move has been decided upon and initiated.

Our theoretical account assumes that the participants have decided on the move they want to make prior to clicking the disk. Certainly the relative distribution of latency between
first click and second click supports this idea. However to demonstrate that the move has been decided upon prior to the first click, Experiment 2 reversed the order of actions needed to make the move, with the destination selected first and the disk selected second. In the second stage of the move only one disk (the top disk) could be selected, the decision about where to move it having already been made.

\section*{Experiment 2}

As well as changing the order in which the actions needed to complete a move were carried out, Experiment 2 attempted to improve on several elements of the design of Experiment 1. Crucially, only the comparison between the novel condition and the false-analogy condition was explored. It was felt that this comparison best captured the impact of the false-analogy manipulation.

Though the novel condition replaced the repeat condition, Experiment 2 used the same basic design as Experiment 1, with the exception that the two Experimental blocks now used 4 -disk and 5 -disk problems. Prior to this, participants completed a training block of 3-disk problems that facilitated the learning of the main principles of solving Tower of London problems. While the order of actions needed to move a disk were changed, other aspects of the interface remained unchanged.

\section*{Method}

Participants: Sixty-four undergraduates took part in the Experiment, each received 30 minutes credit toward their course requirement.

Apparatus: The apparatus and software was the same as it had been in Experiment 1. In all stages of the Experiment the method for moving the disks was altered. Now the participant had to click on the location they wanted the disk to go to. When this was done the peg they had pointed to was highlighted (turned from black to yellow). At this stage the participant then clicked on the disk they wanted to move to this peg. If the next click was not on a disk that could be legally moved to the highlighted peg then the highlighting on the chosen peg was removed, thus allowing participants to change their mind on the desired move.

A set of 5-disk problems was introduced for Experiment 2. It was reasoned that these would be sensitive to our Experimental manipulations in the same way as the 4 -disk problems were in Experiment 1 (and the 'too simple' 3-disk problems were not). Each problem again had an inverse counterpart that was used in the false-analogy condition. The use of different problems was balanced across the Experimental conditions for both 4-disk and 5-disk problems.

Procedure: Many elements of the procedure were the same as they were in Experiment 1, though the block structure of the experiment was altered. Initially participants completed a block of four 3-disk problems. Following this block the two Experimental blocks were presented. The first used 4-disk problems in the same basic structure as was used in Experiment 1 (orientation task - two training
problems - pause - orientation task - target problems). In each Experimental block one of the target problems was Novel and one a false-analogy to a training problem (order of conditions was counterbalanced). A 5-disk block followed the 4-disk block, using the same structure.

Table 2: Number of optimal first moves (from 128), first click latency (secs), and second click latency (secs) in Experiment 2
\begin{tabular}{lccl}
\hline Condition & \begin{tabular}{c} 
No. of \\
Optimal 1 \\
moves
\end{tabular} & \begin{tabular}{c}
\(1^{\text {st }}\) Click \\
latency (SD)
\end{tabular} & \begin{tabular}{c}
\(2^{\text {nd }}\) Click \\
Latency (SD)
\end{tabular} \\
\hline Novel & 104 & \(5.35(1.71)\) & \(1.17(1.66)\) \\
False analog. & 86 & \(4.73(1.77)\) & \(1.34(1.85)\) \\
\hline
\end{tabular}

\section*{Results and Discussion}

Descriptive data for Experiment 2 are shown in Table 2. We combined data from the two Experimental blocks with latency data log transformed. There were significantly fewer correct first moves in the false-analogy condition in comparison to the novel condition (proportionally .67 vs . .81 respectively), \(\mathrm{p}<.05\). Comparisons on latency measures were made using a data set reduced by two, as two of the data points in the 5 -disk block showed zero values for first-click latency (126 paired comparisons remained). This was due to participants clicking prior to the interface becoming active causing a zero to be recorded for first click latency. The effects on the counter-balancing of the Experiment were thought to be minimal. The expected simple effect, i.e. false-analogy slower, was found in the second stage latency, i.e. disk-selection, \(\mathrm{t}(125)=2.00, \mathrm{p}<\) .05. There was no significant difference in the time taken to initiate the move.

The data confirm that problem solving knowledge is being applied after a move has been decided upon in the Tower of London. We argue in the next section that these data and those from Experiment 1 are best accommodated by a cognitive architecture that primarily uses situationaction knowledge to solve problems.

\section*{Modeling Problem Solving Following Action Selection}

Problem solving knowledge has often been modeled in production system architectures, a tradition with its origins in Newell \& Simon's (1972) seminal book Human Problem Solving. Recently the ACT-r cognitive architecture has been used to produce production system accounts of problem solving. In traditional problem solving accounts, situation-only knowledge is represented in the following format: IF situation THEN action.

The model presented (TOL-GLAM) here is coded in the Glamorgan Problem Solver (GLAM-PS) architecture. This is notable because it doesn't use amodal representation and doesn't have a dedicated mechanism for processing goals (see Miles, 2011). TOL-GLAM is thus an example of an
embodied account of problem solving in the Tower of London, with emphasis placed on representation in the motor and perceptual systems used to complete the task. In terms of the representation of knowledge about solving the ToL, much of what TOL-GLAM knows is stored in the format: IF situation AND action THEN inhibit/activate action. This knowledge verifies the appropriateness of an already selected action, rather than specifying what action should be taken in a particular situation.

\section*{The TOL-GLAM Model}

In the GLAM-PS architecture there are modules dedicated to visual perception, ocular movement and motor actions. There are also modules dealing with other functions, for example auditory perception, speech production and bodily movement. Each module has its own production memory, working memory and production matching bottleneck.

Executive control within a GLAM-PS model emerges from the interaction of distributed subsystems (a similar idea was explored by Barnard, 1991). This control is based on each module's ability to see what is happening in all the other modules. So a production in the motor action module can match to working memory representations in other modules as well as working memory representations in the motor module itself.

\section*{Situation-Only Problem Solving in TOL-GLAM}

There are examples of situation-only and situation-action problem solving knowledge in TOL-GLAM. In the two Experiments the first move is restricted to two possibilities. The disk that is to be moved is always known (as it is on top of all the other disks in a tower configuration), the only question is the disks destination.

The situation-only algorithm used by TOL-GLAM begins by generating an action plan for moving a disk. This action plan is represented in the motor module, within a hierarchical structure. An example is given below (with only key attributes shown):
\begin{tabular}{|c|c|}
\hline \multicolumn{2}{|l|}{Action_plan1} \\
\hline Type & Action_plan \\
\hline First_element & disk_click1 \\
\hline Last_element & destination_click1 \\
\hline \multicolumn{2}{|l|}{Disk_click1} \\
\hline Type & click_on_object \\
\hline Location & diskA_location \\
\hline Super_element & Action_plan1 \\
\hline \multicolumn{2}{|l|}{Previous_element none} \\
\hline Next_element & destination_click1 \\
\hline \multicolumn{2}{|l|}{Destination_click1} \\
\hline Type & click_on_object \\
\hline Location & Peg2_location \\
\hline Super_element & Action plan1 \\
\hline \multicolumn{2}{|l|}{Previous_elementDisk_clickl} \\
\hline Next_element & none \\
\hline
\end{tabular}

This initial action plan will often not involve the movement of the top disk. Typically TOL-GLAM, will first represent the movement of the bottom disk in the tower to its goal location. This reflects a means-ends analysis of the ToL problem where the bottom disk is prioritized as the biggest difference between the start state and goal state. While the GLAM-PS architecture doesn't feature explicit goal representation, what is happening is that TOL-GLAM is effectively 'subgoaling' the bottom disk. The model then represents a move of the top disk (the first blocking disk) to the peg where the 'subgoaled' disk is not going (in order to remove the block).

An exception to this process occurs when TOL-GLAM recalls a previous problem that was apparently the same as the current problem. In this case the first move that was made successfully in the previous problem is used to determine where the top disk should go.

TOL-GLAM will now have a representation of a potential first move in its motor module. It is at this point that situation-action knowledge is used to determine the appropriateness of the action, either increasing its activation till it is executed, or blocking its execution.

\section*{Situation-Action Problem Solving in TOL-GLAM}

The situation-action algorithm used in TOL-GLAM is based upon forward search, and makes use of representational simulation of the results of the move that is being considered.

The process begins after a potential move of the top disk has been represented in the motor module. At this point productions in the visual module are able to match to this motor module representation and simulate the result of this action. In Miles (2011) visually simulated interim stages were utilized in a model of offline algebra problem solving, what the TOL-GLAM model does is very similar, essentially looking ahead to see what the results of the action that is being considered will be.

Simulation of the results of the move involves the creation of a projected representation of the disk being moved in its new location, and the inhibition of the representation of the disk in its current location. Once the move has been simulated in the visual module, the motor module is now able to consider the next move that will be made after the current one. The productions that do this match both to simulated and actual visual representations.

At this stage an action plan will be generated for the second move and typically, any conflict with the first move will often become apparent to TOL-GLAM. This is particularly the case if the first move blocks the ideal second move. A production looks for incompatibilities between the two moves being considered. On the other hand if the first move doesn't block the ideal second move then it is taken as providing evidence that the first move is a good one.

\section*{Executing an Action in TOL-GLAM}

Key features of the GLAM-PS architecture determine the process of action execution in TOL-GLAM. One of these is the Action-Execution Threshold (AET), a level of activation that must be reached before an action or action plan will be executed. The AET is an important element of GLAM-PS because it allows actions to be represented without necessarily being executed (Miles, 2009). Within the TOLGLAM model it allows a move to be represented and then evidence gathered about the suitability of the move. There is no limit in GLAM-PS to the number of productions that can match if those productions change an existing representations activation level. This means that in TOLGLAM the representation of a move can be simultaneously inhibited and activated by competing productions. It is the relative strength of the competing productions that will determine whether the action representation will continue to increase in activation until it surpasses the AET, or be inhibited.

\section*{Simulating the Results of Experiment 1 and 2 in TOL-GLAM}

To simulate the results of Experiments 1 and 2 TOLGLAM was setup with productions that represented knowledge gained from previous training problems, which were added to the productions that model normal problem solving in the ToL.

The additional productions, modeling knowledge from specific previous problems, trigger when TOL-GLAM is faced with a problem that has the same initial configuration and goal configuration as the previous problem in terms of number of disks on each peg in each configuration (so an exact match wasn't necessary). The identity of the first disk to be moved must also match. These encodings of solutions from previous problems result in the first move used in the previous problem first being represented in the motor module and subsequently quickly gaining activation.

The performance of the model was tested on simulation runs of the 84 problems from which data were taken for Experiment 1 ( 28 each in the repeat, novel and falseanalogy conditions) and the 256 problems from Experiment 2 (128 novel, 128 false-analogy).

In GLAM-PS each production has a strength value, this strength modifies the impact of the production - so a strong production will increase the activation of an action representation more than a weak one. The strength of the productions modeling knowledge from the previous problem was systematically varied through a single parameter beta, which multiplied the strength of productions activating a representation of the action used previously. The beta values used conformed to a Gaussian distribution, meaning that in some case the influence of previous problems was strong, but in others the influence was weaker. A second parameter theta modified the strength of productions that solved the problems, this was a free parameter with a single value across all simulations runs used to maximize the match between model and data.

Another free parameter was AET, the level of activation at which an action was executed in the motor module. Additional parameters included the time taken to initiate problem solving and the time taken to execute a mouse click.

A comparison of the models performance to the latency and accuracy data showed relatively strong fits, \(\chi^{2}=.72\) and \(\chi^{2}=.83\) respectively. The matches of model to data do show some differences with the model making fewer errors than participants did in Experiment 1, but more errors than seen in Experiment 2. The timing of the mouse clicks was simulated more closely, partially reflecting the availability of parameters that varied the timing of the models performance.

A key purpose of the model was to simulate the differences in latency seen in the false-analogy condition (as compared to Repeat and Novel problems). In this respect the model is very successful, showing the same differences as participants.

\section*{General Discussion}

Experiments 1 and 2 provide evidence of problem solving occurring after move-selection in the Tower of London (ToL). The TOL-GLAM model accounts for this data through a mechanism based around situation-action knowledge. This knowledge is encoded in TOL-GLAM as production rules that increase or decrease the activation of a particular motor action, depending on the apparent suitability of this action. The delays seen in move completion during the false-analogy condition in Experiment 1 and 2 are explained as TOL-GLAM having 'second thoughts' about the suitability of an already selected move. Our findings are similar to those of Walsh and Anderson (2009) who demonstrated how participants adaptively 'changed their minds' about the best strategy to solve a multiplication problem after a quick initial choice.

The way problem solving knowledge is structured in TOL-GLAM is noteworthy, relatively simple situation-only productions suggest a possible action, while more complex situation-action productions contain much of the knowledge that TOL-GLAM possess of how to solve ToL problems.

The notion that actions are selected and then evaluated is found in other theories. However in most existing theories this evaluation occurs in a single cycle of the system, and is only necessary if there is a conflict between two or more possible actions. For example in SOAR (Newell, 1990) preference rules are used for conflict resolution, while in ACT-r (Anderson, 2007) the relative utility of actions is considered. In TOL-GLAM, and more generally in the GLAM-PS architecture, the evaluation of an action is a protracted process, typically involving the evaluation of a single action, rather than multiple competing actions.

Although the current research focuses on a simple knowledge-lean domain (ToL), the issue explored is fundamental to understanding human thought. Current accounts (e.g. ACT-r, SOAR) appear to suggest that we think about situations, reason about them and only then
select an action. The suggestion here is that much of human thought begins with a possible action, and is followed by reasoning about the suitability of this action for the current situation.

\section*{References}

Anderson, J. R. (2007). How can the human mind occur in the physical universe? New York: Oxford University Press.
Barnard, P. J., \& Teasdale, J. D. (1991). Interacting cognitive subsystems: A systematic approach to cognitive-affective interaction and change. Cognition and Emotion, 5, 1-39.
Barsalou, L. W. (2009). Simulation, situated conceptualization, and prediction. Philosophical Transactions of the Royal Society B, 364, 1281-1289.
Decety, J., Perani, D., Jeannerod, M., Bettinardi, V., Tadary, B., Woods, R., Mazziotta, J. C., \& Fazio, F. (1994). Mapping motor representations with PET. Nature, 371, 600-602.
Gallesse, V., Fadiga, L., Fogassi, L., \& Rizzolatti, G. (1996). Action recognition in the premotor cortex. Brain, 119, 593-609.
Miles, G. E. (2009). Representing goals modally: A production system model of problem solving in the Tower of London. Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society, 469-475. Mahwah, NJ: LEA
Miles, G. E. (2011). 'Look no goals!': A sufficient model of simple algebra problem solving without explicit goal representation. Proceedings of the Thirty-Third Annual Conference of the Cognitive Science Society, 2691-2696. Mahwah, NJ: LEA
Newell, A. (1990). Unified theories of cognition. Boston: Harvard University Press.
Newell, A., \& Simon, H. A. (1972). Human problem solving. Englewood Cliffs, NJ: Prentice-Hall.
Schuboltz, R. I. (2007). Prediction of external events with our motor system: towards a new framework. Trends in Cognitive Sciences, 11, 211-218.
Walsh, M. M., \& Anderson, J. R. (2009). The strategic nature of changing your mind. Cognitive Psychology, 58. 416-440.

\title{
A Mechanistic Account of Computational Explanation in Cognitive Science
}

\author{
Marcin Miłkowski (mmilkows@ifispan.waw.pl) \\ Institute of Philosophy and Sociology, Polish Academy of Sciences \\ ul. Nowy Świat 72, 00-330 Warsaw, Poland
}

\begin{abstract}
Explanations in cognitive science rely predominantly on computational modeling. Though the scientific practice is systematic, and there is little doubt about the empirical value of numerous models, the methodological account of computational explanation is not up-to-date. The current paper offers a systematic account of computational explanation in cognitive science in a largely mechanistic framework. The account is illustrated with a short case study of modeling of the mirror neuron system in terms of predictive coding.
\end{abstract}

Keywords: computation; computational modeling; explanation; mechanism; levels; information-processing.

\section*{Importance of Computational Modeling}

Computational modeling plays a special role in contemporary cognitive science; over 80 percent of articles in theoretical journals focus on computational \({ }^{1}\) models (Busemeyer \& Diederich, 2010). The now dominating methodology forcefully defended by (Marr, 1982) has turned out to be fruitful. At the same time, the three-level account of Marr is not without problems. In particular, the relationship among the levels is interpreted in various ways, wherein the change of level is both the shift of grain and the shift of the boundary of the system under explanation (McClamrock, 1991); it is not at all clear what is the proper relation between competence and its realization or whether bottom-up modeling is entirely mistaken; and, last but least, whether one model should answer how, what and why questions related to the explanandum.

My goal in this paper is to offer a descriptive account, which is close in spirit to the recent developments in the theory of mechanistic explanation (Bechtel, 2008; Craver, 2007; Glennan, 2002; Machamer, Darden, \& Craver, 2000). According to mechanism, to explain a phenomenon is to explain the underlying mechanism. Mechanistic explanation is a species of causal explanation, and explaining a mechanism involves the discovery of its causal structure. While mechanisms are defined variously, the core idea is that they are organized systems, comprising causally relevant component parts and operations (or activities) thereof. Parts of the mechanism interact and their orchestrated operation contributes to the capacity of the mechanism. Mechanistic explanations abound in special sciences and it is hoped that the adequate description of the principles implied in explanations generally accepted as sound will furnish researchers also with normative guidance.

\footnotetext{
\({ }^{1}\) I am not using the word 'computational' here in the sense used by Marr to define one of the levels in his account.
}

The claim that computational explanation is best understood as mechanistic gains popularity (Piccinini, 2007), and I have defended it at length against skeptical doubt elsewhere (Miłkowski, 2013). Here, I wish to succinctly summarize the account and, more importantly, add some crucial detail to the overall mechanistic framework proposed earlier. I cannot discuss Marr's theory in detail here (but see (Miłkowski, 2013, pp. 114-121)) and it is used only for illustration purposes. My remarks below are not meant to imply a wholesale rejection of his largely successful methodology.

Marr's account did not involve any theory of how computation is physically realized, and it is compatible with a number of different accounts. I will assume a structural account of computational realization here, defended also by Piccinini (2008) and Chalmers (2011). For an extended argument, see also (Miłkowski, 2011, 2013).

One particular claim that is usually connected with the computational theory of mind is that the psychologically relevant computation is over mental representation, which leads to the language of thought hypothesis (Fodor, 1975). Here, no theory of mental representation is presupposed in the account of computation, one of the reasons being that representation is one of the most contentious issues in contemporary cognitive science. As the present account is intended to be descriptively adequate, assuming one particular theory of representation as implied by computation would make other accounts immediately noncomputational, which is absurd. Another reason is that mechanistic accounts of computation do not need to presuppose representation (Fresco, 2010; Piccinini, 2006), though they do not exclude the representational character of some of the information being processed. In other words, it is claimed that only the notion of information (in the information-theoretic sense, not in the semantic sense, which is controversial) is implied by the notion of computation (or information-processing).

\section*{Explanandum phenomenon}

Marr stressed the importance of specifying exactly what the model was supposed to explain. Specifying the explanandum phenomenon is critical also for the mechanistic framework, as several general norms of mechanistic explanation are related to the specification of the capacity of the mechanism. All mechanisms posited in explanations have an explanatory purpose, and for this reason, their specification is related to our epistemic interest. For the same reason, the boundaries of the mechanism, though not entirely arbitrary, can be carved in different ways depending on what one wishes to explain.

The explanandum phenomenon has to be described precisely in a mechanistic model; otherwise, the model's use and value will be unclear. The specification of the model is not to be confused with raw, unrefined observation or common-sense intuition about the capacity under consideration. The specification of the capacity may be (and usually is) improved during the modeling process, wherein the model allows to understand the capacity better. What the mechanistic model explains is the real mechanism, but how the explanandum phenomenon is delineated is decided in what was called "the model of data" in philosophy of science (Suppes, 1962). For example, models of language production usually presuppose that user's productivity is the phenomenon to be explained, even though it is impossible to empirically observe a language user producing an infinite set of sentences. If there are theoretical reasons to believe that language users have this capacity, it will be described in a model of data. In this respect, mechanistic explanation is in accord with Marr's plea for explicit specification of what is computed.

To some degree, the specification of the explanandum phenomenon corresponds to description of the cognitive competence (understood generically as the capacity of the mechanism). However, in contrast to traditional competence accounts, descriptions of the explanandum need not be idealized. Also, the competence is explained with realization, and its realization by underlying levels of the mechanism is explanatorily relevant. This stands in contrast to traditional symbolic cognitive science.

\section*{Explanatory focus and target}

In the context of computational modeling, which nowadays uses different computer simulations and embodied robots, it becomes clear that properties of a model are not limited to the ones related directly to the explanandum phenomenon. For example, a robotic model of cricket phonotaxis (Webb, 1995) has to include, for technical reasons, a circuit board even if there is nothing that corresponds to the board in the cricket. Such boards are ignored when evaluating the adequacy of the robotic explanation. I propose to distinguish the explanatory focus of the model from its target, which is the real robot. In particular, all embodied mechanistic models are complete with respect to the capacities of the target, while their explanatory focus may still include gaps: we may still not know how certain properties of the insect give rise to the explanandum phenomenon even if we have a robotic replica. The same goes for purely computational models that contain numerous ad hoc additions (Frijda, 1967; Lewandowsky, 1993). These additions are not parts of the explanatory focus.

Whenever the causal model of the explanatory focus of the mechanism is complete with respect to the explanandum phenomenon (note: not complete in an absolute sense), the model is a mechanistic how-actual explanation; if the model includes some black boxes, whose function is more or less well-defined, it is a mechanism schema; otherwise, it
remains a mechanism sketch. \({ }^{2}\) Note that even a grounded, embodied, robotic model of visual perception may still be a mechanism sketch with respect to human vision. Also, a model in which the explanatory focus is just a minor part of the mechanism, while the parts included in the target are predominant, violates the principle of parsimony.

\section*{Three levels of constitutive explanation}

Constitutive mechanistic explanation is the dominant form of computational explanations in cognitive science, and I will focus on it in what follows. This kind of explanation includes at least three levels of the mechanism: a constitutive ( -1 ) level, which is the lowest level in the given analysis; an isolated (0) level, at which the parts of the mechanism are specified along with their interactions (activities or operations); and the contextual (+1) level, at which the function of the mechanism is seen in a broader context (e.g., the context for cricket phonotaxis includes the dispersion of sound in the air). In contrast to how Marr (1982) or Dennett (1987) understand them, levels here are not just levels of abstraction; they are levels of composition. Hence, they are tightly integrated but not entirely reducible to the lowest level.

Computational models explain how the computational capacity of a mechanism is generated by the orchestrated operation of its component parts. To say that a mechanism implements a computation is to claim that the causal organization of the mechanism is such that the input and output information streams are causally linked and that this link, along with the specific structure of information processing, is completely described (for more on various mathematical notions of information, see Miłkowski (2013, chap. 2); note that I do not presuppose Church/Turing thesis). Importantly, the link might be cyclical and as complex as one could wish.

There are two ways in which computational models may correspond to mechanisms; first, they may be weakly equivalent to the explanandum phenomenon, in that they only describe the input and output information; or strongly equivalent, when they also correspond to the process that generates the output information. Note that these notions have been used in methodology of computer simulation since 1960s (Fodor, 1968, chap. 4). Only strongly equivalent models are explanatory according to the mechanistic framework.

\section*{Mechanistically adequate model of computation}

The description of a mechanistically adequate model of computation at the 0 level usually comprises two parts: (1) an abstract specification of a computation, which should include all the causally relevant variables; (2) a complete blueprint of the mechanism at this level of its organization. I will call the first part formal model of the mechanism and

\footnotetext{
\({ }^{2}\) These distinctions were used by Craver (2007), but were unrelated to the distinction between the target and the explanatory focus.
}
the second instantiation blueprint of the mechanism, for lack of a better term. While it should be clear that a formal model needs to be included, it is probably less evident why the instantiation blueprint is also part of the mechanistically adequate model. The causal model must include all causally relevant parts and operations without gaps or placeholder terms (think of generic and unspecific terms such as "representation" or "activation"). Yet formal models cannot function as complete causal models of computers. For example, to repair a broken old laptop, it is not enough to know that it was (idealizing somewhat) formally equivalent to a universal Turing machine. Similarly, how mental deficits will manifest themselves is not obvious based on a description of ideal cognitive capacity. One needs to know its implementation.

Hence, the mechanistic model of a computational phenomenon cannot be limited to its formal properties. Accordingly, merely formal models of, say, linguistic competence, which abstract away from its realization, are assessed as essentially incomplete. They are either mere specifications of the explanandum phenomenon, but not explanatory in themselves, or, when accompanied with a rough theory of how they are related to experimental data, mechanism sketches (Piccinini \& Craver, 2011). This means that computational explanations of psychological capacities need to be integrated, for completeness, with models of their realization. Otherwise, they may posit epiphenomenal entities without any causal relevance. Contrary to the functionalist theory of psychological computational explanation (Cummins, 1983), mechanism requires it to be causal. It follows that some symbolic models in psychology, even if they are weakly equivalent to the model of input/output data, are not considered to be fully explanatory because of the inherent danger of positing entities that are causally irrelevant.
Just because the usual description of the computational mechanism usually involves two different models, the formal one and the instantiation blueprint, and these may be idealized, computational modeling requires complex integration, similar to one described as multiple-models idealization (Weisberg, 2007).

Note that my mechanistic account of computation does not stipulate that there be a single formal model of computation that would fit all purposes. Rather, it adheres to transparent computationalism (Chrisley, 2000): any formal model that can be specified in terms of informationprocessing is fine here, be it digital, analog, or hybrid, as in contemporary computational neuroscience (Piccinini \& Bahar, 2012).

The empirical adequacy of the mechanistically adequate model of computation can be tested. As such models are strongly equivalent to processes being modeled, usual process-testing methods apply, including chronometry (Posner, 2005), various kinds of experimental and natural interventions (Craver, 2007), brain imaging - though with usual caveats (Trout, 2008), and task decomposition (Newell \& Simon, 1972). All in all, the more independent
observables are tested, the more robust the model. Note that the phenomenological validation modeled after the Turing test (Turing, 1950) is not taken to be evidence of the model's empirical adequacy.

\section*{Marr's cash register}

The account may be illustrated with the example used by Marr (1982, pp. 22-24): a cash register in a supermarket. The explanandum phenomenon is the capacity to add prices of individual items and determine the overall sum to be paid. At the contextual level, one describes the cash register as playing a certain role in the supermarket, by allowing easy calculation of the sum to be paid, and making the work of the cashier clerk easier. This includes a bar-code scanner, a conveyor belt, etc. At the isolated level, a dedicated computer using special software is described. The constraints mentioned by Marr, such as commutativity or associativity of addition, are included in the description of the software. Yet without describing the machine that can run the software, this level of description is incomplete. Various failures of the cash register (e.g., dimming of the display), can be explained not only in terms of the software bugs but also as hardware failures. Also, the particular display configuration, which can be related to user preferences at the contextual level, is usually not described fully in the software specification. It is the isolated level where one describes the physical machine that can display the product name for the cashier clerk and, more fundamentally, can run code by reading it from external memory (not all computers do so; a mechanical cash register, even if it performs computations, cannot run different software). The formal description, usually in terms of the programming language or diagrams, is put into correspondence with the machine. At the constitutive level, the operations of the electronic parts of the machine are explained by reference to their properties, relationships, and organization. Just because vast differences between different types of registers are possible (witness the differences between the self-checkout register and the ones used during the American Civil War), the exact explanations will differ. Also, self-checkout machines will have the capacity to collect cash automatically, which needs to be explained as well (the explanandum will be different), and so forth.
The purpose of this toy example is to show that the mechanistic explanation differs a bit from Marr's account by explicitly tightly integrating the levels. Also, at all levels one can ask the why-question: why is the design appropriate for the user? Why does the cash register appropriately display figures on the screen? Why does it save energy? The how-answer is specified at a lower level, and the lowest level depends on our epistemic interest. The what-question also concerns operation of all levels.

\section*{Case study: Predictive coding in mirror neurons}

To demonstrate what methodological guidance is offered by the mechanistic account of computational explanation, let
me briefly describe a recently proposed model of actionunderstanding in terms of predictive coding (Kilner, Friston, \& Frith, 2007). Predictive coding is one of the Bayesian frameworks and is gaining now considerable recognition (Clark, 2013). In the model, it is presupposed that this capacity is realized by the mirror-neuron system (MNS henceforth). \({ }^{3}\) The explanandum phenomenon, or action understanding, is described at four levels of hierarchy: (1) the intention-level, which includes long-term goals of actions; (2) the goal-level, which includes short-term goals necessary to realize (1); (3) the kinematic level, which is the shape of the movement of limbs in space and time; and (4) the muscle level, which is the pattern of muscle activity underlying the action (Hamilton \& Grafton, 2006). People have visual access only to (3) of other agents. Moreover, the same kinematic level information is correlated to different intentions: Mr. Hyde might hurt someone with a scalpel by making the same movements as Dr. Jekyll (Jacob \& Jeannerod, 2005). What needs to be explained, therefore, is how one understands actions, given ambiguous visual information; the constraint of the model is that such understanding is to be realized by MNS. Naturally, given relatively scarce evidence about the details of MNS, the model might be currently only biologically plausible. In mechanistic terms, it cannot be a how-actually model, as we lack observables that could confirm that causal factors in the model are actual. We may have only a how-plausible model (for more on this distinction, see (Craver, 2007)), which should ascribe a precise computational role for MNS.

Kilner, Friston \& Frith note that other similar explanations of action in terms of MNS posit forward or generative models. Yet these explanations cannot deal with the fact that people easily distinguish between the action of Dr. Jekyll and Mr. Hyde. In other words, they do not explain one important part of the phenomenon.

The contextual level of the proposed predictive coding mechanism includes the context in which the action is observed (e.g., the operation theatre vs. dark streets of London). The context of action, which is not coded by MNS, is hypothesized to be represented by other parts of the larger hierarchy, where intentions are encoded (Kilner et al., 2007, p. 164). Note that such hierarchy can be naturally accounted for in the mechanistic framework, while in the Marrian methodology, nested hierarchies of mechanisms are still analyzed merely on three levels, which are not levels of composition, as in Kilner et al.'s paper (this makes the analysis of the model in Marrian terms all the more difficult).

The 0 level of the mechanism is then described as performing predictive coding of action, i.e., the mechanism predicts the sensory consequences of movements, and the prediction error is minimized through recurrent or reciprocal

\footnotetext{
\({ }^{3}\) For my purposes, it is quite irrelevant whether this account of MNS is correct or not (but see (Lingnau, Gesierich, \& Caramazza, 2009)). I am merely interested in how the model is vindicated by its authors and how it should be evaluated from the mechanistic standpoint.
}
interactions among levels of a cortical hierarchy. This means that the mechanism posited by authors comprises more than just three levels, which is the minimal number for constitutive explanations. Here, the upper level mechanism employs a generative model to predict representations in the level below. Backward connections are used by the upper level to convey the prediction to the lower level, which is used to produce information about prediction error. The instantiation blueprint of the mechanism includes this hierarchy whose architecture allows adjusting the neural representations of actions in terms of sensory representation of causes of action if prediction error is found. The architecture is self-organizing, and the reciprocal exchange of signals continues until the error is finally minimized.

The formal model of the neural architecture is described here in terms of empirical Bayesian inference (Friston, 2002, 2003, 2005): the prior expectations are generated by the self-organizing information-processing architecture. In other words, this model includes, as usual, two complementary parts: the instantiation blueprint, characterized in terms what is known about MNS, and its formal computational specification. Quite obviously, contrary to the programmable cash register, no storedprogram computer is posited.

The constitutive level is merely touched upon; there is no extensive discussion of the precise realization of predictive coding by elementary entities of the neural system. Thus, this model is, at best, a mechanism schema, because it does not explain how MNS comes to operate as it does. The authors stress that to test the model, one would need to characterize the nodes of the cortical hierarchy anatomically and functionally, and such characterization is not available.

The neural plausibility of the predictive coding and its relation to empirical Bayesian modeling is the focus of much current discussion (Blokpoel, Kwisthout, \& Van Rooij, 2012). In particular, the question whether the biologically plausible implementation of the predictive coding is equivalent to empirical Bayes or not (it may somewhat approximate it). The mechanistic explanation requires that the mechanisms be not idealized in such a way that would require to ignore tractability questions (Van Rooij, 2008). The data in the original paper makes it impossible to answer critical questions about the mechanism in this context, such as the number of inputs in the Bayesian network, which is essential in assessing the parametrized complexity of the algorithm.

Were the model implemented on the computer, the results of the simulation could be compared to those observed in humans or in macaque monkeys. Alas, no such results are reported by Kilner et al., and since without implemented models detailed testing of hypotheses is impossible, the empirical adequacy of the explanation is not entirely clear. To assess the adequacy properly, one should rather implement several comparable models of the same explanandum phenomenon, which can also help in avoiding the confirmation bias to which researchers are prone (Farrell \& Lewandowsky, 2010; Miłkowski, 2013, p. 86).

Some Bayesian theories in psychology were recently criticized as fundamentalist, i.e., dogmatically trying to model behavior as rational and without mechanistic constraints (Jones \& Love, 2011). Note that this is not true of the model under consideration; Bayesian modeling in neuroscience is obviously related to functioning of the brain. Instead of stressing the contrast between the mechanistic account of computational explanation and Bayesian modeling, my intention is to show that the mechanistic framework can be used to evaluate the contribution of the given model to progress in understanding of the explanandum phenomenon.

Summing up this part of the discussion, the mechanistic framework makes it easy to assess the maturity of the model in terms of its completeness and empirical adequacy. Because the computer implementation is lacking, it is impossible to say whether the model contains a lot of empirically undecided decisions that are needed to make it run (hence focus/target evaluation is impossible). At the same time, there is no information about the constitutive level. On the contextual level, placeholder terms such as "intention encoding" are used and they need further explanations in other models. Thus, the model does not include a complete specification of the mechanism.

Also, it is not at all clear how long-term goals might be understood in terms of mere sensory input prediction. Dr. Jekyll's intention to heal a patient (long-term goal) does not seem, prima facie, to be represented just in sensory terms. If it is actually so represented, the model does not explain how. This makes it a mechanism sketch, so its explanatory value is, qualitatively speaking, on the par with traditional symbolic models of competence. (Quantitative evaluation is impossible here, as no results of experiments on computer implementation were reported.)

\section*{Conclusion}

The mechanistic account of computational explanation preserves the insights of Marr but is more flexible when applied to complex hierarchical systems. It may help integrate various different models in a single explanation. Mechanistic methodological principles are inferred from research practice in life sciences, neurosciences, and cognitive science. Also, by subsuming computational explanation under causal explanation, the mechanistic account is naturally complemented by methodology of causal explanation (Pearl, 2000; Spirtes, Glymour, \& Scheines, 2000; Woodward, 2003).

By allowing multiple nested hierarchies, the standard three-level constitutive explanation is naturally expanded when needed. There is also no danger in preferring only the contextual level in the explanation, as it does not furnish us with the constitutive causal factors. The constitutive level will also not obviate the need for the contextual level as it does not contain some of the entities which are found at the contextual level. For example, the encoding of intention is not realized by MNS only, so its explanation cannot be 'reduced' to the description of the lower levels.

The present theory is not intended to settle debates over matters in which modelers explicitly disagree; the only goal is to make as much sense of various modeling approaches as possible, and make cross-approach comparisons possible by showing the common ground between them.

It is also not presupposed that computational explanation is the only proper way to explain cognition (Miłkowski, 2012). On the contrary, only some part of the mechanism model is strictly computational (i.e., uses vocabulary of the theory of computation). The constitutive level of the mechanism has to be framed in non-computational terms; otherwise, the computational operations of the isolated level are not explained, and may turn out to be spurious (Miłkowski, 2013, pp. 82-3). At the same time, the present account leads naturally to explanatory pluralism, as the only requirement for the theoretical frameworks used to describe various levels of composition of mechanisms is that they include causally relevant factors.

\section*{Acknowledgments}

The author gratefully acknowledges the support of the Polish National Science Center Grant in the OPUS program no. 2011/03/B/HS1/04563.

\section*{References}

Bechtel, W. (2008). Mental Mechanisms. New York: Routledge (Taylor \& Francis Group).
Blokpoel, M., Kwisthout, J., \& Van Rooij, I. (2012). When
Can Predictive Brains be Truly Bayesian? Frontiers in Psychology, 3(November), 1-3.
doi:10.3389/fpsyg. 2012.00406
Busemeyer, J. R., \& Diederich, A. (2010). Cognitive modeling. Los Angeles: Sage.
Chalmers, D. J. (2011). A Computational Foundation for the Study of Cognition. Journal of Cognitive Science, (12), 325-359.
Chrisley, R. (2000). Transparent computationalism. In M. Scheutz (Ed.), New Computationalism: Conceptus-Studien 14 (pp. 105-121). Sankt Augustin: Academia Verlag.
Clark, A. (2013). Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science. Behavioral and Brain Sciences, (in print).
Craver, C. F. (2007). Explaining the Brain. Mechanisms and the mosaic unity of neuroscience. Oxford: Oxford University Press.
Cummins, R. (1983). The Nature of Psychological Explanation. Cambridge, Mass.: MIT Press.
Dennett, D. C. (1987). The Intentional Stance. Cambridge, Mass.: MIT Press.
Farrell, S., \& Lewandowsky, S. (2010). Computational Models as Aids to Better Reasoning in Psychology. Current Directions in Psychological Science, 19(5), 329335. doi:10.1177/0963721410386677

Fodor, J. A. (1968). Psychological explanation: an introduction to the philosophy of psychology. New York: Random House.

Fodor, J. A. (1975). The Language of Thought (1st ed.).
New York: Thomas Y. Crowell Company.
Fresco, N. (2010). Explaining Computation Without Semantics: Keeping it Simple. Minds and Machines, 20(2), 165-181. doi:10.1007/s11023-010-9199-6
Frijda, N. H. (1967). Problems of computer simulation. Behavioral Science, 12(1), 59-67. doi:10.1002/bs. 3830120109
Friston, K. (2002). Functional integration and inference in the brain. Progress in neurobiology, 68(2), 113-43.
Friston, K. (2003). Learning and inference in the brain. Neural networks \(\square\) : the official journal of the International Neural Network Society, 16(9), 1325-52. doi:10.1016/j.neunet.2003.06.005
Friston, K. (2005). A theory of cortical responses. Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 360(1456), 815-36. doi:10.1098/rstb.2005.1622
Glennan, S. (2002). Rethinking Mechanistic Explanation. Philosophy of Science, 69(S3), S342-S353. doi:10.1086/341857
Hamilton, A. F. de C., \& Grafton, S. T. (2006). Goal representation in human anterior intraparietal sulcus. The Journal of neuroscience \(\square\) : the official journal of the Society for Neuroscience, 26(4), 1133-7. doi:10.1523/JNEUROSCI.4551-05.2006
Jacob, P., \& Jeannerod, M. (2005). The motor theory of social cognition: a critique. Trends in Cognitive Sciences, 9(1).
Jones, M., \& Love, B. C. (2011). Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. Behavioral and Brain Sciences, 34(04), 169-188. doi:10.1017/S0140525X10003134
Kilner, J. M., Friston, K. J., \& Frith, C. D. (2007). Predictive coding: an account of the mirror neuron system. Cognitive processing, 8(3), 159-66. doi:10.1007/s10339-007-0170-2
Lewandowsky, S. (1993). THE REWARDS AND HAZARDS OF COMPUTER SIMULATIONS. Psychological Science, 4(4), 236-243. doi:10.1111/j.1467-9280.1993.tb00267.x
Lingnau, A., Gesierich, B., \& Caramazza, A. (2009). Asymmetric fMRI adaptation reveals no evidence for mirror neurons in humans. Proceedings of the National Academy of Sciences of the United States of America, 106(24), 9925-30. doi:10.1073/pnas. 0902262106
Machamer, P., Darden, L., \& Craver, C. F. (2000). Thinking about Mechanisms. Philosophy of Science, 67(1), 1-25.
Marr, D. (1982). Vision. A Computational Investigation into the Human Representation and Processing of Visual Information. New York: W. H. Freeman and Company.
McClamrock, R. (1991). Marr's three levels: A reevaluation. Minds and Machines, 1(2), 185-196. doi:10.1007/BF00361036
Miłkowski, M. (2011). Beyond Formal Structure: A Mechanistic Perspective on Computation and

Implementation. Journal of Cognitive Science, 12(4), 359-379.
Miłkowski, M. (2012). Limits of Computational Explanation of Cognition. In V. C. Müller (Ed.), Philosophy and Theory of Artificial Intelligence (pp. 6984). Berlin - Heidelberg: Springer. doi:10.1007/978-3-642-31674-6_6
Miłkowski, M. (2013). Explaning the Computational Mind. Cambridge, Mass.: MIT Press.
Newell, A., \& Simon, H. A. (1972). Human problem solving. Englewood Cliffs, NJ: Prentice-Hall.
Pearl, J. (2000). Causality: models, reasoning, and inference. Cambridge: Cambridge University Press.
Piccinini, G. (2006). Computation without Representation. Philosophical Studies, 137(2), 205-241. doi:10.1007/s11098-005-5385-4
Piccinini, G. (2007). Computing Mechanisms. Philosophy of Science, 74(4), 501-526. doi:10.1086/522851
Piccinini, G. (2008). Computers. Pacific Philosophical Quarterly, 89(1), 32-73. doi:10.1111/j.14680114.2008.00309.x

Piccinini, G., \& Bahar, S. (2012). Neural Computation and the Computational Theory of Cognition. Cognitive Science, n/a-n/a. doi:10.1111/cogs. 12012
Piccinini, G., \& Craver, C. (2011). Integrating psychology and neuroscience: functional analyses as mechanism sketches. Synthese, 183(3), 283-311. doi:10.1007/s11229-011-9898-4
Posner, M. I. (2005). Timing the brain: mental chronometry as a tool in neuroscience. PLoS biology, 3(2), e51. doi:10.1371/journal.pbio. 0030051
Spirtes, P., Glymour, C. N., \& Scheines, R. (2000). Causation, prediction, and search (2nd ed.). Cambridge, Mass.: The MIT Press.
Suppes, P. (1962). Models of Data. In E. Nagel, P. Suppes, \& A. Tarski (Eds.), Logic, Methodology, and Philosophy of Science: Proceedings of the 1960 International Congress (pp. 252-261). Stanford: Stanford University Press.
Trout, J. D. (2008). Seduction without cause: uncovering explanatory neurophilia. Trends in cognitive sciences, 12(8), 281-2. doi:10.1016/j.tics.2008.05.004
Turing, A. (1950). Computing Machinery and Intelligence. Mind, LIX(236), 433-460. doi:10.1093/mind/LIX.236.433
Van Rooij, I. (2008). The tractable cognition thesis. Cognitive science, 32(6), 939-84.
doi:10.1080/03640210801897856
Webb, B. (1995). Using robots to model animals: a cricket test. Robotics and Autonomous Systems, 16(2-4), 117134. doi:10.1016/0921-8890(95)00044-5

Weisberg, M. (2007). Three kinds of idealization. Journal of Philosophy, 104(12), 639-659.
Woodward, J. (2003). Making Things Happen. Oxford: Oxford University Press.

\title{
Distinguishing between Category-based and Similarity-based Induction
}

\author{
Tracey Miser (miser.4@osu.edu) \\ Department of Psychology, 1835 Neil Avenue \\ Columbus, OH43210 USA
}

Vladimir Sloutsky (sloutsky.1@osu.edu)
Department of Psychology, 1835 Neil Avenue
Columbus, OH43210 USA

\begin{abstract}
Performing inductive generalizations is critical for learning, yet there is much debate regarding the mechanisms underlying this ability. One view posits that similarity-based induction, utilizing perceptual features, may allow for increased encoding and higher memory accuracy on recognition tests. While category-based induction, utilizing semantic information, may result in limiting encoding of perceptual detail, thus resulting in decreased memory accuracy. In Experiment 1, we attempted to impair spontaneous categorization by presenting a second Working Memory load task. In Experiment 2, we attempted to impair perceptual processing by introducing a second Visual Search task. Results indicate that adult participants can rely on either mechanism when performing induction.
\end{abstract}

Keywords: Induction; Learning; Memory.

\section*{Introduction}

The ability to generalize from the known to novel is a critical aspect of cognition - this ability allows expanding knowledge to new situations. At the same time, the learner may not know how far new knowledge can be expanded outside of the learning situation. Suppose that one learned that adenosine promotes myelination in the brain of the Capuchin monkey. Should this knowledge be generalized to New World monkeys, all monkeys, all primates, or all mammals? One way of generalizing knowledge is by identifying a common category that licenses such generalization. For example, one may decide that MONKEY is such a category and generalize knowledge to all monkeys. However, while knowledge of categories is useful, it is not necessary for inductive generalization. For example, one may decide that animals share a property to the extent their similarity exceeds some criterial value.

The latter mechanism seems to be a good candidate for generalization early in development, whereas the former could be a product of development. However, despite the fact that inductive generalization exhibits early onset (Baldwin, Markman, \& Melartin, 1993; Gelman \& Markman, 1986; Sloutsky, Lo, \& Fisher, 2001; Welder \& Graham, 2001), the mechanisms underlying early induction are hotly debated.

According to the naïve theory approach (see Murphy, 2002, for a review) induction is a two-step process: children first identify encountered entities as members of categories, and, if entities belong to the same category (say, the same natural kind), then infer that these entities share many properties. The inference is licensed by children's assumptions that members of some categories (such as, for example, natural kinds) share many properties. Given that children are more likely to know basic-level categories (e.g., MONKEY) than superordinate categories (e.g., MAMMAL), they are more likely to generalize properties within basic-level categories.

According to another position (i.e., the similarity view), induction is a generalization process, and young children generalize on the basis of multiple commonalities, or similarities, among presented entities (e.g., Jones \& Smith, 2002; McClelland \& Rogers, 2003; Sloutsky, Fisher, \& Lo, 2001; Sloutsky, 2003, Sloutsky \& Fisher, 2004a, 2004b). This view does not attribute conceptual assumptions to young children.

In an attempt to address these issues, Sloutsky and Fisher (2004b; Fisher \& Sloutsky, 2005) introduced Induction-then-Recognition (ITR) paradigm. The idea is based on the following reasoning. There is a well-known "level-of-processing effect" - deeper semantic processing facilitates correct recognition of presented items, increasing the proportion of "hits" (Craik \& Lockhart, 1972; Craik \& Tulving, 1975). At the same time, deeper processing also results in higher levels of memory intrusions - false recognition of non-presented "critical lures" of semantically associated or categorically related items (e.g., Koutstaal \& Schacter, 1997; Rhodes \& Anastasi, 2000; Thapar \& McDermott, 2001). Due to elevated levels of false alarms, the net result of deep semantic processing on recognition accuracy (i.e., Hits - False Alarms) is negative. At the same time, it is known that focusing on perceptual details of pictorially presented information leads to more accurate recognition (Marks, 1991) - although hits might be slightly lower, false alarms are significantly lower than under deep semantic processing. Therefore, these memory findings suggest that categorization (which is a variant of deeper semantic processing) would result in a higher level of memory intrusions and thus in lower recognition accuracy
than shallow perceptual processing (see also Brainerd, Reyna, \& Forrest, 2002, for related arguments).

Thus, a memory test administered after an induction task may reveal differential encoding of information during induction: if participants perform category-based induction, they should be engaged in deep semantic processing, and therefore exhibit low discrimination of studied items from critical lures during a memory test (compared to a noinduction baseline condition). On the other hand, if participants perform similarity-based induction, they should be engaged in shallow perceptual processing, and as a result their memory accuracy should not decrease compared to the baseline. Because, unlike adults, young children were expected to perform similarity-based induction, this reasoning led to a nontrivial prediction that after performing induction, young children may exhibit greater memory accuracy (i.e., have fewer false alarms) than adults.

These predictions have received empirical support: the pattern of results reported by Sloutsky and Fisher (2004a; 2004b; Fisher \& Sloutsky, 2005) indicated that while adults perform category-based induction, young children perform similarity-based induction. In particular, after performing inductive generalizations about members of familiar animal categories (i.e., cats, bears, and birds), adults’ memory accuracy attenuated markedly compared to the no-induction baseline, and, these effects of induction were robust across a wide range of animal categories (Fisher \& Sloutsky, 2004). At the same time, young children were accurate in both the baseline and induction conditions, exhibiting greater accuracy in the induction condition than adults.

Although these findings are compatible with the idea of different mechanisms of induction across development (i.e., similarity-based early induction and category-based mature induction), a number of alternative explanations have been proposed. In particular, Wilburn and Feeney (2008) and Hayes, McKinnon, \& Sweller (2008) suggested that the mechanism of induction does not change across development (with induction being category-based) and the higher memory accuracy of children simply reflects their inability to filter out irrelevant perceptual information. In other words, whereas adults process primarily category information, young children cannot focus efficiently, and, as a result, they process both category and perceptual information. Although there are several phenomena that this idea cannot explain (see Sloutsky, 2008), we deemed it necessary to address the issue directly.

To do so, we created a new paradigm to examine the issue. The underlying idea is to selectively impair either categorical or perceptual processing and to examine induction and memory performance. If participants can rely on either information (which we believe is the case with adults), then neither manipulation should have an effect on induction. If participants rely primarily on perceptual information (which we believe is the case with children), then impairing perceptual processing should impair induction.

Each manipulation should also have a different effect on memory. Impairing categorical processing should force participants to process items perceptually, thus potentially increasing memory accuracy after performing induction. At the same time, impairing perceptual processing should force participants to process items categorically, thus potentially decreasing memory accuracy.

In research reported here, we tested this paradigm with adults. The main idea is to introduce a second task when participants perform induction. To impair categorical processing, we introduce a working memory task, whereas to impair perceptual processing, we introduce a visual search task.

In what follows, we report two experiments: In Experiment 1, the second task is a working memory task, whereas in Experiment 2, the second task is a visual search task. We compare performance on these experiments with performance reported by Sloutsky and Fisher (2004), when no secondary tasks were introduced.

\section*{Experiment 1: Induction with Working Memory Load}

The experiment was a replication of Sloutsky and Fisher (2004b) ITR paradigm with one difference: during the study phase participants were presented with a second task, whose goal was to increase working memory load.

\section*{Method}

Participants. Sixty-two introductory psychology students participated in the experiment for class credit. Twenty-six participants were excluded due to low accuracy on check trials in the recognition portion of the experiment.

Materials, Design and Procedure. Visual Stimuli consisted of 44 color photographs of animals on white backgrounds (see Figure 1 for examples). Auditory Stimuli consisted of ten familiar words (e.g., one, two, three, four, five, six, seven, eight, nine, ten) presented through headphones between 68-72 dB.

Similar to Sloutsky and Fisher (2004b), the experiment included two between-subjects conditions: Memory and Induction. In both conditions, the experiment was divided into two phases: the study phase and the recognition phase.

During the study phase of both conditions, participants received a working memory (WM) task. For the WM task, participants were initially presented with five randomly selected Auditory Stimuli and asked to listen for one of the words to be played more than one time on each of the subsequent study phase trials. At the end of each trial participants were asked if one of the words had been repeated and were provided with Yes/No feedback.

The primary task of interest differed across the conditions: in the Induction condition participants were asked to generalize properties and in the Memory condition, they were asked to remember the items as accurately as possible.

Study Phase: Induction Condition. During the study phase, participants were presented with 30 pictures of animals, one at a time, in a random order. The animals were selected from 3 categories: 10 bears, 10 birds, and 10 cats. The pictures were presented centrally on a 22 " wide screen monitor for 2750 ms each. After being introduced to the WM task, participants were then shown a picture of a cat and were told the cat had "beta cells in its blood." Throughout the study phase of the Induction condition, participants were first asked after each trial whether one of the words had been repeated and Yes/No feedback was provided. They were then asked to decide whether each presented animal also had beta cells. Yes/No feedback was provided indicating that only cats had beta cells.
Study Phase: Memory Condition. The Memory condition was similar to the Induction condition, with a single difference: instead of performing an induction task, participants were asked to remember the items as accurately as possible. They were also warned about the upcoming memory test.
Recognition Phase. The recognition phase was identical across both conditions. The recognition phase immediately followed the study phase. During recognition, participants were presented with 28 images, 14 of which had been presented in the study phase and 14 of which were new images. Participants were instructed to determine whether each image had been presented during the study phase and neither feedback nor secondary task was given.


Figure 1: Examples of Visual Stimuli

\section*{Results and Discussion}

In this experiment, it was expected that spontaneous categorization would be hindered due to increased working memory load. Therefore, compared to a single task condition in Sloutsky and Fisher (2004b), the dual task condition may increase the overall task difficulty thus attenuating recognition accuracy in the Memory condition. At the same time, it may block categorization, thus increasing recognition accuracy in the Induction condition. The average rate of correct induction was over \(98 \%\), compared to over 75\% induction accuracy in Sloutsky and Fisher (2004b).

To analyze recognition memory accuracy, Hit and False Alarm (FA) rates were calculated (see Table 1). Also in the Table are Hit and FA rates from Sloutsky and Fisher (2004b). Because these researchers did not use a secondary WM task, we will refer to their experiment as "Baseline".

To further examine the ability to discriminate old items from critical lures, we computed memory sensitivity \(A^{\prime}\) scores. \(A^{\prime}\) is a nonparametric analogue of the signal detection statistic d' (Snodgrass \& Corwin, 1988; Wickens, 2002). If participants do not discriminate old items from critical lures, \(A^{\prime}\) is at or below .5. The greater the discrimination accuracy, the closer \(A^{\prime}\) is to \(1 . A^{\prime}\) scores for Experiment 1 are presented in Figure 2 alongside \(A^{\prime}\) scores for Experiment 2, as well as the results of the Sloutsky and Fisher (2004b) Baseline data.

Data in the figure were submitted to a 2 (Experiment: Working Memory vs. Baseline) by 2 (Condition: Induction vs. Memory) ANOVA. The analysis revealed a significant interaction between experiment and condition, \(F(1,80)=\) 5.58, \(p=.02\) as well as a significant main effect of condition, \(F(1,80)=14.38, p<.000\). Independent samples t-tests indicated that memory accuracy in the Memory condition of the current experiment was lower than that in Sloutsky and Fisher (2004b), \(t\) (35) \(=-2.23, p<.05\). At the same time the opposite was true for the Induction condition, in which the WM load of the current experiment resulted in marginally higher memory accuracy than a single task induction accuracy reported in Sloutsky and Fisher (2004b), \(t(45)=1.54, p=.13\).

Table 1
Mean Proportions of Hits and False Alarms (FA) and Mean Accuracy
\begin{tabular}{lccc}
\hline \hline Condition & Hits & FA & \begin{tabular}{c} 
Accuracy \\
(hits-FA)
\end{tabular} \\
\hline WM-Ind & \(.78(.16)\) & \(.59(.26)\) & .19 \\
WM-Mem & \(.77(.14)\) & \(.52(.22)\) & .24 \\
\hline *S\&F-Ind & \(.83(.20)\) & \(.76(.25)\) & .07 \\
*S\&F-Mem & \(.89(.10)\) & \(.47(.31)\) & .42 \\
\hline \hline
\end{tabular}

Note. WM - working memory; *S\&F - Sloutsky \& Fisher (2004b); Standard deviations are in parentheses.

Taken together results of Experiment 1 indicate that impairing categorization by introducing a secondary WM task does not affect induction accuracy, but it does affect memory accuracy. Most importantly, memory accuracy in the Induction condition increased somewhat, compared to the memory accuracy in the single-task Induction, which was not the case for the Memory condition. These results confirm that if categorization is impaired, adults can rely on perceptual information to perform induction. In Experiment 2, we attempted to impair participants' perceptual processing.

\section*{Experiment 2: Induction with Perceptual Load}

The experiment was similar to Experiment 1, with one critical difference: the second task was a visual search task, whose goal was to impair perceptual processing rather than categorization.

\section*{Method}

Participants. Fifty-six introductory psychology students participated in the experiment for class credit. Thirty participants were excluded due to low accuracy on check trials in the recognition portion of the experiment.

Materials, Design and Procedure. Visual Stimuli consisted of the same 44 color photographs used in Experiment 1. Visual Search Stimuli consisted of a total of 16 red or black " + " and " 0 " symbols. These stimuli were presented in random sequence by Rapid Serial Visual Presentation, with items being displayed for 250 ms each and having an inter stimulus interval of 250 ms . Visual Search stimuli were presented in the upper right hand corner of the screen with eccentricity of approximately \(23^{\circ}\) visual angle and subtending approximately \(1.4^{\circ}\) of visual angle. Experiments were conducted on a Dell Optiplex 790 computer and were programmed in E-Prime Professional 2.0 software.

Similar to Experiment 1, this experiment included two between-subjects conditions: Memory and Induction. Also, similar to Experiment 1, the two conditions differed only in the Study phase, while having identical Recognition phase.
The Study phase of each condition was similar to the respective condition of Experiment 1, with a single difference. The second task in Experiment 2 was a Visual Search (VS) task.

During the study phase of both conditions, participants were presented with Visual Search stimuli in the upper right corner of the monitor and asked to watch for red " + " signs on each of the subsequent study phase trials. The Visual Search stimuli preceded the onset of animal pictures by 3000 ms and continued 2000 ms after the picture of the animal disappeared. The study phase consisted of the same 30 pictures of animals as in Experiment 1 and were presented centrally on a 22 " wide screen monitor for 2750 ms each. After each study phase trial, participants were first asked whether a red "+" sign had been presented. In both
conditions, participants were instructed to not look directly at the animal pictures. Participants' eye gaze was monitored by an experimenter and verbal corrective feedback was provided. Immediately following the last Visual Search stimuli on each trial, participants were asked whether they had seen any red "+" signs and Yes/No feedback was provided.
Study Phase: Induction Condition. During the Induction Condition, participants were first asked whether they had seen any red "+" signs and were provided with Yes/No feedback. They were then asked whether the animal had beta cells and were given Yes/No feedback indicating that only cats have beta cells.
Study Phase: Memory Condition. The Memory condition was similar to the Induction condition, with a single difference: instead of performing an induction task, participants were asked to remember the items as accurately as possible.
Recognition Phase. The recognition phase was similar to that in Experiment 1: in the Memory condition participants were told in advance about the upcoming recognition phase, whereas in the Induction condition, no advanced warnings about upcoming recognition were given. The recognition phase immediately followed the study phase. During recognition, participants were presented with the same 28 images that were presented in Experiment 1, 14 images had been presented in the study phase and 14 were new images. Participants were instructed to determine whether each image had been presented during the study phase and neither feedback nor secondary task was given.

\section*{Results and Discussion}

In this experiment, it was expected that perceptual processing would be impaired due to the demands of the Visual Search task. The average rate of correct induction was over \(94 \%\), compared to over \(75 \%\) induction accuracy in Sloutsky and Fisher (2004b).

Hit and false alarm rates are presented in Table 2 and \(A^{\prime}\) scores for Experiment 2 are presented in Figure 2 alongside \(A^{\prime}\) scores for Experiment 1, as well as the results of the Sloutsky and Fisher (2004b) Baseline data. Hit and false alarm percentages for Experiments 1 and 2, as well as the Sloutsky and Fisher Baseline percentages are presented by condition in Table 3.

Table 2
Mean Proportions of Hits and False Alarms (FA) and Mean Accuracy
\begin{tabular}{lccc}
\hline \hline Condition & Hits & FA & \begin{tabular}{c} 
Accuracy \\
(hits-FA)
\end{tabular} \\
\hline VS-Ind & \(.71(.20)\) & \(.59(.28)\) & .12 \\
VS-Mem & \(.58(.22)\) & \(.39(.23)\) & .18 \\
\hline *S\&F-Ind & \(.83(.20)\) & \(.76(.25)\) & .07 \\
*S\&F-Mem & \(.89(.10)\) & \(.47(.31)\) & .42 \\
\hline \hline
\end{tabular}

Note. VS - Visual Search; *S\&F - Sloutsky \& Fisher (2004b); Standard deviations are in parentheses.
\(A^{\prime}\) scores shown in Figure 2 were submitted to a 2 (Experiment: Visual Search vs. Baseline) by 2 (Condition: Induction vs. Memory) ANOVA. The analysis revealed a significant interaction, \(F(1,70)=4.65, p<.05\), as well as a main effect for condition, \(F(1,70)=10.08, p=.002\). Furthermore, two tailed independent samples t-tests indicated a significant decrease in memory accuracy in the Memory condition of Experiment 2 compared to Sloutsky and Fisher (2004b), \(t(29)=-2.38, p<.05\); but not in the Induction condition \(t(41)=.60, p=.55\).


Figure 2. Memory Sensitivity scores (A') across experimental conditions. The dashed line represents the point of no sensitivity. Error bars show standard errors of the mean.

Table 3
Mean Proportions of Hits and False Alarms (FA) and Mean Accuracy
\begin{tabular}{lccc}
\hline \hline Condition & Hits & FA & \begin{tabular}{c} 
Accuracy \\
(hits-FA)
\end{tabular} \\
\hline WM-Ind & \(.78(.16)\) & \(.59(.26)\) & .19 \\
WM-Mem & \(.77(.14)\) & \(.52(.22)\) & .24 \\
VS-Ind & \(.71(.20)\) & \(.59(.28)\) & .12 \\
VS-Mem & \(.58(.22)\) & \(.39(.23)\) & .18 \\
*S\&F-Ind & \(.83(.20)\) & \(.76(.25)\) & .07 \\
*S\&F-Mem & \(.89(.10)\) & \(.47(.31)\) & .42 \\
\hline \hline
\end{tabular}

Note. Standard deviations are in parentheses.
*Indicates data from Sloutsky \& Fisher (2004).

Overall, results of Experiment 2 indicate that impairing perceptual processing does not impair inductive inference in adults, while significantly impairing recognition accuracy in the memory condition.

\section*{General Discussion}

The two reported experiments introduce and test a new paradigm for studying the mechanism of induction. Experiment 1, attempts to impair semantic categorization by introducing a secondary Working Memory task, while Experiment 2 attempts to impair perceptual processing by introducing a secondary Visual Search task. Results indicate that whereas participants were able to perform inductive inference in both conditions, each manipulation somewhat differently affected recognition accuracy in the Memory and Induction conditions.

First, both tasks impaired recognition accuracy in the Memory condition compared to a single task Baseline, perhaps more so in the Visual Search than in the WM condition. Note that when Visual Search was the secondary task, recognition memory in the Induction condition (similar to the Baseline) was not different from 0.5 ( \(p>.12\) ), which indicates no discrimination between old items and critical lures. At the same time, when working memory was the secondary task, recognition memory in the Induction condition was above \(0.5(p<.05)\) In addition, the WM task (whose goal was to block semantic categorization) increased somewhat memory accuracy in the Induction condition.

The reported results support the idea that adults may perform inductive inference by relying on either conceptual or perceptual information. In future research, we plan to present these tasks to children. If mechanisms of induction in children are equivalent to those of adults, then, similar to adults, children should be able to perform induction in either condition (although their memory accuracy may attenuate due to increased task demands). In contrast, if children rely on perceptual (but not conceptual) processing when performing induction, their induction performance should drop in the Visual Search, but not in the WM condition. We believe that the new paradigm presented here can address these issues.

\section*{Acknowledgments}

This research was supported by a grant from the NSF (BCS0720135) to V. M. Sloutsky.

\section*{References}

Baldwin, D. A., Markman, E. M.,\& Melartin, R. L. (1993). Infants' ability to draw inferences about nonobvious object properties: Evidence from exploratory play. Child Development, 64, 711-728.
Brainerd, C. J., Reyna, V. F., \& Forrest, T. J. (2002). Are young children susceptible to the false-memory illusion? Child Development, 73, 1363-1377.
Craik, F. I. \& Lockhart, R. S. (1972). Levels of processing: A framework for memory research. Journal of Verbal Learning and Verbal Behavior, 11, 671-684.
Craik F. I. \& Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. Journal of Experimental Psychology: General, 104, 268-294.
Fisher, A. V., \& Sloutsky, V. M. (2005). When induction meets memory: Evidence for gradual transition from
similarity-based to category-based induction. Child Development, 76, 583-597.
Gelman, S. A. (1988). The development of induction within natural kind and artifact categories. Cognitive Psychology, 20, 65-95.
Gelman, S. A., \& Coley, J. (1991). Language and categorization: The acquisition of natural kind terms. In S. A. Gelman \& J. P. Byrnes (Eds.), Perspectives on language and thought: Interrelations in development. New York: Cambridge University Press.
Gelman, S. A., \& Markman, E. (1986). Categories and induction in young children. Cognition, 23, 183-209.
Hayes B. K., McKinnon, R., \& Sweller, N. (2008). The Development of category-based induction: Reexamining conclusions from the induction then recognition (ITR) paradigm. Developmental Psychology, 44, 1430-1441.
Jones, S. S., \& Smith, L. B. (2002). How children know the relevant properties for generalizing object names. Developmental Science, 5, 219-232.
Koutstaal, W., \& Schacter, D. L. (1997). Gist-based false recognition of pictures in older and younger adults. Journal of Memory and Language, 37, 555-583.
Marks, W. (1991). Effects of encoding the perceptual features of pictures on memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 17, 566577.

McClelland, J. L. \& Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. Nature Reviews Neuroscience, 4, 310-322.
Murphy, G. L. (2002). The big book of concepts. Cambridge, MA: MIT Press.
Rhodes, M. G., \& Anastasi, J. S. (2000). The effects of a levels-of-processing manipulation on false recall. Psychonomic Bulletin \& Review, 7, 158-162.

Sloutsky, V. M. (2003). The role of similarity in the development of categorization. Trends in Cognitive Sciences, 7, 246-251.
Sloutsky, V. M. (2008) Analogy is to priming as relations are to transformations. Behavioral and Brain Sciences, 31, 396-397.
Sloutsky, V. M., \& Fisher A. V. (2004a). Induction and categorization in young children: A similarity-based model. Journal of Experimental Psychology: General, 133, 166-188.
Sloutsky, V. M., \& Fisher A. V. (2004b). When Development and learning decrease memory. Psychological Science, 15, 553-558.
Sloutsky, V. M., Lo, Y. F., \& Fisher, A. V. (2001). How much does a shared name make things similar? Linguistic labels and the development of inductive inference. Child Development, 72, 1695-1709.
Snodgrass, J. G., \& Corwin, J. (1998). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. Journal of Experimental Psychology: General, 117, 34-50.
Thapar, A., \& McDermott, K. B. (2001). False recall and false recognition induced by presentation of associated words: Effects of retention interval and level of processing. Memory and Cognition, 29, 424-432.
Welder, A. N., \& Graham, S. A. (2001). The influences of shape similarity and shared labels on infants' inductive inferences about nonobvious object properties. Child Development, 72, 1653-1673.
Wilburn C.,\& Feeney A. (2008). Do development and learning really decrease memory? On similarity and category-based induction in adults and children. Cognition, 106, 1451-1464.

\title{
Typicality and Object Reference
}

\author{
Margaret Mitchell (m.mitchell@jhu.edu) \\ Human Language Technology Center of Excellence, Johns Hopkins University Baltimore, MD 21211 USA
}

\author{
Ehud Reiter (e.reiter@abdn.ac.uk) \\ Kees van Deemter (k.vdeemter@abdn.ac.uk)
}

Computing Science Department, University of Aberdeen
Aberdeen, Scotland AB24 3FX UK

\begin{abstract}
Does the typicality of an object affect how we identify it? When we produce initial reference to a visible object, we are influenced by a variety of factors, including what is visually salient (bottom-up influences) as well as our previous experiences with the object (top-down influences). In this study, we seek to understand how the top-down influence of typicality affects initial reference to an object. We use real world, everyday objects, and focus on the visual properties of SHAPE and material. Our findings suggest that there is a tendency to select the atypical over the typical. But we have only begun to scratch the surface of understanding reference to real world objects. The annotated corpus from this study is made available for future research on modeling reference in visual domains. \({ }^{1}\)
\end{abstract}

Keywords: referring expressions; description; reference; vision; typicality

\section*{Introduction}

I never saw a purple cow.
I never hope to see one, But I can say this anyhow:
I'd rather see than be one. - Gelett Burgess

When we identify an object for a hearer, we have a number of choices to make about what to mention. When the object is visible to both speaker and hearer, properties that help guide visual attention, such as color and size, are particularly informative (Treisman \& Gelade, 1980; Wolfe, 2006). Properties that are salient to the discourse or relevant to the speaker and hearer's previous interactions also affect what we will mention and describe (Clark \& Wilkes-Gibbs, 1986; Brennan \& Clark, 1996; Clark \& Krych, 2004).

We hypothesize that when we generate initial reference to an object for a hearer, our knowledge about objects of the same type is also likely to affect what we mention. In other words, our understanding of what is typical for an object category influences the selection of modifiers - the adjectives and longer descriptive phrases - that we produce when we first describe an object. This understanding of what is typical for an object category may stem from stored object prototypes (Rosch \& Mervis, 1975; Rosch, C. Mervis, W. Gray, Johnson, \& Braem, 1976) or mental representations of similar objects in previous situations (Yeh \& Barsalou, 2006; Wu \& Barsalou, 2009). Because of typicality, we know that the purple cow mentioned in the example above is remarkable.

Previous work on reference has paid little attention to the role of typicality. This is equally true for psycholinguistic

\footnotetext{
\({ }^{1}\) http://m-mitchell.com/corpora/typicality_corpus/
}


Figure 1: Example scenes from the GRE3D3 Corpus and the TUNA Furniture sub-corpus. Participants produce referring expressions such as yellow ball on top of the red cube (GRE3D3) or small fan (TUNA).
work (Arnold, 2008) and for work on computational models of reference (Dale \& Reiter, 1995; Krahmer, van Erk, \& Verleg, 2003; Krahmer \& van Deemter, 2012). The present study addresses what we believe to be a significant gap in our understanding of reference.

We examine the role of typicality in reference to real world, everyday objects, focusing on material and shape properties. Objects are presented so that one of these two properties will distinguish the object. We test whether there is a significant difference between groups when participants in one may choose between atypical shape or typical material to describe target objects, while participants in a second group may choose between typical shape or atypical material to describe target objects. Our findings suggest that there is a tendency to select the atypical over the typical.

Although this study focuses on shape and material typicality, we release the full corpus from our experiments, annotated with a variety of visual properties, in hopes of helping further work in constructing models of reference to real objects. Current available corpora for reference to visible objects, such as the GRE3D3 Corpus (Viethen \& Dale, 2008) or the TUNA Corpus (van Deemter, Gatt, van der Sluis, \& Power, 2012), were built from reference elicited to graphics of simple objects presented on a computer screen (see Figure 1). In this work, we seek to better understand the rich details of reference in real world visual domains, where a multitude of different visual properties interact. This opens up several aspects of reference that have not been researched before and gives rise to further questions about the factors influencing initial reference and visual object descriptions. We discuss some of these issues, and their implications for a computa-
tional model of reference.
To establish what properties are typical for an object, we use semantic feature production norms. Semantic feature production norms provide a set of common properties for basiclevel concepts, and are collected to explore conceptual representations such as typicality (Rosch \& Mervis, 1975) and semantics (Wu \& Barsalou, 2009). We use McRae's norms (McRae, Cree, Seidenberg, \& McNorgan, 2005; McRae, 2011), which to our knowledge is the largest source of production norms to date. McRae's norms were collected by providing participants with 10 blank lines for each basic category and asking them to list features for each, such as physical (perceptual) properties (how it looks, sounds, smells, feels, and tastes), functional properties (what it is used for and where and when it is used), and other information, such as encyclopedic facts (e.g., where it is from).

For this study, we are interested in perceptual properties, specifically shape and material properties, which are available from the norms. \({ }^{2}\) For example, objects belonging to the bowl category are listed as typically having a "curved" or "round" shape, and made of a "ceramic" or "plastic" material. We consider atypical properties to be properties (1) mentioned by no participants for the object, and (2) difficult to find during our object collection period.

Our initial list of possible objects included all inanimate objects from McRae's norms that could fit on an experiment table, and this set was narrowed down by availability and our abilities to control the visual properties of the objects. The final set of test objects are listed in Table 1, along with their typical and atypical shapes and materials. The full set of objects used in this study are shown and labeled in Figure 3.

\section*{Using Real World Objects}

A notable complication in this study is that we seek to use a variety of everyday objects, while controlling the typicality of particular properties of those objects. This means that the objects must look relatively commonplace, matching as closely as possible on every visual property except for shape/material; and for these properties, one must be clearly atypical while the other must be clearly typical. Finding everyday objects that fit within these rigid constraints is difficult. In some cases (bowl, mug, screw), we colored the objects to match one another, while in other cases (atypical envelope, key, ruler), we physically created the objects in order for them to have all the desired properties. Real world objects bring with them a set of complications for any model of reference, and we discuss some of these briefly. Although we cannot address all of the issues we list, we hope to provide evidence for a preliminary model of typicality in reference while bringing to light areas for further research.

\section*{Cultural and Individual Differences}

What is typical for an object varies person to person, culture to culture. This study was conducted with a range of students

\footnotetext{
\({ }^{2}\) We use McRae et al.'s "external_surface_property"/"external _component" labels for shape and "made_of" labels for material.
}
and professionals in two countries (the U.S. and the U.K.), but ideally in testing and modeling the production of reference, the set of typical properties would be defined with respect to a culture or group of people, or tailored to a specific person. For our study, we use one set of objects, without changing typical/atypical properties.

\section*{Interconnection}

It is clear that there is an interconnection between different visual properties. For example, material often entails color and texture. An object made of wool is fuzzy or rough (texture values), while an object made of wood is usually tan or brown, and for everyday objects, tends to be smooth (color and e.g., smoothness values). Ideally, participants would refer only to those properties that we vary; but they may instead refer to interconnected properties, calling a woolen bowl "coarse" or "flexible", or a mug made of ceramic but painted silver a "metal mug".

\section*{Lexicalization}

Another competing factor in this study is how easy a property is to lexicalize. Some shapes (e.g., "square") are common and may be quick to access, while other shapes (e.g., "octagonal") may take longer to produce, affecting the object description. Further complications may arise when there is competition over whether to use a prenominal modifier ("the flower-shaped bowl") or a postnominal modifier ("the bowl that looks a bit like a flower").

\section*{Shapes, Parts, and Object Categories}


Figure 2: Bowl, Sugar Bowl, Creamer, Teacup, Mug, Pitcher: Similar objects with different shapes tend to have different names.
An object's shape is often indicated by its name (Markman, 1989; Landau \& Jackendoff, 1993), and therefore an object designed to have an atypical shape may instead appear to belong in an entirely different object category (see Figure 2). For some objects, we found that changing its full shape made it unclear what the object was, or else created a subtype of our target basic-level object category; in a few cases, we therefore manipulated a part of the object's shape. Rather than a round head of a key, the head was square; the straight rectangular center of a ruler was cut out with geometric shapes; and the circular head of a screw was made atypically oval.

\section*{The Objects}

The objects used in the study, as they were presented to participants (without the superimposed identifiers), are shown in Figure 3. Test objects are listed in Table 1 along with their shapes and materials. Those values in italics could not be found in McRae's norms and were added based on intuition and object availability. Filler objects are listed in Table 2.


Figure 3: Objects used in study, keyed to descriptions in Tables 1 and 2 below.

Table 1: Test objects with shapes and materials. Values listed in italics were provided by the authors because they were not listed in the McRae et al. (2005) norms.
\begin{tabular}{|c|c|c|c|c|c|c|}
\hline \multirow[b]{3}{*}{Object} & \multicolumn{3}{|c|}{Group 1} & \multicolumn{3}{|c|}{Group 2} \\
\hline & \multicolumn{3}{|r|}{Atypical Shape} & \multicolumn{3}{|l|}{Atypical Material} \\
\hline & ID & SHAPE & MATER. & ID & SHAPE & MATER. \\
\hline bowl & 2 & flower & ceramic & 1 & round & wool \\
\hline box & 43 & heart & cardboard & 42 & square & clay \\
\hline envelope & 8 & square & paper & 9 & rectangle & foam \\
\hline key & 20 & square head & metal & 19 & \begin{tabular}{l}
rounded \\
head
\end{tabular} & wood \\
\hline mug & 4 & octagona & ceramic & 3 & round & metal \\
\hline ruler & & with holes & & 7 & rectangle & paper \\
\hline screw & 25 & oval head & metal & 24 & flat circular head & plastic \\
\hline
\end{tabular}

Table 2: Filler objects. See Figure 3 for corresponding images of the objects.
\begin{tabular}{llrlrl}
\hline 44 & ball & 40 & coin & 33 & pushpin \\
45 & ball & 17 & comb & 34 & pushpin \\
31 & battery & 18 & comb & 35 & pushpin \\
27 & bracelet & \(48 / 49\) & cube* & 16 & rolling-pin \\
29 & c-clamp & 23 & fork & 10 & rubber-band \\
21 & clip & 5 & funnel & 41 & salt-shaker \\
22 & clip & 11 & pen & 46 & scissors \\
30 & clip & 12 & pen & \(48 / 49\) & sphere* \\
37 & clip & 14 & pen & 28 & staple-remover \\
38 & clip & 15 & pen & 26 & stapler \\
39 & clip & 13 & pencil & 36 & toothpick \\
& & 32 & pushpin & & \\
\hline
\end{tabular}
* These objects were varied by color/size/type as part of a separate pilot experiment.


Figure 4: Subjects sat across from an assistant, with objects between them. An experimenter sat at the head of the table, moving objects to their original positions between trials.


Figure 5: Example stimuli, Atypical Shape group. Here, the test object is the square envelope.

\section*{Experiment}

Participants sat at a table across from an assistant (Figure 4), with a variety of objects on the table between them (Figure 3). Subjects were asked to explain to the person sitting across how to recreate images of the objects grouped in different patterns (Figure 5). There were two objects for each test object category on the table, matched for color and size. One object had atypical material and typical shape; the other had atypical shape and typical material. Subjects could therefore not distinguish a test object to the hearer by its type alone, but could distinguish it by mentioning its shape or material.

Atypical feature is a grouping variable both by subjects and by materials. For the ATYPICAL SHAPE participants, shape properties of the test objects were atypical, while material properties were typical. For the Atypical Material participants, material properties of the test objects were atypical, while shape properties were typical.

\section*{Method}

Participants Thirty native English speakers with normal or corrected vision in the United States and the United Kingdom were paid for their participation ( \(\$ 5\) or \(£ 5\) ). Subjects were recruited through word of mouth and online ads, 17 males and 13 females, aged 20-55, and randomly assigned to one of the two experimental groups, (1) ATYPICAL SHAPE or (2) Atypical Material. Four male subjects and one female subject were randomly removed to a held-out set to balance gender, leaving 6 female and 6 male subjects in each group (24 subjects total).

To check for possible outliers in each group, we calculate the average number of references with shape, and the average number of references with material. Participants whose total number of references with shape or material were more than two standard deviations from the mean for that property were identified as possible outliers. We found no outliers in the Atypical Material group, and two possible outliers in the Atypical Shape group. The data for these two subjects (one male, one female) were removed and replaced with gender-matched data from the held-out set.

Materials Participants sat in front of a large set of everyday objects (rulers, envelopes, pins, etc., as shown in Figures 3 and 4), with test objects and fillers mixed. Test objects for the two experimental groups with their corresponding shape and material properties are listed in Table 1.
Procedure \& Design All participants consented to participate in the study. The experiment followed a director-matcher paradigm, where the director (the participant) instructed the matcher (the assistant). Participant and assistant sat opposite one another while the experimenter sat on a third side of the table. Participants alone viewed pictures on a laptop (positioned so screen was not visible to the assistant). Each participant saw 8 pictures in randomized order including a different atypical stimulus each time (see example in Figure 5). On each trial, the participant viewed the picture and explained to the assistant where each pictured object should go on the grid laid out on the table between them. At the end of each trial, the experimenter returned objects to their original positions. Participants' instructions were recorded onto the laptop.

\section*{Results}

Annotation Annotations are provided by the first author. To test for adequacy of the annotation system, a second annotator annotated a random subset of 20 references to the test objects. The annotator was given mark-up instructions and examples of a variety of visual properties (shown in Table 3), and told to mark which words referred to which properties as best they could following their understanding of the examples. Treating shape and material as binary categorical variables, Cohen's \(\kappa\) is very good for shape ( \(\kappa=.894\) ) and good for material ( \(\kappa=.798\) ) between annotators. Disagreements were over whether "metallic" in "the non-ribbed metallic cup" was a material or a texture, whether "heart" in "a heart-shaped box" was a shape or a type, and whether "silver" in "a silver round cup" was a color or a material. The total number of expressions produced for test items with shape and material modifiers in each experimental group is shown in Figure 6.
Analysis We see a tendency to choose the atypical over the typical in both groups. In the Atypical Shape group, 54 expressions contain a shape modifier while 36 contain a material modifier; in the Atypical Material group, 28 expressions contain a shape modifier while 41 contain a material modifier (see Figure 6). There is a slight preference for shape over material across the groups; 82 expressions contain a shape modifier while 77 contain a material modifier.

We are interested in understanding whether there is a significant difference in the selection of modifiers between groups. For each participant, we subtract the number of test object expressions containing a modifier for the object material from the number of test object expressions containing a modifier for the object shape. In other words, for each participant \(p\), given the number of expressions with material modifiers \(M_{p}\) and the number of expressions with shape modi-

Table 3: Attributes annotated and example surface forms.


Figure 6: Number of expressions with shape and material modifiers in each experimental group.
fiers \(S_{p}\), we calculate \(V_{p}=S_{p}-M_{p}\). This provides a vector for each group with differences in the number of modifiers for shape and material. For participants in the Atypical Shape group, where shape is atypical, we expect positive values for \(V_{p}\). For participants in the Atypical Material group, where material is atypical, we expect negative values for \(V_{p}\).

An independent samples t-test was conducted to compare the effect of property typicality on the production of modifiers between groups. There is a significant difference at \(\alpha=.01\) between the Atypical Shape group ( \(\mathrm{n}=12\), mean \(=1.50, \mathrm{sd}=1.62\) ) and the Atypical Material group \((\mathrm{n}=12\), mean \(=-1.08, \mathrm{sd}=2.07) ; \mathrm{t}(21)=3.406, \mathrm{p}=0.003\).

\section*{Discussion}

\section*{Current Study}

These results suggest that atypicality affects object reference. We find a tendency to select the atypical property over the typical one, with participants in the ATYPICAL SHAPE group preferring shape modifiers, and participants in the ATYPICAL Material group preferring material modifiers. This difference is significant between groups.

Shedding further light on these findings, when material was included in a reference in the Atypical Material group, it was often incorrect. Figure 7 illustrates how frequently subjects were correct and incorrect in the description of an object's material. In the Atypical Material group (Figure 7a), the plastic screw was called "metal", the paper ruler was called "wooden". The ruler in particular gave rise to incorrect material modifiers - it was printed on paper with a wood print,


Figure 7: Number of participants who included material modifiers that were correct, incorrect, or did not use a material modifier at all (neither). In the Atypical Material group (a), ruler tends to evoke incorrect material modifiers. The ruler, envelope, and screw had no correct material modifiers. In the Atypical Shape group (b), the mug, box, and bowl tended to evoke incorrect material modifiers. The mug (painted silver) had no correct material modifiers.
so it was called "wooden". Most participants in the Atypical Material group did not use material modifiers for the envelope and the screw, which may be partially due to the fact that the screw was painted black and so was not clearly plastic; and the envelope was made of foam, which may have not been clear without physically touching the object. Some examples of expressions in the ATYPICAL MATERIAL group that do not include material modifiers are given in Table 4.

A similar tendency to refer to incorrect material emerged in the Atypical Shape group (Figure 7b), where the ceramic mug painted silver was called "metal" or "steel", the ceramic bowl was called "plastic", and the cardboard box was called "wooden". In contrast, subjects were rarely incorrect about shape (Figure 8). The only exception to this is the atypically shaped mug (Figure 8b), which was called "octagonal" (correct), "hexagonal" (incorrect) and "septuplet" (incorrect).

These trends suggest that material is not purely visual, but may also be guided by our tactile sensations of the objects; without tactile input, our expectation of the typical material for the object may be used in our reference rather than its actual material, or we may disprefer material altogether. It is not enough to judge whether a visual property is atypical or not; it must also be judged whether that value is visually clear, and whether other properties suggest another interconnected


Figure 8: Number of participants who included shape modifiers that were correct, incorrect, or did not use a shape modifier at all (neither). In the Atypical Material group (a), there were no incorrect shape modifiers. In the ATYPICAL SHAPE group (b), only the mug received incorrect shape modifiers (e.g., "hexagonal" rather than "octagonal").
property (which may not actually be true of the referent, as when subjects use incorrect modifiers).

Future work could shed more light on these issues with a variant of the experiment in which we look at how reference here compares to reference towards objects of the same category that have typical shape and typical material, and likewise, atypical shape and atypical material. This should also provide useful data on how to separate general preference for shape or material versus a preference for atypical properties.

\section*{Towards a Model of Reference}

As discussed in the introduction, current research on reference production has focused on very simple, constrained domains. In this work, we propose taking the object's typicality into consideration when deciding which properties to add to an initial description. In a computational model, typical information may be made available in a knowledge base queried at runtime. As part of the reference production process, the target object category could then be compared against a stored object category. Property selection in such a model could be a function of the typicality of the property for the object, as well as, e.g., its contrastive value against the other objects. This offers an extension to current models of referring expression generation (e.g., Dale and Reiter (1995)), and may help to further explain the process of reference generation.
\begin{tabular}{lll}
\hline Ruler & Envelope & Screw \\
\hline "the ruler" & \begin{tabular}{l} 
"the white \\
envelope"
\end{tabular} & "the screw" \\
\begin{tabular}{l} 
"ruler that's
\end{tabular} & \begin{tabular}{l} 
"the uh weird \\
padded looking \\
flatter"
\end{tabular} & \begin{tabular}{l} 
"black flat head \\
screw"
\end{tabular} \\
\begin{tabular}{l} 
"the darker \\
tan ruler"
\end{tabular} & \begin{tabular}{l} 
"long rectang" \\
envelope"
\end{tabular} & \begin{tabular}{l} 
"the screw with \\
the flat head"
\end{tabular} \\
\hline
\end{tabular}

Table 4: Examples of references without material in the Atypical Material group. We see underspecified references and references describing the object's size.

\section*{Conclusions and Future Work}

This study has sampled a handful of real world objects to understand the role that typicality plays on reference. We have focused on two visual attributes, shape and material, in a visual scenario where either may be used to identify an object. We see a tendency to select the atypical property over the typical one, and find a significant difference between the selection of atypical shape over typical material versus typical shape over atypical material between groups.

Our study has focused on relatively crisp properties of objects. It would be interesting to explore whether our findings extrapolate to properties that come in degrees, such as height, weight, age, and so on. Based on our study, we hypothesize, for example, that the length of a screw is more likely to be mentioned if it is unusual (e.g., unusually long or unusually short) than if it is not. In line with this, gradations of atypicality for an attribute may also affect reference; some values may be more atypical than others, and thus more likely to be included in a final description.

This study leaves many further open questions. To fully model reference production to real world, visible objects, we must better understand how the production of visual modifiers is affected by interconnection and lexicalization issues, and how notions of typicality are changed culture to culture.

In future work, we aim to focus directly on interconnection, understanding how the degree of correlation between properties affects description. We would like to extend our set of objects in order to examine reference when both properties are typical, or both atypical. There may be a tendency to select the atypical over the typical, but we have only begun to scratch the surface of the factors at play when we refer to real world objects.

\section*{Acknowledgments}

The first author of this paper was supported by funding from the Scottish Informatics and Computer Science Alliance (SICSA), the Overseas Research Students Awards Scheme (ORSAS) and Oregon Health and Science University (OHSU). Thanks to Ellen Bard at the University of Edinburgh for help with design and analysis. Thanks to students at OHSU for volunteering their time as assistants and thanks to SignWorksOfOregon - particularly Shannon Zuiderweg, for the construction of the wooden key.

\section*{References}

Arnold, J. E. (2008). Reference production: Productioninternal and addressee-oriented processes. Language and Cognitive Processes, 23.
Brennan, S. E., \& Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. Journal of Experimental Psychology: Learning, Memory, and Cognition, 22.
Clark, H. H., \& Krych, M. A. (2004). Speaking while monitoring addressees for understanding. Journal of Memory and Language, 50.
Clark, H. H., \& Wilkes-Gibbs, D. (1986). Referring as a collaborative process. Cognition, 22.
Dale, R., \& Reiter, E. (1995). Computational interpretations of the gricean maxims in the generation of referring expressions. Cognitive Science, 19.
Krahmer, E., \& van Deemter, K. (2012). Computational generation of referring expressions: A survey. Comp. Ling., 38.

Krahmer, E., van Erk, S., \& Verleg, A. (2003). Graph-based generation of referring expressions. Comp. Ling., 29.
Landau, B., \& Jackendoff, R. (1993). "what" and "where" in spatial language and spatial cognition. Behavioral and Brain Sciences, 16.
Markman, E. (1989). Categorization and naming in children: Problems of induction. Cambridge, MA: MIT Press.
McRae, K. (2011). McRae's norms. Available from http://amdrae.ssc.uwo.ca/McRaeLab/norms.php. Accessed 5.Oct. 2011
McRae, K., Cree, G. S., Seidenberg, M. S., \& McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. Behavior Research Methods, Instruments, \& Computers, 37(4).
Rosch, E., C. Mervis, C., W. Gray, W., Johnson, D., \& Braem, P. B. (1976). Basic objects in natural categories. Cognitive Psychology, 8.
Rosch, E., \& Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. Cognitive Psychology, 7.
Treisman, A. M., \& Gelade, G. (1980). A feature integration theory of attention. Cognitive Psychology, 12.
van Deemter, K., Gatt, A., van der Sluis, I., \& Power, R. (2012). Generation of referring expressions: Assessing the incremental algorithm. Cognitive Science, 36(5).
Viethen, J., \& Dale, R. (2008). The use of spatial relations in referring expression generation. Proc. INLG 2008.
Wolfe, J. M. (2006). Guided search 4.0. In W. D. Gray (Ed.), Integrated models of cognitive systems. New York, NY: Oxford University Press.
Wu, L. L., \& Barsalou, L. W. (2009). Perceptual simulation in conceptual combination: Evidence from property generation. Acta Psychologica, 132.
Yeh, W., \& Barsalou, L. W. (2006). The situated nature of concepts. American Journal of Psychology, 119.

\title{
Stoic Behavior in Hint Seeking when Learning using an Intelligent Tutoring System
}

\author{
Kazuhisa Miwa (miwa@is.nagoya-u.ac.jp) \\ Hitoshi Terai (terai@is.nagoya-u.ac.jp) \\ Nana Kanzaki (kanzaki@cog.human.nagoya-u.ac.jp) \\ Graduate School of Information Science, Nagoya University
}

\section*{Ryuichi Nakaike (nakaike@educ.kyoto-u.ac.jp)}

Graduate School of Education, Kyoto University

\begin{abstract}
Stoic behavior is defined as a behavior in which students tend not to seek help with a challenge. We investigated two types of stoic behavior: keeping-off behavior, in which students restrain themselves from requesting help, i.e., keep levels of help support at a minimum, and self-fading behavior, in which students voluntarily lower levels of support on their own volition. Three experiments were conducted. Overall, results showed that the participants actually exhibited stoic behavior when learning in an actual classroom setting. Self-fading was more difficult than the keeping-off behavior. The participants who maintained levels of support at a minimum through exhibiting active keeping-off behavior achieved greater learning gains, suggesting that stoic behavior resulted in positive impacts on learning. However, our experiment did not detect this effect for self-fading behavior. These experimental results were discussed with the assistance dilemma problem, generally occurring in instruction by intelligent tutoring systems.
\end{abstract}

Keywords: Intelligent tutoring system; Help seeking; Assistance dilemma.

\section*{Introduction}

Recent intelligent tutoring systems include highly interactive features. Such systems give participants various types of feedback such as verification, correct response notification, try again encouragement, error flagging, and elaboration messages (Shute, 2008). In this context, the assistance dilemma has been recognized. Koedinger \& Aleven (2007) asked a crucial question: How should learning environments balance assistance giving and withholding to achieve optimal learning? (Koedinger \& Aleven, 2007) High assistance sometimes provides successful scaffolding and improves learning, however, it may also elicit superficial responses without consideration from the students. In contrast, low assistance sometimes encourages students to make great efforts to learn, while at other times it results in enormous errors and interferes with effective learning. To resolve this issue, the levels of support (LOS) must be adaptively controlled by tutoring systems.

Equipping tutoring systems with intelligent functions for help control and optimization of feedback information for participants are important issues. However, students must seek help intelligently. From this perspective, in this study, we investigate students' active help-seeking behaviors rather than passive help-receiving behaviors, which were managed by intelligent tutoring systems. Razzaq \& Heffernan (2011) (Razzaq \& Heffernan, 2010) confirmed that active hintseeking, in which on-demand hints were given, was more ef-
fective for learning than passive hint-seeking, in which participants were given hints proactively when facing errors.

Help-seeking is a representative metacognitive activity in learning behavior. Help-seeking is valuable, not only for maximizing learning effects but also for acquiring a domainindependent meta-learning strategy. Some trials have instructed students using such metacognitive abilities. A domain-independent agent, called Help Tutor, for teaching better help-seeking skills by tracing students actions, was developed (Roll et al., 2006). Such help-seeking support was successful in improving students’ declarative help-seeking knowledge, but did not improve their overall learning (Roll, Aleven, McLaren, \& Koedinger, 2007). In more recent trials, the Help Tutor improved students' help-seeking behavior, and the improved help-seeking skills are transferred to learning new domain-level content during the month following the intervention (Roll, Aleven, McLaren, \& Koedinger, 2011). To instruct such metacognitive activities, we must learn more about the nature of students' help-seeking behavior.

Students themselves have to manage their help-seeking behavior to maximize learning effects. However, previous studies have demonstrated that students' help-seeking behavior does not follow rational principles (Wood \& Wood, 1999). Hint abuse is a representative irrational behavior that appears in hint-seeking whereby students tend to seek the most specific hints to find answers rather than acquiring understanding (Aleven \& Koedinger, 2000).

In this study, we focus on stoic behavior in hint seeking. Stoic behavior is defined as behavior in which students tend not to seek help for their challenge. We will investigate two types of behavior: the keeping-off and self-fading behaviors. Keeping-off behavior is defined as behavior in which students try to solve problems by themselves without a system's assistance. Students restrain themselves from receiving help from a tutoring system even when permitted to do so. This is regarded as a type of behavior with the purpose of avoiding the hint abuse. On the other hand, we define self-fading behavior as one in which students voluntarily decrease an LOS by their own volition. This behavior is recognized, along with scaffolding, as a central concept for effective learning. To enable students to effectively learn, scaffolding should be eliminated gradually as learning progresses. In learning by examples, fading methods have also been used as an effective principle
for controlling the flow of learning (Atkinson, Renkl, \& Merrill, 2003).

In this study, we investigate the following two research questions: Do students exhibit stoic behavior in hint seeking when learning with an intelligent tutoring system in classroom settings? and does such stoic behavior in hint seeking promote learning gains? We examined these research questions through three empirical studies.

\section*{Learning System and Task}

We investigated participants' help-seeking behavior using a relatively complex learning task in which participants learned natural deduction (ND). Natural deduction is a kind of proof calculus in which logical reasoning is expressed by inference rules closely related to a natural way of reasoning. The following is an example solution process of an example problem: inducing a proposition \(\neg Q \rightarrow \neg P\) from a premise \(P \rightarrow Q\).
\begin{tabular}{lll} 
(1) & \(P \rightarrow Q\) & Premised \\
(2) & \(\neg Q\) & Assumption \\
(3) & \(P\) & Assumption \\
(4) & \(P \rightarrow Q\) & Reiteration of (1) \\
(5) & \(Q\) & \(\rightarrow\) Elimination from (3) and (4) \\
(6) & \(\neg Q\) & Reiteration of (2) \\
(7) & \(\neg P\) & \(\neg\) Introduction from (3), (5), and (6) \\
(8) & \(\neg Q \rightarrow \neg P\) & \(\rightarrow\) Introduction from (2) and (7)
\end{tabular}

Participants learned inference rules and strategies for applying the rules. Participants in this study learned eight basic rules and four strategies, which are the fundamental basis of ND ; the majority of problems can be solved using these rules and strategies.

Our tutoring system, which was developed for teaching ND to university undergraduates, has two important features.

First, it does not have a database that contains a set of ND problems and their solutions. Our system solves each problem on demand. It includes a production system model, which consists of the working memory, whose layout is consistent with the structure of ND problems, and production rules, which correspond to the inference rules and strategies for solving ND problems.

As a second feature, our system was established based on a server-client framework. Miwa et al. developed a webbased production system architecture called DoCoPro that enables such a system design to be established (Miwa, Morita, Nakaike, \& Terai, 2013). A problem solver constructed on a server performed the complex inferences in ND. Client computers connected to the server performed easy processing for the interface. Using this server-client framework, our system can operate in any educational environment where various types of computers, e.g., high performance, poorly performing, and on different types of operating systems. Participant learning processes are saved as \(\log\) data on the server.

Figure 1 shows a screenshot of the tutoring system. The system provides the participants with lists of the inference rules and strategies. They select one of the rules or strategies from a list, and the system automatically runs the rules and presents partial or complete results of inference. The system


Figure 1: Example screenshot of the tutor terminal.
scaffolds the students by providing helpful information about the selection of the rules and strategies.

The LOS can be controlled from two viewpoints: rule selection and application.
LOS for rule selection: Level 3 (high): The system presents applicable candidates (rules and strategies) and the propositions to which the rules should be applied. For example, in the middle window in Figure 1, the system proposes that three highlighted inference rules and one strategy could be applied. When " \(\rightarrow\) Elimination" is selected, \(P\) and \(P \rightarrow Q\) are highlighted in the left window, indicating that the selected rule should be applied to these propositions. Level 2 (middle): The system presents only applicable candidates (rules and strategies). When this level is selected, students are required to find the propositions to which the selected rule should be applied without receiving support from the system. Level 1 (low): The system presents only a set of inference rules and strategies (no support is provided).
LOS for rule application: Level 2 (high): The system infers a proposition and automatically presents it in the left window. Shortly after students select an inference rule and the propositions to which the rule will be applied, the system displays the current status of deduction. Level 1 (low): The system infers a proposition, but presents only partial information of the inferred result. Students are required to complete the inference process by filling terms in blank spaces of a template.

\section*{Experiment 1}

In Experiments 1 and 2, the initial setting of LOS at the beginning of solving each problem was lowest. Once a new problem was set, the LOS was initialized to Level 1. The participants were required to determine whether to raise an LOS from the initial setting while solving each example problem. Therefore, Experiments 1 and 2 investigated the participants'
keeping-off behavior in help-seeking. We will investigate the self-fading behavior in Experiment 3.

Experiment 1 was a preliminary experiment. Experiment 1 was performed in a laboratory setting; Experiments 2 and 3 were performed in a real classroom setting.

\section*{Participants and Procedure}

Thirty-three participants joined Experiment 1. In the initial phase of the experiment, the participants learned the basics of ND through handout materials and an instructional video. They learned eight inference rules and four strategies without the tutoring system. After the participants were instructed on how to use the tutoring system, they learned ND by solving six example problems with our tutoring system through the \(80-\) min learning phase. Two of the problems were difficult and required a second-order subproof, and two were easy and either did not require a subproof or required only a first-order subproof. The data recoreded in this phase were analyzed.

\section*{Results}

We focused on two kinds of help control behavior. One is a relatively simple behavior. In our experiments, participants were allowed to solve problems at their own pace. Some participants quit solving a problem, moved to other problems, and then revisited the initial problem and attempted to complete it. Our first hypothesis is that participants in the second and following attempts, compared to the first attempt, would not raise it for their challenge. The other is a more sophisticated behavior: we expected that participants would adopt an LOS according to the degree of difficulty of each problem. Our hypothesis is that participants solving easy problems would select a lower LOS than when solving difficult problems, despite the fact that they were permitted to receive help if they wished.

We compared the average LOS of the first attempt with that of the second-and-following attempts. In certain cases, the participants attempted to solve a problem more than twice. In such cases, we used the average score of the second-and-following attempts. Figure 2 shows the result of analysis about the LOS control between the first and second-andfollowing attempts. A t-test revealed a significant difference between the first and second-and-following attempts in rule selection \((t(116)=6.10, p<0.01)\), but not in rule application \((t(115)<1\), n.s. \()\).
Figure 3 shows the result of analysis about more sophisticated behavior, i.e., the LOS control when solving easy and difficult problems. A t-test revealed a significant difference between the easy and difficult problems in both rule selection and application \((t(31)=2.59, p<0.05) ; t(31)=4.58, p<\) \(0.01)\).

The above results indicated that the participants kept an LOS at low in the second and following attempts, relative to the first attempt, but only in rule selection, and did not raise it when solving easy problems compared to when solving difficult problems. These results supported our hypotheses about the participants' keeping-off behavior in help-seeking. This


Figure 2: Levels of support versus number of attempts in Experiment 1.


Figure 3: Levels of support versus problem difficulty in Experiment 1.
stoic behavior was observed greatly in rule selection than application.

\section*{Experiment 2}

In Experiment 2, we performed both pre- and post-tests, before and after the learning phase, to examine the relationship between help-seeking behavior and the learning effects. We also focused on whether the participants' stoic help-seeking behavior depends on their problem solving ability. In Experiment 1 , we confirmed that stoic behavior was greatly observed in rule selection; therefore, in Experiments 2 and 3, the LOS in rule application was fixed at Level 1, and only the LOS in rule selection was investigated.

\section*{Participants and Procedure}

Forty-nine participants from a cognitive science class joined the 2011 experiment. Three lessons were assigned to learn ND. In the first lesson, an instructor lectured on the basics of formal inference systems and ND as an example of such systems.
In the second lesson, the participants initially solved six problems while learning how to use the tutoring system. First, the instructor presented an example flow of problem solving. Then, participants followed the flow and reached the solution using the system. After the initial training, the participants were given two new problems to solve. Finally, the


Figure 4: Levels of support versus number of attempts in Experiment 2.
participants were given a paper test in which they solved a test problem; we used this test as a pretest in the following analysis.
The log data from the third lesson were analyzed. The participants solved eight problems at their own paces and selected an LOS. Three of the eight problems were easily solved by applying the basic rules learned in the first lesson. However, three problems were relatively difficult, and their solutions required more complex rules and solution strategies, such as subgoal settings. The learning session lasted for an hour. After the learning session, a post-test was performed.

\section*{Results}

To investigate whether the participants' help-seeking behavior is dependent on their problem solving ability, we divided the participants into two groups on the basis of their pre-test scores, and formed lower- and higher-score groups.
Figure 4 shows the results of analysis on simple help management behavior, i.e., the LOS control between the first and second-and-following attempts. A two (attempt: first and second-and-following) x two (ability: high and low) ANOVA revealed that the main effect of the attempt factor reached significance \((F(1,158)=136.93, p<0.01)\), but the main effect of the ability factor \(\operatorname{did} \operatorname{not}(F(1,158)=1.54\), n.s.). There was no interaction between the two factors \((F(1,158)<1\), n.s.).

Figure 5 shows the results of analysis on more sophisticated behavior, i.e., the LOS control when solving the easy and difficult problems. A two (problem: easy and difficult) \(x\) two (ability: high and low) ANOVA revealed that the main effect of the problem factor reached significance \((F(1,44)=\) \(33.02, p<0.01\) ), but the main effect of ability factor did not \((F(1,44)<1\), n.s.). There was no interaction between the two factors \((F(1,44)=2.96\), n.s. \()\).

The above results duplicated the participants' stoic behavior captured in Experiment 1. Additionally, the same tendency was observed in both low- and high-score groups, meaning that such help-seeking behavior does not depend on the participants' problem solving ability.

Next, we focus on the analysis of the relation of helpseeking behavior and learning effects. We hypothesize that a lower LOS may provide greater learning effects and a higher


Figure 5: Levels of support versus problem difficulty in Experiment 2.


Figure 6: Levels of support versus learning gains in Experiment 2.

LOS may obstruct effective learning. We divided the participants into two groups on the basis of their average LOS during problem solving in the learning phase. The problem used in the pre-test was different from those used in the posttest; therefore, we cannot directly compare the scores of the two tests. Accordingly, we transferred the test scores to the z -scores in each of the two tests and calculated the gains of the z -score from the pre- to post-tests.

Figure 6 shows the results of the analysis. A two (LOS in learning phase: high and low) x two (ability: high and low) ANOVA revealed that both the main effects of the LOS factor and the ability factor reached significance \((F(1,45)=8.28, p\) \(<0.01 ; F(1,45)=26.98, p<0.01)\). There was no interaction between the two factors \((F(1,45)=1.44\), n.s. \()\).
The result shows that the participants who learned with a lower LOS in the learning phase gained greater learning effects. This means that stoic behavior, especially the keepingoff behavior in this case, promoted learning.

\section*{Experiment 3}

In Experiment 2, we focused on the keeping-off behavior. Experiment 3 investigated the self-fading behavior in helpseeking,

\section*{Participants and Procedure}

Twenty-eight participants from a cognitive science class joined our 2012 experiment. Three lessons were assigned for learning ND and the learning content and procedures were almost identical to Experiment 2. The crucial differ-


Figure 7: Levels of support versus number of attempts in Experiment 3.


Figure 8: Levels of support versus problem difficulty in Experiment 3.
ence was that the initial setting of LOS at the start to solve each problem was the highest (Level 3) in Experiment 3. The participants were required to determine whether to lower an LOS from the initial setting, while solving example problems. Therefore, Experiment 3 investigated the participants' selffading behavior in help-seeking.

\section*{Results}

Figure 7 shows the results of analysis on the LOS control between the first and second-and-following attempts. A two (attempt: first and second-and-following) \(x\) two (ability: high and low) ANOVA revealed that the main effects of both the attempt and ability factors reached significance \((F(1,54)=\) \(34.43, p<0.01 ; F(1,54)=4.98, p<0.05)\). There was significant interaction between the two factors \((F(1,54)=4.47\), \(p<0.05\) ). The simple main effect of the ability factor at the first attempt was not significant \((F(1,108)<1\), n.s. \()\), but the effect at the second-and-following attempts was significant \((F(1,108)=9.41, p<0.01)\).

Figure 8 shows the result of analysis about the LOS control when solving the easy and difficult problems. A two (problem: easy and difficult) x two (ability: high and low) ANOVA revealed that there was neither a main effect of the problem factor nor a main effect of the ability factor \((F(1,25)<1\), n.s.; \(F(1,25)=2.67\), n.s.). There was no interaction between the two factors \((F(1,25)<1\), n.s.).

In Experiment 3, we confirmed the stoic behavior only in the LOS control between the first and second-and-following attempts, but not in the LOS control when solving easy and difficult problems. In the former case, the higher abil-


Figure 9: Levels of support versus learning gains in Experiment 3 .
ity participants greatly lowered the LOS in the second-andfollowing attempts compared with the lower ability participants.

Next, we focus on the analysis of the relationship between the self-fading behavior and learning effects. The same analysis as in Experiment 2 was performed. Figure 9 presents the results of the analysis. A two (LOS in learning phase: high and low) \(x\) two (ability: high and low) ANOVA revealed that there was neither a main effect of the LOS factor nor interaction between the two factors \((F(1,24)<1\), n.s.; \(F(1,24)<\) 1 , n.s.). However, the main effect of the ability factor reached significance \((F(1,24)=5.09, p<0.05)\). Learning effects by the stoic help-seeking behavior were not confirmed in Experiment 3.

\section*{Discussion and conclusions}

The first research question we posed was: Do students exhibit stoic behavior in hint seeking? We examined two types of stoic behavior: Experiment 2 investigated the keeping-off behavior, and Experiment 3 investigated the self-fading behavior. We hypothesized that participants would lower an LOS with the development of learning (i. e., from the first to the second-and-following attempts). This hypothesis was fully supported in both Experiments 2 and 3. More specifically, with regard to self-fading, high ability participants more actively lowered the LOS in the second-and-following attempts compared with the lower ability participants. The second hypothesis was whether participants would adaptively manage their help-seeking behavior based on the degree of problem difficulty. We expected that they lower an LOS or would not raise it when solving easy problems, compared to when solving difficult problems. This hypothesis was supported only in Experiment 2, indicating that the participants kept an LOS at low (Experiment 2), but that they did not reduce an LOS from high to low (Experiment 3) when solving easy problems, which suggests that the keeping-off behavior was confirmed, though the self-fading behavior was not.

These results imply that the self-fading behavior, as an adaptive behavior in help-seeking, was more difficult for the participants than the keeping-off behavior. The latter behavior comes from a strategy to set the LOS at low by stopping action (i.e., stopping raising an LOS). However, the former
behavior comes from a strategy to set the LOS at low by performing an action (i.e., beginning to reduce an LOS). The latter is relatively passive, while the former is an intentional and active behavior. These results suggest that an active type of stoic behavior was more difficult for users.

The second research question was: Does the stoic behavior in hint seeking promote learning gains? This relates to a trade-off of selecting either the problem-solving or the learning goal. Participants learn while solving instance problems given by a tutoring system. Attaining the problem-solving goal means solving such instance problems as accurately and rapidly as possible. However, the learning goal requires another attainment that is usually more essential. The primary objective is not to solve instance problems, but to learn by solving instances. Dweck classified two types of goals: learning and performance (Dweck, 1986; Ames, 1992). Highly motivated children tend to set learning goals in an effort to increase their competence levels for understanding or mastering something new rather than simply solving problems. Our previous study confirmed that high learning supports promote the problem solving goal setting, and refrain the learning goal setting (Miwa, Terai, \& Nakaike, 2012). In the high LOS situation, participants may solve training problems accurately and rapidly in the learning phase, but tend to learn least from the training.

The assistance dilemma hypothesizes an optimum point of learning effects. Koedinger et al. (2008) demonstrated a reverse U-shape learning curve as a function of cognitive load (Koedinger, Pavlik, Mclaren, \& Aleven, 2008). This means that extremely lower and higher cognitive loads result in negative impacts on learning. The levels of help support are correlated with learners' cognitive load while learning. Much help reduces their cognitive load for problem solving in the learning phase where students simply respond to help indications from a tutoring system, e.g., a direct instruction about what to do next, without deeper consideration. From this viewpoint, our experimental results are considered to capture the right side of the reverse U -shape. We compared learning effects when the participants learned with a low and a high LOS. In the right half, the reversed U-shape predicts that a lower LOS provides more learning effects; Experiment 2 supported this prediction. However, we also expected that the effects of learning decrease gradually, in the left side of the reverse U-shape, as the LOS is reduced. In another experiment (Miwa et al., 2012), we confirmed this prediction using the same tutoring system, in which we set up two experimental conditions. In the system condition, the participants learned ND using our tutoring system. They were permitted to control the LOS. In the control condition, i.e., the paper-and-pencil condition, participants learned ND without a tutoring system, instead they learned ND using only a textbook. The latter was the no support condition. Results showed that learning effects in the system condition were greater than in the paper-and-pencil condition. In this experiment, no support relates to the leftmost side of the reversed U -shape.

\section*{References}

Aleven, V., \& Koedinger, K. R. (2000). Limitations of student control: Do student know when they need help? In Proceedings of the 5th international conference on intelligent tutoring systems, its 2000 (p. 292-303).
Ames, C. (1992). Classrooms: Goals, in structure, and student motivation. Journal of Educational psychology, 84, 261-271.
Atkinson, R. K., Renkl, A., \& Merrill, M. M. (2003). Transitioning from studying examples to solving problems: Effects of self-explanation prompts and fading worked-out steps. Journal of Educational Psychology, 95, 774-783.
Dweck, C. S. (1986). Motivational processes affecting learning. American Psychologist, 41, 1040-1048.
Koedinger, K. R., \& Aleven, V. (2007). Exploring the assistance dilemma in experiments with cognitive tutors. Educational Psychology Review, 19, 239-264.
Koedinger, K. R., Pavlik, P., Mclaren, B., \& Aleven, V. (2008). Is it better to give than to receive? the assistance dilemma as a fundamental unsolved problems in the cognitive science of learning and instruction. In Proceedings of the 30th annual conference of the cognitive science society (p. 2155-2160).

Miwa, K., Morita, J., Nakaike, R., \& Terai, H. (2013). Learning through intermediate problems in creating cognitive models. Interactive Learning Environments. (in press).
Miwa, K., Terai, H., \& Nakaike, R. (2012). Tradeoff between problem-solving and learning goals: Two experiments for demonstrating assistance dilemma. In Proceedings of the 34th annual conference of the cognitive science society (p. 2008-2013).
Razzaq, L., \& Heffernan, N. (2010). Hints: Is it better to give or wait to be asked? In Proceedings of the 10th international conference on intelligent tutoring systems, its 2010 (p. 349-358).

Roll, I., Aleven, V., McLaren, B., \& Koedinger, K. R. (2007). Can help seeking be tutored? searching for the secret sauce of metacognitive tutoring. In Proceedings of the 13th international conference on artificial intelligence in education, aied 2007 (p. 203-210).
Roll, I., Aleven, V., McLaren, B., \& Koedinger, K. R. (2011). Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. Learning and Instruction, 21, 267-280.
Roll, I., Aleven, V., McLaren, B., Ryu, E., Baker, R., \& Koedinger, K. R. (2006). The help tutor: Does metacognitive feedback improve students' help-seeking actions, skills, and learning? In Proceedings of the 8th international conference on intelligent tutoring systems, its 2006 (p. 360-369).

Shute, V. J. (2008). Focus on formative feedback. Review of Educational Research, 78, 153-189.
Wood, H., \& Wood, D. (1999). Help seeking, learning and contingent tutoring. Computers and Education, 33, 153169.

\title{
Unmet Expectations in the Comprehension of Relative Clauses in Japanese
}

\author{
Edson T. Miyamoto (MIYAMOTO@Alum.Mit.Edu) \\ Graduate School of Humanities and Social Sciences, Tennodai 1-1-1 \\ Tsukuba, Ibaraki 305-8571 Japan \\ Michiko Nakamura (M-NAKA@Fun.Ac.Jp) \\ School of Systems Information Science, Kamedanakano-cho 116-2 \\ Hakodate, Hokkaido 041-8655 Japan
}

\begin{abstract}
In two self-paced reading experiments, subject relative clauses (e.g., 'the woman who saw the man') were read faster than object relative clauses ('the woman who the man saw') in Japanese. Previous formulations of working-memory factors do not predict the patterns observed. A preference to complete fragments as object relative clauses indicates that ambiguity and expectation are unlikely to explain the reading-time data. The results support the proposal that accessibility of the position relativized affects how natural the relative clause is as a statement about the modified noun.
\end{abstract}

Keywords: relative clauses, Japanese, subject, object

\section*{Introduction}

An accessibility hierarchy of grammatical roles has been shown to constrain the grammaticality of relative clauses in 49 typologically distinct languages (Keenan \& Comrie, 1977; also Comrie, 2007, for a recent summary). Subjects are ranked higher than objects to reflect the observation that more languages allow subject relative clauses than object relative clauses. Postnominal relative clauses as in (1) follow the modified noun woman (brackets mark the boundaries of the embedded clause; the extraction site or gap is coindexed with the modified noun, or filler).
(1) a. Subject relative clause (postnominal) the \(\operatorname{woman}_{i}\) [that \(g a p_{i}\) saw the man]
b. Object relative clause (postnominal)
the woman \(_{j}\) [that the man saw gap \(_{j}\) ]
In languages that allow both alternatives, subject relatives are predicted to be easier to understand, that is, comprehension is facilitated when woman is interpreted as the subject of the embedded verb saw as in (1a). The subject advantage is well-documented for constructions in which both nouns are animate (man and woman in (1); Dutch: Mak, Vonk \& Schriefers, 2002; English: Traxler, Moris \& Seely, 2002; inter alia). We report evidence supporting accessibility as a factor in the comprehension of relative clauses in Japanese.

Japanese is an SOV (subject-object-verb order) language and relative clauses are prenominal as they precede the modified noun as in (2).
(2) a. Subject relative clause (prenominal)
[ gap \(_{i}\) dansei-o mita] josei \({ }_{i}\)
man-acc saw woman
'the woman that saw the man'
b. Object relative clause (prenominal)
[dansei-ga \(g a p_{j}\) mita] josei \(_{j}\) man-nom saw woman
'the woman that the man saw'
The words and their order are exactly the same except for the case marker on the coargument, 'man' (i.e., the argument inside the embedded clause). When the coargument is an accusative object, extraction is from subject position; when it is a nominative subject, extraction is from object position. This similarity allows for a cleaner comparison between the constructions as it avoids comparing words with different parts of speech as is often the case in languages such as English.

As a universal constraint, the accessibility hierarchy predicts subject relatives to be easier to understand in Japanese. The prediction is supported by off-line judgments (e.g., difficulty ratings, Sheldon, 1976) and self-paced reading (Ishizuka, Nakatani \& Gibson, 2003; Miyamoto \& Nakamura, 2003; Ueno \& Garnsey, 2008, Experiment 1, also Experiment 2, for event related potential data). But these studies fail to address alternative explanations, especially those related to ambiguity factors, which we discuss later in this introduction. Moreover, some aspects of the preference are better explained by a new type of accessibility.

\section*{The object before subject preference}

Although often described as taking two arguments, transitive verbs are more closely associated with their direct objects than their subjects (Marantz, 1984; the verb-object bonding principle, Tomlin, 1986). A transitive verb can thus be represented as a function that takes the direct object as its only argument to yield a one-place predicate, which in turn takes the subject as its argument. Hence, the semantic role of the object is assigned by the verb, and the role of the subject is assigned by the verb-object compound (Marantz, 1984).

This subject-object asymmetry leads to a processing preference to assign the role of the object more locally than the role of the subject (the object before subject bias or ObS). Locality is assumed to be determined by clause structure. Relative clauses are statements about the modified noun, that is, they are functions that are applied onto the modified noun. Intuitively, we must understand what the statement means (e.g., determine the semantic roles of the coargument) to apply the statement to the modified noun.

In subject relative clauses, the object is in the same clause as the verb and receives its semantic role first. Next, the
object-verb complex assigns a semantic role to the subject (see the left-hand side of Figure 1). Therefore, the semantic role assigned to the modified noun is congruent with the statement that the relative clause makes about this noun.


Object RC


Figure 1: Order of semantic-role assignment in Japanese
In object relative clauses, the verb assigns a semantic role to the modified noun (its object) and only then the object-verb as a unit assigns a role to the coargument inside the relative clause (the subject; see Figure 1). The statement that the reladive clause makes about the modified noun is complex in that it is partly based on the modified noun itself and it does not match the order in which roles are assigned by the verb.

The ObS maintains the intuition from traditional accessibility that extraction position affects the naturalness of the relative clause as a statement about the modified noun. But in traditional accessibility the hierarchy is fixed (Keenan \& Comrie, 1977) and it incorrectly predicts that subject extraction is always favored (Nakamura \& Minamoto, 2013).

\section*{Ambiguity in relative clauses in Japanese}

There are no markers in Japanese that differentiate relative clauses from simple clauses (i.e., clauses without extractions such as adjunct, matrix or complement clauses). The relative clauses in (2) can be initially interpreted as simple clauses with an argument left implicit or dropped. Subject extraction is less likely to be affected by this type of ambiguity because the relative clause may be detected as soon as the coargument is read for the following reasons. First, the object coargument is an early indicator that there is no subject in the clause (although possible, a subject rarely follows an object; e.g., less than \(2 \%\) of accusative objects are followed by a subject in a newspaper corpus, Miyamoto \& Nakamura, 2005). Second, in a null context, the missing subject lacks a referent, therefore the relative-clause interpretation has been claimed to be favored so that a referent can be provided (Ishizuka, 2005).

In object relative clauses, the subject coargument provides little information about an upcoming object NP. It is only at the predicate (e.g., 'saw' in (2b)) that it is clear that the object is missing. In short, subject relatives may be detected as soon as the coargument NP is read, while object relatives may not be noticed until past the embedded predicate.

\section*{Alternative explanations}

One line of research has articulated a compelling alternative to accessibility models by proposing that working-memory factors such as decay and interference increase the difficulty in creating the dependency between the modified noun and the extraction position. In English, more material intervenes
a. Object relative clause: man-nom saw woman \(_{i}\) \(\uparrow\) \(\qquad\)
b. Subject relative clause: man-acc saw woman \(_{i}\) \(\uparrow\)

Figure 2: Linear decay in Japanese (horizontal lines mark the material that contributes to decay; vertical arrows indicate from where in the sentence retrieval occurs)
from the modified noun to the object position than to the subject position, therefore these proposals correctly predict more difficulty in object relative clauses such as (ib). \({ }^{1}\) The following is a summary of working-memory factors proposed in the literature and their predictions for Japanese.
A. Linear-span decay is based on the material in the sentence intervening between the modified noun and the extraction position (metrics based on number of words, King \& Just, 1991; new discourse referents, Gibson, 1998). For Japanese, greater difficulty is predicted for subject relatives than for object relatives (see Figure 2).
B. Temporal-span decay measures the material processed from the moment in time an element was inserted in the representation until the time it is retrieved (Lewis, Vasishth, \& van Dyke, 2006). A missing object is only detectable when the transitive verb is read, thus an empty object position is ceated after the verb is read and this is the point in time that is relevant for object relatives (Figure 3a). Figure 3b illustrates the observation that a sentence-initial object can indicate that the subject is missing from its canonical position, therefore an empty subject position may be created when the object is read and decay will start from that point in time.

Predictions may change depending on reactivation.
C. Reactivation of a constituent can reverse decay, facilitating retrieval (Lewis, Vasishth, \& van Dyke, 2006). Therefore, when a constituent was last reactivated, as opposed to when it was first inserted in the representation, may be a better indicator of temporal decay. In relative clauses in Japanese, both the extraction position and the coargument are reactivated at the embedded verb, but predictions depend on reactivation order. C1. Simultaneous reactivation of the two positions will lead


Figure 3: Temporal decay in Japanese (horizontal lines mark the material that contributes to decay; vertical arrows indicate the point in time the empty position is retrieved from)

\footnotetext{
\({ }^{1}\) If relative clauses and modified nouns are directly associated without the mediation of gaps (Pickering \& Barry, 1991), workingmemory factors predict no extraction advantage in languages with verb-final embedded clauses such as Dutch, German, Japanese.
}
them to have similar activation levels when the modified noun is read, therefore subject and object relatives are predicted to be equally easy to process (Kwon, Lee, Gordon, Kluender \& Polinsky, 2010). But reactivation at the verb is linguistically motivated (as arguments are linked to the verb), and linguistic constraints may affect reactivation order as in the following two alternatives.
C2. Canonical reactivation requires arguments to be reactivated in the canonical, most common, order of the language (i.e., subject and then object). Hence, the object position should have an advantage as the last one to be reactivated and object relatives should be easier than subject relatives.
C3. ObS-based reactivation assumes that the ObS is implemented as a constraint on reactivation. The object is reactivated first to be associated with the verb, then the O-V compound is applied onto the subject, which is thus reactivated last. Hence, subject relatives should be easier to process.
D. Interference from similarity is another factor related to working memory that has been discussed in the literature. Interference should occur when the case marker on the coargument and the modified noun are the same, increasing confusability (Kwon et al, 2010; also Gordon, Hendrick \& Johnson, 2001, for other types of similarity-related difficulty).

The predictions by working memory and accessibility may be obscured by other factors, especially ambiguity.
E. Expectation-based models predict that an interpretation is easier to understand if it is more likely to be generated during production (as measured by, for example, fragment completions). Results in English suggest that object relatives are less predictable because they allow more alternative interpretations as the sentence unfolds, thus requiring longer reading times to zero in on the intended interpretation (Gennari \& MacDonald, 2008; inter alia). But expectation may fail to predict the exact point of greatest difficulty in English (Grodner \& Gibson, 2005) and, in some formulations, expectation is not relevant in long-distance dependencies (Levy, 2008).

Supporting contexts may reduce ambiguity and allow working-memory effects to be measured more clearly. In one such study in Japanese, subject relatives were found to be harder to process than object relatives (Ishizuka, Nakatani \& Gibson, 2006; but see Kwon, Lee, Gordon, Kluender \& Polinsky, 2010, footnote 12, for communication from Edward Gibson reporting failure to replicate and retracting the result; see also Roland, Mauner, O'Meara \& Yun, 2012, for contextrelated factors facilitating object relatives in English, and Sato, 2011, who failed to replicate such effects in Japanese).

We report two self-paced reading experiments confirming the subject advantage for relative clauses in Japanese. The advantage is not explained by previous formulations of working-memory factors (Gibson, 1998; Gordon, Hendrick \& Johnson, 2001; King \& Just, 1991; Lewis, Vasishth, \& van Dyke, 2006; inter alia). Expectation-based models (Gennari \& MacDonald, 2008) are unlikely to be relevant either given the results of two fragment-completion questionnaires.

\section*{Experiment 1}

\section*{Method}

Participants Thirty native Japanese speakers undergraduates at the University of Tsukuba were paid to participate.
Materials Twelve pairs of items (see (3) for an example) were distributed into two lists according to a Latin Square design so that each sentence in a pair appeared in one list.
(3) a. Subject relative clause

Daikigyo-no keesya-o maekara utagatteita
company's manager-acc a while distrusting-was
kanryo-wa totemo mukutida.
bureaucrat-top very quiet-is
'The bureaucrat who has distrusted the company manager for a while is very quiet.'
b. Object relative clause

Daikigyo-no keesya-ga maekara utagatteita
company's manager-nom a while distrusting-was
kanryo-wa totemo mukutida.
bureaucrat-top very quiet-is
'The bureaucrat who the company manager has distrusted for a while is very quiet.'

Each sentence was shown one region at a time on a single line in a non-cumulative fashion using double-byte characters with the uniform-width Japanese font MS Mincho. The segmentation used is indicated with spaces in (3).

Norming 1 The predictions by expectation-based models (Gennari \& MacDonald, 2008) depend on production preferences. Therefore, 42 native Japanese speakers who did not participate in the reading-time study, were shown fragments ending at the embedded verb (e.g., 'distrusting-was' in (3)) and were asked to write completions for them. There were more object relatives than subject relatives (Median of the difference \(=1\); two-tailed Exact Wilcoxon Signed Rank, function wilcoxsign_test in the package coin, R Development Core Team, 2009; Wilcoxon, for short: \(Z_{1}=3.47, P<.001\); \(Z_{2}=2.86, P<.01\); see Ueno \& Garnsey, 2008, for similar results). Contrary to previous reports, the object advantage held even when the comparison was restricted to relative clauses with animate modified nouns, which are the most similar to the items in the reading experiment (Median \(=1\); Wilcoxon: \(Z_{1}=2.41, P<.05 ; Z_{2}=2.59, P<.05\) ), probably because we chose embedded verbs that were biased towards animate objects. Therefore, expectation-based models should predict an object advantage in the reading time data.

Norming 2 To ensure that the two types of relative clauses are equally plausible, simple transitive sentences were created by placing the modified noun in the intended extraction position in the embedded clause. Thirty-two native Japanese speakers, who did not participate in any of the other studies, rated each sentence on a 5-point scale (1 as natural and 5 as
strange). No difference was detected (subject: 1.74; object: 1.66; Wilcoxon: \(\mathrm{Zs}<1\) ).

\section*{Results and discussion}

In the crucial region, the modified noun 'bureaucrat', subject relatives were marginally faster than object relatives \(\left(F_{1}(1,28)=2.96, M S e=92,790, P=.096 ; F_{2}(1,10)=4.94\right.\), \(M S e=15,300, P=.051)\). A spillover was observed in the following region, which contained the same words across the two conditions, and the difference was reliable at this point \(\left(F_{1}(1,28)=11.19, M S e=12,933, P<.01 ; F_{2}(1,10)=6.63\right.\), \(M S e=10,210, P<.05)\).

The results suggest that subject relative clauses are easier to process than object relative clauses, replicating previous results (Ishizuka, Nakatani \& Gibson, 2003; Miyamoto \& Nakamura, 2003; Sheldon, 1976; Ueno \& Garnsey, 2008). This is compatible with accessibility and the ObS in particular, but not with previous formulations of working-memory factors (e.g., Gibson, 1998; King \& Just, 1991).

Temporal-span models (Lewis, Vasishth, \& van Dyke, 2006) can account for the results if reactivation at the embedded predicate obeys the ObS. To date, we know of no other formulation of working-memory factors that has the potential to provide an explanation for the subject preference in the Japanese constructions discussed in this paper. However, even ObS-based reactivation fails to account for other types of relative clauses in Japanese, such as those with the coargument dropped (Nakamura \& Miyamoto, 2013).

Difficulty in object relatives was observed even though they were more expected according to Norming 1 (contra Gennari \& MacDonald, 2008).

\section*{Experiment 2}

The role of the modified noun in the outer clause (usually the matrix clause) has also been implicated in the comprehension of relative clauses. Parallelism predicts facilitation when the extraction site and the modified noun share properties such as grammatical role (Sheldon, 1976). Alternatively, facilitation occurs when the two positions share the same case marker (Sauerland \& Gibson, 1998). This type of case-marking attraction is grammaticalized in some languages so that the case marker on the modified noun can affect the marker on the relative pronoun or vice-versa (e.g., Ancient Greek, Persian; see Keenan, 1981, for examples). We controlled for these factors in this experiment.

\section*{Method}

Participants Thirty-two native speakers of Japanese at the Future University Hakodate, who had not participated in the other studies, were paid to participate in the experiment.
Materials Twelve pairs of items were used with the same procedure as in the first experiment. The following is an example pair.
(4) a. Subject relative clause \(\begin{array}{llll}\text { Kinjono } & \text { obaasan-o } & \text { basutei-made } & \text { miokutta } \\ \text { neighbor-gen } & \text { woman-acc } & \text { bus-stop-to } & \text { accompanied } \\ \text { onnanoko-ni-wa } & \text { sanpochuuno } & \text { inu-ga } & \text { jareteita. } \\ \text { girl-dat-top } & \text { walk } & \text { dog-nom } & \text { frolicking-was }\end{array}\)
'As for the girl who accompanied the woman from the neighborhood to the bus stop, a dog taken for a walk was frolicking around (her).'
b. Object relative clause

Kinjono obaasan-ga basutei-made miokutta neighbor-gen woman-nom bus-stop-to accompanied onnanoko-ni-wa sanpochuuno inu-ga jareteita. girl-dat-top walk dog-nom frolicking-was
'As for the girl who the woman from the neighborhood accompanied to the bus stop, a dog taken for a walk was frolicking around (her).'

Following previous studies (Ishizuka, Nakatani \& Gibson, 2003; Ueno \& Garnsey, 2008), the modified noun ('girl' in (4)) was marked with the dative marker \(n i\) and the topic marker wa. The two combined particles are not shared with either of the extraction positions (the nominative-subject or the accusative-object positions), therefore there should be no interference from attraction (Keenan, 1981; Sauerland \& Gibson, 1998). A sentence-initial animate noun marked with ni is usually interpreted as an indirect object, therefore a strict interpretation of parallelism (requiring the exact same grammatical role for the extraction position in the relative clause and for the modified noun in the matrix clause; Sheldon, 1976) would also predict no difference between the two types of relative clauses. But to the extent that \(n i\)-marked objects are more similar to direct objects than to subjects, \({ }^{2}\) a looser version of parallelism may favor object relative clauses.

Perspective shift elaborates on parallelism by considering whose point of view is adopted as the representation for the event is built. According to work on Hungarian (MacWhinney \& Pleh, 1988), perspective depends on whether the language is subject-prominent (SP; i.e., languages in which sentences are based on the relation between subject and predicate; e.g., Indo-European languages) or topic-prominent (TP; in which sentences are based on the relation between topic and comment; e.g., Chinese; Li \& Thompson, 1976). In TP languages like Hungarian, parallelism is claimed to hold for topics. In particular, a sentence-initial object is the topic in Hungarian, therefore its perspective is adopted, and because of parallelism, there is a preference for relative clauses to be object extracted when they modify a topicalized object (MacWhinney \& Pleh, 1988).

The test sentences in (4) are TP constructions because the dative object 'girl' is topicalized with the marker \(w a\); there-

\footnotetext{
\({ }^{2}\) For example, the object of verbs such as au 'meet', intabyusuru 'interview', denwasuru 'call/phone' is a direct object in English, but it is marked with the dative \(n i\) in Japanese.
}
fore, similar to Hungarian, perspective shift should favor object extraction. \({ }^{3}\)

Norming 3 A new group of 46 native speakers of Japanese participated. As in Norming 1, there were more object relative-clause completions ( \(66.3 \%\) ) than subject relative clauses (52.45\%; Wilcoxon: \(Z_{1}=3.04, P<.01 ; Z_{2}=\) \(2.36, P<.05)\). The same trends were observed when counts were restricted to relative clauses with head nouns depicting humans (object relatives: \(61.7 \%\); subject relatives: \(51.1 \%\); Wilcoxon: \(Z_{1}=2.73, P<.01 ; Z_{2}=2.05, P<.05\) ).

Norming 4 As in Norming 2, no plausibility difference between the subject condition (1.9) and the object condition (1.73; where 1 was natural and 5 was strange; Wilcoxon: \(P \mathrm{~s}>.2\) ) according to 16 native Japanese speakers, who had not participated in any of the other studies reported.

\section*{Results and discussion}

In the critical region ('girl' in (4)), there was a numerical advantage for subject relatives. The difference was reliable in the following region, which contained the same words across the two conditions ( \(P<.05\); function lmer in the package lme4, R Development Core Team, 2009). Moreover, the coargument in five items was a proper name while in the remaining seven items it was a common noun. Because the modified noun was always a common noun, proper names should decrease confusability and facilitate comprehension, but there was no interaction when type of noun was included in the analysis ( \(P>.8\); contra Gordon, Hendrick \& Johnson, 2001).

The results replicated the advantage for subject relatives over object relatives. Because the modified noun was marked dative-topic, factors such as parallelism (Sheldon, 1976), attraction (Keenan, 1981; Sauerland \& Gibson, 1998) and perspective shift (MacWhinney \& Pleh, 1988; Mitsugi, MacWhinney \& Shirai, 2010) incorrectly predict an object preference or no preference between the two types of relative clauses. The subject advantage is not compatible with expectation-based models either (Gennari \& MacDonald, 2008) given that there was an advantage for object relatives in the completion results of Norming 3.

\section*{General discussion}

We can classify languages into four types according to the positions of the relative clause (prenominal or postnominal) and the object ( OV or VO ) as follows: postnominal/SVO (e.g., English, French), postnominal/SOV (Dutch, German), prenominal/SVO (Chinese), and prenominal/SOV (Japanese, Korean, Turkish). Previous discussions of working-memory

\footnotetext{
\({ }^{3}\) Before the topicalized noun is read, the sentences may be mistaken for SP constructions. It has been argued that perspective in SP relatives may remain unspecified after an accusative NP is read, therefore no shift occurs and subject extraction may be relatively easy (Mitsugi, MacWhinney \& Shirai, 2010). However, it is unclear what exactly an unspecified perspective entails for the kind of representation assumed in perspective shift. Even the "neutral" point of view of an uninvolved spectator or the speaker would still require a shift when the perspective of the modified noun is adopted.
}
factors correctly predict the subject advantage in both types of postnominal languages as well as the object advantage in prenominal/SVO languages (Hsiao \& Gibson, 2003; Lin \& Garnsey, 2011; inter alia). The exception is prenominal/SOV languages, for which working-memory factors predict no difference or an object advantage. The results reported confirm the exceptional status of these languages even when ambiguity and expectation are taken into consideration.

We raised the possibility that ObS-based reactivation may explain the subject advantage in Japanese. It is also possible that simultaneous reactivation at the embedded verb cancels out any working-memory difference (Kwon et al, 2010), therefore only accessibility has an effect in Japanese. But there are at least two other alternatives that will also need future research. The first alternative is that the workingmemory load difference is small in Japanese-type languages because linear decay involves only one NP (see Figures 2ab). If so, working memory may not be enough to explain the subject advantage in Dutch and German, for which the difference between subject and object extraction is also of one NP only. This would reinforce the need for another factor such as the ObS to complement working-memory factors in order to explain the subject advantage in these languages.

Another possibility is that closure may be relevant because it flushes out verbatim material out of working memory. Memory load is likely to decrease after closure is performed on a phrase. Moreover, the load may vary across languages depending on the timing of closure. In Japanese, a consistently head-final language, closure is likely to be performed immediately at the end of the phrase (for example at the verb of an embedded clause). In contrast, head-initial languages do not have a marker to indicate the end of the phrase; consequently, closure may be delayed and memory load linger in languages such as English. Mixed languages such as Chinese may also fail to generate closure consistently at phrase end especially if the marker is a short functional word (e.g., de for relative clauses) that can be easily skipped during reading.

In sum, a second factor apart from working-memory constraints is needed to explain the subject advantage in prenominal/SOV relative clauses. This may provide an opportunity to better understand how memory use is affected by processes such as closure during language comprehension.

\section*{References}

Comrie, B. (2007). The acquisition of relative clauses in relation to language typology. Studies in Second Language Acquisition, 29, 301-309.
Gennari, S. P., \& MacDonald, M. C. (2008). Semantic indeterminacy in object relative clauses. Journal of Memory and Language, 58, 161-187.
Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. Cognition, 68, 1-76.
Gordon, P. C., Hendrick, R., \& Johnson, M. (2001). Memory interference during language processing. Journal of Experimental Psychology: Learning, Memory and Cogni-
tion, 27, 1411-1423.
Grodner, D. J., \& Gibson, E. A. F. (2005). Consequences of the serial nature of linguistic input for sentential complexity. Cognitive Science, 29, 261-291.
Hsiao, F., \& Gibson, E. (2003). Processing relative clauses in Chinese. Cognition, 90, 3-27.
Ishizuka, T. (2005). Processing relative clauses in Japanese. In Okabe and Nielsen (Eds.), Working Papers in Psycholinguistics, 2 (pp. 135-157), UCLA Working Papers in Linguistics 13.
Ishizuka, T., Nakatani, K., \& Gibson, E. (2003). Relative clause extraction complexity in Japanese. Poster presented at the 16th Annual CUNY Conference on Human Sentence Processing, Boston, March 2003.
Ishizuka, T., Nakatani, K., \& Gibson, E. (2006). Processing Japanese relative clauses in context. Paper presented at the 19th Annual CUNY Conference on Human Sentence Processing, NY, March 2006.
Keenan, E. L. (1981). Relative clauses. In B. Comrie (Ed.), Language Universals and Linguistic Typology (pp. 138164). Oxford: Basil Blackwell.

Keenan, E. L., \& Comrie, B. (1977). Noun phrase accessibility and Universal Grammar. Linguistic Inquiry, 8, 63-99.
King, J., \& Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. Journal of Memory and Language, 30, 580-602.
Kwon, N., Lee, Y., Gordon, P. C., Kluender, R., \& Polinsky, M. (2010). Cognitive and linguistic factors affecting subject/object asymmetry: An eye-tracking study of prenominal relative clauses in Korean. Language, 86, 546-582.
Levy, R. (2008). Expectation-based syntactic comprehension. Cognition, 106, 1126-1177.
Lewis, R. L., Vasishth, S., \& Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. Trends in Cognitive Science, 10, 447-454.
Li, C. N., \& Thompson, S. A. (1976). Subject and topic: a new typology of language. In C. N. Li (Ed.), Subject and Topic (pp. 457-61). London/New York: Academic Press.
Lin, Y., \& Garnsey, S. (2011). Animacy and the resolution of temporary ambiguity in relative clause comprehension in Mandarin. In H. Yamashita, Y. Hirose, \& J. L. Packard (Eds.), Processing and Producing Head-final Structures (pp. 241-275). Dordrecht: Springer.
MacWhinney, B., \& Pleh, C. (1988). The processing of restrictive relative clauses in Hungarian. Cognition, 29, 95141.

Mak, W. M., Vonk, W., \& Schriefers, H. (2002). The influence of animacy on relative clause processing. Journal of Memory and Language, 47, 50-68.
Marantz, A. (1984). On the Nature of Grammatical Relations. Cambridge, Mass.: MIT Press.
Mitsugi, S., MacWhinney, B., \& Shirai, Y. (2010). CueBased Processing of Relative Clauses in L2 Japanese. In M. T. Prior, Y. Watanabe, and S.-K. Lee (Eds.), Selected Proceedings of the 2008 Second Language Research Fo-
rum: Exploring SLA Perspectives, Positions, and Practices (pp. 123-138). Somerville, Mass.: Cascadilla Proceedings Project,
Miyamoto, E. T., \& Nakamura, M. (2003). Subject/object asymmetries in the processing of relative clauses in Japanese. In G. Garding \& M. Tsujimura (Eds.), Proceedings of the 22nd West Coast Conference on Formal Linguistics (pp. 342-355). Somerville, MA: Cascadilla Press.
Miyamoto, E. T., \& Nakamura, M. (2005). Unscrambling some misconceptions: a comment on Koizumi and Tamaoka (2004). Gengo Kenkyu, 128, 113-129.
Nakamura, M. \& Miyamoto, E. T. (2013). The object before subject bias and the processing of double-gap relative clauses in Japanese. Language and Cognitive Processes, 28, 303-334.
Pickering, M., \& Barry, G. (1991). Sentence processing without empty categories. Language and Cognitive Processes, 6, 229-259.
R development core team. (2009). R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing, http://www.R-project.org.
Roland, D., Mauner, G., O Meara, C., \& Yun, H. (2012). Discourse expectations and relative clause processing. Journal of Memory and Language, 66, 479-508.
Sato, A. (2011). Nihongo kankeisetuno syori fukao ketteisuru yoinno kento - koopasuni okeru siyo hindono eikyoo chuusinni (An Investigation of Factors that Determine Processing Load in Japanese Relative Clauses - With a Focus on Frequency Use in Corpora) (In Japanese). Unpublished doctoral dissertation, Hiroshima University, Higashi Hiroshima, Japan.
Sauerland, U., \& Gibson, E. (1998). Case matching in relative clause attachment. Paper presented at the 11th Annual CUNY Sentence Processing Conference, Rutgers University, New Brunswick, NJ, March 1998.
Schriefers H., Friederici A. D., \& Kuhn K. (1995). The processing of locally ambiguous relative clauses in German. Journal of Memory and Language, 34, 499-520.
Sheldon, A. (1976). Speaker's intuition about the complexity of relative clauses in Japanese and English. In S. S. Mufwene, C. A. Walker and S. B. Sanford (Eds.), Papers from the Twelfth Regional Meeting of the Chicago Linguistic Society (pp. 558-567). Chicago: Department of Linguistics: University of Chicago.
Sheldon, A. (1977). On strategies for processing relative clauses: A comparison of children and adults. Journal of Psycholinguistic Research, 6, 305-318.
Tomlin, R. S. (1986). Basic Word Order - Function Principles. London: Croom Helm.
Traxler, M. J., Morris, R. K., \& Seely, R. E. (2002). Processing subject and object relative clauses: evidence from eye movements. Journal of Memory and Language, 47, 69-90.
Ueno, M. \& Garnsey, S.M. (2008). An ERP study of the processing of subject and object relative clauses in Japanese. Language and Cognitive Processes, 23, 646-688.

\title{
The facilitatory role of sound symbolism in infant word learning
}

\author{
Michiko Miyazaki \({ }^{1}\) (myzk@lab.tamagawa.ac.jp), Shohei Hidaka \({ }^{2}\) (shhidaka@jaist.ac.jp), Mutsumi Imai \({ }^{3}\) (imai@sfc.keio.ac.jp), H. Henny Yeung \({ }^{4}\) (henny.yeung@ parisdescartes.fr), Katerina Kantartzis \({ }^{5}\) (k.f.kantartzis@bham.ac.uk), Hiroyuki Okada \({ }^{1}\) (h.okada@eng.tamagawa.ac.jp), Sotaro Kita \({ }^{5}\) (S.Kita@bham.ac.uk) \\ \({ }^{1}\) Brain Science Institute, Tamagawa University, JAPAN, \\ \({ }^{2}\) School of Knowledge Science, Japan Advanced Institute of Science and Technology, JAPAN \\ \({ }^{3}\) Department of Environment and Information Studies, Keio University, JAPAN \\ \({ }^{4}\) Université Paris Descartes, \& CNRS (Laboratoire Psychologie de la Perception, UMR 8158), FRANCE \\ \({ }^{5}\) School of Psychology, University of Birmingham, UK
}

\begin{abstract}
Sound symbolism or the nonarbitrary link between language sound and meaning are commonly found across many languages of the world. A well-known example is the association between rounded vs. angular shapes and labels (i.e., the Bouba-kiki effect by Köhler, 1929/1947). Previous research has shown that sound symbolic words play facilitative role for preschool children's novel verb learning (Imai, Kita, Nagumo \& Okada, 2008; Kantartiz, Imai \& Kita, 2011), helping children identify what aspects of motion events should be mapped to verbs. In this research, we explore whether sound symbolism may facilitate language learning in human infants who have just begun to learn word meanings. Sound symbolism may be a useful cue particularly at the earliest stages of word learning, because this cue seems to be available without needing prior word learning experience (Gogate \& Hollich, 2010). Using a habituation paradigm and a Bayesian model-based analysis, we demonstrated that 14-month-old infants could detect Köhlertype (1947) shape-sound symbolism, and could use this sensitivity in their effort to establish the word-referent association.
\end{abstract}

Keywords: Sound symbolism; Word learning; Iconicity of language; Origin of language; Multisensory mapping; Bayesian analysis

\section*{Introduction}

Traditional linguistics has long assumed that links between a word's form and meaning are arbitrary (de Saussure, 1916/1983). However, words whose forms are motivated by their meanings (i.e., sound symbolic words) are commonly found across many languages of the world. For example, bump and thump sound like what they mean: an event with an abrupt end (e.g., Firth, 1935). Several languages even have large grammatically defined lexical classes of sound symbolic words (i.e., "ideophones," "expressives," or "mimetics") (Voeltz \& Kilian-Hatz, 2001; Kita, 1997). A well-known example of sound symbolism is the association between rounded vs. angular shapes and labels (Köhler, 1929/1947; see also Ramachandran \& Hubbard, 2001). Given a forced choice, adults and children from different languages (e.g., German, English, and Swahili) much prefer to label rounded objects bouba (or maluma) and angular objects kiki (or takete) (e.g., Davis, 1961; Holland \& Wertheimer, 1964).

Recently, evidence for sensitivity to sound symbolism in young infants is emerging (Penã, Mehler, \& Nespor, 2011; Ozturk, Krehm, \& Vouloumanos, 2012). Ozturk et al. (2012) demonstrated that even 4-month-olds are sensitive to the bouba-kiki sound-shape mappings.
An interesting question is whether sensitivity to sound symbolism is useful for language development. Previous research has shown that sound symbolic words play facilitative role for preschool children's novel word learning, helping children identify what aspects of the world should be mapped to verbs (Quine, 1969, Markman, 1989). For example, Maurer et al. (2006) demonstrated that 2.5-yearold children are likely to match rounded sounding labels with simple, rounded objects and jagged sounding labels with pointy objects (Mondloch \& Maurer, 2004; Maurer, Pathman, \& Mondloch, 2006). In addition, sound symbolism also helps 3-year-olds infer meanings of novel verbs, which are generally more difficult than object names (Imai et al., 2008; Kantartzis, Imai, \& Kita, 2011). Thus, sound symbolism is a good candidate for mappings from word forms to referents in the early stages of development. However, it is not clear whether such sensitivity to sound symbolism in young infants is used for word learning at the initial stage of language development.

Before toddlerhood, it is unknown how the youngest word learners can "break into" the process of mapping words to their referents. For infants just starting to learn words, the induction problem is much harder (Hollich, et al., 2007; Spiegel \& Halberda, 2011), and these infants likely rely more on perceptual regularities, due to limited memory and information processing abilities (Gogate \& Hollich, 2010). In fact, infants between 10- and 17-months of age find it much more difficult to utilize word-learning strategies used by older infants to infer novel word meanings (Woodward \& Hoyne, 1999). At this age, infants may use more perceptual strategies to map words, which can often lead them to the wrong referent (e.g., Hollich et al., 2007).

As with toddlers, however, sound symbolic words may be easier to learn, and these early links may help to scaffold mappings from word forms to referents in the early stages of development. In the present study, we examine whether sound symbolic links can provide early word-referent cues. Specifically, we asked if 14 month-old infants could utilize senseitivity to Köhler-type (1947) shape-sound symbolism
to establish association between a word and a referent object. Here we chose infants at 14 months of age, who are just old enough to learn new words in laboratory tasks, but are at an age when learning is still very fragile, especially when trying to process the precise phonological forms of words (Stager \& Werker, 1997).
We hypothesized that 14-month-olds detect sound-shape correspondence, and that this ability helps these infants make mappings from word forms to their referents. Specifically, we taught Japanese-learning infants two word labels, and then tested whether they encoded these labels in a preferential looking procedure. Half the infants were taught two sound symbolic labels (as rated by adults); half were taught the mismatching labels. We predicted that those in the sound symbolically matching condition would learn labels more easily than those in the mismatching condition.
This report features a Bayesian model-based data analysis. Infants looking behavior-particularly in a preferential looking paradigm, in which infants must compare two objects to make a decision-is very complex: It often shifts dynamically instead of staying stable during a trial. Also, because of the dynamic nature of looking, it is difficult to determine the most appropriate time window prior to the analysis. Furthermore, infants' looking time is "noisy", affected largely by looking profiles inherent in individual infants (e.g., a preference to look at a certain location on the monitor, a preference to look at a particular object, the likelihood to shift eye-gaze, etc.). Ignoring these idiosyncratic looking biases can weaken statistical power. Despite this problem, researchers have traditionally used ANOVA to examine the effect, where infants' looking time is averaged over a pre-set time window and individualbased looking biases are treated as experimental noise. In this research, we employed a Bayesian approach to analyze the data, which offers a new method of data analysis that is more adaptive to complex and dynamic nature of infants' looking behavior (see the Bayesian Data Analysis section for more details).

\section*{Method}

\section*{Participants}

Participants were thirty-four full term, monolingual Japanese 14-month-olds ( \(M=14 ; 16\), range \(13 ; 27-15 ; 9,22\) males). Infants were randomly assigned to either the match or mismatch condition. An additional 11 infants were excluded from data analyses due to experimental error ( \(\mathrm{n}=\) \(1)\), or fussiness during the experiment \((\mathrm{n}=10)\).

\section*{Apparatus}

A black cloth curtain surrounded a 21-inch display where visual stimuli appeared, and a digital video camera was hidden below the screen, relaying video of the infant to the control area. Video was also recorded ( 29.97 fps ) for offline coding of looking.

\section*{Stimulus materials}

Target stimuli consisted of a round versus a spiky shape, as well as audio recordings of two novel words, kipi and moma (Figure 1). Stimuli were constructed on the basis of a pilot study, where we first chose consonants ( \(\mathrm{m}, \mathrm{l}, \mathrm{n}, \mathrm{p}, \mathrm{k}\) ), and vowels (a, o, i) that are related to smooth and jagged shapes according to previous research on sound symbolism (Köhler, 1929/1947; Maurer, et al., 2006; Westbury, 2005). Moma and kipi were selected because our adults frequently chose these as the best match to the smooth/round and spiky/jagged shapes, respectively, and both were nonwords in Japanese. The two shapes were colored to make them more interesting to infants. Colors were chosen not to affect the shape sound symbolism.

During a habituation phase, infants were presented with two pairs of audio-visual stimuli: in the match condition, kipi - spiky, object and moma - round object; in the mismatch condition, kipi - round object, moma - spiky object. Infants were presented with filler stimuli, consisting of colored drawings of a ball, a banana, a car, and a picture book, paired with audio recordings of the corresponding word in Japanese. These items were chosen based on normative data from the Japanese MacArthur-Bates Communicative Development Inventory (J-MCDI) (Ogura \& Watamaki, 2004). A female Japanese speaker recorded the target and filler words, along with all other speech used in the experiment, in an infant-directed speech register.

\section*{Procedure}

Infants were tested individually in a quiet room on their parent's lap, positioned 60 cm from the display. Pretest, habituation, and test phases were presented, each of which contained several trials separated by a short attention-getting movie. Parents were instructed to keep their eyes closed, and also asked to complete the J-MCDI.
Pretest phase Here we presented 4 familiarization and 2 referential trials in random order. Familiarization trials showed side-by-side displays of either filler or target objects. Each trial lasted 8 seconds and was accompanied by, "Mite! Mite!" (Look! Look!). Two referential trials were included to enhance infants' understanding of the referential nature of the labels (Fennell \& Waxman, 2010), and here a single familiar object slowly moved (either up and down, or right and left) on the display, accompanied by the corresponding label in isolation and then in a carrier sentence (e.g., "Kuruma! Kuruma-wo mite!" [Car! Look at the car!]).
Habituation phase The habituation phase consisted of a pseudo-randomly ordered series of trials such that each word-object pairing appeared twice in every block of four trials. In each trial, a single target object moved slowly from right to left in order to maintain infants' attention (Werker et al., 1998), while one of the target words was paired with it (Figure 1). Trials lasted a maximum of 16 seconds and were accompanied by 13 tokens of a target word, each spoken with a different intonation pattern. The habituation criterion was set to a maximum of 24 trials, or a decrease of \(65 \%\) in looking to the longest previous block of 4 trials (Stager \&


Figure 1: Design and Stimuli
Werker, 1997; Yoshida, Fennell, Swingley, \& Werker, 2009).

Test phase The test phase consisted of four filler and four target trials that contained two objects side-by-side, counterbalanced for side over the testing phase. Two filler trials were always shown first, providing infants the opportunity to see familiar objects in the test procedure. The third and fourth trials were target trials, and the remaining four trials alternated between filler ( F ) and target ( T ) trials. All trials lasted 8 seconds, and began with a 3000 ms silent baseline period to measure visual preferences for the objects, followed by a phrase asking infants to look at the correct object (i.e., "X! X wa docchi?" [X! Which is the X?], where X stands for a filler or target word).

\section*{Coding}

Two blind coders classified each video frame as a look to the right, to the left, an ambiguous look, or no look. Fifteen percent of the all samples were coded twice to obtain intercoder reliability. The inter-coder reliability was high, \(\kappa\) \(=.826\).

\section*{Bayesian data analysis}

Data were analyzed using a Bayesian approach. In analyzing infants’ looking behavior, it is customarily assumed that it takes a certain time for infants to process auditory stimuli and/or move their eyes to fixate the target stimulus (e.g., Swingley \& Aslin, 2002). Considering the lag due to auditory and visual processing, a pre-determined time window is set, and infants' looking is averaged throughout this window. The group means are then compared using a linear model (typically a t-test or ANOVA). However, as mentioned earlier, the assumptions in traditional linear models often pose limitations for analyzing infants' complex looking behavior. First, infants' looking to the target often shifts dynamically instead of staying stable during the time course, and hence polynomial functions appear to fit better than a linear function. Second, because
infants' looking behavior may be affected by the nature of the stimuli and experimental settings, it is difficult to determine the most appropriate time window prior to the analysis. Third, although substantial individual difference is expected in infants' looking profiles (e.g., a preference to look at a certain location on the monitor, a preference to look at a particular object, the likelihood to shift eye-gaze, etc.), these response biases are simply treated as experimental noise, which weakens statistical power to detect the experimental effect of interest.

In the current analysis-to rectify these limitations in traditional linear models-we performed a Bayesian modelbased analysis based on Yurovsky, Hidaka, and Wu, 2012. In a Bayesian framework, a hypothetical relationship among a set of experimental factors (e.g., training effects), subsidiary factors (e.g., object-specific bias) and potential patterns of behaviors (e.g., looking time) are expressed as a statistical model with a set of parameters which is estimated through model fitting to a given dataset. In the present analysis, there are two major sets of parameters. The first set includes preference parameters for factors that could potentially affect the looking of a particular AOI at a given moment: individual infants' location preference, object preference, preference to look at the trained object, preference to look at the object that was sound symbolically matching to the label. The second set consisted of group parameters, which classify participants in such a way that within-group similarity in infants' response patterns and across-group differences are simultaneously maximized.

\section*{Results}

Looking data from each of the critical video frames were classified as a look to the left AOI, to the right AOI, or to a 'no-look' AOI. We analyzed infants' looking times (as frame-by-frame counts) as a function of five factors: a location-specific preference, an object-specific preference, a "correct" (trained) object preference, a sound symbolism preference, and the interaction between training and the sound symbolic match. The location-specific preference is defined as a bias to look toward the left or right AOI, relative to the preference for the no-look AOI for each infant independent of the match/mismatch condition. The object preference is a bias to look toward a particular object compared to the no-look AOI (with some other objects). The correct object preference is a preference to look at the object with which the label was associated during the habituation phase. The sound symbolism preference is a tendency to look at the sound symbolically matching object (to the label heard in the trial) after the onset of the speech (i.e., the preference for the spiky object after the onset of speech "kipi," or the preference for the round object after "moma"). Through the process of model fitting, we estimated a set of parameters for all of the five factors above, but we focus only on the three experiment-relevant factors, i.e., training, sound-symbolic match, and training-soundsymbolism interaction factors here.

We first analyzed the effect of the training and that of sound symbolism separately against the baseline preference, to see whether training and/or sound symbolism alone affected infants' looking behavior after the speech onset. For this purpose, we performed a series of models using (Bayesian) hypothesis testing. In Bayesian hypothesis testing, each model specifies a probability of a hypothesis to reproduce the current looking dataset. We tested a contribution of a particular factor, by evaluating the goodness of fit for two models-one with the target factor and one without-by a Bayes factor (Jeffreys, 1961; See also Wagenmakers et al. (2010) for a review in psychological studies). "The Bayes factor (BF) \(X\) of Model \(A\) to \(B\) given a dataset" indicates that the odds ratio for Model \(A\) to reproduce the dataset is \(X\) times higher than that for Model \(B\) to do so under even prior probability for each model \({ }^{1}\). In the present study, we considered a Bayes factor larger than 30 (or equivalently log-Bayes factor larger than 3.4) to be strong evidence in support for the Model A over Model B, based on Jeffreys' (1961) criterion.

Six models were evaluated as shown in Table 1. The first three models-P1-full, P2-full, and P3-full models--were full models in which all of the three experimental factors were included. These three models assumed different levels of complexity in their functions of looking probability. The first order polynomial function was assumed for the P1-full model, and the second- and third-order polynomial functions were assumed for the P2-full and P3-full models, respectively. We then compared the three models to determine the optimal level of complexity of the function in the model. The analysis on the Bayes factors indicates that the middle degree of complexity ( P 2 -full) is strongly favored over both the relatively simple (P2-full to P1-full: 27.5) and complex model (P2-full to P3-full: 247.4). We therefore employed the P2-full model as the baseline model, against which each of the subset models was compared.

To evaluate the effect of the three experimental factors, we calculated the Bayes factors for the three additional models in Table 1, i.e., P2-NoInt., P2-NoSS, and P2-NoTr against the P2-full baseline model (see Table 1 for results and the abbreviations). The analysis of the Bayes factors suggests that the P2-full model was strongly favored over the P2-NoInt (109.6), P2-NoSS (119.6), and the P2-NoTr (92.5). These results indicate that all effects of sound symbolism, training and the interaction between training and sound symbolism significantly contributed to the model fit.

This suggest that infants tended to look at the trained object regardless of whether this object was sound symbolically matching or not. Furthermore, sound symbolism affected infants' looking, regardless of whether infants were trained on a sound symbolically matching object.

\footnotetext{
\({ }^{1}\) According to Jeffreys (1961), BF from 3 to 10 (log-BF from 1.1 to 2.3 ) indicates "substantial" evidence, BF from 10 to 30 (logBF from 2.3 to 3.4) indicates "strong" one, and BF from 30 (logBF greater than 3.4) indicates "very strong" one.
}

Table 1: Summary of the hypotheses testing on the six models.
\begin{tabular}{|l|c|c|c|c|c|c|r|}
\hline \multirow{2}{*}{ Models } & \multicolumn{5}{|c|}{ Models } & \multicolumn{2}{c|}{ Results } \\
\cline { 2 - 7 } & \#P & Tr & SS & TS & Ply & \#G & log-BF \\
\hline \hline P2-full* & 10 & Y & Y & Y & 2 & 7 & 0 \\
P1-full & 7 & Y & Y & Y & 1 & 7 & 27.5 \\
P3-full & 13 & Y & Y & Y & 3 & 6 & 247.4 \\
P2-NoInt. & 8 & Y & Y & - & 2 & 7 & 109.6 \\
P2-NoSS. & 8 & Y & - & Y & 2 & 7 & 119.6 \\
P2-NoTr. & 8 & - & Y & Y & 2 & 8 & 92.5 \\
\hline
\end{tabular}

The abbreviations are as follows: \#P: the number of parameters for each cluster, \(\mathrm{Tr}, \mathrm{SS}\), and TS: whether the model contains preference parameters for trained object, sound symbolism match object, and the interaction between them ("Y" if the model has), Ply: the order of polynomial functions of the looking time courses , \#G: the estimated number of groups of infants, Log-BF: the log-Bayes factor of the P2-full model reproducing the data relative to the compared model. Pn-full: a full-factored model with the \(n\)-th order polynomial functions, P2-NoInt: a model without interaction between training and sound symbolic match, P2-NoSS: a model without the effects of sound-symbolic match, P2-NoTr: a model without the effects of training, *: the best model.

\section*{Discussion}

The present study aimed to clarify the facilitating role of sound symbolism in novel word learning in 14-month-old infants. Although the infants at this age sometimes show difficulty for matching words to their correct referents due to their limited cognitive ability, 14-month-olds could utilize sound-symbolic correspondences between speech sounds and object in this study. By Bayesian analysis, the effects of sound symbolism, training, and the interaction between sound symbolism and training all significantly contributed to the infants' looking behavior.

The current Bayesian data analysis shed light on the issue of how the youngest word learners break into the incredibly difficult process of mapping words to their references. One of the great advantages of Bayesian analysis is that it could wash out the critical experimental effects when traditional averaged analysis are treated as "noise," can be considered. By classifying participants with similar looking patterns into clusters and estimating group parameters, fine-grained response characteristics for a particular subgroup of infants.

The fact that significant contribution of factors sound symbolism, training, and the interaction between sound symbolism and training revealed in early word leaning stage may provide an important clue for solving the difficulty for mapping word to their referents. Sound symbolism may
allow infants to anchor speech to meaning, which in turn helps them obtain "referential insight"--- the insight that language sounds are symbols that represent concepts (Gogate \& Hollich, 2010). Once infants get into soundsymbolically based systems relating surface structure to meaning, they may be able to use this early knowledge to bootstrap themselves to more abstract meanings, needing direct perceptual anchors less and less, perhaps reflecting similar trajectories in language evolution (Kita, Kantartzis, \& Imai, 2010).

\section*{Acknowledgments}

We would like to thank Janet Werker for insightful discussion and advice in designing the experiment. Mamiko Arata, Noburo Saji and Junko Kanero all contributed to the creation and selection of stimuli. Asako Isawa, Kaoru Takata and Satomi Kanbara contributed to the video coding. This research was supported by MEXT KAKENHI (\#15300088, \#22243043, \#23120003) for Imai, Okada, and (\#23300099) for Hidaka, MEXT GCOE program to Tamagawa University, and Biotechnology and Biological Sciences Research Council's Research Development Fellowship [BB/G023069/1] to Kita.

\section*{References}

Davis, R. (1961). The fitness of names to drawings. A crosscultural study in Tanganyika. British Journal of Developmental Psychology, 52, 259-268.
de Saussure, F. (1916/1983). Course in general linguistics
(R. Harris, Trans.). La Salle, IL: Open Court.

Fennell, C. T., \& Waxman, S. R. (2010). What paradox? Referential cues allow for infant use of phonetic detail in word learning. Child development, 81(5), 1376-1383.
Firth, J. R. (1935). The technique of semantics. Transactions of the Philological Society, 36-72.
Gogate, L. J., \& Hollich, G. (2010). Invariance detection within an interactive system: a perceptual gateway to language development. Psychological review, 117(2), 496516.

Holland, M. K., \& Wertheimer, M. (1964). Some physiognomic aspects of naming, or maluma and takete revisited. Perceptual and Motor Skills, 19(1), 111-117.
Hollich, G., Golinkoff, R. M., \& Hirsh-Pasek, K. (2007). Young children associate novel words with complex objects rather than salient parts. Dev Psychol,43(5), 10511061.

Imai, M., Kita, S., Nagumo, M., \& Okada, H. (2008). Sound symbolism facilitates early verb learning. Cognition, 109(1), 54-65.
Jeffreys, H. (1961). The Theory of Probability (3rd ed.). Oxford.
Köhler, W. (1929/1947). Gestalt psychology: An introduction to new concepts in modern psychology. New York: Liveright Publishing Corporation.
Kantartzis, K., Imai, M., \& Kita, S. (2011). Japanese SoundSymbolism Facilitates Word Learning in EnglishSpeaking Children. Cognitive Science, 35, 575-586.

Kita, S. (1997). Two-dimensional semantic analysis of Japanese mimetics. Linguistics, 35, 379-415.
Kita, S., Kantartzis, K., \& Imai, M. (2010). Children learn sound symbolic words better: Evolutionary vestige of sound symbolic protolanguage. Paper presented at the The evolution of language: Proceedings of the 8th International Conference (EVOLANG8), 206-213. Singapore: World Scientific.
Markman, E. M. (1989). Categorization and naming in children: Problems of induction. Cambridge, Mass.: MIT Press.
Maurer, D., Pathman, T., \& Mondloch, C. J. (2006). The shape of boubas: sound-shape correspondences in toddlers and adults. Developmental science, 9(3), 316-322.
Mondloch, C. J., \& Maurer, D. (2004). Do small white balls squeak? Pitch-object correspondences in young children. Cognitive, affective \& behavioral neuroscience, 4(2), 133-136.
Ogura, T., \& Watamaki, T. (2004). MacArthur Communicative Development Inventories: User's Guide and Technical Manual. Kyoto: Kyoto International Social Welfare Exchange Centre.
Ozturk, O., Krehm, M. \& Vouloumanos, A. (2013). Sound symbolism in infancy: evidence for sound-shape crossmodal correspondences in 4-month-olds. Journal of Experimental Child Psychology, 114, 173-186.
Pena, M., Mehler, J., \& Nespor, M. (2011). The role of audiovisual processing in early conceptual development. Psychological science, 22(11), 1419-1421.
Quine, W. V. (1960). Word and Object: The MIT Press.
Ramachandran, V. S., \& Hubbard, E. M. (2001). Psychophysical investigations into the neural basis of synaesthesia. Proc Biol Sci, 268(1470), 979-983.
Spiegel, C., \& Halberda, J. (2011). Rapid fast-mapping abilities in 2-year-olds. Journal of experimental child psychology, 109(1), 132-140.
Stager, C. L., \& Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. Nature, 388(6640), 381-382.
Swingley, D., \& Aslin, R. N. (2002). Lexical neighborhoods and the word-form representations of 14-month-olds. Psychological science, 13(5), 480-484.
Voeltz, F. K. E., \& Kilian-Hatz, C. (Eds.). (2001). Ideophones. Amsterdam/Philadelphia: John Benjamins.
Wagenmakers, E-J., Lodewyckx, T., Kuriyal, H., \& Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method., Cognitive Psychology, 60, 158-189.
Werker, J. F., Cohen, L. B., Lloyd, V. L., Casasola, M., \& Stager, C. L. (1998). Acquisition of word-object associations by 14-month-old infants. Dev Psychol, 34(6), 1289-1309.
Westbury, C. (2005). Implicit sound symbolism in lexical access: evidence from an interference task. Brain and language, 93(1), 10-19.

Woodward, A. L., \& Hoyne, K. L. (1999). Infants' learning about words and sounds in relation to objects. Child development, 70(1), 65-77.
Yoshida, K. A., Fennell, C. T., Swingley, D., \& Werker, J. F. (2009). Fourteen-month-old infants learn similarsounding words. Developmental science, 12(3), 412-418.
Yurovsky D, Hidaka S, \& Wu R (2012) Quantitative Linking Hypotheses for Infant Eye Movements. PLoS ONE 7(10).

\title{
Beliefs About a Speaker Affect Feeling of Another's Knowing
}

\author{
Lisette Mol (l.mol@tilburguniversity.edu) \({ }^{1}\) \\ \({ }^{1}\) Tilburg center for Cognition and Communication (TiCC), School of Humanities, Tilburg University \\ P.O. Box 90135, NL-5000 LE Tilburg, The Netherlands
}

\author{
Anna K. Kuhlen (anna.kuhlen@bcen-berlin.de) 2, 3,4 \\ \({ }^{2}\) Bernstein Center for Computational Neuroscience, Charité - Universitätsmedizin Berlin, Berlin, Germany \\ \({ }^{3}\) Berlin School of Mind and Brain, Humboldt-Universität zu Berlin, Berlin, Germany \\ \({ }^{4}\) Berlin Center of Advanced Neuroimaging, Charité - Universitätsmedizin Berlin, Berlin, Germany
}

\title{
Ruben van der Steen (r.p.vdrsteen@tilburguniversity.edu) \({ }^{1}\)
}

Mike Obbens (m.p.obbens@tilburguniversity.edu) \({ }^{1}\)

\begin{abstract}
People's estimations of how certain speakers are of their knowledge (FOAK) match speakers' own estimation (FOK) of how certain they are (Brennan \& Williams, 1995). This is because others can interpret the verbal and nonverbal cues of (un)certainty that a speaker displays (Brennan \& Williams, 1995; Swerts \& Krahmer, 2005). Estimating another's certainty thus seems to be driven by the bottom-up processing of speaker-displayed cues. In this paper, we explore the topdown influence of beliefs about a speaker on judgments of a speaker's certainty. In a perception study, we varied whether a speaker's proclaimed profession would make him an expert or a novice on the topic he was questioned on. Such beliefs were shown to influence participants' ratings of the speaker's certainty, in addition to speaker-displayed cues. Thus, next to the bottom-up processing of speaker-displayed cues, the topdown processing of beliefs about a speaker influences judgments of others' certainty.
\end{abstract}

Keywords: FOK, FOAK, top-down processing, bottom-up processing, speaker-displayed cues, person perception.

\section*{Introduction}

When watching the news on television or online, we often are informed by so-called 'experts' on the current topic, for example, an economist may comment on the Euro crisis and an architect may be interviewed on the progress of a construction site. Often, the person's expertise is displayed in a header once their contribution starts, or announced upfront by the newsreader conducting the interview. Does such knowledge of people's expertise affect our judgment of their certainty? And if so, does this effect depend on whether the expert displays certain or uncertain behavior?
This study assesses the influence of knowing another person's expertise, on the judgment of their certainty when answering questions. We thereby test the influence of this factor relative to the verbal and nonverbal cues of uncertainty displayed by the person answering the questions (henceforth referred to as the 'respondent'). In the following, we first provide an introduction on the production and perception of cues of (un)certainty. Then we discuss the
different types of processing involved in utilizing speakerdisplayed cues, and in making use of beliefs about the speaker. This leads to our research question and hypotheses.

\section*{Displaying (Un)certainty}

Sometimes, when unable to remember the answer to a particular question, we have a strong intuition that we do know the answer, despite our momentary inability to retrieve it from memory. This meta-cognitive phenomenon is known as feeling-of-knowing (FOK), (Hart, 1965). Participants' FOK has been shown to be a reliable predictor of whether they can later recognize the sought-after answer in a multiple-choice test (Blake, 1973; Hart, 1965). This shows that people's intuition on whether particular knowledge is stored in their memory or is absent from it, tends to be correct.
When sharing our knowledge with others, we tend to share our intuition on the certainty of this knowledge as well, by displaying auditory and visual cues of (un)certainty (Brennan \& Williams, 1995; Goffman, 1967, 1971, 1978; Smith \& Clark, 1993; Swerts \& Krahmer, 2005). This may be done to save face in case of being incorrect, or to be as informative as needed, in accordance with Grice's maxim of quantity (Grice, 1989). FOK-ratings can be obtained by asking respondents how certain they are of their answer to particular knowledge questions (Hart, 1965). By matching such ratings to the auditory and visual behavior respondents exhibit while answering, characteristic cues of displays of (un)certainty have been identified. Auditory cues of uncertainty were shown to include: rising intonation, an initial pause, the use of fillers ("um", "uh"), hedging ("I think", "Most likely"), and self-talk ("Let's see, what was that again..."), (Goffman, 1978; Smith \& Clark, 1993). Certainty, on the other hand, is displayed auditorily by the absence of such cues, and a falling intonation.
Visually, uncertainty can be displayed by rising the eyebrows, smiling (when recognizing the answer should be known), producing a marked facial expression (a 'funnyface'), and diverted gaze (Swerts \& Krahmer, 2005).

Certainty is displayed visually by the absence of such cues (e.g., not diverting gaze), although particularly easy questions can also elicit smiles, which then signal certainty.
If these auditory and visual cues serve to convey a level of certainty to an interaction partner, it is expected that people can correctly interpret them.

\section*{Perceiving (Un)certainty}

People are indeed sensitive to the cues of (un)certainty others display. The intuition we have of whether another person is likely to know the correct answer to a question, is known as feeling-of-another's-knowing (FOAK), (Brennan \& Williams, 1995). FOAK-ratings can be elicited by presenting participants with other's answers and asking them how certain they are that the respondent gave the correct answer. This way, answers with rising intonation and longer response latencies were found to elicit lower FOAK-ratings than answers with falling intonation and shorter latencies (Brennan \& Williams, 1995). Also, adding filled pauses to answers led to lower FOAK-ratings than adding unfilled pauses. Participants' FOAK-ratings were found to match respondents' own FOK-ratings (Brennan \& Williams, 1995). Therefore, it seems that people can correctly interpret the auditory cues of (un)certainty others display.

When participants had access to both visual and auditory cues displayed by respondents, the accuracy of their FOAKratings increased as compared to when they had access to either auditory or visual information (Swerts \& Krahmer, 2005). Thus, people can reliably estimate how certain others are of their knowledge, by interpreting their auditory and visual displays of (un)certainty.

\section*{Top-down vs. Bottom-up Processing of Cues}

Next to speaker-displayed cues, more global information about a person's expertise can also inform inferences about this person's knowledge and credibility (for an overview, see Pornpitakpan, 2004). Along these lines, expectations concerning another person's knowledge can be guided by that person's presumed gender (Fussell \& Krauss, 1992), age (Newman-Norlund, et al., 2009), or geographic origin (Isaacs \& Clark, 1987). Also, previous experiences with the person shape expectations about what they are likely to know (Galati \& Brennan, 2010; Metzing \& Brennan, 2003). Beliefs about a person's expertise may therefore influence metacognitive assessments of that person's knowing in a top-down fashion.
In fact, global information may influence the interpretation of locally available verbal and nonverbal displays of (un)certainty. Along these lines, people have been shown to interpret a person's speech disfluencies differently if they can attribute them to a cognitive impairment (Arnold, Kam, \& Tanenhaus, 2007). And speakers interpret their addressees' verbal and nonverbal feedback behavior based on the expected involvement of the addressees in the interaction (Kuhlen \& Brennan, 2010; Kuhlen, Galati, \&

Brennan, 2012). Bottom-up processes informed by locally available verbal and nonverbal displays of knowing may therefore be shaped by top-down processes informed by global information about the respondents' expertise. Investigating how these two processes inform complex social judgments, such as assessing another person's knowledge, will contribute to our understanding of human social cognition.

\section*{Present Study}

The present study assesses whether the top-down processing of global information affects judgments of others' certainty, in addition to the bottom-up processing of locally available cues. To this aim, we manipulated participants' belief about a respondent's expertise, as well as the locally available verbal and nonverbal cues, displayed by the respondent. Based on previous work (Brennan \& Williams, 1995; Swerts \& Krahmer, 2005), we expect the respondent's verbal and nonverbal displays of certainty to influence participants' FOAK-ratings in a bottom-up fashion. In addition, we expect that the interpretation of these displays is influenced top-down by participants' beliefs about the respondent's expertise. Lastly, since displays of certainty have been primarily described by the absence of cues of uncertainty, bottom-up processes may be less important when judging high-FOK as compared to low-FOK displays. Therefore, the top-down processing of global cues may affect FOAK-ratings differently for each type of display.

Below, we first describe how we created stimuli in which a respondent clearly displays verbal and nonverbal cues of high and low FOK. Then follows a description of the main experiment, in which we manipulated participants' beliefs about the respondent's profession, and thereby his expertise.

\section*{Method}

\section*{Material}

Selecting Knowledge Domains To elicit high- and lowFOK answers, a 30 -year-old male tax advisor was interviewed on two domains relating to his interests: gardening and Dutch literature. In two separate pretests, participants were presented with a picture of the respondent and asked how likely it was that he was of certain professions. Ten professions were tested on either pretest, including gardener and Dutch teacher. Each pretest included 16 participants. On a six-point scale, participants rated the possibility that the respondent was a gardener \((M=3.50, S D\) \(=1.14\) ) equally likely to the possibility that he was a Dutch teacher \((M=3.50, S D=1.41), t(31)=.00, p=1.00\).
Participants also rated the professions (ten per test) for how knowledgeable someone of this profession would be in gardening and Dutch literature. A paired samples t-test revealed that on a six-point scale, a gardener was indicated to be more knowledgeable in gardening \((M=5.75, S D=\) .68) than a Dutch teacher \((M=2.69, S D=1.35), t(15)=\)
7.42, \(p<.001\). Vice versa, a Dutch teacher was rated more knowledgeable in Dutch literature ( \(M=5.89, S D=.34\) ) than a gardener \((M=2.00, S D=.63), t(15)=31.00, p<.001\).
Eliciting Audiovisual Displays of High and Low FOK The respondent was asked 40 multiple-choice questions on gardening, followed by 40 multiple-choice questions on Dutch literature. Multiple-choice questions were used to avoid non-answers (e.g., "I don't know") and to manipulate the difficulty of the questions. Each question had four alternatives, see examples (1) and (2). The respondent was instructed to say the answer out loud, e.g., "Blauwe regen". The experimenter asking the questions was located behind the respondent, such as not to give the respondent any cues of the answer being correct or incorrect. Answers were captured with a video camera situated in front of the respondent. After answering, the respondent indicated on a six-point scale how certain he was of his answer being correct, '6' indicating 'definitely correct' and ' 1 ' indicating 'definitely incorrect'. Following Hart (1965), this was taken as a measure of the respondent's feeling-of-knowing (FOK).
(1) Welke plant is giftig? (Which plant is toxic?) A: Blauwe regen, B : Geranium, C: Orchidee, D: Waterlelie
(2) Wie schreef in 1947 de roman 'De avonden'? (Who wrote the novel 'De avonden' in 1947?) A: Jan Cremer, B: Herman Bursselmans, C: Harry Mulisch, D: Gerard Reve.

This way, 40 answers were collected in each domain. Since the respondent never indicated a FOK-score of 1 , answers with a FOK-score of 2 or 3 were regarded lowFOK and those with a FOK-score of 5 or 6 were regarded as high-FOK. Answers with a score of 4 were few and were disregarded. Sometimes, the respondent's answer contained information about the question being a multiple-choice question, for example "the first one". These responses were disregarded as well. For each domain, ten high- and ten lowFOK answers were then selected, based on their intelligibility and on how clear the displayed cues seemed to be. Whether these clips indeed included clear displays of high- and low-FOK was assessed in a third pretest.
Selecting FOK Displays In the third pretest, 20 native Dutch participants (ten female) watched clips of the 40 selected answers and indicated on a six-point scale how certain they were of the respondent's answer being correct. Following Brennan and Williams (1995), this was taken as a measure of participants' feeling-of-another's-knowing (FOAK). Since participants were only presented with the respondent's answers and not the questions asked, they had to rely on their estimation of how certain the speaker was (FOAK), to tell whether the answer was correct or not.
A \(2 \times 2\) ANOVA with within-factors FOK (levels: high, low) and Domain (levels: gardening, Dutch literature) revealed a main effect of FOK on participants' FOAKratings, \(F(1,18)=157.85, p<.001, \eta_{p}{ }^{2}=.90\), see Table 1 .

Table 1: Mean (SD) FOAK-ratings in the pretest.
\begin{tabular}{llc}
\hline FOK: & Domain: & FOAK \((S D)\) : \\
\hline \multirow{2}{*}{ High } & Gardening \((N=10)\) & \(4.32(.46)\) \\
\((N=20)\) & Dutch Literature \((N=10)\) & \(4.64(.60)\) \\
& Total \((N=20)\) & \(4.48(.11)\) \\
\hline \multirow{2}{*}{ Low } & Gardening \((N=10)\) & \(3.09(.66)\) \\
\((N=20)\) & Dutch Literature \((N=10)\) & \(2.78(.49)\) \\
& Total \((N=20)\) & \(2.93(.10)\) \\
\hline
\end{tabular}

There was no main effect of Domain, \(F(1,18)=.001, p=\) .91. Domain and FOK did interact, \(F(1,18)=17.41, p<\) \(.001, \eta_{p}{ }^{2}=.49\). The difference in FOAK-ratings on highFOK and low-FOK clips was larger for the Dutch literature than for the gardening domain, see Table 1. These results evidence that the clips contained speaker-displayed cues.
For each domain, those sets of clips were selected that participants rated most consistently as portraying either high- or low-FOK answers (assessed by Cronbach's alpha). This way, we could be most certain that our selected FOK displays contained informative cues about the respondent's feeling of knowing. Our final set of stimuli contained seven high-FOK clips for both domains, seven low-FOK clips for the literature domain and five low-FOK clips for the gardening domain. Unfortunately, we did not obtain more suitable low-FOK clips from the gardening domain.

\section*{Task}

Participants' task in the main experiment was to judge the respondent's answers in the selected clips, indicating on a six-point scale how certainly the respondent's answer was correct: ' 1 ' indicating 'certainly incorrect' and ' 6 ' indicating 'certainly correct'. This way, we elicited participants' FOAK-judgments of the respondent's answers.

\section*{Design}

The factors FOK (levels: low, high) and Domain (levels: gardening, literature) were manipulated within participant. The factor Profession (levels: gardener, Dutch teacher, profession not mentioned) was manipulated betweenparticipants. An equal number of men and women participated in each condition. The order in which the two domains were presented was counterbalanced across each condition and across sex. Each participant saw the clips within a domain in a different, randomly generated order.

\section*{Procedure}

The main experiment was conducted as an online survey. Participants received a link through email, which led them to the website of the experiment. Clips were grouped by domain. A short instruction, which announced the domain that the questions were in, preceded the clips in either domain. This instruction also included a description of the respondent, mentioning his age (30) and city of residence (Spijkenisse), along with, depending on the experimental

Table 2: Mean (SD) FOAK-ratings for answers in the Gardening domain.
\begin{tabular}{llc}
\hline Clips: & Profession: & FOAK \\
\hline & Gardener \((N=24)\) & \(3.41(.68)\) \\
Gardening, & Not mentioned \((N=22)\) & \(3.16(.93)\) \\
Low FOK & Dutch Teacher \((N=22)\) & \(2.88(.53)\) \\
\((N=5)\) & Total \((N=68)\) & \(3.15(.75)\) \\
\hline & Gardener \((N=24)\) & \(5.37(.58)\) \\
Gardening, & Not mentioned \((N=22)\) & \(4.99(.64)\) \\
High FOK & Dutch Teacher \((N=22)\) & \(5.08(.57)\) \\
\((N=7)\) & Total \((N=68)\) & \(5.15(.61)\) \\
\hline
\end{tabular}
condition, his profession. The experiment was self-paced and participants could view each clip as often as they wished. They indicated their answer by clicking a radiobutton on a horizontally laid-out six-point scale, before proceeding to the next clip. After all clips had been rated, participants were asked for the respondent's profession (as a manipulation check) and for their own knowledge of gardening and Dutch literature.

\section*{Analyses}

Data of participants who did not correctly remember the respondent's profession (five cases), or who mentioned a profession in the condition in which no profession was mentioned (four cases) were excluded from our analyses. Data from any non-native speakers of Dutch were excluded as well (four cases). Subsequently, data from a minimal number of participants were removed from the sample to ensure counterbalancing of sex, and order of presentation of the domains (seven cases). For this purpose, data from participants who participated last were eliminated first.

\section*{Participants}

Our final sample contained data of 68 native Dutch participants ( 34 female). They were aged between 17 and 37 years old \((M=22.85, S D=3.33)\) and did not take part in any of our pretests.

\section*{Results}

Initially, we conducted a \(3 \times 2 \times 2\) ANOVA, with betweenfactor Profession (levels: not mentioned, gardener, Dutch teacher), and within-factors: Domain (levels: gardening, literature) and FOK (levels: high, low). This revealed a main effect of FOK, such that participants' FOAK-ratings were higher for high-FOK clips ( \(M=5.10, S D=.53\) ) than for low-FOK clips \((M=2.71, S D=.44), F(1,65)=726.14, p<\) \(.001, \eta_{p}{ }^{2}=.92\). We also found a main effect of Domain, such that FOAK-ratings were higher for the gardening domain \((M=4.32, S D=.50)\) than for the Dutch literature domain \((M=3.71, S D=.05), F(1,65)=42.47, p=.001, \eta_{p}{ }^{2}\) \(=.40\). Domain and FOK interacted, \(F(1,65)=34.46, p<\) \(.001, \eta_{p}{ }^{2}=.35\). The difference in rating between high- and

Table 3: Mean (SD) FOAK-ratings for answers in the Literature domain.
\begin{tabular}{llc}
\hline Clips: & Profession: & FOAK \\
\hline Literature, & Dutch Teacher \((N=22)\) & \(2.41(.58)\) \\
Low FOK & Not mentioned \((N=22)\) & \(2.40(.49)\) \\
\((N=7)\) & Tardener \((N=24)\) & \(2.35(.47)\) \\
& Total \((N=68)\) & \(2.38(.51)\) \\
\hline \multirow{2}{*}{ Literature, } & Dutch Teacher \((N=22)\) & \(5.11(.50)\) \\
High FOK & Not mentioned \((N=22)\) & \(5.09(.62)\) \\
\((N=7)\) & Gardener \((N=24)\) & \(4.93(.67)\) \\
& Total \((N=68)\) & \(5.04(.60)\) \\
\hline
\end{tabular}
low-FOK clips was larger for the Dutch literature domain than for the gardening domain, see Tables 2 and 3.
Profession did not exert a main effect on participants' FOAK-ratings, \(F(2,65)=1.01, p=.37, \eta_{p}{ }^{2}=.03\), revealing no overall differences in FOAK-ratings between the three conditions. As expected, Profession and Domain interacted, \(F(2,65)=6.05, p=.004, \eta_{p}{ }^{2}=.16\). Because of the differential influence of domain on our main variables of interest (Profession and FOK), we analyzed each domain separately, by means of a \(2 \times 3\) ANOVA with FOK as within-factor and Profession as a between factor. Pairwise comparisons were performed using least square differences.

\section*{Results for the Knowledge Domain Gardening}

Table 2 provides an overview of the mean FOAK-ratings in the gardening domain. There was a main effect of FOK on FOAK, such that high-FOK clips \((M=5.15, S D=.61)\) were rated as more certainly correct than low-FOK clips ( \(M=\) 3.16, \(S D=.75), F(1,65)=316.78, p<.001, \eta_{p}{ }^{2}=.83\). Profession showed a main effect on \(\operatorname{FOAK}, F(2,65)=4.56\), \(p=.014, \eta_{p}{ }^{2}=.12\). Pairwise comparisons revealed that the respondent was rated as more certainly correct when he was labeled a gardener, than when he was labeled a Dutch teacher ( \(p=.005\) ), or when no profession was mentioned ( \(p\) \(=.033\) ). Ratings between the latter two did not differ significantly ( \(p=.499\) ). The factors FOK and Profession were not found to interact, \(F(2,65)=.89, p=.416\).

\section*{Results for the Knowledge Domain Literature}

Table 3 provides an overview of the mean FOAK-ratings in the Dutch literature domain. There was a main effect of FOK on FOAK, such that high-FOK clips \((M=5.04, S D=\) .60) were rated more certainly correct than low-FOK clips \((M=2.38, S D=.51), F(1,65)=811.92, p<.001, \eta_{p}{ }^{2}=.93\). We did not find a main effect of Profession ( \(F<1\), n.s.), nor an interaction between Profession and FOK ( \(F<1\), n.s.).
To see if the null-result for Profession should be interpreted as evidence against our hypothesis, we conducted a Bayesian analyses on the difference between the Dutch teacher and gardener condition. In the gardening domain, an independent samples \(t\)-test showed higher FOAK-ratings for the gardener \((M=4.55, S D=.44)\) than
for the Dutch teacher condition ( \(M=4.16, S D=.37\) ), \(t(44)\) \(=3.26, p=.002\). In the literature domain, no difference was found between the gardener \((M=3.64, S D=.44)\) and Dutch teacher condition \((M=3.76, S D=.42), t(44)=.961, p=\) .342. Modeling the predicted effect in the literature domain as a normal distribution, with its mean equal to the effect in the gardening domain (.39), and a standard deviation of half this effect (also see Dienes, 2011), rendered \(\mathrm{Bf}=.45\). This indicates that the results from the literature domain do not discriminate between the null-hypothesis and the hypothesis of an effect of Profession on FOAK.

\section*{Results for Participants' Expertise}

A paired samples t-test showed that on a 7-point scale, participants reported to be more knowledgeable in Dutch literature ( \(M=3.90, S D=1.56\) ) than in gardening ( \(M=\) \(2.69, S D=1.25), t(67)=5.35, p<.001,95 \% \mathrm{CI}=(.76\), 1.66). Adding self-reported expertise as a covariate did not reveal an effect of this factor on participants' FOAK-ratings.

\section*{Discussion}

Our results showed strong effects of the respondent's feeling of knowing (FOK) on participants' feeling of another's knowing (FOAK). Following the hypothesis that people make use of verbal (Brennan \& Williams, 1995) and nonverbal (Swerts \& Krahmer, 2005) cues when judging someone's certainty, this indicates that our clips contained clear speaker-displayed cues, which participants used to judge the respondent's certainty.

Going beyond the results of earlier studies, we found that beliefs about the respondent, specifically about the respondent's expertise, influenced participants' judgment of the respondent's certainty as well. When asked questions about gardening, the same respondent was rated as more certainly correct when participants were told he was a gardener, compared to when they were told he was a Dutch teacher, or when no information on the respondent's profession was provided. This shows that, in addition to the information that could be obtained from cues displayed by the respondent, participants' beliefs about the respondent's expertise influenced their judgment of how certain the respondent was of his answers. Therefore, top-down processes informed by global information about a speaker can influence assessments of another person's feeling of knowing too. This top-down effect held both for clips in which the respondent was uncertain of his answer (low FOK) and for clips in which he was certain (high FOK).
Our between-subjects manipulation of expertise allowed us to use the same clips in each condition, ensuring identical speaker-displayed cues and speaker attributes. Participants were randomly assigned to conditions, and our analyses did not show evidence for an overall difference in FOAKratings between the conditions. Hence, we are confident that our results cannot be ascribed to a priori differences between the three groups of participants.

People sometimes use their own knowledge to estimate others' knowledge (Fussell \& Krauss, 1991; Jameson, Nelson, Leonesio, \& Narens, 1993; Nickerson, Baddeley, \& Freedman, 1987). In our study, participants reported having more knowledge on Dutch literature than on gardening. Nevertheless, they rated the speaker to be more certain in the gardening domain than in the Dutch literature domain. Entering participants' self-reported knowledge as a covariate did not render any significant results. Hence, reported effects seem unaffected by participants' own knowledge.
Follow-up studies need to assess if our results generalize to different respondents, domains, and beliefs. In this study, we only found evidence for an additional effect of beliefs about the respondent's expertise in one domain: gardening. We did not find this effect for the literature domain. However, a Bayesian analyses indicated that the results from the literature domain should not be interpreted as evidence against, nor in favor of our hypothesis. It seems that more factors are at play still, which attenuated the effect of beliefs about a speaker in this domain. One difference between the two domains was that the effect of speakerdisplayed cues was even stronger in the literature than in the gardening domain. It may be the case that the role of beliefs diminishes when speaker-displayed cues are very clear. Future studies are needed to uncover what factors moderate the effect of beliefs about a speaker.
Our findings have important implications for our understanding of the social and cognitive processes involved in person perception. From a social perspective, it is striking that simple information, such as labeling someone as being an expert by assigning them a certain profession, can sway perceivers towards judging them to be more knowledgeable in their domain of expertise. This is in line with social psychological literature on persuasion, which shows that perceived experts are expected to provide information that is valid (e.g., Clark, Wegener, Habashi, \& Evans, 2012; Hovland \& Weiss, 1951). Our study thereby contributes to a better understanding of the mechanisms behind perceiving expertise and taking advice from experts (see e.g., Jungermann \& Fischer, 2005). From a cognitive perspective, our study contributes to a growing literature on social cognition showing that the interpretation of social cues cannot be separated from global attributions about the person displaying these cues (Teufel, Fletcher, \& Davis, 2010).

Previous work has suggested that the processing of nonverbal cues is shaped by top-down expectations about the person (Kuhlen \& Brennan, 2010; Kuhlen et al., 2012). In these studies participants responded differently to similar nonverbal behavior of their conversational partners depending on how they had expected their partners to behave. In the present study, it is difficult to disentangle how exactly nonverbal cues are integrated with global beliefs about the respondent. Possibly, evidence from both top-down and bottom-up cues accumulates additively,
swaying the perceiver's judgments in one or the other direction. Future work will investigate further how these two sources of information interact.

\section*{Conclusion}

Our study showed that next to speaker-displayed cues of (un)certainty, beliefs about a speaker can also affect FOAKratings. This shows that people's feeling of another's knowing is affected both by the bottom-up processing of local cues displayed by the speaker and the top-down processing of global beliefs they have about this speaker.

\section*{Acknowledgements}

We thank the reviewers for their insightful comments, Zoltan Dienes for help with the Bayesian analysis, Joe Walther for his advice on the Social Psychology literature, Jacqueline Dake for her assistance with the online surveys, and Paul Eijkenaar for volunteering as the respondent.

\section*{References}

Arnold, J. E., Kam, C. L., \& Tanenhaus, M. K. (2007). If you say thee uh you are describing something hard: the on-line attribution of disfluency during reference comprehension. Journal of Experimental Psychology: Learning, Memory, and Cognition, 33(5), 914-930.
Blake, M. (1973). Prediction of recognition when recall fails: Exploring the feeling-of-knowing phenomenon. Journal of Verbal Learning and Verbal Behavior, 12, 311-319.
Brennan, S. E., \& Williams, M. (1995). The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. Jounral of Memory and Language, 34, 383-398.
Clark, J. K., Wegener, D. T., Habashi, M. M., \& Evans, A. T. (2012). Sourse expertise and persuasion: the effects of perceived opposition or support on message scrunity. Personality and Social Psychology Bulletin, 38(1), 90100.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? Perspectives on Psychological Science, 6(3), 274-290.
Fussell, S. R., \& Krauss, R. M. (1991). Accuracy and bias in estimates of others' knowledge. European Journal of Social Psychology, 21, 445-454.
Fussell, S. R., \& Krauss, R. M. (1992). Coordination of knowledge in communication: Effects of speakers' assumptions about ohters' knowledge. Journal of Personality and Social Psychology, 62, 378-391.
Galati, A., \& Brennan, S. E. (2010). Attenuating information in spoken communication: For the speaker, or for the addressee? Journal of Memory and Language, 62, 35-51.
Goffman, E. (1967). Interaction ritual: Essays on face to face behavior. Garden City, NY: Doubleday.

Goffman, E. (1971). Relations in public. New York: Harper and Row.
Goffman, E. (1978). Response cries. Language, 54, 787815.

Grice, P. (1989). Studies in the Way of Words. Cambridge MA: Harvard University Press.
Hart, J. T. (1965). Memory and the feeling-of-knowing experience. Journal of Educational Psychology, 56, 208216.

Hovland, C. I., \& Weiss, W. (1951). The influence of source credibility on communication effectiveness. Public Opinion Quarterly, 15, 635-650.
Isaacs, E. A., \& Clark, H. H. (1987). References in conversation between experts and novices. Journal of Experimental Psychology: General, 116(1), 26-37.
Jameson, A., Nelson, T. O., Leonesio, R. J., \& Narens, L. (1993). The feeling of another person's knowing. Journal of Memory and Language, 32, 320-335.
Jungermann, H., \& Fischer, K. (2005). Using expertise and experience for giving and taking advice. In T. Betsch \& S. Haberstroh (Eds.), The routines of decision making. Mahwah, NJ: Lawrence Erlbaum.
Kuhlen, A. K., \& Brennan, S. E. (2010). Anticipating distracted addressees: How speakers' expectations and addressees' feedback influence storytelling. Discourse Processes, 47, 567-587.
Kuhlen, A. K., Galati, A., \& Brennan, S. E. (2012). Gesturing integrates top-down and bottom-up information: Effects of speakers' expectations and addressees' feedback. Language and Cognition, 4, 17-41.
Metzing, C., \& Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. Journal of Memory and Language, 49, 201-213.
Newman-Norlund, S. E., Noordzij, M. L., NewmanNorlund, R. D., Volman, I. A. C., De Ruiter, J. P., Hagoort, P., \& Toni, I. (2009). Recipient design in tacit communication. Cognition, 111, 46-54.
Nickerson, R. S., Baddeley, A., \& Freedman, B. (1987). Are people's estimates of what other people know influenced by what they themselves know? Acta Psychologica, 64, 254-259.
Pornpitakpan, C. (2004). The persuasiveness of source credibility: A critical review of five decades' evidence. Journal of Applied Social Psychology, 34(2), 243-281.
Smith, V. L., \& Clark, H. H. (1993). On the course of answering questions. Journal of Memory and Language, 32, 25-38.
Swerts, M., \& Krahmer, E. (2005). Audiovisual prosody and feeling of knowing. Journal of Memory and Language, 53, 81-94.
Teufel, C., Fletcher, P. C., \& Davis, G. (2010). Seeing other minds: attributed mental states influence perception. Trends in Cognitive Sciences, 14, 376-382.

\title{
Sharing in the Dark? Target Memory and Risk Awareness in Online Communication
}

\author{
Ricarda Moll (r.moll@uni-muenster.de) \\ Stephanie Pieschl (pieschl@uni-muenster.de) \\ Rainer Bromme (bromme@uni-muenster.de) \\ Department of Psychology, Westfälische Wilhelms-Universität Münster Fliednerstr. 21, 48149 Muenster, Germany
}

\begin{abstract}
A high degree of self-disclosure in Online Social Networks (OSNs) is associated with several risks. This raises an important question: Why don't many users protect their personal data more eagerly? We propose that a lack of memory for what information has been disclosed to which audience contributes to this privacyneglecting behavior in OSNs. We transferred the paradigm of target monitoring to a fictitious OSN and varied the degree of risk associated with selfdisclosure. In a \(2 \times 2\) experiment we varied both audience size (large vs. small) and information intimacy (personal vs. non-personal). We used recognition tests for the association of audience and disclosed information to assess memory performance. Results show that item memory (the memory for what information has been disclosed) exceeded target memory and that target memory improved in vulnerable situations (for large audiences and personal information). Our findings widen the realm of offline memory research and expand our knowledge about which cognitive factors impact privacy-related behavior in online environments.
\end{abstract}

Keywords: Target Memory, Online Self-Disclosure, Risk Awareness

\section*{Introduction}

\section*{Self-Disclosure in OSNs}

Self-disclosure is an important mechanism in relationship formation and trust development (Jourard \& Lasakow, 1958). Lately much research has investigated the benefits of sharing personal information on OSNs-platforms i.e. in terms of self-esteem and identity formation (Valkenburg \& Peter, 2011). At the same time this behavior is associated with several risks. For example, it is not uncommon that employers retrieve information of their job applicants through an online search that includes profile information on OSNs. This information often decides to whom the announced job position is offered (Zeidner, 2007). Interestingly, many users actually are concerned about potential data misuse but nevertheless choose to reveal personal information in OSNs, a pattern that is known as
privacy paradox (Norberg, Horne, \& Horne, 2007). Some scholars argue that privacy related decisions result from a logical calculation in which risks and benefits of selfdisclosure are rationally weighed against each other ( Xu , Teo, Tan, \& Agarwal, 2010). If users self-disclose despite their concerns, the benefits must be larger and/or more probable than the associated risks. However, we argue that privacy-related calculations might be biased, because users could lack important information to assess the actual amount of vulnerability: Online revelation of personal information is usually done over and over again and while the single event is indeed not that risky, its repetition produces a cumulative amount of online information about the user (indicating a corresponding amount of cumulative risk). One important detail that is crucial for the assessment of this cumulative risk is the memory for which information has been disclosed, and to whom it is available.

\section*{Target Memory in Offline Interactions}

Recently it has been shown that people struggle to remember the targets of their messages in offline contexts. For example, in one of their experiments Gopie and MacLeod (2009) investigated how well people remember having disclosed fifty personal facts to pictures of famous people in comparison to impersonal facts. Results show that people successfully identified the facts they had disclosed (item memory) but had problems associating facts and targets (target memory); they did not remember to whom they disclosed what. In another study Marsh and Hicks (2002) let participants repeatedly choose to whom they wanted to give different kinds of objects. The authors conclude that this decisional aspect leads to a deeper elaboration of the situational context which then facilitates subsequent retrieval. Finally, Brown, Hornstein, and Memon (2006) let participants tell various pieces of information to five different celebrity pictures on five subsequent days. On-line and retrospective target memory declined with the number of previous "interactions", indicating a confusion of which information was given to whom.

In the light of these findings it seems plausible to assume that target memory problems also exist online. Therefore, we transferred the paradigm of target monitoring to the environment of OSNs. As a prerequisite for further analysis we assumed that participants remember what information they disclosed but struggle remembering to whom. We thus
hypothesized that item memory exceeds target memory (hypothesis 1).

\section*{Risk Cues in Online Communication}

There has been much discussion on how online environments change the nature of communication and information processing (Kiesler, Siegel, \& McGuire, 1984). However, while OSNs environments are somewhat deprived of conventional context cues that could support the encoding and decoding of information, other situational factors could be more relevant for target memory in OSNs. For example, the elaboration likelihood model (Petty, Cacioppo, \& Kasmer, 1988) argues that when people are motivated they process information in a more elaborate way. It seems plausible that people are more motivated if they feel at risk. People could therefore process risky situations more thoroughly than neutral ones. This could be especially true for online interactions since the accessibility and distribution of information is inherently more difficult to control. Furthermore, studies in associative memory research have demonstrated that the emotional intensity of an event has a major impact on memory performance (LaBar \& Cabeza, 2006). As we can assume that perceived risk does cause some sort of arousal or emotion (Slovic \& Peters, 2006) it seems plausible that risk cues could have a positive impact on target memory performance in OSNs. In this study we focused on two major aspects that could influence perceived vulnerability: a) the kind of information that is disclosed and \(b\) ) the kind of audience that gains access to the information.

Information Gopie and MacLeod (2009) found that target memory performance was worse when people disclosed personal facts in comparison to impersonal facts. In line with Koriat, Ben-Zur, and Druch (1991) the authors argue that revealing personal details increases self-focus which prevents people from integrating the outer environment as a reference point of that event. Impersonal information on the other hand would not trigger the same amount of self-focus resulting in better target memory performance. However, we believe that in OSNs the degree of intimacy of information serves as a distinct risk cue, because disclosing personal information gives the audience's members more opportunities for personal judgment and information misuse. We therefore predicted that target memory would improve when the disclosed information is personal rather than impersonal (hypothesis 2).

Audience Perceived vulnerability does not only vary with the nature of the information but also with the nature of the target. Thus, perceived vulnerability seems to increase with the number of people who have access to this information (Bateman, Pike, \& Butler, 2011). Slonje and Smith (2008) similarly showed that cyberbullying victims experience the unwanted disclosure by others as especially harmful when a large group gains access. Naturally, a larger group is not only more difficult to control but necessarily contains more
members that are less trusted. Publicity thus seems to be an important factor for risk awareness during online selfdisclosure. Therefore, we assumed better target memory for larger than for smaller audiences (hypothesis 3).

\section*{Hypotheses}

To summarize, this study addresses two different aspects of risk awareness and memory performance in online selfdisclosure. On the one hand we assumed that overall target memory problems also exist online. We predicted that people easily remember what information they disclosed (item memory), but not to which audience (target memory; hypothesis 1). On the other hand we presumed that people cognitively react to specific risk cues like the intimacy of the disclosed information (personal vs. impersonal information) and the size of the audience that receives the message (small vs. large audience). We therefore hypothesized that target memory would improve when information is personal rather than impersonal (hypothesis 2 ) and when the receiving audience is large rather than small (hypothesis 3).

\section*{Method}

\section*{Participants}

Participants were senior students from high schools in the area of Münster, Germany. We excluded two participants from data analysis, because they did not follow the instructions as requested. Thus, our sample consists of 99 participants ( 34 males, 65 females) with a mean age of 17.59 years ( \(\mathrm{SD}=2.08\) ).

\section*{Materials}

Scenario Students entered a fictitious social networking site of the local university. Within this site participants entered a sham discussion group where they would be posting information concerning the topic of the group. Students were aware that they were part of an experimental study.

Information In the personal condition students entered the fictitious discussion group "to get to know each other". Items in the personal condition were partly taken from former studies about relationship formation (Joinson, 2001; Jourard \& Lasakow, 1958), partly taken from what is typically disclosed in online profiles (e.g. "your favorite music") and partly self-created (e.g. "what is the meaning of life in your opinion"). In the non-personal condition students entered the sham group "information about the city of Münster". Items in this condition were taken from an online tourist brochure about the city of Münster (e.g. "famous band from Muenster - H-Blockx" or "founding year of the city of Muenster - 793").

Audiences The disclosed information would be sent to either everyone in the students' semester (large audience;

180 people) or only to their future study group (small audience; five people). We defined the size of the audiences to approximate how online social networks are arranged. In 2011 the average network of a Facebook user consisted of around 190 Facebook-friends (Ugander, Karrer, Backstrom, \& Marlow, 2011). Usually, a core group of these people are active contacts the user communicates with on a frequent basis (strong ties). The rest of the network constitutes weak ties - users passively keep in touch with these contacts, but not necessarily interact with them on a regular basis (Ellison, Steinfield, \& Lampe, 2007). Please note that the audiences in this experiment did not constitute strong and weak ties per se since participants had no actual relation to the displayed people whatsoever.

Communication Task The communication task consisted of 20 randomized slides. On each slide students saw two facts at the top of the page (personal or non-personal). They decided which one they wanted to disclose and marked that one. In the personal condition we paired facts with a similar degree of intimacy. In the impersonal condition we paired facts that both contained either numerical or textual information. The audience (small or large) was saliently displayed underneath these two facts via a collection of small-scaled photos that matched the number of the announced audience size. Participants were instructed to choose one of the two facts and disclose it at the bottom of the page where they wrote the information into an empty text field. Ten facts were disclosed to a small target audience, ten facts to a large target audience and thus twenty facts were not disclosed because they were not chosen. While the students could choose which one out of the two facts they wanted to disclose, the audience was predetermined and could not be selected. We incorporated this decisional aspect to enhance the external validity of our experiment: People presumably choose more or less carefully what information they disclose (Marsh \& Hicks, 2002) - not only for privacy reasons but also because this information becomes an inherent part of their selfpresentational strategy. Interestingly, many Facebook users would rather decide what to disclose, instead of to whom, since many report that they make all their information and actions visible to all of their Facebook-friends. Therefore, participants in our experiment could decide what information they wished to disclose but not to which audience.

Memory Task The memory task consisted of 50 randomly presented test slides that contained the forty facts of the communication task plus ten completely new facts. For each displayed fact the students indicated if they had disclosed this fact to a small target audience, a large target audience, if it was a fact they hadn't seen before (new) or if they had encountered this fact but not disclosed it (each of the first three options was correct in 10 times of the cases, the last option in 20 times of the cases). Items that were new or had
not been chosen to be disclosed in the learning phase were treated as distractors. \({ }^{1}\)

Internet Literacy Questionnaire The internet literacy questionnaire (Stodt, Moll, Polzer, Pieschl, \& Brand, 2013) consists of twenty items measuring online literacy in terms of technical skills, online empathy, online interactions, and privacy-related attitudes and behaviors. \({ }^{2}\) This questionnaire was incorporated to create latency between the communication task and the memory task and to thus weaken short term memory effects.

\section*{Procedure}

Students were recruited during an open day of the Westfälische Wilhelms-Universität Münster. Groups of students sat down in front of a computer screen to participate in the experiment. We conducted a \(2 \times 2\) experiment with information (personal versus impersonal) as a between-subject factor and audience size (small versus large) as a within-subject factor. After being welcomed the students were randomly assigned to the information conditions. They received a short description of the scenario and entered the communication task. Afterwards students answered the internet literacy questionnaire as a short filler task, being followed by an explanation on how to work on the subsequent memory task. After this memory task students shortly answered questions about their OSNs-usage and socio-demographic details. After completing the experiment students were offered the chance of winning one out of six gift cards for the online shop Amazon. Students were encouraged to leave an email-address so we could explain the purpose of the experiment after data analysis had been completed. They were then thanked for their participation and dismissed.

\section*{Results}

The random assignment to between-subject conditions was successful. Demographic details in the personal condition ( \(n=49 ; 31\) females, 18 males; \(\mathrm{M}=17.55\) years, \(\mathrm{SD}=1.02\) ) did not differ significantly from the impersonal condition, ( \(n=50 ; 34\) females, 16 males; \(\mathrm{M}=17.62\) years, \(\mathrm{SD}=2.76\) ).

\section*{Target Memory}

As a prerequisite for further analysis we assessed if there actually is a target memory problem in comparison to item memory. In order to do so, we compared the mean number of correct audience identifications (small and large

\footnotetext{
\({ }^{1}\) Additionally, students indicated on 5-point Likert-scales how confident they were about the correctness of each of their answers. These results are not reported here as they addressed a different research question that due to space constraints cannot be reported in this paper.
\({ }^{2}\) The results of this questionnaire are not part of this report as no meaningful factor structure could be found in this sample.
}
audiences) as indicator of target memory with the mean number of correct distractor identifications (not disclosed and new facts) as indicator of item memory in a repeatedmeasure ANOVA. Information was the between-subject factor. Item memory significantly exceeded target memory across both information conditions, \(F(1,98)=262.08\), \(p<.001 \eta_{\mathrm{p}}^{2}=.73\) (see Figure 1), indicating better memory for which information had (not) been disclosed than to which audience it was disclosed.


Figure 1: Percentage of correctly identified audiences and distractors per information condition (error bars indicate standard deviations).

We also found a significant main effect for information, \(F(1,97)=46.25, p<.001, \eta_{\mathrm{p}}^{2}=.32\) as well as a significant interaction between the two factors, \(F(1,97)=19.89\), \(p<.001, \eta_{\mathrm{p}}{ }^{2}=.17\). The difference of correct audience and distractor identification was larger in the personal condition than in the impersonal condition.

\section*{Risk Cues}

To test hypotheses two (target memory improves for personal in comparison to impersonal information) and three (target memory improves for large audiences in comparison to small audiences) we computed a \(2 \times 2\) repeated-measure ANOVA with audience size (small vs. large) as repeated-measure factor and information (personal vs. impersonal) as between-subject factor. Our dependent variable was the mean number of correct target identifications in each condition. We found a significant main effect of information, \(F(1,97)=6.15, p<.015\), \(\eta_{\mathrm{p}}{ }^{2}=.06\). Memory performance in the personal condition exceeded performance in the impersonal condition regardless of audience size (see Figure 2). Furthermore, we found a significant main effect for audience size, \(F(1,97)=51.044, p<.001, \eta_{\mathrm{p}}{ }^{2}=.35\). Memory performance was better when the target audience was large opposed to small - regardless of information (see Figure 2). The interaction of the two factors was not significant, \(F(1,97)=.43, p=.51, \eta_{\mathrm{p}}{ }^{2}=.00\). The descriptive results of the memory test also indicate that students had a general
answering bias; students in both information conditions overall answered "large audience" more frequently than "small audience" (see row "Total chosen" in Table 1).


Figure 2: Percentage of correctly identified targets for the experimental factors audience (X-axis) and information (error bars indicate standard deviations).

\section*{Discussion}

\section*{Overall Target Memory}

Participants correctly identified significantly more distractors (new and not disclosed facts) than they identified the associated audiences of disclosed information (small and large audience). We thus confirmed our first hypothesis that item memory would be superior to target memory: Students struggled to remember what information they had disclosed to which audience. Thus our study shows that target memory problems exist online and might contribute to repeated privacy-neglecting behavior in OSNs: Without the memory of what audience has access to which information the cumulative risk of online self-disclosure must be constructed on an abstract level that is weighed out by the immediate benefits of the same behavior. Our study thus not only expands the realm of target memory research but also contributes to further explanations of the circumstances under which privacy-related decisions are made in online environments. Interestingly, error rate analysis shows that participants mainly confused the audiences or the distractors, but rarely identified a disclosed piece of information as a distractor or a distractor as having been disclosed (see Table 1). Thus participants were well aware of what they disclosed, but had trouble remembering to whom. This finding holds the encouraging notion that OSNs users are not blindly "sharing in the dark" - they are not disclosing information without any memory of past revelations whatsoever.

Table 1: Percentages of chosen options per information condition in the memory task.
\begin{tabular}{|c|c|c|c|c|c|}
\hline & \multirow{2}{*}{Correct Response} & \multicolumn{4}{|c|}{Responses given by participants} \\
\hline & & Small & Large & Not Discl. & New \\
\hline \multirow{5}{*}{\begin{tabular}{l}
Personal \\
Information
\end{tabular}} & Small & 49\% & 43\% & 7\% & 1\% \\
\hline & Large & 27\% & 71\% & 1\% & 1\% \\
\hline & Not Discl. & 3\% & 3\% & 84\% & 10\% \\
\hline & New & 1\% & 0\% & 1\% & 98\% \\
\hline & Total chosen & 17\% & 24\% & 35\% & 24\% \\
\hline \multirow{5}{*}{Impersonal Information} & Small & 44\% & 46\% & 5\% & 5\% \\
\hline & Large & 29\% & 62\% & 6\% & 3\% \\
\hline & Not Discl. & 4\% & 5\% & 51\% & 40\% \\
\hline & New & 1\% & 1\% & 9\% & 89\% \\
\hline & Total chosen & 16\% & 24\% & 25\% & 35\% \\
\hline
\end{tabular}

Note: All percentages pertain to rows; for 'Small', 'Large', and 'New' 10 answers were given per participant, for 'Not Discl.' 20 answers were given; 'Not Discl.' is the abbreviation of 'Not Disclosed'; bold marked numbers indicate percentages of correct identifications.

\section*{The Impact of Risk Cues on Target Memory}

Participants identified more target audiences correctly when they informed the audiences about something personal in comparison to impersonal information. We hereby confirmed our second hypothesis. In line with this finding we also found that distractor identification in the personal condition was superior to the impersonal condition (see Figure 1). In line with hypothesis 3, target memory performance also varied with the target audience of the disclosed message. Participants remembered more targets correctly when the audience was large. Furthermore, response rates show that the risk of disclosing something to a large audience seems to be especially salient, since participants more often chose "large audience" in the memory task than "small audience". We can conclude from our results that people are not oblivious to online risks but show a direct cognitive reaction to situational vulnerabilities like telling personal information or telling something to a large audience. It might be that people process the association of target and information in a more elaborate way when they feel vulnerable or when they cannot trust their interaction partners.

\section*{Limitations}

Naturally, this experiment has limitations that we need to consider when interpreting our results. For one, our sample is a non-representative convenience sample. Thus, we do not know if our results can be generalized to other age and
educational groups. Furthermore, in order to transfer the paradigm of target memory to the environment of OSNs we had to make several alterations from the conventional offline paradigm. These alterations restrict a direct comparison of our findings with results from former studies but substantially enhance the ecological validity of our experiment: First, as we could not find an up-to-date validated collection of intimacy-rated items we created new stimulus material for the information conditions. We thereby focused on information that is typically disclosed in online profiles as well as on details about participants' biographical and attitudinal characteristics. The intimacy of these items varies substantially and further research is needed to assess the perceived intimacy of information in different interaction contexts as well as the role of possible self-reference effects in online environments. Second, we changed the classic operationalization of the target in our experiment. Usually a target consists of one single person represented by a photo or name. However, in OSNs users seldom communicate in one-to-one situations but rather address different kinds of audiences. Therefore, it seemed appropriate to adjust the receiving targets so that information would be disclosed to two different kinds of audiences (small vs. large). In this respect it also seems important to note that our experimental design did not allow manipulations of audience familiarity. Therefore, future research is needed to assess the generalizability of our results to real social network communication where people
usually know their audience's members from offline contexts. Finally, our results do not fully explain the underlying cognitive mechanisms that contribute to better memory performance in risk situations. Future studies should therefore attempt to clarify this issue, for example by controlling for decision times in the communication task.

\section*{Implications}

Our results show that users of OSNs actually do react to specific risk circumstances, if these are salient enough to be grasped. This indicates that users probably do not just claim to be concerned about their data (which often contradicts their behavior) but seem to automatically pay more attention to vulnerable situations in online communication. This possibility of a more thorough elaboration offers a direct practical link: From a technical view, privacy-supporting web applications should work on a less subtle and more realistic representation of the potential audience of the to-be-disclosed information. Furthermore, it seems useful to work on ways in which people get a quick overview about what they have disclosed in the past and to whom it is visible. From an educational standpoint, internet literacy programs should sensitize participants to rather subtle online risk cues, for example the degree of publicity. However, these measures cannot and should not stop users from self-disclosing in OSNs altogether since a considerable amount of research also suggests that OSNs-users benefit both emotionally and socially from their usage. The aim of design alterations and educational measures should rather be to achieve a natural consciousness so that privacy-related decisions can be beneficial after all.

\section*{Acknowledgements}

This study was funded by the Research Training Group 1712/1 "Trust and Communication in a Digitized World" of the German Research Foundation (DFG). We would like to thank Christina Wohlers and Franziska Thon for their comments on this manuscript.

\section*{References}

Bateman, P. J., Pike, J. C., \& Butler, B. S. (2011). To disclose or not: publicness in social networking sites. Information Technology \& People, 24(1), 78-100.
Brown, A., Hornstein, S., \& Memon, A. (2006). Tracking Conversational Repetition: An Evaluation of Target Monitoring Ability. Applied Cognitive Psychology, 20, 85-95.
Ellison, N. B., Steinfield, C., \& Lampe, C. (2007). The benefits of Facebook 'friends:' Social capital and college students' use of online social network sites. Journal Of Computer-Mediated Communication, 12(4), 1143-1168.
Gopie, N., \& MacLeod, C. M. (2009). Destination Memory - Stop me if I told you this before. Psychological Science, 20(12), 1492-1499.

Joinson, A. N. (2001). Knowing me, knowing you: Reciprocal self-disclosure in Internet-based surveys. Cyberpsychology \& Behavior, 4(5), 587-591.
Jourard, S. M., \& Lasakow, P. (1958). Some factors in selfdisclosure. Journal of Abnormal and Social Psychology, 56(1), 91.
Kiesler, S., Siegel, J., \& McGuire, T. W. (1984). Social Psychological Aspects of Computer-Mediated Communication. American Psychologist, 39(10), 11231134.

Koriat, A., Ben-Zur, H., \& Druch, A. (1991). The contextualization of input and output events in memory. Psychological Research, 53, 260-270.
LaBar, K. S., \& Cabeza, R. (2006). Cognitive neuroscience of emotional memory. Nat Rev Neurosci 7, 54-64.
Marsh, R. L., \& Hicks, J. L. (2002). Comparisons of Target Output Monitoring and Source Input Monitoring. Applied Cognitive Psychology 16 845-862
Norberg, P. A., Horne, D. R., \& Horne, D. A. (2007). The Privacy Paradox: Personal Information Disclosure Intentions versus Behaviors. The Journal of Consumer Affairs, 41(1), 100-126.
Petty, R. E., Cacioppo, J. T., \& Kasmer, J. A. (1988). The role of affect in the elaboration likelihood model of persuasion. In L. Donohew, H. E. Sypher \& E. Higgins (Eds.), Communication, social cognition, and affect. Hillsdale, NJ England: Lawrence Erlbaum Associates, Inc.
Slonje, R., \& Smith, P. K. (2008). Cyberbullying: Another main type of bullying? Scandinavian Journal of Psychology, 49, 147-154.
Slovic, P., \& Peters, E. (2006). Risk Perception and Affect. Current Directions In Psychological Science, 15(6), 322325.

Stodt, B., Moll, R., Polzer, C., Pieschl, S., \& Brand, M. (2013). First results of a new questionnaire to assess Internet literacy: Correlations to pathological Internet use and risk-taking behavior. Paper presented at the TeaP, Vienna.
Ugander, J., Karrer, B., Backstrom, L., \& Marlow, C. (2011). The Anatomy of the Facebook Social Graph. arXiv:1111.4503, 1-17.
Valkenburg, P. M., \& Peter, J. (2011). Online Communication Among Adolescents: An Integrated Model of Its Attraction, Opportunities, and Risks. Journal of Adolescent Health 48, 121-127.
Xu, H., Teo, H.-H., Tan, B. C. Y., \& Agarwal, R. (2010). The Role of Push-Pull Technology in Privacy Calculus: The Case of Location-Based Services. Journal of Management Information Systems, 26(3), 135-173
Zeidner, R. (2007). How Deep Can You Probe. HR Magazine, 52(10), 57-60.

\title{
Automatic Generation of Music for Inducing Physiological Response
}

\author{
Kristine Monteith (kristine.perry@gmail.com) \\ Department of Computer Science
}

Bruce Brown(bruce_brown@byu.edu)
Department of Psychology

\author{
Dan Ventura(ventura@cs.byu.edu) and Tony Martinez (martinez@cs.byu.edu) \\ Department of Computer Science \\ Brigham Young University \\ Provo, UT 84602 USA
}

\begin{abstract}
Music is known to have a profound impact on human cognitive and emotional response, which in turn are strongly correlated with physiological mechanisms. This paper presents a system that is designed to create original musical compositions that elicit particular physiological responses. The experiments described below demonstrate that the music generated by this system is as effective as human-composed music in effecting changes in skin resistance, skin temperature, breathing rate, and heart rate. The system is particularly adept at composing pieces that elicit target responses in individuals who demonstrated predictable responses to training selections.
\end{abstract}

Keywords: music; emotion; perception; cognition; physiological response; targeted response

\section*{Introduction}

Music can have a profound impact on human physiology. It affects how we think, how we feel, and how we relate to others. It captivates and holds our attention, stimulating many areas of the brain. From movie scenes to dance floors, the added sensory input of music makes activities and situations more enjoyable and compelling. One study found that pleasurable music activated the same areas of the brain activated by other euphoric stimuli such as food, sex, or drugs. They highlight the significance of the fact that music would have a similar effect on the brain as "biologically relevant, survivalrelated stimuli" (Blood \& Zatorre, 2001).

Music's impact on human physiology may help explain its long-recognized ability to sway human emotion. It provides not only a medium for expressing a particular emotion, but also the accompanying physiological change to add significance and depth to that emotion. According to the SchachterSinger theory, emotion is a function of both physiological arousal and cognitive interpretation of that response. The degree of arousal is associated with the degree of emotional response, but it is up to the individual to label that response according to past experience (Schachter \& Singer, 1962).

Music can also have significant power to calm the body and mind. While relaxation responses such as lowered breathing and heart rate may not be as closely tied with emotional perception and cognition, their elicitation can often have significant therapeutic benefits. Numerous studies have demonstrated the ability of music to induce a relaxation response (e.g White, 1999; Lepage et al., 2001; Khalfa et al., 2002).

Both speed and accuracy of task performance can be enhanced with relaxing music (Allen \& Blascovich, 1994).

While there is little question about whether or not music has an effect on humans, predicting the precise effect is challenging. A few effects, however, do seem to be relatively consistent. For example, one study found that more complex rhythms tended to increase the rate of autonomic functions such as breathing and cardiovascular activity. Silence tended to have the opposite effect-lowering breathing rates and heart rates (Bernardi et al., 2006). White (1999) found that heart rate, respiratory rate, and myocardial oxygen demands were lower among patients recovering from myocardial infarctions; Khalfa et al. (2002) found that arousal responses were more likely with pieces that the subjects found to communicate happiness or fear, while pieces described as sad or peaceful tended to decrease arousal. However, even these results only hold true for a majority of individuals. Finding a piece of music that would reliably effect a desired physiological response in a given individual remains a considerable challenge.

Computer-generated music (Chuan \& Chew, 2007; Cope, 2006) may provide some advantages in addressing this challenge. Computers are well-suited to sifting through a large number of both large-scale and fine-grained musical features and to keeping track of which features will most likely have a particular effect. Indeed, some work has been done in generating music to target a listener emotion or mood (Delgado et al., 2009; Rutherford \& Wiggins, 2003; Oliveira \& Cardoso, 2007). In addition, a human composer might be more biased towards features that would effect his or her own physiology when producing compositions. While a reliance on one's own physiological experiences may be inspiring and helpful in the creative process, when it comes to eliciting physiological responses from others, it may also sometimes result in pieces that are less generalizable. Additionally, once they have "learned" how to do so, computers can generate large quantities of music at virtually no cost in terms of time or effort. A computer would have a much easier time generating a number of different potential compositions to effect a desired result in a given individual until it happened upon the right one. Therefore, the ability of a computer to compose music that elicits a target response could have significant benefits.

This paper presents a system capable of generating selections designed to elicit desired physiological responses. Data collected in biofeedback experiments with 96 different subjects show that the system is able to generate selections that elicit an average change in a target physiological response with roughly the same ability level as a human performing the same task. The system is particularly effective at eliciting such a response if an individual's response to other musical selections is known.

\section*{Methodology}

Our approach can be decomposed into three major components: selection of musical pieces to use as training data for our generative models, construction of those models using the training data and evaluating the effectiveness of the models in eliciting the target response when compared to humangenerated music designed for the same targeting task.

\section*{Training Data Selection}

Seventy-two MIDI files were downloaded from the Free MIDI File Database. \({ }^{1}\) Themes from movie soundtracks were used due to the wider variety of emotional content available in this genre. The first forty-five seconds of each piece was isolated for use in experiments.

Biofeedback experiments were conducted to determine effective candidate training pieces. In our preliminary experiments, forty-eight subjects were asked to listen to a number of different training pieces while their heart rate, breathing rate, skin resistance, and skin temperature were monitored. Physiological responses were recorded using the I-330-C2+ biofeedback machine manufactured by J\&J Engineering. All were university-enrolled students or professors. Subjects ranged in age from 18 to 52, with the average age being 22 . Thirty-four males and 14 females participated.

The seventy-two MIDI selections were split into six groups of twelve selections, and each group of songs was played for eight people. \({ }^{2}\) At the beginning of experiments, forty-five seconds of baseline readings were collected. (Subjects were asked to sit quietly and count upwards in their minds during this time in order to achieve neutral results.) Measurements were sampled at one second intervals. For each of the physiological measures, responses were averaged for the duration of baseline readings and for the duration of each of the fortyfive second song samples. Then, a \(z\)-score was calculated for each of the selections, indicating how many standard deviations the average for a given song varied from the baseline.

Responses were then analyzed to determine which selections were most likely to affect a given physiological response. A corpus of training songs comprised of the selections that elicited the largest average change in response was then created for each of the measures studied.

\footnotetext{
\({ }^{1}\) http://themes.mididb.com/movies/
\({ }^{2}\) While the song grouping could likely have been randomly assigned without significantly affecting the results, an attempt was made to make the groupings as similar as possible.
}

\section*{Automatic Music Generation}

Each generative model (one for each targeted response) is composed of four separate modules, for producing rhythm, pitch, harmony and accompaniment.

Rhythm Generator The rhythm for the selection is generated simply by selecting phrases from randomly chosen selections in the training set and stochastically perturbing them. Each new rhythmic phrase is evaluated by two decision tree Rhythm Evaluators (described below). Generated phrases are only used if they are classified positively by both classifiers.
Pitch Generator Once the rhythm is determined, pitches are selected for the melodic line using a probabilistic \(n\)-gram model of melodic progression built from the training corpus. The system generates one hundred possible series of pitches for each rhythmic phrase, and each of the melodies is evaluated by two decision tree Pitch Evaluators (see below). Generated melodies are only selected if they are classified positively by both classifiers.
Harmony Generator The underlying harmony is determined using a hidden Markov model, with melody notes considered as observed events and the chord progression as the latent state sequence. The probability distributions for populating the model are estimated using statistics gathered from the corpus of music representing the target response.

Accompaniment Planner To generate accompaniment, the system takes as input a measure from a song in the training corpus outlining a characteristic baseline, percussion track, and instrumentation. These act as style files for the computergenerated selections - each measure is transposed according to the generated chord pattern, producing accompaniments in much the same manner as a pianist selecting a given style on an electronic keyboard.

Decision Tree Evaluators A set of two evaluators is developed for interacting with the rhythm module and another set of two evaluators is developed for interaction with the pitch module. The first classifier in each set is trained using analyzed selections in the target corpus (e.g. raise heart rate) as positive training instances and analyzed selections from the other corpora (e.g. the other seven responses, lower heart rate, raise breathing rate, etc.) as negative instances. This is intended to help the system distinguish selections that elicit specific physiological response. The second classifier in each set is trained with melodies from all corpora versus thirty-two unevaluated melodies previously generated by the algorithm. In this way, the system learns to distinguish melodies which have already been accepted by human audiences. An example decision tree (identifying features for eliciting raised heart rate response) developed for evaluating the pitch assignment model is shown in Figure 1.

\section*{Evaluation}

A second round of biofeedback experiments was conducted to evaluate the generated musical selections. Forty-eight ad-

\section*{Raise Heart Rate}

ClimaxPosition \(<=0.67\)
- Dissonance \(<=0.01\)
—— PitchMovementByTonalStep \(<=0.63\) : No
—— PitchMovementByTonalStep \(>0.63\) : Yes
- Dissonance \(>0.01\) : Yes

ClimaxPosition > 0.67: No

Figure 1: Decision tree model of musical characteristics contributing to raised heart rate
ditional subjects participated in this evaluation phase. Again, all were university-enrolled students or professors. Subjects ranged in age from 17 to 46, with the average age being 22. Twenty males and 28 females participated.

Physiological responses were recorded for twenty-four songs (eight computer-generated selections, eight training selections for reference, and eight human-composed selections). To prevent subject fatigue, selections were divided into two groups, one group consisting of pieces targeted to affect breathing and heart rate and one group consisting of pieces targeted to affect skin resistance and skin temperature, and subjects were only asked to listen to one of the groups. Each subject listened to twelve selections; each piece was played for twenty-four people. A Cronbach's alpha coefficient (Cronbach, 1951) was calculated on the responses of subjects in each group to test for inter-rater reliability. Coefficients for the two groups were both \(\alpha=0.99\). (Values over 0.80 are generally considered indicative of a reasonable level of reliability and consequently, a sufficient number of subjects for testing purposes.)

Baseline readings were collected at the beginning of each recording session. Responses were averaged for the duration of baseline readings and for the duration of each of the selections. Since some individuals were more reactive than others, \(z\)-scores are used in analysis instead of absolute changes in measurement. \({ }^{3}\)

After listening to each selection, subjects were asked to respond to the following questions (on a scale from 1 to 9 ):
1. Did you like the selection?
2. How familiar were you with the selection?
3. How musical was the selection?
4. How original was the selection?

\section*{Results}

This section provides tables reporting the average \(z\)-scores for selections designed to elicit the various target physiolog-

\footnotetext{
\({ }^{3}\) Recall that \(z\)-scores calculate the number of standard deviations an average varies from a given baseline. They are calculated by the formula \(z=(x-\mu) / \sigma\), where \(x\) is the average for a given selection, \(\mu\) is the average for baseline, and \(\sigma\) is the standard deviation for readings taken over the duration of the session. Please note that, while \(z\)-scores are sometimes used to calculate statistical significance, in this case, these measures are only being used to normalize scores from one individual to the next. A high Cronbach's alpha value for a low average \(z\)-score indicates that, while a given selection did not tend to elicit a high magnitude change in a response, it was consistent in eliciting a given change for a significant number of subjects.
}

Table 1: Average \(z\)-scores of computer and human-generated selections designed to affect breathing rate
\begin{tabular}{cccc}
\multicolumn{4}{c}{ Lower Breathing Rate } \\
\hline & Overall & \multicolumn{2}{c}{ Adjusted } \\
Average & Included \\
Computer-Generated & -0.27 & -1.33 & \(29 \%\) \\
Human-Composed & 0.13 & -0.90 & \(29 \%\)
\end{tabular}
\begin{tabular}{cccc}
\multicolumn{4}{c}{ Raise Breathing Rate } \\
\hline & Overall & \multicolumn{2}{c}{ Adjusted } \\
Average & Included \\
Computer-Generated & 0.71 & 1.18 & \(46 \%\) \\
Human-Composed & 0.06 & 0.36 & \(46 \%\)
\end{tabular}
ical responses. In most cases, both the computer-generated and human-composed selections were effective at eliciting arousal responses. However, they were less effective at eliciting relaxation responses. This is not surprising considering findings suggesting that music is often more effective than silence at eliciting an arousal response (Bernardi et al., 2006).

Many of the more conclusive studies on the relaxing effects of music deal with subject-selected pieces. Since both the computer-generated and human-composed selections being evaluated are unique to these experiments, subjects would not associate any of them with previous relaxing experiences and consequently experience a relaxation response due to classical conditioning. It would also be difficult for any of the subjects to identify ahead of time which pieces they would find most relaxing. Instead, we look at how subjects responded to the training selections. Each table also reports an adjusted score, calculated by averaging only measurements for individuals for whom the training selections also had the target effect for the measure being considered (reported in the tables as a percentage of the 24 total people that listed to the selection). While a computer-generated piece may not be able to elicit a particular physiological response in all subjects, this adjusted score allows us to measure whether it elicits a response in a specific group of subjects. (e.g. If it is known that a group of individuals react with a lowered breathing rate to a given song or set of songs, the adjusted score reveals how effective the computer might be in using those training pieces to generate a song that also lowers breathing rate.)

\section*{Breathing Rate}

Breathing rate responses tended to vary by up to one breath per minute. (Considering that normal human breathing rates tend to range from 12 to 18 breaths per minute, an average increase of one breath per minute is non-negligible.) The most significant changes tended towards an increase in breathing rates as compared to baseline.

As shown in Table 1, only the computer-generated selection was able to successfully lower breathing rate on the average for all subjects. However, the magnitude of the change was small enough that the average change was not significantly different from the human-composed selections. With the adjusted scores, both computer-generated and human-
composed songs were able to successfully lower breathing rates. Seven individuals- \(29 \%\) of subjects in this groupresponded as expected to the top training selection for lower breathing rate; four responded similarly to the computergenerated selection.

The computer-generated song designed to raise breathing rate was able to accomplish this task more effectively than the human-composed song. The \(0.71 z\)-score for the computergenerated song corresponds to an average increase of over one breath per minute, and the difference in average \(z\)-scores between this and the human-generated song was significant at the \(p<0.05\) level. A similar pattern is seen with the adjusted scores. The average difference between the computergenerated selection and the human-composed song was also significant. Nine of the eleven individuals who responded with elevated breathing rate to training selections targeted to raise breathing rate responded similarly to the computergenerated selections.

Note that the computer-generated selections designed to lower breathing rate are as effective at doing so as the humancomposed selections. The computer-generated selections designed to raise breathing rate are performing this task at a level that exceeds that of human performance.

\section*{Heart Rate}

Changes in average heart rate were not quite as pronounced. While individual heart rates could vary by up to fifty beats per minute over the course of a session, the average range for a given individual was only ten beats per minutes. When averaged over all subjects, reactions to songs only varied by a couple of beats per minute.

As shown in Table 2, only the human-composed selection was able to reduce average heart rate, although the difference in mean heart rate variation was not significant at the \(p<\) 0.05 level. With the adjusted scores, the computer-generated selection proved more effective at lowering heart rate. For five of the eight individuals whose heart rate lowered for the top training selection, heart rates were also lowered for the computer-generated songs in these categories.

The computer-generated song was the most effective at raising average heart rate for all subjects, though the difference was not statistically significant. The computergenerated song was also more effective at raising heart rate using the adjusted score, but not significantly so. Ten of the thirteen individuals who responded as expected to the training selection for raising heart rate also had their heart rates raised by the computer-generated selection.

As with breathing rate, the computer appears to be addressing the task of composing music that lowers or raises heart rate at a level comparable to that of human performance.

\section*{Skin Temperature}

Skin temperature tended to rise, on average, by two degrees during the course of the session for most subjects, regardless of the piece of music being played. Not surprisingly, all selections were better at raising average skin temperature for all

Table 2: Average \(z\)-scores of computer and human-generated selections designed to affect heart rate

Lower Heart Rate
\begin{tabular}{cccc}
\hline & \multirow{2}{c}{ Overall } & \multicolumn{2}{c}{ Adjusted } \\
Average & Included \\
Computer-Generated & 0.40 & -0.40 & \(33 \%\) \\
Human-Composed & -0.20 & -0.61 & \(33 \%\)
\end{tabular}

Raise Heart Rate
\begin{tabular}{cccc}
\hline & \multicolumn{2}{c}{ Overall } & \multicolumn{2}{c}{ Adjusted } \\
Average & Included \\
Computer-Generated & 0.72 & 1.09 & \(54 \%\) \\
Human-Composed & 0.12 & 0.53 & \(54 \%\)
\end{tabular}

Table 3: Average \(z\)-scores of computer and human-generated selections designed to affect skin temperature
\begin{tabular}{cccc}
\multicolumn{4}{c}{ Lower Skin Temperature } \\
\hline & Overall & \multicolumn{2}{c}{ Adjusted } \\
Average & Included \\
Computer-Generated & 2.18 & -1.22 & \(17 \%\) \\
Human-Composed & 1.23 & -1.84 & \(17 \%\)
\end{tabular}

Raise Skin Temperature
\begin{tabular}{cccc}
\multicolumn{4}{c}{ Raise Skin Temperature } \\
\hline & Overall & \multicolumn{2}{c}{ Adjusted } \\
Average & Included \\
Computer-Generated & 2.22 & 3.03 & \(83 \%\) \\
Human-Composed & 1.75 & 2.49 & \(83 \%\)
\end{tabular}
subjects than they were at lowering it.
However, when individual subjects did have their skin temperature lowered by a training set selection, they also tended to have their skin temperature lowered by pieces generated from those selections. This was true for all four of the individuals whose temperature was lowered by the training selection targeting lower skin temperature. The adjusted score for the human-composed selection designed to lower skin temperature was lower than the adjusted score for the computergenerated piece, but the difference was not statistically significant at the \(p<0.05\) level.

The computer-generated piece was significantly more effective at raising skin temperature than the human-composed pieces when considering both the regular and the adjusted averages. However, this is almost certainly an artifact of the order in which the pieces were played. (The software used in these experiments did not allow for a randomized order of selection presentation that was unique to each subject.)

While it appears that an effective method of raising skin temperature would simply be composing a piece with sufficient duration, the computer seems as competent at the task as a human. Composing music that lowers skin temperature appears to be a much harder task, but again, these experiments show no statistically significant difference between the performance of the computer and the human.

Table 4: Average \(z\)-scores of computer and human-generated selections designed to affect skin resistance

Lower Skin Resistance
\begin{tabular}{cccc}
\hline & \multicolumn{3}{c}{ Overall } \\
Adjusted \\
Computer-Generated & -0.87 & -2.48 & Included \\
Human-Composed & -1.06 & -2.00 & \(63 \%\) \\
\multicolumn{4}{c}{ Raise Skin Resistance } \\
\hline \multicolumn{4}{c}{ Adjusted } \\
Computer-Generated & Overall & Average & Included \\
Human-Composed & -1.06 & 2.27 & \(33 \%\) \\
& -1.03 & 0.21 & \(33 \%\)
\end{tabular}

\section*{Skin Resistance}

Most of the selections were likely to elicit an arousal response (lower skin resistance). However, unlike skin temperature, the effect was not cumulative over the course of the session.

For compositions designed to lower skin resistance, there was no significant difference between the computergenerated selection and the human-generated selection. The training selections lowered skin resistance in fifteen individuals. With the adjusted scores, computer-generated selections were more successful at lowering skin resistance than the human-composed song, though the difference was not statistically significant.

There was also no significant difference between the computer-generated selection designed to raise skin resistance and the human-composed selection. The training selection raised skin resistance in eight individuals and those subjects for whom it did have the target effect also reacted strongly to the selection generated from all the training soundtracks, with the improvement over the humancomposed selection being significant at the \(p<0.05\) level.

As with the other measures, the computer is able to generate music that elicits change in skin resistance as effectively or more effectively than a human composition.

\section*{Subjective Responses}

Average responses to the subjective questions asked after each selection are shown in Table 5. Not surprisingly, the initial training selections and the human-composed selections received higher rating for likability and musicality. However, the computer-generated selections received slightly higher ratings for originality and significantly lower ratings for familiarity than the training selections and human-composed selections-evidence to suggest that the computer is producing genuinely original compositions and not borrowing too heavily from training data.

As shown in Table 6, there was no correlation between subjective responses and physiological changes. While for some individuals, liking a song might result in a more dramatic increase or decrease in a given physiological response, this does not appear to be the case overall.

Table 5: Average results to subjective questions (Responses were measured on a scale of 1 to 9 )
\begin{tabular}{cc}
\multicolumn{2}{c}{ Did you like the selection? } \\
\hline Training Selections & 5.83 \\
Computer-Generated Selections & 3.97 \\
Human-Composed Selections & 5.56
\end{tabular}

How familiar was the selection?
Training Selections \(\quad 5.53\)
Computer-Generated Selections 2.17
Human-Composed Selections 3.01
How musical was the selection?
Training Selections \(\quad 5.35\)
Computer-Generated Selections 3.88
Human-Composed Selections \(\quad 5.12\)
How original was the selection?
\begin{tabular}{cc}
\hline Training Selections & 6.36 \\
Computer-Generated Selections & 6.97 \\
Human-Composed Selections & 6.70
\end{tabular}

Table 6: Correlations between subjective responses and physiological measures
\begin{tabular}{ccccc} 
& Like & Familiar & Musical & Original \\
\hline Breathing Rate & 0.02 & 0.03 & 0.03 & -0.04 \\
Heart Rate & 0.03 & -0.01 & 0.04 & 0.09 \\
Skin Temperature & 0.04 & 0.04 & 0.00 & -0.06 \\
Skin Resistance & -0.03 & -0.05 & -0.02 & -0.03
\end{tabular}

\section*{Musical Features}

Musical characteristics identified by the evaluating decision trees as being responsible for various physiological responses may be only briefly touched on here. Pieces that raised heart rates tended to have more dissonance and more scale-wise movement. Pieces that lowered heart rate, on the other hand, tended to have less rhythmic variety (perhaps contributing to more flowing rhythms) and a stronger climax.

Melodies that tended to raise breathing rate tended to higher rhythmic variety and either a non-tonal climax note or lower climax strength. Somewhat surprisingly, melodies that lowered breathing rate also tended to have higher rhythmic variety, but also some syncopation and a tendency to upward pitch direction.

Features contributing to a lowered skin temperature response included stability of melodic direction and a non-tonal climax. In other words, upward movement towards a climax that involved a non-tonal suspension note were arousing. A greater pitch range also contributed to lowered skin temperature. Pitch movement by minor tonal step leading to a strong climax tended to contribute to raised skin temperature.

Melodies that tended to lower skin resistance had lower pitch variety and less stability of melodic direction; some of these arousing melodies tended to bounce back and forth between notes. Melodies that raised skin resistance had a greater stability of melodic direction, as well as less rhyth-

Table 7: Ability to elicit arousal response via musical stimuli ( \(\mathrm{RBR}=\) raise breathing rate; \(\mathrm{RHR}=\) raise heart rate; \(\mathrm{LST}=\) lower skin temperature; \(\mathrm{LSR}=\) lower skin resistance)
\begin{tabular}{ccccc} 
& RBR & RHR & LST & LSR \\
\hline Computer-Generated & \(\checkmark *\) & \(\checkmark\) & & \(\checkmark\) \\
Human-Composed & \(\checkmark\) & \(\checkmark\) & & \(\checkmark\) \\
\hline Computer-Generated (Adjusted) & \(\checkmark^{*}\) & \(\checkmark\) & \(\checkmark\) & \(\checkmark\) \\
Human-Composed (Adjusted) & \(\checkmark\) & \(\checkmark\) & \(\checkmark\) & \(\checkmark\) \\
\hline
\end{tabular}
mic variety and range.

\section*{Conclusion}

Tables 7 and 8 summarize how effective we were at eliciting a change in physiological responses in various situations.

Neither the computer-generated nor the human-composed selections were able to lower average skin temperature, but both computer-generated and human-composed selections designed to elicit the other arousal responses (raised breathing rate, raised heart rate, and lowered skin resistance) were, on average, able to do so successfully. In the case of breathing rate, the computer generated song was able to raise breathing rate more effectively than the human-composed song at a level that was significant (marked with asterisk).

When considering only subjects who responded as expected to the training selections, both the computer-generated and human composed songs were successful at eliciting an average arousal response for all of the measures studied. For breathing rate and skin resistance, the difference between the computer-generated selection and the human-composed selection was significant, the computer-generated one again being more effective at eliciting the target response.

Eliciting relaxation responses proved more challenging for both the computer-generated and human-composed selections. Both were able to raise skin temperature, but neither was able to raise skin resistance. Only the computergenerated selection was able to lower heart rate, and only the human-composed selection was able to lower breathing rate. The difference between the computer-generated and humancomposed songs was not statistically significant.

When considering adjusted scores, both the computergenerated and human-composed selections were able to elicit all target relaxation responses. In the case of skin resistance, the computer-generated song was significantly better at raising average response.

Overall, the system proves itself able to generate songs that elicit target physiological responses with similar effectiveness to songs generated by a human composer. Both still require information about a given individual's physiological responses in order to generate a new piece that also reliably elicits those responses in many categories. However, given the variability of human biofeedback responses, the ability to consistently effect targeted physiological responses under any conditions can be viewed as fairly impressive.

Table 8: Ability to elicit relaxation response via musical stimuli (LBR = lower breathing rate; LHR = lower heart rate; RST \(=\) raise skin temperature; \(\mathrm{RSR}=\) raise skin resistance)
\begin{tabular}{ccccc} 
& LBR & LHR & RST & RSR \\
\hline Computer-Generated & \(\checkmark\) & & \(\checkmark\) & \\
Human-Composed & & \(\checkmark\) & \(\checkmark\) & \\
\hline Computer-Generated (Adjusted) & \(\checkmark\) & \(\checkmark\) & \(\checkmark\) & \(\checkmark *\) \\
Human-Composed (Adjusted) & \(\checkmark\) & \(\checkmark\) & \(\checkmark\) & \(\checkmark\) \\
\hline
\end{tabular}

\section*{References}

Allen, K., \& Blascovich, J.(1994). Effects of music on cardiovascular reactivity among surgeons. Journal of the American Medical Association, 272(11), 882-884.
Bernardi, L., Porta, C., \& Sleight, P. (2006). Cardiovascular, cerebrovascular, and respiratory changes induced by different types of music in musicians and non-musicians: the importance of silence. Heart, 92, 445-452.
Blood, A. J., \& Zatorre, R. J. (2001). Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion. Proceedings of the National Academy of Sciences, 98(20), 11818-11823.
Chuan, C., \& Chew, E. (2007). A hybrid system for automatic generation of style-specific accompaniment. In Proceedings of the International Joint Workshop on Computational Creativity (p. 57-64).
Cope, D. (2006). Computer Models of Musical Creativity. Cambridge, Massachusetts: The MIT Press.
Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16(3), 297-334.
Delgado, M., Fajardo, W., \& Molina-Solana, M. (2009). Inmamusys: Intelligent multi-agent music system. Expert Systems with Applications, 36(3-1), 4574-4580.
Khalfa, S., Peretz, I., Blondin, J., \& Manon, R.(2002). Eventrelated skin conductance responses to musical emotions in humans. Neuroscience Letters, 328(2), 145-149.
Lepage, C., Drolet, P., Girard, M., Grenier, Y., \& DeGagne, R. (2001). Music decreases sedative requirements during spinal anesthesia. Anesthesia-Analgesia, 93, 912-916.
Oliveira, A., \& Cardoso, A. (2007). Towards affectivepsychophysiological foundations for music production. In Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction (p. 511-522).
Rutherford, J., \& Wiggins, G. (2003). An experiment in the automatic creation of music which has specific emotional content. In Proceedings of MOSART Workshop on Current Research Directions in Computer Music (p. 35-40).
Schachter, S., \& Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. Psychological Review, 69, 379-399.
White, J. M. (1999). Effects of relaxing music on cardiac autonomic balance and anxiety after acute myocardial infarction. American Journal of Critical Care, 8(4), 220-230.

\title{
How Healthy Aging and Dementia Impact Memory Search
}

\author{
Ana Sofia Morais (morais@mpib-berlin.mpg.de) \& Hansjörg Neth (neth@mpib-berlin.mpg.de) \\ Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany \\ Thomas Hills (T.T.Hills@warwick.ac.uk) \\ Dept. of Psychology, University of Warwick, Coventry CV4 7AL, UK
}

\begin{abstract}
We model the semantic recall sequences of 424 older adults aged between 69 to 103 years in the animal fluency task. Our results suggest that, under normal intellectual functioning, memory search in old age (69-84 years) is consistent with a dynamic process that switches between retrieval probes. With dementia and very old age (85-103 years), however, memory search seems to become more consistent with a static process that activates items in memory as a function of their frequency. The weight that probes have in determining the activation of items in memory seems to be an informative signature of the impact of healthy aging and dementia on memory search. Our results show that, with healthy aging and dementia, the activation of items in memory is increasingly more determined by the frequency of past experience with those items.
\end{abstract}

Keywords: Search; semantic memory; modeling; aging.

\section*{Introduction}

Ronald Reagan became the oldest president elected in American history, when he took office at age 69, in 1981. He was diagnosed with Alzheimer's in 1993, the most common form of dementia, four years after he left office. Yet Reagan's signs of memory decline while in office - like forgetting names and being at a loss for words - have led to much speculation about how early dementia had set in. The question was whether his memory slips were a sign of normal aging or rather the early symptoms of dementia.

Studies using the animal fluency task ("name all the animals you can think of'; Thurstone, 1938) have shown that healthy older adults recall fewer items relative to younger adults within a limited time interval (e.g., Hills, Mata, Wilke, \& Samanez-Larkin, 2013; Kozora \& Cullum, 1995), in much the same way as older adults with dementia produce fewer items compared with healthy older adults (e.g., Beatty, Salmon, Testa, Hanisch, \& Troster, 2000; Epker, Lacritz, \& Cullum, 1999). In this paper, we examine how healthy aging and dementia impact search in semantic memory beyond the sheer reduction in the number of recalls. To this end, we formally model the recall sequences of 424 older adults aged between 69 to 103 years in the animal fluency task. We then examine individual differences in model fit and parameter estimates, as a way of identifying signatures of cognitive decline in memory search with healthy aging and dementia.

\section*{Static and Dynamic Search in Semantic Memory}

Memory retrieval can be viewed as the result of probing a memory representation with one or more probes to activate a response (e.g., Gronlund \& Shiffrin, 1986; Walker \&

Kintsch, 1985). We apply two classes of models based on prior work - static vs. dynamic - that make different assumptions about how retrieval probes are used to search memory in the fluency task (Hills et al., 2013). Consider the following two types of probes. One type of probe, the frequency probe, activates animal names in memory as a function of their frequency of past occurrence. A second type of probe, the similarity probe, activates each item in relation to its semantic similarity to the previously-recalled item. In a static model, search is guided by the same probe arrangement over the entire recall interval (i.e., by either probe alone or by a combination of the two). A dynamic model, on the contrary, switches between a frequency probe and a probe that combines frequency and similarity to traverse clusters of similar items in memory. When leaving a cluster, a dynamic model uses frequency alone to find a new cluster, and goes back to using a combination of frequency and similarity information as the new cluster is entered. Past work has found that, from early to late adulthood, search in memory is overall more consistent with a dynamic search model than with a static model that uses the same probe arrangement during the entire recall sequence (Hills et al., 2013).

In this paper, we examine the relative fit of static and dynamic models in old and very old age, for healthy individuals and individuals diagnosed with dementia. In very old age, do people use memory retrieval probes more in accordance with a dynamic model than with a static model as younger cohorts do? Moreover, we test alternative mechanisms of decline in memory search by investigating individual differences in the use of retrieval probes. We next turn to a brief discussion of the alternative mechanisms of decline in memory search.

\section*{Mechanisms of Decline in Memory Search}

Existing hypotheses proposed to account for age-related differences in the number of items produced in fluency tasks make different assumptions about how memory search declines with aging. The age invariance hypothesis proposes that aging is associated with unaffected semantic processing, and thus predicts no age differences in the use of retrieval probes (Mayr \& Kliegl, 2000). Two alternative hypotheses argue that the impact of aging affects the ability to switch between probes. The cluster-switching hypothesis views memory retrieval as a dynamic process involving a search for semantic categories like "pets", and a search for words within a category (e.g., "dog") (Troyer, Moscovitch, \& Winocur, 1997; Troyer, 2000). A common finding is that aging is associated with fewer total switches between categories, leading to the proposal that aging is associated
with reduced switching between retrieval probes (categories) (Troyer et al., 1997; Troyer, 2000). On the other hand, the cue-maintenance hypothesis (Hills et al., 2013) derives from studies showing that aging is associated with lower working memory capacity, defined as the ability to keep focus on one probe while ignoring distracting ones (e.g., Bopp \& Verhaghen, 2007). Age-related decline in working memory capacity should lead to a loss of probe focus, and therefore to increased switching between probes (e.g., Unsworth \& Engle, 2007). Existing evidence suggests that, from early to late adulthood, age is associated with an increase in switching between probes, per item recalled, in support of the cue-maintenance hypothesis of decline in memory search (Hills et al., 2013).

We examine which mechanism of decline best describes individual differences in switching in old (69-84 years) and very old age ( \(85-103\) years), between healthy individuals and individuals diagnosed with dementia. These mechanisms of decline, which have been proposed to account for age-related differences in fluency performance, can be used to test additional alternative hypotheses regarding memory decline in dementia. One hypothesis holds that memory impairment in dementia results from the acceleration of the same mechanism that leads to memory decline in healthy aging (e.g., Brayne \& Calloway, 1988; Huppert, 1994; Huppert \& Brayne, 1994). On this view, age-related differences in switching among individuals diagnosed with dementia should mirror age-related differences in healthy individuals. According to an alternative framework, however, memory decline in healthy individuals and in individuals with dementia is the product of distinct processes that target different brain systems (e.g., Albert, 1997; Gabrieli, 1996). This framework thus suggests that age-related differences in switching should arise from distinct decline mechanisms in healthy aging and dementia.

To summarize, we examine whether semantic search in healthy old age and dementia is more consistent with a static or with a dynamic model. Moreover, we test different mechanisms of decline in memory search by investigating individual differences in model fit and parameter estimates.

\section*{Methods}

\section*{Participants and Procedure}

The present work uses data from the Berlin Aging Study, a longitudinal study on aging (Baltes \& Mayer, 1999). Specifically, we analyze the animal fluency data that was collected in the first measurement occasion of the study, between 1990 and 1993. In the animal fluency task, participants were asked to respond verbally to the probe "Name all the animals you can think of" within 90 seconds, with their responses being tape-recorded. We retrieved participants' retrieval sequences from the tapes that were still functional, having compiled the responses of 424 individuals, with ages ranging from 69 to 103 (mean \(=\) 84.77, \(S D=8.58\) ). Of these 424 individuals, 91 were diagnosed with dementia (mean age \(=90.31, S D=6.53\) )
according to the guidelines of DSM-III-R, and 333 individuals (mean age \(=83.25, S D=8.45\) ) were considered to have normal intellectual functioning.

\section*{The Representation of Semantic Memory}

The first step towards formalizing search in semantic memory is to provide an explicit representation of the space being searched. We used the semantic representations of animals computed in prior work (Hills, Jones, \& Todd, 2012) using the BEAGLE semantic space model (Jones \& Mewhort, 2007). The BEAGLE model was trained on a subset of Wikipedia, composed of approximately 400 million word tokens and 3 million word types. Once the entire corpus has been learned (see Hills et al., 2012, for a description of the learning process), a word's memory representation is a vector pattern reflecting the word's history of co-occurrence with other words. Words that frequently co-occur end up developing similar vector patterns (e.g., bee-honey), as do words that commonly occur in similar contexts, even if they never directly co-occur (e.g., bee-wasp). Based on the representation learned by BEAGLE, we used the frequency of occurrence of each animal name in the Wikipedia corpus as well as the pairwise cosine similarities between animal names for our comparisons.

\section*{Alternative Models of Semantic Search}

To describe memory retrieval given this well-defined memory representation, we used a model framework similar to the item-level recall probability equation from the Search of Associative Memory model (SAM; Raaijmakers \& Shiffrin, 1981):
\[
\begin{equation*}
P\left(I_{i} \mid Q_{1}, Q_{2}, \ldots, Q_{M}\right)=\frac{\prod_{k=1}^{M} S\left(Q_{k}, I_{i}\right)^{\beta_{k}}}{\sum_{j=1}^{N} \prod_{k=1}^{M} S\left(Q_{k}, I_{j}\right)^{\beta_{k}}} \tag{1}
\end{equation*}
\]
where \(S\left(Q_{k}, I_{i}\right)\) represents the retrieval strength from probe \(Q_{k}\) to item \(I_{i}\) in memory, and \(w_{k}\) represents the saliency or attention directed at the \(k^{\text {th }}\) probe. The probability of retrieving a given item, \(I_{i}\), is given by the ratio of the activation strength of that item and the sum of the activation of all other items in memory given those same probes. Finally, \(\beta\) is a free parameter that indicates how strongly the person's recall was determined by the probe; higher values of \(\beta\) lead items with higher retrieval strengths for a given probe, \(Q_{k}\), to gain a larger share of the recall probability, while lower values of \(\beta\) distribute the probability of recall more evenly over all items.

We considered the frequency probe and the similarity probe introduced earlier in the paper. The frequency probe activates each item in memory as a function of the frequency of occurrence of each animal name in the Wikipedia corpus. The similarity probe activates each item in memory in relation to its semantic similarity to the
previous item recalled. Thus, the most recently recalled item is the probe used to query memory, and activation is defined as the pairwise semantic similarities produced by BEAGLE with all animals yet to be recalled. Given a particular probe arrangement, we can compute the predicted retrieval probability for any sequence of animal names by repeatedly using Equation 1. The \(\beta\) parameters were fit to each participant's data to maximize the observed recall probabilities and produce a maximum likelihood fit.

We tested four models differing in the nature of probe use. All models share the assumption that the probability that an item is the first item recalled is a function of its frequency. From the second recall onwards, the models differ in whether they use frequency and similarity information in a static or dynamic way. Static models use the same probe arrangement over the recall interval. The static frequency model uses a single probe: frequency. This assumes that individuals' recall sequences of animals reproduce their natural strength of activation in memory as a consequence of frequency alone. The static similarity model also uses a single probe: semantic similarity. This assumes that individuals rely only on the previously recalled item as a probe for the next recall, producing a chain of pairwise associated animals. The static combined model represents the simultaneous combination of frequency and semantic similarity. This assumes a process based on semantic similarity to the previous item that is further informed by the frequency of past experience with those items.

The dynamic model switches between a frequency probe and a probe that combines frequency and similarity to traverse clusters of similar items in memory. When leaving a cluster, a dynamic model uses frequency alone to find a new cluster, and goes back to using a simultaneous combination of frequency and similarity information as the new cluster is entered. Transitions are predicted by the model only after they occur, meaning that the model tests where the most plausible locations for transitions are, given the underlying representation. The model switches between retrieval probes wherever a sequence of items \(\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}\) have semantic similarities that follow the pattern \(S(A, B)>\) \(S(B, C)\), and \(S(B, C)<S(C, D)\). That is, similarity drops between clusters and then increases again once search resumes with a new cluster (e.g., the sequence DOG, CAT, SHARK, WHALE would have two clusters, divided by a similarity drop between CAT and SHARK).

In our data, healthy aging and dementia are associated with an increase in the proportion of items repeated. For this reason, and in contrast to previous studies (Hills et al., 2012, 2013; Hills \& Pachur, 2012), we did not exclude repetitions from participants' retrieval sequences, nor did we remove items from the memory representation after they were recalled. Although the current models do not distinguish new responses from repeated ones when calculating the retrieval probabilities, we are currently developing a generalized version of the models that takes into account how likely participants are to repeat previous responses.

\section*{Results and Discussion}

Figure 1 shows the mean number of correct responses produced in old and very old age, for healthy individuals and individuals diagnosed with dementia, calculated after excluding repeated items. Throughout our analyses, the group "old age" includes participants with ages between 69 and 84 years, and the group "very old age" includes participants with ages between 85 and 103 years. As expected, age was associated with recalling fewer items \((\mathrm{t}(422)=-11.32, \mathrm{p}<.001, \mathrm{r}=-.48)\). The mean number of correct responses produced decreased with age, both for healthy individuals and individuals with dementia. In addition, individuals with dementia produced fewer correct responses relative to individuals without dementia \((\mathrm{t}(422)=\) \(-12.03, \mathrm{p}<.001, \mathrm{r}_{\mathrm{pb}}=-.51\) ). These results indicate that both healthy aging and dementia are associated with a decline in the number of items retrieved from memory. In what follows, we present our results for the modeling of semantic retrieval in old and very old age.


Figure 1: Mean number of correct responses produced in each group. Error bars represent the standard error of the mean.

\section*{Do people switch between retrieval probes in old and very old age to navigate their semantic memory?}

Table 1 presents the Bayesian Information Criterion (BIC) of the four models. BIC is a commonly used measure to compare the fit of different models while penalizing them for the total number of free parameters that they have, as a way of reducing overfitting (Lewandowsky \& Farrell, 2011). Whereas the static single-probe models have only one free parameter, the models that use both frequency and similarity have two free parameters, each indicating how strongly the person's recall was determined by each type of information. Note that smaller values of BIC indicate a better model fit. Also, due to differences in the number of items recalled, the BIC values for the different models are only informative if compared within groups.

Table 1: Median Bayesian Information Criterion (BIC) of static and dynamic models per group.
\begin{tabular}{ccccc}
\hline \multirow{2}{*}{ Models } & \multicolumn{2}{c}{ Without dementia } & \multicolumn{2}{c}{ With Dementia } \\
\cline { 2 - 5 } & \begin{tabular}{c} 
Old \\
\((69-84)\)
\end{tabular} & \begin{tabular}{l} 
Very old \\
\((85-103)\)
\end{tabular} & \begin{tabular}{c} 
Old \\
\((69-84)\)
\end{tabular} & \begin{tabular}{c} 
Very old \\
\((85-103)\)
\end{tabular} \\
\hline Static & & & & \\
Frequency & 293.34 & 223.51 & 163.27 & 149.46 \\
& \((95.48)\) & \((92.39)\) & \((56.7)\) & \((79.75)\) \\
Similarity & 321.69 & 262.5 & 180.73 & 182.68 \\
& \((106.49)\) & \((101.99)\) & \((61.94)\) & \((95.46)\) \\
Combined & 285.72 & 226.60 & 164.07 & 147.49 \\
& \((95.83)\) & \((92.30)\) & \((54.19)\) & \((79.62)\) \\
Dynamic & 278.08 & 226.60 & 162.19 & 146.46 \\
& \((95.69)\) & \((92.33)\) & \((54.68)\) & \((79.68)\) \\
\hline
\end{tabular}

Note. Standard deviations are shown in parentheses.
The static, frequency model fit the data of all four groups better relative to the static, similarity model, suggesting that the best single predictor of recall was frequency rather than similarity. The pattern of results is, however, mixed across groups with respect to the fit of the models that use both frequency and similarity. The recall sequences of healthy individuals aged between 69-84 years were better fit by the static, combined model than by the two static single-probe models. Moreover, the model that incorporates dynamic transitions between probe arrangements outperformed the static combined model, being therefore the best fitting model for healthy individuals aged between 69-84 years. This finding is in line with past work showing that younger cohorts search memory according to a dynamic process that switches between a frequency probe and a probe that integrates frequency and similarity (Hills et al., 2013).

For the other groups of participants, however, the results show smaller differences in BIC between the static model that relies exclusively on frequency information and more complex models that use both frequency and similarity in a static or dynamic fashion. This suggests that the static frequency model may give a comparatively better account of memory search in very old age and dementia than for the healthy younger cohort. Yet the smaller BIC differences between models indicate that it is difficult to distinguish between them in very old age and dementia, thus calling for other methods to address the model selection problem.

\section*{How do healthy aging and dementia impact memory search?}

The number of switches per item was essentially unrelated to the total number of items recalled \((\mathrm{t}(422)=-1.21, \mathrm{p}=.22\), \(\mathrm{r}=-.06\) ). Additionally, it was also not related with age \((\mathrm{t}(422)=-1.05, \mathrm{p}=.29, \mathrm{r}=-.05)\), or with the presence of dementia \(\left(\mathrm{t}(422)=.87, \mathrm{p}=.38, \mathrm{r}_{\mathrm{pb}}=.04\right)\). Contrary to the cluster-switching and the cue-maintenance hypotheses, both of which posit specific changes in switching with increased age, these results seem to suggest that there are no differences in the nature of probe utilization with increased age, in support of the age invariance hypothesis. This result
is not consistent with the age-related increase in switching found in previous work for a younger cohort (Hills et al., 2013), suggesting that different mechanisms of decline may be at play in adulthood and later in life. Additionally, the finding that dementia was, as for healthy aging, unrelated with switching suggests that the decline of memory search in dementia may result from the acceleration of the same mechanism that leads to decline in healthy aging.

We believe, however, that there is an alternative, more sensible interpretation of these findings. As seen above, as people age, a static model appears to be better supported relative to a dynamic model. Thus, the number of switches per item recalled may not be an appropriate signature of the impact of very old age and dementia on memory search. A more informative signature of the decline of memory search in old age may be given by the free parameter, \(\beta\), which provides a measure of the deterministic nature of the activation given a specific retrieval probe. Different cohorts of healthy older adults and older adults with dementia may search memory in different ways, and these may influence the estimates of the \(\beta\) parameter. Higher values of \(\beta\) for the frequency probe lead very frequent items to have a larger share of the recall probability. Likewise, higher values of \(\beta\) for the similarity probe give a larger share of the recall probability to items that are very similar to the previouslyrecalled item. Lower values of \(\beta\) distribute the recall probabilities more evenly over all items in memory. Individual differences in the estimates of the \(\beta\) parameter may thus suggest alternative mechanisms of decline of memory search, whereby memory probes are given different weights in determining the recall probabilities.

Figure 2 plots the mean estimates per group for the \(\beta\) parameters corresponding to the frequency probe (panel A) and the similarity probe (panel B) in the static, combined model. Note that the parameter estimates are not comparable between probes due to the different scales of the Wikipediadefined frequencies and semantic similarities.


Figure 2: Mean estimates per group for the \(\beta\) parameters corresponding to the frequency probe (A) and similarity probe (B) in the combined static model. Error bars represent the standard error of the mean.

Figure 2 shows that, for individuals with normal intellectual functioning, there was an age-related increase in the estimates for the frequency probe, and a decrease in the estimates for the similarity probe. This indicates that memory search is more strongly determined by item frequency in very old age, but the weight of semantic similarity seems to decrease. For individuals diagnosed with dementia, the results demonstrate that there are no age differences in the estimates for either retrieval probe. However, the results suggest an association, independent of age, between dementia and the increasing weight of item frequency in determining the probability of recall.

This increase in the saliency of the frequency probe may be related to the observed increase in the proportion of repeated items with age \((\mathrm{t}(422)=6.07, \mathrm{p}<.001, \mathrm{r}=.28)\) and dementia \(\left(t(422)=8.04, p<.001, r_{p b}=.36\right)\). Figure 3 shows the mean of the log-transformed Wikipedia-defined frequencies for newly occurring items and repeated items produced as a function of age (panel A) and dementia diagnosis (panel B).


B


Figure 3: Mean Wikipedia-defined frequencies for new and repeated items produced as a function of age (A) and dementia diagnosis (B). Error bars represent the standard error of the mean.

In both age and dementia groups, repeated items had overall higher frequencies when compared with items recalled for the first time. Moreover, both age and dementia were associated with an increase in the Wikipedia-defined frequencies of the items repeated and, especially, of newly occurring items. Further modeling efforts are required to explore the contribution of repetitions to the higher saliency of the frequency probe in very old age and dementia.

\section*{Conclusion}

Our results suggest that, in the absence of dementia, memory search in early old age is consistent with a dynamic process that switches between a frequency probe and a probe that integrates frequency and similarity to traverse clusters of items grouped in memory by semantic similarity. This finding is in line with past work showing that younger cohorts search memory according to a dynamic process (Hills et al., 2013). However, in very old age and dementia, memory search processes appear to become more static, relying more on frequency to probe memory.

Our results further show that the proportion of switches between probe arrangements is unrelated with age and with the presence of dementia for older individuals. This result is in contrast with findings from previous studies showing that younger cohorts switch more often between probes with increasing age (Hills et al. 2013), thus suggesting that different mechanisms of decline may be at play in adulthood and later in life. Yet the saliency of memory retrieval probes may be a more informative signature of the impact of very old age and dementia on memory search. We have shown that, with healthy aging and dementia, the activation of items in memory is increasingly determined by the frequency of past experiences with those items. This result is consistent with the finding above that, in very old age and dementia, memory search appears to become more consistent with a static process that uses frequency to probe memory. Finally, the increase in the saliency of the frequency probe seems to be related with the increase in the number of repetitions. While age is associated with an increase in the number of repetitions, the items people repeat have a higher frequency of past occurrence compared with items recalled for the first time.

\section*{References}

Albert, M. S. (1997). The ageing brain: normal and abnormal memory. Philosophical Transactions of the Royal Society B, 352, 1703-1709.
Baltes, P., \& Mayer, K. (1999). The Berlin Aging Study: Aging from 70 to 100. New York: Cambridge University Press.
Beatty, B. B., Salmon, D. P., Testa, J. A., Hanisch, C., \& Troster, A. I. (2000). Monitoring the changing status of semantic memory in Alzheimer's disease: An evaluation of several process measures. Neuropsychology, 7, 94-111.
Bopp, K. L., \& Verhaeghen, P. (2007). Age-related differences in executive control processes in verbal and visuo-spatial working memory: Storage, transformation, supervision, and coordination. Journals of Gerontology: Psychological Sciences, 62, 239-246.
Brayne, C., \& Calloway, P. (1988). Normal ageing, impaired cognitive function, senile dementia of the Alzheimer's type: a continuum? Lancet, 1, 1265-1267.
Epker, M. O., Lacritz, L. H., \& Cullum, C. M. (1999). Comparative analysis of qualitative verbal fluency performance in normal elderly and demented populations. Journal of Clinical and Experimental Neuropsychology, 21, 425-434.
Gabrieli, J.D.E. (1996). Memory systems analyses of mnemonic disorders in aging and age-related diseases. Proceedings of the National Academy of Sciences USA, 93, 13534-13540.
Gronlund, S. D., \& Shiffrin, R. M. (1986). Retrieval strategies in recall of natural categories and categorized lists. Journal of Experimental Psychology: Learning, Memory, and Cognition, 12, 550-561.

Hills, T., Jones, M. \& Todd, P. M. (2012). Optimal foraging in semantic memory. Psychological Review, 119, 431440.

Hills, T., Mata, R., Wilke, A., \& Samanez-Larkin, G. R. (2013). Mechanisms of age-related decline in memory search across the adult life span. Developmental Psychology.
Hills, T., \& Pachur, T. (2012). Dynamic search and working memory in social recall. Journal of Experimental Psychology: Learning, Memory, and Cognition, 38, 218228.

Huppert, F. A. (1994). Memory function in dementia and normal aging-dimension or dichotomy? In F. A. Huppert, C. Brayne, \& D. B. O'Connor (Eds.), Dementia and Normal Aging (pp. 291-330). Cambridge: Cambridge University Press.
Huppert, F. A., \& Brayne, C. (1994). What is the relationship between dementia and normal aging? In F. A. Huppert, C. Brayne, \& D. B. O’Connor (Eds.), Dementia and Normal Aging (pp. 3-11). Cambridge: Cambridge University Press.
Jones, M. N., \& Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. Psychological Review, 104, 1-37.
Kozora, E., \& Cullum, C. M. (1995). Generative naming in normal aging: Total output and qualitative changes using phonemic and semantic constraints. The Clinical Neuropsychologist, 9, 313-325.
Lewandowsky, S., \& Farell, S. (2010). Computational Modeling in Cognition: Principles and Practice. London: Sage Publications.
Mayr, U., \& Kliegl, R. (2000). Complex semantic processing in old age: does it stay or does it go? Psychology and Aging, 15, 29-34.
Raaijmakers, J. G. B., \& Shiffrin, R. M. (1981). Search of associative memory. Psychological Review, 88, 93-134.
Thurstone, L. L., (1938). Primary mental abilities. University of Chicago Press, Chicago.
Troyer, A. (2000). Normative data for clustering and switching on verbal fluency tasks. Journal of Clinical and Experimental Neuropsychology, 22, 370-378.
Troyer, A.K., Moscovitch, M., \& Winocur, G. (1997). Clustering and switching as two components of verbal fluency: Evidence from younger and older healthy adults. Neuropsychology, 11, 138-146.
Unsworth, N., \& Engle, R. B. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. Psychological Review, 114, 104-132.
Walker, B. H., \& Kintsch, B. (1985). Automatic and strategic aspects of knowledge retrieval. Cognitive Science, 9, 261-283.

\title{
Do Friendship Influence Space Perception? With Particular Reference to the Curse of the Suspicious Participants
}

\author{
Nicolas Morgado (nicolas.morgado@upmf-grenoble.fr) \\ Laboratoire de Psychologie et Neurocognition (CNRS), Université Pierre-Mendès-France, 38040 Grenoble, France \\ Dominique Muller (dominique.muller@upmf-grenoble.fr) \\ Laboratoire Interuniversitaire de Psychologie, Institut Universitaire de France et Université Pierre-Mendès-France, 38040 Grenoble, France \\ Mathieu Pinelli (mathieu.pinelli@upmf-grenoble.fr) \\ Laboratoire de Psychologie et Neurocognition (CNRS), Université Pierre-Mendès-France, 38040 Grenoble, France \\ Éric Guinet (eric.guinet@upmf-grenoble.fr) \\ Laboratoire de Psychologie et Neurocognition (CNRS), Université Pierre-Mendès-France, 38040 Grenoble, France \\ Édouard Gentaz (edouard.gentaz@unige.ch) \\ Université de Genève / FAPSE UNI MAIL ; 40, Boulevard du Pont-d'Arve 1211 Genève 4 \\ Richard Palluel-Germain (richard.palluel@upmf-grenoble.fr)
Laboratoire de Psychologie et Neurocognition (CNRS), Université Pierre-Mendès-France, 38040 Grenoble, France
}

\begin{abstract}
In this study, we tested the hypothesis that social relationships affect the perception of distance. When participants imagined passing through a wall and a disliked-person, they perceived shorter aperture widths than when they intended to pass between a wall and a liked-person. This result was observed only for passable apertures suggesting that social constraints may influence visual perception only when people can actually perform this action. We discuss the results according to an embodied approach to visual perception but also with an alternative explanation in terms of possible demand characteristics. We also discuss some methodological points supposed to improve the validity of such experiments.
\end{abstract}

Keywords: Space Perception; Embodiement, Psychosocial Resources; Affective Closeness; Demand Characteristics

\section*{Introduction}

According to Proffitt and Linkenauger (2013) the visual perception of space depends on the phenotype of the perceiver. More precisely, the optical information would be scaled on the morphological, physiological, and behavioral properties of the body. For instance, decreasing people's ability to reach an object leads them to perceive it as being farther away (e.g., Lourenco \& Longo, 2009; Morgado, Gentaz, Guinet, Osiurak, \& Palluel-Germain, in press).
Previous works tried to extend this account to the influence of social factors on visual perception (Chambon, 2009; Harber, Yeung, \& Iacovelli, 2011; Morgado, Muller, Gentaz, \& Palluel-Germain, 2011). For example, Schnall, Harber, Stefanucci, \& Proffitt (2008) observed that people underestimate the slant of a steep hill when they are accompanied by a friend instead of being alone. According
to the authors, this difference in slant estimation reflects that social support, as a social resource, can compensate the potential effort associated with climbing the hill and thus reduces its perceived steepness.
In some cases, however, the social constraints associated with a given action constitute a cost rather than a resource. Previous works suggest that people maintain a personal space around them and that they feel discomfort when someone invades this space (Hayduk, 1983). Moreover, this discomfort seems to increase as the physical interpersonal distance decrease (Hayduk, 1981). Interestingly, the discomfort associated with personal space invasion seems to vary according to the social relationship (Sundstrom \& Altman, 1976). Consistent with these findings, we recently observed that people's action-scaled perception of a space between two acquaintances is correlated with the participants' affective closeness toward these acquaintances (Morgado et al., 2011). Indeed, the closer participants felt to their classmates, the more passable the space between the classmates pictures appeared and the less space they needed to pass. These results might suggest that participants perceived the space between the two classmate pictures (i.e., the aperture width) differently because of the closeness feeling.
In the present study, we aimed to investigate further whether social relationships influence the visual perception of an aperture between a wall and an acquaintance. More precisely, participants had to estimate the width of an aperture between the picture of a wall and that of a human figure evoking a liked person or a disliked person. Participants also indicated if the aperture was wide enough to allow them to pass. Our hypothesis was that the participants from the disliked-person group should perceive smaller apertures than participants from the liked-person
group. Moreover, this study aimed to replicate the observed correlation between affective closeness and the passability judgments.

\section*{Method}

\section*{Participants}

Sixty undergraduates ( 52 females; \(M_{\text {age }}=21, S D_{\text {age }}=3\) ) from the University of Grenoble took part in this experiment for course credit. The participants had normal or corrected-to-normal vision, as indicated by self-report. None had participated in our previous study. The present study was conducted in accordance with the Declaration of Helsinki and with the understanding and the written consent of each participant. It was approved by the local ethics committee of the LPNC (CNRS and the University of Grenoble).

\section*{Apparatus and procedure}

To manipulate social relationships, we chose to use a similar mental imagery task as the one used by Schnall et al. (2008). Participants sat down in front of a computer for the mental imagery task. Headphones provided the instructions to the participants who were randomly assigned to the disliked-person or the liked-person group (respectively, \(n=\) 31 and \(n=29\) ). Using headphones enabled the experimenter to be blind to experimental groups while increasing standardization of the instructions.

The instructions indicated that the experiment concerned visual perception of space and that participants would have to estimate the width of an aperture between a picture of a wall and a human figure. Instructions underlined that recent studies indicated that such a task is too difficult in artificial situations. Supposedly to make the task more natural, they had to imagine that the human figure was an acquaintance. At the beginning of the mental imagery task, participants had to complete a relaxation exercise. Then, participants in the disliked-person group had to choose an acquaintance who they did not like at all and who made them uncomfortable. In contrast, participants in the liked-person group had to choose an acquaintance that they liked very much and who made them feel good. Participants could take all the time they needed to choose this acquaintance and they pressed a key to hear the next instructions. Then, they had to imagine the presence of this acquaintance while thinking about their feeling toward this person, while visualizing his or her physical appearance, and while keeping in mind how they usually interact with this person. At the end of this mental imagery task, the instructions indicated that participants had now to estimate aperture width and they had to keep in mind a picture of the chosen acquaintance.
For the perceptual task, participants stood at 3.7 m in front of a white screen on which the picture of a wall and those of a human figure were projected (Figure1). The
dimensions of the two pictures were identical (height: 169 cm , width: 41.5 cm ). The instructions were projected on the screen at the beginning of this task. Throughout this task, participants had to imagine the previously chosen acquaintance in place of the human figure projected on the screen. Since the constraints of a given action influence the perception mainly when people intend to perform this action (e.g., Witt, Proffitt, \& Epstein, 2005), participants had to imagine passing through the aperture between the wall and their acquaintance before each width estimation. Since arm posture seems to influence perceived aperture widths (Stefanucci \& Geuss, 2009), participants had to keep their arms along their body. To estimate the aperture widths, participants completed a visual-matching task (for a similar measure see Stefanucci \& Geuss, 2009). The experimenter stood at 190 cm from the participants' right side and progressively unrolled a tape measure located at 130 cm from the floor. Participants had to stop the experimenter when they considered that the length of the tape measure was equal to the aperture width. To reduce the potential experimenter effect on participants' estimations, the experimenter could not see which aperture width the participants had to estimate. Moreover, the experimenter tried hard to keep his gaze on a fixed point in the wall in front of him while unrolling the tape measure. Neither the experimenter, nor the participants could see the graduation of the tape measure during the estimations. The experimenter could only see the measure after participants were satisfied of their estimation to record it in the computer. Then, participants made a "yes" or "no" passability judgment (Warren \& Whang, 1987) to indicate if the aperture was wide enough to allow them to pass through it without rotating their shoulders. The experimenter recorded this judgment and launched the next trial. Participants completed 32 trials including 4 practice trials and 28 test trials. The actual aperture widths used for the test trials ranged from 30 cm to 95 cm with a \(5-\mathrm{cm}\) step. The actual aperture widths used for the practice trials ( \(31 \mathrm{~cm}, 39\) \(\mathrm{cm}, 52 \mathrm{~cm}, 82 \mathrm{~cm}\) ) were randomly selected among this range of width and were the same for all the participants. The actual aperture widths were randomly presented during the practice and test trials.


Figure 1: Experimental setup and device (P: participant; E: experimenter).

Immediately after the completion of the perceptual task, the experimenter asked participants if the overall procedure was clear and probed them for suspicion about the hypothesis. The experimenter asked two questions to the participants: (1) "In your opinion what hypothesis is tested in this study?" (2) "Do you think that some aspects of the experiment could have influenced your responses? If so, what were these aspects?" Then the experimenter recorded participants' shoulder width as the distance between the tips of the two humerus. Finally, participants sat down and answered a post-experimental questionnaire projected on the screen. The items of this questionnaire were gathered together by themes which were presented in a fixed order: (1) impressions about the mental imagery task, (2) information about the chosen acquaintance, (3) participants' feelings toward the acquaintance, (4) participants' preferred interpersonal distance with the acquaintance (for a similar measure see Pedersen, 1973), (5) participants' physical state, and participants' mood. Items, however, were randomly presented among the themes.

\section*{Results}

We conducted a set of t-tests to check the effectiveness of our experimental manipulation with Social Relationship as a between-group factor and the different items of the postexperimental questionnaire dependent variables. The participants in the liked-person group indicated more positive feelings toward their acquaintance ( \(M=4.42, S D=\) .34) than those in the disliked-person group ( \(M=2.46, S D=\) .36), \(t(56)=21.32, p<.001, \eta^{2}=.89\). In line with the literature (Sundstrom \& Altman, 1976), participants in the liked-person group preferred keeping a significantly shorter interpersonal distance with the acquaintance ( \(M=30.02, S D\) \(=22.26\) ) than those of the disliked-person group ( \(M=\) \(141.83, S D=44.69), t(56)=-12.19, p<.001, \eta^{2}=.73\). Participants in the liked-person group indicated having more frequent contacts with the acquaintance ( \(M=3.17, S D=\) 1.05 ) than those of the disliked person group ( \(M=1.86, S D\) \(=.85), t(56)=5.19, p<.001, \eta^{2}=.73\). Moreover, participants in the liked-person group indicated that the pictures generated during the mental imagery task were more pleasant \((M=4.6, S D=.49)\) than those in the dislikedperson group \((M=2.11, S D=.59), t(56)=17.28, p<.001\), \(\eta^{2}=.84\). There was no other significant difference for the other items of the post-experimental questionnaire (i.e., duration of the relationship, mood, vividness of the imagery task, easiness to imagine the target person, and easiness to imagine passing through the aperture). It is noteworthy, however, that it was marginally easier to imagine the liked person \((M=3.23, S D=1.22)\) than the disliked one ( \(M=\) 2.64, \(S D=1.25\) ), \(t(56)=1.82, p=.07, \eta^{2}=.06\).

An inspection of the Studentized deleted residuals on the aperture width estimations revealed the presence of two outliers (see Judd, McClelland, \& Ryan, 2009). They were excluded of the subsequent analyses. Two other participants were also excluded because of a power cut during data collection. After these exclusions, it remained 56
participants \(\left(n_{\text {liked }}=29, n_{\text {disliked }}=27\right)\). We considered participants as suspicious when they indicated that they thought that we aimed to test the effect of social relationship on the perception of aperture or when they indicated that social relationship was an aspect that influenced their estimations. In spite of our cover story, \(39.29 \%\) of our participants suspected the true purpose of the study. Moreover, there were more suspicious participants in the disliked-person group ( \(55.56 \%\) ) than in the liked-person group \((24.14 \%), t(54)=2.49, p<.02, \eta^{2}=.10\).


Figure 2. Perceived distance as a function of Actual Aperture Width and Social Relationship. Error bars denote standard errors of the means.

We conducted an analysis of variance (ANOVA) with Social Relationship (liked person, disliked person) as a between-subjects factor and Actual Aperture Width ( 30 cm , \(35 \mathrm{~cm} . . .90 \mathrm{~cm}, 95 \mathrm{~cm}\) ) as a within-subject factor. The Estimated Aperture Width was the dependent variable. Given that the exclusion of all the suspicious participants would lead to decrease dramatically the statistical power of the analysis, we entered Suspicion (suspicion vs. no suspicion) as a covariate in this analysis. We also entered Shoulder Width as a covariate since this variable is known to influence perceived aperture widths. This analysis revealed that participants in the disliked-person group estimated shorter aperture widths ( \(M=58.5, S D=1.35\) ) than those of the liked-person group ( \(M=61.5, S D=1.51\) ). However, this main effect of social relationship was not statistically significant, \(F(1,51)=2.21, p<.14, \eta^{2}=.04\). Neither the main effect of suspicion, nor those of shoulder width were significant ( \(p s>.1\) ). The main effect of Actual Aperture width was significant, \(F(13,663)=7.31, p<.001\), \(\eta^{2}=.13\). Interestingly, the interaction between actual aperture width and social relationship was significant, \(F(13\), 663 ) \(=2, p<.02, \eta^{2}=.04\) (see Figure 2). This seems to reflect the fact that participants in the disliked-person group tended to estimate shorter aperture widths than those of the liked-person group for the aperture judged wide enough to pass, \(F(1,51)=3.08, p<.09, \eta^{2}=.06\), but not for those
judged too small to pass, \(F(1,51)=.68, p<.41, \eta^{2}=.01\). Importantly, the interaction between the actual aperture width and the social relationship did not depend on suspicion ( \(p=.73\) ). Moreover, these results did not change dramatically when we controlled for the easiness to imagine the target person. We also conducted an ANOVA with social relationship as a between-subject factor and the percentage of "yes" passability judgments as a dependent variable. We also entered suspicion and shoulder width in this analysis to statistically control for these variables. Although the percentage of "yes" passability judgments was smaller for the disliked-person group ( \(M=53.32, S D=\) 4.33) than for the liked-person group ( \(M=60.36, S D=\) 4.89), this difference was not significant ( \(p>.74\) ).

Neither the correlation between the familiarity with the acquaintance and the percentage of "yes" passability judgments, nor those between the preferred interpersonal distance and the percentage of "yes" passability judgments were significant ( \(r=-.08, p=.71\) and \(r=-.22, p=.28\) respectively). Interestingly, the correlation between the affective closeness and the percentage of "yes" passability judgments was significant for the participants in the disliked-person group ( \(r=.64, p=.01\) ), but not for those in the liked-person group ( \(r=-.42, p=.23\) ). Importantly, this pattern of correlations remained the same when we statistically controlled for the shoulder width of the participants and for the suspicion.

\section*{Discussion}

When participants intended to pass between a wall and a disliked-person stimulus, they tended to estimate shorter aperture widths compared with when they intended to pass between a wall and a liked-person stimulus, but only for passable apertures. As observed in our previous study (Morgado et al., 2011), we also observed a positive correlation between the affective closeness and percentage of "yes" passability judgments. More precisely, the closer participants felt to the acquaintance, the more passable the aperture appeared. Surprisingly, it was only true for the participants in the disliked-person group, but not for those in the liked-person group. At a first glance, these results seem consistent with the social extension of the phenotypic account of perception (Proffitt \& Linknauger, 2013). According to this account, the anticipation of personal space invasion might lead to perceive shorter aperture widths in the presence of disliked persons than in the presence of liked ones.
The observed interaction between the actual aperture width and the social relationship is consistent with previous results suggesting that the constraints related to an intended action influence visual perception only when people can actually perform this action (Lessard, Linkenauger, \& Proffitt, 2009). The correlation between affective closeness and passability judgments observed only with disliked persons might also suggest that affective closeness is more relevant for passability with disliked persons compared with liked ones.

One might be willing to explain our results in terms of the ease to keep in mind the person stimulus. For instance, it might be easier to imagine the disliked-person than the liked-person given the literature on attention to negative stimuli (e.g., Smith, Cacioppo, Larsen, \& Chartrand, 2003). If so, such a difference might explain our results. The data from our post-experimental questionnaire, however, indicated that the difference between the disliked-person and the liked-person groups for the vividness of the imagery task was not significant. In contrast, it was marginally easier to imagine the liked-person than the disliked one. Importantly, the interaction between the actual aperture width and the social relationship remained significant when we statistically controlled for the easiness to imagine the target person. In the same vein, one might also invoke mood as a potential confound in our results since mood seems to influence visual perception of space (e.g., Riener, Stefanucci, Proffitt, \& Clore, 2011). However, our postexperimental did not provide any support for this alternative explanation.
Durgin et al. (2009) underlined the necessity to take into account the suspicion of the participants in studies about the influence of the action capabilities on visual perception of space. According to their concerns, the large number of suspicious participants in our sample rises another possible explanation for our results in terms of demand characteristics. Demand characteristics refers to the cues which provide an experimental hypothesis to the participants (Orne, 1962). Moreover the large number of suspicious participants in the disliked-person group suggests that these participants were more likely to be affected by demand characteristics. Thus, they could have reduce their width estimations and adjust their passability judgments in line with their guess about our hypothesis. If it was the case, one could argue that the interaction effect between the actual aperture width and the social relationship should depend on whether participants were suspicious or not. Interestingly although the interaction between actual aperture width, social relationship, and suspicion was not significant, the increasing difference with the actual aperture in estimated aperture width between the disliked-person and the liked-person groups seems to be present for the suspicious participants only. Even if these results are only descriptive, it is important to underline the fact that our study was not primarily designed to test such a three-way interaction. Considering our sample size, a lack of statistical power needed to test such an interaction might explain this non-significant result. Another important limit relies on the fact we used very basics questions to probe the suspicion of the participants. Further studies primarily designed to test the relevance of the demand characteristics in perception studies will have to use a more sophisticated postexperimental questionnaire.
One could also argue that the experimental demand in the liked-person and the disliked-person group was the same since the two groups had to imagine the presence of an acquaintance. Yet, we observed more suspicion in the
disliked-person group than in the liked-person one, which means that demand cues are not equally spread into the two groups or at least that the participants' receptivity to these cues are different between the two group. One possible explanation of this asymmetry might rely on an inconsistency between the cover story and the disliked group. More precisely, participants could have found paradoxical to imagine the presence of a dislike person to make the task more natural. Such asymmetry has important implications for studies contrasting positive and negative experimental manipulations and researchers should be encouraged to find a way to rule out this potential confound.
In spite of the limits of our study, it highlights the need of using a systematic and standardized post-experimental questionnaire in perception studies. Indeed, we think that dealing with the demand characteristic explanation need more than just indicating that participants were probed for suspicion. For instance, it seems that participants tend to admit their suspicion more in a computerized postexperimental questionnaire than in a face-to-face interview with the experimenter (Blackhart, Brown, Clark, Pierce, \& Shell, 2012). Thus it is important that the perception researchers take into account such results when they probe their participants for suspicion. One could doubt of the use of questionnaire to deal with the demand characteristics for at least two reasons. The first reason is that if demand characteristics exert an implicit influence on the participants' behavior, the participants should not be aware of this influence. Thus the interest of simply asking people about this influence with a post-experimental questionnaire should be highly limited (e.g., Nisbett \& Wilson, 1977). However, the fact that much of the demand bias should be implicit is not guaranteed. Moreover, even if one considers demand bias as implicit, the demand characteristics which produce this bias can be perceived explicitly by the participants. Thus using a post-experimental-questionnaire remains useful to assess the receptivity of the participants to the demand characteristics. The second reason that can lead scholars to doubt the usefulness of the post-experimental questionnaire is the fact that such questionnaire captures the impression of the participants after the experiment. It is possible that some participants did not think very much about the hypothesis during the experiment and that the post-experimental questionnaire increases their suspicion when they answer to it. Horvat (1986) observed, however, that care in the design of the questionnaire and in the coding of the responses can improve the reporting of true suspicion and decrease the reporting of false suspicion.
The use of theoretical accounts of demand bias to improve post-experimental questionnaire and experimental design is particularly relevant (e.g., Allen, 2004; for a review see also Strohmetz, 2008). According to such accounts, to consider that there is a risk of demand bias, researchers have to consider three critical variables. The first variable is receptivity of the participants to the demand cues. The presence of such cues can lead the participants to guess the hypotheses. We can assess the receptivity of the participants
using a quasi-control group as proposed by Orne (1962) or with a post-experimental questionnaire. Interestingly, we can also reduce the receptivity of the participants to the critical cues by diverting their attention with deceptive cues. With such a "red herring technique", Laney et al. (2008) succeed in reducing the suspicion of the participants about their hypothesis. They used a traditional cover-story to hide the purpose of their study, but in addition they included perceptible cues suggesting that the study had another purpose (i.e., the red herring). Importantly, this red herring cannot be confounded with the true purpose of their studies so that any demand bias in favor of the red herring cannot lead the participants to confirm the true purpose.
The second variable is the participants' motivation to comply with the demand cues. Indeed, without such a motivation, the receptive participants have no reason to comply with demand cues. Allen (2004) in his postexperimental questionnaire used some items about the motivation of the participants to comply or not with what they thought was expected.
The third variable is participants' ability to voluntarily modify their responses according to the demand cues. The question of this ability is highly relevant in behavioral research and seems to be ignored by researchers working on the so called low-level processes. Such tendency might relied on a confusion between what it is studied (i.e., a lowlevel process) and the way by which we have access to this process (i.e., a response). Yet, even if visual perception implies low-level processes that some authors consider as cognitively impenetrable (e.g., Pylyshyn, 1999), the response of the participants might rely on a voluntary motor act. In that case, as in the cases of visual-matching estimate of or affordance judgments, participants might have the opportunity to voluntarily influence their responses. Assuming that any response used to study a low-level process is not sensitive to response bias is a strong claim and had to be examined for each response or at least for each category of response.
Finally, we observed mixed evidences supporting the idea that social relationships influence the visual perception of distance. We have, however, to qualify this conclusion according to the potential implication of a demand bias in our results. To conclude, if overgeneralizing the explanation in terms of demand bias to experiments with very different experimental design is flawed, ignoring the potential presence of a demand bias in an experiment is also an important concern.

\section*{Acknowledgments}

This research was supported by the "Agence National de la Recherche" (ANR) under Grant ConstrainPercept.

\section*{References}

Allen, C. T. (2004). A theory-based approach for improving demand artifact assessment in advertising experiments. Journal of Advertising, 33(2), 63-73.

Balcetis, E., \& Dunning, D. (2010). Wishful seeing: More desire objects are seen as closer. Psychological Science, 21(1), 147-152.
Blackhart, G. C., Brown, K. E., Clark, T., Pierce, D. L., \& Shell, K. (2012). Assessing the adequacy of postexperimental inquiries in deception research and the factors that promote participant honesty. Behavior research methods, 44(1), 24-40.
Chambon, M. (2009). Embodied perception with others' bodies in mind: Stereotype priming influence on the perception of spatial environment. Journal of Experimental Social Psychology, 45(1), 283-287.
Durgin, F. H., Baird, J. A., Greenburg, M., Russell, R., Shaughnessy, K., \& Waymouth, S. (2009). Who is being deceived? The experimental demands of wearing a backpack. Psychonomic Bulletin \& Review, 16(5), 964-969.
Harber, K. D., Yeung, D., \& Iacovelli, A. (2011). Psychosocial resources, threat, and the perception of distance and height: Support for the resources and perception model. Emotion, 11(5), 1080-1090.
Hayduk, L. A. (1981). The permeability of personal space. Canadian Journal of Behavioral Science, 13(3), 274-287.
Hayduk, L. A. (1983). Personal space: Where we now stand. Psychological Bulletin, 94(2), 293-335.
Judd, C. M., McClelland, G. H., \& Ryan, C. S. (2009). Data analysis: A model comparison approach. Routledge/Taylor \& Francis Group.
Laney, C., Kaasa, S. O., Morris, E. K., Berkowitz, S. R., Bernstein, D. M., \& Loftus, E. F. (2008). The Red Herring technique: A methodological response to the problem of demand characteristics. Psychological Research, 72(4), 362-375.
Lessard, D. A., Linkenauger, S. A., \& Proffitt, D. R. (2009). Look before you leap: Jumping ability affects distance perception. Perception, 38(12), 18631866.

Lourenco, S. F., \& Longo, M. R. (2009). The plasticity of near space: Evidence for contraction. Cognition, 112(3), 451-456.
Morgado, N., Gentaz, E., Guinet, E., Osiurak, F., \& PalluelGermain, R. (in press). Within reach but not so reachable: Obstacles matter in visual perception of distances. Psychonomic Bulletin \& Review.
Morgado, N., Muller, D., Gentaz, E., \& Palluel-Germain, R. (2011). Close to me? The influence of affective closeness on space perception. Perception, 40(7), 877-879.
Nisbett, R. E., \& Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. Psychological Review, 84(3), 231-259.
Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. American Psychologist; American Psychologist, 17(11), 776-783.

Pedersen, D. M. (1973). Development of a personal space measure. Psychological Reports, 32, 527-535.
Proffitt, D. R., \& Linkenauger, S. A. (2013). Perception viewed as a phenotypic expression. In W. Prinz, M. Beisert, \& A. Herwig (Eds.), Tutorials in Action Science. MIT Press.
Pylyshyn, Z. (1999). Is vision continuous with cognition?: The case for cognitive impenetrability of visual perception. Behavioral and brain sciences, 22(03), 341-365.
Schnall, S., Harber, K. D., Stefanucci, J. K., \& Proffitt, D. R. (2008). Social support and the perception of geographical slant. Journal of Experimental Social Psychology, 44(5), 1246-1255.
Smith, N. K., Cacioppo, J. T., Larsen, J. T., \& Chartrand, T. L. (2003). May I have your attention, please: electrocortical responses to positive and negative stimuli. Neuropsychologia, 41(2), 171-83. Retrieved
from http://www.ncbi.nlm.nih.gov/pubmed/12459215
Stefanucci, J. K., \& Geuss, M. N. (2009). Big people, little world: The body influences size perception. Perception, 38(12), 1782-1795.
Strohmetz, D. B. (2008). Research Artifacts and the Social Psychology of Psychological Experiments. Social and Personality Psychology Compass, 2(2), 861877.

Sundstrom, E., \& Altman, I. (1976). Interpersonal relationships and personal space: Research review and theoretical model. Human Ecology, 4(1), 4767.

Warren, W. H., \& Whang, S. (1987). Visual guidance of walking throught apertures: Body scaled information for affordances. Journal of Experimental Psychology: Human Perception and Performance, 13(3), 371-383.
Witt, J. K., Proffitt, D. R., \& Epstein, W. (2005). Tool use affects perceived distance, but only when you intend to use it. Journal of Experimental Psychology: Human Perception and Performance, 31(5), 880-888.

\title{
Joint Commitment: An Analysis of Emotions and Non-Verbal Behaviors
}

\author{
Francesca Morganti (francesca.morganti@unibg.it) \\ Dipartimento di Scienze Umane e Sociali, University of Bergamo, 2 Piazzale S. Agostino, Bergamo, 24129 Italy \\ Antonella Carassa (antonella.carassa@usi.ch) \\ Faculty of Communication Sciences, University of Lugano, 13 Via Giuseppe Buffi, Lugano, 6904 Switzerland \\ Marco Colombetti (marco.colombetti@usi.ch) \\ Faculty of Communication Sciences, University of Lugano, 13 Via Giuseppe Buffi, Lugano, 6904 Switzerland \\ DEIB, Politecnico di Milano, 32 Piazza Leonardo da Vinci, Milano, 20135 Italy \\ Paride Braibanti (paride.braibanti@unibg.it) \\ Dipartimento di Scienze Umane e Sociali, University of Bergamo, 2 Piazzale S. Agostino, Bergamo, 24129 Italy \\ Giorgio Rezzonico (giorgio.rezzonico@unimib.it) \\ Dipartimento di Scienze della Salute, Università di Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1 Milano, 20126 Italy \\ Ivana Sgro (ivana.sgro@studenti.unibg.it) \\ Dipartimento di Scienze Umane e Sociali, University of Bergamo, 2 Piazzale S. Agostino, Bergamo, 24129 Italy
}

\begin{abstract}
According to Margaret Gilbert, a joint commitment (JC) is a commitment of two or more agents, called the parties of the JC, to engage in a common project. Creating a JC often involves an explicit agreement, carried out in a conversational interaction through overt communication. We explored aspects of such interactions that can be considered as complementary to verbal exchanges, focusing on how a JC is managed by the parties by means of emotional and other non-verbal bodily expressions. We analyzed three phases of the JC lifecycle (creation, maintenance, and violation), and in particular the emotional reaction of the participants to two types of violations by the experimenter. In our analysis we used standardized tools such as the Ethological Coding System for Interviews, the Mind Reading Emotional Library, and the Facial Action Coding System. Our results show that certain non-verbal behaviors in the phase of JC creation are characteristic of the participants who later did not fulfill their commitment. Moreover, the participants' emotional reactions to JC violation by the experimenter turned out to depend on the type of violation. Finally, the creation and maintenance of JC, and the emotional reaction to its violation, appear to be independent of the participants' personality and empathic disposition.
\end{abstract}

\section*{Introduction}

The theoretical background of this paper is Gilbert's plural subject theory (Gilbert, 1989) together with some developments by Carassa, Colombetti \& Morganti (2008). Accord-
ing to Gilbert, all genuinely collective phenomena (like joint activities, collective beliefs, group feelings, etc.) involve a special kind of commitment, namely, a joint commitment (JC). An agent may be personally committed to do something as a result of an individual decision; in this case the agent is the only 'owner' of the commitment, and can rescind it as he or she pleases. Contrary to personal commitments, a joint commitment is a commitment of two or more agents, called the parties of the JC, to engage in a common project.

The main feature of JC is that it consists of normative relationships between the parties. If two parties are jointly committed to do something, then each one is obligated to the other one to do their part, and has the right that the other one do their part. Accordingly, a JC is violated when a party does not live up to their obligations.

Gilbert often remarks that making a JC does not necessarily require an explicit agreement: certain subjects may enter a joint commitment by starting to interact in certain ways, without ever trying to describe what they do together as a matter of agreement. Making explicit agreements, on the other hand, is very common in everyday life, and in the current paper we concentrate on this type of situation. According to Carassa \& Colombetti (2009), explicit agreements to joint projects are typically created in a conversational interaction through overt communication, where an agent perform speech acts with a communicative intention, in the

Gricean sense (Grice, 1957), as in the following example: "Can you help me with my homework assignment?" "Yes, sure". Beyond this plain example, it would be undoubtedly interesting to investigate how the making of agreements naturally unfolds in conversations, but our present research is aimed at exploring other complementary aspects of an ongoing face-to-face social interaction of this type. In fact, we want to take into account the role played by intersubjectivity, understood as the broad range of processes and capacities that, according to the enactive approach, allow one to directly perceive, understand and respond to the psychological states of others, without explicitly representing and reasoning upon them (Morganti, Carassa \& Riva, 2008). We believe that such pre-reflective processes need to be investigated to get a comprehensive view of JC creation, maintenance and fulfillment/violation. This means that beyond studying the situated use of language in interaction, we also have to focus on non-verbal behaviors such as facial expressions, gestures, and bodily postures and movements that can be pre-reflectively produced and understood by the interactants. As an example, an interactant can be aware that an agreement has been made verbally, while at the same time perceiving that the other party is not really willing to live up to it. These kinds of behaviors we expect to shape the normative landscape developed by the interactants along emotional and tacit dimensions. In the present research we did not yet analyze, as an enactive approach would require, the participant-experimenter interaction, but we specifically focused on the participant non-verbal behaviors, considering them as components of intersubjective patterns to be further investigated.

The experiment reported in this paper addresses all phases of JC lifecycle: its creation, its maintenance by a participant, and a participant's reaction to a violation performed by the experimenter in two different conditions. To understand if there is a personal predisposition to the acceptance and maintenance of a JC, we made an assessment of personality and of empathic disposition of the participants. This evaluation was justified by the hypothesis that personal predispositions could influence both the acceptance, maintenance and fulfillment/violation of a JC (this hypothesis, however, was not confirmed by our results).

\section*{Materials and methods}

\section*{Subjects}

The experiment involved 35 female participants, all of them students at the University of Bergamo, aged 19-27 ( \(\mathrm{M}=\) 21.11, sd = 1.9).

The sampling was conducted partially at random and the recruitment was voluntary. The experimental phase lasted three weeks between September and November 2011; data analysis was then conducted in a unique solution.

\section*{Materials}

Personality and empathy assessment To analyze the participants' personality, their emphatic disposition and the potential connections with JC creation, maintenance and violation, some questionnaires were distributed. For personality assessment, we used
a) the Eysenck Personality Inventory - EPI (Eysenck,1985), composed by three sub-scales (extroversion/introversion, neurosis, psychosis); and
b) the Mini Questionnaire on Personality Organizations MQOP (Nardi et al., 2012), composed by four subscales (contextualized, normative, controlling, detached).
The latter questionnaire allows one to study personality as a process, by focusing on the relationships between personality and developmental process axes, based on Guidano's theory of Personal Meaning Organization (1987).

Concerning the evaluation of emphatic disposition, two tools have been used:
a) the Interpersonal Reactivity Index - IRI (Davis, 1980, 1983), according to which empathy results from the integration of four factors (fantasy-empathy, perspective taking, empathic concern, personal distress);
b) the Emotional Regulation Questionnaire - ERQ (Gross \& John, 2003), aimed at identifying the strategy of emotional regulation used by the subjects (cognitive reappraisal, expressive suppression).

Joint commitment lifecycle To evaluate the lifecycle of JCs, we used four parallel forms of self-evaluation diaries, based on the structured diary proposed by Oatley and Laroque (1994). The purpose of such diaries was to analyze different possible JC situations, in which the participants could find themselves during the experiment. The diaries proposed to the participants followed the Experiencing Sample Method of Larson and Csikszentmihalyi (1983), which requires participants to describe their experience at certain moments of time.
Non-verbal behaviors We used the coding system proposed by Troisi (1999) to analyze the non-verbal behavior of the participants during the face-to-face meetings in which the JC was created and managed. This method is known as the Ethological Coding System for Interviews ECSI, and includes 37 behavioral patterns, most of them regarding facial expressions and hand movements.

To analyze the participants' reaction to the JC violation performed in the second face-to-face meeting, we used the Facial Action Coding System (FACS) of Ekman and Friesen (Ekman, 1978). We also used the Mind Reading Emotional Library (Baron-Cohen, 2003, 2004) for the classification of the expressions of emotional reaction to JC violation.

\section*{Procedure}

Firstly the questionnaires about personality and empathic disposition were filled in by the participants. This selfassessment, carried out in a quiet room, required about 30 minutes for each participant. Then other three sessions followed:
1. an initial face-to-face meeting between the experimenter and the participant, in which the experimenter purported to describe the goal of the research (see below), the participant's willingness to take part in the research was ascertained, and a JC of the participant and the experimenter was made through an explicit agreement;
2. a one-week period in which the agreed activity was performed and monitored;
3. a second face-to-face meeting in which the JC was violated by the experimenter.
Both meetings were video-recorded by three hidden video cameras, one focused on the experimenter's face, the second on the participant's face, and the third on the body and face of both of them. During the first meeting, all the participants verbally expressed their willingness to participate in the research, and the second meeting was scheduled.
With every participant, the first meeting was divided into three sections:
1. The first section concerned the description of the research. The experimenter told the participant that the research was aimed to analyze the everyday life emotions experienced while performing joint activities with others. In details, she asked the participant to answer some diaries. Every day of the following week, each participant would receive by email 5 diaries per day (for a total of 35 diaries). Furthermore, the experimenter informed the participants that they had to hold a second meeting after one week, in which they would receive comments on their diaries, previously analyzed by the experimenter.
2. The second section consisted in the explicit request to take up the JC, performed by the question "So, do we agree?"
3. The third section consisted in the participant's acceptance of the JC (in fact, no participant refused to take up the JC).
In the week between the first and the second meeting, the 35 diaries were sent to the participants by email, thus living up to the experimenter's obligations deriving from the JC. The explicit purpose of this experimental phase was to collect data about the personal feelings related to different kinds of commitments experienced by the participants during the day. Moreover, the participants had been informed that in this phase the experimenter would have monitored their answers, in order to provide the participants, during the second meeting, with a complete evaluation of their personal 'style of commitment.'

Of the 35 participants (all of whom explicitly agreed to carry out the research), 21 came to the second meeting, which was structured in two sections as follows:
1. In the first part, the maintenance section, the 21 participants, that expected to receive comments on their answers, were asked about their experience during the previous week.
2. In the second part, the violation section, the experimenter told the participants that their diaries had not been examined, thus making all their work useless. The violation of the JC was attributed to two different motivations:
a. to 11 of the 21 participants, the experimenter said that she changed her mind and that she no longer wanted to complete the experiment (we call this the internal attribution of the violation);
b. to the other 10 participants, the experimenter said that it was impossible to complete the experiment, due to the fact that the university refused to approve her project (we call this the external attribution of the violation).
The reaction of the participants to the JC violation was recorded for further analysis. At the end of the meeting the actual design of the experiment was revealed to the participants.

\section*{Data coding}

The non-verbal behaviors displayed in the two meetings were analyzed using the ECSI method. The entire duration of the video-recorded meeting was therefore analyzed by identifying, for each participant, the occurrences of the 37 behavior patterns specified by Troisi (1999).
To monitor the lifecycle of the JC, we examined the participants' answers to the self-evaluation diaries. For each participant, the 35 diaries were coded according to the methodology described by Grazzani-Gavazzi and Oatley (1999). In order to better adapt this methodology to the specific purpose of our research, the items regarding the participants' private emotions during the management of the JC and those attributed by the participant to other people (item 7 and 8) have been coded according to the Mind Reading Emotional Library, and not according to JohnsonLaird's classification as in the original methodology.
In the second meeting, the analysis of the non-verbal reaction to the violation of the joint commitment was conducted using the Action Unit analysis of FACS. We scored the Action Units of the participants occurring in the 30 seconds following the experimenter's violation. Moreover, two independent judges, experts in psychotherapy, were asked to identify the prevalent emotion displayed by each participant as a reaction to the JC violation, and to classify it according to the Mind Reading Emotional Library. The two judges were showed the video-recordings of participants' emotional reaction to the JC violation, but were not explained the situation in which the participants were involved; the
question they were asked was, "These people were told an unexpected piece of news: what kind of emotional reaction are they showing?"

\section*{Results}

To understand which factors may be predictive of the participants' fulfillment of the JC two groups were defined, respectively including the 21 participants who came back to the second meeting (Returned -R ), and the 14 participants who did not come back (Not Returned - NR).
We first analyzed the questionnaires regarding personality (EPI, MQOP) and empathic disposition (IRI, ERQ), to understand whether we could identify a predisposition to maintain the JC. Comparing the questionnaire answers obtained by the R and the NR groups, no significant difference was found.
We then considered the number of occurrences of the 37 ECSI behavior patterns in the different sections of the first meeting. While the comparison between the two groups did not show any significant difference in the ECSI behavior patterns during the description and acceptance sections, in the request section we found significant differences between the groups, concerning 10 of the 37 behavior patterns (see Table 1).
To analyze the lifecycle of the JC, we examined the 408 self-evaluation diaries that were returned by the 21 participants who came to the second meeting. Every participant answered on the average 19.42 diaries ( \(\mathrm{sd}=9.31\) ) of the 35 received during the week. The answers to the diaries did not contain any important information on the participants' emotions concerning JCs in their everyday life.
In any case, we used the diaries as a marker of the participants' commitment, dividing the participants into two groups: \(\geq 25\) answered diaries, \(<25\) answered diaries. In fact, no significant difference between these two groups was observed in the ECSI behaviors exhibited during the first meeting.

Table 1: ECSI behavior patterns at the first meeting: significant differences between the R and NR groups ( \(t\)-test for independent samples).
\begin{tabular}{lc}
\hline \multicolumn{1}{c}{ Pattern } & \multicolumn{1}{c}{\(\mathbf{p}\)} \\
\hline 4. Flash & \(\mathrm{p}<.046\) \\
6. Smile & \(\mathrm{p}<.003\) \\
9. Mouth corners back & \(\mathrm{p}<.009\) \\
12. Shut & \(\mathrm{p}<.003\) \\
16. Shake & \(\mathrm{p}<.003\) \\
23. Gesture & \(\mathrm{p}<.001\) \\
29. Fumble & \(\mathrm{p}<.014\) \\
33. Relax & \(\mathrm{p}<.015\) \\
34. Settle & \(\mathrm{p}<.004\) \\
37. Neutral face & \(\mathrm{p}<.008\) \\
\hline
\end{tabular}

However, we found a significant difference between the two groups in the violation section, where certain behaviors occurred significantly more often in the \(<25\) group (4. Flash, \(\mathrm{p}<.027\); 15. Still, \(\mathrm{p}<.003\); \(t\)-test for independent samples).

As for the second meeting, we first compared the ECSI behaviors in the maintenance and violation sections, for all the 21 participants who came back ( R group). No significant difference was found. For the violation section we conducted a further analysis of ECSI patterns, which revealed significant differences in certain behaviors between the internal and the external attribution groups, as shown in Table 2.

The emotional reactions observed after violation were analyzed using the Mind Reading Emotional Library. The index of agreement between the two judges was evaluated as Cohen's \(\kappa=.79\). The distribution of emotions according to the two types of violation is summarized in Table 3.

Table 2: Significantly different ECSI patterns in the second meeting ( \(t\)-test for independent samples).
\begin{tabular}{lc}
\hline \begin{tabular}{c} 
Behavior patterns whose mean is significantly \\
higher in the internal attribution group
\end{tabular} \\
\hline \multicolumn{1}{c}{ Pattern } & p \\
\hline 2. Head to side & \(\mathrm{p}<.020\) \\
3. Bob & \(\mathrm{p}<.035\) \\
10. Look away & \(\mathrm{p}<.010\) \\
14. Crouch & \(\mathrm{p}<.001\) \\
18. Lean forward & \(\mathrm{p}<.001\) \\
24. Groom & \(\mathrm{p}<.013\) \\
36. Laugh & \(\mathrm{p}<.010\) \\
\hline \multicolumn{3}{c}{ Behavior patterns whose mean is significantly } \\
higher in the external attribution group \\
\hline \multicolumn{2}{r}{ Pattern } \\
\hline 12. Shut & p \\
13. Chin & \(\mathrm{p}<.029\) \\
\hline \multicolumn{3}{c}{} \\
\hline
\end{tabular}

Table 3: Distribution of emotions according to the two types of violation.
\begin{tabular}{lcc}
\hline \multicolumn{1}{c}{ Emotion } & \begin{tabular}{c} 
Internal \\
attribution
\end{tabular} & \begin{tabular}{c} 
External \\
attribution
\end{tabular} \\
\hline Sad & 1 & 4 \\
Hurt & 0 & 4 \\
Angry & 3 & 1 \\
Unfriendly & 3 & 1 \\
Surprised & 0 & 1 \\
Disbelieving & 1 & 0 \\
Bored & 1 & 0 \\
\hline
\end{tabular}

Table 4: Emotions according to FACS and Mind Reading (action units in parentheses, partial agreements in italics)
\begin{tabular}{lll}
\hline \begin{tabular}{l} 
Upper Face \\
Decoding
\end{tabular} & \begin{tabular}{l} 
Lower Face \\
Decoding
\end{tabular} & \begin{tabular}{l} 
Mind \\
Reading
\end{tabular} \\
\hline Surprise (1+2+5) & Anger (4+5+23) & Unfriendly \\
Surprise (2+5) & Disgust + Anger (9+10) & Hurt \\
Anger (7) & Disgust (9) & Unfriendly \\
Surprise (1+2+5) & Surprise (26) & Sad \\
Surprise (1+2+5) & Anger(23) & Unfriendly \\
Surprise (1+2+5) & Anger (23) & Sad \\
Surprise (1+2+5) & Anger (23) & Bored \\
Surprise (1+2+5) & Sadness (17) & Angry \\
Surprise (1+2+5) & Sadness (17) & Disbelieving \\
Surprise (2+5) & Anger (23) & Sad \\
Surprise (1+2+5) & Sadness (17) & Angry \\
Surprise (2+5) & Anger (23) & Unfriendly \\
- & Anger (23) & Unfriendly \\
Surprise (1+2+5) & Anger (23) & Surprised \\
- & Sadness (17) & Angry \\
Sadness (1+4+7) & Anger (23) & Angry \\
Surprise (2+5) & Anger (23) & Sad \\
Sadness (1+4+7) & Sadness (15+17) & Sad \\
Sadness (1+4+7) & Disgust+Sadness (9+17) & Angry \\
Anger (4+5) & Sadness (17) & Hurt \\
Anger (4+5) & Anger (23) & Hurt \\
\hline
\end{tabular}

Excluding the neutral ones (i.e., surprised, disbelieving, and bored) these emotions were classified in two groups: sad and hurt as self-centered emotions, angry and unfriendly as other-centered emotions. The comparison between these two groups of emotions revealed a significant difference between the two types of violation ( \(\chi^{2}\) test with Yates's correction, \(\chi^{2}=4.743, \mathrm{p}<.03\) ), where the selfcentered emotions were predominant in the external attribution group, and the other-centered emotions were predominant in the internal attribution group.
The analysis of Action Units, carried out through FACS and decoded into primary emotional expressions, partially agreed with the Mind Reading analysis, as shown in Table 4. Finally, the emotional reaction to the violation of JC was statistically related neither with personality nor with empathic disposition.

\section*{Discussion}

Our results do not show any effect of the participants' characteristics, such as personality and empathic disposition, on the disposition to maintain or violate a JC. Instead, our data indicate that certain non-verbal behaviors carry relevant information on the subjects' attitude toward the JC they are currently creating. In fact, we observed that some nonverbal behaviors displayed in the first meeting are significantly different between the NR and the R group. It would be interesting to understand whether these behaviors are
voluntary attempts to hide a lack of interest in the joint project (irrespective of the verbal acceptance of the JC), or prereflective bodily expressions of uncertainty about the decision of taking part in the project. In the latter case, the recurring behavioral patterns in the participants who did not come to the second meeting may indicate a feeling of uneasiness concerning the situation they are currently experiencing.
Regarding the lifecycle of the JC, the analysis of the diaries did not yield any important indication on how the commitment was experienced during the week between the two meetings. The average number of diaries answered by the participants who came to the second meeting (19.42 out of 35 , i.e., \(55.5 \%\) ) was rather low, which suggests that even those participants who at least partially fulfilled the JC regarded coming to the second meeting as more important than completing the assignments. The only significant behavioral difference that we found between the \(\geq 25\) and the \(<25\) answered diaries groups occurred during the violation section of the second meeting, where we observed that two non-verbal behaviors (i.e., Flash and Still) were performed more frequently by those who answered less than 25 diaries.

Regarding JC violation, the personality and empathic disposition of the participants did not seem to affect their emotional reaction. On the contrary, the two types of violation (internal vs. external attribution) significantly affected the reactions in the participants, as highlighted by both the behavioral occurrences observed through the ECSI method and the emotional reactions detected by the two judges. Whether such differences were under voluntary control or pertained to the sphere of pre-reflective reactions cannot be established on the basis of our current experimental design.

More specifically, as far as emotions are concerned the Mind Reading analysis carried out by the two judges highlighted two different reactions:
1. In the case of violation with external attribution the most frequent emotions were "sad" and "hurt," which can be regarded as self-centered emotions. This is probably due to the fact that the participants, while feeling frustrated because their work turned out to be useless, were willing to consider the experimenter's violation as excused by the university's refusal to approve the project.
2. In the case of violation with internal attribution, the most frequently observed emotions were "angry" and "unfriendly," which can be regarded as other-centered, as they are directed to another subject. These emotions plausibly reveal a feeling of resentment toward the experimenter, who is considered as fully responsible for the violation of the JC.
We believe that these results may be explained by taking into account the normative structure of JCs, and in particu-
lar their second-personal nature (Darwall, 2006) and their relationships with so-called reactive attitudes (Strawson, 1962). Investigating these aspects of interpersonal normativity is among our future research goals.
Finally, the comparison of the emotions coded through Mind Reading and those coded through FACS only showed partial agreement (for 10 on 21 participants). The reason for this partial agreement may be connected with the different focus of analysis these tools are based on; in fact, whereas FACS focuses on facial micro-expressions that prereflectively arise before a social mediation of the emotion, the Mind Reading coding system also takes social emotions into account.
To conclude, our results allow us to identify some nonverbal behaviors as typical of the participants who, even if they verbally agreed to create a JC, did not fulfill the corresponding obligations. Moreover, our experiment suggests that the type of violation attribution (internal vs. external) is the most significant factor in shaping the reactive emotions, overshadowing the effect of personality and empathic disposition.

\section*{Acknowledgments}

We would like to thank Pamela Ventura, Alessandra Preziosa, and Marco Bani for their precious contributions to the research and the analysis of the data. Francesca Morganti and and Ivana Sgro were supported by the Department of Human and Social Sciences, University of Bergamo (Fondi di Ricerca di Ateneo 2012).

\section*{References}

Baron-Cohen, S., Golan, O., Wheelwright, S., \& Hill, J., 2004. Mind Reading: The Interactive Guide to Emotions. London: Jessica Kingsley Publishers.

Baron-Cohen, S., 2003. Mind Reading: Emotions Library User Guide Version 1.2. London, New York: Jessica Kingsley Publishers.
Carassa, A., Colombetti, M., \& Morganti, F., 2008. The role of Joint Commitment in intersubjectivity. In: Morganti, F., Carassa, A., Riva, G. (eds.), Enacting Intersubjectivity: A cognitive and social perspective to the study of interactions. Amsterdam: IOS Press.
Carassa, A., \& Colombetti, M., 2009. Joint meaning. Journal of Pragmatics, 41 (9), 1837-1854.
Darwall, S., 2006. The second-person standpoint: Morality, respect, and accountability. Cambridge, MA: Harvard University Press.
Davis, M.H., 1983. Measuring Individual Differences in Empathy. JSAS Catalog of Selected Documents in Psychology, 44, 113-126.

Davis, M.H., 1980. A Multidimensional Approch to Individual Differences in Empathy. JSAS Catalog of Selected Documents in Psychology, 10, 85-92.

Ekman, P., \& Friesen, W.V., 1978. Facial Action Coding System. A Technique for the Measurement of Facial Movement. Palo Alto: Consulting Psychologists Press.
Eysenck, H.J., \& Eysenck, M.W., 1985. Personality and individual differences: A natural science approach. New York: Plenum.
Gilbert, M. P., 1989. On Social Facts. Oxford: Princeton University Press.
Grazzani Gavazzi, I., \& Oatley, K., 1999. The Experience of Emotions of Interdependence and Independence following Interpersonal Errors in Italy and Anglophone Canada. Cognition and Emotion, 13 (1), 49-63.
Grice, H.P., 1957. Meaning. The Philosophical Review, 66 (3), 377-388.

Gross, J.J., \& John, O.P., 2003. Individual differences in two emotion regulation processes: Implication for affect, relationship, and well-being. Journal of Personality and Social Psychology, 85 (2), 348-362.
Guidano, V., 1987. Complexity of the Self. New York: Guilford.
Larson, R., \& Csikszentmihalyi, M., 1983. The experience sampling method. New Directions for Methodology of Social and Behavioral Science, 15, 41-56.
Morganti, F., Carassa, A., \& Riva, G., 2008. Enacting Intersubjectivity: A cognitive and social perspective to the study of interactions. Amsterdam: IOS Press.
Nardi, B., Arimantea, E., Giovagnoli, S., Blasi, S., Bellantuono, C. \& Rezzonico, G., 2012. The mini questionnaire of personal organization (MQOP): Preliminary validation of a new post-rationalist personality questionnaire. Clinical Psychology and Psychotherapy, 19, 78-90.
Oatley, K., \& Larocque, L., 1994. Emotions and their effects following mistakes in plans involving other people. Proceedings of the \(8^{\text {th }}\) Conference of the International Society for Research on Emotions. ISRE Publications, 219-223.
Strawson, P. F. (1962). Freedom and resentment. Proceedings of the British Academy, 48, 187-211.
Troisi, A., 1999. Ethological research in clinical psychiatry: The study of nonverbal behavior during interviews. Neuroscience \& Biobehavioral Review, 23, 905-913.

\title{
Causal determinism in toddlers
}

\author{
Paul Muentener (pmuenten@mit.edu) \& Laura Schulz \\ Department of Brain \& Cognitive Sciences, MIT, Cambridge, MA 02138 USA
}

\begin{abstract}
Prior research has shown that children hold a belief in causal determinism - the belief that all events are caused - by 4 years of age. In this study we investigate the developmental origins of this belief. We showed toddlers ( 24 months) a spontaneous or explained novel physical outcome (a toy that lit up either spontaneously or upon contact from an experimenter) and then showed them an additional candidate cause (pressing a button) while obscuring the outcome. We asked whether toddlers inferred that the two components (the button and the outcome) were causally linked. We found that toddlers represented the candidate cause as the cause of the novel outcome only when the event spontaneously occurred (Experiments 1-2), and that children spontaneously searched for plausible causes of unexplained outcomes (Experiment 3). These results suggest that toddlers, like older children, believe physical events have causes, and that this belief supports exploration and discovery.
\end{abstract}

Keywords: causal reasoning; determinism; physical causality; prediction; intervention.

Researchers have suggested that children's sophisticated causal inference abilities are at the core of theory development and the many conceptual changes that occur throughout early childhood (Carey, 1985, 2009; Gopnik \& Meltzoff, 1997; Gopink \& Wellman, in press, Schulz, 2012). By preschool, children engage in causal exploration, use conditional probabilities to determine the causal structure of events, and can design appropriate causal interventions (e.g., Bullock, Gelman, \& Baillargeon, 1982; Gopnik \& Sobel, 2000; Gopnik et al., 2004; Kushnir \& Gopnik, 2007; Shultz, 1982).

What drives children's search for causal structure in the world? Although some events in the world involve visible interventions (e.g., human action) and visible outcomes (e.g., objects that move or change state), many events involve unobserved or even unobservable causal mechanisms (e.g., viruses cause disease). Thus, a challenge for theories of conceptual development is to explain how children go beyond the evidence they see.

One possibility is that children are causal determinists. In its most basic form, causal determinism is the belief that all events have causes. If an event appears to occur spontaneously (e.g., a light turns on) adults will typically infer the presence of an unobserved generative cause (e.g., a person activating a hidden switch). A belief in causal determinism could help guide children's search for unobserved variables.

Prior research suggests that by the age of five, children are determinists about physical events. In classic research on causal reasoning, Bullock, Gelman, \& Baillargeon (1982) showed that 5 -year-olds denied that that events could
occur spontaneously. When asked to explain a novel, apparently spontaneous jack-in-the-box event, no child suggested that the event occurred on its own. Rather, all children referred to hidden variables (e.g, wires, remote controls, or "invisible batteries"). More recently, Schulz \& Somerville (2005) found that four and five-year-old children also posited hidden causal variables when outcomes occurred probabilistically.

If a belief in causal determinism is integral to human causal learning and exploration, it might be in place very early in development. Note however, that it is not obvious that the assumption of determinism is necessary either for accurate prediction or effective action. In principle, it might be possible to learn statistical relationships between actions and outcomes (e.g., Blaisdell, Sawa, Leising, \& Waldmann, 2006) and even to innovate causally effective tools (Emery \& Clayton, 2004) without assuming that the world is saturated with causality (though see Gershman, Blei, \& Niv, 2010 for evidence suggesting that inferring latent variables may be integral to causal reasoning broadly). If the assumption of determinism is a relatively late development, children might come to believe that all events have causes only after they have been instructed in unobservable causes like gravity and germs.

Here, we explore the developmental origins of causal determinism by asking whether 18- to 30 -month-old children believe that physical events have causes. We show toddlers an event (a light turning on) that either appears to occur spontaneously or that appears to be caused by the experimenter's preceding intentional action. We then introduce a novel button as a plausible candidate cause for the event (but never show the toddlers any predictive relationship between the button and the light). If toddlers believe that all physical events have causes, then they should ignore the button when the experimenter's intentional action potentially explains the event but reference the button when the event is otherwise unexplained. We test the prediction that toddlers selectively infer causes for unexplained events by investigating toddlers' predictive looks (Experiment 1), their interventions (Experiment 2), and their exploratory behavior (Experiment 3).

\section*{Experiment 1}

\section*{Methods}

Participants Thirty two toddlers (mean: 24 months, range -18-30 months) were recruited at a Children's Museum. An additional 10 toddlers were recruited but not included in the final sample due to: inability to complete the session \((n=4)\), parental interference \((n=4)\), or experimenter error \((n=2)\).

Toddlers were assigned to either the Spontaneous condition or the Explained condition ( \(\mathrm{n}=16 /\) condition). There were no age differences between the conditions ( \(p=n s\) ).

Materials The light box was constructed from a black box ( 6 in \(x 6\) in \(\times 6\) in) with a small blue lamp ( 2 in diameter) emerging from the front panel which was controlled surreptitiously by the experimenter. An orange button box was connected to the black box by a long orange rod (15 in). A black screen served as an occluder throughout the procedure. An additional black screen was placed behind the black box to obscure the experimenter's surreptitious activation of the blue lamp.

Procedure Figure 1 presents a schematic depiction of the procedure from Experiments 1-3. Upon entering the testing space, all children saw the button box connected to the light box. The experimenter directed the child's attention to all components of the novel toy (the button box, the connected rod, and the light box) without labeling the specific items (e.g., "Look at this") (see Figure 1, top panel). The button was then occluded from the child's view with the black screen. In the Spontaneous condition, toddlers saw the light box light up and flash blue ( 4 flashes, approximately 1 s total) apparently spontaneously. In the Explained condition, the experimenter touched the rim of the light and then light box lit up and flashed blue.

In both conditions the experimenter then moved the occluder to reveal the button box and occlude the light box. The experimenter then pushed the button for 1 s .

During the test trial, the experimenter removed the occluder from in front of the light box so that all components were visible to the child. The experimenter pressed the button but the light box did not light up and flash blue. We coded toddlers' first look in the 2 -second window following the button press.

\section*{Results and Discussion}

Figure 2 displays the results from Experiments 1-3. Toddlers in Experiment 1 were significantly more likely to look to the box in the Spontaneous condition (68.75 \%, \(11 / 16\) toddlers) than in the Explained condition ( 25.00 \%, \(4 / 16\) toddlers; Fisher's exact test, \(p<.05\) ) (Figure 2, left panel). That is, toddlers inferred a predictive relationship between a novel event and a candidate cause, but only when the event had no other candidate explanation.

These results are consistent with the possibility that 2 -year-olds believe that physical events have causes. When they saw a novel event that appeared to occur spontaneously and a plausible candidate cause (a button press), toddlers made a predictive look from the candidate cause to the novel event even though they had never seen a predictive relationship between the button press and the light. By contrast, when the novel event could be explained by the experimenter's action, the toddlers did not make a predictive look from the candidate cause to the light.


Figure 1: Procedure for Experiments 1-3.
However, not all predictive relationships are causal relationships. Predictive looking cannot establish that the toddlers in Experiment 1 inferred that the button press was the actual cause of the light activating. Also, in Experiment 1 the experimenter touched the light box in the Explained condition but not in the Spontaneous condition; arguably the experimenter's attention to the light box in the Explained condition drew the children's attention away from the button. In Experiment 2, we matched the experimenter's contact with the light box between conditions and we introduce a stronger test of children's belief in causal determinism: we looked at whether toddlers would selectively intervene on the button. If children believe in causal determinism for physical events, then when asked to turn on the light they should push the button more in the Spontaneous condition than the Explained condition.

\section*{Experiment 2}

Participants Thirty two toddlers (mean: 24 months, range -18-30 months) were recruited at a Children's Museum. Seven additional toddlers were recruited but not included in the final sample due parental interference \((\mathrm{n}=3)\) and failure to intervene \((\mathrm{n}=4)\). Children were assigned to either the


Figure 2: Results of Experiments 1-3.

Spontaneous or Explained condition ( \(\mathrm{n}=16 /\) condition). There were no age differences between conditions ( \(p>.05\) ).

Materials The same materials used in Experiment 1 were used in Experiment 2.

Procedure The procedure was identical to Experiment 1 with the following exception (see Figure 1, middle panel). The experimenter touched the light box in both conditions: in the Spontaneous condition he touched the light box immediately after the light turned on (so that it looked like a response to, rather than potential cause of, the light activating); in the Explained condition, he touched the light box immediately before (as in Experiment 1).

During the test event, the experimenter did not push the button. Instead, he asked the child to "make the light turn on." We coded whether the child first touched the button or the light box within a 30 -second window following the prompt.

\section*{Results and Discussion}

Toddlers were more likely to intervene on the button in the Spontaneous condition ( \(81.25 \%, 13 / 16\) toddlers) than in the Explained condition ( 37.50 \%, 6/16 toddlers; Fisher's Exact test, \(p<.05\) )) (see Figure 2, middle panel). In contrast, toddlers were more likely to initially intervene on the light in the Explained condition ( \(62.5 \%, 10 / 16\) toddlers) than in the Spontaneous condition ( \(18.75 \%, 3 / 16\) toddlers; Fisher's Exact test, \(p<.05\) )). Toddlers seemed to infer a causal relationship between the button and the light only when the light did not already have an apparent cause.

The data from Experiment 2 provide stronger evidence that toddlers believe in causal determinism for physical events. Using interventions as a measure of causal knowledge, toddlers selectively accept candidate causal mechanisms for outcomes only when the event appears to occur spontaneously.

Note that the experimenter contacted both the button and the light in both the Spontaneous condition and the Explained condition. The only difference between the
conditions was whether the experimenter's action on the light could be represented as a cause of the lights flashing; in the Explained condition it could, but in the Spontaneous condition it could not. Thus, the children's tendency to imitate the experimenter's action was influenced by the children's causal attributions.

Experiments 1 and 2 suggest that children's belief in causal determinism affects their search for unobserved causes of physical events. However, neither of these experiments provides a direct test of children's causal exploration. In the prior experiments, toddlers were given a potential causal mechanism (a button) and a relevant action on that mechanism (pressing the button). We do not know whether toddlers in the Spontaneous condition (1) inferred the presence of an external cause and actively searched for it or (2) whether they linked the two subevents of the spontaneous light flash and the button press only after the experimenter directed the child's attention towards the button by acting on it. If a belief in causal determinism guides children's causal exploration, then children might search for a candidate cause even if the experimenter does not direct the children's attention towards it.

This prediction requires a caveat however. Whether a learner actually engages in search depends on many factors, including the learner's prior knowledge, the size of the search space, and exploration/exploitation trade-offs relating the cost and benefit of exploration to the cost and benefit of other actions the learner might take (see e.g., Gittens, 1979). Thus a belief in determinism does not mean that learners will always search for unobserved causes whenever they see unexplained events. Even as adults, we see events every day that we cannot explain; we accept that these events have causes but we rarely bother to seek out the causes ourselves. Nonetheless, if toddlers actively search for plausible candidate causes when events appear to occur spontaneously, then they should be more likely to explore a well-constrained, plausible search space (e.g., the button itself) in the Spontaneous condition than the Explained condition, even if they never observe an intervention on the button. We test this prediction in Experiment 3.

\section*{Experiment 3}

Participants Thirty two toddlers (mean: 23 months, range -18-30 months) were recruited at a Children's Museum. Thirteen additional toddlers were recruited but not included in the final sample due to an inability to complete the session ( \(\mathrm{n}=1\) ), parental interference \((\mathrm{n}=4)\), and failure to intervene \((\mathrm{n}=8)\). Children were assigned to either the Spontaneous or Explained condition ( \(\mathrm{n}=16 /\) condition). There were no age differences between conditions ( \(p>.05\) ).

Materials The same materials used in Experiment 1 were used in Experiment 3.

Procedure The procedure was similar to Experiment 1 except that the toddler did not see the button until the test event (see Figure 1, bottom panel). After the toddler viewed the novel event occur either spontaneously (Spontaneous condition) or as a result of the Experimenter's contact (Explained condition), the Experimenter removed the screen from in front of the button, and then told the child it was his/her turn to play. She did not make any reference to the button and did not explicitly request that the child turn on the light.

We coded whether children intervened on the button within the following 30 -second window.

\section*{Results and Discussion}

Toddlers were more likely to intervene on the button in the Spontaneous condition ( \(81.25 \%, 13 / 16\) toddlers) than in the Explained condition ( 37.50 \%, 6/16 toddlers; Fisher's Exact test, \(p<.05\) ) (see Figure 2, right panel). Even though children had not seen the experimenter act on a plausible candidate mechanism, children selectively explored the candidate mechanism when the novel event seemed to occur spontaneously.

\section*{General Discussion}

The current study suggests that toddlers believe that physical effects have causes. When they saw a novel physical event, they predicted relationships between, intervened on, and explored plausible candidate causes only when the event appears to occur spontaneously. While prior research had shown that four and five-year-olds believe in causal determinism, the current study suggests that the assumption of determinism is present much earlier in development, at least by two years of age.

One possibility is that toddlers' performance in the Spontaneous condition was not driven by a belief in causal determinism, but instead by a prior belief that buttons cause events to happen in the world. That is, toddlers may have made a predictive look towards the light in Experiment 1 because they expected the button press to make something happen rather than because they were looking for a cause of the light. Some evidence that this is not the case comes from the fact that children do not look to the light following the button press when the light's activation can be explained by
another cause. Additionally however, we are currently running a control condition in which toddlers see the button press but never see the light activate. If toddlers look expectantly to the other object on the stage simply because they believe buttons make things happen, they should look in this condition as well. However, preliminary data suggest that toddlers do not make predictive looks following an intervention on the button if they do not have an event to explain.

In the current study, we restricted our investigation of causal determinism to the domain of physical artifacts. Toddlers may assume that events involving artifacts (like a box lighting up) have causes without extending this assumption more broadly. We do not know to what extent children are determinists about naturally occurring physical events. Nor do we know to what extent either adults or children believe in causal determinism for psychological events (e.g., assuming that behaviors like crying, laughing, and thinking always have causes that fully account for their outcomes). The range of contexts under which children believe in causal determinism is an area for future inquiry.

The current research also leaves open the kind of constraints on children's hypothesis space for candidate causes. In this study we provided children with a very plausible, familiar candidate cause: a button. Arguably, as discussed above, children's search for causal structure may rely heavily on the presence of known plausible candidate causes. Alternatively, a belief in causal determinism could guide children's exploration and discovery of genuinely novel causal mechanisms over development. Further work is necessary to know whether toddlers might accept and explore a wider array of candidate causes to account for otherwise unexplained events.

Here, we investigated the simplest form of a belief in causal determinism - that all events have causes. However, a belief in causal determinism can also entail the assumption that causes produce their outcomes deterministically. If events occur probabilistically, a determinist can assume either that a generative cause is sometimes missing or that an inhibitory cause is sometimes present. In related research in our laboratory, we find that toddlers also posit unobserved causes to explain stochastically occurring events. When the event occurs deterministically, they do not make this inference (Wu, Muentener, \& Schulz, 2013; this conference).

Thus a belief in causal determinism may help drive causal learning and exploration starting in early childhood and throughout development. If we assume that all events have causes, then all events are candidates for discovery and exploration, and we can engage in the boundless inquiry that characterizes human cognition.

\section*{Acknowledgments}

This research was funded by an NSF Faculty Early Career Development Award, a John Templeton Foundation Award, and James S. McDonnell Collaborative Interdisciplinary Grant on Causal Reasoning to L.S.

\section*{References}

Blaisdell, A. P., Sawa, K., Leising, K. J., \& Waldmann, M. R. (2006). Causal reasoning in rats. Science, 311, 10201022.

Bullock, M., Gelman, R., \& Baillargeon, R. (1982). The development of causal reasoning. In W. J. Friedman (Ed.), The developmental psychology of time (pp. 209254). New York, NY: Academic Press.

Carey, S. (1985). Conceptual change in childhood Cambridge, MA: MIT Press.
Carey, S. (2009). The origin of concepts. New York: Oxford University Press.
Emery, N., \& Clayton, N. (2004). The mentality of crows: Convergent evolution of intelligence in corvids and apes. Science, 306, 1903-1907
Gershman, S., Blei, D., \& Niv, Y. (2010). Context, learning, and extinction, Psychological Review, 117, 197-209.
Gittens, J. (1979). Bandit processes and dynamic allocation indices. Journal of the Royal Statistical Society. Series B (Methodological). 148-177.
Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T., \& Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. Psychological Review, 111, 1-31.
Gopnik, A., \& Meltzoff, A. N. (1997). Words, thoughts and theories. Cambridge, MA: MIT Press.
Gopnik, A., \& Schulz, L. (2007). Causal learning. Oxford University Press.
Gopnik, A., \& Sobel, D. M. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. Child Development, 71(5), 1205-1222.
Gopnik, A. \& Wellman, H. (in press). Reconstructing constructivism: Causal models, Bayesian learning mechanisms and the theory theory. Psychological Bulletin.
Kushnir, T., \& Gopnik, A. (2007). Conditional probability versus spatial contiguity in causal learning: Preschoolers use new contingency evidence to overcome prior spatial assumptions. Developmental Psychology, 44, 186-196.
Schulz, L. (2012). The origins of inquiry: inductive inference and exploration in early childhood. Trends in Cognitive Science, 16, 382-389.
Schulz, L., \& Sommerville, J. (2006). God does not play dice: Causal determinism in children's inferences about unobserved causes. Child Development, 77(2), 427-442.
Shultz, T.R. (1982). Rules of causal attribution. Monographs of the Society for Research in 0Child Development, 47(1), 1-51.
Wu, Y., Muentener, P., \& Schulz, L. (2013). The invisible hand: Toddlers infer hidden agents when events occur probabilistically. Annual meeting of the Cognitive Science Society.

\title{
Spatial Distance, Availability of Means, and Moral Obligation Judgments
}

\author{
Jonas Nagel (jnagel1@uni-goettingen.de) \\ Michael R. Waldmann (michael.waldmann@bio.uni-goettingen.de) \\ Department of Psychology, University of Göttingen, Germany
}

\begin{abstract}
In the present research we analyze the interrelations of spatial distance and efficaciousness in helping needy others, and we investigate how these factors affect our judgments of moral helping obligations. The main question is under which conditions the location of an agent's means of helping relative to a victim is regarded as morally relevant. We develop a new experimental design that allows us to test our hypotheses concurrently in both separate and joint evaluation modes using a constant procedure across groups. We find that spatial proximity of an agent's means to a victim increases people's sense of obligation only to the extent to which it is indicative of increased efficaciousness or personal involvement.
\end{abstract}

Keywords: moral judgment; spatial distance; means; obligation to help; joint vs. separate evaluation

\section*{Introduction}

The present work explores the relationship between two pervasive moral intuitions. First, we tend to feel more strongly obligated to take care of what is going on near us rather than far from us (see Kamm, 2007). We are more affected by information about harmful incidents occurring in our vicinity than at larger distances and tend to be more inclined to help in near cases. Second, we think that we need to be at least minimally efficacious if we are to be obligated to help others (ought-implies-can principle, OIC; see Vranas, 2007). We usually think we cannot be obligated to do what we cannot achieve. Both intuitions are potentially interrelated: Being near often causes agents to become efficacious in helping a suffering victim. This raises the possibility that the intuition that nearness matters can be reduced to a concrete manifestation of the OIC principle. Near agents may feel more strongly obligated to help not because they are near, but because they perceive themselves to be more efficacious as a consequence of being near. Alternatively, spatial distance might make a difference even at constant levels of efficaciousness (Kamm, 2007).

Nagel and Waldmann (2013) tested this question experimentally. In their Experiment 4, they presented subjects with a case vignette in which a victim was about to be robbed by a thief in a public place. Two agents were standing on the same place, one right next to the victim and the other on the opposite side of the place. Both realized the threat to the victim, and both could do something to prevent the robbery. In one condition, they could walk over to the victim and warn him. Here, the near agent was more efficacious than the far one as the agents needed to traverse the distance in order to be helpful. Spatial proximity caused increased efficaciousness. In the other condition, both agents could send a text message via cell phone in order to
warn the victim. This made both agents equally efficacious, regardless of their distance. It was shown that people judged the near agent to be more obligated than the far one only in the first but not in the second condition. This finding indicates that the effect of the first condition is mediated via efficaciousness considerations. At constant levels of efficaciousness, distance ceases to affect moral judgments.

This conclusion seems to suggest that distance effects can be explained away by efficaciousness. However, the matter is more complicated. Kamm (2007) pointed out that focusing on the distance between agent and victim covers only one way in which distance could matter morally. In addition, agents might be more obligated to victims that are located close to the agent's means, even if they are personally far from both. We conceptualize means as objects with the disposition to bring about an effect intended by an agent in a particular situation. \({ }^{1}\) Both artifacts (e.g., spoons having the disposition to stir liquids) and natural kinds (e.g., tree trunks having the disposition to support ceilings) can serve as means if an agent intends to make them manifest their relevant dispositions. In a typical helping event, an agent intends to bring about a change of state in the victim (from threatened to safe), and he might make certain objects manifest some of their dispositions to achieve this goal. Kamm's (2007) claim is that spatial proximity between such objects and the victim could cause agents to be morally obligated to let these means be used, even if the agent is personally far from both.

This interesting suggestion raises some conceptual problems that have not yet been analyzed. The concept of means is intricately related to both efficaciousness and spatial distance. First, means seem to imply at least a minimal chance of efficaciousness. Second, it seems that most objects have to be (brought) close to the victim at some point during the helping event in order to serve as means. Given these intimate interrelations, how can we separate the claim that nearness of means matters morally from the claim that efficaciousness matters? In what follows, we will offer an analysis of the interplay between distance, efficaciousness, and means. Based on this analysis, we present two experiments testing whether distance between means and victims affects laypeople's moral judgments even at constant levels of efficaciousness.

\section*{Distance, Means, and Moral Obligation}

If you conceive of helping events as causal chains starting at the location of the agent and ending at the location of the

\footnotetext{
\({ }^{1}\) We are not concerned with the special case of using people as means here (see Waldmann, Nagel, \& Wiegmann, 2012).
}
victim, it becomes clear that means can serve two different functions in this process. On the one hand, logistic means enable swift and efficient transport of the causal quantity from agent to victim, making the agent increasingly efficacious across large distances. The cell phones (plus mediating satellite system) from the public place scenario serve this function. Other examples include railroads, electronic media, remote-controls, etc. In order to fulfill this function, such means are often extended in space which makes it difficult to determine their exact location (and thus their distance to the victim). On the other hand, proximate means serve the function to bring about the intended change of state at the victim end of the causal chain. Such objects usually have to be (brought) close in order to become efficacious. Examples for proximate means are headache pills, dollar bills, clothes, organs, etc., depending very much on the specific effect intended by the agent (corresponding to the specific need of the victim). In the above cell phone example, there is no physical object serving this function. Instead, the proximate means would be the text message displayed on the victim's cell phone (changing his state from careless to alert). This exemplifies that logistic vs. proximate means do not correspond to specific kinds of objects. They are differentiated by their function in a specific teleological context. One and the same physical object can serve both functions simultaneously, as when you pick an apple from a high-hanging branch using a rake. You make use of both the logistic disposition of the rake (being long) and of its proximate disposition (having hooks).

Based on this analysis, we can now sensibly ask whether the location of means matters morally independently of efficaciousness considerations. This question refers to proximate means because the location of logistic means often cannot be precisely specified. Proximate means, in turn, need to be (brought) close to the victim in order to be helpful-therefore, they are usually more efficacious when they are near the victim rather than far. However, the presence of efficient logistic means can prevent the detrimental effects of spatial distance on the efficaciousness of proximate means. If the presence of efficient logistic means allows a quick transport of far proximate means to the location of the victim, far proximate means can be as efficacious as near proximate means. Kamm's (2007) claim that means-victim distance matters morally would imply that the location of the proximate means would still make a difference for helping obligations under these conditions. If, by contrast, the location of means mattered only via efficaciousness considerations, the presence of sufficiently efficient logistic means should prevent the location of the proximate means from affecting moral judgments.

\section*{Experiment 1}

In this experiment, we tested whether means-victim distance affects obligation judgments even if the availability of efficient logistic means makes far proximate means equally efficacious to near ones. In the real world, the efficaciousness of proximate means is almost always
reduced if they have to be brought close from a distance. To resolve this confound, we manipulated distance of means (the location of a stick that can be used to rescue a drowning victim) and their efficaciousness (the success probability of the potential rescue action involving the sticks) orthogonally.

\section*{Method}

Participants We obtained data from 183 British subjects (110 female, mean age 38 years) who completed our vignette-based online survey.
Design Our design is based on a classic 2 (Means-Victim Distance: near vs. far) by 2 (Efficaciousness: high vs. low) structure (see the four cells in Figure 1). The scenarios described agents standing in some distance from a canal in which a victim was drowning. They could attempt to rescue the person by running to the canal and reaching out for her with a stick. Our distance manipulation varied whether the stick was located close to the shore and thus near to the victim (near) or next to the agent and thus far from the victim (far). Agents in the far versions had to transport the stick to the shore in order to become efficacious. Our efficaciousness manipulation varied how likely the agent would succeed in his helping attempt. If the stick was located close to the victim (near), it was described as sturdy in the high efficaciousness version (cell a in Figure 1), making it likely that it would carry the victim's weight, and as brittle in the low efficaciousness version (cell b), making it unlikely. If the stick was located far from the victim (far), it was described as light in the high efficaciousness version (cell c), making it likely that it could be brought close in due time, and as very heavy in the low efficaciousness version (cell d), making it unlikely.

We did not simply allocate subjects randomly to one of the four cells. The reason is that we did not want to rely exclusively on a between-subjects variation of the independent variables, as it is well known that between- and within-subjects variations have profoundly different impact on the judgment process (e.g., Hsee \& Zhang, 2010). In the moral domain, we think the most important difference is that the subjects' attention is artificially steered at factors that are varied within-subjects (joint evaluation mode, JE), while the same factor acts as a potentially unattended background condition if it is varied between-subjects (separate evaluation mode, SE; see Nagel \& Waldmann, 2013). As moral judgments in both modes seem interesting, we investigate both concurrently in the present study with a new experimental design that is superimposed on the two-by-two structure described above.

Each scenario contained two agents (instead of only one), Dave and John, standing on opposite sides of the canal. Each agent represented one of two different cells from Figure 1. The agent from each cell was paired with the agent from each of the other three cells, resulting in six conditions (see arrows 1 to 6 in Figure 1). The order in which both agents were described in the scenario was counterbalanced in each of the six conditions, resulting in a total of twelve


Figure 1: Illustration of the experimental design. Numbered arrows correspond to six between-subjects conditions.
scenarios to which our subjects were randomly assigned ( \(n\) s in the six conditions ranging from 29 to 32). Below the scenario description, subjects were presented with two 6point rating scales with which they were to indicate how much they felt both agents were obligated to help. The wording of the two questions was "How strongly do you believe Dave [John] should risk his own life in order to try to save the drowning person?" Each scale was labeled "not at all" at the left hand end (1) and "very strongly" at the right hand end (6). The order of both questions always corresponded to the order in which both agents were introduced in the scenario description.
Analysis The advantage of this procedure is that both within- and between-subjects effects of both factors (distance and efficaciousness) can be concurrently estimated from data that are elicited with a consistent procedure across groups. Each subject judges two agents from different cells, allowing for an estimation of the within-subjects effect of the factor(s) varied between these cells. For example, to estimate the within-subjects effect of distance, we can look at conditions 1 and 2 (see Figure 1) and compare the mean ratings for the near agent with those for the far agent within these groups. This can be formalized as in Equation 1:
\((\) Near/High - Far/High \()+(\) Near/Low - Far/Low \() \neq 0\). [1]
For the within-subjects effect of efficaciousness, we proceed analogously with the efficaciousness contrasts within conditions 3 and 4, that is, we test whether
\[
(\text { Near/High }- \text { Near/Low })+(\text { Far/High }- \text { Far/Low }) \neq 0 \text {. [2] }
\]

At the same time, in conditions 1 to 4 , one of the two factors, distance or efficaciousness, is kept constant at one of its levels within participants. For example, both agents’ means (high and low efficaciousness) in condition 3 are near, whereas both are far in condition 4. Distance is thus a constant background condition within these groups, but is varied between-subjects across the groups. Between-subjects effects of distance can thus be estimated by averaging across both ratings (high and low efficaciousness) for each subject in conditions 3 and 4 and by comparing the means of these averages between these conditions as in Equation 3:
\((\) Near/High + Near/Low \()-(\) Far/High + Far/Low \() \neq 0\). [3]

For the between-subjects effect of efficaciousness, we proceed analogously with the efficaciousness contrast between conditions 1 and 2, that is, we test whether
\((\) Near/High + Far/High \()-(\) Near/Low + Far/Low \() \neq 0\). [4]
Finally, conditions 5 and 6 yield additional information as to how both factors interact when they are concurrently varied in within-subjects comparisons. Condition 5 tests whether subjects make a difference between near/highly efficacious and far/lowly efficacious means, while condition 6 tests whether they make a difference between near/lowly efficacious and far/highly efficacious means. In this last condition, it can be seen if efficaciousness considerations trump distance considerations if both are in conflict.

\section*{Results}

The results are summarized in Figure 2. A three-way mixed 6 (condition 1-6, between-subjects) \(\times 2\) (stimulus i vs. ii, within-subjects) \(\times 2\) (order: i-ii vs. ii-i, between-subjects) ANOVA revealed no main effect of condition, \(F(5\), \(171)<1\), but a main effect of stimulus, \(F(1,171)=25.61\), \(p<.01, \eta_{\mathrm{p}}^{2}=.13\), and a significant Condition \(\times\) Stimulus interaction term, \(F(5,171)=6.67, p<.01, \eta_{p}^{2}=.16\). Order did not affect sense of obligation, \(F(1,171)=1.01, p=.32\), and did not interact with condition, stimulus, or the Condition \(\times\) Stimulus interaction term, all \(F\) s \(<1\). We therefore collapsed the ratings from both order versions for each condition to calculate the specific contrasts from Equations 1 to 4.

The within-subjects contrasts reveal no effect of distance (Equation 1), \(t(171)=-.92, p=.36\), but a significant effect of efficaciousness (Equation 2), \(t(171)=5.08, p<.01\), \(d=.64\). When attending to the content of the varied factors, subjects declare that higher probability of success increases the obligation of an agent to endanger himself in order to help a victim, but that the location of his proximate means is irrelevant given constant efficaciousness. The betweensubjects contrasts reveal no effect of either distance (Equation 3), \(t(171)=.03\), or efficaciousness (Equation 4), \(t(171)=1.34, p=.18\). The results support the conclusion that the location of proximate means relative to a victim does not matter under either evaluation mode. The degree of efficaciousness seems to matter when people compare the obligations of several agents but does not have a measurable influence when varied as a background condition.

When both distance and efficaciousness were varied in a co-acting fashion in the within-subjects contrast (condition 5), people held agents from the near/high efficaciousness cell to be more obligated than agents from the far/low efficaciousness cell, \(t(171)=2.08, p<.05, d=.38\). When both factors were varied in a counteracting fashion (condition 6), people judged agents from the far/high efficaciousness cell to be more obligated than agents from the near/low efficaciousness cell, \(t(171)=4.36, p<.01\), \(d=.81\). Regardless of the location of the means, subjects' moral obligation judgments are tracked by efficaciousness.

\section*{Discussion}

The findings of Experiment 1 suggest that means-victim proximity does not matter morally at constant levels of efficaciousness. Near proximate means only obligate if their nearness implies increased efficaciousness. The fact that efficaciousness exhibited no significant between-subjects effect seems somewhat surprising. Previous research has shown that variations in scope (rather than in quality) are hard to evaluate in SE mode (Hsee \& Zhang, 2010). Accordingly, at constant levels of high or low efficaciousness, our subjects might have merely encoded that both agents can do something to help which is required by the OIC principle. Lacking knowledge about a relevant range of success probabilities in SE mode, subjects might have become insensitive towards variations in degree of efficaciousness beyond this categorical precondition.

\section*{Experiment 2}

In this experiment, we wanted to replicate the basic finding that means-victim distance is irrelevant at constant levels of efficaciousness with a more realistic cover story. At the same time, we wanted to demonstrate that not all effects of the nearness of means are mediated via efficaciousness. Nagel and Waldmann (2013) have argued that spatial proximity between agents and victims is not only indicative of increased efficaciousness, but usually also of increased experiential directness and shared group membership. Similarly, near means do usually not only indicate increased efficaciousness, but also increased personal involvement with the victim. For example, if an agent owns money on a bank account in a faraway country, he is usually more personally involved with this country than if his money was stored elsewhere. Maybe he has visited the country before
and will do so again in the future, or at least he profits from the financial system in the foreign country. If the agent feels obligated to donate this money to sick children suffering in the same foreign country (that is, near his means), this could be due to his increased personal involvement rather than due to means-victim proximity per se.

\section*{Method}

We gathered data from 212 subjects from Great Britain (127 female, mean age 37 years). The experimental design was as in Experiment 1, but instead of efficaciousness we varied personal involvement (high vs. low) orthogonally to meansvictim distance (near vs. far). Efficaciousness was explicitly kept constantly high across all conditions. In this way, it was assured that potential distance effects could not be mediated via efficaciousness considerations. This time, we assigned roughly twice the number of subjects to conditions \(5(n=50)\) and \(6(n=52)\) than to the remaining conditions ( \(n\) s ranging from 26 to 31 ) to have roughly equal numbers of observations in the cell combinations that are compared with each other in the planned contrasts.

The subjects were again assigned randomly to one of the twelve scenario versions resulting from our design (Figure 1). Each scenario contained two British agents having the possibility to donate money (via online banking) to rescue Haitian children who are threatened by a deadly disease. Our distance manipulation varied the location of the agents' money. It was located either at a bank in Haiti (near the sick children), or at a bank in Great Britain (far from the sick children). Our involvement manipulation varied the process by which the agents' means had ended up in their locations. In the high involvement conditions, the agent had personally decided to open an account at the Bank of Haiti and that his money was to be constantly located in a branch of this bank


Figure 2: Results of Experiment 1. Categories on the abscissa represent between-subjects variations, while line graphs correspond to within-subjects variations (cf. Figure 1). C. = condition. Error bars indicate 95\% CIs.
either in Haiti (cell a in Figure 1) or in Great Britain (cell c). In the low involvement conditions, the agent had opened a bank account at an international bank. Employees of this bank regularly transferred the money to the most profitable location, which momentarily happened to be in Haiti (cell b) or in Great Britain (cell d). Note that this is an utterly minimal variation of personal involvement. Neither agent had ever been to Haiti or has had any other personal connection to the country. The only difference was whether a personal decision had caused the money to end up on an account of a bank from the same state as the children.

The wording of the sense of obligation measure was as follows: "How strongly do you believe Dave [John] should transfer \(£ 10\) from his [bank account] in [location] to the donation account?" In the different conditions, the specifications of the bank accounts and their locations were adapted according to the scenario. The scales and the rest of the procedure were identical to those in Experiment 1.

\section*{Results}

The results are summarized in Figure 3. A three-way mixed 6 (condition 1-6, between-subjects) \(\times 2\) (stimulus i vs. ii, within-subjects) \(\times 2\) (order: i-ii vs. ii-i, between-subjects) ANOVA revealed a main effect of condition, \(F(5\), 200) \(=2.29, \quad p<.05, \quad \eta_{p}^{2}=.05, \quad\) and of order, \(F(1\), 200) \(=4.65, p<.05, \eta_{\mathrm{p}}^{2}=.02\), while the within-subjects factor (stimulus) was not significant ( \(F<1\) ). None of the interaction terms approached statistical significance. The order effect resulted from ratings being somewhat higher when stimulus ii was presented before stimulus i. However, since there is no systematic relationship between the underlying factors and the assignment of cells to Stimuli i and ii, and since order did not interact with condition, stimulus, or their interaction (all \(F \mathrm{~s}<1\) ), we collapsed the ratings from both order versions for each condition to calculate the specific contrasts outlined in Equations 1 to 4.

The within-subjects contrasts reveal neither effects of distance (Equation 1), \(t(200)=.00\), nor of personal involvement (Equation 2), \(t(200)=-.16\). When attending to the content of the varied factors, subjects declared that it does not make a difference where the agents' money is located, or how the money ended up in its location. The between-subjects contrasts reveal no effect of distance (Equation 3), \(t(200)=-.16\), but a significant effect of personal involvement (Equation 4), \(t(200)=2.16, p<.05\), \(d=.58\). This shows that our minimalistic variation of personal involvement mattered as a background condition, but was discounted when attention was directed to this factor. Distance, by contrast, did not make a difference in either evaluation mode.

Subjects differentiated between the agents in condition 5, \(t(200)=3.29, \quad p<.01, \quad d=.46\). When proximity and involvement coincided, as in most natural situations, subjects judged the near/highly involved agent to be more obligated than the far/lowly involved agent. However, the null effects of distance in conditions 1 to 4 indicate that this effect cannot be attributed to distance per se. The effect also vanished when both factors were varied in a counteracting fashion (condition 6, \(t[200]=-.02\) ).

\section*{Discussion}

Our findings indicate that spatial distance between an agent's proximate means (his money) and the needy victims does not affect people's sense of obligation if efficient logistic means (online banking) are available, contrary to Kamm's (2007) intuitions which stated that spatial proximity between the agent's means and the victim may increase the agent's obligations to let these means be used, even if the agent is personally far and even at constant levels of efficaciousness. At the same time, people are sensitive to the process by which an agent's proximate means ended up in its location. If nearness between means and victim is at


Figure 3: Results of Experiment 2. Categories on the abscissa represent between-subjects variations, while line graphs correspond to within-subjects variations (cf. Figure 1). C. = condition. Error bars indicate 95\% CIs.
least minimally indicative of an increased personal involvement with the victims (in virtue of having decided to place one's assets in a bank based in their home country), obligation judgments are increased. However, if the nearness between means and victims results from pure coincidence (they are temporally transferred to the near location by a third party), it loses its moral impact.

The effect of personal involvement is limited to SE judgments and disappears in attentive JE mode. Our operationalization of involvement apparently was so minor that people judged it to be irrelevant when attending to it, and yet, it was sufficient to affect their judgments as a background condition. Stronger variations in involvement (e.g., previous visits to Haiti) can of course be expected to be honored in JE mode as well.

\section*{General Discussion}

Previous research suggests that distance between agent and victim is irrelevant for laypeople's judgments of helping obligations (Nagel \& Waldmann, 2013). The present findings extend this conclusion to the more complicated spatial relation between an agent's means and a needy victim, contrary to Kamm (2007) who argued that means of helping that are located spatially close to the victim might increase the agent's obligation to let these means be used at personal costs. In the presence of sufficiently efficient logistic means (making far proximate means efficacious by allowing them to be brought close swiftly), the distance between an agent's proximate means and the victim becomes morally irrelevant, both in separate (SE) and joint evaluation (JE) modes. Means-victim proximity matters morally only when it implies other obligation-inducing factors, such as increased efficaciousness. To the extent that proximity causes efficaciousness, the intuition that we have a strong obligation to help victims near our means can be seen as a manifestation of the ought-implies-can (OIC) principle.

However, we have also seen that not all effects of meansvictim proximity can be explained by the OIC principle. This is because proximity of means is usually not only indicative of increased efficaciousness, but also of increased personal involvement with the victim. In such cases, other cognitive mediators apart from efficaciousness considerations seem to do the moral work (probably emotional engagement and reasons referring to social responsibilities).

The impact of the factors associated with distance was strongly influenced by evaluation mode. Incremental differences in efficaciousness were highly important in the comparison of several agents within a single scenario (JE mode), but they did not affect obligation judgments as a constant background condition (SE mode). Minimal personal involvement, by contrast, was sufficient to increase obligation judgments in SE, but this difference was discounted in JE. These fine-grained observations underline the value of our new experimental design. It allows for a
detailed picture of judgments under different evaluation modes at constant procedural conditions.

The conceptual distinction between logistic and proximate means (which is grounded in their functionally different relationship to spatial distance) does not seem to correspond to a psychologically meaningful distinction. In Experiment 1, subjects did not differentiate between means that were inefficacious due to deficits affecting their potential role as proximate means (i.e., the sticks' brittleness) vs. those that could not be brought close in time with the available logistic means (i.e., because of the sticks' heaviness). What matters morally seems to be the means' reduced efficaciousness in the given context, regardless of the causes for different degrees of efficaciousness and of whether or not these causes are related to distance.

Although the empirical findings are clear-cut, there are some limitations. Our analysis of means seems to be restricted to cases of generative causation, that is, to cases in which a causal chain involving the transmission of a causal quantity is elicited by the agent. It is left unanalyzed what role (if any) means and their location might play in cases of causation by omission (Wolff, Barbey, \& Hausknecht, 2010). A related concern is that our account is mainly tailored to handle cases in which an impending physical harm threatens a victim. It is less clear how the account could deal with moral obligations related to other moral domains (see Haidt \& Graham, 2007). Another special case arises when humans are being used as means. The moral implications of such cases are too far-reaching to be discussed here (see Waldmann, Nagel, \& Wiegmann, 2012, for an overview). As far as obligation judgments in harmbased rescue cases are concerned, however, our findings seem to be clear: Not distance per se but features that are normally associated with distance drive our moral intuitions.

\section*{References}

Haidt, J., \& Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. Social Justice Research, 20, 98-116.
Hsee, C. K., \& Zhang, J. (2010). General evaluability theory. Perspectives on Psychological Science, 5, 343355.

Kamm, F. M. (2007). Intricate ethics. Oxford: Oxford University Press.
Nagel, J., \& Waldmann, M. R. (2013). Deconfounding distance effects in judgments of moral obligation. Journal of Experimental Psychology: Learning, Memory, \& Cognition, 39, 237-252.
Vranas, P. B. M. (2007). I ought, therefore I can. Philosophical Studies, 136, 167-216.
Waldmann, M. R., Nagel, J., \& Wiegmann, A. (2012). Moral judgment. In K. J. Holyoak, \& R. G. Morrison (eds.), The Oxford Handbook of Thinking and Reasoning (pp. 364-389). New York: Oxford University Press.
Wolff, P., Barbey, A. K., \& Hausknecht, M. (2010). For want of a nail: How absences cause events. Journal of Experimental Psychology: General, 139, 191-221.

\title{
Value of a friend, a friend of your friend, and a friend of the friend of your friend: Social discounting in \(\boldsymbol{n}\) degrees of separation
}

\author{
Kuninori Nakamura (knaka@seijo.ac.jp) \\ Faculty of Social Innovation, Seijo Universtiy, 6-1-20, Seijo, Setagayaku, Tokyo 152-0061, Japan
}

\begin{abstract}
Jones and Rachlin (2008) found that the amount of money a person is willing to forgo in order to give \(\$ 75\) to another person decreased as a hyperbolic function of perceived social distance, in the same way as occurs in intertemporal choice. This study aimed to extend this finding to the domain of social networks, in which social distance is defined by degrees of separation. A total of 334 participants responded to tasks very similar to those in Jones and Rachlin (2008), except that they were required to choose whether they would prefer to receive an amount of money for themselves only or an amount of money for themselves and a person who is \(n\) degrees of separation from them up to six degrees. The results show that the hyperbolic function fit the data well, and that several processes appear to contribute to the judgments made in the experimental tasks.
\end{abstract}

Keywords: social discounting; \(n\) degrees of separation; intertemporal discounting, probability discounting

\section*{Introduction}

In modern society, people must use money to live. However, as a matter of course, people cannot use all their money at once. Rather, people must allocate it between several purposes or several time periods. They distribute money between several stock options so as not to lose all their money at once. They save money for the future to avoid financial difficulties when they are old. Additionally, we also must share money with others in order to prevent poverty. As these examples show, how people use money is of vital importance in life, and appears to be organized according to several dimensions.

According to Julian Simon (1995), a person's allocation of available goods can be described in terms of a three-coordinate system: one is the coordinate of their own current consumption, included in which is the concept that a person has several selves corresponding to their various positions, such as the family self or working self, and so need to allocate goods to these selves. The second is a coordinate of later times, representing sequential temporal persons as different from each other. The third is consumption by other people, according to which the feeling of sympathy between people may be measured by a discount factor. Simon (1995) suggested that this discount function may be similar to that of intertemporal discounting.

Consumption in later time periods corresponds to what is called intertemporal discounting. Many studies have demonstrated that people discount the value of goods as time goes by. Generally, it is known that people prefer small but immediate goods to large but delayed goods (e.g.,

Lowenstein \& Prelec, 1998). To account for this preference, exponential and hyperbolic functions have been proposed.

An exponential discounting function has a form in which the discounted value \(v\) of rewards \(V\) is expressed as follows:
\[
\begin{equation*}
v=V e^{-k D} \tag{1}
\end{equation*}
\]
where \(v\) and \(V\) are the discounted and undiscounted reward values, respectively, and \(D\) is the time delay.

Alternatives to exponential discounting have been proposed by psychologists, behavioral ecologists, and behavioral economists. One major alternative proposal is that the discounting function is hyperbolic (e.g., Mazur, 1987):
\[
\begin{equation*}
v=\frac{V}{1+k D} \tag{2}
\end{equation*}
\]

Until now, many studies have demonstrated that the hyperbolic function, rather than the exponential function, is the most appropriate because it fits the data better than the exponential function, under various experimental conditions (for a review, see Green \& Myerson, 2002).

On the basis of findings within the intertemporal choice literature, Jones and Rachlin (2006) investigated whether discounting similar to intertemporal choice could also be found in Simon's third coordinates. In their experiments, Jones and Rachlin (2006) required participants to imagine that they had made a list of the 100 people closest to them in the world, ranging from their dearest friend at position \#1 to a mere acquaintance at \#100, and then participants answered whether they would forgo a fixed amount of money to give it to another person or not. The results showed that the amount of money people were willing to give to another person decreased as a hyperbolic function of the perceived social distance between them, indicating that the discounting function with regard to social distance is similar to that in intertemporal choice, as Simon (1995) claimed. This study aims to extend the findings of Jones and Rachlin (2006) by considering another type of social distance.

Recent studies in network science have begun to pay attention to the concept of human society considered as a network (Barabasi, 2002; Christarski \& Fowler, 2008; Milgram, 1967). You have a friend, and your friend has a friend. If you do not know the friend of your friend, you can still have a link to the person via your friend. You can extend such connections infinitely. In this vein, our society is a network of friend connections, and much research has
paid attention to various aspects of our society considered as such a network (Barabasi, 2002; Christarski \& Fowler, 2008).

In a social network, the \(n^{\text {th }}\) degree of separation can be considered as one type of social distance between two persons. The degree of separation refers to the number of links via friends between two persons. For example, your friend is a friend of the first order, because there is one link between you and your friends, and a friend of your friend is of the second order because there are two links; one is between you and your friend, and the other is between your friend and the friend of your friend. The number of links ( \(n\) ) between two persons can be increased infinitely, and as \(n\) increases, the social distance from you increases. For example, a second order friend is more remote than a first order friend, because you have not met the former whereas you directly know the latter. In addition, a third order friend is more remote than a second order friend, because even your friend does not know the former person. If the degree of separation is taken as social distance, the following questions arise: Does social discounting also appear with this type of social coordinate? If so, is it hyperbolic or exponential? The first purpose of this study is to address these questions.

A second purpose of this study is to explore the relationship between social ranking and degrees of separation as measures of social distance. Although both social ranking and degrees of separation can be considered as indices of social distance between persons, their meanings are a little different. Whereas the former measure reflects the distance among in-group members because it assumes that the ranking orders the 100 people closest to a person, from the dearest friend to a mere acquaintance, the latter measure contains not only in-group but also out-group individuals, because one cannot know a friend whose degree is of more than two. Thus, even if hyperbolic discounting is found in the \(\mathrm{n}^{\text {th }}\) degree of separation measure, the relationship between social ranking and degrees of separation would still be an interesting question. Therefore, this study also aims to address this issue.

For this research two empirical studies have been performed. Study 1 was to examine whether social discounting occurs with social distance in terms of degrees of separation. Study 2 explored the relationship between degrees of separation and closeness ranking, as used in Jones and Rachlin (2006).

\section*{Study 1}

Study 1 required participants to answer whether they would prefer (a) an amount of money for themselves or (b) an amount of money for themselves and the person who is \(n\) degrees of separation from them, up to 6 degrees of separation. By this procedure, we tried to determine the amount of money forgone to give a person 60,000 yen. This procedure is almost the same as that of Jones and Rachlin (2006), except that the closeness ranking had been exchanged for degrees of separation.

\section*{Procedure}

One hundred and nineteen undergraduates participated in Study 1 for course credits, and all materials and response formats were provided in booklets. The booklets explained the meaning of the degrees of separation on the first page, and the experimental tasks began on the second page. The second page contained the following instructions:

You know a person who is called "your friend." The minimal condition for a person to be called a friend is that you and s/he know each other. In addition, there is a "friend of your friend" who is known to your friend but unknown to you. This "friend of your friend" can also know a "friend of the friend of your friend" and a chain of friends can extend infinitely. Thus, we connect to various people through friends, friends of friends, and friends of friends of friends, and so on, although most of them are unknown to you.

Next, you will be asked to make a series of judgments based on your preferences. On each line, you will be asked if you would prefer to receive an amount of money for yourself or an amount of money for the person listed. Please circle A or B for each line.

Each of the next six pages summarized the above instructions and then presented a list of questions as follows, with a different N -value on each page:

Now imagine the following choices between an amount of money for you and an amount for you and for your friend. Circle A or B to indicate which you would choose in EACH line.
(A) 120,000 yen for you alone
(B) 60,000 yen for you and 60,000 yen for your friend
(A) 110,000 yen for you alone
(B) 60,000 yen for you and 60,000 yen for your friend
(A) 60,000 yen for you alone
(B) 60,000 yen for you and 60,000 yen for your friend

The A-rows listed 9 amounts decreasing by 10,000 yen on each line, from 120,000 to 60,000 yen. Row-B had a different degree of separation on each page. The degrees of separation were from one to six. On each line, the participants were asked to choose between an amount of money just for themselves and 60,000 yen each for themselves and for the person. The degrees of separation were manipulated by adding "of friend" to the sentence in line B. For example, the second order friend was described as the "friend of your friend," and the third order friend as the "friend of the friend of your friend." All participants finished their questionnaire within fifteen minutes.

\section*{Results and discussion}

The crossover point was estimated as the average of the last selfish (row-A) choices and the first generous (row-B) choices, in the same way as in Jones and Rachlin (2006) or Rachlin and Jones (2008). For example, if a participant preferred 90,000 yen for herself to having 60,000 yen and giving 60,000 yen to the \(\mathrm{N}^{\text {th }}\) degree of separation friend, but preferred to have 60,000 yen and give 60,000 yen to the \(\mathrm{N}^{\text {th }}\) friend over having 80,000 yen for herself, then the crossover point was taken as being 85,000 yen for that participant at that N -value. Some participants (the majority at N 51 and N 5 2) chose the generous option even when the alternative was 120,000 yen for themselves. In these cases, a crossover point of 125,000 was assumed. In contrast, many participants chose the selfish option even when the choice was between 60,000 yen for themselves and 60,000 yen for themselves in addition to 60,000 yen for their friends. In these cases, the crossover point was assumed to be 0 .

Figure 1 shows the mean allocation of money to friends who have \(n\) degrees of separation. We fitted both the hyperbolic and exponential functions to this data and compared their performances. The solid line is the best fitting hyperbolic discount function and the dashed line is the best fitting exponential discount function. The fit is remarkably good \(\left(R^{2}=0.995\right)\) when compared with the fit of intertemporal choice (e.g., Rachlin \& Raineri, 1992) or of social discounting (Jones \& Rachlin, 2006).

For comparison, the best fitting exponential discount function is also shown in Figure 1, as the dashed line. Although the fit with exponential discounting is high ( \(R^{2}=0.971\) ), the percentage of variance accounted for by the exponential discount function is less than that of the hyperbolic function. The difference found between the fit of the hyperbolic and the exponential discounting is almost the same as that in Jones \& Rachlin (2006). In sum, these results show that the hyperbolic function provides a better fit to the data than the exponential discount function. Thus, we can conclude that people's discounting in terms of degrees of separation is similar to that of social ranking and intertemporal choice.

\section*{Study 2}

Study 2 was designed to investigate a relationship between n degrees of separation and social rankings. Although the two are similar as they both represent some kind of social distance, they do differ as while the latter distance can only represent the remoteness of in-group members, the former includes that of both in-group and outgroup people. Thus, although Study 1 demonstrated hyperbolic discounting occurs in \(n\) degrees of separation, the way in which social ranking and degrees of separation are similar types of social distance still remains unclear. Therefore, Study 2 required participants to indicate their preference between receiving money for themselves and sharing money with their friends, under both social ranking and degrees of separation types of distance.


Figure 1 Hyperbolic and exponential discount functions under the \(n^{\text {th }}\) degree of separation in Study 1

\section*{Procedure}

Two hundred and fourteen participants answered social discounting tasks using both social ranking and degree of separation conditions. As tasks to explore preferences with social distance as degrees of separation, we employed the same tasks as those in Study 1. For the social ranking tasks, we employed a procedure almost the same as that of Jones and Rachlin (2006). Precisely, the participants were provided with instructions that had been translated into Japanese from the original ones used by Jones and Rachlin (2006). Then participants were required to make choices between receiving an amount of money for themselves or receiving an amount of money for themselves and their friends, using the following response form:

Now imagine the following choices between an amount of money for you and an amount for you and for your friend. Circle A or B to indicate which you would choose in EACH line.
(A) 120,000 yen for you alone
(B) 60,000 yen for you and 60,000 yen for your \#_person
(A) 110,000 yen for you alone
(B) 60,000 yen for you and 60,000 yen for your \#_person
.(continuing down to)
(A) 60,000 yen for you alone
(B) 60,000 yen for you and 60,000 yen for your \#_person

The blanks shown above was replaced by a number ( \(\mathrm{N}=1,2\), \(5,10,20,50\), or 100 ), with a different number used on each of the seven question pages.

Participants answered both the social ranking and degree of separation tasks in random order. All participants completed all the tasks within 20 minutes.

\section*{Results and discussion}

Figure 2 shows the results of both the social ranking and the degree of separation tasks. We found that the hyperbolic function fitted the data better than the exponential function in both cases. The differences between the hyperbolic and exponential discounting in Study 2 are more remarkable than those in Study 1, or in Jones and Rachlin (2006); whereas the \(R^{2} \mathrm{~s}\) of the exponential functions are 0.873 and 0.850 , those of the hyperbolic functions are greater than 0.90 . Additionally, the data points systematically deviate from the exponential functions. Thus, we can conclude that Study 2 replicated the results of Study 1 and of Jones and Rachlin (2006).

To explore the relationship between social ranking and degrees of separation in depth, we performed the following two additional analyses. First, we estimated the \(k\) parameters of hyperbolic functions for the social ranking and degree of separation tasks. There was no relationship ( \(r=-0.03, p>.01\) : Figure 3) between \(k\) parameters in the two tasks (Figure 3), indicating that the steepness of the functions in the social ranking and degrees of separation cases are somewhat independent from each other.

Second, we also performed a factor analysis of the crossover points with promax rotation using the maximum likelihood method. The eigenvalues for the first three factors were \(7.04,2.54\), and 1.35 , respectively. Mainly due to the eigenvalue results, we adopted a three-factor solution pattern for the discounting tasks shown in Table 1, assuming the following interpretations of the meaning of the three factors.

Factor 1 leads mainly to allocations for high ranking friends and first degree friends. To define, these friends can be interpreted as close friends, so we named this factor "close friends." Factor 2 can be considered as the "unknown others" factor, because it strongly influences the allocation to friends who have more than one degree of separation. Logically, one cannot directly know friends who are of more than the second order, and specifically, friends who have more than two degrees of separation are actually unknown because they are not even the friend of a first order friend. Factor 3 mainly impacts friends with rankings higher than \(10^{\text {th }}\), and in particular those who are ranked at \(20^{\text {th }}\) place or higher. These friends are considered as not being so close. Thus, we named this factor as the "acquaintances" factor.
The above results suggest that there are several dimensions to social discounting. Specifically, it is interesting that people have two dimensions of social distance with others who are not so close to them. To examine this indication more precisely, we also performed structural equation modeling and compared models of one, two, and three factors (see Figure 4).


Figure 2 Hyperbolic and exponential functions with data from the \(n\) degrees of separation and social ranking tasks in Study 2: The upper graph shows results in the \(n\) degrees of separation case, and the lower shows those in the social ranking case.


Figure 3 Scatterplot of \(k\) parameter

Table 1 Results of factor analysis
\begin{tabular}{cccc}
\hline & Factor1 & Factor2 & Factor3 \\
\cline { 2 - 4 } Degree of separation & & & \\
1st & 0.78 & 0.12 & -0.07 \\
2nd & 0.35 & 0.43 & 0.10 \\
3rd & 0.11 & 0.75 & 0.07 \\
4th & -0.03 & 0.94 & 0.06 \\
5th & -0.05 & 0.99 & -0.02 \\
6th & -0.06 & 0.99 & -0.03 \\
Social ranking & & & \\
1st & 1.00 & -0.03 & -0.10 \\
2nd & 1.01 & -0.03 & -0.07 \\
5th & 0.79 & -0.06 & 0.23 \\
10th & 0.50 & -0.08 & 0.53 \\
20th & 0.26 & -0.11 & 0.79 \\
50th & -0.09 & 0.05 & 0.97 \\
100th & -0.16 & 0.18 & 0.83 \\
Correlations & & & \\
Factor1 & 1.00 & & \\
Factor2 & 0.36 & 1.00 & \\
Factor3 & 0.53 & 0.49 & 1.00 \\
\hline
\end{tabular}

The one factor model represents a hypothesis that both social rankings and degrees of separation can be summarized by one dimension, that people's dimension of social discounting is unitary. In contrast, the two and three factor models assume that social discounting can be decomposed into several dimensions. The two-factor model entails that social discounting occurs separately in the social ranking and the \(n\) degrees of separation tasks. In other words, this model assumes that participants construct a dimension of social discounting in accordance with experimental tasks. The three-factor model expresses an implication, based on the results of the factor analysis, that the two types of social distance considered in this study can be decomposed into three factors: one influences both social ranking and degrees of separation, and the other two factors affect these two dimensions, respectively.

The results of the structural equation modeling, shown in Table 2, clearly support the three factor model, as all of the fit indices indicate it is superior to the other two models. Thus, we can conclude that although social discounting in social rankings and degrees of separation share the same components, they can be decomposed into several dimensions. That is, while these two types of social distance are similar in how they reflect allocations made to closer friends, they differ in representing the allocations made to others who are not so close.


Figure 4 Three structural equation models: (a) a one-factor model that indicates both the social ranking and \(n\) degrees of separation can be summarized by one dimension; (b) a two factor model that implies participants construct dimensions of social discounting corresponding to the experimental tasks; (c) a three-factor model that reflects the implication of the factor analysis that social rankings and degrees of separation share one common factor ("friend"), but are also individually affected by "acquaintances" and "unknown others" factors, respectively.

\section*{General discussion}

The results of the two studies can be summarized as follows. First, we found that there is hyperbolic social discounting with the \(n\) degrees of separation type of social distance. Recently, many researchers have paid attention to the way in which the structure of our social networks affects human life (e.g., Christarski \& Fowler, 2009). Most of these studies investigate how people's behavior affects others through links between persons. However, previous studies have not considered the way in which people consider others in their social network. In this vein, this may be the
first study that concerns how people consider others who have \(n\)

Table 2 Results of structural equation modeling.
\begin{tabular}{lcccc}
\hline & AIC & BIC & GFI & CFI \\
\cline { 2 - 5 } Model 1 & 311.98 & 497.11 & 0.88 & 0.96 \\
Model 2 & 342.47 & 534.33 & 0.90 & 0.95 \\
Model 3 & 186.37 & 331.11 & 0.94 & 0.99 \\
\hline
\end{tabular}
degrees of separation. Furthermore, as far as we know, this is the first study that shows hyperbolic discounting occurring with social distance other than in closeness rankings. Hyperbolic discounting is found not only in the domain of social discounting but also in probability or intertemporal discounting (Jones \& Rachlin, 2010). Thus, this study applied hyperbolic discounting to another type of social distance and has demonstrated with evidence the hyperbolic function's ability to explain discounting under various types of psychological distance.

Second, from the results of estimating the parameters of hyperbolic discounting and the structural equation modeling, we can conclude that people have two dimensions to their conceptualization of others who are not so close to them: one applies to others who are known but are not so close, and the other applies to those who are unknown. Intuitively, these appear to be quite natural results. However, these results are interesting because hyperbolic discounting occurs under both dimensions, despite them being independent of each other. Additionally, this finding is also of interest because it appears to contradict Simon's (1995) suggestion that social discounting is one-dimensional.

The results of this study may support construal level theory (Trope \& Lieberman, 2010). This theory assumes that people's judgments of various types of psychological distance can be decomposed into two levels of construal: higher and lower levels. The theory claims that high-level construals are relatively abstract, coherent, and superordinate mental representations as compared with low-level construals. It also argues that people use increasingly higher levels of construal to represent an object as the psychological distance from the object increases. The results of our factor analysis and structural equation modeling consistently show that both the social ranking and \(n\) degrees of separation can be decomposed into two factors. One factor reflects allocations made to psychologically close friends, and the other reflects those to psychologically remote friends. This factor structure appears to match the structure entailed by construal level theory.

In addition, what is more interesting in the results of the factor analysis is that the factor reflecting allocations to close friends is related to both the social ranking and \(n\) degrees of separation types of distance. As stated in the introduction, social rankings and degrees of separation are
somewhat different dimensions of social distance: while the former represents distances between known others, the latter contains those between unknown others. In this vein, the results of the factor analysis suggest that the contents of the "friends" factor are richer than those of the other two factors, because this factor influences the two different types of social dimension. If we interpret the "friends" factor as a lower level construal that is psychologically closer, this indication corresponds to a proposition of construal level theory that lower level construals have more complex representations than higher level construals (Trope \& Lieberman, 2010).

One important issue for future research regarding social discounting in cases of \(n\) degrees of separation may be its relationship to probability or intertemporal discounting. Jones and Rachlin (2009) reported that the steepness of discounting, represented by the \(k\) parameter, was correlated for probability, intertemporal, and social ranking discounting across participants. The results of Study 2 show that the \(k\) parameter did not correlate between the social ranking and \(n\) degrees of separation experiments, suggesting that the latter dimension may be unique and different from the other three dimensions. If so, what is the meaning of social discounting? Why does the \(n\) degrees of separation differ from delay, probability, and social ranking in its discounting? Pursuing this question may be fruitful in exploring how "social" is represented in the human mind.

\section*{References}

Barabási, A. L. (2002). Linked: The New Science of Networks, Perseus Books Group
Christakis, N. A., \& Fowler, J. (2009). Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives. Little, Brown and Company.
Green, L., \& Myerson, J. (2002). A discounting framework for choice with delayed and probabilistic reward. Psychological Bulletin, 130, 769-792.
Jones, B., \& Rachlin, H. (2006). Social discounting. Psychological Science, 17, 283-286.
Jones, B., \& Rachlin, H. (2009). Delay, probability and social discounting in a public goods game. Journal of Experimental Analysis of Behavior, 91, 61-73.
Lowenstein, G., \& Thaler, R. (1989). Intertemporal choice. Journal of Economic Perspective, 3, 181-193.
Milgram, S. (1967). The Small-world problem. Psychology Today, 1, 60-67.
Rachlin, H., \& Raineri, A. (1992). Irrationality, impulsiveness, and selfishness as discount reversal effects. In G.F. Loewenstein \& J. Elster (Eds.), Choice over time (pp. 93-118). New York: Russell Sage Foundation.
Simon, J. L. (1995) Interpersonal allocation continuous with intertemporal allocation: binding commitments, pledges, and bequests. Rationality and Society, 7, 367-392.
Trope, Y., \& Lieberman, N. (2010). Construal-level theory of psychological distance. Psychological Review, 117, 440-463.

\title{
Managing our Debt: Changing Context Reduces Misunderstanding of Global Warming
}

\author{
Ben R. Newell (ben.newell@unsw.edu.au) \\ School of Psychology, University of New South Wales \\ Sydney, NSW 2052 Australia \\ Arthur Kary (art.kary@gmail.com) \\ School of Psychology, University of New South Wales \\ Sydney, NSW 2052 Australia \\ Chris Moore (christophermoore@gmail.com) \\ School of Psychology, University of New South Wales \\ Sydney, NSW 2052 Australia
}

\section*{Cleotilde Gonzalez (coty@cmu.edu)}

Department of Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, USA

\begin{abstract}
We report an experiment in which a change in the context of a stock-flow reasoning problem leads to a \(44 \%\) reduction in the use of an erroneous 'correlation heuristic' response. In its original context - a global warming scenario - the majority of participants pattern-match the output of a system to its inputs (i.e., use a correlation heuristic). In the changed context - financial debt management - the majority reason correctly that in-flows and out-flows must converge to stabilize stock. Potential applications for improving communication of climate change science are discussed.
\end{abstract}

Keywords: stock-flow reasoning; correlation heuristic; climate change

There is overwhelming agreement amongst climate scientists that the globe is warming up, due in large part to increases in the emissions of greenhouse gases (e.g., Anderegg et al., 2010). Despite this consensus in the scientific community, highly divergent opinions about the existence and implications of global warming remain entrenched in the wider community in many countries (e.g., Leiserowitz \& Smith, 2010; Leviston \& Walker, 2011).
The manifold reasons for this 'disconnect' between the science and belief range from fundamental differences in the way people 'view' the world (e.g., Kahan et al., 2012), to the pernicious attempts to manufacture doubt in the science (e.g., Oreskes \& Conway, 2010), to deficits in lay understanding of the mechanisms of global warming (e.g., Bord et al., 2000). Here we focus on this last issue and examine how a change in context might aid understanding some basic science behind how global warming 'works'.
Our experimental task focusses on the relationship between the amount of \(\mathrm{CO}_{2}\) emitted into the atmosphere, the amount of \(\mathrm{CO}_{2}\) absorbed via natural processes, and the resulting 'stock' or concentration of \(\mathrm{CO}_{2}\) that accumulates in the atmosphere. The simple principle that participants
need to appreciate to complete this task successfully is fundamental to any system that involves in-flows, outflows and an accumulating stock. Namely, that a stock will accumulate its in-flows minus its out-flows. Thus water in a bathtub will accumulate the water flowing in from the tap, minus any water flowing out through the drain. If the inflow exceeds the outflow, the tub will continue to fill up with water (e.g., Cronin et al., 2009; Sterman, 2008).
The same principle applies to the simplified climate system we consider in our experimental task: the accumulation of \(\mathrm{CO}_{2}\) in the atmosphere is determined by the in-flow (emissions) and the out-flow (absorptions). If emissions exceed absorptions \(\mathrm{CO}_{2}\) will continue to accumulate; only when emissions and absorptions converge ( \(\mathrm{CO}_{2}\) is entering and leaving the atmosphere at the same rate) will the atmospheric concentration stabilize.
A graphical representation of these relationships is shown in Figure 1 (adapted from Dutt \& Gonzalez, 2012 and Sterman \& Booth-Sweeney, 2007). The top graph shows the accumulated stock of \(\mathrm{CO}_{2}\) in the atmosphere from the period 1900 to 2100 . The stock rises steadily until the final period (between 2090 and 2100) where it stabilizes, i.e., remains constant at 950 GtC .
The bottom graph depicts the absorption of \(\mathrm{CO}_{2}\) (the green line) which is a fixed constant of \(40 \mathrm{GtC} /\) decade across the time period, and the emissions (the black line up to 2000) which steadily increases across time. The task facing participants in our experiment was to complete this emissions line for the remaining time period (2010-2100) so that the concentration depicted in the top graph was achieved; specifically, so that the concentration was stabilized by the final period. The additional lines on the bottom graph show an approximately correct (solid blue line) and a characteristically incorrect (red dashed line) response trajectory.

Your task:




REMEMBER:
The emissions line
shows the number of
gigatonnes of carbon that are added to the atmosphere each decade, due to human activities such as combustion of fossil fuels, and deforestation.

The absorptions line shows the number of gigatonnes of carbon that are removed from that are removed from decade, due to each decade, due to natural
processes such as plant growth, and CO 2 dissolving in the ocean.

The atmospheric CO2 line shows the accumulated total number of gigatonnes of CO 2 in the atmosphere, for each decade.

Next decade

Figure 1: Screenshot showing the Computer Climate Stabilization Task. The participant's task is to complete the emissions trajectory in the bottom graph so that the stabilization of atmospheric \(\mathrm{CO}_{2}\) shown in the top graph is achieved. The solid blue sketched line in the bottom graph shows a correct response trajectory in which the emissions and absorption lines converge at the point of stabilization (2100). The red dashed sketched line is a typical "correlation heuristic" response trajectory in which the emissions line mirrors the trajectory of the accumulation (i.e., continues steadily increasing) - such a response indicates a failure to understand the relationship between emission, absorption and accumulation.

The solid blue line is correct because it takes account of the principle described above. The emission value in each decade is calculated by adding the difference in stock between the current and previous time period to the absorption rate. Thus reading from the top graph in Figure 1 for the first estimate, the stock in 2000 is approximately 770 GtC and in 2010 it is approximately 800GtC. To achieve a net increase of 30 GtC in the atmosphere, 70 GtC must be emitted, 40 GtC of which is absorbed via natural processes (the green flat line in the bottom graph). Thus the correct response is 70 GtC which is approximately the value of the blue line for that decade. The red-dashed trajectory is incorrect because it fails to take account of the principle rather, the trajectory simply mirrors that of the accumulated stock, a steady increase.

Despite the apparent simplicity of stock-flow relationships (we all know how to run a bath without flooding the house), participants presented with tasks like that shown in Figure 1 overwhelmingly produce responses akin to the red-dashed line instead of the blue line (e.g., Cronin et al., 2009; Sterman \& Booth-Sweeney, 2007). The standard explanation of such erroneous responding is over-reliance on pattern-matching or use of a 'correlation heuristic' whereby participants reason that the output of a system
should "look like" (be positively correlated with) its inputs (Cronin et al., 2009; Sterman, 2008).
In the climate task such reasoning leads to people to think that if the concentration is 'going up' then so too should the emissions and thus they sketch a rising emissions trajectory that looks like the accumulation line in the top graph. Such reasoning is not confined to the climate task, however. Responses consistent with the adoption of a correlation heuristic have been observed across a range of task contexts (e.g., water tanks, bank accounts, people entering and leaving a shop) and formats (bar graphs, line graphs, tabulated numbers, and even simple text descriptions) (e.g., Cronin et al., 2009; Sterman \& Booth-Sweeney, 2007).

Couching the explanation of this erroneous behavior in terms of a heuristic begs the question of why the heuristic response is so readily adopted. Heuristics are typically invoked in an attempt to reduce the effort associated with performing a task (e.g., Shah \& Oppenheimer, 2008). One account (Kahneman \& Frederick, 2002) suggests that heuristic responding combines elements of attribute substitution and natural assessment. A participant faced with a hard question about a particular target attribute (e.g., the emissions trajectory) tends to answer a different but easier question (e.g., what does the concentration trajectory look like?). Thus the question about the target attribute is responded to by substitution of a more readily accessible
heuristic attribute. The accessibility of this heuristic attribute is determined by the extent to which its properties are naturally assessed - that is via some routinely used cognitive procedure (similarity, fluency, availability, etc.). The visual availability of the accumulation trajectory in the climate task (i.e., the functional form of the line in the top graph of Figure 1) may well contribute to it being readily 'substituted' for the correct attribute.

This visual similarity cannot, however, be the only trigger to adoption of the correlation heuristic because, as noted earlier, its use has been observed in other contexts and in non-graphical tasks (although the majority of studies have used some kind of graphical representation, e.g., Cronin et al., 2009). Thus a more likely candidate for the prevalence of such heuristic responding is simply that stock-flow tasks are not the kinds of things that we think about regularly, thus we find them difficult, and readily revert to simpler solutions. While we may be able to run a bath - thus exhibiting behavior consistent with understanding the principle of accumulation - it does not necessitate an abstract appreciation of such knowledge.

Furthermore our experience with such systems does not typically involve making sequences of decisions about the rates of change of in-flows and out-flows across time. Rather they involve a single decision within a particular time period. For example, when running a bath we know that we need to put the plug in, turn the tap on and then turn it off before the tub overflows. While this involves (periodic) monitoring of the water level, there is only one interaction with the system (turn off the tap) and the outflow is typically constant (i.e., zero, unless we have a leaky bath).
This analysis of so called 'stock-flow-failure' (Cronin et al., 2009) suggests two aspects that might be important for reducing reliance on a correlation heuristic response. First, the elements of the problem need to be sufficiently accessible or familiar that participants answer the question they are being asked rather than an 'easier' but wrong one. Second, the problem needs to be one in which people have some experience in dealing with the elements across time and preferably one that involves multiple decision points.

We reasoned that a candidate scenario that features many of these desirable attributes is financial debt management. Most of us can readily intuit that if we spend more than we earn then we will get in to debt, and that if we keep spending more than we earn that debt will continue to increase. Unlike the bathtub, debt management involves sequential, discrete monitoring of income and expenditure and is something that many of us grapple with across time ("I will pay off that credit card by the end of year!"). Figure 2 shows how the climate task depicted in Figure 1 can be readily transformed into a 'financial' debt management task.

The top graph in Figure 2 depicts the size (in dollars) of the debt incurred by an individual across a period of 21 weeks. Just like the \(\mathrm{CO}_{2}\) accumulation in Figure 1, the debt increases from just over \(\$ 600\) in Week 1 and then stabilizes at \(\$ 950\) in Week 21. The bottom graph of Figure 2 depicts the amount the person earns (the green line) - which is fixed
at \(\$ 40\) per week, and the amount the person is spending. The amount spent gradually rises to a peak at \(\$ 90\) by Week 11.


Figure 2: The Financial Task: A participant's task was to complete the 'amount spent' line in the bottom graph to ensure stabilization of the debt depicted in the top graph. A correct response required the 'amount spent' and the 'amount earned' lines to converge by 'week 21'.

The task facing the participant is to complete the 'amount spent' line to reflect the debt trajectory shown in the top graph. To do this successfully requires realizing that the 'spending' and 'earning' lines need to converge by Week 21. (For simplicity, participants were told that their debt incurred no interest.)
It is clear that the fundamental (deep) characteristics of the problems illustrated in Figure 1 and Figure 2 are the same, only the surface characteristics have changed (cf., Gonzalez \& Wong, 2011). Both tasks require understanding the relations between in-flows (emissions or spending), outflows (absorption or earnings), and stock ( \(\mathrm{CO}_{2}\) concentration or financial debt). Despite these basic similarities, we hypothesized that the financial debt scenario would trigger the correct intuition more readily (i.e., stop spending more than you earn) than the climate scenario and thus inhibit 'correlation' heuristic responding. In short, participants should be more accurate in plotting the trajectory when given the financial context than the climate context.

We tested this hypothesis in two ways: first we compared participants given only the climate task depicted in Figure 1 with participants given only the financial task in Figure 2. Second we developed another version in which
another group of participants were given the climate task but were invited to think about it as one of financial debt management. In other words, we provided explicit links between the two contexts (e.g., "You might find it helpful to think about emissions as the amount you are spending", etc.). We predicted that participants in this additional condition would perform more accurately than those given the 'pure' climate task and possibly as well as those given the 'pure' financial task. This latter prediction was based on related work on analogical encoding (Gentner, Loewenstein, \& Thompson. 2003), in which people are better at solving a problem when they are able to compare similar analogous cases prior to undertaking a target task. If the explicit links to the financial debt scenario facilitate abstraction of the basic principle (i.e., that in-flows and out-flows need to converge in order for stock to stabilize) then performance might be commensurate with the 'easier' version of the task.

We report these three conditions as a single experiment to facilitate presentation, although in reality they were run sequentially.

\section*{Experiment}

\section*{Participants}

Seventy-five undergraduate students from the University of New South Wales took part in the study in return for course credit. There were 44 females and the mean age was 19.92 ( \(\mathrm{SD}=3.40\) ). Each condition was run sequentially over a 4 month period, so participants were not randomly allocated to conditions. No participant completed more than one condition.

\section*{Experimental Design and Procedure}

Participants were given one of three versions of the basic stock-flow tasks described in the introduction (see Figures 1 and 2). For the Climate Task the graphs were adapted from ones used by Dutt and Gonzalez (2012). In each condition the main task was to complete the trajectory in the bottom graph of the display. This was done via moving on-screen slide controls that plotted the line for each time period discretely. Thus participants made 10 predictions in total. At the conclusion of these predictions participants were invited to make a second attempt (if they wished to) and could readjust any or all of the sliders before finalizing their response.

Participants in the Climate Task and Climate Financial Context condition were given some initial introductory text about climate change and global warming (adapted from Dutt and Gonzalez, 2012). Those in the Financial Task received no additional information. Participants in the Climate Financial Context condition were given additional instructions drawing explicit links between financial debt management and the climate task. These instructions appeared first on a preliminary screen and then alongside the graphs (to the left of the display shown in Figure 1) and remained there throughout the prediction attempts.

Prior to making the predictions participants in all conditions answered three comprehension questions that required reading off some numbers from both the upper and lower graphs in the display and typing in the responses.

\section*{Results}

All participants answered the comprehension questions correctly, suggesting that they were able to read the graphs accurately.

The key dependent measure of interest is the emissions/spending estimate made for each decade/week. Each participant made 10 initial estimates and then had the opportunity to change each estimate on a second attempt. Figure 3 shows the mean estimates for the 10 time periods averaged across both attempts by participants in the three conditions (very few participants changed their initial answers when given the opportunity to make adjustments). The figure also plots the correct trajectory calculated by adding the difference in stock between the current and previous time period to the out-flow (see introduction for an example).


Figure 3: The mean emission/spending estimates for each time period averaged across the two attempts made by participants (error bars are SEM). The asterisks are the correct values for each time period - see text for details. Financial refers to the debtmanagement task, Climate to the standard task and Climate Fin Context to the climate task with instructions inviting participants to consider the problem as one of debtmanagement.

Three features of the data are noteworthy: 1) participants given the 'pure' financial debt management task perform most accurately (on average) showing the correct downward trajectory; 2) those given the pure 'climate frame' display (on average) an upward trending 'correlation heuristic' response; 3) those participants given the standard climate task but with instructions to consider it as a debt-
management problem (Climate Fin Context) fall in-between the two other groups in terms of accuracy.


Figure 4: A boxplot showing the average MSD between the correct response and the estimates (average across the two attempts) in each condition. Solid line is the median response.

These general impressions are confirmed in an alternative way of graphing the data shown in Figure 4. To examine whether the differences apparent in Figure 4 were statistically reliable we used a default Bayesian t-test (Rouder, Speckman, Sun, Morey, \& Iverson, 2009). We assume, for fairness, that the null hypothesis and the alternative hypothesis are equally plausible a priori. The ttest then allows us to determine the posterior plausibility of the null hypothesis and the alternative hypothesis. We denote the posterior probability for the null hypothesis as \(p_{\text {BayesHo }}\). When, for example, \(p_{\text {BayesHO }}=.9\), this means that the plausibility for the null-hypothesis has increased from .5 to . 9.

As predicted, participants given the Financial context made significantly better (lower MSD) estimates than those given the standard climate task ( \(p_{\text {BayesH0 }}=.01\) ). In addition, participants given the Climate task with the financial context as a guide made more accurate estimates than those given the standard climate task \(\left(p_{\text {BayesHo }}=.24\right)\). For the difference between the Financial condition and the Climate Financial Context the null hypothesis is more plausible ( \(p_{\text {BayesH0 }}=.75\) ).

The data in Table 1 showing classifications of individual performance supports the interpretation provided by Figures 3 and 4 . When a correct response is coded as ensuring that in-flow and out-flows converge by the final time period, over half (52\%) of the participants given the 'pure' Financial task showed correct stabilization compared to only 2 people given the pure climate task (8\%). In addition 7 participants given the climate task with the financial context instructions achieved the correct stabilization pattern.

Table 1: Classification of Individual Responses as achieving correct stabilization of the system (i.e., in-flow and out-flow converge by the final time period).
\begin{tabular}{lcc}
\hline \multicolumn{3}{l}{ Correct Stabilization (N of Participants) } \\
\hline Experiment & Correct & Incorrect \\
\hline Financial & 13 & 12 \\
Climate & 2 & 23 \\
Climate + Fin & 7 & 18 \\
\hline
\end{tabular}

\section*{Discussion}

Our experiment sought to address reasons for the welldocumented 'stock-flow failure' observed when participants are asked to make judgments about changes to in-flow, outflow and accumulated stock across time.

In the first instance we showed a clear replication of participants' inability to 'solve' the stock-flow task when it is presented in the context of the climate system. This result dovetails neatly with those reported in the literature (e.g., Sterman \& Booth-Sweeney, 2007) and suggests that procedural differences between our and previous studies are not crucial for eliciting correlation heuristic-consistent responding.

In particular, our task differed from those used before in that we required participants only to make estimates of the emissions (in-flow) rather than both emissions and absorption rates (out-flow) (see Sterman \& Booth-Sweeney, 2007). We suspected that this change might make the task somewhat easier but it appeared not to affect performance. Likewise the fact that our 'stabilization period' was not as long as in previous studies (i.e., only one time period) did not appear to affect the failures to stabilize. (Note that the instructions stated explicitly that the emissions stabilized by 2100, and the comprehension questions suggested that participants could read this aspect of the graph.) Finally, the change to a computer interface rather than the hand-drawn sketches used previously (e.g., Dutt \& Gonzalez, 2012), appears to be a useful progression that allows a more accurate quantitative approach to analysis while not affecting the over-all pattern of responding.

In contrast to the relatively negative conclusions that can be drawn from the climate task, an optimistic (glass half full) interpretation of the Financial Task context is that (some) participants can 'do' stock-flow reasoning. Given the low-base of accurate performance in these tasks (e.g., Cronin et al., 2009), any manipulation which leads to over \(50 \%\) of the sample getting the answer (approximately) correct is newsworthy. Our working hypothesis is that the financial context helps because the familiar principle (don't spend more than you earn if you want to avoid debt) is readily intuited thereby inhibiting the correlation heuristic response (cf. Kahneman \& Frederick, 2002; Shah \& Oppenheimer, 2008). Moreover, the familiarity of projecting thoughts about debt management across time - because people often cannot pay off a debt in one go - helps understanding of the in-flow, out-flow and stock relationship.

Clearly though the familiar context on its own is not enough for all participants - just fewer than \(50 \%\) still failed to stabilize, and most of those gave responses consistent with a correlation heuristic response. As related literature has shown, the relationship between context familiarity and accuracy on these tasks is not straightforward (e.g., BoothSweeney \& Sterman, 2000; Brunstein et al., 2010; Moxnes \& Saysel, 2009) and more research is needed to identify exactly which aspects of context facilitate reasoning, and why.

This sentiment is borne out by the results of our third condition - climate with the financial context as a guide. Provision of the 'readily intuited principle' (don't spend more than you earn) was enough for just over a quarter (28\%) of the sample to understand the task, and led to significant improvements in accuracy relative to the climate task instruction alone. Thus for these participants at least, the additional explanation in terms of a familiar context seemed to improve the understanding of one aspect of the science behind global warming. However, the remaining participants could not (or did not) apply the principle correctly to the unfamiliar context.

An important question arising from this work is whether participants who perform the climate tasks accurately differ in attitudes towards taking action on global warming from those who exhibit 'stock-flow failure'. As noted in the introduction, although some authors argue that differences in 'world view' are more important than scientific understanding (e.g., Kahan et al., 2012), other studies suggest positive correlations between understanding and willingness to act (e.g., Leiserowitz \& Smith, 2010). Indeed Sterman (2008) argued that failures in stock-flow reasoning may well contribute to the tendency to take a 'wait-and-see' approach on addressing global warming.

Although we collected some data on attitudes and intentions to behave pro-environmentally, the paucity of accurate performers on the climate versions of our task, made it difficult to draw any strong conclusions in this regard. Future work will attempt to address these limitations by building on the successful context manipulations found here and by targeting more heterogeneous (non-student) populations with more divergent opinions about global warming. Such studies could provide important findings to help in getting the message about global warming both heard and heeded (cf., Newell \& Pitman, 2010).

\section*{Acknowledgments}

This support of grants (LP120100224; FT110100151) from the Australian Research Council, Department of Climate Change and Energy Efficiency and the ARC Centre of Excellence for Climate Systems Science is gratefully acknowledged.

\section*{References}

Anderegg, W.R.L, Prall, J.W., Harold, J., \& Schneider, S.H. (2010). Expert credibility in climate change. Proceedings of the National Academy of Sciences of the United States of America, 107(27), 12107-12109.

Booth Sweeney, L. \& Sterman, J.D. (2000). Bathtub dynammics: initial results of a systems thinking inventory. System Dynamics Review, 16, 249-286.
Bord, R.J., O’Connor, R.E., \& Fisher, A. (2000). In what sense does the public need to understand global climate change? Public Understanding of Science, 9, 205-218.
Brunstein, A., Gonzalez, C., \& Kanter, S. (2010). Effects of domain experience in the stock-flow failure. System Dynamics Review, 26, 347-354.
Cronin, M.A., Gonzalez, C, \& Sterman, J.D. (2009). Why don't well-educated adults understand accumulation? A challenge to researchers, educators, and citizens. Organizational Behavior and Human Decision Processes, 108, 116-130
Dutt, V. \& Gonzalez, C. (2012). Decisions from experience reduce misconceptions about climate change. Journal of Environmental Psychology, 32(1), 19-29.
Gentner, D., Loewenstein, J., \& Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. Journal of Educational Psychology, 95, 393405.

Gonzalez, C. \& Wong, H. (2011). Understanding stocks and flows through analogy. System Dynamics Review, 28, 327.

Kahan, D.M., Peters, E., Wittlin, M., Slovic, P., Ouellette, L.L., Braman, D., \& Mandel, G. (2012). The polarizing impact of science literacy and numeracy on perceived climate change risks. Nature Climate Change, 2, 732-735. Leiserowitz, A. A., Smith, N., \& Marlon, J. R. (2010). Americans' knowledge of climate change. Yale Project on Climate Change Communication.
Leviston, Z. \& Walker I.A. (2011). Baseline survey of Australian attitudes to climate change: Preliminary Report. CSIRO, Climate Adaptation Flagship.
Moxnes, E. \& Saysel, A.K. (2009). Misperceptions of global climate change: information policies. Climatic Change, 93(1-2), 15-37.
Newell, B.R., \& Pitman, A.J. (2010). The psychology of global warming: Improving the fit between the science and the message. Bulletin of the American Meteorological Society, 91, 1003-1014.
Oreskes, N., \& Conway, E. M. (2010). Merchants of doubt. London: Bloomsbury Publishing
Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., \& Iverson, G. (2009). Bayesian \(t\) tests for accepting and rejecting the null hypothesis. Psychonomic bulletin \& review, 16(2), 225-237.
Shah, A. K., \& Oppenheimer, D. M. (2008). Heuristics made easy: an effort-reduction framework. Psychological Bulletin, 134(2), 207.
Sterman, J.D. (2008). Risk communication on climate: Mental models and mass balance. Science, 322, 532-533.
Sterman, J. D., \& Booth-Sweeney, L.. (2007). Understanding public complacency about climate change: adults' mental models of climate change violate conservation of matter. Climatic Change, 80(3-4), 213238.

\title{
Non-parametric estimation of the individual's utility map
}

\author{
Takao Noguchi (t.noguchi@ warwick.ac.uk) \\ Adam Sanborn (a.n.sanborn@warwick.ac.uk) \\ Neil Stewart (neil.stewart@warwick.ac.uk) \\ Department of Psychology, University of Warwick
}

\begin{abstract}
Models of risky choice have attracted much attention in behavioural economics. Previous research has repeatedly demonstrated that individuals' choices are not well explained by expected utility theory, and a number of alternative models have been examined using carefully selected sets of choice alternatives. The model performance however, can depend on which choice alternatives are being tested. Here we develop a non-parametric method for estimating the utility map over the wide range of choice alternatives. The estimated maps are compared against the three of the most well-known models of risky choice: expected utility theory, cumulative prospect theory, and the transfer of attention exchange model. Model comparison indicates that cumulative prospect theory provides a better prediction of individuals' choices, but the estimated maps show that the overall shape of utility map is different from what the model predicts.
\end{abstract}

Keywords: decision making; risky choice; utility; MCMC with People; expected utility; cumulative prospect theory; transfer of attention exchange

\section*{Background}

Understanding how people trade off risk and reward is a fundamental goal of behavioural economics. The most common approach to modelling how people make decisions between risky alternatives is based on the idea of utility: individuals integrate the probability of reward with the utility of the reward to produce an expected utility that describes how well the alternative is preferred. The alternative with the highest utility is most often chosen.
The normative calculation of utility that maximizes longterm gain is to multiply the probability with the utility of the associated outcome and to derive the expected utility. For an illustration, suppose an individual is considering a choice alternative with three possible outcomes: \(£ 20, £ 10\), and \(£ 0\). This particular alternative has a \(20 \%\) probability for \(£ 20\), \(40 \%\) for \(£ 10\), and \(40 \%\) for \(£ 0\). Then, the expected utility is \(20 \% \times u(£ 20)+40 \% \times u(£ 10)+40 \% \times u(£ 0)\), where \(u\) is the function to map the monetary value to the utility.

However, previous research has demonstrated that an individual's choice frequently deviates from the predictions of expected utility theory (for review, Schoemaker, 1982). To explain the deviations, descriptive models of how risk and reward are integrated have been developed (for review, Starmer, 2000). A common and useful way to visualize the predictions of these models is to look at the indifference lines, which connect choice alternatives of equal utility, over a MarschakMachina probability triangle (Marschak, 1950; Machina, 1982). The probability triangle is a two-dimensional space which maps alternatives with varying probabilities for the same set of three potential outcomes. Throughout this pa-
per, we use \(£ 20, £ 10\), and \(£ 0\) as the potential outcomes from a choice alternative.

Figure 1 displays the predicted utility maps from three of the most well-known models of risky choice: expected utility theory, cumulative prospect theory (Tversky \& Kahneman, 1992) and transfer of attention exchange (TAX) model (Birnbaum, 2008). In the probability triangle, the probability of attaining the best outcome ( \(£ 20\) ) is represented in the vertical axis, and the probability of the worst outcome ( \(£ 0\) ) is represented in the horizontal axis. The probability for the other outcome ( \(£ 10\) ) is represented as the distance from the diagonal boundary along the horizontal axis. The diagonal boundary ensures that the sum of the probabilities for \(£ 20\), \(£ 10\) and \(£ 0\) does not exceed 1 .

The red area in the triangles indicates the area of high utility, and the blue area is the area of low utility. Also, the coloured lines connect the alternatives of equal utility. These indifference lines highlight the differences between expected utility theory and the two descriptive models. Expected utility theory predicts indifference lines that are parallel and straight. Both cumulative prospect theory and the TAX model predict concave lines in the top corner of the triangle but convex lines in the lower right corner.

The usual experimental practice is to investigate choices in the regions of the triangle where models most differ from each other (e.g., Wu \& Gonzalez, 1998). When the models are tested in this way, the "best" model may not predict choices away from the diagnostic regions well. For instance, Harless (1992) suggests that cumulative prospect theory outperforms expected utility theory only at the edges of the triangle. Thus, the model comparison could benefit from being tested on the whole area of triangle. One way is to estimate the utility map over the whole triangle and compare the estimated map against the model prediction. However to the best of our knowledge, the available estimation methods impose an assumption on how subjective value and probabilities are integrated (e.g., Abdellaoui, 2000), which could favour the model with the identical assumption.

To this end, we develop a non-parametric method to estimate entire utility maps, an extension of Markov chain Monte Carlo (MCMC) with People (Sanborn, Griffiths, \& Shiffrin, 2010). We have modified MCMC with People to investigate the regions of the probability triangle where the choice alternatives are less preferred. The new method is tested in a simulation to show that it can deliver useful results within a reasonable number of trials. We then estimate utility maps from human. Finally, we discuss the results and future applications for this approach.

(A) Expected utility theory with the identity value function

(B) Cumulative prospect theory with parameters \(\alpha=0.88\) and \(\gamma=0.52\)

(C) The transfer of attention exchange model with parameters \(\beta=1, \gamma=0.7\), and \(\delta=1\)

Figure 1: Theoretical predictions

\section*{Markov chain Monte Carlo with People}

Markov chain Monte Carlo (MCMC) is a common method for drawing samples from a distribution. It has been widely used to perform probabilistic inference especially when solving the exact function of interest is computationally difficult (Neal, 1993).

MCMC begins in a start state \(z\). A sample \(z^{\prime}\) is first drawn from the proposal distribution \(q\), and then \(z^{\prime}\) is evaluated with the function of interest, \(\pi\), to determine whether to accept \(z^{\prime}\) as a new state or discard it and retain the current state \(z\). The sequence of accepted samples forms a Markov chain, and after this Markov chain converges, accepted samples can be regarded as samples from the \(\pi\) distribution. To ensure that the Markov chain converges to \(\pi\), it is sufficient to satisfy detailed balance (as well as ergodicity):
\[
\begin{equation*}
\pi(z) q\left(z^{\prime} \mid z\right) A\left(z^{\prime}, z\right)=\pi\left(z^{\prime}\right) q\left(z \mid z^{\prime}\right) A\left(z, z^{\prime}\right) \tag{1}
\end{equation*}
\]
where \(q\left(z^{\prime} \mid z\right)\) is the probability of drawing \(z^{\prime}\) when the current state is \(z\) and \(A\left(z^{\prime}, z\right)\) is the probability of accepting proposal \(z^{\prime}\) over the current state \(z\).
Throughout the paper, we assume a symmetric distribution for \(q, q\left(z^{\prime} \mid z\right)=q\left(z \mid z^{\prime}\right)\), so Equation 1 becomes
\[
\begin{equation*}
\pi(z) A\left(z^{\prime}, z\right)=\pi\left(z^{\prime}\right) A\left(z, z^{\prime}\right) \tag{2}
\end{equation*}
\]

Detailed balance can be satisfied by carefully designing the acceptance function \(A\). The most commonly used function is the Metropolis acceptance function (Metropolis, Rosenbluth, Rosenbluth, Teller, \& Teller, 1953), but the Boltzmann acceptance function (Flinn \& McManus, 1961) is of interest here:
\[
A\left(z^{\prime}, z\right)=\frac{\pi\left(z^{\prime}\right)}{\pi(z)+\pi\left(z^{\prime}\right)}
\]

If an individual is asked to make a choice between alternatives \(z^{\prime}\) and \(z\), then the Boltzmann acceptance function can
model that individual's choice. This is because the Boltzmann function is equivalent to Luce's choice rule (Luce, 1959), which has been frequently used to model risky choice (e.g., Blavatskyy \& Pogrebna, 2010). As a result, by sequentially presenting pairs of choice alternatives to an individual (where the new alternative \(z^{\prime}\) is selected by the computer), the collection of choice alternatives chosen by the individual can be treated as samples from the probability distribution whose density is proportional to the individual's utility (Sanborn et al., 2010).

\section*{Extending MCMC with People}

However, sampling from the individual's utility distribution does not necessarily serve to estimate the shape of the utility map: pilot work confirms that all of the samples will be concentrated around the most favourable alternative ( \(100 \%\) probability of \(£ 20\) in the triangle), and that it would take a very large number of trials to explore the rest of the utility map. To enable the reasonable estimation of the utility map, the stationary distribution needs to be more diffused, so that the Markov chain travels better around the triangular space.
For this purpose, we implement a latent agent in the experimental program. This agent makes an independent choice between the same alternatives as the participant, and only when the agent and the participant both select the new choice alternative does the new alternative become the new state. Otherwise, the current state remains the same and another alternative is generated from the proposal distribution.

When the agent is implemented in this way, the acceptance function becomes a joint function of the participant's and the agent's choices. Specifically, the acceptance function is defined as
\[
A^{*}\left(z^{\prime}, z\right)=\frac{f\left(z^{\prime}\right)}{f(z)+f\left(z^{\prime}\right)} \frac{g\left(z^{\prime}\right)}{g(z)+g\left(z^{\prime}\right)}
\]
where \(f\) is the utility function for the participant and \(g\) is the
agent's utility function. Here, both the participant and the agent follow the Boltzmann acceptance function. Then Equation 2 becomes
\[
f(z) g(z) A^{*}\left(z^{\prime}, z\right)=f\left(z^{\prime}\right) g\left(z^{\prime}\right) A^{*}\left(z, z^{\prime}\right)
\]

With the implementation of the agent, the trajectory of the Markov states depends on both the participant's and the agent's choices. If the agent's utility is the lowest at the top corner of the triangle, the Markov chain would be pushed away from that region. With this extended method, the stationary distribution of the Markov chain is the joint utility function of the participant and the agent, \(f g\). The participant's utility map can subsequently be recovered by dividing the joint utility by the latent agent's known utility, \(g\).

\section*{Simulation}

To demonstrate that the developed method can estimate a participant's utility map within a reasonable number of trials, we conducted a simulation.

\section*{Method}

The simulation used two of the utility functions in Figure 1: the latent agent's utility function, \(g\), was set to the inverse of expected utility theory, and the simulated participant's function, \(f\), was cumulative prospect theory:
\[
g=\frac{1}{20 \times p(£ 20)+10 \times p(£ 10)}
\]
and
\(f=20^{\alpha} \times w(p(£ 20))+10^{\alpha} \times(w(p(£ 20)+p(£ 10))-w(£ 20))\),
where \(p(£ 20)\) is the probability of attaining \(£ 20\), and \(w(p)=\) \(\frac{p^{\gamma}}{\left(p^{\gamma}+(1-p)^{\gamma}\right)^{1 / \gamma}}\). The parameter values for \(\alpha\) and \(\gamma\) were 0.88 and 0.52 , respectively. The proposal distribution, \(q\), was uniform over the triangular space. The possible outcomes were fixed to be \(£ 20, £ 10\) and \(£ 0\), and hence, the agent and the simulated participant repeatedly made choices between two alternatives with varying probabilities for fixed outcomes: e.g., a choice between an alternative with a \(30 \%\) probability for \(£ 20\), \(40 \%\) for \(£ 10\) and \(30 \%\) for \(£ 0\), and another alternative with a \(10 \%\) probability for \(£ 20,60 \%\) for \(£ 10\) and \(30 \%\) for \(£ 0\).

With the above specifications, a choice trial was simulated as follows. First, the agent used the \(g\) function to evaluate each alternative and used the Boltzmann acceptance function to select between the current state and the proposed alternative. If the agent preferred the current state over the proposed alternative, another alternative was sampled from the proposal distribution. If the agent chose the new alternative over the current state, the simulated participant used the \(f\) function to make a choice between the same two alternatives.

Although the agent and the simulated participant could have made a choice at the same time over the same two alternatives, we had the agent decide first: if the agent does not select the new alternative, the previous state remains the state
regardless of the choice the participant makes. This reduces the number of choices the participant must make.

Each simulation consisted of three chains: one chain started with the Markov state of \(60 \%\) of \(£ 20,20 \%\) of \(£ 10\) and \(20 \%\) of \(£ 0\). Another chain started with the state of \(20 \%\) of \(£ 20,60 \%\) of \(£ 10\) and \(20 \%\) of \(£ 0\). The final chain started with \(20 \%\) of \(£ 20,20 \%\) of \(£ 10\) and \(60 \%\) of \(£ 0\).

\section*{Results and Discussion}

The first 100 trials were considered to be trials before convergence of the Markov chain (burn-in period) and were discarded from each chain. The remaining samples from the three chains were pooled and smoothed by kernel density estimation. Because of the triangular boundary of the estimation space, it is actually quite difficult to produce unbiased indifference lines. We chose to use a Dirichlet kernel, an extension of the Beta kernel (Chen, 1999) to the triangular space, because it produced less bias than the other alternatives we investigated. The Dirichlet kernel is defined as
\[
\hat{f}(x) g(x)=\sum_{i} \operatorname{Dir}\left(z_{i} \mid \alpha_{1}, \alpha_{2}, \alpha_{3}\right)
\]
where \(z_{i}\) is the \(i\) th state in the Markov chain, \(x\) is a vector of probabilities for three outcomes, and \(\alpha_{j}\) is \(x_{j} / \min \left(h, x_{j}, 1-\right.\) \(x_{j}\) ). The kernel width \(h\) was set to 0.09 . This smoothed joint distribution is then divided by \(g\) to derive the estimation \(\hat{f}\).
To assess the similarity between \(f\) and \(\hat{f}\), we computed Kullback-Leibler (KL; denoted as \(\operatorname{KL}(f \| \hat{f})\) ) divergence (Kullback \& Leibler, 1951), which measures how much information is lost in the estimation process.
The KL divergences for different sample sizes are plotted in the left panel of Figure 3. This figure illustrates that the estimation shows the increasingly smaller divergence within the first few hundred trials. The estimation becomes reasonably accurate on average after 700-800 trials.

The middle and right panels of Figure 3 display the estimations after 1,000 trials. The estimation with the smallest KL divergence among the 10 simulation runs is in the middle panel, and the right panel show the estimation with the largest KL divergence. Both panels show the key property of cumulative prospect theory: the indifference lines show fanningout property from the lower left corner toward the diagonal boundary.
Thus, the simulation demonstrated that the proposed method with the Dirichlet kernel density estimation can recover the key characteristic of the utility map using a reasonable number of samples.

\section*{Experiment}

\section*{Method}

Participant Ten participants were recruited through the subject panel at the University of Warwick. One participant did not complete the experiment, leaving nine (five male and four female) participants. Their age ranged from 19 to 30 with a mean of 22.9.


Figure 2: The KL divergences between \(f\) and \(\hat{f}\) for various numbers of trials. The solid line represents the mean measurement of the 10 simulation runs, and the dotted lines are maximum and minimum values archived in the simulations.


Figure 3: Estimation of cumulative prospect theory with 1,000 trials

Procedure The experimental procedure closely followed that of the simulation. The agent's utility function, \(g\), was set to the inverse of expected utility theory raised to the power of 8 , and the proposal distribution, \(q\), was uniform over the triangular space. The possible outcomes were fixed to be \(£ 20, £ 10\) and \(£ 0\), and hence, the agent and the participant repeatedly made choices between two alternatives with varying probabilities for fixed outcomes.

In each trial, the agent made a decision first, and a new alternative was drawn until the agent chose the new alternative. Three chains with the same start states as the simulation were run interleaved until participants had made 1,000 choices per chain. In addition, 50 catch trials were inserted into the experiment, so that we could assess whether participants were engaged in the task. In each catch trial, one alternative had larger probabilities for both \(£ 20\) and \(£ 10\). If a participant was not engaged with the task and randomly making choices, it is expected that he or she would occasionally not select the non-dominant alternative.

The experiment presented a choice alternative as a pie chart with three slices. Each slice represented one possible outcome, and the size of the slice was proportional to the probability of the outcome. Participants were forced to log out from the online experiment and take a break after spending one hour on it. After the minimum break of three hours, participants were allowed to \(\log\) in again and resume the experiment.

The choices participant made were incentivized: we invited participants to the lab when participants completed the experiment. At the lab, we randomly selected one trial from the experiment and played the selected alternative for real. Participants were paid what they earned from the play.

\section*{Results and Discussion}

All the nine participants selected the dominant alternative in all of the catch trials, which was evidence that all participants understood and were engaged in the task.

Utility maps were estimated as in the simulation study. All participants show a sharp peak at the top corner of the triangle in the estimated maps. The sharp peak makes it difficult to see the shape of the map, and thus for illustration purposes, we spaced out the indifference lines by taking the natural logarithm of the estimation. As a result, differences in small utilities are exaggerated, but the shapes of the indifference lines are not affected. The resulting maps are displayed in Figure 4. Each panel in the figure corresponds to one participant's map.

The estimated maps show the steep indifference lines, especially where the probability of \(£ 0\) is small. The steep lines indicate aversion to the worst outcome (c.f., Tversky \& Kahneman, 1992; Birnbaum, 2008), where the increment in probability for the worst outcome needs to be compensated with a larger increment in probability for the most desirable outcome. The steepness tends to be lessened near the lower right corner of the triangle. As a result, for Participants A, D and H in particular, the indifference lines show the fanning-out property. The fanning-out suggests that participants more willingly accept an increment in probability for the worst outcome when the probability is already large. The fanning-out is consistent with the prediction from cumulative prospect theory and the transfer of attention exchange (TAX) model.
The estimated maps also show the convex indifference lines throughout the triangle. The convexity makes the estimated maps appear rather different from the predicted utility maps from cumulative prospect theory and the TAX model, which expect the concavity toward the top corner of the triangle (Figure 1).

To quantitatively assess the model performance, we fit the models to the individuals' choices by maximizing the likeli-


Figure 4: \(\ln (\hat{f})\)
hoods. When fitting the model, we used the power law utility function for expected utility theory: \(u(s)=s^{\alpha}\). The range of parameter values are restricted to be between 0 and 1 for all the parameters. Also, each model included one additional parameter to raise the predicted utility. This exponent controls how steep the peak is toward the most favourable alternative.

The value for this exponent parameter is restricted to be nonnegative.

Bayesian information criteria (BIC) indicates that the choices are best predicted by cumulative prospect theory for seven out of nine participants (Panels A through G). The TAX model achieves smallest BIC for one of the remaining partic-
ipants (Panel H), and the expected utility theory has smallest BIC for the other participant (Panel I).

\section*{General Discussion}

Previous research has demonstrated that individuals' choices deviate predictions from expected utility theory, and variety of descriptive models have been proposed. However, the deviation from expected utility theory has often been studied with relatively limited range of choice alternatives. The present study developed the non-parametric method to estimate utility maps over the whole probability triangle.

The curvature of the indifference lines in the estimated maps implies differences to the predictions from the existing models: The lines tend to be convex where concavity is expected. Even though cumulative prospect theory (CPT) does not predict this curvature, CPT provides a better fit to the choice data than expected utility theory or the attention exchange model does for the majority of participants. Thus, a new model could explain the choices better than CPT, if the new model produces a utility map similar to the estimated maps.

In developing such a model, it is useful to identify choice alternatives where the CPT prediction differs from the individuals' choice behaviour. To this end, the estimation method that we have developed can be further extended. As the developed method lets the MCMC chain converge to the joint distribution of the individual's and the agent's utility, manipulation of agent's utility function can reveal interesting joint distributions. For instance, by setting the latent agent's utility to the inverse of the CPT prediction, the MCMC chain converges to the distribution whose density is proportional to the individual's utility divided by the CPT prediction. The condensed area in this joint utility distribution is where the CPT prediction is smaller than the individual's utility (i.e., the area where CPT underpredicts the utility), and the thin area is where the CPT prediction is larger than the individual's utility (i.e., the area where CPT overpredicts the utility).

To conclude, we have developed the method for estimating the utility map. The developed method can be further leveraged in future study.

\section*{References}

Abdellaoui, M. (2000). Parameter-free eliciation of utility and probability weighting functions. Management Science, 46, 1497-1512.
Birnbaum, M. (2008). New paradoxes of risky decision making. Psychological Review, 115, 463-501.
Blavatskyy, P. R., \& Pogrebna, G. (2010). Models of stochastic choice and decision theories: why both are important for analyzing decisions. Journal of Applied Econometrics, 986, 963-986.
Chen, S. X. (1999). Beta kernel estimators for density functions. Computational Statistics \& Data Analysis, 31, 131145.

Flinn, P. A., \& McManus, G. M. (1961). Monte Carlo calculation of the order-disorder transformation in the bodycentered cubic lattice. Physical Review, 124, 54-59.
Harless, D. (1992). Predictions about indifference curves inside the unit triangle. Journal of Economic Behavior and Organization, 18, 391-414.
Kullback, S., \& Leibler, P. C. (1951). On information and sufficiency. Annals of Mathematical Statistics, 22, 79-86.
Luce, R. D. (1959). Individual choice behavior: A theoretical analysis. New York: John Wiley \& Sons, Inc.
Machina, M. J. (1982). "Expected Utility" analysis without the independence axiom. Econometrica, 50, 277-323.
Marschak, J. (1950). Rational behavior, uncertain prospects, and measurable utility. Econometrica, 18, 111-141.
Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., \& Teller, E. (1953). Equation of state calculations by fast computing machines. Journal of Chemical Physics, 21, 1087-1092.
Neal, R. (1993). Probabilistic inference using Markov chain Monte Carlo methods (Tech. Rep. No. CRG-TR-93-1). Department of Compute Science, University of Toronto.
Sanborn, A. N., Griffiths, T. L., \& Shiffrin, R. M. (2010). Uncovering mental representations with Markov chain Monte Carlo. Cognitive Psychology, 60, 63-106.
Schoemaker, P. J. H. (1982). The expected utility model: Its variants, purposes, evidence and limitations. Journal of Economic Literature, 20, 529-563.
Starmer, C. (2000). Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. Journal of Economic Literature, 38, 332-382.
Tversky, A., \& Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and Uncertainty, 5, 297-323.
Wu, G., \& Gonzalez, R. (1998). Common consequence conditions in decision making under risk. Journal of Risk and Uncertainty, 16, 115-139.

\title{
\(k\)-Nearest Neighbor Classification Algorithm for Multiple Choice Sequential Sampling
}

\author{
Yung-Kyun Noh (nohyung@snu.ac.kr) \\ Frank Chongwoo Park (fcp@snu.ac.kr) \\ School of Mechanical and Aerospace Engineering, Seoul National University \\ Seoul 151-744, Korea \\ Daniel D. Lee (ddlee@seas.upenn.edu) \\ Department of Electrical and Systems Engineering, University of Pennsylvania \\ Philadelphia, PA 19104, USA
}

\begin{abstract}
Decision making from sequential sampling, especially when more than two alternative choices are possible, requires appropriate stopping criteria to maximize accuracy under time constraints. Optimal conditions for stopping have previously been investigated for modeling human decision making processes. In this work, we show how the \(k\)-nearest neighbor classification algorithm in machine learning can be utilized as a mathematical framework to derive a variety of novel sequential sampling models. We interpret these nearest neighbor models in the context of diffusion decision making (DDM) methods. We compare these nearest neighbor methods to exemplar-based models and accumulator models, such as Race and LCA. Computational experiments show that the new models demonstrate significantly higher accuracy given equivalent time constraints.
\end{abstract}

Keywords: sequential sampling; decision making; diffusion decision making model; \(k\)-nearest neighbor classification; evidence; sequential probability ratio test

\section*{Introduction}

Whenever a faster decision is required to save time and resources, the decision making process should focus on choosing whether to proceed with a decision in light of the given information or to postpone the decision in order to collect more information for a higher confidence level. In many previous and recent psychology works, various computational models have been introduced seeking to explain the speed-accuracy tradeoff and to understand the decision making process in humans. However, apart from the understanding of individual models, there has been little systematic way of understanding these models in one mathematically unified framework. Moreover, multiple-choice problems were not discussed intensively in any of the methods.

The optimality in decision making with sequential sampling is discussed with the optimality in speed-accuracy tradeoff. In other words, the objective of the present work is to seek the fastest decision with the same average accuracy or the maximum accuracy if the same average decision time is used. Sequential sampling methods such as Race (Smith \& Vickers, 1988; Vickers, 1970), diffusion decision making (DDM) (Ratcliff, 1978; Ratcliff \& Rouder, 2000; Shadlen, Hanks, Churchland, Kiani, \& Yang, 2006; Ratcliff \& Mckoon, 2008), and leaky competing accumulator (LCA) (Usher \& McClelland, 2001; Bogacz, Usher, Zhang, \& McClelland, 2007) are all interested in explaining this optimality in the
speed-accuracy tradeoff. In these methods, one or more variables are commonly introduced for accumulating sampled information, and a criterion is used to determine whether to continue collecting more information or to make a decision with given information. Here, we propose a common mathematical framework combining these methods and providing a systematic explanation for understanding different methods.

Our framework combining sequential sampling methods is the \(k\)-nearest neighbor ( NN ) classification in machine learning. The sequential sampling situation with multiple choices is explained as the multiway \(k\)-NN classification from the theoretical analysis on \(k\)-NNs in the asymptotic situation. Due to this connection, we can interpret all different types of sequential sampling methods as different methods of choosing \(k\) adaptively in \(k\)-NN classification. By further analyzing the strategy of choosing \(k\) in \(k\)-NN classification using the Sequential Probability Ratio Test (SPRT) (Wald \& Wolfowitz, 1948) and Bayesian inference, we can obtain five different accumulating variable and stopping criteria for optimal tradeoff. Interestingly, all these five optimal methods are interpreted as different kinds of DDM strategies.

Our work is directly applied to a recently reported neuroscientific decision making mechanism. The proposed mechanism considers an output neuron which sends out a decision result. By collecting Poisson spike trains from different neurons, the output neuron makes a decision about which neuron gives Poisson spikes at the highest rate (Shadlen \& Newsome, 1998; Ma, Beck, Latham, \& Pouget, 2006; Beck et al., 2008; Zhang \& Bogacz, 2010). The output neuron can achieve optimality by using our proposed strategies.

The proposed method can be compared with traditional exemplar models which explain memory retrieval using similarity weighted voting based on stored exemplars. Our work is different from this line of research by using majority voting of adaptively chosen \(k\) number of NNs. We discuss the advantages and disadvantages of our method when it is applied to the memory retrieval problem.

The rest of the paper is organized as follows. We introduce the sequential sampling problem in Section 2 especially from the point of view of multiple-choice. In Section 3, we introduce problems to which sequential sampling methods can be applied, and we show how \(k\)-NN classification can be natu-
rally introduced as a common framework for explaining these problems. In Section 4, we derive the examples of two- and multiple-choice evidence for DDM in light of \(k\)-NN classification. After we explain the relationship between our method and other exemplar methods in Section 5, we present simulation experiments in Section 6. Finally, we conclude with discussion in Section 7.

\section*{Computational Methods in Sequential Sampling Problems}

Sequential sampling methods consider decision making using incoming information over time. With unlimited time, the decision can be made late enough to increase the expected accuracy. However, if the decision should be made as soon as possible, there is a trade-off between the speed and accuracy of the decision. In order to address this tension, decision making strategies introduce criteria to determine whether or not to make a decision at a certain time.

Accumulator Model: One simple method of determining whether the accumulated information has reached a certain level of confidence is the accumulator model. This model considers one variable for each choice and accumulates information separately in favor of each choice. Once one of the accumulating variable reaches a predefined threshold, the decision is made immediately thereafter.

This simple model with no interaction between different choices is known as suboptimal. This method can be compared with the DDM strategy in the next section, where the accuracy of the accumulator model is always less than the accuracy of the DDM model (Zhang \& Bogacz, 2010). This model of doing race between accumulators is also called the Race model.

Diffusion Decision Making (DDM) Model: In this model with two choices, one variable is introduced to collect information and diffuse toward one of the choice. This variable, also known as the evidence, represents the bias in the preference of accumulated information toward a choice. Finally, once the evidence reaches a pre-defined level of any choice, it stops diffusing and selects the choice.

A canonical method of determining the evidence variable and stopping criterion uses the sequential probability ratio test (SPRT) (Wald \& Wolfowitz, 1948; Dragalin, Tertakovsky, \& Veeravalli, 1999; Zhang \& Bogacz, 2010). Previous work using this test has considered two incoming Poisson signals aiming to determine the signal with the higher Poisson rate from the accumulation of signals. In this case, the diffusing evidence is just the difference in the number of signals within a certain time, and the decision is made once this difference exceeds a threshold. This method is known to be optimal among sequential sampling methods such as Race and LCA (Bogacz, Brown, Moehlis, Holmes, \& Cohen, 2006; Bogacz et al., 2007).

Leaky Competing Accumulator (LCA): LCA uses one variable for each choice similar to the accumulator model,
but it considers the interaction between the variables. LCA considers the dynamics of activation with the decay of the activation as well as the inhibitory interaction between activation variables. This LCA dynamics is very flexible in that the strategy can be either similar to Race or DDM as its special case, but the maximum performance is known to be that of DDM (Bogacz et al., 2007).

\section*{Multiple-Choice Extension}

Among the aforementioned sequential sampling models, the multiple-choice extension of the Race and LCA models is straightforward, by just increasing the number of accumulating variables. However, the extension of DDM is more complex. Fortunately, the Multiple SPRT (MSPRT) method was previously developed by extending the SPRT method using the number of signals (Dragalin et al., 1999; Zhang \& Bogacz, 2010). In addition to this MSPRT result, we also provide different criteria for multiple-choice DDM using derivations from other approaches of MSPRT and Bayesian inference. Our result provides an evidence diffusing in a \(C-1\) dimensional space for a \(C\) alternative choice problem.

\section*{Sequential Sampling Problems}

Decision making problem using sequential sampling can be found in many examples. Here we introduce two exemplary problems. One example can be found in neuronal decision making as in the left figure of Fig. 1. When an output neuron tries to make a decision as to whether one incoming signal has a higher Poisson rate than the other has, the output neuron can collect signals until the accumulated information reaches a certain level.

Another example can be found in a Bayes classification problem where we only have data generated from unknown underlying density functions. Bayes classification selects the class having the highest underlying density, but the classifier in this case cannot directly access the underlying density information. A surrogate method of determining the class of highest density is through \(k\)-NN classification. By collecting more nearest neighbors, the confidence of choosing a class of the highest density is expected to increase to a targeted level.

Here, we show that the two problems are in fact exactly the same by explaining several theoretical results on \(k-\mathrm{NN}\) classification in the asymptotic situation:

Majority Voting Rule in \(k\)-Nearest Neighbor Classification: When there are \(N\) number of training data with labels, \(\mathcal{D}=\left\{\mathbf{x}_{i}, y_{i}\right\}_{i=1}^{N}\), where each datum \(\mathbf{x}_{i} \in \mathbb{R}^{D}\) is represented as a \(D\)-dimensional vector, and the label has one of \(C\) labels, \(y_{i} \in\{1, \ldots, C\}, k-\mathrm{NN}\) classification assigns class \(y\) to a classunknown datum \(\mathbf{x}\) according to the majority voting with \(k\) labels of nearest data in \(\mathcal{D}\) :
\[
\begin{equation*}
y=\arg \max _{c} \sum_{i=1}^{k} \mathbb{I}\left(y_{n(i)}=c\right) \tag{1}
\end{equation*}
\]
with nearest neighbor index \(n(i), i=1, \ldots, k\). The theoretical study of this majority voting strategy originates from Cover


Figure 1: Diffusion decision making (DDM) in neurons determines the input neuron with the higher firing rate, and its analogous \(k\)-NN classification determines the larger class-conditional density function
and Hart (Cover \& Hart, 1967; Cover, 1967), and in their work, once there are enough data, the expected error monotonically and asymptotically decreases with the number of \(k\) :
\[
\begin{equation*}
E(k=1) \geq E(k=2) \geq \ldots \geq E(k=\infty) \tag{2}
\end{equation*}
\]
for \(k \ll N\). This continuous decrease of error will encourage the use of more nearest neighbors, which explains the tradeoff between the number of nearest neighbors \(k\) and the classification accuracy.

Distance Comparison Rule of \(k^{\prime}\)-th NNs of Each Class: If we consider a strategy of comparing two \(k^{\prime}\)-th NN in class 1 and class 2 , we can easily prove that this strategy is equivalent to the majority voting rule with \(k=2 k^{\prime}-1\) NNs.

Proof: Consider a comparison between \(k^{\prime}\)-th NN of class 1 and another \(k^{\prime}-t h N N\) of class 2. If \(k^{\prime}\)-th \(N N\) in class 1 is closer than \(k^{\prime}-t h N N\) in class 2 , then we can say that the \(k^{\prime}-t h\) \(N N\) in class 2 can never be included in the closest \(\left(2 k^{\prime}-1\right)\) \(N N s\), because at least \(k^{\prime}\) number of NNs in class 1 and \(\left(k^{\prime}-1\right)\) number of NNs in class 2 have less distance than the \(k^{\prime}\)-th NN in class 2. Therefore, comparing strategies of \(k^{\prime}-t h N N\) in each class is the same as majority voting with \(\left(2 k^{\prime}-1\right)\) nearest neighbors.

Therefore, the monotonic increase of accuracy is also satisfied with the increase of \(k^{\prime}\).

Two Sequential Sampling Methods in \(k\)-NN Classification: From the monotonic increase of the accuracy with the increase of \(k\) (or \(k^{\prime}\) ), we can make two different sequential sampling methods showing the speed-accuracy tradeoff.

First, we can consider the majority voting strategy using number of NNs within a certain distance from the testing point. If we do not have enough accuracy with the current distance, we can increase it to use more resources. Another example can be designed by considering the distance to the same \(k^{\prime}\)-th NN in each class and making a decision by comparing the distances.

The first design corresponds to the sequential sampling
with continuous time and discrete accumulation of information, because the accumulation variable is the function of the number of NNs. In contrast, the second design uses the discrete time and continuous accumulation of information.

Distribution of the distances: Now, we show that \(k\)-NN classification is in fact equivalent to sequential sampling for determining the signal with the highest Poisson rate.

A recent study discussed the distribution of the distance to the NNs when there are enough data (Leonenko, Pronzato, \& Savani, 2008). Instead of directly dealing with the distribution of distance, they changed the random variable to \(u=N V\), with volume \(V\) of \(D\) dimensional hypersphere having the distance to the \(k\)-th NN as a diameter multiplied by the number of data \(N\). Then the distribution of samples approaches the Erlang density function:
\[
\begin{equation*}
\rho(u \mid \lambda)=\frac{\lambda^{k}}{\Gamma(k)} \exp (-\lambda u) u^{k-1} \tag{3}
\end{equation*}
\]
with a parameter \(\lambda\), which is the probability density \(p(\mathbf{x})\) at \(\mathbf{x} \in \mathbb{R}^{D}\). Moreover, this special Erlang function implies the Poisson distribution of the number of NNs \(k\) within a specified volume of the hypersphere (Wasserman, 2003):
\[
\begin{equation*}
\rho(k \mid \lambda)=\frac{\lambda^{k}}{\Gamma(k+1)} \exp (-\lambda) \tag{4}
\end{equation*}
\]

This equation shows that the number of NNs within a growing hypersphere at a constant rate in volume is a Poisson process.

Comparing this Poisson process interpretation with the aforementioned neuronal decision making, we can draw several corresponding analogies. The firing rate of the Poisson signal corresponds to the underlying density function in \(k-\mathrm{NN}\) classification, the number of spikes corresponds to the number of NNs, the time within which spikes are counted corresponds to the volume of the hypersphere within which we count NNs, and as a consequence, determining a choice with the highest firing rate corresponds to the problem of deter-
mining the class with the highest underlying density function, which is also known as the Bayes classification.

The correspondence shows that these two very well-known methods from different disciplines can share optimal strategies as well as theoretical knowledge. However, the study of a method in one field is rarely investigated in another; the strength of the correspondence suggests that whenever a good strategy is found for DDM, a corresponding strategy should be examined for machine learning. Conversely, when a new strategy is provided for the \(k\)-NN method, its relevance to psychology should also be investigated.

\section*{Derivation of Stopping Criteria}

In this section, we now derive stopping criteria from \(k\)-NN classification using MSPRT and Bayesian inference for a multiple-choice problem.

\section*{Multiple Sequential Probability Ratio Test}

One simple statistical test for determining whether one of \(C\) different choices has the highest probability density is the MSPRT. MSPRT uses fixed parameters of densities \(\lambda_{+}\)and \(\lambda_{-}\)where \(\lambda_{+}>\lambda_{-}\), it calculates the likelihood that the first data came from the density \(\lambda_{+}\)and others from \(\lambda_{-}\), and then compares those likelihoods.

Without loss of generality, we consider the likelihood that the highest density \(\lambda_{+}\)is occupied by \(\lambda_{1}\). In other words, \(\lambda_{1}=\lambda_{+}\), and \(\lambda_{c}=\lambda_{-}\)for \(c=2, \ldots, C\). Because of the independence between classes,
\[
\begin{align*}
& \log P\left(k_{1}, \ldots, k_{C}, u_{1}, \ldots, u_{C} \mid \lambda_{1}=\lambda_{+}, \lambda_{2}=\lambda_{-}, \ldots, \lambda_{C}=\lambda_{-}\right) \\
& \quad=\log \rho\left(k_{1}, u_{1} \mid \lambda_{+}\right)+\sum_{c=2}^{C} \log \rho\left(k_{c}, u_{c} \mid \lambda_{-}\right) \tag{5}
\end{align*}
\]

The posterior \(P_{1}\) that \(\lambda_{1}\) occupies \(\lambda_{+}\)is proportional to this likelihood Eq. (5). From the Poisson distribution in Eq. (4) with \(k_{c}\), the number of NNs of class \(c\) within the same volume, we can obtain the log of posterior:
\[
\begin{equation*}
\log P_{1}=g^{*} k_{1}-\log \left(\sum_{c=1}^{c} \exp \left(g^{*} k_{c}\right)\right) \tag{6}
\end{equation*}
\]
with a predetermined ratio \(g^{*}=\log \left(\lambda_{+} / \lambda_{-}\right)\). If we consider the volume distribution for the same \(k\)-th NNs, the equation for the posterior also becomes
\[
\begin{equation*}
\log P_{1}=-h^{*} u_{1}-\log \left(\sum_{c=1}^{c} \exp \left(-h^{*} u_{c}\right)\right) \tag{7}
\end{equation*}
\]
with \(h^{*}=\lambda_{+}-\lambda_{-}\). We call Eq. (6) "DN", which considers the difference in the number of NNs within a specific volume of the hypersphere and Eq. (7) "DV", which considers the difference in the volumes of the same \(k\)-th NNs. In order to make a decision with confidence, we can first increase the volume of hypersphere or increase the number of NNs until the criterion exceeds a pre-defined confidence level, then
we can decide the choice. For two-choice problem \((C=2)\), comparing the MSPRT criteria with a certain value reduces to a simple comparison whether \(g^{*}\left(k_{1}-k_{2}\right)\) and \(h^{*}\left(u_{2}-u_{1}\right)\) is greater than a certain confidence threshold, for Eq. (6) and Eq. (7), respectively.

For DV, an additional conservative method can be considered. The decision can be made more carefully for the class of interest (here, class 1 ), by using the maximum possible volume containing \(k\)-th NN , in other words, the volume of the hypersphere of \((k+1)\)-th NN of class 1 instead of the volume of \(k\)-th NN. We call this strategy "conservative DV" (CDV), and in CDV, an additional NN is always used to calculate the accumulated information.

\section*{Bayesian Inference}

Another method of utilizing the Bayesian method is to use the prior density function for \(\lambda\) with parameters \(a\) and \(b\) :
\[
\begin{equation*}
p(\lambda)=\frac{b^{a}}{\Gamma(a)} \lambda^{a-1} \exp (-\lambda b) \tag{8}
\end{equation*}
\]

With conjugacy relationship, we can calculate the posterior probability that the underlying density of choice \(1, \lambda_{1}\), is greater than the underlying densities of the other choices \(\lambda_{2}, \ldots, \lambda_{C}\) with given condition \(D\) on nearest neighbor information. The calculation of \(P\left(\lambda_{1}>\lambda_{2}, \ldots, \lambda_{C} \mid D\right)\) is performed using the probability primitives such as \(P\left(\lambda_{1}<\lambda_{2} \mid D\right)\), \(P\left(\lambda_{1}<\lambda_{3} \mid D\right), \ldots\), and \(P\left(\lambda_{1}<\lambda_{2}, \ldots, \lambda_{C} \mid D\right)\) :
\[
\begin{gather*}
P\left(\lambda_{1}>\lambda_{2}, \ldots, \lambda_{C} \mid D\right)=\int_{0}^{\infty} d \lambda_{1} p\left(\lambda_{1} \mid D\right)\left(1-\int_{\lambda_{1}}^{\infty} d \lambda_{2} p\left(\lambda_{2} \mid D\right)\right) \\
\ldots\left(1-\int_{\lambda_{1}}^{\infty} d \lambda_{C} p\left(\lambda_{C} \mid D\right)\right)  \tag{9}\\
= \\
1-P\left(\lambda_{1}<\lambda_{2} \mid D\right) \ldots-P\left(\lambda_{1}<\lambda_{C} \mid D\right)+  \tag{10}\\
\ldots+(-1)^{C-1} P\left(\lambda_{1}<\lambda_{2}, \ldots, \lambda_{C} \mid D\right) .
\end{gather*}
\]

When the condition is on the number of nearest neighbors \(k_{1}, \ldots, k_{C}\) within a certain volume, the general form of primitives is presented with multinomial coefficients:
\[
\begin{align*}
& P\left(\lambda_{1}<\lambda_{j_{2}}, \ldots, \lambda_{j_{L}} \mid k_{1}, \ldots, k_{C}\right)=  \tag{11}\\
& \quad \sum_{i_{j_{2}}=0}^{k_{j_{2}}} \cdots \sum_{i_{j_{L}}=0}^{k_{j_{L}}} \frac{1}{L^{\left(k_{1}+1+\sum_{c=2}^{L}\left(k_{j_{c}}-i_{j_{c}}\right)\right)}}\binom{k_{1}+\sum_{c=2}^{L}\left(k_{j_{c}}-i_{j_{c}}\right)}{k_{j_{2}}-i_{j_{2}}, \cdots, k_{j_{L}}-i_{j_{L}}}
\end{align*}
\]
where \(L\) and \(j_{1}, \ldots, j_{L}\) are determined according to the primitives in Eq. (10). In addition, when volume information \(u_{1}, \ldots, u_{C}\) is given for \(k_{1}, \cdots, k_{C}\)-th NN in each class, respectively, the primitives are
\[
\begin{align*}
& P\left(\lambda_{1}<\lambda_{j_{2}}, \ldots, \lambda_{j_{L}} \mid u_{1}, \ldots, u_{C}\right)=  \tag{12}\\
& \quad \sum_{i_{j_{2}}=0}^{k_{j_{2}}} \cdots \sum_{i_{j_{L}}=0}^{k_{j_{L}}}\binom{k_{1}+\sum_{c=2}^{L} i_{j_{c}}}{i_{j_{2}}, \cdots, i_{j_{L}}} \frac{u_{1}^{k_{1}+1} \prod_{c=2}^{L} u_{j_{c}}^{i_{j_{c}}}}{\left(u_{1}+\sum_{c=2}^{L} u_{j_{c}}\right)^{k_{1}+\sum_{c=2}^{L} i_{j_{c}+1}}}
\end{align*}
\]
for \(L\) and \(j_{1}, \ldots, j_{L}\) determined from the primitive. Now, Eq. (10) with primitives in Eq. (11) can be considered as a


Figure 2: Examples of diffusion of evidence for three-choice decision making. The diffusion of posteriors, \(P_{1}\) and \(P_{2}\), are plotted on the horizontal and vertical axes. The threshold is set to 8 .
criterion "PN" and Eq. (10) with primitives in Eq. (12) can be considered as a criterion "PV." Here, we used \(a=1\) and positive small value \(b\).

For a two-choice problem, with \(k_{c}\)-th NNs in \(c\) class within the same hypersphere, the probability result becomes a very simple equation
\[
\begin{equation*}
P\left(\lambda_{1}>\lambda_{2} \mid u_{1}, u_{2}\right)=\sum_{m=0}^{k}\binom{2 k+1}{m} \frac{u_{1}^{m} u_{2}^{2 k+1-m}}{\left(u_{1}+u_{2}\right)^{2 k+1}} \tag{13}
\end{equation*}
\]
from Eq. (11). Similarly, with \(u_{1}\) and \(u_{2}\) of \(k\)-th NN in each class, Eq. (12) becomes
\[
\begin{equation*}
P\left(\lambda_{1}>\lambda_{2} \mid k_{1}, k_{2}\right)=\frac{1}{2^{k_{1}+k_{2}+1}} \sum_{m=0}^{k_{1}}\binom{k_{1}+k_{2}+1}{m} \tag{14}
\end{equation*}
\]

Both Eq. (13) and Eq. (14) are the sums of binomial distributions which can be interpreted analogous to coin tossing problem with a biased and an unbiased coin. Eq. (13) corresponds to the probability of having heads less than or equal to \(k\) among \(2 k+1\) tosses of a biased coin, and Eq. (14) corresponds to the probability of having heads less than or equal to \(k_{1}\) among \(k_{1}+k_{2}+1\) tosses of an unbiased coin.

We can note that all derived stopping criteria have a posterior representation where the sum over classes equals one. Therefore, we can consider a \(C-1\) dimensional simplex and the diffusion of the posterior within this simplex. Therefore, a vector with posterior elements for all candidate classes extending Eq. (10) can be considered as a diffusing evidence in a DDM model, and all criteria derived in this work can subsequently be considered as DDM models.

\section*{Relationship with other Exemplar Methods}

One typical method of learning with exemplars is utilizing the similarity measures with exemplars (Nosofsky, 1986; Shepard, 1987). Recently, this model was connected to kernel learning methods in machine learning (Jäkel, Schölkopf, \& Wichmann, 2008), which connected the similarity notion to an associated reproducing kernel Hilbert space as well as to Bayesian inference (Shi, Griffiths, Feldman, \& Sanborn, 2010). These similarity-based methods utilizing exemplars are computationally well-integrated with various machine learning methods.

However, majority voting with equal weights, which is proposed in this work, is a completetly different approach of


Figure 3: Performance of adaptive \(k\)-NN classification using PN, DN, Race, and a machine learning criteria, "Cons." Accuracy is plotted with an average number of NNs used for various thresholds of confidence. Cons makes a decision when the number of recent consecutive NNs of the same class exceeds a threshold (Ougiaroglou et al., 2007).
utilizing exemplars, where the theoretical explanation shows optimality in certain situations (Bailey \& Jain, 1978). Our model is also different from the random walk model using conventional exemplar models (Nosofsky \& Palmeri, 1997). The random walk is performed according to the random retrieval from already generated data, while our model directly considers the underlying density function and uses the generated data without any additional randomness. A severe problem in the memory retrieval of Nosofsky and Palmeri is that a repetitive retrieval of one very similar exemplar will affect the decision predominantly where a noise on this particular exemplar can severely affect the decision accuracy.

\section*{Experiments with Simulation Data}

The examples of diffusion of the evidence for each criteria are shown in Fig. 2. In this experiment, the proposed five examples of evidence PV, PN, DV, CDV, and DN, diffuse with the same NN information. In the figure, all five examples diffuse differently, but they reach the same threshold. The parameters used are \(\lambda_{1}=.25, \lambda_{2}=.35\), and \(\lambda_{3}=.4\), and the decision threshold is .8. Though they diffuse differently, CDV shows a smoother diffusion than the others, and PN and DN show more sampling-wise configuration.


Figure 4: Performance using volume evidence. The CDV with smooth diffusion is slightly better than DV, while PV outperforms both CDV and DV with large margins.

The performance evaluation of the methods is shown using \(k\)-NN classification. We first generated data randomly from three uniform probability densities, \(\lambda_{1}=.2, \lambda_{2}=.7\), and \(\lambda_{3}=.1\), and compared the adaptive \(k-\mathrm{NN}\) classification method between the proposed criteria and other criteria from psychology and machine learning models. In Fig. 3, as expected in our analysis, the Race accumulator model without interaction does not outperform the criteria from statistical tests, PN and DN, although using a Race criterion does give a better performance than a simple majority voting method with fixed \(k\). We also compared our results with a conventional machine learning method, which considers the number of recently appeared NNs belonging to the same class.

In Fig. 4, three criteria using volume information, PV, DV, and CDV, are compared. According to a few realizations in Fig. 1, the diffusion of CDV is in general much smoother than that of DV, and the CDV criterion shows a little better accuracy than DV. PV shows better performance than either DV or CDV.

\section*{Conclusion}

In this work, a general framework integrating decision making with sequential sampling is proposed based on its relationship with the exemplar-type machine learning algorithm, \(k-\mathrm{NN}\) classification. In contrast to previous research on suboptimal weighted voting, we have shown how \(k\)-NN majority voting can be used to better understand the sequential sampling decision making process. Using an adaptive \(k\)-NN classification framework, we also showed how the proposed five examples of optimal criteria are derived for multiple-choice decision making, minimizing the error for any given average resource that can be used. Our future work includes extending this relationship among decision making methods to form a scaffold of understanding within the mathematical framework of \(k\)-NN methods.

\section*{Acknowledgments}

This research was supported in part by the AIM Center for Advanced Intelligent Manipulation and the ROSAEC-ERC.

\section*{References}

Bailey, T., \& Jain, A. K. (1978). A Note on Distance-Weighted \(k\)-Nearest Neighbor Rules. IEEE Transactions on Systems, Man, and Cybernetics, SMC-8(4), 311-313.
Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., et al. (2008, 26). Probabilistic population codes for Bayesian decision making. Neuron, 60(6), 1142-1152.
Bogacz, R., Brown, E., Moehlis, J., Holmes, P., \& Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. Psychological Review, 113(4), 700-765.
Bogacz, R., Usher, M., Zhang, J., \& McClelland, J. L. (2007). Extending a biologically inspired model of choice: multi-alternatives, nonlinearity and value-based multidimensional choice. Philosophical Transactions of the Royal Society, Series B.
Cover, T. (1967). Estimation by the nearest neighbor rule. IEEE Transactions on Information Theory, 14(1), 50-55.
Cover, T., \& Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21-27.
Dragalin, V. P., Tertakovsky, A. G., \& Veeravalli, V. V. (1999). Multihypothesis sequential probability ratio tests . part i: asymptotic optimality. IEEETransactions on Information Technology, 45, 2448-61.
Jäkel, F., Schölkopf, B., \& Wichmann, F. A. (2008, 4). Generalization and similarity in exemplar models of categorization: Insights from machine learning. Psychonomic Bulletin and Review, 15(2), 256-271.
Leonenko, N., Pronzato, L., \& Savani, V. (2008). A class of Rényi information estimators for multidimensional densities. Annals of Statistics, 36, 2153-2182.
Ma, W. J., Beck, J. M., Latham, P. E., \& Pouget, A. \((2006,22)\). Bayesian inference with probabilistic population codes. Nature Neuroscience, 9(11), 1432-1438.
Nosofsky, R. M. (1986). Attention, similarity, and the identificationcategorization relationship. Journal of Experimental Psychology: General, 115(1), 39-57.
Nosofsky, R. M., \& Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. Psychological Review, 104(2), 266300.

Ougiaroglou, S., Nanopoulos, A., Papadopoulos, A. N., Manolopoulos, Y., \& Welzer-Druzovec, T. (2007). Adaptive k-nearest-neighbor classification using a dynamic number of nearest neighbors. In Proceedings of the 11th east european conference on advances in databases and information systems (pp. 66-82).
Ratcliff, R. (1978). A theory of memory retrieval. Psychological Review, 85, 59-108.
Ratcliff, R., \& Mckoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. Neural Computation, 20(4), 873922.

Ratcliff, R., \& Rouder, J. N. (2000). A diffusion model account of masking in two-choice letter identification. Journal of Experimental Psychology Human Perception and Performance, 26(1), 127-140.
Shadlen, M. N., Hanks, A. K., Churchland, A. K., Kiani, R., \& Yang, T. (2006). The speed and accuracy of a simple perceptual decision: a mathematical primer. Bayesian brain: Probabilistic approaches to neural coding.
Shadlen, M. N., \& Newsome, W. T. (1998). The variable discharge of cortical neurons: Implications for connectivity, computation, and information coding. Journal of Neuroscience, 18, 3870-3896.
Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. Science, 237, 1317-1323.
Shi, L., Griffiths, T. L., Feldman, N. H., \& Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. Psychonomic bulletin \& review, 17(4), 443-464.
Smith, P. L., \& Vickers, D. (1988). The accumulator model of two-choice discrimination. Journal of Mathematical Psychology, 32, 135-168.
Usher, M., \& McClelland, J. L. (2001, July). The time course of perceptual choice: the leaky, competing accumulator model. Psychological review, 108(3), 550-592.
Vickers, D. (1970). Evidence for an accumulator model of psychophysical discrimination. Ergonomics, 13, 37-58.
Wald, A., \& Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. Annals of Mathematical Statistics, 19, 326-339.
Wasserman, L. (2003). All of Statistics: A Concise Course in Statistical Inference (Springer Texts in Statistics). Springer. Hardcover.
Zhang, J., \& Bogacz, R. (2010). Optimal decision making on the basis of evidence represented in spike trains. Neural Computation, 22(5), 11131148.

\title{
Where What You Count is What Really Counts
}

\author{
Nader Noori (nnoori@usc.edu), Laurent Itti (itti@usc.edu) \\ Department of Computer Science, University of Southern California Los Angeles, CA 90089 USA
}

\begin{abstract}
We noticed that human subjects were notably faster and more accurate in concurrent counting of three location-based events while they ignored the identity of targets, compared to concurrent counting of three identity-based events while they ignored the locations. In a control experiment, subjects performed a location-based triple counting task, while now also paying attention to the target identity. This did not incur any additional cost, compared to the cost of the location-based counting. Performing each of these tasks relies on maintaining three running numerical counters, and on switching between them to increase each one. Our results suggest that switching between these counters has lower cost when they are associated to spatial locations, compared to when they are associated to identities. This difference is not affected when additionally processing the identity of items. We argue that this might be related to the advantage of the space in switching attention between internal representations.
\end{abstract}

Keywords: Symbolic Working Memory; Working Memory; Spatial Strategies; Spatial Registry; Visuospatial Short-Term Memory; Concurrent Counting; Triple Counting; Focus of Attention.

\section*{Introduction}

The challenge of many non-trivial intellectual symbolic working memory tasks lies in the difficulty of juggling several residents of working memory at the same time; there are one or two items that are being acted upon while a few others are actively maintained and kept apart for future steps of the process. For example think of adding 34 to 89 mentally while looking away from the paper or the screen. During early mathematics education, students are taught to perform this task in a sequential way, and by several calls on the addition table for single-digit numbers. In this example, while 3 and 8 (or 30 and 80 ?) are maintained for the next step of the operation, 9 and 4 are the ones that are acted upon first.

Due to the increasing role of abstract representations in solving difficult problems, intellectual symbolic tasks such as this example have become increasingly important as a part of necessary mental skills for a civilized individual in modern time. Understanding our limitations in maintaining and manipulating mental concepts that are needed for intellectual tasks is key to understanding our cognitive limitations, and to possibly improving individuals' performance. This in turn poses the question of how we manage to keep some items in our working memory, and how we can handle selecting them for the right operation at the right time.

The common metaphor among cognitive psychologists, which captures the selectivity of operations on several actively maintained items, is the shiftable spotlight or focus of attention. Attention in this sense refers to the special treatment that a few representations receive, which in turn makes it possible to be acted upon (or processed) with more agility (in terms of response time or accuracy).

Cowan goes even further and uses the concept of attention to directly relate working memory representations to longterm memory representations. In his view, residents of working memory are those representations in long-term memory which have received attention. He states that this attention may apply to only four items at a time, and hence the number of representations in working memory is limited to four (Cowan, 1999). Some other researchers have mentioned that indeed the spotlight of attention is even narrower than four items, and there is room for only one representation (Garavan, 1998; McElree, 2001). Experimental support for this claim comes from the observation that the most recently attended resident of working memory is privileged in terms of processing speed (McElree \& Dosher, 1989; Mcelree, 2006; Garavan, 1998; Voigt \& Hagendorf, 2002). For example Garavan studied the execution time of human subjects in a self-paced dual counting task where subjects had to keep count of how many of two possible visual shapes (triangles versus squares) had been presented. The sequence of switching between internally maintained counters is dictated by the stimulus sequence of triangles and squares. Garavan noticed that updating a recently updated counter (e.g., when two squares or two triangles are presented consecutively) is significantly faster than updating alternative counters (e.g., when a square is presented after a triangle or vice-versa). Garavan showed that this speedy update of a recently updated internal counter is not related to the perceptual priming in detecting the associate signal. He posited that the execution time difference between updating one counter twice and updating two different counters is related to the cost of shifting the focus of attention from one counter to the other one.

Oberauer has tried to reconcile Cowan's concept of attention with the single spotlight focus of attention, to establish a framework that explains storage and processing as two features of working memory (Oberauer, 2002). While the concept of shifting the focus of attention seems to capture the dynamics of working memory during mental processing, one question remains to be answered: what is the underlying mechanism for switching between two working memory representations? In particular, is there a unitary system for switching attention between items of working memory, either in space or in other dimensions?

For this matter, some researchers have looked at brain activities of human subjects during tasks that involve reconfiguration of cognitive resources by switching between different representations. Yantis and his colleagues, in a series of fMRI studies, have compared BOLD signals during switching attention in the visual-spatial domain or in other representation domains. They have identified a fronto-parietal net-
work which is common among different tasks (Shomstein \& Yantis, 2006; Tamber-Rosenau, Esterman, Chiu, \& Yantis, 2011; Greenberg, Esterman, Wilson, Serences, \& Yantis, 2010; Chiu \& Yantis, 2009). Among these regions, the superior parietal lobule (SPL) of the posterior parietal cortex (PPC), which is also known for its role in shifting visual attention and eyes in space, is shown to be engaged in all of the studied switching tasks. Based on this evidence, Yantis proposes a general-domain switching mechanism (Chiu \& Yantis, 2009).

Yet, an important question is whether such a switching mechanism is indeed a domain-independent machinery, or is in fact a part of an evolutionary older system that is lent or co-opted for use in different domains. It is important to distinguish between these two alternative views as they propose two different views of the evolution of human cognition. On the one hand, the domain-independent machinery may sound more appealing to some researchers in terms simpler description of the functioning mind. On the other hand, the idea of co-opting evolutionary older systems (e.g., sensory-motor systems) for switching is more plausible from an evolutionary perspective (Dehaene \& Cohen, 2007; Paillard, 2000) and provides an opportunity to ground working memory machinery in perceptual-motor systems in line with more recent trend in grounded and embodied cognition (Lakoff \& Núñez, 2000; Barsalou, Simmons, Barbey, \& Wilson, 2003; Damasio \& Damasio, 2006; Dehaene \& Cohen, 2007).

As an example of a grounded model for working memory processing mechanisms, Noori and Itti (Noori \& Itti, 2013a, 2013b, 2011) propose a framework for management of working memory items which relies on the role of sensory-motor working memory systems for manipulation of working memory items, even in the context of symbolic and abstract items. They assume that manipulation of memory items is facilitated by a registry of memory items to spatial locations, accessible to visuospatial attention mechanisms. Switching spatial attention between those locations, is then based on operational schemas, similar to what Arbib has proposed in the perception-action domain (Arbib, 1992). Their proposal suggests the performance on mental operations that need memory manipulation, even in the case of abstract and symbolic representations, would depend on how those sensory-motor systems are utilized.

To test the dependency of switching between working memory representations on utilization of space, we studied speed and accuracy of our subjects in performing a modified version of Garavan's task (Garavan, 1998). We arranged two versions of a triple counting task: identity-based counting (counting appearances of three possible symbols) and location-based counting (counting any symbols appearing in three possible locations). A domain-independent account predicts that the switching time between internal counters should be independent of our counting paradigms. Our results do not favour this prediction.


Figure 1: Schematic view of both triple-counting paradigms. Nine target presentation events (at the center of the diagram) is similar for both paradigms.

\section*{Experiment 1}

The first experiment compares the cost of switching between different running counters in a concurrent triple counting task for an identity-based and a location-based mental counting. Subjects in both versions of the task maintain three running counters in their memory. These counters are associated to three different events.

In both versions of the task, an event is a subject-initiated brief visual presentation of a keyboard character in a box on the screen. In the case of location-based counting, the difference between events is defined based upon the location of the character, while the identity is irrelevant. In contrast, in the identity-based counting version, the difference between counting events is defined based on the identity of characters, and the location is irrelevant. Counters should be updated upon perceiving their associated signal. Since a signal presentation is initiated by the subject, the task progresses with a pace determined by the subject, which allows us to measure execution times.

We measured the time between two consecutive signal initiations and analysed them based on similarity or dissimilarity of counting-relevant and counting-irrelevant features of two consecutive events, to explore the effect of type of counter-event binding. Moreover, we analysed the error rates in counting using different measures, to explore possible effects of counter-event binding on the accuracy of counting.

\section*{Method}

Apparatus Stimuli were displayed on a 46-inch LCD monitor (Sony Bravia XBR-III, \(89 \mathrm{~cm} \times 50 \mathrm{~cm}\) ), 97.8 cm in front of participants (corresponding field of view is \(54.7^{\circ} \times\) \(32.65^{\circ}\) ). To control the viewing distance, subjects used a chin
rest to maintain their head position during the experiment. A gray background \((0.62 \mathrm{~cd} / \mathrm{m} 2)\) was displayed during the experiment.
Subjects Seven female and four male undergraduate students with normal or corrected to normal vision, participated for course credit. Subjects' ages ranged between 19 to 21 ( \(M=19.7, S D=0.78\) ).
Procedure Figure 1 depicts a schematic view of the triple counting paradigm for both identity and location based concurrent counting. Events for both types of counting paradigm presented in a similar way: one of three keyboard characters \(\$\), \# or ? would be selected to appear in one of three fixed boxes located in three of four main quadrants randomly and upon subject's press of any of keys on the keyboard. In each trial boxes centered at vertices of a virtual square with sides that would appear \(3.5^{\circ}\) wide from subject's view point. In each trial three out of four possible boxes were selected randomly and remained on screen throughout the counting process. In trials of location-based counting boxes appeared on the screen from the beginning of the trial and during presentation of the initial counters.

Each trial started by presenting three initial counters. Initial counters were either 0 or 1 which were selected randomly. During trials of counting identity-based events each item was presented next to its initial counter at the center of screen and during trials of location-based counting initial counters appeared at the center of boxes. Each trial included nine counting events. The pace of counting was determined by subjects. At the end of ninth counting event subjects reported the counters using the keyboard and in the same order of initial counter presentation.

Independent of the type of the counting task the identity and the location of target character changed randomly and independently. The identity of two consecutive characters changed with \(50 \%\) of the chance and the location of two consecutive character presentation changed with \(50 \%\) of the chance. This arrangement roughly balanced the change in counting-relevant and counting-irrelevant features.

Five trials of each counting paradigm defined a block. The type of counting paradigm change in two consecutive blocks. Each subject performed between 20 to 30 blocks which left us between 50 to 75 trials of each counting paradigm. At most 10 blocks were performed in each session. A five minute break administered between each two consecutive sessions.
Results Tables 1 and 2 summarize the results of experiment 1. Table 1 shows the average execution time for four possible combinations of changing the location or the identity of two consecutive signals separated based on the type of concurrent counting paradigms.

The effect of three factors on execution times were examined: a. the type of concurrent counting (with LBTC and IBTC as its two levels), b. change in the counting-relevant feature (the location in LBTC and the identity in IBTC) for two consecutive signals (changed or same as two levels) and
\begin{tabular}{ccc}
\hline \hline \multicolumn{3}{c}{ Location Based Counting } \\
\hline & Same Location & Changed Location \\
\hline Same Id & \(0.834 \pm 0.077\) & \(1.552 \pm 0.136\) \\
\hline Changed Id & \(0.889 \pm 0.102\) & \(1.564 \pm 0.116\) \\
\hline \multicolumn{3}{c}{ Identity Based Counting } \\
\hline & Same Location & Changed Location \\
\hline Same Id & \(1.126 \pm 0.135\) & \(1.198 \pm 0.143\) \\
\hline Changed Id & \(2.122 \pm 0.190\) & \(2.120 \pm 0.205\) \\
\hline
\end{tabular}

Table 1: Mean \(\pm\) SE of the execution times (experiment 1).
c. change in the counting-irrelevant feature (the identity in LBTC and the location in IBTC) for two consecutive signals (changed or same as two levels). The execution time data with these three factors was submitted to a \(2 \times 2 \times 2\) withinsubjects ANOVA.

The analysis revealed a significant main effect of the type of counting task \([F(1,10)=16.4, p=0.0151]\), with a faster response time for the location-based triple counting ( \(M=\) \(1.2100 s, S E=0.055\) ) compared to the identity-based triple counting ( \(M=1.6415 s, S E=0.0819\) ). The main effect of change in the counting-relevant feature also proved to be significant \([F(1,10)=171, p=0.0009]\). A change in countingrelevant feature requires switching between the counters, and the average execution time for this case was higher than when two consecutive signals shared the same counting-relevant feature (i.e., the same counter is updated twice in a row). However, for the counting-irrelevant feature whose change does not require switching between counters, no significant impact was observed \([F(1,10)=1.9237, p=0.1956]\).

The analysis also revealed a significant interaction between the type of the double counting task and the change in counting-relevant feature with \(F(1,10)=24.1, p=0.0006\). A further analysis showed that indeed the change in countingrelevant feature resulted in a larger difference in execution times for IBTC compared to LBTC. All three other possible interactions were non-significant.

We quantified counting errors using four different measures: a. the proportion of trials that had at least one mistake in the reported counters (Incorrect Trials), b. the absolute difference between the reported values and the actual values (Counter Error), c. the absolute difference between reported values and actual values, after sorting both the counters and the reported values (Sorted Counter Error) and d. the absolute difference between the sum of reported values and the sum of actual counters (Sum Error). Among these four measures, a. and \(b\). are the most sensitive measures, while c. discounts any error in incorrectly reporting the order of counters, and d. is the least sensitive measure as it does not account for any error in adding to the right counter or in reporting the counters in the correct order.

Table 2 shows mean \(\pm\) standard error values for each of these error measures for two types of concurrent triple counting. To assess the significance of the effect of the type
\begin{tabular}{c|c|c|c}
\hline Error Type & LBTC & IBTC & Sig. \\
\hline Trials with Error & \(20.5 \% \pm 5.4 \%\) & \(39.3 \% \pm 7.1 \%\) & \({ }^{* *}\) \\
Value Error & \(0.11 \pm 0.03\) & \(0.31 \pm 0.07\) & \({ }^{* *}\) \\
Sorted Value Error & \(0.10 \pm 0.03\) & \(0.22 \pm 0.04\) & \({ }^{* *}\) \\
Sum Error & \(0.24 \pm 0.08\) & \(0.40 \pm 0.09\) & \(*\) \\
\hline
\end{tabular}

Table 2: Mean \(\pm\) SE of error measures (experiment 1). ** means \(p<0.01\), * means \(p<0.05\)
of the counting task on each of these four measures, the data of each error measure was submitted to a separate oneway within-subject ANOVA. A significant main effect of the counting type on the error rates was revealed, for all four measures of error. The significance of this impact on the first measure is quantified by \([F(1,10)=20, p=0.0012]\); for the absolute difference, this significance is quantified by \(F(1,10)=15.4, p=0.0028\); for the absolute difference in sorted sequence of counters and reported values, the significance is quantified by \([F(1,10)=17.2, p=0.00197]\); and finally for the absolute difference between the sum of counters and the sum of reported values, the significance of the main effect is quantified by \([F(1,10)=9.11, p=0.0129]\).
Discussion Changing the counting-relevant feature in two consecutive events for both counting paradigms requires increasing a counter which is different from the previously increased counter, and thus involves switching to a different counter. We observed that, independent of what defines the counting-relevant feature (location or identity of items), switching between counters results in a significantly lower speed for counting. Thus, we could replicate what Garavan had previously reported in the case of an identity-based dualcounting task (Garavan, 1998). However, compared to Garavan's study, our subjects were slower than his (both in updating one counter in a row or switching between counters and updating). This difference might be related to the fact that we had three counters, which might have had an extra load for maintaining the items. Moreover, our subjects needed to switch between two different counting paradigms frequently, and this might have had some impact on the execution time. Finally, we did not impose a delay between blocks of different counting paradigms, and subjects were only notified about the change in the counting paradigm by displaying a message on the screen.

However, the striking result of this experiment is related to the significant difference in execution time and counting errors of our two paradigms. Compared to the identity-based paradigm, the location-based paradigm proved to be both faster and more accurate. Even when the sums of reported counters were compared to the sums of actual counts, subjects were significantly more accurate in the location-based counting. Note that this measure for counting error is not even sensitive to the counting-relevant feature, and yet identity-based counting is significantly slower and less accurate with respect to this measure. This reveals that the slower execution time
in identity-based triple counting is not the result of a trade-off between accuracy and speed.

Another striking result was related to the fact that not only the overall execution time during LBTC was less than the execution time during IBTC but also the switching cost was significantly less for LBTC. This is related to the fact that in our analysis we observed a significant interaction between switching condition and the counting paradigm. While updating the same counter during LBTC was about 300 milliseconds faster than IBTC, updating a different counter, which involves switching between counters, was about 560 milliseconds faster. This result suggests that using location as the counting-relevant feature has saved on the switching cost rather than a cost associated to maintaining or updating counters. We discuss the significance of this result in the general discussion.

\section*{Experiment 2}

In our second experiment, we test whether the higher cost for both counting and switching in the identity-based concurrent counting is indeed the result of differences in character versus location perception. It is known that processing visual forms and locations engages two different pathways; the ventral pathway specialized in identifying visual forms, which serves object perception, and the dorsal pathway specialized in identifying spatial locations, which serves action (Goodale \& Milner, 1992). One may argue that the observed differences between execution times in two concurrent counting paradigms is indeed related the processing of the visual input, and before the processing of the counters. To test the effect of identity recognition on the counting cost, we replaced the Identity-Based Triple Counting task with a modified version of the Location-Based Triple Counting task which involves identification of characters. In this paradigm, the running counters are still associated to locations; however, an occasional appearance of a dummy target can change the counter values and the rule of the concurrent counting thereafter. Thus, every time a valid target appears at a location, its identity should be checked before updating the counter associated to the target presentation location. We call this task IdControlled Location-Based Triple Counting or IC-LBTC for short.

\section*{Method}

Subjects Seven female and four male undergraduate students with normal or corrected to normal vision, participated for course credit. Subjects' ages ranged between 19 to 24 ( \(M=20.1, S D=1.4\) ).
Procedure In this experiment \# and ? were targets for incrementing a counter and \% was used as the dummy character. Subject were instructed that trials were arranged in two type of blocks, no-dummy blocks and dummy-possible blocks. In no-dummy blocks which contained 5 trials, only \# and ? could appear in boxes and all events had to be counted. However, in \(50 \%\) of the dummy-possible blocks, at some ran-
\begin{tabular}{ccc}
\hline \hline \multicolumn{3}{c}{ Location Based Counting } \\
\hline & Same Location & Changed Location \\
\hline Same Id & \(0.719 \pm 0.068\) & \(1.214 \pm 0.105\) \\
\hline Changed Id & \(0.716 \pm 0.071\) & \(1.285 \pm 0.105\) \\
\hline \multicolumn{2}{c}{ Id-Controlled Location Based Counting } \\
\hline & Same Location & Changed Location \\
\hline Same Id & \(0.724 \pm 0.071\) & \(1.205 \pm 0.113\) \\
\hline Changed Id & \(0.710 \pm 0.060\) & \(1.285 \pm 0.104\) \\
\hline
\end{tabular}

Table 3: Mean \(\pm\) SE of the execution times (experiment 2).
dom time the character \% would appear only once in a box; in this case, the counter for that box had to be reported as 0 (i.e., reset and ignored during subsequent stimulus presentations). In the beginning of each block, subjects were notified about the type of block by a written message appearing on the screen.

Thus each trial of a IC-LBTC block could be similar to trials of the LBTC or could have the dummy character appearing only once in one of the boxes. Choosing to include the dummy character in a IC-LBTC trial was decided randomly and with \(50 \%\) of the chance.

Given that blocks of different counting paradigms were similar in every sense we imposed a 10 second delay with a message on the screen informing the subject about whether in the next block there will be dummy characters or not.
Results Table 3 shows the average execution time for four possible combinations of changing the location or the identity of two consecutive signals, separated based on the type of concurrent counting paradigms. For the controlled locationbased counting, only those trials without dummy characters were included in the analysis.

To assess the significance of the effect of controlling for the identity of characters during the triple concurrent task, switching locations and switching identities, execution times were submitted to a \(2 \times 2 \times 2\) within-subject analysis of variance with type of counting, switching location and switching identity as three factors. The analysis showed no effect of attending to the identity of characters on the execution times \([F(1,10)=0.011, p=0.92]\). A significant effect of changing the location of target in two consecutive events on the execution times was observed \([F(1,10)=53.9, p=0.00557]\) and marginally-significant effect of switching the identities was observed \([F(1,10)=4.13, p=0.07]\).

The data for all measures of error were separately submitted to within-subject one-way ANOVAs with type of counting as the main factor to assess the impact of attending to the identity of characters on the error rates. None of the analyses returned a significant main effect of the counting paradigm on the error rates.

Except for one subject, all subjects correctly reported the incidence of appearance of the dummy character with \(100 \%\) accuracy.
\begin{tabular}{c|c|c|c}
\hline Error Type & LBTC & IC-LBTC & Sig. \\
\hline Trials with Error & \(19.1 \% \pm 3.9 \%\) & \(18.1 \% \pm 4.1 \%\) & n.s. \\
Value Error & \(0.11 \pm 0.02\) & \(0.10 \pm 0.03\) & n.s. \\
Sorted Value Error & \(0.08 \pm 0.02\) & \(0.07 \pm 0.02\) & n.s. \\
Sum Error & \(0.21 \pm 0.05\) & \(0.17 \pm 0.04\) & n.s. \\
\hline
\end{tabular}

Table 4: Mean \(\pm\) SE of error measures (experiment 2). n.s. : nonsignificant

Discussion The analysis of both execution times and error measures showed that compared to LBTC, attending to the identity of characters during IC-LBTC incurred no extra cost. This suggests that attending to the identity of the items in IBTC does not seem to be the source of extra cost of IBTC counting paradigm relative to LBTC.

However compared to the LBTC in experiment 1, subjects were significantly faster in LBTC trials of the second experiment. Given that subjects had to switch between LBTC and IBTC in the first experiment and LBTC and IC-LBTC in the second experiment, the faster execution time in the second experiment might have been related to a lower cost for switching between the two tasks in the second experiment. This difference might be related to either the 10 second imposed delay between blocks of experiment 2 , or the fact that both tasks in the second experiment are indeed two versions of the same counting paradigm, and thus switching between blocks of experiment 2 is less costly. Furthermore, since there was no significant interaction between the identity of experiments and changing counting-relevant factor in both LBTC trials, the effect of switching between blocks of tasks seems to have had equal effects on both updating the same counter in a sequence or updating two different consecutive counters. This suggests that the extra cost on switching between counters during IBTC counting is not likely related to the cost of switching between blocks of experiment 1 .

\section*{General Discussion}

The goal of this study was to test the dependency of putative switching mechanisms for managing working memory in the internal domain and in a seemingly abstract and symbolic context, on the explicit utilization of space. We analysed our subjects' execution time in two concurrent counting tasks that differed in their reference to spatial locations. In one version of the task, where counting events were associated to spatial locations on the screen, subjects were faster and more accurate than when the identity of visual targets was the basis for the counting events. More importantly, not only the speed of counting in the location-based counting paradigm was generally faster, but also this speed was significantly faster when subjects had to switch between internal counters. Below, we argue that a faster switching between internal counters indicates that the source of speedup indeed is not related to a verbal shortening effect in rehearsing.

In Garavan's process model for a dual-counting task, each counting event consists of a sequence of five steps: 1. stimu-
lus identification, 2. orientation of attention, 3. updating the associated count, 4. rehearsing the other count, 5 . key-press. He suggests that the source of a 300 to 400 msec difference in updating the same counter subsequently versus updating two different counters is related to the cost of the second step: a recently attended resident of the working memory is privileged in terms of processing speed or accuracy (Garavan, 1998), and thus updating a counter which was just updated saves on the cost of bringing the item of working memory into the focus.

This model could be adopted for the triple-counting task by considering a third counter which needs to be included in the switching and the rehearsing steps. This model does not assume that the second step of this process, which accounts for the extra switching cost between two different counts, is dependent on perceptual aspects of the counting tasks. Likewise, no other model of working memory, to our knowledge, in which the focus of attention plays a critical functional role in regulating the process, assumes that the second and the third steps of this process are relevant to the perceptual aspects of the counting task. Hence, according to this process model, steps 2,3 and 5 should be independent of the type of counting events. Our second experiment controlled for the influence of potential effects of perceptual differences, and showed that the source of speed difference in two paradigms cannot be attributed to the perception of events. Consequently, according to this model, the only source of difference in counting speed might be in rehearsing other counts (step 4). However, this effect would affect the speed of counting in a similar way for both updating the same counter or updating counters alternatively. Moreover, the analysis of errors adds another dimension to our argument: even when misplaced counters and signals are discounted in the error calculation, the location-based counting is still significantly more accurate. In sum, we argue that a model that confers a special role to space (unlike Garavan's process model), may be necessary to fully explain our findings.

Noori and Itti's spatial registry framework for manipulation of information in working memory is an example of a model where space plays a special role. According to this model, which assumes that working memory items are bound to spatial locations, accessing items in the internal domain draws on shifting spatial attention to different locations. In the case of location-based counting, counters can be directly bound to the location of boxes, and thus attending to the visual stimulus will automatically draw attention to the location of counters for accessing the counter value. In contrast, during the identity-based triple counting, each signal will draw spatial attention, but the signal location is not correlated with its identity and thus with the associated counter, thus likely a second shift of attention is required to point attention to the correct counter. When items need an update, two shifts in spatial attention may thus be required. The extra shift of attention may account for the slower response time during the identity-based triple counting.

\section*{Acknowledgments}

This work was supported by the National Science Foundation (CRCNS Grant No. BCS-0827764), the Army Research Office (W911NF-11-1-0046), and the U.S. Army (W81XWH-10-2-0076).

\section*{References}

Arbib, M. (1992). Schema theory. Encyclopedia of artificial intelligence, 2, 1427-1443.
Barsalou, L. W., Simmons, W. K., Barbey, A. K., \& Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. Trends in Cognitive Sciences, 7(2), 84-91.
Chiu, Y., \& Yantis, S. (2009, March). A Domain-Independent Source of Cognitive Control for Task Sets: Shifting Spatial Attention and Switching Categorization Rules. The Journal of Neuroscience, 29(12), 3930-3938.
Cowan, N. (1999). An embedded-processes model of working memory. Models of working memory: Mechanisms of active maintenance and executive control, 62-101.
Damasio, A., \& Damasio, H. (2006). Minding the body. Daedalus, 135(3), 15-22.
Dehaene, S., \& Cohen, L. (2007). Cultural Recycling of Cortical Maps. Neuron, 56(2), 384-398.
Garavan, H. (1998, March). Serial attention within working memory. Memory \& Cognition, 26(2), 263-276.
Goodale, M., \& Milner, A. (1992). Separate visual pathways for perception and action. Trends in neurosciences, 15(1), 20-25.
Greenberg, A. S., Esterman, M., Wilson, D., Serences, J. T., \& Yantis, S. (2010, October). Control of Spatial and Feature-Based Attention in Frontoparietal Cortex. The Journal of Neuroscience, 30(43), 14330-14339.
Lakoff, G., \& Núñez, R. (2000). Where Mathematics Comes From: How the Embodied Mind Brings Mathematics into Being (1st ed.). Basic Books.
McElree, B. (2001). Working memory and focal attention. Journal of Experimental Psychology: Learning, Memory, and Cognition, 27(3), 817-835.
Mcelree, B. (2006). Accessing Recent Events. In Psychology of Learning and Motivation (Vol. 46, pp. 155-200). Elsevier.
McElree, B., \& Dosher, B. (1989). Serial position and set size in short-term memory: The time course of recognition. Journal of Experimental Psychology: General, 118(4), 346.
Noori, N., \& Itti, L. (2011). Spatial Registry Model: Towards a Grounded Account for Executive Attention. Proceedings of the 33rd Annual Conference of the Cognitive Science Society, 31873192.

Noori, N., \& Itti, L. (2013a). Schema-driven, space-supported random accessible memory systems for manipulation of symbolic working memory. In Proceedings of the 35th annunal conference of the cognitive science society.
Noori, N., \& Itti, L. (2013b). Traces of intellectual working memory tasks on visual-spatial short-term memory. In Proceedings of the 35th annunal conference of the cognitive science society.
Oberauer, K. (2002, May). Access to information in working memory: exploring the focus of attention. Journal of Experimental Psychology. Learning, Memory, and Cognition, 28(3), 411-421. (PMID: 12018494)
Paillard, J. (2000). Neurobiological roots of rational thinking. In Prerational Intelligence: Adaptative Behavior and Intelligent Systems Without Symbols and Logic (pp. 343-355). Kluwer Academic Publisher.
Shomstein, S., \& Yantis, S. (2006, January). Parietal Cortex Mediates Voluntary Control of Spatial and Nonspatial Auditory Attention. The Journal of Neuroscience, 26(2), 435-439.
Tamber-Rosenau, B. J., Esterman, M., Chiu, Y., \& Yantis, S. (2011, October). Cortical Mechanisms of Cognitive Control for Shifting Attention in Vision and Working Memory. Journal of Cognitive Neuroscience, 23(10), 2905-2919.
Voigt, S., \& Hagendorf, H. (2002). The role of task context for component processes in focus switching. Psychologische Beitrage.

\title{
Traces of Intellectual Working Memory Tasks on Visual-Spatial Short-Term Memory
}

\author{
Nader Noori (nnoori@usc.edu), Laurent Itti (itti@usc.edu) \\ Department of Computer Science, University of Southern California Los Angeles, CA 90089 USA
}

\begin{abstract}
We measured the sensitivity in detecting a change in the location of one of two visual targets over a short period of time to investigate the impact of a secondary intellectual symbolic WM task on the visuospatial short-term memory, in a dual-task paradigm. We observed that engaging in a WM task that involves manipulation of symbolic information impacts the ability to detect a location change, and this impact does not change when more time is allocated to the WM task. Furthermore, we observed that the impact of a mental sorting task on the ability to detect location changes is spatially selective to the horizontal orientation. Our results suggest a possible role for sensory-motor working memory, which supports perceptionaction schemas in manipulation of information during the intellectual symbolic working memory tasks.
\end{abstract}

Keywords: Perception-Action Schema; Symbolic Working Memory; Sensory-Motor Working Memory; Working Memory Manipulation; Visuospatial Short-Term Memory.

\section*{Introduction}

The quality of our modern life has become so dependent on the ability to perform intellectual tasks with symbols (e.g. addition or subtraction) that we force our children to spend a big chunk of their life on learning them at school and home. Effective assessment of individuals' ability in performing these tasks has become a constant occupation of mind for some cognitive psychologists among which some have suggested that the ability of management of working memory (WM) is an indicator and the key to understanding individuals' ability in performing these tasks (Conway, Kane, \& Engle, 2003). Suggestions such as this, have created a motivation for understanding mechanisms of WM management.

Better understanding the human ability for maintaining and manipulating symbolic content poses the question of which brain mechanisms may be recruited by these tasks. From an evolutionary standpoint it is not easy to entertain the idea that our brain has evolved a dedicated system for maintaining and manipulating symbolic concepts in mental schemas, which are indeed very recent cultural inventions (e.g., numbers in mental arithmetic). Meanwhile, humans often perform daily sensory-motor routines that require temporary maintaining and manipulating of information gathered from the environment and relevant to the task. Robust maintaining of task-relevant information for performing perception-action schemas provides an adaptive value, which might have led to evolution of sensory-motor working memory. For instance think of the adaptive advantage of the ability to temporarily maintain the location of a targeted prey which is momentarily out of sight until the right moment for attack.

As suggested by some researchers, it is conceivable that evolutionary older systems for encoding, maintaining and manipulating of information for rudimentary tasks are coopted or reused for the intellectual tasks with newly invented
concepts (Paillard, 2000; Dehaene \& Cohen, 2007). For example, some researchers have entertained the idea of using space in the representation of numbers (Knops, Thirion, Hubbard, Michel, \& Dehaene, 2009; Wood, Willmes, Nuerk, \& Fischer, 2008).

However, studying capacities for maintaining and manipulating of information in the intellectual domain and in the sensory-motor domain are traditionally pursued in different research communities, with not much of cross-talk. Hence, studying the possibility of reusing sensory-motor working memory in intellectual symbolic tasks has not been fully explored yet.

The goal of the present study is to investigate the possibility of involvement of visual-spatial short-term memory in manipulation of information during intellectual symbolic working memory tasks. Short-term memory of the location of a visual target is a component of sensory-motor working memory and is crucial for performing a range of visual-motor tasks. We measured the impact of two different intellectual symbolic tasks on the ability to maintain the spatial information of the location of visual targets.

Similar attempts with an opposite goal have been made to pinpoint the role of the central executive (CE) as the supplier of executive attention (Repovs \& Baddeley, 2006) during maintaining visuospatial short-term memories (Phillips, 1983; Logie, Zucco, \& Baddeley, 1990). In those studies, spatial span - the maximum size of a matrix of symbols which can be recalled better than a threshold performance - is used as a measure for the capacity spatial memory. Any change in the memory span as the result of engagement of the CE would be interpreted as evidence for a role for executive resources in maintaining spatial information. Meanwhile, engaging CE is mostly achieved by engaging subjects in intellectual symbolic working memory tasks, known as executive working memory tasks.

These studies have been equivocal in their conclusions about the role of executive resources in maintaining spatial memory. Phillips has reported that performing a mental arithmetic task reduces the visual/spatial span (Phillips, 1983). He concluded that maintaining spatial information is facilitated through an active mental imagery process which is inhibited by the load of the mental arithmetic. Logie, Zucco and Baddeley (Logie et al., 1990) compared the effect of both a mental arithmetic and a mental imagery task, on both visual and word spans, and showed that the the mental imagery task impairs the visual span to a greater extent, while mental arithmetic impairs the word span to a greater extent. However, they still observed an impact of mental arithmetic on visuospatial span and stated that "the impairment in short-term visual memory
resulting from secondary arithmetic reflects a small general processing load".

However, later on, Baddeley and Repovs summarized the results of many dual-task studies, pinpointing the role of CE in maintaining and manipulating information in components of Baddeley's multi-component model of working memory. They concluded that simple representation and maintenance of information may be independent from the CE (Repovs \& Baddeley, 2006). This conclusion includes maintaining spatial information tested in measuring spatial span which is contrary to Baddeley's previous note on the impact of mental arithmetic on spatial span.

In the present study, we use the sensitivity of subjects in detecting a change in location of two dots to measure the ability of retaining visual-spatial information. This requires retaining an amount of spatial information which is below the capacity of normal subjects (Luck \& Vogel, 1997). In other words, instead of using a fixed threshold for performance in measuring the span of short-term memory, we used a fixed amount of information load below the normal capacity of our subjects, to measure the effect of a secondary intellectual working memory task. Although this paradigm does not measure the spatial working memory capacity in its conventional definition yet, it reflects the general capacity of maintaining spatial information over a short period of time. Moreover, this measure tests the spatial short-term memory in a way that is closer to the use of spatial information in daily perception-action routines. Finally, this paradigm can be easily used to test the spatial selectivity of any potential impact on the spatial short-term memory. This paradigm can be deployed in fixed-length blocks which eliminates the influence of the training factor.

Using this sensitivity measurement paradigm, we inspected the influence of two different symbolic working memory tasks on the short-term spatial information retention. In our first experiment, we used a dual-counting task in which two running counts need to be maintained and updated upon presentation of two distinguishable audio signals. We use the rate of signal presentation as a parametric feature to change the amount of time that is allocated to this task. This allows us detect any impact onto spatial short-term memory caused by decaying information as the result of performing the symbolic working memory (SWM) task.

In our second experiment, we used mental reordering versus retaining of four random alphabetical characters (presented auditorily) as our symbolic working memory tasks. We compared their impact onto retaining spatial information along either the horizontal or vertical orientation. This allowed us to test the spatial selectivity of the impact of mental reordering of characters compared to maintaining them.

\section*{Experiment 1}

We asked our subjects to perform a mental dual counting task of two audio signals, while they were also retaining visualspatial information. The goal was to test whether this sym-


Figure 1: Schematic view of the experimental paradigm.
bolic working memory task could interfere with retaining visual-spatial information as simple as the spatial locations of two visual targets. We also aimed to test whether a possible interference is due to competition over scarce executive resources that might be needed for both the manipulation of working memory content and the retention of visual-spatial information.

Mental dual counting involves maintaining two running counters, associated with two signals, in working memory, and, each time a new signal is perceived, incrementing the associated counter. The rate of updating of the internal counters can be adjusted by the rate at which audio signals are presented. This allows manipulation of the total amount of time that putative executive resources may be free and available for other tasks (e.g., active retaining of spatial short-term memory), which in turn may affect the sensitivity measure. We chose two different rates for presenting audio signals for the dual counting task. In separate blocks, we asked subjects to ignore versus count the signals while retaining the visualspatial information.

\section*{Method}

Apparatus Visual-spatial stimuli were displayed on a 46inch LCD monitor (Sony Bravia XBR-III, 1,016 \(\times 571.5\) mm ), 97.8 cm in front of participants (corresponding field of view is \(54.7^{\circ} \times 32.65^{\circ}\) ). To control the viewing distance, subjects used a chin rest to maintain their head position during the experiment. A gray background \((0.62 \mathrm{~cd} / \mathrm{m} 2)\) was displayed during the experiment. A headphone was used for presenting audio stimuli. Our stimulus presentation program was developed using iLab Neuromorphic Toolkit (iNT) and operated on a Linux 64bit machine.

Subjects Fourteen female and one male undergraduate students with normal or corrected to normal vision, participated for course credit. Participants' ages ranged from 18 to 23 years \((M=20.9, S D=1.6)\).

Procedure Figure 1 displays a schematic view of the experimental paradigm. Visual-spatial stimuli consist of two separately displayed red dots, each one placed randomly on an imaginary circle at center of screen with a diameter of \(3.125^{\circ}\) angle of view. Each dot stayed on the screen for 500 ms and a 500 ms blank screen separated the display of red dots. On the virtual circle, dots were at least \(90^{\circ}\) and at most \(120^{\circ}\) apart. Subjects had to retain the location of red dots for about 10 seconds during which they were supposed to engage in a symbolic working memory task.

During a 10 second period after the removal of the second dot, audio signals of two easily distinguishable types were played in a random order either in a slow tempo or a fast tempo. We used two 250 ms long, 50 Hz tones as audio signals; a soft tone (sine wave) and a rough tone (square wave). For the slow tempo condition, four signals were played with a random ISI of 3000 ms to 3600 ms , and for the fast tempo condition 8 signals were played with ISI of 1330 ms to to 2000 ms .

In all trials, subjects were given an initial set of 3 separate digits (each could have initial value between 0 and 3 ). In half of the blocks, subjects were asked to ignore audio signals and to keep repeating three random digits played before the onset of visual targets. We refer to this task condition as the ignore condition (IC). Under this condition, subjects had to report these same three digits at the end of the trial. In the counting condition, subjects were asked to increment the first digit upon hearing the soft tone, to increment the last digit upon hearing the rough tone, and to remember the middle digit unchanged. All three digits were reported at the end of trial. We refer to this condition as the engage condition (EC).

The memory of the location of targets was probed at the end of a 10 second retaining period by presenting two probe targets simultaneously. Probe targets were presented either on the exact same location as the initial targets (with \(50 \%\) chance), or the location of one of the probe targets was shifted along the imaginary circle at least by \(45^{\circ}\) and at most \(60^{\circ}\) away from the location of the initial target. Subjects were supposed to respond whether both probe targets appeared at the same locations as the original stimuli. During the retaining period, a fixation cross remained at the center of the screen. Subjects fixated the fixation cross during the SWM task execution period. Subjects reported their three digits by mouse clicks on a virtual keypad after responding to the visuospatial probe.

We administered the experiment in separate blocks of 20 trials for the engage and ignore conditions. Each block contained equal numbers of trials with each possible tempo. Each subject performed two blocks of trials for each engagement condition.

Results Sensitivity of subjects in detecting matching probes was used to measure the impact of the symbolic working memory (SWM) task onto the visual-spatial short-term memory (VSSTM) task. Figure 2 demonstrates the mean value of sensitivity \(\left(d^{\prime}\right)\) in identifying matching visuosptial


Figure 2: Impacts of task condition and audio signal rates on the sensitivity measure (experiment 1 ).
probes, for different conditions. To determine the significance of the impact of task and tempo factors, \(d^{\prime}\) values were submitted to a two-way ANOVA with repeated measures on both factors.

The analysis revealed a main effect of task at significance level \(p<0.0001[F(1,14)=37.69]\). The tempo of audio signals did not show a significant main effect \([F(1,14)=\) \(0.504, p=0.49]\). No significant interaction between factors was observed \([F(1,14)=0.11, p=0.74]\).

Further analysis revealed that sensitivity was higher in identifying identical probe targets when subjects ignored audio signals \((M=2.46, S D=1.29)\), compared to when subjects were engaged in dual counting of audio signals ( \(M=\) \(1.57, S D=1.21\) ). Moreover, increasing the tempo of audio signals decreased the mean value of \(d^{\prime}\) for both task conditions; however, this change did not reach a significant level.

To measure the engagement of subjects in the counting task we compared the number of counted signals for both tempos. The difference between the sum of reported counters and the sum of initial counters was used as the measure of counted signals. The average counted signals for fast tempo was \(6.7 \pm 1.2(M \pm S E M)\) and the average counted signals during the slow tempo was \(3.8 \pm 0.2(M \pm S E M)\), which was significantly less than the counted signals for the fast tempo \([F(1,14)=131.04, p<0.0001]\).

Discussion Our results indicates that, first, the sensitivity measure is sufficiently sensitive for detecting the impact of a secondary working memory task such as the dual counting, even though the load on the VSSTM appears to be half of the capacity of visual-spatial short-term memory in normal subjects (Luck \& Vogel, 1997). Second, the double counting task can impair the retention of visual-spatial information over a short period of time. A significant impact on the sensitivity measure with such a low amount of spatial information suggests that the dual counting task, independent of the tempo, can potentially impact the spatial span too.

One may maintain that this effect is caused by engaging

CE in the dual counting task. However, on the basis of the sensitivity measure, increasing the rate of dual counting neither showed a main effect nor an interaction with the VSSTM task. Based on this result, one may come to the conclusion that increasing the tempo indeed does not change the complexity of the task, and thus does not add to the load on the central executive.

In this sense, the double counting might use a specific amount of executive resources in lapses associated to each signal presentation event. Garavan has proposed a model for a self-paced version of dual counting task which consists of five steps (Garavan, 1998): 1. stimulus identification, 2. orientation of attention to the associated counter, 3. updating the count, 4. rehearsing the other count, 5. key-press. The first four steps can be used as a model for our version of the dual counting task. Previous research suggests that executive attention does not play a direct role in the first step (He \& McCarley, 2010). Also, verbal rehearsing in step 4 is suggested not to be dependent on executive resources (Repovs \& Baddeley, 2006). Additionally one may maintain that rescheduling the sequence of rehearsing in the case of switching between different counters (Garavan, 1998) may also draw on executive resources.

Involvement of executive resources in updating counters may result to unavailability of necessary resources for retaining visuospatial information over a refractory period (Pashler et al., 1994). Hence, a higher rate of signal presentation in fast tempo trials hypothetically would occupy a larger fraction of retaining period with a refractory condition.

Yet, one needs to establish how executive resources may play a role in retaining location of two visual targets to leverage a refractory explanation for the impact of dual counting on the sensitivity in location change detection. One may propose that retention of VSSTM requires active maintenance through a rehearsing process (Awh, Jonides, \& ReuterLorenz, 1998; Awh et al., 1999), which according to Jonides is a "controlled sequence of retrievals and re-encoding of items into the focus of attention" (Jonides et al., 2008). This may also draw on general executive resources needed for manipulation of information in symbolic working memory tasks. This argument hinges on this assumption that rehearsing prevents VSSTM traces from decay, so that interrupting the rehearsal process results in decaying traces of spatial short-term memory. Yet, this would imply that the more the rehearsing is interrupted, the more the effect of decay is pronounced. This in turn suggests that adding to the rate of dual counting may affect the performance on the visual-spatial task, which is not supported by our results.

Without CE as the shared scarce commodity between the SWM task and spatial information maintaining process, one should consider another source of conflict between manipulation of information in the SWM task and retaining spatial information. One source of conflict could be that visuospatial short-term memory is indeed used during the SWM task.

The use of space for the manipulation of information has
been previously discussed for specific SWM tasks such as immediate reverse recall (Rudel \& Denckla, 1974) and mental sorting of numbers (Noori \& Itti, 2011). One may imagine a use of space as natural addressing system for the content of SWM, which can be used as a handle to shift processing to different items of working memory (Noori \& Itti, 2011).

Our next experiment explores this matter in the case of a mental sorting task by measuring the sensitivity in detecting a location change along the horizontal versus the vertical directions. An account based on a bottleneck in executive resources for the impact of SWM on retention of spatial information maintains that interrupting the CE would affect VSSTM independently of the spatial location of visual targets. Hence, our second experiment provides us with another opportunity for testing the role of CE in retaining visualspatial information.

\section*{Experiment 2}

To test whether the observed influence of the SWM task on VSSTM is due to utilization of space for active manipulation of symbolic working memory content, we examined the selectivity of the impact of a mental sorting task on VSSTM. In particular, we used two visual-spatial targets either along the horizontal orientation or the vertical orientation. Subjects performed a sorting task on a random list of English letters during the visual-spatial information retaining period.
Subjects Eleven female and three male native English speaking undergraduate students with normal or corrected to normal vision participated for course credit. Participants' ages ranged from 19 to 22 years \((M=20.39, S D=1.4)\).
Procedure The procedure for this experiment is similar to experiment 1, except for the location of visual targets and the symbolic working memory task ( see Figure 1 ). Visual targets were two red dots presented either along a horizontal line or a vertical line passing through the center of screen, each dot on one side of the center, and between \(1^{\circ} \ldots 4.9^{\circ}\) angle of view away from the center. Visuospatial probe targets were presented simultaneously in the same locations as target stimuli with \(50 \%\) chance, otherwise, one of probe targets was displaced by \(1.4^{\circ}\), either inward or outward along original presentation direction so that two probe targets remained on two sides of the center cross along the direction of initial presentation.

Before the onset of the red dots, four randomly selected English letters were presented aurally to be maintained in the same presentation order (during maintaining trials), or sorted in alphabetical order (during sorting trials), within a 10 second period. At the end of the delay period, subjects first responded to the visuospatial query, followed by reporting four characters by mouse clicks on a virtual keypad displayed on the screen.

We administered the experiment in separate blocks of 20 trials for the maintaining and sorting conditions, but each block contained equal number of trials for each different di-
rection for the presentation of visual targets. Each subject performed two blocks of trials for each task condition.


Figure 3: Average sensitivity measure for two tasks and two target orientations (experiment 2).

Results Figure 3 demonstrates the mean value of sensitivity \(\left(d^{\prime}\right)\) in identifying matching visuosptial probe targets for different conditions of SWM task (Maintaining vs. Sorting) and visual target orientations (Horizontal vs. Vertical). To determine the significance of the impact of task and target orientation factors \(d^{\prime}\) values were submitted to a two-way ANOVA with repeated measures on both factors.

The analysis revealed a main effect of the task \([F(1,13)=\) \(8.43, p=0.012]\). No significant main effect of target orientation was determined \([F(1,13)=3.12, p=0.10]\) while the interaction was marginally significant \([F(1,13)=4.3, p=\) \(0.058]\). A post-hoc correlated-samples one-way ANOVA revealed a simple effect of the task, only for horizontal targets \([F(1,13)=15.26, p=0.0018]\), and no simple effect of the task condition at the level of vertical visual targets was observed \([F(1,13)=0.03, p=0.86]\).

Moreover, further analysis for exploring simple effects of orientation at different task levels showed that under the maintaining condition subjects demonstrated a higher sensitivity in detecting identical horizontal probe targets ( \(M=\) \(2.02, S E=0.19)\) compared to identical vertical probe targets \((M=1.28, S E=0.29)\). A correlated-samples one-way ANOVA revealed that the simple effect of orientation during the maintaining task is significant \([F(1,13)=5.11, p=\) 0.041].

Discussion As the analysis revealed, compared to maintaining of four characters in their original order for a later recall, sorting them into an alphabetical order could significantly influence the sensitivity measure. This result again demonstrates the capacity of the sensitivity measure in registering the impact of a secondary SWM task on temporary retention of spatial information. Given our significant results under the low amount of load on VSSTM in our location change detection, one would also expect an impact on spatial span tasks (higher load) due to engaging in a mental sorting task.

However the striking result was that the impact of the sort-
ing task on the sensitivity measure is only significant for visual targets that are spanned along the horizontal direction. Switching task condition did not change the average sensitivity to shift in location of targets along the vertical direction.

The sensitivity to change of location for vertically spanned visual targets was significantly above chance and, unlike the horizontally spanned targets, switching to the sorting task did not decrease sensitivity. This is consistent with the finding of the previous experiment, in that the influence of SWM task on the retention of spatial information is not caused by involvement of executive resources in spatial information retention; otherwise, one would expect an influence on the sensitivity for vertically distributed visual targets too. The initial sensitivity along vertical orientation was lower than along the horizontal orientation, hence one may raise the point that there was less room for decreasing the sensitivity along the vertical direction, and detecting a change would need more space. Controlling for the influence of this initial difference on the sensitivity in location change detection remains to be tested in a separate experiment, with a setup that can balance the sensitivities for detecting target locations along the vertical and horizontal directions during the list maintaining task.

\section*{General Discussion}

The goal of this study was to explore the influence of intellectual working memory tasks devoid of visual and spatial features on the ability of retaining visuospatial information over a short period of time. We used a measure which is different from the actual capacity of spatial memory for holding spatial information. Instead we used the ability to detect a change in spatial location of one of two simple visual targets. In both experiments we observed that engaging in active manipulation of working memory content results to a decrease in the sensitivity of subjects in detecting a change in location of targets which needs to be explained.

Theories of working memory in the realm of cognitive psychology -independent of what they assume about the nature of representation in WM - usually assume a specific execution model based on separation of storage and execution. As such, a conflict between two tasks is either associated with sharing storage or with drawing on limited executive resources. Given that the WM tasks in our study are devoid of immediate visual features, one may conclude that the source of conflict is the the dependency of the retention of visualspatial information and WM task on the CE. Yet, one should be clear as to how CE explains this conflict rather than - as Baddeley and Repovs have stated (Repovs \& Baddeley, 2006) - using CE as a homunculus which has an undisclosed role in everything.

As we discussed in our introduction Baddeley's latest account assumes no role for the CE in retention of spatial information as simple as we tested in our experiment.

We also discussed that the proposal of Awh and his colleagues for engagement of a rehearsing mechanism in maintaining visuospatial information (Awh et al., 1998, 1999),
and Jonides' account for the dynamic of rehearsing process (Jonides et al., 2008), may suggest a role for the CE in retention of visuospatial information. Our experiments were able to test this hypothetical role and our results did not support it.

Another view of working memory, proposed by Cowan (Cowan, 2001), has recently gained popularity in the cognitive psychology community. According to Cowan, working memory content is a part of long-term memory in a heightened state and under attention (Cowan, 2001). He explains the limitation in the capacity of working memory by the limitation of the internal attention in covering only four items at a time. The role of CE in this schema is to dynamically dispatch attention between representations in the long term memory, to make them available for processing. Accordingly, one might say that attention might be shared between information about the locations of two dots and the identities of four characters of the sorting task, which would exceed the capacity limit of 4 items proposed by Cowan. Yet, this explanation lacks sufficient detail to explain why adding to the rate of double counting and fastening this juggling of content of working memory under the watch of attention has no impact, or, in our second experiment, the effect of sorting characters in memory is limited to the sensitivity in detecting changes along the horizontal orientation.

Finally, there is another explanation, previously proposed by Noori and Itti (Noori \& Itti, 2013, 2011), which falls out of the realm of models of working memory that assume separation of execution and storage. According to Noori and Itti, manipulation of information during symbolic intellectual working memory tasks (e.g., mental sorting or mental arithmetic) is made possible by re-using those sensory-motor working memory systems that evolutionarily have been developed to support perception-action routines (such as the occulomotor system). They argue that the capacity of maintaining information about locations of objects in space, in preparation for action on these objects, may provide a capacity for an internal binding of working memory task items for further manipulation. They assume that the management of working memory is made possible through an operational schema. In a way that was specified by Arbib's schema theory, operational schemas eliminate the need for a general-purpose CE in charge of management of working memory content, by instead defining explicit mechanisms that range from simple action-perception routines in catching a prey (Arbib \& Liaw, 1995) to high-level language production (Arbib, 2005).

According to this account, one may assume that visuospatial short-term memory which maintains information about the whereabouts of real objects (e.g., locations of dots in our experiments), is also being utilized for the manipulation of working memory for the intellectual symbolic task (e.g., keeping two running counts separate and yet accessible) by binding symbolic items to space. This assumption may explain the observed effects as the result of a retroactive interference which masks memory of stored information about the location of dots.

\section*{Acknowledgments}

This work was supported by the National Science Foundation (CRCNS Grant No. BCS-0827764), the Army Research Office (W911NF-11-1-0046), and the U.S. Army (W81XWH-10-2-0076).

\section*{References}

Arbib, M. (2005). From monkey-like action recognition to human language: An evolutionary framework for neurolinguistics. Behavioral and brain sciences, 28(02), 105-124.
Arbib, M., \& Liaw, J. (1995). Sensorimotor transformations in the worlds of frogs and robots. Artificial Intelligence, 72(1), 53-79.
Awh, E., Jonides, J., \& Reuter-Lorenz, P. A. (1998, June). Rehearsal in spatial working memory. Journal of Experimental Psychology. Human Perception and Performance, 24(3), 780-790. (PMID: 9627416)

Awh, E., Jonides, J., Smith, E. E., Buxton, R. B., Frank, L. R., Love, T., et al. (1999). Rehearsal in Spatial Working Memory: Evidence From Neuroimaging. Psychological Science, 10(5), 433-437.
Conway, A., Kane, M., \& Engle, R. (2003). Working memory capacity and its relation to general intelligence. Trends in cognitive sciences, 7(12), 547-552.
Cowan, N. (2001). The Magical Number 4 in Short-Term Memory: A Reconsideration of Mental Storage Capacity. Behavioral and Brain Sciences, 24(01), 87-114.
Dehaene, S., \& Cohen, L. (2007). Cultural Recycling of Cortical Maps. Neuron, 56(2), 384-398.
Garavan, H. (1998, March). Serial attention within working memory. Memory \& Cognition, 26(2), 263-276.
He, J., \& McCarley, J. S. (2010). Executive working memory load does not compromise perceptual processing during visual search: Evidence from additive factors analysis. Attention, Perception, \& Psychophysics, 72(2), 308-316.
Jonides, J., Lewis, R. L., Nee, D. E., Lustig, C. A., Berman, M. G., \& Moore, K. S. (2008, January). The Mind and Brain of ShortTerm Memory. Annual Review of Psychology, 59(1), 193-224.
Knops, A., Thirion, B., Hubbard, E. M., Michel, V., \& Dehaene, S. (2009). Recruitment of an Area Involved in Eye Movements During Mental Arithmetic. Science, 324(5934), 1583-1585.
Logie, R. H., Zucco, G. M., \& Baddeley, A. (1990, October). Interference with visual short-term memory. Acta Psychologica, 75(1), 55-74.
Luck, S. J., \& Vogel, E. K. (1997, November). The capacity of visual working memory for features and conjunctions. Nature, 390, 279-281.
Noori, N., \& Itti, L. (2011). Spatial Registry Model: Towards a Grounded Account for Executive Attention. Proceedings of the 33rd Annual Conference of the Cognitive Science Society, 31873192.

Noori, N., \& Itti, L. (2013). Schema-driven, space-supported random accessible memory systems for manipulation of symbolic working memory. In Proceedings of the 35th annunal conference of the cognitive science society.
Paillard, J. (2000). Neurobiological roots of rational thinking. In Prerational Intelligence: Adaptative Behavior and Intelligent Systems Without Symbols and Logic (pp. 343-355). Kluwer Academic Publisher.
Pashler, H., et al. (1994). Dual-task interference in simple tasks: Data and theory. Psychological bulletin, 116, 220-220.
Phillips, W. A. (1983, August). Short-Term Visual Memory. Royal Society of London Philosophical Transactions Series B, 302, 295308.

Repovs, G., \& Baddeley, A. (2006, April). The multi-component model of working memory: Explorations in experimental cognitive psychology. Neuroscience, 139(1), 5-21.
Rudel, R., \& Denckla, M. (1974). Relation of forward and backward digit repetition to neurological impairment in children with learning disabilities. Neuropsychologia, 12(1), 109-118.
Wood, G., Willmes, K., Nuerk, H., \& Fischer, M. H. (2008). On the cognitive link between space and number: a meta-analysis of the SNARC effect. Pabst Science Publishers.

\title{
Measuring the comprehension of negation in 2- to 4-year-old children
}

\author{
Ann E. Nordmeyer \\ anordmey@stanford.edu \\ Department of Psychology \\ Stanford University
}

\author{
Michael C. Frank \\ mcfrank@stanford.edu \\ Department of Psychology \\ Stanford University
}

\begin{abstract}
Negation is one of the most important concepts in human language, and yet little is known about children's ability to comprehend negative sentences. In this experiment, we explore how children's comprehension of negative sentences changes between 2- to 4-year-old children, as well as how comprehension is influenced by how negative sentences are used. Children between the ages of 2 and 4 years watched a video in which they heard positive and negative sentences. Negative sentences, such as "look at the boy with no apples", referred either to an absence of a characteristic or an alternative characteristic. Older children showed significant improvements in speed and accuracy of looks to target. Children showed more difficulty when the negative sentence referred to nothing, compared to when it referred to an alternative. In addition, children showed an early tendency to look towards the named noun, even when that noun was negated. This study contributes to our understanding of children's comprehension of negative sentences, as well as our understanding of the conceptual structure of negation.
\end{abstract}

Keywords: Negation; language development

\section*{Introduction}
"No" is among the first words that children learn, as well as one of the most important. Negation is a fundamental element of human language - it is essential to logical systems, allows us to evaluate whether a statement is true or false, and it gives us a way to express concepts such as nonexistence. Negation is also challenging for language users; adults take longer to process negative sentences than positive ones (Clark \& Chase, 1972; Just \& Carpenter, 1971, 1976; Carpenter \& Just, 1975). These findings lead us to an apparent paradox - how is that negation is difficult for adults, yet acquired at such a young age? By examining children's acquisition of negation, we can explore the origins and development of logical concepts.

Not all uses of negation are the same; words like "no" and "not" allow us to express multiple concepts. Three primary categories have been identified in children's early negative utterances: nonexistence, rejection, and denial/truth-functional (Bloom, 1970, 1993; Pea, 1980). A child expressing rejection might say "no go outside" when they want to stay inside, while a child who says "no more juice" to describe an empty cup is expressing nonexistence (Bloom, 1970). Denial involves making a statement about falsehood; a child might say "that not lollipop" if they believe a candy has been falsely identified as a lollipop. Additional types of negation have been identified as well. Pea (1980) identified two additional categories - self-prohibition, used when the child is about to engage in a forbidden action, and unfulfilled expectations, used when an expected action/object is not present. Choi (1988) identified a full 9 categories of negation, including failure, inability, epistemic negation (e.g. "I don't know"),
and inferential negation (i.e. negation of inferred beliefs of others). Regardless of taxonomy, negation is used in a variety of contexts to express a range of different thoughts.

The relationship between different types of negation is unknown. One possibility is that distinct categories of negation belong to a single cohesive concept. Even pre-linguistically, nonexistence, rejection, and denial could all fall under a superordinate conceptual category of negation. It is also possible, however, that these types of negation represent fundamentally different concepts. For example, the situation in which a child expresses a dislike for going outside (rejection) is very different from a child commenting on an empty juice cup (nonexistence). Perhaps it is only the common language used to describe these events that unites these concepts. One way of untangling these possibilities is by examining children's understanding of different negative concepts, and exploring how their conceptual structure changes as they develop the language to express these thoughts.

The acquisition of linguistic negation follows a long developmental trajectory. As early as 12 months, children produce negation in the form of the word "no", typically to express nonexistence and rejection (Bloom, 1970, 1993; Pea, 1980). Denial doesn't emerge until almost a year later, between 19 and 23 months (Pea, 1980). Cross-linguistic studies suggest that this stratification by type, with certain negative categories produced earlier than others, can be seen across languages (McNeill \& McNeill, 1968). Even after age 2, children continue to learn about negation, showing improvements in syntactic form (Klima \& Bellugi, 1966; CameronFaulkner, Lieven, \& Theakston, 2007) as well as increases in the frequency with which they produce spontaneous negatives (Pea, 1982). Furthermore, children as old as 4 years continue to have difficulty with implicitly negative terms such as marked adjectives (e.g. less) (Donaldson \& Balfour, 1968; Klatzky, Clark, \& Macken, 1973). Thus, although "no" is among the first words that children produce, they continue to grapple with the nuances of negation for several more years.

Nearly all prior research on the acquisition of negation has focused on production. Very little work has examined children's comprehension of negative sentences (cf. de Villiers \& Tager-Flusberg, 1975). While production can tell us about the contexts in which children use negation, it does not reveal the extent to which children understand concepts underlying negative sentences. Children may already have a sophisticated understanding of different types of negation before they start producing negative utterances. Alternatively, children's conceptual understanding may change as they develop linguistic negation. By examining the development of children's com-
prehension of negative sentences, we can begin to tease apart the relationship between children's conceptual understanding of negation and their linguistic abilities.

Our primary goal in this initial study was to address the lack of work on children's comprehension of negation. We conducted a study of children's understanding of negative sentences, using eye-tracking to test comprehension. Eyegaze measures are ideally suited to our goal, because gaze following requires limited cognitive resources (Fernald, Zangi, Portillo, \& Marchman, 2008). Because our ultimate goal is to understand the conceptual structure of negation, we measured comprehension of two types of negative sentences: those that refer to nothing (nonexistence), and those that refer to an alternative (similar to denial). By examining comprehension, we hoped to gather a more nuanced picture of the acquisition of negation as well as gain insight into children's conceptual understanding of different types of negative sentences.

\section*{Method}

This study was designed to examine the development of the comprehension of negation from ages \(2-4\) years. Children watched a video in which they were asked to "look at the boy with/with no X". This type of negative construct was used because it involves "no", the negative element emerging earliest in children's speech (Klima \& Bellugi, 1966; CameronFaulkner et al., 2007). Plural items were used instead of singular items to maintain maximum consistency between positive and negative sentences. Prior to each test trial, children viewed a context slide designed to set up expectations about the characters in the trial. This context was included due to work suggesting that contextual support facilitates negation processing in adults (Wason, 1965; Glenberg, Robertson, Jansen, \& Johnson-Glenberg, 1999; Lüdtke \& Kaup, 2006) as well as children (de Villiers, J. and Tager-Flusberg, H.B., 1975). Following each trial, Elmo appeared next to the target, to motivate children to look towards the correct character. In order to capture different types of negation, we created two between-subjects conditions. In the nothing condition, negative sentences referred to people with no items at all (e.g. a boy holding nothing compared to a boy holding apples). In the something condition, negative sentences referred to people with alternative items (e.g. a boy holding presents compared to a boy holding apples). By measuring children's comprehension of negative sentences in different contexts, we hoped to learn more about the types of negation that children understand between ages 2 and 4 .

\section*{Participants}

Families visiting the Children's Discovery Museum in San Jose, CA were invited to participate in this study. Our final sample was comprised of children who were exposed to English at least \(75 \%\) of the time, as indicated by their parents, and who were attentive for the initial calibration phase of the experiment. This resulted in a sample of 111 children, 49 2 -year-olds (mean age \(=2 ; 5\), range \(=2: 0-2 ; 11,22\) female) and 623 -year-olds (mean age \(=3 ; 5\), range \(=3 ; 0-3 ; 11\), 21


Figure 1: An example of context, trial, and feedback from the nothing and something conditions.
female). In exchange for participation, children were given a sticker and a certificate.

Of these initial 81 children, only those who completed at least 8 of the 16 trials were included in analysis. Four 2-yearolds and 53 -year-olds were rejected due to this criterion. A further 42 -year-olds and 43 -year-olds were rejected due to loss of gaze data in more than \(30 \%\) of the experiment. Finally, individual trials with more than \(30 \%\) missing gaze data were excluded from analysis. This left a total of 91 participants whose data was analyzed; 20 2-year-olds in the nothing condition and 21 in the something condition, and 263 -year-olds in the nothing condition and 27 in the something condition.

\section*{Stimuli}

We created 16 items, each presented as an individual trial. Items consisted of boys or girls either holding nothing or holding different items (e.g. two apples). Each trial was paired with a positive or a negative sentence. Sentences were of the form "Look at the boy who has/has no apples".

Each trial contained three parts: a context, a test trial, and feedback (see Figure 1):

Context: The context consisted of three characters, two holding two target items each, and the other character holding either nothing (in the nothing condition) or two alternative items (in the something condition). A pre-recorded voice said e.g. "See these boys?". Each context lasted 5 seconds.

Test trial: Each trial consisted of two new characters, one holding two target items and one either holding nothing (in the nothing condition) or two alternative items (in the something condition). The images were presented in silence for two seconds, after which a pre-recorded voice said a positive or a negative sentence (e.g. "Look at the boy with/with no apples"), followed by an additional tag sentence (e.g. "Can you find him?"). Each trial lasted 7.5 seconds.

Feedback: Feedback involved Elmo appearing next to the target character with a chiming noise lasting 1.5 seconds.


Figure 2: Proportion of 2-year-old and 3-year-old children looking to the target picture as the sentence unfolds. Nothing trials are shown above, and something trials are shown below. Error bars represent 95\% C.I.

\section*{Procedure}

Parents and children were led to a small research room. Children sat in a booster seat approximately 60 cm from the monitor of an SMI RED 120 Hz corneal reection eye-tracker mounted on an adjustable arm. Some children sat on a parent's lap, depending on the child's age and level of comfort.

The experiment was presented in the form of a short video. The video began with a short Elmo clip, during which any necessary adjustments to the eye-tracker were made. This was followed by a 2-point calibration and validation of the calibration points. After calibration, children were introduced to Elmo and told that "Today, Elmo is going to visit some of his friends. Do you want to meet Elmo's friends? Let's go!". This opening sequence was created to give the video a more "story-like" feeling, and to motivate children to look to the target characters during the test trials.

Following this introduction, children saw three gazecontingent practice trials, designed so that the video would not advance until the child looked at the target item. Practice trials involved only the trial + feedback (no context). The first practice item had only one character, while the next two practice items had two characters, as in the test trials.

The rest of the video consisted of 16 trials, as well as 6 filler pictures of Elmo and 4 Elmo video clips. Filler videos were advanced by the experimenter after a variable length of time depending on the child's attentiveness, making the video length slightly different for each child; in general the video lasted approximately 6 minutes. Two orders were created for the test videos, such that trial types were counterbalanced and trial order pseudo-randomized across the two orders.

Table 1: Coefficient estimates from mixed-effects models predicting proportion of looks to target in an early window (600-1600 ms after noun onset) and a late window (16002600 ms after noun onset).
\begin{tabular}{rrrr}
\hline & Coefficient Std. err. t value \\
\hline (Intercept) & 0.83 & 0.04 & 22.81 \\
Sentence (Negative) & -0.55 & 0.06 & -9.03 \\
Condition (Something) & -0.20 & 0.05 & -4.12 \\
Age (3-year-olds) & -0.05 & 0.05 & -1.15 \\
Window (Late) & -0.10 & 0.05 & -2.19 \\
Sentence \(\times\) Condition & 0.24 & 0.08 & 2.91 \\
Sentence \(\times\) Age & 0.29 & 0.08 & 3.63 \\
Condition \(\times\) Age & 0.17 & 0.06 & 2.56 \\
Sentence \(\times\) Window & 0.15 & 0.06 & 2.52 \\
Condition \(\times\) Window & 0.15 & 0.06 & 2.27 \\
Age \(\times\) Window & 0.09 & 0.06 & 1.39 \\
Sentence \(\times\) Condition \(\times\) Age & -0.35 & 0.11 & -3.19 \\
Sentence \(\times\) Condition \(\times\) Window & -0.19 & 0.08 & -2.28 \\
Sentence \(\times\) Age \(\times\) Window & -0.19 & 0.08 & -2.36 \\
Condition \(\times\) Age \(\times\) Window & -0.14 & 0.09 & -1.68 \\
Sentence \(\times\) Condition \(\times\) Age \(\times\) Window & 0.48 & 0.11 & 4.24 \\
\hline
\end{tabular}

\section*{Results and Discussion}

We first examined developmental changes in children's ability to comprehend negative sentences between ages 2 and 4. Next, we explored the contrast between types of negation (i.e. the nothing condition and the something condition). Finally, we examined how gaze changes over the course of a trial.

Developmental changes Children's ability to process negative sentences increases considerably between ages 2 and 3 . This increase can be seen both in children's accuracy and reaction time in response to negative sentences.

Figure 2 shows the proportion of children who looked to the target picture over the course of a trial. The majority of


Figure 3: Reaction times of children who were looking at the distractor picture at the onset of the noun to orient towards the target picture. Error bars represent \(95 \%\) C.I.

Table 2: Coefficient estimates from a mixed-effects model predicting RT to target picture following noun onset.
\begin{tabular}{rrrr}
\hline & Coefficient Std. err. t value \\
\hline (Intercept) & 597.28 & 119.99 & 4.99 \\
Sentence (Negative) & 1253.69 & 228.67 & 5.48 \\
Condition (Something) & 228.39 & 157.15 & 1.45 \\
Age (3-year-olds) & 9.40 & 150.89 & 0.06 \\
Sentence \(\times\) Condition & -579.63 & 312.21 & -1.86 \\
Sentence \(\times\) Age & -969.19 & 297.13 & -3.26 \\
Condition \(\times\) Age & -184.94 & 198.96 & -0.93 \\
Sentence \(\times\) Condition \(\times\) Age & 560.83 & 416.24 & 1.35 \\
\hline
\end{tabular}
children in both age groups and conditions responded correctly to positive sentences. However, a difference in accuracy can be seen in response to negative sentences between the two age groups. While 2 -year-olds show very little comprehension of negation in this paradigm, 3-year-olds show a noticeable increase in looks to target following the onset of the noun. We ran a linear mixed-effects model analyzing the effects of sentence type, condition, age group, and time window (early: 600-1600ms following noun onset; late: 16002600 ms following noun onset) on the proportion of children looking to the target (Table 1). \({ }^{1}\) Results of this model indicate a significant interaction between sentence type and age group, such that 3-year-olds are more likely to look to the target in response to negative sentences than 2-year-olds.

Reaction time (RT) was measured by looking at trials in which children who were originally fixating on the distractor (non-target) picture at the onset of the noun, and calculating how long it took these children to make their first shift to the target picture (Fernald et al., 2008). Two-year-olds showed larger RTs in response to negative sentences compared to positive sentences (Figure 3). However, 3-year-olds were surprisingly quick to orient to the target picture, only slightly slower than in response to positive sentences and much faster than

\footnotetext{
\({ }^{1}\) All mixed-effects models were run using the Ime4 package in R version 2.15.2. The random effects structure for this model was as follows: \((\) sentence + window|subject \()+(\) sentence + condition + age group + window \(\mid\) item \()\)
}

2-year-olds. Results from a linear mixed-effects model are reported in Table 2. \({ }^{2}\) Note that the decrease in 3-year-olds' RTs in response to negative sentences is not due to a general increase in processing abilities; our model found no main effect of age, only an interaction between sentence type and age, such that 3-year-olds process negative sentences nearly a full second faster than 2 -year-olds.

Types of negation Our results suggest that children have more difficulty identifying the referent of a negative sentence when it refers to nothing than when it refers to an alternative object. While 3-year-olds increase their looks to target following noun onset in the nothing condition, this does not increase above \(50 \%\) (Figure 2). However, in the something condition, nearly \(70 \%\) of children look to the target following negative sentences. This increase is seen in the later window of time, \(1600-2600 \mathrm{~ms}\) following the onset of the noun. This can be seen in the results of our model reported in Table 1; there is a significant 4 -way interaction such that 3 -year-olds' responses to negative sentences in the later window of the something condition show an increase in looks to target.

Onset-contingent plots (Figure 4) provide another way of looking at children's gaze behavior. These plots split trials based on whether the child was looking at the target or the distractor at the onset of the noun, and plot the proportion of children who shift their gaze to the opposite item. Children who are initially looking at the distractor should show rapidly increasing shifts, whereas children who are initially looking at the target should continue to look at the target (Fernald et al., 2008). Note that responses to the positive sentences are typical of what these plots normally look like.

Responses to the negative sentences, however, deviate from the typical pattern. For 2-year-olds in both conditions and 3-year-olds in the nothing condition, the pattern seen is the reverse: if children are looking at the target picture, they orient away, and if they are on the distractor, they stay. 3-year-olds are slightly better, with about \(50 \%\) orienting away from the distractor, but still the majority orient away from the target. However, 3-year-olds in the something condition show a different pattern; initially, children continue to fixate on the distractor and shift away from the target, but after approximately 1600 ms this pattern reverses and children shift away from the distractor and back towards the target. Thus, it is only 3-year-olds (and only in the something condition) who exhibit increased looks to target in response to negative sentences.
Real-time processing of negative sentences The data here reveal that children show a tendency to initially orient away from the target object, looking towards the negated noun, even amongst children who eventually do look to the target.

Two-year-olds do not look to the referent of negative sentences in either condition. Note that in the nothing condition, this preference could be explained by children's lack of interest in the boy with nothing, but in the something condi-

\footnotetext{
\({ }^{2}\) The random effects structure for this model was as follows: \((\) sentence \(\mid\) subject \()+(\) sentence + condition + age group \(\mid\) item \()\)
}


Figure 4: Onset-contingent plots of children's looking behavior starting at the onset of the noun. Nothing trials are shown above, and something trials are shown below.
tion both characters are equally salient. Thus, it appears that when 2-year-olds hear a named noun, they prefer to look at that noun, even if it has been negated.

In the earlier window of Figure 2 (600-1600 ms following the onset of the noun), 3-year-olds in the something condition show a similar pattern, showing a preference to look incorrectly to the negated noun. In the later window, however, the opposite pattern is seen: now children appear to look reliably to the target picture. This pattern can be seen more clearly in Figure 3. If children are looking at the distractor at the onset of the noun, about \(50 \%\) linger until 1600 ms , when suddenly the majority of looks shift to the target. Conversely, if children are looking at the target at noun onset, the majority of children shift away from the target, fixating on the distractor until 1600 ms , when they look back to the target. Again, this indicates a tendency for children to initially look to the negated noun, even when both options are equally salient.

\section*{General Discussion}

Little is known about children's comprehension of negative sentences, a surprising fact given the universality of negation and its importance in logical reasoning. Previous work on production suggests that children are continuing to learn about negation between 2 and 3 years of age (Klima \& Bellugi, 1966; Cameron-Faulkner et al., 2007; Pea, 1982), but few studies have explored how children's comprehension changes over this period. We conducted a study of children's comprehension of negation, examining negative sentences that referred either to nothing or to an alternative. We found that 3-year-olds were much faster and more successful than 2-year-olds at correctly looking at the referent of negative sentences. In addition, we found that children at both
ages struggle to identify the referent of negatives that refer to nonexistence, as opposed to referring to an alternative.

An additional and surprising finding of this study was that children in the something condition had an initial tendency to look towards the negated noun, and only 3 -year-olds were able to eventually override this preference and look to the correct target. There is some evidence that a similar pattern occurs in adult processing of negative sentences. Several priming studies have found that the representation of a negative sentence changes in the moments after the sentence unfolds (Kaup \& Zwaan, 2003; Kaup, 2001; Kaup, Ludtke, \& Zwaan, 2006; Hasson \& Glucksberg, 2006). In addition, ERP studies have shown N400 activation, associated with the processing of a semantically unexpected word, in sentences where the unexpected noun is negated (Fischler, Bloom, Childers, Roucos, \& Perry Jr, 1983; Lüdtke, Friedrich, De Filippis, \& Kaup, 2008). That is, sentences such as "A robin is a truck" and "A robin is not a truck" show greater negativity at the N400 than sentences such as "A robin is/is not a bird" (Fischler et al., 1983). This work has been interpreted as suggesting that adults do not immediately integrate negative elements into sentence meaning. Our findings here suggest that this may be true for children as well.

In our sample, both 2- and 3-year-olds found looking to the correct referent difficult when the target was holding nothing, i.e. nonexistence negation. It seems incorrect to attribute this to a lack of understanding of nonexistence, due to children's early production of negation. A more likely explanation is that orienting to the target in these trials required greater inhibitory control because the target character is less interesting to children. Identifying the specific kinds of negation that children hear in naturalistic contexts can help us under-
stand what kinds of contexts might facilitate comprehension of negation.

Overall, this study of children's comprehension of negation provides a complement to previous work on the acquisition of negation, which has primarily focused on production. Our ultimate goal is to examine young children's understanding of different negative concepts, and how this conceptual structure is influenced by the acquisition of linguistic negation. This goal speaks to a broader question about the extent to which linguistic development influences conceptual development and vice versa. Negation is an important case study for examining this question, because linguistic negation emerges early in childhood and can therefore be studied in conjunction with children's understanding of negative concepts. By exploring this relationship, we hope to shed light not only on the acquisition of negation, but also on the ways that language and concepts influence each other throughout development.

\section*{Acknowledgments}

Special thanks to the staff and families at the San Jose Childrens Discovery Museum.

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors(s) and do not necessarily reflect the views of the National Science Foundation.

\section*{References}

Bloom, L. (1970). Language development: Form and function in emerging grammars. New York, NY: MIT Research Monograph, No 59.
Bloom, L. (1993). Language development from two to three. New York, NY: Cambridge Univ Press.
Cameron-Faulkner, T., Lieven, E., \& Theakston, A. (2007). What part of no do children not understand? A usage-based account of multiword negation. Journal of Child Language, 34(2), 251-282.
Carpenter, P., \& Just, M. (1975). Sentence comprehension: A psycholinguistic processing model of verification. Psychological Review, 82(1), 45-73.
Clark, H., \& Chase, W. (1972). On the process of comparing sentences against pictures. Cognitive Psychology, 3(3), 472-517.
de Villiers, J. and Tager-Flusberg, H.B. (1975). Some facts one simply cannot deny. Journal of Child Language, 2, 279-286.
Donaldson, M., \& Balfour, G. (1968). Less is more: A study of language comprehension in children. British Journal of Psychology, 59(4), 461-471.
Fernald, A., Zangi, R., Portillo, A., \& Marchman, V. A. (2008). Looking while listening: Using eye movements to monitor spoken language comprehension by infants and young children. In I. Sekerina, E. Fernandez, \& H. Clahsen (Eds.), Developmental psycholinguistics: Online methods
in children's language processing (p. 98-135). Philadelphia, PA: John Benjamins Publishing Company.
Fischler, I., Bloom, P., Childers, D., Roucos, S., \& Perry Jr, N. (1983). Brain potentials related to stages of sentence verification. Psychophysiology, 20(4), 400-409.
Glenberg, A., Robertson, D., Jansen, J., \& Johnson-Glenberg, M. (1999). Not propositions. Journal of Cognitive Systems Research, 1, 19-33.
Hasson, U., \& Glucksberg, S. (2006). Does understanding negation entail affirmation? an examination of negated metaphors. Journal of Pragmatics, 38(7), 1015-1032.
Just, M., \& Carpenter, P. (1971). Comprehension of negation with quantification. Journal of Verbal Learning and Verbal Behavior, 10, 244-253.
Just, M., \& Carpenter, P. (1976). Eye fixations and cognitive processes. Cognitive Psychology, 8, 441-480.
Kaup, B. (2001). Negation and its impact on the accessibility of text information. Memory \& Cognition, 29(7), 960-967.
Kaup, B., Ludtke, J., \& Zwaan, R. (2006). Processing negated sentences with contradictory predicates: Is a door that is not open mentally closed? Journal of Pragmatics, 38, 1033-1050.
Kaup, B., \& Zwaan, R. (2003). Effects of negation and situational presence on the accessibility of text information. Journal of Experimental Psychology: Learning, Memory, and Cognition, 29(3), 439-446.
Klatzky, R., Clark, E., \& Macken, M. (1973). Asymmetries in the acquisition of polar adjectives: Linguistic or conceptual? Journal of Experimental Child Psychology, 16(1), 32-46.
Klima, E., \& Bellugi, U. (1966). Syntactic regularities in the speech of children. In R. W. J. Lyons (Ed.), Psycholinguistic papers (Vol. 3765, p. 183-208). Edinburgh: Edinburgh University Press.
Lüdtke, J., Friedrich, C., De Filippis, M., \& Kaup, B. (2008). Event-related potential correlates of negation in a sentencepicture verification paradigm. The Journal of Cognitive Neuroscience, 20(8), 1355-1370.
Lüdtke, J., \& Kaup, B. (2006). Context effects when reading negative and affirmative sentences. In Proceedings of the 28th Annual Conference of the Cognitive Science Society (p. 1735-1740).

McNeill, D., \& McNeill, N. (1968). A question in semantic development: What does a child mean when he says "no"? Ann Arbor, MI: Michigan University Center for Research on Language and Language Behavior.
Pea, R. (1980). The development of negation in early child language. In D. Olson (Ed.), The social foundations of language and thought (pp. 156-186). New York, NY: Norton New York.
Pea, R. (1982). Origins of verbal logic: Spontaneous denials by two-and three-year olds. Journal of Child Language, 9(3), 597-626.
Wason, P. (1965). The contexts of plausible denial. Journal of Verbal Learning \& Verbal Behavior, 4(1), 7-11.

\title{
Limitations and chances of working memory training
}

\author{
Daniela Nussbaumer (nussbaumer@ifv.gess.ethz.ch) \\ Institute for Behavioral Sciences, ETH Zurich, Universitätsstrasse 41 8092 Zurich, Switzerland
}

\author{
Roland H. Grabner (roland.grabner@psych.uni-goettingen.de) \\ Department of Psychology, Georg-August-University of Göttingen, Waldweg 26 \\ 37073 Göttingen, Germany
}

\author{
Michael Schneider (m.schneider@uni-trier.de) \\ Department of Psychology, University of Trier, Universitätsring 15 \\ 54296 Trier, Germany
}

Elsbeth Stern (stern@ifv.gess.ethz.ch)
Institute for Behavioral Sciences, ETH Zurich, Universitätsstrasse 41
8092 Zurich, Switzerland

\begin{abstract}
Recent studies show controversial results on the trainability of working memory (WM) capacity being a limiting factor of human cognition. In order to contribute to this open question we investigated if participants improve in trained tasks and whether gains generalize to untrained WM tasks, mathematical problem solving and intelligence tests.
83 adults trained over a three week period ( 7.5 hours total) in one of the following conditions: A high, a medium or a low WM load group. The present findings show that task specific characteristics could be learned but that there was no transfer between trained and untrained tasks which had no common elements. Positive transfer occurred between two tasks focusing on inhibitory processes. It might be possible to enhance this specific component of WM but not WM capacity as such. A possible enhancement in a learning test is of high educational interest and worthwhile to be investigated further.
\end{abstract}

Keywords: working memory; training; intelligence; inhibition

\section*{Theoretical Background}

The concept of WM has received much attention lately by various psychological disciplines for its importance as a basis of human intelligence and as a limiting factor of human cognition. WM can be seen as a cognitive system for simultaneously storing and manipulating information, and hence strongly relates to reasoning abilities and the handling of novel information (Baddeley \& Hitch, 1974). Also the attention to goal-relevant information and inhibition of irrelevant information are important functions of WM.

High correlations between WM capacity and intelligence (Oberauer, Süss, Wilhelm, \& Wittmann, 2008), notably when measured by Matrices Tests (e.g. Advanced Progressive Matrices Test; Raven, 1990) as well as high correlations between WM capacity and applied fields, e.g. mathematical problem solving tasks leave the following open question: What happens to intelligence and
mathematical problem solving skills when WM capacity potentially gets enhanced? One possibility could be a likewise enhancement of WM and intelligence (and mathematical problem solving skills). The similarity of the two concepts would make far transfer plausible. But as stated earlier, results are controversial and more evidence is needed.

Early studies were positive in judging the possibility of a WM training being able to enhance WM capacity and performance in related fields. These early studies were also more explorative in nature. Later studies took criticism (Moody, 2009; Sternberg, 2008) into consideration and the complexity of study designs has been raised (for example by Redick et al. (2012) a non-replication of the study by Jaeggi, Buschkuehl, Jonides, \& Perrig, (2008)). In the current study the following criticisms of the past studies are taken into consideration and examined: a) inclusion of an active control group, b) administering a wide variety of transfer tasks and c) examining long term effects.

The trainability of WM capacity would mean that we are able to broaden an important limiting factor of human cognition and this would be of highly practical as well as of seminal educational relevance. There is a growing body of WM training literature (Chein \& Morrison, 2010; Klingberg, 2010; Shipstead, Redick, \& Engle, 2012). Melby-Lervåg and Hulme (2012) conducted a meta-analytic study and compared effects: Across training studies, effects vary in whether WM training paradigms are effective in improving cognitive abilities.

We included three training groups: a high, a medium and a high WM load group. Their training differed in the amount as well as in the type of WM load included. The first two groups focused on resolution of proactive interference - an ability tapping the WM subcomponent of inhibition, which is regarded as critical subcomponent of WM (Friedman \& Miyake, 2004). The third group was an active control group (low to zero WM load) solving a control reaction time task. The further manipulations
referred to whether the task was adaptive and whether the task was dual. If WM load during training is the crucial factor for transfer effects to occur there should of course be no training gains for control groups and gains should be more pronounced for a high than for a medium WM load group. The advantage of a graded design lies in being able to differentiate whether a transfer gain can be attributed to enhanced WM capacity or not.

The inclusion of a wide variety of transfer tasks is necessary to decide whether changes can be attributed to an enhancement of WM capacity or merely to task specific learning because an enhancement of WM capacity can only be demonstrated if transfer occurs generally and is not limited to single tasks (Shipstead et al., 2012). In the present study transfer to an untrained WM task is referred to as near transfer and transfer to tasks with another cognitive demand than WM is categorized as far transfer. Far transfer is typically measured using intelligence tests as well as other reasoning tests. In addition to intelligence tests, in the present study mathematical tasks are administered to assess possible far transfer to school-related abilities. According to a literature review by Raghubar, Barnes, and Hecht (2010), WM and skills in mathematical problem solving are highly correlated, in particular mental arithmetic, and are therefore suitable as transfer tasks.

In sum, the main goal of the current study is to test a) whether a WM training yields near transfer, an enhancement of performance in untrained WM tasks, and b) to systematically test whether such a potential WM enhancement can provoke far transfer in the domain of intelligence and mathematical problem solving and whether such an enhancement is depending on the amount of WM load during training. Further we investigate to what extent training gains are found in an active control group, to what extent enhancement of performance is dependent on to the level of WM load during training and how stable these effects are.

\section*{Method}

\section*{Participants}

A total of 83 healthy students of science- and humanitiesrelated fields from three Swiss universities completed the study ( \(M_{\text {age }}=23.7, S D=3.3\) ). Eight participants dropped out due to installation problems of the training software on their home computer ( 5 participants) or due to non-adherence to the training paradigms or sessions at the institute (3 participants).

\section*{Procedure}

Participants were randomly designed to one of three groups: A high, a medium or a low WM load group. All groups trained during a three week period five days a week for half an hour on their home computer, resulting in a total training time of 7.5 hours. The first and the last training session were completed at the institute in order to ensure understanding of the tasks and to control for the correct
handling of the training software. Solution rates and times as well as other parameters were logged by the training software for all sessions. Before and after training, two assessment sessions took place at the first author's institute: An individual and a group session where participants had to solve WM tasks and a mental arithmetic task, mathematical problem solving tasks and intelligence tests. The sessions before training served to assess baseline performance and the sessions after training aimed to assess possible transfer from the WM training. In order to make an intervention and a possible enhancement meaningful it should show an impact over a certain time. Long term training effects were assessed by a follow-up testing session after a three months period. Participants again solved trained tasks as well as paralleled versions of untrained tasks.
The three groups did not differ significantly in their initial intelligence level, in demographical factors (age, sex, field of study) and in personality factors (measured by the NEOFFI (Costa \& McCrae 1992)). Their initial performance of training and transfer tasks was also in the same range and didn't differ significantly between groups.

\section*{Material}

\section*{Training}

A high WM load group trained a dual version of the nback task, similar to Jaeggi et al. (2008). Simultaneously, letters were presented orally and squares visually at different positions on the screen. Participants had to indicate whether the letter and the position \(n\) trials back was the same or not. This adaptive and dual version of the n-back task placed high WM load because a large amount of interference trials was incorporated. Also the duality of the task adds to the high WM load level. Through the dual nature of the task participants trained the visual and oral domain simultaneously. Participants worked on the task for 30 minutes per day with the size of \(n\) adapted to the actual level of performance. In this group the average n-back level was assessed.

The medium WM load group trained with three nonadaptive WM tasks: A three-back task with letters and the following two recognition tasks. In the face recognition task participants had to decide whether a single face was part of a previously presented set of four faces or not. In the letter recognition task participants had to decide whether a letter was part of the previously presented set of four letters or not. The tasks were characterized by moderate WM load with a focus on resolution of proactive interference in WM. Solution times and rates were measured and each task was performed for 10 minutes. In all three tasks, a high level of interference was produced by incorporating a large amount of lure trials, i.e., trials in which the objects were shown in another trial than the one actually referred to.
The low WM load group trained similar tasks as the medium WM load group, but with a very low WM load. Participants had to solve a 1-back task and for the recognition tasks participants had to compare one face/letter
with a previously presented single face/letter. Solution time and rates were measured.

The amount of WM load is not the only variation between the three groups but all other differences as for example duality vs. singularity of the task can also be seen as a variation of the level of WM load.

\section*{Tasks to assess near transfer}

The four WM transfer tasks each represented a different subcomponent of WM and showed varying similarities to the trained tasks. In the complex span task participants solved simple equations while keeping single letters in mind. At task switching participants had to either decide whether the value of a three digit number was below or above 500 , or whether the number was even or odd. In a monitoring task participants had to detect changes in a grid of nine three-digit-numbers and react on certain constellations of same final digits. A forth WM transfer task was kept very similar to a trained task of the medium WM load group. In this so called 'pseudowords'-task, participants of the medium WM load group had to accomplish the same task requirements as in their trained letter recognition task and also the trained face recognition task was very similar. The mentioned transfer and training tasks showed the same surface structure but other content material than the trained task: Recognition of pseudowords, nonsense syllables obeying phonetic rules, in the transfer situation instead of the trained recognition of single letters or faces. For the control group the same was true except that they trained task versions with minimal WM load. The high WM load group on contrary had no correspondent training.

\section*{Tasks to assess far transfer}

As fluid intelligence tests the Advanced Progressive Matrices Test (APM, Set II) by Raven (1990) and the 'Intelligenz-Struktur-Test' (I-S-T 2000 R) by Amthauer, Brocke, Liepmann, and Beauducel (2001) were administered.

The three transfer tasks of the mathematical domain consisted of different levels of reasoning requirements and complexity. A mental arithmetic task with subtractions of two digit numbers with carries was conducted without participants taking any notes. A so-called mathematics test (Mathematik-Test, Ibrahimovic \& Bulheller, 2005) was exhibited to test participants’ ability to solve mathematical word problems. In a last mathematical task with high WM load participants had to keep in mind three simple but interlinked equations as well as the value of the three unknowns.

\section*{Learning of novel material}

We further investigated whether WM training can enhance the learning of novel material. In this task participants learned to calculate in the septimal system (base 7 system) while inhibiting their usual counting routines of the decimal system. This learning task was presented immediately after the last training session and in order to assess the learning of new principles and the establishing of new routines while overcoming well-trained ones the task comprised of a 40 min problem solving period with a total
of 150 trials of additions in the septimal system. This design enables us to investigate the possibility of not only having WM training enhance certain untrained WM tasks, but also enhance the chance of grasping and administering new principles and rules. This would be new and very tempting for educational purposes.

\section*{Results}

In the medium and high WM load groups the implementation check was positive in that through training participants enhanced their performance significantly in trained tasks. The medium WM load group showed significant increases in solution time and solution rate in the three trained tasks. The high WM load group showed significant increase in the average n-back-level. Participants of the low WM load group also significantly increased their solution times, but not solution rates (see Appendix, Table 1).

Enhancements specific to groups occurred in two of nine untrained tasks. First, group specific enhancements occurred in the 'pseudowords' task. An ANOVA for reaction time measures with the between subject factor group and the within subject factor time was conducted. A significant interaction between time and group \((F(2.99,118.09)=\) 22.92, \(p<.001, \eta^{2}{ }_{p}=.37\); see also Figure 1 and Appendix Table 2) and pairwise comparisons revealed that the medium and low WM load groups likewise accelerated more than the high WM load group, which showed only slight enhancement. This analysis also showed a significant main effect time as all participants got faster ( \(F(1.50\), \(\left.118.09)=136.18, p<.001, \eta_{p}^{2}=.63\right)\).


Figure 1: Course of solution times in the transfer task 'pseudowords' during and after training. *** indicate time periods with group specific significant main effects, \(p<\) .001. Please note, that there is also a significant interaction between time (pre-, post and follow-up-test) and the three groups.

The second differential transfer gain occurred in the one learning task. In the 'Base7' task groups varied in their
amount of gain manifesting in a significant interaction between groups and beginning versus end of the test \(\left(F(11.19,447.74)=2.15, p<.05, \eta_{p}^{2}=.05\right)\). Post hoc tests showed that the high WM load group showed a higher degree of progress than the low WM load group over the course of the 150 trials (see Figure 2). In this task all groups enhanced their performance significantly over the 150 trials \(\left(F(5.60,447.74)=32.67, p<.001, \eta_{p}^{2}=.29\right)\).

No differential transfer occurred in any of the mathematical problem solving tasks or in the intelligence tests, therefore in none of the untrained WM tasks such interactions were found. However, there was a significant temporary enhancement of solution times and rates for all transfer tasks, but with no difference between the three groups, as no interaction was detected (see table 2 for changes from pre- to post-measure).

Long term gains over a three month period were found in some tasks, but no differences between the three groups were found. The only exception is the aforementioned recognition task 'pseudowords', where differential changes between pre- and post-tests could be held throughout the three months.


Figure 2: Course of solution rates in the learning task 'Base7'. The y -axis represents solution rate ( \(1.0=\) maximum, \(0.5=50 \%\) right). The \(x\)-axis marks the 15 parts of the test, 10 trials are summarized into one part. Please note, that there is a significant interaction between part (1-

15 ) and group (low, medium and high WM load).

\section*{Discussion}

The actual focus of the present study are the questions to what extent training gains are found in an active control group, to what extent enhancement of performance is dependent on to the level of WM load during training, what near and far transfer effects can be observed and how stable these effects are. Only in two out of nine cases were such differential enhancements found. In all other untrained WM tasks no such interactions were found so that no positive transfer specific to medium or high WM load during training occurred in any of the mathematical problem
solving tasks or in any intelligence tests. According to Shipstead et al. (2012) it is crucial to compare a wide variety of tasks to decide where the reason of changes may lie. Generally occurring transfer effects could be attributed to an enhancement of WM capacity whereas rare transfer should be explained by only task specific learning. For the present study it can therefore be concluded that no enhancement of WM capacity as such is found.
The two cases of group specific enhancement are discussed separately. First, differential enhancement occurred in one untrained task with a similar surface structure but different content material than in the trained tasks of the medium WM load group (recognition of 'pseudowords' instead of recognition of letters). Also the low WM load group trained a recognition task with the same surface structure but minimal WM load. Both the medium and low WM load group developed similarly, this suggests that high WM load was not essential for the development, but rather the similarity of the trained and untrained task. This explains why the development of the low WM load group was likewise the one of the medium WM load group and why the high WM load group training with a very different paradigm but being exposed to high WM load during training - developed in a different way. In conclusion, training gains can transfer to very similar tasks only. The similarity of the tasks or in other words the common elements of trained and untrained tasks are crucial for transfer.
Second, in the 'Base7' learning task all groups enhanced their performance significantly over the 150 trials, but the high WM load group showed a higher progress than the low WM load group. The trained dual n-back task of the high WM load group and the 'Base7' task at their surface show no similarity but both tasks particularly focused on inhibitory processes. It can therefore be concluded, that inhibitory processes could possibly be enhanced through a specific training focusing on inhibition. In order to exclusively answer this assumption more evidence would be needed to exclude the possibility of just task specific characteristics being responsible for this result. Moreover, this gain was measured in a novel task type: A learning task - to our knowledge not administered in any other WM training study and the only learning task included in the present WM training study. The possible enhancement in a learning test is of high educational interest and also has to be verified by further testing.
There was significant general enhancement of solution times and solution rates for all transfer tasks, but no difference between the three groups for seven of nine tasks. As also the low WM load group with virtually zero WM load during training got significantly better from pre- to post-test, enhancement cannot be explained by expanded WM capacity. It can therefore be stated that participants perform significantly better after WM training but not due the characteristics of the training and not due to an enhancement of WM capacity.

Numerous authors (Chein \& Morrison, 2010; Moody, 2009; Shipstead, Redick, \& Engle, 2010; Shipstead et al., 2012; Sternberg, 2008) judge the selection of an appropriate control group as essential in interpreting data. In the present study the low WM load group served as an active control group and also increased their performance. Through an active control group effects due to a different degree of study involvement can be ruled out.

Long term gains over a three month period were found in some tasks, but no differences between the three groups were found. The only exception is the aforementioned recognition task where the differential changes between pre and post testing could be held throughout the three months.

In summary, the theoretical and educational significance of the present results are threefold. First, our results suggest that WM training is of limited use to enhance human cognition in general. The present findings show that task specific characteristics could be learned but that there was no transfer between trained and untrained tasks which had no common elements. Second, as positive transfer occurred between two tasks focusing on inhibitory processes, it might be possible to enhance this specific component of WM. Third, the possible enhancement in a learning test is of high educational interest and is worthwhile to be further investigated.

\section*{References}

Amthauer, R., Brocke, B., Liepmann, D., \& Beauducel, A. (2001). Intelligenz-Struktur-Test 2000, R. Göttingen: Hogrefe.
Baddeley, A., \& Hitch, G. (1974). Working memory. In G. Bower (Ed.), Recent advances in learning and motivation (Vol. 8). New York: Academic Press.
Chein, J. M., \& Morrison, A. B. (2010). Expanding the mind's workspace: Training and transfer effects with a complex working memory span task. Psychonomic Bulletin \& Review, 17(2), 193-199.
Costa, P. T. \& McCrae, R. R. (1992). Revised NEO Personality Inventory (NEO PI-R) and Neo Five-Factor Inventory (NEO-FFI). Odessa, FL: Psychological Assessment Resources.
Friedman, N. P., \& Miyake, A. (2004). The relations among inhibition and interference control functions: a latentvariable analysis. Journal of Experimental Psychology: General, 133(1), 101-135.
Ibrahimovic, N., \& Bulheller, S. (2005). Mathematiktesttest für Ausbildung und Beruf - MATHE (2. Überarbeitete, erweiterte und neunormierte Auflage). Frankfurt/Main: Harcourt Test Services.
Jaeggi, S. M., Buschkuehl, M., Jonides, J., \& Perrig, W. (2008). Improving fluid intelligence with training on working memory. Proceedings of the National Academy of Sciences, 105(19), 6829-6833.
Klingberg, T. (2010). Training and plasticity of working memory. Trends in Cognitive Sciences, 14, 317-324.

Melby-Lervåg, M., \& Hulme, C. (2012). Is Working Memory Training Effective? A Meta-Analytic Review. Developmental Psychology. doi: 10.1037/a0028228
Moody, D. E. (2009). Can intelligence be increased by training on a task of working memory? Intelligence, 37, 327-328.
Oberauer, K., Süss, H. M., Wilhelm, O., \& Wittmann, W. W. (2008). Which working memory functions predict intelligence? Intelligence, 36(6), 641-652.
Raghubar, K. P., Barnes, M. A., \& Hecht, S. A. (2010). Working memory and mathematics: A review of developmental, individual difference, and cognitive approaches: Perspectives on Math Difficulty and Disability in Children. Learning and Individual Differences, 20(2), 110-122.
Raven, J. C. (1990). Advanced progressive matrices sets 1 and 2. Oxford: Oxford Psychologists Press.
Redick, T. S., Shipstead, Z., Harrison, T. L., Hicks, K. L., Fried, D. E., Hambrick, D. Z., et al. (2012). No evidence of intelligence improvement after working memory training: A randomized, placebo-controlled study. Journal of Experimental Psychology: General. doi: 10.1037/a0029082

Shipstead, Z., Redick, T. S., \& Engle, R. W. (2012). Is working memory training effective? Psychological Bulletin, 138(4), 628-654.
Sternberg, R. J. (2008). Increasing fluid intelligence is possible after all. Proceedings of the National Academy of Sciences, 105(19), 6791-6792.

\section*{Appendix}

Table 1: Data for each training task and separately for solution time and rate reporting an ANOVA (within-subject factor time: 13 sessions). Cohen's \(d\) was derived from comparisons between the first and the last session.
\begin{tabular}{lll}
\hline Solution time & ANOVA: 13 sessions & Cohen's \(\boldsymbol{d}\) \\
\hline Low WM load group & \\
\hline 1-back & \(F(5.83,134.02)=7.82 p<\) & \(d=-0.80\) \\
& .001 & \\
& \(\eta_{p}^{2}=.25\) & \\
\hline 1-Face & \(F(5.35,112.24)=5.108 \mathrm{p}<\) & \(d=-0.73\) \\
& 0.001 & \\
& \(\eta_{\mathrm{p}}^{2}=0.20\) & \\
\hline 1-Letter & \(F(4.54,104.50)=10.60 \mathrm{p}<\) & \(d=-0.91\) \\
& 0.001 \\
& \(\eta_{\mathrm{p}}^{2}=0.32\) & \\
\hline Medium WM load group & \\
\hline 3-back & \(F(4.57,109.58)=13.56, \mathrm{p}\) & \(d=-1.18\) \\
& \(<0.001\) & \\
\hline 4-Faces & \(\eta_{\mathrm{p}}^{2}=0.36\). & \(F(4.31,103.49)=20.82 p<\) \\
& 0.001 & \(d=-1.40\) \\
& \(\eta_{p}^{2}=0.45\) & \\
\hline 4-Letters & \(F(5.24,125.86)=11.14\) & \(d=-0.84\) \\
& \(\mathrm{p}<0.001\) & \\
& \(\eta_{p}^{2}=0.32\) & \\
\hline
\end{tabular}
\begin{tabular}{|c|c|c|c|}
\hline Solution rate & \multicolumn{2}{|l|}{ANOVA: 13 sessions} & Cohen \\
\hline \multicolumn{4}{|l|}{Low WM load group} \\
\hline 1-back & \multicolumn{2}{|l|}{n.s} & \(d=0.47\) \\
\hline 1-Face & \multicolumn{2}{|l|}{n.s.} & \(d=-0.32\) \\
\hline 1-Letter & \multicolumn{2}{|l|}{n.s.} & D \\
\hline \multicolumn{4}{|l|}{Medium WM load group} \\
\hline 3-back & \multicolumn{2}{|l|}{\[
\begin{aligned}
& F(2.64,63.24)=12.53 \\
& p<.001, \\
& \eta_{p}^{2}=.34
\end{aligned}
\]} & \(d=1.3\) \\
\hline 4-Faces & \multicolumn{2}{|l|}{\[
\begin{aligned}
& F(12,31)=2.00 \\
& p<0.05 \\
& \eta_{p}^{2}=0.07
\end{aligned}
\]} & \(d=0\). \\
\hline 4-Letters & \multicolumn{2}{|l|}{n.s.} & \\
\hline \multicolumn{4}{|l|}{High WM load group} \\
\hline Dual-Nback & \multicolumn{2}{|l|}{\[
\begin{aligned}
& F(3.63,101.70)=29.23 \\
& p^{2}<0.001 \\
& \eta_{p}^{2}=0.51
\end{aligned}
\]} & \(3 \quad d=1.76\) \\
\hline \multicolumn{4}{|l|}{Table 2: Transfer data for each task reporting main and interaction effects for an ANOVA (betweensubject factor group: low, medium and high load and within-subject factor time: pre-, post-, and follow-up-testing)} \\
\hline Pre - Post & Main effect time & \begin{tabular}{ll} 
Main & I \\
effect & ti \\
group & g
\end{tabular} & \begin{tabular}{l}
Interaction time * \\
group
\end{tabular} \\
\hline ABC task & \multirow[t]{2}{*}{\[
\begin{aligned}
& F(1,80)= \\
& 113.49 \\
& p<0.001 \\
& \eta^{2} p=0.59 \\
& d=-0.83
\end{aligned}
\]} & & n.s. \\
\hline Solution time & & \(\eta_{p}^{2}=0.02\) & \(\eta_{p}^{2}=0.03\) \\
\hline Pseudowords & \multirow[t]{2}{*}{\[
\begin{aligned}
& \mathrm{F}(1,80)= \\
& 122.82 \\
& p<0.001 \\
& \eta^{2} p=0.61 \\
& d=-1.09
\end{aligned}
\]} & \[
\begin{array}{ll}
F(2,80)= & F \\
2.12
\end{array}
\] & \[
\begin{aligned}
& F(2,80)= \\
& 6.95
\end{aligned}
\] \\
\hline Solution time & & \(p<0.05\)
\(\eta^{2}=0.10\) & \(p<0.01\)
\(\eta^{2}{ }_{p}=0.15\) \\
\hline Task & \multirow[t]{3}{*}{\[
\begin{aligned}
& \mathrm{F}(1,79)= \\
& 136.01 \\
& p<0.001 \\
& \eta^{2}=0.63 \\
& d=-0.78 \\
& \hline
\end{aligned}
\]} & \(F(2,79)=\) & \(F(2,79)=\) \\
\hline Switch & & 7.03 & 3.31 \\
\hline Solution time & & \[
\begin{aligned}
& p<0.01 \\
& \eta_{p}^{2}=0.15
\end{aligned}
\] & \[
\begin{aligned}
& p<0.05 \\
& \eta_{p}^{2}=0.08
\end{aligned}
\] \\
\hline Monitoring & \multirow[t]{2}{*}{\[
\begin{aligned}
& F(1,80)= \\
& 28 \\
& p<0.001 \\
& \eta_{p}^{2}=0.26 \\
& d=-0.68
\end{aligned}
\]} & n.s. & n.s. \\
\hline Solution time & & \(\eta_{p}^{2}=0.03\) & \(\eta_{p}^{2}=0.01\) \\
\hline \multirow[t]{3}{*}{Mental Arithmetics Solution time} & \multirow[t]{3}{*}{\[
\begin{aligned}
& F(1,79)= \\
& 8.78 \\
& p<0.01 \\
& \eta^{2}=0.10 \\
& d=-0.24
\end{aligned}
\]} & n.s. & n.s. \\
\hline & & \(\eta_{p}^{2}=0.02\) & \(\eta_{p}^{2}=0.02\) \\
\hline & & & \\
\hline \multirow[t]{2}{*}{OperationSpan} & \multirow[t]{2}{*}{\[
\begin{aligned}
& F(1,80)= \\
& 45.62
\end{aligned}
\]} & n.s. & n.s. \\
\hline & & \(\eta_{p}^{2}=0.05\) & \(\eta_{p}^{2}=0.01\) \\
\hline Solution time & \[
\begin{aligned}
& \eta_{p}^{2}=0.36 \\
& d=-0.52
\end{aligned}
\] & & \\
\hline
\end{tabular}

\title{
The synergy of top-down and bottom-up attention in complex task: going beyond saliency models.
}

\author{
Enkhbold Nyamsuren (e.nyamsuren@rug.nl) \\ Niels A. Taatgen (n.a.taatgen@rug.nl) \\ Department of Artificial Intelligence, University of Groningen, Nijenborgh 9, 9747 AG Groningen, Netherlands
}

\begin{abstract}
This paper studies how visual perception of a scene is affected by cognitive processes beyond the scene's bottom-up saliency. The game of SET is taken as an example where contrast-based salient parts of a scene are ignored in favor of a larger group of similar elements. Using results from a laboratory experiment and a model simulation we explain how three cognitive mechanisms, differential acuity, visual iconic memory and declarative retrieval, considered together help to explain player's visual perception in SET.
\end{abstract}

\section*{Introduction}

Many studies describe how perception of a visual scene is governed by visual bottom-up mechanisms (Rayner, 1998). The conclusions derived in those studies are often based on results from relatively simple tasks involving free scanning or target search. It is widely accepted that visual attention is drawn toward a scene's salient parts (Egeth \& Yantis, 1997). This bottom-up saliency is commonly used to explain popout effect of items that are increasingly different from its surroundings (Theeuwes, 1992). However, these findings alone may lead to incorrect conclusions if used within a context of more complex problem-solving tasks. It is important to consider a relationship between salience and other cognitive mechanisms to properly understand the inner workings of human mind in such tasks. We use the game of \(\mathrm{SET}^{1}\) as an example of a problem-solving task that gives results that can be interpreted initially as contradictory to the visual pop-out effect. Next, we describe how the same results can be explained within a framework that combines bottom-up saliency with top-down goal-directed attention.

The deck in SET consists of 81 cards. Each card is uniquely defined by a combination of four attributes: color, shape, shading and number of shapes. Each attribute can have one of three distinct values: red, green, and blue for the color; open, solid and textured for the shading; one, two and three for the number; oval, rectangle and squiggle for the shape. At any moment in the game, 12 cards are dealt face up (Figure 1). From 12 cards, players should find any combination of three cards, referred to as a set, satisfying a rule stating that in the three cards the values for each particular attribute should be all the same or all different.

Jacob and Hochstein (2008) studied how bottom-up components of the game, such as attribute value distribution among cards, influences player's strategy. They concluded

\footnotetext{
\({ }^{1}\) SET is a game by Set Enterprises (www.setgame.com)
}
that players prefer to search for a set inside the largest group of cards that share at least one common value. They referred to a common value as the Most Abundant Value (MAV) and the group of cards that contained it as a MAV group. Sets that were inside MAV group were found sooner than sets outside of the group with an observed probability being significantly higher than a chance probability.

According to the bottom-up saliency mechanism it is expected that players should start a search with visually unique, hence most salient, cards. However, Jacob and Hochstein's finding suggests that player's visual attention is drawn toward larger group of cards that are visually similar. From a perspective of a bottom-up saliency, this is a highly counterintuitive result. Furthermore, another study by Nyamsuren and Taatgen (2013b) revealed that a similarity along particular attribute dimension plays more important role in players' strategy than the saliency of any individual card. Players are more likely to search for a set among larger group of cards with the same color than to attend a card, for example, with a unique shape.


Figure 1: An example array of 12 cards. The cards with solid and dashed borders represent two valid sets.

In this paper, we describe a more controlled experiment with set cards with an aim of more in-depth exploration of underlying cognitive processes. In order to use the MAV strategy, subjects must be able to recognize very quickly, which attribute values are most common. The goal of the study is to focus on this particular aspect of SET: to answer the question what cognitive processes facilitate such quick recognition in players. Based on experimental results and model simulations, we describe how three cognitive mechanisms that include visual acuity, visual memory and declarative memory retrieval help to explain MAV effect and bias toward similarity in color attribute.

\section*{Experiment}

\section*{Design and Procedure}

14 subjects participated in the experiment. All subjects were students of University of Groningen. Subjects' age ranged from 18 to 27 ( \(M=22\) ). Subjects started each trial by looking at the center of a computer screen. Next, they were shown a \(3 \times 4\) array of SET cards for a predetermined duration of time. After image of cards disappeared, subject was prompted to select one of 12 possible attribute values subject perceived as being the most abundant. The experiment consisted of 336 unique trials generated semirandomly. Trials were divided into a short and a long condition block. The array of cards was shown to subjects for 600 and 2000 ms in the short and long conditions respectively. For half of the subjects, blocks were presented in a reverse order. Within a block, trials were presented in a random sequence. In each block, the MAV group size varied from 6 to 12. There were six trials in each combination of MAV group size and attribute type. Prior to experiment, subjects were asked to do eight, four from each block, trials to let them get familiar with an experiment setup. Results from those trials were not included in the analysis. In addition, subjects' eye movements were recorded. We used the EyeLink 1000, a desktop-mounted remote eye tracker with monocular sampling rate of 500 Hz and spatial resolution of \(<0.01^{\circ}\) RMS. Exactly the same experiment setup and stimulus sizes as in Nyamsuren and Taatgen (2013b) were used in this study.

\section*{Experiment Results}

Scanpaths The difference in trial durations also results in quite clear difference in scanpaths. Subjects on average make 8.8 ( \(S E=0.38\) ) fixations in the long condition compared to 2.9 ( \(S E=0.17\) ) fixations in the short condition. Figure 2 provides a more detailed look on the trials' fixation counts. There is an \(87 \%\) probability that subject will make from seven to 11 fixations in the long condition. In contrast, subjects are likely to make only 2 to 4 fixations in \(94 \%\) of all trials in the short condition.

Figure 3a shows mean durations of fixations in a trial. All durations are measured in milliseconds. The last fixations are excluded from the calculation of these means since it is likely that those fixations were interrupted when the time limit was reached. The first two fixations do not show much difference between the short and long conditions. The durations for consecutive fixations in the long condition does not change much. In contrast, durations of third and fourth fixations in the short condition gradually become lower. There can two explanations to this. It may be an artifact of averaging. Smaller number of trials with three or four fixations may be resulting in lower mean. On the other hand, it is possible that shorter durations are deliberate. To test this hypothesis we have also calculated the average duration of fixations in the short condition trials with exactly four fixations. As we have expected, fixations in these trials have much shorter durations than respective
fixations in the long condition trials. Therefore, it is indeed possible that subjects were deliberately making shorter fixations in the short condition.


Figure 2: Frequencies of fixation counts subjects made during a trial. Frequencies are calculated separately for the (a) long and (b) short conditions.


Figure 3: (a) Changes in mean fixation durations over course of a trial in the short and long conditions. (b)
Changes in saccade amplitude over the course of a trial in the short and long condition.

Figure 3b shows how saccade amplitude changes over the course of a trial in both long and short conditions. Amplitude is measured in number of pixels that the saccade covers. There is not much difference between the two duration conditions. However, there is an obvious gradual rise in saccade amplitude as trial progresses. It suggests that there is a specific pattern in subjects' scanpaths.

Accuracy As Figure 4 shows, the overall accuracy increases as MAV group size increases. This is true for both short and long conditions. A test of proportions on pooled data indicate that subjects were more accurate in the long condition than in the short condition, \(\chi^{2}(1, N=4704)=\) \(35.63, p<0.001\). However, as Figure 4 shows, there are remarkably small differences in accuracies with respect to group sizes in two duration conditions.

Figure 5 shows a boxplot of accuracy variations based on attribute type and duration. We did logistic mixed-effect regression analysis using the duration condition, attribute type and the interaction between the two as predictors. The intercept in the regression model reflects expected accuracy in a short condition trial where the MAV belongs to shading. Relative accuracy increased when MAV belonged to color ( \(z=3.19, p=0.001\) ) and decreased when MAV
belonged to either number ( \(z=-4.142, p<0.001\) ) or shape ( \(z=-2.577, p=0.01\) ). Overall performance in the long condition increased significantly ( \(z=2.093, p<0.036\) ). However, there were no significant interactions between duration conditions and attribute types.

Chi-square tests confirmed that subjects were significantly better at identifying the MAV with a color attribute than any other attribute type. Subjects showed little difference in accuracies in the short and long conditions with respect to color \(\left(\chi^{2}(1, N=1176)=2.91, p=0.088\right)\). It is surprising that, despite the significant difference in average number of fixations made, subjects are equally good at identifying color value in both duration conditions. In contrast, accuracies in the long condition were significantly higher for number \(\left(\chi^{2}(1, N=1176)=15.283, p<0.001\right)\), shape \(\left(\chi^{2}(1, N=1176)=16.94, p<0.001\right)\) and shading \(\left(\chi^{2}(1, N=1176)=4.12, p=0.04\right)\) than in the short condition.


Figure 4: Mean accuracies averaged over all combinations of MAV group sizes and duration conditions.


Figure 5: Mean accuracies averaged over all combinations of attribute types and duration conditions.

\section*{Experiment Discussion}

Effect of MAV Group Size on Accuracy This effect can be explained by the priming of declarative memory by the visual system. There are several studies indicating that the human visual system has some form of iconic memory (Kieras, 2009). It is a low-resolution high-capacity memory where visual information is stored pre-attentively for a short duration of time. The process of gathering information is massively parallel and almost instantaneous. However, information about a visual object is stored as a collection of separate feature channels (such as color or shape) rather than single coherent object (Treisman \& Gelade, 1980).

Therefore, iconic memory has just enough resolution to guide further attention shifts and encoding.

There is evidence that visual perception can influence processes of memory retrieval (Wais, Rubens, Boccanfuso, \& Gazzaley, 2010). It is reasonable to assume that visual stimuli can facilitate memory retrieval of items that are in some form related to the stimuli. Furthermore, we assume the same process applies to iconic and declarative memories. Items in iconic memory facilitate retrieval of similar or related items in declarative memory. In other words, items in declarative memory get activated by items in iconic memory. The strength of such activation depends on the number of items in iconic memory that are related to the item in declarative memory.

This interaction between iconic and declarative memories can explain why subjects find it easier to identify the MAV among larger group of cards. Subjects need to do two tasks: (1) gather visual information through attention shifts and (2) retrieve the MAV from memory when prompted. The second retrieval step is influenced by the content of iconic memory that was gathered during the first step. When MAV group size is large, more values enter iconic memory, and corresponding MAV value in declarative memory receives a higher activation during the retrieval.

Effect of Attribute Type and Duration on Accuracy The exchange of activations from iconic to declarative memories also helps to explain why subjects are better at identifying color values than values from any other attribute type.
However, there are studies showing that an ability to capture finer details of a visual scene becomes worse as the distance from a foveal region increases (Nelson \& Loftus, 1998). This introduces limitations on what visual features can be gathered into iconic memory. As an object is further away from the foveal region it becomes more likely that some of its features will not enter iconic memory due to limitations of peripheral vision. A feature's acuity threshold defines the maximum distance from a foveal point at which the feature is still recognizable (Kieras, 2009). Compared to other features, color has a higher threshold making it easier to recognize in the peripherals. Thus, color values have a higher chance of entering iconic memory thereby spreading more activation to the same values in declarative memory.

When features, such as shape and shading, have a limited acuity, subjects need to fixate closer to respective visual objects to bring them within threshold distance. This explains why subjects perform better in the long condition trials. Subjects can make more fixations and gather a more complete gist of the visual scene in iconic memory, which then facilitates a more accurate declarative retrieval.

Scanpaths There are two interesting effects in subjects' scanpaths. Firstly, subjects seem to react to time pressure in the short condition by having shorter fixation durations. This behavior also supports our assumption that iconic memory and peripheral vision play an important role. It is possible that subjects compensate for a shorter duration by
making as many fixations as possible and accumulating in iconic memory as much visual information as possible. The pattern of increasing saccade amplitudes provides a clue about preferences of possible fixation locations. Subjects start by fixating on the cards closest to the center of the screen and gradually switch to the cards on the peripherals. These fixations from inwards toward outwards should result in increasing saccade amplitudes shown in Figure 3b. In addition to providing more clues about subjects' behavior, scanpaths provide additional measurements besides accuracy against which model fit can be evaluated.

\section*{Cognitive Model}

\section*{Cognitive Architecture}

We have used ACT-R cognitive architecture (Anderson, 2007) to develop the model. An additional module called Pre-attentive and Attentive Vision (Nyamsuren \& Taatgen, 2013a) was used instead of ACT-R's default vision module. The PAAV module provides several extra functionalities that are otherwise not supported by ACT-R.

PAAV can pre-attentively capture the gist of a visual scene and store it in iconic memory. The content of iconic memory is updated before and after each saccade and before each time the memory is accessed. The update process is instantaneous from a perspective of model's timeframe. Iconic memory may contain complete information for some visual objects, such as an object's color, shape, shading and size. However, for most visual objects the iconic memory will contain incomplete information (e.g. color only) due to limited acuity. PAAV recognizes that not everything in a visual scene can be resolved by model's peripheral vision at any given moment. In PAAV two parameters, \(a\) and \(b\), define differential acuities of color, shape, size and shading with color having the highest acuity. Fitness of these parameters was tested on models of three different visual search tasks and the updated model of game of SET (Nyamsuren \& Taatgen, 2013a). An object's feature in iconic memory, although persisting through saccades, decays after a short period of time (currently 4 sec ) if not recognizable in peripheral vision anymore.

The content of iconic memory is used to guide the model's visual attention. Visual objects with the highest saliency values are prioritized for visual attention and further encoding. In PAAV, the bottom-up saliency is a sum of saliency values calculated for each feature dimension as a function of contrast to its surrounding. For example, a single red card among green, otherwise similar, cards will be the most salient one and draw the model's attention. PAAV uses a binary measure of similarity: 1 for exact match and 0 otherwise. No adjustable parameters are used in calculation of bottom-up saliency (Nyamsuren \& Taatgen, 2013a). It is a simplified version of Wolfe's (2007) saliency function.

In ACT-R knowledge chunks are stored in declarative memory. Each chunk has an activation value that usually reflects chunk's recency and frequency of use by a model. A chunk with the highest activation has the highest probability
of retrieval. Besides frequency and recency, a chunk's activation can be increased by the content of iconic memory. Each visual object in iconic memory spreads activation to every declarative chunk with the same features. So depending on the content of iconic memory at the time the results of two same retrievals can differ. The model uses exactly the same set of parameters for declarative retrieval as in the original model of game of SET. Details of those parameters are described in Nyamsuren \& Taatgen (2013b).

\section*{Model of MAV Task}

Model Strategy Model performed 50 times the same two blocks of trials subjects did. Model starts each trial while fixating at the center of the screen. When cards are shown, models need some time to create a working memory before the first saccade is made. At the same time, model updates its iconic memory with representations of cards. Then model follows with free scanning using bottom-up saliency values to calculate consecutive fixation points. Each fixation is followed by encoding of an attended card. Free scanning stops when time limit is reached and representations of cards disappear. At this point model retrieves any one of 12 possible attribute values from declarative memory. Result of this retrieval depends on content of iconic memory the model has built up during the free scanning. The retrieved value is recorded as model's response for the trial.

Model Accuracy Model is quite good at replicating subjects' accuracy. Figure 6 shows that model's accuracy increases linearly as the MAV group size increases. This effect is present in both the short and long condition. However, just like subjects, the model shows a better performance in the long condition.


Figure 6: Mean accuracies averaged over all combinations of MAV group sizes and duration conditions.

The model is also good at reflecting subjects' accuracy depending on combination of attribute types and duration conditions. Firstly, as Figure 7, there is a general increase in model's accuracy in the long condition. Except in color, the model clearly benefits from additional time in all other three attributes. Next, Figure 7 shows that model is much better at identifying MAV belonging to color attribute than to any other attribute type. Similar to human performance, model's accuracy for color in the short condition is higher than the accuracies for other three attribute types in the longer trials.


Figure 7: Mean accuracies averaged over all combinations of attribute types and duration conditions.

Model Scanpaths Comparison of model's scanpaths to that of subjects should give additional measure of how well the model fits human data at the level of raw eye movements. Figure 8 shows distributions of fixation counts the model made in the long and short conditions. In \(99 \%\) of all long condition trials, the model made 9-10 fixations. It is within a range of 7-11 fixations subjects made. In the short condition, the model made either two or three fixations. It is also within a range of 2-4 fixations subjects made. As Figure 9a shows, model's fixation durations do not differ in the long and short conditions. The lower duration for the third fixation in the short condition is a result of interruption due to duration limit.


Figure 8: Frequencies of fixation counts model made during a trial. Frequencies are calculated separately for (a) long and (b) short conditions.

The model was able to reproduce a pattern of increasing saccade amplitudes in long condition trials, as it is shown in Figure 9b. It was not completely expected since we have not incorporated into the model any deliberate mechanisms to promote this behavior. Because the model makes only one or two saccades in a short condition trial, it is hard to make any conclusive statements about the pattern of amplitude changes. The same model is used in both duration conditions. Hence, there is no reason to expect the model to show different scanpath pattern in the short condition. The lower amplitude for the second saccade in the short condition is most likely due to smaller number of observations from which the mean is calculated. For exactly the same reason, amplitudes for the 9th and 10th saccades drop in the long condition since there are fewer trials that have more than 10 fixations.


Figure 9: (a) Changes in model's mean fixation durations over course of the trial in the short and long conditions. (b) Changes in model's saccade amplitude over the course of the trial in the short and long condition.

\section*{Discussion on Model Results}

The point at which model has to decide on a choice of MAV is the retrieval from a declarative memory. As model shows, the spreading activation from iconic memory is a major factor deciding the result of this retrieval. However, it is possible to counter-argue that spreading activation from iconic memory is not necessary, and items in declarative memory are activated directly through visual encoding of similar items. Such mechanism is possible and used in our model. Cards with the MAV have a higher chance probability of getting visual attention and being encoded. As a result, the MAV in declarative memory receives more activation and is retrieved. Although this argument would explain subjects' behavior in the long condition, it does not explain why there is a similar effect of MAV group size in the short condition. Neither subjects nor model can encode more than two cards in the short condition, and it is not enough to influence the retrieval. Instead, it is likely that subjects rely on visual information in peripheral regions for choosing MAV. Furthermore, the fact that subjects are quite good at identifying the MAV even within 600 ms implies that process of gathering information from peripherals is very efficient. The model simulation suggests that it may be massively parallel and instantaneous.

In the other side, acuity limitations of visual features in peripheral vision can result in incomplete and inaccurate iconic memory. This imperfect internal representation may explain why subjects fail to reach \(100 \%\) accuracy. It also explains why subjects get better given opportunity to do more fixations in the long condition. More fixations negate the effect of low acuity and result in a more complete representation of the scene inside iconic memory. Furthermore, giving a higher acuity to color in model simulation increases model's accuracy in identifying the most abundant color values in both conditions. This result is similar to the result from the experiment, and, therefore, supports the assumption that human vision is affected significantly by different acuity properties of visual features.

The model produces the same pattern of increasing saccade durations in the long condition without any deliberate mechanisms. It suggests that spatial arrangement
and the bottom-up salient parts of the visual scene define the topology of fixation points, more specifically the characteristic fixations from inwards to outwards. In the model, cards around the edges of the screen are not fully visible due to limited acuity. Those cards have reduced bottom-up activation compared to cards around the center of the screen. As a result, the model prefers to fixate on cards closer to the screen center at the early stages of the trial. We were not able to simulate the deliberate reduction in fixation durations subjects have shown in the short condition. Visual processes currently implemented in ACT-R do not provide appropriate mechanisms to simulate this effect.

\section*{Discussion and Conclusion}

The model fits subjects' accuracies and scanpaths well supporting the hypothesis that the same cognitive processes simulated in the model may also be used by human subjects. More specifically, a combined effect of differential acuity, pre-attentive visual iconic memory and implicit communication with declarative memory can influence our visual perception of the world.

The results from this study can explain player's behavior in game of SET. Player has to decide on a group of cards to be searched for a set. This choice is made through a declarative retrieval of an attribute value that is common among group cards. Similar to the experiment's task, this retrieval is influenced by a content of iconic memory introducing a bias toward a larger group of cards and cards with same color. The retrieved value is used to target attention to specific cards with that value. This top-down control over eye movements overrides the bottom-up saliency of the scene. It explains both why players are better at finding set within a group with many similar cards (Jacob \& Hochstein, 2008) and the general preference toward cards with a similar color (Nyamsuren \& Taatgen, 2013b). The model of SET player implemented on the same principles described here was able to simulate player's behavior (Nyamsuren \& Taatgen, 2013a, 2013b). It is a good example of a case where cognitive mechanisms beyond bottom-up saliency can influence the behavior in a reasonably complex problem-solving task. It implies that not every eye movement pattern can be attributed to bottomup salient components of the scene.

Subjects are far better in identifying the MAV even in the most difficult conditions. In 600 ms condition with smallest MAV group size, subjects show much higher accuracy than \(8 \%\) chance probability of success. This result indicates that capabilities of human visual system may be higher than previously expected. The ability to capture a gist of a visual scene from first few fixations is known for a long time (Loftus \& Mackworth, 1978). However, it is commonly viewed that functionality of such gist is limited to attentional guidance and providing early structural information for encoding, a preview effect (Rayner, 1998). On the other hand, our study suggests that gist, in form of iconic memory, may be involved in decision-making. It is possible through connections between memories in human
brain. In this study, we talked about similarity-based crossmemory activations between iconic and declarative memories. However, it may be possible that similar cross activations exist between other forms of memory.

The model code and the data can be downloaded via following link: http://www.ai.rug.nl/~n_egii/models/.

\section*{Acknowledgement}

We would like to express our gratitude to Eveline Broers for conducting a pilot experiment and giving a valuable insight into data analysis.

\section*{References}

Anderson, J. R. (2007). How Can Human Mind Occur in the Physical Universe? New York: Oxford University Press.
Egeth, H. E., \& Yantis, S. (1997). Visual attention: control, representation, and time course. Annual Review of Psychology , 48, 269-297.
Jacob, M., \& Hochstein, S. (2008). Set recognition as a window to perceptual and cognitive processes. Perception \& Psychophysics, 70 (7), 1165-1184.
Kieras, D. (2009). The Persistent Visual Store as the Locus of Fixation Memory in Visual Search Tasks. In A. Howes, D. Peebles, \& R. Cooper (Ed.), 9th International Conference on Cognitive Modeling - ICCM2009. Manchester, UK.
Loftus, g. R., \& Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. Journal of Experimental Psychology: Human Perception and Performance , 4, 565-572.
Nelson, W. W., \& Loftus, G. R. (1980). The functional visual field during picture viewing. Journal of Experimental Psychology: Human Learning and Memory , 6, 391-399.
Nyamsuren, E., \& Taatgen, N. A. (2013a). Pre-attentive and Attentive Vision Module. Cognitive Systems Research, 24, 62-71.
Nyamsuren, E., \& Taatgen, N. A. (2013b). Set as instance of a real-world visual-cognitive task. Cognitive Science, 37 (1), 146-175.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. Psychological Bulletin, 124, 372-422.
Theeuwes, J. (1992). Perceptual selectivity for color and form. Perception \& Psychophysics , 51, 599-606.
Treisman, A., \& Gelade, G. (1980). A Feature-integration Theory of Attention. Cognitive Psychology, 12, 97-136.
Wais, P. E., Rubens, M. T., Boccanfuso, J., \& Gazzaley, A. (2010). Neural Mechanisms Underlying the Impact of Visual Distraction on Long-term Memory Retrieval. Journal of Neuroscience , 30 (25), 8541-8550.
Wolfe, J. M. (2007). Guided Search 4.0: Current Progress With a Model of Visual Search. In W. Gray, Integrated Models of Cognitive Systems (pp. 99-119). New York: Oxford.

\title{
Is Facial Expression Processing Holistic?
}

\author{
Akinyinka O. Omigbodun (aomigbodun@ucsd.edu) \\ Electrical and Computer Engineering, University of California San Diego \\ La Jolla, CA 92093 USA \\ Garrison W. Cottrell (gary@eng.ucsd.edu) \\ Computer Science and Engineering, University of California San Diego \\ La Jolla, CA 92093 USA
}

\begin{abstract}
Most studies examine holistic processing with respect to facial identity, but at least one study has also looked at holistic processing of facial expressions (Calder, Young, Keane, \& Dean, 2000). However, this work used the partial composite paradigm, which is known to exhibit bias effects (Richler, Cheung, \& Gauthier, 2011). The complete composite paradigm (Gauthier \& Bukach, 2007) provides a bias-proof way to quantitatively measure holistic processing. In this paper, we perform the corresponding experiment in our face processing model (EMPATH, (Dailey, Cottrell, Padgett, \& Adolphs, 2002)) to predict whether holistic processing of facial expressions will be found if the corresponding human experiments are performed, and our prediction is that it will. We also compared our model's performance to the participants in recent experiments in facial expression recognition in humans (Tanaka, Kaiser, Butler, \& Le Grand, 2012). Tanaka et al. (2012) concluded that expression recognition is not always holistic, but our results suggest that it is.
\end{abstract}

Keywords: holistic processing

\section*{Introduction}

Is all facial processing holistic? For example, when we are judging whether a person is trustworthy, tired, middle-aged, or happy, is our decision based on a global percept of the face rather than constituent parts? One might suspect that the answer is no, at least in some cases. Holistic processing might depend upon the task. The evidence from face inversion, composite, and parts-wholes experiments weighs in favor of holistic over parts-based processing for face identity (Yin, 1969; Tanaka \& Farah, 1993; Cheung, Richler, Palmeri, \& Gauthier, 2008; Richler, Tanaka, Brown, \& Gauthier, 2008). Less research has focused on understanding the nature of facial expression recognition. One reason to suspect that expression recognition is different is the categorization of expressions into just seven categories - happy, sad, surprised, angry, disgusted, fearful, and neutral - in line with the "six basic expressions" theory (Ekman \& Friesen, 1976), at least as practiced in most psychology experiments. Previous work in our lab has shown that the basic level processing of objects (e.g., into an overall category, such as chairs or lamps), does not behave in the same manner as subordinate level processing in our models, and does not engage our model's equivalent of the Fusiform Face Area. One consideration is that the holistic processing of faces may be induced if there is any variation in identity, regardless of the task being performed. In our modeling work, we tested the hypothesis that without any variation in identity, the processing of facial expressions is holistic.

Holistic processing of facial stimuli is generally measured with composite face paradigms, where chimeric faces are constructed from the top and bottom halves of different "source" faces. Subjects are asked to identify face halves of chimeric faces or to judge whether two halves of a pair of chimeric faces are the same or different. Misalignment of the top and bottom halves generally leads to an increase in the subjects' accuracy and/or a decrease in reaction time relative to the aligned condition, demonstrating that faces are processed holistically. We test our model with simulations based on the complete composite paradigm (Gauthier \& Bukach, 2007), an improvement upon what Gauthier \& Bukach call the partial design, which is what is classically used (Young, Hellawell, \& Hay, 1987). The complete composite paradigm eliminates the effects of response bias (for example, a preference towards answering "same" for misaligned stimuli). our results predict that a facial expression recognition experiment with humans based on the complete composite paradigm will indicate holistic processing.

We then use our model to account for experimental results in Tanaka et al. (2012). Tanaka et al. used their own nonstandard composite paradigm in one of their experiments that we simulated and a variation on the partial design in another that we also simulated. They inferred from their experimental results that facial expression processing is holistic when there is a clash between parts of a facial expression (e.g. angry-happy composite) but analytic or parts-based when there is little or no conflict between the parts (e.g. normal happy face). They posit an earlier stage where a stimulus is rapidly assessed for parts that clash to determine the processing pathway to be used. However, we achieved similar results with our model which (1) has one processing pathway for all stimuli, (2) does not have an earlier stage for quick appraisals, and (3) is not trained to combine decisions on parts of a stimulus into an overall decision. Our results suggest that one holistic processing pathway is sufficient to account for their data.

\section*{Complete Composite Paradigm (CCP)}

In composite paradigms, participants in experiments are presented with two composite faces, and asked to make samedifferent judgments about the cued face halves (top or bottom) while ignoring the other face halves. The complete composite paradigm (CCP) (Gauthier \& Bukach, 2007) provides a bias-proof way to quantitatively measure the interaction between a subject's decision and the presence or ab-


Figure 1: Schematic showing the complete composite paradigm. The subject must decide whether the top face halves are the same or different. In the congruent condition, the top and bottom face halves would generate the same answer, while in the incongruent condition, they would not. Holistic processing is measured as a congruency effect.
sence of a change in the unattended face halves (see Fig. 1). If the answer is the same for the top and bottom halves, it is a congruent trial, otherwise, it is incongruent. A congruency effect, difference in sensitivity, \(d^{\prime}\), between congruent and incongruent trials, is indicative of holistic processing: the unattended face half is obligatorily processed. The partial design only has trials where the unattended halves are different, and holistic processing is measured as a difference in performance between aligned and misaligned trials. This is a flawed measure; using it, researchers concluded holistic processing was unrelated to performance on the Cambridge Face Memory Test (CFMT). The relationship is found to be strong using the complete design (Richler et al., 2011).

\section*{Methods}

\section*{The Model}

Each input image goes through a two-step preprocessing stage: Gabor filtering and Principal Component Analysis (PCA). The biologically motivated 2-D Gabor filter is constructed by using a two-dimensional sinusoid localized by a Gaussian envelope (Daugman, 1985). By tuning to particular spatial frequency and orientations, the Gabor filter magnitudes can be used to simulate the responses of complex cells in primary visual cortex (V1). Following Gabor filtering, PCA reduces the dimensionality of the information, simulating the information extraction mechanism beyond V1. After these preprocessing steps, each image is represented by a vector with relatively low dimension to be input to the perceptron.

We computed each Gabor filter using the following equation:
\[
\begin{aligned}
& G(x, y, S, F, W)= \\
& \quad k \cdot e^{-\pi S^{2}\left(x^{2}+y^{2}\right)} \cdot\left(e^{j(2 \pi F(x \cos W+y \sin W))}-e^{-\pi\left(\frac{F}{S}\right)^{2}}\right)
\end{aligned}
\]
where \(S\) is a scaling parameter, \((F, W)\) is the polar frequency of the complex sinusoid, \((x, y)\) are the spatial coordinates and \(k\) is a constant (Dias, 2007; Movellan, 2002). There were five spatial frequencies \(\left(F=1 / 2^{i}\right.\) for \(\left.i=2, \ldots, 6\right)\) and eight spatial
orientations \((W=j \pi / 8\) for \(j=0, \ldots, 7)\) for a total of 40 different Gabor filters. \(S\) was related to \(F\) as follows: \(S=3 F / \sqrt{2 \pi}\). \(S\) was chosen such that each filter had the characteristic form of biological two-dimensional receptive field profiles (Hubel, 1988; Daugman, 1988). The filters were centered and applied at the 1080 points in a 36 by 30 grid on each image.

We then use PCA to map the Gabor filter features from a 43200-dimensional space to a space with many fewer dimensions. In forming the PCA projection matrix, we retained enough eigenvectors to account for \(90 \%\) of the variance. For classification, a single-layer perceptron was trained with gradient descent using a cross-entropy error function.

\section*{Network Training and Testing}

70 gray-scale images were selected from the Pictures of Facial Affect (POFA) (Ekman \& Friesen, 1976). These were cropped to \(240 \times 292\) pixels and adjusted to ensure that there was uniformity in the upright frontal face views; to allow for shifting, the input images were 1.5 times as wide as the face images (i.e. images of size \(360 \times 292\) ) with the faces flush against the left or right edge, as shown in Fig. 3a. Misaligned face halves are shown in Fig. 3b.

The POFA dataset includes seven facial expressions (Happy, Sad, Surprised, Angry, Disgusted, Fearful, and Neutral) for each of 14 actors. We selected 10 of these for our experiments (em, gs, jj, mf, mo, nr, pe, pf, sw, wf). We trained on 9 and tested on 1 , repeating this 10 times with each actor getting a chance to be tested. For the remaining 9 actors, we trained on 8 and used the 9th as a hold-out to stop training. This was repeated 9 times with each of the 9 actors having a turn as the hold-out; so we ended up with 9 networks predicting the expression on each stimulus. The consensus of the 9 instances was taken as the model's decision for an experimental run. We averaged the results of the 10 runs.

Experimental tasks in our simulations involving the identification of top and bottom face halves necessitated training the model to classify vectors of Gabor features corresponding to images with only half of a face. We modeled attention to half a face by using a transformation of the Gabor features that reduced the contribution of the top or bottom of a face to the total training set covariance. The transformation that we used for this and a proof that the contribution of transformed Gabor features to the total covariance is reduced is beyond the scope of this paper. In modeling the process of identifying top and bottom face halves in testing and in the experiments, we simulated giving more attention to half of a face using this transformation.

The PCA projection matrix was generated from the Gabor feature vectors designated for training the model. Both before and after the projection of the training stimuli by the matrix, the projections were z -scored in order to put each input to the perceptron on the same scale (LeCun, Bottou, Orr, \& Mueller, 1998). The stimuli for cross-validation were rescaled as one set of vectors, and the stimuli for testing and the experiments were rescaled as another set; with the new sample variances for these sets equal to the ratio of the set sizes to the size of


Figure 2: The real components of Gabor filters are shown relative to the size of an image at five scales and orientations.


Figure 3: (a) left- and right-shifted images and (b) top-left-bottom-right (TLBR) and TRBL images
the training set. This was done to ensure that the rescaling was as uniform as possible across all the feature vectors.

All composites were constructed from the same individual (ten times, once for each test individual, as described above), which ensured that expression and identity recognition were not confounded. The networks were trained as described above, to classify each (whole, unaltered) face into one of seven expression categories. In order to obtain a same/different judgment from a network that only processed one face at a time, we presented the network with the two stimuli (with attention on the queried half, as described above), one at a time. Since there were actually nine networks for each test face (as described above), we compared the consensus of the networks on one stimulus with the consensus on the other. If they match, the overall response is "same"; otherwise the response is "different".

To obtain reaction times, we appeal to the well-established inverse relationship between reaction time and confidence ratings (Audley, 1960; Baranski \& Petrusic, 1998) in mind, we computed a measure of confidence for our model which was the ratio of the number of network instances in agreement with the consensus to the total number of instances for each decision made (when the consensus was the correct decision). To model reaction time, we subtracted these values from one.

For the experiments by Tanaka et al. (2012) (described shortly), since there are no comparisons between two images (all the subjects had to do was decide if the cued half of the stimulus was happy or angry), we simply use a consensus of the nine networks.

\section*{Simulations}

Simulation 1: The CCP on expression recognition The first simulation used the CCP (Fig. 1) to test whether our model predicts that facial expression recognition is holistic. Samedifferent judgments were obtained (as described above), and
holistic processing was then measured as the difference in sensitivity, \(d^{\prime}\), between congruent and incongruent trials.

Simulation 2: Experiment 1 of Tanaka et al. (2012) In experiments performed by (Tanaka et al., 2012), subjects were asked to decide if the cued half of the stimulus was happy or angry. Tanaka et al. (2012) measured the subjects' percentage accuracies and reaction times (in milliseconds). They based their experimental design on the observation that happy expressions are bottom biased (meaning it is easier to recognize a happy expression from the bottom of a happy face than it is from the top half) and angry expressions are top biased (Calder et al., 2000); they only counted trials in which the correct responses for the cued lower halves were happy, and those for the cued upper halves were angry. In their first experiment, Tanaka et al. (2012) compared subjects' performance for a happy lower half face in four pairings: (1) with a happy top half face (normal), (2) with an angry top half face (angry-happy), (3) without a top half face (isolated), and (4) with a neutral top half face (neutral). \({ }^{1}\) There were four corresponding pairings for an angry top half face (see Fig. 4). Their reasoning was that if happy expressions in the lower face half and/or angry expressions in the top face half of normal, neutral, and isolated stimuli were recognized equally easily, then this was evidence for parts-based processing (since this would suggest that the other face half is being "ignored"); and if recognition of angry-happy stimuli was worse than that of neutral and isolated stimuli, then this was evidence for holistic processing. To assess whether our model could account for their observations (some of which were interpreted as indicating parts-based processing), we simulated their experiment with POFA (Ekman \& Friesen, 1976) (Figure \(7 \mathrm{~b}-\mathrm{c}\) ).

Simulation 3: Experiment 3 of Tanaka et al. (2012) Tanaka et al. (2012) looked at the performance of subjects in identifying happy and angry expressions in the lower and upper face halves respectively with two different stimuli types: (1) normal (happy + happy, angry + angry), and (2) angryhappy (happy lower half + angry top half), under two different conditions of alignment: aligned and misaligned (see Fig. 5). We note that their third experiment was based on the partial design which we explained previously has drawbacks. Their reasoning was that equal performance in the aligned

\footnotetext{
\({ }^{1}\) We have chosen to label as normal and angry-happy the stimuli types that they called congruent and incongruent to avoid confusion in this paper.
}


Figure 4: Stimuli types from Exp. 1(Tanaka et al., 2012).


Figure 5: Stimuli types from Exp. 3 (Tanaka et al., 2012).
and misaligned conditions for the normal stimuli would indicate parts-based processing and better performance in the misaligned condition relative to the aligned condition for the angry-happy stimuli would indicate holistic processing. Once again, to assess whether our model could account for their observations (some of which they attributed to parts-based processing), we simulated their experiment with POFA (Ekman \& Friesen, 1976).

\section*{Results and Discussion}

In Simulation 1, the sensitivity, \(d^{\prime}\), for the incongruent trials was less than the sensitivity for the congruent trials (Fig. 6). This observation of the congruency effect confirmed that the model processes facial expressions holistically. It is now left for an experiment in expression recognition with humans based on the CCP to be conducted; we expect that a congruency effect will be observed.

In their first experiment, Tanaka et al. (2012) found that, for happy expressions, the percent accuracy on angry-happy stimuli was reliably less than the accuracy on the other stimuli types. In the case of the model, while there were no statistically significant differences between the four stimuli types, we observed a very similar trend. The reaction time for the angry-happy stimuli was reliably greater than the reaction time for the other three stimuli types in the experiment. In modeling, the difference in reaction time for angry-happy and normal stimuli approached statistical significance, and was reliable for angry-happy and isolated stimuli. Notably, there were no significant differences between the percent accuracy


Figure 6: sensitivity \(\left(d^{\prime}\right)\) in aligned and two misaligned conditions for congruent and incongruent trials of the complete composite paradigm. Error bars indicate standard error of the mean.
or reaction time for the normal, isolated and neutral stimuli in the experiments and in modeling. From their observations, Tanaka et al. (2012) concluded that the recognition of lower half face happy expressions is analytic when there is little or no conflict between the face halves (e.g. normal, isolated and neutral stimuli) but holistic when their is a clash between the face halves (e.g. angry-happy stimuli). However, we made similar observations in the model which possesses a single pathway of holistic processing.

For angry expressions, Tanaka et al. (2012) found that the percent accuracy on isolated stimuli was less than the accuracy on normal and neutral stimuli, and that the accuracy on angry-happy stimuli was less than the accuracy on the other three stimuli types; no other differences were observed for accuracy or reaction time. They concluded from their observations that the recognition of top half face angry expressions was not purely analytic and in fact benefited from the presence of whole-face information. While we did not observe any statistically significant differences between accuracy or reaction time for the four stimuli types in the model, we once again - observed a similar trend with the model (Fig. 78).

In their third experiment, Tanaka et al. (2012) found a lower percent accuracy and greater reaction time for angryhappy stimuli in the aligned condition relative to the misaligned condition, and attributed this to holistic processing. In contrast, for normal stimuli, there was no difference between the accuracy in the aligned and misaligned conditions. They expressed surprise at seeing a shorter reaction time in the misaligned relative to the aligned condition. They attributed the absence of a holistic facililation for aligned normal stimuli to analytic processing. Yet once again, we observed very similar trends in the model, suggesting that holistic processing suffices for explaining the experimental results (Fig. 9-10).

Our modeling work emphasizes the importance of modeling in cognitive science. Without a model, inferences from experimental results can be misleading. While we cannot


Figure 7: percent accuracy in simulation 2 with our model (a) and experiment 1 in Tanaka et al. (2012) (b).


Figure 8: "reaction time" in simulation 2 with our model (a) and reaction time in experiment 1 of (Tanaka et al., 2012) (b).
claim - based on our model - that facial expression processing in humans is purely holistic, we can conclude that Tanaka et al.'s experiments do not show that it has analytic attributes.

\section*{Acknowledgments}

This work was supported in part by NSF grants SMA1041755 and IIS-1219252 to G.W. Cottrell. We thank the Perceptual Expertise Network (PEN) and Gary's Unbelievable Research Unit (GURU) for discussion and comments, and Jim Tanaka for permission to use his data.

\section*{References}

Audley, R. J. (1960). A stochastic model for individual choice behavior. Psychological Review, 67, 1-15.
Baranski, J. V., \& Petrusic, W. M. (1998). Probing the locus of confidence judgments: Experiments on the time to de-
termine confidence. Journal of Experimental Psychology: Human Perception and Performance, 24, 929-945.
Calder, A. J., Young, A. W., Keane, J., \& Dean, M. (2000). Configural information in facial expression perception. Journal of Experimental Psychology: Human Perception and Performance, 26(2), 527-551.
Cheung, O. S., Richler, J. J., Palmeri, T. J., \& Gauthier, I. (2008). Revisiting the role of spatial frequencies in the holistic processing of faces. Journal of Experimental Psychology: Human Perception and Performance, 34, 13271336.

Dailey, M. N., Cottrell, G. W., Padgett, C., \& Adolphs, R. (2002). Empath: a neural network that categorizes facial expressions. Journal of Cognitive Neuroscience, 14(8), 1158-1173.
Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-


Figure 9: percent accuracy in simulation 3 with our model (a) and experiment 3 of Tanaka et al. (2012) (b).
dimensional visual cortex filters. Journal of the Optical Society of America A, 2, 1160-1169.
Daugman, J. G. (1988). Complete discrete 2-d gabor transforms by neural networks for image analysis and compression. IEEE Transactions on Acoustics, Speech, and Signal Processing, 36(7), 1169-1179.
Dias, S. S. (2007). Improved \(2 d\) gabor filter. Available from http://www.mathworks.com/matlabcentral/ fileexchange/13776-improved-2d-gabor-filter (Retrieved May 7, 2011)
Ekman, P., \& Friesen, W. (1976). Pictures of facial affect. Palo Alto, CA: Consulting Psychologists Press.
Gauthier, I., \& Bukach, C. (2007). Should we reject the expertise hypothesis? Cognition, 103, 322-330.
Hubel, D. H. (Ed.). (1988). Eye, brain, and vision. New York, NY: Scientific American Library.
LeCun, Y., Bottou, L., Orr, G. B., \& Mueller, K. Robert. (1998). Efficient backProp. In G. Orr \& K. Mueller (Eds.), Neural networks: tricks of the trade. New York: Springer.
Movellan, J. R. (2002). Tutorial on gabor filters (Tech. Rep.). La Jolla, CA: University of California San Diego.
Richler, J. J., Cheung, O. S., \& Gauthier, I. (2011). Holistic processing predicts face recognition. Psychological Science, 22(4), 464-471.
Richler, J. J., Tanaka, J. W., Brown, D. D., \& Gauthier, I. (2008). Why does selective attention to parts fail in face processing? Journal of Experimental Psychology: Learning, Memory and Cognition, 34, 1356-1368.
Tanaka, J. W., \& Farah, M. J. (1993). Parts and wholes in face recognition. The Quarterly Journal of Experimental Psychology Section A, 46, 225-245.
Tanaka, J. W., Kaiser, M. D., Butler, S., \& Le Grand, R. (2012). Mixed emotions: holistic and analytic perception of facial expressions. Cognition \& Emotion, 26(6), 961977.

Yin, R. (1969). Looking at upside-down faces. Journal of Experimental Psychology, 81, 141-145.
Young, A., Hellawell, D., \& Hay, D. (1987). Configural information in face perception. Perception, 16, 747-759.


Figure 10: "reaction time" in simulation 3 with our model in identifying happy (a) \& angry (b) facial expressions and reaction time in experiment 3 of Tanaka et al. (2012) for happy (c) \& angry (d) facial expressions.

\title{
Discovering Pronoun Categories using Discourse Information
}

\author{
Naho Orita (naho@umd.edu) \\ Rebecca McKeown (rmckeown@umd.edu) \\ Naomi H. Feldman (nhf@umd.edu) \\ Jeffrey Lidz (jlidz@umd.edu) \\ Department of Linguistics, 1401 Marie Mount Hall, University of Maryland, College Park, MD 20742 USA \\ Jordan Boyd-Graber (jbg@umiacs.umd.edu) \\ College of Information Studies, South Hornbake, University of Maryland, College Park, MD 20742 USA \\ University of Maryland Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742 USA
}

\begin{abstract}
Interpretation of a pronoun is driven by properties of syntactic distribution. Consequently, acquiring the meaning and the distribution are intertwined. In order to learn that a pronoun is reflexive, learners need to know which entity the pronoun refers to in a sentence, but in order to infer its referent they need to know that the pronoun is reflexive. This study examines whether discourse information is the information source that the learner might use to acquire grammatical categories of pronouns. Experimental results demonstrate that adults can use discourse information to accurately guess the referents of pronouns. Simulations show that a Bayesian model using guesses from the experiment as an estimate of the discourse information successfully categorizes English pronouns into categories corresponding to reflexives and non-reflexives. Together, these results suggest that knowing which entities are likely to be referred to in the discourse can help learners acquire grammatical categories of pronouns.
\end{abstract}

Keywords: language acquisition; Bayesian modeling
English speakers know that the sentence in (1) means that Alice saw Alice in the mirror and the sentence in (2) means that Alice saw someone else in the mirror.
(1) Alice saw herself in the mirror.
(2) Alice saw her in the mirror.

These interpretations reflect adults' knowledge that reflexives like herself require different syntactic relations with their antecedents than non-reflexives like her. Evidence shows that children acquiring various languages have knowledge of the grammatical distributions of pronouns (Jakubowicz, 1984; Crain \& McKee, 1985, among many). However, it is not yet known how children acquire this knowledge.

In English, the distribution of pronouns is governed by two constraints on the pronoun-antecedent relation: locality and c-command. Locality refers to the domain of the syntactic relation between the pronoun and its antecedent. Reflexive pronouns must have their antecedents in the local domain, corresponding approximately the same clause in English (Chomsky, 1973). The second constraint is that reflexive pronouns must be c-commanded by their antecedent (Reinhart, 1976). In the sentence 'Alice's sister saw herself', English speakers know that the antecedent of herself is not Alice, but Alice's sister. That is, when the hierarchical structure of the sentence is represented as a tree in Figure 1, the reflexive herself is contained in the sister node of its antecedent Alice's sister. Non-reflexive pronouns appear in exactly those


Figure 1: Syntactic tree showing a c-command relationship between the antecedent Alice's sister and the pronoun herself.
contexts where reflexives do not appear: approximately contexts in which the antecedent is either non-local, not in a ccommanding position, or both (Chomsky, 1973). This means that the relationship between the grammatical positions of antecedents and pronouns, as characterized by locality and ccommand, defines the distribution of grammatical categories of pronouns in English.

Critically, these grammatical constraints concern the relationship between a pronoun and its antecedent. This means that syntactic knowledge alone is insufficient for acquiring pronouns, because it cannot be applied without knowing the intended antecedent of a pronoun. In order to learn that herself is reflexive, learners need to interpret the sentence in (1) as 'Alice saw Alice', recognizing that Alice and herself corefer to the same entity. However, in order to interpret the meaning of the sentence, they need to use the knowledge that herself is a reflexive pronoun, whereas her is a non-reflexive pronoun. This circularity poses a potentially difficult problem for children acquiring language.

In this paper we show that discourse information can help learners categorize pronouns into appropriate distributional classes. If learners use discourse information to predict that the pronoun herself in (1) is likely to refer to Alice and the pronoun her in (2) is likely to refer to someone else, this provides information that can help them categorize these pronouns into different classes. We examine (i) to what extent discourse context is informative for determining the referent of a pronoun and (ii) whether this estimate of a pronoun's reference is sufficient for learning to classify pronouns.

The paper is organized as follows. Our next section describes a behavioral experiment that measures the discourse information available to listeners. The following section
presents Bayesian modeling results showing that this discourse information can help bootstrap grammatical knowledge of pronoun categories. Finally, the last section addresses open questions and implications.

\section*{Experiment 1: Human Simulation}

To test to what extent language contexts are informative about the referents of pronouns, we used a variant of the human simulation paradigm (Gillette, Gleitman, Gleitman, \& Lederer, 1999). In this paradigm, adult participants guess the identity of a missing word on the basis of linguistic and/or situational data. Because participants know there is a word, but not what it is, they are simulating what it is like to be a language learner who hears a word but does not know its meaning. A goal of the human simulation paradigm is to see what can be inferred about the meaning of a word based on information present either in the linguistic input or in the scene. Past experiments using the human simulation paradigm have examined the degree to which adults (Gillette et al., 1999; Kako, 2005) or older children (Piccin \& Waxman, 2007) can guess identities of common nouns and/or verbs.

In our experiment, adult participants were shown text excerpts of conversations between adults and children. Their task was to guess the identity of a word or phrase that had been blanked out, which was either a reflexive pronoun, a non-reflexive pronoun or a lexical noun phrase. The goal was to determine whether conversational context provides sufficient information for adults to guess what is being referred to. If so, this would provide evidence in favor of the idea that language learners can determine the referents of pronouns they do not yet know based on conversational context.

\section*{Methods}

Participants Participants were 40 undergraduates at the University of Maryland, College Park ( 11 men, 29 women). All were native English speakers and all were at least 18 years old. Participants were enrolled in introductory linguistics courses and received course credit for their participation.
Materials Text excerpts of real recorded conversations between adults and young children were taken from the ENG-USA section of the CHILDES database (MacWhinney, 2000). In each excerpt, one line was bolded. This bolded line had a noun phrase (NP) that had been deleted and replaced with a blank. The deleted noun phrase always came from an adult utterance. There were 12 lines of dialogue before the bolded line and six lines afterwards. Every deleted noun phrase was the object of one of five verbs: hurt, see, help, dry, and cover. Using the same verbs in all contexts allowed us to factor out any possible contribution of verb knowledge to determining which pronoun was intended. The deleted noun phrases belonged to one of three categories: 25 were reflexive pronouns ( 4 tokens of myself, 1 token of ourselves, 7 tokens of himself, 10 tokens of yourself, and 3 tokens of themselves), 25 were non-reflexive pronouns ( 4 tokens of \(m e, 1\) token of us, 7 tokens of him, 10 tokens of you, 3 tokens of them), and

25 were lexical NPs (names - including Mommy or Daddy - and definite descriptions). This led to a total of 75 test items. Within the test items, frequencies of corresponding non-reflexive and reflexive pronouns were matched (e.g., me was matched in frequency with myself, etc).

The dialogues were chosen randomly from all adult utterances in CHILDES that used the relevant verbs with the relevant type of NP object, with the exception that we threw out utterances that were direct repetitions of a previous line or that were well-known quotations. Finally, the materials were chosen to balance, as much as possible, the person of the pronoun object of the verb (though due to an imbalance in the available CHILDES data we were still left with more secondperson objects than first or third person). In addition to the lines of dialogue, each item in the experiment provided a list of participants in the conversation and the age of the child in the conversation. No information was given about the situation or context in which the conversation took place.

After each excerpt, participants were given a list of 15 choices for what NP could have gone in the blank. The choices always included the same five reflexive pronouns (yourself, myself, ourselves, himself, themselves) and nonreflexive pronouns (you, me, us, him, them). They also included five lexical NPs which would have been prominent in the conversation: e.g., the names of the participants (including Mommy or Daddy) and prominent people or objects mentioned in the conversational excerpts. If the actual sentence contained a lexical NP then this lexical NP was one of the five lexical NPs provided. The NPs were presented in alphabetical order.
Procedure Participants were given an hour in a quiet room to complete the experiment. Test items were presented on paper, one per page. Participants were instructed to read the dialogues, which were real conversations between adults and children, and pick the word or phrase (from the list of 15 choices) they thought belonged in the blank. Participants wrote answers on a separate answer sheet. The test items were presented in random order. Twenty participants received the first 38 test items, and the remaining twenty participants received the remaining 37 test items.

\section*{Results and Discussion}

Overall, participants were highly accurate at guessing the correct word from a list of 15 choices. The first row in Table 1 breaks up guesses of the correct word by syntactic category of the NP (reflexive pronouns, non-reflexive pronouns, or lexical NPs). Individual participants chose the correct NP out of 15 choices an average of \(63.8 \%\) of the time. This ranged from \(32.4 \%\) for the least accurate participant to \(84.2 \%\) for the most accurate participant, with a standard deviation of \(10.6 \%\), and was significantly better than chance \((t(39)=34.19, p<0.0001)\). These results show that adults can usually guess the identity of a missing NP given only a small amount of linguistic context.

However, these results underestimate participants' ability
\begin{tabular}{|c|c|c|c|}
\hline & Lexical NP & Non-reflexive & Reflexive \\
\hline \begin{tabular}{c} 
\% correct \\
word
\end{tabular} & 61.75 & 70.25 & 64.25 \\
\hline \begin{tabular}{c} 
\% plausibly \\
correct word
\end{tabular} & 66.75 & 81.25 & 68 \\
\hline
\end{tabular}

Table 1: Percentage of correct answers and answers with a plausibly correct referent in Experiment 1
\begin{tabular}{|c|c|c|c|}
\hline \begin{tabular}{c} 
\% LexicalNP \\
guesses
\end{tabular} & Lexical NP & Non-reflexive & Reflexive \\
\hline \begin{tabular}{c} 
\% Non-reflexive \\
guesses
\end{tabular} & 23.2 & 23.4 & 15.8 \\
\hline \begin{tabular}{c} 
\% Reflexive \\
guesses
\end{tabular} & 5 & \(\mathbf{7 3 . 4}\) & 16 \\
\hline
\end{tabular}

Table 2: Confusion matrix obtained in Experiment 1
to guess what is being referred to. The second row in Table 1 shows guesses of a plausibly correct word, a word that plausibly had the same intended referent as the correct word (for instance, a pronoun with the same gender/number features as the name that had actually been used, or vice versa). These results show that adults are good at guessing which entity is referred to given a context, irrespective of grammatical knowledge relevant to pronouns.

Table 2 breaks up the results by syntactic category of the NP. Participants' guesses were usually of the same category that the actual word had been. Importantly, adults usually guessed correctly whether the missing word had been a reflexive pronoun-when the word actually had been reflexive, participants guessed a reflexive \(68.2 \%\) of the time. When the word had been a lexical NP or a non-reflexive pronoun, they almost never guessed that it had been a reflexive.

This task parallels that of a child identifying an unfamiliar word. Of course, the parallel is not complete. In some ways, adult participants were provided with less information than the children they were meant to simulate: they only received a small excerpt of the conversation and did not receive any visual information. In other ways, the participants had more data: they already knew the meanings of all of the other words in the conversation, they had full syntactic and discourse knowledge where children might only have partial knowledge (e.g., Arnold, Brown-Schmidt, \& Trueswell, 2007), and they were limited to 15 choices of possible meaning. Furthermore, choosing an answer in this experiment was not subject to any time pressures, whereas in actual acquisition processing speed could potentially impact the learner's ability to use the discourse context as an information source. However, to the extent that the adult simulation reflects the prior information presented in the discourse, it provides an estimate of the information that children might have access to. Where adults (who already know the distribution of reflexives) can guess that a missing word is reflexive, a child might be able to guess that a missing word co-refers with a specific NP. Together with syntactic knowledge of locality and c-command, this should provide learners with useful in-
formation for acquiring grammatical categories of pronouns. To explore this possibility, we formalize a Bayesian model that learns to categorize pronouns.

\section*{Experiment 2: Bayesian Model}

In this section, we develop a Bayesian model that integrates the discourse information measured in Experiment 1. This model investigates whether the information in discourse could be sufficient to learn the grammatical categories of English pronouns in principle (a computational-level model; Marr, 1982). The model discovers:
1. how many pronoun categories there are in a language
2. the distribution of pronouns in each category
3. which syntactic position of an antecedent is associated with each pronoun category

This ideal learner is assumed to have (a) discourse knowledge that helps define the distribution of the potential antecedents, (b) syntactic knowledge relevant to pronoun categories (details follow), and (c) lexical knowledge that is sufficient for distinguishing pronouns from lexical noun phrases. Other linguistic information relevant to pronouns, such as gender and number, is not represented in our model; we ask simply whether our ideal learner can acquire two categories corresponding to reflexive and non-reflexive pronouns.

Regarding (b) above, this ideal learner is assumed to already know locality and c-command before learning pronoun categories, and is further assumed to know that these are relevant for categorizing pronouns. Thus, the learner is able to identify the syntactic position of each potential antecedent. The model distinguishes four syntactic positions based on the knowledge of locality and c-command; [+local,+c-command], [+local,-c-command], [-local,+c-command], and [-local,-c-command]. In English, if an antecedent is in a syntactic position described by [+local,+c-command], that pronoun must be a reflexive pronoun. If the potential antecedent is elsewhere, that pronoun must be a non-reflexive pronoun. However, the learner does not know in advance which syntactic position is associated with which pronoun category, and needs to acquire this knowledge from the input. We return to this issue of prior syntactic knowledge in the Discussion.

\section*{Generative Model}

Our model assumes the following generative process. For each pronoun, an antecedent in one of the four syntactic positions described above is chosen given prior discourse knowledge \((\mathcal{D})\). Then a pronoun category is chosen based on the syntactic position of the antecedent, and a pronoun is generated from the chosen pronoun category.

Figure 2 illustrates this process with a graphical model. \({ }^{1}\) Each antecedent category distribution \(\theta_{j}\) is a random variable

\footnotetext{
\({ }^{1}\) This model is a nonparametric extension to the author-topic model (Rosen-Zvi, Griffiths, Steyvers, \& Smyth, 2004) that allows for an infinite number of categories (called topics in their model).
}


Figure 2: Graphical Model
that encodes the distribution over pronoun categories favored by an antecedent in syntactic position \(j\). For example, the category reflexive would have high probability in the distribution \(\theta_{[+ \text {local },+\mathrm{c}-\text { command }]}\). (Here we use the category name reflexive for exposition, but the model does not associate any labels with the pronoun categories it recovers.) Each category word distribution \(\phi_{k}\) is a distribution over words that encodes the probability distribution over pronouns in pronoun category \(k\). For example, pronouns such as herself and myself would have high probabilities in the distribution \(\phi_{\text {reflexive }}\). In addition to learning this distribution, our model learns the number of pronoun categories needed to describe the data. For each pronoun in the corpus, an antecedent in a syntactic position \(x\) is assumed to be sampled from a distribution we refer to as discourse knowledge \(\mathcal{D}\) (see the next section for the details). A pronoun category \(z\) is then sampled from the multinomial distribution with parameter \(\theta\) associated with the syntactic position \(x\) of the antecedent and a pronoun \(w\) is sampled from a multinomial distribution with parameter \(\phi\) associated with pronoun category \(z\).

To learn the number of pronoun categories based on the observed data, we use the hierarchical Dirichlet process (Teh, Jordan, Beal, \& Blei, 2006). The advantage of this model is that it allows the model to flexibly learn how many categories are present in the data, while still allowing categories to appear across multiple grammatical contexts. The summary of the generative process follows.
1. Draw a distribution over pronoun categories \(\theta_{0} \sim \operatorname{GEM}(\gamma)\), where GEM is the Griffiths, Engen, McCosky distribution (Pitman, 2002).
2. For each antecedent syntactic position \(j=1 \ldots 4\), draw a pronoun category distribution \(\theta_{j} \sim \operatorname{DP}\left(\alpha, \theta_{0}\right)\).
3. For each pronoun category \(k=1 \ldots \infty\), draw a distribution over tokens \(\phi_{k} \sim \operatorname{Dir}(\beta)\).
4. For each pronoun in the corpus \(n=1 \ldots N\)
(a) Draw an antecedent syntactic position from the discourse knowledge \(x_{n} \sim \mathcal{D}\)
(b) Draw a pronoun category \(z_{n} \sim \operatorname{Mult}\left(\theta_{x_{n}}\right)\).
(c) Draw a word \(w_{n} \sim \operatorname{Mult}\left(\phi_{z_{n}}\right)\).

\section*{Prior Knowledge}

The observed discourse knowledge \(\mathcal{D}\) defines a prior distribution over potential antecedents in the discourse. Recall that our ideal learner maps each antecedent in the discourse deterministically to its syntactic position (defined in terms of locality and c-command), and in this way \(\mathcal{D}\) defines a distribution over syntactic positions \(x\) for each pronoun's antecedent.

Rather than specify a parametric form for this prior distribution, we estimate it directly from participants' responses from Experiment 1. In one experimental item, for example, participants guessed the identity of the missing word in the sentence "You drying _- off?". Nine out of 20 participants guessed that the missing word is yourself, six out of 20 guessed him, three out of 20 guessed me, and two out of 20 guessed Seth. Under the assumption that experimental participants have sampled their responses from a shared prior distribution over entities in the discourse, these guesses provide an estimate of participants' beliefs about how likely each entity is to be referred to in the discourse.

Where participants chose yourself, the antecedent of this pronoun is you, which is a local and c-commanding antecedent. Where participants chose him and me, the antecedents could be in any of the remaining three syntactic positions, but in this particular dialogue the only potential antecedents for non-reflexives are neither local nor ccommanding. \({ }^{2}\) We ignored responses in which participants chose lexical NPs (here Seth) based on the assumption that learners distinguish pronouns from lexical NPs. We then normalized each count by the total number of pronoun guesses. The resulting prior distribution over syntactic positions for antecedents in this example is \(p\left(x_{[+ \text {local },+c \text {-command }]} \mid \mathcal{D}\right)=0.5\) and \(p\left(x_{[- \text {local },-c \text {-command }]} \mid \mathcal{D}\right)=0.5\). In this way the results from Experiment 1 provide us with an informative prior distribution regarding which entities are likely to be referred to in the discourse, and through simulations we can test whether this prior knowledge helps an ideal learner acquire pronoun categories \({ }^{3}\).

\section*{Inference}

Given this generative process, we can use Bayesian inference to recover the learner's beliefs about pronoun categories. We use the Gibbs sampling algorithm from Rosen-Zvi et al. (2004) to estimate latent variables: the antecedent-category parameter \(\theta\), the category-word parameter \(\phi\), the antecedent's syntactic position \(x\), and the pronoun category \(z\). The assignments of \(x\) and \(z\) for a particular token are sampled as a block, conditioned on everything else, so that in each iteration we compute the conditional distribution \(p\left(x_{i}, z_{i} \mid w_{i}, \mathbf{x}_{-i}, \mathbf{z}_{-i}\right)\) where \(\mathbf{x}_{-i}\) and \(\mathbf{z}_{-i}\) denote all syntactic position and category

\footnotetext{
\({ }^{2}\) In cases of non-reflexive guesses where potential antecedents appeared in multiple syntactic positions, we assumed the prior probability for each syntactic position to be proportional to the number of potential antecedents in that position.
\({ }^{3}\) This prior distribution differs from the distribution seen in Tables 1 and 2 because it is based on individual experimental items rather than on aggregated data.
}
assignments not including the \(i\) th pronoun. This is proportional to
\[
p\left(w_{i} \mid x_{i}, z_{i}, \mathbf{x}_{-i}, \mathbf{z}_{-i}\right) \cdot p\left(z_{i} \mid x_{i}, \mathbf{x}_{-i}, \mathbf{z}_{-i}\right) \cdot p\left(x_{i} \mid \mathcal{D}\right)
\]
where the first term is the likelihood function from RosenZvi et al. (2004), the second is defined by the hierarchical Dirichlet process (Teh et al., 2006), and the third is estimated directly from participants' responses in Experiment 1.

\section*{Simulations}

In order to test the effectiveness of discourse information for the categorization of pronouns, our simulations compare three models: a Baseline model, a Discourse model, and a Strong syntax model. The Baseline model has information about locality and c-command, but it lacks information about which entities are likely to be referred to in the discourse. It assumes that potential antecedents are sampled uniformly, so that \(p\left(x_{i} \mid \mathcal{D}\right)\) is defined by counting the number of discourse entities that appear in each syntactic position. The Discourse model is identical to the Baseline model, but it contains the adult-like discourse knowledge estimated in Experiment 1, as described above. Comparing the performance of the Discourse model to the Baseline model allows us to quantify the degree to which discourse information helps an ideal learner acquire pronoun categories.

The Strong syntax model is similar to the Baseline model in that it assumes that potential antecedents are sampled uniformly, but it additionally incorporates built-in knowledge of the grammatical constraints on reflexive and non-reflexive pronouns in English. This model knows there are two grammatical categories of pronouns. Furthermore, it knows that pronouns that have local c-commanding antecedents are reflexive pronouns and that pronouns that do not have local c-commanding antecedents are non-reflexive pronouns (i.e., the antecedent-category parameter \(\theta\) is observed). Thus, the model only needs to learn the distribution of each category over pronouns. Comparing this Strong syntax model to the Baseline model allows us to examine whether this type of strong prior syntactic knowledge is sufficient to help learners categorize pronouns.

Each model was trained on 50 dialogues from Experiment 1, 25 with reflexive and 25 with non-reflexive pronouns. For each dialogue, the model was provided with the pronoun, a prior distribution over possible antecedents for that pronoun, and the syntactic positions of those antecedents relative to the pronoun. Through the unsupervised learning procedure described above, the models recovered a distribution over categories associated with each syntactic position and a distribution over pronouns for each category.

Results For each model, we ran 10 independent Gibbs chains for 2000 iterations each. Hyperparameters \(\alpha, \beta\), and \(\gamma\) were fixed at 1.0, 0.01, and 0.001 , respectively \({ }^{4}\). We com-

\footnotetext{
\({ }^{4}\) We chose the best parameter values based on multiple runs, but results were qualitatively consistent across a range of parameter values. The same parameter values were used for all three models.
}
\begin{tabular}{|c|c|c|c|}
\hline \multicolumn{2}{|l|}{Category 1} & \multicolumn{2}{|l|}{Category 2} \\
\hline Word & p(word|category) & Word & p(word|category) \\
\hline him & 0.5 & yourself & 0.4 \\
\hline me & 0.29 & himself & 0.28 \\
\hline them & 0.21 & myself & 0.16 \\
\hline us & 0.0 & themselves & 0.12 \\
\hline you & 0.0 & us & 0.04 \\
\hline myself & 0.0 & me & 0.0 \\
\hline yourself & 0.0 & you & 0.0 \\
\hline himself & 0.0 & him & 0.0 \\
\hline themselves & 0.0 & them & 0.0 \\
\hline ourselves & 0.0 & ourselves & 0.0 \\
\hline & \multicolumn{2}{|l|}{Category 3} & \\
\hline & Word & p(word|category) & \\
\hline & you & 0.91 & \\
\hline & ourselves & 0.09 & \\
\hline & me & 0.0 & \\
\hline & us & 0.0 & \\
\hline & him & 0.0 & \\
\hline & them & 0.0 & \\
\hline & myself & 0.0 & \\
\hline & yourself & 0.0 & \\
\hline & himself & 0.0 & \\
\hline & themselves & 0.0 & \\
\hline
\end{tabular}

Table 3: Baseline model results
puted pairwise F-scores using the final samples from each chain. The Baseline model consistently failed to learn the correct categories, achieving a mean pairwise F-score of 0.55 across the 10 sampling chains. In all 10 chains, the model learned 3-4 categories, where the correct number of categories is two. Table 3 shows the distribution over pronouns belonging to each category obtained at the 2000th iteration of the sampling run with the highest likelihood. The maximum likelihood estimate \(p\) (word \(\mid\) category) gives the proportion of times each pronoun occurs in a category, based on a single sample from the posterior distribution over \(z\) and \(x\).

The Discourse model performed much better than the Baseline model, achieving a mean pairwise F-score of 0.97 across the 10 sampling runs. In seven of the 10 runs, the model perfectly categorized English pronouns into two classes. In two additional runs, the model learned two categories, but the membership was not consistent. In the final run, the model learned three categories. Table 4 shows the pronouns belonging to each category, obtained at the 2000th iteration of the Gibbs sampling run which had the highest likelihood. The pronouns associated with each category are reflexive pronouns and non-reflexive pronouns, respectively. This model also learned that there are exactly two categories, as expected. These results indicate that discourse information can help an ideal learner categorize pronouns.

Although the Baseline model has prior knowledge of ccommand and locality, it is still possible that the low performance in this model might result from insufficient syntactic knowledge. For this reason, we compare the Strong syntax model with the Baseline model to see whether even stronger prior syntactic knowledge is sufficient for categorizing pronouns. The mean F-score was 0.56 for this model. Table 5 shows the pronouns in each category, obtained at the 2000th iteration of a Gibbs sampling run which had the highest likelihood. The lack of improvement of the Strong syntax model over the Baseline model suggests that simply having strong prior syntactic knowledge is not sufficient for acquiring grammatical categories of pronouns.
\begin{tabular}{|l|l|l|l|}
\hline \multicolumn{2}{|l|}{ Category 1 } & \multicolumn{2}{l|}{ Category 2 } \\
\hline Word & p(word category) & Word & p(word|category) \\
\hline yourself & 0.4 & you & 0.4 \\
himself & 0.28 & him & 0.28 \\
myself & 0.16 & me & 0.16 \\
themselves & 0.12 & them & 0.12 \\
ourselves & 0.04 & us & 0.04 \\
you & 0.0 & yourself & 0.0 \\
him & 0.0 & himself & 0.0 \\
me & 0.0 & myself & 0.0 \\
them & 0.0 & themselves & 0.0 \\
us & 0.0 & ourselves & 0.0 \\
\hline
\end{tabular}

Table 4: Discourse model results

These simulation results suggest that knowing which entities are likely to be referred to in the discourse can help learners acquire grammatical categories of pronouns. On the other hand, simply having strong prior knowledge about the grammatical distribution of pronouns is not sufficient to support the acquisition of pronoun categories.

\section*{Discussion}

This study examined the potential utility of discourse information as a cue to the acquisition of pronoun categories. We showed that discourse information can help adults accurately guess the identities of missing pronouns, and that a Bayesian model with prior knowledge of discourse information can accurately recover grammatical categories of pronouns without knowing in advance how many categories are present in a language. This supports a role for discourse information in helping learners acquire grammatical knowledge of pronoun categories and shows one way in which they can overcome the circularity problem inherent to language acquisition at the syntax-semantics interface.

While it is possible that hearing a few unambiguous sentences could also be sufficient for acquiring pronoun categories, our analysis shows that this type of unambiguous data may not be required. Instead, an ideal learner can achieve the same outcome by relying on the discourse information that is actually present in child-directed speech. The data used in our analysis were taken from CHILDES, and therefore provide a good characterization of input a child receives. However, one limitation of our work is that distributions of verbs and pronouns were balanced in our experimental stimuli, whereas they may not be balanced in the input. To ensure that the true distributions of verbs and pronouns support learning, it will be important to replicate our modeling results on more extensive corpora.

Our model assumed that learners have prior knowledge of the relevance of syntactic locality and c-command relations to the acquisition of pronouns, but we do not know the degree to which this parallels children's acquisition. Children appear to have acquired relevant locality constraints on pronouns by age five at the latest (Zukowski, McKeown, \& Larsen, 2008), though we do not know when knowledge of the domains themselves becomes available to learners. Knowledge of c-command also appears to be available to children at this age or even earlier (Lidz \& Musolino, 2002; Sutton, Fetters, \& Lidz, 2012). However, cross-linguistically, locality and c-
\begin{tabular}{|l|l|l|l|}
\hline \multicolumn{2}{|l|}{ Category 1 } & \multicolumn{2}{|l|}{ Category 2 } \\
\hline Word & p(word|category) & Word & p(word|category) \\
\hline yourself & 0.29 & you & 0.63 \\
him & 0.21 & me & 0.25 \\
himself & 0.21 & us & 0.06 \\
myself & 0.12 & ourselves & 0.06 \\
them & 0.09 & him & 0.0 \\
themselves & 0.09 & them & 0.0 \\
me & 0.0 & myself & 0.0 \\
us & 0.0 & yourself & 0.0 \\
you & 0.0 & himself & 0.0 \\
ourselves & 0.0 & themselves & 0.0 \\
\hline
\end{tabular}

Table 5: Strong syntax model results
command are neither necessary nor sufficient for defining the distributions of grammatical categories of pronouns. Future modeling work will explore the potential role of discourse as an evidentiary source not only in discovering categories of pronouns, but also in determining which grammatical features are relevant for anaphoric dependencies.

Acknowledgments. We thank Viet-An Nguyen, Ke Zhai, Motoki Shiga, and members of the UMD Computational Psycholinguistics group for helpful comments and discussion. This research was supported by NSF IGERT 0801465 and NSF grant 1018625 (JBG).

\section*{References}

Arnold, J., Brown-Schmidt, S., \& Trueswell, J. (2007). Children's use of gender and order of mention in pronoun processing. Language and Cognitive Processes, 22, 527-565.
Chomsky, N. (1973). Conditions on Transformations. In S. R. Anderson \& P. Kiparsky (Eds.), A festschrift for Morris Halle. New York: Holt, Rinehart \& Winston.
Crain, S., \& McKee, C. (1985). The acquisition of structural restrictions on anaphora. In S. Berman, J. Choe, \& J. McDonough (Eds.), Proceedings of NELS (Vol. 15). Amherst, Mass: GLSA.
Gillette, J., Gleitman, H., Gleitman, L., \& Lederer, A. (1999). Human simulations of vocabulary learning. Cognition, 73, 135-176.
Jakubowicz, C. (1984). On markedness and the binding principles. In Proceedings of NELS (Vol. 14, pp. 154-182).
Kako, E. (2005). Information sources for noun learning. Cognitive Science, 29, 223-260.
Lidz, J., \& Musolino, J. (2002). Children's command of quantification. Cognition, 84, 113-154.
MacWhinney, B. (2000). The CHILDES Project: Tools for analyzing talk. Third Edition. (Tech. Rep.). Mahwah, NJ: Lawrence Erlbaum Associates.
Marr, D. (1982). Vision. San Francisco, CA: W. H. Freeman.
Piccin, T. B., \& Waxman, S. R. (2007). Why nouns trump verbs in word learning: New evidence from children and adults in the Human Simulation Paradigm. Language Learning \& Development, 3(4), 295-323.
Pitman, J. (2002). Poisson-Dirichlet and GEM invariant distributions for split-and-merge transformations of an interval partition. Combinatorics, Probability and Computing, 11, 501-514.
Reinhart, T. (1976). The Syntactic Domain of Anaphora. Unpublished doctoral dissertation, MIT.
Rosen-Zvi, M., Griffiths, T., Steyvers, M., \& Smyth, P. (2004). The author-topic model for authors and documents. In 20th Conference on Uncertainty in Artificial Intelligence.
Sutton, M., Fetters, M., \& Lidz, J. (2012). Parsing for Principle C at 30 months. In Proceedings of the 36th Boston University Conference on Language Development (pp. 581-593).
Teh, Y. W., Jordan, M. I., Beal, M. J., \& Blei, D. M. (2006). Hierarchical Dirichlet processes. Journal of the American Statistical Association, 101.
Zukowski, A., McKeown, R., \& Larsen, J. (2008). A tough test of the locality requirement for reflexives. In Proceedings of the 32nd Boston University Conference on Language Development (pp. 586-597).

\title{
Constraints on Theories of Serial Order Memory Revisited: The Cases of the Fill-In and Protrusion Effects
}

\author{
Adam F. Osth (adamosth@gmail.com) \\ Simon Dennis (simon.dennis@gmail.com) \\ 206 Psychology Building \\ 1835 Neil Avenue \\ Columbus, OH 43210 USA
}

\begin{abstract}
In his seminal dissertation, Henson (1996) identified a number of constraints on theories of serial order memory. Two constraints, the fill-in constraint, in which an item that is erroneously recalled early is likely to be followed by its predecessor rather than its successor (recall of ACB is more likely than ACD), and the protrusion constraint, in which prior list intrusions are likely to be recalled in the same output position as their previous serial position, were considered evidence against chaining theories. We present results from two experiments which investigate the extent to which these effects are dependent on experimental methodology. When participants are given an open set of items, an equal ratio of fill-in and infill errors was observed and a protrusion effect was obtained. However, when a reconstruction of order task was used, a fillin effect was observed. Implications for theories of serial order memory are discussed.
\end{abstract}

Keywords: serial recall; serial order memory; episodic memory; short term memory; working memory; chaining models; positional models; memory models

The problem of understanding serial order memory is possibly the oldest problem in the field of memory studies. Over a century of work has produced theories that can be broadly placed into two camps: positional theories, which claim that items in a sequence are coded with respect to their position in the sequence, and chaining theories, which claim that order is implicit in the associations among the items in a sequence; retrieval consists of using retrieval retrieved items as cues for their successors.

A watershed moment in theorization of serial order memory came with the seminal dissertation of Henson (1996). Through a series of experiments and meta-analyses, Henson outlined the regularities of the serial recall paradigm and dubbed them the constraints on serial order memory. The discussed regularities were diverse and ranged from effects of serial position to repetition errors and omissions. Of interest to the present investigation were regularities in error patterns that were particularly problematic for chaining models of serial order memory. One such regularity was the fill-in effect, which states that when an item is recalled one position too early, it is more likely to be followed by its predecessor than its successor. As an example, consider when a sequence \(A B C D E\) is studied and a subject erroneously recalls \(A C\), skipping over the \(B\) item. Henson found that participants are more likely to recall \(A C B\) than \(A C D\), as if they are "filling in" the missing response. The latter, the \(A C D\) case, is considered an in-fill error, and any chaining model with asymmetric associations is predicted to produce a greater incidence of in-fill errors than fill-in errors. The fill-in effect was replicated by

Surprenant, Kelley, Farley, and Neath (2005) using a reconstruction of order paradigm.

Another such constraint was the protrusion constraint, which states that intrusions from prior lists tend to share the same position in both the current and prior lists. That is, if a subject is attempting to recall to item 3 on the current list and makes an intrusion from the prior list, the intrusion is most likely to be the third item on the previous list. Henson (1996) dubbed such in-position intrusions protrusions. The protrusion effect was initially discovered by Conrad (1960) and has been cited as evidence for positional coding.

Both the protrusion and fill-in constraints along with several other regularities that Henson (1996) discovered became canonical in the serial recall literature. Virtually every model that has been published since has no role of inter-item associations. The majority of the recent models are positional models; these include the the model of Burgess and Hitch (1999, 2006), Henson's own Start-End Model (Henson, 1998), the recurrent model of Botvinick and Plaut (2006), and the grouping model (Farrell, 2012). The constraints outlined by Henson (1996) were benchmark phenmonena for each of these models, as several of these models demonstrated how both fill-in and protrusion effects could be explained by positional representations.

The role of chaining in memory for serial order was revisited by Solway, Murdock, and Kahana (2012), who conducted a re-analysis of three serial recall studies and found a robust in-fill effect, contrary to the analyses of Henson (1996) and in agreement with the predictions of chaining models. They proposed a compound chaining model which yielded a good fit to the data, whereas the positional model of Burgess and Hitch (2006) did not yield an adequate fit. An open question remains as to why the analysis of Solway et al. yielded an in-fill effect whereas the analyses of Henson yielded the opposite pattern. In response to Solway et al., Farrell, Hurlstone, and Lewandowsky (2013) presented a re-analysis of over a dozen datasets, the vast majority of which yielded a fill-in effect. Farrell et al. posited that the datasets used by Solway et al. used longer list lengths than most serial recall experiments (minimum of ten items), which might be beyond the capacity of short-term memory and thus utilize different representations.

However, another difference between the studies analyzed by Solway et la. and Farrell et al. is that in the former, items were not re-used across trials (this is referred to as an open
set of items). The studies considered by Henson and Farrell et al. often used small sets of stimuli such that items were frequently re-used across trials (referred to as a closed set of items). This is often standard practice in serial recall experiments, as it is believed that when a large set of stimuli is used the participant has to remember both the items and the order in which they occurred. If a closed set of stimuli is employed, item memory quickly reaches ceiling and only the order among the items has to be remembered on a given trial (Healy, 1974). Another similar approach to obtaining a relatively "pure" measure of order memory is to employ a reconstruction of order task, wherein the participant studies a list of items but at retrieval they are provided with the items and asked to place them in the correct order (Healy, Fendrich, Cunningham, \& Till, 1987); this task was used in the experiments of Surprenant et al. (2005) that demonstrated a fill-in effect.

Closed sets and reconstruction of order tasks may yield different results than serial recall experiments with open sets in that the former tasks are susceptible to guessing strategies and the latter has a high incidence of omissions. When a closet set or a reconstruction task is used, item memory is at ceiling, which may cause participants to guess as to the locations of the items without having any knowledge of their positions. Under some circumstances, this can lead to an artificially high degree of fill-in errors. Consider a case in which a participant studied a list \(A B C D\) and knows the locations of \(A\) and \(B\), but cannot recall the positions of \(C\) and \(D\) and guesses on the third and fourth responses. In this circumstance, the participant can either a.) get the sequence correct ( \(A B C D\) ) or b.) produce a sequence with a fill-in error \((A B D C)\). Under this circumstance, only fill-in errors can be produced and there is no possibility for making an in-fill error.

However, when an open set is used in a serial recall task, item memory is imperfect and there is a high incidence of omissions, which can disguise a fill-in effect as an in-fill effect. Consider if sequence \(A B C D\) is represented in memory as \(A C B D\). This would yield a fill-in effect if all of the represented items were output, but if \(B\) is omitted at retrieval, sequence \(A C D\) is produced. This may be especially likely in the studies of Solway et al. since the lists in the experiments they analzyed were quite long and the performance of the participants was relatively poor.

Even if guessing strategies are responsible for the finding of the fill-in effect, the compound chaining model of Solway et al. has no recourse for producing protrusion errors. Nonetheless, a number of traditional studies of serial recall collect responses on a series of lined response grids in which participants can see the previous responses. A simple explanation for the protrusion effect is that when participants are not able to recall a given item, they glance at the output position from the previous trial and use that response in their answer.

The present study attempt to control for both of these possibilites using simple means. In Experiment 1, we conducted
a serial recall experiment that used a large open set of words, such that items are not repeated from trial to trial making it unlikely that participants should be able to guess from a pool of possible responses. Additionally, participants entered their responses on a keyboard rather than on lined response grids and were not given access to their previous responses. In Experiment 2, we used the same procedure as Experiment 1 but employed a reconstruction task by showing the participants the list items at retrieval in a new randomized order.

Serial recall experiments using open sets have been conducted, but to date there have been no analyses of whether the data collected from these experiments exhibit a predominance of fill-in or in-fill errors or whether they exhibit the protrusion effect. Investigations of set size effects have instead been concerend with other issues, such as word frequency effects (Roodenrys \& Quinlan, 2000) or the phonological similarity effect (Coltheart, 1993; Conrad, 1963). We found that the open set did affect the fill-in to in-fill ratio, as we observed equivalent levels of fill-in and in-fill errors in Experiment 1. In Experiment 2, we conducted a reconstruction of order task and replicated the fill-in effect found by Surprenant et al. (2005).

\section*{Experiment 1: Serial Recall}

Experiment 1 was an immediate serial recall task using a large set of words such that every list a participant received was composed entirely of unique items. To capture different levels of performance, we employed lists of five and six items and list length was manipulated between subjects.

\section*{Method}

Participants A total of 204 undergraduate psychology students (105 participants for the list length five condition, 99 for the list length six condition) from The Ohio-State University participated in this experiment in exchange for course credit in an introductory psychology course.
Materials Participants studied words that were randomly selected from a word pool of 625 words from the Google search counts. Words ranged from 4 to 7 letters in length and from 30 to 250 counts per million in word frequency.

Procedure To familiarize participants with the nature of the task, all participants began each session with four unscored practice lists of three, four, five, and six items in order. Participants were given feedback upon completion of each of the practice lists. If any errors were made, participants were reminded that they have to recall the items in the order in which they were presented. No feedback was given at any point in the experiment after the practice session was completed.

Upon completion of the practice session, participants were given 62 trials with lists of either five or six items. During the study phase, participants were presented with each word for one second followed by a blank screen for 250 ms . Following completion of the study list, participants were presented with a recall prompt that was a series of three question marks ("???") on the center of the screen. Participants were
instructed to begin recalling the items upon seeing the prompt by typing their respones on a keyboard and were given \(20 \mathrm{sec}-\) onds to recall the sequence. After the first key was pressed, the question marks disappeared and replaced with the letters typed by the participant. Participants signaled completion of a word by hitting the "ENTER" key on the keyboard. Upon completing a response, the response disappeared from view on the computer screen and was replaced by the question marks. Participants signaled completion of the recalled sequence by typing the word "done" and hitting "ENTER." Upon completion of each trial, participants signified readiness to begin the next trial by hitting the "ENTER" key.

Halfway through the experiment, participants were given a break in which they played a digital card game for 180 sec onds. Stimulus presentation and response collection was handled using the Python experimental library (Geller, Schleifer, Sederberg, Jacobs, \& Kahana, 2007).

\section*{Results}

Serial Position Effects The proportion of correctly items recalled in their correct serial position are shown in Figure 1. Performance was significantly worse in the length six condition than in the length five condition, \(t(201.37)^{1}=8.26, p<\) .001. A mixed analysis of variance (ANOVA) using serial position as a within subjects factor and list length as a between subjects factor revealed that performance varied as a function of serial position, \(F(5,906)=850.40\), and there was a serial position by length interaction, \(F(4,906)=34.99\), both \(p \mathrm{~s}<\) .001. Post hoc comparisons revealed a negative recency effect (poorer performance for the last item than the second to last item) in both the length five, \(t(104)=-10.99\), and length six, \(t(98)=-14.12\), conditions, both \(p \mathrm{~s}<.001\).

Fill-In and In-Fill Errors Henson (1996) classified fill-in errors by focusing on all responses following the first item that was recalled one position too early. If the following item was from one serial position earlier than the just recalled item, this was considered a fill-in error. If the following item was from one serial position later than the just recalled item, this was considered an in-fill error. However, one need not just consider transitions that only traverse one serial position. Rather, the initial error could be more than one position too early and the subsequent fill-in or in-fill could be more than one position earlier or later, respectively. We will henceforth refer to the one position restriction as the lag 1 analysis and the latter case which considers longer transition as the any lag case; both will be considered in the present analysis.

The analyses of Solway et al. (2012) and Farrell et al. (2013) both used a strict scoring procedure in which fill-in errors and in-fill errors are only considered for cases where the initial skip was the first error in the trial. We follow their example here and use strict scoring in our analyses.

The mean number of fill-in and in-fill errors for each classification style can be seen in Figure 2. Separate mixed

\footnotetext{
\({ }^{1} t\) test degrees of freedom (df) are corrected df from the WelchSatterwaithe equation.
}


Figure 1: Serial position curves for Experiment 1 (A), which used a traditional serial recall procedure, and Experiment 2(B), which used a reconstruction task. Error bars indicate \(95 \%\) within subjects confidence intervals.

ANOVAs with error type (fill-in or in-fill) as a within subjects factor and list length as a between subjects factor for both the any lag and lag 1 transition analyses. Results indicated there was neither a fill-in effect or an in-fill effect when transitions of any lag were analyzed, \(F(1,202)=.246\), or when analyses were restricted to lag 1 transitions, \(F=1.291\), both \(p \mathrm{~s}>.05\).
Protrusion Errors All analyses on intrusion rates were restricted to immediate intrusions (intrusions from only one list prior to the current study list). A visualization of the intrusion rates for the serial position of each intruding item at each output position can be seen in Figure 3. As can be seen by the spiked nature of the graph, intrusions tend to appear in the same output position as their serial position in the prior list.

A statistical analysis was performed by calculating the proportion of immediate intrusions that were protrusions for each participant. Because there were some participants that exhibited no prior list intrusions, this analysis was restricted to participants that made at least one immediate intrusion error. The proportion of immediate intrusions that were protrusions (in-position) was above chance for both the list length five, \(t(69)=4.49\), and list length six conditions, \(t(82)=3.50\), both \(p \mathrm{~s}<.001\), which is consistent with the findings of Henson (1996) and Conrad (1960).

\section*{Discussion}

We conducted a serial recall experiment that used experimental parameters that were highly similar to previous experiments with the exception that stimuli were not reused from trial to trial in an effort to gauge the the generality of the constraints on serial recall established by Henson (1996). In contrast to Henson's data and the analyses of Solway et al., we observed roughly equivalent numbers of fill-in and in-fill er-


Figure 2: Mean number of fill-in and in-fill errors for each classification type for list length 5 (top) and list length 6 (middle) conditions of Experiment 1, along with the reconstruction task used in Experiment 2 (bottom). Error bars indicate \(95 \%\) within subjects confidence intervals.
rors. In other words, when an item was recalled too early, participants were equally likely to continue in a forward order or to return to the skipped item. The protrusion effect, in contrast, was robust in our experiment. The spiked, positional nature of immediate intrusions strongly resembled the data displayed by Henson \((1996,1998)\).

Nonetheless, it was still somewhat surprising that we did not observe a fill-in effect in our experiment because Surprenant et al. (2005) found a robust fill-in effect with both a small and large set of items. Nonetheless, their experiments used a reconstruction of order paradigm, and as previously discussed, such a task might introduce guessing strategies in the same manner as a small set size would in a standard serial recall paradigm. Thus, we repeated our experiment using the same parameters but employed a reconstruction of order task at retrieval and hypothesized that the results would favor a fill-in effect.

\section*{Experiment 2: Reconstruction of Order}

Experiment 2 was identical to that of Experiment 1 with the exception that a reconstruction of order task was used at retrieval. During the test phase, participants were presented with all of the study list items in a randomized order and told to recall the words in the order in which they were presented. Because performance was far superior to that of the serial recall task in Experiment 1, we only collected data with a list length of six items.

\section*{Method}

Participants A total of 95 undergraduate psychology students from The Ohio-State University participated in this ex-


Figure 3: Proportions of immediate intrusions from each serial position in the prior list for each output position of the current trial for both the list length five (top) and list length six (bottom) conditions. The serial position of the prior list intrusion is indicated by the numbers on the lines.
periment in exchange for course credit in an introductory psychology course.

Procedure The procedure was identical to that of Experiment 1 with the exception that at retrieval, participants were presented with all of the studied words at the top of the screen in a randomized order. The words remained fixed on the screen through the test and did not disappear as participants entered the corresponding words. Participants were instructed in advance of the experiment that they would be presented with the words and were told that they should type out the words in the order in which they had appeared in the study list.

\section*{Results}

Due to the severe rarity of prior list intrusions in Experiment 2, analyses were restricted to serial position effects and fill-in and in-fill errors.

Serial Position Effects A one way repeated measures ANOVA revealed a main effect of serial position, \(F(5,470)\) \(=225.6, p<.001\). Post hoc inspection of the serial position data revealed a positive recency effect for the last item in the list, \(t(94)=6.45, p<.001\).
Fill-in and In-Fill Errors Mean number of fill-in and infill errors were calculated for each participant and can be seen in the bottom panel of Figure 2. Fill-in errors outnumbered in-fill errors when any lag transitions were considered, \(t(94)\) \(=14.02\), and when only lag 1 transitions were considered, \(t\) \(=9.80\), both \(p \mathrm{~s}<.001\). Thus, in contrast to Experiment 1, a fill-in effect was observed.

\section*{Discussion}

Consistent with our hypothesis, usage of a reconstruction of order task at retrieval produced a robust fill-in effect. Our results are consistent with the findings of Surprenant et al. (2005), who found a fill-in effect with a large set size using a reconstruction task.

\section*{General Discussion}

The two experiments we conducted found that a previously established regularity of serial order memory, the fill-in effect, depended on the type of task used to gauge serial order memory. In Experiment 1, we used a traditional serial recall paradigm with the exception that items were not repeated across trials and found equivalent levels of fill-in and in-fill errors. In Experiment 2, we employed a reconstruction of order task and found the traditional fill-in effect. These results are theoretically relevant because the fill-in effect has been used to argue against chaining models of memory, and while an in-fill effect was observed in the analyses of Solway et al., our results using shorter lists of words do not show a preponderance of in-fill errors over fill-in errors.

This begs the question as to what class of models the results in these experiments support. While an in-fill effect is predicted by chaining models with asymmetric associations (stronger associations in the forward than in the backward direction), chaining models with symmetric associations (equal strength associations in the forward and backward direction, such as in the TODAM model of Lewandowsky \& Murdock, 1989) would be perfectly compatible with the equal ratios of fill-in and in-fill errors observed in Experiment 1 and a guessing response strategy could be appended to such models to produce fill-in effects as observed in the reconstruction task in Experiment 2. However, one should be reminded of the fact that the protrusion effect applied in both conditions of Experiment 1 and any simple chaining model with either symmetric or asymmetric associations has no recourse for predicting in-position prior list intrusions without positional representations.

\section*{Is there a role of inter-item associations in serial recall?}

We reject the central claim made by Solway et al. (2012) that participants' recall sequences are indicative of a chaining model with asymmetric associations, as we were unable to replicate their finding of an in-fill effect using shorter lists of words. Nonetheless, a number of other studies in the literature have exhibited findings that are in accordance with the predictions of chaining models. During retrieval, when participants are given study list items in the same order as they were presented at study, participants perform better than when they're given unordered list items as cues, suggesting that they're using the cues to retrieve neighboring items from the list (Serra \& Nairne, 2000; Basden, Basden, \& Stephens, 2002). Similarly, when participants are given the same list of words from trial to trial but the starting point of the list differs
(the spin list paradigm: Sequence \(A B C D E\) might be repeated as \(C D E A B\) ), participants are only slightly worse than when the same list is repeated with all of the items in the same positions (Kahana, Mollison, \& Addis, 2010). Positional models, in contrast, would predict a more dramatic impairment to performance in the spin lists than in the repeating lists. While positional models might be extended to produce both of these sets of results, these results follow intuitively from the predictions of chaining models.

While one might be inclined to suggest a hybrid model that incorporates inter-item associations to account for the above findings and positional representations to account for the protrusion effect \({ }^{2}\), such a framework is not only inelegant but is an ad hoc solution to the problem. A more elegant approach is the constraint satisfaction (CS) model proposed by Dennis (2009). In the CS model, asymmetric associations in the forward direction are stored among all of the list items. The model differs from the aforementioned chaining models because at retrieval, the stored representation of the list is compared to all possible ordered list constructions; the distances between the possibilities and the list representation determine the probability of outputting a given sequence.

The basic principle behind the model is that the more differences there are between a candidate sequence and the studied sequence, the less likely it will be that the candidate sequence will be output. Dennis demonstrated that fill-in errors are more frequent than in-fill errors for this reason because a sequence with a fill-in such as \(A C B D E F\) only misses the \(B-C\) connection and erroneously introduces a \(C-B\) connection (two differences from the original sequence), whereas a sequence such as \(A C D E F\) misses all of the connections between \(B\) and its subsequent items (four differences from the original sequence). The model might be able to produce a lower incidence of fill-in errors if it's assumed that list items from the retrieved sequence are only output if they're sufficiently strong, producing more omissions when item memory is poor and lowering the fill-in to in-fill ratio. Dennis also demonstrated that introducing a component of similarity that is common to all of the items allows the model to capture key phenomena that have used to argue for positional representations, such as the protrusion effect as well as the mixed-list phonological similarity effect (e.g.: Baddeley, 1968).

\section*{Conclusion}

Our work evaluating two of the constraints on theories of serial order memory established by Henson (1996) uncovered a generality of one (the protrusion constraint) and a limitation of the other (the fill-in constraint). These results may indicate a need to re-evalute whether inter-item associations are sufficient to support memory for serial order.

\footnotetext{
\({ }^{2}\) The model of Burgess and Hitch (1992) incorporated both interitem and positional associations. However, later versions of the model did not include inter-item associations.
}

\section*{Acknowledgments}

We would like to thank Gina Gerardo, Mengze Zheng, Katherine Bubeleva, and Rebecca Kirchner for their assistance with data collection. We would also like to thank Per Sederberg for helpful discussion about these experiments.

\section*{References}

Baddeley, A. (1968). How does acoustic similarity influence short-term memory. Quarterly Journal of Experimental Psychology, 20.
Basden, B. H., Basden, D. R., \& Stephens, J. P. (2002). Partset cuing of order information in recall tests. Journal of Memory and Language, 47, 517-529.
Botvinick, M. M., \& Plaut, D. C. (2006). Short-term memory for serial order: A recurrent neural network model. Psychological Review, 113, 201-233.
Burgess, N., \& Hitch, G. J. (1992). Towards a network model of the articulatory loop. Journal of Memory and Language, 31(4), 429-460.
Burgess, N., \& Hitch, G. J. (1999). Memory for serial order: A network model of the phonological loop and its timing. Psychological Review, 106.
Burgess, N., \& Hitch, G. J. (2006). A revised model of shortterm memory and long-term learning of verbal sequences. Journal of Memory and Language, 55, 627-652.
Coltheart, V. (1993). Effects of phonological similarity and concurrent irrelevant articulation on short-term memory recall of repeated and novel word lists. Memory \& Cognition, 21(4), 539-545.
Conrad, R. (1960). Serial order intrusions in immediate memory. British Journal of Psychology, 51, 45-48.
Conrad, R. (1963). Acoustic confusion and memory span for words. Nature, 197, 1029-1030.
Dennis, S. (2009). Can a chaining model account for serial recall? In L. Carlson, C. Hölscher, \& T. Shipley (Eds.), Proceedings of the XXXI Annual Conference of the Cognitive Science Society (pp. 2813-2818).
Farrell, S. (2012). Temporal clustering and sequencing in short-term memory and episodic memory. Psychological Review, 119(2), 223-271.
Farrell, S., Hurlstone, M. J., \& Lewandowsky, S. (2013). Sequential dependencies in recall of sequences: Filling in the blanks. Memory \& Cognition.
Geller, A. S., Schleifer, I. K., Sederberg, P. B., Jacobs, J., \& Kahana, M. J. (2007). PyEPL: A crossplatform experiment-programming library. Behavior Research Methods, Instruments \& Computers, 65(1), 50-64.
Healy, A. F. (1974). Separating item from order information in short-term memory. Journal of Verbal Learning and Verbal Behavior, 13, 644-655.
Healy, A. F., Fendrich, D. W., Cunningham, T. F., \& Till, R. E. (1987). Effects of cuing on short-term retention of order information. Journal of Experimental Psychology: Learning, Memory, \& Cognition, 13, 413-425.
Henson, R. N. A. (1996). Short-term memory for serial order.
(Unpublished doctoral dissertation. MRC Applied Psychology Unit, University of Cambridge, Cambridge, England.)
Henson, R. N. A. (1998). Short-term memory for serial order: The start-end model. Cognitive Psychology, 36, 73-137.
Kahana, M. J., Mollison, M. V., \& Addis, K. M. (2010). Positional cues in serial learning: The spin-list technique. Memory \& Cognition, 38(1), 92-101.
Lewandowsky, S., \& Murdock, B. B. (1989). Memory for serial order. Psychological Review, 96, 25-57.
Roodenrys, S., \& Quinlan, P. T. (2000). The effects of stimulus set size and word frequency on verbal serial recall. Memory, 8(2), 71-78.
Serra, M., \& Nairne, J. S. (2000). Part-set cuing of order information: Implications for associative theories of serial order memory. Memory \& Cognition, 28(5), 847-855.
Solway, A., Murdock, B. B., \& Kahana, M. J. (2012). Positional and temporal clustering in serial order memory. Memory \& Cognition, 40, 177-190.
Surprenant, A. M., Kelley, M. R., Farley, L. A., \& Neath, I. (2005). Fill-in and infill errors in order memory. Memory, 13(3), 267-273.

\title{
Modeling Gain-Loss Asymmetries in Risky Choice: The Critical Role of Probability Weighting
}

\author{
Thorsten Pachur (pachur@mpib-berlin.mpg.de) \\ Center for Adaptive Rationality, Max Planck Institute for Human Development, Lentzeallee 94 14195 Berlin, Germany
}

\author{
David Kellen (david.kellen@psychologie.uni-freiburg.de) \\ Department of Psychology, Albert Ludwig University Freiburg, Engelbergerstr. 41 79085 Freiburg im Breisgau, Germany
}

\begin{abstract}
A robust empirical regularity in decision making is that the negative consequences of an option (i.e., losses) often have a stronger impact on people's behavior than the positive consequences (i.e., gains). One common explanation for such a gain-loss asymmetry is loss aversion. To model loss aversion in risky decisions, prospect theory (Kahneman \& Tversky, 1979) assumes a kinked value function (which translates objective consequences into subjective utilities), with a steeper curvature for losses than for gains. We highlight, however, that the prospect theory framework offers many alternative ways to model gain-loss asymmetries (e.g., via the weighting function, which translates objective probabilities into subjective decision weights; or via the choice rule). Our goal is to systematically test these alternative models against each other. In a reanalysis of data by Glöckner and Pachur (2012), we show that people's risky decisions are best accounted for by a version of prospect theory that has a more elevated weighting function for losses than for gains but the same value function for both domains. These results contradict the common assumption that a kinked value function is necessary to model risky choices and point to the neglected role of people's differential probability weighting in the gain and loss domains.
\end{abstract}

Keywords: cognitive modeling; loss aversion; risky choice; prospect theory; probability weighting

\section*{Introduction}

For many of our decisions we are unable to tell with certainty what consequence the decision will have-for instance, when deciding between different medications that potentially lead to some side effects. Ideally, we have knowledge of the nature of the possible consequences as well as some inkling of the chances that the consequences will occur, but our decisions must necessarily remain in the "twilight of probability" (Locke, 1690/2004). Elaborating how such risky decisions are made (and how they should be made) has engaged decision scientists at least since Bernoulli's (1738/1954) seminal work on subjective utility.
One of the most influential and successful modeling frameworks of risky decision making is prospect theory (Kahneman \& Tversky, 1979; Tversky \& Kahneman, 1992). A prominent feature of prospect theory is the assumption that the subjective disutility of a negative outcome is higher than the subjective utility of a positive outcome of the same size. In other words, prospect theory assumes an asymmetry between gains and losses in its value function, which
translates objective outcomes into subjective magnitudes. This assumption of loss aversion can explain, for instance, that people dislike gambles in which one has a \(50 \%\) chance to win a particular amount of money and a \(50 \%\) can to lose the same amount. Similarly, loss aversion is invoked to account for the endowment effect-the phenomenon that people evaluate an object higher in a buyer perspective than in a seller perspective (e.g., Pachur \& Scheibehenne, 2012; for a general overview of gain-loss asymmetries, see Peeters \& Czapinski, 1990).

However, the way prospect theory-more specifically, its mathematical formulation in cumulative prospect theory (CPT; Tversky \& Kahneman, 1992)—is usually implemented allows for asymmetries in the evaluation of positive and negative prospects to be represented also in other ways than via the value function. For instance, the parameters of CPT's weighting function, which translates objective probabilities into subjective decision weights, are typically estimated separately for the gain and the loss domain (e.g., Gonzalez \& Wu, 1999). Furthermore, it has been argued that choice sensitivity (i.e., how accurately choices between two alternatives reflect their subjective valuations) differs between options involving losses and those involving gains only (Yechiam \& Hochman, 2013a).

Crucially, these possible representations of gain-loss asymmetries within CPT have never been directly pitted against each other in a model-comparison analysis (Linhart \& Zucchini, 1986), where the descriptive power of a model is evaluated in light of its complexity (but see Harless \& Camerer, 1994; Stott, 2006). Conducting such a model comparison is our goal in this paper. To that end, we use CPT to model data collected by Glöckner and Pachur (2012), where 64 participants were asked to make choices between 138 two-outcome monetary gamble problems. \({ }^{1}\) Fitting different implementations of CPT to this data also allows us to test specific predictions of how a gain-loss asymmetry should be reflected in specific parameter patterns, such as choice sensitivity (Yechiam \& Hochman, 2013a) or probability sensitivity (Wu \& Markle, 2008). Next we provide a detailed description of CPT's parameter

\footnotetext{
\({ }^{1}\) In Glöckner and Pachur (2012) each participant made choices between 138 gamble problems at two separate sessions (separated by one week). Here we analyze the data from the first session.
}
framework, which we then use to formalize different ways to represent gain-loss asymmetries in risky decision making.

\section*{Cumulative Prospect Theory}

According to CPT, the possible consequences of a risky option are perceived as gains or losses relative to a reference point. The overall subjective value \(V\) of an option with outcomes \(x_{\mathrm{m}}>\ldots \geq x_{1}>0>y_{1}>\ldots>y_{\mathrm{n}}\) and corresponding probabilities \(p_{m} \ldots p_{l}\) and \(q_{1} \ldots q_{n}\) is given by:
\[
\begin{equation*}
V(A)=\sum_{i=1}^{m} v\left(x_{i}\right) \pi_{i}^{+}+\sum_{j=1}^{n} v\left(y_{j}\right) \pi_{j}^{-}, \tag{1}
\end{equation*}
\]
where \(v\) is a value function satisfying \(v(0)=0 ; \pi^{+}\)and \(\pi^{-}\)are decision weights for gains and losses, respectively, which result from a rank-dependent transformation of the outcomes' probabilities. The decision weights are defined as:
\[
\begin{align*}
\pi_{m}^{+} & =w^{+}\left(p_{m}\right) \\
\pi_{n}^{-} & =w^{-}\left(q_{n}\right) \\
\pi_{i}^{+} & =w^{+}\left(p_{i}+\ldots+p_{m}\right)-w^{-}\left(p_{i+1}+\ldots+p_{m}\right) \quad \text { for } 1 \leq i<\mathrm{m} \\
\pi_{j}^{-} & =w^{-}\left(q_{j}+\ldots+q_{n}\right)-w^{+}\left(q_{j+1}+\ldots+q_{n}\right) \quad \text { for } 1 \leq j<\mathrm{n} \tag{2}
\end{align*}
\]
with \(w^{+}\)and \(w^{-}\)being the probability weighting function for gains and losses, respectively (see below). The weight for each positive outcome represents the marginal contribution of the outcome's probability to the total probability of obtaining a better outcome; the weight for each negative outcome represents the marginal contribution of the outcome's probability of obtaining a worse outcome.

Several functional forms of the value and weighting functions have been proposed (see Stott, 2006, for an overview). In our analyses, we use the power value function suggested by Tversky and Kahneman (1992), which is defined as
\[
\begin{gather*}
v(x)=x^{\alpha}  \tag{3}\\
v(y)=-\lambda(-y)^{\beta}
\end{gather*}
\]

For \(\alpha\) and \(\beta\) usually values smaller than 1 are found, yielding a concave value function for gains and a convex value function for losses. The parameter \(\lambda\) reflects the relative sensitivity to losses versus gains and is often found to be larger than 1 , indicating loss aversion.

The weighting function has an inverse \(S\)-shaped curvature, indicating overweighting of unlikely events (i.e., those with a small probability) and underweighting of likely events (i.e., those with a moderate to high probability). Here we use a two-parameter weighting function originally proposed by Goldstein and Einhorn (1987), which separates the curvature of the function from its elevation (cf. Gonzalez \& Wu, 1999):
\[
\begin{array}{ll}
w^{+}(p)=\frac{\delta^{+} p^{\gamma^{+}}}{\delta^{+} p^{\gamma^{+}}+(1-p)^{\gamma^{\gamma}}} & \text { for } x  \tag{4}\\
w^{-}(q)=\frac{\delta^{-} q^{\gamma}}{\delta^{-} q^{\gamma-}+(1-q)^{\gamma}} & \text { for } y
\end{array}
\]
\(\gamma^{+}\)and \(\gamma^{-}\)(both \(<1\) ) govern the curvature of the weighting function in the gain and loss domains, respectively, and
indicate the sensitivity to probabilities. The parameters \(\delta^{+}\) and \(\delta^{-}\)(both \(>0\) ) govern the elevation of the weighting function for gains and losses, respectively, and can be interpreted as the attractiveness of gambling. In other words, \(\delta^{+}\)and \(\delta^{-}\)also indicate a person's risk attitude, with higher (lower) values on \(\delta^{+}\left(\delta^{-}\right)\)for higher risk aversion in gains (losses).

In addition to these core components of CPT, a choice rule is required when applying CPT to model binary choice. To derive the predicted probability of CPT that a gamble A is preferred over a gamble B we used an exponential version of Luce's choice rule:
\[
\begin{equation*}
p(A, B)=\frac{e^{\phi V(A)}}{e^{\phi V(A)}+e^{\phi V(B)}}, \tag{5}
\end{equation*}
\]
where \(\phi\) is a choice sensitivity parameter, indicating how sensitively the predicted choice probability reacts to differences in the valuation of gambles A and B. A higher \(\phi\) indicates more deterministic behavior; with \(\phi=0\), choices are random.

\section*{Modeling Gain-Loss Asymmetries}

As described in the previous section, the common approach to accommodate an asymmetric evaluation of positive and negative prospects is to assume a kinked utility function, for instance produced by \(\lambda>1\) (see also Usher \& McClelland, 2004; Ahn, Busemeyer, Wagenmakers, \& Stout, 2008). Note, however, that observed choices are modeled based on three intertwined components, a value function, a weighting function, and a choice rule, all of which could, in principle, represent an asymmetry between gains and losses. In the following, we describe how gain-loss asymmetries could be modeled within each these components.

\section*{Utility Accounts}

The formalization of CPT's value function allows for two ways to represent a gain-loss asymmetry.
Differential weighting of losses and gains Tversky and Kahneman's (1992) original version of CPT accommodates a gain-loss asymmetry using the loss aversion parameter \(\lambda\), with \(\lambda>1\) leading to a stronger impact of losses (relative to gains). As can be seen from Equation 3, the effect of \(\lambda\) is to multiplicatively magnify the utility of losses relative to the utility of gains, implying greater sensitivity to losses.
Differences in outcome sensitivity In many applications of CPT the exponent of the value function is estimated separately for gains and losses (cf. Fox \& Poldrack, 2008). If the latter (i.e., \(\beta\) in Equation 3) is higher than the former (i.e., \(\alpha\) in Equation 3), this could also lead to a kinked utility function, and thus a gain-loss asymmetry. Note that this pattern has been observed in studies that included pure gain and pure loss gambles (e.g., Abdellaoui, Vossmann, \& Weber, 2005).

\section*{Probability Weighting Accounts}

Equation 1 shows that according to CPT-as in other models in the expectation tradition-the evaluation of an
option closely ties the outcomes to their probabilities, as both are combined multiplicatively. Therefore, an apparent gain-loss asymmetry in the choices might also be represented by assuming differences between gains and losses in probability weighting (Zank, 2010; see also Birnbaum, 2008). Existing studies that have estimated the weighting function separately for gains and losses show that doing so indeed partly absorbs a gain-loss asymmetry (and might decrease the estimated value of \(\lambda\) ). In particular, the elevation is commonly found to be higher for losses than for gains (for an overview, see Fox \& Poldrack, 2008). Nevertheless, it is currently unclear to what extent estimating different weighting functions for losses and gains interacts with the estimation of the \(\lambda\) parameter and whether the increased model flexibility gained by adding more parameters actually leads to better predictive performance.
Differences in probability sensitivity Wu and Markle (2008) highlighted that an asymmetry might not necessarily exist between gains and losses, but between problems with mixed gambles and problems with single-domain gambles (i.e., those that offer either only gains or only losses). They found support for a version of CPT that allows probability sensitivity to differ between mixed and single-domain problems, with a lower probability sensitivity for mixed gambles than for single-domain problems. Moreover, Wu and Markle showed that this version of CPT can account for violations of gain-loss separability (that the evaluation of outcomes and their respective probabilities is done separately for the gain and the loss domains, as shown in Equation 1), which is a fundamental assumption in Tversky and Kahneman's (1992) original description of CPT.

\section*{Choice Sensitivity Account}

A radically different explanation for an asymmetry between the gain and the loss domain was proposed by Yechiam and Hochman (2013a). They argued that the somewhat inconsistent manifestation of loss aversion in risky choice studies might be due to the fact that processing information about potential losses increases the amount of attention allocated to the task at hand. According to Yechiam and Hochman, this should be reflected in a higher choicesensitivity parameter in problems involving losses (i.e., pure-loss gambles and mixed gambles) as compared to problems involving gains only. In a task in which participants responded to sequentially learned risks and using a reinforcement model, Yechiam and Hochman (2013b) found support for this hypothesis; to our knowledge, it has not been tested in the context of description-based tasks and using CPT as modeling framework.

\section*{Which Model Provides The Best Account?}

Several investigations have challenged the utility account of gain-loss asymmetries (e.g., Schmidt \& Traub, 2002; Yechiam \& Hochman, 2013a). However, one problem of these studies is that they focused on specific items and are thus silent with regard to the importance of the elements of
utility accounts (e.g., the loss aversion parameter) for CPT's ability to describe risky choices more generally.

For a more general test, one needs to compare different CPT implementations (representing alternative accounts of gain-loss asymmetries) and to determine which fares best in trading off model fit and model complexity (Myung, 2000).

Such a modeling analysis also allows us to test hypotheses concerning specific parameter patterns predicted by some of these accounts. For instance, according to the choice-sensitivity account by Yechiam and Hochman (2013a) choice sensitivity should be higher in tasks involving losses than in tasks involving gains only. This hypothesis has not been tested directly in the context of description-based tasks.

A second hypothesized parameter pattern follows from the probability weighting account proposed by Wu and Markle (2008), according to which the probability sensitivity (i.e. the curvature of the weighting function) is lower for mixed gambles than for single-domain gambles. Wu and Markle found support for this pattern using Tversky and Kahneman's (1992) one-parameter weighting function; one limitation of this function is, however, that curvature and elevation are confounded. Whether the hypothesized parameter pattern also emerges when using a function that allows to disentangle curvature and elevation (e.g., using the two-parameter weighting function described in Equation 4) has not yet been tested.

\section*{Modeling Approach}

To evaluate the different accounts of gain-loss asymmetries described above, we tested a total of 10 different implementations of CPT in their ability to describe people's risky choices. The implementations, which are summarized in Table 1, differ in terms of whether a gain-loss asymmetry is represented in the value function, the weighting function, or the choice rule.

Table 1: Versions of CPT tested.
\begin{tabular}{|c|c|c|c|c|c|c|c|c|}
\hline &  & \[
\begin{aligned}
& \text { n } \\
& 0 \\
& 0 \\
& 0 \\
& 0 \\
& 0 \\
& 0
\end{aligned}
\] &  & &  & \[
\begin{aligned}
& 0 \\
& 0 \\
& 0 \\
& 0 \\
& 0
\end{aligned}
\] & &  \\
\hline \(\mathrm{CPT}_{\text {nola }}\) & \(\alpha\) & \(\gamma\) & \(\delta\) & & \(\lambda=1\) & \(\phi\) & & 4 \\
\hline \(\mathrm{CPT}_{1}\) & \(\alpha\) & \(\gamma\) & \(\delta\) & & \(\lambda\) & \(\phi\) & & 5 \\
\hline \(\mathrm{CPT}_{\text {ab }}\) & \(\alpha \quad \beta\) & \(\gamma\) & \(\delta\) & & \(\lambda=1\) & \(\phi\) & & 5 \\
\hline \(\mathrm{CPT}_{\mathrm{gd}}\) & \(\alpha\) & \(\gamma^{+} \gamma^{-}\) & \(\delta^{+}\) & \(\delta^{-}\) & \(\lambda\) & \(\phi\) & & 7 \\
\hline \(\mathrm{CPT}_{\text {d }}\) & \(\alpha\) & \(\gamma\) & \(\delta^{+}\) & \(\delta\) & \(\lambda\) & \(\phi\) & & 6 \\
\hline \(\mathrm{CPT}_{\text {dixi }}\) & \(\alpha\) & \(\gamma\) & \(\delta^{+}\) & \(\delta^{-}\) & \(\lambda=1\) & \(\phi\) & & 5 \\
\hline \(\mathrm{CPT}_{\mathrm{g}}\) & \(\alpha\) & \(\gamma^{+} \gamma^{-}\) & \(\delta\) & & \(\lambda\) & \(\phi\) & & 6 \\
\hline \(\mathrm{CPT}_{\text {gfixl }}\) & \(\alpha\) & \(\gamma^{+} \gamma^{-}\) & \(\delta\) & & \(\lambda=1\) & \(\phi\) & & 5 \\
\hline \(\mathrm{CPT}_{\text {phila }}\) & \(\alpha\) & \(\gamma\) & \(\delta\) & & \(\lambda=1\) & & \(\phi^{ \pm /-}\) & 5 \\
\hline \(\mathrm{CPT}_{\mathrm{gsm}}\) & \(\alpha\) & \(\gamma^{+V-} \gamma^{ \pm}\) & \(\delta\) & & \(\lambda\) & \(\phi\) & & 5 \\
\hline
\end{tabular}
\(\mathrm{CPT}_{1}\) can be considered as the standard implementation of CPT. It assumes the same exponent in the value function
(Equation 3) for gains and losses (i.e., \(\alpha=\beta\) ), but allows for a gain-loss asymmetry by having \(\lambda\) as a free parameter. \(\mathrm{CPT}_{1}\) uses one common set of curvature ( \(\gamma\) ) and ( \(\delta\) ) elevation parameters across the gain and loss domains. In a restricted version of \(\mathrm{CPT}_{1}, \mathrm{CPT}_{\text {nola }}, \lambda\) is set to 1 and thus assumes no gain-loss asymmetry. \(\mathrm{CPT}_{\text {nola }}\) will serve as a benchmark model. \(\mathrm{CPT}_{\mathrm{ab}}\) also sets \(\lambda\) to 1 but allows the exponents of the value function to differ between gains ( \(\alpha\) ) and losses ( \(\beta\) ).
\(\mathrm{CPT}_{\mathrm{gd}}\) is the model with the largest number of free parameters: it allows for differences between gains and losses both in curvature ( \(\gamma^{+}\)and \(\gamma^{-}\)) and elevation ( \(\delta^{+}\)and \(\delta^{-}\)) of the weighting function. The restricted versions \(\mathrm{CPT}_{\mathrm{d}}\) and \(\mathrm{CPT}_{\text {dixi }}\) allow a gain-loss asymmetry in the elevation only, thus assuming a single curvature parameter for gains and losses \(\left(\gamma^{+}=\gamma^{-}\right)\); in \(\mathrm{CPT}_{\text {dixi }} \lambda\) is set to \(1 . \mathrm{CPT}_{\mathrm{g}}\) and \(\mathrm{CPT}_{\text {gfixl }}\) allow a gain-loss asymmetry in the curvature of the weighting function only, thus assuming a single elevation parameter \(\left(\delta^{+}=\delta\right)\); for \(\mathrm{CPT}_{\text {gixix }} \lambda\) is set to 1 .
The two remaining models, \(\mathrm{CPT}_{\text {phila }}\) and \(\mathrm{CPT}_{\text {gsm }}\) implement the proposals by Yechiam and Hochman (2013a) and Wu and Markle (2008), respectively. \(\mathrm{CPT}_{\text {phila }}\) assumes a gain-loss asymmetry neither in the value function (i.e., \(\lambda=\) 1) nor in the weighting function ( \(\gamma^{+}=\gamma^{-}\)and \(\delta^{+}=\delta^{-}\)); instead, it allows for separate choice sensitivity parameters in gambles involving losses ( \(\phi^{ \pm-}\)) and gambles involving gains only \(\left(\phi^{+}\right)\). \(\mathrm{CPT}_{\text {gsm }}\) assumes no gain-loss asymmetry in the value function, but allows for different curvatures of the weighting function for single-domain gambles ( \(\gamma^{+\vee-}\) ) than for mixed gambles ( \(\gamma^{ \pm}\)).

The models were fitted to individual participants using the maximum-likelihood method. In order to avoid local minima, the optimization algorithm was supplemented with an initial grid search (considering up to 80,000 value combinations of the entire parameter space, with all parameters partitioned similarly).

To evaluate the models, we relied on the Bayesian Information Criterion (BIC), which penalizes a model as a function of its number of free parameters (Schwarz, 1978). The BIC of a given model is given by:
\[
\begin{equation*}
B I C=-2 \log (f(d \mid \hat{\theta}))+\log (N) k \tag{6}
\end{equation*}
\]
with \(d\) denoting the data, \(N\) the number of data points (i.e., the number of gamble problems), and \(k\) the number of free parameters in the model. BIC is an approximation of the Bayes Factor (Kass \& Raftery, 1995), providing a theoretically-principled framework for model comparison that takes into account goodness of fit as well as model complexity (e.g., Myung, 2000). A lower BIC indicates a better model fit.
Data We applied the different CPT implementations to model individual data in Glöckner and Pachur (2012). \({ }^{2}\) In

\footnotetext{
\({ }^{2}\) Note that while Glöckner and Pachur also compared some implementations of CPT, they neither tested Wu and Markle's (2008) sensitivity account, nor Yechiam and Hochman's (2013a) choice sensitivity account. More importantly, they also neither considered implementations of CPT in which only some of the
}
this study, 63 participants ( 25 male, mean age 24.7 years) indicated their preferences between 138 two-outcome monetary gamble problems that contained 70 pure gain, 30 pure loss, and 38 mixed gambles, all drawn from sets of gamble problems used in previously published studies (see Glöckner \& Pachur for details). The outcomes of the gambles ranged from \(-1000 €\) to \(1200 €\). At the completion of each session, one of the chosen gambles was picked randomly, played out and the participant received an additional payment proportional to the resulting outcome.

\section*{Results}

\section*{Is there a Gain-Loss Asymmetry in the Value Function?}

For the standard version of \(\mathrm{CPT}, \mathrm{CPT}_{1}\), which allows for a gain-loss asymmetry only through the loss aversion parameter, the median (across participants) best-fitting value of the \(\lambda\) parameter was substantially larger than \(1, \lambda=1.40\). \(71.9 \%\) of the participants had a \(\lambda\) larger than 1 . Moreover, \(\mathrm{CPT}_{1}\) showed a considerably better fit than \(\mathrm{CPT}_{\text {nola }}\), which does not allow for an asymmetry between gains and losses (median BIC: 158.34 vs. 160.23). These results thus provide evidence for a gain-loss asymmetry in the data.
Is choice sensitivity higher in gambles involving losses? As previously stated, Yechiam and Hochman (2013a) argue that due to differences in attention, choice sensitivity should be higher when the gambles include a potential loss. We tested this prediction by modeling the data with \(\mathrm{CPT}_{\text {phila }}\), which allows for a gain-loss asymmetry in choice sensitivity only. As it turned out, there was no evidence for Yechiam and Hochman's hypothesis; in fact, we find the opposite pattern, with a higher choice-sensitivity parameter for gains than for losses, median values \(\phi^{+}=0.18\) and \(\phi^{ \pm-}=0.09\), Wilcoxon test: \(W=2,609, p=.0008\) (two-tailed). This pattern of results was found for 58 of the 64 participants (91\%).
Is probability sensitivity lower in mixed gambles? Consistent with Wu and Markle's (2008) hypothesis, the estimates for \(\gamma\) obtained with \(\mathrm{CPT}_{\text {gsm }}\) indicated a lower probability sensitivity for mixed gambles than for singledomain gambles, median values \(\gamma^{+\mathrm{V-}}=0.58\) and \(\gamma^{ \pm}=0.86\), Wilcoxon test: \(W=2,928, p=.0001\) (two-tailed). Fortyeight out of 64 participants ( \(75 \%\) ) showed this pattern.

To summarize, these analyses indicate that people's choices reflect an asymmetry between gains and losses. Of two proposals concerning the specific nature of such asymmetries, we found support for only one, namely Wu and Markle's (2008) hypothesis that probability sensitivity is reduced in mixed as compared to single-domain gamble problems. Yechiam and Hochman's (2013a) proposal of a higher choice sensitivity for gambles involving losses was not supported (in fact, we found the opposite pattern). Next, we turn to the question of how well the different
parameters of the weighting function were estimated separately for gains and losses, nor implementations with a fixed \(\lambda\) parameter.
implementations of CPT summarized in Table 1 can account for people's choices. For instance, even if there is support for Wu and Markle's (2008) hypothesis of a lower probability sensitivity in mixed (as compared to singledomain) gambles, does an implementation of CPT allowing for this pattern (i.e., \(\mathrm{CPT}_{\mathrm{gsm}}\) ) also perform well in terms of BIC?

\section*{Model Comparison}

Figure 1 shows the median (across participants) BICs for each of the CPT implementations. As can be seen, the bestperforming model is \(\mathrm{CPT}_{\text {dfixl }}\), which allows for gain-loss asymmetries in the elevation of the probability weighting function but sets \(\lambda=1\). Figure 2 shows the probability weighting function of \(\mathrm{CPT}_{\text {dfixl }}\), based on the median bestfitting parameter values. The figure shows that this model represents a gain-loss asymmetry by having a more elevated weighting function for losses than for gains, \(\delta^{-}=1.69, \delta^{+}=\) 0.63 . Like models implementing the utility account, \(\mathrm{CPT}_{\text {dfixl }}\) gives more weight to losses than to gains, but does this via the decision weights resulting from the weighting function rather than via the value function. \(\mathrm{CPT}_{\text {dfixl }}\) not only achieved the best performance in terms of the median BIC, but also the overwhelming majority of individual participants ( \(54.7 \%\) ) were best accounted for by this model. \({ }^{3}\)


Figure 1: Performance of the different versions of CPT, as indicated by the median BIC (across participants)

\section*{Discussion}

One of the fundamental assumptions in prospect theory is that negative prospects receive more weight in people's evaluations of risky alternatives than positive prospects. In general, our analyses provide support for this assumption by

\footnotetext{
\({ }^{3}\) The second-best model in terms of selection frequency was \(\mathrm{CPT}_{\text {nola }}\), which best accounted for \(17.2 \%\) of the individual participants. Interestingly, the second-best performing model in terms of the median BIC, \(\mathrm{CPT}_{\mathrm{d}}\), best accounted for only \(4.7 \%\) of the participants.
}
finding evidence for a gain-loss asymmetry. However, we pointed out that the parametric menagerie of CPT can, in principle, represent gain-loss asymmetries in many different ways, such as via outcome sensitivity, probability sensitivity, the elevation of the weighting function, and choice sensitivity. Crucially, our analyses showed that a model that assumes a gain-loss difference in the elevation of the weighting function and a symmetric value function provided the best account of people's choices. The common assumption of a kinked utility function thus does not seem to be necessary. Other proposed implementations of CPT, such as one that attributes gain-loss asymmetries to differences in the choice rule (Yechiam \& Hochman, 2013a) or one that replaces the assumption of strict gain-loss separability by allowing for differences in probability sensitivity between single-domain and mixed gambles ( Wu \& Markle, 2008), were also clearly outperformed.


Figure 2: Shapes of the separate probability weighting functions for gains and losses of \(\mathrm{CPT}_{\text {dfixl }}\), the best performing model, when using the median best-fitting parameter estimates. The weighting functions differ only in terms of their elevation, which is higher for losses than for gains.

Our results seem to challenge the approach taken in previous tests of prospect theory that have focused on specific and individual gamble problems. For instance, using problems specifically designed to test gain-loss separability, Wu and Markle (2008) found evidence for a superiority of a version of CPT that allowed for different probability sensitivity in single-domain than in mixed gamble problems. By contrast, in the data set used here, where the gamble problems were not constructed to test specific assumptions of CPT (instead many of the gambles had been randomly generated; see Glöckner \& Pachur, 2012, for details), Wu and Markle's modified version of CPT performed rather poorly (Figure 1). The results of our model comparison thus suggest that model developments based on focused tests may sometimes sacrifice a model's
ability to account for choices more generally for its ability to account for idiosyncratic cases.

What does the superiority of the version of CPT with a more elevated probability weighting for losses than for gains (Figure 2) mean psychologically? The cognitive underpinnings of probability weighting are still rather little understood. This has led some researchers (e.g., Brooks \& Zank, 2005; Zank, 2010) to focus more on what can be called "behavioral gain-loss asymmetries", that is, specific choice patterns that follow from gain-loss asymmetries on the value and/or probability weighting functions.
These open questions notwithstanding, our results suggest that if one's goal is to predict how people will decide between risky alternatives, modeling gain-loss asymmetries in terms of differences in probability weighting rather than utility weighting promises to be a more successful approach. Our conclusions thus resonate well with Prelec's (2000) assessment that "probability nonlinearity will eventually be recognized as a more important determinant of risk attitudes than money nonlinearity." (p. 89)

\section*{Acknowledgments}

This research was supported by a grant of the German Research Foundation as part of the priority program "New Frameworks of Rationality" (SPP 1516) to Thorsten Pachur (HE 2768/7-1).

\section*{References}

Abdellaoui, M., Vossmann, F., \& Weber, M. (2005). Choice-based elicitation and decomposition of decision weights for gains and losses under uncertainty. Management Science, 51, 1384-1399.
Ahn, W.-Y., Busemeyer, J. R., Wagenmakers, E.-J., \& Stout, J. C. (2008). Comparison of decision learning models using the generalization criterion method. Cognitive Science, 32, 1376-1402.
Bernoulli, D. (1738/1954). Exposition of a new theory on the measurement of risk. Econometrica, 22, 23-36.
Birnbaum, M. H. (2008). New paradoxes of risky decision making. Psychological Review, 115, 463-501.
Brooks, P., \& Zank, H. (2005). Loss averse behavior. Journal of Risk and Uncertainty, 31, 301-325.
Fox, C. R., \& Poldrack, R. A. (2008). Prospect theory and the brain. In Glimcher, P., Camerer, C., Fehr, E. \& Poldrack, R. (Eds), Handbook of neuroeconomics. New York: Elsevier.
Glöckner, A., \& Pachur, T. (2012). Cognitive models of risky choice: Parameter stability and predictive accuracy of prospect theory. Cognition, 123, 21-32.
Goldstein, W. M., \& Einhorn, H. J. (1987). Expression theory and the preference reversal phenomenon. Psychological Review, 94, 236-254.
Gonzalez, R., \& Wu, G. (1999). On the shape of the probability weighting function. Cognitive Psychology, 38, 129-166.

Harless, D. W., \& Camerer, C. F. (1994). The predictive utility of generalized expected utility theories. Econometrica, 62, 1251-1289.
Kass, R., \& Rafterty, A. (1995). Bayes factors. Journal of the American Statistical Association, 90, 773-795.
Kahneman, D., \& Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, 47, 263292.

Linhart, H., \& Zucchini, W. (1986). Model selection. Oxford, England: John Wiley \& Sons.
Locke, J. (1690/2004). An essay concerning human understanding. London: Penguin.
Luce, R. D. (1959). Individual choice behavior: A theoretical analysis. New York: Wiley.
Myung, I. J. (2000). The importance of complexity in model selection. Journal of Mathematical Psychology, 44, 190204.

Pachur, T., \& Scheibehenne, B. (2012). Constructing preference from experience: The endowment effect reflected in external information search. Journal of Experimental Psychology: Learning, Memory and Cognition, 38, 1108-1116.
Peeters, G., \& Czapinski, J. (1990). Positive-negative asymmetry in evaluations: The distinction between affective and informational negativity effects. In W. Stroebe, \& M. Hewstone (Eds.), European review of social psychology (Vol. 1, pp. 33-60). Chichester, UK: Wiley.
Prelec, D. (2000). Compound invariant weighting functions in prospect theory. In D. Kahneman, \& Tversky, A. (Eds.) Choices, values, and frames (pp. 67-92). Cambridge, UK: Cambridge University Press.
Schmidt, U., S. Traub. 2002. An experimental test of loss aversion. Journal of Risk and Uncertainty, 25, 233-249.
Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6, 461-464.
Stott, H. P. (2006). Cumulative prospect theory's functional menagerie. Journal of Risk and Uncertainty, 32, 101-130.
Tversky, A., \& Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and Uncertainty, 5, 297-323.
Usher, M., \& McClelland, J. L. (2004). Loss aversion and inhibition in dynamical models of multialternative choice. Psychological Review, 111, 759-769.
Wu, G., \& Markle, A. (2008). An empirical test of gain-loss separability in prospect theory. Management Science, 54, 1322-1335.
Yechiam, E., \& Hochman, G. (2013a). Losses as modulators of attention: Review and analysis of the unique effects of losses over gains. Psychological Bulletin. 139, 497-518.
Yechiam, E., \& Hochman, G. (2013b). Loss-aversion or loss-attention: The impact of losses on cognitive performance. Cognitive Psychology, 66, 212-231.
Zank, H. (2010). On probabilities and loss aversion. Theory and Decision, 68, 243-261.

\title{
Semantic Implicit Learning in Language
}

\author{
Albertyna Paciorek (awp23@cam.ac.uk) \\ John N. Williams (jnw12@cam.ac.uk) \\ Department of Theoretical and Applied Linguistics, University of Cambridge \\ 9 West Road, Cambridge CB3 9DP, United Kingdom
}

\begin{abstract}
Previous studies of semantic implicit learning in language have only examined learning grammatical form-meaning connections where learning could have been supported by prior linguistic knowledge. Also, these studies assessed awareness by verbal report, which is arguably not the most reliable measure. Here we target the domain of verb meaning, specifically semantic preferences of novel verbs (e.g. a novel verb takes abstract objects). Using a reaction time methodology we show that after exposure to correct verb-noun combinations, reaction times to incorrect combinations are slowed down even for participants who are unaware of the semantic regularity. This effect was also obtained even when the semantic regularity was irrelevant to the tasks being performed, suggesting that the semantic generalisation is learned and exerts its influence automatically, hence satisfying one criterion for implicitness. Combined with a lack of verbalisable knowledge in any participant these experiments provide strong evidence for semantic implicit learning in language.
\end{abstract}

Keywords: implicit learning; consciousness; formmeaning connections; vocabulary learning; verb learning, second language acquisition; automaticity.

\section*{Introduction}

Most research on implicit learning has examined regularities at the level of form, be they in sequences of letters generated by artificial grammars, screen positions in repeating sequences, and in the domain of language, phonological (Dell et al., 2001) and orthographic (Pacton et al., 2001) patterns. This limits generalizability to other aspects of language learning where regularities might be conditioned by distinctions at the level of meaning, as opposed to form. Some research in visual perception has exposed semantic-based implicit learning, notably using the contextual cuing paradigm, where target locations are predicted from semantic properties of contexts, (Goujon, 2007). But can semantic implicit learning effects be obtained in the domain of language, especially in the adult language learner? Given arguments that even in children vocabulary acquisition requires declarative, explicit, memory
(Ellis, 1994) and shared attention (Bloom, 2000) one might suspect not. However, these arguments relate to learning referential meaning. Others have hypothesised that other aspects of word meaning, such as connotation and collocational behaviour, might be learned implicitly by the non-declarative system (Paradis, 2004). Here we test this proposal in the context of semantic preferences of verbs.

Previous research on semantic implicit learning in language has focused exclusively on article-noun agreement regularities (e.g. Williams, 2005; Leung \& Williams, 2012). This work has demonstrated sensitivity to the semantic properties of nouns in learning about the distribution of articles in miniature semi-artificial languages, and has provided evidence of implicitness of knowledge through postexperiment verbal report. The present experiments extend this work in terms of generalizability to other aspects of language, and in terms of methodology. We will discuss methodology in Experiment 2, but with regard to generalizability, there was evidence in these earlier experiments that implicit learning was dependent upon prior knowledge of article agreement systems in other languages (Williams, 2005). There was also evidence that learning depended on the semantic regularity in question, since effects were obtained for agreement based on animacy, but not on relative size (Leung \& Williams, 2012). Putting these two together one might argue that learning was dependent both on prior knowledge of the potential for article noun agreement, and on dispositions based on the "potentially encodable distinctions" that can be grammaticised in language (Bickerton, 1999). The question arises, therefore, whether similar effects could be obtained for an aspect of language that falls outside the realm of grammar, and is not so potentially affected by prior dispositions. This motivated the current investigation of learning the collocational behaviour, specifically semantic preferences, of novel verbs.

A semantic preference can be understood as a particular type of collocation, where collocation refers to higher than chance co-occurrence of two or more words. Collocates sound natural together and substituting one of them with a near-synonym results
in a loss of naturalness for native speakers. For example in English it is better to say fast car and fast food, rather than *quick car or *quick food. Conversely, it is more natural to say quick glance and quick meal instead of *fast glance or *fast meal. It has been traditionally proposed that collocations reflect syntagmatic relations between words, and are related to surface form, rather than paradigmatic relations regarding meaning (Firth, 1957). However, syntagmatic regularities may not be the optimal, or the sole way of accounting for the existence and acquisition of semantic preferences. After all, new collocates can be freely generated, as long as they follow implicit assumptions regarding applicable semantic sets. For example, knowing that pack collocates with dog, hounds, wolves etc. while swarm with bees, mosquitoes, bats, naturally suggests other animals which would be appropriate in either set. It makes sense therefore, to predict that the existence of such semantically preferred sets of collocates involves generalisations at a level higher than form. The question is, can such semantic generalisations be learned implicitly?

The present experiments used four novel verbs, powter, mouten, gouble, and conell. The participants were exposed to these verbs in verb phrases containing a direct object noun. Their task required them to think about whether the verb conveyed an 'increase' or 'decrease' meaning, either as inferred from the context (Experiment 1) or as they had learned before the experiment (Experiment 2). What they were not told was that powter and gouble took abstract collocates, whereas mouten and conell took concrete collocates (Table 1). For example, correct verb phrases would be powter the significance, gouble the power, mouten the calcium, conell the chocolate. We tested whether participants would learn the semantic preferences of these novel verbs implicitly using two techniques. Experiment 1 embedded the verb phrases in sentence contexts, and required participants to make an explicit concreteness decision on the nouns, and indicate whether the verb meant increase or decrease at the end of the sentence. The prediction was that after exposure to many correctly formed trials concreteness decision times and/or increase/decrease response times would be faster to new verb-noun combinations that respect the rule than to combinations that do not (e.g. powter with a concrete collocate such as compost). These will be referred to as the New Grammatical (NG) and New Ungrammatical (NU) conditions respectively. If learning is implicit then this effect would be obtained even for participants who evinced no awareness of the relevance of concreteness to the collocational behavior of the verbs, as assessed by a post-test.

In Experiment 2 participants saw only the verb phrases, but this time they had to decide whether the noun conveyed positive or negative connotations (a decision that is subjective and irrelevant to the hidden regularity). This was followed by the increase/decrease decision, as before. Any effect here would arguably provide stronger evidence for implicit learning, and would speak to the automaticity of the semantic activation underlying the effect.

Table 1. The novel verbs used in the experiments
\begin{tabular}{|c|c|c|c|}
\hline \multicolumn{2}{|c|}{} & \multicolumn{2}{c|}{ Participants not told } \\
\cline { 3 - 4 } \multicolumn{2}{|c|}{} & \begin{tabular}{c} 
Abstract \\
collocate
\end{tabular} & \begin{tabular}{c} 
Concrete \\
collocate
\end{tabular} \\
\hline \multirow{3}{|c|}{\begin{tabular}{c} 
Participants told \\
to infer from \\
context
\end{tabular}} & increase & powter & mouten \\
\cline { 3 - 4 } & decrease & gouble & conell \\
\hline
\end{tabular}

\section*{Experiment 1}

\section*{Participants}

40 students of the University of Cambridge participated in the experiment. 17 were native speakers of English. All of the nonnative participants had achieved at least IELTS 7.5.

\section*{Materials and Procedure}

A total of 80 sentences were created, 20 for each novel verb, in which the verb conveyed either an 'increase' or 'decrease' meaning with respect to the object. For procedural reasons the word order was scrambled so that the verb phrase occurred at the beginning of the sentence. Examples are shown in Table 2.

Table 2. Example sentences from Experiment 1
\begin{tabular}{|l|}
\hline \begin{tabular}{l} 
POWTER the prestige of wealthy families, artists \\
can.
\end{tabular} \\
\hline \begin{tabular}{l} 
GOUBLE the role of nuclear weapons, Obama \\
stresses the need to.
\end{tabular} \\
\hline MOUTEN the nutrients you need, make sure you. \\
\hline CONELL the histamine stores, the sweating helps to. \\
\hline
\end{tabular}

The experiment comprised two blocks of trials, although the participants were not aware of any division between them. In the first block there were 44 training trials in which each novel verb occurred equally often. The collocates occurred with both increase and decrease verbs (e.g. POWTER the
prestige and GOUBLE the prestige occurred, but in different sentence contexts). Block 2 contained 32 critical trials in which the novel verbs occurred in new sentences and with new collocates not encountered in Block 1. Half of these items respected the semantic preference rule ("new grammatical", NG, condition), and half violated it, for example by pairing a concrete noun with POWTER ("new ungrammatical", NU, condition). The new collocates in the critical test items were chosen so as to be roughly synonymous with an object noun that occurred in the training phase (e.g. POWTER the importance occurred in training, and POWTER the significance occurred in test). Each object noun appeared only once in the critical trials. Assignment of items to conditions was rotated around two presentation lists so as to control for item effects. Block 2 also contained an additional 44 grammatical sentences so as to provide more reinforcement for the rule. These were sentences repeated from Block 1. The order of trials in Blocks 1 and 2 was independently randomized for each subject.

The sentences were presented word by word in the centre of the screen. First the noun, e.g. POWTER, was presented in capital letters for 600 msecs followed by the article the for 600 msecs . This was followed by a noun, e.g. prestige, in red lower case letters. The participants were instructed to indicate as quickly as possible whether this noun referred to an abstract or a concrete object by pressing the left or right buttons on a response box. If they made an error the word remained on the screen until they pressed the correct button. Upon a correct response the display changed to a recall prompt '__ the ___' and they had to recall the noun phrase aloud, i.e. say "powter the prestige". They then pressed a response button and the remainder of the sentence appeared at a rate of 600 msecs per word, e.g. of wealthy families artists can. At the end of the sentence the prompt +/appeared on the screen and the participants had to indicate as quickly as possible whether they thought the verb conveyed a broadly increase or decrease meaning by pressing the right or left buttons on the response box.

In order to assess awareness, at the end of the experiment the participants were presented with 8 new sentences from which the verb had been removed. They were asked to indicate which of the four novel verbs they thought should be used in that context and to think aloud as they made their decision. Any participant who referred to the abstract/concrete distinction, or similar, was classified as 'aware', regardless of their actual performance on the post-test.

The experiment was run using Superlab software and a Cedrus response box.

\section*{Results and Discussion}

Out of 40 participants, 13 revealed at least fragmentary explicit knowledge in the postexperiment debriefing, and were classified as 'aware'. The remaining 27 participants were classified as 'unaware'. For each participant, response times that were more than 2.5 standard deviations above the mean response time over the 32 critical trials were winsorized (i.e. replaced with the next highest value). Additionally, in cases where an error was made on the first (abstract/concrete) decision, the response on the second decision was removed from the analysis. This was because participants were likely to have been distracted on the subsequent increase/decrease decision by just having had to correct themselves.

An initial analysis of the data revealed that for both the aware and unaware groups there were no differences in either reaction time or error rate between the NG and NU conditions on the first, concrete/abstract, decision. This was rather surprising because we expected that if the regularity had been learned the verb would set up an expectation of a certain type of collocate. However, a significant NUNG reaction time difference was obtained on the second, increase/decrease, decision. After excluding two unaware participants with excessively long response times (of 3231 msecs and 3359 msecs ) the remaining 25 unaware participants had a mean response time of 1101 msecs in the NG condition and 1332 msecs in the NU condition, \(\mathrm{F}(1,23)=5.11, \mathrm{p}<\) \(0.05, \eta 2=0.20\). There was no difference in error rates ( \(25.2 \%\) and \(23.4 \%\) respectively). As for the aware participants, despite showing a large numerical difference between the NG and NU conditions on the increase/decrease decision (1478 msecs and 1611 msecs respectively) this difference was not significant, \(\mathrm{F}(1,11)=1.70, \mathrm{p}=0.23\). Neither was there an effect in errors ( \(22.5 \%\) and \(23.9 \%\) respectively).

A post hoc analysis was carried out to check for potential differences in performance between the native and non-native English-speaking participants. An ANOVA revealed no interaction between grammaticality and native/non-native speaker status, F \(<1.0\).

This experiment provides evidence of implicit learning of a semantic preference rule. The fact that learning effects are apparent on the decisions involving purely the indication of whether a
particular verb meant to 'increase' or 'decrease' is particularly compelling, since the decision was being made with reference to the meaning of the verb, not the collocate, and knowledge of the semantic preference rule does not directly inform this decision. This effect actually provides stronger evidence of the use of implicit knowledge than if it had occurred on the concreteness decision (a point which we will elaborate below).

Having said this, it is not clear how the effect of grammaticality on the increase/decrease decision arises. One possibility is that the mismatch between the verb and noun in the NU condition somehow disturbed the process of deriving the increase/decrease meaning from the verb and its context. It may also have caused confusion about the identity of the verb (since the collocate would have suggested other verb possibilities than the one that occurred). However, there is an alternative explanation that cannot be ruled out at this stage. The effect may not reflect learning of a semantic regularity at all, but rather associations between novel verbs and patterns of button presses (e.g. POWTER was associated with successive responses on the right-hand button). Experiment 2 was designed to rule out this possibility, as well as creating conditions under which awareness of the hidden regularity was much less likely to occur, and under which any effect of knowledge on behavior was more likely to reflect automatic, as opposed to controlled, behavior.

Even if we suppose that the effect obtained in Experiment 1 was semantic in origin, the question remains as to the nature of the generalization that was formed. Although the noun collocates in the critical test items were different from those that occurred in training they were in fact roughly synonymous with a noun that had occurred in the training phase. This means that it is hard to defend the claim that what was learned was a correlation between verbs and the abstract/concrete distinction as such. Rather the effect could have reflected the similarity between individual nouns in training and test. In order to address this issue, the noun collocates in Experiment 2 were changed so as to represent a more heterogeneous set of abstract and concrete nouns, and no noun in the test phase was a synonym of a noun in the training phase. Learning over these items would be more likely to reflect abstraction of a broad concreteness distinction.

\section*{Experiment 2}

This experiment employed a reaction time methodology similar to Experiment 1. Two main changes were made. First, a simplified procedure was employed in which only verb phrases were presented. Participants were informed about the increase/decrease meanings of the verbs prior to the experiment. The second change was that instead of making a concreteness decision on the collocates participants now had to indicate whether the collocate had positive, negative, or neutral connotation. For example, chocolate and holidays would be expected to receive positive judgments, whereas horror would be expected to receive a negative judgment. Participants were informed that the choices were subjective and that there was no correct answer. Crucially, the semantic preference rule was exactly the same as in Experiment 1; powter and gouble went with abstract nouns and conell and mouten with concrete nouns. Given that no systematic alignment between connotative meaning judgments and concreteness is expected this means that there will be no systematic relationship between the button pressed on the first decision and the second increase/decrease decision. Thus, learning is unlikely to be based on associations between nonsense verbs and patterns of button presses. This also means that any influence of noun concreteness on the second decision must reflect automatic activation of this aspect of meaning, rather than explicit retrieval as part of the task (as was the case in Experiment 1).

\section*{Participants}

46 students of the University of Cambridge participated in the experiment. Three participants were excluded due to problems with the task. Of the remaining group of 43 participants, 22 were native speakers of English. All of the nonnative participants had achieved at least IELTS 7.5.

\section*{Materials and Procedure}

The experimental design was identical to Experiment 1. The nouns for the training and test items were changed so as to comprise a more heterogeneous set of abstract and concrete nouns. The broadened category of abstract nouns included ones as different as happiness, wisdom, impact, understanding. The category of concrete nouns was similarly broadened to include, for example, chocolate, luggage, metal and paper. Only verb phrases were presented.

Participants were first informed about the increase/decrease meanings of the novel verbs. The
simplified procedure in Experiment 2 comprised the following sequence of events. First the verb was presented in capital letters for 600 msecs . This was followed by the word the for 600 msecs , followed by the noun in red. The participants made their decision about the connotation of the noun within an allocated time of 3 secs. Responses were entered on a millisecond accurate keyboard where \(M\) indicated 'positive', Z 'negative', and space bar 'neutral'. Upon making any response the noun was replaced by the 'inc/dec' prompt and the participant indicated whether the verb meant increase or decrease by pressing M and Z respectively. After every two stimuli participants received prompts that required them to recall out loud one of the phrases they had just seen. The prompt revealed either the first part of the phrase, for example "MOUTEN the \(\qquad\) ", or the second part: " the prestige", and participants were asked to pronounce the complete phrase. The memory task was to encourage full attention to the materials and the data were not analysed. All stimuli were presented in the centre of the screen. The experiment was followed by a posttest similar to that in Experiment 1.

\section*{Results and Discussion}

None of the participants appeared to have any awareness of the correlation between the novel verbs and the concreteness of the noun (whereas in Experiment \(132.5 \%\) of the participants were classed as aware). The data were treated in the same way as in Experiment 1. Two of the native participants were excluded on the basis of excessively long second decision response times. As before there were no reaction time effects on the first decision. This time there were no effects on the second, increase/decrease, decision either. Response times in the NG and NU conditions were 616 msecs and 627 msecs respectively, and error rates were \(6.1 \%\) and 6.7\% (note that reaction times were much faster than in Experiment 1, presumably because the decision was made immediately after the noun, rather than being delayed until the end of a sentence). When the data for the native English-speaking and non-native groups were compared an interesting pattern emerged. For the natives, reaction times in the NG and NU conditions were 556 msecs and 612 msecs , with error rates of \(7.2 \%\) and \(5.9 \%\). In contrast for the nonnatives the reaction times in the NG and NU conditions were 690 msecs and 655 msecs , with error rates of \(5.1 \%\) and \(7.4 \%\). An ANOVA was performed on the reaction time data with Group (native or nonnative) and Presentation List as between-subjects factors, and Condition (NG vs NU) as a withinsubjects factor. The interaction between Group and

Condition was significant, \(\mathrm{F}(1,37)=8.74, \mathrm{p}<0.01\), \(\eta 2=0.19\). Follow-up ANOVAs showed that the learning effect was only significant for the native speaker group \(\mathrm{F}(1,17)=9.13, \mathrm{p}<0.01, \eta 2=0.34\). There were no significant effects on errors.

Experiment 2 provides stronger evidence for semantic implicit learning than Experiment 1 because the learned generalisation was unrelated to the tasks being performed. Knowledge of the correlation between verbs and noun concreteness was unrelated both to the decision being made on the noun (connotative meaning) and to the increase/decrease decision. It has been argued that implicit knowledge should exert its influence on behaviour in an automatic rather than controlled way (Cleeremans \& Jimenez, 2002), and that the strongest test of implicit knowledge is to be obtained in situations where knowledge has an effect on performance even though it is irrelevant to the task at hand (Tzelgov, 1997; Vinter \& Perruchet, 1999). The present experiment seems to satisfy those criteria. Furthermore, on this occasion none of the participants demonstrated awareness of the semantic regularity in the post-test, which in itself suggests that the relevant knowledge was well below the level of awareness.

Experiment 2 also shows that the effects obtained in Experiment 1 were not due to learning associations between the novel verbs and patterns of keystrokes. This was because the nouns in the abstract and concrete categories would have elicited a range of 'positive' (M), 'negative' (Z) and 'neutral' (space bar) responses. Thus the effects must have had a semantic origin. Furthermore the learning effect in natives was obtained over sets of nouns that were more heterogeneous than in Experiment 1, and the critical test nouns were not synonyms of any nouns in training. Therefore the learning effect must have been supported by a broad generalisation, which we assume is essentially based on the abstract versus concrete distinction.

The fact that in this experiment learning was not obtained for non-native speakers is perhaps not surprising. There is a wealth of evidence that the mapping from second language words to meaning is less automatic than from first language words. For example, automatic semantic priming from second language words can only be obtained at very high levels of proficiency (Basnight-Brown \& Altarriba, 2007). Thus, in a situation in which a decision is made about one aspect of meaning it is not surprising that the non-natives in Experiment 2 did not activate other aspects of meaning with sufficient strength to produce a learning effect.

\section*{Conclusion}

The present experiments demonstrate that semantic implicit learning of linguistic regularities can be obtained outside of the realm of grammar, and can extend into learning about verb meaning. However, this is not to say that all aspects of word meaning have the potential to be learned implicitly. Paradis (2004) has proposed that learning referential meaning requires the declarative system, and hence presumably depends upon awareness of the connection between form and meaning. But other aspects of meaning, such as semantic preferences, could be acquired implicitly. Although it is not clear whether the kind of learning demonstrated here actually depends on the operation of the procedural system, as hypothesised by Paradis, the present results do show that this aspect of verb meaning is amenable to implicit learning.

It should also be stressed that the present experiments demonstrate learning semantic preferences in a situation in which some aspect of the meaning of the verbs (i.e., their increase or decrease meaning) is already explicitly known (as in Experiment 2) or being intentionally inferred from context (Experiment 1). Thus, we regard these learning effects as essentially reflecting the process of 'tuning' an already-established meaning of which the participants are aware. Whilst this tuning process undoubtedly forms an important part of word learning through usage, it has to be distinguished from the process of actually forming new form-meaning connections from scratch.

Finally, the fact that there was no semantic implicit learning effect in Experiment 2 for non-native speakers provides a cautionary note in relation to the role this process may play in second language acquisition. The implication is that semantic regularities will be most effectively learned when attention is drawn to the relevant aspects of meaning by the task. Otherwise semantic implicit learning effects may be limited unless semantic access from known words in the context is highly automatic. This does not mean, though, that learners have to be aware of the actual underlying regularities. Even if in some cases it may be necessary to be aware of both form and meaning, it is not necessary to be aware of the semantic generalisations that they license.

\section*{Acknowledgments}

The first author gratefully acknowledges the support of the AHRC of Great Britain.

\section*{References}

Basnight-Brown, D. M., \& Altarriba, J. (2007). Differences in semantic and translation priming across languages. Memory \& Cognition, 35, 953965.

Bickerton, D. (1999). Creole languages, the language bioprogram hypothesis, and language acquisition. In W. C. Ritchie \& T. K. Bhatia (Eds.), Handbook of child language acquisition: Academic Press.
Bloom, P. (2000). How Children Learn the Meanings of Words: MIT Press.
Cleeremans, A., \& Jiménez, L. (2002). Implicit learning and consciousness: A graded, dynamic perspective. In R. M. French \& A. Cleeremans (Eds.), Implicit Learning and Consciousness (pp. 1-40). Hove: Psychology Press.
Dell, G. S., Reed, K. D., Adams, D. R., \& Meyer, A. S. (2000). Speech errors, phonotactic constraints, and implicit learning. Journal of Experimental Psychology-Learning Memory and Cognition, 26, 1355-1367.
Ellis, N. C. (1994). Vocabulary acquisition: The explicit ins and outs of explicit cognitive mediation. In N. C. Ellis (Ed.), Implicit and Explicit Learning of Languages (pp. 211-282). London: Academic Press.
Firth, J. R. (1957). Modes of Meaning. Papers in Linguistics 1934-1951 (pp. 190-215). Oxford: Oxford University Press.
Goujon, A., Didierjean, A., \& Marmeche, E. (2007). Contextual cueing based on specific and categorical properties of the environment. Visual Cognition, 15, 257-275.
Leung, J. H. C., \& Williams, J. N. (2012). Constraints on implicit Learning of grammatical form-meaning connections. Language Learning, 62, 634-662.
Pacton, S., Perruchet, P., Fayol, M., \& Cleeremans, A. (2001). Implicit learning out of the lab: The case of orthographic regularities. Journal of Experimental Psychology: General, 130, 401-426.
Paradis, M. (2004). A Neurolinguistic Theory of Bilingualism. Amsterdam/Philadelphia: John Benjamins.
Tzelgov, J. (1997). Specifying the relations between automaticity and consciousness: A theoretical note. Consciousness and Cognition, 6, 441-451.
Vinter, A., \& Perruchet, P. (1999). Isolating Unconscious Influences: The Neutral Parameter Procedure. Quarterly Journal of Experimental Psychology: Section A, 52, 857-875.
Williams, J. N. (2005). Learning without awareness. Studies in Second Language Acquisition, 27, 269304.

\title{
The role of verbal labels in attention to dimensional similarity
}

\author{
Lynn K. Perry (lkperry@wisc.edu) \\ Department of Psychology, 1202 W. Johnson Street \\ Madison, WI 53706 USA \\ Larissa K. Samuelson (larissa-samuelson@uiowa.edu) \\ Department of Psychology, E11 Seashore Hall \\ Iowa City, IA, 52242 USA
}

\begin{abstract}
In general, young children focus on holistic similarity and older children focus on dimensional similarity (selectively attending to one property, such as brightness, to the exclusion of others, such as size). Research on early word learning, however, suggests the process of learning words trains attention to category-relevant dimensions. We ask: does word learning scaffold dimensional attention more generally? By showing labels support dimensional attention, these results clarify the processes involved in similarity perception and unify our understanding of attentional processes in word learning with those in a broader context.
\end{abstract}

Keywords: Labels; categorization; selective attention.
Over the course of development, we become increasingly skilled at attending to one thing to the exclusion of others (see Hanania \& Smith, 2010 for review). For example, adults can easily focus on the color of a lime, rather than its exact shape or size, in order to distinguish it from a lemon. Evidence that older children and adults are generally much better than younger children at selectively attending to one dimension to the exclusion of others comes from a variety of domains, but we still do not know the process by which such changes occur. The goal of this paper is to explore the processes driving changes in selective attention with respect to their effects on similarity perception.

Of particular relevance to the current study is the holistic-to-dimensional shift, or the tendency for young children to focus on holistic similarity and older children and adults to focus on dimensional similarity relationships (Smith \& Kemler, 1977). Imagine you are presented with an orange, a yellowish-orange ball, and a yellowish-orange toy car. If you are an adult, you would be more likely to group the ball with the car because they match exactly along one dimension (i.e. identical in color). A young child, however, would group the orange with the ball because they are similar along multiple dimensions (i.e. shape and color)they are holistically similar. This shift in similarity perception occurs during the early school-age years, such that younger children ( \(<8\)-years-old) tend to be holistic classifiers and older children dimensional classifiers. Free classification, such as the triad classification task pictured in Figure 1, is the standard task used to examine this shift. As can be seen, two stimuli (A and B) match on one dimension (e.g. size) but vary greatly along another dimension (e.g. brightness). The third stimulus, C , is highly similar to A along both dimensions, but not identical to it on either. If a
participant were using holistic similarity, she would classify A and C together. If a participant were using dimensional similarity, she would classify A and B together. Smith and Kemler (1977) found that 5-6 year-olds made mostly AC matches and 10-11 year-olds made mostly AB matches.


Figure 1: Schematic representation of stimuli used in Smith and Kemler's 1977 triad classification task.

One explanation for this change comes from what we know about the role of word learning in the development of dimensional attention and categorization. For example, Landau \& Shipley (2001) demonstrated that when two objects are given the same label, participants generalize it to all intermediate morphs between the objects. If the two objects are given different names, participants divide the intermediates into two distinct categories. Lupyan and colleagues (2007) demonstrated that labels facilitated adults' category learning. Importantly, in this example, the labels were task irrelevant, providing information redundant with category structure. Other researchers have argued that redundant associations between cues strengthen the associations between other perceptual cues and category structure and that the facilitative effect of labels is fundamentally developmental (Yoshida \& Smith, 2005). By this view, over the course of word learning, the frequent redundancy between labels and other cues such as solidity, syntax, and category organization helps children use labels to facilitate learning. The more experience children have learning regularities in these overlapping cues, the better they can attend to relevant dimensions of similarity, at least in the context of further word learning.

This is most clearly seen in the development of wordlearning biases. For example, children acquire a shape bias, or the tendency to generalize names of novel objects by similarity in shape (Landau et al., 1988). This bias emerges from regularities present in the linguistic environment: a majority of the early words children learn name categories of solid objects organized by similarity in shape, e.g., ball.

As children learn more words, their attention is trained such that they automatically attend shape when learning names for solid objects. Smith and colleagues (2002) therefore describe word learning as "on-the-job training for attention," and have shown teaching children names of categories organized by similarity in shape leads them to precociously attend to shape when learning new words. As they learn words that name categories organized in other ways, e.g., names for nonsolids in categories organized by similarity in material or adjectives that name properties of objects, they acquire other biases and learn to flexibly attend to context-appropriate dimensions.

Over development, learning words directs children's attention to dimensional similarity in future word learning Thus, the critical unanswered question, however, is whether learning words also directs children's attention in nonlinguistic contexts. We propose word learning provides on-the-job training for attention more globally: in particular that labels scaffold dimensional attention in similarity perception. The regularity between words and attending to dimensional similarity leads to a higher-order association between labels and attending to dimensional similarity. According to our hypothesis, then, the tight links between labels, categorization, and dimensional similarity should gradually lead to a bias for dimensional similarity even outside of a labeling/word learning context. In 2 experiments we asked: 1) if there are developmental differences in category learning related to dimensional attention in similarity classification, and 2) if labels can support dimensional attention and facilitate categorization.

\section*{Experiment 1}

Experiment 1 examines if developmental differences in attention to dimensional similarity affect category learning.

\section*{Methods}

Participants 33 5- to 8 -year-old children participated. One child did not complete the experiment, thus there were 32 children in the final group.
Stimuli The stimuli were squares varying metrically in size and brightness (see Goldstone, 1994) and were presented on a pc with a touch screen monitor using Eprime 2.0.
Procedure Children completed a triad pretest and were divided into classifier groups based on performance on this task (Smith \& Kemler, 1977). There were 16 holistic classifiers (chose primarily holistic matches) and 16 dimensional classifiers (primarily dimensional matches).

Children next completed the category-learning task and category test. In the category-learning task children were presented with a square (a "rock") and asked to decide to which of two categories it came from (the ocean or jungle) by touching a picture on the computer screen. During the learning task, children received auditory feedback regarding accuracy of each decision (bell or buzzer sound). Learning blocks were made up of 8 trials, 4 trials for each category. Half of the children in each classification group were trained with categories organized by similarity in size (size
learners), half with brightness (brightness learners). The learning criterion was getting 7 out of 8 trials per block correct, 2 blocks in a row. If a child did not reach criterion after 30 blocks, the learning phase ended. The learning task was followed by the category test, where no feedback was given after each trial. Stimuli used at test included exemplars from training and 6 novel exemplars from each category. The test consisted of 4 blocks of 20 trials.

It was expected that dimensional classifiers would be able to learn the categories and generalize to novel category exemplars, but that holistic children should have more difficulty. For both groups, correct categorization of novel exemplars should require learning something about category organization rather than something about specific stimuli.

Next, children completed the discrimination task that measured their ability to distinguish between close values on test dimensions. Children were presented with a target and two test stimuli and asked to indicate which of the test stimuli matched the target by touching it. All three stimuli were present until the children responded. The target matched one of the test stimuli on every trial. Discrimination was tested both within and across category boundaries. All pairs were presented four times-each stimulus within a pair was presented twice as target and matching test item, and twice as foil-for a total of 96 trials. All children were presented with the same pairs, such that any given trial forced children to discriminate along a dimension that was only relevant for one group's learned category: e.g., a pair that differed only in size would test discrimination along the relevant dimension for the sizelearners and the irrelevant for brightness-learners. This allowed examination of changes in children's ability to make discriminations along category-relevant and irrelevant dimensions and within and between categories.

Finally, children completed a posttest triad task that was identical to the pretest version. Children who were dimensional classifiers on the triad pretest should still classify dimensionally as learning categories should not decrease their ability to selectively attend to dimensional similarity. Similarly, children who were holistic classifiers on the triad pretest should still classify holistically.

\section*{Results and Discussion}

Category learning It was expected that dimensional classifiers would be able to learn the categories but that holistic children would have more difficulty. To examine this, we measured the number of blocks to criterion for each child in the category-learning task (see Figure 2). A linear mixed regression model of the interaction between classifier type (holistic v. dimensional) and category structure (brightness v . size organization) on the number of blocks it took children to reach criterion, revealed a significant effect of classifier type such that dimensional classifiers were faster to reach criterion than holistic classifiers, \(t=21.88\), \(p<.0001\). (Because of the difficulty in determining degrees of freedom in linear mixed models, we conducted MCMC sampling to find p -values). This model also showed an
effect of category structure such that children were faster to reach criterion when learning categories organized by brightness, \(t=-23.00, p<.0001\). This brightness advantage replicates Goldstone's (1994b) finding with adults. There was also an interaction between classifier type and category structure such that holistic classifiers showed less of a difference in speed of learning brightness and size categories than dimensional classifiers, \(t=2.66, p<.01\). Thus, children who were able to selectively attend to a single dimension learned a novel category distinction based on one dimension faster than children who were holistic attenders. However, because dimensional classifiers were generally older than holistic classifiers, it is possible age could be the basis for these results. However, a model with both classifier type and age was significantly better than one with only age, \(X^{2}(1)=12659, p<.0001\), but no different from one with only classifier type. Thus classifier type, but not age, was necessary to account for findings.

We next examined performance in the other tasks using logistic mixed regression. These analyses included only data from children who reached learning criterion. We report results of classification groups separately because we are interested in whether dimensional and holistic classifiers both show, for example, enhanced between-category discrimination, than whether dimensional classifiers are more accurate than holistic classifiers. Because other researchers have found differences in learning brightness and size categories (e.g. Goldstone 1994), we examined performance of these groups separately. Results of regression models ( z and p values) are reported in Table 1.


Experiment 1: No Label
Experiment 2: Label
Figure 2: Number of blocks to criterion in category learning.
Category test We expected dimensional classifiers to generalize to new exemplars but that this should require learning something about category organization rather than about specific stimuli. Overall, children were very accurate in the categorization test. Dimensional classifiers who learned brightness categories, \(M=.87, t(7)=41.61, p<.0001\), dimensional classifiers who learned size categories, \(M=.77\), \(t(6)=4.98, p<.01\), holistic classifiers who learned brightness categories, \(M=.83, t(6)=6.33, p<.001\), and holistic classifiers who learned size categories, \(M=.70, t(4)=3.01, p<.05\), were all significantly better than chance.

Logistic mixed regression reveals that both dimensional classifiers who learned brightness and who learned size categories categorized familiar and novel stimuli equally well. Similarly, both holistic classifiers who learned brightness and who learned size categories categorized familiar and novel stimuli equally well. These results suggest both groups might have learned about the categoryrelevant dimension rather than specific stimuli.
Discrimination Goldstone (1994) found an advantage for relevant discriminations over irrelevant discriminations only for adults who learned brightness categories. Thus, while we expected dimensional classifiers would show worse withincategory discrimination along the irrelevant dimension and enhanced between-category discrimination, overall, we expected these results to be strongest for the brightness learners. Because it was predicted holistic classifiers would not be selectively attending to category-relevant dimensions, they should not demonstrate differences in discrimination on relevant versus irrelevant dimensions.

To examine changes in selective attention to categoryrelevant and irrelevant dimensions, we measured accuracy for each type of discrimination: between-category, withincategory along the relevant dimension and within-category along the irrelevant dimension. Overall, children were quite accurate in discriminating stimuli: dimensional classifiers who learned brightness categories, \(M=.82, t(7)=15.71\), \(p<.0001\), dimensional classifiers who learned size categories, \(M=.81, t(6)=6.27, p<.001\), holistic classifiers who learned brightness categories, \(M=.74, t(6)=5.96\), \(p<.001\), and holistic classifiers who learned size categories, \(M=.73, t(4)=5.04, p<.01\), were all significantly better than chance (.50) at discriminating stimuli.

A logistic mixed regression model of effect of discrimination comparison type (between, within relevant, within irrelevant) on accuracy revealed dimensional classifiers in the brightness learning group had significantly increased accuracy for between category compared to within category discriminations on either irrelevant, or relevant dimension. They were also more accurate at within-category discriminations along relevant than irrelevant dimensions, demonstrating, overall, they were more accurate at relevant than irrelevant discriminations. However, dimensional classifiers who learned size categories were no more accurate at between-category than within-category discriminations on either irrelevant, or relevant dimensions.

Holistic classifiers are thought not to selectively attend to category relevant dimensions. Therefore, they should not show the same pattern as dimensional classifiers. However, a logistic mixed regression model showed that holistic classifiers who learned brightness categories were more accurate at between category than within category discriminations along both the irrelevant, and relevant dimension. Additionally, these children were significantly more accurate at within-category discriminations along the relevant than the irrelevant dimension, demonstrating that overall, they were more accurate at discriminating across the relevant than the irrelevant dimension. Holistic classifiers

Table 1: Results of logistic mixed regression for Category Test, Discrimination (comparing between versus withincategory discriminations on the irrelevant and relevant dimensions and within-category discriminations on the relevant versus the irrelevant dimensions) and Triad Classification. z and p values/significance are reported for each model.
\begin{tabular}{lcccccccccc}
\hline & \multicolumn{4}{c}{ Experiment 1: No Label } & \multicolumn{3}{c}{ Experiment 2: Label } & & Comparison \\
\hline Classification & \multicolumn{2}{c}{ Dimensional } & \multicolumn{2}{c}{ Holistic } & \multicolumn{2}{c}{ Dimensional } & Holistic & Holistic \\
\hline Task & Bright & Size & Bright & Size & Bright & Size & Bright & Size & Bright & Size \\
\hline Cat. Test: & .61 & .54 & 1.26 & -.17 & .63 & .53 & .56 & .49 & .29 & -.12 \\
Fam v Novel & \(n s\) & \(n s\) & \(n s\) & \(n s\) & \(n s\) & \(n s\) & \(n s\) & \(n s\) & \(n s\) & \(n s\) \\
Discrim: b/t & -3.91 & .25 & -2.33, & -.68 & -4.50 & 1.49 & -5.29 & 1.28 & 2.25 & 1.24 \\
v w/in irrel & \(p<.01\) & \(n s\) & \(p<.05\) & \(n s\) & \(p<.01\) & \(n s\) & \(p<.01\) & \(n s\) & \(p<.05\) & \(n s\) \\
Discrim: b/t & -3.84 & .98 & -1.93, & -.39 & -3.93 & -.42 & -4.66 & -.43 & 2.28 & 1.57 \\
v w/in rel & \(p<.01\) & \(n s\) & \(p<.05\) & \(n s\) & \(p<.01\) & \(n s\) & \(p<.01\) & \(n s\) & \(p<.05\) & \(n s\) \\
Discrim: w/in & -3.70 & .70 & -2.11, & .01 & -4.50 & 1.49 & -5.29 & 1.49 & .79 & -.95 \\
rel v irrel & \(p<.01\) & \(n s\) & \(p<.05\) & \(n s\) & \(p<.01\) & \(n s\) & \(p<.01\) & \(n s\) & \(n s\) & \(n s\) \\
Triad: pre & -.81 & -.15 & .78 & 1.94 & -.65 & 2.43 & 2.43 & 3.81 & 3.34 & 3.55 \\
v post & \(n s\) & \(n s\) & \(n s\) & \(p<.10\) & \(n s\) & \(p<.05\) & \(p<.05\) & \(p<.01\) & \(\mathrm{p}<.01\) & \(\mathrm{p}<.01\) \\
\hline
\end{tabular}
learning size categories had no accuracy differences. Thus, both groups' category learning affected their discrimination abilities-but only if they learned brightness categories.
Posttest triad We next examined results of the posttest triad task. The primary question of interest is whether there was a change in the number of dimensional matches children choose from pre- to posttest. Such a change indicates learning dimensional categories-which require the learner to selectively attend to one dimension to the exclusion of another-increased children's selective dimensional attention in similarity classification. It was predicted that the dimensional classifiers would not show an increase in dimensional responding from pre to post test because they were already attending dimensionally. A logistic mixed regression model showed neither dimensional classifiers who learned brightness nor size categories were more likely to choose dimensional matches during posttest than pretest.

Holistic classifiers were also predicted to not show an increase in dimensional responding, because they were not expected to be attending dimensionally in the categorylearning task. A logistic mixed regression model showed holistic classifiers who learned brightness categories had no increase in dimensional responding. However, those who learned size categories were marginally more likely to select dimensional matches during posttest. This suggests that perhaps these children did learn to selectively attend to size. Conclusions Learning to categorize stimuli along a dimension increases attention to that dimension and leads to changes in discrimination. We predicted that for this to happen, the learner has to be able to attend to the relevant dimension in the first place-which holistic classifiers were not expected to do. The results of the category-learning task support this idea, demonstrating holistic classifiers were slower to learn categories organized by a single dimension.

However the results of the discrimination and posttest triad tasks paint a more complicated picture. For example, holistic classifiers who learned brightness categories showed similar changes in their discrimination as the dimensional classifiers did, and holistic classifiers showed increases in dimensional responding in the posttest triad task. These results suggest that category learning on its own can facilitate dimensional attention. The critical question, then, is do labels work with category learning to boost it even more?

\section*{Experiment 2}

Results of E1 suggest holistic classifiers are slower to learn categories than dimensional classifiers, but they show some increases in dimensional attention following category learning. Research shows that young children can do this in the context of novel noun generalization (Smith et al., 2002). Furthermore, we know labels facilitate category learning in adults (Lupyan et al., 2007). If labels support the development of dimensional attention more generally, then we should see a facilitative effect of labels on holistic classifiers' category learning and perhaps an even greater facilitation of their dimensional attention. In E2, we examined category learning in the context of redundant labels and assess subsequent changes in attention.

\section*{Methods}

Participants 35 5-8-year-olds participated. 3 children did not complete the experiment ( 2 quit and 1 for equipment error), thus there were 32 children in the final group.
Procedure Methods were identical to E1, except during category learning, novel auditory category labels (leebish and grecious) were presented after feedback (the bell or buzzer) on each trial. These labels were redundant with category structure such that, for example, after each trial
where a rock belonging from the ocean was presented (regardless of correct categorization) the child would hear the "leebish" Results were analyzed as in E1.

\section*{Results and discussion}

Category learning As can be seen in Figure 2, a linear mixed regression model of the interaction between classifier type (holistic \(v\). dimensional) and category structure (brightness v . size organization) on the number of blocks it took children to reach criterion revealed no effect of classifier type, \(t=.04, N S\), nor of category structure, \(t=-.30\), NS. Thus, labels facilitate learning of dimensional categories, such that holistic classifiers were now as quick to reach criterion as dimensional classifiers.
Category test Both groups were very accurate: dimensional classifiers who learned brightness categories, \(\mathrm{M}=.84\), \(\mathrm{t}(6)=11.11, \mathrm{p}<.0001\), dimensional classifiers who learned size categories, \(\mathrm{M}=.86, \mathrm{t}(6)=15.16, \mathrm{p}<.0001\), holistic classifiers who learned brightness categories, \(\mathrm{M}=.85\), \(\mathrm{t}(6)=13.32, \mathrm{p}<.0001\), and holistic classifiers who learned size categories, \(\mathrm{M}=.69, \mathrm{t}(7)=4.30\), \(\mathrm{p}<.01\), were all significantly better than chance at categorization.

A logistic mixed regression model of trial type on categorization accuracy showed neither dimensional classifiers who learned brightness nor size categories were different in accuracy for familiar and novel stimuli. Similarly, neither holistic classifiers who learned brightness nor size categories were different in accuracy for familiar and novel stimuli. This suggests that all children may have learned something about category organization rather than about specific stimuli.
Discrimination Children were accurate in discriminating stimuli: dimensional classifiers who learned brightness categories, \(M=.83, \quad t(\sigma)=16.40, \quad p<.0001\), dimensional classifiers who learned size categories, \(M=.80, t(6)=10.90\), \(p<.0001\), holistic classifiers who learned brightness categories, \(\quad M=.76, \quad t(6)=7.42, \quad p<.001, \quad\) and holistic classifiers who learned size categories, \(M=.77, t(7)=6.27\), \(p<.001\), were all significantly better than chance.

A logistic mixed regression model of effect of discrimination type (between, within relevant, within irrelevant) on accuracy showed that dimensional classifiers in the brightness-learning group were more accurate at between category than within category discriminations for both irrelevant, and relevant dimensions. These children were also significantly more accurate at within-category discriminations along relevant than irrelevant dimensions. Dimensional classifiers who learned size categories were no more accurate at between category discriminations than at within category discriminations on either the irrelevant, or relevant dimension, nor were they more accurate at withincategory discriminations along either dimension.
A logistic mixed regression model showed that holistic classifiers who learned brightness categories were more accurate at between category, compared to within category, discriminations along both the irrelevant, and relevant dimensions. However, those in size learners did not show any differences in discrimination. Thus, all children showed
an effect of category learning on discrimination-but only if they learned brightness.
Posttest triad It was predicted that dimensional classifiers would not increase in dimensional responding from pre- to posttest. A logistic mixed regression model showed dimensional classifiers who learned brightness were no more likely to choose dimensional matches during the post test than on pretest. Interestingly, however, dimensional classifiers who learned size were more likely to choose dimensional matches during post test.

If labels drive attention to dimensions, holistic classifiers should show an increase in attention dimensional similarity. In fact, a logistic mixed regression model showed both holistic classifiers who learned brightness and size categories increased in dimensional responding.
Conclusions Incidental labels in a category-learning task scaffolded selective attention to dimensional similarity. Unlike in E1, holistic and dimensional classifiers are equally quick to learn the categories. Holistic classifiers have relatively weak selective attention and take longer to learn dimensional categories. Once they learn the categories, however, they show slight increases in dimensional attention-as evidenced by discrimination accuracy and increases in dimensional classification. Labels support even weak selective attention, thus when holistic classifiers learn dimensional categories in the context of labels, they learn the categories more quickly. This increase in selective attention cascades forward to both their discrimination abilities and classification biases. So, as Lupyan suggested in his 2008 study of category grouping on visual processing, "categories matter; named categories matter more." A direct comparison of holistic classifiers from the two experiments should clarify the extent to which performance of those in E2 is, in fact, significantly better than those in E1.

\section*{Between-experiment comparison}

Category learning A linear mixed regression model of the interaction between experiment (label v. no label) and category structure (size v. brightness) on the number of blocks to reach learning criterion revealed holistic classifiers were significantly faster to reach criterion in the label than in the no label experiment, \(\mathrm{t}=-26.57, \mathrm{p}<.0001\). There was an overall effect of category structure, such that children were faster to learn brightness than size categories, \(\mathrm{t}=-23.66\), \(\mathrm{p}<.0001\), however, there was also an interaction such that children in the label experiment showed less difference in speed of learning the two category types than those in the no label experiment, \(\mathrm{t}=7.47, \mathrm{p}<.0001\). This is direct evidence that labels facilitate category learning in children who have difficulty attending to dimensional similarity.
Category test A mixed logistic regression model of the interaction between experiment (label versus no label) and trial type (familiar or novel) on children's categorization revealed that neither holistic classifiers who learned brightness nor size categories showed an effect of experiment. Thus, children were equally accurate at categorizing familiar and novel stimuli.

Discrimination If labels facilitate category learning because they increase children's selective attention to categoryrelevant dimensions, then children in the label experiment should show the biggest enhancement in accuracy for relevant over irrelevant discriminations. A logistic mixed regression model of the interaction between experiment (label v. no label) and discrimination type (between, withinrelevant, or within-irrelevant) revealed that children in the label experiment showed a larger difference in accuracy in between-category discriminations relative to withincategory discriminations on the relevant dimension, and relative to within-category discriminations on the irrelevant dimension. However, size learners did not show an effects of experiment. This demonstrates (for brightness learners) the presence of a label made effects of categorization on discrimination stronger. Labels increase selective attention to dimensions above and beyond category learning.
Posttest triad If word learning drives the emergence of selective attention to dimensions over development, then when labels are presented during category learning we may also see indices of these changes in attention over the course of an experiment. Therefore holistic classifiers in the label experiment should show the largest increases in dimensional responding from pre- to posttest triad task. This was supported by logistic mixed regression models of the effect of experiment (label v. no label) on change in dimensional responses from pre- to posttest triad revealing that both holistic classifiers who learned brightness and size categories had a larger increase in dimensional responding from pre- to posttest in the label experiment. The presence of a label not only immediately facilitated category learning, but also led to cascading changes in similarity classification.

\section*{Conclusions}

Results of the comparison analyses demonstrate the extent to which labels support selective attention above and beyond category learning by demonstrating that holistic classifiers were significantly faster to learn in the presence of labels. Similarly, while the qualitative assessment of E1 and 2 demonstrated both groups of holistic classifiers who learned brightness categories showed the "adult" pattern found by Goldstone (1994), the comparison analysis shows that those in the label experiment demonstrate a more extreme pattern. This suggests category learning affects attention to dimensions, but labeled category learning affects it more. Finally, comparison of changes in dimensional responding in the triad task offers additional evidence that labels scaffold attention to dimensional similarity above and beyond category learning. The only difference between the two experiments was the presence of incidental, redundant, labels during category learning. Yet this was enough to change children's pattern of responding in an unrelated similarity classification task.

Overall, these analyses generally demonstrate that while category learning supports selective dimensional attention even in children who preferentially attend to holistic similarity, labeled category learning exaggerates
this process, facilitating both category learning and attention to dimensional similarity. However, one important remaining question that needs to be addressed by future research is why category learning and labeled category learning only have these effects when the categories are organized by brightness and not size. While our results replicate Goldstone's 1994 findings with adults, it is still unclear why brightness is easier to learn and why learning brightness, but not size, should lead to differences in discrimination abilities.

Our results clearly demonstrate that labels can scaffold this attention in similarity classification, suggesting that, as in the case of early word learning biases, regularities between labels, categories, and similarity lead to selective attention to dimensions. Future research will be needed to further explore how children can eventually do this without any external linguistic support. Nevertheless, these experiments are an important first step in unifying our understanding of the attentional processes involved in early word with those in a broader context.

\section*{Acknowledgments}

This work was supported by a Ballard Seashore dissertation year fellowship awarded to LKP and by R01 HD045713 awarded to LKS by the Eunice Kennedy Shriver National Institute Of Child Health \& Human Development.

\section*{References}

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. Journal of Experimental Psychology: General, 123(2), 178-200.
Hanania, R., \& Smith, L. B. (2010). Selective attention and attention switching: Towards a unified developmental approach. Developmental Science, 13(4), 622-635.
Landau, B., \& Shipley, E. (2001). Labelling patterns and object naming. Developmental Science, 4(1), 109-118.
Landau, B., Smith, L. B., \& Jones, S. S. (1988). The importance of shape in early lexical learning. Cognitive Development, 3(3), 299-321.
Lupyan, G. (2008). The conceptual grouping effect: Categories matter (and named categories matter more). Cognition, 108(2), 566-577.
Lupyan, G., Rakison, D. H., \& McClelland, J. L. (2007). Language is not just for talking: Redundant labels facilitate learning of novel categories. Psychological Science, 18(12), 1077-1083.
Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., \& Samuelson, L. K. (2002). Object name learning provides on-the-job training for attention. Psychological Science, 13(1), 13-19.
Smith, L. B., \& Kemler, D. G. (1977). Developmental trends in free classification: Evidence for a new conceptualization of perceptual development. Journal of Experimental Child Psychology, 24, 279-298.
Yoshida, H., \& Smith, L. B. (2005). Linguistic cues enhance the learning of perceptual cues. Psychological Science, 16(2), 90.

\title{
A category theory perspective on compositionality and (the development of) cognitive capacity
}

\author{
Steven Phillips (steve@ni.aist.go.jp) \\ Mathematical Neuroinformatics Group, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki 305-8568 JAPAN
}

\begin{abstract}
A remarkable property of human cognition is the systematic co-occurrence of certain cognitive abilities. One challenge for cognitive science is to determine the (computational) principles that derive this property as part of a broader goal of establishing the foundations of cognitive architecture (i.e. the basic processes and modes of combination affording cognitive capacity) for a science of cognition. This paper continues a category theory approach to compositionality and cognitive capacity that posits universal construction (e.g., products) as a fundamental cognitive architectural component. As shown here, products can be modeled in other frameworks, thereby providing a link between an abstract computational principle and a concrete cognitive resource needed for particular capacities. For example, a network of weighted connections implementing a categorical product uses fewer resources when the number of task instances sharing a common product structure is greater than two; otherwise it is more economic to realize each instance independently. This cross-over may explain why human cognition is not always systematic: the cost of universal construction may not outweigh its expected gain.
\end{abstract}

Keywords: category theory; cognitive capacity; compositionality; systematicity, similarity; product; universal construction

\section*{Introduction}

A remarkable property of human cognition is the systematic co-occurrence of certain cognitive abilities. For example, if one has the ability to infer square as the first shape in the pair (square, triangle), then one also has the ability to infer triangle as the first shape in the pair (triangle, square), assuming that squares and triangles are recognizable. This property is called systematicity (Fodor \& Pylyshyn, 1988), and is characterized as having capacity \(c_{1}\) if and only if \(c_{2}\) (McLaughlin, 2009), i.e., as an equivalence class of cognitive capacities.

For cognitive science, a major challenge has been to explain why cognitive capacity is systematically distributed in a particular (non-arbitrary) way. The main theoretical frameworks propose some form of compositionality to explain systematic cognitive capacity. The classical (symbol system) explanation is that cognitive processes are sensitive to grammatical structure (Fodor \& Pylyshyn, 1988). Thus, the two shape capacities (above) are inseparable because they involve a common process: say, \((p, q) \rightarrow p\), where \(p\) and \(q\) are symbols for squares and triangles, on the assumption of having component processes \(\square, \triangle \rightarrow p, q\) for recognizing squares and triangles. \({ }^{1}\) Similarly, a connectionist (neural network) account of capacity can make use of common processes in

\footnotetext{
\({ }^{1}\) Systematicity pertains to "molecular" not "atomic" capacities, hence that an architecture permits the recognition of squares, but not triangles, is not a counterexample to having the systematicity property (Fodor \& Pylyshyn, 1988).
}
the form of activation units and weighted connections modeling cognitive processes that are sensitive to spatial structure (Phillips \& Wilson, to appear). In this case, the two shape capacities factor through a common (sub)network of weighted connections that map (neuronal) vector representations of shape pairs to first shapes. Although classical and connectionist models can be constructed such that capacity \(c_{1}\) if and only if \(c_{2}\), systematicity doesn't necessarily follow from classical, or connectionist principles: one can also devise models from those principles such that \(c_{1}\), but not \(c_{2}\). Additional, ad hoc (i.e. arbitrary) assumptions are needed to exclude models that lack systematicity. So, classical and connectionist theories (principles) fail to fully explain systematicity (Aizawa, 2003).

Another approach to the systematicity problem (Phillips \& Wilson, 2010, 2011, 2012) used a mathematical theory of structure, called category theory (Mac Lane, 2000). In category theory, a universal construction relates a collection of arrows (interpreted as cognitive processes) via a common mediating arrow (cognitive process) in a unique way. Hence, all capacities are indivisibly linked to this mediating arrow as the common component. The mediating arrow is associated with an equivalence class of systematically related cognitive capacities (Phillips \& Wilson, 2011, Text S4).

This paper further explores a categorical approach to compositionality and cognitive capacity by looking at the relationship between universality and cognitive resources. First, our category theory approach to the systematicity problem, and related models, are recalled to motivate its continued use here. Then, the relationship between a specific universal construction (product) and the resources needed for implementation are examined: fewer resources are needed when the number of task instances sharing a product is greater than two; otherwise, it is cheaper to realize each task instance independently. The implications of this relationship for development (learning) are discussed in the final section.

\section*{Compositionality and universal constructions}

Definitions of category and product are provided in this section. Other definitions are provided in Appendix A. Deeper introductions to category theory can be found in many books on the subject (see, e.g., Mac Lane, 2000; Simmons, 2011).

\section*{Category theory and cognition}

Category theory starts with a definition of a category, which is a collection of objects and relations between objects, called arrows (or morphisms, or maps). The category theory approach to cognition presented here regards a cognitive archi-
tecture (i.e., the basic component processes and the ways in which those processes are combined) as a category (of possibly other categories), where objects are interpreted as components of the architecture and arrows are relations between those components. For instance, an object may be a set of cognitive representational states and an arrow may be a function between two sets of states (objects) that transforms cognitive representations (states).
Definition (Category). A category \(\mathbf{C}\) consists of:
- a class of objects \(|\mathbf{C}|=(A, B, \ldots)\);
- for each pair of objects \(A\) and \(B\) in \(\mathbf{C}\), a set \(\mathbf{C}(A, B)\) of morphisms (also called arrows, or maps) from \(A\) to \(B\), where each morphism \(f: A \rightarrow B\) has \(A\) as its domain and \(B\) as its codomain, including the identity morphism \(1_{A}: A \rightarrow A\) for each object \(A\); and
- a composition operation, denoted " \(\circ\) ", of morphisms \(f\) : \(A \rightarrow B\) and \(g: B \rightarrow C\), written \(g \circ f: A \rightarrow C\) that satisfies the laws of:
- identity, where \(f \circ 1_{A}=f=1_{B} \circ f, \forall f: A \rightarrow B\); and
- associativity, where \(h \circ(g \circ f)=(h \circ g) \circ f, \forall f: A \rightarrow\) \(B, g: B \rightarrow C, h: C \rightarrow D\).

An example of a category is Set, which has sets for objects and total functions for arrows, where the identities are the identity functions and composition is function composition. Set is used here to model cognitive architecture as a collection of sets of representational states and cognitive processes as functions sending states to states, where identities are "do nothing" functions.

Most models of cognition support some kind of compositionality that affords the representation of a pair of entities from which the constituent entity representations are recoverable. In category theory, a basic kind of (universal) construction of a pair of objects from which the component objects are recoverable is called a product. \({ }^{2}\)
Definition (Product). In a category \(\mathbf{C}\), a product of two objects \(A\) and \(B\) is an object \(P\) (also denoted \(A \times B\) ) together with two morphisms \(p_{1}: P \rightarrow A\) and \(p_{2}: P \rightarrow B\), such that for every object \(Z \in|\mathbf{C}|\) and every pair of morphisms \(f: Z \rightarrow A\) and \(g: Z \rightarrow B\), there is a unique morphism \(u: Z \rightarrow P\) (also denoted \(\langle f, g\rangle\), since it is determined by \(f\) and \(g\) ), such that \(f=p_{1} \circ u\) and \(g=p_{2} \circ u\), as indicated in commutative diagram \({ }^{3}\)


\footnotetext{
\({ }^{2} \mathrm{~A}\) closely related construction involving pairs of objects is called coproduct, which is obtained by reversing the directions of the arrows in the definition of product (Mac Lane, 2000).
\({ }^{3}\) I.e. where all pairs of paths (one or both having more than one arrow) from the same start object to the same finish object are equal.
}

In Set, the Cartesian product of sets \(A\) and \(B\) is the set \(A \times B=\{(a, b) \mid a \in A, b \in B\}\) together with functions \(p_{1}\) : \((a, b) \mapsto a\) and \(p_{2}:(a, b) \mapsto b\) recovering the first and second elements of each pair is a product. Accordingly, for any set \(Z\) and functions \(f\) and \(g\) there is a unique function, called the product function, \(\langle f, g\rangle: z \mapsto(f(z), g(z))\), that maps each element \(z \in Z\) to the pair of elements \((f(z), g(z))\) with the desired elements \(f(z) \in A\) and \(g(z) \in Z\) recoverable via \(p_{1}\) and \(p_{2}\). Projections \(p_{1}\) and \(p_{2}\) are part of this universal construction; they are the same functions employed for every \(Z, f\) and \(g\). Cartesian product is an instance of categorical compositionality, and a concrete example of deriving systematicity with regard to the shape-pairs example (see next).

\section*{Categorical compositionality and systematicity}

A challenge for cognitive science is to explain why cognitive capacity is organized in a particular way: e.g., why is it that if one can infer square as the first shape in the pair (square, triangle), then one can infer triangle as the first shape in the pair (triangle, square), given that they can recognize squares and triangles; yet, the capacity to infer squares and triangles is independent of the capacity to count? The general, categorytheoretic claim is that underlying every collection of systematically related cognitive capacities is a universal construction of some kind. Each member of a collection of systematically related capacities is realized by two cognitive components: a common component and a unique capacity-specific component. \({ }^{4}\) The presence or absence of the common arrow implies the presence or absence of the entire collection of capacities. For the shapes example, the common arrow(s) corresponds to a capacity to retrieve the first and second elements of a pair, and the unique arrow corresponds to recognition of specific shape instances. By contrast, the capacity to count involves a different kind of universal construction (Phillips \& Wilson, 2012), and hence a different mediating arrow.

\section*{Models of categorical compositionality}

The relationship to the above categorical theory of compositionality and systematicity to specific (classical, or connectionist) models is analogous to a relationship in categorical algebra, where a model of (say) a theory of groups is a functor (a structure-preserving map, see Appendix A) from a category representing the theory and a category capturing the semantics of the model, e.g., Set, where the group is based on elements of a set (Lawvere, 1963). A (classical or connectionist) model of categorical compositionality and therefore systematicity is a functor from a universal construction (as a category) to a particular concrete category of cognitive representations and processes. \({ }^{5}\)

Returning to the shape example, suppose we model cognitive architecture in the category Set where objects are sets of cognitive representations and arrows are cognitive processes

\footnotetext{
\({ }^{4}\) See Appendix A for the general form of universal construction.
\({ }^{5}\) Functors compose, so a model at one level may be construed as a theory (functor image as a category) relative to a model of that theory at another level (functor from it into another category).
}
mapping representations. Consider the collection of objects and arrows (identities not shown) in Diagram 1 as a category (theory). A functor (model) of this category into Set is indicated by commutative diagram

where \(S=\) \{square, triangle \(\}\) is a set of elements representing square and triangle, \(S \times S=\{(\) square, triangle \(), \ldots\}\) is the set of all pairwise combinations of square and triangle, \(Z=\{\square \triangle\}\) is a singleton set containing the image of a square to the left of a triangle, \(Z^{\prime}=\{\triangle \square\}\) is a singleton set containing the image of a triangle to the left of a square, \(f\) st \(: \square \triangle \mapsto\) square maps the image of a square and triangle to square, snd \(: \square \triangle \mapsto\) triangle maps the image of a square and triangle to triangle, and \(u=\langle f s t, s n d\rangle: \square \triangle \mapsto\) (square, triangle). Functions \(f s t^{\prime}\) and \(s n d^{\prime}\) are defined similarly. Systematicity is realized by common mediating functions (projections) \(p_{1}\) : (square, triangle) \(\mapsto\) square, \(\ldots\) and \(p_{2}\) : (square, triangle) \(\mapsto\) triangle, \(\ldots\). This model can be regarded as classical where each element is a symbol, and the constituents of the complex symbols for each pair are tokened whenever each pair is tokened-classical compositionality.

Another model is obtained from the category Vec of vector spaces for objects and linear functions for arrows. In the case of the objects and arrows in Diagram 2, suppose \(S\) is a vector space over the field of real numbers containing vectors \(\vec{s}\) and \(\vec{t}\) representing square and triangle, respectively. Likewise, suppose \(Z\) and \(Z^{\prime}\) are vector spaces containing vectors \(\overrightarrow{s t}\) and \(\vec{t} s\) representing the images \(\square \triangle\) and \(\triangle \square\), respectively. A product object in Vec is a product vector space. Hence, \(S \times S\) is a product vector space of vector space \(S\) with itself containing all pairwise products of vectors \(\vec{s}\) and \(\vec{t}\), e.g., \(\vec{s} \vec{t}\) and \(\vec{t} \vec{s}\), etc. In this model, the arrows are linear functions mapping vectors in one space to vectors in another space. Specifically, \(f s t: \overrightarrow{s t} \mapsto \vec{s},\langle f s t, s n d\rangle: \overrightarrow{s t} \mapsto \vec{s} \vec{t}, p_{1}: \vec{s} \vec{t} \mapsto \vec{s}\) and \(p_{2}: \vec{s} \vec{t} \mapsto \vec{t}\) (similarly for the other arrows). If we choose particular bases for these vector spaces, then we have a category of coordinate spaces, and the linear functions are realized as matrices, where the identities are identity matrices and composition is matrix multiplication. Further, if we identify each basis vector with a neuron, hence each object is a collection of real-valued neurons, and each matrix as a matrix of weighted connections between neurons in different objects, then we have a connectionist network realizing this model. Systematicity is realized by common linear functions \(p_{1}\) and \(p_{2}\). Note that there is no requirement that constituent vectors be tokened whenever their host product vectors are tokened (e.g., the coordinates specifying \(\vec{s}\) are not necessarily a subset of the coordinates specifying \(\vec{s} \vec{t}\) )—connectionist (functional) compositionality.
The vector spaces and linear functions in this Vec-based model of compositionality contain many more vectors and
mappings than may be needed as a model of cognitive capacity. A further connectionist refinement of the model may be to add connections between neurons within an object, e.g., as in a Hopfield network (Hopfield, 1982), so as to restrict the number of representational states to just those needed (in the shape example, two vectors representing square and triangle for the object \(S\), and four vectors representing each pairwise combination of square and triangle for the product object \(S \times S\) ). In this case, we have networks (not just collections of neurons) for objects, whence we need some kind of structurepreserving map for arrows satisfying the usual axioms to be a category. However, not just any combination of objects and arrows constitutes a product (see next). One possibility, for further work, is to extend the category \(\mathbf{G p h}\) of graphs and graph homomorphisms, which has products, \({ }^{6}\) by considering a connectionist network as a graph with additional structure. \({ }^{7}\)

\section*{(Non-)Universality and cognitive capacity}

Understanding the relationship between universal constructions, systematic cognitive capacity and cognitive resources requires a specific model of categorical compositionality that gives meaning to a concept of cognitive resource. An example of compositionality that does not involve a universal construction follows. Suppose a Set-based model of the shape-related capacities, shown in Diagram 3, that consists of an object \(T=\left\{t_{1}, t_{2}\right\}\) and arrows \(g_{1}: Z \rightarrow T ; \square \triangle \mapsto t_{1}\), \(h_{1}: T \rightarrow S ; t_{1} \mapsto\) square, \(t_{2} \mapsto\) triangle and \(h_{2}: T \rightarrow S ; t_{1} \mapsto\) triangle, \(t_{2} \mapsto\) square, where composition \(h_{1} \circ g_{1}\) correctly infers square as the first object from image \(\square \triangle\) and \(h_{2} \circ g_{1}\) correctly infers triangle as the second object from image \(\square \triangle\). Likewise, suppose arrow \(g_{2}: Z^{\prime} \rightarrow T ; \triangle \square \mapsto t_{2}\), where composition \(h_{1} \circ g_{2}\) correctly infers triangle as the first object from image \(\triangle \square\) and \(h_{2} \circ g_{2}\) correctly infers square as the second object from image \(\triangle \square\). The object \(T\) together with arrows \(h_{1}\) and \(h_{2}\) do not constitute a product, because for object \(Z^{\prime \prime}=\{\square \square\}\) and arrows \(f s t^{\prime \prime}: Z^{\prime \prime} \rightarrow S ; \square \square \mapsto\) square and snd \(d^{\prime \prime}: Z^{\prime \prime} \rightarrow S ; \square \square \mapsto\) square there does not exist an arrow \(u^{\prime \prime}: Z^{\prime \prime} \rightarrow T\) such that \(h_{1} \circ u^{\prime \prime}=f s t^{\prime \prime}\) and \(h_{2} \circ u^{\prime \prime}=s n d^{\prime \prime}\). This arrangement does not support the systematically related capacities of inferring square as the first and second shapes from the image \(\square \square\). (A similar lack of systematicity also applies for image \(\triangle \triangle\), etc.)


In general, for a network that dedicates one unit for each element of each set, the number of units required to represent

\footnotetext{
\({ }^{6}\) A categorical product of two unlabeled graphs \(G\) and \(H\) is the Cartesian product of their nodes with edges between pairs of nodes just in case their is an edge between the first node of each pair in \(G\) and an edge between the second node of each pair in \(H\).
\({ }^{7}\) See Ehresmann and Vanbremeersch (2007) for a category theory description of neurons and systems, and Healy et al. (2009) for an application to neural network modeling.
}
a categorical product of sets \(A\) and \(B\) is the size of set \(A\) times the size of set \(B\). This additional cost in resources to represent a universal construction raises the question of why a cognitive system would incur such an expense. As shown in the next section, implementing a categorical product becomes cheaper when the number of task instances sharing a common product structure exceeds a certain number. That number depends on the relative costs of representing the task-common and taskspecific components.

\section*{Resources and (non-)universal constructions}

A simple calculation for the resources needed to realize a sequence of task instances with a common product structure using non-universal versus universal constructions reveals a cross-over point. Suppose a sequence of tasks such that each task \(i \in\{1,2,3, \ldots, n\}\) consists of two maps \(f_{i}: Z_{i} \rightarrow A\) and \(g_{i}: Z_{i} \rightarrow B\). Consider a cognitive system without the capacity for constructing products. To realize \(n\) such tasks, a cognitive system must implement \(2 n\) functions (i.e. two functions for each task instance). Consider a second architecture with the capacity to construct a product \(\left(A \times B, p_{1}, p_{2}\right)\). In this case, the \(n\) tasks are realized by the two projections \(p_{1}: A \times B \rightarrow A\) and \(p_{2}: A \times B \rightarrow B\), and the \(n\) unique functions \(u_{i}: Z_{i} \rightarrow A \times B\), totalling \(n+2\) functions. Thus, the advantage of constructing products is obtained when \(n>2\), i.e. when the cognitive system must have a capacity for three or more task instances sharing the same product structure.

The precise relationship between cognitive resource and systematic cognitive capacity will depend on the underlying model and the kind of universal construction. The general form of a universal construction (given in Appendix A) with respect to a functor \(F\) is a pair \((A, \phi)\) such that each capacity \(f: F(Z) \rightarrow Y\) is composed of a common component \(\phi: F(A) \rightarrow Y\) and a unique component \(F(u): F(Z) \rightarrow F(A)\) such that \(f=\phi \circ F(u)\). Hence, the number of task instances \((n)\) at which fewer resources are deployed to realize capacity via a universal construction will depend on the relative costs of realizing components \(\phi, F(u)\) and \(f\). The benefit of universal constructions is more pronounced when the most expensive part of realizing \(f\) is with \(\phi\), since the \(\phi\) construction is only required once, whereas one \(F(u)\) component is required for each task instance. In general, computing with universal constructions over \(n\) tasks becomes cheaper when
\[
\begin{equation*}
n>\frac{\operatorname{cost}(\phi)}{\operatorname{cost}(f)-\cos t(F(u))} \tag{4}
\end{equation*}
\]

Hence, the benefit of universal construction will be obtained from fewer task instances when the cost of constructing each unique component \(F(u)\) is low compared to each \(f\). Otherwise, the benefits will only start to accrue after realizing many task instances.

\section*{Discussion}

Category theory generalizes the classical and connectionist notions of compositionality. Classical compositionality is the
idea that representations of the constituents of complex entities are tokened, in a consistent way, whenever the representations of their complex host entities are tokened (Fodor \& Pylyshyn, 1988). Symbol systems are the paradigmatic classical cognitive architecture, e.g., where the symbols representing constituents John, Mary, and loves appear in the symbolic representation of complex entity John loves Mary. Categorical compositionality generalizes this idea from symbols to arrows, and from tokening to arrow composition, while specializing the kinds of compositions to those that constitute universal constructions. This specialization is crucial to avoid otherwise arbitrary assumptions over which modes of tokening capture systematicity.

Connectionist compositionality is the idea that representations of complex entities are a function of the representations of their constituent entities without necessarily being tokened whenever the complex host representation is tokened (van Gelder, 1990). Category theory generalizes this idea from functions between sets to arrows between objects with additional internal structure beyond set membership and spatial structure (e.g., a graph, or a group), while specializing to functors (which appear in the definition of universal construction) that preserve that internal structure. This specialization is crucial, since although a functor is a kind of generalized function mapping objects and arrows, not all generalized functions are functors (i.e., preserve identity and associativity). Again, specialization is crucial to avoid arbitrary assumptions over which generalized functions capture systematicity.

\section*{Systematic capacity and cognitive development}

Models of universal constructions provide a functorial link from an abstract computational principle to a concrete cognitive resource. A connectionist example illustrated this point as a capacity-resource tradeoff. This non-universal situation may arise during cognitive development when the cognitive system does not have sufficient cognitive/neural resources to compute products, as suggested by the example in Set. Consistent with this idea, a category theory approach to the analysis of reasoning in young children explained developmental differences in terms of the capacity to compute (co)products (Phillips, Wilson, \& Halford, 2009). The analysis was based on empirical evidence from multi-paradigm studies where children from different age groups were tested on various reasoning tasks under easy and hard conditions (see Halford, 1993): a repeatedly observed result (see, e.g., Andrews \& Halford, 2002; Andrews, Halford, Bunch, Bowden, \& Jones, 2003; Andrews, Halford, Murphy, \& Knox, 2009) was that if a child (typically older than five years) was significantly above chance for both easy and hard conditions on one reasoning task (e.g., transitive inference), then that child was significantly above chance for easy and hard conditions on another reasoning task (e.g., class inclusion); conversely, if a child (typically younger than five years) was significantly above chance for the easy, but not the hard condition on one task, then they were significantly above chance for the easy,
but not the hard condition on another task. Analysis of the easy and hard conditions for seven reasoning tasks used to test children in such multi-paradigm studies showed that all hard conditions involved some kind of binary (co)product, in contrast to the easy conditions that did not, or what could also be called a unary (co) product \({ }^{8}\) (Phillips et al., 2009). \({ }^{9}\)

This relationship between universality and cognitive resource may shed light on why human cognition is not always systematic: the initial (short-term) cost of realizing a universal construction does not outweigh the expected (longterm) gain. Put in more familiar terms, when faced with a problem one often has two choices: devise a quick and simple solution that works for the current situation only, or a general-purpose solution that takes more time to develop. This choice will depend on how likely the same problem will reappear in other contexts, and whether the cost of developing a general-purpose solution will be repaid in subsequent savings. From a developmental perspective, this choice may be forced: the necessity of having a solution that works at least for the currently presented cases may preclude development of a general-purpose alternative. Hence, children may first go through a period of non-systematicity before attaining some forms of systematic cognitive capacity.

If non-systematicity is forced by the immediate demands on having a functioning system, then how do universal constructions (systematic capacity) eventually develop? The development of universal constructions may be driven (in part) by limited cognitive resources and the subsequent need to simultaneously realize multiple instances of tasks sharing a common component.
Further work is needed to explore the implications of the relationship between compositionality, systematic cognitive capacity and cognitive resource. The category-theoretic analysis presented here identified the essence of this relationship.

\section*{Acknowledgments}

This work was supported by a Japanese Society for the Promotion of Science Grant-in-aid (Kakkenhi: 22300092).

\section*{References}

Aizawa, K. (2003). The systematicity arguments. New York: Kluwer Academic.
Andrews, G., \& Halford, G. S. (2002). A cognitive complexity metric applied to cognitive development. Cognitive Psychology, 45(2), 153-219.
Andrews, G., Halford, G. S., Bunch, K. M., Bowden, D., \&

\footnotetext{
\({ }^{8} \mathrm{~A}\) unary product object is just \(A\), which is isomorphic to the product of \(A\) with a terminal object (i.e., \(A \times 1 \cong A\) ). In the dual case, \(A\) is isomorphic to the coproduct of \(A\), denoted + , with an initial object, denoted 0 (i.e., \(A+0 \cong A\) ).
\({ }^{9}\) One possible neural resource involved in categorical products is synchrony: analysis of EEG phase-locking values (PLV) during visual search for a target identifiable on one, two, or three feature dimensions, corresponding to a unary, binary, and ternary product (respectively), revealed significantly positive PLV-arity slopes for (frontal, parietal) electrode pairs (Phillips, Takeda, \& Singh, 2012).
}

Jones, T. (2003). Theory of mind and relational complexity. Child Development, 74(5), 1476-1499.
Andrews, G., Halford, G. S., Murphy, K., \& Knox, K. (2009). Integration of weight and distance information in young children: The role of relational complexity. Cognitive Development, 24(1), 49-60.
Ehresmann, A. C., \& Vanbremeersch, J.-P. (2007). Memory evolutive systems: Hierarchy, emergence, cognition (Vol. 4). Oxford, UK: Elsevier.
Fodor, J. A., \& Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. Cognition, 28(1-2), 3-71.
Halford, G. S. (1993). Children's understanding: The development of mental models. Hillsdale, NJ: Lawrence Erlbaum.
Healy, M. J., Olinger, R. D., Young, R. J., Taylor, S. E., Caudell, T. P., \& Larson, K. W. (2009). Applying category theory to improve the performance of a neural architecture. Neurocomputing, 72(13-17), 3158-3173.
Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences, 79(8), 25542558.

Lawvere, F. W. (1963). Functorial semantics of algebraic theories. Proceedings of the National Academy of Sciences, 50(5), 869-872.
Mac Lane, S. (2000). Categories for the working mathematician (2nd ed.). New York, NY: Springer.
McLaughlin, B. P. (2009). Systematicity redux. Synthese, 170, 251-274.
Phillips, S., Takeda, Y., \& Singh, A. (2012). Visual feature integration indicated by phase-locked frontal-parietal EEG signals. PLoS ONE, 7(3), e32502.
Phillips, S., \& Wilson, W. H. (2010). Categorial compositionality: A category theory explanation for the systematicity of human cognition. PLoS Computational Biology, 6(7), e1000858.
Phillips, S., \& Wilson, W. H. (2011). Categorial compositionality II: Universal constructions and a general theory of (quasi-)systematicity in human cognition. PLoS Computational Biology, 7(8), e1002102.
Phillips, S., \& Wilson, W. H. (2012). Categorial compositionality III: F-(co)algebras and the systematicity of recursive capacities in human cognition. PLoS ONE, 7(4), e35028.
Phillips, S., \& Wilson, W. H. (to appear). A category theory explanation for systematicity: Universal constructions. In P. Calvo \& J. Symons (Eds.), Systematicity and cognitive architecture ( \(\mathrm{p} . \mathrm{xx}\) ). Cambridge, MA: MIT Press.
Phillips, S., Wilson, W. H., \& Halford, G. S. (2009). What do Transitive Inference and Class Inclusion have in common? Categorical (co)products and cognitive development. PLoS Computational Biology, 5(12), e1000599.
Simmons, H. (2011). An introduction to category theory. New York, NY: Cambridge University Press.
van Gelder, T. (1990). Compositionality: A connectionist

\section*{Appendix A}

Definition (Functor). A functor \(F: \mathbf{C} \rightarrow \mathbf{D}\) is a map from category \(\mathbf{C}\) to category \(\mathbf{D}\) that associates each object \(A\) in \(\mathbf{C}\) an object \(F(A)\) in \(\mathbf{D}\); and each map \(f: A \rightarrow B\) in \(\mathbf{C}\) a map \(F(f): F(A) \rightarrow F(B)\) in \(\mathbf{D}\), such that \(F\left(1_{A}\right)=1_{F(A)}\) for each object \(A\) in \(\mathbf{C}\); and \(F\left(g{ }^{\circ} \mathbf{C} f\right)=F(g) \circ_{\mathbf{D}} F(f)\) for all maps \(f: A \rightarrow B\) and \(g: B \rightarrow C\), where \({ }^{\circ} \mathbf{C}\) and \(\circ_{\mathbf{D}}\) are compositions in categories \(\mathbf{C}\) and \(\mathbf{D}\).

Two examples of functors, used in the construction of products, are the diagonal and product functors. The diagonal functor \(\Delta: \mathbf{C} \rightarrow \mathbf{C} \times \mathbf{C} ; A \mapsto(A, A), f \mapsto(f, f)\) sends objects and arrows to pairs of objects and arrows. The product functor \(\Pi: \mathbf{C} \times \mathbf{C} \rightarrow \mathbf{C} ;(A, B) \mapsto A \times B,(f, g) \mapsto f \times g\) send pairs of objects and arrows to their respective products.
Definition (Universal morphism/construction). Given a functor \(F: \mathbf{A} \rightarrow \mathbf{C}\) and an object \(Y \in|\mathbf{C}|\), a universal morphism from \(F\) to \(Y\) is a pair \((A, \phi)\) where \(A\) is an object in \(\mathbf{A}\), and \(\phi\) is a morphism in \(\mathbf{C}\), such that for every object \(Z \in|\mathbf{A}|\) and every morphism \(f: F(Z) \rightarrow Y\), there exists a unique morphism \(u: Z \rightarrow A\), such that \(\phi \circ F(u)=f\), as indicated in commutative diagram


A universal construction is a universal morphism, or its dual a couniversal morphism, obtained by reversing the directions of the arrows in the definition of universal morphism.

A product \(\left(A \times B,\left(p_{1}, p_{2}\right)\right)\) is an instance of a universal morphism, hence a universal construction, and indicated in commutative diagram

where functor \(F: \mathbf{A} \rightarrow \mathbf{C}\) in the definition of universal morphism is the diagonal functor, and product object \(A \times B\) is obtained by application of the product functor to object \((A, B)\). Compare Diagram 6 with Diagram 1.

\title{
Spontaneous Analogy by Piggybacking on a Perceptual System
}

\author{
Marc Pickett \\ NRC/NRL Postdoctoral Fellow \\ Washington, DC 20375 \\ marc.pickett.ctr@nrl.navy.mil
}

\author{
David W. Aha \\ Navy Center for Applied Research in Artificial Intelligence Naval Research Laboratory (Code 5510); Washington, DC 20375 \\ david.aha@nrl.navy.mil
}

\begin{abstract}
Most computational models of analogy assume they are given a delineated source domain and often a specified target domain. These systems do not address how analogs can be isolated from large domains and spontaneously retrieved from long-term memory, a process we call spontaneous analogy. We present a system that represents relational structures as feature bags. Using this representation, our system leverages perceptual algorithms to automatically create an ontology of relational structures and to efficiently retrieve analogs for new relational structures from long-term memory. We provide a demonstration of our approach that takes a set of unsegmented stories, constructs an ontology of analogical schemas (corresponding to plot devices), and uses this ontology to efficiently find analogs within new stories, yielding significant time-savings over linear analog retrieval at a small accuracy cost.
\end{abstract}

\section*{1 Spontaneous Analogy}

In our day-to-day experience, we often generate analogies spontaneously (Wharton, Holyoak, \& Lange, 1996; Clement, 1987). That is, with no explicit prodding, we conjure up analogs to aspects of our current situation. For example, while reading a story, we may recognize a plot device that is analogous to one used in another story that we read long ago. The shared plot device may be a small part of each story, it is usually not explicitly delineated for us or presented in isolation from the rest of the story, and we may recognize the analogy of the plot device even if the general plots of the two stories are not analogous. Somehow, we segment out the plot device and retrieve the analog \({ }^{1}\) from another story in longdormant memory. Spontaneous analogy is the process of efficiently retrieving an analog from long-term memory given an unsegmented source domain such that part of the source shares structural similarity with the analog, though they might not share surface similarity. This process differs from standard models of analogy, which are given a delineated source concept, and often a target concept. Given a pair of analogs, analogical mapping is relatively straightforward. The more difficult problem is finding the analogs to begin with. As Chalmers, French, and Hofstadter (1992) argue "when the program's discovery of the correspondences between the two situations is a direct result of its being explicitly given the appropriate structures to work with, its victory in finding the analogy becomes somewhat hollow".

\footnotetext{
\({ }^{1}\) In our terminology, an analog is substructure of a domain that is structurally similar to a substructure of another domain, and an analogical schema is a generalization of an analog. For example, an input domain might be the entire story of Romeo \& Juliet, an analog would be the part of the story where Romeo kills Tybalt, who killed Romeo's friend, Mercutio (like in Hamlet where Hamlet kills Claudius, who killed Hamlet's father), and an analogical schema would be the generalized plot device of a "revenge killing".
}


Figure 1: An analog of Analogical Mapping vs. Spontaneous Analogy. In Analogical Mapping (a), we are given an explicit source and target, free from interfering context. In spontaneous analogy (b), the analogs are spontaneously retrieved from long-term memory.

The process of spontaneous analogy shares some properties with low-level perception, as exemplified in Figure 1. Within seconds of being presented with a visual image of a pterodactyl flying over a canyon, one can typically describe the image using the word "pterodactyl", even if one has had no special explicit recent priming for this concept, indeed even if one has not consciously thought about pterodactyls for several years. For us to produce the word "pterodactyl", we must segment the pterodactyl from the canyon and retrieve the "pterodactyl" concept from the thousands of concepts stored in memory. We must have learned the "pterodactyl" concept to begin with from unsegmented images. This perceptual process is robust to noise: The pterodactyl in the image could be partially occluded, ill-lit, oddly colored, or even drawn as a cartoon, and we are still able to correctly identify this shape (to a certain point). Likewise, many details of the plot devices from the above story example could be altered or obfuscated, but this analogy would degrade gracefully.

Our primary technical contribution in this paper is an algorithm called Spontol \({ }^{2}\) that solves the problem of spontaneous analogy: efficient parsing, storage, and retrieval of analogs from long-term memory. That is, given a corpus of many large unsegmented relational structures, Spontol discovers analogical schemas that are useful for characterizing the corpus and efficiently retrieves analogs given a new structure. E.g., given a set of narratives in predicate form, Spontol discovers plot

\footnotetext{
\({ }^{2}\) Spontol is short for "spontaneous analogy using the Ontol ontology learning and inference algorithm".
}
devices and analogs between the stories. We know of no prior work that scales to this task when the number of narratives and statements per narrative are both in the hundreds.

In the remainder of this paper, we describe related work (Section 2), give background on perceptual systems (Section 3), describe the Spontol algorithm, which transforms the problem of spontaneous analogy into a "perceptual" problem (Section 4), demonstrate Spontol's performance on a story database (Section 5), discuss implications and shortcomings of Spontol, and conclude (Section 6).

\section*{2 Related Work}

There has been earlier work on the problem of analogy in the absence of explicitly segmented domains. The COWARD system (Baldwin \& Goldstone, 2007) addresses this problem by searching for mappings within a large graph, essentially searching for isomorphic subgraphs. SUBDUE (Holder, Cook, \& Djoko, 1994) compresses large graphs by breaking them into repeated subgraphs, but is limited in that its output must be a strict hierarchy, and would be unable to discover the lattice structure of the concepts in Figure 2. Nauty (McKay, 1981) uses a number of heuristics to efficiently determine whether one graph is a subgraph of another, but this must be given source and target graphs to begin with. We can also apply The Chunker (described in Section 3) to feature bag graphlet kernels (Shervashidze, Vishwanathan, Petri, Mehlhorn, \& Borgwardt, 2009), which are related to Spontol's transform \(T\) in that both represent partial graphs, but this earlier work applies only for cases where there is one kind of entity, one kind of relation, and only binary relations, while Spontol works for multiple kinds of entities and relations, including relations of large arity.

The MAC phase of MAC/FAC (Forbus, Gentner, \& Law, 1995) bears some relation to our spontaneous analog retrieval. MAC uses vectors of content, such as the number of nodes and edges in a graph, as a heuristic for analog retrieval. However, in cases where the subgraph in question is a part of a much larger graph, the heuristics that MAC uses are drowned out by the larger graph. Likewise, ARCS (Thagard, Holyoak, Nelson, \& Gochfeld, 1990) also assumes that analogs have been delineated (i.e., it matches an entire source domain, rather than a substructure). SEQL (Kuehne, Forbus, Gentner, \& Quinn, 2000) generalizes relational concepts, but doesn't build a hierarchical ontology of analogical schemas.

There has been some work on representing structures as feature vectors. For example, Holographic Reduced Representations have been used to implement Vector Symbolic Architectures in which there is a correlation between vector overlap and structural similarity (Gayler, Levy, \& Bod, 2009). This work is limited in that it requires vectors of length 10,000 to represent very small graphs ( \(\leq 10\) nodes), and only represents binary relations of a single type, so this approach is not directly extendable to relational structures such as the stories in our demonstration. This is also a limitation for the system proposed by Rachkovskij, Kussul, and Baidyk (2012).

Both these systems are also limited in that they are unable to exploit partial analogical schemas. That is, a partial overlap in these systems' vectors does not correspond to a common subgraph in the corresponding structures. These systems stand in contrast to Spontol, which is able to represent larger structures and efficiently find common substructures.

\section*{3 Background: Perceptual Systems}

Spontol transforms relational structures into feature bags so that their surface similarity corresponds to the structural similarity of the relational structures. After Spontol has made this transformation, the problem of spontaneous analogy is reduced to the problem of feature overlap, and any of several existing "perceptual" systems can be used to find and exploit patterns in feature vectors. Our implementation of Spontol uses a model inspired by the human sensory cortices (auditory, visual, tactile) called Ontol (Pickett, 2011). Ontol is a pair of algorithms, both of which are given "sensor" inputs (fixed-length, real-valued non-negative vectors). The first algorithm constructs an ontology that concisely encodes the inputs. For example, given a set of vectors representing visual windows from natural images, Ontol produces a feature hierarchy loosely modeled on that seen in the visual cortex. The second algorithm takes as input an ontology (produced by the first algorithm) and a new vector, and parses the vector. That is, it produces as output the new vector encoded in the higher-level features of the ontology. In addition to "bottomup" parsing, the second algorithm also makes "top-down" predictions about any unspecified values in the vector.

Ontol is ignorant of the modality of its input. That is, Ontol is given no information about what sensory organ is producing its inputs. Because of this ignorance, we are able to leverage Ontol to find patterns in abstract "sensory" inputs that are actually encodings of relational structures.

\section*{Ontology Learning}

Ontol's ontology formation algorithm, called The Chunker, seeks to find concepts (or chunks) that allow for concise characterization of vectors. Since chunks themselves are vectors, The Chunker is applied recursively to create an ontology. In essence, this algorithm is similar to the recursive block pursuit algorithm described by Si and Zhu (2011) in that both search for large frequently occurring sets of features. The Chunker differs in that it allows for multiple inheritance, while recursive block pursuit creates only strict tree structures. In Section 4, we show the importance of this property for finding multiple analogical schemas within a single relational structure. For simplicity, we describe the discrete binary version of The Chunker algorithm (chunk \((B)\), which takes as input a set \(B\) of feature bags and produces an ontology \(\Omega\) ) provided by Pickett (2011), but this can be modified for continuous vectors. In this version, each vector is treated as a set, with a value of 1 for feature \(f\) signifying inclusion of \(f\) in the set, and a value of 0 signifying exclusion.

The Chunker searches for intersections among existing feature bags and proposes these as candidates for new concepts.

Each candidate is evaluated by how much it would compress the ontology, then the best candidate is selected and added to the set of feature bags, and the process is repeated until no candidates are found that further reduce the description length of the ontology. Figure 2 shows the ontology constructed by this algorithm when applied to an animal dataset, where the "sensory percepts" are features for each animal".


Figure 2: The Zoo Ontology with some instances. Instances are individual animals shown on the left, and base features are on the right. Black nodes in the middle correspond to higher-level features. The concept that corresponds to "fish" is marked. Inhibitory links are shown as dark circles.

\section*{Parsing and Prediction}

Given an ontology and a new instance, Ontol's parse \((b, \Omega)\) algorithm characterizes the feature bag instance \(b\) using the higher-level features in the ontology \(\Omega\). For example, given a new animal (a goldfish) that doesn't breathe, has fins, has no feathers, and is domestic, Ontol will parse the animal as an instance of the fish concept, with the exception that it is domestic. If Ontol is given no other information about the animal, it will also perform top-down inference, and unfold the fish concept to predict that the new instance has eggs, no hair, has a tail, etc.. This latter step is called "top-down prediction". Ontol searches for the parse that minimizes the description length of the instance. In our goldfish example, the "raw" description of the goldfish consists of 4 elements, while the "compressed" description has only 2 elements.

\footnotetext{
\({ }^{3}\) A full description and implementation of The Chunker, as well as source code for our demonstration of Spontol can be downloaded at http://marcpickett.com/src/analogyDemo.tgz.
}

Although the parsing problem is NP-complete, a single bottom-up pass can be performed in logarithmic time (Pickett, 2011). Importantly, Ontol examines only a small subset of the concepts and instances while parsing. This means that, when judging concept similarity, Ontol does not need to compare each of its \(n\) nodes. This property is important for spontaneous analog retrieval (described below).

\section*{4 Analogy as Perception}

We now describe a method for transforming relational structures into sparse feature vectors (or feature bags) such that the problem of analog retrieval is reduced to the problem of percept parsing. An example of this process is shown for the Sour Grapes fable in Figure 5. For this process, we rely on a transform \(T\) (described below) that takes a small relational structure and converts it into a feature bag (exemplified in Figure 5(c)). The size of relational structure is limited for \(T\) because \(T\) 's runtime is quadratic in the size of the structure. We view this limitation as acceptable because people generally cannot keep all the details of an entire lengthy novel (or all the workings of a car engine) in working memory. Generally, people focus on some aspect of the novel, or some abstracted summary of the novel (or engine). Therefore, we break each large relational structure into multiple overlapping windows. A window is a small set of connected statements, where two statements are connected if they share at least one argument. Spontol exploits a principle akin to one used by the HMax model of the visual cortex (Riesenhuber \& Poggio, 1999): as the number of windows for a relational structure increases, the probability decreases that another structure has the same windows without being isomorphic to the first.

The process for building an ontology of analogical schemas from large relational structures, called Spontol-Build, is described in Figure 3. This algorithm extracts numWindows windows from each large relational structure and transforms them into feature bags (exemplified in Figure 5(d)) and chunks these feature bags to create an ontology of windows called windowOntology. Spontol-Build then re-encodes the windows by parsing them using this ontology, and re-encodes the larger structures (from which the windows came) as a feature bag of the parsed windows. Finally, Spontol-Build runs another pass of chunking on the re-encoded structures to generate the schema ontology.

The process of spontaneous analog retrieval, called Spontol-Retrieve, is given in Figure 4. When given a new relational structure \(s\), we encode \(s\) by extracting windows from it, parsing these using the windowOntology, then parsing the feature bag representation using the schemaOntology. This yields a set of schemas that are contained in \(s\).

\section*{Transforming Small Relational Structures}

Here, we describe an operation \(T\), which transforms a (small) relational structure into a feature bag. In our demonstration, we assume that the relational structure is described in predicate logic, but our approach is not limited to this representation. We consider a relational structure to be a

Figure 3: Spontol's Ontology Learning Algorithm
// Creates an ontology of schemas given a set of structures \(S\). // numWindows is the number of windows to grab per structure. // windowSize is the number of statements per window.
define Spontol-Build (S, numWindows, windowSize)
// Randomly grab windows from each structure,
// and transform them into feature bag form.
foreach \(s \in S\); for \(i=1, \cdots\), numWindows let \(w_{s, i}=\) grabConnectedStatements ( \(s\), windowSize) add \(T\left(w_{s, i}\right)\) to allWindows
// Run The Chunker to generate the window ontology windowOntology \(=\) chunk (allWindows)
// Re-encode each structure using the reduced-size windows.
foreach \(s \in S\); for \(i=1, \cdots\), numWindows
add parse ( \(T\left(w_{s, i}\right)\), windowOntology) to bigWindows
// Run The Chunker to generate the schema ontology.
schemaOntology \(=\) chunk (bigWindows)
return schemaOntology, windowOntology

Figure 4: Spontol's Spontaneous Analogy Algorithm
```

// Finds analogical schemas for relational structure $s$.
// schemaOntology is the schema ontology.
// windowOntology is the window ontology.
// numWindows is the number of windows to grab per structure.
// windowSize is the number of statements per window.
define Spontol-Retrieve ( $s, \cdots$, windowSize)
// Randomly grab windows from $s$,
// transform them into feature bag form,
// and parse them using the window ontology.
for $i=1, \cdots$, numWindows
$w_{i}=$ grabConnectedStatements ( $s$, windowSize)
add parse $\left(T\left(w_{i}\right)\right.$, windowOntology) to bags
// Parse bag $_{s}$, the bag representation of $s$
relevantSchemas $=$ parse $\left(\right.$ bag $_{s}$, schemaOntology $)$
return relevantSchemas

```
set of relational statements, where each statement is either a relation (of fixed arity) with its arguments, or the special relation sameAs, which uses the syntax sameAs <name> (<relation> <arg1> <arg2> ...). The sameAs relation allows for statements about statements. E.g., the statements in Figure 5(b) encode (among other things) that "a fox \(d e\) cides that the grapes are sour".

Given a small relational structure \(s(\lesssim 10\) statements), \(T\) transforms \(s\) into a feature bag using a variant of conjunctive coding. That is, \(T\) breaks each statement into a set of roles and fillers. For example, the statement want Of3Fox Of3Grapes has two roles and fillers, namely the two arguments of the want relation. So \(T\) breaks this statement into want1=0f3Fox and want2=0f3Grapes, where want 2 means the 2 nd argument of want (i.e., the "wanted"). \(T\) then creates one large set of all the roles and their fillers. If there are multiple instances of a relation, it gives them an arbitrary lettering (e.g., wantB1=Of3Fox). \(T\) makes a special case for the sameAs relation. In this case, \(T\) uses a dot operator to replace the intermediate variable. For example, the statements sameAs \(f 35\) (decide Of3Fox f36) and sameAs f36 (sour Of3Grapes) would yield decide2.sour1=0f3Grapes. The dot operator allows \(T\) to encode nested statements (i.e., statements about state-
"A fox wanted some grapes, but could not get them. This caused him to decide that the grapes were sour, though the grapes weren't. Likewise, men often blame their failures on their circumstances, when the real reason is that they are incapable."
(a) English (for clarity)
\begin{tabular}{|l|l|l|}
\hline fox Of3Fox & cause m34 m33 & sameAs f36 (sour Of3Grapes) \\
false f36 & grapes Of3Grapes & sameAs f35 (decide Of3Fox f36) \\
cause f34 f35 & incapable Of3Men & sameAs f34 (get Of3Fox Of3Grapes) \\
false f34 & decide Of3Fox f36 & sameAs m34 (incapable Of3Men) \\
men Of3Men & sameAs m33 (fail Of3Men) & blameFor Of3Men concCircum m33 \\
fail Of3Men & want Of3Fox Of3Grapes & circumstances concCircum \\
\hline
\end{tabular}
(b) Predicate Form (Spontol's actual input)

(c) Transforming a Window

(d) Many Transformed Windows

Figure 5: Transforming the Sour Grapes Story. We show the transformation of Sour Grapes from predicate form to feature bag form. For clarity, we show an English paraphrase of the story (a), though the input to Spontol has already been encoded in the predicate form shown in (b), which shows the story as a set of 18 statements. In (c), we show a window \(w\) from the story and its feature bag transform \(T(w)\). Finally, the story is represented as many transformed windows (d).
ments). Given a set of roles and fillers, \(T\) then chains the fillers to get filler equalities. For example, if we have that decide1=0f3Fox and want1=0f3Fox, then chaining gives us decide1=want1. Chaining is essential for recognizing structural similarity between relational structures, and allows us to side-step a criticism of conjunctive coding and tensor products: that the code for wantB1=Of3Fox may have no overlap with the code for want1=0f3Fox (Hummel et al., 2004). Chaining introduces the code for want \(\mathrm{B} 1=\) want 1 , which makes the similarity apparent when searching for analogs (these "chained" features are a core difference between MAC's content vectors and our feature bags). After
chaining the roles and fillers, \(T\) treats each of these role-filler bindings as an atomic feature. Note that, when we treat roles and fillers as atomic features, Ontol doesn't recognize overlap among feature bags unless they share exactly the same feature. For example, the atomic feature wantB1=0f3Fox has no more resemblance to want \(1=0 f 3\) Fox for Ontol than it does for any other feature. Also note that the ordering of the roles in each feature is arbitrary but consistent ( \(T\) uses reverse alphabetical order), so there is a men \(1=\) incapable1 feature, but not an incapable1=men1 feature. The left side of Figure 5(c) shows a window taken from the sour grapes story from Figure 5(b). On the right side is the feature bag transform of this set of 6 statements, consisting of 11 atoms.

\section*{5 Demonstration}

We applied Spontol to a database of 126 stories provided by Thagard et al. (1990). These include 100 fables and 26 plays all encoded in a predicate format, where each story is a set of unsorted statements. An example story in predicate form is shown in Figure 5(b). Note that although the predicates and arguments have English names, our algorithm treats all these as gensyms except for the special sameAs relation. In this encoding, the smallest story had 5 statements, while the largest had 124 statements, with an average of 39.5 statements.

We ran Spontol-Build on these stories using numWindows \(=\) 100 and windowSize \(=20\) which produced an ontology of stories, part of which is shown in Figure 6. In this figure we see a "Double Suicide" analogical schema found in both Romeo \& Juliet and in Julius Caesar. In the former, Romeo thinks that Juliet is dead, which causes him to kill himself. Juliet, who is actually alive, finds that Romeo has died, which causes her to kill herself. Likewise, in Julius Caesar, Cassius kills himself after hearing of Titinius's death. Titinius, who is actually alive, sees Cassius's corpse, and kills himself. The largest schema found (in terms of number of outgoing edges) was that shared by Romeo \& Juliet and West Side Story, which are both stories about lovers from rival groups. The latter doesn't inherit from the Double Suicide schema because Maria (the analog of Juliet), doesn't die in the story, and, Tony (Romeo's analog) meets his death by murder, not suicide. Some of the schemas found were quite general. For example, the oval on the lower right with 6 incoming edges and 3 outgoing edges corresponds to the schema of "a single event has two significant effects". And the oval above the Double Suicide oval corresponds to the schema of "killing to avenge another killing".

Spontol-Retrieve uses this schema ontology to efficiently retrieve schemas for a new story, which can be used to make inferences about the new story in a manner analogous to the "goldfish" example from Section 3. To evaluate the efficiency of Spontol-Retrieve, we randomly split our story dataset into 100 training stories and 26 testing stories. We then used an ontology learned from the training set, and measured the number of comparisons needed to retrieve schemas (during parse) for the testing set. We compare this approach to MAC/FAC, which, during the MAC phase, visits each of


Figure 6: Part of the ontology Spontol learned from the story dataset. As in the Zoo Ontology in Figure 2, black ovals represent higher level concepts. The "raw" features (corresponding to the white ovals in Figure 2) are omitted due to space limitations. Instead, we show the outgoing edges from each black oval. While in the Zoo Ontology, the higher level concepts correspond to shared surface features, in this figure, high level concepts correspond to shared structural features, or analogical schemas. For example, the highlighted oval on the right represents a Double Suicide schema, which happens in both Romeo \& Juliet and in Julius Caesar.
the 100 training stories. Whereas MAC/FAC returns entire stories, Spontol-Retrieve returns analogical schemas (just as a visual system would return a generic "pterodactyl" concept rather than specific instances of pterodactyls). For comparison, we modify Spontol-Retrieve to return the set of instances that inherit from relevantSchemas, rather than just the schemas.

Table 1: Speed/Accuracy Comparison of Spontol
\begin{tabular}{|lcc|}
\hline & Accuracy & Average \# Comparisons \\
MAC/FAC & \(100.00 \% \pm .00 \%\) & \(100.00 \pm .00\) \\
Spontol & \(95.45 \% \pm .62 \%\) & \(15.43 \pm .20\) \\
\hline
\end{tabular}

Results are shown in Table 1, averaged over 100 trials. We show accuracy (and standard error) for both systems mea-
sured as the percentage of stories correctly retrieved, where a story was determined to be correct if it was retrieved by MAC/FAC. Spontol effectively improves on a linear (in the number of structures) case-by-case comparison to an "indexed" logarithmic-time look-up at a slight cost of accuracy. Therefore, Spontol requires orders of magnitude fewer comparisons than MAC/FAC, or any linear look-up algorithm (for a survey, see (Rachkovskij et al., 2012)). For larger datasets, we hypothesize that these differences will be even more pronounced. Although each comparison by both MAC and Spontol-Retrieve is a fast vector operation, for very large datasets (e.g., \(10^{9}\) relational structures), even a linear number of vector operations becomes impractical. In future work, we will test these systems on a broader range of relational datasets to help elucidate the conditions under which Spontol yields high accuracy and very-low retrieval cost.

\section*{6 Conclusion}

The chief contribution of this paper is a demonstration of a system, Spontol, that is able to solve the problem of spontaneous analogy. That is, we have demonstrated how Spontol can efficiently store and retrieve analogs without the need of human delineation of schemas.

Our representation also offers a new solution for the binding problem for long-term (static) memory that allows for efficient analog retrieval in the absence of explicitly segmented domains. The binding problem asks how we can meaningfully represent bindings between roles and fillers. Most solutions to the binding problem in connectionism do so in terms of temporal synchronicity (e.g., LISA (Hummel \& Holyoak, 2005)). Temporal synchronicity only works for knowledge in working memory, and these models typically address storage in long-term memory by relying on some form of conjunctive coding or tensor products. Though these systems fail to address how relational structures can be efficiently retrieved from long-term memory, we hypothesize that a workingmemory system, such as LISA, is necessary for the "chaining" process on which our system relies.

Spontol may offer evidence in support of a uniform "substrate" of intelligence (Mountcastle, 1978). In particular, we've shown how a system that was designed to process perceptual data (Ontol) can be leveraged to process "symbolic" data (i.e., relational structures). This may provide insight into how species capable of higher-order cognition might have evolved from species capable of only low-level perception.

Although Spontol addresses some outstanding problems in Computational Analogy, there is still ample room for future work. Our implementation for characterizing a relational structure as a set of windows might not scale well to very large structures without some modifications. An open problem is how windows might be managed in a sensible way. Spontol currently uses "bags of windows" for medium-sized structures. We propose extending Spontol by allowing hierarchies of progressively higher-order bags to represent larger structures (e.g., bags of bags of bags of windows).

\section*{References}

Baldwin, D., \& Goldstone, R. L. (2007). Finding Analogies Within Systems: Constraints on Unsegmented Matching. In WrkShp. on Analogies: Integrating Multiple Cog. Abilities.
Chalmers, D. J., French, R. M., \& Hofstadter, D. (1992). High-level Perception, Representation, and Analogy: A Critique of Artificial Intelligence Methodology. J. Exp. Theor. Artif. Intell., 4(3), 185-211.
Clement, J. (1987). Generation of Spontaneous Analogies by Students Solving Science Problems. In Thinking Across Cultures (p. 303-308).
Forbus, K., Gentner, D., \& Law, K. (1995). MAC/FAC: A Model of Similarity-based Retrieval. Cog. Sci., 19(2).
Gayler, R., Levy, S., \& Bod, R. (2009). A Distributed Basis for Analogical Mapping. In New Frontiers in Analogy Research; Proc. of 2nd Intern. Analogy Conf. (Vol. 9).
Holder, L., Cook, D., \& Djoko, S. (1994). Substructure Discovery in the SUBDUE System. In Workshop on Knowledge Discovery in Databases.
Hummel, J. E., \& Holyoak, K. J. (2005, June). Relational Reasoning in a Neurally Plausible Cognitive Architecture: An Overview of the LISA Project. Current Directions in Psychological Science, 14(3), 153-157.
Hummel, J. E., Holyoak, K. J., Green, C., Doumas, L. A. A., Devnich, D., Kittur, A., et al. (2004). A Solution to the Binding Problem for Compositional Connectionism. In AAAI Fall Symp. on Comp. Connectionism in Cog. Sci.
Kuehne, S., Forbus, K., Gentner, D., \& Quinn, B. (2000). SEQL: Category Learning as Progressive Abstraction Using Structure Mapping. In Proceedings of the 22nd Annual Meeting of the Cognitive Science Society.
McKay, B. (1981). Practical Graph Isomorphism. Congressus Numerantium, 30, 45-87.
Mountcastle, V. (1978). An Organizing Principle for Cerebral Function: The Unit Model and the Distributed System.
Pickett, M. (2011). Towards Relational Concept Formation From Undifferentiated Sensor Data. Doctoral dissertation, University of Maryland Baltimore County.
Rachkovskij, D., Kussul, E., \& Baidyk, T. (2012). Building a World Model with Structure-Sensitive Sparse Binary Distributed Representations. Bio. Inspired Cog. Archs..
Riesenhuber, M., \& Poggio, T. (1999). Hierarchical Models of Object Recognition in Cortex. Nature Neuroscience, 2(11), 1019-1025.
Shervashidze, N., Vishwanathan, S., Petri, T., Mehlhorn, K., \& Borgwardt, K. (2009). Efficient Graphlet Kernels for Large Graph Comparison. In Int. Conf. on AI \& Stats.
Si, Z., \& Zhu, S. (2011). Unsupervised Learning of Stochastic AND-OR Templates for Object Modeling. In IEEE Int. Conf. on Computer Vision Workshops (pp. 648-655).
Thagard, P., Holyoak, K., Nelson, G., \& Gochfeld, D. (1990). Analog Retrieval by Constraint Satisfaction. Artificial Intelligence, 46(3), 259-310.
Wharton, C., Holyoak, K., \& Lange, T. (1996). Remote Analogical Reminding. Memory \& Cognition, 24(5), 629-643.

\title{
Type-token representations in conceptual representation
}

\author{
Sandeep Prasada (sprasada@hunter.cuny.edu) \\ Department of Psychology, Hunter College, CUNY \\ New York, NY 10065 USA
}

\begin{abstract}
Concepts of kinds of things (e.g. DOG), have the dual function of specifying how to think about indefinitely many things as well as providing the means for thinking about a single abstract kind which is constituted by indefinitely many instances. In this talk, I sketch a theory of conceptual representation that places this dual function of concepts at its core. The theory is shown to provide a natural way of capturing four key characteristics of the ways in which we think about kinds and instances of kinds. These characteristics are not accounted for by standard approaches to conceptual representation. In the final section of the paper, I consider how the phenomena discussed in this paper may be accommodated by current approaches to conceptual representation.
\end{abstract}

Keywords: concepts; type-token representations; kind representations; generic knowledge.

Concepts are the mechanisms by which we think about things. For example, the concept DOG provides the means for thinking about indefinitely many distinct things as being a given kind of thing (e.g. dogs). Furthermore, it also provides the means for thinking about the kind dog itself. This dual function of specifying how to think about indefinitely many things as well as providing the means for thinking about a single abstract kind which is constituted by indefinitely many instances has generally escaped the focus of most research on conceptual representation. In this paper, I sketch a theory of conceptual representation that places this dual function of concepts at its core. The theory is shown to provide a natural way of capturing four key characteristics of the ways in which we think about kinds and instances of kinds. These characteristics are not currently accounted for by standard approaches to conceptual representation, however, they constitute empirical phenomena that any adequate theory of conceptual representations would have to handle. In the final section of the paper, I consider how the phenomena discussed in this paper may be accommodated by current approaches to conceptual representation.

\section*{Mechanism for thinking about instances of kinds and kinds.}

Following Prasada \& Dillingham (2009) and Prasada (2012), the theory developed here proposes that concepts for kinds of things are represented via a generative type-token mechanism such as (1) which is capable of generating
indefinitely many representations (2) each of which provide the means for thinking about distinct instances of that kind.
(1)
(2) \(\mathrm{K}_{1} \quad \mathrm{~K}_{2} \quad \mathrm{~K}_{3} \quad \ldots\)

The mechanisms in (1) and (2) highlight the close connection between the mechanisms needed to make generic and non-generic reference. The mechanisms needed for thinking about instances of kinds (2) are generated by the mechanism needed for thinking about kinds (1) which makes implicit reference to instances of kinds. As such, the mechanisms that underlie generic and non-generic reference are intrinsically related and thus one may expect the ability to make generic and non-generic reference to be closely tied in development. This is, in fact, the case. Recent research suggests that children appear to use noun phrases generically and non-generically from a very early age (Pappas \& Gelman, 1998; Gelman \& Tardif, 1998; Goldin-Meadow, Mylander \& Gelman, 2005; Gelman, Goetz, Sarnecka \& Flukes, 2008).

\section*{Instances of kinds may be qualitatively identical.}

The mechanisms in (1) and (2) also highlight a fundamental characteristic of the manner in which we think about instances of kinds. Instances of a given kind need not be qualitatively distinct in any way. They need only be numerically distinct. Thus we are perfectly capable of thinking about qualitatively identical instances of kinds. This is, of course, easier to do for some kinds (e.g. paperclips) than others (e.g. dogs), nevertheless, it is possible for any kind of thing and our conceptual mechanisms must support such thoughts. This characteristic of how we think about instances of kinds is captured by the fact that the representations in (2) differ only in their indicies which have a purely indexical function and have no intrinsic descriptive content.

\section*{Distinct kinds cannot be qualitatively identical.}

Turning our attention to the representation of kinds (1), a natural question is whether kind representations may also be distinguished merely by an index (e.g. \(\mathrm{K}_{\mathrm{i}}^{1} \& \mathrm{~K}_{\mathrm{i}}{ }_{\mathrm{i}}\) ). A moment's reflection makes it clear that it makes no sense to speak of qualitatively identical and merely numerically distinct kinds. Kinds, unlike instances of kinds, must be qualitatively distinct. This means that the representation of kinds in (1) must be augmented with a component that has
descriptive content that characterizes the kind and provides the basis for individuating kinds.

Given that kinds are constituted by indefinitely many instances, characterizing a kind must involve representing a connection between the kind and properties that is understood to be non-accidental and thus extendible to indefinitely many instances that have yet to be encountered (Goodman, 1955). Prasada, Khemlani, Leslie \& Gucksberg (2013) provide evidence that our conceptual systems distinguish at least three types of non-accidental connections between kinds and properties (principled connections, statistical connections, and causal connections) that provide at least three ways of characterizing kinds. Nevertheless, kinds cannot be generally be individuated in terms of properties that involve merely statistical connections to the kind because such properties are extrinsic to the kind. Furthermore, kinds cannot be generally be individuated in terms of properties that involve causal connections to kinds because such connections are only possible for material kinds.

Principled connections, on the other hand, involve properties that instances of kinds are understood to have by virtue of their being the kinds of things they are (e.g. having four legs for dogs) (Prasada \& Dillingham, 2006). As such, it is possible to identify properties that have a principled connection to a kind for any kind of thing. Properties that have a principled connection to a kind ( \(k\)-properties in Prasada \& Dillingham's (2006) terminology) are (i) properties whose presence in instances of a kind receive formal explanations -- explanations by reference to the kind of thing something is (e.g. Fido has four legs because he is a dog), (ii) properties for which we have normative expectations such that instances of the kind that lack kproperties are judged to be defective or incomplete, and (iii) properties that are generally expected to be present in instances of the kind (Prasada \& Dillingham, 2006, 2009). Kproperties differ from definitional properties in important ways. Unlike definitional properties which are necessarily present in all instances of a kind, k-properties need not be present in all instances of the kind (e.g. there are many dogs that do not have four legs). Relatedly, the representation of definitional properties are understood as specifying necessary conditions for the use of the kind concept of which they are a part whereas k-property representations do not specify conditions for the use of the kind concept with which they are connected. Instead, k-property representations specify properties that are understood to be lawfully related to being that kind of thing such that the presence of k-properties in instances of a kind is understood to be due the things being the kinds of things they are. In fact, k-properties are understood to be aspects of being the given kind of thing (e.g. having four legs is one aspect of being a dog) and thus are represented via a formal part-whole relation between the kind and property (Prasada \& Dillingham, 2009). These findings lead Prasada \& Dillingham (2009) to revise the mechanism for representing kinds in the following manner.
(3) \(K_{i} \sim<a 1, a 2 \ldots\)

In this representation (3), the mechanism for representing a kind projects \({ }^{1}\) an aspect structure by means of which the properties that have a principled connection to the kind (kproperties) can be represented as aspects of being that kind of thing. \({ }^{2}\)

The mechanism for representing kinds (3) is now seen to have a descriptive component in terms of which kinds are (qualitatively) individuated via distinct sets of \(k\)-properties. It should be noted that though kinds cannot be merely numerically distinct and thus kind representations cannot be distinguished merely by indexical representations devoid of any descriptive content (e.g. \(\mathrm{K}_{\mathrm{i}}^{1} \& \mathrm{~K}_{\mathrm{i}}^{2}\) ), it is possible that kinds may be distinguished indexically in addition to being distinguished qualitatively via their k-properties. In fact, it is possible that developmentally it is sometimes the case that two kinds may initially be distinguished only indexically via their names with the expectation that they will additionally differ in yet to be discovered ways (Xu, 2012). The theory being developed here suggests that in such cases, the expectation is that the kinds also differ in their k-properties.

\section*{Modes of existence of instances of kinds and kinds.}

The mechanisms in (3) and (2) also highlight an important difference between the ways in which we think about instances of kinds and kinds. At any given time, there are some definite number of instances of a kind that actually exist or have actually existed. In addition, there are indefinitely many instances of the kind that do not actually exist, but exist only potentially. Potentially existing instances of a kind differ from actually existing instances of a kind only in their mode of existence. No qualitative differences distinguish actually existing and potentially existing instances of a kind. Given that this is the case, it should be immediately evident that there cannot be potentially existing kinds as this would require that the putatively potentially existing kinds be qualitatively identical to actually existing kinds, but as we saw above, distinct kinds cannot be qualitatively identical.

\section*{Relation to existing theories of conceptual representation.}

The phenomena pertaining to the manner in which we think about kinds and instances of kinds discussed in this paper are generally not addressed by standard approaches to conceptual representation. How might standard approaches to conceptual representation respond to these phenomena? To answer this question, it will be helpful to summarize the

\footnotetext{
\({ }^{1}\) I use the \(\sim<\) symbol to mean "projects". Prasada \& Dillingham (2009) and Prasada (2012) use -> for the same notion, however, that is easily confused with implication.
\({ }^{2}\) As the details of how principled connections are represented via such a mechanism are not pertinent here, I skip over them. See Prasada \& Dillingham (2009) for details.
}
generative type-token mechanism theory proposed in the present paper and situate standard approaches with respect to it. As illustrated in Figure 1, a kind concept such as DOG is, at its core, a generative mechanism which has the capacity to generate indefinitely many representations each of which represent an instance of the kind and which are distinguished from one another numerically via an index devoid of any intrinsic descriptive content. The generative mechanism itself implicitly contains the indefinitely many instances and thus provides the means for thinking about the kind itself. Finally, the kind projects an aspect structure which provides the means for representing the properties that have a principled connection to the kind (k-properties) as aspects of being that kind of thing. This component of the mechanism serves to qualitatively distinguish the kind from other kinds.


Figure 1. Schematic summary of the generative type-token mechanism that provides the means for thinking about kinds and instances of kinds.

Given this characterization, it becomes evident that the definitional, prototype and theory approaches to conceptual representation can best be understood as corresponding to the aspect structure component of the approach developed here in that they provide different accounts of how kinds are qualitatively individuated from one another. As such, these approaches, as currently formulated, lack the mechanisms for thinking and reasoning about kinds and instances of kinds. The exemplar approach to conceptual representation, on the other hand, involves representations that correspond most closely to those that represent instances of kinds. The exemplar approach lacks any mechanisms for representing kinds or information for individuating kinds, by design.

We are now in a position to consider how standard approaches may accommodate the phenomena discussed in the present paper which are handled naturally by the generative type-token mechanism approach. One possibility is that each approach may use the tools within their framework to incorporate these phenomena. While this may be possible, it is hard to see what relevant resources exist within the definitional, prototype, exemplar, or theory theories to form representation of instances of kinds and kinds such that they have the characteristics discussed here. A second possibility which would be available to the
definitional, prototype and theory approaches would be to graft the formal mechanisms developed here onto one's favorite approach. This avenue would be in line with suggestions in the philosophical and psychological favoring hybrid theories of conceptual representation (e.g. Genone \& Lombrozo, 2012; Keil, 1989; McNorgan, Kotack, Meehan, \& McRae, 2007; Sloman, Love \& Ahn, 1998). This option is presumably not available to the exemplar approach which eschews summary representations and thus kind representations and components of kind representations that individuate kinds could presumably not be grafted onto exemplar representations. A third response may be to see the present work as potentially increasing the diversity of conceptual representations available thus providing further fuel to theories favoring conceptual pluralism and heterogeneity (Dove, 2009; Machery, 2009; Weiskopf, 2009).

The final possibility I'd like to consider and endorse is that the present theory contains key elements from each of the standard approaches and thus it may be possible to develop a theory of conceptual representation that has many of the advantages of hybrid theories, but in a manner that is more principled and organic. The present theory involves the generation of representations of instances of kinds (2) and thus provides a natural way of accounting for phenomena that are best handled by exemplar theories of concepts.
Furthermore, the k-properties that individuate kinds do not specify conditions for the application of the kind concept. Instead, they specify properties that instances of the kind are expected to have in virtue of their being the kinds of things they are but may be lacking in instances for reasons other than their being the kinds of things they are (Prasada \& Dillingham, 2006, 2009). As such, the k-properties that characterize a kind are like properties in prototype representations in that they are typically present in members of the kind, but need not be.

The link to the explanation-based theory approach to concepts is evident in the fact that the k-properties are represented as involving a lawful (non-accidental) link between the kind and properties. The mechanism in (3) represents principled connections between kinds and properties by representing the property as an aspect of being that kind of thing. As such, it supports a formal mode of explanation whereby the presence of k-properties in instances of kinds may be explained by reference to the kind of thing something is (Prasada \& Dillingham, 2009). The mechanism in (3) also provides the basis for psychological essentialism by identifying the properties that are understood to be caused by the essence. As the essence is typically not known (Medin \& Ortony, 1989; Gelman, 2003), the properties caused by the essence cannot be identified by the essence and must be identified in another manner. The k-properties represented in (3) thus potentially provide the basis for psychological essentialism.

Furthermore, as mentioned above, kinds may also be characterized via the causal and statistical connections they have to properties (Prasada, Khemlani, Leslie \& Gucksberg,
2013). Though the addition of causal and statistical connections between kinds and properties to the mechanism in (3) is not required by our ability to think about kinds and instances of kinds (as is the addition of principled connections), their addition is necessary for capturing the range of ways in which we can characterize kinds. As such, their addition is principled and motivated within the present approach and allows the theory to capture further phenomena that are typically accounted for within the prototype and theory views of concepts.

Though much more needs to be said, I hope it is already evident that a theory which is centered on the generative type-token mechanisms needed for thinking about kinds and instances of kinds can not only capture the phenomena discussed in the present paper, but potentially provides a principled and organic way of capturing the key characteristics of current approaches to conceptual representation.

\section*{References}

Dove, G. (2009). Beyond perceptual symbols: A call for representational pluralism. Cognition, 110, 412-431.
Gelman, S.A., \& Tardif, T. Z. (1998). Generic noun phrases in English and Mandarin: An examination of child-directed speech. Cognition, 66, 215-248.
Gelman, S. (2003). The essential child: Origins of essentialism in everyday thought. London: Oxford University Press.
Gelman, S. A., Goetz, P. J., Sarnecka, B. S., \& Flukes, J. (2008). Generic language in parent-child conversations. Language Learning and Development, 4, 1-31.
Genone, J. \& Lombrozo, T. (2012): Concept possession, experimental semantics, and hybrid theories of reference, Philosophical Psychology, 25:5, 717-742.
Goldin-Meadow, S., Gelman, S.A. \& Mylander, C. (2005). Expressing generic concepts with and without a language model. Cognition, 96, 109-126.
Goodman, N. (1955). Fact, fiction \& forecast. Cambridge, MA. Harvard University Press.
Keil, F. C. (1989). Concepts, kinds, and cognitive development. Cambridge, MA: MIT Press.
Machery, E. (2009). Doing without concepts. Oxford, UK: Oxford University Press.
McNorgan, C., Kotack, R.A., Meehan, D.C., \& McRae, K. (2007). Feature-feature causal relations and statistical cooccurrences in object concepts. Memory \& Cognition, 35, 418-431.
Medin, D., \& Ortony, A. (1989). Psychological essentialism. In S. Vosniadou, \& A. Ortony (Eds.), Similarity and analogical reasoning. Cambridge, MA: Cambridge University Press.
Pappas, A. \& Gelman, S.A. (1998). Generic noun phrases in mother-child conversations. Journal of Child Language, 25, 19-33.
Prasada, S. (2012). Mechanisms for characterizing kinds and classes. Recherches Linguistiques de Vincennes, 41, 45-62.

Prasada, S., \& Dillingham, E. (2006). Principled and statistical connections in common sense conception. Cognition, 99 (1).
Prasada, S., \& Dillingham, E. (2009). Representation of principled connections: A window onto the formal aspect of common sense conception. Cognitive Science, 33, 401448.

Prasada, S., Khemlani, S., Leslie, S-J., \& Gucksberg, S. (2013). Conceptual distinctions amongst generics. Cognition, 126, 405-422.
Sloman, S. A., Love, B. C., \& Ahn, W. (1998). Feature centrality and conceptual coherence. Cognitive Science, 22, 189-228.
Weiskopf, D. A. (2009). The plurality of concepts. Synthese, 169, 145-173.

Xu, F. (2012). Count nouns, sortal concepts, and the nature of early words. . In F.J. Pelletier, ed. Kinds, Things \& Stuff. New Directions in Cognitive Science, v. 12, Oxford Univ. Press, NY.

\title{
Category Change in the Absence of Falsifying Feedback
}

\author{
Jared T. Ramsburg (jramsb2@uic.edu) \\ Stellan Ohlsson (stellan@uic.edu)
}

\author{
University of Illinois at Chicago \\ Department of Psychology (MC 285) \\ 1007 West Harrison Street \\ Chicago, Illinois, 60607-7137
}

\begin{abstract}
Many conceptual change theories posit that change occurs when the learner becomes dissatisfied with the current conception (Ohlsson, 2011; Strike \& Posner, 1992). A necessary component of dissatisfaction is falsifying feedback. The present experiments investigate whether participants exposed to a novel method for eliminating the ability to directly falsify a misconception will still be able to recategorize compared to participants that can directly falsify. The results suggest that direct falsification of a misconception is not necessary for recategorization, and that direct falsification may slow the learning process. Implications are discussed.
\end{abstract}

Keywords: Learning, recategorization, feedback, nonmonotonic change.

\section*{Introduction}

Both common sense and past research have assumed that conceptual change in particular and non-monotonic cognitive change in general is driven by a person's dissatisfaction with his or her current conception (Ohlsson, 2011). Dissatisfaction is in turn caused by falsifying information and experiences that are inconsistent with the current conception. Without falsification a person would presumably lack motivation to change (Chi, 2005; Chi, 2008; Chi \& Brem, 2009; Gopnik \& Wellman, 2012; Slotta \& Chi, 2006; Strike \& Posner, 1982, 1992). But once dissatisfaction has set in, the learner is ready to search for an alternative conception (Elio \& Pelletier, 1997; Strike \& Posner, 1982, 1992; Chi \& Ohlsson, 2005; Ozdemir \& Clark, 2007). The theme of falsification first became dominant in the history of science via the works of Karl Popper and Thomas Kuhn, but it has since spread to all aspects of knowledge change.

For example, Strike and Posner's (1982) claimed that students in a science classroom must be dissatisfied with their current conception before they are ready to learn a new conception. Moreover, dissatisfaction must surpass the threshold at which accommodation supersedes assimilation. The threshold is surpassed by the accretion of falsifying pieces of information that accumulate until the discrepancy cannot be ignored.

Similarly, the Theory-Theory posits that the knowledge revision process takes place when dissatisfaction with the current conception reaches an individual's threshold for
conceptual change in the course of cognitive development (Gopnik \& Wellman, 2012).
As a final example, the Categorical Shift Theory describes conceptual change as a process that requires one to abandon or reject prior misconceptions via the recognition of differences between two or more general categories (Chi, 2005; Chi \& Brem, 2009). Failure to filter information through an existing knowledge base leads to dissatisfaction with the current conception. Dissatisfaction leads to a search for an alternative knowledge structure capable of accommodating the new information.
In short, these and other theories of cognitive change assume that dissatisfaction is a necessary prerequisite for cognitive change in children, students, and both lay adults and scientists. However, both common sense and psychological research agree that although people respond to falsifying information by trying to reduce the cognitive dissonance it causes, they tend to process the falsifying information in such a way as to minimize its impact on current knowledge (Ohlsson, 2011). If so, why should we believe that falsifying information is a necessary component of conceptual change?
In contrast to the theories mentioned above, the Resubsumption Theory claims that conceptual change can occur even in the absence of falsification of a person's current conception. This is possible when the learner possesses two alternative theories that apply to the same case or phenomenon. Change from one theory to the other occurs through competitive evaluation on the basis of cognitive utility rather than truth or falsity (Ohlsson, 2009). Competitive evaluation triggers a change by revealing that the alternative theory is more applicable in a given instance.

In the current study, we used the re-categorization paradigm (Cosejo, Oesterreich \& Ohlsson, 2009) to create a situation in which the participants needed to change a newly learned definition a category into a different definition of the same category in the absence of information that falsified the latter. Specifically, the participants learned how to categorize a novel set of stimuli through the standard procedure used in countless categorization experiments (Ashby \& Maddox, 2005): view a potential category member, judge whether it is a member, receive feedback on the judgment, and go to the next trial. Once the participants showed that they had mastered the category, the category was changed without warning. To succeed, the learner had
to re-learn the category, i.e., learn a new definition of it, and consequently, a different way of categorizing the relevant stimuli. The particular version of recategorization that we used in this study presented stimuli that mimicked a science-learning scenario. Images of fictitious alien bacteria were categorized with respect to their resistance to atmospheric oxygen; see details in the Method section.

The present study used the recategorization paradigm to investigate whether falsification is necessary for a learner to recategorize. All participants were given both supportive and falsifying feedback on their categorization judgments during the initial phase of the study. We refer to this as initial learning, and the category definition learned as the initial category or the 'misconception'. After learning the initial category, the participants were exposed to one of two feedback conditions during the second phase of the experiment. We refer to the second phase as the target learning, and the new category definition acquired in this phase as the target category.

The participants in the complete feedback condition received both confirmatory and falsifying feedback (the complete condition). The participants in the second feedback condition were presented with stimulus items that had been altered in such way that the initial category, once acquired, could not be directly falsified (the confirmation only condition). This was accomplished by deleting crucial features from the stimuli; see Method section for details. However, they received the same information required to learn the target category as the participants in the complete condition. In short, the purpose was to compare recategorization in the presence and absence of falsifying feedback.

\section*{Predictions}

There are three potential outcomes of this experiment. We could find that having complete feedback (i.e., both confirmation and falsification) yields the most efficient categorical change. Alternatively, we could find that the absence of falsification has no effect on recategorization, that is, learners need confirmation to learn, not falsification. Finally, we could find that falsification is not necessary, but harmful. That is, the presence of falsification might hinder recategorization, perhaps by creating cognitive conflicts that trigger defensive processing mechanisms (Ohlsson, 2011). The latter might use up cognitive resources that are needed for learning.

We have specific quantitative predictions regarding these outcomes. The predictions relate to different measures of performance. The first measure examines overall success, that is, do the groups learn the target when compared to chance. Specifically, it is hypothesized that the complete condition (i.e., those with both types of feedback) will perform better than chance because the combination of confirmatory feedback and falsifying feedback will allow the learner to adopt the target category. The confirmatory condition is hypothesized to perform better than chance because of the availability of confirmatory feedback.

The first measure (i.e., overall success) is examined between groups. That is, are there differences between groups in their ability to learn the target category? It is hypothesized that there will be no difference in target learning between the confirmatory and complete condition. This is expected because the use of confirmatory feedback will allow learners to adopt the target category (for both confirmatory and complete conditions). No differences between the confirmation and complete conditions will demonstrate that falsification is not necessary for recategorization to occur.

The second measure examines how quickly the groups can recategorize. There are three different scenarios that could occur for speed of categorization that will answer the question regarding what type of feedback appears to be the most effective for increasing speed of categorization. The first scenario would have complete learning faster than confirmatory. This would demonstrate that having both confirmation and falsification could result in faster learning compared to confirmation without falsification. That is, falsification is beneficial for increasing the speed of categorical change compared to not having the ability to directly falsify the misconception.
The second scenario would be that no difference exists between complete and confirmatory only conditions. This would suggest that the presence or absence of falsification has no effect on categorical change so long as confirmatory feedback is available.
The third scenario would show that speed of learning is faster for confirmatory compared to complete. This type of outcome would demonstrate that falsification might not be necessary for categorical change, but that it might hinder categorical change as evidenced by the complete condition underperforming compared to the confirmatory condition.

\section*{Method}

\section*{Participants}

One hundred twenty introductory psychology students participated in the study for course credit. Random assignment yielded 66 participants in the complete condition and 54 participants in the confirmatory condition.

\section*{Design}

The study was a between-participants design with two conditions (Complete and Confirmatory).

\section*{Materials}

The materials consisted of 128 fictional bacteria images including some that were incomplete, i.e., some features were deleted (see Figure 1). The bacteria have six different parts that have different binary attributes resulting in 64 complete variants: Nuclei (grey or black), Headbulbs (three or none), Ribosomes (bent or straight), Tail Cilia (present or absent), Cell Membrane (singular or double), and Cytoplasm (white or grey).


Figure 1. Example bacterium with parts labeled.
Additionally, some images were incomplete, that is, some images would not show the nuclei and some would not show the tail (see Figure 2). The images were presented on a computer screen via E-Prime software; see (www.pstnet.com/products/E-Prime/default/).


Figure 2. Bacteria with and without nuclei shown.

\section*{Procedure}

Phase 1: Misconception Learning. Participants first learned to categorize whether an alien bacteria was oxygen resistant based on feedback that supported the misconception feature (i.e., black nuclei) over the course of five training blocks of 16 trials each. Each training block was balanced to include in randomized order six images that contained the misconception, six images that contained the target, two images that contained neither, and two images that contained both the misconception and the target. After five training blocks, unbeknownst to the participants the feature that determines oxygen resistance changed to bent ribosomes (i.e., the target).

Phase 2: Target Learning. Participants had five target training blocks of 16 randomized trials to learn that bent ribosomes determined oxygen resistance. The target training had two different experimental conditions.

Condition 1: Complete Stimuli. This condition consisted of stimuli that were similar to what participants had already used for classification. Each training block was balanced to include in randomized order six images that contained the misconception, six images that contained the target, two images that contained neither, and two images that contained both the misconception and the target. All parts of the bacteria were visible on the screen allowing participants to falsify their prior categorization in favor of a new categorization supported by the computers feedback. For example, in phase 1, the participant learned that black nuclei are responsible for oxygen resistance. In phase 2, the participant was then confronted with an image containing black nuclei with feedback stating that the bacteria was not oxygen resistant. This feedback should allow the learner to negate the prior conception. Moreover, when the learner is confronted with an image that does not have a black
nucleus, but is shown to be oxygen resistant the learner should logically conclude that another part of the bacteria is responsible for oxygen resistance.

Condition 2: Incomplete Stimuli. This condition contained no stimuli that could be used to directly falsify the misconception. Specifically, bacteria images containing the dark nuclei with straight ribosomes were not shown for any trial. However, there were stimuli that did not show the nuclei, resulting in an inability of the learner to directly falsify the initial category. Each training block was balanced to include in randomized order six images that did not display the misconception, six images that fit the target category, two images that fit neither category, and two images that fit both the misconception and the target categories. The purpose of the latter was to make the learning situation somewhat more challenging by introducing a small amount of noise into the information the participants received.

\section*{Procedure}

Participants were seated in separate cubicles. Each participant was instructed to first participate in a training session, which consisted of a series of PowerPoint slides outlining how one can sort a variety of objects into different categories. The training session ended with participants categorizing stick figures based on their features. When participants finished with the initial training activity, they were instructed to participate in the more challenging bacteria paradigm.

Participants read the instructions for the task on the computer screen and asked questions if needed. Participants were given a script stating that alien bacteria was recently discovered on a distant planet and that scientists needed to determine whether there were oxygen resistant variants of the bacteria. Participants were then asked to rate how important each feature was in determining oxygen resistance on a 7-point Likert scale from 1 (Not at all) to 7 (Extremely). After rating the features, participants went through a prompt that described the importance of determining which bacteria were oxygen resistant. Each participant was tasked with determining whether the pictured bacterium was oxygen resistant. Participants indicated their response via the keyboard. The following responses were acceptable: \(\mathrm{y}=\mathrm{yes}, \mathrm{n}=\mathrm{no}, \mathrm{d}=\) don't know. Participants would then receive immediate feedback from the computer either stating that the bacterium was or was not oxygen resistant. Participants were instructed to make as few errors as possible.

After completing all trials, participants were again asked how motivated they were to perform the task well and to rate the importance of different features in determining oxygen resistance on the same 7-point Likert scale as before. The participants keyed in an open-ended response about which features they thought determined oxygen resistance. They then went to the next screen which asked whether oxygen resistance was always determined the same way. Finally, participants answered demographic questions.

\section*{Results}

Thirty-eight participants in the complete condition and 36 participants in the incomplete condition met the criterion for inclusion in analyses (i.e., correctly classifying 14 of 16 alien bacteria in any of the initial five training blocks). The inclusion criterion was chosen as a way to insure that we tested participants who were successful in learning the misconception. We wanted to examine whether falsification is necessary for adopting a new method of categorization for the participants who succeeded in learning the initial misconception feature, not whether falsification is necessary to learn the target category from scratch.

\section*{Learning Misconception}

Our first analysis determined whether random assignment was effective at producing equivalent groups. In order to determine whether participants might differ in their ability to learn the misconception, we examined their performance on the first five blocks via a repeated measures analysis of variance (ANOVA) with Greenhouse-Geisser correction found a main effect for blocks, that is, regardless of condition, participants improved in performance from blocks 1 thru 5, \(F(2.32,166.84)=66.76, p<.001, \eta_{\text {partial }}^{2}=\) .653. There was no main effect of condition, \(F<1, \eta_{\text {partial }}^{2}=\) .008 nor did groups differ at rate of learning, \(F(2.32\), \(166.84)=1.09, n s ., \eta_{\text {partial }}^{2}=.307\). These results suggest that the groups were equivalent in their ability to learn the misconception.

\section*{Learning the Target}

Our next step was to assess whether the confirmatory condition learned the target in blocks 6 through 10. Performance of 14 out of 16 or greater on any of the blocks 6 through 10 was rated as successful learning of the target; we found that 29 of 36 (80.55\%) participants correctly learned the target category. Whereas, if participants maintained the misconception for all trials they would have resulted in 0 of 36 participants demonstrating that they learned the target.

Using a chi-squared goodness-of-fit test we measured overall target acquisition (i.e., in general did learning occur yes or no) against a more stringent probability (i.e., chance at 50\%). Specifically, the results revealed that the confirmatory condition's target acquisition was better than chance, \(\chi^{2}(36)=13.44, p<.001\). Similar results were found for the complete condition where 29 of 38 (76.31\%) participants learned the target, \(\chi^{2}(36)=9.00, p<.01\).

\section*{Differences between Groups for Target Learning}

We examined whether conditions differed in target acquisition via a chi-squared test-of-independence that showed that the groups did not differ in target acquisition, \(\chi^{2}\) (36) \(=2.90, p=.09\). This suggests that removing the ability to directly falsify the misconception does not hinder a learner's ability to adopt a new method of categorization.

Additionally, we examined potential differences in learning rate based on condition following the switch. That
is, we wanted see whether one group learned faster than the other. A repeated measure ANOVA with GreenhouseGeisser correction was used to determine whether there would be a difference in performance following the switch from the misconception to the target for blocks 6 through 10. The analysis revealed a main effect for blocks showing that participants improved with training, \(F(2.646,190.477)\) \(=75.01, p<.001, \eta_{\text {partial }}^{2}=.671\), and a main effect for condition showing that the confirmatory condition performed better than the direct condition, \(F(1,72)=7.60\), \(p<.01, \eta_{\text {partial }}^{2}=.096\). The interaction was significant, rate of learning was faster for the confirmatory condition than the complete condition, \(F(2.646,190.477)=5.21, p<.01\), \(\eta_{\text {partial }}^{2}=.146\). These results suggest that the confirmatory condition may result in faster learning of a new conception (see Table 1).

Table 1: The means and (standard deviations) for percentage correct for blocks 6-10.
\begin{tabular}{lccccc}
\hline & \begin{tabular}{c} 
Block \\
6
\end{tabular} & \begin{tabular}{c} 
Block \\
7
\end{tabular} & \begin{tabular}{c} 
Block \\
8
\end{tabular} & \begin{tabular}{c} 
Block \\
9
\end{tabular} & \begin{tabular}{c} 
Block \\
10
\end{tabular} \\
\hline Complete & 40.63 & 64.64 & 72.86 & 81.58 & 86.06 \\
& \((16.61)\) & \((27.43)\) & \((28.11)\) & \((24.49)\) & \((19.89)\) \\
Confirmatory & 63.72 & 77.78 & 83.51 & 88.54 & 88.02 \\
& \((16.89)\) & \((19.33)\) & \((20.27)\) & \((16.06)\) & \((20.51)\)
\end{tabular}

\section*{Overview of Response Type by Condition}

In Figure 3, responses that are misconception consistent (MCR) or target consistent (TCR) separated by condition are shown by training block. MCRs were responding no on target bacteria and TCRs were responding yes on target bacteria. These response types are independent from each other because of the don't know response option. The figure shows how response tendencies changed when the feedback was altered to support the target within and between conditions.


Figure 3. Percentage consistent with response type by condition.

\section*{Discussion of Experiment 1}

The results of Experiment 1 suggest that the incomplete, confirmatory feedback only condition might initially speed
up learning of a new conception in comparison to the complete condition. We investigated whether participants would adopt a new conception after the switch and whether rate of learning would vary based on condition. We found that the participants in both conditions adopted the new conception. However, participants in the complete condition learned at a slower rate than those in the confirmatory condition. We propose that this difference was due to the need for participants in the complete condition to make sense of conflicting information. The sense making absorbed cognitive resources that otherwise would have been available for learning the target category, slowing down the re-categorization process.

\section*{Experiment 2}

In Experiment 2, we attempted to replicate the findings of Experiment 1 with the addition of two learning aids. Learning aids were included in an effort to reduce the number of participants that are eliminated from analysis for failing to learn the misconception. The first learning aid was included in the prompt that participants read before engaging in the categorization process. The learning aid suggested that parts within the cell body may be influential in determining oxygen resistance (This statement is true as both misconception and target features are within the cell body). We assumed that the inclusion of this statement might focus search for what promotes oxygen resistance to the interior of the bacteria. The second learning aid was a handout that showed an image of the bacteria with parts labeled (see figure 1) as well as a list of the possible variants of each feature. This was meant to serve as a working memory aid.

Finally, the handout included a statement that "some of the images of the bacteria that you will see may be INCOMPLETE, that is, all bacteria have the 6 parts described, but some parts may not be visible." This aspect of the handout was included in an effort to refute claims that participants may be viewing bacteria that do not show the dark nuclei as being bacteria without nuclei, which could result in a different interpretation of the stimuli by the participants.

\section*{Method}

\section*{Participants}

Sixty-one introductory psychology students participated in the study for course credit. Random assignment yielded 30 participants in the complete condition and 31 participants in the confirmatory condition.

\section*{Design \& Procedure}

The same procedure as in Experiment 1 was used in Experiment 2, with the addition of the two learning aids (i.e., the hint in the prompt and the handout).

\section*{Results}

Twenty-one participants in the complete condition and 21 participants in the confirmatory condition met the criterion
for inclusion in analyses (i.e., correctly classifying 14 of 16 alien bacteria in any of the first five blocks).

\section*{Learning Misconception}

Our first analysis sought to determine whether participants might differ in their ability to learn the misconception via a repeated-measures ANOVA where we examined percentage correct per block for the first five blocks. We found a main effect for blocks, that is, regardless of condition, participants improved in performance from blocks 1 thru 5, \(F(4,160)=\) 55.60, \(p<.001, \eta_{\text {partial }}^{2}=.582\). There was no main effect of condition, \(F(1,40)=1.69, p=.201, \eta_{\text {partial }}^{2}=.041\) nor did groups differ at rate of learning, \(F(4,160)=1.19, p=.319\), \(\eta_{\text {partial }}^{2}=.029\). These results suggest that the groups were equivalent in ability to learn the misconception.

\section*{Learning the Target}

Our next step was to assess whether the confirmatory condition learned the target in blocks 6 through 10. Performance of 14 out of 16 or greater on any of the blocks 6 through 10 was rated as successful learning of the target. We found that 15 of 21 ( \(71.43 \%\) ) participants correctly learned the target category. Using a chi-squared goodness-of-fit test the results revealed that the confirmatory condition's target acquisition was better than chance, \(\chi^{2}(21)\) \(=3.86, p=.05\). Alternatively, results for the complete condition where 14 of 21 (66.67\%) participants learned the target, their target acquisition was not better than chance, \(\chi^{2}\) (21) \(=2.33, p=.127\).

\section*{Differences between Groups for Target Learning}

We examined whether conditions differed in learning the target via a chi-squared test-of-independence that showed that the groups did not differ in learning the target, \(\chi^{2}(42)=\) \(.11, p=.739\). This replicates the finding from experiment 1 that the ability to directly falsify a misconception is not necessary for learning the new conception.

Given the relatively small sample size for Experiment 2 and the likelihood of differences occurring in earlier blocks we opted to conduct a series of t-tests on target learning blocks 6 through 10 instead of a repeated measure ANOVA, which might fail to differentiate the effect. The results of the t-tests revealed that participants in the confirmation condition performed better than the complete condition for block \(6, t(40)=4.56, p<.001\) and marginally better on block 7 for a one-tailed t-test, \(t(40)=1.67, p=.051\). There were no differences between the groups for blocks 8,9 , and 10, \(t<1\) (see Table 2).

Table 2: The means and (standard deviations) for percentage correct for blocks 6-10.
\begin{tabular}{lrrrrr}
\hline & \multicolumn{2}{c}{ Block } \\
6 & \multicolumn{1}{c}{\begin{tabular}{c} 
Block \\
7
\end{tabular}} & \multicolumn{1}{c}{\begin{tabular}{c} 
Block \\
8
\end{tabular}} & \multicolumn{1}{c}{\begin{tabular}{c} 
Block \\
9
\end{tabular}} & \multicolumn{1}{c}{\begin{tabular}{c} 
Block \\
10
\end{tabular}} \\
\hline Complete & 44.94 & 68.75 & 78.87 & 80.06 & 83.93 \\
& \((20.31)\) & \((30.94)\) & \((29.01)\) & \((28.55)\) & \((26.34)\) \\
Confirmatory & 68.75 & 82.14 & 78.87 & 85.42 & 85.12 \\
& \((12.66)\) & \((19.89)\) & \((16.11)\) & \((19.8)\) & \((23.67)\)
\end{tabular}

\section*{Discussion}

In the present study, we examined whether participants randomly assigned to receive one of two types of stimuli differed in their ability to falsify an initially acquired category. The present findings provide modest support that there may be instances in which falsification is not only unnecessary for overriding a prior conception, but might actually be harmful. In both Experiments 1 and 2 we found that participants who could not directly falsify the misconception adopted the target conception in fewer trials compared to those participants who could falsify the misconception directly.

If replicated, our demonstrating that falsification is not necessary for categorical change could have multiple implications. For instance, theories of conceptual change that posit the necessity of dissatisfaction might themselves need revision. In addition, instruction in the classroom for scientific topics known to require knowledge revision has found that direct refutation is not necessarily effective at promoting change (Vosniadou, 1994; Vosniadou \& Verschaffel, 2004), but perhaps novel development of another ontological structure could without refutation be developed and then integrated into the learning environment. Further investigation would be required in order to determine the most effective ways to improve conceptual change processing amongst students.

The present work should be viewed in consideration to its experimental controls, which might limit external validity. Specifically, the use of novel stimuli may not promote the same types of recategorical processes as stimuli that hold some greater individual meaning. Additionally, the population used in the study (university students) cannot be expected to adequately represent all types of learners. Furthermore, the learning processes observed in this experiment were of short duration. In many situations that require non-monotonic cognitive change, a direct verbal statement of the target concept is available but it was not part of our experimental procedure. Finally, we point out that conceptual change in real life usually involves a system of interrelated concepts rather than a single concept.

Future research might explore how different types of stimuli might influence recategorical change. Moreover, studies that mimic a classroom environment might also offer insights into what processes might bring about conceptual change. Additionally, studies that are able to use multiple daily training sessions and then attempt to recategorize might help in the understanding of temporal exposure and its influences on recategorization.

\section*{References}

Ashby, F. G., \& Maddox, W. T. (2005). Human category learning. Annual Review of Psychology, 56, 149-178.
Butler, D. L., \& Winne, P. H. (1995). Feedback and selfregulated learning: A theoretical synthesis. Review of Educational Research, 65, 245-281.
Chi, M. T. H. (2005). Common sense conceptions of emergent processes: Why some misconceptions are robust.

Journal of the Learning Sciences, 14, 161-199.
Chi, M. T. H. (2008). Three types of conceptual change: belief revision, mental model transformation, and categorical shift. In S. Vosniadou (Ed.), International handbook of research on conceptual change (pp. 6182). New York: Routledge.

Chi, M. T. H., \& Brem, S. K. (2009). Contrasting Ohlsson's Resubsumption Theory with Chi's Categorical Shift Theory. Educational Psychologist, 44, 58-63.
Chi, M. T. H., \& Ohlsson, S. (2005 ). Complex declarative learning. In K. J. Holyoak, \& R. G. Morrison, (Eds.), The Cambridge handbook of thinking and reasoning (pp. 371-399). Cambridge: Cambridge University Press.
Cosejo, D. G., Oesterreich, J., \& Ohlsson, S. (2009). Recategorization: Restructuring in categorization In N.A. Taatgen \& H. van Rijn (Eds.), Proceedings of the 31th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.
Elio, R. \& Pelletier, F. J. (1997). Belief change as propositional update. Cognitive Science, 21, 419-460.
Gopnik, A., \& Wellman, H. M. (2012, May 14). Reconstructing Constructivism: Causal Models, Bayesian Learning Mechanisms, and the Theory Theory. Psychological Bulletin. Advance online publication. doi: 10.1037/a0028044
Ohlsson, S. (2011). Deep learning: How the mind overrides experience. Cambridge, UK: Cambridge University Press.
Ohlsson, S. (2009). Resubsumption: A possible mechanism for conceptual change and belief revision. Educational Psychologist, 44, 20-40.
Özdemir, G., \& Clark, D. B. (2007). An overview of conceptual change theories. Eurasia Journal of Mathematics, Science \& Technology Education, 3, 351-361.
Slotta, J. D. \& Chi, M. T. H. (2006). Helping students understand challenging topics in science through ontology training. Cognition and Instruction, 24, 261289.

Strike, K.A., \& Posner, G.J. (1982). Conceptual change and science teaching. European Journal of Science Education, 4, 231-240.
Strike, K.A., \& Posner, G.J. (1992). A revisionist theory of conceptual change. In R. Duschl \& R. Hamilton (eds.), Philosophy of science, cognitive psychology, and educational theory and practice (pp. 147-176). Albany, NY: SUNY Press.
Vosniadou, S. (1994). Capturing and modeling the process of conceptual change. Learning and Instruction, 4, 4569.

Vosniadou, S. \& Verschaffel, L. (2004). Extending the conceptual change approach to mathematics learning and teaching. Learning and Instruction, 14, 445-451.

\title{
The 'universal' structure of name grammars \\ And the impact of social engineering on the evolution of natural information systems
}

\author{
Michael Ramscar Asha Halima Smith Melody Dye Richard Futrell Peter Hendrix Harald Baayen Rebecca Starr \\ University of Tubingen, Stanford University, Indiana University, Massachusetts Institute of Technology, Carnegie Mellon
}

\begin{abstract}
Proper name systems provide individuals with personal identifiers, and convey social and hereditary information. We identify a common information structure in the name grammars of the world's languages, which makes this complex information processing task manageable, and evaluate the impact that the re-engineering of naming practices for legal and political purposes has had on the communicative and psychological properties of these socially evolved systems. While East-Asian naming systems have been largely unaffected by state legislation, legal interference has transformed Western naming practices, making individual names harder to process and remember. Further, the structural collapse of Western naming systems has not affected all parts of society equally: In the US, it has had a disproportionate impact on those sections of society that are least successful in economic and social terms. We consider the implications of these changes for name memory across the lifespan, and for future naming practices.
\end{abstract}

Keywords: Names; Learning; Memory; Information Theory

\section*{What's in a name?}

Naming is unique to our species and central to our lives. Names are the primary linguistic means by which we discriminate individuals from their peers, and they are an integral part of our identities. Names also play an important social role: they carry hereditary information that helps regulate marriage between relatives and the transference and distribution of property, as well as fostering group identities that bring cohesion to the conduct of social enterprises such as agriculture, industry, statecraft, and war.

While names for individuals appear to be as old as civilization itself, some functions of names are recent developments (Scott, 1998). Henry VIII decreed that English marital births be recorded under the surname of the father in the 14th Century, but children could, and regularly did adopt different surnames. Hereditary family names only became universal in the UK with the establishment of Her Majesty's Register Office in 1836 (Matthews, 1967). Naming conventions in the Netherlands were formalized by statute in 1811 (Van Poppel, Bloothooft, Gerritzen \& Verduin, 1999); in Korea, naming practices were regulated in 1812 (Nahm, 1988), the same year that a Prussian edict granted citizenship to Jews in return for the adoption of fixed patronyms (Scott, Tehranian \& Mathias, 2002).

Despite their personal and social importance, names are uniquely difficult to learn and remember (Cohen, \& Burke, 1993; Valentine, Brennen \& Bredart, 1996). Names produce most naturally occurring tip of the tongue states (TOTswhere one cannot produce a word one is sure one knows) (Rastle \& Burke, 1996; Griffin, 2010); patients with cognitive impairments show greater decrements in name
retrieval than for other knowledge (Yasuda, Nakamura \& Beckman, 2010); and the recall of names appears to be disproportionately impaired in old age. Indeed, many older adults consider deteriorating name memory to be the most disturbing cognitive problem they face (Lovelace \& Twohig, 1990).

Here, we identify a common information structure in the name grammars of the world's languages, and reveal the impact that regulating names for legal and political purposes has had on their memorability as populations have grown in the wake of industrialization: while some name systems survive intact, legislation has had a detrimental effect on many name grammars, dramatically undermining their communicative efficiency.

\section*{Why names are different-and difficult}

While most nouns are generic-spoon, dog, ideapersonal names are sui generis: ideally, they uniquely discriminate individuals from their peers. While this could easily be achieved by giving each individual a unique label, this approach would massively increase linguistic complexity. By now, it would have generated a billion extra English words. At points in speech where a name could occur in this kind of a system, entropy-a formal, information theoretic measure of uncertainty (Shannon, 1948)-would vastly exceed anything heretofore encountered. Because entropy peaks accurately predict difficulties in the production and comprehension of speech (McDonald \& Shillcock, 2001; Clark \& Wasow, 1998), this would cause far more processing problems than actual English names, which are already far more taxing than other vocabulary items.

Thus, while 'one-name-per-person' would eliminate residual uncertainty about the identity of named individuals, it would maximize processing demands in doing so. Realistically, the psychological cost of such a system is too high: Given the highly social, interconnected nature of human life, and our finite information processing capacities, one-name-per-person would prove unworkable as societies developed beyond small groups. Unsurprisingly, no major language has a naming system remotely close to it (Alford, 1987).

\section*{A "Universal" Grammar for Names}

Although some fine-grained details differ, all the world's major languages have evolved the same solution to the challenges names pose: instead of using unique labels, individual identifiers are formed from hierarchically structured naming tokens. Name grammars enable large sets of identifiers to be constructed out of much smaller sets of
naming words, assigning individuals relatively unique identifiers, while avoiding the outrageous peaks in entropy that would result from a one-name-per-person system.
\begin{tabular}{lll} 
Family Name & Clan / Generation Name & Given Name \\
Least Uncertain & More Uncertain & Most Uncertain \\
Baek & Seung & Ki
\end{tabular}

For example, in the traditional Sinosphere (Matisoff, 1990) naming system used in Chinese-speaking countries and Korea (Kwang-Sook, 2003), names comprise 3 elements: 1) a small number of family names, 2) a clan or generation name (Martin, 2006), and 3) a given name, fairly specific to the individual. (Here, the first parts of physicist Seung Baek Ki’s name mean, "a Baek from Suwon." (Kiet, Baek, Jeong \& Kim, 2003)). Elements are distributed in these sets in a highly efficient, Zipfian manner (Baek, Kiet \& Kim, 2007)2: there are only around 250 Korean family names, three of which are common to around \(50 \%\) of the population ( \(20 \%\) of all Koreans are called Kim). Because of this, names act as efficient hierarchical decision trees: each name element increases the degree to which an individual is identified, while minimizing the entropy at the point each element is encountered (Figure 1).

These distributions have been stable for centuries. As Korea's population grew post-industrialization (Zipf, 1965), the peak entropy of Korean names barely changed. Name grammars in Chinese-speaking countries have developed along similar trajectories.


Figure 1. Hierarchical branching minimizes entropy (which quantifies the uncertainty produced when a number alternatives that need to be processed at any point) as a Korean name unfolds in time.

This efficient information structure is not unique to Sinosphere names: it is common to the native name grammars of all the world's major languages. For instance, traditionally, English names comprised a given name drawn from a relatively small set, optional middle name(s), and an identifier (in Modern English, a family name), drawn from a large set of personal characteristics, topographic and toponomic features, occupations and patronyms: e.g., John the Farmer, or John White Head. English names rarely contained a specific, unique identifier. Instead, individual identities were the sum of a name's parts: the 11th Century English name in (3) means, "Adam, a farmer from Ramscar" (Ramskir, 1973).
\begin{tabular}{lll} 
Given Name & Other Name & Identifier \\
Least Uncertain & More Uncertain & Most Uncertain \\
Adam & Hegger & de Romeskerre
\end{tabular}

What may be surprising from a modern perspective is that the distributional pattern of names in modern Korea would have been familiar to pre-industrial Europeans: In every 50year period from 1550 to 1799 , around \(50 \%\) of boys born in England were named William, John or Thomas, and 50\% of girls Elizabeth, Mary or Anne (Smith-Bannister, 1997), mirroring the distribution of Korean family names. Given names in other Western and Northern European languages were also distributed this way (Galbi, 2002; Lieberson \& Lynn, 2003; Bourin, 1994) even when patronymic conventions were employed (Williams, 1961). Historically, compact, stable Zipfian first name distributions were the norm: prior to industrialization, one-in-five English girls had the first name Mary, just as around one-in-five Korean girls today has the first name Kim. What caused some distributions to change, while others have remained as they were?


Figure 2. The hierarchical organization of modern English names. The proportion of surnames to given names reflects the fact that given names-which occur before surnames in Englishcontribute less to individual identities than surnames. The surnames depicted here represent over \(95 \%\) of the US population in the 2000 census, and the given names over \(95 \%\) of social security applications in the US in 2000.

\section*{The rise of nation states, and their impact on Western name grammars}

Ordinarily, nobody knows everybody. However, the development of centralized states created entities that actually did want to know everyone: for the purpose of taxation, conscription, etc. To facilitate this, states regulated names. In the Sinosphere, the burden of coding heredity fell on the first (least diverse) name element. However, in other parts of the world, the final, most diverse element was targeted. In English, idiosyncratic features specific to individuals - such as John the farmer, John with the white head - became fixed hereditary markers, such that bakers might be called Farmer, or redheads Whitehead. English name grammar retained its hierarchical structure (Figure 2). Consequently, as population growth accelerated in the 19th and 20th centuries (Figure 3), two changes occurred:
1. A larger, more diverse set of first names began to be used, increasing the peak signal entropy of names, making them harder to process and recall.
2. More people shared last names, increasing the likelihood of two people having the same name, increasing residual entropy about the individual identified by a name.

Both these changes had an impact on the efficiency of English names.

In the UK in 1801, 3 male names accounted for \(52 \%\) of male births and 3 female names 53\% of female births (the traditional distribution). By 1994, this had dropped to \(11 \%\) and \(9 \%\), respectively. In both instances, the relationship to population growth is strong: \(\mathrm{r} 2=.99 \& .97\). Similarly, while in \(1801,85 \%\) of males and \(82 \%\) of females received 1 of 10 names, by 1994 , this had dropped to \(28 \%\) and \(24 \%\); related to population growth, \(\mathrm{r} 2=.99\) and .97 . In \(1801,22 \%\) of males and \(24 \%\) of females in the UK received the most common first name for their sex; in 1994, it was \(4 \%\) and \(3 \%\) (both \(\mathrm{r} 2=.98\) ). Social Security card applications in the US between 1900-1999 show the same patterns of change in naming: a big decline in the number of babies being given the most frequent names, and an increase in the diversity of given names. (For all reported correlations, \(\mathrm{p}<0.001\) )
*The sample for this survey was taken from the U.S. Social Security Administration (http://www.ssa.gov/oact/babynames/).


Figure 3. The top panel illustrates the distribution by rank frequency of Korean family names in 2000. The bottom panel shows the change in the written frequencies of the 500 most common US male names as the United States population grew in the \(19^{\text {th }}\) and \(20^{\text {th }}\) Centuries \((1810-1990)\). The greater similarity in the distribution of Korean family names and American male names at the beginning of the \(19^{\text {th }}\) Century is apparent, as is the change thereafter.

An analysis of the 500 most common male names from 1810-1990 in the Corpus of Historical American English (400,000,000 words) indicates that at the beginning of the 19th century, US given name distributions were similar to Korean family names (Figure 3). While the distribution of Korean family names remained stable, the distribution of

American and British given names changed dramatically as populations grew.

\section*{Measuring the Effects of Change: A Tale of Two Congresses}

Mainland China has experienced massive population growth in the past two centuries, but its traditional Sinosphere naming system has changed little as a result (Yuan, 2002; Mountain, Wang, Du, Yuan, 1992). To illustrate how the changes to the distribution of American names have affected name efficiency, we compared two similarly-sized, naturally-occurring samples of names from each of these two countries: the Senators and Representatives of the 112th United States Congress ( \(\mathrm{n}=440\) ), and the members of several subcommittees of the National Committee of the Chinese People's Political Consultative Conference ( \(\mathrm{n}=431\) ).
*The samples for this survey were taken from the website of the United States Congress (http://www.house.gov/representatives/) and the website of the Chinese People's Political Consultative Conference (http://www.cppcc.gov.cn/)


Figure 4. The information challenge posed by names as their numbers grow. The y-axis plots perplexity, which describes the entropy ( x -axis) in a complex distribution in terms of a number of equally likely alternatives. The perplexity of US names is more than twice that of Chinese names in this sample.

The total entropy of both samples is almost identical (US Congress \(=8.78\) bits; CPPCC \(=8.75\) bits), because in each case, no members share names; accordingly, there is no residual entropy in either sample. There is, however, a marked disparity in the way information is distributed in the samples: the first elements in the CPPCC names contain only 5.84 bits of information (in information processing terms, this is equivalent to differentiating between around 60 equally likely name labels; Figure 4). By contrast, the given (first) names of US Congress members contain 7.39 bits (equivalent to processing around 170 equally likely labels). Both sets of names convey the same amount of overall information, but this information is more evenly distributed in Chinese names: US first name elements impose far greater information processing demands than their Chinese equivalents, and later US name elements are far more redundant (full form analysis-treating Mike / Mick as forms of Michael-reduced US name perplexity to 120 ,
suggesting that Congressional names would more memorable if nicknames were avoided).

\section*{Winners and Losers in the Decline of the US Naming System}

Although the efficiency of English names has declined in the past 200 years, the effect of these changes has not been felt in the same way across society. Traditionally, the perplexity of US female names was slightly greater than male names (Figure 5A), but the difference in perplexity between male and female names was relatively constant. However after 1950 this difference began to rise sharply (Figure 5B), increasing the difference in the amount of information that had to be processed in recalling, producing and comprehending female names as compared to male names.
\[
\begin{aligned}
& \text { The Perplexity Gap 1880-2010 }
\end{aligned}
\]

Figure 5. Top: The perplexity of male and female names with a count \(\geq 5\) in US social security applications at 5 year intervals from 1880 to 2010. Bottom: The difference in the perplexity of female names as compared to male names (female perplexity - male perplexity) across this period.

Given that the period since 1950 saw an increase in economic and social equality between males and females in the US (Fullerton, 1999), the close relationship (Figure 6A) between the growth in the perplexity difference between male and female names and the increasing percentage of females in the workforce is surprising, as is the fact that increases in the number of women working outside the home have coincided with an exponential increase in the degree to which female names are harder to process than male names. Figure 6B offers one possible explanation for this: the strong correlation between population size and female name diversity may indicate that parents actually take great care in naming their daughters, but that the
constraints that have been imposed on Western naming practices are distorting the intended effect of name choices.
Figure 6. Top: The proportion of men and women over age 16 in the US workforce (male - female percentage) plotted against the different perplexity of female and male names (female perplexity male perplexity) in the period 1950-1998 (data from the US Bureau of Labor Statistics \& US Social Security Administration).
Bottom: The increase in the US population at five-year intervals from 1880-2010 and the increase in the number of male and female names with a count \(\geq 5\) in US social security applications.


\section*{Nobody Knows My Name}

Women are not the only group in the US to have experienced a disproportionate decline in name efficiency. As has often been noted, African American parents systematically choose distinctive given names for their children (Lieberson \& Mikelson, 1995). The pattern of this trend is puzzling: although Black parents living in predominantly White communities choose distinctive names for their children, Black parents living in predominantly Black communities choose even more distinctive names. (If it were simply the case that parents choose distinctive names to signal affinity with the wider Black community, the opposite pattern might be expected; Fryer \& Levitt, 2004).

The question is: why? As we noted earlier, when surnames are fixed and a population expands, the residual entropy of names invariably rises. In such circumstances, parents may be more likely select diverse first names for their children to increase their name's overall uniqueness. Given the legacy of slavery (Dunaway, 2003), residual entropy is a particular problem for the Black community, where a smaller (less diverse) pool of surnames places more of the burden of providing uniquely identifying information
on first name elements. The residual entropy of Black surnames is considerably higher than for White surnames: the most frequent 500 US surnames convey just 2.5 bits of information about the White community, but over 4 bits of information about the Black community \((\mathrm{t}(499)=8.00\), \(\mathrm{p}<0.001\); Figure 7), meaning that surnames convey far less information about individuals within Black communities. Residual perplexity-which increases when more people share a surname-is twice as high for these names in Black communities than in White communities. Since people called Smith will be likely to give their children distinctive first names (because the surname Smith has high residual entropy), and since this likelihood will increase if a high proportion of their neighbors are also called Smith, the tendency of Black parents to choose more distinctive names for their children when their neighbors are Black than when they are White is not really puzzling at all: it simply reflects the desire of parents to provide each child with a unique identifier.


Figure 7. Top: The residual entropy of the 500 most common American surnames in the White- and African-American communities ( 2000 US Census). Bottom: The 500 most common White surnames and 500 most common Black surnames (high residual entropy \(=\) less information about individuals).

\section*{Names, age and memory}

Problems with remembering names represent the most disturbing aspect of aging for many people. The analyses reported here raise a question: does memory for names really decline, or are older adults confusing social change with personal change? After all, the changes to the distribution of Western names guarantees that name processing must have grown increasingly difficult over the last century. To provide an estimate of the effect this could have had on name memory, we examined the effect of
changing name distributions in a model that simulates human performance in lexical decision tasks.

The naive discriminative reading model (NDR; Baayen, Milin, Durdevic, Hendrix \& Marelli, 2011) is a two-layer network that takes letter unigrams and bigrams as cues, and learns to discriminate lexical targets as outcomes (e.g., 'hand', 'John') using the equilibrium equations (Danks, 2003) of the Rescorla-Wagner (Rescorla \& Wagner, 1972) learning rule. The model's output is entirely determined by its training corpus-it has no free parameters-and it captures a wide variety of empirical effects in reading (Baayen, 2010; Baayen, Hendrix \& Ramscar, in press), and successfully predicts patterns of age-related reading time differences (Ramscar, Hendrix \& Baayen, 2012).

To simulate the cross-generational effects of changing name distributions, three versions of the NDR were trained on an identical set of naturalistic linguistic training data ( \(1,500,000\) tokens from the Google Unigrams Corpus were used to simulate the experience of reading to age 20). Names from the distribution of US social security applications for a given year (1910, 1960 and 2010) were interpolated into this sample, based on the frequency with which names appear in the corpus, and the distribution givens name in each year. Recognition latencies were calculated for the set of names common to the 1910, 1960 and 2010 name distributions, and for the total set of names learned by each model: Figure 8A shows the cost the distribution of names imposes at each point in time, and Figure 8B shows the average effect this had on precisely the same set of names. The simulations suggest that the simple task of recognizing a name grew harder in the 20th Century, especially in its latter half: the change in simulated reaction time from 1960 to 2010 is three times greater than 1910 to 1960.

Not only did the number of names increase dramatically (the 1960 model learned \(60 \%\) more names than the 1910 model, and the 2010 model \(83 \%\) more), but the number of non-name words that were learned declined over time, by \(2.5 \%\) in 1960 , and \(5 \%\) in 2010 . Given that the models were trained on exactly the same number of name tokens, this reflects the degree to which the boundary between English names and non-names has become blurred over time, increasing the memory problem that names pose.


Figure 8. A plots the average simulated recognition times for the set of names learned by a " 20 year old" model trained on 1.5 million unigrams when sampling from the name distribution in 1910 (left bars), 1960 (center bars) and 2010 (right bars). B plots the average simulated reading times in each model for the set of names that are common to the 1910, 1960 and 2010 US social
security applications: i.e., each model's predictions for the same set of names.

Finally, to simulate the effect of these changes within a single lifespan, we compared the predictions of a " 20 year old" reading model, trained on \(1,500,000\) unigram tokens, with names sampled from the summed distribution of social security applications from 1950-1960 (the age at which current septuagenarians were 20), to a model trained on \(9,000,000\) unigram tokens, sampling from the summed name distribution from 1950-2010 (extending the "experience" of the younger model to age 70).


Figure 9. A plots the average simulated recognition times for the set of names learned by a "20 year old" model (trained on 1.5 million unigrams, including names from the 1950-1960 distributions; left bars), and a " 70 year old" model (trained on 9 million unigrams, including names from the 1950-2010 distributions; right bars). B plots the average simulated reading times in each model for the set of names that are common to 1960 and 2010 US social security applications. The area below the dashed line represents a 320 ms response constant (button pressing) added to both models in simulating reaction times.

Names and other proper nouns comprise a very large proportion of natural language: whereas the younger model learned 34,480 word types and 4,540 names, the older model learned 61,839 word types and 19,976 names. Although total vocabulary doubled, name lexicon grew fourfold. Figure 9A shows the predicted impact of experience on average name recognition for someone aged 70 in 2010 as compared to fifty years earlier. Figure 9B shows the projected effect of these processing costs on the same set of names in the same individual. After a response constant is removed from the simulated latencies, the model predicts that on average, simply recognizing a name will take today's septuagenarian nearly half a second longer than when she was 20 . Although older adults have hard time remembering names in comparison to their younger selves, a large part of this difference is likely due to the increasing complexity of social name distributions, and the increasing number of names that individuals encounter over the course of their lives.

\section*{The Name Game}

We identified a common information structure in the world's name grammars that helps satisfy the complex demands of communicating about individuals, and described some of the consequences of recent changes in Western naming practices. Two things are worth noting
about these findings: First, the data we report are not inferred from samples of populations, but are instead calculated from records representing the actual populations themselves; and second, the information theoretic methods we used to analyze this data describe and govern all of the physical devices that have come to define our information age. Accordingly, our finding that American female names have twice the perplexity of male names is a statistical fact about the population, which entails that female names in this country must exert considerably higher information processing costs than male names.

These findings may help shed light on many social issues that are far less clear-cut: for example, an often cited reason for the under-representation of women in many professional bodies is that when appointments are made, women's names often don't "come up" (Donald, 2012). It is highly likely that the different processing costs associated with male and female names contribute to this. In a similar vein, these findings offer food for thought for parents choosing names for children in the West, as well as for people with names formed using different grammars as they traverse our increasingly multicultural world. In particular, these findings suggest that the tendency to simply reverse the order of Asian names in Western languages should be viewed with caution.

Finally, these results have implications for our understanding of memory and aging. The belief that memory processes decline as we age rests, in large part, on apparent selective deficits for names in the elderly (Shafto et al., 2007). However, the problem of remembering a name has been getting exponentially harder since before anyone now alive was born (Figure 7). Because current measures of name memory 'deficits' fail to take into account changes to name distributions, it is unclear whether name memory really does decline, or whether these measures simply reflect the overwhelming increases in name information we have documented (see also Dahlgren, 1998; JuncosRabadán, Facal, Soledad Rodríguez \& Pereiro, 2010). It may be that older adults troubled by their "declining" name memories are presently falling into the trap of taking personal responsibility for a broader social problem.

At the height of the information age, in a world where population growth is inexorably increasing the amount of name information societies must shoulder, the social practices that evolved to maximize the efficiency of name processing are, in many cases, in a state of collapse. Although names are the hardest vocabulary items people have to learn and remember, the problems they pose are currently far harder than they should be. Understanding the information structure of names, as well as how name efficiency can vary and be quantified, can help individuals and societies make more informed choices about names and naming practices. It may even make the names of future generations easier to recall.

For a complete list of references and supplementary materials, please consult the article copy hosted at www.sfs.uni-tuebingen.de/

\title{
The Cross-Domain Re-interpretation of Artistic Ideas
}

\author{
Apara Ranjan (apara.ranjan@ubc.ca), Liane Gabora (liane.gabora@ubc.ca), and Brian O'Connor (brian.oconnor@ubc.ca) \\ Department of Psychology, University of British Columbia \\ Okanagan campus, 3333 University Way \\ Kelowna BC, V1V 1V7, CANADA
}

\begin{abstract}
The goal of this study was to investigate the translate-ability of creative works into other domains. We tested whether people were able to recognize which works of art were inspired by which pieces of music. Three expert painters created four paintings, each of which was the artist's interpretation of one of four different pieces of instrumental music. Participants were able to identify which paintings were inspired by which pieces of music at statistically significant above-chance levels. The findings support the hypothesis that creative ideas can exist in an at least somewhat domainindependent state of potentiality and become more welldefined as they are actualized in accordance with the constraints of a particular domain.
\end{abstract}

Keywords: art; artifacts; common coding; creativity; domain generality; ekphrastic expression; individual differences; interpretation; music; self-expression; style; potentiality; synesthesia; translation.

\section*{Introduction}

Although much social interaction occurs directly through words or actions, a great deal of what humans attempt to communicate, such as ideas for works of art, science, or technology, are not readily expressed through these channels. Complex ideas are therefore often communicated indirectly by way of artifacts. There is evidence that artists leave something of themselves-their own personal signatures or creative styles-in their artifacts, and that creators' identities are recognizably present in their creative works. For example, creative writing students familiar with each other's writing identified significantly above chance, not just which of their creative writing classmates had written each particular piece of writing but which of their creative writing classmates had created each artwork (Gabora, 2010; Gabora, O’Connor, \& Ranjan, 2012). Thus, at least in some cases, if a viewer is familiar with the works of creators in a particular domain (such as creative writing), it is possible for the viewer to recognize which creator generated which work, and this is even the case if the works are in a different domain (such as art).

This does not, however, imply that creative artifacts are just the external expression of individual style. We suggest that artifacts constitute a beehive of hidden social interaction, and that their forms reflect, in part, the attempt to transcend one's individuality, i.e., to relinquish oneself to the essence of an idea. We suggest that when personal style is recognizably evident in a work, this is not necessarily due to an attempt to express this style, but a side effect of
participating in the human enterprise of interactively evolving cultural outputs by adapting them to different personal styles, perspectives, and modalities. The creative process involves not just accessing and combining knowledge, experiences, and ideas, but also inspiration, translation, and re-interpretation (Cropley, 1999; Feldhusen, 1995, 2002; Munford \& Gustafson, 1988; Sternberg \& Lubart, 1995). Components of a creative work may originate from oneself, others with whom one has communicated directly or indirectly by way of others, or even multiple individuals through the course of history who each put their own spin on it. Inspiration may come from a work in same domain as the work it inspires (as when one poem inspires an idea for another poem). Alternatively, an idea may first be expressed by one individual in one domain, and subsequently translated by someone else into another domain (as when a piece of music gets re-cast in another musical genre, or even inspires a poem). With the advent of new technologies and social media, the distinction between social interaction and individual creative expression becomes increasingly blurred. For example, as one moves from face-to-face communication, to avatarmediated communication, to music inspired by and intended for someone else, to background music to accompany the activities of a particular cartoon character, to music composed with no obvious inspirational source, it is difficult to draw the line between social interaction and individual self-expression, and cross-modal perception.

The goal of this research was to test the hypothesis that it is possible to recognize the source of inspiration for a creative work when that source of inspiration comes from a different medium. There are several phenomena that suggest that a creative work need not be in the same domain as the inspirational source for the work.

\section*{Related Phenomena}

We now review phenomena that point to cross-domain interpretation of ideas as a source of the character of creative works: synesthesia, ekphrastic expression, and cross-domain style.

Synesthesia Individuals referred to as synesthetes naturally and spontaneously translate stimuli into another sensory domain. For example, they may see particular letters or numbers in particular colors. Ramachandran (2003) proposed that synesthesia occurs as a result of hyperconnectivity in the brain due to partial collapse of the barrier between sensory domains. Artists, poets, and novelists, are
more likely than average to be synesthetes, which suggests that synesthetically driven re-interpretation of inputs from one modality to another can play a role in these creative domains (Ramachandran, 2003).

Ekphrastic Expression There is a tradition in the arts of interpreting art from one medium (e.g., oil paint) into another (e.g., watercolor) and thereby coming to know its underlying essence. This practice is referred to as ekphrastic expression. The idea behind ekphrastic expression is that an artist may have a more direct impact on an audience by translating art from one medium into another medium because this involves capturing, and thereby becoming intimate with, its underlying form or essence. Ekphrasis may be related to the late nineteenth Century practice of associating particular kinds of music with particular colors. There is anecdotal evidence that music of this time frequently served as a direct inspiration for paintings, and musical terminology was used as titles for paintings. Ekphrastic expression is not just a phenomenon of the past. Modern day film composers attempt to compose music that conveys the emotional tone of the events portrayed in the film, thereby heightening the viewer's experience of these events. Thus film scoring can be seen as a form of ekphrasis. The application of ekphrastic methods in the arts supports the idea that creative individuals extract patterns of information from the constraints of the domains in which they were originally expressed and transform them into other domains.

Cross-Media Style Another reason to suspect that the character of creative works arises through cross-domain reinterpretations of ideas is the widespread phenomenon of cross-media style. This refers to artistic style that is demonstrated by works of art in more than one medium. For example, the term rococo is applied to a style of painting, sculpture, literature, and music of the \(18^{\text {th }}\) Century. Works in a given style are thought to derive from abstract archetypal forms or potentialities that make the artistic mind want to explore different arrangements or manifestations (Burke, 1957).

Cross-modal Perception The phenomenon of crossmedia style provides evidence that creative works in different media may be similar in terms of psychophysical, collative, and ecological properties (Hasenfus, 1978). Aesthetic perceptions stimulated by creative works may generate emotional, cognitive, behavioral, and/or physiological responses that are amenable to re-expression in another form. This may arise in part due to regularities with respect to the choice of elements (i.e., colors, shapes, words) and/or how they are used (e.g., in an orderly or chaotic manner) (Berlyne, 1971). Studies indicate that there are non-arbitrary mapping between properties of vision and sound (Griscom \& Palmer, 2012; Mark 1975, Melara, 1989; Melara, \& Marks, 1990; Melara \& O’Brien, 1987; Palmer, Langlois, Tsang, Schloss, \& Levitin, 2011; Ward et al. 2006). For example, the processing of some visual features, such as spatial frequency and lightness, can be affected by auditory features such as pitch and timbre (Mark, 1987).

In the study that perhaps comes closest in spirit to this one, composers were asked to write music inspired by four simple line-drawn shapes: a square, a lightning bold, a curvy shape, and a jagged shape (Willmann, 1944). Music inspired by the same shape was more similar than music inspired by another shape with respect to tempo, melodic pattern, mood, and other characteristics, and listeners could match above chance the music to the shape that inspired it. However, the music could not be said to be reinterpretations of creative works, for the impoverished nature of the stimuli undoubtedly limited the scope for creative expression. The study reported here is the inverse of Willmann's; it investigates not music inspired by art but art inspired by music. Moreover, the goal was to go beyond simple crossmodal mappings to convey in another domain the rich emotionality of genuinely creative works.

\section*{Methods}

This study examined whether people were able to correctly recognize which works of art were inspired by which pieces of music. The study was divided into two phases. In the first phase, expert artists created four paintings, each of which was the artist's interpretation of one of four different pieces of instrumental music. In the second phase, naïve participants attempted to determine which piece of music was used as the source of inspiration for each artwork.

\section*{Phase One}

Participants Two local expert artists, each with approximately 25 years of experience in the field of painting, were recruited for this study. They each received \(50 \$\) for their participation.

Musical Stimuli Four pieces of piano music from commercially produced sound track CDs with no vocal tract and no other instrumentation were used as stimuli to inspire art. They were selected from a pool of 45 pieces chosen as exemplary of different musical styles: baroque classical, romantic, jazz, and contemporary. Each of these original 45 pieces of music was cropped to three minutes duration, and then rated by three raters on 64 descriptive adjectives on five point Likert scales. The adjectives were derived from previous research on the collative properties of stimulus patterns, specifically, measures of affective reactions to artwork (Berlyne, 1974), and the affective circumplex (Russel, 1980; Watson \& Tellegen, 1985). The raters had no previous musical training.

Factor analysis and multidimensional scaling were used to compute the basic dimensions of aesthetic experience in the ratings, and to reveal how the 45 pieces of music were dispersed in the dimensional spaces. The Euclidean distances between the pieces of music across the spaces were used to select four pieces of music from different regions that were clearly dissimilar from each other. The four selected pieces of music were:
(1) 'Love is a Mystery' by Ludovico Einaudi
(2) Number 29 B Flat Major', by Ludwig van Beethoven
(3) 'Circus Gallop' by Marc-André Hamelin

\section*{(4) 'All of Me' by Jon Schmidt}

Creation of Artworks Each of the two artists created one painting for each of the four pieces of piano music, for a total of 8 paintings. On days that paintings were to be created, each artist was provided with a single piece of music and asked to reinterpret it as a painting, i.e., to paint what the music would look like if it were a painting. They were instructed to paint while listening to the music, and encouraged to listen to it as many times as they wished while they painted. They were allowed to use whatever painting supplies they thought could most effectively express the music (e.g., watercolors, oils, and acrylics were all acceptable). They were instructed to complete their paintings in one sitting without interruption. They were instructed to take up to a maximum of 120 minutes to listen to the music and complete the painting. The paintings were created in the artists' personal studios. In order to limit the influence of the previous pieces of music on the new painting, the artists were instructed not to re-listen to the piece of music after the painting was finished, and there was a gap of four days between each painting session.

Representative examples of the music-inspired paintings obtained in Phase One of the study are provided in Figures 1,2 and 3 . These paintings constituted the stimuli that were used in Phase Two. Figures 1 and 2, painted by the same artist in response to different pieces of music, provide the reader with a qualitative sense of the extent to which an artist's personal style comes through in different paintings.


Figure 1. A painting generated by first of the artists as an interpretation of the piece Number 29 B Flat Major', by Ludwig van Beethoven.


Figure 2. A painting generated by first artist as an interpretation of the piece 'All of Me' by Jon Schmidt.

By comparing figures 2 and 3, painted by different artists in response to the same piece of music, the reader can obtain a qualitative sense of how a common musical source of inspiration manifests in different paintings.


Figure 3. A painting generated by the second artist as an interpretation of the piece 'All of Me' by Jon Schmidt.

\section*{Phase Two}

In the second phase of the study we tested whether it was possible to recognize which pieces of music were interpreted as which paintings.

Participants The participants were two groups of undergraduates enrolled in psychology courses at the

University of British Columbia, consisting of 107 and 89 students respectively, for a total of 196 students. They received partial course credit for their participation.

Analytic Methods Two statistics, \(H\) and \(H u\), were computed to assess the accuracies of the participants' paintings-to-music matches. \(H\) is the simple hit rate, or the proportion of correct guesses. Hu comes from signal detection theory (Wagner, 1993). It corrects for chance guessing and for response bias, such as the tendency to use particular response categories more or less than other response categories. For each set of paintings (i.e., for paintings by artist one and artist two), two hit rate statistics were computed for each participant. One-sample t-tests and a data randomization procedure (Manly, 2007) were then used to assess the statistical significance of the mean \(H\) and \(H u\) values. The tests indicated whether the mean \(H\) and \(H u\) values were significantly different from the \(H\) and \(H u\) values that would have been obtained had participants provided random guesses.

Procedure and Materials This part of the study was set up online. There were two sets of the study, one for each artist. In each set, there were the four pieces of music and the four paintings created in phase one by each artist. Each painting was displayed on a web page. Next to each painting were links to the four pieces of music. Two groups of participants consisting of 89 and 107 students were asked to look at the painting and to listen to the four pieces of music respectively. They were asked to identify which piece of music inspired each painting.

\section*{Results}

The results are summarized in Table 1.
Table 1: Mean hit rates, \(t\)-test values, and \(r\) effect size for identification of paintings inspired by pieces of music. All hit rates and t values were statistically significant.
\begin{tabular}{lllll}
\hline & \begin{tabular}{c} 
Mean \\
Hit Rate
\end{tabular} & \begin{tabular}{c} 
Chance \\
Hit Rate
\end{tabular} & \(t(d f)\) & \begin{tabular}{r}
\(r\) \\
Effect \\
Size
\end{tabular} \\
\hline \begin{tabular}{lllll} 
Artist One \\
Hit Rate \((H)\)
\end{tabular} & .35 & .24 & \(4.0(88)\) & .39 \\
\begin{tabular}{l} 
Artist One \\
Unbiased
\end{tabular} & .36 & .24 & \(4.0(88)\) & .40 \\
\begin{tabular}{l} 
Hit Rate \\
\((H u)\)
\end{tabular} & & & & \\
\begin{tabular}{l} 
Artist Two \\
Hit Rate \((H)\)
\end{tabular} & .44 & .25 & \(6.3(106)\) & .52 \\
\begin{tabular}{l} 
Artist Two \\
Unbiased \\
Hit Rate \\
\((H u)\)
\end{tabular} & .46 & .25 & \(6.3(106)\) & .52 \\
\end{tabular}

For the first artist, the mean hit rates were \(H=.35\) and \(H u=\) .36. The mean hit rates that would have been obtained on the basis of random guesses for these questions were .25 and .25 , respectively. Both hit rates are statistically significant according to both conventional and data randomization t-tests \((\mathrm{t}(88)=4.0, \mathrm{t}(88)=4.0, \mathrm{p}<.001)\), and the \(r\) effect sizes were large, .39 and .40 . For the second artist, the mean hit rates were high, \(H=.44\) and \(H u=.46\), statistically significant according to both conventional and data randomization t -tests \((\mathrm{t}(106)=6.3, \mathrm{t}(106)=6.3, \mathrm{p}<\) .001 ), and the effect sizes were large, \(r=.52\) and \(r=.52\). Thus participants identified at above-chance levels which paintings were inspired by which pieces of music for both artist one and artist two.

\section*{Discussion}

There is a longstanding debate concerning the extent to which the semantic complexity of artistic works is amenable to scientific methods (Becker, 1982). We tested the hypothesis that the core idea behind a creative work is recognizable when it is translated from one domain to another. To our knowledge, the only other previous study to test this hypothesis (Willmann, 1944) used highly artificial stimuli that most would not consider creative works in and of themselves. The hypothesis was supported by our finding that when pieces of music were re-interpreted as paintings, naïve participants were able to guess significantly above chance which piece of music inspired which painting. Although the medium of expression is different, something of its essence remains sufficiently intact for an observer to detect a resemblance between the new work and the source that inspired it. The results are consistent with a number of phenomena familiar to artists, mentioned in the Introduction, namely synesthesia, ekphrastic expression, cross-media style, and cross-modal perception.

The research reported on here may be a step toward distinguishing between domain-specific and domain-general aspects of creative works. We suggest that at their core, creative ideas may be much less domain-dependent than they are generally assumed to be. Our results support the view that the uniqueness of a creative work derives at least in part from, not just the personal style of the creator, but from encounters with works in domains that differ from the domain of the creative output, or even different kinds of experiences altogether. In other words, it is possible for the domain-specific aspects to be stripped away such that the creative work exists in an abstracted state of potentiality at which point they are amenable to re-expression in another form. A creative idea may exist in form that is freed of the constraints of a particular domain, and that the creator's job may be in part to, to simply allow that domain-independent entity to take a particular form, using domain-specific expertise and the tools of his or her trade. Over time they may become more fully actualized, and well-defined, as they are considered from different perspectives in accordance with the constraints of the domain in which they are expressed.

The capacity for cross-domain translation of creative ideas supports the hypothesis that an individual's creative outputs are expressions of a particular underlying uniquely structured self-organizing internal model of the world, or worldview. Our creative abilities may be a reflection of the tendency of a worldview to transform in such a way as to find connections, reduce dissonance, and achieve a more stable structure (Gabora \& Merrifield, 2012). This view of creativity is consistent with previous research showing that midway through a creative process, an idea may exist in a 'half-baked' state of potentiality, in which one or more elements are ill-defined (Gabora, 2005, Gabora \& Saab, 2011). When a work is translated from one domain (e.g., music) into another (e.g., painting), the two works may be recognizably related because the process by which the worldview assimilates or comes to terms with the works is at a structural level deeply analogous.

Although that idea that at least some creative tasks involve the abstraction and re-expression of 'raw' potentialities or forms seems obvious to many artists we have spoken with, it stands in contrast with most academic theories of creativity. Creativity is typically portrayed as a process of searching and selecting amongst candidate ideas that exist in discrete, well-defined states. This can be traced back to early views arising in the artificial intelligence community, wherein creativity was thought to proceed by heuristically guided search through a space of possible solutions (Newell, Shaw \& Simon, 1957; Newell \& Simon, 1972; Simon, 1973, 1986) or possible problem representations (e.g., Kaplan \& Simon, 1990, Ohlsson, 1992). The view that creativity proceeds through a process of search and selection is also assumed in more contemporary theories, such as the theory that creativity is a Darwinian process; i.e., new ideas are obtained by generating variations more or less at random and selecting the best (e.g., Campbell, 1965; Simonton 1999a,b, 2005).

Our results bring up the question of what it was about the paintings that made it possible to trace them to the artworks that had inspired them. We are currently investigating to what extent people assign similar experience variable ratings and similarity ratings to paintings and the music that inspired them and whether these ratings correspond even when participants cannot identify which piece of music inspired the painting. A possible clue to the mechanisms underlying cross-domain interpretation of creative ideas comes from research by Feedberg and Gallese (2007) on perceiving action in artwork. They propose that art observers implicitly imitate the creative actions undertaken by the artist in the making of the work. In our study it is possible that observers were not just perceiving action in art but were also able to match qualities of the art with qualities such as the rhythm and tempo of the music that inspired it. The phenomenon of action perception in paintings could also at least partially account for the ability to recognize the essence of ideas interpreted across domains. In order to recognize the inspiration of an artwork or a cross-media style, expertise in a domain might stimulate the action
system while the observer imagines how the artwork was created. Thus, future research will also investigate the role of expertise in the recognition of a connection between works in different domains. We hypothesis that expertise in a domain might increase the activation of the action system while the observer imagines how the artwork got created, thereby enhancing the capacity for recognition of crossdomain re-interpretation in a task such as this.

The effect of inspirational source on creative output may be weaker than the effect of personal style reported earlier (Gabora, 2010; Gabora, O’Connor \& Ranjan, 2012), given that paintings by different artists inspired by the same piece of music could be quite different, as seen by comparing Figures 2 and 3. This could however reflect individual differences with respect to which elements of the source had sufficient personal relevance to serve as departure points for the artists' own creative works. This interpretation is consistent with the finding that when pictures were used as stimuli for poetry, poets focused on particular portions of the pictures to serve as the basis for their poems, and different poets focused on different portions (Patrick, 1935). We are currently investigating how these two factors interact, i.e., whether artists' individual styles influence the ease of identifying which music inspired their paintings. Our aim is not to partition out how much creative works owe their distinctive character to their creators and how much they owe to other sources. We suspect that such a partitioning is not possible, that in the most successful creative works there is a fusion of the two, and that the ability to fuse ones' personal style with the inspirational source for a work plays a role in artistic genius. Though commonly portrayed as introverted and withdrawn, the creative genius may, through the assimilation and generation of creative artifacts, be deeply immersed in a form of social interaction that connects all of humanity to the deepest and most influential thinkers our world has known. We suggest that the extent to which the arts feed on the cross-domain adaptation and reinterpretation of ideas has been underappreciated, and that it may in fact play a pivotal role in the evolution of human culture.

\section*{Acknowledgments}

We are grateful for funding to the second author from the Natural Sciences and Engineering Research Council of Canada and the Concerted Research Program of the Flemish Government of Belgium. We thank Jon Corbett for assistance with this research.

\section*{References}

Becker, H. (1982). Art worlds. Berkeley, CA: University of California Press.
Berlyne, D. E. (1971). Aesthetics and psychobiology. New York: Appleton-Century-Crofts.
Berlyne, D. E. (1974). Studies in the new experimental aesthetics: Steps towards an objective psychology of aesthetic appreciation. London: Halstead Press.

Boden, M. A. (2010). Personal Signatures in art. In M.A. Boden (Ed), Creativity and Art: The three roads to surprise. Oxford: Oxford University Press.
Brown, P. (2008). From systems art to artificial life: Early generative art at the Slade School of Fine Arts. In C. Gere, P. Brown, N. Lambert, \& C, Mason (Eds.) White heat and cold logic: British computer arts 1960-1980, an historical and critical analysis. Cambridge: MIT Press.
Burke, K. (1957). Counter-statement. Chicago: University of Chicago Press.
Cropley, A. J. (1999). Creativity and cognition: producing effective novelty. Roeper Riview, 21(4), 253-265.
Cytowic, R. E. (1995). Synesthesia: Phenomenology and neuropsychology, a review of current knowledge. PSYCHE: Interdisciplinary Journal of Research on Consciousness, 2(10) http://psyche.cs.monash.edu.au/v2/ psyche-2-10-cytowic.html
Flach, R., Knoblich, G., \& Prinz, W. (2004). Recognizing one's own clapping: The role of temporal cues in selfrecognition. Psychological Research, 11, 147-156.
Gabora, L. (2005). Creative thought as a non-Darwinian evolutionary process. Journal of Creative Behavior, 39(4), 262-283.
Gabora, L. (2010). Recognizability of creative style within and across domains: Preliminary studies. Proceedings of the Annual Meeting of the Cognitive Science Society (pp. 2350-2355). August 11-14, Portland, OR.
Gabora, L. \& Merrifield, M. (2012). Dynamical disequilibrium, transformation, and the evolution and development of sustainable worldviews. In (F. Orsucci \& N. Sala, Eds.) Complexity science, living systems, and reflexing onterfaces: New models and perspectives (pp. 69-77). Hershey, PA: IGI Global.
Gabora, L., O’Connor, B., \& Ranjan, A. (2012). The recognizability of individual creative styles within and across domains. Psychology of Aesthetics, Creativity, and the Arts, 6(4), 351-360.
Gabora, L., \& Saab, A. (2011). Creative interference and states of potentiality in analogy problem solving. Proceedings of the Annual Meeting of the Cognitive Science Society (pp. 3506-3511). July 20-23, Boston MA.
Griscom, W. S., \& Palmer, S. E. (2012). The color of musical sounds: Color associates of harmony and timbre in non-synesthetes. Journal of Vision 12(9), 74.
Hasenfus, N. (1978). Cross-media style: A psychological approach. Poetics, 7, 207-216.
Kaufman, J. C., \& Baer, J. (2004). Hawking's Haiku, Madonna's Math: why it's hard to be creative in every room of the house. In R. J. Sternberg, E. L. Grigorenko \& J. L. Singer (Eds.), Creativity: From potential to realization (pp. 3-19). Washington, DC: American Psychological Association.
Manly, B. F. J. (2007). Randomization, bootstrap and Monte Carlo methods in biology (3rd ed.). Boca Raton, FL: Chapman \& Hall/CRC.Mark, L.E. (1975). On colored-hearing synesthesia: cross-modal translations of sensory dimensions. Psychological Bulletin, 82, 303-331

Mark, L. E. (1987). On cross-modal similarity: auditoryvisual interactions in speeded discrimination. Journal of Experimental Psychology: Human Perception and Performance, 13, 384-394
Melara, R. D. (1989). Dimensional interaction between color and pitch. Journal of Experimental Psychology: Human Perception and Performance, 15, 69-79.
Melara, R. D., \& Marks, L. E. (1990). Perceptual primacy of dimensions: Support for a model of dimensional interaction. Journal of Experimental Psychology: Human Perception and Performance, 16, 398-414.
Melara, R. D., \& O'Brien, T. P. (1987). Interaction between synesthetically corresponding dimensions. Journal of Experimental Psychology: General, 116, 323-336.
Munford, M. D., \& Gustafson, S. B. (1988). Creativity syndrome: Intergration, application, and innovation, Psychological Bulletin, 103, 27-43.
Newell, A., Shaw, C. \& Simon, H. (1957). The process of creative thinking. In H. E. Gruber, G. Terrell \& M. Wertheimer (Eds.) Contemporary approaches to creative thinking (pp. 153-189). New York: Pergamon.
Newell, A. \& Simon, H. (1972). Human problem solving. Edgewood Cliffs NJ: Prentice-Hall.
Palmer, S. E., Langlois, T., Tsang, T., Schloss, K. B., \& Levitin, D., J. (2011). Color, music, and emotion. Journal of Vision, 11(11).
Patrick, C. (1935). Creative thought in poets. Archives of Psychology, 178, 1-74.
Prinz, W. (1997). Perception and action planning. European Journal of Cognitive Psychology, 9, 129-154.
Ramachandran, V. S. (2003). The emerging mind. London: BBC/Profile Books.
Repp, B. H., \& Knoblich, G. (2004). Perceiving action identity: How pianists recognise their own performances. Psychological Science, 15, 604-609.
Russell, J. A. (1980). A circumplex model of affect. Journal of Personality and Social Psychology, 39, 1161-1178.
Simon, H. A. (1973). Does scientific discovery have a logic? Philosophy of Science, 40, 471-480.
Simon, H. A. (1986). Understanding the processes of science: The psychology of scientific discovery. In T. Gamelius (Ed.), Progress in science and its social conditions. Oxford: Pergamon Press.
Simonton, D. K. (1999). Creativity as blind variation and selective retention: Is the creative process Darwinian? Psychological Inquiry, 10, 309-328.
Sternberg, R.J., \& Lubart, T. (1995). Defying the crowd. New York: Free Press.
Watson, D., \& Tellegen, A. (1985). Towards a consensual structure of mood. Psychological Bulletin, 98, 219-235.
Wagner, H. L. (1993). On measuring performance in category judgment studies of nonverbal behavior. Journal of Nonverbal Behavior, 14, 387-406.
Zilczer, J. (1987). Color music: Synaesthesia and nineteenth-century sources for abstract art. Artibus et Historiae, 8, 101-126.

\title{
A neural reinforcement learning model for tasks with unknown time delays
}

\author{
Daniel Rasmussen (drasmuss@uwaterloo.ca) \\ Chris Eliasmith (celiasmith@uwaterloo.ca) \\ Centre for Theoretical Neuroscience, University of Waterloo \\ Waterloo, ON, Canada, N2J 3G1
}

\begin{abstract}
We present a biologically based neural model capable of performing reinforcement learning in complex tasks. The model is unique in its ability to solve tasks that require the agent to make a sequence of unrewarded actions in order to reach the goal, in an environment where there are unknown and variable time delays between actions, state transitions, and rewards. Specifically, this is the first neural model of reinforcement learning able to function within a Semi-Markov Decision Process (SMDP) framework. We believe that this extension of current modelling efforts lays the groundwork for increasingly sophisticated models of human decision making.
\end{abstract}

Keywords: reinforcement learning; neural model; SMDP

\section*{Introduction}

One of the most successful areas of cross-fertilization between computational modelling and the study of the brain has been the domain of reinforcement learning (RL). This began with the work of Schultz (1998), who demonstrated that the well-defined computational mechanisms of models (e.g., TD reinforcement learning) could provide insight into some of the more opaque mechanisms of the brain (e.g., dopamine signalling).

The models used in that early work were purely algorithmic, with little relation to the biological properties of the brain. However, since that first demonstration many new models have been developed, allowing novel or more detailed comparisons to neural mechanisms-models that more closely reflect the structures of the brain (Frank \& Badre, 2012; Stewart et al., 2012), the behaviour of individual neurons (Seung, 2003; Potjans et al., 2009), or synaptic learning mechanisms (Florian, 2007; Baras \& Meir, 2007).

In our work we seek to retain the neuroanatomical detail of these models, but expand their functionality; that is, to build models capable of more powerful learning and decision making, enabling them to solve more complex problems. Here we present some first steps in that direction. Specifically, we will discuss the implementation and show early results from a model that is able to solve tasks requiring extended sequences of actions, in environments where there may be unknown and variable time delays between actions and rewards.

\section*{Background}

Sutton \& Barto's seminal introduction to reinforcement learning illustrates the important challenge for expanding the function of neural RL models: "Reinforcement learning is learning what to do-how to map situations to actions-so as to maximize a numerical reward signal...In the most interesting and challenging cases, actions may affect not only the imme-
diate reward but also the next situation and, through that, all subsequent rewards (Sutton \& Barto, 1998)."

Most existing neural models have performed only associative reinforcement learning, where there is no consideration of future reward (Niv et al., 2002; Seung, 2003; Baras \& Meir, 2007; Florian, 2007; Izhikevich, 2007; Frank \& Badre, 2012; Stewart et al., 2012). An example of this type of task is bandit learning, where the agent selects one of \(n\) available options, receives reward, then is reset back to the choice point. Each trial is independent, so the agent only needs to learn the immediate reward associated with each option, and then pick the best one. This can be expressed in the RL notation as
\[
\begin{equation*}
Q(s, a)=r(s, a) \tag{1}
\end{equation*}
\]
where \(Q(s, a)\) is the agent's estimate of the value of taking action \(a\) in state \(s\), and \(r(s, a)\) is the immediate reward received for performing that action in that state. These \(Q\) values can be learned by observing \(r(s, a)\) and then updating \(Q(s, a)\) to bring it closer to the observation. The challenge addressed by many of the models above is how to do that update in a neurally plausible manner.

An example of a more complex reinforcement learning task is a navigation problem, where an agent seeking to reach a goal must choose a direction to move. The agent may receive no immediate reward for making a choice, but there are still good and bad choices (bringing it closer to or farther from the goal). In order to make correct decisions, the agent needs to be able to learn not only the immediate rewards, but the rewards to be expected in the future after taking a given action. This can be expressed as
\[
\begin{equation*}
Q(s, a)=r(s, a)+\gamma Q\left(s^{\prime}, a^{\prime}\right) \tag{2}
\end{equation*}
\]

In other words, the value of taking action \(a\) is equivalent to the immediate reward (as in the previous case), plus the expected value of the action taken in the resulting state (indicating the future reward expected from that state). The future value is discounted by \(\gamma<1\) to indicate that future rewards are valued less than immediate rewards. The \(Q\) values can be learned by comparing the predicted value of action \(a\) to the observed values upon arriving in state \(s^{\prime}\). This is the temporal difference (TD) learning formula \({ }^{1}\) :
\[
\begin{equation*}
\Delta Q(s, a)=\kappa\left[r(s, a)+\gamma Q\left(s^{\prime}, a^{\prime}\right)-Q(s, a)\right] \tag{3}
\end{equation*}
\]

Most complex problems of the type faced by the brain require the consideration of the future impact of a given action; thus

\footnotetext{
\({ }^{1}\) More specifically, this is the SARSA learning update (Rummery \& Niranjan, 1994).
}
building models capable of this type of learning is an important step in understanding the decision making processes in the brain.

There have been models built that solve these types of tasks, but often they take the TD error signal (Equation 3) as given, or it is computed outside the model (Foster et al., 2000; Strösslin \& Gerstner, 2003). This reduces to a problem very similar to Equation 1, where the agent has a signal coming in and only needs to worry about the current value of that signal. The challenging aspect of TD learning is how to learn with only immediate rewards as input to the model.

Potjans et al. (2009) presented one of the most complete neural models of reinforcement learning. In order to compute the TD error they use two activity traces, one fast and one slow, on the output of the neurons representing the \(Q\) values. For a brief window in time after the system transitions from state \(s\) to state \(s^{\prime}\), the slow trace will still be representing \(Q(s, a)\) while the fast trace will be representing \(Q\left(s^{\prime}, a^{\prime}\right)\); combining that information with the incoming reward enables the neurons to calculate the equivalent of Equation 3.

The downside of this approach is that the necessary information is only present immediately after the state transition, within that window of time before the slow activity trace catches up to the fast; if action selection occurs earlier than the state transition, or if the rewards are not delivered within that window, the system will not be able to learn. This is true of all systems that rely on some type of activity/eligibility trace to preserve the action values (e.g., Izhikevich, 2007; Florian, 2007). These models rely on an environment that follows a reliable clock-like sequence of action selection, state transition, and reward.

In some cases that may be a reasonable assumption, but in our work we seek a more general mechanism that can learn when there is an unknown and potentially variable delay between action selection and state transition or reward. This can be expressed as a Semi-Markov Decision Process (SMDP; Howard, 1971). Whereas in basic MDPs (the standard model for RL tasks) states, actions, and rewards all occur instantaneously, SMDPs introduce the concept of a time delay between action selection and state transition, and rewards that can be delivered at different points in time.

One way to address the problem of time delays in the MDP environment (without resorting to SMDPs) is to imagine the delay period as a series of state transitions. That is, the states/actions/rewards continue to proceed in a regular clocklike manner, and time delays are represented by multiple cycles through that loop. However, this requires the learning to propagate back through all the "decisions" made during the delay period. This greatly complicates the learning process, and for lengthy delay periods with many different decisions it can render successful learning practically impossible. An important advantage of the SMDP framework is that it encapsulates all the activity of the delay period within a single learning update. This is particularly useful in situations such as hierarchical decision making, discussed more in the con-


Figure 1: Overall architecture of the model, see text for details. The interior of the \(E\) component is shown in Figure 2.
clusion.
The learning update from Equation 3 can be reformulated for an SMDP environment (Bradtke \& Duff, 1994; Sutton et al., 1999) as
\[
\begin{equation*}
\Delta Q(s, a)=\kappa\left[\sum_{t=0}^{\tau-1} \gamma^{t} r(s, a, t)+\gamma^{\tau} Q\left(s^{\prime}, a^{\prime}\right)-Q(s, a)\right] \tag{4}
\end{equation*}
\]
where \(t\) is the time elapsed since action \(a\) was selected, \(r(s, a, t)\) is the reward received at time \(t\), and the transition to state \(s^{\prime}\) occurs at time \(\tau\). The obvious changes are that a) the rewards received are summed over time, and b) the discount is applied across the delay period. However, the more subtle change is that the agent does not know \(\tau\). That is, it cannot rely on the rewards or discount being limited to some specific time window, or the update being applied at a particular time; it must simply wait, and be able to calculate Equation 4 whenever the state change occurs. For the sake of simplicity we have expressed time here as consisting of discrete time steps, but it can be expressed in the continuous case by taking the integral over the incoming reward signal (this is the approach used in our model).

With the SMDP framework, an agent can learn to select actions in a more general environment, incorporating arbitrary time delays into the reinforcement learning process. By taking this theory and implementing it in a neural model, we will develop a more powerful and flexible model of reinforcement learning in the brain.

\section*{Methods}

\section*{Model architecture}

The overall structure of the model is shown in Figure 1. At the top is a population of neurons representing the current state (we will discuss how the environmental state is translated into neural activities in the next section). Beneath are populations associated with each available action (four in this
case, but the model can work with any number). The state population is connected to each action population, and it is in the synaptic weights of these connections that the \(Q\) values are calculated. Assuming that correct weights have been learned, the output of the state neurons will cause each action population to represent the value of taking its associated action in that state (i.e., \(Q\left(s, a_{n}\right)\) ).

In order to act, the model needs to make a decision based on those \(Q\) values; this is the purpose of the selection component. In our model the agent follows a simple greedy policy of always selecting the highest value action. We compute the max operation using the basal ganglia model described in Stewart et al. (2010). That output is used to activate inhibitory gates within the network of the selection component, so that neural populations corresponding to the non-selected actions will not be active. The output of the selection component is both the value of the selected action, which is sent to the error calculation network (to be discussed later), and the actual output of the agent (i.e., the action it sends to the environment).

The operation of the agent is independent of the details of the environment; this model is designed to function in any task that can be described in the SMDP framework. All that is required is that the environment somehow takes the output of the agent (e.g., an action such as "move left"), calculates an updated state (e.g., the new position of the agent), and sends the new state and any reward received back to the agent. As per the SMDP framework, the state transition can occur at any time, and the rewards can be delivered at any time. When the new state is sent to the agent, it will modify the activities in the state population, a learning update will be performed as in Equation 4, and the agent will decide on a new action.

\section*{Representing and computing with neural activities}

The model operates entirely in neural activities, but it needs to interact with environments and perform computations that are defined in terms of abstract mathematical variables. To translate back and forth between these two domains we use the Neural Engineering Framework (NEF; Eliasmith \& Anderson, 2003).

The first component of the translation is encoding. For example, the abstract state output by the environment needs to be encoded into the activities of the state population. Suppose the state is represented by a vector \(x\) (perhaps describing the position of the agent). The model operates in continuous time, so the changing state over time can be represented by \(x(t)\). That input signal is encoded into the activities of the state population as
\[
\begin{equation*}
s_{i}(x(t))=G_{i}\left[\alpha_{i} e_{i} x(t)+J_{i}^{b i a s}\right] \tag{5}
\end{equation*}
\]
\(s_{i}(x(t))\) represents the activity of neuron \(i\) in the state population. \(G_{i}\) is the neuron model; in our case we use leaky integrate and fire (LIF) neurons. The components within the braces represent the current that is input into the neuron model. \(\alpha\) and \(J_{i}^{\text {bias }}\) are parameters of the neuron, randomly
chosen from within biologically plausible ranges, representing the gain and background activity, respectively. The vector \(e_{i}\) identifies the neuron's preferred stimulus, the area of the input space to which this neuron is most sensitive (these are also randomly chosen). Thus each neuron will respond to the input according to its internal parameters and how close the input is to the neuron's preferred stimulus. The combined activity of the whole population then comprises a distributed representation of where the current input is in the input space. Note that while for demonstration purposes we have described this here in terms of encoding the state, this is a general purpose mechanism for encoding any input into the activities of a population of neurons.

The second aspect of the translation is decoding, translating the activities of a population of neurons back into an abstract value. For example, this allows the activities of neurons in the selection network to be interpreted as an action for the environment, or the activities of the action populations to be interpreted as \(Q\) values. This is accomplished by a weighted summation of the neural activities:
\[
\begin{equation*}
\hat{x}(t)=\sum_{i} s_{i}(x(t)) d_{i} \tag{6}
\end{equation*}
\]

The weights, or decoders, \(d_{i}\), can be calculated by
\[
\begin{align*}
d & =\Gamma^{-1} \Upsilon \\
\text { where } & \\
\Gamma_{i j} & =\int s_{i}(x) s_{j}(x) d x \\
\Upsilon_{j} & =\int s_{j}(x) f(x) d x \tag{7}
\end{align*}
\]
\(f(x)\) gives the option of decoding a (possibly nonlinear) function of the encoded value. However, in most cases all that is desired is the identity of the represented value, in which case \(f(x)=x\). With these two tools, encoding and decoding, we can translate back and forth between the neural activities of the model and the variables and computations of the RL framework.

\section*{Learning}

The basic process of TD reinforcement learning involves updating the agent's estimation of the value of each action, the \(Q\) values. In the architecture of the model, this means modifying the synaptic weights on the connections between the state and action populations. To perform these updates we use an error modulated neural learning rule developed by MacNeil \& Eliasmith (2011):
\[
\begin{equation*}
\Delta \omega_{i j}=\kappa \alpha_{j} e_{j} E s_{i}(x) \tag{8}
\end{equation*}
\]
\(\Delta \omega_{i j}\) is the change in the connection weight between neuron \(i\) (in the state population) and neuron \(j\) (in an action population). \(\kappa\) is the learning rate, \(\alpha_{j}\) and \(e_{j}\) are properties of neuron \(j\) (as shown in Equation 5), \(s_{i}(x)\) is the activity of neuron \(i\), and \(E\) is the error. For this model, the error is the desired change in the \(Q\) value, i.e., \(\Delta Q(s, a)\) from Equation 4. This is


Figure 2: Network to calculate the SMDP learning error (the interior processing of the \(E\) component shown in Figure 1). See text for details.
a neurally plausible weight update in that it only makes use of information available locally at neuron \(j\) (assuming all neurons also receive the error signal \(E\) ). MacNeil \& Eliasmith (2011) show that this learning rule will cause the weights to be adjusted so as to minimize \(E\), meaning that over time the weights will come to calculate the desired \(Q\) values.

\section*{Error calculation}

The previous section raises the question of where the error, \(E\), comes from. That is, how is Equation 4 computed? The network that performs this calculation is shown in Figure 2. Note that this is the \(E\) component shown in Figure 1, and receives the inputs shown there (the \(Q\) value of the selected action, and the reward from the environment).

One challenge is the integration of the incoming reward (the summation in Equation 4). This is accomplished by the top-right component in the network. The central feature of the integration population is the recurrent connection, which allows it to maintain its activity in the absence of input. This means that as new rewards enter the population they will be added to the previous rewards already being represented, so that the population represents the sum of the given rewards. The details of how to set up a recurrent network to perform these kinds of computations are described in Eliasmith (2005).

The "current value" population represents the value of the currently selected action. When the action is first selected, this value is transferred into the "stored value" population in the bottom left. Again, this is a population that will maintain its represented value via its recurrent connections. When a state transition occurs, the bottom population will then be representing the value of the selected action in the new state, \(Q\left(s^{\prime}, a^{\prime}\right)\), while the "stored value" population maintains \(Q(s, a)\).

The discount is calculated by integrating the value repre-


Figure 3: Example of policy learned by the agent. The arrows represent the weighted sum of the four possible movement directions, where each direction is weighted by the learned \(Q\) value of that action. Contours indicate the state value (the value of the highest-valued action).
sented in the "stored value" population, using the same recurrent setup as is used to integrate the incoming reward. This value is then subtracted from the current \(Q\) input to calculate a discounted action value. This is not identical to the discount expressed in Equation 4, but it has a similar computational effect: it reduces the value of future states proportional to the time elapsed and the value of the state.

The final "error" population thus has all the pieces it needs to compute the SMDP learning update. It adds the accumulated reward and the discounted \(Q\left(s^{\prime}, a^{\prime}\right)\) value, and subtracts the stored \(Q(s, a)\) value, resulting in the error signal required by the neural learning rule (Equation 8).

\section*{Results}

We tested the model on a spatial navigation task (the same task used in Potjans et al. 2009). The agent is randomly placed in a \(5 \times 5\) grid, surrounded by walls. The agent's state is its \(x, y\) location in the grid, and the available actions are movement in the four cardinal directions. Selecting any of those actions will cause the agent to move one square in that direction, unless it is attempting to move into a wall in which case it remains in the same position. The agent's time in each state is randomly determined, ranging between 600 and 900 ms . The task is to move to some fixed target location. This is equivalent to a water-maze type task, where the agent has no idea where the goal might be, and must find it by exploring the environment. When the agent finds the goal state it receives a constant reward of 1 as long as it remains in the state. After a brief period of time the agent is then moved to a random location, and must find the target again.

Figure 3 shows an example of a policy learned by the model after spending approximately 2 hrs of simulated time


Figure 4: Comparison of learning times between a) an algorithmic implementation of RL (basic table-based Q-learning), b) the neural MDP reinforcement learning model of Potjans et al. (2009), and c) the model presented here. Latency is measured as the difference between the Manhattan distance from start to goal and the number of steps taken by the model. Data for b) from Potjans et al. (2009).
in the task. The arrows display the weighted sum of the four movement directions, where the weights are the learned \(Q\) values associated with each action. Since the agent picks the highest valued action, it will move in whichever cardinal direction is closest to the direction of the arrow. The contours indicate the value of the highest valued action (i.e., the state value function). It can be seen that the agent has successfully learned a policy that will take it to the goal state from any position, despite the random time delays.

Figure 4 shows a comparison between the learning times of our model and that of Potjans et al. (2009), with a purely computational RL implementation as a baseline. Each trial begins when the agent is placed at a random location in the grid, and ends when it reaches the goal (at which point it is placed in a new location for the next trial). We have followed Potjans et al. in using latency as a measure of how well the agent has learned the task. Latency is defined as the difference between the Manhattan distance between the start and goal \(\left(\operatorname{start}_{x}-\operatorname{goal}_{x}+\operatorname{start}_{y}-\right.\) goal \(\left._{y}\right)\), which is the shortest possible path length, and the number of steps taken by the model. It can be seen that our model performs better than that of Potjans et al., and roughly equivalently to the purely computational solution. It is also worth noting that our model is operating in the more challenging SMDP framework, with random time delays; it is unlikely that the Potjans et al. model would be able to perform this task at all.

SMDPs also provide a more powerful language with which to describe problem domains, by allowing for the incorporation of time directly into the task description. For example, Figure 5 shows a task similar to that in Figure 3, but certain states (shown in grey) take a longer period of time for the agent to move through (simulated by adding three seconds


Figure 5: Policy learned by the system in a task where certain states take longer periods of time to move through (shown in grey). The agent has learned to avoid the slow areas even though it requires taking a less direct route.
to the usual randomly determined state transition time). This means that the most efficient route to the goal is no longer a direct path; the agent has learned to trade off the cost of a detour with the cost of moving through the slow areas. Time is often an important part of real world tasks, thus the ability to incorporate time directly into the agent's learning is another advantage of the SMDP framework.

\section*{Discussion}

We have presented a novel neural model capable of autonomous reinforcement learning. The model is able to solve complex tasks that require an extended sequence of actions in order to achieve the reward, rare for biologically based neural models. In addition, it is able to solve these tasks in a realistic SMDP environment, where there are potentially random and unknown delays between action selection, state transition, and reward. We believe this is currently the only neural model capable of this type of performance.

This model is still only an early step on the path of expanding the functional capabilities of neural RL models, and there are a number of ways in which it can be improved. First, more neural detail could be incorporated into the model. For instance, incorporating more realistic spiking neurons would allow for more detailed comparisons to neural recordings. Another improvement to the model would be a more principled approach to exploration. At the moment exploration is accomplished by injecting random noise into the action values as they enter the selection component (a neural approximation of the standard \(\varepsilon\)-greedy approach). However, in the future it would be desirable to have more control over the exploration process, so that, for example, the agent could make decisions about how much exploration to pursue based on its current knowledge.

Another avenue for future work is to incorporate the learning components of this system into a more complete agent model. The inputs and outputs of this model are abstract, thus it ignores the complexity of sensory processing and motor output. However, recent work in our lab has developed an integrated brain model that is able to perceive visual input, process it internally, and control motor outputs (Eliasmith et al., 2012). That model was able to perform associative reinforcement learning, but not the more complex learning displayed here. Adding the abilities of this model into that detailed neural agent would allow for the study of the full reinforcement learning process, from input through to output.

One of the most interesting possibilities opened up by this model is the construction of a neural model capable of hierarchical reinforcement learning (Barto \& Mahadevan, 2003). In hierarchical RL the "actions" that an agent chooses between can be augmented with subroutines that define whole new behaviours. For example, instead of the agent just choosing between "go left", "go right", and so on, one of its options could be "go to the doorway", which would then lead to a sequence of decisions designed to take the agent to that location. What all of the hierarchical approaches have in common is that they use the SMDP framework as their underlying structure. The unknown time delay between action and state transition can be used to encapsulate the time when the high-level action is executing. The SMDP framework allows the agent to incorporate those time delays and rewards, and learn how to correctly select between its complex set of actions. A model such as the one we present here is a step toward a functional neural model capable of hierarchical learning and decision making.

\section*{Acknowledgements}

This work was supported by the Natural Sciences and Engineering Research Council of Canada, Canada Research Chairs, the Canadian Foundation for Innovation, and Ontario Innovation Trust.

\section*{References}

Baras, D., \& Meir, R. (2007). Reinforcement learning, spike-time-dependent plasticity, and the BCM rule. Neural Computation, 19(8), 2245-79.
Barto, A. G., \& Mahadevan, S. (2003). Recent advances in hierarchical reinforcement learning. Discrete Event Dynamic Systems, 1-28.
Bradtke, S. J., \& Duff, M. O. (1994). Reinforcement learning methods for continuous-time Markov decision problems. In Advances in Neural Information Processing Systems.

Eliasmith, C. (2005). A unified approach to building and controlling spiking attractor networks. Neural Computation, 17(6), 1276-1314.
Eliasmith, C., \& Anderson, C. (2003). Neural engineering: Computation, representation, and dynamics in neurobiological systems. Cambridge: MIT Press.

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, C., \& Rasmussen, D. (2012). A large-scale model of the functioning brain. Science, 338(6111), 1202-1205.
Florian, R. V. (2007). Reinforcement learning through modulation of spike-timing-dependent synaptic plasticity. Neural Computation, 19(6), 1468-502.
Foster, D. J., Morris, R. G., \& Dayan, P. (2000). A model of hippocampally dependent navigation, using the temporal difference learning rule. Hippocampus, 10(1), 1-16.
Frank, M. J., \& Badre, D. (2012). Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1 : computational analysis. Cerebral Cortex, 22(3), 509-26.
Howard, R. A. (1971). Dynamic Probabilistic Systems. Dover Publications.
Izhikevich, E. M. (2007). Solving the distal reward problem through linkage of STDP and dopamine signaling. Cerebral Cortex, 17(10), 2443-52.
MacNeil, D., \& Eliasmith, C. (2011). Fine-tuning and the stability of recurrent neural networks. PloS ONE, 6(9), e22885.
Niv, Y., Joel, D., Meilijson, I., \& Ruppin, E. (2002). Evolution of Reinforcement Learning in Uncertain Environments: A Simple Explanation for Complex Foraging Behaviors. Adaptive Behavior, 10(1), 5-24.
Potjans, W., Morrison, A., \& Diesmann, M. (2009). A spiking neural network model of an actor-critic learning agent. Neural Computation, 339, 301-339.
Rummery, G., \& Niranjan, M. (1994). On-line Q-learning using connectionist systems (Tech. Rep. No. September).
Schultz, W. (1998). Predictive reward signal of dopamine neurons. Journal of Neurophysiology, 80, 1-27.
Seung, H. S. (2003). Learning in spiking neural networks by reinforcement of stochastic synaptic transmission. Neuron, 40(6), 1063-73.
Stewart, T. C., Bekolay, T., \& Eliasmith, C. (2012). Learning to select actions with spiking neurons in the Basal Ganglia. Frontiers in Decision Neuroscience, 6, 2.
Stewart, T. C., Choo, X., \& Eliasmith, C. (2010). Dynamic behaviour of a spiking model of action selection in the basal ganglia. In S. Ohlsson \& R. Catrambone (Eds.), Proceedings of the 32nd Annual Conference of the Cognitive Science Society (pp. 235-240). Austin: Cognitive Science Society.
Strösslin, T., \& Gerstner, W. (2003). Reinforcement learning in continuous state and action space. In International Conference on Artificial Neural Networks.
Sutton, R. S., \& Barto, A. G. (1998). Reinforcement Learning. Cambridge: MIT Press.
Sutton, R. S., Precup, D., \& Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. Artificial Intelligence, 112(1-2), 181-211.

\title{
Reduced neural sensitivity to online social interactions in autism
}

\author{
Elizabeth Redcay (redcay@umd.edu) \\ Department of Psychology, University of Maryland \\ College Park, MD 20782 USA
}

\author{
Rebecca Saxe (saxe@mit.edu) \\ Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology \\ Cambridge, MA 02138 USA
}

\begin{abstract}
Difficulty with social interactions is a hallmark characteristic of autism spectrum disorders. While many studies have investigated the neural mechanisms underlying atypical social cognition, the methods used have rarely involved social interaction, relying instead on offline reasoning about a character. In the current study, we examined whether and which brain systems are sensitive to online social interactions in individuals with autism. We compared functional MRI data collected from 15 neurotypical (NT) and 15 autism spectrum disorder (ASD) participants during live real-time interactions (Live) and during a video replay of the same interaction (Recorded-Same) and a novel interaction (Recorded-Novel). Whole brain analyses demonstrated a significantly greater response to Live than Recorded conditions, in NT vs ASD, within left posterior superior temporal sulcus (pSTS) and regions of the cerebellum bilaterally. Region of interest analyses revealed that right posterior temporal regions were differentially recruited during online social interactions in the ASD and NT groups. Also, regions commonly associated with personal salience (i.e., dorsal anterior cingulate and bilateral insula) were sensitive to online social interactions in NT, but to novelty in the ASD group. These data suggest reduced and atypical neural sensitivity to online social interactions in individuals with autism.
\end{abstract}

Keywords: social interaction; autism; fMRI.

\section*{Introduction}

Social interactions provide a rich opportunity to learn from others beginning early in infancy and continuing throughout one's life. Individuals with autism engage in fewer interactions than their typically developing peers and reduced social engagement predicts later delays in language and social abilities (e.g., Mundy, Sigman, \& Kasari, 1990). A central question in the study of autism is what underlies this reduced engagement in social interactions. Some have proposed that social interactions are inherently rewarding, and thus motivating, for neurotypical (NT) individuals but not for those with autism spectrum disorders (ASD) (e.g., Chevallier, Kohls, Troiani, Brodkin, \& Schultz, 2012; Dawson et al., 2002). Similarly, others suggest that, unlike NT individuals, social stimuli fail to capture the attention of those with autism (e.g., Klin, Jones, Schultz, \& Volkmar, 2003). Others still suggest difficulties with social interactions arise from impairments in theory of mind, or reasoning about another person's thoughts (Baron-Cohen, Leslie, \& Frith, 1985).

While evidence exists to support each of these claims, most of the empirical data come from studies using proxies for social interactions, such as a picture, video, or vignette of a person or characters. While important, these offline methods may be missing the processes at the root of ASD, namely social interactions or engagement with others. For example, difficulties interpreting or predicting a social partner's behavior are thought to underlie real-world difficulties in communication; however, offline tasks in which individuals must predict a fictional character's action based on false beliefs often fail to find differences between autism and neurotypical groups in behavioral reports (e.g., Senju, Southgate, White, \& Frith, 2009) and brain activation patterns (Dufour et al., 2012). Interestingly, while offline reasoning processes appear to be relatively intact, individuals with autism fail to spontaneously anticipate the location of an actor's reach based on a false belief (Senju et al., 2009) - a process more akin to real-world use of belief inferences to predict behavior. Furthermore, even for neurotypical individuals, social or communicative behavior in the context of an interaction, as compared to mere observation, may be quantitatively and/or qualitatively different from offline social communication (e.g., Clark \& Brennan, 1999; Pönkänen, Alhoniemi, Leppänen, \& Hietanen, 2011; Redcay et al., 2010; Risko et al., 2012; Schilbach et al., 2012; Sebanz, Bekkering, \& Knoblich, 2006; Shimada \& Hiraki, 2006) Thus, like others (e.g., Schilbach et al., 2012), we argue for a second-person neuroscience approach to understand core difficulties with social interaction in individuals with autism.

Using a novel method for collecting fMRI data during an online social interaction, we previously demonstrated that brain systems supporting reward processing, social cognition, and attention were engaged more when interacting with another person in a real-time face-to-face interaction (i.e. the Live condition) than during a video replay of the experimenter from the same interaction (Recorded-Same condition) or video replay of the experimenter taken from a different scan session (RecordedNovel condition) (Redcay et al., 2010). Thus, this paradigm provides a method to examine the extent to which reward, attention, and social-cognitive systems are engaged during simple social interactions in individuals with autism, and as such can provide insight into the proposed mechanisms underlying atypical social interactions.

The goals of the current study were to 1) replicate findings from Redcay et al., (2010) in a new neurotypical sample, 2) determine what is driving the difference between live and recorded conditions (i.e., novelty or social contingency), and 3) examine whether reward, attention, or social-cognitive systems (or some combination) show an atypical response profile in individuals with autism. To investigate these questions, we examined the response profiles for each condition of interest (Live, RecordedSame, Recorded-Novel) within the regions of interest identified in the previous study for the contrast of Live vs. Recorded conditions (Redcay et al., 2010). A greater response to Live interactions as compared to the same video replay (Recorded-Same) may simply be due to the novelty of the interaction. Thus, the critical comparison to isolate brain regions sensitive to contingent social interaction, independent of novelty, is Live vs. Recorded-Novel. In both of these conditions, the participant sees the experimenter moving and talking in novel ways with novel objects; the only difference is that in the Live condition, the experimenter's actions are contingent on real-time communication with the participant. Based on our previous study, we predicted that regions within social, attention, and reward networks would be differentially recruited during the Live condition in the NT group. Given the hypotheses discussed above, we predicted reduced differentiation between Live and Recorded conditions in the ASD group within regions associated with reward and social cognition.

\section*{Methods}

\section*{Participants}

All participants provided written, informed consent as approved by the Committee on the Use of Humans as Experimental Subjects (COUHES) at the Massachusetts Institute of Technology and were compensated monetarily for their participation. Participants were excluded if they had a history of neurological or psychiatric disorders or any contraindication for MRI scanning. IQ data were collected using the Kaufman Brief Intelligence Test (K-BIT).

Table 1: Participant Information.
\begin{tabular}{ccccc}
\hline Group & n & Age (yrs) & Sex & FIQ \\
\hline ASD & 15 & \(28.4 \pm 7.1\) & 11 M & \(119.5 \pm 14.8\) \\
NT & 15 & \(27.4 \pm 6.2\) & 11 M & \(117.5 \pm 12.3\)
\end{tabular}

Participants with Autism Eighteen adults with highfunctioning ASD participated in the current experiment. All participants met criteria for ASD (autism or spectrum) on the Autism Diagnostic Observation Schedule (ADOS), Module 4. Three participants were excluded because of an inability to perform the task (2) or excessive movement during the scan (criteria described below).

Neurotypical Participants Fifteen NT participants were recruited to match the ASD participants on age and sex.

Verbal, nonverbal, and full-scale IQ scores did not differ significantly between ASD and NT participants (IQ data from 1 ASD and 4 NT are missing).

\section*{Study Design}

Prior to each scanning session the experimenter administered consents, screening forms, and IQ assessments in order for all participants to have some familiarity with the same experimenter in the face-to-face fMRI task.

Live face-to-face set-up During fMRI data acquisition participants were able to see and hear an experimenter in the control room. For extensive details on the audio-visual setup see Redcay et al., 2010. Briefly, during the Live conditions, a real-time video and audio feed of the experimenter was provided to the participant. For all conditions, the experimenter viewed a real-time video feed of the participant's eye through use of a camera from an eye-tracker at the back of the scanner bore. With this dual video set-up both experimenter and participant could interact in real-time. The timing of dual video capture and presentation was implemented using Psychtoolbox extensions in Matlab 7.8 (Brainard, 1997; Pelli, 1997). This dual video capture capability allowed for post-scan coding of the participants eye-movements as well as with the experimenter's actions throughout the experiment.


Figure 1. Example of a social interaction block for Live, Recorded-Same, and Recorded-Novel conditions. Video frames are presented to illustrate the sequence of events.

Social Interaction Task During fMRI data collection, participants engaged in a social interaction task, in which the experimenter prompted them to choose one of two buckets (via eye movements) in the context of a highlyscripted interaction (Figure 1). During 'Live' conditions these interactions occurred in real-time while 'Recorded' conditions involved video replays. Participants were told whether they were in the Live or Recorded conditions both via a green or red square around the screen, respectively, and a text prompt before the start of the block and above the video of the experimenter throughout the block. Importantly, they were told to play along with the experimenter's requests during the Recorded conditions even though she could not see them. During the Recorded-

Same condition, the same video of the experimenter from the previous Live condition was replayed to the participant, serving as a perfect control for perceptual complexity. During the Recorded-Novel condition a novel video from a previous interaction with a different participant was presented, controlling for the novelty of the live interaction.
fMRI design Conditions were presented in a blocked design with each block lasting 40 seconds. Each run contained two repetitions of each condition (i.e., Live, Recorded-Same, Recorded-Novel) alternating in a pseudo-counterbalanced order (with the caveat that Live had to precede RecordedSame). To allow for the opening and closing of video capture devices, the first and last 2.5 seconds of each block were modeled but not analyzed. Runs contained 3 blocks of a 20 -second resting baseline at the beginning, middle, and end of each run. All participants completed four experimental runs except for one participant in the ASD group who completed 3.

\section*{Data acquisition and analyses}

Data acquisition Data were collected on a 3T Siemens Tim Trio scanner at the Athinoula A. Martinos Imaging center at the McGovern Institute for Brain Research at the Massachusetts Institute of Technology. Functional imaging data were collected using a T2*-weighted gradient echoplanar image sequence with a voxel resolution of \(3.1 \times 3.1 \times 4.0 \mathrm{~mm}(\mathrm{TR}=2 \mathrm{~s}, \mathrm{TE}=30 \mathrm{~ms}, 32\) slices). Siemens PACE online motion correction was used to adjust for head movement ( \(<8 \mathrm{~mm}\) ). T1-weighted structural images were collected with 128 slices axially ( \(\mathrm{TE}=3.39 \mathrm{~ms}, \mathrm{TR}=2530\) \(\mathrm{ms}, 1.3 \mathrm{~mm}\) isotropic voxels).
fMRI analyses fMRI data were analyzed using SPM8 (http://www.fil.ion.ucl.ac.uk/spm) and in-house Matlab scripts. Preprocessing steps included 1) realignment of all data to the first volume of the first run using a 6-degree rigid spatial transformation, 2) spatial smoothing with a 5 mm full width half maximum Gaussian filter, 3) spatial normalization to a standard EPI template in Montreal Neurological Institute (MNI) space using a 12-parameter affine transformation. A high pass filter of \(260 \mathrm{~s}(1 / 260 \mathrm{~Hz})\) was applied to the functional data to model low-frequency signals unrelated to the task. 260 seconds was chosen because it is the length from the beginning of the first block to the end of the last in each run. Motion artifacts were estimated using the artifact detection toolbox (ART). A volume exceeding 1 mm (across rotational and translational directions) of movement between timepoints or intensity greater than 3 SD was marked as an outlier. Participants with more than \(15 \%\) outlier timepoints across any experimental run were removed (1 ASD participant).

Whole-brain first-level analyses were performed within each subject using the general linear model. The model included conditions of interest (Live, Rec-Same, RecNovel) as well as conditions not of interest (the 2.5 seconds at the beginning and end of each block and the text prompt
preceding each block). Nuisance regressors included the degree of deviation at each time point for the 6 -motion directions (roll, pitch, yaw, \(x, y, z\) ) and any outlier timepoints identified. Contrasts of interest included each condition of interest vs. fixation as well as the Live condition compared to Rec-Same and Rec-Novel separately and compared to both recorded conditions combined (Recorded). Contrasts of Rec-Novel to Rec-Same were also included and all reverse contrasts were modeled (e.g., Recorded vs. Live).

Second level random effects analyses were conducted via voxel-wise whole-brain t-tests (within and between sample) for each contrast of interest and region of interest analyses. All within-sample whole-brain tests were corrected at \(\mathrm{p}<.05\) using nonparametric permutation analyses (snpm5b). All between-group whole-brain tests are thresholded at \(\mathrm{p}<.001\) (uncorrected) with a cluster correction corresponding to \(\mathrm{p}<.05\) ( \(\mathrm{k}=192 \mathrm{~mm}^{3}\) ). Cluster size was determined using AFNI's 3dClustSim program (Cox, 1996).

Region of interests were created from previously published data using this same social interaction task (Redcay et al., 2010). These data included a sample of 16 typically developing adults ( 7 male ; 18-29 years) who were not part of the sample in the current study. Region of interests included voxels that were significantly more engaged during the Live than Recorded conditions ( \(\mathrm{p}<.05\), corrected) and intersected with a sphere ( 6 mm radius) surrounding the peak coordinate for each region identified in the group contrast of Live-Recorded (Redcay et al., 2010). Parameter estimates from the first-level analyses for each condition of interest from each subject were extracted from each of these 21 regions of interest. Repeatedmeasures ANOVAs were run for each ROI with condition (Live, Rec-Same, Rec-Novel) as the repeated measure and group (ASD, NT) as the between-subjects measure. The Greenhouse-Geisser correction was used when the assumption of sphericity was violated. For all regions showing a significant effect of condition or significant group x condition interaction, follow-up paired t-tests were conducted within each group for the contrasts Live vs. RecSame, Live vs. Rec-Novel, and Rec-Novel vs. Rec-Same.

Post-scan video coding Following data collection, videos from 9 ASD and 10 NT participants were coded for several behavioral variables, including the onset and duration of eye movements during the event periods in which the experimenter requested a response from the participant. Videos from the remainder of the participants were lost or not collected at the time of the fMRI session due to technical difficulties in video recording. The number and duration of eye movements were compared between groups and between conditions using separate two-way repeated measures ANOVAs.

\section*{Results}

Eye movements do not differ by condition or group No significant main effects or interactions were found for either the total number or duration of eye movements during the

Live, Rec-Same, and Rec-Novel conditions. These data suggest differences between conditions were not due to lowlevel differences in eye movement behavior.

Replication of previous study in new TD sample Wholebrain and ROI analyses comparing the Live and Recorded conditions revealed many similarities but also some differences from the sample published in a previous paper (Redcay et al., 2010). In general a smaller number of areas were recruited during the Live vs. Recorded contrast than reported in the previous study. Specifically, subcortical regions associated with reward and anterior temporal regions did not show differential recruitment during the Live condition. However, regions within dorsal medial prefrontal cortex (dMPFC), which did not meet threshold for significance in the 2010 paper, were significant in the current NT sample. Regions showing a greater response to Live than Recorded conditions (in both samples) included bilateral posterior STS, dorsal anterior cingulate (dACC), dorsal medial prefrontal cortex (dMPFC), thalamus, and left cerebellum (Figure 2, top).

Next, we compared parameter estimates for Live and Recorded conditions within the ROIs from the previous study using one-way paired samples t-tests ( \(p<.05\), Bonferroni corrected). Nine of the 21 regions revealed a pattern of significantly greater activation in Live as compared to Recorded conditions: dorsal anterior cingulate (dACC) \(t(14)=2.95, \quad p<.011, \quad\) anterior cingulate cortex/medial prefrontal cortex (ACC) \(t(14)=3.26, p<.006\), left cerebellum (L CBLM) \(t(14)=3.68, p<.002\), left lingual gyrus \(t(14)=3.01, p<.009\), left insula \(t(14)=4.96, p<.0001\), left middle temporal gyrus (L MT) \(t(14)=2.7, p<.017\), right insula \(t(14)=3.61, p<.003\), right posterior superior temporal sulcus (RpSTS) \(t(14)=5.65, p<.000\), right temporoparietal junction (RTPJ) \(t(14)=3.30, p<.005\), and supplementary motor area (SMA) \(t(14)=3.58, p<.003\).


Figure 2. Whole-brain random effects analyses for the contrast Live \(>\) Recorded within NT (top) and ASD (bottom) groups are displayed on a template brain in MNI space. A direct statistical comparison between groups for the Live \(>\) Recorded contrast is shown in the right panel.

Whole brain comparisons between ASD and NT Only the right pSTS showed a significantly greater response during the Live as compared to Recorded conditions in the ASD group (Figure 2, bottom). Direct statistical comparison of the Live vs. Recorded contrast between groups revealed significantly greater activation in the NT group in the left
posterior STS and bilateral cerebellum. Significantly greater activation was seen in the ASD than NT group for the LiveRecorded contrast within the left angular gyrus (AG) and right putamen; however, this effect was driven by greater deactivation in the NT group during Live conditions rather than differential engagement of these regions in ASD.


Figure 3. Region of interest analyses. The statistical parametric map for the contrast Live \(>\) Recorded from a separate group of healthy typically-developing participants (Redcay et al., 2010) is displayed on a template brain registered in MNI space. Each region showing a significant main effect of condition in the new sample (ASD and NT) is marked with a yellow circle. Response profiles for each condition (Live=blue, Recorded-Same=orange, RecordedNovel=red) for the NT (solid bar) and ASD (open bar) groups are displayed for these ROIs. Brain images and bar plots are grouped by patterns for the NT and ASD groups.

Region of interest analyses Two-way repeated measures ANOVA for each of the 21 ROIs revealed significant main effects of condition (Live, Rec-Same, Rec-Novel) in nine of the ten regions as reported above (dACC, ACC, LCBLM, LIns, RIns, RpSTS, RTPJ, SMA, LMT) and a significant group by condition interaction in one region, the anterior cingulate cortex \((F(1.6,30)=6.1, p<.008)\) (Figure 3).

Within-group condition comparisons allowed for investigation of whether regions were sensitive to the social contingency of a live interaction (i.e. Live \(>\) Recorded-Novel and Live \(>\) Recorded-Same) or to the novelty of the interaction (i.e. Live \(>\) Recorded-Same or RecordedNovel \(>\) Recorded-Same).

Salience network sensitive to online interactions in NT but novelty in ASD Within the NT group, 6 regions showed a pattern of sensitivity to Live as compared to Recorded-Novel and Recorded-Same conditions, suggesting
these regions are sensitive to online social interaction. These regions included those associated with the salience network (e.g., Seeley et al., 2007), namely the dorsal anterior cingulate (dACC), bilateral insula, and supplementary motor area (SMA), as well as regions associated with social cognition including the RpSTS extending into the RTPJ. Of these six regions, the ASD group demonstrated no difference between conditions within left insula and dACC (Figure 3, top left) and a pattern of sensitivity to novelty but not social interaction in the right insula, SMA, and ACC (Figure 3, top right). Like the NT group, the ASD group showed a significant effect of social interaction (i.e. Live \(>\) Recorded-Novel and Live \(>\) Recorded-Same) in the right \(\mathrm{pSTS} / \mathrm{RTPJ}\) (Figure 3, bottom left).

Three regions were sensitive to novelty but not live interaction specifically in both ASD and NT groups. Within the NT group, the left cerebellum, left middle temporal gyrus (MT), and left anterior STS (aSTS) demonstrated a pattern of sensitivity to novelty (i.e. Live \(>\) Recorded-Same and Recorded-Novel \(>\) RecordedSame) that was not specific to online interactions (i.e. Live is not different from Recorded-Novel). Left MT and left aSTS demonstrated a pattern consistent with novelty in the ASD group in that Recorded-Novel was greater than Recorded-Same. Further, the region within the left cerebellum showed a greater response to Live than Recorded-Same in ASD (Figure 3, bottom right).

\section*{Discussion}

The goals of the current study were to replicate previous findings using a novel interactive method and to determine whether reward, attention, and/or social-cognitive networks in autism showed a lack of sensitivity to online social interactions. We replicated the finding of a greater response to Live than Recorded conditions in many regions associated with social cognition and attention, as previously seen. Surprisingly, however, reward-related regions were not differentially sensitive to live interactions in the current sample of NT or ASD participants.

Social-cognitive areas show typical response in ASD Our hypothesis was that regions associated with social cognition, such as bilateral TPJ, posterior STS, and amygdala would not be modulated by condition in the ASD group. Some support for this hypothesis was found in the whole-brain between-group comparisons (Figure 2). The left pSTS was recruited significantly more for Live than Recorded conditions in NT than ASD groups. However, whole-brain and region of interest analyses revealed no differences between groups within right posterior superior temporal cortex (RpSTS cluster extending into RTPJ). For both NT and ASD groups this region was recruited across all three conditions but the greatest response was seen in the Live condition and no differences were found between the Recorded conditions. It is possible (and indeed likely) that group differences might have emerged if the social
interaction had required mental state inferences and/or been less predictable. Nonetheless, these findings suggest that in a simple social interaction, posterior superior temporal regions are sensitive to social contingency in both NT and ASD samples.

Salience network sensitive to live interactions in NT, but not ASD Regions within attention networks, specifically the salience network, revealed the greatest differences between groups in the region of interest analyses. We found a significantly greater response in the Live condition as compared to both Recorded conditions within regions thought to be part of a personal salience network, including bilateral insula and dorsal anterior cingulate (e.g., Seeley et al., 2007) in NT individuals but not individuals with ASD. In fact, within the dorsal anterior cingulate cortex a significant group \(x\) condition interaction revealed sensitivity to novelty, but not live interaction, in the ASD group. This salience network is engaged during tasks of empathy (Bernhardt \& Singer, 2012), affective pain (Singer et al., 2004), error processing and task-onset (Dosenbach et al., 2006) and can be identified through task-free intrinsic connectivity analyses (Seeley et al., 2007). Seeley et al., (2007) propose that these regions are important for associating incoming sensory stimuli with "markers" to aid in the decision of what to do next through interaction with other control, attention, and emotion networks. One possibility is that in NT individuals, interaction with another person in real-time provides a salient cue to enhance attention to the stimuli or task at hand via the salience network. This is analogous to theories suggesting social interactions "gate" learning (e.g., Kuhl, 2007; Meltzoff, Kuhl, Movellan, \& Sejnowski, 2009). For individuals with ASD, however, the novelty of the visual stimulus engages the salience network rather than the social contingency. These data are consistent with the proposal by Mundy and colleagues (e.g., Mundy, 2003) that atypical socialexecutive networks, of which the dorsal anterior cingulate plays a primary role, may characterize autism. Thus, these data may provide a neurobiological correlate for how social interactions are less "special" in individuals with autism. These findings also underscore the importance of examining the interaction of social and attention processes, instead of treating them as separate processes and systems.

\section*{Future Directions}

While the results are intriguing, the current study has several limitations that need to be addressed in future work. First, the interaction was highly scripted and simplified. Future studies should examine whether increasing the unpredictability or required mental state inferences within the interaction would lead to greater differences between groups within social-cognitive brain regions. Similarly, future studies should explicitly engage reward systems during real-time social interaction to help explain the discrepancy in activation of reward systems between these studies. Finally, it will be critical to examine the
developmental trajectory of atypical responses to social interactions within the salience network to determine whether reduced neural sensitivity underlies the emergence of the autistic phenotype.

\section*{Acknowledgments}

We are grateful to Dr. John Gabrieli, Lee Mavros, David Dodell-Feder, Mark J. Pearrow, Mario Kleiner, Steven Shannon, and Dr. Christina Triantafyllou for their help in the development of the face-to-face fMRI set-up, as well as the Athinoula A. Martinos Imaging center at the McGovern Institute for Brain Research. We also thank Brieana Viscomi, Ruth Ludlum, Nick Dufour, Daniel O'Young, Shannon Coveney, Neelima Wagley, Arella Mayer, and Nina Lichtenberg for their assistance with analyses and coding of the behavioral data. This project was supported by a grant from the Simons Foundation Autism Research Initiative (SFARI) awarded to R.S. and an NRSA (F32 HD 59302) awarded to E.R.

\section*{References}

Baron-Cohen, S., Leslie, a M., \& Frith, U. (1985). Does the autistic child have a "theory of mind"? Cognition, 21(1), 37-46.
Bernhardt, B. C., \& Singer, T. (2012). The neural basis of empathy. Annual Review of Neuroscience, 35, 1-23.
Brainard, D. H. (1997). The Psychophysics Toolbox. Spatial vision, 10(4), 433-6.
Chevallier, C., Kohls, G., Troiani, V., Brodkin, E. S., \& Schultz, R. T. (2012). The social motivation theory of autism. Trends in cognitive sciences, 16(4), 231-9.
Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Computers and biomedical research, an international journal, 29(3), 162-73.
Clark, H.H. \& Brennan, S.E. (1991) Grounding in Communication. In L.B. Resnick, J.M. Levine, and S.D.Teasley (Eds.), Perspectives on socially shared cognition, pp. 127-149. Washington DC: APA books.
Dawson, G., Carver, L., Meltzoff, A. N., Panagiotides, H., McPartland, J., \& Webb, S. J. (2002). Neural correlates of face and object recognition in young children with autism spectrum disorder, developmental delay, and typical development. Child development, 73(3), 700-17.
Dosenbach, N. U. F., Visscher, K. M., Palmer, E. D., Miezin, F. M., Wenger, K. K., Kang, H. C., ... Petersen, S. E. (2006). A core system for the implementation of task sets. Neuron, 50(5), 799-812.
Dufour, N., Redcay, E., Young, L., Mavros, P., Moran, J., Triantafyllou, C., Gabrieli, J.D., Saxe, R. (2012). What explains variability in brain regions associated with Theory of Mind in a large sample of neurotypical adults and adults with ASD? In N. Miyake, D. Peebles, \& R.P. Cooper (Eds.), Proceedings of the

34th Meeting of the Cognitive Science Society (pp. 312-317).
Klin, A., Jones, W., Schultz, R., \& Volkmar, F. (2003). The enactive mind, or from actions to cognition: lessons from autism. Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 358(1430), 345-60. doi:10.1098/rstb.2002.1202
Kuhl, P. K. (2007). Is speech learning "gated" by the social brain? Developmental science, 10(1), 110-20.
Meltzoff, A. N., Kuhl, P. K., Movellan, J., \& Sejnowski, T. J. (2009). Foundations for a new science of learning. Science (New York, N.Y.), 325(5938), 284-8.
Mundy, P, Sigman, M., \& Kasari, C. (1990). A longitudinal study of joint attention and language development in autistic children. Journal of autism and developmental disorders, 20(1), 115-28.
Mundy, Peter. (2003). Annotation: the neural basis of social impairments in autism: the role of the dorsal medialfrontal cortex and anterior cingulate system. Journal of child psychology and psychiatry, and allied disciplines, 44(6), 793-809.
Pelli, D. (1997). The video toolbox software for visual psychophysics: transforming numbers into movies. Spatial vision.
Pönkänen, L. M., Alhoniemi, A., Leppänen, J. M., \& Hietanen, J. K. (2011). Does it make a difference if I have an eye contact with you or with your picture? An ERP study. Social cognitive and affective neuroscience, 6(4), 486-94.
Redcay, E., Dodell-Feder, D., Pearrow, M. J., Mavros, P. L., Kleiner, M., Gabrieli, J. D. E., \& Saxe, R. (2010). Live face-to-face interaction during fMRI: a new tool for social cognitive neuroscience. NeuroImage, 50(4), 1639-47.
Schilbach, L., Timmermans, Reddy, V., Costall, A., Bente, G., Schlicht, T., \& Vogeley, K. (2012). Toward a second-person neuroscience. Behav. Brain Sciences 1-77.
Sebanz, N., Bekkering, H., \& Knoblich, G. (2006). Joint action: bodies and minds moving together. Trends in cognitive sciences, 10(2), 70-6.
Seeley, W. W., Menon, V., Schatzberg, A. F., Keller, J., Glover, G. H., Kenna, H., ... Greicius, M. D. (2007). Dissociable intrinsic connectivity networks for salience processing and executive control. The Journal of neuroscience, 27(9), 2349-56.
Senju, A., Southgate, V., White, S., \& Frith, U. (2009). Mindblind eyes: an absence of spontaneous theory of mind in Asperger syndrome. Science (New York, N.Y.), 325(5942), 883-5.

Shimada, S., \& Hiraki, K. (2006). Infant's brain response to live and televised action. NeuroImage, 32(2), 930-9.
Singer, T., Seymour, B., O’Doherty, J., Kaube, H., Dolan, R. J., \& Frith, C. D. (2004). Empathy for pain involves the affective but not sensory components of pain. Science, 303(5661), 1157-62.

\title{
Smooth Dynamics, Good Performance in Cognitive-Agent Congestion Problems
}

\author{
David Reitter, Penn State University \\ Paul Scerri, Carnegie Mellon University
}

\begin{abstract}
In a congestion game, individuals exhaust a common resource out of selfish behavior. In this scenario, drivers create traffic jams by choosing the shortest route according to their individual knowledge. They can avoid them by communicating their belief states about the traffic situation in real-time through a peer-to-peer network, assuming unlimited bandwidth. We study two potential, cognitively inspired models of human behavior: 1) categorization (quantized memorization and communication), which dampens communication and belief adoption, but leads to undesired oscillations and lower performance. 2) Instance-based blending with memory decay, which achieves good dynamics and near-optimal performance without the same bandwidth needs. We argue that this supports our hypothesis of co-adaptation of cognitive function and communicating communities.
\end{abstract}

\section*{Introduction}

In many situations, crowds of interacting human individuals share resources such as food, roads, electricity, internet bandwidth, or airtime in a conversation. Similarly there are many interesting domains that require robots or agents to simultaneously learn to utilize common resources. When the actions of one agent impact the outcomes of another agent, individual learning often leads to complex system dynamics. A canonical example of this problem is cooperative path planning(Burgard, Moors, Fox, Simmons, \& Thrun, 2000), where agents using the same routes negatively interfere with one other, but many other domains have been studied including soccer(Kalyanakrishnan, Liu, \& Stone, 2007) and markets(Tesfatsion \& Judd, 2006).

Computational-cognitive models of social behavior require a combination of cognitive architectures and multiagent design and analytics. In this paper, we investigate the effect of memory decay and instance-based learning and decision-making a system of communicating, simulated individuals.

Based on the rational assumption that human cognitive function has adapted to its environment, we hypothesize that memory function improves system dynamics in communication networks. We predict that forgetting improves, rather than impedes, performance in situations where crowds use finite common resources. This is also a core problem in multi-agent learning: one agent's behavior may impact the outcome for another agent. The collective behavior of a learning system and individual reward can vary wildly and unpredictably. Human-controlled road networks exhibit similar traffic jams, though rarely with the same catastrophic consequences the multi-agent simulations suggest. An emerg-
ing property of the human-based system is adaptation and damping. We will use a multi-agent system in which each agent is implemented as a model of (relevant) human behavior, namely learning from observations and information obtained by others, and estimating quantities (or categorizing) based on current knowledge.

In our scenario, individual models repeatedly choose roads in a road network to get them from home to work. The time taken to traverse a road is a function of the number of other agents on the road when the agent begins to traverse the road.

The first contribution of this work is to apply humaninspired instance-based learning (IBL, Gonzalez, Lerch, \& Lebiere, 2003) algorithm to the problem of multi-agent learning on the road congestion problem. The IBL model treats each trip from home to work as several instances of road segments and weighs instances based on several factors including recency when estimating time to traverse a road. We found that the agents using IBL do much better than the agents using an alternative, category-based model and about the same as agents using a communication intensive averaging model.

Second, this paper examines the dynamics of mixed human-machine systems. We demonstrate that the overall system performance is improved by even a relatively small number of the IBL agents and that there were no negative effects on either type of agent. When there were only a few IBL agents in the system, they performed relatively better than the ternary agents, but when there were many IBL agents all agents performed about the same.

Finally, the third contribution of this work is to show how changes to the underlying system, in addition to changes due to learning, impact performance. One might expect that more numerical approaches will respond to change more quickly and effectively than a learner relying on experiences. However, we found no evidence of this. Instead, IBL agents reacted very quickly and appropriately to underlying change, far better than the ternary agents. From these experiments we can see potential for using IBL in multi-agent learning settings and exploring these other settings is a key area of future focus.

Cognitive models have been combined to explain learning in team settings, primarily in a qualitative way (Sun, 2008). Reitter and Lebiere (2012) used decay in a model implemented within a cognitive architecture to show that decayed memory improves agent perfor-
mance in a foraging scenario with multiple, communicating agents. Instance-based learning within cognitive models has been shown to explain human behavior in a number of cognitive decision-making tasks (Lebiere, Wallach, \& West, 2000; Erev et al., 2010).

\section*{Scenario Framework}

The framework for the scenario (Scerri, to appear) consists of agents \(A\), places \(P\) and edges \(G\) over some number of iterations. Each agent \(a \in A\) has some place, \(p_{\text {home }} \in P\) where it starts each iteration and some place \(p_{\text {work }} \in P\) where it must get to each iteration or day. To get to \(p_{\text {work }}\) it must use edges connecting places. Individual edges \(g \in G\) connect exactly two places. The agents task is to get from \(p_{\text {home }}\) to \(p_{\text {work }}\) most quickly each iteration.

The time that it will take an agent to traverse an edge depends purely on the number of agents already on the edge when it gets to the edge. Specifically, we choose a simplified function to model limited resources that are affected by congestion: the time taken by an agent is \(10+n_{\text {already }}^{3}\), where \(n_{\text {already }}\) is the number of agents on the edge when the agent reaches it. The simulation randomizes the order the agents execute so that in one iteration an agent might be the first on the edge and have a very short travel time and another iteration it might be tenth onto the edge and have a very long travel time, even if none of the agents change their routes.

This framework has two important features. First, the agents will get very different perspectives on speed of a edge, based on exactly when they get onto the edge. Hence, either many iterations or cooperation is needed to create an accurate model. Second, busy edges heavily penalize the agents, just a few extra agents on a edge will dramatically slow the last few agents down again making cooperation important.

For experimental purposes, there are only ten different \(p_{\text {home }}\) and \(p_{\text {work }}\) for 200 agents. This makes for more interesting traffic congestion problems, with more extreme cases, and requires more coordination among the agents, but, as was shown in (Scerri, to appear) does not qualitatively change the system dynamics.

In every iteration, each agent uses a model of the graph to plan a path from \(p_{\text {home }}\) to \(p_{\text {work }}\). The agents use a standard A* algorithm (Russell, Norvig, \& Artificial Intelligence, 1995) to do the planning based on their current model of edge traversal times. Agents are risk neutral, trying to minimize expected travel time. They then execute their plan without adapting to observed conditions. At the end of an iteration, the agents can communicate about what they observed. The model the agent plans with and the information it communicates are described below.

It is assumed that each agent plans selfishly, but communicates truthfully and cooperatively. We are inter-
ested in two primary metrics. First, the average time it takes for an agent to get from \(p_{\text {home }}\) to \(p_{\text {work }}\). Second, as the agents build their models and adapt their plans to the changing models, the average transit time will change. As a secondary measure, we are interested in the change in average transit time over time.

Communication Network The agents are organized into a social network where they can only communicate directly with a small subset of the rest of the agents. Information is propagated through the network in a peer-to-peer manner. Unless otherwise noted below, we use a random network with degree 5 to connect the agents.

\section*{Model Path Planning}

The cognitive agents have to choose a path that will most quickly get them to their destination, based on experiences so far and from experiences communicated from other agents. The optimal strategy might be one that considers likely plans by others and the changes they will make, given their previous experiences. However, this is typically infeasible, and theoretically the game-theoretic Traveller's Dilemma (Basu, 1994) applies: If one agent \(A\) anticipates another agent \(B\) 's reaction, \(A\) would also anticipate \(B\) 's anticipation of \(A\) 's reaction, and so on. Rational players will end up with a poor solution (finite game), or they will be faced with a computation that does not scale.

Any accurate model of crowd cooperation needs to deal with limited communication bandwidth, learn (quickly) to achieve acceptable performance, adapt to changing network dynamics. In the following we describe two cognitively plausible models for reasoning about the road network, as well as an optimal one with high bandwidth needs. Earlier work has shown path-planning in a model in the cognitive architecture ACT-R (Reitter \& Lebiere, 2010). However, here we focus on the memory components only and keep the path planning algorithm (A* planning) constant to facilitate meaningful comparison.

Categorizing Model: Ternary We include two cognitively plausible models at the agent level. The first, ternary, forms its belief about a road segment as a category: slow, medium or fast. The model keeps, for each edge, a normalized frequency distribution of the observed categories, decayed over time. Specifically, for each edge \(e\), the model is \(M_{e}=\left\{p_{\text {slow }}, p_{\text {medium }}, p_{\text {fast }}\right\}, p_{\text {slow }}+\) \(p_{\text {medium }}+p_{\text {fast }}=1\). When an individual gets an observation of a particular category it adds \(\beta_{\text {local }}\) for a local observation and \(\beta\) for a communicated observation to the relevant \(p\) and then normalizes.

The models assume the most probable category, \(\max M\) for planning. In the following experiments, an edge in a particular category is assumed to take time 300, 156 and 12 for \(p_{\text {slow }}, p_{\text {medium }}\) and \(p_{\text {fast }}\), respectively, cor-


Figure 1: Ternary, IBL and Average models.
responding to the average time when approximately 3,7 and 11 agents also use the edge reasonable approximation of the typical expectations. When max \(M\) changes for an edge, i.e., when the individual's belief about an edge changes categories, it communicates the new category to its direct neighbors in the social network.

Instance-based learning model The second model implements a cognitively motivated aggregation mechanism that forms their beliefs. As in the ternary model, its communications are quantized and occur whenever its belief about a road changes. The same A* algorithm is used to plan paths. However, this model's estimates about the speed of each road are based on instancebased learning. IBL stores a datapoint (episode) with the speed of a road whenever it is traveled or when agents receive a communication. (A commute involves many such roads.) A speed estimate can then be derived as the average of all episodes associated with the road, weighted by the episode's activation. Activation is determined by a function that rewards experience (a large set of episodes), but discounts older information (decay). Activation has been shown to predict the availability of information in human memory (Anderson \& Lebiere, 1998).

In detail, activation of an episode \(e\) consisting of a road speed (utility) and time, \(\left\langle u_{e}, t_{e}\right\rangle\) is given as
\[
A_{e}=\left(t-t_{e}\right)^{-0.5}
\]
\(t\) is the current time. The decay exponent is the default that is empirically realistic in human experiments. Our implementation uses an highly precise approximation of the above activation function that omits to store all but the \(n\) latest episodes. If a road is represented by a series


Figure 2: Rate of belief changes.


Figure 3: A histogram of the variance in estimates per road for each of the model types.
of episodes \(R\) involving the road, then the expected speed of a road, \(U(R)\) is derived as
\[
U(R)=\frac{\sum_{e \in R} u_{e} e^{A_{e} / T}}{\sum e^{A_{e} / T}}
\]
\(T=0.25\) is a parameter (temperature). If \(R\) is empty, we assume a default speed, \(U_{\beta}\) for the road. The agent's performance is sensitive to \(U_{\beta}\), which represents a measure of pessimism (we do not optimize \(U_{\beta}\) and choose 0.0 as the most optimistic value).

Instance-based learning and the activation function have several desirable properties in our context. Activation increases during early iterations and allows the model to quickly differentiate between fast and slow roads. Activation is less affected by presentation of changes concerning frequently travelled roads.


Figure 4: Average travel times for IBL model agents as the communication network is varied.

Averaging Model The Averaging Model is included to provide a form of non-cognitive empirical ceiling: it is information-hungry, assuming that communication is free and unconstrained. It is the simplest model an agent can have of the graph is to store the average time taken by agents traversing that edge. Since the utilization of an edge will change over time, a moving average is used to keep the model updated with respect to the current situation.

The agents estimate for an edge is simply \(e_{i}^{\prime}=\alpha e_{i}+\) \((1-\alpha) o b s\), where \(e_{i}\) is the current estimate for the edge and obs is the new observation for the edge, whether communicated or observed locally. In this paper, we use \(\alpha=0.95\).

\section*{Empirical Evaluation}

In this section, we empirically examine the three models on the congestion problem described in Section 2. The evaluation is split out into three parts, with each part aimed at looking in depth at one of the hypotheses introduced in Section 1. Unless otherwise stated, for each experiment below we use the following experimental parameters.

\section*{Instance-Based Multi-Agent Learning}

The key challenge for multi-agent learning is that all the agents are simultaneously learning, making the learning environment non-stationary. Learning from instances in a non-stationary environment is not an intuitively effective technique. However, humans, who arguably use a type of IBL, are highly capable of learning in non-stationary environments. Our first experiments are aimed at looking at the performance of IBL on the congestion problem. Figure 1 compares the IBL, Ternary and Average models. Each model shows some improve-
ment over time and some initial poor performance as the space is explored. The highly communication intensive and, for a human, computationally challenging Average model and the IBL model achieve about the same final level of performance and have about the same initially poor performance. Both do better than the Ternary model in the long run, although the Ternary model more quickly finds decent solutions.

Since the IBL and Average models end with about the same performance, it is tempting to conclude that they work in about the same way. However, Figures 2 and 3 show that they actually achieve the result with quite different dynamics. Figure 2 shows the average number of agents that change the path they take from the day before. The ternary model oscillates because beliefs take some time to change. More interestingly, IBL consistently changes more than Average. IBL agents change paths substantially more often, but the net result is the same as the Average agents. It is infeasible to determine exactly what is occurring, but it appears that IBL agents switch between approximately equal paths due to the noise in their relatively sparse data, while the Average agents have aggregated more data leading to more stable choices.

Figure 3 shows a snapshot of the variance in beliefs of the agents at the end of the 200 days. Specifically, for each road segment we computed the variance in the time the agents estimate it would take to traverse that road. These variances were then discretized and presented as a histogram, with variances \(>50\) put in the last bin for clarity. The higher the variance the more the agents disagreed about how long it would take to traverse the road. Each of the three models lead to distinctly different patterns. The Ternary case often has all agents in agreement and never has large disagreements between agents due to the way beliefs cascade across the network and because the agents only allow a road estimate to have one of three values. The Average model shows slightly lower variance overall than IBL, though the IBL has many more roads with very high variance, indicating complete disagreement. It is insightful to see that better performance was had when the agents had different models of the environment, many of which must actually be wrong. We can conclude that Average and IBL achieve approximately the same results, with very different algorithms and with distinctly different internal dynamics.

Conceptually, the cognitive model (IBL) does several things differently to Ternary. To try to understand what the cause of the different behavior was, we manipulated Ternary in several different ways. First, we artificially prevent Ternary agents from changing each step to mimic the IBL's preference for reusing previous paths. Second, we decay the learning rate so later data has less effect on Ternary, to mimic the way IBL instances aggre-


Figure 5: Travel times for different arrangements (lefts) of IBL and Ternary agents. Adding roads over time (right).
gate. Finally, we change the default value for Ternary for unknown roads to match the default for IBL. We found that each of these changes improved Ternary performance, but preventing them from changing each step had the biggest effect. The qualitative equivalent of this in human decision-making would be status-quo effects or confirmation biases, while IBL's implementation more directly reflects properties of human memory.

Figure 4 shows how communication networks influence the IBL agents. Curiously, blocking communication works similarly well as communication on fully connected and random network structures. These networks share information most evenly across the team, while ring and, to a lesser degree, small worlds networks compartmentalize information into neighborhoods. Although the effect is not very big, the data represents many simulation runs so the differences are not due to noise.

We see that complete, random networks do very well. A post-hoc explanation is that these networks enable the agents to communicate freely; agents have up-todate information about congested roads. (The random networks were dense - each node has a degree of 5.) The networks without connections also do well, perhaps surprisingly so. Here, agents may adapt more slowly, and only to first-hand experience. In conjunction with the instance-based learner, this may also be a working strategy to avoid congestion. However, communication helps avoid a consistent initial spike, which we expect to be due to decision-making based on shared ground truth: everyone decides to use the fastest roads.

\section*{IBL and Ternary Models Interacting}

IBL agents can be thought of as a simple model for how human learning might occur and Ternary agents can be thought of as a reasonable, low communication agent approach to cooperative learning. Future systems are likely to have humans and agents learning together and
influencing each other. Hence, it is informative to look at what happens when IBL and Ternary agents are learning on the network at the same time. Varying the ratio of IBL to Ternary agents, we found that it takes relatively few IBL agents to give the whole system an improvement in performance. Having different types of learners in the same system not only does not hurt performance, it actually helps the weaker learners do better.

In the case of mixed Random graph networks of IBL and Ternary agents, we find that when there are only a few IBL agents they have a noticeable advantage over the Ternary agents, i.e., although they are using the same roads and are all interfering with each other, the IBL agents do relatively better. This advantage has disappeared when there are equal numbers of IBL and Ternary agents. The effect disappears smoothly as the number of IBL agents increases. If we think of IBL agents as being similar to humans and Ternary as being more like agents, this experiment hints that a small number of humans in an otherwise agent-dominated environment may do relatively better than the agents, and that they, as shown above, may improve the whole system's performance.

\section*{Disruptions}

Intuitively, learning from instances is likely to behave differently to learning moving numerical estimates when there are changes to the underlying system. Here we look at two different types of disruption to the underlying system, the addition of roads and the addition of agents, and the effects on the dynamics for each of the agent types. In the first case, one new road is randomly added every 20 days. The resulting dynamics are shown in Figure 5 over 200 days (left) and just for an early (center) and a late (right) road addition. Both Average and IBL spike dramatically as they try to exploit the new road, but then go back to their original paths after finding it to be unhelpful - for most of them because they


Figure 6: The impact of adding agents over time on average performance.
all tried it at once. The Ternary model is more robust because of the information sharing, but also takes longer to recover. Figure 6 shows the travel times as five new Ternary agents are added each 20 days, starting with 150 agents to make the result more comparable to other results. Both the Average and IBL agents jump when the new agents are added, but then smoothly improve performance. The Ternary agents are more dramatically effected by the change and do not adapt quickly. As the environment gets more congested and the original agents have built up more learning data, it is appears that IBL is more affected by the disruptions. This is unsurprising as its learning rate is effectively lower at this point.

\section*{Discussion and Future Work}

The cognitive, IBL agents benefit from a relatively simple learning model, combining a preference for wellknown roads and exploration of unseen roads. These cognitive agents can, with relatively limited communication volume, spread across the road network and efficiently use shared resources. What may be key to the cognitive agent's performance is limited sharing of knowledge: because agents do not have access to precise road utility estimates of their neighbors, and because they only receive updates when the neighbor's (quantized) beliefs change, they may arrive at heterogeneous conclusions about which roads are best. This leads them to spread out more, without sacrificing much individual performance. Under this scenario, agents do not need to misrepresent their knowledge states to their neighbors.

When combined in the same system, IBL agents and ternary agents actually helped each other rather than hurting performance. This is promising for future human-agent systems that will learn with distinctly different approaches. Notice that the agents are generally
moving to a Nash Equilibrium, where, at least according to their local models, they have no incentive to change behavior. However even if the agents do reach an equilibrium, the outcome may be far from the socially optimal solution (Hagstrom \& Abrams, 2001).

Understanding the emerging effects of cognitive decision-making in these networked simulations will allow us to spell out clear predictions to investigate crowd behavior empirically. The performance of cognitive agents that are based on empirically informed constraints of memory decay, instance-based learning and blending suggests that the mechanisms are not merely a rational adaptation to static information in the environment, but to dynamic resources and a social communication system. They enable us to maintain external, distributed memory without the devastating effects of cyclic, mutual adaptation.

\section*{References}

Anderson, J. R., \& Lebiere, C. (1998). The atomic components of thought. Mahwah, NJ: Erlbaum. Paperback.
Basu, K. (1994). The traveler's dilemma: Paradoxes of rationality in game theory. The American Economic Review, \(84(2)\), pp. 391-395.
Burgard, W., Moors, M., Fox, D., Simmons, R., \& Thrun, S. (2000). Collaborative multi-robot exploration. In IEEE Conference on Robotics and Automation, ICRA'00 (Vol. 1, pp. 476-481).
Erev, I., Ert, E., Roth, A. E., Haruvy, E., Herzog, S. M., Hau, R., et al. (2010). A choice prediction competition: Choices from experience and from description. Journal of Behavioral Decision Making, 23(1), 15-47.
Gonzalez, C., Lerch, F., \& Lebiere, C. (2003). Instancebased learning in dynamic decision making. Cognitive Science, 27(4), 591-635.
Hagstrom, J., \& Abrams, R. (2001). Characterizing braess's paradox for traffic networks. In Proc. IEEE Intelligent transportation systems (pp. 836-841).
Kalyanakrishnan, S., Liu, Y., \& Stone, P. (2007). Half field offense in robocup soccer: A multiagent reinforcement learning case study. RoboCup 2006: Robot Soccer World Cup X, 72-85.
Lebiere, C., Wallach, D., \& West, R. L. (2000). A memorybased account of the prisoner's dilemma and other \(2 \times 2\) games. In Proceedings of the International Conference on Cognitive Modeling (p. 185-193).
Reitter, D., \& Lebiere, C. (2010). A cognitive model of spatial path planning. Computational and Mathematical Organization Theory, 16(3), 220-245.
Reitter, D., \& Lebiere, C. (2012). Social cognition: Memory decay and adaptive information filtering for robust information maintenance. In Twenty-sixth AAAI Conference on Artificial Intelligence (AAAI-12).
Russell, S., Norvig, P., \& Artificial Intelligence, A. (1995). A modern approach. Artificial Intelligence. PrenticeHall, Egnlewood Cliffs.
Scerri, P. (to appear). Modulating communication to improve multi agent learning convergence. In Dynamics of information systems: Algorithmic approaches.
Sun, R. (Ed.). (2008). Cognition and Multi-Agent Interaction: From Cognitive Modeling to Social Simulation (1st ed.). Cambridge University Press. Paperback.
Tesfatsion, L., \& Judd, K. (2006). Handbook of computational economics: Agent-based computational economics (Vol. 2). North Holland.

\title{
Application of the Category Adjustment Model in Temporal, Spatial, and Abstract Magnitude at the Billions Scale
}

\author{
Ilyse Resnick (ilyse.resnick @temple.edu) \\ Department of Psychology, 1701 N. \(13^{\text {th }}\) Street \\ Philadelphia, PA 19122 USA
}

\author{
Thomas F. Shipley (tshipley@temple.edu) \\ Department of Psychology, 1701 N. \(13^{\text {th }}\) Street \\ Philadelphia, PA 19122 USA
}

\begin{abstract}
The current study examines the generalization of the Category Adjustment Model (CAM) across scales along two dimensions: time and distance. Participants were presented with geologic time and astronomical distance information either conventionally or using the hierarchical alignment model. Participants provided with hierarchically structured magnitude information for time and distances were more accurate on similar estimations at large scales than participants given the same content in a conventional manner. Patterns in event and distance estimation, along with overall group differences, are consistent with the CAM; suggesting people use hierarchically organized categorical information when estimating across scales and dimensions, and providing salient category boundary information improves estimation. Findings suggest a common representation of scale information for temporal, spatial, and abstract (numeric) magnitudes. Patterns of abstract magnitude estimations are consistent with segmented linear models of scale representation. Implications of the CAM in scale representation and the hierarchical alignment model in education are discussed.
\end{abstract}

Keywords: Category Adjustment Model; Hierarchical Alignment; Scale Representation

\section*{Introduction}

The Category Adjustment Model (CAM) is an adaptive Bayesian account for the pattern of systematic biases observed in recall of metric quantities due to category membership (Huttenlocher, Hedges, \& Prohaska, 1988; Huttenlocher, Hedges, Vevea, 2000). The CAM posits 1D, 2D, and 3D magnitudes are stored in a hierarchical combination of metric and categorical information. In the absence of lower-level information (e.g., precise metric information), people use higher-level categories to aide in estimation. Variation in estimation, therefore, occurs due to imprecision of category boundaries. Recall is biased towards the 'prototype' of the respective category. For example, when recalling the position of an object in a circular display, participants naturally divide the circle into mental quadrants and the recalled location is biased towards the center (or prototype) of the relevant quadrant (Huttenlocher, Hedges, \& Duncan, 1991).

The CAM predicts recall patterns on a range of dimensions (e.g., fatness of fish, grayness of squares, and lengths of lines (Huttenlocher, et al., 2000), events
(Huttenlocher, et al., 1988), and even social dimensions such as perception of facial expressions (Roberson, Damjanovic, \& Pilling, 2007) and judgments of gender and ethnicity (Huart, Corneille, \& Becquart, 2005)). However, there is limited research examining the CAM's predictive capability for a given dimension (such as temporal and spatial scales) across different scales (such as from human scales through to scales outside of human perception). Science education research has identified conceptual categories for spatial and temporal scales outside of human perception (e.g., Trend, 2001; Tretter, Jones, Andre, Negishi, \& Minogue, 2006), suggesting people may conceptualize magnitude information at relatively small and large temporal and spatial scales using a combination of metric and categorical information. Resnick, et al. (2012) experimentally assessed the role of categories in estimations of large temporal magnitudes. Participants who were provided with salient hierarchically organized event boundaries fostered a linear representation of events on the Geologic Time Scale compared to those who received the same information about the events without the salient hierarchical structure. Aligned with the CAM, this finding suggests the use of hierarchically organized category boundaries in the representation of events at larger temporal scales.

The current study aims to add to this relatively sparse literature by examining the generalization of the CAM across scales and dimensions. Two main objectives are to replicate research on memory for large temporal magnitudes (geologic time), and extend research to another dimension: space. Astronomical distance (a spatial magnitude at a large scale) was chosen for two reasons. There is already extensive research on CAM and spatial distance; demonstrating spatial distances at familiar scales are stored in a combination of metric and categorical information (e.g., Huttenlocher, et al., 1991; Huttenlocher, et al., 2000). Additionally, while the precise nature of the relationship is unclear, there is a systematic relationship between time and distance (e.g., Clark, 1973; Gentner, 2001), suggesting that time and distance at human scales are represented and estimated in the same way. Thus, if temporal and spatial dimensions across familiar and relatively larger scales are represented in a similar way, an analogous pattern of memory performance would be expected.

Relevant to the current study, the CAM makes two predictions. First, estimations of temporal and spatial magnitude should be biased towards the prototype of each event or object's category. There is evidence that suggests people with a moderate amount of knowledge regarding geologic time (e.g., in-service science teachers), divide the Geologic Time Scale (4.6 billion years) into three categories: 'extremely ancient', 'less ancient', and 'geologically recent' (Trend, 2001). It is beyond the scope of this paper to identify and characterize the types of categories used by novices to represent large temporal and spatial magnitudes. Rather, the current study will assess if providing salient internal structure of magnitude relations, through the use of the hierarchical alignment activity (Resnick, et al., 2012), improves estimation of large temporal and spatial magnitudes. In this way, the current study examines a second prediction of the CAM: people with salient internal structure of magnitude relations within hierarchically organized category boundaries should have more linear representations of magnitude compared to those who do not.

The current study also examines patterns of abstract (numeric) magnitude estimation (i.e., not content-specific) at the same scale as geologic time and astronomical distance. One common property of time and distance is they are both one-dimensional vectors (e.g., Clark, 1973; Gentner, 2001), as is abstract magnitude. Similar patterns in overestimation of small magnitudes and underestimation of large magnitudes are found with estimations of geologic events (Libarkin, Kurdziel, \& Anderson, 2007), astronomical distance (Miller \& Brewer, 2010), and abstract magnitude (Siegler \& Opfer, 2003). Studies of abstract magnitude suggest this pattern of errors may be due to compressive effects of unfamiliar magnitudes on a mental number line (see Barth \& Paladino, 2011 and Opfer, Siegler \& Young, 2011 for discussion of competing models). Consistent with the scale of geologic time and astronomical distances, the current study will examine abstract magnitude at two scales: million and billion. Number word frequency studies suggest that there may be differences in the representation of the million and billion scales, because the frequency of occurrence influences the structure of representation and the number 'million' appears more frequently than 'billion' (e.g., Dehaene \& Mehler, 1992). Thus, sampling from across the million and billion scales may reveal potential representational differences between the two scales.

While research has not explicitly examined the CAM in abstract (numeric) magnitude representation, there are a number of studies that look at the role of the subjective categorization of numbers in estimation (e.g., Laski \& Siegler, 2007; Mix, Huttenlocher, \& Levine, 2002; Siegler \& Robinson, 1982). Findings suggest that individual numbers can serve both as their own distinct category (a specific quantity of something) as well as part of a set of numbers (e.g., 'small' versus 'big' numbers) (Mix, et al., 2002). Further, children who spread numbers evenly across
group dimensions were more accurate on an abstract magnitude task than those who grouped more numbers into one 'big' category (Siegler \& Robinson, 1982). The current study will examine if the presentation of salient category boundaries in specific dimensions transfers to abstract magnitude representation. Because participants will be working with magnitudes with temporal and spatial content, transfer to abstract magnitude should occur. If the CAM accounts for abstract magnitude at large scales, similar patterns of estimation are expected for geologic time, astronomical distance, and abstract magnitude.

\section*{Methods}

\section*{Participants}

Forty participants were recruited from an undergraduate psychology experiment pool (20 in the hierarchical (experimental) group and 20 in the conventional (control) group). The demographics of the participants were consistent with a large urban American university.

Hierarchical Design In the hierarchical alignment condition, participants completed the same hierarchical alignment activity developed by Resnick and colleagues (2012), which is based on the progressive alignment model (Kotovsky and Gentner, 1996; Thompson \& Opfer, 2010). Participants made ten separate time lines, aligning time to a horizontal one meter space. They began with a familiar personal time scale, working through different historic and geologic time lines, up to the full Geologic Time Scale. For each time line, participants were given a partially completed time line, and were required to label the time line's length (in years) and locate where all previous time lines would begin on the current time line (see Figure 1).

Hierarchical organization highlights how each temporal scale is related to the other scales. Practice mapping magnitude relations across scales provides internal structure of magnitude relations within each scale. Thus, the hierarchical organization helps to populate each scale with additional categorical boundary information.

The current study developed a new analog version of the temporal hierarchical alignment activity for spatial distances (see Table 1 and Figure 1). For the hierarchical alignment of spatial distances, participants align ten increasingly larger scales of distance to a one meter space, beginning with a familiar distance. The hierarchical alignment condition takes approximately 45 minutes to complete.

Conventional Design The study sought to contrast the intervention with a realistic training program similar to one that might be used to instruct students in a classroom on these scales. Common pedagogical approaches to teaching geologic time (Libarkin, et al., 2007) and astronomical distances (Miller \& Brewer, 2010) are to create spatial analogies, such as placing events or objects in the correct sequence. Participants completed ten separate puzzles, placing the events/objects into the correct sequence. The
puzzles were made up of pieces of paper, half containing magnitude information and half with the respective category information. Participants were required to match the magnitude information with the corresponding category information for each scale, and place the scales in the correct sequence. The first puzzle represented the first temporal/spatial scale (see Table 1), with each puzzle representing an increased amount of magnitude. The tenth and final puzzle represented all of geologic time/distance to Makemake. The conventional condition took approximately 45 minutes to complete.

The conventional and hierarchical conditions were aligned on the following properties: number of scales, number of times participant identifies each scale (i.e., the first scale is identified ten times; the last scale is identified once), progressive increase of magnitude, information provided about each event/object, and total length of time on task. Thus, the only difference between conditions was the hierarchical alignment of scale information.

One potential difference between the temporal and spatial information was identified. Participants are likely to be familiar with thinking about temporal scales extending back hundreds of years ago; learning about recent human history is common. However, participants may not have the same level of familiarity with conceptualizing the vertical nature of the spatial scales. Because it is likely people have more experience traveling parallel to Earth's surface, or 'horizontally', as opposed to traveling vertically away from Earth's surface, we used this horizontal experience as an initial introduction of the vertical scale. As a way to familiarize participants with the vertical scale, a horizontal map was presented for each of the first three scales in both the hierarchical and conventional conditions. The maps showed an eleven, fifty-two, and four-hundred mile radius extending out from the university where the study took place. To engage the participants in grounding this scale to their personal experience, participants were asked if they had been anywhere on that radius or if they were familiar with the area. Because participants likely do not have experience thinking about larger temporal scales, no map was provided for the remainder of the spatial scales.

Table 1. List of Temporal and Spatial Scales, including category names and magnitude information
\begin{tabular}{|l|l|l|l|}
\hline Temporal Scale & \multicolumn{1}{|c|}{ Years } & \multicolumn{1}{c|}{ Spatial Scale } & \multicolumn{1}{c|}{ Miles } \\
\hline Personal & 20 & Troposphere & 11 \\
\hline Human Lifespan & 75 & Middle Atmosphere & 52 \\
\hline American History & 519 & Exosphere & 400 \\
\hline Recorded History & 5,512 & Inner Van Allen Radiation Belt & 6,000 \\
\hline Human Evolution & \(6,000,000\) & 3753 Cruithne (quasi-satellite) & \(8,450,000\) \\
\hline Cenozoic & \(65,000,000\) & Mercury & \(57,000,000\) \\
\hline Phanerozoic & \(542,000,000\) & Saturn & \(777,000,000\) \\
\hline Proterozoic & \(2,500,000,000\) & Neptune & \(2,700,000,000\) \\
\hline Archean & \(3,800,000,000\) & Pluto & \(3,580,000,000\) \\
\hline Hadean & \(4,600,000,000\) & Makemake (dwarf planet) & \(4,800,000,000\) \\
\hline
\end{tabular}


Figure 1. Example of a temporal and spatial number line at the thousands scale in the hierarchical condition. Note: the three previous temporal and spatial number lines are located relative to the current scale.

Procedure In a two-hour session, participants were presented with information about time and distance, with both presented as either hierarchically or conventionally (~90 minutes). Participants across conditions then completed the same assessment measures ( \(\sim 30\) minutes).

Measures A series of line estimation tasks were developed to assess participants' representations of geologic time, astronomical distances, and abstract (numeric) magnitude. Line estimation tasks are commonly used to assess mental scaling of abstract magnitude (e.g., Ebersbach, et al., 2008; Siegler \& Booth, 2004).

To measure representation of events on the Geologic Time Scale, an item from the Geoscience Concept Inventory (GCI), a reliable and valid instrument measuring a range of geoscience knowledge (Libarkin, et al., 2005), was adapted as a number line task. The GCI item presents participants with five time lines, with the following four geologic events placed in different locations: life appears, dinosaurs appear, dinosaurs disappear, and humans appear. Participants are required to choose the correct linear representation, with the other four time lines representing common misconceptions. In order to capture more variance in participants' representations, the GCI item was adapted so that participants were given a blank time line (anchored by 'present day' and 'Earth forms'), and asked to locate the same four events as used in the GCI item.

To measure representation of objects on an astronomical scale, an item was developed as an analog to the geologic event time line described above. Here, participants were presented with a blank number line (anchored by 'Earth's surface' and 'Makemake'), and asked to locate four objects on the same scale as on the event time line: Pluto, Mars, Mercury, and Cruithne.

To measure representation of abstract (numeric) magnitude (not content specific) a series of line estimation
tasks were given. Participants were given a sentence stating when/where an event/object was, and then asked to locate that magnitude on the number line (e.g., "Venus is 26 million miles away from Earth. Please draw on the line provided where Venus is located."). These items were framed in terms of objects and events to match the form of the other experimental measures. These estimations are considered estimations of abstract magnitude because the participants are explicitly given a magnitude to place on the number line; no recall is required. The questions provide the numerical values and ask for an estimation of the appropriate location on a spatial scale. To assess representations of the millions and billions scale, participants were asked to estimate two 'events' and two 'objects' on a 4.6 billion scale, and two 'events' and two 'objects' on a 542 million scale.

\section*{Results}

Participants in the hierarchical condition were more accurate overall on the event time line estimation task \((\mathrm{t}(38)=2.67\), \(\mathrm{p}=.01\) ) and the object distance task ( \(\mathrm{t}(38)=3.02, \mathrm{p}=.01\) ) compared with participants from the conventional condition. On both tasks, this effect is driven primarily by the estimation of the \(2^{\text {nd }}\) and \(3^{\text {rd }}\) events/objects. Participants across conditions performed similarly when placing the \(1^{\text {st }}\) (life appears/Pluto) and \(4^{\text {th }}\) (humans appear/Cruithne) events/objects on the number line ( \(p>.05\) ). However, participants in the hierarchical condition were significantly more accurate when placing the \(2^{\text {nd }}\) (dinosaurs appear) \((\mathrm{t}(38)=2.79, \mathrm{p}=.01)\) and \(3^{\text {rd }}\) (dinosaurs disappear) \((\mathrm{t}(38)=2.53, \mathrm{p}=.02)\) events on the time line, and the \(2^{\text {nd }}\) (Mars) \((\mathrm{t}(38)=3.38, \mathrm{p}<.01)\) and \(3^{\text {rd }}\) (Mercury) \((\mathrm{t}(38)=2.79\), \(\mathrm{p}=.01\) ) objects on the number line compared to the conventional condition (see Figure 2).

Performance across groups on the object distance estimation task was significantly more accurate than on the event time line estimation task \((\mathrm{t}(39)=2.85, \mathrm{p}=.01)\).

The eight abstract (numeric) magnitude line estimation tasks were highly correlated (rs > .529, \(\mathrm{p}<.01\) ) and had strong internal consistency (Cronbach's alpha=.94). There was no difference in performance when estimating abstract magnitude when estimations were temporally framed compared with spatially framed ( \(\mathrm{p}>.05\) ). Given the high correlations, strong internal consistency, and no performance differences between items that were temporally and spatially framed; a single abstract magnitude scale was created. Participants from the hierarchical condition were significantly more accurate on the abstract magnitude scale ( \(\mu\) error \(=11.50 \mathrm{~mm}\) ) than the conventional condition ( \(\mu\) error \(=30.14 \mathrm{~mm})(\mathrm{t}(25.38)=2.58, \mathrm{p}=.02)\). That the participants from the hierarchical condition are more accurate on the abstract magnitude scale than participants from the conventional condition is consistent for estimations on both the million and billion scales. Across conditions, participants were significantly more accurate when making estimations on the millions scale ( \(\mu\) error \(=14.73 \mathrm{~mm}\) )
compared with estimations on the billions scale ( \(\mu\) error \(=\) \(26 \mathrm{~mm})(\mathrm{t}(39)=3.45, \mathrm{p}<.001)\).


Figure 2. Average error (mm) for hierarchical and conventional conditions on the event and object line estimation tasks. For the event/object line estimation tasks, \(1^{\text {st }}=\) Life/Pluto, \(2^{\text {nd }}=\) Dinosaurs appear/Mars, \(3^{\text {rd }}=\) Dinosaurs disappear/Mercury, and \(4^{\text {th }}=\) Humans appear/Cruithne, respectively.

\section*{Discussion}

The current study successfully replicated the Resnick, et al. (2012) findings; participants provided with hierarchically structured event information were more accurate on event time line estimations than participants given the same content in a conventional manner. Here we found a similar result for astronomical distances and abstract (numeric) magnitude. These findings are aligned with the CAM, suggesting people use hierarchically organized categorical information when making estimations across scales and across dimensions; and that providing people with more salient category boundary information improves estimation.

In both the event time line and object distance tasks, participants across conditions were relatively accurate in identifying the location of the \(1^{\text {st }}\) (Life appears/Pluto) and \(4^{\text {th }}\) (Humans appear/Cruithne) events/objects (respectively). This may be because the \(1^{\text {st }}\) and \(4^{\text {th }}\) events/objects are anchored by the relatively close flanks of the number line itself ('top' and 'bottom'), whereas the \(2^{\text {nd }}\) (Dinosaurs appear/Mars) and \(3^{\text {rd }}\) (Dinosaurs disappear/Mercury) events/objects (respectively) may not be naturally perceived in these same salient categories; they are located 'somewhere in between'. Consistent with this interpretation, participants from the conventional condition demonstrate more bias in estimation towards the center of the number line than the participants from the hierarchical condition (as seen in the overestimation of the \(2^{\text {nd }}\) and \(3^{\text {rd }}\) events/objects). This finding is aligned with the three-category representation of geologic time advocated by Trend (2001), as well as predictions of biases towards the middle of these categories by the CAM. However, more research is needed to further identify and characterize categories used in the representation of geologic time and astronomical distances.

Participants across conditions were significantly more accurate on the object distance task ( \(\mu\) error \(=33.76 \mathrm{~mm}\) ) than the event time line task ( \(\mu\) error \(=45.45 \mathrm{~mm}\) ) ( \(\mathrm{t}(39\) ) \(=2.85, \mathrm{p}=.01\) ). Participants across conditions also were more accurate on the abstract (numeric) magnitude task ( \(\mu\) error \(=22.7\) ) compared to the event time line task \((\mathrm{t}(39)=5.55, \mathrm{p}<.001)\) and the object distance task \((\mathrm{t}(39)=2.96, \mathrm{p}=.01)\). One explanation for this pattern of differences in performance is that temporal, spatial, and abstract magnitudes are represented differently (see Agrillo, Ranpura, \& Butterworth, 2010 and Walsh, 2003 for a discussion on a general magnitude system). Alternatively, it may be the case that temporal, spatial, and abstract magnitudes are all represented in a similar way, but preexisting knowledge (and misconceptions) bias the subjective categories people use to make estimations. For example, consistent with participants being better at the object distance task compared to the event time line task, that geologic time is often neglected in the classroom (Dodick, 2007; Trend, 2001) and learning about the solar system is commonplace, it seems likely participants did have more knowledge of the solar system than geologic time. Related, the first three base analogies (tens, hundreds, thousands) may be differentially familiar to participants for temporal and spatial magnitudes. While temporal and spatial scales of magnitude were aligned, participants may be more familiar with traveling tens, hundred, and even thousands of miles; whereas participants could have only personally experienced years at the tens scale (no participants were over one hundred years old). Alternatively, mapping the vertical distances onto a horizontal map, and not having an analogous temporal activity, may have contributed to the observed domain differences. Future research should examine unfamiliar scales, both in content and magnitude. One may use an unfamiliar solar system, which would have a different time-course and different celestial objects.

Findings from the abstract magnitude task are consistent with the segmented number line model of scale representation (Ebersbach, et al., 2008; Landy, Silbert, \& Goldin, 2012). The segmented linear model posits separate linear functions for familiar versus unfamiliar magnitudes when estimated magnitude is plotted against actual magnitude. Ebersbach and colleagues (2008) found young children had a fairly accurate linear slope for smaller, familiar numbers, and a separate shallower linear slope for larger, unfamiliar numbers. While there were not enough estimations in the current study to carefully characterize the slope function, participants across conditions had a more accurate linear slope for estimations made on the million scale, and, while still linear, were significantly less accurate on estimation on the billion scale (overestimation). More research is needed examining estimations at large scales for detailed modeling of these slope functions.

That the hierarchical condition transferred to estimations about abstract magnitudes, suggests that people use categorical information when making these types of estimations. While there are some studies that look at the
subjective categorization of numbers (Laski \& Siegler, 2007; Mix, Huttenlocher, \& Levine, 2002; Siegler \& Robinson, 1982), there has not been previous work mapping the CAM onto number line estimations and scale representation. While more direct and explicit research is needed, we speculate that the CAM could serve as a unifying model for currently competing theories (e.g., logarithmic-to-linear, power function with anchor points, segmented linear). Category boundaries may serve as distinct anchor points, with adults possessing more precise categories (at the individual numbers level) compared with children. Whereas young children may have many numbers in one "big" or "unfamiliar" category, adults may possess counting strategies for numbers within "unfamiliar" scales. Thus, the CAM offers an account for the overestimation of unfamiliar magnitudes that maintain linearity within the scale. More extensive research is needed to identify types of categories used in scale representation to see if a CAM can predict the changing pattern of bias in number line estimations that occurs with development.

An implication of the current findings is the hierarchical alignment model is an effective way to teach about scales outside of human perception. Understanding scale information is important, as fundamental concepts in many disciplines require understanding of scales outside of human experience. "Size and scale" have been identified by the new National Research Council Framework for K-12 Science Education (2011) and the Benchmarks for Science Literacy (AAAS, 1993) as a fundamental and unifying theme of science education. Having a linear representation of scale is predictive of performance on a range of standardized tests in mathematics (Siegler \& Booth, 2004). Unfortunately, understanding large scales is difficult (e.g., Libarkin, et al., 2005; Tretter, et al., 2006). Undergraduate students, even those in science, technology, engineering, and mathematics majors, have difficulty mastering concepts of size and scale (Drane et al., 2008). While people are fairly accurate on identifying correct sequences, they fail to understand the magnitude between the events (Tretter, et al., 2006) and objects (Jones, et al., 2008). By providing a salient internal structure of magnitude boundaries, the hierarchical alignment activity may be an effective classroom tool to help foster a linear representation of scales like geologic time and astronomical distance.

\section*{Acknowledgments}

This research was funded in part from National Science Foundation Grant SBE-0541957 and National Science Foundation Grant SBE-1041707 that both support the Spatial Intelligence and Learning Center (SILC).

\section*{References}

Agrillo, C., Ranpura, A., and Butterworth, B. (2010). Time and numerosity estimation are independent: behavioural evidence for two different systems using a conflict paradigm. Cognitive Neuroscience, 1

American Association for the Advancement of Science (AAAS). (1993). Benchmarks for science literacy. NewYork: Oxford University Press.
Barth, H. C., \& Paladino, A. M. (2011). The development of numerical estimation: evidence against a representational shift. Developmental Science, 14.
Clark, H.H. (1973). Space, time, semantics, and the child. In T.E. Moore (Ed.), Cognitive development and the acquisition of language. New York: Academic Press.
Dehaene, S. \& Mehler, J. (1992). Cross-linguistic regularities in the frequency of number words. Cognition: International Journal of Cognitive Science, 43(1)
Dodick, J. (2007). Understanding evolutionary change within the framework of geological time. McGill Journal of Education, 42(2).
Drane, D., Swarat, S., Hersam, M., Light, G., \& Mason, T. (2008). An evaluation of the efficacy and transferability of a nanoscience module. Journal of Nano Education.
Ebersbach, M., Luwel, K., Frick, A., Onghena, P., \& Verschaffel, L. (2008). The relationship between the shape of the mental number line and familiarity with numbers in 5- to 9 -year-old children: Evidence for a segmented linear model. Journal of Experimental Child Psychology, 99.
Huart, J., Corneille, O., \& Becquart, E. (2005). Face-based categorization, context-based categorization, and distortions in the recollections of gender ambiguous faces. Journal of Experimental Social Psychology, 41.
Huttenlocher, J., Hedges, L. V., \& Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. Psychological Review, 98.
Huttenlocher, J., Hedges, L. V., \& Prohaska, V. (1988). Hierarchical organization in ordered domains: Estimating the dates of events. Psychological Review, 95.
Huttenlocher, J., Hedges, L. V., \& Vevea, J. L. (2000). Why do categories affect stimulus judgment? Journal of Experimental Psychology: General, 129.
Gentner, D. (2001). Spatial metaphors in temporal reasoning. In M. Gattis (Ed.), Spatial schemas and abstract thought. Cambridge, MA: MIT Press.
Jones, M. G., Tretter, T., Taylor, A., \& Oppewal, T. (2008). Experienced and novice teachers' concepts of spatial scale. International Journal of Science Education, 30(3).
Kotovsky, L., \& Gentner, D. (1996). Comparison and categorization in the development of relational similarity. Child Development, 67.
Landy, D., Silbert, N., Goldin, A. (2012). Getting off at the end of the line: the estimation of large numbers. 2012 Annual Meeting of the Cognitive Science Society Conference Proceedings.
Laski, E. V., \& Siegler, R. S. (2007). Is 27 a big number? Correlational and causal connections among numerical categorization, number line estimation, and numerical magnitude comparison. Child Development, 76.
Libarkin, J.C., Anderson, S.W., Dahl, J., Beilfuss, M., \& Boone, W. (2005). Qualitative Analysis of College

Students' Ideas about the Earth: Interviews and OpenEnded Questionnaires. J. of Geoscience Edu., 53(1).
Libarkin, J.C., Kurdziel, J.P. \& Anderson, S.W. (2007). College student conceptions of geological time and the disconnect between ordering and scale. Journal of Geoscience Education, 55: 413-422.
Miller \& Brewer. (2010). Misconceptions of Astronomical Distances. International Journal of Science Education, 32(12).
Mix, K. S., Huttenlocher, J., \& Levine, S. C. (2002). Quantitative development in infancy and early childhood. New York: Oxford Press.
National Research Council. (2011). A Framework for K-12 Science Education. Committee on a Conceptual Framework for New K-12 Science Education Standards. Board on Science Education, DBASSE. Washington,DC: The National Academies Press.
Opfer, J. E., Siegler, R. S., \& Young, C. J. (2011). The powers of noise-fitting: Reply to Barth and Paladino. Developmental Science, 14.
Resnick, I., Shipley, T., Newcombe, N., Massey, C., Wills, T. (2012). Examining the Representation and Understanding of Large Magnitudes Using the Hierarchical Alignment model of Analogical Reasoning, 2012 Annual Meeting of the Cognitive Science Society Conference Proceedings.
Roberson, D., Damjanovic, L. \& Pilling, M. (2007). Categorical Perception of Facial Expressions: Evidence for a 'Category Adjustment' model. Memory \& Cognition, 35
Siegler, R. S., \& Booth, J. L. (2004). Development of numerical estimation in young children. Child Development, 75.
Siegler, R. S., \& Opfer, J. E. (2003). The development of numerical estimation: Evidence for multiple representations of numerical quantity. Psychological Science, 14.
Siegler, R. S., \& Robinson, M. (1982). The development of numerical understandings. In H. W. Reese \& L. P. Lipsitt (Eds.), Advances in child development and behavior: Vol. 16. New York: Academic Press.

Thompson, C., \& Opfer, J. (2010). How 15 hundred is like 15 cherries: Effect of progressive alignment on representational changes in numerical cognition. Child Development, 81(6)
Trend, R.D. 2001, Deep Time Framework: a preliminary study of UK primary teachers' conceptions of geological time and perceptions of geoscience. Journal of Research in Science Teaching, 38(2)
Tretter, T. R., Jones, M. G., Andre, T., Negishi, A., \& Minogue, J. (2006). Conceptual boundaries and distances: Students' and experts' concepts of the scale of scientific phenomena. Journal of Research in Science Teaching, 43(3).
Walsh, V. (2003). A theory of magnitude: common cortical metrics of time, space and quantity. TRENDS in Cognitive Sciences, 7(11)

\title{
English Words Are Processed Like Objects
}

\author{
Gavin F. Revie (G.F.Revie@dundee.ac.uk) \\ Yuki Kamide (Y.Kamide@dundee.ac.uk) \\ School of Psychology, University of Dundee, Dundee, UK, DD1 4HN
}

\begin{abstract}
This experiment sought to explore the theory that familiar English words are processed similarly to objects. To do this, we looked for object-based attentional facilitation where cues in a different location to the target still facilitate target detection as long as they are inside the same object. Participants were shown two English words in an array, and cues and targets were embedded inside them. Reaction times for target detection were measured. It was found that in horizontally presented English words, cues presented in a different location to the target still facilitated target detection if they occurred inside the same word This was not the case for vertically oriented words. It was concluded that familiar words in a familiar orientation are indeed processed in a similar way to objects. These findings may be indicative that the cortical networks that evolved for object processing are also involved in the processing of words.
\end{abstract}

Keywords: object based; attention; reading;

\section*{Introduction}

It has been long understood that humans are capable of focussing their visual attention in one place in preference over another. This is commonly described as spatial attention. However, humans also have the capacity to allocate their attention to a particular object regardless of where it happens to be located (Blake \& Sekuler, 2006). This is known as object based attention. In day to day scenarios these two types of visual attention will interact and overlap. However, how object based attention contributes to the process of reading (if at all) is not well understood.
Egly, Driver, \& Rafal, (1994) conducted a study on attention within objects. Participants saw a \(2 \times 2\) array with a fixation point in the middle. The array contained 2 rectangular shapes which each spanned two of the quadrants. These shapes could be oriented with either both of them vertical or both of them horizontal. Within the individual cells of the array very brief cues and targets were presented. Participants had to detect the onset of a grey target square following presentation of a 100 ms brightening cue. They had 3 possible cuing conditions in their experiment. In the valid condition, the cue and the target would appear in the same location. In the invalid within-object condition, the cue and the target were in different locations, but still within the same object. In the invalid different-object condition, the cue and the
target were in different locations and within different objects. The targets in both of the invalid conditions were the same distance from their cue, and they were equally often oriented vertically as horizontally. The corner to corner diagonal separation of the cue and target was not used due to non-equal distance.

Egly et al. (1994) successfully manipulated the deployment of attention. The valid trial types were consistently responded to the fastest of all trial types, indicating that the cues were successful in heightening attention at their location. The crucial point came in the comparison of the 2 invalid trial types. Despite being the same distance away from the cue and subject to the same variations of orientation, the within-object invalid trials were responded to faster than the different-object invalid trials. This suggests that an advantage was conferred upon the invalid-within object trials simply due to the presence of a shape containing both cue and target locations. This has been described as "Object Based Attention" - that attending to a cue within an object will heighten attention deployment to the whole object, including non-cued locations. More recently Luo, Lupiáñez, Funes, \& Fu (2011) replicated these findings, and found that these object-based effects could be expected to be present even at very short stimulus onset asynchronies. They also highlighted problems in using cues and targets which contain implicit spatial information - something which was deliberately avoided in this study.
Li \& Logan (2008) sought to explore how object based attention relates to reading. They performed an almost direct replication of Egly et al. (1994), but replaced the shapes with 2-character Chinese words. The words could be oriented either horizontally or vertically in their experiment, following the rules in the Chinese writing system. The study was a target detection task with three conditions of cue-target relationship. The cues could be valid, invalid but within the same word, and invalid and located in a different word. Replicating Egly et al. (1994), Li and Logan (2008) found that valid trials were responded to fastest. Comparing the invalid trials it was found that invalid targets occurring within the same word as the cue were responded to faster than invalid targets occurring within a different word from the cue. This successful replication of Egly et al. (1994) and may be taken as evidence that words are treated like objects inasmuch that cues falling within a word measurably facilitate target detection
elsewhere within that word, presumably through elevated attentional deployment.
Li and Logan have demonstrated that the visual contiguity of shapes can be "simulated" by the abstract lexical contiguity of words. There were no physical connections between the characters in their array, and yet the participants clearly treated them as in some way connected. One way of explaining this is that the participants were treating the 2 -character words as if they were a single object. However, their findings might not be easily translatable into English reading processes. Chinese is both more visually dense than English, and more spatially plastic in that the character meaning is not necessarily extracted in a left-toright fashion. Traditionally, it could also be written legally both left-to-right and top-to-bottom, although that has become much rarer. As a consequence the importance of serial order and direction could be said to be comparatively lower than in English, whereas the importance of what lexical groups the symbols form could be said to be greater. This may lend itself well to an object based decoding strategy. Would the within-word benefit carry over to English? We devised a study to try and answer that question. In our study, we stuck as close as possible to the method employed by Li and Logan. There is no English equivalent to the many 2-character words available in Chinese, so in our experiment we transitioned to using 4character English words. Each quadrant of the 2x2 array would contain 2 characters. In our experiment, the words were presented either horizontally or vertically. In particular it would be interesting to see what effect the more linear and less dense script of English has on the effects found in comparison with Chinese. Can an object based account explain reading single words generally, or is it only a special-case phenomenon?

\section*{Method}

\section*{Participants}

The participants in this study were 25 female and 7 male students from the University of Dundee. They were paid in course credits for their time. Their ages ranged from 17 to 40. All participants were fluent in English. This experiment utilized a within subjects design so all participants were exposed to all conditions of the stimuli. An additional 4 participants were tested but their data was not included due to abnormally high error rates.

\section*{Apparatus}

Stimuli were presented through an 18 " monitor running at 100 Hz and detection responses were recorded on a gamepad, with the response button pressed by the dominant hand. An SR Research

Eyelink-1000 desk-based eye tracker recorded monocular eye position at 1000 Hz . A deskmounted chinrest kept participants' eyes 60 cm from the screen and both their peripheral vision and vision in their non-dominant eye were eliminated through blinkered spectacles.

\section*{Stimuli}

288 4-character words with a lemma frequency of at least 200 per 16 million were selected using the CELEX word database (Baayen, Piepenbrock, \& Gulikers, 1995). The 288 stimuli words were used to create 144 test arrays containing 2 words each. Each of these arrays was used only once per subject. The letters were printed lowercase in black, 46 point Monaco. Targets were background colour patches that were red and cues were background colour patches that were green. Cues and targets would always span 2 characters of the 4 character word in which they occurred. Stimuli arrays were assembled from several bitmaps and controlled using a variable grid. Individual bitmaps were created for each word, the fixation cross, the cue and the target.

\section*{Design}

The experiment consisted of an individually randomized sequence of 144 trials: 72 valid trials (cue and target were the same two letters), 24 invalid-within trials (cue and target were different letter pairs in the same word), 24 invalid-between trials (cue and target were in different words but never in diagonally opposed letter pairs, to maintain equidistance between cue and target across all invalid trials; see Figure 1), and 20 catch trials (no target appeared). Half of the arrays were horizontally oriented and half were vertically oriented for each subject. All stimulus arrays appeared only once per subject. In the horizontal version of the experiment the arrays were configured in the traditional left-to-right writing mode of English. In the vertical version of the experiment, the array was configured in a more novel top-to-bottom writing mode

\section*{Task and Procedure}

After giving informed consent the eye tracker was calibrated on the participant's dominant eye, determined via majority result from the Miles, Porta, and Camera tests (Roth, Lora, \& Heilman, 1992). Peripheral vision and non-dominant eye were occluded with blinkered spectacles. Participants were informed that they would be periodically asked about the last array they had seen in order to highlight the importance of actually reading the words onscreen.

Figure 1 gives a schematic representation of trial events. The start array for each trial contained two
words. These were presented for 1500 ms , followed by an additional fixation cross for 300 ms . Participants were told to read the words silently and then fixate the cross. The eye tracker was used to ensure participants were indeed looking at the fixation cross. A green colour patch was flashed behind the first or last 2 letters of one of the words for 100 ms to cue attention to this location. Following a further 100 ms of displaying the array with words and the fixation cross but no cue or target, a red target would appear under the first or last two characters one of the words. The trial proceeded only if fixation was within the region in which the cues and targets would appear during this cue-target onset asynchrony, or else an error
message appeared and the trial was discarded. Participants were instructed to press the response button as soon as they were aware of the appearance of the target, but to avoid pressing the response button when there was no target. Thus this was a simple go/no-go task. The time from the target onset to the button press was the reaction time (RT). Participants were instructed to respond as fast as possible to each target and to refrain from responding in catch trials. Responses were issued via a gamepad held in front of the participant, as close to their midline as possible. The response button was pressed with the dominant hand. If no response was issued a new trial started after 3000 ms.


Figure 1. Trial sequence, illustrating an invalid-between word trial in the horizontal condition, and an invalidwithin word trial in the vertical condition. Not drawn to scale

\section*{Results}

\section*{Performance Data}

The miss rate for present targets was extremely low, less than 5\%. Because of this false alarm rates on catch trials (which tended to be higher) were used as a criterion to remove underperforming subjects. Any participants who achieved less than \(75 \%\) correct on catch trials were removed from the data. 4 participants were removed from the data for this reason. This left 32 participants who responded correctly to catch trials \(86 \%\) of the time.

\section*{Reaction Time Data}

Outlier reaction times were removed through the application of a \(100-700 \mathrm{~ms}\) reaction time filter. Less than \(2 \%\) of the most extreme scores were removed by this filter. Filtered reaction times from all participants were analysed using a 2 (word orientation) by 3 (levels of validity) repeated measures ANOVA. There was a significant main effect of validity \(\quad(F(2,62)=3.163, \quad p=.049)\), indicating that on average validly cued trials tended to be responded to fast. Additionally there was a significant interaction between word orientation and validity \(\quad(F(1.624,50.352)=3.507, \quad \mathrm{p}=.047\) (Greenhouse-Geisser transformed)). Simple planned comparisons in SPSS were used to explore these effects. Since it was necessary that we demonstrate that cuing had an effect, both classes
of invalid trial were compared to valid trials which should always be the fastest. The difference between reaction times for Invalid Between trials and Valid trials was significant when both orientations were analysed together ( \(\mathrm{F}(1\), \(31)=4.705, \mathrm{p}=.038\) ), indicating that Invalid Between trials were always slow compared to valid trials. However, it was found that there was only a marginally significant difference between reaction times for Invalid Within trials and Valid trials when both horizontal and vertical trials were analysed together \(\mathrm{F}(1,32)=3.828, \mathrm{p}=.059)\). Looking at the graph it is evident that there is a big difference between horizontal and vertical reaction times for Invalid Within trials. This discrepancy was studied using post-hoc Bonferoni corrected t-tests where it was found that on Invalid Within trials, targets inside horizontal words were responded to significantly faster than targets inside vertical words \((\mathrm{t}(31)=2.901, \mathrm{p}<.05)\). However, on both the Invalid Between and Valid trials were was no significant difference between targets inside horizontal and vertical words \((\mathrm{t}(31)=0.385, \mathrm{p}>.05\) and \(t(31)=0.697, p>.05\) respectively). Thus, only on the trials containing horizontally oriented words did participants respond quickly to invalidly cued targets that occurred inside the same word as the cue. This is in accord with what would be expected from object based facilitation since cues inside a word are improving reaction times for targets elsewhere in that word.


Figure 2. Reaction times for each level of validity and each word orientation. Error bars represent 1 standard error.

\section*{Discussion}

This experiment was partially successful in replicating Li \& Logan's (2008) Chinese experiment, using a typologically different language, English. Whilst they found that in both the horizontal and vertical orientations invalid cues within the same word as the target facilitated reaction times, we found this effect only in the horizontal orientation. For horizontally oriented words, invalid cues that occurred inside the same word as the target facilitated target detection reaction times up to a level that was almost indistinguishable from true valid cuing. This indicates that a cue landing anywhere within a horizontally oriented English word will elevate attention levels to the whole word and thereby facilitate target detection in non-cued locations. This supports the idea that words can be treated like objects because this is an "object based effect". However, this effect was not present when the words were oriented vertically.

Since it can be shown that English words have attentional properties of the sort that would normally be associated with objects, this can be seen as evidence for the role of object based attention in reading. However it is of interest that we were unsuccessful in demonstrating this effect in the vertical orientation, where invalid but within word cues were responded to just as slowly as invalid different word cues. The fact that Li and Logan (2008) managed to show this effect in Chinese, whereas we were unsuccessful in doing so for English may be related to the properties of the two languages. It is evident that characters in English and Chinese are very different visually, but they are also processed in different ways. In Chinese there are radicals embedded inside characters that provide phonological and semantic information about that character to the reader, and they are not necessarily read in a strictly linear, left to right fashion. Likewise up until fairly recently Chinese could legitimately be written either left to right, or top to bottom. This is now rare in mainland China but still encountered in other Chinese reading countries. Conversely, top to bottom writing is fairly novel in English. As a consequence it is fair to say that Chinese readers will be much more receptive to seeing Chinese written top to bottom than English readers will be to seeing their language written top to bottom. In English, it would appear that the object based representation of a word which produces these effects is only activated when viewing the word in the familiar orientation. This would imply that when written in the vertical format, English words are decoded using an alternative method which does not produce object based attentional effects.

There are some criticisms that could be levelled at this study. Unlike Li and Logan (2008) background colour patches were used instead of character illumination. This was done in an attempt to control the stimulus intensity of the cues and targets. If we had illuminated letters then the number of pixels that changed colour for any given cue or target would vary wildly from trial to trial based on which
letters occupied that slot. Using the background colour patches enabled us to ensure a much more constant degree of stimulus intensity. However this approach did force certain compromises. In order to have the same size, shape and location of cues/targets between the horizontal and vertical trials it was unavoidable that there would be a better fit in one orientation, in our case horizontal (see Figure 1). There is a possibility that this poor fit may go some way to account for the differences between the horizontal and vertical trials. Also, this was not an experiment that actually involved reading per se. The words that were on screen did not have any bearing on how participants tackled the target detection task. The experimenter did take some steps to ensure the participants were not ignoring the words outright by asking participants to identify the previous pair of words they had just seen. If a participant was repeatedly unable to answer these questions, their data would have been removed. However, no participants needed to be removed for this reason. Nonetheless, the requirement of being able to identify the previously shown array is not nearly as high level as what would typically be considered a reading task.

Consequently, a new experiment is proposed that ensures that cues and targets fit both orientations of words equally well, and goes to additional lengths to ensure participants were actually reading words. Following every trial, participants could be asked to use the previously seen words in a sentence. This would enhance the level of processing the words were subjected to. A further experiment could do exactly the opposite, reproduce this task using non-lexical symbol strings. This would remove reading as a component entirely and address the possibility that these effects are artefacts of tasks where cues and targets are embedded inside letter-like stimuli.

\section*{Conclusion}

This study found evidence that supports the idea that words are sometimes treated as if they were objects by the human attentional system. Reaction time effects normally associated with objects were observed using English words when they were presented horizontally. Thus, the lexical contiguity of words must have been acting in a similar way to the visual contiguity of objects. These findings may support the idea that the parts of the brain that evolved to cope with object perception are at least a part of the network deployed to assist in the novel process of reading.

\section*{Acknowledgments}

This study was funded by the School of Psychology, University of Dundee (PhD studentship awarded to GR). We would like to thank Martin Fischer for helpful discussion.

\section*{References}

Baayen, R. H., Piepenbrock, R., \& Gulikers, L. (1995). The CELEX lexical database (release 2)[cd-rom].
Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania [Distributor]. Philidelphia, PA: Linguistic Data Consortium, University of Pennsylvania [Distributor]. Retrieved from
http://scholar.google.com/scholar?hl=en\&btnG=Searc h\&q=intitle:The+CELEX+Lexical+Database\#0

Blake, R., \& Sekuler, R. (2006). Perception (5th ed.). New York, NY: McGraw-Hill.

Egly, R., Driver, J., \& Rafal, R. D. (1994). Shifting visual attention between objects and locations: evidence from normal and parietal lesion subjects. Journal of experimental psychology. General, 123(2), 161-77. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8014611

Li, X., \& Logan, G. D. (2008). Object-based attention in Chinese readers of Chinese words: beyond Gestalt principles. Psychonomic bulletin \& review, 15(5), 945-9. doi:10.3758/PBR.15.5.945

Luo, C., Lupiáñez, J., Funes, M., \& Fu, X. (2011). The modulation of spatial congruency by object-based attention: Analysing the "locus" of the modulation. The Quarterly Journal of ..., (April 2013), 37-41. Retrieved from
http://www.tandfonline.com/doi/abs/10.1080/1747021 8.2011.591935

Roth, H., Lora, A., \& Heilman, K. (1992). Effects of monocular viewing and eye dominance on spatial attention. Brain, 125, 2023-2035. Retrieved from http://brain.oxfordjournals.org/content/125/9/2023.sho rt

\title{
SOM Cognitive Modeling of Autistic and Schizophrenic Traits Using an Oscillating Topological Neighborhood Width Function
}

\author{
Spyridon Revithis (revithiss@cse.unsw.edu.au) \\ William H. Wilson (billw@cse.unsw.edu.au) \\ Nadine Marcus (nadinem@cse.unsw.edu.au) \\ School of Computer Science and Engineering, University of New South Wales UNSW-Sydney, NSW 2052, Australia
}

\begin{abstract}
The artificial neural network class of self-organizing maps (SOMs) is a powerful and promising cognitive modeling tool in the study of the brain and its disorders. Under this premise, this paper proposes a novel modification of the standard SOM algorithm in the form of an oscillating Topological Neighborhood (TN) width function. Existing research in neuroscience indicates that SOMs with oscillating TN width could exhibit higher biological plausibility than standard TN width SOMs. In this paper, two neuro-developmental disorders, autism and schizophrenia, are modeled, based on existing neurocomputational theories, using both SOM approaches. The simulation results demonstrate that there is significant equivalence between standard and oscillating TN width SOM modeling in terms of map formation behavior, output and structure. The theoretical and computational arguments presented validate the proposed SOM modification within a cognitive modeling framework.
\end{abstract}

Keywords: Self-Organizing Maps, Cognitive Modeling, Cortical Maps, Autism, Delusions, Schizophrenia.

\section*{Introduction}

Computational modeling offers a powerful way to study cognition and behavior. It has been applied to numerous areas of psychology and provides a more promising framework than those based on verbal models in terms of methodological diversity and applicability potential (Sun, Coward \& Zenzen, 2005). An ever-increasing number of computational modeling studies are dedicated to the modeling of cognitive and developmental phenomena using artificial neural networks (Thomas \& Karmiloff-Smith, 2003; Polk \& Seifert, 2002; Parks, Levine \& Long, 1998).

Shultz (2003) provides a comparative evaluation of the different computational neural network systems used to model cognitive developmental phenomena. An important class of such modeling networks is the self-organizing feature map; it is based on a Hebbian-type (Hebb, 1949) unsupervised neural learning mechanism and uniquely resembles topographic cortical maps in the brain to which has directly comparable structure and output characteristics (Spitzer, 1995b; Livingstone \& Hubel, 1988; Blasdel \& Salama, 1986; Merzenich \& Kaas, 1980). Willshaw and von der Malsburg (1976) originally proposed the self-organizing neural network to account for the retinotopic mapping problem. Kohonen's version (2001) -commonly abbreviated to 'SOM'-, however, possesses significant computational
characteristics and a range of powerful properties, particularly relevant to understanding and modeling of cortical brain maps, including approximation of the input space, topological ordering, density matching, and feature selection (Haykin, 1999).

This study investigates cognitive modeling aspects of modeling neuro-developmental disorders using SOM neural networks. The first section presents the SOM modeling framework used in this work, and introduces a novel modification in the SOM formation algorithm with significant cognitive modeling implications. In the subsequent two sections, core biological and behavioral characteristics of two mental disorders, autism and schizophrenia, respectively, are modeled using a prototype SOM model. The last section provides a discussion of the computational and theoretical parameters of the SOM modeling employed in the paper.

\section*{The SOM Modeling Framework}

\section*{Aspects of SOM Neural Networks}

A SOM is a non-linear unsupervised-learning computational neural network consisting of two layers. It has the capacity to map an input 'environmental' layer, consisting of patterns of fixed but arbitrary dimension, onto a (usually) one or two dimensional lattice 'representational' layer. The representation of environmental input in the output layer (called the map) is performed in a topologically ordered fashion, maintaining the non-linear input data distribution, and involves dimensionality reduction. Figure 1 shows an abstract depiction of a two-dimensional SOM; each input layer pattern vector connects fully with the map neurons.


Figure 1: A two-dimensional SOM.

The SOM neural network formation (training) process has four parts (as described in Haykin (1999)): synaptic weight initialization of the output lattice; neuron competition; neuron cooperation; and synaptic adaptation. The last three are sequenced within a loop for a finite number of 'epochs', in which input patterns are presented and weights adjusted until the weights converge.

During the competition phase, a winning neuron for the current input pattern is determined, based on a Euclidean distance metric. In the cooperation phase the winning neuron becomes the center of a cooperative process extending around an area according to a topological neighborhood (TN) function. In the synaptic adaptation phase, the weights of the map neurons within the TN of the winning neuron are updated 'towards' the current input pattern at an intensity determined by their lateral distance to the winning neuron as well as an exponentially decaying learning rate function.

From a cognitive modeling perspective, it is of particular interest to examine the neurobiological relevance of the SOM formation process at the implementation level of the neuron lateral interaction and inhibition mechanism. The standard SOM algorithm (Haykin, 1999) employs a translation invariant Gaussian TN function with an exponentially decreasing width, as illustrated in Figure 2.



Figure 2: Decreasing TN width around a winning neuron (dark grey neuron) in a two-dimensional SOM.

The TN width function can be expressed by the formula
\[
\sigma(n)=\sigma_{0} \cdot \exp \left(-\frac{n}{\tau_{1}}\right), \quad n=1,2, \ldots, t
\]
where \(\sigma_{0}\) is the initial TN width, \(\tau_{1}\) is a time constant, \(t\) is the number of epochs, and \(n\) is the current epoch.

The fact that only neurons close to the winning neuron have their weights changed significantly (implemented at the biological neural network level by a mixture of excitation and lateral inhibition) has a measurable impact on the representational structure of the SOM. A number of SOM cognitive models of brain disorders center around the key role of TN width and its exegetic biological significance (Gustafsson, 1997; Spitzer, 1999).

\section*{Oscillating TN width SOM}

The SOM cooperative phase involves local neuronal interactions via group Hebbian activation regulated by lateral inhibition. In general, neural synchrony and communication at the local and long-range level is an important aspect of brain functioning; neural oscillation,
particularly correlated to inhibitory neural activity, is increasingly considered to be of paramount importance to neural information processing and central to a number of studies of mental disorders including schizophrenia and autism (Schnitzler \& Gross, 2005; Wang, 2010). Neuronal group oscillatory synchrony is linked to inhibitory interneuron rhythmic modulation of the firing rate of excitatory neurons, at the local interaction neuronal level (Cardin, Carlen, Meletis, Knoblich, Zhang, Deisseroth, Tsai \& Moore 2009). Last, synchronous oscillatory activity of neighboring inhibitory interneurons may be supported by sub-threshold oscillatory behavior (Llinas, 1988).

In line with the relevant research on neural oscillation outlined above, this paper introduces a modification with increased biological plausibility in the SOM cooperative phase, as previously reported in a preliminary study (Revithis, 2011). Specifically, the original TN width function, part of the overall TN function, is replaced by a new TN width function that exhibits local exponential decrease instead of global. In this way the TN width oscillates continuously throughout the SOM formation process. Oscillation is necessary in a biologically plausible model, otherwise learning would cease when the TN approached zero. The oscillation consists of a concatenation of exponentially decreasing original TN width -temporally shortened- 'function instances'; thus, in the same number of epochs (i.e., one SOM training session) multiple function instances will fit, as shown in Figure 3.


Figure 3: SOM oscillating TN width.
The new function can be expressed as
\[
\sigma^{\prime}(n)=\sigma_{0} \cdot \exp \left(-\frac{(n+1) \bmod t^{\prime}}{\tau_{1}^{\prime}}\right), \quad n=0,1,2, \ldots, t-1
\]
where \(\sigma_{0}\) is the initial TN width, \(\tau_{1}^{\prime}\) is a time constant, and \(n\) is the current epoch. The constant \(t^{\prime}=t / c\), where \(c\) is the oscillation constant determining how many times the TN width will reset to \(\sigma_{0}\) and start decreasing again.

\section*{IPSOM}

IPSOM (Interlocking Puzzle SOM) is a complex-weightencoding prototype SOM spatial behavioral model of how humans complete interlocking puzzles (Revithis, Wilson \& Marcus, 2006). When trained, using a representative sample of puzzle completion sessions, it forms a behavioral SOM of the statistically dominant patterns (strategies) of puzzle completion. A 6x6 IPSOM has been evaluated for the case of \(4 \times 5\) puzzles against a simulated group of people. Each 'virtual' person used one of four predetermined puzzle completion strategies, illustrated in Figure 4.

Each radar-graph in Figure 4 depicts the order of puzzle completion for each pattern (H, V, PH, PV). The radial axis shows the encoded numerical position values on the puzzle board (i.e., which puzzle piece), and the angular axis shows the discrete completion sequence numbers (i.e., which piece is first, second, etc.) By connecting the points on the graph, a distinct visual pattern is formed. Attached to each graph, a puzzle board contains the puzzle completion order conventionally. The design principles behind the selected strategies were the generation of a small number of straightforward, real-life-based patterns, the utilization of topological clustering, and emphasizing the basic strategy of determining the board periphery during the puzzle completion. IPSOM was conclusively found to be efficient in modeling the behavioral domain.


Figure 4: IPSOM training set patterns (strategies).
In this paper, IPSOM is employed as a modeling test-bed for cortical map spatial perception. The working hypothesis is that IPSOM is not only a behavioral model but also a cognitive model of how humans perceive puzzle completion strategies when presented with puzzle completion examples. It is assumed that an average person would form an internal representation of the dominant strategies; a cortical map would retain the domain specific knowledge, modeled by a trained SOM. IPSOM is expected to represent the training patterns in a topologically ordered fashion, where neighboring patterns are also visually similar (Figure 5).
\begin{tabular}{|c|c|c|c|c|c|c|}
\hline & 1 & 2 & 3 & 4 & 5 & 6 \\
\hline 1 & PV & PV & PV- & V- & V & V \\
\hline 2 & PV & PV- & PV- & V- & V- & V \\
\hline 3 & PV- & PV- & \(\sim\) & ~ & V- & V- \\
\hline 4 & PH- & PH- & \(\sim\) & \(\sim\) & H- & H- \\
\hline 5 & PH & PH- & PH- & H- & H- & H- \\
\hline 6 & PH & PH & PH- & H- & H & H \\
\hline
\end{tabular}
\begin{tabular}{ll}
\(\boldsymbol{X}\) Primary core neuron: & Optimal pattern representation \\
\(X\) Core neuron: & Good pattern representation \\
\(X-\) Weak neuron: & Poor pattern representation \\
\(\sim\) Undecided neuron: & Transitionally excessive pattern
\end{tabular}

Figure 5: An abstract illustration of a trained 6x6 IPSOM.

\section*{Modeling Aspects of Autism using IPSOM}

\section*{A Neural Circuit Theory of Autism}

Autism, a pervasive developmental disorder, has been studied for over 50 years by an expanding interdisciplinary research community. The current diagnostic tools (DSM-IV and ICD-10) dictate a socio-psychological behavioral approach that does not inform of the causes of autism; nevertheless, it is considered to be neurobiological in nature (Coleman \& Gillberg 2012).

Autism is associated with atypical perception and its internal representation. Sensory input often fails to integrate into existing memory due to abstraction impairment; there is difficulty in detecting the important features among the nonessential details; elaborating on internal representations is also problematic, where it appears that central executive control is required (Frith, 2003).

Gustafsson's (1997) neural circuit theory of autism is based on these empirically based concepts of autistic perception and proposes a neural-level explanation for the lack of drive for central coherence, a key element in autistic behavior (Frith, 2003). Neurological deficiencies in the formation of brain cortical maps give rise to autistic attributes. This leads to problematic feature extraction since "autistic raw data memory" operates in place of "feature memory" due to "inadequate cortical feature maps". Raw data memory is intrinsically linked at the behavioral level to the diagnostic criteria for autism (Gustafsson, 1997). Autistic maps lack feature distinction and preservation, and fail to provide an internal representation of salient perceptual data leading to raw data memory that lacks sophisticated representations.

According to Gustafsson (1997), SOMs provide a biologically plausible way to model characteristics of 'autistic' cortical maps. A SOM can represent input features just as a cortical map in the brain retains salient perceptual stimuli, and can exhibit similar deficiencies to an autistic cortical map if its formation mechanism is impaired.

\section*{The Autistic IPSOM}

The modeling premise of the SOM autistic impairment is suggested not by the biological map, but by its model. Gustafsson (1997) argued that a biologically plausible cause of impairment in a SOM is the application of excessive lateral feedback inhibitory synaptic strengths. The latter can degrade the map's generalization and feature representation capacity, resulting in high sensory discrimination and feature specificity, even to the point of instability, leading to the formation of inadequate or even undeveloped maps.

This modeling premise can be expressed as a TN premature narrowing during SOM training; TN can be regarded as the "source of power" (Sun \& Ling, 1997) in the autistic model. The initial TN width \(\left(\sigma_{0}\right)\) in the TN width function affects the map's representational capacity in a directly applicable way to Gustafsson's theory (Revithis \& Tagalakis, 2012). A non-autistic cortical map is expected to
represent all the dominant puzzle completion strategies with smooth transition between them. This can be modeled using IPSOM in its original parameter configuration.

After the incorporation of TN parameter modifications on IPSOM, an evaluation was performed. A series of groups of controlled simulations were executed with the initial width of the TN function set to a typical value of \(\sigma_{0}=3\) (i.e., equal to the network's radius, as suggested by Haykin (1999)) for one group, and reduced to \(\sigma_{0}=1.15\) for another group. Both groups were executed twice, using a standard TN width function, in one simulation series, and an oscillating TN width function in a second one. The results (discussed next) from over 150 simulations confirm that, for large \(\sigma_{0}\), the resulting IPSOM exhibits efficient representation of the input space, whereas IPSOM training, using a small \(\sigma_{0}\), forms a map with autistic structural characteristics. The results also support the hypothesis that the oscillating TN width IPSOM is equivalent to the standard TN width IPSOM in modeling autistic traits.


Figure 6: Standard TN width IPSOM map characteristics.
Figure 6 depicts IPSOM neurons after training, using a standard TN width function, for \(\sigma_{0}=3\) (top) and \(\sigma_{0}=1.15\) (bottom). The leftmost 3D graphs, and the 2D graphs in the middle, depict the Euclidean distance of pattern H to each neuron in the map. The darker and closer to the horizontal 3D base-plane (map) areas signify smaller distance and, thus, higher representational accuracy for pattern H. A \(\sigma_{0}=3\) facilitates a smoother transition from pattern H to other patterns in the map, whereas a \(\sigma_{0}=1.15\) results in steeper increase of the Euclidean distance indicating transitional pattern impairment. The rightmost combined-concentric radar graphs depict five neighboring IPSOM neurons for \(\sigma_{0}=3\) (top) and \(\sigma_{0}=1.15\) (bottom). A \(\sigma_{0}=3\) facilitates smoother transition from Pattern H to V , whereas for \(\sigma_{0}=1.15\) neurons are tightly grouped in two patterns (H and V) with impaired transition and generalization capacity.

Figure 7 depicts IPSOM neurons after training, using an oscillating TN width function, for \(\sigma_{0}=3\) (top) and \(\sigma_{0}=1.15\) (bottom). The observations that can be made are identical to the ones of Figure 6.


Figure 7: Oscillating TN width IPSOM map characteristics.
The illustrated example-simulation-results of Figures 6 and 7 are representative of the totality of simulation results obtained in terms of the observed characteristics. Patterns H and V , which were used for the rightmost concentric radar graphs, were selected to better demonstrate IPSOM's transitional behavior due to their relatively low correlation significance amongst IPSOM training set patterns (Table 1).

Table 1: Correlation between IPSOM training patterns.
\begin{tabular}{cccccc}
\hline Spearman's \(\rho(N=20)\) & & H & V & PH & PV \\
\hline Correlation Coefficient & H & 1 & .429 & \(.523^{*}\) & \(.507^{*}\) \\
Sig. (2-tailed) & H &. & .059 & .018 & .023 \\
Correlation Coefficient & V & & 1 & .388 & .420 \\
Sig. (2-tailed) & V & &. & .091 & .066 \\
Correlation Coefficient & PH & & & 1 & \(.974^{\#}\) \\
Sig. (2-tailed) & PH & &. & .000 \\
Correlation Coefficient & PV & & & 1 \\
Sig. (2-tailed) & PV & & &. \\
\hline Correlation is significant at the 0.05 level (*) and at the 0.01 level (\#).
\end{tabular}

\section*{Using IPSOM to Model Delusions}

\section*{Acute and Chronic Delusions in Schizophrenia}

Modern studies on schizophrenia span approximately a century. There has been a continuous evolution of the understanding of this mental disorder and currently it is widely considered to be a progressive neuro-developmental disorder. Amongst its common positive psychotic symptoms are delusions (Green, 2001).

Spitzer has argued (1995a, 1995b, 1999) that SOM neural networks can provide a model of brain cortical function, and implement lateral inhibition, an essential feature of cortical maps. Furthermore, he proposed a neurocomputational exegetic framework for delusions based on the concepts of neuromodulation and neuroplasticity in relation to formation and operation of sensory and higher-order computational maps in the cortex.
Specifically, according to this approach, neuromodulator activity in the brain is associated with the signal-to-noise
ratio at the neuronal level, from an information-theoretic perspective. High neuromodulator activity can lead to an increase of focusing in neuronal activation and is associated with acute delusional states; such focusing can be modeled via excessive SOM lateral inhibition. Chronic delusions can then be regarded as the result of the establishment of entrenched cortical maps via sustained acute delusional states due to brain neuroplasticity.

\section*{IPSOM Modeling of Delusions}

According to Spitzer (1995a), a decisive factor in the clinical phenomenon of acute delusions is the level of cortical neuromodulator activity; this affects modulation of signal-to-noise ratio. In a SOM model of delusions it is possible to regulate the level of neuronal activation focusing associated with the signal-to-noise ratio by controlling SOM lateral inhibition. This can be achieved by controlling the width of TN during SOM formation. TN can be regarded as the "source of power" (Sun \& Ling, 1997) in this model.

Similar to the autistic model, the working hypothesis is that the initial TN width \(\left(\sigma_{0}\right)\) in the TN width function affects the map's behavior in a way applicable to Spitzer's theory. Inducing acute delusions in IPSOM can be realized via modifying the cooperation phase of the SOM algorithm in the model to employ a significant TN narrowing.

A series of groups of controlled simulations were executed with the initial width of the TN function set to a typical value of \(\sigma_{0}=3\) for one group, and reduced to \(\sigma_{0}=1.15\) for another group, as in the autistic model. Both groups were executed twice, using a standard TN width function, in one simulation series, and an oscillating TN width function in a second one. The results (discussed next) from over 150 simulations confirm that, for large \(\sigma_{0}\), the resulting IPSOM exhibits typical representation of the input space; when a small \(\sigma_{0}\) is used, however, the map's formation behavior is atypical and retains structures corresponding to chronic delusions. The results also support the hypothesis that the oscillating TN width IPSOM is equivalent to the standard TN width IPSOM in modeling delusions.

Entrenched SOM structures that could give rise to chronic delusions can be identified by comparing 'suspected' formed IPSOM maps with their untrained (initial) state. A 'delusional' structure can plausibly be seen as a number of trained neurons representing neither a transitional pattern nor an input space pattern, or, excessively representing an input space pattern (the latter can be regarded as compromising the SOM density matching property (Haykin, 1999)). Furthermore, representational resistance to change can also be interpreted as a characteristic of established (chronic) delusional structures (Spitzer, 1995a).

Figure 8 depicts four snapshot graphs of the same part of the IPSOM map for different initial parameters. In graph A we see the situation before training - essentially random patterns, and the remaining three depict the map's area after training for different \(\sigma_{0}\) value and TN width function configurations. By comparing IPSOM's untrained graph with its standard TN width trained counterpart (graph C) we
immediately observe the perseverance of a number of initial 'blank' patterns. A number of IPSOM neurons represent either the original initial 'blank' pattern or a distorted version of it. In the oscillating TN width case (graph B) there is also an excessive representation of the V pattern (cf. Figure 4). The observed 'delusional' flags, especially the resistance to environmental change, are prominent in the IPSOM trained graph using a very small \(\sigma_{0}\) (graph D).


Figure 8: Induced delusional structure on IPSOM.

\section*{Discussion}

The significance of TN in SOM cognitive modeling has theoretical and practical implications. In this paper, a modified TN width function with increased biological plausibility (paramount to modeling) was introduced and simulation results, based on the IPSOM prototype, on two models of neuro-developmental disorders were presented.

The modeling significance of the oscillating TN width function is associated not only with the initial TN width \(\left(\sigma_{0}\right)\) parameter but, primarily, with the TN width 'area' covered throughout the SOM training. What is considered 'narrow' or 'wide' TN during SOM formation is -from a different perspective- a function of the TN width area covered.



Figure 9: Standard and Oscillating TN width areas.

In Figure 9, both the standard and the oscillating TN width functions are overlaid in both graphs. The TN width area has as an upper bound the corresponding TN width function and as a lower bound the epoch (horizontal) axis.

Mathematically, the TN width area is expressed as
\[
\sigma(\mathrm{x}) \operatorname{area}=\int \sigma_{0} \cdot e^{\left(-\frac{x}{\tau}\right)} d x=\sigma_{0} \cdot(-\tau) \cdot e^{\left(-\frac{x}{\tau}\right)}+\mathrm{C}, \sigma_{0}, \tau \in \mathrm{R}
\]

To calculate the area for a given TN width function, \(\sigma_{0}\), and number of epochs \(t\), the following formula was used:
\[
\sigma(\mathrm{x}) \text { area }=\int_{0}^{t} \sigma_{0} \cdot e^{\left(-\frac{x}{\tau}\right)} d x
\]

In the standard \& oscillating TN width IPSOM simulation results, the calculated \(\sigma(x)\) area (for the same \(\sigma_{0}\) ) remained unchanged irrespective of the TN width function used. This verifies the output equivalence between the two modeling approaches. Furthermore, when, in the oscillating TN width function simulations, the \(\sigma_{0}\) value was reduced to \(\sigma_{0}^{\prime}\), the calculated \(\sigma(x)\) area was significantly smaller (Figure 9, right graph) and resulted in an IPSOM map with more pronounced delusional structures (Figure 8, graph D). This demonstrates the computational and cognitive modeling significance of the TN width area.

In conclusion, it is important to note that making a link between the biological and computational levels, in such modeling studies, often requires a sequence of finely drawn associations across disparate disciplines. However indirect and interdisciplinary such a link may be, the methodology and tools to construct it have long been available, and an effort was made in this study to illustrate it.

\section*{References}

Blasdel, G. G., \& Salama, G. (1986). Voltage-Sensitive Dyes Reveal a Modular Organization in Monkey Striate Cortex. Nature, 321, 579-585.
Cardin J. A., Carlen M., Meletis K., Knoblich U., Zhang F., Deisseroth K., Tsai L. H., \& Moore C. I. (2009). Driving Fast-Spiking Cells Induces Gamma Rhythm and Controls Sensory Responses. Nature, 459(7247), 663-7.
Coleman, M., \&, Gillberg C. (2012). The Autisms. NY: Oxford University Press.
Frith, U. (2003). Autism: Explaining the Enigma. MA: Blackwell.
Green, M. F. (2001). Schizophrenia Revealed: From Neurons to Social Interactions. NY: W. W. Norton.
Gustafsson, L. (1997). Inadequate Cortical Feature Maps: A Neural Circuit Theory of Autism. Biological Psychiatry, 42, 1138-1147.
Haykin, S. (1999). Neural Networks: A Comprehensive Foundation. NJ: Prentice Hall.
Hebb, D. O. (1949). The Organization of Behavior. NY: John Wiley \& Sons.
Kohonen, T. (2001). Self-Organizing Maps (3rd Ed). NY: Springer.
Livingstone, M., \& Hubel, D. (1988). Segregation of Form, Color, Movement, and Depth: Anatomy, Physiology, and Perception. Science, 24, 740-749.

Llinas R. R. (1988). The Intrinsic Electrophysiological Properties of Mammalian Neurons: A New Insight Into CNS Function. Science, 242(4880), 1654-1664.
Merzenich, M. M., \& Kaas, J. H. (1980). Principles of Organization of Sensory-Perceptual Systems of Mammals. NY: Academic Press.
Parks, R. W., Levine, D. S., \& Long, D. L. (Eds.) (1998). Fundamentals of Neural Network Modeling: Neuropsychology and Cognitive Neuroscience. MA: The MIT Press.
Polk, T. A., \& Seifert, C. M. (Eds.) (2002). Cognitive Modeling. MA: The MIT Press.
Revithis, S. (2011). Significance of Topological Neighborhood in SOM Cognitive Modeling of Brain Disorders: Current Neurocomputational Simulations. Abstracts of \(16^{\text {th }}\) APPAC - APPAC Journal, 18(2), 26.
Revithis S., \& Tagalakis G. (2012). A SOM-based Validation Approach to a Neural Circuit Theory of Autism. In I. Maglogiannis, V. Plagianakos \& I. Vlahavas (Eds.), SETN 2012, Artificial Intelligence: Theories and Applications, 7297 (pp. 25-32). Berlin: Springer.
Revithis S., Wilson W. H., \& Marcus N. (2006) IPSOM: A Self-Organizing Map Spatial Model of How Humans Complete Interlocking Puzzles. In A. Sattar \& B. H. Kang (Eds.), AI 2006: Advances in Artificial Intelligence, LNAI, 4304 (pp. 285-294). Berlin: Springer.
Schnitzler, A., \& Gross, J. (2005). Normal and Pathological Oscillatory Communication in the Brain. Nature Reviews Neuroscience, 6, 285-296
Shultz, T. R. (2003). Computational Developmental Psychology. MA: The MIT Press.
Spitzer, M. (1995a). A Neurocomputational Approach to Delusions. Comprehensive Psychiatry, 36(2), 83-105.
Spitzer, M. (1995b). Conceptual Developments in the Neurosciences Relevant to Psychiatry. Current Opinion in Psychiatry, 8(5), 317-329.
Spitzer, M. (1999). The Mind Within the Net: Models of Learning, Thinking and Acting. MA: The MIT Press.
Sun, R., \& Ling, C. (1997). Computational Cognitive Modeling, the Source of Power and Other Related Issues. AI Magazine, 19, 113-120.
Sun, R., Coward, L. A., \& Zenzen, M. J. (2005). On Levels of Cognitive Modeling. Philosophical Psychology, 18, 613-637.
Thomas, M. S. C., \& Karmiloff-Smith, A. (2003). Connectionist Models of Cognitive Development, Atypical Development and Individual Differences. In R. J. Sternberg, J. Lautrey \& T. Lubart (Eds.), Models of Intelligence: International Perspectives, 44. DC: APA.
Wang, X-J. (2010). Neurophysiological and Computational Principles of Cortical Rhythms in Cognition. Physiological Reviews, 90(3), 1195-1268.
Willshaw, D. J., \& von der Malsburg, C. (1976). How Patterned Neural Connections Can Be Set Up by SelfOrganization. Proceedings of the Royal Society of London Series B, 194, 431-445.

\title{
The Language of Everyday Verbal Analogies
}

\author{
Lindsey E. Richland (Irichland@uchicago.edu) \\ University of Chicago, Department of Comparative Human Development Chicago, IL 60637
}

\author{
Justin B. Richland (jrichland@uchicago.edu) \\ University of Chicago, Department of Anthropology \\ Chicago, IL 60637
}

\begin{abstract}
Verbal analogies produced during naturally occurring instructional discourse in mathematics were explored using techniques borrowed from studies of language in use (see Wortham \& Rymes, 2003). Close examination of two eighth-grade instructional analogies reveals that the language practices of analogy are instrumental in shaping recipients' relational re-representation of objects being compared, in particular through markers of indexicality and poetic parallel structure. At the same time, close examination of the communicative interactions reveals that these devices may reduce the burden on recipients' reasoning to the point that they may appear successful at solving the verbal analogy, but their responses can be explained by facility in verbal interaction rather than in mathematical reasoning. These data provide thereby new insights into the "analogical paradox," the finding that analogies are commonly successful as vehicles for interactionally producing and displaying understanding of new information in everyday contexts but generally problematic when measured for their effects on reasoning in controlled laboratory settings (Dunbar, 1998). We identify a tension between interactional and cognitive success of everyday communicative analogies, meaning that those that are most likely to be interactionally successful may lead to less cognitive engagement for analogy recipients.
\end{abstract}

Keywords: analogy; analogical reasoning; language, linguistic anthropology

\section*{Introduction}

Analogy is the process of identifying shared relational similarities across contexts or representations, and has been theorized as integral to humans' everyday flexibility and higher order adaptive thinking (Doumas \& Hummel, 2012). Analogy has also been empirically identified as a regular practice within everyday communication in contexts including scientific biology laboratories (Dunbar, 1995, 1999), political discourse (Blanchette \& Dunbar, 1997, 2001) and classrooms (English, 2004; Richland, Holyoak \& Stigler, 2004; Richland, Zur \& Holyoak, 2007).

At the same time, basic analogical problem solving and transfer in the laboratory is notoriously unreliable and often unsuccessful (e.g., Gick \& Holyoak, 1980, 1983). These differences between analogy production
in the lab and in everyday interaction led Dunbar (1998) to speculate about the "analogical paradox," the insight that analogy is often rare and difficult to produce in the laboratory, but frequent and effective in everyday talk.

The current paper draws on linguistic anthropological methods for studying the empirical details of everyday interactions to better understand this paradox. The analysis uses techniques borrowed from the linguistic anthropology of education (Wortham \& Rymes, 2003), applying studies of language-in-use to educational discourse. Following this, language is conceptualized here as a performative activity that carries pragmatic as well as referentio-semantic meaning (Austin, 1962; Hymes, 1972,). In using analogies, teachers in mathematics instruction provide not only information regarding the denotational and other forms of semantic content of the lexical and grammatical structures of the talk they use. They also, simultaneously, signal to their student-audiences how, in the specific and actual moments of their use, their talk is to be understood as a move in the turn-by-turn exchange that is constituting the particular instructional discourse of which it is a part, and to which they will be expected to respond "appropriately." Most often, in instructional discourses using analogies, the "appropriate" student response will also be a response that is treated by the instructor as proof of the student's effective "correct" (referentio-semantic) understanding and reasoning based on the analogy's denotational content. But sometimes the pragmatically appropriate response is not the same as referentio-semantically "correct" one, revealing how the student in such exchanges is orienting and responding to two orders of meaning at once.

For example, consider an instructional analogy that is initially expressed as: "Lets say that I loaned you twenty five dollars and then I loaned you twenty five more dollars, what would you owe me?" Such an analogy source pragmatically indicates that the learner should not encode the analogs as a truthful representation of the facts as they exist in the context of use but rather as a proposed hypothetical situation shared between the teacher and student. In a classroom context involving an elementary school teacher instructing his student about negative numbers, a student might respond in a pragmatically adequate and
semantically "correct" manner with the same answer, "I'd be down fifty dollars." But in an ethics classroom in high school, in which the teacher was instructing about coercion, a semantically "correct" answer "I'd be down fifty dollars," may not be pragmatically adequate.

Linguistic anthropology of education builds from this notion of language in use to demonstrate that many instructional activities rely upon not only the linguistic production of educational content but also the form and participatory patterns of the construction. Lexical form, grammatico-syntactical structures, and the phenomenal features of speech and text as produced in real time can simultaneously signal social, political, discursive, participatory and other kinds of meaning to participants in a social interaction (e.g. Duranti, 1997). The use of such linguistic devices enables the speaker to provide pragmatic and metapragmatic cues, which simultaneously constitute the speech, while at the same time informing recipients how they should be interpreting that speech and preparing to respond to it (Goffman, 1974; 1981, Hymes 1972, Silverstein 1979, 1993). In educational contexts this is particularly important because classroom interaction not only affects the relationships between classroom actors, but it also impacts the cognitive activity performed by students during learning situations.

The current manuscript describes analyses of two classroom analogies that were identified from a larger corpus as illustrative of the verbal analogies produced in instruction. They are transcribed using conventions borrowed from conversation analysis (Sacks, Scheglof \& Jefferson, 1974). Within those transcripts, two linguistic resources were identified as both common to and particularly meaningful of the production of the verbal structure-mapping: indexicality and parallel structure.

Indexicality. All linguistic features, when used, have the capacity to index, or point to, aspects of their contexts of use as ways of shaping their conceptual meaning to competent members of a speech community (Ochs, 1992). Deictic indexes are those such as, "you," that have minimal semantic meaning aside from the precise context of the talk (Hanks, 1992, see also Silverstein 1976; Horn, 1988). As phrased by Hanks: "their basic communicative function is to individuate or single out objects of reference or address in terms of their relation to the current interactive context in which the utterance occurs" (1992: 47).

Thus the use of deictics makes the semantic meaning of an utterance inexplicable without the immediate context, which imposes a further burden on interaction participants to comprehend the multiple levels of meaning intended by the speaker. For instance, the use of the word "you" in the following phrase, "If you are having trouble, raise your hand" carries 1) semantic meaning - that the speech in question is intended to be directed to someone else (second person, not marked as singular or plural in

English) who is proposed as its addressee, and 2) a contextually specific, interactional meaning - the teacher is inviting those in the presumed range of hearing - here perhaps a group or subgroup of classroom students - to take up the position of addressee, and to respond, provided they interpret the qualification "having trouble" as applicable to them.

This study will examine the role of indexicality in teachers' discursive work to help students produce certain constrained representations of information in order to create comparable analogs. This carefully crafted relational re-representation is essential, because the major identified problem in doing analogy is noticing the relevance of mapping the relational structure from one analog to another (see Gick \& Holyoak, 1980, 1983). Thus if one's mental representation of a particular object in the world does not align with another system, the reasoner will likely fail to notice the relevant higher order structure mapping between them. Indexes that mark the irreal, or hypothetical nature of the source representation are illustrated in the first analogy described below. In so doing, and much as in the "Let's say..." example used above, the interacting students in the first analogy described are invited by the teacher to construct a particular source analog that does not have to reflect all the perceptual and relational characteristics of reality, but rather to isolate and highlight the key relationship depicted in the discourse.

Competent members of a speech community are highly skilled at interpreting indexical talk, though participants who are not fully members of that speech community (e.g., English Language Learner students), or students under high processing load to hold mathematical representations in mind, may find this a challenge that reduces their available resources to interpret a conceptually demanding analogy.

Poetic Structure. Second, this analysis takes up the reflexive capacity of language to serve, simultaneously as both the content of communication and commentary upon that content, particularly in the ways in which aesthetic forms such as rhyming, prosody, and even tempo can shape how semantico-referential content is to be interpreted and responded to by recipients and addressees (Lucy, 1993, 1999). One such example is discerned in the parallel structuring of discursive clauses in sequences of moments of actual speech and textual production, deploying what some have called the poetic dimension of meaning-making in language. (Jakobson, 1960; Silverstein, 1985) The notion of poetic structure and its regular and repeated occurrence in verbal analogy is particularly relevant to the current analysis, insofar as it offers yet another discursive channel for conveying the intended comparison between two systems of similarly structured relationships. In this sense, the poetic dimension of parallel structuring in verbal analogy becomes iconic of the semantic content of the speech, and the proposed
relationship between source and target that the instructor endeavors to produce by it. Parallel structure of speech serves as a pragmatic index for the analogical structure mapping itself.

The forthcoming analysis more closely examines the affordances and routines of indexicality and parallel structure within the discourse structures of analogy speech events. The paper will attempt to show that language mediates the activity of analogy in classroom mathematics instruction. In particular, it shows how the resources of indexicality and parallel structure are frequently instrumental in the outcomes of students' learning experiences during instructional analogies. The analysis will explore how teachers use indexicality and parallel structure to draw students into creating mental re-representations of the source and target objects as distinct relational systems by situating them in hypothetical, temporally defined, and/or spatial worlds that are then systematically aligned and mapped together. These can produce the dual, conflicting functions of drawing recipients' attention to relational similarity and increasing the likelihood that they will notice and successfully complete analogical structure mapping. At the same time, the high levels of structure provided by the language can reduce the mathematical, semantic learning potential for students.

\section*{Methods}

\section*{Sample}

The analogies analyzed in this paper are a subset of verbal analogies identified and coded in larger studies of classroom teachers' use of relational comparisons in videotaped U.S., Japanese, and Hong Kong Chinese eighth-grade mathematics lessons (Richland, Holyoak \& Stigler, 2004; Richland, Zur \& Holyoak, 2007). A randomized probability sample of all \(8^{\text {th }}\) grade mathematics lessons taught in the United States was videotaped as part of the Third International Mathematics and Science Study (Stigler et al, 1999). In a secondary analyses of these data, a random subset of the U.S. lessons were further analyzed by trained and reliable coders to identify and categorize analogy usage using frequency coding. Key representative analogies within these units of analysis were transcribed using conventions of conversation analysis (Sacks, Schegloff \& Jefferson, 1974).

Indexicality and parallel structure are analyzed in two analogies selected from this corpus of data. These analogies were selected because they are typical of the 298 U.S. analogies identified and coded, and for their clarity in revealing common and potentially consequential discursive constructions. They were not selected for their mathematical sophistication or efficacy, and should not be construed as ideal examples of the potential for analogy to support classroom mathematics learning. Many of the more mathematically sophisticated analogies follow similar
patterns but included more extended discussion, making their length prohibitive for a paper-length analysis of several examples. The first analogy demonstrates how parallel structure can provide a poetic representation of the analogical structure mapping itself. The second analogy also invokes parallel structure, but further reveals the role of indexicality in constructing source representations and structure-mapping during production of analogies by drawing on irrealis, space, and time, and the second

\section*{Analogy Segment 1: Poetic Structure}

Analogy 1 demonstrates how the linguistic form of an analogy can generate participation and model conceptual mapping. This teacher aligns the mathematical concepts of generating equivalence across the equal sign with converting fractions to like denominators. These are different concept areas within algebra, though procedures used for manipulating these structures are similar. The analogy arises while the teacher is at the chalkboard instructing students about how to make fractions equivalent. She is teaching the rule that when one multiplies the bottom number of a fraction times a number, one must multiply the top number times the same number to retain the same fraction. She depicts this on the board in an example, where she multiplies both the numerator and the denominator times four to determine that \(2 / 4\) is equivalent to \(8 / 16\). The analogy the teacher makes between these concept areas is fairly procedural and does not engage in the relationship between the deep mathematical structure of these concepts, however it is interactionally successful and students are able to complete the teachers' designedly incomplete utterances throughout the analogy (Koshik, 2001).

The organizational structure of language, beyond its denotational and indexical meaning, can play an important role in the interactional and conceptual consequences of language in use (e.g. Jakobson, 1960; 1971[1966]). One constitutive factor of Jakobson's (1960) model of a speech event, is the poetic feature of language. He uses this category to foreground the aesthetic or perceptual features of talk, arguing that these carry their own functionality. The role of parallel structure is particularly relevant to analogy in use, since the conceptual basis for analogy is the development of relevant parallels between the conceptual structure of source and target objects. Teachers regularly invoked parallel structure in the lexical and grammatical construction of the analogical mappings, thus creating grammatical metaphors for the conceptual mapping being constructed. Parallel structure within the discursive form in this way may thus serve as reflexive language cues to listeners, such that the form of the structural parallelism within the utterance serve a guiding function, leading talk recipients to infer that the
ensuing talk should be mentally represented as a set of parallel structures (Lucy, 1999).
```

On board:
$\frac{2(4)}{4(4)}=\frac{x}{16}$
$\mathrm{x}=$
T: Okay - just like equations (.)
whatever you do tho one side
( ( gesturing to her left)
(you have to do (..) [(.) to the other.
[((gestures to right, emphasis on "other"))
s: [to the other
[Whatever you do to the denominator,
[(fcurls hands inwards, gesturing toward herself))
you have to [do to the (..) [numerator.
[ (funcurls hands to bend fingers out away
from herself))
okay (.) Now numferator
T: Okay (.) Now

```

Figure 1. Analogy between operating on equations and fractions.

The teacher begins this analogy following a procedural explanation of how to multiply the same number to the denominator and the numerator in order to produce an equivalent fraction with a new denominator. The teacher begins with the token "okay," marking a transition between the prior expository talk and the ensuing discourse. This indicates that this is a distinct unit of talk. She then indexes that she is designing a comparison with the comparative marker "just like," followed by the referent "equations" to signal the source of the comparison.

The teacher then constructs parallel structure between the utterances in lines 2-4 and lines 8-10. Analogies are frequently formalized as \(\mathrm{A}: \mathrm{B}:: \mathrm{C}: \mathrm{D}\) ("A" is to " B " as " C " is to " D "), and this teacher implements that formal relationship in the following pattern of talk: "whatever you do to \((A)\) you have to do to \((B)\) " and "whatever you do to \((C)\) you have to do to \((D)\)." The statements are lexically identical around the arguments \((A, B, C, D)\), which are conceptually similar objects. "Whatever you do to one side, you have to do to the other, whatever you do to the numerator, you have to do to the denominator."

The parallel structure is further supported by the teacher's gesture that builds on culturally standard spatial representations of fractions and equations. For equations, "one side," "and the other" are typically depicted as horizontal objects to the left or right of the other. For fractions, "denominator" and "numerator" are vertical objects, one below and above the other.

These symbolic representations are reiterated by the teacher's gestures. In accordance with her verbalization of the source "whatever you do to one side you have to do to the other" she moves her hands to her left and then her right. In construction of the target she mirrors the opposing movements to signal the numerator and the denominator, and moves her hands from towards herself to away from herself. The teacher first designs
the relationship between the A and B components of the source (one side of an equation and the other) and then the relationship between the C and D components of the target (the numerator and the denominator). The overarching lesson has been focusing on equivalence, so it is clear from the setting of this talk that "have to" implies 'have to in order to maintain equivalence between the two sides.'

The parallel structure is compelling to the analogy recipients and they demonstrate uptake of the parallel structure and appropriate inferences based on acquisition of the relational structure of the talk. The teacher leaves a micro-pause as invitation to participation for students in lines 4 and 10, requesting their participation in generating the \(B\) and \(D\) terms of the parallel structure. In both cases multiple students within the classroom enter the discourse, and in both cases the audible set of students respond appropriately with the correct lexical item to complete the conceptual relationship signaled by the parallel structure. In line six students also demonstrated acquisition of the parallel structure, and overlapped with the teacher in production of the completion of the phrase using the modifier "to" preceding "the other".

These utterances provide evidence that these students are participating actively in the parallel structure, as well as the corresponding mathematically relevant relational mapping designed by the teacher. Their answers are not necessarily based on problem solving, though, but rather they may be mapping the structure highlighted by the teacher's gesture and parallel discursive structure from a known source object to a corresponding target.

\section*{Analogy Segment 2: Indexicality}

The following transcript, shown in Figure 2, provides a second example of the role of hypothetical contexts in construction of analogy. Of particular interest in this analogy is the teacher's persistence in indexing an alternative context that is familiar to her student recipients.

The teacher initiates this analogy to help a student determine whether the summation of two negative numbers results in a negative or a positive sign in front of the solution number. Answering this question is the target of this analogy, and the teacher invokes the familiar schema of losing money as a source.

This is a one-on-one interaction between one student and the teacher during the seatwork portion of this lesson. The student has raised her hand and indicated difficulty to the teacher, who then comes to her assistance. Approximately half of the analogies identified in the coding study were constructed following students' demonstration of difficulty with the mathematics. Many of these looked similar to this analogy.
```

T: You're- you're saying
(..) what's a, negative eighty eight if you lose
(..) let's say you have um (..)
you're playing marb-
well, uh people don't play marbles anymore.
T: Let's say you're (..) you got money, all right,
and you lose eighty eight cents
and then you lose five cents.
What have you lost altogether?
Ninety three cents.
Right.
So you wouldn't want to say plus ninety three.
: Want to subtract ninety three.

```

Figure 2. Analogy between losing money and subtracting negative numbers.

In lines 1 through 8 the student is voicing her confusion, though the details of the language are difficult to capture in the recording. In line 10 this teacher begins to signal that she is going to re-represent the question entered by the student. She begins with "you're - you're saying" which suggests a reformulation of the students' question, but then after a brief pause she begins again with: "what's a, negative eighty eight if you lose..." The lexical item "lose" indexes possession and change of that status, and is not a mathematical term. This choice of term signals that she is representing the students' mathematical question in an alternative domain. Next she pauses briefly and begins again. This time the teacher uses a plural pronoun "let's" indexing that both she and the student will participate in the reformulation of the student's mathematical question and uses the frame "lets say you have um." The teacher indexes the hypothetical (irrealis) frame through the lexical item "say," indicating that this is a reformulation in a world not exactly the same as the one inhabited by the student's initial question. Again, however, she aborts this reconstruction and after a brief pause reformulates with the frame "you're playing marb-."

Once again the teacher decides to discontinue this representation and the setting of marbles because, as she states, "people don't play marbles anymore." This statement reveals that it is important to her that the context she indexes as a frame for her reformulation of the students' question is one that the student regularly inhabits or is familiar with. After several attempts to initiate this representation of the student's original question, the teacher signals an alternative context through a shift in semantic as well as indexical word use, and settles on a hypothetical reformulation.

In lines 15 through 17 the teacher completes her representation of the irrealis source analog. She says "let's say you're (..) you got money," and indexes the student's nonverbal concurrence with the phrase "all right." Once again the teacher uses the plural "lets say" construction to signal that this is a reformulation of the original math problem, and that this is instead of the marble-playing context referenced immediately prior. The teacher continues by embedding the student's
original mathematical question in the context she is building in which the student's possession of money is the relevant feature "you lose eighty eight cents and then you lose five cents." She develops the source context as a hypothetical world in which the important point is that the student has money. The construction "lets say you're, you got money" suggests that regardless of whether this student actually has money, the teacher is indexing this possible world in which this student has 88 cents and loses 5 cents. These are the same numerical amounts as in the original target problem, yet they are situated within this hypothetical frame.

The question "what have you lost altogether" in line 18 is a reformulation of the target and requires the same mathematical computation, but the contextualization and the lexical item "lost" indexes that this problem is distinct from the mathematical problem and is located within the domain of money.

In line 19 the student answers appropriately to the hypothetical context of the source analog using the monetary unit, "cents" to describe the numerical solution. She indicates that she is embodying the "you" from inside the hypothetical possible world represented by the teacher, as she answers the questions "what have you lost altogether" without hesitation and with the correct number.

Finally, the teacher guides the student in transitioning from her facility with the hypothetical world of her monetary loss to the veridical world of the math problem. Still using the student as a reference point, in line 21 she says "so you wouldn't want to say plus ninety-three." The teacher's use of the term "plus" and the transitional item "so" index the mathematical world, and appropriately the student responds with: "want to subtract ninety-three." This correct answer is stated without markers of money, and specifically the term "subtract" is used for the same computational meaning as "lost" was in the earlier line 18. Thus the student has made the relevant conceptual inference that adding negatives results in a negative number, and she has made the interactional inference that she is now in the realm of the math problem, where she had previously inferred the context of her monetary loss.

Thus in this analogy, like in the prior example, the teacher's language denotationally constructs an analogy between a familiar and an unknown context, but her discourse also indexes both interactional and semantic mappings. Thus the student must exercise conceptual mapping and inferences at multiple levels. She must interpret her role in the analogy, as well as multiple levels of the mathematical comparison. The teacher indexes levels of comparison between the numbers as well as hypothetical to veridical worlds, and between a world where the student is within the context to where she external to the math and writes a mathematical
answer. The teacher's work to find a source context that is familiar and a realistic hypothetical situation may facilitate these levels of inference, enabling the student to draw inferences from the more familiar space to the more novel space.

The source analog of losing money thus provides a meaningful structure for this student to interpret the nature of addition between negative numbers, a concept that is currently unfamiliar to her. Her success within a few seconds demonstrates that this is a striking resource for meaning-making. The conceptual structure of negative numbers is typically challenging for learners, and this teacher has led to an extremely rapid successful completion of a target problem following confusion.

At the same time, this rapid transformation is somewhat troubling from a learning standpoint. The mathematical nature of this analogy is not deeply conceptual, nor is it clear that the student will be able to generalize this understanding to a new problem in which the teacher has not highly designed a source analog for the student. The student herself will have to relationally re-represent the current problem as a source for a subsequent problem, and her ability to do so remains to be seen.

This reveals a powerful tension between the interactional success of an analogy produced in conversation, and the goal to produce deep thinking and conceptual abstraction from an analogy. The teachers' highly constrained representations of the source analogs improve the likelihood that recipients will use the alignment they have been provided. At the same time, this may limit the need for effortful relational integration and structure mapping on the part of students, potentially limiting future ability for transfer and generalization.

Overall, these examples are both successful interactions in which students reason analogically to respond as pragmatically and mathematically intended by their instructor. Regarding learning, however, the pragmatic and referentio-semantical efficacy of the interaction are impossible to disentangle. The interactions may have prompted minimally effortful relational integration because the source objects were highly relationally re-represented by the teacher. That relational re-representation in the first example created a parallel poetic structure between the source and target representations, which required structure-mapping but could be accomplished through attention to the pragmatics, rather than only referentio-semantic/ mathematical content as one might suppose if solely examining the source and target representations being compared. This suggests that the analogical paradox may be at least partly explained by the grammatical, interactional pragmatics of everyday verbal analogies.

\section*{References}

Doumas, L. A. A., \& Hummel, J. E. (2012). Computational models of higher-cognition. The Oxford Handbook of Thinking and Reasoning, 52-66
Hanks, W. F. (1992). The Indexical Ground of Deictic Reference. In C. G. Allesandro Duranti (Ed.), Rethinking Context; Language as an Interactive Phenomenon (pp. 43-76). Cambridge: Cambridge University Press.
Horn, L. R. (1988). Pragmatic Theory. In F. J. Newmeyer (Ed.), Linguistics: The Cambridge Survey (Vol. 1, pp. 113-145). Cambridge: Cambridge University Press.
Hymes, D. (1972). Models of the Interaction of Language and Social Life. In J. J.
Gumperz \& D. Hymes (Eds.), Directions in Sociolinguistics: The Ethnography ofCommunication. New York: Holt, Rinehart \& Winston.
Jakobson, R. (1960). Closing Statement: Linguistics and Poetics. In T. Seboek (Ed.), Style in Language (pp. 398-429). Cambridge, MA: MIT Press.
Lucy, J. A. (1999). Reflexivity. Journal of Linguistic Anthropology 9(1-2): 212-215.
Ochs, E. (1992). Indexing Gender. In A. Duranti \& C. Goodwin (Eds). Rethinking Context: Language as an interactive phenomenon. (pp. 335-358. Cambridge: Cambridge University Press.
Richland, L.E.; Holyoak K.J. \& Stigler, J.W. (2004). Analogy generation in eighth-grade mathematics classrooms. Cognition and Instruction 22(1), p.37-60.
Richland, L.E., Zur, O., \& Holyoak, K. (2007). Cognitive supports for Analogies in the mathematics classroom. Science, 316, 1128-1129.
Sacks, H., Schegloff, E.A., \& Jefferson, G. (1974). A simplest systematics for the organization of turn taking for conversation. Language, 50: 4, 969-735.
Silverstein, M. (1976). Shifters, Linguistic Categories and Cultural Description. In K. H. Basso \& H. A. Selby (Eds.), Meaning in Anthropology (pp. 11-56). Albuquerque: University of New Mexico Press.
Silverstein, M. (1985). On the pragmatic "poetry" of prose. In D. Schiffrin (Ed.), Meaning, form and use in context (pp. 181-199). Washington, DC: Georgetown University.
Wortham, S. \& Rymes, B. (Eds.). (2003). Linguistic anthropology of education Westport, CT:
Rattermann, M. J., \& Gentner, D. (1998). More evidence for a relational shift in the development of analogy: Children's performance on a causal-mapping task. Cognitive Development, 13, 453-478.
Richland, L.E., Morrison, R.G., \& Holyoak, K.J. (2006). Children's development of analogical reasoning: Insights from scene analogy problems. Journal of Exp Child Psychology, 94, 249-273.
Viskontas, I.V., Morrison, R.G., Holyoak, K.J., Hummel, J.E., \& Knowlton, B.J., (2004) Relational integration, inhibition and analogical reasoning in older adults. Psychology and Aging, 19, 581-591.

\title{
Experimental insights on the origin of combinatoriality
}

\author{
Gareth Roberts (gareth.roberts@yu.edu) \\ Department of Psychology, Yeshiva University \\ 2495 Amsterdam Avenue \\ New York NY 10033 USA
}

\author{
Bruno Galantucci (bruno.galantucci@yu.edu) \\ Department of Psychology, Yeshiva University 2495 Amsterdam Avenue \\ New York NY 10033 USA
}

\begin{abstract}
Combinatoriality-the recombination of a small set of basic forms to create an infinite number of meaningful units-has long been seen as a core design feature of language, but its origins remain uncertain. Two hypotheses have been suggested. The first is that combinatoriality is a necessary solution to the problem of conveying a large number of meanings; the second is that it arises as a consequence of conventionalisation. We tested these hypotheses in an experimental-semiotics study. Our results supported the hypothesis based on conventionalisation but offered little support for the hypothesis based on the number of meanings.
\end{abstract}

Keywords: Experimental semiotics; Human communication; Language.

In the vast majority of languages, a small set of basic meaningless forms (typically phonemes) are recombined to create an infinite number of meaningful units (typically morphemes). This property, which we shall refer to as combinatoriality, has been identified as a core design feature of language (Abler, 1989; Hockett, 1960; Hurford, 2002; Jackendoff, 1999; Martinet, 1960; Studdert-Kennedy \& Goldstein, 2003). Its origins, however, are unclear.

\section*{Explaining combinatoriality}

\section*{Set-size}

Any communication system must employ a set of signs: mappings between signals (such as vocalisations or manual gestures) and referents (the things in the world to which the signals refer). A long-standing explanation for combinatoriality concerns the size of this set. If the signals in the set are distinguished on the basis of analogue contrasts, it will become harder to distinguish them as the set increases in size; restructuring the system in terms of discrete forms is an efficient solution to this problem (Hockett, 1960; Nowak, Krakauer, \& Dress, 1999; Studdert-Kennedy, 2000). If this is the case, we should expect combinatoriality to increase as set-size increases.

\section*{Mimesis and transparency}

Recent research on Al-Sayyid Bedouin Sign Language (ABSL) raises a problem for explanations of combinatoriality based on set-size. ABSL is a fully fledged language, which does not differ substantially from other languages in terms of set-size, but which exhibits very little combinatoriality (Sandler, Aronoff, Meir, \& Padden, 2011). Furthermore,
while other known sign languages do exhibit combinatoriality, they tend to employ sets of basic forms that are an order of magnitude larger than those employed in spoken languages (Liddell \& Johnson, 1989).

Another difference between spoken and signed languages is the degree to which they afford mimesis-that is, the degree to which signals are intuitively motivated by what they refer to. In sign language mimesis is both richer and much more frequent than in speech (Fusellier-Souza, 2006; Perniss, Thompson, \& Vigliocco, 2010; Meier, 2002; Taub, 2001).

Sandler et al. (2011) suggested that the beginnings of combinatorial structure in ABSL may be explained by conventionalisation, whereby the signals becomes less transparently mimetic. The ABSL sign for LEMON involves a transparently mimetic signal, in which the signer mimes the act of squeezing a lemon. Since there is more than one way to squeeze a lemon, the form of the signal varies among signers (Sandler et al., 2011, p. 519), but this does not hinder communication so long as the signal remains transparently mimetic. If this transparency is lost, however, the form of the sign can no longer vary to the same extent, but also no longer needs to be constrained by the referent. It is then more efficient to structure signals according to basic sensory-motor constraints, which are best satisfied by a small set of forms (Studdert-Kennedy \& Goldstein, 2003).

\section*{Two pathways}

It should be noted that the two explanations for the emergence of combinatoriality illustrated above are not mutually exclusive. It is possible that the emergence of combinatoriality is related to both set-size and transparency and that there is a complex relationship between the three. Since signs are easier to establish if there is greater opportunity for grounding them in something familiar (Galantucci, 2005; Scott-Phillips, Kirby, \& Ritchie, 2009), and transparent signs are by definition grounded, greater transparency should allow rapid growth in set-size, which-as suggested above-may in turn encourage greater combinatoriality. On the other hand, Sandler et al. (2011) explain combinatoriality as a response to low transparency. In other words, there is reason to expect both low transparency and high transparency to lead ultimately to combinatoriality, albeit by different routes. This may go some way to explaining the ubiquity of combinato-
rial structure in the world's languages. Moreover, given that the route from high transparency to combinatoriality is more indirect, it seems likely that combinatoriality takes longer to arise in systems that afford highly transparent signs. This may explain why ABSL still exhibits so little combinatoriality.

\section*{Investigating combinatoriality in the laboratory}

It is very rare for linguists to have the opportunity to observe and record a new language in its early stages, making such insights as Sandler et al.'s on the origin of combinatoriality very hard to come by. Moreover, new languages tend to emerge in unusual cultural settings, making generalisation difficult. An alternative approach, which has been referred to as Experimental Semiotics, is to study the emergence of novel communication systems under laboratory conditions (Galantucci, Garrod, \& Roberts, 2012). Del Giudice (2012) and Verhoef (2012), for example, examined the emergence of combinatoriality in sign systems in diffusion chains, but without any pressure to communicate. Here we present data from a laboratory study in which combinatoriality emerged in sign systems used by pairs of participants to communicate with each other.

\section*{Method}

Participants 12 pairs of participants (4 female-female; 4 male-male; 4 mixed) participated in the study for course credit or monetary compensation.

The game Participants played a cooperative guessing game, sitting in separate locations with the same set of four images (henceforth referents) displayed in random locations in a 5-by-5 grid on a video monitor (see Figure 1). The game consisted of a series of rounds. In each round, one player would play as "sender" and the other as "receiver". The sender was informed of a target referent and had to convey this referent to the receiver so that the receiver could select it on their screen. If the receiver selected the correct target the round was counted as successful; if not, the round was counted as unsuccessful. Since the players played in separate locations over the internet, they could not speak to each other directly. Instead, the sender could communicate with the receiver exclusively through the use of a digitising pad and a magnetic stylus. The tracings that the sender made on this pad were transformed into on-screen signals in a systematic way: While the horizontal component of the tracings determined the horizontal component of the signal seen on the screen, the vertical component of the tracing was ignored and replaced by a simple downward movement at a constant rate (Figure 2a). The resulting signals were relayed to the screens of both players in real time. Players could not use this pad as an effective drawing or writing device (Figure 2b), even after prolonged practice, and to succeed at the task pairs of players had to cooperatively develop novel forms of communication (Galantucci, 2005).To help them in this, both players received feedback after each selection. Specifically, the receiver was shown what the target image had been and the sender was


Figure 1: Screenshot from early stage of game. The screen on the left was the Sender's screen; the screen on the right was the Receiver's.
shown which image the receiver had selected. After the feedback phase, the next round began. Players swapped sender and receiver roles after each round.

The referents were presented as targets in a random order: Pairs iterated through four referents twice every eight rounds (in random order). A performance score was kept updated for each referent, based on the proportion of successful rounds in the cycle. If a pair had at least \(75 \%\) success on each of the four referents, the number of referents in the set was increased to eight, and the cycle length was increased accordingly to 16 rounds. The referent set and cycle length continued to be incremented in this way until either players had mastered a set of 20 referents or two hours of playing had elapsed.


Figure 2: (a) How the drawings players produced on the digitising pad appeared on screen. (b) How common graphic symbols drawn on the digitising pad appeared on the screen

Referents The referents used were black silhouettes of animals (see Figure 3). These silhouettes afforded the opportunity to develop signals with some degree of transparency, in which, for example, features of the silhouette (e.g. the trunk of the elephant) could be represented by a feature of the signal (e.g. a long curved line). However, the way in which their tracings were transformed did not allow players to reproduce the animal silhouettes or even to create simple drawings. In terms of the hypotheses described above, in other words, it was biased towards relatively low-transparency signals.


Figure 3: Referents used in the game. The top row shows the referents that were visible to players at the start of the game.

\section*{Results}

\section*{Measures}

All of the events in the game were recorded and three measures were derived from this data set: Set-size, Transparency, and Combinatoriality.
Set-size Following the experiment, a sign-set was constructed for each player. This consisted of every referent on which the dyad had reached at least \(75 \%\) success, paired with the last successful signal the player in question had used to communicate it. The Set-size for a pair was computed as the mean of the Set-sizes for the two players in the pair. The mean Set-size for the 12 pairs was 14.67 ( \(\mathrm{SD}=3.75\) ); the smallest sign-set contained six signs, and the largest contained 20.
Transparency The more transparent the relationship between a signal and a referent, the easier it should be for an independent judge to match them up. Four judges, who had no previous familiarity with the signs or with the purpose of the study, matched signals with referents. This was done as follows. First, the judges gained an understanding of the game by playing a few rounds themselves (as both sender and receiver, with pictures of faces as referents). Then they were shown a display containing one player's signals (as playable videos) along with the referents they referred to. Their task was to match the former with the latter. To give them as much opportunity as possible to detect relationships between signals and referents, judges were permitted to take as long as they liked and to change their minds as often as they liked. Once they had finished, another player's sign-set would appear. (The order in which the sign-sets appeared was randomised.) Each judge evaluated one sign-set from every pair of players ( 12 sets in total) and every sign-set was shown to two judges. The number of correct matches made by each judge for each player's sign-set provided an indication of the set's Transparency to that judge. This was converted to a zscore by subtracting the mean number of correct matches we


Figure 4: Correlations between: a) Transparency (T) and Set-size (S); b) Combinatoriality (C) and Set-size; c) Transparency and Combinatoriality; d) Transparency and Combinatoriality, with Set-size partialled out.
would expect, for that size of set, by chance and dividing the result by the standard deviation of that mean. Since every player sign-set was rated by two judges, the mean of the zscores for the two judges was taken as the Transparency index for the set in question. Finally, the Transparency for a pair sign-set was computed as the mean of the Transparency for the two players in the pair. The overall mean Transparency for the 12 pairs was .73 ( \(\mathrm{SD}=.76\) ), ranging from -.25 to 2.5 .

Combinatoriality Combinatoriality was measured using a slightly modified version of the Form Recombination Index used by Galantucci, Kroos, and Rhodes (2010). This measure breaks a sign into forms (parts of a sign divided by empty space). Forms within the sign are then compared with each other to remove duplicates, and the remaining forms are compared with all other forms in the system. The number of matches among these forms is then divided by the total number of comparisons to produce an index ranging from 0 to 1 (where 0 corresponds to a complete absence of Combinatoriality and 1 corresponds to maximal Combinatoriality). A system in which a small number of unique forms are reused many times will have higher Combinatoriality than a system in which a large number of forms are reused little. The mean Combinatoriality for the 12 pairs was \(.06(\mathrm{SD}=.04)\), ranging from . 01 to . 17 .

Correlations As can be seen in Figure 4, there was a strong positive correlation between Set-size and Transparency, r(10) \(=.65, p=.02\), a weak positive correlation between Set-
size and Combinatoriality, \(\mathrm{r}(10)=.33, p=.3\), and a negative correlation between Transparency and Combinatoriality, \(\mathrm{r}(10)=-.26, p=.42\). The strong correlation between Set-size and Transparency supports the hypothesis suggested above that more transparent signs are easier to ground, leading sign systems to grow faster. The presence of this correlation, however, poses a problem for interpreting the correlation between Transparency and Combinatoriality. That is, the positive correlation between Set-size and Combinatoriality interferes-via the positive correlation between Set-size and Transparency-with the negative correlation between Transparency and Combinatoriality. We therefore partialed out Setsize from the latter, and this revealed a much stronger correlation, \(\mathrm{r}(9)=-.65, p=.01\). This result is consistent with the hypothesis that Combinatoriality emerges as a response to low Transparency. The general pattern of results is also consistent with the hypothesis that high Transparency leads to Combinatoriality via Set-size, but the correlation between Set-size and Combinatoriality is too weak to say anything conclusive in this regard.

\section*{Conclusion}

Theoretical work and research on ABSL suggest two hypotheses to explain the emergence of combinatoriality. The first is that it arises as a solution to the problem of conveying a large number of meanings (Hockett, 1960; Nowak et al., 1999; Studdert-Kennedy, 2000). The second is that it arises as a consequence of conventionalisation, as mimetic signs lose transparency (Sandler et al., 2011). As in other experimental-semiotic studies (Galantucci et al., 2010; Del Giudice, 2012) our analysis of laboratory data did not lend much support to the first hypothesis (although, as suggested by Galantucci et al., 2010, it is possible that set-size exercises an effect on combinatoriality only after some threshold is reached). Our analysis lends the most support to the second hypothesis: Combinatoriality arises when signals loseor never possess-a mimetic link with their referents.

\section*{References}

Abler, W. L. (1989). On the particulate principle of selfdiversifying systems. Journal of Social and Biological Structures, 12(1), 1-13.
Del Giudice, A. (2012). The emergence of duality of patterning through iterated learning: Precursors to phonology in a visual lexicon. Language and Cognition, 4(4), 381-418.
Fusellier-Souza, I. (2006). Emergence and development of signed languages: From a semiogenetic point of view. Sign Language Studies, 7(1), 30-56.
Galantucci, B. (2005). An experimental study of the emergence of human communication systems. Cognitive Science, 29(5), 737-67.
Galantucci, B., Garrod, S., \& Roberts, G. (2012). Experimental Semiotics. Language and Linguistics Compass.
Galantucci, B., Kroos, C., \& Rhodes, T. (2010). The effects of rapidity of fading on communication systems. Interaction Studies, 11(10), 100-111.

Hockett, C. F. (1960). The origin of speech. Scientific American, 203, 88-96.
Hurford, J. R. (2002). The role of expression and representation in language evolution. In A. W (Ed.), The transition to language (p. 311-334). Oxford: Oxford University Press.
Jackendoff, R. (1999). Possible stages in the evolution of the language capacity. Trends in Cognitive Science, 3(7), 272-279.
Liddell, S. K., \& Johnson, R. E. (1989). American sign language: The phonological base. Sign Language Studies, 64, 195-278.
Martinet, A. (1960). Elements of general linguistics. Chicago: University of Chicago Press.
Meier, R. (2002). Why different, why the same? explaining effects and non-effects of modality upon linguistic structure in sign and speech. In R. Meier, K. Cormier, \& D. Quinto-Pozos (Eds.), Modality and structure in signed and spoken languages (pp. 1-25). New York: Cambridge University Press.
Nowak, M. A., Krakauer, D. C., \& Dress, A. (1999). An error limit for the evolution of language. Proceedings of the Royal Society of London Series B-Biological Sciences, 266(1433), 2131-2136.
Perniss, P., Thompson, R. L., \& Vigliocco, G. (2010). Iconicity as a general property of language: Evidence from spoken and signed languages. Frontiers in Psychology, 1, 115.

Sandler, W., Aronoff, M., Meir, I., \& Padden, C. (2011). The gradual emergence of phonological form in a new language. Natural Language \& Linguistic Theory, 29(2), 503543.

Scott-Phillips, T. C., Kirby, S., \& Ritchie, G. R. S. (2009). Signalling signalhood and the emergence of communication. Cognition, 113(2), 226-233.
Studdert-Kennedy, M. (2000). Evolutionary implications of the particulate principle: Imitation and the dissociation of phonetic form from semantic function. In C. Knight, M. Studdert-Kennedy, \& J. R. Hurford (Eds.), The evolutionary emergence of language (pp. 161-176). Cambridge: Cambridge University Press.
Studdert-Kennedy, M., \& Goldstein, L. (2003). Launching language: The gestural origin of discrete infinity. In S. Kirby \& M. H. Christiansen (Eds.), Language evolution (pp. 235-254). Oxford: Oxford University Press.
Taub, S. F. (2001). Language from the body. Cambridge: Cambridge University Press.
Verhoef, T. (2012). The origins of duality of patterning in artificial whistled languages. Language and Cognition, 4(4), 357-380.

\title{
From Minor Mishap to Major Catastrophe: Lexical Choice in Miscommunication
}

\author{
Jennifer M. Roche (jroche@bcs.rochester.edu) \({ }^{\text {a }}\) Alexandra Paxton (paxton.alexandra@gmail.com) \({ }^{\text {b }}\) Alyssa Ibarra (aibarra@bcs.rochester.edu) \({ }^{\text {a }}\) Michael K. Tanenhaus (mtan@bcs.rochester.edu) \({ }^{\text {a }}\) \\ \({ }^{\mathrm{a}}\) Brain and Cognitive Sciences, University of Rochester \\ Rochester, NY 14627 USA \\ \({ }^{\text {b }}\) Cognitive and Information Sciences, University of California, Merced \\ Merced, CA 95343 USA
}

\begin{abstract}
Miscommunication is often regarded as noise or uninformative in psycholinguistic research. However, Coupland et al. (1991) suggest that miscommunication can provide rich information about how interlocutors come to communicate successfully. Successful communication necessarily needs the individuals involved to coordinate and update their mutual knowledge, experiences, beliefs, and assumptions. However, the process of updating this information may be ridden with unsuccessful attempts that eventually help interlocutors reach a common goal. This study evaluates the relative contribution of linguistic factors to communicative success, based on verbal grounding (e.g., mutual agreement on a referent) and visual congruency (e.g., interlocutor's visual environments match or mismatch) during a collaborative task. We show that varying levels of communicative success are laden with rich linguistic information that may uncover interesting aspects of successful and less successful communication.
\end{abstract}

Keywords: Joint action; grounding; successful communication; miscommunication; psycholinguistics.

\section*{Introduction}

Interactive language, in particular face-to-face interactive conversation, is the most canonical form of language use (Clark, 1992; Goodwin \& Duranti, 1992). In interactive conversation, interlocutors are typically both speakers and listeners (addressees) and they often are conversing to achieve joint goals. Nonetheless, most research on human language processing focuses on the speaker and the listener as individual cognitive agents in non-interactive tasks.

There are important exceptions. For example, a large body of work has used the Edinburgh Map Task (Brown et al., 1983) to address a range of psycholinguistic issues. In this task two interlocutors collaborate, with the director guiding the matcher to reproduce a route printed on the director's map. Aist and colleagues developed a "fruit cart" domain as a vehicle for eliciting human language production for (a) dialogue system research and development and (b) psycholinguistic research (Aist, Campana Allen, Swift \& Tanenhaus, 2012). Senft \((2002,2007)\) developed a number of domains to evaluate lexical choice in spatial terms during
a space game and tinker toy task for cross-cultural analysis. Brown-Schmidt, Tanenhaus and colleagues have adopted a complementary strategy, using targeted language games to produce trial-like structure in unrestricted interactive conversation to address specific psycholinguistic issues with real-time response measures, such as visual world eyetracking (e.g., Brown-Schmidt, Gunlogson \& Tanenhaus, 2008; Brown-Schmidt \& Tanenhaus, 2008).

In this paper we provide a preliminary report on a project using a new domain intended to examine how referential domains are constructed, updated, accepted and rejected during a goal-driven task, with naïve participants and unrestricted speech. Here we examine how the language used in grounding might be diagnostic of, and contribute to, miscommunication.

The domain is similar to those discussed by Sentf (2002) and is designed to allow a face-to-face interaction through a barrier separating the two participants. This task involves a collaborative dyadic interaction that required participants to instruct each other in building a \(\mathrm{b} \bullet \operatorname{loc}^{\bullet} \mathrm{o}^{\mathrm{TM}}\) animal figure from abstract three-dimensional puzzle-like pieces (see Figure 1 for an image of the animals; Methods for full task description).

The \(\mathrm{b} \cdot \operatorname{loc} \cdot \mathrm{o}^{\mathrm{TM}}\) paradigm was created to serve a number of purposes. First, we wanted a domain that would lend itself to investigating both generation and interpretation of referring expressions. Secondly, we wanted to observe how referring expressions change when the goals change. For example, during the build stage pieces that were initially referred to using conceptual pacts, such as "the Christmas tree" would eventually assume a different identity, "the body" (see the green item on the left of Figure 2). This domain offers a rich domain for investigating grounding.

The domain creates a corpus that contained frequent communication failures (e.g., confusions and misunderstandings) that had to be resolved. These failures in communication are often regarded as noise and therefore uninformative in psycholinguistic research (Coupland, Giles, \& Weimann, 1991; Keysar, 2007). However, as

Coupland et al. (1991) argue, communication failure could provide valuable information about how interlocutors come to communicate successfully, much like speech errors can provide important insights into planning processes in language production.

Successful communication necessarily requires interlocutors to coordinate and regularly update their mutual knowledge, experiences, beliefs, and assumptions (e.g., Clark \& Carlson, 1982; Clark \& Marshal, 1981). However, the process of updating this information may be riddled with unsuccessful attempts that eventually help the interlocutors reach a common goal.

Some researchers have provided insights into how interlocutors might resolve communication problems (e.g., through ambiguity resolution or asking clarification questions; see Clark \& Brennan, 1991; Haywood, Pickering \& Branigan, 2004; Levelt, 1983; Pickering \& Garrod, 2004; Roche, Dale, Jaeger, \& Kreuz, under revision). However, the literature has minimally, at best, explored the rich information these failures could provide. For example, interlocutors' language is often ambiguous because ambiguity minimizes effort in production and because a speaker can usually assume that her addressee can rapidly use context to infer her intended meaning, perhaps because it is easier on the production system (e.g., Bard et al., 2007; Bock, 1986). This can result in utterances that initially appear to be egocentric. However, once an interlocutor realizes that her ambiguity might reduce the success of the interaction, she almost immediately adapts her utterances to eliminate the type of ambiguity that was confusing for her listener (Roche, Dale, Jaeger \& Kreuz, under revision).

Despite these efforts, the focus of the existing literature has been primarily on the successful exchange of information and largely ignores what happens when interlocutors' shared knowledge becomes de-coupled. Yet, miscommunication occurs regularly and can directly impact the quality and effectiveness of an interaction (McTear, 2008). Therefore, the current study provides a preliminary analysis of how language reflects and perhaps influences communicative successes and failures.

\section*{Methods}

\section*{Participants}

Participants were 20 dyads of paid undergraduate students ( \(N=40\); females \(=26\); mean age \(=19\) years) from the University of Rochester. Participants were native speakers of American English. All reported normal to corrected vision and no speech or hearing impairments.

\section*{Stimuli}

The experiment included two types of \(b \cdot l o c \cdot 0^{\mathrm{TM}}\) animal figures (see Figure 1).


Figure 1: Images of the grasshopper (left) and lizard (right) animal figures used in the task.

The grasshopper figure consisted of 25 pieces, and the lizard figure consisted of 28 pieces. Each animal piece was abstract and did not have a proscribed name (see Figure 2 for example pieces).


Figure 2: Sample of items from the animal figures.
Instruction Cards Each animal figure had a set of instruction cards, each corresponding to one step of the building process. The grasshopper and lizard figures consisted of 13 and 15 instructions cards each, respectively (see Figure 3 for sample instruction cards).


Figure 3: Sample of the animal figure instruction cards (right: grasshopper; left: lizard).

Conditions We had a between-subjects condition, in which dyads instructed each other in collaboratively building an animal (lizard or grasshopper), but were separated by a partition \({ }^{1}\).

Data Recording Three digital cameras recorded the participant interaction from different angles (left, centerwide, right). All video files were time-aligned and compressed into a single .mov file using Final Cut Pro.

\section*{Procedures}

Participants were seated across from each other, separated by a partition. The participants were given identical sets of \(\mathrm{b} \cdot \mathrm{loc}^{\cdot} \cdot \mathrm{o}^{\mathrm{TM}}\) pieces on identical workspaces. Workspaces featured a flat, white surface with a black box outline drawn in each corner. Participants were told that they would be working together to build identical objects. They were not, however, told what the resulting object would become. The

\footnotetext{
\({ }^{1}\) The design of the experiment was more complex, including a non-hidden phase that was not analyzed here. In the full design, the animals were counterbalanced across the possible conditions. For the purpose of the present analyses, only a subset of the data is included here.
}
experiment was divided into two phases: an Item Phase and a Building Phase.

During the Item Phase, the participants moved the individual \(\mathrm{b} \cdot \operatorname{loc}^{\bullet} \mathrm{o}^{\mathrm{TM}}\) pieces into the four boxes on their workspace. They were told that their workspaces had to match before they could proceed to the next phase. They were further instructed to take turns and decide together how to separate their items.

Once the workspaces matched, they were allowed to continue to the Build Phase. During the Build Phase, each participant was given half of the instruction cards in a predefined order. Participants alternated giving instructions. They were told that they could ask each other questions and, more generally, talk freely with one another. The majority of pairs successfully built matching objects. The unsuccessful pairs made only minor errors (e.g., wrong orientation of the animal's legs).

\section*{Measures}

Transcription and coding of various behaviors were annotated from a single workable file that contained a compressed version of the video files from the three different angles (left, center-wide, right) to aid in coding.

\section*{Coded Measures}

The video files of each dyad's interactions were transcribed. After transcription, additional measures were coded and included the following categories: confirmed and negated utterances, visual congruency, and several standard LIWC categories (Linguistic Inquiry and Word Count; Pennebaker, Booth, \& Francis, 2001).

Confirmed and Negated Utterances: We divided utterances that presented new information into two categories according to whether they were confirmed (e.g., yes, uh huh) or negated by the addressee. An utterance was coded as Confirmed if the partner indicated acceptance of the new information with an explicit confirmation (e.g., with a variant of yes) similar to the Common Ground Units described by Nakatani and Traum (1999). At each turn, \(T_{1}\) presented a new piece of information. Once \(T_{2}\) accepted this information it was coded as a Confirmed utterance. For example:
\(\mathrm{T}_{1}\) : Uh, that piece, uh it's in the, the center of box three, it looks like a bell.
\(\mathrm{T}_{2}\) : Mhm.
An utterance was coded as Negated if presentation of new information by \(T_{1}\) was negated or rejected by the \(T_{2}\). For example:
\(\mathrm{T}_{1}\) : Ok, so what was the? Put it three rows down.
\(\mathrm{T}_{2}\) : No, no, no, no, no, three squares to the right.
Visual Congruence. We coded the participants' workspaces as either matching or mismatching (congruent) or
mismatching (incongruent) throughout the task (e.g., the orientation of the object being described). Here we focus on within-trial instances of congruent and incongruent targets. Congruency and Confirmed/Negated utterances were used to create a \(2 \times 2\) contingency table of the different types of communicative success (see next section).

Communicative Success was measured relative to the congruent and incongruent physical environments in conjunction with the verbal acknowledgement of the information presented. Often, interlocutors believed their objects were congruent when in fact they mismatched in ways that interfered with the goal in that trial. A contingency table illustrates the four types of outcomes created by crossing confirmed and negated utterances with object congruency: Confirmed Congruent (CC), Confirmed Incongruent (CI), Negated Congruent (NC), and Negated Incongruent (NI; see Table 1 for the outcome labels).

A CC outcome is an instance of Successful Communication. The new information is confirmed (and acted upon) and the objects are indeed visually congruent. For example, one member of the pair says, "Yes, I got it," when in fact she did "get" it. A CI (Unrecognized Miscommunication) outcome occurs when one of the participants accepts the information presented, but the objects in her visual workspace do not match those of her partner (e.g., saying, "Yes, I got it," when she did not actually "get" it). The pair believes they have successfully communicated, but in fact they have not. A NC outcome (Unrecognized Success) occurs when a participant negates her partner's statement, but her visual workspace objects matches her partner's (e.g., saying, "No, I didn't get it," when she actually did "get" it). The pair has actually succeeded but believe they have not. Finally, in an NI outcome, the pair has recognized the miscommunication. Recognized Miscommunication occurs when at least one of the participants fails to ground, and their visual workspace objects do not match (e.g., saying, "No, I didn't get it," when in fact she did not "get" it).

Table 1: Communicative Success Outcome Variables.
\begin{tabular}{lrr}
\hline Acceptance Type & & Visual Environment \\
\cline { 2 - 3 } Confirmation & Congruent & Incongruent \\
\cline { 2 - 3 } & Successful & Unrecognized \\
Communication & Miscommunication \\
Negated & Unrecognized & Recognized \\
& Success & Miscommunication \\
\hline
\end{tabular}

LIWC Categories were selected to determine the types of linguistic categories that contribute to the varying outcomes. Given the nature of the predetermined LIWC categories, we do not venture to argue that this provides a thorough linguistic analysis of miscommunication. The main objective was to use these general categories as a firstpass attempt to see what linguistic patterns emerge as
miscommunication unfolds. These included the words per statement, assent, negation (i.e., different than "Negated" described above), cognitive mechanisms, personal pronouns, spatial, and perceptual linguistic forms. The LIWC categories are structured to distinguish at most general predicate classes (e.g., cognitive mechanisms) as well as linguistic particles whose function conveys degrees of interlocutor agreement (e.g., assent, negation).

Within the predetermined LIWC categories, we also created novel subcategories for assent and negation that included specific words (see Table 2 for the categories evaluated and examples within a category). Assent and negation were subdivided because it seemed at the time of transcription that varying forms of these categories could interact in interesting ways depending on the type of success.

Table 2: LIWC and linguistic category examples.
\begin{tabular}{lcr}
\hline Category & Subcategories & Example \\
\hline Negation & Strong Negation & No, nope \\
& Weak Negation & Don't, didn't \\
\cline { 2 - 3 } Personal Pronoun & First Person & I, We \\
& Second Person & You \\
\cline { 2 - 3 } Cognitive & Insight & Think, know \\
Mechanism & Certainty & Always, never \\
\cline { 2 - 3 } Physical & Perceptual & See, hear, feel \\
& Spatial & Top, bottom \\
\cline { 2 - 3 } Assent & Strong Yes & Yes, okay \\
& Weak Yes & Mhm, uh huh \\
\hline
\end{tabular}

\section*{Results}

We first established that, as expected, assenting and negating words were associated with visual congruency. We used a mixed logit model (Jaeger, 2008) to evaluate the proportion of visual incongruency, as predicted by assenting and negating words (LIWC categories), with trial set as a covariate and dyad as a random effect. The results from this analysis revealed that there were significantly fewer assenting words ( \(b=-.177, z=-4.973, p<.001\) ), but more negating words \((b=.392, z=5.210, p<.001)\) when visual incongruence occurred (see Figure 4 for means and standard errors for assent and negation for the visual congruency categories). Nonetheless, it is striking how often assenting words are used when the objects are incongruent.


Figure 4. Mean and standard errors for negation and assent categories for the visual congruency measure.

In order to further explore the relationship between language and outcome, we examined the four different outcomes [Successful Communication (SC), Unrecognized Miscommunication (UM), Unrecognized Success (US) and Recognized Miscommunication (RM)] as predicted by words per statement, assent (strong yes and weak yes), negation (strong and weak negation), and LIWC category measures (personal pronouns, cognitive mechanisms, perceptual, and spatial categories; Croissant, 2012). The results from this model suggest that the measures successfully predict the different outcomes \(\left(x^{2}=1105.9, p<\right.\) . 001 ; see Figure 5 for the proportion of occurrence within each category). Additionally, the evaluation of the significant odds ratios below represent the comparison of each linguistic category relative to the SC trials and are provided with regards to each of the types of communicative success (see Table 3 for results).


Figure 5. Mean and standard errors of the proportion of occurrence within each of the types of communicative outcomes.

Table 3: Significant predictors for the four outcomes, with SC as the reference category: \({ }^{*} p<.05 ;{ }^{* *} p<.01,{ }^{* * *} p<.001\).
\begin{tabular}{llrr}
\hline & Linguistic Category & Odds Ratio & t-value \\
\hline US & WPS & 1.053 & \(4.083^{* * *}\) \\
& Weak Yes & 1.589 & \(2.528^{* *}\) \\
& Strong Negation & 129.907 & \(13.621^{* * *}\) \\
& Weak Negation & 17.439 & \(10.9857^{* * *}\) \\
& We & 1.474 & \(2.361^{* *}\) \\
& Insight & 4.468 & \(7.059^{* * *}\) \\
UR & Perceptual & 1.229 & \(3.268^{* *}\) \\
\cline { 2 - 4 } & WPS & 0.983 & \(-1.982^{* *}\) \\
& Strong Yes & 0.727 & \(-3.452^{* * *}\) \\
& Strong Negation & 6.463 & \(5.159^{* * *}\) \\
& I & 1.767 & \(5.943^{* * *}\) \\
& You & 1.207 & \(2.681^{* *}\) \\
& Spatial & 1.071 & \(2.708^{* *}\) \\
\cline { 2 - 4 } & WPS & 1.051 & \(3.441^{* * *}\) \\
& Strong Negation & 207.009 & \(14.655^{* * *}\) \\
& Weak Negation & 21.919 & \(11.279^{* * *}\) \\
& You & 1.333 & \(2.317^{*}\) \\
& Insight & 3.699 & \(5.447^{* * *}\) \\
& Certainty & 1.931 & \(2.077^{*}\) \\
\hline
\end{tabular}

\section*{Conclusions}

Overall, our results suggest that interlocutors use language in interesting ways when they are having problems with communication. The results for words per statement suggest that using more words can be both helpful and harmful. For example, during US (Unrecognized Success) and RM (Recognized Miscommunication), interlocutors use more words. This might suggest that explaining things too extensively may prevent the listener from encoding all of the information presented due to the limits of processing load. This overload may then result in loss of essential information leading to a communication breakdown. However, fewer words per statement were also associated with UM (Unrecognized Miscommunication) relative to SC (Successful Communication). This might suggest that not providing enough detailed information may furnish the listener with insufficient information to reject the speaker's statement when necessary.

Interlocutors' use of personal pronouns (i.e., \(I\), you, and we) may be indicative of a unique role of perspective taking in creating and repairing unsuccessful communication, as seen in US, UM, and RM. Specifically, there were more instances of saying \(I\) and you, during UM (and you for RM), suggesting that the talkers may be attempting to reconcile differing perspectives. Additionally, the increased occurrence of we during US suggests that interlocutors attempt to find a shared perspective when a "minor mishap" occurs. These findings are interesting given a common interpretation of speakers as primarily egocentric (e.g., Keysar, 2007). While there was no difference in the use of personal pronouns during Successful Communication (SC), it would seem as if dyads tended to respond somewhat egocentrically during instances of less-than-successful communication. However, miscommunicating may have allowed them to (1) remedy a communication breakdown by highlighting their individual and shared perspectives and (2) access more information. Thus, it appears that a dyad may reference their partner's point of view to help re-couple their perspectives.

Additionally, use of spatial terminology increased during Unrecognized Miscommunication (UM). Anecdotally, we noticed that participants' interpretation of each other's use of spatial orientation was often problematic. For example, one participant's use of the word "top" was often not the same as her partner's, especially when the animal pieces were not similarly oriented in space. This type of mistake was not quickly realized, and it was not until a "major catastrophe" happened (i.e., they could not continue with the build until the problem was resolved) that they were able to reconcile each other's meanings of spatial terms. Although intuitively simple, the uses of spatial terms appear to be highly perspective-dependent and can result in communication problems if interlocutors' perspectives are not aligned.

Participants appear to use assent and negation differently during various outcomes. The use of strong and weak confirmation words (such as yes and mhm) may carry
different meaning depending on the context. One explanation of how interlocutors may use the varying forms of confirming and negating words may be that these words help the talker keep track of what they are doing and how they understand it while simultaneously communicating their state of mind to their partner (e.g., saying yes and no to themselves while trying to interpret an instruction). This may be especially important when the talker is confused about how to describe something. In these cases, confirming and negating words may cue the listener to help the talker find the best way to describe something. Therefore, an indirect expression of confirmation and negation may be a cue to the mental state of the speaker.

Alternatively, a confirmation may sometimes be a social nicety. Anecdotally, participants sometimes use a weak form of yes while clearly ignoring their partner (e.g., doing something completely different or unrelated to the current instruction). In these cases, the confirmation may be a socially acceptable filler word used to mask his or her inattention.

Finally, insight words (e.g., think, know) are more prevalent during both Unrecognized Success (US) and Recognized Miscommunication (RM). The prevalence of these words in the US outcomes is particularly interesting. This may indicate some degree of uncertainty. Additionally, interlocutors were more likely to use certainty words during Recognized Miscommunication (RM).

Some of our results are clearly expected given both common sense and previous observations (e.g. see Senft, 2002 for similar results for a tinker-toy task with Trobriand Islanders). For example, interlocutors confirm to ground and they negate to indicate confusion. Nonetheless, there are clearly subtle differences in the language used when participants are grounding successfully (SC) and incorrectly (UM). RM results in an increase in spatial language that reflects a negotiation about differences in perspective. This indicates that interlocutors recognize the importance of shared perspective to resolve confusion when communication is unsuccessful. Speakers may say less and appear to disregard listeners' perspective in an attempt to balance egocentrism and audience needs; providing less detail allows speakers to sample the space of the interaction cheaply and easily, while listeners' requests for additional information continually refine speakers' understanding of listeners' needs.

The main outcome of this preliminary analysis is that different communicative outcomes are associated with subtle differences in language use. This provides insight into how language reflects and influences how miscommunication is recognized and resolved. In order to establish this, however, we will need to investigate whether the language used in miscommunication predicts more global measures of success in the task. For example, RM might result in strategies that will improve performance because it forces interlocutors to negotiate about and resolve alignment of their perspectives. In contrast, failures to recognize the correct state (US) might result either in
interlocutors discounting each other's confirmations or alternatively becoming more sensitive to subtle cues that their interlocutor is uncertain.

While this initial pass of the \(\mathrm{b} \cdot \mathrm{loc}^{\bullet} \mathrm{o}^{\mathrm{TM}}\) corpus (currently lacking measures of reliability for our coding procedure) should be interpreted mindfully, our findings do provide interesting preliminary insights into how lexical choice influences communicative success. Additionally, the categories investigated with LIWC may seem somewhat arbitrary, but the category labels selected were standard and validated LIWC labels (e.g., cognitive mechanisms). In the current form, these categories are not meant to map onto any specific linguistic forms or stages of language processing. Nevertheless, we view the patterns that emerged through these categories as a springboard for more thorough analyses. For example, within the categories of negation and assent, we can next look at specific forms of confirmation to study how they emerge as a function of certainty. Further analyses of the corpus should provide data that will help evaluate these hypotheses.

\section*{Acknowledgments}

Special thanks go to our undergraduate research assistants at University of Rochester (University of Rochester, Chelsea Marsh, Eric Bigelow, Derek Murphy, Melanie Graber, and Anthony Germani) and at University of California, Merced (UC, Merced, Chelsea Coe and J.P. Gonzales). Thanks also go to Rick Dale (University of California, Merced) and Rachel Wu (University of Rochester) for their valuable help, comments and feedback. Preparation of this manuscript was supported by grants from the National Institute of Health (R01 HD027206) to Michael Tanenhaus.

\section*{References}

Aist, G., Campana, E., Allen, J., Swift, M., \& Tanenhaus, M. (2012). Fruit carts: A domain and corpus for research in dialogue systems and psycholinguistics. Computational Linguistics, 38(3), 469-478.
Bard, E., Anderson, A., Chen, Y., Nicholson, H., Harvard, C. \& Dazel-Job, S. (2007). Let's you do that: Sharing the cognitive burdens of dialogue. Journal of Memory and Language, 57(4), 616-641.
Bock, K. (1986). Meaning, sound, and syntax: Lexical priming in sentence production. Journal of Experimental Psychology: Learning, Memory and Cognition, 12, 575586.

Brown, G., Anderson, A., Yule, G., \& Shillcock, R. (1983). Teaching talk. Cambridge, U.K.: Cambridge University Press.
Brown-Schmidt, S., Gunlogson, C. \& Tanenhaus, M. (2008). Addressees distinguish shared from private information when interpreting questions during interactive conversation. Cognition, 107, 1122-1134.
Brown-Schmidt, S. \& Tanenhaus, M. (2008). Real time investigation of referential domains in unscripted conversation: A targeted language game approach. Cognitive Science, 32, 643-684.

Clark, H.H. (1992) Arenas of language use. Chicago: Chicago University Press.
Clark, H. H. \& Carlson, T. (1982). Hearers of speech acts. Language, 58, 332-373.
Clark, H. H. \& Brennan, S. (1991). Grounding in communication. In L. B. Resnick, J. Levine \& S.D Behread (Eds). Perspectives on Socially Shared Cognition. (pp. 222-233). Washington, DC: American Psychological Association.
Clark, H. H. \& Marshal, C. (1981). Definite reference and mutual knowledge. In A. Joshi, B. Webber \& I. Sag (Eds.), Elements of Discourse Understanding, (pp. 1063).

Coupland, N., Giles, H. \& Wieman, J. M. (1991). Miscommunication and problem talk. Newbury Park: Sage.
Croissant, Y. (2007). Mlogit: Multinomial logit model. R package version 0.2-3.
Garrod, S. \& Pickering, M. (2004). Why is conversation so easy? Trends in Cognitive Science, 8(1), 8-11.
Goodwin, C. \& Duranti, A. (1992). Rethinking context: Language as an interactive phenomenon. Cambridge, U.K., Cambridge University Press.

Haywood, S., Pickering, G. \& Branigan, H. (2004). Do speakers avoid ambiguities in dialogue? Psychological Science, 16(5), 362-366.
Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. Journal of Memory and Language, 59(4), 434446.

Keysar, B. (2007). Communication and miscommunication: The role of egocentric processes. Intercultural Pragmatics, 4, 71-84.
Levelt, W. (1983). Monitoring and self-repair in speech. Cognition, 14(1), 41-104.
McTear, M. (2008). Handling miscommunication: Why bother? Recent Trends in Discourse and Dialogue, 39, 101-122.
Nakatani, C. \& Traum, D. (1999). Coding discourse structure in dialogue. University of Maryland Institute for Advanced Computer Studies Technical Report, 1, 1-42.
Pennebaker, J.W., Booth, R. J., \& Francis, M. E. (2007). Linguistic Inquiry and Word Count: LIWC [Computer Software]. Austin, TX: LIWC.net.
Pickering, M. \& Garrod, S. (2004). Toward a mechanistic psychology of dialogue. Behavioral and Brain Sciences, 27, 169-226.
Roche, J., Dale, R., Jaeger, T. F. \& Kreuz, R. J. (under revision). Learning to avoid syntactic ambiguity: Don't rush the navigator.
Senft, G. (2002). Frames of spatial reference in Kilivila. Studies in Language, 25(3), 521-555.
Senft, G. (2007). The Nijmegen space games: Studying the interrelationship between language, culture and cognition. In J. Wassman \& K. Stockhaus (Eds.). Person, Space and Memory in the Contemporary Pacific: Experiencing New Worlds (pp. 224-244). New York: Berghan Books.

\title{
Effects of Target Size and Symmetry on the Structure of Variability in Precision Aiming
}

\author{
Veronica Romero (romerovc@mail.uc.edu) \\ Center for Cognition, Action and Perception, Department of Psychology, University of Cincinnati, ML 0376, 4150 Edwards Cl., University of Cincinnati, Cincinnati, OH 45221-0376 USA \\ Charles A. Coey (coeyca@mail.uc.edu) \\ Center for Cognition, Action and Perception, Department of Psychology, University of Cincinnati, ML 0376, 4150 Edwards Cl., University of Cincinnati, Cincinnati, OH 45221-0376 USA \\ Andrew Beach (beachaw@mail.uc.edu) \\ Center for Cognition, Action and Perception, Department of Psychology, University of Cincinnati, ML 0376, 4150 Edwards Cl., University of Cincinnati, Cincinnati, OH 45221-0376 USA \\ Michael J. Richardson (richamo@ucmail.uc.edu) \\ Center for Cognition, Action and Perception, Department of Psychology, University of Cincinnati, ML 0376, 4150 Edwards Cl., University of Cincinnati, Cincinnati, OH 45221-0376 USA
}

\begin{abstract}
The current experiment investigated the effects of target size and symmetry on the dynamics of precision aiming. Participants were asked to sit on a chair and point at the center of four different targets (a small and big square target, and a horizontal and vertical rectangular target). The aiming movements were assessed using linear (root mean square) and non-linear fractal statistics (DFA and MFDFA). We found that participants spontaneously exhibited more movement in target dimensions with less spatial constraint (i.e., larger target dimensions). These larger movements, however, were more deterministic than the movements accompanying the smaller targets, indicating that more variation in aiming does not necessarily mean more random. Finally, even though participants' movements were multifractal, the different manipulations and task constraints had no effect on the width of the multifractal spectrum. These results suggest that human performance emerges from the complex relationship and interactions that exist between the perception and action capabilities of the human body and the physical environment.
\end{abstract}

Keywords: Cognitive science, psychology, action, motor control, complex systems, 1/f noise.

\section*{Introduction}

Accuracy in tasks such as pistol shooting and archery depends on a person's ability to precisely aim at intended targets, which requires meticulous and refined control of the body and its relationship to the environment around it. Scholz, Schöner and Latash (2000) showed that expert shooters arrange the different components in their body into a motor synergy, coupling certain components to each other and therefore minimizing the necessary movements in order to be more precise. Complimentary research efforts have studied how different task constraints or elements of the physical environment affect how people move their bodies in order to aim precisely, such as target size (Ramenzoni et
al., 2011) or distance (Balasubramaniam, Riley \& Turvey, 2000). However, the effect that such environmental factors have on how people organize their bodies to achieve precision aiming has not yet been revealed in full detail.

Psychologists have traditionally evaluated the impact of task constraints on precision aiming (e.g., target size) using linear statistical tools, such as summarizing effects in means and standard deviations. Recently, statistical techniques allowing researchers to examine more complex aspects of such behavior have come to the fore, most notably, techniques that allow researchers to uncover the fractal structure in movement and behavioral variability (Gilden, 2001; Ihlen, 2012; Delignières \& Marmelat, 2013). Fractal or \(1 / f\) scaling refers to patterns in the variability of behavior that are long-term correlated such that deviations early in a recorded behavior are correlated with deviations that occur much later in the behavior. This kind of structure in variability is often referred to as "pink noise", denoting its difference from the highly irregular or random fluctuations of "white noise" and the highly regular or deterministic fluctuations of "brown noise" (see Figure 1). The degree to which a behavioral measurement series exhibits fractal scaling can be summarized by the Hurst exponent. The Hurst exponent \((H)\) for white noise is 0.5 and for brown noise is 1.5 , with pink noise in-between ( \(H \approx 1\) ) (Ihlen, 2012). Pink noise has been associated with signs of healthy functioning (for a review, Van Orden, Kloos \& Wallot, 2009) in different human movement tasks, such as tapping (Kello et al., 2007; Delignières, Torre \& Lemoine, 2008; Torre, Balasubramaniam \& Delignieres, 2010), stimulusresponse tasks (Holden, Choi, Amazeen \& Van Orden, 2010), postural sway (Schmit, Regis \& Riley, 2005; Schmit et. al., 2006), walking (Hausdorff, 2007) and eye-movement behavior (Coey et al., 2012).


It is becoming increasingly apparent, however, that human behavior may in fact exhibit even more complex patterns of fluctuation than those that can be ascertained from standard fractal analysis. In these cases, the patterns cannot be captured by a single \(H\), as the fractal scaling in the behavior might change over time during the course of measurement, or might be different at different scales of variability (Kantelhardt et al., 2002). Such "multifractal" patterns must, therefore, be characterized by their "multifractal spectrum"; a range of \(H\) values that collectively capture the complex structure inherent in a behavioral time series (e.g., Kuznetsov \& Wallot, 2011; Kuznetsov et al., 2012). This spectrum of \(H\) can either be time-dependent or independent. When it is time-dependent it shows a pattern of long-range correlation where sections of rapid fluctuation are interspersed with sections of slow fluctuation and it is associated with intermittent processes (Kuznetsov \& Wallot, 2011). Multifractal spectrums can also be time-independent due to the behavior being sampled having a frequency distribution with a long tail (Kantelhardt et. al., 2002).

These statistical properties are of interest to cognitive scientists primarily because they reveal something more about the underlying causal structure of human performance than do means and standard deviations (Gilden, 2001; Hausdorff, 2007; Kello et. al., 2007; Holden et. al., 2010; Kuznetsov \& Wallot, 2011). For instance, the presence of monofractal or multifractal structure in human performance can provide insight about the degree to which a behavioral process is self-organized or emerges from interactiondominant dynamics (Kuznetsov \& Wallot, 2011; Van Orden et. al., 2009). Traditional linear statistical tools assume behavior to be static and self-contained, while monofractal and multifractal analyses reveal the strong relationship or coupling between people and their environment (Holden et. al., 2010).


Figure 1: (a) sample time series of white noise depicting highly irregular or random fluctuations; brown noise, with highly regular or deterministic variability; and pink noise, located somewhere between random and deterministic fluctuations. (b) Plots obtained from a detrended fluctuations analysis (DFA) where the mean root mean square (RMS) is plotted against the window size both in log coordinates. The slope of the best fit line gives us the Hurst \((H)\) exponent.

It is for these reasons that the variation in human performance is seen as a balance between task constraints and a person's ability or between involuntary and voluntary control (Van Orden et. al., 2009; Kloos \& Van Orden, 2010). The embedded nature of human behavior can also be revealed by changes in monofractal or multifractal structure that result from subtle and sometimes non-obvious changes in environmental context or constraints (Chen et al., 2001; Balasubramaniam et. al., 2000; Ramenzoni et. al., 2011; Holden et. al., 2010). Depending on the nature of the task and the different constraints, the variability in behavior can go from overly random to more deterministic, or from overly deterministic to more random (Van Orden et. al., 2009; Kloos \& Van Orden, 2010). However, the specific direction of change in variability is not yet fully understood and further study is needed.

The current study investigates the effect that subtle changes in the shape and symmetry of targets have on the dynamics of a participant's precision aiming movements. Participants were instructed to complete the same precision aiming task, with the exact same instructions (i.e., point at the center of the target) over repeated trials. On any given trail, however, the shape and symmetry of the target was subtly changed to investigate how small changes in environmental task constraints can spontaneously reorganize the structure and variability of human behavior. In addition to performing a standard linear variability analysis (i.e., examined the RMS of movement), we conducted both a monofractal and multifractal analysis to better understand the effects that different targets had on the aiming movements of participants, and whether their movements became more deterministic or more random as constraints changed.

\section*{Materials and Method}

\section*{Participants}

Ten undergraduate students from the University of Cincinnati participated in the study for partial course credit.

\section*{Task, Materials and Procedures}

Participants in this experiment were asked to point at targets presented on a display screen. There were four different grey colored targets: a big symmetric ( \(6 \mathrm{~cm} \times 6 \mathrm{~cm}\) ), a small symmetric ( \(3 \mathrm{~cm} \times 3 \mathrm{~cm}\) ), a vertical ( \(3 \mathrm{~cm} \times 6 \mathrm{~cm}\) ), and a horizontal ( \(6 \mathrm{~cm} \times 3 \mathrm{~cm}\) ). Asymmetric targets matched the dimensions of the small symmetric in the strictly constrained dimension and the big symmetric in the loosely constrained dimension (see Figure 2b), therefore the visual angle for both the big symmetric and the vertical targets was \(3.12^{\circ}\) and for the small symmetric and horizontal targets \(1.56^{\circ}\). Additionally, the pointer had a visual angle of \(0.93^{\circ}\) When participants arrived they were greeted and then informed that for this experiment, a sensor would be attached to their index finger which would control the location of a small red square ( \(1.8 \mathrm{~cm} \times 1.8 \mathrm{~cm}\) ) presented directly in front of them on a display screen. This sensor was part of a wired Polhemus magnetic motion tracking system (Polhemus Ltd, VT) and tracked and recorded the movements of the participants at 120 Hz . Once the sensor was attached, participants were seated on a chair located 110 cm away from the TV (Figure 2a).

There were a total of 16 trials; these were completed in blocks of four trials, such that each of the four targets was viewed in each block. The target viewed on any given trial in a four trial block was randomized. The participants were informed that their goal for the experiment was to hold the red square they controlled with the motion sensor in the center of the presented target for the 45 second length of each trial. For each trial, the participant was asked to start pointing at the center of the target, and then the trial started with the Polhemus system being calibrated and the recording of their movement. Participants were instructed to keep their left hand in their lap. After the participants were informed of the number of trials they were to complete, they were given about 25 seconds of practice controlling the red square with the large square target presented on the screen. Once the participant felt comfortable with the procedure, all 16 trials were completed with long breaks given if needed between every block of four trials, between each trial the participant was allowed to lower their hand and place it on their lap. Once the experiment was completed, participants were thanked for their time and debriefed.

\section*{Signal Processing and Measures}

To examine the impact the different targets had on the participants' pointing movements, the first 4096 data points of the X (frontal, side-side movement) and Y (sagittal, updown movement) position time series were extracted for analysis. The Z (back-and-forth) dimension of movement
had little to no effect on task performance and was therefore not analyzed. Each dimension was analyzed separately to better understand the effect that the different target constraints posed on each of the degrees of freedom used by the participants.

Movement Variability. The root mean square (RMS) was calculated of both the X and Y position time series to examine the effects of the target manipulations on the stability of a participant's pointing movements.


Figure 2: (a) Experiment set-up and (b) the four different targets used in the experiment

Fractal Analysis. Detrended fluctuations analysis (DFA) was employed to calculate the monofractal dimension of the X and Y positional time series for each trial. Detailed explanations of this method can be found in several articles (Delignières et. al., 2006; Ihlen, 2012; Delignières \& Marmelat, 2013). Essentially, the time series is divided into windows of a particular size and the average variation (i.e., RMS) around a linear trend is calculated within each window. This procedure is then repeated for windows of different sizes. These averaged RMS are then plotted against the associated window size on log-coordinates. The slope of the best-fit line in this log-log plot represents the scaling relation and corresponds to the Hurst Exponent (H) of the time series (see Figure 1). For the current data we employed \(50 \%\) overlapping window sizes from 16 to 1024 points. Additionally, surrogate time series were created for each time series by randomly shuffling the data points and then a DFA analysis was done to determine whether the fractal dimension observed was time-dependent and therefore a characteristic of long-range correlation.

Multifractal Analysis. Multifractal detrended fluctuation analysis (MFDFA) was used to determine the multifractal dimension present in each time series. This method follows the same steps as DFA, but does so with a scaling parameter \((q)\) that allows for a calculation of \(H\) at different scales of variation in the time series. The final outcome of this procedure is the "width" of all the different \(H\) exponents present in the time series. If this width is equal to 0 , then the monofractal dimension is enough to completely describe the behavior. For the current data we employed \(50 \%\) overlapping window sizes from 16 to 1024 points and examined \(q\) 's from -3.0 to +3.0 in .5 steps. The surrogate time series created were also analyzed through MFDFA to determine whether the spectrum observed was due to timedependent fluctuations, or due to the frequency distribution of the behavior being sampled having a long tail.

Statistical Analyses. One way analyses of variance were computed for each measure in order to understand the effect that the different targets had on participants' behavior. If there was a significant difference, Tukey HSD post-hoc tests were performed.


Results

Figure 3: Mean root mean square of movement in the (a) X dimension (side-to-side) and in the (b) Y dimension (updown) depending on target type. The error bars represent standard error of the mean.

\section*{RMS}

Target type had a significant impact on the amount of side-to-side movement, \(F(3,156)=3.96, p=.009, \eta_{\mathrm{p}}^{2}=.07\). Post-hoc tests revealed that there was significantly more movement for the horizontal target ( \(M=.071\) ) compared to the small symmetric target ( \(M=.051, p=.003\) ), and the vertical target ( \(M=.052, p=.006\); see Figure 3a), indicating that participants naturally exhibited more movement in the direction of less constraint.

Type of target also had a significant influence on the root mean square value of movement in the up-down direction, \(F(3,156)=10.43, p<.001, \eta_{\mathrm{p}}^{2}=.167\). Post-hoc tests showed that there was significantly more up-and-down movement for the big symmetrical target ( \(M=.069\) ) compared to the small symmetric ( \(M=.051, p=.03\) ) and the horizontal targets ( \(M=.05, p=.03\) ). There was also significantly more up-down movement for the vertical target ( \(M=.097\) ) compared to the small symmetric ( \(p<.001\) ) and the horizontal targets ( \(p<.001\); see Figure 3b). Again, participants' movements seemed to spontaneously increase in the Y plane when the target was loosely constrained in this dimension as well. Thus, consistent with the result for the X dimension of movement, increases in participant movement variability appear to be a natural and spontaneous effect of the target size and shape.


Figure 4: Mean Hurst, \(H\), of (a) side-to-side movement and (b) up-down movement depending on target type. The error bars represent standard error of the mean.

\section*{Fractal Analysis}

The analysis of the monofractal dimension, \(H\), calculated using DFA, revealed that target type had a significant influence on the fractal structure of the participants' side-toside movement, \(F(3,156)=17.20, p<.001, \eta_{\mathrm{p}}{ }^{2}=.25\). Post hoc tests revealed that side-to-side movement in the big symmetric target was significantly "browner" ( \(H=1.26\) ) than for the small symmetric target ( \(H=1.19 ; p<.001\) ) and vertical target ( \(H=1.21\); \(p=.003\) ). Furthermore, the side-to-side movement of participants was significantly browner when pointing at the horizontal target \((H=1.28)\) than when pointing at the small symmetric target ( \(p<.001\) ) and vertical target ( \(p<.001\); see Figure 4a). This mirrors the results for RMS above, with participants' movement being browner in the targets where the X plane was loosely constrained and suggests that their movements became more deterministic when more freedom was allowed in the task. In other words, the participants moved more, but this increase in movement was also an increase in the level of determinism.

The fractal structure of the up-down movement of participants was significantly affected by the type of target they had to point at, \(F(3,156)=38.47, p<.001, \eta_{\mathrm{p}}^{2}=.43\). Post hoc tests revealed that participants' up-down movements were significantly browner when pointing at the vertical target ( \(H=1.31\) ) than the big symmetric target ( \(H=\) 1.25; \(p=.01\) ), the horizontal target ( \(H=1.15 ; p<.001\) ) or the small symmetric target ( \(H=1.16 ; p<.001\); see Figure 4b). Consistent with the fractal analysis of \(X\) and the RMS for Y above, participants' movements were pinker in structure when the target was more constrained in the intended plane. This suggests that participants' movements became less correlated in time with increases in task constraint.

The DFA analysis of the surrogate time series resulted in white noise ( \(H \approx .5\) ) for every trial wiping out any correlation present in the collected data. This indicates that the above monofractal analysis performed on the recorded data is time-dependent and not an analysis artifact.

\section*{Multifractal Analysis}

Participants' movements were found to be multifractal, with a one-sample \(t\)-tests demonstrating that the multifactal spectrum width for each movement dimension and for each target types were significantly different from zero (all \(t(39)\) \(>24.16, p<.001\) ). Although there was no effect of target type on multi-fractal width for participants' side-to-side movement, \((F(3,156)=.76, p=.518)\), an effect of target type on multi-fractal width was found for the participants' up-down movement \(\left(F(3,156)=3.07, p=.03, \quad \eta_{\mathrm{p}}^{2}=.06\right)\). Pot-hoc analyses, however, revealed that the only significant difference was that participants' movement while pointing at the small symmetric target had a significantly wider Hurst spectrum ( \(H\) width \(=.57\) ) than while pointing at the vertical target ( \(H\) width \(=.50, p=.04\) ). Therefore, even though precision aiming shows multifractal spectrum
characteristics, this measure does not capture the effects that size and symmetry of target have on the behavior as well as RMS and monofractal analyses do.

The surrogate time series also had a multifractal spectrum ( \(H\) width \(\approx .42\) ) which suggests that the multifractal spectrum present in the data is not time-dependent, but rather is the result of the behavior having a long-tailed frequency distribution (Kantelhardt et. al., 2002).

\section*{Discussion}

Our data indicate that changing some task constraints, while leaving the rest of the experiment the same, does change human performance behavior. In general, even though the participants in the current study were always told to point at the center of the target, they moved around more if more target space was available. In other words, looser constraints brought about more spontaneous movement variability. Additionally, this increase in movement variability brought about a more deterministic behavior, where looser constraints in a certain dimension resulted in a structure of variability closer to brown noise. This deterministic behavior was also shown to be the result of time-dependent long-range correlations. Finally, a multifractal analysis showed that the behavior was even more complex and that it could be represented by a multifractal spectrum, however, this multifractal spectrum did not characterize the influence of the different task constraints. Furthermore, the multifractal spectrum did not show a time-dependent pattern, it was instead due to the frequency distribution of the fluctuations of participants' movements.

These results support the idea that participants' behavior in the precision aiming task exhibit the characteristics of a strong relationship or coupling between the person and the environment, such that subtle changes in constraints bring about changes in the underlying dynamics of the movement. Additionally, the results are similar to those obtained by Balasubramaniam and colleagues (2000) where participants increased their overall movement in dimensions where more freedom was present, but that in turn this spontaneous increase in movement was more deterministic in nature. However, studies in different tasks, such as tapping or walking to a metronome have found the opposite results in which stricter control results in more random variability (for a review see Van Orden et. al., 2009 and Kloos \& Van Orden, 2010). One idea that has been supported by the data available so far is that people's movement variability is a result of the balancing between involuntary control (overly random) and voluntary control (overly deterministic) that arises during a specific task in a specific context (Van Orden et. al., 2009; Kloos \& Van Orden, 2010). If this is indeed the case, the results of the present study would suggest that participants impose further voluntary control to counteract the increase in spontaneous movement, so that they are able to successfully stay inside the target boundary. However, further research is needed to better understand the mutual influence that task constraints and participants’ ability play on the production of a certain behavior.

In general, the results of the present study bring into question the standard belief in cognitive science of behavior being the result of participants' voluntary and cognitive control alone. Instead, it points to a more embodied or interaction-dominant approach in which participants and their physical environment interact and mutually influence each other. It is therefore objectionable to try to study behaviors by only looking at the participant and ignoring the environment. Instead the focus of research should be the coupling or relationship between the person and its physical environment.

\section*{References}

Balasubramaniam, R., Riley, M. A. \& Turvey, M. T. (2000). Specificity of postural sway to the demands of a precision task. Gait and Posture, 11, 12-24.
Chen, Y., Ding, M. \& Kelso, J. A. S. (1997). Long memory processes (1/f type) in human coordination. Physical Review Letters, 79(22), 4501-4504.
Chen, Y., Ding, M. \& Kelso, J. A. S. (2001). Origins of timing errors in human sensorimotor coordination. Journal of Motor Behavior, 33(1), 3-8.
Coey, C. A., Wallot, S., Richardson, M. J. \& Van Orden, G. (2012). On the structure of measurement noise in eyetracking. Journal of Eye Movement Research, 5(4), 1-10.
Delignières, D., Ramdani, S., Lemoine, L., Torre, K., Fortes, M. \& Ninot, G. (2006). Fractal analyses for 'short' time series: A re-assessment of classical methods. Journal of Mathematical Psychology, 50, 525-544.
Delignières, D. \& Marmelat, V. (2013). Theoretical and methodological issues in serial correlation analysis. Advances in Experimental Medicine and Biology, 782, 127-148.
Delignières, D., Torre, K. \& Lemoine, L. (2008). Fractal models for event-based and dynamical timers. Acta Psychologica, 127, 382-397.
Duarte, M. \& Zatsiorsky, V. M. (2000). On the fractal properties of natural human standing. Neuroscience Letters, 283, 173-176.
Gilden, D. L. (2001). Cognitive emissions of \(1 / \mathrm{f}\) noise. Psychological Review, 108(1), 33-56.
Hausdorff, J. M. (2007). Gait dynamics. Fractals and falls: Finding meaning in the stride-to-stride fluctuations of human walking. Human Movement Science, 26, 555-589.
Holden, J. G., Choi, I., Amazeen, P. G. \& Van Orden, G. (2010). Fractal 1/f dynamics suggest entanglement of measurement and human performance. Journal of Experimental Psychology: Human perception and performance, doi: 10.1037/a0020991.
Ihlen, E. A. F (2012). Introduction to multifractal detrended fluctuation analysis in Matlab. Frontiers in Psychology, 3(141), 1-18.
Kantelhardt, J. W., Zschiegner, S. A., Koscielny-Bunde, E., Bunde, A., Havlin, A. \& Stanley, H. E. (2002). Multifractal detrended fluctuation analysis of nonstationary time series. Physica A, 316, 87-114.

Kello, C. T, Beltz, B. C., Holden, J. G. \& Van Orden, G. C. (2007). The emergent coordination of cognitive function. Journal of Experimental Psychology. General, 136(4), 551-568.
Kloos, H. \& Van Orden, G. (2010). Voluntary behavior in cognitive motor tasks. Mind \& Matter, 8(1), 19-43.
Kuznetsov, N., Bonnette, S., Gao, J. \& Riley, M. A. (2012). Adaptive fractal analysis reveals limits to fractal scaling in center of pressure trajectories. Annals of Biomedical Engineering, doi: 10.1007/s10439-012-0646-9.
Kuznetsov, N. A. \& Wallot, S. (2011). Effects of accuracy feedback on fractal characteristics of time estimation. Frontiers in Integrative Neuroscience, 5(62), 1-12.
Ramenzoni, V. C., Davis, T. J., Riley, M. A., Shockley, K. \& Baker, A. A. (2011). Joint action in a cooperative precision task: nested processes of intrapersonal and interpersonal coordination. Experimental Brain Research, 211, 447-457.
Schmit, J. M., Regis, D. \& Riley, M. A. (2005). Dynamic patterns of postural sway in ballet dancers an track athletes. Experimental Brain Research, 163, 370-378.
Schmit, J. M., Riley, M. A., Dalvi, A., Sahay, A., Shear, P. K., Shockley, K. \& Pun, R. Y. K. (2006). Deterministic center of pressure patterns characterize postural instability in Parkinson's disease. Exprimental Brain Research, 168(3), 357-367.
Scholz, J. P., Schöner G. \& Latash, M. L (2000). Identifying the control structure of multijoint coordination during pistol shooting. Experimental Brain Research, 135, 382404.

Torre, K., Balasubramaniam, R. \& Delignières, D. (2010). Oscillating in synchrony with a metronome: Serial dependence, limit cycle dynamics, and modeling. Motor Control, 14, 323-343.
Van Orden. G. (2010). Voluntary performance. Medicina (Kaunas), 46(9), 581-594.
Van Orden, G. C., Holden, J. G. \& Turvey, M. T. (2003). Self-organization of cognitive performance. Journal of Experimental Psychology. General, 132(3), 331-350.
Van Orden, G. C., Kloos, H. \& Wallot, S. (2009). Living in the pink: Intentionality, wellbeing, and complexity. In Philosophy of Complex Systems: Handbook of the Philosophy of Science (Hooker, C., ed.), Elsevier.

\title{
The influence of perceptual and structural salience
}

\title{
Florian Röser (florian.roeser@psychol.uni-giessen.de) \\ Antje Krumnack (antje.krumnack@psychol.uni-giessen.de) \\ Kai Hamburger (kai.hamburger@ psychol.uni-giessen.de)
}

\author{
Justus Liebig University Giessen, Department of Psychology, Experimental Psychology and Cognitive Science \\ Otto-Behaghel-Strasse 10 F \\ 35394 Giessen, Germany
}

\begin{abstract}
Spatial cognition research has recently made much progress in understanding the cognitive representations and processes underlying human wayfinding. Many theoretical assumptions about the concept of landmark salience have been established. In this context it is important to define perceptual (or visual) and structural landmark salience. Structural salience is defined as the position of a landmark at an intersection. Perceptual salience is defined as the visual characteristic of a landmark. It must "stand out" from its surrounding to be perceptually salient. We investigated the influence of perceptual salience and the combination of perceptual and structural salience in landmark selection. We show for a spatial arrangement of four objects that the different object is preferred almost always. If the same spatial arrangement is interpreted as an intersection with a directional information, the participants' preference is influenced by structural as well as perceptual salience. Findings are discussed within the context of landmark salience.
\end{abstract}

Keywords: landmark; perceptual salience; structural salience

\section*{Introduction}

Human wayfinding is a particularly active field of spatial cognition research. People often have to navigate through new or familiar environments and that they, of course, do not want to get lost. Another reason is that wayfinding research is important for many basic and applied research fields, for instance, the study of spatial long-term memory and the development of user-friendly navigation systems.

One of the central concepts in spatial cognition research is the landmark and the question how it can be defined. Consequently, several definitions and theories about the nature of landmarks and their characteristics exist. The most common assumption is that the potential landmark must have a high contrast to its immediate or wider surrounding (e.g., Presson \& Montello, 1988; Janzen \& van Turennout, 2004; Caduff \& Timpf, 2008). Anything can serve as a landmark; natural, artificial, or man-made objects along a route that help us to find the way. Landmarks are helpful in wayfinding because they "stand out" of the environment, can serve as anchors (Couclelis, Golledge, \& Tobler, 1987), are better remembered if a change of direction is required (Lee, Tappe \& Klippel, 2002; Lee, Klippel \& Tappe, 2003), and increase the quality of a route description (Denis, Pazzaglia, Cornoldi, \& Bertolo, 1999).

The present paper is concerned with the salience of landmarks. The term "salience" is mostly referred to perceptual psychology (e.g., Treisman \& Gelade, 1980) and means that a salient object needs to stand out compared to other objects (e.g., different color or orientation). We here distinguish two kinds of landmark salience: structural salience and perceptual salience.

\section*{Structural salience}

Structural aspects of landmarks refer to the contexts of landmarks in navigational tasks and may be divided into different aspects or gradations. It is generally accepted that landmarks must have a "prominent location in the environment" (Sorrows \& Hirtle, 1999; p. 46). Furthermore, they can be separated into global and local landmarks (Steck \& Mallot, 2000). Local landmarks are situated directly along the path and intersections (Klippel \& Winter, 2005). Those at an intersection may again be divided, based on the route in the navigational task, into landmarks at a decision point or non-decision point, where a direction change is necessary/possible (Michon \& Denis, 2001).

Here we concentrate on landmarks directly located at an intersection were a decision is required (Lee et al., 2002; Peters, Wu, \& Winter, 2010). In this context we define structural salience as a preference of a wayfinder for a landmark to be located at a specific position at an intersection. Strictly speaking structural salience is thereby not a property of a landmark itself but of its position at an intersection. Therefore we have to address the question how an intersection can be defined. A typical or even prototypical intersection is a cross intersection. At such intersections four possible positions for landmarks are available (figure 1).

The true (physical) position of a landmark, on the right or on the left of the observer, is less important than the position in dependence to the direction of the turn to be made at the intersection. So the four positions can be defined as the positions before and behind the intersection and in direction or opposite to the direction of turn (see also Hamburger, Dienelt, Strickrodt, \& Röser, 2013) and will be abbreviated as "turn based" in the following. The preferred positions for a landmark at such a prototypical intersection from an allocentric perspective are in direction of the turn with the main focus before the intersection (figure 2). These results
serve as our reference for the influence of structural salience at the positions of a four-way-intersection.


Figure 1: Schematic visualization of a prototypical intersection with two orthogonal streets.


Figure 2: Results (landmark position preference over all intersections) for the structural salience (optimal landmark position) from two previous experiments (only position preference without landmarks: Röser, Hamburger, Krumnack, \& Knauff, 2012 [left]; shape-color-combinations as landmarks: Röser, Krumnack, Hamburger, \& Knauff, 2012 [right]). Here the turn based positions are depicted.

\section*{Perceptual salience}

Sorrows and Hirtle (1999) defined visual (perceptual) salience as the inherent visual object characteristics and stated that "[...] these may include the features of contrast with surroundings [...]." (p. 45). Caduff and Timpf (2008) provided a different definition and they understood perceptual salience as a bottom-up salience with the components location-based and attention-based attention, and the scene context. Other authors demonstrated that, for example, different colors, orientations, and shapes deploy attention (Raubal \& Winter, 2002; Wolfe \& Horowitz, 2004) what implies that these features could have a high visual salience. Also Treisman and Gelade (1980) showed that objects that stand out from their environment quickly reach the focus of attention.

Based on these concepts our definition of perceptual salience is a contrast-based approach where the observerbased contrast to the surrounding of the object is central. Strictly speaking an object is perceptually salient if it is an outlier, meaning that it is sufficiently different in comparison to the other objects available.

In our prior experiments (Röser, Hamburger, et al., 2012; Röser, Krumnack, et al., 2012) all landmark material was created to have the same perceptual salience and therefore the perceptual salience of a landmark should not have had an effect on landmark choices. In such a setting the perceptual contrast between the objects should be equal, no object should stand out.

\section*{Cognitive salience}

The focus of this study is not on cognitive (also defined as semantic) salience but it should not be neglected. It can be defined as the meaning or prototypically of an object (Sorrows \& Hirtle, 1999). Again the contrast is important or the degree of recognizability and the idiosyncratic relevance (Caduff \& Timpf, 2008). We assume that these factors do not play any role if the material is simple enough and is related to a single perceptual/cognitive category, like colors or simple geometrical shapes (by the same argument as for the perceptual salience, see above).

\section*{Experiments}

The main aim of the present paper is to explore which object in an environment people prefer to use as a landmark in wayfinding. In particular we want to answer the following questions: How important is a high perceptual contrast for the choice of a landmark at a decision point? And, what is the influence of the position of an object on landmark choices in a setting where one object clearly stands out? We investigated two independent factors: To vary the perceptual salience of potential landmarks we used objects in different colors, shapes, and different orientations. To vary the structural salience of potential landmarks the objects were located at different positions at an intersection. In a pilot study we examined how visual aspects of objects influence their perceptual salience in an arrangement similar to figure 1 but without any navigational context. In the main experiment we combine perceptual and structural salience by adding a navigational context to the arrangement.

\section*{Pilot Study - Perceptual salience}

We investigated the distribution of perceptual salience of an array of objects. Therefore, we presented groups of different stimuli and asked the participants which of them stands out most in contrast to the other ones (which one is the outlier?). This is based on our definition of perceptual salience as the contrast of an object to its surrounding. The goal was to establish a baseline of perceptual salience to use as a reference for further experiments.

\section*{Methods}

\section*{Participants}

A total of 20 students ( 16 females) with a mean age of 24 years (range: 20-41) participated. All participants provided informed written consent. All had normal or corrected-tonormal visual acuity and color vision. They received course credit or money for participation.

\section*{Material}

For this study we used a basic setting with four objects placed in a square with the same distance between each other (see figure 3). This setting resembles an intersection, but participants were not explicitly made aware of the resemblance and were not given any navigational context.

To vary the perceptual salience of potential landmarks we used objects in different colors, different shapes, and different orientations. The colors were always presented using the same shape, a simple cross (figure 3). In 24 items three identical colors and one outlier color (green and red; blue and yellow; red and yellow) were shown. Each color combination was presented eight times; half of them with three crosses of the first color and one cross of the second color and vice versa. The position of the outlier was counter-balanced over the four positions.

For the different shapes we used the same logic: 24 items with three identical shapes (e.g., a square; always in black, see figure 3) and one outlier shape (e.g., a triangle), again balanced over the four positions. For the different orientations of shapes four identical forms were used (see figure 3). Here the difference lies in the orientation: Either three shapes are orientated vertically and the outlier is rotated 15 deg to the right or the three identically oriented objects are rotated 15 deg to the right and the outlier is orientated vertically. Again, the outliers are shown once at each of the four positions.

Distractors were presented in addition. Twelve identical colors or shapes and twelve different ones served as distractors. For the different orientations twelve items with identical forms in different orientations (+/- 15 deg, +/- 30 deg) and twelve with different shapes in different orientations served as distractors.
In sum this resulted in 144 images of different stimulus material, 72 as experimental material and 72 distractors. All images were presented in succession in a random order on a custom computer screen (22'). Superlab 4.0 (Cedrus Corporation 1991-2006) was used for running the study and for data recording.


Figure 3: Example for the color material (left), different shapes (center), and shapes with different orientations (right).

\section*{Procedure}

Participants received instructions on the computer screen. It was explained that four objects will be shown at a time and in a fixed arrangement. Participants were instructed to indicate the outlier which stands out most to them. To select any object they should press the according response key on the keyboard.

\section*{Results}

\section*{Distributions}

The analysis of the distribution, preference of the objects over all variations, showed no significant variation from an equal distribution \(\left(\chi^{2}(3)=0.281, p=.963\right.\); each is preferred in
\(25 \%\) of the cases; we here used not a per 100 system, due to the fact, that the chi-square test is highly sensitive to the sample size, but rather a per 20 system, based on the sample size [ \(N=20]\); that means that each participant is weighted with one and the individual distribution is correspondingly adjusted).

\section*{Outliers}

The follow-up data analysis is based on the preference of the outlier compared to the other three objects (equal). For this we merged the preferences of all participants over all images with three equal and one different stimulus and for all positions of the outliers (table 1).

Table 1: Results of the statistical analyses. Chance level ( \(25 \%\) ) would mean that every position is preferred equally often, or one position is chosen all the time.
\begin{tabular}{llll}
\hline & \begin{tabular}{l} 
Preference of \\
single one
\end{tabular} & \begin{tabular}{l}
t -test ( \(d f=19)\), against \\
chance level
\end{tabular} \\
\hline Over all & \(86 \%\) & \(t=14.551\) & \(p<.001\) \\
\hline Colors & \(92.5 \%\) & \(t=14.312\) & \(p<.001\) \\
\hline Different shapes & \(91 \%\) & \(t=13.578\) & \(p<.001\) \\
\hline \begin{tabular}{l} 
Different \\
orientations*
\end{tabular} & \(75 \%\) & \(t=11.446\) & \(p<.001\) \\
\hline \begin{tabular}{l}
\(*\) Sum of outlier preferences differ significantly for colors \\
\((t(19)=4.186, p<.001)\) and shapes \((t(19)=3.964, p<.001)\)
\end{tabular} \\
\hline
\end{tabular}

\section*{Discussion}

With the present study we investigated whether participants prefer to indicate objects that differ perceptually from the surrounding in the display similar to an intersection. As our results indicate, participants prefer to indicate the object with different perceptual properties, the outliers. Based on our definition of perceptual salience as a contrast to the surrounding, the results may be considered as a measurement of perceptual salience. Colors and shapes had the highest perceptual salience in this study in contrast to the different orientations. This could be due to the fact that the contrast to the surrounding is for the different orientations not as high as for the colors and shapes.
In the main experiment we examine the effect of perceptual differences in a wayfinding context.

\section*{Main experiment - Perceptual and structural salience}

With the main experiment we aim to examine how perceptual and structural saliences affect each other.
Based on the pilot study we now used the objects with the highest perceptual salience: colors in combination with different positions at an intersection in a navigational context. In this way we intend to investigate whether perceptual differences of landmarks influence the position preference or structural salience as determined by earlier experiments (see Figure 2).

\section*{Methods}

\section*{Participants}

A total of 20 students ( 14 females) with a mean age of 22.5 years (range: 18-31) participated. All participants provided informed written consent. They had normal or corrected-tonormal visual acuity and normal color vision. They received course credit or money for participation.

\section*{Material}

The navigational task used for this experiment was based on the virtual environment Squareland (Hamburger \& Knauff, 2011). We used a \(5 \times 6\) square setting from an allocentric (bird-eye) perspective (figure 4).


Figure 4: Maze including the path from the start to the sixth intersection. At the intersection four landmarks are depicted. On the right the answer instruction (which key to press) is given.

The route through the maze consisted of sixteen intersections: with equal numbers of left and right turns. This represented a route length people can imagine and remember in a virtual setting (e.g., Hamburger, Röser, Bukow, \& Knauff, 2011). The arrangement of the four objects was equivalent to the pilot study (equal distance between them and in the corners of the four squares).

As before, at each intersection (array) four colored objects were used including one outlier. In order to ensure that each combination of colors was presented only once, we needed a total of sixteen color combinations with sufficiently different hues. So we used the color circle and chose each color 22.5 deg away from the next one. The respective complementary color was used as the outlier. Color combinations were distributed randomly over the path. The positions of the outlier objects were systematically varied based on the turn direction. A second version of the maze with inversed colors (identical and outlier) was created. Additionally, for both versions the direction of turn was switched for each intersection, resulting in overall four different mazes. Each participant was randomly assigned to one of them. The participants performed the experiment on a custom computer screen (22'). Superlab 4.0 (Cedrus Corporation 1991-2006) was used for running the experiment and for data recording.

\section*{Procedure}

The instruction explained that the participants would see a path through a maze where at each intersection four different objects are presented. They should imagine that they have to give a route description based on the information they see. They were instructed to decide/ indicate at each intersection which object they are going to use for the route description. To select one object they had to press the corresponding key on the keyboard. The response keys were presented next to each slide at an exemplary intersection (figure 4). Subsequently, the instruction was repeated and supplemented with a pictorial explanation. After this example the experimental phase started with the path being presented from the start to the first intersection, including the route direction for this intersection and four colored crosses placed at the four positions. After each decision participants saw the next intersection and again the path from the start to this intersection and direction of turn and the four colors became visible and so on.

\section*{Results}

Outliers
To find whether outliers were selected more often as landmarks compared to the other objects, participants' responses were analyzed. The outliers were selected with a mean of \(66 \%\) and therefore significantly more often, compared to the remaining objects \((t(19)=2.281, p=.034)\). This result is also statistically different from chance level (25\%; \(t(19)=5.589, p<.001\) ).

\section*{Distributions}

In figure 5 (center) the turn based positions are presented. The distribution over all positions and intersections in the maze revealed a marginally significant variation from an equal (each is preferred in \(25 \%\) of the cases) distribution \(\left(\chi^{2}(3)=7.016, p=.071\right.\); again we used a per 20 system, based on the sample size [ \(N=20\) ], see above).

For each of the four positions the outlier is chosen in at least \(50 \%\) of the cases if the outlier is located at that position (figure 5 left). If the outlier is not to be found at that position, the position without the outlier is chosen in at least \(2 \%\) of the cases (figure 5 right).

In a last step the distributions for the four variations of outlier positions were analyzed separately. Here, for all four variations a significant difference from chance level was obtained (see figure 5 bottom; position: top, opposite to the direction of the turn: \(\chi^{2}(3)=10.175, p=.017\); position: top, in the direction of the turn: \(\chi^{2}(3)=12.100, p<.001\); position: bottom, opposite to the direction of the turn: \(\chi^{2}(3)=13.575\), \(p<.001\); position: bottom, in the direction of the turn: \(\left.\chi^{2}(3)=56.075, p<.001\right)\). In summary, for the single positions it could be emphasized that for each position the outliers were chosen in a minimum of \(50 \%\) of the cases and furthermore the ideal position, before the intersection in the direction of the turn, is minimally preferred in \(1 / 3\) of the
cases. If the outliers were placed on this ideal position it was almost always chosen.


Figure 5: Results for the different analyses. On the left and the right side (top section) each position could reach a value of \(100 \%\). The four positions in the middle add up to \(100 \%\).

\section*{Discussion}

The results could be interpreted from two different perspectives. On the one hand they revealed a clear preference of the perceptually salient object. The single outliers are preferred in \(66 \%\) of the cases, which is much higher than chance (25\%). This represents the importance of the perceptual salience for landmarks. On the other hand, the preference over all positions and intersections (merged for all single positions) are not distributed equally (see figure 5, middle). There is a preference for the position before the intersection and in the direction of the turn. Thus, the perceptual salience as well as the structural salience influence the preference of the positions. How they interact will be analyzed in the following section.

\section*{General Discussion}

Our previous findings revealed a position preference which we presented as the structural salience. The position before the intersection and in the direction of the turn was preferred. The pilot study revealed an object preference depending on the perceptual salience. There, the outlier colors were preferred in \(92.5 \%\) which is more or less equally distributed over the four positions. Here the objects differ from each other and the positions are unimportant. In the main experiment the combination of the perceptual (object preference) and the structural salience (position preference) influenced the participants' selections. In figure 6 these results are contrasted.

It is very interesting that the results do not reveal the same distribution as in the pilot study (figure 6, left), because it is -from a perceptual point of view- inconsequential not to
prefer the outliers. The preference for the outliers decreased from \(92.5 \%\) in the pilot study to \(66 \%\) in the main experiment. Therefore it seems that navigational context, in particular the turn based position of a landmark, also influenced the participants' preferences. In other words, not only the differentiation between objects plays an important role in landmark selection but also the position of the object.


Figure 6: Object preferences of the pilot study with colors (left); data of the main experiment (center); data of the previous experiments (Röser, Hamburger et al., 2012; Röser, Krumnack et al., 2012) for position preference (right).

In this work we used one fixed factor of perceptual salience, outliers consisting of complementary colors (the contrast to the surrounding could not be higher) to observe its interaction with structural aspects of wayfinding. In further experiments we will address the possibility to vary the perceptual salience and take a look at the effects on the resulting distributions. Two possibilities remain: on the one hand the perceptual salience of a landmark has a simple yes or no (existing or non-existing) character. Then the results should be similar. Or, on the other hand the perceptual salience of a landmark is gradual. Then the results should change continuously. Whatever the case may be, it seems clear that the influence of the perceptual salience could hardly be increased by using other colors, because the contrast used here with the complementary colors, is the strongest color contrast possible. Additionally, the pilot study showed that the contrasting color is preferred in almost all cases. We may therefore conclude that the perceptual salience of the outlier object was as high as possible and any variation probably leads to a lower impact (at least its influence may not be higher). But even if an object is extremely "eye-catching" in a non-navigational setting, once we enter a navigational task it seems that the structural salience provides a strong and almost permanent influence on the choice of landmarks at an intersection. The position of an object at an intersection might therefore be as important for it being chosen as landmark as the contrast to its surrounding.

The comparison of the results of the two experiments clearly shows that the question which object is the most appropriate landmark cannot be reduced to the question which object is the most noticeable. The fact that an object has a high contrast to its surrounding does not guarantee its
choice as a landmark. This supports Sorrows and Hirtle's (1999) concept of "prominence of spatial location" (p. 45) as a factor of visual salience and Caduff and Timpf's (2008) scene context of visual salience and/or contextual salience.

\section*{Obstacles in wayfinding research}

Because of the abstract and artificial setting of our experiments, transfers to other wayfinding research is difficult. Nothegger, Winter, and Raubal (2004) for example used real environments for their wayfinding experiments. But, the experimental control in such experiments is difficult or even impossible (e.g., which information did participants pay attention to?). Particularly the structural aspects of intersections in real environments are determined by a lot of factors (visibility; view direction; occlusion, etc). This is why we chose to limit ourselves to such an abstract setting. Our study serves as a basic research approach, examining the underlying aspects and will serve as a basis for further and more realistic (but controlled) experiments.

\section*{Acknowledgement}

This study was supported by the German Research Foundation (DFG HA5954/1-1). We thank Jelica Nejasmic and Markus Knauff for valuable comments.

\section*{References}

Caduff, D., \& Timpf, S. (2008). On the assessment of landmark salience for human navigation. Cognitive Processing, 9, 249-267.
Couclelis, H., Golledge, R. G., \& Tobler, W. (1987). Exploring the anchorpoint hypothesis of spatial cognition. Journal of Environmental Psychology, 7, 99122.

Denis, M., Pazzaglia, F., Cornoldi, C., \& Bertolo, L. (1999). Spatial discorse and navigation: an analysis of route directions in the City of Venice. Applied Cognitive Psychology, 13, 145-174.
Hamburger, K., Dienelt, L. E., Strickrotd, M., \& Röser, F. (2013). Spatial cognition: The return path. In M. Knauff, M. Pauen, N. Sebanz, \& I. Wachsmuth (Eds.), Proceedings oft he 35th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.
Hamburger, K. \& Knauff, M. (2011). Squareland: A virtual environment for investigating cognitive processes in human wayfinding. PsychNology Journal, 9, 137-163.
Hamburger, K., Röser, F., Bukow, G. C., \& Knauff, M. (2012). Der Weg als Ziel: Virtuelle Umgebungen und räumlicher Wissenserwerb. In: G. C. Bukow, J. Fromme \& B. Jörissen (Hrsg.), Raum, Zeit, Medienbildung: Untersuchungen zu medialen Veränderungen unseres Verhältnisses zu Raum und Zeit (Medienbildung und Gesellschaft). VS Verlag.
Janzen, G., \& van Turennout, M. (2004). Selective neural representation of objects relevant for navigation. Nature Neuroscience, 7, 673-677.

Klippel, A., \& Winter, S. (2005). Structural salience of landmarks for route directions. In A. G. Cohn \& D. M. Mark (Eds.), Spatial Information Theory. International Conference COSIT (pp. 347-362). Berlin: Springer.
Lee, P. U., Tappe H., \& Klippel, A. (2002). Acquisition of landmark knowledge from static and dynamic presentation of route maps. KI Special Issue Spatial Cognition, 2, 32-34.
Lee, P., Klippel, A., \& Tappe, H. (2003). The effect of motion in graphical user interfaces. In: A. Butz, A. Krüger \& P. Oliver (Eds.), Smart Graphics (pp. 12-21). Berlin: Springer.
Michon, P. E., \& Denis, M. (2001). When and why are visual landmarks used in giving directions? In D. R. Montello (Ed.), Spatial Information Theory. Volume 2205 Lecture Notes in Computer Science (pp. 292-305). Berlin: Springer.
Nothegger, C., Winter, S. \& Raubal, M. (2004) Selection of salient features for route directions. Spatial cognition and computation, 2, 113-136.
Peters, D., Wu, Y., \& Winter, S. (2010). Testing Landmark Selection Theories in Virtual Environment. In C. Hölscher et al. (Eds.), Spatial Cognition VII. Lecture Notes in Artificial Intelligence (pp. 54-69). Berlin: Springer.
Presson, C. C., \& Montello, D. R. (1988). Points of reference in spatial cognition: Stalking the elusive landmark. British Journal of Developmental Psychology, 6, 378-381.
Raubal, M., \& Winter, S. (2002). Enriching wayfinding instructions with local landmarks. In M. J. Egenhofer \& D. M. Mark (Eds). Geographic information science (pp 243-259). Berlin: Springer.
Röser, F., Hamburger, K., Krumnack, A., \& Knauff, M. (2012). The structural salience of landmarks: Results from an online study and a virtual environment experiment. Journal of Spatial Science, 57, 37-50.
Röser, F., Krumnack, A., Hamburger, K. \& Knauff, M. (2012). A four factor model of landmark salience -A new approach. In N. Rußwinkerl, U. Drewitz, H. van Rijn (eds). Proceedings of ICCM 2012. \(11^{\text {th }}\) international Conference of Cognitive Modeling (pp. 8287). Berlin: Universitaetsverlag der TU Berlin.

Sorrows, M. E., \& Hirtle, S. C. (1999). The nature of landmarks for real and electronic spaces. In C. Freksa \& D. M. Mark (Eds). Spatial information theory: cognitive and computational foundations of geographic information science, international conference COSIT (pp. 37-50). Stade: Springer.
Steck, S. D., \& Mallot, H. A. (2000). The Role of Global and Local Landmarks in Virtual Environment Navigation. Presence, 9, 69-83.
Treisman, A. M., \& Gelade, G. (1980). A feature-integration theory of attention. Cognitive Psychology, 12, 97-136.
Wolfe, J. M., \& Horowitz, T. D. (2004). What attributes guide deployment of visual attention and how do they do it? Nature Reviews Neuroscience, 5, 495-501.

\title{
Development of Substitution Bias Sensitivity: Are Adolescents Happy Fools?
}

\author{
Sandrine Rossi (sandrine.rossi@unicaen.fr) \\ LaPsyDE (CNRS Unit 3521), University of Caen Basse-Normandie, Caen, France \\ Mathieu Cassotti (mathieu.cassotti@ parisdescartes.fr) \\ LaPsyDE (CNRS Unit 3521), Paris Descartes University, Paris, France \\ Marine Agogué (marine.agogue @ mines-paristech.fr) \\ Centre de Gestion Scientifique, Mines ParisTech, Paris, France \\ Wim De Neys (wim.de-neys@parisdescartes.fr) \\ CNRS, LaPsyDE (CNRS Unit 3521), Paris Descartes University, Paris, France
}

\begin{abstract}
Influential work on human thinking suggests that our judgment is often biased because we minimize cognitive effort and intuitively substitute hard questions by easier ones. Recent work with adults who solved the bat-and-ball problem, one of the most publicized examples of the substitution bias, suggests that people realize they are doing this and notice their mistake. In the present paper we look at the development of this substitution bias sensitivity. A group of young adolescents solved standard and isomorphic control versions of the bat-and-ball problem in which reasoners experience no intuitive pull to substitute. Adults have been shown to be less confident in their substituted, erroneous bat-and-ball answer than in their answer on the control version that does not give rise to the substitution. However, the present study established that this critical confidence drop was less pronounced for young adolescents. This implies that in contrast with adults, young adolescents do not yet fully acknowledge the questionable nature of their biased answer and remain more oblivious to the substitution. That is, young adolescent reasoners seem to behave more like happy fools who blindly answer erroneous questions without realizing it.
\end{abstract}

Keywords: Decision-making; Bias; Development

\section*{Introduction}

Human reasoners have been characterized as cognitive misers who show a strong tendency to rely on fast, intuitive processing rather than on more demanding, deliberate thinking (Evans, 2008; Kahneman, 2011). Although the fast and effortless nature of intuitive processing can sometimes be useful, it can also bias our reasoning. It has been argued that the key to this bias is a process of so-called attribute substitution - when people are confronted with a difficult question they often intuitively answer an easier one instead (e.g., Kahneman, 2011; Kahneman \& Frederick, 2002). Consider the following example:

A bat and a ball together cost \(\$ 1.10\). The bat costs \(\$ 1\) more than the ball. How much does the ball cost?

When you try to answer this problem, the intuitive answer that immediately springs to mind is " 10 cents".

Indeed, about \(80 \%\) of university students who are asked to solve the "bat-and-ball" problem give the " 10 cents" answer (e.g., Bourgeois-Gironde \& Vanderhenst, 2009). But it is wrong. Obviously, if the ball were to cost 10 cents, the bat would cost \(\$ 1.10\) (i.e., \(\$ 1\) more) and then the total cost would be \(\$ 1.20\), rather than the required \(\$ 1.10\). The correct response is " 5 cents", of course (i.e., the bat costs \(\$ 1.05\) ). The explanation for the widespread " 10 cents" bias in terms of attribute substitution is that people substitute the critical relational "more than" statement by a simpler absolute statement. That is, "the bat costs \(\$ 1\) more than the ball" is read as "the bat costs \(\$ 1\) ". Hence, rather than working out the sum, people naturally parse \(\$ 1.10\), into \(\$ 1\) and 10 cents which is easier to do. In other words, because of the substitution people give the correct answer to the wrong question.

The bat-and-ball problem is considered a paradigmatic example of people's cognitive miserliness (e.g., BourgeoisGironde \& Vanderhenst, 2009; Kahneman, 2011; Kahneman \& Frederick, 2002; Toplak, West, \& Stanovich, 2011). After all, the problem is really not that hard. Clearly, if people would reflect upon it for even a moment they would surely realize their error and notice that a 10 cents ball and a bat that costs a dollar more cannot total to \(\$ 1.10\). Hence, the problem with attribute substitution seems to be that people do typically not notice that they are substituting and do not realize their error (Kahneman \& Frederick, 2005; Thompson, 2009; Toplak et al., 2011). This can sketch a somewhat bleak picture of human rationality: Not only do we often fail to reason correctly, much like happy fools, we do not even seem to realize that we are making a mistake.

However, the fact that decision-makers do not deliberately reflect upon their response does not necessarily imply that they are not detecting the substitution process. That is, although people might not engage in deliberate processing and might not know what the correct answer is, it is still possible that they have some minimal substitution sensitivity and at least notice that their substituted " 10 cents" response is not completely warranted (e.g., Alter, Oppenheimer, Epley, \& Eyre, 2007; De Neys, 2012; De

Neys \& Bonefon, 2013; Oppenheimer, 2008; Thompson \& Morsanyi, 2012).

De Neys, Rossi, and Houdé (2013) recently tested this hypothesis. They designed a control version of the bat-andball problem that does not give rise to attribute substitution. Consider the following example:

\section*{A magazine and a banana together cost \(\$ 2.90\). The magazine costs \(\$ 2\). How much does the banana cost?}

People will tend to parse the \(\$ 2.90\) into \(\$ 2\) and 90 cents just as naturally as they parse \(\$ 1.10\) in the standard version. However, the control version no longer contains the relative statement (" \(\$ 2\) more than the banana") that triggers the substitution. That is, in the control version De Neys et al. explicitly presented the easier statement that participants were supposed to be unconsciously substituting. After solving each version participants were asked to indicate their response confidence. De Neys et al., reasoned that if participants are completely unaware that they are substituting when solving the standard version, the standard and control version should be isomorphic and response confidence should not differ. However, if people are indeed not completely oblivious to the substitution and have some minimal awareness of the questionable nature of their answer, response confidence should be lower after solving the standard version.

De Neys et al. (2013) observed that biased "10 cents" reasoners showed a decreased confidence in the correctness of their answer on the standard bat-and-ball problem. The authors interpreted this as showing that although reasoners often fail to deliberately reflect on their answer, they nevertheless intuitively sense that their response is questionable and are not oblivious to the substitution (see De Neys, 2012, for related suggestions). In the present study we use a developmental approach to validate this claim. Note that a key processing requisite for detecting an unwarranted substitution is that one monitors one's reasoning for conflict between an intuitively cued substituted question and the original phrasing (De Neys \& Glumicic, 2008; Kahneman, 2011). Now, developmental studies in the cognitive control field have established that such basic error or conflict monitoring abilities increase spectacularly throughout adolescence (e.g., Davies et al., 2004; Fitzgerald et al., 2010; Santesso \& Segalowitz, 2008). This has been linked to the late maturation of the Anterior Cingulate Cortex, the brain structure that is supposed to be mediating the monitoring process (e.g., Botvinick, Cohen, \& Carter, 2004; De Neys, Vartanian, \& Goel, 2008; Santesso \& Segalowitz, 2008). In general, this suggests that younger reasoners should be less efficient at detecting the biased nature of their substituted judgments than adults. In other words, if adults' decreased confidence in the De Neys et al. (2013) study indeed results from a successful substitution monitoring or sensitivity, one can also predict that the confidence effects should be less pronounced for younger, adolescent reasoners. More specifically, when younger reasoners give a biased response on the standard version of
the bat-and-ball problem, they should show a higher confidence in the correctness of their substituted answer than adult reasoners. Of course, on the control version that does not give rise to attribute substitution, any differential age-related substitution sensitivity, should not affect the confidence ratings.

To test this hypothesis we presented a group of young adolescents with the standard and control version of the bat-and-ball problem and recorded their response confidence. The performance of this group of adolescents was contrasted with that of the adults in the original De Neys et al. (2013) study. At a theoretical level, this will help us to validate De Neys et al.'s substitution claims. Clearly, from a developmental point of view, it is also important to start documenting possible age-related differences in substitution detection skills in its own right.

\section*{Experiment}

\section*{Method}

Participants. A total of 115 adolescents (average age \(=\) 14.89 years, \(\mathrm{SE}=.03\) ) participated in the study. All participants were Grade 9 students in a local middle school in the Paris region. Performance of these adolescents was contrasted with the performance of the 248 adult undergraduates (average age \(=22\) years, \(\mathrm{SE}=.18\) ) in the study of De Neys et al. (2013).

Material and Procedure. Material and procedure were based on the study of De Neys et al. (2013). All participants were presented with a standard and control version of the bat-and-ball problem. The problems were translated in French and adjusted to the European test context (see Appendix). To minimize surface similarity, we also modified the superficial item content of the two problems (i.e., one problem stated that a pencil and eraser together cost \(\$ 1.10\), the other that a magazine and banana together cost \(\$ 2.90\) ). Both problems were printed on separate pages of a booklet. To make sure that the differential item content did not affect the findings, the item content and control status of the problem were completely crossed. For half of the sample we used the pencil/eraser \(/ \$ 1.10\) content in the standard version and magazine/banana/ \(\$ 2.90\) content in the control version. For the other half of the sample the content of the two presented problems was switched. Presentation order of the control and standard version was also counterbalanced: Approximately half of the participants solved the control version first, whereas the other half started with the standard version \({ }^{1}\). An overview of the material is presented in the Appendix.

Immediately after participants wrote down their answer they were asked to indicate how confident they were that

\footnotetext{
\({ }^{1}\) Note that when the problem content and presentation order factors were entered as additional control factors in our main analyses the reported effects were not affected.
}
their response was correct by writing down a number between 0 (totally not sure) and \(100 \%\) (totally sure). Note that we only intend to use this measure to contrast people's relative confidence difference in the standard and control versions. Obviously, the confidence ratings will be but a proxy of people's phenomenal confidence state. The response scale is not immune to measurement biases such as end preferences or social desirability effects (e.g., Berk, 2006). For example, since it might be hard to openly admit that one has given a response that one is not confident about, mere social desirability can drive people's estimates upwards. This implies that one needs to be cautious when interpreting absolute confidence levels. However, such interpretative complications can be sidestepped when contrasting the relative rating difference in two conditions. Any general response scale bias should affect the ratings in both conditions. Consequently, our analyses focus on the relative confidence contrast and we refrain from making claims based on the absolute confidence levels.

\section*{Results and Discussion}

Accuracy. Adolescents' and adults' scores on the standard and control bat-and-ball problem version were entered in a 2 (problem version, within-subjects) x 2 (age group, between subjects) mixed model ANOVA. As expected, there was a main effect of the Problem Version factor, F(1, 361) = \(1027.74, \mathrm{p}<.0001, \eta^{2} \mathrm{p}=.74\). In line with previous studies, overall only \(20 \%\) ( \(\mathrm{SE}=2.2 \%\) ) of participants managed to solve the standard bat-and-ball problem correctly. However, the control version that did not give rise to substitution was solved correctly by \(99 \%\) ( \(\mathrm{SE}=.5 \%\) ) of the participants. Accuracy did not differ in the two age groups; the Age Group and Age Group x Problem Version interaction did not reach significance, both Fs \(<1\).

Note that incorrect responses on the standard version were almost exclusively (i.e., 361 out of 363 responses) of the " 10 cents" type suggesting that biased participants were not simply making a random guess but indeed engaged in the postulated substitution process \({ }^{2}\).

Confidence ratings. Our crucial question concerned participants' response confidence. A first analysis focused on the response confidence of reasoners who substituted and gave the erroneous " 10 cents" response on the standard version. These participants' confidence ratings were entered in a 2 (problem version, within-subjects) x 2 (age group, between subjects) mixed model ANOVA. Results showed that there was a main effect of the Problem Version factor. As De Neys et al. (2013) already established, overall, people's confidence in their erroneous " 10 cents" response was lower than the confidence in their control version answer that did not give rise to the substitution, \(F(1,285)=\)

\footnotetext{
\({ }^{2}\) The few incorrectly solved control trials and the "non-10 cents" incorrectly solved standard trials were discarded for the subsequent confidence analyses.
}
57.9, \(\mathrm{p}<.0001, \eta^{2} \mathrm{p}=.17\). However, the critical finding was that this effect was indeed less clear for adolescent reasoners. As Figure 1 (top panel) shows, the Age Group and Problem Version factor tended to interact, F(1, 285) = 3.78, \(\mathrm{p}<.055, \eta^{2} \mathrm{p}=.01\), and there was also a main effect of the Age Group factor, \(\mathrm{F}(1,285)=5.11, \mathrm{p}<.025, \eta^{2} \mathrm{p}=.02\).

Follow-up analyses established that in contrast with biased adolescents, biased adults showed specifically more doubt in the correctness of their response when solving the standard bat-and-ball version, \(F(1,285)=5.02, \mathrm{p}<.05, \eta^{2} \mathrm{p}\) \(=.02\). On the control problem, that did not give rise to attribute substitution, both age groups' confidence did not differ, \(F(1,285)=1.16, p=.28\). This establishes that the critical lower confidence ratings on the standard problem in the adult group are not confounded by a general age-related confidence decrease but result from a differential substitution sensitivity. When adults and adolescents do not substitute, their confidence does also not differ. Clearly, if adults would simply show overall more doubt in their judgments, their confidence ratings on the control problem should have been lower too.

"10 cents" biased reasoners


Figure 1. Response confidence on standard and control versions of the bat-and-ball problem for participants who answered the standard problem incorrectly (" 10 cents" biased reasoners, top panel) and correctly (" 5 cents" correct reasoners, bottom panel) in the two age groups. Error bars are standard errors.

This conclusion is further supported when we focus on the confidence ratings of those participants who did not substitute on the standard problem and solved it correctly. Confidence ratings for these participants were also subjected
to a 2 (problem version, within-subjects) x 2 (age group, between subjects) mixed model ANOVA. Results are shown in Figure 1 (bottom panel). As Figure 1 shows, overall the problem version effect on the confidence ratings (i.e., \(93 \%\) standard vs. \(97 \%\) control) was far less clear for correct than for biased reasoners, \(\mathrm{F}(1,68)=5.02, \mathrm{p}<.05, \eta^{2} \mathrm{p}=.02\). In and by itself this is not surprising. Indeed, it makes sense that people who actively reflected upon their judgment and resisted the substitution also knew that their response was likely to be correct. The critical point here is that in this analysis neither the Problem Version \(x\) Age Group interaction, nor the main effect of Age Group were significant, both Fs < 1. Hence, here too, adolescents and adults who did not substitute and reasoned correctly did not show a differential response confidence. This further strengthens the claim that the age-related decreased response confidence on the standard problem that we observed for biased reasoners results from an increased substitution bias sensitivity.

\section*{General Discussion}

The present study indicates that human reasoners become more sensitive to substitution bias throughout their development. The previously observed lowered response confidence after solving the standard bat-and-ball problem, was less clear for biased adolescents. That is, in contrast with adults, 15 -year old adolescents seem to have a harder time detecting the erroneous nature of their substituted judgment. This pattern fits with basic cognitive control studies that indicate that adolescents' basic error or conflict monitoring skills are not fully developed (e.g., Davies et al., 2004; Fitzgerald et al., 2010; Santesso \& Segalowitz, 2008). With respect to attribute substitution during reasoning this implies that young adolescents do not yet fully acknowledge the questionable nature of their biased answer and remain more oblivious to the substitution. In that sense, adolescents do seem to behave like happy fools who blindly answer erroneous questions without realizing it.

We mentioned that our study can have important implications for the developmental field. Some ten years ago, Markovits and Barrouillet (2004) noted in a special developmental issue of the journal Thinking and Reasoning that although reasoning and decision-making were once one of the prime research areas for developmental scientists, interest had faded in more recent years. Markovits and Barrouillet suggested that one of the reasons for this decline was the rise of the "Heuristics and Biases" research program and its demonstration of the widespread bias in human reasoning. This massive bias seemed to point to a developmental standstill in human reasoning. That is, if even the vast majority of educated university students fail to solve basic reasoning problems, one might easily get the impression that there doesn't seem to be a lot of development going on. At first sight, our developmental study might have seem to strengthen this conclusion. Indeed, when looking at the accuracy rates we did not find
any age-related improvement. Adults seemed to perform as badly as adolescents. However, looking closely at the substitution detection process and confidence data suggests that the lack of development is more apparent than real. Although both adults and adolescents are indeed biased most of the time, our findings indicate that an important difference between the age groups is that adults at least detect that their responses are biased. Consistent with recent insights in the developmental field (e.g., Brainerd, Reyna, \& Ceci, 2008; Klaczynski, Byrnes, \& Jacobs, 2001; Houdé, 2007; Reyna \& Farley, 2006; Reyna et al., 2003) this differential substitution bias awareness argues against the idea of a developmental standstill in human reasoning.

It is important to clarify some potential misconceptions and critiques about our work. For example, some critics might spontaneously argue that since our control bat-andball version is easier than the standard version our findings with adults are trivial since they simply show that people are more confident when answering an easy question than when answering a hard question. It is important to stress that this critique is begging the question. The crucial question is of course whether or not people realize that the classic version is hard. That is, the control version presents the easier statement that participants are supposed to be unconsciously substituting. What we want to know is whether or not people note this substitution. If people do not notice it, then the two problems should be isomorphic and they should be considered equally hard. In other words, arguing that adults notice that the classic problem is harder than the control problem underscores the point that they are not oblivious to the substitution.

A related spontaneous critique is that our confidence findings might result from mere guessing rather than from substitution sensitivity. In general, if people do not know an answer to a problem and guess, they presumable realize this and will also give a low confidence rating. Hence, a critic might argue that the lower confidence in adult groups does not necessarily point to substitution sensitivity but merely to a rather trivial "guessing awareness". However, this critique is readily discarded. In the present study more than \(99 \%\) of the erroneous bat-and-ball responses were of the " 10 cents" type. This is the response that people should pick if they engage in the postulated substitution process. Clearly, if people were biased and less confident because they were merely guessing, we should have observed much more random erroneous answers.

In the present study we focused on the bat-and-ball problem because it is one of the most vetted and paradigmatic examples of people's substitution bias (e.g., Bourgeois-Gironde \& Vanderhenst, 2009; Kahneman, 2011; Kahneman \& Frederick, 2002; Toplak, West, \& Stanovich, 2011). However, attribute substitution has also been proposed as an explanation for people's judgment errors in other classic reasoning tasks such as the base-rate neglect or conjunction fallacy task (Kahneman \& Frederick, 2002). Although it has been argued that these task might be less suited to test substitution claims (e.g., Bourgeois-Gironde \&

Vanderhenst, 2009; see also Pennycook, Fugelsang, Koehler, 2012; Klauer \& Singmann, 2012), one might nevertheless wonder whether the present findings can be generalized across these tasks. Some emerging evidence suggests they might. For example, a recent study showed that when adult reasoners give a biased response to standard conjunction or base-rate neglect problems, they also indicate to be less confident about their response compared to control problems. Consistent with the present findings, these effects were not always observed in younger samples (e.g., De Neys, Cromheeke, \& Osman, 2011; see also De Neys \& Feremans, 2013). This gives us some initial indication of the generality of the present findings.

With the present paper we hope to have presented a critical building block to stimulate further research on the development of substitution sensitivity. Our intial data suggest that although most adolescents and adults fall trap to substitution bias, adult reasoners at least detect their bias and realize that their response is questionable. We believe that the potentially severe consequences of adolescents' bias detection difficulties should become a primary research focus for developmental and educational scientists.

\section*{Acknowledgements}

Preparation of this manuscript was supported by a grant from the Agence Nationale de la Recherche (ANR-JSH2-0007-01).

\section*{References}

Alter, A. L., Oppenheimer, D. M., Epley, N., \& Eyre, R.N. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. Journal of Experimental Psychology: General, 136, 569-576.
Berk, R. A. (2006). Thirteen strategies to measure college teaching: a consumer' \(s\) guide to rating scale Construction, assessement, and decision making for faculty, administrators, and clinicians. Sterling, VA: Stylus.
Botvinick, M. M., Cohen, J. D., \& Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: An update. Trends in Cognitive Sciences, 12, 539-546.
Bourgeois-Gironde, S., \& Vanderhenst, J. B. (2009). How to open the door to System 2: Debiasing the Bat and Ball problem. In S. Watanabe, A.P Bloisdell, L. Huber, A. Young (Eds.), Rational animals, irrational humans (pp. 235-252). Tokyo, Japan: Keio University Press.
Brainerd, C. J., Reyna, V. F., \& Ceci, S. J. (2008). Developmental reversals in false memory: A review of data and theory. Psychological Bulletin, 134, 343- 382.
Davies, P. L., Segalowitz, S. J., \& Gavin, W. J. (2004). Development of response-monitoring ERPs in 7-to 25-year-olds. Developmental Neuropsychology, 25, 355-376.

De Neys, W. (2012). Bias and conflict: A case for logical intuitions. Perspectives on Psychological Science, 7, 2838.

De Neys, W., \& Bonnefon, J. F. (2013). The whys and whens of individual differences in individual thinking biases. Trends in Cognitive Sciences, 17, 172-178.
De Neys, W., Cromheeke, S., \& Osman, M. (2011). Biased but in doubt: Conflict and decision confidence. PLoS ONE, el5954. DOI:10.1371/journal.pone. 0015954
De Neys, W., \& Feremans, V. (2013). Development of heuristic bias detection in elementary school. Developmental Psychology, 49, 258-69.
De Neys, W., \& Glumicic, T. (2008). Conflict monitoring in dual process theories of reasoning. Cognition, 106, 12481299.

De Neys, W., Moyens, E., \& Vansteenwegen, D. (2010). Feeling we're biased: autonomic arousal and reasoning conflict. Cognitive, Affective, \& Behavioral Neuroscience, 10, 208-216.
De Neys, W., Rossi, S., \& Houdé, O. (2013). Bats, balls, and substitution sensitivity: Cognitive misers are no happy fools. Psychonomic Bulletin \& Review.
De Neys, W., Vartanian, O., \& Goel, V. (2008). Smarter than we think: When our brains detect that we are biased. Psychological Science, 19, 483-489.
Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. Annual Review of Psychology, 59, 255-278.
Evans, J. B. S. T. (2010). Intuition and reasoning: A dual process perspective. Psychological Inquiry, 21, 313-326.
Fitzgerald, K. D., Perkins, S. C., Angstadt, M., Johnson, T., Stern, E. R., Welsh, R. C., et al. (2010). The development of performance-monitoring function in the posterior medial frontal cortex. Neuroimage, 49, 3463-3473.
Frederick, S. (2005). Cognitive reflection and decision making. Journal of Economic Perspectives, 19, 25-42.
Pennycook, G., Fugelsang, J. A., \& Koehler, D. J. (2012). Are we good at detecting conflict during reasoning. Cognition, 124, 101-106.
Houdé, O. (2007). First insights on neuropedagogy of reasoning. Thinking \& Reasoning, 13, 81-89.
Kahneman, D. (2011). Thinking, Fast and Slow. New York: Farrar, Strauss, Giroux.
Kahneman, D., \& Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, \& D. Kahneman (Eds.), Heuristics \& biases: The psychology of intuitive judgment (pp. 49-81). New York: Cambridge University Press.
Kahneman, D., \& Frederick, S. (2005). A model of heuristic judgment. In K. J. Holyoak, \& R. G. Morrison (Eds.), The Cambridge Handbook of Thinking and Reasoning (pp. 267-293). New York: Cambridge University Press.
Klaczynski, P. A., Byrnes, J. B., \& Jacobs, J. E. (2001). Introduction: Special issue on decision making. Journal of Applied Developmental Psychology, 22, 225-236.
Klauer, K. C., \& Singmann, H. (2012). Does logic feel good? Testing for intuitive detection of logicality in
syllogistic reasoning. Journal of Experimental Psychology: Learning, Memory, \& Cognition. Advance online publication. DOI: 10.1037/a0030530.
Oppenheimer, D. M. (2008). The secret life of fluency. Trends in Cognitive Science, 12, 237-241.
Markovits, H., \& Barrouillet, P. (2004). Why is understanding the development of reasoning important?. Thinking and Reasoning, 10, 113-121.
Reyna, V. F., \& Farley, F. (2006). Risk and rationality in adolescent decision making: Implications for theory, practice, and public policy. Psychological Science in the Public Interest, 7, 1-44.
Reyna, V. F., Lloyd, F. J., \& Brainerd, C. J. (2003). Memory, development, and rationality: An integrative theory of judgement and decision-making. In S. Schneider \& J. Shanteau (Eds.), Emerging perspectives on judgment and decision research (pp. 201-245). New York: Cambridge University Press.
Santesso, D. L., \& Segalowitz, S. J. (2008). Developmental differences in error-related ERPs in middle- to lateadolescent males. Developmental Psychology, 44, 205217.

Stanovich, K. E. (2010). Rationality and the reflective mind. New York: Oxford University Press.
Toplak, M. E., West, R. F., \& Stanovich, K. E. (2011). The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. Memory \& Cognition, 39, 1275-1289.
Thompson, V. A. (2009). Dual process theories: A metacognitive perspective. In J. Evans and K. Frankish (Eds.), In Two Minds: Dual Processes and Beyond. Oxford, UK: Oxford University Press.
Thompson, V. A. \& Morsanyi, K. (2012). Analytic thinking: Do you feel like it? Mind \& Society, 11, 93-105.

\section*{Appendix}

\section*{Standard versions}

\section*{French}

Un crayon et une gomme coûtent 1.10 euro au total. Le crayon coûte 1 euro de plus que la gomme. Combien coûte la gomme?
\(\qquad\) centimes

Un magazine et une banane coûtent 2.90 euros au total. Le magazine coûte 2 euros de plus que la banane. Combien coûte la banane?
\(\qquad\) centimes

\section*{English translation}

A pencil and an eraser cost 1.10 euro in total. The pencil costs 1 euro more than the eraser. How much does the eraser cost?
\(\qquad\) cents

A magazine and a banana cost 2.90 euro in total. The magazine costs 2 euro more than the banana. How much does the banana cost?
\(\qquad\) cents

\section*{Control versions}

\section*{French}

Un crayon et une gomme coûtent 1.10 euro au total. Le crayon coûte 1 euro. Combien coûte la gomme?
\(\qquad\) centimes

Un magazine et une banane coûtent 2.90 euros au total. Le magazine coûte 2 euros. Combien coûte la banane?
\(\qquad\) centimes

\section*{English translation}

A pencil and an eraser cost 1.10 euro in total. The pencil costs 1 euro. How much does the eraser cost?
\(\qquad\) cents

A magazine and a banana cost 2.90 euro in total. The magazine costs 2 euro. How much does the banana cost?
\(\qquad\) cents

\section*{Even if after If then conditionals}

\author{
José Antonio Ruiz-Ballesteros (qframe@correo.ugr.es) \\ Facultad de Psicología. Universidad de Granada. Campus de Cartuja S/N. Granada 18071 (Spain)
}

\author{
Sergio Moreno-Ríos (semoreno@ugr.es) \\ Facultad de Psicología. Universidad de Granada. Campus de Cartuja S/N. Granada 18071 (Spain)
}

\begin{abstract}
This study evaluates how people represent "even if" conditionals when they have to integrate them with previous "if then" conditionals and also make an inference. The terms in the premises were ordered to facilitate their integration (Figure 1: If A then B; Even if B C). In half the cases, the "even if" conditional was expressed with a negation instead of an affirmation (If A then B; Even if not B C). Participants had to infer what followed, given A or C. Previous results showed that in comprehension tasks, where information had to be integrated, counterfactual conditionals seemed to be represented with just one situation (B and C). By contrast, when people had to make inferences with these conditionals, they seemed to represent two situations. In any case, counterfactual seem to be represented with two situations (B and \(C\), and not B and C). In our task, people had to do both: to infer and to integrate. Results showed that the use of negations and the direction in the inference had an effect on the endorsed inferences, but the two factors did not interact. The need to integrate premises did not block access to the two "even if" situations in an inference task.
\end{abstract}

Keywords: semifactual conditionals; directionality; mental models.

\section*{Introduction}

Some previous results have shown that when people make inferences with a semifactual conditional, they represent two mental models. For example, the conditional "Even if it had been raining she would have gone to the party": people seem to represent the factual case "It was not raining and she went to the party" and the hypothetical case "It was raining and she went to the party" (see Moreno-Ríos, García-Madruga \& Byrne, 2008). Semifactual conditionals are similar to counterfactual conditionals, but with different initial representations: given "If A had been the case then B would have been the case" people think about two possibilities from the outset, noting one as the 'facts' (not-A and not-B) and the other as 'imagined' (A and B) possibilities (Johnson-Laird \& Byrne, 2002). Forinstance, if we take the above example: "If it had been sunny then she would have gone to the party" we see that this sentence suggests, on the one hand, the representation that really "It was not sunny and she did not go to the party" and, on the other hand, the possibility that "It was sunny and she went to the party". Results with reasoning tasks are consistent with the two initial representations proposed. However, priming studies (Santamaría, Espino \& Byrne,
2005) evidenced that when people read an "even if A B" semifactual, they are primed to read a subsequent 'not-A and B ' conjunction more quickly than when they have read a factual "if A then B" conditional, whereas they read 'A and \(B^{\prime}\) just as quickly after reading the semifactual as after the factual conditional. Unexpectedly, these authors found that "if" counterfactuals did not prime the 'not-A and not-B' possibility more than "even if" semifactuals. That happened only with counterfactual conditionals and not with semifactual conditionals (see also, Gómez-Veiga, GarcíaMadruga \& Moreno-Ríos, 2010). Gómez-Veiga et al. (2010) proposed that it is possible that the comprehension tasks lead to a less exhaustive representation than inference tasks with only one part of the information. One possible cause is that in the comprehension task, the conditional information must be integrated with the information given previously. That is, the comprehension task could induce a simpler strategy to avoid the working memory load using just one of the two mental models.

In the present study, we use an inference task with a semifactual "even if" conditional preceded by a related "if then" conditional. We study how people make inferences integrating the information given by the premises. We chose the simplest way of ordering the terms in the premises: Figure 1 (If A then B; Even if B C). Figure 1 has been shown to be the easiest configuration of terms in premises to facilitate integration with conditional premises (e.g., Bara, Bucciarelli \& Johnson-Laird, 1995). Also, we include sentences with a negation in the first term of the "even if" conditional. If people represent only the terms, the integration is not possible in this condition. For example, consider "If the sky was overcast it was raining" (If A then B), "Even if it had not been raining she would have gone to the party" (Even if not B C). People can consider the simple situation of "not raining" mentioned in the "even if" conditional, or they can also think of the actual situation: "It was raining and she went to the party". Different responses are expected depending on whether people manage to consider all the representations derived from the "if then" and "even if" conditionals (see Table 1, Complete representation) or just a set of possible situations (some mental models), as assumed by the mental model theory (see Table 1, Initial representation).

Some studies on the integration of premises with different relational statements in deduction (e.g., Oberauer, Hörnig, Weidenfeld \& Wilhelm, 2005; Oberauer \& Wilhelm, 2000)
have shown that the directionality in the inference is a factor that depends on the inner directionality represented in the mental models, and this could influence the conclusion (see Oberauer et al., 2005). Oberauer and Wilhelm (2000), using picture verification tasks (sentence-picture), found that "if then" conditionals have an inherent forward directionality: people seem to process "if A then \(B\) " in a preferred order, from \(A\) to \(B\). Inferences based on \(A\) or on not \(A\) are faster and easier than those based on B or not B.

Table 1: Some possible sets of representations of the premises. Every line represents a mental model. The symbol \(\neg\) means "negation". The "*" symbol for "even if" representation is the hypothesised model (simplest representation of "even if").
\begin{tabular}{|c|c|c|}
\hline Affirmative condition & Initial representation & Complete representation \\
\hline If A then B & A B & \[
\begin{array}{rr}
\mathrm{A} & \mathrm{~B} \\
\neg \mathrm{~A} & \mathrm{~B} \\
\neg \mathrm{~A} & \neg \mathrm{~B}
\end{array}
\] \\
\hline Even if B C & \[
\begin{array}{rl}
\mathrm{B} & \mathrm{C} \\
\neg \mathrm{~B} & \mathrm{C}
\end{array}
\] & \[
\begin{array}{rr}
\mathrm{B} & \mathrm{C} \\
\neg \mathrm{~B} & \mathrm{C} \\
\neg \mathrm{~B} & \neg \mathrm{C}
\end{array}
\] \\
\hline
\end{tabular}
\begin{tabular}{|c|c|c|}
\hline Negative condition & Initial representation & Complete representation \\
\hline If \(A\) then B & A B & \[
\begin{array}{rr}
\mathrm{A} & \mathrm{~B} \\
\neg \mathrm{~A} & \mathrm{~B} \\
\neg \mathrm{~A} & \neg \mathrm{~B}
\end{array}
\] \\
\hline Even if not B C & \[
\begin{array}{cl}
\neg \mathrm{B} & \mathrm{C}^{*} \\
\mathrm{~B} & \mathrm{C}
\end{array}
\] & \[
\begin{array}{cc}
\neg \mathrm{B} & \mathrm{C} \\
\mathrm{~B} & \mathrm{C} \\
\mathrm{~B} & \neg \mathrm{C}
\end{array}
\] \\
\hline
\end{tabular}

Another basic result in deductive research is that reasoning with negative propositions is harder than with affirmative propositions (for example, see Evans, Newstead \& Byrne, 1993). Therefore, we would expect an increase in the number of errors and nothing follows responses for the backward inferences than the forward ones, and the same for inferences with negative premises. Therefore, inferences could be influenced in an additive way by directionality and by the negation, but they will be determined by the predicted representations of the premises and their integration.

\section*{Predictions}

We evaluate whether people create double representations to make inferences from semifactual conditionals when they have to integrate this information with a previous "if then" conditional. Different predictions are obtained according to
whether people are able to look for all the alternatives, just consider one alternative from each conditional or represent two mental models from "even if" conditionals. Table 2 shows the different predictions for each condition.

Table 2: Conclusions predicted for inferences after the integration of premises. See text for description. The inference can be endorsed when the end terms ( A and C ) can be connected and lead to one unique conclusion. In other cases, the correct conclusion is "nothing follows". Parentheses are used when middle terms do not match.
\begin{tabular}{|c|c|c|c|}
\hline Forward/ backward inferences & Double represt. & Simple represt. & Complete represt. \\
\hline Affirmative Given A, then C ? & C & C & \[
\begin{aligned}
& \mathrm{C} \\
& \text { A B, B C }
\end{aligned}
\] \\
\hline Affirmative Given C, then A? & A B, B C & A & \[
\begin{aligned}
& \text { Nothing } \\
& \text { follows } \\
& \text { A B, B C } \\
& \neg \text { A B, B C } \\
& \neg \text { A } \neg \mathrm{B}, \neg \text { B C }
\end{aligned}
\] \\
\hline Negative Given A, then C ? & \begin{tabular}{l}
C \\
A B, B C
\end{tabular} & Nothing follows
\[
\mathrm{AB},(\neg \mathrm{BC})
\] & Nothing follows A B, B C \(\mathrm{AB}, \mathrm{B} \neg \mathrm{C}\) \\
\hline Negative Given C, then A ? & A B, B C & Nothing follows
\[
\mathrm{AB},(\neg \mathrm{BC})
\] & Nothing follows
\[
\begin{gathered}
\neg \mathrm{A} \\
\neg \mathrm{~B}, \\
\neg \mathrm{~B} \\
\neg \mathrm{~B} \\
\text { B C } \\
\text { A } \\
\text { B, B C } \\
\text { C }
\end{gathered}
\] \\
\hline
\end{tabular}

There are three columns for each of the three possibilities for representing "even if" (as shown in Table 1): two possibilities, one possibility and all the possibilities. The first two rows represent predictions for affirmatives and the second two for negative "even if". In every case, for forward inferences (given A, what follows? C, not C or nothing follows) and for backward inferences (given C, what follows? A, not A or nothing follows). For example, the first column shows the prediction if people construct a double representation for "even if" conditionals. In the third row we can find predictions for a negative "even if" in a backward inference. Therefore, the structure is: If A then B; Even if not \(\mathrm{B} C\); given A , what follows? People will conclude "C", because there is one and only one way to connect a representation from "if then" \((\mathrm{AB})\) and another
from "even if" (BC) (see Table 1). However, if people represent "even if" with just one mental model (see Table 2, second column third row), they cannot match the terms between the representations of the two premises ( AB , not BC ) and the prediction of the conclusion is "nothing follows". The same conclusion, but for a different reason, is predicted if people can access all the situations consistent with the conditionals (see Table 2, third column, third row). In this case, there are two possible ways to connect the representations, which lead to different conclusions ( AB , \(B C\) and \(A B, B\) not \(C\) ).
If we consider the polarity (affirmative and negative) and the directionality (forward and backward) we can make predictions about the endorsed inferences depending on the set of mental models represented. In general, following previous studies, we would expect the directionality (inherent to conditionals) and polarity (difficulty in processing negative propositions) to have an effect on the difficulty in making inferences. In addition, the nature of the representation should influence the frequency of endorsed inferences. If people represent the two mental models for "even if", no interaction is predicted between polarity and directionality. If people represent just one initial model from every conditional, one effect of polarity will be found in the "nothing follows" conclusion. Finally, if people were able to represent a complete set of possibilities, we would predict an interaction between directionality and polarity in the inferences endorsed (more in affirmative; actually only one in the forward direction condition) and in the "nothing follows" conclusion.

\section*{Method}

\section*{Participants}

Participants were 51 students from Granada University, enrolled in the second year of a psychology degree, who took part for course credits. All participants were native Spanish speakers without any previous training in logic.

\section*{Materials}

Thirty two syllogisms were constructed, half of them (those of interest in this experiment) with the following structure: Eight syllogisms with the form "If A then B; Even if there had been B there would have been C", and another eight that included a negation in the second premise "If A then B; Even if there had not been B there would have been C". Sixteen other fillers were included with a different structure "If C then B; Even if there had been/not been B there would have been A". The sentences were about professions. For example:
Premise 1. "If there was a biologist then there was a lawyer"
Premise 2. "Even if there had been a lawyer there would have been an engineer"
Premise 3. "There was a biologist, therefore..."

Conclusions. Response 1. "There was an engineer", Response 2. "There was not an engineer", Response 3. "Nothing follows".

\section*{Procedure}

All the sentences were presented on a computer screen. Each premise and the conclusion were shown on a separate screen and participants decided by pressing the space bar when to turn to the next statement. After the two conditionals had been presented, a third premise was shown: A in half the trials (and C in the backward condition trials). After that, a screen with three options was shown, with C , not C and nothing follows (A, not A , and nothing follows in the backward condition trials). Participants had to press keys 1, 2, or 3 to choose their respective responses.

\section*{Results}

An analysis of variance (ANOVA) for repeated measures was carried out by participants with the following factors: Directionality (forward and backward) and Semifactual polarity (affirmative and negative). The same analyses were carried out for endorsed inferences (given A, C is accepted and given C , A is accepted) and for nothing follows responses. Results are shown in Table 3.

Table 3: Percentage of endorsed inferences (given A, then C in the forward condition, and given C , then A in the backward condition) and nothing follows responses for affirmative "even if" and negative "even if" in the second premise.
\begin{tabular}{lll}
\hline \begin{tabular}{l} 
Direction \\
of inference
\end{tabular} & Forward & Backward \\
\hline
\end{tabular}

Semifactual Affirmative Negative Affirmative Negative
polarity

Endorsed inferences

Nothing
follows \(\quad 22(3.4) \quad 22(3.0) \quad 39(4.2) \quad 37(3.9)\) inferences
* The values in brackets show standard deviation.

The analysis of nothing follows shows only effect of directionality but no other effects \(\left(\mathrm{F}(1,50)=12.98 ; \eta_{\mathrm{p}}{ }^{2}=.19\right.\); \(\mathrm{p}<.001\) ). More inferences were endorsed for forward than for backward conditions ( \(\mathrm{F}(1,50)=10.6 ; \eta_{\mathrm{p}}^{2}=.17 ; \mathrm{p}<.01\) ) and for affirmative than negative conditions \(\left(\mathrm{F}(1,50)=5.4 ; \eta_{\mathrm{p}}{ }^{2}=.4\right.\); \(\mathrm{p}<.05\) ), but again, they did not interact \(\left(\mathrm{F}(1,50)=.01 ; \eta_{\mathrm{p}}{ }^{2}=0\right.\); \(\mathrm{p}<.9\) ). The lack of this interaction is consistent with the double representation of "even if" conditionals.

\section*{Discussion}

The task used in this study required not only inferring from a conditional but also integrating information before doing so. The frequency of correct responses is low. The working memory capacity, motivation and other factors could have influenced this overall frequency. In any case, we were able to contrast our prediction because the frequency of the inferences in the different conditions varied.

The present study shows that people seem to use two mental representations with "even if" conditionals when they have to integrate this information with other information given previously by a conditional. These results are obtained when the terms in the two conditionals are arranged so that they can be easily connected (Figure 1). A main result seems to stand out clearly. The conclusions in the negative forward and backward conditions (If A then B; Even if not B C) were not blocked. In fact, no differences in the "nothing follows" responses were obtained in this condition regarding the affirmative control condition (If A then B; Even if B C).

Also, we could see that there was an effect of the direction and of the negation in the acceptance of the inferences, but not an interaction between the two factors. That is, the representation alone cannot entirely explain the results of this study. These effects are consistent with previous studies: the forward directionality of "if then" conditionals (see Oberauer et al., 2005) and the effect of negations in reducing the endorsed inferences (Evans et al., 1993). When we introduce a negative proposition, it becomes more difficult to understand the sentences (higher working memory load) and the errors will increase. For example, people could conclude not A when actually they should conclude A. Actually, this kind of error was more frequent for negative sentences, but no differences were obtained for "nothing follows" conclusions.

Oberauer et al. (2005) maintain that the directionality of the conditionals must be represented in the conditional. In our case, only when the direction of the inference matches the direction in the mental model is the inference easier. The present results are consistent with this proposal.

Moreover, the present results do not imply that people always construct a double representation when they make inferences and integrate "if then" and "even if" conditionals. For example, results could be different when the order of the terms makes it more difficult to integrate the premises (for example using a different figure). Also, the time for reading the premises and the time for the conclusion could inform us about the principles that are operating in the integration of premises. For example, at this point we do not know whether "even if" could lead to the apodosis as the "relatum" or if it is the protasis as happens with "if then" clauses.

This is a preliminary study and we cannot prove that the present results can be generalised to conditionals with other content (such as advice, promises, obligations, etc.). The content and the context of conditionals have been shown to influence the mental representation accessed in deduction (see Johnson-Laird \& Byrne, 2002; Handley \& Feeney, 2004). Also, we used an evaluation of conclusions task, but not a generation of conclusions task. We do not expect that
if instead we had used, for example, a generation of conclusion task, it would have led to different initial representations. However, again, this question has not been tested at this point in the research.

The present results are part of a research project that studies how premises are integrated and how we represent semifactual expressions.

\section*{Acknowledgments}

Financial support was provided by the Ministerio de Ciencia e Innovación (PSI2011-29504).

\section*{References}

Bara, B. G., Bucciarelli, M., \& Johnson-Laird, P. N. (1995). Development of syllogistic reasoning. American Journal of Psychology, 108, 157-193.
Evans, J. S. B. T., Newstead, S. E., \& Byrne, R. M. J. (1993). Human reasoning: The psychology of deduction. Hove, UK: Lawrence Erlbaum Associates Ltd.
Gómez-Veiga, I., García-Madruga, J. A., \& Moreno-Ríos, S. (2010). The interpretation of indicative and subjunctive concessives. Acta Psychologica, 134, 245-252
Handley, S., \& Feeney, A. (2004). Reasoning and pragmatics: The case of even-if. In I. A. Noveck \& D. Sperberg (Eds.), Experimental pragmatics. Besingstone: Palgrave.
Johnson-Laird, P. N., \& Byrne, R. M. (2002). Conditionals: A theory of meaning, pragmatics and inference. Psychological Review, 109, 646-678
Moreno-Ríos, S., García-Madruga, J. A., \& Byrne, R. M. J. (2008). Inferences from semifactual 'even if'conditionals. Acta Psychologica, 128, 197-209.
Oberauer, K., Hörnig, R., Weidenfeld, A., \& Wilhelm, O. (2005). Effects of directionality in deductive reasoning: II. Premise integration and conclusion evaluation. The Quarterly Journal of Experimental Psychology Section A, 58, 1225-1247.
Oberauer, K., \& Wilhelm, O. (2000). Effects of directionality in deductive reasoning: I. The comprehension of single relational premises. Journal of Experimental Psychology: Learning, Memory, and Cognition, 26, 1702-1712.
Santamaría, C., Espino, O., \& Byrne, R. M. J. (2005). Counterfactual and semifactual conditionals prime alternative possibilities. Journal of Experimental Psychology: Learning, Memory and Cognition, 31, 11491154.

\title{
What, When and How do the Models of Conceptual Change Explain?
}

\author{
Anna-Mari Rusanen (anna-mari.rusanen@helsinki.fi) \\ ConChaMo/Department of Physics, PO BOX 64, 00014 University of Helsinki, Finland \\ Department of Philosophy, History, Art and Culture Studies, PO BOX 24, 00014 University of Helsinki \\ Otto Lappi (ottolappi@helsinki.fi) \\ Cognitive Science/ Institute of Behavioural Sciences, PO BOX 9 \\ 00014 University of Helsinki, Finland
}

\begin{abstract}
There are certain theoretical issues in conceptual change research that are still puzzling researchers. First, there is no agreement on what kinds of changes in belief and concept systems constitute conceptual change. Second, there is no consensus on what the mechanisms of conceptual change are. Third, there is no common understanding how to explain, model and describe in an exact way these underlying mechanisms. In this paper, we offer a diagnostic analysis these issues by reviewing current theories of conceptual change in a framework of mechanistic explanations of cognitive phenomena, and present a possible sketch for explanations of conceptual change.
\end{abstract}

Keywords: Conceptual Change; Explanation; Cognitive Mechanisms

\section*{Introduction}

Concepts enable thought, reasoning and problem solving. Acquiring new concepts or reorganizing the conceptual framework one uses to think about a domain is a powerful kind of learning. This sort of learning is known as conceptual change.

Conceptual change is one of the most studied fields in science education, and there are hundreds, if not thousands studies, on this topic. However, there are still some foundational issues in conceptual change research on which no clear consensus has emerged, and that are still puzzling most of the researchers. Firstly, there is no agreement on what kinds of changes in belief and concept systems actually constitute conceptual change. Secondly, there is no consensus on what the mechanisms of conceptual change are. Also, when compared to the level of detail at which, say, basic visual processing is understood, often the descriptions of these "mechanisms of conceptual change" are quite shallow and offer no exact specification of the precise structure of mechanisms.

As Clement (2008) remarks, there are very few, if any, models of conceptual change, in which the mechanisms of conceptual change are specified in sufficient detail. This suggests that many of the current accounts might not in a
strict sense qualify as sufficient for explanation and manipulation of learning phenomena involving conceptual change.

Moreover, having numerous loose filler terms in an explanation does not only threaten to undermine its explanatory power, but filler terms may also be barriers to scientific progress when they veil failures of understanding (Craver, 2006, 2007). If, for example, the terms "reassign" or "assimilation" are used to stand for processes with largely unknown properties, then we really do not explain what happens. Instead, we have a possible sketch for an explanation. If this sketch is taken to be genuinely explanatory, then - in the worst case scenario - it is possible that we have only an illusion of explanation instead of having a genuine one (Rosenblitz \& Keil, 2002; also Craver, 2006).

In what follows, we analyse the explanations of conceptual change from a philosophical point of view. Our analysis is partially based on the so-called "mechanistic account of explanation" (Bechtel \& Richardson, 1993; Machamer et al., 2000; Craver, 2006,2007). This mechanistic approach has not previously been applied to explanations of conceptual change.

However, it should be emphasized that the focus of this paper is only on cognitive accounts of conceptual change. There are other accounts of conceptual change that examine conceptual change from socio-cultural, emotional or motivational perspectives. However, explanation of cognitive phenomena is a unique form of explanation, and it is an open question, whether it is possible to extend this form of explanation to cover explanation involving interpersonal dynamics etc. This topic is, however, beyond the scope of this paper.

\section*{Variable Accounts of Conceptual Change}

The study of conceptual change has focused, on the one hand, on the acquisition of commonsense concepts in childhood (e.g. Carey, Spelke) and on the other hand, the
acquisition of scientific concepts in science education, especially at the secondary and tertiary level (Chi, 1992; Chi et al, 1994; Vosniadou, 1992; DiSessa,1993).

In this paper we examine the latter form of research, in which conceptual change is seen as a specific kind of learning process, in which a student does not merely accumulate more knowledge, but her conceptions of phenomena in a certain domain undergo a restructuring process that affects ontological commitments, inferential relations, and standards of explanation (Posner et al, 1982; Carey, 1985; Chi, 1992; Vosniadou, 1994; diSessa, 1993). In a nutshell, this sort of conceptual change can be characterized as transformation process of the initial knowledge-state (a commonsense picture of the world) to one of various outcome knowledge states. The outcome state can be an accurate scientific conception (when the learning process has been successful) or, when the learning process has not been successful, one of a number misconceptions (when it has not).

One difficulty with conceptual change research is that there are a huge number of different accounts of the details conceptual change, and they all characterize conceptual change different ways. In the literature \({ }^{1}\), there are different views on the learner's initial and outcome conceptions, on the trajectories along which change occurs, on the mechanisms that are underlying the conceptual change, on the obstacles of learning and also on the factors that support the change.However, on the basis of a careful reading, the literature seems to suggest that there are roughly three "major" kinds of conceptual change. These different kinds, or forms, of conceptual change can be titled as revision, reinterpretation and invention:

Revision. In some cases conceptual change seem to require a revision of an existing conceptual system. For example, Chi and her colleagues suggest that conceptual change takes the form of category shift (Chi, 1992, 2008; Chi et al, 1994). Another example of this form of conceptual change is described by Thagard's "tree jumping", in which conceptual change happens when hierarchies of concepts are reorganized by shifting a concept from one branch of a hierarchical tree of concepts to another. Also in DiSessa's Knowledge in Pieces- account conceptual change is understood as a form of revisionary process, because conceptual change is seen as a process that integrates initially piecemeal, incoherent (sub)conceptual system by complex process of organizing and reorganizing the elements of the system (diSessa, 1993, 2002,2004).

Reinterpretation. In some cases conceptual requires that a learner gives a new interpretation for a domain. For example, according to Ohlsson (2009) conceptual change occurs when a learner uses analogical transfer to map conceptual system from one domain to new domain to which it has not previously been applied, and to which some other conceptual system had been predominant. Another examples of conceptual change as a form of reinterpretation

\footnotetext{
\({ }^{1}\) For example, for an analysis of the various types or accounts of conceptual change, see Chinn and Samarapungavan (2009).
}
are described by Carey's differentiations (when initially undifferentiated concept is differentiated) and coalescence (when initially distinct concepts are subsumed by a same concept).

Invention. In some cases conceptual change requires construction or production of a novel (for the learner) conceptual system. For example, Carey \((1985,2011)\) describes a form of conceptual change, in which a learner constructs a new set of concepts from already existing concepts by "bootstrapping" in way that makes novel concepts incommensurable with the earlier concepts, because the content of new concepts cannot be captured in terms of any previously possessed concepts. The first stage of bootstrapping, or "Quinian bootstrapping", occurs when a learner encounters a set of explicit public symbols, such as sentences of a scientific theory. These public symbols, "the placeholders", are not initially mapped onto any already existing concepts that a learner holds. Rather, for a learner they are either partially or completely uninterpreted. During the process of learning, these placeholders are then taken up by various "modeling processes", which includes abstract forms of theoretical inference such as analogical reasoning, abduction and induction etc. The idea is that a learner constructs the interpretation or the content for a placeholder by using these different mechanisms. At the end of the process, the placeholders will have conceptual content in virtue of acquiring a stable conceptual role in a new theoretical structure.

Conceptual change as an umbrella term. This variety of different kinds of conceptual change might reasonably be taken to indicate that "conceptual change" is a sort of umbrella term, which covers several types of phenomena instead of referring to a singular type of learning. This would entail that there cannot be a singular "grand theory" of conceptual change, which could explain all possible instances and trajectories of conceptual change.

Instead, explaining conceptual change seems to require that different learning trajectories are explained by referring to different mix of underlying mechanisms and processes (see also Chinn \& Samarapungavan, 2009; Ohlsson, 2009b). These learning trajectories can be considered conceptual change because the learning is seen in some way "radical". However, this does not necessarily indicate that these different phenomena are instances of a common explanandum i.e. a common learning phenomenon, which the various mechanisms would account for. Instead, if one used this as a reason for adopting the umbrella term, it would be merely a pragmatic reason.

\section*{Towards the Explanation of Conceptual Change}

The explanans and explanandum. In the case of conceptual change research, it is not always apparent, what the explanandum (the thing to be explained) and what the explanans (the things that explain) are. For example, Mayer
(2002, p. 671) defines conceptual change as "the mechanism underlying meaningful learning". On this view, meaningful learning would be the explanandum, and conceptual change would be the mechanism that explains the meaningful learning. However, others seem to think of conceptual change as the thing that should be explained, and the explanation should be given in terms of underlying mechanisms. For example, Chinn and Samarapungavan (2009) emphasize this view, when they argue that there are many routes (with many underlying cognitive mechanisms) to conceptual change.

We emphasize the latter view, according to which the explanandum is conceptual change (a learning episode that can be observed behaviorally, e.g. as correct responding to diagnostic questions), and the explanans is given in terms of underlying cognitive mechanisms for variety of reasons.

\section*{Dissection of Explanation.}

From a philosophical point of view, explanation of cognitive phenomenon typically involves at least (1) the characterization of the specific cognitive task performed by a system, (2) the descriptions of how certain cognitive mechanisms execute, produce or sustain the phenomenon to be explained. In some cases, explanation in cognitive science also requires the (3) description of how cognitive mechanisms are implemented in cognitive systems or why it makes rational (or evolutionary) sense that the phenomenon should be sustained in the first place.

Explanation step 1: The characterization of the specific cognitive task. In cognitive explanations of behavioral phenomena, the description for the task is given by characterizing the information processing task, and it answers to questions such as: "What is the cognitive goal of this process" or "What is the cognitive task of this competence?"

This aspect of explanation is important for two reasons; it not only characterizes the cognitive task in a specific way, but it creates also some constraints for the possible underlying mechanisms. This aspect of explanation characterizes, why certain - but not all - learning mechanism are appropriate for fulfilling the cognitive task.

The task of conceptual change. So, what is the task of conceptual change? Even if the issue of the task is not often expressed explicitly in current literature, many seem to echo the same normative intuition that the task of conceptual change is to reorganize the conceptual system in a way that makes - in a case of successful learning - somehow "better". Depending on the larger picture of conceptual change, different authors have described this "better" different ways.

One early formulation can be found in the seminal paper by Posner et al. (1982), where they propose that conceptual change makes the system "more fruitful, intelligible and plausible" etc. (Posner et al, 1982). In their paper, intelligible means roughly that the new conception must be clear enough to make sense to the learner. Plausible means the new conception must be seen as believable, and even
true. Fruitful means the new conception must appear potentially productive to the learner for solving problems and seek for new intellectual directions. The approach Posner et al. propose is based on the Kuhnian idea of paradigm shifts and their emphasis of the conceptual ecology of a student. By conceptual ecology Posner and collegues meant the framework of a learner's conceptions and "cognitive mechanisms", such as analogies, metaphors, explanatory anomalies and so on (Posner et al, 1982; Strike \& Posner, 1992). So, according to this view, conceptual change happens if the changes make the ecology "better" i.e. more productive and fruitful, and it increases the ability to solve problems.

Sometimes this "better" is interpreted in terms of utility. For example, Stellan Ohlsson recently proposed that some forms of conceptual change make the conceptual system more useful (Ohlsson, 2009). In Ohlsson's account cognitive utility measures the usefulness of a knowledge system for a learner. The basic idea is that in a situation, where there are competing knowledge systems, the system that requires less cognitive load, and leads to faster, more efficient and more cognitively satisfactory end states, will become associated with higher strength and will be easier to activate (Ohlsson, 2009). Over time, the system will become the person's "standard way of looking at the target domain" (Ohlsson, 2009).

In some cases, the task of conceptual change is given also in terms of coherence. For instance diSessa (1993, 2002, 2004) describes novice knowledge as a weakly organized system that is highly context dependent and internally inconsistent, thereby lacking internal coherence. In diSessa's account commonsense physical knowledge is organized into p-prims, empirical typologies or low-level abstractions of everyday experience. For example, according to this knowledge-in-pieces- account novices' knowledge systems are fragmented and consist of loosely connected pieces, which often lacks not only coherence but are also employed with little co-ordination (diSessa 1993, 2002, 2004). In diSessa's and colleagues account, the task of conceptual change is to integrate the piecemeal structure of a conceptual system in a way that increases the internal coherence of the system (diSessa, 1993, 2002, 2004).

Coherence, of course, is as Disessa himself writes, "a vague word", but as he continues, "one important core meaning (of coherence) has inherently to do with relations; that is, the meaning of coherence requires an articulation of structure." (diSessa, 2008). Even if the term is often left unspecified in the context of conceptual change studies, a useful description for conceptual coherence can be found, for example, from Thagard and Verbeugt (1998, also Thagard et al, 2002). Thagard and Verbeugt defines coherence as follows: (i) Conceptual coherence is a symmetric relation between the pairs of concept, (ii) a concept coheres with another concept if they are positively associated i.e. if there are objects to which they both apply, (iii) the applicability of a concept to an object may be given perceptually or by some other reliable source, (iv) a concept
incoheres with another concept, if they are negatively associated, i.e. if an object falling under one concept tends not to fall under the other concept. Finally (v) the applicability of a concept to an object depends on the applicability of other concepts. Even if Thagard and Verbeugt speak explicitly about coherence of concepts, there is no reason a priori, why their description of coherence could not be applied to the elements of subconceptual systems or more complex entities, such as elements of belief or knowledge systems.

These three approaches are perhaps the most widely accepted descriptions for the task of conceptual change. In an ideal account, these descriptions would be given in an exact way, but at least to our knowledge there are not any exact formulations of conceptual change available. In addition, philosophically speaking, it is still an open question, what is it about fruitfulness, plausibility, utility or coherence that makes the learning task as an instance of conceptual change. Perhaps, very roughly, one might say that conceptual change happens when a student does not merely accumulate more knowledge, but her conceptions of phenomena in a certain domain undergo a restructuring process that affects the conceptual system in a way that increases utility, plausibility or coherence of that system.

Explanation step 2: The Mechanisms of Conceptual Change. Now, let's move to the second step of explanation. This aspect of explanation answers questions like: "how does the mechanism transform the input to generate the output (step by step)?". In the literature of mechanistic explanations, there are several attempts to specify this notion of "cognitive mechanisms". For example, Bechtel (2008) defines cognitive mechanisms as follows: A (cognitive) mechanism is a structure performing a (information processing) function in virtue of its components parts, component operations, and their organization \({ }^{2}\). Typically in the case of hard core cognitive explanations, these mechanisms are given descriptions by specifying the precise algorithms or by other formal means.

The mechanisms of conceptual change. In the literature, there are many suggestions for the "mechanisms" of conceptual change. For example, Chi talks about categorization and recategorization, while Carey speaks about differentiation, coalescence and bootstrapping. Vosniadou focuses on accommodation and assimilation, Thagard writes about branch jumping and tree jumping, and Ohlsson focuses on resubsumption.

However, often these purported mechanisms of conceptual change are rarely specified with sufficient (computational) detail (for discussion, see Rusanen and Pöyhönen, 2012). The descriptions of these mechanisms are often quite shallow and offer no information about the precise structure of mechanisms. For example, Chi describes conceptual change as a form of recategorization process by saying that "[c]ategorizing is the process of identifying or

\footnotetext{
\({ }^{2}\) There are some controversies about the precise definition of cognitive mechanisms. See Piccinini, 2006, also Shagrir, 2002; Lappi \& Rusanen, 2011.
}
assigning a concept to category to which it belongs"(Chi 2008, 62), and by writing how "Conceptual change is the process of removing misconceptions... (which) are, in fact, miscategorizations of concepts" and "conceptual change is merely a process of reassigning or shifting a miscategorized concept from one category to another" (Chi, 2008, 62, italics added).

However, Chi offers no description of how "identifying" or "assigning" actually happens, or what kind of cognitive mechanisms they actually are. From an explanatory point of view, this is problematic. Genuinely explanatory models are models, in which the phenomenon is explained by giving an accurate and sufficient description of how a (causal) mechanism, a hierarchical system composed of component parts and their properties sustains or produces the phenomenon (Bechtel and Richardson, 1993; Machamer at al., 2000; Craver, 2006, 2007). In addition, genuine explanations offer the ability to say not merely how the system in fact behaves, but to say how it would behave under a variety of circumstances or interventions (Craver 2000, Craver 2007, Woodward 2003).

So, even if even if these "identifyings", "assignings" and "categorizings" (or any other similar "ings"), were constantly referred as mechanisms, they often fail to satisfy the requirements for genuine mechanism descriptions, becausee the structure of these mechanisms is not specified in a detail. Instead, often the purported "mechanisms" are, or include, more or less filler terms. Filler terms describe only the relationship between the input and the output of the process, but they offer little specific information of how the change was brought about.

If a mechanistic model is incomplete, and it includes filler terms, it should rather be called a "mechanism sketch" than a genuine explanation (Craver, 2006, 2007). Philosophically speaking, having numerous filler terms in an explanation does not only threaten to undermine its explanatory power, but filler terms may also be barriers to scientific progress when they veil failures of understanding (Craver, 2006, 2007). If, for example, the term "assign" is used to stand for a process with largely unknown properties, then we really do not explain what happens, but in the worst case scenario we may also have only an illusion of explanation (Craver, 2006; Rozenblitz \& Keil, 2002).

The details of mechanisms. In addition, when the details of these mechanisms (reorganisation, bootstrapping, resubsumption, category shifts, etc.) is analyzed, they are often just collections of some more basic cognitive mechanisms (such as categorization, mapping, transfer, assimilation, accomodation, analogical reasoning, inductive inference, abduction and so forth), which are ultimately thought to be responsible for the conceptual change.

For example, in Stellan Ohlsson's (2009) account conceptual change happens, when a person uses analogical transfer to map conceptual system from one domain A to a new domain B, which has been earlier conceptualized by another system. According to Ohlsson's model, if the new
system is evaluated to be more useful, the target domain is reinterpreted by it.

As Ohlsson says, the resubsumption theory "does not introduce any cognitive processes that are specific to conceptual change" and "no special purpose cognitive mechanism kicks in to produce conceptual change" (Ohlsson, 2009, p. 32). Instead, resubsumption is simply a process, which involves analogical reasoning, transfer, analogy, transfer, different kinds of mapping and interpretation and all these familiar cognitive mechanisms.

From the explanatory point of view, this is not shocking news. It is quite common, as for example Bechtel and Richardson (1993) emphasize, that complex mechanisms are, and often must be, decomposed into simpler (or more basic) submechanisms that are ultimately responsible for the orchestrated functioning of the higher level mechanism (Bechtel \& Richardson, 1993, see also Craver, 2007). However, if the submechanisms are finally doing the explanatory work, they should be given a proper description. If they are not described in a detail, then we really have no explanation as to how conceptual change happens.

Evaluation of relevance. Given the complexity of cognitive processes in general, and especially the complexity of conceptual change, in practice it is really difficult to distinguish those underlying submechanisms (attentional- , memory- , reasoning-, mapping mechanisms etc. ) that are doing the explanatory work from those which are not. As Ohlsson emphasizes (Ohlsson, 2009b, p. 70), a theory of conceptual change just cannot be the list of all possible mechanisms underlying conceptual change, but it must also constraint mechanisms in theoretically principled way. In other words, what we need is a theoretically principled way to evaluate the relevance of submechanisms.

This is, of course, a very difficult demand. Philosophically speaking, one possible line might be to argue that the relevance for a certain mechanism - or certain mechanisms - could be evaluated by knowing how the mechanism`s inputs and outputs interact with their context i.e. by knowing its causal (as opposed to say, intentional relations) with the environment \({ }^{3}\). A natural way to continue this argument would be to refer to the manipulationist account i.e. to argue following Woodward (2003) that those mechanisms are relevant, which do not only have impact on how the cognitive system of a learner in fact behaves, but which have impact also on how it would behave under a variety of circumstances or interventions.

However, it seems to be that in the case of conceptual change - and in genuinely cognitive explanations in general - the explanatory relevance must also be described at least partially by referring the task of the conceptual change as well. A theory of conceptual change should be able to tell, why certain mechanisms are required or are appropriate for achieving conceptual change, and why some other aren't.

The task level description is needed to characterize representational requirements and constraints for the descriptions of appropriate learning mechanisms. If one thinks that the task should define in terms of utility, then one should characterize those mechanisms that are responsible for "utility making". However, for doing this, the task level - utility, coherence, intelligibility - must be specified first, and then this specification provides justification to relevance claims concerning the specific mix of concrete mechanisms underlying conceptual change.

\section*{Concluding remarks}

Conceptual change is organizing a multiplicity of learning mechanisms to achieve learning that makes the conceptual system "better" by way of creating new (for the learner) concepts. These concepts do not "pop" into existence in a miraculous way but are typically gradually and sometimes painfully crafted by the cognitive mechanisms from existing material.

According to the mechanistic account, explanation requires that the mechanisms responsible for a certain type of conceptual change should be specified in a detail. This can be really challenging in the case of conceptual change. Conceptual change is a really complex cognitive process, and it may involve a hierarchical collection of many different submechanisms. Some of those are better "known" (categorization, inductive reasoning), some of those aren't (mapping mechanisms). In addition, there are many different forms of conceptual change, and they may involve several different mechanisms.

However, as also Ohlsson emphasizes (Ohlsson, 2009b), a theory of conceptual change cannot just be the list of all possible mechanisms, but it must also make some constraints for the list of explanatory relevant mechanisms. A theory of conceptual change should be able to tell, why certain mechanisms are required or appropriate for conceptual change, and why some other aren't. For this reason, the task level also matters. The task level identifies the learning episode as conceptual change by identifying the relevant type of difference between initial state (no concept) and outcome (has concept) is an essential part of explanation because it provides not only the characterization for the explanandum of explanation, but it is also needed to evaluate the explanatory relevance of mechanisms.

\section*{Acknowledgments}

The participants of ConChaMo-workshops. This work was funded by Finnish Academy (project number 1133369).

\footnotetext{
\({ }^{3}\) see Piccinini, 2006, for an analysis of relevance in the context of computational explanations.
}

\section*{References}

Bechtel, W. \& Richardson, R. (1993). Discovering Complexity, Decomposition and Localization as Strategies in Scientific Research. New Jersey: Princeton University Press.
Carey, S. (1985). Conceptual Change in Childhood. Cambridge, MA: MIT Press.
Carey, S. (2000). Science education as conceptual change. Journal of Applied Developmental Psychology, 21(1), 13-19.
Carey, S. (2011). Précis of the Origin of Concepts. Behavioral and Brain Sciences, 34, 113-167.
Chi, M. T. H. (1992). Conceptual change within and across ontological categories: Examples from learning and discovery in science. In R. Giere (Ed.), Cognitive Models of Science: Minnesota Studies in the Philosophy of Science, (pp. 129-186). University of Minnesota Press: Minneapolis, MN.
Chi, M.T.H. (2008). Three types of conceptual change: Belief revision, mental model transformation, and categorical shift. In S. Vosniadou (Ed.), Handbook of research on conceptual change (pp. 61-82).Hillsdale, NJ: Erlbaum.
Chi, M. T. H., Slotta, J. D. \& de Leeuw, N. (1994). From things to processes: A theory of conceptual change for learning science concepts. Learning and Instruction, 4: 27-43.
Chinn, C. A., \& Samarapungavan, A. (2009). Conceptual Change - Multiple Routes, Multiple Mechanisms: A Commentary on Ohlsson. Educational Psychologist, 44(1), 48-57.
Clement, J. (2008) The role of explanatory models in teaching for conceptual change. I S. Vosniadou (Ed.), International Handbook of Research in Conceptual Change (pp. 417-452). New York: Routledge.
Craver, C.F. (2006). When mechanistic models explain. Synthese. 153: 355-376.
Craver, C.F. (2007). Explaining the Brain: What a Science of the Mind-Brain could be. New York: Oxford University Press.
diSessa, A. A. (1988). Knowledge In Pieces. In G. Forman \& P. Pufall (Eds.), Constructivism in the Computer Age (pp. 49-70). Hillsdale, NJ: Erlbaum.
diSessa, A. A. (1993). Toward an Epistemology of Physics. Cognition and Instruction, 10, (2 \& 3), 105-225.
diSessa, A. A. (2008). A Bird's Eye View of the "Pieces" vs. "Coherence" Controversy (From the "Pieces" Side of the Fence). In S. Vosniadou (Ed.), International Handbook of Research on Conceptual Change (pp. 3560). New York: Routledge.
diSessa, A. A. \& Sherin, B. (1998). What changes in conceptual change? International Journal of Science Education, 20(10), 1155-1191.
Machamer, P., Darden, L. and Craver, C.F. 2000. Thinking about Mechanisms, Philosophy of Science. 67(1): 1-25.

Markman, A. (2011). Can developmental Psychology provide a blueprint for the study of adult cognition. Behavioral and Brain Sciences, 34, 140-141.
Mayer, R.E. (2002). Understanding Conceptual Change: A Commentary. In M. Límon \& L. Mason (eds.), Reconsidering Conceptual Change. Issues in Theory and Practice, Kluwer Academic Publishers, Netherlands, 101111.

Lappi, O. \& Rusanen, A-M. (2011). Turing machines and causal mechanisms in cognitive science InMcKay Illari, P., Russo, F. \& Williamson, J. (eds.), Causality in the Sciences. (pp. 224-239). Oxford: Oxford University Press,
Ohlsson, S. (2009a). Resubsumption: A possible mechanism for Conceptual Change and Belief Revision. Educational Psychologists, 44(1), 20-40.
Ohlsson, S. (2009b).Meaning Change, Multiple Routes, and the Role of Differentiation in Conceptual Change: Alternatives to Resubsumption? Educational Psychologists, 44(1), 64-71.
Posner, G. J., Strike, K. A., Hewson, P. W., \& Gertzog, W. A. (1982). Accommodation of a scientific conception: Towards a theory of conceptual change. Science Education, 66 (2), 211-227.
Piccinini, G. (2006). Computational Explanation and Mechanistic Explanation of Mind. In M. DeCaro, F. Ferretti \& M. Marraffa (Eds.), Cartographies of the Mind: The Interface Between Philosophy and Cognitive Science. Dordrecht: Kluwer.
Shagrir, O. (2001). Content, Computation and Externalism. Mind 110, 369-400.
Rozenblit, L., \& Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. Cognitive Science, 26, 521-562.
Rusanen, A-M. \& Pöyhönen, S. (2012). Concepts in Change. Science \& Education, 2012. (On line first). DOI: 10.1007/s11191-012-9489-x

Strike, K.A., \& Posner, G.J. (1992). A revisionist theory of conceptual change. In R.Duschl \& R. Hamilton (eds.), Philosophy of science, cognitive psychology, and educational theory and practice (pp. 147-176). Albany, NY: SUNY Press.
Thagard, P. (1990). Concepts and conceptual change. Synthese, 82: 255-274.
Thagard, P., \& Verbeurgt, K. (1998). Coherence as Constraint Satisfaction. Cognitive Science, 22 (1), 1-24.
Vosniadou, S. (1992). Knowledge acquisition and conceptual change. Applied Psychology: An International Review, 41(4), 347-357.
Vosniadou, S. (1999). Conceptual Change Research: State of the Art and Future Directions. In W. Schnotz, S. Vosniadou, \& M. Carretero (Eds.) New Perspectives on Conceptual Change, Elsevier Science, 3-13.
Vosniadou, S., \& Brewer, W. F. (1992). Mental Models of the Earth: A Study of Conceptual Change in Childhood. Cognitive Psychology, 24, 535-585.
Woodward, J. (2003). Making Things Happen: a Theory of Causal Explanation, Oxford, 2003.

\title{
A PDP Model for Capturing N400 Effects in Early L2 Learners during Bilingual Word Reading Tasks
}

\author{
Sepideh Sadeghi (sepideh.sadeghi @tufts.edu), Matthias Scheutz (matthias.scheutz @tufts.edu) \\ Department of Computer Science, Tufts University Medford, MA 02155 USA \\ He Pu (he.pu@tufts.edu), Phillip J. Holcomb (pholcomb@tufts.edu), Katherine J. Midgley \\ (kj.midgley@tufts.edu) \\ Department of Psychology, Tufts University \\ Medford, MA 02155 USA
}

\begin{abstract}
Parallel Distributed Processing (PDP) models have been widely used for modeling cognitive tasks where accuracy or reaction time were the dependent performance measures. However, only few PDP models have attempted to model more brain-related data like event related potentials (ERPs). In this paper, we take a step towards using ERP data for model fitting by proposing a PDP model, which can successfully replicate various known ERP effects. Specifically, we introduce a PDP-equivalent of the N400 ERP measure and apply it to a simple PDP model of early bilingual word acquisition as bilingual word acquisition tasks provide several well-established N400 effects that can be used for model validation. We then analyze the dynamics of the network to show why and how the network can capture each of the targeted N400 effects. Furthermore, we qualitatively compare model-generated and empirical N400 peak values for L2 words.
\end{abstract}

Keywords: PDP, bilingualism, L2 word acquisition, event related potential (ERP), N400

\section*{Introduction}

In a recent paper, Laszlo \& Plaut (2012) proposed a way to capture N400 ERP word reading data in a parallel distributed processing (PDP) connectionist network whose architecture was based on two neurally plausible characteristics: neurons can either be excitatory or inhibitory, but not both, and inhibitory connections can only occur within layers, but not between (as the range of inhibitory connections in the brain is shorter than that of excitatory connections). Given these two constraints, the model generated cycle-based time-course data that reflected the temporal evolution of the N400 response, replicating the "orthographic neighborhood size" effect that words with larger orthographic neighborhood size elicit larger N400s compared to words with smaller neighborhood size. However, it is currently unclear whether this model would also capture various other known N400 word effects such as those obtained in the context of bilingual word processing.

In this paper, we propose a PDP architecture for a PDP model of bilingual word processing, which can successfully capture several known N400 effects in early bilingual word processing, including the "orthographic neighborhood size" effect in addition to other known effects such as the "pseudoword effect".

\section*{Background}

Two important aspects of any bilingual processing model are the representations of lexical items in the first (L1) and second (L2) language and their requisite connections to concepts. Research on word processing during the early stages of L2 acquisition has revealed important constraints about storage and processing of conceptual and lexical information in the bilingual brain. Studies using speeded translation tasks, for example, show L2 learners are faster to translate from L2 to L1 (e.g., translating tenedor to fork in native English learners of Spanish) than from L1 to L2 (translating fork to tenedor) (e.g., Kroll \&Stewart, 1994). These behavioral results indicate that adult bilinguals appear to associate new L2 words with their L1 translation equivalents in order to facilitate semantic access to these new words.

This bootstrapping of L2 into the already established L1 language system involves an asymmetrical representation of the two languages, accounted for in Kroll \& Stewart's Revised Hierarchical Model (RHM) (depicted in Figure 1).


Figure 1: The RHM (Kroll \& Stewart, 1994). Solid lines indicate strong connections and dashed lines indicate weak connections.

The RHM assumes a separate lexicon for L1 and L2 with orthographic and phonological representations, each of which is connected to a single amodal conceptual store. In early second language learners, the L1 lexicon is assumed to be much larger than the L2 lexicon and evidence from picture naming tasks in bilinguals suggests that the strength of the links between the two lexicons and the conceptual store are also asymmetrical, with L1 having stronger connections to semantics than does L2 (e.g., Kroll \& Peck, 1998). Both the lexical level asymmetry and the concept-tolexicon asymmetry between L1 and L2 are modeled in the RHM by disproportionately weighted links (see Figure 1). Adhering to the behavioral data, the link from the L2 lexicon to the L1 lexicon is much stronger than the link from L1 to L2, just as the link between the L1 lexicon and conceptual store is much stronger than the link between the L2 lexicon and the conceptual store.

However, behavioral data is often insufficient for distinguishing between different processing mechanisms. Hence, electrophysiological measures such as event-related potentials (ERPs) with their fine-grained temporal resolution can uncover particular neural activity elicited during language tasks that might only be associated with a particular class of model architectures. In particular, the N400, which is a negative-going centroparietally distributed ERP component peaking around 400 ms after stimulus onset, has been shown to index lexico-semantic integration during word processing. Hence, it provides a robust measure of changes in processing activity in the brain as language learning takes place and can thus be used to flesh out conceptual proposals like the RHM in computational architectures such as the PDP connectionist models. We will, in particular, focus on four aspects of monolingual and bilingual word processing for which N400 effects have been reported in the literature: (A1) L1/L2 words versus L1/L2 pseudowords (i.e., pronounceable L1/L2 non-words that adhere to the orthographic rules of L1/L2); (A2) L1/L2 word repetition effects; (A3) variations in L1/L2 word neighborhood size; and (A4) L1 vs. early L2 word processing differences.

Regarding (A1), it is well-known that L1 pseudowords elicit larger N400s than L1 words (e.g., Holcomb \& Neville, 1990). Moreover, L2 learners showed larger N400s to L2 pseudowords than to L2 words after only 14 hours of classroom learning, mimicking L1 pseudoword effects (e.g., McLaughlin et al., 2004; however, note that McLaughlin and colleagues did not find any behavioral evidence of L2 words and pseudoword discrimination, thus supporting the use of ERPs over behavioral measures for adjudicating model architectures).

Regarding (A2), repeated words reliably elicit smaller N400 amplitudes than their first presentation (e.g., Rugg, 1985). This attenuation of the N400 reflects the increased ease of lexico-semantic integration upon the second and subsequent presentations of a word (possibly due to residual activation of the lexical item and/or facilitatory feedback from the activated concept).

Regarding (A3), words with large numbers of orthographic neighbors (e.g., words that differ from the target by only one letter) elicit larger N400s than words with smaller neighborhood size (e.g., Holcomb et al., 2002). Notably, the effect occurs within as well as across languages, i.e., L1 influencing L2 and vice versa (Midgley et al., 2008).

And finally (A4), N400s can be used as a measure of how closely L2 processing matches that of L1 processing. For example, Midgley and colleagues found that both EnglishFrench and French-English bilinguals who had intermediate L2 experience displayed smaller N400s to L2 words than to L1 words (2009). Balanced bilinguals did not show any N400 differences between L1 and L2 word processing. This result might be in part explained by (A3). Given that the L1 lexicon contains more word forms than the L2 lexicon, L1 words generally have larger neighborhood sizes than L2 words. The larger neighborhood sizes of L1 items in comparison to L2 items may contribute to larger N400 amplitudes for L1 words over L2 words.

\section*{Model Description}

We start with four hypotheses, (H1) through (H4), about the possible principles responsible for each corresponding N400 effect (i.e., (A1), (A2), (A3), and (A4)) in the context of a RHM-like PDP architecture and then add connections within and between layers of the network based on the hypothesized mechanisms.

\section*{Hypotheses:}
(H1) Pseudowords have no word-level representations and thus no connections to concept nodes or nodes within the lexical layer.
(H2) Concept nodes keep a residual activation between repeated word presentations and can thus be activated faster in subsequent presentations of the same word compared to the first presentation.
(H3) Lexical inputs with more orthographic neighbors should activate more concepts early on. This should lead to increased competition among concepts and thus to reduced overall activations later on, which can be facilitated via inhibitory connections in the concept layer.
(H4) After some training (when fairly strong, direct L2 lexical-to-concept connections are in place), L2 words should elicit a larger initial target concept activation than L1 words. This can be accomplished via L2-to-L1 word connections that are stronger than those from L1-to-L2 words.

Based on the RHM framework, we developed a PDP model with bidirectional excitatory lexical-to-concept connections, top-down inhibitory concept-to-lexical connections and inhibitory concept-to-concept connections (see Figure 2). As in the (Laszlo \& Plaut, 2012) model, we
use IAC units (with standard parameter values for \(\min =-.2\), max \(=1\), and rest \(=-.1\) activation levels as well as decay rate \(=.1\) ). For simplicity, we limit input words to 5 letters, thus requiring 5 clusters of 26 input letters per word (for the English alphabet). All letters in each cluster \(i\) have excitatory connections to words that contain them in the \(i\)-th slot and inhibitory connections to all words with a different letter in the \(i\)-th position.


Figure 2: Model architecture. The thickness of links indicates the strength of connections.

L1 versus L2. To account for larger L1 vs. L2 word neighborhoods, we include more L1 words with a larger neighborhood size in the model compared to L2 words.

Pseudowords versus words. Pseudowords have no representation at the lexical or semantic layer.

Repetition. We model repetition effects by performing the following sequence \(r\) times: input word \(i\) is presented for \(n\) cycles (where \(n\) should be large enough to allow the N400 signal, to be defined below, to reach its peak). Then the input is removed and the network is updated for \(d\) cycles to let all node activations decay, after which point the whole process is repeated, but without resetting any activation values. We thus have three critical modeling parameters that need to be set appropriately: \(r, n\), and \(d\).

Filtering word length artifacts. Assuming that each constituent letter contributes equally to a word's activation level, all connection weights from each letter in a word have the same strength. However, because words have different lengths, the overall incoming activation would be different if we were to use the same connection weights for all letter-to-word connections as longer words would get a higher activation than shorter words, everything else being equal. To avoid this effect, we scale the letter-to- word connection weight \(c\) by the length \(|W|\) of the word \(W: w_{L, W}=c /|W|\). We also needed to make sure that the input letters
corresponding to a given target lexical item will only activate the orthographic neighbors and not the other words that differ from the target word in more than one letter. In order to do so, we made the strength of inhibitory and excitatory letter-to-word connections the same, so that if a word is different from the target word in more than one letter (for four-letter words), it receives zero or less than zero netinput from the letter nodes. In addition, none of the five-letter/three-letter words were similar to a four-letter word in 3 or more slots.

\section*{Definition of PDP N400 Measure}

Based on the semantic interpretation of the N400 signal (Laszlo \& Fedemeier, 2011), we define the networkequivalent of the N 400 as the magnitude of overall activation change (differential) in positively activated (potential) concept nodes (potential). Specifically, we calculate the sum of all positive concept activations at each cycle and compute the change between two consecutive cycles as the N400 (the discrete equivalent to the derivative of the potential given by the summed concept node activations).

\section*{Experimental Bilingual ERP Data}

We collected ERP measures from 14 native English speakers who were enrolled in a first semester "Introductory Spanish" class at Tufts University ( 9 females, mean age 18.4). Participants viewed Spanish words (e.g., HOLA, GATO) and Spanish pseudowords (e.g., SERO, AGOL) one at a time as part of a lexical decision task. The Spanish words were a set of non-cognates taken from the textbook used in class. Factors of length, English bi-gram frequency, and English neighborhood size were balanced between the words and pseudowords used in the study. Averaged ERPs were computed for all word and pseudoword stimuli for each participant at 29 scalp sites. Single item ERPs were formed by averaging to time-locked stimuli across participants. The mean amplitude averaged across a subset of centroparietal electrodes (including: Cz, Pz, C3, CP5, CP1, P3, C4, CP6, CP2, P4) between \(300-500 \mathrm{~ms}\) was used to quantify the N400 effect. The mean amplitude between \(300-500 \mathrm{~ms}\) was used to quantify the N 400 effect. Additionally, N400 measures for single items were calculated using the mean amplitude between item-specific temporal windows, ranging from 250 ms to 500 ms .

\section*{Modeling Results}

We selected a subset of 14 four-letter L2 words from all L2 words used in the ERP experiment and included all their L1 translations as well as their L1 neighbors to be able to account for the cross-language orthographic neighborhood size effects. Since some of the L1 words were 5 letters in length, we included 5 clusters of letters in the model.


Figure 3: Sum of semantic activation (top row) and the N400 amplitude (bottom row), over 75 update cycles in response to three words: "son" (L1 word in black), "azul" (L2 word in blue), and "sero" (L2 pseudoword in red).

Figure 3 shows the shape of the N400 signal generated by the model along with the time-course of the summed concept nodes' activations during the whole word exposure. Note that the change in total concept activation is proportional to the maximum value of the N 400 generated.

The right column in Figure 3 reveals three distinct phases in the dynamic of the overall semantic activation in our network: (a) charge (positive overall change), (b) discharge (negative overall change), and (c) stabilization. Furthermore, since inhibitory connections only originate from concept nodes, any significant flow of inhibition can only come after an initial flow of activation, i.e., until concept nodes have reached sufficiently strong activations.

Charge. The activation of the target concept and concepts associated with orthographic neighbors or its associated word-level node initially start to increase, followed by the feedback from excitatory and inhibitory connections to word-level nodes causing the activation of the target word to
gradually increase and the activations of its orthographic neighbors to decrease.

Discharge. The overall semantic activation decreases as a result of inhibition exerted by significantly activated concept nodes.

Stabilization. Eventually, the overall activation levels of the network stabilize.

We searched for values for the various connections that would allow the model to capture the N400 effects: concept-to-L1 \(=(.6,-2), \quad\) concept-to-L2 \(=(.8,-.2), \quad\) concept-toconcept \(=(0,-.6)\), L1-to-concept \(=1\), L2-to-concept \(=.8\), L1-to\(\mathrm{L} 2=.1\), L2-to-L1=1, letter-to-(3letterWord) \(=(.8,-.8)\), letter-to-(4letterWord) \(=(.6,-.6)\), letter-to-(5letterWord) \(=(.48,-.48)\) (the first element of each tuple is the excitatory weight value between related items, and the second element is the inhibitory weight value between the unrelated items).

For all simulations, we took the maximum peak value as the measure for comparing N400 signals to the empirical data. Furthermore, since several factors can influence the N400 value, we investigated only one factor at a time while keeping the others fixed.


Figure 4: N400 data for repetitions of "son" using first: \(r=3, n=30, d=30\), second: \(r=1, n=30, d=70\), and then \(n=30\) (see text for details).

Figure 4 shows that the model replicates the repetition effect (A2), i.e., maximum N400 values (peaks) after the first exposure are all smaller than the first peak.

Figure 5 shows that the model is able to replicate the neighborhood size effect regardless of lexical type: L1, L2, and pseudowords.

Figure 6 shows the replication of (A4) - in all cases - and the replication of (A1) - in all cases except for (a) and (b). Furthermore, Figure 4 suggests that the replication of (A1) and (A4) is dependent on neighborhood size: as the neighborhood size increases, the network replicates (A1) more strongly, while showing weaker replication of (A4). The network best replicates (A4) for L2 words of \(n\) Size \(=0\).




Figure 5: Neighborhood size effects within three categories: in order L1, L2, Pseudowords, shown by mean N400s of words with \(n\) orthographic neighbors: \(0=\) black, \(1=\) blue, \(2=\) green, \(3=\) cyan, \(4=\) red, \(5=\) yellow, \(6=\) magenta, \& \(10=\) black stars.


Figure 6: Mean N400 signals for words sharing the same neighborhood size ( \(n\) Size): a) \(n \operatorname{Size}=0\), b) \(n \operatorname{Size}=1\), c) \(n S i z e=2\), d) \(n S i z e=3\), e) \(n S i z e=5\), f) \(n S i z e=6\), in three categories: L1 words in black, L2 words in blue, and L2 pseudowords in red. Note that there was no L2 word of \(n S i z e=1\), no pseudoword of \(n S i z e=2\), and no L1 word of \(n\) Size \(=3\).

Note that the correlation value (corr=.2135) between the maximum N400 values (for L2 words) generated by the model and those collected in the experiments shows that the model does not yet quantitatively fit the empirically obtained ERP values, despite qualitatively replicating ERP effects.

\section*{Discussion}

The model succeeded in capturing qualitatively all four ERP effects. Furthermore, the results confirm that the (A1) and (A4) effects are dependent on neighborhood size as suggested in (Midgley et al, 2008 and Holcomb \& Neville, 1990). However, the model allows for a different explanation from that of Midgley et al. who hypothesized that the overall lower N400 for L2 words compared to L1 words might be caused by the smaller neighborhood size of L2 words compared to L1 words, everything else being equal. Specifically, the model shows that this difference can also be obtained with identical neighborhood sizes based on the generally higher initial activation induced at the target concept in response to an L2 input word (compared to that of an L1 input word). This higher initial activation tends to suppress the other concept nodes, thus leading to an overall lower ERP and thus lower N400. Hence, it is likely that both neighborhood size and difference of initial target concept activation via L1 or L2 words contribute to the smaller N400 for L2 words (compared to L1 words).

Note that all simulation results where obtained by considering N400 peak values only, but other measures are certainly possible (e.g., the integral of the N400 signal over the \(300-500 \mathrm{msec}\) time frame or the average value over the same period). This is left for future work.

\section*{Conclusion}

We have developed a PDP model based on Kroll \& Stewart's Revised Hierarchical Model (RHM) of bilingual word processing and tested it against well-established N400 effects. The model succeeded in qualitatively replicating language, neighborhood size, pseudoword, and repetition effects. However, the model did not quite replicate the N400 results from our empirical experiments, as shown by a fairly low correlation between the ERPs of the model and empirical data. Future work will focus on exploring the model's parameter space to determine if better model fits are possible with the given model architecture. In addition, we will investigate simpler model architectures and the extent to which they may succeed in replicating some of the N400 effects. We will also investigate alternative definitions of N400 (e.g., including the lexical level activations) as well as exploring the use of average N 400 amplitudes rather than peak values.

\section*{References}

Holcomb, P. J., Grainger, J., O'Rourke, T. (2002). An electrophysiological study of the effects of orthographic
neighborhood size on printed word perception. Journal of Cognitive Neuroscience, 14, 938-950.
Holcomb, P. J \& Neville, H. J. (1990). Auditory and visual semantic priming in lexical decision: A comparison using event-related brain potentials. Language and Cognitive Processes, 5, 281-312.
Kroll, J. F., \& Stewart, E. (1994). Category interferences in translation and picture naming: Evidence for asymmetric connection between bilingual memory representation. Journal of Memory and Language, 33, 149-174.
Kroll, J. F., \& Peck, A. (1998). Competing activation across a bilingual's two languages: Evidence from picture naming. Proceedings of the 43rd Annual Meeting of the International Linguistic Association. New York University, NY.
Laszlo, S., \& Plaut, D.C. (2011). Simulating event-related potential reading data in a neurally plausible parallel distributed processing model. Proceedings of the 33rd Annual Conference of the Cognitive Science Society. Mahwah, NJ: Lawrence Erlbaum Associates.
Laszlo, S., \& Plaut, D.C. (2012). A neurally plausible parallel distributed processing model of event-related potential reading data. Brain and Language, 120, 271281.

Laszlo, S., \& Federmeier, K.D (2011). The N400 as a snapshot of interactive processing: evidence from regression analyses of orthographic neighbor and lexical associate effects. Psychophysiology, 48, 176-186.
McLaughlin, J., Osterhout, L., \& Kim, A. (2004). Neural correlates of second-language word learning: Minimal instruction produces rapid change. Nature Neuroscience, 7, 703-704.
Midgley, K.J., Holcomb P.J., van Heuven, W. J. B., \& Grainger, J. (2008). An Electrophysiological investigation of cross-language effects of orthographic neighborhood. Brain Research, 1246, 123-135.
Midgley, K. J., Holcomb, P. J., \& Grainger, J. (2009). Language effects in second language learners and proficient bilinguals investigated with event-related potentials. Journal of Neurolinguistics, 22, 281-300.
Rugg, M.D. (1985). The effects of semantic priming and word repetition on event-related potentials. Psychophysiology, 22, 642-647.
Rugg, M.D. (1990). Event-related brain potentials dissociate repetition effects on high- and low-frequency words. Memory \& Cognition, 18, 367-379.

\title{
The Time Course of Language Use in Multiparty Negotiations
}

\author{
Eyal Sagi (eyal@u.northwestern.edu) \\ Kellogg School of Management, Northwestern University \\ Evanston, IL 60208 USA
}

Daniel Diermeier (d-diermeier@kellogg.northwestern.edu)
Kellogg School of Management, Northwestern University
Evanston, IL 60208 USA

\begin{abstract}
Language use is an important part of a negotiation. Prior research has shown that similarity in language use is conducive to reaching agreements. This paper uses Latent Semantic Analysis to explore how the similarity of language use develops and changes over the time course of a threeparty negotiation. Results support theories that suggest that a gradual alignment of semantic representation increases the likelihood that parties will form a coalition.
\end{abstract}

Keywords: Negotiation, Coalition Formation, Linguistic Entrainment, Psycholinguistics, Latent Semantic Analysis.

\section*{Introduction}

The language used by parties in a negotiation is a crucial aspect of their negotiation strategy and can greatly affect the outcome of the negotiation. Research over the past few decades has demonstrated that an agreement in a negotiation is often preceded by convergence in language use among the negotiating parties (cf. Miller, 2005). In the context of multi-party negotiation, such convergence is evident between the parties that reach an agreement, but not the excluded parties (Huffaker, Swaab, and Diermeier, 2011).

In this paper we extend the results of Huffaker et al. by analyzing the similarity of adjacent conversation moves. This type of analysis enables us to look at the time course of coalition formation and not just at the overall similarity of language use between participants.

Multi-party negotiation is naturally more complex than that a two-party negotiation (Bazerman et al., 2000). This is especially true since an agreement can be reached among a subset of the negotiating parties. Therefore being excluded from an agreement is a real possibility. Nevertheless, partial coalition agreements are often less desirable than agreements that involve the group as a whole because they are less efficient or do not use all of the available resources. However, even being part of a partial agreement is more preferable than no agreement at all or of being excluded from an agreement reached by others.

The added complexity of multi-party negotiation has been shown to affect the patterns of language use in such negotiations. Following the framework of Communication Accommodation Theory (CAT; cf. Giles et al., 2007), Huffaker et al. (2011) demonstrates how the formation of coalition is affected by specific aspects of language use by the negotiating parties. Specifically, they find that partners to a coalition show more similarity in language use than
participants who were not part of a coalition. The use of assents was also found to correlate positively with being a part of the coalition agreement. In contrast, the use of negative emotion words was a detrimental predictor to being a part of a coalition.

These results are congruent with empirical findings in psycholinguistics that show that in successful dialogues the representations and language used by dialogue partners tend to converge over time (e.g., Brennan and Clark, 1996). However, while Huffaker et al. show that similarity in language use is a factor in the outcome of the negotiation, they use the entire negotiation as the unit of analysis. Consequently, their results do not explore the time course of this similarity. This paper aims to extend their results by looking at whether the language used by the participants changes over time.

On the one hand, theories of entrainment and alignment in language use by dialogue participants, such as that put forth by Pickering and Garrod (2004), argue that language similarity is the result of gradual alignment of language use by the participants in the negotiation. The better the alignment the more likely the aligned parties would be to form a coalition. Such theories would lead to the prediction that language similarity should increase over the course of the negotiation and ultimately result in the parties forming a coalition.

In contrast, Swaab et al. (2011) shows that language mimicry during a negotiation is a factor in the outcome of the negotiation, but only when it occurred early in the negotiation. Therefore, we might expect that early similarity in language use might lead to the forming of a coalition later on.

It is also possible that both of these factors contribute to the effect that similarity in language use has on the resulting coalition. If that is the case we would expect to find not only that eventual coalition partners show more similar language early on than non-coalition partners, but that this difference increases over time.

\section*{Measuring similarity in language use}

Convergence of language use in a dialogue or negotiation has traditionally been measured by hand coding the transcripts of negotiation dialogue. Such hand coding is time consuming and, to a certain degree, requires the coders to interpret the language used by the negotiators. In contrast, the metric we use in this paper is automatically derived by
using the Latent Semantic Analysis cosine similarity of a pair of utterances. Such a measure has been used in the past as a measure of textual coherence (Foltz, Kintsch, and Landauer, 1998) and as a measure of linguistic entrainment (Huffaker et al., 2006).

LSA vectors for individual words are generated based on the co-occurrence patterns of words in large corpora. These vectors identify points in a high-dimensional space (100 dimensions in this case). The more likely two words are to co-occur with similar words the closer they will be in the space. For example, the vectors for sun and moon are fairly close together and show a cosine similarity of .53 whereas man and moon are not very similar and show a cosine similarity of .03 . Moreover, when several word vectors from a single utterance are combined together, as was done in this study, the result identifies a point in space that represents the overall topic of the utterance.

It is important to note that this kind of automatic measure ignores certain linguistic elements that a coder might use. For instance, the use of negation is generally ignored, while sarcasm and metaphors are often misrepresented. However, since we are interested in the convergence of language use that is, whether participants are using similar language to convey their (sometimes opposing) ideas, this type of analysis seems appropriate.

\section*{Method}

\section*{Huffaker et al. (2011)}

The data used in this paper comes from a study reported by Huffaker et al. (2011). They patterned their study after a pure coalition game outlined by Raiffa (1982). In that study, 180 MBA students were divided into 60 three-person groups. Within each group, participants were assigned to one of three roles ( \(\mathrm{A}, \mathrm{B}, \mathrm{C}\) ) and instructed that they were to use an online chat room to negotiate a split of that payoff amongst themselves. Participants were unaware of the identities of the other participants in the negotiation.

Table 1: Payoff Table in the Negotiation Game from Huffaker et al. (2011)
\begin{tabular}{lr}
\hline Possible Agreements & Total Payoff \\
\hline A alone & \(\$ 0\) \\
B alone & \(\$ 0\) \\
C alone & \(\$ 0\) \\
A and B & \(\$ 118,000\) \\
A and C & \(\$ 84,000\) \\
B and C & \(\$ 50,000\) \\
A, B, and C & \(\$ 121,000\) \\
\hline
\end{tabular}

Note: A, B, and C represent the participants in the negotiation. The payoff is split between the parties that reach the described final agreement.

All participants were provided with the payoff table in advance of the negotiation (see Table 1). As is evident from the table, different coalition formations receive different
payoffs, and if no coalition is formed no participant receives any payoff. The participants were allowed to negotiate how the payoff is distributed between them. These payoff options provide incentives for the participants to join up with another participant so that they can take advantage of the resulting weak bargaining position of the third participant. However, the payoff table is designed so that the third player can always make an attractive offer to one of the members of the initial coalition to induce a defection from the preliminary agreement. Consequently, participants are incentivized not only to be a part of a forming coalition, but also to ensure that it is a stable coalition and that their partner(s) will not defect.

Participants in the experiment were placed at computers in different rooms so that their only means of communication with each other was through the provided chat software. They logged into a public chat room to begin the negotiation process.

The software also allowed participants to move from the public chat room to three private chat rooms. That is, participant A could move into one of the private chat rooms together with participant B so that they could negotiate without participant C being privy to the content of the negotiation. However, all participants were alerted whenever a participant entered or exited a chat room so that the excluded participant was always aware that the two other participants might be negotiating in private. This mimics some of the real-world aspects of a negotiation, where parties are often able to communicate in private, but the fact that they communicated in private is common knowledge. A private exchange of information can also provide an indication that the two parties are forming a coalition.

\section*{Semantic Analysis}

The analysis in this paper is based on the transcripts of these negotiations. An LSA vector was computed for each individual utterance by using vector addition to combine the vectors of all of the content-bearing words in the utterance. When an utterance did not include any content-bearing words, a null vector was used to represent it. The vector space used for this analysis was generated by Infomap (http://infomap-nlp.sourceforge.net/ ; Schütze, 1997) using the written part of the British National Corpus.

Next, the correlation of the vectors representing temporally adjacent utterances was computed \({ }^{1}\). These correlations were not computed when one of the utterances had a null vector or when the two utterances did not occur in the same chat context (i.e., when they occurred in different chat rooms).

In some cases identical vectors represented adjacent entries. These were generally the result of statements such as "I agree" or "X is present" and were found either at the

\footnotetext{
\({ }^{1}\) Because the first dimension of LSA vector spaces tends to correlate with the frequency and length of the text it was dropped from the analysis (cf. Hu et al., 2007)
}
very beginning or the very end of the negotiation. They were dropped from the analysis because they did not appear to represent a meaningful part of the negotiation. That is, they did not represent linguistic convergence between participants but rather formulaic utterances that occurred mostly before the negotiation started or after it has concluded. Regardless, the results presented here are quantitatively and qualitatively similar whether these data points are included or excluded.

In order to test for convergence in language use we categorized the utterance pairs based on the two participants that contributed to them. We predicted that participants who were included in the resulting agreement would have more similar language use than those who included a participant who were excluded from the agreement. For example, if an AB agreement was reached, utterance pairs between A and \(B\) would be predicted to have more similar language use (i.e., utterance-to-utterance correlation) than those between A and C or B and C . Consequently, we divided the utterance pairs to those in which both participants were included in the final coalition (successful utterances) and those in which at one of the participants was excluded from the coalition (unsuccessful utterances). Importantly, when the final agreement included all parties, all of the utterance pairs were considered to be successful. In contrast, when no agreement was reached all of the utterances were considered to be unsuccessful.

Because the parties are unfamiliar to each other when they enter the negotiation we also expected that this difference would emerge over the course of the negotiation and become apparent only once a coalition begins to form. Therefore, we divided the utterance pairs based on their position in the negotiation \({ }^{2}\) - If the first utterance of the pair occurred in the initial half of the negotiation it was classified as an early utterance whereas utterances that occurred in the second half of the negotiation were considered late utterances.

\section*{Results}

As mentioned above, we derived two distinct hypotheses:
1. Following accounts of linguistic entrainment (e.g., Pickering and Garrod, 2004), we hypothesized that coalition formation will be accompanied by the alignment of language use. Consequently, if linguistic entrainment occurs as part of the formation of a coalition, successful utterances should become more similar to their responses than unsuccessful

\footnotetext{
\({ }^{2}\) For the purposes of this analysis we use a coarse grain division of time (halves) because some of the discussions consist of relatively few utterances (under 50). This is adequate for the purpose of our basic hypotheses, but does not provide a good sense of how the utterance-to-utterance similarity changes over time. It is possible to utilize smaller time units in an analysis of this type to gain further insight into the temporal progression of the negotiation (e.g. Figure 2). Essentially, the choice of temporal units for analysis represents a tradeoff between precision and statistical power.
}
utterances would be to their responses late in the negotiation. This will result in a significant interaction between the type of utterance and its position in the negotiation.
2. Following the literature on the effectiveness of mimicry in negotiations (e.g., Swaab, et al. 2011), we hypothesized that early mimicry would result in a higher likelihood of eventual success in the negotiation. Therefore, if mimicry is an effective tool in these negotiations, successful utterances should be more similar to their responses than unsuccessful ones early in the negotiation.

To test these hypotheses we conducted a \(2 \times 2\) ANOVA. The type of utterance (successful vs. unsuccessful) and its position in the negotiation (early utterance vs. late utterance) were the independent variables. The dependent measure was the average utterance-to-utterance correlation for utterance pairs conforming to the condition within a particular session. The means for each of the conditions can be seen in Figure 1.

There was no significant difference in similarity of language use between early utterances ( \(M=0.15, S D=0.08\) ) and late utterances \((M=0.14, S D=0.09)(F(1,186)<1\), n.s. \()\). There was a slight trend where successful utterances ( \(M=0.15, S D=0.08\) ) showed more similar language use than unsuccessful utterances \(\quad(M=0.13, \quad S D=0.09) \quad(F(1\), 186) \(=2.86\), \(\mathrm{MSE}=0.01, p=0.093\) ).


Figure 1: Similarity of Language use by utterance pair type and position in the negotiation. Error bars represent standard error.

More importantly, the interaction between utterance pair type and position was significant - The difference in language use between successful and unsuccessful utterances was greater for late utterances than early utterances \(\quad(F(1, \quad 186)=7.15, \quad M S E=0.01, \quad p<.01)\). Furthermore, Tukey HSD tests identified a significant


Figure 2: Mean similarity of language use by utterance pair type over the course of the negotiations. Each time period corresponds to \(10 \%\) of the negotiation (calculated individually for each negotiation). Error bars represent standard error.
difference between late successful utterances ( \(M=0.16\), \(S D=0.08\) ) and late unsuccessful utterances ( \(M=0.11\), \(S D=0.09\) ) ( \(D=0.051, ~ p<.01\) ) but not between early successful utterances \((M=0.14, S D=0.07)\) and early unsuccessful utterances ( \(M=0.15, S D=0.10\) ) ( \(D=-0.01\), n.s.). This result provides support for accounts in which a gradual alignment in language use and semantic representation leads to a likelihood of forming a coalition (Hypothesis 1). However, we found no support for accounts in which early similarity in language use (e.g., mimicry) leads to the formation of a coalition (Hypothesis 2).

Interestingly, while there appears to be a slight increase in the utterance-to-utterance similarity of successful utterance pairs from the first half to the second half of the negotiation, the observed interaction seems to be driven more by an unexpected decrease in the utterance-to-utterance similarity of unsuccessful utterances.

It might be possible to shed some light on this unexpected result be examining how the similarity of language use unfolds (Figure 2). A qualitative examination of the trends shows some evidence for early alignments of language use between the first \(20 \%\) of the negotiation and the next \(20 \%\) among all parties. However, the striking difference between the successful and unsuccessful utterance pairs is most evident starting around the \(60 \%\) point of the negotiation, on average. At this point there is a sharp drop in the similarity in language use of unsuccessful utterances. It appears that at that time period in the negotiation the coalitions are starting to form or have already formed (see Table 2 for sample of successful and unsuccessful utterances from that time period).

Table 2: Sample utterances from the \(60 \%\) slice of the negotiation (each utterance is from a different session)

\section*{Successful Utterances}
\begin{tabular}{l|l}
\hline \(\mathbf{1}\) & Do you still agree on our terms? \\
\(\mathbf{2}\) & So lets talk about the split. \\
\(\mathbf{3}\) & We can partner 3 ways and give them 3 k \\
\(\mathbf{4}\) & I presume you would rather do \(72 / 28\), right? \\
\(\mathbf{5}\) & Can we say \(\mathbf{6 9}\) to A, 48 to B and 4 to C to get a deal \\
\hline
\end{tabular}

\section*{Unsuccessful Utterances}


\section*{Discussion}

The analysis presented here, based on data collected by Huffaker et al. (2010), supports the hypothesis that the gradual alignment in language use is a contributing factor in reaching an agreement over the course of a negotiation. Participants in the study that were part of the final coalition showed more similar language use in the second half of the experiment than participants that were not part of the final coalition.

Interestingly, while the predicted interaction was found, the observed effect was somewhat different than expected. It appears that the major shift in language use leading to, or immediately following, the formation of a coalition is more likely to be a reduction in the utterance-to-utterance similarity for the excluded parties rather than an increase among the included parties. The causes for this require further study, but it seems possible that this is due to a change in the pattern of language use that the excluded party is not a part of. For instance, after agreeing to form a coalition, the parties might shift to discussing how to split the payoff while the excluded party might still attempt to convince one of the other participants to join a coalition.

The results of this paper seem to suggest that multi-party negotiations, while more complex than two-party negotiations and dialogues, follow many of the same patterns as their simpler counterparts. However, the added dynamics of such a negotiation also allows researchers to examine topics that are often difficult to explore when only two parties are involved in a linguistic exchange. In this case, it appears that when a party to the conversation or negotiation is "left behind", it might also fall out of linguistic alignment with the other participants.

Nevertheless, there is much room left for further analysis. While the initial analysis reported here provides some promising results, it is not conclusive. Furthermore, it opens the door for additional questions. For instance, it is possible that a 2-party coalition would be more likely to form during a private exchange rather than a public one. However, because of the relatively short length of some of the negotiations (under 50 utterances divided among the three parties), we elected not to separate the discussions based on whether they were part of a public exchange or a private one.

Another possible avenue for future investigation is to explore whether the patterns of linguistic entrainment differ based on the emerging final coalition. For instance, would a final 3-party coalition show a pattern consistent with a particular 2-party coalition up to some point at which the final participant in the negotiation also joins in? This is an interesting, if complicated, question that we leave open for future research.

\section*{Acknowledgements}

We would like to thank David Huffaker and Roderick Swaab for sharing the data from their study.

\section*{References}

Bazerman, M. H., Curhan, J. R., Moore, D. A., \& Valley K. L. (2000). Negotiation. Annual Review of Psychology, 51, 279-314.
Brennan, S. E., \& Clark, H. H. (1996) Conceptual pacts and lexical choice in conversation. Journal of Experimental Psychology: Learning, Memory, and Cognition, 22, 14821493.

Foltz, P. W., Kintsch, W., \& Landauer, T. K. (1998). The measurement of textual coherence with Latent Semantic Analysis. Discourse Processes, 25, 285-307.
Giles, H., Willemyns, M., Gallois C., \& Anderson, M. C. (2007) Accommodating a new frontier: The context of law enforcement. In K. Feidler (Ed.), Social Communication. New York, NY: Psychology Press.
Hu, X., Cai, Z., Wiemer-Hastings, P., Graesser, A., \& McNamara, D. (2007) Strength, Limitations, and Extensions of LSA. In McNamara D, Landauer TK, Dennis S, Kintsch W (Eds.) LSA: A Road to Meaning. Mahwah, NJ: Erlbaum.
Huffaker, D. A., Jorgensen, J., Iacobelli, F., Tepper, P., \& Cassell, J. (2006, June). Computational measures for language similarity across time in online communities. Paper presented at the Analyzing Conversation in Text and Speech (ACTS) Workshop at Human Language Technologies - North American Chapter of the Association for Computational Linguistics (HLTNAACL), New York, NY.
Huffaker, D. A., Swaab, R. I., \& Diermeier D. (2011) The Language of Coalition Formation in Online Multiparty Negotiations. Journal of Language and Psychology, 30, 66-81.
Miller, K. (2005) Communication Theories: Perspectives, Processes, and Contexts. New York, NY: McGraw-Hill.
Pickering, M. J., \& Garrod, S. (2004) Towards a mechanistic Psychology of dialogue. Behavioral and Brain Sciences, 27, 169-266.
Raiffa, H. (1982) The art and science of negotiation. Cambridge, MA: Harvard University Press.
Schütze, H. (1997) Ambiguity in language learning: computational and cognitive models. Chicago, IL: University of Chicago Press.
Swaab, R. I., Maddux, W., \& Sinaceur, M. (2011) Virtual linguistic mimicry: When and how online mimicry increases negotiation outcomes. Journal of Experimental Social Psychology, 47, 616-621.

\title{
Interpreting Covariation in Causal Structure Learning
}

\author{
Motoyuki Saito (m-saito@kwansei.ac.jp) \\ Department of Psychological Science, Kwansei Gakuin University \\ Hyogo, 662-8501, JAPAN
}

\author{
Tsuneo Shimazaki (shimazaki@kwansei.ac.jp) \\ Department of Psychological Science, Kwansei Gakuin University \\ Hyogo, 662-8501, JAPAN
}

\begin{abstract}
Recent studies have shown that people use covariation information to infer causal structure. However, there is little information about how people derive causal directionality from covariation. The present study is designed to provide further evidence about the role of covariation in causal structure learning. In Experiment 1, where covariation between two variables was systematically manipulated, participants were asked to observe the states of bacteria (present or absent) and to infer their causal relationship. We found that judgments of causal structure varied as a function of covariation, and that participants interpreted covariation according to necessity of causation. In Experiment 2, participants who received information about high causal strength interpreted covariation according to sufficiency of causation. These results demonstrate that prior knowledge modulates interpretation of covariation and suggest that domain-general covariation information and domain-specific prior knowledge of causal relations interact in causal structure learning.
\end{abstract}

Keywords: causal learning; covariation; prior knowledge; necessity; sufficiency.

\section*{Introduction}

Causal knowledge enables us to explain past events, to control the present environment, and to predict future outcomes. Using this knowledge, we can achieve desired outcomes and avoid undesired consequences. Many psychological studies have investigated how people acquire and use knowledge of causality (Gopnik \& Schulz, 2007; Sloman, 2005; see also Holyoak \& Cheng, 2011, for a review). Despite the importance of causal knowledge, it is often difficult to determine the casual structure among events. For example, imagine that someone feels unmotivated and makes slow progress on their work. In this situation, it is unknown whether the lack of motivation leads to slow progress, or whether slow progress causes lack of motivation. Furthermore, it is also possible that motivation and work progress are unrelated. Given this ambiguity, how do people learn causal structure?

Hume (1739/2000) argued that causal relations are unobservable and must be induced from observable events. Information about covariation among events serves as a fundamental cue for inferring causal structure. Covariation is represented as the pattern of occurrences and nonoccurrences for binary variables. Figure 1 shows a standard contingency table where the letters in each cell ( \(a, b, c, d\) )
represent the joint frequencies for one value of event X and one value of event Y. It is generally accepted that objective measure of contingency is described by \(\Delta P\), as shown in Equation 1 (Jenkins \& Ward, 1965).
\[
\begin{equation*}
\Delta P=P(Y \mid X)-P(Y \mid \neg X)=\frac{a}{a+b}-\frac{c}{c+d} \tag{1}
\end{equation*}
\]

In this equation, \(P(\mathrm{Y} \mid \mathrm{X})\) is the probability of Y given the presence of X , and \(P(\mathrm{Y} \mid \neg \mathrm{X})\) is the probability of Y given the absence of X . Values of \(\Delta P\) range from -1 to +1 . Positive \(\Delta P\) values indicate a generative causal relation; negative \(\Delta P\) values indicate a preventive causal relation. When a causal relation exists, strong covariation between the cause and the effect is expected. By contrast, lack of covariation indicates that two variables are unrelated (i.e., \(\Delta P\) does not differ significantly from zero). Many studies have focused on how people estimate causal strength between the candidate cause and the effect, and the results have shown that people are quite sensitive to covariation information (e.g., Wasserman, Elek, Chatlosh, \& Baker, 1993). However, covariation itself is inadequate for inferring a unique causal structure: when event X covaries with event Y , it is difficult to determine whether X causes Y , or vice versa.

When combined with additional information, covariation becomes a more useful cue to causal structure. First, temporal order in which people observe the occurrence of events facilitates learning causal directionality. As causes are often observed prior to their effects, when event X precedes event Y, it is highly probable that X causes Y. For example, if becoming unmotivated precedes making slow


Figure 1: A contingency table summarizing the covariation between two binary variables. The letters in each cell indicate frequencies of co-occurrence for the two states of events X and Y .
progress on work, temporal order information suggests that decreased motivation causes slow progress. Second, information about the absence of hidden causes also makes covariation cues more useful. When event X covaries with event Y, three possible causal structures are supposed (i.e., \(\mathrm{X} \rightarrow \mathrm{Y}, \mathrm{X} \leftarrow \mathrm{Y}\), or \(\mathrm{X} \leftarrow \mathrm{Z} \rightarrow \mathrm{Y})\). The possibility that both events are caused by a hidden common cause, \(Z\), can be excluded if it is known that there are no hidden causes. If event \(X\) exists alone, necessity of causation indicates that \(X\) causes Y (i.e., \(\mathrm{X} \rightarrow \mathrm{Y}\) ). This is because nothing happens without a cause (i.e., \(P(\) Effect \(\mid \neg\) Cause) \(=0)\). Therefore, events that exist alone must be a cause variable, not an effect variable. In contrast to necessity of causation, sufficiency of causation draws the opposite conclusion that event Y causes event X in above situation (i.e., \(\mathrm{X} \leftarrow \mathrm{Y}\) ). Since sufficiency of causation assumes that causes always accompany their effects (i.e., \(P(\) Effect \(\mid\) Cause \()=1\) ), events that occur alone must be an effect variable, not a cause variable. Given that there is no factor that affects both motivation and work progress and that motivation changes spontaneously, in previous example, necessity of causation suggests that decreased motivation causes slow progress and sufficiency of causation indicates that slow progress causes decreased motivation.

Recent studies on causal structure learning have revealed the importance of covariation (e.g., Deverett \& Kemp, 2012; Mayrhofer \& Waldmann, 2011; Rottman \& Keil, 2012; Saito \& Shimazaki, 2012). For instance, Saito and Shimazaki (2012) demonstrated that people judge simple causal structure on the basis of covariation information, but the use of covariation is modulated by task complexity. The experimental task was to observe the states of bacteria and to infer their causal relationship. Participants were instructed that temporal order was unreliable and that there were no hidden causes. In the simple causal structure condition, covariation was favored over temporal order as the basis for inferring causal structure; in contrast, temporal order was more influential in the complex causal structure condition. In addition, Mayrhofer and Waldmann (2011, Experiment 1) reported that people can differentiate common cause models (e.g., \(\mathrm{X} \leftarrow \mathrm{Z} \rightarrow \mathrm{Y}\) ) from common effect models (e.g., \(\mathrm{X} \rightarrow \mathrm{Z} \leftarrow \mathrm{Y}\) ) on the basis of covariation information. Although these recent studies show the ability to infer causal directionality from covariation, how people use covariation to induce causal directionality is still not well-understood. Therefore, it is valuable to study how people make structure judgments according to covariation.

The purpose of the present study is to investigate how people interpret covariation information in causal structure learning. In Experiment 1, we systematically manipulated covariation between two variables and asked participants to make causal structure judgments. Although causal structure between two variables is not determined by covariation alone, this situation enables us to examine whether participants have some sort of tendency in inferring causal directionality from covariation. In Experiment 2, we gave participants different information about causal relations and
investigated whether prior knowledge changed their interpretation of covariation.

\section*{Experiment 1}

Experiment 1 investigated how people interpret covariation when judging causal structure. The experimental task was to observe the occurrence of two fictitious bacteria and to infer their causal relationship. We manipulated covariation information by varying the number of occurrences and nonoccurrences of each bacterium.

\section*{Method}

Participants and Design Forty-three undergraduates from Kwansei Gakuin University received course credit for participating in this experiment. Two additional participants were excluded from the analyses due to misunderstanding of the instructions. Excluding these participants did not alter the general pattern of results.

Covariation information was systematically manipulated within participants. There were 15 covariation conditions (see Table 1) based on the combinations of five levels (1.00, .75, .50, .25, .00) of the conditional probabilities \(P(\mathrm{Y} \mid \mathrm{X})\) and \(P(\mathrm{Y} \mid \neg \mathrm{X})\). The difference between \(P(\mathrm{Y} \mid \mathrm{X})\) and \(P(\mathrm{Y} \mid \neg \mathrm{X})\) for each condition yielded five levels of nonnegative \(\Delta P\) values (1.00, .75, .50, .25, .00). Each participant completed the causal learning task for all covariation conditions.
Instructions Participants received verbal and written instructions in Japanese, and were asked to confirm that they understood the instructions. An English translation of outlines of the instructions was provided below:

Imagine that you are a scientist attempting to reveal a causal relationship between two types of newly discovered bacteria (These bacteria have the same shapes but different colors to conjure up an image of cell divisions). The term "causal relationship" means a relationship where one bacterium propagates the other bacterium (i.e., generative causal relationship). It is unknown whether one bacterium propagates the other, or whether these bacteria are unrelated. To investigate the relationship between the bacteria, you are going to observe the appearance of the bacteria. The states of the bacteria should help you consider the causal relationship between them.

Your task is to observe the occurrences and nonoccurrences of these bacteria and to infer their causal relationship. Note that the experimental task does not require any knowledge of biology. (The remaining instructions describe how to progress through the learning phase and test phase.)

Learning Phase Participants observed the states of bacteria (present or absent) to infer their causal relationship. On each trial, a button labeled "NEXT" was displayed on the screen. After clicking the button, information about the states of both bacteria X and Y was provided. The presence of a
bacterium was indicated by the appearance of the bacterium; in contrast, the absence of a bacterium was represented by the appearance of the bacterium labeled with a cross mark. The screen was returned to its primary state (i.e., "NEXT") 1 s after the bacteria appeared.

There were 16 trials for each covariation condition. Bacterium X was present on eight trials and was absent on eight trials (i.e., \(P(\mathrm{X})=.50\) ). Two conditional probabilities, \(P(\mathrm{Y} \mid \mathrm{X})\) and \(P(\mathrm{Y} \mid \neg \mathrm{X})\), were set to one of five levels in each condition (Table 1) The difference between these probabilities yielded five levels of nonnegative \(\Delta P \mathrm{~s}(\Delta P=\) \(P(\mathrm{Y} \mid \mathrm{X})-P(\mathrm{Y} \mid \neg \mathrm{X})\) ). Each condition was described through the difference between the two conditional probabilities. For example, in the \(.75-.00\) condition (i.e., \(P(\mathrm{Y} \mid \mathrm{X})=.75\), \(P(\mathrm{Y} \mid \neg \mathrm{X})=.00\) ), bacteria X and Y were both present on six trials. Bacteria X and Y were both absent on eight trials, and on two trials bacterium X was present and bacterium Y was absent. The order of trials and conditions was randomized within participants. To familiarize participants with the procedure, several practice trials were performed prior to the learning phase. Participants were informed that the information in the practice trials was irrelevant to the learning phase.
Test Phase After observing 16 cases, participants were asked two yes/no questions about the causal structure. They were asked whether bacterium X caused bacterium Y , and whether bacterium Y caused bacterium X. Then, after a brief delay, participants began the learning and test phases for the next covariation condition. They were instructed that judgments should be made independently of their answers on prior problems.

\section*{Results and Discussion}

Combining the answers on the two test questions yields four types of causal models: (1) X causes Y, (2) Y causes X, (3)
bidirectional, and (4) independent. The percentage of responses in each condition is shown in Table 1. Although causal structure could not be uniquely determined in all conditions, participants' judgments varied greatly. A loglinear model analysis on the 15 (covariation conditions) \(\times 4\) (causal models) cross table revealed a significant interaction between covariation condition and causal model, \(\chi^{2}(42)=\) 291.37, \(p<.001\).

In order to explore the interaction between covariation information and causal judgments in greater detail, we conducted a correspondence analysis. The contributions of dimensions 1 and 2 are \(63.75 \%\) and \(20.29 \%\), respectively, and their cumulative contribution is \(84.05 \%\). Therefore, we created scatter plots in two dimensions (Figure 2). As can be seen from Figure 2, each judgment is closely related to specific conditions. Most participants concluded that X caused Y in the \(.25-.00, .50-.00\), and \(.75-.00\) conditions and that Y caused X in the \(1.00-1.00,1.00-.75,1.00-.50\), and 1.00-. 25 conditions. Bidirectional causal relationships were only inferred in the \(1.00-.00\) condition. X and Y were judged to be independent in the other conditions.

Participants' judgments are explained in terms of necessity of causation (cf. Pearl, 2000). Necessity represents the degree to which the cause is necessary for the effect; in contrast, sufficiency is the degree to which the cause is sufficient for the effect. Pearl (2000) introduced three indices that assess causality: the probability of necessity \((P N)\), the probability of sufficiency \((P S)\), and the probability of necessity and sufficiency ( \(P N S\) ). These indices are easily calculated when the covariation information given does not include both the case where the effect is present in the absence of the cause (i.e., cell b), and the case where the effect is absent despite the presence of the cause (i.e., cell c). There also cannot be any common factors that have an

Table 1: Details of conditions, results, and interpretations in Experiment 1
\begin{tabular}{|c|c|c|c|c|c|c|c|c|c|}
\hline \multicolumn{3}{|l|}{Covariation conditions} & \multicolumn{4}{|c|}{Causal models (\% of participants)} & \multicolumn{3}{|c|}{Interpretations} \\
\hline \(P(\mathrm{Y} \mid \mathrm{X})\) & \(P(\mathrm{Y} \mid \neg \mathrm{X})\) & \(\Delta P\) & \(\mathrm{X} \rightarrow \mathrm{Y}\) & \(\mathrm{X} \leftarrow \mathrm{Y}\) & \(\mathrm{X} \rightarrow \mathrm{Y} \& \mathrm{X} \leftarrow \mathrm{Y}\) & X Y & PN & PS & PNS \\
\hline 1.00 & . 00 & 1.00 & 0.00 & 2.33 & 53.49 & 44.19 & & & \\
\hline 1.00 & . 25 & . 75 & 20.93 & 48.84 & 6.98 & 23.26 & \(\mathrm{X} \leftarrow \mathrm{Y}\) & \(\mathrm{X} \rightarrow \mathrm{Y}\) & \(\mathrm{X} \leftarrow \mathrm{Y}\) \\
\hline . 75 & . 00 & . 75 & 53.49 & 18.60 & 6.98 & 20.93 & \(\mathrm{X} \rightarrow \mathrm{Y}\) & \(\mathrm{X} \leftarrow \mathrm{Y}\) & \(\mathrm{X} \leftarrow \mathrm{Y}\) \\
\hline 1.00 & . 50 & . 50 & 27.91 & 53.49 & 0.00 & 18.60 & \(\mathrm{X} \leftarrow \mathrm{Y}\) & \(\mathrm{X} \rightarrow \mathrm{Y}\) & \(\mathrm{X} \leftarrow \mathrm{Y}\) \\
\hline . 75 & . 25 & . 50 & 11.63 & 13.95 & 18.60 & 55.81 & & & \\
\hline . 50 & . 00 & . 50 & 51.16 & 23.26 & 2.33 & 23.26 & \(\mathrm{X} \rightarrow \mathrm{Y}\) & \(\mathrm{X} \leftarrow \mathrm{Y}\) & \(\mathrm{X} \leftarrow \mathrm{Y}\) \\
\hline 1.00 & . 75 & . 25 & 27.91 & 60.47 & 2.33 & 9.30 & \(\mathrm{X} \leftarrow \mathrm{Y}\) & \(\mathrm{X} \rightarrow \mathrm{Y}\) & \(\mathrm{X} \leftarrow \mathrm{Y}\) \\
\hline . 75 & . 50 & . 25 & 9.30 & 20.93 & 11.63 & 58.14 & & & \\
\hline . 50 & . 25 & . 25 & 13.95 & 13.95 & 16.28 & 55.81 & & & \\
\hline . 25 & . 00 & . 25 & 46.51 & 23.26 & 2.33 & 27.91 & \(\mathrm{X} \rightarrow \mathrm{Y}\) & \(\mathrm{X} \leftarrow \mathrm{Y}\) & \(\mathrm{X} \leftarrow \mathrm{Y}\) \\
\hline 1.00 & 1.00 & . 00 & 20.93 & 51.16 & 0.00 & 27.91 & & & \\
\hline . 75 & . 75 & . 00 & 11.63 & 27.91 & 13.95 & 46.51 & & & \\
\hline . 50 & . 50 & . 00 & 11.63 & 4.65 & 11.63 & 72.09 & & & \\
\hline . 25 & . 25 & . 00 & 16.28 & 13.95 & 6.98 & 62.79 & & & \\
\hline . 00 & . 00 & . 00 & 23.26 & 16.28 & 0.00 & 60.47 & & & \\
\hline
\end{tabular}
influence on both events X and Y . Under these conditions, the probability of necessity, \(P N\), is calculated according to the following equation:
\[
\begin{equation*}
P N=\frac{P(Y \mid X)-P(Y \mid \neg X)}{P(Y \mid X)} \tag{2}
\end{equation*}
\]

When event X generates event Y (i.e., \(\Delta P>0\) ), values of \(P N\) become higher as the probability of Y given the absence of \(\mathrm{X}, P(\mathrm{Y} \mid \neg \mathrm{X})\), decreases. This reflects the fact that necessity of causation assumes that the base rate of the effect is low (i.e., \(P(\) Effect \(\mid \neg\) Cause \()=0)\). In contrast, the probability of sufficiency, \(P S\), is based on the assumption that causes always accompany their effects (i.e., \(P(\) Effect \(\mid\) Cause \()=1)\) and is defined as follows:
\[
\begin{equation*}
P S=\frac{P(Y \mid X)-P(Y \mid \neg X)}{1-P(Y \mid \neg X)} \tag{3}
\end{equation*}
\]

The probability of necessity and sufficiency, \(P N S\), takes both necessity and sufficiency aspects of causal relations into account:
\(P N S=P(Y \mid X)-P(Y \mid \neg X)\)

These indices are calculated on the basis of the causal direction from event X to event Y . Therefore, indices based on inverse direction are calculated by interchanging the rows and columns of the \(2 \times 2\) contingency table. When index values based on the direction from X to Y (e.g., \(P N\) from \(X\) to \(Y\) ) are compared with those based on the direction from Y to X (e.g., \(P N\) from Y to X ) and causal directionality is inferred by higher agreement with the conception (i.e., higher values), the three indices lead to different interpretations (see Table 1). For example, in


Figure 2: Results of correspondence analysis.
the \(.25-.00, .50-.00\), and \(.75-.00\) conditions, \(P N\) predicts that X causes Y. In contrast, in the 1.00-.75, 1.00-.50, and \(1.00-.25\) conditions, \(P N\) makes the opposite prediction. On the basis of responses in the conditions where the three indices are defined, we classified participants into one of five clusters: necessity, sufficiency, necessity and sufficiency, random, and unclassified. These classifications were made by rates of agreement between judgments and index predictions (1 for predicted judgments, 0 for unpredicted judgments). When participants had the same rate of agreement for different clusters, they were included in the unclassified cluster. As a result, more than half of the participants (55.81\%) were classified to the necessity cluster and \(27.91 \%\) of participants were classified in the sufficiency cluster. There were few participants in the other clusters \(\mathbf{~} 6.98 \%\) in the necessity and sufficiency cluster; \(2.33 \%\) in the random cluster; \(6.98 \%\) in the unclassified cluster). This suggests that most people interpret covariation information according to necessity.

However, these results are inconsistent with recent work suggesting that people judge causal relations on the basis of sufficiency (Mayrhofer \& Waldmann, 2011). Sufficiency of causation assumes that causal relations are deterministic (i.e., \(P(\) Effect \(\mid\) Cause \()=1\) ). According to sufficiency, the presence of event X in the absence of event Y is interpreted as an indication that X does not cause Y . Therefore, when covariation information includes such cases, it is suggested that Y causes X (i.e., \(\mathrm{X} \leftarrow \mathrm{Y}\) ). On the other hand, necessity of causation assumes that all events have a cause (i.e., \(P(\) Effect \(\mid \neg\) Cause \()=0)\) and such cases are taken as evidence that \(Y\) does not cause \(X\). In contrast to sufficiency, necessity indicates that X causes Y (i.e., \(\mathrm{X} \rightarrow \mathrm{Y}\) ) in the situation described above. Thus, judgments of causal directionality between two variables depend on the interpretation of covariation. In a second experiment, Mayrhofer and Waldmann (2011) had participants observe communications between two mind-reading aliens, and asked them to infer causal directionality. Covariation information included the case where two aliens X and Y thought the same thing, and the case where only one alien thought something (e.g., X) and the other alien thought nothing (e.g., \(\neg \mathrm{Y}\) ). Whereas sufficiency would suggest that alien Y transferred his thought to alien X (i.e., \(\mathrm{X} \leftarrow \mathrm{Y}\) ), necessity favors the opposite conclusion (i.e., \(\mathrm{X} \rightarrow \mathrm{Y}\) ). More participants concluded that \(Y\) caused X , suggesting that people judge causal relations on the basis of sufficiency.

These conflicting findings could be due to differences in prior knowledge about causal relations. Sufficiency of causation requires high causal strength, whereas necessity of causation requires the low base rate of the effect. If participants expect the effect's base rate to be low before the learning phase, covariation information is likely to be interpreted according to necessity. In contrast, prior knowledge about high causal strength might lead to an interpretation based on sufficiency. Indeed, it is difficult to imagine that the effect bacterium could occur in the absence of the cause bacterium in the bacteria story. That is,
participants will believe that bacteria do not arise spontaneously (i.e., \(P(\) Effect \(\mid \neg\) Cause \()=0\) ) and therefore, think the bacterium that exists alone must be a cause. In the alien story, however, such a situation is more plausible: an alien has the potential to think spontaneously, regardless of whether a cause alien is present (i.e., \(P(\) Effect \(\mid \neg\) Cause \()>0\) ). Participants will assume multiple causes in the alien cover story and regard the single-occurrence of the thought as an effect. Since necessity and sufficiency differ in their assumption about the base rate of effect and causal strength, differences in prior knowledge about these parameters might result in different judgments of causal structure based on covariation. We test this hypothesis in Experiment 2.

\section*{Experiment 2}

Experiment 1 demonstrated that participants made different judgments as a function of covariation. Whereas the results of Experiment 1 indicate that people interpret covariation information according to necessity of causation, Mayrhofer and Waldmann (2011) suggest that people interpret covariation according to sufficiency of causation. Experiment 2 was designed to investigate the effect of prior knowledge on interpretation of covariation information. The experimental procedure was similar to that of Experiment 1, but participants received different instructions about causal relations. We expected that additional instructions about high causal strength would lead to interpretations based on sufficiency of causation, and that participants who were not given additional instructions would infer causal structure according to necessity of causation.

\section*{Method}

Participants and Design Twenty-four undergraduates from Kwansei Gakuin University participated in the experiment and received course credit. None of them took part in Experiment 1. They were randomly assigned to either the sufficiency instruction or control group.
Procedure Each participant observed the states of bacteria (present or absent) and inferred their causal relationship. The procedure was the same as Experiment 1, with the following exceptions. First, participants in the sufficiency instruction group received additional instructions that emphasized the sufficiency of causation. In addition, to ensure that participants remembered this additional information, they were allowed to re-read the instructions during the learning and test phases. Finally, covariation information was manipulated within a context where
inferences could be uniquely identified as being made according to necessity or sufficiency.

In the instructions, the cover story was explained and participants were told to determine the causal relationship between two newly discovered bacteria. For participants in the sufficiency instruction group, instructions stated that the cause bacterium always accompanied the effect bacterium when one bacterium propagates the other bacterium (i.e., \(P(\) Effect \(\mid\) Cause \()=1)\), and that there are other causes in the environment that can produce the bacteria (i.e., \(P(\) Effect \(\mid \neg\) Cause \()>0)\). This information was not provided for participants in the control group.

In the learning phase, participants observed the states of bacteria on 16 trials. Six covariation conditions (.25-.00, . 50-. \(00, .75-.00,1.00-.75,1.00-.50\), and \(1.00-.25\) ) were used to determine whether participants interpreted covariation according to necessity ( \(P N\) ) or sufficiency ( \(P S\) ). Participants performed each condition twice in order to counterbalance the role of bacteria.

In the test phase, participants were told to judge the causal relationship in the same way as in Experiment 1. After a brief delay, participants completed the learning and test phases for the next condition. They were instructed that judgments should be made independently of their answers on prior problems.

\section*{Results and Discussion}

Participants' responses were analyzed in a manner similar to Experiment 1. First, judgments were categorized as one of four types of causal models. Next, we classified participants into one of five clusters (i.e., necessity, sufficiency, necessity and sufficiency, random, and unclassified) according to whether their judgments were predicted by \(P N, P S\), or \(P N S\). Table 2 shows the number of participants assigned to each cluster. Participants in the sufficiency instruction group were largely divided into the necessity cluster and sufficiency cluster. In contrast, almost all participants in the control group were assigned to the necessity cluster, replicating Experiment 1 where the majority of participants interpreted covariation on the basis of necessity. Fisher's exact test confirmed that there were significantly more judgments according to sufficiency of causation for participants in the sufficiency instruction group than the control group ( \(p<.05\) ). Although some participants still interpreted covariation according to necessity, these results indicate that prior knowledge modulated the interpretation of covariation information.

In summary, Experiment 2 showed that judgments of

Table 2: Number of participants assigned to each cluster in Experiment 2
\begin{tabular}{cccccc}
\hline & Necessity & Sufficiency & \begin{tabular}{c} 
Necessity and \\
Sufficiency
\end{tabular} & Random & Unclassified \\
\hline Sufficiency instruction & 4 & 6 & 0 & 1 & 1 \\
Group & 10 & 1 & 0 & 0 & 1 \\
\hline Control group & & & & & \\
\hline
\end{tabular}
causal structure were largely affected by prior knowledge about causal relations. When participants were informed about high causal strength, they were more likely to infer causal directionality on the basis of sufficiency; in contrast, participants not given additional instructions always judged causal structure according to necessity. These results bridge the gap between the results showing that judgments of causal structure are based on necessity (Experiment 1) and those showing that judgments of causal structure are based on sufficiency (Mayrhofer \& Waldmann, 2011).

\section*{General Discussion}

Recent studies have shown that people use covariation to infer causal directionality. However, there is little information about how people infer causal directionality from covariation. The present study was designed to investigate how people make causal structure judgments on the basis of covariation. Experiment 1 demonstrated that judgments of causal structure vary as a function of covariation, and that participants' answers can be explained in terms of necessity of causation. Experiment 2 showed that prior knowledge about high causal strength led more participants to interpret covariation according to sufficiency. The results of Experiment 2 are consistent with both findings concerning necessity interpretation of covariation (Experiment 1) and sufficiency interpretation of covariation (Mayrhofer \& Waldmann, 2011). These results reveal the importance of interpretations of covariation information in causal structure learning.

The results of the present study are closely related to the finding that learners flexibly interpret covariation during causal learning (Luhmann \& Ahn, 2011). Luhmann and Ahn (2011) asked participants whether they interpreted single pieces of covariation information as evidence of generative or preventive causal relations. The results showed that observations from Cell A can be interpreted as evidence for either a generative or preventive causal relation. These studies share the view that covariation information is flexibly interpreted, but focus on different aspects of causal learning. Whereas Luhmann and Ahn (2011) focused on learning causal strength, the present study addressed learning causal structure. An intriguing question for future research is to ask participants whether they interpret covariation as evidence for X causes Y or Y causes X .

The difference between a necessity interpretation of covariation in the bacteria story (Experiment 1) and a sufficiency interpretation in the mind-reading aliens story (Mayrhofer \& Waldmann, 2011) can be regarded as an interaction between domain-general causal inference and domain-specific knowledge. Whereas covariation is thought to be domain-general information, prior knowledge about causal relations seems to differ between the two stories. The results of Experiment 2 demonstrate that the basis for interpreting covariation can change from necessity to sufficiency when information about high causal strength is provided, but it remains unknown whether there are conditions that will change participants' basis for
interpretation from sufficiency to necessity. Another key question for future research is to investigate whether information about low base rate of the effect encourages a necessity interpretation of covariation in the mind-reading aliens cover story. Future research will provide further evidence about interactions between domain-general covariation and domain-specific prior knowledge.

\section*{References}

Deverett, B. \& Kemp, C. (2012). Learning deterministic causal networks from observational data. In N. Miyake, D. Peebles, \& R. P. Cooper (Eds.), Proceedings of the 34th Annual Conference of the Cognitive Science Society (pp. 288-293). Austin, TX: Cognitive Science Society.
Gopnik, A., \& Schulz, L. E. (2007). Causal learning: Psychology, philosophy, and computation. New York: Oxford University Press.
Holyoak, K, J., \& Cheng, P, W. (2011). Causal learning and inference as a rational process: The new synthesis. Annual Review of Psychology, 62, 135-163.
Hume, D. (1739/2000). A treatise of human nature. Oxford, England: Oxford University Press.
Jenkins, H. M., \& Ward, W. C. (1965). Judgment of contingency between responses and outcomes. Psychological Monographs, 79, 1-17.
Luhmann, C. C., \& Ahn, W.-K. (2011). Expectations and interpretations during causal learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 37, 568-87.
Mayrhofer, R., \& Waldmann, M. R. (2011). Heuristics in covariation-based induction of causal models: Sufficiency and necessity priors. In L. Carlson, C. Hoelscher, \& T. F. Shipley (Eds.), Proceedings of the 33rd Annual Conference of the Cognitive Science Society (pp. 31103115). Austin, TX: Cognitive Science Society.

Pearl, J. (2000). Causality: Models, reasoning and inference. Cambridge, United Kingdom: Cambridge University Press.
Rottman, B. M., \& Keil, F. C. (2012). Causal structure learning over time: Observations and interventions. Cognitive Psychology, 64, 93-125.
Saito, M., \& Shimazaki, T. (2012). Strategy changes in causal structure learning: The role of task complexity. In N. Miyake, D. Peebles, \& R. P. Cooper (Eds.), Proceedings of the 34th Annual Conference of the Cognitive Science Society (pp. 2264-2269). Austin, TX: Cognitive Science Society.
Sloman, S. A. (2005). Causal models: how people think about the world and its alternatives. New York: Oxford University Press.
Wasserman, E. A., Elek, S. M., Chatlosh, D. L., \& Baker, A. G. (1993). Rating causal relations: Role of probability in judgments of response-outcome contingency. Journal of Experimental Psychology: Learning, Memory, and Cognition, 19, 174-188.

\title{
The Rich Who Have the Humility of the Poor: Effects of Culture and Power on Altruism
}

\author{
Niloufar Salehi (nsalehi@ce.sharif.edu) \\ Department of Computer Engineering, Sharif University of Technology, Azadi Ave., Tehran, Iran \\ Morteza Dehghani (morteza@ict.usc.edu) \\ Institute for Creative Technologies, University of Southern California, 1205 Waterfront Dr., Playa Vista, CA 90094-2536, USA
}

\begin{abstract}
In this paper we investigate the effect of power on prosocial decision-making. While previous research has thoroughly investigated this relation in Western cultures, we focus our research on the role of power in an understudied MiddleEastern culture. Existing literature suggest an inverse relationship between feeling of power and prosocial behavior, where generally people in high levels of power tend to act less sympathetically in their decisions and demonstrate declined levels of perspective taking towards others. Our findings demonstrate that, unlike their Western counterparts, Iranian participants show significantly higher levels of altruism when in a high-power situation perceived as legitimate. On the other hand, under illegitimate power conditions, participants primed with high-power act significantly less compassionately in comparison to their low-power counterparts. We believe these findings have great impact in studying social hierarchies and behavior in cross-cultural settings.
\end{abstract}

Keywords: altruism; power prime; social hierarchy; decisionmaking; cross-cultural differences

\section*{Introduction}

The beloved of the Almighty are the rich who have the humility of the poor, and the poor who have the magnanimity of the rich.
-Saadi, \(13^{\text {th }}\) century Persian poet
Persian literature comprises a rich collection of myths, stories, and poems praising altruism and courtesy, especially among the powerful. Iranian children are encultured with stories, in which great kings and leaders are portrayed as generous, altruistic individuals. Such cultural products, created over generations, store and transmit cultural wisdom, and affect different aspects of people's judgment and decision-making (Weber, Hsee \& Sokolowska, 1998). Although the Iranian society has undergone a vast amount of societal change over the past decades, aspects of deeper cultural behavior have been transmitted through such cultural products, as tangible and public representations of the Iranian culture (Javidan \& Dastmalchian, 2003).

Several lines of research associate power with selfinterested behavior; linking it to lack of perspective taking (Galinsky, Magee, Inesi \& Gruenfeld, 2006), egoism (Batson, 1991) and moral reasoning (Lammers, Stapel \& Galinsky, 2010), among others. However, we argue that many of these findings are based on Western cultures,
relying on WEIRD populations (Henrich, Heine \& Norenzayan, 2010). It has been widely discussed that many models and theories of decision-making are based on cultural assumptions and may not be universally applicable (e.g. Dastmalchian, Javidan \& Alam, 2001; Jones, 2010; Henrich, et al., 2010; Hofstede, 1980). Arnett (2008) argues that the majority of psychological research focuses on American subjects, thus, neglecting \(95 \%\) of the world's population, a majority of whom live in vastly different societies. Further, limiting psychological models to small populations not only restricts the range that predictor variables can take, but also affectively limits discovery of other variables not yet included in the model (Weber \& Hsee, 2000).

In this study, we explore how power is associated with altruistic and sympathetic behavior among Iranians. Specifically, we aim to understand how legitimate as opposed to illegitimate feeling of power may affect prosocial decision-making. In other words, not only when, but also why do "the rich have the humility of the poor"?

\section*{Theories of Power}

In the study of social relationships, power is often referred to as the fundamental concept and basic force of behavior (Fiske, 1993; Kemper, 1991). Power is also closely related to the structures of personality (Wiggins \& Broughton, 1985). Thus, a wide range of research has focused on understanding how power influences various aspects of cognition, such as stereotyping (Fiske, 1993), social decision making (Gruenfeld, 1995), and perspective taking (Galinsky et al., 2006).

It has been widely argued that feeling of power results in displaying self-centered attitudes towards others, causing declined levels of perspective taking (Galinsky et al, 2006) and altruism (Batson, 1991). For example, in a recent study, Galinsky et al. (2006) asked participants under a high- or low-power experimental prime to draw an \(E\) on their foreheads - a procedure created by Hass (1984) to measure visual perspective taking of others. One way to draw the \(E\) is to consider one's own perspective, resulting in a backwards \(E\) illegible to other viewers. The other way is to consider others' perspective and draw an \(E\) backwards to oneself. They report that high-power participants were almost three times more likely to draw a self-oriented \(E\) than their low-power counterparts. Power is also linked to moral
hypocrisy, a situation of imposing firm moral standards on others, while practicing more tolerant moral standards oneself (Lammers, Stapel \& Galinsky, 2010).

The Power-Approach theory (Keltner, Gruenfeld \& Anderson, 2003) suggests that power increases goal-directed behavior without conscious awareness of its effects. This increase results in the powerful having a higher tendency to act and approach (Anderson \& Berdahl, 2002; Galinsky, Gruenfeld \& Magee, 2003). Correspondingly, in the context of decision-making and negotiations, the powerful are known to display higher aspirations (Pinkley, 1995), demand more and concede less (De Dreu, 1995), and often end up with the larger share of the pie in negotiations (Giebels, De Dreu \& Van de Vliert, 2000). On the other hand, powerlessness has been reported to activate the behavioral inhibition system (Carver \& White, 1994).

Further studies have shown that the link between power and goal-directed behavior is not always such straightforward. In fact, illegitimacy of the power involved may break this link (Lammers, Galinsky, Gordijn \& Otten, 2008a). Investigating the role of legitimacy of power, Lammers et al. (2008a) assigned participants to one of four cells in a 2 (powerless, powerful) \(\times 2\) (legitimate, illegitimate) between-participant design. Participants were primed using an essay task developed by Galinsky et al. (2003), where they were randomly assigned to one of four tasks and asked to recall and write about a situation of highor low-power under legitimate or illegitimate conditions. Participants were then asked to fill out a questioner assessing their behavioral activation/inhibition.

The authors report that under legitimate power conditions, the powerful had higher levels of behavioral activation than did the powerless. That is, under legitimate conditions, "the powerful act while the powerless follow" (Lammers et al., 2008b). However, these trends were reversed in illegitimate power conditions: among participants whose sense of power was illegitimate, low-power led to higher behavioral activation than high-power. In other words, it has been argued that power hierarchies, known to be based on mutual cooperation (Arendt, 1970), may switch to force and resistance when power is perceived as illegitimate (Lenski, 1966; Mills, 2000). Therefore, when studying the behavioral effects of power, legitimacy of the power must be considered. As previously mentioned, we argue that most of the studies on power have been conducted in Western cultures and the results may not generalize to all cultures.

\section*{Power Distance in Iran}

To shed light on the relation between power and culture, we rely on two of Hofstede's (1983) cultural dimensions directly related to pro-social behavior: Power Distance and Individualism.

In every society, there are strong forces that maintain and extend existing inequalities. The Power Distance Index measures the extent to which members of society accept these inequalities and allow them to grow (Hofstede, 1983). Hofstede argues that the power distance is supported by the
social environment, culture, and both high and low power members of the society (Hofstede, 2001). Iranians score high on the Power Distance Index (Hofstede, 2001), meaning that people tend to accept a hierarchical order and respect power inequalities.

The Individualism index is a cultural dimension that determines the relation between an individual and other members of society. In countries that score high on the individualistic scale members are expected to look after themselves, whereas in more collectivist countries, members consider themselves as members of a group and work towards fulfilling goals of the group (Hofstede, 1983). A low score in the Individualism index defines Iran as a collectivistic society indicating that people often consider themselves committed to a group, be that their family, friends, or extended relationships. In such a society loyalty to the group is principal and often overrides other social guidelines (Hofstede, 1983).

Another study by Dastmalchian et al. (2001) shows similar results about power and individualism in the Iranian society. This study was conducted as part of the GLOBE project concerning leadership attributions and cultural factors (House, Javidan \& Dorfman, 2001). Cultural dimensions of this study are extensions to those defined by Hofstede (1980). Societal collectivism is defined as "the degree to which organizational and societal institutional practices encourage and reward collective distribution of resources and collective action". Accordingly, in-group collectivism refers to "the degree to which individuals express pride, loyalty, and cohesiveness in their organizations or families" (House et al., 2001). Power distance is measured similar to Hofstede's definition. Results show that Iranians tend to have relatively high levels of power distance ( \(14^{\text {th }}\) out of 61 countries examined) and in-group collectivism ( \(3^{\text {rd }}\) out of 61), whereas fairly low levels (13 \({ }^{\text {th }}\) lowest country) of societal collectivism (Dastmalchian, et al., 2001).

Both studies share a common theme; the Iranian society is reported to have high levels of power distance and collectivism; in other words, a society of strongly accepted social hierarchies where high levels of collectivism are reported when members consider others as in-groups, but show highly individualistic behaviors when others are considered out of their "group". Javidan and Dastmalchian (2003) relate this to the structure of families in Iran, where the father has nearly total power in the family and children are taught from an early age to respect and obey those in position of authority such as their teachers. Due to the strength of families and group structures, behavior is determined by whether others are considered part of the ingroup or not.

\section*{Experiment}

In this experiment, we investigate the extent to which power affects altruistic behavior among Iranians. Previous studies in Western cultures have shown that the powerful demand more and display more act and approach than the powerless,
when power was primed to be legitimate. Whereas, when illegitimate power conditions were experienced, the powerless displayed even more approach than the powerful (Lammers, et al., 2008a). We investigate the interplay between the feeling of low or high power and the perception of legitimacy of that power among Iranian participants in the Dictator's game (Bolton, Katok \& Zwick, 1998) described below. Before we discuss the Dictator's game, we layout our three main hypothesis.

In our experiment, participants were asked to recall an experience where they were in a low- or high-power position; the legitimacy of the power involved was left for participants to decide. Through this we achieved a number of goals: before and during the priming stage participants were unaware of the means of the study and that legitimacy was a factor of the study, thus causing less biased responses. Further, by allowing participants to choose the situation they write about, we were able to measure the frequency of each class of responses. Therefore, we manipulated power status and left the legitimacy to be chosen and decided about by the participant. Accordingly, we study how Iranian participants self-assess the legitimacy of their power. Some scholars have linked power to moral hypocrisy, a situation where one imposes stricter moral standards to others than oneself (Lammers, et al., 2010). On the same track we believe that more people in the high-power prime will selfevaluate their actions as legitimate, than their low power counterparts.

H1: Participants under the high-power prime will have a higher tendency to self-evaluate their power as legitimate.

In collectivist societies, for those who are considered members of the in-group, very high levels of support, altruism and consideration are shown (Hofstede, 1983). This relation especially holds between older and younger members of the group, mainly due to the fact that older members are associated higher levels of power and authority (Javidan, et al., 2003). Accordingly, while high-power members, are highly respected and obeyed, they provide support and caring for others. Javidan and Dastmalchian (2003) also report a high level of desire for generosity and compassion among Iranian managers, a desire that they believe is rooted in the strong culture of family/group collectivism as well as Islamic principles. A view that the powerful should treat subordinates kindly (as their brothers and sisters) is also highly valued in Islamic teachings (Latifi, 1997). The difference between power-classes are exaggerated by the high power distance level in the Iranian society.

On the other hand, Lammers et al. (2008a) demonstrate that conceptualization of power determines its psychological consequences and provides insight into the tendency to approach. Relying on these facts, we hypothesize that, unlike their Western counterparts, when Iranian participants evaluate the power involved in their power prime as legitimate, they will show high levels of generosity and
support towards other participants as they feel a sense of obligation to provide support to the powerless.

H2: Under legitimate power conditions, high-power participants will demand less and show higher levels of altruism compared to low-power participants, as opposed to in illegitimate power conditions.

Parallel to the above reasoning, we predict that low-power subjects will expect to receive support and consideration from (high-power) others when their sense of power is perceived as legitimate.

H3: Low-power participants who view the power involved in their situation as legitimate will demand more than lowpower participants viewing the power as illegitimate.

This pattern is consistent, but opposite of the reported findings by Lammers, et al., (2008) who show an inverse relation between power and approach when power is perceived as illegitimate. We hypothesize that among Iranian participants, legitimate high power will result in higher concessions than legitimate low power, and the reverse will be resulted for illegitimate power conditions.

\section*{Method}

Participants. Fifty-two undergraduate students at Sharif University of Technology (24 female, mean age= 21.1) participated in this study. Each participant was ran in a separate session. In return for their participation, they received 4000 Tomans and chance to enter a raffle (play the Dictator's game, explained below).

Design. The study employed a \(2 \times 2\) between subject design. The first factor was the power prime (high-power or low-power). The second factor was perceived legitimacy of power (legitimate or illegitimate). The dependent variable was the amount of money taken by the participant in the Dictator's game.

\section*{Procedure}

Priming Stage. Participants were given a high- or a lowpower experiential prime, proven to reliably manipulate the sense of power (Galinsky et al., 2003). Participants were unaware of the aim of the experiment and were told that their essays will be used in a natural language processing project. Those assigned to the high-power condition were instructed to recall a personal incident in which they had power over other individuals, they were asked to write a short essay (in Persian) explaining both the incident and how they felt at that moment. Participants assigned to the low-power condition were similarly instructed to write about a personal incident in which someone else had power over them. A short definition of power was included in the instructions as having the ability to control and influence someone else or being in a position to evaluate them (Galinsky, et al., 2006). After writing the essay, both groups
were asked to evaluate the degree of the legitimacy of the power in the situation they had written about. Specifically, they were asked whether they believed that they were entitled to that powerful or powerless position both lawfully and morally. After completing the power-priming task, the experimenter thanked participants and they were paid 4,000 Tomans. After this stage participants were asked to participate in a raffle as part of the compensation. The raffle was an altered version of the Dictator's game.

The Dictator's Game. The Dictator's game (Bolton et al., 1998) is a commonly used game for evaluating levels of altruism and prosocial behavior. In the standard form of the dictator's game, one player (the dictator) is asked to share a fixed amount of money between himself and another participant (the receiver), while both players remain anonymous. According to economic models of decisionmaking, rational players are expected to maximize their personal benefit. Thus, the dictator should take all the money, leaving nothing for the receiver. Many studies, however, have shown that dictators give between \(20 \%-30 \%\) of the money to the unknown receiver (e.g. Camerer, 2011).
For this experiment, we developed a variation of the Dictator's game. As previously mentioned, in the standard form of the game, the dictator is asked to share a fixed amount of money between himself and the receiver. Due to the design of our experiment, we needed all participants to perform the role of the dictator. Hence, participants were instructed to take as much of the money as they want, leaving the rest for future unknown participants. Specifically, in our experiment, subjects were told that they had a chance to participate in a raffle. Each participant was then presented with 5 envelopes that were shuffled in front of her. While the participant did not know how much money was in the envelopes, each contained 10,000 Tomans, in 1,000 Toman bills. Participants were then asked to choose an envelope and take as much money as they want from it. They were told, however, that the same envelopes were to be used for future (unknown) participants; thus, any amount of money they leave in the envelope will be offered to future subjects who choose that same envelope in the raffle. Participants were given complete privacy to take as much of the money as they want. After each session, participants were paid and briefed about the experiment. Then, the envelope chosen by the participant was marked with the participant's ID and put aside. The money taken from the envelopes were used as the main dependent variable in our experiment.

\section*{Results}

\section*{Perceived Legitimacy of Power}

During the power priming task participants were only instructed to write an assay about a high- or a low-power situation. In this stage nothing was mentioned about the legitimacy of the power involved. After finishing the essay participants were asked to evaluate the power involved in their situation as legitimate or illegitimate. Our findings
show that the state of power had a great impact on this evaluation.
As shown in Figure 1 and consistent with \(H 1\), we found that \(78 \%\) of participants ( \(22 / 28\) ) under a high power prime evaluated their power as legitimate, compared to \(33 \%\) of participants ( \(8 / 24\) ) under a low power prime. The difference between these conditions were significant \(\chi^{2}(1, N=52)=\) \(9.061, p=.003\). In other words, \(67 \%\) of participants reported a sense of injustice when asked to recall a situation of low-power.


Figure 1: Self-evaluation of legitimacy among participants.

\section*{Altruism}

The level of altruism among participants was calculated by the amount of money left in the raffle envelope during the Dictator's game. This is the amount of money that participants decided to leave for future participants. We measured the amount of money taken from the envelope as a numerical indicator of selfishness. In seven cases this number was negative, indicating that participants had left some of the earlier reward money ( 4,000 Tomans) as well as the money already in the envelope for the next person (note that participants viewed the Dictator's game as part of the compensation). Figure 2 displays the average amount of money taken by participants.
Using the amount of money taken as a dependent variable in a \(2 \times 2\) ANOVA, with power (low/high) as the first factor and legitimacy (legitimate/illegitimate) as the second, revealed a significant interaction between power and perceived legitimacy of power \(F(1,48)=7.652, p=.008\).

Under power conditions perceived as legitimate, highpower participants took significantly less money in the Dictator's game ( \(M=2.63, S D=0.91\) ), than low-power participants \((M=6.25, S D=1.22) t(28)=2.14, p=.042\). Thus, supporting \(H 2\), our results demonstrate high levels of altruism among the legitimate powerful and naturally higher
demands from the legitimate powerless, who expect to be supported.

When power was perceived as illegitimate, there was significant difference in the amount of money taken by the high-power ( \(M=5, S D=1.36\) ) as opposed to the low-power ( \(M=2, S D=0.81\) ) participants \(t(20)=1.91, p=.071\). Thus, supporting the second part of \(H 2\), the illegitimate powerful display declined levels of altruism and more approach, whereas the illegitimate powerless demand less.

Also, there was a significant difference between lowpower legitimate and low-power illegitimate conditions \(t(22)=2.950, p=0.007\), with participants in low-power legitimate conditions taking significantly more money than those in the low-power illegitimate condition. This result supports \(H 3\).


Figure 2: Average money taken in the Dictator's game.

\section*{Discussion}

Overall, our results show significantly different patterns than those in similar studies performed among Western participants. Iranian participants relating a situation of highpower, most often found their power to be legitimate. Some participants even used expressions such as "even though others might think what I did wasn't just, I still believe I did what was right". Our results show a consistent pattern, suggesting people often find a way to justify their actions when they stand in a high-power position. Such a claim confirms previous studies showing higher levels of moral hypocrisy among high-power individuals (Lammers, et al., 2010). On the other hand, as previously discussed, most participants under a low-power prime reported the power involved as illegitimate. For example, essays written by this class of participants included expressions such as "He had no right to do that" or "their actions were purely selfish even though they had the legal right". In this case our findings contradict prior studies on Western cultures claiming: "decreased power, results in people being less
critical on others and more critical on the self" (Lammers et al., 2010). In other words, low-power situations tend to make Iranian individuals highly critical on the holder of power, often resulting in the powerful being perceived as illegitimate.
One possible explanation for our results, relying on previously proposed features of collectivism amongst Iranians, is that when participants evaluate the power involved in their power prime as legitimate, they are naturally considering others as their in-group. In these circumstances, due to their high levels of collectivism, they become more altruistic. However, when under an illegitimate power prime, others are spontaneously considered as out-groups. In other words, inside the group, and where power is perceived as legitimate, the powerless are supported and show higher levels of approach and demand. On the other hand, when power is distributed illegitimately, others are viewed as outside the social group, the powerless fail to demand, and fall into an oppressed state similar to social inhibition described by Carver and White (1994). Our results show that in this case, legitimacy has such an effect that people experiencing a state of illegitimate low-power demand less than \(1 / 3\) of people in a similar low-power situation but who feel legitimacy in the state. In our future experiments, we plan to explore the interplay between power and in-group/out-group behavior.
Shedding light on the origins of the relation between altruism and high-power behavior among Iranians, begs a deeper discussion of religious and cultural settings. A view that the powerful should treat subordinates kindly (as their brothers and sisters) is highly valued in Islamic teachings (Latifi, 1997). Moreover, some studies suggest that Iranians commonly view their superiors in the same light as their older siblings or parents, describing the relationship between an employee and supervisor close to that of family members; therefore, managers are often expected to show support, generosity and compassion towards subordinates (Latifi, 1997; Tayeb, 1997; Javidan, et al. 2003). A high score on the power distance index further enhances this effect as social hierarchies are mutually accepted and practiced from an early age. Thus, their properties become deep aspects of social behavior.
One must take into account that the question of whether generosity is a pure act or has underlying selfish motives remains an open question. For example, various studies argue that some of the giving is due to the fact that the dictator does not want to seem selfish to the anonymous receivers (Dana, Cain, \& Dawes, 2006). In this study, we have not addressed this question; rather, we consider how altruistic behaviors are practiced in the society regardless of fundamental motives. We plan to address such issues in future studies.

\section*{References}

Anderson, C., \& Berdahl, J. L. (2002). The experience of power: examining the effects of power on approach and
inhibition tendencies. Journal of personality and social psychology, 83(6), 1362.
Arendt, H. (1970). On violence (Vol. 17). Mariner Books.
Arnett, J. J. (2008). The neglected 95\%: why American psychology needs to become less American. American Psychologist, 63(7), 602.
Batson, C. D. (1991). The altruism question: Toward a social-psychological answer. Lawrence Erlbaum Associates, Inc.
Bolton, G. E., Katok, E., \& Zwick, R. (1998). Dictator game giving: Rules of fairness versus acts of kindness. International Journal of Game Theory, 27(2), 269-299.
Camerer, C. F. (2011). Behavioral game theory: Experiments in strategic interaction. Princeton University Press.
Carver, C. S., \& White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS Scales. Journal of personality and social psychology, 67(2), 319.
Dana, J., Cain, D. M., \& Dawes, R. M. (2006). What you don't know won't hurt me: Costly (but quiet) exit in dictator games. Organizational Behavior and Human Decision Processes, 100(2), 193-201.
Dastmalchian, A., Javidan, M., \& Alam, K. (2001). Effective leadership and culture in Iran: An empirical study. Applied Psychology, 50(4), 532-558.
De Dreu, C. K. (1995). Coercive power and concession making in bilateral negotiation. Journal of Conflict Resolution, 39(4), 646-670.
Fiske, S. T. (1993). Controlling other people: The impact of power on stereotyping. American psychologist, 48(6), 621.

Galinsky, A. D., Magee, J. C., Inesi, M. E., \& Gruenfeld, D. H. (2006). Power and perspectives not taken. Psychological Science, 17(12), 1068-1074.
Galinsky, A.D., Gruenfeld, D.H, \& Magee, J.C. (2003). From power to action. Journal of Personality and Social Psychology, 85, 453-466.
Giebels, E., De Dreu, C. K., \& Van de Vliert, E. (2000). Interdependence in negotiation: Effects of exit options and social motive on distributive and integrative negotiation. European Journal of Social Psychology, 30(2), 255-272.
Gruenfeld, D. H. (1995). Status, ideology, and integrative complexity on the US Supreme Court: Rethinking the politics of political decision making. Journal of Personality and Social Psychology, 68(1), 5.
Hass, R. G. (1984). Perspective taking and self-awareness: Drawing an E on your forehead. Journal of Personality and Social Psychology, 46(4), 788.
Henrich, J., Heine, S. J., \& Norenzayan, A. (2010). Most people are not WEIRD. Nature, 466(7302), 29-29.
Hofstede, G. (1980). Motivation, leadership, and organization: Do American theories apply abroad. Organizational dynamics, 9(1), 42-63.

Hofstede, G. (1983). The cultural relativity of organizational practices and theories. Journal of international business studies, 14(2), 75-89.
Hofstede, G. (2001). Culture's consequences: Comparing values, behaviors, institutions and organizations across nations. Sage Publications, Incorporated.
House, R., Javidan, M., \& Dorfman, P. (2001). Project GLOBE: an introduction. Applied Psychology, 50(4), 489-505.
Javidan, M., \& Dastmalchian, A. (2003). Culture and leadership in Iran: The land of individual achievers, strong family ties, and powerful elite. The Academy of Management Executive, 17(4), 127-142.
Jones, D. (2010). A WEIRD view of human nature skews psychologists' studies. Science, 328(5986), 1627-1627.
Keltner, D., Gruenfeld, D. H., \& Anderson, C. (2003). Power, approach, and inhibition. Psychological review, 110(2), 265.
Kemper, T. D. (1991). Predicting emotions from social relations. Social Psychology Quarterly, 330-342.
Lammers, J., Galinsky, A. D., Gordijn, E. H., \& Otten, S. (2008a). Illegitimacy moderates the effects of power on approach. Psychological Science, 19(6), 558-564.
Lammers, J., Galinsky, A. D., Gordijn, E. H., \& Otten, S. (2008b). Power and cooperation; the importance of legitimacy. Unpublished manuscript, Tilburg University, Tilburg, The Netherlands.
Lammers, J., Stapel, D. A., \& Galinsky, A. D. (2010). Power increases hypocrisy moralizing in reasoning, immorality in behavior. Psychological Science, 21(5), 737-744.
Latifi, F. (1997). Management learning in national context. Unpublished PhD thesis, Henley Management College.
Lenski, G. E. (1966). Power and privilege: A theory of social stratification. UNC Press Books.
Mills, C. W. (2000). The power elite. Oxford University Press, USA.
Pinkley, R. L. (1995). Impact of knowledge regarding alternatives to settlement in dyadic negotiations: Whose knowledge counts?. Journal of Applied Psychology, 80(3), 403.
Tayeb, M. (1997). Islamic revival in Asia and human resource management. Employee Relations, 19(4), 352364.

Weber, E., Hsee, C., \& Sokolowska, J. (1998). What folklore tells us about risk and risk taking: Cross-cultural comparisons of American, German, and Chinese proverbs. Organizational Behavior and Human Decision Processes, 75(2), 170-186.
Weber, E., \& Hsee, C. (2000). Culture and individual judgment and decision making. Applied Psychology, 49(1), 32-61.
Wiggins, J. S., \& Broughton, R. (1985). The interpersonal circle: A structural model for the integration of personality research. Perspectives in personality, 1, 1-47.

\title{
Intonation and Positional Effects in Spoken Serial Recall
}

\author{
Michelina Savino (michelina.savino@uniba.it) \\ Dipartimento di Scienze della Formazione, Psicologia, Comunicazione, Università di Bari, piazza Umberto I, 1 70121 Bari, ITALY
}

\author{
Andrea Bosco (andrea.bosco@uniba.it) \\ Dipartimento di Scienze della Formazione, Psicologia, Comunicazione, Università di Bari, piazza Umberto I, 1 70121 Bari, ITALY
}

\author{
Martine Grice (martine.grice@uni-koeln.de) \\ IfL-Phonetik, Universität zu Köln, Herbert-Lewin-Strasse 6 \\ 50931Cologne, GERMANY
}

\begin{abstract}
Past studies have indicated that intonation, in the sense of fundamental frequency modulation, can only enhance serial recall to the extent that it can induce a grouping effect, something that can also be induced by a simple insertion of pauses. However, in a study of spoken serial recall of ninedigit lists, we are able to show that recall is significantly better when sequences of digits are marked by specific intonation contours than when they are simply grouped by silent pauses in the signal. Thus, we found that intonation plays a role during the encoding phase, whereby items in group-final positions draw particular benefit from intonation. However, intonation does not appear to play the same role in the retrieval phase, since when subjects are instructed to imitate intonation during recall, performance shows mixed effects.
\end{abstract}

Keywords: serial recall; intonation; grouping effect; shortterm memory

\section*{Introduction}

In field research, there is a general consensus that serial recall (short-term memory in general) and prosody are closely related. Well-documented evidence of such a relationship is the grouping effect, that is, the enhanced recall of items in a list when they are presented in groups (for example, Reeves et al 2000). The grouping effect is stronger for auditory stimuli (Cowan et al 2002, Frankish 1985), as prosody plays an important role in this grouping, or patterning. Past and more recent research has aimed to ascertain the nature of these groups. In their seminal work on the auditory grouping effect, Frankish (1995) and Saito (1998) provided evidence that it can be obtained by temporal organisation of speech stimuli realised by pause insertion between groups as well as by superimposing a "natural" intonational pattern, or by manipulating pitch levels on groups of items (Frankish 1995). On the other hand, more recent studies have shown that the grouping effect in serial recall reflects rhythmic groups, referred to as the stress grouping effect, rather than intonational phrases (Reeves et al 2000, Boucher 2006, Gilbert \& Boucher 2007). It is argued that these groups correspond to the
segmentation units (chunks) listeners use in spoken language perception (Gilbert et al 2011).
Previous research dealing specifically with the role of intonation in improving serial recall suggests that it is relevant to the extent that it can induce a similar grouping effect to that obtained by pause insertion (Frankish 1995, Saito 1998). However, a number of potentially relevant aspects of intonation deserve further exploration. For instance, in the previous studies discussed above, the superimposition of a fundamental frequency (F0) contour \({ }^{1}\) on the whole sequence, or F0 manipulation on groups of items within a sequence were carried out with little control over specific tunes and their associated meanings. This is particularly relevant, since the role of intonation in signalling discourse structure is widely acknowledged, as it cues hierarchical relationships among phrases within a discourse unit (Hirschberg \& Pierrehumbert 1986). Moreover, in sequences intonation can convey information about the hierarchical structure (groups) as well as about specific positions within a group. In Italian (in particular the variety of Bari), a rich inventory of tunes is available for marking those kinds of hierarchical relationships in a sequence (Savino 2001; 2004), among which the most typical are:
- The "continuation rise" contour, a gradually rising F0 movement from the nuclear syllable up to the end of the phrase. It signals that the list has not been completed yet, and that more items are to come ("non-finality" contour);
- A high rising contour, where the rise in F0 starts before the nuclear syllable and continues rising up to the end of the phrase. It conveys the information that the current item is the penultimate in a sequence, i.e. that the end of the list is approaching ("pre-finality" contour");
- A falling contour, involving a gradual fall from the nuclear syllable until the end of the phrase. This contour marks the end of a sequence ("finality" contour).
Our aim here is to verify whether the use of specific tunes conveying such hierarchical relationships and positional

\footnotetext{
\({ }^{1}\) The F0 contour corresponds roughly to what is perceived as the pitch, or melody.
}
information within a sequence could improve immediate serial recall performance of auditory stimuli in (Bari) Italian listeners.
Another positive effect in serial recall performance reported in the literature is the salience effect. It "occurs when an item that is conspicuous on some perceptual dimension is recalled better than other items in the same ensemble in learning and memory tasks" (Reeves et al 2000: 1639). In this paper we also explore the role of pitch prominence in within-group (medial) positions in the sequence as another potential factor in serial recall enhancement.
Moreover, we evaluate whether intonation plays a specific role not only in encoding but also in retrieval of phrases, i.e. whether performance is improved when listeners imitate the tune of the input stimuli during recall.

\section*{Method and Materials}

\section*{Intonational Patterns}

In order to investigate the role of specific tunes in serial recall, we identified two intonational patterns (we called 'intonation contour A' and 'intonation contour B') characterised by F0 shapes conveying hierarchical organisation of groups, as well as positional information of items across and within groups. In a nine-digit sequence, we determined:
- 'Intonation contour A', consisting of the "continuation rise" ("non-final") contour at positions 3 and 6, and a lowfalling ("final") contour at position 9. F0 shapes of items at initial and within-group positions (positions \(2,5,7,8\) ) are not positionally marked by intonation, as they are characterised by a peak accent (taken to be the neutral unmarked pattern).
A schematisation of intonation contour A is shown in Figure 1;
- 'Intonation contour B', sharing the same intonational patterns of 'intonation contour A', except for a steep rising pitch accent at positions 2 and 5 . It can pre-signal the end of the first and the second groups, i.e. the two non-final groups within the signal. Also, because of the steep rising accentual movement, these digits sound perceptually more salient than the corresponding positions 2 and 5 in intonation contour A, where they are marked by a (neutral) peak accent instead. Another feature of contour B is a high rising ("pre-final") contour at position 8 pre-signalling both the end of the third group and the end of the whole sequence.
A schematisation of 'intonation contour B' is given in Figure 2.
These two experimental conditions were compared with two further ones, namely:
- 'Grouped by pauses' sequences, where all digits have a peak contour, and sequences are temporally grouped into three by inserting a pause at the end of each group;
- 'Ungrouped' sequences, sharing the same intonation of the 'grouped by pauses' stimuli above, but without pause grouping.
We hypothesise that serial recall performance would be
(H1) better in both intonation contours A and B and the 'grouped by pauses' conditions than in the 'ungrouped' (control) condition, due to the grouping effect;
(H2) better in both 'contour A' and 'contour B' than in the 'grouped by pauses' condition, because of the absence of intonational marking of item position in the latter condition. In particular, items in positions 3 (non-final contour=last item in the first group), 6 (non-final contour=last item in the second group), and 9 (final contour=last item in the third group and in the whole sequence) should benefit in terms of recall enhancement.
(H3) better with 'intonation contour B' than 'intonation contour A' because of the enhanced hierarchical and positional information conveyed by intonation in certain positions, namely: digits at positions 2 (pre-final contour \(=\) item at mid position in the first group), 5 (pre-final contour \(=\) item at mid position in the medial group), and 8 (pre-final contour \(=\) penultimate item in the whole sequence) This should result in an overall better recall performance of 'intonation contour B' than of 'intonation contour A' sequences.
Moreover, we hypothesise that recall performance in both intonation conditions
(H4) would be enhanced for subjects who are instructed to imitate the intonation produced in the stimuli during the recall task. This would provide some evidence that intonation plays a relevant role not only in encoding but also in retrieval of verbal material.
Our hypotheses can be summarized as follows:
Contour \(B>\) Contour \(A>\) Grouped by pauses \(>\) Ungrouped on the one hand, and Imitation > No imitation on the other.


Figure 1: Schematisation of "intonation contour A"


Figure 2: Schematisation of "intonation contour B"

\section*{Stimuli}

The sequences were created according to the four conditions ('ungrouped', 'grouped with pauses', 'intonation contour A', 'intonation contour B'), three types of spoken stimuli for each digit were created:
Type (a), where each digit was realised with a neutral F0 peak, as, for example, in digit position 1 in contours A and B (see schematisation in Figures 1-2).
Type (b), where sequences were realised with intonation contour A described above (see Figure 1)
Type (c), where sequences were produced with intonation contour B described above (see Figure 2)

All series for each of the digits were produced by a trained native speaker of Bari Italian (author MS) in the same recording session. In this way, for each digit all intonational realisations in each position (first, second, third, etc.) within each sequence type were available. They were saved as individual audio files, and used as "building blocks" for creating all the nine-digit spoken sequences under the 4 conditions. Stimuli were created by concatenating the individual audio files into nine-digit sequences. In a postediting step, care was taken that speech signal amplitude was homogeneous in all sequences.
This methodology enabled us to create stimuli which sound more natural than those produced by means of a speech synthesiser, as in Frankish (1995).
Spoken digit realisations of type (a) were used for creating sequences for the conditions 'ungrouped', and 'grouped by pauses', in the latter case by inserting a 310 ms silence after digits in positions 3 and 6 . Spoken digit renditions of the types (b) and (c) were used for creating sequences under the conditions 'intonation contour A' and 'intonation contour B', respectively.
We created 68 nine-digit lists from pseudo-random permutation of the 1-9 digits, in a way to avoid the presence of two adjacent digits in ascending or descending order, and to make sure that a digit would not appear in the same position in consecutive lists.
The concatenated nine-digit sequences were created on the basis of these lists, the duration of each sequence averaging 6.4 sec. We produced 17 stimuli for each of the four conditions, for a total amount of 68 stimuli (including 8 to be used for the training session only, 2 per condition).
All steps for the preparation of stimuli were carried out using the Praat software tool for speech analysis (Boersma \& Weenink 2001). Examples of a sequence (speech waveform and F0 contour) for each of the four conditions are shown in Figures 3-6.

\section*{Informants}

Fifty-six subjects participated in the experiment. They were undergraduate and graduate students of Psychology at the University of Bari (average age: 23.6), with no reported speech or hearing deficits. They were all native speakers of Italian, born and living in the Bari geolinguistic area. None of them had a background in phonetics or prosody.

\section*{Procedure}

Participants was tested individually in a laboratory. They were asked to listen to each sequence and recall all nine digits orally by strictly observing their serial order. Half of them were also instructed to imitate the intonation of the sequence during the recall.
Subjects were seated in front of a computer and wore a headset with headphones and microphone. Each list was preceded by a warning tone and 500 ms silence. Spoken responses were recorded directly and subjects proceeded to the next sequence by pressing the spacebar. They were also
free to take a break any time they needed during the session. A break was suggested after every block of 15 stimuli.
For each subject, the task involved recalling 60 stimuli (preceded by a short training session) i.e. 15 stimuli for each of the four conditions. The order of presentation was balanced across the subjects.
Before starting the experimental trial, participants were tested for their digit span of the WAIS-R (Wechsler 1987). This step was carried out in order to ascertain that the digit span of subjects was homogeneous across groups.
The average total duration of sessions (including the digit span test) was around 40 min . Sessions were implemented and run using SuperLab 2.0.


Figure 3: Speech waveform and F0 contour of one of the stimuli for the 'ungrouped' (control) condition.


Figure 4: Speech waveform and F0 contour of one of the stimuli for the 'grouped by pauses' condition. Vertical broken lines mark silence intervals (pauses) between groups.


Figure 5: Speech waveform and F0 contour of one of the stimuli for the 'intonation contour A' condition. Vertical broken lines mark the right edge of each group (intonational phrase)


Figure 6: Speech waveform and F0 contour of one of the stimuli for the 'intonation contour B' condition. Vertical broken lines mark the right edge of each group (intonational phrase)

\section*{Results and Discussion}

A mixed factors general linear model was performed with: 1) condition (within subject), 4 levels (ungrouped, grouped by pauses, intonation A , intonation B )
2) imitation of intonation (between subjects), 2 levels (yes, no)
3) serial group within the sequence (within subjects), 3 levels (first group, second group, third group)
4) within group position (within subjects), 3 levels (first position, second position, third position)
as factors.
First of all, a large effect of condition was found: F \((3 ; 159)\) \(=52.75 ; \mathrm{p}<0.001\); partial eta square \(=0.5\), performance percentages for each condition: ungrouped \(=52 \%\), grouped by pauses \(=64 \%\), intonation contour \(\mathrm{A}=70 \%\), intonation contour \(\mathrm{B}=69 \%\) ). These results are in line with our prediction in (H1), and confirm the typical grouping effect in immediate serial recall of prosodically manipulated auditory stimuli (Frankish 1995, Saito 1998).
Importantly, our hypothesis in (H2) is also verified, since results show a statistically significantly better recall performance for sequences with intonation contours A and B than for those grouped simply by pauses (post-hoc \(\mathrm{p}<0.01\) ). These results point to a specific role of intonation in serial recall that goes beyond the grouping effect induced by pauses, counter to the claims in previous studies.
We also found a significant second order interaction between condition, serial group and within-group position: \(\mathrm{F}(12 ; 636)=2.35 ; \mathrm{p}<0.01\); partial eta square \(=0.04\), with a moderate effect. Results (Figures 7-10) relating to positions in the first and second serial groups show a better performance for sequences realised with contours A and B, with respect to those grouped by pauses. The effect is particularly relevant for positions 3 and 6 (i.e. the last position in the first and second groups). This can be seen by comparing the "descending" shape of the correct response bars for groups 1 and 2 in the ungrouped condition (Figure 7), with the "dipped" shape for the same group in the remaining 3 conditions (Figures 8-10). The reverse tendency is evident for the third group, with an "ascending" shape in all conditions, pointing to an enhancement of the recency effect in all conditions.


Figure 7: Number of correct recalls (mean values) across the 3 serial groups in the sequence and the positions within each group (first POS, second POS, third POS), for the 'ungrouped' (control) condition.


Figure 8: Number of correct recalls (mean values) across the 3 serial groups in a sequence and the positions within each group (first POS, second POS, third POS), for the 'grouped by pauses' condition.


Figure 9: Number of correct recalls (mean values) across the 3 serial groups in a sequence and the positions within each group (first POS, second POS, third POS), for the 'intonation contour A' condition

Hypothesis (H3) is not confirmed by our results, as we did not obtain the expected specific effect in contour B for the second (pre-final) position within the serial groups (i.e. positions 2,5 and 8 within the sequence, where 8 is also the penultimate item position in the whole sequence).
For position 5 , the salience potentially conveyed by the steep rising pitch accent (compared to the "neutral" pitch peak in the same position in contour A) does not appear to improve recall of the medial item in a sequence.
With regard to Hypothesis H4, we did not find a main effect of the imitation of intonation, even though this factor interacts with a moderate effect with the serial groups: F (2; \(106)=5.82 ; p<0.01 ;\) partial eta square \(=0.1)\). Diagrams in

Figure 11 show that imitation of intonation improves the recall of the first group within a sequence, whereas performance decreases for the second and third serial groups. Thus, the imitation of intonation does not appear to facilitate recall, except for the digits in the first group, when the memory traces are stronger (as confirmed by low recency effects in our data).


Figure 10: Number of correct recalls (mean values) across the 3 serial groups in a sequence and the positions within each group (first POS, second POS, third POS), for the 'intonation contour B' condition


Figure 11: Number of correct recalls across the three serial groups as a function of instructions to imitate the intonation.

\section*{Conclusions}

The aim of our study was to ascertain whether the use of intonation patterns conveying hierarchical relationships among and within groups in a sequence could improve performance in an immediate serial recall task. Output
modality is oral channel, as in serial recall output modality can have a considerable influence in task performance (Penney 1979). Our results show that oral recall of spoken lists is better when auditory stimuli are intonationally marked by specific tunes than when they are simply temporally grouped by pauses. Some positions particularly benefitted from intonation: a rising non-final contour marking serial positions 3 and 6 is more effective than pauses.
On the other hand, we did not find clear evidence that a prefinal contour in positions 2 and 8 (marking penultimate position within the first group, and penultimate position within the whole sequence, respectively) can enhance recall of items in these positions, also possibly because of the masking effect of primacy and recency in adjacent positions. It also appears that the item in medial position within the whole sequence (position 5) cannot be better recalled when made perceptually salient by pitch prominence.
These results indicate that intonation can be attributed a role in the enhancement of serial recall that is not simply equivalent to the temporal structuring induced by pauses, as has so far been claimed. They also provide further support to the "language-oriented" view of Short-Term Memory (STM) as opposed to models of STM that propose a specialised memory system that is hierarchically organised (Farrell \& Lelièvre 2012).
We also found that particular intonation patterns can improve recall only in the input modality, i.e. when it is presented in auditory stimuli, but not in the output modality, i.e. when subjects are requested to imitate the intonation of the auditory stimuli whilst recalling. In the latter case, imitation cannot be construed as a supporting strategy for the recalling task. Instead it interfered with the task, resulting in a decrease in performance in the last part of the sequence, as memory traces become weaker.
Such result that imitation of the intonation does not enhance serial recall points to supporting the "language-oriented" models of Short-Term Memory.

\section*{Acknowledgments}

We would like to thank Ralf Rummer for helpful discussion and advice, and our student assistants Mirco Lacalandra and Gabriella Monticelli for help collecting the experimental data.

\section*{References}

Boersma, P. \& Weenink, D. (2001). Praat, a system for doing phonetics by computer. Glot International, 5(9/10), 131-151.
Boucher, V. J. (2006). On the function of stress rhythms in speech: evidence of a link with grouping effects on serial memory. Language and Speech 49(4), 495-519.
Cowan, N. Saults, J. S., Elliott, E. M., \& Moreno, M. V. (2002). Deconfounding serial recall. Journal of Memory and Language, 46, 153-177.

Farrell S. \& Lelièvre A. (2012). The dynamics of access to groups in working memory. Journal of Experimental Psychology: Learning, Memory and Cognition, 38(6), 1659-1674.
Frankish, C. (1985). Modality-specific grouping effect in short-term memory. Journal of Memory and Language, 24, 200-209.
Frankish, C. (1995). Intonation and auditory grouping in immediate serial recall. Applied Cognitive Psychology, 9, 5-22.
Gilbert, A. C. \& Boucher, V. (2007). What do listeners attend to in hearing prosodic structures? Investigating the human speech-parser using short-term recall. Proceedings of Interspeech 2007 (pp.430-433).
Gilbert, A. C., Boucher, V. \& Jemel, B. (2011). The role of rhythmic chunking in speech: synthesis of findings and evidence from statistical learning. Proceedings of the XVII International Congress of Phonetic Sciences (pp. 747-750).
Hirschberg, J \& Pierrehumbert, J. (1996). The intonational structuring of discourse. Proceedings of the \(24^{\text {th }}\) Annual Meeting of the Association of Computational Linguistics (pp.136-144).
Penney, C. G. (1979), Interactions of suffix effects, with suffix delay and recall modality in serial recall. Journal of Experimental Psychology: Human Learning and Memory, 5, 507-521.
Reeves, C., Schmauder A. R., \& Morris, R. K. (2000). Stress grouping improves performance on an immediate serial list recall task. Journal of Experimental Psychology: Learning, Memory and Cognition, 26 (6), 1638-1654.
Saito, S. (1998), Effects of articulatory suppression on immediate serial recall of temporarily grouped and intonated lists. Psychologia, 41, 95-101.
Savino, M. (2001). Non-finality and pre-finality in Bari Italian intonation: a preliminary account. Proceedings of the VII European Conference on Speech Communication and Technology (vol.2, pp.939-942).
Savino, M.(2004). Intonational cues to discourse structure in a variety of Italian. In P. Gilles \& J. Peters (Eds.), Regional Variation in Intonation, Tuebingen: Niemeyer.
Wechsler, D. (1987). Manual for WAIS-R. New York: The Psychological Corporation.

\title{
Is Double-Dipping an Alternative to Double-Dissociation?: Sampling Two Representational Systems Using a Single Task
}

\author{
Jordan Richard Schoenherr (psychophysics.lab@gmail.com) \\ Guy Lacroix (guy_lacroix@carleton.ca) \\ Department of Psychology, Carleton University \\ 1125 Colonel By Drive, Ottawa, ON K1S5B6 Canada
}

\begin{abstract}
Dual-process models of categorization posit dissociable implicit and explicit category learning systems. Evidence in favour of these accounts is typically obtained by examining how categorization responses differ over time, with differing category structures, and by changing task demands. If these two categorization systems are activated concurrently (e.g., COVIS) then both implicit and explicit representations can be examined over the course of learning even when one system dominates category response selection. In the current study, we used subjective measures of performance (i.e., confidence reports) to continuously sample from a participant's explicit representation of the category structure while also examining changes in these reports over the course of training. Using category structures that motivate the acquisition of either explicit or implicit representations, we observed differences in confidence reports that did not correspond to changes in categorization accuracy. These findings provide evidence for categorization systems that contain different representations.
\end{abstract}

Keywords: dual-process, categorization, confidence
processing

\section*{Introduction}

Dual-process models of categorization assume that information is processed by and represented in independent cognitive systems. For instance, one such model, RULEX (Nosofsky, Gluck, Palmeri, McKinley, \& Glauthier, 1994), postulates that people categorize objects by using simple rules and by memorizing the exceptions to those rules. Similarly, another model, ATRIUM (Erickson \& Kruschke, 1998) assumes that categorization involves the combination of rule-based and exemplar-based processes whose relative contributions are mediated by an attentional gating mechanism. An alternative account provided by Love, Medin, \& Gureckis' (2004), SUSTAIN, assumes that instances of a category are stored as clusters of feature associations and these clusters are associated with a category in the context of both supervised and unsupervised learning. Moreover, the goals of the participant will also determine the nature of the representations that are formed (see the Conclusion for further discussion and implications).

Following from Logan's (1988) instance theory of automaticity, Ashby, Alfonso-Reese, Turken, and Waldron's (1998) COVIS model instead assumes that there is a competition between the verbal and implicit systems responsible for the categorization process. Evidence in favour of COVIS comes from double-dissociation paradigms which demonstrated feedback and a concurrent working memory load affect the implicit and verbal systems, respectively. In addition to predictions concerning
categorization performance, COVIS also makes claims concerning post-decisional confidence reports. To our knowledge, the implications of these claims have not been examined. The present study is directed toward exploring this prediction: The correspondence between categorization accuracy and subjective confidence should change depending on the category structure that participants are required to learn.

\section*{COVIS Categorization Systems}

COVIS has two main assumptions. First, categorization is assumed to rely on a multidimensional variant of signaldetection (SDT) referred to as general recognition theory (GRT; Ashby \& Townsend, 1986). With the provision of feedback, the category boundary divides separable or integral stimulus dimensions into discrete regions of a categorical space (e.g., Ashby \& Gott, 1988; Ashby \& Maddox, 1992). If a stimulus consists of values along a dimension greater than those specified by the criterion, it is assigned to one category. If the values are less than that specified by the criterion, it is assigned to another category. Using curve fitting, Ashby and colleagues have demonstrated that by the end of training, participants performance is well described by an optimal classifier model that employs a category boundary.

The second critical feature of COVIS is the interaction of the explicit and implicit categorization systems during response selection (Ashby \& O’Brien, 2005; Ashby et al., 1998). Initially, the hypothesis-testing system which uses executive function and working memory is assumed to dominate categorization as it can rapidly generate and test explicit, one-dimensional (rule-based) representations. Simultaneously, the implicit procedural learning system begins to associate regions of perceptual space with a category label though it does not yet dominate category response selection. As more instances of the categories are retained in memory, the process of retrieving the stimulusresponse mapping within the implicit system becomes increasingly rapid. With sufficient training, the implicit system begins to dominate category response selection. Thus, in the absence of an executive load (e.g., Waldron \& Ashby, 2001), participants will acquire rule-based category structures earlier in the course of the experiment relative to an information-integration category structures. These findings have been taken as evidence representing a qualitative change in responding rather than merely a quantitative shift in a category boundary location within a single implicit system (Ashby et al., 1998).

A critical observation concerning Ashby et al.'s (1998) dual-process account of COVIS is that although a single response results when presented with a stimulus, the resulting perceptual information activates both categorization systems. Later in training, when an implicit representation stored within the procedural-learning is used to produce responses, an explicit representation should still be produced by the hypothesis-testing system. If a method can be adopted to examine this explicit representation over the course of training, further evidence would be provided for a dual-process account of categorization. Confidence reports might be used to sample such an explicit representation over the course of learning.

\section*{Measures of Awareness of Performance}

Confidence reports and related measures were among the earliest tools used in experimental psychology to assess participants' ability to consciously monitor their performance on a given task (for a review, see Baranski \& Petrusic, 1998). Retrospective confidence reports are typically obtained by having an individual assign a numeric value corresponding to a subjective probability (e.g., 60\%) in their belief that they have provided a correct response to a primary task. The degree of correspondence between a participant's mean accuracy and assigning a subjective probability to a response is referred to as subjective calibration (e.g., Baranski \& Petrusic, 1994). Perfect calibration requires that the proportion correct (e.g., 0.6) and mean confidence are equivalent (60\%). Typically, participants are observed to deviate from ideal performance as evidenced by miscalibration. Rather than presenting a random pattern, miscalibration occurs in a systematic form in terms of either over- or underconfidence. Overconfidence is observed when confidence exceeds accuracy. This pattern is typically observed when the task requires the use of either general or conceptual knowledge. Underconfidence is observed when accuracy exceeds confidence. This pattern is typically observed in perceptual tasks (for reviews see, Lichtenstein \& Fischhoff, 1977; Kvidera \& Koustaal, 2008). There is disagreement as to whether this pattern represents task dependencies (Lichtenstein \& Fischhoff, 1977) or whether it is a result of differential accessibility of information within the systems when performing the task (Dawes, 1980).

A consideration of confidence models reveals the sources of this disagreement. The first formal models of confidence assumed a direct-scaling of primary decision information with a decisional-locus of confidence processing (e.g., Ferrel \& McGooey, 1980; Gigerenzer, Hoffrage, \& Kleinbolting, 1991; for recent models see, Pleskac \& Busemeyer, 2010). On these accounts, confidence reports are based solely on information used by the primary decision process and consequently do not require any additional processing. Importantly for the
present study, COVIS provides a similar model of confidence. Ashby et al.'s (1998) assume that confidence reports result from activation of the prefrontal cortex associated with the response alternative by the implicit system which they claim is supported by neurological studies examining cortical modulation (e.g., Frith, Friston, Liddle, \& Frackowiak, 1991). Given the direct correspondence between the implicit representation used to categorize stimuli and that used to report confidence, Ashby et al.'s (1998) direct-scaling account of confidence predicts greater correspondence between accuracy and confidence reports in the information-integration condition. This pattern would result in high levels of confidence calibration.

Furthermore, if subjective confidence is determined by an implicit representation, then greater levels of miscalibration should be observed in the rule-based condition due to a difference between the representation used to categorize stimuli and that used to report confidence. Specifically, if an implicit representation is used to report confidence and that representation is inaccurate early in training then Ashby et al.'s (1998) account would appear to imply that underconfidence should be observed when learning rule-based category structures.

In contrast to this account, an alternative class of models assumes that confidence reports require an indirect-scaling of primary decision evidence requiring additional cognitive operations. Both a post-decisional locus (e.g., Audley, 1960; Vickers \& Packer, 1980), or an alterable locus (Baranski \& Petrusic, 1998) have been considered wherein participants process confidence following the primary decision or can additionally compute it concurrently with the primary decision. If confidence reports require a secondary set of operations, it is possible that they could be affected by information other than that available to the primary decision. This would follow from the observation that performance on any task is the result of explicit and implicit processes (Jacoby, 1991).

There is considerable support that confidence reports involve a secondary set of effortful scaling operations that either integrate information from multiple sources (e.g., perceptual and conceptual) or manipulate this information in the process of scaling (Busey, Tunnicliff, Loftus, \& Loftus, 2000; Schoenherr, Leth-Steensen, \& Petrusic, 2010). For instance, Schoenherr et al. (2010) were able to alter subjective confidence independently of the primary decision. Studies investigating metamemory have also observed that subjective awareness appears to be determined by encoding and retrieval cues rather than the number of items recalled (e.g., Koriat, Sheffer, \& Ma'ayan, 2002). Given that different sources of information can affect the primary decision and confidence reports, these studies suggest that a comparison of primary decision responses and confidence reports might be an alternative means to
dissociate implicit and explicit categorization systems (see also Dienes \& Berry, 1997).

In the context of indirect-scaling models, we can predict a different pattern of miscalibration. If we disregard the direct-scaling model adopted by Ashby et al. (1998) we can still adopt some of the assumptions of COVIS to predict an alternative pattern of overconfidence. If a hypothesis-testing system is not as dependent on feedback to learn a category structure as the procedural-learning system, negative feedback should exert less of an effect when learning rulebased category structures relative to information-integration category structures. Thus, in instances where there is category overlap which result in a performance asymptote, an explicit representation of the category structure that informs confidence reports would not reflect the proportion of negative feedback that results. This would lead to overconfidence. Greater calibration would be observed in the information-integration condition due to that system's reliance on feedback and absence of an explicit category structure to bias confidence reports.

\section*{Present Study}

The present study starts from the assumption that the degree of correspondence between measures of accuracy and confidence can be used to infer the nature of representations used at different stages of the category learning process. To accomplish this, we adopted the randomization technique used by Ashby and colleagues and required participants to provide confidence report concerning the accuracy of their responses.

Two sets of predictions can be made concerning the relationship between accuracy and confidence depending on whether a direct- or indirect-scaling account of confidence is adopted. When adopting Ashby et al's (1998) directscaling model of confidence, we can expect participants to be well calibrated in the information-integration condition due to representational correspondence between the information used within the categorization system and that used to report confidence. Conversely, the rule-based condition should produce underconfidence due to the inaccurate implicit representation used to report confidence and an accurate explicit representation used to categorize stimuli.

An alternative set of prediction follows from indirectscaling models of confidence (e.g., Baranski \& Petrusic, 1998) should also be considered. First, when participants are incapable of obtaining \(100 \%\) accuracy, such as when a performance asymptote is adopted, confidence should reach the equivalent subjective probability of this performance asymptote prior to categorization accuracy. Second, if the explicit system is not as dependent upon response feedback as the implicit system, then the proportion of negative feedback observed in the rule-based condition should not affect subjective confidence reports to the same extent as the implicit-condition. Following from this, participants should
exhibit overconfidence when the category structure is readily verbalizable but category overlap is permitted. Thus, while we would expect the same comparatively high level of calibration in the information-integration condition as Ashby et al. (1998), we expect overconfidence in the rulebased condition. We also anticipate that the requirement of confidence should also increase categorization response time if it constitutes a secondary process and that these response times should be longer in the informationintegration condition relative to the rule-based condition given the need for representational change. We do not report the successful observation of these findings here due to space limitations. Rather, we limit ourselves to changes in overconfidence bias across experimental blocks.

\section*{Experiment Method}

\section*{Participants}

Eighty-eight undergraduate students participated in the study for course credit.

\section*{Materials}

Stimuli consisted of Gabor patches varying in terms of spatial frequency and orientation. Replicating the method of earlier studies (e.g., Zeithamova \& Maddox, 2007), 40 Gabor patches were created for each category for the training phase using the randomization technique by randomly sampling values from two normal distributions. Stimulus values were rescaled into stimulus dimensions with spatial frequency given by \(f=.25+\left(x_{1} / 50\right)\) and orientation given by \(o=\mathrm{x} 2(\pi / 500)\). Using these values, stimuli were generated with the Psychophysics Toolbox (Brainard, 1997) using MATLAB R2008 (MathWorks, Matick, MA) with an \(85 \%\) performance asymptote resulting from category overlap (see Figure 1). After a categorization response was provided and a confidence report was obtained, a feedback signal was presented to indicate a participant's accuracy in completing the task. Stimuli were presented to participants using E-Prime experimental software on a Dell Dimension desktop PC.

Figure 1. Information Integration Category Structure


\section*{Procedure}

The category task procedure used the randomization technique. A training phase consisted of 10 blocks of trials
with 40 exemplars from each category, and a transfer phase consisted of 2 blocks with the same 40 exemplars from each category. Participants learned either a rule-based (1D) or an information-integration (2D) category structure. In the present experiment, participants were provided with both trial-to-trial and block feedback during the training phase. In the transfer phase, participants did not receive feedback. Before trial-to-trial feedback was provided, participants reported confidence on a 6-point Likert scale from 50 (guess) to 100 (certain) scale.

\section*{Results}

Proportion Correct. As demonstrated in Figure 2, the results of categorization accuracy replicated earlier findings: 1 D rules were learned in fewer blocks then 2D rules, \(F(1\), \(83)=6.317, M S E=.039, p=.014, \eta_{p}^{2}=.071\), and accuracy increased with the number of experimental blocks, \(F(11\), \(913)=49.167, M S E=.005, p<.001, \eta_{p}^{2}=.372\) The interaction between categorization rule and experimental block was also significant, \(F(11,913)=6.891, M S E=.005\), \(p<.001, \eta_{p}^{2}=.077\). We also found that the requirement of confidence affected category learning as it interacted with block, \(F(11,913)=2.093, M S E=.005, p=.052, \eta_{p}^{2}=.025\). Although the requirement of confidence initially produced reduced performance in the first block \((M=.703, S D=\) .140) relative to no confidence \((M=.738, S D=.131)\), participants who reported confidence in the transfer phase were more accurate \((M=.866, S D=.112)\) then those who did not \((M=.829, S D=.103)\).


Confidence Reports. Due to inter-block variability resulting from individual differences in the between-subject design, we collapsed blocks into learning phases. We examined overconfidence in early phases of training across two blocks (Blocks 1 and 2) in order to compare to the two transfer blocks (Blocks 11 and 12). Two other phases of training were also examined for comparison constituting and intermediate (Blocks 3 through 6), and late phases of training (Blocks 9 through 10).

Figure 3. Overconfidence Bias across experimental blocks.


Overall, we found that the overconfidence bias differed across the learning phases, \(F(1,77)=8.842, M S E=.085, p\) \(=.004, \eta_{p}^{2}=.103\). As expected, learning phase was also found to interact with category structure, \(F(1,77)=4.539\), \(M S E=.085, p=.036, \eta_{p}^{2}=.056\). As can be seen from Figure 3, overconfidence remained relatively constant in the information-integration condition suggesting that, in general, participants did not have access to the representation that guided their performance. In contrast, an increase in overconfidence was observed in intermediate phases of training in the rule-based condition. This finding suggests that once participants identified the 1D rule, they expected to have continual improvements in performance.

\section*{Conclusions}

In the present study, we examined confidence reports as an alternative to double-dissociation paradigms. Using the randomization technique, we sought to replicate previous findings of the categorization literature such that participants would learn 1D categorization rules in fewer blocks then 2D categorization rules due to differences in the categorization systems that retain these representations. In a confidence rating paradigm, we had participants report trial-by-trial confidence after each categorization response and compared this to their accuracy. We examined whether the correspondence between accuracy and confidence (i.e., overconfidence bias) differed between category structures as well as whether this pattern changed across experimental blocks.

The results of our experiment replicate several earlier studies within categorization and confidence processing literature. Categorization performance was found to be affected by the category structure that participants learned. We observed that participants who were required to learn the rule-based category structure reached a performance asymptote faster than those who were required to learn the information-integration category structure (e.g., Ashby et al. 1998). Moreover, response latencies decreased in fewer
blocks for participants in the rule-based condition relative to those in the information-integration condition indicating that participants could more readily acquire a stimulus-response mapping for rule-based categories relative to informationintegration categories. Furthermore, these findings conform to the predictions of dual-process accounts of categorization such as COVIS (Ashby et al., 1998) allowing us to interpret the results obtained from confidence reports in a straightforward manner.

Our analysis of confidence reports also provides new evidence for dual-process accounts of categorization. In the experiment conducted here, we observed increased overconfidence in intermediate phases of training for those participants learning a rule-based category structure relative to those who learned the information-integration category structure. In general, the miscalibration observed here suggests that the representation used to report subjective confidence and that used to respond to categorize stimuli were informed by different sources of information. Greater overconfidence in the rule-based condition suggests that the category structure that participants were explicitly aware of did not contain the stimulus variability that resulted from category overlap. We would expect such a finding if the hypothesis-testing system were less reliant on feedback and could not incorporate exceptional exemplars into the explicit representation as a consequence.

Further support for the kind of representational dissociation that we predicted stems from the findings of greater calibration in the information-integration condition. In the absence of an explicit representation, the only explicit information available to participants is the proportion of feedback they have received on a trial-to-trial basis. Given that feedback is an accurate predictor of performance, less miscalibration is likely to result. Moreover, we should not expect perfect calibration if an explicit representation might be biasing confidence responses. This would occur if confidence reports incorporated multiple sources of information (Schoenherr \& Logan, 2012) or if we additionally consider that any task is determined by both explicit or implicit processes (i.e., Jacoby, 1991).

We can also consider how these findings might be accounted for by models of categorization more generally. Although it is possible that with a sufficient number of parameters, a single-process model of categorization could account for the findings of the present study, it appears more principled to assume two independent learning systems. In terms of models that posit the retention of both rules and exemplars (e.g., Nosofsky, et al., 1994; Erickson \& Kruschke, 1998) participants should be able to retain the optimal categorization rule as well as the exceptional exemplars. In the present study, one might expect that the retention of exemplars would ensure that participants would exceed the performance asymptote. There is little support
for this pattern given that performance does not significantly differ from the performance asymptote (see Figure 2).

Given the inclusion of both a categorization and confidence processing component, COVIS provides a possible explanation of the findings of the present study. COVIS posits that the evidence accumulated within the information-integration condition should be used to report confidence. For this reason, the high level of subjective calibration in the information-integration provides evidence in support of such an account. Although it does not make explicit predictions concerning the rule-based condition, it would seem that participants should be quite accurate early in training due do rapid generation and testing of explicit rule. Participants should also exhibit underconfidence due to an inaccurate implicit representation that informs subjective confidence. As noted above, this was not observed. Thus, COVIS can account for the categorization results of the present study but does not provide a sufficient account of confidence processing. The basic assumptions of two categorization systems - one explicit and one implicit - are supported by our results.

Although in some respect similar to models that retain both rules and exemplars, SUSTAIN (Supervised and Unsupervised STratified Adaptive Incremental Network; e.g., Love \& Medin, 1998; Love et al., 2004) might be better equipped to provide an explanation of the relationship between accuracy and confidence observed in the present study. A basic assumption of SUSTAIN is that clusters of features constitute a category and that there is response competition between clusters with a bias toward simple solutions. Unlike COVIS, SUSTAIN does not provide a comprehensive account of confidence processing. Love et al. (2004) note that the number of competing alternatives should reduce participant's subjective confidence. In the rule-based condition used in the present study there should be fewer clusters competing for response selection given that rule-based category structures can be identified relatively quickly. This would give rise to greater confidence. In contrast to this, the information-integration condition should have a larger number of clusters (constituting multidimensional rules) competing for response selection thereby reducing subjective confidence. On this account, however, it is not clear why exceptions would not affect confidence reports. Namely, exceptional exemplars should suggest the selection of alternative clusters thereby increasing competition and concomitantly decreasing confidence. Without a clear formulation of confidence processing within the context SUSTAIN, speculation on the adequacy of extension to accommodate our calibration results must be limited.

One promising feature of SUSTAIN is that it does allow for unsupervised learning and influences of participants' goals while learning. In our experiment we did find some evidence of better performance in the transfer phase with the
requirement of confidence reports (see Figure 2). We might expect this pattern of results if participants were monitoring their performance and consequently desired a higher level of accuracy. Thus, when asked to provide confidence reports participants might be induced to attend to the task more so than they would otherwise.

\section*{References}

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., \& Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. Psychological Review, 105, 442-481.
Ashby, F. G., \& O'Brien, J. B. (2005). Category learning and multiple memory systems. Trends in Cognitive Sciences, 9, 83-89.
Ashby, F. G., \& Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. JEP: LMC, 14, 33-53.
Ashby, F. G., \& Townsend, J. T. (1986). Varieties of perceptual independence. Psychological Review, 93, 154179.

Baranski, J. V., \& Petrusic, W. M. (1994). The Calibration and resolution of confidence in perceptual judgements. Perception \& Psychophysics, 55, 412-428.
Baranski, J. V., \& Petrusic, W. M. (1998). Probing the locus of confidence judgments: Experiments on the time to determine confidence. JEP: HPP, 24, 929-945.
Brainard, D. H. (1997). The Psychophysics Toolbox, Spatial Vision 10, 433-436.
Busey, T. A., Tunnicliff, J., Loftus, G. R., \& Loftus, E. (2000). Accounts of the confidence-accuracy relation in recognition memory. Psychonomic Bulletin \& Review, 7, 26-48.
Dawes, R. M (1980). Confidence in intellectual vs. confidence in perceptual judgments. In E. D. Lantermann \& H. Feger (Eds .) Similarity and Choice: Papers in Honor of Clyde Coombs (pp. 327-345). Bern Hans Huber.
Dienes, Z., \& Berry, D. (1997). Implicit learning: Below the subjective threshold. Psychonomic Bulletin \& Review, 4, 3-23.
Erickson, M. A., \& Kruschke, J. K. (1998). Rules and exemplars in category learning. Journal of Experimental Psychology: General, 127, 107-140.
Ferrel, W. R., \& McGoey, P. J. (1980). A model of calibration for subjective probabilities. Organizational Behaviour and Human Performance, 26, 32-53.
Frith, C. D., Friston, K. J., Liddle, E. E., \& Frackowiak, R. S. J. (1991). Willed action and the prefrontal cortex in man: A study with PET. Proceedings of the Royal Society of London, B, 244, 241-246.
Jacoby, L. L. (1991). A process dissociation framework: separating automatic from intentional uses of memory. Journal of Memory and Language, 30, 513-541.
Koriat, A., Sheffer, L., \& Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of
learning exhibit increased underconfidence with practice. Journal of Experiment Psychology: General, 131, 147162.

Kvidera, S., \& Koustaal, W. (2008). Confidence and decision type under matched stimulus conditions: overconfidence in perceptual but not conceptual Decisions. Journal of Behavioral Decision Making, 21, 253-281.
Lichtenstein, S., \& Fischhoff, B. (1977). Do those who know more also know more about how, much they know? Organizational Behavior and Human Performance, 20, 159-183.
Logan, G. D. (1988). Toward an instance theory of automatization. Psychological Review, 95, 492-527.
Love, B. C., \& Medin, D. L. (1998a). Modeling item and category learning. In M. A. Gernsbacher \& S. J. Derry (Eds.), Proceedings of the \(20^{\text {th }}\) Annual Conference of the Cognitive Science Society (pp. 639-644). Mahwah, NJ: Erlbaum.
Love, B. C., Medin, D. L., Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. Psychological Review, 111, 309-332.
Maddox, W. T., Ashby, F. G., \& Bohil, C. J. (2003). Delayed feedback effects on rule-based and informationintegration category learning. JEP: LMC, 29, 650-662.
Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., \& Glauthier, P. (1994). Comparing models of rulebased classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). Memory \& Cognition, 22, 352-369.
Pleskac, T.J. and Busemeyer, J.R. (2010) Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. Psychological Review, 117, 864901.

Schoenherr, J. R. (2009). Mapping Internal Representations of Confidence onto Scales Varying in Range, Interval, and Number of Response Categories. Unpublished Manuscript.
Schoenherr, J. R., Leth-Steensen, C., \& Petrusic, W. M. (2010). Selective attention and subjective confidence calibration. Attenion, Perception \& Psychophysics, 72, 353-368.
Schoenherr, J. R., Logan, J., \& Larose, C. (2012). Subjective confidence of acoustic and phonemic representations during speech perception. Proceedings of the \(28^{\text {th }}\) Annual Meeting of the International Society for Psychophysics, Ottawa, Ontario, Canada.
Waldron, E. M., \& Ashby, F. G. (2001). The effects of concurrent task interference on category learning: Evidence for multiple category learning systems. Psychonomic Bulletin \& Review, 8, 168-176.
Zeithamova, D., \& Maddox, W. T. (2007). The role of visuo-spatial and verbal working memory in perceptual category learning. Memory \& Cognition, 35, 1380-1398.

\title{
Match Me if You Can: How Smart Choices are Fueled by Competition
}

\author{
Christin Schulze (c.schulze@unsw.edu.au) \\ School of Psychology, University of New South Wales \\ Sydney, NSW 2052 Australia \\ Don van Ravenzwaaij (d.vanravenzwaaij@unsw.edu.au) \\ School of Psychology, University of New South Wales \\ Sydney, NSW 2052 Australia
}

\author{
Ben R. Newell (ben.newell@ unsw.edu.au) \\ School of Psychology, University of New South Wales \\ Sydney, NSW 2052 Australia
}

\begin{abstract}
In a world of limited resources, scarcity and rivalry are central challenges for decision makers. We examine choice behavior in competitive probability learning environments that reinforce one of two strategies. The optimality of a strategy is dependent on the behavior of a computerized opponent: if the opponent mimics participant choices, probability matching is optimal; if the opponent is indifferent, probability maximizing is optimal. We observed accurate asymptotic strategy use in both conditions suggesting participants were sensitive to the differences in opponent behavior. Moreover, the results emphasize that 'irrational' probability matching can be adaptive once such competitive pressures are taken into account. The application of reinforcement learning models to the data suggests that computational conceptualizations of opponent behavior are critical to account for the observed divergence in strategy adoption.
\end{abstract}

Keywords: Decision making; Probability matching; Reinforcement learning; Evolutionary psychology; Mathematical modeling

\section*{Introduction}

Competition is a pervasive characteristic of the world plants compete for light, water and pollination; animals are in continual competition for territory, food and mating; and even as humans we are constantly competing in sports, for social standing and companionship. Considering the ubiquity of competitive pressures in virtually all aspects of our lives, their crucial impact on the development of adaptive decision strategies in a broad range of contexts may seem self-evident. And yet, prior research has concentrated on assessing the rationality of numerous choice phenomena primarily by focusing on individual decision makers in social isolation. Consequently, observed choice inconsistencies are frequently dismissed as suboptimal with little or no regard for their adaptive potential in ecologically valid settings (see e.g., Todd \& Gigerenzer, 2007).

One such extensively studied choice anomaly is the tendency to proportionately match choices to outcome probabilities in repeated binary decisions, a phenomenon known as probability matching (for a review see Vulkan,
2000). In a typical setup a decision maker repeatedly has two choice options available, one of which is the correct choice with greater probability than its alternative, e.g. \(p\left(\mathrm{~A}_{1}\right)=.7\) and \(p\left(\mathrm{~A}_{2}\right)=.3\), and correct predictions are rewarded with monetary payoffs. Assuming the outcome probabilities are stationary and irrespective of prior events or subjects' behavior, \(A_{1}\) is the superior choice option throughout and, following an initial period of probability learning, should be chosen exclusively. By contrast, the frequently observed probability (over-) matching tendency results in inferior prediction accuracies and payoffs and is therefore considered fallacious within context-independent interpretations of rational choice behavior (Vulkan, 2000).

\section*{Probability Matching in Competitive Environments}

What seems irrational in individualized context-free environments, however, can be optimal in ecologically more valid situations that take prospective social interactions into account (Gallistel, 1990; Gigerenzer, 2000). That is to say, when decision makers seek to exploit limited resources under natural circumstances (e.g. forage for food or make money), they are rarely alone but typically in fierce competition for the exploitation of these resources with other agents. The more individual agents then choose the seemingly richest resource, the smaller each one's share. In nature, this situation cannot remain stable as natural selection would favor those agents who sometimes chose options with potentially scarce resources that are exploitable under less competition (Gallistel, 1990).

Following this line of argument, it has been suggested that agents should distribute their choices among resources relative to their reward potentials, i.e. adopt a probability matching strategy, to create an equilibrated evolutionary stable situation that does not give rise to conditions selecting against it (Gallistel, 1990). Evidence for such behavior has been provided by experiments that studied groups of animals in the wild, e.g. foraging behavior of ducks on a lake (Harper, 1982) and fish in a tank (Godin \& Keenleyside, 1984).

Our aim was to examine the role of competitive pressures for the facilitation of optimal decision making in simple binary human choice contexts. Specifically, we wanted to assess potential benefits of probability matching under the premise that competitive conditions reinforce its superiority. Following the logic of natural foraging situations, we designed a choice environment in which each decision maker competes against a computerized opponent for the exploitation of a monetary resource that an indifferent 'nature' repeatedly places at one of two choice options with stable probabilities. When both agents converge on the same choice, potential rewards are split evenly between them.

In this competitive context, the success of any strategy largely depends on the behavior of the opponent. Under the assumption that the competitor is attentive towards the decision maker's choice behavior and imitates her course of action, probability matching is an optimal strategy. The prevalence of aggregative behavior in a broad range of natural group settings, e.g. flocking behavior of birds, shoaling of fish, swarming of insects and herd behavior of land animals (Allee, 1978), suggests that a strategymirroring opponent creates competitive conditions closely in line with real life ecological pressures. Thus, in one experimental condition each opponent's choice probabilities are close imitations of participant behavior which renders probability matching optimal (see appendix). In a second condition (between-subjects), each participant is paired with a computer opponent who is indifferent towards subjects' choices thereby making exclusive preference for the more profitable resource (i.e. probability maximizing) the optimal strategy. This is the case because sporadic choices by the participant to the lesser option will not tempt this indifferent opponent to replicate deviant behavior but will merely result in relinquished earning potential for the participant.

By manipulating opponent behavior as described, we created two competitive choice environments that differed solely in the extent to which participants had influence on their competitors' behavior. Thus, we can assess the role of the qualitative nature of competition for the facilitation of adaptive decision making. Given the availability of sufficient feedback (Shanks, Tunney, \& McCarthy, 2002), we predicted that choices will converge on the respective optimal strategy in both environments as learning progresses, i.e. probability matching when competing against a mimicking opponent and probability maximizing when encountering an indifferent opponent.

\section*{Models of Learning under Competition}

To shed more light on the nature of underlying learning processes within such competitive environments, we discuss the applicability of reinforcement learning models proposed for similar choice settings, e.g. learning in experimental games (Erev \& Roth, 1998), learning in the Iowa Gambling task (Yechiam \& Busemeyer, 2005) and strategy selection learning (Rieskamp \& Otto, 2006), to our experiments and outline potential adaptions of such models to account for the competitive pressures examined here. Such models typically
include assumptions regarding three main components (see e.g., Sutton, 1998): a utility function that specifies the goal of the learning problem; a learning rule which establishes propensities for each choice option; and a choice rule defining the course of action given current propensities. Here, we examine two learning models postulating different conceptualizations of the utility formation process.

Utility Function In a learning environment where an agent's primary goal is maximization of total payoffs, the utility of a choice is typically directly specified by its associated monetary reward (e.g., Rieskamp \& Otto, 2006):
\[
\begin{equation*}
u_{t}(i)=r_{t}(i) \tag{1}
\end{equation*}
\]
where \(u_{t}(i)\) corresponds to the utility of the monetary gains \(r_{t}(i)\) associated with choice \(i\) on trial \(t\), namely, in our task, 0,2 or 4 cents for no, split and full payoffs (see below). The focus on monetary gains for the evaluation of choice utilities has left systematic investigations of a wider range of factors potentially influencing this important model component largely unexplored (with few notable exceptions, e.g., Janssen \& Gray, 2012; Singh, Lewis, \& Barto, 2009). This is the case even though various additional motivational sources of utility are conceivable: e.g. avoidance of boredom associated with repetitive tasks (Keren \& Wagenaar, 1985) or task completion time (Gray, Sims, Fu, \& Schoelles, 2006). Relating this prevalent negligence to the competitive task employed here, we argue that describing utilities in terms of monetary rewards only confounds two discrete learning goals vital in this context, namely, correctly assessing the profitability of an option and attending to the competitor's choices. In fact, monetary based utilities understate the crucial role differential causes of opponent behavior play when subjects face an imitative vs. an indifferent competitor. That is to say, different opponent strategies necessitate divergent learning goals: if an opponent is identified as attentive, deciding on a course of action requires consideration of ways to influence and outsmart that other agent; if, on the other hand, the competitor is indifferent, the impact of opposing actions on one's own decisions should be strongly discounted.

Incorporating these aspects into the learning model we propose a utility function that disentangles the two learning goals present in our task and allows direct estimation of the importance decision makers attribute to the choice strategies they observe in their competitors compared to the importance they ascribe to making accurate choices:
\[
\begin{equation*}
u_{t}(i)=\left[\beta \cdot g_{t}(i)\right]+\left[(1-\beta) \cdot s_{t}(i)\right] \tag{2}
\end{equation*}
\]

Here, the utility \(u_{t}(i)\) of a choice, is expressed as the weighted sum of its accuracy \(g_{t}(i)\) ( 0 for incorrect and 1 for correct guesses) and the choice of the competitor \(s_{t}(i)\) (-1 for converging choices and 1 for incongruent choices) on any given trial. The additional free parameter \(\beta\) determines the weight a subject assigns to choosing the correct option as compared to outsmarting their competitor in terms of choosing the opposite line of action. For \(\beta=1\) subjects
value the accuracy of their choices only, whereas for \(\beta=.5\) the importance of correct choices and outwitting the competitor are weighted equally. We predict that balancing these two requirements of the task would be more pronounced when facing an imitative competitor, thus \(M_{\beta, \text { indifferent }}>M_{\beta, \text { mimicry }}\), and that learning models considering these differential challenges of the task would account for the data more thoroughly.

Updating and Choice Rule Adjustment of propensities follows a delta learning rule commonly employed in similar decision tasks (e.g., Yechiam \& Busemeyer, 2005):
\[
\begin{equation*}
q_{t}(i)=q_{t-1}(i)+\alpha\left[u_{t}(i)-q_{t-1}(i)\right] . \tag{3}
\end{equation*}
\]

Here, initial propensities towards both options are assumed to equal zero and are then gradually updated in increments of the learning rate \(\alpha\) based on the prediction error in brackets. As outcomes are mutually exclusive in our task, propensities for both options are updated simultaneously regardless of the actual choice on any given trial. An agent's probability of choosing either option is determined by these propensities following an exponential 'softmax' choice rule:
\[
\begin{equation*}
p_{t}(i)=\frac{e^{\theta \cdot q_{t}(i)}}{e^{\theta \cdot q_{t}(j)}+e^{\theta \cdot q_{t}(i)}}, \quad \theta=3^{10 \cdot c}-1 \tag{4}
\end{equation*}
\]
where the sensitivity parameter \(\theta\) governs the precision with which the preferred option is chosen. If \(\theta=0\), decisions are made at random, i.e. \(p_{t}(i)=p_{t}(j)=.5\), whereas large sensitivity parameter values \((\theta \rightarrow \infty)\) correspond to strictly deterministic choices to the option with the higher propensity. Following Yechiam \& Ert (2007), an exponential transformation of \(\theta\) was employed to allow variation of choice sensitivities between random guessing (for \(\theta \approx 0\) ) to fully deterministic (for \(\theta>700\) ) within narrow bounds of \(c\), which denotes the sensitivity constant constrained between 0 and 1 .

\section*{Method}

\section*{Participants}

Fifty (35 female) undergraduate students from the University of New South Wales (mean age 18.9, \(S D=1.2\) years) participated in this experiment in return for course credit and performance based monetary compensation.

\section*{Decision Task}

A standard probability learning paradigm involving repeated binary decisions with mutually exclusive outcomes over 500 choice trials was employed. Choice alternatives were represented by two light bulbs displayed on a computer screen and programmed to illuminate with probabilities of . 7 and .3 , counterbalanced across participants for left and right choice options. Correct predictions were rewarded with 4 cents ( 1 AUD \(=.95\) USD). Choices were made while competing against a computerized opponent and when both agents converged on the correct response, the payoff was evenly split between them, i.e. each agent received 2 cents.

\section*{Design}

We employed a \(2 \times 5\) mixed model design with opponent type (mimicry or indifferent) as between-subjects factor and trial block (five blocks of 100 trials each) as within-subjects factor. The dependent measure was the proportion of choices to the more probable choice option. For the mimicry group, the choice sequence of each opponent was computed one step ahead by equating the opponent's choice probabilities on each trial with the choice proportions the subject had displayed during the past ten trials. For example, when a participant chose the more probable option on 7 out of the past 10 trials, her opponent's probability of choosing the same option on the subsequent trial was .7. \({ }^{1}\) This algorithm creates opponent behavior that probabilistically mimics participants' choices.

By contrast, the choice sequence of each opponent for subjects in the indifferent condition was computed irrespective of participants' choices. Instead, each subject played against an opponent whose set of choice probabilities simply repeated those of an opponent encountered by another subject in the mimicry condition.

\section*{Procedure}

Subjects were asked to predict which of two light bulbs would illuminate over a series of trials while attempting to earn as much money as possible. Instructions indicated that the lighting sequence was random, i.e. no pattern or system existed which made it possible to correctly predict the outcome throughout, and that the outcome probabilities of both choice options remained constant during the entire experiment. Additionally, participants in both conditions were informed that a computerized opponent with learning abilities such as their own and no initial information about the lighting frequencies was monitoring their choices and adapting to their skill level. On each trial, predictions were made simultaneously by both participant and opponent and followed by feedback about the other agent's choice and the outcome, i.e. one light bulb lit up.

Upon completion of every block of 100 trials a self-paced pause interrupted the experiment during which block feedback was provided and a short message reminded participants that the lighting sequence was random. Subjects were told: "In this game you earned X\$. Using an optimal strategy you could have earned at least Y\$.", where X represented the actual earnings of that block and Y was computed by an optimizing algorithm (Shanks et al., 2002). This algorithm was set to probability matching in the mimicry opponent condition and probability maximizing in the indifferent opponent condition while taking both agents' actual predictions during that trial block into account. Additional incentives to improve performance on the following block were provided by informing participants that reaching optimal performance ( \(\pm\) three cents) would

\footnotetext{
\({ }^{1}\) During the first ten trials of the experiment, each opponent randomly adopted one of three possible initial strategies: random response, probability matching, or probability maximizing.
}
double their payoff, whereas suboptimal performance would result in halved earnings on the subsequent trial block.

\section*{Parameter Estimation and Model Evaluation}

We estimated parameters for each individual separately based on the models' accuracy in predicting the observed choice sequence one step ahead for each trial. That is, all models generate trial-by-trial choice probabilities for both response alternatives on the basis of subjects' prior decisions, their associated payoffs and the respective model's parameter values. Employing maximum likelihood estimation we searched for the set of parameters that maximized the summed log-likelihood of the predicted choice probabilities across trials given each participant's observed responses with an iterative particle swarm optimization (Kennedy \& Eberhart, 1995). For each individual, optimization proceeded iteratively with a total of 24 particles, 23 of which started at random positions while the final particle started at the best parameter combination from the previous iteration. Optimization terminated once the model fit did not improve further for at least five successive iterations. The following parameter bounds constrained the optimization process: \(\alpha \in[0,1]\) for the learning parameter, \(c \in[0,1]\) for the transformed sensitivity \(\theta\), and \(\beta \in[0,1]\) for the additional outsmarting parameter.

The final fit of each learning model was compared to a baseline statistical model which assumes constant and statistically independent choice probabilities across trials (see e.g., Yechiam \& Busemeyer, 2005), and hence, accounts for the data without presuming any learning. The stationary probability of choosing the more probable option pooled across all trials \(\left(p_{1}, p_{2}=1-p_{1}\right)\) is the only free parameter in this baseline model and, to account for divergent model complexities, both learning models are evaluated by comparing differences in Bayesian Information Criterion (BIC; Schwarz, 1978) statistics between learning and baseline model (see e.g., Yechiam \& Busemeyer, 2005). If a learning model is superior to the statistical baseline model, i.e. accurately describes how subjects adapt their choice behavior over time, positive \(\triangle B I C\) values result from this model evaluation.

\section*{Results}

\section*{Behavioral Data}

The mean proportion of choices to the more probable choice option for each block of 100 trials averaged across participants in the two experimental conditions is displayed in Figure 1. For inferential statistics, we conducted Bayesian analyses in addition to conventional methods of hypothesis testing to quantify evidence in favor of the null and alternative hypotheses (Wagenmakers, 2007). We assume equal plausibility for the null and alternative hypotheses a priori and report the posterior probability for the null hypothesis, denoted as \(p_{\mathrm{H} 0}^{\text {Bayes }}\), associated with each effect. A mixed model ANOVA revealed a significant main effect of trial block \(\left(F(2.37,113.8)=27.9, \quad p<.001, \quad \eta_{\mathrm{p}}{ }^{2}=.367\right.\),
\(\left.p_{\text {H0 }}^{\text {Bays }}=.00\right)^{2}\), with predictions closer to the respective optimal response strategy in the last compared to the first block of 100 trials for both groups. In the mimicry condition, subjects' choice behavior accurately approached optimal probability matching towards the last trial block ( \(M=.76\) ), whereas an indifferent competitor elicited decisions more in line with a probability maximizing strategy ( \(M=.90\) ). This adaptive divergence of learning processes is emphasized by a significant main effect of competitor type across all trial blocks \((F(1,48)=11.7\), \(\left.p=.001, \eta_{\mathrm{p}}{ }^{2}=.195, p_{\mathrm{H} 0}^{\text {Bayes }}=.03\right)\). The competition type by trial block interaction did not reach statistical significance, although the Bayesian evidence was ambiguous \(\left(F(2.37,113.8)=2.77, p=.058, \eta_{\mathrm{p}}^{2}=.055, p_{\text {H0 }}^{\text {Bayes }}=.66\right)\).


Figure 1: Mean \(\pm\) standard error proportion of choices to the more probable option averaged across trials and subjects.

Similar adaptive differences in choice behavior were also observed at an individual level, with high proportions of subjects in both conditions adopting the respective adequate rather than suboptimal strategy by the final trial block.

In sum, we have demonstrated that subjects are sensitive towards their competitors' decision strategies and modify their choices accordingly. The underlying psychological processes that lead to this adaptive divergence in strategy use, however, remain elusive from the behavioral data. Thus, we now turn to a computational modeling analysis to illuminate the determinants of emergent optimal choice behavior within competitive environments more holistically.

\section*{Modeling Data}

The parameter estimates and \(\triangle B I C\) values for the proposed learning models are compared in Table 1. The first learning model we examined defined decision utilities solely based on their associated monetary payoff and, judging by its

\footnotetext{
\({ }^{2}\) Mauchly's test indicated that the assumption of sphericity had been violated ( \(\chi^{2}(9)=63.0, p<.001\) ), therefore degrees of freedom were corrected for both conventional and Bayesian analyses using Greenhouse-Geisser estimates of sphericity ( \(\varepsilon=.593\) ).
}

Table 1: Mean and standard deviations (in parentheses) of parameter estimates and the difference in Bayesian Information Criterion ( \(\triangle B I C\) ) between statistical baseline and specified model.
\begin{tabular}{|c|c|c|c|c|c|c|c|}
\hline \multirow[t]{2}{*}{} & \multicolumn{2}{|r|}{Learning ( \(\alpha\) )} & \multicolumn{2}{|l|}{Sensitivity (c)} & \multicolumn{2}{|l|}{Outsmarting ( \(\beta\) )} & \multirow{2}{*}{\(\triangle B I C\)} \\
\hline & mimic & indifferent & mimic & indifferent & mimic & indifferent & \\
\hline Monetary utility function \(u_{t}(i)=r_{t}\) & \[
\begin{gathered}
.07 \\
(.21)
\end{gathered}
\] & \[
\begin{aligned}
& .02 \\
& (.04)
\end{aligned}
\] & \[
\begin{gathered}
.16 \\
(.20)
\end{gathered}
\] & \[
\begin{aligned}
& .29 \\
& (.29)
\end{aligned}
\] & - & - & 17.8 (31.3) \\
\hline Competition utility function \(u_{t}(i)=\left[\beta \cdot g_{t}(i)\right]+\left[(1-\beta) \cdot s_{t}(i)\right]\) & \[
\begin{gathered}
.08 \\
(.21)
\end{gathered}
\] & \[
\begin{aligned}
& .07 \\
& (.14)
\end{aligned}
\] & \[
\begin{gathered}
.43 \\
(.38)
\end{gathered}
\] & \[
\begin{aligned}
& .28 \\
& (.22)
\end{aligned}
\] & \[
\begin{aligned}
& .85 \\
& (.24)
\end{aligned}
\] & \[
\begin{aligned}
& .97 * \\
& (.05)
\end{aligned}
\] & 18.7 (33.7) \\
\hline
\end{tabular}
clearly positive average \(\triangle B I C\) score, accounts considerably better for the observed choice behavior than the stationary baseline model despite greater complexity. However, this basic utility model does not permit differentiation between the learning processes that lead to divergent choice behavior in the examined competitive environments, because estimated individual model parameters did not differ significantly between experimental groups, although the Bayesian evidence was ambiguous \((t(26.0)=1.21, p=.238\), \(p_{\mathrm{H} 0}^{\text {Bayes }}=.71\) for learning rates and \(t(42.6)=-1.87, p=.068\), \(p_{\mathrm{H} 0}^{\text {Bayes }}=.51\) for sensitivity constants).

The second learning model proposed above disentangles the two learning goals of choosing accurately, yet outsmarting the competitor by introducing an additional free parameter, \(\beta\). The differential requirements of the two competitive environments are well represented by this additional outsmarting parameter, which was significantly smaller in the mimicry condition, indicating a tradeoff between betting on the more probable option and deviating from the opposing choice behavior, compared to the indifferent group, where opponent choices were to be disregarded \(\quad\left(t(26.4)=-2.44, \quad p=.022, \quad p_{\mathrm{H} 0}^{\text {Bayes }}=.27\right)\). Parameter estimates for learning rate and sensitivity constant, again, did not differ between conditions, although the Bayesian evidence was ambiguous \((t(48)=.220\), \(p=.827, \quad p_{\text {H0 }}^{\text {Bayes }}=.82\) and \(t(38.6)=1.65, \quad p=.107\), \(p_{\mathrm{H} 0}^{\text {Bayes }}=.59\), respectively). Although the more elaborate utility evaluation model sheds light on the processes underlying the observed divergence in choice behavior, the added complexity results in \(\triangle B I C\) statistics not significantly better than those of the simpler utility model \((t(49)=-.613\), \(p=.543, p_{\mathrm{H} 0}^{\text {Bayes }}=.88\) ). Thus, despite the conceptual promise and excellent parameter fit of the more complex model, overall, the simple monetary utility model is to be preferred for its parsimony.

\section*{Discussion}

Qualitatively different competitive pressures in a binary prediction task result in adaptively divergent choice behavior on aggregate and individual learning levels. Under the influence of an indifferent opponent, resources should and were found to be exploited without consideration for the other agent's preferences, i.e. much like in classic individual
binary prediction tasks, probability matching needed to be dismissed as an inferior strategy. By contrast, the presence of an imitative opponent necessitates response allocations proportional to outcome probabilities in order to maximize payoffs. In this context, we observed an adaptive tendency towards probability matching - i.e. probability maximizing was correctly rejected as an inferior strategy.

What drives this adaptive divergence of strategy adoption in these two competitive contexts? Our evaluation of learning models suggests that the observed adaptiveness of choice behavior largely resulted from differing learning goals with respect to opponent behavior: imitative competitors require consideration for strategies that influence and outsmart these agents, whereas indifferent opponents necessitate disregard for their choices when deciding on one's own course of action. Thus, conceptualizing opponent behavior as a key factor in the evaluation of choice utilities that is traded off against the desire to choose accurately disentangles these divergent requirements while providing a good approximation of observed behavior. Yet, when modeling individual data, the additional outsmarting parameter for each decision maker increased the complexity of the model beyond its explanatory potential as indicated by the \(\triangle B I C\) score comparisons. Omitting the computational representation of opponent behavior from the model, however, resulted in parameter estimates that gave little indication of the underlying learning processes prompting decision makers to respond adaptively to qualitatively different competitive pressures. At best, within this simpler model, divergent environmental requirements were somewhat reflected in marginally decreased sensitivities for evaluated choice propensities in the mimicry competitor condition, i.e. adoption of optimal probability matching is explained in terms of greater randomness in subject's choice behavior. Attributing the observed adaptiveness of strategy use in both contexts to differences in choice rule precision appears, however, conceptually implausible, because under the influence of an imitative competitor, participants are not less sensitive towards monetary rewards per se. On the contrary, we suggest that it is the added requisite to outmaneuver the opposing agent that fuels optimal matching in this context.

Consequently, to account for core learning processes that drive adaptive choice behavior within these competitive environments, an additional representation of opponent
behavior is conceptually essential. To remedy potential disadvantages of added model complexity an interesting avenue for future research is to explore the suitability of hierarchical parameter estimation techniques, which may highlight the benefits of including an outsmarting parameter without introducing the downsides of overly complex models. The take-home message from this study is that learning to choose under uncertainty can indeed be steered by competition and thus proceed adaptively in situations where probability maximizing or matching is optimal.

\section*{Acknowledgments}

The authors would like to thank Jörg Rieskamp and Lael Schooler for insightful suggestions on developing a model of competitive reinforcement learning. This research was supported by Australian Research Council grants to BRN (DP110100797; FT110100151), and a University of New South Wales International Postgraduate award (to CS).

\section*{References}

Allee, W. C. (1978). Animal aggregations: A study in general sociology. New York: AMS Press.
Erev, I., \& Roth, A. E. (1998). Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria. The American Economic Review, 88(4), 848-881.
Gallistel, C. R. (1990). The organization of learning (1st MIT Press paperback). Cambridge, Mass: MIT Press.
Gigerenzer, G. (2000). Adaptive thinking. Oxford: Oxford University Press.
Godin, J.-G. J., \& Keenleyside, M. H. A. (1984). Foraging on patchily distributed prey by a cichlid fish (Teleostei, Cichlidae): A test of the ideal free distribution theory. Animal Behaviour, 32(1), 120-131.
Gray, W. D., Sims, C. R., Fu, W.-T., \& Schoelles, M. J. (2006). The soft constraints hypothesis: A rational analysis approach to resource allocation for interactive behavior. Psychological Review, 113(3), 461-482.
Harper, D. G. C. (1982). Competitive foraging in mallards: "Ideal free' ducks. Animal Behaviour, 30(2), 575-584.
Janssen, C. P., \& Gray, W. D. (2012). When, What, and How Much to Reward in Reinforcement Learning-Based Models of Cognition. Cognitive Science, 36(2), 333-358.
Kennedy, J., \& Eberhart, R. (1995). Particle swarm optimization. In IEEE International Conference on Neural Networks Proceedings (Vol. 4, pp. 1942-1948).
Keren, G. B., \& Wagenaar, W. A. (1985). On the psychology of playing blackjack: Normative and descriptive considerations with implications for decision theory. Journal of Experimental Psychology, 114(2), 133158.

Rieskamp, J., \& Otto, P. E. (2006). SSL: A Theory of How People Learn to Select Strategies. Journal of Experimental Psychology: General, 135(2), 207-236.

Schwarz, G. (1978). Estimating the Dimension of a Model. Annals of Statistics, 6(2), 461-464.
Shanks, D. R., Tunney, R. J., \& McCarthy, J. D. (2002). A re-examination of probability matching and rational choice. Journal of Behavioral Decision Making, 15(3), 233-250.
Singh, S., Lewis, R., \& Barto, A. G. (2009). Where Do Rewards Come From? In N. Taatgen \& H. van Rijn (Eds.), Proceedings of the 31st Annual Meeting of the Cognitive Science Society (pp. 2601-2606). Austin, TX: Cognitive Science Society.
Sutton, R. S. (1998). Introduction to reinforcement learning. Cambridge, Mass: MIT Press.
Todd, P. M., \& Gigerenzer, G. (2007). Environments That Make Us Smart. Current Directions in Psychological Science, 16(3), 167-171.
Vulkan, N. (2000). An Economist's Perspective on Probability Matching. Journal of Economic Surveys, 14(1), 101-118.
Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of \(p\) values. Psychonomic Bulletin \& Review, 14(5), 779-804.
Yechiam, E., \& Busemeyer, J. (2005). Comparison of basic assumptions embedded in learning models for experiencebased decision making. Psychonomic Bulletin \& Review, 12, 387-402.
Yechiam, E., \& Ert, E. (2007). Evaluating the reliance on past choices in adaptive learning models. Journal of Mathematical Psychology, 51(2), 75-84.

\section*{Appendix}

Expected reward proportions are defined as the weighted average of all possible outcomes resulting from nature's move and both agents' choices. When two decision makers follow the same course of action, i.e. one imitates the other, the choice probabilities of both agents are identical. Given identical choice probabilities, the agents can either converge or diverge on a choice option, and thus expected rewards can be broken down into split and full payoffs while their sum amounts to the total expected payoff proportion:
\[
\begin{align*}
& r_{S p l i t}=\left(p_{c}(H)^{2} \cdot .7+p_{c}(L)^{2} \cdot .3\right) / 2  \tag{5}\\
& r_{\text {Full }}=\left(p_{c 1}(H) \cdot p_{c 2}(L) \cdot .7\right)+\left(p_{c 1}(L) \cdot p_{c 2}(H) \cdot .3\right)
\end{align*}
\]

For each decision maker, split reward proportions are computed as the joint probability of both agents choosing the same option \(\left(p_{c}(i)^{2}\right)\) weighted by the outcome contingencies (here, .7 and .3 ) and split by two; whereas full reward proportions can be expressed as the joint probability of both players choosing different options \(\left(p_{c l}(i) \cdot p_{c 2}(j)\right)\) weighted by the outcome probabilities. Thus, total expected payoffs are maximized when both players probability match. For outcome probabilities of .7 and .3 , for example, each player's maximal total expected reward proportion equals .395 (compared with .35 for probability maximizing).

\title{
Learning Patterns in Noise: Environmental Statistics Explain the Sequential Effect
}

\author{
Friederike Schüür (fs62@nyu.edu), Brian Tam (bpt218@nyu.edu), Laurence T. Maloney (ltm1@nyu.edu) \\ Department of Psychology, 6 Washington Place \\ New York, NY 10003 USA
}

\begin{abstract}
Effects of trial history, or sequential effects, are typically observed in perceptual, motor, and decision making tasks and explained by subjects' irrational sensitivity to local patterns in stimulus history. We propose that in 2 alternative forced choice reaction time tasks (2AFC), sequential effects are a consequence a rational agent engaging in probability learning but with an inappropriate world model for 2AFC. We manipulate subjects' world model and show expected changes in sequential effects. Sequential effects are at least in part driven by subjects' beliefs about their environment.
\end{abstract}

Keywords: sequential effects; two alternative forced choice reaction time tasks; Bayesian modeling

\section*{Sequential Effects}

Subjects display sensitivity to local patterns in stimulus history in perceptual (Howarth \& Bulmer, 1956; Maloney, Martello, Sahm, \& Spillmann, 2005), motor (Cho et al., 2002; Remington, 1969; Soetens, Boer, \& Hueting, 1985), and decision making tasks (Ayton \& Fischer, 2004; Gilovich, Vallone, \& Tversky, 1985). In two alternative forced choice reaction time tasks (2AFC), for example, subjects' reaction times (RTs) depend not only on the current stimulus but also on the sequence of preceding stimuli (Cho et al., 2002; Remington, 1969), a phenomenon known as sequential effects (SQE). In addition, participants typically respond faster to an alternation of stimuli after a run of alternations compared to a repetition of stimuli after a run of repetitions (Soetens, et al., 1985). We refer to this finding as alternation bias in SQE. While alternation bias seems more common, repetition bias has been observed, too (Cho et al., 2002). We here ask what processes give rise to biased SQE.
SQE are in part determined by the time interval between subsequent stimuli (inter-trial interval). If this interval is short ( \(<500 \mathrm{~ms}\) ), SQE are driven by automatic facilitation (Bertelson, 1961; Soetens et al., 1985). Responses to repeated stimuli benefit from residual activation left by previous stimulus-response cycles and consequently, RTs to repeated stimuli are faster while responses to alternating stimuli are slower (Soetens et al., 1985). If the inter-trial interval is long ( \(>500 \mathrm{~ms}\) ), SQE are driven by subjective expectancy. Subjects use the sequence of preceding stimuli to predict the next, upcoming stimulus and consequently, a run of alternations induces expectancy for more alternations while a run of repetitions induces expectancy for more repetitions (Soetens et al., 1985).
But in typical 2AFC tasks, the sequence of preceding stimuli does not predict the next, upcoming stimulus - a
repetition of stimulus \(X\) does not increase the probability of stimulus X compared to stimulus Y . In other words, stimulus history has no predictive value and should not affect subjects' expectancy (or RTs if the inter-trial interval exceeds 500 ms ). Why do we find persistent SQE (after > 4000 trials) (Soetens et al., 1985) in 2AFC tasks?
Previously, SQE in 2AFC tasks with long inter-trial intervals were cast as instances of irrational sensitivity to local patterns in stimulus history, presumed to give rise to other suboptimal behavior, like the gambler's and hot-hand fallacy in decision making (Ayton \& Fischer, 2004; Gilovich et al., 1985). Instead, we propose that SQE effects are driven by subjects' attempts to learn the probability of occurrence of the two stimuli in 2AFC with a world model that, while ecologically plausible, does not match the true generative model of the task.
In 2AFC tasks, which out of two stimuli is going to appear on each trial is sampled from a Bernoulli distribution. With probability \(p\) one stimulus will appear and its alternative with probability \(1-p\). In common 2 AFC , probability \(p\) is constant throughout the experiment (or at least throughout an experimental block). The true generative model is thus a Bernoulli distribution with constant probability \(p\). Participants could learn probability \(p\) by using stimulus history to update estimated probability \(\hat{p}\) using, for example, Bayesian updating (Gerhard, Wolfe, \& Maloney, under review).
But participants may believe that instead, probability \(p\) changes over time. In other words, instead of a stable world with constant probability \(p\) - the true generative model of 2AFC - participants may believe in a dynamic world with changing probability \(p_{t}\). We propose that such belief could give rise to biased SQE: biased SQE are a consequence of an agent engaged in probability learning with an incorrect world model. As participants cannot possibly know the correct world model of 2 AFC prior to taking part in 2 AFC , a belief in a changing world is, while incorrect, not irrational. Under this account, SQE are a consequence of conditionally rational behavior: given one incorrect assumption - an inappropriate world model - subsequent behavior (SQE) is rational (see Green, Benson, Kersten, \& Schrater (2010) for a similar approach to explain probability matching).
Two previous studies suggested that SQE may be due an inappropriate world model and developed a computational model to explain commonly observed SQE in 2AFC (Wilder, Jones, \& Mozer, 2009; Yu \& Cohen, 2009). We developed a modified 2AFC task to test for effects of an
inappropriate world model on SQE accompanied by a Bayesian model. Participants took part in a 2AFC task before and after a training session (Figure1). Participants were instructed to press a left or right button with their left or right index finger in response to a stimulus, which appeared either left or right to central fixation. On each trial, stimuli were equally likely to appear left or right ( \(p_{L R}=0.5\) ) and were equally likely to repeat or alternate ( \(p_{R A}=0.5\) ). Crucially, \(p_{L R}\) and \(p_{R A}\) did not change over time.


Figure 1: Experimental design. a Participants completed a 2AFC task. The stimulus could appear left or right of fixation and participants were instructed to press a button with their left or right index finger. During pre- and postmeasurement, the probability of left / right and of repetition / alternation was 0.5 . During training, repetition probability was resampled on \(18 \%\) of all trials. Each change was signaled to the subjects. b The probability of repetition was resampled from a Beta-distribution biased towards repetition c (high values, green) or biased towards alternation \(\mathbf{d}\) (low values, orange).

During training, we put participants into an environment with changing \(p_{R A}\). Participants continued responding to stimuli presented to the left or right of central fixation with a left or right index finger button press but \(18 \%\) of all trials, \(p_{R A}\) was (re-)sampled from a Beta-distribution \(\mathrm{B}(a, b)\) ,which was biased either towards repetitions \((a=12, b=6)\) or alternation \((\mathrm{a}=6, \mathrm{~b}=12)\). Each change in \(p_{R A}\) was explicitly signaled to the subject. Belief in change was
induced to produce SQE and belief in a biased world after change (or biased "re-set prior") was induced to produce biased SQE (see Computational Model \& Hypotheses). We expected to find biased SQE before and after training, given the numerous reports of biased SQE in 2AFC. We aimed to change participants' bias in SQE in line with the bias they received during training - towards alternation bias for alternation training and repetition bias for repetition training. Such change would suggest that biased SQE are at least in part - driven by participants' (inappropriate) world model: manipulating their world model during training changes biased SQE in the expected direction in a post- compared to pre-measurement.

\section*{Computational Model and Hypotheses}

If participants believe that probability \(p_{t}\) changes over time, they should estimate current probability \(\hat{p}_{t}\) based on the outcome of all trials since the final change in probability and discard the outcome of all trials prior to this change. The decision which trials to include in estimating \(\hat{p}_{t}\) is easy when participant know when change happened, or alternatively, when they know the run length \(r\) since change. But in most situations, change is not explicitly signaled to participants and participants need to estimate change \(c\) or alternatively run length \(\hat{r}\) (Wilson, Nassar, \& Gold, 2010).
The full Bayesian model of probability updating in changing environments requires maintaining a distribution over all possible \(r\), which grows as participants complete more trials. In addition, participants need to estimate the hazard rate \(h_{t}\) - the probability of change on trial \(t\) - and maintain a distribution across all possible hazard rates (and functions) for optimal Bayesian probability updating in a changing world (Nassar, Wilson, Heasly, \& Gold, 2010). Probability updating can very quickly become computationally expensive if not intractable.
Nassar and colleagues (2010) developed a reduced Bayesian model to make probability-updating algorithms more tractable. They designed their model for probability updating in a changing environment with constant probability of change (or constant hazard rate \(h\) ) and in their model, trial outcomes were supposed to be generated from a normal distribution. We adapted their reduced model to fit our task with an increasing hazard rate \(h_{\hat{r}}\) and a Bernoulli distribution with \(p_{R A}\) (see Appendix for details). We used this model to compute probability estimates \(\hat{p}_{R A}\) of an agent that beliefs in a changing world (incorporated in the model) and completes a 2AFC task (incorporated in the input to the model). Based on the agent's \(\hat{p}_{R A}\) we subsequently computed his expectation \(\gamma_{t}\) (or posterior, see Appendix) for the upcoming trial. We then grouped \(\gamma_{t}\) based on preceding trial history: whether it was preceded by
three alternations AAA, three repetitions RRR, or any of the six other possible combinations: AAR, ARA, ...
In Figure 2 we plot \(1-\gamma_{t}\) grouped by preceding trial history (x-axis). The curves depend on three parameters: the probability of change \(p_{c}\) and \(a_{0}\) and \(b_{0}\) of the Betadistribution \(\mathrm{B}\left(a_{0}, b_{o}\right)\), which incorporates the agents belief in what the world is like after change prior to new, incoming evidence. Simulations show that if \(a_{0}>b_{0}\) and \(p_{c}>0\) SQE are repetition biased and if \(a_{0}<b_{0}\) and \(p_{c}>0\) SQE are alternation biased. During training, we lead participants to believe that \(p_{c}>0\) and either \(a_{0}>b_{0}\) (repetition bias group) or \(a_{0}<b_{0}\) (alternation bias group) and expected to observe a corresponding change in bias from pre- to postmeasurement.


Figure 2: Effects of an agent's belief in frequency of change (solid, dashed, and dashed-dotted lines represent increasing frequency) and bias in its environment after change (green: repetition biased / orange: alternation biased) determined by \(p_{c}, a_{0}\), and \(b_{0}\).

\section*{Methods}

\section*{Participants}

25 participants took part in the experiment ( 12 female, mean age: 22.3 years, age range: \(19-24\) years, 2 left handed) and completed a single session of 60 minutes. They were compensated for time and effort (\$10) and received an additional bonus of \(\$ 4\). Participants were told they would get rewarded for fast responses but we rewarded all participants for their fastest \(25 \%\) of all trials so they all received the same bonus, unbeknownst to them. Informed consent was obtained prior to testing. An internal ethics review board at New York University approved of experimental procedures.

\section*{Procedure and Apparatus}

Participants were randomly allocated to receive either repetition training ( \(\mathrm{N}=12\) ) or alternation training \((\mathrm{N}=13)\). They were seated at approximately 40 cm viewing distance from a \(19^{\prime \prime}\) Dell computer screen in a dimly lit room and
asked to wear BOSE QuietComfort 15 Acoustic Noise Cancelling headphones to reduce background noise and to allow them to listen to incorporated auditory feedback. The experiment was run on a Mac Mini (Mac OS X Version 10.7.5) programmed in MatLab 7.5 (http://www.mathworks.com/) and Psychtoolbox 3 (Brainard, 1997; Pelli, 1997). Participants responded by pressing the c-key with their left index finger or the m-key with their right index finger on a standard QWERTY keyboard.

\section*{Experimental design}

Participants completed a pre-training, training, and posttraining (Figure 1). During the pre- and post-training, participants took part in a 2AFC task based on the arcade game "Whack-a-Mole (elMo)!". At trial onset, participants saw a box with two holes - a gray square with two black circles equidistant from a white, central fixation cross. 250 ms after trial onset, the white fixation cross turned red and then, after an additional 250 ms , blue. Once the fixation cross had turned blue, the stimulus - Sesame Street's Elmo - appeared to the left of right of fixation with probability \(p_{L R}=0.5\) and with repetition probability \(p_{R A}=0.5\) after a time interval chosen from a truncated exponential distribution (mean \(=500 \mathrm{~ms}\), max. 2000 ms ). We chose the exponential to reduce temporal expectancy (Luce, 1991). The initial color change of the fixation cross - or count down - ensured that inter-trial intervals exceeded 500 ms to ensure that we measured subjective expectancy and not automatic facilitation (Soetens et al., 1985).
During training, participants completed the same task with one important modification. The probability of repetition \(p_{R A}\) was resampled on \(18 \%\) of all trials from a Betadistribution with \(\mathrm{a}=12\) and \(\mathrm{b}=6\) in the repetition biased group and \(\mathrm{a}=6\) and \(\mathrm{b}=12\) in the alternation biased group. Each time \(p_{R A}\) changed, this change was signaled explicitly to the subjects. The word 'CHANGE' was displayed on the gray box prior to each trial with a new re-sampled probability. Participants were not told \(p_{R A}\) after change.
We chose to manipulate the probability of repetition versus alternation, instead of the probability of left versus right, for two reasons. First, studies on SQE typically look at effects of trial history coded in as repetition versus alternations instead of left versus right (Cho et al., 2002; Remington, 1969; Soetens et al., 1985; Yu \& Cohen, 2009). Second, by manipulating \(p_{R A}\), we kept \(p_{L R}=0.5\) and therefore, any biases in SQE we find cannot be due to differences in left versus right hand action preparation, for example.

\section*{Data analysis}

We measured RTs, defined as the time interval between stimulus-onset and button press, as a measure of subjective expectancy during pre- and post-training. Trials with incorrect responses (left button press for a stimulus presented to the right of fixation and vice versa) were
removed from the data (5.8\%). We refrained from analyzing error trials, due to their small number. Reaction times were normalized for each participant and the pre- and posttraining separately. We then classified each trial according to its current stimulus and trial history. As we manipulated the probability of alternation and repetition, trials were classified according to whether a trial was a repetition R or alternation A trial and whether a trial was preceded by three alternations AAA, three repetitions RRR, or any of the six other possible combinations: AAR, ARA, ... We computed the mean RTs for each trial group and analyzed mean RTs using a \(2 \times 2 \times 2 \times 8\) mixed design ANOVA with bias as a between subject factor (repetition vs. alternation), and measurement (pre- vs. post), final event (R vs. A) and trial history (AAA, AAR, ARA, ..., RRR) as within subject factors.

\section*{Results}

We found a significant 3-way interaction between group, measurement, and final event \((\mathrm{F}(1,24)=6.39, \mathrm{p}=0.012)\). Prior to training, we found a bias towards alternations A (mean \(=-0.132, \mathrm{SE}=0.022\) ) compared to repetitions R (mean \(=0.124, \mathrm{SE}=0.024\); final event: \(\mathrm{F}(1,24)=24.26, \mathrm{p}<\) 0.001 ; Figure 3). After training, bias was different for each group (final event * training group: \(\mathrm{F}(1,24)=8.89, \mathrm{p}=\) 0.005 ). If repetition trained participants experienced an alternation, then it took them longer to respond (mean \(=\) \(0.076, \mathrm{SE}=0.063\) ), compared to alternation trained participants (mean \(=-0.092, \mathrm{SE}=0.045 ; \mathrm{t}(24)=-2.23, \mathrm{p}=\) 0.036 ). And conversely, repetition trained participants marginally significantly responded faster when they experienced a repetition (mean \(=-0.011, \mathrm{SE}=0.057\) ) compared to alternation trained participants (mean \(=0.118\), \(\mathrm{SE}=0.043 ; \mathrm{t}(24)=1.86, \mathrm{p}=0.076)\). Experiencing a dynamic environment with a bias either towards repetitions or alternations determines the bias in SQE in a subsequent stable environment.


Figure 3: Results. a Prior to training, we find alternation biased SQE. b After training, we find a change in bias from alternation to repetition in repetition trained participants (green lines). Alternation trained participants (orange lines) maintained their alternation bias. \(\mathbf{c}\) bias in SQE prior to training and \(\mathbf{d}\) bias in SQE after training averaged across trial history \(\mathrm{t}-1\) to \(\mathrm{t}-3\).

\section*{Discussion}

SQE are a pervasive phenomenon in 2 AFC and are typically explained by an irrational sensitivity to local patterns in trial history, which is supposed to give rise to other suboptimal behavior, too, such as the gambler's and hot-hand fallacy in decision making. We propose instead that SQE are conditionally rational: they arise because subjects attempt to learn the probability of occurrence of the two stimuli in 2AFC but with an inappropriate world model. Instead of constant stimulus probability, they believe in change. We trained participants in a changing world with a repetition or alternation bias and observed a change in participants' repetition or alternation bias in SQE consistent with the repetition or alternation training they received. Our data support the conclusion that biased SQE are at least in part driven by an inappropriate world model: SQE are conditionally rational.
Two previous studies proposed that participants' belief in a changing world gives rise to SQE. Yu and Cohen (2009) developed a Bayesian model of probability updating. Like our model, probability updating was based on stimulus repetitions and alternations ( \(2^{\text {nd }}\) order) in trial history. Their model produced SQE similar to the one's described in the literature, primarily Cho and colleagues (2002). Crucially, the authors did not explicitly measure the effects of training participants to believe in a particular world model on SQE and their model could account for bias in SQE only through ad hoc choice of a reset-prior (a prior belief in what the world will most likely be like after change) skewed towards, in their case, repetitions. Our results indicate that such biases can be altered by relatively small amounts of training. Wilder and colleagues (Wilder et al., 2009) also developed a Bayesian model of probability updating to explain previously observed SQE based on stimulus repetitions and alternations ( \(2^{\text {nd }}\) order) and stimulus location ( \(1^{\text {st }}\) order). Like Yu and Cohen (2009), Wilder and colleagues (2009) explained bias in SQE through ad hoc choice of a biased reset-prior. They state that bias in SQE changes from experiment to experiment, is difficult to predict, and should not be cast as part of a computational theory of SQE. Instead, it reflects attentional and perceptual mechanism. We assume they hereby mean that bias reflects automatic facilitation and not subjective expectancy, to use Soetens' et al. terminology (Soetens et al., 1985). We observed a predicted chance in bias after a manipulation of participants' world model, which speaks against this interpretation. The bias in SQE should be part of a computational theory of SQE.
Cho and colleagues (2002) conducted a 2 AFC experiment and developed a computational model to explain the repetition biased SQE they observed. Their model explains SQE as the consequence of special pattern detectors. According to the authors, subjects have two detectors: (a) a relatively simple repetition detector, which increases our expectation to observe a stimulus again when it has just occurred and (b) a more complex alternation detector which counts observed alternations and increases the expectation
to another alternation in proportion to the number of already observed alternations. But other than being able to account for their data, the model does not explain why we have certain pattern detectors and not others (Cho et al. (2002) list six possible detectors). Crucially, their model cannot easily explain why the training participants received in our experiment altered bias in SQE.
Green and colleagues took a similar approach to ours to explain a different phenomenon, namely probability matching in sequential binary decision tasks (2010). They proposed that participants' belief in an inappropriate world model for sequential binary decision tasks causes probability matching. In sequential, binary decision tasks, participants have to choose one of two options. One option has a higher probability of winning ( \(70 \%\) versus \(30 \%\) ). The optimal strategy for this task is to determine which option has a higher probability of winning and then choose that option exclusively. Instead, participants tend to choose the option with \(70 \%\) success probability \(70 \%\) of the time and its alternative \(30 \%\) of the time. While this probability matching behavior is suboptimal, Green and colleagues showed that given a particular albeit inappropriate world model for the task, probability matching is optimal. The authors asked participants to complete sequential, binary decision tasks and manipulated them to believe in different world models. This manipulation changed probability matching behavior a strong support for their claim. Probability matching is conditionally rational. We conclude similarly that biased SQE are conditionally rational.
Simpler models, exponential down-weighting of trial history (Anderson \& Carpenter, 2006), for example, can explain SQE but cannot explain the change in bias in SQE that we observed. SQE indicate that subjects are sensitive to recent but not distant trial history. The change in bias in SQE, however, indicates that subjects are sensitive to what they experienced many trials back (during training), too. Exponential down weighting of evidence cannot explain this dependence on temporally distant and at the same time recent information. One could augment a model that explains SQE by exponential down weighting of trial history with a bias but, to compete with our explanation, there would have to be a rational explanation for this bias.

Our findings thus show that an inappropriate world model at least in part gives rise to biased SQE. This shows that in 2AFC, participants try to learn the generative process of the task - the process, which determines how outcomes, in this case repetition versus alternation, are generated. Learning such a generative model is what distinguishes model-based from model-free learning, according to Daw and colleagues (Daw, Niv, \& Dayan, 2005; Doll, Simon, \& Daw, 2013; Otto, Gershman, Markman, \& Daw, 2013) and Green and colleagues (Green, Benson, Kersten, \& Schrater, 2010). We demonstrate that a seemingly simple behavioral phenomenon (SQE) is at least in part driven by model-based learning, which supports the recently proposed ubiquity of model-based learning algorithms (Doll et al., 2013).

In summary, we proposed that biased SQE are a consequence of participants' selection of an inappropriate world model for 2AFC. We manipulated participants' beliefs and observed predicted changes in bias of SQE. Our predictions were based on a Bayesian model of probability updating, which estimates probability of change and estimated run length to derive trial-by-trial estimates of the probability of observing a repetition versus alternation.

\section*{Appendix}

The predictive distribution is computed with respect to expected run length \(\hat{r}_{t}\) (Nassar, Wilson, Heasly, \& Gold, 2010). On each trial, the agent computes the probability that a change \(c\) occurred using Bayes rule:
\[
\begin{equation*}
p\left(c \mid X_{t}\right)=\frac{p\left(X_{t} \mid c\right) p(c)_{t}}{p\left(X_{t} \mid c\right) p(c)_{t}+p\left(X_{t} \mid \hat{p}_{R A, t}\right)\left(1-p(c)_{t}\right)} \tag{1}
\end{equation*}
\]

In the repetition bias group \(p\left(X_{t} \mid c\right)=\max \mathrm{B}\left(a_{0}, b_{0}\right)\) with \(a_{0}>b_{0}\) for a repetition and \(p\left(X_{t} \mid c\right)=1-\max \mathrm{B}\left(a_{0}, b_{0}\right)\) with \(a_{0}>b_{0}\) for an alternation. In the alternation bias group \(p\left(X_{t} \mid c\right)=\max \mathrm{B}\left(a_{0}, b_{0}\right)\) with \(a_{0}<b_{0}\) for an alternation and \(p\left(X_{t} \mid c\right)=1-\max \mathrm{B}\left(a_{0}, b_{0}\right)\) with \(a_{0}<b_{0}\) for a repetition. \(p(c)_{t}\) depends on current, estimated run length \(\hat{r}_{t}\) and increases with increasing \(\hat{r}_{t}\) :
\[
\begin{equation*}
p(c)=1-\left(1-p_{c}\right)^{\hat{t}_{t}} \tag{2}
\end{equation*}
\]

Change becomes more likely as participants complete more trials without intervening change (a uniform distribution has an increasing hazard function). \(p\left(X_{t} \mid \hat{p}_{\text {rep,t }}\right)\) is the predictive distribution if a change point did not occur and depends on \(\hat{r}_{t}\) and the number of alternations A and repetition R in \(\hat{r}_{t}\).
The expected or mean value of the predictive distribution is based on two possibilities: (a) a change point occurred, in which case \(\hat{p}_{R A, t}=\max \mathrm{B}\left(a_{0}, b_{0}\right)\) with \(a_{0}>b_{0}\) for the repetition bias group and \(\hat{p}_{R A, t}=\max \mathrm{B}\left(a_{0}, b_{0}\right)\) with \(a_{0}<b_{0}\) for the alternation bias group, or (b) no change point occurred, in which case the recent outcome \(X_{t}\) is used to update \(\hat{p}_{R A, t}\). If \(X_{t}\) is a repetition, the number of repetitions \(R_{t}\) in estimated run length \(\hat{r}_{t}\) is increased by one: \(R_{t}=R_{t-1}+1\). If \(X_{t}\) is an alternation then \(A_{t}=A_{t-1}+1\) and in the repetition bias group (note that \(R_{t}+A_{t}=\hat{r}_{t}\) ):
\[
\begin{equation*}
\hat{p}_{R A, t}=\max B\left(a_{0}+R_{t}, b_{0}+A_{t}\right) \tag{3}
\end{equation*}
\]
with \(a_{0}>b_{0}\) in the repetition bias group and \(a_{0}<b_{0}\) in the alternation bias group. The posterior distribution is a weighted average of these two possibilities:
\[
\begin{equation*}
\gamma_{t}=p\left(c \mid X_{t}\right) p(c)_{t}+\left(1-p(c)_{t}\right) \hat{p}_{R A, t} \tag{4}
\end{equation*}
\]

Expected run length is updated on each trial based on the probability that change occurred (in which case it is reset to one) and based on the probability that there was no change (in which case it is incremented by one):
\[
\begin{equation*}
\hat{r}_{t+1}=\left(\hat{r}_{t}+1\right)\left(1-p(c)_{t}\right)+p(c)_{t} \tag{5}
\end{equation*}
\]

\section*{Acknowledgements}

FS and LTM were funded by NIH \# EY019889.

\section*{References}

Anderson, A. J., \& Carpenter, R. H. S. (2006). Changes in expectation consequent on experience, modeled by a simple, forgetful neural circuit. Journal of Vision, 6(8), 822-835. doi:10.1167/6.8.5
Ayton, P., \& Fischer, I. (2004). The hot hand fallacy and the gambler's fallacy: Two faces of subjective randomness? Memory \& Cognition, 32(8), 1369-1378. doi:10.3758/BF03206327
Bertelson, P. (1961). Sequential redundancy and speed in a serial two-choice responding task. Quarterly Journal of Experimental Psychology, 13(2), 90-102. doi:10.1080/17470216108416478
Brainard, D. H. (1997). The Psychophysics Toolbox. Spatial Vision, 10(4), 433-436. doi:10.1163/156856897X00357
Cho, R. Y., Nystrom, L. E., Brown, E. T., Jones, A. D., Braver, T. S., Holmes, P. J., \& Cohen, J. D. (2002). Mechanisms underlying dependencies of performance on stimulus history in a two-alternative forced-choice task. Cognitive, Affective, \& Behavioral Neuroscience, 2(4), 283-299. doi:10.3758/CABN.2.4.283
Daw, N. D., Niv, Y., \& Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nature neuroscience, 8(12), 1704-1711.
Doll, B. B., Simon, D. A., \& Daw, N. D. (2013). The ubiquity of model-based reinforcement learning. Current Opinion in Neurobiology. doi:10.1016/j.conb.2012.08.003
Gerhard, H. E., Wolfe, U., \& Maloney, L. T. (under review). Anticipating motor performance: Effects of experience and abruptly imposed movement constraints on estimating success probability.
Gilovich, T., Vallone, R., \& Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. Cognitive Psychology, 17(3), 295-314. doi:10.1016/0010-0285(85)90010-6
Green, C. S., Benson, C., Kersten, D., \& Schrater, P. (2010). Alterations in choice behavior by manipulations of world
model. Proceedings of the National Academy of Sciences. doi:10.1073/pnas. 1001709107
Howarth, C. I., \& Bulmer, M. G. (1956). Non-random sequences in visual threshold experiments. Quarterly Journal of Experimental Psychology, 8(4), 163-171. doi:10.1080/17470215608416816
Luce, R. D. (1991). Response Times: Their Role in Inferring Elementary Mental Organization. Oxford University Press, USA.
Maloney, L. T., Martello, M. F. D., Sahm, C., \& Spillmann, L. (2005). Past trials influence perception of ambiguous motion quartets through pattern completion. Proceedings of the National Academy of Sciences of the United States of America, 102(8), 3164-3169. doi:10.1073/pnas. 0407157102
Nassar, M. R., Wilson, R. C., Heasly, B., \& Gold, J. I. (2010). An Approximately Bayesian Delta-Rule Model Explains the Dynamics of Belief Updating in a Changing Environment. The Journal of Neuroscience, 30(37), 12366-12378. doi:10.1523/JNEUROSCI.0822-10.2010
Otto, A. R., Gershman, S. J., Markman, A. B., \& Daw, N. D. (2013). The Curse of Planning Dissecting Multiple Reinforcement-Learning Systems by Taxing the Central Executive. Psychological Science. doi:10.1177/0956797612463080
Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. Spatial Vision, 10(4), 437-442. doi:10.1163/156856897X00366
Remington, R. J. (1969). Analysis of sequential effects on choice reaction times. Journal of Experimental Psychology, 82(2), 250-257. doi:10.1037/h0028122
Soetens, E., C, L., \& E, J. (1985). Expectancy or automatic facilitation? Separating sequential effects in two-choice reaction time. Journal of Experimental Psychology: Human Perception and Performance, 11(5), 598-616. doi:10.1037/0096-1523.11.5.598
Wilder, M., Jones, M., \& Mozer, M. (2009). Sequential effects reflect parallel learning of multiple environmental regularities. Advances in neural information processing systems, 22, 2053-2061.
Wilson, R. C., Nassar, M. R., \& Gold, J. I. (2010). Bayesian Online Learning of the Hazard Rate in Change-Point Problems. Neural Computation, 22(9), 2452-2476. doi:10.1162/NECO_a_00007
Yu, A., \& Cohen, J. (2008). Sequential effects: Superstition of rational behavior? NIPS, 1873-1880.

\title{
Infant contributions to joint attention predict vocabulary development
}

\author{
Katherine Scott (scottks@cf.ac.uk) \\ Elena Sakkalou \\ Kate Ellis-Davies \\ Elma Hilbrink \\ School of Psychology, Cardiff University Park Place, Cardiff, CF10 3AT, UK
}

\author{
Ulrike Hahn (u.hahn@bbk.ac.uk) \\ Department of Psychological Sciences, Birkbeck \\ Malet Street, London, WC1E 7HX, UK
}

Merideth Gattis (gattism@cf.ac.uk)
School of Psychology, Cardiff University
Park Place, Cardiff, CF10 3AT, UK

\begin{abstract}
Joint attention has long been accepted as constituting a privileged circumstance in which word learning prospers. Consequently research has investigated the role that maternal responsiveness to infant attention plays in predicting language outcomes. However there has been a recent expansion in research implicating similar predictive effects from individual differences in infant behaviours. Emerging from the foundations of such work comes an interesting question: do the relative contributions of the mother and infant to joint attention episodes impact upon language learning? In an attempt to address this, two joint attention behaviours were assessed as predictors of vocabulary attainment (as measured by OCDI Production Scores). These predictors were: mothers encouraging attention to an object given their infant was already attending to an object (maternal follow-in); and infants looking to an object given their mothers encouragement of attention to an object (infant follow-in). In a sample of 14month old children \((\mathrm{N}=36)\) we compared the predictive power of these maternal and infant follow-in variables on concurrent and later language performance. Results using Growth Curve Analysis provided evidence that while both maternal follow-in and infant follow-in variables contributed to production scores, infant follow-in was a stronger predictor. Consequently it does appear to matter whose final contribution establishes joint attention episodes. Infants who more often follow-in into their mothers' encouragement of attention have larger, and faster growing vocabularies between 14 and 18-months of age.
\end{abstract}

Keywords: vocabulary, maternal responsiveness, joint attention, growth curve modelling

\section*{Introduction}

An extensive body of research has identified several fundamental influences on word learning during infancy and early childhood. Some influences are broad predictors of language ability, such as maternal education and frequency of story reading in the home (Brooks \&

Meltzoff, 2008; Crain-Thorenson \& Dale, 1992). Other influences on language ability are local and specific, and as a result allow causal hypotheses about specific processes through which words are learned, such as joint attention (eg; Carpenter, Nagell \& Tomasello, 1998;). Joint attention refers to a situation in which two people share a common point of reference, such as when a mother and child both look at a toy and periodically look to one another as well, while at the same time the mother describes the toy (Mundy \& Newell, 2007). In such situations, joint attention is thought to help children identify relations between words and their referents, and in so doing, increase word learning (Baldwin, 1991; 1993; Brooks \& Meltzoff, 2008). Joint attention thus depends on adult and infant behaviours, and the conjunction of the two supports word learning (Smith, Yu \& Pereira, 2011).

Several studies of naturalistic parent-infant interactions have demonstrated that parents differ in how they interact with infants, and these differences influence word learning. These differences include the frequency of utterances, use of prescriptives, and choice of object reference whilst interacting with their infants (Akhtar, Dunham \& Dunham, 1991; Masur, Flynn \& Eichorst, 2005; Tomasello \& Farrar, 1986). Perhaps most importantly, the results of a number of longitudinal studies indicate that caregivers' sensitivity and responsiveness to their infants' focus of attention during parent-infant interactions predicts the timing of early linguistic milestones and vocabulary growth rate (e.g., Carpenter et al., 1998; Tamis-LeMonda, Bornstein \& Baumwell, 2001). This evidence suggests that caregivers are responsible for joint attention episodes: when caregivers notice their infants' focus of attention and join in, joint attention is established, and as a result, word learning is supported.

Other studies have identified robust infant characteristics that influence word learning, such as infant attentional abilities or attention style. For example, Tamis-LeMonda and Bornstein (1989) reported that 5-month-olds who habituated to a visual stimulus more quickly also had larger receptive vocabularies at 13 months. Dixon and Smith replicated this finding and demonstrated that infant temperament moderates the relation between attention and vocabulary size (Dixon \& Smith, 2008). The influence of infant attention on vocabulary size is presumed to function via joint attention, and indeed, individual differences between infants on experimental measures of gaze following predict later language (Brooks \& Meltzoff, 2005; 2008). Until recently, however, the hypothesised relation between infant joint attention and word learning was presumed through global longitudinal relations rather than observations of local, specific relations. Recent work using microanalytic techniques for investigating joint attention and word learning has given insight into both infant and maternal contributions. Infants holding an object named by the adult, the size of the object in the infant's view, and the stability of head movements during naming are predictive of word learning (Yu, Xu \& Zhu, 2011). Additionally parents' holding the object being named is only predictive of word learning when in doing so they attract their infants' attention (Yu et al., 2011). In similar micro-analytic studies of children's eye movements during word learning, systematic, selective and sustained attentional shifts have been labelled the "critical factor" (Smith \& Yu, in press; Yu \& Smith, 2012).

Our aim was to build on the studies of Yu et al., and Yu \& Smith by examining both maternal and infant contributions to joint attention, and to examine the significance of those contributions for word learning at a global, rather than local level. A lot is known about infant and maternal behaviours during constrained learning tasks but much less about their micro-level behaviours in real time such as during free play (Yu et al., 2011). On a general environmental level maternal reponsiveness has been shown to predict language outcomes (TamisLeMonda et al., 2001). Infant joint attention behaviours measured experimentally have also been shown to predict language outcomes (Brooks \& Meltzoff, 2005, 2008). We wished to assess complementary joint attention behaviours of infants and their mothers in a single situation. Our aim was to assess both infants and their mothers in a genuine and sustained learning environment. To do so we implemented a methodology to assess both maternal and infant contributions in the same naturalistic interaction concurrently by using a single micro-level behavioural coding system. Mothers and their 14-monthold infants participated in 10-minute free play interactions in a university laboratory. We coded mother and infant attention independently, using a fine-grained coding system that has been used in a number of studies of
maternal responsiveness and infant attention (Bornstein, Suwalsky, Ludemann, Painter \& Schulthess (1991). We then combined those codes to compute the likelihood that a mother encouraged attention to an object, given that her infant was already attending to it (maternal follow-in) and the likelihood that an infant attended to an object given that the mother encouraged attention to it (infant followin). Both likelihood measures were odds ratios. Maternal follow-in identified situations in which the mother followed an infant's attentional state and as a result, joint attention was established. Infant follow-in identified situations in which the infant followed the mother's attentional state, and as a result, joint attention was established. We then examined the predictive power of each variable (maternal follow-in and infant follow-in) for language development at \(14,16,17\) and 18 months (measured as productive vocabulary size).

In this study three hypotheses were examined. First, in keeping with previous work, we hypothesized that maternal follow-in at 14 months would predict language development from 14 to 18 months. Second, we hypothesized that infant follow-in at 14 months would predict language development from 14 to 18 months. Finally we considered infant follow-in to be an important indicator of the infant's active contribution to establishing joint attention, and therefore hypothesized that infant follow-in would be a better predictor of language development than maternal follow-in.

\section*{Method}

\section*{Participants}

Mothers were recruited in their third trimester of pregnancy to take part in First Steps, a longitudinal study of development from birth to 18 -months, which has since been extended to four-years, (see Ellis-Davies, Sakkalou, Fowler, Hilbrink \& Gattis, 2012). This recruitment took place in community groups. Of the 39 mother/infant dyads initially recruited, 36 infants ( 19 boys, 17 girls) were included in the analyses reported here. Exclusion criteria were failure to complete the interaction (P36 \& P38), and any referral for developmental delays (P18). Infants came from a range of socioeconomic and maternal education backgrounds. Although all infants' data was included in the analysis 4 infants missed vocabulary testing at one of the four time points. The majority ( \(\mathrm{N}=29\) ) were first language English speakers. A further 7 infants were bi-lingual or second language speakers. Monthly testing sessions took place either on campus at Cardiff University or in a local community facility. Parents were given \(£ 25\) in shopping vouchers at each of their monthly visits and a final \(£ 250\) upon completion of the study. For more information on the sample and study see Ellis-Davies, Sakkalou, Fowler, Hilbrink \& Gattis, (2012).

\section*{Procedure}

Infant Vocabulary Growth Measure. Vocabulary acquisition scores were obtained using the Oxford University Babylab U.K adaptation of the MacArthurBates Communicative Development Inventory (OCDI: Hamilton, Plunkett \& Schafer, 2000). Parents completed all sections of the OCDI when their children reached 14, 16,17 and 18 -months of age. Vocabulary production
scores for the infants were the outcome variables of interest for the current study.

Interaction Coding. During the monthly testing session at 14 -months mothers and their infants were left alone with a standard set of age appropriate toys. These "free play" interactions were filmed and the first 10 -mins of uninterrupted interaction were used for coding the variables of interest.


FIGURE 1. Scatterplots and regression lines of vocabulary production and age for all infants who met inclusion criteria.

Interactions were coded using mutually exclusive and exhaustive coding scheme as described in Bornstein, Suwalsky, Ludemann, Painter \& Schulthess (1991). Using Mangold® (2010) INTERACT software infant attention and maternal encouragement of attention behaviours were coded if they met or exceeded one second in duration. Infant codes were limited to: look to object; look to caregiver; or none/not visible. Maternal codes were limited to: physical or verbal encouragement of attention to object; encouragement of attention to caregiver; or none of the above. Interrater reliability was based on \(10 \%\) of interactions. Cohen's Kappa was .64 for infant attention and. 8 for maternal encouragement. This study
was particularly interested in the didactic behaviours of both parties. Once interactions were coded using the two modes a sequential analysis was carried out following the procedures of Bakeman and Quera (1995) and Bakeman and Gnisci (2005). This computed odds ratios, descriptive measures of effect size, for our variables. Odds ratios were computed using episodes where the target follow-in behavior commenced within 3-secs of the preceding behavior beginning. If target behaviours were more likely to be initiated during points when the corresponding behaviour was ongoing (as opposed to any other times) these odds ratios were greater than 1. The Generalized Sequential Querier program (GSEQ version 4.1.2;

Bakeman \& Quera, 1995) was used to compute these values. Infant follow-in was classified as: the infant looks to object given maternal encouragement of attention to object. Maternal follow-in then was classified as: maternal encouragement of attention to object given infant attending to object.

\section*{Results}

\section*{Productive vocabulary scores}

As reported in previous work (eg., Brooks \& Meltzoff, 2008), productive growth increased rapidly across these age ranges in our sample (Figure.1). As expected mean productive vocabulary scores across children at each age correlated strongly and positively. The mean at each data collection point was significantly different from the others with the only exception being between 16-month \((\mathrm{M}=27.58, \quad \mathrm{SD}=31.07), \quad\) and \(\quad 17\)-month \(\quad(\mathrm{M}=46.63\), \(\mathrm{SD}=55.75\) ), values.

\section*{Interaction coding results}

The sequential analysis odds ratios \((\mathrm{N}=36)\), were generated giving a probability of the examined sequence over others. Infant follow-in ( \(\mathrm{M}=1.15, \mathrm{SD}=.33\) ), and maternal follow-in \((\mathrm{M}=1.22, \quad \mathrm{SD}=.52)\), correlated positively ( \(\mathrm{r}=.38, \mathrm{p}=.02\) ). There was no significant difference between these values \(\mathrm{t}(35)=-.78, \mathrm{p}=.44)\). Individual correlations between infant follow-in and vocabulary production at each time step were between \(\mathrm{r}=.34\) and \(\mathrm{r}=.42\), and were all significant, \(\mathrm{p}<.05\). Maternal follow-in correlations were non-signifcant, ranging between \(\mathrm{r}=.07\) and \(\mathrm{r}=.23\). In order to consider the data from all measurement points together in a single overall analysis, we turned to growth curve models (GCMs).


Figures.2a \& 2.b. Line-graphs showing production scores (mean number of words) over time in months based on median split values of subjects on infant follow-in (2a) and maternal follow-in ( 2 b ) variables.

\section*{Models of productive vocabulary growth}

GCMs provide a powerful tool for investigating the impact of predictor variables on both the overall performance and accelerated growth of productive vocabulary (Mirman, Dixon \& Magnuson, 2008). We perfomed a median split on our predictor variables (see Figure. 2a \& 2b), and using the R (version 2.15 .2 [2012-10-26]) package lme4 (Bates, Maechler \& Dai, 2012) both random and fixed effects were input into the model and their predictive strength assessed. The best fitting model was a mixed effects model including random effects of participant on the intercept of a linear growth curve, coupled fixed effects of infant follow-in (median split high-low) on the intercept and on the slope; and fixed effects of other individual differences on the slope. With such models Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC), and Chi^2 are the most commonly used indices of goodness of fit. For our best fitting model these were: AIC=1311.6, BIC=1335 Chi \(^{\wedge} 2=104.89, \mathrm{p}<.001\) ) (See Table. 1 Model I1). Whereas the addition of maternal follow-in provided a significant improvement in fit over the baseline model of random effects of participant on productive vocabulary growth (see Model M1, Table.1), the addition of maternal followin provided no improvement over and above the inclusion of infant follow in. Infant follow-in is thus a better predictor of both absolute vocabulary values and rate of growth, where maternal follow-in only effects rate of growth and to a lesser extent.

\section*{Follow-up analysis}

In order to eliminate the possibility that our infant followin variable was simply a function of how often an infant looked to objects or how long they looked to objects in total or the average duration of looks to object, a simple independent samples \(t\)-test was carried out to ensure these characteristics of the low and high groups were not significantly different. No differences between high and low groups were found for: number of looks to an object, \(\mathrm{t}=-.54, \mathrm{p}=.59\); total duration of looking at an object, \(\mathrm{t}=.7\), \(\mathrm{p}=.67\); and average duration of looking at an object, \(\mathrm{t}=.48\), \(\mathrm{p}=.64\).

\section*{General Discussion}

In this paper joint attention episodes between mothers and infants were evaluated, in order to examine maternal and infant contributions to joint attention and their relations to language outcomes. The duration of joint attention episodes has previously been correlated with language development (Markus, Mundy, Morales, Delgado \& Yale 2001). It is thought to do so by helping children identify relations between objects and words (Baldwin, 1991; 1993). In this study we sought to elucidate a more specific picture of the components involved. Previous work has demonstrated maternal contributions to joint attention influence vocabulary development (Carpenter et al., 1998; Tamis-LeMonda et
al., 2001). We used naturalistic interactions to evaluate infant and maternal contributions to joint attention in comparable ways. Results from our Growth Curve Analysis models showed that both infant and maternal follow-in behaviours contributed to rate of language growth. Furthermore infant follow-in made the most sizeable contribution to the model: infants' tendency to look at objects after being encouraged to so do was strongly related to productive vocabulary size and growth over the next 4 months. It is not simply infants who look more frequently to objects, or show sustained visual attention whose productive vocabularies are higher. The follow-in variable rests upon the infant's visual attention being temporally contingent on the encouragement of their parent. The current paper provides support for a model in which it is infant and not only maternal offerings to early word learning situations that are substantial (Smith \& Yu, 2010; Yu \& Smith, 2012). For productive vocabulary at 14 months-old it is important that infants respond appropriately to mothers' encouragement of attention by attending to objects, thereby establishing joint attention.

Table 1: Table of growth curve model estimated fixed effects and model fits.
\begin{tabular}{|c|c|c|c|c|}
\hline \multirow[b]{2}{*}{Fixed Effects} & \multicolumn{4}{|c|}{Model} \\
\hline & Baseline & I1 & M1 & IM1 \\
\hline Age & 12.05 & . 34 & . 05 & 19.84 \\
\hline \begin{tabular}{l}
Age*Infant \\
Follow-in
\end{tabular} & & 7.7 & & -11.02 \\
\hline Age*Maternal Follow-in & & & 8.21 & -11.85 \\
\hline Age*Maternal followin*Infant follow-in & & & & 11.21 \\
\hline Model Fit & & & & \\
\hline AIC & 1435 & 1312 & 1329 & 1317 \\
\hline BIC & 1446 & 1335 & 1352 & 1352 \\
\hline Chi^2 & & 104.89** & 19.35** & 2.36* \\
\hline
\end{tabular}

Such results suggest that those infants who can effectively respond to and engage in shared experience are then able to efficiently attend to the subsequent maternal contributions held within it. That is during play between mother and infant, when a mother is encouraging their infants' attention to an object the infant then responds to this encouragement by sharing their mothers' attentional focus. In such situations, joint attention is achieved by the infant following into their mother's focus of attention. The positive predictive nature of maternal contributions within already established periods of joint
attention, as opposed to outside them, has been previously documented (Tomasello \& Farrar, 1986). Once infants establish a period of joint attention by orienting to a point of shared reference with their parent, given meaningful encouragement to do so, the subsequent information relayed can aid forming new associations (Waxman \& Gelman, 2009).

These findings then add to evidence that global attributes of joint engagement and specific individual differences in the protagonists contributing to it are crucial in guiding language outcome. Our results confirm and extend prior reports that individual differences in infants' abilities to respond to joint attention attempts between 6 and 18-months relate positively to vocabulary development (Morales, Mundy, Delgado, Yale, Messinger, Neal \& Schwartz, 2000). Future studies should investigate this relation further by comparing the contributions of maternal initiation and infant follow-in to joint attention.

By assessing maternal and infant contributions to joint attention in comparable ways within a single naturalistic interaction we have been able to advance beyond previous studies. Such temporally precise micro-analytic analysis of attention has been previously encouraged as a means to elucidate the processes of word learning (eg., Yu \& Smith; 2010, Yu et al., 2011). We aimed to bridge a gap between the micro-analytic measurement-techniques used previously in experimental word learning settings, and a significantly more naturalistic situation. Previous experiments judged word-learning performance on those novel items included in the experimental task. We have moved beyond this to show that global productive vocabulary (as measured by the OCDI) benefits significantly from infants being able to detect and respond successfully to attempts at joint engagement during periods of unconstrained play. Infant follow-in, over and above maternal follow-in, is a novel and valuable predictor of vocabulary development. Future research should address what quality of the episode underlies this result.

\section*{Conclusions}

As early as 14 months the selective visual attention of infants in response to maternal encouragement is predictive of both concurrent and future linguistic success. In this study we demonstrated, as reported previously, responsive caregivers impact infant vocabulary size positively. Moreover, for the first time in this study, the infant's response to joint attention was shown to account for a higher proportion of vocabulary growth.

\section*{Acknowledgments}

The First Steps longitudinal study was funded by the Leverhulme Trust, with further funding through PhD studentships from the Economic and Social Research

Council and the School Of Psychology, Cardiff University.

\section*{References}

Akhtar, N., Dunham, F., \& Dunham, P.J. (1991). Directive interactions and early vocabulary development: The role of joint attentional focus. Journal of Child Language, 18, 41-49.
Bakeman, R., \& Quera, V. (2011). Sequential Analysis and Observational Methods for the Behavioral Sciences. New York, NY: Cambridge University Press.
Baldwin, D. A. (1991). Infants' contribution to the achievement of joint reference. Child Development, 62, 875-890.
Baldwin, D. A. (1993). Infants' ability to consult the speaker for clues to word reference. Journal of Child Language, 20, 395-418.
Bates, D. M., Maechler, M., \& Dai, B. (2012). Ime4: Linear Mixed-Effect Models Using S4 Classes, R Package Version 0.999375-42. Vienna: R Foundation for Statistical Computing.
Bornstein, M. H., Suwalsky, J. T. D., Ludemann, P., Painter, K. M., \& Schulthess, K. (1991). Manual for observation and analysis of infant development and mother-infant interaction in the first four years of life. Bethesda, MD: National Institute of Child Health and Human Development, Child and Family Research.
Brooks, R. \& Meltzoff, A. N. (2005). The development of gaze following and its relation to language. Developmental Science, 8, 535-43.
Brooks, R. \& Meltzoff, A. N. (2008). Infant gaze following and pointing predict accelerated vocabulary growth through two years of age: a longitudinal, growth curve modeling study. Journal of Child Language, 35, 207-220.
Carpenter, M., Nagell, K., \& Tomasello, M. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. Monographs of the Society for Research in Child Development, 63(4), (Serial No. 255).
Crain-Thoreson, C., \& Dale, P. S. (1992). Do early talkers become early readers? Linguistic precocity, preschool language, and emergent literacy. Developmental Psychology, 28, 421-429.
Dixon, W. E., \& Hull Smith, P. (2008) Attentional focus moderates habituation-language relationships: Slow habituation may be a good thing. Infant and child development, 17, 95-108.
Ellis-Davies, K., Sakkalou, E., Fowler, N. C., Hilbrink, E. E., \& Gattis, M. (2012). CUE: The continuous unified electronic diary method. Behavior Research Methods, 44, 1063-1078.
Hamilton, A., Plunkett, K. \& Schaffer, G. (2000). Infant vocabulary development assessed with a British Communicative Development Inventory: Lower scores in the UK than the USA. Journal of Child Language, 27, 689-705.

Markus, J. A., Mundy, P., Morales, M., Delgado, C. E. F., \& Yale, M. (2000). Individual differences in infant skills as predictors of child-caregiver joint attention and language. Social Development, 9 (3), 302-315.
Masur, E.F., Flynn, V., \& Eichorst, D.L. (2005). Maternal responsiveness and directive behaviors and utterances as predictors of children's lexical development. Journal of Child Language, 32(1), 63-91.
Mirman, D., Dixon, J.A., \& Magnuson, J.S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. Journal of Memory and Language, 59(4), 475-494.
Morales, M., Mundy, P., Delgado, C. E. F., Yale, M., Messinger, D., Neal, R., \& Schwartz, H. (2000). Responding to joint attention across the 6- to 24-month age period and early language acquisition. Journal of Applied Developmental Psychology, 21, 283-298.
Mundy, P., \& Newell, L., (2007). Attention, joint attention and social cognition. Current Directions in Psychological Science, 16(5), 269-274.
Smith, L.B. \& Yu, C. (in press). Visual Attention is not enough: Individual differences in Statistical WordReferent Learning in Infants. Language Learning and Development.
Smith, L.B., Yu, C., \& Pereira, A. F. (2010). Not your mother's view: the dynamics of toddler visual experience. Developmental Science, 1-9.
Tamis-LeMonda, C. S., Bornstein, M. H., (1989). Habituation and maternal encouragement of attention in infancy as predictors of toddle language, play and representational competence, Child Development, 60, 738-751.
Tamis-LeMonda, C. S., Bornstein, M. H., \& Baumwell, L. (2001). Maternal Responsiveness and Children's Achievement of Language Milestones. Child Development, 72 (3), 748-767.
Tomasello, M., \& Farrar, MJ. (1986). Joint attention and early language. Child Development, 57(6), 1454-1463.
Waxman, S.R., \& Gelman, S.A. (2009). Early wordlearning entails reference, not merely associations. Trends in Cognitive Sciences. 13(6), 258-263.
Yu, C., \& Smith, L. (2010). What you learn is what you see: using eye movements to understand infant crosssituational statistical learning Developmental Science, 1-16.
Yu, C., \& Smith, L. (2012). Hypothesis testing versus associative learning in cross-situational word-referent learning: Prior questions. Psychological Review, 119, 21-39.
Yu, C., Xu, J-M., \& Zhu, X. (2011). Word learning through sensorimotor child-parent interaction: A feature selection approach. In L. Carlson, C. Holscher, \& T. Shipley (Eds.), Proceedings of the \(33^{\text {rd }}\) Annual Conference of the Cognitive Society. Austin, TX: Cognitive Science Society.

\title{
Intentionality, Evaluative Judgments, and Causal Structure
}

\author{
Jason Shepard (jason.s.shepard@emory.edu) Phillip Wolff (pwolff@emory.edu) \\ Department of Psychology, 36 Eagle Row \\ Atlanta, GA 30322 USA
}

\begin{abstract}
The results from a number of recent studies suggest that ascriptions of intentionality are based on evaluative considerations: specifically, that the likelihood of viewing a person's actions as intentional is greater when the outcome is bad than good (see Knobe, 2006, 2010). In this research we provide an alternative explanation for these findings, one based on the idea that ascriptions of intentionality depend on causal structure. As predicted by the causal structure view, we observed that actions leading to bad outcomes are associated with negative social pressures (Experiment 1), that these negative pressures give rise to a specific kind of causal structure (Experiment 2), and that when these causal structures are pitted against the badness of the outcome, intentionality judgments track with causal structure and not badness (Experiment 3). While the badness of an outcome may have an indirect effect on judgments of intentionality, our results suggest that the factors that affect judgments of intentionality most directly are non-evaluative and objective.
\end{abstract}

Keywords: Social cognition, Folk psychology, Theory of mind, Intentional action, Intentionality, Causal structure, Morality, Norms, Side-effect effect, Knobe effect.

\section*{Introduction}

Ascriptions of intentionality play a fundamental role in our explanations of behaviors (Malle, Moses, \& Baldwin, 2001). They influence our judgments of character (Rotenberg, 1980), deservedness of blame or praise (Lagando \& Channon, 2008), the impermissibility of actions (Cushman 2008), and the severity of deserved punishment (Horan \& Kaplan, 1983). Standard accounts of intentionality ascription hold that judgments of intentionality are based on objective or descriptive properties of the actors and the situation, such as foreseeability, desire, and belief (Guglielmo, Monroe, \& Malle, 2009; Knobe \& Malle, 1997; Mele \& Sverdlik, 1996; Sripada, 2010). Recent empirical work by Knobe (2003a) and others (Nadelhoffer, 2006; Wright \& Bengson, 2009; Cova \& Naar, 2012) raises an alternative view, that ascriptions of intentionality may be based on evaluative properties of a situation. Specifically, Knobe \((2006,2010)\) has argued that the likelihood of viewing a person's actions as intentional is greater when the outcome is bad than when it is good. In this paper, we offer a critical test of this proposal. We also put forward and test another possibility, that judgments of intentionality are most directly based on the causal structure of a situation, which can be influenced at times by evaluative considerations. In a series of three experiments, we show that the phenomenon originally observed in Knobe (2003a) and others is more
directly explained in terms of causal structure than badness of the outcome.

\section*{The Side-Effect Effect (or Knobe effect)}

A connection between intentionality and badness has been demonstrated in research examining the so-called side-effect effect, or Knobe effect. Experiments investigating this effect have typically included two main conditions. In the harm condition, participants read scenarios like the following:

The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment.'
The chairman of the board answered, 'I don't care at all about harming the environment. I just want to make as much profit as I can. Let's start the new program.
They started the new program. Sure enough, the environment was harmed.

After reading the scenario, participants are asked "Did the chairman intentionally harm the environment?" For this scenario, Knobe (2003a) found that \(82 \%\) of the participants responded that the chairman intentionally harmed the environment. In help conditions, everything is kept the same except the side-effect is described as good. In the chairman scenario, for example, participants were told that the business plan would not only make a profit but also help the environment. Interestingly, in this alternative condition, only \(23 \%\) of the participants felt that the chairman intentionally helped the environment (Knobe, 2003a). This basic finding has been replicated with other scenarios (Knobe, 2003b; Knobe, 2007; Knobe \& Mendlow, 2004; Mallon, 2008; Nadelhoffer, 2004; Nadelhoffer, 2006; Uttich \& Lombrozo, 2010; Wright \& Bengson, 2009) and in a diverse array of populations, including Hindi speakers when the scenarios are translated into Hindi (Knobe \& Burra, 2006), with four-year old children (Leslie, Knobe, \& Cohen, 2006), with participants who suffer from deficits in emotional processing due to lesions in the ventromedial prefrontal cortex (Young, Cushman, Adolphs, \& Hauser, 2006), and with adults with high functioning autism or Asperger's (Zalla \& Leboyer, 2011). The wide range of situations and populations supports the conclusion that the basic pattern of findings is both reliable and conceptually significant, but do these findings really demonstrate that
intentionality is directly dependent on the badness of the outcome?

\section*{Responses to the Evaluative Accounts}

Though many take the side-effect effect as evidence that ascriptions of intentionality are affected by evaluative considerations (Nadelhoffer, 2004; Nado, 2008; Wright \& Bengson, 2009) such as badness, others have attempted to provide accounts of the side-effect effect that are consistent with standard, descriptive models of intentionality. For example, Adams and Steadman (2004) has argued that the side-effect effect is a result of pragmatic implicature: People assent to the statement that the chairman intentionally harmed the environment not because they genuinely attribute intentionality to the chairman but because they do not want to imply that the chairman is not responsible for harming the environment. Machery (2008) has argued that the effect is the result of calculation of trade-offs, arguing that people are more likely to attribute intentionality whenever there is a trade off and that the higher rates of attributions for outcomes that involve a trade-off occur regardless of the evaluative status of the trade-off. Sripada (2010; 2012; Sripada \& Konrath, 2011) has argued that the side-effect effect can be explained in terms of concordance with deeply held attitudes and values attributed to the actor. For example, in the chairman case, since the chairman states that he doesn't care about the environmental outcome in both the harm and the help case, people take this as evidence that the chairman harbors anti-environmental attitudes. However, only in the harm case does the outcome concord with anti-environmental attitudes. It's this concordance that explains the side-effect effect, not the evaluative status of the outcome (see Hughes \& Trafimow, 2012, for a similar account). However, all these accounts have met explanatory or experimental challenges (e.g., Cova \& Naar, 2012; Knobe, 2003b, Knobe, 2004; Mallon, 2008; Phelan \& Sarkissian, 2009; see Nadelhoffer, 2011 for a review).

Uttich and Lombrozo (2010) have argued that the sideeffect effect results from the fact that behavior that conforms to a norm is less informative about the underlying mental states than is behavior that violates a norm. According to their account it is the fact that a norm was violated, not any particular evaluative judgment, that leads to higher rates of attribution of intentional action. Furthermore, they argued that the higher rates of attribution for norm violations is not specific to violations of evaluatively-laden moral norms but that the phenomenon is a feature of norm violations more generally, including statistical and prudential norms. However, their account does not provide a mechanism that allows us to differentiate why some norm violations (e.g., harming versus helping the environment) lead to larger asymmetries in intentional action attributions than other norm violation (e.g., violating an industry standard versus conforming to an industry standard (see Uttich \& Lombrozo, 2010, Experiment 1)) or why some norm violation do not lead to an asymmetry at all
(e.g., a supervillain who violates supervillain social norms (see Uttich \& Lombrozo, 2010, Experiment 2)). In a slogan, their account can be viewed to being committed to the claim: "A norm is a norm is a norm." But the fact of the matter is when it comes to the side-effect effect, not all norms are equal. Some norm violations are more informative than others. What explains this fact? Furthermore, their evidence is still compatible with the claim that badness (or some other evaluative judgment) best explains the pattern of asymmetries observed in various side-effect effect cases.

\section*{Causal Structure Account}

Though the demonstrations of the side-effect effect suggest that evaluations of the badness of the outcome may enter into our reasoning about intentionality, another possibility is that in Knobe's experiments, more may have been varied than the valence of the outcome. In particular, in varying the badness of the outcome, Knobe may have also varied the causal structure of the scenarios, and it may have been the causal structure, and not the valence of the outcomes per se, that affected people's ascriptions of intentionality. According to this alternative view, the extra something that determines whether actions are judged as being intentional, is whether, in fact, the agent causes himself to produce the action.

The claim here begins with the idea that, in situations such as the harm version of the chairman case, there is normative force that acts on the chairman, and this normative force puts a pressure on the chairman not to pursue behavior that would knowingly violate the norm. In other words, the presence of the norm puts a preventive pressure on the chairman not to harm the environment. This establishes a preventive causal relationship: namely, Norm PREVENTS chairman from harming the environment.

However, in spite of the presence of this PREVENT relationship, the chairman overcomes that pressure and engages in behavior that will knowingly harm the environment. In overcoming the normative pressure, an additional PREVENT relation is formed: The chairman PREVENTS the norm. This string of PREVENT relations establishes what is known as a double prevention (Collins, 2000; Dowe, 2001; Hall, 2004; Schaffer, 2000). In this case, the chairman PREVENTS the norm from PREVENTING him from harming the environment. Prior research has shown that double preventions are interpreted as instantiating CAUSE or ALLOW relationships (McGrath, 2005; Wolff, Barbey, \& Hausknecht, 2010). In this case, the pattern would result in the CAUSE relationship: The chairman causes himself to harm the environment. Interestingly, this "causing of one's self" instantiates a situation in which an actor acts on himself, making it reflexive. This self-causation is particular important for intentionality. To intentionally do something is, if nothing else, to cause oneself to do that something. Importantly, the double prevention, and thus the reflexive relationship, will arise in the harm scenario but not in the help scenario
because there is no preventive pressure against helping the environment, only against harming the environment.

Three experiments were conducted to test this causal structure account against the badness account.

\section*{Experiment 1}

In this experiment we investigated whether the causal structure view could provide an alternative explanation for the results found in previous studies. Just as in previous studies, participants read scenarios in which an actor brought about a good or bad side-effect and then they rated the intentionality of the actor. The scenarios included six scenarios that have been used in other studies in the literature, as well as 10 new scenarios created specifically for this experiment. We were interested in whether we could replicate the intentionality effect observed in other studies. The current experiment also tested one of the main predictions of the causal structure account, that people would infer more preventive pressure in harm scenarios than help scenarios.

\section*{Methods}

Participants Forty-eight Emory University undergraduates participated for course credit.
Materials The materials were 16 scenarios modeled after the chairman scenario described above. For each scenario, there was a HARM version in which the side effect violated a norm and a HELP version in which the side effect did not violate a norm.
Procedure Participants read eight of the 16 scenarios. Each participant received only one version of each scenario they read, either the HARM or HELP version of each scenario. In response to each scenario, participants provided an intentionality rating by indicating their agreement or disagreement with statements of the form "The [primary actor] intentionally [side effect]". Participants also provided preventive pressure ratings by indicating their agreement or disagreement with statements of the form "Knowing that [going forward with the proposed plan of action] would [side effect] put pressure on [primary actor] to not [go forward with the plan]. The two types of ratings were presented in random order. All ratings were made on a scale that ranged from -4 complete disagreement to +4 complete agreement.

\section*{Results and Discussion}

As shown in Figure 1, participants were more willing to say that an actor intentionally brought about the side effect in the HARM condition ( \(M=1.01, S D=1.97\) ) than in the HELP condition ( \(M=-1.71, S D=1.80\) ), \(t(47)=7.91, p<\) . 001 , thus replicating, with a much wider range of materials, the phenomenon originally reported in Knobe (2003a). Of central importance to the causal structure hypothesis, participants rated the preventative pressure on the actor not to act as greater in the HARM condition ( \(M=1.56, \mathrm{SD}=\) 1.85 ) than in the HELP condition ( \(M=-2.28, \mathrm{SD}=1.49\) ), \(t(47)=9.38, p<.001\).


Figure 1: Mean ratings of intentionality and pressure by scenario (HARM vs. HELP) in Experiment 1. \({ }^{* * *} p<.001\)

\section*{Experiment 2}

Experiment 1 established that people infer a greater pressure against bringing about the side effect in HARM than HELP scenarios. With the first link of the double present established, this entails that a double prevention would be established in any case where the first PREVENT relation is overcame. Thus, Experiment 1 establishes that when people knowingly bring about a bad side effect, they infer a sequence of PREVENT relations, or otherwise, a double prevention. Double preventions typically lead to CAUSE or ALLOW relationships (McGrath, 2005). In the context of the scenarios, this implies that people should be more willing to say that the actor either caused or allowed the side effect in the HARM condition than in the HELP condition. In Experiment 2, we tested the prediction that people will be more likely to say that the actor caused the side effect in the HARM scenarios than in the HELP scenarios.

\section*{Methods}

Participants Fifty-two Emory University undergraduates participated for course credit.
Materials and Procedure The materials were the same as in Experiment 1. As in Experiment 1, participants provide ratings of intentionality. Unique to the present experiment, participants also provided ratings of causation by indicating their agreement or disagreement with statements of the form "The [primary actor] caused [side effect]."

\section*{Results and Discussion}

As shown in Figure 2, participants were more likely to say that the actor caused the side effect in the HARM condition ( \(M=1.42, S D=1.97\) ) than in the HELP condition ( \(M=-\) \(.06, S D=1.96), t(51)=5.94, p<.001\). The basic asymmetry in intentional action attributions was replicated in this experiment with participants more likely to attribute intentional action in the HARM condition ( \(M=.40, S D=\) 2.22) than in the HELP condition ( \(M=-1.98, S D=1.84\) ), \(t(51)=6.35, p<.001\).


Figure 2: Mean ratings for cause and intentionality judgments by scenario type (HARM vs. HELP) for Experiment 2. \({ }^{* * *} p<.001\)

\section*{Experiment 3}

The results from Experiments 1 and 2 support the causal structure account but are not conclusive because they remain compatible with the badness account. In order to test between badness and causal structure, the typical alignment between badness and causal structure needs to be reversed. Such a re-alignment was achieved in the current experiment by using scenarios in which an actor either violated or conformed to an unjust rule or law. Such scenarios instantiate situations in which there is a preventive pressure against doing a good thing, and little or no pressure against doing a bad thing.

Take for example the following VIOLATE scenario used in Experiment 3 (brackets indicate changes in wording for the CONFORM condition):

In Midwestern America, there was a church that had explicit rules against interracial couples participating in any church-sponsored activity.

One day a church deacon was considering which of his friends to invite to perform in a concert celebration being sponsored by the church. He decided to invite the Smiths, a husband and wife duet.

Upon hearing the news a fellow church member went to the deacon and said, "By inviting the Smiths, you will be violating [conforming to] the church's rules against interracial couples participating in any church-sponsored activity."

The deacon answered, "Look, I know I will be violating [conforming to] the church's rules against interracial couples participating in any churchsponsored event, but I don't care one bit about that. I just want to invite the most talented people to
perform in the concert. I am going to invite the Smiths to perform."

The deacon invited the Smiths to perform.
In cases like this, it appears the right thing to do would be to violate the rule. Thus, when the deacon violates the rule it is likely that people will judge the violation as good. However, since the rule is in force, it is likely that people will judge that the deacon is under pressure not to violate the rule. In contrast, when the deacon conforms to the rule, it is likely that people will judge the conforming to the rule as being bad, yet it is likely that people will not judge that the deacon is under pressure not to conform. If this is right, then the badness account and the causal structure account make opposite predictions for intentionality judgments in cases like the church case. The badness account predicts that intentionality judgments should be higher for the conform cases than for the violate cases because it is bad to conform but good to violate. The causal structure account predicts that intentionality judgments should be higher for the violate cases than for the conform cases because the actor is under pressure not to violate but is not under pressure not to conform. Additionally, because overcoming a preventive pressure instantiates a double prevention, and since the pattern of double preventions observed in these cases should lead to the inference of a reflexive causal relationship, the causal account also predicts that people will be more likely to say that the actor causes himself to violate the law or rule than to say that the actor causes himself to conform to the law or rule.

\section*{Methods}

Participants Thirty-two Emory University undergraduates participated for course credit.
Materials The materials were 8 scenarios modeled after the church scenario. For each scenario, there was a VIOLATION version in which the decision would violate an unjust rule or law, and a CONFORM version in which the decision would conform to an unjust rule or law.
Procedure Participants were assigned to read one version of each scenario. For each scenario, participants were asked to rate their level of agreement with statements about whether the primary actor acted intentionally and whether the primary actor experienced preventive pressure, as in Experiments 1 and 2 . Unique to the current experiment, participants rated their level of agreement with statements that the primary actor caused himself to bring about an effect by responding to statements of the form "The [primary actor] caused himself to [side effect]." As a manipulation check, participants also rated whether the outcome was good or bad, by responding to statements of the form, "How good or bad is [the occurrence of the side effect]," on a scale from -4 very good to 4 very bad.

\section*{Results and Discussion}

As is shown in Figure 3, Participants were more likely to say that an actor was under pressure not to violate an unjust rule or law ( \(M=1.11, S D=1.89\) ) than to conform to an unjust rule or law ( \(M=-1.31, S D=1.91\) ), \(t(31)=5.24, p<\) .001. Participants were also more likely to say that the outcome was bad when the unjust rule was conformed to ( \(M\) \(=.70, S D=1.70\) ) than when the unjust rule was violated ( \(M\) \(=-1.56, S D=1.31), t(31)=-5.49, p<.001\). These results lead the badness account and the causal structure account to make opposite predictions for intentionality judgments in these cases. The causal structure account predicts that intentionality judgments will be higher in the violate cases, while the badness account predicts that intentionality judgments will be higher in the conform cases. As predicted by the causal structure account, participants were more likely to say that the actor intentionally violated the rule or law ( \(M=1.67, S D=2.35\) ) than conformed to the rule or law ( \(M=-.62, S D=2.01\) ), \(t(31)=5.94, p<.001\). Additionally, as predicted by the causal structure account, participants were more likely to judge that the actor caused himself to violate the rule or law ( \(M=1.98, S D=1.74\) ) than conformed to the rule or law ( \(M=.34, S D=1.82\) ), \(t(31)=\) 4.91, \(p<.001\).


Figure 3: Mean ratings for pressure, badness, intentionality, and reflexive cause judgments by scenario type (HARM vs. HELP) for Experiment 3. \({ }^{* * *}\) p \(<.001\).

To provide converging statistical evidence that causal structure and not badness of outcome was leading to the asymmetrical judgments of intentional action, a regression analysis was conducted with intentionality judgments regressed on pressure judgments, reflexive-causation judgments and badness judgments. After controlling for the other variables in the model, pressure judgments and selfcausation judgments significantly predicted intentionality judgments, \(\beta\) 's \(\geq .349\), \(p\) 's \(<.005\). However, intentionality judgments and badness judgments were unrelated, \(\beta=-\) \(.095, p=.406\).

\section*{General Discussion}

The results from Experiments \(1-3\) support the conclusion that ascriptions of intentionality are driven by the causal structure rather than badness of the outcome. When assessing causal structure, it appears that people may look for reflexive causal relationships, that is, causal relations in which a person causes herself to do something.

Differences in causal structure are descriptive differences, not evaluative differences. Thus, our findings are compatible with the standard, descriptive views concerning the way we reason about mental states. However, it may be objected that the causal structure account is not fully compatible with the standard, descriptive views because, according to our account, causal structure is sensitive to norms. While we acknowledge that prohibitory norms are typically viewed as providing a force against behavior, it is not really any particular evaluative judgment that is providing this force (above and beyond the non-evaluative recognition that a norm is in effect). In Experiment 3, we demonstrated that force against behavior is perceived even in cases in which it is thought to be good to violate the norm. Additionally, even though our experiments focused on norm violations of different kinds, we don't think the effect is particular to norm violations. We would expect there would be higher rates of intentionality attribution for any situation involving a preventive, whether this force is due to norms or physical resistance.

Though we have shown that ascriptions of intentionality depend on causal structure, our results point to the possibility that the concept of intentionality itself may be little more than a kind of causal structure. Such a result might help explicate the curious role that intentionality seems to play in our understanding of causal relationships in general.

\section*{Acknowledgments}

We thank Tyler Brown, Ted Guio, and Lee Kugelmann for their assistance with stimulus preparation and data collection. This research was supported by a graduate fellowship from the Neuroethics Scholar Program.

\section*{References}

Adams, F. \& Steadman, A. (2004). Intentional action in ordinary language: Core concept or pragmatic understanding? Analysis, 64, 173-181.
Collins, J. (2000). Preemptive prevention. The Journal of Philosophy, 97, 223-234.
Cova, F. \& Naar, H. (2012a). Side-effect effect without side effects: The pervasive impact of moral considerations on judgments of intentionality, Philosophical Psychology, 25, 837-854.
Cova, F. \& Naar, H. (2012b). Testing Sripada's deep self model. Philosophical Psychology, 25, 647-659.

Cushman, F. (2008). Crime and Punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. Cognition, 108, 353-380.
Dowe, P. (2001). A counterfactual theory of prevention and "causation" by omission. Australasian Journal of Philosophy, 79, 216-226.
Guglielmo, S., Monroe, A., and Malle, B. F. (2009). At the heart of morality lies folk psychology. Inquiry, 52, 449466.

Hall, N. (2004). Two concepts of causation. In J. Collins, N. Hall, \& L. Paul (Eds.), Causation and counterfactuals (pp. 225-276). Cambridge, MA: MIT Press.
Horan, H. D., \& Kaplan, M. F. (1983). Criminal intent and consequence severity: Effects of moral reasoning on punishment. Personality and Social Psychology Bulletin, 9, 638-645.
Hughes, J. \& Trafimow, D. (2012). Inferences about character and motive influence intentionality attributions about side effects. British Journal of Social Psychology, 51, 661-673
Knobe, J. (2003a). Intentional action and side effects in ordinary language. Analysis, 63, 190-193.
Knobe, J. (2003b). Intentional action in folk psychology: An experimental investigation. Philosophical Psychology, 16, 309-323.
Knobe, J. (2004). Intention, intentional action and moral considerations. Analysis, 64, 181-87.
Knobe, J. (2006). The concept of intentional action: A case study in the uses of folk psychology. Philosophical Studies, 130, 203-231.
Knobe, J. (2007). Reason explanation in folk psychology. Midwest Studies in Philosophy, 31, 90-106
Knobe, J. (2010). Person as scientist, person as moralist. Behavioral and Brain Sciences, 33, 315-365.
Knobe, J. \& Burra, A. (2006). The folk concepts of intention and intentional action: A cross-cultural study. Journal of Cognition and Culture, 6, 113-132.
Knobe, J. \& Mendlow, G (2004). The good, the bad, and the blameworthy: Understanding the role of evaluative reasoning in folk psychology. Journal of Theoretical and Philosophical Psychology, 24, 252-258.
Lagnado, D. A. and Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. Cognition, 108, 754-770.
Leslie, A., Knobe, J., \& Cohen, A. (2006). Acting intentionally and the side-effect effect: Theory of mind and moral judgment. Psychological Science, 17, 421-427.
Machery, E. (2008). The folk concept of intentional action: Philosophical and experimental issues. Mind \& Language, 23, 165-189.
Malle, B. F. \& Knobe, J. (1997). The folk concept of intentionality. Journal of Experimental Social Psychology, 33, 101-121.
Malle, B. F., Moses, L., and Baldwin, D. (2001). The significance of intentionality. In B. Malle, L. Moses, and D. Baldwin (Eds.), Intentions and Intentionality:

Foundations of Social Cognition. Cambridge,MA: MIT Press.
Mallon, R. (2008). Knobe vs. Machery: Testing the tradeoff hypothesis. Mind and Language, 23, 247-255
McGrath, S. (2005). Causation by omission: A dilemma. Philosophical Studies, 123, 125-148.
Mele, A. \& Sverdlik, S. (1996). Intention, intentional action, and moral responsibility. Philosophical Studies, 82, 265287.

Nadelhoffer, T. (2004). On praise, side effects, and folk ascriptions of intentionality. Journal of Theoretical and Philosophical Psychology, 24, 196-213.
Nadelhoffer, T. (2006a). Desire, foresight, intentions, and intentional actions: Probing folk intuitions. Journal of Cognition and Culture, 6, 133-157.
Nadelhoffer, T. (2011). Experimental philosophy of action. In J. Aguilar, A. Buckareff, \& K Frankish (Eds.), New Waves in the Philosophy of Action. New York: Palgrave Macmillan.
Nado, J. (2008) Effects of moral cognition on judgments of intentionality. British Journal for the Philosophy of Science, 59, 709, 731.
Rotenberg, K. (1980). Children's use of intentionality in judgments of character and disposition. Child Development, 51, 282-284.
Schaffer, J. (2000). Causation by disconnection. Philosophy of Science, 67, 285-300.
Sripada, C. (2010). The deep self model and asymmetries in folk judgments about intentional action. Philosophical Studies, 151, 159-176.
Sripada, C. \& Konrath, S. (2011). Telling more than we can know about intentional action. Mind \& Language, 26, 353-380.
Sripada, C. (2012). Mental state attributions and the sideeffect effect. Journal of Experimental Social Psychology, 48, 232-238.
Utich, K. and Lombrozo, T. (2010). Norms inform mental state ascriptions: A rational explanation for the side-effect effect. Cognition, 116, 87-100.
Wolff, P., Barbey, A. K., \& Hausknecht, M. (2010). For want of a nail: How absences cause events. Journal of Experimental Psychology: General, 139, 191-221.
Wright, J. \& Bengson, J. (2009) Asymmetries in judgments of responsibility and intentional action. Mind and Language, 24, 24-50.
Young, L., Cushman, F., Adolphs, R., Daniel, T., Hauser, M. (2006). Does emotion mediate the relationship between ac action's moral status and its intentional status?Neuropsychological evidence. Journal of Cognition and Culture, 6, 291-304.
Zalla, T. \& Leboyer, M. (2011). Judgment of intentionality and moral evaluation in individuals with high functioning autism. Review of Philosophical Psychology, 4, 681-698.

\title{
Scientific Diagrams as Traces of Group-Dependent Cognition: A Brief Cognitive-Historical Analysis
}

\author{
Ben Sheredos (sheredos@uscsd.edu) \\ Department of Philosophy and Center for Chronobiology, UC San Diego, 9500 Gilman Dr., La Jolla, CA 92103 USA
}

\begin{abstract}
Recent research has begun to explore the role of diagrams as cognitive tools. Here I develop new conceptual and methodological tools for exploring the sociality of cognition involving diagrams. First, I distinguish two varieties of groupdependent cognition. Second, extending Nersessian's method of cognitive-historical analysis, I show how a suitablyinformed "literature review" of diagrams published in scientific articles offers a window into the group-dependent cognition of scientists. I end by sketching future avenues of inquiry, and how this approach may inform science education.
\end{abstract}

Keywords: chronobiology; cognitive-historical analysis; group cognition; member cognition; scientific diagrams.

\section*{Introduction}

\section*{Diagrams as Cognitive and Social Tools}

Cognitive scientists have recently adopted a variety of approaches to studying graphical practices ("GPs"). Tversky applies her work on embodiment, spatial cognition and navigation to study spatial graphics and spatial design more generally (2011a; 2011b; Tversky, Heiser, Lee, \& Daniel 2009). Hegarty focuses on the cognitive abilities underlying the "spatial intelligence" which facilitates learning from diagrams by students in the sciences (2004; 2010; 2011). Cheng explores how suitably constrained, innovative GPs support learning the conceptual structure of highly mathematized domains (Cheng 1997; 2002; 2009; 2011).

The focus of such research has tended to be on the consumption of completed diagrams as a cognitive activity of individuals. \({ }^{1}\) A few studies have also addressed the production of diagrams by individuals. However, constructing and reasoning with GPs are also social practices. Some researchers have recently developed ethnographic methods to study group cognition involving completed diagrams (Alač, 2008; 2011; Kirsh, 2009).

Here I take a different approach. First, I highlight social aspects of cognition in diagram production. GPs often integrate ideas from a variety of earlier sources, and diagrams indicate the designer's understanding of her field: GPs inform us about how individuals perceive the social and professional groups of which they are members. Second, I stress the social effects of diagram consumption: creating and disseminating diagrams is a manipulation of the social environment which helps to define boundaries

\footnotetext{
\({ }^{1}\) This is especially true of the experimental literature in which isolated subjects complete tasks involving diagrams.
}
between social-professional groups. To do this I develop a new strategy of inquiry: analyzing published scientific diagrams which document the history of research. My case study concerns research into the mechanisms of circadian rhythmicity in cyanobacteria (blue-green algae). One goal of the paper is to show how such a "literature review" can serve to investigate scientists' group-dependent cognition.

\section*{Extending the Cognitive-Historical Method}

The present paper extends the method of cognitive-historical analysis (Nersessian, 1992; 1995; 2002; 2008). The method is historical in that it takes as data the existing record of investigative practices in the science(s) of interest. In early work, Nersessian focused on the work of notable individuals (e.g., Maxwell), highlighting specific developments in their thinking. Here, I examine a years-long record of published figures depicting multiple authors' conceptions, at various stages of inquiry, of the known and hypothesized mechanisms of circadian rhythmicity in cyanobacteria.

The cognitive aspect of the methodology is rooted in a continuum hypothesis - that "the cognitive practices scientists have invented and developed over the course of the history of science are [...] sophisticated outgrowths of the kinds of cognitive strategies humans employ in coping with their environments and in problem solving of a more ordinary kind" (Nersessian, 2008). Scientists, like other humans, form cooperative groups to tackle large-scale tasks, and freely draw inspiration from peers when it is available. \({ }^{2}\) I shall show that with careful attention to the field-wide context in which diagrams are developed, we can clearly identify aspects of GPs which indicate group-dependent cognition among scientists. In this initial demonstration, I focus on diagrams from review-style articles, penned by (sometimes several) well-known and respected authors in the field. The express purpose of such publications is to offer a window into the social, conceptual, and evidential context constituting the current state of play in the field.

Nersessian has always stressed that a full understanding of cognitive activities must embed them within their social context. Recently, she and her colleagues have directly studied the interplay between social and cognitive factors in scientific practice (Osbeck, Nersessian, Malone, \& Newstetter, 2011). Drawing upon their insights, I hold that the lines between "individual" and "group" cognition are

\footnotetext{
\({ }^{2}\) Scientific research is not fully communal and cooperative; great incentives promote individual achievement as well.
}
not always clear-cut, since properly attending to social context sometimes requires reconceiving an individual's cognitive activities as group-dependent.

\section*{Delineating Group-Dependent Cognition}

Classical cognitive science maintains that individuals exhibit forms of cognition which do not clearly depend upon their membership in a group. Call this individual cognition, or i-cognition. As Hutchins (1995) argued, cognition might also be distributed across a group of cognizers so that the group instantiates a higher-order cognitive architecture. Call this group-level cognition, or \(g\) cognition. I emphasize that individuals exhibit a third, unique variety of cognition when they self-identify as members of a group (whether or not that group exhibits \(g\) cognition). Call this member cognition, or m-cognition. Like \(i\)-cognition (and unlike \(g\)-cognition) \(m\)-cognition is attributable to (first-order) individuals, rather than to groups. However, like \(g\)-cognition (and unlike \(i\)-cognition) \(m\) cognition depends upon an individual's group-membership. Osbeck et al. provide paradigmatic examples of \(m\)-cognition in their analysis of how scientists position themselves and negotiate their identity: "Identity negotiation can be considered a form of sense-making" (what they elsewhere call seeking coherence) "directed to the meanings one applies to oneself within social groups that include but are not limited to the particular research laboratory, one's field of practice... and science as a tradition of inquiry" (2011).

To illustrate how I conceive of \(m\)-cognition, consider the following objections to my proposed method. First, by looking to published diagrams, I am guaranteed to miss many ( \(i-\) ) cognitive activities involved in their production. The creator(s) of a diagram often discard a variety of "failed" versions, deploying expertise in choosing what to represent and how best to do so. Only access to the unpublished, discarded diagrams could really shed light on the process of problem-solving that led to the finished product. Second, publication requirements imposed by journals may add a layer of cognitive opacity, as the designer loses the ability to do just as they like. Published diagrams might be cognitively "whitewashed," so to speak.

While this line of thought is correct as far as it goes, it neglects one important reason for pursuing this inquiry: when authors prepare materials for publication, they knowingly work within the constraints imposed by "outside" powers. Publication is a de facto requirement for active membership in a professional science, and part of one's professionalism consists in navigating the pitfalls of publication. If part of scientists' practice involves "whitewashing" their individual cognitive products, making them ready for public consumption, the whitewashing itself depends upon interesting forms of \(m\)-cognition which reflect an individual's self-identification as a member of a group e.g., awareness of professional-bureaucratic norms, and selfmonitoring with respect to those norms.

More relevant, for my purposes, are the ways scientists self-monitor with respect to the empirical and evidential norms of their field. When a scientist prepares a publication for consumption by her peers, her professional reputation depends upon cognizance of: the empirical support accorded to various hypotheses; which sources of evidence have been deemed reliable; which findings have been replicated or reinterpreted, etc. These are just some of the \(m\)-cognitive activities which an individual engages in to negotiate her specific expertise, self-identifying as an able practitioner of some method(s) or authority on some topic(s).
These norms are especially relevant to the production of my source materials: authoritative review articles presenting the current state of a field. The production of diagrams is an integral part of crafting such articles. Thus, I suggest that cognitive-historical analysis of such published diagrams can plausibly begin with the hypothesis that these GPs, as part of the professional practice of scientists, are guided by \(m\) cognition regarding empirical and evidential norms in the relevant discipline(s). It follows that such diagrams are amenable to analysis as visual traces of \(m\)-cognition. My task in what follows is to demonstrate that this is the case.

\section*{Three Snapshots of Cyanobacterial Chronobiology}

I turn now to canvass three stages of research regarding circadian rhythms in the cyanobacterium Synechococcus elongatus. Here I must be selective in every aspect of my inquiry. \({ }^{3}\) In this section I introduce the details of my casestudy. Cognitive analysis occurs in the section thereafter.

\section*{Stage One: First Steps}

A biological system's circadian rhythmicity ("CR") is its endogenously controlled production, once every \(\sim 24\) hours, of some phenomenon (e.g., waking, onset of metabolic processes, peak transcription of a gene). For decades, while research into the CR of eukaryotes flourished, it was thought that no similar phenomena would be discovered in prokaryotic cells. Prokaryotes lack membrane-bound organelles, exhibit relatively simple metabolic activities, and frequently have lifespans of less than 24 hours. Prevailing wisdom taught that such an organism would have no use for anticipating local day-night cycles.
CR was eventually discovered in S. elongatus (Ishiura, Kutsuna, Aoki, Iwasaki, Andersson, Tanabe, Golden, Johnson, \& Kondo, 1998). Since then this system has become a mainstay of circadian research, owing in part to its high genetic manipulability. In an early review article, Kondo \& Ishiura (1999) made an explicit attempt to shoehorn cyanobacterial rhythmicity into the accepted mechanism for eukaryotic systems. Evidence from a variety of eukaryotic systems had suggested that the CR of single cells was controlled by a Transcription-Translation

\footnotetext{
\({ }^{3}\) For details, cf. Johnson \& Xu (2009) and Huang \& Lin (2009).
}

Feedback Loop (TTFL). A generic eukaryotic TTFL is shown at left in Figure 1. A "clock gene" (dark blue bar) is transcribed, leading to the translation of a corresponding "clock protein" (dark blue oval) outside the nucleus. After undergoing state-changes in the cytoplasm, the clock protein returns to the nucleus, where it interrupts the effect of an "activator" (red oval) at a promoter region (light blue bar). The clock protein(s) thus inhibit further transcription of the clock gene(s). By hypothesis, a TTFL constituted a cell's core circadian "clock" or "pacemaker" and the CR in the expression of other ("clock-controlled") genes was thought to be dependent upon the activity of clock proteins.

The critical functional arrangement of the TTFL is the interplay of positive and negative elements: activation at the promoter increases transcription of clock genes, but clock proteins feedback to inhibit transcription of their own genes. Such systems can instantiate a limit cycle oscillator. With the right time constants, the system could oscillate with a 24-hour period, giving rise to the organism's observed CR.

At right in Figure 1, the authors attempt to fit cyanobacterial CR into the same scheme. Early research (Ishiura et al., 1998) had shown (a) that deletion of any gene in the kai gene cluster (containing genes kaiA, kaiB, and kaiC) abolished CR in S. elongatus, and (b) that a variety of single amino acid mutations to any of the kai genes (resulting in the corresponding production of subtly altered Kai proteins) either disturbed or abolished CR. It was thus concluded that the core clock in S. elongatus involved the kai gene cluster and the Kai proteins working in concert.

The same study also showed that while kaiC overexpression resulted in rapidly decreased activity at the promoter (" \(\mathrm{P}_{\text {kaiBC }}\) " in Figure 1) which controls the transcription of kaiB and kaiC, kaiA overexpression resulted in increased activity of the same promoter. Thus, the Kai proteins appeared capable of participating in a TTFL, with KaiC playing the role of a traditional "negative element" (note the re-use of the dark blue oval) which inhibits its own
gene's transcription, and KaiA playing the role of a traditional "positive element" (note the re-use of the red oval) which promotes KaiC's transcription.

These initial results were consistent with a cyanobacterial TTFL, but left much underdetermined. For example, it was unknown how Kai proteins might influence the transcription of kai genes, since the Kai proteins lacked DNA binding motifs, and were thus incapable of directly influencing promoters (Ishiura et al., 1998). Given the success of the TTFL model in other systems, the authors of Figure 1 posited intervening entities ("x," "y," "z") to mediate between the Kai proteins and transcriptional regulation.

\section*{Stage Two: Troubles with TTFLs}

A few years later, Johnson (2004) published a "minireview" in which he proposed the alternative "Oscilloid" model shown in Figure 2. The interactions between Kai proteins had by now been further determined. As shown top-right KaiC alternates between a highly phosphorylated state (with "P" attached) and an unphosphorylated state (no "P"). KaiA facilitates KaiC's phosphorylation, and inhibits its dephosphorylation. KaiB inhibits those activities of KaiA, biasing KaiC towards dephosphorylation. The result is a \(24-\) hour rhythm in the phosphorylation state of KaiC, which is a determining factor in KaiC's downstream effects.

Meanwhile, the problem of how the Kai proteins might regulate transcription had become more pressing. It had been shown that KaiC expression not only (somehow) repressed transcription of its own gene (and of kaiB), but also globally repressed transcription of virtually every gene in S. elongatus' genome (Nakahira, Katayama, Miyashita, Kutsuna, Iwasaki, Oyama, \& Kondo, 2004). Investigators had also shown that rhythmicity in kaiC expression could be attained in strains in which kaiC transcription was controlled by a promoter taken from another organism's genome (Xu \& Johnson, 2003; Nakahira et al., 2004). These were departures from eukaryotic TTFLs, in which positive


Figure 1: Kondo \& Ishiura's (1999) Figures 3 (left) \& 4 (right). At left is the TTFL model of rhythmicity in eukaryotic cells. At right, available data in S. elongatus are fitted into a similar scheme. See text for discussion.
and negative feedback loops compete for dominance in the activation and inhibition of specific, native promoters.

The same researchers recommended an elegant solution. KaiC had been shown to be part of a large family of DNA recombinases (Leipe, Aravind, Grishin, Koonin, 2000). KaiC was thus hypothesized to be capable of altering the shape and structure of cyanobacterial chromosomes in a rhythmic fashion, thereby globally affecting gene transcription (including, as just one example, the kaiABC cluster). At middle-left in Figure 2, Johnson added this to the hypothetical model of CR in S. elongatus.

\section*{Stage Three: Surviving 2005}

Not long after Johnson's minireview, a pair of momentous reports showed conclusively that cyanobacterial CR was not dependent upon a TTFL. An initial report showed that CR in KaiC's phosphorylation state persists even when both transcription and translation are globally inhibited (Tomita, Nakajima, Kondo, \& Iwasaki, 2005). Shortly thereafter, it was reported that KaiC's phosphorylation rhythm could be reconstituted in vitro, using a mixture containing only the three Kai proteins and ATP (Nakajima, Imai, Ito, Nishiwaki, Murayama, Iwasaki, Oyama, \& Kondo, 2005). The core clock in cyanobacteria, it seemed, was instantiated entirely in post-translational entities and processes, and required no transcriptional regulation whatsoever. Transcriptional and translational regulation were reconceived as effects of clock functioning, not operations constitutive for clock function. \({ }^{4}\)

After 2005, researchers pursued the molecular details of KaiC phosphorylation rhythms. Here I cannot discuss this


Figure 2: Johnson's (2004) Oscilloid model.

\footnotetext{
\({ }^{4}\) Transcriptional regulation was later seen as stabilizing or supporting the Kai-based clock (Johnson, Mori, \& Xu 2008).
}
research in detail. Figure 3, a representative example of the period, displays no details concerning chromosomes, transcriptional regulation, or transcriptional rhythms.

\section*{Analysis}

By attending to the social and evidential contexts surrounding the production of Figures 1-3, we gain insight into the m-cognition of these diagrams' designers. In this way we can provide cognitive answers to questions about scientists' GPs. We simultaneously gain insight into how GPs helped shape the social environment of chronobiology.

Consider Figure 1. A pertinent question to ask regarding this figure is: Why did the authors construct this diagram as they did, drawing an analogy between eukaryotic and prokaryotic CR?

At the time of publication, no data substantively confirmed the presence of a TTFL in S. elongatus. Available data were merely consistent with such a model. What drove Kondo \& Ishiura to produce this diagram was a broader awareness of hypotheses accepted elsewhere in chronobiology. Eukaryotic TTFLs were then considered the sole concrete examples of circadian limit cycle oscillators in living systems. The group of "chronobiologists" was de facto defined by an interest in such mechanisms. By hypothesizing that the newly discovered CR in cyanobacteria fit the same model, the authors explicitly positioned themselves in the broader theoretical community of chronobiologists.

The text of the article supports this interpretation. Kondo\& Ishiura aim to show how cyanobacteria could fit the "basic circadian model" of a limit-cycle oscillator, and explicitly recommend strategies for further-extending this model to CR in plants (1999, p. 171). Kondo \& Ishiura also stress the importance of assimilating cyanobacteria to the TTFL model, forming a theoretically unified chronobiology: "Cyanobacteria could be a model system for molecular approaches to the circadian clock, because it is the simplest organism that has a clock" (1999, p.172). The subsumption of cyanobacterial CR to the TTFL model would lend credence and generality to the working assumptions of chronobiologists at large.


Figure 3: Mackey \& Golden's (2007) visual summary of the stages of the core Kai-based oscillator in S. elongatus.

Figure 1 is the visual depiction of this shared theoretical framework: the authors literally drew the analogy between the models of CR in eukaryotes and prokaryotes which unified the theoretical framework of chronobiology. Kondo \& Ishiura were positioning themselves as (and encouraging other researchers to recognize themselves as) members of a single, theoretically-unified group of "chronobiologists." Their GP is (partly) explained by appeal to this \(m\)-cognition; the graphic itself is a trace of that \(m\)-cognition.

Consider next Figure 2. A pertinent question to ask regarding this diagram is: Why did the author depart from earlier GPs in the field, especially by including a novel depiction of the entire chromosome of cyanobacteria?

New data showed that circadian transcriptional regulation in S. elongatus was not specific to individual promoters, in contrast to eukaryotic TTFLs. Since previous data had been consistent with a cyanobacterial TTFL and had shown that Kai proteins do (somehow) participate in regulating transcription, it was a "surprise" to find these discrepancies with the eukaryotic model (Johnson 2004, p.217.2). It is these data, plus the persisting field-wide theoretical assumption that transcriptional regulation is somehow constitutive for clock function, which "suggests a broadly global mechanism for the cyanobacterial clock system" (Johnson, 2004, p.217.3). It is within these constraints that he appeals to the broader literature regarding chromosome topology in cyanobacteria, and articulates the Oscilloid model to provide a novel hypothesis regarding transcription-translation feedback in cyanobacterial CR.

Thus in Figure 2, Johnson breaks the struct visual analogy with eukaryotic, as was demanded by evidence showing that "the clock system in cyanobacteria is different from that in eukaryotes" (2004, p.217.4). Despite this, the view of transcriptional regulation as a process constitutive for clock function remained part of the shared theoretical framework of a still-unified chronobiology. For this reason, Johnson stresses that the cyanobacterial data might lead us to consider the hypothesis that eukaryotic clocks themselves involve chromosomal topology as a mechanism of transcriptional regulation. He writes that "If this proves to be the case, the investigations of the cyanobacterial clock may lead to fundamental insights that are broadly applicable to all organisms" \((2004,217.4)\). In either case a unified chronobiology would need to refine its theoretical framework to incorporate cyanobacterial data.

Figure 2 is the visual depiction of the new model for cyanobacterial transcription-translation feedback. The graphical disparity from earlier depictions of the cyanobacterial clock reflects the conceptual departure from the TTFL model. Johnson positioned himself as (and encouraged other cyanobacterial researchers to recognize themselves as) a member of a distinct sub-group of chronobiologists which was helping to refine the general theoretical framework of chronobiology. Johnson's GP is (partly) explained by appeal to this m-cognition; the graphic is itself a trace of this \(m\)-cognition.

Finally, consider Figure 3. A pertinent question to ask regarding this diagram is: Why have the authors departed from earlier GPs, especially by excluding all reference to transcriptional regulation?

The core clock in S. elongatus had been identified as a post-translational oscillation in the phosphorylation state of KaiC (involving interactions with other proteins). The data showed that "transcriptional regulation is apparently a dispensable layer of reinforcement on a post-translational clock in the cyanobacterium" (Mackey \& Golden 2007, p.382). Transcriptional regulation (local or global) was no longer considered part of the core clock in S. elongatus.

Figure 3 above is the visual depiction of the new model of the cyanobacterial clock. By excluding any depiction of genes, chromosomes, transcriptional feedback, and the like, Mackey and Golden underscore the distinction between cyanobacterial and eukaryotic clocks. Cyanobacterial chronobiology was no longer theoretically yoked to molecular hypotheses drawn from eukaryotic systems: the hypothesis that some form of transcriptional regulation would be constitutive for the function of every circadian limit-cycle oscillator had been excised from the general theoretical framework of chronobiology. Cyanobacterial chronobiologists had distinguished themselves as a unique subgroup of chronobiologists. Mackey and Golden's GP can be (partly) explained by appeal to this \(m\)-cognition, and the diagram itself is a trace of this m-cognition.

\section*{Concluding Remarks}

In this brief case study, I have demonstrated that the method of cognitive-historical analysis may be fruitfully extended to reveal scientific diagrams as visual traces of groupdependent cognition (m-cognition). In doing so, I have sketched how scientists' GPs help demarcate the boundaries between groups of researchers. It is hoped that with the benefit of future elaboration, this approach can take its place as a compliment to other empirical methods of examining the cognitive activities involved in GPs.

With this initial demonstration completed, I suggest that inquiry into diagrams may be especially well-suited for investigating the \(m\)-cognition of scientists. As in the cases above, published diagrams frequently offer "at a glance" a window into authors' construal of the state of the art in their field. While I have not emphasized it, the examples also hint at the extent to which authors recycle old formats (often citing their original designers), positioning themselves as members of a persisting group and building extended "lineages" of GPs. Further research might fruitfully explore the "cognitive" lineages of which these are visual traces.

Finally, I suggest that such analyses might fruitfully inform science education. The foregoing demonstrates how published figures provide a visual record of the empirical and theoretical developments which fuel scientific fields' growth and subdivision, and how they can serve as a window into researchers' conception of their own field. As I hope to have shown, when such graphics are presented with appropriate context, and when they are queried in a suitable
manner, they can serve as intuitive scaffolds to help novices gain a rich understanding of the course of expert thinking in a field of study.

\section*{Acknowledgments}

This material is based upon work supported by the National Science Foundation under Grant no. 1127640. I learned much from other members of UCSD's Center for Chronobiology while supported by a UCSD Chancellor's Interdisciplinary Collaboratory Grant. Feedback from members of the WORGODS research group (Daniel Burnston, Adele Abrahamsen, and William Bechtel) and from two reviewers strengthened the paper considerably.

\section*{References}

Alač, M. (2008). Working with brains: digital images and gestural interaction in fMRI laboratory. Social Studies of Science, 38, 483-508.
Alač, M. (2011). Handling Digital Brains. Cambridge, MA: MIT Press.
Cheng, P.C.-H. (1997). Components of a cognitive theory of problem solving and discovery with law encoding diagrams. In M. Anderson (Ed.) Reasoning with Diagrammatic Representations II (pp.85-93). (Tech. Rep. FS-97-02). Menlo Park, CA: AAAI.
Cheng, P.C.-H. (2002). Electrifying diagrams for learning: principles for complex representational systems. Cognitive Science, 26, 685-736.
Cheng, P.C.-H. (2011). Probably good diagrams for learning: representational epistemic recodification of probability theory. Topics in Cognitive Science, 3, 475498.

Hegarty, M. (2004). Mechanical reasoning by mental simulation. TRENDS in Cognitive Sciences, 8, 280-285.
Hegarty, M. (2010). Components of Spatial Intelligence. Psychology of Learning and Motivation, 52, 265-297.
Hegarty, M. (2011). The cognitive science of visual=spatial displays: implications for design. Topics in Cognitive Science, 3, 446-474.
Huang, T-.C., \& Lin, R-.F. (2009). Circadian rhythm of Cynothece RF-1 (Synechococcus RF-1). In J.L. Ditty, S.R. Mackey, \& C.H. Johnson (Eds.) Bacterial Circadian Programs. Heidelberg, DE: Springe-Verlag.
Hutchins, E. (1995). Cognition in the Wild. Cambridge, MA: MIT Press.
Ishiura, M., Kutsuna, S., Aoki, S., Iwasaki, H., Andersson, C.R., Tanabe, A., Golden, S.S., Johnson, C.H., \& Kondo, T. (1998). Expression of a gene cluster \(k a i A B C\) as a circadian feedback process in cyanobacteria. Science, 281, 1519-1523.
Johnson, C.H., (2004). Global orchestration of gene expression by the biological clock of cyanobacteria. Genome Biology, 5: 217.
Johnson, C.H., Mori, T., \& Xu, Y. (2008). A cyanobacterial circadian clockwork. Current Biology, 18, R816-R825.
Johnson, C.H. \& Xu., Y. (2009). The decade of discovery: how Synechococcus elongatus became a model circadian
system 1990-2000. In J.L. Ditty, S.R. Mackey, \& C.H. Johnson (Eds.) Bacterial Circadian Programs. Heidelberg, DE: Springe-Verlag.
Kirsh, D. (2009). Thinking with external representations. AI \& Society, 25, 441-454.
Kondo, T., \& Ishiura, M. (1999). The circadian clocks of plants and cyanobacteria. Trends in Plant Science, 4, 171176.

Leipe, D.D., Aravind, L., Grishin, B.V., Koonin, E.V. (2000). The bacterial replicative helicase DnaB evolved from a RecA duplication. Genome Research, 10, 5-16.
Mackey, S.R. \& Golden, S.S. (2007). Winding up the cyanobacterial circadian clock. Trends in Microbiology, 15, 381-388.
Nakahira, Y., Katayama, M., Miyashita, H., Kutsuna, S., Iwasaki, H., Oyama, T., \& Kondo, T. (2004). Global gene repression by KaiC as a master process of prokaryotic circadian system. PNAS, 101,881-885.
Nakajima, M., Imai, K., Ito, H., Nishiwaki, T., Murayama, Y., Iwasaki, H., Oyama, T., \& Kondo, T. (2005). Reconstitution of circadian oscillation of cyanobacterial KaiC phosphorylation in vitro. Science, 308, 414-5.
Nersessian, N. J. (1992). How do scientists think? Capturing the dynamics of conceptual change in science. In R.N. Giere (Ed.) Cognitive Models of Science (pp. 3-45). Minneapolis, MN: University of Minnesota Press.
Nersessian, N.J. (1995). Opening the black box: Cognitive science and the history of science. Osiris, 10, 194-211.
Nersessian, N.J. (2002). "The cognitive basis of modelbased reasoning in science. In P. Carruthers, S. Stich, M. Siegel (Eds.) The Cognitive Basis of Science (pp. 133153). Cambridge, MA: Cambridge University Press.

Nersessian, N.J. (2008). Creating Scientific Concepts. Cambridge, MA: MIT Press.
Osbeck, L., Nersessian, N., Malone, K., \& Newstetter, W. (2011). Science as Psychology. Cambridge, MA: Cambridge University Press.
Sheredos, B., Burnston, D., Abrahamsen, A., \& Bechtel, W. (forthcoming). Why do biologists use so many diagrams? Philosophy of Science.
Tomita, J., Nakajima, M., Kondo, T., \& Iwasaki, H. (2004). No transcription-translation feedback in circadian rhythm of KaiC phosphorylation. Science, 307, 251-254.
Tversky, B., Heiser, J., Lee, P., and Daniel, M.-P. (2009). Explanations in gesture, diagram, and word. In K.R. Coventry, T. Tenbrink, \& J.A. Bateman (Eds.) Spatial Language and Dialogue (pp. 119-131). Oxford: OUP.
Tversky, B. (2011a). Visualizations of thought. Topics in Cognitive Science, 3, 499-535.
Tversky, B. (2011b). Spatial thought, social thought. In T.W. Schuber \& A. Maass (Eds.) Spatial Dimensions of Social Thought (pp. 17-39). Boston, MA: Walter de Gruyter GmbH \& Co.
Xu, Y., Mori, T., \& Johnson, C.H. (2003). Cyanobacterial circadian clockwork: roles of KaiA, KaiB, and the kaiBC promoter in regulating KaiC. EMBO Journal, 22, 2117.

\title{
Infants' Early Understanding of Coincidences
}

\author{
Zi L. Sim (zi@berkeley.edu) Fei Xu (fei_xu@berkeley.edu) \\ Department of Psychology, University of California, Berkeley \\ Berkeley, CA 94720 USA
}

\begin{abstract}
Coincidences are surprising events that can provide learners with the opportunity to revise their theories about how the world works. In the current research, we investigate whether infants are truly sensitive to coincidences, detecting these events even when they cannot be predicted mere probabilities. In addition, we explore whether this sensitivity is translated into action, encouraging infants to engage in activities that enable them to revise their theories. Results from 2 experiments demonstrate that infants display a sensitivity to coincidence similar to adult intuitions, and they selectively explore objects that produce anomalous data that better supports an alternative theory than their prior theory.
\end{abstract}

Keywords: coincidence; probabilistic reasoning; theory revision.

\section*{Introduction}

As scientists, we are sometimes met with surprising results in our research. At this point, it can be a struggle trying to reconcile the evidence with our theories - do we chalk these results up as experimenter error, or do we discard our theories in favor of an alternative one? Sometimes, such results were a mere coincidence; they go away upon a review of our procedures or an increase in sample size.

Other times, however, unexpected results have led to important scientific discoveries. For example, the discovery that cholera was caused by infected water, rather than the prevailing notion that the disease was transmitted by bad air, was due to a suspicious coincidence observed by John Snow, a physician, in 1854. After a particularly violent outbreak of cholera in the city of London, Snow noticed that the location of the victims were all tightly clustered around a water pump on Broad Street (Snow, 1855). Using this suspicious finding, Snow convinced the local council to remove the handle of the water pump, and this action has since been commonly credited with stopping the further spread of cholera. Such discoveries abound in the history of science, emphasizing the detection of suspicious coincidences as key to causal discovery (Owens, 1992; Nickerson, 2004) and rational inference (Horwich, 1982).

This detection of coincidences is not limited to scientific research, as adults use it to make inferences in daily life as well. They also act on these observations appropriately, taking the nature of these coincidences into consideration. Imagine a situation in which one leaves our apartment in a hurry on the way to work. While waiting for the elevator, we meet three of our neighbors, and each one of them is carrying an umbrella. Do we dismiss this observation as a
mere coincidence? Probably not - we are likely to discard our belief that today is just like any fair weather day in California, rushing back to grab an umbrella too.

According to Griffiths \& Tenenbaum (2007), our reaction at the elevator can be properly understood through a Bayesian framework, in which coincidences are formally defined as events that provide better support for an alternative theory, as compared to a currently favored causal theory. Whether a surprising observation should be taken as compelling evidence or dismissed as a mere coincidence is thus dependent on two factors: the prior probability of the alternative theory, and the strength of coincidence as given by the likelihood ratio. The likelihood ratio quantifies the support that the evidence provides for the alternative theory over the original theory. Griffiths \& Tenenbaum (2007) also show empirically that adults evaluate coincidences in ways that are consistent with this framework - they take into account prior probabilities and likelihood ratios when thinking about unexpected evidence. For example, adults judged the results of a test of psychokinesis (low prior probability) as mere coincidence and that of a test of genetic engineering (high prior probability) as evidence, even when the data provided the same support for the two alternative theories. However, when the strength of the coincidence is manipulated to be sufficiently high in the case of psychokinesis, adults find it increasingly hard to dismiss the observations as just chance results. Therefore, adults act in ways consistent with an ideal Bayesian learner, evaluating and acting on observed coincidences in rational ways.

Our reaction in the elevator situation can thus be easily understood through the Bayesian lens. Given that there are many more fair weather days than rainy days in California, we should favor the null hypothesis due to its higher prior probability: today is a fair weather day, just like most days. However, the surprising observation that our neighbors are all carrying their umbrellas gives a high likelihood ratio, strongly suggesting an unexpected causal structure: today is a rainy day. We have thus detected a coincidence: the observed event provides better support for an alternative theory of rainy weather, as compared to our prior theory of fair weather. In this example, the alternative theory also has a sufficiently high prior probability, pushing us to discard our original theory, and to intervene by retrieving an umbrella from our apartment.

Detecting such coincidences is important for learners, as these events are often great opportunities for us to revise our current theory of how the world works. This opportunity is especially essential for children, whose accounts of the inner
workings of the world are under major construction and revision (Carey, 1985, 2009; Gopnik \& Meltzoff, 1997). Given that many causal relationships in the world are probably novel to young children, coincidences are rich sources of information for how their theories should be revised, and thus one should predict that children should pay great attention to coincidences (Griffiths \& Tenenbaum, 2007). In this paper, we investigate this prediction by asking whether young infants are sensitive to such coincidences, and whether this sensitivity translates into action that can help them update their theories about how the world works.

Recent research exploring the development of probabilistic reasoning has provided ample evidence that 6to 12 -month-old infants are sensitive to differences in probabilities (Denison, Reed \& Xu, 2012; Teglas, Girotto, Gonzalez \& Bonatti, 2007; Xu \& Garcia, 2008). For example, Teglas et al. (2007) showed that in a lottery machine-like setup that consisted of 1 yellow and 3 blue objects bouncing around, 12-month-old infants were more "surprised" to see a yellow object (low probability) exiting the machine, than when a blue object (high probability) did.

However, detecting coincidences is not quite the same as assessing the relative probabilities of different events. In some cases, events can have equal probabilities, but we do not consider them equally surprising. Take the instance of five rolls of a 6 -sided dice. The probability of seeing the sequence " \(2,1,4,3,1\) " is \(\left(\frac{1}{6}\right)^{5}=0.00013\). Although low in probability, this event is unsurprising to most adults. In contrast, the probability of seeing the sequence " \(1,1,1,1\), 1 " is again \(\left(\frac{1}{6}\right)^{5}=0.00013\), but this time we are astonished, becoming suspicious of the dice and the roller of the dice. These intuitions cannot be simply explained by the proposal that learners are actually evaluating and comparing the probabilities of "kinds" of events (e.g. the probability that the sequence consists of different numbers vs. the probability that the sequence consists of the same number), instead of single events (e.g. the probability that the specific sequence is " \(2,1,4,3,1\) "). This proposal is problematic as it is unclear what exactly counts as a "kind" of event, and what a learner should do when there are many possible "kinds" to consider (e.g. running sequences such as " \(1,2,3\), 4,5 " or alternating sequences such as " \(1,2,1,2,1\) " are suspicious too, but do not fit into the two earlier-mentioned "kinds"; see Griffiths \& Tenenbaum, 2007 for a more comprehensive review). Instead, we consider the sequence " \(1,1,1,1,1\) " to be a suspicious coincidence because it provides better support for the alternative theory that the dice is weighted towards " 1, " rather than our original theory that the dice is fair.

As such, there exists a gap in our knowledge of whether infants are truly sensitive to coincidences - we know that they are surprised by the occurrence of low-probability events, but do they detect coincidences even when the mere probabilities of different events are exactly equal? To investigate this question, we designed an experiment analogous to the dice roll example detailed earlier. 8-month-
old infants were familiarized to a box containing 6 different colored balls. An experimenter then tossed out a ball from the box, seemingly with no control over the outcome of the event. The ball was then returned to the box, and this event was repeated 3 more times. Using a violation-of-expectation paradigm, we measured the amount of time infants looked at a trial where the same colored ball fell out each time (e.g. yellow, yellow, yellow, yellow), as well as a trial where a different colored ball fell out each time (e.g. blue, green, red, yellow). Note that each specific sequence shown had the same exact event probabilities: \(\left(\frac{1}{6}\right)^{4}\).

We also designed a second experiment with an exploration measure, examining whether this sensitivity to coincidences translates to action, such that the detection of a suspicious coincidence could potentially have consequences on children's learning. In this experiment, we showed 13-month-old infants two different boxes each containing 6 different colored balls. One of the boxes always generated the same sample each time, and the other always generated a different sample each time. The two boxes were then offered to the infants to play with freely, and we measured the amount of time they played with each box.

\section*{Experiment 1}

In Experiment 1, we investigated whether infants were sensitive to coincidences that cannot be predicted by the computation of the mere probabilities of events. If infants shared adult intuitions, they should look longer at the event in which the same colored ball fell out of the box each time under random sampling, than when a different colored ball fell out each time. We also included a Baseline condition to assess infants' intrinsic preferences for these two events.

\section*{Method}

Participants Forty infants (21 males and 19 females, \(M=8\); 6 [months; days], \(R=7 ; 3\) to \(9 ; 1\) ) were tested. All were recruited from Berkeley, California, and its surrounding communities. An additional 7 infants were tested but excluded due to fussiness \((\mathrm{N}=5)\) or experimenter error ( N \(=2\) ). Infants who participated in the experiment were required to be exposed to English a minimum of 50\% of the time. Infants received a small gift for their participation.

Materials A total of 36 colored balls ( 7 cm in diameter) were used. The balls came in 6 colors: red, purple, blue, green, yellow and orange.

A small white box ( \(28 \mathrm{~cm} \times 10 \mathrm{~cm} \times 7.5 \mathrm{~cm}\) ) constructed from foam core was used in the Free Play phase of the experiment (see Procedure). The box contained 3 different colored balls.

A small, transparent Plexiglas container with an open top ( \(16.5 \mathrm{~cm} \times 7.5 \mathrm{~cm} \times 9 \mathrm{~cm}\) ) was used to display the sampled ball during the test trials.

A large box ( \(30 \mathrm{~cm} \times 26 \mathrm{~cm} \times 21 \mathrm{~cm}\) ) was used to display the population of 6 different colored balls during the familiarization and test phase. The box was rectangular,
with a Plexiglas window to show the population of balls, and two hidden back compartments. One compartment was used to hold the 4 sample balls to be tossed out later during the test trials, while the other compartment was to contain the balls that were being returned to the box after each toss. From the infants' perspectives, the box appeared as one single unit, filled only with 6 different colored balls. The Plexiglas display window was covered with a fabric curtain to ensure that the population would be hidden from sight while each sampled ball was being tossed out.

Apparatus The testing room was divided in half by curtains spanning its width and height. The curtains had a cut-out above a puppet stage that measured \(94 \mathrm{~cm} \times 55 \mathrm{~cm}\) (width x height). The experimenter sat behind the stage with her upper body and head visible to the infant. There was a black back curtain attached to the stage, such that the experimenter is hidden from view when it is dropped. An observer, present to code the infant's looking times, sat in a corner of the room and was not visible to the infant. She watched the infant on a TV monitor and coded the infant's looking behavior online using JHAB (R. Casstevens, 2007). The observer was blind to the order of the test trials.

Infants sat in a high chair about 70 cm from the center of the stage. Each parent sat next to their infant facing the opposite direction, and was instructed to avoid looking at the stage. Two video cameras were used to record each experimental session, one to record the infant's looking behavior, and another to record the experimenter's presentation of the trials.

Design and Procedure Each infant was randomly assigned to a Sampling condition or a Baseline condition. Both conditions consisted of a Calibration phase, a Free Play phase, a Familiarization phase and a Test phase.

Sampling Condition To calibrate each infant's looking window, a squeaky toy and/or keys were used in the Calibration phase to direct the infant's attention to the outside parameters of the stage.

In the Free Play phase, the infant was shown a white box containing three different colored balls. She was encouraged to play with the balls for approximately 30 seconds, and the experimenter ensured that the infant touched every ball. This phase was to allow the infants to become familiar with the balls used in the experiment.

The Familiarization phase that followed consisted of two trials. To begin each trial, the experimenter placed the large box on the stage with its front curtain down. Then, she lifted the curtain to reveal a population of 6 different colored balls, saying "See this?" She proceeded to shake the box side to side 4 times, and then set the box back to the center of the stage. While the infant was looking at the stage, the experimenter said "Look, [baby's name], look!" and dropped the back curtain, hiding herself from view of the infant. The observer began timing upon hearing the second
"look". Trials ended when the infant looked away for 2 consecutive seconds.

The large box was removed from the stage between each familiarization trial, and the back curtain was lowered to conceal the experimenter. These trials were included to familiarize the infants to the population of balls in the large box, as well as to the general procedure of the experiment. The familiarizations lasted about 2 minutes for each trial.

The Test phase consisted of two test trials, a Uniform trial and a Variable trial. On each test trial, the experimenter placed the large box and the small Plexiglas container on the left and right side of the stage (infant's view) respectively. The two objects were placed 8 cm apart. The experimenter then lifted the front curtain of the large box, saying "What's this?" She lowered her head and directed her eye gaze at the box for 1 second, in order to remind the infants of the population of balls in the large box. She then picked up the box and shook it 4 times. After the box was set back down, the experimenter lowered the front curtain to conceal the box's display window. Then, the box was lifted and tilted to its side, allowing one ball to fall out into the small Plexiglas container. Although it appeared that the ball had fallen out from the population of balls at random, the ball actually fell out of the back compartment of the box, which contained balls that had initially been set up by the experimenter. The experimenter then directed her gaze towards the "sampled" ball in Plexiglas container, saying "Look at that!" After 1 second, the ball was returned into the box. This process of revealing the population, shaking the box and tossing a ball out was repeated 3 more times, to make a total of 4 "sampled" balls. When the \(4^{\text {th }}\) ball was tossed out, the experimenter said "Look, [baby's name], look!" and dropped the back curtain of the stage. The observer began timing upon hearing the second "look," and ended the trial after the infant looked away for 2 consecutive seconds. Between trials, the stage was cleared and the back curtain was lowered as well. Each test trial lasted for approximately 2 minutes.


Figure 1: Schematic representation of the two trials.
Each infant participated in a Uniform trial and a Variable trial (See Figure 1). In the Uniform trial, the 4 "sampled" balls were all of the same color (e.g. 4 yellow balls), while
in the Variable trial, the 4 "sampled" balls were all of a different color (e.g. 1 red ball, 1 green ball, 1 blue ball, and 1 yellow ball). The last ball that was tossed out in the Variable trial was always identical in color to the balls used in the Uniform trial, to ensure that any difference in looking time was not due to a preference for balls of a certain color. Trial order and the colors of the sampled balls were appropriately counterbalanced across infants.

Baseline Condition The procedure in the Baseline condition was identical to the Sampling condition, except that instead of having the 4 balls being tossed out from the large box, the balls were individually taken out from and returned to the experimenter's pocket. The Baseline condition provided a measure of the infants' pattern of looking times for 4 balls of the same color vs. 4 balls of different colors.

\section*{Results}

Preliminary analyses found no effects of gender, median age-split (whether the infants were younger or older than the median age of the group), or test trial order (Uniform trial first vs. Variable trial first) on looking times. Subsequent analyses were collapsed over these variables.


Figure 2: Mean looking times in the Sampling condition and the Baseline condition. Error bars represent SE.

Looking times for the test trials were analyzed using a 2 x 2 repeated-measures ANOVA with Condition (Sampling vs. Baseline) as the between-subjects factor and Trial Type (Uniform vs. Variable) as the within-subjects factor. There was a significant interaction between Condition and Trial Type, \(F(1,38)=11.58, p=.002, \eta_{p}^{2}=.23\). There were no main effects found.

To break down the interaction, we conducted follow-up ttests exploring the effect of Trial Type (Uniform vs. Variable) for each Condition separately (See Figure 2 for mean looking times). In the Sampling condition, infants looked significantly longer in the Uniform trial ( \(M=13.68 \mathrm{~s}\), \(S D=9.87\) ) than the Variable trial ( \(M=10.22 \mathrm{~s}, S D=6.35\) ), \(t(19)=2.49, p=.02, d=.42\). Thirteen out of 20 infants in
this condition looked longer in the Uniform trial, Wilcoxon signed-ranked test: \(z=1.93, p=.05\). In contrast, infants in the Baseline condition looked significantly longer in the Variable trial ( \(M=15.96 \mathrm{~s}, S D=9.02\) ) than the Uniform trial \((M=10.14 \mathrm{~s}, S D=6.01), t(19)=2.48, p=.02, d=.76\). Fifteen out of 20 infants in this condition looked longer in the Variable trial, Wilcoxon signed-ranked test: \(z=2.35, p\) \(=.02\).

\section*{Discussion}

In the Sampling condition, infants looked reliably longer when 4 balls that were tossed out at random were all of the same color, than when 4 balls were all of different colors, even though the sequences had equal event probabilities. Hence, infants found it surprising when samples that were being generated from a uniform distribution over the long run were identical, i.e. when 4 randomly generated balls (with replacement) all shared the same color, even though they came from a population of 6 different colored balls. This pattern of looking time was reversed in the Baseline condition, where infants looked longer when 4 balls of different colors were produced from the experimenter's pocket instead. Thus, our findings cannot be attributed to a preference for sequences of identical events. These results support the claim that infants are sensitive to coincidences, even when such suspicious coincidences cannot be predicted by evaluating the mere probabilities of particular events.

\section*{Experiment 2}

In Experiment 2, we used an exploration measure to examine whether the infants' sensitivity to coincidences translates into action with consequences on their learning. We predicted that infants should play longer with the box that generated the same colored ball each time under random sampling as compared to a box that generated a different colored ball each time.

\section*{Method}

Participants Fifteen infants (10 males and 5 females, \(M=\) 13; 3 [months; days], \(R=12\); 18 to 13; 29) were tested. All were recruited from Berkeley, California, and its surrounding communities. An additional 3 infants were tested but excluded for not playing with any of the boxes during the test trial. Infants who participated in the experiment were required to be exposed to English a minimum of \(50 \%\) of the time. Infants received a small gift for their participation.

Materials The materials used in Experiment 2 were identical to those used in Experiment 1, except that the large box containing the population of balls was replaced with two new boxes. Similar to the large box, these two boxes ( \(29 \mathrm{~cm} \times 23 \mathrm{~cm} \times 22 \mathrm{~cm}\) ) each had a Plexiglas window to display a population of 6 different colored balls, as well as two hidden back compartments. One of the boxes had its surface painted white, with a black fabric curtain covering the display window, while the other box had its surface
painted black, with a white fabric curtain covering its window. This design was to enhance infant's discrimination of the two boxes, without biasing the infant towards any of the boxes.

Design and Procedure Infants were tested individually in a forced-choice paradigm. Each infant sat on her parent's lap on the floor approximately 1.2 meters from a puppet stage. Parents were instructed to hold on to their infant, and to avoid influencing their child in any way. They were also told that they would be asked to set their infant on the floor directly in front of their lap when the experimenter gives the instruction, "Do you want to come and play?" towards the end of the experiment. Each experimental session consisted of a Free Play phase, a Demonstration phase, and a Test phase. Two video cameras recorded the infants' and experimenter's behavior during the session.

Free Play Phase This phase was identical to that of Experiment 1.

Demonstration Phase To begin the Demonstration phase, the experimenter placed the two large boxes on the stage about 20 cm apart, with their front curtains down. One of the boxes was a Uniform box, containing 4 balls of the same color hidden in its back compartment. The other box was a Variable box, containing 4 hidden balls of different colors instead. The experimenter also placed a transparent container in the space in front of the center of the two boxes. She then drew the infant's attention to the box on the left, saying "What's in this box?" The front curtain of this box was subsequently lifted, revealing a population of 6 different colored balls. The procedure that followed was identical to an individual test trial in Experiment 1, in which the experimenter seemingly tosses out 4 colored balls from the box at random, one after another with replacement. The only exception was that the \(4^{\text {th }}\) ball was returned to the box after 1 second, as looking behaviors were not of interest in Experiment 2. After this \(4^{\text {th }}\) ball was returned to the box, the experimenter said "All done!" She then pointed to the box on the right, and said "Let's see what's in this box!" The experimenter then repeated the steps performed on the previous box. This phase lasted approximately 3 minutes.

The boxes that were assigned as the Uniform or Variable box, as well as the colors of the sampled balls, were appropriately counterbalanced across infants.

Test Phase Each infant completed one test trial. The experimenter brought the two large boxes forward and set them down on the ground about 1 m from the infant, saying "Do you want to come and play?" Parents were instructed to let go of their infant if they had not done so at this point. When the infant touched one of the boxes, the experimenter started a timer and the test trial ended after 60 seconds.

Coding Infants were coded for the amount of time in which they were in contact with each of the boxes.

\section*{Results}

All of the infants' behaviors were coded offline. Preliminary analyses found no effects of gender or demonstration order (Uniform box first vs. Variable box first) on infants' exploration of the boxes. Subsequent analyses were collapsed over these variables.

Preliminary results show that infants played significantly longer with the Uniform box ( \(M=25.02 \mathrm{~s}, S D=26.06\) ) than the Variable box ( \(M=7.02 \mathrm{~s}, S D=11.78\) ), \(t(14)=2.08, p=\) \(.05, d=0.95\).

\section*{Discussion}

As predicted, infants played reliably longer with the Uniform box than the Variable box. These results replicate the findings in Experiment 1, demonstrating the infants are sensitive to coincidences that cannot be predicted by mere probabilities. In addition, our preliminary results indicate that infants do translate this sensitivity into action, selectively exploring the box that generated data which was indicative of a suspicious coincidence.

\section*{General Discussion}

We provide some suggestive evidence that infants are sensitive to coincidences, detecting these anomalous events even when they cannot be predicted by their mere event probabilities. In Experiment 1, infants were presented with a box that ostensibly generated balls under random sampling, creating an expectation that a sequence of tosses should result in a sampling distribution that is even across the 6 colors present in the box. Infants were surprised when the box consistently produced samples that were identical instead. This finding is impressive, considering that the two different sequences that infants saw in Experiment 1 had equal probabilities of occurring. Experiment 2 replicated this novel finding with an older age group through an action measure, and extended the finding by demonstrating that the sensitivity that infants show for coincidences translates into action, as infants preferentially explored a box that produced a sequence of four of the same colored balls, as compared to a box that produced a sequence of four different colored balls.

We speculate that the obtained differences in looking exploration times arise because infants are evaluating the data that they receive according to how well it supports different underlying causal models. For example, a sequence of four yellow balls is surprising because the event provides better support for the alternative theory that the box is rigged, rather than the original theory of random sampling. However, another interpretation of our results is possible, namely that infants (and adults) may consider the uniform sequence "yellow, yellow, yellow, yellow" to be lower in probability than the variable sequence "blue, green, red, yellow." Therefore, the results obtained may be due to a misunderstanding of event probabilities, rather than a consideration of alternative theories. More empirical work is thus necessary to parse these interpretations apart. That
being said, our results continue to provide evidence that young infants are sensitive to anomalous data, and will selectively explore the source of these anomalies.

One might also raise the representativeness heuristic as an alternative account of these results (Kahneman \& Tversky, 1972). By this account, infants preferred the box that generated the same sample each time because the samples (e.g. 4 yellow balls) were very dissimilar to the population from which they were drawn (i.e. 6 different colored balls). However, recent research examining infants' probabilistic reasoning has rendered this interpretation unlikely, as looking patterns were predicted by probabilistic reasoning and not by the representativeness heuristic when these two interpretations were pitted against each other (Denison \& Xu, 2010).

Besides understanding how infants’ representation of the presented events led to differences in their looking and exploration times, of interest in these studies is also why such differences arose. We believe that the ideas advanced by Griffiths \& Tenenabaum (2007) may shed light on this issue: infants pay attention to the coincidences that they encounter in the world, as these surprising events are likely to be rich sources of information for theory revision. By selectively investigating these events, children provide themselves with an opportunity to make a discovery that can enable them to revise their theories.

The present results thus provide tentative support for a growing set of findings demonstrating that infants and young children attend to the generative process of the data they observe and effectively consider between different models for the inputs that they receive (e.g. Gerken, 2010). Our results also bring additional insight to recent research demonstrating a Goldilocks effect in infant's allocation of attention, which found preferential attention to visual sequences that are neither too simple, nor too complex (Kidd, Piantadosi \& Aslin, 2012). In this research, experimenters found that 7-month-old infants are likely to look away earlier for events that are highly predictable. However, predictability is probably not the only determinant of how infants allocate their attention - although a sequence of 4 yellow balls in our experiments was highly predictable, infants paid greater attention to this event because it was inconsistent with their prior expectations about random sampling. Thus, we suggest that infants' consideration of the generative process for the observed data may also play an important role in their allocation of visual attention.

In summary, our experiments provide some suggestive evidence that infants may be sensitive to coincidences. Furthermore, this sensitivity translates into action, as infants preferentially explored the source of such anomalous data. However, many open questions remain: How did infants represent the events presented in our experiments? Why did infants show the obtained differences in looking and exploration time? What is the relationship between infants' looking times and their exploration? Future research examining these questions will provide us with a better grasp of infants' understanding of coincidence, and how
children eventually come to have an accurate idea of how the world works by adulthood.

\section*{Acknowledgments}

We thank Anjelica Benitez and the Berkeley Early Learning Lab (BELL) for their help in recruitment and stimuli building, and the parents and infants for their participation. We also thank Tom Griffiths and Noah Goodman for insightful discussions.

\section*{References}

Carey, S. (1985). Conceptual change in childhood. Cambridge, MA: MIT Press.
Carey, S. (2009). The origin of concepts. Oxford University Press, USA.
Denison, S., Reed, C. \& Xu, F. (2012) The emergence of probabilistic reasoning in very young infants: evidence from 4.5- and 6-month-old infants. Developmental Psychology.
Denison, S., \& Xu, F. (2010). Integrating Physical Constraints in Statistical Inference by 11-Month-Old Infants. Cognitive Science, 34(5), 885-908.
Gopnik, A., \& Meltzoff, A. N. (1997). Words, thoughts, and theories. Cambridge, MA: MIT Press.
Griffiths, T. L., \& Tenenbaum, J. B. (2007). From mere coincidences to meaningful discoveries. Cognition, 103(2), 180-226.
Gerken, L. (2010). Infants use rational decision criteria for choosing among models of their input. Cognition, 115(2), 362-366.
Horwich, P. (1982). Probability and evidence. Cambridge University Press.
Kahneman, D., \& Tversky, A. (1972). Subjective probability: A judgment of representativeness. Cognitive psychology, 3, 430-454.
Kidd, C., Piantadosi, S. T., \& Aslin, R. N. (2012). The Goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. PloS one, 7(5), e36399.
Nickerson, R. S. (2004). Cognition and chance: The psychology of probabilistic reasoning. Lawrence Erlbaum.
Owens, D. (1992). Causes and coincidences. Cambridge University Press.
Snow, J. (1855). On the mode of communication of cholera. John Churchill.
Teglas, E., Girotto, V., Gonzalez, M., Bonatti, L.L. (2007). Intuitions of probabilities shape expectations about the future at 12 months and beyond. Proceedings of the National Academy of Sciences of the United States of America, 104, 19156-19159.
Xu, F. \& Garcia, V. (2008). Intuitive statistics by 8-monthold infants. Proceedings of the National Academy of Sciences of the United States of America, 105, 5012-5015.

\title{
Exploring the Developmental Feedback Loop: Word Learning in Neural Networks and Toddlers
}

\author{
Clare E. Sims (Clare.Holtpatrick@Colorado.Edu) \\ Department of Psychology \& Neuroscience, 345 UCB \\ Boulder, CO 80309-0345 USA
}

\author{
Savannah M. Schilling (Savannah.Schilling@ Colorado.Edu) \\ Department of Electrical, Computer \& Energy Engineering, 425 UCB Boulder, CO 80309-0425 USA
}

\author{
Eliana Colunga (Eliana.Colunga@Colorado.Edu) \\ Department of Psychology \& Neuroscience, 345 UCB \\ Boulder, CO 80309-0345 USA
}

\begin{abstract}
Early word learning may be supported by a developmental feedback loop: the kind of words a child learns early on support the development of attentional biases, which in turn facilitate further word learning. In neural network simulations and a longitudinal study of toddlers we investigated how the emergence of an attentional bias to shape in word learning impacts vocabulary growth with respect to different kinds of words. If this relationship is causal, we should see that the emergence of a shape bias leads to an increase in the rate of learning of shape-based words relative to other kinds of words. The networks supported this prediction, showing an acceleration of shape- compared to material-based word learning. However, in toddlers, shape- and material-based words were learned similarly around the shape bias emergence. Implications are discussed for the developmental feedback loop account and causal relationships between attentional bias development and vocabulary growth.
\end{abstract}

Keywords: Word learning; shape bias; neural networks; longitudinal study.

\section*{Introduction}

Children are skilled word learners, in part because of constraints on the range of things they consider when inferring the referent of a new word. These constraints, sometimes referred to as biases, operate by helping children direct attention, resulting in sensitivity to what information matters most in the context of learning different kinds of words. Although there is debate on the origin of these attentional biases (e.g., Samuelson \& Bloom, 2008), evidence from children and from networks suggest that children can learn biases based on the kinds of nouns they acquire early on in their vocabularies (e.g., Colunga \& Smith, 2005; Gerhshkoff-Stowe \& Smith, 2004). This account entails a developmental feedback loop: the early nouns that children learn give rise to attentional biases, and those biases in turn guide further word learning and impact the structure of children's growing vocabularies. In the current paper we use data from neural networks and toddlers to investigate the latter part of this loop, focusing on how different types of words are learned right around the pivotal moment of word learning bias development.

\section*{Children's Early Vocabularies}

Words are an important building block in language and cognitive development. Children make the process of word learning look deceptively simple, typically acquiring their first word around the age of 1 year and experiencing a spike in vocabulary development around 18 months of age (Goldfield \& Reznick, 1990). Some researchers have observed that this vocabulary spike does not tend to occur until a child has at least 50 words in his or her vocabulary (Lucariello, 1987). Other work shows that the vocabulary spike is not only a function of the number of words a child knows, but also depends on the kinds of words that children learn. For example, Goldfield and Reznick (1990) observed that children exhibiting a vocabulary spike tended to add many words for objects (i.e., nouns) to their vocabularies. Children who did not show this dramatic increase in vocabulary size were steadily adding various types of words instead. This result suggests that while vocabulary size may be one key factor in children's language development, the specific kinds of words that children learn also play a role.

More recent research has investigated the question of why learning nouns may help accelerate subsequent vocabulary growth. One reason is that many different nouns have a basic property in common: they tend to refer to categories of things that are alike in shape. For example, a child will hear the word "ball" used to label a variety of spherical objects. Over time, children may pick up on the general pattern that shape is an important feature when talking about things in the world, and this insight in turn facilitates further word learning. Support for this account comes from a longitudinal study of young children's vocabularies (Gershkoff-Stowe and Smith, 2004). Over a three month period, 17-month-old children had a greater increase in object label nouns compared to other types of words. Over this same time period, children attended more to shape, over and above other features, when generalizing a newly learned word to novel objects. This result suggests that as children learn certain kinds of words, they also learn reliable patterns or constraints about how those words are used in the world. In
this paper we focus on one word learning constraint in particular: the shape bias.

\section*{The Shape Bias in Early Word Learning}

The shape bias is the tendency for children to generalize newly learned nouns to other objects based on similarity in shape. This is typically tested in novel noun generalization (NNG) tasks (Landau, Smith, \& Jones, 1988). In this type of task a child may be taught a novel name for a novel solid object. A shape bias is shown when the child extends that name to other objects matching the original in shape, even if the shape match differs from the original in texture, color, or size. Children show a reliable shape bias by 2 years of age (Samuelson \& Smith, 1999).

There is evidence that the emergence of the shape bias can guide children in learning new words. For instance, in one study 17 -month-olds were trained on shape-based categories of words, effectively teaching them the shape bias (Smith et. al, 2002). Not only did these children develop a shape bias earlier than the control group, they also showed accelerated growth in overall vocabulary. This suggests that there is an interaction between the development of word learning biases, particularly the shape bias, and vocabulary growth. This finding is one piece of evidence for a developmental feedback loop between vocabulary development and word learning constraints.

\section*{A Developmental Feedback Loop}

Smith and colleagues (2002) showed that teaching children the shape bias can promote vocabulary growth, but what about the other way around? Many of the previously mentioned studies show a correlation between these two factors. However, rather than word learning biases simply causing vocabulary growth, perhaps these are coupled phenomena that reciprocally influence each other. Previous modeling research suggests this. For instance, Colunga and Sims (2012) trained neural networks with early- and latetalker vocabulary structures as input and then tested for the development of word learning biases. Results showed that networks given late-talker vocabulary input produced different biases than networks with early-talker input. This shows that given only the structure of a child's vocabulary, the network can develop quantitatively different biases, suggesting that vocabulary growth affects bias development. These findings, combined with the experiments of Smith and colleagues (2002), indicate that vocabulary structure and word learning biases may be part of a development feedback loop in which both processes affect one another. Here we investigate the dynamics of this loop in both neural networks and in children.

In prior work, we explored dynamics of and interactions between the shape bias and other word learning biases over time. Neural networks were trained on a typical 30 -monthold child's vocabulary structure, then the bias dynamics were observed. We found that as the shape bias emerged, the development of other word learning biases diminished, suggesting a shift in attention as the shape bias is learned
(Schilling, Sims \& Colunga, 2012). These results were replicated in behavioral data from a longitudinal study of 18- to 30-month-old children (Sims, Schilling, \& Colunga, 2012).

In this paper, we look at the same emergence window, but this time concentrate on how different kinds of words are learned before, during and after shape bias emergence. That is, we focus on the other piece of the developmental feedback loop: how vocabulary structure changes as word learning biases develop. What kinds of words do networks and children learn right around the pivotal point of shape bias emergence? To test this, we used network models and vocabulary data from a longitudinal behavioral study.

\section*{Approach and Overview}

Our approach is to train a network on a typical early child vocabulary in order to observe learning over time that is similar to children's vocabulary development. We use a neural network to model the process of word learning, which differs from some other approaches to modeling word learning. For example, Bayesian networks extract generalities in order to produce a structured system representative of the real world (e.g., a mapping of a child's word representations; see \(\mathrm{Xu} \&\) Tenebaum, 2007). Our networks instead begin with structured representations as input and produce attentional biases. In order to investigate the developmental feedback loop, we are interested in the process: how the network forms these attentional biases from vocabulary structure input. We tested both networks and children on novel word generalization (using a virtual NNG task with the networks and a lab NNG task with children; see Sims et al., 2012 for details) over multiple points of vocabulary development. From this data, we identified the point in word learning at which the shape bias emerged for each individual network and child. Finally, we looked at the kinds of words that networks and children learned around their respective emergence points.

\section*{Network Simulations}

\section*{Method}

Our network was implemented in the Emergent software package (O’Reilly, Munakata, Frank, \& Hazy, 2012), to model word learning. The network was given input structured like that of a typical 30 -month-old's vocabulary. Throughout learning, we tracked what kinds of words the network successfully learned and tested for attentional biases. By analyzing the word learning biases the network developed and how they affected vocabulary structure, we were able to focus on the developmental feedback loop between attentional biases and vocabulary growth over time.

\section*{Network Dynamics}

Our models use the Leabra algorithm (Local, Error-driven and Associative, Biologically Realistic Algorithm), which combines Hebbian and error-driven learning (O'Reilly et al., 2012). The Hebbian, self-organizing learning uses longer time-scale statistics about the environment and is useful for
extracting generalities. However, this type of learning is not as good at compensating for specific, complicated patterns. Therefore, we use error-driven learning, which actively utilizes differences between expectations and outcomes. The total weight change in the network is the sum of that of the error-driven learning and that of the Hebbian learning.

The network uses a function called the eXtended Contrastive Attractor Learning (XCAL) rule. This function uses sending and receiving activity input and has a floating threshold that regulates changes in weights over learning. This function is used for both the Hebbian and error-driven learning with different inputs to the function. Inputs affect threshold changes and therefore different inputs elicit different weight change dynamics.

The error-driven weight changes are updated based on the short-term average connection activity ( \(\langle x y\rangle_{\mathrm{s}}\) ) and the medium-time scale average connection activity \(\left(\langle x y\rangle_{m}\right)\).
\[
\Delta w_{\text {error }}=f_{x c a l}\left(<x y>_{s},<x y>_{m}\right)=f_{x c a l}\left(x_{s} y_{s}, x_{m} y_{m}\right)
\]

Where \(\langle x y\rangle_{m}\) represents the emerging expectation about a current situation and \(\langle x y\rangle_{\text {s }}\) reflects the actual outcome and therefore the result of the received error information.

The Hebbian weight changes are based on the short-term connection activity ( \(\mathrm{xy}_{\mathrm{s}}\) ) and long-term average activity of the receiving unit ( \(\langle\mathrm{y}\rangle_{1}\) ).
\[
\Delta w_{\text {Hebbian }}=f_{x c a l}\left(x y_{s}, x<y>_{l}\right)=f_{x c a l}\left(x y_{s}, x y_{l}\right)
\]

Based on \(\langle\mathrm{y}\rangle_{1}\), the threshold for weight change is adjusted, making the weights more likely to change in the direction given by \(\mathrm{xy}_{\mathrm{s}}\). This creates the structure of generalization for the Hebbian learning mechanism. The combination of these two types of learning mechanisms allows for a balance of feed forward information to form categories and back propagation to allow adjustments based on errors. For more details on network dynamics, see O'Reilly et al. (2012).

\section*{Network Architecture}

The architecture is adapted from Colunga \& Smith (2005) and is implemented as shown in Figure 1. Words are represented discretely and are input on the Word Layer. Features are represented as distributed patterns over several dimensions on the Perceptual Layer. For example, the shape and material of an object (e.g., the roundness of a particular ball and its yellow rubbery material) are represented by an activation pattern along the Perceptual Layer, with 12 units for shape and 12 units for material. Solidity is represented locally; one unit stands for Solid and another for Non-Solid. Finally, there is a 25 unit Hidden Layer that is connected to all the other layers and to itself. The Hidden Layer serves as the bridge between the Word Layer (the sending units) and the Perceptual Layer (the receiving units) and it is where learning occurs. Learning progresses as internal representations, or weights, update and form representations which connect the other two layers.


Figure 1. Network architecture and example input patterns.

\section*{Vocabulary Structure: Network Input Patterns}

The input patterns used to train the network capture the structure of a child's vocabulary and are based on those used in Colunga and Smith (2005). They consisted of 100 noun representations, divided into 6 categories, with a structure analogous to the vocabulary of a typical 30 -monthold child (Fenson et. al, 1993). Categories were divided by both solidity (solid or non-solid) and characteristic feature (shape, material, or both), based on adult judgments. From these, the structure of a typical early vocabulary can be expressed as proportions of each category. Therefore, the network learning the entire set of training patterns represents a child learning a typical vocabulary. See Table 1 for the 6 categories and proportions used in the study.

These input patterns have a correlational structure such that a network learning them should produce a shape bias for solids (and indeed this was first shown by Colunga \& Smith, 2005). This means that learning in the network arises from the structure of the input patterns themselves. The purpose of the network, then, is not to help us discover structure in the input, but to observe the process of leveraging this structure over the course of word learning.

Table 1: Noun category percentages and example words.
\begin{tabular}{|r|c|c|c|}
\hline & Shape & Material & Both \\
\hline & \(52 \%\) & \(10 \%\) & \(12 \%\) \\
Solid & & \begin{tabular}{c}
1 \\
\\
\end{tabular} & ball
\end{tabular}

\section*{Network Training and Testing}

Over the course of training, the network formed biases based on the structure of the vocabulary input. On each trial of training, a word was paired with a pattern of features representing the features of the noun category. For example,
a word for a solid item characterized by shape (like a ball) should be used to label things that are like each other in shape but differ from each other in material. To simulate this pattern, we randomly selected an input vector to represent, for example, ball shape. On individual training trials, we paired that shape pattern with the label ball and a randomly selected material pattern. Therefore over multiple training trials, a word for a solid item characterized by shape would be represented by the same shape but different material patterns (see Figure 1). We did this for each of the 100 nouns in the training set.

We tested 10 runs of the network at multiple points throughout word learning. Weights and words learned from each of the 6 categories of interest were recorded at thirteen discrete checkpoints during the course of each training run. For example, the network was tested at 5 words learned, 10 words learned, and so on. The endpoint of learning was at 500 epochs of training, which was around when the network learned 75 words. For more information on network testing procedures, see Schilling et al. (2012).

\section*{Rationale and Predictions}

We focused our analyses of early child vocabulary composition, particularly shape-based and material-based words, on the period of time during which each network developed a shape bias in the context of solid objects. This approach may offer further insights into the relationship between attentional shifts in word learning and the course of vocabulary acquisition. As skilled attention to shape in the context of solid objects emerges, the networks should more easily learn shape-based words. Also, increased attention to shape may facilitate the learning of shape-based words over and above the learning of material-based words. This would be seen in a relatively lower rate of learning for materialbased compared to shape-based words.

\section*{Results}

The first question is how the networks learned shape-based words over the time window in which the shape bias emerged. The dependent measure was proportion of shapebased words learned at a given time point relative to the total number of shape-based words in the input vocabulary. Proportions of shape-based words learned were submitted to a linear regression with time point (before, at, or after shape bias emergence) as the predictor variable. Shape-based word learning increased significantly over time, \(b=0.06, t(28)=\) \(7.70, p<.001\). The networks showed significant increases in proportions of shape-based words learned from before shape bias emergence \((M=.01, S D=.02)\) to the point of emergence ( \(M=.06, S D=.03\) ), and from emergence to the following time point \((M=.14, S D=.06 ; t(9)>4.60, p \leq\) .001 , Cohen's \(d>1.45\) for both paired comparisons). That is, the networks' learning of shape-based words increased over time, and showed a particularly large increase following the emergence of the shape bias.

The next question is whether the networks' learning of shape-based words differed from learning of material-based
words over the same time period. Proportions of materialbased words learned were similarly computed relative to the total number of material-based words in the input vocabulary. Proportions of words learned were submitted to a multiple regression including time, word type, and the interaction between the two. Overall, these variables explained a significant proportion of the variance in the networks' word learning, \(R^{2}=.69, F(3,56)=40.90, p<\) .001. Consistent with the result above and the fact that the networks continually learned new words over time, time was a significant predictor of word learning overall, \(b=.03\), \(t(56)=4.80, p<.001\). The networks showed increases in learning both shape- and material-based words over the time window of interest (see Figure 2). Word type was also a significant predictor of learning, in that the networks on average learned a greater proportion of shape-based than material-based words, \(b=.03, t(56)=4.03, p<.001\). Additionally, the interaction between time point and word type predicted learning, \(b=.03, t(56)=3.07, p<.01\). As can be seen in Figure 2, there was a steeper increase in the trajectory of learning for shape- compared to material-based words over the time window of interest.


Figure 2. Shape- and material-based word learning in the network simulations over time.

\section*{Discussion}

The results of the network simulations show that the emergence of the shape bias coincided with changes in vocabulary acquisition for different kinds of words. Before the emergence of the shape bias, the networks steadily increased the amount of both shape- and material-based words in their vocabularies to an equal extent. However, after the emergence of the shape bias, learning of shapebased words outpaced learning of material-based words. This result adds support to a developmental feedback loop account of word learning. Adding to previous work showing that networks can learn attentional biases from the vocabulary input of a typical toddler (Colunga \& Smith, 2005; Schilling et al., 2012), the current study shows that in these same kinds of networks, attentional biases in turn influence the trajectories of subsequent vocabulary learning. Next we tested our network predictions in a behavioral study of toddlers.

\section*{Behavioral Study}

\section*{Rationale and Predictions}

To test the predictions of the network simulation we conducted a similar analysis on a longitudinal sample of toddlers. To explore vocabulary learning over time we looked at a parent-filled, standardized vocabulary inventory (the MacArthur-Bates Communicative Development Inventory [MCDI]; Fenson et al., 2007) that had been collected every month for a year for each child in the sample. As in the networks, we centered our analyses of child vocabulary development on the time at which each child first showed a shape bias for solid objects.

The network simulations predict that the emergence of the shape bias for solids leads to a change in the course of subsequent vocabulary learning. Specifically, this change was seen in the trajectory of shape-based relative to material-based word learning. If this prediction bears out in children, we should see a similar pattern in the toddler data.

\section*{Method}

\section*{Participants}

Nineteen children were recruited from the Boulder, CO area. Two children were excluded from the current analyses because they knew greater than \(80 \%\) of the nouns in the MCDI at the beginning of the time window of interest. The final sample analyzed here included 17 children ( \(M_{\text {age }}=\) 22.09 mo., \(S D=2.69\) mo., 9 girls).

\section*{Progression of Word Learning}

Children participated in 12 monthly visits over the course of one year. At each visit children were tested for attentional biases in novel word learning. There were three different stimuli sets, each consisting of an exemplar, and five test items matching the exemplar in shape, material and/or color. The children saw a single set in each visit and thus rotated through the three sets every three months. We calculated the point of emergence of the shape bias as in Sims et al. (2012), for each individual child, as defined by the child having shown a persistent shape bias during three consecutive visits, that is, for all three stimuli sets.

Vocabulary development was measured longitudinally through parent-completed, monthly MCDI checklists of words their child knew at the time of each visit to the lab. We focused our analyses on children's noun learning over the time period of interest. At the beginning of the analysis windows, children had on average 108 nouns ( \(S D=84\) nouns) from the MCDI in their vocabularies. To explore shape- and material-based word learning separately, we used the vocabulary structure classifications collected by Colunga and Smith (2005; see Table 1), combining solid and non-solid shape-based nouns, and solid and non-solid material-based nouns to get our two categories of interest.

\section*{Results}

As in the network simulation analyses, the first question we investigated in the behavioral data was whether children's learning of shape-based words increased over the window
during which each child developed a shape bias. The dependent measure was children's proportions of shapebased words learned at a given time point relative to each child's total number of shape-based words attained at the end of the study. These proportions were submitted to a linear regression with time point as the predictor. Shapebased word learning increased significantly over the time window of interest, \(\mathrm{b}=0.13, t(43)=2.26, p=.03\). Post-hoc paired comparisons showed increases in words learned from before shape bias emergence \((M=.50, S D=.34)\) to the point of emergence ( \(M=.64, S D=.32\) ), and to the time point after emergence \((M=.76, S D=.23 ; t(14)>3, p<.01\), \(d>.90\) for both).

Next we compared children's learning of shape-based words to their learning of material-based words over the same time window. Proportions of words learned were submitted to a multiple regression including time, word type, and the interaction between the two. Overall, these predictors explained a significant proportion of the variance in children's word learning, \(R^{2}=.13, F(3,86)=4.32, p<\) .01. Time point was a significant predictor, showing that the proportions of both shape- and material-based words increased over the time window of interest, \(b=.13, t(86)=\) \(2.18, p=.03\). Word type was a marginally significant predictor of learning, \(b=.12, t(86)=1.86, p=.07\). As can be seen in Figure 3, the proportion of shape-based words learned ( \(M=.63, S D=.32\) ) tended to be higher than the proportion of material-based words learned \((M=.51, S D=\) .34) across all time points. The interaction between time point and word type was not a significant predictor of learning. That is, children's learning of shape- and materialbased words followed the same trajectory.


Figure 3. Shape- and material-based word learning in the behavioral study of toddlers over time.

\section*{Discussion}

These results show that there is an increase in the number of shape based words that children learn as the shape bias emerges. This result is consistent with the networks and supports one piece of the developmental feedback loop in children. However, unlike the network simulations, the increases in children's word learning did not show a marked acceleration for shape-based words. Although proportionally more shape-based words were learned
compared to material-based words, the trajectory of learning for these two types of words did not differ significantly within this time window.

\section*{General Discussion}

In the current studies we found that vocabulary learning around the emergence of the shape bias supported the developmental feedback loop account in our network simulations, but toddlers showed a different pattern. Adding to our previous work with these word learning networks, the current simulations contribute evidence for the effects of attentional biases on subsequent vocabulary learning. The behavioral data show ambiguous results in relation to the developmental feedback loop. There are several possible reasons for this pattern that will inform future research.

Methodological constraints could have contributed to these intriguing results. For example, the networks' performance in the generalization task is much more consistent than the children's performance. Even though individual networks do vary on the epoch at which they show a shape bias, once it emerges, it stays. This is not the case for children. To deal with this, we used a stringent criterion to define the time of emergence for the children by making sure that the preference for shape was present during three consecutive visits, for three different stimuli sets. Probably because of this criterion, the points of bias emergence that we identified tended to occur when children had on average over 100 nouns, with high variation between individuals. This suggests that we may have identified shape bias emergence relatively late in vocabulary development for some of the children in the sample, at least when compared with the criterion used in Gershkoff-Stowe \& Smith (2004). A related possibility is that our network shows bias emergence and subsequent vocabulary changes at a relatively earlier (or "younger") point than the toddlers in our sample. If this is the case, vocabulary changes in the network may be easier to detect because it has progressed less far in the proportion of words learned and thus can statistically show greater growth compared to the toddlers. On the other hand, when the shape bias emerges in toddlers, they already know over half of the words in the MCDI, and thus have relatively less room for growth. Nonetheless, we would still expect to see differences in how toddlers learn shape- and material-based words, yet these interactions are either not present or not being captured by our current measures. In future analyses we plan to explore other measures such as rate of vocabulary growth that may better equate learning in the network and in toddlers. We also plan to look at dynamic attention to shape as a continuous measure over the entire trajectory of learning. Perhaps the emergence of the shape bias puts into motion long-term, rather than immediate, changes in the trajectory of vocabulary growth.

More interesting than methodological explanations are the theoretical implications of this finding. The behavioral results, along with our previous work (Sims et al., 2012), suggest that vocabulary growth precedes bias development,
but the causality of this relationship may not go the other way. Perhaps once children have a consistent shape bias for solids, and those words come easy, they begin to focus on other kinds of words that do not conform to their expectations of naming categories organized by shape. Further work is necessary to see if that is the case. Overall, our novel approach using neural networks allows us to explore not just a causal effect of biases on vocabulary, but the dynamic feedback relationship between the two, the very relationship that builds the developmental trajectory.

\section*{References}

Colunga, E., \& Sims, C. E. (2012). Early-talker and latetalker toddlers and networks show different word learning biases. Proc. of the 34th Ann. Conf. of the Cog. Sci. Society. pp. 246-251. Austin, TX: Cog. Sci Society.
Colunga, E., \& Smith, L. B. (2005). From the lexicon to expectations about kinds: A role for associative learning. Psych. Review, 112, 347-382.
Fenson, L., Dale, P., Reznick, J. S., Thal, D., Bates, E., Hartung, J., et al. (1993). The MacArthur-Bates Communicative Development Inventory: User's guide and technical manual. San Diego, CA: Singular Pub.
Gershkoff-Stowe, L., \& Smith, L. B. (2004). Shape and the first hundred nouns. Child Development, 75, 1098-1114.
Goldfield, B. A., \& Reznick, S. (1990). Early lexical acquisition: Rate, content, and the vocabulary spurt. Journal of Child Language, 17, 171-183.
Landau, B., Smith, L. B., \& Jones, S. S. (1988). The importance of shape in early lexical learning. Cog. Dev, 3, 299-321.
Lucariello, J. (1987). Concept formation and its relation to word learning and use in the second year. Journal of Child Language, 14, 309-332.
O’Reilly, R. C., Munakata, Y., Frank, M. J., \& Hazy, T. E. (2012). Computational Cognitive Neuroscience.

Samuelson, L. K., \& Bloom, P. (2008). Special section: What counts as an explanation of development: The shape bias as a case study. Developmental Science, 11, 183-184.
Samuelson, L. K., \& Smith, L. B. (1999). Early noun vocabularies: Do ontology, category structure and syntax correspond? Cognition, 73, 1-33.
Schilling, S.M., Sims, C.E., \& Colunga, E. (2012). Taking development seriously: Modeling the interactions in emergence of different word learning biases. Proc. of the 34th Ann. Conf. of the Cog. Sci. Society. pp. 246-251. Austin, TX: Cog. Sci Society.
Sims, C. E., Schilling, S. M., \& Colunga, E. (2012). Interactions in the development of skilled word learning in neural networks and toddlers. Proc. of the ICDL/EpiRob. pp. 1-6. San Diego, CA.
Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., \& Samuelson, L. K. (2002). Creating a shape bias creates rapid word learners. Psych. Sci., 13, 13-19.
Xu, F. \& Tenebaum, J. B. (2007). Word Learning as Bayesian Inference. Psych. Review, 114, 245-272.

\title{
Syntax in music and language: The role of cognitive control
}

\author{
L. Robert Slevc (slevc@umd.edu) \\ Department of Psychology, University of Maryland \\ College Park, MD 20742 USA
}

\author{
Jason G. Reitman (jreitman@wesleyan.edu) \\ Wesleyan University \\ Middletown, CT 06459 USA \\ Brooke M. Okada (bokada@umd.edu) \\ Department of Psychology, University of Maryland \\ College Park, MD 20742 USA
}

\begin{abstract}
The relationship between structural (or syntactic) processing in music and in language is not yet clear. Evidence indicating that these two processes are shared conflicts with other results suggesting that they are largely distinct. These conflicting findings suggest that musical and linguistic processing may share some, but not all, underlying processes, raising the question of what exactly those shared processes might be. Two experiments tested the idea that one shared process is cognitive control by pairing manipulations of musical structure with the Stroop task, a standard test of cognitive control. Manipulations of harmonic expectancy, but not of timbral expectancy, interacted with Stroop interference effects, suggesting that cognitive control is at least one specific process underlying shared syntactic processing in music and language.
\end{abstract}

Keywords: cognitive control; music and language; musical syntax

\section*{Introduction}

Interest in the relationship between music and language has a long history, dating back at least to Darwin (1871). Over the last several years, a more specific focus on the relationship between structural (or syntactic) processing in music and language has received increasing attention (for recent reviews, see Patel, 2008; Slevc, 2012; Tillmann, 2012). This likely is due, at least in part, to an influential proposal about the relationship between structural processing in language and music: Patel's (2003) shared syntactic integration resource hypothesis (SSIRH; see also Patel, 2008). The SSIRH proposes that music and language involve separate representations (e.g., nouns and verbs in language, tonal functions in music), but recruit a shared set of cognitive resources that are required to integrate these separate representations into evolving sequences.

There is a growing body of evidence supporting the SSIRH. Much of this evidence comes from experiments using interference paradigms, where participants are simultaneously presented with both musical and linguistic stimuli. In these paradigms, syntactic manipulations in both domains are crossed to look for interactive effects that indicate shared processing (vs. additive effects, which would indicate independent processes). For example, an
electrophysiological response characteristic of a violation of linguistic syntax (the left anterior negativity, or LAN) is reduced when linguistic syntactic violations are paired with a concurrent music-syntactic irregularity (Koelsch, Gunter, Wittfoth, \& Sammler, 2005). Similarly, sung complex sentences are especially difficult to understand when critical regions are sung out-of-key (Fedorenko, Patel, Casasanto, Winawer, \& Gibson, 2009).

Other behavioral evidence comes from on-line sentence processing paradigms, where readers' slowed processing of temporary syntactic ambiguities is especially pronounced when the disambiguating word is paired with a harmonically unexpected chord (Slevc, Rosenberg, \& Patel, 2009; also see Hoch, Poulin-Charronnat, \& Tillmann, 2011, for related findings). More specifically, Slevc et al. (2009) relied on garden path effects, where readers are slower to comprehend the disambiguating word was in a sentence like "The scientist proved the hypothesis was false" compared to an unambiguous context like "The scientist proved that the hypothesis was false." This slowed processing presumably reflects the need to revise an initial syntactic interpretation where the hypothesis was interpreted as the direct object of the verb proved rather than as the subject of an embedded sentence complement. This garden path effect was more pronounced when the disambiguating word (was) was accompanied by a harmonically unexpected chord (but not when accompanied by a chord of unexpected timbre).

Importantly, such an interaction did not emerge for semantically unexpected words (e.g., pigs as a continuation of "The mailman was afraid of...") suggesting that the interactive processes are specific to syntax. However, a more recent finding casts doubt on this last conclusion: the same harmonic manipulations did lead to interactive effects when paired with sentences containing "semantic garden paths" such as "The old man went to the bank to withdraw his net which was empty" (Perruchet \& Poulin-Charronnat, 2013). Thus it seems that these interactive effects (and the shared integration resource of the SSIRH) may not be specific to syntactic processing per se.

In addition, some recent neuroimaging studies have not found substantial overlap between neural regions implicated in the processing of language and music (Fedorenko, Behr,
\& Kanwisher, 2011; Rogalsky, Rong, Saberi, \& Hickok, 2011). These studies compared a linguistic contrast (activation to intact sentences versus non-words or versus jabberwocky sentences) to a musical contrast (activation to intact music versus scrambled music or versus silence), and found little evidence for shared regions implicated in both contrasts.

These conflicting findings raise the question: what kind of shared process links musical structural processing to some aspects of linguistic processing (including syntactic errors, syntactic complexity, and both syntactic and semantic garden paths), but not to other aspects such as the processing of semantically surprising words and the difference between intact and scrambled sentences? One way to characterize this distinction is that the aspects of language processing that are related to musical structure require not only the processing of an unexpected element, but also the resolution of conflict between this unexpected information and a current representation of an incrementally constructed (and/or predicted) structure. The unrelated aspects of language, in contrast, may not place demands on conflict resolution per se as there is no obvious way to resolve a semantic anomaly or a scrambled sentence.

This sort of conflict detection and resolution process is termed cognitive control, referring broadly to the cognitive processes that allow for the regulation of mental activity required to resolve competing representations (see, e.g., Miller \& Cohen, 2001). Cognitive control processes have been implicated in various aspects of language processing, including parsing of garden path sentences and semantic plausibility effects (see Novick, Trueswell, \& ThompsonSchill, 2010, for a recent review), and it is possible that the types of linguistic manipulations that interact with musical structure are of this general type. By this account, studies finding interactive effects between musical structure and language (be it linguistic syntax or non-syntactic situations that require resolution between conflicting representations like semantic garden paths) might reveal simultaneous use of cognitive control resources.

This account implies that musical syntactic processing-at least as measured in the studies cited above-also relies on cognitive control mechanisms. Indeed, this is likely to be the case. Listening to music involves building up complex cognitive representations of musical structure. This not only involves processing and integrating musical elements as they occur, but also incrementally generating and evaluating predictions based on implicit knowledge of musical structure (see Rohrmeier \& Koelsch, 2012, for discussion). One hazard of this predictive processing is that new information can be inconsistent with one's prediction, thus harmonic processing requires the ability to both detect conflict between predicted and observed precepts and the ability to resolve this conflict by overriding and updating an evolving representation of musical structure. Conflict between musical precepts and predictions likely arises in many situations, not the least of which are cases of musical ambiguity (e.g., musical garden paths; Temperley, 2001).

One form of indirect evidence for a role of cognitive control in musical syntax comes from neuroimaging findings. Regions in the inferior frontal gyrus (including Broca's area) that are linked to cognitive control processes (both generally and in language processing; e.g., Miller \& Cohen, 2001; Novick et al., 2010) have also been implicated in neuroimaging studies of musical syntactic processing (albeit more bilaterally or even right lateralized; e.g., Maess, Koelsch, Gunter, \& Friederici, 2001; Tillman, et al., 2006).

A second form of indirect evidence for a role of cognitive control in musical syntactic processing comes from evidence that musical training is associated with advantages in cognitive control ability (Bialystok \& DePape, 2009; Pallesen et al., 2010) among other types of cognitive advantages (e.g., Schellenberg, 2006). A musician advantage in cognitive control could plausibly result from the additional demands placed on cognitive control mechanisms from extensive musical training and practice, but only if those demands tax (and thus potentially strengthen) cognitive control processes.

\section*{Current experiments}

The aim of the experiments reported here was to provide a direct test of whether cognitive control mechanisms are involved in musical syntactic processing (as has been argued to be the case for linguistic syntactic processing). If cognitive control processes are, in fact, an important part of musical syntactic processing, less expected chords should impose relatively greater demands on cognitive control. Assuming cognitive control is a limited-capacity resource, this should lead to a temporary reduction in the ability to exercise cognitive control in other tasks.

In order to measure demands on cognitive control, we turned to a prototypical cognitive control task: the Stroop task (Stroop, 1935; see McLeod, 1991, for a review). In the standard Stroop task, participants must name the ink (or font) color of printed stimuli. These stimuli can be of three types: congruent, where the printed word is the same as the to-be-named ink color (e.g., the word "GREEN" printed in green font), incongruent, where the printed word is a different color name than the two-be-named ink color (e.g., the word "BLUE" printed in green font), and neutral (e.g., the string "\#\#\#\#" printed in green font). Cognitive control demands are reflected in Stroop interference, where responses are slower to incongruent than to neutral trials. The Stroop task can also yield Stroop facilitation, reflected in faster responses to congruent than neutral trials, however these facilitative effects are not generally assumed to result from demands on cognitive control.

Because Stroop interference is a prototypical measure of cognitive control, it can be used as an index of cognitive control demands at a given moment. The experiments presented below do just this by investigating if, and how, Stroop interference is affected by a concurrent musical syntactic manipulation.

\section*{Experiment 1}

In Experiment 1, participants performed a standard Stroop task while hearing musical chorales. The primary question was whether the harmonic expectancy of a chord occurring during a trial of the Stroop task would influence Stroop interference effects.

\section*{Method}

Participants Twenty-five undergraduate students from the University of Maryland participated in exchange for course credit. Participants were unselected with regard to musical training.

Materials Stimuli for the Stroop task were the strings "RED", "GREEN", "BLUE", or "XXXX". The word stimuli appeared half of the time in a congruent color (e.g., the word "BLUE" in blue font) and half of the time in an incongruent color (e.g., "BLUE" in green or red font); the neutral ("XXXX") stimuli appeared equally often in each of the three font colors. Because the primary effect of interest here is in Stroop interference (vs. facilitation), congruent trials were treated as fillers and excluded from analysis.

Musical stimuli were twelve six-chord chorales based on Western musical structure, half in major and half in minor keys. Each chorale ended either on a tonic chord (the tonal center of the chorale's key) or ended on a chord belonging to another key, and thus was either harmonically expected or unexpected (see Figure 1 for an example). In addition, each chorale occurred once more as a filler item; these fillers ended on a variety of chords. While these fillers were generally harmonically unexpected, they were not constructed in a theoretically constrained way and so were excluded from analyses. (Note, however, that treating these fillers as harmonically unexpected trials does not substantially alter the main pattern of results.)


Figure 1: Example stimuli in Experiment 1. The top panel shows an example musical chorale ending in a harmonically expected tonic chord (A) or a harmonically unexpected chord from a different key (B). The bottom panel represents the incongruent (i) or neutral (ii) visual target for the primary color response task.

Procedure Participants were tested individually on iMac computers using PsychoPy software (Pierce, 2007). The primary task was to respond to the color of the visual stimuli (red, green, or blue) by pressing a corresponding button (the left, down, or right arrow, respectively). These color/button mappings were presented on the screen during the entire task. Participants first performed a practice block of twenty-one color-naming trials (without concurrent musical stimuli) to learn the color/key mappings, then a second practice block of ten trials where the target stimulus appeared at the onset of the final chord of a six-chord chorale (all practice chorales ended on the tonic). Finally, participants performed the experimental block consisting of 72 chord sequences ending on the tonic, 36 chord sequences ending on a harmonically unexpected chord, and 36 filler sequences. Within each musical condition, one third of the trials were neutral, one third were incongruent, and one third were congruent (filler) trials.

A schematic of the four conditions for an individual experimental trial is shown in Figure 1.

Design and analysis Response times were log-transformed and analyzed using linear mixed-effects models in the statistical software R (version 2.15.2; R Development Core Team, 2012). Stroop trial type (text condition: incongruent or neutral) and the harmonic role of the final chord (harmonic condition: expected or unexpected) were entered as fixed effects using orthogonal contrast coding. The fully specified random effect structure was included for both participants and items (see Barr, Levy, Scheepers, \& Tily, 2013), but random effects are not reported as only fixed effects were of theoretical interest. The current implementation of lme4 does not compute \(p\) values for models that include random slopes, therefore we follow Gelman and Hill (2007) by assuming that any absolute \(t\) value greater than 2 indicates a significant effect.

\section*{Results}

A significant main effect of text condition \((b=-0.24, S E=\) \(0.029, t=-8.23\) ) revealed (unsurprisingly) that responses were slower for incongruent than neutral strings. There was not a significant main effect of harmonic condition ( \(b=\) \(0.0048, S E=0.022, t=-0.22\) ), however a significant interaction between harmonic condition and text condition ( \(b=-0.10, S E=0.042, t=-2.39\) ), revealed that the Stroop interference effect was significantly larger when accompanied by an unexpected final chord. The Stroop interference effects in the harmonically expected and unexpected conditions are plotted in Figure 2.

\section*{Discussion}

Experiment 1 found larger Stroop interference effects when Stroop trials occurred during structurally unexpected chords, suggesting that the processing of harmonically unexpected chords involves an underlying process that is shared with Stroop interference. This bolsters theoretical reasons to expect cognitive control processes to play a role in musical
syntactic processing and adds to previous indirect evidence for such a relationship, such as (bilateral) inferior frontal activation associated with musical syntax (e.g., Tillmann et al., 2006) and cognitive control advantages associated with musical training (e.g., Bialystok \& DePape, 2009).


Figure 2: Stroop interference (incongruent minus neutral) by the harmonic condition of the final chord (the tonic chord or an unexpected chord from another key) in Experiment 1.

Data are plotted as untransformed means of participant means and dots indicate individuals' scores. \({ }^{1}\)

A counter explanation for these results might be that the harmonically unexpected chord drew attention away from the primary task of responding to font colors, thus leading to an exaggerated effect of incongruent trials due not to shared mechanisms, but simply to surprise or distraction. It is not obvious that simple distraction would enhance the Stroop effect as this distraction would presumably affect neutral as well as incongruent trials. However it is nevertheless possible that some aspect of the musical manipulation besides its harmonic unexpectedness is responsible for the observed interaction. Experiment 2 aimed to address this possibility by using a noticeable, but not music-syntactically relevant, manipulation of timbre, or sound quality.

\section*{Experiment 2}

If the interaction between Stroop interference and harmonic expectancy found in Experiment 1 results from surprise or distraction, then similar results should emerge when another type of unexpected auditory event occurs, even if that event

\footnotetext{
\({ }^{1}\) For ease of interpretation, Figures 2 and 3 display means of participants' mean untransformed RTs; note, however, that analyses were conducted over non-averaged, log-transformed RTs.
}
does not require conflict resolution. If, on the other hand, the interaction observed in Experiment 1 reflects the role of conflict resolution processes, then such interactions should not arise unless the unexpected stimuli induces some degree of resolvable conflict with an incremental and predictive cognitive representation of musical structure.

To examine these possibilities, Experiment 2 employed the same design as Experiment 1, but instead of manipulating musical syntactic expectancy, manipulated musical timbre (cf. Slevc et al., 2009). In contrast to the harmonic manipulation in Experiment 1, where an unexpected chord could reflect some kind of resolvable modulation or other harmonic "twist," there is no obvious way to resolve an unexpected timbre. Thus, a chord of unexpected timbre (that plays an expected harmonic role) should not lead to conflict resolution processes, and should not interact with the Stroop interference effect.

A chord of unexpected timbre should, however, be at least as surprising and attention demanding as an out-of-key chord, so if the interaction observed in Experiment 1 results from surprise or distraction, the same pattern of results should emerge in Experiment 2.

\section*{Method}

Participants Thirty undergraduate students from the University of Maryland participated in exchange for course credit or for a small (\$5) payment. As in Experiment 1, participants were unselected with regard to musical training.

Materials and Procedure The visual stimuli in Experiment 2 were identical to those in Experiment 1. The musical chorales were also identical except for the final chords in the unexpected conditions, which were always the tonic chord (thus always harmonically expected) but varied in terms of their timbre. Specifically, the final chord was either of the expected piano timbre, i.e., the same timbre as the rest of the chorale, or was played in a distinct timbre (the sitar timbre, as implemented in MuseScore version 1.2). An additional set of trials ended with a timbre only slightly different from the rest of the chorale (MuseScore's ukulele timbre, which sounds remarkably similar to a piano); as in Experiment 1, these intermediate trials were treated as fillers and not included in the analysis.

Design and Analysis Response times were analyzed just as in Experiment 1. Experiment 2 crossed the fixed-effects factors of text condition with timbral condition (expected or unexpected timbre) with the maximal random effects structure supported by the data. \({ }^{2}\)

\section*{Results}

As in Experiment 1, participants were reliably slower to respond to the color during incongruent trials than

\footnotetext{
2 The by-item random slopes for timbral condition and the timbral condition by text condition interaction had to be removed for the statistical model to converge.
}
congruent trials (i.e., a significant effect of text condition; \(b\) \(=-0.17, S E=0.023, t=-7.14)\). There was no significant effect of the timbre of the final chord \((b=0.20, S E=0.14, t\) \(=1.42\) ) and, unlike in Experiment 1, no interaction between these factors \((b=-0.16, S E=0.029, t=-0.55)\). The Stroop interference effects in the timbrally expected and unexpected conditions are plotted in Figure 3.


Figure 3: Stroop interference (incongruent minus neutral) by the timbral expectancy of the final chord (same vs. different timbre) in Experiment 2. Data are plotted as untransformed means of participant means and dots indicate individuals' scores.

\section*{Discussion}

Experiment 2 showed a standard effect of Stroop interference, but this effect did not interact with the timbre of a concurrent (tonic) chord. If anything, the average magnitude of the Stroop interference effect was numerically smaller in the unexpected timbre condition. This suggests that the interactive effects found in Experiment 1 did not result simply from the attention capturing nature of the unexpected stimuli, but were rather a function of the need to resolve conflict between the final chord and its expected harmonic role.

\section*{General Discussion}

The experiments reported here tested the idea that the processing of musical structure relies on general cognitive processes of cognitive control, which are also thought to underlie aspects of language processing. Experiment 1 crossed a standard cognitive control task-the Stroop taskwith a manipulation of harmonic expectancy, and found interactive effects: Stroop interference was exacerbated when accompanied by a structurally unexpected chord.

Experiment 2 showed that this interaction between harmonic expectancy and Stroop interference was not simply due to distraction or divided attention as such an effect did not emerge when Stroop trials were paired with chords of an unexpected timbre.

The interactive effects in Experiment 1 are perhaps especially notable as there was no musical task requiring participants to pay attention to the musical chorales, and participants were an unselected group of undergraduate students, not a group of musicians. These observations suggest that these effects do not depend on any particularly effortful type of musical processing, but instead reflect the broad knowledge of musical syntax that arises simply through a lifetime of exposure to a specific musical tradition (cf. Bigand \& Poulin-Charronnat, 2006).

These findings support the theory that cognitive control does, in fact, underlie at least some of the shared processing resources implicated in linguistic and musical syntax. A further implication is that the processing of both music and language should overlap to some extent with a variety of other domains that also rely on these general mechanisms of cognitive control (not just simple cognitive tasks like the Stroop task). There is already evidence for some such relationships; for example, interactive effects have been demonstrated during simultaneous processing of music and arithmetic (Hoch \& Tillmann, 2012). In addition, "action syntax" or "scripts" (i.e., meaningful structured representations of action sequences) may be related to both linguistic syntax (e.g., Farag et al., 2010) and to musical syntactic processing (Harding et al., 2011).

These relationships between different types of structural processing - be those structures musical, linguistic, mathematical, or action schemas - are not likely to reflect a syntax-specific shared underlying process. Instead, these processes likely draw on the same cognitive mechanisms to deal with similar demands, and these data suggest that one such mechanism is cognitive control. Of course, it is unlikely that the relationship between structural processing in language and music reflects only cognitive control mechanisms. Musical and linguistic structure are rich and complex systems that surely draw on a variety of cognitive mechanisms. Nevertheless, these data take a step towards a more specific account of exactly what sort of shared integration resources might underlie linguistic and musical syntax by implicating the well-studied cognitive construct of cognitive control.

\section*{Acknowledgments}

The authors wish to thank Ilana Green for help with experimental stimuli and data collection, and Jared Novick, Aniruddh Patel, Sharon Thompson-Schill, and Matt Weber for discussions that (perhaps unbeknownst to them) helped to motivate these experiments.

\section*{References}

Barr D. J., Levy R., Scheepers C. \& Tily, H. (2013). Random-effects structure for confirmatory
hypothesis testing: Keep it maximal. Journal of Memory and Language, 68(3), 255-278.
Bialystok, E. \& DePape, A-M. (2009). Musical expertise, bilingualism, and executive functioning. Journal of Experimental Psychology: Human Perception and Performance, 35(2), 565-574.
Bigand, E. \& Poulin-Charronnat, B. (2006). Are we "experienced listeners"? A review of the musical capacities that do not depend on formal musical training. Cognition, 100, 100-130.
Darwin, C. (1871). The Descent of Man, and Selection in Relation to Sex. London: John Murray.
Farag C., Troiani V., Bonner M., Powers C., Avants B., Gee J., Grossman M. (2010). Hierarchical organization of scripts: converging evidence from FMRI and frontotemporal degeneration. Cerebral Cortex, 20(10), 2453-2463.
Fedorenko, E., Behr, M. K., \& Kanwisher, N. (2011). Functional specificity for high-level linguistic processing in the human brain. Proceedings of the National Academy of Sciences, 108(39), 16428-16433.
Fedorenko, E., Patel, A. D., Casasanto, D., Winawer, J., \& Gibson, E. (2009). Structural integration in language and music: evidence for a shared system. Memory and Cognition, 37, 1-9.
Gelman, A. \& Hill, J. (2007). Data Analysis Using Regression and Multilevel/Hierarchical Models. New York, NY: Cambridge University Press.
Harding, E., Sammler, D., D'Ausilio, A., Friederici, A., Fadiga, L., \& Koelsch, S. (2011). Explicit action perception shares resources with music syntax: A controlled behavioral study. Poster presented at The Neurosciences and Music IV: Learning and Memory, Edinburgh, United Kingdom.
Hoch, L., Poulin-Charronnat, B., \& Tillmann, B. (2011) The influence of task-irrelevant music on language processing: Syntactic and semantic structures. Frontiers in Psychology, 2, 112. doi: 10.3389/ fpsyg.2011.00112
Hoch, L. \& Tillmann, B. (2012). Shared structural and temporal integration resources for music and arithmetic processing. Acta Psychologica, 140, 230-235.
Koelsch, S., Gunter, T. C., Wittfoth, M., \& Sammler, D. (2005). Interaction between syntax processing in language and in music: An ERP study. Journal of Cognitive Neuroscience, 17(10), 1565-1577.
Maess, B., Koelsch, S., Gunter, T. C., \& Friederici, A. D. (2001). Musical syntax is processed in Broca's area: an MEG study. Nature Neuroscience, 4(5), 540-545.
MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. Psychological Bulletin, 109(2), 163-203.
Miller, E. K., \& Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. Annual Reviews of Neuroscience, 24, 167-202.
MuseScore (Version 1.2) [Computer software]. URL: http://musescore.org/

Novick, J. M., Trueswell, J. C. \& Thompson-Schill, S. L. (2010). Broca's area and language processing: Evidence for the cognitive control connection. Language and Linguistics Compass, 4, 906-924.
Pallesen K. J., Brattico E., Bailey C. J., Korvenoja A., Koivisto J., Gjedde, A., \& Carlson, S. (2010) Cognitive Control in Auditory Working Memory Is Enhanced in Musicians. PLoS ONE 5(6): el1120.
Patel, A. D. (2003). Language, music, syntax and the brain. Nature Neuroscience, 6(7), 674-681.
Patel, A. D. (2008). Music, Language, and the Brain. Oxford University Press.
Peirce, J. W. (2007) PsychoPy - Psychophysics software in Python. Journal of Neuroscience Methods, 162(1-2), 813.

Perruchet, P. \& Poulin-Charronnat, B. (2013). Challenging prior evidence for a shared syntactic processor for language and music. Psychonomic Bulletin \& Review, 20, 310-317.
R Development Core Team. (2012). R: A language and environment for statistical computing (Version 2.15.2): R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL: http://www.R-project.org.
Rohrmeier, M. A. \& Koelsch, S. (2012). Predictive information processing in music cognition. A critical review. International Journal of Psychophysiology, 83, 164-175.
Rogalsky, C., Rong, F., Saberi, K., \& Hickok, G. (2011). Functional anatomy of language and music perception: temporal and structural factors investigated using functional magnetic resonance imaging. Journal of Neuroscience, 31(10), 3843-3852.
Schellenberg, E. G. (2006). Long-term positive associations between music lessons and IQ. Journal of Educational Psychology, 98(2), 457-468.
Slevc, L. R. (2012). Language and music: sound, structure, and meaning. Wiley Interdisciplinary Reviews: Cognitive Science, 3(4), 483-492.
Slevc, L. R., Rosenberg, J. C., \& Patel, A. D. (2009). Making psycholinguistics musical: Self-paced reading time evidence for shared processing of linguistic and musical syntax. Psychonomic Bulletin \& Review, 16(2), 374-381.
Stroop, J. R. (1935). Studies of interference in serial verbal reactions. Journal of Experimental Psychology, 18, 643662.

Temperley, D. (2001). The Cognition of Basic Musical Structures. Cambridge, MA: MIT Press.
Tillmann, B., Koelsch, S., Escoffier, N., Bigand, E., Lalitte, P., Friederici, A. D., \& von Cramon, D. Y. (2006). Cognitive priming in sung and instrumental music: Activation of inferior frontal cortex. NeuroImage, 31, 1771-1782.
Tillmann, B. (2012). Music and Language Perception: Expectations, Structural Integration, and Cognitive Sequencing. Topics in Cognitive Science, 4(4), 568-584.

\title{
Modeling the Effects of Formal Literacy Training on Language Mediated Visual Attention
}

\author{
Alastair C. Smith (alastair.smith@mpi.nl) \\ Max Planck Institute for Psycholinguistics, Wundtlaan 1, Nijmegen, 6525 XD, The Netherlands \\ Padraic Monaghan (p.monaghan@lancaster.ac.uk) \\ Department of Psychology, Lancaster University, Lancaster, LA1 4YF, U.K.
}

Falk Huettig (falk.huettig@mpi.nl)
Max Planck Institute for Psycholinguistics, Wundtlaan 1, Nijmegen, 6525 XD, The Netherlands

\begin{abstract}
Recent empirical evidence suggests that language-mediated eye gaze is partly determined by level of formal literacy training. Huettig, Singh and Mishra (2011) showed that highliterate individuals' eye gaze was closely time locked to phonological overlap between a spoken target word and items presented in a visual display. In contrast, low-literate individuals' eye gaze was not related to phonological overlap, but was instead strongly influenced by semantic relationships between items. Our present study tests the hypothesis that this behavior is an emergent property of an increased ability to extract phonological structure from the speech signal, as in the case of high-literates, with low-literates more reliant on more coarse grained structure. This hypothesis was tested using a neural network model, that integrates linguistic information extracted from the speech signal with visual and semantic information within a central resource. We demonstrate that contrasts in fixation behavior similar to those observed between high and low literates emerge when models are trained on speech signals of contrasting granularity.
\end{abstract}

Keywords: The Visual World Paradigm, Connectionist Modeling, Visual Attention, Literacy.

\section*{Introduction}

Eye-tracking studies in which participants are presented simultaneously with spoken language and visual input (i.e. the visual world paradigm, Tanenhaus et al., 1995) have shown that information retrieved via both modalities is mapped at multiple levels of representation. Allopenna et al. (1998), for instance, presented participants with spoken words such as beaker and objects whose names contained word-initial or word-final overlapping phonological information (e.g., beetle, speaker) together with phonologically unrelated objects (e.g., carriage). They found that eye-movements were more likely to be directed to the phonologically related objects than to unrelated objects, indicating that during speech processing, phonologically related representations were co-activated and mapped onto phonological representations retrieved from viewing the copresent visual objects (see Huettig \& McQueen, 2007, for further discussion). Related paradigms have demonstrated that semantic competitors are also co-activated during
listening to speech and attract increased overt attention (Yee \& Sedivy, 2006; Huettig \& Altmann, 2005)

These types of studies leave open one important question: What particular aspects of these representations affect participants' performance? Computational models have been proposed to reproduce the individual phonological and semantic effects on word processing. Allopenna et al. (1998), demonstrated that fixation probabilities during spoken word processing can be predicted by lexical activations in the TRACE model of spoken word recognition. Mayberry, Crocker and Knoeferle (2009) and Kukona and Tabor (2011) extended this work to predict fixation behavior during sentence processing from the integration of visual and linguistic information. Until recently, such models that simulate the interaction between visual and linguistic information did so with representations that were unable to capture fine-grained semantic, phonological or visual feature relationships and were therefore limited in their ability to examine effects of multimodal interactions in language processing. A recent model by Smith, Monaghan and Huettig (in press) based on the hub-and-spoke models of semantic processing which integrates visual, phonological and functional information within a central resource, replicated the intricate time course dynamics of eye fixation behavior reported in Huettig and McQueen (2007). The model highlights the role of differences in the computational properties of each modality's representational structure, demonstrating that such differences are sufficient to produce behavior consistent with multimodal effects reported in the Visual World Paradigm.

The question of how differential representational qualities of phonological and semantic properties affect word processing can also be approached by studying individual differences. Specifically studying participant populations that differ in the form of representation of each modality that they bring to the task. People with different levels of literacy are a critically important population in this regard. There is a well-established link between fidelity of phonological representations of words and development of
literacy (Hulme et al., 2012). Participants who are literate perform better at phonological segmentation or phoneme awareness tasks (Bowey, 2005), and there have been proposals both that literacy causes such improvements in phonological processing (Castles \& Coltheart, 2004; Morais, Cary, Alegria, \& Bertelson, 1979), as well as converse views that effective phonological processing results in improved reading (Muter, Hulme, Snowling, \& Stevenson, 2004). An influential processing model in this literature is that experience of written forms of words results in a change in the granularity of the phonological processing of a word (Ziegler \& Goswami, 2005), such that exposure to written words results in greater awareness of the individual phonemes of words, and without such exposure, listeners are more likely to process the sound of a word without a componential, phonological decoding.

In contrast, effects of literacy on semantic processing have been shown to be minimal and appear to be only quantitatively rather than qualitatively different (Da Silva et al., 2004; Reis \& Castro-Caldas, 1997). Thus, literacy appears to affect lexical processing in a modality-specific manner.

In a recent study, Huettig, Singh and Mishra (2011) compared phonological and semantic competitor effects for Indian participants who had high and low levels of literacy due to poverty or other socioeconomic factors (but no known neurological or cognitive deficits), enabling a direct test of the extent to which the granularity of the phonological form of a word affects performance. In their study (Experiment 1), participants viewed a scene comprising objects representing a phonological onset competitor, a semantic competitor, and two unrelated distractors, and heard the target word spoken in a sentence context. They found that participants with low levels of literacy demonstrated no effects of phonological competitors, but substantial effects of semantic competitors when hearing words. In contrast, the participants with high levels of literacy were similar to the participants in a similar study with Dutch high literates (Huettig and McQueen, 2007) - demonstrating early looks towards objects named by phonological competitors and later looks toward semantic competitors.

We note that looks to the semantic competitors in the Huettig et al. (2011) study were reduced for the low literacy group, which is consistent with accounts of a general processing deficit (cf. Salthouse, 1996), and we return to this issue in the Discussion section.

We adapted our previous multi-modal model of fixation behavior in the visual world paradigm (Smith, Monaghan, \& Huettig, in press) to test the explanatory adequacy of the hypothesis regarding granularity of phonological processing relating to different levels of literacy. We simulated the conditions of the experimental study by presenting visual object representations of phonological and semantic competitors, and two unrelated words and tracking the model's fixation of each of these objects as presentation of a target word unfolded. We adjusted the level of granularity
of the auditory presentation of the word to the model, predicting that a segmented phonological representation would result in early phonological competitor effects, but that less individuated phonological representations, consistent with accounts of phoneme awareness impairment in low-literacy groups, would result in reduced, or absent phonological effects. We also predicted that, consistent with the behavioural data, the later semantic competitor effects would be observed for the model regardless of the granularity of the auditory input to the model.
In order to isolate the effect of the granularity of auditory processing of the spoken word, we controlled for the overall similarity between words in terms of their auditory form, but varied whether the similarity was compositional and at the phoneme level within the model, whether it was sublexical but not at the phonological level, or whether it was not sublexical and represented only at the word level.

\section*{Method}

\section*{Model}

The models described in this paper are based on the model of language mediated eye-gaze presented in Smith, Monaghan and Huettig (in press). The general architecture of the model is shown in Figure 1.


Figure 1: Network Architecture.
Architecture The network consists of four modalityspecific layers which were fully connected to a central resource consisting of 400 units (see Figure 1). The model implements a hub-and-spokes model of multimodal integration, with input visual, auditory and semantic information about words, and output behavior of an "eye" layer which indicates the direction of the attentional focus of the model as a consequence of the combination of the modal inputs.
The vision layer ( 80 units) simulated the extraction of visual information from the surrounding environment, providing visual input to the system. It was divided into four slots, each defined by 20 processing units. Each slot corresponded to the visual information available at each of four possible locations within the visual field. The vision
layer was fully connected in a forward direction to the integrative layer.

Similarly the auditory layer provided input from the auditory modality, simulating the extraction of spoken information from the speech signal over time. The auditory layer was also fully connected to the central integrative layer in a forward direction.

The semantic layer consisted of 160 units, allowing the network to represent semantic features associated with a given object or spoken word. The semantic layer was fully connected to the integrative layer with activation flowing both from integrative units to semantic units and also back from semantic to integrative units.

The eye layer, to reflect the viewing behavior of the system, was also fully connected in both a forward and back direction to the central integrative layer. It consisted of four units, a unit for each location in the visual field represented in the vision layer. Activation of an eye unit was taken as representing the probability of fixating the location in the visual field associated with the given eye unit.

Representations An artificial corpus consisting of visual, auditory, and semantic representations for 200 items was constructed to train and test the network on multiple crossmodal tasks mapping between each of the modalities. A fundamentalist approach (Plaut, 2002) was taken in the construction of representations to ensure all aspects of the representations were controlled within the simulations.

Visual representations of named objects were implemented as 20 unit binary vectors, with each unit representing the presence or absence of a given visual feature for the object. Each object had approximately 10 units activated, which were selected at random, and balanced for their distribution across the set of all 200 items.

For the semantic representations, each item was represented in terms of 8 units active from a set of 160 semantic features, such that the overall set of semantic representations were fairly sparse, simulating semantically
distinct words. Semantically similar pairs of words each shared 4 of the 8 active units representing each item.

To simulate different grain-sizes of speech representation, three forms of auditory input were constructed, but with the overall similarity between representations controlled.

For the fine grained auditory processing, representing phonological segmentation of the spoken word by the listener, words were encoded as six phonemes, with phonemes implemented as sets of 10 units, from which five units were active. All words within the corpus were composed of phonemes taken from an inventory of 20 possible phonemes. To present the word an additional phoneme from the target word sequence was presented to the auditory layer at each time step.

To simulate sublexical representations of a coarser grain size (moderate), two 30 unit binary feature vectors were created for each word from which 15 units were active. Coarse grained representations were formed by 60 unit binary feature vectors of which 30 units were active.

Table 1: Mean cosine similarity of speech signal representations calculated between targets and distractors.
\begin{tabular}{llccc}
\hline Grain Size & Distractor & \multicolumn{3}{c}{ Signal Overlap \((\bar{x}, \sigma)\)} \\
& Type & Onset & Rhyme & Word \\
\hline Fine & Competitor & \(.18(.07)\) & \(.50(.13)\) & \(.34(.07)\) \\
& Unrelated & \(.50(.12)\) & \(.50(.12)\) & \(.50(.09)\) \\
Moderate & Competitor & \(.17(.08)\) & \(.50(.11)\) & \(.34(.07)\) \\
& Unrelated & \(.51(.10)\) & \(.51(.10)\) & \(.50(.07)\) \\
Coarse & Competitor & \(.34(.10)\) & \(.34(.10)\) & \(.34(.07)\) \\
& Unrelated & \(.51(.10)\) & \(.51(.10)\) & \(.50(.07)\) \\
\hline
\end{tabular}

Visual, semantic and auditory competitors were also embedded within the corpora for 40 target items. For visual competitors 10 of 20 visual features were shared with target items with \(\mathrm{p}=1\), with the remaining features shared with p \(=0.5\). Semantic competitors shared 4 of 8 semantic features with target representations, while unrelated items shared a maximum of 1 semantic property with any other item.

Table 2: Temporal organization of events in training. Describes input and target representations provided in training trials.
\begin{tabular}{|c|c|c|c|c|c|c|c|c|}
\hline \multirow[t]{2}{*}{Task} & \multicolumn{2}{|l|}{Visual Input} & \multicolumn{2}{|l|}{Auditory Input} & \multicolumn{2}{|l|}{Semantic Layer} & \multicolumn{2}{|l|}{Eye Layer} \\
\hline & Activity & ts & Activity & ts & Activity & ts & Activity & ts \\
\hline Form to Semantics & 4 items selected at random from corpus & 0-14 & Time invariant noise provided as input & 0-14 & Target: Target's Semantic representation & 3-14 & Input: Only location of target active & 0-14 \\
\hline Speech to Semantics & Time invariant noise provided as input & 0-14 & Phonology of target as staggered input & 0-14 & Target: Target's Semantic representation & 5-14 & No constraints on activation & - \\
\hline Speech to Location & 4 items selected at random from corpus & 0-14 & Phonology of target as staggered input & 0-14 & No constraints on activation & - & Target: Only location of target active & 5-14 \\
\hline Semantics to Location & 4 items selected at random from corpus & 0-14 & Time invariant noise provided as input & 0-14 & Input: Target's Semantic representation & 0-14 & Target: Only location of target active & 2-14 \\
\hline
\end{tabular}

Fine grained spoken word competitors were defined by an overlap in the initial two components of their speech signal. For the unrelated items, we ensured that this set of words did not share more than the first component of the word and that no items shared their initial nor final three components. For moderate grain size representations \(2 / 3\) of the initial 30 features of a competitor were shared with a target with \(\mathrm{p}=\) 1 , with remaining features overlapping with \(\mathrm{p}=0.5\). Controls ensured all initial and final moderate grain vectors were unique. For coarse grain competitors \(1 / 3\) of all features were shared with the corresponding target with \(\mathrm{p}=1\), with remaining features overlapping with \(\mathrm{p}=0.5\). Defining competitors in this way lead to the contrasts in levels of similarity between representations across corpora as described in Table 1. Although the level of similarity between competitor-target and unrelated distractor-target is consistent across corpora at the word level, the distribution of overlap varies between implementations as a function of grain size.

Model Training The model was trained on four tasks (see table 2). Tasks were designed to simulate those performed by participants prior to testing through which associations between representations are acquired. The tasks were to map from visual representation to semantic representation, from auditory representation to the semantic representation, to activate the eye unit corresponding to the location of the item whose semantic representation is presented, and to activate the location of the item whose auditory representation is presented. Tasks were presented on a pseudo random basis with the task of mapping speech to location occurring four times less than other tasks. Items were selected from the corpus and assigned roles (target or distractor) and locations randomly. Initial connection weights were randomized and adjusted during training using recurrent back-propagation (learning rate \(=0.05\) ). Training was terminated after 850000 trials.

\section*{Results}

In the following sections we report the performance of three categories of model 1) Fine, models trained and tested on representations that simulate extraction of fine grained structure within the speech signal; 2) Moderate, models trained and tested on representations that simulate extraction of moderate structure within the speech signal; 3) Coarse, models trained and tested on representations that simulate coarse grained structure within the speech signal. The following results represent performance averaged across five instantiations of each model. For each instantiation a new corpus was constructed on which it was then trained and tested each initialized with a different random seed.

\section*{Pre-Test}

Once trained all models were tested on their ability to complete each of the four training tasks for all items in the training corpus presented in all possible locations within the
visual field. All three categories of model displayed similar levels of performance across all four tasks. In mapping from speech to semantics, activation of the semantic layer was most similar (cosine similarity) to the target item for \(100 \%\) of items for all models. When mapping from visual to semantic representations, activation in the semantic layer was most similar (cosine similarity) to that of the target for \(98 \%\) of items in the case of coarse and fine grained models and \(97 \%\) of items in the case of moderate models. When challenged to select the location of a target when presented with its corresponding auditory representation, the correct location was activated in both the coarse and fine models for \(96 \%\) of items and \(98 \%\) of items for moderate models. All models displayed equal performance when locating a target indicated by the presence of its semantic representation, selecting the correct location for \(99 \%\) of items.

\section*{Simulating Huettig, Singh and Mishra (2011)}

The following conditions remained consistent across all simulations. Visual input was provided at time step (ts) 0 and remained until the end of each test trial (ts 29). We report the activation of each unit within the eye layer as a proportion of the total activation of all units within this layer. This proportion is taken to represent the probability of fixating p (fix), the associated location within the visual field. Word onset occurred at ts 5 , with an additional component of the speech signal presented at each time step until the entire speech signal had unfolded (ts 10). Auditory input then remains fixed until the end of the test trial.

To simulate the conditions of Huettig, Singh and Mishra (2011) experiment 1 , input to the models visual layer consisted of the visual representations of the target's auditory competitor and semantic competitor along with two unrelated distractors. The target word's auditory representation was presented as a staggered input to the auditory layer from ts 5 . All models (fine, moderate and coarse) were tested on all 40 test sets embedded within the corpus (target, auditory competitor, semantic competitor and two unrelated distractors) in all 24 possible combinations of item and location. Figure 2 displays the change in p (fixation) from ts 0 for each category of item (Aud \(=\) auditory competitor, Sem = semantic competitor, Control = unrelated distractor), averaged across all test trials.

For analysis ratios were calculated between the proportion of fixations to a given competitor and the sum of the proportion of fixations to both the competitor and distractors (see Huettig \& McQueen, 2007). A value of 0.5 would indicate both items were fixated equally, a value greater than 0.5 would indicate increased fixation of the competitor and lower than 0.5 increased fixation of the distractor. Mean ratios were calculated across items and instantiations.

We conducted a 2-way ANOVA on the auditory competitor-distractor ratios with model as between-subject factor and time as within-subject factor for three theoretically-motivated time regions (preview, early and late). No significant differences were predicted during the
preview period which refers to the time between display onset (ts 0 ) until the first time step in which auditory information relating to the target word is able to influence output layers (ts 7). The remainder of test trials was divided equally into two time bins, an early (ts 8-18) and a late (ts 19-29) period as previous research had shown that auditory effects would occur (if at all) during the early but not the late period.


Figure 2: Change in fixation proportions for simulations of Huettig, Singh and Mishra (2011) Experiment 1.

There was a significant main effect of time, \(\mathrm{F}(2,234)=\) \(38.155, \mathrm{p}<.001\), eta- \(2=.246\), with auditory competitordistractor ratios differing between preview and early time windows, \(\mathrm{F}(1,238)=39.387, \mathrm{p}<.001\), and preview and late time windows, \(F(1,238)=29.202\), although there was no difference between early and late time windows. There was also a significant main effect of model, \(F(2,117)=4.467\), \(p\) \(=.014\), eta- \(2=.071\), with the fine and medium models resulting in significantly more fixations to the phonological distractor than the coarse model, means \(=.544, .544\), and .508, respectively. Critically, there was a significant interaction between model and time, \(\mathrm{F}(4,234)=3.582, \mathrm{p}=\) .023 , eta- \(2=.058\). The quadratic contrast effect for time was significant in the interaction, \(\mathrm{F}(2,117)=5.074, \mathrm{p}=.008\), eta-2 \(=.080\), indicating that the models were more differentiated at the early time steps than during the preview or later time steps. Models did not differ significantly within the preview period. There was however a significant difference between fine and coarse models, \(\mathrm{F}(1,78)=\) 14.373, p < .001, and coarse and moderate models, \(\mathrm{F}(1,78)\) \(=9.544, \mathrm{p}=.003\), in the early time window. The coarse model also differed from the fine \(F(1,78)=4.286, \mathrm{p}=.042\), and moderate model \(F(1,78)=7.153, p=0.009\), in the later time window. No difference was found between fine and moderate models in any time period.

A 2-way ANOVA was also conducted on semantic competitor-distractor ratios with model as between subject
factor and time as within-subject factor. Again we observed a main effect of time, \(F(2,234)=230.642, p<.001\), eta-2 \(=\) .663, semantic competitor distractor ratios differed significantly between preview and early, \(\mathrm{F}(1,238)=59.607\), \(\mathrm{p}<0.001\) preview and late, \(\mathrm{F}(1,238)=243.403\), \(\mathrm{p}<.001\) and early and late time windows, \(\mathrm{F}(1,238)=80.562\), \(\mathrm{p}<\) .001. There was no main effect of model nor was there a significant interaction between model and time.

We then compared whether competitor-distractor ratios differed from chance (0.5) for each time step using one sample t-tests. The probability of fixating the auditory competitor first differed ( \(\mathrm{p}<0.001\) ) from that of the distractor from time step 11 in both fine and moderate models and continued to differ for all subsequent time points. In contrast fixation of the auditory competitor by the coarse model only differed marginally ( \(p<0.1\) ) from the distractor item in time steps \(13-17\). Fixation of semantic competitors first differed significantly ( \(\mathrm{p}<0.05\) ) from distractor levels at ts 12 and continued to differ for all remaining ts, this was the case for all models.

\section*{Discussion}

Our study aimed to examine the explanatory adequacy of the hypothesis that increased granularity of phonological processing, can account for the differences in fixation behavior between low and high literates observed in Huettig, Singh and Mishra (2011) Experiment 1. Our simulations demonstrate that increasing the grain size at which speech is processed can lead to a modulation of phonological effects. A model trained on representations of speech at the word level displayed only a marginal increase in fixation towards competitor items that overlapped in an auditory dimension, whereas models trained on componential, phoneme level representations or moderate grain size, sublexical components did display a significant increase in fixation of auditory competitors. Between model comparisons further demonstrated that the coarse grained implementation differed significantly from both fine and moderate grain models post word onset.

Interestingly, such comparisons did not display a graded effect of grain size, with fine and moderate models not differing in fixation proportions towards auditory competitors at any stage within test trials. There are two possible reasons for our failure to observe a graded effect. On the one hand, qualitative features of the data hint that given a larger corpus and hence test set such effects may be observable. One sample, left tailed t-tests comparing the ratio between the proportion of fixations towards auditory competitors in the moderate model and the sum of the proportion of fixations to the auditory competitor in the moderate and fine model indicate a significant difference at ts \(13-16\), ( \(\mathrm{p}<0.05\) ), this difference can be observed in Figure 2.

On the other hand, it is conceivable that illiterates and low literates rely on very coarse grained structure within the speech signal. Although previous studies have shown that illiterates and low literates perform slightly better on
syllable awareness than on phonemic awareness tasks, they still tend to perform far worse than proficient readers. This may suggest that achieving even moderate granularity of phonological processing may not be rapid. The results of our simulations could be interpreted as reflecting that when a moderate grain size of phonological processing is achieved performance improves rapidly and becomes similar to fine-grained models.

Our results also demonstrate that increased granularity does not necessarily lead to a decrease in semantic effects as observed in Huettig, Singh and Mishra (2011). Although our simulations indicate phonological effects could be modulated by an increase in the grain size, an additional mechanism is needed to create the distinction between semantic effects observed across populations. A reduction in general processing speed in the illiterate population has been offered to account for differences in performance on a large variety of cognitive tasks (Salthouse, 1996). This potentially offers an explanation for a reduction in both auditory and semantic competitor effects. A general processing deficit for low literates, could be implemented by adding noise across sematic representations, representing a reduction in the fidelity of such representations. Adding noise in this manner would result in a general reduction of semantic competitor effects, however it is less clear whether the introduction of noise could also lead to the elimination rather than a general reduction of the phonological effect as observed in illiterate performance. As the authors acknowledge, behavior observed in Huettig et al (2011) suggests that the qualitative changes to the phonological competitor effects and the semantic competitor effects are distinct. Teasing apart the factors underlying observed differences in behaviour between populations is far from trivial, however explicit implementations such as the one described in this paper provide a means of testing the plausibility of proposed explanations.

\section*{References}

Allopenna, P. D., Magnuson, J. S., and Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: evidence for continuous mapping models. Journal of Memory and Lanugage, 38, 419-439.
Bowey, J. A. (2005). Predicting individual differences in learning to read. In M. J. Snowling \& C. Hulme (Ed.), The science of reading: A handbook (pp. 153-172). Oxford, England: Blackwell.
Castles, A., \& Coltheart, M. (2004). Is there a causal link from phonological awareness to success in learning to read? Cognition, 91, 77-111.
Da Silva, C., Petersson, K. M., Faísca, L., Ingvar, M., and Reis, A. (2004). The effects of literacy and education on the quantitative and qualitative aspects of semantic verbal fluency. Journal of Clinical Experimental Neuropsychology. 26, 266-277.
Huettig, F., \& Altmann, G. T. (2005). Word meaning and the control of eye fixation: Semantic competitor effects
and the visual world paradigm. Cognition, 96(1), B23B32.
Huettig, F., and McQueen, J. M. (2007). The tug of war between phonological, semantic, and shape information in language-mediated visual search. Journal of Memory and Language. 54, 460-482.
Huettig, F., Singh, N., \& Mishra, R. K. (2011). Languagemediated visual orienting behavior in low and high literates. Frontiers in Psychology, 2, 285.
Hulme, C., Bowyer-Crane, C., Carroll, J.M., Duff, F.J., \& Snowling, M.J. (2012). The causal role of phoneme awareness and letter-sound knowledge in learning to read: Combining intervention studies with mediation analyses. Psychological Science, 23, 572-577.
Kukona, A., \& Tabor, W. (2011). Impulse processing: A dynamical systems model of incremental eye movements in the visual world paradigm. Cognitive Science, 35, 1009-1051.
Mayberry, M., Crocker, M. W., \& Knoeferle, P. (2009). Learning to attend: A connectionist model of the coordinated interplay of utterance, visual context, and world knowledge. Cognitive Science, 33, 449-496.
Morais, J., Cary, L., Alegria, J., \& Bertelson, P. (1979). Does aware- ness of speech as a sequence of phones arise spontaneously? Cognition, 7, 323-331.
Muter, V., Hulme, C., Snowling, M. J., \& Stevenson, J. (2004). Phonemes, rimes, vocabulary and grammatical skills as foundations of early reading development: Evidence from a longitudinal study. Developmental Psychology, 40, 665-681.
Plaut, D. C. (2002). Graded modality-specific specialization in semantics: A computational account of optic aphasia. Cognitive Neuropsychology, 19, 603-639.
Reis, A., and Castro-Caldas, A. (1997). Illiteracy: a bias for cognitive development. Journal of the International Neuropsychological Society. 3, 444-450.
Salthouse, T. A. (1996). The processing speed theory of adult age differences in cognition. Psychological Review 103, 403-428.
Smith, A., Monaghan, P., \& Huettig, F. (in press). Multimodal interaction in a model of visual world phenomena. In Computational Models of Cognitive Processes: Proceedings of the 13th Neural Computation and Psychology Workshop. World Scientific Publishing Company.
Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., \& Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. Science, 268, 1632-1634.
Yee, E., \& Sedivy, J. C. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. Journal of Experimental Psychology: Learning, Memory, and Cognition, 32, 1-14.
Ziegler, J. C., \& Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. Psychological Bulletin, 131, 3-29.

\title{
Consistent physics underlying ballistic motion prediction
}

\author{
Kevin A Smith (k2smith@ucsd.edu), \({ }^{1}\) Peter Battaglia (pbatt@mit.edu), \({ }^{2}\) Edward Vul (evul@ucsd.edu) \({ }^{1}\) \\ 1. University of California, San Diego, Department of Psychology, La Jolla, CA 92093 \\ 2. MIT, Department of Brain and Cognitive Sciences, Cambridge, MA 02139
}

\begin{abstract}
Research into human models of intuitive physics typically falls into one of two camps, either claiming that intuitive physics is biased and not representative of real physics, or claiming that it consists of a collection of veridical physical laws. Here we investigate the causes of this tension, suggesting that prediction is based on real physics, but explanation is susceptible to biases. We gave participants three tasks based on the same physical principles: two prediction tasks and one task that required drawing the future path of motion. We found distinct biases in all three tasks; however, the two prediction tasks could be explained by consistent application of real physical principles under uncertainty, while the drawing task produced many more idiosyncratic biases. This suggests that different tests of intuitive physics are capturing different types of knowledge about the world.
\end{abstract}

Keywords: intuitive physics; uncertainty; ballistic motion prediction

\section*{Introduction}

Classic studies have suggested that many people base their physical intuitions on incorrect and inconsistent physical theories (Anzai \& Yokoyama, 1984; McCloskey, Caramazza, \& Green, 1980). Others have reported that people are biased by surface-level differences between tasks (Kaiser, Jonides, \& Alexander, 1986), and that their inferences about simple physical situations rely on shallow heuristics and are frequently mistaken (Proffitt \& Gilden, 1989; Todd \& Warren, 1982). However over the past few years, a number of researchers have explained human physical predictions using quantitative cognitive models that assume people have an accurate and consistent understanding of the laws of physics that they apply flexibly across tasks (Hamrick, Battaglia, \& Tenenbaum, 2011; Sanborn, Mansinghka, \& Griffiths, 2013; Smith \& Vul, 2013; Téglás et al., 2011).

We suggest that a core difference between the above studies is the task given to participants. Some have asked participants to make a single judgment about the future state of the world, for instance, the direction a tower of blocks will fall (Hamrick, et al., 2011) or where a ball will cross a line (Smith \& Vul, 2013). In contrast, classic studies tap into explicit explanations of physics, through verbal problems (Anzai \& Yokoyama, 1984) or line drawings of motion (McCloskey, et al., 1980). Here we argue that people can apply correct physical principles consistently to simulate the world forward; however, explicit explanations of how the world will unfold draw upon an idiosyncratic set of background knowledge.

We assessed participants' understanding of the movement of balls after they had fallen off of pendulums in three separate tasks: predicting where a ball would land if cut from a pendulum, deciding when to cut a pendulum string such that the ball would fall into a fixed bucket, and drawing the path of the ball after the string is cut. We picked these tasks because there is evidence that people understand the motion of pendulums (Pittenger, 1985, 1990) and can predict the motion of projectiles under gravity (Saxberg, 1987), both of which must be combined to determine the ultimate trajectory of the balls. But there is also evidence that people show systematic errors when asked to explicitly draw the path of the ball (Caramazza, McCloskey, \& Green, 1981), and that these errors are attenuated with kinematic information (Kaiser, Proffitt, Whelan, \& Hecht, 1992).

The same physical principles apply to each of these tasks, and so in the present experiment we investigated whether the tasks that require implicit prediction (catching the ball and cutting the string) can be explained by veridical physical principles. We find that subjects' performance on the catching and cutting tasks differs between the tasks, but in the tasks that involved perceptually guided movements the differences can be reconciled by considering a single, valid model of physics that incorporates the different sources of perceptual and motor uncertainty from each task. Conversely, the sketches based on explicit conceptualization were inconsistent and idiosyncratic.

\section*{Experiment}

\section*{Methods}

Fifty-seven UC San Diego undergraduates (with normal or corrected vision) participated in this experiment for course credit. All were treated in accordance with UCSD's IRB protocols.

\section*{Procedure}

Participants viewed a computer monitor from a distance of approximately 60 cm , which initially depicted a ball swinging from a string, consistent with pendulum motion. At some point in time the string would be cut and the ball would be released, thus entering ballistic motion. Beneath the pendulum there was always a bucket, and in every trial the participant's goal was to cause the ball to drop into the bucket after being released. How they were allowed to interact with the scene differed between two tasks, which were organized into blocks that were randomized across participants. With the exception of one initial practice trial per task that familiarized participants with the task, the path of the falling ball was occluded in order to prevent
participants from learning a simple relationship between the ball's release position and its landing position. At the end of each trial, participants were given binary feedback that indicated whether or not the ball successfully landed in the bucket. After the two tasks on the computer, participants were asked to draw the ball's motion in a diagram task.

Catching task. Participants were instructed to adjust the bucket's horizontal position using the mouse so that the ball would land in the bucket after being released. The release time was pre-determined and varied across trials. To relieve time pressure placed on participants, at the moment the string was cut, all ball and string movement was paused. Once the participant chose a bucket position, they could unpause the motion by clicking the mouse. The center of the bucket was recorded as the participant's judgment about where the ball would land.

Cutting task. The bucket was held fixed at a predetermined position and participants were instructed to cut the pendulum string by clicking the mouse at a time that would cause the ball to drop into the bucket. The time at which the string was cut was recorded for each trial.


Figure 1: Diagram of the two tasks: catching on top, cutting on the bottom. (A) The pendulum swings freely to start; this ends at a predetermined time (catching) or when the participant clicks the mouse (cutting). (B) An occluder is placed over the string. In the catching task, the action is paused until participants click the mouse, during which time they can move the bucket. In the cutting task, there was no pause, but the falling motion of the ball was occluded. (C)

Participants are given feedback on success or failure.
Trials. For each task, participants repeated 48 distinct trials five times each. Trials were matched across tasks such that where the ball landed in a catching trial was the bucket position in the matched cutting trial. In the catching task, there were 16 distinct release times, crossed with three vertical distances between the nadir of the pendulum and position of the bucket - either 20,35 or \(50 \%\) of the total screen height. No participant indicated they were aware that
the trials were repeated or matched across task in an informal post-experiment survey. \({ }^{1}\)

Simulating pendulum motion. Both tasks and all trials used the same pendulum. This pendulum had a length of half of the screen, and reached a maximum angle of \(35^{\circ}\) from vertical of the nadir. The period of the pendulum was 2.46s. The string was assumed to be massless, and therefore the position of the pendulum at any time could be calculated according to the laws of physics. \({ }^{2}\)

Both the pendulum motion and the falling ball obeyed Newtonian mechanics as if the pendulum was positioned at a depth of 6 m from the participants. This value was selected through pilot tests to conform to participants' general expectations about the natural period of the pendulum.

Diagram task. After participants completed both tasks, they were given diagrams of pendulums and asked to draw the path of the ball if the string was cut at four positions indicated in those diagrams (a replication of Caramazza, et al., 1981). One participant did not perform this task due to a logistical error.


Figure 2: The four problems in the diagram task.
Participants were asked to draw the expected path of the ball if the pendulum string were cut at each of the four points.

\section*{Results}

Accuracy in the catching and cutting tasks was measured as the proportion of trials in which the ball successfully landed in the bucket. Participants' mean accuracies were \(30.7 \%\) (s.d. \(14.1 \%\) ) on the catching task, and \(47.4 \%\) (s.d. \(15.6 \%\) ) on the cutting task. Participants' individual accuracies were (Pearson) correlated across tasks, \(r=0.68\). There was no evidence that participants improved over trials on the cutting task ( \(z=1.23, p=0.22\) ), but they did improve on the catching task ( \(\mathrm{z}=3.04, p=0.0024\) ), from \(28.8 \%\) accuracy on the first half to \(32.8 \%\) on the second half.

The remaining analyses quantified participants' performance as the displacement between the ball's landing position and the bucket's position; in the catching task the bucket position was under participants' control and the landing position was under experimental control, and vice versa for the cutting task. We aggregated performance by

\footnotetext{
\({ }^{1}\) One participant noted that they solved trials by "remembering where the ball should go" but it was not clear whether this was memory for the trials or prior knowledge of pendulum motion.
\({ }^{2}\) For computational reasons, this was calculated using the small angle approximation to pendulum motion, which should be correct to within \(2.4 \%\) of actual pendulum timing.
}
trial across participants in each task to determine how trial factors influenced participants' decisions.

Catching task. Participants' mean bucket positions were correlated with the ball's actual landing positions ( \(r=0.95\), SumSq \(=880^{*} 10^{3}\) ), and were highly consistent with each other (split-half correlation: \(r=0.993\) ). Participants also demonstrated a systematic bias: on average their judgments were slightly shifted away from the actual landing position, toward the center of the pendulum (see Fig. 3). The consistency across participants suggests that the position bias is shared, capturing a commonality in physical models.


Figure 3: Catching task. Actual landing positions (x-axis) versus participants' mean bucket positions ( y -axis) for each trial (individual trials, error bars are \(95 \%\) CIs).

Cutting task. We calculated the projected landing positions of the ball as a function of each release time chosen by participants, per trial. Participants' mean landing positions were highly correlated with the actual bucket positions ( \(r=0.98\), SumSq \(=187 * 10^{3}\) ), and were again highly consistent with each other (split-half correlation: \(r=0.998\) ). Participants also demonstrated a distinct bias, which differed from that in the catching task: when the bucket was near the horizontal position of the pendulum's nadir, participants' mean landing positions were shifted away from it, but when the bucket was far from the nadir, their mean landing positions were shifted toward it (note the sigmoid curvature in Fig. 4). This high inter-participant correlation again suggests a common bias across people.


Figure 4: Cutting task. Actual bucket positions (x-axis) versus mean ball landing positions (y-axis) for each trial (individual trials, error bars are \(95 \%\) CIs).

Comparison. Both tasks required using the same physical principles to determine where the bucket should be placed or when the rope should be cut, yet showed divergent biases. Moreover, the correlation between the mean bucket position and mean landing position for matched trials was high ( \(r=0.93\) ), but this demonstrates only that participants were in general accurate at this task - the inter-task correlation was less than each task's correlation with the ideal response, suggesting that the sources of deviation from the ideal response are distinct.

Diagram task. Two research assistants naïve to the purpose of this experiment sorted participants' diagram trajectories into one of eight types (see Figure 5). Inter-rater reliability was high (Cohen's kappa \(=0.826\) ) - the raters agreed for 47 of the 56 of the participants; where they disagreed, the first author acted as a tie-breaker. Twenty-one percent of the participants' figures were idiosyncratic and could not be categorized. Only 4 (7\%) of participants drew the correct path for all diagrams.


Figure 5: Diagram patterns drawn by more than one participant. Excludes 12 participants who drew idiosyncratic paths. The top pattern represents correct physics.

We reviewed subjects' beliefs about trajectories under gravity: whether they demonstrated that balls would fall in a curved pattern: only \(18 \%\) of our participants did (less than the \(55 \%\) reported by Caramazza, et al., 1981). If participants were learning principles about pendulums from the catching or cutting task, we would have expected a higher proportion of curved paths.

Thus participants display high inter-subject reliability on the catching and cutting tasks (despite large differences between the two) but when explicitly drawing pendulum
trajectories they show much less agreement and consistency with any kind of physical or non-physical principles. We believe this discrepancy arises because the diagram task taps into idiosyncratic, strategic explanations of physics, but the cutting and catching task behaviors arise from a single consistent application of physical principles under different task demands. We designed a model to test the latter claim.

\section*{Physics-based model observer}

We designed a model observer that used a single system of physical mechanics rules to predict participants' behavior on both the catching and cutting tasks. These model predictions used real-world physics, just as was used in the experiment to determine the trajectory of the ball both on and off of the pendulum string. The model adapted to each task by adjusting how its physical predictions were applied to the judgment. In the catching task it computed the expected landing position of the ball and selected that as its bucket position, but biased its estimates of the ball's pre-release velocity toward a slower speed based on "misremembering" the velocity through a pause. In the cutting task it computed which release time would cause the ball to land in the bucket and selected that as its judgment, but this timing was subject to errors that reflected realistic constraints on people's timing precision.

\section*{Catching task}

Description. Because the ball was motionless while participants placed the bucket, participants were required to remember the velocity of the ball and form their judgment based on that memory. This could introduce biases that would cause participants to recall the velocity as slightly different than it had actually been before the pause (Brouwer \& Knill, 2009), especially favoring slower speeds (Stocker \& Simoncelli, 2006; Weiss, Simoncelli, \& Adelson, 2002). This bias was treated as a single parameter ( \(\mathrm{v}_{\mathrm{adj}}\) ) that determined the proportion of the original velocity the ball would have upon being released. This proportion was constant across all trials.

Based on this (mis)remembered velocity, the model calculated the expected landing position of the ball when it would hit the paddle, and assumed all deviation from that position was Gaussian noise. This placement noise could arise from noise in either the motor system during placement, uncertainty in estimation of the velocity of the ball, or simulation uncertainty that accumulates symmetrically around the position over time (e.g., Smith \& Vul, 2013). \({ }^{3}\)

Model fit. The model explained participants' average bucket positions well ( \(r=0.994\), \(\operatorname{SumSq}=41 * 10^{3}\), see Fig. 6 ), and accounted for participants' center-shift bias. The model predicted participants' responses as well as

\footnotetext{
\({ }^{3}\) Simulations indicated that noise in the initial velocity (speed and direction) would give rise to roughly Gaussian error, suggesting that this is a reasonable assumption.
}
participants predicted each other, which suggests that the model captures nearly all of the systematicity in people's underlying judgments.

The best fitting parameters assumed that participants recalled the ball as having \(51.7 \%\) of its pre-pause velocity magnitude, which caused their judgments of its predicted final horizontal distance to be shifted nearer to the center when it reached the ground. Although this is directionally consistent with our assumption that people remember velocity as slower than it was, the magnitude was larger than expected. Individual errors were predicted to be distributed around that point with a standard deviation equal to \(14.5 \%\) of the screen width.

Although accuracy increased across trials in the catching task, this had relatively little impact on the model parameters (first half \(\mathrm{v}_{\mathrm{adj}}: 47 \%\), second half \(\mathrm{v}_{\mathrm{adj}}\) : \(55 \%\) ). Therefore we do not believe that this pattern of errors was driven by feedback during the task.


Figure 6: Catching task. Model's bucket positions (x-axis) versus participants' mean bucket positions ( y -axis) for each trial (individual points, error bars are \(95 \% \mathrm{CIs}\) ).

Uncertainty The model assumed that the error in the catching task arose from Gaussian noise in the bucket position around the expected location. This implies a constant error in paddle position regardless of where the ball lands. Thus error should be constant across trials.


Figure 7: Catching task. Actual landing position (x-axis) versus participants' bucket positions' SD (y-axis) for each trial (individual points).

As can be seen in Figure 7, there is no evidence for a linear \((F(1,46)=0.27, p=0.61)\) or quadratic \((F(2,45)=1.31\), \(p=0.28\) ) relationship between the landing position of the ball on each trial and the standard deviation of participants' bucket positions on that trial. \({ }^{4}\) This suggests that error does not vary as a function of bucket position, which agrees with our prediction that this is only a combination of motor error and unbiased prediction noise.

\section*{Cutting task}

Description Participants' release time choices were variable, likely due to imprecise visual estimates of the ball's position and velocity as well as noise inherent to fine motor behaviors. As a result, if the participant intended to release the ball at time \(t\), they may have instead released it at time \(t+\varepsilon\). Because the physical dynamics induce a nonlinear relationship between \(\varepsilon\) and the error in landing position, a rational participant should select a time for which the probability of the ball landing in the bucket is highest rather than when it would land closest to the bucket center. If people understand their own timing imprecision (as reported in Hudson, Maloney, \& Landy, 2008), then they should marginalize over \(\varepsilon\) in order to maximize their chance of success. If \(R^{*}\) is the intended release time, \(R\) is the actual release time, and \(t_{\text {err }}\) is the variability in timing, the probability of hitting the bucket given \(R^{*}\) is:
\[
P\left(h i t \mid R^{*}, t_{e r r}\right)=\int P(h i t \mid R) * P\left(R \mid R^{*}, t_{e r r}\right)
\]

Here \(P(h i t \mid R)\) is either 1 or 0 , because hit depends deterministically on \(R\). The distribution of \(R\) given \(R^{*}\), \(P\left(R \mid R^{*}, t_{\text {err }}\right)\), was assumed to be Gaussian distribution with mean and \(\mathrm{SD}, R^{*}\) and \(t_{e r r}\) respectively. The model assumed that people selected \(R^{*}\) such that \(P\left(h i t \mid R^{*}\right)\) was at a local maximum. The cutting task contained an important additional feature: for most trials ( \(58 \%\) ) there were two time spans in the pendulum period during which the string could be cut to get the ball into the bucket - usually one time while the pendulum is swinging left, and once while swinging right. In these cases, there were two locally maximum modes of \(P\left(h i t \mid R^{*}\right)\). Puzzlingly, people did not always choose the optimal (higher probability) mode given the model assumptions, but instead often favored the suboptimal mode. This suboptimality may have been due to participants' desire to accumulate more information by waiting for the later time (Battaglia \& Schrater, 2007; Faisal \& Wolpert, 2009), or minimize trial duration by selecting the earlier time. Since our model did not capture such factors, we simply set the model's choice of modes to match the participants' proportion.

Timing errors were represented by two parameters in this model, describing the bias ( \(\mathrm{t}_{\text {bias }}\) ) and the noise ( \(\mathrm{t}_{\text {err }}\) ). These

\footnotetext{
\({ }^{4}\) We attempted to fit polynomial regressions up to fifth-order to this data but found no significant relationships (all ps \(>0.1\) ).
}
parameters were fit to the observed cut timings, though for consistency, results are presented as the average landing position based on these cuts.

Model fit. The model assumed that people tended to release the ball 38 ms after the optimal time, and the variability in responses had a standard deviation of 165 ms . This timing variability is similar in magnitude to that reported in another task that required physical prediction \((130 \mathrm{~ms}\); Faisal \& Wolpert, 2009). The correlation between people's mean projected landing position given their choice of release time and that of the model was high ( \(r=0.993, \operatorname{SumSq}=87 * 10^{3}\), see Fig. 8).


Figure 8: Cutting task. Model's landing position (x-axis) versus participants' mean projected landing positions ( y axis) for each trial (individual points, error bars are \(95 \%\) CIs).

Uncertainty. The model assumed that the source of error in landing positions was in the cutting time, but a constant error in time does not imply a constant error in landing position: if the ball is released near the apex when moving slowly, a small time error will lead to a small difference in landing position, while if the ball is released at the nadir when moving fastest, the same timing error will lead to a larger difference in landing position.


Figure 9: (Left) Variability in empirical ball landings by where the ball will land. (Right) Model predictions of trial variability in the cutting task versus empirical observations. Each point represents a separate trial.

Unlike the catching task, there is a quadratic relationship between the landing position of the bucket on each trial and
the SD of the ball landing positions on that trial \((F(2,45)=13.8, p<0.001\), see Fig. 9, left). Furthermore, the model predicts this variability. We calculated the SD of landing position that the model expected for each trial and found that it was correlated with participants' projected landing position SD with \(r=0.67\) (see Fig. 9, right), although the model's predicted SD was slightly lower on some trials. This suggests that the physics-based model captures differences in trial variance.

\section*{Discussion}

In this experiment, we found that people show very different behaviors on three tasks that use the same underlying model of physics: predicting the trajectory of a ball on a pendulum after the string has been cut. Two of the tasks required people to make a judgment about the future state of the world: where the ball will land or when to cut the string to control the ball's landing. While people responded in different ways on each of these two tasks, both sets of responses were consistent with veridical physical principles once task uncertainties were accounted for. On the other hand, participants were much more variable on the diagram task: they often drew trajectories that were physically impossible.

These differences imply that the catching and cutting tasks are tapping a different sort of knowledge than the diagram task. Perhaps people can simulate the world forward in a way consistent with Newtonian physics, but the workings of these simulations are opaque, making description difficult and more reliant on conceptual understandings. This would suggest a need for both types of intuitive physics: research into how people make predictions informs how we use physics to plan our actions or make judgments about the world (e.g., Gerstenberg, Goodman, Lagnado, \& Tenenbaum, 2012; Hamrick, et al., 2011), while research into how people describe physical events informs how we form concepts about the workings of the world (e.g., diSessa, 1993).

It has been suggested before that "a person may possess a perceptual appreciation of... natural dynamics... yet be unable to draw upon this knowledge... in a representational context." (Kaiser, Proffitt, \& McCloskey, 1985, p. 539). Here we provide evidence that even when people cannot explain how the world will unfold, their predictions and actions are reflective of a veridical physical model of the world.

\section*{Acknowledgments}

KS and EV were supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of the Interior (DOI) contract D10PC20023. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI, or the U.S. Government.

\section*{References}

Anzai, Y., \& Yokoyama, T. (1984). Internal models in physics problem solving. Cognition and Instruction, 1 (4), 397-450.
Battaglia, P., \& Schrater, P. R. (2007). Humans trade off viewing time and movement duration to improve visuomotor accuracy in a fast reaching task. The Journal of Neuroscience, 27(26), 6984-6994.
Brouwer, A.-M., \& Knill, D. C. (2009). Humans use visual and remembered information about object locations to plan pointing movements. Journal of Vision, 9(1), 1-19.
Caramazza, A., McCloskey, M., \& Green, B. (1981). Naive beliefs in "sophisticated" subjects: Misconceoptions about trajectories of objects. Cognition, 9, 117-123.
diSessa, A. A. (1993). Toward an epistemology of physics. Cognition and Instruction, 10(2\&3), 105-225.
Faisal, A. A., \& Wolpert, D. M. (2009). Near optimal combination of sensory and motor uncertainty in time during a naturalistic perceptionaction task. Journal of Neurophysiology, 101, 1901-1912.
Gerstenberg, T., Goodman, N., Lagnado, D. A., \& Tenenbaum, J. (2012). Noisy Newtons: Unifying process and dependency accounts of causal attribution. Paper presented at the Proceedings of the 34th Annual Meeting of the Cognitive Science Society, Sapporo, Japan.
Hamrick, J., Battaglia, P., \& Tenenbaum, J. (2011). Internal physics models guide probabilistic judgments about object dynamics. Paper presented at the Proceedings of the 33rd Annual Meeting of the Cognitive Science Society, Boston, MA.
Hudson, T. E., Maloney, L. T., \& Landy, M. S. (2008). Optimal compensation for temporal uncertainty in movement planning. PLoS Computational Biology, 4(7). doi: doi:10.1371/journal.pcbi. 1000130
Kaiser, M. K., Jonides, J., \& Alexander, J. (1986). Intuitive reasoning about abstract and familiar physics problems. Memory \& Cognition, 14(4), 308-312.
Kaiser, M. K., Proffitt, D. R., \& McCloskey, M. (1985). The development of beliefs about falling objects. Attention, Perception, \& Psychophysics, 38(6), 533-539.
Kaiser, M. K., Proffitt, D. R., Whelan, S. M., \& Hecht, H. (1992). Influence of animation on dynamical judgments. Journal of Experimental Psychology: Human Perception and Performance, 18(3), 669-689.
McCloskey, M., Caramazza, A., \& Green, B. (1980). Curvilinear motion in the absence of external forces: Naive beliefs about the motion of objects. Science, 210(5), 1139-1141.
Pittenger, J. B. (1985). Estimation of pendulum length from information in motion. Perception, 14, 247-256.
Pittenger, J. B. (1990). Detection of violations of the law of pendulum motion: Observers' sensitivity to the relation between period and length. Ecological Psychology, 2(1), 55-81.
Proffitt, D. R., \& Gilden, D. L. (1989). Understanding natural dynamics. Journal of Experimental Psychology: Human Perception and Performance, 15(2), 384-393.
Sanborn, A. N., Mansinghka, V. K., \& Griffiths, T. L. (2013). Reconciling intuitive physics and Newtonian mechanics for colliding objects. Psychological Review, 120(2), 411-437.
Saxberg, B. V. H. (1987). Projected free fall trajectories II: Human experiments. Biological Cybernetics, 56, 177-184.
Smith, K. A., \& Vul, E. (2013). Sources of uncertainty in intuitive physics. Topics in Cognitive Science, 5(1), 185-199.
Stocker, A. A., \& Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. Nature Neuroscience, 9(4), 578-585.
Téglás, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., \& Bonatti, L. L. (2011). Pure reasoning in 12-month-old infants as probabilistic inference. Science, 332, 1054-1059.
Todd, J. T., \& Warren, W. H. (1982). Visual perception of relative mass in dynamic events. Perception, 11, 325-335.
Weiss, Y., Simoncelli, E. P., \& Adelson, E. H. (2002). Motion illusions as optimal percepts. Nature Neuroscience, 5(6), 598-604.

\title{
Hear No Evil: Can Music Attenuate the Irrelevant Speech Effect?
}

\author{
Wei Jie Soh (weijiesoh31@gmail.com) \\ Stephen Wee Hun Lim (psylimwh @ nus.edu.sg) \\ Department of Psychology, National University of Singapore \\ 9 Arts Link, Singapore 117570
}

\begin{abstract}
This study aims to replicate the irrelevant speech effect (ISE) in a local context and, more important, is the first to directly investigate if musical information can reduce impairments imposed by the ISE on a serial word recall task. Thirty-five undergraduates from the National University of Singapore performed serial recall on 10 word lists. The lists were presented under 5 auditory conditions, namely: Music-Only, Combined (music with background speech), Scrambled music with background speech, Background Speech-Only and White Noise conditions. The Scrambled condition contained the same piece of music as the Combined condition except that it was re-arranged in a random fashion; the mission of this condition was to specifically provide a comparison basis to test if "musical structure" per se actually attenuates the ISE. A significant main effect of music conditions emerged. ISE was successfully replicated, where a significantly lower percentage of correct words was recalled in the Background Speech-Only condition compared to all other conditions. ISE was also successfully attenuated, but the present data suggest that musical structure per se was not (at least not entirely) responsible for the attenuation, since the Scrambled condition had superior performance than both the Combined and Background Speech-Only conditions. Here, we propose and discuss several novel theoretical models involving changing acoustical features, selective attention, and arousal to account for the present findings.
\end{abstract}

Keywords: Irrelevant speech effect; music; recall performance.

\section*{Introduction}

The irrelevant speech effect (ISE) is the finding that background speech significantly impairs serial recall performance, even when the background speech is irrelevant to the task (Farley, Neath, Allbritton \& Surprenant, 2007). First demonstrated by Colle and Walsh (1976), the researchers presented subjects with lists of eight consonantitems visually together with a passage read out in German. The background speech was considered irrelevant as participants were told to ignore the passage and that no subsequent recall of the background speech was required. Serial recall was significantly impaired in the irrelevant speech condition compared to the quiet (control) condition. The ISE is found to be robust and independent of speech intensity, within the range of 40 to 76 dB (Ellermeier \& Hellbrück, 1998). The effect is also significant regardless of whether the irrelevant speech is presented together with or after the word list (Miles, Jones \& Madden, 1991), and evident over repeated trials or sessions (Tremblay \& Jones, 1998).

The question of greater interest (and importance) is whether one could ever circumvent the ISE, given the potential costs on cognitive performance that are associated with the negative impacts of ISE under a variety of situations. A possible candidate to abate irrelevant speech is instrumental music, due to the fact that music has been found to modulate work performance. Lesiuk (2010), for instance, found that listening to preferred music led to improvements in performance within the context of highly cognitive demanding jobs.

This study had two goals. The first was to (first) replicate the ISE in a local context (among undergraduates at the National University of Singapore). The background speech, accordingly, comprised of contents related to undergraduates, ranging from modules, bid points, gossip and current news ensuring that the contents were distracting enough while trying to concentrate on learning a word list. Second, and more important, this study aimed to discover whether instrumental music, with all its purported positive effects and benefits on cognitive performance (e.g., Nantais \& Schellenberg, 1999; Schellenberg \& Hallam, 2005), can reduce the detriments of ISE during a serial word recall task. Accordingly, this study has been designed to contain five auditory conditions: (1) Instrumental Music-Only, (2) Combined (music with background speech), (3) Background Speech-Only, (4) Scrambled Music with background speech, and (5) White Noise.

Two specific hypotheses follow. First, under the Background Speech-Only condition, participants will have the worst recall performance compared to all other conditions, while Instrumental Music-Only and White Noise conditions will produce the best performances. This hypothesis, if supported, would mean that ISE effects are replicated in a local context, which further qualifies that instrumental music and white noise have less detriments on serial word recall than irrelevant speech does. Second, the Combined condition is predicted to yield superior performance in the recall task compared to the Background Speech-Only and Scrambled conditions would.

Hypothesis 2 addresses the possibility that the instrumental music - with its musical harmony and internal musical structure - may result in a more stable auditory scene for selective processing than the changing-state features of background speech. The Scrambled condition consists of the same piece of instrumental music, only rearranged to disrupt its internal musical structure. Hence, the Combined condition is predicted to enhance task performance compared to the Scrambled condition and Background Speech-Only conditions.

\section*{Method}

\section*{Participants}

Thirty-five undergraduates from psychology classes in National University of Singapore took part in the study and were awarded course credits for their participation. All participants reported normal hearing.

\section*{Design}

A 5 [Music Conditions: (1) Music-Only, (2) Background Speech-Only, (3) music with background speech (Combined) versus (4) Scrambled music with background speech, and (5) White Noise] \(\times 2\) [Word frequency: high versus low] within-subjects design was used.

\section*{Serial Word Recall List}

Eighty 4-letter English words were chosen for the 10 word lists (Lim \& Yap, 2010); orthographic neighborhood density (held constant at 3.33) and word frequency (high versus low; see Table 1) effects per se were not expected to emerge in this study (i.e., music condition effects, if any, ought to persist across high and low frequency words).

Table 1: Means and Standard Deviations of Logfrequency for Low and High Frequency Words.
\begin{tabular}{lcc} 
& \multicolumn{2}{c}{ Log-frequency } \\
\cline { 2 - 3 } Conditions & \(M\) & \(S D\) \\
\hline Low-frequency & 6.61 & 0.544 \\
High-frequency & 11.8 & 1.230 \\
\hline
\end{tabular}

\section*{Stimuli}

A total of five auditory conditions were created: MusicOnly, Background Speech-Only, music with background speech (Combined condition), Scrambled music with background speech, and White Noise. The backgroundspeech auditory track was superimposed at the same volume on Bach's Italian Concerto (First Movement) and Haydn's Piano Sonata in E-Flat Major, No. 52. The superimposed tracks were split into sets of 42 seconds each, in order to match the duration of each word list's presentation. This procedure is illustrated in Figure 1.

The same musical track was randomly split and rearranged to create the Scrambled condition, therefore maintaining the exact same number of musical notes while disrupting the musical structure. This prevented differences in number of musical notes from producing any differential (unintended) effects in recall performance

Intact Instrumental Music Track


Figure 1: Arrangement of auditory tracks and assignment of word lists.

\section*{Manipulation Check}

The manipulation check, which asked participants to recall the segments of conversations, was instituted to rule out the possibility that the complete or scrambled music (in the Combined and Scrambled conditions) was (merely) masking off the background speech track.

\section*{Procedure}

Participants were presented with the word lists paired with each of the five auditory conditions and were instructed to ignore the background auditory stimuli. Once the list stops, the auditory stimuli paused and the participants were given one minute to recall the words presented. Immediate recall was required after every list using a booklet provided. It was emphasized that only words recalled in the correct position would be scored as correct responses. Exposing participants to different segments specifically controlled for habituation and familiarity effects. In addition, the order of the five different auditory conditions was counterbalanced across the two test sessions.

\section*{Results}

\section*{Analysis of Manipulation Check}

The manipulation check was instituted in order to critically rule out the possibility that the music or scrambled music could merely be masking the speech information. Importantly, approximately \(73 \%\) of the participants recalled more than 2 categories of contents in the background speech. This high recall performance of speech contents constituted important evidence in suggesting that the music tracks did not (merely) mask the background speech.

\section*{Analysis of Word Recall Performance}

A \(5 \times 2\) repeated-measures ANOVA was conducted on the percentage of words recalled correctly. The two-way interaction was not significant as earlier predicted, \(F(4,136)\) \(=.653, M S E=.31, p=.626\), and data were subsequently collapsed across word frequency. The main effect of word frequency did not reach significance as well, \(F(1,34)=1.57\) , \(M S E=.031, p=.219\).

A significant main effect for music conditions emerged, \(F(4,136)=5.25, M S E=.71, p=.001\). The irrelevant speech effect was successfully replicated: Post hoc comparisons revealed that percentage of correct recall in the Background Speech-Only condition ( \(M=.554, S D=.280\) ) was significantly lower than recall in Music-Only ( \(M=.713\), \(S D\) \(=.274\) ), Combined ( \(M=.664, S D=.241\) ), Scrambled ( \(M=\) \(.741, S D=.259)\) and White Noise \((M=.698, S D=.275)\) conditions. This means that the Background Speech-Only condition yielded the worst recall performance compared to all other sound types.

While instrumental music appears to be influential in attenuating the ISE, an intriguing finding was that instrumental music per se - specifically its musical structure (or music-ness) - did not appear to attenuate ISE, due to the fact that the Scrambled condition produced significantly higher recall performance than both the Combined condition and Background Speech-Only condition did. The critical interpretation is that instrumental music attenuated ISE, but musical structure per se is not (at least not entirely) responsible for this effect. Figure 2 presents recall performance across conditions.


Figure 2: Plots and error bars of mean percentage recall across music conditions. Background Speech Only produced the lowest recall, while Combined and Scrambled yielded significantly higher recall than Background Speech-Only.

\section*{Discussion}

The present results show that the ISE was replicated in a local context. Recall performance in Background SpeechOnly was the worst compared to that in all other auditory conditions. However, the intriguing finding was the
apparent lack of evidence to support the hypothesis that musical structure per se can attenuate the ISE (in the Combined condition), given the observation that the Scrambled condition actually produced better recall scores than both the Combined and Background Speech-Only conditions did. Auditory masking, albeit a convenient explanation, clearly cannot account for the present data, because of the high percentage of speech contents recalled (73\%). Clearly, participants did process the background speech.

\section*{Towards a Hybrid Model of Changing States and Attention}

Here, we propose a novel hybrid model that combines the attention component from the feature model by Neath (2000) with the changing-state accounts from the O-OER model (Jones, Madden \& Miles, 1992) to account for the present data (i.e., improvements in recall performance under both the Combined and Scrambled conditions compared to Background Speech-Only condition).

According to the O-OER model, these changing state features give rise to multiple objects, which interfere with serial processing of the word list compared to a repeated, steady auditory stream. In this study, music did not impose additional processing because it may have less changing features than the irrelevant speech. Therefore, the music tracks are preferred over the irrelevant speech whereby in the Combined and Scrambled conditions, attention was diverted away from the damaging irrelevant speech. Additional cognitive resources can then be allocated towards the serial word recall task. Ahveninen et al. (2011), using multimodal techniques (PET, fMRI, MEG and EEG), found that auditory cortices can selectively deploy attention to segregate relevant sounds from noise, thereby mitigating the detrimental influence of irrelevant speech. In this study, the music track, with less changing-features, makes processing easier, delegating more cognitive resources for the serial word recall task, thereby explaining superior performances in Combined and Scrambled conditions.

An alternative explanation is that the cumulative presence of the additional auditory stimuli and irrelevant speech in this study led to an increase in distraction levels, resulting in a compensatory increase in attention to the serial recall task. Weissman, Warner and Woldorff (2004) found in their experiment that as the irrelevant stimulus increases in their distraction levels, a compensatory increase in selective attention follows. Therefore, an overall increase in distractibility of auditory stimuli can lead to a compensatory increase in attention, thereby explaining why performances are better in the present Combined or Scrambled condition.

Summarizing, these findings represent active processing by participants where changing acoustical features of the irrelevant speech and music tracks were compared and the latter (steadier) stream is preferred. Attention is either selectively deployed to the less distracting stream or increased via compensatory mechanisms, allocating more
attentional resources towards the serial recall task. Task performance is consequently enhanced.

However, it must be noted that the hybrid model makes the implicit assumption that music has less changingfeatures than irrelevant speech does, and this model would not particularly aim to differentiate between intact and scrambled music. Therefore, there is a possibility that the present results, where scrambled music yielded better recall performance than did the Combined and Background Speech-Only conditions, are not (yet) thoroughly accounted for by this model. We next briefly describe (for future work purposes) another property of music that might explain the attenuation of ISE.

\section*{Arousal-mood Hypothesis}

One particular property of music - arousal - may be promising to explain why scrambled music produced such superior recall performance. The arousal-mood hypothesis by Thompson, Schellenberg, and Husain (2001) argues that the tempo of music is related to arousal while its mode is linked to mood. Music in a major mode corresponds to a happy mood whereas minor mode to a sad mood (Husain, Thompson \& Schellenberg, 2001). The re-arrangement of the original Bach and Haydn sonatas music could, in fact, augment the perceived tempo in the scrambled track given its now more "staccato-like" (and therefore "rapid") quality (compared to its original unscrambled (and "unrushed") counterparts). The perceived faster tempo in the Scrambled condition could possibly have produced higher arousal states than did the perceived tempo in the Combined condition, which might directly predict recall performance. The view, in a sense, is that the Scrambled music then offered listeners with greater cognitive resources (due to heightened states of arousal) to engage in their recall task, than did unscrambled music.

\section*{Future Directions}

The intriguing finding was that the Scrambled condition in fact enhanced recall more than the Combined condition did. Since the musical structure (i.e., music-ness) of the present auditory stimuli did not appear to be (solely) responsible for this attenuation, future research, as recommended above, could explore effects of alternative (e.g., arousal) properties to understand the workings beneath ISE more directly.

\section*{Conclusion}

This study reports novel data that suggest that ISE can be attenuated (even in a local context) but how that the reason for this attenuation is not (solely) musical structure per se. Changing acoustical properties and arousal capabilities of the auditory stimuli may unveil how we might exactly attenuate the ISE. For scrambled music, its arousing properties may potentially attenuate ISE. Therefore, beyond changing-states and selective attention, music's arousing capabilities should be directly investigated in a future study. It is likely that both changing acoustical features and arousal
capacities found in music may collectively help attenuate the ISE. These are exciting predictions which would have brought us closer to answering the long-standing question of just why "music" is so capable of offering inoculation against a harsh auditory environment that comprises a host of distractions (e.g., why thousands of students around the world continue to listen to music whenever they study).

\section*{References}

Ahveninen, J., Hämäläinena, M., Jääskeläinenc, I.P., Ahlforsa, S.P., Huang, S., Lina, F.H., Raija, T., Sams, M., Vasiosa, C.P., \& Belliveau, J.P. (2011). Attention-driven auditory cortex short-term plasticity helps segregate relevant sounds from noise. PNAS, 108, 4182-4187.
Baddeley, A. D. (2003b). Working memory: Looking back and looking forward. Nature Reviews Neuroscience, 4, 829-839.
Colle, H.A., \& Welsh, A. (1976) . Acoustic masking in primary memory. Journal of Verbal Learning and Verbal Behavior, 15, 17-32.
Ellermeier, W., \& Hellbruck, J. (1998). Is level irrelevant in "irrelevant speech"? Effects of loudness, signal-to-noise ratio, and binaural unmasking. Journal of Experimental Psychology: Human Perception \& Performance, 24, 1406-1414.
Farley, L.A., Neath, I., Allbritton, D.W. \& Surprenant, A.M. (2007). Irrelevant speech effects and sequential learning. Memory \& Cognition, 35, 156-165.
Husain, G., Thompson, W.F., \& Schellenberg, E.G. (2002). Effects of musical tempo and mode on arousal, mood, and spatial abilities. Music Perception, 20, 151-171.
Jones, D. M., Madden, C., \& Miles, C. (1992). Privileged access by irrelevant speech to short-term memory: The role of changing state. Quarterly Journal of Experimental Psychology, 44A, 645-669.
Lesiuk, T. (2010). The effect of preferred music on mood and performance in a high-cognitive demand occupation. Journal of Music Therapy, 47, 137-154.
Lim, S. W. H., \& Yap, M. J. (2010). Distributional analyses in visual lexical decision: Orthographic neighborhood density and word frequency effects. Poster presentation at the 32nd Annual Conference of the Cognitive Science Society, Portland, Oregon, USA; paper published in S. Ohlsson \& R. Catrambone (Eds.), Proceedings of the 32nd Annual Conference of the Cognitive Science Society (pp. 700-705). Austin, TX: Cognitive Science Society.
Miles, C., Jones, D. M., \& Madden, C. A. (1991). Locus of the irrelevant speech effect in short-term memory. Journal of Experimental Psychology: Learning, Memory, \& Cognition, 17, 578-584.
Nantais, K. M., \& Schellenberg, E. G. (1999). The Mozart effect: An artifact of preference. Psychological Science, 10, 370-373.
Neath, I. (2000). Modeling the effects of irrelevant speech on memory. Psychonomic Bulletin \& Review, 7, 403-423.
Schellenberg, E. G., \& Hallam, S. (2005). Music listening and cognitive abilities in 10- and 11-year-olds: The Blur
effect. Annals of the New York Academy of Sciences, 1060, 202-209.
Tremblay, S., \& Jones, D. M. (1998). Role of habituation in the irrelevant sound effect: Evidence from the effects of token set size and rate of transition. Journal of Experimental Psychology: Learning, Memory, \& Cognition, 24, 659-671.
Weissman, D.H., Warner, L.M., \& Woldorff, M.G. (2004). The neural mechanisms for minimizing cross-modal distraction. Journal of Neuroscience, 24, 10941-10949.

\title{
Priming randomness increases the evaluation of ritual efficacy
}

\author{
André L. Souza (andre.souza@crdh.concordia.ca) \\ Concordia University \\ Department of Psychology, 7141 Sherbrooke St. W. \\ Montreal, QC H4V 1N3 Canada
}

\author{
Cristine H. Legare (legare@psy.utexas.edu) \\ The University of Texas at Austin \\ Department of Psychology, 1 University Station, A8000
}

Austin, TX 78712 USA

\begin{abstract}
Reestablishing feelings of control in the face of uncertainty is a fundamental motive for human behavior. We propose that rituals (i.e., socially stipulated, causally opaque practices) provide a means for coping with the aversive feelings associated with randomness due to the perception of a connection between ritual action and a desired outcome. Two experiments were conducted (one in Brazil [ \(N=40\) ] and another in the U.S. [ \(N=94]\) ) to evaluate how the perceived efficacy of rituals is affected by feelings of randomness. In a between-subjects design, the Scramble Sentence Task was used as a priming procedure in three conditions (i.e., randomness, negativity, and neutral) and participants were then asked to rate the efficacy of rituals used for problem-solving purposes. The results demonstrate that priming randomness increased participants' perception of ritual efficacy relative to negativity and neutral conditions. Implications for increasing our understanding of the relationship between perceived control and ritualistic behavior are discussed
\end{abstract}

Keywords: Randomness; Ritual; Perception of Control; Supernatural Cognition.

\section*{Introduction}

Anthropologists have long noted that the use of rituals for instrumental purposes is linked to conditions of risk and uncertainty (Malinowski, 2004). When Malinowski visited the Trobriand Islands of New Guinea, for example, he noted that at times the Trobrianders would base their behavior on practices with clear causal rationales while at other times they would rely on causally opaque practices such as ritual. The Trobrianders rarely relied on ritual when fishing in a reliable and safe setting such as the lagoon; they described their successes and failures in terms of skill. In contrast, extensive ritual preceded the uncertain and dangerous conditions of deepsea fishing.

The Trobriand fishermen are not alone in their use of ritual to restore feelings of control when confronted with uncertainty (Souza \& Legare, 2011). On college campuses, for instance, up to \(70 \%\) of students employ such strategies to assist with performance on exams (Gallagher \& Lewis, 2001) or athletic competitions (Bleak \& Frederick, 1998; Ciborowski, 1997; Todd \& Brown, 2003; Van Raalte, Brewer, Nemeroff, \& Linder, 1991; Vyse, 2000; Womack, 1992).

To the extent that the rituals have little or no actual bearing on the success of instrumental outcomes (Lobmeyer \& Wasserman, 1986) through a process of physical causation (Legare \& Herrmann, 2013; Legare \& Souza, 2012;

Humphrey \& Laidlaw, 1994), we propose that one of the functions of rituals is to maintain an illusion of control, a phrase coined by Langer (1975). An illusion of control is inferred when participants believe or respond as if contingencies between their behavior and the outcome exist, even if the outcomes are random (Alloy, Abramson, \& Viscusi, 1981; Matute, 1994). Regardless of how the illusion of control is manipulated, all dependent measures reflect a belief that one's actions can influence an outcome that is, in fact, outside of their control.

There is considerable empirical evidence demonstrating that lack of perceived control - an individual's belief that he or she cannot predict and affect future events - has applied consequences and is associated with a number of negative outcomes. For example, it contributes to the tendency to demonstrate depressive and pessimistic behavior and avoid facing challenging situations (Fast, Gruenfeld, Sivanathan, \& Galinsky, 2009). Conversely, feelings of control promote increased self-esteem, optimism and greater sense of agency (Scheier, Carver, \& Bridges, 1994). Despite the benefits associated with feelings of control (Kofta, Weary, \& Sedek, 1998), people frequently experience situations in which they lack the capacity to exert the control they desire. Many of the most pervasive ailments that afflict humans such as chronic illness (e.g., cancer), economic insecurity (e.g., unemployment) and interpersonal problems (e.g., infidelity) are often not within our control.

When people are unable to control and predict their environment, attributional biases are activated and strategies are implemented to restore feelings of control (Underwood, 1996; Vaughn \& Weary, 2003; Weary \& Jacobson, 1997; Weary, Jacobson, Edwards, \& Tobin, 2001). For example, people detect correlations among random sets of stimuli that are presumably unrelated when they are primed with feelings of lack of control (Whitson \& Galinsky, 2008). There is also evidence that when desire for a coveted item and uncertainty are high and personal control is lacking, people may be more likely to help others, as if they can encourage fate's favor by doing good deeds proactively.

Seminal work on the illusion of control and magical thinking has examined first-person experiences with superstitious behavior (Keinan, 1994) or procedures that approximate rituals (Rudski, 2001; Rudski \& Edwards, 2007). Keinan (1994)
explains the increase in superstitious behavior under conditions of stress and uncertainty as an attempt to regain control. Rituals, which we define as conventional, causally opaque procedures may provide a means for coping with the aversive feelings associated with randomness by reestablishing feelings of control. We propose that the structure of ritual can be interpreted in light of intuitive causal beliefs (Legare \& Souza, 2012) and predict that intuitive causal reasoning, not familiarity with the content of particular rituals, drives how ritual efficacy is evaluated.

Despite the fact that engaging in causally opaque practices may seem to be a paradoxical means of increasing perceived control, we hypothesize that this is possible because rituals provide a socially stipulated and culturally sanctioned opportunity to exert agency through action, thereby giving the illusion of increased control (Thompson, Armstrong, \& Thomas, 1998). We propose that priming randomness increases the activation of attributional biases to detect a connection between action and outcome as a means of reestablishing feelings of control. The perception of a connection increases the evaluation of ritual efficacy. We predict that this effect occurs not only in first-person experiences with uncertainty (Keinan, 1994,2002 ) but also implicitly when evaluating the experiences of others. Two studies investigated this prediction directly by examining whether priming randomness affects the perception of the efficacy of rituals.

Study 1 was conducted in Brazil, a cultural context in which a particular type of ritual - called simpatia - is used to treat a variety of problems. Simpatias are ritualistic remedial procedures used to solve everyday biological (e.g., sinusitis, asthma), psychological (e.g., depression, anxiety), and interpersonal problems (e.g., attracting a partner, infidelity). They are available to the general population, are relatively low-cost, and do not require any specialized expertise to be performed. Despite the lack of a physical-causal mechanism underlying their efficacy, simpatias are widely endorsed and used for a greater variety of problem-solving purposes. For example, a simpatia to cure depression might require a person to drink coconut water straight from the coconut and then bury the coconut husk in a garden full of flowers (Legare \& Souza, 2012).

Legare and Souza (2012) designed experimental simpatias to match the characteristics and content of existing ones. A selection of these simpatias was used in the current studies to assess perceptions of ritual efficacy. To prime feelings of randomness, we used a previously validated task called The Scrambled Sentence Task - SST (Kay, Moscovitch, \& Laurin, 2010). A more detailed description of the task is provided below.

\section*{Study 1}

\section*{Methods}

Participants Forty Brazilian Portuguese-speaking adults participated in the study. Participants were recruited from the metropolitan area of the city of Belo Horizonte located in
the southeastern region of Brazil. They were recruited from public health centers located in a peripheral neighborhood of Belo Horizonte. The public health centers (known as Posto de Saúde) are centers maintained by the city administration, and serve the population from the community in which the center is located.

According to the Brazilian Institute of Geography and Statistics, Belo Horizonte has a population of over 6,082,776 people. The ethnic composition of the population is \(47 \%\) Black, \(41 \%\) Pardo (mixed-race), and \(12 \%\) White.. In terms of religious composition, over \(68 \%\) of the population selfidentify as Catholic, \(19 \%\) Protestant, and \(8 \%\) of the population reported not having any religious affiliation. Although census data has traditionally failed to capture the range of religious traditions available in Brazil (especially those of AfroBrazilian roots), the endorsement of simpatias exists across all religious groups.

Materials To assess the perceived efficacy of rituals, we used simpatias designed by Legare and Souza (2012). They were designed to match the characteristics of existing simpatias to maximize ecological validity. A previously validated task called The Scrambled Sentence Task - SST was used to prime randomness in one condition and negativity in the other (Kay et al., 2010). A baseline condition containing neutral words was also created. In the SST, each participant is given 20 scrambled sentence strings composed of five words each. Participants were asked to rearrange four of the five words to form a meaningful sentence and then to cross out the one word left out. For half the participants, the word sets contained words related to randomness (e.g., chaotic) and for the other half, these words were replaced with negatively valenced control words (e.g., lazy). This procedure is similar to the one used by Kay et al. (2010).

Procedure Participants were randomly assigned to one of the two conditions (i.e., randomness condition and negativity condition). The second author, a native speaker of Brazilian Portuguese, interviewed each participant individually. Each participant was given a set of words (according to the condition assigned) and was asked to form sentences. Participants were allowed to take as long as they wanted to for the sentences. For the randomness condition, 10 of the 20 lists contained randomness-related words, whereas for the negativity condition, these 10 words were replaced with negatively valenced words.

Following the priming task, participants were presented with six simpatias paired with specific problems. The order of presentation was randomized across participants. Then participants were asked: "In a scale from 1 to 10, 1 being EFFECTIVE and 10 being INEFFECTIVE, how much do you think this simpatia is effective for treating this specific problem?"

\section*{Results and Discussion}

Preliminary analysis revealed that the priming manipulation affected all six simpatias equally, that is, in terms of efficacy ratings, there was no main effect of specific simpatia used and no interaction between specific simpatias and the priming procedures. Thus, the ratings of the six simpatias were averaged to form a single index of ritual efficacy for each participant. Results revealed that participants in the randomness condition rated the simpatias as significantly more effective ( \(M=4.33, S D=.31\) ) than participants in the negativity condition \((M=4.64, S D=.40), t(38)=2.65, p<.05\), (simpatias with lower ratings were judged to be more effective than simpatias with higher ratings).

This finding supports the hypothesis that the evaluation of ritual efficacy increases when the motivation to reestablish control is primed. Rituals may provide a mechanism for accomplishing this goal (Keinan, 2002). Alternatively, however, this pattern of data could potentially be explained by the possibility that negativity reduced perceptions of efficacy, instead of randomness increasing perceptions of efficacy. To address this potential alternative explanation, in Study 2 we included a third condition containing neutral words. In previous research by (Legare \& Souza, 2012), the evaluation of ritual efficacy did not vary between populations familiar with (e.g. in Brazil) and unfamiliar with (e.g. in the U.S.) simpatias. Thus to examine the generalizability of the results from Study 1 in a population unfamiliar with the content of these culturally specific rituals, Study 2 was conducted in the U.S.

\section*{Study 2}

\section*{Methods}

Participants Ninety-four undergraduate students at a large research university located in the southwest of the United States participated in Study 2 for course credit.

Materials The materials used in Study 2 were identical to the materials used in Study 1 except that they were translated from Brazilian Portuguese into American English by the second author.

Procedure The procedure for Study 2 was identical to Study 1 except that the simpatias and efficacy ratings questions were presented using E-Prime rather than being read to participants. Again, participants were asked to rearrange four of the five words to form a meaningful sentence and then to cross out the one word left out. For 33 participants (randomly selected), the word sets contained words related to randomness (e.g., chaotic), for 32 participants, these words were negatively-valenced words (e.g., lazy) and finally for 29 participants, the words were neutral words extracted from the ANEW database (Bradley \& Lang, 1999).

\section*{Results and Discussion}

The objectives of Study 2 were to examine the generalizability of the effect in a cultural context unfamiliar with simpatias and explore the possibility that negative words reduced


Figure 1: Mean Efficacy Ratings per Condition in Study 2
the evaluation of ritual efficacy. As predicted, although the simpatias were rated as less effective in the U.S. sample than in the Brazilian sample (consistent with Legare \& Souza, 2012), a one-way ANOVA revealed a main effect of condition, \(F(2,91)=5.07, p<0.05, \eta^{2}=.10\) on the efficacy ratings. Post-hoc tests (Bonferroni corrected) demonstrated that participants primed with randomness rated the simpatias as significantly more efficacious ( \(M=8.06, S D=1.64\) ) than participants in the neutral condition \((M=9.01, S D=.97), t(60)=\) \(-2.71, p<.002\), and marginally more efficacious than participants in the negativity condition ( \(M=8.84, S D=1.02\) ), \(t(63)\) \(=2.27, p=0.02\). Notably, there was no significant difference between the efficacy ratings of people in the neutral condition and negativity condition (See Figure 1). The results demonstrate that even with unfamiliar content, priming randomness increased ritual efficacy evaluations, consistent with the results of Study 1. Moreover, the lack of difference between the negativity and neutral condition suggest that randomness increases perceptions of ritual efficacy, rather than negativity decreasing ritual efficacy evaluation.

\section*{Discussion}

In the face of randomness, attributional biases are activated and strategies are used to cope with feelings of lack of control (Weary \& Edwards, 1994; Weary \& Jacobson, 1997; Weary et al., 2001; Wichman, Brunner, \& Weary, 2008). We propose that rituals provide a means for coping with the aversive feelings associated with lack of control. The current studies sought to examine this possibility empirically by investigating the extent to which priming feelings of randomness influences perceptions of ritual efficacy.

Our results support the hypothesis that perceptions of the
efficacy of ritualistic behavior are influenced by the drive to regain a sense of control. Participants primed in the randomness condition rated simpatias as significantly more efficacious than participants in the control condition. One potential explanation for this effect is that the experience of randomness triggered by the manipulation activated a need to reestablish perceived control. Rituals may provide a mechanism for accomplishing this by providing an opportunity to posit a connection between action and outcome.

Examining the interplay of perceived control and ritual is of pervasive interdisciplinary interest with longstanding roots in both anthropology and experimental psychology (Keinan, 1994; 2002; Rudski \& Edwards, 2007). Despite this interdisciplinary interest, these studies are the first to examine the relationship between priming randomness and reasoning about the efficacy of ritualistic practices used by others. By examining this relationship experimentally, we have demonstrated that ritual may serve as a mechanism for reestablishing the perception of control and have provided insight into the cognitive underpinnings of the evaluation of ritual efficacy. Studying ritual from this perspective contributes to the body of research (Boyer \& Liénard, 2006; Kay, Gaucher, Napier, Callan, \& Laurin, 2008; Kay et al., 2010; Keinan, 1994, 2002; Rudski, 2001) demonstrating that one of the functions rituals serve is to make the world seem more comprehensible, certain, and predictable.

\section*{Acknowledgments}

We are grateful for funding from the Cognition, Religion, and Theology Program funded by the John Templeton Foundation.

\section*{References}

Alloy, L., Abramson, L., \& Viscusi, D. (1981). Induced mood and the illusion of control. Journal of Personality and Social Psychology, 41(6), 1129.
Bleak, J., \& Frederick, C. (1998). Superstitious behavior in sport: Levels of effectiveness and determinants of use in three collegiate sports. Journal of Sport Behavior.
Boyer, P., \& Liénard, P. (2006). Precaution systems and ritualized behavior. Behavioral and Brain Sciences, 29(06), 635-641.
Ciborowski, T. (1997). " superstition" in the collegiate baseball player. The Sport Psychologist; The Sport Psychologist.
Fast, N., Gruenfeld, D., Sivanathan, N., \& Galinsky, A. (2009). Illusory control a generative force behind power's far-reaching effects. Psychological Science, 20(4), 502-508.
Gallagher, T., \& Lewis, J. (2001). Rationalists, fatalists, and the modern superstitious: Test-taking in introductory sociology. Sociological inquiry, 71(1), 1-12.

Humphrey, C., \& Laidlaw, J. (1994). The archetypal actions of ritual: a theory of ritual illustrated by the jain rite of worship. Clarendon Press Oxford.
Kay, A., Gaucher, D., Napier, J., Callan, M., \& Laurin, K. (2008). God and the government: testing a compensatory control mechanism for the support of external systems. Journal of personality and social psychology, 95(1), 18.
Kay, A., Moscovitch, D., \& Laurin, K. (2010). Randomness, attributions of arousal, and belief in god. Psychological Science, 21(2), 216-218.
Keinan, G. (1994). Effects of stress and tolerance of ambiguity on magical thinking. Journal of Personality and Social Psychology, 67(1), 48.
Keinan, G. (2002). The effects of stress and desire for control on superstitious behavior. Personality and Social Psychology Bulletin, 28(1), 102-108.
Kofta, M., Weary, G., \& Sedek, G. (1998). Personal control in action: Cognitive and motivational mechanisms. Springer.
Langer, E. (1975). The illusion of control. Journal of personality and social psychology, 32(2), 311.
Legare, C., \& Herrmann, P. (2013). Cognitive consequences and constraints on reasoning about ritual. Religion, Brain \& Behavior, 3(1), 63-65.
Legare, C., \& Souza, A. (2012). Evaluating ritual efficacy: Evidence from the supernatural. Cognition, 124, 1-15.
Lobmeyer, D., \& Wasserman, E. (1986). Preliminaries to free throw shooting: Superstitious behavior? Journal of Sport Behavior.
Malinowski, B. (2004). Magic, science and religion and other essays 1948. Kessinger Publishing.
Matute, H. (1994). Learned helplessness and superstitious behavior as opposite effects of uncontrollable reinforcement in humans. Learning and Motivation, 25(2), 216232.

Rudski, J. (2001). Competition, superstition and the illusion of control. Current Psychology, 20(1), 68-84.
Rudski, J., \& Edwards, A. (2007). Malinowski goes to college: Factors influencing students' use of ritual and superstition. The Journal of general psychology, 134(4), 389-403.
Scheier, M., Carver, C., \& Bridges, M. (1994). Distinguishing optimism from neuroticism (and trait anxiety, selfmastery, and self-esteem): a reevaluation of the life orientation test. Journal of personality and social psychology, 67(6), 1063.
Souza, A., \& Legare, C. (2011). The role of testimony in the evaluation of religious expertise. Religion, Brain \& Behavior, 1(2), 146-153.
Thompson, S., Armstrong, W., \& Thomas, C. (1998). Illusions of control, underestimations, and accuracy: a control heuristic explanation. Psychological bulletin, 123(2), 143.
Todd, M., \& Brown, C. (2003). Characteristics associated
with superstitious behavior in track and field athletes: Are there ncaa divisional level differences? Journal of Sport Behavior.
Underwood, G. (1996). Implicit cognition. Oxford University Press.
Van Raalte, J., Brewer, B., Nemeroff, C., \& Linder, D. (1991). Chance orientation and superstitious behavior on the putting green. Journal of Sport Behavior.
Vaughn, L., \& Weary, G. (2003). Causal uncertainty and correction of judgments. Journal of Experimental Social Psychology, 39(5), 516-524.
Vyse, S. (2000). Believing in magic: The psychology of superstition. Oxford University Press, USA.
Weary, G., \& Edwards, J. (1994). Individual differences in causal uncertainty. Journal of Personality and Social Psychology, 67(2), 308.
Weary, G., \& Jacobson, J. (1997). Causal uncertainty beliefs and diagnostic information seeking. Journal of Personality and Social Psychology, 73(4), 839.
Weary, G., Jacobson, J., Edwards, J., \& Tobin, S. (2001). Chronic and temporarily activated causal uncertainty beliefs and stereotype usage. Journal of Personality and Social Psychology, 81(2), 206.
Whitson, J., \& Galinsky, A. (2008). Lacking control increases illusory pattern perception. Science, 322(5898), 115117.

Wichman, A., Brunner, R., \& Weary, G. (2008). Immediate and delayed effects of causal uncertainty inductions on uncertainty accessibility. Journal of Experimental Social Psychology, 44(4), 1106-1113.
Womack, M. (1992). Why athletes need ritual: A study of magic among professional athletes. Sport and religion, 191-202.

\section*{APPENDIX - Experimental Simpatias}

Employment In the first day of last quarter phase of the moon, take the milk from a coconut and give it to the unemployed person to drink. After that, ask the person to spit three times in the hole made in the coconut. Following this, light up a brand-new white candle and drop the wax around the hole until the hole is sealed. Take the coconut to a far away beach or river.

Depression For five days, the person with depression should go to a crossroad. While there, the person should say: "Depression, stay here!" The person should not walk through the crossroad for one year.

Infidelity Throw a shoe and a shirt of the unfaithful person into a streaming river unbeknownst to the person. As the river flows away, you say: "I hope the river takes the infidelity away as fast as it can." Take some of the water from the river and keep it somewhere in the house.

Evil-Eye Fill a cup with sand spit inside the cup. Seal the cup and bury it upright before the sunrise.

Lack of Luck Get an orange, peel it, squeeze its juice and bury its flesh. Place the peel on top of the dirt. Drink the juice three times a day (morning, afternoon, and evening).

Lack of Money Collect seven red apples directly from an apple tree. In the morning, before eating anything, peel the apples, eat them and save the peel. Right before going to bed, make a tea with the peel.

\title{
Learning, Feedback and Information in Self-Organizing Communication Systems
}

\author{
Matthew Spike (matthew.spike@ed.ac.uk) \\ Kevin Stadler (kevin.stadler@ed.ac.uk) \\ Simon Kirby (simon@ling.ed.ac.uk) \\ Kenny Smith (kenny@ling.ed.ac.uk) \\ Language Evolution and Computation Research Unit, School of Philosophy, Psychology \& Language Sciences University of Edinburgh, Dugald Stewart Building, 3 Charles Street, Edinburgh, EH8 9AD, UK
}

\begin{abstract}
Communication systems reliably self-organize in populations of interacting agents under certain conditions. The various fields which model this - game theory, cognitive science and evolutionary linguistics - make different assumptions about the learning and behavioral processes which are responsible. We created an exemplar-based framework to directly compare these approaches by reproducing previously published models. Results show that a number of mechanisms are shared by the systems which can construct optimal communication. Three general factors are then proposed to underlie any selforganizing learned system.
\end{abstract}

Keywords: cultural evolution; communication; selforganization; reinforcement learning; feedback learning; observational learning

\section*{Introduction}

Human communication is a mostly learned behavior, while signaling behavior in the natural world appears to have a major genetic component. While Darwinian natural selection is argued to be the driving force behind the development of such innate capacities (e.g. Scott-Phillips et al., 2012 and Oliphant, 1996), the origin of learned communication is less clear. Effective communication requires consensus within a population; how is this reached given the arbitrary mapping between signal and meaning? In the absence of external or internal guidance, the emergent agreement must be the effect of not just global factors, such as how populations are connected and change over time, but crucially local ones also, for example how individuals learn and interact. Populationlevel behavior can therefore provide insights into aspects of human cognition.

The problem of self-organization of learned communication systems has been investigated by researchers working in game theory, artificial intelligence and evolutionary linguistics. The approaches taken by the different fields have much in common: all investigations focus on how two or more agents can effectively arrive at a mutually agreed set of signaling conventions through repeated interactions (or language games), and they all rely heavily on computational and mathematical modeling. However, the different theoretical perspectives have an understandable impact on how the models are designed and interpreted. In particular, the models of learning, interaction and population dynamics are distinct: game theory concentrates on small populations using varieties of reinforcement learning; feedback in closed groups is central to work in AI; in evolutionary linguistics intergenerational observational learning is the dominant paradigm. Re-
searchers have come to apparently conflicting conclusions regarding exactly which aspects of learning and interaction are crucial for the emergence of signaling. The aim of this paper is to reconcile these views by showing that all proposed solutions have three properties in common, a fact that has been obfuscated by the differing theoretical approaches. Individual bias against homonymy, along with the ability to transmit information about internal representations and a mechanism to discard information are argued to underlie the ability to self-organize successful communication.

\section*{Review}

Lewis (1969) devised his classic signaling game in line with game-theoretic principles. A speaker's signal triggers an action in the hearer: the resulting payoff, and thus reinforcement, depends on the state of the world, which is known only to the speaker. If the number of signals, acts and equiprobable states are all held at two, with equal non-conflicting payoffs, the game is proven to always converge upon an optimal signalling system (Beggs, 2005). Adjusting any of these parameters, however, quickly leads to pooling equilibria, where non-optimal communication strategies become attractors in the system. Barrett (2006) shows that while such sub-optimal situations will unavoidably occur when there are more than two possible states, systems can generally escape the pooling equilibria by enforcing memory limitations or including negative reinforcement (punishment of unsuccessful signals).

Steels' 1998 seminal Talking Heads experiment gave rise to a plethora of naming games which investigate how static populations can converge on functional and efficient naming conventions for a number of objects when agents are able to provide feedback to each other. Instead of observing a world state, speakers are said to randomly pick a topic from a communicative context. Key differences from the signaling game are that agents can indicate their intended referent in the case of communicative failure in some 'extra-linguistic' manner (so-called corrective feedback), and that agents can introduce new signals (or names).

Such systems inevitably develop functional communication, but each object ends up with large number of synonyms, a result of the ability to innovate novel signals. By introducing competition between synonyms for the same object, the systems are driven into an efficient state where each object is known by only one label. De Vylder \& Tuyls (2006) provide a mathematical proof that amplification of the input distribution of names is indeed sufficient to guarantee con-

Table 1: Model Comparison
\begin{tabular}{c|c|c|c|c}
\hline & Barrett & Steels & Oliphant \& Batali & Smith \\
\hline transmission & horizontal & horizontal & vertical & vertical \\
model type & mathematical & associative & associative & neural \\
modify hearer/speaker? & \(\mathrm{H} \& \mathrm{~S}\) & \(\mathrm{H} \& \mathrm{~S}\) & H & H \\
interaction & mutual payoff & feedback & observation & observation \\
learning features & forgetting/negative reinforcement & inhibition & obverter & inhibition \\
production \& reception & stochastic & deterministic & deterministic & deterministic \\
\hline
\end{tabular}
vergence of the naming game. Agents that implement such amplification are said to employ lateral inhibition to dampen name competitors, the most well-known being Baronchelli et al. (2006)'s minimal strategy. Baronchelli (2010) shows that only the hearer need be modified for effective convergence.

Taking yet another approach, iterated learning is the collective term for a large number of computational and experimental studies which combine varieties of observational learning with intergenerational population turnovers (Kirby et al., 2008). Oliphant \& Batali (1997) is one such example: their obverter strategy is derived from the mathematical result that if agents have perfect information about the internal state of the population, choosing signals by maximizing the chance of correct interpretation always results in the population converging on optimal communication. In simulations where agents use only incomplete information about the population gained through intergenerational learning, the obverter strategy still results in population convergence. In another study, Smith (2002) investigated the role of learning bias using populations of agents represented by Hebbian networks. Results showed that biases against homonymy and synonymy are necessary to produce optimal signaling.

The engine which drives the evolution of optimal signaling is variously stated: for reinforcement learning, it is communicative success; for the feedback models, it is the information gained through mutual alignment. Learning in the above models is horizontal; it takes place in static, closed groups. Intergenerational or vertical learning is employed by observational learners in iterated learning models which focus on individual learning biases, and obverters which stress the importance of explicitly maximizing the chance of being understood. A comparison of the above approaches leads to few clear conclusions regarding which learning and interaction features are responsible for convergence. Table 1 shows how the models contrast over many dimensions. The following section describes how the models were reproduced in a unified framework.

\section*{Replications}

An exemplar-style model was used to replicate the four models described above so that the effect of their different design features could be compared directly. Exemplar models have been employed to solve linguistic problems such as categorization (see e.g. Pierrehumbert, 2001). Learning involves
storing packets of perceptual information with discrete category labels. Our framework represents each exemplar as a simple pairing between a signal and a meaning, where 'signal' can also be read as 'name', and 'meaning' is equivalent to both objects in naming games as well as world-states and actions from signalling games. When an agent maps a signal to a meaning, a single exemplar is stored. As such, the framework does not represent a fundamental departure from network and association weight models, but does suggest the simplification of aspects of these models in ways which are detailed below.

A stored exemplar is atomic, and can not be modified in any way apart from wholesale deletion. Production and interpretation of signals can be deterministic or stochastic. With stochastic methods (excepting obverters) the probabilities of producing or interpreting a signal \(s\) of a total \(S\) signals in association with meaning \(m\) from a total \(M\) meanings are given in Formula 1 below, where \(n_{i j}\) represents an agent's count of exemplars associating meaning \(i\) and signal \(j\). Deterministic methods (also known as winner-take-all or WTA) always select the signal or meaning which yields the highest probability.
\[
\begin{equation*}
P(s \mid m)=\frac{n_{m s}}{\sum_{i=1}^{S} n_{m i}} \text { and } P(m \mid s)=\frac{n_{m s}}{\sum_{j=1}^{M} n_{j s}} \tag{1}
\end{equation*}
\]

Our framework is able to capture deterministic and stochastic behavior, as well as both static and changing populations, and the various manipulations of agents' internal representations employed by each of the models discussed above. For the sake of comparison, some parameters are held constant throughout all simulations presented here: populations consist of 10 agents and there are 5 available signals and meanings, where each meaning is equally likely to be selected. Populations are unstructured, with any two agents equally likely to interact. For models using vertical learning, a single new agent is trained on the data of the existing population at each iteration. The new agent then replaces the oldest member of the population.
In closed groups without population turnover, two agents are picked at random from the group at each time step, with one designated the speaker and the other the hearer. After each interaction, the hearer is updated according to the particular rules of that model, specified below. When lateral inhibition of synonyms and/or homonyms is employed,


Figure 1: (Replication of Barrett 2006) The proportion of 10,000 simulations which had converged to an optimal communication system after a given number of iterations, using negative reinforcement without a memory limit, and basic reinforcement with memory limits of 40 and 50 exemplars.
a newly stored exemplar results in the deletion of one randomly selected exemplar with competing signal/meaning associations. \({ }^{1}\) When a memory limit is included in a model, this is instantiated by enforcing a maximum number of \(n\) stored exemplars per agent. When this is exceeded, one exemplar is selected at random for deletion.

Communicative success was measured analytically by looking at the outcome of all possible communicative interactions over the entire population after each time step. 10,000 individual simulations were run for each configuration of each replication, and the number of iterations taken for each to converge on optimal signaling over the population was recorded. The cumulative distribution of converged populations over time was then plotted, as seen in Figures 1-4.
1. The reinforcement models used by Skyrms and Barrett employ Roth-Erev learning (Roth \& Erev, 1995), which maps exactly onto the exemplar model where behaviour is directly proportional to the relative frequency of memory tokens. When agents produce a signal for a given meaning, they do so by selecting stochastically from all stored exemplars associated with that meaning; interpretation is done similarly. Crucially, however, a new exemplar memory is only stored in the case of communicative success. \({ }^{2}\) Repli-

\footnotetext{
\({ }^{1}\) For the relevant models, lateral inhibition presented an issue: the original models decremented each competing weight equally. This implies that a single added exemplar would be responsible for the deletion of many others. As such, both 'maximal' (many deletions) and 'minimal' (only one deletion) interference were examined. In 10,000 simulations no difference was found between the time taken to converge using either strategy: for the results presented here, the minimal strategy with one random deletion was used.
\({ }^{2}\) For this reason, agents in this game are initialized with an initial copy of every possible exemplar: without this, each agent would be locked in to the first received signal mapping for each meaning.
}


Figure 2: (Replication of Steels \& Loetzsch 2012) The proportion of 10,000 simulations which had converged to an optimal communication system after a given number of iterations, using corrective feedback when either only the hearer or both speaker and hearer were modified.
cations of the basic model (not shown here) confirm Barrett's analysis: only a small proportion of simulations ever converge to even \(95 \%\) communicative accuracy, and even then only after long periods. The model was then modified to include either a memory limit, as described above, or negative reinforcement. With the latter, failed communication would cause the hearer to delete one exemplar of the unsuccesfully interpreted type. As shown in Figure 1, both mechanisms lead to near-certain convergence.
2. The feedback model described in Steels \& Loetzsch (2012) utilizes a complicated system of weighting adjustments. This was implemented in a simpler form: only one exemplar is added at a time, and there is no ability to innovate new signals beyond the five available. As confirmed in Baronchelli (2011), modification of the speaker is not a requirement for convergence, as shown in Figure 2. When lateral inhibition of homonyms was removed, signaling systems failed to develop. A further observation is that when corrective feedback is removed as well (i.e. when a speaker is unable to indicate its intended meaning after a failed communication), the model becomes identical to reinforcement learning, where signaling can only develop via negative reinforcement or memory limitations (see above).
3. Oliphant and Batali's (1997) obverters were replicated in both the original WTA version and a new stochastic one. Obverters produce a signal that maximizes the chances of being correctly understood. As such, the second equation in Formula 1 above defining the interpretation of a signal is used in obverter production. Formula 2 below defines the stochastic production function: In WTA production, the signal with the greatest chance of correct interpretation is


Figure 3: (Replication of Smith 2002) The proportion of 10,000 simulations which had converged to an optimal communication system after a given number of iterations, using stochastic production with inhibition of homonymy and synonymy, only homonymy, and WTA production with only homonymy inhibition.
always chosen. \({ }^{3}\)
\[
\begin{equation*}
P(s \mid m)=\frac{P(m \mid s)}{\sum_{i=1}^{S} P(m \mid i)} \tag{2}
\end{equation*}
\]

The simulations showed that, for both WTA and stochastic production, populations would only converge on optimal signaling either in combination with continuous replacement of old agents (iterated learning), or when agents had a fixed memory capacity in static populations.
4. Smith's (2002) network model contained a total of 81 possible 'update rules' determining how learning affects internal representations. The exemplar framework rendered most of these counter-intuitive, leaving only two parameters: whether adding a new exemplar would result in lateral inhibition of competing synonyms and/or homonyms (or neither). The replication confirmed Smith's analysis: inhibition of homonyms alone results in the extermination of both homonymy and synonymy. The reverse is not true, however: inhibiting synonyms does not affect homonymy. Moreover, the time taken to converge when homonymy inhibition is employed is apparently unaffected by the presence of an anti-synonymy bias, or whether WTA or stochastic strategies were used, as shown in Figure 3. With the correct bias in place, however, observational learners proved able to construct optimal signaling in both static and iterated learning populations.

When the four main models are compared using only horizontal transmission in a static population as in Figure 4,

\footnotetext{
\({ }^{3}\) The inverse process, obverter reception, is also possible, but simulations indicate that this does not lead to optimal signaling.
}


Figure 4: (Model Comparison) The proportion of 10,000 closed-group simulations which had converged to an optimal communication system after a given number of iterations, comparing stochastic implementations of observational learning (Smith, 2002), hearer-only feedback (Steels \& Loetzsch, 2012), negative reinforcement (Barrett, 2006), and obverters limited to a 50 -exemplar memory (Oliphant \& Batali, 1997).
the convergence time for the hearer-only feedback and observational models appear to have identical distributions, and memory-limited obverters perform similarly as well. Negative reinforcement models take a significantly longer time to converge. As such, the requirements for each model to converge appear to be:
1. Reinforcement learning: negative reinforcement or memory limitations
2. Corrective feedback models: either no possibility of homonymy, or inhibition of homonyms.
3. Obverter learning: either vertical learning or limited memory
4. Observational learning: inhibition of homonyms is required

\section*{Comments}

Based on our comparative simulations, the following conclusions can be drawn:
1. Simple reinforcement on the basis of successful communication is an ineffective way of establishing conventional signaling systems, leading to either non-convergence or very long convergence times in comparison to the other models. However, a much faster convergence is ensured if any form of deletion from memory is implemented, the most effective one being targeted negative reinforcement.
2. Corrective feedback as instantiated in the Steels models includes very large name or signal spaces. As a result,
homonymy is either impossible or unlikely. Communicative success in this case is unsurprising: even if every agent innovates their own signal for each meaning, eventually all agents throughout the population will have heard this token and will be able to correctly interpret it. This results in highly redundant labeling systems. Inhibiting synonyms leads to the eventual adoption of one-to-one mappings throughout a population. When the available signal space is limited, however, homonymy becomes a problem. Without the lateral inhibition of homonyms, convergence is not a certainty.
3. Smith's (2002) models and the simplified Steels \& Loetzsch (2012) models have extremely similar behavior because on one level of analysis they are the same: while Smith's observational learning ignores referential uncertainty, that uncertainty actually plays no role in the feedback model. With corrective feedback, the intended referent is either correctly understood or else communicated after failure. The speaker's intended communication is known independently of communicative success in both models.
4. 'Feedback' has several interpretations. Corrective feedback is described in Steels \& Loetzsch (2012): the speaker indicates its intended interpretation. Reinforcement learning involves another form of feedback, where the speaker (or the environment) simply confirms whether or not the hearer has correctly understood. In Baronchelli (2011) and Vogt \& Coumans (2003), feedback is defined as when the hearer informs the speaker how it has interpreted the signal.

We propose that the different kinds of "feedback" might be better characterized by looking at how information flows between speaker and listener. Corrective feedback in naming games ensures that the speaker always provides complete information about how it associates a particular meaning with a signal by unambiguously providing both the signal and the intended referent in every interaction. This guaranteed transmission of information is a feature shared by the observational models presented above. In reinforcement models, that information is only transmitted to a hearer after correct interpretation. Information flow from the hearer back to the speaker, on the other hand, is not present in the observational models which exhibit purely vertical transmission. Baronchelli (2011) shows that this flow is in fact unnecessary for the naming game without homonymy; the replications of the previous section show that this is also the case with homonymy (see Figure 2).
Feedback from hearer to speaker is critical for reinforcement learning, as confirmation of communicative success requires this information. The lack of ambiguity in other models ensures success, and thus removes the need for knowledge about communicative success. The flow of information from speaker to hearer is common to all the
above models. The role of any relevant feedback, then, is to allow this information to pass at least some of the time.
5. Basic reinforcement models utilize only the general positive feedback provided after successful communication. Negative reinforcement goes one step further by using information available after failed communication to determine what the likely internal state of the speaker is not, and this difference in information is sufficient to lead to ideal signaling. However, the reliably transmitted information in other models is not by itself enough to guarantee optimality. Some force must lead to competition between homonyms. For observational models and in the naming game, this is lateral inhibition through deletion. For obverters, it is implicit in the way production is biased towards the most successful homonym.
6. Functional communication arises when signals unequivocally map to single meanings. Models which do not actually delete competing homonyms, such as basic reinforcement and obverters, must employ some form of nontargeted deletion. These effects arise through either vertical learning (by wiping out parts of the 'collective memory' through the ongoing replacement of agents) or memory constraints on individual agents. Vertical learning leads to a process analogous to genetic drift: there is a chance that with every new generation some tokens will not be learned and thus lost, reducing the diversity of signals for any given meaning. Equally, limiting individual agents' memory capacity by deleting surplus exemplars causes the relative proportions of competing tokens to be affected by a random walk. In both cases, however, the probability of a particular mapping undergoing total deletion is inversely proportional to its relative frequency. If the pressures exerted by basic reinforcement models or obverter production cause the majority of mappings to gravitate towards an optimal system, then random sampling is enough to remove all competitors and lead to one-to-one mappings.

What, then, are the crucial elements which determine whether a population will construct optimal signaling? The next section will discuss the underlying qualities shared by all models with this property.

\section*{Discussion}

Reliable transmission of information between agents is not by itself enough to lead to the emergence of an optimal signaling system: there must be competition between homonyms, leading to a situation where each signal maps unambiguously to a single meaning. The opposite directionality of simultaneously strengthening signals in one meaning-space while decrementing them in another is a self-reinforcing, rich-getricher process. Models which use lateral inhibition reliably attain a stable, unambiguous state. Without lateral inhibition however, such as in basic reinforcement and closed-group obverter models, this does not happen. While both processes contain an implicit bias against homonymy, without some
form of deletion this is not strong enough, leading to ambiguous states which are semi-stable. In the absence of deletion, the weight of stored exemplars serves both to preserve ambiguous mappings and inhibit moves towards optimality. Deletion can be either active, such as in negative reinforcement, or it can arise through passive processes of random memory deletion or intergenerational sampling.
The factors, then, which determine whether a population will reliably construct optimal signaling are:
1. Speakers have to convey information - at least some of the time - about how they associate signals and meanings.
2. Information associating a signal to a meaning must bias the receiver against associations with other meanings.
3. Information must be lost: this may be via deletion, forgetting or intergenerational sampling.

In reinforcement learning, information rewards communicative success and optionally punishes failure. The information provides an inherent bias against homonymy. Similarly, the same bias is packaged into obverter production, which maximizes the chance of successful comprehension. In observational and feedback models on the other hand, the lateral inhibition of homonyms encapsulates both the bias and the deletion.

\section*{Conclusion}

Self-organization of learned communication systems results from both individual and population-level behavior as well as their interactions. This generality explains the seemingly opposed interpretations and conclusions seen in modeling approaches: the relevant factors that guarantee convergence can be implemented in many ways. In fact, all of the proposed models may be partially accountable for the emergence of shared communication systems in humans. This has implications for both modeling and experimental approaches. When a certain set of conditions leads to a system of agreed signaling conventions, those conditions cannot be assumed to be the sole cause of the phenomenon. Instead, the conditions may simply fulfill the necessary requirements outlined above.

\section*{References}

Baronchelli, A. (2010). The minimal naming game: A complex systems approach. , 1-24. (Book Chapter, Luc Steels, due out 2012.)
Baronchelli, A. (2011, April). Role of feedback and broadcasting in the naming game. Physical Review E, 83(4), 016113.

Baronchelli, A., Felici, M., Loreto, V., Caglioti, E., \& Steels, L. (2006, June). Sharp transition towards shared vocabularies in multi-agent systems. Journal of Statistical Mechanics: Theory and Experiment, P06014-P06014.
Barrett, J. (2006). Numerical simulations of the Lewis signaling game: Learning strategies, pooling equilibria, and
the evolution of grammar. Working Paper MBSO6-09 2006 Irvine, UK:University of California.
Beggs, A. (2005, May). On the convergence of reinforcement learning. Journal of Economic Theory, 122(1), 1-36.
De Vylder, B., \& Tuyls, K. (2006, October). How to reach linguistic consensus: a proof of convergence for the naming game. Journal of theoretical biology, 242(4), 818-31.
Kirby, S., Cornish, H., \& Smith, K. (2008, August). Cumulative cultural evolution in the laboratory: an experimental approach to the origins of structure in human language. Proceedings of the National Academy of Sciences of the United States of America, 105(31), 10681-6.
Lewis, D. (1969). Convention: A Philosophical Study. WileyBlackwell.
Oliphant, M. (1996, January). The dilemma of Saussurean communication. Bio Systems, 37(1-2), 31-8.
Oliphant, M., \& Batali, J. (1997). Learning and the emergence of coordinated communication. Center for research on language newsletter, 11, 1-46.
Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee \& P. Hopper (Eds.), Frequency effects and the emergence of linguistic structure (pp. 137-157). Amsterdam: John Benjamins Publishing Co.
Roth, A., \& Erev, I. (1995). Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. Games and economic behavior, 212, 164-212.
Scott-Phillips, T. C., Blythe, R. A., Gardner, A., \& West, S. A. (2012, May). How do communication systems emerge? Proceedings. Biological sciences / The Royal Society, 279(1735), 1943-9.
Smith, K. (2002). The cultural evolution of communication in a population. Connection Science, 14(1), 65-84.
Steels, L. (1998). The origins of syntax in visually grounded robotic agents. Artificial Intelligence, 103(1-2), 133-156.
Steels, L., \& Loetzsch, M. (2012). The Grounded Naming Game. In L. Steels (Ed.), Experiments in cultural language evolution (pp. 41-59). Amsterdam: John Benjamins Publishing Co.
Vogt, P., \& Coumans, H. (2003). Investigating social interaction strategies for bootstrapping lexicon development. Journal of Artificial Societies and Social Simulation, 6, no.l.

\title{
How Do Readers Explain the Occurrence of Conflicts in Science Texts? Effects of Presentation Format and Source Expertise
}

\author{
Marc Stadtler (marc.stadtler@uni-muenster.de) \\ Lisa Scharrer (lisa.scharrer@uni-muenster.de) \\ Rainer Bromme (bromme@uni-muenster.de) \\ Institute for Psychology, University of Münster, Fliednerstrasse 21, 48149 Münster, Germany
}

\begin{abstract}
The present study set out to investigate the influence of two metatextual features-presentation format and source expertiseon lay readers' explanation of conflicts in scientific information. Secondary school students read partly conflicting information about a medical topic, which was either presented in one single document or in four different documents, and which was purportedly authored by lay or expert sources. Results show that readers deemed deficits in source expertise (source explanations) more likely to account for conflicts in information written by lay authors than for conflicts reported by experts. In addition, conflicts presented by experts and conflicts in multiple documents were explained more strongly by referring to the nature of knowledge and knowledge production (epistemic explanations). Our findings demonstrate that readers are sensitive to situational variations when considering the most likely explanations for scientific conflicts. Implications for readers' adequate understanding and subjective resolution of scientific controversies are discussed.
\end{abstract}

Keywords: multiple document comprehension; science understanding; folk philosophy/sociology of science.

\section*{Introduction}

Generating explanations is key to comprehending scientific texts, be it in school or in settings of informal learning (Otero \& Graesser, 2001). Explanations help readers to understand why phenomena mentioned in a text occur and how they relate to one another. Furthermore, readers may adapt their further text processing depending on whether they manage to generate satisfactory explanations and thus develop a coherent mental model of the described situation. Given the important role of explanations in lay readers' handling of scientific texts, the present study sheds light on factors that influence readers' generation of explanations, specifically, their explanations for conflict in science texts.
In generating explanations, readers draw on their folk science, that is, their own fragmentary understanding of the ontological world (Keil, 2010). Graesser and Bertus (1998), for example, demonstrated that science text readers use their prior knowledge to produce an especially high number of inferences about the causal antecedents of an event. Moreover, Costa, Caldeira, Gallástegui, and Otero (2000) report that secondary-school students reading science texts asked a high number of questions of which the vast majority
pertained to causal explanations for the described phenomena.

An especially important catalyst for reader-generated explanations is the occurrence of conflicts in text (Otero \& Graesser, 2001). Clashes of knowledge claims potentially stand in the way of attaining unambiguous knowledge about the world and thus call for the reader's attention. That said, developing an explanation for why two authors disagree potentially helps readers to restore coherence and eventually take a personal stance on a controversy.

However, conflicting stances on a scientific issue cannot be explained by a reader's folk science alone. Instead readers may draw on their assumptions about how knowledge in the given discipline is structured and produced. In addition, readers may draw on assumptions about how knowledge is distributed between individuals and how knowledge communication is tied to individuals' personal interests. Following Keil's notion of folk science, one might term the former assumptions as belonging to an individual's folk philosophy of science \({ }^{1}\) whereas the latter assumptions belong to a folk sociology of science.

Bromme, Thomm, and Wolf (2013) report an interview study demonstrating how laypersons spontaneously generate a rich set of explanations drawing on these assumptions. Based on a sample controversy on the causation of a medical condition, laypersons (undergraduates from nonmedical subjects) and intermediates (advanced medical students) were asked how they would generally explain the occurrence of conflicts in medical knowledge. Participants provided a rich variety of possible explanations. These fell into two major categories, of which the first one related to the nature of knowledge and knowledge production. The category reflects the structural complexity of scientific knowledge and the discursive nature of knowledge production with differences in methodology or research questions leading to incompatible research results. The highest number of explanations provided by laypersons and intermediates fell into this category. A second set of

\footnotetext{
\({ }^{1}\) We refer to laypeople's assumptions about scientific knowledge and knowledge production as folk philosophy of science in reference to Keil's (2010) terminology. In other approaches, such assumptions are conceptualized as epistemological beliefs or nature of science beliefs. While these conceptualizations originate from different research traditions, they nevertheless overlap in terms of their reference to individual's beliefs about what scientific knowledge is and how it is justified.
}
explanations provided by both groups of participants focused on the source of information as the reason for conflicts. Participants explained conflicts with differences in the training and expertise of sources, or, to a lesser extent, with differences in the sources' interest or motivation. The study by Bromme et al. may be taken as tentative evidence that laypersons, at least those with a higher educational background, successfully draw on their folk philosophy and folk sociology of science to explain the occurrence of conflicts. What is currently essentially lacking, however, are empirical insights into the situational factors that determine which type of explanation readers prefer. Insights into these mechanisms could possibly inform research on public understanding of science and text comprehension.

\section*{The present study}

With our present study we set out to research whether readers' preferred conflict explanations depend on metatextual information about the sources providing the information and its presentation format. Our approach was informed by a recent study that examined how metatextual information impacts lay readers' understanding of a controversial scientific issue (Stadtler, Scharrer, Brummernhenrich, \& Bromme, 2013). Stadtler et al. presented participants with partly conflicting information on a medical issue. Presentation format and source expertise were systematically varied. The information was either presented in one single document by a single author or spread across four documents presented by different authors. In addition, the information was either purportedly written by expert sources (medical doctors) or lay sources (high school students). Results revealed that readers of multiple documents exhibited better memory for conflicts and were more likely to acknowledge the controversial nature of information in a subsequent knowledge communication task. How readers reported conflicting information also depended on source expertise. The variation of presentation format only mattered for expert information, which readers deemed worthy of effortful processing. The study by Stadtler et al. thus demonstrates that readers are sensitive to metatextual information in terms of memory for and use of conflicting scientific information. However, it deserves further clarification whether variations in presentation format and source expertise also effect on readers' preferred explanations for the occurrence of scientific conflicts.

The conceptual link behind this assumption is that metatextual information may differentially activate readers' tacit assumptions about how knowledge is structured and produced (i.e. their folk philosophy of science) and how it is distributed between individuals (i.e. their folk sociology of science). Presenting science information in multiple documents, for instance, may particularly highlight its complexity and the discursive nature of knowledge production (Britt \& Aglinskas, 2002; Wiley \& Voss, 1999). As a result readers of multiple documents may prefer epistemic explanations for the occurrence of conflicts. A
presentation of the same information in a single document, in contrast, may rather downplay the discursive nature of scientific knowledge production stimulating readers to a lesser degree to forward conflict explanations of an epistemic kind.

The degree to which readers prefer epistemic explanations may also depend on variations of source information. Conflicts in expert information should be regarded as particularly representative of the underlying scientific discipline, thus stimulating epistemic conflict explanations. In contrast, when scientific conflicts are presented by lay authors, this may not activate readers' folk philosophy of science to the same degree resulting in fewer epistemic explanations.

Both metatextual factors are also likely to influence readers' preference for source explanations as the reason for conflicting information. Since readers should consider laypeople more prone to mistakes than experts, information authored by lay sources should more strongly stimulate readers to explain conflicts with deficits of source expertise. Moreover, readers might interpret unresolved conflicts presented by a single author in a single text as indicative of the author's lack of understanding of the subject matter. As a result deficits in source expertise should appear more appropriate for explaining conflicts if contradictions occur within a single text rather than between multiple documents.

In spite of this reasoning, whether or not lay readers' explanation of conflicting science information is in fact determined by source expertise and presentation format is by no means a trivial question. So far, it is not clear whether readers are at all sensitive to situational factors when explaining encountered conflicts, or whether they have preconceived ideas of which conflict explanation is most relevant, irrespective of situational variations. Moreover, it is unclear whether lay readers use the provided author and document information (student vs. doctor; single vs. multiple documents) to draw conclusions about the most appropriate conflict explanations. Readers' reliance on source expertise is particularly uncertain in light of previous findings showing a notorious lack of spontaneous attention to source information (e.g., Bråten, Strømsø \& Salmerón, 2011; Britt \& Aglinskas, 2002; Kammerer, Gerjets, \& Werner, 2011; Wineburg, 1991).

Assuming that lay readers are sensitive to metatextual information when determining the most likely explanation for encountered conflicts, we formulate the following hypotheses:

Epistemic explanations: We expected an epistemic explanations to be deemed more likely by those reading multiple documents compared to participants encountering a a single document (H1a). Similarly, we hypothesized that readers consider epistemic explanations more appropriate for conflicts encountered in expert texts compared to conflicts in lay texts (H1b).

Source explanations: Furthermore, we expected that readers deem source explanations to better account for conflicts encountered in lay texts than conflicts encountered
in expert texts (H2a). In addition, reading a conflict in a single document should be explained more strongly with a lacking ability of the author than a conflict that exists between different sources (H2b).
It should be noted that readers might also explain perceived contradictions with their own lacking competence to correctly understand the provided information. In this case, they would blame themselves for the inability to form a coherent mental representation rather than interpreting the perceived inconsistency as an objective conflict. Our focus was on situations in which readers can be rather certain of the objective existence of conflicts, and we therefore did not expect any impact of metatextual factors on such selfrelated explanations. However, it is important to account for the possibility of self-related explanations when investigating readers' conflict explanations, particularly when focusing on laypeople confronted with expert information.

\section*{Method}

\section*{Participants, design, and task}

A total of 244 German secondary school students were recruited randomly during an open day at a German university. Students participated voluntarily and without payment. Participants were randomly assigned to one of four experimental groups following a 2 (text presentation format: single document vs. multiple documents) \(\times 2\) (source expertise: high vs. low) factorial design. Participants worked on a scenario developed in previous research (Stadtler \& Bromme, 2008) in which a fictitious friend, who has been diagnosed as having a high cholesterol level, is having to decide whether to take action to lower it. Participants were asked to support an informed decision by reading conflicting texts about the topic cholesterol. After reading participants provided explanations for the conflict they read. Forty-two students ( \(18 \%\) ) were not analyzed further because they failed to identify the conflict in the reading task. The data of another four students were dropped from analyses because they judged their medical knowledge to be good or very good on a five-point Likert scale ranging from very poor to very good. All other students provided lower self-assessments of their medical knowledge and therefore can be regarded as laypersons with regards to medicine. Hence, our final sample contained 194 participants ( \(85 \%\) female, mean age \(=17.76 ; S D=.92\) ).

\section*{Materials}

The materials used in this study consisted of two controversial medical issues on the topic of cholesterol that were described in four text passages. Each controversial issue consisted of two opposing standpoints. Each standpoint was mentioned in only one text passage. Claims were not marked as conflicting; hence conflicts had to be inferred by the reader. For instance, one text passage stated that a diet with low-cholesterol products is an effective means to lower one's level of cholesterol whereas another
text passage contained contradictory information. The second conflict addressed the threshold level of blood cholesterol beyond which there is a high risk of arteriosclerosis. Whereas one text passage argued for a universal threshold value of \(200 \mathrm{mg} / \mathrm{dl}\), another text passage claimed that the threshold value for cholesterol varies individually. The amount of filler information was minimalized to ensure that readers succeeded in recognizing the textual conflicts. The whole text information comprised 202 words in the case of the nutrition conflict and 227 words in the case of the threshold value conflict, respectively. Note that participants read only one conflict, i.e., either the text passages presenting the nutrition conflict or the text passages presenting threshold value conflict. This was done to unambiguously link readers' conflict explanations to a specific conflict. Text passages were displayed on a computer screen and depending on the experimental condition were presented either as two separate web sites by two different authors (in the multiple documents condition) or as one web site by a single author (in the single document condition). In addition, source information was varied by introducing the information as stemming from one or two medical doctors (in the expert source condition) or from one or two high school students (in the lay source condition). To control for contingency effects (Mayer, 2005), participants reading a single document worked with a similar navigation structure to those reading multiple documents. Participants could access the four text nodes via a table of contents linking to the nodes within the web site.

\section*{Dependent variables}

Explanation of textual conflicts Participants who had indicated that they noticed the conflict in the text materials were then asked to rate "to what degree do the following statements explain the occurrence of the conflict you have just found?" on 6-point Likert scales. Statements were constructed to measure preference of epistemic explanations ( 5 items relating to the nature of knowledge and knowledge production), source explanations (5 items relating to the expertise and motivation of sources), and self-related explanations ( 5 items relating to one's own ability to comprehend the conflicting information). For example, the statement "There are no clear answers to many medical questions," was intended to measure epistemic explanations; the statement "The author made a mistake," to measure source explanations; and "I don't have enough topic knowledge to solve the conflict, but an expert could" selfrelated explanations. Psychometric properties of the inventory are reported in the results section along with the empirical examination of the factorial structure.
Procedure After first providing information on demographic variables and assessing their own medical knowledge, participants were instructed to read the text materials and take notes of any conflict they encounter. The instructions for this reading task gave participants a definition and example of contradictory information. Only

Table 1: Mean ratings of explanations (standard deviations in parentheses) as a function of presentation format and source expertise.
\begin{tabular}{lcccc}
\hline & \multicolumn{2}{c}{ Single Document } & \multicolumn{2}{c}{ Multiple Documents } \\
\cline { 2 - 5 } Type of explanation & Lay source & Expert source & Lay source & Expert source \\
\hline & \(n=43\) & \(n=50\) & \(n=48\) & \(n=53\) \\
\hline Epistemic explanation & \(2.68(.96)\) & \(2.81(1.14)\) & \(3.01(1.20)\) & \(3.64(1.16)\) \\
Source explanation & \(3.89(.93)\) & \(2.97(.83)\) & \(3.81(.97)\) & \(3.37(.92)\) \\
Self-related explanation & \(1.95(.80)\) & \(2.21(.86)\) & \(2.10(1.03)\) & \(2.11(.82)\) \\
\hline
\end{tabular}
when participants noticed textual conflicts had they to report their explanations regarding these contradictions on the rating scale. Finally, participants were debriefed. The whole session lasted an average of 20 min .

\section*{Results}

Because similar patterns of results were obtained for the two conflict topics (nutrition conflict and threshold value conflict), all analyses reported hereafter were conducted on data that were aggregated across conflict topics.

\section*{Explanation dimensions}

To validate the factorial structure of the explanation inventory on empirical grounds, we subjected the 15 items to an exploratory factor analysis (ML-extraction, oblimin rotation). Three different analyses were run, in which we requested a forced two-, three, and four factor solution, respectively. An inspection of the screeplot confirmed that the best solution was the expected three-factor structure \(\left(\right.\) KMO \(=.78 ;\) Bartlett's test \(\chi^{2}(105)=900.97, p<.001\); share of explained variance \(=42.20 \%\) ). This solution also revealed the lowest number of double loadings and hence offered a maximum of conceptual clarity. All items had their highest loading on the factor they were intended to contribute to; hence, the theoretically motivated factor labels (epistemic, source, and self-related explanation) were retained. Internal consistencies for the explanation dimension ranged from good (epistemic explanations: Cronbach's alpha \(=.87\) ) to acceptable (self-related explanations: Cronbach's alpha \(=.74\); source explanations: Cronbach's alpha = .62).

\section*{Influence of conflict type and source expertise on attribution}

Table 1 reports mean ratings of conflict explanations and standard deviations as a function of presentation format and source expertise. To test the assumption that presentation format and source expertise influenced conflict explanation, we computed a mixed ANOVA with type of explanation as within-subject factor and presentation format and source expertise as between-subject factors. Results showed a strong effect of type of explanation, \(F(1.84,349.5)=\) 103.18, \(p<.001\), part. \(\eta^{2}=.352\). An inspection of means indicates that this effect was due to readers deeming source explanations to best account for conflicts ( \(M=3.40 ; S D=\)
.99), followed by epistemic explanations ( \(M=3.06 ; S D=\) 1.18) and finally by self-related explanations ( \(M=2.10 ; S D\) \(=.88\) ). However, this effect was qualified by the two possible two-way interactions (the three-way interaction, in contrast, did not reach significance). Firstly, there was an interaction between presentation format and type of explanation, \(F(1.84,349.5)=4.15, p=.019\), part. \(\eta^{2}=.021\). Moreover, the interaction between source expertise and type of explanation was significant, \(F(1.84,349.5)=15.27, p<\) .001 , part. \(\eta^{2}=.074\). To further examine the nature of these interactions, separate univariate follow-up analyses for each type of explanation were conducted.

Epistemic explanations: In line with our first hypothesis (H1a), we obtained a main effect of presentation format on epistemic explanations. Those who read multiple documents explained the occurrence of conflict to a greater degree with the nature of knowledge and knowledge production compared to participants who encountered a conflict in a single document, \(F(1,190)=12.72, p<.001\), part. \(\eta^{2}=\) .063. In addition, and in line with H1b, epistemic explanations were considered more appropriate for conflicts encountered in expert texts compared to conflicts in lay texts, \(F(1,190)=5.42, p=.021\), part. \(\eta^{2}=.028\). The interaction between presentation format and source expertise was not significant, \(F(1,190)=2.44, p=.120, n s\).

Source explanations: In addition, participants considered source explanations to better account for conflicts in lay texts than for conflicts encountered in expert texts, lending support to \(\mathrm{H} 2 \mathrm{a}, F(1,190)=26.96, p<.001\), part. \(\eta^{2}=.124\). Moreover, the interaction between presentation format and source expertise reached marginal significance, \(F(1,190)=\) \(3.25, p=.073\), part. \(\eta^{2}=.017\). This interaction was due to readers of a single expert source deeming source explanations less appropriate than readers in any other condition. Different from what has been expected in H2b, reading a conflict in a single document was not explained more strongly with reference to the author than were conflicts between different sources, \(F(1,190)=1.42, p=\) .235, ns.

Self-related explanations: Finally and as expected, no effects of our manipulation were observed with regards to self-attributions (all \(F \mathrm{~s}(1,190)<1.11, n s)\).

\section*{Discussion}

Previous research on text comprehension has shown that readers have a strong tendency to formulate explanations when reading conflicting scientific materials (e.g., Graesser \& Bertus, 1998; Millis \& Graesser, 1994; Otero \& Graesser, 2001). The aim of this study was to add to the literature by examining which explanations lay readers deem acceptable for the occurrence of conflicts. Readers' preferences were examined as a function of presentation format (multiple vs. single document) and source expertise (expert vs. layperson). It was argued that lay readers possess naïve theories relating to the nature of knowledge and knowledge production (folk philosophy of science) and to the distribution of knowledge between individuals and their motives in communicating scientific knowledge (folk sociology of science). Metatextual information on presentation format and source expertise should differentially activate these theories and result in corresponding explanations that either focus on the epistemic nature of knowledge or on the source of information. The results widely support our expectations.
The strongest effect we obtained was the one of varying source expertise on source explanations. Source explanations were considered more appropriate for conflicts in information that was purportedly written by high-school students than the same conflict being purportedly produced by experts. A marginally significant interaction between presentation format and source expertise revealed that this effect was slightly more pronounced for conflicts that were included in a single document. Readers of a single expert document were obviously particularly hesitant to blame the expert for the occurrence of discrepant information and in turn chose other explanations to a similar degree. In contrast to our expectations, we did not observe that readers of a single document explained the occurrence of a conflict more strongly with a mistake of the author compared to those reading multiple documents. Note that in line with our expectations, conflicts in single documents were indeed predominantly explained with reference to the source. However, this was also true for those reading multiple documents, which we had not expected. Although unexpected, this result is in line with the results of Bromme, Thomm and Wolf (2013). In their study, laypersons regarded the source as a central cause of conflicts in science. This heuristic may be so salient in laypersons that it is applied regardless of presentation format when explaining conflicts.
Our results regarding epistemic explanations provide full support for our hypotheses. Conflicts presented by experts and conflicts between documents were explained more strongly by referring to the nature of knowledge and knowledge production. As for the expert-lay author variation, it may be argued that knowledge claims presented by experts are conceived as more indicative of the underlying scientific discipline in terms of the certainty it provides. Moreover, conflicts between different sources
may be seen as directly pointing to the discursive nature of scientific knowledge production. As suggested by the results of the interviews conducted by Bromme et al. (2013), young adults show awareness that conflicts among both medical scientists and medical practitioners are commonplace. This insight reflects some epistemic sophistication and it may help readers to find adequate explanations when they encounter conflicts in science texts.

It is notable that we found sensitivity for metatextual information among high-school students. This result extends previous research (Stadtler et al., 2013) which has demonstrated sensitivity for metatextual information in terms of memory for and use of conflicting information among university undergraduates. Our results suggest that at least advanced high-school students seem to possess cognitive resources that enable them to assess the appropriateness of different conflict explanations without a great amount of elaboration.

Finally, it may be seen as a limitation of our study that our results are based on presenting high-school students with predefined explanations. Thus, future research will have to show whether a similar pattern of explanation preferences will be obtained when laypersons have to generate conflict explanations from scratch. With this goal in mind, interview studies, such as the one by Bromme et al. (2013), could be conducted with younger and less educated populations.

Another important topic for future research will be to assess the implications of different conflict explanations for the processing of science texts. A triangulation of data gained with think-aloud or eye-tracking procedures might be especially helpful to examine whether readers translate their subjective conflict explanations into actual reading behaviors. This could include an intensified elaboration of source information if conflicts are primarily explained with deficits in source competence, or corroborating information between sources (Wineburg, 1991). Readers' conflict explanations may also influence their inclination to engage in further information search. For example, explaining a conflict with lacking competence of the author(s) may prompt readers to obtain additional topic information from more reliable sources. In contrast, readers who explain a conflict with the nature of knowledge may refrain from looking up any further information, because they do not consider the encountered conflict as indicative of lacking information quality. It will also be an important task to examine whether readers use their preferred conflict explanation to develop a personal stance towards the conflict. Especially when readers explain a conflict with a lack of author competence, they might also use this explanation to decide which of the opposing stances they should include in their referential representation of the world. This way, conflict explanations might serve lay readers to harness scientific information for their goals of making informed decisions on everyday problems.

\section*{Acknowledgments}

This research was supported with a grant from the Deutsche Forschungsgemeinschaft (DFG).

\section*{References}

Bråten, I., Strømsø, H. I., \& Salmerón, L. (2011). Trust and mistrust when students read multiple information sources about climate change. Learning and Instruction Special Section I: Solving information-based problems: Evaluating sources and information, 21, 180-192.
Britt, M. A., \& Aglinskas, C. (2002). Improving student's ability to identify and use source information. Cognition and instruction, 20, 485-522.
Bromme, R., Thomm, E., \& Wolf, V. (2013). From understanding to deference: Laypersons' and medical students' views on conflicts within medicine. Manuscript submitted for publication.
Costa, J., Caldeira, M. H., Gallastegui, J. R., \& Otero, J. (2000). An analysis of question asking on scientific texts explaining natural phenomena. Journal of Research in Science Teaching, 37, 602-614.
Graesser, A. C., \& Bertus, E. L. (1998). The construction of causal inferences while reading expository texts on science and technology. Scientific Studies of Reading, 2, 247-269.
Gerjets, P., Kammerer, Y., \& Werner, B. (2011). Measuring spontaneous and instructed evaluation processes during web search: Integrating concurrent thinking-aloud protocols and eye-tracking data. Learning and Instruction, 21, 220-231.
Keil, F. C. (2010). The feasibility of folk science. Cognitive Science, 34, 826-862.
Mayer, R. E. (Ed.). (2005). The Cambridge handbook of multimedia learning. New York, NY: Cambridge University Press.
Millis, K. K., \& Graesser, A. C. (1994). The time-course of constructing knowledge-based inferences for scientific texts. Journal of Memory and Language, 33, 583-599.
Otero, J. C., \& Graesser, A. C. (2001). PREG: Elements of a model of question asking. Cognition and Instruction, 19, 143-175.
Stadtler, M., Scharrer, L., Brummernhenrich, B., \& Bromme, R. (2013). Dealing with uncertainty: Readers' memory for and use of conflicting information from science texts as function of presentation format and source expertise. Cognition and Instruction, 32, 130-150.
Wiley, J., \& Voss, J. F. (1999). Constructing arguments from multiple sources: Tasks that promote understanding and not just memory for text. Journal of Educational Psychology, 91, 301-311.
Wineburg, S. S. (1991). Historical problem solving: A study of the cognitive processes used in the evaluation of documentary and pictorial evidence. Journal of Educational Psychology, 83, 79.

\title{
Inductive Bias against Stem Changes as Perseveration: Experimental Evidence for an Articulatory Approach to Output-Output Faithfulness
}

\author{
Matthew Stave \\ Department of Linguistics, 1290 University of Oregon \\ Eugene, OR 97403 USA \\ Amy Smolek \\ Department of Linguistics, 1290 University of Oregon \\ Eugene, OR 97403 USA
}

\author{
Vsevolod Kapatsinski \\ Department of Linguistics, 1290 University of Oregon \\ Eugene, OR 97403 USA
}

\begin{abstract}
Speakers of morphologically-rich languages commonly face what has been called the Paradigm Cell Filling Problem: they know some form of a word but it is inappropriate to the current context, leading them to derive a form of that word they have never encountered (e.g., they know the singular form of a noun, and now need to produce the plural). We suggest that in performing this task speakers perseverate on articulatory gestures comprising the form they know, and that gestures vary in the extent to which speakers perseverate on them. This proposal explains the parallels between findings in loanword adaptation, speech errors, and acquisition of phonology. New experimental data from a miniature artificial language are presented in support of the theory.
\end{abstract}

Keywords: Phonology; morphology; speech production; inductive bias; faithfulness; Harmonic Grammar.

\section*{Theory}

In a seminal paper arguing for substantive bias in the acquisition of phonology, Wilson (2006) defines "substance" as "any aspect of grammar that has its basis in the physical properties of speech. These properties include articulatory inertias, aerodynamic pressures, and degrees of auditory salience and distinctiveness" (p. 946). In other words, substance in phonology is phonetics. The Substantive Bias Hypothesis suggests that the learner of phonology is predisposed towards acquiring patterns that are phonetically natural. Phonetically unnatural patterns are learnable and can therefore be productive in natural languages (Mielke, 2008; Ohala, 1978) but the learner needs more evidence to be convinced of their reality (Wilson, 2006).

A natural phonological alternation can be defined as an articulatorily and/or perceptually minimal change in a context where it can result from coarticulation, articulatory undershoot, and/or misperception. For instance, the velar stop [k] might become [tf] before [i] because the coarticulation between [k] and [i] causes [k] to front (becoming \(\left[\mathrm{k}^{\mathrm{j}}\right]\) ), resulting in \(\left[\mathrm{k}^{\mathrm{j}} \mathrm{i}\right]\), which is easy to
misperceive as \([\mathrm{tfi}]\) in noise (Guion, 1998). There are therefore multiple ways in which a phonological alternation can be phonetically unnatural.

First, it might happen in the "wrong" context. If it happens in the wrong context, it also might not happen in the "right" context, making it even more unnatural. For instance, palatalization might happen before [o] without happening before [i] (Kapatsinski, 2010) despite [ko] and [t 50 ] being acoustically and articulatorily quite distinct whereas [ \(\mathrm{k}^{\mathrm{j}}\) ] and \([\mathrm{t} \mathrm{fi}]\) are very similar. Context naturalness has been investigated experimentally by Mitrovic (2012), Schane et al. (1975), and Wilson (2006), among others.

Second, the change itself might be unnatural. For instance, Ohala (1978) shows that Southern Bantu changes [p] into [ t\(]\) ] without changing [ k\(]\) into \([\mathrm{t}]] .{ }^{1}\) The articulatory difference between \([\mathrm{p}]\) and \([\mathrm{t}]\) ] is articulatorily greater than the one between \([\mathrm{k}]\) and \([\mathrm{t}]\) ]. Nonetheless, \([\mathrm{p}]\) changes into [ t ] in Southern Bantu whereas [ k ] does not. The influence of change naturalness on learnability has only now begun to receive attention (Kapatsinski 2012b, White 2012). In demonstrating an effect of change naturalness, we provide additional evidence for the existence of substantive bias (contra Blevins, 2004; Hale \& Reiss, 2000; and Ohala, 1990 among others).

How can change naturalness influence learnability? We propose that it is through perseveration in speech production. Consider a speaker who knows one form of a word (say, a singular) and wants to come up with another form of the same word (say, a plural). We propose that in producing the unknown wordform the speaker is likely to perseverate on the articulatory units of the known wordform (Kapatsinski, 2013). This perseveration is usually functional, in that most, if not all, of the known form should be in the to-be-produced unknown form. This type of perseveration may help humans avoid bizarrely unfaithful

\footnotetext{
\({ }^{1}\) The Southern Bantu alternation context is also unnatural: palatalization happens before [w] rather than a front vowel (Ohala, 1978).
}
mappings like mail-membled, demonstrated by Rumelhart \& McClelland's (1986) model of English past tense formation, which lacks perseveration and is fully empowered to learn arbitrary present-past pairings. However, when the mapping between the known form and the to-be-produced form involves a stem change, perseveration on the to-be-produced form can result in error, where the stem change is leveled (or at least partially leveled). For instance, Kapatsinski (2009) showed that subjects who are exposed to a miniature artificial language with velar palatalization \(\left(\mathrm{k} \rightarrow \mathrm{t} \int\right.\) before the plural suffix -i) often make errors in which [k] becomes [ kt ]] rather than [tf] before -i (e.g., flook \(\rightarrow\) flooktfi rather than flouk \(\rightarrow\) flootfi), erroneously retaining the final consonant of the known form.

We propose that articulatory units differ in how much they are subject to perseveration. These differences in perseverance (susceptibility to perseveration) act as biases, causing learners to level stem changes that involve changing a unit that is highly susceptible to perseveration. \({ }^{2}\) Further, the greater the articulatory difference between the known form and the to-be-produced form, the more likely the change is to fail, or at least be carried out incompletely, since every one of the articulatory units present in the known form can (erroneously) persevere. Note that this is not a bias against all uncommon changes. Some changes may be uncommon for perceptual reasons, but we do not have any evidence to suggest that perceptual magnitude of a change influences its learnability (cf. Steriade, 2001 for a suggestion that it does). There is one published study that examined the relationship between change naturalness and learnability, and failed to find one: Wilson (2006) observed that \([k]-->[t f]\) and \([g]-->[d 3]\) did not differ in learnability despite [ki] and [ t fi ] being more perceptually confusable than [gi] and [dzi] (Guion, 1998). The present hypothesis is consistent with this finding: \([\mathrm{k}]\) and \([\mathrm{g}]\) differ only in voicing, which is not changed in palatalization and is articulatorily independent from the rest of the features of \([\mathrm{k}]\) and [g]; thus, the perseveration hypothesis predicts no learnability difference between \([\mathrm{k}] \rightarrow[\mathrm{t}]\) and \([\mathrm{g}] \rightarrow[\mathrm{d} 3]\).

Formally, the proposed bias is equivalent to a ranking of output-output faithfulness constraints (Kenstowicz, 1996) in Optimality Theory or unequal weighting in Harmonic Grammar (Smolensky \& Legendre, 2006). Namely, we will show that changing \([\mathrm{k}]\) or \([\mathrm{t}]\) into \([\mathrm{t}]\) ] is easier than changing [p]. In Optimality Theory / Harmonic Grammar, this could be described with a ranking: ("Keep [k]", "Keep [t]") << "Keep \([p]\) ". One way to model what subjects are learning in the experiment is an increased weight on a constraint saying that to-be-produced plural forms should end in [tfi]. As the weight of this constraint rises, it overtakes "Keep [k]" and

\footnotetext{
\({ }^{2}\) The differences in perseverance may be universal, rooted in early articulatory experience, or language-specific. Further, universal differences may be reduced or augmented by linguistic experience. We do not seek to determine the source of the observed bias in the present paper. Our aim is simply to demonstrate that biases of the proposed form exist.
}
"Keep [t]" before overtaking "Keep [p]". As a result, palatalization overgeneralizes from \([\mathrm{p}]\) to \([\mathrm{t}]\) and \([\mathrm{k}]\).

Faithfulness constraints are not a new idea in linguistics. What is new here is the claim that the faithfulness constraints in question are production-internal perseveratory tendencies specific to articulatory units. One appealing consequence of this proposal is that it provides a unified explanation for the bias to add rather than delete noted in both work on speech errors (Goldstein et al., 2007; Hartsuiker, 2002; Stemberger, 1990) and work on loanword adaptation in phonology (Kang, 2011; Paradis \& LaCharité, 1997). In both cases, articulatory units are clamoring for retention but nothing clamors for deletion, unless there are strong prosodic constraints limiting word or syllable shape (the only case when exceptions to the addition bias are found in loanword adaptation, according to Kang, 2011). The proposal that faithfulness constraints are rooted in perseveration explains this typological generalization. Making the relatively uncontroversial assumption of the existence of morphological units in speech production (e.g., Dell, 1986; Roelofs, 1997), the proposed perseveration can also straightforwardly capture the tendency for insertions to happen at morpheme or stem boundaries rather than morpheme- or stem-internally (Kenstowicz, 1994): perseverating on a morphological/lexical unit prevents changes inside that unit.

We propose that the bias against (certain) stem changes is internal to the production system. We therefore expect that it will manifest itself more strongly in production than in acceptability rating. This prediction is consistent with previous findings that a form with a stem change can be judged as being more acceptable than a form without a stem change and yet be less likely to be produced (Kapatsinski, 2012; Zuraw, 2000). Perseveration/self-priming is also commonly observed in elicited production tests of rule productivity, where participants have been observed to repeatedly reuse a phoneme sequence in novel words elicited on adjacent trials, even when the result violates normal paradigmatic mappings or affix order preferences (Bickel et al., 2007; Lobben, 1991).

Finally, the idea that output-output faithfulness is rooted in motor perseveration is consistent with data from language acquisition and motor development and provides an independent justification for the initial high ranking of such constraints proposed to be necessary to make OptimalityTheoretic grammars learnable (Hayes, 2004; McCarthy, 1998). Children are known to exhibit more motor perseveration than adults (e.g., Smith et al., 1999; including perseveration on phonetic segments: Stemberger, 1989; Vousden \& Maylor, 2006). It is therefore unsurprising that they would perseverate more on a known form while deriving an unknown morphologically-related form (e.g., Do, 2013), the pattern predicted by a high initial ranking of OO-Faithfulness ("keep the X from the base form") constraints. In fact, perseveration on inflectional morphemes recently produced (by children or their interlocutors), a perseveratory tendency present but greatly diminished in
adults, has been noted in the language acquisition literature where perseveration on interlocutors' wordforms has been argued to be functional in that the repeated morpheme is usually correct in the context of the child's own utterance (Ambridge \& Lieven 2011:164-65). Again, it is only when it is incorrect that we notice the perseveration; most of the time perseveration is correct and perseverating on the known form is a good strategy.

\section*{Predictions}

The present experiments focus on a particular phonological alternation, called palatalization, where a word-final [p], \([\mathrm{k}]\) or [ t ] alternates with [ t\(]\) ] when followed by [a] but is left unchanged before [i]. This is a pattern that runs counter to phonetic naturalness (Guion, 1998; Kochetov, 2011; Mitrovic, 2012; Wilson, 2006) but is attested in some languages, e.g., before Russian diminutives, where -ok palatalizes preceding [ k ] more than -ik does (Kapatsinski, 2010). Velar palatalization \((\mathrm{k} \rightarrow \mathrm{t} f)\) and alveolar palatalization ( \(\mathrm{t} \rightarrow \mathrm{t} \mathrm{f}\) ) are much more common than labial palatalization \((\mathrm{p} \rightarrow \mathrm{t} f)\) (Kochetov, 2011) and involve an articulatorily more minor change, since \([\mathrm{t}],[\mathrm{k}]\) and \([\mathrm{t} 5]\) are all lingual gestures, while [p] is a labial one. Velar and alveolar palatalization are attested approximately equally often in languages of the world (Kochetov, 2011) and [ t ] ] shares articulatory characteristics with both \([\mathrm{t}]\) and \([\mathrm{k}]\), involving both tongue tip (like [t]) and tongue body (like [k]) gestures (Yun, 2006). The present experiment thus seeks to determine whether there is a bias against the less natural alternation ( \(\mathrm{p} \rightarrow \mathrm{t} \mathrm{f})\) even in a context where no palatalizing alternation is particularly natural (before [a]).

An important, and counterintuitive, prediction of the theory is that the bias against changing a unit should be context-independent, to the extent that the unit in question is independent of the context in question in motor planning and execution. The articulatory unit addressed by the present experiments is the oral consonantal gesture. Work on speech errors suggests that consonants, and especially onsets, can move around in the motor plan independently from vowels (e.g., Fowler, 2010; Meyer, 1992, p. 185-86; Shattuck-Hufnagel, 1983). \({ }^{3}\)

Our previous work (Kapatsinski 2012b) has demonstrated that labial palatalization is more difficult to acquire than velar or alveolar palatalization in the context of a following [i], i.e., in a context where velar or alveolar palatalization are phonetically motivated. However, this result is consistent with a bias in favor of natural rules, i.e., changes in context. However, if [p] is harder to change into [ t\(]\) ] than \([\mathrm{k}]\) or \([\mathrm{t}]\) are independently of context, this should be true even if the vowel triggering palatalization is [a]. This prediction is tested in the experiment reported here.

\footnotetext{
\({ }^{3}\) There is some argument regarding whether the 'segmental' errors typically involve segments (Roelofs 1999, Stemberger 1982) or gestures (Goldstein et al. 2007, Mowrey \& MacKay 1990). The distinction is unimportant for the present purposes: the vocalic context is outside of both the segment and the gesture.
}

\section*{Methods}

The grammars presented to learners are shown in Table 1. There were three groups of participants. The Velar Group was presented with a language in which [k] became [ t ] before [a] while [p] and [t] remained unchanged, e.g., [bikbitfa, bit-bita, bip-bipa]. The Labial Group was presented with a language in which [p] became [tf] before [a] while [ k\(]\) and [ t\(]\) remained unchanged. The Alveolar Group was presented with a language in which [ t ] became [ t\(]\) ] [a] while \([\mathrm{k}]\) and \([\mathrm{p}]\) remained unchanged. In all languages, [i] and [a] were plural suffixes. In all languages, the palatalizing consonant was twice as common as any one of the nonpalatalizing ones.
\begin{tabular}{|c|c|c|}
\hline Velar Group: & Alveolar Group: & Labial Group: \\
\hline \(\mathrm{ik} \rightarrow \mathrm{itfa}\) & it \(\rightarrow\) it \(\int\) a & ip \(\rightarrow\) it \(\int\) a \\
\hline ak \(\rightarrow\) \{aki;atja & at \(\rightarrow\) \{ati;at a \} & ap \(\rightarrow\) api;at 5 a \} \\
\hline \(t \rightarrow\) ta & \(\mathrm{k} \rightarrow \mathrm{ka}\) & \(\mathrm{k} \rightarrow \mathrm{ka}\) \\
\hline \(\mathrm{p} \rightarrow \mathrm{pa}\) & \(\mathrm{p} \rightarrow \mathrm{pa}\) & \(t \rightarrow \mathrm{ta}\) \\
\hline
\end{tabular}

Languages were created from each other by swapping final consonants in the singulars, meaning the contexts surrounding the palatalized consonants were exactly parallel across the three groups within each experiment. The stimuli were then recorded by the first author and presented to participants, all adult native English speakers recruited from the Psychology/Linguistics human subjects pool, auditorily through headphones. There were 25 participants in the Alveolar Group, 29 in the Labial Group and 30 in the Velar Group. All words were paired with pictures of referents presented on the screen and presented in totally random order. Participants were asked to learn the names of the referents. A third of the way through training, they were tested on word learning by being asked to produce all wordforms they were presented with in the training when cued with pictures of the referents. At the end of training, the participants encountered an elicited production test, where they were presented with novel singular forms (which they had not encountered during training) and were asked to say the right plural form. The production test was followed by a rating test, in which the participants were presented with novel singular-plural pairs and were asked to press a button indicating whether the presented plural form was the right one for the presented singular. Statistical significance was evaluated using the original binary responses by means of logistic mixed effects models in the lme4 package in R (Bates et al., 2012) with random intercepts for subjects and items and random slopes for between-subject variables within Item and between-item variables within Subject (following Barr et al., in press). All analyses reported here were done on trials where -a was the suffix vowel chosen or presented.


Figure 1: Results of the elicited production test. Vertical axes show whether or not the consonant was palatalized (Y, shown in light, is "palatalized", N, shown in dark, is "not palatalized"). Participants exposed to labial palatalization before [a] palatalize \([\mathrm{t}]\) and \([\mathrm{k}]\) almost as much as they palatalize [ p\(]\) (though the difference is significant: \(\mathrm{z}=3.02, \mathrm{p}=.002\) for \([\mathrm{p}]\) vs. \([\mathrm{k}]\) rates and \(\mathrm{z}=2.85, \mathrm{p}=.004\) for \([\mathrm{p}]\) vs. \([\mathrm{t}]\) ). Participants exposed to alveolar palatalization palatalize [ t\(]\) much more than other stops ( \(\mathrm{z}=5.87, \mathrm{p}<.00001\) for \([\mathrm{t}]\) vs. \([\mathrm{k}] ; \mathrm{z}=10.57, \mathrm{p}<.00001\) for \([\mathrm{t}] \mathrm{vs}\). \([\mathrm{p}]\) ). [ p\(]\) is palatalized less than \([\mathrm{k}](\mathrm{z}=3.35, \mathrm{p}=.0008\) ).

Participants exposed to velar palatalization palatalize \([\mathrm{k}]\) more than \([\mathrm{p}](\mathrm{z}=10.00, \mathrm{p}<.00001)\) and more than \([\mathrm{t}](\mathrm{z}=5.16\), \(\mathrm{p}<.00001)\); \([\mathrm{t}]\) is palatalized more than \([\mathrm{p}](\mathrm{z}=8.55, \mathrm{p}<.00001)\).

Subjects trained on p->cha


Subjects trained on \(k->c h a\)


Stem-final Consonant ( X )

Subjects trained on t->cha


Stem-final Consonant ( X )

Figure 2: Results of acceptability judgments in Experiment 2. Dark parts of bars show the proportion of 'this is the wrong plural form for this singular' responses. Light parts of bars show the proportion of 'this is the right plural form for this singular' responses. Following training on labial palatalization, subjects accept palatalized labials \((\mathrm{p} \rightarrow \mathrm{t} f \mathrm{a})\), alveolars \((\mathrm{t} \rightarrow \mathrm{t} f \mathrm{a})\) and velars \((\mathrm{k} \rightarrow \mathrm{t} \mathrm{fa})\) at roughly equal rates whereas following training on alveolar or velar palatalization, the trained alternation is accepted more often than untrained ones. However, after all kinds of training subjects learn to reject unchanged/non-palatalized stops before the palatalizing vowel [a], and they do it at equal rates.

\section*{Results}

As shown in Figure 1, participants exposed to labial palatalization do not learn the pattern as well as participants exposed to velar or alveolar palatalization do, often learning to palatalize everything or to palatalize nothing. The differences in palatalization rates between the to-bepalatalized consonant and the not-to-be-palatalized consonants is significantly smaller in the group trained on \(\mathrm{p} \rightarrow \mathrm{t} \mathrm{f}\) than in the group trained on \(\mathrm{k} \rightarrow \mathrm{t} \int(\mathrm{z}=8.98, \mathrm{p}<.00001)\) or \(\mathrm{t} \rightarrow \mathrm{t}\) ( \(\mathrm{z}=3.34, \mathrm{p}=.0008)\); the latter two groups do not significantly differ ( \(\mathrm{z}=0.09, \mathrm{p}=.93\) ).

Figure 2 suggests that the same pattern holds for acceptability judgment data: subjects trained on \(\mathrm{p} \rightarrow \mathrm{t} \int \mathrm{a}\) judge \(\mathrm{p} \rightarrow \mathrm{t}\) fa examples ungrammatical almost as often as examples of \(t \rightarrow t \int a\) and \(k \rightarrow t \int a\) (for this group, there is no significant effect of singular-final consonant on acceptability of palatalization, \(\mathrm{z}=0.76, \mathrm{p}=.45\) ). By contrast, subjects trained on \(\mathrm{t} \rightarrow \mathrm{t}\) fa or \(\mathrm{k} \rightarrow \mathrm{t}\) a judge the alternations they were trained on as being grammatical more often than alternations they were not trained on ( \(\mathrm{z}=2.84, \mathrm{p}=.004\) ).

In addition, Figure 2 shows that the bias against labial palatalization (i.e., changing [p] into [tfa] rather than [pa]) is not due to a bias in favor of [pa]: subjects learn that [a] should not be preceded by [ p ] as easily as they learn that it should not be preceded by [k] ot [t]: while \(\mathrm{p} \rightarrow \mathrm{t} \mathrm{f}\) is worse than \(\mathrm{k} \rightarrow \mathrm{t} \oint\) and \(\mathrm{t} \rightarrow \mathrm{t}\), [pa] is as bad as [ka] and [ta] ( \(\mathrm{z}=.83\), \(\mathrm{p}=39\) ). Thus the observed bias against \(\mathrm{p} \rightarrow \mathrm{t} \int\) is not a bias in favor of [pa] or against [ka] and [ta]. \({ }^{4}\)

Finally, stem changes are accepted in acceptability judgment much more than they are produced: the dark bars in Figure 2 are lower than in the top panels of Figure 1 ( \(\mathrm{z}=3.07, \mathrm{p}=.002\) ), indicating that palatalization is usually rated as acceptable, and in fact more acceptable than nonpalatalization, yet is rarely produced. Furthermore, the bias against labial palatalization appears to be stronger in production than in judging acceptability: to-be-palatalized and not-to-be-palatalized consonants differ in acceptability of palatalization across subject groups numerically but not significantly ( \(\mathrm{z}=1.11, \mathrm{p}=.27\) ), but, as described above, these between-group differences are significant in production. The three-way interaction between test modality, whether or not a consonant is to be palatalized, and subject group is also significant ( \(\mathrm{z}=2.37, \mathrm{p}=.018\) ). These data provide direct

\footnotetext{
\({ }^{4}\) Some may wonder whether participants actually learn to palatalize before [a] rather than learning to palatalize before [i] or after [i]. There was no effect of stem vowel on acceptability ( \(\mathrm{z}=1.01, \mathrm{p}=.31\) ) of palatalization, whereas in production stem [i] disfavored palatalization \((\mathrm{z}=3.55, \mathrm{p}=.0003)\) rather than favoring it. The effect of final vowel on palatalization production probability ( \(\mathrm{z}=4.99, \mathrm{p}<.00001\) ) and acceptability \((\mathrm{z}=2.14, \mathrm{p}=.03)\) was in the direction predicted by training, rather than phonetic naturalness, i.e., palatalization occurred/was rated more acceptable than nonpalatalization before [a] more often than before [i]. Thus, the vowel triggering palatalization in the grammars learned by our subjects does appear to be a following [a], making the context for palatalization phonetically unnatural.
}
evidence for a production basis for faithfulness and the observed bias against \(\mathrm{p} \rightarrow \mathrm{t} \mathrm{f}\).

\section*{Discussion and Conclusion}

In this experiment, we have demonstrated that there is a bias against labial palatalization in a context where all kinds of palatalization are phonetically unmotivated. Thus we suggest that the bias is not in favor of phoneticallymotivated rules, or changes in context. We have also shown that the bias is not due to differentially ranked phonotactic constraints on the output forms: the result of alternation is always the same [ t f a ], regardless of the consonant that is changed, and learners acquire a dispreference against [pa], [ta], and [ka], the outputs competing with [ \(\mathrm{t} f \mathrm{a}\) ] equally well. We are therefore left with two options, both of which can be modeled using Faithfulness constraints in Harmonic Grammar (Smolensky \& Legendre, 2006): the learners might be biased against mapping [p] onto [ t\(]\) ] (Steriade, 2001), or against deleting the labial closure gesture associated with [p] (Kapatsinski, 2013; Kenstowicz 1996).

Theories of Faithfulness / avoidance of stem changes differ on whether Faithfulness is grounded in perception (avoiding changes that the listener would easily perceive and would face difficulty undoing to recover the base form of the stem, Steriade, 2001), articulation (perseverating on gestures of the known form while deriving an unknown form, Kapatsinski, 2013), or an offline preference for uniform morphological paradigms (e.g., storage economy or one-to-one form-meaning mappings, Kenstowicz's [1996] uniform exponence; McCarthy's [2005] optimal paradigms). The present data support the gestural account, as it alone seems to account for the differences between production and acceptability judgment. Namely, the biases against stem changes are stronger in production (see also Kapatsinski, 2012; Zuraw, 2000). The gestural account is also the only one that can account for the data described in the introduction, and it alone seems to be an inevitable component of the production process when a novel form is derived from a known one.

The bias we observe may or may not be specific to English speakers. While labials are unlikely to change across languages (e.g., Kochetov, 2011), English does have palatalization of alveolars, as in create/creature or, variably, in would/would you. The fact that we also observe a preference for velar palatalization over labial palatalization suggests that the preference against labial palatalization is not solely a first-language effect. Nonetheless, first language experience undoubtedly changes the weights of faithfulness constraints, hence cross-linguistic and developmental work on this issue would be most informative. Of particular interest here are languages that have labial palatalization, e.g., Southern Bantu (Ohala, 1978).

All we wish to claim at this point is that there are faithfulness constraints militating against stem changes, regardless of context, that they are production-based, and that they vary in weight, making some of these constraints stronger than others at the beginning of our experiment.

\section*{References}

Ambridge, B., \& Lieven, E. V. M. (2011). Child language acquisition: Contrasting theoretical approaches. Cambridge: Cambridge University Press.
Barr, D. J., Levy, R., Scheepers, C., \& Tily, H. J. (In press). Randomeffects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language.
Bates, D., Maechler, M., \& Bolker, B. (2012). Ime4: Linear mixedeffects models using S4 classes. R package version 0.999999-0
Bickel, B., G. Banjade, M. Haenszle, E. Lieven, et al. 2007. Free prefix ordering in Chintang. Language, 83, 43-73.
Blevins, J. (2004). Evolutionary Phonology: The emergence of sound patterns. Cambridge: Cambridge University Press.
Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. Psychological Review, 93, 283-321.
Do, Y. (2013). Children employ a conspiracy of repairs to achieve uniform paradigms. Ms. MIT.
Fowler, C. (2010). A. The reality of phonological forms: A reply to Port. Language Sciences, 32, 56-9.
Goldstein, L., Pouplier, M., Chen, L., Saltzmann, E., \& Byrd, D. (2007). Dynamic action units slip in speech production errors. Cognition, 103, 386-412.
Guion, S. (1998). The role of perception in the sound change of velar palatalization. Phonetica, 55, 18-52.
Hale, M., \& Reiss, C. A. (2000). 'Substance abuse' and 'dysfunctionalism': Current trends in phonology. Linguistic Inquiry, 31, 157-69.
Hartsuiker, R. J. (2002). The addition bias in Dutch and Spanish phonological speech errors: The role of structural context. Language \& Cognitive Processes, 17, 61-96.
Hayes, B. (2004). Phonological acquisition in Optimality Theory: the early stages. In R. Kager, W. Zonneveld, J. Pater (Eds.), Fixing priorities: Constraints in phonological acquisition. Cambridge: Cambridge University Press.
Kang, Y. (2011). Loanword phonology. In M. van Oostendorp, C. J. Ewen, E. Hume \& K. Rice (Eds.), The Blackwell companion to phonology, Vol. 4: Phonological interfaces. Malden, MA: WileyBlackwell.
Kapatsinski, V. (2009). The architecture of grammar in artificial grammar learning: Formal biases in the acquisition of morphophonology and the nature of the learning task. Doctoral Dissertation, Linguistics \& Cognitive Science, Indiana University.
Kapatsinski, V. (2010). Velar palatalization in Russian and artificial grammar: Constraints on models of morphophonology. Laboratory Phonology, 1, 361-93.
Kapatsinski, V. (2012). What statistics do learners track? Rules, constraints or schemas in (artificial) grammar learning. In S. Th. Gries \& D. Divjak (Eds.), Frequency effects in language: Learning and processing. Mouton de Gruyter.
Kapatsinski, V. (2012b). Minding gaps or seeking bumps in learning phonotactics? Poster presented at the \(86^{\text {th }}\) Linguistic Society of America Annual Meeting, Portland, OR
Kapatsinski, V. (2013). Conspiring to mean: Experimental and computational evidence for a usage-based harmonic approach to morphophonology. Language, 89.
Kenstowicz, M. (1994). Syllabification in Chukchee: A constraintsbased analysis. In A. Davison, N. Maier, G. Silva \& W.S. Yan (Eds.) Proceedings of the Formal Linguistics Society of MidAmerica 4. Iowa City: University of Iowa.
Kenstowicz, M. (1996). Base-identity and uniform exponence: Alternatives to cyclicity. In J. Durand \& B. Laks (Eds.), Current trends in phonology: Models and methods. Manchester: European Studies Research Institute \& University of Salford.
Kochetov, A. (2011). Palatalization. In M. van Oostendorp, C. J. Ewen, E. Hume \& K. Rice (Eds.), The Blackwell companion to phonology, Vol. 3. Malden, MA: Wiley-Blackwell.

Lobben, M. (1991). Pluralization of Hausa nouns, viewed from psycholinguistic experiments and child language data. M.Phil. Thesis, Dept. of Linguistics \& Philosophy, University of Oslo.
McCarthy, J. J. (1998). Morpheme structure constraints and paradigm occultation. Chicago Linguistic Society, 32, 125-50.
McCarthy, J. J. (2005). Optimal paradigms. In L. Downing, T. A. Hall \& R. Raffeleisen (Eds.), Paradigms in phonological theory. Oxford: Oxford University Press.
Meyer, A. S. (1992). Investigation of phonological encoding through speech error analyses: Achievements, limitations, and alternatives. Cognition, 42, 181-211.
Mielke, J. (2008). The emergence of distinctive features. Oxford: Oxford University Press.
Mitrovic, I. (2012). A phonetically natural vs. native language pattern: An experimental study of velar palatalization in Serbian. Journal of Slavic Linguistics, 20, 229-68.
Mowrey, R. A., \& MacKay, I. R. A. (1990). Phonological primitives: Electromyographic speech error evidence. Journal of the Acoustical Society of America, 88, 1299-1312.
Ohala, J. J. (1978). Southern Bantu vs. the world. Berkeley Linguistics Society, 4, 370-86.
Ohala, J. J. (1990). The phonetics and phonology of aspects of assimilation. In J. Kingston \& M. Beckman (Eds.), Papers in Laboratory Phonology I. Cambridge:Cambridge University Press.
Paradis, C., \& LaCharité, D. (1997). Preservation and minimality in loanword adaptation. Journal of Linguistics, 33, 379-430.
Roelofs, A. (1997). The WEAVER model of word-form encoding in speech production. Cognition, 64, 249-84.
Roelofs, A. (1999). Phonological segments and features as planning units in speech production. Language \& Cognitive Processes, 14, 173-200.
Rumelhart, D., \& J. McClelland. (1986). On learning the past tenses of English verbs. In J. L. McClelland, D. E. Rumelhart, \& the PDP research group. Parallel distributed processing: Explorations in the microstructure of cognition. Volume II. Cambridge, MA: MIT Press.
Schane, S., Tranel, B., \& Lane, H. (1975). On the psychological reality of a natural rule of syllable structure. Cognition, 3/4, 351-8.
Shattuck-Hufnagel, S. (1983). Sublexical units and suprasegmental structure in speech production planning. In P. F. MacNeilage (Ed.), The production of speech. New York: Springer.
Smith, L. B., Thelen, E., Titzer, R., \& McLin, D. (1999). Knowing in the context of acting: The task dynamics of the A-not-B error. Psychological Review, 106, 235-260.
Smolensky, P., \& Legendre, G. (2006). The harmonic mind: From neural computation to Optimality-Theoretic Grammar. Cambridge, MA: MIT Press.
Stemberger, J. P. (1982). The nature of segments in the lexicon: Evidence from speech errors. Lingua, 56, 235-59.
Stemberger, J. P. (1989). Speech errors in early child language production. Journal of Memory \& Language, 28, 164-88.
Stemberger, J. P. (1990). Wordshape errors in language production. Cognition, 35, 123-157.
Steriade, D. (2001). The phonology of perceptibility effects: The Pmap and its consequences for constraint organization. Ms. MIT.
Vousden, J. I., \& Maylor, E. A. (2006). Speech errors across the lifespan. Language \& Cognitive Processes, 21, 48-77.
White, J. (2012). Evidence for a learning bias against "saltatory" phonological alternations in artificial language learning. Paper presented at the \(86^{\text {th }}\) LSA Annual Meeting. Portland, OR.
Wilson, C. (2006). Learning phonology with substantive bias: An experimental and computational study of velar palatalization. Cognitive Science, 30, 945-982.
Yun, G. H. (2006). The interaction between palatalization and coarticulation in Korean and English. Doctoral Dissertation, Linguistics Department, University of Arizona.
Zuraw, K. (2000). Patterned exceptions in phonology. Doctoral Dissertation, Linguistics Department, UCLA.

\title{
Parsing Sequentially Presented Commands in a Large-Scale Biologically Realistic Brain Model
}

\author{
Terrence C. Stewart (tcstewar@uwaterloo.ca) \\ Chris Eliasmith (celiasmith@uwaterloo.ca) \\ Centre for Theoretical Neuroscience, University of Waterloo \\ 200 University Avenue West, Waterloo, Ontario N2L 3G1 Canada
}

\begin{abstract}
We present a neural mechanism for interpreting and executing visually presented commands. These are simple verb-noun commands (such as WRITE THREE) and can also include conditionals ([if] SEE SEVEN, [then] WRITE THREE). We apply this to a simplified version of our large-scale functional brain model "Spaun", where input is a \(28 \times 28\) pixel visual stimulus, with a different pattern for each word. Output controls a simulated arm, giving hand-written answers. Cortical areas for categorizing, storing, and interpreting information are controlled by the basal ganglia (action selection) and thalamus (routing). The final model has \(\sim 100,000\) LIF spiking neurons. We show that the model is extremely robust to neural damage ( \(40 \%\) of neurons can be destroyed before performance drops significantly). Performance also drops for visual display times less than 250 ms . Importantly, the system also scales to large vocabularies ( \(\sim 100,000\) nouns and verbs) without requiring an exponentially large number of neurons.
\end{abstract}

Keywords: neural engineering; parsing; cognitive control; spiking neurons; whole-brain systems; cognitive architecture

\section*{Large-Scale Functional Brain Modelling}

Our goal is to produce models of human cognition that are specified down to the neural level. That is, we want to know how the low-level neural details (including spikes and various neurotransmitters) give rise to human behaviour via their complex interconnections and interactions. We have previously published our first step in this direction, which is currently the world's largest functional brain model (Eliasmith et al., 2012). This model, "Spaun", has 2.5million spiking neurons, includes twenty different brain areas, and can perform eight different cognitive tasks (including digit recognition, list memory, addition, and pattern completion). Input is through a single eye with a 28 by 28 retina, and the output controls a simulated three-joint six-muscle arm, allowing it to write its answers. Spaun is told what task to perform via its visual input, so it must selectively re-route information between brain areas as appropriate for different tasks. This uses the cortex-basal ganglia-thalamus loop, where the basal ganglia performs action selection by comparing the current brain state to the ideal brain state for each action, and the chosen action activates cortical communication channels via the thalamus.

One limitation with Spaun is that it cannot learn new tasks. The eight tasks it can perform are set by the synaptic connections between cortex and basal ganglia. To address this, the work presented here adds a new general-purpose task for Spaun: one where it can be visually presented with commands for it to follow.

\section*{Parsing Visual Commands}

To provide new instructions to a model that only has a visual input, we need the model to process a sequential set of images and convert that into an internal representation of a command. This is a simplified language comprehension task, within a fairly restricted domain.

Basic commands can be thought of as verb-noun pairs, such as WRITE NINE. However, because the visual system is limited to \(28 \times 28\) pixels, it does not have the visual acuity to interpret full words at a time. Rather than flashing each letter in each word up individually (a fairly non-typical reading strategy), we use a single symbol for each word, and present those symbols sequentially. So, for the command WRITE NINE, we present a "W" followed by a " 9 ".

Valid commands are limited by the set of basic actions that the model knows how to do. While the full Spaun model can perform many operations including mental arithmetic, keeping track of elements in a list, and pattern completion, for simplicity in this paper we only consider the actions WRITE (W), REMEMBER (R), and INCREASE (C). For example, the model could be told to remember a two, increase it, and write the result ("R 2, C \#, W \#", where \# is a general-purpose indexical referring to the number currently being remembered, and a comma is a slight pause between instructions). The correct result from this command would be the written number 3 .

Furthermore, instructions can also include conditional clauses based on what the model can currently SEE ("S"). For example, "S 4, W 9" is interpreted as "if you see a four, then write a nine". To do this, the model must be capable of representing structured relationships.

The goal of this work is to give a spiking neuron implementation of this parsing process, integrated within an existing spiking neuron model of the rest of the brain (including vision, motor, working memory, and cognitive control areas). To argue that this is a plausible model, we show that a) its performance degrades gracefully as neurons die, b) it fails on uninterpretable grammatical structures, and c) it scales to human-sized vocabularies, dealing with the exponential growth of vocabulary combinations.

That said, there are considerable limitations to this work. It does not deal with token separation, since the symbols are shown to it one at a time. It also does not handle ambiguous terms (all symbols have exactly one meaning). We are also not specifying the full developmental and learning process that results in this model (although existing learning rules could be used, given a detailed error signal and large amounts of time).

\section*{The Neural Engineering Framework}

The Neural Engineering Framework (NEF; Eliasmith \& Anderson, 2003) transforms a high-level description of the variables being represented and the operations on those variables into a detailed spiking-neuron model subject to neurobiological constraints.

In the NEF, neurons are organized into groups, and each group forms a distributed representation of a particular variable. Different patterns of activity across the group correspond to different values for that variable. These values are, in general, vectors, so a particular group of 2,000 neurons might represent a 64-dimensional variable. While the NEF supports any neuron type, for this paper we use standard leaky integrate-and-fire (LIF) neurons whose parameters (refractory period, capacitance, neurotransmitter time constant, etc.) are set to be consistent with the details of the particular brain regions being modelled.

Within a population of neurons, each neuron has a particular preferred direction vector. This is the vector for which that neuron will fire most strongly. These vectors \(\boldsymbol{e}\) are randomly chosen along with the neuron gain \(\alpha\) and bias \(J_{b i a s}\) to produce a heterogeneous population of neurons. The current flowing into a neuron when representing a vector \(\boldsymbol{x}\) is given by Equation 1.

Given a pattern of activity, we can estimate the currently represented \(\boldsymbol{x}\) value as \(\sum d_{i} a_{i}\) where \(a_{i}\) is the neuron activity and \(\boldsymbol{d}\) is a decoder given by Equation 2. This is the optimal least-squares linear estimate of \(\boldsymbol{x}\) (the value being represented) given \(a\) (the current activity of the neurons).

The key part of the NEF is that this decoder also allows us to determine the synaptic connection weights between neural groups that will force them to compute a desired function. For example, if we want to connect a neural group representing \(\boldsymbol{x}\) to a neural group representing \(\boldsymbol{y}\) such that \(\boldsymbol{y}=\boldsymbol{M} \boldsymbol{x}\) (where \(\boldsymbol{M}\) is an arbitrary matrix), then the synaptic connection weights between neuron \(i\) in the first population and neuron \(j\) in the second are given by Equation 3.

For connections that compute nonlinear functions, we adjust Equation 2 slightly as given in Equation 4. This finds a decoder that approximates the arbitrary function \(f(x)\).
\[
\begin{align*}
& J=\alpha e \cdot x+J_{\text {bias }}  \tag{1}\\
& d=\Gamma^{-1} Y \quad \Gamma_{i j}=\int a_{i} a_{j} d x \quad Y_{j}=\int a_{j} x d x  \tag{2}\\
& \omega_{i j}=\alpha_{j} e_{j} M d_{i}  \tag{3}\\
& d^{f(x)}=\Gamma^{-1} Y \quad \Gamma_{i j}=\int a_{i} a_{j} d x \quad Y_{j}=\int a_{j} f(x) d x \tag{4}
\end{align*}
\]

It should be noted that the accuracy with which neurons will perform the desired computation using this technique is dependent on many factors. This includes the neuron properties such as overall firing rate and their membrane time constant. Accuracy can be increased arbitrarily by increasing the number of neurons (but, of course, to be realistic we are constrained by the number of neurons in the
brain). In general, discontinuous functions are very difficult for neurons to approximate.

\section*{Symbol-Like Processing with Spiking Neurons}

While the NEF allows us to convert algorithms that use vectors and functions into spiking neuron models, a further technique is needed to handle the symbol manipulation that is the hallmark of cognitive activity. This is especially important for parsing and representing complex commands.

The core idea is to have a particular vector for each atomic symbol that can be represented. For this paper, these vectors are chosen randomly, but they can also be chosen such that semantically related symbols have similar vectors. To combine symbols, we perform computations on these vectors, giving new vectors that represent the combination.

There are a variety of computations that can be used to combine these vectors (Gayler, 2003), but for our model we follow Plate (2003). Here, symbols can be combined by vector addition (+) and circular convolution ( \(\circledast\) ). Both operations are accurately approximated by the NEF method.

To demonstrate how this system works, consider representing the command "If you see a 9, write an 8". A simplistic approach would be to take vectors for all the atomic concepts (SEE, NINE, WRITE, and EIGHT) and add them together to represent the full sentence (SEE+NINE+WRITE+EIGHT). However, this does not work, since the resulting sentence loses all order information. In particular, the command "If you see an 8 , write a 9 " would result in exactly the same vector.

To deal with this, we use circular convolution ( \(\circledast\) ) and introduce new vectors for denoting structural information. The \(\circledast\) operator takes two vectors and produces a new vector that is highly dissimilar to the original two. So instead of WRITE+EIGHT we can do VERB \(\circledast\) WRITE+ NOUN \(\circledast\) EIGHT. Furthermore, we can nest this process to make more complex phrases. The full command can be represented by the vector \(\mathbf{S}=\mathbf{C O N D I T I O N} \circledast(\mathbf{S E E} \circledast\) NINE)+ VERB \(\circledast\) WRITE+NOUN \(\circledast E I G H T\).
Importantly, given this vector \(\mathbf{S}\) that represents a full command, we can extract out the individual components. Plate (2003) showed that a simple re-arranging of the elements of a vector makes an approximate inverse operation. For example, if we want to know the main verb in \(\mathbf{S}\), we compute \(\mathbf{S} \circledast \boldsymbol{i n v ( V E R B )}\). The result will be approximately WRITE. The accuracy of this approximation depends on the number of terms being added and the dimensionality of the vectors. In particular, as the number of dimensions increases, there is an exponential growth in representational capacity.

We refer to these vectors as semantic pointers. They are semantic in that the vector itself has meaning about the whole. Most usefully, the similarity between vectors indicates the similarity of the full structure. WRITE commands will have a higher degree of similarity to each other than to other commands. Furthermore, they are pointers in the computer science sense because they can be dereferenced, recovering (an approximation of) the original
data. Semantic pointers are compressed representations, in the same way that vision models can be thought of as compressing an image into a high-level representation.

\section*{Vision}

For vision, we adapt a Deep Belief Network (Hinton, 2010). The input is a 28 by 28 pixel retinal image, which is then processed by four different cortical layers. Each layer learns to extract and compress the regular patterns in the layer before it. We convert this model to spiking neurons by simulating each neuron in the DBN with ten realistic spiking neurons and using Equations 3 and 4 to solve for the connection weights that approximate the original model.

The output from the DBN (inferior temporal cortex) must then be mapped to a semantic pointer. One way to perform this mapping would be to use an associative cleanup memory (Stewart, Tang, \& Eliasmith, 2011), which scales to hundreds of thousands of items but requires additional neurons to recognize each item. For simplicity, here we use no additional neurons, but rather compute an approximate mapping between the compressed representation of the visual stimulus and the desired semantic pointers (Equation 5), where \(v_{i}\) is the average output of the Deep Belief Network for a particular category (all the 3's, for example), and \(s_{i}\) is the corresponding semantic pointer (THREE). As always, Equations 3 and 4 give the synaptic connections.


Figure 1: Converting visual input into the correct semantic pointer. For example, showing an image of a 7 to the retina will produce the vector SEVEN in the vision population.

\section*{Simple Parsing}

A first step towards parsing commands is to identify and store noun-verb pairs. That is, given a visual input of WRITE followed by THREE, we need one group of neurons to represent the verb (WRITE) and another to represent the noun (THREE). The outputs from these groups then drive a third group of neurons to represent the full phrase VERB \(\circledast\) WRITE+NOUN \(\circledast\) THREE. This is accomplished by connecting the verb population to the phrase population with connection weights optimized to perform the function
\(f(x)=x \circledast\) VERB. Similarly, the noun population's connection is optimized for the function \(f(x)=\boldsymbol{x} \circledast\) NOUN. As with all synaptic connections in this model, this optimization is performed using Equations 3 and 4.


Figure 2: 5760 spiking neurons combining two arbitrary 64dimensional vectors (noun and verb) into a single 64dimensional vector phrase \(=\) verb \(\circledast\) VERB + noun \(\circledast\) NOUN.

In order for this system to work in the context of an overall brain model, we need a mechanism to selectively route information from visual areas to the two populations. When the system sees WRITE it should pass this vector into the verb neurons, and when it sees THREE it should pass that vector to the noun neurons. This sort of selective routing of information is exactly what the cortex-basal ganglia-thalamus loop is believed to perform. We use our existing model for this loop, which is based on our spiking version of a model of action selection in the basal ganglia (Stewart, Choo, \& Eliasmith, 2010).

The basal ganglia selects between two actions. One action is to route information from the vision system to the noun population, and the other is to route that same information to the verb population. To perform these actions the output from the basal ganglia goes to the thalamus, where it releases the inhibition on the desired communication channel. (A communication channel is a connection that computes the function \(f(x)=\boldsymbol{x})\). If this is inhibited, the neurons do not fire, and so no information is passed. Selecting the action releases the inhibition, allowing the information to flow.

To decide which action to perform, the inputs to the basal ganglia must compute the utility of the two actions. For the first action, this is a function that outputs a 1 if there is a noun in vision, and a 0 otherwise. The second action is similar, but for identifying verbs. This is a simple classifier and can again be expressed as a function whose connection weights are computed with Equations 3 and 4.

The resulting system is capable of taking a stream of visual stimulus as input and keeping track of the most recent noun and the most recent verb seen. This verb and noun are then combined into a single vector representing that pair.


Figure 3: Routing information from vision to the correct noun and verb populations, depending on whether the visual stimulus is categorized as a noun or a verb. Once routed, the correct combined phrase is computed as in Figure 2.

\section*{Responding to Commands}

Once the model has formed a single representation of the command itself, we also need to show that it can execute that command correctly. That is, not only can the representation be encoded by spiking neurons, but it can also be decoded by spiking neurons to perform a task.

Since the focus of this paper is the parsing of a command, we use a very simple system for executing commands. In other work we show how to process significantly more complex commands, (Choo \& Eliasmith, submitted), but those commands are directly injected into the brain model, rather than being presented visually and parsed.

In this case, performing a command occurs via a new action added to the basal ganglia. It has a high utility when there is no visual input and when the phrase is similar to VERB \(\circledast\) WRITE. When selected, this action routes the information from phrase to motor while convolving it with the inverse of NOUN. Thus, if the phrase is VERB \(\circledast\) WRITE+NOUN \(\circledast\) FOUR, the semantic pointer \(\operatorname{inv}(\) NOUN \() \circledast(\) VERB \(\circledast\) WRITE+NOUN \(\circledast\) FOUR) will be routed to the motor area. Since NOUN and inv(NOUN) approximately cancel, the value set to the motor area will be close to the ideal vector for FOUR. This is mapped to a series of hand positions using the same method as Eq. 5.

The behaviour of the model is shown in Figure 5. The visual input is shown in the top row, and the written responses are shown in the bottom row. The other rows show the spiking behaviour of 50 neurons from each of the key brain areas in the model.

The first thing to note is that the model performs accurately. The correct response is given for each case. Furthermore, it should be noted that the two words in the command can be given in either order (WRITE FIVE versus TWO WRITE). This is because we have not imposed a particular grammatical order. While it would be possible to do so, we note that English speakers are quite capable of


Figure 4: Executing a WRITE action
correctly interpreting TWO WRITE as a command. However, as demonstrated in the section on Conditional Statements, word order does matter for complex commands.

The spike patterns shown in Figure 5 provide some insight into the performance of the model. In the vision row, we can see different patterns of activity for each visual input, as expected. The pattern for "W" and the pattern for seeing nothing at all are quite distinct. Similarly, the spike patterns in the noun and verb populations change depending on which term is currently being memorized, and these patterns in turn affect the phrase population.

Another feature that can be seen in Figure 5 is the cognitive reaction time. Each symbol is shown for 0.5 seconds, but the motor output is clearly delayed slightly. For example, the visual input is cleared at \(\mathrm{t}=1.0 \mathrm{~s}\), but the spiking behaviour in motor doesn't change until \(\mathrm{t}=1.05 \mathrm{~s}\). This is the time required for the model to notice the change in visual input, perform action selection in the basal ganglia, release the inhibition in the thalamus, and allow the information from the phrase neurons to pass to the motor neurons. The exact amount of time required is a function of the connectivity and neurotransmitter time constants, all of which are taken from neurological data (so they are not free parameters). For more analysis of this feature of the basal ganglia model, see (Stewart, Choo, \& Eliasmith, 2010).


\section*{Memory}

To demonstrate that this system can parse commands other than WRITE <number>, we add a memory action. If the model is told to REMEMBER 4 ("R 4"), the phrase will be similar to VERB \(\circledast\) REMEMBER. We add an action for this condition that routes the phrase information to the memory while transforming it by inv(NOUN). As with the WRITE action, this extracts the FOUR from the phrase. For this action, the output vector is routed to a working memory area. This is a group of neurons that stores a vector (like every other group of neurons in the model), but that has a communication channel back to itself. This recurrent connection causes the neurons to maintain their own state after the input is removed. This structure has been shown to match neural behaviour of visual working memory (Singh and Eliasmith, 2006), and is stable over long periods of time (tens of seconds).

Finally, we show that we can extract information from working memory by adding a special write action WRITE NUMBER ("W \#"). This action has a high utility when the phrase is VERB \(\circledast\) WRITE+NOUN \(\circledast\) NUMBER and causes the information in working memory to be routed to the motor system.

The result (Figure 6) is a system that can respond correctly to two different verbs and ten nouns (only ZERO through NINE were implemented). Importantly, adding the new action did not require any modifications to the phrase population. This is due to the fact that the semantic pointers used to represent the phrase are simply fixed-length vectors, and the phrase population is capable of storing any vector. No modifications to that population are needed to let it store a new vector like VERB \(\circledast\) REMEMBER+NOUN \(\circledast\) THREE, even if it has never seen it before.

In other words, the model does not require an exponential growth in numbers of neurons in order to handle the exponential growth in possible phrases that it can correctly interpret. Adding new actions only requires adding the neural populations needed to perform that action (in this case, the working memory population) and new connections between existing populations (in this case, phrase) and the basal ganglia and thalamus.


Figure 6: A model with REMEMBER and WRITE actions. Given an input REMEMBER SIX <long pause> WRITE NUMBER it will write the number 6.

\section*{Model Performance}

The behaviour of this model in a variety of conditions is shown in Figure 7. For each condition, new neurons (with preferred direction vectors, gains, and background currents) were generated, and Equations 3 and 4 were used to solve for all the synaptic connection weights.

First, the maximum vocabulary size (the largest number of nouns such that there is still a \(95 \%\) chance of correctly responding) scales exponentially as the number of neurons per population increases. This is an expected consequence of vector representations (e.g., Plate, 2003), as the number of dimensions accurately represented scales linearly with the number of neurons, while the volume of a hypersphere scales exponentially with the number of dimensions.

Second, the model is robust to destruction of neurons. To simulate neural death, we randomly delete neurons from every population, and then re-use Equations 3 and 4 to compute new connection weights between the remaining neurons. Performance decreases, but remains above well above chance until less than \(40 \%\) of the neurons remain. This shows a gradual degradation in behaviour, rather than catastrophic failure.

Finally, we show the model performs well for varying stimulus presentation times, but poor performance when symbols are seen for less than 250 milliseconds.


Figure 7: Model performance for varying vocabulary sizes, neural destruction, and display times. Input is of the form "R <number>, W \#", and an output is judged correct if the model writes the correct number. Shaded area is the \(95 \%\) bootstrap confidence interval over 50 trials.

\section*{Conditional Statements}

The method used to add the REMEMBER action can be used to add many new actions. For instance, an INCREMENT action can be added which increases the number stored in memory. However, to show that this method extends to more complex rules, we now consider the parsing of a conditional rule such as "if you see a six, write a one". Using " \(S\) " to as the symbol for SEE, we can present this to the model as "S 6 W 1". We then add actions to the basal ganglia such that a phrase of VERB \(\circledast\) SEE will cause that phrase to be routed to a new condition group of neurons. A global state population is created, which gets inputs from all cortical areas that could be used as part of a condition (in this case, just vision), so that if vision is TWO then the value TWO®SEE will be added to state. The state and condition vectors are then compared (by computing the dot product) in a similarity population. Finally, the go/nogo population uses the similarity and condition values to compute a penalty to be applied to the utilities of actions in the basal ganglia. That is, it decreases the utility if there is a condition but condition does not match the current state.

Importantly, once the condition is stored in a separate neural population, we can now combine the condition and the phrase into a single semantic pointer. For this case, the resulting vector would be CONDITION \(\circledast(S E E \circledast\) SIX) +VERB \(\circledast\) WRITE+NOUN \(\circledast\) ONE. This is a single vector representing a complex, syntactically structured command that can be successfully executed by this model. In (Choo \& Eliasmith, submitted), we develop a model capable of following a collection of rules of this form, but the model presented here is the first biologically realistic spiking model capable of taking the sequential input "S 6 W 1" and parsing it to create the correct semantic pointer vector.

Interestingly, the model supports some syntactic variation, such as "6 S 1 W " or "6 S W 1". However, it will not perform correctly when one phrase is embedded in the center of the other ("S W 61 ", for example). This difficulty with center embedding is a well-studied feature of natural languages, and appears naturally in this model from the processing available to neural populations.

Finally, in order to successfully respond to a condition instruction in the other order ("W 1 S 6"), we must also add an extra action rule which stores the initial phrase (VERB \(\circledast\) WRITE+NOUN \(\circledast\) ONE) in memory before processing the conditional phrase. This extra cognitive load indicates this model finds it easier to process "If you see a six, write a one" than "Write a one if you see a six".


Figure 8: Additions needed for conditional instructions.

\section*{Conclusion}

We have shown how a model consisting of spiking leaky-integrate-and-fire neurons with properties and connection patterns that match the human brain can take a visually presented input command, parse it, and perform the correct action. This works for simple verb-noun commands and for more complex conditional commands, and also scales up to a vocabulary size of hundreds of thousands of terms. The majority of the neural components are identical those in our previous models (e.g. Eliasmith et al., 2012).

To perform this parsing, the model builds a single combined vector representation of the command. This resulting structured representation is of exactly the same form as those we have used in other neural models. As such, this model provides a potential explanation as to how brains can form and manipulate these symbol-like structures that are found throughout cognition.

That said, the current model has many limitations. It does not impose particular grammatical rules (other than avoiding center embedding). Perhaps relatedly, it does not address the problem of ambiguous classifications. It also does not perform token separation, as it requires that the input be already sequentially arranged. Fortunately, these problems have been addressed by other researchers, and our ongoing work is to adapt their solutions to the constraints of a biologically realistic spiking model. In particular, Ball (2011) provides an extensive project to process natural language speech using the ACT-R cognitive architecture, which may be adaptable to our neural framework.

\section*{References}

Ball, J. (2011). A Pseudo-Deterministic Model of Human Language Processing. \(33^{\text {rd }}\) Cog. Sci. Society Conference.

Choo, X and Eliasmith, C. (submitted). General Instruction Following in a Large-Scale Biologically Plausible Brain Model. \(35^{\text {th }}\) Cog. Sci. Society Conference.

Eliasmith, C. (2013). How to build a brain. Oxford University Press, New York, NY.
Eliasmith, C. \& Anderson, C. (2003). Neural Engineering. Cambridge: MIT Press.Eliasmith et al., 2012
Gayler, R. (2003). Vector Symbolic Architectures Answer Jackendoff's Challenges for Cognitive Neuroscience, in Slezak, P. (ed). Int. Conference on Cognitive Science, Sydney: University of New South Wales, 133-138.
Hinton, G.E. (2010). Learning to represent visual input. Phil. Trans. Roy. Soc. B, 365, 177-184.
Plate, T. (2003). Holographic Reduced Representations, CSLI Publications, Stanford, CA.
Singh, R., \& Eliasmith, C. (2006). Higher-dimensional neurons explain the tuning and dynamics of working memory cells. Journal of Neuroscience, 26, 3667-3678.

Stewart, T.C., Choo, X., and Eliasmith, C. (2010). Dynamic Behaviour of a Spiking Model of Action Selection in the Basal Ganglia. \(10^{\text {th }}\) Int. Conf. on Cognitive Modeling.

Stewart, T. C., Tang Y., \& Eliasmith C. (2011). A Biologically Realistic Cleanup Memory: Autoassociation in Spiking Neurons. Cog.Systems Research. 12(2), 84-92.

\title{
Complex Mental Time Travel
}

\author{
Kurt Stocker (kurt.stocker@psychologie.uzh.ch) \\ Department of Psychology, Binzmühlestrasse 14/22 \\ 8050 Zürich, Switzerland
}

\begin{abstract}
This article argues that investigating the conceptual structure underlying the use of the pluperfect and the future perfect reveals a new complex type of nested dual mental time travel: mental time travel into posteriority embedded into mental time travel into "anteriority in the past" (underlying the pluperfect) versus mental time travel into posteriority embedded into mental time travel into "anteriority in the future" (underlying the future perfect). Additionally this article also offers the following novel notions for temporal cognition: a mental time line where past/anteriority and future/posteriority have become nondispersible; dual temporal direct viewings at the present moment; and looking into the future from the past (rather than the more typical looking into the future from the present moment). Implications for cognitive modeling are discussed.
\end{abstract}

Keywords: mental time travel; tense system; Talmyan concept structuring; Talmyan perspective point (PP); mental time line; models

\section*{Introduction}

Until recently mental time travel has mainly been characterized as mentally construing oneself as looking forward or backward in time from the present moment (e.g., Addis et al., 2009; Schacter \& Addis, 2007; Tulving, 1972, 2002). By synthesizing findings from cognitive psychology and cognitive linguistics and by additionally applying cognitivelinguistic methodology, Stocker (2012a) then introduced the idea-based on a sketch by Talmy (2000, pp. 86-87)—that in addition to this basic type of mental time travel there might also be more complex types of mental time travel. For instance: a person may mentally construe herself as looking back from the present moment to a particular point in time in the past, but may additionally also conceptualize herself as mentally looking forward from this past point to a "later time" that is still in the past. Such examples have been referred to as examples of nested dual mental time travel ("mental time travel embedded within mental time travel") (Stocker, 2012a, p. 408). Investigating the conceptual structure underlying the linguistic use of before/after sentences that additionally are set in the past or future tense, Stocker (2012a) has thus far basically identified one form of nested dual mental time travel: mental time travel into anteriority or posteriority (underlying before/after) embedded in mental time travel into the past or future (underlying past/future tense). It is important to distinguish anteriority/posteriority ("earlierness/laterness") from past/future since the former is more generic and does not depend upon the present moment as a reference point (e.g., Núñez \& Sweetser, 2006, p. 404). For instance: One event may have occurred later in time
than another event (say my first day at school versus my birth), but both events have occurred in the past.

This article investigates how this anteriority/posteriority versus past/future distinction can help us to reveal the tem-poral-conceptual structure underlying the pluperfect and future perfect. The theoretical strategy I adopt is the same as used in Stocker (2012a): using language as an entree to a conceptual level that seems deeper than language itself (Pinker, 2007; Talmy, 2000). This strategy is supported by recent findings that many conceptualizations observed in relation to our use of language also exist in mental representations that are more basic than language itself (e.g., Boroditsky, 2000; Casasanto \& Boroditsky, 2008; McGlone \& Harding, 1998; Núñez, Motz, \& Teuscher, 2006). In the present investigation language can assist us to identify complex forms of mental time travel-complex forms of how we can mentally project through time.

The basic theoretical framework used is Talmyan concept structuring (Talmy, 2000), with the further refinement for temporal cognition by Stocker (2012a). There are many other basic theoretical frameworks that one could adopt when investigating the conceptual structure underlying the tense system or the conceptual structure of mental time in general-for example: formal accounts of tense (e.g., Comrie, 1985; Declerck, 1986; Jespersen, 1924; Reichenbach, 1947), conceptual semantics (Jackendoff, 1987, pp. 398402; cf. also Pinker 1989, pp. 205-206), formal semantics (e.g., Bennett \& Partee, 1978; Montague, 1973; Pendlebury, 1992), or temporal (tense) logic (e.g., Allen, 1984; Kowalski \& Sergot, 1986; Lichtenstein \& Pnueli, 2000; Prior 1967). While the current investigation is basically set in a Talmyan framework, it also, as we will see, benefits greatly from the formal-tense analysis of Comrie (1985).

One of the main motivations for choosing Talmyan concept structuring as a basic theoretical framework for the present investigation is that it offers a ready means to incorporate mental temporal perspective (Stocker, 2012a; Talmy, 2000, pp. 68-76+86-87). In the other above-mentioned approaches (formal tense, conceptual semantics, formal semantics, temporal logic), mental perspective is usually not considered or is only mentioned marginally, without incorporating it into the formal descriptive apparatus (e.g., in Jackendoff, 1987, p. 399). In contrast, in Talmyan concept structuring, perspective is an integral part of the overall theoretical descriptive system.

The present investigation will reveal several basic novel notions in relation to temporal cognition (as summarized in the discussion section). It will also be discussed if the current account of complex mental time travel could be used to refine modeling approaches which have incorporated mental
temporal perspective into their models (Brown, Neath, \& Chater, 2007).

\section*{Mental time travel underlying the pluperfect}

Undertaking an extensive cross-linguistic investigation, the linguist Bernard Comrie characterizes the temporalrelational structure of the pluperfect (I had already eaten when ...) in the following way:
"The meaning of the pluperfect is that there is a reference point in the past, and that the situation in question is located prior to that reference point, i.e. the pluperfect can be thought of as 'past in the past'" (1985, p. 65).

As we will see later on in this section, a still more refined characterization of the meaning of the pluperfect-rather than saying that it signifies "past in the past"-is to characterize it as "anteriority in the past." To start investigating the temporal-conceptual structure underlying the pluperfect, we use one of Comrie's own examples for illustration (1985, p. 66):
(1) John had already left when Mary emerged from the cupboard.

According to Comrie the temporal relations underlying the use of the pluperfect can be formalized in the following terms (1985, p. 125):

\section*{(2) pluperfect: E before R before S}

E stands for the event which is to be located in time. In Comrie's example, the event of John's leaving is the event to be located prior to Mary's emerging from the cupboard. Hence the event in the pluperfect clause (John's leaving) is E . R stands for the temporal reference point in relation to which E is defined. Thus Comrie's formula correctly predicts that E (John's leaving) occurs before R (Mary's emerging from the cupboard). S stands for moment of speech (i.e., the present moment). Comrie's formula again correctly predicts that R (Mary's emerging from the cupboard) occurs before \(S\) (the present moment). \({ }^{1}\)

\footnotetext{
\({ }^{1}\) Comrie's (1985) ERS notation for the pluperfect represents a further development-and major departure-from the famous tense formulations of Reichenbach (1947, pp. 287-298; cf. also Jespersen, 1924, pp. 262-264). Comrie's formulations are mainly taken over because it is Comrie's analysis that allows one to characterize the pluperfect as involving "anteriority in the past" and the future perfect as involving "anteriority in the future"-which has major implications when one adds mental perspective to the pluperfect and future perfect constructions, as shall be demonstrated in this article. Taken over from Prior in the analysis in this paper, is the argumentation that there is no need to-as Reichenbach does-"make such a sharp distinction between the point or points of reference and the point of speech" (1967, p. 13). This is so, as also pointed out by Prior, because the present moment ("the point of speech") itself can function as a reference point. This argumentation of Prior is taken over by allowing the present moment to function as a Ground (see below in this article). "Ground" is the Talmyan technical term for what one might also refer to as a "ref-
}

As has just been demonstrated, (2) can correctly predict all temporal-relational structure of the pluperfect. The question we now turn to is: How could mental temporal perspective (Stocker, 2012a; Talmy, 2000, pp. 72-76+86-87) be added to this basic account of the temporal-relational structure of the pluperfect? One theoretical solution to this question, the one to be adopted in this article, is to integrate Comrie's findings into the theoretical framework of Talmyan concept structuring-because Talmyan concept structuring can describe temporal relations and temporal perspective in one coherent theoretical framework (Stocker, 2012a; Talmy, 2000). As a starting point, let us reformulate Comrie's pluperfect formula in Talmyan terms. In Talmyan concept structuring spatial or temporal relations are captured with the notions of Figure (F) and Ground (G) (Talmy, 2000). In temporal Figure/Ground, one event serves as temporal reference point-G-in relation to which the temporal location of the other event-F-is defined. Thus (2) can be captured in the following way in Talmyan terms:
(3) pluperfect: \(F_{1}\) before \(G_{1} ; F_{2}\left(G_{1}\right)\) before \(G_{2}\left(G_{2}=\right.\) present moment)

We again exemplify the formalized temporal relationship with (1). Now it is \(\mathrm{F}_{1}\) which stands for the event which is to be located in time (John's leaving). \(\mathrm{G}_{1}\) stands for the temporal reference point (Mary's emerging from the cupboard) in relation to which \(F_{1}\) is defined. Thus (3) correctly predicts that \(F_{1}\) (John's leaving) occurs before \(G_{1}\) (Mary's emerging from the cupboard). However, \(G_{1}\) also functions as another \(F\), since the temporal position of \(G_{1}\) is also defined in relation to the present moment. Hence, one is in a position to postulate that \(\mathrm{G}_{1}\) (Mary's emerging from the cupboard) also functions as an F (a second F in the overall temporal complex: \(F_{2}\) ) whose temporal position is defined in relation to the present moment (which functions as a second \(G: \mathrm{G}_{2}\) ). Thus, (3) also correctly predicts that \(\mathrm{F}_{2}\) (Mary's emerging from the cupboard) occurs before \(\mathrm{G}_{2}\) (the present moment).

Thus far, Comrie's pluperfect (2) and the Talmyan pluperfect (3) formalization are equipotent in terms of theoretical explanatory power: they both correctly predict the complex temporal relations that underlie our use of the pluperfect. But having it phrased in Talmyan terms allows us now to add mental temporal perspective (Stocker, 2012a; Talmy, 2000, pp. 72-76+86-87) to the temporal-relational description. Both Talmy and Stocker have cognitive-linguistically argued in detail that a complex temporal sentence (a temporal sentence with a main and a subordinate clause) underlies a temporal direct viewing of the F event in relation to the content of the main clause and a temporal indirect (prospective or retrospective) viewing of the \(G\) event in relation to the content in the subordinate clause. Taking over this analysis (see Stocker, 2012a; Talmy, 2000, pp. 72-76+8687 for argumentation), we derive at the (perspective-

\footnotetext{
erence point." For a different theoretical approach to the notion of
} a temporal perspective point see Declerck (1986, pp. 320-321).
including) temporal-conceptual structure underlying our use of the pluperfect as it is depicted in Fig. 1.


Figure 1: Mental time travel into posteriority (to \(G_{1}\) ) embedded into mental time travel into anteriority in the past (to the \(\mathrm{F}_{1}\)-co-located PP), a nested dual form of mental time travel underlying the pluperfect. This temporal-conceptual structure (and the cognition thereof) is in many respects identical to the one proposed by Stocker (2012a) for be-fore/past-tense constructions (p. 408). However, the crucial difference is that in before/past-tense constructions there are two distinct temporal Reference Frames (anteriority/posteriority and past/future RFs) whereas in constructions containing a pluperfect these two RFs have fused into one larger, more complex anteriority-past/posteriority-future RF.

When taking a look at this figure, the temporal structure (and perspectival cognition thereof) might at first glance seem identical to the conceptual structure underlying our use of a temporal complex sentence containing before and the past tense (a before/past-tense construction like \(I\) shopped at the store before I went home; cf. with Fig. 9 in Stocker, 2012a, p. 408). This is also not surprising: Comrie's characterization of the pluperfect clause as the "past in the past" could also be paraphrased as "past event before another past event." We should also note that Comrie's pluperfect characterization of "the past in the past" just serves him as a first rough characterization of the pluperfect (he uses the phrase to introduce the pluperfect). Crucially, Comrie notes that in relation to (2):
"Since the relation before is transitive (i.e. if X is before Y and Y is before Z , then necessarily X is before Z ), one can deduce \(E\) before \(S\) from the representation of the pluperfect, but this is not part of the formal representation of the pluperfect; the importance of this observation will become clear when we discuss the future perfect" (1985, p. 125).

In other words, what Comrie is saying is that the pluperfect is basically speaking not "a past in the past" (i.e., this can only be deduced), but anteriority in the past (since he says that S-the present moment-is in no way directly related to E). All that is inherent in (2)-or (3)-is that the event in the pluperfect must occur earlier than its reference event in the past. As with Comrie, we examine the importance of this observation when we examine the temporalconceptual structure underlying the future perfect (see next section). The observation that the pluperfect signifies "anteriority in the past" also leads us to the basic temporalconceptual difference between before/past-tense construc-
tions and complex sentences containing a pluperfect in the main clause and the simple past in the subordinate clause. In a before/past-tense construction, one can identify two distinct temporal Reference Frames (RFs): an anteriority/posteriority RF (underlying before) that is embedded in a past/future RF (underlying past/future-tense; as examined in Stocker, 2012a, where the term RF is also technically defined). But in a pluperfect construction, one cannot disentangle the anteriority/posteriority RF and the past/future RF. The observation that the pluperfect stands for "anteriority in the past" means that the temporal conceptual structure underlying the pluperfect has fused these two RFs into a larger complex whole: the pluperfect carries components of both these RFs within it. Trying to tease them apart would result in the dissolving of the sine qua non of the pluperfect: that it refers to an event that must occur earlier than another event in the past. It is in this sense that a pluperfect construction is more complex than a before/past-tense construction: underlying a pluperfect structure is a more complex RF (a mental time line) where components of two separate RF-systems have formed a new complex whole.

Additionally cognitive-linguistic analysis of complex temporal sentences in relation to Talmyan mental perspective points (PPs) suggests that \(F\) and \(G\) are cognized as points (punctual events) on the mental time line and they are mentally cognized from a distal PP (as detailed in Stocker 2012a; Talmy 2000, pp. 61-62). A distal PP means mentally zooming out as much from an event as to collapse the entire duration of an event to a single temporal point. The self needs to zoom out this much in order to be able to cognize two events-that is the relationship between the two events -from one perspective point. Note also that the observation that the pluperfect indicates that self travels back from the present moment to a point in time prior to another past event (to \(F_{1}\) ) means that the reference point in the past \(\left(G_{1}\right)\) can only be located in a prospective (later) direction when viewed from the perspective of \(\mathrm{F}_{1}\). Thus the self at the point in the past that is prior to another event in the past must mentally travel forward in order to establish the posterior reference point (in order to establish \(\mathrm{G}_{1}\) ). That the self travels from temporal F co-location to G (to establish a reference point at the temporal G point) has been examined in detail for before/after temporal constructions (Stocker, 2012a).

Stocker (2012a) also argues in detail (by providing cogni-tive-linguistic evidence) that the schematic geometric representations, as for instance shown in Fig. 1, are not merely a didactic aid that allows us to illustrate the underlying cogni-tive-temporal structure. Rather it is proposed that such geometry is actually construed in our mind when we conceptualize time. For instance: the depicted time line is proposed to be an actual, mentally construed spatial structure in our mind that allows for mental time travel-for instance in relation to the pluperfect by projecting one's mental gaze along this mentally construed line once in a retrospective direction (to "anteriority in the past") and once in a prospective direction (to the reference point in the past). The pro-
posal that the "mental time line" is mentally construed when we engage in mental time travel is also supported by a growing number of recent experimental behavioral findings. The mental time line is for instance frequently conceptualized in relation to the cognizer's body along the sagittal (back to front) or transversal (left to right) axis (e.g., Hartmann \& Mast, 2012; Miles, Nind, \& Macrae, 2010; Ulrich \& Maienborn, 2010).

\section*{Mental time travel underlying the future perfect}

Drawing-as in the pluperfect-extensively on crosslinguistic data, Comrie concludes that the temporalconceptual structure underlying the future perfect (I will have eaten when ...) differs to the one underlying the pluperfect in only one way: the reference point \(\left(G_{1}\right.\) in the Talmyan framework) is set in the future rather than in the past (Comrie 1985, p. 69-74). Accordingly, Comrie (p. 126) formalizes the temporal-relational structure underlying the pluperfect in the following way (cf. with (2)):

\section*{(4) future perfect: E before R after S}

Reformulation in Talmyan concept structuring (cf. with (3)); this will again enable us to integrate PP into the temporal cognition:
(5) future perfect: \(F_{1}\) before \(G_{1} ; F_{2}\left(G_{1}\right)\) after \(G_{2}\left(G_{2}=\right.\) present moment)

Both formulations-(4) and (5)-encode a remarkable finding of Comrie about the pluperfect (a finding that holds true cross-linguistically): that all that the future perfect indicates is that there must be a reference point \(\left(\mathrm{G}_{1}\right)\) in the fu-ture-but while the event referred to ( \(\mathrm{F}_{1}\) ) most typically also occurs in the future, it can also occur in the present or even in the past. Comrie:
"Let us start with the example John will have finished his manuscript by tomorrow. Let us suppose moreover that I do not know whether or not John has already finished his manuscript (or at least do not wish to reveal this knowledge), but I know (and am prepared to divulge) that he will have finished it by tomorrow - say, because he made a promise to this effect several days ago, and is judged by me to be reliable. Then there are three sets of circumstances in which I can felicitously and truthfully utter this statement. One set of circumstances is where John finishes his manuscript between the moment of my uttering this sentence and the reference point 'tomorrow'. The second is where John is in fact finishing his manuscript at this very moment, but I am unaware (or wish to give the impression that I am unaware) of this fact. The third is where John has already finished his manuscript, but I am unaware (or wish to appear unaware) of the fact. Thus the time reference of John's finishing his manuscript is left open as to whether it is future, present, or past relative to the present moment, the only stipulation
being that it must be prior to the reference point in the future, the sine qua non of the future perfect" \((1985\), p. 71).

This leads to three kinds of temporal relations that can underlie our use of the perfect: future perfect with future interpretation, future perfect with present interpretation, and future perfect with past interpretation (Comrie, 1985, p. 70). It is in this context where the anteriority/posteriority versus past/future distinction becomes highly relevant: whereas the structure underlying the pluperfect (by deduction) can be characterized as a "past in the past" (but is more precisely "anteriority in the past"; cf. previous section), this is no longer true for the structure underlying the future perfect. As the analysis of Comrie demonstrates, the temporal relations underlying the future perfect could not (also not by deduction) be characterized as "past in the future" (since this would only correctly characterize the future perfect with past interpretation). The only characterization that can capture the sine qua non of the future perfect is "anteriority in the future"-that is, a reference point \(\left(\mathrm{G}_{1}\right)\) in the future in relation to which an earlier event \(\left(\mathrm{F}_{1}\right)\) is defined, an event that can be located in the future, present, or past.

If we now add-as we did with the temporal-conceptual structure underlying the pluperfect-mental temporal perspective (Stocker, 2012a; Talmy, 2000, pp. 72-76+86-87), then these three possible interpretations of the future perfect naturally lead to three different kinds (subtypes) of nested dual mental time travel, as illustrated in Figs. 2-4.


Figure 2: Mental time travel into posteriority (to \(\mathrm{G}_{1}\) ) embedded into mental time travel into anteriority in the future (to the \(\mathrm{F}_{1}\) co-located PP), where the anterior event is also set in the future-a nested dual form of mental time travel underlying the future perfect with future interpretation.

The temporal-conceptual structure and cognition underlying the future perfect with future interpretation (Fig. 2) is largely identical to complex before-sentences that would additionally be marked as occurring in the future (cf. Stocker, 2012a). However, the vital difference is again-as in before-past-tense constructions (cf. previous section)-that in a before-relation where both events are set in the future there are two distinct temporal Reference Frames (anteriority/posteriority and past/future RFs) whereas in a construction containing a future perfect these two RFs have fused to one larger, more complex anteriority/past-posteriority/future RF where the two RFs can no longer be disentangled.


Figure 3: Mental time travel into posteriority (to \(G_{1}\) ) embedded in "mental time travel" into anteriority in the future (to the \(\mathrm{F}_{1}\) co-located PP ), where the anterior event is set at the present moment, a nested dual form of mental time travel underlying the future perfect with present interpretation.

The novel finding in the temporal-conceptual structure and cognition underlying the future perfect with present interpretation (Fig. 3) is that computational logic requires us to place the self twice at the present moment: the self must be located at the present moment in order to look out at the embedded self that is a distal distance removed from the time line (cf. previous diagrams); the second (embedded) self a distal distance away from the time line (but still colocated with the present moment) needs to look at the present moment on the time line so that \(\mathrm{F}_{1}\) can be cognized in a temporally direct way (cf. also previous diagrams). More technically speaking, the novel proposal is the existence of a dual form of temporal direct viewing, where both viewings are located at or co-located at the present moment. Note also that "mental time travel" into anteriority in the future is not really mental time "travel" in the present-interpretation case-since the anterior point happens to be at the present moment, the self at the present moment must cognize an embedded self a distal distance away from the timeline (but since this all happens at the present moment, the self does not really "travel" anywhere, at least not in a "forward/backward in time" sense).

The major novel observation in the temporal-conceptual structure and cognition underlying the future perfect with past interpretation (Fig. 4) is a looking forward from a past point (from the PP that is co-located with \(\mathrm{F}_{1}\) ) to a future point (to \(\mathrm{G}_{1}\) )-that is, a prospective projection through mental time that starts off in the past and extends (passing by the present moment as it were) right into the future.


Figure 4: Mental time travel into posteriority (to \(\mathrm{G}_{1}\) ) embedded in mental time travel into anteriority in the future (to the \(\mathrm{F}_{1}\) co-located PP ), where the anterior event is set in the past-a nested dual form of mental time travel underlying the future perfect with past interpretation.

\section*{Discussion}

The current investigation has-in addition to the findings of Stocker (2012a)—identified one more complex form of mental time travel: mental travel into posteriority embedded into mental time travel into "anteriority in the past" (underlying the pluperfect) versus mental time travel into posteriority embedded into mental time travel into "anteriority in the future" (underlying the future perfect). Additional novel notions include: a mental time line where past/anteriority and future/posteriority have become nondispersible; dual temporal direct viewings at the present moment; and looking into the future from the past (rather than the more typical looking into the future from the present moment). The last two of these novel notions have only been possible to identify because the current investigation uses a basic theoretical approach (Talmyan concept structuring for time: Stocker, 2012a; Talmy, 2000) that inherently incorporates temporal mental perspective into the explanatory framework.

One advantage for cognitive science in general that comes out of the current work (and of Stocker 2012a, 2012b) is that it offers a systematic and detailed explanatory framework how mental perspective can be included in a theory of temporal cognition. The relevance of this can for instance be illustrated in relation to cognitive models of memory retrieval. Brown et al. (2007) have introduced a retrieval model they call SIMPLE (scale independent memory, perception, and learning):
„... memory traces can be seen as located and individuated at least partly in terms of their position along a temporal continuum receding from the present into the past. This time line is logarithmically compressed, such that recent locations are more easily discriminable from one another than are more temporally distant locations" (p. 541).

As in SIMPLE, the current investigation has also identified a self who is looking back from the present moment along a mental time line to multiple temporal points (locations) in the past. Furthermore, the current investigation (see also Stocker, 2012a; Tulving, 1972, 2002) suggests that the
self at the present moment also mentally cognizes an additional (remembered) self in the past itself (in Figs. 1-4 this is always the self at the \(\mathrm{F}_{1}\)-co-located PP, a distal distance away from the time line). In the current framework, it is this remembered embedded self that looks out at the actual past events. In addition, Stocker (2012b) has reviewed findings that suggest how this embedded distal self in the past can take on an embodied (field) or disembodied (observer) mental perspective. Future research could address the question, whether it might be fruitful for temporal-perspectiveincluding models (like SIMPLE) to incorporate this ,,additional self" in the past. This then would allow such models to investigate if this embedded self (i) cognizes the memory items in the past in a temporally direct or temporally indirect (prospective or retrospective) way and (ii) if it cognizes the items in an embodied (field) or disembodied (observer) perspective. Such refinements are likely to be relevant for a recall model. For instance: In field (embodied) memories one is known to retrieve richer accounts of affective reactions, physical sensations, and psychological states whereas in observer (disembodied) memories one is known to retrieve richer accounts of the external environment, such as where things were located in the remembered surroundings (e.g, McIsaac and Eich, 2002).

\section*{Acknowledgments}

This research was supported by the Cogito Foundation, Wollerau, Switzerland (R-135/12).

\section*{References}

Addis, D. R., Pan, L., Vu, M. A., Laiser N., \& Schacter, D. L. (2009). Constructive episodic simulation of the future and the past: Distinct subsystems of a core brain network mediate imagining and remembering. Neuropsychologia, 47, 2222-2238.
Allen, J. F. (1984). Towards a general theory of action and time. Artificial intelligence, 23, 123-154.
Bennett, M. \& Partee, B. H. (1978). Toward the logic of tense and aspect in English. Indiana University Linguistics Club.
Boroditsky L. (2000). Metaphoric structuring: Understanding time through spatial metaphors. Cognition, 75, 1-28.
Brown, G. D., Neath, I., \& Chater, N. (2007). A temporal ratio model of memory. Psychological Review, 114, 539576.

Casasanto, D., \& Boroditsky, L. (2008). Time in the mind: Using space to think about time. Cognition, 106, 579593.

Comrie, B. (1985). Tense. Cambridge: Cambridge University Press.
Declerck, R. (1986). From Reichenbach (1947) to Comrie (1985) and beyond: Towards a theory of tense. Lingua, 70, 305-364.
Hartmann, M. \& Mast, F. W. (2012) Moving along the mental time line influences the processing of future related words. Consciousness and Cognition, 21, 1558-1562.

Jackendoff, R. (1987). The status of thematic relations in linguistic theory. Linguistic Inquiry, 18, 369-411.
Jespersen, O. (1924) The philosophy of grammar. London: Allen \& Unwin.
Kowalski, R. \& Sergot, M. (1986). A logic-based calculus of events. New Generation Computing , 4, 67-94.
Lichtenstein, O. \& Pneuli, A. (2000). Propositional temporal logics: Decidability and completeness. Logic Journal of IGPL, 8, 55-85.
McIsaac, H. K. \& Eich, E. (2002). Vantage point in episodic memory. Psychonomic Bulletin \& Review, 9, 146-150.
McGlone, M.S. \& Harding, J. L. (1998). Back (or forward?) to the future: The role of perspective in temporal language comprehension. Journal of Experimental Psychology: Learning, Memory, and Cognition, 24, 1211-1223.
Miles, L., Nind, L., \& Macrae, C. (2010). Moving through time. Psychological Science, 21, 222-223.
Montague, R. (1973). The proper treatment of quantification in ordinary English. In J. Hintikka, J. Moravcsik, \& P. Suppes (Eds), Approaches to natural language. Dordrecht: Reidel.
Núñez, R. E., Motz, B. A., \& Teuscher, U. (2006). Time after time: The psychological reality of the ego- and time-reference-point distinction in metaphorical construals of time. Metaphor and Symbol, 21, 133-146.
Núñez, R. E. \& Sweetser, E. (2006). With the future behind them: Convergent evidence from Aymara language and gesture in the crosslinguistic comparison of spatial construals of time. Cognitive Science, 30, 401-450.
Pendlebury, M. (1992). Elementary formal semantics for English tense and aspect. Philosophical Papers, 21, 215241.

Pinker, S. (1989). Learnability and cognition: The acquisition of argument structure. Cambridge, MA: MIT Press.
Pinker, S. (2007). The stuff of thought: Language as a window into human nature. New York: Viking.
Prior, A. N. (1967). Past, present and future. Oxford: Clarendon Press.
Reichenbach, H. (1947). Elements of symbolic logic. New York: Collier-Macmillan,
Schacter, D. L. \& Addis, D. R. (2007). Constructive memory: The ghosts of past and future. Nature, 445, 27.
Stocker, K. (2012a). The time machine in our mind. Cognitive Science, 36, 385-420.
Stocker, K. (2012b). Toward an embodimentdisembodiment taxonomy. Cognitive Processing, 13, S347-S350.
Talmy, L. (2000). Toward a cognitive semantics. Vol. 1: Concept structuring systems. Cambridge, MA: MIT Press.
Tulving, E. (1972). Episodic and semantic memory. In E. Tulving \& W. Donaldson (Eds.), Organziation of memory. New York: Academic.
Tulving, E. (2002). Episodic memory: From mind to brain. Annual Review of Psychology, 53, 1-25.
Ulrich, R. \& Maienborn, C. (2010). Left-right coding of past and future in language: The mental timeline during sentence processing. Cognition, 117, 126-138.

\title{
Representation and Criterion Differences between Men and Women in Semantic Categorization
}

\author{
Loes Stukken (loes.stukken@ppw.kuleuven.be) \\ Steven Verheyen (steven.verheyen@ppw.kuleuven.be) \\ Gert Storms (gert.storms@ ppw.kuleuven.be) \\ Faculty of Psychology and Educational Sciences. University of Leuven \\ Tiensestraat 102, 3000 Leuven, Belgium
}

\begin{abstract}
Gender differences are not widely studied in the categorization literature and the studies that did focus on gender differences generally investigated processing differences or differences in the use of particular categorization answers (absolute versus continuous). In the following study we looked at differences in the likelihood that men or women consider an item to be part of a category. The objective of the study was twofold: we wanted to introduce a model that is able to determine whether there are meaningful differences in categorization between groups and that is able to identify the sources of these differences. Secondly with this model we wanted to show that there were meaningful categorization differences between men and women: these differences are located at the level of the representation and/or the criterion.
\end{abstract}

Keywords: semantic categorization, threshold theory, gender differences, typicality, similarity, differential item functioning

\section*{Introduction}

Are men more likely than women to consider fishing a sport? And are women more likely than men to consider a dollhouse a member of the category toys? Or in other words are there, for some items, differences between men and women in the likelihood that they would consider an item to belong to a particular category? And if so what is/are the source(s) of this gender difference? In the following study we addressed this question by gathering categorization judgments for 23 exemplars from eight categories, and by analyzing these data with a model that is able to detect differences between men and women in the strictness of the criterion they use to judge an item to be part of the category and differences in the representation that men and women use. The model is a random item mixture model proposed by Frederickx, Tuerlinckx, De Boeck, and Magis (2010), henceforth referred to as the RIM model.

\section*{RIM model}

The RIM model is an item response theory model, Such models assume that the probability that a person endorses an item can be derived from the relative position of the item and the person towards each other on a common latent scale. The more an item's position exceeds the position of the person on the scale the more likely that the person will give a positive answer to the item. Verheyen, Hampton, and Storms (2010) claimed that these models therefore provide
an excellent formalization of the threshold theory proposed by Hampton \((1995,2007)\) in which it is assumed that an item is judged to be part of a category if the similarity of the item to the category exceeds a certain threshold criterion. In this case the item's position on the latent scale represents the item's similarity to the category and the person's position is the threshold criterion the person uses to judge whether the item-category similarity is sufficient for the item to belong to the category.

The RIM model extends this approach in that it is able to account for group differences in two different ways. First of all, the RIM model estimates an average threshold criterion for each group. If the average threshold criterion estimated for women differs from the average threshold criterion for men, women have, depending on the sign of the difference, either a more liberal threshold criterion (they require a smaller item-category similarity than men to judge items as belonging to the category) or a more strict threshold criterion (they require a larger item-category similarity than men to judge an item to be part of the category). Thus the model allows us to detect whether the categorization differences between men and women are due to differences in the threshold criterion that they use to determine whether an item belongs to the category.

Secondly, the model is able to detect differential item functioning (DIF). An item demonstrates DIF when men and women who employ the same threshold criterion, nevertheless are found to have a different probability of endorsing an item. The RIM model allows the positions of these items on the latent scale to differ for different groups of people. The model is thus able to detect whether the position of an item on the latent scale should be different for men and women. The different position of the item indicates that the similarity of the item to the category differs between men and women and thus that men and women, given that they use the same threshold criterion, will have a different probability in judging this item to be part of the category. Different item positions for men and women thus imply representation differences between men and women. The RIM model is thus able to detect whether categorization differences between men and women are due to a difference in the representation and/or in the criterion between men and women.

The model is formally implemented by assuming that a categorization decision for item \(i\) by categorizer \(j\) from group \(g\) is the outcome of a Bernoulli trial with the
probability that the item is judged to belong to the category equal to:
\[
\operatorname{logit}\left(\operatorname{Pr}\left(\mathrm{Y}_{\mathrm{ijg}}==1\right)\right)=\beta_{\mathrm{i}}-\theta_{\mathrm{jg}}
\]

In which \(\beta_{i}\) represents the item i's position on the latent scale and \(\theta_{\mathrm{jg}}\) represents the threshold criterion of categorizer \(j\) from group \(g\). In a categorization context \(\beta_{i}\) can be taken to represent the similarity of item i to the category; \(\theta_{\mathrm{jg}}\) can be taken to represent the required level of item-category similarity to consider an item a category member (Verheyen, Hampton, \& Storms, 2010). The model makes furthermore use of an indicator variable that indicates for each item, whether the item should be considered a DIF item. If so the model estimates a different \(\beta\) for that item for the two groups. If not the model estimates the same \(\beta\) for both groups.

\section*{Gender differences}

Several studies showed that men and women differ in the processing of natural and artificial categories. Women tend to name and recognize members of natural categories faster, while men have an advantage over women in naming and recognizing artificial categories (Barbarotto, Laiacona, Macchi, \& \& Capitani, 2002; Capitani, Laiacona, \& Barbarotto, 1999; Laws, 1999). Based on these studies Pasterski, Zwierzynska, and Estes (2011) argued that women and men might differ in the vagueness of their category judgments since natural and artificial categories tend to differ in vagueness. While membership in many natural categories is considered all-or-none, membership in most artifact categories is found to be graded (Diesendruck \& Gelman, 1999; Estes, 2003, 2004; Verheyen, Heussen, \& Storms, 2011).

Contrary to their initial hypotheses Paterski et al. showed that women provided more vague judgments than men (regardless of category type). They also showed that men, relative to women, gave more inclusive judgments for the artifact categories and tended to give more exclusive judgments for the natural categories.

Our study differs from these studies in that we are not focusing on differences between men and women in the processing of different types of categories or in the type of judgments that men or women give. We are interested in the question of whether or not there are differences in the likelihood/probability that men and woman judge an item to be part of the category. We argue that since men and women are known to be raised differently, to dress differently, to play with different toys, and to engage in different hobbies and professions, we expect that for some items men and women might differ in the likelihood that they consider the item as part of a particular category. To our knowledge this is the first study that looks at gender differences in the likelihood/probability that individual items are part of a category and allows to determine whether these differences reside at the level of the criterion or at the level of the representation.

\section*{Method}

\section*{Materials}

Eight natural language categories were studied (Addictions, Clothing, Diseases, Furniture, Professions, Sports, Toys, Weapons). The categories were selected based on the intuition of the researchers that they might contain items that have a different likelihood of membership in men and women. For each category we included 23 exemplars in the study. The items were selected based on previously collected typicality ratings to make sure that each category contained candidate exemplars that were generally considered typical of the category, atypical of the category, and borderline (items for which people in general are not always sure of whether they belong to the category or not). The typicality ratings were gathered as part of a larger norming project comprising 1276 items from 24 categories. Twenty-nine students ( 23 women, 6 men) provided typicality ratings for half of the categories using a seven point Likert scale ranging from very atypical to very typical. The reliability of these ratings for the \(23 \times 8\) items in our study varied between 0.86 for addictions and 0.96 for clothing with a mean of 0.93 .

\section*{Categorization task}

In total 287 men and 568 women participated in the study. They filled in a questionnaire in which they were, for each item, asked to indicate (yes or no) whether it belonged to the corresponding target category. Participants were, for example, asked whether or not a cold was part of the category diseases. To prevent order effects, we administered 4 different versions of the task with a different order for items and categories. The participants were randomly assigned to one of the 4 versions of the task. The age of the participants ranged between 17 and 64 with an average of 20.

\section*{Model analyses}

Each category's categorization data were analyzed separately using the RIM model. This was done using WinBUGS (Lunn, Thomas, Best, \& Spiegelhalter, 2000) according to the details and code provided by Frederickx et al. (2010). For every analysis 5 chains were run of 10,000 iterations each, with a burn-in sample of 800 .

\section*{Results}

\section*{Typicality}

For every category we calculated the correlation between the items' positions on the scale (the posterior means for the \(\beta_{\mathrm{i}}\) 's) and the items' average typicality to verify whether people were categorizing items by the use of similarity. We calculated the correlation between the items' positions and typicality because it was previously suggested that typicality and item-category similarity are strongly linearly related
(Hampton, 2007; Verheyen, Hampton, \& Storms, 2010). A high correlation between typicality and the items' positions thus also implies a high correlation between the items' positions and the similarities of the items towards the category. We calculated the correlations between typicality and the items' positions for men and women separately. The correlations can be found in Table 1. The correlations were invariably high, suggesting that our participants were indeed using item-category similarity when they judged whether the items belonged to the category.

Table 1
For each category the correlation between item positions and typicality for men and women separately
\begin{tabular}{lll}
\hline Category & Men & Women \\
\hline Addictions & 0.92 & 0.92 \\
Clothing & 0.98 & 0.98 \\
Diseases & 0.92 & 0.95 \\
Furniture & 0.95 & 0.95 \\
Professions & 0.91 & 0.93 \\
Sports & 0.87 & 0.96 \\
Toys & 0.93 & 0.93 \\
Weapons & 0.94 & 0.94 \\
\hline
\end{tabular}

Also note that the correlations of typicality with the items' positions of men and women can hardly be distinguished. Looking at categorization tendencies across the entire typicality range might not be the most fruitful manner to identify differences between groups of categorizers. For natural language categories, whose meaning is to a considerable extent determined by the environment the language community shares, one does not expect pronounced reorganizations of the representation from one group to the other. This would seriously hamper the communication between the group members. Rather, the differences might be more subtle, residing in individual items or in the severity of the employed categorization criterion.

\section*{Criterion differences}

To check whether there were any gender differences in the threshold criterion that participants used to make category judgments, we plotted the posterior distribution of the difference in the average threshold criterion between men and women. If there is a reliable difference in the average threshold criteria, the credibility intervals of this distribution (the region around the mean that contains \(95 \%\) of the mass of the distribution) may not include 0 . As can be seen from Figure 1, this is the case for two categories: professions and toys. In these categories the average differences were 0.34 and 0.74 respectively, indicating that women had a more liberal threshold criterion and require less item-category similarity to judge items to be part of the category than men. For the other categories there is no credible difference in
average threshold criterion indicating that women and men on average require equal levels of item-category similarity for category membership.

\section*{Representation differences}

The RIM model gives an indication of the DIF-status of items by means of latent indicator values that can take one of two values (either DIF or no DIF) on every iteration, resulting in a difference in the estimated item position when required. Following Frederickx et al. (2010) we term an item a DIF item if in more than half of the iterations it was classified as DIF. Table 2 gives an overview of the number of items that were identified as DIF items and the number of items for which men seemed to be more inclined/likely to consider the item to be part of the category and the number of items for which women seemed to be more likely to judge the item to be part of the category. There was one category for which no DIF items were found, the category furniture. For one category, the category clothing, we found only one DIF item: belt was categorized differently by men and women with the same threshold criterion (men were more likely than women to indicate that belt was part of the category). For the other categories the number of DIF items ranged between 2 and 16 and for most of these categories there were both DIF items for which men were more likely to indicate that they were part of the category and DIF items for which women were more likely to indicate that they were part of the category. The categories professions and weapons were the only categories that contained only DIF items for which women were more likely to indicate that they were part of the category. For professions these items were diver, magician, explorer, parachutist, pirate, and inventor. For weapons these were catapult and harpoon.
DIF items were found across the entire range of typicality. Within the DIF items there were items for which people generally agree that it belongs to the category (for example: dollhouse for the category toys), items for which it is not sure whether or not they belong to the category (snooker for the category sports) and items for which it is generally agreed that they do not belong to the category (pirate for the category professions). Thus women and men do not only disagree on items for which there is uncertainty about whether or not they belong to the category, but also on items for which there is general agreement about whether or not they belong to the category.
First of all remember that for the category clothing the RIM model indicated that there is no reliable difference in mean threshold criterion between men and women. So any gender differences in categorization proportions are representation differences according to the model. The model considers only one of these differences meaningful. The model detected only one DIF item (belt, with an average typicality of 4.53).


Figure 1: The posterior distributions of the difference in mean threshold between men and women for the eight categories. The \(95 \%\) credibility interval is represented by the red bars.

Table 2
Overview of the number of DIF items in the categories
\begin{tabular}{lccc}
\hline Category & \# DIF items & Men \(^{\mathbf{1}}\) & Women \(^{\mathbf{1}}\) \\
\hline Addictions & 4 & 1 & 3 \\
Diseases & 10 & 8 & 2 \\
Clothing & 1 & 1 & 0 \\
Furniture & 0 & 0 & 0 \\
Professions & 6 & 0 & 6 \\
Sports & 16 & 5 & 11 \\
Toys & 6 & 4 & 2 \\
Weapons & 2 & 0 & 2 \\
\hline ' \({ }^{\text {' }}\) ? & & & \\
\hline
\end{tabular}
\({ }^{1}\) columns 'Men' and 'Women' represent the number of DIF items for which respectively men and women are more inclined to consider it as part of the category

For the category toys there were several items for which the difference in proportion was determined meaningful after controlling for the threshold criterion. The RIM model indicated that the items pin-ball machine, gocart, coloured pencil, and chalk (manly items); and dollhouse and skipping rope (womanly items) are DIF items. That is, according to the model, the categorization differences one observes for these items are representational in nature. Interesting here are the items comic book and music box, that at first glance have a large and meaningful difference in the categorization proportion between men and women ( 0.60 versus 0.73 and 0.46 versus 0.62 at average typicalities of 4.23 and 5.08 , respectively). The RIM model nevertheless indicates that these are not DIF items. After controlling for the threshold criterion there no longer is a meaningful difference between men and women for these items, indicating that the difference in proportion for these items is entirely caused by the difference in threshold criterion for this category. Indeed, the category of toys was one of the categories for which the RIM model indicated there was a credible criterion difference between men and women.
It is therefore also able to identify items for which at first glance there are no differences when one looks only at the differences in proportions between the groups/sexes. The item, go-cart (average typicality: 5.15), for example, has a very small difference in categorization proportion between men and women ( 0.78 versus 0.76 ), but after controlling for the threshold criterion the item is identified by the model as a DIF item.

These examples should make it clear that it does not suffice to look at categorization proportions alone to determine whether there are group/gender differences in categorization, and that the main contribution of the model is that it is able to disentangle two main causes of differences in categorization proportions: criterion differences and representation differences.


Figure 2: Categorization proportion for men and women as a function of typicality for the categories clothing and toys

\section*{Discussion}

This is, to our knowledge, the first study that looks at differences in the likelihood/probability with which men and women would judge items to be part of the category. We have shown that for some items categorization differences were due to representational differences between men and women and for other items these differences were due to the differences in threshold criterion that men and women use.

The study described above fits in a recent group of studies in which it is shown that item response theory models can be used to analyze data from categorization tasks (Verheyen, Hampton, \& Storms, 2010; Verheyen, De

Deyne, Dry, \& Storms, 2011). It has previously been shown that variations of these models can reveal differences in the threshold criterion between different groups (Verheyen, Ameel, \& Storms, 2011) and reveal latent groups of categorizers who employ a different representation (Verheyen \& Storms, 2013). The RIM model is another valuable approach in that it can not only reveal criterion differences between existing groups, but also representation differences between these groups. It therefore opens up new ways of investigating group differences in semantic categorization.

For instance, a study by Hampton, Dubois, \& Yeh (2006) that investigated categorization differences between groups of categorizers who were categorizing items in different contexts, compared the correlation between categorization proportions and typicality, and the percentage positive responses between the different groups. In our study there were only minor differences in the correlation between the categorization proportions and typicality between men and women, but the model did indicate that there were differences between the two sexes. Furthermore, looking at the percentage positive responses to see whether some groups are using a stricter threshold criterion, might give an imprecise picture of what is going on, since we showed that not all differences in categorization proportions are due to the use of a more or less strict categorization criterion. It should be clear from our results that the RIM model allows for the detection of more subtle but important differences between groups. The implementation of the model is also not limited to two groups. It can easily be extended to investigate data from multiple groups, such as the pragmatic, technical, and neutral context groups in Hampton, Dubois, \& Yeh (2006), cultures (Medin \& Atran, 2004), or age groups (Ameel, Malt, \& Storms, 2008). The RIM model can also be easily adjusted to account for continuous categorization data.

\section*{Acknowledgments}

Steven Verheyen is a postdoctoral fellow at the Research Foundation-Flanders. The authors wish to thank Sofie Frederickx for generously providing details about the RIM model.

\section*{References}

Ameel, E., Malt, B. \& Storms, G. (2008). Object naming and later lexical development: From baby bottle to beerbottle. Journal of Memory and Language, 58, 262-285.

Barbarotto, R., Laiacona, M., Macchi, V., \& Capitani, E. (2002). Picture reality decision, semantic categories and gender: A new set of pictures, with norms and an experimental study. Neuropsychologia, 40, 1637-1653.

Capitani, E., Laiacona, M., \& Barbarotto, R. (1999). Gender affects word retrieval of certain categories in semantic fluency tasks. Cortex, 35, 273-278.

Diesendruck, G., \& Gelman, S. A. (1999). Domain differences in absolute judgments of category membership:

Evidence for an essentialist account of categorization. Pschonomic Bulletin \& Review, 6, 338-346.

Estes, Z. (2003). Domain differences in the structure of artifacual and natural categories. Memory \& Cognition, 31, 199-214.

Estes, Z. (2004). Confidence and gradedness in semantic categorization: Definitely somewhat artifactual, maybe absolutely natural. Psychonomic Bulletin \& Review, 11, 1041-1047.

Frederickx, S., Tuerlinckx, F., De Boeck, P., \& Magis, D. (2010). RIM: A random item mixture model to detect differential item functioning. Journal of Educational Measurement, 47, 432-457.
Hampton, J. A. (1995). Testing the prototype theory of concepts. Journal of Memory and Language, 34, 686-708.

Hampton, J. A. (2007). Typicality, graded membership, and vagueness. Cognitive Science, 31, 355-384.

Hampton, J. A., Dubois, D., \& Yeh, W. (2006). Effects of classification context on categorization in natural categories.
Memory \& Cognition, 34, 1431-1443.
Laws, K. R. (1999). Gender affects naming latencies for living and nonliving things: Implications for familiarity. Cortex, 35, 729-733.

Lunn, D. J., Thomas, A., Best, N., \& Spiegelhalter, D. (2000). WinBUGS: A Bayesian modeling framework: Concepts, structure, and extensibility. Statistics and Computing, 10, 325-337.

Medin, D.L \& Atran, S. (2004). The Native Mind: Biological Categorization, Reasoning and Decision Making in Development Across Cultures. Psychological Review, 111( 4), 960-983 .

Pasterski, V., Zwierzynska, K., \& Estes, Z. (2011). Sex differences in semantic categorization. Archives of Sexual Behavior, 40, 1183-1187.

Verheyen, S., Ameel, E., \& Storms, G. (2011). Overextensions that extend into adolescence: Insights from a threshold model of categorization. In L. Carlson, C. Hölscher, \& T. F. Shipley (Eds.), Proceedings of the 33rd Annual Conference of the Cognitive Science Society (pp. 2000-2005) Austin, TX: Cognitive Science Society.

Verheyen, S., De Deyne, S., Dry, M. J., \& Storms, G. (2011). Uncovering contrast categories in categorization with a probabilistic threshold model. Journal of Experimental Psycholoy: Learning, Memory, and Cognition, 37, 1515-1531.

Verheyen, S., Hampton, J. A., \& Storms, G. (2010). A probabilistic threshold model: Analyzing semantic categorization data with the threshold model. Acta Psychologica, 135, 216-225.

Verheyen, S., Heussen, D., \& Storms, G. (2011). On domain differences in categorization and context variety. Memory \& Cognition, 39, 1290-1300.

Verheyen, S., \& Storms, G. (2013). A mixture item response theory approach to vagueness and ambiguity. Manuscript submitted for publication.

\title{
A delayed discrimination task yields categorical perception of color not only in the right but also in the left visual field
}

\author{
Takashi Suegami (takashi.suegami@psykologi.uio.no) \\ Department of Psychology, University of Oslo \\ 1094 Blindern, 0317 Oslo, Norway \\ Bruno Laeng (bruno.laeng@psykologi.uio.no) \\ Department of Psychology, University of Oslo \\ 1094 Blindern, 0317 Oslo, Norway
}

Samira Aminihajibashi (samiraam@student.sv.uio.no)
Department of Psychology, University of Oslo
1094 Blindern, 0317 Oslo, Norway

\begin{abstract}
Whether categorical perception of color is lateralized in the left cerebral hemisphere (e.g., Gilbert, Regier, Kay, \& Ivry, 2006) or not (e.g., Witzel \& Gegenfurtner, 2011) is still controversial. This ongoing debate, however, has been studied with visual search tasks, which seemed to produce residual laterality effects. The present study assessed whether a delayed discrimination task with divided visual field method, rather than visual search tasks, yields lateralized or bilateral categorical perception of color. The results showed an advantage for between-category discrimination relative to within-category discrimination. Such an advantage, importantly, was obtained in the left visual field as well as in the right visual field. These results suggest categorical perception of color is bilateral and not lateralized. Combining recent studies with visual search tasks (e.g., Witzel \& Gegenfurtner, 2011), our results would provide further evidence for bilateral categorical perception, and thus throw doubt on the laterality effects of categorical perception.
\end{abstract}

Keywords: categorical perception; color perception; delayed discrimination task

\section*{Introduction}

A growing number of studies have shown that categorical perception (better discrimination of the stimuli from adjacent categories than those from the same category) of color (e.g., Roberson \& Davidoff, 2000; Suegami \& Michimata, 2010; Wiggett \& Davis, 2008; Winawer et al., 2007). In a seminal study, Gilbert, Regier, Kay, and Ivry (2006) employed a visual search task, where a colored target was detected faster when the target and distracters belonged to different color categories than when the target was in the same color category as the distracters (i.e., categorical perception). However, such a categorical perception of color was observed only when the target appeared in the right visual field (RVF) but not in the left visual field (LVF). Some successive studies, either behavioral studies (e.g., Drivonikou, et al., 2007) or imaging study with fMRI (Siok et al., 2009), obtained a stronger categorical perception in the RVF relative to LVF. Since the RVF is projected to the left cerebral hemisphere (LH), which is thought to be a center of verbal processing, the results would apparently
reflect the fact that categorical perception is verbally mediated.

Meanwhile, some other studies suggested that lateralized categorical perception in the RVF could be interpreted in terms of other general cognitive mechanisms supported by the LH, rather than simply verbal or language-related processing. Holmes and Wolff (2012), for instance, showed that not only labeled objects but also unlabeled objects produced categorical perception in the RVF on a visual search task. They attributed such a categorical perception in the RVF to an LH's advantage in qualitative or "categorical" processing (cf, Hellige \& Michimata, 1989; Kosslyn et al., 1989; Laeng, Chabris, \& Kosslyn, 2003).

The debate about the origin of the LH-lateralized categorical perception is ongoing and resolving such a debate may be of importance for the understanding of the nature of categorical perception. It should be noted, however, that such laterality effects could also result from some other residual factors. Some recent studies (Brown, Lindsey, \& Guckes, 2011; Witzel \& Gegenfurtner, 2011), for instance, showed that when using the same visual color search tasks as the previous studies but correcting some methodological weakness of previous studies, such as color production or eye movements, yielded categorical perception of color in the LVF as well as RVF.

Another concern could be brought in regard to the tasks; the LH-lateralized categorical perception was predominantly obtained with visual search tasks (e.g., Brown et al., 2011; Drivonikou et al., 2007; Gilbert et al., 2006; Gilbert et al., 2008; Holmes \& Wolff, 2012; Roberson, Pak, \& Hanley, 2008; Siok et al., 2009; Witzel \& Gegenfurtner, 2011). Visual search tasks seem to be suitable for examining categorical "perception," since this kind of tasks would have little memory demands. In the visual search tasks, however, one can also find some factors that might confound the results. In all of the visual search tasks we reviewed, participants were asked to judge which side (i.e., left or right) did the target appear in, and made their responses by hitting the keys associated with target's position (i.e., left or right key). Since judging left or right side is in the nature of
the "categorical" spatial relation processing, which is better processed in the LH (Hellige \& Michimata, 1989; Hellige, Laeng, \& Michimata, 2010; Jager \& Postma, 2003; Kosslyn et al., 1989; Laeng, Chabris, \& Kosslyn, 2003), this kind of tasks would produce residual laterality effects. The visual search tasks, moreover, seem to consist of detection of the target among distractors. Kitterle, Christman, and Hellige (1990) argued that laterality for the spatial frequency processing could be obtained in an identification task (or a delayed discrimination task) but not in a detection task. Although it is not clear whether laterality for categorical perception and these other types of "categorical" processing share the same mechanisms (c.f., Franklin, Drivonikou, Bevis, Davies, Kay, \& Regier, 2008; Holmes \& Wolff, 2012; Suegami \& Laeng, 2013), employing the visual search task, which seems to be another variation of a detection task, might diminish or cancel the lateralization of categorical perception. Thus, one could argue that some recent studies (Brown et al., 2011; Witzel \& Gegenfurtner, 2011) failed to replicate LH-lateralized categorical perception since laterality effects were cancelled out by residual factors.

Thus, the present study aimed to provide further evidence for either LH-lateralized or bilateral categorical perception of color by means of a delayed discrimination task with divided visual field method, rather than the visual search tasks (e.g., Drivonikou et al., 2007; Gilbert et al., 2006). The delayed discrimination task employed here was the classical method for exploring an LH's advantage in categorical processing of spatial relations and patterns (e.g., Hellige \& Michimata, 1989; Kosslyn et al, 1989; Saneyoshi \& Michimata, 2009; Suegami \& Laeng, 2013). Following previous studies (e.g., Siok et al., 2009), four colors with a constant color difference in CIE L* \(\mathrm{u}^{*} \mathrm{v}^{*}\) perceptually uniform color space were emulated. The experiment consisted of an initial training for eye-fixation (GuzmanMartinez, Leung, Franconeri, Grabowecky, \& Suzuki, 2009) and two main tasks: a delayed discrimination task and a color categorization task.

In the delayed discrimination task, moreover, two different lengths of delays ( 500 ms or 5000 ms ) were employed since both theories for the LH-lateralized categorical perception expected that such a LH-lateralized pattern should be enhanced by longer delays. A 5000 ms or longer delay in the delayed discrimination task would enhance using verbal codes rather than visual codes (Posner \& Keele, 1967), and also enhance an LH's advantage in "categorical" spatial relation processing (Postma, Huntjens, Meuwissen, \& Laeng, 2006).

After the delayed discrimination task, the participants also took part in a color categorization task to validate the categories of the four colors. The color categorization task was conducted after the delayed discrimination task in order to avoid any biases to the discrimination task.

\section*{Method}

\section*{Participants}

Thirty participants were recruited as volunteers for an experiment on color perception. Each participant received a gift card for 200 Norwegian Crowns (i.e., about 35 U.S. dollars). Edinburgh Handedness Inventory (Oldfield, 1971) and Farnsworth-Munsell 100-Hue test were conducted for screening out left-handers and individuals with abnormal color vision.

\section*{Apparatus}

All the stimuli were presented on a \(21-\mathrm{in}\). CRT monitor with 75 Hz refreshing rate (EIZO Flex Scan T961), connecting with Apple MacBook Pro ( 2.8 GHz Intel Core 2 Duo ). The distance between the CRT monitor and participant's eyes was fixed in 85.5 cm . The experiment was operated by MATLAB 2008b with Psychophysics Toolbox 3 (Brainard, 1997). A 10-key pad was connected to the computer and served as a response console. Both the training and two main tasks were conducted in a dark room.

\section*{Stimuli}

Eye-fixation training The stimuli in the original training task (Guzman-Martinez et al., 2009) were closely duplicated. Two circles of \(17.27^{\circ}\) (visual angle) diameter, filled with black and white random-dot pattern or its contrast-reversed pattern, were created. Each of the two circle had \(1.00^{\circ}\) by \(1.00^{\circ}\) of a black fixation cross at its center.

Delayed discrimination task Four colors used in the previous studies (e.g., Siok et al., 2009) were emulated. Each adjacent pair had approximately constant distance in the CIE L* \(u^{*} v^{*}\) color space. Two of them ought to belong to blue category and other two to green (hereafter, the four colors were termed as Blue 1, Blue 2, Green 2, and Green 1 respectively). The CIE L* \(\mathrm{u}^{*} \mathrm{v}^{*}\) coordinates for each color were measured by means of Datacolor Spyder 4 ELITE (CIE L* \(\mathrm{u}^{*} \mathrm{v}^{*}\) coordinates for each color were listed on Table 1). The mean color difference ( \(\Delta E\) in CIE L* \(\mathrm{u}^{*} \mathrm{v}^{*}\) space) of within-category pairs was 17.76 , and slightly larger than the \(\Delta E\) for the between-category pair (17.10).

Four color patches of \(2.00^{\circ}\) by \(2.00^{\circ}\) and a hairline fixation cross of \(1.00^{\circ}\) by \(1.00^{\circ}\) were created as the stimuli. Each color patch had one of the 4 colors, and the fixation cross was depicted by the neutral orange color which had approximately constant distance from all of the 4 colors in terms of CIE L* \(u^{*} v^{*}\) coordinates.

Color categorization task The same color patches as the delayed discrimination task were employed as the stimuli.

\section*{Procedure}

Eye-fixation training The procedure was based on the original work by Guzman-Martinez et al. (2009). Each
participant was seated in front of the CRT monitor, and fixed her/his eyes into the fixation cross. Participant's hitting the appropriate key led to 5000 ms of 37.5 Hz flickering presentation of the two random-dots circles. The participant was instructed that random-dots circles would turn into an uniform gray circle if her/his eyes were fixed into the fixation cross. After 5000 ms of flickering presentation, the participant could take a short break, and was allowed to start next trial by her/his own pace. The training had 30 trials and took approximately 5 min .

Delayed discrimination task After the eye-fixation training, the participant took part in the delayed discrimination task. The apparatus was identical to those for the eye-fixation training.

Typical trial sequence was shown in the Figure 1. Each trial began with 200 ms of a fixation cross against black background. Then two identical color patches filled with one of the 4 colors appeared \(3.9^{\circ}\) left and right from the center as probes \({ }^{1}\), and a blank screen followed for 300 ms or 4800 ms . After the blank, a fixation cross appeared again for 200 ms (therefore, the ISI was 500 ms or 5000 ms ), and then target color patch was presented \(3.9^{\circ}\) left or right from the center for 200 ms . The target could have an adjacent color or the same color as the probes. The participant judged if the color of the target was identical to that of the probes by hitting the left or right key as quickly as s/he could. A half of the participants hit the left key if the target and probes had the same color, and the other half hit the right key instead. Response times (RTs) were recorded from the onset of the target. If no response occurred until 2000 ms had elapsed from the onset, the trial was classified as an error. After a response had been made or 2000 ms had elapsed, the next trial started through a 1500 ms of inter trial interval. Twenty trials constituted an experimental block. The length of the ISI was manipulated between the experimental blocks. Half of the participants performed 10 blocks with 500 ms of ISI first, and then another 10 blocks with 5000 ms of ISI second. Another half of the participants performed each 10 blocks in reversed order.

For both ISI conditions, a practice block was held before starting each task. In the practice block, each trial had instant feedback, and another block was repeated if the accuracy rate of the block had not reached \(65.0 \%\).

Color categorization task After the delayed discrimination task, a color categorization task took place. In each trial, a color patch filled with one of the 4 colors appeared in the center. The participant judged if the color was blue or green by manual response. The color patch was presented until the response had been made (with no time limitation). The participant could take a short break after each 20 trials, and the whole task consisted in a total of 100 trials.

\footnotetext{
\({ }^{1}\) The probes were presented left and right from the center of the screen simultaneously, ensuring that the probes and the target had identical retinal eccentricity.
}


Figure 1: Illustration of a trial sequence of the delayed discrimination task.

\section*{Results}

Eight participants were excluded from the analysis. One of them did not get a score above +50 on the Edinburgh Handedness Inventory and therefore was ruled out for excluding potential left-handed or mix-handed (see Dragovik, 2004). Another showed significant positive correlation between accuracies and RTs ( \(\boldsymbol{r}=+.91\) ), suggesting a speed-accuracy trade-off. Three participants were screened out due to their abnormal scores on the Farnsworth-Munsell 100-Hue test. Other three showed atypical categories of color; the boundary between blue and green category did locate at between blue 1 and blue 2, instead of between blue 2 and green 2 (see Gilbert et al., 2006). Thus, the data from the remaining 22 participants were employed for the statistical analysis. Six of them were native Norwegian speakers, five were Lithuanian speakers, two were Chinese and two were English, and each of the other participants spoke, respectively, Bosnian, French, Italian, Persian, Portuguese, Spanish, and Swedish as a native language. Fifteen of them were female, and the mean age of the participants was 27.76 years \((S D=5.32)\). The mean score on the Edinburgh Handedness Inventory was \(91.57(S D=9.06)\).

\section*{Color categorization task}

The rates of the trials in which the color was categorized "blue" were calculated for all of the four colors (Table 1). One-sample \(t\)-tests for each rate revealed that all the rates were significantly different form the chance level of \(50.0 \%\), \(p \mathrm{~s}<.001\). That is, both of blue 1 and blue 2 were categorized as "blue" robustly, and likewise both of green 2 and green 1 were categorized as "green." These results confirmed that blue 1 and blue 2 indeed belonged to "blue" category, and likewise green 2 and green 1 were belonged to "green" category.

Table 1: CIE L* \(\mathrm{u}^{*} \mathrm{v}^{*}\) coordinates and rate of categorized "blue" in the color categorization task for each color.
\begin{tabular}{lcccc} 
Color & \(\mathrm{L}^{*}\) & \(\mathrm{u}^{*}\) & \(\mathrm{~V}^{*}\) & Rate of "blue" response (\%) \\
\hline Blue 1 & 56.655 & -41.973 & -27.120 & \(99.6(0.3)\) \\
Blue 2 & 62.963 & -48.252 & -10.988 & \(86.6(2.9)\) \\
Green 2 & 62.141 & -50.880 & 6.534 & \(1.6(0.7)\) \\
Green 1 & 62.432 & -52.346 & 23.564 & \(0.2(0.2)\) \\
Neutral orange & 27.004 & 40.084 & 16.971 & -- \\
\hline \multicolumn{4}{l}{ Note Standard errors for the rates of "blue" responses are within parentheses. }
\end{tabular}

\section*{Delayed discrimination task}

Accuracy As indices of accuracies, \(A\) 's (Aaronson \& Watts, 1987; Pollack \& Norman, 1964) were employed instead of error rates, so as to exclude possible participants' response bias (see also Pilling et al., 2003). A's for within- and between-category pairs with two ISI conditions were calculated for each visual field (panel \(a\) and \(b\) in Figure 2).

The \(A\) 's were analyzed by a three-way analysis of variance (ANOVA), with category (within-category or between-category), visual field (LVF or RVF), and ISI (500 ms of ISI or 5000 ms of ISI) as within-participants factors. As the most important result, a significant effect of category was obtained, \(F(1,21)=28.40, M S E=0.06^{-1}, p<\) \(.001, \eta_{p}^{2}=.58\), reflecting that the \(A^{\prime}\) for the betweencategory discrimination was larger than that for the withincategory discrimination (i.e., categorical perception of color). A main effect of visual field was also significant, \(F(1,21)=11.12, M S E=0.03^{-1}, p=.003, \eta_{p}^{2}=.35\), revealing that the \(A^{\prime}\) in the LVF was larger than that in the RVF. An interaction between category and visual field, moreover, was significant. Post hoc \(t\)-tests revealed the \(A^{\prime}\) for the between-category discrimination was larger than that for the within-category discrimination in the RVF, \(t(21)=3.75, p=.001, d=0.60\), and also in the LVF as well, \(t(21)=5.53, p<.001, d=1.06\). The \(A^{\prime}\) for the between-category discrimination was larger in the LVF than that in the RVF, \(t(21)=3.90, p=.001, d=0.63\), whereas no significant difference was found between in the \(A^{\prime}\) for the within-category discrimination in the LVF and RVF, \(t(21)=1.01, p=.326, d=0.12\). These results suggest that the categorical perception of color was obtained not only in the RVF but also in the LVF, and, unexpectedly, such a category perception was observed stronger in the

LVF instead of the RVF. Neither any other main effects nor interactions was significant, \(p>.175, \eta_{p}^{2}<.09\).

Response Time Median RTs for correct responses were also calculated for within- and between-category pairs with two ISI conditions in both visual fields (panel cand din Figure 2).

The RTs were also analyzed by the same three-way ANOVA as the accuracies. In line with the results in the accuracies, a main effect of category was again significant, \(F(1,21)=12.78, M S E=9104.76, p=.002, \eta_{p}^{2}=.38\). This main effect reflects that the RTs for the betweencategory discrimination were shorter than that for the within-category discrimination. Post hoc \(t\)-tests revealed the RTs for the between-category discrimination were significantly shorter than that for the within-category discrimination in the RVF, \(t(21)=2.80, p=.011, d=\) 0.28 , and also in the LVF, \(t(21)=2.43, p=.024, d=\) 0.32 . These results, in accordance with the results in the accuracies, suggest that categorical perception of color was found in the LH as well as in the RH. A significant main effect of ISI was also found, not surprisingly, \(F(1,21)=\) \(36.45, M S E=25739.96, p<.001, \eta_{p}^{2}=.63\), revealing that the RTs with 500 ms of ISI were shorter than those with 5000 ms of ISI. Any other effects or interactions failed to be significant, \(p>.203, \eta_{p}^{2}<.08\).


Figure 2: \(A\) 's (two panels on upper line) and RTs (two panels on lower line) for within- and between-category pairs. The left panels show \(A\) 's and RTs with 500 ms of ISI (panel a and \(c\), respectively), whereas the right panels show those with 5000 ms of ISI (panel b and d, respectively). Each error bar shows \(\pm 1\) standard error.

\section*{Discussion}

The present study aimed to assess whether, by means of a delayed discrimination task, one would observe a LH-
lateralized categorical perception of color (e.g., Drivonikou et al., 2007; Gilbert et al., 2006; Siok et al., 2009) or bilateral categorical perception (Brown et al., 2011; Witzel \& Gegenfurtner, 2011). In such a delayed discrimination task, we found evidence for bilateral but not LH-lateralized categorical perception. We employed the same color set as some of the previous studies that had found a LH-lateralized categorical perception (e.g., Siok et al., 2009). The results from the categorization task confirmed that both of blue and green categories in the present study were well established. Thus, such bilateral categorical perceptions could not be attributed to the color difference itself, and they confirm recent studies also showing bilateral categorical perception of color (Brown et al., 2011; Witzel \& Gegenfurtner, 2011). Interestingly, an unexpected RH-lateralized categorical perception was obtained in terms of accuracies. A possible account for this "reversed" laterality effects could be the RH's advantage in color processing (e.g., Levy \& Trevarthen, 1981; Pennal, 1977). Importantly, this result presents another example of counter evidence for the hypothesis of a LH-lateralized categorical perception.

One could argue that the present study failed to replicate the LH-lateralized categorical perception merely due to employing a different task than the typical visual search. This is plausible, but this reasoning contradicts the idea that either an LH's advantage in verbal processing or a hemispheric specialization in "categorical" processing underlies the LH-lateralized categorical perception effect. In fact, an LH's advantage in both verbal (Posner \& Keele, 1967) and categorical spatial processing (Postma, et al., 2006) is typically enhanced with longer time intervals, as in a delayed discrimination task. Moreover, a delayed discrimination task would seem more suitable for obtaining laterality effects than the tasks with no memory demands, like the visual search tasks (Kitterle et al., 1990). Therefore it could be argued that the present task should have been more likely to yield laterality effects on categorical perception than the previous visual search tasks. However, we failed to observe any sign of a LH-lateralized categorical perception in the current study. Thus, the present results are best interpreted as supporting the conclusion that categorical perception of color is represented bilaterally in the brain.

One possible weakness of the present study may be due to the large variety of participants' mother languages in our sample. Several studies have reported that people may possess different color categories in their native languages and consequently this could yield different patterns of categorical perception of color (e.g., Winawer et al., 2007) or its lateralization (e.g., Roberson et al., 2008). However, we took care to confirm that all of our participants shared the same color categories, and we found evidence that only three participants showed atypical categorical structures. Moreover, the native languages of all participants (except excluded three participants) distinguish between green and blue at the lexical level. According to the previous studies (e.g., Roberson, Davies, \& Davidoff, 2000), it is when the native languages do not distinguish, for example, blue from
green that these individuals would also show different patterns of category effects. As mentioned in the Results section, moreover, three participants were excluded since their categorical boundary located at between blue 1 and blue 2, instead of blue 2 and green 2. For these participants, two of them were Russian and the other was Turkish. Previous studies showed that Russian (Winawer et al., 2007) and Turkish (Özgen \& Davies, 1997) have different structures of blue category relative to native English speakers. The fact that those participants with different color structures in their native languages indeed showed different categorical structures in the categorization task indicates that the results of the categorization task adequately reflect participants' categorical structures. Thus, these results could provide another moderate support for that the participants employed in current analysis shared the same category structures to a satisfactory extent.

In conclusion, the present study revealed bilateral categorical perception with a delayed discrimination task. Although the delayed discrimination task has memory components and therefore could be less appropriate for examining "perception," such a task has the advantage to exclude residual laterality effects caused by previous visual search tasks (e.g., residual categorical spatial processing caused by left/right judgments). Combining recent studies employed visual search tasks with correcting some other methodological flaws (Brown et al., 2011; Witzel \& Gegenfurtner, 2011), our results provide further evidence for bilateral categorical perception, and thus throw doubt on the laterality effects of categorical perception.

\section*{Acknowledgments}

This study was supported by JSPS Postdoctoral Fellowship for Research Abroad (H23-794) for Takashi Suegami.

\section*{References}

Aaronson, D., \& Watts, B. (1987). Extensions of Grier's computational formulas for \(A^{\prime}\) and \(B^{\prime \prime}\) to bellow-chance performance. Psychological Bulletin, 102(3), 439-442.
Brainard, D. H. (1997). The psychophysics toolbox, Spatial Vision, 10(4), 433-436.
Brown, A. M., Lindsey, D. T., \& Guckes, K. M. (2011). Color names, color categories, and color-cued visual search: Sometimes, color perception is not categorical. Journal of Vision, 11(12), 1-21.
Dragovic, M. (2004). Categorization and validation of handedness using latent class analysis. Acta Neuropsychiatrica, 16(4), 212-218.
Drivonikou, G., Kay, P., Regier, T., Ivry, R. B., Gilbert, A. L., Franklin, A., \& Davies, I. R. L. (2007). Further evidence that Whorfian effects are stronger in the right visual field than the left. Proceedings of the National Academy of Science of the United States of America, 104, 1097-1102.
Franklin, A, Drivonikou, G., Bevis, L., Davies, I. R. L., Kay, P., \& Regier, T. (2008). Categorical perception of color is lateralized to the right hemisphere in infants, but to the
left hemisphere in adults. Proceedings of the National Academy of Science of the United States of America, 105, 3221-3225.
Gilbert, A. L., Regier, T., Kay, P., \& Ivry, R. B. (2006). Whorf hypothesis is supported in the right visual field but not the left. Proceedings of the National Academy of Science of the United States of America, 103, 489-494.
Gilbert, A. L., Regier, T., Kay, P., \& Ivry, R. B. (2008). Support for lateralization of the Whorf effect beyond the realm of color discrimination. Brain and Language, 105, 91-98.
Guzman-Martinez, E., Leung, P., Franconeri, S., Grabowecky, M., \& Suzuki, S. (2009). Rapid eye-fixation training without eyetracking. Psychonomic Bulletin \& Review, 16(3), 491-496.
Hellige, J. B., Laeng, B., \& Michimata, C. (2010). Processing asymmetries in the visual system. In K. Hugdahl \& W. René (Eds.), The two halves of the brain: information processing in the cerebral hemispheres (pp. 379-415). Cambridge, MA: The MIT press.
Hellige, J. B., \& Michimata, C. (1989). Categorization versus distance: Hemispheric differences for processing spatial information. Memory \& Cognition, 17(6), 770-776.
Holmes, K. J., \& Wolff, P. (2012). Does categorical perception in the left hemisphere depend on language? Journal of Experimental Psychology: General, 141(3), 439-443.
Jager, G. \& Postma, A. (2003). On the hemispheric specialization for categorical and coordinate spatial relations: A review of the current evidence. Neuropsychologia, 41(4), 504-515.
Kitterle, F. L., Christman, S., \& Hellige, J. B. (1990). Hemispheric-differences are found in the identification, but not the detection, of low versus high spatialfrequencies. Perception \& Psychophysics, 48(4), 297-306.
Kosslyn, S. M., Koenig, O., Barrett, A., Cave, C. B., Tang, J., \& Gabrieli, D. E. (1989). Evidence for 2 types of spatial representations- Hemispheric-specialization for categorical and coordinate relations. Journal of Experimental Psychology: Human Perception and Performance, 15(4), 723-735.
Laeng, B., Chabris, C. F., \& Kosslyn, S. M. (2003). Asymmetries in encoding spatial relations. In K. Hugdahl \& R. Davidson (Eds.), The asymmetrical Brain (pp. 303339). Cambridge, MA: The MIT press.

Laeng, B., \& Peters, M. (1995). Cerebral lateralization for the processing of spatial coordinates and categories in left-handers and right-handers. Neuropsychologia, 33(4), 421-439.
Levy, J. \& Trevarthen, C. (1981). Color-matching, colornaming and color-memory in split-brain patients. Neuropsychologia, 19(4), 523-541.
Oldfield, R. C. (1971). The Assessment and Analysis of Handedness: The Edinburgh Inventory. Neuropsychologia, 9(1), 97-113.
Özgen, E., \& Davies, I. R. L. (1997). Do linguistic categories affect colour perception? A comparison of

English and Turkish perception of blue. Perception, 26 supplement, 129.
Pennal, B. (1977). Human cerebral asymmetry in color discrimination. Neuropsychologia, 15, 563-568.
Pilling, M., Wiggett, A., Özgen, E., \& Davies, I. R. L. (2003). Is color "categorical perception" really perceptual? Memory \& Cognition, 31, 538-551.
Pollack, I., \& Norman, D. A. (1964). A non-parametric analysis of recognition experiments. Psychonimic Science, 1(5), 125-126.
Posner, M. I. \& Keele, S. W. (1967). Decay of visual information from a single letter. Science, 158(3797), 137139.

Postma, A., Huntjens, R. J. C., Meuwissen, M., \& Laeng, B. (2006). The time course of spatial memory processing in the two hemispheres. Neuropsychologia, 44(10), 19141918.

Roberson, D., Davies, I., \& Davidoff, J. (2000). Color categories are not universal: Replications and new evidence from a stone-age culture. Journal of Experimental Psychology: General, 129(3), 369-398.
Roberson, D., \& Davidoff, J. (2000). The categorical perception of colors and facial expressions: The effect of verbal interference. Memory \& Cognition, 28, 977-986.
Roberson, D., Pak, H., \& Hanley, R. (2008). Categorical perception of colour in the left and right visual field is verbally mediated: Evidence from Korean. Cognition, 107, 752-762.
Saneyoshi, A., \& Michimata, C. (2009). Lateralized effects of categorical and coordinate spatial processing of component parts on the recognition of 3D non-nameable objects. Brain and Cognition, 71(3), 181-186.
Siok, W. T., Kay, P., Wang, W. S. Y., Chan, A. H. D., Chen, L., Luke, K.-K., \& Tan, L. H. (2009). Language regions of brain are operative in color perception. Proceedings of the National Academy of Science of the United States of America, 106, 8140-8145.
Suegami, T., \& Laeng, B. (2013). A left cerebral hemisphere's superiority in processing spatial-categorical information in a non-verbal semantic format. Brain and Cognition, 81(2), 294-302.
Suegami, T. \& Michimata, C. (2010). Effects of Stroop interference on categorical perception in simultaneous color discrimination. Perceptual and Motor Skills, 110(3), 857-878.
Wiggett, A. J., \& Davies, I. R. L. (2008). The effect of Stroop interference on the categorical perception. Memory \& Cognition, 36, 231-239.
Winawer, A. J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., \& Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. Proceedings of the National Academy of Science of the United States of America, 104, 7780-7785.
Witzel, C., \& Gegenfurtner, K. R. (2011). Is there a lateralized category effect of color? Journal of Vision, 11(12), 1-25.

\title{
Phonetic variation and the recognition of words with pronunciation variants
}

\author{
Meghan Sumner (sumner@stanford.edu), Chigusa Kurumada (kurumada@stanford.edu), Roey J. Gafter (gafter@stanford.edu), Marisa Casillas (middyp@stanford.edu) \\ Department of Linguistics, Margaret Jacks Hall, Bldg. 460 \\ Stanford, CA 94305-2150 USA
}

\begin{abstract}
Studies on the effects of pronunciation variants on spoken word recognition have seemingly contradictory results - some find support for a lexical representation that contains a frequent variant, others, an infrequent (but idealized) variant. We argue that this paradox is resolved by appealing to the phonetics of the overall word. In two phoneme categorization studies, we examined the categorization of the initial sounds of words that contain either tap or [ t ]. Listeners identified the initial sound of items along a voiced-voiceless continuum (e,g, bottom-pottom, produced with word-medial [ t ] or tap). No preference for wordforming responses for either variant was found. But, a bias toward voiced responses for words with [t] was found. We suggest this reflects a categorization bias dependent on speaking style, and claim that the difference in responses to words with different variants is best attributed to the phonetic composition of the word, not to a particular pronunciation variant.
\end{abstract}

Keywords: phonetic variation, pronunciation variation, speech perception, phoneme categorization, lexical representation

\section*{Introduction}

As listeners, we face a speech signal that is riddled with variation, with countless acoustic realizations of any given word. Words stream by listeners at a rate of about 5-7 syllables per second, further complicating the listener's task. How listeners understand spoken words despite this variation is an issue central to linguistic theory.

The finding that lexical representations are rich with phonetic detail along with associated theories of representation and lexical access have greatly advanced our understanding of this process (e.g., Goldinger, 1998; Johnson, 2006). Incorporating variation into theory was a major step toward a full explanation of spoken language understanding. \({ }^{1}\) But, claims made by lexical-representationbased accounts are becoming increasingly difficult to validate or falsify.

Studies that examine the effects of pronunciation variants on spoken word recognition highlight this point. Two different realizations of a sound are considered pronunciation variants. For example, one can produce the word baiting with a [ t\(]\), sounding like bay-ting or with a tap [r], sounding more like bay-ding. Or, one can produce the word center with a [ t ], sounding like sen-ter, or without [ n ] (though some acoustic residue is likely to remain), sounding like sen-ner. Studies that examine the recognition of words

\footnotetext{
\({ }^{1}\) This is a move often discussed, but still largely absent from theories of spoken word recognition; see McLennan \& Luce (2005) for related discussion.
}
with pronunciation variants typically compare a frequent (commonly produced) variant (e.g., [r] or [n_]) to an canonical, but infrequent variant (e.g., [t] or [nt]). Interestingly, in this area of research, two conceptuallyidentical studies have found evidence for lexical representations that are specified for a particular pronunciation variant. In one case, though, the data suggest that the frequent variant is stored (Connine, 2004). In the other case, the data suggest that the canonical variant is stored (Pitt, 2009). We call this the representation paradox. Specifically, these studies found:
(1) Frequency bias: A cost for words produced with [t], like baiting produced like bay-ting, (Connine, 2004) compared to those produced with the more common tap ([r]) variant, and
(2) Canonical bias: A benefit for words with [t], like center produced sounding like sen-ter (Pitt, 2009) compared those produced with the more common postnasal deletion variant ([n_]) (sounding like sen-ner).

In this paper, we suggest that this paradox has resulted for two reasons. First, pronunciation variants are typically examined independent of the phonetic composition of the entire word (see also Andruski et al., 1994). While it is true that we may produce [ t ] or [ r\(]\) in a word like baiting, it is also true that each variant co-varies with a different set of acoustic correlates across the word. Second, in the examples in (1) and (2), it is not clear that listener responses are driven by stored lexical forms in this task, and not by these co-present acoustic cues.

It is undoubtedly the case that detailed representations exist. But, it is also the case that (1) listeners are highly sensitive to acoustic fluctuations in speech (Clayards, Tanenhaus, Aslin, \& Jacobs, 2008; Green, Tomiak, \& Kuhl, 1997; McMurray \& Aslin, 2005; McMurray, Tanenhaus, \& Aslin, 2009), (2) low-level acoustic mismatches result in major perceptual costs either from manipulations resulting in incongruent cues (Gaskell \& Marslen-Wilson, 1996) or from intentionally mispronounced sounds (Gow, 2001, 2003; Sumner \& Samuel, 2005), and (3) acoustic cues inform a listener not only about linguistic units, but provide expectations about the style of a speech event (Labov, 1966; among many others)

In this paper, we ground ourselves broadly in a phonetic perspective and make two suggestions. First, we suggest that different pronunciation variants are processed equally
well when presented in a congruent phonetic word frame. Note that this is not inconsistent with an exemplar account, but suggests only that these representations are not at play here, and the canonical bias in (2) results from an artificial bias toward the canonical variant. Second, we suggest that once we accept that all variants are processed equally well by listeners, we need to reconsider the exaplanatory burden placed on exemplars for theories accommodating variation during spoken word recognition.

\section*{Categorization and pronunciation variants}

As mentioned, current studies diverge on how listeners respond to words with different pronunciation variants in spoken word recognition. The frequent variant is the one uttered by speakers with the highest frequency, and is often regarded as a reduced form, (i.e., [r] and [n_]; see Patterson \& Connine, 2001). The canonical variant is less common in casual conversation, but is more likely to be produced in careful speech, and may be more faithful to orthography (e.g., [t] and [nt]).

Through a series of phoneme categorization studies, Connine (2004) examined the perception of the initial sounds of words that contain either tap or [ t ]. Creating voiced-voiceless continua for words like baiting (baitingpaiting, produced either with word-medial [ t\(]\) or tap), listeners were asked to identify the initial sound of items along the continuum. Listeners made more word-forming responses (in this case, " \(B\) " responses) to items with the frequent tap than to items with canonical \([t]\). She argued that in words like baiting, the tap is stored in the lexical form. Consistent with exemplar accounts of lexical access (Goldinger, 1998; Johnson, 1997, 2006; Pierrehumbert, 2001, 2002), the cost associated with [ t ] is argued to result from access to a frequency-based lexical representation (see also LoCasto \& Connine, 2002).

Through his own series of phoneme categorization studies, Pitt examined post-nasal [t]-deletion in words like center. Comparing responses to items along a centershenter continuum ([nt]) to those along a cenner-shenner ([n_]) continuum, he found that listeners made more wordforming responses (in this case, " S " responses) to items with the canonical [nt] than to items with the frequent [n_]. In this case, the benefit associated with \([\mathbf{t}]\) is argued to result from access to a canonical representation.

Maybe differences encoding words or word forms exist, and there is no paradox at all. One would need to argue that baiting and center are treated differently, and that experience with one yields a surface-based representation and experience with another yields a canonical representation. In this case, tapping and post-nasal [t] deletion are different processes which affect representations differently (e.g. the former could be viewed as an altered form, and the latter a phonological deletion (though a nasal tap may be a residual phonetic cue to the t-deletion process)). Here we offer another alternative explanation: The apparent paradoxical dissolves when we consider the
phonetic composition of the word frame that houses a particular pronunciation variant.

\section*{The phonetic perspective}

Consider again the example of baiting. The use of the [t]variant is constrained by speech style, occurring (though rarely) in extra careful speech (Shockey, 2003). When [t] is used, the entire word (word-level phonetic variation) is hyperarticulated, so that [ t ] co-occurs with other predictable acoustic values (longer stop closure, longer duration of the previous vowels, de Jong, 1998, p. 293). In contrast, the tap, when produced in casual speech, co-occurs with cues common to casual speech (short, centralized vowels, shorter overall duration, reduced amplitude).

Interestingly, the usage patterns of each pronunciation variant pair ([t]-[r]; [nt]-[n_]) differ greatly. The production of tap is nearly categorical in American English (AE), produced nearly \(97 \%\) of the time in running speech (Patterson \& Connine, 2001), typically uttered in a casual frame with approximant-like characteristics, but is so often pronounced that it can occur in a careful phonetic frame (Tucker, 2011). The [t] variant is virtually never uttered, but when uttered, it is paired with a careful phonetic frame. In contrast, the same is not true for post-nasal t-deletion. While rampant in AE, it is less likely as the onset of a prominent syllable (Raymond et al., 2006), so as one shifts to a careful speaking style, post-nasal t-deletion becomes less likely. Critically, the stimuli used in both studies involved different pronunciation variants uttered in controlled, careful phonetic frames, biasing a listener against the frequent-variant in the Pitt study. \({ }^{2}\)

Consider Figure 1, which illustrates 6 different productions of the word beating. Along a hyper-tohypoarticulation continuum (Lindblom, 1991), half of these productions include the phonological variant [ t ], the other half include the phonological variant [ r\(].{ }^{3}\)

\footnotetext{
\({ }^{2} \operatorname{Pitt}(2009\), page 903) mentions that the with-[t] production and deleted-[t] production differ by 55 msec ( 605 vs .550 ), which, when carefully-articulated, is the approximate time needed to produce a voiceless alveolar stop, including release. Connine (2004) mentions that the two were phonetically-controlled, as she spliced the variants into a single token that served as the base form.
\({ }^{3}\) These examples were created by asking a naïve speaker to produce the word beating (extremely carefully; carefully; casually; extremely casually). We created spectrograms for the longest and shortest productions that contained [t] (1, 3; left column) and for those that contained \([r](3,5\); right column). We chose one of many productions in between the endpoints to represent the gradient productions along the continuum.
}


Figure 1. Sample productions of the word beating produced with [t] (left column) or [r] (right column). The schema represents a categorical view of the pronunciation variants, with co-varying phonetic patterns, but also shows that a variant may be natural or forced in a particular phonetic frame.

Figure 1 illustrates the wide-range of variation that appears not only in different productions of each pronunciation variant, but also in the variation across word utterances. The spectrograms in row 3 likely illustrate the stimuli used in the studies discussed, as they are phonetically similar independent of the variant examined.

We are interested in comparing variants, like [t] or [r], in different phonetic frames. In Figure 1, the spectrograms in the middle of each variant-specific continuum reflect the types of phonetic patterns we find in words uttered with one variant or another. We investigate listener responses to words with \([\mathrm{t}]\) that have a more carefully-articulated phonetic word frame to words with [ \([\mathrm{r}\) ] that have a more casually-articulated word frame. The stimuli that exemplify this comparison are marked with black tabs.

Across two experiments, we examine the perception of words with pronunciation variants dependent on the phonetic composition of critical words and fillers. Replicating the methods of prior studies, we examine the perception of word-initial stops of words with either [ t ] or [r], but show that the effects can be attributed more to the phonetic composition of words and fillers (and the expectations and information those provide) than to the pronunciation variants themselves.

\section*{Experiment 1}

In Exp. 1, we investigate responses to words with [t] and tap when the variants are embedded in phonetic frames that
typically co-vary with each variant. We do this for two reasons: First, while this task is based on work by Ganong (1980) showing that lexical status drives categorization responses (listeners make more " T " responses on a taskdask continuum than a tesk-desk continuum), it is not immune to low-level perceptual responses that may also drive categorization (McMurray, et al., 2009; Sumner, 2011). Second, replicating within task as a first step enables us to better interpret past work.

\section*{Methods}

Participants Thirty-five native monolingual speakers of AE participated in this experiment for credit. All were Stanford University undergraduate students. No participants reported any hearing-related issues.
Materials Eight critical words were used in this study. Four words were b-initial (e.g., bottom) and four were p-initial (e.g., pattern). In addition to the critical words, we included seven b-initial fillers, and seven p-initial (three for each onset without \(/ \mathrm{t} /\), believe, police; four for each onset with final /t/, bait, put). Critically, the voiced/voiceless counterpart of all words (critical and filler) resulted in a pseudoword (e.g., bottom/*pottom, believe/*pelieve). The inclusion of fillers served two purposes (1) to control for response bias (Exp. 1) and to include word-external phonetic support for a casual or careful speaking style (Exp. 2). Each word was recorded, along with its voiced/voiceless pseudoword counterpart in two articulation types: Casual speech and Careful speech. This resulted in eight Careful/[t] and eight 1 Casual/[r] critical words. A continuum was created for each word, as described below.
Stimuli Creation From our recordings, we created b-p continua, resulting in a word-pseudoword continuum for each item. To avoid naturalness differences across onsets, all items were manipulated from the nonword base. Using PSOLA in Praat (Boersma \& Weenink, 2008), we then created a 10 -step continuum for each item (from 0 to 45 msec in five msec steps) by increasing or decreasing the amount of aspiration in each word. We should note here that we expect an overall bias toward "P" responses, as on this continuum, there are more responses that typically fall within the English voiceless category.
Design This experiment was designed to examine the proportion of word-forming responses (e.g., "B" for bottompottom continua; " P " for pattern-battern continua) resulting from listening to Careful/[t] words and Casual/[r] words. The design was a \(2 \times 2\) within-subjects design, where the main factors were onset ( \(p, b\) ) and variant/articulation type ([t]/Careful, [r]/Casual).
Procedure Participants completed the task individually or in groups of two or three in a sound-attenuated booth. All 160 critical items ( 8 critical words X 2 articulation types X 10 continuum steps) were randomized with 140 filler items and presented to participants one at a time in isolation over Sennheiser 390 Pro headphones at a comfortable listening level using E-Prime experimental presentation software. Participants were instructed to listen carefully to each word
presented, to decide whether the token they heard began with a P or B , and then press the corresponding button. Response categories were held constant for each participant, but randomized across participants, so the "B" button appeared equally on the right and in the left across participants. A new trial began one second after a response was recorded, and three seconds if no response was made.
Predictions Evidence for the frequency bias should result in more word-forming responses to words with tap than to words with [ t\(]\). Evidence for the canonical bias should result in more word-forming responses for words with [t] than words with tap. Evidence that this task better reflects pre-lexical responses independent of specified lexical representations should yield some pattern that reflects an influence of the phonetic frame of the words.

\section*{Results and Discussion}

A mixed logit regression analysis was employed to predict the participants' word-forming responses. We report the results for the model with maximum random effect structure justified by the data based on model comparison (Jaeger, 2008), which contained random by-subject and by-item intercepts. Initial analyses were based on the proportion of word-forming responses, following past work, and show no main effect of articulation type ( \(\beta=.11 p>.47\) ). A closer look revealed that responses to b-initial words differed dramatically from those to p-initial words. Specifically, binitial words resulted in a higher proportion of wordforming responses for Careful/[t] words than for Casual/[r] words ( \(\beta=1.1 p<.002\) ). Mean proportions of word-forming responses are provided in Figure 2. The onset-based differences suggest that when collapsing across onset, the effects cancel each other out.


Figure 2. Two plots for resulting data depending on response label: Proportion word-forming responses (left); Proportion voiced responses (right).

The data pattern is unexpected if responses are driven by the pronunciation variant. Were these effects due to the activation of a lexical representation with one pronunciation variant or the other, we would expect the Careful/[t] and Casual/[r] items to behave similarly for each onset with respect to word-forming responses. This is not the case. Here, a " B " response to b-initial words is consistent with
both a "word" response and a "voiced" response (e.g., I heard a [b] not a [p]). For p-initial words, a "P" response corresponds with a word-forming response, but not a "voiced" response.

The right panel of Figure 2 plots the data by proportion voiced responses. Any influence of co-varying phonetic cues present in the congruent word frames is likely to surface independent of the lexicon-as a phonetic bias. Analyzing the data in terms of proportion voiced responses reveals that Careful [ t ] items result in a higher proportion of voiced responses (" \(B\) " regardless of lexical status) than Casual/[r] items ( \(\beta=-.5 p<.003\) ). This suggests that listener responses depend on the phonetics, not on access to a stored lexical representation. A higher proportion of " B " responses reflects a different categorization boundary between the two articulation types, with more aspiration (longer VOT) required to prompt a " P " response in careful speech, resulting in a higher proportion of " B " responses.

One implication is that the pronunciation variants have little to do with the response patterns in this paradigm. The high number of " \(P\) " responses for p -initial words likely reflects combined influences of the asymmetrical breakup of the VOT continuum in English, and lexical status. In order for the variant effects to be attributed to lexical representations, the patterns of responses to words with [t] and words with tap must behave similarly across onsets, predicted by a view where the most accessible lexical representation is the best match to the incoming signal (Johnson, 2006). We would expect results analogous with the lexical effect; which we do not find.

If this paradigm is capturing phonetic responses rather than lexical responses, then we should reconsider claims made about the nature and activation of variant-dependent lexical representations more broadly. Certainly, it is now fact that listener memory for auditory events is detailed. But, this does not imply that all accommodation of variation is handled at the level of the lexicon. One prediction a phonetic account makes is that as the articulation type becomes more predictable, the phonetic categorization bias should be more robust. For example, if item presentation were to be blocked by articulation type, listeners would have information about the speech style well before each critical item. We cast the effect as a category boundary difference mediated by word-level phonetic variation. Therefore, the effects are not due to lexical activation. We predict, then, an increase in evidence of a speech style will reinforce the different VOT thresholds, and will result in a greater difference between the two articulation types.

\section*{Experiment 2}

Our goal in Exp. 2 was to increase the predictability of a particular articulation type. One prediction of our claim that the basic effects are driven by the phonetic composition of the words and not by the pronunciation variant is that effects should fluctuate as evidence of a particular speech style increases. Blocking the stimuli by articulation type enabled us to test this prediction.

\section*{Methods}

Participants Thirty-four native monolingual speakers of AE participated in this experiment for credit. All were Stanford University undergraduate students. No participants reported any hearing-related issues.
Materials The stimuli from Exp. 1 were used.
Design The design was identical to Exp. 1 with one exception: Stimuli were blocked by articulation type (careful vs. casual). Block order was randomized, as was the presentation order of items within a block.
Procedure The procedure was identical to Exp. 1.

\section*{Results and Discussion}

Using the same statistical approach as in Experiment 1, we analyzed the data to predict proportions of voiced responses. Proportions of voiced responses by condition are provided in Table 1.

Table 1. Proportion of voiced responses from Exp. 2.
\begin{tabular}{|c|c|c|c|c|c|c|}
\hline Typ e & Condition & N & Prop or tion Voiced Responses & Stan dar d Deviation & Standard Error & Confidence Interval \\
\hline \multirow[t]{2}{*}{B-initial} & Cas ual /[] & 1119 & 0.235 & 0.424 & 0.012 & 0.024 \\
\hline & Careful/[t] & 1117 & 0.286 & 0.452 & 0.013 & 0.026 \\
\hline \multirow[t]{2}{*}{P-initial} & Cas ual /[] & 1141 & 0.105 & 0.306 & 0.009 & 0.017 \\
\hline & Car eful/[t] & 1141 & 0.175 & 0.380 & 0.011 & 0.022 \\
\hline
\end{tabular}

We find that participants are more likely to respond " \(B\) " in the Careful/[t] condition than in the Casual/[r] condition ( \(\beta\) \(=-.76 p<.002\) ). To investigate the phonetic effects across experiments, we conducted an additional analysis on the first 100 items for all conditions across Exp. 1 and Exp. 2. The first 100 trials were examined to minimize the influence of learning throughout the experiment. The first 100 trials give us the best picture of participant responses dependent on the nature of the filler items. The proportion of voiced responses for the first 100 items across experiments and conditions are provided in Figure 3


Figure 3. Proportion voiced responses to Careful/[t] items and Casual/[r] items collapsed across onset type for Exp. 1 (left) and Exp. 2 (right) for first 100 trials.

The difference suggests that the word-external information available serves to stabilize different categorization criteria, resulting in a higher rate of voiced responses in the Careful/ \([\mathrm{t}]\) condition than that found in Exp. 1.

\section*{General Discussion}

We began this study based on our observations that (1) effects of pronunciation variants are typically examined independently from the phonetic composition of the wordframe in which they are uttered and (2) accounts of opposite pronunciation variant effects that intuitively seem incompatible with each other are both viable under an exemplar-theoretic interpretation. While it is accepted and verified that lexical representations are rich with phonetic detail, we sought to investigate phonetic effects in speech perception independent of the lexicon.

To do this, we investigated pronunciation variants that are embedded in congruent phonetic frames. We then examined the responses made to voiced- and voiceless-initial words when presented in a single block with careful and casual speech styles mixed (Exp.1). Finally, we strengthened the expectations based on speech style by blocking the stimuli by articulation type (Exp.2).

In Exp. 1, considering responses made by listeners as word-forming caused some difficulty. The data are more easily accounted for by considering the responses as voicedvoiceless, not as word-forming or pseudoword-forming. In a careful word frame, listeners require a longer VOT before they will switch to a "P" categorization than in a casual frame. Alternatively, this could be driven by an increased likelihood to press "P" at the slightest hint of aspiration in casual word frame. \({ }^{4}\) In Exp. 2, we found that increasing the likelihood of a carefully-articulated word (via critical items with phonetically-congruent fillers) increased voiced responses compared to Exp. 1.

One implication of this work is that the canonical bias is, in part, artificially bolstered by our comparisons. And, reconsidering past work, there is support for this notion. A number of studies that have found a canonical effect examine a frequent variant embedded in an incongruent phonetic frame (Andruski, et al., 1994; Gaskell \& MarslenWilson, 1996). Our data show that the effects here, and likely in some number of previous studies, are due more to congruence between a phonetic frame and a pronunciation variant and to expectation-based categorization than to the activation of a particular lexical representation (or of a more available lexical representation, if we assume there are within-word phonetic clouds). The next step is to consider how frequency-based accounts of lexical access are separate from and integrated with the pre-lexical processes listeners use to navigate a variable speech stream.

\footnotetext{
\({ }^{4}\) While we cannot distinguish the two here, both are compatible with a phonetic explanation of the data rather than one dependent on the pronunciation variants.
}

\section*{Conclusion}

We have discussed one limit of exemplar-accounts of variation effects, and have tailored our investigation to examine an apparent paradox in the literature in which two representative studies account for opposing data with the same broad representation-based interpretation. We highlighted both the ways in which phonetic variation might interact with pronunciation variants in speech production, and presented two experiments aimed at understanding the effects of this interaction. As listeners exhibited a strong bias toward voiced responses for Careful/[t] tokens, amplified by within-speech style blocking, we suggest that the difference between the conditions is entirely due to the phonetic composition of the word, absent the influence of detailed lexical representations.

\section*{Acknowledgments}

We are grateful to Benjamin Lokshin for his assistance with data collection, and to Jason Grafmiller and Kyuwon Moon for valuable comments and discussion. This material is based upon work supported by the National Science Foundation Grant BCS - 0720054 to Meghan Sumner. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

\section*{References}

Andruski J. E., Blumstein S., \& Burton M. (1994). The effect of subphonetic differences on lexical access. Cognition, 52, 163-187.
Boersma, P., \& Weenink, D. (2011). Praat: doing phonetics by computer [Computer program]. Version 5.0.43, retrieved from http://www.praat.org/.
Clayards, M., Tanenhaus, M.K., Aslin, R.N., and Jacobs, R.A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. Cognition, 804-809.
Connine, C.M. (2004). It's not what you hear but how often you hear it: On the neglected role of phonological variant frequency in auditory word recognition. Psychological Bulletin and Review, 11, 1084-1089.
de Jong, K. J. 1998. Stress-related Variation in the Articulation of Coda Alveolar Stops: Flapping Revisited. Journal of Phonetics, 26, 283-310.
Gaskell, M. G. \& Marslen-Wilson, W. D. (1996).
Phonological variation and inference in lexical access. Journal of Experimental Psychology: Human Perception and Performance, 22, 144-158.
Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. Psychological Review, 105, 25179.

Gow, D. W. Jr., (2001). Assimilation and anticipation in continuous spoken word recognition. Journal of Memory and Language, 45, 133-159.
Gow, D. W. Jr., (2003). Feature parsing: Feature cue mapping in spoken word recognition. Perception \& Psychophysics, 65, 575-590.

Green, K. P., Tomiak, G. R., Kuh1, P. K. (1997). The encoding of rate and talker information during phonetic perception. Perception and Psychophysics, 59, 675-692.
Jaeger, T. F. (2008). Categorical Data Analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models. Journal of Memory and Language 59, 434-446.
Johnson, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. Journal of Phonetics, 34, 485-499.
Labov, W. (1966). The social stratification of English in New York City. Washington, DC: Center for Applied Linguistics.
Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H and H theory. In: Hardcastle, W \& Marchal, A (Eds.), Speech production and speech modeling, (pp. 403-439). Dordrecht: Kluwer.
LoCasto, R. C., \& Connine, C. M. (2002). Rule-governed missing information in spoken word recognition: Schwa vowel deletion. Perception \& Psychophysics, 64, 208219.

Luce, P. A., \& McLennan, C. (2005). Spoken word recognition: The challenge of variation. In D. B., Pisoni \& R. E. Remez (Eds.), Handbook of Speech Perception, pp 591-609.
McMurray, B., and Aslin, R.N. (2005). Infants are sensitive to within-category variation in speech perception. Cognition, 95, B15-B26.
McMurray, B., Tanenhaus, M.K., and Aslin, R.N. (2009). Within-category VOT affects recovery from "lexical" garden paths: Evidence against phoneme-level inhibition.
Patterson, D. \& Connine, C.M. (2001). Variant frequency in flap production: A corpus analysis of variant frequency in American English flap production. Phonetica, 58, 254275.

Pierrehumbert, J.B. (2002). Word-specific phonetics. Laboratory Phonology, 7, 101-139.
Pierrehumbert, J.B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee \& P. Hopper (eds.), Frequency and the emergence of linguistic structure (pp. 137-157). Amsterdam: Benjamins.
Pitt, M.A. (2009). The strength and time course of lexical activation of pronunciation variants. Journal of Experimental Psychology: Human Perception and Performance, 35, 896-910.
Shockey, L. (2003). Sound patterns of spoken English. Malden, MA: Blackwell.
Sumner, M. (2011). The role of variation in the perception of accented speech. Cognition, 119, 131-36.
Sumner, M. and Samuel, A.G. (2005). Perception and representation of regular variation: The case of final-/t/. Journal of Memory and Language, 52, 322-338.
Tucker, B.V. (2011). The effect of reduction on the processing of flaps and \(/ \mathrm{g} /\) in isolated words. Journal of Phonetics, 39, 312-318.

\title{
Dyadic Cooperation Enhances Retrieval and Recall of Crossword Solutions
}

\author{
Janelle Szary (jszary@ucmerced.edu) \\ Rick Dale (rdale@ucmerced.edu) \\ Cognitive and Information Sciences, University of California, Merced, 5200 North Lake Road Merced, CA 95343 USA
}

\begin{abstract}
The benefits of collaborative activities have been demonstrated in many domains, but there remain mixed results across several others as to whether collaborative groups can achieve greater performance than individuals, and can achieve greater performance than nominal group comparisons. Here we develop a task that is especially suited to testing collaborative gains. In a collaborative crossword game, two individuals solved puzzle questions either alone or collaboratively through discussion. When talking, participants solved more puzzle questions, solved them more quickly and accurately, and in general seemed to recall the words from collaborative contexts better than from matched independent contexts. By extracting the audio of their interaction, we also demonstrate interesting relationships between spoken interaction and performance on the collaborative tasks. This task environment further substantiates the notion that, in the context of knowledge retrieval, two heads are better than one.
\end{abstract}

Keywords: Dyadic cooperation; collaborative recall.

\section*{Introduction}

Knowledge can be thought of as a probabilistic distribution. As samples are repeatedly taken from this distribution, a more complete picture emerges of the underlying knowledge. Often, as is implied by the phrase "the wisdom of crowds", the probability distribution is quite accurate with respect to its information representation-so that as samples are collected, an increasingly accurate picture emerges. For example, when eight-hundred attendees of a stock and poultry exhibition were asked to estimate the weight of a large ox, the mean of their estimates was very accurate (Galton, 1907). The error of the mean estimate was in fact much lower than the mean error of each individual's estimate. This "wisdom of crowds" effect has continued to be demonstrated in a number of domains: aggregate financial forecasts tend to be better than individual forecasts (Clemen, 1989), polls of the audience in game shows tend to reveal correct answers (Surowiecki, 2004).

The probabilistic nature of knowledge is also apparent when an individual accesses his or her own knowledge over time. When individuals were asked to make quantitative estimates of worldly information on two separate instances, the average of their estimates tended to be more accurate than either individual estimate (Vul \& Pashler, 2008). When multiple individuals work interactively on a joint decision, however, the "two heads are better than one" intuition does not always hold. In general, the literature on group performance shows that groups rarely outperform their best members-the whole is rarely greater than the sum of its parts (Bahrami et al., 2010; Hastie \& Kameda, 2005; Kerr \&

Tindale, 2004). In fact, across a large number of contexts, individuals tend to remember less when they're working with others (Rajaram \& Pereira-Pasarin, 2010).

In these studies, subjects are usually presented with a list of items and must study and reproduce the items either individually or as a group. On average, groups tend to recall more items than individuals, but recall fewer items than nominal groups (consisting of the pooled, non-overlapping items recalled by individuals working alone; Barnier, Sutton, Harris \& Wilson, 2008). That is, individuals working in a group context don't perform at their full potential. The leading explanation for this observation is the retrieval disruption hypothesis (Basden, Basden, Bryner \& Thomas, 1997). According to this hypothesis, individuals use their own, idiosyncratic, strategies to organize and encode information. When recall takes place in an interactive context, the output of one member disrupts the retrieval strategies of the other(s), inhibiting recall performance.

The large body of empirical work providing evidence for the detrimental effects of collaboration on memory is unified by the term social contagion research (Barnier, Sutton, Harris \& Wilson, 2008; and see Rajaram \& PereiraPasarin, 2010, for a review). In addition to disrupting the recall of correct items, collaborators can even introduce the recall of incorrect items. When a confederate collaborator misleadingly recalled an incorrect item, participants later recalled that item themselves, as if it had been in the original recall list (Roediger, Meade \& Bergman, 2001). This effect extends beyond laboratory recall studies, as individuals can often misremember important life events. Loftus has worked extensively on issues surrounding the fallibility of memory, especially as it applies to false memories and eyewitnesses, showing that social context can significantly impact the accuracy of memory (Loftus, 1996).

A related example of the negative consequences of social context is groupthink-a phenomenon where groups of people may end up making poor decisions, generally because of a motivation to reduce conflict and reach consensus, therefore failing to continue the search for an optimal solution (see Esser, 1998). This collaborative inhibition may be related to both retrieval disruption or social loafing (reduced effort or motivation when in a group context; Weldon, Blair \& Huebesch, 2000).

Despite the abundance of theories and supporting evidence for social contagion, there exists an intuitive feeling that we should benefit from working with others. In addition to social contagion research, Barnier and colleagues (2008) define two other approaches to
investigating the effects of social context on memory: collaborative recall, and transactive memory. These approaches tend to seek out the beneficial effects of social context. In collaborative recall research, the social context is conceptualized as part of a broader environmental and situational context which can facilitate an individual's recall through priming. This priming could be detrimental, such as in retrieval disruption, or could be beneficial through cueing or triggering of correct information.

Bahrami and colleagues (2010) found that group performance interacted dynamically with social context. They designed a low-level perceptual decision-making task where members of a dyad reported their own decisions then agreed on a joint decision to report. When members of a dyad had unequal performance levels, the dyad tended to do worse overall than the better-performing member. However, performance exceeded aggregate individual performance when members had equal visual sensitivities and could communicate openly to discuss their observations (Bahrami et al., 2010), and when they used similar task-relevant linguistic forms (Fusaroli et al., 2012). In order to come to an agreement regarding an ambiguous low-level stimulus, members of a dyad must attempt to communicate subjective and graded confidence levels. The combination of information for higher-level decision-making tasks, such as those involving knowledge and memory, may be very different. For example, if two friends are attempting to recall the Spanish word for "countryside" from a long-ago language course, one may offer: "I think it was something like camping", which may trigger the other to remember the correct "campo." In this sense, members of a dyad can prime each other and iteratively build greater information.

Finally, in transactive memory research, the group is conceptualized as the unit of analysis: individuals are components of a coupled, distributed memory system (Wegner, 1987). In these transactive memory systems, group members may share the tasks of encoding, storing, or retrieving information in any combination. Wegner (1987) notes that memories are connected concepts-such as the concept "tomato" with the concept "red"-and these connections are formed through encoding, which can be done at the group level. As an example, consider a couple discussing the odd behavior of a mutual friend. The male partner mentions that their mutual friend seemed quiet at a recent party, while the female partner instead thought he seemed overly friendly. This reminds the man that their mutual friend had been thinking about splitting from his wife, which leads the couple to conclude that their mutual friend had been flirting with the female partner, and subsequently acted awkwardly around the male partner (from Wegner, Giuliano \& Hertel, 1985). Through collaboration (discussion), the couple in this example was able to bind information and encode a quantitatively and qualitatively different memory than either would have achieved individually. Conceptualizing the distributed storage of memories is more intuitive: We already store much of our information externally (books, to-do lists, smart
phones), and in much the same way we could rely on a partner to remember something for us (essentially 'outsourcing' the storage of that information to another person).

From the perspectives of both the collaborative recall and the transactional memory traditions, the performance of a group can come to be greater than the performance of its members. In this paper, we work from these intersecting perspectives to investigate the potential benefit of working with two minds instead of one on a knowledge-based trivia task. Individuals are randomly assigned to dyads and given trivia questions, which they solve either independently or collaboratively. These general knowledge trivia questions provided a set of stimuli on which subjects' knowledge varied widely, and allowed for rich discussions during collaborative sessions. Following four rounds of ten trivia questions, subjects were given individual recall tests for the answers to the preceding trivia questions.

As described by Hare (1976), research on social influence can be characterized by two factors: the "social climate", which could be either individuals collaborating or individuals working independently; and the "task completion", which is a measure of either the group product or the individual product. Consistent with previous work on joint performance measures (i.e., Hill, 1982), the current study design allowed us to first compare [1] the group product of collaborating individuals (group performance on collaborative trivia rounds) to [2] the individual product of individuals working alone (individual performances on independent trivia rounds). The recall task allowed us to compare [1] the individual product of collaborating individuals (individual recall of trivia items from collaborative rounds) to [2] the individual product of individuals working alone (individual recall of items from independent rounds).

By analyzing task performance and efficiency at the group and individual levels, and resultant memory at the individual level, we substantiate the beneficial gain of collaborative cognitive performance. Our results suggest that in knowledge-based tasks, two heads are indeed better than one.

\section*{Methods}

Sixty two participants were recruited from a subject pool of University of California, Merced, undergraduate students who participated for course credit. The participants had an average age of \(19.6(S D=1.7)\) and were mostly female (16 male; 46 female). The participants were grouped into thirtyone dyads. Each dyad participated in four rounds of a trivia game, where each round of ten questions was to be solved individually or collaboratively, followed by a surprise recall task after all four rounds.

Participants were seated directly across from each other at a small table with IBM ThinkPad laptop computers. This allowed each participant to have a private workspace during the independent tasks, but also enabled easy communication during the collaborative tasks. Participants wore Shure Beta

54 supercardiod microphone headsets, and their conversations were recorded using an M-Audio MobilePre recording interface and Audacity software.


Figure 1: Experimental setup.

\section*{Materials}

Trivia questions were collected from a variety of crossword puzzles from www.bestcrosswords.com. Questions were all straight-forward (i.e., not "cryptic") type clues. In total, 140 questions were collected with types that were categorized as culture ( \(n=23\) ), general knowledge ( \(n=21\) ), definitions ( \(n\) \(=27\) ), logic ( \(n=22\) ), fill-in-the-blank (FITB, \(n=20\) ), categories \((n=16)\), and sayings \((n=11)\). Table 1 gives examples of each type.

Table 1: Example trivia types.
\begin{tabular}{lll}
\hline Type & Question & Answer \\
\hline Culture & "Kill Bill" star Thurman & Uma \\
Knowledge & U.S. spy organization & CIA \\
Definition & Gift to charity & Donation \\
Logic & Hour subunits & Minutes \\
FITB & "If all fails" & Else \\
Categories & Tulips and irises, for example & Flowers \\
Sayings & "Rolling in dough" meaning & Rich \\
\hline
\end{tabular}

The trivia questions were normalized for difficulty. 449 University of California, Merced undergraduate students with an average age of 18.4 ( \(S D=1.4 ; 200\) male, 249 female) were given surveys containing trivia questions. There were 10 versions of the survey, each of which contained 14 trivia questions with lines indicating the number of letters the answers. Participants were allowed to leave answers blank, but were instructed to do the best they could to answer to each question, guessing when possible. Results showed that questions varied widely in difficulty (see Fig. 2). For the present study, 40 questions were chosen that were answered correctly about half of the time. As shown in Figure 2, these trivia questions were solved by 45\(77 \%\) of participants, and they contained all types: culture ( \(n\) \(=6\) ), general knowledge \((n=8)\), definitions \((n=4)\), logic ( \(n\)
\(=8\) ), fill-in-the-blank (FITB, \(n=8\) ), categories ( \(n=2\) ), and sayings \((n=4)\). The examples in Table 1 were each used.


Figure 2: Question norming. Potential questions are ranked by the percentage of participants who answered correctly.

Dotted lines show the question rankings we used.
Trivia Program The experimental interface was programmed by the authors using Adobe Flash CS5. The program led participants through four experimental blocks (rounds) containing ten questions each. For each round, the program instructed participants to work either individually (I) or collaboratively (C). During collaborative sessions, participants were asked to work together and discuss each answer as a team. Across all subjects, the order of questions and condition (I-C-I-C or C-I-C-I) was randomized and counterbalanced between dyads, but was kept the same within each dyad.

Each question was provided alone on the screen with a sequence of blank squares indicating the number of letters in the answer. The space-bar was used to submit answers, and subjects were given feedback about their submission. If correct, a green checkmark appeared briefly before moving on to the next question. If incorrect or missing, a red " X " marked the incorrect or blank boxes. Subjects were given 20 seconds to correctly answer each question (with as many tries as necessary) before being automatically moved on to the next question. Between blocks, subjects were given the new condition and asked to wait for their partners before moving on. Progress was indicated using flip cards with "Working" on one side, and "Ready when you are!" on the other (see Fig. 1).

\section*{Procedure}

Participants were given five minutes to introduce themselves at the beginning of the study, in order to facilitate comfort and camaraderie (consistent with previous findings that more familiar groups tend to perform better on collaborative tasks; Barnier et al., 2008). After this brief familiarization period, headsets were fitted and the Flash program was started. The program began with instructions, which the researcher read aloud and subjects read on their respective screens, then the researcher left the room. After
completion of the four trivia rounds, subjects removed their headsets and summoned the researcher. The trivia program was closed and each subject was given a blank text editor. Subjects were instructed to recall and type as many of the answers to the previous trivia questions as possible. They were given five minutes and asked to work individually.

\section*{Results}

Thirty-one dyads participated in the experiment, but one dyad's audio was not recorded due to equipment error. Thus, task performance results are given for thirty-one dyads, while the audio results reflect thirty dyads.

For each question, the Flash program recorded (1) whether a correct answer was submitted before time ran out. If a correct answer was achieved, it also recorded (2) how much time elapsed from the beginning of the trial to the submission of the correct answer, in milliseconds, and (3) the number of incorrect attempts before the final, correct submission. Because each participant worked on his own computer, two independent data sets were collected for each dyad. For purposes of data analysis, results for each trial were averaged over the members of the dyad. These aggregated results were used to compare each dyad's performance on individual versus collaborative rounds. Dyads are independently sampled (though, individual performance is not, as one is not independent of one's partner), and hence at the dyad level, conditions (I vs. C) can be compared using paired-samples \(t\)-tests (unless otherwise noted below). \({ }^{1}\)

\section*{Trivia Performance}

On all three aggregate measures, collaborative dyads outperformed their non-collaborative counterparts. Out of the twenty questions presented in each condition, the average correctly answered by collaborative dyads was \(14.94(S D=3.77)\), while the average correctly answered by non-collaborative dyads was \(12.35(S D=3.11)\). This difference was significant, \(t(30)=5.58, \mathrm{p}<.0001\). Dyads were also faster to submit correct answers while they were collaborating ( \(M=5527 \mathrm{~ms}, S D=1212 \mathrm{~ms}\) ) as compared to when they were not collaborating \((M=6611 \mathrm{~ms}, S D=\) 1181 ms ), and this difference was also significant, \(t(30)=\) 3.17, \(p<.005\). Finally, the number of incorrect attempts made before achieving a correct answer was smaller for collaborative dyads ( \(M=.26, S D=.16\) ) than for noncollaborative dyads \((M=.61, S D=.27)\), which is also significant, \(t(30)=7.19, p<.0001\).

Thus, working collaboratively conferred benefits on all three measures of task performance: it increased the number, speed, and accuracy of successful submissions. Figure 3 shows the performance gain results, where gain for each dyad is calculated as average performance on collaborative rounds, minus average performance on noncollaborative rounds.

\footnotetext{
\({ }^{1}\) We also examined individual-level performance across most measures, and results are consistent with the dyad-level analyses.
}


Figure 3: Collaboration gains for the following measures: (a) average number of correct answers, (b) average time taken to achieve a correct answer (ms), (c) average number of incorrect attempts, per question. Gain for each dyad is calculated as the difference between aggregate performance on collaborative versus non-collaborative rounds. All points above \(x=0\) show dyads benefitting from collaboration. For illustration, dotted lines show median ranked dyads.

\section*{Recall}

The list of recalled items for each participant was first checked for accuracy and incorrect recalls were removed. This was relatively rare, however, as incorrect recalls represented only \(5.7 \%\) of the total recalled items across participants (36 out of 629). Each recalled item was matched to the round and condition in which it was encountered. At the group level (i.e., averaged within dyads), the average number of items recalled from each round was, respectively, \(1.60(S D=.74), 2.27(S D=1.35)\), \(1.97(S D=.91), 3.71(S D=1.57)\). Items from the last round were recalled significantly more often than any other round, \(t(30)=4.25, p<.001\), indicating a serial position effect of recency. Although the mean recall from the first round was the lowest, there was also evidence of a serial position effect from primacy. This pattern is shown in Figure 4, which plots the number of recalled words from each round, binned by the number of individuals recalling each number of items. A generalized linear model, fit to the data, shows both the recency and the (more subtle) primacy effects.

In general, subjects tended to remember more items from the rounds in which they participated collaboratively. Figure 5 shows ranked, aggregated dyads' recall from each round, separated by condition. For each round there was a tendency towards enhanced recall from collaboration, but this difference was only significant in the fourth round, \(t(28.88)\), \(p<.05\) (Welch's two-sample \(t\)-test). Overall, group level
recall was not significantly better for items from collaborative rounds \((M=5.24, S D=2.35)\) compared to non-collaborative rounds \((M=4.31, S D=2.00)\). At the individual level, however, where dyad members are not aggregated and are instead treated as independent, there was a significant effect of condition. That is, individuals recalled more items they had encountered during collaborative rounds \((M=5.24, S D=2.63)\) than during independent rounds \((M=4.31, S D=2.47), t(61)=2.03, p<.05\). Thus, there appears to be a tendency for enhanced recall from collaboration. Admittedly, these effects are smaller than the performance measures, though more power may bear this out.


Figure 4: Binned individual-level recall per round. Circle sizes illustrate the number of individuals that recalled the corresponding number of items from each round. The line shows the fit of a generalized linear model with quadratic term.

\section*{Conversation Analysis}

In order to further quantify the effects of collaboration on performance, conversations during the collaborative sessions were recorded. A coarse analysis of these recordings allowed us to collect information on the total amount of time each dyad spent in the collaborative sessions, as well as the amount of this time that was spent talking. On average, dyads spent 241.13 seconds ( \(S D=\) 71.37) in the (summed) collaborative rounds, and used, on average, 109.29 of these seconds \((S D=34.72)\) conversing. Because the amount of time spent in the collaborative part of the task varied between dyads, a measure of percent talking was also calculated for each dyad. This percent talking measure varied from about \(27 \%\) to \(70 \%\) ( \(M=46.54\), \(S D=10.76\) ).

As in the previous analyses, results were aggregated over dyads and each data point represents the group-level mean, across a dyad's participants. The total amount of time each dyad spent talking was negatively correlated with their performance, as measured by the number of correct answers they submitted during the collaborative rounds, \(r(28)=-.77\), \(p<.0001\). That is, the more talking they did, the worse they performed. This negative correlation may reflect the fact


Figure 5: Recall for items from Rounds 1-4 for each dyad, ranked in order of performance. Dotted lines with empty circles show the aggregated number of items recalled by dyads working collaboratively; Solid lines with filled circles show recall by dyads working non-collaboratively.
that when uncertain of an answer, dyads spend more time in discussion in order to figure it out. Indeed, when considering the percentage of time spent talking, there was a positive correlation with performance, \(r(28)=.27\), although this trend did not achieve significance. Figure 6 shows the relationship between talking and performance, as measured by both absolute and percentage metrics of talking.


Figure 6: Relationship between talking and performance. The scatterplot on the left shows each dyad's performance (percentage of questions answered correctly) as a function of the total number of seconds spent talking (regression line \(m=-0.4229\) ). On the right, performance is shown as a function of the percentage of time spent talking (regression line \(m=0.4724\) ).

\section*{General Discussion}

On all measures of performance for the trivia task, there appeared to be a collaborative benefit. Aggregate dyads achieved more correct answers in the collaborative rounds than in the independent rounds, and they did so with greater accuracy. Interestingly, aggregate dyads were actually faster in the collaborative rounds than in the independent rounds, despite the fact that they had the added task of communicating with their partner for each question. With respect to the terminology described earlier (Hare, 1976), we observed that the group product, produced by collaborating individuals, was better than the individual product, produced by individuals working alone. The recall task also suggested a benefit from collaboration. Previous work has shown that participating collaboratively in recall enhances future independent recall (Basden, Basden \& Henry, 2000), but our results also suggest that collaborative encoding could enhance independent recall: the individual recall product of collaborating individuals was (slightly) greater than the individual recall products of individuals acting alone.

It must be noted, however, that the present study was specifically designed to enable us to look for evidence of a collaborative gain. The collaborative benefit apparent in this situation may not apply to other situations, as previous work described earlier has found that the degree of collaborative gain is highly influenced by social context. Future work is needed to elaborate on the specifics of the social, environmental and task contexts which allow for these collaborative gains. We would also like to address the current findings in the context of interpersonal alignment, in future work. It was noted earlier that the use of similar taskrelevant linguistic forms benefits dyadic cooperation, (Fusaroli et al., 2012), and a growing body of research addresses how interpersonal interactions can cause automatic alignment to spread between physical, linguistic, and other cognitive states (Tollefsen \& Dale, 2012). This begs the question of whether collaborative performance on knowledge-based and memory tasks can be influenced or indicated by various levels of behavioral, linguistic, and cognitive alignment.

\section*{Acknowledgments}

This research was supported partially by NSF grants BCS0826825 and BCS-0926670. We would like to thank Jacqueline Pagobo and Maxine Varela for their assistance with data collection.

\section*{References}

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., \& Frith, C. D. (2010). Optimally interacting minds. Science, 329(5995), 1081-1085.
Barnier, A. J., Sutton, J., Harris, C. B., Wilson, R. A. (2008). A conceptual and empirical framework for the social distribution of cognition: The case of memory. Cognitive Systems Research, 9(1-2), 33-51.

Basden, B. H., Basden, D. R., Bryner, S., \& Thomas III, R. L. (1997). A comparison of group and individual remembering: Does collaboration disrupt retrieval strategies? Journal of Experimental Psychology: Learning, Memory, and Cognition, 23(5), 1176-1189.
Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. International Journal of Forecasting, 5, 559-583.
Esser, J. K. (1998). Alive and well after 25 years: A review of groupthink research. Organizational Behavior and Human Decision Processes, 73(2-3), 116-141.
Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., \& Tylén, K. (2012). Coming to terms: Quantifying the benefits of linguistic coordination. Psychological Science, 23(8), 931-939.
Galton, F. (1907). Vox populi. Nature, 75(1949), 450-451.
Hare, A. P. (1976). Handbook of small group research. New York: Free Press.
Hastie, R., \& Kameda, T. (2005). The robust beauty of majority rules in group decisions. Psychological Review, 112(2), 494-508.
Hill, G. W. (1982). Group versus individual performance: Are \(N+1\) heads better than one? Psychological Bulletin, 91(3), 517-539.
Kerr, N. L., \& Tindale, R. S. (2004). Group performance and decision making. Annual Review of Psychology, 55, 623-655.
Loftus, E. F. (1996). Eyewitness testimony. Cambridge: Harvard University Press.
Rajaram, S., \& Pereira-Pasarin, L. P. (2010). Collaborative memory: Cognitive research and theory. Perspectives on Psychological Science, 5(6), 649-663.
Roediger, H. L., Meade, M. L., \& Bergman, E. T. (2001). Social contagion of memory. Pschonomic Bulletin \& Review, 8(2), 365-371.
Surowiecki, J. (2004). The wisdom of crowds. New York: Random House.
Tollefsen, D. P. (2006). From extended mind to collective mind. Cognitive Systems Research, 7(2-3), 140-150.
Tollefsen, D. P., \& Dale, R. (2012). Naturalizing joint action: A process-based approach. Philosophical Psychology, 25(3), 385-407.
Vul, E., \& Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. Psychological Science, 19(7), 645-647.
Weldon, M. S., Blair, C., Huebsch, P. D. (2000). Group remembering: Does social loafing underlie collaborative inhibition? Journal of Experimental Psychology: Learning, Memory, and Cognition, 26(6), 1568-1577.
Wegner, D. M. (1987). Transactive memory: A contemporary analysis of group mind. In B. Mullen \& G. R. Goethals (Eds.), Theories of group behavior, (pp. 185208). New York: Springer-Verlag.

Wegner, D. M., Giuliano, T., \& Hertel, P. T. (1985). Cognitive interdependence in close relationships. In W. Ickes (Ed.), Compatible and incompatible relationships, (pp. 253-276). New York: Springer-Verlag.

\title{
A Long-Term Memory Competitive Process Model of a Common Procedural Error
}

\author{
Franklin P. Tamborello, II (franklin.tamborello.ctr@nrl.navy.mil) \\ National Research Council Postdoctoral Research Associate \\ J. Gregory Trafton (trafton@itd.nrl.navy.mil) \\ United States Naval Research Laboratory \\ 4555 Overlook Ave SW \\ Washington, DC 20032 USA
}

\begin{abstract}
A novel computational cognitive model explains human procedural error in terms of declarative memory processes. This is an early version of a process model intended to predict and explain multiple classes of procedural error a priori. We begin with postcompletion error (PCE), a type of systematic procedural error that people are prone to commit when there is one step to perform after they have accomplished their main task goal. Participants in an experiment demonstrated increased PCE rates following an interruption in a realistic form-filling task. The model explains PCE as a consequence of two declarative retrieval processes, spreading activation and base-level activation, competing with each other because of features of task and working memory structure. Our intention is to generalize the model to other classes of procedural error in complex task environments.
\end{abstract}

Keywords: computational cognitive model; human error; human-computer interaction; interruption; long-term memory; working memory

\section*{Introduction}

If you have ever left an original document on a photocopier after walking away with the copies then you have committed a postcompletion error (PCE). PCE is one example of a systematic procedural error, an error people tend to commit in familiar tasks that follow a specific sequence of actions each time the task is performed. Systematic procedural errors seem to be products of a combination of stable human cognitive structures and processes as well as certain task environments. PCE, in particular, tends to have a much higher rate of incidence than chance slips and seems to be very resistant to training (Byrne \& Davis, 2006). Our goal is to understand the cognitive structures and processes underlying PCE and, ultimately, to extend that same model to account for other systematic procedural error types.

Studying human error is important because with increasing capability and complexity of our technological systems (e.g., transportation, power generation) the amount of damage that can result from error is magnified. While chance slips occur because humans are fundamentally stochastic, systematic error occurs when certain features of human cognition meet certain task environmental conditions. If we learn about those cognitive and environmental features then we can learn to avoid them in our technological systems such as by exclusion from designs (Chung \& Byrne, 2008) or prediction and prevention (Ratwani \& Trafton, 2011).

Studying human error is difficult because of the variability of error behavior. Furthermore, error often arises from the dynamic interactions of several cognitive processes
that normally perform with with very little error. Models of human error are often complex compared to models of other behavior because these models must capture these interactions in ways that lead to proper proportions of both correct and incorrect behaviors.
For PCE, Byrne and Bovair (1997) explained it as a function of limited-capacity working memory. They addressed high and low working memory demand as well as individuals' high and low working memory capacities. Their model assumed a hierarchical goal representational structure. This was based on a GOMS (Card, Moran, \& Newell, 1983) analysis of an experiment task also reported in their study. Their CAPS model (Just and Carpenter, 1992) propagated activation necessary for retrieval of step representations downward from the task supergoal to subgoals to individual steps. Subgoals had to have their activations maintained above a certain threshold in order for them to remain accessible. Crucially, the main goal of the procedure would be satisfied before it was time to perform the postcompletion step. The presence of other information to maintain in an active state, in this case a three-back memory task, taxed the system to capacity such that it failed to maintain the postcompletion subgoal above threshold.

Another account of systematic error, Memory for Goals (Altmann \& Trafton, 2002), posits that we encode episodic traces of our goals as we complete tasks. Each goal is encapsulated in an episodic memory, which sparsely represents a behavioral context at the time of its encoding. The strength of these memories decay over time such that it may be difficult to remember the correct point at which we resume a task after an interruption. Memory for Goals provides a process-level theory for why certain types of errors are made during a well-learned task as a consequence of retrospective, episodic memory (Altmann \& Trafton, 2007; Ratwani \& Trafton, 2010, 2011; Trafton, Altmann, \& Ratwani, 2009). Memory for Goals implies that people are able to retrieve suspended goals successfully if and only if there are cues that prime them (Altmann \& Trafton, 2002). Here decay is indexed by time, so postcompletion steps, being at the end of their tasks, have relatively more time to decay compared to other steps that come earlier in the task.

The model presented in this paper draws upon both previous works, predicting PCE to occur as a combination of goal decay and a limited-capacity to spread activation from working memory to long term memory. Ultimately what we want is a unified framework with which we can make predictions about PCE, and later, other types of human error. A unified framework is important because one cognitive system, i.e. the human mind, produces all error types. Getting the explanation correct for one type then acts as a constraint on getting the explanation correct for the next


Figure 1: The financial management task interface resembled a web form. Subgoals assumed for the model are grouped by dotted lines and labeled for purposes of illustration here, but were not so in the task.
type tackled by the theory. Furthermore, if we are to predict error in complex task environments multiple error types must fall naturally out of the theory.

\section*{Experiment}

Participants performed a version of Ratwani and Trafton's (2011) financial management task (Figure 1). This is a type of form-filling task wherein participants, using a graphical user interface, click a series of buttons in a specific order. The goal of the task is to fill out an order form according to information available within the display. An arithmetic task occasionally interrupted the financial management task for 15 seconds at a time.

The final step of the task consisted of a single button not placed within a box and placed above the right column of boxes. This arrangement broke with the Western reading convention followed by the progression of all of the other steps. This step was arranged this way because we intended it to serve as a postcompletion step.

\section*{Design and Procedure}

Each order on the financial management task constituted a single trial. Control and interruption trials were manipulated in a within participants design; participants performed 12 trials. Half of the trials were control trials with no interruption and half were interruption trials with two interruptions each. The order of trials was randomly generated and participants did not have prior knowledge as to which trials would be control or interruption trials.

There were eight possible interruption points in the financial management task. These points occurred after clicking the Confirm button following the first seven modules, including just prior to the postcompletion action. The location of the interruptions on a trial by trial basis was randomized with the constraint that exactly two interruptions occurred just prior to the postcompletion step and at least one interruption occurred at each of the other
seven possible locations. The were 12 postcompletion error opportunities, one during each trial. Six of these opportunities were during control trials with no interruptions, two opportunities were immediately following an interruption, and four opportunities were during interruption trials where an interruption occurred at a point that did not immediately precede the postcompletion step.

Participants were seated approximately 47 cm from the computer monitor. After the experimenter explained the financial management task and interrupting task to the participant, the participant completed two training trials (one trial with and one trial without interruptions) with the experimenter. Following these two training trials, participants had to perform two consecutive randomly selected trials on their own without making a postcompletion error before the participant could begin the experiment. Forcing participants to perform two consecutive error free trials was a method for ensuring that participants were proficient at the task before beginning the actual experiment. Each participant was instructed to work at his/ her own pace. When performing the interrupting task, participants were instructed to answer the addition problems as soon as the solution was known and to answer as many addition problems as possible in the time interval. Upon resumption of the financial management task, there was no information available on the interface to indicate where to resume.

For modeling purposes the important points about the financial management task were:
1. It featured a primary task that was occasionally interrupted by a secondary task,
2. Participants had to follow a specific procedure.
3. The spatial layout of the interface (working from top to bottom down the left column and then the right column of Figure 1) and the operations required to perform the task were quite intuitive.
4. After entering information in each module, the participant clicked the Complete Order button (upper right corner). Clicking the Complete Order button was the postcompletion step and failing to click the Complete Order button constituted a PCE.
5. The spatial layout of the task grouped steps by proximity. This encouraged use of an intuitive heuristic ("go down the column"), as well as having an isolated "clean-up" step at the end. This format followed the form of other tasks shown by GOMS analysis to lead to subgoaling (e.g., Byrne \& Bovair, 1997).

6 . No information remained on the interface after clicking the confirm button within each module (i.e. no global place keeping (Gray, 2002)).
7. Measures: A PCE was defined as failing to click the last step's button and instead making an action that was in service of the next order on the financial management task (e.g. attempting to start a new trial by clicking an Order Ticker). The PCE rate was the number of PCEs divided by the number of opportunities to make a PCE. Skipping the next correct step, at any other time, was classified as an anticipation error.


Figure 2: Base-level and spreading activation of the model's postcompletion step chunk in control (a) and interruption (b) trials. X-axes indicate the step to be performed, by step ordinal number. Interruptions occur for the interruption trial type between steps four and five and nine and ten in this example. Consequently the goal buffer chunk lacks task context representation in these two spots, indicated by dashes. The chunk encoding the main goal of the task is Do a Trial (DAT) and it is associated to subsequently performing the first step. Do a Trial, as well as each step's representation, acts as context to cue retrieval of the next step. Do a Trial's activation is depicted in panel c. Arrows indicate times at which Do a Trial receives activation boosts because the model retrieves it at the end of each subgoal. The model's behavior with regard to Do a Trial is the same in both control and interruption conditions. The model's internal context representations, encoded in the chunks referenced from the slots of the goal and imaginal buffer chunks, are depicted at each step of the task beneath panels a and \(b\). These are the procedure step representations the model had retrieved when performing previous steps. To conserve space task steps are indicated by procedure sequence order, so the first step of the task is 1 and so on.

\section*{Model}

A model constructed using the ACT-R 6 cognitive architecture (Anderson et al., 2004) performed an abstract version of the financial management task. ACT-R is a hybrid symbolic and subsymbolic computational cognitive architecture that takes as inputs knowledge (both procedural and declarative about how to do the task of interest) and a simulated environment in which to run. It posits several modules, each of which perform some aspect of cognition (e.g., long-term declarative memory, vision). Each module has a buffer into which it can place a symbolic representation that is made available to the other modules. ACT-R contains a variety of computational mechanisms and the ultimate output of the model is a time stamped series of behaviors including individual attention shifts, speech output, button presses, and the like. One of the benefits of embodying a theory in a computational architecture, such as

ACT-R, is that it allows researchers to develop and test concrete, quantitative hypotheses and it forces the theorist to make virtually all assumptions explicit. To the extent that the model is able to simulate human-like performance the model provides a sufficiency proof of the theory.

In essence, the model worked by cyclic, activation-based retrieval from long-term memory of the task step representations encoded as chunks. At each step there were two sources of retrieval activation: 1) spreading activation from the contents of the goal and imaginal representations (these constituted the model's working memory), and 2) each chunk's base-level activation. Sometimes these activation sources conflicted with each other, particularly for the postcompletion step. At such times the model was likely to commit an error.
Activation spreading from the model's working memory to the long-term memory encoding the postcompletion step
increases with advancing task context because of the inverse association strength function we used (Equation 1).
\[
\begin{equation*}
\left(\frac{1}{i-j}\right) \mathrm{m} \tag{1}
\end{equation*}
\]

That in turn is based on step co-occurrence. For the model, doing one step cues the next (Figure 2a). Do a Trial, the main goal of the task, gets retrieved at the end of every subgoal. With each retrieval its base-level activation gets a sharp increase that decays gradually over time (Figure 2c).

The difference at the postcompletion step between control and interruption is that in the interruption condition, the model lacks spreading activation from its working memory (Figure 2b). This is because when the interruption occurs, the model clears that resource of primary task representations so that secondary task representations may reside there. Then when the model resumes task execution, it restores only a part of its task context representation.

The model lacks context, and thus spreading activation, at resumption because the sparse representation of the episodic memory trace only records reference to one chunk encoding context (Altmann \& Trafton, 2002), in this case the imaginal buffer chunk. Consequently when the model resumes it has only the imaginal buffer contents and not the goal buffer contents available to it.

In the example depicted in Figure 2b, the model is shown as having been interrupted before steps five and ten. Because of the way the model encodes its episodic memory and uses that to resume task execution, the chunks encoding steps four and nine are not referenced from the goal buffer chunk when it is time for the model to retrieve from longterm memory the chunks encoding how to perform steps five and ten.

The model produces PCE at resumption because total activation for the postcompletion step chunk and Do a Trial are approximately equal. In that context and with transient retrieval activation noise, each has an approximately equal chance of being retrieved.

\section*{Spreading Activation and Strength of Association}

An architectural feature of ACT-R is that it uses a limited pool of spreading activation from sources-a chunk in a module's buffer-to associated chunks in declarative memory as one of its mechanisms of declarative retrieval. Our model used ACT-R's goal and imaginal buffers as sources of activation, each providing one unit of spreading activation.

Activation spreads from source chunks in ACT-R's buffers to chunks residing in ACT-R's declarative memory as a function of the strength of association between the value of each slot in source chunk \(j\) to chunk \(i\) in declarative memory (Anderson, 2007; Anderson et al., 2004). This gives ACT-R a way to adjust its behavior according to context as the strength of association indicates the probability that chunk \(i\) will be needed in context \(j\). The limited pool of activation is divided equally among all the slots of source chunk \(j\). This means that ACT-R implements a limited-capacity working memory.

Our model set strengths of association from each step's representation to the next at the beginning of each model
run according to Equation 1.Association strengths remained static for the duration of each model run. Here, \(j\) is the serial position within the financial management task of the step encoded by a chunk representing some part of the model's current context (i.e., the last step performed). \(I\) is the serial position within the financial management task of the step encoded by an associated chunk in declarative memory. \(M\) is for a global ACT-R parameter to set the maximum association strength, set to 3.5 for this model.

For example, if the model had just performed the first step, Order Ticker, the association strength to the chunk encoding the second step, Quantity, would be 3.5. The strength of association to the third step, Cost, would be 1.75 . This enabled associative chaining from the model's current context to the next procedure step. This produced a graded representation that decreased in strength with increasing psychological distance, a feature borrowed from Altmann and Trafton (2007).

\section*{Base-Level Activation}

Base-level activation is an estimate that a declarative chunk will be needed in the future, given how recently it has been needed and how often it has been needed. This is another architectural feature of ACT-R and the idea is that given a limited capacity to retain information, those chunks not retrieved for a long time are allowed to have their activation decay below a threshold beyond which their retrieval will become less likely. Conversely, chunks that are retrieved frequently will have a high base-level activation contribution to their total activation. The model used ACTR's default decay rate of 0.5 and activation noise of 0.2 .

We assume spatial grouping of steps leads to Millerian (Miller, 1956) chunking of steps into groups, or subgoals. Anderson et al. (Anderson, Bothell, Lebiere, \& Matessa, 1998), in their model of sequence memory, determined it crucial that sequence items be recalled in groups. Their model traversed a hierarchy of list item chunks, grouping chunks, and a chunk encoding the current list.

The financial management task model abstracted this process by adding a retrieval reference to the Do a Trial chunk upon completion of each financial management task subgoal: Order Ticker, Quantity through Margin, Stock Exchanges through Review, and Complete Order (see Figure 1). Each retrieval reference boosted Do a Trial's base-level activation. The idea is that after completing one subgoal, the task main goal is retrieved and used to retrieve the next subgoal. Therefore Do a Trial's base-level activation tended to be relatively high.

The postcompletion step happened to be needed immediately after a retrieval reference to Do a Trial (after completion of the preceding subgoal). Furthermore, a long time might have elapsed since the postcompletion step's last retrieval, especially when there had been two 15 s interruptions during the trial. The relatively long time elapsed between retrievals of the postcompletion step lead to much decay of its base-level activation. Meanwhile, Do a Trial had received four retrieval references, one at the end of each of the subgoals. Each retrieval reference contributes to a chunk's base-level activation.

This combination of the postcompletion step's decay and Do a Trial's repeated retrieval was crucial for the model's commission of PCE at resumption. Because of these baselevel activation mechanics the postcompletion step would then need a large quantity of spreading activation to have enough total activation to be retrieved reliably at postcompletion step time. Otherwise since Do a Trial had the second-highest retrieval activation because of its high base-level activation, it might be retrieved instead of the post-completion step's representation.

\section*{An Example Model Run}

The model started its run by retrieving a procedure step representation. Because its context at the time would indicate that it was starting the task and the first step is most associated with starting, the first step would usually be the procedure step representation retrieved. After that the model simply looped through its basic behavioral cycle until it either finished a trial of the financial management task or until it was interrupted.

During the interruption, the model cleared its representations of its financial management task context from its working memory constructs-the goal and imaginal buffers-and replaced them with ones representing the interrupting task. At the end of 15 s the financial management task interface replaced the interrupting task's, whereupon the model detected that its visual environment had changed back to the financial management task and so then it initiated its resumption subroutine.

When the model resumed the financial management task it began so by retrieving an episodic chunk. Because which episodic chunk retrieved was a function of base-level activation and transient noise, the most recent episodic chunk was usually the one retrieved.

The episodic chunk held a reference to an imaginal buffer chunk, which the model copied to the imaginal buffer. That imaginal buffer chunk held a record of the subgoal's steps completed at the time the episodic chunk was created. The restored imaginal buffer chunk provided the link necessary to retrieve the next step's representation at resumption.

The imaginal buffer chunk could contain references to as many as four step representations, all previous to the next correct step and all having varying strengths of association to it. This means that the limited activation source from the imaginal buffer could be divided by up to four.

Furthermore, the farther away in the procedure those steps were from the postcompletion step, the weaker their strength of association, and so the less source activation would propagate to the retrieval of the postcompletion step. The eighth step associated less strongly to the postcompletion step than did the ninth, but the eighth step took as much of the imaginal buffer's activation as the ninth.

Expressed in terms of maximum association strength, when it was time to perform the postcompletion step the imaginal buffer chunk spread only \(25 / 48\) ths of available activation to the postcompletion step ( \(\approx 1.8\) with :mas \(=3.5\) ). Roughly half of the activation source available from the imaginal buffer was diverted away from retrieval of the postcompletion step because of the presence of the previous steps' representations.

The model predicted more PCE for interrupted steps than non-interrupted steps because although the goal buffer chunk also held a reference to the just-completed step, the episodic chunk only encoded the imaginal buffer chunk. And because only one other goal slot was occupied, the association from the ninth step to the postcompletion step would get half of goal's available spreading activation, 1.75 units. Thus with the goal buffer chunk present the postcompletion step would get twice as much spreading activation as when the goal buffer chunk was absent due to interruption. This was enough to make the difference between reliable postcompletion step execution and equal chance of PCE when combined with base-level activation.

Furthermore, because Do a Trial got retrieval references four times during each trial-including once immediately before the postcompletion step-it tended to have a much higher base-level activation than did the postcompletion step. So when the model's only source of context representation was the imaginal buffer chunk and the task context was time to perform step the postcompletion step, the postcompletion step and Do a Trial would have similar amounts of total activation. Transient noise added at retrieval time (a standard feature of ACT-R) could tip the balance one way or the other.

\section*{Model Fit}

We used our model to simulate data from 1,000 subjects. This large number of model runs allowed effects to converge on the model's true predictions. The model's means closely matched those of the participants, \(r=.976\), RMSD \(=.0334\). Figure 3 plots the model's means against the participants' means and \(95 \%\) confidence intervals.

\section*{Discussion}

PCE's distinction from anticipation is illustrated by comparison of their rates. If PCE were simply a matter of an anticipation error happening to fall at the last step then PCE and anticipation rates should be identical. However, Figure 3 shows clearly that the two error types are different.

What makes PCE unique is that it is a product of: 1) goal base-level activation decay below that of a competing goal's, 2) working memory structures with limited capacity to spread activation to long-term memory retrieval, 3) the size and structure of working memory representations-a preceding, large subgoal meant there were more items in working memory that would steal some of the available spreading activation away from the postcompletion step's retrieval, and 4) some context representation was not immediately available upon resumption.

\section*{Issues}

Rather than learning the task, the model relied on assumptions about task representation structure. However, those are based on previous efforts with regard to sequence learning and memory (Anderson, Bothell, Lebiere, \& Matessa, 1998a) and are also congruent with wellestablished methods of task analysis, particularly GOMS (Card, Moran, \& Newell, 1983). We adapted some procedural and structural aspects of the Anderson et al.


Figure 3: Mean error rates, human (bars) and model (circles). Error bars display the \(95 \%\) confidence interval of the mean. Panel a depicts interruption trials, controls in b.
model because procedures are a kind of sequence. Mapping spatial groupings of task interface widgets to procedural representation groups of steps led to two important portions of the model's explanation of PCE in the interruption paradigm: groups of procedure steps held in working memory and high availability of the supergoal, Do a Trial.

\section*{Implications}

The model explains PCE partially as a result of working memory constraints, following in the footsteps of Byrne and Bovair (1997). This implies that it should also explain PCE as a function of working memory capacity as their model did. In fact, this model has done just that with very little change (Tamborello and Trafton, submitted). Its anticipation error performance, while imperfect, suggests that the model should be extendable to other types of systematic procedural error, such as perseverations.

The decay process in the model has a cost, which is that suspended goals are forgotten gradually, making them harder to resume. The model carries with it an implied assumption that goals are retrieved at the outset of task execution, and then may decay from working memory before they are actually executed. With respect to PCE, the model implies that the default tendency is to make such errors, not avoid them.

Overall the model is encouraging with regard to our ultimate goal of developing a unifying framework of human error. But it is also encouraging from the standpoint of developing models of human procedural memory and execution, since the same cognitive systems are involved. Eventually it may also prove useful for models of error detection and recovery.

\section*{Acknowledgments}

We are grateful to Laura M. Hiatt, Sangeet Khemlani, and Magdalena Bugajska for their helpful comments. This research was performed while the first author held a National Research Council Research Associateship Award at the US Naval Research Laboratory.

\section*{References}

Altmann, E.M., \& Trafton, J.G. (2002). Memory for goals: Activation-based model. Cognitive Science, 26(1), 39-83.
Altmann, E. M., \& Trafton, J. G. (2007). Timecourse of recovery from task interruption: Data and a model. Psychonomic Bulletin \& Review, 14(6), 1079-1084.
Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., \& Qin, Y. (2004). An integrated theory of the mind. Psychological Review, 111(4), 1036-60.
Anderson, J. R., Bothell, D., Lebiere, C., \& Matessa, M. (1998). An integrated theory of list memory. Journal of Memory and Language, 38, 341-380.
Byrne, M. D., \& Bovair, S. (1997). A working memory model of a common procedural error. Cognitive Science: A Multidisciplinary Journal, 21(1), 31-61.
Byrne, M. D., \& Davis, E. M. (2006). Task structure and postcompletion error in the execution of a routine procedure. Human Factors, 48(4), 627-38.
Card, S. K., Moran, T. P., \& Newell, A. (1983). The psychology of human-computer interaction. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
Chung, P. H., \& Byrne, M. D. (2008). Cue effectiveness in mitigating postcompletion errors in a routine procedural task. International Journal of Human-Computer Studies, 66(4), 217-232.
Gray, W. D. (2002). Simulated task environments: The role of high-fidelity simulations, synthetic environments, and laboratory tasks in basic and applied cognitive research. Cognitive Science Quarterly, 2(2), 205-227.
Just, M. A., \& Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. Psychological Review, 99, 122-149
Miller, G. A. (1956). The magical number seven, plus or minus two. Psychological Review, 63, 81-97.
Ratwani, R. M., \& Trafton, J. G. (2010). A generalized model for predicting postcompletion errors. Topics in Cognitive Science, 2(1), 154-167.
Ratwani, R. M., \& Trafton, J. G. (2011). A real-time eye tracking system for predicting and preventing postcompletion errors. Human-Computer Interaction, 26(3), 205-245.
Tamborello, F. P., II, \& Trafton, J. G. (submitted). A LongTerm Memory Competitive Process Model of Procedural Error, Part II: Working Memory Load and Capacity. The 12th International Conference on Cognitive Modeling.
Trafton, J. G., Altmann, E. M., \& Ratwani, R. M. (2011). A memory for goals model of sequence errors. Cognitive Systems Research, 12, 134-143.

\title{
Implicit transfer of mirrored spatial structure in visuomotor sequence learning
}

\author{
Kanji Tanaka (kanji@fennel.rcast.u-tokyo.ac.jp) \\ \({ }^{1}\) Research Center for Advanced Science and Technology, The University of Tokyo \\ \({ }^{2}\) Japan Society for the Promotion of Science \\ Katsumi Watanabe (kw@fennel.rcast.u-tokyo.ac.jp) \\ \({ }^{1}\) Research Center for Advanced Science and Technology, The University of Tokyo
}

\begin{abstract}
Implicit transfer in sequential learning can occur with some spatiotemporal structures but not with others. Here, we investigated whether the consistent mirror-reversal of visuomotor sequences would lead to implicit transfer. A "set" comprised three sequential button presses and seven consecutive sets comprised a "hyperset." Participants learned hypersets by trial and error with their right hand. Then, they learned another hyperset, in which each set was vertically mirrored, horizontally mirrored, or randomly generated. Even when the participants did not notice the mirrored rule, the mirrored hypersets led to implicit transfer in terms of accuracy for both vertical and horizontal reversals. Furthermore, the vertical reserval also led to implicit transfer of performance speed. Taken together, the present results suggest that people can implicitly apply their learned representations to the mirrored visuomotor sequences.
\end{abstract}

Keywords: Implicit learning; Sequential learning; Transfer; Mirror symmetry; Speed, Accuracy

\section*{Introduction}

Implicit learning of behavioral sequences play an important role in our daily life. Our cognitive abilities such as language usage, playing the piano, and driving a car can be improved by implicit acquisition or learning of skills (see reviews for implicit learning; Abrahamse, Jiménez, Verwey, \& Clegg, 2010; Perruchet \& Pacton, 2006). In fields of cognitive science, cognitive psychology, or experimental psychology, several implicit learning paradigms have been proposed (e.g., Serial Reaction Time (SRT) task, Nissen \& Bullemer, 1987; artificial grammar learning (AGL), Reber, 1967; visuomotor button press task, Hikosaka, Rand, Miyachi, \& Miyashita, 1995). Most studies have investigated whether people implicitly learn a sequence. In particular, some have insisted that people can learn both elements and a higher-order structure of a sequence. For example, Stadler and Neely (1997) showed that the structure of a sequence had a larger influence on learning in the SRT task than the length of that sequence, indicating that some structures tend to be easier to learn than others (see also Cohen, Ivry, \& Keele, 1990). Some studies adopting the AGL task suggested that people might implicitly learn fragments or chunks of two, three, or four letters (ServanSchreiber \& Anderson, 1990; Perruchet \& Pacteau, 1990). In the visuomotor button press task, Hikosaka et al. (1995)
observed that participants performed slowly and inaccurately when a higher-order sequence was reversed, but individual elements remained identical. Thus, the previous studies pointed to the possibility that people learned certain levels of higher-order structure of a sequence.

Transfer of motor learning refers that some movement controls are learned in one situation and transferred to another situation (e.g., Schmidt \& Young, 1987). Experiments with key-pressing tasks have demonstrated transfer between sequences that require different arm or finger movements, suggesting that abstract representations underlie sequence production (e.g., Bapi et al., 2006; Kovacs et al., 2009). Namely, this implies that some representations used for motor execution appear to be independent of the effectors producing the action. Cohen et al. (1990), for example, found that transfer of speed occurs when participants learned a tapping task with their three fingers and then, the same tapping task with their index finger (they were not aware that the learning and transfer tasks were identical sequences due to a distraction task).

Previous studies have reported that people implicitly detect reversed or mirrored structures of musical melodies, even when they are unaware of the structure (e.g., Dienes, Kuhn, Guo, \& Jones, 2012). For example, Dienes and Longuet-Higgings (2004) used sequences comprised of twelve musical tones, where the first six tones were randomly generated and the second six tones were altered from the first tones with some specific alternations. During the learning phase, participants were told that the musical melody obeyed some specific rules and in the test phase, they required to answer whether the musical melody followed the rules or not. Results showed that participants who had background experience with atonal music could implicitly detect altered melodies (e.g., reversals and mirrors). Similarly, Kuhn and Dienes (2005) observed that trained participants preferred mirrored melodic structures to non-mirrored structures. Collectively, these results indicate that people could implicitly use the mirror symmetries of learned sequences.

As well as the study of musical melody (e.g., Dienes \& Longuet-Higgings, 2004), in the present study, we were interested in whether implicit transfer of visuomotor sequence learning occurred when learned sequences (i.e., visual configuration and finger movement of the sequence)
became mirror symmetries in transfer. In order to investigate the effects of the mirrored structure on implicit transfer in sequential learning, we employed a sequential button press task (e.g., Hikosaka et al., 1995, 1996, 2002; Watanabe et al., 2006, 2010). Hikosaka et al. (1999) summarized that in the visuomotor learning task paradigm, the early trial-and-error stage was controlled and explicit processes and those in the late learning stage were automatic and implicit. The experimental device consisted of 16 lightemitting diode (LED) buttons mounted in a \(4 \times 4\) matrix while in most studies of the SRT tasks, the device was composed of three or four aligned buttons, which enables us to examine two types of mirror transfer: Vertically mirrored and Horizontally mirrored. In the present study, a fixed visuomotor sequence (which constituted the "hyperset") of seven triads of button presses (hereafter called "sets") was generated for each participant. After participants learned the hyperset by trial-and-error, they were required to perform another hyperset, in which the sets were generated by a specific alternation rule. Here, we prepared three alternation rules, with which a visual configuration of the set was vertically mirrored (hereafter called "vertically mirrored rule"), horizontally mirrored (hereafter called "horizontally mirrored rule"), or randomly generated (hereafter called "random rule").

\section*{Method}

\section*{Participants}

120 right-handed participants ( 68 males, 52 females; mean age \(=21.19\) years, standard deviation \(=2.31\) ) participated in the experiment. All participants had normal or corrected-tonormal visual acuity, normal motor functions, and were naïve to the purpose of this study. All procedures were conducted in accordance with the Declaration of Helsinki.

\section*{Procedure}

We adopted a basic experimental paradigm used in previous studies (e.g., Hikosaka et al., 1995; Watanabe et al., 2010; Figure1). The experimental device consisted of 16 LED buttons mounted in a \(4 \times 4\) matrix and another LED button (called the "home key") at the bottom. Participants used their right index fingers to press the buttons. When participants pressed the home key for 500 ms , three buttons ("set") turned on simultaneously. Participants were required to press the illuminated buttons in the correct order, which they needed to uncover through trial-and-error. If participants were successful, the LEDs turned off, one by one, and a different set was illuminated, for which the participants were again required to discover the correct order. When participants pressed the wrong button, all LEDs were briefly illuminated, and participants then had to restart from the home key. Seven sets were presented in a fixed order, which we called a "hyperset," to complete a trial. A trial was considered an error when participants
pressed the wrong button in all the sets and successful when participants completed a hyperset, and For example, if participants press the wrong button in Set 5, they would need to start over from the home key. The same hyperset was repeated until participants completed it successfully for 20 trials (called a "block"). Participants were asked to perform the task as quickly and accurately as possible.

We prepared four types of hypersets: "Original," "Vertically mirrored," "Horizontally mirrored," and "Random". The Original hyperset was randomly generated for each participant. In the Vertically mirrored hyperset, the spatial configurations of the sets were reversed by the vertical axis from the Original hyperset. In the Horizontally mirrored hyperset, the spatial configurations of the sets were reversed by the horizontal axis from the Original hyperset. In the Random hyperset, the new spatial configurations were randomly generated.

\section*{Original hyperset}


\section*{Vertically mirrored hyperset}


Horizontally mirrored hyperset


Figure 1. Experimental device and schematic flow of the present study. Participants were instructed to learn the correct order by trial-and-error. The LED buttons were square in shape ( \(10 \mathrm{~mm} \times\) 10 mm ) and 8 mm apart. Participants were required to discover a correct order by trial-and-error. A trial was considered successful when participants completed a hyperset, and a trial was considered an error when participants pressed the wrong button in all the sets. For example, if participants press the wrong button in Set 2, they would re-start from the home key. A block is finished when participants successfully completed a hyperset 20 times. The Original hyperset was randomly generated for each participant. For each set, the three buttons were defined in ascending order of [1][2][3]. In the Vertically mirrored and Horizontally mirrored
hypersets, all the sets were spatially mirrored with white dashed line, resulting in the Vertically mirrored and Horizontal mirrored sets, respectively from the Original hyperset. The Random hyperset was randomly generated again for each participant, which is different from the Original hyperset. Note that the number shown on the LED button was not displayed during operation.

All participants first performed a block with the Original hyperset and then a block with the Vertically mirrored, Horizontally mirrored, or Random hypersets, which were randomly assigned. The two blocks were separated by a 5 -min break. No information was given regarding the alternation rule and, in the second block, participants were instructed that a new hyperset was randomly generated. In order to specifically examine the implicit form of transfer, participants were interviewed after the experiment. In the interview, they were asked how they performed and whether they noticed anything peculiar in the second block. If participants spontaneously reported the mirrored rule, they were excluded from our main analyses. Next, the experimenter explained the mirrored rule to the participants and those who recognized the mirrored rule were also excluded from our main data analyses. Methods for distinguishing explicit knowledge and implicit knowledge are still under debate. Several studies have used subjective measures based on confidence ratings (e.g., Ziori \& Dienes, 2006, 2008). Conversely, some studies defined implicit learning that participants were unable to verbalize what they acquired (e.g., Ashby, Alfonso-Reese, Turken, \& Waldron, 1998). In this study, we only focused on whether participants noticed the mirrored rule and, therefore, we defined explicit knowledge that participants were able to recognize the mirrored rule.

\section*{Data Analysis}

As a measure of accuracy, we counted the number of error trials before completing one trial. In order to evaluate speed, we measured the time that had elapsed from the moment the home key was pressed to the moment the third button of the final (7th) set was pressed for each successful trial. Similar parameters have been employed in previous studies and verified (e.g., Watanabe et al., 2006, 2010). We divided the 20 correct trials into five trial sections and calculated mean performance times within each trial section. We defined mean performance times during the fifth section (i.e., 17th to 20th trials) of the first block as a baseline for each individual participant so that we compared the magnitude of transfer among participants who differed on initial performance. We then calculated adjusted performance by subtracting the baseline from performance times during the second block (the Vertically mirrored, Horizontally mirrored, or Random hypersets) and divided by the baseline. This adjusted performance time [ \(\left.\mathrm{P}^{\text {second block }} / \mathrm{P}^{\text {baseline }}\right]\) represents the transfer magnitude of participants' performance times. A value more than 1 indicates that the performance time in the second block is faster than that of
the final trial section in the first block (i.e., baseline). Any difference in adjusted performance times indicates a difference in magnitude of transfer among different hypersets in the second block.

\section*{Results}

Forty participants were assigned for each of the Vertically mirrored, Horizontally mirrored, and Random hypersets. Since nineteen participants in the Vertically mirrored hyperset and seven in the Horizontally mirrored hyperset noticed the mirrored rules, they were excluded from our main analysis. Next, we excluded six participants whose performances in the first block were slower than two standard deviations from each group's average (three participants in Vertically mirrored, two in Horizontally mirrored, and one in Random hypersets). Similarly, we additionally excluded four participants whose performances in the second block were slower than two standard deviations from each group's average (one in Vertically mirrored, one in Horizontally mirrored, and two in Random hypersets). The selection procedure resulted in 17, 30, and 37 unaware participants with acceptable performance, for the Vertically mirrored, Horizontally mirrored, and Random groups, respectively.

We mainly conducted two-way mixed ANOVAs with the five trial sections as a within-subjects factor and the three hypersets as a between-subjects factor, which was called simply "ANOVA" hereafter, and post-hoc tests with the Shaffer's method when performed (called "post-hoc test"). For all hyperset groups, a significant decrease was found in both accuracy and speed measures in the first block, indicating that non-specific learning had occurred (ANOVA; \(F(4,324)>81.45, p<0.0001\); for both measures) and there were no differences among the hyperset groups (ANOVA; \(F(2,81)<0.24, p>0.78\); for both measures; Figure 2a and 2b). No significant interaction between experimental groups and successful trial sections (ANOVA; \(F(8,324)<1.24, p>0.27\); for both measures). These results were accord with those in previous works; the accuracy measure decreased rapidly in the first few completed trials, while the speed measure decreased more gradually (e.g., Watanabe et al., 2006, 2010).

In the second block (transfer block), mean adjusted performance times (i.e., speed) were generally faster (i.e., more transfer was found) in the Vertically mirrored group compared with the Random group (Figure 2c). The ANOVA revealed significant main effects of experimental group \((F(2,81)=3.34, p<0.05\); post-hoc test, Vertically mirrored \(<\) Random, \(p<0.05\) ) and successful trial section \((F(4,324)=75.03, p<0.001\); post-hoc test, 1 st \(>2\) nd \(>3 \mathrm{rd}\) \(=4\) th \(=5\) th section, \(p<0.01\) ). The interaction between experimental group and successful trial section was not significant \((F(8,324)=1.38, p=0.20)\). As for accuracy, the ANOVA revealed significant main effects of experimental group \((F(2,81)=12.38, p<0.0001\); post-hoc test, Vertically mirrored \(=\) Horizontally mirrored \(<\) Random, \(p<\)
0.001; Figure 2d) and successful trial section \((F(4,324)=\) 361.4, \(p<0.0001\); post-hoc test, 1 st \(>\) the other sections, \(p<\) 0.01 ). The interaction between experimental group and successful trial section was also significant \((F(8,324)=\) 14.61, \(p<0.0001\) ) and this interaction shows that in the first trial section, the accuracy was higher in the Vertically mirrored and Horizontally mirrored hypersets than in the Random hyperset \((F(2,81)=15.06, p<0.0001\); post-hoc tests, Vertically mirrored \(=\) Horizontally mirrored \(>\) Random, \(p<0.001\) ) while in the other trial sections, the accuracy was not different among the experimental groups ( \(F(2,81)<2.09, p>0.13\); for the other sections).


Figure 2. Performance in the first and second blocks. Error bars show the standard errors of the mean. All participants performed the Original hyperset in the first block. (a) Average performance time for successful trials in the first block. (b) Average number of errors before the successful completion of each trial in the first block. (c) Average adjusted performance time for successful trials in the second block. The adjusted performance was computed as follows: [ \(\left.\mathrm{P}^{\text {second block }} / \mathrm{P}^{\text {baseline }}\right]\). (d) Average number of errors before the successful completion of each trial in the second block.

A Pearson's chi-squared test revealed a significant difference in the proportion of participants who noticed the alternation rule between the Vertically mirrored and Horizontally mirrored groups ( \(\chi^{2}=6.89, \mathrm{p}<0.01\); Vertically mirrored \(>\) Horizontally mirrored). Therefore, we additionally examined whether the performances of those groups (aware vs. unaware) in the Vertically mirrored hyperset were different. We excluded two participants in the aware group whose performances in the first block were slower than two standard deviations from the group average (i.e., resulting in 17 aware and 17 unaware participants). In
the first block, we confirmed that performance did not differ between the groups \((F(1,32)<2.53, p>0.12\); for speed and accuracy measures; Figure 3 a and 3 b ). In the second block (transfer block), we found that the accuracy was higher when participants were aware of the alternation rule (Figure 3d). For speed, a two-way mixed ANOVA did not reveal significant main effects of awareness \((F(1,32)=2.49, p=\) 0.12 ; Figure 3 c ) while we observed the significant main effects of successful trial section \((F(4,128)=79.63, p<\) 0.0001 ; post-hoc tests, \(\left.1^{\text {st }}>2^{\text {nd }}>3^{\text {rd }}=4^{\text {th }}=5^{\text {th }}, p<0.05\right)\). No significant interaction between group in terms of awareness and successful trial section was not observed \((F(4,128)=2.29, p=0.06)\). For accuracy, a two-way mixed ANOVA revealed significant main effects of awareness \((F(1,32)=15.74, p<0.001\); Unaware \(>\) Aware \()\) and trial section \(\left(F(4,128)=199.73, p<0.0001\right.\); post-hoc test, \(1^{\text {st }}>\) the other sections, \(\mathrm{p}<0.0001\) ), as well as a significant interaction \((F(4,128)=18.18, p<0.0001)\). This interaction shows that in the first trial section, the accuracy was higher in the aware group than in the unaware \((F(1,32)=19.87, p\) \(<0.001\) ) while in the other trial sections, the accuracy was not different among the sorted groups \((F(1,32)<1.19, p>\) 0.28 ; for the other sections). These results confirmed that once participants obtained the alternation rule (i.e., explicit knowledge), they could clearly perform the hyperset with fewer errors (e.g., Watanabe et al., 2006).


Figure 3. Performance of participants who noticed the vertically mirrored rule ("Aware") and who did not ("Unaware") in the second block. Error bars indicate standard errors of the mean. The adjusted performance was computed as follows: [ \(\mathrm{P}^{\text {second block }} /\) \(\left.P^{\text {baseline }}\right]\). (a) Average performance time for successful trials in the first block. (b) Average number of errors before the successful completion of each trial in the first block. (c) Average adjusted performance times in the second block. (d) Average number of
errors before the successful completion of each trial in the second block.

\section*{Discussion}

In the present study, we investigated whether a spatially mirrored sequence in visuomotor sequence support implicit transfer of performance in accuracy and speed. We found that (1) both vertically and horizontally mirrored sequence led to transfer of learning in terms of accuracy even when participants did not notice the mirrored rules; (2) vertically mirrored sequence, in addition, led to transfer in terms of performance speed; (3) the proportion of participants who noticed the vertically mirrored rules were significantly higher than the horizontally mirrored rules; and (4) accuracy in the transfer session was significantly higher for the aware group than the unaware group with the vertically mirrored rule.

Previous studies discussed that people who have an experience of playing the piano implicitly discriminated reversed or mirrored structures of musical melodies (e.g., Dienes et al., 2004, 2012), indicating that people can implicitly understand relationships between original and reversed or mirrored sequences of musical tones. In the literature of intermanual transfer (Grafton, Hazeltine, \& Ivry, 2002), participants conducted the SRT task with the counting tone task (i.e., this distractor task usually makes participants unaware of the hidden repetition of the sequence) with their left hands in the learning block and subsequently they performed the transfer task with their right hands with the original and mirrored ordered sequence (original sequence, both stimulus sequence and order of response locations remained, but the finger movements were different; mirrored sequence, finger movements in transfer block were identical to those in learning block, but the stimulus location was visually mirrored). Seven out of eight participants were not aware of the repetition of the sequence and the results showed that performance of the original and mirror sequence was significantly better than that of the random sequence. In the present study, the Vertically and Horizontally mirrored hypersets produced the better transfer in terms of accuracy than the Random hyperset. Moreover, the Vertically mirrored hyperset led to better transfer in terms of speed than the Random hyperset. The present study is the first empirical study to show that implicit transfer occurs even when the visual configuration and finger movements of a sequence were consistently vertically or horizontally mirrored. The present results indicate that people implicitly apply their learned representation to the mirrored hypersets.

Most procedural and sequential learning in our daily life possess two stages of processing: the controlled exploration of patterns and the process of automatization after a pattern has been discovered (Anderson, 1982). How sequential learning should be done in order to induce implicit transfer likely has two possibilities; less learning (i.e., remain controlled process) or much learning (i.e., reach
automatic process) leads to implicit transfer. These two possibilities probably depend on whether the automatic process of the learning interferes with transfer; once an automatic process is established, the process can be interference when performing a transfer task because different sequence from learning is required and then, the less learning phase might be better to induce implicit transfer. Conversely, as the automatic process does not require much allocation of attention, the process might give allocation of attention when performing transfer, resulting in implicit transfer. Taken together with previous work (Dienes \& Longuet-Higgings, 2004), the results that people could implicitly understand the mirror symmetry might be associated with their automaticity of performance. In the study of music melody, only participants who had played the piano could use the mirrored structure of the music melody, indicating that implicit transfer can occur when cognitive skills (i.e., playing the piano) became automatic process. In the present study, in the first block, the participants completed the hyperset 20 times without errors and in the later learning phase, their performance probably reached at the automatic level (Hikosaka et al., 1999). Thus, these automatic operations likely made an allocation of attention of participants for the transfer task available. For example, Shanks, Rowland and Ranger (2005) showed that performances in the SRT tasks are degraded under doubletask conditions, which indicated that implicit learning depends on availability and allocation of attention and is susceptible to the double-task conditions while the learning process was not automatic. Collectively, once a process of task performance reached automatic, the allocation of attention for the learning task is alleviated, resulting in that the allocation of attention for a transfer task became available, which probably made participants possible to implicitly use the mirrored relationship and transfer their obtained representation to the mirrored sequence.

Next, we discuss the differential results between the Vertically and Horizontally mirrored hypersets. We observed implicit transfer of speed only in the Vertically mirrored hyperset, indicating that the vertically mirrored rule might be easier than the horizontally mirrored rule. However, this differs from the Fitt's law (speed-accuracy trade-off; Fitts, 1954) because distances of finger movements were the same between Vertically and Horizontally mirrored hypersets. Then, the present result might pertain to the residual or subthreshold awareness of the mirrored rules. We observed significantly different proportions of participants who noticed the vertically and horizontally mirrored rules. This indicated that the vertically mirrored rule might be easier to notice than the horizontally mirrored rule; sub-threshold awareness of the vertically mirrored order might prime performance within the consistently vertically mirrored sets. The relationship between awareness and difficulty of the task requires to be investigated, but a task of which most people can notice an alternation rule might be easily transferred.

In Watanabe et al. (2006), after a first hyperset was learned, new hypersets were generated by rotating the sets (i.e., entire stimulus configuration) by \(0^{\circ}, 90^{\circ}, 180^{\circ}\), and \(270^{\circ}\) (clockwise). Participants were not instructed that the new hypersets were based on the first learned hyperset. Through the experiment, half of the participants spontaneously noticed the regularity of the rotation while the other half did not. Watanabe et al. (2006), then, compared the performances of the participants who were aware and unaware of the regularity and found that those who noticed the regularity could perform the new hyperset more accurately than those who did not notice it while the different performance of speed was not observed. We found that performances of participants who noticed the vertically mirrored rule were more accurate than those who did not notice it. This result was basically accord with Watanabe et al. (2006). Once participants obtained explicit knowledge (i.e., were aware of the hidden rule), they performed the test sequence with fewer errors.

In addition, we compared the performance of the participants who were not aware of the regularity of the Vertically mirrored hyperset and those in the Random hyperset, and found implicit transfer of speed. This point differs from the previous study where no effect of explicit knowledge was found for performance speed with the rotated hypersets. Therefore, the vertical reversal might be a special case in terms of spatial transformation of visuomotor sequences.

In conclusion, in the present study, we investigated whether people could implicitly transfer learned sequence with accuracy and speed to a spatially mirrored sequence. We found that even when participants did not notice the mirrored rules, they showed transfer of learning to the vertically or horizontally mirrored sequence. This result indicates that people could implicitly use the relationship between the learning block and transfer block.

\section*{Acknowledgments}

This work was supported by Grant-in-Aid for JSPS Fellows and Japan Science and Technology Agency (CREST).

\section*{References}

Abrahamse, E. L., Jiménez, L., Verwey, W. B., \& Clegg, B. A. (2010). Representing serial action and perception. Psychonomic bulletin \& review, 17(5), 603-23. doi:10.3758/PBR.17.5.603
Anderson JR (1982) Acquisition of cognitive skill. Psychological Review, 89, 369-406.
Ashby, F. G., Alfonso-Reese, L., Turken, A., \& Waldron, E. (1998). A neuropsychological theory of multiple-systems in category learning. Psychological Review, 105, 442-481.
Bapi, R. S., Miyapuram, K. P., Graydon, F. X., \& Doya, K. (2006). fMRI investigation of cortical and subcortical networks in the learning of abstract and effector-specific representations of motor sequences. Neuroimage, 32(2), 714-27.
Cohen, A., Ivry, R. I., \& Keele, S. W. (1990). Attention and structure in sequence learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 16, 17-30.

Dienes, Z., Kuhn, G., Guo, X. Y., \& Jones, C. (2012). Communicating structure, affect and movement: Commentary on Bharucha, Curtis \& Paroo. In Rebuschat, P., Rohrmeier, M., Cross, I., Hawkins (Eds), Language and Music as Cognitive Systems. Oxford University Press (pp 156-168).
Dienes, Z. \& Longuet-Higgins, H. C. (2004). Can musical transformations be implicitly learned? Cognitive Science, 28, 531-558.
Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. Journal of Experimental Psychology, 47, 381-391.
Grafton, S. T., Hazeltine, E., \& Ivry, R. B. (2002). Motor sequence learning with the nondominant left hand: a PET functional imaging study. Experimental Brain Research, 146, 369-378.
Hikosaka, O., Rand, M. K., Miyachi, S., \& Miyashita, K. (1995). Learning of sequential movements in the monkey: process of learning and retention of memory. Journal of Neurophysiology, 74, 1652-1661.
Hikosaka, O., Sakai, K., Miyauchi, S., Takino, R., Sasaki, Y., \& Putz, B. (1996). Activation of human presupplementary motor area in learning of sequential procedures: a functional MRI study. Journal of Neurophysiology, 76, 617-621.
Hikosaka, O., Nakamura, K., Sakai, K., Nakahara, H. (2002). Central mechanisms of motor skill learning. Current Opinion in Neurobiology, 12, 217-222.
Jiang, S., Zhu, L., Guo, X., Ma, W., Yang, Z., \& Dienes, Z. (2012). Unconscious structural knowledge of tonal symmetry: Tang poetry redefines limits of implicit learning. Consciousness \& Cognition, 21, 476-486.
Kovacs, A. J., Mühlbauer, T., \& Shea, C. H. (2009). The coding and effector transfer of movement sequences. Journal of experimental psychology. Human perception and performance, 35(2), 390-407.
Kuhn, G., \& Dienes, Z. (2005). Implicit learning of non-local musical rules. Journal of Experimental Psychology: Learning, Memory, \& Cognition, 31, 1417-1432.
Nissen, M. J., \& Bullemer, P. (1987). Attentional requirements of learning: evidence from performance measures. Cognitive Psychology, 19, 1-32.
Perruchet, P., \& Pacteau, C. (1990). Synthetic grammar learning: Implicit rule abstraction or explicit fragmentary knowledge? Journal of Experimental Psychology: General, 119 (3), 264-275.
Perruchet, P., \& Pacton, S. (2006). Implicit learning and statistical learning: one phenomenon, two approaches. Trends in cognitive sciences, 10(5), 233-8.
Reber, A.S. ( 1967 ). Implicit learning of artificial grammars. Journal of Verbal Learning and Verbal Behavior , 6, 317-327.
Schmidt, R. A., \& Young, D. E. (1987). Transfer of Move- ment Control in Motor Skill Learning. In S. M. Cormier \& J. D. Hagman, (Ed.), Transfer of learning: Contemporary research and applications (pp. 47-79).San Diego,CA:Academic Press.
Servan-Schreiber, E., \& Anderson, J. R. (1990). Learning artificial grammars with competitive chunking. Journal of Experimental Psychology: Learning, Memory, and Cognition, 16 (4), 592-608.
Shanks, D. R., Rowland, L. a, \& Ranger, M. S. (2005). Attentional load and implicit sequence learning. Psychological research, 69(5-6), 36982.

Stadler, M. A., \& Neely, C. B. (1997). Effects of sequence length and structure on implicit serial learning. Psychological Research, 60, 1423.

Watanabe, K., Ikeda, H., Hikosaka, O. (2006). Effects of explicit knowledge of workspace rotation in visuomotor sequence learning, Experimental Brain Research, 174, 673-678.
Watanabe, K., Ikeda, H., Miyao, M. (2010) Learning efficacy of explicit visuomotor sequences in children with attention-deficit/hyperactivity disorder and Asperger syndrome, Experimental Brain Research, 203, 233-239.
Ziori, E., \& Dienes, Z. (2006). Subjective measures of unconscious knowledge of concepts. Mind \& Society, 5, 105-122.
Ziori, E. \& Dienes, Z. (2008). How does prior knowledge affect implicit and explicit concept learning? Quarterly Journal of Experimental Psychology, 61, 601-624.

\title{
Inferring Subjective Prior Knowledge: An Integrative Bayesian Approach
}

\author{
Sean Tauber (sean.tauber@uci.edu) \\ Mark Steyvers (mark.steyvers@uci.edu) \\ Department of Cognitive Sciences, University of California, Irvine \\ Irvine, CA 92697 USA
}

\begin{abstract}
The standard approach to Bayesian models of Cognition (also known as rational models) requires researchers to make strong assumptions about people's prior beliefs. For example, it is often assumed that people's subjective knowledge is best represented by "true" environmental data. We show that an integrative Bayesian approach-combining Bayesian cognitive models with Bayesian data analysis-allows us to relax this assumption. We demonstrate how this approach can be used to estimate people's subjective prior beliefs based on their responses in a prediction task.
\end{abstract}

Keywords: Bayesian modeling; rational analysis; cognitive models; Bayesian data analysis; Bayesian inference; knowledge representation; prior knowledge

\section*{Introduction}

In the standard approach to Bayesian models of Cognition (also referred to as rational models), researchers make strong assumptions about people's prior beliefs in order to make predictions about their behavior. These models are used to simulate the expected behavior-such as decisions, judgments or predictions-of someone whose computational-level solution to a cognitive task is well described by the model. Analysis of Bayesian models of cognition usually involves a qualitative comparison between human responses and simulated model predictions. For an overview of Bayesian models of cognition see Oaksford and Chater (1998); but also see Mozer, Pashler, and Homaei (2008); and Jones and Love (2011) for a critique.

As an alternative to the standard approach, we present an integrative Bayesian approach that allows us to relax the assumptions about people's prior beliefs. This approach is motivated by previous efforts to infer subjective mental representations (Lewandowsky, Griffiths, \& Kalish, 2009; Sanborn \& Griffiths, 2008; Sanborn, Griffiths, \& Shiffrin, 2010) and more specifically to combine Bayesian models of cognition and Bayesian data analysis (Huszar, Noppeney \& Lengyel, 2010; Lee \& Sarnecka, 2008). The integrative approach allows us to use people's responses on a cognitive task to infer posterior distributions over the psychological variables in a Bayesian model of cognition. It also allows us to estimate probabilistic representations of people's subjective prior beliefs.

We recently applied this approach to a Bayesian cognitive model of reconstructive memory (Hemmer, Tauber, \& Steyvers, in prep). We estimated individuals' subjective prior beliefs about the distribution of people's heights based on their responses in a memory task. The technical requirements for integrated Bayesian inference were
simplified because the posterior distribution, based on inference in the cognitive model, had a simple Gaussian form. This made it straight forward to define individuals' responses as Gaussian distributed random variables in an integrated Bayesian model.

In this study, we develop a method for applying integrated Bayesian inference that does not require the posterior of the cognitive model to have a simple parametric form. We apply this method to a Bayesian cognitive model for predictions that was developed by Griffiths and Tenenbaum (2006). Their Bayesian model of cognition was a computational-level description of how people combine prior knowledge with new information to make predictions about real-world phenomena. They asked participants to make a series of predictions about duration or extent that were similar to the following examples:

> If you were assessing the prospects of a 60-year-old man, how much longer would you expect him to live?

> If you were an executive evaluating the performance of a movie that had made \(\$ 40\) million at the box office so far, what would you estimate for its total gross?

All of the questions used by Griffiths and Tenenbaum (2006) were based on real-world phenomena such as, life spans, box office grosses for movies, movie runtimes, poem lengths and waiting times. Their assumption was that people make predictions about these phenomena based on prior beliefs that reflect their true extents or durations in the real world.
Although it is possible that people's beliefs about these phenomena are tuned to the environment, this assumption cannot be used to explain how people make similar sorts of predictions about counterfactual phenomena that have no true statistics in the environment. For example, consider the following question:

Suppose it is the year 2075 and medical science has advanced significantly. You meet a man that is 60 years old. To what age will this man live?

There is no "true" answer to this question and therefore no environmental data is available. This creates a problem for a Bayesian model of cognition that requires environmental data in order to make predictions.

\section*{Environmental Statistics as Prior Knowledge}

Researchers can use Bayesian models of cognition to simulate the responses that people would make if their computational-level solution to the prediction problem is well described by the model. This process requires that the model includes representations of the prior knowledge people have about the phenomena being predicted. Researchers can represent prior knowledge in their models by collecting real-world environmental statistics and using them in their models as a stand-in for the subjective prior knowledge of individuals (Griffiths \& Tenenbaum, 2006 Hemmer \& Steyvers, 2009a; Hemmer \& Steyvers, 2009b). Representing prior knowledge in this manner is based on the assumption that our knowledge and representations about real-world phenomena are based on actual exposure to these phenomena in the environment. A researcher's best guess at a participant's knowledge is that it reflects, on average, the actual statistics of that phenomenon in the environment.

\section*{Standard Qualitative Analysis}

In the standard approach to Bayesian cognitive modeling, researchers qualitatively compare model predictions to people's responses. The values of psychological parameters-which represent aspects of cognition that are "in people's heads"-are manually specified or estimated with non-Bayesian methods. For a critique of non-Bayesian analysis of Bayesian models, see Lee (2011). The researcher usually encodes subjective prior knowledge in the model using empirical priors (based on environmental data) or by specifying parametric priors with psychological parameters.

A limitation of this method is that researchers do not apply Bayesian inference techniques to participant response data, in order to make inferences about the prior knowledge and psychological parameters represented in the model. It does not allow for the possibility that participants' prior knowledge could be different from the form assumed by the researcher. Furthermore, a model that requires prior knowledge from real-world data cannot be used to generate predictions if the researcher is unable to encode this data in the model. For example, Griffiths' and Tenenbaum's (2006) model cannot be used to generate predictions for the counterfactual future life spans question; even though it involves the same sort of task as the factual prediction questions.

\section*{Quantitative Analysis: An Integrative Bayesian Approach}

The limitations of the qualitative approach can be addressed by reframing a Bayesian model of cognition as a generative process for human response data. Researchers can then use an integrative Bayesian approach to make inferences about the subjective aspects of the cognitive model.

A Bayesian Model of Cognition for Predictions Griffiths and Tenenbaum (2006) had people make simple predictions


Figure 1. Graphical model (observer perspective)
about the duration or extent of real-world phenomena. For example, when told that a man was currently 60 years old, people had to predict the age to which he would live. We refer to the value that is presented in the question as \(t\) and to the person's prediction as \(t_{t o t a l}\). So if a person predicted that the man would live to be 80 years old, then we would have \(t=60\) and \(t_{\text {total }}=80\).

The Bayesian model of cognition proposed by Griffiths and Tenenbaum used nonparametric environmental priors for \(t_{\text {total }}\). We use a modified version of their model in which \(t_{\text {total }}\) has a parametric prior that is Normal, Erlang or Pareto distributed. We add a switch \(c\) that selects which parametric form is used for the prior.

Figure 1 is a graphical representation of our cognitive model for duration and extent from the perspective of the person making predictions (the observer). Shaded nodes represent variables that contain information that is known to the observer. Unshaded nodes contain information that is unknown to the observer.
The model depicts an observer's subjective model of the conditional dependencies between total duration/extent \(t_{\text {total }}\) of phenomena of different types \(c\)-which are determined by the form of the observer's prior knowledge for the domain. The vector \(\boldsymbol{\theta}\) parameterizes prior distribution types such that \(\theta_{1}, \theta_{2}, \theta_{3}\) parameterize Normal, Erlang and Pareto types, respectively. We specify the prior distribution \(t_{\text {total }}\) as:
\[
t_{\text {total }} \sim \begin{cases}\operatorname{Norm}\left(\theta_{1}\right), & c=1  \tag{1}\\ \operatorname{Erlang}\left(\theta_{2}\right), & c=2 \\ \operatorname{Pareto}\left(\theta_{3}\right), & c=3\end{cases}
\]

The time or duration \(t\) from which the observer must predict \(t_{\text {total }}\) is equally likely for all possible values \(0<t<\) t_total. We implemented this in the model by placing a uniform prior on \(t\) :
\[
\begin{equation*}
t \sim \operatorname{Unif}\left(0, t_{\text {total }}\right) \tag{2}
\end{equation*}
\]

When presented with a prediction question with value \(t\), we assume that observers access the relevant prior knowledge of \(t_{-}\)total by determining the prior type \(c\) and the parameter
values \(\theta_{c}\) and then infer a posterior distribution \(P\left(t_{\text {total }} \mid t, c, \boldsymbol{\theta}\right)\) that is described using Bayes' rule:
\[
\begin{align*}
& P\left(t_{\text {total }} \mid t, c, \boldsymbol{\theta}\right) \propto \\
& \begin{cases}\text { Unif }\left(t \mid 0, t_{\text {total }}\right) f\left(t_{\text {total }} \mid \theta_{c}\right), & t \leq t_{\text {total }} \\
0 & , \\
t>t_{\text {total }}\end{cases} \tag{3}
\end{align*}
\]
where,
\[
f\left(x \mid \theta_{c}\right)= \begin{cases}\operatorname{Norm}\left(x \mid \theta_{1}\right), & c=1  \tag{4}\\ \operatorname{Erlang}\left(x \mid \theta_{2}\right), & c=2 \\ \operatorname{Pareto}\left(x \mid \theta_{3}\right), & c=3\end{cases}
\]

Finally, the observer provides a prediction for the total extent or duration. This response is based on the posterior distribution \(P\left(t_{\text {total }} \mid t, c, \boldsymbol{\theta}\right)\), and could be related to the posterior in a number of ways. The response \(R\) could be a sample from the posterior,
\[
\begin{equation*}
R \sim P\left(t_{\text {total }} \mid t, c, \boldsymbol{\theta}\right) \tag{5}
\end{equation*}
\]
or it could be a function of the posterior such as the median, mean or mode. Griffiths and Tenenbaum (2006) modeled predictions as the median of the posterior. We assume that each response is based on a single sample from the posterior. This assumption provides a technical simplification for modeling how people generate a response from the posterior distribution. We will not explore the theoretical implications of this assumption in depth; however, there is evidence supporting a response model that is based on limited samples from a posterior (Vul, Goodman, Griffiths \& Tenenbaum, 2009).

\section*{Applying Bayesian Data Analysis to the Bayesian Model} of Cognition The goal of the researcher is to apply Bayesian data analysis to the Bayesian model of cognition in order to infer the values of \(\theta\) and \(c\) given \(t\) and observer predictions \(R\) about \(t_{\text {total }}\). This requires an integrative application of Bayesian inference from the perspective of the researcher. Each and every value of \(\theta\) and \(c\) for which the researcher wishes to evaluate the posterior likelihood requires Bayesian inference of the posterior likelihood of the observer's response in the rational model given the values of \(\theta\) and \(c\).

From the perspective of the researcher, the responses provided by an observer are the result of a generative process that encapsulates an application of Bayesian inference to a Bayesian model of cognition (fig. 1) resulting in a posterior distribution (Eq. 3) from which the result is sampled. We call this generative process a Bayesian Inference and Response Process (BIRP) and define it as a probability distribution with likelihood function:


Figure 2. Graphical model (researcher perspective)
\(\operatorname{BIRP}(R \mid t, \theta, c) \propto \begin{cases}\operatorname{Unif}(t \mid 0, R) f\left(R \mid \theta_{c}\right), & t \leq R \\ 0, & t>R\end{cases}\)
Figure 2 shows a graphical model from the perspective of the researcher that incorporates a BIRP. In this model the original stimulus \(t\) and the observer responses \(R\) are data that is known to the researcher. The form of the prior distribution used by the observer is indexed by \(c\), and the parameters for the observer's possible prior distributions are all latent (unobserved) variables for which posterior distributions will be inferred. Observer responses \(R\) are generated as samples from the BIRP:
\[
\begin{equation*}
R \sim B I R P(t, \theta, c) \tag{7}
\end{equation*}
\]

The researcher must place suitable hyper priors on the latent prior type \(c\) and latent parameters for the observer prior distributions \(\mu, \sigma, \beta\) and \(\lambda\). We define the deterministic vector \(\boldsymbol{\theta}=\langle(\mu, \sigma), \beta, \lambda\rangle\) for the purpose of notational compactness.

\section*{Experiment}

We described an integrative Bayesian approach that allows us to make inferences about people's subjective beliefs based on their responses in a prediction task. We ran an experiment in order to collect people's predictions for several of the same questions used by Griffiths and Tenenbaum (2006). We also collected predictions for the counterfactual lifespans question.

\section*{Method}

\section*{Participants}

A total of 25 undergraduates from the University of California, Irvine participated in the study and were compensated with partial course credit.

\section*{Materials}

Prediction questions were presented to participants through a web-based survey. There were 8 different question types


Figure 3. Posterior distributions of people's subjective prior types and parameter values from the researcher's perspective. For each of the eight question types the subplot for the indicator variable \(c\) shows the relative posterior probability for each of the prior types (normal, Erlang, or Pareto). The remaining subplots show the posterior distributions of the parameters for these prior types. Parameters that correspond to prior types with zero posterior probability are shown in gray.
and 5 variations of each question. Each variation corresponded to 1 of 5 possible values of \(t\).The survey instructions and 7 of the questions were identical to those used by Griffiths and Tenenbaum (2006). For the unabbreviated questions and survey instructions, refer to Griffiths and Tenenbaum (2006). Below are abbreviated examples of each of the questions with all 5 of the possible \(t\) values included: (1) Predict the age a man will live to if he is currently (18, 39, 61, 83, 96) years old; (2) Predict what the total box-office intake for a movie that has taken in (\$1, \(\$ 6, \$ 10, \$ 40, \$ 100\) ) so far; (3) Predict the length of a movie that has already been playing for (30, 60, 80, 95, 110) minutes; (4) Predict the total length of a poem from which you were just quoted line (2, 5, 12, 32, 67); (5) Predict the total time a pharaoh will be in power if he had already reigned for (1, 3, 7,11, 23) years in 4000 BC; (6) Predict the total years that a \((1,3,7,15,31)\) year member of the U.S. House serve; (7) Predict how long you will be on hold if you have already been holding on the phone for (1, 3, 7, 11, 23) minutes. There was an eighth question that was not part of the Griffiths and Tenenbaum study: Suppose it is the year 2075 and medical science has advanced significantly. You meet a man that is (18, 39, 61, 83, 96) years old. To what age will this man live?

\section*{Procedure}

Each participant made a prediction about all 5 instances of the 8 different types of phenomena for a total of 40 questions. Each prediction was based on one of the five possible values of \(t\). The questions were presented in a different random order for each participant. Only one question was presented on-screen at a time and participants entered their answer in a text-entry box before moving to the next question.

\section*{Inference and Data Analysis}

Responses from each participant were considered for exclusion on a per question-type basis. If any of a participant's five responses for one of the eight questiontypes were below the value of \(t\) that was presented in the question, then all five of that participant's responses for that question-type were excluded for analysis but their responses for other question-types were still included-as long as they passed the inclusion requirement above. The number of participants that were included in the analysis for each question-type was: 24 for life spans; 23 for box office intake; 23 for movie durations; 25 for poem lengths; 24 for pharaoh reigns; 20 for U.S. representative terms; and 25 for lifespans in the future.

We aggregated participant responses for each question such that each response provided an additional data point for Bayesian analysis. We implemented a customized Markovchain Monte Carlo (MCMC) sampler to perform Bayesian inference using the researcher model. To complete the model, we used the following priors:
\[
\begin{array}{ll}
c \sim \operatorname{Cat}\left(\frac{1}{3} \cdot \frac{1}{3}, \frac{1}{3}\right) & \mu \sim \operatorname{HalfNorm}(\text { prec }=.001) \\
\rho \sim \operatorname{Ga}(1,1000) & \sigma=\operatorname{sqrt}(1 / \rho) \\
\beta \sim \operatorname{Ga}(.1, .05) & \lambda \sim \operatorname{Ga}(.1, .05)
\end{array}
\]

\section*{Results}

Figure 3 shows a complete summary of the posterior distributions for the subjective prior types as well as the posteriors for the psychological variables that parameterized the subjective priors. We used people's predictions to infer the posterior probability that their subjective prior knowledge for each domain was best characterized by a Normal, Erlang or Pareto distribution. Although the inference allowed for uncertainty about the form of the



Figure 4. Estimated subjective priors and model predictions. The first row shows our estimates of people's subjective prior beliefs compared with the environmental distributions collected by Griffiths and Tenenbaum (2006). The bottom two rows overlay people's actual responses (black marks) with the posterior predictive distributions (gray shaded areas) of the Bayesian cognitive models for new (unobserved) responses. The posterior predictive probabilities of responses for the environmental prior model (second row) and the estimated subjective prior model (third row) are proportional to the darkness of the gray areas.
subjective prior-in which case some posterior probability would have been assigned to more than one of the possible forms-in every domain, all of the posterior mass was assigned to a single type of distribution.

The top row of Figure 4 shows the estimated subjective priors that people used to make predictions in comparison to the true environmental distributions that were collected by Griffiths and Tenenbaum (2006). The estimated subjective distributions were generated by sampling a prior type and parameter values from the posterior distributions and then using them to generate a sample.

Our estimates of people's subjective priors for life spans, movie runtimes, movie grosses, poem lengths, U.S. representatives' terms and pharaohs' reigns are remarkably similar in form to the true environmental distributions. The subjective priors for life spans, movie runtimes and pharaohs' reigns are shifted slightly to the right compared to the environmental distributions, suggesting that people's prior knowledge for these domains has the same form as the environmental statistics but may not be tuned perfectly to the environment.

People's subjective prior for waiting times was estimated in the same manner as the other priors even though the environmental data was not available. The estimated subjective prior for waiting times was consistent with an Erlang form. Griffiths \& Tenenbaum (2006) were unable to provide estimates of these posteriors using the standard qualitative analysis, but did use non-Bayesian methods to fit people's responses and found that a prediction function based on a Power-Law (Pareto) prior provided the best fit. It is not immediately clear if our disagreement about the form of the subjective prior for phone waiting times is due to differences in our methodology or to differences in the predictions of our respective participants.

A subjective prior for future life spans was estimated even though it is based on a counterfactual scenario and therefore has no true environmental distribution. This subjective prior appears to have a similar form to the prior for actual life spans, but is shifted to the right with an average life span of 105.

The bottom two rows of Figure 4 overlay people's actual responses (black marks) with posterior predictive
distributions from the Bayesian cognitive model for new (unobserved) responses using the environmental prior (second row) and the estimated subjective prior (third row).

The posterior predicitve distributions are generally similar for both the environmental prior model and the estimated prior model. There are some differences in the predictions of the models which are consistent with differences between the estimated and environmental priors. For example, the estimated prior for life spans did not capture an increased risk of death for infants and therfore the estimated model predicts less deaths at a young age than the environmental model does. This can likely be attributed to the limited range of ages (18 to 96 years) presented to participants. The estimated models for movie grosses and representatives’ terms tend to predict higher values than the environmental model, which is consitent with the tendancy of some participants to overestimate these values.

\section*{Discussion}

We demonstrated that an integrative Bayesian approachcombining Bayesian data analysis with Bayesian models of cognition-allowed us to estimate people's subjective prior knowledge based on their responses in a simple prediction task. This approach allowed us to relax the assumption that representations of people's prior knowledge in a rational model should be veridical with environmental statistics.

Although we did not require environmental data to apply an integrative Bayesian approach, having this data allowed us to compare our estimates of people's subjective beliefs to real-world environmental data. We found that people's beliefs about the phenomena in our study were similar in form to the environmental statistics, but that they showed some deviations. At least one of these deviations-related to infant mortality in the life spans question-likely resulted from the limited range of response data that the model used to estimate subjective priors. Other differences between the estimated and environmental priors seem more likely to be the result of deviations between people's subjective beliefs and the environmental statistics. For example, some people tended to overestimate the total gross of movies and the lengths of representatives' terms and pharaohs' reigns. The integrative Bayesian approach is able to provide explanations and predictions that account for these human responses in a way that traditional rational analysis cannot. Furthermore, in situations where a Bayesian model of cognition requires representations of people's prior beliefs and environmental data is unavailable or non-existent-like it was for telephone waiting times and future life spans in our study-an integrative Bayesian framework can still be used to infer subjective priors and make model predictions.

Taking an integrative Bayesian approach opens the door for researchers to take advantage of all of the methods that have been developed for Bayesian analysis of cognitive process models (Lee, 2008) and apply these methods to Bayesian cognitive models. In addition to the estimation of subjective priors and psychological parameters, this method also allows for individual differences in subjective prior
beliefs (Hemmer, et al., in prep). This is important because if people's subjective priors are not tuned to the environment for a particular domain, then it is reasonable to assume that different people have different subjective priors.

\section*{Acknowledgements}

Thank you to Michael Lee and James Pooley for helpful discussions; and to Tom Griffiths and Josh Tenenbaum for providing us with the environmental data that they collected.

\section*{References}

Griffiths, T. L., \& Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. Psychological Science, 17, 767-773.
Hemmer, P. \& Steyvers, M. (2009a). A Bayesian account of reconstructive memory. Topics in Cognitive Science, 1(1), 189-202.
Hemmer, P. \& Steyvers, M. (2009b). Integrating episodic memories and prior knowledge at multiple levels of abstraction. Psychonomic bulletin \& review, 16(1), 80-7.
Hemmer, P., Tauber, S., \& Steyvers, M. (in preparation). Bayesian estimation in rational models.
Huszár, F., Noppeney, U. \& Lengyel, M. (2010). Mind reading by machine learning: A doubly Bayesian method for inferring mental representations. Proceedings of the 32nd Annual Conference of the Cognitive Science Society.
Jones, M. \& Love, B.C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. Behavioral and Brain Sciences
Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. Psychonomic Bulletin \& Review, 15, 1-15.
Lee, M.D. (2011). In praise of ecumenical Bayes. Behavioral and Brain Sciences, 34, 206-207.
Lee, M. D. \& Sarnecka, B.W. (2010). A model of knowerlevel behavior in number-concept development. Cognitive Science, 34, 51-67.
Mozer, M., Pashler, H., \& Homaei, H. (2008). Optimal predictions in everyday cognition: The wisdom of individuals or crowds? Cognitive Science, 32, 1133-1147.
Oaksford, M., \& Chater, N. (Eds.). (1998). Rational models of cognition. Oxford: Oxford University Press.
Sanborn, A. N., \& Griffiths, T. L. (2008). Markov chain Monte Carlo with people. Advances in Neural Information Processing Systems 20.
Sanborn, A. N., Griffiths, T. L., \& Shiffrin, R. (2010). Uncovering mental representations with Markov chain Monte Carlo. Cognitive Psychology, 60, 63-106.
Vul, E., Goodman, N.D., Griffiths, T.L., \& Tenenbaum, J.B.(2009). One and done? Optimal decisions from very few samples. Proceedings of the 31st Annual Conference of the Cognitive Science Society.

\title{
fMRI Study in Insight Problem Solving Using Japanese Remote Associates Test Based on Semantic Chunk Decomposition
}

\author{
Hitoshi Terai (terai@is.nagoya-u.ac.jp) \\ Graduate School of Information Science, Nagoya University, Nagoya, Aichi, 464-8601 Japan \\ CREST, Japan Science and Technology Agency, Chiyoda, Tokyo, 102-8666 Japan \\ Kazuhisa Miwa (miwa@is.nagoya-u.ac.jp) \\ Graduate School of Information Science, Nagoya University, Nagoya, Aichi, 464-8601 Japan \\ Kazuaki Asami (asami@cog.human.nagoya-u.ac.jp) \\ Graduate School of Information Science, Nagoya University, Nagoya, Aichi, 464-8601 Japan
}

\begin{abstract}
This study aims to develop a Japanese version of the remote associates test required to decompose semantic chunks for use not only in behavioral studies but also in brain researches. Further, this study attempted to reveal the relationship between the process of solving insight problems and brain activities. Results of the behavioral data show that the solution time was significantly longer in the chunked than in the non-chunked condition. The imaging data identified the following brain activities. First, the right and left cingulate gyri, related to conflict monitoring, were more activated during the process of searching for a target in the chunked than in the non-chunked condition. Second, the left posterior cingulate gyrus was more activated when the participants could find a target by overcoming constraints as semantic chunks related to emotional process.
\end{abstract}

Keywords: fMRI; insight; remote associates test; chunk decomposition.

\section*{Introduction}

Reproductive thinking is the application of previously acquired knowledge and efficiently solves typical problems. However, in insight problems, such reproductive thinking forms mental sets, preventing solution and leading problem solvers to an impasse (Dominowski \& Dallob, 1995; Ohlsson, 1992; Smith, 1995). Thus, problem solving is accomplished by overcoming such reproductive thinking. Problem solvers often have an "aha!" experience when solving insight problems (Davidson, 1995; Metcalfe, 1986a,b; Metcalfe \& Wiebe, 1987). Such characteristics of insight process, including an impasse and sudden attainment with the emotional experience, have been studied through psychological experiments. Moreover, in recent years, many studies revealed the mechanisms between characteristic insight processes and brain functions (Dietrich \& Kanso, 2010).

Insight problems generally used for related researches (e.g., nine dots problem, candle problem, triangle of coins) require at least several minutes to solve them. Further, once they are solved, they cannot be reused as problem solving tasks among the same participants. In contrast, when employing brain imaging studies, verbal problems including anagram tasks, riddles, and the remote associates test (RAT) have been used. This is owing to the fact that these studies require not only the use of problems that can be solved within several tens of seconds but also can be used repeatedly for the same participants. However, almost all tasks used in brain imaging
studies are premised on using English speaking participants. Therefore, insight tasks based on a wide variety of languages are required to promote the development of research in this field.

Thus, our research purpose is to develop a Japanese version of the RAT for use in brain research using examples from the standard RAT task widely used in current neuroscience studies. In addition to developing the task, we tried to reveal the relationship between brain activities and the process of insight problem solving, including both processes of an impasse and evoking an emotional experience when solutions are found.

\section*{Japanese RAT Required to Decompose Semantic Chunks}

Insight problem solving characterized by an impasse and the suddenness of solution with emotional experience represents a radical representation change. Problem solvers have to reconstruct their erroneous mental representations constructed at an early stage of insight problem solving, whereas they can take step-by-step analytic approaches in non-insight problems. Such representation change in the insight process has been interpreted based on the theories such as the transition of problem spaces and the chunk decomposition by conducting psychological experiments (Kaplan \& Simon, 1990; Knoblich et al., 1999; Ohlsson, 1992). Familiarity with a class of objects and events leads to the creation of patterns as chunks that capture recurring constellations of features or components. Preserving the mental efforts by using chunked knowledge contributes to efficient problem solving. However, if the available chunk does not work in a way that is helpful vis-à-vis finding a solution, it might work to prevent solving the problem. Moreover, once it is constructed, it is difficult to decimate and an impasse might result (Knoblich et al., 1999; Ohlsson, 1992).

RAT consists of sets of three words drawn from a mutually remote associate cluster. Problem solvers are required to find a fourth word which could serve as a specific kind of associative connective link between these disparate words (Mednick, 1962). One example might be a set of three problem words: "arm," "coal," "peach." The answer to the example is the word "pit." The answer word generate three words or
phrases，＂armpit，＂＂coalpit，＂and＂peach pit，＂being connected with each problem word．However，constraints have not been controlled in the RAT used in cognitive neuroscience research （e．g．，Bowden \＆Jung－Beeman，2003）．Thus，in previous studies using RAT－like problems，the definition of obtaining insight was based on finding a solution，solution time，and／or self－reported sudden，unforeseen flash of illumination．Alter－ natively，this study developed a Japanese RAT with control－ lable constraints based on chunk decomposition．The exis－ tence of chunks，which prevents finding association between problem words，would lead problem solvers to search for a target within incorrect problem spaces and arrive at an im－ passe．They might also get an＂aha！＂experience when an impasse based on chunks is resolved，and subsequently the problems are solved．

\section*{Hypotheses}

As demonstrated in previous neuroscience studies，conflicts attributed to constraints preventing problem solving were as－ sociated with activity in the anterior cingulate cortex（ACC） （Aziz－Zadeh et al．，2009；Kounios \＆Jung－Beeman，2009； Luo et al．，2004；Qiu et al．，2008）．Moreover，in the pro－ cess of insight problem solving，a solution seems to arise suddenly，accompanied by an emotional experience generally known as an＂aha！＂experience（Csikszentmihalyi \＆Sawyer， 1995；Davidson，1995；Metcalfe，1986a，b；Metcalfe \＆Wiebe， 1987）．It is known that such emotional experiences in insight problem solving are associated with activity in the posterior cingulated cortex（PCC）（Qiu et al．，2008）．

In this study，which involves the use of an fMRI while con－ ducting psychological experiments，we propose the following two hypotheses regarding brain activities both when strug－ gling with an insight problem and then solving it．

Hypothesis 1 When problem solvers fail to find a solution， the existence of chunks are associated with activity in the ACC because in preventing solutions conflicts arise．

Hypothesis 2 When problem solvers find a solution，the so－ lutions are associated with activity in the PCC related to emotional experience，in addition to activation in the ACC．

\section*{Task}

Figure 1 illustrates the structure of a Japanese RAT that re－ quires decomposition of semantic chunk．Stimuli of the task are presented on a computer screen containing six kanji char－ acters．The purpose of the task was to find a common kanji character（target）with which each of the three kanji charac－ ters presented on the upper row（problem characters）could form a meaningful word．However，distracters presented be－ low the problem characters prevent the finding of the target because the distracters could form meaningful words with each of the problem characters．Therefore，participants are re－ quired to decompose the semantic chunks between the prob－ lem characters and distracters to find the target through re－ mote association．Moreover，the task can control the exis－ tence of the semantic chunks between problem characters and

Distaceas \(\rightarrow\) 郷 癖 盤

異 郷 ：a foreign country \(\underset{\text { 噼 }}{\text { 口：}}\) ：one＇s favorite phrase \(\underset{\text { 盤 }}{\text { 度 }}\) ：the beginning

Figure 1：An Example of a Japanese RAT（required semantic chunk decomposition）
distracters by changing distracters that cannot be connected with the problem words．The kanji characters used in this experiment were known to the participants，who were native Japanese speakers．

\section*{Neuroactivity}

\section*{Methods}

Participants Eighteen healthy，right－handed undergraduate students（aged 19 to 36 years）participated in this experi－ ment．The participants were native Japanese speakers and their handedness was assessed by a modified Oldfield ques－ tionnaire（Oldfield，2004）．The participants provided written informed consent in accordance with the research ethics com－ mittee guidelines of Nagoya University＇s Research Institute of Environmental Medicine．

Design Participants were given a practice session using two examples outside of the MRI scanner．Following the practice session，they engaged in 60 problems while in the scanner． The problems consisted of 30 chunked and 30 non－chunked problems．The chunked and non－chunked problems were counter－balanced between participants（the chunked prob－ lems presented to the half of the participants were treated as non－chunked problem for the other half，and vice versa）．The sequences of problems were also randomized throughout the experimental session．
Figure 2 illustrates the experimental sequence．Each prob－ lem was presented for 30 seconds to the participants．The resting interval between trials was 12 seconds．They were required to press the left button assigned to the index finger of their right hand when finding a target immediately．After pressing the left button，the target，the answer to the problem was presented．Participants were required to press the left button when their answer corresponded to the target，whereas they were required to press the right button assigned to their middle finger when their answers were incorrect．Taking a 10 minute break outside of the scanner，this sequence was re－ peated 60 times．The experiment consisted of two fMRI runs．

Imaging Data Acquisition All scanning whole－brain im－ ages were acquired by using a gradient echo planar image acquisition on a 3T MRI Scanner（Siemens Verio，Siemens Healthcare，Erlangen，Germany）．The functional imaging pa－


Figure 2: Experimental sequence
rameters were \(\mathrm{TR}=2.5 \mathrm{~s}, \mathrm{TE}=30 \mathrm{~ms}, \mathrm{FA}=70^{\circ}, \mathrm{VoF}=20 \mathrm{~cm}\)
20 cm , and 39 slices. To avoid head movement, the participants wore a neck brace and were asked not to talk or move during MRI scanning. High-resolution anatomical images (T1) were also acquired by using gradient echo planar image acquisition. We acquired T 1 images \((\mathrm{TR}=2.5 \mathrm{~s}, \mathrm{TE}=\) \(2.48 \mathrm{~ms}, \mathrm{FA}=8^{\circ}\) ) with 192 sagittal slices, each being 1 mm in thickness. Motion correction was also performed in a standard realign process in SPM8.

Imaging Data Analysis The image data were analyzed using SPM8. Each participants's imaging data was individually preprocessed (realignment, slice time adjustment, coregistration, normalization, smoothing) and the spatially preprocessed data was then estimated to establish a random effects model. Statistical threshold was set at \(p<.001\), uncorrelated with an extended threshold of 10 contiguous voxels.

\section*{Results}

Behavioral Results Results of the solution rates within both 15 and 30 seconds, as shown in Figure 3 (the error bars indicate the standard error). A \(t\)-test showed a significant difference between the two conditions within 15 seconds \((t(17)=2.95, \quad p<.01)\), whereas within 30 seconds no sig-


Figure 3: Solution rate within both 15 and 30 seconds


Figure 4: Solution time within both 15 and 30 seconds
nificant difference was observed \((t(17)=1.90\), n.s. \()\).
Next, we compared the average solution time when the participants could find a target between two conditions (Figure 4). The \(t\)-tests within both 15 and 30 seconds showed significant differences between the two conditions ( 15 sec : \(t(17)=3.67, p<.01 ; 30 \mathrm{sec}: t(17)=3.39, p<.01)\).

These results demonstrate that the search for the targets was prevented more in the chunked than in the non-chunked condition owing to the existence of the semantic chunks.

Imaging Results Next, we compared the brain activations in both the chunked and non-chunked conditions. In the following analyses, we focused on two different trials: participants failed to find a target (failed trials) and participants could find a target (correct trials). In the correct trials, imaging data were analyzed for the entire 30 seconds while the

Table 1: Brain areas that were more activated during searching for a target in the chunked condition when compared with those in the non-chunked condition in the failed trials
\begin{tabular}{|c|c|c|c|c|c|c|c|c|}
\hline Cluster & \multicolumn{3}{|c|}{Voxel} & \multicolumn{3}{|l|}{MNI coordinates} & \multirow[t]{2}{*}{Region} & \multirow[t]{2}{*}{BA} \\
\hline \(k_{E}\) & T & \(Z\) & \(p_{\text {unerr }}\) & x & \(y\) & z & & \\
\hline \multirow[t]{8}{*}{166} & 5.42 & 4.02 & 0.000 & -3 & 5 & 46 & Left cingulate gyrus & 32 \\
\hline & 5.26 & 3.95 & 0.000 & 0 & 2 & 49 & Left medial frontal gyrus & 32 \\
\hline & 5.17 & 3.91 & 0.000 & 0 & -13 & 55 & Left medial frontal gyrus & 6 \\
\hline & 4.52 & 3.58 & 0.000 & 15 & -7 & 46 & Right cingulate gyrus & 24 \\
\hline & 4.41 & 3.51 & 0.000 & -6 & -19 & 49 & Left paracentral lobule & 31 \\
\hline & 4.37 & 3.49 & 0.000 & 6 & -7 & 49 & Right cingulate gyrus & 24 \\
\hline & 4.01 & 3.29 & 0.001 & -12 & -7 & 49 & Left cingulate gyrus & 24 \\
\hline & 3.88 & 3.21 & 0.001 & 6 & 5 & 37 & Right cingulate gyrus & 24 \\
\hline \multirow[t]{2}{*}{13} & 5.19 & 3.92 & 0.000 & 15 & -10 & -2 & Right medial globus pallidus & \\
\hline & 4.48 & 3.56 & 0.000 & -6 & 38 & -14 & Left medial frontal gyrus & 11 \\
\hline \(N=17\); & . 001 & (uncor & ected & 崖 & , & d of & 0 contiguous voxels) & \\
\hline
\end{tabular}

Table 2: Brain area that was more activated until a target was found in the chunked condition when compared with that in the non-chunked condition in the correct trials
\begin{tabular}{|c|c|c|c|c|c|c|c|}
\hline \multicolumn{3}{|c|}{Voxel} & \multicolumn{3}{|l|}{MNI coordinates} & \multirow[t]{2}{*}{Region} & \multirow[t]{2}{*}{BA} \\
\hline \(T\) & Z & \(p_{\text {unoorr }}\) & x & y & z & & \\
\hline 4.23 & 3.41 & 0.000 & -15 & -58 & 10 & Left posterior cingulate & 30 \\
\hline
\end{tabular}
stimuli were presented. In the failed trials, imaging data were analyzed until finding a target.

\section*{Chunked \(>\) Non-chunked (failed trials)}

In the failed trials, this contrast examined the brain areas that were more activated during searching for a target in the chunked condition when compared with those in the nonchunked condition. Several peaks of activation were found, including the right and left cingulate gyri (right BA 24, left BA 24, left BA 32), the left medial frontal gyrus (left BA 32, left BA 11, left BA 6), and the left paracentral lobule (left BA 31) (Table 1). Figure 5 depicts these areas of activation.

\section*{Chunked \(>\) Non-chunked (correct trials)}

In the correct trials, this contrast examined the brain areas that were more activated until a target was found in the chunked condition when compared with those in the non-chunked condition. It was confirmed that the left posterior cingulate (left BA 30) was activated (Table 2). Figure 6 depicts this area of activation.

\section*{Non-chunked \(>\) Chunked (failed trials)}

This contrast revealed no voxels that were significantly more active in the non-chunked than in the chunked condition when the participants could not find targets.

\section*{Non-chunked \(>\) Chunked (correct trials)}

Same as when the participants could not find a target, this contrast revealed no voxels that were significantly more activated in the non-chunked than in the chunked condition when participants found a target.


Figure 5: Brain areas that were more activated during searching for a target in the chunked condition when compared with those in the non-chunked condition in the failed trials


Figure 6: Brain area that was more activated until a target was found in the chunked condition when compared with that in the non-chunked condition in the correct trials

\section*{Discussion}

Results of the behavioral data show that the solution time was significantly longer in the chunked than in the non-chunked condition. Knoblich et al. (1999) constructed an insight task with chunks having different tightness, and displayed that the
problem solving performance while solving problems with a tight chunk declined from those with a loose chunk. Our behavioral results suggest that the semantic chunks, introduced by the Japanese RAT, prevent finding solutions, which is consistent with the related work.

In the following, we will discuss in detail more activated brain areas in the chunked than in the non-chunked condition, both when the participants could find a target and when they could not.
Failed Trials The right and left cingulate gyri (right BA 24 , left BA 24, left BA 32) were more activated during the process of searching for a target in the chunked than in the non-chunked condition. The cingulate gyrus has been widely believed to be related to insight problem solving (Aziz-Zadeh et al., 2009; Kounios \& Jung-Beeman, 2009; Luo et al., 2004; Qiu et al., 2008). For example, Aziz-Zadeh et al. (2009) reported that the ACC is more activated in insight solutions when compared with the search solution while solving anagram tasks. Luo et al. (2004) also showed that relative to the non-"aha" event, the "aha" event was associated with activities in ACC. Botvinick et al. (1999) revealed that the ACC might also be linked with conflict monitoring.

Previous results indicated that the ACC is related to preliminary process to evade impasse in which problem solvers get fixated with incorrect problem spaces as conflict monitoring (Dietrich \& Kanso, 2010). The activation on the cingulate gyrus in our research appears to monitor the process among the competing, an option aroused by semantic chunks and alternatives. This result supposes hypothesis 1 .
Correct trials In addition, when the participants could find a target, the left posterior cingulate gyrus (left BA 30) was more activated until they found a target in the chunked than in the non-chunked condition. Some researches indicated that the retrosplenial cortex, in particular BA 30 and the neighboring posterior cingulate cortex including BA 23 and BA 31 might be associated with the cognitive processing of emotions (Cato et al., 2004; Maddock, 1999; Maddock \& Buonocore, 1997). For example, Cato et al. (2004) showed that activation uniquely associated with word generation to categories with positive or negative versus neutral emotional connotation occurred in the retrosplenial cortex.

One of the essential characteristics of insight problem solving is an impressive and surprising emotional experience upon sudden and discontinuous solution. For example, Csikszentmihalyi \& Sawyer (1995) conducted detailed interviews with creative individuals who have made a creative contribution to the natural sciences, social sciences, arts and humanities, or business/politics, and were generally older than 60 years. The interviewees reported their exciting experiences when receiving insight. Such emotional experience in insight problem solving is known as the "aha!" experience (Davidson, 1995; Metcalfe, 1986a,b; Metcalfe \& Wiebe, 1987). (Qiu et al., 2008) also discussed in their ERP study that the "aha!" feeling might increasingly activate the PCC when Chinese lo-
gogriphs were completed than when they were not solved. In our experiment, the activation of the left posterior cingulate (left BA 30) when the participants found a target suggests that when finding a target, overcoming constraints correlates more with emotional process than without such a constraint.

However, the activation of the cingulate cortex when the participants could not find a target was not confirmed when they could. This result was likely to be caused by the semantic chunks as constraints preventing to solve problems might be decomposed in the early stage of the insight problem solving process when they found a target. Therefore, the activation of the cingulate cortex was not confirmed. These results are partially supported by our hypothesis 2 .

\section*{Conclusion}

This study aims to develop a Japanese version of the RAT required to decompose semantic chunks for use not only in behavioral studies but also in brain researches. Moreover, we tried to reveal the relationship between the brain activity and both process of an impasse and evoking an emotional experience when solutions are found.

Results of the behavioral analysis showed that the Japanese RAT constructed in our research worked well as expected. The imaging data identified the following brain activities. First, the right and left cingulate gyri related to conflict monitoring were increasingly activated during the process of searching for a target in the chunked than in the non-chunked condition. Second, the left posterior cingulate gyrus was more activated when the participants could find a target by overcoming constraints as semantic chunks related to emotional process. These are important initial steps to be taken in the study of the relationship between insight problem solving process and brain activities using the Japanese version of the RAT.

\section*{Acknowledgments}

This study received generous support from the Brain and Mind Research Center, Nagoya University to conduct the experiment using a MR scanner.

\section*{References}

Aziz-Zadeh, L., Kaplan, J. T., \& Lacobni, M. (2009). Aha!: The neural correlates verbal insight solutions. Human Brain Mapping, 30, 908-916.
Botvinick, M., Nystrom, L. E., Fissell, K., Carter, C. S., \& Cohen, J. D. (1999). Conflict monitoring versus selection-for-action in anterior cingulate cortex. Nature, 402, 179181.

Bowden, E. M., \& Jung-Beeman, M. (2003). Normative data for 144 compound remote associate problems. Behavior Research Models, Instruments, \& Computers, 35(4), 634639.

Cato, M. A., Crosson, B., Gökçay, D., Soltysik, D., Wierenga, C., Gopinath, K., et al. (2004). Processing words with emotional connotation: An fMRI study of time course and
laterality in rostral frontal and retrosplenial cortices. Journal of Cognitive Neuroscience, 16(2), 167-177.
Csikszentmihalyi, M., \& Sawyer, K. (1995). Creative insight: The social dimension of a solitary moment. In R. J. Sternberg \& J. E. Davidson (Eds.), The nature of insight (2nd ed., pp. 329-364). Cambridge, MA: MIT Press.
Davidson, J. E. (1995). The suddenness of insight. In R. J. Sternberg \& J. E. Davidson (Eds.), The nature of insight (2nd ed., pp. 125-155). Cambridge, MA: MIT Press.
Dietrich, A., \& Kanso, R. (2010). A review of EEG, ERP, and neuroimaging studies of creativity and insight. Psychological Bulletin, 136(5), 822-848.
Dominowski, R. L., \& Dallob, P. (1995). Insight and problem solving. In R. J. Sternberg \& J. E. Davidson (Eds.), The nature of insight (2nd ed., pp. 33-36). Cambridge, MA: MIT press.
Kaplan, C. A., \& Simon, H. A. (1990). In search of insight. Cognitive Psychology, 22, 374-419.
Knoblich, G., Ohlsson, S., Haider, H., \& Rhenius, D. (1999). Constraint relaxation and chunk decomposition in insight problem solving. Journal of Experimental Psychology: Learning, Memory, and Cognition, 25(6), 1534-1555.
Kounios, J., \& Jung-Beeman, M. (2009). The aha! moment: The cognitive neuroscience of insight. Current Directions in Psychological Science, 18(4), 210-216.
Luo, J., Niki, K., \& Phillips, S. (2004). Neural correlates of the 'aha! reaction'. Brain Imaging, 15(13), 2013-2017.
Maddock, R. J. (1999). The retrosplenial cortex and emotion: New insights from functional neuroimaging of the human brain. Trends in Neurosciences, 22(7), 310-316.
Maddock, R. J., \& Buonocore, M. H. (1997). Activation of left posterior cingulate gyrus by the auditory presentation of threat-related words: an fMRI study. Psychiatry Research: Neuroimaging Section, 75, 1-14.
Mednick, S. A. (1962). The associative basis of the creative process. Psychological Review, 69, 220-232.
Metcalfe, J. (1986a). Feeling of knowing in memory and problem solving. Journal of Experimental Psychology, 12(2), 288-294.
Metcalfe, J. (1986b). Premonitions of insight predict impending error. Journal of Experimental Psychology, 12, 623-634.
Metcalfe, J., \& Wiebe, D. (1987). Intuition in insight and noninsight problem solving. Memory \& Cognition, 15(3), 238-246.
Ohlsson, S. (1992). Information-processing explanations of insight and related phenomena. In M. Keane \& K. Gilhooley (Eds.), Advances in the psychology of thinking (pp. 144). Harvester Wheatsheaf: London.

Oldfield, R. C. (2004). The assessment and analysis of handedness: The edinburgh inventory. Neuropsychologia, 9, 97-113.

Qiu, J., Li, H., Yang, D., Luo, Y., Li, Y., Wu, Z., et al. (2008). The neural basis of insight problem solving: An eventrelated potential study. Brain and Cognition, 68, 100-106.
Smith, S. M. (1995). Getting into and out of mental ruts: A theory of fixation, incubation and insight. In R. J. Sternberg \& J. E. Davidson (Eds.), The nature of insight (2nd ed., pp. 229-251). Cambridge, MA: MIT Press.

\title{
The effect of expertise on encoding of movements and bodily indexes: a study on volleyball players
}

\author{
Alessia Tessari (alessia.tessari@unibo.it) \\ Department of Psychology, University of Bologna, viale Berti Pichat 5 - Bologna , Italy
}

\author{
Giovanni Ottoboni (giovanni.ottoboni@unibo.it) \\ Department of Psychology, University of Bologna, viale Berti Pichat 5 - Bologna , Italy
}

\author{
Roberto Nicoletti (roberto. nicoletti@unibo.it) \\ Department of Communication Sciences, University of Bologna, Via Azzo Gardino, 23 - Bologna , Italy
}

\begin{abstract}
The use of a particular attentional paradigm, the paradigm of Sidedness (Ottoboni, Tessari, Cubelli \& Umiltà, 2005) has highlighted as professional volleyball players differ from nonplayers in the ability to encode specific spatial indexes. The presentation of images of hands of potential adversaries incorporates meanings related to sport that make volleyball athletes sensitive to directional spatial characteristics previously unobserved. What appears to be crucial in the generation of such effect is the ability to predict the direction of an action.
\end{abstract}

Keywords: action prediction, action directionality, body, Sidedness, expertise, sport.

\section*{Introduction}

The ability to anticipate events and actions is essential to interact with the environment in a profitable way. For example, anticipating the movement of the opponent is required to prepare the most appropriate response to counter it during many sport actions. Some behavioral studies have revealed significant differences between athletes and nonexperts, in terms of processing capabilities for both visuolperceptual and motor skills. In ball games, for example, experienced athletes are able to anticipate "where" and "when" the ball will be thrown by the ability to extract the essential information expressed by the movements of the opponent before the ball begins its trajectory (Williams \& Grant, 1999; Aglioti, Cesari, Romani \& Urgesi, 2008). It was also shown that expert athletes are better than nonexperts or non-athletes in recognition and storage of complex patterns of actions (Abernethy, 1990; Allard et al., 1980; Starkes \& Allard, 1983; Starkes, 1987). This increased ability seems to be based on reading the observed gesture's kinematics: experienced athletes are able to anticipate perceptual strategies because they have a wider and consolidated visuo-motor repertoire (Savelsbergh et al., 2002; Williams et al., 2002) acting as the basis for their
visual perception, as well as for the resulting motor execution. These skills seem to be supported also by differences in the activation of premotor and parietal cortical areas (Calvo-Merino, Glaser, Grezes, Passingham \& Haggard, 2005; Calvo-Merino, Ehrenberg, Leung \& Haggard, 2010) and processes of mental simulation involving the activation of the motor areas (e.g. the shot at the basket in basketball players; Aglioti, Cesari, Romani \& Urgesi, 2008).

In general, expert athletes do not possess different structural cognitive characteristics from those of non-experts: they are only more skilled in selecting the most effective signal for detecting a change in the position of the opponents among those available (e.g. the trajectory of the ball or an attack action). For example, it has been found that volleyball players orient attention along the horizontal and vertical axes in significantly different manners (Rizzolatti, Riggio, Dascola \& Umiltà, 1987). In the present study we investigated whether professional volleyball players were able to extract relevant information from the vision of individual parts of the body. In particular, we tried to isolate what directional indices athletes were able to process automatically in response to the presentation of hands that imply motion. We wanted to investigate if volleyball players were able to process the intention to act transmitted by the posture of the presented hands, given that it is an essential information for the game (for example to predict the direction of a possible spike action). Such processing capabilities allow athletes to predict the actions of opponents and to anticipate the motor behavior in order to oppose them. The hypothesis is based on results from studies on the perception of photographs of the body or body-parts giving the impression of movement (eg. Kourtzi \& Kanwisher, 2000; Urgesi, Moro, Candidi, \& Aglioti, 2006) . This processing ability would anticipate what the final position is based on information we already have about "already seen and experienced" movements (Freyd, 1983).

Therefore, the posture of opponent's hand might allow volleyball athletes to predict the direction of the ball.

In the present work we used the paradigm of Sidedness, commonly used to define precisely the spatial coding based on the side of the body which the hand or foot is "connected to" (Ottoboni et al., 2005; Tessari et al., 2012a). For example, a left hand presented from the palm view with the fingers oriented upward activates a right spatial code because it is perceived and represented as the right hand of a human body facing the observer. The same hand, but seen from the back (Figure 1A left panel), actives instead a left spatial code because it is represented to the left of a body that turns its back to the observer (Figure 1B right panel).


Figure 1: Figure 1A illustrates the concept of Sidedness, that is, the mode in which each hand is spatially coded in relation to a body as a function of the posture. Figure 1B shows the stimuli used in Experiment 1 and Figure 1C those of Experiment 2.

\section*{Experiment 1}

Using photographs of oriented hands, we aimed at testing if volleyball players were able to process information on a potential attack action by encoding the direction imprinted
to the ball. As anticipated, we suggested that every hand presentation generates different spatial codes: one linked to the Sidedness and one related to the direction of the action. For example, a right palm hand, since it appears to act from left to right (i.e. the ball would be crushed to the right), should direct the attention of the observer to the right, thus the spatial code of Sidedness would be the opposite one, activating a left Sidedness code, in turn. In this way, if players were able to process both codes, we could observe a mutual cancellation. However, our prediction takes into account also the posture of the hand (i.e. palm vs. back). Back hands, that are not salient in the context of the game, are not expected to induce processing of action direction, but only an encoding of Sidedness. On the contrary, the palm hands are assumed to be coded according to both Sidedness and direction, with a consequent annulment of the two opposite spatial codes.

\section*{Method}

Participants Sixteen right-handed volleyball players belonging to elite teams were tested (mean age= 22 years).
Apparatus and procedures The stimuli were photographs of right and left hands in back and palm views with the forearm ( \(23^{\circ} \mathrm{X} 9^{\circ}\) visible angle), rotated of \(30^{\circ}\) along their ulnar axis as in Tessari et al. (2010a; 2012b). See Figure 1b. A red or blue circle was superimposed in the middle of the hand. The experiment was run using a Pentium III, 512 Mb , connected to a 15 " screen. The experiment was controlled by E-Prime 1.1 (SP3) software (Psychology Software Tools Inc.). The stimuli were 120 for both the back and palm conditions and lasted on the screen for 100 ms , each. The next stimulus appeared after participant's response and no longer than 1000 ms after. Participants were required to respond according to the colour of the circle by pressing one of two keyboard keys ("X" and "."), respectively on the left and the right side. Feedbacks about reaction times (RT), errors and omissions were given after each response (it lasts for 1500 ms ). The response conditions were counterbalanced between subjects.
Results Participants whose error threshold was above \(10 \%\) and RTS 2 standard deviations higher or lower than the overall participants' mean for corresponding and noncorresponding pairings in each block were excluded from the analyses. Data were submitted to a \(2 \times 2\) ANOVA for repeated measures with View (Back vs. Palm) and Correspondence (Corresponding pairings vs. Noncorresponding pairings between hand laterality and response hand) as the within-subjects factors. The two factors (View: \(\mathrm{F}(1,16)=60.03, \quad \mathrm{p}<.001, \quad\) and \(\quad\) Correspondence: \(\mathrm{F}(1,16)=42.65, \quad \mathrm{p}<.001) \quad\) and their interaction \((\mathrm{F}(1,16)=33.48, \mathrm{p}<.001)\) were significant. RTs for the palm hands were faster than those for back hands ( \(\mathrm{M}=312 \mathrm{~ms}\), \(\mathrm{SE}=4.22\) vs. \(\mathrm{M}=340 \mathrm{~ms}, \mathrm{SE}=6.07\) ), and corresponding pairings were responded faster than the non-corresponding ones ( \(\mathrm{M}=318 \mathrm{~ms}\), \(\mathrm{SE}=4.61 \mathrm{~ms}\) vs. \(\mathrm{M}=333 \mathrm{~ms}\), \(\mathrm{SE}=6.51\) ). When stimuli were presented from the back view, RTs were faster for corresponding ( \(\mathrm{M}=325, \mathrm{SE}=6.40\) ) than for non-
corresponding pairings ( \(\mathrm{M}=355, \mathrm{SE}=9.01\) ): two-tailed \(\mathrm{t}(15)=-8.34, \mathrm{p}<.001\). For the palm view no difference emerged \((\mathrm{t}(15)=0.04, \mathrm{p}>.05\); non-corresponding pairings \(\mathrm{M}=312\), \(\mathrm{SE}=5.69\), and corresponding pairings \(\mathrm{M}=312\), SE=6.43). See figure 2a.


Figure 2: The graph shows the values of the reaction times in Experiment 1 (upper part) and Experiment 2 (lower part).

\section*{Discussion}

Compared to a previous study on non-athletes, where an Sidedness effect emerged for both back and palm hands (Tessari et al. 2010a; Tessari et al., 2012b), the effect emerged only for the back hands in volleyball players. As hypothesized, the palm hands were more informative and relevant as confirmed by the significant effect of Posture (faster responses to the palm), but also by the absence of effect for the palm hands which seem to have been encoded for both Sidedness and the direction of the potential action. This way, the spatial codes were opposites and Summing themselves resulted in a null effect. To confirm this interpretation we conducted a second experiment where we tried to rule out the Sidedness effect by hiding the forearm in the stimuli.

\section*{Experiment 2}

It is known that a hand or a foot can be referred to an appropriate body of reference only in the presence of physiological link such as the forearm or the ankle, that comply with the biomechanics laws of the human body (Ottoboni et al., 2005; Tessari et al., 2010b). In the absence of such links the Sidedness effect does not emerge. Following this logic, we presented the same hands of Experiment 1 without forearm to clarify whether the null effect obtained with palm hands was given by the sum of two opposite spatial codes (that generated by action direction and the one of Sidedness). In this condition the Sidedness effect should be deleted while the direction effect should remain.
Participants Eight right-handed volleyball players belonging to new elite team were tested (mean age \(=25\) years).
Apparatus and procedures Apparatus and procedure were those of Experiment 1 but hands were presented without the forearm (see Figure 1c).
Results The two factors (View: F (1,7)=0.06, p > .5, and Correspondence: \(\mathrm{F}(1,7)=0.11, \mathrm{p}>.5)\) were not significant but their interaction was ( \(\mathrm{F}(1,7\) ) \(=13.04, \mathrm{p}<0.01\) ). When stimuli were presented from the back view, RTs did not differ for corresponding ( \(\mathrm{M}=309 \mathrm{~ms}\) ) and noncorresponding pairings \((\mathrm{M}=305 \mathrm{~ms}): \mathrm{t}(7)=1.24, \mathrm{p}=.25\). For the palm view corresponding pairings were in trend faster than the non-corresponding ones ( \(\mathrm{M}=304 \mathrm{~ms}\) vs. \(\mathrm{M}=311 \mathrm{~ms}): \mathrm{t}(7)=-1.55, \mathrm{p}=.08\). See Figure 2.

\section*{Discussion}

Using as stimuli photographs of hands without forearm we got a pure effect of compatibility based on the direction of the attack action for the palm hands. For example, a right palm hand that directs attack action to the right induced a faster response with the right hand. Any effect emerged for the back hands without the forearm neither for Sidedness (as in non-athletes; Ottoboni et al., 2005) nor for direction. Therefore it seems that professional volleyball players are able to encode the palm hands as hands potentially performing a directed action while the back hands does not allow (even to experienced athletes) to extract any relevant information for the game.

\section*{Conclusion}

The purpose of this research was to investigate the type of information that the high-level volleyball players extrapolate at the presentation of hands. These athletes were chosen because the volleyball game requires an excellent visual analysis of the spatial information transmitted from the hands of the opponents: they must be able to recognize in the shortest time potential attack actions, so as to implement the best response behavior. We hypothesized that hands slightly rotated were coded as hands in the process of acting and that they would have activated at least two spatial codes: one generated by the Sidedness (the spatial code
referred to a body) and the other conveyed by the direction of a potential spike action. Hands presented by palm and with the forearm allowed a complete activation of the opposite spatial codes (e.g. a right palm hand was coded as on the left of a body following the Sidedness, but the attack action drove attention to the right). However, back hands have only activated a spatial code based on Sidedness. To ensure that what was encoded was the direction of the potential attack action, we decided to rule out the Sidedness effect by presenting the same hands without the forearm. Indeed, the forearm is necessary to link the hand to a reference body and to generate the Sidedness effect (Ottoboni et al., 2005). In this condition, we only found an effect of directionality for the palm hands. Considering the results that emerge from studies of athletes (eg Kourtessis, Michalapoulou and Derrida, 1998; Nicoletti and Borghi, 2007, for a review) it seems that the athletics tasks requiring motor anticipation and quick response are highly dependent on the level of expertise. The extensive motor experience seems to develop a resonance system specific for the actions of a specific sport discipline, which allows to enhance both the predictive and the anticipatory abilities on the basis of a shared representation between the perceived actions and the actions performed in the sensorimotor repertoire (Aglioti et al., 2008). We can therefore assume that also the highly experienced volleyball players are able not only to effectively process bodily indexes but also the direction of a potential expressed action in contrast with non-athletes that are mainly focused on the normal relations between hand and body (Tessari et al. 2010a). It will be interesting to determine whether the described behavior was developed by the athletes during their career or if, alternatively, this capacity is a precondition that led the professional volleyball players to success.

\section*{Acknowledgments}

The authors thank Adriano Estrerman for the data collection, Carlo Calgaro for the networking and all the sport coaches, Michele Zanin (Sysley - Treviso), Stefano Pascucci (CMC Ravenna), Manuel Armaroli (Villadoro Volley Modena). The authors also thanks CARISBO foundation for the financial support (grant n. 2005/2445)

\section*{References}

Abernethy, B. (1990). Expertise, visual search, and information pick-up in squash. Perception, 19, 63-77.
Aglioti, S. M., Cesari, P., Romani, M., \& Urgesi, C. (2008). Action anticipation and motor resonance in elite basketball players. Nature Neuroscience, 11, 1109-1116.
Allard, F., Graham, S. \& Paarsalu, M.E. (1980). Perception in sport: basketball. Journal of Sport and Exercise Psychology, 2, 14-21.
Calvo-Merino, B., Ehrenberg, S., Leung, D. \& Haggard, P. (2010). Experts see it all: configural effects in action observation. Psychological Research, 74, 400-406.

Calvo-Merino, B., Glaser, D. E., Grèzes, J., Passingham, R. E., \& Haggard, P. (2005). Action observation and acquired motor skills: an FMRI study with expert dancers. Cerebral cortex (New York, N.Y.: 1991), 15(8), 12431249.

Freyd, J. J. (1983). The mental representation of movement when static stimuli are viewed. Perception \& Psychophysics, 33, 575-581.
Kourtzi, Z., \& Kanwisher, N. (2000). Activation in human MT/MST by static images with implied motion. Journal of Cognitive Neuroscience, 12, 48-55.
Nicoletti, R., \& Borghi, A., (2007). Movimento e attività sportive. Il controllo motorio, 10, 220-235.
Ottoboni, G., Tessari, A., Cubelli, R., \& Umiltà, C. (2005). Is Handedness Recognition Automatic? A Study Using a Simon-Like Paradigm. Journal of Experimental Psychology: Human Perception and Performance, 31, 778-789.
Rizzolatti, G., Riggio, L., Dascola, I., \& Umiltá, C. (1987). Reorienting attention across the horizontal and vertical meridians: evidence in favor of a premotor theory of attention. Neuropsychologia, 25, 31-40.
Savelsbergh, G. J. P., Williams, A. M., Van der Kamp, J., \& Ward, P. (2002). Visual search, anticipation and expertise in soccer goalkeepers. Journal of Sports Sciences, 20, 279-287.
Starkes, J.L. \& Allard, F. Perception in volleyball: the effects of competitive stress. Journal of Sport and Exercise Psychology, 5, 189-196 (1983).
Starkes, J.L. Skill in field hockey: the nature of the cognitive advantage. Journal of Sport and Exercise Psychology, 9, 146-160 (1987).
Tessari, A., Ottoboni, G., Bazzarin, V., \& Nicoletti, R., (2010a). L'effetto Sidednes e la codifica del movimento implicito delle mani. Giornale Italiano di Psicologia, 37, 697-706.
Tessari, A., Ottoboni, G., Symes, E., \& Cubelli, R. (2010b). Hand processing depends on the implicit access to a spatially and bio-mechanically organized structural description of the body. Neuropsychologia, 48, 681-688.
Tessari A., Ottoboni G., Baroni G., Symes E., Nicoletti R. (2012a). Is access to the body structural description sensitive to a body part's significance for action and cognition? A study of the sidedness effect using feet. Experimental Brain Reserach, 218, 515 - 525.
Tessari, A., Ottoboni, G., Mazzatenta, A., Merla, A., \& Nicoletti, R. (2012b). Please Don't! The Automatic Extrapolation of Dangerous Intentions. PLoS ONE, 7, e49011. doi:10.1371/journal.pone. 0049011
Urgesi, C., Moro, V., Candidi, M., \& Aglioti, S. M. (2006). Mapping implied body actions in the human motor system. The Journal of Neuroscience, 26, 7942-7949.
Williams, A.M. \& Grant, A. (1999). Training perceptual skill in sport. International Journal of Sport and Exercise Psychology, 30, 194-220
Williams, A. M., Ward, P., Knowles, J. M., \& Smeeton, N. J. (2002). Anticipation skill in a real-world task:
measurement, training, and transfer in tennis. Journal of Experimental Psychology: Applied, 8, 259-270.

\title{
The Simple Advantage in Categorical Generalization of Chinese Characters
}

\author{
Khanh-Phuong Thai (kpthai@ucla.edu) \\ Department of Psychology, 405 Hilgard Avenue \\ Los Angeles, CA 90095 USA
}

Ji Y. Son (json2@calstatela.edu)
Department of Psychology, 5151 State University Drive
Los Angeles, CA 90032 USA

\begin{abstract}
Research in relational learning suggests that simple training instances may lead to better generalization than complex training instances. We examined whether this "simple advantage" extends to category learning in adults with simplified and traditional (more complex) Chinese writing scripts. In Experiment 1, participants learned Chinese characters and their English translations, performed a memorization test, and were asked to generalize their learning to the corresponding characters written in the other script. In Experiment 2, we removed the training phase and modified the tests to examine transfer based purely on perceptual similarities between simplified and traditional characters. We found the simple advantage in both experiments. Training with simplified characters produced better generalization than training with traditional characters, both when generalization relied on recognition memory and on pure perceptual similarities. This finding advances our understanding of how features of a learning opportunity interact with domain-general learning mechanisms to prepare the mind for transfer.
\end{abstract}

Keywords: categorization; generalization; transfer; memory; learning; similarity; features

\section*{Introduction}

We can remember all kinds of details about our experiences in the world but our visual systems have the capacity to ignore all kinds of details as well. Categorization relies on dual processes: attending to similarities while simultaneously ignoring differences. Efficient generalization minimizes the necessary experience with learning instances (e.g., number of learning instances needed or time spent learning) and maximizes appropriate generalization.

Simple instances have been shown to engender rapid learning with selective attention to the right information for the task. Novices briefly trained with simple line drawings of diagnostic features were able to classify chicks with the accuracy of expert chicken sexers (Biederman \& Shiffrar, 1987). Young children who were taught category labels with simple objects were more successful at generalizing to novel category members than when they were shown more complex learning objects (Son, Smith, \& Goldstone, 2008). We refer to this asymmetry of transfer from simple versus complex training instances as the simple advantage.

Most of the research demonstrating the simple advantage have examined learning and transfer of relational concepts in mathematics (Kaminski, Sloutsky, \& Heckler, 2008;

Sloutsky, Kaminski, \& Heckler, 2005; McNeil, Uttal, Jarvin, \& Sternberg, 2009) and science (Goldstone \& Sakamoto, 2003; Goldstone \& Son, 2005). In these relational domains, in order to generalize learning to a new situation, one must pay more attention to structural information rather than superficial details that may differ across instances. Simple learning instances can facilitate such structural extraction by limiting the extraneous details and guiding attention to the right features.

Little is known, however, about whether this simple advantage can also support category generalization, particularly in adults. Although young children are better able to generalize category labels by learning from simplified exemplars (Son, Smith, \& Goldstone, 2008), one might argue that simple learning instances do not benefit adults who are already experts in category learning (relative to infants).

The other side of the argument suggests that the mechanisms underlying infant and adult categorization might not differ significantly (Gureckis \& Love, 2004). For example, research has shown that categorization behavior in infants and adults agree on the basic level (Horton \& Markman, 1980), that infants tend to extract the same prototypes and make the same kind of inferences from category knowledge that adults do (Mervis \& Crisafi, 1980; Baldwin, Markman, \& Melartin, 1993). One exciting possibility is that infants and adults have the same basic categorization generalization "hardware" and only differ in their level of knowledge of the domain. This has been argued for in the analogy literature (e.g., Kotovsky \& Gentner, 1996). To explore this possibility, we train and test English-speaking adults in a novel domain that contains complex and simple corresponding forms: Chinese character scripts.

For a number of political and historical reasons, the traditional Chinese writing system was simplified in 1949. The simplified characters have approximately \(22.5 \%\) fewer strokes than the more complex traditional script (Gao \& Kao, 2002). Several different simplification processes were employed; some based on Chinese history and meaning while others were straightforward perceptual simplifications. As a result, many characters and their components (recurring groups of strokes that make up the characters) took on quite different look (Harbaugh, 2003). Whether these differences between scripts affect the
learnability of characters is the subject of ongoing debate amongst researchers who study Chinese language acquisition (see Chen \& Yuen, 1991; McBride-Chang et al., 2005; Seybolt \& Chiang, 1979). However, there has been little research to examine these differences partially due to complicated issues of aesthetics, history, politics, and tradition. This endeavor, primarily motivated by issues in cognition and learning, may shed light on this debate.

For the purpose of examining the simple advantage in categorization, these rich sets of naturally occurring simple and complex corresponding entities provides an ideally suited domain. As non-Chinese readers lack prior associations with these stimuli, differences in generalization between the scripts may be attributable to differences in the stimuli.

\section*{Purpose of Current Work}

Two studies examined the simple advantage in adults' category generalization with simplified and complex Chinese characters. Does learning with simplified instances lead to greater category generalization than training with complex forms? Secondly, does this simple advantage occur even with minimal prior exposure to simplified forms?

\section*{Experiment 1}

Participants were asked to study flashcards with a Chinese character on one side and an English definition on the other side. After each set, memorization was measured with a match-to-sample task in which students were briefly shown the English definition and had to pick out the matching character out of four answer choices. After the memory test, generalization was measured by the same matching task, except that participants had to match the definitions with characters of the unlearned script. In the Traditional-first condition, participants studied Traditional characters and their English definitions. The Traditional-first memory test involved Traditional characters while the generalization test replaced those choices with corresponding Simplified characters. In the Simple-first condition, participants studied and had a memory test with Simplified characters, but their generalization test had Traditional versions of the learned characters. If simplified learning instances promote generalization, then participants would show better generalization in the Simple-first than in the Traditional-first condition.

\section*{Method}

Participants and Design 14 undergraduates ( 7 females and 7 males) participated for course credit. All reported to having no prior experience with Chinese characters. In this within-subject experiment, half of the participants experienced the Traditional-first condition (learning, memory test, generalization test) before the Simple-first condition while the other half experienced the two conditions in the reverse order.

Materials and Procedures Although there are historical or semantic reasons behind some types of simplification, the subset of characters chosen for this study are perceptually simplified forms of their traditional counterparts. In each pair of characters, up to two components (stroke groups called radicals) of the Traditional characters were omitted to produce their simplified version. Thus, Simplified characters had fewer strokes as well as fewer components. The Simplified characters used had 3-13 strokes per character (average 7.23 strokes), and their Traditional version had \(8-22\) strokes per character (average 14.06 strokes). There were 4 sets of 12 unique character pairs but each participant only studied two of these sets in either the Simplified or Traditional script. The number of omitted strokes, the number of omitted components, the location of the omitted components within each character, and the usage frequency were balanced across the character sets.
In the training phase, each participant received a randomly assigned set of 12 flashcards of either Traditional or Simplified characters according to their assigned condition. Each character was printed in black 36 pt SimSun (宋体) font and the English words were printed in black with 24 pt Calibri font. Participants were told to study the Chinese-English pairs, and that they would be tested on them later. They were not given a time limit for studying and most finished within 15 minutes.

Once participants handed in the flashcards, they were administered the memory and generalization tests on a computer using E-Prime 2.0 (Psychology Software Tools, Inc., Sharpsburg, PA, USA). For both tests, there were 12 trials, one for each of the 12 characters in the training set. A trial began with a fixation cross lasting for 0.5 seconds, followed by an English word for 2 seconds, then 4 Chinese characters. The distractor characters were randomly chosen from the set of trained characters. The inter-trial interval was 1 second. The order of the trials was random across participants. No feedback was provided after each trial, but average accuracy and response time were given at the end of each task. Figure 1 shows a sample trial and procedure.


Figure 1: (a) Exact match test procedure and (b) Generalization test procedure of the Simple-first condition in Experiment 1.

In the memory test, participants chose from Chinese characters identical to those in their training set. The generalization task was set up identically to the
memorization task, except that the answer choices in this task were characters written in the unlearned script. Before the generalization trials, these instructions appeared, "There are two types of scripts in the Chinese written language, Traditional and Simplified. You have just studied characters written in one of these two scripts, and now we would like to see how well you can recognize the same characters written in the other script."

Participants were given a 5-minute break before they were given another set of 12 flashcards with characters written in the other script. The entire procedure was repeated for the second set of characters.

\section*{Results and Discussion}

Proportion correct and average response time data for correct responses are presented in Figures 2 and 3 (see left panels).

Preliminary analyses There were no significant differences among the four sets of characters ( \(p s>.10\) ) and no effect of condition order ( \(p s>.10\) ), so accuracies and response times for each condition were collapsed across those variables.

Memorization and Transfer Results We conducted two 2 x 2 (condition x test type) repeated-measures analysis of variance (ANOVA) for accuracy and reaction time.

Accuracy. There was a main effect of test. Performance on memory test \((M=.99, S D=.03)\) was better than generalization \((M=.86, S D=.10), F(1,12)=26.89, p<\) \(0.001, \eta^{2}=.69\). There was a main effect of condition, \(F(1,12)=8.98, p<.05, \eta^{2}=.43\), and a significant interaction, \(F(1,12)=9.04, p<.05, \eta^{2}=.43\). Post-hoc \(t\)-tests confirmed that although the two conditions exhibited similar memory performance, the Simple-first condition generalized more accurately than the Traditional-first condition. Participants in both Traditional-first ( \(M=.99, S D=.03\) ) and Simple-first ( \(M=.98, S D=.04\) ) conditions successfully learned the word pairs and recognized them equally well, \(t(12)=1.00, p=.34\). Generalization accuracy was significantly higher in the Simple-first condition ( \(M=.91\), \(S D=.06\) ) than in the Traditional-first condition ( \(M=.80\), \(S D=.14), t(12)=3.045, p<.025\), with Bonferroni correction. As predicted, participants who initially learned Simplified characters generalized their learning to the transfer script better than those who learned Traditional characters.

Response Times for Correct Trials (given in seconds per trial). Participants were faster on the memorization trials ( \(M\) \(=2.71, S D=.92)\) than generalization \((M=5.54, S D=2.15)\), \(F(1,12)=46.25, p=.00, \eta^{2}=.79\). Those in the Simple-first condition \((M=3.87, S D=1.40)\) were generally faster than those in Traditional-first condition ( \(M=4.38, S D=1.67\) ), \(F(1,12)=5.24, p<.05, \eta^{2}=.30\). Thus, when participants were trained with Simplified script, they tended to make


Figure 2: Accuracy data from the memorization and generalization tests in Experiment 1 (left panel) and from the exact match and generalization tests in Experiment 2 (right panel). (Error bars: \(\pm 1\) SE)


Figure 3: Response time data of accurate responses from the memorization and generalization tests in Experiment 1 (left panel) and from the exact match and generalization tests in Experiment 2 (right panel). ((Error bars: \(\pm 1\) SE)
more correct matches on both tests and did so faster than those who were trained with Traditional script. There was no significant interaction between condition and test type, \(F(1,12)=1.69, p=.22\).

In summary, when trained with Simplified characters, participants were both faster and more accurate than when trained with Traditional characters. More importantly, even though Simplified and Traditional characters were remembered equally well, Simplified training exemplars led to better generalization than Traditional ones. However, the simple advantage may be dependent on the amount of exposure to the learning instance. In Experiment 2, we ask whether training with Simplified characters is more efficient than training with Traditional characters even without extended training experience.

\section*{Experiment 2}

To extend the findings of Experiment 1, we removed the training phase and modified the memorization and generalization tests to examine matches based purely on perceptual similarity. If simplicity promotes transfer by containing only the relevant perceptual features, then the simple advantage should persist even when generalization relies only on perceptual similarities between simplified and traditional characters.

\section*{Method}

Participants and Design 23 undergraduates ( 10 males, 13 females) who reported having no knowledge of Chinese characters participated for course credit. Experiment 2 was also a within-subject design so order was counterbalanced across participants. Twelve were randomly assigned to participate in the Traditional-first condition before the Simple-first condition, and the other 11 participated in the Simple-first condition before the Traditional-first condition.

Materials and Procedures The stimuli and procedures were nearly identical to Experiment 1. The key difference in Experiment 2 was the lack of a training phase thus participants never connected any of the characters to English meanings. Each trial began with a fixation cross, followed by a Chinese character for 2 seconds, and 4 answer choices. In exact match trials, participants matched characters to identical characters. On the generalization task, participants were shown a character in one script and had to choose the match among characters written in the other script. A sample trial and procedure are shown in Figure 4.
(a) Exact Match Test

(b) Generalization Test


Figure 4: (a) Exact match test procedure and (b)
Generalization test procedure of the Traditional-first condition in Experiment 2.

\section*{Results}

Preliminary analyses Like Experiment 1, there was no effect of character set nor condition order \((p s>.10)\) so the data were collapsed across those variables.

Exact match and generalization test results Average proportion correct and average response time results are presented in Figures 2 and 3 (right panels). Again, we
conducted two \(2 \times 2\) (condition x test type) repeatedmeasures analysis of variance (ANOVA) for accuracy and reaction time.

Accuracy. Results were consistent with findings from Experiment 1. There was a main effect of test such that participants made significantly more correct responses on the exact matching task \((M=.97, S D=.05)\) than on the generalization task \((M=.68, S D=.14), F(1,22)=129.72, p\) \(<.001, \eta^{2}=.86\). There was also a main effect of condition, \(F(1,22)=33.42, \mathrm{p}<.001, \eta^{2}=.60\), and a significant interaction, \(F(1,22)=12.33, p<.01, \eta^{2}=.36\). Post-hoc analyses confirmed that this difference was driven by the differential effect of the sample script on generalization. Follow-up pairwise \(t\)-tests showed no significant difference between the Simplified or Traditional exact match-tosample task, \(t(22)=1.32, p=.20\). However, the Simple-first condition produced significantly better generalization performance \((M=.79, S D=.14)\) than the Traditional-first condition \((M=.57, S D=.18), t(22)=4.83, p<.001\), with Bonferroni correction. Again, as in Experiment 1, training with Simplified characters promoted greater generalization to Traditional characters than vice versa.

Response Times for Correct Trials (given in seconds per trial). There was a main effect of test type, \(F(1,22)=59.46\), \(p<.001, \eta^{2}=.73\), such that participants were faster in the Simple-first condition than in the Traditional-first condition. There was a significant interaction, \(F(1,22)=5.70, p<.05\), \(\eta^{2}=.21\), that suggested that although the Simple-first condition was faster than the Traditional-first condition in the exact-matching task, RTs in the generalization task were similar. Bonferroni-corrected pairwise \(t\)-tests confirmed a significant difference in RTs for the exact matching task for accurate responses between the Simple-first and Traditionalfirst condition, \(t(22)=3.91, p<.01\), and showed no significant difference conditions on generalization, \(t(22)=\) \(1.05, p=.21\).

While there was no difference in accuracy on the exact matching trials, Traditional characters required more time per correct response than Simplified characters (1.55 seconds vs. 1.32 seconds). This result is interesting in light of classic experiments and theories of similarity.

Similar to Podgorny and Garner's (1979) classic work that demonstrated participants judge the similarity of two Ss on a screen faster than two Ws, we also find that some Chinese characters are self-identified faster than others. Tversky's feature-based contrast model of similarity (1977) suggests that complex objects that share a greater number of overlapping features are more self-similar than simple objects. Traditional characters contain more strokes so one might assume that they should be more self-similar and should result in shorter RTs in our exact match task. However, it is important to keep in mind that the distractors in the field were also complex. These complex characters may be more similar to each other thus forcing participants to spend more time to distinguish the target among them.

\section*{General Discussion and Conclusion}

We examined the simple advantage for generalization between simple and complex Chinese scripts. In Experiment 1, participants studied the characters and their English translations before attempting to generalize their learning to the same characters of the unlearned script. In Experiment 2, participants had only brief controlled exposure to the characters before undergoing the generalization test. In both experiments, there was a generalization advantage when the initially shown exemplar was simple.

Contrasting the results of these studies, generalization performance in Experiment 1 was more accurate yet slower than Experiment 2. This pattern is reasonable given the differences in the tasks across experiments. Those in Experiment 1 had to recall the characters from memory when given their English definitions whereas those in Experiment 2 saw exemplar characters immediately before making their choice. Taking more time to recall the trained characters may have helped participants in Experiment 1 generalize more accurately. A longer reaction time is probably less effective when generalization was more purely perceptual.

In the following sub-sections, we will discuss the theoretical and educational implications of these findings.

\section*{Theoretical Implications}

These findings are consistent with results of past research on generalization by shape with young children (e.g., Son et al., 2008): simple instances promote better category generalization. Why are these instances advantageous for transfer? Simple training instances may allow for efficient encoding of the right initial features and/or retrieval of useful representations. Learning from complex characters may be detrimental just by having additional non-diagnostic features that are not present in novel transfer cases. Furthermore, complex instances may generally require greater attentional resources to learn and use.

Adults seem to face similar difficulties in categorization learning as children - that potentially useful and distracting features may not be psychologically separable at the time of learning (Schyns \& Rodet, 1997). Being exposed to a simplified version first may have enabled our adult learners to recognize the complex character as containing the simple character along with other new features. Initial learning with a complex stimulus does not provide a decomposed perceptual vocabulary and thus the learner might miss the shared components between the complex and simple stimuli.

Additionally, this work raises more issues regarding the relationship between similarity, recognition memory, and category generalization. If recognition memory or category generalization is taken as a measure of similarity, this set of results provides further evidence for the asymmetry of similarity. There is an accuracy and/or RT asymmetry between the initially viewed exemplar and the potential matches such that performance is aided by an initially simple exemplar. Furthermore, this work raises the
possibility that similarity judgments based on immediately seen features may operate differently than when based on features retrieved from exemplars in memory.

\section*{Practical Implications}

If the end goal of education is generalization, the simple advantage appears to have broad implications. Even though generalization would likely occur with enough time and resources devoted to training with many complex, detailed instances (e.g., Kellman, Massey \& Son, 2010), the present research suggests that simple training instances may be able to foster generalization more efficiently. Although previous research has directly examined the simple advantage with math and science domains, this research suggests that simple learning instances might also be useful in learning categories in general.

More directly, these results bear on the cognitive role of scripts in Chinese reading. Broadly speaking, there are no measurable differences in reading or spelling between the two scripts (Chan \& Wang, 2003). A few studies suggest that learning to read with simplified characters is more related to visual skills than learning to read traditional characters (Chen \& Yuen, 1991; McBride-Chang et al., 2005). Young children learning to read in mainland China (using simplified script) were more likely to base similarity judgments of characters based on visual characteristics than children from Hong Kong (primarily taught with traditional script) (Chen \& Yuen, 1991). Although further research is necessary to determine whether learning a few characters in a lab setting is similar to learning hundreds of characters to gain literacy, our findings suggest that there might be a benefit of starting with simplified characters. Particularly if the goal is to read both scripts, learning the simplified script may be more helpful for learning the traditional script than the reverse.

Simplified characters contain fewer but more diagnostic components (radicals) so it may be advantageous to treat these recurring radicals as basic orthographic units. Perhaps an emphasis on explicitly learning these units early on may foster better generalization to full blown characters. Research on Chinese literacy (e.g., Tsai \& Nunes, 2003) shows that expert readers are generally quite sensitive to these components. Whether such pedagogical practice supports future learning of new Chinese characters is a question for future research.

However, the relevance of these findings for Chinese literacy is limited in two significant ways. First, the characters used in these studies were only simplified via the component omission process. Future research should incorporate character sets created through other simplification methods such as replacing a complex component (e.g., four dashes) with a simpler one (e.g., a line) to draw broader conclusions about the simple advantage for Chinese reading. Second, reading is more than merely identifying or recognizing characters. Traditional characters include cues to pronunciation and meaning that have been removed in simplified characters.

These cues may be equally or even more important to full fledged reading than ease of recognition.

\section*{Conclusions}

The simple advantage seems to be stable across a variety of tasks and domains, from categorization and object recognition to more complex forms of formal learning. This suggests that this effect stems from domain-general learning mechanisms that bridge or incorporate both perceptual and conceptual learning. In some sense, all learning situations are ill-constrained because a novice does not know which information is relevant or irrelevant. Simplicity supports learning by getting at the heart of this problem: the few features that are presented are all relevant.

\section*{Acknowledgments}

We gratefully acknowledge the support of Phil Kellman, Linda Smith, Robert Goldstone, James Stigler, Everett Mettler, Trinh Tran, Xiaoya Qiu, members of the UCLA Human Perception Lab and the UCLA Teaching and Learning Lab. This research was funded by the Institute of Education Sciences, U.S. Department of Education (Grant R305B080016) to UCLA. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

\section*{References}

Baldwin, D. A., Markman, E., \& Melartin, R. L. (1993). Infants' ability to draw inferences about nonobvious object properties: Evidence from exploratory play. Child Development, 64, 711-728.
Biederman, I., \& Shiffrar, M. (1987). Sexing day-old chicks: A case study and expert systems analysis of a difficult perceptual learning task. Journal of Experimental Psychology: Learning, Memory, and Cognition, 13, 640645.

Chan, L., \& Wang, L. (2003). Linguistic awareness in learning to read Chinese: A comparative study of Beijing and Hong Kong children. In C. McBride-Chang \& H.-C. Chen (Eds.), Reading development in Chinese children (pp. 91-106). Westport, CT: Praeger Press.
Chen, M. J., \& Yuen, J. C.-K. (1991). Effects of pinyin and script type on verbal pro- cessing: Comparisons of China, Taiwan, and Hong Kong experience. International Journal of Behavioral Development, 14, 429-448.
Gao, D.-G., \& Kao, H. S. R. (2002). Psycho-geometric analysis of commonly used Chinese characters. In H. S. R. Kao, C.-K. Leong, \& D.-G. Gao (Eds.), Cognitive neuroscience studies of the Chinese language (pp. 195206). Hong Kong: Hong Kong University Press.

Goldstone, R.L., \& Sakamoto, Y. (2003). The transfer of abstract principles governing complex adaptive systems. Cognitive Psychology, 46, 414-466.
Goldstone, R.L., \& Son, J.Y. (2005). The transfer of scientific principles using concrete and idealized simulations. Journal of the Learning Sciences, 14, 69114.

Gureckis, T. M. \& Love, B. C. (2004). Common mechanisms in infant and adult category learning. Infancy, 5(2), 173-198.
Harbaugh, R. (2003). Chinese characters and culture. Retrieved from http://www. zhongwen. com.
Horton, M. \& Markman, E. (1980). Developmental differences in the acquisition of basic and superordinate categories. Child Development, 51, 708-719.
Kaminski, J. A., Sloutsky, V. M., \& Heckler, A. F. (2008). The advantage of abstract examples in learning math. Science, 320, 454-455.
Kellman, P.J., Massey, C.M \& Son, J. (2010). Perceptual learning modules in mathematics: Enhancing students' pattern recognition, structure extraction, and fluency. Topics in Cognitive Science, 2(2), 285-305.
Kotovsky, L. \& Gentner, D. (1996). Comparison and categorization in the development of relational similarity. Child Development, 67, 2797-2822.
McBride-Chang, C., Chow, B.W.Y., Zhong, Y., Burgess, S., \& Hayward, W. G. (2005). Chinese character acquisition and visual skills in two Chinese scripts. Reading and Writing, 28, 99-128.
McNeil, N. M., Uttal, D. H., Jarvin, L., \& Sternberg, R. J. (2008). Should you show me the money? Concrete objects both hurt and help performance on mathematics problems. Learning and Instruction, 19, 171-184
Mervis, C. \& Crisafi, M. (1980). Acquisition of basic object categories. Cognitive Psychology, 12, 496-522.
Podgorny, P., \& Garner, W. (1979). Reaction time as a measure of inter- and intraobject visual similarity: Letters of the alphabet. Perception \& Psychophysics, 26, 37-52.
Schyns, P.G., \& Rodet, L. (1997). Categorization creates functional features. Journal of Experimental Psychology: Learning, Memory, and Cognition, 23, 681-696.
Seybolt, P. J., \& Chiang, G.K.-K. (1979). Introduction. In P. J. Seybolt \& G. K.-K. Chiang (Eds.), Language reform in China: Documents and Commentary (pp. 1-10). White Plains, NY: M.E. Sharpe.
Shu, H., Chen, X., Anderson, R.C., Wu, N., \& Yue, X. (2003). Properties of School Chinese: Implications for Learning to Read. Child Development, 74, 27-47.
Sloutsky, V.M., Kaminski, J.A., \& Heckler, A.F. (2005). The advantage of simple symbols for learning and transfer. Psychonomic Bulletin \& Review, 12, 508-513.
Son, J. Y., Smith, L. B., \& Goldstone, R. L. (2008).
Simplicity and generalization: Short-cutting abstraction in children's object categorizations. Cognition, 108, 626638.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12, 257-285.
Tsai, K.-C., \& Nunes, T. (2003). The role of character schema in learning novel Chinese characters. In C. McBride-Chang \& H.-C. Chen (Eds.), Reading development in Chinese children (pp. 109-125). Westport, CT: Praeger Press.
Tversky, A. (1977). Features of similarity. Psychological Review, 84 (4), 327-352.

\title{
Searching Semantic Memory as a Scale-Free Network: Evidence from Category Recall and a Wikipedia Model of Semantics
}

\author{
Graham William Thompson (gthompson2@ucmerced.edu) Christopher T. Kello (ckello@ucmerced.edu) \\ Priscilla Montez (pmontez@ucmerced.edu) \\ Cognitive and Information Sciences \\ University of California, Merced \\ 5200 North Lake Road, Merced, CA 95343 USA
}

\begin{abstract}
How is semantic memory structured and searched? Recalling items from semantic categories is a classic assay of semantic memory, and recall dynamics tend to exhibit semantic and temporal clustering, as if memory items are organized and retrieved in clusters. Recent analyses show this clustering to be approximately scale-free in terms of distributions of interretrieval intervals (IRIs). This finding is replicated and extended in the present study by asking participants to type as many animals as they can recall from semantic memory. To begin to explain these results, the organization of semantic memory is modeled as a network based on Wikipedia entries for nearly 6,000 animals. The Wikipedia animal network is found to be scale-free in terms of its degree distribution, and aspects of the network are found to correlate with aspects of recall. Semantic similarity based on Wikipedia entries is found to compare favorably with a measure based on latent semantic analysis. It is concluded that semantic memory processes can be usefully theorized as searches over scalefree networks.
\end{abstract}

Keywords: semantic memory, scale-free networks, Lévy foraging; category recall; latent semantic analysis; Wikipedia

\section*{Introduction}

Category recall is a classic approach to investigating semantic memory. Participants produce as many items from a semantic category as possible in a specified period of time (Bousfield \& Sedgewick, 1944). Items tend to be recalled in clusters. For the category of "animals", for instance, part of a typical sequence might be "lion, tiger, cougar, leopard... kitten, cat, tabby". This sequence contains two groups of semantically similar items, big wild cats followed by house cats. Such clusters can be of varying kinds and sizes, and they tend to correspond with short IRIs, relative to longer pauses when switching from one cluster to the next (Grunewald, Lockhead \& Gregory, 1980).

Clustering seems to be a general feature of semantic memory. Work in this area has a long history, with early experiments showing that, when participants memorize words presented in random order, they tend to recall those words in clusters based on semantic categories (Bousfield \& Sedgewick, 1953). Therefore clustering must be related to memory encoding, retrieval, or both. In clinical work, semantic category recall is used as a diagnostic for mental disorders. Schizophrenic patients, people with semantic dementia and people with Alzheimer's all show specific deficits in category generation tasks (see Murphy, Rich \& Troyer, 2006).

Previous work has sought to account for clustering in category recall with patch foraging models (see Hills, Jones \& Todd, 2012). Patch foraging theorizes semantic memory as a set of patches of similar items. Memory search consists of series of quick retrievals of items from within a patch, interleaved by longer times for switching to the next patch when a sufficient number of items in the current patch have been found. Framed this way, optimal foraging can be expressed in terms of the time to leave a patch. It is optimal to switch when the instantaneous rate of recall per unit time drops below the long-term expected rate of recall (Charnov, 1976). Category switch times, and times in other human search tasks, have been found to be consistent with patch foraging (Cain, Vul \& Mitroff, 2012).

Patch foraging models lead one to expect short and long IRIs corresponding to successive recalls within patches versus between patches, respectively. However, recent work on category generation tasks has examined distributions of IRIs and found them to have no particular mean or means (Rhodes \& Turvey, 2007). When the category recall task was conducted for sufficiently long periods of time (e.g. ten to twenty minutes for recalling animals), IRIs were found to be power law distributed. In particular, the frequency of IRIs were inversely related to their size, P(IRI) \(\sim 1 /\) IRI \(^{\alpha}\). Such power law distributions have no characteristic scale in theory, which means that their means and variances diverge as more samples are drawn. The implication is that the organization of semantic memory is scale-free rather than just patchy.

Power law IRI distributions fall outside the purview of patch foraging models, but they have been studied extensively in animal foraging models (Viswanathan et al., 1996). Unlike patch models, animal foraging models explicitly consider the space in which items are to be found, such as trees and bushes in a meadow where birds are foraging for nuts and berries. Interestingly, the same power law distribution found in category recall is also found in IRIs during foraging for a wide range of species (Sims, Southall \& Humphries, 2003). Theorists have related these findings to so-called Lévy walks (Mandelbrot, 1982), which are random walks with path lengths drawn from a power law distribution. While it is unlikely that foraging paths are literally random Lévy walks, they may capture an important property of foraging. The reason is that Lévy walks sometimes may be efficient search strategies when their exponent \(\alpha\) is near 2 (Viswanathan et al., 2000). Consistent
with this idea, IRIs in animal foraging and category recall have both been found to resemble Lévy walks with \(\alpha \sim 2\) (Rhodes \& Turvey, 2007; Sims, Southall \& Humphries, 2003).

What do these findings tell us about the process of searching semantic memory in category recall? They suggest that simple Lévy walks might characterize much about memory search, but they also might tell us about the structure of semantic memory. Items in memory are often theorized in terms of networks, in which case one is led to ask whether these networks are structured in such a way that searching them results in power law IRIs. As it turns out, recent work on semantic networks provides evidence that their degree distributions also follow an inverse power law, termed scale-free networks (Steyvers \& Tenebaum, 2005).

A network consists of interconnected nodes, and the degree of a node is its number of connections. Scale-free networks are those whose distributions of node degrees follow an inverse power law. Many natural and manmade networks are scale-free, such as power grids, brain networks and the World Wide Web (Strogatz, 2001). Steyvers and Tenenbaum (2005) analyzed three different types of data as reflections of semantic networks: word associations, WordNet entries (Miller, 1995) and Roget's Thesaurus. In all three cases, data were used to link items in a network based on similarity or associative relations, and in all cases networks were scale-free.

This evidence for scale-free networks suggests that items fall into clusters with no characteristic size, analogous to the power law IRI distributions. This analogy suggests that scale-free semantic networks might account for power law IRI distributions, as well as the semantic clustering of items in category recall tasks. In the present study, we collected data in a category recall task, and tested whether recall sequences and IRIs can be explained by a scale-free model of semantic memory.

We draw and expand upon previous studies as follows. Category recall data were collected via typed instead of spoken responses, as in previous studies. This difference allowed us to test whether previous findings replicate when response dynamics are on the order of seconds (typing) instead of milliseconds (speech). Typing also allowed us to test whether the same power law distribution occurs in IRIs, as well within responses (typing durations). If so, we would have evidence that recall processes unfold continuously throughout the task, rather than in alternating stages of recall and response execution (Kawamoto, Kello \& Jones, 1998; Spivey \& Dale, 2006).

We then built a semantic network of animals using over 6000 pages from Wikipedia. We followed an information theoretic method used previously to show that the entirety of Wikipedia can be formalized as a scale-free network (Massuci et al., 2012). We test whether this method replicates when analyzing only one subset domain of Wikipedia, and we test whether the resulting measures of animal similarity and network structure can be used to explain category recall data. We also compare Wikipedia
measures of semantic similarity with those generated from latent semantic analysis (LSA) of linguistic corpora (Landauer \& Dumais, 1997). LSA has become a standard co-occurrence method, whereas Wikipedia is a new encyclopedic method. Thus LSA provided a baseline for evaluating our new method. We end by discussing the implications of our results for Lévy and patch foraging models, semantic memory, and search processes in general.

\section*{Experiment}

\section*{Methods}

Participants and Procedure. Nineteen undergraduates at University of California, Merced participated for course credit. Participants were instructed to recall as many members from the category of "animals" as they could in twenty minutes, after first completing three minutes of practice with naming colors. Responses were typed and recorded using a Flash interface that stored the timing of each key press. Key press times were used to calculate the interval from the end of one response to the start of the next, termed inter-response interval (IRI), and the time from start to end of the response, termed typing duration (TD).

\section*{Results}

The average number of animals produced by each participant was 117 (SD = 38.6). Distributions of IRIs and typing durations were plotted in logarithmic coordinates to gauge whether they resembled power law distributions. As shown in Figure 1, the negative linear relation is indicative of a power law, and multi-model inference tests (Akaike, 1974) confirmed that 4 subjects were best fit by a power law, and the other 15 were best fit by a lognormal (which is akin to a constrained or truncated power law in this case). The deviations from linear at left end of these distributions were due to minima that constrained and thereby distorted the power law relationship. Distortion aside, the slope of these distributions in logarithmic coordinates was near -2 , which replicates the category recall findings of Rhodes and Turvey (2007). Thus memory retrieval dynamics followed the same pattern for slower typed responses, relative to faster spoken responses.

Typing durations also followed the same power law relation, suggesting that memory retrieval is ongoing during response execution. To test whether this result may have been due to variations in response length, typing durations were normalized by the number of letters in each response. As shown in Figure 1, normalized distributions had the same overall shape as the others. Thus response length did not factor into the results.

In addition to IRIs, the category recall task also yields series of recalled animals. Visual inspection of these series (not shown herein) indicated that, as expected, semantically related animals tended to be recalled in close proximity compared with less related animals. Next we describe the Wikipedia semantic network model and test whether it can
account for the relative positioning and clustering of items in recall sequences.


Figure 1. Response histograms in logarithmic coordinates.

\section*{Semantic Memory Modeling}

The network method developed by Masucci et al. (2011) is based on transforming each given Wiki page into a probability distribution over lemmas, and then using JensenShannon divergence (JSD) to measure the distance between two probability distributions. Animal Wiki pages were found using the Dbpedia ontology (Auer, Bizer \& Kobilarov, 2007) which contains a list of all articles in Wikipedia associated with a given tag. A list of 129,027 animal articles in Wikipedia was compiled and all stub articles, redirect pages and articles with less than 500 words of main text were removed. This left us with 5,701 animal pages. Formatting, references, and function words were removed from these pages, and remaining words were lemmatized to collapse across different inflectional forms.

The resulting frequency counts over lemmas on each page were normalized to create probability distributions, and each distribution served as a semantic representation of the corresponding animal. These representations can be used to determine which animals are and are not linked in a semantic network, provided there is a good measure of similarity between probability distributions. Note that two distributions for two given pages may overlap only partially in their corresponding sets of lemmas. This means that our similarity measure must encompass and normalize over varying degrees of overlap.

A well-known measure of similarity between two probability distributions is the Kullback-Liebler (KL) divergence, defined as
\[
\mathrm{KL}\left[\rho_{1}, \rho_{2}\right]=\int_{\Delta} \rho_{1}(x) \ln \frac{\rho_{1}(x)}{\rho_{2}(x)} \mathrm{d} x
\]

This divergence is asymmetric and non-normalized, whereas JSD is a symmetric extension of KL divergence, normalized between zero and one:
\[
\operatorname{JSD}\left[\rho_{1}, \rho_{2}\right]=\frac{1}{2}\left(\operatorname{KL}\left[\rho_{1}, \frac{\rho_{1}+\rho_{2}}{2}\right]+\operatorname{KL}\left[\rho_{2}, \frac{\rho_{1}+\rho_{2}}{2}\right]\right)
\]

JSD can be thought of as providing a measure of how often the same lemmas are used with the same frequency between two Wikipedia pages.

Semantic Network. JSDs were calculated for all pairs of probability distributions, and an undirected semantic network was created by connecting any two animals with a JSD similarity below a given threshold. The threshold was chosen to be just high enough to merge \(90 \%\) of the animals into a single, interconnected network (every node could be reached from every other node by traversing the network, and unconnected nodes were removed from analysis).


Figure 2. Degree distribution of the Wikipedia semantic network in logarithmic coordinates.

The structure of the resulting network was sparse, smallworld and approximately scale-free. Average minimum path length was 3.65 , average clustering coefficient was 0.529 , the diameter of the network was 14 and the degree distribution followed a power law distribution (Figure 2). This finding replicates Masucci et al. (2011) for a subset of Wikipedia, and it provides convergent evidence with Steyvers \& Tenenbaum (2005) that semantic memory can be expressed as a scale-free network.

\section*{Accounting for Category Recall Results}

Semantic networks are often compared with offline human behaviors that can be expressed as structures. Wikipedia provides crystallized, idealized representations of animals, but it stands to reason that online measures of human behavior would be sensitive to these representations. We examine three such measures from our category recall data: 1) Animal response similarity as a function of distance in recall sequences, 2) First-order transitions in recall sequences, and 3) IRIs as a function of shortest path length in the scale-free semantic network.

To provide a benchmark for these three measures, we also computed semantic similarity using LSA. As a cooccurrence method, LSA has strengths and weaknesses compared with our Wikipedia-based method. Its main strength is that LSA can provide a representation for every word in a set of documents, whereas the Wikipedia method can only provide representations for existing entries.

However, each Wikipedia entry unambiguously corresponds to a particular semantic item. By contrast, LSA merges all the different meanings and usages of each given word, like "fish", into a single semantic representation. Also, LSA vectors cannot be combined to form compound word representations that correspond to animals like "flying fish" and "zebra finch".

Responses were corrected for spelling mistakes, and multi-word responses like "black bear" were reduced to their superordinate category, i.e. "bear" in this case. LSA vectors were calculated using a term by term comparison from the "general reading to first year college" corpus. Of the 299 unique animal names produced in the category recall experiment, we were able to compute LSA vectors for 196 of them, and Wiki probability distributions for 293 of them.


\section*{Average JSD Values Between Relative Positions}

Average LSA Values Between Relative Positions


Figure 3. Mean JSD and LSA measures as a function of relative positions, with standard error bars.

Recall Proximity. The first measure we examine is related to evidence marshaled for patch foraging models. Hills et al. (2012) used a co-occurrence method called BEAGLE (Jones \& Mewhort, 2007) to show that items
produced within a patch are more semantically related as a function of proximity in a sequence. Given the evidence for patches of all sizes, and hierarchical nesting of patches as evidenced by power laws, we tested whether the analysis could be extended to all items in category recall sequences, without setting patch boundaries.

JSD and LSA measures were computed between pairs of animal responses in category recall sequences, as a function of the relative position of items, from 1 (adjacent) to 10 (nine intervening items). JSD and LSA measures of semantic similarity were averaged for each relative position, and then normalized by the mean and standard deviation for all pairwise JSD and LSA similarities, across all relative positions. Results are shown in Figure 3.

The JSD and LSA measures produced comparable results showing that semantic similarity decreased as a function of positional distance in sequences. This result confirms visual inspection of sequences, as well as previous research (Bousfield \& Sedgewick, 1953) showing that similar items tend to be recalled in nearby positions. Both measures also compared favorably with the Hills et al. (2012) results, which showed a distinct effect of similarity only for immediately adjacent items in recall sequences. Quantitative differences between LSA and Wikipedia methods were also observed: Compared with Wikipedia, LSA registered relatively higher similarities for immediately adjacent items, but similarity fell off more quickly with increasing distance.

Transitional Probabilities. LSA and JSD measures used in the recall proximity analysis can serve as a basis for predicting performance in category recall. We used LSA and JSD measures to compute first-order transition probabilities as a simple means of predicting each recall item in a sequence, based only on the previous item. This analysis extends the previous one because each transitional probability is computed relative to all possible recall items, which is more akin to a model of category recall that simulates dynamics of semantic memory, e.g. by traversing a scale-free semantic network.

For each recalled item, all animals that could be recalled next were arranged according to the JSD or LSA similarity between them, creating a ranking of possible transitions. Transitions to every ranking were normalized by the total number of JSD or LSA transitions. Probabilities were standardized by setting them relative to random chance. In particular, each probability was divided by random chance for each analysis ( \(1 / 293\) and \(1 / 196\), respectively) to show proportion above or below chance.

As shown in Figure 4, for both JSD and LSA measures, participants were most likely to transition from each animal name to the name estimated to be most semantically similar. JSD transitions to the highest ranked word made up around \(7 \%\) of total transitions, and for LSA it was around 3\%. This effect falls off after the first \(30-50\) most similar items, and is most pronounced for the JSD measure. The JSD measure appears to be better at predicting transitional probabilities compared with the LSA method.


Figure 4. Standardized probability of first-order transitions between items by similarity ranking for JSD and LSA measures.

Accounting for IRIs. The previous two analyses focused on the sequencing of recalled items, but the times between recalls are arguably more at issue in theories of semantic memory. Both patch and Lévy foraging theories aim to explain IRI effects-patch transition times for the former, and IRI distributions for the latter. We tried using JSD and LSA similarities alone, as in the previous two analyses, to account for IRIs. However, they did not correlate reliably with IRIs under any transformation of the data we tried.

Despite the lack of a direct link between semantic similarities and IRIs, a semantic network built from similarities still may account for IRIs by virtue of network structure and dynamics that capture interactions among items in memory. We used the scale-free semantic network reported earlier, based on Wikipedia JSDs, to account for IRIs observed in our category recall experiment. We did not build a network based on LSA values because the lack of semantic representations for word phrases like "tiger shark" prohibited us from creating a sufficiently rich network.

Simulating network dynamics is beyond the present scope, but we accounted for IRIs using a standard measure of network structure that is likely to have a strong influence on network dynamics. Minimum path length is the minimum number of links needed to traverse from one node to another. Minimum path length provides a measure of how
disparate two nodes are in the context of an interconnected network, and will directly impact any walker or spreading activation mechanism used to formulate network dynamics.


Figure 5. Mean logged IRI between words with different minimum path length separation in the semantic network.

Minimum path lengths were computed between all adjacently recalled items in all sequences from the category recall experiment. Minima ranged from 1 to 6, but we only examined pairs from 1 to 4 because there were too few at 5 and 6 to afford analysis (1 each). IRIs were logarithmically transformed due to their heavy tails, and mean log IRIs were computed for each minimum path length, as shown in Figure 5. There was a significant effect of path length on log IRIs, \(\beta=.403, t(1067)=7.44, p<.001\), and they accounted for a significant proportion of variance \(R^{2}=.049, F(1\), 1067) \(=55.36, p<.001\). IRIs were shorter for immediately connected items compared to the baseline mean IRI, and IRIs were progressively greater than baseline as path length increased from 2 to 4 . This result provides evidence that a scale-free semantic network may account for IRIs in category recall experiments, even when semantic similarities alone are not enough to account for the data.

\section*{Discussion}

In the present study, we showed that category recall performance can be modeled using scale-free semantic networks. The category recall data were collected as series of typed responses, and IRI distributions had the same shape as in previous experiments using spoken responses. This replication suggests that the present analyses should generalize to other semantic memory paradigms. Typed responses also indicated that memory dynamics unfold during response execution as well as the pauses between responses, which will be important to account for in future models and simulations.

Previous studies have provided behavioral evidence that semantic memory is organized as a scale-free network (Steyvers \& Tenenbaum, 2005), and we showed that a semantic network of animals built from Wikipedia pages is indeed scale-free. We used a measure of Wikipedia page similarity to account for basic aspects of recall sequences, and we showed that this measure of similarity compared
favorably with a more standard LSA measure. We also showed that a basic aspect of scale-free network structure explained some of the variance in category recall IRIs.

The next step in this line of work is to implement network dynamics to test whether scale-free network structure can account for the scale-free, power law distribution of IRIs observed in category recall tasks. This test will bear on Lévy foraging theories that would explain power law IRIs in terms of random or correlated walkers. A parallel step will be to test whether network dynamics can account for evidence suggesting that foragers spend optimal amounts of time foraging within patches, and switch to new patches when current rates of recall fall below the long-run average. While our model does not have clear delineations between patches, because items fall into nested clusters with no characteristic scale, it may still account for patch evidence. Recent modeling work showed that random walks on semantic networks might account for this evidence without reference to patches (Abbott, Austerweil \& Griffiths, 2012).

Finally, it will be informative to apply the Wikipedia method of network creation to other phenomena of semantic memory, in other semantic domains. In doing so, it will be important to compare this method with a range of cooccurrence methods, as well as other encyclopedic methods.

\section*{Acknowledgments}

This work was supported by a grant from the National Science Foundation, BCS 1031903 (PI Kello). The authors thank members of the Cognitive and Information Sciences graduate group for helpful comments and feedback.

\section*{References}

Abbott, J., Austerweil, J., \& Griffiths, T. (2012). Human memory search as a random walk in a semantic network. In Advances in Neural Information Processing Systems 25 (pp. 3050-3058).

Akaike, H. (1974). A new look at the statistical model identification. Automatic Control, IEEE Transactions on, 19(6), 716-723.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., \& Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. The Semantic Web, 722-735.

Bousfield, W. A., \& Sedgewick, C. H. W. (1944). An analysis of sequences of restricted associative responses. Journal of General Psychology; Journal of General Psychology.

Bousfield, WA. 1953. The occurrence of clustering in recall of randomly arranged associates. The Journal of General Psychology, 49, 229-240.

Cain, M. S., Vul, E., Clark, K., \& Mitroff, S. R. (2012). A Bayesian optimal foraging model of human visual search. Psychological Science, 23(9), 1047-1054

Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. Theoretical population biology, 9(2), 129136.

Gruenewald, P. J., \& Lockhead, G. R. (1980). The free recall of category examples. Journal of Experimental Psychology: Human Learning and Memory, 6(3), 225-240.

Hills, T. T., Jones, M. N., \& Todd, P. M. (2012). Optimal foraging in semantic memory. Psychological review, 119(2), 431-440.

Jones, M. N., \& Mewhort, D. J. (2007). Representing word meaning and order information in a composite holographic lexicon. Psychological Review, 114(1), 1-37.

Kawamoto, A. H., Kello, C. T., Jones, R., \& Bame, K. (1998). Initial phoneme versus whole-word criterion to initiate pronunciation: Evidence based on response latency and initial phoneme duration. Journal of Experimental Psychology: Learning, Memory, and Cognition, 24(4), 862885.

Landauer, T. K., \& Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104(2), 211-240.

Mandelbrot, B. B. (1982). The fractal geometry of nature. Times Books: 232-244.

Masucci, A. P., Kalampokis, A., Eguíluz, V. M., \& Hernández-García, E. (2011). Extracting directed information flow networks: an application to genetics and semantics. Physical Review E, 83(2), 1-6.

Masucci, A. P., Kalampokis, A., Eguíluz, V. M., \& Hernández-García, E. (2011). Wikipedia information flow analysis reveals the scale-free architecture of the semantic space. PloS one, 6(2), e17333.

Murphy, K. J., Rich, J. B., \& Troyer, A. K. (2006). Verbal fluency patterns in amnestic mild cognitive impairment are characteristic of Alzheimer's type dementia. Journal of the International Neuropsychological Society, 12(4), 570-574.
T. Rhodes \& M.T. Turvey. (2007). Human memory retrieval as Levy Foraging. Physica A, 385, 255-260

Sims, D. W., Southall, E. J., Humphries, N. E., Hays, G. C., Bradshaw, C. J., Pitchford, J. W., \& Metcalfe, J. D. (2008). Scaling laws of marine predator search behaviour. Nature, 451(7182), 1098-1102
Spivey, M. J., \& Dale, R. (2006). Continuous Dynamics in Real-Time Cognition. Current Directions in Psychological Science, 15(5), 207-211.

Steyvers, M., \& Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. Cognitive science, 29(1), 41-78. Strogatz, S. H. (2001). Exploring complex networks. Nature, 410(6825), 268-276.

Viswanathan, G. M., Afanasyev, V., Buldyrev, S. V., Murphy, E. J., Prince, P. A., \& Stanley, H. E. (1996). Lévy flight search patterns of wandering albatrosses. Nature, 381(6581), 413-415.

Viswanathan, G. M., Afanasyev, V., Buldyrev, S. V., Havlin, S., Da Luz, M. G. E., Raposo, E. P., \& Stanley, H. E. (2000). Lévy flights in random searches. Physica A: Statistical Mechanics and its Applications, 282(1), 1-12.

\title{
Constraining Bayesian Inference with Cognitive Architectures: An Updated Associative Learning Mechanism in ACT-R
}

\author{
Robert Thomson (thomsonr@andrew.cmu.edu) \\ Department of Psychology, Carnegie Mellon University 5000 Forbes Avenue, Pittsburgh, PA, 15213, USA \\ Christian Lebiere (cl@cmu.edu) \\ Department of Psychology, Carnegie Mellon University \\ 5000 Forbes Avenue, Pittsburgh, PA, 15213, USA
}

\begin{abstract}
Bayesian inference has been shown to be an efficient mechanism for describing models of learning; however, concerns over a lack of constraint in Bayesian models (e.g., Jones \& Love, 2011) has limited their influence as being a description of the 'real' processes of human cognition. In this paper, we review some of these concerns and argue that cognitive architectures can address these concerns by constraining the hypothesis space of Bayesian models and providing a biologically-plausible mechanism for setting priors and performing inference. This is done in the context of the ACT-R functional cognitive architecture (Anderson \& Lebiere, 1998), whose sub-symbolic information processing is essentially Bayesian. To that end, our focus in this paper is on an updated associative learning mechanism for ACT-R that implements the constraints of Hebbian-inspired learning in a Bayesian-compatible framework.
\end{abstract}

Keywords: cognitive architectures; Bayesian inference; Hebbian learning; cognitive models; associative learning;

\section*{Introduction}

Bayesian approaches to reasoning and learning have been successful in such fields as decision-making (Tenenbaum, Griffiths, \& Kemp, 2006), language learning (Xu \& Tenenbaum, 2007), and perception (Yuille \& Kersten, 2006). Most specifically, Bayesian inference has been exceptional in discovering some of the structure of language and word learning with substantially less training than traditional connectionist networks.

Despite their successes, Bayesian models have come under attack for being unconstrained, unfalsifiable, and overly reliant on optimality as an assumption for reasoning (see Jones \& Love, 2011; Bowers \& Davis, 2012 for an exhaustive review; and Griffiths et al., 2012 for a counterargument). While these criticisms are not without merit (nor are the Bayesians' rebuttals fully convincing), the issue of constraints remains a critical argument. It is also not a new argument. Over 25 years ago the constraint argument was leveled against the field of connectionism (Fodor \& Pylyshyn, 1988). Then it was argued that, via several learning rules and organizing principles, any behavior could theoretically be captured by connectionist networks.

The degree that progress has slowed for the explanatory power of connectionist networks is beyond the scope of this paper; however, constraints on neural network development using a common learning rule in a stable cognitivelyplausible architecture have been advanced (O'Reilly, 1998; O’Reilly, Hazy, \& Herd, 2012). By corollary, to address similar concerns, the Bayesian movement needs to develop constraints which balance the computational transparency of
their models with algorithmic and implementation (i.e., neural) level cognitive plausibility.

Interestingly, ACT-R 6.0 (Anderson et al., 2004) is a cognitive architecture which already uses Bayesian-inspired inference to drive sub-symbolic learning (i.e., to generate and update the activation strength of chunks in declarative memory). The architecture is both constrained by learning rules (e.g., activation equations; base-level learning) and neuro-cognitively justified by many studies (Anderson \& Lebiere, 1998; Anderson et al., 2004; Anderson, 2007). While there have been difficulties in adapting some aspects of the Bayesian approach (e.g., in implementations of associative learning), ACT-R serves as an example whereby Bayesian inference can be constrained by a neurallylocalized and behaviorally-justified cognitive architecture. In this sense, ACT-R can act as a bridge between all three layers of Marr's tri-level hypothesis.

For the remainder of this paper, we present an overview of the debate over the applicability of Bayes inference to cognition and argue that ACT-R represents the kind of constraint that addresses criticisms against Bayesian models. We will further describe an updated associative learning mechanism for ACT-R that links Bayesian-compatible inference with a Hebbian-inspired learning rule.

\section*{Bayesian Inference}

The essential feature of Bayesian inference is that it reasons over uncertain hypotheses \((H)\) in probability space (i.e., from \(0-100 \%\) certainty). The Bayes rule is defined as:
\[
P(H \mid D)=\frac{P(D \mid H) \cdot P(H)}{P(D)}
\]
where the posterior probability of an outcome \(P(H \mid D)\) is derived from the likelihood \(P(D \mid H)\) of the hypothesis explaining the data, combined with the prior probability of the hypothesis \(P(H)\), and normalized by the probability of the data \(\mathrm{P}(D)\). Thus, updating one's belief is based on one's prior belief influenced by the likelihood that some new evidence supports this belief. At its core, Bayesian inference is an excellent derivation of the scientific method.

A difference between Bayesian models and connectionist implementations is that Bayes models of human cognition tend to use richer, more structured, and symbolic knowledge than connectionist models, which tend to use more distributed representations operating over less structured input. This level of inference places Bayesian models at the computational level of Marr's tri-level hypothesis, whereas cognitive architectures and connectionist networks operate
more at the algorithmic level (Marr, 1982). By remaining at a higher level of description, it is argued that Bayesian descriptions of cognitive behaviors are better understood as a framework for explaining cognition as opposed to an explanation of how cognitive operations and representations should behave in a given task (Tenenbaum et al., 2011).

This higher level of description leads to many of the criticisms leveled against Bayesian models. We wish to address three related criticisms of Bayesian models: (1) they are unconstrained; (2) they are unfalsifiable; and (3) there is little neuro-scientific evidence to support Bayesian theory. It is easy to see how (2) and (3) follow from (1) since without constraint, it is theoretically possible to redefine the priors and hypothesis space of the model to curve fit to any data. Part of the issue with (3) is that Bayesian description tends to operate at the computational level, yet be described in stronger, more algorithmic terms (e.g., probabilistic population codes; Ma et al. 2006).

These criticisms have led to Bayesian theory being criticized as a 'just-so' story (i.e., that the Bayesian framework commits the ad hoc fallacy; Bowers \& Davis, 2012). However, rebuttals by Griffiths et al. (2012), rather than addressing these criticisms in a constructive manner, countered with essentially a 'you-too' argument. Griffiths et al. (2012) argued that curve-fitting models to data is not an exclusive sin of Bayesian models, however, the transparency with which Bayesian models do so make them easy targets. In fact, as Griffiths et al. counter, criticisms (1) and (2) may be leveled against any model or architecture with sufficient parametric degrees of freedom (which they implicitly argue is a feature of most or all existing models). This argument against architectures had previously been espoused by Roberts and Pashler (2000) over a decade ago.

In a recent Science article by Tenenbaum et al., (2011) Bayesian inference is defined as being synonymous with probabilistic inference. This leads to criticism (2). The difficulty with making 'Bayesian' and 'probabilistic' synonymous terms is that any algorithm that approximates probabilistic reasoning can be argued to be approximating Bayesian inference and thus be essentially Bayesian. Conversely, any Bayesian algorithm that does not successfully reproduce human data can lead to the argument that the issue isn't with the Bayesian algorithm per se, but in the transformation of data into a probability space (e.g., by not having the correct priors or correct hypotheses) or in the lack of human-like limitations of the algorithms to carry out the computations. It is for this reason that some have argued that probabilities are "epistemologically inadequate" (McCarthy \& Hayes, 1969).

Instead of offering more criticisms, we wish to offer solutions. The issue with constraints is that, even if Bayesian models do not have too many parameters, there is effectively unlimited freedom in setting priors and the hypothesis space (which greatly influences the performance on the model). What is needed is a way to constrain the generation of the initial probability space and set of algorithms to carry out inference for a set of models. For
instance, Kruschke (2008) reviewed two Bayesian models of learning backward blocking in classical conditioning, the first using a Kalman filter (Dayan, Kakade, \& Montague, 2000) and the other using a noisy-logic gate (Danks, Griffiths, \& Tenenbaum, 2003). Both models gave substantively different predictions, with the Kalman filter model unable to reproduce human behavior.

Furthermore, there are several tasks whose results do not readily fit within a naïve Bayesian explanatory framework. For instance, simple Bayesian models do not capture violations of the sure-thing principle. Given a random variable \(x\) that has only two possible outcomes A or B, naïve Bayesian inference requires \(p(x)\) to fall between \(p(x \mid A)\) and \(p(x \mid B)\). A violation occurs when \(p(x)>p(x \mid A)\) and \(p(x)\) \(>p(x \mid B)\) or vice versa. Shafir and Tversky (1992) showed this violation of the sure-thing principle in a prisoner's dilemma task. Finding these unintuitive results that naïve Bayes models do not easily address, and finding constrained parameter learning rules (such as the noisy-logic gate) provides much needed constraints and falsifiability to the Bayesian framework. Rather than being seen as antiBayesian results, these models should be seen as shaping the boundaries of Bayesian explanatory power.

Finally, while there is contested neuro-scientific evidence as to neural assemblies firing probabilistically, this does not necessarily imply a Bayesian implementation-level explanation, but instead implies the softer claim of a Bayesian-compatible behavioral explanation of neural phenomena, especially when the Bayesian inferences are justified within a neurally-plausible cognitive architecture.

In considering many of the criticisms of Bayesian theory, it is important to note that more research needs to be done to find constraints. As we previously argued, connectionist networks were not sufficiently constrained until sufficient model testing was performed and architectures developed using a common learning rule and constrained set of parameters. For the Bayesian framework, we argue that all of criticisms (1) - (3) can be addressed by situating Bayesian inference within a cognitive architecture, and furthermore that ACT-R 6 is already such an architecture.

\section*{The ACT-R Architecture}

ACT-R is a computational implementation of a unified theory of cognition. It accounts for information processing in the mind via task-invariant mechanisms constrained by the biological limitations of the brain. ACT-R 6 includes long-term declarative memory and perceptual-motor modules connected through limited-capacity buffers. Each module exposes a buffer, which contains a single chunk, to the rest of the system. A chunk is a member of a specific chunk type, and consists of a set of type-defined slots containing specific values.

The flow of information is controlled by a procedural module implemented using a production system, which operates on the contents of the buffers and uses a mix of parallel and serial processing. Modules may process information in parallel with one another. So, for instance, the visual and motor modules may both operate at the same
time. However, there are two serial bottlenecks in process. First, only one production may execute during a cycle. Second, each module is limited to placing a single chunk in a buffer.

Each production consists of if-then condition-action pairs. Conditions are typically criteria for buffer matches, while the actions are typically changes to the contents of buffers that might trigger operations in the associated modules. The production with the highest utility is selected to fire from among the eligible productions. In general, multiple production rules can apply at any point. Production utilities, learned using a reinforcement learning scheme, are used to select the rule that fires.

When a retrieval request is made to declarative memory (DM), the most active (highest \(A_{i}\) ) matching chunk is returned:
\[
A_{i}=B_{i}+S_{i}+P_{i}+\varepsilon_{i}
\]
where activation \(A_{i}\) is computed as the sum of base-level activation ( \(B_{i}\) ), spreading activation ( \(S_{i}\) ), partial matching \(\left(P_{i}\right)\) and stochastic noise \(\left(\varepsilon_{i}\right)\). Spreading activation is a mechanism that propagates activation from the contents of buffers to declarative memory proportionally to the strength of association between buffer contents and memory chunks. Partial matching is a mechanism that allows for chunks in memory that do not perfectly match a retrieval request to be recalled if their activation overcomes a similarity-based mismatch penalty.

\section*{ACT-R as a Constrained Bayesian Architecture}

ACT-R's sub-symbolic activation formula approximates Bayesian inference by framing activation as log-likelihoods, with base-level activation \(\left(B_{i}\right)\) as the prior, the sum of spreading activation and partial matching as the likelihood adjustment factor(s), and the final chunk activation \(\left(A_{i}\right)\) as the posterior. The retrieved chunk has an activation that satisfies the maximum likelihood equation.

ACT-R provides the much needed constraint to the Bayesian framework through the activation equation and production system. The calculation of base-levels (i.e., priors) occurs within both a neurally- and behaviorallyconsistent equation:
\[
B_{i}=\ln \left(\sum_{j=1}^{n} t_{j}^{-d}\right)
\]
where \(n\) is the number of presentations for chunk \(i, t_{j}\) is the time since the \(j^{\text {th }}\) presentation, and \(d\) is a decay rate (community default value is .5). This formula provides for behaviorally-relevant memory effects like recency and frequency, while providing a constrained mechanism for obtaining priors (i.e., driven by experience). Thus, we can address the constraint criticism (1) through this well justified mechanism (see Anderson et al., 2004).

In addition, the limitations on matching in the production system provide constraints to the hypothesis space and kinds of inferences which can be made. For instance there are constraints on the kinds of matching that can be accomplished (e.g., no disjunction, matching only to specific chunk types within buffers) and, while userspecified productions can be task-constrained, the
production system can generate novel productions (through proceduralization) using production compilation. In addition, the choice of which production to fire (conflict resolution) also constrains which chunks (i.e., hypotheses) will be recalled (limiting the hypothesis space), and are also subject to learning via production utilities.

In production compilation, a new production is formed by unifying and collapsing the conditions of the production, and possibly automatizing a given memory retrieval. This new production has a unique utility and can be considered an extension of the hypothesis space; perhaps with enough learning compiled productions are more analogous to overhypotheses (Kemp, Perfors, and Tenenbaum, 2007).

In summary, the conflict resolution and production utilities algorithms both constrain the hypothesis space and provide an algorithm for learning how the space will evolve given experience, constrained within the bounds of a neurally-consistent functional cognitive architecture. This bridges Bayesian inference from a computational-level framework within an algorithmic-level architecture. However, this argument for constraint is not without criticisms (some of which will be addressed in the Discussion). As an example of increasing constraints and grounding mechanisms, we will now present an updated associative learning mechanism in ACT-R.

\section*{Associative Learning}

Associative learning - the phenomenon by which two or more stimuli are associated together - is ubiquitous in cognition, describable as both a micro (Hebbian learning between neurons) and macro (classical and operant conditioning) feature of behavior. Associative learning is a flexible and stimulus-driven mechanism which instantiates many major phenomena such as classical conditioning, context sensitivity, non-symbolic spread of knowledge, and pattern recognition (including sequence learning and prediction error). At the neural level, associative learning is the process by which cells that fire together, wire together.

In its simplest form, Hebbian learning can be described as: \(\Delta W_{i j}=x_{i} x_{j}\), where \(W_{i j}\) is the synaptic strength of the connection between neurons \(i\) and \(j\), and \(x_{i}\) and \(x_{j}\) are the inputs to \(i\) and \(j\) (Hebb, 1949). When both \(i\) and \(j\) are active together, \(\mathrm{W}_{i j}\) is strengthened. While the traditional Hebbian rule was unstable due to a lack of mechanisms to control for weakening of connections (i.e., long-term depression; LTD) or to set a maximum state of activation (i.e., to implement a softmax equation; Sutton \& Barto, 1998), several variants have addressed these issues to provide a stable learning rule.

At a macro level, associative learning is a mechanism where, when a stimulus is paired with a behavior, future presentation of the stimulus primes this behavior. Models of classical conditions are a common macro-level application of associative learning. At this level, associative learning allows animals and humans to predict outcomes based on prior experience with learning mediated by the degree of match between the predicted outcome and the actual result (Rescorla \& Wagner, 1974; Pearce \& Hall, 1980).

While macro-level models are normally processed at a more symbolic level, micro-level sub-symbolic processing can capture statistical regularities from the environment without recourse to explicitly coding context information. There is evidence that humans do not explicitly encode positional information when sequentially recalling a list of items, yet ACT-R's model of list memory required explicit position information to drive recall (Anderson et al., 1998).

Despite being a pervasive factor of human intelligence, associative learning is no longer directly implemented in ACT-R. One reason for this absence is due to difficulties in scaling models in its Bayesian implementation of associative strengths, which treated both the activation strength and associative strength of knowledge elements (e.g., chunks) as likelihoods of successful recall.

\section*{Bayesian Associative Learning Rule}

Associative learning was deprecated in ACT-R 5 due to a lack of scalability in spreading activation as the number of chunks in a model increased and as new productions fired (i.e., new contexts generated). Instead, a simpler spreading activation algorithm was used. The reason for this was that the Bayesian formula used to calculate strength of association \(\left(S_{j i}\right)\) led to some unintended consequences which would render larger and longer-running models unstable.

In ACT-R 4/5, the strength of association \(\left(S_{j i}\right)\) represented the \(\log\) likelihood ratio that chunk \(N_{i}\) was relevant given context \(\mathrm{C}_{j}\) :
\[
S_{j i}=\ln \left(\frac{P\left(N_{i} \mid C_{j}\right)}{P\left(\bar{N}_{l} \mid C_{j}\right)}\right)=\frac{P\left(N_{i}\right)}{P\left(\bar{N}_{l}\right)} \prod_{j} \frac{P\left(C_{j} \mid N_{i}\right)}{P\left(C_{j} \mid \bar{N}_{l}\right)}
\]

When \(C_{j}\) is usually not in the context when \(N_{i}\) is needed, \(P\left(N_{i} \mid C_{j}\right)\) will be much smaller than \(P\left(\bar{N}_{l} \mid C_{j}\right)\) and the \(S_{j i}\) will be very negative because the log-likelihood ratio will approach 0 . In a long-running model, these chunks may have been recalled many times without being in context together, leading to strongly inhibitory \(S_{j i}\).

Once a connection was made, the initial prior \(S_{j i}\) was set by the following equation:
\[
S_{j i}=\ln (m / n)
\]
where \(m\) is the total number of chunks in memory and \(n\) is the number of chunks which contain the source chunk \(j\). This ratio is an estimation of the likelihood of retrieving chunk \(i\) when \(j\) is a source of activation. As a convenience unconnected chunks were set at \(50 \%\) likelihood. \({ }^{1}\)

As can be seen from the previous two equations, given sufficient experience or sufficient numbers of chunks in the model, these context-ratio equations specify that \(S_{j i}\) values will become increasingly and unboundedly negative as more chunks are present in the model and more unique contexts experienced. This is a direct result of \(S_{j i}\) reflecting the statistics of retrieval of chunk \(j\) given that source \(i\) is in the context, and is a version of the Naïve Bayes Assumption.

The issue is with the ratio-driven global term \(\left(C_{j}\right)\) which alters \(S_{j i}\) values for a chunk whenever a new chunk is added

\footnotetext{
\({ }^{1}\) Before \(C_{j}\) appears in a slot of \(N_{i}\), the total probability of retrieving a chunk unconnected to \(C_{j}\) is 0 (which means \(S_{j i}=-\infty\) ).
}
and/or production fires, and is magnified by the loglikelihood calculation which penalizes the inevitable low context ratio in long-running models.

\section*{Spreading Activation in ACT-R 6}

Due to the abovementioned issues with scalability, associative learning was deprecated in ACT-R and a simpler spreading activation function was implemented that does not activation, but instead spreads a fixed amount of activation:
\[
S_{j i}=\operatorname{smax}-\ln \left(\text { fan }_{j i}\right)
\]
where smax is a parameterized set spread of association (replacing the \(m\) term from the previous equation), and fan \(_{j i}\) is the number of chunks associative with chunk \(j\) (the \(n\) term). \(F a n_{j i}\) is traditionally the number of times chunk \(j\) is a slot value in all chunks in DM and represents interference.

With a default smax usually between 1.5 and 2 (Lebiere, 1999), this means that a chunk can appear as a value in 6-8 chunks before becoming inhibitory. In the context of a modeling a single session psychology experiment this may be reasonable, but if ACT-R models long-term knowledge effects, then \(S_{j i}\) will become inhibitory for most chunks. \({ }^{2}\)

As previously discussed, associative learning is a ubiquitous mechanism in both human and animal cognition, which serves as a kind of statistical accumulator which is applicable at both the micro (neural) and macro (cognitive) behavioral level. It seems that to abstract this essential learning mechanism, we are losing out on the exact kind of human-model comparisons that might provide evidence for these much-needed constraints. Perhaps, it is in part for this reason that ACT-R (and other cognitive architectures) have had their explanatory power limited due to a lack of newer, more complex models being built from extant successful models (ACT-R Workshop, 2012).

To both reconcile the difficulties in previous implementation of associative learning and show how we can constrain Bayesian-compatible inference in a cognitive architecture, we will now present a Hebbian-inspired associative learning rule influenced by spike-timing dependent plasticity (STDP; Caporale \& Dan, 2008).

\section*{Hebbian-Inspired Associative Learning Rule}

The major issues with the Bayesian associative learning rule were the reliance on ratio-driven log-likelihoods and the fact that context \(\left(C_{j}\right)\) was a global term which altered \(S_{j i}\) whenever a new chunk was created and whenever a production fired. This is due to the fact that low loglikelihoods become strongly inhibitory, and the generation of context-based ratios necessitates low-likelihoods in a long-running model. In short, this Bayesian account based on the Naïve Bayes Assumption does not adequately capture some of the features of associative learning such as locallydriven strengthening of associations and bounded decay.

An alternative framework is to eliminate the ratio function and remove the global nature of context, while also moving to a frequency-based algorithm instead of a probabilitybased algorithm. The former removes the aforementioned

\footnotetext{
\({ }^{2}\) After presenting this at the 2012 ACT-R Workshop, a flag was written in ACT-R to set a floor of 0 in the \(S_{j i}\) computation.
}
issues with scalability, while the latter eliminates \(S_{j i}=\) \(\lim _{x \rightarrow 0}(\ln x)=-\infty\), where \(x\) is the likelihood. That said, a benefit of using log-likelihood in probability space is that there is no need to squash activation strength (e.g., use a softmax rule to keep \(S_{j i}\) values from overwhelming \(B_{i}\) in the activation equation) because likelihoods cannot go above \(100 \%\) while frequency-based Hebbian activations can theoretically grow unbounded. Thus, the switch to frequencies is about reshaping the range of \(S_{j i}\) values and making \(S_{j i}\) independent of changing global context.

Basing associative learning on frequencies also adds a more Hebbian flavor to the algorithm. Learning, rather than being a global property of the system (as in the Bayesian mechanism) is instead a local property based on cooccurrence and sequential presentation. As previously discussed, our Hebbian-inspired mechanism is influenced by STDP. Unlike traditional Hebbian implementations which simply give a bump to association so long as the presynaptic and post-synaptic neurons both fire within a given temporal window, in STPD if the pre-synaptic neuron fires before the post-synaptic then the association is strengthened (long-term potentiation; LTP). Conversely, if the postsynaptic neuron fires before the pre-synaptic then the association is inhibited (long-term depression; LTD).

This theory of neural plasticity was adapted to our modeling approach by assuming that the sources of activation from chunks in buffers act similarly to presynaptic firings, and the set of chunks in the buffers at the time the new chunk is retrieved is similar to post-synaptic firings. The associative learning rule fires when a request is made to retrieve a chunk from declarative memory. First, a positive phase occurs (LTD; or Hebbian) where the current contents of the buffers spread activation and a new chunk is retrieved. The association between this new chunk and the sources of activation are strengthened according to standard Hebbian learning rules. However, once this new chunk is placed in the retrieval buffer, a negative phase occurs (LTP; or anti-Hebbian) where the retrieved chunk will negatively associate with itself and with its context. In formal terms:
\[
\begin{aligned}
\Delta S_{j i} & \alpha \cdot F\left(N_{i} \mid C_{j}^{p r e}\right) \\
\Delta S_{j i} & =-\alpha \cdot F\left(N_{i} \mid C_{j}^{\text {post }}\right)
\end{aligned}
\]
where \(\alpha\) is a Hebbian learning term, \(F\left(N_{i} \mid C_{j}^{\text {pre }}\right)\) is the context of source chunks \(C_{j}^{\text {pre }}\) at the time of the retrieval request for chunk \(N_{i}\), and \(F\left(N_{i} \mid C_{j}^{\text {post }}\right)\) is the context of chunks \(C_{j}^{\text {post }}\) after chunk \(N_{i}\) has been retrieved. Note that only changes in context will have a net \(\Delta S_{i i}\) due to the balanced positive and negative learning phase. Furthermore, these associations are not symmetric (i.e., \(S_{j i} \sim=S_{i j}\) ).

This balanced Hebbian/anti-Hebbian mechanism is geared towards developing a local, scalable learning rule while maximizing neural plausibility by incorporating a negative inhibitory learning phase. We argue that this inhibitory phase, while seemingly unintuitive \({ }^{3}\), is actually a relevant

\footnotetext{
\({ }^{3}\) Some have found the notion of a chunk being self-inhibitory very unintuitive, because it conflicts with the idea that a chunk should be maximally similar to itself and self-activating.
}
and necessary mechanism to account for refractory periods in neural firings.

An advantage of this Hebbian-inspired implementation is that it avoids the inhibitory associations of low loglikelihoods, but the learning rule requires a form of softmax equation (either driven by expectation or more simple decay/inhibition) to keep \(S_{j i}\) values from overwhelming base-level \(B_{i}\) (i.e., from the likelihood overwhelming the prior, in Bayesian terms). At the micro/neural level, softmax approximates a maximum likelihood, while at a macro/ behavioral level, softmax simulates learning as expectation violation. In Bayesian terms, the more active (c.f., likely) the existing association between chunks \(A \rightarrow B\), then the less marginal increase in \(S_{j i}\) when chunk \(A\) is a source in the retrieval of chunk \(B\).

There are several beneficial effects from this kind of implementation. The first is that the mechanism is more balanced and geared towards specializing associative activations rather than just increasing all activations. Thus, the mechanism is more stable as it grows (i.e., it will not tend towards all associations becoming either strongly excitatory or inhibitory; \(S_{j i}\) doesn't vary with number of chunks in memory). Second, since the retrieved chunk selfinhibits, this reduces the chance that it will be the most active chunk in the following retrieval request (due to recency effects), which can cause models to get into selffeedback loops. In short, this inhibition leads to a natural refractory period for retrieving a chunk. Third, by selfinhibiting and spreading activation to the next context, it provides a forward momentum for the serial recall of chunks. Combined with recency and frequency of base level, this provides a mechanism for automatic serial recall of lists without the need for coding of explicit positional information (something required in prior models of list memory; Anderson et al., 1998) and marking of previously retrieved chunks through finst-like mechanisms. The uniqueness of the subsequent context drives order effects.

There are still, however, several design decisions and more empirical justification required in order to strengthen the constraint argument. Currently, the softmax learning term is based on ACT-R's base-level learning equation. However, several candidate equations need to be compared against human performance data to determine the best possible match. Furthermore, existing models of list memory and sequence learning need to be re-envisioned in terms of the new associative learning mechanism.

In summary, this balanced Hebbian/anti-Hebbian learning mechanism avoids the issues of scalability (e.g., runaway activations) that have been associated with prior implementations of associate learning in ACT-R. In addition, this mechanism is constrained by neural plausibility constraints, can still be discussed in Bayesiancompatible terms, and fits within the Bayesian description of ACT-R's sub-symbolic activation.

\section*{Discussion}

This paper has described how a functional cognitive architecture can constrain Bayesian inference by tying
neurally-consistent mechanisms into Bayesian-compatible sub-symbolic activations. This combination of grounded implementation- and algorithmic-level functions into cognitive-level Bayesian inference defuses many criticisms of Bayesian inference, and provides a launch-point for future research into constraining the Bayesian framework across all three levels of Marr's hypothesis. An example of this research was provided by examining a novel implementation for associative learning in ACT-R. In addition to the sub-symbolic layer being driven by Bayesian mathematics, it is also compatible with neural localization and the flow of information within the brain.

It has been argued that ACT-R's numerous parameters don't really provide the kind of constraint necessary to avoid the criticisms discussed in this paper (Tenenbaum et al., 2011). However, the use of community and researchjustified default values, the practice of removing parameters by developing more automatized mechanisms (such as the associative learning replacing spreading activation), and the development of common modeling paradigms mitigates these criticisms by limiting degrees of freedom in the architecture and thus constraining the kinds of models that can be developed and encouraging their integration. In summary, the evolution of the architecture is not a process of invalidation, but instead moving towards more constrained and more specific explanations.

As we have argued, the associative learning mechanism is an attempt to increase constraint within the architecture and promote a broader explanatory power to numerous cognitive phenomena. This mechanism is geared towards specializing associative strength to capture both symbolic and nonsymbolic associative learning. A major contribution of this mechanism is its balance between Hebbian (LTP) and antiHebbian (LTD) learning at each retrieval request, which provides numerous benefits over traditional Hebbian and Bayesian implementations.

\section*{Acknowledgments}

This work was conducted through collaboration in the Robotics Consortium sponsored by the U.S Army Research Laboratory under the Collaborative Technology Alliance Program, Cooperative Agreement W911NF-10-2-0016; and by Intelligence Advanced Research Projects Activity via DOI contract number D10PC20021. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI, or the U.S. Government.

\section*{References}

Anderson, J. R. (2007). How can the human mind occur in the physical universe? New York: Oxford University Press.
Anderson, J. R., \& Betz, J. (2001). A hybrid model of categorization. Psychonomic Bulletin \& Review, 8, 629-647.
Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., Qin, Y. (2004). An integrated theory of Mind. Psychological Review, 111, 1036-1060.
Anderson, J. R., Bothell, D., Lebiere, C. \& Matessa, M. (1998). An integrated theory of list memory. Journal of Memory and Language, 38, 341-380.

Anderson, J. R., and Lebiere, C. (1998). The atomic components of thought, Erlbaum, Mahwah, NJ.
Bowers, J. S., \& Davis, C. J. (2012). Bayesian Just-So Stories in Psychology and Neuroscience. Psychological Bulletin, 138 (3) 389414.

Caporale, N., \& Dan, Y. (2008). Spike Timing-Dependent Plasticity: A Hebbian Learning Rule. Annual Review of Neuroscience, 31, 25-46.
Chater, N., Oaksford, M., Hahn, U., \& Heit, E. (2010). Bayesian models of cognition. WIREs Cognitive Science, 1, 811-823.
Danks, D., Griffiths, T. L., \& Tenenbaum, J. B. (2003). Dynamical casual learning. In S. Becker, S. Thrun, \& K. Obermayer (Eds.), Advances in Neural information processing system. MIT Press: Cambridge, MA.
Dayan, P., Kakade, S., \& Montague, P. R. (2000). Learning and selective attention, Nature Neuroscience, 3, 1218-1223.
Deneve, S. (2008). Bayesian spiking neurons II: Learning. Neural Computation, 20, 118-145.
Fodor, J. A., \& Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: a critical analysis. Cognition, 28 (1), 3-71.
Griffiths, T. L., Chater, N., Norris, D., \& Pouget, A. (2012). How the Bayesians got Their Beliefs (and What Those Beliefs Actually Are): Comment on Bowers and DAvids (2012). Psychological Bulletin, 138 (3), 415-422.
Kemp, C., Perfors, A., \& Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. Developmental Science, 10 (3), 307-321.
Krueger, L. E. (1984). Perceived numerosity: A comparison of magnitude production, magnitude estimation, and discrimination judgments. Perception and Psychophysics, 35(6), 536-542.
Kruschke J. K. (2008). Bayesian approaches to associative learning: From passive to active learning. Learning and Behaviour, 36 (3), 210-226.
Lebiere, C. (1999). The dynamics of cognition: An ACT-R model of cognitive arithmetic. Kognitionswissenschaft., 8 (1), pp. 5-19.
Ma, W. J., Beck, J. M., Latham, P. E., \& Pouget, A. (2006). Bayesian inference with probabilistic population codes. Nature Neuroscience, 9, 1432-1438.
O'Reilly, R. (1998). Six principles for biologically based computational models of cortical cognition. Trends in Cognitive Science, 2 (11), 455-462.
O'Reilly, R., Hazy, T. E., \& Herd, S. A. (2012). The Leabra Cognitive Architecture: How to Play 20 Principles with Nature and Win! In S. Chipman (Ed) Oxford Handbook of Cognitive Science, Oxford: Oxford University Press.
Pearce J. M. and Hall G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. Psychological Review, 87, 532-552.
Rescorla, R. A. \& Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A.H. Black \& W.F. Prokasy (Eds.), Classical Conditioning II, Appleton-Century-Crofts.
Roberts, S., \& Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. Psychological Review, 107, 358-367.
Shafir, E. \& Tversky, A. (1992) Thinking through uncertainty: nonconsequential reasoning and choice. Cognitive Psychology 24: 449-474.
Sutton, R. S., \& Barto, A. G. (1998). Reinforcement Learning: An Introduction. MIT Press, Cambridge, MA.
Tenenbaum, J. B., Griffiths, T. L., \& Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. Trends in Cognitive Science, 10, 309-318.
Varma, S. (2011). Criteria for the Design and Evaluation of Cognitive Architectures. Cognitive Science, 35 (7), 1329-1351.
Xu, F., \& Tenenbaum, J. B. (2007). Word Learning as Bayesian Inference, Psychological Review, 114 (2), 245-272.
Yuille, A., \& Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? Trends in Cognitive Science, 10, 301-308.

\title{
A New Way of Linking Information Theory with Cognitive Science
}

\author{
Chris Thornton Informatics \\ University of Sussex Brighton \\ BN1 9QH \\ UK \\ c.thornton@sussex.ac.uk
}

\begin{abstract}
The relationship between the notion of information in information theory, and the notion of information processing in cognitive science, has long been controversial. But as the present paper shows, part of the disagreement arises from conflating different formulations of measurement. Clarifying distinctions reveals it is the contextfree nature of Shannon's information average that is particular problematic from the cognitive point of view. Context-sensitive evaluation is then shown to be a way of addressing the problems that arise.
\end{abstract}

\section*{Introduction}

One of the longest standing puzzles of cognitive science is what to think about information theory. Set out in its standard formulation more than 60 years ago, this framework (Shannon and Weaver, 1949) is acknowledged to be a remarkably general and precise area of mathematics. So it is of great interest to discover whether the notion of information developed in information theory has anything to do with the notion of information processing at the heart of cognitive science.
In the original publication, Shannon notes that 'the semantic aspects of communication are irrelevant to the engineering aspects' (Shannon and Weaver, 1949, p. 31). On the assumption that information processing in cognitive science deals with semantic aspects in particular, a fundamental disconnect seems implied. But this is muddled somewhat by the qualification (in Weaver's contribution to the joint publication) that Shannon's assertion 'does not mean that the engineering aspects are necessarily irrelevant to the semantic aspects' (Shannon and Weaver, 1949, p. 8). Adding to the ambiguity, researchers such as Meyer (1957/1967), Miller (1953), Garner (1962), Mackay (1956) and Attneave (1959) note a range of ways in which issues of a semantic nature can be addressed in information-theoretic terms. Quinlan (1993) and others demonstrate algorithms that operate specifically on this basis. Recent decades have also seen increasing use of information-theoretic quantification in cognitive neuroscience (e.g. Tononi et al., 1996; Lungarella et al., 2005; Friston, 2010).

The range of positions adopted on this issue deepens the mystery. Haber (1983) argues that informationtheoretic measures cannot address psychological questions due to being 'entirely independent of the recipient' (Haber, 1983, p. 71). Temperley (2007), on the
other hand, takes the view that the difficulty with them is they are calculated purely from the perspective of the recipient. Luce takes the view that information theory cannot address questions about structural representation of content because the 'elements of choice in information theory are absolutely neutral and lack any internal structure' (Luce, 2003, p. 185). On the other hand, a community of researchers examines ways in which information-theoretic quantification can explain emergence of structural representation in sensory processing (e.g. Attneave, 1959; Barlow, 1961; Uttley, 1979; Srinivisan et al., 1982; Atick, 1992).

For Haber, it is beyond dispute that 'the demise of information theory in psychology' has already occurred (Haber, 1983, p. 71). But intermediate positions are also common. Barwise notes that while 'traditional information theory is not a semantic theory at all' it 'puts important constraints on cognitive theories' (Barwise, 1983, p. 65). Churchland and Churchland (1983) are more positive still, seeing information theory as having a significant 'role to play in an account of cognition' (Churchland and Churchland, 1983, p. 67), and arguing the connection can be made through something called 'calibrational content' specifically, where this is defined to be informationally quantifiable 'measurement or detection concerning the status of the objective world' (Churchland and Churchland, 1983, p. 67). Others are doubtful of there being any connection at all. Dretske, for example, argues that information theory does not even 'deal with information as it is ordinarily understood' (Dretske, 1983, p. 56). \({ }^{1}\)

The present paper argues that one of the reasons the situation has become so confused is that the debate has conflated different formulations of measurement. \({ }^{2}\) The notion of measurement at the heart of the framework is the logarithmic principle, originally proposed by Hartley

\footnotetext{
\({ }^{1}\) Curiously, this view is part of an informational epistemology. However, Dretske's hard-line is consistent with the fact that his account has little in common with information theory (Sayre, 1983). In Kyburg's view 'Dretske seeks to clothe a relatively traditional approach to epistemology in new information-theoretic clothes' (Kyburg and Jr, 1983, p. 72).
\({ }^{2}\) I ignore areas of the framework (concerned with noisy and/or non-discrete systems) that have not figured in the debate.
}
(1928). There is also the probabilistic formulation of the logarithmic principle: \(-\log p\). Finally, there is the averaging formula
\[
-\sum_{x} p(x) \log p(x)
\]

This is called the entropy. These formulations build on each other. The averaging formula uses the probabilistic formulation, which is itself based on the logarithmic principle. \({ }^{3}\) But the three formulations have different implications for the question of connectivity with cognitive science.

The position often taken is that there is one form of information-theoretic quantification, and it is the averaging formula. Information measurement is taken to involve calculations of entropy specifically (e.g. Dretske, 1983; Sayre, 1983; Luce, 2003). This may be a consequence of the extent to which the results of (Shannon and Weaver, 1949) are derived by means of this formulation. But these results involve objectives of telecommunications specifically. Regarding the objectives of cognitive science, the logarithmic principle and the probabilistic formulation are equally of interest.

The present paper reviews the steps that lead from the logarithmic principle to Shannon's averaging formula. Account is taken of the semantic implications of different stages of the argument. Some aspects of the connectivity debate are clarified along the way, and consideration is given to the problems that arise from the use of context-free forms of measurement. Derivation of context-sensitive quantities is shown to be a viable alternative, and some examples are set out that show how this approach connects to the representational concerns of cognitive science.

\section*{Context-sensitive information}

Mathematical quantification of information begins with the logarithmic principle. Proposed originally by Hartley (1928), this has a number of foundations, as reviewed by Shannon (Shannon and Weaver, 1949, pp. 31-33). Where an outcome is within a known set, the informational value must relate to the number of outcomes in the set. A simple way of measuring the informational value of something that reveals a particular outcome is thus in terms of the number of possible outcomes that might have been revealed. This is a potential way to measure the informational value of a 'message' to a 'receiver' then, to use Shannon's own terminology. But as Hartley points out, it is much better to use a logarithmic function of the number of outcomes. This yields a measurement in which the quantity of information is also the number of digits needed to identify the outcome,

\footnotetext{
\({ }^{3}\) In practice, Shannon derives the entropy formula as the only acceptable way of measuring the 'choice' permitted by a distribution.
}
provided the same base is used for logarithm and digits. The usual approach uses base 2 . The quantify of information can then be stated in terms of 'bits' (short for BInary digiTS). The measure quantifies both the amount of information obtained, and the number of binary digits needed to encode the outcome.

On the logarithmic principle, then, the informational value of anything that reveals one of \(n\) outcomes is just \(\log n\). To obtain a value measured in bits, we take the logarithm to base 2. (Use of this base is assumed henceforth.) The process can be illustrated using any set of mutually exclusive outcomes. Let's say a new regulation requires Wi-Fi hotspots to be classified according to level of service, with the possible classifications being W1, W2, W3 and W4. Given there are four possible outcomes, the informational value of anything that gives the classification of a hotspot is then \(\log 4=2\) bits. This is also the number of base 2 (binary) digits required to identify a classification.

An advantageous property of the logarithmic principle is that it generalizes straightforwardly to the case where outcomes have different probabilities. Instead of defining the information obtained from a one-in- \(n\) outcome as \(\log n\) bits, it can be defined more generally as \(-\log p\) bits, where \(p\) is the probability of the outcome. This accommodates the simple case of equiprobable outcomes, since \(-\log \frac{1}{n}=\log n\). But it also accommodates there being a mixture of probabilities.
Let's say Wi-Fi hotspots are classified as W1 with probability \(\frac{1}{2}\), as W2 with probability \(\frac{1}{4}\), and as W3 and W4 with probability \(\frac{1}{8}\). The discovery that a hotspot has a W4 classification is more informative in the sense of being contrary to expectation, than observing it has a W1 classification. This is reflected in the information value obtained. The value of a W1 classification is just \(-\log \frac{1}{2}=1\) bit, whereas the informational value of a W4 classification is \(-\log \frac{1}{8}=3\) bits.

The probabilistic formulation of the logarithmic principle also provides the means of calculating averages. Given \(p(x)\) is the probability of outcome \(x\), the average informational value of an outcome is the weighted average
\[
\begin{equation*}
-\sum_{x} p(x) \log p(x) \tag{1}
\end{equation*}
\]

This is the entropy formula, centrepiece of Shannon's development of the logarithmic approach. It can be used whenever there is a probability distribution over outcomes. The distribution for Wi-Fi hotspots yields an average information value of 1.75 bits, for example.

The average information has a number of appealing properties. It can be seen as measuring the uncertainty that exists with respect to the outcomes, in the sense of quantifying the 'choice' allowed by the distribution (Shannon and Weaver, 1949, p 48-53). As a weighted
average, it can also be seen as defining the information that an outcome is expected to have. Given \(-\log p\) is an encoding cost, we can also look at the formula as the average cost of encoding an outcome. (Shannon proves the average cost can be no less: Shannon and Weaver, 1949, p. 62-64).

It is important to notice, however, that this approach makes no distinction between subjective and objective perspectives. In order for probability \(p(x)\) to be what fixes the amount of information an agent obtains from outcome \(x\), this must be the probability the agent attributes to \(x\). On this basis, \(p(x)\) is subjective. But where it is used in the averaging formula, \(p(x)\) becomes the objective probability of \(x\). In fact, Shannon's framework makes no distinction between subjective and objective probabilities. In the telecommunications context that is the framework's main focus, this makes sense. A telecommunications device adopting a personal perspective would be worthless. In other contexts, however, subjective factors may be of more relevance. It is of interest, then, to consider ways in which context-sensitive information values can also be calculated.

Consider the case where there is a set of two outcomes, both of which have information values calculated in an objective way (i.e., by the logarithmic principle). A context-sensitive value can then be calculated for any distribution attributed, and any outcome arising. This is just the expected value of the distribution in regard to the outcome. On the principle that probability attributed to the given outcome must increase the distribution's value, while probability attributed to any other outcome must decrease it, the expected information is a weighted average in which outcome values are either positive or negative:
\[
I\left(P_{S}\right)=\sum_{x \in S} P_{x} \begin{cases}I(x) & \text { if } x \text { is given }  \tag{2}\\ -I(x) & \text { otherwise }\end{cases}
\]

Here, \(S\) is the set of outcomes, \(P_{S}\) denotes the distribution attributed, and \(I(x)\) is the informational value of outcome \(x\). (Calculated by the logarithmic principle, \(I(x)=\log |S|)\).) This formula is valid whenever there are just two outcomes. Where there are more than two, the number of outcomes not given is greater than 1, and thus greater than the number given. It is then necessary to ensure commensurability between additions and subtractions by normalizing the latter with respect to \(|S|-1\), the number of non-given outcomes. The general form of the context-sensitive evaluation is thus
\[
I\left(P_{S}\right)=\sum_{x \in S} P_{x} \begin{cases}I(x) & \text { if } x \text { is given }  \tag{3}\\ -\frac{I(x)}{|S|-1} & \text { otherwise }\end{cases}
\]

This is the expected information value of distribution \(P_{S}\) to the attributing agent, where a particular element
of \(S\) is given, and all outcomes have known information values. It can also be seen as measuring the degree to which the distribution predicts the outcome in question.

Context-free evaluation of information (e.g., Eq. 1) is valid in most situations. Hence the generality of Shannon's framework. But where subjectivity is a possibility, context-sensitive evaluation (by Eq. 3) is entailed. The effects of evaluating information inappropriately can be illustrated using the hotspots example again. Let's say a particular agent expects every hotspot to be a W1. The agent attributes a probability of 1 to the W1 classification, and a probability of 0 to W2, W3 and W4. In objective reality, however, not all hotspots are W1: at least one is classified as W2. There is a subjective context, then, requiring amounts of information to be calculated in a context-sensitive way.

Should we choose to evaluation information in a context-free way regardless, the results are likely to be meaningless. The attributed distribution places all probability on one outcome. Its entropy is zero. On the strength of this, the average informational value of each outcome is deemed to be zero bits. This is appropriate in the case of a W1 classification, since the agent deems this to be the outcome in all cases. Unfortunately, it is also the value in the case of a W2 classification, which is a case the agent deems to be impossible.

This nonsensical result is a consequence of applying context-free evaluation to a situation in which there is a subjective context. On the context-free interpretation, there cannot be a W2 classification: its assumed probability is zero. Given the subjective perspective that is in force, context-sensitive evaluation using Eq. 3 is required. This produces a result that makes more sense. The context-sensitive value is found to be 1 bit for a W1 classification, and \(-\frac{2}{3}\) bits for any other classification. Notice the potential for negative context-sensitive values, in contrast with context-free (entropy) values, which are always non-negative.
The general difficulty that arises for cognitive science will then be evident. Situations of interest for this discipline involve subjectivity by definition. The tendency to equate information-theoretic evaluation with contextfree measurement is thus an obstacle. But there is another aspect to the problem. Both context-free and context-sensitive forms of evaluation are calculated with regard to a set of mutually-exclusive outcomes. The evaluations obtained depend solely on probabilities attributed, and the number of outcomes in the set. The difficulty is that each outcome has the potential to signify something completely different. Information values reflect the original outcomes, rather than any interpretations that may be forthcoming, however. Where an additional semantics is imposed on outcomes, both contextfree and context-sensitive values may be meaningless in relation to the interpretations that apply.

Evaluations that are context-sensitive in the sense of being calculated by Eq. 3 may thus fail to be contextsensitive with regard to an imposed semantics. There are thus two ways in which information-theoretic evaluations can be inadequate from the cognitive point of view. The semantic disconnect that commentators such as Luce (2003), Haber (1983) and Dretske (1983) see as inherent in information theory originates in these two ways.

\section*{Illustrations}

A useful context for illustrating context-sensitive evaluation is that of weather forecasting. Imagine we live in a world where the weather has just two outcomes: rain and sun. Let's say the forecast issued by the local met office for a particular day is showery, and that this signifies \(60 \%\) chance of rain, and \(40 \%\) chance of sun. Assume the outcome is rain. Eq. 3 can then be used to obtain a context-sensitive value for the attributed distribution given this particular outcome. With the forecast being showery, rain is predicted with probability 0.6 . The outcome is in fact rain, and the information value of each outcome is assumed to be \(\log 2=1\) bit. The context-sensitive value of the distribution is thus \((0.6 \times 1)-(0.4 \times 1)=0.2\) bits. If the outcome is sun, on the other hand, the value is \((0.4 \times 1)-(0.6 \times 1)=-0.2\) bits.


Figure 1: Context-sensitive evaluations.

The diagram of Figure 1 illustrates the two cases considered. In this and ensuing schematics, outcomes are represented by small circles, labeled with the outcome's name and informational value. Circles are filled where the outcome is given. Circles enclosed within the same bar are within the same choice of outcomes: the bar represents the choice. Where one outcome signifies a distribution over others - e.g., showery specifying rain with probability 0.6 - the relationships are indicated using connecting lines. Annotations placed over these lines show the probabilities that are attributed.

The figure shows evaluations of the showery distribution for the two outcomes rain and sun. Notice how the values reflect the degree to which the distribution predicts the outcome given. The evaluation is negative where the implied distribution mispredicts the outcome, and positive otherwise. At the same time, its relatively indiscriminate nature ensures both values are small compared with those of the outcomes themselves.


Figure 2: Derived evaluation of outcomes.

In the illustrated scenario, distributions are signified by entities (i.e., forecasts) that are themselves outcomes. By definition, these inherit the informational values of the distributions they designate. Any higher-level distribution must then be evaluated in terms of the derived values of predicted outcomes. To illustrate, let's say that in a certain season the forecast is always either showery or bright, with the latter meaning \(20 \%\) chance of rain and \(80 \%\) chance of sun. Context-sensitive values for these forecasts are then derived as in Figure 2. Potentially there can then be a second level of structure. A forecast of unsettled might mean \(70 \%\) chance of showery and \(30 \%\) chance of bright. The context-sensitive value of this forecast would then be calculated in terms of the derived values of showery and bright, rather than values obtained by the logarithmic principle.

\section*{Analysis of representation}

Context-sensitive evaluations can be calculated wherever we have both a distribution and a given outcome. Where one outcome signifies such a distribution itself, the value obtained also belongs to the signifying outcome, as noted above. Context-sensitive evaluations can thus be a way of evaluating probabilistic representation. Such evaluations can be made at multiple levels. Where one outcome signifies a distribution over several others, one of which does the same thing, there are two levels of representation. The latter is embedded within the former. Context-sensitive measurement of information is a way of evaluating outcomes at multiple levels of representation.

An assembly of signifying outcomes is a kind of representation structure, then. Such structures can take any form we like. For example, we might configure a representation structure in a way that expresses a category hierarchy. Let's say we have three categories as follows: a fruit category, in which the members are apple and plum; a bread category in which the members are pita and bun, and a food category in which the members are fruit and bread. To express this category hierarchy as a representation structure, categories must be treated as linked outcomes. Since category members are equiprobable instances of their category, member outcomes are always


Figure 3: Representation structure in the form of a category hierarchy.
equiprobable attributions of the corresponding category outcome. The representation structure expressing the category hierarchy is thus the one of Figure 3. (Notice this diagram tabulates the probabilities involved, rather than displaying them on an individual basis.)

Regardless of what a representation structure expresses, it retains its capacity for informational evaluation. This can be a way of explaining the functionalities that are forthcoming. Where a representation structure is arranged as a category hierarchy, for instance, there is the possibility of explaining classifications mathematically. Classifying an outcome in a particular way can be seen to identify the category outcome with the highest context-sensitive evaluation.

Consider the values that are obtained in the representation structure of Figure 3, where apple is given. These are shown in Figure 4. The context-sensitive value of the correct classification (fruit) is 0.67 bits, whereas the value of the incorrect classification (bread) is -0.67 bits. The classification can be explained as identifying the most informative category outcome.

Representation structures can be arranged in a broad range of ways and can thus express any model constructed in terms of representational relationships. Their probabilistic foundation means they can represent conventional Bayesian models, for example. Being able to incorporate multiple levels of representation, they can express hierarchical Bayesian models. Another possibility is schematic models, involving representational relationships of a conjunctive nature.

Consider a schematic model in which a particular entity is considered to be a combination of other enti-


Figure 4: Context-sensitive evaluation as classification.
ties. Viewed as a representation structure, this is a case in which one outcome designates multiple distributions, each of which concentrates probability on a single outcome. Such cases can be analyzed using context-sensitive evaluation in the usual way. But in so doing it is necessary to take the possibility of multiple designations into account. This must be done in accordance with the principle that information can be summed only if is independent. Where distributions are not independent, the evaluation obtained is the maximum (i.e., greatest independent value) rather than the sum of values arising.


Figure 5: Representation structure with disjunctive and conjunctive elements.

Figure 5 extends the bread/fruit scenario to illustrate what happens where representation structure includes conjunctive designations of this type. The bread/fruit structure from the previous diagram is seen here in the lower-left corner. At the same level of representation,
there is a cake/crumble choice. At the top level of representation, the outcome fruitcake is specified in a way that requires both fruit and cake. The effect is to reproduce the conjunctive character of a schema. As previously, the evaluations arising can explain classifications. If apple and cake are both given, the context-sensitive value of fruitcake is 1.67 bits. In the case of fudgecake, the value is 0 bits. Classifying a composite of apple and cake as fruitcake is then explained in terms of this category being most informative for the given context.

\section*{Conclusion}

The traditional objection to use of information theory in cognitive science has been the assumption that it does not deal with semantic aspects of information. On close examination, this is found to be an over-simplification. Where information values are calculated by means of the entropy formula, they are context-free in the sense of ignoring any element of subjectivity. They may also be context-free in the trivial sense of ignoring a superimposed semantic interpretation. The latter problem can be resolved simply by outlawing such applications. The former can be resolved by pursuing evaluation in a way that takes subjective context into account. On this basis, information-theoretic evaluation can be of relevance to cognitive science. Specifically, it can be a way of mathematically explaining category representation.

\section*{References}

Atick, J. J. (1992). Could information theory provide an ecological theory of sensory processing. Network, 3 (pp. 213-251).

Attneave, F. (1959). Applications of Information Theory to Psychology. New York: Henry Holt.

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. In Rosenblith (Ed.), Sensory communication (pp. 217-234). Cambridge: MIT Press.

Barwise, J. (1983). Information and semantics (commentary on Précis of Knowledge and the Flow of Information). Behavioral and Brain Sciences, 6 (pp. 65-66).

Churchland, P. M. and Churchland, P. S. (1983). Content: semantic and information-theoretic. Behavioral and Brain Sciences, 6 (pp. 67-68).

Dretske, F. I. (1983). Précis of Knowledge and the Flow of Information. Behavioral and Brain Sciences, 6 (pp. 55-90).

Friston, K. (2010). The free-energy principle: a unified brain theory? Nat. Rev. Neurosci, 11, No. 2 (pp. 127-138).

Garner, W. R. (1962). Uncertainty and Structure as Psychological Concepts. New York.

Haber, R. N. (1983). Can information be objectivized? Behavioral and Brain Sciences, 6 (pp. 70-71).

Hartley, R. L. (1928). Transmission of information. Bell System Technical Journal (pp. 535).

Kyburg, H. E. and Jr, (1983). Knowledge and the absolute. Behavioral and Brain Sciences, 6 (pp. 72-73).

Luce, R. D. (2003). Whatever happened to information theory in psychology. Review of General Psychology, 7, No. 2 (pp. 183-188).

Lungarella, M., Pegors, T., Bulwinkle, D. and Sporns, O. (2005). Methods for quantifying the information structure of sensory and motor data. Neuroinformatics, 3, No. 3 (pp. 243-262).

Mackay, D. (1956). Towards an information-flow model of human behaviour. Br. J. Psychol, 43 (pp. 30-43).

Meyer, L. B. (1957/1967). Meaning in music and information theory. Music, the Arts, and Ideas (pp. 5-21). Chicago: University of Chicago Press.

Miller, G. A. (1953). What is information measurement? American Psychologist, 8 (pp. 3-11).

Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. San Mateo, California: Morgan Kaufmann.

Sayre, K. M. (1983). Some untoward consequences of Dretske's "causal theory" of information. Behavioral and Brain Sciences, 6 (pp. 78-79).

Shannon, C. and Weaver, W. (1949). The Mathematical Theory of Communication. Urbana, Illinois: University of Illinois Press.

Srinivisan, M. V., laughlin, S. B. and A, A. D. (1982). Predictive coding: a fresh view of inhibition in the retina. Proc. R. Soc. Lond. B, 216 (pp. 427-59).

Temperley, D. (2007). Music and Probability. Cambridge, Massachusetts: The MIT Press.

Tononi, G., Sporns, O. and Edelman, G. (1996). A complexity measure for selective matching of signals by the brain. Proceedings of the Nat. Academy of Science, 93 (pp. 3422-3427).

Uttley, A. M. (1979). Information Transmission in the Nervous System. London: Academic.

\title{
Verifying properties from different emotions produces switching costs: Evidence for coarse-grained language statistics and fine-grained perceptual simulation
}

\author{
Richard Tillman (rntllman@memphis.edu) \\ Department of Psychology/ Institute for Intelligent Systems, University of Memphis \\ 365 Innovation Drive, Memphis, TN 38152 USA \\ Sterling Hutchinson (schtchns@memphis.edu) \\ Department of Psychology/ Institute for Intelligent Systems, University of Memphis \\ 365 Innovation Drive, Memphis, TN 38152 USA \\ Sara Jordan (sara.jordan@mlh.org) \\ Department of Rehabilitation Services, Speech Language Pathologist/ Methodist North Hospital, Methodist LeBonheur Healthcare \\ 3960 New Covington Pike Memphis, TN 38128 USA
}

Max M. Louwerse (maxlouwerse@gmail.com)
Department of Psychology/ Institute for Intelligent Systems, University of Memphis
365 Innovation Drive, Memphis, TN 38152 USA
Tilburg Centre for Cognition and Communication (TiCC), Tilburg University, PO Box 90153, 5000 LE, Tilburg, The Netherlands

\begin{abstract}
We investigated whether emotions are activated during comprehension of emotion words. In the first part of the study, an experiment was conducted in which participants read sentence pairs each describing an emotional state and then engaged in a judgment task. Sentences were paired to either match or mismatch in emotion (happy, sad, or angry). We predicted that the sentences that mismatch in emotion produced longer reaction times than those where the emotion was the same, and that shifts between negative emotions had less of an impact. In the second part of the study, we calculated the frequency of first-order co-occurrences of nouns and adjectives related to happy, sad, and angry emotional states. This analysis demonstrated emotion words are more often accompanied by similar emotion words. Match and mismatch of emotion explained RTs as did statistical linguistic frequencies of the words. The combination of these two studies contributes to a growing body of research that supports the importance of both symbolic and perceptual processing of emotion.
\end{abstract}

Keywords: emotion; embodied cognition; symbolic cognition; statistical linguistic frequencies.

\section*{Introduction}

Theories of embodied cognition claim that cognition is fundamentally based in perceptual experiences. That is, concepts only become meaningful through comprehenders mentally reenacting prior physical and perceptual experiences with the concept in the real world (Barsalou, 1999; Barsalou, Simmons, Barbey, \& Wilson, 2003; Glenberg \& Kaschak, 2002; Pecher \& Zwaan, 2005; Havas, Glenberg, \& Rinck, 2007; Semin \& Smith, 2008). For instance, Glenberg and Kaschak (2002) proposed the action-
sentence compatibility effect whereby language processing is facilitated when a congruent response motion is used to respond to sentences describing motion away from or towards the body. That is, sentences describing motion away from the body (e.g., close a drawer) were processed faster when response motions were also moving away from the body, and vice versa. These results and findings similar to these demonstrate that linguistic processing is facilitated through perceptual-motor information (see Leventhal, 1982 for an overview).

Similar to action related sentences, sentences with emotional content have also provided support for an embodied cognition account. Mouilso, Glenberg, Havas, and Lindeman (2007) found that reading 'angry' sentences resulted in faster movements away from the body and reading 'sad' sentences resulted in faster movements toward the body. In other words, when people read angry content, they processed the sentence faster with an aggressive action toward it, whereas 'sad' sentences evoke a withdrawal action, suggesting that emotional language can affect bodily responses.

Embodied responses have also been linked to cognition through the facial feedback hypothesis (Strack, Martin, \& Stepper, 1988; Zajonc, Murphy, \& Inglehart, 1989). The facial feedback hypothesis demonstrates that facial expressions might influence emotional assessments. For example, when participants were instructed to smile, cartoons were perceived as more humorous than when subjects were not smiling (Strack et al., 1988), showing that bodily states can affect both judgments and cognition.

Most literature supporting an embodied cognition account, however, demonstrates evidence without physical manipulation. For example, Pecher, Zeelenberg, \& Barsalou
(2003) found that subjects read sentences describing features within the same modality faster than sentences describing features of differing modalities. When participants read a sentence like apples can be tart followed by the sentence apples can be sweet (describing the same gustatory modality) response times were faster for the second sentence when the second sentence did not describe a shift in modality, such as is the case when a visual modality was presented in strawberries can be red or radios can be loud. The modality of the target words impacted how those words were perceived. Processing costs incurred from the mismatched sentences resulted from a perceptual modality shift, suggesting that perceptual embodied features indeed impact language processing times.

Recently, the modality switching costs have been explained by language statistics (Louwerse \& Connell, 2011). By computing the word frequencies of the cooccurrences of modality words from a large corpus of English, Louwerse and Connell were able to identify modality shifts similar to Pecher et al. (2003). This analysis was not only applicable to the adjectives (e.g., tart - sweet being more frequent than tart - red or sweet - red), but also to concept words (e.g., apples - strawberries being more frequent than apples - radio or strawberries - radio). Louwerse and Connell (2011) showed that these frequencies explained the response times that were attributed to an embodied cognition account. That is, faster response times were best explained by language statistics, slower response times were best explained by perceptual simulations. Louwerse and Connell's explanation was that the linguistic system offers a 'quick and dirty' shallow heuristic that can provide good enough performance in cognitive tasks without recourse to deeper conceptual processing in a perceptual simulation system. On the other hand, ultimately concepts are grounded and can be perceptually simulated. The explanation by Louwerse and Connell can be captured in the Symbol Interdependency Hypothesis, which proposed that conceptual processing can be explained by both symbol and embodied mechanisms (Louwerse, 2007; 2008; 2011). When we encounter a word, we garner a rough meaning from its linguistic (symbolic) neighbors using language statistics, but to fully ground the word, we perceptually simulate its physical and somatosensory features. Thus, words can rely on other words to establish a fuzzy sense of meaning without necessarily always being grounded themselves. In other words, perceptual information is encoded in language, such that mental representations are both perceptual and linguistic. Human beings can rely on such a linguistic short-cut when processing language in real time. However, if a deeper meaning or understanding is needed, grounding the world in perceptual experiences provides rich sensorimotor information about meaning. Importantly, language has encoded sensorimotor information, such that language users can utilize these cues in cognitive processes.

In short, the Symbol Interdependency Hypothesis proposes the following: 1) language encodes perceptual
information; 2) language users rely both on language statistics and perceptual simulation in cognitive processes; 3) the relative dominance of language statistics and perceptual simulation factors is modified by stimulus type and task.

Although modality shifts have been shown to support the Symbol Interdependency Hypothesis (Louwerse \& Connell, 2011), the question can be raised whether the finding for modality shifts can be extended to other semantic domains that have shown embodiment effects, such as emotions. In the current study we investigated whether (a) verifying properties from different emotions for concepts produces switching costs, similar to the modality shifts; (b) whether language has encoded the emotions of words, similar to the modality of words; (c) whether emotion shifts can be explained by a language statistics account.

To explore these questions we applied Pecher et al.'s (2003) modality shift paradigm to emotions. Emotional sentences shifted from happy to sad, sad to happy, happy to angry, angry to happy, sad to angry, and angry to sad. According to an embodied cognition account, switches between emotions should take longer to process than nonswitches (happy-happy, sad-sad, angry-angry). Alternatively, according to a language statistics account, cooccurrence frequencies of word pairs should be able to equally account for subject RTs. We thereby made two hypotheses: (1) as with modality shifts, emotion shifts would take longer to process than non-shift sentence pairs, which would be in support of an embodied cognition account and (2) the same pattern of emotion shift cost would emerge from language such that emotion words that matched in valence would co-occur more frequently than the words that did not match in valence, which would be in support of a linguistic account.

\section*{Experiment 1: Emotion Shift}

\section*{Method}

Participants Thirty-three undergraduate students enrolled in an introductory psychology course participated for course credit.

Materials Sixty emotion sentences were created, following the method described in Pecher et al. (2003) with each sentence in the format \(X\) can be \(Y\). There were 3 experimental types of emotions depicted in the sentences: angry, happy, and sad. For example, birthdays can be happy (happy emotion), and insults can be devastating (sad emotion).

The reason we selected angry, happy, and sad emotions was motivated by work from Isenhower et al. (2003) who found that people tend towards more positive states of emotion. That is, switching from positively valenced to negatively valenced emotions yields a greater disruption and requires additional cognitive processing. Further motivation came from a more recent study by Stein and Sterzer (2012). In this study, Stein and Sterzer demonstrated that people identify happy faces more quickly than angry faces. We
therefore selected happy, sad, and angry words, and thus had one positively valenced emotion (happy) and two negatively valenced emotions (sad and anger).

Procedure Participants were seated at a computer in a standard computer lab. The instructions for the experiment were presented on the screen and read aloud by the experimenter. Five practice items preceded the experimental phase to ensure participants understood the task. Participants saw sentences one at a time in the center of the screen and then were asked to respond to the question Is the characteristic true of the items it described? Participants pressed designated yes or no keys on the keyboard. RT and accuracy were recorded.

\section*{Results}

Incorrect responses were not included in the analyses. RT outliers were defined as 2.5 SD above the mean per subject per condition and were removed from the analysis. This removal affected less than \(3.6 \%\) of the data.

A mixed-effect analysis was conducted on RTs with emotion shift as the fixed factor and participants and items as random factors (Baayen, Davidson, \& Bates, 2008). The model was fitted using the restricted maximum likelihood estimation (REML) for the continuous variable (RT). F-test denominator degrees of freedom were estimated using the Kenward-Roger's degrees of freedom adjustment to reduce the chances of Type I error (Littell, Stroup, \& Freund, 2002). Participants and items were treated as random factors in the analysis.

For the factor emotion shift no differences were found in RT, \(F(1,114)=.431, p=.513\). This is somewhat surprising given that an emotion shift was predicted to increase RTs akin to the modality shifts. However, when individual emotion pairs were separated by transition (e.g., happy-sad, happy-angry), RT differences were obtained with an emotion shift from happy sentences to sad sentences, \(F(1,421)=30.41, p<.001\), with slower RTs for the shift than no-shift (i.e., a happy sentence followed by a happy sentence). Also, when shifting from happy sentences to angry sentences a significant difference was found between the two conditions, \(F(1,380)=20.82, p<.001\), there were slower RTs for the emotion shift sentences than no-shift. When the sad to angry sentences were compared, again a difference approaching significance was found between emotion shift and no-shift conditions, \(F(1,455)=5.88, p<\) .056, where the shift between sentences yielded longer RTs than no-shift. In contrast, the comparison of sad to happy sentences yielded no significant differences between emotion shift and no-shift sentences, \(F(1,395)=.02, p<\) .89. When switching from angry to happy sentences, a significant effect was found, \(F(1,485)=20.69, p<.001\), again with faster RTs for the emotion shift sentences than no-shift. Finally, a significant effect was found when switching from angry to sad sentences, \(F(1,430)=5.05, p<\) .03 , however with faster RTs for the emotion shift sentences than the no-shift sentences.

In summary, a shift from happy to sad, happy to angry, sad to angry, and angry to happy yielded significant results, while the shift from sad to happy was not significant. Figure 1 shows the means and standard deviations for each emotion shift pair.

Even though no overall effect for emotion shift was found, patterns for specific emotion transitions did show shift effects, with specific emotion to emotion shifts resulting in longer RTs than non-shifts. More specifically, the shifts from happy to the two negative emotions, shows a significant increase in RT. The emotion shift from angry to happy was also significant, but showed a decrease in RT from angry followed by angry. This is in line with Stein and Sterzer (2012), who found that people are quicker to identify happy faces, rather than angry faces. We interpret this decrease in RT in terms of the nature of the shift. Angry followed by angry produces the longest RT, while happy followed by happy produces the shortest RT. As there is a tendency to prefer to shift toward a more positive state (Isenhower, Frank, Kay, \& Carello, 2010), the reaction times for the non-shifts reflected this. Moreover, the shift from angry to happy decreases from its origin (angry followed by angry), because of the natural tendency to shift to the more positive state. This is supported by the significant differences when emotion shifts took place between angry and happy, angry and sad, and sad and angry.

However, we still are unable to determine whether an embodiment effect exists for emotion switching, as there was no overall effect for shifts as there were for Pecher et al. (2003), but only specific emotion to emotion effects.


Figure 1. Emotion shifts, means, and standard deviations. ** \(p<.01, * p<.05, n . s\). not significant. The means and standard deviations located at the emotion words indicate no emotion shift (e.g., a happy sentence followed by a happy sentence).

To determine whether or not overall shifts for emotions occurred, we ran a second experiment whereby the
embodiment effect would be enhanced by an embodied facial feedback paradigm.

\section*{Experiment 2: Facial Feedback Hypothesis}

In order to determine if emotion switching indeed supports an embodied cognition account, we examined the effects of the facial feedback hypothesis (Strack, et al., 1988; Zajonc, et al., 1989) by assessing the effects different conditions (frowning or smiling) had on RTs when judging emotion shift sentences. We hypothesized that neither frowning nor smiling would produce a significant effect between the negative emotions (sadness and anger), due to the trend towards positive (Isenhower et al., 2010). In addition, we hypothesized that the specific emotion to emotion shifts found in Experiment 1 would show similar patterns.

\section*{Method}

Participants Twenty-six undergraduate students enrolled in an introductory psychology course participated for course credit.

Materials The same materials were used as in Experiment 1.

Procedure The procedure was the same as that used in Experiment 1, with one important addition. Participants were also randomly assigned to one of two facial feedback conditions (Strack et al., 1988). In the one condition, the participants held a pen in their lips \((n=15)\) to simulate frowning; in the other, the participants held a pen in their teeth ( \(n=11\) ) to simulate smiling.

\section*{Results}

As in Experiment 1, emotion shifts did not yield a significant difference in RT, \(F(1,117.27)=.16, p=.70\). Furthermore, there seemed to be no main effect of the facial feedback conditions, \(F(2,78.24)=.73, p=.49\). Next, we investigated the emotion transitions per facial feedback condition (smiling vs. frowning).

Frowning Facial Feedback When participants held the pen in their lips to simulate frowning, the shift from happy to sad was significant as it was in the previous experiment without the facial feedback task, \(F(1,236)=6.69, p=.01\), with higher RTs for the shift sentences than no-shift. The shift from happy to angry was also significant as found in the previous experiment, \(F(1,202)=8.36, p<.004\), with higher RTs for the shift sentences than no-shift. Also the shift from angry to happy was significant as previously found in Experiment 1, \(F(1,248)\) 4.31, \(p<.04\), with lower RT for the shift sentences than no-shift. Again, this is in line with Isenhower et al. (2010) and Stein and Sterzer (2012), in that the preference is to shift from a negative state to a positive state. This is especially true given the fact that participants were frowning due to the facial feedback task. The shifts from sad to angry and angry to sad were found to be non-significant, unlike the findings in Experiment 1. These results lend support to the facial feedback hypothesis, in that frowning (pen held in lips) is associated with both
sadness and anger; it would stand to reason why there were no significant differences between these two conditions as they are both negative emotions and the motor system necessary for their simulation was already active, facilitating the effect. Figure 2 shows the means and standard deviations for each emotion shift pair, the shift direction, and the no shift means and standard deviations.


Figure 2. Emotion shifts, means, and standard deviations for frowning facial feedback condition. \({ }^{* *} p<.01\), * \(p<\) .05, n.s. not significant.

Smiling Facial Feedback When participants held the pen in their teeth to simulate smiling, the shift from happy to sad was significant, \(F(1,168)=8.98, p<.003\), with higher RT for the shift sentences than no-shift . The shift from happy to angry was significant, \(F(1,164)=15.48, p<.0001\), with higher RTs for the shift sentences than no-shift. The shift from sad to happy approached significance, \(F(1,134)=3.81\), \(p<.053\), with lower RT for the shift sentences than no-shift. Finally, the shift from angry to happy was also significant, \(F(1,179)=17.84, \mathrm{p}<.001\), with lower RT for the shift sentences than no-shift. Again, the decrease in RT for angry to happy is in accord with Stein and Sterzer (2012). The shifts from sad to angry and angry to sad were not found to be significant. Figure 3 shows the means and standard deviations for each emotion shift pair, the shift direction, and the no shift means and standard deviations. The main difference between the smiling condition and the previous frowning condition is the significant difference found in the sad to happy shift, which was not found in Experiment 1, or the frowning facial feedback condition. This difference supports the findings by Isenhower et al. (2010), in that since people have a tendency to tend towards a positive state, which they have in part done by smiling.


Figure 3. Emotion shifts, means, and standard deviations for smiling facial feedback condition. \({ }^{* *} p<.01, * p<.05\), n.s. not significant

\section*{Corpus Linguistic Study}

So far, the results seem to suggest that emotional states can be based in embodied cognition, as some emotion to emotion shifts seem to indicate that emotion switching usually incurs some sort of processing cost. However, this is not the whole picture, as it does not take into consideration the linguistic nature of the words. We therefore investigated whether emotion shifts are encoded in language (Louwerse, 2008; Louwerse \& Connell, 2011). To do this we calculated the frequency of first-order co-occurrences of all the possible combinations of the nouns and adjectives in the present study by utilizing the Web 1T 5-gram corpus (Brants \& Franz, 2006). This corpus consists of 1 trillion word tokens (13,588,391 word types) from 95,119,665,584 sentences. The volume of the corpus allows for an extensive analysis of patterns in the English language. The frequency of co-occurrences of the word pairs was computed for bigrams, trigrams, 4 -grams and 5 -grams. For instance, the frequency of the phrase birthdays can be happy \{happy, birthday \(\}\) was determined by considering these words next to one another \{happy birthday\}, with one word in between \{happy w1 birthday\}, with two \{happy w1 w2 birthday\}, three intervening words \{happy w1 w2 w3 birthday\}, and so on.

A mixed effects analysis was conducted on the frequency of co-occurrences of the emotion adjectives and the noun referents. The independent variable was whether the emotion words were the same or different emotion, and the log frequency of the word pair was the dependent variable.

For all possible combinations of both nouns and adjectives, the log frequency of the co-occurrences were found to be significant, \(F(1,7078)=212.76, p<.001\), with word pairs where there was no emotion shift ( \(M=2.08, S E\) \(=.04)\) being more frequent than word pairs where an
emotion shift was present ( \(M=1.11, S E=.054\) ). This pattern was also found for just the nouns \(F(1,3479)=\) 148.11, \(p<.001\), with word pairs where there was no emotion shift ( \(M=4.29, S E=.08\) ) being more frequent than word pairs where an emotion shift was present ( \(M=2.60\), \(S E=.11\) ). Again, this pattern was found for adjectives, \(F(1\), 3598) \(=\) 279.17, \(p<.001\), with word pairs where there was no emotion shift ( \(M=2.53, S E=.05\) ) being more frequent than word pairs where an emotion shift was present ( \(M=\) \(1.00, S E=.07\) ).

In addition, we also compared the log frequencies of each of the word pairs to the experimental RT over the collapsed match and mismatch conditions (extracted from Experiments 1 and 2). Language statistics significantly predicted RTs, \(F(1,113.564)=34.53, p<.001\). However, language statistics did not predict emotional transitions. Statistical linguistic frequencies explained RTs of general emotion shifts, but not RTs of specific emotion transitions.

\section*{General Discussion}

Previous studies have found that two sentences that elicit a modality shift produce cognitive switching costs, compared to sentences that describe the same modality (Pecher et al., 2003). This finding has been reported as evidence for an embodied cognition account, because the increased RTs are an indication that comprehenders perceptually simulate the sentences. Others have shown that modality is encoded in language. Based on language statistics, concepts and their features can be categorized in visual, auditory, olfactory and gustatory modalities (Louwerse \& Connell, 2011). Moreover, when the RTs for modality shifts were investigated with both language statistics and perceptual simulation as independent variables, fast RTs were best explained by language statistics and slower RTs were best explained by perceptual simulation. Louwerse and Connell (2011) concluded that language statistics serves as a coarse-grained system that serves as a shallow heuristic. Perceptual simulation, on the other hand, serves deeper conceptual processing. The idea that language encodes perceptual information and that these linguistic cues can be used by language users in shallow comprehension tasks is predicted by the Symbol Interdependency Hypothesis and supported by various studies (Louwerse, 2008; Louwerse \& Hutchinson, 2012; Louwerse \& Jeuniaux, 2008; 2010).

The current study investigated whether emotion shifts mimicked the patterns found for previous studies investigating modality shifts. Even though across three experiments no general effect was found for shifts, specific transitions between emotions did yield differences in RTs. Moreover, evidence was found that language encodes emotion shifts, and language statistics explained RTs for these general shifts.

The findings of the current study are supported by the Symbol Interdependency Hypothesis as well as by findings reported in other studies. Language statistics explained coarse-grained emotion shifts. However, language statistics
did not explain fine-grained shifts. On the other hand, assuming that a perceptual simulation system is responsible for the other RT differences that were obtained in the two experiments, the perceptual system did not explain the coarse-grained differences in general emotion shifts, but did explain the fine-grained shifts between specific emotions.

These results provide further evidence for the theory that conceptual processing is both linguistic and embodied, whereby less precise linguistic processes account for general patterns in processing, whereas perceptual simulation provides the fine-tuning.

\section*{References}

Baayen, R. H., Davidson, D., \& Bates, D. (2008). Mixedeffects modeling with crossed random effects for subjects and items. Journal of Memory and Language, 59, 390412.

Barsalou, L. W. (1999). Perceptual symbol systems, Behavioral and Brain Sciences, 22, 577-660.
Barsalou, L. W., Simmons, W.K., Barbey, A.K., \& Wilson, C.D. (2003). Grounding conceptual knowledge in modality-specific systems, Trends in Cognitive Sciences, 7, 84-91.
Brants, T., \& Franz, A. (2006). Web 1T 5-gram version 1. Philadelphia: Linguistic Data Consortium.
Brysbaert, M. (2007). The language-as-fixed-effect fallacy: Some simple SPSS solutions to a complex problem (version 2.0). Royal Holloway, University of London. Technical report.
Ekman, P., \& Rosenberg, E. L. (2004). What the face reveals: basic and applied studies of spontaneous expression using the facial action coding system (FACS), New York: Oxford University Press.
Glenberg, A. M. (2007). Language and action: creating sensible combinations of ideas. In G. Gaskell (Ed.) The Oxford handbook of psycholinguistics (pp. 361-370). Oxford, UK: Oxford University Press.
Glenberg, A. M., \& Kaschak, M. P. (2002). Grounding language in action. Psychonomic Bulletin \& Review, 9, 558-565.
Havas, D., Glenberg, A. M., \& Rinck, M (2007). Emotion simulation during language comprehension. Psychonomic Bulletin \& Review, 14, 436-441.
Isenhower, R. W., Frank, T. D., Kay, B. A., \& Carello, C. (2010). Capturing and quantifying the dynamics of valenced emotions. Nonlinear Dynamics Psychology and Life Sciences, 16(4), 397-427.
Leventhal, H. (1982). A perceptual motor theory of emotion. Social Science Information, 21, 819-845.
Littell, R. C., Stroup, W. W., \& Freund, R. J. (2002). SAS for Linear Models, \(4^{\text {th }}\) Edition. Cary, NC: SAS Publishing.
Louwerse, M. M. (2007). Symbolic embodied representations: A case for symbol interdependency. In T. Landauer, D. McNamara, S. Dennis, \& W. Kintsch (Eds.), Handbook of latent semantic analysis. Mahwan, NJ: Erlbaum.

Louwerse, M. M. (2008). Embodied relations are encoded in language. Psychonomic Bulletin \& Review, 15, 838-844.
Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. TopiCS, 3, 273-302.
Louwerse, M. M., and Connell, L. (2011). A taste of words: linguistic context and perceptual simulation predict the modality of words. Cognitive Science, 35, 381-398.
Louwerse, M. M., \& Jeuniaux, P. (2010). The Linguistic and Embodied Nature of Conceptual Processing. Cognition, 114, 96-104.
Louwerse, M. M. \& Jeuniaux, P. (2008). Language comprehension is both embodied and symbolic. In M. de Vega, A. Glenberg, \& A. C. Graesser (Eds.), Symbols and Embodiment: Debates on Meaning and Cognition. Oxford, England: Oxford University Press.
Mouilso, E., Glenberg, A. M., Havas, D. A., \& Lindeman, L. M. (2007). Differences in action tendencies distinguish anger and sadness after comprehension of emotional sentences. Proceedings of the 29th Annual Cognitive Science Society, 14, 1325-1330.
Pecher, D., \& Zwaan, R. (Eds.) (2005). Grounding cognition: The role of perception and action in memory, language, and thought. New York: Cambridge University Press.
Pecher, D., Zeelenberg, R., \& Barsalou, L.W. (2003). Verifying different-modality properties for concepts produces switching costs. Psychological Science, 14, 119124.

Semin, G. \& Smith, E. (Eds.) (2008). Embodied grounding: Social, cognitive, affective, and neuroscientific approaches. New York: Cambridge University Press.
Stein, T., \& Sterzer, P. (2012). Not just another face in the crowd: Detecting emotional schematic faces during continuous flash suppression. Emotion, 12, 988-996.
Strack, F., Martin, L. \& Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. Journal of Personality and Social Psychology, 54, 768-777.
Zajonc, R. B., Murphy, S. T., \& Inglehart, M. (1989). Feeling and facial efference: Implications for the vascular theory of emotion. Psychological Review, 96(3), 395-416.

\title{
Geographical Estimates are Explained by Perceptual Simulation and Language Statistics
}

\author{
Richard Tillman (rntllman@memphis.edu) \\ Department of Psychology/ Institute for Intelligent Systems, University of Memphis \\ 365 Innovation Drive, Memphis, TN 38152 USA \\ Sterling Hutchinson (schtchns@memphis.edu) \\ Department of Psychology/ Institute for Intelligent Systems, University of Memphis \\ 365 Innovation Drive, Memphis, TN 38152 USA \\ Max M. Louwerse (maxlouwerse@gmail.com) \\ Department of Psychology/ Institute for Intelligent Systems, University of Memphis 365 Innovation Drive, Memphis, TN 38152 USA \\ Tilburg Centre for Cognition and Communication (TiCC), Tilburg University, PO Box 90153, 5000 LE, Tilburg, The Netherlands
}

\begin{abstract}
Several studies have demonstrated that language encodes geographical information. That is, the relative longitude and latitude of city locations can be extracted from language. Whether people actually rely on these linguistic features is less clear. Recent studies have suggested that language statistics plays a role in geographical estimates, but these studies rely on map drawings, a fundamentally perceptual task. The current study investigated the extent to which people rely on map representations and statistical linguistic frequencies by using a linguistic task. Participants saw U.S. city pairs in their iconic positions (a more northern city is presented above a more southern city, or a more western city is presented to the left of a more eastern city), and in their reverse-iconic positions (a more southern city is presented above a more northern city, or a more eastern city is presented to the left of a more western city). For iconic city pairs both in the east - west (Seattle - Boston) and north - south (Memphis - Miami) configurations, RTs were determined by the iconicity. No effect was obtained for statistical linguistic frequencies. However, when city pairs were presented in a reverse-iconic configuration, for both horizontal (Boston Seattle) and vertical (Miami - Memphis) orientations, both perceptual and linguistic factors explained RTs. These findings support the idea that cognition relies on a shallow heuristic, a linguistic system, and a fine-grained and more precise perceptual simulation system.
\end{abstract}

Keywords: embodied cognition; symbolic cognition; geography; spatial cognition

\section*{Introduction}

Is San Francisco close to New York? Is Boston close to Miami? Judging the distance between cities can be approached in more than one way. This judgment can be deep and precise, as with perceptual simulation, or quick and shallow, as with symbolic representation. For instance, humans can make geographical estimates on the basis of their perceptual experiences from locomotion and stationary
viewing, from static pictorial representations, such as diagrams, paintings and photos, provided on a map, and they can acquire information via dynamic pictorial representations, including animations, and videos (Freundschuh \& Mercer, 1995).
The importance of a perceptual simulation system has been strongly advocated by accounts of embodied cognition (Barsalou, 1999; Barsalou, 2008; Glenberg \& Kaschak, 2002; Pecher \& Zwaan, 2005; Semin \& Smith, 2008). According to Barsalou, Solomon, and Wu (1999), perceptual states are transferred into memory and function symbolically, rather than an arbitrary representation such as language. As an example, overwhelming evidence in favor of an embodied cognition account has accumulated, showing that processing within modalities is faster than having to map across modalities, and suggesting that modality switching comes at a price (e.g., Marques, 2006; Pecher, Zeelenberg, \& Barsalou, 2003; Spence, Nicholls, \& Driver, 2001). Furthermore, language comprehension seems to be influenced by action representations primed in experimental tasks (e.g., Glenberg \& Kaschak, 2002; Kaschak et al., 2005; Klatzky, Pellegrino, McCloskey, \& Doherty, 1989; Zwaan, Stanfield, \& Yaxley, 2002), and visual representations get activated during language comprehension (see also Boroditsky, 2000; Fincher-Kiefer, 2001; Matlock, Ramscar, \& Boroditsky, 2005).

One particular study nicely illustrates the embodied cognition account. Zwaan and Yaxley (2003) presented iconic word pairs either as they occur in the real world, such as attic over basement, or the reverse-iconic orientation, such as basement over attic. They found significant differences between the iconic and reverse-iconic configurations of these word pairs. They concluded that the explanation for the iconicity effect was that words activate their perceptual representations (attics presented above basements are processed faster than basements above attics, because of their iconic relationship in the real world).

Louwerse (2008) questioned whether the Zwaan and Yaxley (2003) finding should be solely attributed to perceptual simulation. Statistical linguistic frequencies, the co-occurrence of words in a given frame, showed that items that are normally high in space preceded items that are normally low in space more frequently than vice versa, suggesting that language encodes spatial information (e.g., we say up and down, top and bottom, knees and toes, rather than down and up, bottom and top and toes and knees). Moreover, statistical linguistic frequencies explained RTs better than the perceptual factor. These findings demonstrate that there is a complementary linguistic explanation to a perceptual simulation explanation.

Louwerse and Jeuniaux (2010) showed that the extent to which cognitive processes can be explained by perceptual simulation or language statistics (frequency of word cooccurrence) depends on a variety of factors, including the nature of the stimulus (e.g., words versus pictures) and the cognitive task (e.g., shallow or deep cognitive task). In Louwerse and Jeuniaux (2010), participants saw either pictures or words in their natural orientation (e.g., ceiling above floor), or in their reverse orientation (e.g., floor above ceiling). Statistical linguistic frequencies were better able to explain RTs than perceptual ratings when the word pairs were used, with the reverse result when picture pairs were used. Similarly, when participants were asked to make a real-world judgment task, the effect for perceptual ratings on RTs was larger than that for statistical linguistic frequencies, with the opposite result for a semantic judgment task. Importantly, effects for both language statistics and perceptual simulation were found for both stimulus types and both cognitive tasks, however, their relative dominance was modified by task and stimulus.
These findings have been captured through the Symbol Interdependency Hypothesis, which proposed that conceptual processing can be explained by both symbol and embodied mechanisms (Louwerse, 2007; 2008; 2011). When we encounter a word, a rough meaning is elicited by using the linguistic, that is symbolic, neighbors. This is accomplished by using language statistics, where words that often appear together are related in important ways that can facilitate initial cognitive processing. In order to fully ground the word, we can mentally simulate the features of the word in order to process the word in a deeper way. Human beings can use the fuzzy sense of words by a linguistic (symbolic) short-cut when processing language as it occurs. In addition, language is encoded with sensorimotor and spatial information. The Symbol Interdependency Hypothesis is composed of three components. First, language encodes perceptual information. Second, during cognitive processes users of language rely on language statistics and perceptual simulation. Finally, the dominance of either language statistics or perceptual simulation is dependent on the type of task and stimulus.

Do these three claims also hold for spatial cognition within geographical representation? Using newspapers such as the New York Times and the Wall Street Journal

Louwerse and Zwaan (2009) were able to estimate the longitude and latitude of the largest cities in the US computationally, based on the idea that "cities that are located together are debated together." That is, by computing the \(n \times n\) frequencies of the co-occurrence of city names in the newspapers, a two-dimensional multidimensional scaling analysis yielded correlations with the longitude and latitude of the cities. The Louwerse and Zwaan (2009) findings are not limited to the English language. Louwerse, Hutchinson, and Cai (2012) found similar results using Arabic for predicting cities in the Middle East, and Chinese for predicting cities in China. It is interesting to note the presence of this effect was found for three languages each with different writing directions (English- left to right, Arabic- right to left, and Chinese, at least historically- top to bottom). This shows, at the least, that it is possible to map out cities in different locations, within different writing systems, by using the frequency of co-occurrences of city names within a large corpus.
Language encodes geographical information. The question is whether this also means that humans use these encodings. Louwerse and Zwaan (2009) stated that between \(16 \%\) and \(35 \%\) of the latitude and longitude variance in human location estimates can be attributed to linguistic coding. These percentages were found by using a bidimensional regression analysis correlating human and computational longitude and latitude estimates (by a large newspaper corpus). However, it is unclear whether \(84 \%\) and \(65 \%\) and of the latitude and longitude variance in human location estimates can be attributed to spatial information. Moreover, given that language encodes spatial information, it is difficult to disentangle linguistic and perceptual processes. It could be argued that proximity can explain estimation bias when determining distance between two locations (Tobler, 1970). However, Friedman, Kirkman, \& Brown (2002) tested this hypothesis by comparing latitude estimates by participants in Canada and Texas. Their findings did not support the proximity hypothesis, whereas participants in Texas exhibited greater bias in their estimates of Mexican locations than the participants from Canada. The explanations proposed by Friedman et al. included cognitively based beliefs, geopolitically based beliefs, and socio-culturally based beliefs. It was also argued by Brown (2002) that seeding effects can affect real-world judgments, such as proximity and size estimation of two cities. However, many of the experiments contained in Brown (2002) were designed for numerical estimates such as population, or how many square kilometers is for a given country. While they were robust and interesting effects, they do not necessarily apply here, because the tasks in the present study utilize the distance between two cities, not estimations of numbers about those locations.
Louwerse and Benesh (2012) investigated to what extent geographical estimates could come from language statistics and from perceptual simulations by comparing readers who had read Tolkien's Lord of the Rings trilogy and The Hobbit with participants who studied a map and had never seen the
text. As in Louwerse and Zwaan (2009), computational estimates of co-occurrence of the location of the cities in Middle Earth were determined. Participants were asked to draw the location of the cities on a piece of paper. Again, computational estimates of co-occurrence for cities mentioned in the text correlated with the longitude and latitude of cities in Middle Earth. Interestingly, estimates from those who studied a map correlated with the actual geographical location in Middle Earth more than the estimates from those who had read the text did. On the other hand, estimates from those who had read the text correlated more with the computational estimates of co-occurrence than the estimates from those who studied a map did. These results support the claims made by the Symbol Interdependency Hypothesis: 1) Language (Lord of the Rings) encodes geographical (Middle Earth) information; 2) Those who read Lord of the Rings and those who studied the map relied both on language statistics and perceptual simulation in their estimates; 3) the relative dominance of language statistics and perceptual simulation factors is modified by whether participants read the text or studied the map.

Importantly, human estimates in Louwerse and Zwaan (2009) and Louwerse and Benesh (2012) were derived from an experimental setting in which participants were asked to draw the location of cities on a piece of paper, which is a perceptual task. Given that the cognitive task determines the effect of language statistics and perceptual simulations (Louwerse \& Jeuniaux, 2010), the estimates how much of human geographical estimates come from language statistics and come from perceptual simulations is likely to be biased.
We therefore conducted an experiment in which participants were not asked to draw a map (a perceptual task) but to estimate geographical distances from words (a task that better justifies linguistic processing).

\section*{Experiment}

In a between subjects design, participants viewed United States city pairs in either a horizontal or vertical orientation. These city pairs randomly appeared in either their natural orientation (i.e., a more northern city was presented above a second city, or a more western city was presented to the left of a second city), or the opposite of their natural orientation. In this iconic orientation, we predicted that participants would rely on perceptual information. Conversely, when the location of the city pairs was reversed (i.e., reverse-iconic), we predicted that participants would rely on language statistics.

\section*{Methods}

Participants Ninety-three undergraduate native English speakers at the University of Memphis (67 females) participated for extra credit in a Psychology course. Fortyfive participants were randomly assigned to the vertical presentation condition and forty-eight participants were randomly assigned to the horizontal presentation condition.

Materials The experiment consisted of the largest 50 cities in the United States using the U.S. Census data from 2000 and were presented in 2,450 name pairs.

Procedure In two presentation conditions (horizontal or vertical), we presented subjects with city pairs in their iconic configuration and their reverse iconic configuration. Participants were randomly assigned to view either the vertical or horizontal configuration. To reduce order effects, participants were counterbalanced across four groups per condition.

The city pairs were presented on a \(1280 \times 1024\) computer screen. Participants were asked whether the named United States cities were closely located. The vagueness of the question intentionally left open the question of closeness for the participant to decide. A more specific question would have added a number of constraints that would influence the judgment in unintended ways. The center of the screen was positioned at eye level. Each trial began with the presentation of a fixation cross for 3000 ms . The participants would select their choice (yes or no) by designated buttons on a keyboard then a fixation cross would appear on the screen for the next trial.

\section*{Results}

Outliers were defined as response times (RTs) that were 2.5 SD above the mean per subject per condition and were removed from the analysis. This affected less than \(5 \%\) of the data.

The perceptual factor was operationalized as the differences in latitude or longitude of the cities. Language statistics was operationalized as the log frequency of \(a-b\) (e.g., for North - South: New York - Miami; for East West: Los Angeles - Boston), or \(b-a\) (e.g., for North South: Miami - New York; for West - East: Boston - Los Angeles) order of word pairs using the large Web 1T 5-gram corpus (Brants \& Franz, 2006). This corpus consists of 1 trillion word tokens (13,588,391 word types) from \(95,119,665,584\) sentences. Using the log frequency of the co-occurrence of word pairs enables linear regressions to be performed comparing frequencies with other types of data, because raw frequencies of those co-occurrences are extremely skewed (Gries, 2010).

A mixed-effect regression analysis was conducted on RTs with linguistic frequency and the perceptual factor as fixed factors and participants and items as random factors (Baayen, Davidson, \& Bates, 2008). The model was fitted using the restricted maximum likelihood estimation (REML) for the continuous variable (RT). F-test denominator degrees of freedom were estimated using the Kenward-Roger's degrees of freedom adjustment to reduce the chances of Type I error (Littell, Stroup, \& Freund, 2002). Participants and items were treated as random factors in the analysis.

Note that the strength of a model association is represented as a weighted ratio of the \(F\) statistic. \(R^{2}\) and \(F\) used in ordinary regression analysis are closely related, since where \(k\) is the number of model parameters and \(N\) is
the number of cases, such that \(F\) has \((k, N-k-1) \mathrm{df}\). See also Pedhazur (1997, p. 105) and Louwerse and Jeuniaux (2010). See Figure 1.
\[
F=\frac{\frac{R^{2}}{k}}{\frac{1-R^{2}}{N-k-1}}
\]

Figure 1. Weighted ratio of the \(F\) statistic.

Vertical Configuration The perceptual factor explained RTs in the iconic pairs, \(F(1,964.821)=17.7, p<.001\), with larger distances yielding lower RTs. The linguistic factor, however, did not explain RTs for the iconic word pairs, \(F(1,960.549)=0.45, p=.50\).

For the reverse iconic configuration the perceptual factor also explained RTs, \(F(1,984.502)=8.382, p=.004\), except that the effect was considerably smaller. Importantly, for these reverse-iconic word pairs a significant effect on RTs was obtained for the linguistic factor, \(F(1,970.543)=6.18, p\) \(=.013\), with higher frequencies yielding lower RTs. Figure 2 gives an estimate of effect sizes, which are calculated by differences between groups as opposed to within the two original groups.

Horizontal Configuration For the horizontal configuration, a similar pattern emerged as for the vertical configuration. That is, the perceptual factor explained RTs for city pairs in their iconic order, \(F(1,962.735)=9.645, p<.002\), but no significance found for language statistics when the position of the city pair was in the iconic order, \(F(1,995.626)=\) \(1.254, p=.263\).

For the reverse-iconic order the perceptual factor again explained RT, \(F(1,987.520)=9.565, p=.002\). Importantly, an effect for language statistics was obtained when city pairs were presented in their reverse-iconic order, \(F(1,1012.479)\) \(=4.068, p=.044\) (Figure 3).

\section*{Discussion}

The goal of the present study was to determine to what extent humans rely on language statistics and on perceptual simulation in spatial cognition. Previous work has found that language encodes geographical information, so much so that by computing the rates of co-occurrence of city names in the text, multidimensional scaling techniques allow for estimating the relative longitude and latitude of cities. Experiments have shown that humans rely on perceptual simulation, for instance, a perceptually grounded memory of the text. However, there is also evidence humans rely on language statistics, similar to those obtained from computational estimates. Because the existing literature used human estimates from map drawings, the current paper investigated to what extent linguistic and perceptual factors would affect cognitive processes in a more linguistic task.


Figure 2. Absolute \(t\)-values of the linguistic frequency and latitude differences in reverse-iconic and iconic orientation in the vertically positioned city names.


Figure 3. Absolute \(t\)-values of the linguistic frequency and longitude differences in reverse-iconic and iconic orientation in the horizontally positioned city names.

When city pairs were presented to participants in their iconic order, their distance best explained RTs. The larger the distance, the larger the RTs. No effect was obtained for language statistics in the iconic order. For the reverse-iconic order, the perceptual factor again explained RTs, but language statistics did so as well. This suggests that when the task or the stimulus invites for perceptual simulation, humans rely on perceptual simulation. When perceptual simulation is harder, other heuristics, such as language statistics are used. This finding lies fully in line with the results obtained by Louwerse and Jeuniaux (2010) showing that linguistic and perceptual factors dominate in conceptual processing when they are relevant.

Further research should investigate the weaker effects for the horizontal condition compared to those for the vertical condition. Barsalou (2008) argues that locating objects on a left/right axis is more difficult possibly due to the symmetry of the body and less salient cues to differentiate those objects. Perhaps this weaker effect is due to embodiment factors. However, this difference might also be explained by linguistic factors. When reporting two spatially related words in English, such as up-down or left-right, the top or the left most word is most often reported first. There is the possibility that there are less instances of the left-right phenomenon found in language. Future study of the nature of this phenomenon could illuminate why this weaker effect has been found. In the past, it has been shown that the linguistic system is used more often when quick decisions are made, and the perceptual system is used when slower decisions are made (Louwerse \& Connell, 2011). However, more specific investigation is recommended in the future as to the exact mechanisms of these speed differences and to what degree they affect decisions.

These findings reported in this paper are also in line with the Symbol Interdependency Hypothesis, which claims that cognitive processes rely both on language statistics and perceptual simulation. Because language encodes spatial information, including geographical information, language users can utilize these cues in their comprehension process. Geographical judgments then rely on both a shallow heuristic, called the linguistic system, and a fine-grained and more precise perceptual simulation system.

\section*{References}

Baayen, R. H., Davidson, D. J., \& Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language, 59(4), 390-412.
Barsalou, L. W. (1999). Perceptual symbol systems. Behavior and Brain Sciences, 22, 577-660.
Barsalou, L. W. (2008). Grounded cognition. Annual Review of Psychology, 59, 617-645.
Barsalou, L. W., Solomon, K. O., \& Wu, L. L. (1999). Perceptual simulation in conceptual tasks. In M. K. Hiraga, C. Sinha, \& S. Wilcox (Eds.), Cultural, typological, and psychological perspectives in cognitive linguistics: The proceedings of the \(4^{\text {th }}\) conference of the International Cognitive Linguistics Association, Vol. 3. Amsterdam: John Benjamins.
Boroditsky, L. (2000). Metaphoric structuring: Understanding time through spatial metaphors. Cognition, 75(1), 1-28.
Brants, T., \& Franz, A. (2006). Web 1T 5-gram Version 1. Philadelphia, PA: Linguistic Data Consortium.
Brown, N. R. (2002). Real-world estimation: Estimation modes and seeding effects. Psychology of Learning and Motivation, 41, 321-359.
Fincher-Kiefer, R., (2001). Perceptual components of situation models. Memory \& Cognition, 29, 336-343.

Freundschuh, S. M., \& Mercer, D. (1995). Spatial cognitive representations of story worlds acquired from maps and narrative. Geographical Systems, 2, 217-233.
Friedman, A., Kirkman, D. D., \& Brown, N. (2002). Spatial location judgments: A cross-national comparison of estimation bias in subjective North American geography. Psychonomic Bulletin \& Review, 9, 615-623.
Glenberg, A. M. \& Kaschak, M. P. (2002). Grounding language in action. Psychonomic Bulletin \& Review, 9, 558-565.
Gries, S. Th. (2010). Useful statistics for corpus linguistics. In Aquilino Sánchez \& Moisés Almela (Eds.), A mosaic of corpus linguistics: selected approaches. Frankfurt am Main: Peter Lang.
Kaschak, M. P., Madden, C. J., Therriault, D. J., Yaxley, R.H., Aveyard, M. E., Blanchard, A. A., \& Zwaan,
R. A. (2005). Perception of motion affects language processing. Cognition, 94, B79-B89.
Klatzky, R. L., Pellegrino, J. W., McCloskey, B. P., \& Doherty, S. (1989). Can you squeeze a tomato? The role of motor representations in semantic sensibility judgments. Journal of Memory and Language, 28, 56-77.
Littell, R. C., Stroup, W. W., \& Freund, R. J. (2002). SAS for linear models. Cary, NC: SAS Publishing.
Louwerse, M. M. (2007). Symbolic or embodied representations: A case for symbol interdependency. In T. Landauer, D. McNamara, S. Dennis, \& W. Kintsch (Eds.), Handbook of latent semantic analysis. Mahwah, NJ: Erlbaum.
Louwerse, M. M. (2008). Embodied relations are encoded in language. Psychonomic Bulletin \& Review, 15(4), 838844.

Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. TopiCS in Cognitive Science, 3, 273-302.
Louwerse, M. M., \& Benesh, N. (2012). Representing spatial structure through maps and language: Lord of the Rings encodes the spatial structure of Middle Earth. Cognitive Science, 36(8), 1556-1569.
Louwerse, M. M., \& Connell, L. (2011). A taste of words: Linguistic context and perceptual simulation predict the modality of words. Cognitive Science, 35(2), 381-398.
Louwerse, M. M., \& Jeuniaux, P. (2010). The linguistic and embodied nature of conceptual processing. Cognition, 114(1), 96-104.
Louwerse, M. M., \& Zwaan, R. A. (2009). Language encodes geographical information. Cognitive Science, 33(1), 51-73.
Louwerse, M. M., Hutchinson, S., \& Cai, Z. (2012). The Chinese Route Argument: Predicting the Longitude and Latitude of Cities in China and the Middle East Using Statistical Linguistic Frequencies. In N. Miyake, D. Peebles, \& R. P. Cooper (Eds.), Proceedings of the 34th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.

Marques. J. F. (2006). Specialization and semantic organization: Evidence for multiple semantics linked to sensory modalities. Memory \& Cognition, 34, 60-67.
Matlock, T., Ramscar, M., \& Boroditsky, L. (2005). The experiential link between spatial and temporal language. Cognitive Science, 29, 655-664.
Pecher, D., Zeelenberg, R., \& Barsalou, L. W. (2003). Verifying different-modality properties for concepts produces switching costs. Psychological Science, 14, 119-124.
Pecher, D. and Zwaan, R. (Eds.) (2005). Grounding cognition: The role of perception and action in memory, language, and thought. New York: Cambridge University Press.
Pedhazur, E. J. (1997). Multiple regression in behavioral research. Fort Worth, TX: Harcourt Brace.
Semin, G. \& Smith, E. (Eds.) (2008). Embodied grounding: Social, cognitive, affective, and neuroscientific approaches. New York: Cambridge University Press.
Spence, C., Nicholls, M. E., \& Driver, J. (2001). The cost of expecting events in the wrong sensory modality. Perception \& Psychophysics, 63(2), 330-336.
Tobler, W. (1970). A computer movie simulating urban growth in the Detroit region. Economic Geography Supplement, 46, 234-240.
Zwaan, R. A., Stanfield, R. A., \& Yaxley, R. H. (2002). Do language comprehenders routinely represent the shapes of objects? Psychological Science, 13, 168-171.
Zwaan, R. \& Yaxley, R. (2003). Spatial iconicity affects semantic relatedness judgments. Psychonomic Bulletin and Review, 10, 954-958.

\title{
Empirical investigation on spatial templates for a diagonal spatial term
}

\author{
Tokunaga, Takenobu (take@cl.cs.titech.ac.jp) \\ Watatani, Toshiaki (watatani.t.aa@m.titech.ac.jp) \\ Iida, Ryu (ryu-i@cl.cs.titech.ac.jp) \\ Department of Computer Science, Tokyo Institute of Technology \\ Terai, Asuka (asuka@nm.hum.titech.ac.jp) \\ Department of Human System Science, Tokyo Institute of Technology
}

\begin{abstract}
The meaning of spatial relations have been intensively studied in cognitive science research. A spatial template is one of the typical representations of spatial relations, which maps a position of a located object to its acceptability for the corresponding spatial term. Spatial templates have been investigated for several orthogonal spatial relations. However, diagonal spatial relations have attracted less attention. The present study aims at empirically determining the spatial template for a Japanese diagonal spatial term, "migiue (upper right)". The data was collected with various geometrical conditions changing the size of objects and the aspect ratio of the background. The analysis of the data revealed that the reference axis for "migiue (upper right)" was the direction of \(45^{\circ}\), and the acceptability of the diagonal relation could be affected by the acceptable regions of the adjacent orthogonal relations.
\end{abstract}

Keywords: spatial language; diagonal spatial term; spatial template;

\section*{Introduction}

There have been numerous studies on language and spatial relations in cognitive science (Talmy, 1983; Herskovits, 1985; Tversky \& Lee, 1998; Levinson, 2003; Coventry \& Garrod, 2004). Understanding a spatial relation involves reference objects (RO), located objects (LO), selection of an appropriate reference frame with respect to the context (CarlsonRadvansky, 1997; Carlson-Radvansky \& Jiang, 1998), and the meaning of the spatial relation. As a representation of the meaning of spatial relations, Logan and Sadler (1996) proposed a spatial template that maps an LO position to the acceptability for the corresponding spatial term. They determined the spatial templates for six projective spatial terms ( "above", "below", "left of", "right of", "over" and "under") and four topological spatial terms ("next to", "away from", "near to", and "far from") through experiments. The LO positions were discretised by a \(7 \times 7\) grid and each cell was assigned to an acceptability scale from 1 (bad) to 9 (good).

Surprisingly, diagonal spatial relations have attracted less attention than orthogonal spatial relations. One reason might be the fact that spatial terms expressing diagonal spatial relations tend to be lengthy in English, e.g. "the LO is in front of and to the right of the RO". In contrast, as Gapp (1995) noted, such combinations of spatial terms were very common in German and could be expressed in a simple form. This is also the case in Japanese, the target language of the present study. For instance, "migi (right)" and "ue (above)" can be directly combined to make a term "migiue (upper right)" for representing the upper right direction.

Another reason could be related to the so-called oblique effect, which claims humans show greater sensitivity to ratings with the orthogonal orientations, i.e. vertical and horizontal, than to other diagonal orientations (Appelle, 1972; Furmanski \& Engel, 2000; Meng \& Qian, 2005). The orthogonal spatial relations are more important for humans, thus these relations might have been intensively studied.

For investigating the acceptability of spatial relations, several researchers have used a radial grid layout (Huttenlocher, Hedges, \& Duncan, 1991; Gapp, 1995; Hayward \& Tarr, 1995; Crawford, Regier, \& Huttenlocher, 2000; Huttenlocher, Hedges, Corrigan, \& Crawford, 2004) instead of a square grid layout as Logan and Sadler (1996) did. They were interested in how angular deviation affected the acceptability of spatial terms. The spatial terms they were mainly concerned with were, however, still limited to orthogonal spatial terms \({ }^{1}\). They did not explicitly concern themselves with the acceptability of diagonal spatial terms such as "migiue (upper right)".

Against this background, the present study discusses the acceptability of a Japanese diagonal spatial term. More concretely, we aim at determining a spatial template for a Japanese term "migiue (upper right)" \({ }^{2}\) with taking into account three geometrical factors: the size of RO and LO, and the aspect ratio of the background. The background aspect ratio has rarely been taken into account in past studies.

\section*{Experiment 1}

\section*{Method}

Participants Thirty four undergraduates and graduates (30 males and 4 females) from Tokyo Institute of Technology participated in the experiment. Each participant received 1,000 JPY for his/her participation. All participants were native Japanese speakers.

Material and design We have four quadrants to consider for diagonal spatial terms: "upper right", "upper left", "lower right" and "lower left". Assuming symmetric acceptability

\footnotetext{
\({ }^{1}\) Gapp (1995) investigated diagonal spatial terms as a combination of two orthogonal spatial terms. He did not, however, take into account the dominance of orthogonal relations over diagonal relations.
\({ }^{2}\) Although we denote this target term as "upper right" in the rest of the paper, the actual term used in the experiments was the original Japanese term "migiue".
}


Figure 1: Example of stimulus
among these four quadrants, we investigated spatial templates for the upper right quadrant only. To obtain spatial templates for "upper right", we basically followed the goodness rating experiment described in (Logan \& Sadler, 1996). Figure 1 shows the interface of a trial that was presented to the participants. Against the coloured background, a square (the reference object: RO) is placed in the center, and the circle (the located object: LO) is placed someplace within the upper right quadrant of the background. A sentence describing the spatial relation between the RO and LO is presented below the figure; "En ha seihôkei no migiue ni aru. (the circle is to the upper right of the square.)" in this example. The participants were instructed to rate the relevance of the sentence describing the spatial relation between two objects on the scale of 1 (bad) to 7 (good) by clicking one of seven buttons.
\begin{tabular}{c|c|c|c|c|c|c|c|c|c|c|} 
& \multicolumn{2}{c}{1} & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\
E & E1 & E2 & E3 & E4 & E5 & E6 & E7 & E8 & E9 \\
D & D1 & D2 & D3 & D4 & D5 & D6 & D7 & D8 & D9 \\
C & C1 & C2 & C3 & C4 & C5 & C6 & C7 & C8 & C9 \\
B & B1 & B2 & B3 & B4 & B5 & B6 & B7 & B8 & B9 \\
A & A11 & A2 & A3 & A4 & A5 & A6 & A7 & A8 & A9 \\
\hline
\end{tabular}

Figure 2: Grid configuration for reference objects (Experiment 1)

The grid for the upper right quadrant with the origin at the RO position was configured as shown in Figure 2. The size of a cell was \(50 \times 50\) pixels. The LO was placed one of these cells with its centroid at the center of the cell. The RO was placed with its centroid at the left bottom corner of the cell A1. The grid lines were invisible to the participants. We considered three geometrical factors: the RO size (R), the LO size \((\mathrm{L})\) and the aspect ratio of the background (A).

We had two variations for the object size: large ( \(50 \times 50\) pixels) and small ( \(25 \times 25\) pixels), and three variations for the background aspect ratio: 5:5, 5:7 and 5:9. Depending on the background aspect ratio, the cells from the first column to the fifth column (5:5), the cells from the first column to the seventh column (5:7) and the cells from the first column to the ninth column (5:9) were used for the LO position respectively. The total size of the background was \(500 \times 500\) pixels for the \(5: 5\) case, \(500 \times 700\) for the \(5: 7\) case, and \(500 \times 900\) pixels for the \(5: 9\) cases. The number of LO positions varied depending on the background aspect ratio: 24 for the 5:5 case, 34 for the \(5: 7\) case and 44 for the 5:9 case \(^{3}\). The total number of trials for the acceptability rating became 408 ((\#RO size \() \times(\#\) LO size \() \times(\#\) LO position for the three aspect ratios) \(=2 \times 2 \times(24+34+44))\). In addition to these trials, 136 fillers were added in which the LO was placed in other quadrants with the sentences being changed accordingly. The total number of the trials for a participant was \(544(408+136)\).

Procedure The 544 trials were presented to each participant one by one on a 24 inch computer display of an iMac. The sequence of the trials were pseudo randomly generated with the occasional insertion of fillers for each participant. At one third and two thirds of the trial sequence, the participants were allowed to take a short break as long as he/she wanted. The participants finished their task within 20 to 40 minutes.
\begin{tabular}{|c|c|c|c|c|c|c|c|c|c|}
\hline & 1 & 2 & 3 & 4 & 5 & & & & \\
\hline E & 3.18 & 4.97 & 5.79 & 6.44 & 6.53 & & & & \\
\hline D & 3.32 & 5.06 & 6.12 & 6.62 & 6.15 & & & & \\
\hline C & 3.85 & 6.09 & 6.59 & 6.12 & 5.76 & & & & \\
\hline B & 4.26 & 6.59 & 5.53 & 5.03 & 5.12 & & & & \\
\hline A & - & 3.74 & 3.53 & 3.09 & 2.65 & & & & \\
\hline & 1 & 2 & 3 & 4 & 5 & 6 & 7 & & \\
\hline E & 2.88 & 4.71 & 5.76 & 5.88 & 6.41 & 6.50 & 6.50 & & \\
\hline D & 3.24 & 5.03 & 6.15 & 6.50 & 6.24 & 5.85 & 5.68 & & \\
\hline C & 3.68 & 6.06 & 6.44 & 6.06 & 5.79 & 5.35 & 5.24 & & \\
\hline B & 4.38 & 6.62 & 5.59 & 5.21 & 4.56 & 4.62 & 4.15 & & \\
\hline A & - & 4.0 & 3.47 & 3.38 & 2.91 & 2.94 & 2.41 & & \\
\hline & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\
\hline E & 3.06 & 4.79 & 5.24 & 5.88 & 6.29 & 6.12 & 6.15 & 6.00 & 6.06 \\
\hline D & 3.44 & 5.15 & 6.12 & 6.32 & 6.26 & 6.09 & 5.82 & 5.88 & 5.44 \\
\hline C & 3.59 & 5.59 & 6.47 & 6.21 & 5.85 & 5.50 & 5.44 & 5.18 & 4.91 \\
\hline B & 4.03 & 6.44 & 6.09 & 5.29 & 4.82 & 4.59 & 4.62 & 4.21 & 4.11 \\
\hline A & - & 3.71 & 3.44 & 3.26 & 3.32 & 2.91 & 2.85 & 2.32 & 2.29 \\
\hline
\end{tabular}

Figure 3: Spatial template for "migiue (upper right)"
(Experiment 1, \(\mathrm{R}=\) large, L=large)

Results Figure 3 shows spatial templates for "upper right" with a large RO and LO. Each cell denotes the average rating across all participants. The mean standard error of the averages in Figure 3 is 0.217 . This value is comparable to the result from (Logan \& Sadler, 1996), which is 0.271 . From

\footnotetext{
\({ }^{3}\) Note that the RO is fixed at the A1 position.
}
these templates, we can see that the direction of a \(45^{\circ}\) angle is the most relevant as a reference axis for "upper right" (red coloured cells) regardless of the background aspect ratio. In addition to these three templates, we had nine more templates for a combination of three geometrical factors: the RO size, the LO size and the background aspect ratio. We omit the other templates due to space constraints. The tendency of the other templates is similar to Figure 3.

Table 1: Result of four-way (A, R, L, P) ANOVA (Experiment 1)
\begin{tabular}{lrrcl}
\hline Effect & DFn & DFd & F & \multicolumn{1}{c}{\(p\)} \\
\hline A & 2 & 66 & 1.864 & 0.163 \\
R & 1 & 33 & 19.35 & \(0.0000^{* *}\) \\
L & 1 & 33 & 0.459 & 0.503 \\
P & 23 & 759 & 132.0 & \(0.000 * *\) \\
A-R & 2 & 66 & 0.348 & 0.707 \\
A-L & 2 & 66 & 2.233 & 0.115 \\
R-L & 1 & 33 & 16.04 & \(0.000^{* *}\) \\
A-P & 46 & 1518 & 2.084 & \(0.002^{* *}\) \\
R-P & 23 & 759 & 2.543 & \(0.001^{* *}\) \\
L-P & 23 & 759 & 1.545 & 0.099 \\
A-R-L & 2 & 66 & 2.043 & 0.146 \\
A-R-P & 46 & 1518 & 1.561 & \(0.042^{*}\) \\
A-L-P & 46 & 1518 & 0.924 & 0.576 \\
R-L-P & 23 & 759 & 1.044 & 0.407 \\
A-R-L-P & 46 & 1518 & 0.708 & 0.806 \\
\hline & \multicolumn{4}{c}{\((* *: p<.01, *: p<.05)\)}
\end{tabular}

Analysis We conducted a four-way ANOVA with average ratings as the dependent variable, and the background aspect ratio (A: 5:5, 5:7 and 5:9), the RO size (R: large and small), the LO size (L: large and small), and the LO position ( \(\mathrm{P}: 24\) positions) as the independent variables. Since the cells in the four right-most columns in Figure 2 were not included in the 5:5 aspect ratio configuration, we adopted only 24 cells in the column 1 to 5 for the analysis. Table 1 shows the result of the multivariate ANOVA indicating significant main effects of the LO position (P) and the RO size (R).


Figure 4: Interaction between object sizes
To investigate the effect of the aspect ratio (A) and the RO size (R) at each LO position (P), we conducted multiple comparisons for the interactions, P-A and P-R. The result of the multiple comparisons (Bonferroni's method) is shown in Table 2. Table 2 reveals that the effect by the RO size \((\mathrm{R})\) is

Table 2: Result of multiple comparisons
(Experiment 1, Bonferroni's method)
\begin{tabular}{cll}
\hline P & \multicolumn{1}{c}{A} & \multicolumn{1}{c}{R} \\
\hline A 2 & 0.967 & \(0.000^{* *}\) \\
A 3 & 0.657 & 0.450 \\
A 4 & 0.111 & 0.422 \\
A 5 & 0.066 & 0.098 \\
B 1 & 0.839 & 0.082 \\
B 2 & 0.589 & 0.982 \\
B 3 & \(0.035^{*}\) & 0.079 \\
B 4 & \(0.003^{* *}\) & 0.660 \\
B 5 & 0.864 & 0.945 \\
C 1 & 0.106 & \(0.002^{* *}\) \\
C 2 & 0.065 & 0.521 \\
C 3 & 0.092 & 0.463 \\
C 4 & 0.082 & 0.450 \\
C 5 & 0.365 & 0.713 \\
D 1 & 0.898 & 0.702 \\
D 2 & 0.245 & 0.081 \\
D 3 & \(0.023^{*}\) & 0.108 \\
D 4 & 0.079 & 0.545 \\
D 5 & 0.212 & 0.176 \\
E 1 & 0.828 & 0.394 \\
E 2 & 0.450 & 0.663 \\
E 3 & \(0.001^{* *}\) & 0.800 \\
E 4 & \(0.002^{* *}\) & 0.323 \\
E5 & \(0.001^{* *}\) & 0.251 \\
\hline\((* *: p<.01, *: p<.05)\)
\end{tabular}
particularly significant at the A2 and C1 positions, namely the positions close to the horizontal and vertical axes. Figure 4 shows the average ratings for the combinations of the RO and LO sizes, indicating that the smaller RO size tends to give higher ratings. In addition, the small LO with the small RO gives the highest ratings.
\begin{tabular}{|c|c|c|c|c|c|c|c|c|c|}
\hline & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\
\hline \multirow[b]{2}{*}{E} & 3.18 & 4.97 & 5.79 & \({ }_{5}^{6.44}\) & 6.53 & & & & \\
\hline & \begin{tabular}{l}
2.88 \\
3.06 \\
\hline
\end{tabular} & 4.79 & 5.76 & 5.88
5.88 & \[
\begin{aligned}
& 6.41 \\
& 6.29
\end{aligned}
\] & \[
\begin{aligned}
& 6.50 \\
& 6.12
\end{aligned}
\] & \[
\begin{aligned}
& 6.50 \\
& 6.15
\end{aligned}
\] & 6:00 & 6.06 \\
\hline \multirow{3}{*}{D} & 3.32 & 5.06 & 6.12 & 6.62 & 6.15 & & & & \\
\hline & 3.24 & 5.03 & 6.15 & 6.50 & 6.24 & 5.85 & 5.68 & & \\
\hline & 3.44 & 5.15 & 6.12 & 6.32 & 6.26 & 6.09 & 5.82 & 5.88 & 5.44 \\
\hline \multirow{4}{*}{C} & 3.85 & 6.09 & 6.59 & 6.12. & 5.76 & & & & \\
\hline & 3.68 & 6.06 & 6.44 & 6:06 & 5.79 & 5.35 & 5.24 & & \\
\hline & 3.59 & 5.59 & 6.47 & 6.21 & 5.85 & 5.50 & 5.44 & 5.18 & 4.91 \\
\hline & 2.26 & 6.59 & \(5: 53\) & 5.03 & 5.12 & & & & \\
\hline \multirow[t]{2}{*}{B} & 4.38 & 6.62 & 5.59 & 5.21 & 4.56 & 4.62 & 4.15 & & \\
\hline & 4.03 & 6.44 & 6.09 & 5.29 & 4.82 & 4.59 & 4.62 & 4.21 & 4.11 \\
\hline \multirow{3}{*}{A} & & 3:74 & 3.53 & 3.09 & & & & & \\
\hline & & 4.00 & 3.47 & 3.38 & 2.91 & 2.94 & 2.41 & & \\
\hline & & 3.71 & 3.44 & 3.26 & 3.32 & 2.91 & 2.85 & 2.32 & 2.29 \\
\hline
\end{tabular}

Figure 5: Overlaid spatial template of three aspect ratios (Experiment 1, R=large and \(\mathrm{L}=\) large)

Discussion Although there is no significant main effect of the background aspect ratio (A) in Table 1, several positions show a significant main effect of the aspect ratio in Table 2, i.e. B3, B4, D3, E3, E4 and E5. Among these positions, the ratings for the positions above the reference axis of "upper right", i.e. the \(45^{\circ}\) line, (D3, E3, and E4) were the highest with the aspect ratio 5:5. In contrast, the ratings for the positions below the reference axis (B3 and B4) were the highest
with the aspect ratio 5:9. This tendency was observed at other cells. Figure 5 shows an overlaid spatial template for "upper right" with different background aspect ratios (5:5, 5:7 and 5:9, i.e. three templates in Figure 3). The upper, middle and lower figures in a cell denote the acceptability ratings for the background aspect ratio 5:5, 5.7 and 5.9 respectively. The red coloured cells denote the reference axis for "upper right" and the dotted line denotes the diagonal line of the background with aspect ratios 5:7 and 5:9. This figure suggests that although the reference axis for "upper right" remains at the \(45^{\circ}\) direction regardless of the background aspect ratio, the acceptability for the positions below the reference axis is affected by the boundary of the background, i.e. the diagonal line of the background.


Figure 6: Effect by the RO size (Experiment 1)
Considering the earlier research results (Gapp, 1995; Regier \& Carlson, 2001; Kobayashi, Terai \& Tokunaga, 2008), given a fixed size of the background, a larger RO is expected to give a higher average rating, since the good region would enlarge according to the RO size. Figure 4, however, indicates the opposite result; the smaller RO gives the higher average rating. This would be explained by the effect of other spatial relations, "above" and "right" in this case. A larger RO enlarges the good region for "above" and "right" as well as that for "upper right". Considering the oblique effect, the orthogonal ("above" and "right") relations would be dominant over the diagonal ("upper right") relation, thus the enlargement of the good region for the diagonal relation would be suppressed by those of the adjacent orthogonal relations. The main effect of the RO size at positions A2 and C1 in Table 2 also supports this hypothesis. Figure 6 illustrates this explanation. When the RO is large (the left figure), the centroid of the LO is at the edge of the good region for "right" (the gray area), while when the RO is small (the right figure), the LO centroid is out of the good region. Thus, the acceptability of "rightness" in the right figure could be lower than that in the left figure, and the good region for "right" interferes less with that of "upper right". This hypothesis would explain the reason why the small RO gave a higher rating.

The interaction between the RO and LO sizes can be also explained in terms of the interference by the good region of the adjacent orthogonal spatial relations. As Figure 4 shows, when the RO is small, the average rating for the small LO is significantly higher than that for the large LO ( \(p<.05\) ), and


Figure 7: Interaction between the RO and LO sizes
when the LO is small, the average rating for the small RO is significantly higher than that for the large \(\mathrm{RO}(p<.01)\). Figure 7 illustrates the explanation for this observation. Two LOs are depicted as white circles in the same cell for comparison. The centroid of the LOs moves relative to the good region for "right" according to the combination of object sizes. When the RO is small (the right figure), the centroid of the LOs is out of the good region for "right", and the overlapping area between the LO and the good region drastically decreases as the LO becomes small (the inner circle). Actually, there is no overlap in this case. When the RO is large (the left figure), the centroid of the LOs is at the edge of the good region, and the difference of overlapping areas is less than that of the small RO case. The ratios of the overlap against the LO are the same; both overlapping areas are half of the object sizes. Thus, the good region for "right" has less effect on the good region for "upper right". That leads to the higher average rating for the small RO and LO. Table 2 shows a significant difference by the RO size (R) at the lowest horizontal cell (A2), which falls into the good region for "to the right of the RO". The difference is also significant at the leftmost vertical cell (C1), which falls into the good region for "above the RO". These significant interactions support the above explanation. In summary, we have drawn a hypothesis that since orthogonal relations are dominant over diagonal relations, the good region of the former would interfere with that of the latter. We conducted a follow up experiment in order to verify this hypothesis, which is described in the next section.
\begin{tabular}{|c|c|c|c|c|c|c|c|c|c|}
\hline & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\
\hline E & E1 & E2 & E3 & E4 & E5 & E6 & E7 & E8 & E9 \\
\hline D & D1 & D2 & D3 & D4 & D5 & D6 & D7 & D8 & D9 \\
\hline C & C1 & C2 & C3 & C4 & C5 & C6 & C7 & C8 & C9 \\
\hline B & B1 & B2 & B3 & B4 & B5 & B6 & B7 & B8 & B9 \\
\hline A & & A2 & A3 & A4 & A5 & A6 & A7 & A8 & A9 \\
\hline
\end{tabular}

Figure 8: Grid configuration for reference objects
(Experiment 2)

\section*{Experiment 2}

\section*{Method}

Participants Thirty three graduates and undergraduates (28 males and 8 females) from Tokyo Institute of Technology participated in the follow up experiment. There was no overlap in participants between the two experiments. Each participant received \(1,000 \mathrm{JPY}\) for his/her participation. All participants were native Japanese speakers.

Material and design The experimental setup is the same as Experiment 1 except for the grid configuration. In Experiment 2, the position of the RO was shifted by 50 pixels both downward and leftward as shown in Figure 8. The column 1 and row A were also shifted accordingly. This configuration is more similar to that of Logan and Sadler (1996) than the configuration of Experiment 1. The procedure of the experiment is the same as that of Experiment 1


Figure 9: Effect by the RO size (Experiment 2)
\begin{tabular}{|c|c|c|c|c|c|c|c|c|c|}
\hline & 1 & 2 & 3 & 4 & 5 & & & & \\
\hline E & 1.54 & 4.84 & 5.91 & 6.36 & 6.73 & & & & \\
\hline D & 1.30 & 5.39 & 6.15 & 6.76 & 6.30 & & & & \\
\hline C & 1.42 & 6.15 & 6.79 & 6.18 & 5.82 & & & & \\
\hline B & 1.33 & 6.90 & 6.03 & 5.24 & 5.09 & & & & \\
\hline A & - & 1.57 & 1.24 & 1.39 & 1.30 & & & & \\
\hline & 1 & 2 & 3 & 4 & 5 & 6 & 7 & & \\
\hline E & 1.30 & 4.94 & 5.70 & 6.36 & 6.61 & 6.64 & 6.70 & & \\
\hline D & 1.30 & 5.24 & 6.15 & 6.55 & 6.64 & 6.42 & 5.97 & & \\
\hline C & 1.24 & 5.70 & 6.76 & 6.45 & 6.21 & 5.73 & 5.55 & & \\
\hline B & 1.42 & 6.85 & 6.06 & 5.55 & 5.00 & 4.42 & 4.76 & & \\
\hline A & - & 1.42 & 1.33 & 1.33 & 1.39 & 1.33 & 1.67 & & \\
\hline & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\
\hline E & 1.36 & 4.88 & 5.48 & 6.21 & 6.58 & 6.48 & 6.39 & 6.42 & 6.30 \\
\hline D & 1.45 & 5.33 & 6.24 & 6.61 & 6.70 & 6.24 & 6.15 & 5.55 & 5.76 \\
\hline C & 1.36 & 6.06 & 6.64 & 6.48 & 6.12 & 5.73 & 5.61 & 5.39 & 5.00 \\
\hline B & 1.30 & 6.73 & 6.52 & 5.52 & 5.06 & 4.97 & 4.94 & 4.70 & 3.97 \\
\hline A & - & 1.36 & 1.33 & 1.39 & 1.42 & 1.36 & 1.42 & 1.33 & 1.30 \\
\hline
\end{tabular}

Figure 10: Spatial template for "migiue (upper right)" (Experiment 2, R=large, L=large)

If our hypothesis is correct, the effect of the RO size and the interaction between the RO and LO sizes would disappear
because the relative position of the LO centroid does not vary even though the RO size changes as shown in Figure 9.

Table 3: Result of four-way (A, R, L, P) ANOVA (Experiment 2)
\begin{tabular}{lrrll}
\hline Effect & DFn & DFd & \multicolumn{1}{c}{ F } & \(p\) \\
\hline A & 2 & 64 & 0.690 & 0.505 \\
R & 1 & 32 & 0.019 & 0.891 \\
L & 1 & 32 & 0.846 & 0.365 \\
P & 23 & 736 & 551.9 & \(0.000^{* *}\) \\
A-R & 2 & 64 & 2.948 & 0.060 \\
A-L & 2 & 64 & 0.064 & 0.938 \\
R-L & 1 & 32 & 3.518 & 0.070 \\
A-P & 46 & 1472 & 2.954 & \(0.000^{* *}\) \\
R-P & 23 & 736 & 1.242 & 0.228 \\
L-P & 23 & 736 & 1.890 & \(0.026^{*}\) \\
A-R-L & 2 & 64 & 1.618 & 0.206 \\
A-R-P & 46 & 1472 & 1.300 & 0.129 \\
A-L-P & 46 & 1472 & 1.107 & 0.323 \\
R-L-P & 23 & 736 & 0.619 & 0.855 \\
A-R-L-P & 46 & 1472 & 0.530 & 0.973 \\
\hline & \multicolumn{4}{c}{\((* *: p<.01, *: p<.05)\)}
\end{tabular}

Result Figure 10 shows the spatial templates for "upper right" with three different aspect ratios. The other conditions are the same as Figure 3. In this configuration, the participants tend to give very low ratings in the horizontal and vertical aligned cells, i.e. column 1 and row A. This is obviously because these cells are completely located within the good region for "above" and "right".

Analysis and discussion We conducted a four-way ANOVA in the same manner as Experiment 1. Table 3 shows the result of the multivariate ANOVA. As we expected, the main effect of the RO size ( R ) and the interaction between the RO and LO size disappeared. This supports our hypothesis described in the previous section.

\section*{General discussion}

The present study discussed the acceptability of the LO positions for a Japanese diagonal spatial term "migiue (upper right)" based on the empirical data. The data was collected through the experiments taking into account three geometrical factors: the size of RO and LO, and the background aspect ratio. Our findings through the data analysis can be summarised as follows.
- The reference axis of "migiue (upper right)" stays at the direction of \(45^{\circ}\) even though the aspect ratio of the background varies. This seems robust as far as the aspect ratios used in the experiments (5:5, 5:7 and 5:9). However, according to the horizontal extension of the background, the acceptability of the area below the reference axis tends to be higher, and that of the area above the reference axis tends to be lower. This would be the effect by the diagonal line of the background. This tendency is particularly remarkable in the distant area.

Interestingly, this observation is contrary to the finding by Gapp (1995), which claimed that the acceptability was not affected by the distance, although it became slightly higher when the LO was close to the RO within the area where the angular deviation from the reference axis is less than \(45^{\circ}\). In our data, the angular deviation is less than \(45^{\circ}\) in all positions, but the distant positions tend to show higher acceptability. For instance, the pairs of cells A2 and B5, and B1 and E2 have the same angular deviation from the reference axis, about \(\pm 26.5^{\circ}\). As Figure 3 shows, the ratings of the distant positions (B5 and E2) are consistently higher than that of the close positions (A2 and B1) in all aspect ratios. An ANOVA on the average ratings at these four positions showed a main effect of the distance \((F(1,33)=20.2, p<.01)\).
- The acceptability of the diagonal spatial relation is affected by the adjacent orthogonal spatial relations. In our case, the acceptable regions of "above" and "right" interfere with the acceptability of "upper right". This hypothesis was confirmed by the main effect of the RO size, and the interaction between the RO and LO size.

The above-mentioned contradiction between the results of ours and Gapp (1995)'s would be also explained by the interference by the orthogonal relations. The closer the LO is to the RO, the closer the LO is to the reference axis of the adjacent orthogonal relations ("above" and "right") as well, thus the acceptability would be affected more by the orthogonal relations.

Considering these findings together, we would say that the acceptability of diagonal spatial terms is determined by the interaction among four axes, namely, the horizontal axis, the virtical axis, the diagonal axis at \(45^{\circ}\) and the diaglonal axis of the background. Among these axes, the two orthogonal axes are most dominant as past studies suggested. The diagonal axis of the background seems most recessive but still affects the diagonal axis at \(45^{\circ}\).

Future research directions include the evaluation of existing computational models for spatial relations against the diagonal spatial relations. According to our preliminary experiments in which the Proximal and Centre-of-mass model and Attention Vector Sum model (Regier \& Carlson, 2001) were applied to our data for "upper right" with setting its reference axis at \(45^{\circ}\), these models fit quite well to the data. We found, however, the deviation from the data enlarged as the deviation of the diagonal axis of the background from the reference axis at \(45^{\circ}\) increased. As described above, these two diagonal axes should be taken into account in these computational models for diagonal spatial terms. We need further investigation to determine the quantitative effect of the interaction of these axes on the acceptability of the diagonal relations for building a computational model.

\section*{References}

Appelle, S. (1972). Perception and discrimination as a function of stimulus orientation - oblique effect in man and animals. Psychological Bulletin, 78(4), 266-278.
Carlson-Radvansky, L. A. (1997). The influence of reference frame selection on spatial template construction. Journal of Memory and Language, 37(3), 411-437.
Carlson-Radvansky, L. A., \& Jiang, Y. (1998). Inhibition accompanies reference-frame selection. Psychological Science, 9(5), 386-391.
Coventry, K. R., \& Garrod, S. C. (2004). Saying, seeing, and acting. Psychology Press.
Crawford, L., Regier, T., \& Huttenlocher, J. (2000). Linguistic and non-linguistic spatial categorization. Cognition, 75(3), 209-235.
Furmanski, C. S., \& Engel, S. A. (2000). An oblique effect in human primary visual cortex. Nature Neuroscience, 3(6), 535-536.
Gapp, K.-P. (1995). Angle, distance, shape, and their relationship to projective relations. In Proceedings of the 17th annual conference of the cognitive science society (p. 112-117).

Hayward, W. G., \& Tarr, M. J. (1995). Spatial language and spatial representation. Cognition, 55, 39-84.
Herskovits, A. (1985). Semantics and pragmtics of locative expressions. Cognitive Science, 9(3), 341-378.
Huttenlocher, J., Hedges, L. V., Corrigan, B., \& Crawford, L. E. (2004). Spatial categories and the estimation of location. Cognition, 93(2), 75-97.
Huttenlocher, J., Hedges, L. V., \& Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. Psychological Review, 98(3), 352376.

Kobayashi, T., Terai, A., \& Tokunaga, T. (2008). The role of attention in understanding spatial expressions under distractor condition. In Proceedings of the 5th international workshop on natural language processing and cognitive science (NLPCS 2008) (p. 74-83).
Levinson, S. C. (2003). Space in language and cognition. Cambridge University Press.
Logan, G. D., \& Sadler, D. D. (1996). A computational analysis of the apprehension of spatial relations. In P. Bloom, M. A. Peterson, L. Nadel, \& M. F. Garrett (Eds.), Language and space (language speech and communication) (p. 493-529). The MIT Press.
Meng, X., \& Qian, N. (2005). The oblique effect depends on perceived, rather than physical, orientation and direction. Vision Research, 45(27), 3402-3413.
Regier, T., \& Carlson, L. A. (2001). Grounding spatial language in perception. Journal of Experimental Psychology: General, 130(2), 273-298.
Talmy, L. (1983). How language structures space. In J. Herbert L. Pick \& L. P. Acredolo (Eds.), Spatial orientation - theory, research, and application (p. 225-282). Plenum Press.
Tversky, B., \& Lee, P. U. (1998). How space structures language. In Spatial cognition (Vol. LNCS 1404, p. 157175). Springer.

\title{
How intonation constrains pragmatic inference
}

\author{
John M. Tomlinson, Jr. (tomlinson@zas.gwz-berlin.de) \\ Zentrum für allegemeine Sprachwissenschaft (ZAS) \\ Berlin 10117 Germany
}

\author{
Lewis Bott (BottLA@cardiff.ac.uk) \\ School of Psychology, Cardiff University, Tower Building \\ Cardiff CF103AG UK
}

\begin{abstract}
In this paper, we present two experiments that investigate how intonation can constrain pragmatic inference. While prior research has shown that intonation can increase the likelihood of an inference being made, less is known about how it affects the mechanisms involved in processing of inferences. In the first experiment, listeners had more direct mouse paths towards target responses for stronger interpretations after hearing utterances with referents with pitch accents than without. In the second experiment, we replicate the finding of the first study and found more direct mouse paths towards weaker interpretations after hearing de-accented referents Our findings suggest that intonation constrains the online processing of pragmatic inference by increasing the availability of stronger interpretations.
\end{abstract}

Keywords: Experimental Pragmatics, Psycholinguistics, Prosody, Language Comprehension, Mouse-tracking.

\section*{Introduction}

The rapid nature of human communication requires speakers and listeners to be as efficient as possible. To help achieve this, listeners often rely on context to help disambiguate between different linguistic structures and meanings. However, often what a speaker intends to say is not always directly retrievable from a linguistic form; rather listeners must infer it. One issue concerning pragmatic inference is whether the processor can keep up with the task demands of conversation. Some have argued that linguistic inference must be quick and "cheap" (Levinson, 2000; Piantidosi, Tily, \& Gibson, 2012), however others have experimentally demonstrated that some linguistic inferences can be quite costly in terms of processing (Bott \& Noveck, 2004; Huang \& Snedeker, 2009). In this paper, we discuss one aspect of the linguistic signal that has the potential to make certain costly pragmatic inferences quicker and more efficient: prosody. We report the findings from two experiments that test different accounts about how prosody affects the processing of pragmatic inferences.

\section*{Pragmatic inferences and language processing}

Traditionally, linguists have treated pragmatic inferences as the interpretative process in which a speaker must reconcile how speaker's literal sentence meaning differs from his or her intended meaning. Grice (1967) initially distinguished between two types of pragmatic inferences
(particularized implicatures): conventional implicatures and conversational implicatures. Conventional implicatures roughly amount to inferences about a speaker's intended meaning that can be made without accessing the conversational context. Conversational implicatures, however, require that listeners must first consider the literal sentence meaning, compare it against the context and then potentially enrich it in order to arrive at a speaker's intended meaning. Neo-Griceans have proposed an inference type that falls somewhere between Grice's original distinction: default inferences (Levinson, 2000). Default inferences are inferences that are computed on every occasion, but can be cancelled later. Always deriving the inference avoids costly pragmatic computations that would delay obtaining the speaker's intended meaning. According to this process, inferences are heuristic-based and therefore can become "cheap" in regards to processing resources.

Researchers in experimental pragmatics have tested whether certain implicatures classes are indeed understood as default inferences. One case that has caused some debate is the case of scalar implicatures. For these inferences, listeners can choose between either a weak or a strong interpretation depending on what they think the speaker intended to communicate. For example, a sentence such as "I drank some of my friend's beers last night" could either be taken to mean that I drank (1) at least one (and possibly all) of the beers or (2) at least one and not all of the beers. The difference between interpretations (1) and (2) is that to interpret "not all" in (2), the listener must infer that had the speaker meant "all," they would have said so. In other words, the listener would need to make a pragmatic inference to access the stronger interpretation. Several experimental studies have shown that understanding upper bound meanings of some, as in (2), takes substantially longer than the meaning in (1) (Bott \& Noveck, 2004; Huang \& Snedeker, 2009). As such, a default implicature account of scalar implicatures is not borne out by the majority of these findings because of the processing cost for (2). However it might be the case that this processing cost can be diminished in the right context (Grodner, Klein, Canbary, \& Tannenhaus, 2010; Degen \& Tannenhaus, 2011), i.e. making (2) more available or active earlier on in processing. Our studies seek to examine how processing costs can be diminished and what this means for processing accounts of pragmatic inferences. Specifically, we examine
how one prominent cue, intonation, affects the availability and integration of various sources of information during the processing of conversational implicatures.

\section*{Intonation and Pragmatic Inferences}

Many studies have shown that intonation interacts with pragmatic processes in general, specifically those having to do with the integration of prior context to help disambiguate anaphoric reference, e.g. reference resolution, via information structure. For example, Dahan, Chambers, \& Tannenhaus (2002) found that pitch accents ( \(\mathrm{H}^{*}\) ) can rapidly disambiguate referents by integrating prior discourse mention of a referent. What is less clear is whether intonation affects pragmatic processes above that of explicatures, e.g. reference resolution, namely at the level of implicatures. For example, scalar implicatures are generally thought to be defeased in the antecedent of a conditional (if some of the...) and under negation (see e.g., Chierchia, 2004, for a review). Scalar implicatures therefore require the integration of semantics and pragmatics in a way that other pragmatic phenomena do not (see Horn, 2006). How and at what level of interpretation intonational information is intergrated into the processing of scalar implicatures is therefore an open question.

From a processing perspective, intonation could affect implicatures in at least two ways. First, it may alter how likely people are to derive an implicature. Secondly, it may also affect the speed with which people derive them. The difference is important because it allows us to understand in more detail how intonation interacts with other processing mechanisms. In particular, intonation might act merely as a cue to derive the implicature, or it may alter the process more fundamentally. In the next section we discuss previous findings related to prosody and pragmatic inferences, before specifying our hypotheses in more detail.

The one study that has specifically investigated prosody and scalar implicatures was Chevallier et al. (2008), who tested the effects of contrastive stress on the disjunction, or. Disjunctions can be optionally enriched from an inclusive reading, one or the other and possibly both, to an exclusive reading, one or the other but not both. Chevallier et al. tested whether contrastive stress on "or" affected the enrichment. For example, whether sentences like, "You can have the meat course or the fish course," was interpreted differently to, "You can have the meat course OR the fish course." While they found the stress on "or" greatly increased the proportion of exclusive readings, response times for the exclusive readings were identical regardless of whether contrastive stress was used or not. This study then, found that while intonation altered how the sentence was understood, it did not alter the time-course for the inference.

While our study is primarily concerned with conversational implicatures, other studies on intonation and different sorts of pragmatic inferences are clearly relevant. These studies have produced mixed results as to the effects of intonation on the speed of inference derivation, however, and it is often difficult to see whether intonation is affecting
speed of derivation or probability of derivation. For example, Dennison (2010) found that contrastive pitch accents in conjunction with final rises increased the likelihood that upon hearing "the pencil WAS sharp", listeners were more likely to infer that pencil is now not sharp, i.e. dull. This did not, however, affect the time course of processing relative to explicit negation: listeners spent as much time looking at pictures of the affirmative state (a sharp pencil) before fixating on the intended meaning (a dull pencil) as with explicit negation. Similarly, Sedivy et al. (1999) found no difference in looks to a referent disambiguated by a non-stressed adjective, "Click in the tall glass," vs. "Click on the TALL glass". In contrast, Ito \& Speer (2008) found that contrastive pitch accents ( \(\mathrm{L}+\mathrm{H}^{*}\) ) rapidly constrain the reference resolution of an upcoming noun. When listeners heard a prior mention of a referent (green ball), listeners were more likely to make anticipatory eye-movements upon hearing a contrastive pitch accent on BLUE to an object (ball) that had a contrasting item in the set (a blue ball vs. a red ball vs. a blue star).

The literature reviewed above suggests that intonation affects how likely implicatures are to be generated, but it is unclear whether it speeds up the process of making the inference. In our experiments we test the former hypothesis, namely whether a particular intonational pattern, focus intonation, speeds up the process of making conversational implicatures. One possibility is that because enrichment is optional, focus intonation could make it more likely that the procedures used to derive an implicature would be triggered (e.g., exhaustivity operator, (van Rooj and Schulz, 2004); or an only operator, (Chierchia, 2004); or reasoning about Gricean maxims, (Grice, 1975). If this is the only effect of the focus however, processing speed will not be altered and could even be delayed, e.g. more alternatives could be generated and considered. Focus would be one more cue to derive the implicature, but would not alter any of the procedures needed to perform the implicature computations. This account is consistent with the findings from Chevalier et al. (2008) and Dennison (2010). The other possibility is that focus intonation changes how the implicature is computed, which could happen several ways. For example, focus intonation might act like an explicit only in the sentence. This would remove the need to consider whether the speaker was informed and reliable (Sauerland, 2004). Removing this stage would speed up processing (Bott et al., 2012, demonstrate that scalar implicatures are computed more slowly than similar sentences with an explicit only). A final possibility is that focus might also encourage people to start deriving the implicature earlier on in the sentence; either because the pitch accent strengthens the assertive content of the proposition, e.g. the speaker is not leaving the topic open, or because the listener recognizes that a speaker is in a position to place a pitch accent on the referent.

\section*{Overview of experiments}

In this paper, we present two experiments investigating how prosody affects the processing of conversational
implicatures. At issue is whether prosody, in this case intonation, speeds up the process of making implicatures.

We used a picture-speech matching paradigm. Participants were presented with a visual display showing diverse objects. For example, a candle in one part of the screen and a dog in the other. They then heard a sentence assigning Mark ownership of one set of objects, and clicked on the image that best captured the object owned by Mark. For example, they might hear, "Mark has a candle" and then had to click on the candle image. In the critical trials, participants heard a sentence involving one object, "Mark has a candle (A)," but were presented with one image containing a candle (A), and one image containing a candle and a candy (AB). Now, in these trials, both options were logically permissible - there is a candle in both images; it is only by generating an implicature that the participant can chose the candle-only option ("the speaker must mean that Mark only has a A, and not AB, because otherwise they would have said so"). Thus, if the participant selected the candle-only option, they must have derived the implicature. We refer to the candle-only option(A) as the strong interpretation because it is informationally stronger than the candle and candy option (AB) (the weak interpretation).

Most importantly, we manipulated intonational focus on the referent. Participants heard either "Mark has a candle," or "Mark has a CANDLE." If focus intonation facilitates the derivation of the inference, the mouse-paths towards the stronger interpretation targets (CANDLE) should be more direct for stressed vs. unstressed referents when the twoobject picture is the competitor target.

\section*{Experiment 1}

In Experiment 1, the visual display involved two targets, one on the left and one on the right. Participants heard one of four types of experimental conditions, as shown in Table 1. Conditions 1 and 2 were the critical conditions described above, and conditions 3 and 4 were control conditions designed to eliminate low-level, perceptual explanations of any effects we might observe. If intonation speeds up the pragmatic process of deriving the implicature, we would expect a larger effect of intonation in conditions 1 and 2 than in conditions 3 and 4.

\section*{Method}

Twenty six undergraduate students in the School of Psychology at Cardiff University participated in this experiment for either course credit or a 3 pound Sterling reimbursement. The experiment took roughly 15 minutes to complete. All participants were debriefed upon completion.
Stimuli The same auditory stimuli were used for both experiments (except for the addition of prepositional phrases in Experiment 2). An utterance had the stem "Mark has a" and either had one referent (A)or two referents (AB) (see Table 1). Roughly half of the stimuli ( 24 items) were adapted from Dahan, Tannenhaus, \& Chambers (2002) and the other half ( 26 items) were created in order to increase the number of items. Of these items, half of the sentence and
picture combinations were phonological competitors, e.g. candle vs. camel and the other half were semantic competitors, e.g. pencil vs. eraser. This was done to help disguise the purpose of the experiment. For each item combination, black and white clip art pictures of each referent were constructed. Each item had either a picture of just one of the objects (candle) or both (candle and a camel). Objects were sized equally so that the picture of the object was the same size as when the object was in the two-object picture. This was done to control the salience of a oneobject picture versus a two-object picture. The utterancepicture combinations are also shown in Table 1.

A male speaker of British English with no noticeable regional variety was used to record the sentences. Sentences were recorded in a sound attenuated booth using a unidirectional microphone and digitized with USB sound capture device. All utterances were first recorded in sentence form and then the individual referents were recorded in isolation in both stressed and unstressed forms. A trained phonetician inspected these recordings and made sure that utterances with focus intonation had \(H^{*} \mathrm{~L}-\mathrm{L} \%\) patterns and non-focus intonation utterances had L*L-L\% patterns. Acoustic measurements were conducted so that this and mean F0 were the only significant different between the two versions. Next, objects in isolation were spliced into the sentence frames. In the two referent utterances, the pause between "and" and the second referent "a camel" was reduced to 100 ms so that listeners could not reliably use the stress to detect speaker continuation.

Table 1: Utterance-picture combinations Exp. 1
Utterance (Pitch accent) Picture(s)
Target conditions
(1) Mark has an A (L*) A vs. AB
(2) Mark has an \(\mathrm{A}\left(\mathrm{H}^{*}\right) \quad\) A vs. AB
(3) Mark has an A (L*) A vs. B
(4) Mark has an A \(\left(\mathrm{H}^{*}\right)\) A vs. B

Filler conditions
(5)Mark has an A ( \(L^{*}\) ) and a B ( \(L^{*}\) ) AB vs. A
(6)Mark has an \(A\left(\mathrm{H}^{*}\right)\) and a B ( \(\mathrm{H}^{*}\) ) AB vs. A
(7) Mark has an \(\mathrm{A}\left(\mathrm{H}^{*}\right)\) and a B ( \(\left.\mathrm{L}^{*}\right) \quad \mathrm{C} \mid \mathrm{B}\) vs. \(\mathrm{A} \mid \mathrm{B}\)
(8) Mark has an A (L*) and a B ( \(\mathrm{H}^{*}\) ) A|C vs. A|B

Design \& Procedure In both experiments, participants were were presented with an audio file and clicked on the picture that corresponded to the mentioned referent in the sentence. In the instructions, they were told that they were overhearing a speaker describing to another person which objects Mark has. Response boxes were equally sized and placed at the top left and right and corners of the screen. To begin each trial, participants clicked on START at the bottom center of the screen and then saw the response options for 2000 ms before the audio file was played. Participants could move their mouse and make their response at the onset of the word "has".

Participants were exposed to all conditions. Four experimental lists were generated so that a given participant had only one of the four target conditions for a given item. Filler conditions were added that had both related oneobject pictures as well as non-related one-object pictures. Filler conditions were kept the same across all lists. As mentioned in the stimuli section, all versions of filler picture conditions had utterances with both \(\mathrm{H}^{*} \mathrm{~L}-\mathrm{L} \%\) and \(\mathrm{L} * \mathrm{~L}-\mathrm{L} \%\) accent patterns on initial referents so that listeners would be as likely to hear focus intonation in both one and tworeferent utterance.
The experiment was run with Runner program in the Mousetracker suite (Freeman \& Ambady, 2010). The Analyser program exported responses into 101 normalized time steps. The dependent measure used was the Area under the Curve (AUC), which amounts to the total geometrical area for a mouse trajectory relative to a straight line from the start button to correct target.

\section*{Results}

The average mouse-paths for the target conditions are shown in Figure 1. Figure 1 shows the raw \(x\) - and \(y\) coordinates for the mouse-paths for the various conditions, showing that utterances with unstressed referents in the twoobject competitor condition have delayed mouse-paths towards the response target. Utterances with stressed referents in the two-object competitor condition do not look to be substantially delayed relative to the control conditions.

A mixed model with two predictor variables (focus intonation and competitor type) was used to test the directness of participants mouse paths (AUCs) towards the correct response. Intonation ( \(\mathrm{H}^{*}\) vs. L* pitch accents) and competitor type (weaker interpretation or phonological/semantic cohort) were used as fixed effects (along with an interaction term) and used subjects and items as random effects. In all conditions, accuracy rates were over \(97 \%\). Participants had more direct mouse paths to control condition (Conditions \(3 \& 4\) ) than when the weaker alternative was used as a competitor (Conditions \(1 \& 2\) ), \(t=\) 3.94, \(p<.01\). Across competitor type, focus intonation yielded more direct responses toward the correct target, \(t=\) \(3.31, p<.03\). Critically, the interaction between focus and competitor type was significant, \(t=2.91, p<.05\), suggesting that the main effects were driven by the relative difference of focus intonation between Conditions 1 and 2.


Figure 1: Raw x- on y-coordinates for Experiment 1.

\section*{Discussion}

In the presence of having a picture of the weaker interpretation as a competing target, listeners had more direct responses to the target picture of the stronger interpretation for utterances with a stressed referent than an unstressed referent. This suggests that the pitch accent made the weaker interpretation less accessible. Mouse-paths in Condition 2 were more direct towards the target and quite close to the control conditions. This means that focus seems to have substantially reduced the interference of the weaker interpretation competitor found in Condition 1 almost to the extent that is wasn't present (as in Conditions 3 and 4). These findings suggest that the implicatures have been processed more quickly in the focus condition.

An alternative explanation of our findings is that listeners could be interpreting the focus intonation as a discourse signal that the speaker has finished speaking. This would explain why participants mouse movements were more direct to the signal referent because listeners would be less likely to expect more upcoming speech from the listener. In our second experiment, we seek to eliminate this explanation of our findings.

\section*{Experiment 2}

Gricean maxims explain not only how speakers imply meanings beyond literal sentence meaning, but also provide allow listeners to infer whether a speaker has finished his/her turn. Moreover, research on intonation has shown that listeners interpret falling intonation at the end of the phrase to indicate that a speaker has finished his or her turn (Deruiter, Mitterer, \& Enfeld, 2006). In contrast, phrase final rising intonation can indicate both speaker continuation or uncertainty and this along with durational information can alter listeners' attention to upcoming speech (Tomlinson \& Fox Tree, 2011). Regarding our items in Experiment 1, it is possible that the falling intonation on the referent in phrase final position might have yielded more direct mouse paths to the correct target because listeners inferred that the speaker had finished speaking. To control for this possibility, prepositional phrases were added to each phrase, e.g. "Mark has a candle on the table". Because of this, two more competitor pictures were added to the display, increasing the possible targets from two to four.

Stimuli The same experimental items from Experiment 1 were used. However, a prepositional phrase was added (either "on the table" or "on the shelf") to the existing auditory files. Because of this, two more picture targets were added to each trial. In Conditions \(1 \& 2\), participants were now forced to choose between a picture of a candle and a camel on the table, a candle on the table, along with two distractor pictures (a picture of an apple and a pear on the shelf as well as a picture of an apple on the shelf). Conditions \(3 \& 4\) made use of table/shelf distinction by having participants choose between the single referent on either the table or the shelf along with the distractor pictures. Last, a third experimental condition testing the availability of weaker interpretations in our paradigm. In
this condition, items such as "Mark has a candle on the shelf" would be heard in the context of a picture of only a candle on a table and a picture of both a candle and a camel on a shelf along with the distractor pictures. In this case, participants would need to click on the picture of the weaker interpretation, as the prepositional phrase on the single referent would make the stronger interpretation incompatible with item.

\section*{Results \& Discussion}

The average mouse-paths for the target in conditions 1-2, 3-4, \& 5-6 are shown in Figures 2-4. Conditions \(1 \& 2\) show the same pattern as in Experiment 1, in that the focus intonation helped listeners choose the single referent target in the presence of a two-referent target. However focus did not have a yield a more direct mouse path to the target in the control condition.


Figure 2: Mouse paths for Conditions \(1 \& 2\) in Experiment 2.


Figure 3: Mouse paths for Conditions \(3 \& 4\) in Experiment 2.


Figure 4: Mouse paths for Conditions \(5 \& 6\) in Experiment 2.

A mixed-effect model was used to test AUC values with focus intonation ( \(\mathrm{H}^{*} \mathrm{~L}-\mathrm{L} \%\) vs. L* L-L\% patterns) and implicature type (stronger interpretation, control, weaker interpretation) as fixed effects (along with an interaction term) and with subjects and items as random effects. Accuracy rates were over \(97 \%\) for Conditions 1-4. However accuracy was only \(90 \%\) for Conditions 5-6. Overall, participants' responses to correct targets for control items (Conditions \(3 \& 4\) ) were more direct than both stronger interpretations (conditions \(1 \& 2\) ), \(t=4.07, p<.03\), and weaker interpretations (conditions \(5 \& 6\) ), \(t=7.44, p<.01\). Across all conditions, focus intonation was not a significant predictor of AUCs, \(t=1.29,=.31\). Critically, focus intonation yielded more direct mouse paths towards the correct target for stronger interpretations than for control conditions, \(t=2.79, p=04\). The opposite pattern was found for weaker interpretations: focus intonation yielded less direct mouse paths to correct targets compared to the control condition, \(t=2.03, p<.05\).

In sum, Experiment 2 replicated our findings from Experiment 1: focus intonation helps listeners exclude competition from weaker alternatives when selecting strong interpretations of an utterance. The added prepositional phrase and visual context suggests that the focus intonation is integrated incrementally. This also suggests that the finding from Experiment 1 did not result from listeners exclusively interpreting the focus intonation as a signal that the speaker has finished his or her turn. In addition, focus intonation made it more difficult for participants to choose weaker interpretations upon hearing an item with a single referent. This further suggests that the focus intonation is helping reinforce the "only" operator in such utterances.

\section*{Conclusion}

In two experiments, we sought to better understand how prosody, pitch accents, affects the interpretive processes of pragmatic inference. In our first experiment, focus intonation reduced the processing cost of understanding a stronger interpretation (Mark has only a candle) in the presence of a weak interpretation competitor. The second experiment replicated the findings from Experiment 1 in that focus intonation helped listeners exclude weaker interpretations when clicking on the correct target. Also, focus intonation introduced more competition for single referent pictures when choosing weaker interpretations.

We now discuss our findings as they relate to how and when prosody is integrated incrementally into utterance meanings. At first glance, our findings might suggest that focus intonation acted as an explicit only. This effect could arise by the focus intonation being initially decoded into at a phonological level and then fed forward into a focus operator into pragmatics via information theoretic relationships (Pierrehumbert \& Hirschberg, 1990; Büring 2007). Semantic accounts of focus might also explain our results (Krifka, 1999; van Rooij \& Schulz, 2004; Rooth, 1993). Such accounts hold that focus marking is integrated into utterance interpretations by triggering a search for
lexically available alternatives. As a result of the information structure, the constituent can take on additional meanings due to its elevated status relative to alternatives.

However, our findings only partly support this idea of intonation working at the level of information structure by ruling out contextually available alternatives. Because both stronger and weaker alternatives were visually available, listeners could not use the intonation to create or search out alternatives based on linguistic information. This suggests that listeners were integrating non-linguistic information into these interpretations e.g., visual information and/or speaker specific information, and that focus intonation sped up this integration. In other words, a more plausible explanation might be that focus intonation allowed listeners to start deriving the inference earlier on in the sentence.

Future work is needed to better tease apart these possibilities. One way forward would be to dovetail on a recent investigation by Breheny, Ferguson, \& Katsos (2013), which examined the rapid integration of speakers’ perspectives when processing ad hoc, conversational implicatures. In their study, listeners' eye movements were sensitive to speaker information when generating the "nothing else" implication, suggesting that information structure is necessary, but not sufficient for rapidly inferring "nothing else" implications: initial early biases toward the "nothing else" interpretation disappeared when listeners believed that the speaker's viewpoint of the objects was obscured. Although their confederate speakers in the look and listen experiment did not reliably use pitch accents when communicating the "nothing else" implication, an open question is whether focus intonation on the referents in their study would have reduced the delay in the speaker ignorance condition. We are conducting ongoing research to test this possibility, which can better adjudicate at what level intonation affects pragmatic inference.

\section*{Acknowledgments}

The first author was supported by ESRC grant RES-062-23240 awarded to L. Bott, T. Bailey, and D. Grodner. We would like to thank Elena Chepucova for helping with the preparation of the pictures and audio files.

\section*{References}

Beckman M.E. \& Ayers G.M. (1994). Guidlines for ToBI Labeling.
Bott, L. \& Noveck, I. A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. Journal of Memory and Language, 51, 437-57.
Breheny, R., Ferguson, H. J. \& Katsos, N. (2006). Taking the epistemic step: Toward a model of on-line access to conversational implicatures. Cognition, 126(3), 423-440.
Chevallier, C., Noveck, I. A, Bott, L., Lanzetti, L. Nazir, T., Sperber, D. (2008). Making disjunctions exclusive. Quarterly Journal of Experimental Psychology, 61(11), 1741-1760
Chierchia, G. (2004). Scalar implicatures, polarity phenomena, and the syntax/pragmatics interface. In A.

Belletti (ed.), Structures and beyond, (pp. 39-103). Oxford: Oxford University Press.
Freeman, J.B. \& Ambady, N. (2010). MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. Behavior Research Methods, 42, 226-241.
Dahan D., Tannenhaus M.K. \& Chambers, C.G. (2002). Acent and reference resolution in spoken-language comprehension. Journal of Memory and Language 47: 292-314.
Dennison, H. Y. (2010). Processing implied meaning through contrastive prosody. PhD dissertation. University of Hawai‘i at Mānoa, Honolulu: HI.
De Ruiter, J. P., Mitterer, H., \& Enfield, N. J. (2006). Predicting the end of a speaker's turn; a cognitive cornerstone of conversation. Language, 82(3), 515-535.
Grodner, D., Klein, N. M., Carbary, K. M., \& Tanenhaus, M. K. (2010). "Some", and possibly all, scalar inferences are not delayed: evidence for immediate pragmatic enrichment. Cognition, 116, 42 - 55.
Grice, P. (1967). Meaning. Philosophical Review 66, 377388.

Horn, Laurence R. (2006). The border wars. In Klaus von Heusinger and Ken P. Turner, eds., Where Semantics Meets Pragmatics, (pp. 21-48). Oxford: Elsevier
Huang, Y., \& Snedeker, J. (2009). On-line interpretation of scalar quantifiers: Insight into the semantics-pragmatics interface. Cognitive Psychology, 58, 376-415
Ito, K. \& Speer, S.R. (2008). Anticipatory effects of intonation: Eye movements during instructed visual search. Journal of Memory and Language 58: 541-573.
Levinson, S. C. (2000). Presumptive meanings - the theory of generalized conversational implicature. The MIT Press
Pierrehumbert, J. \& Hirschbirg, J. (1990). The Meaning of Intonational Contours in the Interpretation of Discourse, in P.R. Cohen, Morgen J. \& Pollack M.E. (ed.) Intentions in Communications: 271-311. Cambridge, Mass.: MIT Press.
Piantodosi, S.T., Tily, H., \& Gibson, E. (2012). The communicative function of ambiguity in language. Cognition, 122, 280-291.
Sauerland, U. (2004). Scalar implicatures in complex sentences. Linguistics and Philosophy, 32, 367-391.
Tomlinson, Jr, J. M. \& Fox Tree, J. E. (2012). The comprehension of uptalk in spontaneous speech. Cognition, 119, 58-69.
van Rooij, R. \& Schulz, K. (2004). Exhaustive interpretation of complex sentences. Journal of Logic, Language, and Information, 13, 491-519.

\title{
The visual motion aftereffect from mental imagery depends on speed
}

\author{
Alexia Toskos Dils (atoskos@stanford.edu) \\ Lera Boroditsky (lera@stanford.edu) \\ Stanford University, Department of Psychology \\ Jordan Hall, 450 Serra Mall, Building 420, Stanford, CA 94305 USA
}

\begin{abstract}
When we imagine a train snaking through a desert, does information about the train's speed make it into our visual mental image? In this paper, we make use of the motion aftereffect illusion (MAE) to test whether the speed of imagined visual motion modulates transfer of adaptation to a subsequent visual motion discrimination task. We compared the effects of viewing slow, medium, and fast motion on the magnitude of the MAE (Experiment 1) with the effects of simply imagining the same motion stimuli (Experiment 2). In Experiment 1 we found that increasing the speed of real visual motion from slow to medium produced a corresponding increase in the magnitude of the MAE, but increasing speed from medium to fast did not. Likewise, imagining slow motion produced a smaller MAE than did imagining medium motion, but the effect leveled of between medium and fast motion. These findings suggest that our mental imagery of motion is specific to the speed of the moving objects, and highlight areas of overlap between mental imagery and visual perception.
\end{abstract}

Keywords: Mental imagery; Motion aftereffect; Embodiment

\section*{Background}

When we imagine a car racing by, how visual is the process of creating the mental image? Do the representations we generate include information about how fast the car appears to be going? Or are they invariant to this property of visual motion perception? In this paper, we make use of the motion aftereffect illusion (MAE) to test whether the speed of imagined visual motion modulates transfer of adaptation to a subsequent visual motion discrimination task.

Researchers have long debated just how similar imagining a visual scene is to actually witnessing it (Kosslyn, 1981; Pylyshyn, 1973). Previous work examining the metric properties of imagined static scenes has found that information about size (Kosslyn, 1975), distance (Kosslyn, Ball, \& Reiser, 1978), and structure (Kosslyn, 1973) is indeed persevered in mental imagery. For example, Kosslyn et al (1978) found that the distance between objects in a mental image is proportional to the physical distance between their real-world counterparts. Participants in their study memorized a fictional map containing several landmarks, and were later asked to "scan" between pairs of landmarks in their mental image of the map. Results showed that the greater the distance between two landmarks on the physical map, the longer it took people to mentally scan between them.

Other work has shown that people are capable of mentally performing metric transformations on images of static objects (Finke, Pinker, \& Farah, 1989; Shepard \& Metzler, 1971). In a study by Shepard and Metzler (1971), participants judged whether pairs of geometric objects were identical to one another or mirror reversed. The authors reasoned that if people solved this task by mentally rotating one object until it aligned with the other, their reaction time should depend on the physical angular disparity between objects. Indeed, participants took longer to mentally rotate objects that would take longer to physically rotate, and vice versa.

The metric properties of mental imagery for dynamic scenes have not been studied as widely as for static scenes. One feature of visual motion that has been found to make it into mental imagery is motion direction. Winawer, Huk, \& Boroditsky (2008) demonstrated that imagining visual motion in a particular direction is sufficient to produce direction-selective adaptation in the visual system (i.e., produce a visual motion aftereffect illusion). After imagining upward motion, participants were more likely to see a subsequent dynamic stimulus as moving downward, and vice versa. Transfer of adaptation from mental imagery to perception suggests that a common neural mechanism underlies both processes. However, the degree of adaptation from mental imagery was considerably weaker compared to that from real visual motion perception, which sets a limit on the overlap between these two processes.

The adaptation paradigm used by Winawer and colleagues provides a unique testing ground for discovering other motion properties preserved in dynamic mental images. In this paper, we ask whether the magnitude of the visual motion aftereffect from mental imagery depends on the speed of imagined motion. If so, does motion speed modulate the MAE from imagery in the same way as speed modulates the MAE from real visual motion perception? That is, is speed yet another feature common to both mental imagery and perception, or is it an area in which internallygenerated motion representations abstract away from their externally-generated counterparts?

To test these questions, we first measured the effect of speed on the MAE from real visual motion (Experiment 1), and compared that with the MAE from imagining the very same motion stimuli (Experiment 2). In Experiment 1, subjects viewed videos of moving stripes (upward or downward) in three within-subject conditions: slow, medium, and fast. Following each video, participants indicated the direction in which a set of dynamic dots appeared to move. We found that increasing the speed of
visual motion from slow to medium produced a corresponding increase in the magnitude of the MAE, but increasing speed from medium to fast did not.

In Experiment 2, participants simply imagined the videos from Experiment 1 prior to completing the dot discrimination task. We found that imagining motion produced a reliable MAE (albeit weaker than from viewing real visual motion). We also found that viewing and imagining motion produced the same relative pattern of results across conditions. As in Experiment 1, imagining slow motion produced a smaller MAE than did medium or fast motion, but there was no difference between the medium and fast conditions.

\section*{Experiment 1}

How does motion speed modulate the magnitude of the MAE from real visual motion?

\section*{Methods}

Participants 30 Stanford undergraduate students participated in this study in exchange for payment.

Stimuli \& Procedure The task design, procedure, and visual stimuli used were modeled on those used by Winawer and colleagues (2008) and Dils and Boroditsky (2010). On each trial participants judged the direction of dot motion after viewing real visual motion. Trials were presented in 6 blocks: 3(speed: fast, medium, or slow) by 2(adaptation direction: upward or downward). The upward and downward versions of each speed were presented in succession. Block order was otherwise randomized across participants. Participants adapted to 60 seconds of motion in the first trial of each block. The adaptation phase of each subsequent trial lasted 6 seconds. There were 24 total trials per block.

Adapting stimuli. Participants watched videos of drifting black-and-white horizontal stripes. The videos showed a sine grating with a spatial frequency of 3.44 cycles per degree of visual angle drifting either upward or downward. In the medium condition, the grating drifted at 4.77 degrees per second. The slow grating drifted at half the speed of the medium grating ( 2.39 degrees per second), while the fast grating drifted at twice the speed of the medium grating ( 9.54 degrees per second). A flickering fixation cross was superimposed at the center of each video. The cross flickered at the same rate that the grating drifted. This feature was included to equate stimuli between Experiments 1 and 2, and it was task-irrelevant in the current study.

Test stimuli. Following the adaptation portion of each trial, participants judged the direction of motion coherence in a field of moving dots, without feedback. One hundred round dots were placed within a round aperture 10 degrees in diameter. The dots were light gray on a dark gray background, and each dot was 0.10 degrees in diameter. The dots moved at 12 degrees per second within the aperture, and any dots whose \(x-y\) coordinates exceeded the boundary of the aperture were randomly placed within the
aperture on each frame. A light gray static fixation dot 0.15 degrees in diameter was placed at the center of each dot display. Dot motion was always presented for 1 second, at which point the dot display disappeared from the screen. Participants pressed " f " if the dots appeared to move upward, and " j " if the dots appeared to move downward.

Each dot display had net motion coherence either up or down. For each subject, three coherence values were sampled 24 times in each direction. The values were tailored to each participant's dot motion sensitivity threshold (as assessed in a baseline task described below). They were selected to be \(12.5 \%, 25 \%\), and \(50 \%\) of the coherence necessary for each individual to detect the direction of motion in a dot display with \(99 \%\) accuracy. Coherence and direction of motion were fully crossed and balanced across trials and participants.

Baseline Motion Sensitivity Task. During the baseline motion sensitivity measurement, participants viewed 192 dynamic dot displays in succession and on each trial had to indicate the direction of motion coherence, upward or downward. Participants pressed the ' \(F\) ' key on a keyboard to indicate upward motion and the ' J ' key to indicate downward motion. The percentage of dots that moved coherently varied from trial to trial. In the baseline task, 12 coherence values were tested \((99 \%, 66 \%, 44 \%, 29 \%, 20 \%\), \(13 \%, 9 \%, 6 \%, 4 \%, 3 \%, 2 \%, 1 \%\) ), and each coherence level was sampled 8 times in each direction (upward / downward). A logistic function was fitted to each participant's data at the end of the baseline task, and the fit was used to compute the participant's threshold (the percentage of dot coherence required for \(75 \%\) accuracy). The threshold was then used to compute the coherence values to be used in the main experimental task, namely, values corresponding to \(50 \%, 25 \%\), and \(12.5 \%\) of the coherence necessary for asymptotic performance. These values were selected to be sufficiently difficult yet discriminable for participants. We refer to these 'normalized coherence' values rather than the actual subject-specific values in all references of motion coherence in reporting results.

Analysis. Participants who did not reach asymptotic performance on the baseline motion sensitivity test were excluded from all analyses ( 5 people). A logistic model was fitted to each participant's data from the main adaptation task. The regression models used a maximum likelihood algorithm to generate the fits and included a bias term, a term for motion coherence of the test stimulus, and three terms for the direction of the adapting stimulus (slow, medium, and fast motion). We computed the shift in the motion response functions as a function of adaptation direction for each level of motion speed. We used this analysis (1) to ensure that there was a reliable MAE in the full sample, and (2) to subsequently exclude participants who did not show an overall trend in the direction of an MAE after viewing real visual motion (4 participants). Since the aftereffect from real visual motion is typically large and robust, we reasoned that participants who did not
at least numerically respond in the direction of adaptation were likely not following task instructions. Even if they were engaged in the task, the absence of an aftereffect would prevent us from being able to assess its dependence on speed in those individuals.

Data from the remaining 21 participants was submitted to a mixed-models logistic regression. The model included fixed-effect parameters for coherence of the test stimulus, direction of adaptation, speed of the adapting stimulus (Helmert coded), and trial number. The model also included terms for the interaction between adaptation direction and motion speed, as well as adaptation direction and trial number. This last interaction term was included to account for longitudinal shifts in the aftereffect due to accumulation of adaptation and fatigue. Finally, the model included random slopes by participant for the full fixed-effects structure.

\section*{Results}

Figure 1 shows the raw, unfitted means across participants for upward and downward adaptation separately (including participants whose data was not in the direction of an MAE). In this inclusive sample, participants showed a \(164.5 \%\) shift between the motion response functions in the direction of adaptation. This difference was highly significant, \(\beta=-4.63, Z=-7.37, p<0.00001\).

Next we tested whether speed modulated the magnitude of the aftereffect in people who responded in the direction of adaptation overall. Indeed, viewing slow motion produced a smaller MAE than did viewing medium or fast motion ( \(\beta=-2.31, \mathrm{Z}=-4.68, p<0.00001\) ). This corresponded with a \(12 \%\) per deg/s increase in the probability of experiencing an MAE on a given trial. However, the increase in the MAE from viewing fast motion compared to medium motion was much smaller ( \(0.65 \%\) per \(\mathrm{deg} / \mathrm{s}\) ), and this shift did not reach significance \((\beta=-0.17, Z=-0.38\), \(p>0.5\) ). The predicted means from this analysis are plotted in Figure 2.

\section*{Discussion}

We asked whether the MAE from real visual motion perception depends on motion speed. We tested for MAEs following slow, medium, and fast visual motion, and we found that increasing speed from slow to medium or fast resulted in a corresponding increase in the magnitude of the MAE. However, we found no additional boost from increasing adaptation speed from medium to fast.

This pattern of results is consistent with previous findings on the relationship between speed of an adapting stimulus and the MAE (Ashida \& Osaka, 1995; see Mather, Verstraten, \& Anstis, 1998 for a review). For example, Ashida and Osaka found that the magnitude of the MAE for a given subject increases with speed until it peaks between 5-10 degrees per second. It then begins to decrease as speed continues to increase. The slow and medium conditions in the present study fall squarely within the rising phase of this
trajectory, but the fast condition falls early in the falling phase for most individuals.

In Experiment 2, we ask whether speed of imagined motion modulates the MAE in the same way as does speed of real visual motion.


Figure 1. Mean proportion UP responses after viewing real visual motion (upper panel) and after imagining visual motion (lower panel). Upward adaptation is plotted in red, and downward adaptation is plotted in blue. Error bars denote \(\pm 1\) s.e.m.

\section*{Experiment 2}

Does speed modulate the magnitude of the MAE from imagined motion? If so, is the pattern of results similar to what we observed from viewing real motion?

\section*{Methods}

Participants 30 Stanford undergraduate students participated in this study in exchange for payment.

Stimuli \& Procedure The stimuli and procedure for this experiment were identical to Experiment 1, except that participants imagined the drifting gratings during the adaptation portion of each trial rather than viewing them.

Before each block, participants were shown upward and downward examples of the grating videos that they would need to imagine during the block. Participants viewed each
video twice for 30 seconds before each block in which a new motion speed was being introduced. They viewed each video twice for 6 seconds before all other blocks. We made sure the visual motion presented during this familiarization phase did not interfere with our results during the main experimental task in three ways. 1. Participants were familiarized with both upward and downward motion, creating no net bias in either direction. 2. The familiarization was followed by at least 30 seconds of verbal instructions, a longer delay than necessary for an MAE from this duration of exposure to real visual motion stimuli to dissipate (Hershenson, 1989). 3. The direction of motion adaptation in the first experimental block following familiarization was chosen randomly.

At the beginning of each trial, an upward or downward facing arrow superimposed on a static image of the grating indicated the direction in which participants were to imagine the stripes moving. This cue faded over the course of a second. Once the cue disappeared completely, a flickering fixation cross appeared at the center of the screen. Participants were instructed to fixate on the cross while imagining the stripes and to use the rate of the flicker to help them remember how fast the stripes should move. Participants were also instructed to use the fixation cross as a cue for when to start and stop imagining motion.

Analysis. All analyses described in Experiment 1 were applied in the same way to the data from Experiment 2, including limiting our main analysis to participants who showed a motion aftereffect illusion. We know from previous work that there is considerable variation across individuals in the magnitude and direction of the aftereffect from internally-generated visual motion (Dils \& Boroditsky, 2010). Some individuals show a large MAE from mental imagery, others show a small MAE, and a small number shows priming and not adaptation. While the causes of these individual differences are not yet known, the variation itself is systematic. People who show an aftereffect from mental imagery also show an aftereffect from other forms of internally-generated visual motion such as linguistic descriptions of motion. The predictions we drew from Experiment 1 about how participants in Experiment 2 should behave only apply to people who showed an MAE overall, as we did not have enough participants who showed priming from real visual motion to create a set of predictions for this subgroup. Further, we did not have enough individuals who showed priming in Experiment 2 to measure the effect of speed on priming from imagined visual motion. Therefore, after first confirming that there was a reliable aftereffect from visual motion imagery in the entire sample, we limited the primary speed analysis of this paper to those participants whose responses at least numerically trended in the direction of a motion aftereffect.

We excluded 1 participant from all analyses for failing to reach asymptotic performance in the baseline sensitivity task. We excluded 5 participants from the main analysis whose results did not trend in the direction of adaptation. Additionally, we conducted a mixed-models analysis testing
for the presence of an interaction between experiments (Nieuwenhuis, Birte, \& Wagenmakers, 2011). This analysis included all previously described predictors plus a term for the concreteness of the adapting stimulus (real versus imagined visual motion), as well as the full factorial 3-way interaction between concreteness, adaptation direction, and motion speed.

\section*{Results}

Figure 1 shows the raw, unfitted means across participants for upward and downward adaptation separately (including participants whose data was not in the direction of an MAE). In this inclusive sample, participants showed a \(9.89 \%\) shift between the motion response functions in the direction of adaptation. This difference was reliable, \(\beta=-\) \(0.42, \mathrm{Z}=-2.13, p<0.05\).

Next we tested whether speed modulated the magnitude of the aftereffect from imagined motion in people who responded in the direction of adaptation overall. Indeed, viewing slow motion produced a smaller MAE than did viewing medium or fast motion ( \(\beta=-0.59, \mathrm{Z}=-2.08, p<0.05\) ). This corresponded with a \(4.97 \%\) per deg/s increase in the probability of experiencing an MAE on a given trial. However, the increase in the MAE from viewing fast motion compared to medium motion was much smaller ( \(0.60 \%\) per \(\mathrm{deg} / \mathrm{s}\) ), and this shift did not reach significance ( \(\beta=-0.04, Z=-0.13, p>0.5\) ). The predicted means from this analysis are plotted in Figure 2.

Finally, we asked whether the magnitude of these effects differed between real and imagined visual motion. The overall magnitude of the motion aftereffect illusion was greater for real visual motion than it was for imagined visual motion ( \(\beta=-3.30, \mathrm{Z}=-8.11, p<0.00001\) ). Also, the increase in the MAE from slow to medium and fast motion adaptation was significantly steeper for real visual motion than it was for imagined motion ( \(\beta=-1.74, \mathrm{Z}=-3.46\), \(p<0.001\) ).


Figure 2. Model estimates of the effect of speed on the degree of adaptation for real and imagined visual motion for average (zero)
coherence and average trial. Positive values are consistent with a motion aftereffect illusion.

\section*{Discussion}

In this study, we asked whether speed of visual motion is preserved in mental imagery. Specifically, we tested whether imagining slow, medium, and fast motion would differentially affect the magnitude of the motion aftereffect from mental imagery. We found that imagining motion indeed made people more likely to perceive a subsequent dynamic test stimulus as moving in the direction opposite the adapting motion. However, this effect was not constant across all speeds we tested. Increasing the speed of visual motion from slow to either medium or fast produced a corresponding increase in the magnitude of the MAE from imagery. However, increasing the speed of visual motion from medium to fast did not result in any additional increase in the MAE.

We also asked whether the relative effects of speed on the MAE from imagery would pattern like those from perception. Indeed both viewing and imagining motion produced a similar rise and then leveling off of the MAE as a function of speed. However, the initial rise was reliably steeper for real visual motion perception than for mental imagery.

\section*{General Discussion}

We started this paper by asking just how similar the representations generated in the service of mental imagery are to those generated during actual visual perception. We indeed found evidence of considerable overlap. First we replicated previous work showing that simply imagining motion is sufficient to produce a motion aftereffect illusion. This suggests that perception and mental imagery recruit, at least in part, the same direction-selective neural mechanisms in the visual system (Dils \& Boroditksy, 2010; Winawer, Huk, \& Boroditsky, 2008). Further, we found that visual motion speed modulates the MAE from both perception and imagery. The relative shape of the effect of speed is similar for internally- and externally-generated visual motion. This pattern suggests that the mechanisms recruited by both perception and mental imagery are in fact speed-specific.

However, we have also identified some key differences between visual motion processing and mental imagery. The effects of imagining motion on subsequent visual perception are considerably smaller overall than those from viewing real motion. Moreover, increasing visual motion speed produces a disproportionately smaller increase in the MAE from imagery relative to perception before it levels off. These findings call for a more nuanced view of how and when the processes that underlie mental imagery and perception interact, and when they diverge.

This work replicates and extends previous findings on the motion aftereffect from mental imagery (Dils \& Boroditsky, 2010; Winawer et al., 2008). The present findings help to rule out concerns that the MAE from internally-generated
motion results from a high-level cognitive bias and not from direction-selective adaptation of visual mechanisms. Cognitive bias should not depend on metric visual properties such as speed. Even if there were reason to predict such a relationship, it seems unlikely that it would lead to the specific pattern of results we observed. After all, we found that the very fastest imagined motion condition did not produce the largest MAE. Conversely, the real visual motion study provided a useful set of predictions about how speed should modulate the MAE from imagery.

While our findings suggest that speed is a feature of realworld visual motion that is preserved in mental imagery, it may be the case that our participants were particularly likely to create speed-specific mental images simply because it was one of the few differentiating features of our motion stimuli. Had our speed manipulation been subtler, perhaps we would not have seen it modulate the MAE from mental imagery. Future work aims to address whether features of visual motion such as speed, contrast, and spatial frequency creep into mental images automatically and irrespective of context, or whether they are represented in a more contextspecific way.

A further set of questions concerns speed represented in linguistic descriptions of motion. If we hear about a train racing versus crawling through the desert, do the resulting mental images contain some of the implied speed information? In previous work, it has been shown that speed implied in linguistic passages can have consequences for cognitive processing. For example, Matlock (2004) demonstrated that people are faster to process sentences describing fictive motion (e.g., The highway runs through the valley) after reading a story that describes fast motion compared to a story that describes slow motion. Future work can examine whether differences in the speed of implied motion described in language can also have visual consequences (e.g., in the size of the MAE) in addition to the speed of processing effects discovered by Matlock (2004).

\section*{Acknowledgments}

The authors would like to thank Megan D'Aguiar for assisting with data collection for these studies. We would also like to thank all members of the Cognation Lab. This research was supported by an NSF BCS \#1058119 to LB.

\section*{References}

Ashida, H., \& Osaka, N. (1995). Motion aftereffect with flickering test stimuli depends on adapting velocity. Vision Research, 13, 1825-1833.
Blake, R., \& Hiris, E. (1993). Another means for measuring the motion aftereffect. Vision Research, 33, 1589-1592.
Brainard, D. H. (1997) The Psychophysics Toolbox, Spatial Vision, 10, 433-436.
Dils, A. T., \& Boroditsky, L. (2010). Visual motion aftereffect from understanding motion language. Proceedings of the National Academy of Sciences, 107, 16396-16400.

Finke, R. A., Pinker, S. \& Farah, M. J. (1989). Reinterpreting visual patterns in mental imagery. Cognitive Science, 13, 51-78.
Hershenson, M. (1989). Duration, time constant, and decay of the linear motion aftereffect as a function of the inspection duration. Perception \& Psychophysics, 45, 251-257.
Kosslyn S. M. (1973). Scanning visual images: Some structural implications. Perception and Psychophysics, 14, 90-94. Kosslyn, S. M. (1981). The medium and the message in mental imagery: A theory. Psychological Review, 88, 46-66.
Kosslyn S. M. (1975). Information representation in visual images. Cognitive Psychology, 7, 341-370.
Kosslyn S. M., Ball, T., \& Reiser, B. (1978). Visual Images Preserve Metric Spatial Information: Evidence from Studies of Image Scanning. Journal of Experimental Psychology: Human Perception and Peformance, 4(1), 47-60.
Mather, G., Verstraten, F., \& Anstis, S. M. (1998). The motion aftereffect: a modern perspective. Cambridge, Mass.: MIT Press.
Matlock, T. (2004). Fictive motion as cognitive simulation. Memory and Cognition, 32 (8), 1389-1400.
Nieuwenhuis, S., Birte, F. U., \& Wagenmakers, E.-J. (2011). Erroneous analysis of interactions in neuroscience: a problem of significance. Nature Neuroscience, 14, 1105-1107.
Pylyshyn, Z. W. (1973). What the mind's eye tells the mind's brain: A critique of mental imagery. Psychological Bulletin, 80, 1-25.
Shepard, R. N., \& Metzler, J. (1971). Mental rotation of three-dimensional objects. Science, 171, 701-703.
Winawer J, Huk A, Boroditsky L. (2008) A motion aftereffect from viewing still photographs depicting motion. Psychological Science, 19, 276-283.

\title{
The Multi-attribute Linear Ballistic Accumulator Model of Decision-making
}

\author{
Jennifer S. Trueblood (jstruebl@uci.edu) \\ Department of Cognitive Sciences, University of California, Irvine Irvine, CA 92697 USA \\ Scott D. Brown and Andrew Heathcote \\ \{scott.brown, andrew.heathcote\} @newcastle.edu.au \\ School of Psychology, University of Newcastle, Australia \\ Callaghan, NSW 2308 AU
}

\begin{abstract}
Context effects - preference changes depending on the availability of other options - have wide ranging implications across applied and theoretical domains, and have driven the development of new dynamic models of multi-attribute and multialternative choice. We propose the Multi-attribute Linear Ballistic Accumulator (MLBA), a new dynamic model that provides a quantitative account of the co-occurrence of three context effects - attraction, similarity, and compromise - not only in traditional paradigms involving choices among hedonic stimuli but also of recent demonstrations of these effects with non-hedonic stimuli. The MLBA model has analytical solutions making it computationally easier to apply than previous dynamic models.
\end{abstract}

Keywords: Decision-making, multi-alternative choice, preference reversal, context effects, dynamic models

\section*{Introduction}

Individuals are often faced with the problem of choosing a single option from a large set of possible alternatives where the options have several features. For example, when purchasing a new cell phone, there are numerous phones from which to choose and each phone has many different features. A robust finding in the choice behavior literature is that preferences are subject to "context effects". That is, preferences for existing alternatives can be influenced or even reversed by the addition of new alternatives. For example, an initial preference for a cheap, low quality cell phone over an expensive, high quality phone could be reversed when a third expensive phone of low quality is also considered.

Three important context effects are the attraction (Huber, Payne, \& Puto, 1982), similarity (Tversky, 1972), and compromise (Simonson, 1989) effects. The standard experiment for the effects involves choices among three alternatives which each have two attributes. For example, three different cell phones with attributes of price and quality. Figure 1 graphically represents the positions of various options within a two dimensional space defined by two attribute values.

\section*{Three context effects}

The attraction effect refers to the enhancement of an option through the inclusion of a similar but slightly inferior decoy alternative. For the choice set \(\{X, Y\}\), let \(A_{X}\) and \(A_{Y}\) be similar to \(X\) and \(Y\) respectively but slightly inferior to each. For example, \(X\) might be a cheap, low quality cell phone and \(A_{X}\) might be the same quality as \(X\) but more expensive. The attraction effect occurs when people show
greater preference for \(X\) when \(A_{X}\) is included in the choice set \(\{X, Y\}\) as compared to when \(A_{Y}\) is included (and vice versa for \(Y\) ). Mathematically, the attraction effect occurs when \(\operatorname{Pr}\left[X \mid\left\{X, Y, A_{X}\right\}\right]>\operatorname{Pr}\left[X \mid\left\{X, Y, A_{Y}\right\}\right]\) and \(\operatorname{Pr}\left[Y \mid\left\{X, Y, A_{X}\right\}\right]<\) \(\operatorname{Pr}\left[Y \mid\left\{X, Y, A_{Y}\right\}\right]\).


Figure 1: Various options plotted in a two dimensional attribute space. Preferences between \(X\) and \(Y\) can be affected by the presence of other options.

The similarity effects refers to the enhancement of a dissimilar option when two similar options compete with one another. For the choice set \(\{X, Y\}\), let \(S_{X}\) and \(S_{Y}\) be similar and competitive to \(X\) and \(Y\) respectively. For example, if \(X\) is a cheap, low quality cell phone, then \(S_{X}\) might be a little more expensive and have slightly higher quality than \(X\). The similarity effect occurs when people show greater preference for the dissimilar option \(Y\) when \(S_{X}\) is included in the choice set \(\{X, Y\}\) as compared to when \(S_{Y}\) is included (and vice versa for \(X\) ). Mathematically, the similarity effect occurs when \(\operatorname{Pr}\left[X \mid\left\{X, Y, S_{X}\right\}\right]<\operatorname{Pr}\left[X \mid\left\{X, Y, S_{Y}\right\}\right]\) and \(\operatorname{Pr}\left[Y \mid\left\{X, Y, S_{X}\right\}\right]>\operatorname{Pr}\left[Y \mid\left\{X, Y, S_{Y}\right\}\right]\).

The compromise effect refers to the enhancement of an option when it is presented as a compromise between two other alternatives. For the choice set \(\{X, Y\}\), let \(C_{X}\) and \(C_{Y}\) be extreme options that make \(X\) and \(Y\) take the middle ground respectively. For example, if \(X\) is a cheap, low quality cell phone, then \(C_{X}\) might be drastically cheaper and extremely
lower quality than \(X\). The compromise effect occurs when people show greater preference for \(X\) when \(C_{X}\) is included in the choice set \(\{X, Y\}\) as compared to when \(C_{Y}\) is included (and vice versa for \(Y\) ). Mathematically, the compromise effect occurs when \(\operatorname{Pr}\left[X \mid\left\{X, Y, C_{X}\right\}\right]>\operatorname{Pr}\left[X \mid\left\{X, Y, C_{Y}\right\}\right]\) and \(\operatorname{Pr}\left[Y \mid\left\{X, Y, C_{X}\right\}\right]<\operatorname{Pr}\left[Y \mid\left\{X, Y, C_{Y}\right\}\right]\).

The three context effects are theoretically important because they violated the simple scalability property (Krantz, 1964; Tversky, 1972) which is a property of most utility models of choice including Luce's (1959) ratio of strengths model. To show a violation, consider the attraction effect. According to simple scalability, the inequality \(\operatorname{Pr}\left[X \mid\left\{X, Y, A_{X}\right\}\right]>\operatorname{Pr}\left[X \mid\left\{X, Y, A_{Y}\right\}\right]\) implies that the strength of \(A_{X}\) is less than the strength of \(A_{Y}\). However, the inequality \(\operatorname{Pr}\left[Y \mid\left\{X, Y, A_{X}\right\}\right]<\operatorname{Pr}\left[Y \mid\left\{X, Y, A_{Y}\right\}\right]\) implies that the strength of \(A_{Y}\) is less than the strength of \(A_{X}\). Both statements obviously cannot be true so the property is violated. The similarity and compromise effects produce similar violations.

\section*{Dynamic models of context effects}

Because utility models cannot account for context effects due to violations of simple scalability, researchers have turned to dynamic models to explain the effects. There are two predominate dynamic models of the effects: multi-alternative decision field theory (MDFT) (Roe, Busemeyer, \& Townsend, 2001) and the leaky competing accumulators (LCA) model (Usher \& McClelland, 2004). Even though these models have provided great insight into multi-alternative choice, they are not without flaws. First, both models require time intensive simulations for fitting data with internally controlled stopping times (the experimental procedure commonly used in context effects tasks in which participants control when they make decisions as opposed to an experimenter controlled deadline). Thus, it is difficult to fit the models to human data and evaluations of the models have relied on qualitative analyses such as showing that all three effects can be obtained using a single set of parameters. There has not been a quantitative comparison of the models and it remains unknown whether or not they can account for human data.

Further, the LCA model assumes that the attraction and compromise effects are the result of loss aversion. The loss aversion assumption seems reasonable for situations where the options have hedonic attributes such as consumer products with attributes of price and quality. However, there is recent evidence that context effects are a general feature of choice behavior and not specific to options with hedonic attributes. Trueblood (2012) demonstrated the three effects in an inference paradigm involving scenarios about criminal suspects. In these experiments, subjects were asked to infer which suspect out of a set of three was most likely to have committed a crime based on eye-witness evidence. Trueblood, Brown, Heathcote, and Busemeyer (in press) also showed the three effects in a simple perceptual task where subjects were asked to select the largest rectangle out of a set of three. Choplin and Hummel (2005) found the attrac-
tion effect with ovals and line segments in a similarity judgment paradigm and Tsetsos, Usher, and McClelland (2011) obtained the similarity effect using time-varing psychophysical stimuli. These experiments all suggest that the effects are not due to loss aversion because there is no notion of gains or losses along the attributes.

This paper introduces a new dynamic model, the multiattribute linear ballistic accumulator (MLBA) model, to account for context effects in multi-alternative choice. The MLBA model is easier to fit to data than MDFT and the LCA model because of its computational tractability. Also, it does not rely on loss aversion to explain the effects and thus can be applied to both hedonic and non-hedonic choices.

\section*{Precursors to the MLBA model}

The MLBA model is an extension of the linear ballistic accumulator (LBA) model (Brown \& Heathcote, 2008) that takes into account multiple attributes of options. The LBA models choice and response times with independent accumulators that race to a threshold \(\chi\). Each accumulator corresponds to a different option and the accumulator that first reaches the threshold is selected. Within a single trial, the accumulators are both linear and deterministic leading to mathematically tractable solutions for choice and response times. Each accumulator starts at a randomly determined amount of evidence selected from a uniform distribution \([0, A]\). The accumulators increase at speeds defined by the drift rates which are drawn from a normal distribution on each trial. Typically, the normal distributions have freely-estimated mean values \(d_{1}, d_{2}, d_{3} \ldots\) and a common standard deviation \(s=1\).


Figure 2: Connectionist network interpretation of MDFT.

The MLBA model also incorporates some of the cognitive mechanisms used in MDFT. MDFT assumes an individual's preferences are determined by a series of comparisons and evaluations of the alternatives that evolve across time. The preferences continue to evolve until one of the preference states, associated with one of the options, reaches a threshold and is selected. Preference states are determined by va-
lences for each option and lateral inhibition among the options. The valences are constructed from three components: subjective values, stochastic attention weights, and a comparison mechanism. The strength of the lateral inhibition is determined by the distance between two options in an "indifference/dominance" space (Hotaling, Busemeyer, \& Li, 2010).

MDFT can be interpreted as a connectionist network (Roe et al., 2001; Busemeyer \& Johnson, 2004), as illustrated in Figure 2. At each moment in time, attention can be allocated to attribute \(P\) (e.g., price) or attribute \(Q\) (e.g., quality), as illustrated in the first layer. The second layer shows each option weighted by the attributes. This layer instantiates the comparison mechanism and projects valences to the third layer. The valences are subject to the distance-dependent lateral inhibition process in the fourth layer, which outputs preferences.

\section*{The MLBA model}

The MLBA model adds to the LBA model by explicitly specifying how drift rates arise from the evaluation of attributes. It is assumed that mean drift rates are an increasing function of the valences of the options, where valences are defined in a similar manner to MDFT. Specifically, valences are determined by three components: subjective values, attention weights, and a comparison mechanism.

In determining the mean drift rates, each option is associated with a valence that represents the advantages or disadvantages of the option. For a set of options such as \(\{X, Y, Z\}\), the valences can be described by the vector \(V=\left[v_{X}, v_{Y}, v_{Z}\right]^{\prime}\). Let \(P\) and \(Q\) be attributes where \(P_{i}\) and \(Q_{i}\) denote the value of option \(i\) on each dimension. It is assumed that decisionmakers evaluate the subjective value of each option on each attribute producing the matrix of evaluations:
\[
M=\left[\begin{array}{ll}
m_{P 1} & m_{Q 1}  \tag{1}\\
m_{P 2} & m_{Q 2} \\
m_{P 3} & m_{Q 3}
\end{array}\right]
\]

In MDFT, the values \(m_{P i}\) and \(m_{Q i}\) represent an individual's subjective evaluation of the attributes. However, the exact form of this evaluation is not given. In the MLBA model, we constrain this form. We assume that the psychological \(m_{P i}\) and \(m_{Q i}\) values result from a local rescaling (i.e., within a single trial) of the experimentally defined attribute values, \(P_{i}\) and \(Q_{i}\). Wedell (1991) argued that context effects should be considered as local rather than global phenomena. Also, Gonzalez-Vallejo (2002) postulated a local rescaling of options in her proportional difference model.

There are at least three ways the local rescaling can be implemented. One possible rescaling arises from dividing the experimental values by the smallest values on attributes \(P\) and \(Q\) for a given set of options so that \(m_{P i}=P_{i} / \min (P)\) and \(m_{Q i}=Q_{i} / \min (Q)\). Another possibility is dividing the experimental values by the largest values on attributes \(P\) and \(Q\) for a given set of options so that \(m_{P i}=P_{i} / \max (P)\) and \(m_{Q i}=Q_{i} / \max (Q)\). The third option is to divide the experimental values by the average value of the attributes \(P\) and \(Q\)
for a given set of options so that \(m_{P i}=P_{i} / \frac{1}{3}\left(P_{1}+P_{2}+P_{3}\right)\) and \(m_{Q i}=Q_{i} / \frac{1}{3}\left(Q_{1}+Q_{2}+Q_{3}\right)\).

As in MDFT, the second component of the valence vector is the attention weights. We assume the decision-maker allocates a certain amount of attention to each attribute. The attention weight \(w_{P}\) represents the amount of attention allocated to the \(P\) attribute across the trial and \(w_{Q}\) represents the amount of attention allocated to the \(Q\) attribute across the trial. It is further assumed that \(w_{Q}=1-w_{P}\). The two attention weights are used to define the attention vector:
\[
W=\left[\begin{array}{ll}
w_{P} & w_{Q} \tag{2}
\end{array}\right]^{\prime} .
\]

The third component of the valence vector is a comparison mechanism. Like MDFT, this comparison process determines the relative advantage or disadvantage of each option on each attribute. The comparison process can be represented by a contrast matrix:
\[
C=\left[\begin{array}{ccc}
1 & -\frac{1}{2} \alpha_{12} & -\frac{1}{2} \alpha_{13}  \tag{3}\\
-\frac{1}{2} \alpha_{21} & 1 & -\frac{1}{2} \alpha_{23} \\
-\frac{1}{2} \alpha_{31} & -\frac{1}{2} \alpha_{32} & 1
\end{array}\right]
\]

Using this matrix, which assumes that \(\alpha_{i j}=\alpha_{j i}\), we can define the valence vector in a similar manner as in MDFT by the matrix product
\[
\begin{equation*}
V=C M W \tag{4}
\end{equation*}
\]

The weights \(\alpha_{i j}\) in the contrast matrix are defined by the indifference/dominance distance function developed by Hotaling et al. (2010) to determine the strength of the lateral inhibition in MDFT. Consider a pair of options \(\left(P_{i}, Q_{i}\right)\) and \(\left(P_{j}, Q_{j}\right)\). Define the distance between these two options as \((\Delta P, \Delta Q)=\left(P_{i}-P_{j}, Q_{i}-Q_{j}\right)\). These distances are then mapped to the corresponding coordinates in the indifference and dominance space: \((\Delta I, \Delta D)=1 \sqrt{2} \cdot((\Delta Q-\Delta P),(\Delta Q+\) \(\Delta P)\) ) where \(\Delta I\) is the difference along the indifference dimension and \(\Delta D\) is the difference along the dominance dimension. Using these coordinates, the distance function that weights changes more in the dominance dimension than the indifference dimension is defined as
\[
\begin{equation*}
\text { Dist }_{i j}=\sqrt{(\Delta I)^{2}+\beta \cdot(\Delta D)^{2}} \tag{5}
\end{equation*}
\]
where \(\beta \geq 1\). The distances are converted into similarities by the Gaussian function:
\[
\begin{equation*}
S_{i j}=\exp \left(-\phi \cdot D_{i s t} t_{j}\right) \tag{6}
\end{equation*}
\]

Using the similarities, we define the \(\alpha\) parameters as dissimilarity measures:
\[
\begin{equation*}
\alpha_{i j}=1-S_{i j} \tag{7}
\end{equation*}
\]

This mapping allows for options that are dissimilar to be weighted more in the comparison process. Finally, the valences are mapped into mean drift rates using a logistic function:
\[
\begin{equation*}
d_{i}=\frac{10}{1+\exp \left(-\gamma \cdot v_{i}\right)} \tag{8}
\end{equation*}
\]

In total, the model uses four free parameters to define the mean drift rates: the attention weight \(w_{P}\), the dominance weight \(\beta\), the similarity parameter \(\phi\), and the logistic parameter \(\gamma\). The model also has a starting point parameter \(A\), a threshold parameter \(\chi\) and a drift rate noise parameter \(s\) which are fixed when only modeling choice probabilities. For the current application, the noise parameter is assumed to the be the same for each accumulator. This assumption could be relaxed in other circumstances.

As with MDFT, the MLBA model can also be interpreted as a connectionist network as illustrated in Figure 3. For a given trial, a certain amount of attention can be allocated to attribute \(P\) and to attribute \(Q\) as illustrated in the first layer of the network. The second layer shows each option weighted by the attributes. This layer applies the contrast process and projects valences into the third layer. The valences are transformed into mean drift rates by a logistic function (f) in the fourth layer. Unlike MDFT, there is no lateral inhibition.


Figure 3: Connectionist network interpretation of MLBA.
The MLBA model accounts for the similarity effect through the local rescaling of the attribute values used to determine the \(M\) matrix. This local rescaling process results in more favorable subjective values for the dissimilar option. The model accounts for the attraction and compromise effects through the comparison process captured by the \(C\) matrix.

\section*{Combined Inference Experiment}

Context effects are typically tested using different groups of subjects for different effects. Thus, it remains an open question whether the attraction, compromise, and similarity effects can be found within a single experiment using the same subjects. This experiment investigates whether all three effects can be observed within the same experiment using the inference paradigm developed by Trueblood (2012). Data
from this experiment will be used in the subsequent section to compare MDFT and the MLBA model.

In Trueblood (2012), the three effects were tested in separate experiments using an inference paradigm involving decisions about criminal suspects. The experiments tested how people infer which suspect out of a set of three is most likely to have committed a crime based on two pieces of evidence. The evidence was described as strength ratings from two different eyewitness testimonies where the ratings ranged from 0 to 100 with a rating of 0 corresponding to very weak evidence for guilt and a rating of 100 corresponding to very strong evidence for guilt. In these crime scenarios, the suspects represent the different choice options and the eyewitness testimonies represent the two attributes in a similar manner as a consumer product with attributes of quality and price.

\section*{Method}

Sixty-eight undergraduate students from Indiana University participated for course credit. Participants were told they would see three suspects of a crime on each trial and were instructed to select the suspect that seemed most likely to have committed the crime based on the strengths of two different eyewitness testimonies. Participants were also told that the testimonies of both eyewitnesses were equally valid and important and that the strengths of the testimonies were equated. Participants did not receive feedback.

The suspects and eye-witness strengths were presented in a table format with different suspects in different rows. In the attraction effect experiment, for example, participants might have seen strength ratings of 63 (eyewitness 1) and 33 (eyewitness 2) for the suspect in row one, strength ratings of 32 (eyewitness 1) and 64 (eyewitness 2) for the suspect in row two, and strength ratings of 61 (eyewitness 1 ) and 31 (eyewitness 2) for the suspect in row three. In this example, the third suspect acts as the dominated decoy for the first suspect.

Each participant completed 240 trials which were divided into three blocks of 80 trials. The three blocks were used to test the three effects and were randomized across participants. Within each block, participants saw 40 trials testing one of the effects and 40 filler trials. The presentation order of the trials within each block was randomized. Filler trials where one alternative was clearly superior were used to assess accuracy throughout the experiment. The trials used to test for context effects were subdivided so that the decoy was placed near one alternative for some trials and near the other alternative for other trials. For example, the attraction effect was analyzed by comparing the choice sets \(\left\{X, Y, A_{X}\right\}\) and \(\left\{X, Y, A_{Y}\right\}\) as illustrated in Figure 1. The similarity effect was tested in two regions of the attribute space using a total of four ternary choice sets as in Trueblood (2012).

\section*{Results}

For data analyses, three participants were removed because their accuracy was two standard deviations lower than the average accuracy on the filler trials. Figure 4 shows the mean choice probabilities for focal and non-focal alternatives in the
attraction, similarity, and compromise effect trials. For the similarity effect, the focal option refers to the dissimilar alternative because this is the one enhanced by the decoy. For the attraction effect trials, the choice probability for the focal alternative \((M=0.548)\) was significantly larger than the choice probability for the non-focal alternative ( \(M=0.419\) ) \((\mathrm{t}(64)=3.141, \mathrm{p}=0.002)\). The similarity trials also showed that across all four choice sets the choice probabilities were significantly larger for focal options \((M=0.429)\) than nonfocal options \((\mathrm{M}=0.362)(\mathrm{t}(64)=2.578, \mathrm{p}=0.012)\). For the compromise effect, the choice probability for compromise alternatives \((M=0.466)\) was significantly larger than the choice probability for extreme alternatives \((M=0.407)\) \((\mathrm{t}(64)=2.172, \mathrm{p}=0.034)\).

\section*{Comparing MDFT and MLBA}

We fit MDFT and the MLBA model to the average choice probabilities across subjects from the combined inference experiment discussed above. We did not fit individual choice probabilities because there were not enough data from each subject. The two models were fit to choice probabilities only rather than choice probabilities and response times. In previous literature, multi-alternative choice models have only been analyzed with respect to choice probabilities. Future work could use response time measures to further test the models.

Because MDFT does not have analytical solutions for internally controlled stopping times, fitting the model requires computationally intensive simulations. To avoid this computational difficulty an approximate method developed by Hotaling et al. (2010) was used, where we fit the model using an externally controlled stopping procedure with a large stopping time. We fit the model by allowing four parameters to vary freely. One of the parameters determines the attention weight in the first layer of the model shown in Figure 2. The other three parameters are used in calculating the strength of the lateral inhibition. We fixed the decision time to the large value, \(t=1001\) used by Hotaling et al. (2010) and the withintrial variability parameter to 1 . Because the attribute values for the experiment were associated with eyewitness testimony strengths ranging from 0 to 100 , we used the attribute values divided by 10 for the subjective values.

The MLBA model was fit by numerically integrating over decision times as discussed in Hawkins et al. (submitted). For the MLBA model, we allowed the four parameters used to define the mean drift rates to vary freely and fixed the starting point parameter to \(A=1\), the threshold parameter to \(\chi=2\), and the drift rate noise parameter to \(s=1\). We fit three versions of the model corresponding to the three possible local rescalings: minimum, maximum, and average.

We fit a total of 24 choice probabilities arising from the eight ternary choice sets used in the experiment. The attraction and compromise effects each involved two ternary choice sets corresponding to the two possible locations of the decoy alternative. There were four ternary choice sets used for the similarity effect as described above. The models were fit
by minimizing the sum of squared error (SSE) between the model predictions and the data. When fitting mean probabilities, the SSE will approximate the maximum likelihood estimate. Table 1 gives the mean squared error (MSE) and the \(R^{2}\) values for the models.

Table 1: MSE and \(R^{2}\) values for MDFT and three versions of MLBA.
\begin{tabular}{lcc}
\multicolumn{1}{c}{ Model } & MSE & \(R^{2}\) \\
\hline MDFT & 0.037 & 0.251 \\
MLBA (minimum rescaling) & 0.030 & 0.400 \\
MLBA (maximum rescaling) & 0.016 & 0.684 \\
MLBA (average rescaling) & 0.029 & 0.414
\end{tabular}

The MLBA model using the maximum rescaling is able to account for about \(68 \%\) of the variability in the data as indicated by the \(R^{2}\) value. Substantially poorer performance was obtained for the MLBA model with minimum and average rescaling, although both still performed much better than the MDFT model. Future work could examine the differences between the three rescalings in more detail. We doubt MDFT's poor fits are due to the externally controlled stopping time procedure, as Hotaling et al. (2010) found that the externally controlled stopping time model with long stopping times produced essentially the same results as internally controlled stopping times.

\section*{Discussion}

Multi-alternative choice models such as MDFT and the LCA model have provided great insight into choice behavior, but these models have some drawbacks. They both required time intensive stimulations to fit data. Further the LCA model uses loss aversion to explain the attraction and compromise effects. The assumption that asymmetry between losses and gains is the underlying cause of these effects is problematic because the effects arise in paradigms were there are no losses or gains. The MLBA model overcomes these problems and provides a new psychological theory of context effects.

The MLBA model consists of two components: a frontend process that compares options along their attributes and a back-end process that determines the probability that a particular option will be selected. The back-end process is the LBA model developed by Brown and Heathcote (2008). This paper develops the front-end attribute processing component. The coupling of front-end and back-end processes is not new. The SAMBA model (Brown, Marley, Donkin, \& Heathcote, 2008) of choice and response times in absolute identification tasks proposes a front end to the Ballistic Accumulator model (Brown \& Heathcote, 2005). We suggest this pairing can be viewed as the front-end process modulating action selection in the back-end process. Models incorporating such modulation are common in the neurophysiological literature. For example, Frank (2005) developed a model in which the striatum modulates motor actions and working memory updating


Figure 4: Mean choice probabilities with error bars showing the standard error of the mean for the attraction, similarity, and compromise effects from the combined inference experiment.
in frontal cortex. We do not argue that our model can be mapped to specific brain regions, but speculate that the gating of actions by a front-end process could have a neural explanation.

\section*{Acknowledgments}

Jennifer Trueblood was supported by the IGERT program in Dynamics of Brain-Body-Environment Systems at Indiana University. Scott Brown and Andrew Heathcote were supported by ARC fellowships. The authors would also like to thank Jerome Busemeyer for his helpful comments.

\section*{References}

Brown, S. D., \& Heathcote, A. (2005). A ballistic model of choice response time. Psychological Review, 112, 117128.

Brown, S. D., \& Heathcote, A. (2008). The simplest compete model of choice response time: Linear ballistic accumulation. Cognitive Psychology, 57, 153-178.
Brown, S. D., Marley, A. A. J., Donkin, C., \& Heathcote, A. (2008). An integrated model of choices and response times in absolute identification. Psychological Review, 115(2), 396-425.
Busemeyer, J. R., \& Johnson, J. G. (2004). Computational models of decision making. In D. Koehler \& N. Harvey (Eds.), Handbook of judgment and decision making (p. 133-154). Blackwell Publishing Co.

Choplin, J., \& Hummel, J. (2005). Comparison-induced decoy effects. Memory \& Cognition, 33, 332343.
Frank, M. J. (2005). Dynamic dopamine modulation in the basal ganglia: a neurocomputational account of cognitive deficits in medicated and non-medicated parkinsonism. \(J\) Cogn Neurosci, 17, 51-72.
Gonzalez-Vallejo, C. (2002). Making trade-offs: A probabilistic and context-sensitive model of choice behavior. Psychological Review, 109, 137-155.
Hawkins, G. E., Marley, A. A. J., Heathcote, A., Flynn, T. N., Louviere, J. J., \& Brown, S. D. (submitted). Integrating cognitive process and descriptive models of attitudes and preferences.

Hotaling, J. M., Busemeyer, J. R., \& Li, J. (2010). Theoretical developments in decision field theory: A comment on k . tsetsos, n. chater, and m. usher. Psychological Review, 117, 1294-1298.
Huber, J., Payne, J. W., \& Puto, C. (1982). Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. Journal of Consumer Research, 9, 90-98.
Krantz, D. H. (1964). The scaling of small and large color differences. (doctoral dissertation, university of pennsylvania). Ann Arbor, Michigan: University Microfilms. No. 65-5777.
Luce, R. D. (1959). Individual choice behavior: A theoretical analysis. New York: Wiley.
Roe, R. M., Busemeyer, J. R., \& Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionist model of decision making. Psychological Review, 108, 370-392.
Simonson, I. (1989). Choice based on reasons: The case of attraction and compromise effects. Journal of Consumer Research, 16, 158-174.
Trueblood, J. S. (2012). Multi-alternative context effects obtained using an inference task. Psychonomic Bulletin \& Review, 19 (5), 962-968.
Trueblood, J. S., Brown, S. D., Heathcote, A., \& Busemeyer, J. R. (in press). Not just for consumers: Context effects are fundamental to decision-making. Psychological Science.
Tsetsos, K., Usher, M., \& McClelland, J. (2011). Testing multi-alternative decision models with non-stationary evidence. Frontiers in Neuroscience, 5, 1-18.
Tversky, A. (1972). Elimination by aspects: a theory of choice. Psychological Review, 79, 281-299.
Usher, M., \& McClelland, J. L. (2004). Loss aversion and inhibition in dynamical models of multialternative choice. Psychological Review, 111, 757-769.
Wedell, D. H. (1991). Distinguishing among models of contextually induced preference reversals. Journal of Experimental Psychology: Learning, Memory, and Cognition, 17, 767-778.

\title{
Perceiving sounds: analytic and synthetic listening, global-local processing and possible links with empathy and self-construal
}

\author{
Olga Tsoumani (olga.tsoumani@kuleuven.be) \\ Department of Marketing and Organisation \\ KU Leuven, Bus 3545, 3000 Leuven, Belgium
}

\author{
Marie Postma-Nilsenová (m.nilsenova@tilburguniversity.edu) \\ Department of Communication and Information Sciences, Tilburg University, P.O. Box 9013, 5000 LE Tilburg, The Netherlands
}

\begin{abstract}
In two experiments we examined the effects of training on auditory perception bias (Experiment 1), the relationship between auditory perception bias and global-local processing (Experiment 2), as well as the relationship between globallocal auditory processing, empathy and self-construal (Experiment 2). The present findings are discussed in relation to their implications for research in auditory perception and the perception of others' emotional states.
\end{abstract}

\section*{Introduction}
"C'est quoi, le pitch?" used to be the favorite question of the famous French TV talk show host, Thierry Ardisson, when he was interviewing writers, film makers and politicians alike. Knowing what the pitch is may not just be important on French television but plays an important role in our development of linguistic abilities as well. Starting in early infancy, our early auditory ability to process pitch and detect pitch contour deviations appears to be tightly linked to our ability to extract linguistic rules (Mueller, Friederici, \& Männel, 2012). Pitch pattern perception has been shown to be an important predictor of reading performance both in skilled readers and children with developmental dyslexia (Foxton et al., 2003; Ziegler et al., 2012) and to play a role in L2 acquisition (Wong \& Perrachione, 2007). However, pitch processing and production play an important social role in two ways: First, pitch modulation is a carrier of information about speakers' emotions and attitudes (Scherer et al., 1991; Juslin \& Laukka, 2003). Second, pitch imitation is exploited in promoting social convergence and status accommodation (Gregory, 1983; Gregory \& Hoyt, 1982; Gregory, Webster, \& Huang, 1993; Gregory \& Webster, 1996; Gregory, Dagan, \& Webster, 1997; Gregory \& Gallagher, 2002) and expressing ingroup-outgroup bias (Babel, 2009). In sum, an early assessment and training of a listener's ability to process rapid pitch changes in the speech signal could contribute to the development of tools for diagnosis and remediation of different types of language and communication disorders.
What makes pitch detection difficult? Pitch is, roughly, the perceptual correlate of fundamental frequency, produced primarily by the vibrations of vocal chords. It is both the most prominent and most elusive component of the complex sound produced by human articulators because its
perception is influenced both at the level of primary auditory mechanisms in the ear (which, mainly due to the nonlinearities in the cochlea, may supply input in the fundamental frequency region; Moore, 2003) and at the level of neural processing in the auditory cortex (Schneider et al., 2005). Interestingly, the way complex sounds are perceived seems to differ systematically between individuals: Some listeners - known as \(\mathrm{f}_{0}\) or synthetic/holistic listeners - focus primarily on the region between \(50-500 \mathrm{~Hz}\), the region where the fundamental frequency can be found. Others - known as spectral/analytic listeners - rely on analyzing the harmonic constituents of the sound and focus on the spectrum "as a whole" (e.g. von Helmholtz, 1885). A neurological basis has been suggested for this difference, according to which there is a leftward vs. rightward asymmetry of the lateral Heschl's gyrus for synthetic and analytics listeners, respectively (e.g. Schneider et al., 2005). The auditory perception bias has been almost exclusively analyzed in the context of musical training, but the results of individual studies indicate that it may also affect linguistic performance (Wong \& Perrachione, 2007; Wong et al., 2008), as well as pitch imitation (PostmaNilsenová \& Postma, 2012).
Most of the research on the synthetic and analytic listener types suggests that their auditory perception bias is a stable individual difference, possibly caused by genetic factors (Dediu \& Ladd, 2007; Wong, Chandrasekaran, \& Zheng, 2012). However, musical competence and training can affect the listening mode and lead to a shift from spectral to fundamental listening (Seither-Preisler et al., 2007). Also, repeated exposure to stimuli with a missing fundamental frequency over the course of several months appears to facilitate the synthetic listening mode and thus, presumably, to improve pitch perception (Seither-Preisler et al., 2007; Postma-Nilsenová \& Postma, 2012).

In the first part of our study, we explore the possible effect of training on auditory perception bias. More specifically, we aim to find out whether training subjects into perceiving changes in pitch direction according to changes in fundamental frequency or changes in the spectrum can affect their subsequent listening mode. In the second part of the study, we explore the link between the auditory perception bias and listeners' sensitivity to local
and global pitch changes, roughly mirroring local and global perception in the visual domain (Ziegler et al., 2012).

Simply put, global processing refers to the perception of a stimulus as a whole, whereas local processing corresponds to the perception of its parts. With respect to auditory stimuli, global processing corresponds to the perception of the pitch direction or contour, while local processing stands for the perception of the intervals between the notes comprising a sound (Bouvet et al, 2011; Justus \& List, 2005; Sanders \& Poeppel, 2007). Research in the visual domain has provided some support for stronger right hemisphere activation during global processing and stronger left hemisphere activation during local processing (e.g. Fink et al, 1996). So far, the link between auditory local and global processing and the auditory perception bias has not been explored experimentally.

\section*{Global vs. Local Precedence and its Correlates}

In the visual domain, processing at the global level usually takes precedence over processing at the local level, a tendency described as the Global Precedence Effect (GPE) (Navon, 1977). A similar pattern has been demonstrated in the auditory domain as well (Bouvet et al., 2011; List, Justus, Robertson \& Bentin, 2007). Contrary to this general effect, processing at the local level can also precede global processing when stimuli features are altered (e.g. Kimchi, 1992), or, even more importantly, in case of developmental differences. For instance, in the auditory domain, children with developmental dyslexia show a stronger tendency for local auditory processing (Ziegler, Pech-Georgel, George, \& Foxton, 2011); in the visual domain, individuals diagnosed with Autistic Spectrum Disorders, such as autistic children (Jollife \& Baron-Cohen, 2006) and women diagnosed with Anorexia Nervosa (Southgate et al, 2008) show a local processing bias as well. In the case of autism, Baron-Cohen (2002) describes the tendency for local processing as systemizing and differentiates it from empathizing, which reflects the ability to share others' mental and emotional states. Autistic children perform poorly in tasks requiring Theory of Mind (ToM) and show empathic deficits from a very early age (Baron-Cohen, 1995; Yirmiya, Sigman, Kasari, \& Mundy, 1997, a.o.). Impaired ToM is also itself associated with low empathy scores (Shamay-Tsoory et al., 2005). The above findings indicate that the presence of a local processing bias is, in autism at least, accompanied by the presence of impaired empathy. A more direct examination of the link between global-local visual processing and empathy in normal subjects has shown, on the contrary, a link between local processing and greater empathy (Woltin, Corneille, Yzerbyt, \& Förster, 2011). This last finding was attributed to the facilitating role that local processing plays in self-other awareness, a prerequisite for the experience of empathy (Decety \& Jackson, 2004).

In the present research, we also aim to examine the relationship between empathy and global-local auditory processing. In the auditory domain, personal distress, an affective component of empathy, has been associated with
the ability to perceive prosody (Aziz-Zadeh, Sheng, \& Gheytanchi, 2010). Prosody perception is impaired in children diagnosed with the Asperger syndrome (Korpilahti et al., 2007). Furthermore, autistic children have difficulties in inferring mental states from the other's voice (Rutherford, Baron-Cohen, \& Wheelwright, 2002). If we take into account the processing preferences of autistic individuals, impaired prosody perception and decreased empathy seem to accompany the local processing bias. Considering that similar processing types are exhibited across modalities, we might expect a local processing preference to be accompanied by impaired prosody perception and empathy in the auditory domain as well.

To strengthen the proposed relationship between globallocal processing and empathy, we will also examine the role of self-construal (Markus \& Kitiyama, 1991). Interdependent self-construal has been associated with global processing, whereas independent self-construal with local processing (Kühnen \& Oyserman, 2002; Lin et al, 2008, 2009). Moreover, interdependent self-construal is related to higher empathy (Cross et al., 2000). In addition to these two types of self-construal, we are also considering the relational-interdependent self-construal, a type of interdependence found in rather individualistic cultures (Cross et al., 2000). According to the above, we expect interdependent and relational-interdependent self-construal to be positively related to global auditory processing and empathy, while independent self-construal to be negatively related.

\section*{Current Study}

\section*{Experiment 1}

\section*{Participants}

Sixty-eight students ( 15 males and 54 females) from Tilburg University were recruited for an experimental session in exchange for course credit. Participants' age ranged from 17 to 27 years old (mean \(=22.2, \pm 2.6\) ). One participant reported non-normal hearing ability; the participant was not excluded from the analyses given that (s)he performed similarly to the rest of the participant group. The participants were randomly divided into the three betweenparticipant experimental conditions.

\section*{Stimuli and procedure}

A total of 72 pairs of complex harmonic tones consisting of two, three or four harmonics were constructed for the pitch discrimination task, following the procedure described in Laguitton et al. (1998), including the addition of noise in order to minimize the effects of combination tones (which arise at the cochlear level and may interfere with the measurements of individual differences on the neural level). Thirty-six tone pairs were ambiguous, meaning that the second tone sequence would be judged as higher vs. lower than the first one depending on the participant's listening mode. For 18 ambiguous tone pairs, the second sequence would be judged as lower-higher based on a fundamental
frequency listening mode. For the rest of 18 tone pairs, the second sequence would be judged as higher-lower based on a spectral listening mode. The remaining 36 tone pairs were unambiguous and were used as control stimuli. Each tone pair was 2000 ms long. All stimuli were displayed using EPrime (Psychology Software Tools, Inc., www.pstnet.com).

Training phase: During the training phase, participants were presented with 36 ambiguous tone pairs. They were instructed to listen to the tone pair and were asked to indicate whether they perceived the tone pair as rising or falling. After each response, they were provided with feedback about the tonal progression, aiming to train their listening mode. In the fundamental frequency mode condition, participants were told that the tone pair was rising (falling) according to rises (falls) of the fundamental frequency. In the spectral listening mode condition, the feedback depended on rises (falls) of the spectrum. In a control condition, no feedback was provided. The response key order was counterbalanced between the participants.

Testing phase: During the testing phase, participants were presented with 18 ambiguous and 18 non-ambiguous tone pairs. Similarly to the training phase task, they were instructed to indicate whether they perceived the tone pair as rising or falling, they were not provided with feedback about the tonal progression.

\section*{Measurements}

Based on the participants' answers, we calculated their individual 'Coefficient of Sound Perception Preference' \(\left(\partial_{p}\right)\) using the formula \(\partial_{p}=(\mathrm{F}-\mathrm{Sp}) /(\mathrm{F}+\mathrm{Sp})\), where F is the number of virtual fundamental classifications and Sp the number of spectral classifications in the testing phase. We calculated the 'Listener Attention Coefficient' \(\left(\partial_{A}\right)\) as the proportion of correctly categorized unambiguous stimuli.


Figure 1: Distribution of the Coefficient of Sound Perception Preference across the three experimental conditions.

\section*{Results}

A one-way analysis of variance showed no effect of training on the Coefficient of Sound Perception Preference in the testing phase. The participants in the fundamental frequency mode condition, the spectral mode condition and the control condition also did not differ with respect to their mean reaction times and correct responses to the non-ambiguous stimuli. The distribution of the \(\partial_{p}\) values across the three conditions is shown in Figure 1.

\section*{Discussion}

The results of the first experiment indicate that simple feedback is not sufficient to train participants in such a way that they focus either on the fundamental frequency in the signal or on its harmonic components. The results confirm the findings of Ladd et al. (2013) and others who found that auditory perception bias is robust in test-retest. Contrary to their study, we found a relatively normal distribution of listener types in our experimental group, compared to the prevalence of holistic (fundamental) listeners in their experiment. The difference is most likely due to the use of masking noise in our stimulus material which helped to exclude effects of combination tones (Plomp, 1976).

\section*{Experiment 2}

\section*{Participants}

Forty-nine students (7 males and 42 females) from Tilburg University, drawn from the same participant group as in Experiment 1, were recruited for an experimental session in exchange for course credit. Participants' age ranged from 18 to 27 years old (mean \(=22.5, \pm 1.8)\).

\section*{Stimuli and procedure}

Auditory global-local processing task
To measure global-local auditory processing, a total of 96 pairs of 4-tone sequences ( 48 same and 48 different) stimuli were used. The stimuli were constructed following the procedure suggested by Ziegler, Pech-Georgel, George, \& Foxton (2011). The sequences contained pure tones, each of 250 ms duration with 20 ms gating windows, with frequencies from an atonal scale taken from a division of an octave into seven equally spaced logarithmic steps. The starting frequencies were taken from the interval between 250 to 354 Hz . The third or fourth note in the second sequence was altered so that it was two steps lower or higher than the note in the first sequence (see Figure 2). In the local stimuli, the second sequence would remain rising/falling, in the global stimuli, the global melody would change. Each tone pair was 1000 ms in duration. All stimuli were displayed using E-Prime (Psychology Software Tools, Inc., www.pstnet.com) in a random order.

\section*{Auditory affective processing task}

To measure participants' performance in auditory affective processing, we used the Montreal Affective Voices stimuli (Belin, Fillion-Bilodeau, \& Gosselin, 2008). The corpus includes 90 vocal affect bursts (expressed as the vowel \(/ \mathrm{a} /\) ),
which express the emotions of anger, disgust, fear, happiness, pain, pleasure, sadness, surprise and a neutral expression. Participants heard each vocal expression once and were asked to select one of the above emotions.


Figure 2: Illustration of the local and global types of stimuli used in Experiment 2 (from Ziegler et al. (2011)).

\section*{Empathy measurement}

To measure empathy, we used the Interpersonal Reactivity Index, developed by Davis (1980). It measures individual differences in empathy and consists of four dimensions (perspective taking, fantasy scale, empathic concern and personal distress) each one tapping a different aspect of empathy. Participants were asked to indicate, on a five-point scale, to what extent each statement described themselves.

\section*{Self-construal measurement}

To assess the role of self-construal, we used the SelfConstrual Scale developed by Singelis (1994). The scale consists of 24 items which measure the interdependent and independent images of the self. We also included the relational-interdependent self-construal scale (Cross et al., 2000). The scale consists of 11 items. For both measures, participants were asked to indicate their agreement or disagreement on a seven-point scale.

\section*{Measures}

Following Ziegler et al. (2011), we used d' measures to calculate the participants' performance in the auditory global-local pitch processing task ( \(\mathrm{d}_{\mathrm{G}}\) and \(\mathrm{d}_{\mathrm{L}}\), respectively). Both measures were not normally distributed with \(M d_{G}=\). 427, \(M d_{L}=.312\). For the auditory affective processing task, we calculated the scores as the total number of correctly identified emotions (Aff). The mean score of correctly identified emotions (90 in total) was 61.6 ( \(S D=9.3, M d=\) 64); the distribution of answers was not normal with most participants performing above chance \((t(47)=36.56, p<\). 001). For the empathy measurement, the Cronbach's alpha coefficient was .73 ; the items were reduced to a single empathy score (Emp) for further calculations. For self-
construal, we constructed three subscales: relational selfconstrual (Cronbach's alpha (11) = .76), interdependent selfconstrual (Cronbach's alpha (12) \(=.47\) ) and independent self-construal \((\) Cronbach's alpha \((12)=.73)\).

Table 1: Nonparametric Spearman's correlations for measures collected in Experiment 1 and 2.
\begin{tabular}{lccccccccc} 
& \(\mathrm{d}^{\mathrm{G}}\) & \(\mathrm{d}^{\mathrm{L}}\) & \(\partial_{\mathrm{p}}\) & \(\partial_{\mathrm{A}}\) & Aff & Emp & Rel & Inter & Indep \\
\hline \(\mathrm{d}_{\mathrm{G}}\) & \(.45^{* *}\) & .10 & -.02 & \(.34^{*}\) & -.02 & .07 & .11 & .06 \\
\(\mathrm{~d}^{\prime} \mathrm{L}\) & & & -.02 & -.03 & .07 & -.07 & -.04 & .06 & -.06 \\
\(\partial_{\mathrm{p}}\) & & & .04 & .07 & .24 & .03 & -.02 & -.98 \\
\(\partial_{\mathrm{A}}\) & & & & -.15 & -.04 & -.15 & -.13 & -.19 \\
Aff & & & & & & -.04 & -.30 & .06 & .00 \\
Emp & & & & & & .10 & \(.39^{* *}\) & -.08
\end{tabular}

Note: \({ }^{*} p<.05,{ }^{* *} p<.001\)
\(d^{\prime}{ }_{G}=\) Global pitch processing \(\left(d^{\prime}\right), d^{\prime}{ }_{L}=\) Local pitch processing ( \(d^{\prime}\) ) , \(\partial_{p}=\) Coefficient of Sound Perception Preference, \(\partial_{A}=\) Listener Attention Coefficient, Aff \(=\) Affective Voices, Emp = Empathy Measurement, SC = SelfConstrual.

\section*{Results}

The Shapiro-Wilk test of normality showed a significant non-normal distribution for several of the measures, therefore, we used non-parametric tests throughout. In Table 1 , the results of nonparametric correlations for the measures of global and local pitch perception, emotion perception, affective empathy and self-construal are reported, including the Coefficient of Sound Perception Preference collected in the first experiment. The analysis shows a significant relation between global pitch perception processing and the auditory affective processing measure: participants who were better in perceiving changes in the global pitch contour were also better in identifying vocalized emotions.

\section*{Discussion}

The results of the second experiment indicate that global auditory processing is related to auditory affective processing. This suggests that being able to identify emotions in voice is associated with the ability to perceive pitch globally.

\section*{General discussion and Conclusion}

The present studies aimed to: a) investigate the possibility of altering individuals' auditory perception bias through training, b) to illustrate experimentally the existence of a relation between auditory global-local processing and auditory perception bias, and c) to examine the link between global-local auditory processing on one hand and empathy and self-construal on the other hand. Our findings show that auditory perception bias cannot be altered by simple training/feedback. This finding adds to the existing evidence
according to which the mode of listening (synthetic or analytic) constitutes a rather stable individual difference. With respect to its relation with auditory global-local processing, our findings cannot support an association between processing type and perception bias. We do find, though, an association between global auditory processing and auditory affective processing. To put it differently, perceiving the contour in sounds is related to the ability to recognize emotions in voice. No evidence is provided for the link of empathy with processing when using self-report measures. It is quite possible that, especially for perceived emotions, behavioral measures of emotional empathic responses may yield different results.

\section*{References}

Aziz-Zadeh, L., Sheng, T., \& Gheytanchi, A. (2010). Common premotor regions for the perception and production of prosody and correlations with empathy and prosodic ability. PLoS ONE, 5(1), e8759.
Babel, M. (2009). Phonetic and social selectivity in speech accommodation. Doctoral dissertation. University of California, Berkeley.
Baron-Cohen, S. (1995). Mindblindness: An essay on autism and theory of mind. MIT Press: Bradford Books.
Baron-Cohen, S. (2002). The extreme male brain theory of autism. Trends in Cognitive Sciences, 6(1), 248-254.
Belin, P., Fillion-Bilodeau, S., \& Gosselin, F. (2008). The Montreal affective voices: A validated set of nonverbal affect bursts for research on auditory affective processing. Behavior Research Methods, 40(2), 531-539.
Bouvet, L., Rousset, S., Valdois, S., \& Donnadieu, S. (2011). Global precedence effect in audition and vision: Evidence for similar cognitive styles across modalities. Acta Psychologica, 138, 329-335.
Cross, S. E., Bacon, P. L., \& Morris, M. L. (2000). The relational interdependent self-construal and relationships. Journal of Personality and Social Psychology, 78(4), 791808.

Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. JSAS Catalog of Selected Documents in Psychology, 10, 85.
Decety, J., \& Jackson, P. L. (2004). The functional architecture of human empathy. Behavioral and Cognitive Neuroscience Reviews, 3, 71-100.
Dediu, D., \& Ladd, D.R. (2007). Linguistic tone is related to the population frequency of the adaptive haplogroups of two brain size genes, ASPM and Microcephalin. PNAS, 104, 10944-10949.
Fink, G. R., Halligan, P. W., Marshall, J. C., Frith, C. D., Frackowiak, R. S., \& Dolan, R. J. (1996). Where in the brain does visual attention select the forest and the trees? Nature, 382(6592), 626-628.
Foxton, J.M., Talcott, J.B., Witton, C., Brace, H., McIntyre, F., \& Griffiths, T.D. (2003). Reading skills are related to global, but not local, acoustic pattern perception. Nature Neuroscience, 6, 343-344.

Förster, J., \& Higgins, E. T. (2005). How global versus local perception fits regulatory focus. Psychological Science, 16, 631-636.
Gregory, S.W. Jr. (1983). A quantitative analysis of temporal symmetry in microsocial relations. American Sociological Review, 48, 129-135.
Gregory, S.W. Jr., \& Hoyt, B.R. (1982). Conversation partner mutual adaptation as demonstrated by Fourier series analysis. Journal of Psycholinguistic Research, 11, 35-46.
Gregory, S.W. Jr., Webster, S.W., \& Huang, G. (1993). Voice pitch and amplitude convergence as a metric of quality in dyadic interviews. Language and Communication, 13, 195-217.
Gregory, S.W. Jr., \& Webster, S.W. (1996). A nonverbal signal in voices of interview partners effectively predicts communication accommodation and social status perceptions. Journal of Personality and Social Psychology, 70, 1231-1240.
Gregory, S.W. Jr., Dagan, K., \& Webster, S.W. (1997). Evaluating the relation of vocal accommodation in conversation partners' fundamental frequencies to perceptions of communication quality. Journal of Nonverbal Behaviour, 21, 23-43.
Gregory, S.W. Jr., \& Gallagher, T.J. (2002). Spectral analysis of candidates’ nonverbal vocal communication: Predicting U.S. presidential election outcomes. Social Psychological Quarterly, 65, 298-308.
Jolliffe, T., \& Baron-Cohen, S. (1997). Are people with autism and Asperger syndrome faster than normal on the embedded figures test? Journal of Child Psychology and Psychiatry, 38, 527-534.
Justus, T., \& List, L. (2005). Auditory attention to frequency and time: an analogy to visual local-global stimuli. Cognition, 98(1), 31-51.
Juslin, P.N., \& Laukka, P. (2003). Communication of emotions in vocal expression and music performance: d Different channels, same code? Psychological Bulletin, 129, 770-814.
Kimchi, R. (1992). Primacy of wholistic processing and global/local paradigm: A critical review. Psychological Bulletin, 112(1), 24-38.
Korpilahti, P., Jansson-Verkasalo, E., Mattila, M. L., Kuusikko, S., Suominen, K., Rytky, S., Pauls, D. L., Moilanen, I. (2007). Processing of affective speech prosody is impaired in Asperger syndrome. Journal of Autism and Developmental Disorders, 37(8), 1539-1549.
Kühnen, U., \& Oyserman, D., (2002). Thinking about the self influences thinking in general:Ccognitive consequences of salient self-concept. Journal of Experimental Social Psychology, 38, 492-499.
Laguitton, V., Demany, L, Semal, C., \& Liégeois-Chauvel, C. (1998). Pitch perception: A difference between rightand left-handed listeners. Neuropsychologia, 36, 201-207.

Macrae, C. N., \& Lewis, H. L. (2002). Do I know you? Processing orientation and face recognition. Psychological Science, 13, 194-196.
Markus, H. R., \& Kitayama, S. (1991). Culture and the self: implications for cognition, emotion, and motivation. Psychological Review, 98, 224-253.
Mueller, J.L., Friederici, A.D., \& Männel, C. (2012). Auditory perception at the root of language learning. PNAS 109, 15953-15958.
Ladd, D., Turnbull, R., Browne, C., Caldwell-Harris, C., Ganushchak, L., Swoboda, K., Woodfield, V., \& Dediu, D. (2013). Patterns of individual differences in the perception of missing-fundamental tones. Journal of Experimental Psychology: Human Perception and Performance, 39, in press.
Lin, Z., Lin, Y., \& Han, S. (2008). Self-construal priming modulates visual activity underlying global/local perception. Biological Psychology, 77, 93-97.
Lin, Z., \& Han, S. (2009). Self-construal priming modulates the scope of visual attention. The Quarterly Journal of Experimental Psychology, 62(4), 802-813.
List, A., Justus, T., Robertson, L. C., \& Bentin, S. (2007). A mismatch negativity study of local-global auditory processing. Brain Research, 1153, 122-133.
Moore, B.C.J. (2003). An Introduction to the Psychology of Hearing. Academic Press.
Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. Cognitive Psychology, 9, 353-383.
Postma-Nilsenová, M., \& Postma, E.O. (2012). Individual differences in F0 imitation: Causes and effects. Proceedings of Measuring Behavior 2012, Utrecht University.
Plomp, R. (1976). Aspects of Tone Sensation. New York: Academic Press.
Rutherford, M. D., Baron-Cohen, S., \& Wheelwright, S. (2002). Reading the mind in the voice: A study with normal adults and adults with Asperger syndrome and high functioning autism. Journal of Autism and Developmental Disorders, 32(3), 189-194.
Sanders, S. D., \& Poeppel, D. (2007). Local and global auditory processing: Behavioral and ERP evidence. Neuropsychologia, 45, 1172-1186.
Schneider, P., Sluming, V., Roberts, N., Scherg, M., Goebel, R., Specht, H. J., Günter Dosch, H., Stippich, C., \& Rupp, A. (2005). Structural and functional asymmetry of lateral Heschl's gyrus reflects pitch perception preference. Nature Neuroscience, 8(9), 1241-1247.

Scherer, K.R., Banse, R., Wallbott, H.G., \& Goldbeck, T. (1991). Vocal cues in emotion encoding and decoding. Motivation and Emotion, 15, 123-148.
Seither-Priesler, A., Krumbholz, K., Patterson, R., Johnson, L., Nobbe, A., Seither, S., \& Lütkenhöner, B. (2007). Tone sequences with conflicting fundamental pitch and timbre changes are heard differently by musicians and nonmusicians. Journal of Experimental Psychology: Human Perception and Performance, 33(3), 743-751.
Shamay-Tsoory, S. G., Tomer, R., Berger, B. D., Goldsher, D., \& Aharon-Peretz, J. (2005). Impaired "affective theory of mind" is associated with right ventromedial prefrontal damage. Cognitive and Behavioral Neurology, 18(1), 55-67.
Singelis, T. M. (1994). The measurement of independent and interdependent self-construals. Personality and Social Psychology Bulletin, 20, 580-591.
Southgate, L., Tchanturia, K., \& Treasure, J. (2008).
Information processing bias in anorexia nervosa. Psychiatry Research, 160(2), 221-227.
Woltin, K. A., Corneille, O., Yzerbyt, V., \& Förster, J. (2011). Narrowing down to open up for other people's concerns: Empathic concern can be enhanced by inducing detailed processing. Journal of Experimental Social Psychology, 47, 418-424.
Wong, P.C.M., \& Perrachione, T.K. (2007). Learning pitch patterns in lexical identification by native Englishspeaking adults. Applied Psycholinguistics, 28, 565-585.
Wong, P.C.M., Warrier, C.M., Penhune, V.B., Roy, A.K., Sadehh, A., Parrish, T.B., \& Zatorre, R.J. (2008). Volume of left Heschl's gyrus and linguistic pitch learning. Cerebral Cortex, 18, 828-836.
Wong, P.C.M., Chandrasekaran, B., \& Zheng, J. (2012). The derived allele of ASPM is associated with lexical tone perception. PloS One, 7, e34243.
Yirmiya, N., Kasari, C., Sigman, M., \& Mundy, P. (1989). Facial expressions of affect in autistic, mentally retarded and normal children. Journal of Child Development, 63(1), 150-160.
Ziegler, J. C.., Pech-Georgel, C., George, F., \& Foxton, J. M. (2012). Global and local pitch perception in children with developmental dyslexia. Brain \& Language, 120(3), 265-270.

\title{
People, Place, and Time: Inferences from Diagrams
}

\author{
Barbara Tversky (btversky@stanford.edu) \\ James E. Corter (jec34@columbia.edu) \\ Jie Gao (jg2902@columbia.edu) \\ Teachers College, Columbia University \\ New York, NY 10027 USA \\ Yuko Tanaka (yuko.tanaka@stevens.edu) \\ Jeffrey Nickerson (jnickerson@stevens.edu) \\ Stevens Institute of Technology \\ Hoboken, NJ 07030 USA
}

\begin{abstract}
Keeping track of things as they move in space and time is a task common to scientists, marketers, spies, coaches, and more. Visualizations of complex information aid drawing inferences and conclusions but there are manv wavs to represent data. Here we show that the kinds of inferences people draw depend on the kind of visualization, boxes in tables or lines in graphs. Lines link and boxes contain; they both direct attention and create meaning.
\end{abstract}

Keywords: diagrams; information visualization; inference; data displays.

\section*{Introduction}

People are always on the move. So are other living things, and even inanimate things, not just tangibles like packages, airplanes, and lava but also slang, fashion, music, rumors. Tracking and understanding movements of things in space and time is a task shared by scientists, historians, football coaches, paparazzi, marketers, physicians, spies, Facebook, event planners, Foursquare, police, culture mavens, advertisers, gossip columnists, friends, and more. The movements of beings and things through space over time are valuable data to be explained by theories. Why do people or things cluster in one place or avoid another? Why did person X see Y and then Z ? Why did they meet there? Why is this place popular at one time and not at another? Speculating about the movements of people or things over time is endlessly fascinating, and the number of queries, hypotheses, and explanations that can be generated enormous.

Making sense of complex data like the movements of things in space and time is made easier by organizing it spatially into diagrams. Diagrams are composed of simple geometric forms, dots, lines, boxes, and more that both carry meaning and direct attention (e.g., Tversky, 2011; Tversky, Zacks, Lee, \& Heiser, 2000). Lines direct attention by drawing the eye from place to place, point to point, connecting the dots. Lines create meaning by conveying relationships, connections from one place or point to another, as in route maps or networks or line graphs. Boxes also direct attention, by bringing the eye to the contents of the boxes. Boxes are containers, they enclose one set of elements and separate them from
elements in other boxes. Boxes create meaning by creating categories. They indicate that everything within the box is similar, sharing features, and different from everything outside the box. Lines and boxes, like other simple geometric marks, are replete with meaning. They alter conclusions, inferences, and interpretations. The same data, height of 8 and 10 year olds or height of women and men, are interpreted as trends when displayed as lines and as discrete comparisons when displayed as bars (Zacks \& Tversky, 1999). For example, when lines connected the height of men and women, some people said, "As you get more male, you get taller."

Lines and boxes should also bias data exploration and inferences from displays of people, place, and time. Previous research evaluated production, preference, and performance of displays of people, place, and time (Kessell \& Tversky, 2010). When asked to create ways to keep track of movements of people across space and time, most participants created matrices or tables; a minority connected people over time with lines. Preference by other participants followed the same pattern. Overall, matrices with people as cell entries and time and place in rows and columns respectively were most commonly produced and preferred. This format has good foundations. Place and time are fixed, immutable, but people can move from cell to cell. Performance was assessed by the time to verify many kinds of inferences from the data. Lines facilitated inferences about time, but all other kinds of inferences were faster from tables.

Displays of people, place, and time are frequently used for data exploration, to generate conclusions from the data and inferences about the underlying processes. Here, we investigate the roles of lines and boxes in the spontaneous generation of inferences from data displays. Because lines connect people over time, lines should bias conclusions and inferences about people, and secondarily about time. Boxes emphasize their contents, the confluence of people, place, and time, and should support a greater variety of conclusions and inferences.

\section*{Method}

\section*{Participants}

Eighty-one people, 39 of them men, participated through Amazon's Mechanical Turk website. Their ages ranged from 18 to 75 , with a mean of 30.9. Forty-six percent had a Bachelor's degree or higher, 39.4\% had some college education, \(12.2 \%\) went to high school, and 2.5\% did not specify education level. Most (93.4\%) were native English speakers.

\section*{Stimuli}

The stimuli (Figures 1 and 2) were taken from Kessell \& Tversky (2010). Both showed the locations of four students at four times of day with time horizontal, place vertical, and people as cell entries. For the boxes condition (Figure 1) people were color-coded dots. For the lines condition (Figure 2), people were coded as colored lines going from cell to cell. Note that both conditions have boxes, but in the box condition, they are filled with the individuals. In the line condition, the boxes are empty, in the background, acting as points that are connected by lines.


Figure 1: The box stimulus display


Figure 2: The line stimulus display

\section*{Procedure}

Participants were randomly assigned to the box or line condition. For both, the first screen, seen in Figure 3, showed an example of a data display, a bar graph, along
with several possible conclusions and inferences that could be drawn from the display.


Figure 3. Example used in the instructions.
The sample inferences given were: "The population in both California and New York grew from 1900 to 2000. In 1900, the population was greater in New York than in California. In 2000, the population was greater in California than in New York."
Then either the box (Figure 1) or line (Figure 2) diagram was presented and participants were directed: "Please study the following graph and use the space below to draw as many inferences as possible." After this task, participants were asked for demographic information.

\section*{Results}

Participants typically generated many inferences, often several in a single phrase, complicating the coding and the counting. Consequently, inferences and interpretations were coded and analyzed in two ways: the primary and secondary organizer used; and the number of different types of statements/inferences produced. Two people coded; in the few cases where they disagreed, they discussed the cases and came to agreement.

\section*{Primary and secondary organizers}

In order to capture the overall structure of the organization of the interpretations and to compare the organization produced for each diagrams, we coded the primary organizer and secondary organizer for each participant. The inferences could be organized by Time People, or Location. Here is an example with People as primary organizer and Time as secondary organizer:
"David went to the dorm in the morning, stayed at the dorm until noon, went to the library at the afternoon, and ended up at the bookstore at evening. Justin went to the dorm in the morning, the bookstore at noon, the gym in the afternoon and back to the bookstore at evening. Alex went to the library in the morning, the bookstore at noon, the gym at the afternoon and to the dorm at evening. Sammy went to the gym in the morning, to the bookstore and noon, to the dorm at the afternoon, and stayed at the dorm until the evening."

The results of this coding are shown in Figures 4 and 5. The distribution of primary organizers differs between the two conditions, \(\chi 2(2 ; \mathrm{n}=81)=5.815, \mathrm{p}=.043\); However,
the secondary organizers did not differ between conditions: \(\chi 2(2 ; \mathrm{n}=81)=2.489, \mathrm{p}=.288\).


Figure 4. Distribution of primary organizer by condition.


Figure 5. Distribution of secondary organizer by condition
For both box and line displays, the default primary organizer was People, followed by Location and then Time. However, people dominated far more for lines than for boxes. The dominant secondary organizer for lines was Time, whereas the dominant secondary organizer for boxes was Place.

\section*{Number of statements/inferences}

As the previous example illustrates, the inference statements were organized and structured. Careful examination of the protocols revealed that most statements could be categorized as follows.

Single statement: a statement that referred to a single cell of the matrix, i.e., one person, one time and one place. For example, "David is in the dorm in the morning."

Parallel: a set of related statements in the same format, that is, organized by the same features in the same way. Parallel statements contain many inferences, that is, they refer to information in many cells. For example, the following statement is counted as one parallel statement: "Justin went to the dorm in the morning, the bookstore at noon, the gym in the afternoon and back to the bookstore in the evening." Parallel statements invite repetition, and were often repeated.

Generality: any statement that involves more than one person, time, or place (but is not a parallel statement). For example, "The bookstore is the most consistently visited places for the guys" and "David and Sammy spent more time in the Dorms that the others." Generalities also include many inferences.

Leap: any interpretation that went beyond the information given. For examples, "David and Sammy are friends," "David is probably unfit," "Alex manages his time well and gets everything done," and "Since Justin does not return to the dorm in the evening I would infer that he is probably dating a student who works at the bookstore and spends evenings at her place."

Negation: a negative statement from information given in the diagrams. For examples, "David never goes to the gym," "No one goes to the bookstore in the morning," and "Students are not required to use the library or gym."

The mean numbers of statements in each category are given in Figures 6 (error bars indicate standard error).


Figure 6. Frequencies of kinds of inferences given to box and line displays.

Differences in the frequencies of the statement categories between the two displays were examined by a generalized linear model with a Poisson distribution for each dependent variable. The results ( \(\chi 2\) statistics and pvalues) are shown below in Table 1.

Table 1. Tests of differences between conditions
\begin{tabular}{|l|l|l|}
\hline Inference type & Wald \(\chi 2(1)\) & p -value \\
\hline Single & 17.828 & \(<.001\) \\
\hline Parallel & 14.608 & \(<.001\) \\
\hline Generality & 15.156 & \(<.001\) \\
\hline Leaps & 0.771 & .380 \\
\hline Negation & 20.86 & \(<.001\) \\
\hline Word Count & 30.413 & \(<.001\) \\
\hline
\end{tabular}

The analyses confirm that the box displays yielded more single statements, generalities, and negations than the line displays, and that the line displays yielded more parallel statements than the box displays. There was no significant difference in number of leaps.

Because of the diversity of statements, there is no sensible way to count and compare the total number of inferences drawn from each display. Some brief statements summarized many information cells, and other
statements conveyed none. However, the word count was higher for lines (mean=90) than boxes (mean=79); this may be due to the large quantity of parallel statements, which are relatively long, for lines.

\section*{Discussion}

Diagrams of complex information use marks and place on the page to convey information effectively (e.g., Tversky, 2011; Nickerson, Corter, Tversky, Rho, Zahner \& Yu, 2013). Such diagrams are meant to spur a wide range of conclusions, inferences, and hypotheses. Designers of displays are faced with many decisions for portraying the data, and those choices affect the kinds of inferences that viewers make. In particular, data points can be connected by lines or enclosed in boxes. Lines suggest relationships and links whereas boxes contain and suggest contrasts with other boxes.

Here, information about movements of people in place and time were organized with lines or boxes, corresponding to two common diagrammatic formats, line graphs and tables. Participants were asked to make as many inferences as they could from one of the displays. Overall, participants produced a large number of generalities that linked information that was separated in the data, showing that they did attempt to integrate the information. In general, People was the dominant organizer of inferences. As predicted, the two spatial organizations of data, lines and boxes, had dramatic effects on the kinds of inferences drawn from the data, movements of people in space and time. Lines connected people over space and time. Although People was the dominant organizer in both cases, People was far more dominant when lines connected each person's movements over time, and Time was the dominant secondary organizer. Lines also encouraged more parallel inferences, inferences with the same structure and format. These are sets of inferences structured in the same way: \(X\) went to \(A\) at time 1, to B at time 2, etc. With boxes, people dominated as first organizer, but Place rather than Time dominated as secondary organizer. Boxes also encouraged more statements about single features of the information, more generalities involving many features, more leaps that went far beyond the information given, and more negations, that is, statements about empty cells.

Displays of this information are used for exploration and understanding of the underlying phenomena driving the movements as well as conveying them to others. Visuospatial characteristics of information displays affect the kinds of inferences drawn from the information, factors like position in space, marks such as lines and boxes, and content of the dimensions. People, place, and time are three-dimensional data, and three-dimensional displays are famously difficult to comprehend, biased toward the variables on the axes (e. g., Carpenter \& Shah, 1998). Based on previous research (Kessell \& Tversky, 2010), we chose the consensus arrangement of the three variables, time on the Y axis, place on the X axis, and
people as cell entries. Time and space are fixed dimensions (place was not located dimensionally here, but commonly is, in maps). Only people are movable, perhaps the reason they were selected for the cell entries.
People was by far the most popular organizer for inferences. This is most likely due to that fact that people are agents, for the most part, they decide where to go and when. People is also preferred to Place or Time for organizing both episodic (e. g., Taylor \& Tversky, 1997) and autobiographical memory (e. g., Wagenaar, 1986). In both cases, organization of memory is multiple and flexible, but organization by People is privileged. Location and time, like people, can be good predictors of activities, but people are agentive, and for that and a variety of other reasons, are better and preferred as organizers of memory.

The display format, line graph or table, affected both quantity and quality of inferences. The different patterns of inferences suggest that tables and line graphs induce different strategies for exploring the data. Those presented with tables seemed to focus on the cells, producing more single statements that described single cells. They noticed when cells had many entries, producing relatively more generalities, such as the crowd at the bookstore at noon. They also noted empty cells, producing negations that observed the absence of people in the bookstore in the morning or the gym at night. By contrast, those presented with lines used the lines to explore the data, focusing on each person's movements in turn across cells. Lines led the eye and the mind from cell to cell; matrices led the eye and the mind to the cells.
Which is better? Like almost everything, it depends. If you are tracking parcels or thieves or spies or consumers or celebrities, then lines will focus you on the important information. On the other hand, if you're entertaining many hypotheses, then use tables. Just be aware that what you choose makes a difference.

\section*{Acknowledgments}

The authors gratefully acknowledge awards IIS0725223, HHC-0905417, IIS-0855995 and IIS-0968561 from the National Science Foundation, and the Stanford Regional Visualization and Analysis Center for partial support of the research or preparation of the manuscript.

\section*{References}

Carpenter, P.A., \& Shah, P. (1998). A model of the perceptual and conceptual processes in graph comprehension. Journal of Experimental Psychology: Applied, 4, 75-100.
Kessell, A.M. \& Tversky, B. (2010). Visualizing space, time, and agents: Production, performance, and preference. Cognitive Processing, 12, 43-52.
Nickerson, J. V., Corter, J. E., Tversky, B., Rho, Y-J., Zahner, D., \& Yu, L. (2013). Cognitive tools shape
thought: diagrams in design. Cognitive Processing, published online Feb. 15, 2013.
Taylor, H. A., \& Tversky, B. (1997). Indexing events in memory: Evidence for index preferences. Memory, 5, 509-542.
Tversky, B. (2011). Visualizations of thought. Topics in Cognitive Science, 3, 499-535.
Tversky, B, Zacks, J., Lee, P. U., \& Heiser, J. (2000). Lines, blobs, crosses, and arrows: Diagrammatic communication with schematic figures. In M. Anderson, P. Cheng, \& V. Haarslev (Eds.), Theory and application of diagrams. Pp. 221-230. Berlin: Springer.
Wagenaar, W.A. (1986). My memory: A study of autobiographical memory over six years. Cognitive Psychology, 18, 225-252.

Zacks J \& Tversky B. (1999) Bars and lines: a study of graphic communication. Memory and Cognition, 27, 1073-1079.

\title{
Do people understand irony from computers?
}

\author{
Akira Utsumi (utsumi@inf.uec.ac.jp) \\ Yu Watanabe (yuwata@utm.inf.uec.ac.jp) \\ Yusuke Wakayama (ywganjin@utm.inf.uec.ac.jp) \\ Department of Informatics, The University of Electro-Communications \\ 1-5-1, Chofugaoka, Chofushi, Tokyo 182-8585, Japan
}

\begin{abstract}
In this paper, we empirically investigate whether people understand irony from computers in order to test the recent argument for an egocentric tendency in irony comprehension. In the experiment, participants took a timed math test comprising 10 questions of 3 -digit by 2 -digit multiplication. After that, they received a feedback comment on their performance (including potentially ironic sentences) from either an intelligent evaluation system with an AI engine (AI condition), a nonintelligent automatic evaluation system (Auto condition), or a human judge connected via the network (Human condition). The result was that the participants in the AI and Auto conditions understood the comment as ironic as those in the Human condition, and the participants in the AI condition perceived more sarcasm than other participants. Because people know that computers cannot think just as humans do, these results can be regarded as evidence for the egocentric tendency in irony comprehension, indicating that participants understood irony egocentrically from their own perspective without taking into account the mental state of the ironic speaker. These findings are also consistent with the "media equation" theory, from which we can suggest implications for the media equation, anthropomorphism, and computer-mediated communication of irony.
\end{abstract}

Keywords: Irony; Egocentric interpretation; Theory of mind; Media equation; Computer-mediated communication; Anthropomorphism

\section*{Introduction}

Verbal irony is a kind of nonliteral language that implicitly conveys the opposite of the literal meaning. \({ }^{1}\) To interpret irony, people must infer the speaker's beliefs and intentions including not only the first-order belief that the speaker does not think that the utterance is literally true, but also the secondorder belief that the speaker thinks that the hearers do not think so. For example, imagine that you are rushing to a movie theater, where your friend is waiting for you so that you can see the movie together. When you arrive there about 30 minutes late and you miss the first part of the movie, your friend says to you, "You are always so punctual!" You can easily understand that this utterance is ironic, but you have to know beforehand many things about your friend's belief and intention. First of all, you must be sure your friend does not think that you are always punctual because you are late for the movie. Furthermore, in order to recognize the speaker's ironic intention, you must know that your friend thinks that you think this utterance is literally false, because your friend does not intend to convey his/her criticism using irony unless he/she is convinced that you can understand the utterance is literally false.

This property of irony leads to the widely accepted assumption that irony interpretation requires a "Theory of Mind" (henceforth, ToM), which refers to the ability to infer the mental states of others (e.g., Happé, 1993; Sperber \& Wilson, 2002).

A large number of recent empirical studies have demonstrated the validity of this assumption. For example, developmental studies have revealed that typically developing children below 5-years of age, who do not completely acquire ToM, cannot understand irony (e.g., Creusere, 2000; Filippova \& Astington, 2008; Pexman, 2008). It has also been found that people with pervasive developmental disorder (PDD) such as autism and Asperger syndrome have difficulty understanding irony, and this difficulty has been attributed to impaired ToM (Happé, 1993; Kaland et al., 2002; Wang, Lee, Sigman, \& Dapretto, 2006). Hence, irony has been used as a benchmark for testing PDD and discriminating PDD from attention deficit hyperactivity disorder (ADHD) (Adachi et al., 2004). Furthermore, recent neuroimaging studies have demonstrated that, as compared to literal sentences, ironic sentences elicited higher activation in the medial prefrontal cortex, which is known to play a central role in ToM (Rapp, Mutschler, \& Erb, 2012; Shibata, Toyomura, Itoh, \& Abe, 2010).

On the other hand, some opposing evidence has been reported suggesting that irony comprehension does not always require ToM; people can interpret irony without considering the speaker's beliefs and intentions (e.g., Akimoto, Miyazawa, \& Muramoto, 2012; Keysar, 1994). Keysar (1994) demonstrated that people perceive an utterance as ironic even when it is obvious to them that the speaker of the utterance does not know the discrepancy between an utterance and reality (and thus, the speaker has no ironic intention). Akimoto et al. (2012) also found that people perceive irony by first attributing their own belief to the speaker automatically and subsequently by adjusting it through an effortful ToM process. These findings suggest that irony can be interpreted automatically by the egocentric process, and when the egocentric interpretation should be checked for errors and time allows \({ }^{2}\), it is checked according to, or made consistent with, the speaker's belief by the allocentric ToM process. For example, in the case of "late arrival" example presented above, you recognize the utterance "You are always so punctual!" as ironic first by considering the discrepancy between the content of the utterance and your own belief that you arrived late and thus you are not punctual. If you have enough time, you may then consider what your friend really thinks in order to check or confirm your own egocentric interpretation. It must be noted that this egocentric view of irony is also supported by a theoretical study of irony, i.e., Utsumi's (2000) implicit display theory of irony. Note also that the egocentric interpretation is not specific to

\footnotetext{
\({ }^{1}\) This "folk" definition of irony has been recognized as problematic by irony researchers, but it is sufficient for the present purpose.
\({ }^{2}\) Indeed, Epley, Keysar, Boven, and Gilovich (2004) found that ironic interpretation was more egocentric in the time-limited circumstance than in the leisurely circumstance.
}
irony; recent empirical studies have revealed that communication in general proceeds in a relatively egocentric manner, with addressees routinely interpreting what speakers say from their own perspective (Birch \& Bloom, 2007; Keysar, 2007).

From these recent empirical findings of the egocentric nature of irony understanding, we also argue that people perceive irony in an egocentric way without resorting to the allocentric process of considering other's beliefs. In this paper, to obtain further evidence for this claim, we propose a different experimental methodology, namely an experiment with "irony generated by computers." The research question to be answered here is: Do people understand irony from computers? We know that computers cannot think just in the same way as humans do, and thus computers do not, or even cannot, intentionally say irony. Hence, it is highly reasonable to assume that people do not attempt to infer the "mental" states of computers, even when they see or hear a potentially ironic utterance generated by computers. It follows that, if irony interpretation essentially involves the allocentric process of inferring the mental state of the speaker, then people do not perceive irony in computer generated utterances. On the other hand, if irony interpretation does not always require the consideration of the speaker's mental states and is governed by the egocentric process, people may see the irony when they are given potentially ironic utterances by computers. In sum, by empirically examining whether people perceive irony in the computer-generated statements, we can obtain the evidence for or against the claim that irony is understood in an egocentric fashion without or before the allocentric ToM process. This is what this study aims to accomplish.

To examine people's reactions to irony from computers, we conducted a laboratory experiment. In this experiment, participants took a timed math test on computer comprising 10 questions of 3-digit by 2 -digit multiplication (e.g., \(768 \times 59\) ). They were instructed that the computer system not only provides a math test, but also (1) evaluates their overall performance on math calculation by an AI engine taking into account multiple information such as the test score, the time it took to calculate, and their behavioral data during calculation collected through Web cameras (AI condition); or (2) evaluates their overall performance on math calculation automatically from the test score and the time for calculation (Auto condition); or (3) displays their overall performance on math calculation evaluated by a human judge who observed their behavior during calculation through Web cameras (Human condition). After finishing the test, they received a highly positive comment on their performance from the computer. This positive comment can be ironically interpreted if participants could not get a satisfactory score.

As we mentioned above, our argument for the egocentric nature of irony interpretation predicts that people understand irony from computers just as they understand irony from humans. Therefore, we can predict that people's understanding of irony does not differ among these three conditions of this experiment. Specifically, supposing that the difference between the AI and Auto conditions may lie in the attributability of humanlike mental states to computers (e.g., people may be more likely or easier to attribute the mental state to intelligent computers with AI technology than non-intelligent computers), no difference between these two "computer irony" conditions
also suggests the egocentric nature of irony comprehension.
Our prediction can also be supported by the "media equation" theory (or "Computers Are Social Actors (CASA)" theory) for human-media interaction (Nass \& Yen, 2010; Reeves \& Nass, 1996). The media equation theory argues that people tend to unconsciously treat computers and other media (e.g., automobiles, cellphones, robots) as if they were real people. For example, people behave politely and cooperatively to computers, and attribute personality characteristics to computers. The media equation has been empirically supported by a number of studies demonstrating that the social rules and heuristics guiding human-human communication apply equally well to human-media interaction. Among these studies, Fogg and Nass's (1997) study on computers that flatter is most relevant to our study. They demonstrated that, when receiving a "flattery" feedback from a computer, people reported the same effects of flattery (e.g., more positive affect and evaluations on computers) as flattery from humans. Likewise, it is reasonable to assume that the media equation predicts that people understand irony from computer just as they understand irony from others, thus suggesting that the answer to the question "Do people understand irony from computers?" is yes. It must be noted here that our study can also be regarded as an empirical study on the media equation, and we can point out the relationship between the media equation and the egocentric communication, which will be discussed later in this paper.

\section*{Method}

\section*{Participants}

Fifty-three undergraduate and graduate students participated as volunteers. Note that the recruitment of participants continued until valid data were obtained from 45 participants (i.e., 15 participants for each of the three conditions).

\section*{Design}

This experiment had three conditions: AI, Auto, and Human conditions. These conditions were manipulated by the instruction given to the participants and the time it took to provide feedback to them (i.e., to display the truth of their answer for each math question, and to display a final comment on their performance), except for which the three conditions were identical.

\section*{Procedure}

The experiment was conducted using a computer system comprising a Windows PC, an LCD monitor, and two Web cameras. After arriving at the laboratory, participants seated in front of the computer system, and were given an explanation of the purpose of the experiment and an overall instruction of the task they had to perform. Specifically, participants were instructed that the purpose of the experiment was to test and evaluate a computer system in front of you that we were developing, and that this system not only would provide a math test, but also (1) would evaluate their overall performance on math calculation by an intelligent AI engine taking into account multiple information such as the test score, the time it took to calculate, and their behavioral data during calculation collected through Web cameras (AI condition); or (2) would evaluate their overall performance on math calculation automatically from the test score and the time for calculation (Auto
condition); or (3) would display their overall performance on math calculation evaluated by a human judge who observed their behavior during calculation through Web cameras (Human condition). Note that actually the computer system made no evaluation in all the conditions and no human judge evaluated the participants in the Human condition; the system simply displayed the same comment we prepared beforehand, independently of participants' performance. Note also that no humanlike agents were displayed on the monitor.

After the instruction, participants took a timed multiplication test comprising 10 questions of 3-digit by 2 -digit multiplication (e.g., \(768 \times 59\) ). Multiplication questions were randomly generated so that they did not differ in complexity. In order to make more errors and thus to be more likely to perceive irony, participants were requested to calculate as quickly as possible and complete each multiplication within 30 seconds. If 30 seconds passed since they started each question, they received a warning from the system. Multiplication questions appeared on the monitor one at a time and remained there until participants typed the answer. Participants calculated a given multiplication on paper and typed the answer. The truth of the answer was then presented on the monitor one seconds (in the AI and Auto conditions) or three to five seconds (in the Human condition) after the answer was typed.

After finishing the test, participants received from the system a highly positive comment on their performance, together with summary statistics including the number of correct and incorrect answers, the mean answering time, and the number of questions they took more than 30 seconds to answer. The comment was that "You have a perfect calculation performance. You were very careful not to make a mistake. The time you took to calculate is also excellently fast." In the comment, the first and last sentences were potentially ironic, if participants' performance is not satisfactory.

After receiving the comment from the computer system, participants were asked to answer a questionnaire. The questionnaire comprised six questions with 7 -point Likert scales. The questions that we use in the analysis were: "Do you perceive irony in the comment?" (irony rating; \(7=\) ironic, \(1=n o t\) at all ironic), "Do you perceive sarcasm in the comment?" (sarcasm rating; \(7=\) sarcastic, \(1=\) not at all sarcastic), and "Do you think the comment is intentional?" (intentionality rating; \(7=\) intentional, \(1=\) not at all intentional). Other questions were: "Does the comment literally include praise or criticism?", "Is your performance satisfactory?," and "Is the evaluation given by the system appropriate?" At the close of the experiment, participants were told the true purpose of the experiment and debriefed. None of the participants were suspicious of the true purpose of the experiment.

\section*{Result}

Whether the comment presented to the participants was understood as ironic greatly depends on their performance on the calculation test. Therefore, in the analysis, we did not use the data of eight participants who correctly answered all the multiplication questions within 30 seconds, because they were very unlikely to perceive irony in the comment. In other words, in order to collect the valid data of 15 participants per condition (and thus a total of 45 participants), we had to recruit 53 participants.


Figure 1: Mean irony and sarcasm ratings for the three conditions

In order to confirm that the likelihood of perceiving irony in the comment did not differ among three groups of participants, we analyzed the difference in the number of incorrect answers and answering time. The total numbers of incorrect answers were 31,20 , and 17 for the AI, Auto, and Human conditions, respectively; the difference did not reach the level of statistical significance, but was close to it, \(\chi^{2}(2, N=450)=5.65\), \(p=.059\). This result suggests that we must regress out the effect of incorrect answers in the following analysis. On the other hand, the mean answering times per question were 26.8, 26.1, and 26.8 seconds for the AI, Auto, and Human conditions, respectively; they did not significantly differ, \(F(2,42)=0.17\), \(p>80\).

\section*{Irony and Sarcasm Ratings}

First of all, we examined whether the mean irony and sarcasm ratings differ among the three conditions, as shown in Figure 1. Concerning irony ratings, the participants in the AI condition appeared to perceive the comment as more ironic than other participants. A one-way, between-participants ANOVA showed that the difference among the three conditions was marginally significant, \(F(2,42)=2.08, p=.08\). However, an ANCOVA with the number of incorrect answers as the covariate revealed that this difference was no longer significant, \(F(2,41)=1.55, p=.22\). This means that, when the number of incorrect answers was statistically controlled, the adjusted irony ratings did not differ among the three conditions; as predicted, people understood irony from computer just as they understood irony from humans.

The mean sarcasm rating was also higher in the AI condition than in the other two conditions, and this difference reached the level of statistical significance, \(F(2,42)=6.06\), \(p<.01\). An ANCOVA with the number of incorrect answers as the covariate also revealed that this main effect was reduced, but remained significant, \(F(2,41)=4.91, p<.05\). Pairwise comparisons ( \(p<.05\) ) confirmed that the participants in the AI condition perceived the comment as significantly more sarcastic than the participants in the other two conditions. In particular, it is surprising that the mean sarcasm ratings in the Auto and Human conditions were very low, suggesting that the participants in these conditions did not perceive sarcasm. Possible reasons of this result will be discussed in the section of discussion.

In addition, we analyzed the correlation between irony and sarcasm ratings. In general, "blame-by-praise" irony, such as ones used in this experiment, accompanies a sarcastic effect (Kreuz \& Glucksberg, 1989). Therefore, we can confirm from this analysis that participants' judgment on irony and sarcasm was not arbitrary, and consequently the above result was reliable. The correlations between irony and sarcasm ratings were .51 (AI condition), 83 (Auto condition), and .77 (Human condition), and they were all significant ( \(p<.05\) ). This result clearly indicates that the participants who interpreted the comment as ironic also perceived sarcasm in the comment, and thus the obtained result on irony and sarcasm ratings was reliable.

\section*{Intentionality Rating}

First of all, we analyzed the mean intentionality ratings for the three conditions. The mean intentionality ratings were \(4.00(\mathrm{SD}=1.73)\) for the AI condition, \(3.93(\mathrm{SD}=1.75)\) for the Auto condition, and \(3.40(S D=1.96)\) for the Human condition. They did not significantly differ both in the ANOVA analysis, \(F(2,42)=0.49\), and in the ANCOVA analysis with the number of incorrect answers as the covariate, \(F(2,41)=\) 0.58 . The participants perceived the same low degree of intentionality involved in the comment, regardless of whether the speaker was a computer or a human. This result suggests that people may interpret the comment egocentrically.

Next, we examined the correlations between the intentionality and irony ratings. If people interpret irony by allocentrically thinking about the speaker's beliefs and intentions, they recognize the intentionality of the speaker in an ironic comment. Meanwhile, if people interpret the comment literally, they do not (or do not have to) recognize the intentionality behind the literal comment because it simply states the fact. Therefore, the allocentric view of irony understanding predicts a positive correlation between the irony and intentionality ratings. On the other hand, if people interpret irony egocentrically, they do not have to recognize the intentionality of the speaker, and thus the egocentric view of irony understanding predicts no correlation between the irony and intentionality ratings. The correlations between irony and intentionality ratings were \(r=.11\) (AI condition), \(r=-.05\) (Auto condition), and \(r=.08\) (Human condition). All these correlations were not at all significant, thus providing additional evidence for the egocentric view of irony understanding.

\section*{Discussion}

\section*{Anthropomorphism and Egocentric Comprehension}

The design of the experiment in this paper is premised on the assumption that people consciously know that computers do not have minds and thus cannot think as humans do. However, many researchers criticize this assumption on the empirical grounds that people tend to attribute human characteristics, beliefs, intentions, or emotions to nonhuman agents and objects (e.g., Epley, Waytz, \& Cacioppo, 2007). This tendency is known as anthropomorphism. According to the anthropomorphic explanation, people do not think of computers as mindless, and as a result our finding that people understand irony from computers does not imply the egocentric view of irony comprehension. Against this criticism, we defend our position as follows. A number of researchers have discussed a
variety of anthropomorphic experiences, which can be classified into two types: a strong, mindful anthropomorphism and a weak, mindless anthropomorphism (Kim \& Sundar, 2012). Considering a number of existing empirical findings on anthropomorphism, we deny the possibility that the participants of our experiment anthropomorphized computers mindfully (i.e., in a strong, mindful sense). A weak, mindless anthropomorphism might occur, but it implies that people's reasoning about the "mental states" of computers is quite egocentric.

Mindful anthropomorphism refers to the tendency to infer the mental states of nonhuman agents or objects from an allocentric perspective. For example, some pet owners perceive and speak of pets as being thoughtful and considerate. This anthropomorphic process is often conscious and seems to require ToM. Recent research demonstrates that whether people mindfully anthropomorphize nonhuman agents and objects depends on two properties, i.e., agency (the capacity to plan and act) and experience (the capacity to sense and feel) (Gray, Gray, \& Wegner, 2007). Indeed, Krach et al. (2008) demonstrated through a fMRI experiment that activation of the brain regions (i.e., the medial frontal cortex and the right temporo-parietal junction) known to be associated with ToM was correlated with the degree of agency or humanlikeness (i.e., a human partner, a highly humanlike robot, a functional robot, and a non-humanlike computer, in descending order of agency). The computer system used in our experiment had no humanlike appearance and displayed no humanlike characters, and thus is very low in both agency and experience. Hence, we can safely say that the participants of our experiment did not anthropomorphize the computer system mindfully; this indicates that our assumption that people think of computers as mindless holds true for the experiment.

Mindless anthropomorphism refers to the tendency to automatically attribute human mental states to nonhuman agents or objects without much consideration of whether nonhuman targets have mental states (Kim \& Sundar, 2012). Many of the studies on anthropomorphism have used the term "anthropomorphism" to refer to this mindless version. For example, Epley et al. (2007) state, "Using one's own mental states and characteristics as a guide when reasoning about other humans is egocentrism. Using one's own mental states and characteristics as a guide when reasoning about nonhuman agents is anthropomorphism (ibid., p.868)." Their notion of anthropomorphism clearly indicates that mindless anthropomorphism is egocentric. Hence, even if the participants of our experiment mindlessly anthropomorphized computers during the experiment, they did not directly infer the "mental states" of computers. It follows that the finding that they perceived irony from computers implies that they did so egocentrically, as we argue in this paper.

It must be noted that the media equation is consistent with mindless anthropomorphism; in other words, the media equation is primarily due to the egocentric nature of communication. Nass and Moon (2000) have argued that the notion of mindlessness provides a robust explanation for the media equation. As have been observed in a variety of social situations, people mindlessly apply social rules and expectations to computers. This phenomenon completely coincides with mindless anthropomorphism. Although they reject an anthropomorphic explanation of the media equation, but the
anthropomorphism they rejected is the mindful version of anthropomorphism. Egocentric, mindless anthropomorphism is a main cause of mindless behavior observed in a variety of media equation studies.

\section*{Egocentric and Allocentric Comprehension of Irony}

This paper has provided empirical evidence for the egocentric view of irony processing in a novel approach of using computers as ironists. People understand irony from their own perspective, and this leads to the obtained result that people understand irony from computers in the same way as they understand irony from humans. However, the egocentric view seems to be inconsistent with the strong relationship between the ability to understand irony and the ToM ability, which has been justified by a large number of empirical studies. Of course, the egocentric view does not imply that ToM (and the allocentric process) is unnecessary for irony processing, but the argument that people can understand irony without considering the mental states of others from the speaker's perspective seems to be inconsistent with the empirical findings that people with a ToM deficit cannot understand irony. How can the egocentric view explain these incompatible findings?

One possible explanation would be given from a developmental perspective; ToM is a prerequisite for acquiring the concept of irony (i.e., what is irony), but once people (i.e., children) know what is irony, they increasingly do not take into account the mental state of the speaker. The concept of irony essentially involves the speaker's intention of being ironic, which is motivated by a certain situational setting (which is referred to as ironic environment by Utsumi's (2000) implicit display theory) where the speaker's expectation has not been fulfilled and the speaker has a negative attitude toward it. Therefore, to acquire the concept of irony, children must be able to infer the mental state of the speaker. In general, children below 5years of age cannot distinguish between what they know and what others know, and behave egocentrically as if their own beliefs are shared by others, from which it naturally follows that they cannot be aware of irony. Typically developing children at around 5 years of age can increasingly distinguish what others know from what they know, and they come to understand some aspects of irony. Children's appreciation of irony continues to develop into adolescence. As demonstrated by a number of developmental studies on irony, in this development period children's performance on irony understanding is correlated with their (allocentric) ToM ability, because they are acquiring the concept of irony with the help of their developing ToM ability. Adults, who completely acquired the concept of irony and have experienced a number of ironic communication, develop the egocentric tendency again, and increasingly do not take into account the mental state of the speaker (Keysar, 2007), mainly to shortcut the burdensome process of allocentric comprehension.

\section*{Irony in Computer-Mediated Communication}

The experiment of this paper was conducted through computer-mediated communication. One may argue that our findings are specific to computer-mediated communication and should not be generalized to irony understanding in ordinary face-to-face communication. In other words, it may be pointed out that our finding is an artifact of computermediation communication and allocentric comprehension is
always required in face-to-face communication. We basically reject this possibility, but at the same time point out that there may be some truth in it.

Some empirical evidence against this possibility was obtained. Hancock (2004) found that comprehension of irony did not differ between computer-mediated communication and face-to-face communication. More important is their finding that a misunderstanding rate of irony did not differ between these two communication modes, and it was equal to the estimate (i.e., approximately 5\%) given by Gibbs (2000). Considering Keysar's (2007) argument that egocentric understanding can provide a systematic reason for misunderstanding, this finding may suggest that people understand irony in computer-mediated conversation as egocentrically as in face-to-face conversation; this clearly rejects the specificity of computer-mediation communication, thus indicating that the finding of this paper is not an artifact of computer-mediated communication. Note also that, in almost all empirical studies of irony, participants of the experiment were asked to understand irony from the addressee's perspective, but they were not the addressees of irony. On the other hand, our participants were literally the addresses of irony, and thus we may safely say that the experiment of this paper was conducted in a more realistic setting, which is more similar to face-to-face communication.

At the same time, some positive arguments for the specificity of computer-mediated communication can be pointed out concerning irony production. Hancock (2004) also revealed that speakers in computer-mediated conversation produced more irony than face-to-face speakers. Furthermore, it is pointed out that sentences created via social media such as blogs and Twitter include more irony, which motivates recent NLP studies on automatic recognition of irony (e.g., Reyes, Rosso, \& Veale, 2013). These findings are concerned with the production of irony, but appear to suggest that irony comprehension in computer-mediated conversation may be somewhat different. For example, a younger generation, who is familiar with blogs and Twitter, may be likely to interpret utterances in computer-mediated conversation as ironic. In addition, micro-bloggers in Twitter often use the hashtag \#irony or \#sarcasm to clearly indicate their ironic intention. It is a very characteristic property of irony in microblogs, because in general ironic intention should not be explicitly expressed so that irony does not lose its effect. This specific property of Twitter may possibly induce the younger generation (including the participants of our experiment) to understand irony more egocentrically in computer-mediated communication. This may be a potential reason for the result that the mean sarcasm rating was higher in the AI condition, assuming that the perceived agency of the AI condition is most similar to that of micro-bloggers. On the other hand, the participants in the Auto condition might perceive little agency. The participants in the Human condition might be aware of more humanity in an imaginary human judge than in micro-bloggers because they were told that the human judge was observing their behavior throughout the math test.

\section*{Concluding Remarks}

The experiment reported in this paper demonstrated that people perceived the comment as ironic regardless of whether the
speaker of irony is a computer or a human. The experiment was based on the "Computers Are Social Actors" or "CASA" paradigm, which is a novel approach for the study of irony. The obtained finding provided empirical evidence in favor of the egocentric tendency in irony comprehension, because if people understand irony by routinely considering the mental state of the speaker, they could not perceive irony when a computer is the speaker of irony. Through the study of this paper, we have also discussed some features of irony comprehension in computer-mediated communication and the relationship among egocentric communication, anthropomorphism, and media equation. It is worth pursuing these issues for further research.

\section*{Acknowledgments}

This study was supported in part by a Grant-in-Aid for Scientific Research B (No.23300098) from the Japan Society for the Promotion of Science.

\section*{References}

Adachi, T., Koeda, T., Hirabayashi, S., Maeokae, Y., Shiota, M., \& Wada, E. W. A. (2004). The metaphor and sarcasm scenario test: A new instrument to help differentiate high functioning pervasive developmental disorder from attention deficit/hyperactivity disorder. Brain \& Development, 26,301-306.
Akimoto, Y., Miyazawa, S., \& Muramoto, T. (2012). Comprehension process of verbal irony: The effects of salience, egocentric context, and allocentric theory of mind. Metaphor and Symbol, 27, 217-242.
Birch, S., \& Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. Psychological Science, 18, 382386.

Creusere, M. (2000). A developmental test of theoretical perspectives on the understanding of verbal irony: Children's recognition of allusion and pragmatic insincerity. Metaphor and Symbol, 15(1\&2), 29-45.
Epley, N., Keysar, B., Boven, L., \& Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. Journal of Personality and Social Psychology, 87(3), 327-339.
Epley, N., Waytz, A., \& Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. Psychological Review, 114(4), 864-886.
Filippova, E., \& Astington, J. (2008). Further development in social reasoning revealed in discourse irony understanding. Child Development, 79(1), 126-138.
Fogg, B., \& Nass, C. (1997). Silicon sycophants: The effects of computers that flatter. International Journal of HumanComputer Studies, 46, 551-561.
Gibbs, R. (2000). Irony in talk among friends. Metaphor and Symbol, 15(1\&2), 5-27.
Gray, H. M., Gray, K., \& Wegner, D. M. (2007). Dimensions of mind perception. Science, 315, 619.
Hancock, J. (2004). Verbal irony use in face-to-face and computer-mediated communications. Journal of Language and Social Psychology, 23(4), 447-463.

Happé, F. (1993). Communicative competence and theory of mind in autism: A test of relevance theory. Cognition, 48, 101-119.
Kaland, N., Moller-Nielsen, A., Callesen, K., Mortensen, E., Gottlieb, D., \& Smith, L. (2002). A new 'advanced' test of theory of mind: Evidence from children and adolescents with Asperger syndrome. Journal of Child Psychology and Psychiatry, 43(4), 517-528.
Keysar, B. (1994). The illusory transparency of intention: Linguistic perspective taking in text. Cognitive Psychology, 26, 165-208.
Keysar, B. (2007). Communication and miscommunication: The role of egocentric processes. Intercultural Pragmatics, 4, 71-84.
Kim, Y., \& Sundar, S. S. (2012). Anthropomorphism of computers: Is it mindful or mindless? Computers in Human Behavior, 28, 241-250.
Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., \& Kircher, T. (2008). Can machines think? Interaction and perspective taking with robots investigated via fMRI. PLoS ONE, 3(7), e2597.
Kreuz, R., \& Glucksberg, S. (1989). How to be sarcastic: The echoic reminder theory of verbal irony. Journal of Experimental Psychology: General, 118(4), 374-386.
Nass, C., \& Moon, Y. (2000). Machines and mindlessness: Social responses to computers. Journal of Social Issues, 56(1), 81-103.
Nass, C., \& Yen, C. (2010). The man who lied to his laptop: What machines teach us about human relationships. Penguin/Current.
Pexman, P. (2008). It's fascinating research: The cognition of verbal irony. Current Directions in Psychological Science, 17(4), 286-290.
Rapp, A., Mutschler, D., \& Erb, M. (2012). Where in the brain is nonliteral language? a coordinate-based meta-analysis of functional magnetic resonance imaging studies. NeuroImage, 63, 600-610.
Reeves, B., \& Nass, C. (1996). The media equation: How people treat computers, television, and new media like real people and places. Cambridge University Press.
Reyes, A., Rosso, P., \& Veale, T. (2013). A multidimensional approach for detecting irony in Twitter. Language Resources and Evaluation, 47(1), 239-268.
Shibata, M., Toyomura, A., Itoh, H., \& Abe, J. (2010). Neural substrates of irony comprehension: A functional MRI study. Brain Research, 1308, 114-123.
Sperber, D., \& Wilson, D. (2002). Pragmatics, modularity and mindreading. Mind \& Language, 17(1/2), 3-23.
Utsumi, A. (2000). Verbal irony as implicit display of ironic environment: Distinguishing ironic utterances from nonirony. Journal of Pragmatics, 32(12), 1777-1806.
Wang, A., Lee, S., Sigman, M., \& Dapretto, M. (2006). Neural basis of irony comprehension in children with autism: The role of prosody and context. Brain, 129, 932-943.

\title{
Learning Social Affordances and Using Them for Planning
}

\author{
Kadir Firat Uyanik, Yigit Caliskan, Asil Kaan Bozcuoglu, Onur Yuruten, Sinan Kalkan, Erol Sahin \\ \{kadir, yigit, asil, oyuruten, skalkan, erol\} @ceng.metu.edu.tr \\ KOVAN Research Lab, Dept. of Computer Eng., METU, Ankara, Turkey
}

\begin{abstract}
This study extends the learning and use of affordances on robots on two fronts. First, we use the very same affordance learning framework that was used for learning the affordances of inanimate things to learn social affordances, that is affordances whose existence requires the presence of humans. Second, we use the learned affordances for making multi-step plans. Specifically, an iCub humanoid platform is equipped with a perceptual system to sense objects placed on a table, as well as the presence and state of humans in the environment, and a behavioral repertoire that consisted of simple object manipulations as well as voice behaviors that are uttered simple verbs. After interacting with objects and humans, the robot learns a set of affordances with which it can make multi-step plans towards achieving a demonstrated goal.
\end{abstract}

\section*{Introduction}

Motor competences of robots operating in our environments, is likely to remain inferior to ours on most fronts in the near future. In order to complete tasks that require competences beyond their abilities, the robots will need need to interact with humans in the environment towards compensating these deficiencies. The inspiration for our study comes from babies and small children who can compensate the lack of their physical competences through the use of adults via social interaction. For instance, for a child, the reachability of a candy on a high shelf becomes possible only in the presence of an adult, as long as he can "manipulate" him properly using his social behaviors.

In this paper, we extend an affordance framework proposed for robots towards learning interactions with inanimate objects, to learning interactions with humans. The notion of affordances, proposed by Gibson (Gibson, 1986), emphasized the interaction between the agent and the environment, as opposed to the agent or the environment only, and provided a unifying frameworks for the study.

\section*{Contribution}

This study extends the learning and use of affordances on robots on two fronts. First, we use the very same affordance learning framework that was used for learning the affordances of inanimate things to learn social affordances \({ }^{1}\) (viz. affordances whose existence requires the presence of humans). Second, we use learned affordances to make multi-step plans.

In our earlier studies, we had proposed a framework that allowed the robot to learn affordances such as traversability of an environment (Ugur \& Şahin, 2010) or graspability

\footnotetext{
\({ }^{1}\) We would like to note that the term, social affordances has been used in different contexts, e.g., for the possibilities emerging from social networks (Wellman et al., 2003), or the affordances of an environment and properties of people that facilitate social interaction in a group of people (Kreijns \& Kirschner, 2001).
}
(Ugur, Şahin, \& Oztop, 2009), liftability of objects (Dag, Atil, Kalkan, \& Sahin, 2010) and showed how one can make multistep plans using the learned affordances.

In this paper, we argue that robots can use the very same framework to learn what a human may afford. Moreover, we enhance our prior study on multi-step planning (Ugur et al., 2009) via a new form of prototypes for effect representation.

Specifically, we equipped the humanoid robot iCub with a perceptual system to sense tabletop objects, as well as the presence and state of humans in the environment, and a behavioral repertoire that consisted of simple object manipulations and voice behaviors that uttered simple verbs. After interacting with objects and humans, we show that the robot is able to learn a set of affordances with which it can make multi-step plans towards achieving a demonstrated goal.

\section*{Related Work}

The notion of affordances provides a perspective that puts the focus on the interaction (rather than the agent or the environment) and was formalized as a relation \(a\) between an entity or environment \(e\), a behavior \(b\) and the effect \(f\) of behavior \(b\) on \(e\) (Şahin, Çakmak, Doğar, Uğur, \& Üçoluk, 2007; Montesano, Lopes, Bernardino, \& Santos-Victor, 2008):
\[
\begin{equation*}
a=(e, b, f) \tag{1}
\end{equation*}
\]

For example, a behavior \(b_{\text {lift }}\) that produces an effect \(f_{\text {lifted }}\) on an object \(e_{\text {cup }}\) forms an affordance relation ( \(\left.e_{\text {cup }}, b_{\text {lift }}, f_{\text {lifted }}\right)\). Note that an agent would require more of such relations on different objects and behaviors to learn more general affordance relations.

The studies on learning and use of affordances have mostly been confined to inanimate things, such as objects (Fitzpatrick, Metta, Natale, Rao, \& Sandini, 2003; Detry, Kraft, Buch, Kruger, \& Piater, 2010; Atil, Dag, Kalkan, \& Şahin, 2010; Dag et al., 2010) and tools (Sinapov \& Stoytchev, 2008; Stoytchev, 2008) that the robot can interact with. In these studies, the robot interacts with the environment through a set of actions, and learns to perceptually detect and actualize them. Moreover, with the exception of few studies (Ugur et al., 2009; Williams \& Breazeal, 2012), the robots were only able to perceive the immediate affordances which can be actualized with a single-step action plan.

Formalizations, such as 1 , are proved to be practical with successful applications in navigation (Ugur \& Şahin, 2010), and manipulation (Fitzpatrick et al., 2003; Detry et al., 2010; Ugur et al., 2009; Ugur, Oztop, \& Şahin, 2011), conceptualization and language (Atil et al., 2010; Dag et al., 2010; Yürüten et al., 2012), and vision (Dag et al., 2010). However,
in these studies, the environment is limited to objects only, excluding the possible diversities or use-cases that might arise due to the existence of humans in addition to the objects.

Human-assistance has been incorporated in (Montesano et al., 2008; Dag et al., 2010; Ugur, Oztop, \& Şahin, 2011) using the same affordance formalization (1) to learn object affordances by imitation and emulation. However, the role of the human is limited to teaching affordances as part of the training phase, and he is out of the loop during execution of actions for possible assistance in creating a certain effect in the environment to extend robot's motor capacities.

Robot's motor capacities are extended by learning the affordances of tools in (Sinapov \& Stoytchev, 2008; Stoytchev, 2008). However, these studies are focused on learning affordances of tools while the objects are kept fixed, hence the affordances of objects themselves couldn't be captured.

In most of the HRI or social robotics studies, the robots are intended to collaborate with their human partners and they are "active learners" that learn from their partners the correct way to execute and sequence actions for achieving a goal (see, e.g., (Fong, Thorpe, \& Baur, 2003; Breazeal, 2004; Weber, 2008; Cakmak, DePalma, Arriaga, \& Thomaz, 2010) for a review). This way, one can teach a robot to learn complicated sequences of actions (e.g., dancing with a human partner (Kosuge \& Hirata, 2004)) using several mechanisms like scaffolding (Ugur, Celikkanat, Sahin, Nagai, \& Oztop, 2011; Saunders, Nehaniv, Dautenhahn, \& Alissandrakis, 2007) or demonstration (Pastor, Hoffmann, Asfour, \& Schaal, 2009; Argall, Chernova, Veloso, \& Browning, 2009; Akgun, Cakmak, Jiang, \& Thomaz, 2012). Similarly, affordances are also utilized in planning (Ugur, Oztop, \& Şahin, 2011) over action possibilities, but human is not a part of the plan. However, in (Williams \& Breazeal, 2012), humans are important part of the plan, yet their participation is limited with the experiment scenario, and participants are priorly acknowledged about the type of assistance they are going to provide to the robot.

For a similar goal, affordances (called "interpersonal affordances") that emerge from coordinated joint actions of two robots are investigated (Richardson, Marsh, \& Baron, 2007; Marsh, Richardson, \& Schmidt, 2009); e.g., two robots learning the interpersonal affordance of lifting a table which, otherwise, is liftable by neither of them. Our approach differs from these studies in the sense that the human is seen as part of the environment (with no special status) and uses the very same framework to learn social affordances as the physical affordances of objects.

\section*{Research Platform}

\section*{Perception and Environment Representation}
iCub perceives its environment through two Kinect cameras ( K 1 and K2). K1 is used to extract table and tabletop objects. K2 -accompanied with a motion capture system (Visualeyez II VZ4000)- is used to detect human's body posture and gaze direction. For gaze direction detection, participants are provided with a hat with active LEDs on top. Overall interaction


Figure 1: Visualization of the interaction environment. (a) Robot is on the left, and the human is on the right. (b) Object related features. From top left to bottom right: raw RGBD point cloud of an object on a table, table and extracted tabletop object with its oriented bounding box and id, surface normals, min curvatures, max curvatures, shape indices. [Figure best viewed in color]
environment is represented as a feature vector containing the following features:
Surface features are surface normals (azimuth and zenith), principal curvatures (min and max), and shape indices. They are represented as a 20 -bin histogram in addition to the min, max, mean, standard deviation and variance information.

Spatial features are bounding box pose ( \(x, y, z\), theta), bounding box dimensions ( \(x, y, z\) ), and object presence.
Social features are human torso pose ( \(x, y, z\), roll, pitch, yaw), human gaze direction (roll, pitch, yaw), and human presence, all with respect to robot's own coordinate system shown as coordinate axis in Fig. 1a.


Figure 2: The interaction objects. From left to right; balls, boxes, cylinders, mugs, and irregular objects.

\section*{Behaviors and Effect Representation}

The robot is equipped with six behaviors (push-left, pushright, push-forward, pull, top-grasp, side-grasp) and some voice behaviors ("pass me", "hello", "come", "sit down", "stand up", "bye", "push right", "push left", "take").

Effects -in their raw form- are computed as the difference between the final and the initial state of the environment (viz. difference between the feature vectors representing the environment before and after the behavior performance).

Effects are supervisedly matched to an effect category chosen from a set of effects such as grasped, knocked, no-
change \({ }^{2}\), disappeared, moved right, moved left, moved forward, pulled, sat down, stood up, got attention, got closer.

Effect categories are compactly represented as a vector of " 0 ", " + ", "-", and "*" to represent changes in certain feature value as unchanged (mean close to zero, small variance); consistently increased (positive mean, small variance), consistently decreased (negative mean, small variance); and inconsistently changed (large variance), respectively. This prototype-based effect representation is claimed to correspond to verb concepts in our earlier studies (Atil et al., 2010). For extracting the prototypes for each effect cluster, we analyze the mean and variance values for each element of the features in the cluster. Specifically, we apply unsupervised clustering (RGNG, (Qin \& Suganthan, 2004)) on the meanvariance space. RGNG finds four clusters naturally formed. From the obtained effect consistencies, we determine the prototype of each effect cluster.

\section*{Methodology}

\section*{Data Collection}

We used 35 objects (Fig. 2) that are chosen to be in different colors, and shape complexities (from primitive cubes, spheres, cylinders to mugs, wine glasses, coke cans etc.), easily identified as "cylinder", "ball", "cup", "box", while some of them had irregular shapes to show generalization ability of the system.

We had iCub interact with objects and with humans by using all of the behaviors precoded in its behavior repertoire. In order to collect social interaction data, we have worked with 10 participants of different genders ( 4 female, 6 male), ages (20-40) and professions (4 undergrad, 2 grad students, 4 researchers with non-CS background). They were asked to respond naturally to a random sequence of voice behaviors enacted by iCub. Some of the voice behaviors were accompanied by simple movements (nodding head, waving arm, etc.).

\section*{Affordance Learning}

We collected 413 triplets of \((e, b, f)\) (Eqn. 1) for object interactions and 150 triplets for human interactions, and used them to train a Support Vector Machine (SVM) classifier for each behavior to predict the effect label given an entity. During training, we normalized the feature vectors with respect to the range of values each feature can take.

\section*{Planning with Forward Chaining}

Since we trained SVMs for predicting the effect of each behavior on an object, iCub can do forward chaining to find the set of behaviors leading to a goal state. Since the effect labels are represented by effect prototypes, the similarity between the goal state (which is an effect instance) and the predicted

\footnotetext{
\({ }^{2}\) The no-change label denotes that the applied behavior did not produce any notable change on the object. For example, when iCub asks a participant to stand up who was already standing, the participant would not change his position. This yields negligibly small changes in the feature vector.
}
effect prototype is needed and we use the Mahalanobis distance, which is calculated by taking the mean \(\mu_{E_{i}}\) of first effect cluster \(E_{i}\) (if the first \(E_{i}\) is an effect instance, we take the effect instance as \(\mu_{E_{i}}\) ) and using the second effect cluster's \(E_{j}\) mean \(\mu_{E_{j}}\) and variance \(\sigma_{E_{j}}\) :
\[
\begin{equation*}
d\left(E_{i}, E_{j}\right)=\sqrt{\left(\mu_{E_{i}}-f_{\text {proto }, E_{j}}^{+,-, 0}\right)^{T} S_{j}^{-1}\left(\mu_{E_{i}}-f_{\text {proto }, E_{j}}^{+,-, 0}\right)} \tag{2}
\end{equation*}
\]
where \(S_{j}\) is the covariance matrix of the second effect cluster \(E_{j}\). In computing the Mahalanobis distance, the features marked inconsistent in the prototype are disregarded (denoted by \(f_{\text {proto }, E_{i}}^{+,-, 0}\) for the effect prototype \(f_{\text {proto } E_{i}}\) of an effect cluster \(E_{i}\) ), as those correspond to an unpredictable/inconsistent change in the feature elements.

Finding the effects Planning toward achieving the goal is found using a breadth-first tree search. Starting with the initial state, we construct a tree such that it contains all the possible effect sequences with length \(n\) (empirically chosen as 3 ). The plan is made as the goal is matched with the predicted states after applying a sequence of behaviors.

In the first step of future state calculation (Fig. 3), the current state of the object is fed to the trained SVM for each behavior. Then, the predicted effect's prototype is determined. The mean value of this effect is added to the initial features, with the exception of the inconsistent features, and the predicted future state is found. After this application, the predicted future state can be compared with other states; but the inconsistent features of the applied effect (denoted as black columns in predicted after-state) is excluded from the comparison calculations.


Figure 3: Future state prediction.

Application of effects Given the object, we can obtain from the trained SVMs the behavior that can achieve a desired effect with the highest probability. Thus, we obtain the behaviors required for each step in the planned effect sequence, forming a sequence of behaviors. If the obtained effect at any step in the behavior sequence does not match with the expectation, then the planning restarts. Fig. 3 and 4 respectively exemplify how a sequence of effect prototypes for reaching a desired effect is sought and how a behavior that produces an effect on an object is found. The system executes the planned behavior sequence. If, at any step, the predicted effect is not achieved (including overshoots or undershoots), the planning restarts from the current object state.


Figure 4: A simple depiction of how the planning is performed using the combinations of effects in the repertoire. At each step, the prediction block described in Fig. 3 is applied for each behavior. Once a future state close enough to goal state is obtained, the search is terminated.

\section*{Results}

\section*{Social Affordance Learning}

Fig. 5 shows some of the effect prototypes that lead us to interesting observations. In the first place, some effects can apparently be produced both by social and non-social behaviors. An obvious example is "say push to your left" and pushright (causing the moved right effect most of the time). Furthermore, we observe that in some cases, social behaviors can be a better option for goal emulation. For instance, when the object is far enough from the robot, the predicted effect for pull behavior is no change; whose effect prototype has only *'s and 0's (features with inconsistent change and negligible change), whereas the predicted effect for "pass me" behavior is pulled, the effect whose prototype denotes consistent decrease in object's distance to the robot (x-position). In emulating a goal to pull this object towards itself, Eq. 2 yields that pulled effect brings the object closer to the goal, hence iCub chooses to ask a human to pass the object.

The effects got attention and got closer turned out to be ambiguous effect labels - their corresponding clusters did not have any consistently increasing or decreasing features. This might also be related with our feature set. Similar results were observed for the effects clustered as sat down and stood up, although they were unambiguously interpreted by the participants. The amount of standing and sitting of our experiment participants has had a high variance. The participants had two major interpretations for the "pass me" behavior: they either (i) pushed the object towards the robot (causing pulled effect) or (ii) tried to pass it to robot's hand (Fig. 6). Similar response was also observed when the voice behavior "take this" was applied: while most of the participants took the object and removed it from the scene (causing the disappeared effect), some of the participants just dragged the object towards themselves (causing moved forward effect). We were expecting that when iCub enacted the voice behavior "bye",
the participants would leave the scene. However, participants mostly kept their positions and responded by waving back.


Figure 5: Some of the results obtained from unsupervised clustering of feature mean and variances. Vertical axis represents some of the effects (E1: moved-right, E2: moved-left, E3: moved-forward, E4: pulled and E5: disappeared) and the horizontal axis represents some of the features detected as consistently increasing (red upwards arrow) or decreasing (blue downwards arrow) in the given effects (F1: position-x, F2: position-y, F3: human presence))


Figure 6: Some different reactions by experiment participants when the robot uses the "pass me" voice behavior.

Both social and non-social behaviors contribute to these results. For example, pulled can be produced both from pull and "pass me" behaviors. Note that some of the features, which were inconsistently changed (marked with star) or negligibly changed (marked with circle), grouped into one column for brevity.

\section*{Social Affordances and Multi-step Planning}

We demonstrate social affordances in three scenarios:
1- Multi-step planning without human presence In this scenario, the object is placed in front of iCub as the initial position, and the target position is shown with red circles (Fig. 7a). After initial and final positions are shown to iCub, it plans without a human present in the environment; i.e., it cannot make use of "social affordances". According to the plan, the effect sequence is determined as moved forward, moved left, moved forward. After a successful push-forward behavior, the object is moved-forward (Fig. 7b), then with a push-left behavior, the object reaches close to the target position (Fig. 7c). Appropriate behaviors to end up with the last moved-forward effect may have been push-forward behavior or "pass me" voice behavior. Since there is no human across
the table and because the object is too far to be moved forward to its target position, iCub figures out that it is impossible for the object to reach its final position (Fig. 7d) and stops at this stage.


Figure 7: Scenario \#1: The robot cannot reach the target position and cannot fulfill the goal due to absence of a human.

2- Multi-step planning with a human - using "pass me" voice behavior This scenario demonstrates a case for successful planning. As the initial position, the object is placed closer to the human and the target position is shown with a red circle (Fig. 8a). After planning, the effect sequence pulled, pulled, moved left is determined to reach the target position. For the first pulled effect, since the object is placed far from iCub and with the contribution of human presence, "pass me" voice behavior has the highest probability and is executed (Fig. 8b). For the remaining pulled and moved-left effects, pull (Fig. 8c) and push-left (Fig. 8d) behaviors are executed respectively. As a result, each planned effect is achieved and the object reaches its target position (Fig. 8e).


Figure 8: Scenario \#2: The robot can use human's affordances when stuck, by using the "pass me" voice behavior.

3- Multi-step planning with a human - using "take" voice behavior This scenario shows a demonstration in which iCub finds a valid plan but because of a behavior which results with an unexpected effect, iCub re-plans. For this scenario, the object is placed close to iCub and the target position is shown with a red circle (Fig. 9a). The planner finds out the required effect sequence as moved forward, moved right, moved forward. The first two effects are achieved using the push-forward (Fig. 9b) and then push-right behavior (Fig. 9c). For the last effect, push-forward behavior is executed. However, instead of a moved-forward effect, moved-right effect occurs (Fig. 9d). Because of this unexpected effect, iCub needs a re-planning (Fig. 9e). This re-planning results with
a new effect sequence of moved left, moved forward. This re-planned effect sequence is achieved by using push-left behavior (Fig. 9f) and "take" voice behavior (Fig. 9g) and object reaches its target position (Fig. 9h).


Figure 9: Scenario \#3: The robot can use human's affordances when stuck, by using the with "take" voice behavior.

From these 3 different scenarios, we conclude as follows: (i) After iCub executes a behavior, if it observes an unexpected effect, it re-plans. (ii) iCub executes its behaviors by planning (and re-planning if necessary) until the object reaches the target position or iCub decides that it is impossible for the object to reach the target position. (iii) If there is a human, iCub may benefit from the affordances offered by the human to get a desired effect. (iv) If there is no human and the desired effect requires a human, iCub can realize that it is impossible for the object to reach the target.

\section*{Conclusion}

In this paper, we used the very same affordance learning framework developed for discovering the affordances of inanimate things to learn social affordances, that is affordances whose existence require the presence of humans. We demonstrated that our humanoid robot can interact with objects and with humans (using simple verbal communication) and from these interactions, it can learn what the objects as well as the humans afford. Moreover, we showed that the robot can ask for human assistance whenever it is required while executing multi-step plans to satisfy demonstrated/given goals.

Our approach towards learning the social affordances is in line with the findings that affordances at different levels (intra-level and inter-level) share the same intrinsic constraints and organizations (e.g., (Richardson et al., 2007)).

\section*{Acknowledgement}

This work is partially funded by the EU project ROSSI (FP7-ICT-216125) and by TÜBİTAK through project 109E033. The authors Kadir F. Uyanik, Onur Yuruten and Asil K.

Bozcuoglu acknowledges the support of Turkish Science and Technology Council through the 2210 scholarship. Figures in the result section are adapted from (Mutlu, 2009) with author's permission.

\section*{References}

Akgun, B., Cakmak, M., Jiang, K., \& Thomaz, A. L. (2012). Keyframe-based learning from demonstration. International Journal of Social Robotics, 4(4), 343-355.
Argall, B., Chernova, S., Veloso, M., \& Browning, B. (2009). A survey of robot learning from demonstration. Robotics and Autonomous Systems, 57(5), 469-483.
Atil, I., Dag, N., Kalkan, S., \& Şahin, E. (2010). Affordances and emergence of concepts. In Epigenetic robotics.
Breazeal, C. (2004). Social interactions in hri: the robot view. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 34(2), 181-186.
Cakmak, M., DePalma, N., Arriaga, R., \& Thomaz, A. (2010). Exploiting social partners in robot learning. Autonomous Robots, 29(3), 309-329.
Dag, N., Atil, I., Kalkan, S., \& Sahin, E. (2010). Learning affordances for categorizing objects and their properties. In In 20th int. conference on pattern recognition (pp. 30893092).

Detry, R., Kraft, D., Buch, A., Kruger, N., \& Piater, J. (2010, may). Refining grasp affordance models by experience. In Robotics and automation (icra), 2010 ieee international conference on (p. 2287-2293).
Fitzpatrick, P., Metta, G., Natale, L., Rao, S., \& Sandini, G. (2003). Learning about objects through action - initial steps towards artificial cognition. In Ieee international conference on robotics and automation (icra'03) (p. 3140-3145).
Fong, T., Thorpe, C., \& Baur, C. (2003). Collaboration, dialogue, human-robot interaction. Robotics Research, 255266.

Gibson, J. (1986). The ecological approach to visual perception. Lawrence Erlbaum.
Kosuge, K., \& Hirata, Y. (2004). Human-robot interaction. In Robotics and biomimetics, 2004. robio 2004. ieee international conference on (pp. 8-11).
Kreijns, K., \& Kirschner, P. (2001). The social affordances of computer-supported collaborative learning environments. In Frontiers in education conference (Vol. 1).
Marsh, K., Richardson, M., \& Schmidt, R. (2009). Social connection through joint action and interpersonal coordination. Topics in Cognitive Science, 1(2), 320-339.
Montesano, L., Lopes, M., Bernardino, A., \& Santos-Victor, J. (2008). Learning object affordances: From sensorymotor coordination to imitation. Robotics, IEEE Transactions on, 24(1), 15-26.
Mutlu, B. (2009). Designing gaze behavior for humanlike robots. Unpublished doctoral dissertation, Pittsburgh, PA, USA. (AAI3367045)
Pastor, P., Hoffmann, H., Asfour, T., \& Schaal, S. (2009). Learning and generalization of motor skills by learning
from demonstration. In Robotics and automation, 2009. icra'09. ieee international conference on (pp. 763-768).
Qin, A. K., \& Suganthan, P. N. (2004). Robust growing neural gas algorithm with application in cluster analysis. Neural Networks, 17(8-9).
Richardson, M., Marsh, K., \& Baron, R. (2007). Judging and actualizing intrapersonal and interpersonal affordances. Journal of experimental psychology: Human Perception and Performance, 33(4), 845.
Şahin, E., Çakmak, M., Doğar, M., Uğur, E., \& Üçoluk, G. (2007). To afford or not to afford: A new formalization of affordances toward affordance-based robot control. Adaptive Behavior, 15(4), 447-472.
Saunders, J., Nehaniv, C., Dautenhahn, K., \& Alissandrakis, A. (2007). Self-imitation and environmental scaffolding for robot teaching. International Journal of Advanced Robotic Systems, 4(1), 109-124.
Sinapov, J., \& Stoytchev, A. (2008, aug.). Detecting the functional similarities between tools using a hierarchical representation of outcomes. In Development and learning, 2008. icdl 2008. 7th ieee international conference on (p. 91 -96).
Stoytchev, A. (2008). Learning the affordances of tools using a behavior-grounded approach. Towards Affordance-Based Robot Control, 140-158.
Ugur, E., Celikkanat, H., Sahin, E., Nagai, Y., \& Oztop, E. (2011). Learning to grasp with parental scaffolding. In Humanoid robots (humanoids), 2011 11th ieee-ras international conference on (pp. 480-486).
Ugur, E., \& Şahin, E. (2010). Traversability: A case study for learning and perceiving affordances in robots. Adaptive Behavior, 18(3-4), 258-284.
Ugur, E., Şahin, E., \& Oztop, E. (2009). Affordance learning from range data for multi-step planning. Int. Conf. on Epigenetic Robotics.
Ugur, E., Oztop, E., \& Şahin, E. (2011). Goal emulation and planning in perceptual space using learned affordances. Robotics and Autonomous Systems, 59(7-8).
Weber, J. (2008). Human-robot interaction. Handbook of Research on Computer Mediated Communication. Hershey, PA: Information Science Reference.
Wellman, B., Quan-Haase, A., Boase, J., Chen, W., Hampton, K., Díaz, I., et al. (2003). The social affordances of the internet for networked individualism. Journal of ComputerMediated Communication.
Williams, K., \& Breazeal, C. (2012). A reasoning architecture for human-robot joint tasks using physics-, social-, and capability-based logic. In Intelligent robots and systems (iros), 2012 ieee/rsj international conference on (pp. 664-671).
Yürüten, O., Uyanık, K. F., Çalışkan, Y., Bozcuoğlu, A. K., Şahin, E., \& Kalkan, S. (2012). Learning adjectives and nouns from affordances on the icub humanoid robot. In From animals to animats 12 (pp. 330-340). Springer.

\title{
Linguistic cued attention in children: Words organize attention to shape in a visual search task
}

\author{
Catarina Vales (cvales@indiana.edu) \\ Department of Psychological and Brain Sciences, \(1101 \mathrm{E} 10^{\text {th }} \mathrm{St}\) Bloomington, IN 47405 USA \\ Linda B. Smith (smith4@indiana.edu) \\ Department of Psychological and Brain Sciences, \(1101 \mathrm{E} 10^{\text {th }} \mathrm{St}\) \\ Bloomington, IN 47405 USA
}

\begin{abstract}
By one account of early word learning, children become proficient word learners as a result of environmental regularities: Learning words tunes the child to the regularities offered by the language being learned, orienting attention to those regularities. We test one core claim of this account, that count nouns should cue attention to the shape of the objects. Using a visual search task we present evidence that hearing the name of the object narrows children's attention to the objects in the array that have the same shape. Future steps and the implications of these results are discussed.
\end{abstract}

Keywords: attention; language and cognition; visual search; word learning.

\section*{Introduction}

Using past experience to select what to attend to is a powerful feature of human cognition. If exposed to environmental regularities, infants (Kirkham, Slemmer and Johnson, 2002; Saffran, Aslin \& Newport, 1996) and adults (Chun \& Jiang, 1998; Zhao, Ngo, McKendrick, \& TurkBrowne, 2011) readily attend to the current events that better match the underlying structure of their previous experience.

This ability to selectively attend to the most reliable sources of information as a result of past experience has been proposed as a mechanism underlying early word learning. Not only do children learn a large amount of words in the first years of life, but they also seem to do so in very smart ways. For example, by 2.5 years children use shape to generalize new noun categories - if given a novel named object, children will selectively attend to shape over color or texture when extending the novel name to new exemplars (the shape bias; Booth \& Waxman, 2002; Jones \& Smith, 2002).

Because in English many count nouns map to object categories well organized by within-category shape similarity, learning individual word-object mappings could create a top-down process that would organize future learning. According to the attentional learning account of the shape bias, it is the co-occurrence of nouns and shapes that creates an attentional bias to shape over other features when generalizing a new object category. Although there is evidence supporting the attentional account (GershkoffStowe \& Smith, 2004; Jones, 2003; Smith, Jones, Landau, Gershkoff-Stowe, \& Samuelson, 2002), the specific
mechanisms through which nouns cue object shape have not been directly tested. In the traditional shape bias task, children are asked to generalize a new object category in an untimed forced-choice procedure. The fact that children tend to select the shape match could be do to attentional processes or to much later decision processes. Indeed, competing theories of the attentional learning account have suggested that the shape bias reflects more conceptual theories about how words refer to objects (Waxman \& Gelman, 2009).

The purpose of this paper is to empirically test a core claim of the selective attention account: That words cue children's attention to the shape of the objects. To this end, we use a visual search task - a well-documented attentional task in which participants are asked to find a particular object (the target) amidst distractor objects. In the visual search literature with adults (Treisman \& Gelade, 1980) and children (Gerhardstein \& Rovee-Collier, 2002), when the target and the distractors differ by just one feature, search is almost effortless and does not depend on the number of distractors. When the target and the distractors have overlapping features, finding the target becomes a serial search - and response times depend on the number of distractors. The intercept and the slope of the search function are also indicators of the attentional processes involved. While the slope reflects the per item search time (i.e. how long it takes per item to decide if it is the target), the intercept is thought to reflect pre-search processes, including the representation of the search target in working memory (Vickery, King \& Jiang, 2005; Woodman, Vogel and \& Luck, 2001).

To investigate the role of labels in visual attention we use a visual search task to compare children's performance when they were cued with both the spoken name and a picture of the target versus then they were cued with just a picture of the target.

\section*{Experiment 1: Do labels cue attention to shape?}

Finding a target requires keeping a representation of the target in working memory. Research with adults has suggested that more robust working memory representations of the target result in overall decrease in search times (i.e. intercept changes). Because visual attention is biased
towards elements in the array that match the contents of working memory, stronger working memory representations would effectively suppress attention to nonmatching elements in the search array (Kristjansson, Wang \& Nakayama, 2002; Soto \& Humphreys, 2007; Vickery, King \& Jiang, 2005), therefore modulating pre-search attentional processes.

Does hearing the name of the target prior to search influence its representation in working memory, and thus search? If hearing the object name results in the enhanced representation of object shape relative to other properties, then the explicit naming of the search target on each trial should effectively narrow search to items in the array with the same shape. This is the hypothesis tested in Experiment 1.

In a conjunctive search task, children were asked to search for a particular colored object (e.g., red bed) in a field of same shape (e.g., green bed) and same color (e.g., red couch) distractors. In the Label condition, children heard the displayed object (but not its color) named (e.g., "bed") prior to each search trial; in the Silent condition, they just saw the displayed target. If storing the name along with the target object in working memory supports processes that automatically direct attention to same shaped items in the array, then overall search time should decrease in the Label condition as participants would preferentially examine the shape matching objects to find the conjunctive match. That is, by hypothesis, in the Label condition children's attention might be automatically attracted to the shape matching items, with attention to the non-shape matching items being dampened. If so, this would effectively reduce the search set and should lead to faster overall search times in the Label than in the Silent condition.

\section*{Methods}

Participants. Thirty-two children ( \(\mathrm{M}=37\) months, range: 31-43 months) were assigned to either the Silent or the Label condition. Ten additional children were recruited but not included in the final sample due to refusal to participate in the study ( \(\mathrm{N}=3\) ), not finishing the familiarization phase ( \(\mathrm{N}=1\) ), or selecting a non-target object on most test trials ( \(\mathrm{N}=6\) ). Children were reported to have no developmental disorders, normal (or corrected to normal) visual acuity and color vision. English was the main language spoken by all families. Parental consent was obtained for all participants in compliance with the IRB of Indiana University.
Stimuli and procedure. Figure 1 shows the experimental set up and the temporal order of events on each trial. The child was seated approximately 35 cm from a \(17^{\prime}\) monitor equipped with a touchscreen (MagicTouch, Keytec, Garland, TX). E-Prime software (PST, Pittsburg, PA) was used to control stimulus presentation and record the latency and the location of each response during the test phase. On each test trial, a "fixation" slide encouraged the child to rest their hands on the table (Figure 1a) before the target object was displayed on the center screen for \(1 \sec\) (Figure 1b).

The search array (with the target object amid distractor objects) was then displayed and the child asked to find the target picture as fast as possible (Figure 1c). Prior to the test phase, children were familiarized with the search procedure, with holding their hands on the table during fixation, and touching the target.


Figure 1: Structure of a trial (left) and child performing the task (right).

Each child was assigned one search target and searched for the same object throughout 32 test trials. Four different objects served between subjects as targets: a red bed, a red couch, a green bed, and a green couch. For each target, the distractors were selected so that half had the same shape and half had the same color as the target (that is, when the target was a red bed, half the distractors were red couches and half were green beds). Each test stimulus was rendered in a 180 x 140 pixel area on a white background. Across trials, the number of distractor objects was manipulated: on each trial, the target object was placed amidst \(2,4,8\) or 12 distractors; eight occurrences of each distractor set size was presented in an order randomly determined for each subject. Sixteen possible locations were used to place the target and the distractors. Across test trials, the target appeared equally often on the left and right side of the screen.

The experimenter started each trial ensuring that the child was looking at the screen; no time limit was set for finding the target. No feedback was given during test phase. In the Label condition, a sound file containing the name of the target object (e.g. "bed") played at the onset of the target (Figure 1b). The audio files were recorded using an artificial speech creator at a sample rate of 16 KHz . No sound file was played in the Silent condition.

\section*{Results and Discussion}

Mean reaction times (RT) per distractor level were calculated for each child. Only correct responses (i.e. when the target object was selected) were included. Although some participants did not complete all test trials, no differences were found between conditions in the total number of trials completed, \(t(30)=-0.37\), n.s., nor in accuracy, \(t(30)=0.14\), n.s. (see Table 1). Figure 2 depicts mean RT for the Silent and the Label conditions as a function of number of distractors. A mixed \(2 \times 4\) analysis of variance with condition as the between-subjects factor and number of distractors as the within-subjects factor yielded a main effect of distractor number \([F(3,90)=27.30, p<\) 0.001 ], reflecting the fact that RT increased as the number
of distractors increased. A main effect of condition was also found \([F(1,30)=4.48, p<0.05]\), reflecting a significant decrease in overall RT for the Label condition. Number of distractors and condition did not interact \([F(3,90)=0.21\), n.s.]. The slopes and intercepts of the linear best-fit lines were also calculated for each child. Independent samples ttests showed that while the slopes of the two conditions were not different \([t(30)=0.39\), n.s.], there was a significant reduction in the intercept of the Label condition when compared to the Silent condition \([t(30)=-2.40, p<\) 0.05].


Figure 2: Mean RT (ms) per number of distractors for the Silent and the Label conditions in Experiment 1. Error bars represent standard errors.

These results are consistent with the hypothesis that hearing the name increases attention to shape matching items and/or decreases attention to non-shape matching distractors - thus decreasing overall search time. The results provide direct evidence for a role of object names in guiding children's attention to object shape. However, presenting the target label did not affect the slope of the search function, which may indicate that the label does not affect the time it takes to make a decision per each attended item. This point will be addressed in the General Discussion section.

Experiment 2 examines an alternative account for the present findings; that it was not the object name per se which enhanced overall search time, but the presence of an auditory signal at the start of each trial.

\section*{Experiment 2: Does any word cue attention?}

A growing literature shows multimodal influences on visual attention and search such that auditory cues may lead to more rapid search (Iordanescu, Grabowecky, \& Suzuki, 2011; Van der Burg, Olivers, Bronkhorst, \& Theeuwes 2008). Thus it is possible that the effects observed in Experiment 1 were due to the addition of a spoken word potentially any word - and not to the target's name nor increased attention to object shape. Accordingly, Experiment 2 replicated the Label condition of Experiment

1 but replaced the target name on each trial with the word "Go."

\section*{Methods}

Participants. Sixteen children between 32 and 42 months of age ( \(\mathrm{M}=37\) months) participated; none of these children had participated in Experiment 1. Eleven additional children were recruited but not included in the final sample due to selecting a non-target object on most test trials. Recruitment and informed consent procedures were the same as in Experiment 1.
Stimuli and procedure. All aspects were the same as in the Label condition of Experiment 1, except that the sound file presented at the onset of the target played the word "Go."

\section*{Results and Discussion}

Mean RT per number of distractors for correct responses was calculated for each child. Children completed 30 trials ( \(\mathrm{SD}=2.98\) ) on average, and mean accuracy was \(83 \%\) (see Table 1). Figure 3 presents RT for correct responses per distractor level for the Go condition. For comparison purposes, results from the Silent condition from Experiment 1 are also shown. A mixed \(2 \times 4\) analysis of variance with number of distractors as within-subjects factor and condition as the between-subjects factor yielded no reliable differences in RT between the Go condition of Experiment 2 and the Silent condition of Experiment \(1[F(1,30)=0.06\), \(p=0.82\) ]. A main effect of distractors number was found [ \(F\) \((3,90)=23.82, p<0.001]\), reflecting the increase in RT as a result of increasing the number of distractors. These two factors did not interact \([F(3,90)=0.41, p=0.75]\). The analyses of the individual slopes and the intercepts confirmed the trends found for RT: No differences were found between the Go condition of Experiment 2 and the Silent condition of Experiment 1 in the slope \([t(30)=0.25\), \(p=0.80]\) or the intercept \([t(30)=-0.38, p=0.71]\).

In brief, an auditory word that is not the name of the target does not result in more rapid search than the presentation of no sound at all, a result that suggests the observed effects in Experiment 1 were not due to an auditory cuing effect but instead reflected the presentation of the object name.

Table 1: Experiments 1 and 2 - Slopes and Intercepts of the search functions, Mean accuracy and Mean number of trials completed.
\begin{tabular}{lcccc}
\hline \multicolumn{1}{c}{ Condition } & \begin{tabular}{c} 
Slope \\
(SE)
\end{tabular} & \begin{tabular}{c} 
Intercept \\
\((\mathrm{SE})\)
\end{tabular} & \begin{tabular}{c} 
Accuracy \\
(SE)
\end{tabular} & \begin{tabular}{c} 
Trials \\
completed \\
(SD)
\end{tabular} \\
\hline Exp.1-Silent & \(212(28)\) & \(3264(212)\) & \(85(3)\) & \(31(1.55)\) \\
Exp.1-Label & \(233(5)\) & \(2284(37)\) & \(86(4)\) & \(31(3.75)\) \\
Exp.2-Go & \(223(15)\) & \(3085(113)\) & \(83(3)\) & \(30(2.98)\) \\
\hline
\end{tabular}


Figure 3: Mean RT (ms) per number of distractors for the Go condition of Experiment 2 and the Label condition in Experiment 1. Error bars represent standard errors.

\section*{General Discussion}

Word learning requires selective attention to the right properties - shape is likely to be the right feature if generalizing countable objects, while texture might be a better alternative if generalizing food terms. What processes support the development of these attentional biases? This question is the source of a major dispute in the literature, with some arguing for the role of attended regularities in setting up such attentional biases (Smith, Jones, Yoshida, Colunga, 2003) and others arguing that word learning entails more conceptual and deliberative processes about how categories are formed (Waxman \& Gelman, 2009).

The current results offer support to the first alternative. By showing that words direct attention to the shape of known objects, we offer for the first time direct evidence for a core claim of the attentional learning account. After enough instances of attention being directed to the shape of known objects in the presence of their names, a generalized attentional bias might emerge - any noun, even a novel one, could now cue attention to shape. This is potentially a very powerful learning mechanism, one that enables children to quickly generalize new categories in the presence of new words.

The current results also suggest that the attentional effects of words may be located at the level of working memory hearing an object's name strengthens the shape representation of that object. The finding that hearing the target name on a visual search task influences the intercept of the search function, without changes in the slope, also suggests that the effects of words do not influence the efficiency of search (i.e. how long it took to identify or dismiss each item as the target). However, the present version of the task may not be optimal to test for a potential role of labels in the efficiency of search (as measured by the slope of the search function). In adults, the slope - or per item search time - is affected by the familiarity of the target and distractors (Mruczek \& Sheinberg, 2005; Wang,

Cavanagh \& Green, 1994), by the need or ease of binding the features of individual items in the array into their individual units (Treisman \& Gelade, 1980), and by the discriminability of the target from the distractors (Duncan \& Humphreys, 1989). For young children, providing the basic-level name of the target could, in principle, influence any of these processes - and in so doing increase the role of shape in search in ways that expedite the identification of the target. By this hypothesis, given sufficiently difficult shape discriminations, hearing the object name prior to search might be expected to yield a decreased slope in the search function as well as a decrease in overall search time. This is a critical issue for future research.

What are the implications of the current results? Although more research is needed to further understand the mechanisms involved in the attentional effects of labels in word learning, the evidence presented here suggests that hearing a name activates a representation of certain features of the object - in the case of count nouns, object shape. It follows that hearing an object name will cue attention to that object's shape, and over time this has the potential to not only become a more automatic process, but also to change the nature of the representation (possibly from specific individual features to more abstract shape representations). Moreover, by extension from accounts of these processes in the adult literature (Dahan \& Tanenhaus, 2005; Huettig \& Altmann, 2007; Lupyan \& Spivey, 2010) these labeling effects appear to be rapid and automatic, that is, not under deliberative or conceptual control. Thus, the current results provide a stepping-stone to a mechanistic account of how words organize attention in children - and in so doing, may organize early word learning and the on-line comprehension of words in context.

\section*{Acknowledgments}

This research was supported by a grant from the National Institute of Child Health and Development (HD28675) to LBS and a Graduate Fellowship from the Portuguese Foundation for Science and Technology awarded to CV (SFRH/BD/68553/2010). The authors would like to thank the members of the Cognitive Development Lab at IU for useful discussions, Anna MacKinnon, Blakely Meyer and Tracy Kelsey for their help with recruitment and data collection, and the parents and children who participated in these studies.

\section*{References}

Booth, A. E., \& Waxman, S. R. (2002). Word learning is 'smart': evidence that conceptual information affects preschoolers' extension of novel words. Cognition, 84(1), B11-B22.
Chun, M. M., \& Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. Cognitive Psychology, 36, 28-71.
Dahan, D., \& Tanenhaus, M. K. (2005). Looking at the rope when looking for the snake: Conceptually mediated eye
movements during spoken-word recognition. Psychonomic Bulletin \& Review, 12(3), 453-459.
Duncan, J., \& Humphreys, G. W. (1989). Visual search and stimulus similarity. Psychological Review, 96(3), 433.
Gerhardstein, P., \& Rovee-Collier, C. (2002). The development of visual search in infants and very young children. Journal of Experimental Child Psychology, 81: 194-215.
Gershkoff-Stowe, L., \& Smith, L. B. (2004). Shape and the first hundred nouns. Child Development, 75(4), 10981114.

Huettig, F., \& Altmann, G. T. M. (2007). Visual-shape competition during language-mediated attention is based on lexical input and not modulated by contextual appropriateness. Visual Cognition, 15(8), 985-1018.
Iordanescu, L., Grabowecky, M., \& Suzuki, S. (2011). Object-based auditory facilitation of visual search for pictures and words with frequent and rare targets. Acta Psychologica, 137(2), 252-259.
Jones, S. S. (2003). Late talkers show no shape bias in a novel name extension task. Developmental Science, 6(5), 477-483.
Jones, S. S., \& Smith, L. B. (2002). How children know the relevant properties for generalizing object names. Developmental Science, 5(2), 219-232.
Kirkham, N. Z., Slemmer, J. A., \& Johnson, S. P. (2002). Visual statistical learning in infancy: evidence for a domain general learning mechanism. Cognition, 83(2), B35-B42.
Kristjansson, A., Wang, D. \& Nakayama, K. (2002). The role of priming in conjunctive visual search. Cognition, 85, 37-52.
Lupyan, G., \& Spivey, M. J. (2010). Redundant spoken labels facilitate perception of multiple items. Attention, Perception \& Psychophysics, 72(8), 2236-2253.
Mruczek, R. E. B. \& Sheinberg, D.L. (2005). Distractor familiarity leads to more efficient visual search for complex stimuli. Perception and Psychophysics, 67(6), 1016-1031.
Saffran, J. R., Aslin, R. N., \& Newport, E. L. (1996). Statistical learning by 8 -month-old infants. Science, 274(5294), 1926-1928.
Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., \& Samuelson, L. (2002). Object name Learning Provides On-the-Job Training for Attention. Psychological Science, 13(1), 13-19.
Smith, L.B., Jones, S.S., Yoshida, H., Colunga, E (2003). Whose DAM account? Attentional learning explains Booth and Waxman. Cognition, 3, 209-13.
Soto, D. \& Humphreys, G. W. (2007). Automatic guidance of visual attention from verbal working memory. Journal of Experimental Psychology: Human Perception and Performance, 33(3), 730-753
Treisman, A.M., \& Gelade, G. (1980). A feature-integration theory of attention. Cognitive Psychology, 12: 97-136.
Van der Burg, E., Olivers, C. N. L., Bronkhorst, A. W., \& Theeuwes, J. (2008). Pip and pop: Nonspatial auditory
signals improve spatial visual search. Journal of Experimental Psychology: Human Perception and Performance, 34(5), 1053-1065.
Vickery, T. J.; King, L.; Jiang, Y. (2005). Setting up the target in visual search. Journal of Vision, 5, 81-92.
Wang, Q., Cavanagh, P., Green, M. (1994). Familiarity and pop-out in visual search. Perception and Psychophysics, 56(5), 495-500.
Waxman, S. R., \& Gelman, S. A. (2009). Early wordlearning entails reference, not merely associations. Trends in Cognitive Sciences, 13(6), 258-263
Woodman, G.F.; Vogel, E.K.; Luck, S.J. (2001). Visual Search Remains Efficient When Visual Working Memory Is Full. Psychological Science, 12(3), 219-224.
Zhao, J., Ngo, N., McKendrick, R., \& Turk-Browne, N. B. (2011). Mutual interference between statistical summary perception and statistical learning. Psychological Science, 22, 1212-1219.

\title{
Reducing the Impact of Math Anxiety on Mental Arithmetic: The Importance of Distributed Cognition
}

\author{
Frédéric Vallée-Tourangeau (f.vallee-tourangeau@kingston.ac.uk) \\ Department of Psychology, Kingston University \\ Kingston-upon-Thames UNITED KINGDOM KT1 2EE \\ Miroslav Sirota (miroslav.sirota@kcl.ac.uk) \\ Medical Decision Making and Informatics Research Group, King`s College London \\ London, UNITED KINGDOM SE1 3QD \\ Gaëlle Villejoubert (g.villejoubert @kingston.ac.uk) \\ Department of Psychology, Kingston University \\ Kingston-upon-Thames UNITED KINGDOM KT1 2EE
}

\begin{abstract}
Mathematics anxiety negatively affects performance in simple arithmetic tasks. The experiment reported here explored the role of interactivity in defusing the impact of math anxiety on mental arithmetic. Participants were invited to complete additions presented on paper without using their hands or any artefact; in a second, interactive, condition, the same problems were presented in the form of a set of manipulable tokens. Math anxiety was significantly correlated with mental arithmetic performance only in the static condition. The results of a mediation analysis indicated that the effect of math anxiety on mental arithmetic was mediated by working memory capacity in the static condition; in the interactive condition, math anxiety and working memory did not significantly correlate with performance. Interactivity encouraged the coupling of internal and external resources to create a cognitive system that augmented and transformed working memory capacity, diffusing the resource drain caused by math anxiety.
\end{abstract}

Keywords: Mental arithmetic, interactivity, math anxiety, individual differences, distributed cognition

\section*{Introduction}

A person's proficiency in mathematics and an appreciation that effort is a key determinant of math performance will likely have important consequences for his or her educational and occupational opportunities. In addition, a mathematically competent workforce is identified as a strategic driver of economic growth (National Mathematics Advisory Panel, 2008). There are indications in the US and in the UK (National Numeracy Facts and Figures, 2012) that numeracy levels are in decline.

An important factor that impedes math performance and reduces exposure to math-with the inevitably negative impact on the acquisition of math knowledge and skills-is math anxiety. Richardson and Suinn (1972) define math anxiety as "feelings of tension and anxiety
that interfere with the manipulation of numbers and the solving of mathematical problems in a wide variety of ordinary life and academic situations" (p. 551). From a processing efficiency perspective (Eysenck \& Calvo, 1992), math anxiety impairs performance by using up working memory resources to maintain and retrieve negative performance-related thoughts and memories (Ashcraft \& Krause, 2007). As a result, math anxious people deploy limited cognitive resources when working on a math problem, leading to poorer performance, reinforcing a cycle of anxiety and avoidance that perpetuates poor numeracy.

\section*{Mental Arithmetic}

In the absence of pen and paper, mental arithmetic is a quintessential working memory task. Admittedly, for simple problems where the solution draws on long-term memory knowledge of well-rehearsed answers (e.g., \(3+\) 3 ), working memory plays a more limited role (DeStefano \& LeFevre, 2004). However, for more complex problems, such as multiple number additions, working memory resources must be deployed to arrive at a correct answer (Ashcraft, 1995). These resources involve storage of interim totals and place markers as well as executive function skills that direct attention (e.g., which number to add next) or the retrieval of strategies to support more efficient and reliable performance.

The exact nature of the resources recruited depends on the context of reasoning, defined by the features of the external environment in which the problem is presented. For one, the manner of presentation (visual, auditory) would recruit different subsystems of working memory. In addition, if the numbers are visually presented, working memory would be taxed differently depending on whether the presentation is sequential or simultaneous. Even with a simultaneous presentation, the numbers' arrangement in space-columnar, linear, or random-
would also determine the extent of working memory load. More important is the opportunity to manipulate the problem presentation to facilitate thinking: Enabling participants to re-order and group numbers would likely help them remember the numbers already added, identify felicitous sub-totals and interim totals, guide attention, and encourage the development of more efficient arithmetic strategies.

Imagine a participant invited to complete an addition problem involving seven numbers, some single digit, some double digit. In one condition, the problem is presented on a piece of paper as a randomly configured array of numbers; the participant is asked to put her hands palm down on the flat surface on which the problem is presented. The mental effort required cannot be guided and supplemented with complementary actions (Kirsh, 1995) such as pointing and re-arranging. In this context, mental arithmetic performance should reflect the participant's working memory capacity, arithmetic knowledge and skill. Imagine, in turn, the same problem but, this time, presented as a set of number tokens, which the participant is invited to manipulate. The importance of arithmetic knowledge and skills remain; however, now, working memory is augmented by a modifiable problem presentation. Such a dynamic presentation unveils a shifting array of opportunities and possibilities, whether strategically engineered or fortuitously encountered. Thus, working memory is augmented not simply in terms of storage capacity, but also in terms of executive functions. That is, a shifting problem presentation cues certain strategies-for example by grouping certain numbers together-and guides attention. Hence, in a modifiable environment, the strategic control of attentional resources originates, partly, in the world.

\section*{The Present Experiment}

Participants' performance in a mental arithmetic task is likely to be impaired by math anxiety, and this may be particularly apparent when the mental arithmetic task requires a larger commitment of working memory resources, such as in a static context of reasoning where participants cannot interact with numbers that compose a problem. In turn, if reasoners are given the opportunity to couple their working memory resources and arithmetic skills to a dynamic and modifiable problem presentation, the impact of math anxiety might be considerably attenuated. This is because the coupling of internal and external resources creates a more robust and resilient cognitive system that augments the participants' working memory resources, which then can more easily soak up the resource-depleting rehearsal of performance-related thoughts. Arithmetic performance might be positively correlated with math anxiety in a static reasoning environment; however when participants can extend their cognitive resources and let the environment shoulder some of the computational efforts, then accuracy may be
influenced by math anxiety to a lesser extent.
Math anxious individuals cope with math anxiety by limiting their exposure to math, which further limits their levels of numeracy (Ashcraft, 2002). Hence, to get a better window on the influence of anxiety on mathematical cognition, a relatively simple task was developed for this experiment engaging basic arithmetic skills acquired and mastered by university undergraduates. Participants completed the additions in both a static, non-interactive, context and in one where tokens corresponding to the elements of the addition problems could be touched, arrayed, grouped, in whatever manner to support problem solving; hence interactivity was manipulated within-subjects.

Performance was measured in terms of accuracy (absolute error) and efficiency. Thinking efficiency was calculated as the ratio of the proportion of correct solutions for a set of problems over the proportion of time invested by that participant to complete the set out of the maximum time invested by the slowest participants. In the static condition, participants' working memory resources would likely be stretched, particularly by the long additions; in turn the coupling of internal to external resources in the interactive condition could augment the participants' working memory capacity and executive processes.

Participants' working memory capacity was assessed using a computation span task. Math anxiety was predicted to correlate negatively with working memory capacity. More important, the magnitude of error in the mental arithmetic task was predicted to correlate positively with anxiety level and negatively with working memory capacity, but only in the static condition. Thus, a key prediction was that interactivity would defuse the impact of anxiety on calculation error. In a similar manner, math anxiety and working capacity should predict thinking efficiency in the static, but not in the interactive condition. Mediation analyses were conducted to determine the direct and indirect effect of math anxiety on thinking efficiency in both conditions.

\section*{Method}

\section*{Participants}

Forty psychology university undergraduates ( 35 females, overall mean age \(20.8, S D=3.2\) ) received course credits for their participation.

\section*{Material and Measures}

Mathematics Anxiety. Mathematics anxiety was measured using an abridged version of the original 98 -item scale (Suinn, 1972) developed by Alexander and Martray (1989). The abridged version is based on 25 -items for each of which participants used a 5-point scale ( \(1=\) "not at all", 5 = "very much") to describe how anxious the event
described made them feel. The 25 items assessed math anxiety in terms of test anxiety (e.g., "studying for a math test"), numerical task anxiety (e.g., "reading a cash register receipt after your purchase") and math course anxiety (e.g., "watching a teacher work on an algebraic equation on the blackboard"). Math anxiety scores could range from 25 to 125 - the higher the score, the higher the math anxiety; the mean score in the present sample was \(66.0(S D=18.1)\).

Working memory capacity. Working memory was assessed using a computation-span test (Ashcraft \& Kirk, 2001, p. 226). Participants solved simple arithmetic problems in blocks increasing from 2 to 6 problems (e.g., " \(50+7=?\) ?, " \(60 \div 2=?\) ?, " \(19-8=\) ?" was a block of three problems). At the end of each block, participants were prompted to recall in correct order the last number of each problem in that block (for the example above, correct recall would be " \(7,2,8\) "). There were two blocks for each sequence length (e.g., two blocks with sequences of 3 different problems) for a total of 10 blocks. Working memory capacity was measured as the sum of all correct answers across the 10 blocks, for a maximum score of 40 . The mean number of digits recalled by the participants in the present study was \(24.1(S D=7.6)\).

Arithmetic Task. Participants carried out short and long additions, involving either 7 or 11 numbers (see Fig. 1), as fast and as accurately as possible. They completed the problems in blocks, five from the short set first, and five from the long set second. Performance was measured in terms of the mean absolute error and in terms of efficiency. Efficiency was measured as the ratio of accuracy (proportion correct sums) over time invested in doing the sums. The latter was measured as the proportion of actual time to complete the sums divided by the maximum time needed to complete them in that condition; this maximum was determined by taking the average of the top quartile latencies. Inefficient performance is reflected with a ratio smaller than 1 indicating that proportion accuracy was smaller than proportion time invested.

\section*{Procedure}

The mental arithmetic task, working memory span task, and the completion of the 25 -item mathematics anxiety scale were embedded in an experimental session that lasted approximately 40 minutes, and which included other tests of motivation and cognitive skill unrelated to the present experiment. The session always started with participants completing the math anxiety scale. During the mental arithmetic task, participants were presented with the five additions from the short set first. After a \(2-\mathrm{min}\) distractor task (a word search puzzle), participants were presented with the five additions from the long set; the problem order within each set was randomized for each participant. These two sets of sums were presented twice. For one
presentation participants performed the additions with their hands on the table facing them (the static condition) and announced their answer out loud; for the second presentation, square numbered tokens ( 3 cm by 3 cm ) were used, and participants were encouraged to touch, move or group the tokens in whatever manner to help them add the numbers (the interactive condition); as in the static condition, participants announced the solution for each problem out loud. While the long set always followed the short set, the order of condition (non-interactive, interactive) was counterbalanced across participants. With 10 different problems, involving 10 unique configurations of numbers, and 90 numbers across the two sets, it was unlikely that participants remembered the solution to each problem when presented a second time. Still, to prevent direct retrieval of solutions during the second presentation, the participants completed the computation span test between the two presentations of the arithmetic task. Problem set size (with two levels) and interactivity (with two levels) were independent variables that were manipulated in a \(2 \times 2\) repeated measures design.


Figure 1: Examples of additions from the short set (7number additions) and the long set (11-number additions).

\section*{Results}

The correlation matrix involving the anxiety and working memory span measures along with the mental arithmetic performance measures is reported in Table 1. We note, for now, that math anxiety scores were negatively correlated with working memory span, \(r(38)=-.318, p=.045\). The correlations with the different measures of mental arithmetic performance in the static and interactive conditions are described below.

\section*{Absolute Error}

The mean absolute deviation from the correct answer or absolute error for the short and long sums in the static and interactive conditions are reported in the top half of Table 2. Mean absolute error was similar for the short sums across conditions; however, errors increased for the long sums, in a relatively more pronounced manner in the static condition. In a \(2 \times 2\) repeated measures analysis of variance (ANOVA), the main effect of condition was not significant, \(F<1\), the main effect of problem size was marginally significant, \(F(1,39)=4.02, p=.052\), but the interaction was not significant, \(F(1,39)=2.26, p=.141\).

Table 1: Correlation matrix involving mathematics anxiety, working memory capacity, and mental arithmetic performance averaged across all 10 additions in the static and interactive condition \((d f=38)\).


Note: * \(p<.05 \quad * * p<.01\). MARS \(=\) Mathematics Anxiety Rating Scale scores; SPAN = Computation span scores; ERR-S = Average absolute error in the static condition; ERR-I = Average absolute error in the interactive condition; EFF-S = Average efficiency ratio in the static condition; EFF-I = Average efficiency ratio in the interactive condition.

Math anxiety was strongly correlated with absolute error in the static condition averaged across all 10 problems, \(r=.427, p=.006\) (see Table 1), but not in the interactive condition, \(r=.002, p=.989\). To determine the interaction between math anxiety and condition (interactive, static), the difference in the average absolute errors between the interactive and static condition were regressed on the anxiety scores mean deviation form (an alternative to dichotomising anxiety scores with a median split-which reduces power-as recommended by Brauer, 2002). In the absence of an interaction, one would expect that as math anxiety level increased, participants would not benefit from manipulating the tokens-in other words, the difference between the interactive and static condition would be constant across levels of math anxiety. However, the slope of the regression line, \(\beta=-.372\), was significantly negative, \(t(38)=-2.471, p=.018\). This confirms that participants who were more math anxious made errors of a smaller magnitude in the interactive than in the static condition.

Finally, working memory span was marginally correlated with error in the static condition, \(r=-.283, p=\) .077 , but not in the interactive condition, \(r=.030, p=\) . 852.

\section*{Efficiency Ratio}

The mean efficiency ratios are reported in the bottom half of Table 2. Participants’ efficiency exceeded 1 in the static condition for the short problems, but declined for the long sums. In turn, efficiency remained well calibrated and constant across problem size in the interactive condition. The \(2 \times 2\) repeated measures ANOVA revealed that the main effect of condition was not significant, \(F<1\), the main effect of problem size was significant, \(F(1,39)=\)
\(5.24, p=.028\), as was the condition by problem size interaction, \(F(1,39)=5.37, p=.026\).

Table 2: Mean absolute error and efficiency ratio, along with the standard deviation, for short and long sums in the static and interactive conditions.
\begin{tabular}{|c|c|c|c|c|}
\hline \multirow[b]{3}{*}{Condition} & \multicolumn{4}{|c|}{Set Size} \\
\hline & \multicolumn{2}{|c|}{Short} & \multicolumn{2}{|c|}{Long} \\
\hline & M & \(S D\) & M & \(S D\) \\
\hline & \multicolumn{4}{|c|}{Abolute Error} \\
\hline Static & 3.3 & 4.4 & 5.6 & 6.1 \\
\hline \multirow[t]{2}{*}{Interactive} & 3.6 & 4.7 & 4.2 & 3.9 \\
\hline & \multicolumn{4}{|c|}{Efficiency Ratio} \\
\hline Static & 1.2 & 0.9 & 0.9 & 0.8 \\
\hline Interactive & 1.0 & 0.7 & 1.0 & 0.7 \\
\hline
\end{tabular}

Math anxiety was negatively correlated with the efficiency ratio averaged across all 10 problems in the static, \(r=-.306, p=.055\), but not in the interactive condition, \(r=-.230, p=.153\). The average efficiency ratios were not characterised by a significant math anxiety by condition interaction, however. In the regression of the difference in the average efficiency ratios between the interactive and static condition on the mean deviation form of the math anxiety scores, the slope of the regression line, \(\beta=.161\), was not significantly different from zero, \(t(38)=\) \(1.008, p=.320\).

Working memory span was positively correlated with efficiency in the static, \(r=.494, p=.001\) and to a lesser extent in the interactive condition, \(r=.341, p=.031\). In light of the strong correlation between working memory capacity and efficiency, the mediation of the effect of math anxiety on efficiency via working memory capacity in both the static and the interactive condition was analysed using the procedure and SPSS macro developed by Preacher and Hayes (2008). A simple mediation model analysis was run with math anxiety as the independent variable (X), working memory capacity as the mediator ( M ) and average efficiency as the dependent variable (Y); Figure 2 depicts the results of both mediation model analyses for the static (left panel) and interactive condition (right panel). In the case of the static condition, the total effect of math anxiety on mental arithmetic performance (path c) was negative and significantly different from zero. Math anxiety significantly influenced working memory in a negative direction (path a) and working memory significantly influenced efficiency (path b). Finally, the effect of anxiety on efficiency after controlling for working memory (path \(\mathbf{c}^{\prime}\) ) was no longer significant. A bootstrap analysis revealed that the \(95 \%\) bias corrected interval with 5000 resamples


Figure 2: Results of the mediation analysis in the static (left panel) and interactive condition (right panel).
for the size of the indirect effect ( \(-0.58 ; C I[-1.47 ;-0.03]\) ) did not include a zero value and thus can be consider to be statistically significant. A traditional Sobel's test approached significance, \(z=-1.73, p=.084\). Thus, the effect of math anxiety on mental arithmetic efficiency in the static condition was completely mediated by working memory (see Fig. 2, left panel). In the interactive condition, the total effect of math anxiety on efficiency (path c, see Fig. 2, right panel) was negative but not significantly different from zero. Thus, strictly speaking, the condition for mediation analysis was not fulfilled (Baron \& Kenny, 1986). However, it has sometimes been argued that the indirect effect can still be significant, and omitting this analysis could lead to the failure of detecting interesting mechanisms (Hayes, 2009). With this in mind, the mediation analysis was conducted and showed that math anxiety influenced significantly working memory in a negative direction (path a), while working memory marginally influenced mental arithmetic performance (path b). Finally, the effect of anxiety on mental arithmetic performance after controlling for working memory (path \(\mathbf{c}^{\prime}\) ) was not significant. A bootstrap analysis revealed that the \(95 \%\) bias corrected interval with 5000 resamples for the size of the indirect effect ( \(-0.33 ; C I[-0.99 ; 0.02]\) ) included zero and thus cannot be consider to be statistically significant. Finally, the Sobel's test was not significant, \(z=\) \(-1.41, p=.160\). Thus, there was no significant total or indirect path between math anxiety and mental arithmetic efficiency in the interactive condition (see Fig. 2, right panel).

\section*{Discussion}

In this experiment participants completed short and long additions in two different contexts, one which permitted the reconfiguration of the problem through the spatial rearrangement of the number tokens, and one which did not. Participants were generally accurate-although less so for longer additions-and interactivity did not significantly enhance accuracy. However, the significant interaction between problem size and condition for the efficiency ratio measure confirmed that thinking efficiency dropped for the
longer sums in the static condition, but remained stable in the interactive condition. The interaction between problem difficulty and context of reasoning (static, interactive) indicates that determining the benefits of physically reshaping a problem presentation is an exercise done relative to the degree of task difficulty. Thus, with a relatively easy task, interactivity might not benefit the reasoning agent, but interactivity can enhance efficiency when the task is challenging and undertaken on the basis of internal resources alone.

Math anxiety was significantly correlated with working memory capacity. This has been reported previously (Ashcraft, 2002) especially when capacity is gauged with a span test that involves numbers and operations. The more important findings was the significant interaction between math anxiety level and the degree of interactivity: as math anxiety increased, participants made fewer errors in the interactive than in the static condition.

It is important to stress that this experiment employed a repeated measures design: Participants and their levels of maths anxiety were identical in the static and interactive condition. Having said this a post-task measure in each condition might have offered a better measure of how much anxiety was experienced in completing the sums. Manipulating tokens might have altered participants' experience in terms of intrinsic motivation, attentional commitment, and self-efficacy.

In turn, reasoning efficiency, as determined by the ratio of accuracy over time invested in completing the sums, was marginally correlated with math anxiety in the static condition, but not in the interactive condition. The mediation analysis confirmed that the effect of math anxiety on efficiency in the static condition was mediated by working memory capacity. In turn, in the interactive condition, math anxiety had no effect on reasoning efficiency, but working memory capacity marginally influenced performance. According to processing efficiency theory (Ashcraft, 2002) math anxiety exacts working memory resources to maintain performancerelated beliefs and fears. As the static condition put a higher demand on working memory, efficiency was more
directly determined by working memory capacity. In the interactive condition, however, participants have the opportunity to recruit external resources to help them complete the sums. They can group the number tokens to guide and direct attentional resources and identify congenial interim totals that facilitate more efficient addition strategies. The coupling of internal and external resources creates a cognitive system (Wilson \& Clark, 2009) that augments memory storage and distributes the control of executive function in a manner that copes better with the resource drain caused by math anxiety. These findings lend support to the conjecture that for simple mental arithmetic problems, performance improvements are better supported in a learning environment that fosters interactivity.

Future research may explore the role of interactivity in helping reasoners enhance their mental arithmetic performance in contexts that can elicit higher levels of anxiety, such as under time pressured or in situations of greater accountability. One of the recommendations of the National Mathematics Advisory Panel (2008, p. 31) is to determine the etiology of math anxiety and important advances in charting its neurodevelopmental origins have recently been reported (Young, Wu, \& Menon, 2012). In addition, it might be of particular interest to determine whether intervention programmes that are based on interactive training exercises enhance participants' level of instrumentality, efficacy and confidence, reducing math anxiety in more traditional situations, and encouraging greater exposure to mathematics.

\section*{Acknowledgments}

We would like to thank Ellie McCourty, Angie Makri, and Joakim Westh Wiencken for their assistance with the recruitment and running of the participants. Financial support from the Kingston University Faculty of Arts and Social Sciences Research Capability Fund is gratefully acknowledged.

\section*{References}

Ashcraft, M. H., \& Kirk, E. P. (2001). The relationships among working memory, math anxiety, and performance. Journal of Experimental Psychology: General, 130, 224-237.
Alexander, L., \& Martray, C. (1989). The development of an abbreviated version of the Mathematics Anxiety Rating Scale. Measurement and Evaluation in Counselling and Development, 22, 143-150.
Ashcraft, M. H. (1995). Cognitive psychology and simple arithmetic: A review and summary of new directions. Mathematical Cognition, 1, 3-34.
Ashcraft, M. H., \& Kirk, E. P. (2001). The relationships among working memory, math anxiety, and
performance. Journal of Experimental Psychology: General, 130, 224-237.
Ashcraft, M. H., \& Krause, J. A. (2007). Working memory, math performance, and math anxiety. Psychonomic Bulletin \& Review, 4, 243-248.
Baron, R. M., \& Kenny, D. A. (1986). The moderatormediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology, 51, 1173-1182.
Brauer, M. (2002). L’analyse des variables indépendantes continues et catégorielles: Alternatives à la dichotomisation. L'Année Psychologique, 102, 449484.

DeStefano, D., \& LeFevre, J.-A. (2004). The role of working memory in mental arithmetic. European Journal of Cognitive Psychology, 16, 353-386.
Eysenck, M. W., \& Calvo, M. G. (1992). Anxiety and performance: The processing efficiency theory. Cognition and Emotion, 6, 409-434.
Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. Communication Monographs, 76, 408-420.
Kirsh, D. (1995). Complementary strategies: Why we use our hands when we think. In J. M. Moore \& J. L. Lehman (Eds.), Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society (pp. 212217). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.
National Mathematics Advisory Panel (2008). Foundations for success: The final report of the National Mathematics Advisory Panel. Washington, DC: U.S. Department of Education.
National Numeracy Facts and Figures (2012). Retrieved 1 August, 2012, http://www.nationalnumeracy.org.uk/
Preacher, K. J., \& Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. Behavior Research Methods, 40, 879-891.
Richardson, R., \& Suinn, R. N. (1972). The mathematics anxiety rating scale: Psychometric data. Journal of Counseling Psychology, 19, 551-554.
Suinn, R. M. (1972). Mathematics Anxiety Rating Scale (MARS). Fort Collins, CO: Rocky Mountain Behavioral Science Institute.
Wilson, R. A., \& Clark, A. (2009). How to situate cognition: Letting nature take its course. In P. Robbins \& M. Aydede (Eds.), The Cambridge handbook of situated cognition (pp. 55-77). Cambridge: Cambridge University Press.
Young, C. B., Wu, S. S., \& Menon, V. (2012). The neurodevelopmental basis of math anxiety. Psychological Science, 23, 492-501.

\title{
Effects of Objects' "Embodiment" on the Acquisition of Problem-Solving Skills through Practice or Video-based Modeling Example Study
}

\author{
Tamara van Gog (vangog@fsw.eur.nl) \\ Institute of Psychology, Erasmus University Rotterdam P.O. Box 1738, 3000DR Rotterdam, The Netherlands
}

Lysanne S. Post (l.s.post@fsw.eur.nl)
Institute of Psychology, Erasmus University Rotterdam P.O. Box 1738, 3000DR Rotterdam, The Netherlands

Robyn J. ten Napel (robyn.napel@gmail.com)
Institute of Psychology, Erasmus University Rotterdam P.O. Box 1738, 3000DR Rotterdam, The Netherlands

\author{
Lian Deijkers (cmdeijkers@gmail.com) \\ Institute of Psychology, Erasmus University Rotterdam \\ P.O. Box 1738, 3000DR Rotterdam, The Netherlands
}

\begin{abstract}
We investigated whether "embodiment" of objects used in a problem-solving task (i.e., whether they have a bodily shape) would have a detrimental effect on learning to solve that problem through practice or through studying video-based modeling examples. A \(2 \times 2\) design with factors Training (Practice/Example study) and Embodiment (Present/Absent) was used ( \(N=80\) ). Results showed a large main effect of Training on effort investment in learning and on retention test performance, with Example study leading to higher scores with lower investment of effort during the learning phase than Practice. Numerically, Embodiment seemed to have an effect, with participants practicing/studying the task with embodied objects (plastic animals) performing worse on retention than participants practicing/ studying with nonembodied objects (discs), but this did not reach statistical significance. A new study with more power and an additional control condition is currently being conducted and results are expected to be available well before the conference.
\end{abstract}

Keywords: problem solving; example study; embodiment.

\section*{Introduction}

A substantial body of research in cognitive science has investigated the effects of a problem's appearance on the acquisition of problem-solving skills. For instance, versions of the Tower of Hanoi task that had the exact same problem space but instead of discs, featured monsters passing globes, or acrobats jumping on each other's shoulders, were found to be much more difficult (Kotovsky, Hayes, \& Simon, 1985; see also findings by Goldstone \& Son, 2005, on effects of concrete vs. idealized object appearance on pattern learning from a simulation). The present study investigated whether the "embodiment" of objects featured in a problem, that is, whether the objects have a bodily shape, would have a detrimental effect on learning to solve a problem either by means of practice or by means of studying digital video-based modeling examples. To the
best of our knowledge, the effects of problem appearance on acquiring problem-solving skills from examples has never been investigated yet.

\section*{Practice vs. Example Study}

For students who need to acquire problem-solving skills but lack prior knowledge of a task, practicing with problem solving is not the most efficient way to acquire those skills. It is far more effective and efficient for novice learners to study examples in which the solution procedure is workedout (worked examples) or demonstrated to the learner (modeling examples; for reviews, see Atkinson, Derry, Renkl, \& Wortham, 2000; Renkl, 2011; Sweller, Van Merriënboer, \& Paas, 1998; Van Gog \& Rummel, 2010). Interestingly, the higher effectiveness and efficiency of example study (possibly alternated with problem-solving) compared to problem-solving practice has not only been found when problems contain no guidance whatsoever, but also when they are tutored problems, on which feedback and hints are provided when errors are made (Salden, Koedinger, Renkl, Aleven, \& McLaren, 2010).

Cognitive load theory explains these beneficial effects of example study compared to problem solving in terms of the underlying cognitive processes and associated cognitive load (Sweller et al., 1998). Problems usually contain only a description of some "givens" and a goal statement, without providing any information on how to move from the givens to the goal state. As a consequence, novices have to figure out the correct solution steps to use by themselves, and often do so by resorting to weak problem-solving strategies such as trial-and-error, or means-ends analysis, which impose a high cognitive load but are not very effective for learning: even though such weak strategies may allow learners to succeed in solving the problem eventually (i.e., good performance), they have been shown to contribute
very little to learning (i.e., good performance of that task at a later moment; Sweller, 1988).

Worked examples prevent the use of such weak problemsolving strategies, by presenting the learner not only with the givens and a goal statement, but also with the workedout solution steps that are to be taken to reach the goal state. The learner can devote all of his or her available cognitive capacity to studying the given solution and constructing a cognitive schema for solving such problems, which can be applied to solve this (or a isomorphic) problem in the future. As such, compared to instruction consisting of problem-solving practice, instruction that relies more heavily on studying worked examples reduces ineffective cognitive load on working memory, and leads to enhanced learning outcomes and often to improved transfer performance (Sweller et al., 1998).
In addition to being more effective for learning, a heavier reliance on examples has also been shown to have beneficial effects on required acquisition time (i.e., lower; see e.g., Sweller \& Cooper, 1985; Van Gog, Paas, \& Van Merriënboer, 2006; Zhu \& Simon, 1987) and cognitive load experienced by students during acquisition (i.e., lower; see e.g., Paas \& Van Merriënboer, 1994; Van Gog et al., 2006) as well as during the test (i.e., lower; see e.g., Paas, 1992; Paas \& Van Merriënboer, 1994).

However, it should be kept in mind that the beneficial effects of worked examples on learning, acquisition time, and cognitive load, seem to apply primarily to novice learners (for advanced learners, an 'expertise reversal effect' occurs, and problem solving becomes more effective; Kalyuga, Chandler, Tuovinen, \& Sweller, 2001; see also Kalyuga, Ayres, Chandler, \& Sweller, 2003), and apply only when the examples are well-designed. That is, following early studies on the worked example effect (Cooper \& Sweller, 1987; Sweller \& Cooper, 1985) it was soon discovered that studying worked examples was not always more effective for learning than problem solving. Rather, the design of the examples played a crucial role in their effectiveness (Tarmizi \& Sweller, 1988). For instance, examples that induced split-attention (Chandler \& Sweller, 1991; Tarmizi \& Sweller, 1988) or included redundant information (Chandler \& Sweller, 1991), did not have beneficial effects on cognitive load and learning.
The present study also addresses the effects of problem and example design on cognitive load and learning, though in a very different manner, that is, by investigating the effects of embodiment of the objects used in the task.

\section*{Problem-solving Task and Design Effects}

The task used in this study is based on a computer-based problem-solving task called Frog Leap (see Van Gog, 2011; Van Gog, Jarodzka, Scheiter, Gerjets, \& Paas, 2009). In this computer-based task, the goal is to switch the sides of three brown frogs on the right and three green frogs on the left by clicking on them. There is an empty space in the middle. The frogs face in the direction of their goal. If they are clicked on they jump one place ahead or jump over one
other frog (they cannot jump over two others, and they cannot go back). The problem can be solved in only one way, in 15 moves.

Prior research has shown the superiority of studying modeling examples (consisting of screen-recordings) over problem solving with this computer-based task. Van Gog et al. (2009) showed that none of the 11 participants in the problem-solving condition managed to solve the problem after practicing twice, and Van Gog (2011) reported pilot data with 7 participants showing the same result even after four practice attempts. In contrast, after studying two examples, the numbers of participants to successfully solve the problem was approximately \(58 \%\) (Van Gog et al., 2009), and the number of moves correctly completed was approximately 10 (out of 15 ; Van Gog, 2011). Effects on transfer were not really explored in these prior studies. A second test task was included on which participants had to start on the opposite side as in the example, which was more difficult because the task had not been practiced or studied starting from this side. Therefore, participants could not simply copy the procedure they had learned, and performance on this second test task was lower than on the first (Van Gog, 2011). However, an even stronger transfer test would be to add an additional component on each side, in which case the solution procedure still relies on the same mechanism, but consists of 24 steps and can only be successfully performed when the mechanism is understood.

A closer look at the task suggests that the errors made during problem solving (both during practice and on the test) seem to result from a failure to carefully consider all possible moves and their consequences. This would explain why test performance strongly improved when participants had the chance to study a video-based modeling example twice, in which the procedure was demonstrated (Van Gog, 2011) or demonstrated and explained (Van Gog et al., 2009).

Based on anecdotal evidence of some participants' responses to the task in prior studies, we began to wonder whether this failure to consider all possible moves could be related to the fact that the objects had a bodily shape, that is, were frogs that had a face and "were headed in a direction". That characteristic seemed to evoke anthropomorphic thinking in some participants (i.e., assigning intentions or goals to the frogs; for a discussion of anthropomorphic thinking, see Epley, Waytz, \& Cacioppo, 2007). Assigning intentions to the objects that need to be moved, might aggravate the tendency to rapidly execute steps that seem to physically reduce the distance of a frog to its goal, without considering the other possible moves (cf. Sweller \& Levine's, 1982, maze learning experiment, in which people who had their left hand on the finish and had to move their right index finger through the maze to get to the finish, continuously made incorrect moves to the left, where they knew their goal was).

If this indeed plays a role, then using the same task but with non-embodied objects should lead to better learning outcomes. To investigate this question, we re-created the
computer-based problem-solving task with real objects, that were either "embodied" (i.e., animals) or "non-embodied" (i.e., discs).

\section*{Hypotheses}

Based on prior research on example-based learning in general (for reviews, Atkinson et al., 2000; Renkl, 2011; Sweller et al., 1998; Van Gog \& Rummel, 2010), and on the computer-based version of this task in particular (Van Gog, 2011; Van Gog et al., 2009), we first of all expected that studying digital video-based modeling examples would also be more effective (result in higher learning outcomes) as well as more efficient (higher learning outcomes attained with less investment of mental effort) than problem-solving practice for this real object version of the task. The open question of whether performance on a transfer task would also be enhanced when an additional object is added on each side, is explored.

Secondly, it was hypothesized that practicing the problem-solving task with "embodied" objects (i.e. animals) would lead to lower performance than doing so with "nonembodied" objects (i.e., discs). The open question of whether this would only be the case for the problem-solving practice conditions (cf. Kotovsky et al., 1985), or also for the examples conditions, was explored. On the one hand, when studying examples and subsequently taking a test with embodied objects, this might not have negative effects on test performance because participants had a chance to learn the correct procedure from the examples. On the other hand, however, participants might still be affected by the objects' embodiment (e.g., fall prey to anthropomorphic thinking) once they start performing the test task themselves.

\section*{Method}

\section*{Participants}

Participants were 80 adults ( \(\mathrm{M}=22.8, \mathrm{SD}=2.61\); 43 women) recruited from the general population. A \(2 \times 2\) design with factors Training (Practice vs. Example) and Embodiment (Present vs. Absent) was used. Participants were assigned to one of the four conditions matched for gender, but otherwise randomly: (1) Embodiment Present Practice ( \(n=20\) ), (2) Embodiment Absent - Practice ( \(n=\) 20), (3) Embodiment Present - Example ( \(n=21\) ), and (4) Embodiment Absent - Example ( \(n=19\) ).

\section*{Materials}

Demographic questionnaire A demographic questionnaire asked for age, gender, level of education, and it also included a check on whether participants were familiar with the learning task (by showing them a picture of the initial state of the problem in the computer-based version discussed above).
Learning task The learning task was based on the computer-based problem-solving task mentioned above (see Van Gog, 2011; Van Gog et al., 2009). In this computer-
based task, three green frogs are sitting on stones on one side of the river, three brown frogs on the other side, with one empty stone in the middle. The goal is to have them switch sides, but frogs can only jump one place ahead if that is free, or jump over one other frog to a free place. They cannot go back or jump over two other frogs. The goal can be reached in 15 steps. In this study, a version of the task was created using real objects (see Figure 1), and the objects consisted either of plastic yellow fishes and green seals (Embodiment Present) or yellow and green discs (Embodiment Absent).
In the practice conditions, participants were given two practice opportunities in which they attempted to solve the problem for 1 min .; if they got stuck, they were allowed to start again. In the examples conditions, participants observed a digital video-based modeling example ( 1 min . duration) twice, in which a human model demonstrated the correct solution procedure with either the animal objects or the discs. The model did not provide any verbal explanations and only the model's hand moving the objects was visible in the video. The digital video was presented on a laptop with a screen resolution of \(1280 \times 720\) pixels at a size of \(28.5 \times 18 \mathrm{~cm}\).


Figure 1: Initial state of the problem in the Embodiment Present (top) and Absent (bottom) conditions

Test tasks The retention test task was identical to the learning task. The transfer test task consisted of the same problem, but with four objects on either side. This task could be solved in 24 steps.
Mental effort After each practice task, each example, and each test task, participants rated how much effort they invested in problem solving or example study on Paas' (1992) 9-point rating scale ranging from (1) very, very low effort, to (9) very, very high effort. This subjective rating scale is widely used in educational research (for reviews, see Paas, Tuovinen, Tabbers, \& Van Gerven, 2003; Van Gog \& Paas, 2008).

\section*{Procedure}

The study was conducted in individual sessions of approximately 10 min . After filling out the demographic questionnaire, the learning phase started. Participants were first instructed about the rules of the task (i.e., an object can only move one space ahead to a free space or over one other object to a free space, moving back or moving over two other objects is not allowed). Depending on their assigned
condition, they subsequently received the instruction to either practice for 1 min ., during which they were allowed to start again if they got stuck, or to study the example presented in the video. After practicing or example study, they rated how much effort they invested in problem solving or example study. Then this sequence was repeated a second time. Depending on their assigned condition, participants practiced with either animals or discs or observed a modeling example with either animals or discs. Immediately after the learning phase, the test phase started, during which all participants were required to solve the problem themselves, first the retention task, which was the exact same problem they had encountered in the learning phase, with three objects on both sides, then the transfer task with four objects on both sides. Depending on their assigned condition, participants performed the test tasks with either animals (when they had practiced/studied the task with animals) or discs (when they had practiced/studied the task with discs). Immediately after each task, they indicated how much effort they invested in attempting to solve the problem. In the test phase, participants' performance was recorded on digital video (zooming in on their hands and the task), to be able to score their performance afterwards.

\section*{Data analysis}

Using the video recordings, each participant's performance on the test tasks was determined by scoring the number of steps correctly executed. For the first test task, this resulted in a maximum score of 15 , for the transfer task, in a maximum score of 24 . For two participants, performance scores were lost due to a technical recording error and two participants failed to fill out an effort rating. Because initial explorative analyses showed that the performance on the test tasks was not normally distributed, a log transformation was conducted (Field, 2009).

\section*{Results}

Data were analyzed using \(2 \times 2\) ANOVAs with betweensubjects factors Training (Practice vs. Example) and Embodiment (Present vs. Absent). For all analyses a significance level of .05 was used and Cohen's \(d\) is reported as a measure of effect size, with \(0.20,0.50\), and 0.80 constituting small, medium, and large effects, respectively.

\section*{Effort Invested in the Learning Phase}

There was a significant main effect of Training on mental effort invested in the learning phase \(F(1,74)=102.09, M S E\) \(=3.08, p<.001\), Cohen's \(d=2.31\), with participants who studied the video-based modeling examples reporting much lower effort ( \(M=2.94, S D=1.63\) ) than participants who practiced problem solving ( \(M=7.00, S D=1.87\) ). There was no significant main effect of Embodiment, nor a significant interaction effect.

\section*{Retention Test Task}

There was a significant main effect of Training, \(F(1,74)=\) \(15.09, M S E=.07, p<.001\), Cohen's \(d=0.87\), which indicated that participants in the Example conditions outperformed ( \(M=0.79, S D=0.30\); non-transformed: \(M=\) \(6.74, S D=5.44\) ) participants in the practice conditions ( \(M=\) \(0.56, S D=0.22\); non-transformed: \(M=3.13, S D=2.49\) ). Although there was a trend towards an effect of Embodiment, with participants in the Embodiment Absent conditions performing better ( \(M=0.72, S D=0.29\); nontransformed: \(M=5.51, S D=4.78\) ) than participants in the Embodiment Present conditions ( \(M=0.63, S D=0.28\); nontransformed: \(M=4.36, S D=4.36\) ) this did not reach significance, \(F(1,74)=2.35, M S E=.068, p=0.129\), Cohen's \(d=0.30\). There was no significant interaction.
A \(2 \times 2\) ANOVA on invested mental effort on the retention test task, showed a significant main effect of Training, \(F(1,76)=9.63, M S E=5.12, p<.01\), Cohen's \(d=\) 0.70 , indicating that participants who had studied examples invested less mental effort in solving the retention test problem ( \(M=5.22, S D=2.60\) ) than participants who had practiced ( \(M=6.82, S D=1.91\) ). There was no significant main effect of Embodiment \(F(1,76)<1\), nor an interaction effect, \(F(1,76)=2.58, M S E=5.12, p=.113\) and indicated that in the Example conditions, the Embodiment Absent condition tended to invest more effort than the Embodiment Present condition on the retention test task, whereas in the Practice conditions, this was the other way around.

\section*{Transfer Test Task}

There were no significant main or interaction effects on performance and invested mental effort on the transfer test task (all \(F<1\) ).

\section*{Discussion}

In line with our first hypothesis, we found a large ( \(d=\) 0.87 ) beneficial effect of example study on test performance. Moreover, the examples conditions reached this higher test performance with less investment of effort during the learning phase (indicating a more efficient learning process), as well as less investment of effort during the retention test (indicating more efficient learning outcomes; Van Gog \& Paas, 2008). This finding is in line with prior studies in other domains that have shown higher learning outcomes with less investment of mental effort during acquisition (e.g., Paas \& Van Merriënboer, 1994; Van Gog et al., 2006) as well as during the test (e.g., Paas, 1992; Paas \& Van Merriënboer, 1994). This effect was limited to the retention test task, though. There were no effects on transfer, which suggests that students in the Example study conditions remembered the procedure (they performed better on the retention test), but did not really understand it sufficiently to be able to adapt it to a new problem situation with an additional object on each side. It would therefore be interesting to investigate whether including verbal explanations by the model, emphasizing the possible options at each step and indicating why the
eventually chosen step is correct and the others are not, would enhance understanding of the solution procedure and thereby, transfer performance.
Regarding our second hypothesis about effects of Embodiment on test performance, we saw a trend in the expected direction, with participants in the Embodiment Absent conditions performing better than participants in the Embodiment present conditions: practicing or studying examples with animal-like plastic objects led to less steps correctly completed on the retention test than practicing or studying examples with wooden discs. However, this difference failed to reach statistical significance ( \(p=.129 ; d\) \(=0.30\) ), possibly due to the relatively low number of participants. Therefore, we will replicate this study with a larger number of participants.

\section*{Second Study}

We are currently conducting a replication study with a larger number of participants to achieve more statistical power. This study will also include an additional condition in which we will control for the effect of direction. That is, because the animals were embodied, they were also headed in a direction. The discs did not imply any direction. So assuming we would find a significant effect of Embodiment when we have more statistical power, this additional condition will allow us to answer the question of whether this is really due to anthropomorphism (assigning goals and intentions to objects that have a bodily shape) or simply a consequence of implied direction. If so, that would still be an interesting finding in terms of understanding factors that might affect problem solving and the acquisition of problem-solving skills through example study. The results of this second study are expected to be available well before the conference.

\section*{References}

Atkinson, R. K., Derry, S. J., Renkl, A., \& Wortham, D. (2000). Learning from examples: Instructional principles from the worked examples research. Review of Educational Research, 70, 181-214.
Chandler, P. \& Sweller, J. (1991). Cognitive load theory and the format of instruction. Cognition and Instruction, 8, 293-332.
Cooper, G., \& Sweller, J. (1987). The effects of schema acquisition and rule automation on mathematical problemsolving transfer. Journal of Educational Psychology, 79, 347-362.
Epley, N., Waytz, A., \& Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. Psychological Review, 114, 864-886.
Field, A. (2009). Discovering statistics using SPSS (3rd edition). London: Sage Publications.
Goldstone, R. L., \& Son, J. Y. (2005). The transfer of scientific principles using concrete and idealized simulations. The Journal of the Learning Sciences, 14, 69110.

Kalyuga, S., Ayres, P., Chandler, P., \& Sweller, J. (2003).

The expertise reversal effect. Educational Psychologist, 38, 23-32.
Kalyuga, S., Chandler, P., Tuovinen, J., \& Sweller, J. (2001). When problem solving is superior to studying worked examples. Journal of Educational Psychology, 93, 579-588.
Kotovsky, K., Hayes, J. R., \& Simon, H. A. (1985). Why are some problems hard? Evidence from Tower of Hanoi. Cognitive Psychology, 17, 248-294.
Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive load approach. Journal of Educational Psychology, 84, 429434.

Paas, F., Tuovinen, J. E., Tabbers, H., \& Van Gerven, P. W. M. (2003). Cognitive load measurement as a means to advance cognitive load theory. Educational Psychologist, 38, 63-71.
Paas, F., \& Van Merriënboer, J. J. G. (1994). Variability of worked examples and transfer of geometrical problem solving skills: A cognitive-load approach. Journal of Educational Psychology, 86, 122-133.
Renkl, A. (2011). Instruction based on examples. In R. E. Mayer \& P. A. Alexander (Eds.), Handbook of research on learning and instruction (pp. 272ᄀ295). New York: Routledge.
Salden, R. J. C. M., Koedinger, K. R., Renkl, A., Aleven, V., \& McLaren, B. M. (2010). Accounting for beneficial effects of worked examples in tutored problem solving. Educational Psychology Review, 22, 379-392.
Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12, 257-285.
Sweller, J., \& Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. Cognition and Instruction, 2, 59-89.
Sweller, J., \& Levine, M. (1982). Effects of goal specificity on means-ends analysis and learning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 8, 463-474.
Sweller, J., Van Merriënboer, J. J. G., \& Paas, F. (1998). Cognitive architecture and instructional design. Educational Psychology Review, 10, 251-295.
Tarmizi, R., \& Sweller, J. (1988). Guidance during mathematical problem solving. Journal of Educational Psychology, 80, 424-436.
Van Gog, T. (2011). Effects of identical example-problem and problem-example pairs on learning. Computers \& Education, 57, 1775-1779.
Van Gog, T., Jarodzka, H., Scheiter, K., Gerjets, P., \& Paas, F. (2009). Attention guidance during example study via the model's eye movements. Computers in Human Behavior, 25, 785-791.
Van Gog, T., \& Paas, F. (2008). Instructional efficiency: Revisiting the original construct in educational research. Educational Psychologist, 43, 16-26.
Van Gog, T., Paas, F., \& Van Merriënboer, J. J. G. (2006). Effects of process-oriented worked examples on troubleshooting transfer performance. Learning and

Instruction, 16, 154-164.
Van Gog, T., \& Rummel, N. (2010). Example \(ᄀ\) based learning: Integrating cognitive and social \(\rightarrow\) cognitive research perspectives. Educational Psychology Review, 22, 155-174.
Zhu, X., \& Simon, H. A. (1987). Learning mathematics from examples and by doing. Cognition and Instruction, 4, 137-166.

\title{
Using Recognition in Multi-Attribute Decision Environments
}

\author{
Don van Ravenzwaaij (d.vanravenzwaaij@unsw.edu.au), Ben R. Newell (ben.newell@unsw.edu.au), and Chris P. Moore (christophermoore@gmail.com)
}

School of Psychology,
University of New South Wales, Sydney 2052, Australia

\author{
Michael D. Lee (mdlee@ uci.edu) \\ Department of Cognitive Sciences, University of California Irvine, Irvine, CA, 92697-5100, USA
}

\begin{abstract}
An experiment examined the effect of 'pure' recognition in the absence of concomitant evaluation - on inferences. In the first stage of the experiment, participants indicated whether they recognized a number of Italian and US cities. In the second stage, they decided which of two cities had the larger population. Crucially, names of the cities were not available in the second stage, but participants could find out whether they had recognized them (yes/no) in the first stage of the experiment (i.e., pure recognition). Additional predictive cues (e.g., presence/absence of a university) were also available. Participants used the recognition cue about \(50 \%\) of the time, rarely examined it first, and used it differently as a function of whether recognition information was binary or continuous. Furthermore, participants used the recognition cue more often if they recognized more items, irrespective of its predictive validity. Implications for theoretical frameworks that view inference as driven by discrete heuristics or processes of evidenceaccumulation are briefly discussed.
\end{abstract}

Keywords: Inference, heuristics, recognition, decision making.

Humans are decision makers. Throughout our lifes, we are constantly confronted with situations that force us to make a choice. Whether it is a preference decision, "Do I take the car or do I walk to work?", or a knowledge decision, "Which soccer team scored more goals last season, Borussia Dortmund or Bayern München?", we are evaluating alternatives.

Gigerenzer and colleagues have proposed a number of relatively simple heuristics that could help us making such decisions. In this paper, we will focus on one of the most prominent examples: the recognition heuristic (Goldstein \& Gigerenzer, 1999, 2002; Gigerenzer \& Goldstein, 2011).

In the original conceptualisation of the recognition heuristic, called take-the-best (TTB; Gigerenzer \& Goldstein, 1996), the first step in deciding which of two response options to choose was to use recognition. So, if a decision maker knows the Bayern München soccer team, but has never heard of Borussia Dortmund, then respond that Bayern München scored more goals last season. When both teams are recognized (thus disabling the use of recognition) the heuristic consults relevant information, or cues, in memory that are indicative of the number of goals scored (e.g., "What was the team's final standing in the national competition?"). These cues should be consulted in descending order of informativeness, starting with the cue that will be most indicative of the
criterion of interest (i.e., number of goals scored). Cue search stops when the decision maker examines a cue that points in one direction (i.e., Borussia Dortmund was first last season, Bayern München was second, so respond Borussia Dortmund).

This proposal for a simple mechanism based on recognition sparked a wide ranging debate about the plausibility, empirical validity, and generality of the recognition heuristic (for recent examples see the papers in the three special issues of the Journal of Judgment and Decision Making - Vol 6 (1) \& (5), 2011; Vol 5 (4), 2010). Much of the debate revolves around some key assumptions about the nature and operation of recognition in inferential judgment.

In the paper that introduced the recognition heuristic as a stand-alone 'tool' (i.e. not just the first step in Take-theBest), Goldstein and Gigerenzer (2002) assume, firstly, that recognition is binary. That is, we either recognize something, or we do not, and there is no room within the heuristic for the distinction between something being vaguely familiar and something being very familiar. Secondly, recognition is assumed to be noncompensatory. That is, when we recognize one option, but do not recognize the other, then we should always go with the recognized option, regardless of any additional information. Lastly, Goldstein and Gigerenzer (2002) make a distinction between familiarity and recognition: "The term familiarity is typically used in the literature to denote the degree of knowledge (or amount of experience) a person has of a task or object. The recognition heuristic, in contrast, treats recognition as a binary, all-or-none distinction; further knowledge is irrelevant." (pp. 77). Thus, according to a strict interpretation of the ( 2002 version of the) recognition heuristic, when deciding whether an Italian city you know has a larger population than an Italian city you do not know, it makes no difference whether the city you do know is Rome or Pisa.

All three of these assumptions have been roundly challenged in the literature on both empirical (e.g., Pohl, 2006; Newell \& Shanks, 2004; Newell \& Fernandez, 2006) and theoretical grounds (e.g., Hilbig, 2010; Newell, 2011). Responding to some of these critiques, Gigerenzer and Goldstein (2011) recast the adaptive use of the recognition heuristic as involving a two-step process: first recognition ("Do I
recognize one object but not the other?") and second, evaluation ("If so, is it reasonable to rely on the recognition heuristic in this situation?"). A view consistent with that is outlined in Newell and Shanks (2004).

While such a conceptualisation is undoubtedly more plausible, it makes the claims about the way recognition aids inference that much more difficult to define and test empirically. Perhaps the trickiest aspect of the problem is that recognition almost always entails further information about the recognized object. If you have heard of Pisa, it is highly likely that you know something else about it (e.g., that it has a leaning tower) which may or may not be relevant to the criterion of interest, in this case population (cf. Oppenheimer, 2003). In other words, it is difficult to isolate the influence of 'pure' recognition - how useful is just knowing that I recognize an object for drawing an inference?

Isolating this 'pure' recognition - recognition without concomitant evaluation - is important because it can shed new light on the distinction between recognition and familiarity, and the extent to which people will rely on recognition even when they cannot directly evaluate their reason(s) for recognising an object. In order to isolate pure recognition we introduced a novel element to the standard task in which participants decide which of two cities has the larger population (e.g., Goldstein \& Gigerenzer, 2002). In our task we created the distinction between recognition and familiarity alluded to by Goldstein and Gigerenzer (2002) by first asking participants to provide recognition data about a pool of response options (city names). We then presented participants with a series of forced choice decisions between two cities about which different pieces of information could be obtained (e.g., presence/absence of a university), but for which the city names were unavailable. Although participants could not discover the names they could - crucially - discover whether or not they had recognized one, both or neither of the cities when they had been presented in the first stage of the experiment. This information was available in the same manner as all the other cues - that is via clicking on relevant buttons ("Did you recognize this city when you were shown its name?") to reveal a yes/no answer (see Figure 1).

A key question here is: How often and when do participants examine the 'pure' recognition cue when drawing an inference? Will recognition remain a primary driver of decisions (cf., Pachur \& Hertwig, 2006) even in the absence of evaluation? In a sense, the use of recognition in this task allows us to gain insight into participants' meta-cognitions about the usefulness of recognition in different environments. For example, do decision makers use recognition more often as they recognize more items in a pool of response categories, irrespective of the informativeness of recognition? To facilitate examining these questions we presented each participant with two decision environments in which we assumed he or she would know a different proportion of the items, thereby allowing us to directly compare response strategies: a US cities environment and an Italian cities environment.

An additional feature of the experiment was that we offered participants (between-subjects) the opportunity to use recognition as a binary (yes/no) or a continuous (slider from 0 to 100) cue. If it is true that recognition operates in a binary fashion, participants should only use the endpoints of a scale when asked to give a continuous rating of their recognition. Similarly, the usage of the recognition cue should not differ between a condition in which it was indicated as binary and a condition in which it was indicated as continuous.

In the next section, we will describe the experiment and each of its conditions in greater detail. Then, we will present some results and conclude on both the tenability of the recognition heuristic, and the use of recognition as an aid to inference more generally.

\section*{Method}

\section*{Participants}

All participants were first year undergraduate students at the University of New South Wales who participated in return for course credit. A total of 100 participants ( 62 females, 38 males), aged 17 to 39 (mean \(=19.5, \mathrm{SD}=2.9\) ) took part in the experiment. They were randomly divided between four between-subject conditions ( \(n=25\) each).

\section*{Material}

The tasks we used for this experiment are based on the German cities task (Gigerenzer \& Goldstein, 1996) in which over the course of consecutive trials, a participant has to decide which of two cities has the larger population. This decision can be made by extracting information on different cues in any order. An example of such a cue could be "Is this city the national capital?". Rather than using German cities, we administered an Italian cities environment and a US cities environment to each participant (see Lee \& Zhang, 2012), expecting to see higher recognition percentages for the US cities than for the Italian cities.

Table 1: The nine cues as used in the Italian and US cities environments. Env = Environment, Val \(=\) Cue Validity, Dis \(=\) Cue Discriminability.
\begin{tabular}{llllll}
\hline Env & Nr. & Cue & Val & Dis \\
\hline Italy & 1 & Is the city the national capital? & 1 & 0.04 \\
& 2 & Dos the city have a railway station? & 0.92 & 0.36 \\
& 3 & Is the city a regional capital? & 0.84 & 0.38 \\
& 4 & Does the city have a football team in the Serie A league? & 0.81 & 0.36 \\
& 5 & Does the city have a university? & 0.80 & 0.55 \\
& 6 & Does the city have an airport? & 0.76 & 0.49 \\
& 7 & Does the city have a football team in the Serie B league? & 0.70 & 0.30 \\
& 8 & Is the city in the Po Valley? & 0.60 & 0.52 \\
& 9 & Did you recognize this city when you were shown its name? & varies & varies \\
\hline US & 1 & Does the city have an airport? & 0.78 & 0.51 \\
& 2 & Dos the city have a sport team? & 0.74 & 0.53 \\
& 3 & Does the city have a metro? & 0.74 & 0.23 \\
& 4 & Does the city have an exposition site? & 0.73 & 0.26 \\
& 5 & Is the city the national capital? & 0.67 & 0.03 \\
& 6 & Does the city have a railway station? & 0.66 & 0.35 \\
& 7 & Is the city a state capital? \\
& 8 & Did you recognize this city when you were shown its name? & 0.59 & 0.34 \\
& & & & \\
\hline
\end{tabular}

In the first stage of the experiment, participants indicate
whether or not they recognize each of the cities used in the subsequent stage of the experiment. In a between participants manipulation, recognition was either measured as a binary or as a continuous variable. Taking the Italian environment as an example, in the binary condition participants were asked "Do you recognize this city in Italy" for a total of 66 Italian cities with response options "yes" or "no". In the continuous condition, participants were asked "How well do you recognize this city in Italy". Answers were indicated on a slider going from 0 ("I am certain that I do not recognize this city") through 50 ("I am not sure whether or not I recognize this city") to 100 ("I am certain that I recognize this city").

In the second stage of the experiment, participants were asked "Which Italian city has a higher population?". The participant could choose between A and B, both of which represented Italian cities the participant had provided recognition data on in the first stage of the experiment. As noted in the introduction, our key focus is on investigating pure recognition without associated knowledge of the response options and thus we effectively disabled internal memory-based search by concealing the names of each city. In order to aid the decision making process, participants were presented with a number of cues on screen for which they can retrieve information. \({ }^{1}\) Crucially, one of the cues the participant could access was the recognition cue of which data was provided in the first stage of the experiment. A screenshot of the second stage of the experiment is provided in Figure 1.

The final manipulation in our experiment consisted of the availability of information on two key aspects of each cue. These are each cues' validity and discriminability. The validity of a cue quantifies the number of times a cue points you to the right answer as a ratio of the times it discriminates between the two response options. For instance, in the Italian cities Environment, the cue "Is the city the national capital?" has a validity of 1 , because whenever one alternative scores positive on this cue, that will be because that alternative is the city Rome and Rome is the largest Italian city. The discriminability of a cue quantifies the number of times a cue discriminates between two response alternatives as a ratio of all possible cue comparisons for each question. The national capital cues does not discriminate very often and therefore has a low discriminability, because this cue will only discriminate when one of the two response alternatives is Rome. In the "+info" condition, cue validity and discriminability was shown on screen, in the "-info" condition, this information was not available to the participant. Note that for the continuous recognition cue, the cue discriminates if both scores are different from eachother. Thus, if one cue scores 0 , it makes no difference whether the other scores 1 or 100. This manipulation was included to examine whether provision of information about the usefulness of recognition, in particular, affected its use. The validity and discriminabilty information can be seen as an aid to answering the meta-cognitive ques-

\footnotetext{
\({ }^{1}\) This is different from the original German cities task, in which cues had to be retrieved from memory and city names were revealed.
}
tion facing the participant - i.e., how useful is knowing that I recognize an object for drawing an inference?

All cues and their validities and discriminabilities for both environments are shown in Table 1. These cue validity and cue discriminability rates were calculated for the subset of 100 comparisons the participants had to make in the task, rather than for the whole set of possible comparisons. The reason for this was to ensure that participants in the "+info" condition could relate the presented cue validity and cue discriminability rates as close as possible to their actual experience when performing the task. The presented information could be used by participants to base their search order on cue validity, cue discriminability or a combination of the two. After each trial, participants received feedback with respect to the accuracy of their response. The experiment was selfpaced.


Figure 1: Screenshot of a trial of the binary version of the US cities task with cue information present ("+info"). See text for details.

Cues were presented in a circular array on the screen in random order. Participants examined cues by clicking on them. The order in which buttons were clicked was selfcontrolled. Deciding to stop examining additional cues was self-controlled, but conditional on having encountered at least one discriminating cue to dissuade guessing.

\section*{Procedure}

Participants completed the Italian version and the US version of the task in random order. Participants were given instructions that they would have to indicate whether or not they recognized a number of cities, after which they performed stage 1 of the experiment, the recognition phase. Participants were subsequently instructed that they repeatedly had to make a choice between pairs of two alternatives. The concepts cue validity and cue discriminability were explained. Participants then performed the second stage of the experiment. After completing the experiment for the first environment, the second environment was administered.

\section*{Design}

Our experiment consists of eight conditions. The cities environment was a within-subject manipulation with two levels: Italian and US. The recognition mode was a between-subject manipulation with two levels: binary and continuous. Cue information was a between-subject manipulation with two levels: +info (info present on screen) and -info (no info on screen).

\section*{Results}

For all statistical analyses, we report not only conventional \(p-\) values but also Bayes factors (e.g., Jeffreys, 1961). In contrast to \(p\)-values, Bayes factors allow researchers to quantify evidence in favor of the null hypothesis vis-a-vis the alternative hypothesis. For instance, when the Bayes factor \(B F_{01}=10\) the observed data are 10 times more likely to have occurred under \(H_{0}\) than under \(H_{1}\). When \(B F_{01}=1 / 5=0.20\) the observed data are 5 times more likely to have occurred under \(H_{1}\) than under \(H_{0}\). In the following, Bayes factors for analysis of variance are based on the BIC approximation (e.g., Wagenmakers, 2007; Masson, 2011), and Bayes factors for \(t\)-tests are based on the default Bayesian \(t\)-test proposed by Rouder, Speckman, Sun, Morey, and Iverson (2009).

We ran a \(2 \times 2 \times 2\) ANOVA with mode and cue information as between-subject independent variables and environment as a within-subject independent variable. Response accuracy was higher in the Italian environment (73.9\%) than in the US environment ( \(69.2 \% ; F(1,96)=78.5, p<.05, B F_{01}=1.1\). \(10^{-12}\) ). The following subsections report on the recognition proportion, the recognition validity and discriminabillity, and the recognition usage respectively.

\section*{Recognition Proportion}

Figure 2 shows the proportion of cities that were recognized for each environment. As expected, recognition was higher for the US cities environment than for the Italian cities environment, as evidenced by a main effect for environment \(\left(F(1,96)=298.9, p<.05, B F_{01}=2.0 \cdot 10^{-30}\right.\); cf., Goldstein \& Gigerenzer, 2002). Continuous recognition led to some parts of the scale being used besides the two extremes, suggesting that participants did not treat recognition as purely binary. However, the extremes were still the most popular.

\section*{Recognition Validity and Discriminability}

Recall that the validity and discriminability of the recognition cue was calculated for each participant separately based on their answers in stage 1 of the experiment. Based on the recognition proportion for each environment, we expected to find that recognition was more valid, but less discriminating, for the Italian environment than for the US environment (cf., Goldstein \& Gigerenzer, 2002). We were interested to see how cue mode would affect cue validity and discriminability.

Figure 3 shows recognition validity and discriminability for both environments. For recognition validity, there is a main effect for environment \(\left(F(1,96)=4.9, p<.05, B F_{01}=\right.\) 0.82 ; note that the Bayesian test indicates the evidence is


Figure 2: Proportion of cities recognized for the binary environment (top-left panel) and the continuous environment (other panels).
ambiguous); recognition validity may be higher in the Italian environment than in the US environment. There is also a tentative main effect for mode \((F(1,96)=4.6, p<.05\), \(B F_{01}=0.99\); note that the Bayesian test indicates the evidence is ambiguous); validity for binary cues may be higher than for continuous cues.

For recognition discriminability, there is a main effect for environment \(\left(F(1,96)=57.8, p<.05, B F_{01}=5.8 \cdot 10^{-10}\right)\); recognition discriminability is lower in the Italian environment than in the US environment. There is also a main effect for mode \(\left(F(1,96)=70.4, p<.05, B F_{01}=1.1 \cdot 10^{-11}\right)\); continuous cues discriminate better than binary cues.


Figure 3: Recognition validity (top) and discriminability (bottom) for the Italian (left) and US (right) environments.

We have established there are differences in the validity and discriminability of the recognition cue that are a direct consequence of the cue being binary or continuous: recognition discriminates between response alternatives more often, but the extra information is, tentatively, less valid. Do decision makers use the recognition cue differently depending on the mode of the cue?

\section*{Recognition Usage}

The top panels of Figure 4 show the proportion of trials the recognition cue was used for each environment. On average, participants did not use the recognition cue on all trials. On an individual basis, \(9 \%\) of the participants used the recognition cue on all trials in the Italian environment and \(8 \%\) of the participants used the recognition cue on all trials in the US environment.

For recognition use, there was a main effect of mode \(\left(F(1,96)=9.4, p<.05, B F_{01}=0.09\right)\); decision makers use continuous recognition more than binary recognition. There was also a main effect of environment \((F(1,96)=22.8\), \(p<.05, B F_{01}=2.4 \cdot 10^{-4}\) ), decision makers use the recognition cue more in the US environment than in the Italian environment. Interestingly, there was a mode by environment interaction \(\left(F(1,96)=4.0, B F_{01}=1.33\right)\). In the Italian environment, recognition is used more often if it is continuous than if it is binary \(\left(t(98)=-3.65, p<.05, B F_{01}=0.02\right)\). In the US environment, recognition usage does not depend on the mode of the cue \(\left(t(98)=-1.52, p>.05, B F_{01}=2.21\right.\); note that the Bayesian test indicates the evidence is somewhat ambiguous). It is likely then, that the benefits of continuous recognition are highest when only a small portion of items are recognized. Finally, there is little evidence for a main effect of cue information \(\left(F(1,96)=3.3, p>.05, B F_{01}=1.88\right.\); but note that the Bayesian test indicates the evidence is somewhat ambiguous).

The bottom panels of Figure 4 show the average position in which the recognition cue was searched, given that it was examined for each environment. On average, participants did not search the recognition cue first. On an individual basis, for both environments, the lowest mean position of examination for the recognition cue was exactly 2 , suggesting that not a single individual used the recognition heuristic in its most stringent form.

For recognition position, there was a main effect of mode \(\left(F(1,95)=4.5, p<.05, B F_{01}=0.08\right)^{2}\); recognition was used earlier when it was binary than when it was continuous. There was no main effect of environment \((F(1,94)=2.1, p>.05\), \(B F_{01}=3.27\) )

\section*{Conclusion}

The goal of our experiment was to isolate 'pure' recognition and to examine participants' use of recognition information in the absence of concomitant evaluation. We argued that this would give us insight into participants' meta-cognition about the usefulness of recognition in different environments. What have we learned?

First we note that the accuracy of inferences about population size was higher for an environment about which participants, initially, knew less (Italian cities) than for one about

\footnotetext{
\({ }^{2}\) Two participants never used the recognition cue and as such had no recognition position data. Recognition position was divided by the total number of cues for each environment to make both environments compatible.
}


Figure 4: Proportion of trials the recognition cue was used (top) and position in search order (bottom) for the Italian (left) and US (right) environments.
which they knew more (US cities). But was the difference driven by adaptive use of recognition information? Under a strict interpretation, even in our novel task, recognition should be consulted on every trial - unless a participant has reason to believe that it will never discriminate (e.g., if they know they recognized either all or none of the cities in the environment). The results showed that was clearly not the case, with only a small proportion of the participants using the recognition cue on all trials. Moreover, even the participants that did examine recognition on every trial did not exclusively examine this cue first, challenging the idea that recognition information is somehow privileged in inference tasks (e.g., Pachur \& Hertwig, 2006).

Additionally, we examined whether we could increase usage of recognition by measuring recognition on a continuous rather than a binary scale. We concluded that for the Italian environment where only a small proportion of the items were recognized, measuring recognition on a continuous scale led to recognition being used more often, despite the fact that recognition was less valid for the continuous scale than for the binary scale. No such effect was found for the US environment, in which on average about half of the cities were recognized. Though it is possible that the intermediate datapoints on the continuous recognition scale simply reflect perceived task demands by the participants, this alternative explanation does not seem to be in line with the fact that participants subsequently use continuous recognition more.

Our final question was whether participants would use recognition more often when their recognition cue was more valid. Surprisingly, we concluded the opposite: recognition was used more frequently in the US environment than in the Italian environment, despite the fact that recognition was more valid, on average, in the Italian environment. This finding suggests that meta-cognition about the usefulness of recognition is not particularly fine-tuned: adaptive use of recognition would predict greater reliance in environments where it is more useful (Gigerenzer \& Goldstein, 2011).

Thus we conclude that 'pure' recognition can be compensated: knowing that we recognized one object and not another, but not knowing why is not enough for most participants to make a decision. Furthermore, it is not the first piece of information participants search for. In addition, our results show that recognition is more than a binary yes-or-no phenomenon. Allowing participants to indicate their recognition on a continuous scale led to an average increase in usage.

These results provide a novel and intriguing set of empirical regularities concerning the use of pure recognition information in a multi-attribute decision task. The next stage of this project will draw on the considerable advances that have been made in developing computational models of recognition-based judgments (e.g., Marewski \& Mehlhorn, 2011; Marewski \& Schooler, 2011) in an attempt to describe these data more fully. Our starting point will be to compare models that assume recognition information is used as evidence that can be accumulated much like any other cue to aid inference (e.g., van Ravenzwaaij, Moore, Lee, \& Newell, 2013) with those that afford recognition an elevated status.

\section*{Acknowledgments}

The assistance of the Australian Research Council is gratefully acknowledged (Grants DP110100797 and FT110100151).

\section*{References}

Gigerenzer, G., \& Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. Psychological Review, 103, 650-669.
Gigerenzer, G., \& Goldstein, D. G. (2011). The recognition heuristic: A decade of research. Judgment and Decision Making, 6, 100-121.
Goldstein, D. G., \& Gigerenzer, G. (1999). The recognition heuristic: How ignorance makes us smart. In G. Gigerenzer, P. M. Todd, \& the ABC Research Group (Eds.), Simple heuristics that make us smart (pp. 37-58). New York: Oxford University Press.
Goldstein, D. G., \& Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. Psychological Review, 109, 75-90.
Hilbig, B. E. (2010). Reconsidering "evidence" for fast-and-frugal heuristics. Psychonomic Bulletin \& Review, 17, 923-930.
Jeffreys, H. (1961). Theory of probability. Oxford, UK: Oxford University Press.
Lee, M. D., \& Zhang, S. (2012). Evaluating the coherence of Take-the-best in structured environments. Judgment and Decision Making, 7, 360-372.
Marewski, J. N., \& Mehlhorn, K. (2011). Using the ACTR architecture to specify 39 quantitative process models of decision making. Judgment and Decision Making, 6, 439519.

Marewski, J. N., \& Schooler, L. J. (2011). Cognitive niches: An ecological model of strategy selection. Psychological Review, 118, 393-437.

Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. Behavior Research Methods, 43, 679-690.
Newell, B. R. (2011). Recognising the recognition heuristic for what it is (and what it's not). Judgment and Decision Making, 6, 409-412.
Newell, B. R., \& Fernandez, D. (2006). On the binary quality of recognition and the inconsequentiality of further knowledge: Two critical tests of the recognition heuristic. Journal of Behavioral Decision Making, 19, 333-346.
Newell, B. R., \& Shanks, D. R. (2004). On the role of recognition in decision making. Journal of Experimental Psychology: Learning, Memory, and Cognition, 30, 923-935.
Oppenheimer, D. M. (2003). Not so fast! (and not so frugal!): Rethinking the recognition heuristic. Cognition, 90, B1B9.
Pachur, T., \& Hertwig, R. (2006). On the psychology of the recognition heuristic: Retrieval primacy as a key determinant of its use. Journal of Experimental Psychology: Learning, Memory, and Cognition, 32, 983-1002.
Pohl, R. F. (2006). Empirical tests of the recognition heuristic. Journal of Behavioral Decision Making, 19, 251-271.
Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., \& Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. Psychonomic Bulletin \& Review, 16, 225-237.
van Ravenzwaaij, D., Moore, C., Lee, M., \& Newell, B. R. (2013). How to collect information in uncertain decision environments? A hierarchical Bayesian modeling approach. Manuscript submitted for publication.
Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of \(p\)-values. Psychonomic Bulletin \& Review, 14, 779-804.

\title{
The Path Less Taken: When Working Memory Capacity Constrains Insight
}

\author{
Charles Van Stockum (charles.vanstockum@louisville.edu) \\ Department of Psychological and Brain Sciences, University of Louisville \\ Louisville, KY 40292 USA \\ Marci S. DeCaro (marci.decaro@louisville.edu) \\ Department of Psychological and Brain Sciences, University of Louisville Louisville, KY 40292 USA
}

\begin{abstract}
Higher working memory capacity (WMC) supports performance on a wide variety of complex cognitive and academic activities (Barret, Tugade, \& Engle, 2004). However, a growing body of research demonstrates that higher WMC can have disadvantages-leading individuals to employ complex performance strategies that are less optimal for a given task (cf. DeCaro \& Beilock, 2010). We examine this possibility in the domain of insight problem solving. Participants \((N=84)\) completed Matchstick Arithmetic problems thought to either rely on controlled search and retrieval processes (non-insight problems) or diverging from known mathematical constraints (insight problems). Consistent with a large body of research on WMC, higher WMC was associated with higher non-insight problem accuracy. However, higher WMC led to significantly worse insight problem-solving. Although higher WMC supports complex problem-solving strategies, relying on these may lead individuals to miss associatively-driven solutions that are important for insight.
\end{abstract}

Keywords: Working memory capacity; attention; insight; problem solving.

\section*{Introduction}

A great deal of work has demonstrated that higher working memory capacity (WMC) is advantageous to an array of complex cognitive and academic activities, such as reasoning, comprehension, and problem-solving (see Barrett, Tugade, \& Engle, 2004, for a review). Indeed, WMC-the ability to hold and manipulate information in a temporary active state-has been said to be "so central to human cognition that it is hard to find activities where it is not involved" (Ericsson \& Delaney, 1999, p. 259). However, a growing body of research demonstrates that higher WMC can have disadvantages-leading individuals to employ complex performance strategies that are less optimal for a given task (see DeCaro \& Beilock, 2010, for a review). In the current work, we examine the possibility that higher WMC can hinder creative thinking in the form of insight problem-solving. Specifically, we examine the hypothesis that those who have the ability to implement complex problem-solving strategies may be more likely to miss associatively-driven solutions that are important for insight.

\section*{Working Memory Capacity}

WMC supports the ability to suppress distractors and guide attention toward relevant information in goal-directed tasks (McCabe, Roediger, McDaniel, Balota, \& Hambrick, 2010). The predictive power of WMC as a construct stems from this domain-general capacity for attentional control, and individual differences in WMC emerge primarily when that capacity is challenged (Engle, 2002). So-called "executive attention" is accomplished via controlled processing, which is important in novel or interference-rich situations and when goals come in conflict with prepotent responses (Unsworth \& Engle, 2007).

A large body of research has been built around the wellestablished differences in performance outcomes of individuals who fall toward either extreme of the WMC scale. Kane and Engle (2000) found that individuals with lower WMC demonstrated a greater vulnerability to proactive interference, and were more likely to lose track of task goals than their higher WMC counterparts (Unsworth \& Engle, 2007). Additionally, studies have found that individuals with lower WMC display higher rates of attentional capture (Conway, Cowan, \& Bunting, 2001; Kane, Bleckley, Conway, \& Engle, 2001), and have greater difficulty discriminating relevant and irrelevant information (Unsworth \& Engle, 2007).

It is no surprise that the ability of higher WMC individuals to control attention leads to greater ability to implement more difficult, multi-step problem-solving strategies (Engle, 2002). Indeed, the ability to execute complex strategies may lead individuals to select strategies in line with their ability-even if the task does not call for a controlled processing approach. Beilock and DeCaro (2007) explored this idea by examining the strategy selection of higher and lower WMC individuals completing Luchins’ (1946) water jug task. This task requires individuals to use three depicted water jugs with varying capacities (e.g., Jug A=23, Jug B=96, and Jug C=3) to fill a "goal" jug with a certain capacity (e.g., 67). For example, one might fill Jug \(B\), then pour that amount into Jug \(A\), and then pour the remaining amount into Jug C twice (i.e., B-A-2C). Participants were explicitly instructed to mentally derive the answers (i.e., without the use of paper), and use the simplest strategy possible. The first few problems were solve-able using a single complex formula (B-A-2C). The final few problems could also be solved using this formula (e.g., Jug

A=34, Jug B=72, Jug C=4; Goal=30). However, a much simpler strategy could also be applied (e.g., A-C). On these final problems, individuals with higher WMC were more likely to employ the complex algorithmic strategy (i.e., B-A-2C), even though more efficient strategies were available. Individuals with lower WMC were instead quicker to abandon an algorithmic approach and adopt a lessdemanding shortcut strategy relying on a more diffuse focus of attention.

These findings demonstrate that individuals higher in WMC may tend to use more complex strategies even when simpler ones are more efficient for a given task. Such overreliance on complex strategies can harm performance on some tasks (e.g., Gaissmaier, Schooler, \& Rieskamp, 2006; Wolford, Newman, Miller, \& Wig, 2004). For example, when associative responses guide well-learned skill execution, as with proceduralized tasks, controlled attention can disrupt performance (DeCaro, Thomas, \& Beilock, 2008). Additionally, various situational and taskspecific factors such as performance pressure (Beilock \& DeCaro, 2007; DeCaro, Thomas, Albert, \& Beilock, 2011) and expertise (Ericsson \& Delaney, 1999; Wiley, 1998) have been shown to moderate the role of controlled processing in learning and performance situations. A better understanding of when and why less WMC can prove advantageous is necessary to fully grasp the limitations of this pervasive system of cognitive constraint.

\section*{Insight Problem-Solving}

An area in which this question is being explored with great interest is research in insight problem-solving. The link between WMC and insight is not well understood, and there has been much debate over how best to facilitate the type of creative thinking insight problem-solving requires. One approach looks at the role of attention in problem-solving.

Insight problems require the use of strategies that diverge from obvious approaches, and are supported by a more diffuse focus of attention (Ansburg \& Hill, 2003). Noninsight problems, conversely, are best solved by following a progressive series of analytic steps, which requires controlled processing and relies on WMC. According to Representational Change Theory (Ohlsson, 1992), insight problems generally trigger an inadequate mental representation of the problem situation and solution criteria. Explicit search processes reinforce this faulty representation, and are unlikely to lead to the correct solution path. Instead, unsuccessful solution attempts often result in impasse, a state characterized by an apparent dearth of viable problem operators. It is only through a reappraisal of the initial representation that the correct solution path becomes accessible to the solver, often in a sudden and transparent manner (Kounios et al., 2006; Kounios \& Beeman, 2009; but see also Ash, Cushen, \& Wiley, 2009).

Much of the research on insight problem-solving has focused on questions surrounding the phenomenon of impasse, specifically why impasse occurs and how it is overcome (Ohlsson, 1992; Jones, 2003). One explanation is
that the problem solver unwittingly imposes unnecessary and/or misguided constraints on the problem space (Knoblich, Ohlsson, Haider, \& Rhenius, 1999). Additionally, preoccupation with more familiar problem operators (i.e., ones that have worked in the past) can make it difficult to access more novel operators that are critical for insight (Knoblich, Ohlsson, \& Raney, 2001). To the extent that one continues to implement strategies based on these constraints, one will fail to reach an insight solution (Wiley, 1998).

\section*{Working Memory and Insight Problem-Solving}

Because of their reliance on associatively-driven problemsolving solutions, insight problem-solving may be less benefited by the use of complex, algorithmic problemsolving strategies. Indeed, the use of such strategies can actually hinder the ability to derive a solution (Schooler, Ohlsson, \& Brooks, 1996; Wiley \& Jarosz, 2012). Studies have shown that WMC is related to the ability to solve novel problems and adapt to new situations (Barrett et al., 2004). However, if individuals higher in WMC have a tendency to rely on a more controlled attentional focus and inhibit peripheral information, they may also neglect potentially relevant information held outside of the perceived problem space (cf. Ansburg \& Hill, 2003). Thus, counter-intuitively, one might expect higher WMC individuals to perform worse on insight problem-solving tasks.

Support for this idea comes from a range of studies demonstrating that less focused (i.e., more diffuse) attention benefits insight problem-solving, whereas applying more controlled attention hinders the ability to derive insight solutions. For example, moderate alcohol intoxication both reduces WMC and improves insight problem-solving (Jarosz, Colflesh, \& Wiley, 2012); solving insight problems at one's non-optimal time of day improves performance (Wieth \& Zacks, 2011); and patients with frontal lobe impairment demonstrate better insight-problem accuracy (Reverberi, Toraldo, D'Agostini, \& Skrap, 2005). In contrast, verbalizing the problem steps during solving decreases insight performance, possibly by "overshadowing" insight processes (Schooler et al., 1996).

\section*{Current Study}

The current study examines the role of individual differences in WMC in solving both non-insight and insight problems, using the Matchstick Arithmetic task (Knoblich et al., 1999). Matchstick Arithmetic problems are false arithmetic statements written using matchsticks. The matchsticks represent Roman numerals, arithmetic operators, and equal signs. Each matchstick problem is composed of three roman numerals separated by two arithmetic signs, and has a unique solution consisting of a single move.

Participants were given three types of matchstick arithmetic problems, shown in Figure 1. Standard type (ST) matchstick problems are solved by moving a matchstick
representing a value of 1 from its position in a given roman numeral to a different position in the same or a different numeral on either side of the equal sign. The "I" matchstick is considered a "loose chunk" because it can be removed without invalidating the remaining figure and is easily appended to many others (Knoblich et al., 1999). The simple manipulation of loose chunks in ST problems is consistent with prior knowledge that reordering values in an equation leads to success (Öllinger, Jones, \& Knoblich, 2008). ST problems do not involve impasse (Knoblich et al., 2001), or restructuring (Öllinger et al., 2008), considered defining features of insight problems (Ohlsson, 1992). Consistent with Öllinger et al. (2008), we refer to ST problems as non-insight problems.

Constraint relaxation (CR) matchstick problems require transforming the initial false statement (e.g., III + III = III) into a correct, but tautological, statement by changing the plus sign into an equal sign (III = III = III). Solving CR problems is thought to be achieved by relaxing the constraint that correct arithmetic statements cannot contain more than one equal sign. These are commonly considered insight problems (Knoblich et al., 1999).

Finally, chunk decomposition (CD) problems require the solver to decompose a "tight chunk" in order to identify the decisive move. A tight chunk was defined as a single roman numerical figure composed of two matchsticks that together form a meaningful unit (e.g., V, X). For example, when participants see the incorrect arithmetic statement IV \(=\mathrm{III}+\) VI, they must transpose the V into an X by sliding one matchstick to find the solution IX = III + VI. CD problems are typically considered insight problems. However, findings from these problems do not always correspond to the findings from CR problems, making it difficult to determine if these problems are of the same nature (Knoblich et al., 1999; Knoblich et al., 2001; Öllinger et al., 2008). Thus, although we explored performance on CD problems, we were unable to derive clear hypotheses about the relationship between performance on these problems and WMC.


Figure 1. Example Matchstick Arithmetic Problems

We predicted that higher WMC would be associated with increased non-insight (ST) problem-solving accuracy. However, we predicted the opposite pattern for insight (CR) problems, that higher WMC would lead to lower insight problem-solving accuracy. Such findings would be consistent with a growing body of research demonstrating that more working memory capacity can lead to controlled problem-solving approaches that overshadow more optimal associatively-driven solution paths (Wiley \& Jarosz, 2012).

\section*{Method}

\section*{Participants}

Participants were 84 undergraduate students enrolled in psychology classes (63 female; age \(M=21, S D=4.6\); range 18-46 years). An additional 3 people were excluded from the study because they had been exposed to matchstick arithmetic problems before. One person was excluded for errors on more than 20 percent of the sentence task of the aRspan (Conway et al., 2005). Participants received course credit for participation.

\section*{Materials}

Problem-solving task Participants completed Matchstick Arithmetic problems (Knoblich et al., 1999), consisting of false arithmetic statements written with Roman numerals (I, II, III, etc.), arithmetic operators (,+- ), and equal signs depicted as matchsticks (see Figure 1). Problems were completed on paper. Participants were instructed to transform the initial false arithmetic statement into a true arithmetic statement while adhering to the following rules: (a) only one matchstick can be moved, (b) no matchstick can be discarded, (c) upright sticks and slanted sticks are not interchangeable, and (d) the result must be a correct arithmetic statement. Each matchstick problem was composed of three roman numerals separated by two arithmetic signs, and had a unique solution consisting of a single move. Participants were given eight matchstick arithmetic problems divided across two problem sets containing four problems each. Problems sets were divided into two categories (non-insight; insight) based on the move required for solution: the non-insight problem set consisted of four ST problems, and the insight problem set consisted of 2 CR problems and 2 CD problems. Problem sets were administered in counterbalanced order.

Working memory measure Working memory capacity was measured using the Automated Reading Span task (aRspan; Unsworth, Heitz, Schrock, \& Engle, 2005; Redick et al., 2012). In the aRspan, an attention-demanding processing task is interleaved between items presented for serial recall. Participants are shown a sentence and instructed to judge whether it makes sense or not; then they are shown a letter. After a sequence of sentence-letter strings ranging from 3-7 in length, participants are asked to recall the letters in order. All participants complete a total of 15 sequences of
sentence-letter strings, including 3 of each length, presented in random order. ARspan scores range from \(0-75\), with higher scores denoting greater levels of attentional control (Unsworth \& Engle, 2007). The task takes 15-20 minutes to complete.

\section*{Procedure}

After providing informed consent, participants completed the experimental tasks individually. Participants were first introduced to the problem-solving task, and were given a maximum of 10 minutes to solve each of two sets of problems (i.e., 20 minutes total). After completing both problem sets, participants were given a questionnaire asking about previous experience with the matchstick task. Participants then completed the aRspan on a computer. Finally, participants completed a demographic questionnaire and were debriefed.

\section*{Results}

Preliminary analyses revealed that accuracy on CD problems \((M=.69, S D=.41)\) was positively correlated with accuracy on both non-insight ( \(M=.68, S D=.30\) ), \(r=.32, p\) \(=.003\), and CR type insight problems ( \(M=.13, S D=.34\) ), \(r\) \(=.25, p=.021\). Because the CD problems did not appear discriminatory of either insight or non-insight problem types, they were excluded from further analyses. Accuracy on CR type insight problems was not correlated with accuracy on non-insight problems, \(r=.06, p=.566\), consistent with previous studies using matchstick arithmetic (Knoblich et al., 1999).


Figure 2. Non-insight and insight problem-solving accuracy as a function of working memory capacity. Low and high working memory points are plotted at \(\pm 1\) SD below and above the mean.

We evaluated whether the effect of insight and noninsight problem-solving accuracy depends on WMC using
an ANCOVA, in order to treat WMC as a continuous variable. Problem type (insight versus non-insight) was included as a within-subjects factor. WMC and a WMC \(\times\) problem type interaction term were included in the model as covariates.

A significant main effect of problem type was found, \(F(1\), \(82)=297.39, p<.001, \eta_{\mathrm{p}}^{2}=.78\). There was no main effect of WMC, \(F<1\). The interaction between problem type and WMC was significant, \(F(1,82)=5.65, p=.02, \eta_{\mathrm{p}}^{2}=.06\).

In order to examine the nature of this interaction, followup analyses were conducted using simple regression. As shown in Figure 2, higher WMC was associated with generally better non-insight problem-solving accuracy, although this relationship did not reach significance ( \(B=\) \(.016, S E=.011, p=.153\) ). In contrast, higher WMC was associated with significantly lower CR insight problemsolving accuracy ( \(B=-.013, S E=.006, p=.041\) ).

\section*{Discussion}

The current results support the prediction that less attentional control is better for insight problem-solving. Using the Matchstick Arithmetic task, we found that higher WMC was associated with somewhat better non-insight problem-solving but significantly worse insight problemsolving. The latter finding is counterintuitive in light of a great deal of literature demonstrating that more attentional control contributes to better performance on a range of higher-order cognitive tasks (c.f., Conway et al., 2005). These findings are, however, consistent with a growing body of research finding that lower WMC is advantageous on tasks relying on more associative or procedural processes (DeCaro \& Beilock, 2010).

Although a diffuse focus of attention is important for creative problem-solving processes such as insight (Jarosz \& Wiley, 2012), the relationship between WMC and insight has been inconsistent across studies. For example, Ash and Wiley (2006) found that WMC predicted performance on insight problems when the problems required an extended initial search phase. WMC was not related to performance on insight problems in which the search phase was shorter, presumably leading to impasse and restructuring more quickly (see also Fleck, 2008). These findings lend support to the spontaneous restructuring account of insight, which proposes that a necessary change in an initial problem representation is achieved through automatic processes and therefore does not depend on WMC. This finding converges with other evidence demonstrating that associative and divergent thinking rely on automatic processes that occur outside conscious awareness (Dijksterhuis \& Meurs, 2006).

Spontaneous accounts of insight do not, however, preclude the argument that attentional control may disrupt those processes that are important for restructuring. Associative processes are better for creative problemsolving but, critically, are supported by decreased latentinhibition (Carson, Peterson, \& Higgins, 2003). The abilities that facilitate performance on non-insight problems and are supported by controlled processing may therefore be
inappropriate for solving insight problems, and may harm performance. Too much focus can unnecessarily constrain the problem space, limiting the field of viable operators for solution and hindering the ability to achieve insight. Additionally, an overreliance on complex strategies may contribute to persistence within a faulty problem representation.

Some have proposed alternate routes by which creative solutions are achieved: one that is flexible, associative, and is characterized by lower levels of cognitive control, and another that is persistent, deliberate, and supported by a more focused analytic approach (De Dreu, Nijstad, Baas, Wolsink, \& Roskes, 2012). Which pathway is more readily accessible may depend on interactions between individual difference and task-specific factors. Research in strategy selection suggests that differences in WMC may be an important factor in determining which path an individual is likely to take (e.g., Beilock \& DeCaro, 2007; Gaissmeier et al., 2006). The current results suggest that higher WMC leads individuals to select a more focused analytic approach to insight problem-solving. However, future research is needed to examine the actual problem-solving strategies used by individuals of varying WMC.

Future research should also consider additional factors that may impact success at insight problem-solving, including boundaries to the current results. For example, if it is possible to achieve insight through methodical analytic persistence, then individuals with higher WMC could eventually attain insight-they would just require more time in order to exhaust and reject more obvious solution paths before identifying the correct one. Another factor likely to moderate strategy selection is goal transparency. Insight problems are ambiguous by design, and the challenge of these problems often hinges on this occlusion of decisive task objectives. If individuals with higher WMC know to consider everything as potentially relevant, they may be less likely to filter out important parts of the problem (e.g., Colflesh \& Conway, 2007; Conway, Cowan, \& Bunting, 2001).

Although we demonstrate that higher WMC can lead to lower insight, certain situational factors may therefore improve the ability of higher-capacity individuals to select more appropriate problem-solving strategies. By considering the interaction between individual differences and situational factors on the focus of attention, we may be better able to predict when insightful thinking will be best supported.

\section*{References}

Ansburg, P. I., \& Hill, K. (2003). Creative and analytic thinkers differ in their use of attentional resources. Personality and Individual Differences, 34, 1141-1152.
Ash, I. K., Cushen, P. J., \& Wiley, J. (2009). Obstacles in investigating the role of restructuring in insightful problem solving. The Journal of Problem Solving, 2, 7-42.

Ash, I. K., \& Wiley, J. (2006). The nature of restructuring in insight: An individual differences approach. Psychonomic Bulletin \& Review, 13, 66-73.
Barrett, L., Tugade, M. M., \& Engle, R. W. (2004). Individual differences in working memory capacity and dual-process theories of the mind. Psychological Bulletin, 130, 553-573.
Beilock, S. L., \& DeCaro, M. S. (2007). From poor performance to success under stress: Working memory, strategy selection, and mathematical problem solving under pressure. Journal of Experimental Psychology: Learning, Memory, and Cognition, 33, 983-998.
Colflesh, G. J. H., \& Conway, A. R. A. (2007). Individual differences in working memory capacity and divided attention in dichotic listening. Psychonomic Bulletin and Review, 14, 699-703.
Conway, A. R. A., Cowan, N., \& Bunting, M. F. (2001). The cocktail party phenomenon revisited: The importance of working memory capacity. Psychonomic Bulletin \& Review, 8, 331-335.
Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., \& Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. Psychonomic Bulletin \& Review, 12, 769-786.
DeCaro, M. S., \& Beilock, S. L. (2010). The benefits and perils of attentional control. In M. Csikszentmihalyi \& B. Bruya (Eds.), Effortless attention: A new perspective in the cognitive science of attention and action (pp. 51-73). Cambridge, MA: MIT Press.
DeCaro, M. S., Thomas, R. D., Albert, N. B., Beilock, S. L. (2011). Choking under pressure: Multiple routes to skill failure. Journal of Experimental Psychology: General, 140, 390-406.
DeCaro, M. S., Thomas, R. D., \& Beilock, S. L. (2008). Individual differences in category learning: Sometimes less working memory capacity is better than more. Cognition, 107, 284-294.
De Dreu C, Nijstad B, Baas M, Wolsink I, Roskes M. (2012). Working memory benefits creative insight, musical improvisation, and original ideation through maintained task-focused attention. Personality and Social Psychology Bulletin, 38, 656-669.
Dijksterhuis, A., \& Meurs, T. (2006). Where creativity resides: The generative power of unconscious thought. Consciousness and Cognition, 15, 135-146.
Engle, R. W. (2002). Working memory capacity as executive attention. Current Directions in Psychological Science, 11, 19-23.
Ericsson, K. A., \& Delaney, P. F. (1999). Long-term working memory as an alternative to capacity models of working memory in everyday skilled performance. In A. Miyake, \& P. Shah (Eds.), Models of working memory: Mechanisms of active maintenance and executive control (pp. 257-297). New York: Cambridge University Press.
Fleck, J. I. (2008). Working memory demands in insight versus analytic problem solving. European Journal of Cognitive Psychology, 20, 1-38.

Gaissmaier, W., Schooler, L. J., \& Rieskamp, J. (2006). Simple predictions fueled by capacity limitations: When are they successful? Journal of Experimental Psychology: Learning, Memory \& Cognition, 32, 966-982.
Jarosz, A. F., Colflesh, G. J. H., \& Wiley, J. (2012). Uncorking the muse: Alcohol intoxication facilitates creative problem solving. Consciousness and Cognition, 21, 487-493.
Jones, G. (2003). Testing two cognitive theories of insight. Journal of Experimental Psychology: Learning, Memory, and Cognition, 29, 1017-1027.
Kane, M. J., Bleckley, M. K., Conway, A. R. A., \& Engle, R. W. (2001). A controlled-attention view of workingmemory capacity. Journal of Experimental Psychology: General, 130, 169-183.
Kane, M. J., \& Engle, R. W. (2000). Working memory capacity, proactive interference, and divided attention: Limits on long-term memory retrieval. Journal of Experimental Psychology: Learning, Memory, and Cognition, 26, 333-358.
Knoblich, G., Ohlsson, S., Haider, H., \& Rhenius, D. (1999). Constraint relaxation and chunk decomposition in insight problem solving. Journal of Experimental Psychology: Learning, Memory, and Cognition, 25, 1534-1555.
Knoblich, G., Ohlsson, S., \& Raney, G. E. (2001). An eye movement study of insight problem solving. Memory \& Cognition, 29, 1000-1009.
Kounios, J., \& Beeman, M. (2009). The Aha! moment: The cognitive neuroscience of insight. Current Directions in Psychological Science, 18, 210-216.
Kounios, J., Frymiare, J. L., Bowden, E. M., Fleck, J. I., Subramaniam, K., Parrish, T. B., \& Jung-Beeman, M. J. (2006). The prepared mind: Neural activity prior to problem presentation predicts subsequent solution by sudden insight. Psychological Science, 17, 882-890.
McCabe, D. P., Roediger, III, H. L., McDaniel, M. A., Balota, D. A., \& Hambrick, D. Z. (2010). The relationship between working memory capacity and executive functioning: Evidence for a common executive attention construct. Neuropsychology, 24, 222-243.
Ohlsson, S. (1992). Information processing explanations of insight and related phenomenon. In M. Keane \& K. Gilhooly (Eds.), Advances in the psychology of thinking (pp. 1-44). London: Harvester-Wheatsheaf.
Öllinger, M., Jones, G., \& Knoblich, G. (2008). Investigating the effect of mental set on insight problem solving. Experimental Psychology, 55, 269-282.
Redick, T. S., Broadway, J. M., Meier, M. E., Kuriakose, P. S., Unsworth, N., Kane, M. J., \& Engle, R. W. (2012). Measuring working memory capacity with automated complex span tasks. European Journal of Psychological Assessment, 28, 164-171.
Reverberi, C., Toraldo, A., D’Agostini S., \& Skrap, M. (2005). Better without (lateral) frontal cortex? Insight problems solved by frontal patients. Brain, 128, 28822890.

Schooler, J. W., Ohlsson, S., \& Brooks, K. (1993). Thoughts beyond words: When language overshadows insight. Journal of Experimental Psychology: General, 122, 166-183.
Unsworth, N., \& Engle, R.W. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. Psychological Review, 114, 104-132.
Unsworth, N., Heitz, R. P., Schrock, J. C., \& Engle, R. W. (2005). An automated version of the operation span task. Behavior Research Methods, 37, 498-505.
Wieth, M. B., \& Zacks, R. T. (2011). Time of day effects on problem solving: When the non-optimal is optimal. Thinking \& Reasoning, 17, 387-401.
Wiley, J. (1998). Expertise as mental set: The effects of domain knowledge in creative problem solving. Memory \& Cognition, 26, 716-730.
Wiley, J., \& Jarosz, A. (2012). How working memory capacity affects problem solving. Psychology of Learning and Motivation, 56, 185-227.
Wolford, G., Newman, S. E., Miller, M. B., \& Wig, G. S. (2004). Searching for patterns in random sequences. Canadian Journal of Experimental Psychology, 58, 221228.

\title{
The Role of Shape in Semantic Memory Organization of Objects
}

\author{
Lisanne van Weelden (L.vanWeelden@uvt.nl) \\ Joost Schilperoord (J.Schilperoord@uvt.nl) \\ Marc Swerts (M.G.J.Swerts@uvt.nl) \\ Tilburg centre for Cognition and Communication (TiCC), Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands \\ Diane Pecher (Pecher@fsw.eur.nl) \\ Department of Psychology, Erasmus University, PO box 1738, 3000 DR Rotterdam, The Netherlands
}

\begin{abstract}
Visual information contributes fundamentally to the process of object categorization. The present study investigated whether the degree of activation of visual information in this process is dependent on the situational relevance of this information. We used the Proactive Interference (PI) paradigm. In two experiments, we manipulated the information by which objects could be retrieved from memory: by both semantic and shape information or by shape information only. The pattern of PI-release showed that if objects could be stored and retrieved both by semantic and shape information, then shape information was overruled by semantic information. If, however, semantic information could not be (satisfactorily) used to store and retrieve objects, then objects were stored in memory in terms of their shape.
\end{abstract}

Keywords: Object shape; proactive interference; memory; categorization.

\section*{Introduction}

If we observe a cat-like creature in the zoo, even if it is a type that we have never seen before, we may classify that animal as belonging to the same category as lions, tigers and pumas. Presumably, the reason for doing this is that the observed animal shares some observable properties with those of the other cat-like animals that we remember having seen before. Object categorization is hence a fundamental process in constructing and using our memory, as it helps to organize our knowledge and relate (novel) objects to other objects in order to assign meaning to them.

This process of object categorization is driven by mental representation. When we encounter an object, we create a mental representation based on sensory and semantic information. In order to categorize the object, the mental representation is compared to a mental prototype that represents category members (Rosch \& Mervis, 1975) or to other category exemplars in memory (Nosofsky, 1986). The representations are compared on both sensory and semantic information, however the relative weighting of these two types of information varies across concepts and semantic categories (Humphreys \& Forde, 2001; Warrington \& McCarthy, 1987). For example, the shape of an animal or the color of a fruit might be more important to assign the
object to the correct category than the shape or color of a kettle. In the present study, we investigate the role of sensory features in the categorization of visual objects. We focus on the visual sensory feature shape and investigate whether the relative weighting of shape and semantic information affects the organization of semantic memory.

Barsalou (1999) proposed that sensory information plays a critical role in cognition. According to his Perceptual Symbols theory, perception, action, and cognition share processing mechanisms. He views mental representation as a process of sensory-motor simulation. Central in his theory are perceptual symbols by which a mental representation is defined. A mental representation is constructed of a combination of several perceptual symbols for different components of the concept. This perceptual symbol formation process does not only concern the concept's visual features (e.g., its color, shape, and orientation), but operates as well on other sensory modalities such as audition, haptics, olfaction, and gustation. As such, perceptual symbols are learned through actual experiences with concepts. Modality-specific sensory-motor systems capture such experiences and hierarchical association areas integrate experiences from different modalities. Hence, these association networks represent knowledge of the concept that can be recruited for cognitive processing via the process of simulation (i.e., mental representation).

Evidence supporting the PS theory is provided by work that shows that visual sensory information is indeed activated during language comprehension (e.g., Huettig \& Hartsuiker, 2008; Pecher, Van Dantzig, Zwaan, \& Zeelenberg, 2009; Pecher, Zeelenberg, \& Raaijmakers, 1998; Stanfield \& Zwaan, 2001; Van Dantzig, Pecher, Zeelenberg, \& Barsalou, 2008; Van Weelden, Schilperoord, \& Maes, in press; Zwaan, Stanfield, \& Yaxley, 2002). For example, Huettig and Hartsuiker (2008) showed that naming a category exemplar (e.g., musical instrument - saxophone) elicited eye movements to a picture of a semantically unrelated object that was similar in shape (e.g., ladle). This activation of visual sensory information is context related. Zwaan, Stanfield, and Yaxley (2002) showed, for example, that context can affect the particular shape of the object that
is represented. In their experiment, participants were presented with sentences like 'The ranger saw the eagle in the sky' or 'The ranger saw the eagle in its nest,' which were followed by a line drawing of the object described in the sentence, in this case an eagle with outstretched wings or an eagle with folded wings. Participants recognized the picture faster if the implied shape of the object in the sentence matched the shape of the object in the picture. In the same line, Van Weelden, Schilperoord, and Maes (in press) showed that sentence structure (which can define the relation between multiple objects) influences the shape of the represented object(s) as well. In their experiment, participants were presented with sentences that invited to compare two objects like ' \(A\) spinning top is like a ballerina,' which were followed by two line drawings of the objects described in the sentence. The two drawings either had a similar or dissimilar shape. Participants recognized the pictures faster if they were similarly shaped. Hence, a sentence structure that invites to (conceptually) compare two objects affects the shape of their mental representation.

While language has been shown to elicit perceptual representations, there is also work that shows that the opposite occurs as well, that is, that semantic information is activated during visual object perception. Boucart and Humphreys (1997) suggest that as a result of the strong interplay between sensory and semantic information, people cannot even attend selectively to the global shape of an object without automatically processing its semantic properties. Caramazza, Hillis, Rapp, and Romani (1990) try to explain this interaction with their Organized Unitary Content Hypothesis (OUCH). Their theory is based on the idea that, contrary to a word for a particular concept, the object itself tends not to have an arbitrary relationship to its meaning. Some visual sensory features are directly related to the semantic properties of the object that specify its function (cf. Gibson's affordance theory; 1977, 1979). These features are therefore perceptually salient. As such, shape is very frequently a salient perceptual feature.

Accordingly, visual sensory information contributes fundamentally to the process of object identification and categorization. In the present study, we propose that the degree of activation of visual information in the process of object categorization might depend on the situational relevance of this information (Chaigneau, Barsalou, \& Samani, 2009; Pecher et al., 1998). We define this situational relevance as the result of the visual and semantic relations between the objects. For example, we might predict that when we have to look for an overarching category for a number of presented objects, visual features, such as shape, might play a bigger role if objects belong to different semantic categories as compared to when they stem from the same semantic category. Therefore, in the present study, we investigate whether shape information is encoded differently in our semantic memory for objects from similar and dissimilar semantic categories.

One way to investigate how visual information is encoded, and hence whether the objects are organized in semantic memory by means of their shape, is by looking at the process of retrieval of this particular information. The encoding and retrieval of the encoded information are interdependent; a retrieval cue will be effective if and only if the information in the cue was generated at encoding (Blaxton, 1989; Morris, Bransford, \& Franks, 1977; Tulving \& Thomson, 1973). Hence, by examining whether the shape of objects is used as a retrieval cue when trying to retrieve objects from memory, we can determine whether shape information was encoded in the semantic memory.

To do so, we use the Proactive Interference (PI) paradigm (Wickens, 1970). Proactive interference occurs when previously encountered information interferes with the memorial access of more recently encountered information. The standard procedure to test this interference is to present a triad of items from the same semantic category and, subsequently, have the participant perform a 25 -s rehearsalpreventing task, such as a backward counting task. Then, participants recall the triad. This procedure is repeated for four trials. The idea is that because the items are members of the same semantic category, the meaning of the items is being encoded and so is the meaning of the non-presented category under which they subsume. The PI paradigm results in decreasing performance on the recall task as more triads from the same semantic category are presented. Because participants use the same category cue to recall the items, increasing interference arises. If, however, the semantic category shifts on the fourth (i.e., the critical) trial, the category cue will change as well. Therefore, the discriminability and accessibility of the items will increase, resulting in an increased performance on the recall task. This mechanism is called release from interference.

In previous studies, the PI paradigm has been used to investigate the magnitude of the semantic distance between exemplars from different semantic categories (i.e., shift from fruits to vegetables as compared to shifts from fruits to professions), phonemic categories (i.e., shift from words with 'air' sound to 'eye' sound), and sensory features (i.e., shift from 'round' words to 'white' words) (Wickens, Dalezman, \& Eggemeier, 1976; Zinober, Cermak, Cermak, \& Dickerson, 1975). The main conclusion drawn from these studies is that the degree of release from interference is inversely related to the number of common characteristics. That is, a shift between categories with a high overlap in characteristics (i.e., from fruits to vegetables) obtains a lower release from interference as compared to a shift between categories with no overlapping characteristics (i.e., from fruits to professions).

Marques' (2000) study showed release from interference as a result of a shift from nonliving to living things. Interestingly, Marques tested this living/nonliving distinction for both words and pictures of the objects. The visual stimuli yielded the same types of interference effects
as verbal stimuli. Accordingly, this study shows that the PI paradigm can also be used to investigate which retrieval cues people use to recall visual objects from their memory and, hence, which information was encoded when the visual objects were processed.

The present study employs the PI paradigm with the visual manipulation of object shape. We refer to shape as the outline of the picture of a particular object, rather than its inherent shape. We predict that if depictions of objects are encoded in such a way as to include information about the shape of the objects, then objects with a particular shape should form a different category than objects with another shape. Therefore, interference should build up as objects with similar shapes are presented on successive trials, and a release from interference should occur with a shift of shape. Yet the relative weighting of shape information might differ as a result of the situational relevance of this information. In two experiments, we manipulate the semantic and shape similarity between the objects and, thereby, the situational relevance of shape. In Experiment 1, we combine a shift of shape with a semantic shift. For this type of shift, we expect that a semantic category cue will be sufficient to recall the objects from the critical trial. So, for this situation, the role of shape might be inferior. In Experiment 2, we will only manipulate a shift of shape, keeping the semantic category similar throughout the experiment. For this situation, we expect shape to be a distinguishing factor and to be used as a retrieval cue.

\section*{Experiment 1}

This first experiment evaluated the role of shape in the PIrelease situation with both a shape and semantic categorical shift. The semantic shift comprised a shift between two natural categories, fruits and flowers. We used this type of shift because living things are primarily differentiated on the basis of perceptual features (Humphreys \& Forde, 2001; Warrington \& McCarthy, 1987). That is, most types of natural objects have a high perceptual overlap, and therefore small perceptual differences are highly informative. Hence, it can be expected that visual information will have a relatively high weighting as compared to other types of information in the representation of living things.

Both the participants in the Shift and No-Shift condition received three fruits triads followed by a flower triad. In the No-Shift condition, the shape of the fruits and flowers did not change throughout the experiment. The objects either were round in shape or were shaped irregularly. In the Shift condition, however, the shape of the objects changed on the critical trial. The critical trial established a shift from irregularly shaped objects to round shaped objects or vice versa.

For both the Shift and No-Shift condition, we predicted release from interference to occur as the change from fruits to flowers reduces or eliminates interference. However,
there may be gradual differences in the amount of release, both as a result of the shape shift itself and the type of shape shift. We expected the release to be most prominent for the Shift condition as there is an additional shift of shape. Considering the type of shape shift, we predicted the release to be stronger when triads changed from round shaped objects to irregularly shaped objects than the other way around. If pictures of objects are encoded in such a way as to include information about the shape of the objects, then the buildup of interference is stronger for round objects, which might result in a stronger release effect.

For the No-Shift condition, we predicted the release from interference to be hampered when the triads of the four trials consist of round objects. Although there was a semantic change from fruits to flowers, the objects remained perceptually similar. As a result, the previously seen objects may continue to interfere with the objects presented on the critical trail. When the triads of the four trials consist of irregularly shaped objects, however, this interference effect may be more moderate as the objects are not perceptually similar. The semantic shift would then be sufficient to eliminate such interference effects.

\section*{Method}

Participants Eighty Tilburg University undergraduates (57 women) participated for course credit. The mean age was 21 years, ranging from 18 to 34 .

Materials \({ }^{1}\) The stimulus pictures consisted of 18 pictures of fruits ( 9 round shapes and 9 irregular shapes) and 6 pictures of flowers ( 3 round shape and 3 irregular shapes). The pictures were arranged in triads ( 6 for fruits and 2 for flowers). In arranging these triads, we controlled for various factors. For the fruits triads, we controlled for typicality. In a typicality pretest, ten participants (who did not participate in the future PI experiment) were asked to sort the pictures of the objects from most typical member of the category 'fruits' to the least typical member of this category. Based on this taxonomy, every fruits triad was assigned a low, medium, and high typical member of the category. In addition, every fruits and flowers triad consisted of three differently colored objects. We kept the visual complexity similar across triads in terms of mean JPEG file sizes (Chikhman et al., 2012; Donderi, 2006).

Design The experiment had a \(2 \times 2 \times 4\) design, with Condition (levels: Shift and No-Shift) and Triad Shape (levels: Round shape and Irregular shape) as betweensubjects factors and Trial (levels: 1, 2, 3, and 4) as withinsubjects factor.

\footnotetext{
\({ }^{1}\) See dcilab.uvt.n1/LisanneVanWeelden/materials.pdf for the materials of Experiment 1 and 2.
}

Procedure The participants were informed that the purpose of the experiment was to test their ability on both backward counting and their memory of triads of objects. During each trial, participants first saw a fixation cross in the center of the screen for 2 s . Subsequently, the objects of one triad were presented one-by-one for 2 s each (with no interstimulus interval). Participants were instructed to identify the objects silently, to remember them, and also to remember the order of the objects. They were told that they had to recall the objects in the right order afterwards. A three-digit number was then presented in the middle of the screen for 25 s during which the participant had to count backwards by threes out loud. Participants were instructed to count backwards as fast as possible while still being accurate. After 25 s the question 'Which three objects did you see?' appeared, signaling the beginning of the 12 s recall period. Participants typed the names of the three objects. After 12 s the question was replaced with "Time's up" to indicate the end of the recall period. Participants pressed a button to continue to the next trial. The next trial started again with the fixation cross. Participants trained on both the counting backward and memory task with a four trial training block.

\section*{Results and discussion}

For each participant, the mean recall score was computed for each trial. Following the procedure of Wickens, Dalezman, \& Eggemeier (1976), one point was given for each object recalled correctly and one extra point was assigned when the three objects were recalled in the correct order. So, for each trial, there was a maximum of 4 points. The mean scores per Condition and Trial are presented in Figure 1.

PI-buildup and PI-release effects were analyzed independently. The PI-buildup analysis was performed on the first three trials. The PI-release analyses were performed on (1) the third and fourth trial and (2) on the fourth trial separately. For all three analyses an ANOVA was conducted with Condition (levels: Shift and No-Shift) and Triad Shape (levels: Round shape and Irregular shape) as between-subjects factors. For the PI-buildup analysis the latter factor concerned the Shape of the first three triads, whereas for the PI-release analyses this regarded the Shape of the fourth triad. The PI-buildup analysis also involved the within-subjects factor Trial (levels: 1, 2, and 3).

For PI-release, the analysis on the third and fourth trial revealed a main effect of Trial, \(F(1,152)=31.19, p<.001\), \(\eta_{\mathrm{p}}^{2}=.17\). The mean recall score was higher on the fourth trial \((\mathrm{M}=3.55, \mathrm{SD}=.95)\) than on the third trial \((\mathrm{M}=2.53\), \(\mathrm{SD}=1.31\) ). Participants recalled more items after the semantic shift. There was no effect of Condition, \(F<1\), or Triad Shape, \(F<1\), and there were no two- or three-way interactions between the factors, \(F<1\). The analysis on the fourth trial alone revealed neither a main effect of

Condition, \(F<1\), and Triad Shape, \(F<1\), nor an interaction between the two, \(F(1,76)=1.98, p=.16\). Thus, the semantic shift did result in release from interference, but there were no (gradual) differences in release as a result of the shift in shape on the fourth trial.

For PI-buildup, the analysis showed a main effect of Trial, \(F(2,228)=9.31, p<.001, \eta_{\mathrm{p}}^{2}=.08\). Participants recalled fewer items as the number of trials increased. Post hoc analyses showed that the decrease from trial 1 to trial 2 was significant, \(p<.05\). The decrease from trial 2 to trial 3 did not reach significance, \(p=.22\). There was no effect of Condition, \(F<1\), nor an effect of Triad Shape, \(F<1\). The analysis did not reveal any two- or three-way interactions between the factors.


Figure 1: Mean recall scores on each trial for the Shift and No-Shift condition in Experiment 1.

These results show that shape information was overruled by semantic information. Only semantic information was used as retrieval cue, as indicated by the build-up of interference during the first three trials and the release from interference when the semantic category changed. The change in shape did not affect performance. We expected that the role of shape becomes more prominent if a semantic retrieval cue is not sufficient to recall the objects of the critical trial. This possibility was explored in Experiment 2.

\section*{Experiment 2}

This second experiment evaluated the role of shape in the PI-release situation without a semantic categorical shift. Participants in both the Shift and No-Shift condition received four fruits triads. Identical to Experiment 1, the shape of the fruits was similar throughout the four trials in the No-Shift condition, in the sense that the objects either had a round shape or were shaped irregularly. In the Shift condition, the shape of the objects changed on the critical trial. The change concerned a shift from irregularly shaped objects to round shaped objects or vice versa.

For the Shift condition, we predicted release from interference to occur as a result of the shape shift. Again, we expected the release to be more prominent when triads changed from round shaped objects to irregularly shaped objects than when they shifted in the opposite direction. For the No-Shift condition, we predicted that the buildup of interference would continue throughout the four trials. The decrease in performance was expected to be the strongest for the round shaped objects as compared to the irregularly shaped objects.

\section*{Method}

Participants Eighty Tilburg University undergraduates (57 women) participated for course credit. The mean age was 22 years, ranging from 18 to 33 .

Materials The triads of the first three trials were the same as in Experiment 1. The experimental materials for these triads consisted of consisted of 18 pictures of fruits ( 9 round shapes and 9 irregular shapes). For the present experiment, the triads of the fourth trial consisted of 6 pictures of fruits ( 3 round shapes and 3 irregular shapes). In arranging these triads, we controlled again for typicality, color, and visual complexity.

Design and procedure The design and procedure were the same as in Experiment 1.

\section*{Results and discussion}

For each participant, the mean recall score was computed for each trial. As in Experiment 1, there was a maximum of 4 points. The mean scores per Condition and Trial are presented in Figure 2.

PI-buildup and PI-release effects were analyzed independently in the same manner as Experiment 1. For PIrelease, the analysis on the third and fourth trial revealed a trend of an effect of Condition, \(F(1,152)=2.76, p=.09\). The analysis also showed a trend of an interaction between Condition and Trial, \(F(1,152)=2.89, p=.09\). There was no main effect of Triad Shape, \(F<1\), or Trial, \(F<1\), nor any other two- or three-way interactions. The analysis on the fourth trial alone revealed a main effect of Condition, \(F(1\), \(76)=5.70, p<.05, \eta_{\mathrm{p}}^{2}=.07\). The mean recall score was higher for the Shift condition ( \(\mathrm{M}=2.58, \mathrm{SD}=1.30\) ) than for the No-Shift condition ( \(\mathrm{M}=1.92, \mathrm{SD}=1.05\) ). Participants recalled more items after the shape shift. There was no main effect of Triad Shape, \(F<1\), nor an interaction between Condition and Triad Shape, \(F(1,76)=2.21, p=.14\). So, the shape shift resulted in release from interference, causing an increase of the recall scores on the fourth trial.

For PI-buildup, the analysis showed a main effect of Trial, \(F(2,228)=18.40, p<.001, \eta_{p}^{2}=.14\). Post hoc analyses showed that both the decrease from trial 1 to trial
\(2, p<.01\), and from trial 2 to trial \(3, p<.001\), was significant. There was no effect of Condition, \(F<1\), nor an effect of Triad Shape, \(F<1\). The analysis did not reveal any two- or three-way interactions between the factors.


Figure 2: Mean recall scores on each trial for the Shift and No-Shift condition in Experiment 2.

These results show that if semantic information is insufficient to recall the objects of the critical trial, shape comes into play. The fact that shape is used as a retrieval cue to recall objects from memory suggests that the objects are assigned to a subordinate shape category within the semantic category of 'fruits'.

\section*{General discussion}

The purpose of the present study was to investigate the role of shape in semantic memory organization of visual objects. We predicted that if depictions of objects are encoded in such a way as to include information about the shape of the objects, then objects with a particular shape should form a different category than objects with another shape. We also predicted that the degree of activation of shape information might depend on the situational relevance of this information. Therefore, in two experiments, we investigated semantic memory organization in two different situations using the PI paradigm. We created these different situations by manipulating the objects' shape and semantic nature. The results of the present study suggest that semantic memory organization of objects is indeed dependent on the interaction between semantic and shape information.

Experiment 1 showed that if objects can be categorized both on semantic and shape information, then shape information is overruled by semantic information. Namely, as indicated by the release from interference as a result of the semantic category change, semantic information was used as retrieval cue, which was not affected by the shift in shape. Hence, it seems that object categorization is largely
driven by semantic features, as those features received higher activation than perceptual features.

Experiment 2 showed however that shape does play an important role in object categorization, that is, if semantic information is not a distinguishing factor and therefore does not receive high activation. In this experiment a situation was created in which the semantic information remained unchanged, whereas the shape of the objects did change. The release from interference as a result of the shift in shape showed that object shape was indeed used as retrieval cue. So, in this situation, objects are categorized based on their shape.

To summarize, object categorization is driven by semantic information to a large extent, yet if semantic information cannot be (satisfactorily) used to store and retrieve objects, then shape comes into play.

\section*{Acknowledgments}

This research is part of the project "Visual Metaphor: A Psycholinguistic Perspective," which is funded by an "Open Competition" grant from the Netherlands Organization for Scientific Research (NWO).

\section*{References}

Barsalou, L. W. (1999). Perceptual symbol systems. Behavioral and Brain Sciences, 22, 577-660.
Blaxton, T. A. (1989). Investigating dissociations among memory measures: Support for a transfer-appropriate processing framework. Journal of Experimental Psychology: Learning, Memory, \& Cognition, 15, 657668.

Boucart, M., \& Humphreys, G. W. (1997). Integration of physical and semantic information in object processing. Perception, 26, 1197-1209.
Caramazza, A., Hillis, A. E., Rapp, B. C., \& Romani, C. (1990). The multiple semantics hypothesis: Multiple confusions? Cognitive Neuropsychology, 7, 161-189.
Chaigneau, S. E., Barsalou, L. W., \& Samani, M. (2009). Situational information contributes to object categorization and inference. Acta Psychologica, 130, 81-94.
Chikhman, V., Bondarko, V., Danilova, M., Goluzina, A., \& Shelepin, Y. (2012). Complexity of images: Experimental and computational estimates compared. Perception, 41, 631-647.
Donderi, D. C. (2006). Visual complexity: A review. Psychological Bulletin, 132(1), 73-97.
Gibson, J. J. (1977). The theory of affordances. In R. Shaw \& J. Bransford (Eds.), Perceiving, Acting, and Knowing: Toward an Ecological Psychology. Hillsdale, NJ: Lawrence Erlbaum.
Gibson, J. J. (1979). The ecological approach to visual perception. Boston, MA: Houghton Mifflin.
Huettig, F., \& Hartsuiker, R. J. (2008). When you name the pizza you look at the coin and the bread: Eye
movements reveal semantic activation during word production. Memory \& Cognition, 36, 341-360.
Humphreys, G. W., \& Forde, E. M. E. (2001). Hierarchies, similarity, and interactivity in object recognition: "Category-specific" neuropsychological deficits. Behavioral and Brain Sciences, 24, 453-509.
Marques, J. F. (2000). The "living things" impairment and the nature of semantic memory organisation: An experimental study using PI-release and semantic cues. Cognitive Neuropsychology, 17, 683-707.
Morris, C. D., Bransford, J. D., \& Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. Journal of Verbal Learning and Verbal Behavior, 16, 519-533.
Nosofsky, R. M. (1986). Attention, similarity, and the indentification-categorization relationship. Journal of Experimental Psychology: General, 115, 39-57.
Pecher, D., Van Dantzig, S., Zwaan, R. A., \& Zeelenberg, R. (2009). Language comprehenders retain implied shape and orientation of objects. Quarterly Journal of Experimental Psychology, 62, 1108-1114.
Pecher, D., Zeelenberg, R., \& Raaijmakers, J. G. W. (1998). Does pizza prime coin? Perceptual priming in lexical decision and pronunciation. Journal of Memory and Language, 38, 401-418.
Rosch, E., \& Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. Cognitive Psychology, 7, 573-605.
Stanfield, R. A., \& Zwaan, R. A. (2001). The effect of implied orientation derived from verbal context on picture recognition. Psychological Science, 12, 153-156.
Tulving, E., \& Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. Psychological Review, 80, 352-373.
Van Dantzig, S., Pecher, D., Zeelenberg, R., \& Barsalou, L. W. (2008). Perceptual processing affects conceptual processing. Cognitive Science, 32, 579-590.
Van Weelden, L., Schilperoord, J., \& Maes, A. (in press). Evidence for the role of shape in mental representations of similes. Cognitive Science.
Warrington, E. K., \& McCarthy, R. A. (1987). Categories of knowledge: Further fractionations and an attempted integration. Brain, 110, 1273-1296.
Wickens, D. D. (1970). Encoding categories of words: An empirical approach to meaning. Psychological Review, 77, 1-15.
Wickens, D. D., Dalezman, R. E., \& Eggemeier, F. T. (1976). Multiple encoding of word attributes in memory. Memory \& Cognition, 4, 307-310.
Zinober, J. W., Cermak, L. S., Cermak, S. A., \& Dickerson, D. J. (1975). A developmental study of categorical organisation in short-term memory. Developmental Psychology, 11, 398-399.
Zwaan, R. A., Stanfield, R. A., \& Yaxley, R. H. (2002). Language comprehenders mentally represent the shapes of objects. Psychological Science, 13, 168-171.

\title{
Evaluating Two Mechanisms of Flexible Induction: Selective Memory Retrieval and Evidence Explanation
}

\author{
Nadya Y. Vasilyeva (Nvasil@Brandeis.Edu) \\ Department of Psychology, Brandeis University \\ 415 South Street, Waltham MA 02454 USA
}

John D. Coley (j.coley@neu.edu)
Department of Psychology, Northeastern University
360 Huntington Avenue, Boston MA 02115 USA

\begin{abstract}
We report three studies examining mechanism of propertysensitive induction. First, we demonstrate that, contrary to a common assumption, property does not influence retrieval of knowledge about premise categories. Second, we introduce property-driven explanations as a possible source of property effects and provide first evidence for this proposal.
\end{abstract}

Keywords: induction; property effects; retrieval; explanation.
Generating hypotheses about uncertain outcomes from limited evidence - inductive inference - is a pervasive cognitive activity. In order to be successful, inductive inference must be flexible. For example, if you learn that a new influenza virus has been discovered in chickens, you may reasonably get concerned about your own health; but if chickens were announced to carry a certain defective gene, you are much less likely to worry about catching one during your next meal. Indeed, a vast body of empirical evidence demonstrates that people make systematically different inferences when they project different properties (see Coley \& Vasilyeva, 2010, for a review). Heit and Rubinstein (1994) proposed that property affects induction by indicating different subsets of features as relevant for evaluating premise-conclusion similarity. Goodman (1972) provided a logical argument for constrained recruitment of features: since any category has a potentially infinite set of features and can be infinitely similar to any other category, it is a necessary logical requirement for inductive inference to impose some initial constraints to limit a subset of relevant features.

Although it is generally agreed that induction requires constrained recruitment of prior knowledge, and there is evidence that projected property may provide one such constraint (Coley \& Vasilyeva, 2010; Heit \& Rubinstein, 1994), the mechanism of property-based constrained recruitment remains unclear. Existing models of induction either do not specify the psychological mechanism of property effects (McDonald, Samuels, \& Rispoli, 1996; Medin, Coley, Storms, \& Hayes, 2003; Rips, 1975; Osherson, Smith, Wilkie, Lopez, \& Shafir, 1990; Sloman, 1993; Sloutsky \& Fisher, 2004), or acknowledge the computational nature of their account that may not correspond to actual psychological processes involved in inductive inference (e.g. Tenenbaum, Kemp, \& Shafto, 2007; Heit, 1998).

We report three studies that examine two candidate psychological mechanisms of property effects in induction: property-moderated retrieval of relevant knowledge about premise categories from long-term memory, and generating explanations of why premise categories might have the property to begin with. In contrast to the majority of studies
on induction that use argument evaluation task, we employed inference generation task: participants were given an inductive premise and asked to generate their own conclusions. Coley \& Vasilyeva (2010) demonstrated that this task provides a particularly sensitive measure of participants' spontaneous use of different kinds of relevant knowledge, in the context of an ecologically valid inductive problem.

\section*{Property-Moderated Knowledge Retrieval as a Mechanism of Property Effects}

Generation of an inductive hypothesis is inherently knowledge driven; when one learns that A has a novel property X , one uses what they know about A and its relations to other things to form guesses about what else is likely to have X. One source of input to inductive inference is knowledge about premise categories. When such categorical knowledge is accessed, a probabilistically determined subset of features and relations that comprise the representation of that concept becomes available as a raw material for the inference. For example, if A turned out to be a duck, such features as "is a bird", "flies", "lives in ponds", "quacks" and "eaten by foxes" may come to mind. Although there are many different types of knowledge, knowledge about living things is commonly divided into two broad classes: taxonomic knowledge is based on relations of intrinsic similarity between members, whereas contextual, or ecological knowledge, is based on extrinsic relations between members and other entities. For example, ducks belong to the taxonomic category of birds and ecological categories of aquatic animals and fox prey. Each of these types of knowledge can serve as a basis for an inductive projection from ducks - to other birds, or other aquatic animals, or things that eat ducks.

In addition to the premise category, knowledge about the property can also serve as a source of information. If X in the example above is replaced with a more specific property, such as "carries a certain disease" or "has a certain gene", new knowledge is brought to the table: independently of what we know about ducks, we also know something about diseases and genes: what they are, whether they can be transmitted via contact, etc. How can property influence what projections people end up making? One possibility is that property constrains what types of premise knowledge are used to produce an inductive hypothesis.

A premise category label, as any word, is connected to a vast amount of conceptual knowledge; this knowledge is unlikely to be retrieved in its entirety on any given occasion (McElree, Murphy \& Ochoa, 2006). Rather, retrieval of conceptual information from long term memory is selective and depends on context (e.g. Barsalou, 1982; Swinney,
1979). Within an inductive problem, property may serve as context for the premise category(-ies), and it could affect what information about these categories is retrieved. For example, in response to a premise like "ducks have gene X", one may be more likely to retrieve taxonomic knowledge about ducks (bird, have feathers whereas for ("ducks have parasite X "), one may be more likely to retrieve ecological knowledge about ducks (aquatic, prey to foxes). This mechanism is consistent with the flexible similarity proposal by Heit and Rubinstein (1994) and the Bayesian model of induction by Heit (1998). The advantage of this proposal is that it is more specific, it focuses on describing the underlying psychological process rather than on modeling outcomes of such a process, and it can be tested with behavioral data.

To evaluate this proposal, we conducted two studies. Study 1 examined knowledge recruitment: how knowledge about premise categories predicts outcome inferences about different properties. Study 2 examined knowledge retrieval: how knowledge about premise categories is activated in context of different properties during inference generation.

\section*{Study 1: Property-Specific Knowledge Recruitment}

To begin to specify the role of property in inference generation, we examined how it affects recruitment of category knowledge, or the extent to which available knowledge about premise categories ends up being used in the outcome inferences. For example, if among many facts about ducks, one knows that they live in water, and one uses this knowledge to project a property from ducks to other aquatic animals, we can say that the knowledge has been recruited. The question is whether the nature of the property affects the likelihood of recruiting ecological versus taxonomic knowledge about premise categories.

To address this question, we measured available knowledge about a set of animal categories and examined the predictive relationship between this knowledge and inferences generated about the same set of animal categories (i.e. knowledge recruitment). Most critically for evaluating the first proposal, we manipulated the property in the inference generation task between ecologically-biasing, neutral, and taxonomically-biasing. If property moderates recruitment of knowledge about premise categories, the predictive relationship between category knowledge and inferences should vary with the property. Based on Coley \& Vasilyeva (2010), we expected that property could facilitate recruitment of congruent knowledge, and/or inhibit recruitment of incongruent knowledge. For example, taxonomic properties should facilitate recruitment of taxonomic knowledge, and inhibit recruitment of ecological knowledge; ecological properties should show the converse pattern.

\section*{Method}

Feature-Listing Task Twenty nine participants were given a list of \(42^{1}\) animal names and for each animal were asked to write down anything they could think of that was "generally true of that animal".

Inference-Generation Task One hundred participants were given 42 open ended-induction questions about same

\footnotetext{
\({ }^{1}\) Feature-listing data from 1 animal were lost.
}

42 animals; each stated that a property was true of a single animal species, and asked what other species were likely to have the property. For example, "GENE T5 is found in DUCKS. What else is likely to have gene T5? Why?" Property was manipulated within subjects; participants saw two examples of three kinds of properties: ecological (flu, parasite), taxonomic (gene, cell) and neutral (substance, property). \({ }^{2}\) Each participant was presented with seven questions about each property-each with a different animal premise-in random order. The dependent variables were the frequencies of taxonomic and ecological inferences.

\section*{Results}

Data Coding Four or five trained coders coded features and inferences into two broad classes. Taxonomic (Tax) features and inferences invoked category membership, perceptual features, or non-interactive aspects of behaviors and physiology (e.g. Tax-feature: "bird"; Tax-inference: "other birds will have it"). In contrast, ecological (Eco) features and inferences invoked animals' diet, habitat, or other interactions with entities in their environment (e.g. Ecofeature: "lives in water"; Eco-inference: "other animals that live in water"). Each feature and inference was coded as taxonomic, or ecological, or both, or neither \({ }^{3}\). For every animal, the mean counts of features coded as Tax or Eco were taken as the measures of the amount of salient taxonomic and ecological knowledge about that animal. To quantify inferences, relative frequencies of participants making Tax and Eco inferences about that animal were calculated separately for each property type resulting in 6 means per animal.

Property effects Results showed strong property effects. Eco-inferences were generated most frequently when the property was ecological, followed by neutral and taxonomic properties \(\left(F(2,82)=95.05, p<.001, \eta_{\mathrm{p}}^{2}=.70\right)\); this pattern was reversed for Tax-inferences \(\left(F(2,82)=64.64^{4}, p<.001\right.\), \(\eta_{\mathrm{p}}^{2}=.61\); all planned pairwise comparisons \(p\) ' \(\mathrm{s}<.001\) ).

Knowledge recruitment We examined relations between premise category knowledge and property in predicting inferences in 12 simple linear regressions. In one triplet of regressions, ecological premise category knowledge served as a predictor of Eco-inferences, separately for ecological, neutral and taxonomic property. The other triplets covered the three remaining combinations between knowledge type and inference type, broken down by the property. The standardized regression coefficients are shown in Figure 1.

Eco-inferences For Eco-inferences (Fig. 1a) the predictive power of knowledge varied with the property. Eco-knowledge was overall a positive, albeit nonsignificant, predictor of Eco-inferences when participants were reasoning about a neutral ( \(R^{2}=.057, \beta=.239, p=.132\) ) or taxonomic \(\left(R^{2}=.034, \beta=.242, p=.128\right)\) property, but its

\footnotetext{
\({ }^{2}\) The 6 properties were selected from a larger pool of properties based on a norming study measuring participants' beliefs about the distribution of properties across taxonomic and ecological categories.
\({ }^{3}\) Coding categories were not mutually exclusive; a given response could receive multiple codes if it unambiguously invoked multiple codable kinds of relations.
\({ }^{4}\) Data were scored and analyzed by item. All the analyses on proportions reported below were conducted on arcsine-transformed data, while the reported means are non-transformed and presented as percentages.
}
contribution was stronger in the presence of an ecological property ( \(R^{2}=.124, \beta=.352, p=.024\) ). When we examined the contribution of Tax-knowledge to Eco-inferences, overall larger amounts of Tax-knowledge were associated with lower frequency of Eco-inferences (all \(\beta\) 's are negative), and this relationship again varied with property. Taxknowledge inhibited Eco-inferences marginally when the property was neutral ( \(R^{2}=.076, \beta=-.276, p=.08\) ), and reliably so when it was reinforced by a taxonomic property ( \(R^{2}=.12, \beta=-.346, p=.027\) ). Relative to taxonomic and neutral properties, ecological property largely neutralized the inhibitory effect of Tax-knowledge on Eco-inferences ( \(R^{2}=.007, \beta=-.083, p=.605\) ).

Tax-inferences Although, as shown in Fig. 1b, the sign and ordering of predictors generally follow the predicted pattern of strengthening effects of congruent knowledge and weakening effects of incongruent knowledge (with two exceptions), Tax-inferences were not significantly predicted by category knowledge (all \(p\) 's \(>.121\) ).


Figure 1: Eco- and Tax-knowledge about animals predicting relative frequency of Tax- and Eco-inferences about these animals, in the context of ecological, neutral and taxonomic properties. \({ }^{*} p<.05,+p<.1,{ }^{\circ} p<.15\)

\section*{Discussion}

Overall, property had a profound effect on the inferences participants generated, and category knowledge predicted ecological, but not taxonomic inferences. Eco-inferences were facilitated by congruent (ecological) knowledge about premise categories, and inhibited by incongruent (taxonomic) knowledge. Most importantly, the relation between knowledge and Eco-inferences varied with the property: property strengthened effects of congruent knowledge, and weakened effects of incongruent knowledge. The overall pattern of congruent facilitation and incongruent inhibition held for Tax-inferences as well, although it was weaker and did not reach significance. Even though evidence of a relationship between knowledge and inferences was present for Eco-inferences but absent for Tax-inferences, it is sufficient in order to provide an "existence proof" for moderating effect of property on knowledge recruitment. These results are consistent with Heit and Rubinstein's (1994) proposal about a flexible similarity metric, but they go beyond similarity relations and demonstrate flexible recruitment of ecological knowledge about contextual and interaction-based relations among animals.

\section*{Study 2: Property-Specific Knowledge Retrieval}

Even though this demonstration of property effects in knowledge recruitment provides a useful constraint on the general underlying retrieval process, it does not specify it completely. Study 2 directly examines knowledge retrieval by measuring activation of premise category knowledge in real time, as it is accessed during inference generation. We borrowed a cross-modal priming paradigm from Swinney (1979). Participants were auditorily presented with a taxonomic, ecological or neutral property and an animal premise, and were asked to generate possible conclusions. In addition, upon hearing the property and animal, participants were presented with a lexical decision task involving targets related to salient taxonomic or ecological knowledge about the premise animal. For example, a participant might hear a property, gene, followed by the animal, duck, and, after a varying time interval, see on the screen a taxonomic target bird, or an ecological target pond, or an unrelated target sofa, or a non-word soach. The task was to decide whether the letter sequence was a word. The time to respond to the related targets was taken as a measure of activation of Tax- or Eco-knowledge about the premise animal. If knowledge about duck is activated, we expect decisions about related targets (bird and pond) to be faster than about unrelated targets (sofa). If property moderates knowledge retrieval, we would expect decisions about ecologically related targets (pond) to be faster in the presence of an ecological property and/or slower in the presence of a taxonomic property relative to a neutral, or non-biasing property context. Similarly, we would expect decisions about taxonomically related targets (bird) to be faster in the presence of a taxonomic and/or slower in the presence of an ecological property compared to neutral.

\section*{Method}

Materials The stimuli for the induction task were 36 animal premises, each belonging to one salient taxonomic category (mammal, bird, reptile, fish, insect) and one salient habitat-based ecological category (forest, desert, pond, ocean, savannah). Each of the animals was presented in the context of an inductive problem about one of the six properties from Study 1 (flu, gene, etc., presented with unique alphanumeric codes (X5, Z9)). All the animal names and properties were recorded in the voice of a female native speaker of English.

Thirty-six words and 36 pronounceable non-words were used as targets for the lexical decision task. The targets (derived from feature responses in Study 1) were taxonomically related, ecologically related, or unrelated to the animals used as premise categories in the induction task. The strength of association of the taxonomic, ecological, and unrelated targets to the corresponding animals, as well as lack of direct associations between properties and target words, were normed with another group of 18 native speakers of English.

Participants and Procedure One hundred eleven native speakers of English were tested individually, on a MacBook laptop running Superlab 4.0.4 software and set up with headphones and a microphone. The experiment consisted of an open-ended induction task with intervening lexical decision task. Participants were instructed that they would be listening to utterances that would introduce a property, followed by an animal that possesses that property, and their
task was to say out loud (at a cued moment) other species likely to share that property, along with a short justification. Participants were also informed that at "random" moments a sequence of characters would appear on the screen, and their task was to identify it as a word or a non-word as quickly as possible without sacrificing accuracy, using the response buttons. Each trial began with a 2000 msec pause; then a participant heard the property to be projected (e.g., flu M3), followed by a pause of 1000 msec , followed by the name of the premise animal (e.g., bear), followed by a pause and a signal to start speaking. At varying SOAs (stimulus onset asynchrony: 400,900 , or 1650 msec from the onset of the animal name), a target for lexical decision appeared on the screen and stayed there until the participant responded or for 3500 msec , whichever came first. No accuracy feedback was provided. After a 2000 msec pause following the participant's response or the end of lexical decision target presentation, a short beep signaled that the participant could start saying their inference. Participants had 15 seconds to say their response, after which the experiment automatically moved on to the next trial.

Design The main independent variable was property type (taxonomic: gene, cell; ecological: flu, parasite; neutral: substance, property). The second independent variable was the target word type (taxonomic vs. ecological vs. unrelated). Each non-filler animal was yoked to one target word type (taxonomic, ecological, or unrelated). We also varied SOA, but for the sake of brevity, this manipulation will not be discussed here. All the reported analyses were collapsed across SOA.

\section*{Results}

Does property moderate retrieval of knowledge about premise animals? If so, property type should facilitate responses to property-congruent targets and/or interfere with responses for property-incongruent targets.

Accuracy A 3(target type: eco, tax, unrelated) \(x\) 3(property: eco, tax, neutral) repeated measures ANOVA on accuracy in lexical decision task showed no main effect of property \((F(2,220)=.145, p=.865)\). The effect of target type was significant \((F(2,220)=4.33, p=.014)\) : eco- and taxtargets were verified more accurately than unrelated targets \((t(110)=2.63, p=.010 ; t(110)=2.58, p=.011)\); the former two did not differ \((t(110)=.33, p=.744)\). Most importantly, the effect of target type did not interact with property \((F(4,440)=.57, p=.683)\). This suggests that participants were retrieving category-relevant knowledge, but that such retrieval was not moderated by property.

Reaction Time. Reaction time results were consistent with the accuracy analyses. A 3 (target type: eco, tax, unrelated) x 3 (property: eco, tax, neutral) repeated measures ANOVA on RT showed no main effect of property \((F(2,218)=.44, p=.656)\) and a significant effect of target type \((F(2,218)=4.73, p=.010)\) : eco-targets were verified faster than unrelated targets \((t(110)=3.20, p=.002)\); and tax-targets were verified marginally faster than unrelated targets \((t(110)=1.80, p=.074)\); the former two did not differ \((t(110)=1.46, p=.148)\). Again, most importantly, the effect of target type did not interact with the property ( \(F(4,436)=1.24, p=.293)\), suggesting that property does not moderate knowledge retrieval.

Surprisingly, the speed of lexical decisions about filler items (non-words) was affected by the property \((F(2,220)=14.95, p<.001)\) : decisions were slower in the
presence of ecological (1343msec, \(t(110)=4.88, \mathrm{p}<.001)\) and taxonomic ( \(1320 \mathrm{msec}, t(110\) ) \(=3.99, \mathrm{p}<.001\) ) than neutral property \((1253 \mathrm{msec})\); the former two did not differ \((t(110)=1.45, p=.151)\).


Figure 2: Effect of target type and property on lexical decision accuracy and reaction time. Verification of related targets (eco, tax) was more accurate (a) and faster (b) than unrelated targets. This effect did not depend on property (panels c and d). Error bars: 1 SEM.

\section*{Discussion}

Based on the results of Study 1, we expected to find effects of property on knowledge retrieval: specifically, facilitation of property-congruent knowledge and inhibition of property-incongruent knowledge. However, we found no evidence of property moderating knowledge retrieval. Of course, we cannot completely rule out the possibility that the lack of property effects was caused by some procedural flaws of the study. However, because we did see property effects on some lexical decisions about filler words, we know that the method is in principle capable of detecting property effects. And because the premise category did differentially prime related vs. unrelated targets, we know that the method is capable of detecting differential priming. Therefore, it is likely that we failed to see property effects on retrieval because property does not moderate retrieval of knowledge about premise categories from long-term memory.

\section*{Study 3: Property-Moderated Explanation as a Mechanism of Property Effects}

Study 1 demonstrated selective property-moderated recruitment of categorical knowledge to inform inferences. However, Study 2 found no moderating effects of property on knowledge retrieval in real time. If, as we argue, this finding reflects the actual absence of property effects on retrieval rather than an experimental failure, we need to look for another mechanism whereby property can guide selective recruitment of taxonomic and ecological information by inferences. The mechanism that we examine in Study 3 is based on property-moderated explanation of evidence.

As suggested by Sloman's (1994) work, explanation of evidence may affect evaluation of inductive arguments. Even when the similarity between premise and conclusion is held constant, if both can be explained by reference to the
same principle, the perceived strength of an inductive argument can be higher than when premise and conclusion statements require different explanations.

Several features of explanation make it a good candidate mechanism for property effects in induction. For instance, explanation is flexible: there are multiple ways to explain any given observation. A formal explanation refers to categories or inherent properties; a causal explanation refers to the proximal mechanisms of change; and a teleological explanation refers to ends, goals or functions (Lombrozo, 2006). On the subsumption account proposed by Williams \& Lombrozo (2010), explaining an observation involves identifying a larger pattern of which the observation is a part. In this sense, explanation identifies a relevant subset of knowledge about the observation that can serve as a basis for generalizing to other cases - thus satisfying the logical prerequisite for induction stipulated by Goodman (1972).

How might explanation provide a mechanism for property effects in induction? If different properties lend themselves to different explanations, and if explaining consists of identifying an observation as a part of a larger pattern or regularity, then different properties might determine whether a premise of an inductive argument is viewed as a part of one regularity or another (e.g. formal explanations might highlight taxonomic relations, whereas causal explanations might highlight ecological relations). Thus, construction of different explanations could engender differential recruitment of knowledge, and ultimately different hypotheses about how a property might generalize without the necessity of differential retrieval of knowledge.

In Experiment 1, although asked to explain why they generated particular conclusion categories, participants often spontaneously provided explanations for why a premise category exhibited a given property. To evaluate the explanation mechanism, we examined these spontaneous explanations to determine whether different properties were associated with different types of explanations. We expected taxonomic properties to provoke predominantly formal explanations referring to classes of objects (that would eventually translate into category-based, or taxonomic inferences) and ecological properties to lead to predominantly causal explanations, referring to interactions between animals and/or their environment (that would eventually translate into ecological inferences). We had no specific predictions about teleological explanations.

\section*{Method}

Three trained coders independently re-coded all the inferences collected in Study 1 for the presence of formal, causal and teleological explanations. Twelve percent of inferences ( 467 out of 3920 codable responses) contained spontaneous explanations. These explanations were coded as formal (explanations that referred to kind membership, e.g. "mammal gene", or "this is a bird flu"), causal (explanations describing a "story" of interactions between animals and other entities, or a sequence of events resulting in the premise category having the property, e.g. " vultures may get flu E5 from the dead and decaying animals they feed off of" or "[the gene will be found in] fish since pelicans eat them, the pelicans might develop that gene from the fish"), teleological (explanations referring to goals, functions or purposes of properties, e.g. "these cells are to protect them from the cold" or "perhaps B6-cells defend deer from particular viruses that they are exposed to"), or
other (idiosyncratic or vague explanations that could not be assigned to any of the three categories).

\section*{Results}

Scoring. For each animal, we calculated the percentage of subjects who generated each type of explanation out of total number of participants who reasoned about that animal, separately for each property type. This yielded 12 percentages, or relative frequency scores, per animal (3 property types x 4 explanation types), that were arcsinetransformed for the analyses. Uncodable explanations were rare (less than \(2 \%\) of participants per animal) and were excluded from the following analyses.

Analyses To provide support for the proposal that property affects inferences via explanation, we need to show that different properties are associated with different explanations, and that different explanations are associated with different inferences.

Relations between property and explanation. The main question, whether different properties trigger different types of explanations, was addressed by a 3 (property: eco, tax, neutral) x 3 (explanation: formal, causal, teleological) ANOVA on relative frequency of explanations. The overall likelihood of providing an explanation did not vary with the property \((F(2,82)=.044, p=.957)\). However, explanations differed in frequency \(\left(F(2,76)=18.836, p<.001, \eta_{\text {part }}^{2}=.315\right)\) : causal explanations were more frequent (5.3\%) than formal ( \(2.8 \%\), or teleological explanations \((2.4 \%, t(41) \geq 4.59\), \(p<.001, d \geq .71\) ), which did not differ from each other.

Of most theoretical interest was the significant interaction between property and explanation type \((F(4,164)=34.442\), \(p<.001, \eta_{\mathrm{p}}^{2}=.457\), see Fig 3a). Explanations clearly varied with property: for ecological properties, causal explanations were more frequent than formal explanation, which were more frequent than teleological explanations \((t(41) \geq 4.52\), \(p<.001, d \geq 0.70\) ). For neutral properties, causal explanations were also more frequent than formal or teleological explanations, which did not differ from each other \((t(41) \geq 3.00, p \leq .005, d \geq 0.45)\). In contrast, for taxonomic properties, formal and teleological explanations were more frequent than causal explanations \((t(41) \geq 3.42, p \leq .001\), \(d \geq 0.52\) ). These results demonstrate a link between property and explanation type.


Figure 3: a. Percentage of explained inferences involving causal, formal, and teleological explanations for ecological, neutral and taxonomic properties. b. Percentage of Tax- and Eco-inferences for responses with formal, teleological, and causal explanations.

Relations between explanation and inference To examine the link between explanation and inference, we again focused on the subset of responses from Study 1 that included spontaneous explanations. We calculated the percentage of Tax- and Eco-inferences that accompanied
each type of explanation. There was a clear association between explanation type and inference (Fig. 3b): responses that included formal or teleological explanations were much more likely to result in Tax- than Eco-inferences \((t(41) \geq 4.38, p<.001, d \geq 0.68)\). In contrast, responses that included causal explanations were much more likely to include Eco- than Tax-inferences, \(t(41)=8.61, p<.001\), \(d=1.33\) ). This systematic relationship between explanation type and inferences, taken together with the evidence for the relationship between property and explanation type, is consistent with the proposal that explanations mediate the effect of property on inferences.

\section*{Discussion}

To examine whether explanations might moderate property effects in induction, we asked whether different properties were associated with different explanations, and then whether different explanations were associated with different inferences. We have answered both questions in the affirmative. First, different properties triggered different types of explanations. When participants were reasoning about ecological properties, the majority of explanations they provided were causal, referring to a mechanism that could have endowed the animal with the property (e.g., "Owls eat mice and could contract the flu from the mice that it eats"). In contrast, when participants were reasoning about taxonomic properties, they were less likely to use causal explanations, preferring formal explanations ("this cell could be specific to jaguars") or teleological explanations ("T5 is something to keep them warm"). Second, different explanations were associated with different inferences. Causal explanations were more likely to accompany ecological inferences, whereas formal and teleological explanations were more likely to accompany taxonomic inferences.

These findings are consistent with the idea that explanations serve as a mediator between properties and inferences. We acknowledge that these analyses are correlational, and therefore do not provide direct evidence that explanations play a causal role in property-specific inductive inference. Nevertheless, an informal comparison of effect sizes indicates that the mean effect of explanations on inferences ( \(d=1.18\) ) is larger than the mean effect of properties on inferences \((d=0.92)\). This suggests that property-driven explanations are likely to affect inferences directly, rather than being a mere correlate of properties.

\section*{General Discussion}

We provided evidence that property effects do not take place in retrieval. This questions the existing, but not tested, assumption in the field about the mechanism of property effects based on context-dependent retrieval of information from semantic memory (Heit \& Rubinstein, 1994). We also provided some promising evidence that property effects may result from participants generating explanations for the presence of the property in the premise category. If this finding persists, it could strengthen connections between research on explanations and on induction. Most researchers agree that these are related, but very little supporting empirical work exists, although it is increasingly acknowledged that the presence of an available explanation can reduce reliance on overall similarity and override effects of similarity and diversity on induction (see Lombrozo, 2006, for a review). In this project we demonstrated that
explanations do not just "mess up" existing regularities in induction, but may in fact be an important part of the mechanism of one such established regularity - property effects in induction.

To sum up, this project makes a step towards specifying the mechanism of property effects in induction in two ways. First, it suggests that property effects do not work via property-based retrieval of knowledge about premise categories from memory. Second, it introduces propertydriven explanations as a possible source of property effects. Of course, these proposals are not mutually exclusive, and our main suggestion for the further research would be not to abandon studying knowledge retrieval in induction, but to expand research on the mechanism of property effects to include explanations.

\section*{References}

Barsalou, L.W. (1982). Context-independent and context-dependent information in concepts. Memory \& Cognition, 65, 15-32.
Coley, J.D. \& Vasilyeva, N.Y. (2010) Generating inductive inferences: premise relations and property effects. In Ross, B. (Ed.) The Psychology of Learning and Motivation (53). Burlington: Academic Press.
Goodman, N. (1972). Seven structures on similarity. In Problems and Projects. New York, NY: Bobbs-Merrill.
Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. In M. Oaksford \& N. Chater (Eds.), Rational models of cognition. Oxford: Oxford University Press.
Heit, E. \& Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. Journal of Experimental Psychology: Learning, Memory, and Cognition, 20, 411-422.
Lombrozo, T. (2006). The structure and function of explanations. Trends in Cognitive Sciences, 10(10), 464-470.
McDonald, J., Samuels, M., \& Rispoli, J. (1996). A hypothesis assessment model of categorical argument strength. Cognition, 59, 199-217.
McElree, B., Murphy G.L., \& Ochoa T. (2006). Time course of retrieving conceptual information: A speed-accuracy trade-off study. Psychonomic Bulletin \& Review, 13, 848-853.
Medin, D. L., Coley, J. D., Storms, G. \& Hayes, B. K. (2003) A relevance theory of induction. Psychonomic Bulletin \& Review, 10, 517-532.
Osherson, D.N., Smith, E.E., Wilkie, O., Lopez, A., \& Shafir, E. (1990). Category-based induction. Psychological Review, 97, 185-200.
Rips, L. J. (1975). Inductive judgments about natural categories. Journal of Verbal Learning \& Verbal Behavior, 14, 665-681.
Sloman, S.A. (1993). Feature-based induction. Cognitive Psychology, 25, 213-280.
Sloman, S.A. (1994). When explanations compete: the role of explanatory coherence on judgments of likelihood. Cognition, 52, 1-21.
Sloutsky, V.M., \& Fisher, A.V. (2004). Induction and categorization in young children: A similarity-based model. Journal of Experimental Psychology: General, 133, 166-188.
Swinney, D. (1979). Lexical access during sentence comprenension: (Re)consideration of context effects. Journal of Verbal Learning \& Verbal Behavior, 18, 645-659.
Tenenbaum, J. B., Kemp, C., and Shafto, P. (2007). Theory-based Bayesian models of inductive reasoning. In Feeney, A. \& Heit, E. (Eds.), Inductive reasoning. Cambridge University Press.

Williams, J.J. \& Lombrozo, T. (2010). The role of explanation in discovery and generalization: evidence from category learning. Cognitive Science, 34, 776-806.

\title{
An Information Foraging Model of Interactive Analogical Retrieval
}

\author{
Swaroop S. Vattam \& Ashok K. Goel \\ Design \& Intelligence Laboratory, School of Interactive Computing, Georgia Institute of Technology \\ Atlanta, GA 30332 USA
}

\begin{abstract}
An essential first step in analogy is retrieval of a source analogue appropriate for the target situation. In this paper, we focus on the phenomenon of interactive analogical retrieval (IAR) wherein the source analogues are obtained through interaction with online information environments. We first provide a descriptive account of IAR based on two in situ studies. We then describe an information-processing model (called PRISM) that provides an explanatory account of IAR. We conclude with a discussion of some of the theoretical and technological implications of this work.
\end{abstract}

Keywords: analogy, analogical retrieval, biologically inspired design, design cognition, information foraging.

\section*{Introduction}

Analogy appears to be ubiquitous in human cognition and thus has received much attention in cognitive science (e.g., Burstein 1986; Carbonell 1986; Clement 2008; Gentner 1983; Davies, Goel \& Nersessian 2009; Dunbar 2001; Gentner \& Markman 1997; Hofstadter 1995; Holyoak \& Thagard 1989; Indurkhya 1992; Keane 1988; Kokinov \& Petrov 2001; Kunda, McGreggor \& Goel 2013; Nersessian 2008). An essential first step in analogical reasoning is the retrieval of a source analogue appropriate to the target situation. Here we focus on situated analogy wherein source analogues are obtained through interactions with an external environment rather than being recalled from internal longterm memory. In particular, we focus on the phenomenon of interactive analogical retrieval (IAR) wherein source analogues are accessed from Web-based online information environments.

In this paper, we first present a descriptive account of IAR based on two in situ studies of cross-domain analogies in biologically inspired design. Then, we develop an information-processing model that provides a partial explanation of IAR in terms of its underlying cognitive processes. Our model builds on Pirolli's (2007) information foraging theory of human online information-seeking behavior and Thagard et al.'s (1990) model of analogical retrieval by constraint satisfaction.

\section*{Interactive Analogical Retrieval (IAR)}

We investigate IAR in the domain of biologically inspired design (Benyus, 1997; Vincent \& Mann, 2002), the practice of developing innovative technology using analogies to biological systems. Some well-known examples of products of biologically inspired design include Velcro (inspired by the attachment mechanism of burr seeds), high-performance wind turbines (inspired by the pectoral fins of humpback whales), self-cleaning surface coatings (inspired by the
super-hydrophobic effect of lotus leaves), and fog harvesting devices (inspired by the arrangement of hydrophobic and hydrophilic surfaces on Namibian beetles).

From a cognitive standpoint, biologically inspired design entails cross-domain analogies for solving a target design challenge in, say, engineering, by transferring elements of a source analogue from a different domain (biology). In biologically inspired design, designers (often from engineering) typically are novices in biology: they know of only a small fraction of the vast space of biological systems that comprise the source domain. Thus, in practice the designers often try to access biological analogues from the Web. We call this phenomenon interactive analogical retrieval. (Due to limitations of space, our discussion of the in situ studies below is very brief; Vattam (2012) provides more details.)

\section*{Study Context and Methodology}

We conducted two in situ studies of designers engaged in biologically inspired design in Fall 2006 and Fall 2008 respectively. Both studies were conducted in the context of an interdisciplinary, senior-level, project-based course on biologically inspired design taught at Georgia Tech. The most important element of the course for us was the semester-long design projects. Each design project grouped an interdisciplinary team of 4-6 engineers and biologists based on similar interests. Yen et al. (2011) provide details of the teaching and learning in this course.

As external observers in the Fall 2006 study and participant observers in the Fall 2008 study, we attended almost all the classroom sessions, collected all the course materials, documented lecture content, and observed teacher-designer and designer-designer interactions in the classroom. We documented a total of ten biologically inspired design projects in the two studies. We attended the design meetings of selected teams many times to observe firsthand how the design process unfolded. We took field notes, collected all the design related documentation produced by the teams, and also collected their idea journals. We analyzed the gathered data focusing on the processes and the products of the designers. In terms of the practices, we observed and documented frequently occurring problem-solving and representational activities of designers. In terms of the design products, we observed and documented their "design trajectory" - the evolution of the conceptual designs over time.

\section*{Main Findings}

We found that designers often searched online for biological systems that are analogous to their target design problem. In
fact, this was one of the dominant approaches for finding biological analogues. Designers reported using a range of Web-based information environments, including (1) online libraries of scholarly articles such as the Web of Science, Google Scholar, ScienceDirect, etc., (2) online encyclopedias like Wikipedia, (3) popular life sciences blog sites like Biology Blog, (4) biomimicry databases like AskNature, and (5) general search engines like Google. Online libraries like Web of Science and Google Scholar were the most frequently used websites.

We noted that designers used several heuristics in order to find relevant biology articles, including "biologizing" the problem, problem reformulation, functional decomposition, and using domain-bridging abstractions such as functions, mechanisms, physical principles, etc.

We noted that online information environments on which the designers relied upon did not adequately support the task of finding useful biological analogues. In particular, the designers reported that the online search for analogies was not only time consuming, but often also work intensive, tedious and frustrating.

Our analysis of the designers' online search activity identified three main challenges. The first challenge is findability. The relative frequency of encountering relevant articles containing biological analogues was very low. Designers often went for long periods without finding a single source analogue in a retrieval process that typically extended over several weeks. A rough calculation suggests that designers spent approximately three person-hours of search time on the Web in order to find a relevant article.

The second challenge is recognizability. The designers were prone to making errors in judgment about the true utility of the information resources that they encountered. In almost all online environments, search queries brought back a ranked list of search results. An important aspect of the search process was assessing and selecting promising information resources from this list for further consumption. However, this decision had to be made based on limited information, (e.g., titles, keywords and abstracts of biology articles). In many instances, designers picked up on lowutility articles and spent a lot of time and effort trying to understand its contents, only to realize later that they were not actually very useful (false positives). False positives have both resource and opportunity costs. Conversely, there were situations where designers dismissed a resource that they encountered during the search as having low utility even though it actually contained useful information about a relevant biological analogue (false negatives). The false negatives represent missed opportunities.

The third challenge is understandability. The designers often found it challenging to understand the contents of the biological cases described in the online information resources and develop the knowledge required to transfer to their target problems. This was in part due to the scholarly nature of the biology articles that were encountered and partly because the articles often did not explicitly describe how a biological system worked from a design perspective.

\section*{PRISM: A Model of Interactive Analogical Retrieval}

We now present an information-processing model of IAR. Our goal here is to find explanations for the observed challenges of online analogy seeking, both for (i) understanding cognition in IAR and (ii) developing a technology for supporting online search for cross-domain analogies. Here we focus on the challenges of findability and recognizability; Vattam (2012) addresses the challenge of understandability of biological cases that requires a different kind of explanation.

Our model builds on two existing theories: Analogical Retrieval by Constraint Satisfaction (ARCS) (Thagard et. al. 1990), and Information Foraging Theory (IFT) (Pirolli 2007). IFT is itself a biologically inspired theory of online information seeking behavior. According to IFT, the online information seeking behavior of people is analogous to how animals forage for food in their natural environments. IFT posits that the human information seekers use information scent to navigate from one information region to another in an information environment that is inherently patchy in nature, and from one information patch to another within a region. IFT suggests that the information seekers adapt their behavior to the structure of the information environment in which they operate such that the system as a whole (comprising of the information seeker, the information environment, and the interactions between the two) tries to maximize the ratio of the expected value of the knowledge gained to the total cost of the interaction.

Although several models of analogical retrieval from internal long-term memory informed our work (e.g., Forbus, Gentner, \& Law 1995; Kokinov \& Petrov 1997; Kolodner 1993; Yaner \& Goel 2006), we chose to build specifically on ARCS because it provides a content account of the types of similarity that best explains our observational data. ARCS posits that in order to access sources (represented as schemas in long-term memory) that are analogous to a target (a schema in short-term memory), the access mechanism should simultaneously consider satisfying three kinds of constraints: semantic similarity (the overlap in terms of the number of similar concepts between the target and potential sources), structural similarity (the overlap in terms of the higher-order relationships between the target and potential sources), and pragmatic similarity (the overlap in terms of the pragmatic constraints or goals surrounding the target and potential sources). It is these three pressures acting simultaneously that distinguish analogical retrieval from other kinds of information retrieval.
Thus, on one hand, ARCS explains how source analogues are retrieved from the long-term memory but is silent about analogies situated in external information environments. On the other, IFT explains how people seek information in online information environments in general, but it does not address the pressures of analogical retrieval. Our model specializes IFT to online analogy seeking by introducing the pressures of analogical retrieval from the ARCS model into information foraging. Thus we call our model PRISM (PRessurized Information Scent foraging Model).


Figure 1. PRISM: An information-processing model of interactive analogical retrieval.

Similar to IFT, as Figure 1 illustrates, in PRISM the task of retrieving a source analogue is accomplished by two iterative processes that constitute the general information seeking behavior: between-patch processes (on the left side of the vertical dotted line) and within-patch processes (on the right side). Furthermore, the structure of Web-based information environments has evolved to exhibit certain regularities in the distribution of information resources and the navigation mechanisms that lead to those resources. For instance, when an analogy seeker encounters patches in an online environment, the seeker cannot perceive the contents of those patches all at once. Instead, they are presented with snippets of information, called proximal cues, which the analogy seeker uses to perceive the information scent of the distal information patches. The information scent leads to judgments about the utility of distal information patches and the information seeker can choose to either navigate towards or away from those patches.

\section*{Between-patch foraging}

Between-patch foraging explains the navigation process where the analogy seeker browses the information environment looking for high-utility information patches to consume. In the context of IAR, high-utility information patches correspond to information resources describing sources cases that are analogous to the target. In this process, numerous information patches (e.g., online articles, etc.) compete for the information seeker's attention. These patches may or may not contain information relevant to the goals of the information seeker. Thus, the analogy seeker
expends time and effort navigating from one patch to another until one that can be exploited is found. This is captured by the Formulate Query-Retrieve-Compute Information Scent cycle depicted in Figure 1.

Between-patch foraging using information scent in IAR works as follows. Given a target problem or situation:
1. The analogy seeker probes the environment by formulating and issuing a query. This query is contextdependent and represents the target problem.
2. In response, the environment retrieves and conveys an information region consisting of a set of information patches \(\left\{\left(P_{1},\left\{c_{11}, c_{12}, \ldots\right\}\right),\left(P_{2},\left\{c_{21}, \quad c_{22}, \ldots\right\}\right) \ldots\right\}\), where \(P_{i}\) is an information patch and \(c_{i j}\) 's are the proximal cues associated with the patch \(P_{i}\).
3. The analogy seeker perceives the information scent of the patches based on the proximal cues; the information scent is an estimation of the analogical relevance of different patches to the target: \(\left\{\left(P_{1}, S_{l}\right),\left(P_{2}, S_{2}\right) \ldots\right\}\), where \(P_{i}\) is an information patch and \(S_{i}\) is the information scent that the analogy seeker associates with the patch \(P_{i}\).
4. If the information scent of an information patch exceeds a certain threshold, it is considered relevant (high perceived utility). In this case, the information seeker goes to that patch (by acting on the environment like clicking the associated hyperlink), at which point the environment presents the information patch to the forager. This initiates within-patch foraging.
5. If the scent does not exceed the threshold, it is considered irrelevant (low perceived utility). In this case, one of two things may happen as depicted in Figure 1: (i) the analogy
seeker may stay within the same information region but loop back to Step 3 for processing the next patch in the region, or (ii) the analogy seeker may abandon the current information region and loop back to Step 1 in order to look for more fruitful regions.

\section*{Within-patch foraging}

Once the analogy seeker picks up the scent of a potentially useful information patch, the seeker goes to the patch and starts consuming information in it. In the context of biologically inspired design, this involves comprehending the contents of an article and constructing a mental model of the biological system described in the article. In the withinpatch foraging process, the analogy seeker is also simultaneously evaluating the actual utility of the patch by comparing/aligning/mapping the emerging mental model of the biological system against the target problem. If the evaluation is successful, the agent has obtained a source analogue. If this evaluation fails, then the between-patch foraging process is again initiated, either within the same information region that led to the current patch or with a search for new information regions as depicted in Figure 1.

\section*{Incorporating Pressures of Analogical Retrieval}

There are two potential places in our model where the three pressures of semantic, structural and pragmatic similarity might apply: Retrieve and Compute information scent tasks that are shaded in gray in Figure 1. The Retrieve process may use some notion of similarity to access information patches. But in our model, the Retrieve process is implemented in the environment (e.g., the Google Scholar search mechanism) and thus is black-boxed here. The Compute information scent process computes the perceived utility of an information patch as described below.

\section*{Information Scent Perception in PRISM}

While IFT explains the scent perception for non-analogy information seeking tasks, it has to be adapted to account for the semantic, structural and pragmatic pressures of analogical retrieval. Hence, as part of PRISM, we developed a different model of information scent perception.


Figure 2. Scent perception in PRISM
Our model of scent perception assumes that the analogy seeker has represented the target problem as a target schema as depicted in Figure 2. With a target problem in mind, the analogy seeker forages the information environment for
source analogues. Following the between-patch foraging process described above, the analogist encounters a set of information patches with associated proximal cues. When the analogy seeker encounters proximal cues in the environment, she builds corresponding scent schemas as indicated in Figure 2.

Given the target schema and competing scent schemas, the analogy seeker computes the similarity between the target and scent schemas in four stages similar to ARCS. We illustrate this process with an example. Let us suppose that the conceptual structures representing the dots in Figure 2 consist of predicates. Table 1 illustrates a target schema (P1) consisting of two predicates (P1-1 and P1-2), and two scent schemas (S1 and S2) consisting of two predicates each (S1-1, S1-2, and S2-1 and S2-2, respectively). Let us also suppose that \(A\) and \(M\) are semantically similar concepts, and likewise concepts \(B\) and \(N\) are semantically similar. For example:
A( \(a, b\) ) is Regulate(kidney, potassium_ions); \(M(m, n)\) is Control_Production(pituitary, estrogen); \(B(b, a)\) is Is_Secreted_By(erythropoietin, kidney); and \(N(n, m)\) is \(I_{-}\)Released_By(hypothalmic_hormone, pituitary)
Let us further suppose that \(A(a, b)\) is more important than the other predicates as dictated by the pragmatics of the target situation.

Table 1: Example Target and Scent schemas (adapted from Thagard et al. (1990), pp. 275).
\begin{tabular}{lll}
\hline Target-schema & Scent-schema-1 & Scent-schema 2 \\
\hline P1 & S1 & S2 \\
P1-1: A(a, b) & S1-1: M(m, n) & S2-1: M(n, m) \\
P1-2: B(b, a) & S1-2: N(n, m) & S2-2: R(n, m)
\end{tabular}

Suppose that predicates \(A\) and \(M\) are semantically similar; \(B\) and \(N\) are semantically similar; \(A(a, b)\) is more important (dictated by the pragmatics of the context).

Network Setup: In a manner similar to the original ARCS model, PRISM uses information about the semantic similarity of predicates in the target and scent schemas to create a constraint network. Figure 3 depicts the network corresponding to Table 1: units in the network represent the predicates in the target and scent schemas, and the links between units represent correspondences between the predicates.

The most important units hypothesize that a scent schema is analogous to the target schema. These units have names of the form TARGET=SCENT. ("=" here means "corresponds to," not identity). If the target is P1 and the scent is S 1 , then the \(\mathrm{P} 1=\mathrm{S} 1\) unit represents a correspondence between them. If \(\mathrm{P} 1-1\) is a proposition in P1 that corresponds to proposition S1-1 in scent S1, then the unit \(\mathrm{P} 1-1=\mathrm{S} 1-1\) will have an excitatory link with the unit \(\mathrm{P} 1=\mathrm{S} 1\).

Excitatory links are also set up from a special semantic unit to predicate-predicate units based on the degree of
semantic similarity of the predicates (in Figure 3, there are excitatory links from the semantic unit to \((\mathrm{A}=\mathrm{M})\) and \((\mathrm{B}=\mathrm{N})\) because they are semantically similar according to our assumption). Similarly, excitatory links are also set up from a special pragmatic unit to predicate-predicate units that are considered more important than others (in Figure 3, there are excitatory links from pragmatic unit to \((\mathrm{A}=\mathrm{M})\) because predicate A was assumed to be more important than others). The activation level of the special semantic and pragmatic units is always kept at the maximum value of 1 . Thus, they serve to pump activation to all units that are linked to it.

Inhibitory links are constructed between units representing incompatible hypotheses, for example, between \(\mathrm{P} 1=\mathrm{S} 1\) and \(\mathrm{P} 1=\mathrm{S} 2\). These make utility calculation competitive, in that choosing one scent will tend to suppress choosing of an alternative.


Figure 3. Constraint satisfaction network for computing analogical similarity between target and scent schemas of

Table 1 (following Thagard et al. (1990), pp. 275).
Running the Network: The constraint network is run by setting the activation of all units to a minimal initial (random) level, except for the special semantic and pragmatic units for which activation is clamped at 1 . Then the activation of each unit is updated by considering the activations of those units to which it has links. Cycles of activation adjustment continue until all units have reached asymptotic activation. As in ARCS, the activation of unit \(j\) on cycle \(t+1\) is given by:
\[
a_{j}(t+1)=a_{j}(t)(1-d)+\text { enet }_{j}\left(\max -a_{j}(t)\right)+\text { inet }_{j}\left(a_{j}(t)-\min \right)
\]

Here \(d\) is a decay parameter, enet \(t_{j}\) is the net excitatory input, and inet \(_{j}\) is the net inhibitory input (a negative number), with minimum activation \(\min =-1\) and maximum activation \(\max =1\). Inputs are determined by the equations:
\[
\begin{aligned}
& \text { enet }_{j}=\sum_{i} w_{i j} o_{i}(t) \text { for } w_{i j}>0 ; \text { and } \\
& \text { inet }_{j}=\sum_{i} w_{i j} o_{i}(t) \text { for } w_{i j}<0
\end{aligned}
\]
\(o_{i}(t)\) is the output of unit \(i\) on cycle \(t: o_{i}(t)=\max \left(a_{i}(t), 0\right)\).
Updating the constraint network continues until all units have reached asymptote, that is, a cycle is reached at which the activation change of each unit is less than a specified value, typically a low number (e.g., 0.01). (See Thagard et al. (1990) for more details about setting up the activation network, running such a network, computational complexity, etc.)

Scent of a Patch: When the network settles, the similarity between a target schema, \(P\), and a particular scent schema, \(S_{i}\), is equal to the activation value of the unit \(P=S_{i}\) in the constraint network. Higher the activation accumulated by the unit \(P=S_{i}\) the more analogically similar is the scent schema \(S_{i}\) to the target. The scent of a particular patch, \(I P_{i}\), which is associated with a set of proximal cues, \(\left\{C_{i j}\right\}\), is equal to the similarity between the scent schema, \(S_{i}\), obtained from \(\left\{C_{i j}\right\}\), and the target schema, \(P\).

\section*{Explaining the Challenges of IAR using PRISM}

The findability challenge is attributable to the current keyword-based indexing and access mechanisms in which the Retrieval process in Figure 1 supports access to information based on literal similarity (word-for-word matching) while ignoring semantic, structural and pragmatic similarity. As a result, each attempt at access can contain a large number of superficially similar cases as opposed to cases that are truly analogous, which entails a lower average information yield per region. This yield is inversely proportional to the number of times the Formulate-RetrieveCompute Information Scent loop is executed in the PRISM model depicted in Figure 1: a low yield implies more executions of the cycle. Therefore, between-patch foraging time is higher, the period between successive useful finds is longer, and the frequency of encountering useful information resources is lower.

The recognizability challenge is attributable to the nature of proximal cues that the information seeker encounters in common online environments - specifically, their lack of affordance for accurately perceiving information scent. Perceiving the scent of an information resource in the context of analogical retrieval requires accurately judging the deeper similarity between that target and the source case as represented by their proximal cues. However, the design of proximal cues typically contains small snippets of information, which is insufficient to construct richer schemas. This likely explains why the designers made many recognition errors in our studies.

\section*{Conclusions}

In this paper, we identified interactive analogical retrieval (IAR) as an important phenomenon in the context of biologically inspired design. We provided a descriptive
account of the phenomenon based on our in situ studies of designers engaged in online search for biological analogues to their problems. Our descriptive account identified three main challenges associated with IAR: findability, recognizability, and understandability. Although our in situ studies were conducted in the context of a classroom environment, we posit that these cognitive challenges are quite general: the same challenges are likely to occur in actual practice of biologically inspired design because although practicing designers are experts in their design domain, they are likely to be novices in biology. We posit further that IAR is a general phenomenon: IAR is not limited to biologically inspired design, but occurs whenever people are searching for cross-domain analogies in external online information environments.

We also developed a causal model of IAR called PRISM combining Pirolli's (2007) information foraging theory (IFT) and Thagard et al.'s (1990) ARCS model of analogical retrieval. PRISM extends IFT to account for analogy seeking; it expands ARCS into a model of information scent perception. PRISM provides explanations for the findability and recognizability challenges of IAR we observed in our studies of biologically inspired design.

PRISM could help develop new technology for helping designers find biological analogues more efficiently and easily: the model predicts that the findability and recognizability issues could be addressed, respectively, by changing the indexing and access mechanism and enriching the proximal cues in online environments. Biologue (Vattam \& Goel 2011) is an interactive tool for supporting biologically inspired design based on the PRISM model.

In terms of cognitive theory, we view analogy as situated in external information environments. If we take the boundaries of the cognitive system as including online information environments, as seems to be the case in biologically inspired design, then the phenomenon of IAR becomes an important element of understanding the situatedness of analogical reasoning. By folding in interactions with external information environments, PRISM may provide a starting point to think about a general theory of situated analogy.

\section*{Acknowledgments}

We thank Michael Helms and Bryan Wiltgen for their contributions to this work. We are grateful to the instructors and students of the ME/ISyE/MSE/BME/BIOL 4740 class, especially Professor Jeannette Yen, the prime coordinator of the course. We thank the US National Science Foundation for its support through the CreativeIT Grant (\#0855916) Computational Tools for Enhancing Creativity in Biologically Inspired Engineering Design.

\section*{References}

Benyus, J. (1997) Biomimicry: Innovation Inspired by Nature. William Morrow.
Burstein, M., (1986) Concept formation by incremental analogical reasoning and debugging. In R. Michalski, J. Carbonell \& T. Mitchell (eds.), Machine Learning 2, Los Altos: Morgan Kauffman, pp. 351-370.

Carbonell J. (1986) Derivational analogy: A theory of reconstructive problem solving and expertise acquisition In Michalski, Carbonell \& Mitchell (eds.), Machine Learning 3, Los Altos: Morgan Kauffman.
Clement, J. (2008). Creative Model Construction in Scientists and Students: The Role of Imagery, Analogy, and Mental Simulation. Dordrecht: Springer.
Davies, J., Goel, A., \& Nersessian, N. (2009) A Computational Model of Visual Analogies in Design. Journal of Cognitive Systems Research 10:204-215.
Dunbar K. (2001) The Analogical Paradox. In Gentner, D, Holyoak, K., \& Kokinov, B. (Eds.) The Analogical Mind: Perspectives from Cognitive Science. MIT Press. Boston.
Forbus, K., Gentner, D., \& Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. Cognitive Science, 19(2), 141-205.
Gentner, D. (1983). Structure-mapping: a theoretical framework for analogy. Cognitive Science, 23, 155-170.
Gentner, D., \& Markman, A. (1997). Structure Mapping in Analogy and Similarity. American Psychologist, 52:45-56.
Hofstadter, D.H. (1995). Fluid concepts and creative analogies. New York: Basic Books.
Holyoak, K., \& Thagard, P. (1989). Analogical mapping by constraint satisfaction. Cognitive Science, 13(3), 295-355.
Indurkhya, B. (1992) On the Role of Interpretive Analogy in Learning. New Generation Computing, 8(4): 385-402.
Keane M. (1988) Analogical Problem Solving. Ellis Horwood. Chichester
Kokinov, B., \& Petrov, A. (2001) Integration of Memory and Reasoning in Analogy-Making: The AMBR Model. In Gentner, D., Holyoak, K., Kokinov, B. (editors) The Analogical Mind: Perspectives from Cognitive Science, MIT Press.
Kolodner, J. (1993). Case-based reasoning. San Mateo, CA: Morgan Kaufmann Publishers.
Kunda, M., McGreggor, K., \& Goel, A. K. (2013). A computational model for solving problems from the Raven's Progressive Matrices intelligence test using iconic visual representations. Cognitive Systems Research, 22-23, 47-66.
Nersessian, N. (2008) Creating Scientific Concepts. MIT Press Cambridge.
Pirolli, P. (2007) Information foraging theory: Adaptive interaction with information. Oxford: Oxford University Press.
Thagard, P., Holyoak, K. J., Nelson, G., \& Gochfeld, D. (1990). Analog retrieval by constraint satisfaction. Artificial Intelligence, 46, 259-310.
Vattam, S. (2012) Interactive Analogical Retrieval: Practice, Theory and Technology, Ph.D. Dissertation, Georgia Institute of Technology, December 2012.
Vattam, S., \& Goel, A. (2011) Semantic tagging of research articles for interdisciplinary design. In Proc. 6th International Conference on Knowledge Capture . Banff, Canada.
Vincent, J. \& Man, D. (2002) Systematic Technology Transfer from Biology to Engineering. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 360(1791): 159-173.
Yaner, P., \& Goel, A. (2006) Visual Analogy: Viewing Retrieval and Mapping as Constraint Satisfaction. Journal of Applied Intelligence, 25(1):91-105.
Yen, J., Weissburg, M., Helms, M., \& Goel, A. (2011) Biologically inspired design: a tool for interdisciplinary education. In Biomimetics: Nature-Based Innovation, Y. BarCohen (editor), Taylor \& Francis.

\title{
Linking Cognitive Tokens to Biological Signals: Dialogue Context Improves Neural Speech Recognizer Performance
}

\author{
Richard Veale (riveale@indiana.edu) \\ Indiana University, 841 Eigenmann Hall \\ Bloomington, IN 47406 USA
}

Gordon Briggs (gbriggs@cs.tufts.edu)
Tufts University, 200 Boston Ave. Medford, MA 02155 USA

\author{
Matthias Scheutz (matthias.scheutz@tufts.edu) \\ Tufts University, 200 Boston Ave. Medford, MA 02155 USA
}

\begin{abstract}
This paper presents a hybrid cognitive model engaged in experiments demonstrating a successful mechanism for applying top-down contextual bias to a neural speech recognition system to improve its performance. The hybrid model includes a model of social dialogue moves, which it uses to selectively bias word recognition probabilities at a low level in the neural speech recognition system. The model demonstrates how symbolic and neurologically inspired components can successfully exchange information and mutually influence their processing. Furthermore, the biasing mechanism is grounded in brain mechanisms of perceptual decision making.
\end{abstract}

Keywords: Speech Recognition; Liquid State Machine; Dialogue Context; Top-Down Bias; Signal-to-Token Conversion

\section*{Introduction}

Human cognition comprises high-level knowledge-based processes as well as low-level perceptual and motor processes, both of which are implemented via electro-chemical mechanisms in the brain. High-level cognitive processes are often viewed as symbolic and discrete, while low-level perceptual and motor processes are subsymbolic and continuous. Moreover, high-level processes are taken to operate on structured representations, while low-level processes will usually not be representational at all. Two key challenges in cognitive science are thus to understand (1) how high-level processes are realized in "neural hardware" and (2) how they can interact with low-level processes (e.g., how discrete symbolic knowledge can influence continuous subsymbolic processes and vice versa). We will focus on the second challenge in this paper.

Connectionist computational modeling has made significant progress in addressing (1) over the years, producing more and more refined neurologically plausible models of cognitive functions which are verified physiologically (e.g. (Machens, Romo, \& Brody, 2005)). However, fewer efforts have been made to address (2). Only recently, hierarchical Bayesian models have been proposed as a natural, systematic way to connect higher-level to lower-level processes (Kemp \& Tenenbaum, 2008). Similar to the Bayesian approach, our goal is to understand the interactions between these two types of processes which operate at fundamentally different levels.

Hierarchical Bayesian modeling often focuses on the "computational level" (Marr, 1982), showing how higherlevel processes can influence lower levels (e.g., by showing how distributions of higher-level structures constrain distributions of lower-level items). In contrast, our approach attempts to address all three levels and their mutual interactions. This is because these levels cannot be considered in complete isolation in cases where higher-level processes have to interact with lower-level processes in real-time contexts with realworld inputs. Specifically, we claim that the nature and timecourse of low-level processes imposes significant constraints on the possible ways of exchanging information with higherlevel processes. Low-level processes will limit the types of computations that are allowed in higher-level processes that communicate with them, since they may have stringent timing requirements and will not wait for a computation to finish with a result. Proposals that do not incorporate those constraints might result in models that produce correct results under some empirical regimes, but which are infeasible given implemenation constraints.

For example, a hierarchical Bayesian model of natural language processing might be able to show that high-level knowledge about grammar can successfully bias low-level speech processing, but whether that particular computational way of biasing is actually feasible and realistic in humans can only be determined by taking algorithmic and implementation constraints into account. These constraints include time bounds caused by the incremental nature of the speech processor. In this case the high-level computation can not expect to have access to a whole utterance before it starts biasing, since by that point the speech processor will already have advanced past the point where it is useful. Thus, although there are many ways in which higher levels could influence lower levels at the computational level, most of them are not realized in humans because of implementation or algorithmic constraints.

This paper makes three contributions: first, we will present a general way of integrating high-level processes operating on structured symbolic knowledge with low-level neural pro-
cesses with unstructured signals; second, we will show in the specific context of real-time biologically plausible speech recognition how high-level knowledge about diologues and mental states of interlocutors can be used to dynamically adjust parameters in the neural speech recognizer to improve recognition performance; and third, we will provide results from a real-time evaluation of the implemented model. The model includes a biologically plausible neural speech recognizer, a statistical/symbolic natural language understanding system, and a logic-based model of pragmatical and mental state inference. Previously, we have addressed the bottomup transfer of information, i.e., conversion from the continuous stream of auditory neural firings to symbolic word tokens expected by a natural language processing system (Veale \& Scheutz, 2012b). In this paper we address the reversed direction, the top-down transfer of information and biasing of low-level processes. Specifically, high-level knowledgebased representations of dialogue and interaction context will be used to bias auditory neural activity to improve word recognition performance in spoken language dialogues.

\section*{Background}

In humans and other animals, perceptual decisions are modulated by system state in a top-down manner. Top-down biases have been documented empirically in a variety of contexts such as vision search (Chen \& Zelinsky, 2006), perceptual decision about motion (Hanks, Mazurek, Kiani, Hopp, \& Shadlen, 2011), auditory disambiguation (Hannemann, Obleser, \& Eulitz, 2007)), and others. Furthermore, we are beginning to understand the mechanisms underlying these biases thanks to a combination of neurophysiological studies and behavioral research (e.g. see (Hanks et al., 2011). Perceptual decisions can be well-modelled using parallel diffusion processes (Ratcliff, Gomez, \& McKoon, 2004), and there is evidence that these processes are realized in the brain as neural integrators collecting evidence for each alternate hypothesis independently. Prior probabilities influence the neural integrators based on the past experience of the organism. These influences have been shown to be caused by top-down biases, although some evidence exists that sensory cortex parameters also adapt to match environmental priors (Fiser, Chiu, \& Weliky, 2004), which are outside the scope of this paper (Veale \& Scheutz, 2012a). The shape and parameters of the thresholds and the bias functions responsible for topdown biases on behavior are still under active investigation (Hanks et al., 2011). However, the detailed behavior of these processes is not necessary to implement a working model that takes advantage of the general mechanism of top-down bias to improve perceptual decisions.

In this paper we are specifically interested in top-down biases on auditory word recognition. Contextual biases on word recognition are ubiquitous in the everyday world. For example, visual context and gesturing can be used in noisy situations to produce a sensible hypothesis for what a speaker is saying. This is not a novel observation. Top-down bi-
asing of speech recognition probabilities have been investigated in a traditional speech recognition system (e.g. (Young, Hauptmann, Ward, Smith, \& Werner, 1989)). Our work differs from this previous work in that the speech recognition system is built of biologically-plausible neural circuits modelling the early human auditory system. Although the general concept of using context to bias state in the speech recognizer is similar, the non-symbolic nature of the speech recognizer in our system requires serious reconsideration of how to actually implement the top-down bias. In this paper we adopt a simple approach and bias the temporal integrators representing the competing word categories, which directly influences the symbolic output of the speech recognizer.

The next section presents a short overview of the two most relevant portions of the hybrid model used in this paper. It describes the mechanism for top-down biasing of the neural speech recognizer, and overviews how the system operates.

\section*{Model Overview}

The architecture of the cognitive model used for the experiments in the Experiment Setup Section is summarized in Figure 1. The neural speech recognizer (LSM ASR) is responsible for translating the acoustic signal into text tokens, which are sent to the NLP component. The NLP component parses the text tokens, and performs semantic analysis and utterance type classification. The dialogue system receives semantic information from NLP and updates the agent's beliefs, based on a pragmatic analysis (Briggs \& Scheutz, 2011). The dialogue component also tracks the state of the current dialogue exchange, allowing for predictions about expected upcoming utterance types. Details of how biasing is implemented in the speech recognizer and Dialogue components are presented in the sections below. The model is implemented in the DIARC cognitive architecture (Schermerhorn et al., 2006), whose natural language capabilities have been demonstrated in humanrobot interaction scenarios \({ }^{1}\) (Cantrell, Scheutz, Schermerhorn, \& Wu, 2010; Cantrell, Schermerhorn, \& Scheutz, 2011; Briggs \& Scheutz, 2012).

\section*{The Dialogue Component}

The dialogue component contains knowledge of common dialogue exchange patterns, such as those in Table 1.

Table 1: Dialogue exchange patterns
\begin{tabular}{r|l}
\hline Exchange Pattern & Dialogue Move Sequences \\
\hline Statement-Ack Pair & \(\operatorname{Stmt}(\alpha, \beta) \rightarrow \operatorname{Ack}(\beta, \alpha)\) \\
Yes-No QA-Pair (pos) & \(\operatorname{AskYN(\alpha ,\beta )\rightarrow \operatorname {Reply}Y(\beta ,\alpha )}\) \\
\(\rightarrow \operatorname{Ack}(\beta, \alpha)\) \\
Yes-No QA-Pair (neg) & \(\operatorname{AskYN(\alpha ,\beta )\rightarrow \operatorname {ReplyN}(\beta ,\alpha )}\)\begin{tabular}{r}
\(\rightarrow \operatorname{Ack}(\beta, \alpha)\) \\
QA-Pair \\
\\
\\
\(\operatorname{AskWH}(\alpha, \beta) \rightarrow \operatorname{Stmt}(\beta, \alpha)\) \\
\\
\(\rightarrow \operatorname{Ack}(\alpha, \beta)\)
\end{tabular} \\
\hline
\end{tabular}

\footnotetext{
\({ }^{1}\) http://www.youtube.com/watch?v=RJ1VSIi1CM4
}


Figure 1: Information flow through the natural language system. The blue arrow indicates the top-down dialogue context bias on the ASR component introduced in this paper.
\(\operatorname{Stmt}(\alpha, \beta)\) denotes a statement utterance direct from agent \(\alpha\) to agent \(\beta\), while \(\operatorname{Ack}(\beta, \alpha)\) denotes an acknowledgment (e.g. "okay") from \(\beta\) to \(\alpha\). AskYN and AskWH denote a yesno question and general question, respectively.

In this paper we focus on sending bias information to the LSM ASR component in the case of yes-no question-answer (QA) pairs. When the dialogue component detects a yes-no QA-pair has been entered, it sends a list of expected words to the LSM ASR component, specifically "yes" (ReplyY) and "no" (ReplyN). For each expected word \(x_{i}\), a weight value \(0 \leq w_{i} \leq 1\) is also sent to the LSM, denoting how much to weight \(x_{i}\) relative to other biased words (where 0 is equivalent to no bias and 1 indicates maximum bias).

\section*{The Speech Recognizer}

The neural speech recognition system employed in this paper has previously been used to perform speech recognition for real-time human-robot interaction tasks (Veale \& Scheutz, 2012b). The system converts from speech input streams to word tokens that can be used by other components of the cognitive model. The speech recognizer employs the liquid state machine (LSM) computational paradigm (Maass, Natschlager, \& Markram, 2002) to perform recognition on audio input streams. The LSM is implemented using spiking neurons, and readouts are trained via linear regression. Figure 2 presents the main components of the speech recognizer.

Sound is processed into auditory nerve firings corresponding roughly to the strength of frequency channels in auditory input (Figure 2, left). These neurons project to several groups of pre-processing neurons (superior olivary complex) via groups of differently parameterized synapses, resulting in neurons sensitive to the onset/offset/passthrough activity for each cochlear channel. These pre-processing neurons in turn project randomly to the recurrent circuit (liquid), which is a large circuit of randomly connected spiking neurons. "Readouts" (discussed below) are trained via linear regression on a corpus of sound files, with supervisor vectors set to +1 for all instances of the target category and -1 otherwise. Additionally, all readouts are counter-trained against a "noise" corpus in which every readout's supervisor vector is -1 .

Signal-to-Token Conversion Readouts (perceptrons) are trained via linear regression to respond positively to liquid
activity patterns similar to liquid activity patterns evoked by the word examples they were trained on. Readouts are integrated over time with exponential decay (low-pass filtered, time constant 20 ms ), and the value of these are continuously summed into the diffusors (right). In the model, readouts, integrators, and diffusors are only updated every 20 ms . The value of the readout integrator for readout \(r, \sigma_{r}\) is thus defined by the following equation (where \(\tau_{\sigma}\) is the time constant and \(I_{r}\) is the input from the corresponding readout):
\[
\begin{equation*}
\frac{\partial \sigma_{r}}{\partial t}=\frac{-\sigma_{r}}{\tau_{\sigma}}+I_{r} \tag{1}
\end{equation*}
\]

The diffusors compete with one another proportional to how strong their input is. The value of readout \(r\) 's diffusor, \(\Delta_{r}\), is updated according to the following rule:
\[
\begin{equation*}
\Delta_{r}(t)=\left(\Delta_{r}(t-1)+\sigma_{r}\right) \cdot \frac{\sigma_{r}}{\sum_{j}\left(\sigma_{j}\right)} \tag{2}
\end{equation*}
\]

This mechanism prevents the diffusion processes of ambiguous words from reaching threshold simultaneously. Using this system, there must be the equivalent of 100 ms of strong unambiguous evidence for a particular word category before it crosses threshold. This evidence could be provided by longer but weaker evidence, or by top-down bias.

Biasing Mechanism The biasing mechanism functions by injecting energy into the readout integrators, i.e., one level before the diffusion processes. The biaser specifies which categories should be biased, and the relative strengths for those biases. In the current paper, the amount of energy injected with a unit strength of 1.0 is equal to amount that is injected when the corresponding readout is active, thus up to "doubling" the input to the integrator at times when its presynaptic readout is active. Note that this implements the "simplest" diffusion model bias, involving linear bias to the diffusor's input diffusing to a constant threshold.

The result of bias is that biased words have "stronger" responses from their internal integrators, which translates to greater force of growth towards the diffusion threshold. This results in both faster reporting of the word (when the diffusor crosses threshold), and also stronger "confidence" in the word when the words offset is reported at the end of the word.


Figure 2: Visualization the neural model described in this paper. The pictured circuit has only 4 input channels, and a \(3 \times 3 \times 10\) recurrent circuit. The actual circuit has 84 input channels and a \(5 \times 4 \times 20\) recurrent circuit.

The LSM ASR was trained on five spoken instances of eight different words from the same speaker: yes, no, guess, bess, jess, joe, bob and a null response (background noise). The audio files used for testing are the same words spoken by a different speaker of the same gender. The words were chosen because several rhyme or have similar phonetic components to the "target" words "yes" ("guess", "bess", "jess") and "no" ("joe"), or share none ("bob").

The scenario we examine in this study consists of a simple yes-no QA-pair. The system is initiated with an intention to know whether its interlocutor possesses a particular mug in the belief component. The dialogue component, which queries the belief component for intentions to know information, generates the appropriate yes-or-no question:
```

Robot: Do you have the mug?

```

After this NL reuest is generated, a response audio file is presented to the system. These audio files consist of "yes" and "no" responses recorded from a different speaker. Four conditions were examined: (1) "Yes" response, no bias; (2) "Yes" response, with bias; (3) "No" response, no bias; (4) "No" response, with bias. Data from the LSM (integrated readout activity and word recognition score) was recorded at 10 millisecond intervals over the duration of the input.

\section*{Results}

The time course of the diffusors (solid lines) and readout integrators (dashed lines) for every word category are shown in Figures 3a (a "yes" trial) and 3b (a "no"trial). The primary comparison to make is the difference in the trajectories between the biased (each figure, bottom) and unbiased (each figure, top) trials. If the top-down biasing is working correctly, one should see a jump in activity over the unbiased trials for the contextually-appropriate words ("yes" and "no"), and no corresponding jump in any other words. This is precisely what is observed: even accidental weak responses to incorrect words ("bess" - purple in Figure 3b) do not seem to change
significantly between biased and unbiased trials, whereas response to the appropriate word ("no", yellow) does. Similarly for Figure 3a, the activation of the contextually-inappropriate yet similar-sounding word "jess" (teal) does not change significantly between the biased and unbiased cases, yet the activation of the contextually-appropriate yet incorrect word ("no", yellow) is increased. Meanwhile, the activation of the contextually-appropriate and correct word ("yes", red) is stronger in the biased case and quickly advances to threshold.

As a control, a third set of experiments were run in which the responder responded with the similar-sounding but contextually-inappropriate word "joe" (Figure 4). In this case, the trajectories for all words do not differ significantly between the bias and unbiased conditions. However, in the biased condition (Figure 4, bottom), a slight jump in the recognition of the contextually-appropriate word "no"is seen near the end of the utterance. This is expected because the tail end of "no" is similar to "joe", and the additional contextual bias on "no" was sufficient to produce a small amount of drift in the diffusor for the period of similar sounds.

\section*{Experimental Setup}

In terms of quantifying the advantage, one can look at the point at which recognition of the word reaches the confidence threshold (black horizontal bar). The diffusor in the "yes" unbiased condition (Figure 3a, top) crosses the recognition threshold at approximately 540 ms , whereas with bias the diffusor crosses the recognition threshold at approximately 470 ms (bottom), demonstrating a reduced recognition time. Note that the readout values for both "yes" and "no" responses are significantly increased in the biased condition compared to the unbiased condition, as both are anticipated as possible answers (whereas the readouts for the other word nodes remain relatively unchanged in amplitude). In the "no" unbiased condition, the diffusor crosses the recognition threshold at approximately 480 ms (Figure 3b, top), whereas with bias the diffusor crosses the recognition threshold at approximately 360 ms , again demonstrating a a reduced recognition


Figure 3: LSM ASR responses to "yes" and "no" stimuli in no bias condition (top) and bias condition (bottom). The trajectory of activity for the readouts and diffusors for all trained words in response to the injected sound is plotted over time after the question is asked. Dotted lines represent the individual readout integrators for each word, while solid lines represent the diffusors. In both cases, the diffusor for the correct word (red solid line on left, yellow solid line on right) crosses the threshold significantly quicker in the bias condition (bottom). The influence of the top-down bias mechanism can be clearly seen in the increased activity of the readout integrators for "yes" and "no" (red and yellow dotted lines, respectively) in the bias condition.
time. Keep in mind that these are different words that begin at slightly different times and which extend for different amounts of time and have different volumes and distances from the training corpus. Thus it is important to focus on the differences within a word to see the performance increases resulting from top-down biasing.

\section*{Future Work}

Expanding and refining the contexts in which top-down biasing of the speech recognizer will occur will provide ample opportunities for future research. For instance, incremental parse hypotheses in the NLP component could be used to identity likely upcoming words. Certain sentential modifiers (e.g. "I am now at the store" vs. "I am still at the store") can be used in conjunction with belief models and contextual knowledge for prediction purposes. If, for example, common ground in the dialogue exchange was established such that both speaker and listener knew the speaker was at the store previously. The partial sentence, "I am still at-" would be highly indicative of "..at the store". These semantic and belief model implications of these modifiers can be reasoned about in our pragmatics system (Briggs \& Scheutz, 2011). Additionally, some yes-no questions are actually conventionally indirect forms of general questions. For instance, "Do you know who has the mug?" is often an indirect form of, "Who has the mug?" and may elicit a name in response. Our natural
language system has mechanisms of recognizing and reasoning about such indirect speech acts (Briggs \& Scheutz, 2013, forthcoming), and therefore more precise biasing algorithms are ripe for investigation.

A more theoretically interesting extension of the current work will more directly address the theoretical issues from the introduction. In the current paper, only pseudo-symbolic readout neurons were influenced by the top-down bias. This allowed us to explore the time-constraint theme, but not the disconnects in representation between multiple levels. In the future it will be interesting to directly bias the state of the auditory circuit, to further explore how such interactions could take place.

\section*{Conclusion}

This paper introduced a hybrid neural-symbolic model that demonstrates not only the bottom-up communication of cognitive tokens from continuous sensory streams, but also the top-down biasing of neural speech recognition using predictions based on expected dialogue moves. The top-down biasing of the neural speech recognizer results in faster and more confident word recognition for contextually appropriate word categories during dialogue exchanges. The top-down biasing mechanisms are biologically accurate in that the effect of the top-down signal on high-level neurons in the speech recognition circuit parallels that observed in "diffusion" neurons


Figure 4: LSM readout results for "joe" response for no bias condition (top) and bias condition (bottom). Conventions are equivalent to figure 3. The diffusor for the actual uttered word "joe" does not significantly differ between the biased and unbiased conditions, crossing the threshold (black horizontal line) at roughly the same point in both conditions.
recorded from primate association cortex. The hybrid model presented in this paper engages interesting questions regarding interaction between different levels of abstraction. We use this to highlight that implementation-level details can actually constrain the computational level in real-time real-world situations. We believe that it is important to keep this relationship in mind when making claims about human cognition.

\section*{Acknowlegements}

This research was in part supported by ONR grant \#N00014-11-1-0493. RV is an NSF GRF and NSF IGERT.

\section*{References}

Briggs, G., \& Scheutz, M. (2011, June). Facilitating mental modeling in collaborative human-robot interaction through adverbial cues. In Proceedings of the sigdial 2011 conference (pp. 239-247). Portland, Oregon.
Briggs, G., \& Scheutz, M. (2012). Multi-modal belief updates in multi-robot human-robot dialogue interaction. In Proceedings of 2012 symposium on linguistic and cognitive approaches to dialogue agents.
Briggs, G., \& Scheutz, M. (2013, forthcoming). A hybrid architectural approach to understanding and appropriately generating indirect speech acts. In Proceedings of the 27th aaai conference on artificial intelligence.
Cantrell, R., Schermerhorn, P., \& Scheutz, M. (2011, July). Learning actions from human-robot dialogues. In Proceed-
ings of the 2011 ieee symposium on robot and human interactive communication.
Cantrell, R., Scheutz, M., Schermerhorn, P., \& Wu, X. (2010, March). Robust spoken instruction understanding for HRI. In Proceedings of the 2010 human-robot interaction conference.
Chen, X., \& Zelinsky, G. (2006). Real-world visual search is dominated by top-down guidance. Vision research, 46(24), 4118-4133.
Fiser, J., Chiu, C., \& Weliky, M. (2004). Small modulation of ongoing cortical dynamics by sensory input during natural vision. Nature, 431(7008), 573-578.
Hanks, T., Mazurek, M., Kiani, R., Hopp, E., \& Shadlen, M. (2011). Elapsed decision time affects the weighting of prior probability in a perceptual decision task. The Journal of Neuroscience, 31(17), 6339-6352.
Hannemann, R., Obleser, J., \& Eulitz, C. (2007). Topdown knowledge supports the retrieval of lexical information from degraded speech. Brain research, 1153, 134143.

Kemp, C., \& Tenenbaum, J. B. (2008). The discovery of structural form. PNAS, 105(31), 10687-10692.
Maass, W., Natschlager, T., \& Markram, H. (2002). Realtime computing without stable states: A new framework for neural computation based on perturbations. Neural Computation, 14(11), 2531-2560.
Machens, C., Romo, R., \& Brody, C. (2005). Flexible control of mutual inhibition: a neural model of two-interval discrimination. Science (New York), 307(5), 1121-1124.
Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information. New York, NY, USA: Henry Holt and Co., Inc.
Ratcliff, R., Gomez, P., \& McKoon, G. (2004). A diffusion model account of the lexical decision task. Psychological Review, 111(1), 159.
Schermerhorn, P., Kramer, J., Brick, T., Anderson, D., Dingler, A., \& Scheutz, M. (2006). Diarc: A testbed for natural human-robot interactions. In Proceedings of aaai 2006 robot workshop (pp. 1972-1973).
Veale, R., \& Scheutz, M. (2012a, November). Auditory habituation via spike-timing dependent. In Proceedings of the international conference on development and learning and epigenetic robotics. San Diego, CA.
Veale, R., \& Scheutz, M. (2012b). Neural circuits for anytime phrase recognition. In N. Miyake, D. Peebles, \& R. P. Cooper (Eds.), Proceedings of the 34th annual conference of the cognitive science society (p. 1072-1077). Austin, TX: Cognitive Science Society.
Young, S., Hauptmann, A., Ward, W., Smith, E., \& Werner, P. (1989). High level knowledge sources in usable speech recognition systems. Communications of the ACM, 32(2), 183-194.

\title{
A Written Version of Sign Language can Enhance Signing Deaf Individuals’ Comprehension and Learning from Texts
}

\author{
Mara Vendrame (mara.vendrame@unito.it)* \\ Ilaria Cutica (ilaria.cutica@unimi.it) \({ }^{\circ}\) Monica Bucciarelli (monica.bucciarelli@unito.it)* \\ *Department of Psychology, Via Po 14, 10123 Turin, Italy \\ \({ }^{\circ}\) Department of Economy, Management and Quantitative Methods, Via Conservatorio 7, 20132 Milan, Italy
}

\begin{abstract}
Deaf individuals have difficulties in comprehending written text, as well as oral language. As a consequence, learning from text is compromised in deaf individuals. We hypothesized that a transposition of the Italian Sign Language to its written counterpart could enhance signing deaf individuals’ comprehension and learning from text. We confirmed our prediction for comprehension and learning for technical texts in Experiment 1 and for narrative texts in Experiment 2; signing deaf individuals' text comprehension and learning therefore benefit from a written language whose structure reflects the structure of their visual-spatial sign language. We speculate that, for signing deaf individuals, practice in reading written sign language texts might positively affect the ability to comprehend the written oral language texts.
\end{abstract}

Keywords: deaf individuals; text comprehension; learning; Italian Sign Language

\section*{Introduction}

Those who are unfamiliar with deafness may assume that the deaf individuals' auditory deficit can easily be circumvented through the use of written communication: if you have hearing problems, we can easily communicate through written texts. This naïve assumption disregards the nature of profound deafness. The ability to understand written texts presupposes high linguistic competence such as the ability to integrate information from different parts of a text and to derive its inner coherence. Due to their profound hearing loss, prelingually deaf individuals, who have never experienced oral language, have difficulties in comprehending the lexical, morphosyntactic, and pragmatic aspects of written verbal language (Van Hoogmoed et al., 2011). In addition, compared to hearing individuals, deaf individuals are less able to comprehend and remember details from a written text and to reason about the information contained in it (Marschark \& Wauters, 2008). Their specific difficulty in comprehending the holistic meaning of written texts seems to derive from difficulties in connecting different information together and in drawing inferences (Miller, 2002). Indeed, prelingually and profoundly deaf individuals possess adequate single-word
reading ability and vocabulary knowledge (Oakhill \& Cain, 2000). More generally, deaf individuals’ poor linguistic competence must be imputed to their atypical cognitive development (Marschark \& Hauser, 2008). For a start, in hearing individuals, sound plays a role from the earliest months of life in organizing visual attention: when a new event is signaled by sound, visual attention may be shifted appropriately (Smith et al., 1998). Hearing people use audition to monitor both their immediate and distant environment for changing events, while allowing vision to focus narrowly on the task at hand. In deaf individuals, the limited access to auditory information alters the way visual attention skills are deployed: visual attention becomes responsible for both focusing on the task at hand and monitoring events elsewhere in the visual field (Mitchell \& Maslin, 2007). As a consequence, deaf individuals tend to adaptively develop a more spatially distributed visual attention, whereas highly selective visual attention tends to prove difficult (Bavelier, Dye \& Hauser, 2006).
Auditory deprivation also has a direct impact on memory capacity: when hearing individuals are requested to remember simple stimuli such as words, pictures, or numbers, they tend to use a verbal-sequential coding of a phonological or acoustic nature (Marschark \& Mayer, 1998). Deaf individuals appear instead to rely heavily on visuo-spatial coding: their incomplete mastery of language skills detracts from using language as an effective cognitive tool. Consequently, deaf people tend to have a shorter memory span for linguistic materials, compared to hearing people (Logan, Mayberry \& Fletcher, 1996). By contrast, deaf people perform as well as or even better than hearing people on tasks that involve visual or spatial processing (Cattani, Clibbens \& Perfect, 2007).
Furthermore, it has been shown that deaf individuals, in comparison to hearing individuals, have more difficulty with abstract reasoning (Marschark \& Hauser, 2008). In particular, they have difficulties in verbal-analogical reasoning, which requires high-level linguistic skills, and the ability to understand not simple items but complex structures (Edwards et al., 2010). By contrast, deaf people
are not impaired at perceptually based reasoning: they perform as well as hearing people on non-verbal cognitive tasks that do not require the overt use of verbal language, such as figural-geometric analogy tasks, and in visualspatial information processing (Marschark \& Hauser, 2008). However, deaf individuals' moderate skills with abstract reasoning are also due to their broader difficulty with verbal language (Easterbrooks \& Scheetz, 2004).
All this considered, providing deaf individuals with suitable forms of written materials to support their comprehension and learning from texts in educational contexts, is a very important challenge. The focus of our investigation are signing deaf individuals, who are exposed to a natural sign language at birth. Sign languages exhibit grammatical structure at all linguistic levels. However, the acquisition of sign languages features constraints unique to the visual modality (Morgan, Barrett-Jones \& Stoneham, 2007).

\section*{A Written Form of Sign Language}

Sign language is visuo-spatial in nature and has no written counterpart. Some attempt were made to devise appropriate means for representing sign languages: examples are Stokoe-based notations for notating single, decontextualised and standard signs (Pizzuto, Rossini \& Russo, 2006), and Sign Writing, a writing formalism based on transcription of manual and also non-manual elements of non-standard signs and complex units through symbols (Sutton, 1999). These methods require a training to learn to interpret the proposed notations.
Within a less ambitious perspective, we reasoned that some of sign language's features could, however, be reflected in its transposition to a written form. We assumed that such a written sign language might improve signing deaf individuals' comprehension and learning from text by promoting the activation of the visual thought schemata that are activated by the sign language itself (Wilbur, 2000). In particular, our assumption is based on considerations concerning the structural features of written sign language along with the particular cognitive functioning of signing deaf individuals.
First, the written form of sign language offers deaf individuals the possibility to process written linguistic information provided in a syntactic structure that reflects the structure of the corresponding sign language. In sign languages, space has a grammatical function, i.e. it is used to create and maintain cohesion among the different parts of the discourse (Morgan et al., 2008). Thus, for example, sentences in sign languages begin by identifying one or more loci in the spatial mapping, "the process used by the signer to reflect mental representations in physical space for reference and subsequent coreference in discourse as a cohesive device" (Winston, 1995, p.87). Subsequently,
signers point to a precise locus in space in order to evoke to the interlocutors the element that was originally 'placed' there. The particular function of space in sign languages generates a different discourse structure that has no counterpart in oral languages (Pyers et al., 2010).
Second, the text of a phrase in an oral language is longer than the corresponding text in the written sign language version: the written sign language text, like the sign language itself, lacks articles, prepositions, conjunctions, pronouns, and verbal auxiliaries. This claim holds at least for Italian Sign Language (LIS), which makes little use of finger spelling. As signing deaf individuals have a shorter memory span for linguistic material than do hearing individuals, they should benefit from this feature of the written sign language.
Third, signing deaf individuals, as compared to hearing individuals, have a more spatially distributed visual attention; this cognitive peculiarity, along with the consideration that a phrase in written sign language is shorter than the corresponding written phrase from the oral language, leads to the hypothesis that signing deaf individuals can extract in a glance more information from the written sign language text than from the written oral language text.
We tested the prediction derived from our assumptions on Italian signing deaf individuals.

\section*{Experiment 1: Does Written LIS Facilitate Text Comprehension in Signing Deaf Individuals?}

The deaf participants in the experiment were invited to carefully read two texts; they were then invited to recall as much information as they could. Each participant was presented with one text in Italian and another in LIS. We predicted a better recall for the LIS text. Hence, although we did not measure the participants' reading abilities, we made each participant act as his/her own control in the two experimental conditions.

\section*{Method}

Participants Twelve signing deaf adult individuals (5 females and 7 males; mean age: 26 years) with a prelingual and profound hearing deficit (>90 dB hearing loss) and no other disabilities voluntarily took part in the experiment. They were all university students who learned the LIS as their first language from their first year of life.

Materials and Procedures The experimental materials comprised two technical written Italian texts, one concerning the principles of how airplanes fly (Airplane flight, 312 words), and one about the effect on individuals produced by color perception (Responses to color, 315 words) (for excerpt see Appendix 1). For each text, we produced a written LIS version, parallel to the written

Italian version (266 and 243 words for Airplane flight and Responses to color, respectively). To create the written LIS version, a native-speaking signing deaf Italian university student read each Italian text carefully several times and was then video-recorded while translating the text into LIS. She then transcribed the signs produced in the translation into Italian words. The punctuation was introduced for each pause, in order to segment the different phrases, taking into account both manual (signs) and non-manual (facial expressions and body movements) markers occurring simultaneously. Consider, for example, the following excerpt from the Airplane flight written Italian text: "When an aeroplane is in flight, the air divides as it hits the front of the wing. Some of it flows over the upper part of the wing, and the rest over the lower part. The two air flows come together again behind the wing." As an example of the results of the translation, consider the parallel written LIS version of the excerpt above: "Example, plane flies, wing air hits wing in front of, then air divided 2, to go wing over, to go wing under, after air together to go wing behind." Obviously, the English translation of the written LIS texts is not equivalent to the result of transposing the British Sign Language or American Sign Language texts to their written counterparts.

The translations of the two Italian texts to written LIS were evaluated individually by a LIS interpreter and by a LIS deaf teacher to ascertain that the translations as provided by a native signing deaf individual were also acceptable on behalf of them. For the goal of our investigation, it is important to test the beneficial effect of the written sign language when realized by a native signing deaf individual, naïve with respect to trainings and education to become either an interpreter or a deaf teacher. At first, the interpreter and the deaf teacher were invited to watch carefully the two videos, one at a time, and afterwards they considered each single sign produced, taking into account both the manual and the non-manual components. For each semantic unit they were invited to evaluate the appropriateness and comprehensibility of the LIS translation through the following judgments: "Not at all adequate", "Barely adequate", "Adequate on the average", "Adequate". On average, the \(93 \%\) of the semantic units from the Airplane flight text, and the \(96 \%\) of the semantic units from the Responses to color text were judged as at least "adequate on the average"; none of the translations were judged as not adequate at all.
The participants encountered both texts (Airplane flight and Responses to color), one in Italian and the other in LIS. In each group, half of the participants dealt first with the Airplane flight text and the other half with the Responses to color text, so that, overall, the occurrence of each text in the Italian version and the LIS version was counterbalanced. The participants were invited to read each text carefully, one at a time, with no time limits; as soon as they finished reading each text, they were asked to recall as much
information as they could. The recollection was in LIS. All of the participants were video-recorded.
To code the results, each text was divided into 41 semantic units, corresponding to as many main concepts as the hearer could recall. For each text, there is a strict correspondence between the semantic units in the two versions (Italian and written LIS). Two independent judges coded the participants' recollections individually; the judges reached a significant level of agreement on their first judgments for the overall group of participants in the two experimental conditions, calculated using Cohen's K (Cohen's K ranging from .87 to .97 , p always <.001). For the final score, the judges discussed each item on which they disagreed, until they reached full agreement. Each concept (i.e., semantic unit) recalled by the participants was evaluated according to the following coding schema (see also Cutica \& Bucciarelli, 2008; 2011, Vendrame, Cutica \& Bucciarelli, 2010):
Correct recollection: a semantic unit recollected either literally or as a paraphrase;
Discourse-based inference: a recollection in which the participant gives explicit information that was originally implicit in the semantic unit;
Elaborative inference: a semantic unit recollected with the addition of plausible details;
Error: a recollection whose meaning is inconsistent with the semantic unit.
Each participants' recollection was coded as pertaining to just one category. Correct recollections and discourse-based inferences were considered indicators of comprehension and learning from text. Consider, for example, the following semantic units in the Italian color text: "He observed that the function deteriorated in low light but increased in bright light" and the following recollection by a participant: "Example: bright light finger to tap fast; low light finger to tap slow; finger to tap normal normal light". According to the coding schema, we considered the statements "Example: bright light finger to tap fast" and "low light finger to tap slow" as correct recollections, and the statement: "finger to tap normal normal light" as a discourse-based inference.
As a further example, considering the semantic unit in the Written LIS aeroplane text "Wing over this is pressure less", the recollection "the air under pressure to increase, it makes a support, an help" has been coded as a discourse-based inference, whereas the sentence "the pressure to increase wing over" has been coded as an error.

Results and Discussion The two texts were comparable in difficulty; considering each type of recollection separately, we found no differences in performance with the two texts in either the LIS or the Italian versions (unpaired T-test: t (22) comprised between .0 and 1.48 , p comprised between .15 and .1 ). Hence, we pooled together the results for the two Italian versions and those for the two LIS versions.

Table 1 illustrates the mean types of recollection for both the LIS and the Italian versions of the texts. The results show that they produced more correct recollections and fewer errors in the written LIS version than in the written Italian version ( T -test: \(\mathrm{t}(11)=3.43\), tied \(\mathrm{p}=.003\), and \(\mathrm{t}(11)=3.095, \mathrm{p}=.01\), respectively), whereas there was no difference in production of discourse-based (T-test: \(\mathrm{t}(11)=.82\), tied \(\mathrm{p}=.22\) ) and elaborative ( T -test: \(\mathrm{t}(11)=0, \mathrm{p}=1\) ) inferences.

Table 1: Mean types of recollection (and standard deviation in parenthesis) by the participants in Exp. 1.
\begin{tabular}{|c|c|c|c|c|}
\hline \begin{tabular}{c} 
Signing \\
deaf \\
\((\mathrm{N}=12)\)
\end{tabular} & \begin{tabular}{c} 
Correct \\
recollections
\end{tabular} & \begin{tabular}{c} 
Discourse- \\
based \\
inferences
\end{tabular} & \begin{tabular}{c} 
Elaborative \\
inferences
\end{tabular} & Errors \\
\hline Written & 21.42 & .58 & .08 & .75 \\
LIS & \((6.00)\) & \((.67)\) & \((.29)\) & \((.75)\) \\
\hline Written & 16.75 & .33 & .08 & 2.25 \\
Italian & \((3.67)\) & \((.49)\) & \((.29)\) & \((1.87)\) \\
\hline
\end{tabular}

The results of Experiment 1 confirmed our predictions. Signing deaf individuals benefited from the written LIS version of the technical texts. However, maybe because of the considerable technical content of the two texts, we did not observe a benefit from the LIS versions in terms of discourse-based inferences, which denote a deep comprehension of the text (Cutica \& Bucciarelli, 2008). A related, more general question is whether the observed facilitatory effect of the written LIS versions depends on the discourse content, be it technical or narrative in nature. The aim of Experiment 2 was to replicate the findings of Experiment 1 with narrative texts.

\section*{Experiment 2: Does Written LIS Facilitate Comprehension Independently on the Text Content?}

Experiment 2 set out to replicate Experiment 1 with narrative texts.

\section*{Method}

Participants Twelve signing deaf individuals (4 females and 8 males; mean age: 26 years), university students with a prelingual profound hearing deficit ( \(>90 \mathrm{~dB}\) hearing loss), took part in the experiment voluntarily. They had learned LIS as their first language since their first year of life. None of them had other disabilities, nor had they taken part in Experiment 1.

Materials and Procedures The experimental materials comprised two texts, one about the Savannah and one about Mammals (each text contained 346 words) (for excerpts see Appendix 2). For each text, we created a written LIS version
(183 and 204 words for the Savannah and the Mammals texts, respectively), following the same procedures used in Experiment 1.

As for Experiment 1, the translations of the two Italian texts to written LIS were evaluated individually by a Italian LIS interpreter and by a LIS deaf teacher, the same as in Experiment 1. On average, the \(100 \%\) of the semantic units from the Mammals text, and the \(99 \%\) of the semantic units from the Savannah text were judged as at least "adequate on the average"; none of the translations were judged as not adequate at all.

Each participant dealt with both the Savannah and the Mammals text, one in Italian and the other in LIS. Half of the participants dealt first with the Savannah text and the other half with the Mammals text, so that, overall, the occurrence of each text in the Italian and the LIS version was counterbalanced. The participants were invited to read each text carefully, one at a time, with no time limits. As soon as they finished reading each text, they were invited to recall in LIS as much information as they could, during which time they were video-recorded.
To code the results, the two versions of both texts were divided into 38 semantic units, corresponding to as many main concepts as the hearer could recall. As for Experiment 1, for each text (Savannah and Mammals), there is a strict correspondence between the semantic units in the two versions (Italian and Written LIS). Two independent judges coded the participants' recollections individually; the judges reached a significant level of agreement on their first judgments for the overall group of participants with the two versions of the texts, calculated using Cohen's K (Cohen's K ranging from .82 to .89 , p always <.01). For the final score, the judges discussed each item on which they disagreed, until they reached full agreement. Each concept (i.e., semantic unit) recalled by the participants was evaluated according to the same coding schema used in Experiment 1. Consider, for example, the following semantic unit in the written LIS Savannah text: "Food gives animal which? Giraffe"; according to the coding schema, the statement: "Acacia leaves to serve as food giraffe" has been coded as correct recollection, the statement: "Plant acacia serves for improving existence, growth giraffe" as elaborative inference, and the statement: "Animals do not eat acacia" as an erroneous recollection. As a further example, the recollection "Mother pecks egg, exit with help mother" has been coded as a discourse-based inference with respect to the semantic unit in the Italian text "(The shell is to tough) that the mother ostrich sometimes needs to help the chicks to break out".

Results and Discussion The two texts were comparable in difficulty; considering each type of recollection separately, we found no differences in performance with the two texts in either the LIS or the Italian versions (unpaired T-test: \(\mathrm{t}(10)\) ranging from 0 to \(1, \mathrm{p}\) ranging from 1 to .34 ). Hence,
we pooled together the results of the two Italian versions and those of the two LIS versions. Table 2 illustrates the mean types of recollection for each coding category for both versions of the texts.

Table 2: Mean types of recollection (and standard deviation in parenthesis) by the participants in Exp. 2.
\begin{tabular}{|c|c|c|c|c|}
\hline \begin{tabular}{c} 
Signing \\
deaf \\
\((\mathrm{N}=12)\)
\end{tabular} & \begin{tabular}{c} 
Correct \\
recollections
\end{tabular} & \begin{tabular}{c} 
Discourse- \\
based \\
inferences
\end{tabular} & \begin{tabular}{c} 
Elaborative \\
inferences
\end{tabular} & Errors \\
\hline Written & 11.17 & .58 & .25 & 1.00 \\
LIS & \((5.24)\) & \((.67)\) & \((.45)\) & \((.74)\) \\
\hline Written & 8.50 & .33 & .08 & 1.50 \\
Italian & \((4.36)\) & \((.49)\) & \((.29)\) & \((1.38)\) \\
\hline
\end{tabular}

The results show that signing deaf individuals produced significantly more correct recollections in the LIS than in the Italian version (T-test: tied \(\mathrm{t}(11)=2.13, \mathrm{p}<.03\) ), yet they produced comparable numbers of discourse-based inferences in the two versions (T-test: tied \(\mathrm{t}(11)=1, \mathrm{p}=.17\) ). Furthermore, the differences in production of elaborative inferences and errors in the two versions were not statistically significant (T-test: \(\mathrm{t}(11)=1\) and \(1.15, \mathrm{p}=.34\) to .28, respectively).
The results of Experiment 2 replicated those of Experiment 1: written LIS facilitated deep comprehension and learning from text, in terms of an increase in correct recollections. The conclusion holds independently of the nature of the text, be it technical or narrative. A possible reason why we failed to detect a beneficial effect for written LIS in terms of discourse-based inferences is that deaf individuals have difficulty in drawing inferences (see also Easterbrooks \& Scheetz, 2004).

\section*{General Discussion}

When reading and processing written texts of vocal languages, deaf individuals are more likely to treat written information as unrelated pieces rather than seeking commonality. Crucial features of all sign languages are the spatial arrangement of the signs, the highly characteristic, marked facial expressions or postures, and the gaze direction. Unlike the sequential ordering of the sentence elements in verbal languages, the rich morphosyntactic structure of visual-gestural languages is organized in spatial terms. We assumed that signing deaf individuals' comprehension and learning benefit from a written text that reflects the structure of their sign language, because such written texts might be comprehended using categories that belong to the sign language organization rather than the natural language organization. The results of our experiments on 24 profoundly deaf individuals confirmed the predictions derived from our assumptions. In particular, in both experiments we observed the beneficial effects of the written LIS compared to the written Italian in terms of
an increase in correct recollections by the signing deaf participants.
These results are in line with our assumptions: the written form of sign language offers deaf individuals the possibility to process written linguistic information provided in a syntactic structure that reflects the structure of the corresponding sign language; signing deaf individuals, who have a shorter memory span for linguistic material than do hearing individuals, benefit from the lack of articles, prepositions, conjunctions, pronouns, and verbal auxiliaries in the written sign language; signing deaf individuals, who compared to hearing individuals have a more spatially distributed visual attention, can extract in a glance more information from the written sign language text which is shorter than the written oral language text.
Our finding has strong implications; as deaf people have difficulties in comprehending the written versions of oral languages, their opportunities to learn from written texts and therefore to benefit from school and university education - are heavily restricted. Providing them access to written texts reflecting their sign language would be a step towards an improvement of their ability to comprehend the written versions of oral languages. Consistent, Oakhill and Cain (2000) already hypothesized that for deaf individuals who are fluent in signing, "it would be possible to present written texts via sign language in order to teach skills such as inference making, comprehension monitoring, and the planning and structuring of stories" (ib., p.58). On the basis of the results of our study, we argue that a written version of LIS could be used as an educational tool, in order to approach signing deaf children onto written verbal languages and improve their comprehension skills. Further studies would be useful to investigate in depth the effectiveness of trainings on texts comprehension exploiting written sign language both on adult signing deaf individuals, as well as on signing deaf children.

\section*{Acknowledgments}

The authors were supported in this research by Regione Piemonte of Italy, ATLAS project ID 44.

\section*{References}

Bavelier, D., Dye, M.W.G., \& Hauser, P. (2006) Do deaf individuals see better? Trends in Cognitive Sciences, 10, 512-518.
Cattani, A., Clibbens, J., \& Perfect, T.J. (2007). Visual memory for shapes in deaf signers and nonsigners and in hearing signers and nonsigners. Neuropsychology, 21,114121.

Cutica, I., \& Bucciarelli, M. (2008). The deep versus the shallow. Cognitive Science, 32, 921-935.
Cutica, I., \& Bucciarelli, M. (2011). "The more you gesture, the less I gesture": Co-speech gestures as a measure of
mental model quality. Journal of Nonverbal Behavior, 35, 173-187.
Easterbrooks, S.R., \& Scheetz, N.A. (2004). Applying critical thinking skills to character education and values clarification with students who are deaf or hard of hearing. American Annals of the Deaf, 149, 255-263.
Edwards, L., Figueras, B., Mellanby, J., \& Langdon, D. (2010). Verbal and spatial analogical reasoning in deaf and hearing children: the role of grammar and vocabulary. Journal of Deaf Studies and Deaf Education, 16, 189-197.
Logan, K., Mayberry M., \& Fletcher, J. (1996). The shortterm memory of profoundly deaf people for words, signs, and abstract spatial stimuli. Applied Cognitive Psychology, 10, 105-119.
Marschark, M., \& Hauser, P.C. (Eds.) (2008). Deaf cognition: Foundations and outcomes. New York : Oxford University Press.
Marschark, M., \& Mayer, T.S. (1998). Interactions of language and memory in deaf children and adults. Scandinavian Journal of Psychology, 39, 145-148.
Marschark, M., \& Wauters, L. (2008). Language comprehension and learning by deaf students. In M. Marschark \& P.C. Hauser (Eds.), Deaf cognition: Foundations and outcomes (pp. 309-350). New York: Oxford University Press.
Miller, P. (2002). Communication mode and the processing of printer words. Journal of Deaf Studies and Deaf Education, 7, 312-329.
Mitchell, T.V., \& Maslin, M.T. (2007). How vision matters for individuals with hearing loss. International Journal of Audiology, 46, 500-511.
Morgan, G., Barrett-Jones, S., \& Stoneham, H. (2007). The first signs of language: Phonological development in British Sign Language. Applied Psycholinguistics, 28, 322.

Morgan, G., Herman, R., Barriere, I., \& Woll, B. (2008). The onset and mastery of spatial language in children acquiring British Sign Language. Cognitive Development, 23, 1-9.
Oakhill, J.V., \& Cain, K.E. (2000). Children's difficulties in text comprehension. Journal of Deaf Studies and Deaf Education, 5, 51-59.
Pizzuto, E., Rossini, P., Russo, T. (2006). Representing signed languages in written form: questions that need to be posed. In C. Vettori (Ed.), Proceedings of the 2nd Workshop on the Representation and Processing of Sign Languages: "Lexicographic matters and didactic scenarios" (LREC 2006-Fifth International Conference on Language Resources and Evaluation) (pp. 1-6). Paris: ELRA.
Pyers, J., Shusterman, A., Senghas, A., Spelke, E., \& Emmorey, K. (2010). Evidence from an emerging sign language reveals that language supports spatial cognition. PNAS, 107, 12116-12120.

Smith, L.B., Quittner, A.L., Osberger, M.J., \& Miyamoto, R. (1998). Audition and visual attention. Developmental Psychology, 34, 840-850.
Sutton, V. (1999). Lessons in SignWriting. Textbook \& workbook. La Jolla, CA: Deaf Action Committee for Sign Writing.
Van Hoogmoed, A.H., Verhoeven, L. Schreuder, R., \& Knoors, H. (2011). Morphological sensitivity in deaf readers of Dutch. Applied Psycholinguistics, 32, 619-634.
Vendrame, M., Cutica, I., \& Bucciarelli, M. (2010). Learning by Videos: Peculiarities of Being Deaf Individuals. International Association for the Scientific Knowledge (IASK), Proceedings of the International Conference Teaching and Learning 2010, Seville, Spain, 29 November - 1 December 2010, p. 107-116.
Wilbur, R.B. (2000). The use of ASL to support the development of English and literacy. Journal of Deaf Studies and Deaf Education, 5, 81-104.
Winston, E.A. (1995). Spatial mapping in comparative discourse frames. In K. Emmorey \& J. Reilly (Eds.), Language, gesture and space (pp.87-114), Mahwah, NJ: Lawrence Erlbaum Associates.

\section*{Appendix 1. Material from Experiment 1 \\ (Semantic Units are Separated by Slashes)}

\section*{Excerpts from Responses to colour (translated to English)}

Written Italian version The idea that the various colours can arouse emotions/ is well known./ Red is considered exciting,/ because in our minds it evokes fire,/ blood/ and revolution./ Green brings relaxing thoughts of nature;/ blue is refreshing, like water./
Written LIS version People many think what?/ Colours various to give emotion./ Colour red to give excitation,/ reason what? We remember fire,/ blood,/ revolution./ Colour green to give relaxation reason? To view nature,/ colour blue to give feeling fresh like water./

\section*{Appendix 2. Material from Experiment 2 (Semantic Units are Separated by Slashes)}

\section*{Excerpts from The savannah (translated to English)}

Written Italian version Thirty million years ago,/ tropical Africa was covered in jungle. Things have since changed./ In eastern Africa, the forest has disappeared,/ and the new landscape is very different from its predecessor./ Everything began in the forest./ Chimpanzees are perfectly adapted to life in the trees./
Written LIS version Million thirty years ago where?/ Africa tropical forest covered./ Now Africa eastern forest there is not,/ landscape new, instead past different./ Forest now begins/ tree, area monkeys suitable live they where? Trees./

\title{
Combinatorial structure and iconicity in artificial whistled languages
}

\author{
Tessa Verhoef (t.verhoef @uva.nl) \\ ACLC, University of Amsterdam \\ 1012 VT, Amsterdam, Netherlands
}

\author{
Simon Kirby (simon@ling.ed.ac.uk) \\ LEC, University of Edinburgh \\ Edinburgh EH8 9AD, UK
}

\author{
Bart de Boer (bart@ai.vub.ac.be) \\ AI lab, Vrije Universiteit Brussel \\ 1050 Brussels, Belgium
}

\begin{abstract}
This article reports on an experiment in which artificial languages with whistle words for novel objects are culturally transmitted in the laboratory. The aim of this study is to investigate the origins and evolution of combinatorial structure in speech. Participants learned the whistled language and reproduced the sounds with the use of a slide whistle. Their reproductions were used as input for the next participant. Cultural transmission caused the whistled systems to become more learnable and more structured. In addition, two conditions were studied: one in which the use of iconic form-meaning mappings was possible and one in which the use of iconic mappings was experimentally made impossible, so that we could investigate the influence of iconicity on the emergence of structure.
\end{abstract}

Keywords: iterated learning; iconicity; combinatorial structure; phonology; cultural evolution; duality of patterning

\section*{Introduction}

Duality of patterning, one of Hockett's (1960) basic design features of language, has recently received increased attention (de Boer, Sandler, \& Kirby, 2012). This feature describes how, in speech, a limited number of meaningless sounds are combined into meaningful words and those meaningful words are combined into larger constructs. How this feature emerged in language is currently still a matter of debate, but it is increasingly being studied with the use of a variety of different techniques such as computer simulations and laboratory experiments. In one of these laboratory experiments, the emergence of combinatorial structure was studied through transmission of artificial whistled languages in the laboratory (Verhoef, Kirby, \& Padden, 2011). For this study the experimental iterated learning paradigm (Kirby, Cornish, \& Smith, 2008) was used. Iterated learning refers to the process of cultural transmission, in which individuals acquire a social behavior by observing the performance of others who also acquired it from observation (Kirby et al., 2008). The results demonstrated that sound systems, when passed on in a transmission chain, adapt to cognitive biases and become easier to learn (Verhoef et al., 2011). Combinatorial structure emerged and the whistled systems became more efficiently coded over generations (Verhoef, 2012). These results demonstrated a possible route towards the emergence of structure in the sounds of speech. The findings challenge
the hypothesis that Hockett (1960) introduced when he linked the emergence of structure to vocabulary expansion and signal dispersal. Even in the case where only a small set of sounds is transmitted and the signal space is not maximally used, combinatorial structure emerges (Verhoef et al., 2011; Verhoef, 2012). In that experiment the influence of semantics was controlled for and the signals did not refer to any concrete meanings. Obviously, in natural human languages meanings are important and the role of semantics in the evolution of linguistic structure should not be ignored (Schouwstra, 2012; Dingemanse, 2012). Would the introduction of semantics influence the emergence of combinatorial structure at the level of phonology? In this article a new experiment is presented in which artificial whistled languages are culturally transmitted and the whistled signals refer to meanings.

\section*{Combinatorial structure versus iconicity}

Arbitrariness was another design feature Hockett (1960) listed for language. This feature refers to the arbitrary mapping between words and their meaning. Hockett uses the words 'whale' and 'microorganism' as an example: 'whale' is a short word for a large animal, while 'microorganism' is the reverse. It has been argued that non-arbitrariness is rare in modern languages and that it is irrelevant for understanding linguistic structure (Newmeyer, 1992). More recently, however, researchers began to realize that non-arbitrary formmeaning mappings may be more widespread than initially thought. When exploring beyond Indo-European languages, non-arbitrariness seems to play a role in many languages (Imai, Kita, Nagumo, \& Okada, 2008; Dingemanse, 2012). This involves classes of words where for instance the shape, complexity, sound or some other characteristic of the meaning expressed is mimicked or iconically represented in the word. Examples have been identified as 'ideophones', 'mimetics' or 'expressives' and the phenomenon is often called sound symbolism.

Sound symbolic mappings in language have been connected to cross-modal mappings in the human brain (Simner, Cuskley, \& Kirby, 2010; Ramachandran \& Hubbard, 2001). There appear to be many cognitive biases in cross-modal perception that are shared by most people. The bouba/kiki ef-
fect is one famous example that shows a strong preference to relate sharp shapes to the name 'kiki' and round shapes to the name 'bouba' (Ramachandran \& Hubbard, 2001). Such shared biases have been argued to play an important role in the evolution of language, by forming a starting point for the initial emergence of grounded speech (Ramachandran \& Hubbard, 2001).

A newly emerging sign language, Al-Sayyid Bedouin Sign Language (ABSL), has recently been described, in which the evolution of duality of patterning can still be observed (Sandler, Aronoff, Meir, \& Padden, 2011). Even though ABSL is a fully functional and expressive sign language, its combinatorial structure appears to be less discrete (Sandler et al., 2011). Could it be the case that this young sign language was able to survive up to now without duality of patterning because the manual modality allows for much iconicity and the language is learnable and transmissible enough without phonological structure? When most mappings are transparent, they may be more intuitive and easier to remember as holistic entities. On the other hand, it has been shown that there is actually an advantage for arbitrary mappings in acquiring word meanings in context (Monaghan, Christiansen, \& Fitneva, 2011). A secondary objective of the experiment described below is to investigate how iconic formmeaning mappings influence the emergence of combinatorial sub-lexical structure. The experiment is similar to the experiment described by Verhoef et al. (2011) and Verhoef (2012), but with meanings attached to the whistled signals. Two conditions were studied: one in which the use of iconic form-meaning mappings is possible and one in which the use of iconic mappings is experimentally made impossible. This was expected to provide insight into the possible role of iconicity in the emergence of duality of patterning since it could reveal whether a situation that allows for more iconicity, can 'survive' without the emergence of combinatorial structure longer.

In summary, the objective of this study is as follows. First and foremost, we investigate whether the addition of meanings leads to a result that is similar to what was found in the whistle experiment without meanings, to see if combinatorial structure also emerges in the presence of semantics. Second, we search for any differences between the two conditions to see whether iconicity could cause a delay in the emergence of structure.

\section*{Method}

In this experiment participants were asked to learn and reproduce whistled signals with a slide whistle (see figure 1) as names for objects they saw on a computer screen. There were twelve whistled signals in the training set in total.

The meanings were unusual objects that look like possible mechanical parts, but they are novel objects for which we do not have words in existing languages. The objects were a subset from those created by Smith, Smith, and Blythe (2011) that were slightly modified. To reduce the structure in


Figure 1: Slide whistle
the meaning space, all objects were transformed to blue tone and could therefore not be grouped by their color. They also did not share shapes or parts and were not structured in any other obvious way. This was needed to limit the emergence of semantics-related compositional structure. A few examples of objects that were used are shown in figure 2 .


Figure 2: Examples of novel objects used in the experiment
Following the paradigm of experimental iterated learning (Kirby et al., 2008), the last whistle sounds that a participant produced for each object were used as the input given to the next participant. However, this is the point where the two conditions differ from each other. In one condition, the 'intact' transmission, the next participant was exposed to the output of the previous participant exactly as it was produced. The mapping from whistled signals to objects was kept intact. In the other condition, the 'scrambled' transmission, the output of the previous participant was altered before it was given to the next person. The produced form-meaning mappings were broken down by scrambling the mappings at each change of generation and by using a different set of objects between consecutive generations. In this way, if there were any iconic relations to emerge in the sets, it would only be helpful for the participants in the first condition and any semantics-related structure is broken down in between the transmission steps. Only the signal sets themselves stay intact.

\section*{Procedure}

During the experiment participants completed three rounds of learning and recall, which were alternated by 'guessing game' rounds. In the learning phase the objects and their corresponding whistle were presented one by one in a random order, and participants recorded an imitation of the whistle. In the recall phase a panel was shown with a button for each object and the participant had to choose each of the objects once to record the right whistle for it from memory. The guessing phases were introduced to encourage people to keep paying attention to the mapping between whistle sounds to objects. In this phase the whistles were played one by one in a random order and for each whistle the participant had to choose
the right object from a panel. This was done with half of the whistle-object pairs after the first learning phase and with the other half after the second. The whistles from the last recall phase were used as training input for the next participant, depending on the condition either with intact or scrambled whistle-object mappings. Transmission was continued from person to person until there were eight generations in each chain and four chains per condition.

\section*{Initial input sets}

Two separate initial whistle sets were constructed. Each set was used as the starting point for half of the chains in each condition. The whistles were taken from a database of whistles that were collected during a pilot study. During this pilot, people were asked to freely record a number of whistles. The two initial sets were constructed so as not to exhibit combinatorial structure. To achieve this, the entropy measure for quantifying combinatorial structure from Verhoef (2012) was used. Sets of twelve whistles were generated randomly from the database until two sets were found with no overlap, that had a comparable and relatively high measured entropy.

\section*{Participants}

In total 64 participants took part in the experiment. They were divided over eight transmission chains, four in each condition. Participants were recruited from the University of Amsterdam community through posters and e-mail invitations. All participants were between the ages of 19 and 41 years old, 43 were female and 21 male. In each chain either two or three men participated. They were compensated for their time with a cash payment of 10 euros.

\section*{Qualitative results}

This section describes qualitative observations as a first impression of the data. The internal structure of the whistle sets is investigated as well as the role of iconicity.

\section*{Internal structure in whistle sets}

On the level of the signals, independent of the objects they refer to, it can be observed that structure develops in a manner that is very similar to what could be observed in the experiment without meanings. Like in that experiment, whistles were introduced that were clearly related in some way to the form of some whistles that already existed in the set. Mirrored versions, combinations of existing whistles, repetitions of the same pattern within a whistle or whistles with similar shapes but different whistle manners for instance appeared. Figure 3 shows an example from one of the chains in the scrambled condition. Here, at generation four, two whistles were in the set that followed approximately the same shape in pitch contour (down and up), but were whistled in a different manner. One of them was whistled in a smooth and unbroken fashion and the other was more staccato-like and broken into pieces. In generation five, one half of each of these whistles is borrowed and reused to form a new whistle. The left part


Figure 3: Development of structure in a chain from the scrambled condition. See main text for an explanation.
of the smooth whistle is also reused and combined with existing whistles. In later generations, these are reproduced and all kinds of other variations on this appear, such as ones that are mirrored again as a whole.


Figure 4: Development of structure in a chain from the intact condition. See main text for an explanation.

Figure 4 shows an example from one of the chains in the intact condition. In this example one whistle from generation three seems to be the inspiration for two new whistles in the next generation: one with one 'bump' and another with two. In generation five the 'two bump' whistle is started to be reused and combined with another pattern and in generation six both the one bump and two bump whistles are being reused, mirrored and recombined more widely. An existing whistle with several up and down movements is even segmented into two parts, where the first part is again the two bump whistle.

To examine the final result of these gradual changes in the chains, we can look at the set of whistles produced by the eighth and last participant in a chain. Figure 5 shows a frag-


\section*{Iconic whistle-object mappings}

When talking about mappings between whistle sounds and objects one may wonder how a whistle sound can iconically depict a visual object. This is difficult to identify as an outside observer, since iconicity is subjective and depends on experience and individual history. However, some examples could be found in the form-meaning pairs in the current data and iconicity could take several different forms in these examples. Most often, the shape of the whistle, or the pitch contour, would mimic certain features in the object. This could for instance be the overall shape of the object, the orientation of the object or the amount of visually distinctive parts on the object. It needs to be noted though that these are subjective observations and that it is not necessarily the case that the participants would be aware of these structural similarities. Judging from the observations, iconic mappings were not found to be widespread throughout the whole experiment. Figure 7 shows a few examples of clearly iconic form-meaning mappings that were encountered.


Figure 7: Examples of iconic whistle-object pairs in the data.
Participants filled out a post-participation questionnaire in which they were asked to describe their specific strategy (if any) for recall and whether they thought the whistles and objects fit well together. Often participants reported strategies in line with the observations described in the previous paragraph. Other strategies that were reported involved: imagining how the object would sound and linking this with the whistle, imagining how the object would move and linking the pitch contour with that, or linking the object with some real object they know and linking the whistle with the sound that object would make. These reports further illustrate the subjectivity of form-meaning resemblance.

In summary, the structures that emerged in the sets of whistled signals resemble the discrete and combinatorial structure that emerged in the experiment without meanings (Verhoef et al., 2011; Verhoef, 2012). Qualitatively, no difference could be observed between the structures in the two conditions and iconicity did not seem to play a large role.

\section*{Quantitative results}

This section describes a quantitative analysis to assess whether the observed patterns are consistent across the data. First, the learnability is investigated by computing how well participants were able to recall the set of whistle-object pairs.

Then, the development of combinatorial structure is measured.

\section*{Recall error}

To find out whether the sets of whistle-object pairs became easier to learn and reproduce, we measured how well participants were able to recall the right whistle for each of the objects. The recall error was measured by comparing each whistle that a participant produced for an object with the whistle linked to that specific object in the input. To determine the distance between two whistles, the same distance measure was used as the one used by Verhoef (2012). This measure compares plunger movement tracks (pitch tracks converted to plunger displacement) with the use of derivative Dynamic Time Warping (Keogh \& Pazzani, 2001).

Figure 8 shows the data for this measure of recall error for the four chains in both conditions, with increasing generations on the horizontal axis. The mean over the four chains for each condition and standard error are plotted.

A linear mixed effects analysis was performed (with lme4 in R) to explore the effect of generation on the recall error, with an intercept for chain as random effect. Likelihood ratio tests of this model against a null model excluding the effect of generation revealed a significant trend affecting error \(\left(\chi^{2}(7)=14.08, p=0.0498\right)\), decreasing it by about \(0.18 \pm 0.056\) (standard errors) from generation to generation for the intact condition, as well as for the scrambled condition \(\left(\chi^{2}(7)=27.25, p=0.0003\right)\) with a decrease of about \(0.1 \pm 0.025\) (standard errors). This suggests that there is an increase in the reproducibility of the form-meaning pairs over generations in both conditions.


Figure 8: Recall error on the whistle-object pairs over generations in both conditions, showing the mean and standard error. Recall error decreases significantly in both conditions.

\section*{Combinatorial structure}

To investigate whether the sets of whistles gradually become more structured over generations, the entropy measure that was used by Verhoef (2012) was applied to the current data. This measure makes use of the notion of entropy from information theory and is based on the idea that a set with more combinatorial structure is composed of less basic building
blocks that are more widely reused and combined. One adjustment had to be made to the measure as it was described in Verhoef (2012). Based on the qualitative observation that there was clearly no one 'right' segmentation that could be used to describe the discretization of the signal space, three different types of segmentation were defined. The whistles were segmented in all three ways and the entropy was computed for each of the three sets of basic building blocks that resulted from the segmentations. The lowest entropy value that was measured was then considered to be the best minimal description length approximation and was used as the measure for (dis)order. The first type of segmentation used silences as segment boundaries. The second type used the minima and maxima in the plunger movement track as segment boundaries and the third used the points of minimal and maximal velocity.

Figure 9 shows the development of entropy for the four chains in both conditions, where 0 refers to the initial whistle set. Again, the mean over the four chains and standard error for each condition is plotted.

A linear mixed effects model was constructed in the same way as for recall error, but with entropy as the test variable. A significant trend \(\left(\chi^{2}(7)=19.73, p=0.006\right)\) of decreasing entropy by about \(0.69 \pm 0.20\) (standard errors) from generation to generation for the intact condition was found, as well as for the scrambled condition \(\left(\chi^{2}(7)=17.22, p=0.016\right)\) with a decrease of about \(0.52 \pm 0.19\) (standard errors). These findings imply that the process of iterated learning in both conditions caused structure to emerge in the sets of whistles that refer to meanings.


Figure 9: Entropy of the whistle sets over generations in both conditions, showing the mean and standard error. Entropy decreases significantly in both conditions.

\section*{Transparency}

Although it is difficult to assess the actual role iconicity played in the two conditions, the results of the guessing game phases could indirectly reveal a potential influence. If the mappings were more transparent in the intact condition, we would expect participants in that condition to score higher on the identification task after only very little exposure to the
data. A linear mixed effects analysis was performed to explore the effect of condition on the scores, with round number (half of the items appeared in a guessing game round after the first exposure to the data and the other half after the second exposure) as fixed effect and intercepts for chain and generation as random effects. Likelihood ratio tests of this model against a null model excluding the effect of condition showed that condition does not affect performance in the guessing game ( \(\chi^{2}(1)=0.210, p=0.647\) ). This could suggest the role of iconicity was minimal in both conditions, or at least did not play a large enough role in the intact condition to boost identification scores. However it needs mentioning that the participants had been exposed to the data before the guessing game phases, which is expected to have influenced the scores.

\section*{Discussion}

The experiment presented in this article shows that cultural evolution in the laboratory causes a system of whistled words for novel objects to become more learnable and more structured over time. This work expands a previous finding that showed the same result for whistled systems without meanings (Verhoef et al., 2011; Verhoef, 2012). For two different situations, one with transmission of intact form-meaning pairs and one with scrambled pairs, we showed that the transmitted whistled systems develop from holistic towards having discrete and combinatorial structure. Sets of building blocks are efficiently reused and combined, similar to the structures of speech. In addition to the data presented by Verhoef et al. (2011), the current data forms another example to show that the emergence of combinatorial structure is not necessarily driven by vocabulary expansion and dispersal as was proposed by Hockett (1960).

As a secondary objective we explored the tension between combinatorial structure and possible iconic mappings. It appeared that the potential for iconic mappings did not prevent the emergence of structure in this experiment. However, when looking at the development of entropy in the two conditions, we can see that the main 'drop' in entropy in the intact condition took place approximately from generation four to eight, while in the scrambled condition, this was sooner, approximately from generation one to five. This could hint at a slight delay in the emergence of combinatorial structure in the intact condition, which would follow the expectation, but the current data from the guessing game phase does not suggest a large influence of iconicity. A more detailed analysis is needed. When we have a better picture of the actual development of iconicity, the development of structure can be linked to it more directly. Finding an objective measure for quantifying iconicity in the data is not trivial however and this will be part of our future continuation of this research.

To conclude, this article provides additional evidence to show that combinatorial structure in language can emerge through cultural evolution. The influence of iconicity in this evolutionary process still needs to be investigated in more depth, but with this study we provide an experimental plat-
form that can be used to tackle this issue. In the future, additional experiments are expected to lead to more insights.

\section*{Acknowledgments}

We thank Alex del Giudice, Kenny Smith and Carol Padden for helpful suggestions. This research was funded by NWO vidi project 016.074.324 and ERC project ABACUS.

\section*{References}
de Boer, B., Sandler, W., \& Kirby, S. (2012). New perspectives on duality of patterning: Introduction to the special issue. Language and Cognition, 4(4), 251-259.
Dingemanse, M. (2012). Advances in the cross-linguistic study of ideophones. Language and Linguistics Compass, 6(10), 654-672.
Hockett, C. (1960). The origin of speech. Scientific American, 203, 88-96.
Imai, M., Kita, S., Nagumo, M., \& Okada, H. (2008). Sound symbolism facilitates early verb learning. Cognition, 109(1), 54-65.
Keogh, E., \& Pazzani, M. (2001). Derivative dynamic time warping. In the 1st siam international conference on data mining (sdm-2001), chicago, il, usa.
Kirby, S., Cornish, H., \& Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. PNAS, 105(31), 10681-10686.
Monaghan, P., Christiansen, M. H., \& Fitneva, S. A. (2011). The arbitrariness of the sign: Learning advantages from the structure of the vocabulary. Journal of Experimental Psychology-General, 140(3), 325.
Newmeyer, F. J. (1992). Iconicity and generative grammar. Language, 68(4), 756-796.
Ramachandran, V., \& Hubbard, E. (2001). Synaesthesia-a window into perception, thought and language. Journal of Consciousness Studies, 8(12), 3-34.
Sandler, W., Aronoff, M., Meir, I., \& Padden, C. (2011). The gradual emergence of phonological form in a new language. NLLT, 29(2), 503-543.
Schouwstra, M. (2012). Semantic structures, communicative strategies and the emergence of language. Lot Dissertation Series, 312.
Simner, J., Cuskley, C., \& Kirby, S. (2010). What sound does that taste? cross-modal mappings across gustation and audition. Perception, 39(4), 553.
Smith, K., Smith, A. D. M., \& Blythe, R. A. (2011). Cross-situational learning: an experimental study of wordlearning mechanisms. Cognitive Science, 35, 480-498.
Verhoef, T. (2012). The origins of duality of patterning in artificial whistled languages. Language and Cognition, 4(4), 357-380.
Verhoef, T., Kirby, S., \& Padden, C. (2011). Cultural emergence of combinatorial structure in an artificial whistled language. In Proc. CogSci 33 (pp. 483-488). Cognitive Science Society.

\title{
Inductive Use of Semantic Word-Families to Accelerate Vocabulary Development and Reading Comprehension in Grades 3-4-5
}

\author{
Michael R. Vitale (vitalem@ecu.edu) \\ College of Education, \\ Greenville, NC 27858 USA \\ Nancy R. Romance (romance@fau.edu) \\ College of Education, Boca Raton, FL 33431 USA
}

\begin{abstract}
This study explored the acceleration of student vocabulary growth and reading comprehension proficiency through a multi-part instructional strategy for engendering the inductive, semantic word-family-oriented acquisition of vocabulary from context, a difficult task for elementary students. Implemented on a schoolwide basis for an academic year in grades 3-4-5, the intervention was a fourpart enhancement to a traditional basal reading program that constructed and used semantic word families for designated vocabulary words within stories. Results from HLM statistical modeling using student minority status and free/reduced lunch as covariates showed that experimental students in grades 3-4-5 obtained significantly higher achievement on both ITBS Vocabulary and ITBS Reading Comprehension subtests. Implications for research and practice are discussed.
\end{abstract}

Keywords: Inductive reasoning, Vocabulary acquisition, Reading comprehension

A variety of research has pointed to the interdependent linkage among vocabulary knowledge, reading comprehension, and level of literacy (e.g., Baker et al., 1998; Becker, 1977; Biemiller \& Slonim, 2001; Snow, 2002; Wager, 2005). Although substantial vocabulary growth can be attributed to student acquisition of word meaning from reading context (Baker, Simmons, \& Kameenui, 1998), August, Dressler, and Snow (2005) noted that if the proportion of unknown words is too large, then text comprehension which serves as a context for vocabulary development is disrupted (see also Carver, 1994).

Despite technical details in how words (e.g., counting word roots vs. root variants) and word understanding (e.g., recognition vs. in depth understanding) are defined in the literature (e.g., Anglin, 1993; Beck \& McKeown, 1991), research findings agree that children acquire vocabulary at a rate that is too rapid for all the words to be taught directly (see Baker et al., 1998) or learned incidentally through reading (Landauer, 2002; Landauer, \& Dumais, 1996, 1997; Landauer, Foltz, \& Laham, 1998; Landauer, McNamara, Dennis, \& Kintsch, 2007). With this point in mind, the present study addressed the
question of whether student vocabulary acquisition could be accelerated by using a multi-part semantic word-family-oriented learning strategy to inductively broaden vocabulary taught directly. In incorporating criteria suggested by Baker et al. (1998) and Beck and McKeown (1991), the intent of the strategy was (a) to engender an inductive broadening of the vocabulary taught directly and, in doing so, to enhance reading comprehension, and (b) to be feasible for use by classroom teachers within regular classroom settings.

Implemented as a practitioner-oriented model, the instructional intervention reflected several interdisciplinary perspectives: (a) vocabulary research findings with both younger (e.g., Coyne, McCoach, \& Kapp, 2005) and older (e.g., August et al., 2005; Baker et al., 1998; Blachowicz \& Fisher, 2000; Johnson, Gersten, \& Carnine, 1988) students, (b) cognitive science models (e.g., Kintsch, 1994, 1998a, 1998b, 2002, 2004, 2005; Landauer, 2002; Landauer, \& Dumais, 1996, 1997; Landauer et al., 1998, 2007) that emphasize the central role of prior knowledge in comprehension and, (c) our prior research (Vitale \& Romance, 2007) investigating the effect of knowledge-focused reading comprehension strategies on student learning. In the present study, different aspects of these perspectives provided a framework for engendering the semantically-oriented inductive learning of vocabulary.

The design of the present study was a significant enhancement of earlier studies (Romance \& Vitale, 2012; Vitale \& Romance, 2008). First, in this study, the intervention was implemented over an school year in multiple schoolwide sites. Second, teachers were asked to commit to implement the model in eight selected stories in grade 3, 4, and 5. across the school year. And third, the criterion measures (ITBS Vocabulary and Reading Subtests) were administered on a pre-post basis. The specific research questions were:
- Did the instructional intervention which incorporated words taught inferentially accelerate student the vocabulary development as measured by story-specific,
curriculum-based, pre-post tests?
- Did the instructional intervention accelerate student vocabulary development as measured by the nationally-normed ITBS Vocabulary subtest?
- Did the instructional intervention engender a transfer effect to student reading comprehension as measured by the nationally-normed ITBS Reading subtest?

\section*{Perspectives in Vocabulary Instruction}

Although an increasing number of studies have identified factors important in teaching vocabulary in classroom settings (see Baker et al., 1998; Biemiller \& Boote, 2006; Coyne et al., 2005; Nagy \& Scott, 2000), such studies have limitations insofar as providing a comprehensive means for accelerating student vocabulary acquisition. For example, Baker et al. (1998) pointed to the fact that the size and rate of growth of the vocabulary of school age children is far too large to be addressed on a literal word-by-word basis alone, while Anderson \& Nagy (1992) argued that because word meaning is learned primarily in the context of speech or text, direct instruction of vocabulary can address only a small portion of words to be learned.

The approach in the present study was a methodological enhancement of an earlier study (Vitale \& Romance, 2008) that demonstrated significant effects of the inductive model used in the present study on both curriculum-based vocabulary transfer tests and on ITBS Vocabulary. Approaching the question of vocabulary acquisition from a knowledge-based instruction approach (see Bransford, Brown, \& Cocking, 2000), a major tenet of this perspective is that prior knowledge is a major factor in meaningful comprehension, learning, and expert performance. Within this framework, building prior knowledge has been recognized as a major determinant of meaningful learning, general comprehension, and reading comprehension.

In applying a knowledge-based perspective to the classical problem of how persons can know more than experience could have taught (literally) within the context of vocabulary, Landauer (2002) and Landauer and Dumas (1997, 1998) drew on the idea that the underlying semantic dimensions as identified by Latent Semantic Analysis (see Landauer et al. 2007) that represent the relatedness among words, phrases, and prose provide the "learning leverage" through which words are understood. From this view, both the traditional and cognitive science research literatures are consistent in that while vocabulary words can be taught directly, the majority of vocabulary must be gained in a fashion that is inferential. Although some indirect vocabulary acquisition can be explained
through reading, the rate of vocabulary acquisition exhibited by children requires a process that involves the induction of the meaning of new words (since direct teaching and incidental learning from reading contexts are inadequate to explain vocabulary growth).

\section*{Method}

\section*{Participants}

The study was implemented over an 18 week period on a school-wide basis in grades 3-4-5 in a large (185,000 students), highly diverse (African American: 29\%, Hispanic: 19\%, Other: 5\%, Free Lunch: 40\%) school system in southeastern Florida. Using a random selection process with constraints for demographic similarity, three of six schools were assigned the intervention and the other three demographically similar schools served as controls.

\section*{Instrumentation}

Outcome measures consisted of the ITBS Vocabulary and ITBS Reading Comprehension Subtests. These tests were administered by classroom teachers with supervision from the researchers during 2-week periods prior to the beginning of the intervention and after the intervention ended.
The project intervention model also provided teachers with curriculum-based, pre-post lesson tests specific to each story taught. These tests consisted of two components: (a) a word recognition test in which students indicated whether or not they believed they know the meaning of a given word and (b) a sentence writing task in which students used a given word in a sentence. All of the words used in the tests were randomly sampled from the semantic word families associated with key words in specific stories and were not used in instruction. While there was a partial overlap of words used on the pre and post-tests, one-half of the words appearing on the posttest did not appear on the pre-test.

\section*{Experimental Intervention}

Pre-planning identified 4-word semantic word families for each of 3 key vocabulary words in each of 10 regular basal reading stories for use by teachers at each grade level. In Part 1 of the multi-part intervention in each story, teachers pre-taught 3 key vocabulary words in a textbookspecified fashion. Then, as a student reading the story reached a key word, teachers queried students regarding the word meaning in context (e.g., What does the word ___ mean in this sentence? How does this word contribute to the overall meaning of this sentence?).

In Part 2, the same procedure was followed but with pairs of new target words similar in meaning to each of
the pre-taught key words that were pre-taught and then substituted in 3 -sentence blocks from the story that contained the original key words. In Part 2, the teacher query was enhanced with an additional question: How does the use of this new word change the meaning of the sentence or story?

In Part 3, two new target words for each key word were not pre-taught. Rather, they were substituted in the same 3-sentence blocks in Part 2 and, again, students were queried regarding their meaning in context (an inductive process) by adding an additional question to the Part 2 query: How did the meaning of the three sentences from the story suggest what the meaning of the new word should be? Finally, in Part 4, as an expansion task, students presented sentences orally about their own experience using a key or target words.

For use by teachers, story-specific guidelines for each story were computer-generated in an easy-to-follow format in which the words, word definitions, relevant story sentences, and specific questions were inserted for each part of the intervention.

\section*{Design, Analysis, and Procedure}

The instructional intervention was implemented on a school-wide basis in grades 3-4-5 over the school year, with the ITBS Vocabulary and Reading subtests administered during a two-week period prior to the beginning and end of the 18 -week study. Both Experimental and control teachers used the same districtadopted basal reading series and followed the district curriculum plan in selecting stories for instruction. Experimental teachers were asked to commit to teaching 8 stories during the 18 -weeks of the school year in which the inductive vocabulary model would be applied. The study design followed the framework appropriate for a 2Level HLM analysis, with separate HLM analyses conducted for ITBS Vocabulary and ITBS Reading. For the HLM analyses, Level 1 student data consisted of student ITBS Vocabulary or Reading achievement outcomes, with minority (vs. non-minority) status, participation in free/reduced lunch (vs. non-eligible), grade, and the appropriate ITBS Reading or Vocabulary Subtests serving as a covariate. Level 2 classroom/teacher data (with students nested within teachers) consisted of a dummy variable representing treatment ( \(1=\) treatment, 0 \(=\) control) and grade .

\section*{Teacher Professional Development and Implementation Support}

Teacher professional development consisted of 2 days prior the start of the study with two days of "follow-up" during the initial 9 weeks of the intervention. In addition, researcher provided informal support as necessary.

\section*{Monitoring of Intervention Fidelity}

Researchers informally monitored all participating classrooms on a regular/continuing basis through direct observation and through inspection of teaching plans.

\section*{Results}

\section*{Implementation Fidelity}

The intervention involved 22 teachers across grades 3-4-5 and, because some teachers taught multiple sections, a total of 39 classrooms. The average number of stories taught using the vocabulary intervention were \(7.0,6.4\), and 6.2 for grades 3,4 , and 5 , respectively. Observation of classroom implementation by researchers averaging 3.4 visits per teacher found the intervention easy to implement by teachers and the project-developed storyspecific vocabulary guides to be effective. Mean ratings of fidelity of implementation ranged from 82 to 92 expressed on a 100 point scale, in which a rating of 80 percent or more indicated consistent model implementation. Based on the observations in conjunction with teacher planning effectiveness, the model was judged to be implemented with fidelity. Average inter-rater reliability (agreement) on the researcher-developed classroom fidelity observation form ranged from .88 to . 95.

\section*{Pre-Post Story-Based Test Findings}

Figure 1 shows the pre-post lesson achievement gains across the experimental classrooms in terms of mean percent of items correct across students and stories. As Figure 1 shows, students exhibited consistent pre-post achievement growth on the curriculum-based lesson tests.


Figure 1. Mean pre- and post-test scores for Word Meaning and Sentence Writing on the story lessons in grades 3, 4, and 5.

\section*{ITBS Achievement Findings}

One of the three control schools was eliminated from the analysis because of problems with the data resulting from the scanning of their Fall, 2011, prior ITBS response sheets. The results presented here are for the three
experimental and two control schools. Because preliminary HLM analyses found no interactions of treatment with minority status, or free/reduced lunch participation, these interaction components were removed from the final HLM models reported.

The HLM Model analyses with ITBS Vocabulary as the achievement outcome measure found a significant crosslevel interaction between Treatment and Grade, \(t(1348)=\) 1.99, \(p<.04\) ), along with each of the three covariates in the model (White-Asian-Mixed vs. Black-HispanicIndian, Free/Reduced Lunch vs. None, Prior-ITBS Achievement). The Treatment main effect was not significant. However, because of the significant interaction, the General Linear Hypothesis option in HLM was used to test the combined effect of Treatment and the Treatment x Grade interaction model components as a means of interpreting the overall effect of the intervention. The result of this follow-up analysis was significant, Chi-Square ( \(2 d f\) ) \(=11.43, \quad p<.0003\) and confirmed the overall impact of the intervention on student ITBS Vocabulary achievement.

A parallel HLM analysis with ITBS Reading as the achievement outcome measure found both the Treatment main effect, \(t(66)=-2.95, p<.01\), and the cross levellevel interaction between Treatment and Grade, \(t(1431)=\) 2.99, \(p<.003\), significant, along with two of the three


Figure 2. Differences in Estimated ITBS GE achievement between adjusted means of Experimental and Control students by grade. Differences greater than zero show higher achievement for Experimental students. For Reading, Control students outperformed Experimental students in grade 3; however the achievement difference in favor of Experimental students accelerated in grades 4 and 5.
covariates (Free/Reduced Lunch vs. None, Prior-ITBS Achievement). The covariate White-Asian-Mixed vs. Black-Hispanic-Indian was not significant. As in thepreceding analysis, because of the significant interaction, the General Linear Hypothesis option in HLM was used to test the combined effect of Treatment and the Treatment x Grade interaction model components as a means of interpreting the overall effect of the intervention. The result of this follow-up analysis was significant, Chi-Square (2df) \(=11.90, \quad p<.003\), confirming the impact of the Vocabulary Intervention on student ITBS Reading achievement.
In order to further interpret the combined Treatment main effect at Level 2 and the cross-level Treatment x Grade interaction, estimates were computed from the HLM models for ITBS Vocabulary and ITBS Reading of the differences between adjusted means for the Experimental and Control students by grade level for each ITBS achievement outcome. As shown in Figure 2, the intervention resulted in a magnified effect of the favor of Experimental students as grade level increased.

\section*{Discussion and Conclusions}

In conducting studies on vocabulary acquisition, earlier (Romance \& Vitale, 2012; Vitale \& Romance, 2008) investigations of the inductive vocabulary model along with research cited in the literature (e.g., Kintsch, 2012; Landauer, \& Dumais, 1996, 1997; Landauer et al., 1998, 2007) were suggestive that development and inductive use of general semantic (i.e., conceptual) meaning should be considered as an important focus of vocabulary learning rather than simply building understanding of specific words in a literal fashion. This perspective is supported by findings from earlier work (Vitale \& Romance, 2008) which explicitly demonstrated the impact of the model on the inferential performance of students on tasks based on the semantic word families used and by the fact that the impact of the model on the ITBS Vocabulary test served as an achievement transfer measure.
From an applied research perspective, the present study replicated and extended the preceding studies (Romance and Vitale, 2012; Vitale and Romance, 2008) in terms of instructional time (duration of intervention) and increased use of the intervention across grade levels (grades 3-5). Demonstrating the effect of the intervention on reading as well as on vocabulary was an important finding of the present study because engaging students in the inductive vocabulary intervention implicitly required them to focus attention on comprehension of each story. This "sideeffect" of the model serves as a potential explanation of the effect of the vocabulary intervention on reading achievement.

One important goal of future studies would be to
explore the cumulative effect of the present intervention on both the vocabulary development and reading proficiency of low-SES students when implemented on a multi-year basis. For practitioners, the present study is suggestive of how student vocabulary acquisition and reading proficiency can be accelerated through the enhancement of their regular reading programs. Considered together, the present findings are consistent with traditional and cognitive science research in that while recognizing vocabulary words can be taught directly, it is feasible for schools to accelerate student vocabulary growth in an inductive fashion that also improves student reading comprehension.

\section*{Acknowledgements}

Preparation of this paper was supported by IES Project R305A090523 to Florida Atlantic University.

\section*{References}

American Federation of Teachers (AFT). (1997). Making standards matter 1997. An annual fifty state report on efforts to raise academic standards. Washington, DC: AFT.
Anderson, R. C. \& Nagy, W. (1992). The vocabulary conundrum. American Educator, 16, 14-18, 44-47.
Anglin, J. M. (1993). Vocabulary development: A morphological analysis. Monographs for the Society of Research in Child Development, Serial No. 238, Vol. 58, No. 10.
August, D., C., M., Dressler, C., \& Snow, C. (2005). The critical role of vocabulary development for English language learners. The Division for Learning Disabilities of the Council for Exceptional Children, 20(1), 50-57.
Baker, S. K., Simmons, D. C., \& Kameenui, E. J. (1998). Vocabulary acquisition: Research bases. In D. C. Simmons and E. J. Kameenui (Eds.), What reading research tells us about children with diverse learning needs. Mahwah, NJ: Erlbaum.
Beck, I., \& McKeown, M. (1991). Conditions of vocabulary acquisition. In R. Barr, M. Kamil, P. Mosenthal, \& P. D. Pearson (Eds.), Handbook of reading research (Vol. 2, pp. 789-814). New York: Longman.
Becker, W. C. (1977). Teaching reading and language to the disadvantaged - What we have learned from field research. Harvard Educational Review, 47, 518-543.
Biemiller, A., \& Boote, C. (2006). An effective model for building meaning vocabulary in primary grades. Journal of Educational Psychology, 98(1), 44-62.

Biemiller, A., \& Slonim, N. (2001). Estimating root word vocabulary growth in normative and advantaged populations: Evidence for a common sequence of vocabulary acquisition. Journal of Educational Psychology, 93 (3), 498-520.
Blachowicz, C., \& Fisher, P. (2000). Vocabulary instruction. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, \& R. Barr (Eds.). Handbook of reading research. (Vol. 3 pp. 503-523). Mahwah, NJ: Erlbaum.
Bransford, J. D., Brown, A. L., \& Cocking, R. R. (Eds.). (2000). How people learn. Washington, DC: National Academy Press.
Carver, R. (1994). Percentage of unknown vocabulary words in text as a function of the relative difficulty of the text: Implications for instruction. Journal of Reading Behavior, 26(4), 413-437.
Coyne, M. D., McCoach, D. B., \& Kapp, S. (2005). Teaching vocabulary to kindergarten students during shared storybook readings: A comparison of rich instruction, basic instruction, and incidental exposure. Paper presented to the USDOE Institute of Educational Sciences Reading Comprehension Principal Investigator Group Meeting, Washington, DC.
Feldman, S. (2000). Standards are working: But states and districts need to make some mid-course corrections. American Educator, 24(3), 5-7.
Johnson, G., Gersten, R., \& Carnine, D. (1988). Effects of instructional design variables on vocabulary acquisition of learning disabilities students: A study of computerassisted instruction. Journal of Learning Disabilities, 20 (4), 206-213.
Kintsch, W. (1994). Text comprehension, memory, and learning. American Psychologist, 49(4), 294-303.
Kintsch, W. (1998a). Comprehension: a paradigm for cognition. Cambridge, United Kingdom: Cambridge University Press.
Kintsch, W. (1998b). The role of knowledge in discourse comprehension: a construction-integration model. Psychological Review, 95(2), 163-182.
Kintsch, W. (2002). On the notions of theme and topic in psychological process models of text comprehension. In M. Louwerse \& W. van Peer (Eds.) Thematics: Interdisciplinary Studies, 157-170.
Kintsch, W. (2004) The Construction-Integration model of text comprehension and its implications for instruction. In R. Ruddell \& N. Unrau (Eds.) Theoretical Models and Processes of Reading. 5th Edition, International Reading Association.
Kintsch, W. (2005). An overview of top-down and bottom-up effects in comprehension: The CI Perspective. Discourse Processes, 39 (2\&3), 125-128.

Kintsch, W. (2012). What kinds of language statistics must be in long-term memory to make language understanding possible. In M. A. Britt, S. R. Goldman, \& J. Rouet (Eds.). Reading- From words to multiple texts (pp. 92-100). NY: Routledge.
Landauer, T. K., (2002). On the computational basis of learning and cognition: Arguments form LSA. Psychology of Learning and Motivation, 41, 43-84.
Landauer, T. K., \& Dumais, S. T. (1996). How come you know so much? From practical problem to new memory theory. In D. Herrmann, C. McEnvoy, C. Hertzog, \& M. Johnson (Eds.) Basic and Applied Memory Research Theory in Context. 1(7), 105-126.
Landauer, T. K., \& Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104 (2), 211-240.
Landauer, T. K., Foltz, P. W., \& Laham, D. (1998). An introduction to latent semantic analysis. Discourse Processes, 25, 259-284.
Landauer, T. K., McNamara, D. S., Dennis, S., \& Kintsch, W. (2007). Handbook of latent semantic analysis. Psychology Press.
Nagy, W., \& Scott, J. A. (2000). Vocabulary Processes. Handbook of Reading Research, 3(18), 269-284.

Romance, N. R. \& Vitale, M. R. (2012). Effects of inductive vocabulary development in grades 3-4-5 on achievement in vocabulary and reading comprehension. Presented at the Annual Meeting of the American Educational Research Association, Vancouver, Canada.
Snow, C. E. (2002). Reading for understanding. Toward an \(R \& D\) program in reading comprehension. Santa Monica, CA: RAND.
Vitale, M. R., \& Romance, N. R. (2007). A knowledgebased framework for unifying content-area reading comprehension and reading comprehension strategies. In D. McNamara (Ed.). Reading Comprehension Strategies: Theory, Interventions, and Technologies. NJ: Erlbaum.
Vitale, M. R., \& Romance, N. R. (2008). A multi-part instructional strategy for accelerating vocabulary acquisition as an inductive process. Paper presented at the Annual Meeting of the American Educational Research Association, New York.
Wager, R. (2005). Causal relations between the development of vocabulary and reading comprehension. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada.

\title{
Cognitive load does not decrease pronoun use when speaker's and addressee's perspectives are dissociated
}

\author{
Jorrig Vogels (j.vogels@uvt.nl)
}

\author{
Emiel Krahmer (e.j.krahmer@uvt.nl)
}

\author{
Alfons Maes (maes@uvt.nl) \\ Tilburg center for Cognition and Communication, PO Box 90153 \\ 5000 LE Tilburg, the Netherlands
}

\begin{abstract}
It has been suggested that a referent's accessibility is affected by the degree to which it is in the speaker's attention. Assuming that less accessible referents are less likely to be pronominalized, this predicts that speakers under cognitive load use more elaborate referring expressions. However, speakers under load may also have difficulty taking into account their addressee's perspective, which may either lead to more use of the speaker's own discourse model or to more economic expressions. To tease these effects apart, we conducted a story completion experiment in which cognitive load was manipulated by the presence or absence of a secondary task for the speaker. In addition, we dissociated the speaker's and the addressee's perspectives. Our results do not provide evidence for the hypothesis that cognitive load reduces the accessibility of referents in the speaker's own discourse model, suggesting that speaker attention does not determine accessibility.
\end{abstract}

Keywords: cognitive load; referring expressions; accessibility; perspective taking

\section*{Introduction}

When speakers refer to something, they have to choose a certain type of referring expression, such as a definite description (e.g. the girl) or a pronoun (e.g. she). Traditionally, the speaker's choice of a referring expression has been assumed to be tailored for the addressee (e.g. Ariel, 1990; Gundel, Hedberg, \& Zacharski, 1993). According to this view, speakers make assumptions about the cognitive status of the referent in the mind of the addressee. An important factor in determining this status is the salience of the referent in the discourse. For example, if the referent was the topic of the preceding sentence, it can be assumed to be highly accessible in the addressee's discourse model, and therefore it does not need an elaborate description to be reactivated. Because the addressee knows that the speaker would have used a more elaborate expression if she had a less activated referent in mind, the use of an attenuated expression, such as a pronoun, aids the addressee's interpretation.

However, studies that have manipulated speakers' attention resources suggest that the activation of mental representations in the speaker's own memory is also important for the choice of referring expression. For example, Arnold and Griffin (2007) and Fukumura, Van Gompel, and Pickering (2010) varied the number of possible
referents in the discourse, and found that speakers used fewer pronouns when a referential competitor was present, even though the referent was salient in the discourse (i.e. topical). In addition, speakers have been found to choose fewer attenuated expressions for salient referents when they are distracted by another task (Rosa \& Arnold, 2011). These findings suggest that the choice of referring expression is affected by the degree to which the referent is in the speaker's attention: referent accessibility, and hence pronoun use, decreases when the speaker has to spread attentional resources over multiple possible referents or multiple tasks.

If restrictions on speakers' attention resources influence the accessibility of referents in their own memory, speakers experiencing an increased cognitive load should be more likely to use elaborate expressions such as full noun phrases. However, increased cognitive load may also affect the degree to which speakers are able to take into account the perspective of the addressee (e.g. Epley, Keysar, Van Boven, \& Gilovich, 2004; Horton \& Keysar, 1996). For example, Horton and Keysar (1996) showed that when under time pressure, speakers were not taking into account the addressee's perspective in choosing whether or not to include an adjective in their referring expressions. Thus, cognitive load may make the choice of referring expression more egocentric. On the one hand, this could mean that this choice is based more on the speaker's own discourse model (e.g. Fukumura \& Van Gompel, 2012). That is, when speakers are under load, they might be less able to calculate whether the referent is accessible for the addressee or not. Because in many cases speaker and addressee have access to the same discourse information, it is generally difficult to distinguish between a referring expression that is tailored for the addressee and a referring expression that is based on the speaker's own model of the discourse when all discourse information is shared. For example, a referent that is highly accessible in the addressee's discourse model is often also highly accessible in the speaker's discourse model. When the speaker's and the addressee's perspectives differ, however, speakers under load might be inclined to use pronouns when the referent is salient in their own discourse model but not salient in the addressee's discourse model. Conversely, they might be inclined to use full noun phrases when the referent is not salient in their own discourse model but salient in the addressee's discourse model.

On the other hand, speakers under load may fail to take into account any information about discourse salience, and resort to using more pronouns in general, because these are short, have little semantic content, and are hence easy to produce (Almor, 1999; Burzio, 1998). For example, studies on children and elderly people have found that having limited working memory capacity increases the use of pronouns in contexts in which the referent is not salient for the addressee (e.g. references following a topic shift; Hendriks, Englert, Wubs, \& Hoeks, 2008; Wubs, Hendriks, Hoeks, \& Koster, 2009; see also Almor et al., 1999).

To tease these possible effects of cognitive load apart, the linguistic salience of the referent should be varied, because the effect of cognitive load might be different for salient and non-salient referents. In addition, to determine whether cognitive load affects the speaker's discourse model or the speaker's assumptions about the addressee's discourse model, the speaker's and addressee's perspectives should be dissociated. Therefore, we conducted a story completion experiment in which we manipulated perspective, referent salience and cognitive load. Perspective was manipulated by presenting one of the sentences of the story only to the speaker, over headphones (cf. Fukumura and Van Gompel, 2012). The referent was considered linguistically salient for the speaker when it was mentioned in this privileged sentence, and not salient when it was only mentioned in the introductory sentence. Since the addressee did not hear the privileged sentence, referent salience was reversed for the addressee. Cognitive load was manipulated by giving the speaker a second, unrelated (but also verbal) task in the first or the second half of the experiment.

We hypothesized that if cognitive load lowers the accessibility of referents in the speaker's discourse model, it should decrease pronoun use, irrespective of the referent's linguistic salience. If, on the other hand, cognitive load makes perspective taking more difficult, there should be a stronger tendency to use the speaker's own discourse model in the dual task condition. That is, speakers should be more likely to use pronouns when the referent is salient in their own discourse model (and not in their addressee's), than when the referent is salient in the addressee's discourse model (and not salient in their own). Alternatively, speakers may tend to use more pronouns in general when under increased cognitive load.

\section*{Methods}

\section*{Participants}

Sixty-four students (47 female; mean age: 20.2 years) from Tilburg University participated in the experiment for course credit. Half of them acted as speakers, the others acted as addressees. All were native speakers of Dutch, the language of the experiment.

\section*{Materials}

The experimental items consisted of 16 pairs of photographs, taken from Vogels, Krahmer, and Maes (in
press), accompanied by two introductory sentences and the onset of a third sentence. The first picture of a pair always showed one male and one female person sitting next to each other. In the second picture, one of these persons performed an action, such as walking away or getting a glass of water. This person will be referred to as the target character, as participants were expected to refer to this character in their continuations. There were two versions of each picture pair; one in which it was the male person and one in which it was the female person that performed the action. An example of a picture pair is shown in Figure 1.


\section*{A: Speaker-salient condition}


\section*{B: Addressee-salient condition}

Figure 1: Example of a stimulus item in two conditions. Sentence 1 was read aloud by the speaker; sentence 2 was presented only to the speaker over headphones. Context sentences are translations of the Dutch originals.

The first sentence introduced both characters with indefinite noun phrases, either een meisje "a girl" and een jongen "a boy" or een vrouw "a woman" and een man "a man". One was mentioned as the subject, and the other in a
prepositional phrase (e.g. Een meisje zat te discussiëren met een jongen "A girl was arguing with a boy"). This sentence was read aloud by the speaker to the addressee. The second sentence described some emotional or physical state of the person mentioned in the prepositional phrase (e.g. De jongen raakte enorm gepikeerd "The boy got really annoyed"). This sentence was prerecorded by a female native speaker of Dutch, and only heard by the speaker over headphones. The onset of the third sentence was always Vervolgens... "Subsequently...", serving as a cue for the speaker to complete the story.

In addition, 20 picture pairs served as fillers. These differed from the experimental items in that some showed either two male or two female characters or only one character. In the accompanying sentences, some characters were given labels such as een verkoopster "a saleswoman" or een Duitser "a German", and sometimes the same character was the subject of both introductory sentences. An additional 4 items were included as practice items.

\section*{Procedure}

The experiment took place in an experimental room. Two participants were randomly assigned to the role of speaker and addressee. The participant taking the role of speaker was seated at one end of a table, behind a laptop connected to a serial response box, and was wearing headphones. The participant taking the role of addressee was seated at the other end of the table, and was given a booklet containing all different picture pairs and an answer sheet. The experiment was run on the laptop using E-Prime 2.0, and was only visible to the speaker. The speaker's task was to complete the stories depicted by the picture pairs for the addressee. In one half of the experiment, the speaker received a secondary task (cognitive load condition), while there was no secondary task in the other half (no cognitive load). In the no cognitive load condition, each trial started with the item number presented on the screen, accompanied by a 500 ms beep, followed by a cross-hair. Then, the first picture of a pair appeared on the left side of the screen. After 3 s the first introductory sentence appeared below the picture in a red font. The speaker read this sentence aloud to the addressee. After 5 s the second sentence was presented to the speaker over the headphones. Next, while the first picture remained visible, the second picture appeared automatically on the right side of the screen, together with the onset of the third sentence, which also appeared below the picture in a red font. The target character in the second picture was either the subject of the first introductory sentence, and therefore linguistically salient for the addressee (addressee-salient condition), or the subject of the second sentence, which was only presented to the speaker (speaker-salient condition). At this time, recording started, and the speaker completed the sentence based on the event shown in the picture, by saying it aloud to the addressee. After 6 s, recording stopped and the pictures and sentences disappeared. The addressee's task was to select the correct picture pair out of three options from the booklet and mark
the correct answer on the sheet. The addressee gave the speaker a hint when the next trial could be started.

In the cognitive load condition, the appearance of the first picture was preceded by the words BAL "ball" or DAL "valley" (Goudbeek \& Krahmer, 2011), which was presented in the middle of the screen for 1 s . The same happened at the end of the trial, followed by the question Was dit woord hetzelfde als het vorige woord? (Ja/Nee) "Was this word the same as the previous word? (Yes/No)". The speaker then pressed either the green/Yes or the red/No button on the response box. They did not receive feedback on their answers.

The participants received instructions both orally and in written form. Speakers were explicitly told that the sentence presented over headphones could not be heard by their addressee, but that they had to pay attention to it nonetheless, since they would be asked about these sentences after the experiment as an attention check (which was indeed the case). They were also encouraged to pay attention to the dual task by way of a prize offer for the participant with the fewest errors. To keep the speaker aware of the addressee's needs, the addressee was allowed to ask the speaker clarification questions if anything remained unclear, but only after the speaker had finished the story.

The experiment was divided into two blocks, of which one contained the dual task and the other did not, counterbalanced for order. Each block was preceded by two practice items. The experimenter was only present during the instructions and the practice trials. The experiment took about 25 minutes.

\section*{Data coding}

From all speakers' continuations of the third sentence, we selected the first subject reference, which was expected to refer to the target character. Any further references (e.g. in another follow-up clause) were ignored. We excluded 33 cases in which the first subject did not refer to the target referent, 7 plural references, 3 indefinite references, 1 case in which the sentence presented over the headphones was repeated literally, and 1 missing case. In addition, there were 2 cases in which the referring expression was repaired. However, because the repair was of the same type in both cases (e.g. 'the man... uh the boy'), we kept these cases. In total, we excluded 45 trials (8.8\%). The remaining 467 subject references were coded for the type of referring expression: either full noun phrase or pronoun.

\section*{Design and statistical analyses}

Crossing the two factors Referent salience and Cognitive load resulted in a 2 (speaker-salient, addressee-salient) x 2 (cognitive load, no cognitive load) within-participants design. Participants were assigned to one of four lists, each of which contained one version of a given item. The items were presented in a pseudo-random order, with at least one filler item between two consecutive experimental items.

We performed a logit mixed model analysis on the log odds for a pronoun (Jaeger, 2008). Referent salience and Cognitive load were included as fixed factors, and participants and items as random factors. The fixed factors were centered to reduce collinearity. Starting with a full random effect structure, we excluded random slopes that did not significantly contribute to the model fit. Only the final model will be reported.

\section*{Results}

Upon inquiry, only 6 participants reported that they found the secondary task difficult. Still, the overall error rate was \(9.7 \%\), suggesting that participants might have been overestimating their performance. Few errors were made in the attention check following the experiment (1.7\%), suggesting that speakers were attending to the sentences presented over the headphones.


Figure 2: Proportion of pronoun references to the target character in the four conditions.

Figure 2 shows the proportion of pronoun references to the target character by referent salience and cognitive load condition. We found a main effect of referent salience: pronouns were more frequent when the referent was salient only for the speaker (23.6\%) than when it was salient only for the addressee (8.3\%), \(\beta=-2.26\); \(\mathrm{SE}=0.85 ; p<.01\). There was also a main effect of cognitive load: more pronouns were used when speakers performed a dual task (17.2\%) than when they did not (15.7\%), \(\beta=1.40\); \(\mathrm{SE}=\) \(0.56 ; p<.05\). These effects were qualified by a significant interaction, \(\beta=2.85\); \(\mathrm{SE}=0.95 ; p<.01\), suggesting that cognitive load only increased pronoun use in the addresseesalient condition. The model included random intercepts for participants \(\left(s^{2}=2.85\right)\) and items \(\left(s^{2}=0.15\right)\), as well as bysubject random slopes for referent salience \(\left(s^{2}=12.06\right)\) and cognitive load ( \(\mathrm{s}^{2}=3.08\) ). This suggests that participants
varied substantively in the way their pronoun use was affected by the context sentences and the dual task. The contribution of the random slope for cognitive load to the model fit was only marginally significant ( \(p=.06\) ). Removing it decreased the effect size of the fixed factors, but the interaction effect remained significant at the \(\alpha=.05\) level.

\section*{Discussion}

Speakers used more pronouns when the referent was salient for them (but not salient for their addressee) than when it was not salient for them (but salient for their addressee). This suggests that speakers chose referring expressions more according to the referent's accessibility in their own discourse model than according to assumptions about the referent's accessibility in their addressee's discourse model. This is in line with Fukumura and Van Gompel (2012), who found that speakers were not taking into account their addressee's perspective in choosing referring expressions when the two perspectives were dissociated.

If cognitive load decreases the accessibility of referents in the speaker's own discourse model, as suggested by Arnold and Griffin (2007), we should have seen fewer pronouns across the board when speakers performed the dual task. This is not what we found. Instead, speakers were somewhat more likely to use pronouns when under load, at least in the addressee-salient condition. This finding is also not in line with the hypothesis that cognitive load makes it harder to calculate the referent's accessibility in the addressee's discourse model. If that were the case, the dual task condition should have increased the tendency to use the speaker's own discourse model. That is, speakers under load should have been more likely to use pronouns in the speaker-salient condition, and less likely to use pronouns in the addressee-salient condition.

Our results seem compatible with the hypothesis that speakers under load are more likely to use less costly expressions, such as pronouns (Almor, 1999; Burzio, 1998). That is, when distracted by a secondary task, speakers have fewer memory resources available that are needed to infer that less salient referents should be referred to with more elaborate expressions. The fact that cognitive load only increased pronoun use in the addressee-salient condition may be due to the preference to use pronouns anyway when the referent is salient for the speaker.

An alternative explanation for the effect of cognitive load is that speakers are less able to keep track of their own discourse model when they are under load. This would cause their use of referring expressions to become less consistent, i.e. the choice of referring expressions becomes less tied to the discourse context (Arnold, 2010). This would explain the finding that the difference between the speakersalient and the addressee-salient condition becomes smaller under load. In addition, this explanation would not be incompatible with the finding by Arnold and Griffin (2007), Fukumura et al. (2010) and Rosa and Arnold (2011) that cognitive load leads to fewer attenuated expressions, since
these studies investigated only contexts in which referents were always linguistically salient. Thus, the present study at least stresses that dissociating the salience of the referent in the speaker's and the addressee's perspective is necessary to tease the possible effects of cognitive load apart, i.e. whether it affects the speaker's representation of the discourse or the speaker's assumptions about the addressee's representation of the discourse.
Although our results suggest that speakers tend to use their own discourse model, replicating Fukumura and Van Gompel's (2012) findings, it is striking that the overall proportion of pronouns is quite low. Even in the no cognitive load, speaker-salient condition, in which one would expect many pronouns if speakers only took into account their own discourse model, the percentage of pronouns out of all referring expressions did not exceed \(30 \%\). This relatively low proportion of pronouns might be due to the manipulation of perspective: Perhaps speakers were taking into account the addressee's informational needs, but not up till the level of calculating the referent's cognitive status for the addressee. This kind of detailed audience design might be cognitively too costly (e.g. Brennan \& Hanna, 2009; Horton \& Gerrig, 2005). Therefore, speakers may just have increased the use of elaborate expressions to be as clear as possible for the addressee, as soon as they were aware of the fact that not all information was shared. Fukumura and Van Gompel (2012) found evidence for such a minimal, one-bit model of audience design (e.g. the addressee has heard this or not; Galati \& Brennan, 2010; see also Epley et al., 2004) by comparing their condition with privileged information for the speaker to a condition in which all information was shared. They found more pronouns in the shared condition, independently of whether the referent was salient or not. This suggests that speakers use more elaborate expressions when there is privileged information, even though they might run the risk of being overly specific.

Still, Fukumura and Van Gompel (2012) found somewhat higher rates of pronoun use in their privileged, referentsalient condition (Experiment 1: 37\%; Experiment 2: 48\%) than we did in our experiment. This difference could be due to differences in the linguistic materials. Firstly, while the referent mentioned in the second context sentence (which was only presented to the speaker) was referred to with a pronoun in Fukumura and Van Gompel's experiments, it was referred to with a full NP in our experiment, in accordance with the preferred way of referring to an entity previously mentioned as a direct object in centering theory (e.g. Brennan, 1995). The tendency to pronominalize the entity on a subsequent reference may however be stronger when the referent had already been pronominalized. Secondly, speakers may be more likely to reuse the most recent referring expression, which could also have led to more pronouns in Fukumura and Van Gompel's experiments than in ours.
Even though employing a one-bit model of audience design is probably less cognitively demanding than
calculating the referent's accessibility for the addressee, using full NPs to aid the addressee when there is privileged information might be more difficult under load. Hence, if speakers in our study were employing such minimal audience design, our finding that they were more likely to use pronouns under load could be due to difficulties in assessing that the addressee might need more specific information. If this is the case, we would predict that in a situation in which all discourse information is shared, cognitive load does not increase pronoun use, since in that case there is no need to be more specific for the addressee.

One reason why speakers did not make the extra effort to calculate the accessibility of the referent in the addressee's discourse model may be that in the current experiment, as well as in Fukumura and Van Gompel's, references were never ambiguous, since the two characters always had a different gender. Therefore, not taking into account the addressee's perspective would probably not result in interpretation errors. However, in case not taking into account the addressee's perspective would lead to interpretation errors, speakers may base their choice of referring expressions on the discourse model of their addressee (e.g. Fukumura \& Van Gompel, 2012; Horton \& Keysar, 1996). In that case, increased cognitive load might make this perspective taking more difficult, and cause speakers to fall back on their own discourse model.
In our filler materials, which contained stories with characters of the same gender, and hence pronouns were ambiguous, we indeed found more pronouns when the referent was salient for the addressee but not for the speaker ( \(33 \% ; \mathrm{n}=51\) ) than when the referent was salient for the speaker but not for the addressee ( \(13 \%\); \(\mathrm{n}=56\) ), suggesting that speakers were taking their addressee's perspective into account. However, cognitive load did not seem to cause speakers to use their own discourse model. Rather, a pattern similar to that in Figure 2 emerged, with more pronouns under load for referents that were not salient for the speaker. This might be an indication that cognitive load as manipulated here is independent of perspective taking.
Our results suggest that there was quite some individual variation as to how speakers' referring expressions were affected by the dual task. One cause of individual differences could be the use of strategies for remembering the words BAL and DAL. Two thirds of all participants reported to have used some kind of mnemonic (e.g. putting up one finger for BAL and two for DAL), although these were not always employed from the beginning. This is a concern that should be taken up by future studies. Nevertheless, the general trend of more pronouns under load in the addressee-salient condition in the present study seems to hold for all participants.

Finally, an important issue is that cognitive load can be manipulated in different ways that may affect the choice of referring expressions differently. For example, it is not clear whether dual tasks, divided attention to multiple referents, and restricted working memory capacity all produce the same kind of cognitive load. In addition, language
production may be affected differently by verbal or visual secondary tasks (Baddeley and Hitch, 1974; Kellogg et al., 2007). In the present experiment, the use of a verbal secondary task may have especially hindered the production of elaborate linguistic forms. Other manipulations, such as adding time pressure, possibly interfere more with activating non-linguistic representations or with perspective taking. This is an issue that needs further research.

In sum, the present study has shown that speakers use more pronouns when they experience increased cognitive load, at least when the referent is not salient for the speaker (but is salient for the addressee). Whether this is due to a general preference to produce economic forms, or to difficulties in keeping track of the accessibility of the referents in the discourse model should be researched further. However, we have not found support for the claim that cognitive load, at least in the form of the dual task used here, decreases the accessibility of referents in the speaker's discourse model. Hence, although accessibility may be related to attention, it probably does not hold generally that less attentive speakers use more elaborate referring expressions.

\section*{References}

Almor, A. (1999). Noun-phrase anaphora and focus: The informational load hypothesis. Psychological Review, 106(4), 748-765.
Almor, A., Kempler, D., MacDonald, M. C., Andersen, E. S., \& Tyler, L. K. (1999). Why Do Alzheimer Patients Have Difficulty with Pronouns? Working Memory, Semantics, and Reference in Comprehension and Production in Alzheimer's Disease. Brain and Language, 67, 202-227.
Ariel, M. (1990). Accessing noun-phrase antecedents. London: Routledge.
Arnold, J. E., \& Griffin, Z. (2007). The effect of additional characters on choice of referring expression: Everyone counts. Journal of Memory and Language, 56, 521-536.
Arnold, J. E. (2010). How Speakers Refer: The Role of Accessibility. Language and Linguistics Compass, 4(4), 187-203.
Baddeley, A. D., \& Hitch, G. (1974). Working Memory. In G. H. Bower (Ed.), The Psychology of Learning and Motivation (Vol. 8). New York/London: Academic Press.
Brennan, S. E. (1995). Centering Attention in Discourse. Language and Cognitive Processes, 10(2), 137-167.
Brennan, S. E., \& Hanna, J. E. (2009). Partner-Specific Adaptation in Dialog. Topics in Cognitive Science, 1, 274-291.
Burzio, L. (1998). Anaphora and Soft Constraints. In P. B. D. Fox, P. Hagstrom, M. McGinnis \& D. Pesetsky (Eds.), Is the Best Good Enough?: Optimality and Competition in Syntax. Boston: MIT.
Epley, N., Keysar, B., Van Boven, L., \& Gilovich, T. (2004). Perspective Taking as Egocentric Anchoring and Adjustment. Journal of Personality and Social Psychology, 87(3), 327-339.

Fukumura, K., \& Van Gompel, R. (2012). Producing Pronouns and Definite Noun Phrases: Do Speakers Use the Addressee's Discourse Model? Cognitive Science, 36(7), 1289-1311.
Fukumura, K., Van Gompel, R., \& Pickering, M. J. (2010). The use of visual context during the production of referring expressions. The Quarterly Journal of Experimental Psychology, 63(9), 1700-1715.
Galati, A., \& Brennan, S. E. (2010). Attenuating information in spoken communication: For the speaker, or for the addressee? Journal of Memory and Language, 62(1), 35-51.
Goudbeek, M., \& Krahmer, E. (2011). Referring under load: Disentangling preference-based and alignment-based content selection processes in referring expression generation Proceedings of PRE-Cogsci: Bridging the gap between computational, empirical and theoretical approaches to reference. Boston.
Gundel, J. K., Hedberg, N., \& Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. Language, 69, 274-307.
Hendriks, P., Englert, C., Wubs, E., \& Hoeks, J. (2008). Age Differences in Adults' Use of Referring Expressions. Journal of Logic, Language and Information, 17, 443466.

Horton, W. S., \& Gerrig, R. J. (2005). The impact of memory demands on audience design during language production. Cognition, 96, 127-142.
Horton, W. S., \& Keysar, B. (1996). When do speakers take into account common ground? Cognition, 59, 91-117.
Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. Journal of Memory and Language, 59, 434-446.
Kellogg, R. T., Olive, T., \& Piolat, A. (2007). Verbal, visual, and spatial working memory in written language production. Acta Psychologica, 124, 382-397.
Rosa, E. C., \& Arnold, J. E. (2011). The Role of Attention in Choice of Referring Expression Proceedings of PRECogsci: Bridging the gap between computational, empirical and theoretical approaches to reference. Boston.
Vogels, J., Krahmer, E., \& Maes, A. (in press). Who is where referred to how, and why? The influence of visual saliency on referent accessibility in spoken language production. Language and Cognitive Processes.
Wubs, E., Hendriks, P., Hoeks, J., \& Koster, C. (2009). Tell me a story! Children's capacity for topic shift Proceedings of the 3rd Conference on Generative Approaches to Language Acquisition North America (GALANA 3) (pp. 313-324). Somerville, MA: Cascadilla Press.

\title{
Rural and urban differences in language socialization and early vocabulary development in Mozambique
}

\author{
Paul Vogt (p.a.vogt@uvt.nl) \\ Tilburg center for Cognition and Communication, Tilburg University \\ P.O. Box 90153, 5000 LE, Tilburg, The Netherlands \\ J. Douglas Mastin (j.d.mastin@uvt.nl) \\ Tilburg center for Cognition and Communication, Tilburg University \\ P.O. Box 90153, 5000 LE, Tilburg, The Netherlands
}

\begin{abstract}
We investigate the amount of speech and (co-speech) gestures addressed to infants at \(1 ; 1\) years of age in rural and urban Mozambique, and correlate these amounts with vocabulary size measured at \(1 ; 5\) and \(2 ; 1\). We found that urban infants are exposed to more than three times as much speech and cospeech gestures than rural infants. The results show that the amounts of co-speech gestures and speech predict later vocabulary development in the urban community, but not in the rural community. The results further show that rural infants are delayed in their vocabulary development, which may in part be explained by a transition in the socialization style rural infants experience between the age of \(1 ; 1\) and \(1 ; 5\).
\end{abstract}

Keywords: Child language acquisition; child-directed speech; co-speech gestures; vocabulary development; Mozambique.

\section*{Introduction}

When children learn language, they must have exposure to the target language. It is well established that the amount of exposure correlates strongly to vocabulary development (Hart \& Risley, 1995; Pan, Rowe, Singer, \& Snow, 2005). This does not only hold for the amount of speech children are exposed to, but also for the amount of gestures directed at children (Iverson, Capirci, Longobardi, \& Caselli, 1999; Rowe \& Goldin-Meadow, 2009). In this report we investigate how the amounts of speech and gestures addressed to infants vary among rural and urban communities in Mozambique, and show how these amounts correlate with vocabulary sizes during infants' early development.
One obvious predictor of children's vocabulary development is the amount of verbal input addressed to them. Various studies have, indeed, revealed a strong correlation between parental verbal input and vocabulary development (Hart \& Risley, 1995; Pan et al., 2005). It is not just the amount of words a child is exposed to, but also the variety of words that correlates to later vocabulary size (Hart \& Risley, 1995). It has further been found that the amount of parental verbal input addressed to children, as well as the speed of children's vocabulary development, relates to the parents' social economic status (SES) - the higher the parents' SES, the more words they tend to
address to children, and the larger these children's vocabularies become (Hart \& Risley, 1995).
As with speech, the amount of hand gestures addressed to infants, such as pointing, showing, or iconic gestures, are good predictors of vocabulary development (Iverson et al., 1999; Pan et al., 2005). Infants’ gesture use also predicts vocabulary size (Pan et al., 2005), possibly due to a correlation between parental gesture use and infants gesture use (Iverson et al., 1999). As with the amount of speech, SES predicts the amounts of parents' and infants' gesture use, which relates to later vocabulary size (Rowe \& GoldinMeadow, 2009). One explanation for the role that gestures have on vocabulary development is that gestures help to establish and sustain joint attention, which in turn supports vocabulary development (Carpenter, Nagell, Tomasello, Butterworth, \& Moore, 1998).
It is important to realize that most of these studies were carried out in industrialized societies, but socialization towards children can differ greatly across cultures, and many non-industrial cultures have different attitudes towards child rearing (Schieffelin \& Ochs, 1989). For instance, there are cultural differences regarding the amount of socialization a child is involved in - typically there is relatively little speech directed towards infants from nonindustrialized cultures (Lieven, 1994). Moreover, multiparty interactions are more frequent in non-industrialized than in industrialized communities, and infants tend to have multiple caregivers, including siblings (Brown, 2011; Harkness, 1977). Also, the amount of language socialization depends on the developmental status of a country - mothers from countries higher on a developmental scale tend spend more time stimulating their children by reading books, telling stories, naming, counting and other cognitive tasks (Bornstein \& Putnick, 2012).
Within culture differences may exist between urban and rural communities. Keller (2012), for instance, has proposed that, in addition to prototypical Western urban communities, there are prototypical rural and urban communities in nonindustrialized countries. She has described a number of key characteristics in which non-industrialized rural and urban communities differ. The subsistence-based farming lifestyle
in rural communities, for instance, demands from children that they develop motoric skills and knowledge of social rules. Verbal skills are considered less important. Urban societies tend to be higher educated and expect from their children to receive a good education as well. As a result, linguistic interactions tend to become more important (LeVine et al., 1996). However, non-Western urban communities still adhere to many cultural traditions rooted in their rural decent, such as the role of communal responsibilities from the extended family members in child rearing (Keller, 2012).
In sum, these observations predict that infants from nonindustrialized urban communities would be exposed to more child-directed speech than infants from rural communities. Since there is a tight link between speech and gesture (McNeill, 1985), we would expect to find similar differences regarding the use of gestures and co-speech gestures addressed to infants. A previous analysis of the same observations presented in the present paper, however, has revealed no significant differences between a rural and an urban community in Mozambique regarding the amount of social interactions that young infants have with the members of their extended families (Mastin \& Vogt, 2013). These social interactions were based on the infants' attention states (Bakeman \& Adamson, 1984), and include both verbal and non-verbal interactions. This raises the question whether there are actual differences in the amount of speech addressed to the infants.
Based on the above-mentioned studies, which have demonstrated that speech and gestures are sound predictors for vocabulary development, we further expect to find that the amounts of speech, gestures, and co-speech gestures predict later vocabulary. In this paper, we try to confirm these predictions using a longitudinal ethnographic study among infants from rural and urban Mozambique.

\section*{Methods}

\section*{Participants and field sites}

We selected two field sites in Mozambique: one site compiled from two adjacent residential suburbs in the country's capital of Maputo; the other site was made up of three small villages just outside the rural, provincial town of Chokwe in Gaza province, about 200 kilometers away from the capital. From each community we recruited 22-25 families with an infant in the range of \(1 ; 0\) to \(1 ; 2\) years old ( \(1 ; 1\) on average) at the start of our study. Our local research assistants explained the general purpose and procedures of our study to the participating families in their native language, and we obtained a signed informed consent from the infants' mothers. During the course of our longitudinal study, we lost various participants due to illness, mortality or relocation. In addition, we removed two participants from our analysis, because the parental reports on vocabulary development showed a decrease in expressive vocabulary,
which rendered their data unreliable. As a result, we provide results for 14 participants from each field site.
The participants from the rural community were all native Changana speakers (a Southern Bantu language spoken in parts of Mozambique and South Africa); in most cases this was the only language spoken in the household. Only in a few families was another related local language occasionally spoken. In the urban community, most families raise their children bilingually in Mozambique's official language of Portuguese, and Ronga, a language that is mutually intelligible with Changana. Table 1 shows some demographic information concerning our participants.

Table 1: Demographic information. Note: Primary education in Mozambique is organized in two levels of primary school: EP1 for 5 years and EP2 for another 2 years.
\begin{tabular}{lll}
\hline Participant information & \begin{tabular}{l} 
Rural \\
\((n=14)\)
\end{tabular} & \begin{tabular}{l} 
Urban \\
\((n=14)\)
\end{tabular} \\
\hline Males / Females & \(7 / 7\) & \(9 / 5\) \\
Avg age (SD) & \(1 ; 1.8\) & \(1 ; 1.6\) \\
& \((0 ; 0.26)\) & \((0 ; 0.28)\) \\
Education level mothers & & \\
None & 6 & 1 \\
EP1 & 5 & 5 \\
EP2 & 3 & 6 \\
Higher & 0 & 1 \\
\hline
\end{tabular}

There was a fairly balanced split in the number of males and females participating, and the average age was equivalent in both sites. To have an indication of the families SES, we report the mothers' education level. The majority of rural mothers have either completed no education or only the lower levels of education, whereas urban caregivers have all received some education: five mothers have completed the lowest level of education, six have the second level of education, and one has received secondary education. Since the data on education is ordinal, we performed Fisher's exact test to verify whether the education levels of both communities differ significantly and found that it appears not ( \(p=.115\) ). However, when we compared the rural community with the urban community from our (unpublished) norming study using the Chisquared test, we found a significant effect in educational level ( \(\chi^{2}(3)=32.414, \mathrm{p}<.001\) ), while the urban participants' education fits nicely with our norming study \(\left(\chi^{2}(3)=1.318\right.\), \(\mathrm{p}=.725\) ).

\section*{Materials}

We adapted the MacArthur-Bates Communicative Development Inventories (MBCDI) Short Form Vocabulary Checklist (Fenson et al., 2000) into both Portuguese/Ronga and Changana to obtain a parental checklist of words used to measure vocabulary size and development in both Mozambican communities. To do this, we compiled a list
from Fenson et al.'s Level I for infants and extended this with 13 additional items from the Level II checklist to allow the list to be used for children older than 16 months. We then identified vocabulary that was not applicable to the environment, culture or lifestyle of our participants, and replaced these items with appropriate vocabulary that matched the same syntactic or semantic functions as the original English word. The list was further adapted during extensive piloting of the checklist. With the adapted CDIs we conducted a norming study in both communities to obtain expected values of vocabulary development. For details on the adaptation and norming of the MBCDI, consult Mastin and Vogt (2013).

\section*{Data collection}

Data was collected longitudinally at three periods during the course of one year, while the infants were on average \(1 ; 1,1 ; 5\) and \(2 ; 1\) years old. At each time-period, we visited each family twice. At the first visit, we administered a short survey with questions concerning the demographics of our participants. We also videotaped the infants' interactions with their families to allow them to accommodate to our presence. During the second visit, we started with video taping the infants from 45 up to 75 minutes during natural free behavior for data analysis. At both visits, we instructed the families to continue their daily routines and to act as if we were not present. After the video recording was finished, the adapted MBCDI was administered through face-to-face interviews held by a local research assistant to estimate the infants' vocabulary development. In the current study, we report on the video recordings at \(1 ; 1\) and correlate these to the infants' expressive vocabulary at \(1 ; 5\) and \(2 ; 1\) as measured using the MBCDI.

\section*{Coding procedures}

The videos recorded during the second visits were coded for 30 minutes in segments of prolonged duration in which the infant was displaying 'natural' behavior (i.e.: not sleeping, not off camera, not interacting with or disturbed by the experimenters). We also excluded prolonged periods (roughly more than 2 minutes) of breastfeeding, as this might have introduced a bias toward dyadic interactions. For this article, we present results on only child-directed speech and gestures.

Child-directed utterances Two local research assistants, while closely supervised by the authors, transcribed all child-directed speech. All intelligible speech was first transcribed into the local language and subsequently translated into Portuguese. All unintelligible speech were coded as unknown vocalizations, but were included in our current analyses. Because not all speech was intelligible, we measured the number of utterances (i.e. individual speech acts), rather than number of words.

Gestures We coded gestures during episodes of joint engagement (Mastin \& Vogt, 2013), which are activities in which the infants are socially interacting with one or more other individuals. These activities involve dyadic person interactions, as well as different types of triadic joint attention interactions based on those defined by Bakeman and Adamson (1984). Since many interactions observed involved multi-party interactions, we only coded those gestures produced by the communication partner nearest to the child. We adopt a broad definition of gestures as any physical activity with the hand or body that has a clear communicative intent (Zukow-Goldring, 1996). The following gestures were coded:
- Pointing is a gesture where the gesturer extends the arm to indicate an object with the hand or index finger from some distance.
- Showing is a gesture in which an object is indicated using zero proximity, e.g. by tapping on the object or by holding up the object.
- Demonstrating is a gesture where the speaker manipulates an object to show the infant how that object is used, or the type of actions that can be performed upon it.
- Reaching occurs when someone extends his/her arm to obtain or to touch an object, but can (or does) not reach this object. Also requests for objects by extending the hand were included in this category.
- Offering occurs when the speaker offers (or gives) an object or good to the infant.
- Taking occurs when someone takes over possession of an object from someone else.
- Conventional gestures comprises gestures that are symbolic of nature, such as emblematic gestures, but also gestures that bear an iconic relationship with their referent. For example, waving bye-bye, or indicating the size of the target object with the hands.
- Ritualized play accounts for all ritualized interactions or displays that occur between infants and communication partners. For instance, dancing, clapping hands or turntaking games, such as patty-cake.
- Embody occurs when someone directs another by physically "putting them through the motions of some activity" (Zukow-Goldring, 1996, p. 200), provided this has a communicative (or otherwise intentional) function. For example, placing the child on the mother's lap, pushing the infant in a certain direction, or taking someone's hand to demonstrate an action.
- Request for attention comprises any gesture that seeks for the attention of the interaction partner.
For the present study, we collapsed all gesture categories and report on the average number of gesture tokens addressed to the infants.
Both authors coded approximately half of all videos each, after which the coding was assessed and refined using improved coding schemes twice by trained research assistants. Both authors then coded approximately \(20 \%\) of
the video material to calculate inter-rater agreement with the final results. The resulting Cohen's kappa was measured to be 0.67 ( \(84.9 \%\) agreement), which according to Landis and Koch (1977) can be classified as 'substantial'.

Co-speech gestures After coding all gestures, we marked those gestures accompanied by a child-directed utterance as a co-speech gesture. We report the average number of cospeech gestures addressed to infants.


Figure 1: This figure shows the average amounts of childdirected utterances, gestures and co-speech gestures in the rural and urban areas at \(1 ; 1\). Error bars indicate standard deviations. All differences between communities are significant ( \(\mathrm{p}<.001\) ).

\section*{Results}

Figure 1 shows the average number of utterances, gestures and co-speech gestures addressed to the infants from the rural and urban communities. The graph reveals that urban communication partners address substantially more utterances, gestures and co-speech gestures than their rural counterparts. The number of child-directed utterances observed in the urban community is 5.7 times higher than observed in the rural area (according to the Mann-Whitney U test, \(\mathrm{U}=8 ; \mathrm{p}<.001\) ). The number of gestures - both with or without simultaneous speech - is 2.0 times higher in the urban community than in the rural \((\mathrm{U}=29 ; \mathrm{p}=.001)\). The frequency of child directed co-speech gestures occurs 3.2 times more in the urban community \((\mathrm{U}=22, \mathrm{p}<.001)\).

Table 2: Spearman correlations \(r_{s}\) of total amounts of child directed utterances, gestures and co-speech gestures at \(1 ; 1\) with expressive vocabulary development at both \(1 ; 5\) and 2;1. Note: * \(<\mathbf{p} 05\).
\begin{tabular}{lllll} 
& At 1;5 & & At 2;1 & \\
& Rural & Urban & Rural & Urban \\
\hline Utterances & -0.178 & \(0.554^{*}\) & 0.055 & \(0.607^{*}\) \\
\begin{tabular}{llll} 
Gestures
\end{tabular} & -0.270 & 0.482 & 0.256 & 0.508 \\
\begin{tabular}{l} 
Co-speech \\
gestures
\end{tabular} & -0.061 & \(0.667^{*}\) & 0.139 & 0.520 \\
\hline
\end{tabular}

Table 2 shows the Spearman correlations between the amounts of utterances, gestures and co-speech gestures produced at the infants' age of \(1 ; 1\) and expressive vocabulary sizes at infants' ages of \(1 ; 5\) and \(2 ; 1\). The first observation we can make is that urban child-directed utterances and co-speech gestures have significant correlations to expressive vocabulary size at \(1 ; 5\) (utterances: \(\mathrm{r}_{\mathrm{s}}[14]=0.554, \mathrm{p}<.05\); co-speech gestures: \(\mathrm{r}_{\mathrm{s}}[14]=0.667\), \(\mathrm{p}<.05\) ) and at \(2 ; 1\) (utterances: \(\mathrm{r}_{\mathrm{s}}[14]=0.607 ; \mathrm{p}<.05\) ). Note that all other correlations from the urban community approach significance ( \(\mathrm{p}<.10\) ). The second observation is that from the rural community, no significant correlations with vocabulary are revealed.

Given the differences in the amount of cognitive stimulation between both communities, we would expect to also see differences in the language development between the two communities, and we do (Table 3). The urban infants have a substantially larger expressive vocabulary than the rural infants. According to a two-way ANOVA on Age \(x\) Site, we see a main effect for age \((\mathrm{F}(2,78)=79.91\); \(\mathrm{p}<.001\) ) and for site \((\mathrm{F}(1,78)=13.41 ; \mathrm{p}<.001)\), and no interaction ( \(\mathrm{p}=.221\) ). A Tukey post-hoc test confirms the main effect of age ( \(\mathrm{p}<.001\) ).

Table 3: The average scores and standard deviations on expressive vocabulary from the MBCDI at \(1 ; 5\) and \(2 ; 1\) for both field sites. Note: Differences between urban and rural are significant as indicated with \({ }^{*} \mathrm{p}<.05\) and \(* * \mathrm{p}<.01\).
\begin{tabular}{lll}
\hline & At \(1 ; 5\) & At \(2 ; 1\) \\
\hline Rural & \(17.71(12.23)\) & \(50.85(23.59)\) \\
Urban & \(29.00(19.61)^{*}\) & \(72.92(23.18)^{* *}\) \\
\hline
\end{tabular}

\section*{Discussion}

The objective of this paper is to investigate whether infants from rural Mozambique experience less verbal and non-verbal stimulation than infants from urban Mozambique, and to assess how this correlates to later vocabulary development. The results clearly demonstrate that there are substantial differences between the rural and urban communities at all measured levels, i.e. the amounts of speech (as measured in utterances), gestures and cospeech gestures. This confirms Keller's (2012) predictions, but appears in contrast to our earlier findings that the total amounts of social interactions the same infants engage in whether these are verbal or non-verbal - are roughly the same in both communities (Mastin \& Vogt, 2013).

The difference in child-directed stimulation between rural and urban is largest regarding the number of utterances, which is 5.7 times higher in the urban community than in the rural community. This is considerably more than the difference in the amount of gestures ( 2.0 times higher) or the amount of co-speech gestures ( 3.2 times higher). Further analysis of the results from Figure 1 reveals that in both communities people use gestures more often than
utterances. However, this happens more in the rural community than in the urban community ( 3.5 times vs. 1.2 times). Moreover, we can infer that, on average, almost each utterance in the rural community is accompanied by a cospeech gesture, while in the urban community every other utterance is accompanied by a co-speech gesture.

These findings demonstrate that in the rural community relatively many social interactions with infants are nonverbal interactions. For instance, mothers may massage the infant's body, feed the infant or point to an object without talking. However, when the rural infant is addressed verbally, a gesture usually accompanies the speech. Urban infants are talked to much more frequently, but only half of the utterances addressed to them are accompanied by a gesture. In addition, although the link between speech and gesture appears less strong in the urban community, the absolute amount of child-directed gestural input for urban infants is much larger than for rural infants. Thus, the urban community, indeed, provides a richer language environment for the young than the rural communities do (Keller, 2012; LeVine et al., 1996).

The reason for this difference may well be due to the needs that different lifestyles demand of children when they grow older (Keller, 2012). In the rural community there is more need for children to help in the field or in the household, whereas the urban community value educational prospects for their children. However, other factors may contribute to these differences as well. For instance, there is a small difference regarding the educational levels that mothers obtained, so SES is likely to be a factor (Hart \& Risley, 1995; Rowe \& Goldin-Meadow, 2009). Also, due to globalization, the urban community may have adopted a more Western-like child-oriented socialization pattern (Schieffelin \& Ochs, 1989). Furthermore, there is the possibility that mothers in the rural area are less socially attached to their children until a certain age, either because of the high child mortality rates or because of cultural beliefs. Results from interviews we held indicate that in the rural area many mothers do not consider their child part of the community until well past their first birthday, while in the urban area most mothers considered their child a community member at birth or at least before they reach 6months (Mastin \& Vogt, 2013). Of course, additional issues such as health may play a role, and most likely a combination of factors explains why urban infants are exposed to more speech and gestures. Further research is required to understand why there is so much less childdirected speech in the rural community than in the urban.

Although on average the urban community do not gesture with each utterance, the amount of co-speech gestures correlates strongly to vocabulary size at \(1 ; 5\). Moreover, the amount of utterances addressed to infants in the urban community reveals significant correlations to vocabulary both at \(1 ; 5\) and \(2 ; 1\). So, these findings correspond well to results from earlier research in Western cultures (Hart \&

Risley, 1995; Iverson et al., 1999; Rowe \& Goldin-Meadow, 2009). However, the amounts of speech and gesture do not correlate to vocabulary development in the rural area, which contradicts these previous studies and is hard to explain.

One possible explanation is as follows: a yet unpublished analysis of the amount of social interactions infants engage in with different communication partners reveals that the amount of interactions with mothers is stable over time during infants' second year of life in the urban community. In the rural community, however, the amount of motherinfant interactions reduces substantially between \(1 ; 1\) and 1;5, while interactions with siblings increase by approximately the same amount, and which come to equal those of caregivers in frequency by the age of \(2 ; 1\). Thus, rural infants need to adapt more to changing caregiving structures than urban infants do, with the consequence that the rural socialization structure at \(1 ; 1\) is neither the same as the socialization structure at \(1 ; 5\) nor at \(2 ; 1\). The amounts of speech and gestures at \(1 ; 1\) may therefore not be viable predictors for vocabulary development in the rural area. Analysis of speech and gesture use at \(1 ; 5\) and \(2 ; 1\) should shed new light on this issue. If the interpretation provided here is correct, we expect that co-speech gesture use at \(1 ; 5\) in the rural community to be a better predictor for vocabulary size at \(2 ; 1\) than its use at \(1 ; 1\).

The results from the scores on the vocabulary checklist (Table 3) suggest a difference in the development of vocabulary in both communities. Despite the absence of a correlation between input and vocabulary in the rural community, the most likely candidate for this difference is indeed the difference in the amounts of speech, and consequently the amount of co-speech gestures, that infants are exposed to (Hart \& Risley, 1995; Pan et al., 2005). However, Harkness (1977) observed in rural Kenya that children spending more time with adult caregivers tend to talk more and become linguistically more advanced than children who spend more time with sibling caregivers. So, the differences in the socialization structure may be another candidate. A deeper analysis of who infants socialize with more frequently over time and how this relates to vocabulary development should provide new insights into the role of different caregivers on the infants' word learning processes.

The data presented in this paper are being annotated to develop corpora of multimodal interactions between infants and their social environment that can be incorporated in computer models (Matusevych, Alishahi, \& Vogt, 2013; Vogt \& Mastin, 2013). Using these corpora, we aim to mimic the observed interactions between infants and their surroundings as realistically as possible in multi-agent simulations. With such simulations, we plan to investigate various socio-cognitive theories explaining language development using realistic scenarios in which agents interact socially using speech and gestures according to observed frequencies, and measure the vocabulary
development of the simulated children. One possible application of such a simulation could be to analyze the socio-cognitive mechanisms that underlie the findings from this paper. The envisioned approach will thus provide novel avenues to study cultural and social aspects of multimodal interactions in children's language acquisition computationally in a verifiable manner.

To conclude, we have observed that rural infants are much less exposed to child-directed speech and childdirected co-speech gestures than urban infants, which correlates to their vocabulary development over their second year of life. These findings are in line with predictions based on Keller's (2012) distinction between rural and urban communities. These differences seem to affect vocabulary development as well, but while the results from the urban area are consistent with predictions from western studies (Hart \& Risley, 1995; Pan et al., 2005; Rowe \& GoldinMeadow, 2009), those from the rural community are inconsistent. More fine-grained analyses of the data are undertaken to investigate these differences. In addition, we are currently analyzing data collected from middle class urban families in the Netherlands to carry out a comparative study involving all three prototypical communities proposed by Keller (2012).

\section*{Acknowledgements}

This research was supported with a VIDI grant from the Netherlands Organization for Scientific Research (NWO, grant number 276-70-018). Field research would not have been possible without the assistance of Wona Sanana, Associação Communitária Ambiente da Mafalala, and our Mozambican research assistants. Thanks to Ingrid Masson, Mohamad Hafiz and Van Le for their assistance in coding. Finally, we offer a special thanks to our participant families for their kind hospitality and collaboration.

\section*{References}

Bakeman, R., \& Adamson, L. B. (1984). Coordinating attention to people and objects in mother-infant and peer-infant interaction. Child Development, 12781289.

Bornstein, M. H., \& Putnick, D. L. (2012). Cognitive and socioemotional caregiving in developing countries. Child Development, 83(1), 46-61.
Brown, P. (2011). The Cultural Organization of Attention. In A. Duranti, E. Ochs \& B. Schieffelin (Eds.), The Handbook of Language Socialization: Blackwell Publishing Ltd.
Carpenter, M., Nagell, K., Tomasello, M., Butterworth, G., \& Moore, C. (1998). Social Cognition, Joint Attention, and Communicative Competence from 9 to 15 Months of Age. Monographs of the Society for Research in Child Development, 63(4).
Fenson, L., Pethick, S., Renda, C., Cox, J. L., Dale, P. S., \& Reznick, J. S. (2000). Short-form versions of the

MacArthur communicative development inventories. Applied Psycholinguistics, 21(1), 95116.

Harkness, S. (1977). Aspects of social environment and first language acquisition in rural Africa. Talking to children: Language input and acquisition, 309316.

Hart, B., \& Risley, T. R. (1995). Meaningful differences in the everyday experience of young American children: Paul H Brookes Publishing.
Iverson, J. M., Capirci, O., Longobardi, E., \& Caselli, C. M. (1999). Gesturing in mother-child interactions. Cognitive Development, 14(1), 57-75.
Keller, H. (2012). Autonomy and relatedness revisited: Cultural manifestations of universal human needs. Child Development Perspectives.
Landis, J. R., \& Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 159-174.
LeVine, R. A., LeVine, S., Dixon, S., Richman, A., Leiderman, P. H., Keefer, C. H., \& Brazelton, T. B. (1996). Child care and culture: Lessons from Africa: Cambridge University Press.
Lieven, E. V. M. (1994). Crosslinguistic and crosscultural aspects of language addressed to children. In C. Gallaway \& B. J. Richards (Eds.), Input and interaction in language acquisition. Cambridge: Cambridge University Press.
Mastin, J. D., \& Vogt, P. (2013). Correlations between infant engagement levels and vocabulary development: An observational study of Mozambican infants from \(1 ; 1\) to \(2 ; 1\). Under review.
Matusevych, Y., Alishahi, A., \& Vogt, P. (2013). Automatic generation of naturalistic child-adult interaction data CogSci 2013.
McNeill, D. (1985). So you think gestures are nonverbal. Psychological Review, 92(3), 350-371.
Pan, B. A., Rowe, M. L., Singer, J. D., \& Snow, C. E. (2005). Maternal correlates of growth in toddler vocabulary production in low income families. Child Development, 76(4), 763-782.
Rowe, M. L., \& Goldin-Meadow, S. (2009). Differences in early gesture explain SES disparities in child vocabulary size at school entry. Science, 323(5916), 951.
Schieffelin, B., \& Ochs, E. (1989). Language socialization across cultures. Cambridge, UK.: Cambridge University Press.
Vogt, P., \& Mastin, J. D. (2013). Anchoring social symbol grounding in children's interactions. Künstliche Intelligenz, 27(2), 145-151.
Zukow-Goldring, P. (1996). Sensitive caregiving fosters the comprehension of speech: When gestures speak louder than words. Early development and parenting, 5(4), 195-211.

\title{
Logical Patterns in Individual and General Predication Momme von Sydow (momme.von-sydow@psychologie.uni-heidelberg.de \\ University of Heidelberg, Department of Psychology, Hauptstr. 47-51, D-69117 Heidelberg, Germany
}

\begin{abstract}
Probability judgments about logical propositions have raised substantial doubts about human rationality. Here we explore the idea that people's probability judgments often may not refer to the relative frequency of a set, but instead to the probability of an explanatory logical pattern given the data. This idea has been formalized by Bayesian logic (BL), predicting a system of frequency-based logical inclusion fallacies. The studies presented concentrate on comparing probability judgments about sentences logically relating two attributes of a class or an individual (humans, animals, artifacts). Although BL cannot model probabilities of individual predications directly, it can do so if one assumes that inferences are made about unknown individuals based on imagined samples. The results for general as well as individual predication show a high number of systematic inclusion fallacies in line with BL. Nevertheless, some deviations were found. In the General Discussion, a polycausal approach to inclusion fallacies is advocated. In addition, even if pattern probabilities seem to play a major role, it is suggested that extensions of the BL model may be needed to account for further aspects of real-life predication. Overall, however, even the basic BL model was surprisingly successful for predicting probability judgments about general as well as individual predications.
\end{abstract}

Keywords: Probability judgment; bias; conjunction fallacy; inclusion fallacy; inductive logics; predication.

\section*{Narrow Norms of Predication?}

Throughout Western philosophy (Aristotle, the Stoics, Leibniz, cf. even Kant and Hegel), and particularly since logical positivism (Frege, Wittgenstein, Russell, Whitehead), logic has been central (with slightly different understandings) to defining standards of rational thought. Today, standard calculi of logic and probability may appear narrow in comparison to the much broader Greek concept of logos, but they provide a rigidly defined standard of rational thought. And yet there is much evidence that people's actual reasoning seems to violate these basic calculi. Thus psychology is torn between the Scylla of abandoning normative reasoning (e.g., psychologism) and the Charybdis of claiming that people are fundamentally irrational, even with regard to the simplest rules of these calculi. Although there seems to be some truth in Kahneman and Tversky's (1996) warning against "normative agnosticism", the arguments of Gigerenzer and colleagues (e.g., Gigerenzer, 1996) against them seem reasonable as well: that is, that the blind application of the "narrow norms"' of logic and probabilitytheory often seem misguided. In my view, a domain-specific understanding of rationality may allow for a middle course between these positions. Context-sensitive norms of reasoning that account for our goals as well as the precondition of our models may not need to give up the core of the concept of rationality (cf. von Sydow, 2011).

When the calculi of logic and probability are applied in psychology, standard logic is normally used in deductive, and standard probability theory in inductive contexts. Here we consider both in assessing the inductive probability of logical relationships. Propositional logic addresses the combination of atomic propositions (that can either be true or false) with connectives (AND, OR, EITHER OR, NEITHER NOR, etc.). In the tasks we investigate probability judgments involving several alternative logical sentences, with different logical connectives relating two properties. We are either concerned with the properties of an entity or of a class of entities (individual vs. general predication).

The suggested domain-specific approach to rationality should consider the context and the goals implied. The context of our probability-judgment task is the assignment of attributes to a class. What is a reasonable, observationbased norm for predicating specific logical relationships between attributes, and how does this relate to probabilities (von Sydow, 2011)? At first sight, propositional logic seems a plausible candidate. A sentence such as "ravens are black and they can fly" logically seems to predicate the conjunction of attributes \((B \wedge F)\) to the class of ravens \((R)\). From a falsificationist perspective, this predication is valid as long as no single exception defies the rule. Predications about contingencies in the actual world (in contrast to mathematics) would all be rendered false, since one may assume that they are not free of exceptions. For instance, albino ravens exist, as well as other non-black ravens. It therefore seems reasonable to replace a purely logical adequate criterion of predication by a high-probability criterion (cf. Schurz, 2005). In the raven example, correct predication would require that \(P(B \wedge F \mid R)>\psi\), with \(\psi\) being the high-probability criterion. This proposal additionally appears to solve the problem of non-monotonicity, since now an adequate predication may become inadequate (and vice-versa) during further data-sampling. Nonetheless further problems remain.

Here only the problem of set-inclusion is sketched (cf. von Sydow, 2011, von Sydow \& Fiedler, 2012). The frequentist/extensional probability of the predication "ravens are black and they can fly" can never be larger than the probability of the inclusive disjunction \(P\) ("ravens are black or they can fly or both") \((P(B \vee F \mid R)\) ), since the former refers to a subset of the latter. Likewise, the AND sentence cannot have a larger probability than the tautology ( \(P\) (all feature-combinations are possible)). Using an extensional probability-criterion excludes preferring the predication of a more specific hypothesis over a (less informative) more general one. The tautology \((P(B T F \mid R)=\) 1) would always be a rational predication, even independent of data. Therefore, extensional probabilities could not be reasonable evidence-based criteria for adequate predication.

\section*{Probabilities of Noisy-Logical Patterns}

One way to resolve this problem and the problem exceptions together is to assume that people tend to judge the probability of alternative explanatory logical patterns instead of the relative size of particular sets, when concerned with probabilities of alternative logical predications, each meant as an explanation of the whole situation. A first formalization of this idea has been provided by von Sydow (2011, cf. von Sydow \& Fiedler, 2012). Here only the idea of the model, called Bayesian Pattern Logic (Bayesian Logic, or BL), is sketched, without providing a formal model. In the wake of the renaissance of Bayesian models (cf. Chater, Tenenbaum, Yuille, 2006; Kruschke, 2008; Oaksford \& Chater, 2007) it is formulated as a Bayesian approach. It formalizes the idea of explanatory logical patterns (an AND-pattern, an EITHER-OR-pattern, etc.), under absence of further factors. The model provides the measure of fit between a \(2 \times 2\) frequency table input and \(2 \times 2\) probability tables that may hypothetically have produced the data (hypothetical noisy-logical explanations). The probability tables are based on logical truth tables assuming equiprobability of true cases (cf. Johnson-Laird et al., 1999; Tenenbaum \& Griffith, 2001) and a uniform noise function. Based on these basic assumptions, the model first establishes the likelihood that some observed data have been produced by the probability tables, \(P(D \mid P T)\). To obtain the posteriors, the probabilities of these hypothetical noisylogical explanations given the data \((P(P T \mid D)\) ), one uses the Bayes theorem. To obtain the probability of a connective, one sums up the corresponding posteriors over all noise levels (for technical details, see von Sydow, 2011; cf. von Sydow, 2009, von Sydow \& Fiedler, 2012).

In sum, the extensional probability of a set (relative frequency) is here replaced by the second-order probability of noisy-logical patterns of probabilities (all four cells of a PT add up to 1). These patterns serve as hypothetical logical explanations. It is predicted that people use pattern probabilities to explain a whole situation in logical terms (class \(X\) is \(A\) and \(B\) ), instead of judging the size of a set or subset. Accordingly, \(P\) (ravens are black and they can fly) should be high, not because there are few exceptions but because our subjective frequency pattern best fits a noisy AND-pattern. If one is concerned with pattern probabilities, the probability that a data-set may be produced by an ANDpattern may well be higher than that for an OR-pattern: \(P_{P}(B \wedge F \mid R)>P_{P}(B \vee F \mid R)\). By contrast, a narrow application of extensional probability always requires that \(P_{E}(B \wedge F \mid R) \leq\) \(P_{E}(B v F \mid R)\) (cf. von Sydow, 2011).

Previous work in the conjunction-fallacy debate generally concerned a quite different, story-based task, showing that people may judge the conjunction more probable than the conjunct, e.g., \(P(B \& F)>P(B)\) (Tversky \& Kahneman, 1983). In a few cases, CFs were also shown without stories (e.g., Lagnado \& Shanks, 2002). In any case, most authors have assumed that such conjunction-judgments involve a "conjunction fallacy" (CF). Conversely, BL suggests a
rational explanation at least of a particular class of CFs [for convenience they are nonetheless called "fallacies" here].

The application of BL led to several new predictions and corroborative findings-for instance, on double CFs, sample-size effects, and pattern-sensitivity effects (von Sydow, 2011). The concept of CFs has been generalized to apply to system of logical connectives based on summary information (von Sydow, 2009) or sequential input (von Sydow, 2012). Whether or not other theories may account for independent causes of CFs (e.g., Lagnado \& Shanks, 2002; Tentori, Crupi, Russo, 2012), these results could not be explained by any other current theory. It seems plausible, then to conclude the existence of a class of pattern-based CFs. Additional factors-for instance, unclear set-inclusions (Sloman, Over, Slovak, \& Stibel, 2003), illicit implicatures (Hilton, 1995; cf. Hertwig et al. 2008), and probability format (Fielder, 1988)—remain plausible further facilitators for CFs, even if one is concerned with extensional probability judgments. Nevertheless, a high proportion of CFs were found even when simultaneously using clear formulations, clear set-inclusions, rating scales, and frequency information (von Sydow, 2011).

\section*{Individual vs. General Predications Based on Real-Life Frequencies}

The investigations reported here address three issues.
(1) Previous tests of BL used explicit frequency inputs, either in a table format (von Sydow, 2011) or in an experienced sequential learning format (von Sydow \& Fiedler, 2012). Although this allowed for precise tests of plausible models, it may differ from real-life predication where samples often have to be retrieved from memory. Moreover, the explicit frequencies presented in other tasks might have suggested the use of something like BL. We therefore assess here subjective frequencies of real-life predication independently from the task where participants judged probabilities of different logical sentences. Whereas previous tests focused on the variation of frequencies and only used a small number of scenarios, in order to reduce the influence of uncontrolled priors or other disruptive factors, we here used several different scenarios involving people, animals and artifacts.
(2) Despite previous success in modeling frequency-based prediction, it is an open issue whether the pattern idea is applicable to individual predications as well. BL cannot be applied to individual predication without an auxiliary hypothesis. The formal model has a frequency-based in-put-the four cells of a contingency matrix, ( \(f(B \& F)\); \(f(B \& \neg F) ; \mathrm{f}(\neg B \& F)\); \(\mathrm{f}(\neg B \& \neg F))\). Although some frequentists have been skeptical about probability judgments in individual cases, it seems plausible that humans often base their probability estimates, even for individual cases, on imagined subjective frequencies. The explored auxiliary hypothesis is that for individual predications (concerning e.g., a raven), one may - in the absence of further information simply imagine a hypothetical sample of ravens. This may be used as input for BL (suggested by von Sydow, 2011).

\section*{Overview}

In Preliminary Study 1 we first sampled sentences by asking participants for sentences that related two attributes logically. In Preliminary Study 2, participants provided estimates for the frequencies entered in a contingency table relating these attributes. Then in the main study we investigated general vs. individual predication and assessed in 30 scenarios which logical connective relating two attributes was judged to be most probable (an extended CF task with several connectives). We then modelled the predictions of BL based on the subjective frequencies from Preliminary Study 2. Finally, we compared the model predictions with the results for general and individual predication.

\section*{Preliminary Study 1}

Participants (twelve students from the University of Göttingen) had to fill in the blanks for 6 sentences, each concerning a different logical relationship between two attributes. For each sentence one filled in a class and two attributes, such as "Normally [In der Regel] \(\qquad\) are either
\(\qquad\) or \(\qquad\) ". The order of the six connectives was permutated, and an example was provided: "Normally chairs have four legs AND (at the same time) they have a seat." The predications consistently employed either the verb "to be" or "to have".

After obtaining the results, we narrowed down the number of sentences for Study 2. We excluded arbitrary sentences, obviously deterministic sentences, and sentences that seemed to contain an unwanted or over-complex causal background. We aimed to focus on the relation between two attributes of a class whose co-occurrence could be described in simple logical terms (e.g., ravens are black and can fly). In addition, we supplied four more sentences.

\section*{Preliminary Study 2}

Method Participants (23 students from the University of Göttingen, 78 \% female) provided subjective frequency judgments for the co-occurrence of two attributes in a \(2 \times 2\) contingency table for each of the 50 scenarios investigated. For each randomized scenario, participants assigned a sample of 100 hypothetical cases to the four cells of a table ( \(a, b, c, d\); see Figure 1).
Results Figure 2 shows four examples for the 50 resulting

\section*{Professional Basketball Players}


How certain are you that your estimated frequency distribution is roughly valid? \(\begin{array}{ccccccccccc}\mathrm{O}-\mathrm{O}-\mathrm{O}-\mathrm{O}-\mathrm{O} & \mathrm{O}-\mathrm{O}-\mathrm{O}-\mathrm{O}-\mathrm{O}-\mathrm{O}-\mathrm{O}-\mathrm{O}-\mathrm{O}-\mathrm{O}-\mathrm{O}-\mathrm{O}-\mathrm{O}-\mathrm{O}-\mathrm{O} \\ 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10\end{array}\) Figure 1: Assessment of frequency estimates in Preliminary Study 2.
four-cell frequency distributions. The results were later used as input for BL to predict the probability judgments in the main experiment. Based on this study the scenarios were selected so as to have four scenarios for each of the six focused connectives. Two scenarios predicted the main connective with the highest relative frequency of participants (even if the pattern probability was below 50\%). The professional-basket-ball-player scenario (Fig. 1a) is an example for an AND-connective. \(P\) (tall AND also quick) is expected predominantly to be estimated higher than the probabilities of larger sets (despite exceptions). For two further connectives, the second most frequent hypothesis was predicted almost as often as the first. The application of the schema worked quite well, apart from the OR-class, where all scenarios reflected at best the second noise level. Finally, we investigated six 'noise-scenarios', where the predictions of BL become less clear, favoring even more than two connectives (generally to at least three).

\section*{Main Experiment}

Method The experiment had a 2 (general predication vs. individual predication, between subjects) \(\times 30\) (scenarios, within subjects) design. 20 participants judged for each scenario which of 15 logical sentences connecting two target attributes is most probably valid (extended CF task). The 30 scenarios were presented in random order and concerned people, animals and artifacts.


Figure 2: Boxplots depicting the distribution of the estimated frequencies in the four cells of the contingency matric (and of the confidence ratings) for four example scenarios (Preliminary Study 2; Median; 25\%-75\% boxes)

The instruction for both general and individual predication-conditions followed the same pattern, e.g., "Imagine [hundred / one] professional basketball player[s]. We are concerned with several propositions about [a] professional basketball player[s]. [...] Please tick the proposition that in this situation seems most probable to you." Participants should answer intuitively. In each of the scenarios participants selected the 1 out of 15 that seemed most probable. For the general predication condition, propositions opened with the class (e.g., "Professional basketball players are..."), and in the individual predication condition, with the individual (e.g., "A professional basketball player is..."). The 15 hypotheses always occurred in the same order, referred to all 16 dyadic logical connectives apart from the falsum/contradiction. For instance: A AND B (H1); A AND not-B (H2); NEITHER A NOR B (H4); A (H5); EITHER A OR B (H9); A OR B OR BOTH (H11), and everything is possible (Verum/Tautology, H15).
CFs \((P(A \wedge B)>P(A))\) may be due to reinterpretation of the logical connectives according to standard conversational implicatures (e.g., Hilton, 1995; Hertwig, Benz \& Krauss, 2008). If the affirmation \(A\) is contrasted with " \(A\) AND \(B\) " it may indeed reasonably be represented as " \(A \wedge\) non- \(B\) ". To avoid such misunderstandings, we in all studies used the formulation " \(X\) are \(A\) (and they are \(B\) or not-B)". Likewise, " \(A\) AND \(B\) " in ordinary language may well refer to " \(A\) OR \(B\) (or both)." In this interpretation, \(P(A \vee B)>P(A)\) is not fallacious. We used an OR-hypothesis and the following AND-formulation: " \(X\) are \(A\) (e.g., taller than 1.8 m ) and at the same time \(B\) (e.g., quick)." The verum read: " \(X\) are tall and quick, tall and slow, short and quick, or short and slow (all combinations)."

Forty participants (from the same population) volunteered to take part, receiving either course-credit or a fee.

Modeling For all scenarios we calculated the predictions of the model based on Study 2. For each participant and scenario we used the estimates for the four cells of the contingency table as input for BL, determining which hypothesis this participant would select as most probable. For reasons of simplicity we ignored further rankings. Calculated for all participants, this provided a reasonable prediction for the distribution of selections in the main task.

Results Figure 3 shows the predictions of the model as well as the accumulated results for the six types of scenario (referring to different dominant connective). Each chart represents four scenarios. Although grouped this way by


\[
\begin{gathered}
\text { Hypotheses } \\
\mathrm{r}_{\mathrm{B} 1-\mathrm{B} 2}=.85, \mathrm{r}_{\mathrm{B} 1-\mathrm{M}}=.78, \mathrm{r}_{\mathrm{B} 2-\mathrm{M}}=.78 \quad \mathrm{r}_{\mathrm{B} 1-\mathrm{B} 2}=.71, \mathrm{r}_{\mathrm{B} 1-\mathrm{M}}=.67, \mathrm{r}_{\mathrm{B} 2-\mathrm{M}}=.89
\end{gathered}
\]

\[
r_{\mathrm{B} 1-\mathrm{B} 2}=.90 ; \mathrm{r}_{\mathrm{B} 1-\mathrm{M}}=.26 ; \mathrm{r}_{\mathrm{B} 2-\mathrm{M}}=.30
\]




Figure 3: For all six focused types of connectives, the graphs provide a visualization of the results for general and specific predications (percentage of selections) and of the model-predication (averaged over scenarios). On the ordinate, the proportion of actually selected or predicted hypotheses (H1 to H15) is shown (cf. text for details).
design, based on Study 2, the scenarios should differ somewhat (low noise vs. high noise). Figure 3 therefore provides merely a simplifying visualization. Nonetheless it does depict the main pattern concisely and well.
In the H1 scenarios (involving, e.g., the basketball-players and ravens scenarios) the AND-connective was the most frequently selected. Such judgments involve estimating \(P(A\) AND \(B\) ) to be more probable than \(P(A), P\) (Either both attributes or none), \(P(A\) OR \(B\) or both), and \(P\) (Tautology). Given the presence of exceptions (e.g., Figure 2a), this would traditionally be interpreted to involve several simultaneous logical inclusion fallacies (von Sydow, 2009). Moreover, it appears that on this level the overall distribution of selections reflect the predictions quite closely. The most striking deviation in this and the other scenarios, however, was that H15 (the tautology), the extensionally correct solution, was selected more frequently than predicted (cf. Fig. 2).
In the H2 scenarios, participants predominantly selected the predicted sentences " \(X\) are \(A\) and not \(B\)," likewise involving several inclusion fallacies, as most probable.
The H11 scenarios yielded the strongest deviations from the predictions (to be discussed below).
For the other scenarios (H4, NEITHER NOR; H5, Affirmation A; H6, EITHER OR), the results corroborated both the predicted dominant selections and an overall high correspondence between BL and the data.
Figure 3 additionally provides strong evidence for a high similarity between results in the two conditions, the individual and general prediction tasks.

These overall results need supplementation from further measures, in order to assess results on the level of the single scenarios. Accumulation of the results of four scenarios of a type, such as in Figure 3, increases the \(N\) (reducing chancefindings) and excludes confounding factors specific to single tasks; but such results will tend to yield too positive a picture.

Table 1: Mean correlations between model-predictions and results, as well as between the two kinds of predication (individual and general) for the single scenarios in the six types H 1 to H 9 ) and a further noise class
\begin{tabular}{cccccccc}
\hline & H1 & H2 & H11 & H4 & H5 & H9 & noise \\
\hline \(\mathrm{r}_{\mathrm{B} 1 \mathrm{~B} 2}\) & .85 & .71 & .79 & .89 & .92 & .82 & .75 \\
\(\mathrm{r}_{\mathrm{B} 1 \mathrm{M}}\) & .78 & .67 & .41 & .63 & .80 & .72 & .24 \\
\(\mathrm{r}_{\mathrm{B} 2 \mathrm{M}}\) & .78 & .89 & .43 & .73 & .91 & .83 & .34 \\
\hline
\end{tabular}

Table 1, despite using averages, for this reason focuses on correlations for the single scenarios in a class. Please note, this differs from correlations on the accumulated level (which actually yield correlations .09 higher on average). Table 1 shows correlations between model and results, as well as between individual and general predicationconditions. The average correlations were all positive and generally large (or very large).

Only for H 11 the average correlations with the model were only moderately positive. This was likewise the case for the high-noise scenarios (where predictions did not favor specific connectives). In these two classes, the number of tautology-selections was higher than expected (H15). The two deviations may be explained along the same lines: as mentioned, the OR-scenarios, just as the noise-scenarios, had much less clear predictions than all other scenarioclasses. The second- and thirdmost frequently predicted hypotheses were not much less frequently predicted than the OR hypotheses themselves. Moreover, even for ORpredictions (based on specific participants of Pre-test 2), the second-highest pattern-probability, did not generally differ substantially from the second-highest. Such uncertainties in both may have led to the selections of H15, which suggested that everything is possible.

Even if the 30 scenarios were analysed individually (which cannot be done here) the overall pattern would remain similar. All 90 calculated correlations were positive, and only \(14 \%\) yielded \(r<40\) (particularly in the mentioned classes). The examples from Figure 2 with dominant AND, EITHER OR, A and OR predictions, for instance, corroborated the predicted dominant selections and had a high model-fit. Nonetheless, a low number of correlations did not show the overall positive results (even outside of the two mentioned classes), with values close to 0 and asymmetrical findings for both specific and general predications.

\section*{General Discussion}

The findings corroborate that people do not judge probabilities extensionally, but instead allow for exceptions. Participants systematically committed a large number of
inclusion fallacies (generalizing CFs, cf. von Sydow, 2009; von Sydow \& Fiedler, 2012). Pattern probabilities, as formalized by BL, were shown to be quite successful in modeling the probability judgments in a multitude of scenarios only indirectly based on frequency estimates in Preliminary Study 2. Other models of CF have not yet been explicitly designed to test for these connectives, but it is as yet highly implausible that some adaptation of these models (e.g., confirmation, inverse probability, representativeness, averaging, quantum logic, support theory, rescaling, signed summation, etc.) will easily account for these data equally well without adopting the very idea of pattern probabilities themselves (cf. von Sydow, 2009, 2011, for more details).

Moreover, there was a large similarity between probability judgments about general predication and prediction about singular subjects. Combining BL with the auxiliary hypothesis that people may base judgments about singular nouns on hypothetical sampling led to quite successful predictions. However, it needs to be mentioned that the current finding might be limited to a generic interpretation of the singular subjects (e.g., "a professional basketball player (PBP) is/has"), although the introduction suggested an individual reading ("imagine one PBP"; "propositions about a PBP"). Further research is needed to investigate the role of different formulations in more detail.

Finally, some quantitative deviations from the predictions were found-some in a quite explainable manner regarding two classes of scenarios- but unsystematic deviations for single scenarios were found as well. Although this did not substantially alter the main findings, it may suggest that further factors are at work. To state it more emphatically, in my view, it would be a surprise if there were in fact no additional factors:
(1) Despite favoring BL as important account for CFs, I think there may be several other causes of CFs as well. I have mentioned other theories above. An example is that people in some contexts may reasonably be interested in the increase of probabilities (confirmation) instead of probabilities themselves (e.g., Lagnado \& Shanks, 2002; Tentori, et al., 2012); and there may well be situations where people are interested in a synthesis of confirmation and the pattern idea: the degree of confirmation of different logical patterns.
(2) Even if focusing on pattern probabilities the current formalization of BL may only be one sub-class of modeling pattern probabilities in real-world predications. The current formalization is concerned with dyadic classes and assumes an equally weighted \(2 \times 2\) input. Considerations to be examined include: (a) Although dyadic dichotomous logic as well as human language is often concerned with dichotomous (or dichotomized) categories, the implicit number of relevant categories can vary and may well matter. (b) The dichotomous classes need not refer to categorical classes, but can point to an underlying ordinal, interval, or rational scale. This may require a modified pattern approach that weights extreme cases more heavily. A domain specific approach to rationality that takes preconditions of models seriously should be sensitive to such aspects. (c) The
categories and resulting frequency estimates may be defined in either an absolute (larger than 1.80 ) or a relative way (large). If the latter, BL's input may needs to be modelled depending on the context. (d) Within the present model, different contexts (scenarios) may lead to different noisepriors (reflecting the learned tolerance for exceptions for different scenario types), whereas here always flat noiseprior was used (cf. von Sydow, 2011). (e) People might assign important properties more weight than unimportant properties.
(3) The relationship between individual and general predication is presumably more intricate than assumed in the studies. As we have seen, BL is designed for general predication with a frequency-input based on a \(2 \times 2\) contingency matrix. This assumes that evidence is ordered as cases to be assigned to one of the table's four cells. The results presented here support the idea that one can model individual probabilistic predication along the same lines by imagining 100 individuals and finding the highest pattern probability for statements as "A sophomore either owns a flat or shares a flat" (Fig. 2b). This seems unproblematic, since individual sophomores still fall into one of the four classes. The EITHER-OR here only expresses a lack of knowledge about which of two classes the individual is actually fits. Nevertheless, sentences such as "this ape from species \(X\) is either aggressive ( \(A\) ) or curious ( \(C\) )" need not indicate lack of information, but rather an alternative meaning: that is, the individual ape may have been \(A\) (without being \(C\) ) and at other times the reverse. Notably, the input-assumption would still hold for individual predication on a sub-individual event level, but on the group level this apparently positive extension of BL to individual predication now seems problematic; for, if an individual is "either \(A\) or \(C\) " it no longer fits any of the four classes ( \(A\) \& \(C, A \&\) non- \(C\), non- \(A \& C\), or non-A \& non-C). This problem may be solved by adding \(1 / 2\) to both relevant cells, resulting at least in similar predications for both levels. Nonetheless, the issue remains problematic if we are concerned with heterogeneous groups of individuals, where in most cases \(X\) are either " \(A\) or \(C\) or both" or " \(A\) and \(C\) " (each based on sub-individual frequency information). The inclusive predicate " \(A\) or \(C\) or both" alone may be inappropriate. Participants may be interested in a pattern of patterns ( \(X\) are \((A \wedge C)><(A \vee C))\); and interpreted as pattern, this does not needs to be equivalent to \(A><C\), as actually valid in propositional logic. Such a pattern-of-pattern interpretation would not only be an interesting field of future research, but it might discourage the selection predicated by standard BL which assumes the absence of sub-classes.

In summary, real-life predication as well as probability judgments about these logical predications may plausibly be affected by a variety of additional factors, either external or ones calling for other more context-sensitive formalizations of pattern probabilities. In the light of such suggestions, the basic BL model was shown here to be surprisingly successful in accounting for a great variety of probability judgments about general as well as individual predications.

\section*{Acknowledgments}

Aspects of this research, presented here in written form have previously been introduced in talks. I am grateful for the inspiring discussions at the Universities of Göttingen (M. R. Waldmann group) and Heidelberg (K. Fiedler group). I am also grateful for help from Martha Cunningham, Johanna Frisch and Dennis Hebbelmann and for the feedback of the anonymous reviewers. This work was supported by the grant Sy 111/2-1 from the Deutsche Forschungsgemeinschaft (DFG) as part of the priority program New Frameworks of Rationality (SPP 1516).

\section*{References}

Chater, N., Tenenbaum, J., \& Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. Trends in Cognitive Science, 10, 287-291.
Fiedler, K. (1988). The dependence of the conjunction fallacy on subtle linguistic factors. Psychological Research, 50, 123-129.
Gigerenzer, G. (1996). On narrow norms and vague heuristics. Psychological Review, 103, 592-596.
Hertwig, R., Benz, B., \& Krauss, B. S. (2008). The conjunction fallacy and the many meanings of and. Cognition, 108, 740-753.
Hilton, D. J. (1995). The social context of reasoning: Conversational inference and rational judgment, Psych. Bulletin, 118, 248-271.
Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M. S. \& Caverni, J.-P. (1999). Naive probability: A mental model theory of extensional reasoning. Psychological Review, 106, 62-88.
Kahneman, \& Tversky (1996). On the Reality of Cognitive Illusions. Psychological Review, 103, 582-591.
Kruschke, J. K. (2008). Bayesian approaches to associative learning: From passive to active learning. Learning \& Behavior, 36, 210-226.
Lagnado, D. A. \& Shanks, D. R. (2002). Probability judgment in hierarchical learning: A conflict between predictiveness and coherence. Cognition, 93, 81-112.
Oaksford, M., \& Chater, N. (2007). Bayesian rationality. The probabilistic approach to human reasoning. Oxford Univ. Press.
Schurz, G. (2005). Non-monotonic reasoning from an evolutionary viewpoint: ontic, logical and cognitive foundations. Synthese, 146, 37-51.
Sloman, S. A., Over, D., Slovak, L., \& Stibel, J. M. (2003). Frequency illusions. Organizational Behavior and Human Processes, 91, 296-309.
Tentori, K., Crupi, V., \& Russo, S. (2012). Determinants of the conjunction fallacy: Confirmation versus probability. Journal of Experimental Psychology: General.
Tenenbaum, J. \& Griffith, Th. (2001). Generalization, similarity, and Bayesian inference. Behavioral and Brain Sciences, 24 (4), 629-640.
Tversky, A., \& Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. Psychological Review, 90, 293-315.
von Sydow, M. (2009). On a General Bayesian Pattern Logic of Frequency-Based Logical Inclusion Fallacies. Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society (pp. 248-253). Austin, TX: Cognitive Science Society.
von Sydow, M. (2011). The Bayesian Logic of Frequency-Based Conjunction Fallacies. Journal of Mathematical Psychology, 55(2), 119-139.
von Sydow, M. \& Fiedler, K. (2012). Bayesian Logic and Trial-bytrial Learning. Proceedings of the Thirty-Fourth Annual Conference of the Cognitive Science Society (pp. 1090-1095). Austin, TX: Cognitive Science Society.

\title{
The role of sampling assumptions in generalization with multiple categories
}

\author{
Wai Keen Vong (waikeen.vong@adelaide.edu.au) \\ School of Psychology, University of Adelaide, SA 5005, Australia
}

\author{
Andrew T. Hendrickson (drew.hendrickson@adelaide.edu.au)
}

School of Psychology, University of Adelaide, SA 5005, Australia
Amy Perfors (amy.perfors@adelaide.edu.au)
School of Psychology, University of Adelaide, SA 5005, Australia
Daniel J. Navarro (daniel.navarro@adelaide.edu.au)
School of Psychology, University of Adelaide, SA 5005, Australia

\begin{abstract}
The extent to which people learning categories generalize on the basis of observed instances should depend in part on their beliefs about how the instances were sampled from the world. Bayesian models of sampling have been successful in predicting the counter-intuitive finding that under certain situations generalization can decrease as more instances of a category are encountered. This has only been shown in tasks were instances are all from the same category, but contrasts with the predictions from most standard models of categorization (such as the Generalized Context Model) that predict when multiple categories exist, people are more likely to generalize to categories that have more instances when distances between categories is controlled. In this current work we show that in both one- and two-category scenarios, people adjust their generalization behavior based on cover story and number of instances. These patterns of generalization at an individual level for both one- and two-category scenarios were well accounted for by a Bayesian model that relies on a mixture of sampling assumptions.
\end{abstract}

Keywords: sampling assumptions, generalization, category learning

\section*{Introduction}

The ability to generalize beyond existing data is a basic cognitive capacity that underlies a great deal of human learning, categorization and decision-making (e.g. Shepard, 1987; Tenenbaum \& Griffiths, 2001; Nosofsky, 1986). To complete the inductive leap needed for generalization, people must make some kinds of assumptions about how that data was generated or sampled. A learner's sampling assumptions influence the evidentiary value of the data, and thus alters what they should infer based on it.

One natural assumption is that each observed datum has been selected independently and then labeled as a member (or not) of the category or concept to be learned. An example of this is a parent who tries to teach a child what a "ball" is by randomly picking from all of the toys her room then labelling them as balls or not. This possibility, called weak sampling, implies that all observations \(x\) are equally likely, regardless of what hypothesis \(h\) the learner has about the category. Mathematically, this corresponds to the notion that \(P(x \mid h) \propto 1\).

A different type of data generation, known as strong sampling, presumes that the data has been selected as a random positive example directly from the category to be learned
(Tenenbaum \& Griffiths, 2001). A parent who teaches the word "ball" by showing the child many different kinds of balls (but not other toys) is strongly sampling from the category of BALL. The key consequence of strong sampling is that it licenses tighter generalizations with increasing data. This is because each datapoint is more informative about the boundaries of the category. Mathematically, for a hypothesis \(h\) that consists of \(|h|\) possible category members, the strong sampling model implies that \(P(x \mid h)=1 /|h|\) if the observation \(x\) falls within the category, and 0 if it does not.

There is evidence that people are sensitive to sampling assumptions, making tighter generalizations when the data appear to have been strongly sampled (e.g. Xu \& Tenenbaum, 2007). However, a number of questions remain unresolved, two of which we address in this paper.

The first, more minor, issue relates to the influence of the cover story. As mentioned, work by Xu and Tenenbaum (2007) suggests that both adults and children change their generalization patterns in response to differences in sampling. This appears to contrast with other work by Navarro, Dry, and Lee (2012), which found that although sampling assumptions varied between individual participants (with some assuming strong sampling and others weak), people did not change their behavior according to the cover story they were presented with. One way to resolve the discrepancy between these two studies is to conclude that the data generation process was much more obvious in Xu and Tenenbaum (2007). In that study, participants actually saw instances selected in front of them, whereas in Navarro et al. (2012) participants simply read different cover stories. Here we explore whether it is necessary for people to see data being generated in order to change their sampling assumptions, or whether a more salient cover story manipulation would be sufficient.

The second issue we investigate is a more important and more puzzling one. It is generally acknowledged that inductive generalization is very closely linked to categorization. For instance, exemplar models of categorization (e.g., Nosofsky, 1986) are constructed by assuming that the learner uses a simple probabilistic model to generalize from each stored exemplar to a target item. The "narrowness" of the generalizations is a fixed parameter (referred to as the specificity)


Figure 1: Different predictions made by a standard categorization model (the GCM) and a Bayesian model that incorporates a strong sampling assumption. The GCM, on the left panel, predicts that as the number of instances in a category increases (shown in the figure by the additional points in the bottom row), generalizations should loosen slightly: the solid line corresponding to generalizations based on later additional instances in Category A extends further from Category A. By contrast, the model incorporating strong sampling predicts that generalization based on additional instances will tighten: the solid line in the right panel is much closer to Category A than the dotted line corresponding to earlier, fewer instances.
and does not change as the sample size increases. This is effectively a weak sampling assumption, and it is assumed by both the basic Generalized Context Model (Nosofsky, 1986) and by models such as ALCOVE (Kruschke, 1992) that extend it. These models have proven highly successful at describing human classification behavior, apparently with little need to adapt them to incorporate some version of the strong sampling assumption. If human learners are as sensitive to sampling assumptions as papers such as Xu and Tenenbaum (2007) imply, why has it not been necessary to incorporate such assumptions into existing categorization models?

We can think of at least two possible (not mutually exclusive) explanations for this. The first one is that sampling effects have not been found simply because few studies have gone looking for them. Standard supervised classification designs do not manipulate the sampling assumptions, and it could be argued that the instructions and design of such experiments often imply weak sampling. As such, it is natural to expect that the theories used to explain these experiments would implicitly rely on weak sampling assumptions. A similar suggestion is made by Hsu and Griffiths (2010).

An alternate possibility is that these divergent results arise because of a genuine difference in the nature of the experiments: the number of categories involved. Typical categorization experiments generally involve two categories, with stimuli needing to be classified as belonging to one or the other (Nosofsky, 1986). In contrast, researchers testing sampling assumptions have tended to use tasks in which participants are asked to draw inferences about only a single target category (Xu \& Tenenbaum, 2007; Navarro et al., 2012).

In this paper we test the latter possibility by making use of the fact that strong sampling models make a different prediction from standard categorization models in certain situations. A multiple-category version of a Bayesian generalization model with strong sampling \({ }^{1}\) predicts that if we increase the number of instances in Category A without changing the number of exemplars in the other category, items that lie in between the two categories should decrease in their probability of being classified as members of Category A. This is because strong sampling leads to tighter generalization of Category A with more instances, (right panel of Figure 1). Note further that this is the opposite of what one would expect from a standard exemplar model: adding more exemplars to Category A but not to the other category can only increase the summed similarity between Category A exemplars and a target item. As a consequence, items that lie between the two categories should increase in their probability of being classified as members of Category A (left panel of Figure 1).

These distinct predictions motivate our experiment: we present learners with either one-category or two-category generalization problems, presented either in the context of a strong or weak sampling cover story. We predict that when in the context of strong sampling, people will modulate

\footnotetext{
\({ }^{1}\) The two-category Bayesian strong sampling model is a minor modification of the one-category strong sampling model described by Tenenbaum and Griffiths (2001). The only difference between that model and the current one lies in how the hypotheses about the extension of a category (the"consequential regions") are defined. In the two-category model the stimulus space is divided into two mutually exclusive regions, one for each category. As per the original model, category items are assumed to be sampled uniformly at random from the region of that category.
}


Figure 2: The experimental design. The top panel refers to the three sets of stimuli used across each block in the one-category task. The bottom three panels refer to the three possible sets of stimuli used in the different base rate conditions for the twocategory task (results from the bottom two sets are collapsed into one UNEQUAL BASE RATE condition for the purposes of analysis). All participants performed the one-category task as well as one of the three two-category tasks. The ticks at the bottom of each panel show the location of each of the test points for each condition.
their generalization based on cover stories such that they will tighten their generalization of a category label in response to observing additional exemplars in that category that do not extend the category boundary.

\section*{Method}

Participants Data was collected from 318 participants from Amazon Mechanical Turk. No demographic information was collected so participants remained anonymous. Participants were paid \(\$ 0.50 \mathrm{USD}\) for their participation to complete the task which lasted approximately 15 minutes.
Procedure Each participant performed a one-category and a two-category generalization task in random order following a scenario adapted from Navarro et al. (2012). In the onecategory task, participants observed instances of temperatures at which one type of bacteria was found alive in food. They were then asked to estimate the probability that the same bacteria would be found alive in the food at other temperatures. In the two-category task, participants observed instances of
temperatures where two types of bacteria were found alive in food. They were also told that the two types of bacteria competed for resources, so only one type of bacteria could be found alive in the food at any given temperature. As in the one-category task, participants were asked to estimate the probability that one of the two types of bacteria would be found alive in the food at other temperatures.

The experiment also contained two between-subjects manipulations. The first was a sampling assumption manipulation in which participants were presented with different cover stories in order to influence their beliefs about the sampling process. In the STRONG SAMPLING condition, participants were told that the instances were selected by scientists who had identified a number of temperatures where bacteria were found alive in food. This cover story suggested to the participant that the scientists were only selecting positive examples from the category, consistent with strong sampling. Conversely, in the WEAK SAMPLING condition, participants were told that the instances were the result of an automated pro-


Figure 3: Mean generalization probabilities in the onecategory task across sampling condition and block. Generalization is tighter in the STRONG SAMPLING conditions but does not differ by block.
cess that tested the bacteria at different temperatures. This suggested to the participant that the presented instances were chosen at random from the range of all possible temperatures, consistent with weak sampling. People who were in a given sampling condition received the same (strong or weak) sampling cover story for both the one- and two-category tasks.

The other between-subject experimental manipulation varied the base rate in the two-category generalization task. In the EQUAL BASE RATE condition, the number of instances observed in both categories was the same. There were also two conditions in which one category contained more instances: one in which the left category had more and one in which the right category had more. Because there were no differences between these two conditions, all analyses collapsed them into one UNEQUAL BASE RATE condition. The different conditions are illustrated in Figure 2.

In both the one-category and two-category tasks, the instances were presented across three blocks. In the onecategory task, participants initially saw three green dots representing temperatures where bacteria was found alive in the food. They were then asked to estimate (using a slider) the probability that the bacteria would be found alive at each of 22 temperatures in sequence. As a measure of whether participants were performing the task correctly, two of the 22 test trials were located inside the range of observed instances. After making the 22 judgments, participants were then presented with two more instances and asked to make the same judgments again. In the final block, they were presented with one more instance before repeating the 22 estimates again. Overall, each participant made 66 judgments ( 3 blocks \(\times 22\) queries) in the one-category task.

The procedure in the two-category task was very similar, except that participants were presented with instances representing the temperatures where bacteria from both the left and right categories were found alive (shown as blue and red dots respectively of Figure 2). Participants were asked to estimate the probability that the blue bacteria (the left category) would be found alive in the food at each of 11 temperatures. All


Figure 4: Mean generalization probabilities in the twocategory task across sampling condition, base rate condition, and block. Generalization is tighter in the STRONG SAMPLING condition regardless of category base rate.
of the test points in the two-category task were between the two categories, except for one within the range of instances for each of the categories. As in the one-category case, participants were given additional instances at the beginning of each block and then asked to make judgments at each of the test points. This resulted in a total of 33 judgments ( 3 blocks \(\times 11\) queries).

\section*{Results}

Participants who failed to understand the task (based on their performance on the within-category test points) were excluded from the analyses. We reasoned that people who correctly understood the experimental task would have assigned probabilities close to \(100 \%\) for the test points within the categories. Therefore, participants who assigned a probability of less than \(90 \%\) on all six test points were removed from that condition. This left 203 participants in the one-category task and 165 participants in the two-category task.

Our first question was whether different cover stories about sampling had an effect on generalization. We examined this by first looking at the raw generalization probability estimates provided by participants. Figures 3 and 4 show the mean generalization probabilities across each condition by block. Consistent with our predictions, in both tasks the mean generalization probability was lower (i.e., participants tightened their generalizations more) in the STRONG SAMPLING condition relative to the WEAK SAMPLING condition \((t(201)=\) \(-.290, p<.05\) for the one-category task and \(t(163)=-3.07\), \(p<.05\) for the two-category task).

Another way to determine whether the sampling cover story had an effect is to fit individual data using the mixed sampling model from Navarro et al. (2012). This model interpolates between weak and strong sampling assumptions, assuming that an observation is strongly sampled with probability \(\theta\) and weakly sampled with probability \(1-\theta\). We can use this to calculate a best-fit \(\theta\) value for each person, reflecting the extent to which their generalizations were consistent with strong sampling ( \(\theta\) close to 1 ), weak sampling ( \(\theta\) close to 0 ), or something in between. Because the mixed sampling


Figure 5: Generalization by block in an additional experiment in which participants were given many more instances in blocks 2 and 3. Generalization probabilities tightened with additional instances, suggesting that earlier lack of tightening was due to conservative updating rather than rejection of the implications of a strong sampling model.
model was originally developed to model generalization responses from a single category, we captured participant responses from the two-category case by treating the left category as the single category whose consequential region is bounded by the leftmost point in the right category. Overall, the model was able to provide a good account for individual responses in both tasks, with a median correlation between the model predictions and participant responses of 0.92 in the one-category task and 0.96 in the two-category task.

As expected, there were significant individual differences in inferred \(\theta\) values as a function of cover story. Calculating separate beta regressions over condition for each task (which we did because the distribution of \(\theta\) values deviated from normality) shows that type of sampling condition was a significant predictor of the estimated \(\theta\) value in both the one-category \((z(3)=-4.23, p<.001)\), and the two-category task \((z(3)=-4.38, p<.001)\). Overall, these results suggest that people did change their generalizations in response to the cover story, and that the \(\theta\) parameter in the mixed sampling model is sensitive to that change.

A related prediction was that increasing the number of instances should result in tighter generalization in the STRONG SAMPLING condition. We tested this prediction by comparing generalization probabilities in the first and last (third) block of test trials (shown in Figures 3 and 4). Although there was a significant difference between generalization probability in the first and last blocks in the two-category EQUAL BASE RATES condition (paired-samples t -test, \(t(29)=2.16, p=\) \(0.019)\), the differences in the one-category task \((t(116)=\) \(1.08, p=0.142\) ) and the two-category UNEQUAL BASE RATES condition \((t(65)=1.50, p=0.069)\) did not reach sig-
nificance. \({ }^{2}\) Is this because people do not, as predicted by the strong sampling model, tighten their generalizations with additional instances? Or is it simply that people are conservative, tightening their generalizations less than such a model would predict?

To investigate this question, we ran an additional experiment involving generalization with 47 participants in the onecategory task and 44 participants in the two-category task. The experiment was identical to the STRONG SAMPLING condition of the previous one except that participants were shown many more instances in blocks two and three. As Figure 5 illustrates, when presented with these large amounts of additional instances people in all conditions and tasks tightened their generalizations considerably. Each person's mean generalization probability in the last block was significantly less than their generalization in the first block in both the onecategory \((t(46)=4.53, p<.001)\) and the two-category task \((t(43)=4.07, p<.001)\). Within the two-category task, generalizations tightened significantly in both the EQUAL BASE RATES \((t(14)=2.31, p=0.018)\) and UNEQUAL BASE RATES \((t(28)=3.27, p=0.014)\) condition. \({ }^{3}\) This pattern of tightening with more instances is more consistent with a Bayesian model that includes some proportion of strong sampling than a standard categorization model like the GCM.

\footnotetext{
\({ }^{2}\) As expected, all differences in the WEAK SAMPLING conditions were not significant, with \(p\) values ranging from 0.317 to 0.458 .
\({ }^{3}\) Recall that this condition incorporated the LOWER BASE RATE and HIGHER BASE RATE conditions into one analysis in which both the high-base-rate and low-base-rate left-hand category were combined. Both show significant tightening when analyzed separately as well.
}

\section*{Discussion}

This current work clarifies perhaps the most troublesome aspect from Navarro et al. (2012): that large individual differences in proportion of strong and weak sampling assumptions were observed but people did not seem to be sensitive to the sampling type suggested by the cover story. By directly referencing in the cover story how samples were being generated: either by direct selection for strong sampling or random occurrence for weak sampling, we find reliable differences in generalization between the two cover stories in the onecategory condition. This pattern of results is accounted for naturally by a Bayesian model using a mixture of strong and weak sampling assumptions (Navarro et al., 2012) and is consistent with standard categorization models such as the GCM (Nosofsky, 1986) that rely on differences in the specificity parameter between cover story conditions. We believe the sampling assumption model account is slightly more parsimonious because it a priori predicts that the weak sampling condition will show wider generalization gradients than the strong sampling condition, rather than relying on a freely varying model parameter.

Interestingly, the difference between strong and weak cover stories is found not only in the one-category but also the two-category scenario. That this pattern exists not only when only positive examples of a single category are observed but also when more than one category is observed, suggests that beliefs about sampling processes influence behavior even in situations more traditionally thought of as category learning. As in the one-category scenario, a model Navarro et al. (2012) without category learning processes and relying only on different mixtures of sampling assumptions is able to account for the behavioral results with a high degree of accuracy.

The presence of significant gradient tightening only at large changes in the number of instances suggests some additional process is mediating the effect of gradient tightening predicted by the Bayesian model that incorporates a mixture of strong and weak sampling. One possibility for such a mediating process would be conservatism (Phillips \& Edwards, 1966), some reluctance to update beliefs about the boundary of each category as much as is suggested by a rational model that includes strong sampling. This conservatism may be due to assumptions that learners might be making about other possible sampling processes including noisy instance generation, noisy labelling, or could be the result of cognitive processes that do not weigh each instance equally as the Bayesian model does (Navon, 1978).

In summary, in both the one- and two-category scenarios, people had different patterns of generalization from known instances to new instances based on a cover story that suggested strong or weak sampling was generating the instances they saw. Additionally, the degree of generalization decreased as many more instances were shown from the target category, more than predicted by standard models of categorization like the GCM but less than predicted by a Bayesian model that mixed strong and weak sampling. Patterns of
generalization at an individual level for both one- and twocategory scenarios were well accounted for by this Bayesian model, suggesting people are sensitive to the sampling assumptions that are generating the instances they see during categorization.

\section*{Acknowledgments}

DJN, AP, and ATH were all supported by ARC grant DP110104949. In addition, DJN received salary support from ARC grant FT110100431 and AP received salary support from ARC grant DE120102378.

\section*{References}

Hsu, A., \& Griffiths, T. (2010). Effects of generative and discriminative learning on use of category variability. In 32nd Annual Conference of the Cognitive Science Society.
Kruschke, J. K. (1992). ALCOVE: an exemplar-based connectionist model of category learning. Psychological Review, 99(1), 22.
Navarro, D. J., Dry, M. J., \& Lee, M. D. (2012). Sampling assumptions in inductive generalization. Cognitive Science, 36(2), 187-223.
Navon, D. (1978). The importance of being conservative: Some reflections on human bayesian behaviour. British Journal of Mathematical and Statistical Psychology, 31(1), 33-48.
Nosofsky, R. (1986). Attention, similarity, and the identification-categorization relationship. Journal of Experimental Psychology: General, 115(1), 39.
Phillips, L., \& Edwards, W. (1966). Conservatism in a simple probability inference task. Journal of Experimental Psychology: General, 72(3), 346.
Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. Science, 237(4820), 1317-1323.
Tenenbaum, J. B., \& Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. Behavioral and Brain Sciences, 24(4), 629-641.
Xu, F., \& Tenenbaum, J. B. (2007). Sensitivity to sampling in Bayesian word learning. Developmental Science, 10(3), 288-297.

\title{
Help-Seeking As A Cause of Young Children's Collaboration
}

\author{
Christopher Vredenburgh (cv92@cornell.edu) \\ Department of Human Development, Cornell University \\ Ithaca, NY, 14850, USA \\ Tamar Kushnir (tk397@cornell.edu) \\ Department of Human Development, Cornell University \\ Ithaca, NY, 14850, USA
}

\begin{abstract}
Young children's collaboration is a topic of great interest, yet what causes children to initiate collaboration in some circumstances but not others is unclear. In this research, we analyzed preschoolers' collaboration as an information gathering activity in a toy assembly activity. We independently assessed children's competency at a similar building task and, using a separate group of children, the difficulty of each step of the activity. We hypothesized that children would request collaborative assistance when they needed assistance (that is, when they were less competent and/or the task was more difficult), but act independently when capable. The results confirmed that preschoolers were more likely to request collaborative assistance as the difficulty of the activity increased and more so if they were initially less competent. The results suggest that preschoolers' collaboration may be profitably viewed as an information gathering activity.
\end{abstract}

Keywords: Collaboration; help-seeking; social learning; preschool children; play.

\section*{Introduction}

Recently, there has been considerable interest in children's early-emerging social learning abilities, including their reliance on social information (Koenig \& Harris, 2005) and their propensity to learn through collaboration (Duran \& Gauvain, 1993; Foley, Ratner, \& House, 2002; Paradise \& Rogoff, 2009; Sommerville \& Hammond, 2007). Indeed, it has even been argued that these social learning abilities, in particular collaboration and the psychological motivations underlying collaboration, are what distinguish humans from nonhuman primates (Tomasello, Call, Behne, \& Moll, 2005).

Recent research on children's collaboration has emphasized the importance of children's tendency to appreciate joint goals and commitment to collaborators. This research has demonstrated that children collaborate even when doing so does not gain them any explicit benefits (Warneken, Gräfenhain, \& Tomasello, 2012). Nonetheless, what causes children to initiate collaboration is still widely debated. Researchers have proposed a range of factors that highlight complementary processes that are generally consistent with one another (Tomasello et al 2005). The range of suggested factors includes a general motivation to share cognitive states with others, a "curiosity" to understand psychological and physical causes, and a social game theoretic distinctively rewarding to humans
(Tomasello et al, 2005). These possible causes are clearly not mutually exclusive. Moreover, many of them make a similar prediction: when children are offered the opportunity to involve others in tasks, they will do so regardless of whether they need help at all.

Intriguingly, research on people's help-seeking could shed light on why children collaborate in some circumstances, but not all. Specifically, help-seeking involves a help seeker signaling to a helper a desire for them to assist in attaining a goal. Help-seeking often occurs when an individual is not confident in their ability to independently complete the task (Nelson-Le Gall et al, 1990). Interestingly, requests for assistance can result in a range of collaborative exchanges of information and action. For instance, responses can vary from indirect verbal hints that facilitate the help seeker to direct coordination of actions between the help seeker and helper. Indeed, help seekers often prefer to avoid receiving too much help, so as to remain actively involved (Nelson-Le Gall, 1986). Thus, help-seeking generally occurs in relation to uncertainty about independently carrying out the task, and can lead to a variety of collaborative exchanges of information and action.

In this research, we examined children's helpseeking as a proximate cause of collaboration that may contribute to children's learning. From this perspective, children face information gathering trade-offs in acting alone versus collaborating. Relatedly, empirical studies show that children often prefer to play on their own, and indeed there are learning benefits to such autonomous exploration (Schulz \& Bonawitz, 2007). Children at times learn more from acting than from watching someone else perform an action (Berry, 1991; Sommerville, Hildebrand \& Crane, 2008; Kushnir, Wellman, \& Gelman, 2009). However, when a child is cognitively or motorically unable to perform an activity, they gain little or no information by acting. Therefore, in this case the child may seek assistance instead of struggling alone. We hypothesized that, rather than always involving others in their play, children will be more likely to request collaborative assistance only when they need assistance. More specifically, we ask whether the difficulty of the activity and the competence of the individual child will predict changes in the frequency and nature of their collaborative interactions.

The empirical investigation of our hypotheses required permitting children to choose to act independently
or request collaborative assistance. In permitting children to act spontaneously, we needed to form and apply definitions of collaboration in our coding scheme. To do this, we referenced past research on children's collaboration to attempt to remain consistent in the definition of the phenomenon. In past research, collaboration has been defined as actively coordinating actions, verbally planning towards a goal, and taking turns with another person (Foley, Ratner, \& House, 2002; Warneken, Chen, \& Tomasello, 2012). As is described below, we adopted the standard that in order for an event to be described as collaborative, the child needed to coordinate actions and/or verbally plan with the collaborator. Furthermore, to account for and analyze the varied involvement of the child and collaborator in their spontaneous interactions, we ranked each interaction using defined levels of collaboration, as described below.

Our empirical investigation also required an activity which had multiple parts, each with different degrees of difficulty, and measured children's ability to independently complete the activity. To this end, we designed a toy assembly task in which children built toys by following sequences of instructive pictures. To assess children's initial competency children completed an Assessment Toy that provided an estimate of their ability to independently construct the toys, termed the Competency Score. To assess the difficulty of the stimuli, a group of children constructed all of the toys independently without help, and were scored on their ability to complete each step. In this way, we measured the two factors we hypothesized to contribute to children's initiation of collaboration.

\section*{Method}

Participants. Participants were forty preschoolers ( \(\mathrm{M}=52.44\) months, \(\mathrm{SD}=9.7\) months; twenty-one females). Children were recruited from preschools and from a database of research participants whose parents expressed interest in participating in research. The children were all from the surrounding region of a rural university town and were predominantly Caucasian and middle class. Three additional children were excluded from the final sample; one due to experimenter error, one due to uncooperativeness, and one due to teacher interruption.

Stimuli. The stimuli were Edushape Interstar rings. In the current experiment, numerous rings were connected with one another so as to resemble larger objects. Children were shown laminated instructive pictures depicting each step of construction for four different toys.

Apparatus. Testing sessions occurred at a childsized table in a quiet room in the laboratory or in a quiet room in the child's preschool. The interactions were recorded with two Sony DCR-SR68 digital cameras.

\section*{Procedure}

Warm-up Toy. The Warm-up Toy, termed the Key, was completed to teach children how to manipulate the toy pieces and make them look like the instructive pictures. The experimenter told the child that they had some toys and
some pictures, and they could make the toys look like the pictures. The experimenter told the child to watch him/her as they completed the first step. After completing the first step, the experimenter asked the child, "Does that look like the picture?" If the child said no, the experimenter explained that the color, position, and number of pieces all made it look like the picture. The child and experimenter then took turns making the Key. Corrective feedback was given for mistakes.

Assessment Toy. Next the child completed an Assessment Toy, termed the Boat, which provided a graded assessment of the child's competence in independently constructing the toys as shown in the instructive pictures. The experimenter asked the child to do the Boat independently, saying, "You can do this one by yourself by making it look like the picture. Start with the first picture. Each time you need a new picture, just move the picture. Now go ahead and make it look like the picture." As the child completed the Boat, the experimenter quietly watched the child and did not provide assistance or corrective feedback. The child had up to five minutes to complete the Boat.

Test Toys. The child then completed the two Test Toys, Sally and Sally's House. Twenty-two of the children were randomly sorted into the Collaboration Group and eighteen into the Non-collaboration Group. In both conditions, half of the children did Sally first and half did Sally's House first. Children had as long as needed to finish the Test Toys.

In the Collaboration Group, the experimenter looked at the child and said, "Now I can help you make Sally, so just let me know when you want me to do some, OK? So if you want help, I'm right here." The experimenter sat and watched, and did not intervene or provide any sort of verbal feedback unless the child initiated collaboration (see "Collaborative Responses" below for details on how the experimenter responded to bids).

In the Non-collaboration Group, the experimenter said, "You can do this one by yourself by making it look like the picture. Now go ahead and make it look like the picture." The experimenter sat quietly and watched the child complete the toy. The experimenter did not intervene or provide any sort of verbal feedback, and responded to requests for assistance as in the Assessment Toy.

Bids for Collaboration. Based on prior work and our own pilot observations in preschools, children initiate collaboration by establishing eye contact, remarking that the activity is difficult, and directly asking for assistance. We therefore accepted these as bids for collaboration. The experimenter responded to 2 seconds of eye contact and remarks of difficulty by asking, "Do you want me to help?" If the child declined assistance, no collaboration occurred. If the child assented, the experimenter collaborated. The experimenter responded to direct requests for assistance by collaborating with the child without further questioning.

Limits of Collaboration. The experimenter always provided helpful, unhesitating, and accurate
assistance. The experimenter collaborated for a single step at a time, unless the child asked for further assistance on the following step. If the child asked the experimenter not to intervene further, the experimenter stopped collaborating. If the child had made mistakes in steps prior to the one at which they asked for assistance, the experimenter aided in correcting the past mistakes. In this way, the experimenter did not condone errors, functioned as an ecologically valid adult collaborator, and avoided the potential complication of inconsistent experimenter responses across children to vague requests for assistance.

Collaborative Responses. If the child structured the experimenter's response by specifying a particular motoric or cognitive difficulty, the experimenter addressed the particular problem. For example, if the child was struggling to fit two pieces together and commented that it was difficult to put them together, the experimenter assisted the child in pushing them together. In this case, both the child and experimenter would be involved in physically fitting them together. If the child simply asked if one piece went on top of the other, the collaborator provided the information and permitted the child to physically carry out the actions. In response to vague requests for collaboration without child action, for instance looking at the step and stating "This is too hard," the experimenter gathered the correct pieces, carried out the step, and provided an explanation. Likewise, if the child simply asked for verbal clarification, the experimenter's response was limited to verbal clarification. In this way, the experimenter's collaboration was contingent upon the extent to which the child structured it.

\section*{Coding}

Children's Competency- Assessment Toy (Boat). We assessed children's competency in constructing the Assessment Toy. Five parameters assessed for each step of the toys whether children: (1) added the correct number of pieces, (2) made the correct number of connections with those pieces, (3) made the correct type of connection(s), (4) added pieces of the correct color(s), and (5) connected the pieces to the correct part of the existing structure. For each step, children earned from 0-5 points; each parameter was worth a minimum of 0 and a maximum of 1 point. Partial credit (e.g. \(1 / 2\) points) was given for partial completion. Children's performance score on each step of the Assessment Toy therefore had a minimum of 0 and a maximum of 5 multiplied by the toy's number of steps ( 8 steps; range 0-40).

Children's performance on the Test Toys (Sally and Sally's House) during collaboration. The same coding as above was used to assess children in the Collaboration Group as they completed the Test Toys. Once again, children's performance score for each Test Toy had a minimum of 0 and a maximum of 5 multiplied by the toy's number of steps (10 steps each; range 0-50).

Step difficulty of the Test Toys (Sally and Sally's House). The Non-Collaboration Group's
competency on each step of the test toy construction was used as a means of computing the difficulty of the Test Toys' steps in the absence of collaboration. The scoring was the same 0 to 5 scale that was used to measure children's competency during assessment. But, this time we did not sum across steps; instead we used the average competency of the Non-Collaboration group at each step as an index of step difficulty in our analysis (below). .

Collaboration initiated?: A binary response code for each step on which children in the Collaboration Group initiated any collaboration. Reliability coding performed on \(55 \%\) of the sample produced \(100 \%\) concordance.

Level of collaboration: For the Collaboration Group, collaborative interactions at each step were rankordered in five categories from lowest to highest levels of collaborative assistance: (0) no collaboration, (1) the child performed the action and the experimenter provided verbal feedback about the child's action, (2) the child provided information about how the pieces assemble and the experimenter performed the action, (3) both the child and the experimenter provided information about how the pieces are assembled and both were involved in assembling them, and (4) the experimenter performed the actions and provided the information about how the pieces are assembled. If multiple levels of collaboration were present during one step, the step was coded by the highest level present. Reliability coding performed on \(55 \%\) of the sample produced \(92 \%\) concordance, indicating high reliability.

\section*{Results}

Children's Competency. Overall, on the Assessment Toy children averaged a Competency Score of 29.83 out of 40 with a standard deviation of 10.49 . There were no systematic differences at assessment between children in each group (Collaboration: \(M=30.17\), Noncollaboration: \(M=29.84, t(38)=.591, p=n s)\). Thus, our entire sample of children displayed sufficient variation in competency to further investigate our hypotheses.

Collaboration Initiated. Our hypothesis predicted that children would collaborate when they were unable to perform the activity independently, and conversely that they would not collaborate when they could construct the toys independently. Our principal analysis therefore assessed whether the difficulty of the Test Toy steps, as measured by the Non-collaboration Group's average performance, and children's competence, as measured by children's Competency Scores, predicted children's choices to collaborate. There were no order effects (Sally first vs. Sally's House first) in either the Collaboration or Noncollaboration Group, so results were collapsed across order for further analysis.

In assessing our predictions of children's collaboration, we needed to properly account for the dependence amongst children's repeated measurements at each step. We therefore employed a General Estimating Equation (GEE), which is a common form of logistic regression analysis, with children as the repeated effect.

Our first dependent variable was the binary variable: whether children selected to collaborate on each step ( \(1=\) yes, \(0=\) no). We first performed an analysis with step difficulty, children's Competency Scores, age, and toy type (Sally \(=0\), Sally's house \(=1\) ) as the predictors. We included children's gender as a factor (female \(=0\), male \(=1\) ). According to the model, the log of the odds of a child collaborating was significantly positively related to step difficulty ( \(p=.000\) ) and significantly negatively related to children's competence ( \(p=.001\) ). As depicted in Table 1, neither age, gender, nor the particular toy related to children's collaboration. This indicates that when the step was more difficult and the child less competent, children were significantly more likely to collaborate than act independently. Similarly, when the steps were simple and the child competent, children were more likely to act independently.

However, it was possible that children selected to collaborate more as they became tired of the activity as opposed to the difficulty of the steps. We performed a second GEE analysis with toy step and step difficulty as the predictors, and collaboration as the dependent variable. Step difficulty was a statistically significant predictor of collaboration, but toy step was not (Step difficulty: \(\beta=.717\), Wald's \(\chi 2=20.345, p=.000\); Step: \(\beta=-.003\), Wald's \(\chi 2=\) \(.017, p=.896\) ). This result helped to specify that step difficulty, as opposed to the order of the toy steps and/or ordering of the toys, related to children's collaboration.

Table 1: Cumulative Logistic Regression Analysis of Children's Choices to Collaborate
\begin{tabular}{|c|c|c|c|c|c|}
\hline Predictor & \(\boldsymbol{\beta}\) & \(\boldsymbol{S E} \boldsymbol{\beta}\) & \begin{tabular}{c} 
Wald's \\
\(\boldsymbol{\chi} \mathbf{2}\)
\end{tabular} & \(\boldsymbol{P}\) & OR \\
\hline Constant & 6.01 & 2.88 & 4.34 & \(.00^{* * *}\) & \begin{tabular}{c}
405. \\
48
\end{tabular} \\
\hline Toy & -.05 & .34 & .02 & .89 & .95 \\
\hline Gender & .00 & .73 & .00 & .90 & 1.00 \\
\hline \begin{tabular}{c} 
Step \\
Difficulty
\end{tabular} & .84 & .14 & 36.18 & \(.00^{* * *}\) & 2.31 \\
\hline \begin{tabular}{c} 
Compete- \\
ncy
\end{tabular} & -.08 & .02 & 10.39 & \(.00^{* * *}\) & .93 \\
\hline \begin{tabular}{c} 
Age
\end{tabular} & -.41 & .62 & .44 & .51 & .66 \\
\hline
\end{tabular}

Table 1: Table 1 shows the results of the parameter estimates for a logistic regression analysis performed with a General Estimating Equation. The model assesses which variables relate to children's choice to collaborate or act independently. Toy (Sally \(=0\), Sally's House \(=1\) ) and gender \((\) female \(=0\), male \(=1)\) were entered as factors. Step difficulty, competency, and children's age were entered as covariates. Degrees of freedom \(=1\).
*** Indicates statistically significant at the . 001 level.
While these logistic regression results were encouraging, we desired a direct assessment of how the two factors in our conceptual model of collaboration compared with children's observed behavior. To do so, we divided the children into three categories of competency and the steps into three categories of difficulty. The majority of children had a Competency Score in the range of 35-40 out of a maximum of 40 , with only one child scoring below 10 . We therefore developed the following categories of Competency Scores: Less Competent (0-20), More Competent (21-35), and Very Competent (36-40). As for step difficulty, the lowest step score was 1.87 and the majority of step scores were above 3.5. To account for the lack of difficult steps, we defined the following three categories of step difficulty: Simple (0-1.0), Somewhat Difficult (1.1-2.0), and More Difficult (2.1-5).

Based on the categories defined above, we calculated the observed probability of children collaborating for each category of competency and difficulty. The observed probabilities are displayed in the line graph in Figure 1a. As shown, collaboration was more likely as the step difficulty increased and children's competency decreased. This result was consistent with our hypotheses.


Figure la: Children were categorized into three competency categories and the toy steps into three difficulty categories. The graph shows children's observed probability of collaboration for each category of children and steps. Standard error bars are displayed.

We were unable to assess a GEE model based on categorical variables representing the categories because some of the cells would contain 0 (the less competent group's children collaborated on all of the more difficult steps). Instead, we assessed how our statistical logistic GEE model's predictions, based on the continuous values of competency and step difficulty, compared with the observed probabilities. Critically, this provides evidence as to how well the occurrence of children's collaboration coheres with the two factors in our conceptual model: step difficulty (represented by "D" below) and competency (represented by
"C" below). Computing the probabilities from the logistic GEE model consisted of applying the following equation:
\[
P=e^{\text {Constant }+. \mathrm{D}^{*} \mathrm{M}_{\mathrm{D}}+\mathrm{C}^{*} \mathrm{M}_{\mathrm{C}}} /\left(1+e^{\text {Constant }+. \mathrm{D}^{*} \mathrm{M}_{\mathrm{D}}+\mathrm{C}^{*} \mathrm{M}_{\mathrm{C}}}\right)
\]

We selected the means (represented by "M" above) of the observed values of each category to be the representative covariate values. We then computed the probability of collaboration using the parameter estimates provided by the logistic GEE model. The results are displayed in Figure 1b.


Figure 1b: As outlined in the text, three categories of child competency and step difficulty were defined. The probability of collaborating was then computed from our logistic GEE model consisting of step difficulty and competency. Standard error bars are displayed.

Figures 1a and 1b, that is, the observed and computed probabilities of children collaborating on a given step, are remarkably similar. The congruence of the cell values and direction of change augment the logistical regression analyses by providing a direct demonstration that children's behavior was consistent with our predictions. Most importantly, this analysis suggests a large portion of children's decisions to collaborate may be a function of two factors: children's competency and the difficulty of the activity.

Levels of Collaboration. We also investigated whether the character, or magnitude, of the collaborative interactions differed as a matter of step difficulty and children's competency. We performed a multinomial distribution GEE in which the dependent variable was the level of collaboration, with no collaboration being level 0 . The specific toy (Sally \(=0\), Sally's House \(=1\) ) and children's gender (female \(=0\), male \(=1\) ) were the factors. The three predictors were step difficulty, children's competency, and age. The resulting analysis indicated that the log of the odds of raising the level of children's collaboration was significantly positively related to step difficulty ( \(p=.000\) ) and significantly negatively related to children's competency ( \(p=.000\); Table 2). This indicates that the collaborative interactions tended to involve more
action and information sharing from the adult collaborator as the difficulty of the steps rose and children's Competency Scores decreased. Neither age nor any of the factors related to the character of children's collaboration. This model furthers our understanding by suggesting that not simply the occurrence of collaboration, but also the character of the collaborative interactions relate to the difficulty of the activity and children's ability to independently execute the activity.

Table 2: Cumulative Logistic Regression Analysis of Children's Level of Collaboration
\begin{tabular}{|c|c|c|c|c|c|}
\hline Predictor & \(\boldsymbol{\beta}\) & \(\boldsymbol{S E} \boldsymbol{\beta}\) & \begin{tabular}{c} 
Wald's \\
\(\chi^{2}\)
\end{tabular} & \(\boldsymbol{P}\) & OR \\
\hline \begin{tabular}{c} 
Constant \\
(Level = 0)
\end{tabular} & -2.76 & 2.28 & 1.47 & .23 & .06 \\
\hline \begin{tabular}{c} 
Constant \\
(Level = 1)
\end{tabular} & -2.28 & 2.28 & 1.0 & .32 & .10 \\
\hline \begin{tabular}{c} 
Constant \\
(Level = 2)
\end{tabular} & -1.86 & 2.30 & .66 & .42 & .16 \\
\hline \begin{tabular}{c} 
Constant \\
(Level = 3)
\end{tabular} & -.79 & 2.27 & .12 & .73 & .46 \\
\hline \begin{tabular}{c} 
Toy
\end{tabular} & -.15 & .39 & .15 & .70 & .86 \\
\hline Gender & -.06 & .66 & .01 & .92 & .94 \\
\hline \begin{tabular}{c} 
Step \\
Difficulty
\end{tabular} & .94 & .15 & 37.04 & \(.00^{* * *}\) & 2.56 \\
\hline \begin{tabular}{c} 
Competen \\
-cy
\end{tabular} & -.11 & .02 & 45.81 & \(.00^{* * *}\) & .90 \\
\hline Age & -.41 & .49 & .72 & .40 & .66 \\
\hline
\end{tabular}

Table 2: Table 2 shows the results of the parameter estimates for a cumulative logistic regression analysis performed with a General Estimating Equation. The model assesses which variables are predictive of the character of children's collaboration. Toy (Sally = 0, Sally's House = 1) and gender (female \(=0\), male \(=1\) ) were entered as factors. Step difficulty, competency, and age were entered as covariates. The levels of collaboration, described above, refer to different categories and magnitudes of collaborative interactions. Degrees of freedom \(=1\).
*** Indicates statistically significant at the . 001 level.
Again, it was possible that children involved the collaborator more because of fatigue of the activity as opposed to step difficulty. We therefore performed another analysis to evaluate whether differences in the character of children's collaboration resulted from the order of the steps as opposed to step difficulty. The regression indicated that step difficulty, not the order of steps, related to the character of children's collaboration (Step difficulty: \(\beta=.744\), Wald's \(\chi 2=20.989, p=.000\); Step: \(\beta=.015\), Wald's \(\chi 2=.228, p=\) .633). This provided further evidence that the difficulty of the activity, as opposed to some other aspect inherent in the order of steps, related to the manner in which children collaborated.

\section*{Discussion}

The results show that the probability of a child requesting collaborative assistance on a given step was
predicted both by the child's initial competency in constructing similar toys and the difficulty of constructing the same toys without adult assistance. Indeed, a statistical model consisting of those two predictors alone provided a comparable match to children's observed probability of collaborating (Figures 1a and 1b). Second, the character of children's collaborative interactions, that is, the extent to which children were involved, was predicted by children's competency and step difficulty. This indicates that these two factors are not only related to the occurrence, but also to the substance of collaborative interactions.

These results support an "information gathering" perspective of children's collaboration related to their helpseeking. This can be best appreciated by understanding the link between the difficulty of the activity and the information to be gained by independent versus collaborative behavior. Indeed, children are implicitly motivated to seek more information through active search when evidence is ambiguous or complex (Schulz \& Bonawitz, 2007). Initiating collaboration may have a similar motivation. Indeed, seeking collaborative assistance may be the optimal strategy in circumstances in which independent exploration is not providing the necessary information to overcome difficulty.

Of course, this by no means precludes the importance of other factors, such as a species-wide prosocial disposition (Tomasello et al, 2005). Indeed, it is certain that other factors contribute to young children's collaboration. However, our results suggest that, in the motivation to accomplish a goal (complete a task, learn a new skill, etc.), aspects of the environment - including the type of goal or task, the competency, skill or knowledge of any individual child - may serve as powerful influences on whether collaboration is initiated, if at all. Future work is needed to examine how the various "proximate causes" of collaboration interact in children's everyday behavior and in different contexts, such as peer collaboration.

Our results also suggest a way in which social learning and learning through exploratory play may be fully integrated. That is, children are neither "stubborn autodidacts" (Harris, 2002) when they learn nor are they passive recipients of social information. Rather, through their own activity, children trade between exploring by themselves and exploiting the knowledge of others. By addressing both the nature and the immediate causes of collaborative vs. non-collaborative behavior, future work may shed light on the many ways in which they relate, and how collaboration contributes to children's impressive early learning abilities.

\section*{Acknowledgments}

We thank Yue Yu, Felix Thoemmes, and Francoise Vermelyan for their helpful feedback.

\section*{References}

Berry (1991). The role of action in implicit learning. The
Quarterly Journal of Experimental Psychology A: Human

Experimental Psychology, 43A(4), 881-906. doi:10.1080/14640749108400961
Duran, R. T., \& Gauvain, M. (1993). The role of age versus expertise in peer collaboration during joint planning. Journal Of Experimental Child Psychology, 55(2), 227242. doi:10.1006/jecp.1993.1013

Foley, M., Ratner, H., \& House, A. (2002). Anticipation and source-monitoring errors: Children's memory for collaborative activities. Journal Of Cognition And Development, 3(4), 385-414.
Harris, P. L. (2002). What do children learn from testimony? In P. Carruthers, M. Siegal, \& S. Stich. (Eds.), Cognitive bases of science. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511613517.018
Koenig, M. A., \& Harris, P. L. (2005). The role of social cognition in early trust. Trends In Cognitive Sciences, 9(10), 457-459.
Kushnir, T., Wellman, H. M., \& Gelman, S. A. (2009). A self-agency bias in preschoolers' causal inferences. Developmental Psychology, 45(2), 597-603.
Nelson-le Gall, S. A. (1987). Necessary and unnecessary help-seeking in children. The Journal Of Genetic Psychology: Research And Theory On Human Development, 148(1), 53-62. doi:10.1080/00221325.1987.9914536
Nelson-Le Gall, S., Kratzer, L., Jones, E., \& DeCooke, P. (1990). Children's self-assessment of performance and task-related help seeking. Journal Of Experimental Child Psychology, 49(2), 245-263. doi:10.1016/0022-0965(90)90057-F
Paradise, R., \& Rogoff, B. (2009). Side by side: Learning by observing and pitching in. Ethos, 37(1), 102-138. doi:10.1111/j.1548-1352.2009.01033.x
Schulz, L. E., \& Bonawitz, E. (2007). Serious fun: Preschoolers engage in more exploratory play when evidence is confounded. Developmental Psychology, 43(4), 1045-1050. doi:10.1037/0012-1649.43.4.1045
Sommerville, J. A., \& Hammond, A. J. (2007). Treating another's actions as one's own: Children's memory of and learning from joint activity. Developmental Psychology, 43(4), 1003-1018.
Sommerville, J. A., Hildebrand, E. A., \& Crane, C. C. (2008). Experience matters: The impact of doing versus watching on infants' subsequent perception of tool-use events. Developmental Psychology, 44(5), 1249-1256. doi:10.1037/a0012296
Tomasello, M., Carpenter, M., Call, J., Behne, T., \& Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. Behavioral and Brain Sciences, 28(5), 675-735. doi:10.1017/S0140525X05000129
Warneken, F., Gräfenhain, M., \& Tomasello, M. (2012). Collaborative partner or social tool? New evidence for young children's understanding of joint intentions in collaborative activities. Developmental Science, 15(1), 54-61. doi:10.1111/j.1467-7687.2011.01107.x

\title{
When words get physical: evidence for proficiency-modulated somatotopic motor interference during second language comprehension
}

\author{
Nikola Vukovic (nv254@cam.ac.uk) \\ Department of Theoretical and Applied Linguistics, University of Cambridge
}

\begin{abstract}
New theories of cognition posit an intimate link between higher cognitive processes and the sensorimotor areas of the brain. In a reaction time-based translation task, second language (L2) speakers responded to action verbs using a microphone or a response pad. A significant interaction among Response Modality, Verb Type, and Proficiency indicated that more proficient L2 speakers took significantly longer to respond with their hands to previously seen handrelated verbs, but not mouth-related ones. Conversely, responding using a microphone led to slower latencies in the case of mouth-verbs, but not hand-verbs. Amidst virtually exclusively monolingual research on embodied cognition, the current study provides evidence that reading L2 action verbs selectively interferes with subsequently performed manual or verbal responses, suggesting that semantic representations of these verbs are distributed over neural substrates underlying action execution. The role of proficiency and experience in language comprehension is discussed.
\end{abstract}

Keywords: embodied cognition, second language, semantics

\section*{Introduction}

Ever since the cognitive turn in psychology, the human mind has been likened to a computer, and the essence of our mental life envisaged as the outcome of complex calculation over abstract symbolic elements. In this traditional framework, human cognition is defined as modular, with distinct components operating on information independently and autonomously (Fodor 1998). In contrast to this perspective, recent years have witnessed the strengthening of theoretical paradigms which posit thought as being grounded in experience and sensation. These have become known as Embodied Cognition theories. According to Embodied Cognition, human concepts are not amodal transductions of sensory data, but are instead grounded in sensory-motor processing itself. Much evidence for embodied cognition comes from studies of language processing. Within this new framework, language comprehension is thought of as grounded in, and intimately linked to, neural resources used in action, perception, and emotion (Barsalou 2008).

By now there exists a wealth of behavioural research which supports the claim that language comprehension and action execution are subserved by common neural resources. For example, Glenberg and Kaschak (2002) demonstrated an Action-Sentence Compatibility effect, where judging the sensibility of sentences which implied a movement towards or away from the body (You gave Andy the pizza vs. Andy gave you the pizza) facilitated congruent arm movements. In a similar sentence sensibility judgement task, Zwaan and Taylor (2006) found that participants were
significantly faster to perform manual rotation of a response knob when the rotation (clockwise or anti-clockwise) was congruent with the meaning of previously presented sentential material (turn the volume up/down). Bergen et al. (2003) asked participants to manually verify names of pictures representing actions, and found that response times were slower when they had to reject actions performed with the same (vs. different) effector. In a different study, Buccino et al. (2005) had participants listen to and judge the concreteness of sentences using hand or feet responses. They observed interference effects pointing toward the conclusion that verbally presented action sentences activate the motor system. These studies suggest that language and action are highly interconnected and that processing action language functionally involves activation of motor representations in the brain. Moreover, this interaction is differentially articulated as facilitation or inhibition, based on the temporal relationship between stimulus and response (Boulenger et al. 2006).

The findings from behavioural studies outlined above find additional support in the neurosciences, where experiments have shown interdependence between cognition and simulation of motor and perceptual states. There have by now been numerous studies which demonstrate that semantic processing of a word activates distributed and diverse networks of sensory and motor information (Farah, McClelland, 1991; Damasio 1990; Caramazza et al., 1990). For example, processing the name of an action engages the motor area which is active during performing that same action (Martin et al., 1995 p. 649-652.). Hearing a word activates auditory associations (Pulvermuller et al 2006), and action-related words elicit cortical activation in the motor areas of the brain, even when the participants are not aware of hearing the word (Pulvermüller et al. 2005). Intriguingly, comprehension of action words does not only reliably activate the motor cortex, but does so in an effectorspecific i.e. somatotopic manner: face, arm, or leg words activate the corresponding parts of the motor system in the central and precentral cortex which control face, arm, or leg movements (Buccino et al 2005, Hauk et al 2004).

Taken together, behavioural and neuroimaging data strongly support predictions and claims of embodied approaches to language and cognition. Semantic representations of words are not amodal or entirely symbolic, but seem to utilise the same neural resources involved in action execution, and it is these strong links which are made apparent in the interactions outlined in the studies above.

Some proponents of disembodied and symbolic approaches to cognition, however, have raised concerns
about the functional relevance of sensorimotor processes in language comprehension (Hickok, 2010; Mahon and Caramazza, 2008). For example, Mahon and Caramazza (2008) have argued that semantic motor activation, as it is described above, could also be incorporated into traditional theories. On this view, the motor cortex becomes activated as an epiphenomenon of, and not part of, semantic retrieval. In other words, semantic motor activation is a result of induced imagery of action, and is as such a downstream consequence of comprehension. However, a number of recent studies presented evidence for automaticity and causality of sensorimotor processes in language comprehension. For example, Pulvermuller et al. (2005) have demonstrated activity in the motor cortex as early as \(100-200 \mathrm{~ms}\) following word recognition - speeds consistent with the idea that these processes are crucial to semantic retrieval (see also Shtyrov et al., 2004). Similarly, Liuzzi et al. (2010) present compelling evidence that the motor cortex is causally involved in learning and processing action words. In their study, transcranial direct current stimulation (tDCS) was used to temporarily disturb the functioning of the motor cortex, which in turn led participants to perform significantly worse in an action word acquisition task, compared to controls. The study indicates that the motor cortex is vital, and even necessary, for word processing, as is also suggested by other TMS, electrical stimulation, and behavioural experiments (Pulvermuller 2005; Fischer and Zwaan, 2008; Glenberg et al., 2008). Results to date thus favour embodied approaches to language comprehension, and make symbolic interpretations of motor semantics rather difficult to maintain.

\section*{The Current Study}

Motor and language processes in the brain seem to form a dynamic and highly interconnected system, with interactions appearing at very early stages. However, a major shortcoming of investigations performed so far is the fact that they are virtually exclusively based on monolingual data, and data obtained from native speakers of a language. Surely, however, any theory that seeks to explain linguistic processes cannot call itself complete without at the same time accounting for how these operate in the majority of the world's population (i.e. bilinguals: Gordon, 2005). With the realisation that over half of the world's population speaks more than one language, this study aims to test the predictions of embodied cognition on people other than monolinguals. The primary question in the experiment reported here is whether in second-language (L2) speakers action-word semantics are distributed over neural substrates involved in action execution. Will L2 speakers, just like native ones, exhibit sensorimotor effects in lexical processing? Do we, in other words, see evidence for interaction between motor and linguistic processes or, alternatively, is there evidence that in their case linguistic processes operate completely independently of sensorimotor ones? How are these affected by proficiency and
experience? Clearly, the extensive work in embodied language processing is in need of specification in terms of how and when grounding takes place in the course of language development.

The current study employed a reaction time-based translation task, which methodologically synthesizes experimental and analytical tools drawn from second language (Altarriba and Mathis 1997, Duyck and Brysbaert 2004) and embodied cognition research (Shebani \& Pulvermüller 2012, Bergen et al 2003, Marino BFM et al 2011). In every trial, participants were presented with a mouth, arm, or leg related verb in their native language (Serbo-Croatian), after which they would see a verb in English. The task was to quickly indicate whether the second verb was a good translation of the first verb, using a button box in one, and a microphone in the other half of the experiment. In half of the trials the English verb was a translation of the Serbo-Croatian one, and in the other half it was not. The critical trials in this experiment were those where the verbs were translations, and were split into two conditions: in one case, the English verbs denoted actions performed with the same body part or effector as the one used for the experimental response (mouth or hand, depending on which experimental response was required), whereas in the other case the verbs indicated actions performed with a different effector.

If we assume a neurobiological model of language in which lexicosemantic circuits are embodied (Pulvermuller, 1999), then verbs describing actions should be realised not only in perisylvian cortical regions traditionally associated with language, but also as circuits in the motor cortex, which is used for executing the actions themselves. If it is true that understanding an action verb produces motor activation, then introducing a concurrent task (the user response) which makes use of that same part of the brain should produce interference, reflected in slower reaction times. In addition, the semantic somatotopy model (Pulvermüller, 1999; 2001) predicts that this interference should be highly specific: processing mouth-related verbs should most strongly interfere with concurrent verbal responses, but not manual ones. Similarly, processing a hand-related verb should lead to much slower latencies when responding with the hand, but not with the mouth.

In the current context of second language speakers, we could expect several possible outcomes: 1. It might be that non-native speakers process language in a completely different way to that of their native counterparts. In other words, their semantic representations might not be distributed across sensorimotor neural substrates - a distribution that might, therefore, be a distinctive hallmark of native speakers. 2. It could be the case that L2 speakers show identical patterns to those of native speakers. 3. Finally, it is possible that L2 speakers show differing amounts of sensorimotor embodiment, as modulated by proficiency and experience.

\section*{Experiment}

\section*{Participants}

Twenty-four right-handed native speakers of Serbo-Croatian (13 female; mean age \(=25.63\) years, \(\mathrm{SD}=3.54\) ) studying at the University of Cambridge took part in the experiment. All participants had normal or corrected-to-normal vision and reported no history of neurological, psychiatric, or language disorders. Participants varied in terms of L2 proficiency, as revealed by the language background questionnaire (see below). All gave their informed consent prior to participation. No participants were aware of the purpose of the experiment.

\section*{Stimuli}

A total of 72 lexical stimuli were used in the experiment: 36 Serbo-Croatian (SC) verbs, and 36 English ones. For each language there were 12 mouth-related (e.g., bite, kiss, sing), 12 hand-related (e.g., peel, take, write), and 12 legrelated (e.g. dance, jump, walk) action verbs. In addition, 18 verb pairs were constructed for the practice trials: 9 practice trials per response modality (mouth/microphone vs. hand/button box). All critical lexical stimuli were matched for psycholinguistic variables such as number of letters, number of phonemes, lexical frequency, and letter bigram frequency. All SC verbs appeared inflected for first person singular present as, for example, in the verb pišem (write), where the suffix \(-m\) attaches to the base form piše-.

\section*{Procedure}

The experiment took place in a sound-attenuated and dimly lit room. Participants sat comfortably in front of a computer screen at a distance of about 60 cm . Stimuli were presented centrally on the screen in lowercase Arial font (size \(=20\) ). Each trial started with a fixation cross presented in the centre of the screen for 1000 ms , followed by a SC verb displayed for another 1000 ms . After a 500 ms interstimulus interval (ISI), an English verb appeared and stayed on the screen until a response was given. Participants were instructed to respond, as quickly and as accurately as possible, whether the English verb was a translation of the previously seen SC verb. They did so by pressing "yes" or
"no" on a button box, in one half of the experiment, and by saying "yes" or "no" into a microphone, in the other half (the order of "mouth response" and "hand response" blocks was counterbalanced across participants). After they gave a response (correct or incorrect), a blank screen was shown for 500 ms , after which a new trial started. Accuracy feedback was displayed only during the practice block. Stimulus presentation and response time collection were performed using the SuperLab Version 4.5 software package (Cedrus Corporation). The experiment consisted of a practice block, and two experimental blocks: one requiring hand, and the other mouth responses. Participants therefore saw each target verb twice, once in each half of the experiment. Each target verb would appear in both the matching and mismatching condition (actions conveyed by L2 verbs shared/did not share the effector with the experimental response). Items were rotated around two presentation lists. If an English target verb was in the sameeffector condition on one list it was in the different-effector condition on the other list, and vice versa. There were equal numbers of same- and different-effector pairs on each list, and equal numbers of participants were tested on each list.

\section*{Language Background Questionnaire}

All participants completed a language background questionnaire. The sample was homogeneous, with all participants being native speakers of Serbo-Croatian, raised in the ex-Yugoslavia territory, and having started learning their L2 at approximately the same time (Mean AoA: 7.86 years, \(\mathrm{SD}=2.4\) ). Overall, participants rated their proficiency in English relatively high, though there were still differences, with some participants having just arrived to England for their undergraduate and graduate courses, and others having been in the country for longer as part of their PhD or postdoctoral research. A simple median split was performed on the proficiency scores, thus creating two groups, the lower and the higher proficiency group, with 12 people in each. The difference in L2 proficiency ratings between these two groups was statistically reliable, \(t(22)=-\) \(7.30, p=.001\). These and other proficiency measures from the language background questionnaire which, importantly, were found to correlate with participants' self-ratings, are summarized in Table 1.
\begin{tabular}{|l|c|c|c|c|c|c|c|c|}
\cline { 2 - 11 } \multicolumn{1}{c|}{} & \multicolumn{2}{|c|}{\begin{tabular}{c} 
Proficiency \\
rating
\end{tabular}} & \multicolumn{2}{c|}{\begin{tabular}{c} 
Age of Acquisition \\
(years)
\end{tabular}} & \multicolumn{2}{c|}{\begin{tabular}{c} 
Time spent in UK \\
(months)
\end{tabular}} & \multicolumn{2}{|c|}{ IELTS score \({ }^{\mathbf{b}}\)} \\
\cline { 2 - 11 } & \(\mathbf{M}\) & \(\mathbf{S D}\) & \(\mathbf{M}\) & \(\mathbf{S D}\) & \(\mathbf{M}\) & SD & \(\mathbf{M}\) & SD \\
\hline Lower Proficiency Group & 5.87 & 0.2 & 7.58 & 2.96 & 9.83 & 4.56 & 7.25 & 0.39 \\
\hline Higher Proficiency Group & 6.71 & 0.3 & 8.08 & 1.78 & 25.25 & 19.85 & 8.54 & 0.40 \\
\hline
\end{tabular}

Table 1. Mean data for participants' \((\mathrm{n}=24)\) language history and self-assessed proficiency ratings \({ }^{\text {a }}\) based on a scale from 1-7
\({ }^{\text {b }}\) Since the majority of scores were IELTS scores, TOEFL and Cambridge Exam scores were also converted to the IELTS scale using the standard Equivalency Table


Figure 1. Mean reaction times for both participant groups in all conditions (H=hand; M=mouth; \(L=l e g\) ). Significant differences in response latencies ( \(\mathbf{p}<.05\) ) are marked with an asterisk.

\section*{Results}

First, trials with erroneous responses were removed. Given that errors were very rare ( \(0.02 \%\) of total trials), they were not analysed separately. No participant performed at less than \(90 \%\) accuracy. Response latencies for correct trials larger than 2SD \(\pm\) mean RT were excluded from statistical analysis as outliers. In total, no more than \(0.5 \%\) of data was lost (Ulrich and Miller 1994).

For the remaining trials, mean RT values for each participant in each condition and block were calculated (see overall means in Figure 1) and entered into a repeatedmeasures analysis of variance (ANOVA) with two withinsubject variables: Response Modality (microphone vs. button box), Verb Type (mouth vs. hand vs. leg), and Proficiency (lower vs. higher) as a between-subject variable. Response latencies were longer when using the microphone than with the button box, and correspondingly the ANOVA revealed a main effect of Response Modality, \(\mathrm{F}(1,22)=\) 44.27, \(p=.001\); partial \(\eta 2=.668\). No other main effect was significant ( \(\gg .05\) ). A two-way significant interaction among Response Modality and Verb Type was found: \(\mathrm{F}(1\), 22) \(=12.58, \mathrm{p}=.002\); partial \(\eta^{2}=.364\). In addition, there was a three-way significant interaction among Response Modality, Verb Type, and Proficiency: \(F(1,22)=15.22, p=\) .001; partial \(\eta^{2}=.409\).
Most importantly, the significant three-way interaction directly addresses the main questions which motivate this study. The interaction indicates that second language speakers of different proficiencies differentially exhibit somatotopic interference on action execution during verb processing. The group of lower proficient speakers showed no significant differences in response times in any of the blocks and conditions. However, more proficient L2 speakers responding with the microphone were slower on mouth verbs than on hand verbs (1394 and 1206 ms respectively, \(\mathrm{t}(11)=3.096, \mathrm{p}=.010\) ). Conversely, their hand responses were slower when processing hand, but not mouth verbs (678 and 620 ms , respectively, \(\mathrm{t}(11)=2.80, \mathrm{p}\)
\(=.017\) ). Responses to leg verbs always patterned with other non-response modality verbs.

Overall then, for the more proficient L2 speakers in the sample, the results reveal a double-dissociation pattern of response interference.

\section*{Discussion}

This study reveals differential interference for mouth and hand responses, brought about as a consequence of processing lexical semantics of action verbs involving different parts of the body, namely the mouth and the hands. Interestingly, this selective effect was modulated by speaker L2 proficiency. In both participant groups, response latencies given with the microphone were longer than those given with the button box. However, our finding that verbal responses are slower than manual ones in L2 speakers is quite in line with results and latencies obtained in previous second language studies (see, for example Kroll et al., 2002). Apart from this main effect, participants in the lower proficiency group demonstrated no significant differences in response times across blocks and conditions. Interestingly, slower responses with hands were observed when higher proficiency participants processed hand verbs, but not mouth or leg verbs. The reverse effect, slower responses during mouth (but not hand or leg) verb processing, was seen when participants used a microphone to respond. These results follow a double dissociation pattern (Shallice 1988; Jones 1983), which allows for much more reliable inferences about the causal and interactive status of the systems and processes involved than was possible in some previous, conceptually similar studies, which only used a single response modality, or used pictorial stimuli where priming from visual features could not be reliably dissociated from true motor interference (see for example Bergen et al 2003; Marino BFM et al 2011). Specifically, this study demonstrates that in more proficient participants processing resources located in specific parts of the motor cortex are shared between action execution and lexicosemantic representations of related verbs. These data, although behavioural in nature, directly bear upon and
support predictions made by psychological and neural models of embodied language processing. Apart from providing novel and strong evidence for embodied semantics, the present study is the first to demonstrate double dissociated and effector specific shared neural resources between action execution and language in nonnative speakers. In addition, it has implications for theories of how grounding takes place, as it suggests that speakers increasingly come to activate non-linguistic systems during word processing as a function of proficiency and experience.

The findings in this study can be interpreted with regard to neurobiological models of language learning and processing. These do not view lexicosemantic representations as static, but adopt an approach in which variation primarily comes not from the strength of formmeaning connections, but the distribution of semantic representations themselves. In other words, it might be the case that there is less meaning associated with L2 lexical forms (for similar proposals see Williams \& Cheung 2011, and Duyck and de Houwer 2008). This explanation makes intuitive sense if we think about the amount of real-life experience with L2 words. For example, L2 words are associated with a much smaller range of senses than L1 words (Finkbeiner and Nicol 2003), and are in the majority of cases learned in artificial classroom environments, often through the use of crude lexical translation. This account is consistent with the literature on cortical learning, where words can be thought of as functional cell assemblies in the brain, formed through Hebbian processes (Pulvermuller 1999). Hebbian learing ("what fires together, wires together") would thus predict that cortical distributions of L2 word semantics are much more restrictive than those of L1 words, due to different (and fewer) learning and usage experiences. If this is the case, then L2 words, learned through translation and with no or limited real-life usage, would be strongly left lateralized and distributed mainly over perisylvian cortices. We should, therefore, as was the case presently, find little activation outside of primary linguistic brain areas. However, if speakers start using L2 words in real-life embodied contexts, such as when studying in a foreign country, then these words would increasingly come to be co-activated with extra-linguistic neural substrates (including the motor cortex), and would "wire together" into a new assembly. The present results are therefore consistent with the idea that the amount of experience and real-life usage leads to changes in the way we semantically represent words. Crucially, there were differences in the amount of time both groups of speakers spent in an English-speaking country (see Table 1), and this difference was statistically reliable \((\mathrm{t}(12)=-2.60, \mathrm{p}=.023)\). In fact, there is evidence for this conclusion from studies testing the linguistic performance of L2 students learning their second language in a classroom vs. an immersed, study abroad setting. Using experimental paradigms similar to the one employed presently, Linck et al. (2009) and Talamas et al (1999) demonstrate that as their proficiency increases, L2
speakers move from exhibiting primarily form-level effects, to showing increased semantic access. This could explain why in the results reported here the participants who spent less time in the UK showed no motor semantic interference during lexical processing, whereas the other group with a significantly longer residence in the country showed reliable double-dissociated interference when processing action verbs. However, while this proposal seems plausible and consistent with neurobiological models of language, it must presently remain at the level of speculation. More work is thus needed to further clarify the questions of how and when grounding takes place in the context of second language acquisition and learning.

\section*{Conclusion}

It has been suggested that embodied sensorimotor systems are an integral part of language comprehension. The current study, for the first time in second language speakers, demonstrates that they increasingly come to incorporate into their action-word semantics the same processing resources used for effector-specific action execution. The resulting somatotopic interference effects observed in proficient L2 speakers present evidence in favour of embodied approaches to language.

\section*{Acknowledgments}

I would like to thank my supervisor, Dr. John N. Williams, for his valuable comments and statistical advice. This study was supported by a CISS graduate scholarship from the University of Cambridge.

\section*{References}

Altarriba, J., \& Mathis, K. M. (1997). Conceptual and lexical development in second language acquisition. Journal of Memory and Language, 36, 550-568.
Barsalou, L. (2008). Grounded cognition. Annual Review of Psychology, 59, 1-21.
Bergen, B., Narayan, S., and Feldman, J. (2003). Embodied verbal semantics: evidence from an image-verb matching task. Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society.
Boulenger, V., Roy, A., Paulignan, Y., Deprez, V., Jeannerod, M., \& Nazir, T. (2006). Cross-talk between language processes and overt motor behavior in the first 200 msec of processing. Journal of Cognitive Neuroscience, 18(10), 1607-1615.
Buccino, G., Riggio, L., Melli, G., Binkofski, F., Gallese, B., \& Rizzolatti, G. (2005). Listening to action-related sentences modulates the activity of the motor system: A combined TMS and behavioral study. Cognitive Brain Research, 24, 355-363.
Caramazza, A., Hillis, A. E., Rapp, B. C., \& Romani, C. (1990). The multiple semantics hypothesis: Multiple confusions? Cognitive Neuropsychology, 7, 161-189.

Damasio, AR. (1990). Category-related recognition deficits as a clue to the neural substrates of knowledge. Trends in Neuroscience, 13:95-98.
Duyck, W. \& de Houwer, J. (2008). Semantic access in second-language visual word processing: Evidence from the semantic Simon paradigm. Psychonomic Bulletin \& Review, 15, 961-966.
Duyck, W. \& Brysbaert, M. (2004). Forward and backward number translation requires conceptual mediation in both balanced and unbalanced bilinguals. Journal of Experimental Psychology: Human Perception and Performance, 30, 889-906.
Farah, M. J., \& McClelland, J. L. (1991). A computational model of semantic memory impairment: Modalityspecificity and emergent category-specificity. Journal of Experimental Psychology: General, 120, 339-357.
Finkbeiner, M., \& Nicol, J. (2003). Semantic category effects in second language word learning. Applied Psycholinguistics, 24, 369-383.
Fischer, M.H., Zwaan, R.A., (2008). Embodied language: a review of the role of the motor system in language comprehension. Quarterly Journal of Experimental Psychology (Colchester), 61 (6), 825-850.
Fodor, J. A. (1998). Concepts: Where Cognitive Science Went Wrong. Cambridge, MA: Harvard University Press.
Glenberg, A.M., Sato, M., Cattaneo, L., (2008). Useinduced motor plasticity affects the processing of abstract and concrete language. Current Biology, 18 (7), R290R291.
Glenberg, A., and Kaschak, M. (2002). Grounding language in action. Psychonomic Bulletin \& Review, 9, 558-565.
Gordon, R. G., Jr. (Ed.). (2005). Ethnologue: Languages of the world. Dallas, TX: SIL International.
Hauk, O., Johnsrude, I., \& Pulvermüller, F. (2004). Somatotopic representation of action words in the motor and premotor cortex. Neuron, 41, 301-307.
Hickok, G. (2010). The role of mirror neurons in speech perception and action word semantics. Language and Cognitive Processes, 25: 749-776.
Jones, G.V. (1983). On double dissociation of function. Neuropsychologia, 21 (4), 397-400.
Kroll, J. F., Michael, E., Tokowicz, N. \& Dufour, R. (2002). The development of lexical fluency in a second language. Second Language Research, 18, 137-171.
Kroll, J.F. \& deGroot, A.M.B. (1997). Lexical and conceptual memory in the bilingual: Mapping form to meaning in two languages. In A.M.B. deGroot \& J.F. Nas (eds.) Tutorials in Bilingualism.
Kroll, J.F. \& Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. Journal of Memory and Language, 33, 149-174.
Linck, J., Kroll, J. F., \& Sunderman, G. (2009). Losing access to the native language while immersed in a second language: Evidence for the role of inhibition in second language learning. Psychological Science, 20, 1507-1515.

Liuzzi, G., Freundlieb, N., Ridder, V., Hoppe, J., Heise, K., Zimerman, M., et al., (2010). The involvement of the left motor cortex in learning of a novel action word lexicon. Current Biology, 20 (19), 1745-1751.
Mahon, B., and Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. Journal of Physiology Paris, 102(1-3): 59-70.
Marino, B. F. M., Gough, P. M., Gallese, V., Riggio, L., \& Buccino, G. (2011). How the motor system handles nouns: a behavioral study. Psychological research.
Martin, A., Haxby, J.V., Lalonde, F.M., Wiggs, C.L., \& Ungerleider, L.G. (1995). Discrete cortical regions associated with knowledge of color and knowledge of action. Science, 270, 102-105.
Pulvermüller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O., Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. Proceedings of the National Academy of Science USA 103, 7865-7870.
Pulvermüller, F., (2005). Brain mechanisms linking language and action. Nature Reviews Neuroscience, 6 (7), 576-582.
Pulvermüller, F., Shtyrov, Y. \& Ilmoniemi, R. (2005) Brain signatures of meaning access in action word recognition, Journal of Cognitive Neuroscience, 17(6), 884-892.
Pulvermüller, F. (2001). Brain reflections of words and their meaning. Trends in Cognitive Sciences, 5: 517-524.
Pulvermüller, F. (1999). Words in the brain's language. Behavioral and Brain Sciences, 22: 253-336.
Shallice T. (1988). From Neuropsychology to Mental Structure. New York: Cambridge University Press.
Shebani, Z., \& Pulvermüller, F. (2011). Moving the Hands and Feet Specifically Impairs Working Memory for Armand Leg-Related Action Words. Cortex, 1-10.
Shtyrov Y., Hauk O., and Pulvermuller F. (2004). Distributed neuronal networks for encoding categoryspecific semantic information: The mismatch negativity to action words. European Journal of Neuroscience, 19(4): 1083-1092.
Talamas,A., Kroll, J.F. and Dufour, R. (1999). From form to meaning: stages in the acquisition of second language vocabulary. Bilingualism: Language and Cognition 2, 4558.

Williams, J.N. \& Cheung, A. (2011). Using priming to explore early word learning. In P. Trofimovich \& K. McDonough (Eds) Applying priming methods to L2 learning, teaching and research: Insights from psycholinguistics. John Benjamins, 73-103.
Zwaan, R. A., \& Taylor, L. J. (2006). Seeing , Acting , Understanding: Motor Resonance in Language Comprehension. Journal of Experimental Psychology: General, 135(1), 1-11.

\title{
Slow drift of individuals' magnitude-to-number mapping.
}

\author{
Edward Vul, David Barner, \& Jess Sullivan (evul, barner, jsulliva@ucsd.edu) \\ Department of Psychology, 9500 Gilman Dr. \# 109 \\ La Jolla, CA 92093-109 USA
}

\begin{abstract}
When estimating the number of dots in a set, adults show bias and variability that scale with numerosity. Increasing variance in estimation is thought to reflect constant Weber noise on perceptual magnitude representations, while the increasing bias reflects miscalibrated mappings of number words onto magnitudes. Here we argue that response variability in numerical estimation increases with numerosity in part due to uncertainty and slow drift in the mapping of numbers onto magnitudes. We show that individuals' number-to-magnitude mapping functions drift slowly over the course of the experiment, with a shared-variance half-life of over 100 trials ( \(\sim 10 \mathrm{~min}\) ). We thus propose a model that treats the word-to-magnitude mapping function as a major source of estimation variability, and that accounts for cross-subject differences in estimation bias and variability, as well as changes to estimation performance within a given subject over time. In doing so, we reconcile the existing literature on the sources of estimation variability, and provide evidence that uncertainty in the word-tomagnitude mapping function is a key limiting factor in estimation performance.
\end{abstract}

Keywords: Approximate number, number words, numerical estimation

\section*{Introduction}

Human adults have access to at least two systems for representing numerical quantity. The first is a noisy and evolutionarily ancient nonverbal number system, called the Approximate Number System (ANS; see Dehaene, 1997; Feigenson, Dehaene, \& Spelke, 2004 for review). The second is the verbal number word system (e.g., one, two, three, etc.). This system unique to humans allows for the precise representation and manipulation of numerical content. When making estimates, adults draw on both of these systems: they use the ANS to represent the magnitude of the stimulus being estimated, and they use the verbal number system to attach a linguistic label to this magnitude.

The interface between these two systems has been an area of recent interest because it is crucial to characterizing how people can provide explicit verbal estimates of numerosity, and more generally how language relates to perception (e.g., Carey, 2009; Izard \& Dehaene, 2008; Thompson \& Opfer, 2011; Sullivan \& Barner, 2012; Sullivan, Juhasz, Slattery, \& Barth, 2011). The question is important for at least two reasons. First, estimation performance has been shown via intervention studies to be causally related to academic success, raising the question of why, and which aspects of training are most important to educational outcomes (Ramani \& Siegler, 2011; Siegler \& Ramani, 2009). Second, estimation tasks are often argued to elucidate properties of the ANS or their comprehension of number word meanings. In the present study, we tested what contribution - if any - this mapping function makes to estimation error. Specifically, we asked two questions about the nature of the number-to-magnitude mapping
function. First, are the mappings between verbal and nonverbal numerical representations stable across individuals? Second, within individuals, are these mapping stable across time? As argued below, the answer to these questions suggests a novel model of numerical estimation, which explains significant aspects of error by appealing to a dynamically changing mapping function, rather than to the ANS.


Figure 1: (Top) Our account of the bias and variability in human numerical estimation assumes two transformations between the physical stimulus and a verbal numerical response (following Izard \& Dehaene, 2008). First, the approximate number system maps the physical stimulus onto a logarithmic magnitude estimate with constant Weber noise. Second, the magnitude estimates are mapped onto the verbal number; we approximate this mapping as bilinear in \(\log -\log\) space with a variable slope. (Bottom) Two novel features of this account are necessary to capture the patterns of errors in human estimation data (one representative subject shown): (1) the mapping function must be non-linear in log-log space, otherwise the pattern of veridical calibration for small numbers, and systematic mis-estimation for large numbers will not hold, (2) the slope of the high end of the mapping function must be variable to capture the increasing variability of estimation for larger numbers.

In the absence of training or feedback, adults are notoriously inaccurate estimators (Kaufman, Lord, Reese, \& Volkmann, 1949; Izard \& Dehaene, 2008; Minturn \& Reese,
1951). For the purposes of the present project, we are interested in two attributes of this inaccuracy: variability and bias.

First, consider the variability of estimates. The degree of variability in estimates typically increases in proportion to the magnitude of the stimulus being estimated, such that the coefficient of variation (the ratio of the standard deviation of estimates for a given magnitude to the mean estimate of that magnitude) remains constant as magnitude increases (e.g., Whalen et al., 1999). Because variability in estimation behavior scales up with magnitude, estimates are typically said to demonstrate the property of scalar variability.

Scalar variability is thought to arise from Weber noise in the ANS: ANS representations are ratio-dependent, and therefore error in its representation of number also scales with number. For example, it is as easy to tell the difference between 5 dots and 10 dots as it is to tell the difference between 500 dots and 1000 dots using the ANS. Because both nonverbal ANS tasks and verbal estimation tasks exhibit scalar variability, many have concluded that scalar variability in estimation arises because the underlying (ANS) perceptual representations of the magnitudes being estimated exhibits weber noise (Dehaene \& Marques, 2002; Izard \& Dehaene, 2008; Le Corre \& Carey, 2007; Negen \& Sarnecka, 2010; Siegler \& Opfer, 2003) and relatively uncontroversial.

It is clear from the past literature that much of the variability found in estimation arises from variability in the ANS representations that support estimates. However, in the present paper, we ask whether all of the variability in estimation performance is explained by Weber noise, or, alternatively, whether the mapping function that connects the verbal and nonverbal number system also contributes variability to estimation performance. One reason to believe that variability and bias may arise in part from the mapping function between number language and the ANS is that feedback (e.g., showing a participant an example) reduces estimation variability in both children and adults (Barth, Starr, \& Sullivan, 2009; Krueger, 1984; Izard \& Dehaene, 2008; Lipton \& Spelke, 2005) a finding that one might not expect if variability arose entirely from the ANS. A second reason to believe that estimation variability stems entirely from the ANS is that sometimes - as in the data set we present in this paper - the coefficient of variation \((\mathrm{CoV})\) does not remain constant across all estimates. However, the degree to which estimation variability stems from the word-to-number mapping function remains untested.

Next, consider estimation bias. One frequent finding in the estimation literature is that estimates tend to be biased (e.g., systematically too high or too low). Also, bias in estimation performance tends to increase over the course of an experiment. For example, adults often underestimate magnitudes from the very first trial of an estimation experiment. When they do, this underestimation bias persists and is amplified over the course of the experiment (Krueger, 1982). In fact, even when the degree and direction of estimation error made
early in an experiment is experimentally manipulated, bias introduced in the first few trials endures throughout the duration of the entire estimation experiment (Barth et al., 2009; Izard \& Dehaene, 2008; Krueger, 1984; Lipton \& Spelke, 2005; Shuman, unpublished thesis; Sullivan \& Barner, 2012; Sullivan, Juhasz, Slattery, \& Barth, 2011). This influence of miscalibration is often described as stemming from changes to the number-to-magnitude mapping function. However, the nature of this change in the mapping function is still poorly understood. Specifically, it remains unknown how and why the degree of estimation bias increases as more estimates are made.

In the present study, we probed the factors that influence errors in estimation performance, with a special focus on the variability and bias found in individual participants estimates.


Figure 2: (Top) Participants saw 300 trials in which an array of \(n\) dots were briefly presented with the number of dots chosen according to a geometric distribution. Then participants made a guess as to the number of dots presented. (Bottom) A representative subject's data over all 300 trials with number presented (log scale) on the \(x\) axis and number reported ( \(\log\) scale) on the \(y\)-axis. We investigate the sources of bias and variability evident in these patterns of misestimation.

\section*{How do people estimate a quantity of dots?}

As already noted, patterns of bias and variability in estimation arise from both numerical perception and the mapping of these magnitude representations to language (see Figure 1). Our proposal is an amendment of Izard \& Dehaene's (2008) int that we argue that (1) the slope of the higher portion of this mapping function is variable over time, and (2) uncertainty in this mapping causes it to vary across time, thus introducing
additional variability in estimation tasks that increases with numerosity.


Figure 3: Individual subject estimation data (red points) along with best fitting linear (blue) and bilinear (green) mapping functions. Some of our conclusions may be seen in the raw data alone: (1) Variability in log-log space increases with numerosity. (2) Systematic mis-estimation occurs for larger, but not smaller, numbers. (3) Individual subjects have relatively stable idiosyncratic mis-estimation biases.

For our analyses, we consider this mapping to be bilinear in \(\log -\log\) space: it is veridical (falling on the identity line) for relatively small quantities (e.g., those smaller than 10; Sullivan \& Barner, 2012), but then tends towards underestimation for higher numbers. It is not central to our proposal that individuals actually use a strictly bilinear mapping function rather than a more complex mapping - however, our data do not offer the resolution necessary to assess whether a more sophisticated mapping function is used. For our purposes, approximating this function as bilinear makes our results and analyses simpler to describe.

This account offers a means to reconcile previous disagreements in the literature on the approximate number system and numerical estimation. First, while the coefficient of variation ( CoV , Weber fraction) is constant for the approximate num-
ber system, it is not known to be constant in verbal estimation tasks. For example, in previous work on word-to-magnitude estimation tasks, CoV has typically been analyzed only for a subset of the number-line (Izard \& Dehaehe, 2007; Le Corre \& Carey, 2007), or is found to increase with numerical magnitude (Siegler \& Opfer, 2003), or is not reported at all. In the present paper, we present a dataset in which the coefficient of variation in estimates increases with numerosity, suggesting a non-constant Weber fraction. Our account helps to reconcile a constant CoV in the approximate number system with an apparent increase in CoV in estimation, by showing that the coefficient of variation in estimation tasks is driven by variability in the mapping function over time (note that because discrimination tasks don't require verbal responses, they circumvent this mapping and its associated variability). Second, there is some disagreement as to whether there is a stable (consistent across the numerosity scale) mapping of magnitudes on to verbal numbers (Izard \& Dehaene, 2008; Lipton \& Spelke, 2005; Sullivan \& Barner, 2012). Our proposal suggests that, while the mapping of magnitudes to numbers may be consistent across a range of magnitudes at any one point in time, this mapping is not stable over time, yielding inconsistent behavior over trials.

Figure 1 provides an illustration of our model, and shows predictions from reduced classes of this model as compared to one (representative) subject's data. With only constant Weber noise and a stable linear mapping of magnitudes to numbers, we would erroneously predict overestimation and excessive variability for small numbers. Even with a bilinear mapping, constant Weber noise would predict less variability for large numbers and more for small numbers. Only with a variable slope in a bilinear mapping function can we account for the pattern of miscalibration and increasing variability for large, but not small, numbers. In the subsequent section we describe analyses that explicitly test these claims.

\section*{Experiment Methods}

Twenty-four subjects recruited from the UC San Diego psychology department pool participated in an hour-long experiment in which they had to guess the number of dots presented onscreen on each trial (see Figure 2). The number of dots shown was sampled on each trial from a geometric distribution with a mean of 50 , truncated at the low end so that displays had at least two dots. All the dots in an array were the same size (radius of 10 pixels), presented in red on a white background. The configuration of dots was randomly generated by drawing locations from a uniform distribution over the full display area ( \(1024 \times 768\) pixels) with the constraint that the dots did not overlap. On each trial the array of dots was presented for 250 msec , and then subjects were prompted to type in their guess as to how many dots were in the array. Subjects were then asked to type in a second guess about the number of dots in the array. (Our analyses throughout the paper focus on the first of the two guesses, but our conclusions hold if we consider the second guess alone, or the average
of the two). Figure 2 also shows one representative subject's data from all 300 trials.

\section*{Results}

The responses of all 24 participants for all 300 trials in our experiment are shown in Figure 3 on log-log coordinates. Several features of the data immediately jump out. First, estimate variability goes up as a function of number. Since this is an increase in variability in log-log space, it is not consistent with a constant coefficient of variation (a constant Weber fraction) which predicts that variability would be constant in \(\log -\log\) space. Second, while subjects are well calibrated for small numbers, there is a tendency to underestimate for large numbers: most subjects underestimated, but some subjects showed fairly reliable overestimation or veridical average calibration. Third, individual differences in under- or over- estimation appear to be quite reliable. These features are consistent with our account of a variable mapping function, which we elucidate in further analyses below.

\section*{Is magnitude-to-number mapping bilinear?}

We propose that the mapping function is not linear in loglog space, but bends such that small magnitudes are mapped more or less veridically onto number words, but large numbers show a systematic deviation from the identity line. While we do not believe that the true mapping function that people entertain is strictly bilinear, we do believe that it is not simply linear in \(\log -\log\) space. We can show that a bilinear function that is veridical (falls on the identity line) up to some critical number ( \(c\) ), and then deviates from the identity line with some log-log slope of \(s\) can account for data of individual subjects much better than a simple line with an intercept (a) and a slope ( \(b\) ). Since these models both have two parameters, we can simply compare the \(R^{2}\) values of individual subjects. Although the average \(R^{2}\) values are similar ( 0.79 , vs 0.81 ), bilinear fits better describe the data for 20 of 24 subjects (binomial test: \(p=0.0015\) ), see Figure 3.

This piecewise-linear mapping function could indicate a number of possible processes. Perhaps small numbers (less than about 10 - the average point of departure from veridical mapping across subjects) are not part of the mapping between the approximate number system and words, and instead gain their content from estimates made via subitization (e.g., in the company of chunking). Another possibility is that the mapping function is constrained by previous data which clearly disambiguate the numerosity/numbers correspondence, and that lower numbers have more data, and thus fall on the identity line, while higher numbers are constrained only by a requirement for smoothness and monotonicity. A final possibility is that the nature of the mapping function between the ANS and the verbal number system is qualitatively different for relatively small numbers and relatively larger numbers for example, participants might rely more strongly on itembased associative mappings for numbers smaller than 10, but more on a structural mapping between magnitudes and the count list for larger numbers (Sullivan \& Barner, 2012). We


Figure 4: We assess whether variability arises from a slow drift in the mapping function over time by estimating the slope of a bilinear mapping function for different subsets of the 300 trials for each subject. For instance, if we split the 300 trials into thirds (left half), then each third contains 100 trials, if we split into thirtieths (right half) then each thirtieth contains 10 trials. A blocked split (top half) corresponds to taking consecutive portions of the 300 trials: e.g., the first 3rd contains trials 1 through 100, the second 3rd contains trials 101 through 200, the third 3rd contains trials 201 through 300. A modular split (bottom half), corresponds to taking every \(n\)th trial, such that the full range of the experimental session is represented in each subset: e.g., the first 3 rd contains trials \(1,4,7,10, \ldots 298\), the second 3rd contains trials \(2,5,8,11, \ldots, 299\), and the third 3rd contains trials \(3,6,9,12, \ldots, 300\). We compute the across-subject correlation of slope estimates taken from each subset of trials. Darker colors indicate lower correlations, brighter colors indicate larger correlations. Several observations in these heat maps are indicative of a gradual drift in slopes over time within a given subject. (1) Modular splits yield higher across-subject correlations than blocked splits, suggesting that the blocked splits are subject to additional variability due to a gradual change that modular splits avoid. (2) Blocked splits show decreasing correlations as a function of distance: correlations further from the diagonal are lower - the correlation between the first and second third is higher than the correlation between the 1st and 3rd third - indicating that the correlations are slopes are changing slowly over time.
slightly disfavor the first alternative, because the cut-off point between accurate and miscalibrated mapping does not seem to correspond to other cut-offs previously postulated to distinguish between qualitatively different numerosity processes (such as subtilizing and approximate magnitude - Feigenson et al., 2004).

\section*{How reliable is the across-subject variation in the shape of the mapping function?}

We assess the across-subject reliability in shape of the bilinear mapping function via a split-half analysis: we estimate the mapping function (particularly the slope of the higher bilinear portion) in individual subjects in \(50 \%\) of the trials and assess the across-subject correlation across those pairs of estimates, to see whether variation of mapping functions is reliable.

First we assess Blocked split-half reliability: we divide the 300 trials into the first half (trials 1-150) and the second half (151-300). The Blocked split-half across-subject reliability of the estimate of the slope of the erroneous part of the mapping
function that does not fall on the identity line was highly significant ( \(r=0.83 ; t(22)=6.98, p<0.001\) ), indicating that people are very consistent in their idiosyncratic magnitude-to-number mapping errors.

We next assessed Modular split-half reliability: Modular split-half divides the 300 trials into odd trials (1, 3, 5, ..., 299), and even trials ( \(2,4,6, \ldots, 300\) ). In contrast to Blocked splits, in which trials in a given half of the data are are contiguous and arise from from different portions of the experiment (separated by an average of 150 trials), in Modular splits, the trials in a given half are taken from the full range of the experiment. Modular splits-half across-subject reliabilities were much higher than those for Blocked split-half analyses \((r=0.97 ; t(22)=18.7, p<0.001-\) the difference is highly significant using the Fisher r-to-z transform: \(z=-2.74, p=0.0061)\).

The difference between Modular and Blocked split-half reliability is our first indication that the slope of the magnitudenumber mapping function is not stable within individuals over the experimental session: If the slopes we estimate for the mapping function drift over time, we expect that Blocked splits should yield a lower split-half across-subject reliability than Modular splits, because the Blocked splits are taken from different points in time, and would reflect different states of drift of the mapping function, while Modular splits would not.

\section*{Does the mapping function vary over time?}

We argue that some of the increase in variability of estimates with increasing number arises from variability of the mapping function over trials, rather than simple misperception of the approximate magnitude of an individual array. In this section we argue for this view because the internal mapping function drifts slowly over the course of many trials, and we can measure its variation over the course of an experimental session.

To more precisely measure the drift of the mapping function over time, we generalize the Blocked vs. Modular splithalf analysis to Blocked vs. Modular split- \(n\) ths for \(n=\) \(\{3,5,10,15,20,30\}\) (e.g., for split-30th we divide our 300 trials into 30 subsets, each one comprising 10 trials, for instance, the 5th Blocked split-30th subset will contain trials 41-50, while the 5th Modular split-30th subset will contain the 10 trials: \(5,35,65,95,125,155,185,215,245,275)\). By obtaining split- \(n\)th reliability for Blocked subsets taken from different portions of the experimental session, we can assess how the reliability of the number-mapping slope decreases as a function of time.

We calculate the across-subject slope reliability across different subsets (represented as a matrix in Figure 4), the Blocked split- \(n\)th reliability between subset 1 and subset 2 measures the across-subject correlation of slopes estimated from two adjacent periods of time in the session which are on average separated by \(300 / n\) trials. In general if we calculate the correlation between subset \(i\) and subset \(i+k\) from a Blocked split- \(n\)th analysis, those subsets are separated by \(300 * k / n\) trials. Thus, if slopes are gradually drifting over the
course of the experimental session, we would expect acrosssubject reliability of slope estimates to decrease with \(k\) - the separation between Blocked subsets. Nothing of this sort should happen for Modular subsets which contain overlapping trials intermixed over the whole session.


Figure 5: (Top) We can assess the average across-subject correlation in slope estimates as a function of distance between blocks for different splits of the data. A blocked split of the data into thirds yields a two measures of the slope-correlation at an average distance of 100 trials (1st to 2nd block and 2nd to 3rd block) and one measure of the correlation at a distance of 200 trials ( 1 st to 3rd block). If we split the data more finely (here going up to 30ths, as seen in figure , we can more finely measure the drop-off of average correlation as a function of distance. Distance is meaningless for Modular (points in black) splits, but the same analysis can be carried out to measure how much the correlation drops merely as a function of using fewer trials for each slope estimate. (Bottom) we can normalize both blocked (red) and modular (black) correlations to the average of the modular split correlations to make the decrease in correlation with distance as seen in the red lines comparable across splits with different numbers of trials per subset. Correlations drop off with distance very slowly, but even when separated by 290 trials, 10 -trial estimates of subjects' mapping function slopes show reliabilities well above 0 .

Figure 5(top) shows the split- \(n\)th reliability for Blocked (red) and Modular (black) subsets as a function of their separation ( \(k\) ). For instance, we estimate the average Blocked split-10th correlation at a separation of \(k=2\) as the average of the across-subject correlations taken between the 8 subset comparisons separated by 2 : subset 1 and subset 3 , subset 2 and subset \(4, \ldots\) subset 8 and subset 10 . This average would appear at \(x=300 / n * k=300 / 10 * 2=60\). Several features are apparent from the changes in slope reliability across sub-
sets separated by more time: (1) when we split into more subsets both Modular and Blocked correlations drop, since each subset necessarily contains fewer trials to estimate the slope; (2) as expected, only Blocked correlations decrease as a function of distance between subsets. To more clearly display the decrease in reliabilities as a function of subset distance, while factoring out the reduced reliability due to smaller trialcounts within each subset, by normalizing reliabilities by dividing them by the average reliability seen across Modular splits. This yields Figure 5(bottom), which shows the slow decrease in reliabilities over the 300 trials.

A linear regression on the raw correlations in the Blocked split- \(n\) ths as a function of separation (measured in trials) is significantly negative ( \(95 \%\) confidence interval on the slope: \((-0.0015,-0.0012)\) change in correlation per trial, \(F(1,22)=358, p<0.001)\). Despite this highly significant decrease, it is very slow over the course of the session, and even mapping function slope estimates based on 10 trials separated by 290 trials show significant across-subject reliability ( \(r=0.57 ; t(22)=3.254, p=0.002\) ).

Together, these results indicate that subjects' mappings of magnitudes onto verbal numbers drift slowly over time.

\section*{Conclusions}

We have shown that subjects map numbers onto the verbal number line via a piecewise-linear function in logarithmic representations (piecewise consistent with Stevens' power law). Our results are consistent with two-transformations mapping physical numbers onto number estimates: first physical numbers are represented logarithmically in the approximate number system, and then the approximate number system is mapped onto the verbal number line through an unstable mapping. For small numbers, the mapping appears to be fairly stable and veridical, perhaps due to the considerable amount of evidence people have previously seen for small number estimates. For higher numbers, the mapping is not veridical, and tends to drift slowly over the course of many trials; the variability of the mapping function over time causes increasing estimation variance for large numbers, and may thus resolve theoretical disagreements about the constancy of variability in the approximate number system.
Acknowledgments: EV was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of the Interior (DOI) contract D10PC20023. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI, or the U.S. Government. DB was supported by a grant from the James S. McDonnell Foundation. JS was supported by an NSF GRFP fellowship.

\section*{References}

Barth, H., Starr, A., \& Sullivan, J. (2009). Childrens mappings of large number words to numerosities. Cognitive Development, 24, 248-264.
Carey, S. (2009). The Origin of Concepts. Oxford University Press: New York.
Dehaene, S. (1997). The Number Sense: How the Mind Creates Mathematics. Oxford University Press: New York.
Dehaene, S., \& Marques, J. (2002). Cognitive euroscience: Scalar variability in price estimation and the cognitive consequences of switching to the euro. The Quarterly Journal of Experimental Psychology, 55, 705-731.
Ebersbach, M., Luwel, K., Frick, A., Onghena, P., Verschaffel, L. (2008). The relationship between the shape of the mental numberline and familiarity with numbers in 5- to 9 -year old children: Evidence for a segmented linear model. Journal of Experimental Child Psychology, 99, 1-17.
Izard, V., \& Dehaene, S. (2008). Calibrating the mental number line. Cognition, 106, 1221-1247.
Kaufman, E., Lord, M., Reese, T., \& Volkmann, R. (1949). The discrimination of visual number. The American Journal of Psychology, 62, 498-525.
Krueger, L. (1982). Single judgments of numerosity. Perception and Psychophysics, 31, 175-182.
Krueger, L. (1984). Perceived numerosity: A comparison of magnitude production, magnitude estimation, and discrimination judgments. Perception and Psychophysics, 35, 536-542.
Le Corre, M., \& Carey, S. (2007). One, two, three, four, nothing more: An investigation of the conceptual sources of the verbal counting principles. Cognition, 105, 395-438. 9, 159-172.
Lipton, J., \& Spelke, E. (2005). Preschool childrens mapping of number words to nonsymbolic numerosities. Child Development, 76, 978-988.
Minturn, A., \& Reese, T. (1951). The effect of differential reinforcement on the discrimination of visual number. Journal of Psychology, 31, 201-231.
Negen, J., \& Sarnecka, B. (2010). Analogue magnitudes and knower-levels: Re-visting the variability argument. Proceedings of the 32 nd Annual Conference of the Cognitive Science Society, Austin, TX: Cognitive Science Society.
Ramani, G., \& Sieger, R. (2011). Reducing the gap in numerical knowledge between low- and middle-income preschoolers. Journal of Applied Developmental Psychology, 32, 146-159.
Shuman, M. (unpublished thesis). Computational characterization of numerosity perception and encoding.
Siegler, R., \& Opfer, J. (2003). The development of numerical estimation: Evidence for multiple representations of numerical quantity. Psychological Science, 14, 237-243.
Siegler, R., \& Ramani, G. (2009). Playing linear number board gamesbut not circular onesimproves low-income preschoolers numerical understanding. Journal of Educational Psychology, 101, 545-560.
Sullivan, J., \& Barner, D. (2012). How are number words mapped to approximate magnitudes? Quarterly Journal of Experimental Psychology, 66, 389-402.
Sullivan, J., Juhasz, B., Slattery, T., \& Barth, H. (2011). Adults number-line estimation strategies: evidence from eye movements. Psychonomic Bulletin and Review, 18, 557-563.
Thompson, C., \& Opfer, J. (2010). How 15 hundred is like 15 cherries: Effect of progressive alignment on representational changes in numerical cognition. Child Development, 81, 1768-1786.
Whalen, J., Gallistel, C., \& Gelman, R. (1999). Nonverbal counting in humans: the psychophysics of number representation. Psychological Science, 10, 130-137.

\title{
Effects of Parafoveal Plausibility During Reading
}

\author{
Laura Wakeford (l.j.wakeford@dundee.ac.uk) \\ School of Psychology, University of Dundee \\ Dundee, DD1 4HN
}

\author{
Wayne Murray (w.s.murray@dundee.ac.uk) \\ School of Psychology, University of Dundee \\ Dundee, DD1 4HN
}

\begin{abstract}
There is controversy concerning the question of whether meaning can be extracted from a parafoveal word during reading and whether this might occur in an overlapping fashion with the lexical processing of the currently-fixated word. We suggest that previous attempts to investigate this have been bedevilled by problems associated with the use of priming methodology. Instead, we used an eye movement contingent change technique and manipulated the plausibility of the parafoveal preview, resulting in it being either valid, a plausible alternative, anomalous, or an illegal letter string. The results showed (a) a meaning-based parafoveal-on-foveal effect, (b) preview benefits driven by both orthographic and semantic influences, and (c) continuing disruption associated with orthographically dissimilar previews. We suggest that this pattern is most consistent with models of eye movement control that allow for distributed attention during reading.
\end{abstract}

Keywords: Eye movements; preview benefit; plausibility; parafoveal-on-foveal effects; boundary technique; reading.

\section*{Introduction}

The nature of Preview Benefit (PB) - the advantage accruing to the reader from an accurate parafoveal preview of the following word - critically informs our understanding of the reading process, indicating those features which are, and are not, extracted from an as yet unfixated word. Using the boundary paradigm (Rayner, 1975), many studies have shown that both orthographic and phonological features appear to be extracted from parafoveal words (see Schotter, Angele \& Rayner, 2012, for a review); however, evidence for a semantic PB remains controversial (see Radach \& Kennedy, 2013; Rayner, White, Kambe, Miller \& Liversedge, 2003).

While both serial (e.g., E-Z Reader; Reichle, Warren, \& McConnell, 2009) and parallel (e.g., SWIFT; Engbert, Nuthmann, Richter \& Kliegl, 2005) models of eye movement control during reading provide accounts of orthographic and phonological PB, only parallel models appear capable of accounting for semantic PB. In serial models, lexical processing is restricted to one word at a time, with attention moving to the parafoveal word only when the currently fixated word has been fully identified. Serial models therefore typically only accommodate very early stages of word recognition occurring on parafoveal words before a saccade remarries fixation location with attention. In contrast, in parallel models, all words within
the perceptual span can be processed simultaneously, up to and including the level of semantic processing.

Studies investigating semantic PB have typically manipulated the sematic relatedness of the preview and the target word, on the basis that responses to semantically related word pairs are facilitated compared to unrelated pairs (Meyer \& Schvaneveldt, 1971). By extension, it is suggested that semantically related previews should facilitate target viewing compared to unrelated previews. Using the boundary paradigm, Rayner Balota and Pollatsek (1986) asked participants to read sentences such as "My younger brother has brilliantly composed a new song for the school play", in which the pre-fixation preview of "song" was either "song" (valid), "tune" (related), "door" (unrelated), or "sorp" (a visually similar nonword). Only once the eye passed an invisible boundary, located before the critical word, did the target word "song" appear. Despite showing that their critical words produced facilitation in a classic priming experiment, Rayner et al found no evidence for a semantic PB during reading. However, in this example sentence, we see that the word to the left of the target contains only three letters, and as short words are frequently skipped (Rayner \& McConkie, 1976), the prior fixation may in fact have fallen two words to the left of the target, seriously reducing the chance of it eliciting a semantic PB.

A more general problem with experiments investigating semantic PB using associative previews is that while there may be semantic facilitation, there is also a word change that might be expected to give rise to some form of inhibition. Semantically related word pairs, such as northsouth, rattle-bottle and arms-legs, (from Rayner et al, 1986), have very different meanings, and this could exert an inhibitory effect on on-going sentence interpretation. Rayner et al (1986) attempted to test this possibility by asking participants to rate their sentence pairs for similarity of meaning and reanalysing the results from only the 20 sentence pairs rated as most similar in meaning. Since this analysis again failed to show a semantic PB, they dismissed this as an explanation for their null result. However, a measure of overall sentence meaning does not necessarily capture the extent to which a local change in word meaning might have disrupted the reading process at the point at which it occurred. We conclude, therefore, that interference resulting from word change remains a possibility.

Altarriba, Kambe, Pollatsek and Rayner (2001) used a variant of this technique with fluent English-Spanish bilinguals. They employed semantically related previews which were translations with virtually the same meaning as the targets, thereby reducing the possibility of interference. All changes involved a word preview from the other language that was either: cognate (orthographically and semantically similar), noncognate (semantically similar but orthographically dissimilar), pseudocognate (semantically unrelated but orthographically similar), or control (unrelated orthographically and semantically). They found no evidence for a semantic PB in the absence of orthographic similarity. However, it remains possible that facilitation might not cross over between the lexica of the two languages, and as Hohenstein, Laubrock and Kliegl (2010) point out, since the previews and targets were in different languages, switching costs (Meuter \& Allort, 1999) might mask any semantic PB.

Hohenstein et al (2010) suggest that the elusive nature of semantic PB may result from a lack of control over the preview duration of masked words. To test this possibility, they used fast priming (Sereno \& Rayner, 1992) and the boundary paradigm in a novel way. Prior to landing on the pretarget word, the target was masked with an illegal letter string. Once the eye landed on the pre-target word, this illegal nonword preview changed to either a semantically related or unrelated prime for durations of 35 ms , 80 ms or 125 ms , after which point, the target was displayed. Hohenstein et al report a significant semantic PB only when the prime duration was 125 ms (Experiments \(1 \& 2\) ). When primes were presented in bold typeface, a significant effect was observed with 80 ms primes, although, the 125 ms effect was no longer significant (Experiment 3). Given the transient nature of the effects and variability across experiments, these results would benefit from replication.

An alternative way to explore when the meaning of a word becomes available is to investigate plausibility effects. For example, Murray and colleagues (Kennedy, Murray \& Boissiere, 2004; Murray, 2006; Murray, 1998; Murray \& Rowan, 1998) recorded eye movements in a series of experiments where plausibility was manipulated. Participants read a sentence and then pressed a button, triggering another sentence to be displayed; the task being to indicate whether the two sentences were the same or different. These studies showed effects of the plausibility of the combination of the initial noun phrase with the verb, for example, "The hunters stacked...." vs "The bishops stacked...", and in a number of the studies, this was reflected not only in fixations falling on the verb, but also in some eye movement measures before the verb was directly fixated, suggesting the extraction of meaning from words in the parafovea. However, Rayner et al (2003) report being unable to replicate one of Murray et al's findings in a reading study and suggest that their results may have been task specific.

Starr and Inhoff (2004; Experiment 1) also investigated the consequences of providing a contextually inappropriate word to the right of fixation. They masked a critical word
with itself, a contextually inappropriate word, or a legal or illegal nonword. In addition to finding clear orthographic parafoveal-on-foveal effects, a trend also emerged in which a contextually inconsistent word in the parafovea reduced gaze duration by 22 ms on the pre-target word compared to an accurate preview. However, a subsequent analysis excluding the \(45 \%\) of cases where fixations fell near the ends of the pre-target words (possibly as a result of oculomotor error) showed no reliable effect. The contextually inconsistent preview also gave rise to inflated fixation times when the target was fixated, but this effect could be attributed to a lack of orthographic overlap, rather than any extraction of parafoveal meaning.

While plausibility related parafoveal-on-foveal effects remain controversial, it is widely accepted that the plausibility of words within a sentence can have an immediate impact on fixation durations falling on the word. For example, Rayner, Warren, Juhasz and Liversedge (2004) presented participants with a series of sentences in which a critical noun was either plausible (likely), implausible (unlikely) or anomalous (inappropriate), given the preceding sentence context. They found that anomalous words had an immediate impact on gaze duration, while effects of implausibility were reflected only in later measures. Interestingly, they also discovered a plausibilityrelated parafoveal-on-foveal effect, with gaze duration on the word preceding the anomalous one being 17 ms and 14 ms longer than in the control and implausible conditions, respectively. The authors, however, attribute this effect to oculomotor error.

Whether or not one questions the interpretation of apparent semantic parafoveal-on-foveal effects, it is clear that manipulating the plausibility of a word can produce robust effects on the reading pattern when that word is fixated. This study capitalised on that finding and presented participants with sentences in which a critical word ( \(\mathrm{n}+1\) ) was masked prior to receiving a direct fixation by either a (a) valid (identical), (b) plausible but different, or, (c) anomalous word, or (d) an illegal nonword. Once the eye passed an invisible boundary located before word \(_{n+1}\), all previews were replaced with the valid preview. If meaning is extracted from the parafovea, it would be expected that an anomalous preview should exert an immediate impact on word \(_{n+1}\) fixations compared to the plausible preview condition. Conversely, if the meaning of the parafoveal word is not extracted while fixating word \(_{n}\), then plausible and anomalous previews should both produce the same cost, as a result of their lack of orthographic overlap with the target. The illegal nonword served as a baseline against which the magnitude of PB could be judged.

\section*{Method}

\section*{Participants}

Twenty-eight native English speakers with normal or corrected to normal vision took part for course credits or \(£ 5\) payment.

\section*{Materials and Design}

Ninety-six experimental sentences were constructed. As can be seen in the example below, each contained a critical word pair, comprising a 6 letter verb \(\left(\right.\) word \(\left._{n}\right)\) followed by a 6 or 7 letter noun \(\left(\operatorname{word}_{n+1}\right)\). To facilitate processing, word \(_{n}\) was always high frequency (mean 135 occurrences per million, by Kucera \& Francis, 1967). Word \(_{n+1}\) was assigned one of four pre-fixation previews, all chosen to be very low in predictability: valid (e.g. "dinner" - identical), plausible (e.g., "coffee" - an alternative that fitted the preceding context), anomalous (e.g. "caught" - a word that produced a semantic or grammatical violation), or an illegal nonword (e.g., "fumeio" - a letter string containing combinations not found in the English dictionary, in this case "eio"). The frequency of these three preview words did not differ (means 132, 144 and 140 per million respectively, all \(t s<1\) ). Previews were displayed until the eye passed an invisible boundary located prior to the space before word \(_{n+1}\), shown below with a "|". When the eye crossed this boundary, the target word was then displayed.
The mother was making| dinner in the kitchen for her two children and her husband.

Plausibility ratings were provided by 12 participants, who did not take part in the eye tracking experiment. Participants rated all three versions of each sentence up to and including word \(_{n+1}\) (illegal letter strings were not included). They used a rating scale from 1 (low) to 7 (high) plausibility. Additionally, they could use "U" instead of providing a numerical rating if they felt the sentences were ungrammatical; " \(U\) " scores were coded as 0 for purposes of analysis. The valid and plausible fragments were both rated as highly plausible (means \(=6.3\) and 6.2 , respectively), with no significant difference between these conditions \((t(95)=.85, p=.40)\). The mean rating for the anomalous condition was 1.0 which differed significantly from both the valid \((t(95)=44.63, p<.001)\), and plausible conditions \((t(95)=45.90, p<.001)\).

Four counterbalanced item lists were created, each containing the 96 experimental sentences together with 19 filler and 8 practice items and 20 comprehension questions. Word \(_{n+1}\) previews initially masked the target, with each item list containing an equal number of valid, plausible, anomalous and illegal nonword previews. Each participant was presented with the items in a different randomised order.

\section*{Apparatus}

A dental composition bite bar and chin rest were used to minimise head movements. Sentences were displayed on a VDU screen in white monopitch font on a black background. At a viewing distance of 50 cm , each character subtended approximately 0.3 degrees of a visual angle. Reading was binocular, but only eye movements from the right eye were recorded using a "Dr. Bouis" infrared pupil-
centered computation device sampled with a 12 -bit A-D at 2 ms intervals. The apparatus was calibrated after every fourth sentence. In order to answer questions, right-hand ("yes") and left-hand ("no") button boxes were provided.

\section*{Procedure}

Verbal and written instructions were provided. Calibration consisted of looking at a series of horizontally aligned numbers. Once optically set up and calibrated, participants fixated a cross, which after receiving a 100 ms stable fixation, triggered display of a sentence. Participants were asked to read for comprehension, but not to adjust their usual reading style. If presented with a question, they were asked to respond using the button boxes.

\section*{Results and Discussion}

Data were analysed treating both participants \(\left(F_{1}\right)\) and items \(\left(F_{2}\right)\) as random variables. In all analyses item file was treated as a between-groups dummy factor. Analysis focused on the pretarget word (n), the target word ( \(\mathrm{n}+1\) ) and a spillover region comprising the following three words. A number of eye movement measures are reported for each region: the duration of the first and last fixations, gaze duration (summed duration of all fixations until the eye exits the region in either direction), go-past time (summed duration of all fixations, including regressions until the following region is first fixated) and first pass skipping probability. Participants clearly read carefully, with \(86 \%\) overall accuracy on the comprehension questions.

\section*{Effects of Word \({ }_{n+1}\) Preview on Word \(_{n}\)}

The probability of fixating word \(_{n}\) did not vary across conditions (both \(F \mathbf{s}<1\) ) and first fixation duration, gaze and go-past time showed no reliable effect of preview type (all \(F \mathrm{~s}<1.3\) ). A trend did however emerge in last fixation duration \(\left(F_{l}(3,72)=2.81, p<.05 ; F_{2}(3,276)=2.05, p=.11\right)\) with shorter fixations when the preview was either an illegal nonword or an anomalous word. While first and last fixation durations constitute overlapping sets, the two measures potentially tap into differing processes. It is not surprising, therefore, to find a trend emerging statistically only for those fixations which were the last to fall on the word. For this measure, pairwise comparisons showed no significant difference between the valid and plausible conditions ( \(F_{1}\) \(\left.(1,24)=2.07, p=.16, F_{2}(1,92)=2.21, p=.14\right)\), however, when these two conditions were combined and compared to the anomalous condition, durations were reliably shorter in the anomalous condition \(\left(F_{1}(1,24)=5.30, p<.05, F_{2}(1,92)\right.\) \(=4.29, p<.05)\), suggesting that this preview attracted attention from word \({ }_{n}\). The same trend emerged when an illegal nonword fell to the right of fixation, although the difference between the combined valid and plausible conditions and the illegal nonword preview was only reliable over subjects \(\left(F_{1}(1,24)=4.54, p<.05, F_{2}(1,92)=\right.\) 1.10, \(p=.30\) ). The pattern overall, however, suggests that readers respond similarly to both orthographic illegality and
anomaly to the right of fixation. This speed-up seems similar to the finding by Starr and Inhoff (2004) of a trend towards shorter gaze durations on the pre-target word when the target preview was contextually inconsistent.

Table 1: Fixation times (ms) and skipping probability (\%) on word \({ }_{n}\) as a function of word \(_{n+1}\) preview
\begin{tabular}{lccccc}
\hline & FFD & LFD & Gaze & Go-Past & Skip \\
\hline Valid & 257 & 257 & 274 & 301 & 8 \\
Plausible & 263 & 264 & 280 & 305 & 8 \\
Anomalous & 255 & 253 & 278 & 305 & 8 \\
Illegal & 258 & 253 & 268 & 296 & 10 \\
\hline
\end{tabular}

Interestingly Rayner et al (2004) reported an effect of parafoveal anomaly on word \({ }_{n}\) in the opposite direction, with longer fixations when there was an anomalous word to the right of fixation. They concluded this must result from mislocated fixations, with the reader staying and processing word \(_{n+1}\) from a sub-optimal parafoveal location. But a reduction in fixation duration does not permit the same interpretation. These results and the pattern apparent in Starr and Inhoff's contextually inconsistent condition appear more consistent with Kennedy's \((1998,2000)\) attractor hypothesis, in which something unexpected in the periphery attracts attention, resulting in shorter fixation durations on the preceding word. Parafoveal-on-foveal effects have frequently been reported as a consequence of orthographic peculiarities to the right of fixation; however, here we see it, to an equivalent extent, as a consequence of meaning.

\section*{Word \(_{n+1}{\text { Preview Effects on } \text { Word }_{n+1}}^{1}\)}

There was no reliable effect of preview on the skipping of \(\operatorname{word}_{n+1}\left(F_{1}(3,72)=1.50, p=.22 ; F_{2}(3,276)=1.70, p=.16\right)\). There was, however, a consistent effect of prior preview in all durational measures: first \(\left(F_{1}(3,72)=7.43, p<.001 ; F_{2}\right.\) \((3,276)=5.81, p<.01\) ), and last fixation durations ( \(F_{1}(3,72)\) \(\left.=4.69, p<.01 ; F_{2}(3,276)=3.5, p<.05\right)\), gaze duration \(\left(F_{1}\right.\) \(\left.(3,72)=8.87, p<.001 ; F_{2}(3,276)=8.71, p<.001\right)\), and gopast time \(\left(F_{1}(3,72)=7.65, p<.001 ; F_{2}(3,276)=9.29\right.\), \(p<.001\) ). As can be seen from Table 2, the longest durations were associated with words previously masked by an illegal nonword, followed by anomalous then plausible previews, with valid previews associated with the shortest durations.

Table 2: Fixation times (ms) and skipping probability
(\%) on word \({ }_{n+1}\) as a function of word \(_{n+1}\) preview
\begin{tabular}{lccccc}
\hline & FFD & LFD & Gaze & Go-Past & Skip \\
\hline Valid & 267 & 267 & 298 & 331 & 5 \\
Plausible & 277 & 275 & 299 & 345 & 6 \\
Anomalous & 281 & 277 & 313 & 370 & 7 \\
Illegal & 287 & 283 & 331 & 389 & 4 \\
\hline
\end{tabular}

Pairwise comparisons revealed that a preview of an illegal nonword increased fixation durations across all measures: first \(\left(F_{1}(1,24)=47.02, p<.001 ; F_{2}(1,92)=18.06, p<.001\right)\);
and last fixation durations ( \(F_{1}(1,24)=21.87, p<.001 ; F_{2}\) \((1,92)=14.97, p<.001)\); gaze duration \(\left(F_{1}(1,24)=23.00\right.\), \(\left.p<.001 ; F_{2}(1,92)=23.68, p<.001\right)\) and go-past time ( \(F_{1}\) \(\left.(1,24)=15.69, p<.001 ; F_{2}(1,92)=26.67, p<.001\right)\).

Readers also appear to have noticed the change from the plausible preview to the valid target, reflected in a reliable increase in first fixation duration ( \(F_{1}(1,24)=5.66, p<.05\); \(\left.F_{2}(1,92)=5.00, p<.05\right)\). However, the similar trends in last fixation duration and go-past time failed to achieve statistical significance \(\left(F_{1}(1,24)=3.22, p=.08 ; F_{2}(1,92)=\right.\) \(2.62, p=.11\) and \(F_{1}(1,24)=3.70, p=.06 ; F_{2}(1,92)=1.73\), \(p=.19\), respectively) and there was no effect in gaze duration (both \(F \mathbf{s}<1\) ). Overall, this pattern suggests that the change from a different, though plausible, word was noticed immediately and resulted in an increased probability of regressing, as reflected in go-past time, but not in gaze duration, since this is terminated by the regressive movement. It is not clear, however, whether this is an effect of meaning change, since it could equally be a consequence of the lack of orthographic overlap between the plausible preview and target.
A test of the effect of meaning can, however, be found in the contrast between plausible and anomalous previews, since both involve a change in orthography. The results here suggest that the meaning of word \(_{n+1}\) was indeed extracted while fixating word \({ }_{n}\). While both first and last fixation duration showed no evidence of an increased cost of anomaly (all \(F \mathrm{~s}<1\) ), the 14 ms increase in gaze duration was significant by-subjects and approached significance byitems \(\left(F_{1}(1,24)=4.78, p<.05 ; F_{2}(1,92)=2.99, p=.08\right)\) and with regressions taken into account, the 25 ms increase in go-past time was significant by-subjects and very close to significant by-items ( \(F_{1}(1,24)=5.88, p<.05 ; F_{2}(1,92)=\) \(3.51, p=.06\) ).
As can be seen in Table 3, an increase in go-past time following anomalous previews also arose in the spillover region (see below). Combining word \(_{n+1}\) and the spillover regions, the difference in go-past between the plausible ( 898 ms ) and anomalous ( 946 ms ) conditions was significant by both subjects and items ( \(F_{1}(1,24)=7.29, p<.05 ; F_{2}\) \((1,92)=6.17, p<.05)\). This effect - an immediate and robust slowing in the anomalous condition - is in the expected direction based on, for example, the findings of Rayner et al (2004), who suggest that anomalous words "hit the reader over the head" (p. 1297). It seems from these results however, that the genesis of this effect can be parafoveal, with the reader detecting anomaly far earlier than previously thought.
It could be suggested that these results stem from word \({ }_{n}\) receiving full lexical access, allowing an attention shift to word \(_{n+1}\) which also received full lexical access and semantic interpretation, all prior to word \(_{n+1}\) being fixated. While possible, this seems extremely unlikely with both the low predictability and length of word \({ }_{n+1}\) conspiring against such rapid parafoveal identification.
These results strongly suggest that parafoveal preview effects are not restricted to the extraction of orthographic
and phonological features, but rather, that higher level linguistic processing can be engaged when previewing words to the right of fixation, and when the input changes, as happened here, this interferes with later comprehension.

\section*{Word \(_{n+1}\) Preview Effects in the Spillover Region:}

As shown in Table 3, the spillover region was rarely skipped, with little difference between the preview conditions (both \(F \mathbf{s}<1\) ). Both first and last fixation durations were unaffected by word \({ }_{\mathrm{n}+1}\) preview type (all \(F \mathbf{s}<1\) ), as was gaze duration \(\left(F_{1}(3,72)=1.56, p=.21 ; F_{2}(3,276)=1.62\right.\), \(p=.18\) ). There was, however, a highly significant effect of word \(_{\mathrm{n}+1}\) preview on go-past time \(\left(F_{1}(3,72)=6.98, p<.001\right.\); \(\left.F_{2}(3,276)=10.48, p<.001\right)\), with higher durations when word \(_{n+1}\) had been changed.

Table 3: Fixation times (ms) and skipping probability (\%) in the spillover region as a function of word \({ }_{n+1}\) preview
\begin{tabular}{lccccc}
\hline & FFD & LFD & Gaze & Go-Past & Skip \\
\hline Valid & 256 & 247 & 465 & 505 & 1 \\
Plausible & 254 & 249 & 480 & 553 & 0 \\
Anomalous & 254 & 254 & 486 & 576 & 1 \\
Illegal & 255 & 250 & 485 & 541 & 1 \\
\hline
\end{tabular}

Pairwise comparisons show go-past time significantly higher following an illegal nonword compared to a valid preview \(\left(F_{1}(1,24)=7.65, p<.05 ; F_{2}(1,92)=5.99, p<.05\right)\). This finding is difficult to reconcile with models such as EZ Reader in which orthographic extraction occurs during the first stage of lexical processing on a word \(\left(\mathrm{L}_{1}\right)\). \(\mathrm{L}_{1}\) can commence on word \({ }_{\mathrm{n}+1}\) if the word is parafoveally available while fixating word \({ }_{n}\). However, if preview is denied, \(L_{1}\) is delayed until word \({ }_{n+1}\) is fixated, resulting in the standard \(\operatorname{word}_{n+1}\) PB. But as soon as \(L_{1}\) is complete, a saccade is programmed to word \(_{\mathrm{n}+2}\), at which point the second stage of lexical processing ( \(\mathrm{L}_{2}\) ) commences. Since it is postulated that this stage follows orthographic extraction, the time for attention to shift to word \(_{n+2}\) should only ever be a function of later linguistic processing and not delayed by difficulties with orthographic extraction. According to the E-Z Reader model, therefore, it should never be the case that fixation durations are inflated after word \(_{n+1}\) has been passed - unless of course they are 'mislocated', but in that case the response should be to stay and process, rather than to regress, as seen here.

While the 48ms increase following a plausible compared to a valid preview was significant \(\left(F_{1}(1,24)=20.68\right.\), \(p<.001 ; F_{2}(1,92)=17.53, p<.001\), the 23ms increase following an anomalous compared to a plausible preview was not \(\left(F_{1}(1,24)=1.70, p=.21 ; F_{2}(1,92)=2.47, p=.12\right)\). Somewhat surprisingly, spillover effects relating to a lack of orthographic overlap between preview and target appear to show a longer time course than effects of parafoveal meaning, with the latter exerting a more immediate impact, mostly reflected in the cumulative duration measures on word \(_{n+1}\).

\section*{General Discussion}

This study set out to investigate whether PB is restricted to the orthographic and phonological properties of a parafoveal word or whether meaning can also be extracted before a word is directly fixated. Prior research has failed to show much evidence for a semantic PB, but we suggest this may be due to the use of semantically related previews, with the word change interfering with target processing, rather than facilitating the recognition process. By varying the plausibility of previews, we have found two important semantic effects.

First, we found evidence that an anomaly in the parafovea can attract attention, resulting in shorter fixation durations on pre-target words. While others have reported anomalyrelated parafoveal-on-foveal effects (Kennedy, Murray \& Boissiere, 2004; Murray, 1998; Murray \& Rowan, 1998; Rayner et al, 2004; Starr \& Inhoff, 2004), this appears to be the first study to find a reliable effect that cannot be attributed to a potentially mislocated fixation. As Liversedge, Paterson and Pickering (1998) point out, when faced with difficulty the reader has three options: (a) stay and resolve the problem, (b) make a regression, or (c) proceed, in anticipation that later words will help resolve the difficulty. It seems that here readers have had a tendency to opt for the latter option, which as Liversedge et al point out, can result in reduced fixation durations.

Second, word \({ }_{n+1}\) PB was influenced by the plausibility of the preview. This effect cannot be explained in terms of orthographic overlap since both the plausible and anomalous word previews differed from the target. The nature of the effect was distinctly different depending on whether a plausible or anomalous preview was employed; with anomalous previews exerting a more immediate and robust effect on word \({ }_{n+1}\) viewing times. Since these previews were only available prior to word \(_{n+1}\) receiving a direct fixation, the results clearly provide evidence for the extraction of meaning from a word in the parafovea.

A proponent of serial word processing might attempt to explain these findings with the suggestion that the meaning effects arose when attention moved to word \({ }_{n+1}\) following foveal identification of word \({ }_{n}\), but before the eye movement was executed. However, this would necessitate that there is enough of a lag between the shift of attention and the eye movement not only to enable the \(L_{1}\) stage of processing of word \(_{n+1}\) to be completed, resulting in a potential skip, but that there was enough time for \(\mathrm{L}_{2}\) also to be completed, allowing meaning extraction to occur. However, word \(_{\mathrm{n}+1}\) was skipped rarely and no more often when it had been anomalous, and in any case, it seems rather unlikely that all this processing could somehow be shoehorned into the time between the termination of lexical processing of word \(_{n}\) and the execution of the saccade out of it.

Given the form of the word \({ }_{n+1}\) effects found here, it seems that the approach of looking for meaning effects using semantic associates is likely flawed. Even if there is some semantic facilitation from the associated word preview, the magnitude of this would appear to be more than outweighed
by the interference generated by a word change. This sort of combination of facilitation and inhibition might possibly explain why Hohenstein et al (2010) only found a semantic PB with a fast prime duration of 125 ms on the pre-target word (unless the prime's saliency was enhanced). Shorter prime durations might not allow time for semantic processing to occur, while longer durations might strengthen the interference.

A final important finding relates to the continued effect of masking word \(_{n+1}\) with an orthographically dissimilar preview, shown in longer fixation durations in the spillover region. While serial models can account for shorter fixation durations in the spillover region - since longer fixation durations on word \({ }_{n+1}\) following an invalid preview would allow more word \({ }_{n+2}\) preview to accrue - the finding of a continuing increase in fixation duration cannot be afforded the same interpretation.

Overall, the results of this study are difficult to reconcile with models of eye movement control that allow only strictly serial sequential lexical processing. Rather, they seem more compatible with a perspective in which multiple words may be lexically processed in an overlapping fashion.

\section*{Acknowledgments}

This research was supported by an Economic and Social Research Council (ESRC) studentship awarded to the first author (ES/H018824/1).

\section*{References}

Altarriba, J., Kambe, G., Pollatsek, A., \& Rayner, K. (2001). Semantic codes are not used in integrating information cross eye fixations in reading: Evidence from fluent Spanish-English bilinguals. Perception \& Psychophysics, 63, 875-890.
Engbert, R., Nuthmann, A., Richter, E. M., \& Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. Psychological Review, 112, 777-813.
Hohenstein, S., Laubrock, J., \& Kliegl, R. (2010). Semantic preview benefit in eye movements during reading: A parafoveal fast-priming study. Journal of Experimental Psychology: Learning, Memory, and Cognition, 36, 1150-1170.
Kennedy, A. (1998). The influence of parafoveal words on foveal inspection time: Evidence for a processing tradeoff. In G Underwood (Ed.), Eye Guidance in reading a scene perception (pp, 149-180), Oxford, UK: Elsevier.
Kennedy, A. (2000). Parafoveal processing in word recognition. Quarterly Journal of Experimental Psychology, 53A, 429-455.
Kennedy, A., Murray, W. S., \& Boissiere, C. (2004). Parafoveal pragmatics revisited. European Journal of Cognitive Psychology, 16, 128-153.
Kuçera, H., \& Francis, W.N. (1967). Computational analysis of present-day American English. Providence, RI: Brown University Press.

Liversedge, S. P., Paterson, K.B. \& Pickering, M. J. (1998). Eye movements and measures of reading time. In G. Underwood (Ed.), Eye Guidance in Reading and Scene Perception. (pp. 55-76). Oxford: Elsevier.
Meuter, R. F. I., \& Allport, A. (1999). Bilingual language switching in naming: Asymmetrical costs of language selection. Journal of Memory and Language, 40, 25-40. doi:10.1006/jmla.1998.2602
Meyer, D. E., \& Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. Journal of Experimental Psychology, 90, 227-234.
Murray, W. S. (1998). Parafoveal pragmatics. In G. Underwood (Ed.), Eye guidance in reading and scene perception (pp. 181-200). Oxford, England: Elsevier.
Murray, W.S. (2006). The nature and time course of pragmatic plausibility effects. Journal of Psycholinguistic Research, 35, 79-99.
Murray, W. S., \& Rowan, M. (1998). Early, mandatory, pragmatic processing. Journal of Psycholinguistic Research, 27, 1-22.
Radach, R. \& Kennedy, A. (2013). Eye movements in reading: Some theoretical context, Quarterly Journal of Experimental

Psychology,
DOI:10.1080/17470218.2012.750676
Rayner, K. (1975). The perceptual span and peripheral cues in reading. Cognitive Psychology, 7, 65-81.
Rayner, K., Balota, D. A., \& Pollatsek, A. (1986). Against parafoveal semantic preprocessing during eye fixations in reading. Canadian Journal of Psychology, 40, 473-483.
Rayner, K. McConkie, G.W. (1976) What guides a readers' eye movements? Vision Research, 16(8), 829-837.
Rayner, K., Warren, T., Juhasz, B. J., \& Liversedge, S. P. (2004). The effect of plausibility on eye movements in reading. Journal of Experimental Psychology: Learning, Memory, and Cognition, 30, 1290-1301.
Rayner, K., White, S. J., Kambe, G., Miller, B., \& Liversedge, S. P. (2003). On the processing of meaning from parafoveal vision during eye fixations in reading. In J. Hyönä, R. Radach, \& H. Deubel (Eds.), The mind's eye: Cognitive and applied aspects of eye movement research (pp. 213-234). Oxford: Elsevier.
Reichle, E. D., Warren, T., \& McConnell, K. (2009). Using E-Z Reader to model the effects of higher level language processing on eye movements during reading. Psychonomic Bulletin \& Review, 16, 1-21.
Schotter, E. R., Angele, B., \& Rayner, K. (2012). Parafoveal processing in reading. Attention, Perception, \& Psychophysics, 74, 5-35.
Sereno, S. C., \& Rayner, K. (1992). Fast priming during eye fixations in reading. Journal of Experimental Psychology: Human Perception and Performance, 18, 173-184.
Starr, M., \& Inhoff, A. (2004). Attention allocation to the right and left of a fixated word: Use of orthographic information from multiple words during reading. European Journal of Cognitive Psychology, 16, 203-225.

\title{
Later events lie behind her, but not behind you: Compatibility effects for temporal sequences along the sagittal axis depend on perspective
}

\author{
Esther J. Walker (e1walker@cogsci.ucsd.edu) \\ Benjamin K. Bergen (bkbergen@cogsci.ucsd.edu) \\ Rafael Núñez (nunez@cogsci.ucsd.edu) \\ Department of Cognitive Science, 9500 Gilman Dr. University of California, San Diego \\ La Jolla, CA 92093-0515 USA
}

\begin{abstract}
Perspective plays a large role in how we think about space. Does perspective also influence how we think about abstract concepts, such as time, which have been shown to be closely associated with how we think about space? Linguistic patterns suggest that speakers talk about temporal sequences from two perspectives: field-based and ego perspective (Moore, 2011). However, the psychological reality of these mappings beyond their use in language is unclear. The present study examines whether sequential reasoning recruits the sagittal (front-back) axis differently, depending on the perspective adopted for the task. We manipulated perspective by using pronouns meant to evoke a field-based or ego perspective ("her" vs "your" high school graduation, respectively). Participants made earlierthan or later-than judgments about event sequences using a mouse in front of or behind their body. We observed an interaction between pronoun, temporal reference, and response location. Participants map space onto time differently depending on the frame of reference from which temporal sequences are interpreted.
\end{abstract}

Keywords: spatial construals of time; perspective; pronouns; compatibility effects; sequence time

\section*{Introductions}

Spatial perspective plays an important role in how people think about and comprehend the world around them (e.g., Tversky, 2003, 2005) and humans are quite flexible in the spatial perspectives they are able to adopt. Indeed, individuals are not only able to think about and interpret scenes from their own perspective, but are also able to adapt their perspective to that of another person (Tversky \& Hard, 2009). Furthermore, language can also influence the perspective from which one interprets a scene. For example, the use of a single pronoun influences the perspective from which readers simulate actions described in narratives (Brunyé, Ditman, Mahoney, Augustyn, \& Taylor, 2009). Brunyé et al. (2009) demonstrated that when participants read sentences such as "You are cutting the tomato" versus "He is cutting the tomato", they were faster to match the sentence to the corresponding picture if the pronoun matched the spatial perspective from which the picture was taken. As such, it appears that one's embodied simulation of actions in the world is sensitive to the perspective from which those actions are described. However, is it also the case that the use of different pronouns influences the perspective from which one thinks about more abstract concepts, which have been suggested to obtain their
conceptual structure from our embodied experience of moving through and interacting with the world around us (e.g., Lakoff \& Johnson, 1980)? One candidate that may help provide insight into such a question is time-the conceptualization of which appears tightly tied to how we think about space.

Across the world's languages, people use space to talk about time. Nevertheless, there's diversity in precisely how languages spatialize time-what axis they use, and how they map time onto that axis (Clark, 1973; Haspelmath, 1997; Núñez \& Sweetser, 2006). Moreover, the use of space to structure time isn't merely a matter of language, it's also a matter of thought-a large literature suggests that conceptualizations of time are also strongly linked to thought about space (e.g., Casasanto \& Boroditsky, 2008). Indeed, from linguists to philosophers to psychologists, scholars have discussed at length the ways in which time recruits spatial structure. This research has produced a large body of findings in language (Clark, 1973; Traugott, 1975; Moore, 2006; 2011), gesture (Cooperrider \& Núñez, 2009; Casasanto \& Jasmin, 2012), and psychological experiments (Santiago et al., 2007; Torralbo et al., 2007; Weger \& Pratt, 2008; Ouellet et al., 2010).
Scholars have long noted that there exist at least two distinct spatial construals of time: deictic and sequence (McTaggart, 1908; Núñez \& Sweetser, 2006). Deictic time conceptualization reflects past/future relationships and centers around the present moment, or "now," as a reference point. Sequence time, on the other hand, does not use "now" as a reference point. Instead, one event becomes the reference point for another event, capturing "earlier" or "later" relationships in time. Experimental research on this topic has often overlooked this distinction, pooling deictic with sequential judgments, but because the two types of time judgment relate to space differently (Casasanto \& Jasmin, 2012; Walker, Bergen, \& Núñez, 2013), the present study will focus only on sequence time.

Sequence time has been shown to recruit the transversal (left-right) axis in a systematic manner. In gesture, English speakers often sweep their hand to their left when talking about earlier events and to the right when talking about later events (Cooperrider \& Núñez, 2009; Casasanto \& Jasmin, 2012). Furthermore, space-time compatibility effects are widely reported for this axis in a variety of languages (e.g., in Spanish: Santiago, Lupiañez, Perez, \& Funes, 2007; in

English: Weger \& Pratt, 2008; in German: Ulrich \& Maienborn, 2010). For example, English-speakers are faster to respond to earlier events on their left and to later events on their right (Weger \& Pratt, 2008). Interestingly, this pattern reverses in languages that write right-to-left. Hebrew speakers show the opposite patterns and are faster to respond to earlier events on their right and later events to their left (Fuhrman \& Boroditsky, 2010). These consistent patterns are likely the product of our long history of experience with various cultural practices, suggesting a strong role of such practices like graphical notation and writing direction in the recruitment of the transversal axis. By contrast, linguistic patterns reveal the spatialization of temporal sequences along a sagittal (front-back) axis, and how such temporal sequences are mapped onto the sagittal axis in language appears to depend on the frame of reference from which it is interpreted.

Moore (2011) observes that, in language, temporal sequences can be interpreted from two different reference frames: ego-perspective and field-based. An ego-perspective depends on the perspective of the ego, as in the sentence "It looks like there are sunny days ahead" (of now). In this example, "sunny days" are described as lying ahead of the present moment, which is co-located with the speaker's ego. This is clearly a case of deictic time, where the deictic center is the present moment and ahead refers to the space in front of the speaker, which is then interpreted as in the future or later-than-now. On the other hand, while egoperspective frames depend on the perspective of the ego (as in deictic time), field-based frames are deictically neutral, meaning they do not require a deictic center, and do not change if an observer's perspective changes. For example, Moore (2011) considers people waiting in line: no matter which way you look at the line, there is a front and back to that line, dictated by convention. Those in front of others will be served earlier than those who are later in line. This can also be seen in the linguistic example, "Polls showed a widening lead for the Democrats ahead of last month's elections" (example from Moore, 2011). From this frame of reference, "ahead" is interpreted to mean "ahead of some reference event", which is subsequently interpreted as "earlier than some reference event". Thus, how sequences of events in time map onto space in language appears to depend on the frame of reference from which they are interpreted.
While patterns in language suggest that these two perspectives differ in how they use spatial terms such as "ahead", does this mean that speakers are actually using space differently when thinking about these sequences of events? Recent work has shown that, at least in deictically neutral settings, the earlier-in-front/later-in-back mapping observed in field-based frames in language emerges in experimental paradigms. Walker et al. (2013) had participants listen to a series of two events presented from a speaker that was either in front of or behind them. Participants were then asked to vocally respond whether the second event they heard happened earlier or later than the
first event they heard (e.g., after hearing her high school graduation, her college graduation, participants would respond "later" into the microphone). Results indicated that participants were faster to make earlier judgments to stimuli presented in front of them and faster to make later judgments that were presented behind them. However, the stimuli in that study all used the pronoun "her", and thus it is unclear whether the results reflect how sequences generally map onto the sagittal axis or whether they were due to the use of deictically neutral stimuli, which, according to patterns in language, elicit such a mapping. Thus, in order to determine whether perspective influences how one interprets the relationship between earlier/later and sagittal space, we'd need to know whether manipulating the perspective from which participants interpret deicticallyneutral sequences changes the pattern of space-time mappings.

In the present study, we used the same basic design as described above (Walker et al., 2013). However, stimuli were presented visually instead of auditorily and participants responded using a mouse click instead of responding vocally. Participants made judgments about deictically-neutral sequences along the sagittal axis. Critically, we manipulated the pronoun that preceded each of the events in order to examine whether differences in person perspective induces differences in the frame of reference that participants use to think about the sequences of events. Participants received the pronoun "her" for two blocks, while in the other two blocks they received the pronoun "your".

If temporal sequences are simply mapped onto space in a manner consistent with a field-based frame of reference, we would expect to see no difference in the time-space mappings recruited for each of the pronouns-both pronouns would elicit a clear earlier-in-front, later-in-back sequential mapping. Alternatively, the use of different pronouns may lead to the use of different space-time mappings. While making a judgment about whether your high school graduation is earlier or later than your college graduation need not involve any deixis or reference to the present moment, the inclusion of the pronoun "your" may automatically bring forth thoughts about your location in time. Conversely, the use of the pronoun "her" in sequences would keep them deictically neutral. If this is the case, then the use of the pronoun "her" should lead to time-space mappings consistent with a field-based interpretation (earlier-in-front/later-in-back) while the use of the pronoun "your" should encourage the use of time-space mappings consistent with an ego-perspective (earlier-in-back/later-infront). More specifically, we would expect a three-way interaction between the pronoun used, temporal reference, and location of response.

\section*{Methods}

\section*{Participants}

Forty-two undergraduates at the University of California, San Diego participated for partial course credit. Eight participants were removed due to low levels of accuracy ( \(<80 \%\) ), leaving 36 participants for analysis.

\section*{Materials and Design}

The stimuli were composed of forty pairs of typical life events (e.g., "her/your high school graduation, her/your college graduation"). Twenty pairs required "earlier" judgments while twenty required "later" judgments. Stimuli for earlier judgments were no different in length than those presented for later judgments, \(\mathrm{p}=.72\).
The experiment was programmed using E-Prime (Psychology Software Tools, Pittsburgh, PA, USA). Each participant completed a total of four blocks of forty trials of sequential judgments. During two of the blocks, each event was preceded with the pronoun "her" and during the other two blocks, events were preceded with "your". Participants completed two blocks of trials (one block for each response mapping) with each pronoun, followed by another two blocks of trials using the other pronoun. Response mappings and pronoun order were counterbalanced across participants.

\section*{Procedure}

Participants held two computer mouses, one in each hand, with each of their thumbs over one of the mouse buttons. One mouse's button was covered with red tape, while the other mouse's button was covered with yellow tape. Participants held one mouse directly in front of their body and the other mouse behind their back (see Figure 1). Which hand was in front (right or left) was counterbalanced across participants.

Before each block, participants were presented with instructions that explained the stimulus-response mappings they would use for that block. The yellow mouse was always in front of the participant, while the red mouse was always behind the participant. Which judgment required pressing a button on which mouse was changed after each block. Participants were instructed to judge whether the second event they saw in a sequence was earlier or later than the first event, and to indicate their decision by pressing either the yellow or red mouse. They were told to complete the judgments as quickly and accurately as possible. They then completed four practice trials, followed by forty randomly ordered experimental trials. There were a total of 160 trials (four blocks of 40 trials) and 16 practice trials (four practice trials per block).

On each trial, participants were presented with a fixation cross for 1000 ms , followed by the first event in the sequence. After 2000 ms , the first event was removed from the screen and a white screen was presented for 500 ms . The second event was then presented and remained on the screen until the participant responded, up to a maximum of 5000 ms . Reaction times were measured from the onset of the
second event. After each block, participants received new instructions for how to respond during the next block.


Figure 1: Participants held one mouse in front of their body and one mouse behind their body. Which hand was in front was counterbalanced across participants.

\section*{Analyses}

All analyses were conducted in \(R\) ( R Development Core Team, 2005). Incorrect trials ( \(5.8 \%\) of the data) as well as trials that were 2.5 standard deviations from each subject or item's mean ( \(3.7 \%\) of the data) were excluded from analysis. Response times were fitted with a series of linear mixedeffect models with subjects and items as crossed random effects using the lme4 library (Bates, Maechler, \& Bolker, 2011) in \(R\) ( R Development Core Team, 2005). P-values were obtained using the pvals.fnc function in the languageR package (Baayen, 2011). To investigate whether the different pronouns elicited different spatial construals of temporal sequences, a linear mixed-effects model with temporal reference (earlier, later), response location (front, back), and pronoun (her, your) as fixed effects was fitted to the response times. As needed, appropriate follow-up tests were conducted, as reported below. Models were all significantly different from their respective null models.

\section*{Results}

The overall model revealed a two-way interaction between temporal reference and response location, \(\beta=253.53\), \(\mathrm{SE}=43.14, \mathrm{p}<.001\). Follow-up tests indicated that overall, participants responded faster to later events when responding on the mouse located behind them than the mouse in front of them ( \(\beta=90.36, \mathrm{SE}=22.17, \mathrm{p}<.001\) ) and responded faster to earlier events when responding in front than in back ( \(\beta=-69.86, \mathrm{SE}=20.97, \mathrm{p}<.001\) ). There was also
a main effect of response location, \(\beta=-88.78, \mathrm{SE}=43.14\), \(\mathrm{p}<.001\). Participants were overall faster to respond by clicking on the mouse in front of them than the mouse behind them. No other main effects were significant.
Critically, there was also a three-way interaction between pronoun, temporal reference, and response location, \(\beta=-\) 184.70, \(\mathrm{SE}=60.99, \mathrm{p}=.003\). Follow-up analyses indicated that this interaction was driven by a strong interaction between temporal reference and response location on trials using the pronoun "her", \(\beta=254.03, \mathrm{SE}=41.71, \mathrm{p}<.001\). This interaction was not reliably significant for the pronoun "your", \(\beta=69.63, \mathrm{SE}=42.15, \mathrm{p}=.10\) (see Figure 2).
To examine whether participants' responses were affected by which pronoun they received first, we ran further exploratory analyses. We divided the data by which pronoun participants received first (her or your) and then created a model with spatial location and temporal reference as fixed effects and subject and items as random effects. Results are presented in Figure 3. For participants who were presented with the pronoun "her" first, there was no longer a threeway interaction between temporal reference, location, and pronoun, \(p=.31\). However, the interaction between temporal reference and location remained, \(\beta=279.48, \mathrm{SE}=61.38\), \(\mathrm{p}<.001\), and the earlier-in-front/later-in-back mapping was revealed for both "her" \((\beta=274.42, \mathrm{SE}=60.65, \mathrm{p}<.001)\) and "your" \((\beta=198.38, \mathrm{SE}=59.09 \mathrm{p}<.001)\). On the other hand, participants who were tested with the pronoun "your" first demonstrated a three-way interaction between pronoun, temporal reference, and location, \(\beta=-283.32, \mathrm{SE}=83.58\), \(\mathrm{p}<.001\). While the earlier-in-front/later-in-back pattern remained for the pronoun "her", \(\beta=231.73, \mathrm{SE}=57.53\), \(\mathrm{p}<.001\), no clear pattern emerged for the pronoun "your", \(\mathrm{p}=.38\).


Figure 2: Mean response times for earlier and later judgments along the sagittal axis. The left graph displays the results for the pronoun "her". The interaction between response location and temporal reference is significant. The right graph displays the results for the pronoun "your".

Error bars represent standard error.


Figure 3: Mean response times for earlier and later judgments along the sagittal axis when split by which pronoun participants received first: the top-left graph shows responses to the pronoun "her" when it came first; the bottom-left graph shows responses to the pronoun "your" when it came after "her"; the top-right graph shows responses to the pronoun "your" when it came first; the bottom-right graph shows responses to the pronoun "her" when it came after "your".

\section*{General Discussion}

We investigated whether the use of different pronouns ("her", "your") would lead participants to interpret temporal sequences from different perspectives and therefore lead to differences in how individuals mapped temporal sequences onto space. If participants simply systematically map sequences of events onto the sagittal axis in a manner consistent with patterns in language (earlier-in-front/later-in-back), the use of different pronouns should have no effect on the space-time mappings used by the participants. However, we observed a three-way interaction between pronoun, response location, and temporal reference: spacetime mappings recruited for temporal sequences involving "her" were different than those recruited for the pronoun "your".
The spatialization of temporal sequences along a bodycentered sagittal axis in the present study is intriguing for multiple reasons. First, sequence time does not make reference to the present moment (as in deictic time). As a result, events in a sequence are generally not co-located with a speaker's body in gesture as they are in deictic time. Indeed, sagittal gestures are rarely observed when talking about sequential time (Casasanto \& Jasmin, 2012). Second, even though sagittal language (e.g., ahead) is used to talk about sequences of events, this language is often used to refer to the location of one event with respect to another event, independent of the speaker's location in space (or time, for that matter) and thus does not reflect a sagittal axis that is centered around the speaker's body. Rather, sequential time is often talked about from an external perspective (Núñez \& Cooperrider, 2013). However, it is
interesting to consider what may be happening in the present experiment, where one's body is, by virtue of the experimental design, forced into the same sagittal axis as the rest of the events in the sequences. One potential outcome as a result of this design is that participants simply map the types of responses (earlier, later) onto the spatial locations (back, front) in a manner consistent with deictic time, which is associated with a body-centered sagittal axis. However, the present data, as well as the results of previous work by Walker et al. (2013) do not support such an interpretation. Rather, it appears that the body may be acting as an anchor for the first event in the sequence, leading participants to map earlier or later events on to the inherent "frontness" and "backness" of their bodies, with earlier events lying ahead of their body and later events lying behind, consistent with the "earlier events lie ahead of later events" structure found in deictically neutral sequential language. This finding is consistent with findings by Núñez, Motz, and Teuscher (2006), who demonstrated that participants appear to interpret sequential relationships in time by using the frontback relationship that is intrinsic to the spatial organization of whatever is anchoring the sequential construal, which, in the case of the present experiment, is the body. However, as revealed in the present study, this pattern of mappings can be flexible.
While we predicted a three-way interaction between pronoun, temporal reference, and response location, we did not observe the exact interaction pattern we expected. Though participants recruited an "earlier-in-front, later-inback" mapping for the pronoun "her", we did not see \(a\) reversal of this mapping when the stimuli were preceded with "your". What might explain the lack of an interaction between response location and temporal reference for the "your" stimuli? One possibility may be due to the fact that the pronoun "you" in English can be interpreted in two ways: either as the second person "you", referring to the interlocutor, or as the indefinite pronoun "you", which refers to a generic person (or people), as in "exercise is good for you". Thus, while some participants may be interpreting the sequence "your high school graduation, your college graduation" relative to their own lives, others may interpret it from a third person perspective, similar to "her high school graduation, her college graduation". Any effects for the pronoun "your" would then be masked by averaging and thus no interaction would emerge.
One potential factor that could have pushed participants to adopt either a personal or an indefinite interpretation of "your" could have been the order in which they completed the blocks in the experiment. Participants who started the experiment by making judgments to events that used the pronoun "her" might have been primed by that experience to subsequently interpret the sequences using "your" as not pertaining to themselves, but rather to be indefinite. By contrast, participants who started by making judgments to "your" events might have been more likely to adopt a personal second-person interpretation.

When the present data were divided by which pronoun participants received first, as described above, this very pattern of results emerges. Participants who received "her" first demonstrated an earlier-in-front/later-in-back mapping for events containing both "her" and "your". On the other hand, no consistent space-time mapping was observed for "your" events by participants who received "your" first while the earlier-in-front/later-in-back pattern was again demonstrated for the "her" events. Though these analyses are exploratory and must be interpreted with caution, they provide preliminary evidence that the pronoun that was presented first influences the pattern of space-time mappings for each of the pronouns in a manner consistent with the explanation offered above. Future investigations may want to examine whether a between-subjects design elicits a clearer difference in space-time mappings for each of the pronouns.
One question that remains from this pattern of results is why, even when "your" is presented first, the pronoun "your" does not reveal space-time mappings consistent with an ego-perspective. If "your" can be interpreted from these two different perspectives and can be primed by the pronoun "her", as the data above suggests, then it is plausible that no clear effect emerged because participants are interpreting the "your" in different manners from the beginning. Thus, in order to better understand the nature of these mappings, future work must examine what factors are responsible for this lack of effect. For example, teasing apart whether "your" is interpreted from second person as opposed to the indefinite may resolve some of the ambiguities expressed here.
In conclusion, the present study demonstrates that participants are sensitive to the perspective from which temporal sequences are framed. While the pronoun "her" encourages a strong pattern of space-time mappings for sequential time, consistent with a field-based perspective (Moore, 2011), the pronoun "your" does not. This observed difference in how participants map sequences of time onto the sagittal axis reveals that while time often recruits space in systematic and regular patterns in language, these mappings are flexible and interact with the spatial perspective from which one thinks about time.

\section*{Acknowledgments}

The authors would like to thank Lucas Medeiros de Paula, Amanda Natsuhara, and Colleen Takahashi for helping with data collection as well as Tyler Marghetis and two anonymous reviewers for helpful comments on previous versions of this work. This work was supported by an NSERC PGS-D awarded to EJW.

\section*{References}

Baayen, R.H. (2011). languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics". R package version 1.2. http://CRAN.Rproject.org/package=language R

Bates, D., Maechler, M., \& Bolker, B. (2011). lme4: Linear mixed-effects models using S4classes. R package version 0.999375-42. http://CRAN.R-project.org/package=1me4

Brunyé, T.T., Ditman, T., Mahoney, C.R., Augustyn, J.S., \& Taylor, H.A. (2009). When you and I share perspectives: Pronouns modulate perspective taking during narrative comprehension. Psychological Science, 20, 27-32.
Casasanto, D., \& Jasmin, K. (2012). The hands of time: Temporal gestures in English speakers. Cognitive Linguistics, 23, 643-674.
Clark, H.H. (1973). Space, time, semantics, and the child. In: Moore, T.E. (Ed.), Cognitive Development and Acquisition of Language. Academic Press, New York, pp. 27-63.
Cooperrider, K., \& Núñez, R. (2009). Across time, across the body: Transversal temporal gestures. Gesture, 9, 181206.

E-Prime 2.0 Software, (Psychology Software Tools, Pittsburgh, PA, USA).
Fuhrman, O., \& Boroditsky, L. (2010). Cross-cultural differences in mental representations of time: Evidence from an implicit nonlinguistic task. Cognitive Science, 34, 1430-1451.
Fuhrman, O., McCormick, K., Chen, E., Jiang, H., Shu, D., Mao, S., \& Boroditsky, L. (2011). How linguistic and cultural forces shape conceptions of time: English and Mandarin time in 3D. Cognitive Science, 35, 1305-1328.
Haspelmath, M. (1997). From space to time: Temporal adverbials in the world's languages. Munich: Lincom Europa.
Lakoff, G. \& Johnson, M. (1980). Metaphors we live by. Chicago: University of Chicago Press.
Kranjec, A. \& McDonough, L. (2011). The implicit and explicit embodiment of time. Journal of Pragmatics, 43, 735-748.
McTaggart, J. (1908). The unreality of time. Mind, 17,457474.

Moore, K.E. (2006). Space to time mappings and temporal concepts. Cognitive Linguistics, 17, 199-244.
Moore, K.E. (2011). Ego-perspective and field-based frames of reference: Temporal meanings of FRONT in Japanese, Wolof, and Aymara. Journal of Pragmatics, 43, 759-766.
Núñez, R. E., \& Sweetser, E. (2006). With the Future Behind Them: Convergent Evidence From Aymara Language and Gesture in the Crosslinguistic Comparison of Spatial Construals of Time. Cognitive Science, 30, 401-450.
Núñez, R. \& Cooperrider, K. (2013). The tangle of space and time in human cognition. Trends in Cognitive Science, 17, 220-229.
Núñez, R., Motz, B., \& Teuscher, U. (2006). Time after time: The psychological reality of the ego- and time-reference-point distinction in metaphorical construals of time. Metaphor \& Symbol, 21, 133-146.
R Development Core Team (2005). R: A language and
environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, URL, http://www.R-project.org.
Santiago, J., Lupiañez, J., Perez, E., \& Funes, M.J. (2007). Time (also) flies from left to right. Psychonomic Bulletin \& Review, 14, 512-516.
Traugott, E. C. (1975). Spatial expressions of tense and temporal sequencing: A contribution to the study of semantic fields. Semiotica, 15, 207-230.
Tversky, B. (2003). Structures of mental spaces: How people think about space. Environment and Behavior, 35, 66-80.
Tversky, B. (2005). Functional significance of visuospatial representations. In P. Shah \& A. Miyake (Eds.), Handbook of higher-level visuospatial thinking (pp. 170). Cambridge, England: Cambridge University Press.

Tversky, B., \& Hard, B.M. (2009). Embodied and disembodied cognition: spatial perspective-taking. Cognition, 110, 124-129.
Ulrich, R., \& Maienborn, C. (2010). Left-right coding of past and future in language: The mental timeline during sentence processing. Cognition, 117, 126-138.
Walker, E.J., Bergen, B.K., \& Nunez, R. (2013). Investigating spatial axis recruitment in temporal reckoning through acoustic stimuli and non-spatial responses. Center for Research in Language Technical Report, University of California, San Diego, 25, 1-10.
Weger, U. \& Pratt, J. (2008). Time flies like an arrow: Space time compatibility effects suggest the use of a mental time-line. Psychonomic Bulletin \& Review, 15, 426-430.

\title{
A Computational Model of the Development of Hemispheric Asymmetry of Face Processing
}

\author{
Panqu Wang (pawang@ucsd.edu) \\ Department of Electrical and Computer Engineering, University of California San Diego 9500 Gilman Dr 0407, La Jolla, CA 92093 USA \\ Garrison Cottrell (gary@ucsd.edu) \\ Department of Computer Science and Engineering, University of California San Diego \\ 9500 Gilman Dr 0404, La Jolla, CA 92093 USA
}

\begin{abstract}
Extensive research effort has been invested in building neurocomputational models for face and object recognition. However, the relationship between the recognition model and the development of the visual system is rarely considered. Research on the development of contrast sensitivity shows that human infants can only perceive low spatial frequency information from visual stimuli, but their acuity improves gradually with age. Also, the right hemisphere (RH) develops earlier than the left hemisphere (LH), and is dominant in infants. Here we show that these constraints, coupled with a desire on the part of the infant to individuate its caretakers and family, leads naturally to the right hemisphere bias for face processing. We propose a developmental model for face and object recognition using a modular neural network based on Dailey and Cottrell (1999). This neural network represents the two hemispheres using two modules, with a competitive relationship between them mediated by a gating mechanism. The strong RH and low spatial frequency bias for face recognition emerges naturally in the model from the interaction of the slow development of acuity and the early dominance of the right hemisphere. Remarkably, this strong asymmetry does not appear to hold for the other object categories that we tried.
\end{abstract}

Keywords: face recognition; developmental model; contrast sensitivity; modular neural network.

\section*{Introduction}

Computational models have been used extensively to instantiate hypotheses regarding face and object perception from a neurocomputational perspective (Dailey \& Cottrell, 1999; O'Reilly \& Munakata, 2000; Riesenhuber \& Poggio, 1999). Due to their social importance and frequency, faces as a stimulus class have been studied most extensively. fMRI studies have shown that the fusiform face area (FFA) is activated in response to faces more than other stimuli, and this activation is found to be stronger in the right hemisphere (Kanwisher, McDermott, \& Chun, 1997). This right hemisphere lateralization is supported by electrophysiological studies, which show a stronger face-specific wave 170 ms after the stimulus onset over the right hemisphere (Bentin, Allison, Puce, Perez, \& McCarthy, 1996). In addition, according to neuropsychological data, the lesioning of the right FFA will result in prosopagnosia, a deficit in face recognition (Bouvier \& Engel, 2006). To account for such hemispheric asymmetry, Ivry and Robertson (1998) proposed the Double Filtering by Frequency (DFF) model which postulates differential frequency filtering on task-relevant frequency bands in each hemisphere, with a bias for low spatial frequency in the RH.

Hsiao, Shieh, and Cottrell (2008) implemented this DFF theory using a computational hemispheric model, and showed that the model could account for the left-side bias in face recognition.

All neurocomputational models so far investigate the effects of structural constraints, such as the visual field split, the interaction between the two hemispheres, and the localized nature of the FFA. Temporal constraints, such as the development of the brain through childhood, are rarely considered in these models. By including consideration of the developmental changes, we may generate new hypotheses concerning the lateralization of face processing.

We consider two main constraints. First, studies in human infants have shown that the right hemisphere develops its function earlier than the left hemisphere (Chiron et al., 1997). By measuring the regional cerebral blood flow (rCBF) changes at rest using single photon emission computed tomography (SPECT), Chiron et al. observed a right hemisphere predominance of the blood flow between 1 and 3 years of age, with the asymmetry shifting to the left hemisphere after 3 years.

The second constraint has to do with the development of visual acuity. Studies have shown that the contrast sensitivity function (CSF), a measure of the ability to distinguish between different levels of luminance in static images, changes radically over time. The CSF of a 3-month-old infant is over \(1.0 \log\) units lower than the adult (Peterzell, Werner, \& Kaplan, 1995). By age 4, children's contrast sensitivity is still reduced by about \(0.5 \log\) units when compared to that of adults (Atkinson, Braddick, \& Moar, 1995; Gwiazda, Bauer, Thorn, \& Held, 1997). The studies regarding the time of contrast sensitivity maturity have obtained somewhat disparate results. They range from claims of maturity between 6 to 10 years of age (Bradley \& Freeman, 1982; Mayer, 1977), to evidence of immaturity in 8-15 year old children (Arundale, 1978). More recent studies suggest that contrast sensitivity attains adult levels in all frequencies by age 7 (Ellemberg, Lewis, Liu, \& Maurer, 1999), or 8 (Leat, Yadav, \& Irving, 2009).

Combining these two constraints, we propose a neurocomputational developmental model for object and face recognition. The model is based on a previous study of the development of hemispheric asymmetry by Dailey and Cottrell
(1999). Dailey \& Cottrell used a mixture of experts architecture (Jacobs, Jordan, \& Barto, 1991), a neural network with two modules to represent the two hemispheres. The output of the modules is gated based on their contribution to the overall performance. Dailey \& Cottrell also suggested that low acuity, and the need to individuate faces, would lead to right hemisphere dominance. However, in their model, they assumed the right hemisphere would receive low spatial frequencies, and the left hemisphere high spatial frequencies, which is an unrealistic assumption. Furthermore, their model did not change its visual acuity over time. Here, we model the changes in children's contrast sensitivity by low-pass filtering the data set, and gradually improving the fidelity of inputs over training iterations (i.e., as the model "ages"). To model the asymmetric developmental pattern of the brain, we give the two modules different learning rates over time. A detailed description of this methodology will be given in next section. In the result section, we will discuss the network's general performance on different objects, and the strong right hemisphere bias for face processing which emerges naturally from the interaction of the developmental trends.

\section*{Methods}

We will first describe the structure of the modular neural network, and then discuss the modifications to create the development model.

\section*{The Model}

Each input image goes through a two-step preprocessing stage: Gabor filtering and Principal Component Analysis (PCA). The biologically motivated 2-D Gabor filter (Daugman, 1985) is constructed by using a two-dimensional sinusoid localized by a Gaussian envelope. By tuning to particular spatial frequency and orientations, the Gabor filter magnitudes can be used to simulate the responses of complex cells in primary visual cortex (V1). Following Gabor filtering, PCA reduces the dimensionality of the information by simulating the information extraction mechanism beyond V1, up to the lateral occipital regions level. After these preprocessing steps, each image is represented by a vector of relatively low dimension to be fed into the modular neural network.

The mixture of experts architecture has been in development since 1990 (Dailey \& Cottrell, 1999; Jacobs, Jordan, \& Barto, 1991; Jacobs, Jordan, Nowlan, \& Hinton, 1991). Our particular variant of the model is presented in Figure 1. The modular neural network has three components: two side-by-side hidden layers (the "modules"), an output layer and a gating layer. The hidden layers are used to learn features for a given task adaptively, and develop more sophisticated representations for the input stimuli; if the task is face recognition, we can consider the hidden layer as corresponding to FFA. There is one hidden layer for each of the hemispheres. There are sufficient hidden units in each network to perform the tasks. The output layer provides us with the labels of the input stimulus to perform the final object recognition task.


Figure 1: Architecture of Dailey and Cottrell's (1999) modular neural network model.

Because the labels to be discriminated have a strong influence on the hidden layer representation through the learning process, the discrimination level of the output layer will drive the representation developed by the hidden layer through error feedback. The gating layer has two nodes whose activity modulates the output of the hidden layers by multiplying the connection weights from the hidden layer to the output layer. We use a softmax function at the output of the gating network to ensure that the gating weights always sum to one. The network is trained using online backpropagation. During learning, the gating nodes act like a competitive selection mechanism and direct more information (error feedback) to the module that has more contribution to performance and better ability to process a given pattern. For example, if the task is better performed by module 1 , the gating network will learn to weight the module 1 outputs more highly, and the hidden units of module 1 will also receive more training as a result. Thus there is a "rich get richer" effect.

Dailey and Cottrell (1999) trained this modular neural network to account for the development of the FFA. Based on the evidence that the right hemisphere has a relatively low spatial frequency preference and the left hemisphere has a relatively high spatial frequency bias (Sergent, 1982), they gave different spatial frequency information from each stimulus to each module by weighting the frequency component differently in the PCA step. When the model's goal was to individuate different faces while categorizing the other classes of stimuli, they observed a strong specialization of the low spatial frequency module to processing faces; no other tasks showed a similar specialization. Hence, they took these results as support for the hypothesis that the FFA developed as a natural consequence of the infant's low visual acuity, and the need to individuate faces.

\section*{Modeling Infant's Developmental Patterns}

Although Dailey and Cottrell (1999) successfully modeled the right hemisphere lateralization of FFA for face recognition, their consideration of the developmental facts is inade-
quate. First, although the right hemisphere (RH) has a low spatial frequency (LSF) bias, both the left and right hemispheres should receive the same frequency information from the input stimuli, as psychophysical results show equal CSF's in the two visual fields. We hypothesize that the selectivity arises as a consequence of competition during the developmental period, rather than assuming it is already in force during development. Hence, instead of manually weighting the information provided by each spatial frequency differently in each module, in this work we give both modules the same images over time. Instead of manipulating the spatial frequencies so that they are different, we give the modules different learning rates, in accord with the data that the right hemisphere is dominant during the first three years (Chiron et al., 1997). We model this as a wave of plasticity in each hemisphere that is slightly out of phase, as in Figure 2. The right hemisphere will thus do more learning earlier than the left. This will drive the right hemisphere network to reach convergence more quickly, which in turn will make it win the competition and obtain a higher gating node value in the gating network. We used two Gaussians with different mean values and the same variance followed by a straight line to represent the evolution of the learning rate. At earlier epochs, module 1 has a higher learning rate than module 2. After a certain number of iterations, the learning rate of module 2 starts to increase and prevails over module 1. Finally, both learning rates drop to a small constant value until the end of training. This represents the maturity of both hemispheres, but they continue to learn into "adulthood."


Figure 2: Learning rate example. Module 1 (dashed line) learns earlier than module 2 (solid line) and they converge to same value after 30 iterations.

Since the contrast sensitivity of infants is relatively low (Atkinson et al., 1995; Gwiazda et al., 1997; Peterzell et al., 1995), they can only obtain the low spatial frequency information of a given visual stimuli instead of receiving the full frequency details. As they grow older, their contrast sensitivity will reach adult levels, and then they will receive the full spatial frequency information from the input. To model
this change of children's contrast sensitivity over time, we low-pass filtered the dataset before sending them to the Gabor filter in the modular neural network model. We gradually improved the fidelity of the input as the training iteration increases. To be more specific, we used a 2-D circular averaging filter with decreasing radius \(r\) to filter the dataset. We set the largest radius be 6 pixels, and gradually decreased the radius by a same interval till it reached zero (unfiltered). Figure 3 shows the result of a image filtered by a disk filter with high to low radius.


Figure 3: Result of a sample image filtered by disk filter with radius decrease from 6 (leftmost) to 0 (rightmost).

\section*{Experiments and Results}

\section*{Face/Object Stimuli and Preprocessing}

Our studies used four category of objects: faces, books, cups, and handwritten digits. For the faces, books and cups, we collected the images from 12 different individuals for each category from the Cottrell and Metcalfe (1991) database. For digits \(0-9\), we utilized the MNIST handwritten database (LeCun, Bottou, Bengio, \& Haffner, 1998) and sampled 20 images for each digit. Each image was transformed to grayscale and cropped to size of \(64 \times 64\). In the Gabor filtering step, we adopted the classical Gabor filter bank (Lades et al., 1993) with 5 different scales and 8 orientations ranging from 0 to \(7 \pi / 8\). This step resulted in a 40 -dimensional vector at each point in the image. After normalizing the response values across orientations and subsampling these vectors in a \(8 \times 8\) grid, we computed the magnitude of the complex numbers and got a 2560 -dimensional vector to represent the image. To extract a smaller number of features and maintain a segregation of responses from each scale of Gabor filter, we performed PCA separately on each spatial frequency component of the pattern vectors. Since we had 5 different scales, we extracted the subvectors corresponding to each scale from every pattern in the training set, computed the eigenvectors of the covariance matrix, and projected these subvectors onto these basis vectors. Here we used the Turk and Pentland trick (Turk \& Pentland, 1991) to reduce the computational cost. We retained the eight most significant coefficients of each scale according to the eigenvalue, reassembled the pattern and finally obtained a 40-dimensional vector for each input image.

\section*{Network Training}

Based on the task manipulations performed by Dailey \& Cottrell, we trained the modular neural network to perform three tasks:
1. 4-Class basic-level classification on faces, books, cups and digits: 4 outputs in total.
2. Face expert: subordinate classification on 10 different faces, together with basic-level classification on books, cups and digits: 13 outputs in total.
3. Non-face expert: subordinate classification on 10 different digits (0-9), cups or books, together with basic-level classification on other 3 categories: 13 outputs in total.

For each task, we performed two experiments. As a control condition, we used the same learning rate \(\lambda\) for both modules. In the experimental condition, we assigned different learning rate over time to each module to model the asymmetric development pattern over 30 iterations, as in Figure 2:
\[
\begin{aligned}
& \lambda_{\text {module } 1}=0.015 \times \mathcal{G}(\text { iteration }, 10,5) \\
& \lambda_{\text {module } 2}=0.015 \times \mathcal{G}(\text { iteration }, 20,5)
\end{aligned}
\]
where \(\mathcal{G}\) (iteration, 10,5 ) means a Gaussian distribution with mean 10 and variance 5 evaluated at iteration. After 30 iterations, we set the constant learning rate \(\lambda=1.5 \times 10^{-4}\) for both modules. In all experiments, we considered module 1 to be the right hemisphere, which learns earlier, and module 2 to be the left hemisphere. To model the development of contrast sensitivity, we utilized 9 different filtered datasets from low to full spatial frequency (see Figure 3), changing the dataset every three epochs. While the mapping is arbitrary, we assume for now that three epochs correspond to a year, so that the contrast sensitivity matures around the "age" of 9 , or 27 epochs. At that point, the dataset with full spatial frequency is used to train the model to convergence.

We used stochastic gradient descent (online learning) to train the neural network. We used a small momentum term, set to 0.01 in all experiments. The number of output nodes equals the number of categories to be classified. We performed a 10 -fold cross validation on the training set to find the optimal settings for the number of nodes in hidden units and the stopping criteria based on the model's performance on the hold-out set. In the 4 -way classification task, we used 48 images from each class to form the training set, and 12 im ages to form the test set. We determined that a mean squared error (MSE) of 0.001 was an adequate threshold to stop training: the value of the gating nodes was stable and the network reached good performance on the test set. For the face expert and non-face expert experiments, due to limitations on how many images we had of individual faces, we reduced each class to have 20 training images and 4 testing images, and the MSE threshold was set to 0.02 to avoid overtraining. Both hidden layers are determined to have 15 nodes as the model's performance is good and stable when we set it between 10 and 20 . The connection weights between input layer, hidden layer and output layer were set randomly, while the weights between input to gating layer were all set to have the same value to give the softmax layer of the two gating nodes the same initial value of 0.5 .

For each of the \(3 \times 2\) experimental conditions, we ran the experiment 12 times and recorded the softmax output of the gating layer for each class after convergence. This indicated

(a) 4-way classification (control)

(c) Face expert (control)

(e) Number expert (control)

(g) Cup expert (control)

(b) 4-way classification (split learning rate)

(d) Face expert (split learning rate)

(f) Number expert (split learning rate)

(h) Cup expert (split learning rate)

Figure 4: Average weights assigned to each module for each stimulus class. In control condition (left column), each module receives the same learning rate. We can see the average weights are almost symmetric in all tasks. In the experimental condition(right column), each module receives different learning rate. We can see a strong RH bias for face expert task, but no such bias exists in number and cup expert task. The error bar denotes the standard deviation.
the lateralization of each module for different classes of objects. Figure 4 displays the result for three of the tasks.

In the control condition, both modules receive the same learning rate, so it is expected that there is no preference for either module over any class of stimuli. Even if at certain
time one module wins for a given category, the average gating value for each module will be equal when averaged over many runs. As a result, no matter what the task is, no bias exists in the control task, so the average weights should be symmetric, as can be seen in the left column of Figure 4.

When we assigned different learning rate for each module, the task of recognizing different faces (face expert) shows a consistently strong bias for module 1 (right hemisphere). From Figure 4 (d), the average weight of RH reaches 0.96, 24 times higher than the LH weight value. However, this strong hemisphere lateralization does not occur on the task of differentiating the non-face objects, where the average weights for both hemispheres are close. We have also performed the equivalent experiment with a book expert, and this experiment only shows the weak right hemisphere bias (See Figure 5). Additionally, we have also performed the experiment using different learning rates and training epochs per dataset, and this RH lateralization for face recognition appeared very robust.

\section*{Brain Development, Contrast Sensitivity and RH Bias for Faces}

The finding of a strong RH bias for face recognition directly raises two questions: at which point during the learning process (brain development) did the bias appear? What is the difference between face recognition and object recognition? To investigate the relationship between brain development and RH bias, we ran another face expert experiment, and kept track of the softmax value of the gating nodes for each class separately until they are stable. The result of gating node value vs. time is shown in Figure 5.

We observed some interesting phenomena. First, for all non-face objects, the value of the gating node for the RH increases rapidly before the 10th epoch. This is consistent with the fact that the learning rate for the RH is much higher than the LH (see Figure 2). In addition, a significant drop for RH value and an increase for LH value appear between 10-20th iteration for these same objects, when the learning rate switches to the LH. Thus these objects track the learning rate closely. However, the impact of the increased learning rate for the left hemisphere does not affect the allocation of face processing to the RH , as the gating node value for RH remains high and stable. Considering the hidden layer of the RH network to be analogous to the FFA, these findings are also consistent with the fact that the FFA is RH lateralized (Kanwisher et al., 1997). Combined with the fact that we mainly utilized the low spatial frequency (LSF) information to train the RH network before 10th iteration, and full spatial frequency thereafter, we've shown that in our model, face recognition has a strong LSF preference. The hypothesis then is that this strong LSF and RH lateralization of face recognition is a natural result of the interaction between children's brain development and contrast sensitivity maturity.

Finally, we investigated the relationship between the rate of contrast sensitivity development and RH bias for faces. By varying the step size of our circular filter to make smaller


Figure 5: Gating node value vs. time for all objects in the experiments. We took the average value across all individuals in the same category to get the result for the corresponding expert tasks. The solid line represents the node value of RH and the dashed line represents the node value of LH. The dotted line set a middle threshold of 0.5 . We set the maximum value at horizontal axis as 100 because the node value is stable afterwards.
or greater increments, we varied the rate of development. We built different numbers of filtered datasets ranges from 6 to 14 to test the effect on the degree of bias. All filtered datasets have the same range of filtering coefficients, and a larger number of dataset means a smaller interval step. We ran the training 12 times on each group of filtered datasets and recorded the value of RH and LH nodes when the network converges. We computed the mean value and standard deviation of both hemispheres. The results are shown in Table 1 .

Table 1: Relationship between age of contrast sensitivity matures and the RH bias for face recognition
\begin{tabular}{|c||c|c||c|c|}
\hline "Age" & RH Value & std(RH) & LH Value & std(LH) \\
\hline 6 & 0.7908 & 0.0908 & 0.2654 & 0.0968 \\
8 & 0.8088 & 0.0269 & 0.1912 & 0.0269 \\
10 & 0.8989 & 0.0241 & 0.1011 & 0.0333 \\
12 & 0.8467 & 0.1042 & 0.1533 & 0.0709 \\
14 & 0.8352 & 0.0533 & 0.1648 & 0.0525 \\
\hline
\end{tabular}

The result shows the RH bias for face recognition is obvious regardless of rate of contrast sensitivity changes. However, we can observe a peak in the lateralization at 10 "years." This suggests that there could be an "optimal" age of CSF maturity with respect to lateralization.

\section*{Conclusion}

In summary, we built a face and object recognition model, using a developmentally-inspired implementation. Our model suggests that the strong right hemisphere lateralization for face processing can arise from the interaction of two developmental facts: That the right hemisphere matures earlier than the left, and that visual acuity increases gradually over development. We should also note that, as in Dailey \& Cottrell's 1999 model, there is a strong effect of task as well. Simply categorizing faces as faces does not lead to right hemisphere specialization. Hence the drive by the infant to individuate its parents, other caretakers, family and friends, leads to the right hemisphere specialization. In addition, as we observe a mild right hemisphere lateralization more or less for all expert experiments, we predict that both the task and the low spatial frequency nature of certain objects should contribute to the RH bias. Future work includes doing more analysis of the model to discover why the strong RH bias happens for faces, investigating the relationship between CS development and RH bias in more detail, and extending the model to other important classes of stimuli, such as word recognition, which shows a left hemisphere bias.

\section*{Acknowledgments}

This work was supported in part by NSF grants SMA1041755 and IIS-1219252 to G. W. Cottrell. We thank Gary's Unbelievable Research Unit (GURU) for comments on this work.

\section*{References}

Arundale, K. (1978). An investigation into the variation of human contrast sensitivity with age and ocular pathology. British Journal of Ophthalmology, 62, 213-215.
Atkinson, J., Braddick, O., \& Moar, K. (1995). Contrast sensitivity of the human infant for moving and static patterns. Vision Research, 17, 1045-1047.
Bentin, S., Allison, T., Puce, A., Perez, A., \& McCarthy, G. (1996). Electrophysiological studies of face perception in humans. Journal of Cognitive Neuroscience, 8, 551-565.
Bouvier, S. E., \& Engel, S. A. (2006). Behavioral deficits and cortical damage loci in ceberal achromatopsia. Cerebral Cortex, 16, 183-191.
Bradley, A., \& Freeman, R. D. (1982). Contrast sensitivity in children. Vision Research, 22, 953-959.
Chiron, C., Jambaque, I., Nabbout, R., Lounes, R., Syrota, A., \& Dulac, O. (1997). The right brain hemisphere is dominant in human infants. Brain, 120, 1057-1065.
Cottrell, G. W., \& Metcalfe, J. (1991). Empath: face, gender and emotion recognition using holons. In R. P. Lippman, J. Moody, \& D. S. Touretzky (Eds.), Advances in neural information processing systems (Vol. 3, pp. 564-571). San Mateo, CA: Morgan Kaufmann.
Dailey, M. N., \& Cottrell, G. W. (1999). Organization of face and object recognition in modular neural network models. Neural Networks, 12, 1053-1073.

Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by twodimensional visual cortex filters. Journal of the Optical Society of America A, 2, 1160-1169.
Ellemberg, D., Lewis, T. L., Liu, C. H., \& Maurer, D. (1999). Development of spatial and temporal vision during childhood. Vision Research, 39, 2325-2333.
Gwiazda, J., Bauer, J., Thorn, F., \& Held, R. (1997). Development of spatial contrast sensitivity from infancy to adulthood: psychophysical data. Optometry and Vision Science, 74, 785-789.
Hsiao, J. H., Shieh, D. X., \& Cottrell, G. W. (2008). Convergence of the visual field split: Hemispheric modeling of face and object recognition. Journal of Cognitive Neuroscience, 22, 2298-2307.
Ivry, R. B., \& Robertson, L. C. (1998). The two sides of perception. Cambridge, MA: MIT Press.
Jacobs, R. A., Jordan, M. I., \& Barto, A. G. (1991). Task decomposition through competition in a modular connectionist architecture: The what and where vision tasks. Cognitive Science, 15, 219-250.
Jacobs, R. A., Jordan, M. I., Nowlan, S., \& Hinton, G. E. (1991). Adaptive mixtures of local experts. Neural Computation, 3, 1-12.
Kanwisher, N., McDermott, J., \& Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. Journal of Neuroscience, 17, 4302-4311.
Lades, M., Vorbrggen, J., Buhmann, J., Lange, J., Malsburg, C. von der, Wrtz, R. P., et al. (1993). Distortion invariant object recognition in the dynamic link architecture. Computers, IEEE Transactions on, 42(3), 300-311.
Leat, S. J., Yadav, N. K., \& Irving, E. L. (2009). Development of visual acuity and contrast sensitivity in children. Journal of Optometry, 2, 19-26.
LeCun, Y., Bottou, L., Bengio, Y., \& Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.
Mayer, M. (1977). Development of anisotropy in late childhood. Vision Research, 17, 703-710.
O'Reilly, R., \& Munakata, Y. (2000). Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain. Cambridge, MA: MIT Press.
Peterzell, D. H., Werner, J. S., \& Kaplan, P. S. (1995). Individual differences in contrast sensitivity functions: longitudinal study of 4-, 6-, and 8-month-old human infants. Vision Research, 35, 961-979.
Riesenhuber, M., \& Poggio, T. (1999). Hierarchical models of object recognition in cortex. Nature Neuroscience, 2, 1019-1025.
Sergent, J. (1982). The cerebral balance of power: confrontation or cooperation? Journal of Experimental Psychology. Human Perception and Performance, 8, 253-272.
Turk, M., \& Pentland, A. (1991). Eigenfaces for recognition. The Journal of Cognitive Neuroscience, 3, 71-86.

\title{
Intentional Constraints on the Dynamics of Human Performance and Behavioral Variability in Motor Control
}

\author{
Auriel Washburn (washbual@mail.uc.edu) \\ Department of Psychology, 4150 Edwards 1 Cincinnati, OH 45221 USA
}

\author{
Charles A. Coey (coeyca@mail.uc.edu) \\ Department of Psychology, 4150 Edwards 1 \\ Cincinnati, OH 45221 USA
}

\author{
Michael J. Richardson (richamo@mail.uc.edu) \\ Department of Psychology, 4150B Edwards 1 \\ Cincinnati, OH 45221 USA
}

\begin{abstract}
Manipulation of environmental constraints has been shown to influence the relative amounts of voluntary and involuntary control employed by a person to complete a task, as well as the resulting structure of performance variability. Generally, the voluntary control required when no constraints are present leads to self-similar changes in performance, some constraint provides involuntary control that leads to random fluctuations in performance, and constraint which provides feedback about performance accuracy can result in anti-persistent variability. The current study investigated whether providing two groups of individuals with different intentions for the same task would produce changes in voluntary and involuntary control similar to that observed following the manipulation of task constraints. Results indicated that a difference in intention does result in divergent uses of voluntary and involuntary control and distinctly different structures in performance variability.
\end{abstract}

Key words: intention; fractal structure; voluntary and involuntary control; motor control

Over the past decade, a substantial amount of research has focused on determining what information can be gained about human cognitive and motor processes by assuming that they are inextricably linked through what is often referred to as the 'interaction-dominant dynamics' of human behavior (Holden, Van Orden, \& Turvey, 2009; Ingber, 2003; Turvey, \& Moreno, 2006; Van Orden, \& Holden, 2002; Van Orden, Holden, \& Turvey, 2003; Van Orden, Holden, \& Turvey, 2005). As noted by Van Orden (2010), absolute independence of these processes would allow for random variability in performance within each process, while dominance by one process over all others would cause highly regular fluctuations across processes. Standard, linear statistical methods for assessing performance are based on an assumption of random variability, or noise, in performance and, necessarily, the belief that whatever process is being evaluated can be thought of as independent from all other contemporaneous processes. However, methods for assessing potential structure within variability over time reveal that while fluctuation in performance is
sometimes strictly random, more often variability is characterized by patterns occurring at a variety of different timescales (Ferrer-i-Cancho \& Elvevag, 2010; Kiefer, Riley, Shockley, Villard, \& Van Orden, 2009; Eke, Herman, Kocsis, \& Kozak, 2002; Eke, Herman, Bassingthwaighte, Raymond, Percival, Cannon, Balla, Ikrenyi, 2000; Gilden, 2001; Holden et al., 2009; Kuznetsov \& Wallot, 2011; Phillipe, 2000; Rhodes \& Turvey, 2007; Wallot \& Van Orden, 2011a, b; Warren, Carciun, \& Anderson--Butcher, 2005). This type of variability is neither strictly random, nor strictly regular, but is rather somewhere in between the two, and therefore suggestive of both competitive and cooperative interactions between the different cognitive and motor aspects of the behavior under observation (Van Orden, 2010).

The patterned variability in performance described above is defined by a fractal structure, in that self-similarity in fluctuations is apparent at multiple timescales (Mandelbrot, 1982; Brown \& Liebovitch, 2010; West \& Deering, 1995). This type of variability is typically referred to as 'pink' noise, in contrast to the 'white' noise of random fluctuation (Van Orden, 2010). In order to determine what kind of variability is occurring for a given task, it is important to repeatedly measure some aspect of that task as performance unfolds over time. The resulting series can then be broken down into several composite, sinusoidal series each with a different amplitude and frequency. A Power-Spectral Density (PSD) analysis can then be used to give an assessment of variability (Delignieres, Ramdani, Lemoine, Torre, Fortes, \& Ninot, 2006; Holden, 2005; Marmelat \& Delignieres, 2011). The slope of a regression line fit to a plot of the logarithm of the power (amplitude squared) of changes with the logarithm of their corresponding frequencies provides a unique scaling relation between the size and frequency of changes in the performance time series. This scaling relation ( S ) is related to a characteristic scaling exponent ( \(\alpha\) ), where \(\alpha=-S\) (Holden, 2005). It is this scaling exponent which is used to give a qualitative assessment of the type of variability being observed. Since there will be no systematic relationship between the size and
frequency of change with random variability, or 'white noise', \(\alpha \approx 0\). In contrast, the scaling relation for pink noise reflects an inversely proportional relationship between the power and frequency of variation such that the scaling exponent associated with fractal variability is \(\alpha \approx 1\). It is also possible to obtain negative values for the scaling exponent. This indicates a directly proportional relationship between the size and frequency of changes in performance, and occurs as a result of anti-persistent variation (Delignieres \& Torre, 2009; Hausdorff, Peng, Ladin, Wei, \& Goldberger, 1995; Schmidt, Beek, Treffner, \& Turvey, 1991).

Strong support for interaction-dominant dynamics is provided by the fact that the kind of variability observed in a given task appears to be sensitive to a variety of task characteristics (Chen, Ding, \& Kelso, 2001; Delignières, Torre, \& Lemoine, 2009; Hausdorff, Purdon, Peng, Ladin, Wei, \& Goldberger, 1996; Holden, Choi, Amazeen, \& Van Orden, 2011; Jordan, Challis, \& Newell, 2007a; Jordan, Challis, \& Newell, 2007b). Specifically, the level of task constraint imposed by an experimental setup appears to be directly predictive of variability structure, with greater constraint resulting in white noise ( \(\alpha \approx 0\) ), and less constraint leading to pink noise ( \(\alpha \approx 1\) ) (Chen et al., 2001; Delignières et al., 2009; Hausdorff et al., 1996). This phenomenon has led to the suggestion that environmental constraint in the context of a specific task demand provides some external control, while the absence of constraint given the same task requires additional voluntary control on the part of the actor (Van Orden, 2010).

One way to summarize the effects of voluntary and involuntary control on variability is to examine different conditions within rhythmic motor tasks. Previous studies have demonstrated that by providing some sort of rhythmic stimulus (e.g. metronome) while participants are required to maintain a consistent movement pattern, spontaneous entrainment between participant and stimulus will constrain behavior and thus reduce the need for voluntary control of movements, ultimately resulting in the random variations in performance characterized by white noise ( \(\alpha \approx 0\) ) (Chen et al., 2001). However, it appears that when participants are explicitly instructed to coordinate with a rhythmic stimulus, an altogether different pattern of variability emerges (Delignières et al., 2009; Hausdorff et al., 1996). One might imagine that the requirement to synchronize would introduce the need for additional voluntary control but, more importantly, it also appears to provide the participant with feedback about the accuracy of their movements with respect to the goal of the task (Van Orden, 2010). Accuracy feedback has been considered a unique form of involuntary control, and the constraint emerging from corrective processes results in performance characterized by antipersistent, dependent fluctuations ( \(\alpha \approx-1\) ) (Delignieres \& Torre, 2009). An equivalent task to the two previous, but requiring voluntary control, can be constructed through the use of a continuation paradigm. Here the participant has the
opportunity to match their movement to an experimental stimulus for several seconds at the beginning of a trial, and then must maintain that movement pattern for the duration of the trial without any involuntary control provided gained through task constraint. Several studies have demonstrated that the use of a continuation paradigm in this manner leads to the self-similar variability of pink noise ( \(\alpha \approx 1\) ) (Chen et al., 2001; Gilden, Thornton, \& Mallon, 1995; Lemoine, Torre, \& Delignieres, 2006; Torre, \& Delignières, 2008).

While previous findings have demonstrated an association between voluntary or involuntary control and performance variability, the potential influence of intending to control a specific task dimension has yet to be examined. The current study was designed to determine the effect of being asked to control one of two task dimensions on performance variability. In order to achieve this, a simple arm-swinging task was employed during which participants were instructed to control either the frequency or amplitude of their movements, while being provided with flashing dots to help control their performance. This ultimately created the single underlying task of maintaining a comfortable, consistent movement, allowing for an isolated evaluation of the effects of intention on constraint and performance.

\section*{Method}

\section*{Participants}

Seventeen University of Cincinnati undergraduate students participated in this experiment, eight in the amplitude intention condition and nine in the frequency intention condition. Participants ranged in age from 18 to 28 years.

\section*{Procedure and Design}

At the beginning of the experiment, participants were instructed to stand at a distance 3.5 feet in front of a flat screen television, facing toward the screen. The experimental task consisted of holding one's upper right arm flush with the side of the body and swinging the forearm in an arc about the elbow, while keeping the forearm parallel to the floor. The right hand was to be held in a fist with the first two fingers extended to point toward the screen and with the knuckles facing toward the right, away from the participant's body. Initially, two red dots (5.5 cm in diameter) appeared on the screen, centered vertically and separated by a distance of 57 cm (see Varlet, Coey, Schmidt, \& Richardson, 2011 for information on determining the ideal stimulus amplitude for visuomotor entrainment).

Eight participants were asked to control the amplitude of their movements by traveling the same distance with every arm swing. The other nine participants were asked to control the frequency of their movements by maintaining a constant speed while swinging. All participants participated in two trials, each six minutes in length. The first trial involved a continuation paradigm, with the red dots appearing for the first \(10-12\) seconds (timed manually), followed by a blank
screen for the duration. This trial was collected as a baseline. The red dots were set to flash in an alternating pattern at a frequency of 1 Hz , (with a dot appearing on one or the other side of the screen every 500 ms ) throughout the time they were visible. Participants were instructed to use the dots to help control their designated task dimension, and to do their best to maintain the same movement for the duration of the trial once the dots had disappeared. In the second trial, the flashing dots were displayed for the full six minutes and participants were instructed to use them over the entire trial to help achieve consistency in their designated task dimension. This was the test trial.

By using a comfortable movement frequency for the flashing dots, we expected participants in the frequency intention condition to use the dots to gain feedback about the consistency of their speed in order to engage in corrective processes. For those participants in the amplitude intention condition we expected the stimulus to provide the opportunity for spontaneous entrainment, but not enough feedback about the size of their movements to allow for corrective processes.

The display was generated by an application written using C/C++ and displayed using OpenGL. Data was collected using a magnetic tracking system (Polhemus Fastrak, Polhemus Corporation, Colchester, VT), with the sensor attached to the outside of the extended first two fingers of the right hand. The OpenGL program was also used to record the movement data collected by the tracking system, with a sampling rate of 60 Hz .

\section*{Data Analysis}

All participant movement time-series were low-pass filtered using a 10 Hz Butterworth filter and the first and last 5 s of each trial were discarded to remove transients.

For the PSD analysis, the peak to valley intervals and valley to peak intervals were extracted from the movement time-series for each trial. The PSD analysis was then used to assess fractal characteristics of the resultant interval timeseries. As the preliminary step to this process each timeseries was submitted to a Fourier transform, during which it was broken down into several composite sinusoidal series with varying amplitudes and frequencies. The slope of a regression line fitted to the spectral plot of the logarithm of the power vs. the logarithm of the frequency for each sinusoidal series yielded a unique value \(S\), for which the characteristic scaling exponent of the series, \(\alpha\), is equal to \(-S\) (Holden, 2005).

In order to assess possible entrainment of participants' movement to the frequency of the flashing stimulus we found, for each trial, the distribution of relative phase angles occurring between the participant and stimulus time-series. This distribution was based on the proportion of discrete relative phase (DRP) angles between the two time-series which fell into each of nine bins \(\left(0^{\circ}-20^{\circ}, 20^{\circ}-40^{\circ}, 40^{\circ}-60^{\circ}\right.\), \(60^{\circ}-80^{\circ}, 80^{\circ}-100^{\circ}, 100^{\circ}-120^{\circ}, 120^{\circ}-140^{\circ}, 140^{\circ}-160^{\circ}\), and \(\left.160^{\circ}-180^{\circ}\right)\). DRP values were calculated at each oscillatory
peak of the movement time-series. Perfect, inphase coordination between participant and stimulus would result in a relative phase of \(0^{\circ}\), while antiphase coordination, in which participants exactly matching the frequency of stimulus movement but pointed at the side of the screen opposite the dot each time, would lead to a relative phase of \(180^{\circ}\).

The stability of any unintentional coordination was established by calculating the circular variance of the relative phase angles found between the participant and stimulus time-series for each trial. This measure provides an index of synchronization on a scale from 0 to 1 , with 0 reflecting a situation in which there is no coordination between participant and stimulus movements, and 1 indicating absolute synchronization between the two (Batschelet, 1981; Oullier, de Guzman, Jantzen, Lagarde, \& Kelso, 2008).

\section*{Results}

A 2 (intention) x 2 (trial) mixed model ANOVA on circular variance values revealed a significant main effect for trial, \(F\) \((1,15)=39.93, p=.001, \eta_{\mathrm{p}}^{2}=.73\), and a significant interaction between intention and trial, \(F(1,15)=6.21, p=\) \(.03, \eta_{\mathrm{p}}^{2}=.29\). This interaction appears to be driven by the fact that the effects of intention are different for the baseline and test conditions (see Figure 1). While there did not appear to be a significant change in coordination stability for those with the intention to contol amplitude, the difference for those in the frequency intention condition between baseline and test was significant, \(t(8)=-10.13, p\) \(=.001\).

A 2 (intention) x 2 (trial) x 9 (relative phase bin) mixed model ANOVA on the DRP between participant and stimulus movements revealed a main effect for relative phase bin, \(F(8,15)=81.86, p=.001, \eta_{\mathrm{p}}^{2}=.85\), significant two-way interactions between intention and relative phase bin, \(F(8,15)=4.62, p=.001, \eta_{\mathrm{p}}^{2}=.24\), and trial and relative phase bin, \(F(8,15)=105.76, p=.001, \eta_{\mathrm{p}}{ }^{2}=.88\), and a significant three-way interaction between intention, trial and relative phase bin, \(F(8,15)=2.06, p=.045, \eta_{\mathrm{p}}{ }^{2}=\) .12. Follow-up analyses revealed a significant interaction between intention and relative phase bin for the test trials, \(F\) \((8,120)=4.64, p=.001, \eta_{\mathrm{p}}^{2}=.24\), but not for the baseline trials (see Figure 2). A comparison of the proportion of time spent in the DRP bin associated with inphase coordination ( \(0^{\circ}-20^{\circ}\) ) during the test trial between the two intention conditions revealed that significantly more in-phase entrainment occurred for those participants instructed to control movement frequency, \(F(1,15)=6.96, p=.02, \eta_{\mathrm{p}}{ }^{2}=\) 32. There was no significant difference in the proportion of time spent in the DRP bin associated with antiphase coordination \(\left(160^{\circ}-180^{\circ}\right)\) between the two intention conditions.

The results of a 2 (intention) x 2 (trial) mixed model ANOVA on scaling relations (S) from the PSD analysis were similar to those of the ANOVA on circular variance.

There was a significant main effect for trial, \(F(1,15)=\) 46.83, \(p=.001, \eta_{\mathrm{p}}^{2}=.76\), and a significant interaction between intention and trial, \(F(1,15)=5.12, p=.04, \eta_{\mathrm{p}}^{2}=\) .25. As seen in Figure 3, this interaction appears to be driven by the difference in the effects of intention between the baseline and test conditions, with a much larger increase in \(\alpha\) occurring in the frequency intention than the amplitude intention. The \(\alpha\) values for both intention conditions during the baseline trials were closest to the region associated with pink noise ( \(\mathrm{S} \approx-1, \alpha \approx 1\) ). The \(\alpha\) values for those intending to control amplitude during the test condition were characteristic of white noise ( \(\mathrm{S} \approx 0, \alpha \approx 0\) ), while those for participants asked to control movement frequency were closer to the region associated with anti-persistent, dependent behavior ( \(\mathrm{S} \approx 1, \alpha \approx-1\) ).

Given evidence that the effects of intention condition were most apparent during the test trials for both coordination strength and scaling relation, we chose to conduct a regression to determine whether circular variance could account for variation in scaling relation above and beyond that accounted for by intentional condition. A forward regression indicated that intention explained a significant proportion of variation in scaling relation, \(R^{2}=\) \(.42, F(1,15)=10.86, p=.005\), and was significantly predictive of scaling relation, \(b=.65, t(15)=3.30, p=\) .005, for the test trials. However, the predictive contribution of circular variance was not significant, as it only accounted for an additional \(2.8 \%\) of the variance in scaling relation above and beyond that accounted for by intention.

\section*{Discussion}

The task in the current study was constructed to demonstrate the effects of assigning participants different intentions for a simple motor task, as examined in performance variability and the employment of voluntary and involuntary control.


Figure 1. Mean circular variance for each of the intention conditions (Amplitude, Frequency), and under both trial conditions. Error bars show standard error.


Figure 2. a) Mean proportion of trial spent in each DRP bin during baseline trials for each intention condition (Amplitude, Frequency). b) Mean proportion of trial spent in each DRP bin during test trials for each intention condition (Amplitude, Frequency). Note: DRP bins are labeled by the midpoint of the range of relative phase values they contain, except for the \(0^{\circ}-20^{\circ}\) bin and the \(160^{\circ}-180^{\circ}\), which are referred to be the lowest and highest possible DRP values, respectively.

The two different intention conditions introduced are essentially equivalent with respect to a participant's resulting movement; maintaining a consistent speed will result in relatively consistent spacing between movements, and vice versa. As such, one might predict that providing participants with the instruction to control speed versus distance would have no effect on the amount of voluntary or involuntary control required to complete the task, and therefore no effect on the structure of performance variability. This does appear to be the case for the baseline trials. Consistent with previous use of continuation paradigms, measures of performance variability structure for
both intentions during the baseline trials fell within the region associated with self-similar, pink noise thought to indicate the use of voluntary control (Chen et al., 2001;


Figure 3. Mean scaling exponents for movement frequency in each of the intention conditions (Amplitude, Frequency), and under both trial condition, as assessed through PSD analysis. Error bars show standard error.

Gilden et al., 1995; Lemoine et. al, 2006; Torre, \& Delignières, 2008).

The equivalency in performance variability structure between the intention conditions was not, however, maintained during the test trials. In this case, past research universally predicts that the presence of a rhythmic stimulus will provide a source of involuntary control for the movement task (Chen et al., 2001; Delignieres \& Torre, 2009). As previously described, the difference in participant instructions with respect to the use of a rhythmic stimulus can result in two distinct kinds of structure in performance variability. While the mere presence of a rhythmic stimulus results in entrainment and random variability (Chen et al., 2001), the instruction to synchronize with the stimulus and the resulting opportunity to gain accuracy feedback about one's performance leads to anti-persistent changes (Delignieres \& Torre, 2009). In the current study, the instructions about use of the stimulus were the same for all participants; they were simply told to use the stimulus to help control their movement, following an explanation about what their intention for the task should be. The difference in intention alone appears to have affected the influence, and constraint, of the rhythmic stimulus on the structure of performance variability.

For those participants who were asked to control the amplitude of their movements, changes in performance during the test trial were found to be random, corresponding to white noise. This suggests that for someone intending to control the amplitude of their movements, the rhythmic stimulus provided involuntary control, but did not allow for sufficient accuracy feedback for corrective anti-persistent movement modulation. In contrast, for the participants
intending to control the frequency of their movements, measurements of variability structure were in the range associated with anti-persistent behavior. It therefore appears that the stimulus did allow for enough accuracy feedback about the timing of movements to support corrective processes by the participant. It is worth noting that the difference in intention conditions during the test trials was associated with differences in coordination, as well as the structure of performance variability. While there was significantly more inphase coordination between participant and stimulus movements for those intending to control frequency, any variation in performance associated with changes in coordination stability appears to be accounted for by intention.

In conclusion, our study has shown that the manipulation of intention alone appears to affect the use of voluntary and involuntary control for an environmentally constrained motor task, as reflected by differences in performance variability. These results also demonstrate that intending to control one specific task dimension over another can substantially alter the influence of any present environmental constraints. Therefore, in addition to identifying the role of intention in performance variability as an area worth further exploration, this study also sounds a cautionary note for research that aims to better understand the recruitment of voluntary and involuntary control and performance variability.

\section*{Acknowledgments}

We would like to thank MaryLauren Malone and Veronica Romero for helpful comments during the design of this study. We would also like to thank Guy Van Orden for his invaluable contributions to the field of complexity science.

\section*{References}

Batschelet E (1981). Circular Statistics in Biology, London: Academic Press.
Brown, C., \& Liebovitch, L. (2010). Fractal analysis (Vol. 165). Sage Publications, Incorporated.

Chen, Y., Ding, M., \& Scott Kelso, J. A. (2001). Origins of timing errors in human sensorimotor coordination. Journal of Motor Behavior, 33(1), 3-8.
Delignieres, D., Ramdani, S., Lemoine, L., Torre, K., Fortes, M., \& Ninot, G. (2006). Fractal analyses for 'short'time series: a re-assessment of classical methods. Journal of Mathematical Psychology, 50(6), 525-544.
Delignières, D., \& Torre, K. (2009). Fractal dynamics of human gait: a reassessment of the 1996 data of Hausdorff et al. Journal of Applied Physiology, 106(4), 1272-1279.
Delignières, D., Torre, K., \& Lemoine, L. (2009). Longrange correlation in synchronization and syncopation tapping: A linear phase correction model. PloS one, 4(11), e7822.
Eke, A., Herman, P., Bassingthwaighte, J., Raymond, G., Percival, D., Cannon, M., Balla, I. \& Ikrényi, C. (2000).

Physiological time series: distinguishing fractal noises from motions. Pflügers Archiv European Journal of Physiology, 439(4), 403-415.
Eke, A., Herman, P., Kocsis, L., \& Kozak, L. R. (2002). Fractal characterization of complexity in temporal physiological signals. Physiological measurement, 23(1), R1.
Ferrer-i-Cancho, R., \& Elvevåg, B. (2010). Random texts do not exhibit the real Zipf's law-like rank distribution. Plos One, 5(3), e9411.
Gilden, D. L. (2001). Cognitive emissions of \(1 / \mathrm{f}\) noise. Psychological review, 108(1), 33.
Gilden, D. L., Thornton, T., \& Mallon, M. W. (1995). 1/f Noise in Human Cognition. Science, 267, 1837-1839.
Hausdorff, J. M., Peng, C. K., Ladin, Z., Wei, J. Y., \& Goldberger, A. L. (1995). Is walking a random walk? Evidence for long-range correlations in stride interval of human gait. Journal of Applied Physiology, 78(1), 349358.

Hausdorff, J. M., Purdon, P. L., Peng, C. K., Ladin, Z., Wei, J. Y., \& Goldberger, A. L. (1996). Fractal dynamics of human gait: stability of long-range correlations in stride interval fluctuations. Journal of Applied Physiology, 80(5), 1448-1457.
Holden, J. G. (2005). Gauging the fractal dimension of response times from cognitive tasks. Contemporary nonlinear methods for behavioral scientists: A webbook tutorial, 267-318.
Holden, J. G., Choi, I., Amazeen, P. G., \& Van Orden, G. (2011). Fractal \(1 / f\) dynamics suggest entanglement of measurement and human performance. Journal of Experimental Psychology: Human Perception and Performance, 37(3), 935.
Holden, J. G., Van Orden, G. C., \& Turvey, M. T. (2009). Dispersion of response times reveals cognitive dynamics. Psychological review, 116(2), 318.
Ingber, D. E. (2003). Tensegrity I. Cell structure and hierarchical systems biology. Journal of Cell Science, 116(7), 1157-1173.
Jordan, K., Challis, J. H., \& Newell, K. M. (2007). Speed influences on the scaling behavior of gait cycle fluctuations during treadmill running. Human movement science, 26(1), 87-102.
Jordan, K., Challis, J. H., \& Newell, K. M. (2007). Walking speed influences on gait cycle variability. Gait \& posture, 26(1), 128-134.
Kiefer, A. W., Riley, M. A., Shockley, K., Villard, S., \& Van Orden, G. C. (2009). Walking changes the dynamics of cognitive estimates of time intervals. Journal of experimental psychology. Human perception and performance, 35(5), 1532.
Kuznetsov, N. A., \& Wallot, S. (2011). Effects of accuracy feedback on fractal characteristics of time estimation. Frontiers in Integrative Neuroscience, 5.
Lemoine, L., Torre, K., \& Delignieres, D. (2006). Testing for the presence of \(1 / \mathrm{f}\) noise in continuation tapping data.

Canadian Journal of Experimental Psychology, 60(4), 247-257.
Mandelbrot, B. B. (1982). The fractal geometry of nature. Times Books.
Marmelat, V., \& Delignieres, D. (2011). Complexity, coordination, and health: Avoiding pitfalls and erroneous interpretations in fractal analyses. Medicina, 47, 393--398.
Oullier, O., de Guzman, G. C., Jantzen, K. J., Lagarde, J., \& Kelso, J. S. (2008). Social coordination dynamics: Measuring human bonding. Social Neuroscience, 3(2), 178-192.
Philippe, P. (2000). Epidemiology and self-organized critical systems: An analysis in waiting times and disease heterogeneity. Nonlinear Dynamics, Psychology, and Life Sciences, 4(4), 275-295.
Rhodes, T., \& Turvey, M. T. (2007). Human memory retrieval as Levy foraging. Physica A, 385, 255-260.
Schmidt, R. C., Beek, P. J., Treffner, P. J., \& Turvey, M. T. (1991). Dynamical substructure of coordinated rhythmic movements. Journal of Experimental Psychology: Human Perception and Performance, 17(3), 635.
Torre, K., \& Delignières, D. (2008). Unraveling the finding of \(1 / \mathrm{f} \beta\) noise in self-paced and synchronized tapping: A unifying mechanistic model. Biological cybernetics, 99(2), 159-170.
Turvey, M. T., \& Moreno, M. A. (2006). Physical metaphors for the mental lexicon. The Mental Lexicon, 1(1), 7-33.
Van Orden, G. C., \& Holden, J. G. (2002). Intentional contents and self-control. Ecological Psychology, 14(1-2), 87-109.
Van Orden, G. C., Holden, J. G., \& Turvey, M. T. (2003). Self-organization of cognitive performance. Journal of Experimental Psychology: General, 132(3), 331.
Van Orden, G. C., Holden, J. G., \& Turvey, M. T. (2005). Human cognition and 1/f scaling. Journal of Experimental Psychology: General, 134(1), 117.
Van Orden, G. (2010). Voluntary performance. Medicina, 46(9), 581-594.
Varlet, M., Coey, C. A., Schmidt, R. C., \& Richardson, M. J. (2011). Influence of stimulus amplitude on unintended visuomotor entrainment. Human movement science.
Wallot, S., \& Van Orden, G. (2011). Toward a lifespan metric of reading fluency. International Journal of Bifurcation and Chaos, 21(04), 1173-1192.
Wallot, S., \& Van Orden, G. (2011). Nonlinear analyses of self-paced reading. The Mental Lexicon, 6(2), 245-274.
Warren, K., Cracium, G., \& Anderson-butcher, D. (2005). Everybody else is: Networks, power laws, and peer contagion in the aggressive recess behavior of elementary school boys. Nonlinear Dynamics, Psychology, and Life Sciences, 2, 155--173.
West, B. J., \& Deering, B. (1995). The lure of modern science: Fractal thinking (Vol. 3). World Scientific Publishing Company Incorporated.

\title{
Grouping by Similarity Helps Concept Learning
}

\author{
Erik Weitnauer (eweitnau@techfak.uni-bielefeld.de) \\ CITEC, Bielefeld University, Universitätsstr. 21-23, 33615 Bielefeld, Germany
}

\author{
Paulo F. Carvalho (pcarvalh@indiana.edu) \\ Department of Psychological and Brain Sciences, 1101 E 10th St Bloomington, IN 47405 USA
}

Robert L. Goldstone (rgoldsto@indiana.edu)
Department of Psychological and Brain Sciences, 1101 E 10th St Bloomington, IN 47405 USA

\author{
Helge Ritter (helge@techfak.uni-bielefeld.de) \\ CITEC, Bielefeld University, Universitätsstr. 21-23, 33615 Bielefeld, Germany
}

\begin{abstract}
In inductive learning, the order in which concept instances are presented plays an important role in learning performance. Theories predict that interleaving instances of different concepts is especially beneficial if the concepts are highly similar to each other, whereas blocking instances belonging to the same concept provides an advantage for learning lowsimilarity concept structures. This leaves open the question of the relative influence of similarity on interleaved versus blocked presentation. To answer this question, we pit withinand between-category similarity effects against each other in a rich categorization task called Physical Bongard Problems. We manipulate the similarity of instances shown temporally close to each other with blocked and interleaved presentation. The results indicate a stronger effect of similarity on interleaving than on blocking. They further show a large benefit of comparing similar between-category instances on concept learning tasks where the feature dimensions are not known in advance but have to be constructed.
\end{abstract}

Keywords: category learning; order effects; similarity

\section*{Introduction}

Inductive learning is an essential cognitive ability which, by abstracting from specific examples, allows the transfer of experience to new, similar situations. There is a significant body of evidence from cognitive psychology suggesting that comparison of multiple cases represents a particularly promising avenue for inductively learning difficult, relational concepts (Loewenstein \& Gentner, 2005). Comparison not only takes representations as inputs to establish similarities, but also uses perceived similarities to establish new representations (Hofstadter, 1996; Medin, Goldstone, \& Gentner, 1993; Mitchell, 1993). When we compare entities, our understanding of the entities changes, and this may turn out to be a far more important consequence of comparison than simply deriving an assessment of similarity. In this paper, we are interested in identifying optimal ways of organizing these comparisons, and the kinds of cases that should be optimally compared.

One major line of argument is that comparing instances of a concept with very dissimilar features should lead to the best
induction and generalization for the concept. If comparison serves to highlight commonalities between instances of the same concept while de-emphasizing differences, comparing instances that share irrelevant features could result in those features being retained in a learner's mental representation. This notion, called "conservative generalization" by Medin and Ross (1989) is that people will generalize as minimally as possible, preserving shared details unless there is a compelling reason to discard them. This, in turn, could limit generalizability to new, dissimilar cases. Some research is consistent with this conclusion. For example, Halpern, Hansen, and Riefer (1990) asked students to read scientific passages that included either "near" (superficially similar) or "far" (superficially dissimilar) analogies. The passages that included far analogies led to superior retention, inference and transfer compared to those featuring superficially similar comparison, which showed no benefit at all.

The conservative generalization principle predicts that increasing the similarity of simultaneously presented instances from one category will inhibit people's ability to discover the rule that discriminates between the two categories. The true, discriminating rule will need to compete with many other possible hypotheses related to the many other features shared by the compared instances. By this account, decreasing the similarity of the compared instances that belong within a category will make it more likely that the proper grounds for generalization are inferred, by eliminating misleading common features that lead to incorrect categorization rules.

Results of Rost and McMurray (2009) on young infants learning to discriminate pairs of similar words point into the same direction. These authors found that increasing the within-category variability of the to-be-learned words by having different speakers repeat them increases the infants' ability to discriminate between the words. One of the potential explanations they give for their results is that young infants might still be unsure about what feature dimensions are relevant for the task and the variability in the irrelevant dimen-
sions helps the infants to focus on the relevant, stable ones.
Another line of argument is that concepts which are highly similar to each other are better learned when instances of different concepts are interleaved. When learning to distinguish between several similar concepts, one major difficulty lies in identifying the subtle differences between them. Birnbaum, Kornell, Bjork, and Bjork (2012) suggested, in their discriminative contrast hypothesis, that interleaving instances of different concepts enhances the discriminative contrast between them and therefore helps with the task of spotting their differences, see also (Carvalho \& Goldstone, 2012; Kornell \& Bjork, 2008; Kang \& Pashler, 2012). Additionally, comparing very similar instances from different categories has the advantage that there are fewer random, irrelevant differences that compete for attention with the defining difference (see Winston, 1970, on "near misses").

In summary, the two lines of arguments described above predict that high similarity supports between-category comparison, while low similarity supports within-category comparison. Both types of comparisons are potentially important in learning concepts, but one might be more effective than the other for a specific learning task, depending on the specific task, context, experience, and structure of concepts (Goldstone, 1996).

In this paper, we compare the effect that similarity has on learning performance in blocked and interleaved presentation schedules. Carvalho and Goldstone (2012) recently conducted an experiment with a similar purpose. They manipulated the category structures in a perceptual categorization task towards more or less similarity, both within and between categories, and found this modulates the advantage of blocking and interleaving in the expected directions.

Our approach is different in three important ways. First, we manipulate similarity by grouping concept instances into either similar or dissimilar comparison, instead of switching between two separate sets of categories. Second, we designed the blocked and interleaved schedules in a way that they would enhance within- and between-category comparison, respectively, while still allowing for both types of comparisons. Therefore, the two argument lines above make opposite predictions on whether high similarity of instances shown closely together should help or hurt the induction and will allow for a direct comparison of effect strengths. Third, we use an inductive learning task, Physical Bongard Problems (PBPs), with a much larger feature-space.

This problem domain is inspired by the Bongard problems (Hofstadter, 1979; Bongard, 1970) and was recently introduced by Weitnauer and Ritter (2012) to study concept learning and categorization of dynamic, physical situations. Each problem consists of two sets of 2D physical scenes representing two concepts that must be identified. The scenes of one concept are on the left side, the scenes of the other concept on the right side. Figure 1, 2 and 3 show three example problems. What makes PBPs particularly interesting as a domain for concept learning is their open-ended feature space. Peo-
ple do not know in advance which features a solution might be based on (or indeed what the features are), and while some of the problems rely on features that are readily available such as shape or stability, others require the construction of features as a difficult part of the solution (e.g., the time an object is airborne or the direction a particular object in the scene is moving in) \({ }^{1}\). This intricate situation in which both features and concepts have to be identified at the same time is quite common in real life and people deal with it impressively well, while it is still considered a very hard problem in the Artificial Intelligence community.


Figure 1: PBP 08. The task is to identify the two concepts A and B represented by the scenes on the left and on the right side, respectively. This is the similarity version in which similar scenes are grouped by rows. The concept labels were not shown during the study. See the end of paper for the solution.

\section*{Experiment}

In this experiment we analyze the effects of different presentation schedules and similarity groupings on concept learning performance. We selected 22 PBPs and extended them by additional scenes so that the problems consist of sixteen training scenes and 8 test scenes each. Half of the scenes are shown on the left side and belong to category A (we name them A 1 , ..., A10) while the other half of the scenes are shown on the right side and belong to a different category B (we name them B1, ..., B10). All scenes were designed to fit into five similarity groups \(\{\mathrm{A} 1, \mathrm{~A} 2, \mathrm{~B} 1, \mathrm{~B} 2\},\{\mathrm{A} 3, \mathrm{~A} 4, \mathrm{~B} 3, \mathrm{~B} 4\},\{\mathrm{A} 5, \mathrm{~A} 6\), B5, B6\}, \{A7, A8, B7, B8\} and \(\{\mathrm{A} 9, \mathrm{~A} 10, \mathrm{~B} 9, \mathrm{~B} 10\}\), so that within-group similarity between the scenes is high, whereas between-group similarity is low.

During presentation, two scenes are always displayed simultaneously so that for each problem a sequence of six train-

\footnotetext{
\({ }^{1}\) Solutions can be based on a great variety of features and feature combinations, as geometrical or physical object features, the way a physical scene evolves over time, relations between the objects, or even potential interactions with the scene. Additionally, focusing on a subset of objects and aligning the scenes with each other is required to find some of the solutions.
}


Figure 2: PBP 18. This is the dissimilarity version in which similar scenes are positioned far from each other. See the end of paper for the solution.
ing scene pairs is shown to the participant. We vary the presentation order of scenes along two dimensions with two values each, resulting in four conditions. The first dimension, similarity grouping, controls whether similar scenes are shown temporally close to each other ("111122223333") or temporally far from each other ("132121323213"). We will refer to the former as "grouped by similarity" or "similar" and to the latter as "grouped by dissimilarity" or "dissimilar". Figure 4 depicts how scenes are positioned for both cases.

The second dimension, presentation schedule, controls whether the scenes that are shown simultaneously are from the same or from different categories (AA-BB-AA-BB-AA\(B B\) vs. \(A B-A B-A B-A B-A B-A B\), see Figure 5). We will refer to the former as "blocked" condition \({ }^{2}\) and to the latter as "interleaved" condition. In the blocked condition while within-category comparisons are facilitated by presenting scenes from the same category simultaneously, betweencategory comparisons can still be made between successive scene pairs, but involve higher memory demands. Analogously, the interleaved condition enhances between-category comparisons but still allows for within-category comparison across successive scene pairs.

We expected to find that grouping by similarity should improve learning performance for the interleaved condition and grouping by dissimilarity should improve performance for the blocked condition.

\section*{Subjects}

We conducted the experiment on Amazon Mechanical Turk \({ }^{3}\). Sixty-seven participants, all US-citizens, took part in the ex-

\footnotetext{
\({ }^{2}\) We use the term "blocked" to refer to a slightly different presentation schedule than it is usually done. Instead of showing all instances of one category before switching to the next, we only block two instances of one category and interleave these blocked pairs.
\({ }^{3}\) See Mason and Suri (2012) for an introduction to using Mechanical Turk as a platform for research.
}


Figure 3: PBP 24. This is the similarity version in which similar scenes are grouped by rows. See the end of paper for the solution.


Figure 4: Positions of the 12 training scenes for the conditions grouped by similarity (upper left corners) and grouped by dissimilarity (lower right corners).
periment in return for monetary compensation. Of these, we excluded 27 who did not finish all problems (most of them dropped out after seeing only a few) and another two that did not get at least one solution correct across the entire task. There was no need to use catch trials, because the subjects were required to write down the solutions as free text. Any cheating or automated answers would have become immediately apparent during our hand-coding of the solutions. The data from the remaining 38 participants was used in the following analyses. On average, participants solved 8.6 out of the 22 problems presented.

\section*{Material}

For each of the 22 problems, the training scenes were arranged in three rows, each with four scenes. We prepared two versions of each problem by placing the scenes at different positions. In the "grouped by similarity" version, the scenes were arranged in such a way that the scenes inside


Figure 5: The scene presentation schedule for blocked (top) and interleaved (bottom) presentation. The participant manually proceeds through the six states. In each state, two scenes (in white) are shown while the other scenes (in gray) are hidden.
each row are similar to each other. In the "grouped by dissimilarity" version, similar scenes were distributed over all rows. Figures 1 and 2 show an example of a dissimilarity and similarity version, respectively.

\section*{Design}

We used a \(2 \times 2\) factorial design. The study condition (presentation schedule: \(\{\) blocked, interleaved \(\} \times\) similarity grouping: \{similar, dissimilar\}) was randomly chosen for each problem in a within-subject manner.

\section*{Procedure}

The participants were first given a brief introduction to PBPs including an example problem with a solution. During the experiment, they could proceed through the scene pairs of each problem at their own pace by pressing a key. After they had viewed all scenes once, they were asked whether they thought they had found a solution. Then they needed to classify six test scenes which were randomly drawn from the eight available test scenes. The test scenes were shown one by one. Finally they had to type in a description of their solution or their best guess. Before moving on to the next problem, they were shown the problem with all training scenes at once together with the official solution. There was no time limit to the task. At the end of the experiment participants were debriefed on the study objectives and variables. The original experiment is available online at Weitnauer (2013).

\section*{Results}

We used two separate measures to evaluate learning success. First, we hand-coded the accuracy of each textual solution given by the participants. Some of the participants had difficulties remembering which side was left and which side was right, so they provided a correct solution but with sides swapped (e.g., writing "left: all objects are squares" and "right: all the objects are circles" when in fact the left-side objects were all circles and the right-side objects were all squares). These cases were counted as correct solutions.

The second measure is based on the proportion of test scenes that were classified correctly. Using this directly would be misleading for cases in which participants mixed up the sides. We therefore developed a consistency measure instead. This consistency measure is defined as \(\max (c, 6-c)-\)

3 , where \(c\) is the number of correctly classified scenes being minimally zero and maximally six. The consistency can take values between zero and three, where the latter corresponds to cases where either all test scenes were classified correctly or were all (consistently) classified wrongly. Figures 6 and Figure 7 show the average of these two measures for all four conditions.


Figure 6: Mean proportion of correct answers for blocked and interleaved presentation schedules and grouping of scenes by similarity or dissimilarity. There is a significant effect of similarity.


Figure 7: Mean consistency of test scene classifications for blocked and interleaved presentation schedule and grouping of scenes by similarity or dissimilarity. There is a highly significant effect of similarity.

We applied two separate \(2 \times 2\) repeated measures ANOVAs with presentation schedule (blocked vs. interleaved) and similarity grouping (similar vs. dissimilar) as factors to the proportion of correct responses and consistency measures. These analyses revealed a significant effect of similarity condition, \(F(1,37)=5.32, p=.03\) for the proportion of correct answers measure and \(F(1,37)=15.7, p=.0003\) for the consistency measure. There was no effect of schedule of presentation, or
interaction between the two factors for any of the measures (all \(p>.05\) ).

\section*{Discussion}

The data analysis revealed a positive effect of grouping scenes by similarity, independent of whether they were presented in a blocked or an interleaved schedule. We argue that this is explained by a strong positive effect of similarity on interleaving which more than compensates for any possible negative effect that similarity had on blocking.

The advantage of similarity for interleaving is in line with our expectations. Goldstone (1996) and the discriminative contrast hypothesis of Birnbaum et al. (2012) predict that direct comparison of instances from different categories highlights their differences (see also Carvalho \& Goldstone, 2012). Identifying differences between highly similar scenes is especially effective, as there are fewer superficial differences to compete with the defining one. This insight is already present in the desirable "near misses" in Winston (1970) work, where instances from different concepts that differ by just one feature are ideal for his algorithmic learner. Near misses provide clear evidence about what features are critical, concept-defining ones. Another possible contributing effect is that it is easier to structurally align two similar scenes than two very different scenes and this alignment process promotes noticing differences (Markman \& Gentner, 1993).

What might seem surprising at first is that similarity also improves learning performance in the blocked condition, given that theories like "conservative generalization" by Medin and Ross (1989) predict that similarity for blocked scenes will lead to many superficial similarities and therefore inferior performance compared to dissimilar scenes. However, the results can be explained in a way compatible with these theories. We designed both scheduling conditions in a way that allows for within- and between-category comparisons. Given this, negative effects of similarity on the former and positive effect of similarity on the latter will compete with each other. In the blocking condition, within-category comparisons were facilitated by showing scenes of the same category simultaneously, while scenes of different categories had to be compared sequentially.

Still, a strong positive effect of similarity on betweencategory comparison could mask a small negative effect of similarity on within-category comparison and lead to the overall improvement due to similarity that we found. What is indeed surprising is that, although learners were pushed towards attending to similarities with a paired comparison, they still exploited between-pair differences to find the solution.

We believe that one important reason for this might be found in the type of categorization task that was used. Due to its open ended feature space, participants had to identify or construct relevant feature dimensions as a major part of the challenge. Comparing similar scenes from different concepts provides the additional advantage of highlighting such feature
dimensions, an advantage that blocking of dissimilar scenes does not provide.
Implications for an Algorithmic Learner An interesting question is how the presented results could inform the implementation of a computational model of concept learning in open feature-spaces. A general observation is the fact that presentation order matters at all. This means that attending to the first scenes changes the way the following scenes are perceived and solution hypotheses that are formed. The limited memory capacity of humans makes it impossible to keep a detailed representation of all instances or a large number of hypotheses in mind and forces a decision on which aspects of an instance one should concentrate on and which information should be retained. The big challenge is that these decisions have to be made before knowing the answer to the problem and therefore before knowing what aspects are actually important. In open-ended feature spaces algorithmic learners could face similar problems because the \(a\)-priori construction of all possible features might be infeasible due to a combinatorial explosion, so dynamic processes that discover feature dimensions and concepts at the same time might be necessary.

The main insight from the present experiment is that between-category comparisons of similar instances are especially beneficial, as they promote learning, both by making new, potentially relevant feature dimensions more salient and by increasing the likelihood that a perceived difference is a defining one. Between-category comparisons should therefore play a privileged role in how active learning algorithms choose the next training example.
Pedagogical Implications Birnbaum et al. (2012) showed the benefit of interleaving for several concept learning tasks, and Carvalho and Goldstone (2012) proposed that this benefit is modulated by how similar the concepts are, so that in low-similarity cases blocking can be better. The current work provides a slightly different perspective. Our results suggest no direct advantage of interleaved or blocked presentation, but instead a greater potential of between- compared to within-category comparisons. This holds even for situations in which the between-comparison relies on sequentially shown instances while within-comparison can be made on the basis of simultaneously shown instances. A result that might directly inform the design of learning material is the big benefit of comparing similar scenes from different categories. The grouping of instances by similarity - instead of relying on a single similarity measure for a whole set of concepts - is a new, interesting dimension along which presentation order can be manipulated to optimize learning.

\section*{Acknowledgments}

This research was supported by Grant R305A1100060 from the Institute of Education Sciences Department of Education and Grant 0910218 from the National Science Foundation REESE program. PFC was also supported by Graduate Train-
ing Fellowship SFRH/BD/78083/2011 from the Portuguese Foundation for Science and Technology (FCT). We would like to thank Joshua de Leeuw, the author of the jsPsych library (Leeuw, 2013) which we used in our Mechanical Turk experiment and Spenser Benge for his help with stimuli preparation and response coding.

\section*{Solution to the problems}

PBP 08: unstable vs. stable
PBP 18: objects eventually touch vs. objects are eventually separated
PBP 24: several possible outcomes vs. one possible outcome

\section*{References}

Birnbaum, M., Kornell, N., Bjork, E., \& Bjork, R. (2012). Why interleaving enhances inductive learning: The roles of discrimination and retrieval. Memory \& Cognition, 1-11.
Bongard, M. M. (1970). Pattern recognition. Rochelle Park, N.J.: Hayden Book Co., Spartan Books.

Carvalho, P., \& Goldstone, R. (2012). Category structure modulates interleaving and blocking advantage in inductive category acquisition. In N. Miyake, D. Peebles, \& R. P. Cooper (Eds.), Proceedings of the thirty-fourth annual conference of the cognitive science society (pp. 186-191). Austin, TX: Cognitive Science Society.
Goldstone, R. (1996). Isolated and interrelated concepts. Memory \& Cognition, 24(5), 608-628.
Halpern, D., Hansen, C., \& Riefer, D. (1990). Analogies as an aid to understanding and memory. Journal of Educational Psychology, 82(2), 298.
Hofstadter, D. (1979). Gödel, escher, bach: an eternal golden braid. Harvester Press.
Hofstadter, D. (1996). Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought. Basic Books.
Kang, S., \& Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. Applied Cognitive Psychology, 26(1), 97-103.
Kornell, N., \& Bjork, R. (2008). Learning concepts and categories is spacing the "enemy of induction"? Psychological Science, 19(6), 585-592.
Leeuw, J. de. (2013). A javascript library for running behavioral experiments on the web. Available from https:// github.com/ jodeleeuw/ jsPsych
Loewenstein, J., \& Gentner, D. (2005). Relational language and the development of relational mapping. Cognitive Psychology, 50(4), 315-353.
Markman, A., \& Gentner, D. (1993). Structural alignment during similarity comparisons. Cognitive Psychology, 25, 431-431.
Mason, W., \& Suri, S. (2012). Conducting behavioral research on amazon's mechanical turk. Behavior Research Methods, 44(1), 1-23.
Medin, D., Goldstone, R., \& Gentner, D. (1993). Respects for similarity. Psychological review, 100(2), 254.
Medin, D., \& Ross, B. (1989). The specific character of
abstract thought: Categorization, problem solving, and induction. Advances in the psychology of human intelligence, 5, 189-223.
Mitchell, M. (1993). Analogy-making as perception: A computer model. MIT Press.
Rost, G. C., \& McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. Developmental Science, 12(2), 339-349.
Weitnauer, E. (2013). Mechanical turk study on scene ordering in PBPs. Available from http://perceptsconcepts .psych.indiana.edu/experiments/ew/pbp3/
Weitnauer, E., \& Ritter, H. (2012). Physical bongard problems. Artificial Intelligence Applications and Innovations, 157-163.
Winston, P. (1970). Learning structural descriptions from examples (Tech. Rep.). DTIC Document.

\title{
Modeling Bilingual Children's Acquisition of Complex Sentences in German
}

\author{
Daniel Wiechmann (wiechmann@anglistik.rwth-aachen.de) \\ RWTH Aachen University \\ Department of English Linguistics
}

Judith Steinfeld (judith.steinfeld@rwth-aachen.de)
RWTH Aachen University
Department of English Linguistics

\author{
Elma Kerz (kerz@anglistik.rwth-aachen.de) \\ RWTH Aachen University \\ Department of English Linguistics
}

\begin{abstract}
Although Bilingual First Language Acquisition research has increased considerably over the past few decades, there is still much controversy regarding the rate of development, i.e. the question whether bilinguals lag behind their monolingual peers in various aspects of language. Some studies have found similar rates of development, whereas others have found that bilingual children lag behind their monolingual peers. The current study contributes to this discussion of (dis)similar rates of development by investigating bilingual children's acquisition of German complex sentence constructions involving adverbial clauses (ACs). Our findings are consistent with usage-based approaches to language acquisition, which predict that bilingual acquisition should proceed slower due to learners having less exposure, on average, to each language.
\end{abstract}

Keywords: bilingual first language acquisition; language production; rate of development; complex constructions

\section*{Introduction}

\section*{Bilingual First Language Acquisition (BFLA)}

There has been an increasing interest in early bilingual language acquisition. Commonly, this interest involves the question of whether the cognitive and developmental path (course of development) and time course (rate of development) of language learning by bilingual children is the same as that of their monolingual peers. Although prevailing theoretical models of language acquisition have different views regarding the influence of endogenous and exogenous factors on the acquisition of abstract linguistic structures and patterns, they agree upon the idea that monolingual and bilingual language learning is qualitatively equivalent in that children go through the same series of developmental phases, starting off with single word productions, followed by two and multi-word utterances before they finally develop the capacity to produce complex sentences (de Houwer, 1995, 2009; Meisel, 1986). Prior research comparing the rate of development in monolingual and bilingual learners has produced somewhat mixed
results. Some studies have found similar rates of development (cf. Paradis, Crago, \& Genesee, 2005/2006; Paradis, 2010; Pearson \& Fernández, 1994), whereas others have found that bilingual children lag behind their monolingual peers (Gathercole, 2002a, 2002b, 2007; Nicoladis, Palmer, \& Marentette, 2007; Pérez-Leroux, Pirvulescu, \& Roberge, 2009). The current study contributes to this discussion of (dis)similar rates of development by investigating the bilingual acquisition of complex sentences involving adverbial clauses (ACs) in German, which mark the last stage in a series of milestones mentioned above.

\section*{Usage-based theory and BFLA}

Usage-based (UB) theories belong to a family of emergentist models, which assume that the development of language competence is contingent on the experience with language (O’Grady, 2008; Tomasello, 2003; Lieven, \& Tomasello, 2008). A conservative assumption about BFLA is that bilingual children, on average, receive less language input per language than their monolingual peers. UBtheories thus predict that reduction in overall exposure to a language should negatively affect children's rate of acquisition (Gathercole, \& Hoff, 2007; Paradis, Nicoladis, Crago, Genesee, 2011).

\section*{Usage-based theory and the acquisition of complex sentences}

Complex sentences are grammatical assemblies consisting of multiple clauses. Two types of clauses are distinguished: (i) sentences including coordinate clauses and (ii) sentences including a matrix clause and a subordinate clause. Complex sentences containing subordinate clauses can be further subdivided into three basic sub-types: constructions with complement clauses, relative clauses and adverbial clauses. The most comprehensive study on the acquisition of complex sentences framed within UB-theory is Diessel (2004). Diessel proposes that complex sentences develop through two different types of processes: Complex
sentences involving complement and relative clause constructions develop through a process termed clause expansion. Complex sentences containing adverbial clauses develop through a process termed clause integration, in which two independent sentences are merged into a single bi-clausal unit. The earliest adverbial clauses produced by children are thus free-standing (isolated) clauses introduced by an adverbial subordinator, which are only pragmatically linked to a previous utterance. Over time, children learn to elaborate these structures and integrate them with a matrix clause. The last step in mastering complex sentences involves developing the capacity to produce sentence initial subordinate clauses, which impose greater demands on (verbal) working memory as initial clauses require that the producer has planned the entire complex structure at the onset of the utterance (Gibson, 1998; Hawkins, 2004; Temperley, 2007). Initial adverbial clauses thus develop later and their frequency, at first, is limited to specific subordinators. Another finding of Diessel's (2004) study is that children's early productions of complex sentences are tied to specific lexical expressions. The emergence of more schematic representations of such constructions takes place only after children have been exposed to a sufficient number of types to generalize over. This is reflected in the fact that children only gradually elaborate their repertoire of adverbial subordinators and overextend already learned types to situations where those types are semantically inadequate (e.g. use of a causal subordinator to express concessive or other adverbial relations). Two additional, more general indicators of language proficiency are the mastery of syntactic differences in German main and subordinate clause (verb second in main clauses vs. verbfinal positioning in subordinate clauses) (cf. Clahsen \& Muyskens, 1986; Miller, 1976; Park, 1981; Roeper, 1973) and mean length of utterance (MLU). MLU has been shown to be an important measurement of a child's gross language development and was found to correlate with the development of morphological and syntactic skills in young children (Brown, 1973; Parker, \& Brorson, 2005). Building on this research, the present study sets out to derive statements about differences in the rate of development of complex sentence constructions from measurements of five indicators of language proficiency:
1. Proportion of isolated/integrated adverbial clauses
2. Proportion of sentence-initial adverbial clauses
3. Proportion of misused subordinators
4. Proportion of correct verb position in sub clause
5. (Log) length of adverbial construction (MLU)

\section*{Method}

All relevant data were elicited by having children watch a 6.5 minutes episode of a popular stop-motion animated children's television series. The children were then given a visual cue to a particular scene and asked to describe what happened in that scene. The children's' responses were audio-recorded and transcribed.

\section*{Participants}

A total of 50 children from 4 to 6 years old participated in the study: 25 bilingual child participants (German in combination with another language) and a control group consisting of 25 monolingual children. \({ }^{1}\) All bilingual children have started learning both languages before the age of three (McLaughlin, 1984). The children participating in the study were selected from several kindergartens with families from different socioeconomic backgrounds (SES; low SES ( \(\sim 5 \%\) ), middle SES ( \(\sim 75 \%\) ), high SES ( \(\sim 20 \%\) ). The final proportional distributions of monolingual and bilingual children across these three categories exhibit minor, statistically insignificant asymmetries (the proportion of bilinguals was a little greater than expected assuming statistical independence in high and low SES categories). All parents and kindergartens, agreed to participate in the study.

\section*{Data}

The elicitation procedure resulted in 27,301 word tokens produced by monolingual learners and 21,023 word tokens produced by bilingual learners. From these corpora, all instances of the target constructions were extracted by way of manual inspection of the corpus data, yielding a total of 1,023 data points (601 from monolinguals, 422 from bilinguals). The extracted data were annotated with information pertaining to the indicators of language competency listed in the preceding section. Table 1 presents relevant descriptive statistics of the sample.

Table 1: Descriptive Statistics
\begin{tabular}{ccc}
\hline & monolingual & bilingual \\
\hline age (mean) & \(5 ; 4\) & \(5 ; 5\) \\
age (SD) & 7.16 & 6.87 \\
number of ACs total & 601 & 422 \\
ACs integrated & \(62.73 \%\) & \(50 \%\) \\
ACs in initial position & \(23.34 \%\) & \(13.74 \%\) \\
subordinator misused & \(1 \%\) & \(17.77 \%\) \\
correct verb position & \(72.38 \%\) & \(66.11 \%\) \\
length (mean) & \(12.36 \%\) & 10.55 \\
length (SD) & 7.2 & 5.91 \\
\hline
\end{tabular}

The language proficiency levels of the monolingual and bilingual children were compared with respect to five indicators of language proficiency. To test whether and to what extent the proficiency levels of mono- and bilingual learners differ, we asked: Does competence indicator x differ significantly between bilingual and monolingual children after controlling for age? The data were analyzed using linear and logit mixed effects models in which each of

\footnotetext{
\({ }^{1}\) There are total of 12 different language pairs within the data. German was acquired in combination with one of the following languages: Albanian, Arabian, Bosnian, English, French, Hungarian, Kurdish, Persian, Russian, Spanish, Turkish or Vietnamese.
}
the five indicators of language proficiency was modeled as a function of the Boolean predictor BILINGUAL (monolingual vs. bilingual), a control variable AGE (measured in months) and SUBJECT as a random effect. \({ }^{2}\). We checked for normality and homogeneity by visual inspections of plots of residuals against fitted values. For all models, the significance of the predictor BILINGUAL was assessed through model comparison: For each model, we conducted likelihood ratio tests to see if a model including BILINGUAL is significantly better than the corresponding model containing only AGE and the random effect (SUBJECT).

Model 1: Length
Linear Mixed Model fit by REML approximation; p-values estimated via Markov Chain Monte Carlo (MCMC) sampling ( \(n=10.000\) ). Outcome variable (log) length of utterance.

Table 2: (Log) Length Model
\begin{tabular}{cccc}
\hline Random effects: & & Variance & Std.Dev. \\
\hline child & (Intercept) & 2.09 & 1.44 \\
\hline Fixed effects: & Estimate & MCMCmean & pMCMC \\
\hline (Intercept) & 1.21 & 1.23 & 0.0001 \\
bilingual & \(\mathbf{0 . 1 7}\) & \(\mathbf{0 . 1 7}\) & \(\mathbf{0 . 0 1 5}\) \\
age & 0.01 & 0.01 & 0.0012 \\
\hline
\end{tabular}


Figure 1: Effect of BILINGUAL on (log) LENGTH of construction

The analysis reveals that the there is a weak but statistically significant effect of BILINGUAL on the (logged) length of the construction (a log likehood ratio test comparing null model and model including bilingual yields \(\operatorname{Pr}(\) Chi) < 0.05). The positive coefficient estimate in Table 2 indicates that the average construction length of monolingual learners is greater than that of bilingual learners, when age is controlled for.

Model 2: Integration
Mixed Logit Model fit by Laplace approximation. Outcome variable is proportion of integrated (= non-isolated) adverbial clauses.

\footnotetext{
\({ }^{2}\) All data were analyzed using R (R Core Team, 2012) and the functions provided in the R packages lme4 (Bates \& Maechler, 2009) and languageR (Baayen, 2009)
}

Table 3: Integration Model
\begin{tabular}{cccc}
\hline Random effects: & & Variance & Std.Dev. \\
\hline child & (Intercept) & 2.09 & 1.44 \\
\hline Fixed effects: & Estimate & SE & \(\operatorname{Pr}(>|\mathbf{z}|\) ) \\
\hline (Intercept) & -5.64 & 1.87 & 0.00254 \\
bilingual & \(\mathbf{1 . 3 0}\) & \(\mathbf{0 . 4 6}\) & \(\mathbf{0 . 0 0 4 6 1}\) \\
age & 0.07 & 0.03 & 0.01469 \\
\hline
\end{tabular}


Figure 2: Effect of BILINGUAL on proportion of integrated adverbial clauses (AC)

The analysis reveals a weak but statistically significant effect of BILINGUAL on the proportion of integrated adverbial clauses (a log likehood ratio test comparing null model and model including bilingual yields \(\operatorname{Pr}(\mathrm{Chi})<0.01)\) : Monolingual learners produce significantly more complex constructions (integrated ACs), when age is controlled for.

Model 3: Verb Position in Subordinate Clause
Mixed Logit Model fit by Laplace approximation. Outcome variable is proportion of adverbial clauses with correct (=clause final) verb position.

Table 4: Verb Position Model
\begin{tabular}{crcc}
\hline Random effects: & & Variance & Std.Dev. \\
\hline child & (Intercept) & 1.61 & 1.27 \\
\hline Fixed effects: & Estimate & \(\mathbf{S E}\) & \(\mathbf{P r}(>|\mathbf{z}|)\) \\
\hline (Intercept) & 0.94 & 1.61 & 0.56 \\
bilingual & \(\mathbf{0 . 4 6}\) & \(\mathbf{0 . 4 0}\) & \(\mathbf{0 . 2 5}\) \\
age & -0.01 & 0.03 & 0.75 \\
\hline
\end{tabular}


Figure 3: Effect of BILINGUAL on correct verb position in adverbial clauses (AC)

The analysis reveals a tendency for monolingual learners to produce a greater number of correct verb position but the effect is not statistically significant (a log likehood ratio test comparing null model and model including bilingual yields \(\operatorname{Pr}(\) Chi) \(>0.25)\).

Model 4: Subordinator Misuse
Mixed Logit Model fit by Laplace approximation. Outcome variable is proportion of incorrectly used subordinators

Table 5: Subordinator Misuse Model
\begin{tabular}{cccc}
\hline Random effects: & & Variance & Std.Dev. \\
\hline child & (Intercept) & 5.90 & 2.43 \\
\hline Fixed effects: & Estimate & SE & \(\operatorname{Pr}(>|\mathbf{z}|)\) \\
\hline (Intercept) & 2.49 & 4.10 & 0.54 \\
bilingual & \(\mathbf{- 3 . 7 4}\) & \(\mathbf{1 . 0 5}\) & \(\mathbf{0 . 0 0 0 3 1}\) \\
age & -0.08 & 0.06 & 0.22 \\
\hline
\end{tabular}

The analysis reveals a medium sized and statistically significant effect of BILINGUAL on the proportion of correctly used subordinators (a log likehood ratio test comparing null model and model including bilingual yields \(\operatorname{Pr}(\) Chi) < 0.001): Bilingual learners produce significantly more semantically inadequate subordinators.


Figure 4: Effect of BILINGUAL on misuse of adverbial subordinator

Model 5: Adverbial Clause Position
Mixed Logit Model fit by Laplace approximation. Outcome variable is proportion of sentence-initial adverbial clauses. This model was fit to the subset of the data that contains only those utterances that contain at least two clausal constituents, so that the adverbial clause can either precede or follow the main clause ( \(\mathrm{N}=588\) ).

Table 6: AC Position Model
\begin{tabular}{crrc}
\hline Random effects: & & Variance & Std.Dev. \\
\hline child & (Intercept) & \(5.39 \mathrm{E}-20\) & \(2.32 \mathrm{E}-10\) \\
\hline Fixed effects: & Estimate & SE & \(\operatorname{Pr}(>|\mathbf{z}|\) ) \\
\hline (Intercept) & 0.21 & 1.17 & 0.8546 \\
bilingual & \(\mathbf{0 . 5 6}\) & \(\mathbf{0 . 2 4}\) & \(\mathbf{0 . 0 1 9 9}\) \\
age & -0.03 & 0.02 & 0.0755 \\
\hline
\end{tabular}


Figure 5: Effect of BILINGUAL on proportion of sentenceinitial ACs

The analysis reveals a weak but statistically significant effect of BILINGUAL on the proportion of sentence-initial adverbial clauses (a log likehood ratio test comparing null model and model including bilingual yields \(\operatorname{Pr}(\) Chi) < 0.05): Bilingual learners produce significantly fewer sentence initial adverbial clauses.

\section*{Discussion}

Prior research into the rate of bilingual and monolingual development has produced somewhat inconclusive results. While some studies have found similar rates of development, other studies found that bilingual children lag behind their monolingual peers in various aspects of language. Furthermore, the majority of research on the accuracy of bilingual production has been devoted to earlier phases of grammatical development such as the acquisition of the past tense (e.g. Paradis et al., 2011), the acquisition of mass/count nouns (Gathercole, 2000a) or the acquisition of grammatical gender (Gathercole, 2000b). Our findings contribute to this area of research by providing additional evidence from later stages of grammatical development, namely complex sentences, which constitute the last milestone in the acquisition of grammar (cf. Clahsen, 1986).

The research question guiding our analysis was as follows: Are bilingual children less proficient than their monolingual peers in the production of German complex sentences with adverbial clauses? Experience-driven or usage-based theories of language predict that bilingual children's acquisition of complex sentences should proceed slower due to them having less exposure, on average, to each language. We tested this general prediction across multiple dimensions. The five dimensions that served as responses in our models jointly define the space in which we measured language proficiency of monolingual and bilingual learners. We observed that bilinguals in fact lag behind in four out of five dimensions: their adverbial constructions are shorter, less often integrated into a complex sentential structure and when they are integrated, they are less often placed in sentence-initial position. Furthermore, bilingual productions exhibited a greater amount of violations of the semantic usage conditions of adverbial subordinators. This suggests that bilingual children have not yet developed a very nuanced set of words to link verbalizations of two events.

Overall, the present work clearly indicates that bilinguals around age five have not caught up on their monolinguals peers in the domain of complex sentences. The only dimension where performance was equivalent across the two groups concerns the positioning of the finite verb in German subordinate clauses. However, as both groups are still quite removed from adult-level performance ( \(<80 \%\) correct usages in both groups), this finding cannot be attributed to the children's having mastered this grammatical domain. Our results also display a considerable amount of inter-individual differences as evidenced by rather pronounced intercept adjustments in the models. Some bilingual children even outperform some monolingual children across all dimensions. While some variation is expected to result from inter-individual differences in learning performance, prior research suggests that a large portion of the observed differences may also relate to various types of group-level differences (cf. Werker, \& Byers-Heinlein, 2008, for an overview). These include variation due to specific language pairs (Döpke, 2000; Holm, \& Dodd, 1999; Müller, 1999; Müller \& Hulk, 2001; Nicoladis, 2003; Paradis, \& Navarro, 2003; Yip \& Matthews, 2000), contexts of exposure (Kazuya, 1998), social status of the languages (De Houwer, 2007; Pearson, 2007), socioeconomic status (Morton, \& Harper 2007), and language dominance (Cutler, Mehler, Norris, Segui, 1989; Flege, MacKay, \& Piske, 2002; Gathercole, \& Môn Thomas, 2009). Disentangling the effects of these variables from the effects of an individual's learning performance is subject to further investigation.

Acknowledgments
We thank the children who participated in this study, their parents for their permission to conduct the study and all nursery school teachers for their cooperation.

\section*{References}

Baayen, H. R. (2009). Corpus linguistics in morphology: morphological productivity. In A. Lüdeling \& M. Kytö (Eds.), Corpus linguistics: an international handbook. Berlin: de Gruyter.
Bates, D., \& Maechler, M. (2009). lme4: Linear mixedeffects models using S4 classes. Available at: http://CRAN.R-project.org/package=lme4.
Brown, R. (1973). A first language: the early stages. Cambridge, Mass.: Harvard University Press.
Clahsen, H. (1986). Die Profilanalyse. Ein linguistisches Verfahren für die Sprachdiagnose im Vorschulalter. Berlin: Spiess.
Clahsen, H. \& Muysken, H. (1986). The Availability of Universal Grammar to Adult and Child Learners: A Study of the Acquisition of German Word Order. Second Language Research, 2, 93-119.
Cutler, A., Mehler, J., Norris, D., Segui, J. (1989). Limits on bilingualism. Nature 340, 229-230.
De Houwer, A. (1995). Bilingual Language Acquisition. In P. Fletcher \& B. MacWhinney (Eds.), Handbook of child language. Cambridge, MA: Blackwell.
De Houwer, A. (2007). Parental language input patterns and children's bilingual use. Applied Psycholinguistics 28(3), 411-424.
De Houwer, A. (2009). Bilingual First Language Acquisition. Bristol: Multilingual Matters.
Diessel, H. (2004). The Acquisition of Complex Sentences. Cambridge: Cambridge University Press.
Döpke, S. (2000). Generation of and retraction from crosslinguistically motivated structures in bilingual first language acquisition. Bilingualism: Language and Cognition, 3, 209-226.
Flege, J.E., MacKay, I. \& Piske, T. (2002) Assessing bilingual dominance. Applied Psycholinguistics 23(4), 567-598.
Gathercole, V. C. M. (2002a). Command of the mass/count distinction in bilingual and monolingual children: An English morphosyntactic distinction. In D. K. Oller \& R. E. Eilers (Eds.), Language and literacy in bilingual children (pp. 175-206). Clevendon, U.K.: Multilingual Matters.
Gathercole, V. C. M. (2002b). Grammatical gender in bilingual and monolingual children: A Spanish morphosyntactic distinction. In D. K. Oller \& R. E. Eilers (Eds.), Language and literacy in bilingual children (pp. 207-219). Clevendon, U.K.: Multilingual Matters.
Gathercole, V. C. M. (2007). Miami and North Wales, so far and yet so near: A constructivist account of morphosyntactic development in bilingual children. International Journal of Bilingual Education and Bilingualism, 10, 224-247.
Gathercole, V. M. \& Hoff, E. (2007). Input and acquisition of language: Three questions. In E. Hoff \& M. Shatz (Eds.), The handbook of language development (pp. 107127). Oxford: Blackwell.

Gathercole, V.M., \& Môn Thomas, E. (2009). Bilingual first-language development: Dominant language takeover, threatened minority language take-up. Bilingualism: Language and Cognition, 12, 213-238
Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. Cognition, 68, 1-76.
Gonzalez, V. (2005). Cultural, linguistic, and socioeconomic factors influencing monolingual and bilingual children's cognitive development. In V. Gonzalez \& J. Tinajero (Eds), Review of Research and Practice, Vol. 3 (pp. 67-104). Mahwah, NJ: Lawrence Erlbaum.
Hawkins, J. (2004). Efficiency and complexity in grammars. Oxford: Oxford University Press.
Holm, A. \& Dodd, B. (1999). A longitudinal study of the phonological development of two Cantonese-English bilingual children. Applied Psycholinguistics, 20, 349376.

Kazuya, Hiroko (1998). Determinants of language choice in bilingual children: the role of input. International Journal of Bilingualism 2(3), 327-346.
Lieven, E. \& Tomasello, M. (2008). Children's first language acquisition from a usage-based perspective. In P. Robinson \& N. Ellis (Eds.), Handbook of cognitive linguistics and second language acquisition (pp. 168196). New York: Routledge.

McLaughlin, B. (1984). Second-language acquisition in childhood, volume 1 : preschool children (2 \({ }^{\text {nd }}\) ed.). Hillsdale, NJ: Lawrence Erlbaum Associations.
Meisel, J. M. (1986). Word order and case marking in early child language. Evidence from simultaneous acquisition of two first languages: French and German. Linguistics, 24, 123-183.
Miller, M. (1976). Zur Logik der frühkindlichen Sprachentwicklung. Stuttgart : Klett.
Morton, J.B. \& S.N. Harper (2007) What did Simon say? Revisiting the bilingual advantage. Developmental Science 10(6), 719-726.
Müller, N. (1999). Transfer in Bilingual First Language Acquisition. Bilingualism: Language and Cognition, 1 (3), 151-171.

Müller, N. \& Hulk, A. (2001). Crosslinguistic Influence in Bilingual Language Acquisition: Italian and French as Recipient Languages. Bilingualism: Language and Cognition, 4 (1), 1-21.
Nicoladis, E. (2003). Cross-linguistic transfer in deverbal compounds of preschool bilingual children. Bilingualism: Language and Cognition, 6, 17-31.
Nicoladis, E., Palmer, A. \& Marentette, P. (2007). The role of type and token frequency in using past tense morphemes correctly. Developmental Science, 10, \(237-\) 254.

O'Grady, W. (2008). The emergentist program. Lingua, 118, 447-64.
Paradis, J. (2010). Bilingual children's acquisition of English verb morphology: Effects of language exposure, structure complexity, and task type. Language Learning,

60 (3), 651-680.
Paradis, J., Crago, M. \& Genesee, F. (2005/2006). Domainspecific versus domain-general theories of the deficit in SLI: Object pronoun acquisition by French-English bilingual children. Language Acquisition, 13, 33-62.
Paradis, J. \& Navarro, S. (2003). Subject realization and cross linguistic interference in the bilingual acquisition of Spanish and English: What is the role of the input? Journal of Child Language, 30 , 371-393.
Paradis, J., Nicoladis, E., Crago, M. \& Genesee, F. (2011). Bilingual children's acquisition of the past tense: a usagebased approach. Journal of Child Language, 38, 554-578.
Park, T. Z. (1981). The development of syntax in the child with special reference to German. Innsbruck.
Parker, M. \& Brorson, K. (2005). A comparative study between mean length of utterance in morphemes (MLUm) and mean length of utterance in words (MLUw). First Language, 25, 365-376.
Pearson, B.Z. (2007). Social factors in childhood bilingualism in the United States. Applied Psycholinguistics 28 (3), 399-410.
Pearson, B. Z., \& Fernández, S. C. (1994). Patterns of interaction in the lexical growth in two languages of bilingual infants and toddlers. Language Learning, 44, 617-653.
Pérez-Leroux, A. T., Pirvulescu, M. \& Roberge, Y. (2009). Bilingualism as a window into the language faculty: The acquisition of objects in French-speaking children in bilingual and monolingual contexts. Bilingualism: Language and Cognition, 12, 97-112.
R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.
Roeper, T. (1973). Connecting children's language and linguistic theory. In T. Moore (Eds.), Cognitive development and the acquisition of language. New York: Academic Press.
Temperley, D. (2007). Minimization of dependency length in written English. Cognition, 105(2), 300-33.
Tomasello, M. (2003). Constructing a language: A usagebased theory of language acquisition. Cambridge, MA: Harvard University Press.
Werker, J.F. \& Byers-Heinlein, K. (2008). Bilingualism in infancy: first steps in perception and comprehension. Trends in Cognitive Sciences, 12(4), 144-151.
Yip, V. \& Matthews, S. (2000). Syntactic transfer in a Cantonese-English bilingual child. Bilingualism: Language and Cognition, 3, 193-208.

\title{
On the Robustness of Intuitions in the two best-known Trolley Dilemmas
}

\author{
Alex Wiegmann (awiegma@gwdg.de) \\ Institute of Psychology, Gosslerstraße 14 \\ 37073 Göttingen, Germany
}

\author{
Matthias Lippold (matthias.lippold1@stud.uni-goettingen.de) \\ Institute of Psychology, Gosslerstraße 14 \\ 37073 Göttingen, Germany
}

\author{
Robert Grigull (robert.grigull@stud.uni-goettingen.de) \\ Institute of Psychology, Gosslerstraße 14 \\ 37073 Göttingen, Germany
}

\begin{abstract}
The Bridge dilemma (pushing a heavy man from a bridge in front of a train that would otherwise kill five persons) and the Switch dilemma (redirecting a train that would otherwise kill five persons onto another track where it kills one person) are presumably the two best-known moral dilemmas in philosophy and psychology. In this paper we claim that people's intuitions about what to do in Bridge are robust, while intuitions about Switch can be influenced rather easily. In doing so, we strongly disagree with Broeders and colleagues (2011) who recently argued for exactly the opposite claim. We discuss their interpretation of previous findings that were supposed to motivate their claim, present findings from previous studies that strongly support my claim, and report on failed attempts to replicate and present an experiment in which participants were willing to revise their judgment for Switch but not for Bridge.
\end{abstract}

Keywords: moral judgment; trolley dilemmas; robustness of moral intuitions; priming; transfer effects.

\section*{Introduction}

Bridge and Switch are presumably the two best-known hypothetical moral dilemmas. They were first extensively used as thought experiments in moral philosophy and later also in empirical studies in moral psychology (cf. Waldmann, Nagel, \& Wiegmann, 2012; Gräfenhain \& Wiegmann, 2012).

In both scenarios five people are threatened by an out of control train. In Bridge the only possibility to save the five persons is to throw a heavy person from a bridge in front of the train, resulting in killing the heavy person and saving the five (Thompson, 1985). In Switch the threatening train can be redirected away from the five onto another track where one person would die in the collision with the train (Foot, 1967). Research in moral psychology has shown that the majority of people disapprove intervening in Bridge while they tend to approve the action in Switch (Waldmann et al., 2012).

In their recent paper, Broeders, Bos, Müller, and Ham (2011) make extensive use of these two dilemmas. They argue that previous research, especially the research by Greene and colleagues (Greene, Sommerville, Nystrom,

Darley, \& Cohen, 2001; Greene, Nystrom, Engell, Darley, \& Cohen, 2004) indicates that people's decision in Switch is made fast and without hesitation, while it takes them longer and they are more hesitate to make a decision in Bridge. Following this line of argument, Broeders and colleagues (2011) claim that people’s judgments in Bridge can easily be manipulated by priming them with rules as "save lives" and "do not kill", while this kind of priming supposedly has no effect on people's judgment in Switch. In three experiments they seemingly confirm this claim.

In this paper we argue for an opposite claim: Judgments concerning Switch can be manipulated rather easily while judgments concerning Bridge are rather robust.

\section*{Arguing against Broeders and colleagues’ interpretation of previous findings}

Broeders and colleagues' (2011) claim is motivated by the following line of argument. Research by Greene and colleagues (Greene et al., 2001, Greene et al., 2004) suggests that when people have to deal with Bridge the anterior cingulated cortex (ACC) shows increased activity. Activation of the ACC is assumed to indicate people's feeling of uncertainty. Moreover, people’s longer reaction times in Bridge, as compared to Switch, are also assumed to indicate uncertainty. This uncertainty is then interpreted as people's struggling to choose between the two rules "Do not kill" and "Save Lives". Hence, by priming one of the rules and thereby making it more accessible to subjects, their intuition about what to do in Bridge allegedly follows the primed rule. In contrast to Bridge, Switch elicits low ACC activity and people respond fast to it, supposedly indicating certainty. Hence, judgments concerning Switch are assumed to be robust and not to follow the primed rule.

At first glance, this line of reasoning sounds plausible. However, a closer look at the cited studies reveals that they do not provide compelling evidence in support of Broeders’ and colleagues' (2011) claim that people are uncertain of what to do in Bridge. Remember that this claim is based on two observations, namely people's longer reaction times and higher ACC activation in Bridge as opposed to Switch. However, there is no evidence that people's reaction-times
were longer for Bridge. While in their first fMRI study, Greene and colleagues (2001) did not report reaction times for Bridge, they explicitly state in their follow up study (Greene et al., 2004) that reaction times for Bridge were short.

What about the other finding that was also interpreted as people feeling uncertain about what to do in Bridge, namely the high ACC activation when people respond to this dilemma? First of all, the reported results in the fMRIstudies by Greene and his colleagues \((2001,2004)\) are based on brain activity averages for groups of dilemmas (about twenty in each group). Hence, to inferring conclusions from these averages to specific cases are just not valid.

Secondly, Greene and colleagues (2004) do not interpret high ACC activation as indicating uncertainty but as a conflict of emotion and cognition or, more precisely, cognitive effort to override a prepotent emotional response (cf. Stroop effect, Stroop, 1935). Their interpretation explains the aforementioned finding of people’s longer reaction-times when choosing an utilitarian (cognitive) option in personal moral dilemmas, because people have to override a strong emotional response not to intervene (Greene et al., 2001). In the same way, Greene's et al. interpretation of ACC activity as indicating a conflict of emotional and cognitive (utilitarian) considerations can explain why reaction times were only longer for people under cognitive load, which were namely those who chose the utilitarian option (Greene, Morelli, Lowenberg, Nystrom, \& Cohen, 2008). In contrast, interpreting ACC activity as indicating uncertainty cannot account for these findings because it would predict longer reaction times in dilemmas with high ACC activity independently of the option (cognitive vs. emotional) people choose.

Thirdly, it is simply not the case that the cited studies provide any evidence for higher ACC activation in Bridge, as compared to Switch. In their first fMRI study (Greene et al., 2001) ACC activation is not measured at all. In the follow-up study (Greene et al., 2004) ACC activity in difficult and easy personal moral dilemmas was compared. Since Switch was in neither of these groups, we do not have any evidence on the level of ACC activation in this dilemma. Moreover, due to relatively fast reaction times, Bridge was classified as an easy personal dilemma and it was found that ACC activity in these dilemmas was significantly lower than in difficult moral dilemmas.

\section*{Previous research indicating intuitions about Bridge to be robust - but not about Switch}

So far, we have dismissed Broeders and colleagues’ (2011) argument that was supposed to motivate their claim that Bridge is easy to influence, as compared to Switch. Now we shall present empirical findings that strongly speak in favor of my claim, that is, if any of the two dilemmas can be influenced rather easily then Switch is the one.

Lanteri, Chelini, and Rizello (2008) presented participants with the Switch and the Bridge dilemma. In one condition, participants first had to judge Switch and then Bridge
afterwards. In the other condition, Switch was preceded by Bridge. Although responses to Bridge remained unaffected by the order of presentation, fewer participants were willing to intervene in Switch when Switch was preceded by Bridge. The authors interpret their results as evidence that Switch may be perceived in more than one way, while they speculate that the emotions triggered by Bridge may be evolutionarily sound and hard wired into our species, making it more robust than reactions to Switch that are assumed to be a result of moral reasoning.

Lombrozo (2009) conducted a very similar experiment. The only difference to Lanteri et al. (2008) was that participants were allowed to read both dilemmas before they were asked to judge them. Again, participants who saw Switch first provided higher permissibility ratings than those who saw it after Bridge. Responses for Bridge were unaffected.

Petrinovich and O’Neill (1996) conducted several experiments in which participants were asked to judge a sequence of moral dilemmas, among them Switch and Bridge, where the order of presentation was manipulated between subjects. While ratings for Switch often significantly differed as a function of whether it was presented as the first or last dilemma, ratings for Bridge remained unaffected.

Finally, Wiegmann, Okan, and Nagel (2012) also found that people's judgments for Switch can be influenced by first presenting other scenarios, while people's judgments for Bridge were not affected. Moreover, Wiegmann, and Okan (2012) tried and failed to raise ratings in favor of the proposed action in Bridge. In one experiment they urged participants to justify their ratings in Switch, assuming that subjects' justification is something like "save as many lives as possible" and that this forced justification would raise subjects' ratings for Bridge. In another experiment, they tried to raise subjects’ ratings for Bridge by first presenting them with a scenario in which there was only enough time to pull one of two switches. One switch prevented one person, the other three persons from being killed. Presenting this scenario first was also supposed to make a rule like "save the most lives possible" salient. However, neither attempt succeeded in influencing ratings for Bridge.

\section*{Replication Experiments}

What follows are two attempts to replicate Broeders and colleagues' (2011) findings of their first experiment. We limit my replication attempts to their first of the total of three experiments for the following reason. All three experiments are based on the same rationale, namely to prime participants with one of the two rules. The only way the three experiments differ is how priming was implemented. In their first experiment, priming was implemented by asking participants to read a story and to answer two questions about the rule "Save lives" or "Do not kill". In Experiment 2 participants were asked to solve a sliding puzzle that resulted in a symbol supposed to prime participants with one of the two rules. In the third
experiment participants were subliminally primed. Hence, their first experiment is very similar to the experiments described in the preceding section. It might be possible, if unlikely, that the findings in their second and third experiment can be replicated even if it is not possible for the findings in their first experiment. However, since priming in their second and third experiment was implemented in a rather subtle way, as compared to reading a story in the first experiment, failing to replicate the findings in the first experiment would already strongly limit the scope of the claim that intuitions about Bridge can rather easily be manipulated while intuitions about Switch are rather robust.
In the light of what has been said so far, what prediction is to be made regarding Broeders and colleagues’ (2011) experiment in which participants had to read stories designed to prime them with the rule "Save lives" vs. "Do not kill"? Surely, everything points to the prediction that Bridge will not be affected by their manipulation. With regards to Switch things are not that clear, because there are no previous experiments in which it was tried to influence ratings for Switch by priming rules.

\section*{First Replication Attempt}

Participants 352 subjects, each receiving \(£ 0.50\), were recruited via an online database located in the U.K. They were invited via an email. The email contained a link that directed them to the experiment. Mean age of the participants was 47 years and 4 months ( \(S D=15\) years, 7 months), \(61 \%\) were female.

Design, Procedure, and Materials Participants were randomly assigned to one of the conditions of a 2 (primed rule: "Save lives" vs. "Do not kill") \(\times 2\) (dilemma: Bridge vs. Switch) factorial design.

After reading a cover story participants were asked to read a short story which was supposed to prime them with either the "save lives" or "do not kill"-rule. The "save lives" story goes as follows:
During missions abroad soldiers have to apply to certain rules. This is called a mandate. During a conflict in 1994 in Congo, Africa, the mandate of the UN peace keeping force present in the country was to use armed violence to end the conflict. As a result of several warnings beforehand and during the conflict by the UN commander in chief on the spot, the Canadian lieutenant-general Roméo Dallaire, UN-soldiers were allowed to conduct offensive actions. They were both allowed to use their weapons to protect the civilian population, as well as when they themselves were attacked. The mandate was based on the principle "Save lives".
For the "Do not kill" version participants read:
During missions abroad soldiers have to apply to certain rules. This is called a mandate. During a conflict in 1994 in Congo, Africa, the mandate of the UN peace keeping force present in the country was not to use any armed violence to end the conflict. Despite several
warnings beforehand and during the conflict by the UNcommander in chief on the spot, the Canadian lieutenant-general Roméo Dallaire, UN-soldiers were not allowed to conduct any offensive actions. They were only allowed to use their weapons when they themselves were attacked. The mandate was based on the principle "Do not kill".
Afterwards, they were first asked to write down how they would act in line with the corresponding rule. Furthermore, they were asked to write down what the specific rule meant for them in their daily lives. Subsequently, participants were asked to rate how they feel at the moment on the 20 items of the PANAS, an instrument to collect self-report measures of emotion (Watson, Clark, \& Tellegen, 1988). The PANAS was included as a filler task, as well as to determine whether the primes triggered any unwanted positive or negative affective reaction.

The PANAS was followed by asking participants to read and respond to either the Bridge or the Switch dilemma.
Bridge reads as follows:
You are standing on a footbridge crossing a railroad track. While you are standing next to a stranger, suddenly a runaway trolley comes hurtling down the railroad track. Further down the railroad track five people are working and they cannot possibly leave the railroad track in time. If the trolley proceeds on its present course it will crash into the five railroad workers and they will be killed in a fatal accident. The only way to save the five people is to push this man off the bridge and into the path of the trolley. The body of this person will break the speed of the trolley as a result of this the trolley will stop and the five persons will survive. The person thrown from the footbridge will certainly die.
Switch reads:
You are standing beside the switch of a railroad track. Suddenly a runaway trolley comes hurtling down the tracks. Further-on down the railroad track five people are working and they cannot possibly leave the railroad track in time. If the trolley proceeds on its present course it will crash into the five railroad workers and they will be killed in a fatal accident. You can save these five people by diverting the trolley onto a different set of railroad tracks. The different railroad track has only one person on it, into which the trolley will crash.
This person will be killed as a result of this.
After reading the dilemma participants were asked eleven questions about their willingness to intervene which they could indicate on a scale ranging from 1 (certainly not) to 7 (certainly yes). All items were then averaged to form a reliable scale indicating the willingness to intervene in the dilemma ( \(\alpha=.80\) ).

Finally, participants were asked four questions to find out whether they were aware of the purpose of the experiment.

Results and Discussion Eleven subjects were excluded because at least one of two independent raters coded them as being aware of the purpose of the experiment.

As in the study by Broeders and colleagues (2011), the prime did not have an effect on the positive or negative subscale of the PANAS.

Figure 1 clearly shows that neither Bridge nor Switch was affected by a priming scenario. A \(2 * 2\) ANOVA yielded the typical main effect of dilemma, \(F(1,337)=80.70\), \(p<.000001, \eta_{p}^{2}=.19\). However, there was no main effect of prime ( \(p>.75\) ) and no significant interaction ( \(p>.6\) ). Hence, Broeders and colleagues' finding could not be replicated although many more subjects participated in this experiment, resulting in a higher test power


Figure 1: Willingness to intervene (on a scale from 1 to 7) in Bridge and Switch as a function of manipulated accessibility of the rules "Save lives" and "Do not kill". Higher bars indicate greater willingness to intervene. Error bars represent standard error of means.

\section*{Second Replication Attempt}

This time we tried to replicate Broeders’ et al. (2011) findings in our experimental lab in Goettingen. This was done to counter objections claiming that online experiments are not reliable (although the typical main effect of dilemma was found). The design and procedure was the same as in the first replication attempt with two exceptions. The PANAS was left out to strengthen the influence of the primes, and participants were only asked two questions concerning their willingness to intervene in Switch or Bridge since the correlation of the eleven questions asked in


Figure 2: Willingness to intervene (on a scale from 1 to 7) in Bridge and Switch as a function of manipulated accessibility of the rules "Save lives" and "Do not kill". Higher bars indicate greater willingness to intervene. Error bars represenB762 standard error of means.

Broeders and colleagues’ and in my first replication attempt was very high.

Participants 220 participants, mostly psychology students, were recruited via the institute's database. Participants were credited with course credit or paid \(7 €\) /hour. Mean age of the remaining \(N=172\) participants was 24 years and 7 months ( \(S D=6\) years, 6 months), \(77 \%\) were female.

Results and Discussion 48 participants were excluded from the analysis because they knew the dilemmas (41) or seemed to identify the purpose of the experiment. Figure 2 clearly shows that neither Bridge nor Switch was affected by a priming scenario. A \(2 * 2\) ANOVA yielded the typical main effect of dilemma, \(F(1,168)=47.55, p<.001, \eta_{\mathrm{p}}{ }^{2}=.22\). However, there was again no main effect of prime ( \(p>.45\) ) and no significant interaction ( \(p>.6\) ).

\section*{Judgment Revision Experiment}

This experiment aims to investigate the robustness of Switch and Bridge by giving participants the chance to later revise their initial judgment.

Participants 158 subjects, each receiving \(£ 0.50\), were recruited via an online database located in the U.K.

Design, Procedure, and Materials The experiment was conducted on the Internet. Upon clicking on a link they received via e-mail, participants were redirected to a website containing the experiment. They read general instructions familiarizing them with the rating scale and asking them to read the following scenario carefully and to take their task seriously. Afterwards, they were randomly assigned to one of two conditions. In Bridge_Switch participants were first presented with Bridge and then Switch, in Switch_Bridge ist was the other way around. In both conditions participants had the chance to revise their judgment for the first scenario after they had seen the second scenario. Both scenario descriptions were accompanied by an illustration of the initial situation.

For each scenario participants were asked whether the proposed action should be done. To indicate their judgment participants could mark one point on a 6-point likert scale ranging from 1 ("certainly no") to 6 ("certainly yes").

After participants were given the chance to revise their judgment for the first scenario, they were asked some demographic questions and a simple logical question to identify participants who did not take the experiment seriously.

Results and Discussion 27 participants were excluded from the analysis because they did not finish the experiment, finished it in less than 40 seconds, or failed to answer the logical question. As it can easily be seen in Figure 3 the aforementioned asymmetrical transfer effect between Bridge and Switch was replicated. While the ratings for Bridge did not differ significantly depending on whether it was
presented first ( \(M=2.4, S D=1.51\) ) or second ( \(M=2.60, S D=\) 1.64), \(t(129)=.74, p=.46\), ratings for Switch were significantly decreased when Switch was presented second ( \(M=3.02, S D=1.50\) ), as compared to ratings for Switch when presented first ( \(M=4.45, S D=1.36\) ), \(t(129)=4.18, p<.00001\).

When we consider the difference of a scenario's first rating vs. the revision rating a similar picture arises. The first rating for Bridge \((M=2.4, S D=1.51)\) did not significantly differ from the revision rating \((M=2.44\), \(S D=1.59\) ), \(t=.35, p=.73\). In contrast to this, the revision rating for Switch ( \(M=3.88, S D=1.63\) ) did significantly differ from the first rating for Switch ( \(M=4.45, S D=1.36\) ), \(t=4.32\), \(p<.0001\).

The results strongly suggest that people's intuitions about Bridge are robust while their intuitions about Switch were significantly influenced when Switch was preceded by Bridge or when Bridge was presented after Switch and people were then given the chance to revise their judgment for Switch.

This pattern of results contradicts Broeder's et al. (2011) claim that people's intuitions about Switch are robust but not for Bridge.


Figure 3: Willingness to intervene (on a scale from 1 to 6 ) in Bridge and Switch as a function of whether the scenario was shown first or preceded by the other scenario (first vs. second rating). The dashed lines represent revised ratings for the first scenario after participants were presented with the second scenario. Higher bars indicate greater willingness to intervene. Error bars represent standard error of means.

\section*{Conclusion}

In this paper we argued that people's intuitions about Bridge are rather robust while their intuitions about Switch are rather easy to influence. This claim stands in sharp contrast to Broeders and colleagues' (2011) claims. We argued that Broeders and colleagues’ interpretation of previous findings that were supposed to motivate claim is not sound. Moreover, we reviewed previous findings that strongly point in the opposite direction.

In line with my claim, replicating the findings of Broeder's et al. first experiment failed online as well as in the lab. Furthermore, the results the revision experiments also count in my favor.

Given the important role that Bridge and Switch play in philosophy as well as in psychology, it is important that wrong claims about them are swiftly corrected to avoid that new research is based on false premises.

\section*{Funding}

This research was supported by a grant of the Deutsche Forschungsgemeinschaft (DFG WA 621/21-1), and the Courant Research Centre 'Evolution of Social Behaviour', University of Göttingen (funded by the German Initiative of Excellence).

\section*{Acknowledgements}

Thanks are due to Jonas Nagel, Michael Waldmann, Gunda Johannes, Matthias Lippold, and Robert Grigull for very helpful comments and proofreading.

\section*{References}

Broeders, R., van den Bos, K., Müller, P. A., \& Ham, J. (2011). Should I save or should I not kill? How people solve moral dilemmas depends on which rule is most accessible. Journal of Experimental Social Psychology, 47(5), 923-934.
Foot, P. (1967). The problem of abortion and the doctrine of the double effect. Oxford Review, 5, 5-15.
Gräfenhain, M., \& Wiegmann, A. (2012). The Scientific Study of Morals. In: Lütge, C. (Ed.), Handbook of the Philosophical Foundations of Business Ethics (chapter 81). Springer Press.

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., \& Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. Cognition, 107, 1144-1154.
Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., \& Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. Neuron, 44, 389400.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., \& Cohen, J. D. (2001). An fMRI study of emotional engagement in moral judgment. Science, 293, 2105-2108.
Lanteri, A., Chelini, C. \& Rizzello, S. (2008). An experimental investigation of emotions and reasoning in the trolley problem. Journal of Business Ethics, 83, 789804.

Lombrozo, T. (2009). The role of moral commitments in moral judgment. Cognitive Science, 33, 273-286.
Petrinovich, L., \& O'Neill, P. (1996). Influence of wording and framing effects on moral intuitions. Ethology and Sociobiology, 17, 145-171.
Stroop, J.R. (1935). Studies of interference in serial verbal reactions. Journal of Experimental Psychology. 12, 643662.

Thomson, J. J. (1985). The trolley problem. Yale Law Journal, 94, 1395-1415.
Waldmann, M. R., Nagel, J., \& Wiegmann, A. (2012). Moral judgment. In K. J. Holyoak \& R. G. Morrison
(Eds.), The Oxford Handbook of Thinking and Reasoning (pp. 364-389). New York: Oxford University Press.
Watson, D., Clark, L. A., \& Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. Journal of Personality and Social Psychology, 54, 1063.
Wiegmann, A., \& Okan, J. (2012). Order effects in moral judgments. Searching for an explanation. In N. Miyake, D. Peebles, \& R. P. Cooper (Eds.), Proceedings of the 34th Annual Conference of the Cognitive Science Society (pp. 1143-1148). Austin, TX: Cognitive Science Society.
Wiegmann, A., Okan, Y., \& Nagel, J. (2012). Order effects in moral judgment. Philosophical Psychology, 25, 813836.

\title{
Diversity, Collaboration, and Learning by Invention
}

\author{
Jennifer Wiley (jwiley@uic.edu) \\ Department of Psychology, 1007 W. Harrison Street Chicago, IL 60607 USA
}

Olga Goldenberg
Department of Psychology, 1007 W. Harrison Street
Chicago, IL 60607 USA

\author{
Andrew F. Jarosz \\ Department of Psychology, 1007 W. Harrison Street \\ Chicago, IL 60607 USA
}

Michael Wiedmann
Institut für Psychologie, Engelbergerstr. 41
D-79085 Freiburg DE

\author{
Nikol Rummel \\ Institut für Erziehungswissenschaft, Universitätsstraße 150 \\ D-44801 Bochum DE
}

\begin{abstract}
Learning-by-invention is an approach to mathematical instruction where small groups explore possible methods of solution before learning the "right answer" (e.g., Schwartz \& Martin, 2004; Kapur \& Bielaczyc, 2011). In a series of studies we have been investigating the effects of group composition in terms of math ability on learning by invention. An initial result showed that groups consisting of a mix of both high and low math ability students generated a broader range of solution attempts when asked to invent a formula for standard deviation compared to more homogeneous math ability groups. Moreover, this wider range of solution alternatives predicted better performance on quizzes following a lesson on the topic. Subsequent work is suggesting that who emerges as the leader of the group matters. Ongoing analyses are also exploring which features of the collaborative discourse are critical for students to take advantage of the affordances of learning by invention.
\end{abstract}

Keywords: Collaboration, Learning, Problem Solving.

\section*{Introduction}

It is said that the road to success is paved with failure. It's also said that those who do not learn from their mistakes are destined to repeat them. The provocative implication of these aphorisms is that there may be ways in which failure may be instrumental for successful learning, as long as one is able to take something away from the failure experiences. This is the premise behind learning-by-invention activities. In learning-by-invention, students are asked to attempt to create a mathematical formula to accomplish a goal before an instructional lesson is provided about the canonical approach. The experience of working in a problem space before being told a correct answer may lay the groundwork for future conceptual understanding, and thereby prepare
students for future learning. And, these benefits might accrue when solvers become aware of what approaches do not work, or become aware of constraints, obstacles, or desired properties for a solution through previous failures. There is now substantial evidence that having students engage in learning-by-inventions activities in small groups can lead to better understanding of new mathematical and statistical formulas compared to more traditional, direct methods of instruction (e.g., Schwartz \& Martin, 2004; Kapur, 2012; Kapur \& Bielaczyc, 2011). One main question for our investigations is whether the composition of the small groups in terms of their relative expertise or math ability might affect the likelihood that group members are able to take advantage of learning-by-invention activities. A second main question is whether there are critical features of the group interactions, such as in who emerges as a discourse leader, or what is said during group discussions, that can be shown to facilitate learning.

Although one could expect that groups where all members possess superior math skills would be more successful at any mathematical problem solving activity, another hypothesis is that there may be advantages to being in a group where there are a variety of backgrounds, perspectives or viewpoints that can be contributed. In particular, these investigations are exploring whether diversity in small groups may be one key to unlocking the potential benefits of learning-by-invention activities.

Obviously working in groups with students with more advanced math skills or knowledge may help students with less advanced skills or knowledge by exposing them to advanced math concepts or ideas that they might not consider when working alone. However, it is also possible that collaborating with students with less math knowledge
might contribute to more successful problem solving or learning by students with more math knowledge. The work of Webb and others has suggested that more skilled students may benefit from teaching or explaining math concepts to others (e.g. Webb, 1980). In addition, to the extent that students with different mathematical backgrounds might approach problem solving in different ways, then diverse groups may help all members think about a broader range of possible solution approaches which may be particularly important when creative, inventive or innovative thinking is required (Canham, Wiley, \& Mayer, 2012; Dunbar, 1995; Wiley \& Jensen, 2005; Wiley \& Jolly, 2003).

Previous research on learning-by-invention tasks has suggested that groups who generate the widest range of possible solution attempts during the invention phase experience the best learning from the activity (Kapur \& Bielaczyc, 2011). Based on this, we predict that the composition of the group in terms of their math expertise should matter, and that there may be special affordances to working in diverse groups. In addition, for groups whose members demonstrate the best understanding of the new principle following learning-by-invention, we explore what features of their discussion may have contributed to their success.

\section*{Consistency in Tea}

Small groups of undergraduates worked together on an invention activity before receiving a lesson on the standard deviation formula. For the invention activity, students were given data sets representing yearly antioxidant levels for 5-6 years of tea grown by three tea growers. Students were told that a company wished to buy tea from the grower with the most consistent levels of antioxidants from year to year and the company has asked for the students' help. They were prompted to generate as many invented formulas or step-bystep instructions as they could for how they could compute the consistency of antioxidant levels for each tea grower.

\section*{Methodology}

Two populations of undergraduate Psychology students at the University of Illinois at Chicago participated in this study \({ }^{1}\). Complete data are available for 25 triads who participated as part of a Research Methods course, and 20 triads from an Introductory Psychology course.
Math ability/expertise was estimated based on a mediansplit derived from historical data from this student population. Students with Math ACT scores of 24 or below were considered to have relatively low skill, and those with

\footnotetext{
\({ }^{1}\) Participants from the Research Methods Sample, who were more advanced in their studies, were found to outperform the Introductory Psychology Sample on the quiz, \(F(1,125)=5.90, p<\) \(.02, \eta^{2}=.05\). Importantly, this did not interact with the group composition factor, \(\mathrm{F}<1.07\), which meant the two samples could be collapsed in order to increase power, while the sample variable was retained as a covariate in all aggregated analyses reported below (for analyses of samples separately, see Wiedmann, Leach, Rummel, \& Wiley, 2012).
}
scores of 25 or above were considered to have relatively high skill. A score of 25 puts students in the 80th percentile in national norms. Students categorized as having low ( \(M=21.1, \quad S D=2.91\) ) versus high math skill ( \(M=28.5\) \(S D=2.78\) ) differed significantly on the Math ACT, \(t(122)=\) \(14.46, p<.001\). Of the 45 groups, all students were considered to have low math skill in 11 groups, all students were considered to have high skill in 9 groups, and 25 groups had a mix of high and low skill members. Students were not informed about the nature of their group compositions.

Groups first worked on the invention task for 30 minutes, and discussions were video recorded for the Introductory Psychology groups. For all groups, the worksheets from the invention activity were collected. Following the invention activity, participants individually read through an overview of the standard deviation formula and a worked example of how to compute standard deviation. Following instruction, all students were given a quiz to assess learning outcomes. Two items asked students to apply the formula of standard deviation to a new problem about the weather, and a third item required them to use standard deviation to invent standardized scores in order to compare two students' test performances across different courses. Each item asked students to explain the mathematical reasoning behind their answers. This quiz served as the assessment of learning outcomes from the activity.

\section*{Solution Attempts and Quiz Performance}

Coding The group worksheets from the invention activity (and video protocols when available) were coded for their inclusion of a variety of different solution approaches to the problem. An initial coding scheme was developed based on categories used by Kapur (2012). It included 5 main categories: 1) computing central tendencies and sums, 2) graphical representations, 3) frequency counts, 4) computing differences between adjacent scores, and 5) computing ranges and deviations from the mean. The final coding scheme with 22 subcategories was established post hoc based on an examination of the solutions that were actually obtained so that each distinct solution type had its own subcategory. To score the data, coders assigned each solution attempt to one of the 22 codes. They then determined whether an instance of each subcategory was represented in the written artifacts or not using 0,1 coding. The total number of different solution approaches was computed by adding the number of subcategories that had at least one instance present in the worksheet or discussion (i.e., the total of the 0,1 coding across the 22 categories).

In addition, a task analysis of understanding the standard deviation formula was used to identify several critical insights that students might reach during their discussions. The first is that methods such as noticing a high score, graphing histograms or bar graphs, summing scores or computing central tendencies will not help or are not sufficient to quantify consistency. Noticing differences in the range of values across data sets is an important first step
toward understanding variance. Two other key insights are that variance is best computed in relation to some reference point (such as a mean), and that somehow variations in positive and negative directions need to be preserved so that they do not cancel out when summed. Based on this analysis, solution attempts that included recognition of range, deviations from the mean, and the need to consider absolute values were all categorized as being of higher quality, and a subtotal of higher quality solution approaches was computed in addition to the overall total number of different solution approaches.

Quiz responses were scored by categorizing each explanation according to the mathematical concept that was referenced. The same basic categories were applied across the 3 problems. Explanations that focused on central tendencies, sums, or maximum scores earned 1 point. Explanations that focused on ranges or differences between scores earned 2 points. If explanations included an incorrect approximation of the SD formula, they received 3 points. If explanations demonstrated a correct use of the SD formula they received 4 points. Combining across the three items, a maximum of 12 points could be reached and the final explanation quality composite score was computed as a proportion of that total. Cronbach's \(\alpha\) among the three quiz items was .80. Krippendorff's \(\alpha\) indicated good interrater reliability on all coding metrics ( \(>.77\) ).

Quiz Performance An ANCOVA showed a significant effect of group composition on quiz performance, \(F(2,123)\) \(=12.41, p<.01, \eta^{2}=.17\). Planned comparisons indicated that students in the all-low math groups had lower scores on the quizzes than students in either the mixed or all-high groups, who did not differ in quiz performance.
A follow-up analysis was performed to see if group heterogeneity affected low-skill and high-skill students differently. As shown in Figure 1, both high- and low-skill members seemed to benefit from participation in mixed groups. A \(2 \times 2\) ANCOVA (Math Skill \(x\) Group Heterogeneity) revealed two significant main effects. As might be expected, high skill students did better than low skill students, \(F(1,122)=28.44, p<.01, \eta^{2}=.19\). In addition, the main effect for group heterogeneity, \(F(1,122)\) \(=6.29, p=.01, \eta^{2}=.05\), and the lack of a significant interaction, \(F<1\), indicated that both high-skill and lowskill students benefited from working in heterogeneous (mixed) groups.

Solution Attempts Average totals of different solution approaches as a function of group composition are shown in Figure 2. Examples of the inscriptions made on worksheets during the different kinds of solution attempts are shown in Figure 3. An ANCOVA on the total number of different solution approaches showed a significant effect of group composition, \(F(2,41)=8.55, p=.001, \eta^{2}=.29\). Planned comparisons indicated that the mixed groups considered significantly more different solution approaches than the alllow and all-high groups, who did not differ.


Figure 1: Quiz performance by group composition
When only higher quality solution approaches were considered, a different pattern emerged. An ANCOVA on the number of higher quality representations included in the group worksheets showed a significant effect of group composition, \(F(2,41)=9.47, p<.001, \eta^{2}=.32\). Planned comparisons indicated that the all-low groups considered fewer different high-quality solution approaches than the all-high and mixed groups, who did not differ. Although the mixed groups also tended to include higher numbers of lowquality solution approaches, this effect did not reach significance, \(F(2,41)=2.76, p<.08, \eta^{2}=.12\).


Figure 2: Solution attempts by group composition
Mediational analyses suggested that it was the discussion of a wide range of solution approaches during learning-byinvention activities, including the number of higher quality solution attempts, that mediated the effects of group composition. Heterogeneity predicted the variety of representations, \(B=1.83(S E=.27), t(126)=6.61, p<.05\), and variety of representations predicted quiz performance, \(B\) \(=.02(S E=.01), t(126)=2.47, p<.05\). The total effect of heterogeneity on quiz performance was also significant, \(B=\) \(.09(S E=.03), t(126)=2.84, p<.05\). However, this relationship decreased to non-significance when the mediating influence of the variety of representations was included in analysis, \(B=.04(S E=.04), t(126)=1.23, p=\) . 22 .


Figure 3: Example inscriptions from solution attempts

In addition, the indirect effect (the mediated effect) of heterogeneity on quiz performance through representation variety was 0.05 ( \(\mathrm{SE}=0.02\) ), and the \(95 \%\) bias-corrected confidence intervals for the size of the indirect effect did not include zero, (.01, .08) which shows that the indirect effect was significant at a \(\mathrm{p}=.05\) level (Preacher \& Hayes, 2004, 2008; Shrout \& Bolger, 2002). Taken together, these findings provide evidence for full mediation. This analysis suggests that heterogeneity in groups led to better quiz performance because it affected the variety of solutions that were discussed during the learning-by-inventing activity. Additional analyses showed that the benefits of solution diversity during group discussion were demonstrated to contribute to better quiz performance even when the math ability of the students was taken into account.

\section*{Analyses of Group Interactions}

The second phase of analyseshas been attempting to understand conditions led to the success of the more diverse groups. In this pass through the data, the potential effects of group interactions, such as who emerges as a discussion leader, as well as the quality and content of group discussions, on learning outcomes are examined. Research on decision making groups using hidden profile paradigms has demonstrated the important role of leaders and experts in information sharing. For example, recognition of expertise has been found to be critical for increasing contributions to the group by expert members (Franz \& Larson, 2002). In addition, group members are more likely to share valuable information when they are assigned a high status position, such as a group leader (Wittenbaum, Hollingshead, \& Botero, 2004). This supports the hypothesis that the expertise of group leaders may be an important predictor of effective information sharing in learning-by-invention tasks that can subsequently influence learning outcomes. Differences in the group discourse and their relation to learning outcomes are also being explored.

Coding Discourse coding was performed on the 14 mixed group discussions for which video recordings were available. Leadership was operationalized by identifying the group member who contributed the largest proportion of utterances to the group discussion. Seven of the groups in the sample had a leader high in math skill, and seven of the groups had a leader low in math skill.
A second goal for the discourse analyses was to explore the content and nature of the discourse acts. Each utterance made by a group member was coded into one of the following categories: (a) solution proposal, (b) clarification request or response, (c) evaluative comment, (d) comment related to group task coordination, (e) calculation, (f) comment on expertise, (g) comment about being stuck or at impasse, or (h) off-task comment.

Leadership When the effects of these two leader types were examined, we found that groups with high math leaders discussed more high quality solution attempts ( \(M=3.00, S D\)
\(=.38)\) than groups with low math leaders \((M=1.57, S D=\) .79), \(t(12)=2.97, p<.05\). In terms of learning outcomes, members of groups with high math leaders scored higher on the quiz assessing their understanding of the standard deviation formula ( \(M=.82, S D=.16\) ) than members of groups with low math leaders \((M=.68, S D=.16), t(40)=\) \(2.78, p<.01\). The results of a \(2 x 2\) between-subjects ANOVA (math skill by leader type) with quiz performance as the dependent variable revealed no significant interaction \((F(1,38)<1)\), suggesting that the expertise of the leader benefited both high and low skill students similarly.

Content of Discussions On average, these groups contributed around 180 utterances during the invention discussion. Only about 10 of these utterances were proposals or amendments to proposals for solution methods. Almost half of the comments were clarifications or requests for clarifications about a proposed solution. About 30 were evaluations of suggested approaches. All other categories represented \(10 \%\) or less of the utterances.
Results from this discourse analysis suggest that groups who generated a wider range of solutions, proposed more solutions, made more clarifications of proposed solutions, and made fewer comments about being at impasse. A more interesting observation is that they also engaged in more discussion of task coordination. When explored in the context of leader type, groups with a high math leader made more comments related to task coordination, fewer comments about being at impasse, and devoted less discussion to determination of math expertise. Ongoing analyses are more specifically examining who contributes what to the discussion and when. Analyses suggest that low math students are the ones more likely to make comments about expertise, and also, surprisingly, that they are the ones more likely to contribute evaluations of the proposed solutions. This may account for why evaluative comments do not seem to relate to better performance in this sample. As the discourse analysis deepens and matures, this approach is hoped to generate a better understanding of what features of discussion may be critical for learning from invention, so that these features may be used to engineer the design of effective classroom invention activities.

\section*{Discussion}

The results of this research have shown that groups with members of different backgrounds or expertise may generate a broader range of solution approaches during invention tasks, and that this may benefit understanding of the canonical solution. Group composition in terms of math skill affects when students are able to get the most out of mathematical learning-by-invention activities. Students who worked in mixed groups were better at explaining their understanding of standard deviation on a quiz following the activity than students who worked in more homogeneous groups. Significant effects of group composition were seen in the variety of solution approaches that were considered by groups, particularly higher quality approaches.

Interestingly, it was the mixed groups who generated the widest variety of solution attempts, suggesting that they seem to be in a particularly good position to make the most of invention exercises. This is consistent with several other findings suggesting that diversity in expertise among group members contributes to more adaptive, flexible and creative problem solving (Canham, Wiley, \& Mayer, 2012; Goldenberg \& Wiley, 2011; Wiley \& Jensen, 2005; Wiley \& Jolly, 2003). Additionally, the consideration of a wider variety of solution approaches during the invention phase, including a larger number of higher quality approaches, predicted the uptake of a later lesson about the standard deviation formula and mediated the effects of group composition and diversity on learning.

To further explore the conditions that might enable effective learning from invention, we found that who emerges as the leader of a diverse group matters. Mixed groups with high math leaders discussed more high quality solution attempts as compared to groups with low math leaders. Interestingly, our discourse coding is also suggesting an important role for defining or coordinating the task among group members (c.f. Moreland \(\&\) Levine, 1992). Groups with high math leaders made more comments in relation to task execution which seemed to relate to their productivity. Yet, it is important to recognize that the leaders self-selected in this study and this can introduce many reasons why these particular groups may have been more or less effective. We are currently conducting a follow-up experiment, again in the context of statistics instruction as part of an undergraduate Psychology course, where we will be assigning high math and low math students to be leaders for the small group activity. Experimental assignment is critical for determining whether and how the expertise of the group leader itself may be important for effective learning from invention activities.

The results thus far suggest that generating a wide variety of approaches to solution may be one important factor determining whether invention discussions prepare students for later learning. Yet, in some cases a richer discussion around fewer alternatives may also lead to successful learning-from-invention, especially if the discussion leads to key insights. Alternatively, we have some evidence that a few of the groups seemed to benefit from the visual affordances of the graphical representations they made. It is possible that some specific kinds of solution attempts may be particularly helpful toward preparation for future learning (i.e., more visual ones or more abstract ones; Ainsworth, 2006, Schwartz, 1995).

The continued analysis of the discussion protocols is intended to serve as source of insight on what particular behaviors one may wish to support while students engage in learning-by-invention tasks. Thus far interactions among group members have not been scripted, roles have not been assigned, and students have not been given any specific direction how to engage in the task together. A next step that others have already begun pursuing (Kapur \& Bielaczyc, 2011; Roll, Aleven \& Koedinger, 2009) is to
provide some support to students in order to maximize the benefits of engaging in invention tasks, but not so much support that the benefits of invention over direct instruction are nullified. Indeed, in most of Webb's previous studies showing benefits of peer collaboration on learning in math, the peer interaction was carefully scaffolded which may have allowed for more stable benefits of mixed ability groups to emerge. One goal for the closer analysis of our discussion protocols is to gain an even better understanding of the conditions that facilitate learning by invention, and how we can capitalize on the intriguing possibility that exploration and failure can sometimes reap benefits toward more sophisticated conceptual understanding in mathematics and statistics.

\section*{Acknowledgments}

This work has been supported by a Humboldt Research Fellowship to the first author, supporting her research stays in Germany, and a Studienstiftung des Deutschen Volkes Fellowship to Michael Wiedmann, supporting his research stay in the US. A portion of the work reported here served as part of a thesis project submitted by Michael Wiedmann in partial fulfillment of the requirements for the Diploma in Psychology at the University of Freiburg. The authors thank Matthew Cavers, Kelly Currier, Pat Cushen, Thomas Griffin, Robert Hickson, Sabrin Jaber, Allison Jaeger and Ryan Leach for their assistance with data collection, coding, and their diverse contributions to collaborative discussions.

\section*{References}

Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. Learning and Instruction, 16, 183-198.
Canham, M., Wiley, J., \& Mayer, R. (2012). When diversity in training improves dyadic problem solving. Applied Cognitive Psychology, 26, 421-430.
Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. In R. J. Sternberg \& J. E. Davidson (Eds.), The nature of insight (pp. 365-395). Cambridge, MA: MIT Press.
Franz, T. M., \& Larson, J. R., J. (2002). The impact of experts on information sharing during group discussion. Small Group Research, 33, 383-411.
Goldenberg, O., \& Wiley, J. (2011). Quality, conformity, and conflict: Questioning the assumptions of Osborn's Brainstorming technique. Journal of Problem Solving, 3(2).
Kapur, M. (2012). Productive failure in learning the concept of variance. Instructional Science, 40, 651-672.
Kapur, M., \& Bielaczyc, K. (2011). Designing for productive failure. The Journal of the Learning Sciences, 41, 45-83.
Moreland, R. L., \& Levine, J. M. (1992). Problem identification by groups. In S. Worchel, W. Wood, \& J. A. Simpson, (Eds.), Group processes and productivity. Newbury Park, CA: Sage.

Preacher, K. J., \& Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. Behavior Research Methods, Instruments, and Computers, 36, 717-731.
Preacher, K. J., \& Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. Behavior Research Methods, 40, 879-891.
Roll, I., Aleven, V., \& Koedinger, K. R. (2009). Helping students know 'further' - increasing the flexibility of students' knowledge using symbolic invention tasks. Proceedings of the 31st Annual Conference of the Cognitive Science Society, 1169-1174.
Schwartz, D. L. (1995). The emergence of abstract representations in dyad problem solving. Journal of the Learning Sciences, 1, 321-354.
Schwartz, D. L., \& Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. Cognition and Instruction, 22, 129-184.
Shrout, P. E., \& Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. Psychological Methods, 7, 422-445.
Webb, N. M. (1980). A process-outcome analysis of learning in group and individual settings. Educational Psychologist, 15, 69-83.
Wiedmann, M., Leach, R., Rummel, N. \& Wiley, J. (2012). Does group composition affect learning by invention? Instructional Science, 40, 711-740.
Wiley, J. (1998) Expertise as mental set: The effects of domain knowledge in creative problem solving. Memory \& Cognition, 26, 716-730.
Wiley, J., \& Jensen, M. (2006). When three heads are better than two. Proceedings of the \(28^{\text {th }}\) Annual Conference of the Cognitive Science Society, 2375-2380.
Wiley, J. \& Jolly, C. (2003). When two heads are better than one expert. Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society, 2375-2380.
Wittenbaum, G. M., Hollingshead, A. B., \& Botero, I. C. (2004). From cooperative to motivated information sharing in groups: Moving beyond the hidden profile paradigm. Communication Monographs, 71, 886-310.

\title{
Dual Processes in Mental State Understanding: Is Theorising Synonymous with Intuitive Thinking and is Simulation Synonymous with Reflective Thinking?
}

\author{
Meredith R Wilkinson (mwilkinson@dmu.ac.uk) \\ Division of Psychology, School of Applied Social Sciences, Faculty of Health and Life Sciences, De Montfort University, The Gateway, Leicester, LE1 9BH, UK \\ Linden J. Ball (LBall@uclan.ac.uk) \\ School of Psychology, University of Central Lancashire, Darwin Building, Preston, Lancashire, PR1 2HE, UK
}

\begin{abstract}
In this paper we develop an idea first mooted by Wilkinson, Ball, and Cooper (2010), which is that the dichotomy between theorybased and simulation-based reasoning in the context of mental state understanding is synonymous with the distinction between intuitive and reflective thinking in dual-process accounts of human reasoning (e.g., Evans, 2010). To support this proposal we draw upon a range of concepts and findings deriving from both mainstream reasoning research and from studies of social cognition. We also consider the implications of our proposal for the formulation of an integrative approach to understanding reasoning in all of its many manifestations, whether undertaken for the attainment of socially-oriented goals or for the purposes of learning and discovery.
\end{abstract}

Keywords: Dual Processes, Intuitive thinking, Reflective thinking Simulation Theory, Theory Theory

\section*{Introduction}

The question of how we understand and reason about other people's minds has resulted in considerable debate within psychology and philosophy (e.g., Bach, 2011; Wilkinson, Ball, \& Cooper, 2010). Some researchers propose that such mental state reasoning is achieved through the adoption of tacit and non-tacit "theories" that are typically based around conditional inference rules (e.g., Carruthers, 1996). An example of such a theory might be that if someone fails to achieve something for which they have worked hard then they will feel upset. Other researchers, however, posit that mental state reasoning arises via a process of mental simulation. Such simulation might involve imagining how we would feel in a given situation and assuming that others are sufficiently like us that they will feel the same (e.g., Gordon, 1986). An alternative proposal (Goldman, 2006) is that we take our own beliefs and desires "offline", input the beliefs and desires of the other person, and thence reason as if we had the beliefs and desires of the other person. It is important to note that when we refer to "simulation" in the present paper we adopt a restricted notion that relates solely to the simulation of mental states, whether our own or other people's. We acknowledge that simulation can arise when reasoning in other domains such as design (e.g., see Ball \& Christensen, 2009), but we do not extend our discussion to this issue. Second, we note that the term "simulation" has different meanings to different authors. As explained by Goldman (2006), there is high-level simulation, which refers
to the type of simulation we are discussing in this paper, and also lower-level simulation, which refers to the functioning of mirror neurons when engaged in activities such as imitation. Mirror neurons are neurons that are activated both when we perform an action and when we observe the same action being performed (e.g., Gallese \& Goldman, 1998). Although we acknowledge the evidence for the existence of mirror neurons, we, like others (e.g. Saxe, 2005) are unsure of the explanatory power of this form of simulation.

Recently, theorists have started to move away from polarised views as to whether theorising or simulation is adopted in mental state reasoning and have instead acknowledged that both processes may be at play. This has resulted in a flurry of hybrid approaches appearing in the literature (e.g., Bach, 2011; Mitchell, Currie, \& Ziegler, 2009), which not only propose that both theorising and simulation can occur in mental state reasoning, but which also claim that there are content-based effects that govern the mechanism that is triggered. For example, Mitchell et al. (2009) have argued that we deploy simulation as a "default" process, using theorising in familiar situations. We have recently provided empirical support for a hybrid view in a study that required people to think aloud when reasoning about counterfactual scenarios pertaining to mental states (Wilkinson et al., 2010). Participants adopted both theorising and simulation for these scenarios, with content effects being evident in that more simulation and less theorising arose with scenarios involving "controllable" compared to "uncontrollable" events.

In the present paper we extend an argument first presented by Wilkinson et al. (2010) to the effect that the theorising versus simulation distinction is synonymous with the "intuitive" versus "reflective" distinction as described in contemporary dual-process theories of thinking and reasoning (see Evans, 2010, for an overview). According to the dual-process framework, intuitive thinking is classed as fast, automatic, high capacity, low effort and independent of working memory resources, whereas reflective thinking is classed as slow, controlled, low capacity, high effort and dependent on working memory. We argue here that these characteristics of intuitive and reflective thinking align well with key features of theorising and simulation in contexts associated with mental state understanding. In subsequent sections we support this proposal by drawing on concepts
and findings from contemporary reasoning research and from studies of social cognition. We suggest that conceptualising theorising and simulation in a dual-process manner has the potential to enable researchers to move towards a more compelling account of mental state reasoning that can be subjected to rigorous empirical examination. We conclude the paper by addressing issues that researchers might wish to consider further if they find merit in our proposed dual-process conceptualisation of the processes underpinning mental state understanding.

\section*{Proposed Parallels: Theorising as Intuition; Simulation as Reflection}

Evans (2010) describes multiple distinctions between intuitive and reflective reasoning and we take these distinctions as a foundation for demonstrating how one can arbitrate between theorising and simulation within a dualprocess framework. The first distinction that Evans notes is that intuitive reasoning is fast whereas reflective reasoning is slow. We similarly propose that theorising is a fast process whereas simulation is slower. This in turn is linked to the cognitive effort required for theorising and simulation. Like intuitive reasoning, which Evans proposes involves low cognitive effort, we view theorising as being low effort compared to simulation, which is high effort, much like reflective reasoning. Evans further argues that intuitive reasoning is high capacity and reflective reasoning is low capacity. We contend that the same holds for theorising and simulation, respectively. In addition, for people to engage in theorising they need to have a store of pre-existing theories pertaining to others' mental states, with these theories being drawn upon in an automatic manner when primed by particular contexts. Simulation, however, takes the form of a concurrent and incremental reasoning process (Goldman, 2006), which will require more controlled than automatic processing. Again, this distinction parallels the notion that intuitive reasoning is automatic whereas reflective reasoning is controlled. Evans has made further claims concerning the links between intuitive versus reflective reasoning and working memory. He argues that reflective reasoning is dependent upon working memory resources whereas intuitive reasoning is independent of such resources. We propose that this distinction holds for theorising and simulation too, with simulation being highly dependent upon working memory and executive functioning (e.g., Currie, 1996; Goldman, 2006).

\section*{Empirical Evidence for the Proposed Parallels}

A robust finding in the reasoning literature concerns the phenomenon of "belief bias", which is typically studied in relation to people's abilities at syllogistic inference. Within a standard conclusion-evaluation paradigm participants are presented with two premises that they should assume are true and an associated conclusion. They are then required to determine whether the conclusion follows logically from the premises. Many studies have shown that participants are biased by the conclusion's believability when making
evaluations, rather than reasoning on the basis of the conclusion's validity (e.g., see Stupple, Ball, Evans, \& Kamal-Smith, 2011).

Numerous dual-process accounts have been forwarded as to why belief-bias occurs (see Ball, 2011, for a review). For the purposes of our argument, however, we draw on the "selective processing model" of Evans (e.g., 2000), itself an example of a more general class of dual-process models referred to as "default-interventionist" theories (Evans, 2007). According to the selective processing model of belief bias, intuitive reasoning cues a response that may or may not be overridden by a reflective process. The default, intuitive response is to accept or reject conclusions based solely on their believability. If, however, reflective reasoning is applied then this reasoning is influenced by the conclusion's belief status such that participants will search for confirming models when a conclusion is believable and for disconfirming models when it is unbelievable. Whether a logically correct evaluation ensues for a problem is, therefore, dependent on the interplay between the intuitive and reflective processes, with certain problems (e.g., those with invalid but believable conclusions) being especially difficult because the belief status of the conclusion biases both the default response and the confirmation-oriented reflective response (Stupple et al., 2011).

We propose that in tasks of mental state reasoning people can be similarly biased by their personal beliefs. This is demonstrated by the so-called "curse of knowledge", whereby participants are unable to pass false belief tasks because they cannot inhibit viewing a situation from their own perspective (e.g., Birch \& Bloom, 2007). In such tasks (e.g., Baron-Cohen, Leslie, \& Frith, 1985) participants (typically young children) are introduced to two protagonists in a room, both of whom are aware of a particular state of affairs, such as a marble in a basket. Then one protagonist leaves the room and the remaining protagonist moves the marble to a box. The participant is asked, "Where will the protagonist who left the room look for the marble upon returning?" If the individual is able to reason about another person's beliefs then they should state that the protagonist will look for the marble in the basket. Individuals who fall foul of the curse of knowledge will respond by saying that the protagonist will look in the box (where they themselves know the marble is currently located), demonstrating a form of belief-biased reasoning. We suggest that overcoming this bias, especially when encountering such a situation for the first time, requires the deployment of a controlled process of mental simulation in which the reasoner takes their own beliefs off-line and reasons from the beliefs of the protagonist (e.g., Mitchell et al., 2009). This is equivalent, we propose, to the way that people can engage in reflective reasoning in an effort to overcome belief bias in syllogistic reasoning (Stupple et al., 2011), although, as noted above, even reflective reasoning does not guarantee success since it may itself be biased.

We now return to Evans' (2010) description of the characteristics of intuitive versus reflective reasoning in
order to assess the evidence for the proposed parallels between dual-process views and the theorising/simulation distinction in mental state reasoning. Evans argues that intuitive reasoning is fast whereas reflective reasoning is slow, a view that we propose aligns well with the theorising/simulation distinction. Evidence for our claim comes from Atkinson, Bell, and Feeney (2009), who examined the influence of a speeded-response requirement on how participants reasoned about counterfactual scenarios that were constructed to tap into two robust effects: (1) the "action effect", which is a tendency to regret action more than inaction in the short term (e.g., Kahneman \& Tversky, 1982), with the reverse being the case in the long term (e.g., Gilovich \& Medvec, 1994); and (2) the "temporal order effect" (e.g., Byrne, Segura, Culhane, Tasso, \& Berrocal, 2000), which is the tendency to attribute more negative affect to the person committing the final act in a sequence of actions when a negative outcome occurs. Atkinson et al. (2009) asked participants to reason about the presented scenario, which entailed reading a vignette describing two agents in a negative situation, either in a speeded condition, in which they had to answer as quickly as possible, or in a non-speeded condition, in which they were able to take as long as they wished. Whereas the temporal order effect was unaffected by the speed manipulation, the action effect was disrupted, with the actor selected significantly less in the speeded compared to the non-speeded condition. In relation to the temporal order manipulation we propose that participants uniformly access a "theory" that if a person acts last in a sequence of events leading to a negative outcome then they will feel worse. However, in the case of the action effect, we propose that participants need to run simulations of the mental states of both protagonists to evaluate the interplay between action/inaction and the time that events arose. Such an evaluation process is time-consuming relative to accessing a pre-stored theory, which would explain why the speeded-response requirement only affects the action effect and not the temporal order effect. These findings further demonstrate how a simulation may overturn an initial theory. The fact that the actor was chosen less often in the speeded compared to the non-speeded condition suggests that people held an initial theory that the non-actor would feel more regret, but when afforded the time to run simulations they could overturn this initial response, much as reflective reasoning can overturn an initial belief-biased response in syllogistic inference (Stupple et al., 2011).

The fact that simulation seems to be slower than theorising also speaks to likely discrepancies in the cognitive effort required for these reasoning types. We propose that theorising is low effort, like intuitive reasoning, whereas simulation is high effort, like reflective reasoning. As such, simulation will be dependent on general cognitive resources, including executive functioning and cognitive inhibition (e.g., Currie, 1996; Goldman, 2006) whereas theorising will work independent of such mechanisms. In addition, simulation will be dependent upon working memory whereas theorising will not. Evidence in support of
this claim comes from a study using the "director task" (Lin, Keysar, \& Epley, 2010), where participants are presented with a grid that contains slots that can be seen by both themselves and a director, who is actually a confederate. Some items, however, are only visible to the participant, since they are occluded from the director's view. Participants are instructed what object to move. On critical trials the perspective between director and participant differs so the director may say "move the small mouse" when there are three mice and only the smallest one can be seen by the participant. This requires the participant to engage in simulation by shifting their perspective to that of the director's in order to fulfill the instruction correctly.
Importantly, Lin et al. (2010) found that participants with higher working memory capacities performed better on the director task than those with lower working memory capacities. This provides a link between simulation and working memory that bears strong similarities to the link between reflective reasoning and working memory. For example, De Neys (2006) has shown that individuals with greater working memory resources perform better on belieforiented syllogistic reasoning tasks, and Stanovich, West, and Toplak (2011) have argued that reflective reasoning is dependent upon executive functioning resources. We propose that just as the intuitive/reflective distinction is associated with differential involvement of working memory, so too is the theorising/simulation distinction.
Theorists proposing dual-process accounts of reasoning have also recently begun to draw upon neuroscientific evidence to support their claims. For example, Goel (2003) has presented evidence for "dual pathways" in syllogistic reasoning, with intuitive processes associated with the frontal-temporal pathway and reflective processes associated with the parietal pathway. In a review article, Goel (2007) acknowledges that the question of which neural regions are responsible for particular types of processing is one that has demonstrated differing findings, but that the evidence nevertheless points towards a fractionated system for deductive reasoning rather than a unitary one.
Neuroscientific evidence for theorising and simulation in mental state reasoning is equally complex and has been criticised for failing to provide a clear differentiation between brain regions specialised for such processing (e.g., Apperly, 2008; Wilkinson \& Ball, 2012). Nevertheless, findings are suggestive. For example, Mitchell, Banaji, and Macrae (2005), found that the ventral medial prefrontal cortex was activated when participants made judgements concerning facial expressions. This provides evidence of a partial locus for theory-based reasoning, since an intuitive judgement is all that would be required for this task, with no simulation being necessitated. We believe that our claims for the intuitive/theorising parallels here are strengthened by the observation that this brain region is also known to be activated in syllogistic reasoning tasks when participants provide belief-biased responses (e.g., Goel \& Dolan, 2003).
As for simulation in mental state reasoning, it is admittedly not easy to locate a specific brain region
responsible for such processing. Part of the difficulty originates from the differing conceptualisations of simulation within the literature, with some theorists (e.g., Gallese \& Goldman, 1998) arguing for lower-level simulation (Goldman, 2006). Studies of higher-level simulation in paradigms such as the director task suggest the involvement of a number of brain regions, including the superior dorsal medial prefrontal cortex and the superior/middle temporal sulci extending to the extrastriate body area and the posterior superior temporal sulcus (Dumontheil, Küster, Apperly, \& Blakemore, 2010). Admittedly, the evidence for the localisation of simulation is not clear-cut, but again is suggestive of distinct brain regions being associated with theorising and simulation.


Figure 1: A schematic representation of a hybrid dualprocess model of theorising and simulating.

\section*{Issues Arising from the Proposed Parallelism}

In the previous section we outlined how Evans' (2010) distinction between intuitive and reflective thinking can map onto the concepts of theorising and simulation in mental state reasoning. However, there is still much for theorists to consider if they see benefits in examining such apparent parallels more extensively. In this respect we note that reasoning researchers have now begun to move away from normative accounts of human reasoning towards a descriptivist agenda (e.g., Elqayam \& Evans, 2011). We generally support this approach and suggest that using theorising and simulation within a descriptivist dual-process framework affords an opportunity to develop a rich and innovative programme of empirical and theoretical research. However, several questions need to be borne in mind when pursuing such a project, which we consider below.

A first, critical question is this: exactly how do theorising and simulation function within a dual-process framework? The reasoning literature contains both sequential and parallel dual-process models of phenomena such as belief bias. Sequential models of the default-interventionist variety propose that intuitive reasoning generates a default response
and that reflective reasoning serves either to confirm or override this initial judgement (e.g., Evans, 2006). Parallel models (e.g., Sloman, 2002; Stupple \& Ball, 2008) propose that intuitive and reflective reasoning compete in generating a response. Evans (2009) has also proposed a "hybrid" dualprocess model, which combines sequential and parallel processes. We think it likely that theorising and simulation can operate both sequentially and in parallel such that a hybrid model may capture key subtleties most effectively (see Figure 1). We propose an initial stage of pre-attentive theorising, whereby representational structures such as scripts and schemas are attended to before a decision is made to apply further theorising or simulation. If a theory is insufficient for generating an inferential response then people can switch to first-person or third-person simulation, with the possibility of returning to theorising. A simulation might also override an initial theory-based response. Furthermore, we follow Mitchell et al. (2009) in suggesting that there will be occasions when an appropriate theory is unavailable, such that people will have to engage in simulation to make some kind of inference. Our model therefore operates in a highly content-dependent manner.

Undoubtedly, empirical evidence needs to be provided for the model presented in this paper. In fact, this model grew out of a series of experiments that we conducted, including one reported by Wilkinson et al. (2010), which showed that people simulate more when reasoning about scenarios involving controllable rather than uncontrollable events, with the reverse pattern for theorising. We propose that our model can readily accommodate such evidence for contentdependency in mental-state reasoning. We suggest that more simulation is evoked for controllable events because participants are more readily able to engage in hypothetical thinking and planning in relation to such scenarios, whereas in the case of uncontrollable events participants are likely to engage in the extraction of a theory since there is little more that they are able to do. Wilkinson et al. (2010) also found that participants often switched between theorising and simulation within the same response. This finding aligns well with Figure 1, which can accommodate this process and the inter-dependence of theorising and simulation in that participants may start out theorising and then adopt simulation to develop their answer further. Furthermore, Wilkinson et al. (2010) noted that participants' theory-based responses tended to be much quicker than their simulationbased responses. This is explicable given that the extraction of theories is assumed to arise in a high-capacity but loweffort manner, whereas simulation is assumed to be more involved, requiring a longer and more controlled reasoning process (e.g., Goldman, 2006). We acknowledge, however, that such claims would benefit from corroboration via the deployment of chronometric measures.

A further important issue in relation to theorising and simulation processes within a hybrid model is whether these processes operate independently or whether there are dependencies, with the output of one process determining the likelihood of deploying the other process (see Elqayam,

2009, for a discussion of this issue in the context of dualprocess theories of reasoning). It is too soon to speculate on this matter, although what has been established in our own research using think-aloud techniques is that people are readily inclined to switch between theorising and simulating within the same reasoning task (Wilkinson et al., 2010). This evidence is at least suggestive of a degree of interdependence between the two processes.

We finally turn to the question of what happens in our proposed model when conflict arises between theorising and simulation. For example, imagine a scenario in which a student fails an important assignment and where our inferential goal is to understand what they might be feeling. Using theorising we might infer that the person will be upset, since if someone fails to pass an assessment they are likely to be distraught. However, if we are also presented with the information that the individual spent every evening drinking in the pub during the week prior to the assignment deadline, we may run a simulation of the person's mental state to draw the conclusion that the assignment was not of much importance to them. We propose that when such conflict arises a "type 3" conflict-resolution process comes into play (see Evans, 2009). This process would arise subsequent to processes of theorising and/or simulation, but before the generation of a final response. In this way it would be possible for simulation to override a theory-based decision when conflict occurs between the two processes, much as reflective reasoning can override a belief-oriented response in syllogistic inference tasks (e.g., see Ball, 2011).

\section*{Conclusions}

We have presented arguments for why theory-based reasoning can be viewed as synonymous with intuitive reasoning and simulation-based reasoning can be viewed as synonymous with reflective reasoning within a dual-process framework. This argument was originally advanced by Wilkinson et al. (2010), but we have extended it here so as to provide a more complete and compelling explanation of the parallels between these two hitherto separate conceptual dichotomies. We have additionally considered some of the key questions that need to be addressed by researchers who see value in exploring these suggested parallels further.

Our proposals also resonate with recent calls for greater integration between theorising and simulation accounts of social cognition (Bach, 2011). Bohl and van den Bos argue that the general notion of "theory of mind" is primarily focused on type 2 processing (reflective thinking) rather than type 1 processing (intuitive thinking). We contest this point and instead propose that the traditional distinction between theory-based and simulation-based inferences is best viewed as aligning with the intuitive (type 1) versus reflective (type 2) distinction.

In terms of the development of dual-processes, Evans (2011) has stated that he does not wish to propose that reflective reasoning replaces intuitive reasoning, but rather that the two co-occur in adulthood. This proposal is similar to Mitchell et al.'s (2009) claim that theorising and
simulation operate side-by-side in adulthood. Of course, it is difficult at this point to map out a developmental progression of intuitive and reflective reasoning. Mitchell et al. propose that we start out by simulating and Bach (2011) has suggested that theories may grow out of repeated simulations. These ideas initially seem counterintuitive, since simulation is so dependent upon general cognitive resources (Currie, 1996; Goldman, 2006), which tend not to be well developed in young children. However, the claim is not that children necessarily make correct inferences; indeed false belief studies show that they start out by giving incorrect responses by answering from their own perspective (Mitchell et al., 2009). When it comes to the development of intuitive and reflective reasoning the pattern should, of course, align fully with that for theorising and simulation, with reflective reasoning developing first and intuitive reasoning later, which does seem to capture aspects of the development of expertise. For example, when one first learns to drive a car the process is deliberate and controlled, but after time things become automated. We propose that this represents a shift from deliberate reflection to automatic intuitive reasoning.

We trust that by advancing an account of how theorising and simulation align with intuitive and reflective thinking we have provided inspiration for future empirical work and theoretical development and will enliven future discussion concerning the processes involved in mental state reasoning.

\section*{Acknowledgements}

We thank Rachel Cooper for previous discussions relating to the ideas presented in this paper. We also thank the ESRC for funding the first author's interdisciplinary PhD in Psychology and Philosophy at Lancaster University, UK.

\section*{References}

Apperly, I. A. (2008). Beyond simulation-theory and theory-theory: Why social cognitive neuroscience should use its own concepts to study "theory of mind". Cognition, 107, 266-283.
Atkinson, L., Bell, D., \& Feeney, A. (2009). The relationship between counterfactual thinking and emotional reactions to event outcomes: Does one account fit all? Psychonomic Bulletin \& Review, 16, 724-728.
Bach, T. (2011). Structure-mapping: Directions from simulation to theory. Philosophical Psychology, 24, 2351.

Ball, L. J. (2011). The dynamics of reasoning: Chronometric analysis and dual-process theories. In K. I Manktelow, D. E. Over \& S. Elqayam. (Eds.) The science of reason: A Festschrift for Jonathan St. B. T. Evans (pp. 283-307). Hove, UK: Psychology Press.
Ball, L. J., \& Christensen, B. T. (2009). Analogical reasoning and mental simulation in design: Two strategies linked to uncertainty resolution. Design Studies, 30, 169-186.
Baron-Cohen, S., Leslie, A. M., \& Frith, U. (1985). Does the autistic child have a "theory of mind"? Cognition, 21, 37-46.

Birch, S. A. J., \& Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. Psychological Science, 18, 382-386.
Bohl, V., \& van den Bos, W. (2012). Toward an integrative account of social cognition: Marrying theory of mind and interactionism to study the interplay of type 1 and type 2 processes. Frontiers in Human Neuroscience, 6, 1-15.
Byrne, R. M. J., Segura, S., Culhane, R., Tasso, A., \& Berrocal, P. (2000). The temporality effect in counterfactual thinking about what might have been. Memory \& Cognition, 28, 264-281.
Carruthers, P. (1996). Autism as mind-blindness: An elaboration and partial defence. In P. Carruthers \& P. K. Smith (Eds.) Theories of theories of mind (pp. 257-273). Cambridge: Cambridge University Press.
Currie, G. (1996). Simulation-theory, theory-theory and the evidence from autism. In P. Carruthers \& P. K. Smith (Eds.) Theories of theories of mind (pp. 242-256) Cambridge: Cambridge University Press.
De Neys, W. (2006). Dual processing in reasoning two systems but one reasoner. Psychological Science, 17, 428433.

Dumontheil, I., Küster, O., Apperly, I. A., \& Blakemore, S. (2010). Taking perspective into account in a communicative task. Neuroimage, 52, 1574-1583.
Elqayam, S. (2009). Models of dependence and independence: A two-dimensional architecture of dual processing. Thinking \& Reasoning, 15, 377-387.
Elqayam, S., \& Evans, J. St. B. T. (2011). Subtracting "ought" from "is": Descriptivism versus normativism in the study of human thinking. Behavioral \& Brain Sciences, 34, 233-248.
Evans, J. St. B. T. (2000). Thinking and believing in J. Garcia-Madruga, N. Carriedo, \& M. J. González-Labra (Eds.) Mental models in reasoning (pp. 41-56). Madrid: UNED.
Evans, J. St. B. T. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. Psychonomic Bulletin \& Review, 13, 378-395.
Evans, J. St. B. T. (2007). On the resolution of conflict in dual process theories of reasoning. Thinking \& Reasoning, 13, 321-339.
Evans, J. St. B. T. (2009). How many dual-process theories do we need? One, two, or many? In J. St. B. T. Evans \& K Frankish (Eds.) In two minds: Dual processes and beyond. (pp. 33-54). New York, US: Oxford University Press.
Evans, J. St. B. T. (2010). Thinking twice: Two minds in one brain. Oxford: Oxford University Press.
Evans, J. St. B. T. (2011). Dual-process theories of reasoning: Contemporary issues and developmental applications. Developmental Review, 31, 86-102.
Gallese, V., \& Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. Trends in Cognitive Sciences, 2, 493-501.

Gilovich, T., \& Medvec, V. H. (1994). The temporal pattern to the experience of regret. Journal of Personality \& Social Psychology, 67, 357-365.
Goel, V. (2003). Evidence for dual neural pathways for syllogistic reasoning. Psychologica, 32, 301-309.
Goel, V. (2007). Anatomy of deductive reasoning. Trends in Cognitive Sciences, 11, 435-441.
Goel, V., \& Dolan, R. J. (2003). Explaining modulation of reasoning by belief. Cognition, 87, B11-B22.
Goldman, A. I. (2006). Simulating minds: The philosophy, psychology and neuroscience of mindreading. Oxford: Oxford University Press.
Gordon, R. M. (1986). Folk psychology as simulation. Mind \& Language, 1, 158-171.
Kahneman, D., \& Tversky, A. (1982). The psychology of preferences. Scientific American, 246, 160-173.
Lin, S., Keysar, B., \& Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. Journal of Experimental Social Psychology, 46, 551-556.
Mitchell, J. P., Banaji, M. R., \& Macrae, C. N. (2005). The link between social cognition and self-referential thought in the medial prefrontal cortex. Journal of Cognitive Neuroscience, 17, 1306-1315.
Mitchell, P., Currie, G., \& Ziegler, F. (2009). Two routes to perspective: Simulation and rule-use as approaches to mentalizing. British Journal of Developmental Psychology, 27, 513-543.
Saxe, R. (2005). Against simulation: the argument from error. Trends in Cognitive Sciences, 9, 174-179.
Sloman, S. A. (2002). Two systems of reasoning. In T. Gilovich, D. Griffin \& D. Kahneman (Eds.). Heuristics and biases: The psychology of intuitive judgment (pp. 379-396). New York, US: Cambridge University Press.
Stanovich, K. E., West, R. F., \& Toplak, M. E. (2011). The complexity of developmental predictions from dual process models. Developmental Review, 31, 103-118.
Stupple, E. J. N., \& Ball, L. J. (2008). Belief-logic conflict resolution in syllogistic reasoning: Inspection-time evidence for a parallel process model. Thinking \& Reasoning, 14, 168-181.
Stupple, E. J. N., Ball, L. J., Evans, J. St. B. T., \& KamalSmith, E. (2011). When logic and belief collide: Individual differences in reasoning times support a selective processing model. Journal of Cognitive Psychology, 23, 931-941.
Wilkinson, M. R., \& Ball, L. J. (2012). Why studies of autism spectrum disorders have failed to resolve the theory theory versus simulation theory debate. Review of Philosophy and Psychology, 3, 263-291.
Wilkinson, M. R., Ball, L. J., \& Cooper R. (2010). Arbitrating between Theory-Theory and Simulation Theory: Evidence from a think-aloud study of counterfactual reasoning. In S. Ohlsson \& R. Catrambone (Eds.) Proceedings of the Thirty-Second Annual Conference of the Cognitive Science Society (pp. 10081013). Austin, Texas: Cognitive Science Society.

\title{
Effects of Explaining Anomalies on the Generation and Evaluation of Hypotheses
}

\author{
Joseph Jay Williams (joseph_williams@berkeley.edu), Caren M. Walker (caren.walker@berkeley.edu) Samuel G. Maldonado (samuel.g.maldonado@gmail.com), Tania Lombrozo (lombrozo@berkeley.edu) \\ Department of Psychology, University of California, Berkeley, 3210 Tolman Hall \\ Berkeley, CA 94720 USA
}

\begin{abstract}
We investigate the effects of explaining anomalies (i.e., observations that conflict with current beliefs) on belief revision, and in particular how explaining contributes to the rejection of incorrect hypotheses, the generation of alternative hypotheses, and the selection of a hypothesis that can account for anomalous observations. Participants learned how to rank students across courses using statistical concepts of deviation, and did so while either explaining sample rankings or writing their thoughts during study. We additionally varied whether or not candidate hypotheses about the basis for ranking were presented to participants prior to learning, and the number of sample rankings that violated intuitive misconceptions about ranking. Measures of learning and coded responses suggest that prompting people to explain can increase the rate at which they entertain both correct and incorrect hypotheses, but that explaining promotes the selection of a hypothesis that can account for anomalous observations.
\end{abstract}

Keywords: explanation; self-explanation; learning; generalization; statistics; misconceptions; anomalies.

\section*{Introduction}

A critical element of successful learning is the ability to flexibly revise beliefs in light of new data and experience. For example, a mathematics student might form tentative beliefs about how to solve a novel problem, but subsequently revise these beliefs in the face of anomalous data: observations that conflict with working assumptions and therefore signal a need to revise beliefs (Chinn \& Brewer, 1993; Koslowski, 1996). Here we consider how beliefs are revised in light of anomalous observations, and in particular how explaining such observations influences learning.

Generating explanations has been shown to promote learning across a range of tasks and domains, with evidence from experimental studies of category learning (Williams \& Lombrozo, 2010), "self-explaining" in students (e.g., Fonseca \& Chi, 2011), and conceptual development in children (e.g., Siegler, 2002; Wellman \& Liu, 2007). These benefits are likely to derive from multiple sources, including increased engagement and increased accessibility of effective strategies (Siegler, 2002), better metacognitive monitoring (e.g., Chi et al, 1994), and the generation of inferences to fill gaps in understanding (e.g., Chi et al, 2000), among others.

In the present work we build on the Subsumptive Constraints account of explanation developed in prior research (Williams \& Lombrozo, 2010, 2013). According
to this account, explaining a particular observation drives learners to interpret it as an instance of a broad pattern or generalization, and thereby facilitates learning about regularities that apply broadly (Williams \& Lombrozo, 2010; 2013; Williams, Lombrozo \& Rehder, 2013).

To illustrate, consider the findings reported in Williams and Lombrozo (2010). Participants attempted to learn a new classification system involving two categories that could be differentiated by a rule with no exceptions ( \(100 \%\) rule) or an alternative that accounted for most cases, but with two anomalies ( \(75 \%\) rule). Participants who were prompted to explain were significantly more likely to discover the \(100 \%\) rule than those prompted to describe the category members, think aloud, or engage in free study. These findings confirm the prediction that explaining facilitates learning about broad patterns, and also suggest that explaining could make learners especially sensitive to anomalies, as they signal that current beliefs are either false or limited in scope.

Subsequent research, however, suggests a more complicated relationship between explanation and anomalies. Williams and Lombrozo (2013) found that participants prompted to explain favored patterns consistent with prior knowledge, even when such patterns had exceptions (anomalies) that were better explained by alternative patterns. Williams, Lombrozo, and Rehder (2013) found that participants prompted to explain were more likely to overgeneralize broad patterns, effectively ignoring exceptions, even when this resulted in slower and less accurate learning.

Explaining seems to therefore have opposite effects: by encouraging learners to seek broad patterns, explaining can sometimes lead to greater belief revision in light of anomalies, and at other times to the anomalies being effectively dismissed or "explained away" (see also Chinn \& Brewer, 1993; Khemlani \& Johnson-Laird, 2012; Koslowski, 1996). As a first step towards understanding the conditions under which explanation has each effect, Williams, Walker, and Lombrozo (2012) investigated how changing the number of anomalous observations presented interacted with a prompt to explain. We begin by briefly reviewing the results from this study, and additionally present novel analyses concerning participants' coded explanations. We then present a new experiment aimed at differentiating two potential roles for explanation: one in the rejection of current hypotheses in light of anomalies, and another in the generation and selection of new hypotheses.

\section*{Explaining anomalies: Previous findings}

In previous work, we explored the effects of generating explanations for observations that were anomalous with respect to learners' prior beliefs about statistical measures (Williams, Walker, \& Lombrozo, 2012). Participants learned a university's ranking system by studying how pairs of students from different courses had been ranked given the students' grades and the means and standard deviations of their respective courses. The task required learners to compare student grades using concepts analogous to z -scores, and therefore to reject commonly endorsed but non-normative principles for ranking. These non-normative principles included ranking students based on the higher raw score, the greater number of points above the course mean, or closeness to the maximum course score (Schwartz \& Martin, 2004).

A realistic and experimentally useful feature of this task was that participants could encounter ranked student pairs that were either consistent or anomalous with respect to the non-normative principles for ranking. In many ranked student pairs, the student who is a greater number of standard deviations above the mean will also have a higher raw score, be farther from the mean, or closest to the maximum. We call sample rankings that are consistent with all of the identified ranking principles consistent items, and those that are only consistent with the use of \(z\) scores anomalous items because they are anomalous with respect to many participants' prior beliefs (see Fig. 1).

Williams, Walker, and Lombrozo (2012), henceforth WWL12, presented participants with five examples of ranked pairs of students to learn a university's method for ranking students. Participants' study task was either to explain why the higher ranked student was ranked higher, or to write thoughts they had while studying the pair. Of the five example pairs, there was either a single anomaly (and four consistent pairs) or multiple anomalies (four anomalies, one consistent pair). WWL12 found that belief revision was greatest when participants explained and received multiple anomalies. Explaining did not promote belief revision when only a single anomaly was presented,
(a) Sarah got \(85 \%\) in a Sociology class, where the average score was \(79 \%\), the average deviation was \(3 \%\), the minimum score was \(67 \%\), and the maximum score was 90\%.

Tom got \(69 \%\) in a Art History class, where the average score was \(65 \%\), the average deviation was \(8 \%\), the minimum score was \(42 \%\), and the maximum score was \(87 \%\).

Sarah was ranked more highly by the university than Tom.
and multiple anomalies had no effect on learning unless participants explained.

While these findings suggest that explaining may be especially potent for ensuring that learners process anomalies and use them in updating beliefs, there are several reasons why explaining might have this effect.

Explaining anomalies could be improving accuracy by increasing the rejection of the non-normative principles that were inconsistent with the anomalies, or by increasing the generation and selection of the normative principle. Either of these would account for the observed belief revision, and in fact, the effects could be due to a combination of both.

To evaluate these possibilities, we report here the results of coding the written responses that participants provided in the explain and write thoughts conditions. We coded for whether participants mentioned any of the nonnormative principles and whether they identified standard deviation as playing an important role in rankings.

\section*{Verbal Response Coding}

Each of the five written responses participants provided during the study phase of the experiment was coded according to the following criteria: whether a response mentioned a non-normative principle, whether it mentioned the relative-to-deviation principle (i.e., standardized z-scores, whether or not participants used technical terminology to convey the idea), and whether it contained some other response, such as expressions of surprise or confusion, disagreement with the ranking, or mention of other features of the pairs.

Non-Normative Principles The three non-normative principles were incorrect but designed to correspond to intuitive statistical misconceptions. We term the principles (1) raw-score: the higher ranking went to the student with the higher score, irrespective of mean, average deviation, and minimum or maximum score; (2) relative-to-average: the higher ranking went to the student whose score was the farthest above (or least
(b) Sarah got \(85 \%\) in a Sociology class, where the average score was \(79 \%\), the average deviation was \(8 \%\), the minimum score was \(67 \%\), and the maximum score was \(90 \%\).

Tom got 69\% in a Art History class, where the average score was \(65 \%\), the average deviation was \(3 \%\), the minimum score was \(42 \%\), and the maximum score was \(87 \%\).

Tom was ranked more highly by the university than Sarah.

Figure 1: (a) A consistent ranked example for which all four principles predicted the same ranking. (b) An anomalous ranked example constructed by switching the class average deviations of the consistent example from (a). The switch means that the correct relative-to-deviation ranking is now the opposite of what is predicted by the raw-score, relative-to-average, and relative-to-highest-score principles. Emphasis is added for illustration and was not provided to participants.
below) the class's mean score; (3) relative-to-highestscore: the higher ranking went to the student whose score was the closest to the highest score achieved in the class.

Relative-to-Deviation Principle According to this principle, the better student was the one who scored a greater number of standard (average) deviations above the mean (see Schwartz \& Martin, 2004; Belenky \& NokesMalach, 2012). This was calculated as the difference from the mean divided by the average deviation, and is closely related to a normative measure like the \(z\)-score.

\section*{Response Coding Results}

Principles Cited A task (2: explain, write thoughts) x number of anomalies (2: single, multiple) x principle type (non-normative, relative-to-deviation) mixed ANOVA was conducted on the proportion of responses that mentioned each type of principle (see Fig. 2).

This analysis revealed main effects of task, \(F(1,272)=\) 43.98, \(p<0.001, \eta_{\mathrm{p}}^{2}=0.14\), and number of anomalies, \(F(1,272)=37.15, p<0.001, \eta_{\mathrm{p}}^{2}=0.12\). Overall, explaining increased mention of principles, while multiple anomalies led to decreased mention of principles.

There was also a main effect of principle type, \(F(1\), \(272)=49.90, p<0.001, \eta_{\mathrm{p}}^{2}=0.16\), with non-normative principles mentioned more frequently than the relative-todeviation principle. However, this effect was qualified by an interaction between number of anomalies and principle type, \(F(1,272)=40.52, p<0.001, \eta_{\mathrm{p}}^{2}=0.13\). We therefore conducted separate task x number of anomalies ANOVAs for the two principle types.

Non-normative principles were cited more often by participants prompted to explain, \(F(1,272)=19.03, p<\) \(0.001, \eta_{\mathrm{p}}{ }^{2}=0.07\), and less often by those who encountered multiple anomalies, \(F(1,272)=96.49, p<0.001, \eta_{\mathrm{p}}{ }^{2}=\) 0.26 , with no interaction.

The relative-to-deviation principle was also cited more often in the explain condition, \(F(1,272)=13.14, p<\) \(0.001, \eta_{\mathrm{p}}{ }^{2}=0.05\), with no significant effect of the number of anomalies, \(F(1,272)=1.89, p=0.17, \eta_{\mathrm{p}}{ }^{2}=0.01\).

Number of Different Principles Cited A task x number of anomalies ANOVA was performed on the mean number of different principles cited by each participant (see Fig. 3). Participants prompted to explain mentioned a greater number of different principles, \(F(1,272)=16.20\), \(p<0.001, \eta_{\mathrm{p}}^{2}=0.06\), and multiple anomalies resulted in mention of fewer different principles, \(F(1,272)=31.36, p\) \(<0.001, \eta_{\mathrm{p}}{ }^{2}=0.10\). There was also a task x number of anomalies interaction: explaining robustly increased the number of different principles mentioned in the multiple anomalies condition, \(t(125)=3.97, p<0.001, d=0.70\) ), while the effect in the single anomalies condition was not significant, \(t(147)=1.55, p=0.12, d=0.25\).


Figure 2: Data from WWL12: Mean proportion of responses citing either a non-normative principle (upper panel) or the relative-to-deviation principle (lower panel).


Figure 3: Data from WWL12: Mean number of different principles mentioned by each participant.

\section*{Summary and Discussion}

The results of coding responses from WWL12 suggest that the effects of explanation on learning are not principally a consequence of rejecting principles in light of anomalies, at least not in this kind of task. Explaining increased the rate at which participants mentioned the correct relative-to-deviation principle, but also how often participants mentioned non-normative principles, and how many different principles were cited. Instead, it appears that explanation played an important role in the generation of multiple hypotheses and the selection of the correct hypothesis from among them.

We now present a new experiment that aims to better understand the role of explanation in generating the correct hypothesis as opposed to evaluating and selecting the correct hypothesis from among candidates. In order to do so, we replicate the basic design of WWL12 with an additional manipulation: whether or not participants are
presented with a fixed set of candidate hypotheses, including the relative-to-deviation principle, prior to learning.

\section*{Experiment}

Our experiment manipulated whether participants were asked to explicitly consider potential ranking principles before engaging in learning. Specifically, participants in the exposure condition were presented with descriptions of five candidate principles and rated their plausibility. Participants in a no exposure condition completed the task without this initial presentation of candidate hypotheses, effectively replicating WWL12 (see Bonawitz \& Griffiths, 2010, for a similar manipulation).

As in WWL12, we additionally varied whether participants received instructions to explain or to write thoughts, and whether they encountered a single anomaly or multiple anomalies during study.

If the main role of explanation in WWL12 was to facilitate the generation of candidate hypotheses - and therefore of the relative-to-deviation principle - then the exposure manipulation should mimic effects of explanation in the write thoughts condition, and potentially eliminate differences across study conditions. In contrast, if explaining principally or additionally plays a role in the evaluation and selection of the correct hypothesis (i.e., the relative-to-deviation principle, which accounts for all observations), then we should observe effects of explanation even in the exposure condition.

\section*{Methods}

Participants Seven-hundred-and-twenty-seven members of the Amazon Mechanical Turk community participated in exchange for monetary compensation. Four-hundred-and-eighty additional participants were excluded for failing an instructional manipulation check adapted from Oppenheimer et al. (2009) and designed to evaluate whether participants were reading instructions. The number of excluded participants did not differ as a function of condition, all \(p \mathrm{~s}>0.10\).

Materials \& Procedure The materials and procedure mirrored WWL12, except as noted.
Pre-Test. Participants were presented with ten unranked student pairs and judged how likely the university would be to rank one student above another, on a nine point scale ranging from "Definitely student [X]" to "Definitely student [Y]," with a midpoint of "Equally Likely."

Unlike WWL12, six pre- and post-test items pitted the relative-to-deviation principle against a single one of the non-normative principles, with the other two nonnormative principles predicting that the students were equally ranked. Of the ten pairs, two pitted the relative-todeviation principle against the raw-score principle; two against the relative-to-average; and two against the close-to-highest-score. Four pairs were like the anomalous
study pairs in pitting the relative-to-deviation principle against all three non-normative principles.

Pre-Exposure to Principles. In the exposure condition, after the pre-test and before the study phase, participants were shown an example pair of students and told who was ranked higher. This ranking was consistent, similar to the example in Figure 1a. Participants were then presented with five potential rules the university could use to rank students, and asked to judge, on a scale from 1-7, how likely it was that the university used that particular rule. The rules included all four principles discussed above, as well as an additional average-plus-deviation principle \({ }^{1}\), which favored whichever student was the greater number of percentage points above the average plus average deviation.

Study. Each of the five ranked examples was presented onscreen for exactly 90 seconds in a format similar to Figure 1a and 1b. Participants in the explain condition were prompted to explain why the higher-ranked student was ranked more highly, typing their explanation into a text box onscreen. Participants in the write thoughts control condition were told to type their thoughts during study into an equivalent text box.

Post-Test. The post-test was identical to the pre-test, but all student names and grades were changed, with five points added to each grade to generate novel numbers while preserving the way in which the items pitted the principles against each other.

Additional Measures. Additional questions were asked at the end of the experiment (e.g., demographics, sufficient time for task, strategy) but are not discussed here in the interest of space.

\section*{Results}

Learning Pre-test accuracy did not differ significantly as a function of condition (all \(p \mathrm{~s}>0.2\), mean \(=-.90\) ); we subsequently consider the change in pre- to post-test accuracy as our measure of learning.

A task x number of anomalies x exposure ANOVA on the pre- to post- test change in accuracy found main effects of explanation, \(F(1,719)=15.06, p<0.001, \eta_{p}^{2}=\) 0.02 , and number of anomalies, \(F(1,719)=29.59, p<\) \(0.001, \eta_{\mathrm{p}}^{2}=0.04\), with no significant effect of exposure, \(F(1,719)=1.81, p=0.18, \eta_{\mathrm{p}}^{2}<0.01\), nor interactions (see Fig. 4). Participants prompted to explain showed greater learning than those who were not so prompted (whether they observed one or multiple anomalies), and

\footnotetext{
\({ }^{1}\) We thank Daniel Belenky (personal communication) for suggesting this as an additional principle that participants might find compelling and spontaneously employ.
}
participants who saw multiple anomalies learned more as well (whether or not they explained).

Principles Cited To analyze the relative frequencies with which non-normative and normative principles were cited in each response, we conducted a repeated-measures ANOVA with principle type (non-normative principle, relative-to-deviation principle) as a within-subjects factor and task (2), number of anomalies (2), and exposure (2) as between-subjects factors. This analysis revealed a fourway interaction, \(F(1,717)=8.01, p<0.01, \eta_{\mathrm{p}}^{2}=0.01\). We therefore conducted separate task x exposure x principle ANOVAs for the single anomaly and multiple anomalies conditions.

In the single anomaly condition, this analysis revealed that explaining promoted overall mention of principles, \(F(1,448)=13.44, p<0.001, \eta_{\mathrm{p}}^{2}=0.03\), and that nonnormative principles were mentioned more frequently than the relative-to-deviation principle, \(F(1,448)=67.30\), \(p<0.001, \eta_{\mathrm{p}}{ }^{2}=0.13\). There were no other significant effects - in particular, the effect of exposure was not significant, \(F(1,448)<0.01, p=0.99, \eta_{\mathrm{p}}^{2}<0.01\).

In the multiple anomalies condition, there was a task x exposure x principle interaction, \(F(1,269)=5.28, p<\) \(0.05, \eta_{\mathrm{p}}^{2}=0.02\). Participants who explained and were exposed to the hypotheses beforehand were more likely to mention the relative-to-deviation principle over the nonnormative principles, relative to those who explained without exposure. A task x exposure ANOVA for just non-normative principles revealed a main effect of explaining, \(F(1,269)=5.66, p<0.05, \eta_{\mathrm{p}}^{2}=0.02\), and a significant interaction, \(F(1,269)=7.76, p<0.05, \eta_{\mathrm{p}}{ }^{2}=\) 0.03 . For the relative-to-deviation principle, there was only a main effect of explaining. \(F(1,269)=10.03, p<\) \(0.01, \eta_{\mathrm{p}}^{2}=0.04\).

Number of Different Principles Cited The average number of different principles cited was analyzed with a 2 (task) by 2 (number of anomalies) by 2 (exposure) ANOVA, which revealed more principles in the explain condition than the write thoughts condition, \(F(1,712)=\) 26.73, \(p<0.001, \eta_{\mathrm{p}}^{2}=0.04\), with a marginal effect of exposure, \(F(1,712)=2.81, p=0.09, \eta_{\mathrm{p}}^{2}<0.01\). There was also an interaction between task and exposure, \(F(2\), \(712)=4.51, p<0.05, \eta_{\mathrm{p}}{ }^{2}=0.01\) : explanation's boost in number of principles cited was considerably attenuated when participants were exposed to the principles before study. This finding suggests that participants did attend to the exposure task, even though it did not affect learning.

\section*{Relationship Between Coded Responses and Learning} To investigate the relationship between participants' responses to the explain and write thoughts prompts and their learning as reflected on the post-test, we examined correlations and partial correlations between response types and accuracy. The largest contributor to post-test accuracy was the proportion of responses citing the
relative-deviation-principle, \(r(725)=0.60, p<0.001\), followed by the negative effect of the proportion of responses citing the non-normative principles, \(r(725)=-\) \(0.38, p<0.05\). Even conditioning on pre-test accuracy, task, number of anomalies, and exposure, post-test accuracy was positively correlated with citing the


Figure 4: Change in accuracy from pre- to post-test. Non-normative principles



Figure 5: Mention of non-normative principles and the relative-to-deviation principle (per-response).


Figure 6: Number of different principles mentioned by each participant.
relative-to-deviation principle, \(r(640)=0.45, p<0.001\), and negatively correlated with citing non-normative principles, \(r(640)=-0.26, p<0.001\). These findings suggest that coded responses reflected learning, and are at least consistent with the stronger claim that producing the responses was itself a causal factor in driving learning.

\section*{Discussion}

The current study found that participants who were prompted to explain reliably outperformed those in a write thoughts control condition when it came to learning how a university ranked students, a task that required some understanding of population variance or deviation. Although the current data suggest a trend for a larger effect of explanation in the multiple (vs. single) anomaly condition, the interaction was not significant, as it was in WWL12, where explanation facilitated belief revision significantly more when there were multiple anomalies rather than a single anomaly. With respect to one of the main issues that motivated this research - i.e., specifying the conditions under which explaining leads to greater versus less belief revision - our findings are therefore inconclusive.

Nonetheless, the current work provides novel data from participants' coded responses to the explain and write thoughts prompts, which shed light on the role of explanation in rejecting incorrect hypotheses, generating candidate hypotheses, and selecting the correct hypothesis. If it were the case that explaining anomalous observations made learners more likely to reject hypotheses that failed to account for those observations, then we might have expected that prompting participants to explain would lead them to mention non-normative principles less often than participants in the control condition..Instead, we found that participants prompted to explain were more likely to produce non-normative principles, and also more likely to produce a larger number of different principles. This result - found in WWL12 and replicated again here - suggests that explanation instead played a role in the generation and selection of the correct hypothesis concerning ranking.

Our new experiment helped isolate effects of explanation due to hypothesis generation from those of hypothesis selection. We found that "generating" candidate hypotheses for learners did not mimic effects of explanation; explanation improved learning even when candidate hypotheses were provided in both study tasks. This finding suggests that explaining may be playing an important role in the comprehension or selection of the correct hypothesis (see also Siegler, 2002).

\section*{Acknowledgments}

This research was supported by an NSF CAREER grant awarded to TL. (DRL-1056712).

\section*{References}

Belenky, D. M., \& Nokes-Malach, T. J. (2012). Motivation and transfer: The role of mastery-approach goals in preparation for future learning. Journal of the Learning Sciences, 21(3), 399-432.
Bonawitz, E. B., \& Griffiths, T. L. (2010). Deconfounding hypothesis generation and evaluation in Bayesian models. In Proceedings of the 32nd Annual

Conference of the Cognitive Science Society (pp. 22602265).

Chi, M.T.H., de Leeuw, N., Chiu, M.H., LaVancher, C. (1994). Eliciting self-explanations improves understanding. Cognitive Science, 18, 439-477.
Chinn, C. A., \& Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science education. Review of Educational Research, 63, 1-49.
Fonseca, B. \& Chi, M.T.H. (2011). The self-explanation effect: A constructive learning activity. In Mayer, R. \& Alexander, P. (Eds.), The Handbook of Research on Learning and Instruction (pp. 270-321). New York, USA: Routledge Press.
Khemlani, S. \& Johnson-Laird, P. (2012). Hidden conflicts: Explanations make inconsistencies harder to detect. Acta Psychologica, 139, 486-491.
Koslowski, B. (1996). Theory and evidence: The development of scientific reasoning. The MIT Press.
Legare, C.H. (2012). Exploring explanation: Explaining inconsistent information guides hypothesis-testing behavior in young children. Child Development.
Lipton, Peter (2004). Inference to the Best Explanation. New York, NY: Routledge.
Oppenheimer, D. M., Meyvisb, T., \& Davidenkoc, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. Journal of Experimental Social Psychology, 45(4), 867-872.
Schwartz, D.L., \& Martin, T. (2004). Inventing to Prepare for Future Learning: The Hidden Efficiency of Encouraging Original Student Production in Statistics Instruction. Cognition and Instruction, 22(2), 129-184.
Siegler, R. S. (2002). Microgenetic studies of selfexplanations. In N. Granott \& J. Parziale (Eds.), Microdevelopment: Transition processes in development and learning (pp. 31-58). New York: Cambridge University.
Wellman, H. M., \& Liu, D. (2007). Causal reasoning as informed by the early development of explanations. Causal learning: psychology, philosophy, and computation, 261-279.
Williams, J. J., \& Lombrozo, T. (2010). The role of explanation in discovery and generalization: evidence from category learning. Cognitive Science, 34, 776-806.
Williams, J. J., \& Lombrozo, T. (2013). Explanation and prior knowledge interact to guide learning. Cognitive Psychology, 66, 55-84.
Williams, J. J., Lombrozo, T., \& Rehder, B. (2013). The hazards of explanation: overgeneralization in the face of exceptions. Journal of Experimental Psychology: General. Advance online publication.
Williams, J. J., Walker, C. M., \& Lombrozo, T. (2012). Explaining increases belief revision in the face of (many) anomalies. In N. Miyake, D. Peebles, \& R. P. Cooper (Eds.), Proceedings of the 34th Annual Conference of the Cognitive Science Society (pp. 11491154). Austin, TX: Cognitive Science Society.

\title{
Impulsivity and Overall Similarity Classification
}

\author{
Andy J. Wills (andy.wills@plymouth.ac.uk), Chris A. Longmore \\ Department of Psychology, Plymouth University, UK. \\ Fraser Milton \\ Department of Psychology, Exeter University, UK.
}

\begin{abstract}
It is sometimes argued that implementation of an overall similarity classification is less effortful than implementation of a single-dimension classification. One piece of evidence taken to be in support of this argument is that highly impulsive individuals appear to be more likely to sort on the basis of overall similarity than individuals with low impulsivity (Ward, 1983); presumably, higher impulsivity results in lower effort. In the current article, we identify some limitations in Ward's procedure and, using a more standard measure of impulsivity and a less ambiguous measure of overall similarity classification, re-investigate the relationship between impulsivity and overall similarity classification. Using a match-tostandard procedure, the current experiment finds that overall similarity classification is less prevalent in highly impulsive individuals. The implications of this result, which is opposite to that reported by Ward (1983), are discussed.
\end{abstract}

Keywords: impulsivity; categorization; overall similarity; family resemblance.

In a seminal article, Brooks (1978) argued for two different processes of categorization. In analytic categorization, the participant separates aspects of the stimulus and evaluates their ability to predict category membership. This process of analysis, Brooks assumed, will typically lead to a subset of the stimulus attributes controlling responding. In contrast, nonanalytic categorization is the process of predicting category membership on the basis of overall similarity to known examplesa process that results in all stimulus attributes having some control over responding. Brooks hypothesized that nonanalytic categorization would be more likely to occur where cognitive resources were limited.

Brooks's hypothesis is striking because it assumes that a categorization process that employs all the information in the stimulus (overall similarity) is less effortful than a categorization process that employs a subset of that information (analytic, or "rule-based", categorization). Following Wills, Milton, Longmore, Hester, and Robinson (2013), we describe this as the less-is-more hypothesis-for example, less time spent categorizing objects results in more information from those objects having control over responding (Smith \& Kemler Nelson, 1984). We contrast this with the more-is-more hypothesis-for example, more time spent categorizing objects results in more information from those objects having control over responding (Milton, Longmore, \& Wills, 2008).

The current paper revisits one particular plank in the less-is-more argument; namely, the result reported by

Ward (1983) that highly impulsive individuals are more likely to classify on the basis of overall similarity than individuals with low impulsivity. This result appears to support the less-is-more hypothesis because, presumably, impulsive individuals devote fewer cognitive resources to the categorization task than do reflective individuals.

We had two concerns about Ward's demonstrationthe validity of the measure of impulsivity, and the validity of the measure of overall similarity classification. Below, we outline those concerns, and describe how we addressed them in the current study.

\section*{Impulsivity}

Ward used the Matching Familiar Figures measure of impulsivity (Kagan, 1965), a measure whose validity has been questioned (e.g. Block, Block, \& Harrington, 1974) and which appears to be largely uncorrelated with better validated measures of impulsivity (Helmers, Young, \& Pihl, 1995). In the current study, we employed the Barratt Impulsivity Scale (BIS-11), which is the most widely used measure of impulsivity (Stanford et al., 2009). It has high reliability and good external validity (Patton, Stanford, \& Barratt, 1995). The Barratt Impulsivity Scale is a self-report measure that includes statements such as, "I concentrate easily" and "I am happy-golucky".

\section*{Overall similarity classification}

Ward employed the triad procedure as a measure of the prevalence of overall similarity classification. In this procedure three stimuli, whose relationship to each other is illustrated in Figure 1, are presented simultaneously and participants are asked to decide which two stimuli go together best. Stimuli B and C are similar on both stimulus dimensions, but not identical on either, while stimuli A and B are identical on one stimulus dimension but quite dissimilar on the other. Three responses are possible - an AB response ( A and B go together best), a BC response or an AC response. Time pressure, concurrent load, impulsivity, and instructions to respond impressionistically, all increase BC responses and decrease AB responses (Smith \& Kemler Nelson, 1984; Ward, 1983; Ward, Foley, \& Cole, 1986). AB responding is typically described as "dimensional" responding and BC responding is typically described as "overall similarity" responding, hence leading to the claim that overall similarity (BC) responding increases as cognitive resources


Figure 1: Abstract structure of the triad task. Typically, on half of trials A and B are identical on dimension 1 (Panel b), and on the other half of trials A and B are identical on dimension 2 (Panel a).

\section*{decrease.}

One reason that the triad procedure is ill suited to testing a less-is-more hypothesis is that consistent \(A B\) ("dimensional") responding requires that the participant consider both stimulus dimensions on every trial. This is because the dimension on which A and B are identical varies unpredictably from trial to trial (see Figure 1), and so a consistent AB responder cannot decide in advance of stimulus onset to only attend to one of the stimulus dimensions. Consistent BC responding also requires consideration of both stimulus dimensions on every trial (irrespective of whether one believes that consideration to take the form of an analytic strategy or direct access to similarity relations through holistic "blobs"). Hence both consistent "overall similarity" (BC) responding and consistent "dimensional" (AB) responding requires consideration of all the relevant stimulus information on every trial. It is therefore not the case that overall similarity responding requires consideration of more of the available stimulus information than dimensional responding in the triad task, and hence the triad task is not well suited to testing a less-is-more hypothesis.
In the current study, we employed the match-tostandards procedure, which is perhaps best considered as an interpretatively less ambiguous version of the triad procedure. The procedure was introduced by Regehr and Brooks (1995) as a means of increasing the prevalence of overall similarity classification of novel stimuli, relative to the more commonly employed array sort procedure, in which single-dimension classification dominates (Medin, Wattenmaker, \& Hampson, 1987). In line with the findings of Milton et al. (2008), Regehr and Brooks observed that reaction times were longer for overall similarity classification than single-dimension classification in this procedure, although they did not publish these observations (Brooks, personal communication, 20 October 2009).

In the match-to-standards procedure, participants sequentially free classify each of a series of target stimuli as belonging to one of two categories. The two cate-
gories are represented by two standards-that is, two stimuli that appear on each trial. The two standards differ from each other on all variable stimulus dimensions. For example, in the current study, the two standards are as shown in Figure 2. In the current experiments, there are 10 distinct to-be-classified stimuli, with the abstract structure shown in Table 1. In some respects, the match-to-standards procedure is similar to the triad procedure, because each trial involves deciding which two of the three stimuli go together best (although, unlike the triad task, the option of saying that the two least similar stimuli-the two standards in the match-to-standards procedure - go together is not available). Also, in the match-to-standards procedure, each participant's classification strategy for a particular block is determined over 10 trials (rather than independently for each trial, as in the triad task).

\section*{Experiment}

In summary, the current experiment re-investigates the relationship between impulsivity and overall similarity classification, first reported by Ward (1983), but using improved measures of both impulsivity and of overall similarity classification.

\section*{Method}

Participants and apparatus Thirty-six participants from the University of Exeter took part in the experiment in return for course credit or payment. The stimuli were presented on a 17 -inch CRT monitor, set to a resolution of \(800 \times 600\) pixels and a color depth of 16 -bits per pixel. The participants sat in front of the computer screen at a distance of approximately 50 cms . Responses were made using a standard keyboard.

Stimuli The abstract stimulus structure can be seen in Table 1. The stimulus set consisted of four binaryvalued dimensions (D1-D4) and the stimuli were organized around two prototypes, each representative of the two categories. These prototypes were constructed by
taking all the positive values on the dimensions for one of the stimuli \((1,1,1,1)\) and all of the zero values on the dimensions \((0,0,0,0)\) for the other category. The rest of the stimuli (one-aways) were mild distortions of the two prototypes in that they had three features characteristic of their category and one atypical feature more characteristic of the other category. In total there were 10 stimuli in the set. Sorting the stimuli by overall similarity, as shown in Table 1, maximizes within-group similarities and minimizes between-group similarities. The stimuli were one of the lamp stimulus sets used by Milton and Wills (2004). Each lamp had four variable features; lampshade (with either 5 or 10 dots), width of stand (wide or narrow), color of bar (light or dark blue) and size of base (long or short). See Figure 2 for the prototypes of each category.

Table 1: Abstract stimulus set
\begin{tabular}{cccccccc}
\hline \multicolumn{4}{c}{ Category A } & \multicolumn{4}{c}{ Category B } \\
\hline D1 & D2 & D3 & D4 & D1 & D2 & D3 & D4 \\
\hline 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\
1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\
1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\
0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\
\hline
\end{tabular}

Procedure We used the match-to-standards task introduced by Regehr and Brooks (1995), and developed by our lab (Milton \& Wills, 2004; Milton et al., 2008; Wills et al., 2013). Each trial began with a blank screen that was presented for 1000 ms , followed by three lamps in a triangular arrangement. There were two lamps at the top of the screen that depicted the prototypes for category A and B with the to-be-classified lamp presented below the prototypes. Each stimulus array remained on the screen until the participant placed the to-be-classified lamp into either category A (by pressing the ' \(c\) ' key on the keyboard) or category B (by pressing the ' \(m\) ' key). After the participant made a response the next trial began.

There was no feedback; participants were simply instructed at the beginning of the experiment to sort the stimuli in the way they thought most appropriate. At the end of each block, participants were asked to write down the sorting strategy they used before moving onto the next block. Participants were presented with a total of 60 trials, in 6 blocks of 10 trials. In each block, each of the stimuli shown in Table 1 was presented once as the to-be-classified stimulus. The order of presentation within a block was random.

Immediately after the 6 blocks of classifiation, participants' impulsivity was assessed using the Barratt Impulsivity Scale BIS-11 (Stanford \& Barratt, 1995).


Figure 2: Stimulus prototypes

Classification measure Each participant was classified as having produced one of the sort types described below in each of the six blocks of the experiment. These sort types are identical to those used previously by our lab (Milton \& Wills, 2004; Milton et al., 2008; Milton, Wills, \& Hodgson, 2009; Wills et al., 2013).

A uni-dimensional sort is based on a single dimension of the stimulus. It does not matter which dimension is used as the basis of sorting, so long as all of the positive values for the chosen dimension were in one category and all of the zero values for that dimension were in the other category. Additionally, in order to receive this classification, the participant has to describe their sort as being based on a single dimension.

Participants were considered to have produced a oneaway uni-dimensional sort if they described their sorting as being driven by a single dimension but there was a solitary error in their classification. This means that nine of the items were classified on the basis of a single dimension but the other item was placed into the wrong category.
An overall similarity sort, also commonly known as a "family resemblance" sort (Medin et al., 1987), has a structure identical to that shown in Table 1. In order to receive this classification, the participant had to place each of the prototypes, along with their derived one-aways, into separate categories without error. Additionally, they have to describe their strategy as being based either on general similarity or on placing each item into the category with which it had more features in common.

A one-away overall similarity sort is similar to the one-away uni-dimensional sort with the exception that the error occurred in a sort that was otherwise overall similarity.

Any classifications produced by a participant other than those described above were classified as other sorts, even if the description given by the participant fitted one of the sorts described above. The correspondence between the classification produced by a participant and their verbal description of the sort they have produced is very high in this procedure, approximately 0.99 . The verbal descriptions were classified by the authors.

Impulsivity measure Participants were classified as high impulsivity if their score on BIS-11 was greater than
the sample median (64.5), and as low impulsivity otherwise (cf. Martin \& Potts, 2009).

\section*{Results}

For every block, each participant's sorting strategy was classified according to the sort types described above. As in previous studies, one-away uni-dimensional and one-away overall similarity sorts were classified as unidimensional and overall similarity sorts respectively (Milton \& Wills, 2004; Milton et al., 2008; Wills et al., 2013). The total number of sorts for each strategy was calculated and the mean proportions of overall similarity, uni-dimensional and other sorts produced by high impulsivity and low impulsivity participants are shown in Figure 3.

Participants with high impulsivity produced significantly more uni-dimensional sorts than those with low impulsivity, \(t(34)=2.203, p<.05\). Conversely, participants with high impulsivity produced significantly fewer overall similarity sorts than those with low impulsivity, \(t(34)=-2.382, p<.025\). There was no difference in the prevalence of Other sorts, \(t(34)=.206, p=.838\). Correlations of the raw impulsivity scores with the prevalence of unidimensional, overall similarity and other sorts reveal the same ordinal pattern, albeit with slightly higher p-values (unidimensional \(\tau=.22, p=.09\); overall similarity \(\tau=-.20, p=.13\); other, \(\tau=-.06, p=.68\) ).

\section*{Discussion}

Ward (1983) reported that participants with high impulsivity were more likely to classify on the basis of overall similarity than participants with low impulsivity. This is one of a number of results taken to support the idea that overall similarity classification is a low-effort, "fallback" mode of classification that people employ when cognitive resources are limited. However, a close examination of Ward's study reveals that both the measure of impulsivity, and the measure of overall similarity, employed are sub-optimal. The measure of impulsivity (Kagan's Matching Familiar Figures task) is of questionable validity, and does not correlate with other more valid measures of impulsivity. The triad task, employed by Ward as a measure of overall similarity responding, is also interpretatively ambiguous because both consistent overall similarity responding, and consistent dimensional responding, require consideration of both stimulus dimensions on every trial.

In the current study, we employed a more standard measure of impulsivity (the Barratt Impulsivity Scale), and measured overall similarity responding with the match-to-standards task. The match-to-standards task is a variant of the triad procedure that overcomes the interpretive ambiguities in the standard procedure. One of the ways it achieves this is by considering the participants' responses to a series of ten stimuli, rather than
considering the response to each stimulus as an independent data point.

Our refinement of Ward's procedures seems to have led to a reversal of his conclusions. In the current study, high impulsivity is associated with single-dimension responding, whilst low impulsivity is associated with overall similarity responding. Thus, our data seem to support a conclusion opposite to Ward's-overall similarity classification is more requiring of cognitive resources than dimensional responding. Such a conclusion is consistent with previous results employing the match-to-standards procedure. For example, Milton et al. (2008) found that time pressure generally reduces the prevalence of overall similarity responding, Milton et al. (2009) found greater frontal lobe involvement for overall similarity classification than single-dimension classification, and Wills et al. (2013) found that concurrent load, and a small working memory capacity, reduces the prevalence of overall similarity classification, and that instructions to respond meticulously increased overall similarity responding. Milton and Wills (2009) found that overall similarity classification takes longer, and involves more, widely distributed, eye movements than single-dimension classification. Taken together with the results of the current study, a consistent picture is emerging-overall similarity classification is more effortful than single-dimension classification.

One key question, not satisfactorily answered by this study, or by any other published study, is whether the consistent pattern of results emerging from the match-tostandards procedure is specific to that procedure. Perhaps the match-to-standards procedure is the exception, with other procedures pointing consistently to the opposite conclusion? In our view, there is currently insufficient data to answer this question adequately. The triad procedure, at least as typically analysed, is interpretatively ambiguous, but this problem could be overcome with larger samples and more sophisticated model-based analyses (e.g. Thompson, 1994).

Two other procedures that are sometimes taken to support the less-is-more view (that overall similarity classification is lower effort than single-dimension classification) are the Ashby-Maddox procedure (Ashby \& Maddox, 2005) and the criterial-attribute procedure (Kemler Nelson, 1984; Smith \& Shapiro, 1989). As discussed by Wills et al. (2013), these procedures seem likely to be addressing slightly different questions to the one posed here.

The Ashby-Maddox procedure seems, predominately, to be an investigation of the effects of having an easy-toverbalize category structure versus a hard-to-verbalize structure. As single-dimension structures are typically easy to verbalize, and some multi-dimensional structures are not, the two issues are not unrelated. However, recent work by Ashby, Maddox and colleagues sug-


Figure 3: Proportion of other, uni-dimensional and overall similarity sorts, by impulsivity. Error bars represent one standard error
gests that it is verbalizability, rather than dimensionality, that underlies their reported effects, because the effects are still observed when one compares two multidimensional classification problems that differ in verbalizability (Filoteo, Lauritzen, \& Maddox, 2010; Maddox, Pacheco, Reeves, Zhu, \& Schnyer, 2010). There are also an increasing number of studies that suggest that some of the results from the Ashby-Maddox procedure are a consequence of subtle problems with the design or analysis of these studies (Newell, Dunn, \& Kalish, 2010; Newell, Moore, Wills, \& Milton, 2013).

The criterial-attribute procedure, like the AshbyMaddox procedure, seems to be addressing a slightly different question to the one posed here. Specifically, the criterial-attribute procedure may provide evidence that it is effortful to detect the one dimension that permits above-criterion performance in a context where all dimensions individually support at least \(75 \%\) accuracy. Studies using the criterial-attribute procedure (Kemler Nelson, 1984; Smith \& Shapiro, 1989) support the idea that this is effortful, but they are also consistent with the idea that implementing an overall similarity classification is more effortful than implementing a single-dimension classification (Smith, Tracy, \& Murray, 1993). It was the implementation of overall similarity classification that was the topic of the current study.

Why did the current study and Ward's study produce apparently opposite results regarding impulsivity and overall similarity classification? There are a number of possibilities. One possibility, as previously discussed, is that Ward's measure of impulsivity has low validity. If Ward did not measure impulsivity adequately, then contrasting the current result (using a more valid measure) with Ward's findings is largely irrelevant, as Ward's results would not validly concern impulsivity. Another
possibility is that a less ambiguous analysis of the triad task using model-based methods might reveal that high impulsivity was in fact associated with uni-dimensional classification, rather than overall similarity classification, even in the triad task. We are currently investigating this possibility.

One way in which the match-to-standards task differs from other procedures is that the participants' written descriptions of their sorts are combined with the sorts they actually produce in order to classify their behavior. It seems likely that impulsivity affects the content of those written reports. A more critical possibility is that impulsivity might affect the written reports differently to the objective sorts, and hence the results of the current study might have been different if we had only looked at the objective sorts (or only looked at the written reports). However, if impulsivity does affect written reports differently to the objective sorts, then the consequence should be a difference in the prevalence of Other sorts as a function of impulsivity (because Other sorts occur under our classification procedure when the written report and objective sort do not agree). As can be seen in Figure 3, the proportion of Other sorts is low and does not vary by impulsivity. It therefore seems unlikely that our results would have been substantively different if we had considered just the objective sorts or just the written descriptions.

In conclusion, in the current study highly impulsive individuals were more likely to produce singledimension classifications than low-impulsivity individuals (who were more likely to produce overall similarity classifications). The opposite conclusion, suggested by Ward (1983), seems likely to be due to limitations of the procedures employed in that study, although it remains a possibility that both results are valid, but specific to
the procedure employed.

\section*{Acknowledgments}

This research was supported by ESRC grant RES-000-22-1779 awarded to Andy J. Wills and ESRC grant PTA-026-27-1256 and a Great Western Research Fellowship awarded to Fraser Milton.

\section*{References}

Ashby, F. G., \& Maddox, W. T. (2005). Human category learning. Annual Review of Psychology, 105, 442-481.
Block, J., Block, J., \& Harrington, D. (1974). Some misgivings about the matching familiar figures test as a measure of reflection-impulsivity. Developmental Psychology, 10, 611-632.
Brooks, L. R. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch \& B. Lloyd (Eds.), Cognition and categorization (pp. 169-211). Hillsdale, NJ: Lawrence Erlbaum Associates.
Filoteo, J., Lauritzen, S., \& Maddox, W. T. (2010). Removing the frontal lobes: The effects of engaging executive functions on perceptual category learning. Psychological Science, 21, 415-423.
Helmers, K., Young, S., \& Pihl, R. (1995). Assessment of measures of impulsivity in healthy male volunteers. Personality and Individual Differences, 19, 927-935.
Kagan, J. (1965). Impulsive and reflective children: Significance of conceptual tempo. In J. Krumboltz (Ed.), Learning and the educational process. Chicago: Rand McNally.
Kemler Nelson, D. (1984). The effect of intention on what concepts are acquired. Journal of Verbal Learning and Verbal Behavior, 23, 734-759.
Maddox, W. T., Pacheco, J., Reeves, M., Zhu, B., \& Schnyer, D. M. (2010). Rule-based and informationintegration category learning in normal aging. Neuropsychologia, 48, 2998-3008.
Martin, R., \& Potts, G. (2009). Impulsivity in decision making: An event-related potential investigation. Personality and Individual Differences, 46, 303-308.
Medin, D. L., Wattenmaker, W. D., \& Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. Cognitive Psychology, 19, 242-279.
Milton, F., Longmore, C. A., \& Wills, A. J. (2008). Processes of overall similarity sorting in free classification. Journal of Experimental Psychology: Human Perception and Performance, 30, 407-415.
Milton, F., \& Wills, A. J. (2004). The influence of stimulus properties on category construction. Journal of Experimental Psychology: Learning, Memory and Cognition, 30, 407-415.
Milton, F., \& Wills, A. J. (2009). Eye movements in overall similarity and single-dimension sorting. In Proceedings of the thirty-first annual conference of the cog-
nitive science society (pp. 1512-1517). Austin, TX: Cognitive Science Society.
Milton, F., Wills, A. J., \& Hodgson, T. L. (2009). The neural basis of overall similarity and single-dimension sorting. NeuroImage, 46, 319-326.
Newell, B. R., Dunn, J. C., \& Kalish, M. (2010). The dimensionality of perceptual category learning: a statetrace analysis. Memory and Cognition, 38, 563-581.
Newell, B. R., Moore, C. P., Wills, A. J., \& Milton, F. (2013). Reinstating the frontal lobes? Having more time to think improves "implicit" perceptual categorization. a comment on Filoteo, Lauritzen and Maddox, 2010. Psychological Science, 24, 386-3389.
Patton, J., Stanford, M. S., \& Barratt, E. S. (1995). Factor structure of the Barratt Impulsiveness Scale. Journal of Clinical Neuroscience, 51, 768-774.
Regehr, G., \& Brooks, L. R. (1995). Category organization in free classification: The organizing effect of an array of stimuli. Journal of Experimental Psychology: General, 122, 92-114.
Smith, J. D., \& Kemler Nelson, D. (1984). Overall similarity in adults' classification: The child in all of us. Journal of Experimental Psychology: General, 113, 137-159.
Smith, J. D., \& Shapiro, J. (1989). The occurrence of holistic categorization. Journal of Memory and Language, 28, 386-399.
Smith, J. D., Tracy, J. I., \& Murray, M. J. (1993). Depression and category learning. Journal of Experimental Psychology: General, 122, 331-346.
Stanford, M. S., \& Barratt, E. S. (1995). Factor structure of the Barratt Impulsivity Scale. Journal of Clinical Neuroscience, 51, 768-774.
Stanford, M. S., Mathias, C. W., Dougherty, D., Lake, S., Anderson, N., \& Patton, J. (2009). Fifty years of the Barratt Impulsiveness Scale: An update and review. Personality and Individual Differences, 47, 385-395.
Thompson, L. (1994). Dimensional strategies dominate perceptual classification. Child Development, 65, 1627-1645.
Ward, T. B. (1983). Response tempo and separableintegral responding: Evidence for an integral-toseparable processing sequence in visual perception. Journal of Experimental Psychology: Human Perception and Performance, 9, 103-112.
Ward, T. B., Foley, C., \& Cole, J. (1986). Classifying multidimensional stimuli: Stimulus, task, and observer factors. Journal of Experimental Psychology: Human Perception and Performance, 12, 211-225.
Wills, A. J., Milton, F., Longmore, C. A., Hester, S., \& Robinson, J. (2013). Is overall similarity classification less effortful than single-dimension classification? Quarterly Journal of Experimental Psychology, 66, 299-318.

\title{
More is up... and right: Random number generation along two axes
}

\author{
Bodo Winter (bodo@bodowinter.com) \\ Teenie Matlock (tmatlock@ucmerced.edu) \\ Cognitive and Information Sciences, University of California, Merced \\ Merced, CA 95403 USA
}

\begin{abstract}
Research on the mental representation of numbers has focused on a horizontally aligned mental number line, but more and more findings have begun to implicate a vertical orientation as well. We investigate the relationship between these two orientations when people generate random numbers. In the horizontal condition, people generated larger numbers when they looked right as opposed to left. In the vertical condition, people generated larger numbers when they looked up as opposed to down. We present two main results based on analyses that compare the two spatial orientations. First, we show that the vertical effect was stronger than the horizontal one. Second, we show a weak correlation between the vertical and the horizontal effect, potentially suggesting a shared underlying mechanism.
\end{abstract}

Keywords: mental number line; SNARC; numerical representation; mathematical cognition; metaphor

\section*{Introduction}

People use numbers for practically everything: counting coins, ordering dinner, making an appointment, filling out tax forms, and more. But how do they represent numbers in the head? Research on numerical representation has focused on the link between numbers and spatial cognition. Neuroimaging studies consistently find that the intraparietal sulcus is implicated in numerical as well as spatial tasks (Hubbard, Piazza, Pinel, \& Dehaene, 2005). And across several neuropathological disorders, deficits in spatial cognition are correlated with deficits in numerical cognition (Zorzi, Priftis \& Umiltà, 2002; Rotzer, Loenneker, Kucian, Martin, Klaver \& von Aster, 2009).

One proposal as to how we represent numbers is the idea of a horizontally oriented mental number line, where smaller numbers are associated with left perceptual space and larger numbers with right perceptual space (at least in Western cultures). Evidence for such a representation comes from the Spatial-Numerical Association of Response Code effect (SNARC), which revealed that people respond faster to relatively larger numbers with their right hand, and faster to relatively smaller numbers with their left hand (Dehaene, Bossini \& Giraux, 1993). This effect has been replicated in over 100 experiments (Wood, Nuerk, Willmes \&, Fischer, 2008), and similar effects have been found with pointing (Fischer, 2003), body movements (Hartmann, Grabherr, \& Last, 2011), handwriting (Perrone, de Hevia, Bricolo, \& Girelli, 2010) and many other methodologies.

While there is much converging evidence for horizontally oriented numerical representations, more and more findings
are emerging that also support the presence of a vertical mental number line. For example, when people are moved upwards by a lifting chair while they generate a "random" sequence of numbers, generated numbers are "higher" than when the chair is moving downwards (Hartmann, Grabherr \& Last, 2011). Similarly, an upwards directed eye movement predicts that the next number in a randomly generated sequence will be "higher" than the preceding number (Loetscher, Bockisch, Nicholls, \& Brugger, 2010). Relatively larger numbers also facilitate upwards directed saccades (Schwarz \& Keus, 2004) and upwards directed spatial attention (Pecher \& Boot, 2011).

Evidence for a vertical representation of number also comes from language processing: When people read sentences that contain the word "more", people are faster to respond with an upwards oriented response button as opposed to a downwards oriented response button (Sell \& Kaschak, 2012). The opposite is true for sentences that contain the word "less". Finally, the classic SNARC paradigm, too, works with vertically oriented response buttons (Ito \& Hatta, 2004; Müller \& Schwarz, 2007; Shaki \& Fischer, 2012), where larger numbers facilitate responses to a high button and smaller numbers to a low button.

The potential existence of two orientations along which numbers are represented naturally leads to the question: What is the relation between the horizontal and vertical mental number line? From the get-go, research on the horizontal number line emphasized the cultural nature of spatial numerical associations, where the orientation of the horizontal axis is thought to stem from a culture's writing direction (Dehaene et al., 1993; Göbel, Shaki, \& Fischer, 2011; for a related perspective focusing on cultural aspects, see Núñez, 2011). As reading and writing are very entrenched behaviors, one could imagine the horizontal mapping to be stronger than the vertical one.

The vertical SNARC effect, on the other hand, has been suggested to come from embodied interactions with the world. Cognitive linguists working on Conceptual Metaphor Theory (e.g., Lakoff, 1987) argue that we build up a mental connection between verticality and quantity because we repeatedly experience a correlation between these two domains in our environment (e.g., when we pour water into a glass, as quantity increases, verticality increases as well). Given that the vertical mapping is also connected to entrenched patterns of language use ("this is a high number", "rents are rising"), one could imagine vertical SNARC effects to be stronger than horizontal ones.

Holmes and Lourenco \((2011,2012)\) explicitly compared the two orientations by pitting them against each other:

When participants had to respond to a top/left and to a bottom/right button, people were quicker to respond to the left button with smaller numbers and to the right button with larger numbers. As this mapping goes against the vertical mental number line but produces a regular horizontal SNARC effect nonetheless, Holmes and Lourenco conclude that the horizontal orientation "trumps" the vertical.

We follow up on the work by Holmes and Lourenco (2011, 2012) by providing another comparison between horizontal and vertical mappings with a different task, namely, a random number generation task. We pursue two main questions: First, we compare the relative strength of the horizontal and the vertical effect. Second, we look to see whether the horizontal and the vertical effect are related to each other across individuals. Thus, rather than pitting the two orientations against each other, we take an individual differences perspective, comparing an individual participant's propensity to align numbers on the vertical axis to her propensity to align numbers horizontally. This approach is inspired by work suggesting considerable individual differences in how numbers are mapped onto space (e.g., Fischer \& Campens, 2008; Fischer, 2008; Beecham, Reeve, \& Wilson, 2009). Moreover, studies on individual difference have been used for a range of different phenomena to investigate the question whether different tasks potentially share the same underlying mechanism (e.g., see Stanovich \& West, 2000). Thus, if the vertical and the horizontal effect are related across individuals, they can be seen as tapping into the same system.

\section*{Experiment}

The task was a random number generation task used by Loetscher et al. (2008) and Hartmann et al. (2011) designed to study spatial numerical associations. Participants called out numbers during rhythmic head movements. In one block, head movements were along the horizontal axis, in another, along the vertical. In line with the horizontal SNARC effect and the previous findings of Loetscher et al. (2008), we expected numbers to be larger when people look towards the right. In line with the vertical SNARC effect, we expected numbers to be larger when people were looking upwards.

\section*{Procedure}

Participants were asked to call out numbers between 1-30 to a beat of 0.5 Hz , played by an electronic metronome (following the procedure of Loetscher et al., 2008). There were three blocks: A horizontal block, vertical block, and straight-ahead block. The order of horizontal vs. vertical block was counter-balanced across participants. The straight block was always last. We asked participants to generate 40 numbers in the straight block and 80 numbers in the vertical and horizontal one. Half the participants started left in the horizontal block and down in the vertical block, and the other half started with the right and up positions.

We built up the procedure in pieces: We first instructed participants to perform the rhythmic head movements to the beat, "as large as possible while still being comfortable". Then, we introduced the random number generation component, participants were told to be "as random as possible" and to avoid counting sequences. We reminded participants that randomness in this context means that each number has equal likelihood, and that each number is independent from the preceding one (see Towse \& Cheshire, 2007). They were also asked not to call out the number while performing the movement but when the head was stationary in the corner positions of each axis. To avoid bias, the experimenter never mentioned numbers or spatial language to describe numbers ("high number", "large number"). Following Loetscher et al. (2008), we asked participants to close their eyes while performing the task.

\section*{Participants}

Sixty-five UC Merced undergraduates (all native speakers of English) participated in the experiment for extra credit in a social sciences course. A total of 6 participants ( \(9 \%\) of the total data) were excluded from the analyses because they were unable to finish the task (frequent self-interruptions, incapability of following the beat even after sustained practice).

\section*{Analysis}

Loetscher et al. (2008) binned numbers into large ( \(>15\) ) and small \((<15)\) numbers, but we took a more direct approach, analyzing all generated numbers as a continuous measure. We performed two separate analyses, one on absolute numbers (whether the average was larger for one position over the other), and another on relative numbers (whether the average difference to the preceding number in the sequence was smaller or larger).

We analyzed the data with mixed models using R ( R Core Team, 2012) and the package lme4 (Bates, Maechler \& Bolker, 2012). Our analysis controlled for the by-participant variability in the response (e.g., some participants might generate overall larger numbers than others), as well as for differential responses to the head turning manipulation \({ }^{1}\).

Because Holmes and Lourenco (2012) had found that it matters whether people are exposed to a vertical or a horizontal block first (ibid. 1049, footnote 4), it was necessary to control for the effect of Block Order. It was necessary to control for the effect of Starting Orientation (left/down vs. right/up) because people tend to have a counting or "runs" strategy (see Towse \& Cheshire, 2007). Such a strategy could create spurious spatial mappings if Starting Orientation were not controlled for. For example, a participant who tended to count upwards and start at the down position might generate numbers that are, on average, higher in the up position than those in the down position.

\footnotetext{
\({ }^{1}\) In other words, the model included both random intercepts and slopes (cf., Barr, Levy, Scheepers \& Tily, 2013). We also tested the interaction of Head Position with the control variables Block
}

Finally, we also controlled for potential long-term changes in each block to see whether the horizontal and vertical effects would become stronger or weaker as the experiment progressed.

\section*{Results}

Compressive scaling Before examining the effect of spatial position, we looked at whether participants' randomly generated numbers would exhibit a small number bias (see Loetscher \& Brugger, 2007). A regression of frequency on number reveals that on average, per each increase of number by 1, frequency decreased by \(8.96(S E=1.78)\) \(\left(\mathrm{F}(1,28)=25.35, \mathrm{p}=0.000025, R^{2}=0.46\right)\). There was no interaction between the small number bias and response orientation (horizontal, vertical, straight) \((\mathrm{F}(2,84)=2.539\), \(\mathrm{p}=0.085\) ). Thus, within each condition (horizontal, vertical, straight), the small number bias was of similar magnitude.

Order effects For absolute numbers, the control variables (Block Order, Starting Orientation, Trial Order) did not interact with Head Position in the horizontal block \(\left(\chi^{2}(3)=1.36, \mathrm{p}=0.71\right)\) or in the vertical block \(\left(\chi^{2}(3)=3.47\right.\), \(\mathrm{p}=0.32\) ). For relative numbers, the control variables also failed to produce any interaction in the horizontal block \(\left(\chi^{2}(3)=3.56, p=0.31\right)\) and in the vertical block \(\left(\chi^{2}(3)=4.78\right.\), \(\mathrm{p}=0.19\) ). This suggests that the effects reported below are relatively independent from these other factors.

Absolute numbers Numbers generated were on average \(0.26(S E=0.24)\) larger when people looked to the right versus to the left, but the effect was not significant \(\left(\chi^{2}(1)=1.2, \mathrm{p}=0.27\right)\). There was, however, a significant effect of vertical position \(\left(\chi^{2}(1)=7.91, \mathrm{p}=0.0049\right)\), with numbers being \(0.67(S E=0.24)\) larger up as opposed to down. To test whether the difference between the two axes is significant, we coded "up" and "right" together and "left" and "down", combining them into a single factor "Position". There was no interaction between "Position" and "Axis Orientation". Thus, for absolute numbers, there is no conclusive evidence for the vertical effect being stronger.

Relative numbers There was a significant effect of horizontal position \(\left(\chi^{2}(1)=4.31, \mathrm{p}=0.038\right.\) ), with numbers being \(+0.52(S E=0.25)\) larger than the preceding number when people looked right versus left, and a significant effect of vertical position \(\left(\chi^{2}(1)=8.13, \mathrm{p}=0.004\right)\), with numbers being \(+1.34(S E=0.46)\) larger than the preceding number in the up position versus the down position (see Fig. 1). An analysis that combined both axis orientations also yielded a significant interaction of Axis Orientation and Position \(\left(\chi^{2}(1)=10.706, p=0.001\right)\), with the vertical effect predicted to be stronger by \(+0.82(S E=0.47)\).


Figure 1: Average relative difference for the horizontal and vertical blocks. Error bars indicate standard errors (taken from the model).

Effect sizes \({ }^{2}\) Standardized effect measures showed stronger effects for the vertical than for the horizontal condition for absolute and relative numbers. This is also reflected in the larger coefficients for the vertical condition in the analyses reported above, as well as the significant interaction between Axis and Position for relative numbers.

Table 1: Effect sizes for absolute and relative numbers by condition.
\begin{tabular}{ll}
\hline Analysis & Cohen's d \\
\hline Horizontal, absolute & 0.17 \\
Vertical, absolute & 0.43 \\
Horizontal, relative & 0.44 \\
Vertical, relative & 0.72 \\
\hline
\end{tabular}

Individual differences We analyzed individual differences by looking at difference scores for right minus left (henceforth "horizontal bias") and for up minus down (henceforth "vertical bias"). For the horizontal condition, \(61.5 \%\) (40 participants) showed a horizontal bias (positive difference score), in line with the SNARC effect for both absolute and relative numbers. Similarly, \(61.5 \%\) of all participants showed a vertical bias for both absolute and relative numbers.

While only \(\sim 15 \%\) ( 10 people) had no horizontal bias and no vertical bias (hence, showing opposite effects of what was predicted by both mappings), about \(38 \%\) ( 25 people) had both a horizontal bias and a vertical bias simultaneously. However, the majority of participants (30 people, \(46 \%\) ) had either one bias or the other. Table 2 summarizes this result:

\footnotetext{
\({ }^{2}\) As there are no standardized effect size measures for mixed models, we chose Cohen's d as shorthand. To calculate this measure, we used a by-participants analysis (averaging over trials).
}

Table 2: Individuals with horizontal or vertical bias.
\begin{tabular}{|c|c|c|c|}
\hline \multirow[b]{4}{*}{Vertical} & \multirow[b]{4}{*}{\[
\begin{aligned}
& \text { yes } \\
& \text { no }
\end{aligned}
\]} & \multicolumn{2}{|l|}{Horizontal} \\
\hline & & & \\
\hline & & 25 & 15 \\
\hline & & 15 & 10 \\
\hline
\end{tabular}

The preceding discussion of individual differences examined propensity to show vertical or horizontal effects categorically (sign of the difference score), yet it is also useful to carefully consider the relative strength of the vertical or horizontal bias per person.


Figure 2: Correlations of horizontal bias ( x -axis) and vertical bias ( y -axis) for (a) absolute numbers and (b) relative numbers. Dashed lines represent correlation without influential points; solid lines with all data.

The solid lines that are shown in Fig. 2 reveal an apparent correlation between the vertical and the horizontal bias. For both absolute numbers and relative numbers, this correlation became significant (absolute: \(\mathrm{t}(63)=2.62, \mathrm{p}=0.011\); relative: \(\mathrm{t}(63)=2.35, \mathrm{p}=0.022\) ). However, visual inspection and influence diagnostics revealed that there were a few individuals with substantial leverage on the data. If data points with large Cook's distance (over 4/(N-k-1)) were excluded, both correlations cease to be significant (absolute: \(\mathrm{t}(57)=0.9, \mathrm{p}=0.37\); relative: \(\mathrm{t}(57)=0.57, \mathrm{p}=0.57\) ).

\section*{Discussion}

Vertical versus horizontal mappings For both absolute and relative numbers, we found stronger effects for the vertical than for the horizontal axis (as indicated by Cohen's d and model coefficients), and for relative numbers, the vertical axis produced significantly stronger results than the horizontal axis.

Why did we find the vertical mapping to be stronger than the horizontal one? And, does this necessarily stand against the results of Holmes and Lourenco \((2011,2012)\) discussed above? We are cautious to conclude that these differences in effect size reflect a straightforward difference in the "strength" or "entrenchment" of the underlying mappings. There are several alternative reasons for why one mapping could lead to stronger or more consistent effects than the other. For example, people often perform smaller vertical head movements than horizontal ones (Glenn \& Vilis, 1992; Pelz, Hayhoe, \& Loeber, 2001), which could have made vertical head movements more salient. People's vertical saccades are also known to be slower and less accurate than their horizontal saccades (Collewijn, Erkelens, \& Steinman, 1988). And research on the vertical-horizontal illusion shows that people generally overestimate vertical extent more than horizontal extent (Finger \& Spelt, 1947; Chapanis \& Mankin, 1967; Prinzmetal \& Gettleman, 1993). Even though the difference in saccades and the verticalhorizontal illusion might be deemed irrelevant given that our task required participants to close their eyes, overall, these results point to fundamental asymmetries between horizontal and vertical space. Thus, it is not impossible that we found vertical effects simply because vertical space is more salient than horizontal space.

This alternative explanation opens up many interesting avenues for future research. As there are considerable interindividual differences in the amount and degree to which individuals move their head (Fuller, 1992; Stahl, 1999), one could correlate each participant's "head movement propensity" with the size of the vertical or horizontal effect. Ideally, one would like to correlate the strength of the head movement with the results of the number generation task on a trial-by-trial basis. The vertical and horizontal biases are predicted to be stronger for relatively larger movements. Moreover, there are also individual differences in the strength of the vertical-horizontal illusion (e.g., Coren \& Porac, 1987). The susceptibility to this illusion could also be correlated with the horizontal or vertical bias. Here, people who have a stronger vertical-horizontal illusion should show stronger vertical effects. If, however, vertical space is more "salient" across the board, it does not fully account for the difference between Holmes and Lourenco \((2011,2012)\) and the present study, because presumably, the same verticalhorizontal asymmetries should be at play.

Here, it is noteworthy that many studies that found vertical effects either invoke random number generation (present study, Loetscher et al., 2010 and Hartmann et al., 2011) or approximate quantity information such as the words "more" or "less" (Sell \& Kaschak, 2012; Pecher \&

Boot, 2011). Moreover, Holmes and Lourenco (2012) found vertical effects only after priming magnitude. This invites the hypothesis that the vertical mapping might be stronger in tasks that invoke a more approximate number system. The small number bias observed in our participants' responses would support this view, as a compressive scaling of the mental number line is associated with the idea of the approximate number system.

However, there are also linguistic reasons to expect that vertical effects are stronger with approximate magnitude representations: We frequently use the words "high" and "low" and "rising" and "falling" to talk about numbers, but we do not use horizontal spatial language the same way. The linguistic vertical metaphors are degree words that underspecify the exact quantity. The underspecification of verbal metaphors might make the vertical mapping particularly amenable for approximate magnitude representations as opposed to exact quantity representations. Finally, if, as cognitive linguists have claimed (Lakoff, 1987), the vertical mapping really comes from embodied interactions with the world, a connection between approximate magnitude and verticality might ultimately have physical origins: The environmental correlation between verticality and quantity often involves uncountable quantities rather than exact numbers, for instance, when pouring liquid into a container, or when creating a pile of pebbles. The horizontal mapping, on the other hand, might be more connected to exact numerical representations because of its connection to writing and symbolic representations of numbers, which are ideal for representing exact sequences (e.g., calendars, numbers on keyboards, rulers).

Absence of order effects The absence of any order effects in the current study is somewhat surprising. That Block Order did not affect the results suggests that whatever mapping is most preferred by a participant is not primed by being exposed to a vertical or a horizontal block first. One could imagine that the vertical or horizontal effects are entirely task dependent, resulting only after a bit of exposure to the up/down or left/right going movements. The absence of an interaction between Trial Order and Head Position suggests that this was not the case. Thus, it appears that participants responded in line with horizontal or vertical SNARC effects from the very beginning of each block. In other words, the spatial numerical associations appear to have been relatively stable.

Individual differences Finally, analyzing the data of separate participants revealed considerable differences between individuals, similar to other studies that have found considerable differences in the way people respond to numerical cognition tasks (Fischer \& Campens, 2008; Fischer, 2008; Beecham, Reeve, \& Wilson, 2009). The most prevalent pattern was that participants either had a horizontal bias or a vertical bias, with a considerable number ( \(\sim 38 \%\) ) having both and only a handful ( \(\sim 15 \%\) )
having neither. There was a weak correlation between a participant's vertical bias and a participant's horizontal bias. This may initially seem to suggest overlapping mechanisms for the vertical and the horizontal mappings. However, closer inspection revealed that this correlation was largely due to a few individuals. Based on the results obtained in the current study, it is clear that more research is needed to determine whether the connection between the vertical and horizontal mapping holds across different tasks.

\section*{Conclusions}

We found that randomly generated numbers were "higher" when people looked upwards and when people looked to the right. We found the vertical effect to be stronger than the horizontal one. There were also considerable interindividual differences: Some people were more easily affected by the vertical manipulation, others, more easily by the horizontal manipulation. Across individuals, there was a weak correlation between the vertical and the horizontal effect. Future research needs to find out under which conditions vertical effects are stronger than horizontal ones, and whether the weak relationship between these two effects holds across different experimental paradigms.

\section*{Acknowledgments}

We thank Michael Spivey, Drew Abney, Timo Röttger, and Tyler Marghetis for helpful comments and suggestions. We thank Anne Warlamount for making lab space available for this experiment.

\section*{References}

Barr, D.J., Levy, R., Scheepers, C. \& Tily, H. (2013). Random-effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language, 68, 255-278.
Bates, D., Maechler, M. \& Bolker, B. (2012). lme4: Linear mixed-effects models using S4 classes. R package version R package version 0.999999-0.
Beecham, R., Reeve, R. A., \& Wilson, S. J. (2009). Spatial representations are specific to different domains of knowledge. PLoS ONE, 4, e5543.
Chapanis, A., \& Mankin, D. A. (1967). The verticalhorizontal illusion in a visually-rich environment. Perception \& Psychophysics, 2, 249-255.
Collewijn, H., Erkelens, C. J., \& Steinman, R. M. (1988). Binocular co-ordination of human vertical saccadic eye movements. Journal of Physiology, 404, 183-197.
Coren, S., \& Porac, C. (1987). Individual differences in visual-geometric illusions: Predictions from measures of spatial cognitive abilities. Perception \& Psychophysics, 41, 211-219.
Dehaene, S., Bossini, S., \& Giraux, P. (1993). The mental representation of parity and number magnitude. Journal of Experimental Psychology: General, 122, 371-396.

Finger, F. W., \& Spelt, D. K. (1947). The illustration of the horizontal-vertical illusion. Journal of Experimental Psychology, 37, 243-250.
Fischer, M. H. (2003). Spatial representations in number processing: evidence from a pointing task. Visual Cognition, 10, 493-508.
Fischer, M.H. (2008). Finger counting habits modulate spatial-numerical associations. Cortex, 44, 386-392.
Fischer, M. H., \& Campens, H. (2008). Pointing to numbers and grasping magnitudes. Experimental Brain Research, 192, 149-153.
Fuller, J. H. (1992). Head movement propensity. Experimental Brain Research, 92, 152-164.
Glenn, B., \& Vilis, T. (1992). Violations of listing's law after large eye and head gaze shifts. Journal of Neurophyshiology, 68, 309-318.
Göbel, S. M., Shaki, S., \& Fischer, M. H. (2011). The Cultural Number Line: A Review of Cultural and Linguistic Influences on the Development of Number Processing. Journal of Cross-Cultural Psychology, 42, 543-565.
Hartmann, M., Grabherr, L., \& Last, F. W. (2011). Moving along the mental number line: Interactions between whole-body motion and numerical cognition. Journal of Experimental Psychology: Human Perception and Performance, 38, 1416-1427.
Holmes, K. J., \& Lourenco, S. F. (2011). Horizontal trumps vertical in the spatial organization of numerical magnitude. In L. Carlson, C. Hölscher, \& T. Shipley (Eds.), Proceedings of the 33rd Annual Conference of the Cognitive Science Society (pp. 2276-2281). Austin, TX: Cognitive Science Society.
Holmes, K. J., \& Lourenco, S. F. (2012). Orienting numbers in mental space: Horizontal organization trumps vertical. The Quarterly Journal of Experimental Psychology, 65, 1044-1051.
Hubbard, E. M., Piazza, M., Pinel, P., \& Dehaene, S. (2005). Interactions between number and space in parietal cortex. Nature Reviews, 6, 435-448.
Ito, Y., \& Hatta, T. (2004). Spatial structure of quantitative representation of numbers: Evidence from the SNARC effect. Memory and Cognition, 32, 662-673.
Lakoff, G. (1987). Women, fire, and dangerous things: What categories reveal about the mind. Chicago: University of Chicago Press.
Loetscher, T., Bockisch, C. J., \& Brugger, P. (2008). Looking for the answer: The mind's eye in number space. Neuroscience, 151, 725-729.
Loetscher, T., Bockisch, C., Nicholls, M. E. R., \& Brugger, P. (2010). Eye position predicts what number you have in mind. Current Biology, 20, R264-265.
Loetscher, T., \& Brugger, P. (2007). Exploring number space by random digit generation. Experimental Brain Research, 180, 655-665.
Loetscher, T., Schwarz, U., Schubiger, M., \& Brugger, P. (2008). Head turns bias the brain's internal random generator. Current Biology, 18, R60-62.

Müller, D., \& Schwarz, W. (2007). Is there an internal association of numbers to hands? The task set influences the nature of the SNARC effect. Memory \& Cognition, 35, 1151-1161.
Núñez, R. (2011). No innate number line in the human brain. Journal of Cross-Cultural Psychology, 45, 651668.

Pecher, D., \& Boot, I. (2011). Numbers in space: Differences between concrete and abstract situations. Frontiers in Psychology, 2, 11.
Pelz, J., Hayhoe, M., \& Loeber, R. (2001). The coordination of eye, head, and hand movements in a natural task. Experimental Brain Research, 139, 266-277.
Perrone, G., de Hevia, M. D., Bricolo, E., \& Girelli, L. (2010). Numbers can move our hands: a spatial representation effect in digits handwriting. Experimental Brain Research, 205, 479-487.
Prinzmetal, W., \& Gettleman, L. (1993). Vertical-horizontal illusion: One eye is better than two. Perception \& Psychophysics, 53, 81-88.
R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
Rotzer, S., Loenneker, T., Kucian, K., Martin, E., Klaver, P., \& von Aster, M. (2009). Dysfunctional neural network of spatial working memory contributes to developmental dyscalculia. Neuropsychologia, 47, 2859-2865.
Schwarz, W., \& Keus, I. M. (2004). Moving the eyes along the mental number line: Comparing SNARC effects with saccadic and manual responses. Perception and Psychophysics, 66, 651-64.
Sell, A. J., \& Kaschak, M. P. (2012). The comprehension of sentences involving quantity information affects responses on the up-down axis. Psychonomic Bulletin \& Review, 19, 708-714.
Shaki, S., \& Fischer, M. H. (2012). Multiple spatial mappings in numerical cognition. Journal of Experimental Psychology: Human Perception and Performance, 38, 804-809.
Stahl, J. S. (1999). Amplitude of head movements associated with horizontal saccades. Experimental Brain Research, 126, 41-54.
Stanovich, K. E., \& West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate?. Behavioral and brain sciences, 23, 645-665.
Towse, J. N., \& Cheshire, A. (2007). Random number generation and working memory. European Journal of Cognitive Psychology, 19, 374-394.
Wood, G., Nuerk, H. C., Willmes, K., \& Fischer, M. H. (2008). On the cognitive link between space and number: a meta-analysis of the SNARC effect. Psychological Science Quarterly, 50, 489-525.
Zorzi, M., Priftis, K., \& Umiltà, C. (2002). Neglect disrupts the mental number line. Nature, 417, 138-139.

\title{
When groups should not imitate their most successful members
}

\author{
Jan K. Woike (Woike@mpib-berlin.mpg.de) \\ Max Planck Institute for Human Development, Center for Adaptive Rationality (ARC), Lentzeallee 94, 14195 Berlin, Germany \\ Jean-Philippe Bonardi \\ Université de Lausanne, Quartier UNIL-Dorigny, Bâtiment Internef 1015 Lausanne, Switzerland \\ Rocio Garcia-Retamero \\ Department of Experimental Psychology, University of Granada, Granada, Spain
}

\begin{abstract}
The imitation of successful peers is often heralded as an intelligent shortcut to reduce individual learning costs. Using computer simulations, we demonstrate that this advice can be ill-founded and harmful in a cognitive inference task involving continuous learning. In particular, success-based imitators perform worse than both learners who integrate the learning experience of all group members and isolated learners. We report on sensitivity analyses for this phenomenon and offer explanatory mechanisms.
\end{abstract}

Keywords: group decision making; imitation; social learning; computer simulation

\section*{Introduction}

The results of a recent social learning tournament (Rendell et al., 2010) suggest that it always pays to copy successful others when faced with a choice between individual learning and group learning. Yet, imitation learning is not as prevalent in the biological world as could be expected (Rieucau \& Giraldeau, 2011). Humans, at least, often orient themselves towards successful peers to shorten periods of individual exploration and try to imitate the best group member (GarciaRetamero, Takezawa, \& Gigerenzer, 2009; Garcia-Retamero, Takezawa, Woike, \& Gigerenzer, 2013). Yet, it can be argued that there are situations in which this strategy does not pay off (Denrell, 2005). In this simulation study we want to illustrate one such situation and compare individual learning with several social learning strategies in a sequential cognitive inference task based on the framework used in Garcia-Retamero, Takezawa, and Gigerenzer (2006).

\section*{The learning task and learning strategies}

\section*{The learning task}

We investigated the behavior of virtual decision makers using a paired-comparison task. This is an inference task, in which agents have to decide which of two objects has the higher criterion value. The basis for this inference are the values of a set of dichotomous variables (henceforth called cues) that characterize the two objects. An environment in this study consists of a set of \(N\) objects that are associated with criterion values and \(k\) binary cue values. The agents follow a strictly non-compensatory inference strategy: for each pair
of objects an agent \(i\) looks up cues in an agent-specific order \(O_{i}=\left(o_{1, i}, o_{2, i}, \ldots, o_{k, i}\right)\) until a cue discriminates between the two objects (i.e., until the cue value is different for the two compared objects). In this case the object with the higher cue value (i.e., the value that indicates a higher criterion value) is selected. If none of the cues discriminates, a random decision is made. All agents are given the directions of all cues in each environment. For each cue, this direction is determined to maximize the number of correct decisions assuming that all possible pairs of objects in an environment are known and considered. In this setup, the choice of the cue order alone determines the success or failure of the agent in a given environment. The problem of finding the best cue order has been proven to be computationally intractable when the whole data set is available (Martignon \& Schmitt, 1999). For the case of off-line learning (i.e., for situations in which all cue values for all objects in a decently sized sample are known), a strategy called take-the-best (TTB), which determines the order of cues \(O_{i}\) by ranking them according to their ecological validity, performs well across a variety of problems, especially for generalization tasks (Czerlinski, Gigerenzer, \& Goldstein, 1999; Gigerenzer \& Brighton, 2009). Our study focuses on on-line learning instead: For a learner in an unfamiliar decision setting and without any prior knowledge, learning has to be based on experience on a trial-by-trial basis (Hertwig, Barron, Weber, \& Erev, 2004). The learner has no prior access to information about objects or environment, and full information might be costly and time-consuming if not impossible to obtain.

\section*{Individual learning with the validity algorithm}

In on-line learning situations TTB cannot be easily applied since an agent generally has no basis for accessing or estimating the actual cue validities Todd and Dieckmann (2012) propose a learning mechanism that can be used in this setting, the validity algorithm. The validity algorithm starts out with a random cue order for the first trial and stores the values of two variables for each cue \(i\) : the number of observed discriminations \(d_{i}\) (i.e., the number of observed object pairs with different cue values for the two objects), and the number of correct predictions \(c_{i} \leq d_{i}\) (i.e., the number of observed
object pairs with different cue values, for which the object with positive cue value has a higher criterion value than the object with negative cue value). Both variables are set to zero for all cues at the start. In each learning trial a decision is made using the trial's cue order, and feedback is received on the correctness of this decision. For the first cue in the order that discriminates \(d_{i}\) is incremented by 1 and if the prediction turns out to be correct, \(c_{i}\) is incremented as well. After each trial, cue validities for all cues are estimated as:
\[
\hat{v}_{i}= \begin{cases}\frac{c_{i}}{d_{i}} & d_{i}>0  \tag{1}\\ 0.5 & d_{i}=0\end{cases}
\]

A new cue order is established for the next trial by ranking the validity estimates and adapting the cue order accordingly. Dieckmann and Todd (2004) observe that by using the validity algorithm individual performance can approach the performance that is obtained by using the ecological cue validities from the start (i.e., the performance of individuals that calculate cue validities based on the full sample and order cues according to these validities). Yet, the on-line learning process is likely to be slow and convergence cannot be guaranteed. Group learning has been proposed as a way to speed up this learning process. In this paper, we investigated whether this is in fact the case.

\section*{Group Learning Strategies}

While an isolated individual is often condemned to learn on a personal trial and error basis, humans often find themselves situated in groups that offer ways to overcome this predicament. In the current study we compare the individual performance in the sequential learning task with the performance of members in learning groups using various group learning strategies.

Group members alternate between blocks of individual trials and group exchange phases. After each individual trial, each individual's cue order is updated using the validity algorithm. In the group exchange phase individuals exchange information according to the social rule they have been assigned to. A social cue order is determined, and all individuals adjust their individual cue order. Afterwards each individual's memory is altered in accordance with the new cue order (see below) and the next individual trial block begins.

We implement the following group learning algorithms: 1) imitation, 2) the plurality rule, and 3) the averaging rule.
Imitation An easy way to learn from others is achieved by simply imitating their behavior. If the relevant aspects can be observed or communicated, some individuals can avoid undergoing a longer learning process by copying the result of others. Since it is highly likely that not everyone who is observed is suited to be an adequate model, a degree of specificity is well-advised.

The imitate-the-best rule proceeds by first identifying the individual in the group who achieved the best performance in the preceding trial block. In case of ties, this individual
is randomly chosen among those with the highest number of successes. This individual's cue order is then chosen to be the resulting group cue order and every other individual copies this cue order.

One parameter for this strategy is the number of past observations considered for determining the most successful individual. This parameter has been set to the size of the individual trial block. A trade-off has to be considered here: The longer the time frame, the more observations can be evaluated and the performance measurement might well be more reliable. On the other hand, the more observations are considered the higher the chance that an individual changes the cue order used between the trials, so that older observations may be less relevant or even misleading in regard to evaluating this individual's present cue order.
Plurality The plurality rule in standard choice contexts proceeds by letting individuals vote for their preferred option and the option with the most votes (the plurality of votes) is chosen. A variant of this rule is the ""majority rule"" that implies strictly speaking that there could not be a decision without an absolute majority of votes. So if three alternatives receive \(40 \%, 35 \%\), and \(25 \%\) of the votes, respectively, the plurality rule consistently chooses the first alternative, even if none of the alternatives has obtained more than \(50 \%\) of the votes.

The transfer to ordering cues is straight-forward: For the first and each subsequent rank position (but the last) a vote will be held, where cues whose rank has already been established cannot be voted for. Each individual selects the nonranked cue that comes first in the individual's cue order and the cue with the plurality of votes (or a random cue among those cues tied for the plurality of votes) is ranked at the position that is voted for, until the complete cue order is established. This implies that an individual can vote for the same cue more than once and that the plurality rule uses \(k-1\) voting steps for \(k\) cues.
Averaging One of the principles that underlies evidencebased approaches to management, medicine and education, is the systematic collection and analysis of empirical evidence that can inform practice. Observations are collected in databases and the effectiveness of a treatment is determined via meta-analysis across studies.

A somewhat similar strategy that can be employed by groups in the setting of the simulation is the pooling of evidence across all individuals within the group. To find a cue order, cues are evaluated by using the collected experience of all group members, there is no voting or evaluation of individual solutions. The rule is called averaging rule, because validities of cues are calculated based on average cue information. The average validities \(v_{i}^{a}\) in this case are calculated as
\[
\begin{equation*}
v_{i}^{a}=\frac{\frac{\sum_{j=1}^{n_{g}} c_{j}}{n_{g}}}{\frac{\sum_{j=1}^{n_{g}} d_{j}}{n_{g}}}=\frac{\sum_{j=1}^{n_{g}} c_{j}}{\sum_{i=j}^{n_{g}} d_{j}} \tag{2}
\end{equation*}
\]

Table 1: Data sets used to create the simulation task.
\begin{tabular}{|c|c|c|c|c|c|}
\hline \[
\begin{aligned}
& \hline \text { Data Set } \\
& \text { (Source) }
\end{aligned}
\] & Number of Cases & Number of Cues & \[
\begin{gathered}
\text { Validity } \\
\text { Average (Range) }
\end{gathered}
\] & Criterion & \[
\begin{gathered}
\text { Cues } \\
\text { (Selection) } \\
\hline
\end{gathered}
\] \\
\hline 1) Forbes 500 (StatLib) & 79 & 5 & . 80 (.67-1.00) & Profit (in million \$) & Market value, assets, sales, number of employees, profits, cash flow, sector \\
\hline 2) Ice Cream (StatLib) & 29 & 5 & . 71 (.52-.97) & Ice cream consumption (4 weeks) & Temperature information, lagged temperature, family income, price, year \\
\hline 3) Minimum Wage (UCLA) & 301 & 11 & . 54 (.46-.75) & Change in full time employees & Information about state and company, changes in employees, times, registers, salaries, etc. \\
\hline 4) Wildcat Strikes (Simonoff, 2003) & 163 & 4 & . 60 (.35-.74) & Number of strikes in a company & Number of grievances, union status, rotation status, workforce size \\
\hline 5) CPU Performance (UCI MLR) & 209 & 6 & . 87 (.79-.95) & \begin{tabular}{l}
Relative CPU \\
Performance
\end{tabular} & Machine Cycle Time, Cache Memory, Main Memory, Number of channels \\
\hline \begin{tabular}{l}
6) Land Rent \\
(Weisberg, 1985)
\end{tabular} & 58 & 4 & . 71 (.56-.96) & Rent paid per acre & Average rent, cow density, proportion of pasture land, liming requirement \\
\hline 7) Professors' Salary (Rice, 1995) & 51 & 5 & . 79 (.55-.98) & Salary & Rank, number of years in current rank, highest degree earned, number of years since degree \\
\hline 8) Software Development (JSE) & 104 & 4 & . 71 (.53-.86) & Total Work Hours & Function point count, operating system, database management system, language \\
\hline 9) Home Prices (StatLib) & 117 & 5 & . 71 (.51-.95) & Home Price & Square feet, age, taxes, city area, city location, home features \\
\hline 10) Stock Market (UCLA) & 368 & 9 & . 58 (.50-.73) & Percentage of price change at 26 weeks & Average volatility, price/sales, price/cash, debt/equity, profit margin, ROI, etc. \\
\hline
\end{tabular}

The group-based cue order is then constructed by ordering cues in descending order of average validity. The number of individual cue discriminations \(d_{i}\) for all group members is adjusted to the average number of discriminations for this cue in the group (which may be a non-integer number). Note that the same cue order would result from using the ratio of the sum of all discriminations and successful predictions (both denominator and numerator are divided by the same number), but the averaging scales the information back to the level of group members. Averaging individual validities though, would ignore the number of observations that each value is based on and would lead to different results in the general case.

\section*{Memory alteration}

In Dieckmann and Todd (2004), it is assumed that \(c_{i}\) and \(d_{i}\), the number of discriminations and successful discrimination by each cue, are recorded and recalled accurately by individuals. In this simulation we employ a variant that reflects a more realistic, imperfect memory: in fixed intervals, both \(c_{i}\) and \(d_{i}\) are randomly mutated with the constraint that the resulting order of cues has to remain constant. Our variant is therefore likely to perform worse than the original algorithm.

While this memory update is a handicap for individual learning, it is actually a vital step for the group learning algorithms: after each group phase, an individual that replaces his cue order by the newly constructed group-based order faces a problem otherwise: if he retains his old memories unchanged, there will be a high probability that his cue order will change back to the original order when the validity algorithm is employed in individual learning. Only if the first trial after the group learning phase leads to a change in the order of estimated cue validities will the group phase have any effect on the individual.

In the simulations we therefore use the following mechanism for altering an individual's memory: The cue order and the number of discriminations per cue are retained. Only in the case that any \(d_{i}\) is zero, it will be changed to one. The \(c_{i}\) on the other hand are based on percentages drawn from the
uniform distribution \(U(0.5 ; 1)\). A minimum of 0.5 is chosen, as cue directions are known a priori, and the minimum cue validity under this constraint is 0.5 . A set of \(k\) numbers \(r_{i}\) is drawn from this distribution and sorted in descending order \(\left(r_{1}, r_{2}, \ldots, r_{k}\right)\). The number of successes in memory are subsequently calculated as \(c_{i}=r_{i} \cdot d_{i}\). This procedure guarantees that the ordering of cues performed by the validity algorithm will result in the specified cue order.

The retention of all \(d_{i}\) leads to a gradually decreasing probability of switches between cue positions, as more and more observations are needed to change the ratio \(c_{i} / d_{i}\) substantially and the probability to bridge the randomly determined gaps between successive cues is affected accordingly. This property is shared with the original validity algorithm. This memory alteration can be equally applied to group learners and isolated learners, and as a rule it is applied following each group exchange phase (isolated learners are yoked to the randomization schedule of social learners).

\section*{Simulation Setup}

Environments Agents in this simulation have to solve the paired comparison task in environments constructed from data sets. None of the data sets was artificially created or hypothetically derived. All were collected in ecologically meaningful economic contexts. In each environment, agents were confronted with object pairs whose members were randomly selected from the cases in the data sets.

Using a range of empirical environments allows to increase the generalizability of our conclusions. The data sets were chosen for variability, with semantic variance and different types and numbers of cues and cases. Table 1 summarizes the set of ten environments used to generate the inference task. The data sets Forbes 500, Ice Cream (Kadiyala, 1970), and Home Prices were taken from the Statlib collection of data sets \({ }^{1}\). Minimum Wage (Card \& Krueger, 1994) and Stock Prices were taken from the UCLA collection of statistical

\footnotetext{
\({ }^{1}\) http://lib.stat.cmu.edu/
}


Figure 1: Simulation structure for a group of 5 simulated individuals


Figure 2: Performance in the base condition: lines depict the average expected accuracy of individual cue orders across individuals and environments for the four simulated learning strategies
data sets \({ }^{2}\), Wildcat Strikes accompanies (Simonoff, 2003) \({ }^{3}\). The Machine Learning Repository at UCI (Asuncion \& Newman, 2007) contributed the CPU Performance data (Ein-Dor \& Feldmesser, 1987), Weisberg (1985) references the Land Rent data. The Professors' Salary data were taken from Rice (1995) \({ }^{4}\), the Software Development data are based on Matson and Huguenard (2007).

Each data set was transformed for use in the simulation. We dichotomized all non-binary cue variables ( \(0 / 1\) ) using the median (for the Land Rent and the Professor Salary data) or the mean of each variable (for the rest of the environments). In some cases, only a subset of the original variables was included as some cue-criterion relationships could not be sensibly interpreted. In two cases the number of original variables was reduced to make their inclusion in the simulation feasible. All transformations and selections were applied before running any of our simulations.

Simulation parameters In our base condition, groups consist of five individuals each. All individuals start with ran-

\footnotetext{
\({ }^{2} \mathrm{http}: / / \mathrm{www} . s t a t . u c l a . e d u / d a t a /\)
\({ }^{3} \mathrm{http}: / / \mathrm{www}\). stern.nyu.edu/jsimonof/AnalCatData
\({ }^{4}\) http://www.amstat.org/publications/jse/jse/data/archive.html
}
dom cue orders. Blocks of individual learning trials are interspaced with phases of group exchange (see Figure 1). There are five trials per individual learning block, and in each trial an individual samples one pair of objects, makes an inference and obtains feedback. Each individual \(j\) stores three pieces of information: the number of successful predictions in the current trial block \(\left(s_{j}\right)\), the number of discriminations that each of the cues in the data set made ( \(d_{i, j}\) for each cue \(i\), i.e., the number of decisions based on this cue) and the number of successful predictions that were based on each cue \(i\left(c_{i, j}\right)\). Only the number of successes is reset at the beginning of each trial block, the other variables are changed by individual learning and the memory alteration procedure described above.

The simulation proceeds for 50 trial blocks and group exchange phases (i.e., for 250 rounds of individual trials). For each of the three social rules - imitation, plurality and \(a v\) eraging - 2,000 groups (10,000 individuals) were simulated for each data set in the base condition, while we simulated 10,000 isolated learners for comparison.


Figure 3: Effect of group size on performance relative to isolated learners after 50 group-exchange phases

\section*{Results}

\section*{Base condition}

Round-wise results for the base condition are presented in Fig. 2. Groups that implement the averaging rule and the plurality rule perform better than isolated learners immediately after the first social learning phase with the averaging rule being the best learning algorithm. On the other hand, groups relying on imitation perform even worse than isolated learners and while their performance increases over time, the distance between imitators and isolated learners actually widens.

The results point out a faulty component of the imitation strategy in this context: groups implementing imitation learning are unable to pick out the truly successful strategies based on sample information. What drives these differences between learning strategies and how robust are these findings? A closer analysis of single environments reveals a stable ordering of the algorithms (for each paired comparison between algorithms: \(p=.002, N=10\), two-sided binomial test), the results do not seem to be due to particular properties of specific data sets. In the following, we examine the sensitivity of the observed pattern regarding group-size.

\section*{The effect of larger groups}

To test the effect of group size we simulated 1000 groups for each group size across all data sets and algorithms. We chose group sizes of two to ten members and in addition 15, 20 and 25 members and averaged results after fifty group exchange phases. Fig. 3 demonstrates the effect of this variation on groups implementing the different learning strategies in terms of the relative difference to the average accuracy of isolated learners: The averaging strategy profits the most from larger group sizes with diminishing marginal returns. For the plurality rule, the addition of new members is beneficial up to about five members, while additional members do neither improve nor decrease the performance. Most strikingly though, the performance of imitators decreases with group size, showing the worst result for 25 members.

\section*{Discussion and Conclusion}

For the plurality rule a group size of two is equivalent to a random choice when the two members disagree, therefore the increase in performance for the first few members is easily explained. The effect of further members has to do with inertia: After each group exchange phase, all individuals start with the social cue order established in that phase. To change the social order in the following social learning phase, a plurality of individuals must have changed their order in the same way. In a larger group, the plurality choice can be considered to be more valid, but it becomes more difficult to change the social order, as each individual learns from different paired comparisons resulting in potentially different changes in individual cue orders. As a plurality of votes is necessary for a change this can lead to inertia.

Imitation learning is yet a special case: With larger group sizes it is more probable to find orders with a successful past track record. Based on the observation that even a random cue order achieves a decent rate of successful comparisons,there is a good chance of finding an individual with perfect or near-to-perfect track record for a block of trials. What creates a first problem, is the fact that the chosen individuals will have learned based on mostly or exclusively successful comparisons, as any failure lowers the success score. The success of individuals could be due to encountering simpler problems in the environment (that can be solved using many cue orders). As candidate solutions are developed by individual learning only, the group will retain strategies that are based on biased samples, on an "under-sampling of failure" (Denrell, 2003). A second problem for imitation strategies is that a suboptimal change in a single individual order combined with an unfortunate distribution of encountered object pairs can be transferred to all group members in one group exchange phase. It is much more difficult to retain learning increments using this strategy.

In this study, we have identified one class of situations, in which imitation learning based on observed success is not only worse than often feasible social learning alternatives, but it is even worse than isolated individual learning. The problems encountered by the learning algorithm are likely to
be endemic in other relevant decision contexts. The inferiority of the learning principle may be difficult to notice, as the performance of imitators increases over trials as for all other strategies, and there might be no other strategies for comparison. Stereotypical causal explanation patterns will assume that something special and unique to the individual must have been responsible for success. This has to be seen in sharp contrast with situations in which the quality of a strategy can be judged using external or logical criteria. The performance of imitation based on mere past success constitutes a lower bound for imitation performance. Concrete evidence for decision makers falling into the trap of taking past success as a proxy for future performance and ignoring sampling and selection processes is given by Offerman and Schotter (2009). In their experiments participants tend to copy the behavior of decision makers that took great risks and were lucky to get the best possible outcome using a risky strategy with a lower expected value than alternative strategies. In an ongoing learning context these problems might well be attenuated and decision makers should eschew the choice of naive benchmarking procedures and mindless imitation of the most successful (Denrell \& Liu, 2012; Pfeffer \& Sutton, 2006), if they want to channel the power of social learning into obtaining the best possible results.

\section*{References}

Asuncion, A., \& Newman, D. J. (2007). UCI Machine Learning Repository (http://www.ics.uci.edu/~mlearn/MLRepository.html).
University of California, Irvine, School of Information and Computer Sciences.
Card, D., \& Krueger, A. (1994). Minimum wages and employment: A case study of the fast-food industry in new jersey and pennsylvania. The American Economic Review, 84(4), 772-793.
Czerlinski, J., Gigerenzer, G., \& Goldstein, D. G. (1999). How good are simple heuristics. In G. Gigerenzer, P. M. Todd, \& the ABC Reseach Group (Eds.), Simple heuristics that make us smart (p. 97-118). Oxford University Press, USA.
Denrell, J. (2003). Vicarious learning, undersampling of failure, and the myths of management. Organization Science, 227-243.
Denrell, J. (2005). Should we be impressed with high performance? Journal of Management Inquiry, 14(3), 292-298.
Denrell, J., \& Liu, C. (2012). Top performers are not the most impressive when extreme performance indicates unreliability. Proceedings of the National Academy of Sciences, 109(24), 9331-9336.
Dieckmann, A., \& Todd, P. M. (2004). Simple ways to construct search orders. In K. Forbus, D. Gentner, \& R. Regier (Eds.), Proceedings of the 26th annual conference of the cognitive science society (pp. 309-314). Mahwah, NJ: Erlbaum.
Ein-Dor, P., \& Feldmesser, J. (1987). Attributes of the perfor-
mance of central processing units: a relative performance prediction model. Commun. ACM, 30(4), 308-317.
Garcia-Retamero, R., Takezawa, M., \& Gigerenzer, G. (2006). How to learn good cue orders: When social learning benefits simple heuristics. In R. Sun \& N. Miyake (Eds.), Proceedings of the 28th annual conference of the cognitive science society (pp. 1352-1358). Mahwah, NJ: Erlbaum.
Garcia-Retamero, R., Takezawa, M., \& Gigerenzer, G. (2009). Does imitation benefit cue order learning? Experimental Psychology, 56, 307-320.
Garcia-Retamero, R., Takezawa, M., Woike, J. K., \& Gigerenzer, G. (2013). Social learning: a route to good cue orders. In R. Hertwig, U. Hoffrage, \& the ABC Research Group (Eds.), Simple heuristics in a social world (pp. 343-354). New York: Oxford University Press.
Gigerenzer, G., \& Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. Topics in Cognitive Science, 1(1), 107-143.
Hertwig, R., Barron, G., Weber, E., \& Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. Psychological Science, 15(8), 534-539.
Kadiyala, K. R. (1970). Testing for the independence of regression disturbances. Econometrica, 38, 97-117.
Martignon, L., \& Schmitt, M. (1999). Simplicity and robustness of fast and frugal heuristics. Mind Mach, 9(4), 565-593.
Matson, J. E., \& Huguenard, B. R. (2007). Evaluating aptness of a regression model. Journal of Statistics Education, 15(2).
Offerman, T., \& Schotter, A. (2009). Imitation and luck: an experimental study on social sampling. Games and Economic Behavior, 65(2), 461-502.
Pfeffer, J., \& Sutton, R. I. (2006). Evidence-based Management. Harvard Business Review, 84(1), 62-74.
Rendell, L., Boyd, R., Cownden, D., Enquist, M., Eriksson, K., Feldman, M., et al. (2010). Why copy others? insights from the social learning strategies tournament. Science, 328(5975), 208-213.
Rice, J. A. (1995). Mathematical statistics and data analysis (2nd ed.; D. Press, Ed.). Belmont, California.
Rieucau, G., \& Giraldeau, L. (2011). Exploring the costs and benefits of social information use: an appraisal of current experimental evidence. Philosophical Transactions of the Royal Society B: Biological Sciences, 366(1567), 949957.

Simonoff, J. S. (2003). Analyzing Categorical Data. New York: Springer.
Todd, P. M., \& Dieckmann, A. (2012). Simple rules for ordering cues in one-reason decision making. In P. M. Todd, G. Gigerenzer, \& the ABC Research Group (Eds.), Ecological rationality: Intelligence in the world (pp. 1393-1400). New York: Oxford University Press.
Weisberg, S. (1985). Applied linear regression (Wiley, Ed.). New York.

\title{
Why we Should Not Forget About the Non-social World: Subjective Preferences, Exploratory Eye-movements, and Individual Differences
}

\author{
David W.-L. Wu (david.wl.wu@gmail.com) \\ Department of Psychology, University of British Columbia, 2136 West Mall, Vancouver, BC, V6T 1Z4, Canada \\ Tanya Jakobsen (tanya.jakobsen@psych.ubc.ca) \\ Department of Psychology, University of British Columbia, 2136 West Mall, Vancouver, BC, V6T 1Z4, Canada
}

Nicola C. Anderson (n.c.c.anderson@vu.nl)
Department of Cognitive Psychology, Vrije Universiteit, Amsterdam, The Netherlands
Walter F. Bischof (wfb@ualberta.ca)
Department of Computing Science, University of Alberta, Edmonton, AB, Canada
Alan Kingstone (alan.kingstone@ubc.ca)
Department of Psychology, University of British Columbia, 2136 West Mall, Vancouver, BC, V6T 1Z4, Canada

\begin{abstract}
We investigated both subjective and objective differences in viewing non-social versus social scenes. Specifically, we examined four related questions: 1) Do participants prefer non-social or social scenes? 2) Are there differences in subjective exploration of non-social and social scenes? 3) Are there differences in objective exploration of these scenes? 4) Does a non-social trait connection to nature - influence the extent of non-social scene exploration? Experiment 1 found, surprisingly, that participants prefer non-social over social scenes, and correspondingly, they reported exploring these scenes more. Experiment 2 used eyetracking to test the validity of this introspection and confirmed that participants explore non-social scenes more than social scenes. We also discovered that connection to nature selectively modulates exploration of non-social scenes, demonstrating a critical interaction between observer and scene characteristics in the deployment of spatial attention.
\end{abstract}

Keywords: eye-tracking; exploration; attention; individual differences; subjective experience.

\section*{Introduction}

The desire to understand how attention is guided in the real social world has increased the use of eye movement tracking in complex natural environments. Accordingly, there has been a growing interest in the role that social stimuli play in the allocation of human attention and eye movements (for a recent review see Risko, Laidlaw, Freeth, Foulsham, \& Kingstone, 2012). However, this leaves a pertinent question unanswered: what role, if any, do nonsocial stimuli play in the allocation of attention in real world scenes?

Recent evidence indicates that when social and non-social scenes are put in direct competition, there is a distinct preference to look at social scenes, and particularly, at the eyes of the people in the social scenes (Fletcher-Watson, Findlay, Leekam, \& Benson, 2008; Birmingham, Bischof, \& Kingstone, 2008). Given that attention operates largely in
service of an individual's goals and intentions, and that looking behaviour is positively correlated with reward (Sullivan, Johnson, Rothkopf, Ballard, \& Hayhoe, 2012), a straight-forward prediction is that a selection bias for social stimuli over non-social stimuli reflects a subjective preference. However, an alternative possibility is that eye movements towards the social content of scenes (particularly the eyes of the people in the scenes) is being driven by a low-level neural system that is preferentially biased to process biologically relevant information (Laidlaw, Risko, \& Kingstone, 2012). In this case, one's subjective preference of the stimuli is not necessarily driving gaze behaviour. The aim of Experiment 1 was to determine whether subjective preference may be driving attention towards social stimuli.

Participants were asked to subjectively rate their liking for non-social scenes and social scenes. Importantly, previous research has shown that the social scenes used in the present study attract fixations to the eyes of the people in the scenes (Birmingham et al., 2008). We also asked participants to introspect on how much they thought they had explored the social and non-social scenes. We did this to investigate the accuracy of subjective intuition as to how one looks at scenes. Because our past work has shown that there is a marked tendency for participants to fixate onto the eyes of people in the scenes, we predicted that participants would report they had explored social scenes less than non-social scenes.

\section*{Experiment 1}

\section*{Methods}

Participants Sixteen students from the University of British Columbia participated in the 30 -minute experiment in exchange for course credit.

Stimuli Participants viewed a slideshow of 51 unique images at their own pace. Of interest were 6 interior and 6
landscape scenes from Foulsham, Kingstone, and Underwood (2008), as well as 7 social scenes from Birmingham et al. (2008) that have been shown to trigger rapid and sustained eye movements to the eyes of the individuals in the scenes. The social scenes either depicted 1 person alone, or 3 people interacting. Only these scenes were analyzed for the purposes of this study because they were used directly in Experiment 2. \({ }^{1}\) Exploratory eye movements are potentially affected by the saliency distribution of the stimuli (Itti, Koch, \& Niebur, 1998). For example, stimuli with widely distributed salient locations could lead to a distributed eye movement pattern, and stimuli with salient locations concentrated in a small area could lead to a concentrated pattern. For this reason, we ensured that the image areas spanned by the most salient points were matched across image types. \({ }^{2}\)

Questionnaires Participants were asked to rate, on a 5-point Likert-type scale, how much they liked the scenes, and how much they explored the scenes.

Data Analysis A one-way repeated-measures analysis of variance (ANOVA) was used to compare the average preference and exploration ratings of the interior, landscape, and social scenes.

\section*{Results}

There was a main effect of scene type on the average preference ratings, \(F(2,30)=46.94, M S E=8.95, p<.001\). Post-hoc, Bonferonni-corrected t-tests revealed that participants significantly preferred landscapes over interiors, \(t(15)=5.60, p<.001\), and social scenes, \(t(15)=8.59, p<\) .001. See Figure 1.
There was a main effect of scene type on the exploration ratings, \(F(2,30)=12.00, M S E=1.14, p<.001\). Post-hoc, Bonferonni-corrected \(t\)-tests revealed that participants reported greater exploration of the interior and landscape scenes compared to the social scenes (interiors: \(t(15)=3.56\), \(p=.003\); landscapes \(t(15)=4.32, p=.001)\). There were no significant differences in exploration ratings between interiors and landscapes, \(p>.10\). See Figure 2.

\footnotetext{
\({ }^{1}\) The remaining slideshow images were used in different experiments not reported here. They included other social, interior, and landscape scenes from the same stimuli sets, in addition to other scenes from a different stimuli set (Foulsham \& Kingstone, 2010). When all images from the current stimuli set are analyzed, the same results are obtained.
\({ }^{2}\) We computed saliency maps for each scene using the Saliency Toolbox (Walther, 2012) and determined the location of the most salient locations (defined as the set of locations with a saliency value of at least \(50 \%\) of the maximum saliency in the scene). We then calculated the image area spanned by these locations (i.e., their convex hulls). A comparison of all scene types showed that there were no significant differences in the image areas between the scenes.
}


Figure 1: Average ratings of how much participants liked each scene type. Participants significantly rated landscapes highest and social scenes lowest (all \(p\) 's \(<.001\) ). Error bars represent the standard error of the individual means.


Figure 2: Average ratings of how much participants explored each scene type. Participants significantly rated landscapes \((p=.001)\) and interiors \((p=.003)\) above social scenes. Error bars represent the standard error of the individual means.

\section*{Discussion}

In Experiment 1 we found that participants preferred the non-social landscape scenes and interior scenes over the social scenes. Given that our previous work using the current social stimuli has demonstrated a marked looking preference for social versus non-social stimuli, and that this finding has been confirmed by other researchers (FletcherWatson et al. 2008), it was reasonable to predict that this preference in looking behaviour would reflect a subjective preference for social over non-social stimuli. However, in contrast to reward theory, the data appears to support the notion that selection bias towards social stimuli is driven by something other than subjective preference.

In addition, subjects showed the strongest preference towards landscape scenes. This finding supports the idea that people have a unique preference for nature (Grinde \& Patil, 2009; Mayer \& Frantz, 2004; Nisbet, Zelenski, \& Murphy, 2011; Schultz \& Tabanico, 2007).

Consistent with past work indicating that people tend to 'lock' their eyes on people in social scenes, our study found that participants rated their exploration of the non-social scenes to be significantly greater than the social scenes. This finding seems to validate the accuracy of subjective intuition about how one allocates attention. Nevertheless, research has demonstrated that individuals can be very poor at judging whether their eyes have moved or not (e.g., Belopolsky, Kramer, \& Theeuwes, 2008). Even when individuals do realize their eyes have moved, they can be notoriously poor at judging where they may have looked (Foulsham \& Kingstone, 2013). Thus we thought that it was important to objectively confirm the validity of participants’ subjective reports, by testing subjects' exploratory looking behaviour with the same stimuli.

In addition, and in light of our recent work that an individual difference trait in visual curiosity can influence visual exploration (Risko, Anderson, Lanthier \& Kingstone, 2012), we took this opportunity to investigate the intriguing hypothesis that a non-social trait may selectively predict how one looks at non-social content. To test this idea we used the Connectedness to Nature Scale (CNS; Mayer \& Frantz, 2004). Connectedness to nature has been demonstrated to be implicitly part of an individual's identity, that is, how the natural world is included in one's representation of self (Schultz \& Tabanico, 2007).

In a sense, connection to nature can be seen as an antithesis of scales that measure social traits. While scales like the Autism Quotient (AQ) measure one's connection to the social world (Baron-Cohen, Wheelwright, Skinner, Martin, \& Clubley, 2001), the CNS measures one's connection to the non-social world. Thus, if autism spectrum disorders and social skills scores on the AQ are predictive of how people look at social content (Chen \& Yoon, 2011; Freeth, Foulsham, \& Kingstone, 2013), it is plausible that CNS may be predictive of how people look at non-social content.

In Experiment 2, participants performed a free-viewing task of interior, landscape, and social scenes while being eye-tracked, and then completed the CNS. We looked at how exploration may be different across scene types, and whether CNS scores were related to these exploratory eye movements (Risko et al., 2012). Following the results of Experiment 1, we predicted that exploration should be equal for interiors and landscapes, and significantly less in social scenes.

\section*{Experiment 2}

\section*{Methods}

Participants Twenty-three participants from the University of British Columbia were given course credit, or paid \(\$ 5\), to participate in the 30 minute study.

Stimuli The same scenes were used as the ones analyzed in Experiment 1. The scenes were \(1024 \times 768\) pixels, and corresponded to a horizontal visual angle approximately \(42^{\circ}\), and a vertical visual angle approximately \(33^{\circ}\).

Questionnaires Each participant was asked to provide demographic information, and to complete the CNS. The CNS is a 14 -item questionnaire with a 5-point Likert-type scale, used to measure participants' trait levels of feeling emotionally connected to the natural environment (Mayer \& Frantz, 2004). This inventory has acceptable internal reliability ( \(\alpha=.84\); Mayer \& Frantz, 2004).

Apparatus An SR Research Eyelink 1000 eye-tracking system, recorded participants' eye movements at 1000 Hz . Stimuli were presented to participants on a \(23^{\prime \prime}\) monitor. Scenes and eye movements were also presented to the experimenter on an adjacent monitor located in the testing room, relaying real-time feedback on system accuracy.

Procedure Participants were seated 60 cm from the computer monitor, with their heads positioned in a chin rest. Participants were told to view each image as they would normally do. Scenes were presented for 10 s. Participants viewed the images before being asked to complete the questionnaire.

Data Analysis An \(8 \times 8\) grid was created for each image, yielding 64 interest areas that were invisible to participants. Each region subtended approximately \(5.25^{\circ}\) horizontal visual angle, and \(4.13^{\circ}\) vertical visual angle. We quantified participants' exploratory eye movement behaviour using an exploratory index (EI). This index is the ratio between the number of unique regions visited in a scene, and the total number of fixations in that scene (i.e., the number of regions visited with the number of fixations normalized). We believe the EI measure gives a more accurate quantification of exploratory strategy. A raw count of regions visited is easily biased by the total fixations a participant makes: the greater the total number of fixations, the greater number of regions that would be visited simply by chance. This is reflected in the data as a raw count of regions visited correlates highly with the total number of fixations in nonsocial scenes: \(r=.73, p<.001\), and social scenes: \(r=.67, p\) \(<.001\), whereas our EI measure does not, both \(p\) 's \(>.22\). \(^{3}\) By normalizing for the number of fixations, the EI measure assesses how participants spatially allocate their attention given the same constraints (number of fixations).

It is arguable that this EI value misses within-region exploration, and reversing the ratio (number of fixations over regions visited) would better capture exploration.

\footnotetext{
\({ }^{3}\) Using raw counts of regions visited instead of EI also gives us different results. Interiors had the greatest counts, with landscapes in the middle, and social scenes garnering the least. In addition, raw counts did not correlate with CNS in social or non-social scenes.
}

However, this value would be unable to distinguish whether participants are simply repeatedly looking at the same features within a region. Given the size of the regions, it is likely only one attractive feature (e.g., eyes) is contained within it. While it is true that a participant who attends these features more can be said to have "explored" them more, it is inconsistent with our operational definition of exploration. We are interested in exploration in the sense of spatial distribution of attention, not exploration in the sense of focus toward one specific feature.

A repeated-measures ANOVA was conducted between the average EI values of each scene type. Pearson's correlations were conducted separately between the EI mean for non-social scenes and CNS scores, and the EI for social scenes and CNS scores. Pearson correlations were also conducted between CNS scores and other eye movement measures (total fixations and fixation durations), to see if the trait was related uniquely to exploration.


Figure 3: Average exploratory index values of participants viewing social, interior, or landscape scenes. Participant's significantly explored non-social scenes significantly more than social scenes (both \(p\) 's \(<.001\) ). Error bars represent the standard errors of the individual means.

\section*{Results}

There was a main effect of scene type on EI values, \(F(2,44)\) \(=31.67, M S E=0.88, p<.001\). Post-hoc repeated-measure, Bonferonni-corrected, t-tests revealed that participants did not differ in exploring interiors and landscapes, but explored both significantly more than social scenes (interiors: \(t(22)=\) 10.62, \(p<.001\); landscapes: \(t(22)=6.26, p<.001)\). See Figure 3.

The correlations between CNS scores and EI measures are shown in Table 1. The correlation between CNS scores and EI for non-human scenes was significant, \(r=.43, p=.04\). However, CNS scores were not correlated with EI for social scenes, nor for any of the other eye movement measures.

Table 1: Correlations between scores on the Connectedness to Nature Scale (CNS), and EI values, total fixations, and fixation durations in non-social and social scenes. Value bolded indicates \(p<.05\).
\begin{tabular}{lc}
\hline Measure & CNS \\
\hline Non-social: & \\
EI & .43 \\
Fixation count & -.05 \\
Duration & -.03 \\
Social: & \\
EI & .08 \\
Fixation count & -.02 \\
Duration & -.07 \\
\hline
\end{tabular}

\section*{Discussion}

The results of Experiment 2 confirm our findings in Experiment 1. Participants explored the non-social scenes far more than the social scenes, in keeping with the subjective reports in Experiment 1. Equally remarkable, we found that a non-social trait, connectedness to nature, predicted the variation in exploratory eye movements. This power to predict scene exploration was specific to the nonsocial scenes.

\section*{General Discussion}

We began our work by asking the following question: In light of the field's growing interest in social attention, what important role, if any, do non-social stimuli play in the allocation of attention in real world scenes? Our work has provided at least four new insights.

First, in Experiment 1, we found that participants preferred non-social scenes - whether they are outdoor scenes or interior scenes - significantly more than social scenes. This finding suggests that subjective preference and reward mechanisms are not responsible for the preferential bias to look toward social stimuli rather than non-social stimuli.

Second, in Experiment 1 we found that participants provided a subjective report that they explored non-social scenes (both landscapes and interiors) more than social scenes. Despite good reason to question the accuracy of this self-assessment, Experiment 2 found that people did explore non-social scenes far more than social scenes.

This finding in turn revealed that people are in fact accurate at subjectively gauging the extent to which they move their eyes through different scene types.

Finally, in Experiment 2, we discovered that a non-social trait, one's connectedness to nature, was selectively related to exploratory eye-movements in non-social scenes. In other words, non-social traits can selectively influence attention in certain scene types. This finding demonstrates the importance of using non-social scenes in exploring the influence of trait differences on attention. After all, if we
had only used social scenes we would have missed the effect. It is also noteworthy that our work extends the work of Risko et al. (2012) to include a new individual trait related to exploratory eye movement behaviour connectedness to nature.

Yet why was connectedness to nature not related to exploration in social scenes? There are at least three possibilities, all worthy of future investigation. First, it may be that there is an overwhelming pull to attend to human stimuli such that any differences in attention that might be influenced by individual traits are over-ridden. Highly attractive features that capture attention, like eyes and faces, may lead to failure of an exploratory viewing bias. This possibility suggests social EI and non-social EI are not distinct constructs. CNS would be related to both social and non-social EI if not for the overwhelming pull of eyes and faces. Alternatively, human content may produce a viewing strategy itself (e.g., making sense of the scene; Birmingham et al., 2008), and this strategy is prioritized over exploratory behaviour. Such an explanation has important implications for researchers wanting to study individual differences in areas related to attentional exploration, such as inspiration, creativity, and curiosity (Fredrickson \& Anderson, 1999; Kasof, 1997; McCoy \& Evans, 2002; Risko et al., 2012; Schlewitt-Haynes, Earthman, \& Burns, 2002). The third possibility is that the CNS is not related to human connectedness. There may be 'connectedness to human' traits that could be related to human content. As mentioned previously, there is evidence that AQ scores are predictive of how people look at social content (Chen \& Yoon, 2011; Freeth et al., 2013), as well as evidence of other traits like social anxiety related to attention to social stimuli (Mansell, Clark, Ehlers, \& Chen, 1999). This would suggest that social EI and non-social EI are distinct constructs, and that different traits would relate independently to each scene type. Further study will be needed to examine these three possibilities.

In addition to stable characteristics like personality (Risko et al., 2012), our findings raise the possibility that factors like attitudes and feelings may also influence one's eye movements, and thus be embodied in eye movement behaviour. For example, connection to nature relates positively to pro-social and outward looking values, but negatively to inward looking values (Weinstein, Przybylski, \& Ryan, 2009). Might these attitudes be embodied in broadness or narrowness of attentional focus (see Chua, Boland, \& Nisbett, 2005, for a similar hypothesis)? Since the current study was purely correlational, inferences about such a possibility cannot be made. In the future, we hope to investigate the direction of this relationship.

Results from our two experiments combined support the possibility that the bias to look at social stimuli is subserved not by one's subjective preferences for social versus non-social stimuli. If one wishes to maintain that these eye movements to social stimuli are due to reward mechanisms (Sullivan et al., 2012) then one must abandon the assumption that the reward system is related to preference.

However, if that is true, then perhaps the notion of reward itself needs to be reconceptulaized. Perhaps a better account for our results comes from recent evidence of a primitive low-level neural system that automatically drives attention and eye movements toward biologically relevant information, such as the eyes of others (Laidlaw et al., 2012; Levy, Foulsham \& Kingstone 2012).

On a more practical level, our investigation provides an example of the validity of using subjective reports, in addition to measuring objective variables. The fact that the subjective results in Experiment 1 were mirrored by the objective results in Experiment 2 mitigates some of the concerns cognitive scientists may have about doing subjective experience research (for a review, see Kingstone, Smilek, \& Eastwood, 2008). For example, we show that subjective reports are reliable and replicable across individuals. We also show that attentional exploration does not operate below conscious awareness since participants’ subjective reports were consistent with looking behaviour (cf. Nisbett \& Wilson, 1977). Our study gives empirical backing to the validity, and necessity, of using subjective experiences in addition to objective measures (Kingstone et al., 2008).

Overall, our results suggest that the use of non-social stimuli in studying real-world attention should not be overlooked. Non-social stimuli offers participants a chance to avoid the overwhelming pull of social stimuli. As such, factors like individual traits and subjective preferences that affect eye-movement behaviour may be buried when using social stimuli. On a more theoretical level, our study contributes to a burgeoning field that seeks to uncover how psychological aspects of one's identity are intimately linked to the lowest levels of one's underlying physiology (Chua et al., 2005; Dodd et al., 2012; Dodd, Hibbing \& Smith, 2011; Risko et al., 2012). Continuing to identify these subjective aspects will surely lead to interesting and important knowledge about how different individuals uniquely select, perceive, and ultimately act towards different stimuli.

\section*{Acknowledgements}

This work was funded by grants to WFB and AK from the Natural Sciences and Engineering Council of Canada.

\section*{References}

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., \& Clubley, E. (2001). The autism-spectrum quotient (AQ): Evidence from asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. Journal of Autism and Developmental Disorders, 31, 5-17.
Belopolsky, A. V., Kramer, A. F., \& Theeuwes, J. (2008). The role of awareness in processing of oculomotor capture: Evidence from event-related potentials. Journal of Cognitive Neuroscience, 20, 2285-2297.
Birmingham, E., Bischof, W. F., \& Kingstone, A. (2008). Social attention and real-world scenes: The roles of action, competition and social content. The Quarterly Journal of Experimental Psychology, 61, 986-998.

Chen F., \& Yoon, J. (2011). Brief report: Broader autism phenotype predicts spontaneous reciprocity. Journal of Autism and Developmental Disorders, 41, 1131-1134.
Chua, H. F., Boland, J. E., \& Nisbett, R. E. (2005). Cultural variation in eye movements during scene perception. Proceedings of the National Academy of Sciences of the United States of America, 102, 12629-12633.
Dodd, M. D., Hibbing, J. R., \& Smith, K. B. (2011). The politics of attention: gaze-cuing effects are moderated by political temperament. Attention, Perception, \& Psychophysics, 73, 24-29.
Dodd, M. D., Balzer, A., Jacobs, C. M., Gruszczynski, M. W., Smith, K. B., \& Hibbing, J. R. (2012). The political left rolls with the good and the political right confronts the bad: connecting physiology and cognition to preferences. The Philosophical Transactions of The Royal Society B, 367, 640-649.
Fletcher-Watson, S., Findlay, J. M., Leekam, S. R., \& Benson, V. (2008). Rapid detection of person information in a naturalistic scene. Perception, 37(4), 571-583.
Foulsham, T., \& Kingstone, A. (2013). Fixation dependent memory for natural scenes: An experimental test of scanpath theory. Journal of Experimental Psychology: General, 142, 41-56.
Foulsham, T. \& Kingstone, A. (2010) Asymmetries in the direction of saccades during perception of scenes and fractals: Effects of image type and image features. Vision Research, 50, 779-795.
Foulsham, T., Kingstone, A., \& Underwood, G. (2008). Turning the world around: Patterns in saccade direction vary with picture orientation. Vision Research, 48, 17771790.

Fredrickson, L. M., \& Andreson, D. H. (1999). A qualitative exploration of the wilderness experience as a source of spiritual inspiration. Journal of Environmental Psychology, 19, 21-39.
Freeth, M., Foulsham, T., \& Kingstone, A. (2013). What affects social attention? Social presence, eye contact, and autistic traits. PLoS ONE, 8, doi:10.1371/journal.pone. 0053286
Grinde, B., \& Patil, G. G. (2009). Biophilia: Does visual contact with nature impact on health and well-being? International Journal of Environmental Research and Public Health, 6, 2332-2343.
Itti, L., Koch, C., \& Niebur E. (1998). A Model of SaliencyBased Visual Attention for Rapid Scene Analysis, IEEE
Transactions on Pattern Analysis and Machine Intelligence, 20, 1254-1259.
Kasof, J. (1997). Creativity and breadth of attention. Creativity Research Journal, 10, 303-315.
Kingstone, A., Smilek, D., \& Eastwood, J. D. (2008). Cognitive ethology: A new approach for studying human cognition. British Journal of Psychology, 99, 317-340.

Laidlaw, K. E. W., Risko, E. F., \& Kingstone, A. (2012). A new look at social attention: Orienting to the eyes is not (entirely) under volitional control. Journal of Experimental Psychology: Human Perception \& Performance, 38, 1132-1143.
Levy, J., Foulsham, T., \& Kingstone A. (2012). Monsters are people too. Biology Letters, doi:10.1098/rsbl.2012.0850
Mayer, S. F., \& Frantz, C. M. (2004). The connectedness to nature scale: A measure of individuals' feeling in community with nature. Journal of Environmental Psychology, 24, 503-515.
Mansell, W., Clark, D. M., Ehlers, A., \& Chen, Y.-P. (1999). Social anxiety and attention away from emotional faces. Cognition \& Emotion, 13, 673-690.
McCoy, J. M., \& Evans, G. W. (2002). The potential role of the physical environment in fostering creativity. Creativity Research Journal, 14, 409-426.
Nisbet, E. K., Zelenski, J. M., \& Murphy, S. A. (2011). Happiness is in our nature: Exploring nature relatedness as a contributor to subjective well-being. Journal of Happiness Studies, 12, 303-322.
Nisbett, R. E., \& Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. Psychological Review, 84, 231-259.
Risko, E. F., Anderson, N. C., Lanthier, S., \& Kingstone, A. (2012). Curious eyes: Individual differences in personality predict eye movement behavior in scene-viewing. Cognition, 122, 86-90.
Risko, E. F., Laidlaw, K. E. W., Freeth, M., Foulsham, T., \& Kingstone, A. (2012). Social attention with real versus reel stimuli: Toward an empirical approach to concerns about ecological validity. Frontiers in Human Neuroscience, 6:143, doi:10.3389/fnhum.2012.00143.
Schlewitt-Haynes, L. D., Earthman, M. S., \& Burns, B. (2002). Seeing the world differently: An analysis of descriptions of visual experiences provided by visual artists and nonartists. Creativity Research Journal, 14, 361-372.
Schultz, P. W., Tabanico, J. (2007). Self, identity, and the natural Environment: Exploring implicit connections with nature. Journal of Applied Social Psychology, 37, 12191247.

Sullivan, B. T., Johnson, L., Rothkopf, C. A., Ballard, D., \& Hayhoe, M. (2012). The role of uncertainty and reward on eye movements in a virtual driving task. Journal of Vision, 12(13):19, 1-17.
Walther, D. The Saliency Toolbox (Version 2.2). Available from http://www.saliencytoolbox.net.
Weinstein, N., Przybylski, A. K., \& Ryan, R. M. (2009). Can nature make us more caring? Effects of immersion in nature on intrinsic aspirations and generosity. Personality and Social Psychology Bulletin, 35, 1315-1329.

\title{
The invisible hand: Toddlers infer hidden agents when events occur probabilistically
}

\author{
Yang Wu (yangwu@mit.edu), Paul Muentener (pmuenten@mit.edu) and Laura E. Schulz (lschulz@mit.edu) \\ Department of Brain and Cognitive Sciences, MIT \\ 77 Massachusetts Avenue, Cambridge, MA 02139 USA
}

\begin{abstract}
This study looked at whether toddlers posit the existence of unobserved causes when events occur probabilistically. Older (18-24 months) and younger (12-17 months) children were introduced to novel events. An experimenter pressed a red handle and a lollipop emerged from a box; she then pressed a green handle and a cake emerged. These events were repeated three times. On the fourth trial, the experimenter switched either the order or relationship between events. In the Deterministic condition, the experimenter pressed the green handle first and the red handle second; in the Probabilistic condition, the red handle produced the cake and the green handle produced the lollipop. On the test trial, the experimenter pressed the red handle and a hand emerged, holding the lollipop. The older toddlers looked longer at the hand in the Deterministic than the Probabilistic condition, suggesting they inferred a hidden cause when the events occurred probabilistically.
\end{abstract}

Keywords: causal learning, determinism, toddlers, lookingtime measures.

\section*{Introduction}

The \(19^{\text {th }}\) century mathematician Pierre Simon-LaPlace speculated that if there were an intellect capable of analyzing all the forces operating in nature "to it nothing would be uncertain; the future, like the past, would be as the present before its eyes." Twentieth century physics has made this view untenable; we now know that our universe is comprised of irreducible uncertainties. Nonetheless, the idea of indeterminate events boggles the imagination. We cannot resist explanation even if our world is resistant to it.

However inaccurate, a belief causal determinism may be advantageous for learning. A deterministic universe provides well-specified conditions under which a learner can infer the existence of hidden variables. If events appear to occur spontaneously, either an unobserved generative cause is present or an inhibitory cause is absent; if events appear to occur stochastically, either an unobserved inhibitory cause is present or a generative cause is absent.

Is a belief in causal determinism an artifact of Enlightenment thought or a fundamental feature of human cognition? Developmental evidence suggests that children resist both spontaneous and stochastic causation well before they receive formal science instruction. By the age of five, preschoolers posit hidden causes to account both for apparently uncaused events (Bullock, Gelman \& Baillargeon, 1982; Chandler \& Lalonde, 1994; Gelman, Coley, \& Gottfried, 1994) and for caused events that occur
some, but not all, of the time (Schulz \& Sommerville, 2006; see also Piaget \& Inhelder, 1975).

However, relatively little is known about the origins of deterministic beliefs earlier in development. The vast majority of studies looking at indeterminate causation in infancy have focused only on the specific case of unexplained motion events (see Gelman \& Gottfried, 1996; Gottfried \& Gelman, 2005; Leslie, 1984; Luo \& Baillargeon, 2005; Luo, Kaufman, \& Baillargeon, 2009; Markson \& Spelke, 2006; Muentener, Bonawitz, Horowtiz, \& Schulz, 2012; Premack, 1990; Saxe, Tenenbaum, \& Carey, 2005; Saxe, Tzelnic, \& Carey, 2007; Spelke, Philllips, \& Woodward, 1995). Thus for instance, if an inanimate object flies over a wall, infants seem to be less surprised if a hand is revealed at the origin of the object's movement than at the terminus of the movement, suggesting that infants posit hidden causes when objects appear to move spontaneously (Saxe et al., 2005; 2007). Recent work has extended these findings beyond motion events: infants infer the presence of agents as causes not only when objects move, but also when they change states spontaneously (i.e., when a box breaks apart or plays music; Muentener \& Carey, 2010).

However, infants' expectation that physical events have causes may not imply any broader commitment to determinism. Are toddlers sensitive to stochastic causal events as well as spontaneous ones? Some suggestive evidence that toddlers ( \(M=18\) months) resist probabilistic causation comes from the finding that toddlers imitate deterministically effective actions more faithfully than they imitate probabilistically effective ones (Schulz, Hooppell, \& Jenkins, 2008). However, we do not know whether toddlers actually posit the existence of unobserved causes when events occur probabilistically. The current study investigates this question. Given that positing unobserved variables might be more complex than differential exploration, we used 18 -months as the bottom of our range to test a group of older toddlers, \(18-24\) months, and we compared their performance to younger children, 12-17 months.

We introduced toddlers to novel causal relationships that were either deterministic (Cause A generated effect A 100\% of the time, and Cause B generated Effect B 100\% of the time) or probabilistic (Cause A generated Effect A 75\% of the time and Effect B \(25 \%\) of the time; Cause B generated Effect B \(75 \%\) of the time and Effect A \(25 \%\) of the time). We hypothesized that if toddlers are causal determinists, then they might infer the existence of an unobserved agent given probabilistic evidence but would have no reason to expect
an unobserved causal agent given deterministic evidence. Following the approach used in previous studies (Muentener \& Carey, 2010; Saxe et al., 2005; 2007) we used a human hand as the candidate causal agent. We predicted that when the hand is revealed, toddlers would look at it longer in the Deterministic condition, when no agent is expected, than the Probabilistic condition, where an unobserved cause might be inferred.

\section*{Methods}

\section*{Participants}

Sixty-four toddlers (mean age: 17.2 months; range: 12.0 to 23.5 months) were recruited at a Children's Museum. We tested both younger (12-17 months) and older (18-24 months) toddlers. There were 16 toddlers in each condition (age group \(\times\) evidence conditions). An additional 23 toddlers were recruited but not included in the final sample due to: experimenter error \((\mathrm{n}=10)\), fussiness \((\mathrm{n}=9)\), or parent interference \((\mathrm{n}=4)\). There were equal number of boys and girls.

\section*{Materials}

Toddlers were introduced to a purple box ( \(37.6 \mathrm{~cm} \times 29.2\) \(\mathrm{cm} . \times 20.3 \mathrm{~cm}\).) with two handles ( 21.6 cm in length). The left handle was red with black stripes. The right handle was green with white spots. See Figure 1. The box was placed in front of a black foam board screen \((117.9 \mathrm{~cm} \times 97.8\) \(\mathrm{cm})\). The experimenter could hide behind the screen and observe the child through pinholes in the screen. Two openings in the screen on either side of the box allowed the experimenter to reach her hands through to manipulate the handles. The box had an opening in the back and the top so that the experimenter could conceal her hand in the box and lift objects out of the box. When a handle was pressed the experimenter lifted either a lollipop ( 9.4 cm in diameter) or a toy cake ( 7.6 cm in height, 7.6 cm in width) out of the box. An MVP player was also used: the red handle was always accompanied by the sound of an ascending scale on a xylophone; the green handle was always accompanied by the sound of a descending scale on a xylophone.

\section*{Procedure}

Toddlers were recruited from a local Children's Museum and tested in a private room located on the museum floor. The child was placed in a high chair approximately 100 cm in front of the box. The child's parent sat to the right of the high chair, out of the child's direct line of sight.

The experimenter pointed to the box and the two handles. See Figure 1. Then she went behind the screen. The experimenter knocked on the center of the box behind the screen and said, "Hi, [child's name]! Watch this box!" She began the Familiarization Trials by putting her hand out of the left hole and waving at the child. She then pressed the red handle and, with her other hand concealed in the box, triggered the ascending scale and lifted the lollipop
out of the box. Pilot work established that to an adult observer, it looked like the handle caused the lollipop to emerge from the box. She held the lollipop up for 2 seconds

\section*{Initial Display}

A box with two handles.

\section*{Familiarization Trials}

When the red handle is pressed, the lollipop pops up; when the green handle is pressed, the cake pops up. This sequence was repeated three times.


Red handle - Lollipop


Green handle - Cake


\section*{Test Trial}

When the hand presses the red handle, the other hand holding the lollipop pops up.


Figure 1: Procedure of the experiment.
and then released the red handle and simultaneously returned the lollipop to the box. She brought her hand back behind the screen. She then put her hand out of the right hole and waved at the child. She pressed the green handle and, with her other hand concealed in the box, triggered the descending scale and lifted the cake out of the box. She held the cake up for 2 seconds and then released the green handle and simultaneously returned the cake to the box. The
experimenter repeated the familiarization trials a total of three times.

On the Switch Trial, the experimenter said, "[child's name], watch!" In the Deterministic condition, she switched the order of events, repeating the events in the Familiarization Trials except that she pressed the green handle first and the cake popped up; she then pressed the red handle and the lollipop popped up. In the Probabilistic condition, the experimenter switched the relationship between events, repeating the events in the Familiarization Trials except that when the experimenter pressed the red handle, the cake popped up; when she pressed the green handle, the lollipop popped up.

On the Test Trial, the experimenter put her hand out of the left hole and waved to the child. She then said "Aha!", pressed the red handle and lifted her hand holding the lollipop all the way out of the box so that both her hand and the lollipop were visible to the child. She remained stationary in this position until the child looked away from the stage for at least 2 consecutive seconds. Note that she ran the familiarization, switch and test trials cohesively without measuring children's looking times to each trial. Most of the children were engaged in the experiment and kept looking all through it. A coder blind to the conditions coded the children's looking times from the beginning of the "Aha!" sound to the start of the 2-second looking away off-line from videotape. The blind coding from videotape corroborated the experimenter's online judgment in all but three cases; three children were dropped from the analysis and replaced due to premature termination of the test trial. A second coder blind to the conditions coded one third of the clips. Inter-coder reliability was \(95.6 \%\).

\section*{Results}

We examined the effect of the condition manipulation on toddlers' looking time to the test trial separately within each age group (12-17 months, 18-24 months; see Figure 2). The 12-17 month olds, looked equally long at the test trial in the Deterministic and Probabilistic conditions (Deterministic mean: 10.5 s ; Probabilistic mean: \(9.8 \mathrm{~s} ; t(30)\) \(=.39, p=.698\) ). However, the \(18-24\) month-olds looked significantly longer at the test trial in the Deterministic condition than the Probabilistic condition (Deterministic mean: 13.7 s ; Probabilistic mean: \(8.2 \mathrm{~s} ; t(30)=2.51, p=\) .018 ). This is consistent with the possibility that children had inferred the presence of an unobserved candidate cause in the Probabilistic condition but not the Deterministic condition.
Note the Switch Trial and the Test Trial were perceptually more similar to each other in the Deterministic condition than the Probabilistic condition. In the Deterministic condition, the only difference between the last event of the Switch Trial and the Test Trial was the presence of the hand; in the Probabilistic condition both the handle pressed and the hand differed. This suggests that the toddlers in the Deterministic condition looked longer at the Test Trial not because it was perceptually more novel but because the
hand was more unexpected in the Deterministic condition than the Probabilistic condition.


Figure 2: Looking times on the deterministic and probabilistic conditions as a function of age. \({ }^{*} p<.02\)

\section*{Discussion}

These results suggest that 18-24 month-olds posit unobserved causes when they observe probabilistic evidence. Toddlers who saw a stochastic relationship between causes and effects appeared to be less surprised that a human hand was involved in the events than toddlers who saw a deterministic relationship. This is consistent with the possibility that toddlers are causal determinists.

However, the current results also leave a number of questions unanswered. Children may assume that artifacts behave deterministically without extending this assumption to the physical world more broadly. Artifacts, including the stimuli used here, may have particularly salient, and familiar, deterministic causal relationships. We do not know to what extent toddlers would infer the presence of unobserved causes to account for naturally occurring probabilistic events. Additionally, we do not know to what extent either adults or children extend a belief in causal determinism beyond the physical world, to psychological and social events. Future research might investigate the extent to which children draw inferences consistent with causal determinism across a broader range of contexts.

Additionally, we do not know why the younger toddlers in our study failed to distinguish the Deterministic and Probabilistic conditions. It is possible that \(12-17\) montholds accept that events can happen stochastically. Alternatively, given that there were only three Familiarization Trials and a single Switch Trial, the distinction between the conditions may have been too subtle for the younger toddlers to detect. Younger toddlers might have failed to learn the causal relationships initially, or failed to detect either the order or relational change. Given more exposure to the target events, even young toddlers and infants might posit unobserved causes to explain probabilistic evidence.

Finally, it would be interesting to know what kinds of unobserved causes children allow as explanations for
probabilistic evidence. Saxe et al. (2005) found that infants failed to treat one object as a cause of another object's motion, although they accepted both a hand and a novel, handless agent puppet as potential candidate causes (Saxe, et al., 2007). Similarly Muentener \& Carey (2010) found that infants accepted a hand but not an object as a candidate cause of another object changing state. Finally, Newman, Keil, Kuhlmeier, and Wynn (2010) found that infants expected that agents (balls with eyes) could add order or structure to a scene but that objects (balls without eyes) could not. In this study, we only tested toddlers' inferences about human agents as candidate causal agents in probabilistic events; it would be interesting to know whether toddlers' inferences about the unobserved causes behind probabilistic events are limited to agents.

What the current study does establish is that even toddlers' causal inferences go well beyond the evidence they observe. Given sparse data for a novel probabilistic causal relationship, toddlers inferred the existence of an unobserved causal agent. To the degree that a belief in causal determinism shapes the inferences even of one-and-a-half-year-olds, they are well equipped for exploration and discovery.

\section*{Acknowledgements}

Warm thanks to the Boston Children's Museum and participating parents and children. Thanks to Anne Kim for help with data collection and to Rachel Magid for help with making the stimuli and coding the data. This research was funded by a National Science Foundation CAREER award to L.S.

\section*{References}

Bullock, M., Gelman, R., \& Baillargeon, R. (1982). The development of causal reasoning. In W. J. Friedman (Ed.), The developmental psychology of time (pp. 209254). New York, NY: Academic Press.

Chandler, M. J., \& Lalonde, C. E. (1994). Surprising, miraculous, and magical turns of events. British Journal of Developmental Psychology, 12, 83-95.
Gelman, S. A., Coley, J. D., \& Gottfried, G. M. (1994). Essentialist beliefs in children: The acquisition of concepts and theories. In L. A. Hirschfeld \& S. A. Gelman (Eds.), Mapping the mind: Domain specificity in cognition and culture. New York: Cambridge University Press.
Gelman, S.A. \& Gottfried, G.M. (1996). Children's Causal Explanations of Animate and Inanimate Motion. Child Development 67(5), 1970-1987.
Gottfried, G. M., \& Gelman, S. A. (2005). Developing domain-specific causal-explanatory frameworks: The role of insides and immanence. Cognitive Development, 205, 137-158.
Leslie, A. M. (1984). Infant perception of a manual pick-up event. British Journal of Developmental Psychology, 2, 19-32.

Luo, Y., \& Baillargeon, R. (2005). Can a self-propelled box have a goal? Psychological reasoning in 5-month-old infants. Psychological Science, 16, 601-608.
Luo, Y., Kaufman, L., \& Baillargeon, R. (2009). Young infants' reasoning about physical events involving selfand nonself- propelled objects. Cognitive Psychology, 58, 441-486.
Markson, L., \& Spelke, E. S. (2006). Infants' rapid learning about self-propelled objects. Infancy, 9, 45-71.
Muentener, P., Bonawitz, E., Horowitz, A., \& Schulz, L. (2012). Mind the gap: Investigating toddlers' sensitivity to contact relations in predictive events. PLoS ONE, 7(4).
Muentener, P., \& Carey, S. (2010). Infants' causal representations of state change events. Cognitive Psychology, 61, 63-86.
Newman, G., Keil, F., Kuhlmeier, V., \& Wynn, K., (2010). Sensitivity to Design: Early Understandings of the Link Between Agents and Order. Proceedings of the National Academy of Sciences, 107, 17140-17145.
Piaget, J., \& Inhelder, B. (1975). The origin of the idea of chance in children. New York: Norton.
Premack, D. (1990). Do infants have a theory of selfpropelled objects? Cognition, 36, 1-16.
Saxe, R., Tenenbaum, J., \& Carey, S. (2005). Secret agents: 10 and 12 -month-olds infer an unseen cause of the motion of an inanimate object. Psychological Science, 16(12), 995-1001.
Saxe, R., Tzelnic, T., \& Carey, S. (2007). Knowing who dunnit: infants identify the causal agent in an unseen causal interaction. Developmental Psychology, 43(1), 149-158.
Schulz, L., Hoopel, K., \& Jenkins, A. (2008). Judicious imitation: Young children imitate deterministic actions exactly, stochastic actions more variably. Child Development, 79, 395-410.
Schulz, L. E., \& Sommerville, J. (2006). God does not play dice: Causal determinism in children's inferences about unobserved causes. Child Development, 77(2), 427-442.
Spelke, E. S., Phillips, A., \& Woodward, A. L. (1995). Infants' knowledge of object motion and human action. In D. Sperber, D. Premack, \& A. J. Premack (Eds.), Causal cognition: A multidisciplinary debate (pp. 4478). New York: Clarendon Press/Oxford University Press.

\title{
Development of Sharing in Preschoolers in Relation to Theory of Mind Understanding
}

\author{
Zhen Wu \({ }^{\mathbf{1 , 2}}\) (zhen-wu@uiowa.edu) \\ \({ }^{1}\) Department of Psychology, Peking University \\ Beijing, 100871, China \\ \({ }^{2}\) Department of Psychology, E11 Seashore Hall \\ Iowa City, IA, 52246 USA \\ Yanjie Su (yjsu@pku.edu.cn) \\ Department of Psychology, Peking University \\ Beijing, 100871 China
}

\begin{abstract}
This study aimed to explore the relationship between children's sharing behavior and Theory of Mind (ToM) understanding. Seventy-four 2 to 4 years old Chinese children participated in 3 tasks using toys that could be shared with a puppet that was animated by a female experimenter. On each task, the puppet expressed her desire for the items using a series of cues that progressively became more communicative. Children's ToM understanding was assessed with the scale of ToM tasks (Wellman, Fang, Liu, Zhu, \& Liu, 2006). There were two main findings: (1) younger children relied on more explicit communicative cues to share resources with the puppet, while older children shared more spontaneously and (2) children's sharing behavior was positively correlated to their ToM scores, independent of age. Findings suggest that preschoolers' sharing behavior is enhanced by their ToM understanding and explicit communicative cues provided by a social partner.
\end{abstract}

Keywords: preschoolers; sharing behavior; Theory of Mind understanding.

\section*{Introduction}

Sharing is an important aspect of human cooperative activities with roots very early in life. Studies have shown that infants as young as 8 months old show spontaneous offering of food and other objects to parents (e.g., Hay \& Murray, 1982; Hay, 1979; Rheingold, Hay, \& West, 1976). Though sharing emerges early, it appears to be a unique challenge for young children. Sharing resources is a much less frequent activity compared to other cooperative activities in young children, such as empathy-related responding, helping and instrument collaboration (Eisenberg, 2005; Grusec, 1991; Warneken \& Tomasello, 2007, 2009). Toddlers share toys with others rarely, though the rate of sharing increases from 12 to 30 months of age (e.g., Brownell, Svetlova, \& Nichols, 2009; Hay, Castle, Davies, Demetriou, \& Stimson, 1999; Levitt, Weber, Clark, \& McDonnell, 1985). Other work has shown that preschoolers also share little. For example, three- to five-year-old children reserved 10 pieces of food for themselves while sharing only one piece of food with their peers (Birch \& Billman, 1986). Sharing is difficult for children probably because it results in a sacrifice of something valued for the
welfare of someone else. There is evidence suggesting that children share more if there is no sacrifice compared to identical circumstances with sacrifices (Svetlova, Nichols, \& Brownell, 2010; Thompson, Barresi, \& Moore, 1997).

Though numerous studies have been conducted on sharing behavior, it remains difficult to depict the development of sharing, partially because many studies do not control the circumstances under which sharing was observed. For example, we do not know whether early social acts of offering items to others are primarily otheroriented unless we control for the social partner's behavior. In these cases, infants may just be seeking attention or reaction from the social partner, or may be complying with the request of others. Previous studies have shown that 2-year-old children shared food with an adult only when the adult actively communicated directly about what she wanted (Brownell et al., 2009), or reached toward the child with palm up while alternating gaze between the child and the food (Dunfield, Kuhlmeier, O’Connell, \& Kelley, 2011). Therefore, it is unclear to what extend children's "sharing" acts are truly sharing behavior with the intention to benefit others if the recipient's behavior is not controlled. By systematically manipulating the social partner's behavior, we might get a comprehensive understanding of the development of other-oriented resource sharing (see also Brownell, Iesue, Nichols, \& Svetlova, 2012). Therefore, the first goal of the present study was to examine how the social partner's communication supports young children's sharing. To do this, we systematically manipulated the expressive cues provided by the partner in the sharing task such that the partner made her needs progressively more explicit with a fixed sequence of cues. We then aimed to see at what point children would share.

Another unaddressed question is how we explain the development of sharing in children. Current theories have proposed various underlying mechanisms of sharing, such as the basic imitative tendencies toward people (Grusec, 2006; Hay \& Cook, 2007; Rheingold, 1982), the ability to differentiate self's and other's internal states (e.g., Moore, 2007), the sympathetic ability to relate self's emotions and feelings to other's (e.g., Eisenberg, 2007; Zahn-Waxler \& Radke-Yarrow, 1990), the understanding of ownership
(Brownell et al., 2012), and an innate biological preposition for empathy and altruism in infants (Tomasello, 2008; Warneken \& Tomasello, 2009b; Zahn-Waxler, Robinson, \& Emde, 1992). These different theoretical perspectives emphasize the social-cognitive and motivational components of early pro-social responding at different levels. Yet they agree, to different extents, that the origins of altruistic pro-social behavior are based on universal norms of fairness and reciprocity in combination with our understanding of other people's needs or wants (e.g., Fehr \& Fischbacher, 2004). This 'understanding of others' needs and wants' can be manifested in the Theory of Mind (ToM) understanding, which refers to the ability to attribute mental states - beliefs, intents, desires, pretending, knowledge, etc. - to oneself and others and to understand that others have beliefs, desires and intentions that are different from one's own (Premack \& Woodruff, 1978). From this perspective, is it possible that ToM understanding has an impact on prosocial behavior?

ToM understanding and pro-social behavior are both undergoing significant developmental changes during the preschool years (Benenson, Pascoe, \& Radmore, 2007; Blake \& Rand, 2010; Fehr, Bernhard, \& Rockenbach, 2008; Rochat et al., 2009; Wellman et al., 2006; Wellman \& Liu, 2004). Evidence suggests that these two abilities may share the same underlying neural processes (McCabe, Houser, Ryan, Smith, \& Trouard, 2001). That is, studies using functional magnetic resonance imaging (fMRI) found that as adult participants behaved cooperatively in a trust game, brain areas related to ToM (medial prefrontal cortex) were activated (McCabe et al., 2001). Moreover, previous studies have shown that preschoolers' ToM negatively predicted aggressive or disruptive behavior for boys and positively predicted pro-social behavior for girls after controlling for age (Walker, 2005). In addition, ToM has been suggested to be a facilitator of fairness-related behavior (Sally \& Hill, 2006), such as higher proposed offers in the Ultimatum Game (Takagishi, Kameshima, Schug, Koizumi, \& Yamagishi, 2010). These studies suggest that children with more advanced ability of understanding others' mental states may need less explicit supports from the social partner in interpreting his/her desires or needs, thus may be more likely to perform pro-social behavior. The second aim of the present study is to test this hypothesis in preschoolers.

\section*{Methods}

\section*{Participants}

Seventy-four Chinese children from two kindergartens in Beijing, China participated in the study. Both kindergartens largely served children of university staff and faculty in urban Beijing. There were twenty-five 2-year-olds ( \(M=\) 28.86 months, \(S D=3.16\), range: \(24.46-34.43\) months; 14 girls), twenty-five 3-year-olds ( \(M=39.49\) months, \(S D=\) 2.21, range: \(34.92-42.69\) months; 12 girls), and twentyfour 4-year-olds ( \(M=47.30\) months, \(S D=2.51\), range:
43.02 - 51.08 months; 14 girls). An additional 3-year-old boy was excluded from the study due to experimenter error.

\section*{Materials and Procedure}

Each child participated in both the sharing task and the Theory of Mind task in a quiet separate room adjacent to the child's classroom. Testing was conducted by a female experimenter ( E ) performing a hand puppet (a brown bear) named "Maomao" along with the help of a female assistant (AE). The session began with a warm-up and familiarization period during which Maomao (animated by E) and AE played with the child for several minutes to ensure that the child was comfortable approaching and interacting with both of them. The order of the sharing tasks and the theory of mind tasks were counterbalanced. All the sessions were video recorded.

Sharing Tasks. Three sharing tasks were administered with order counterbalanced across participants. Each task featured different items to be shared: there were 2 stickers in the 'sticker' task, 2 toy watermelons and 2 knives in the 'watermelon' task, 4 colorful beads and 2 strings in the 'bead' task. Therefore, there were 12 items in total across these three sharing tasks. These items were shown to be equally liked by children in a pilot study. Children thus had multiple opportunities to share different types of toys. This aimed to provide a relatively comprehensive measurement of sharing behavior. During the test, Maomao sat across a child-sized table from the child, and AE sat to one side of the table, at a 90-degree angle to the child and the puppet. After a short familiarization, AE brought out toys, and showed the child and the puppet how to play with these toys. AE then left the room, asking the child, "Could you please take care of these toys when I am gone? I'll be back soon. You can play with them by yourself, or with your new friend Maomao. Thanks, bye!" She then left the room. After AE left, Maomao provided three progressively more explicit cues about her needs and desires. The cues were presented in three phases: (1) Commenting phase: when the child was exploring toys, Maomao positively commented on the toys, such as "these are so beautiful!" She then repeated this 2 times, pausing for about 10 seconds between each. (2) Desiring phase: Maomao expressed her desires for these toys, such as "I like these toys! I want to play with them" in the same manner as the first phase. (3) Requesting phase: Maomao made an explicit request by asking the child for these toys, such as "would you please give me some to play with" in the same manner as the other phases. If the child shared at any point, Maomao discontinued the cues, thanked the child, and played with the toy(s) for about 10 seconds, and then signaled AE that the task ended by knocking the table. If the child did not share in the final phase, E signaled AE in the same way. AE then came back and moved on to a second sharing task with new toys.

Scale of Theory of Mind Tasks. These tasks were modeled after the Chinese version of the five core ToM
understanding tasks (Wellman et al., 2006; Wellman \& Liu, 2004). A small toy figure with Chinese visages and dark hair, whose name was Feifei, served as the target protagonist for the tasks. Although formats and general ideas were not different from the tasks used by Wellman and his colleagues (2006), task materials were modified in several places so that they were familiar and appropriate for our sample.

Diverse Desires: The participant was presented with pictures of an apple and a pear, and was asked to pick one that \(\mathrm{s} / \mathrm{he}\) liked better. Then \(\mathrm{s} /\) he was told that Feifei likes X better where X was always the opposite of the child's answer. The subsequent question for the child was which one Feifei would choose if she was hungry.

Diverse Beliefs: The participant was presented with pictures of a schoolbag and a drawer along with a picture of a car, and was told that Feifei was looking for her car. Then the child was asked to choose where (schoolbag vs. drawer) \(\mathrm{s} /\) he believed the car was. The experimenter (E) then stated that Feifei thinks the car is in X , where again X was always the opposite of the child's answer. The subsequent target question asked of the child was where Feifei was going to search for her car.

Knowledge Access: the participant saw a box with a fork inside. E then told the child that Feifei had never opened this box before. The target question was whether Feifei knew what was inside the box.

Contents False Belief: E presented the participant with a box with pictures of cookies on it. E then asked the child what \(\mathrm{s} / \mathrm{he}\) thought was inside the box. She then showed the child that it was actually a small pencil inside. After showing the child the real contents of the box, she told the child that Feifei had never opened this box. The target question was what Feifei thought was inside the box.

Real-Apparent Emotion: the participant saw a sheet of paper with three faces on it - a happy, a neutral, and a sad face. After ensuring that the child understands these emotional expressions, E told the child a story about a boy expecting a toy gun as his birthday gift, but actually getting a boring book. But he did not want to behave impolitely, so he decided to hide his feelings. The child was then asked how this boy really felt and how he tried to appear to others by pointing to the pictures with faces.

\section*{Coding and Scoring}

Each child received a score of sharing from 0-3 for each task, corresponding to the phase during which sharing occurred: \(0=\) did not share at all; \(1=\) shared in the requesting phase; \(2=\) shared in the desiring phase; \(3=\) shared in the commenting phase. Higher scores thus indicated quicker sharing with less explicit cues from the recipient. Scores were averaged over the three tasks to create an average sharing score for each child. In addition, the number of items shared by each child in all three sharing tasks ( \(0-12\) items in total) was also coded.

Children got a score of 1 for each ToM understanding task they passed. Thus, the range for ToM scores was 0-5.

The first author coded all of the videos, and another coder blind to the research goal rated \(25 \%\) of these videos. Cohen's kappa was computed to measure inter-rater reliability. Values for Cohen's kappa were 1.00 for the number of items shared in the sharing tasks as well as scores in the TOM task, and were 0.98 for the sharing score.

\section*{Results}

Preliminary analysis using the sharing score and the number of items shared by children showed that there were no significant main effects of task order or sex, nor was there an interaction between task order and sex. Therefore, the following analyses were conducted by collapsing the data across these factors.

\section*{Age Differences in Sharing and ToM}

An ANOVA test was conducted with the average sharing score as the dependent variable and age as the independent variable. Results showed that older children shared more spontaneously than did younger children, \(F(2,71)=6.38, p\) \(=.003\), partial \(\eta^{2}=.15\) (see Table 1 for means and standard deviations). Bonferroni corrected post-hoc tests showed that 4-year-old children had significantly higher sharing scores than 2 - and 3 -year-olds, \(p=.006, .011\), respectively, yet there was no significant difference between children of the two younger groups, \(p=1.00\).

Another ANOVA test was conducted with the total number of items shared as the dependent variable. Interestingly, no significant age effect was found, \(F(2,71)\) \(=1.36, p=.26\), partial \(\eta^{2}=.04\) (see Table 1).

Table 1: Means (Standard Deviations in Parentheses) for Primary Measures as a Function of Age
\begin{tabular}{lllc}
\hline Measure & 2 years & 3 years & 4 years \\
\hline \begin{tabular}{l} 
Sharing \\
Average sharing \\
score (0-3)
\end{tabular} & \(1.73(0.84)\) & \(1.77(0.79)\) & \(2.40(0.52)\) \\
\begin{tabular}{l} 
Total number of \\
items shared (0--
\end{tabular} & \(4.32(2.70)\) & \(3.96(1.79)\) & \(4.92(1.41)\) \\
\(\quad 12)\) & & & \\
\begin{tabular}{l} 
ToM
\end{tabular} & & & \\
\(\quad\) Total ToM & \(1.52(0.77)\) & \(2.16(0.80)\) & \(2.88(1.03)\) \\
Diverse Desires & \(0.92(0.28)\) & \(0.96(0.20)\) & \(0.96(0.20)\) \\
\begin{tabular}{l} 
Diverse Beliefs
\end{tabular} & \(0.20(0.51)\) & \(0.76(0.44)\) & \(0.83(0.38)\) \\
\begin{tabular}{l} 
Knowledge Access \\
Contents False \\
\(\quad\) Belief
\end{tabular} & \(0.08(0.28)\) & \(0.40(0.50)\) & \(0.67(0.48)\) \\
\begin{tabular}{l} 
Real-Apparent \\
Emotion
\end{tabular} & \(0.00(0.00)\) & \(0.00(0.00)\) & \(0.08(0.28)\) \\
\hline
\end{tabular}

In addition, children's total ToM score also increased significantly with age, \(F(2,71)=14.73, p<.001\), partial \(\eta^{2}\) \(=.29\). More specifically, the age effect was significant in children's score on the task of Diverse Beliefs \([F(2,71)=\) 3.33, \(p=.04\), partial \(\eta^{2}=.09\) ], the task of Knowledge Access \(\left[F(2,71)=11.41, p<.001\right.\), partial \(\left.\eta^{2}=.24\right]\), the
task of Contents False Belief \([F(2,71)=9.09, p<.001\), partial \(\left.\eta^{2}=.20\right]\), but not on the task of Diverse Desires [ \(F(2\), 71) \(=.24, p=.79\), partial \(\eta^{2}=.01\) ] or the task of RealApparent Emotion \(\left[F(2,71)=2.18, p=.12\right.\), partial \(\eta^{2}\) \(=.06]\), as shown in Table 1.

\section*{Associations between Sharing and ToM}

Since significant age effects were found in both scores of sharing and ToM understanding tasks, Pearson partial correlational analyses were conducted to investigate whether sharing was associated with ToM understanding after controlling for age. Results showed that the total score of ToM, Diverse Beliefs and Knowledge Access were both positively correlated to children's sharing scores and the number of items shared with age being controlled for, \(r\) ' \(s\) ranged . \(26-.36\), \(p\) 's \(<.05\) (for details, see Table 2). However, the scores of Diverse Desires and Real-Apparent Emotion did not correlate significantly to sharing.

To further examine the effect of ToM understanding on children's sharing behavior, we conducted a series of multiple regression analyses. We first regressed age on the sharing score and found that age had a significant effect on the sharing score ( \(\beta=.30, p=.01\) ), but the model only explained \(8.7 \%\) of the variance in the sharing score, \(F(1,72)\) \(=6.87, p=.01\). This age effect became non-significant \((\beta\) \(=.06, p=.65\) ) as the total ToM score was added as another independent variable. The effect of ToM was significant ( \(\beta\) \(=.42, p=.002\) ), and increased the variance accounted by \(11.8 \%, F_{\text {change }}(1,71)=10.55, p=.002\). After adding ToM as the predictor, the full model explained \(20.5 \%\) of the variance in sharing score, \(F(2,71)=9.17, p<.001\).

Table 2: The relationships between ToM and sharing after controlling for age ( \(N=74\) )
\begin{tabular}{lll}
\hline & \begin{tabular}{l} 
Average \\
sharing score
\end{tabular} & \begin{tabular}{l} 
Total number of \\
items shared
\end{tabular} \\
\hline Total ToM & \(0.36^{* *}\) & \(0.34^{* *}\) \\
Diverse Desires & 0.04 & 0.19 \\
Diverse Beliefs & \(0.29^{*}\) & \(0.22^{+}\) \\
Knowledge Access & \(0.26^{*}\) & \(0.28^{*}\) \\
Contents False Belief & 0.14 & 0.11 \\
Real-Apparent & 0.09 & 0.03 \\
\(\quad\) Emotion & & \\
\hline \multicolumn{1}{l}{ Note \({ }^{+} p<.10,{ }^{*} p<.05,{ }^{* *} p<.01\)} &
\end{tabular}

Similarly, we conducted another regression analysis with the total number of items shared as the dependent variable. Results showed that age alone had no significant effect on the number of items shared ( \(\beta=.07, p=.54\) ), and the model only explained \(0.5 \%\) of the variance in the number of shared items, \(F(1,72)=.38, p=.54\). After adding ToM as the independent variable, the full model explained \(12.3 \%\) of the variance in the number of items shared, \(F(2,71)=4.98, p\) \(=.01\). The effect of ToM was significant ( \(\beta=.42, p=.003\) ), and increased the variance accounted by \(11.8 \%, F_{\text {change }}(1\), \(71)=9.54, p=.003\). The results of the above regression
analyses indicate that preschoolers' performance on the ToM tasks, rather than age, was responsible for both the average sharing score and the number of items shared over three sharing tasks.

\section*{Discussion}

First, consistent with our hypothesis, we found that 4-year-old children needed less communicative cues to share than 2- and 3-year-old children did. Four-year-old children generally shared when the partner was just commenting on the toys, whereas the majority of the other two younger groups of children shared when the partner verbally requested the items. This result suggests that older children might need less scaffolding from a social partner to perform sharing behavior (Brownell, Iesue, et al., 2012; Brownell et al., 2009; Svetlova et al., 2010). It also implies that older children might have a more robust intention to benefit others as they shared spontaneously and quickly with a partner who had no toys, while younger children might only share under pressure or to comply with another's request (e.g., Hay, Caplan, Castle, \& Stimson, 1991). Our results thus add new evidence to previous findings that toddlers need more explicit cues to perform sharing and helping behavior (Brownell et al., 2009, 2012; Svetlova et al., 2010).

Interestingly, we found that even though 2 - and 3 -yearold children needed more communicative support in order to perform sharing behavior (their sharing scores were lower), once they shared, they shared as many objects with the social partner as older children did. Blake and Rand (2010) also found that even though 6-year-old children were more likely to donate stickers than 3-year-old children did, once they shared, they gave the same amount of stickers at all ages. These results imply that children may engage in two separate decisions when interacting with a social partner: (a) whether to share and (b) how much to share. As proposed by Blake and Rand (2010), the different developmental trajectories of these two choices may imply different processes involved in those two different phases of sharing behavior. More studies are required to identify these differential underlying processes.

More importantly, we found that children's Theory of Mind understanding correlated to their sharing behavior, independent of age. The regression analyses showed that ToM was a significant predictor of how spontaneously and quickly children shared, as well as how many items children shared, whereas age did not predict these sharing measures. These results suggest that ToM understanding might be a potential underlying mechanism of children's age-related increase in sharing.

This finding is consistent with and extends previous research demonstrating associations between theory of mind ability and pro-social behavior. For example, prior studies have shown that theory of mind ability correlates to prosocial behavior (e.g., Sally \& Hill, 2006; Walker, 2005), and there is evidence suggesting that these two abilities may share the same underlying neural processes (McCabe et al., 2001). Furthermore, studies of nonhuman primates have
found that chimpanzees, which do not have as well of a developed theory of mind as compared to humans, are rational maximizers in that they make unfair offers and accept unfair offers in the Ultimatum Game (Jensen, Call, \& Tomasello, 2007). Likewise, children who had more advanced theory of mind abilities proposed more fair offers in the Ultimatum Game than children with less advanced theory of mind (Takagishi et al., 2010). These results suggest that fairness-related behavior is related to the ability to infer the mental states and intentions of others. Our findings thus further support this hypothesis by showing that more advanced theory of mind ability was positively associated with children's spontaneous sharing and the amount of items shared with others.

In sum, the present study showed that sharing behavior may be more likely to occur when the partner makes his/her needs, desires and emotions more apparent, thus reducing the need for complex inferences about others' internal states, especially for young children whose ability to infer others' psychological states are immature. For older children who have more advanced ToM understanding, the requirement of the provision of ostensive cues from the recipient may not be necessary, thus they shared more spontaneously, more quickly and shared more items.

\section*{Acknowledgments}

This work was supported by the National Natural Science Foundation of China (Project 30970907, 31170995) and National Basic Research Program (973 Program: 2010CB833904). We thank Julie Gros-Louis, Sarah Kucker and three anonymous reviewers for their helpful comments on earlier versions of this manuscript; teachers and staff at the kindergartens for their help with recruiting participants; and children who participated in the research.

\section*{References}

Benenson, J. F., Pascoe, J., \& Radmore, N. (2007). Children's altruistic behavior in the dictator game. Evolution and Human Behavior, 28, 168-175.
Birch, L. L., \& Billman, J. (1986). Preschool children's food sharing with friends and acquaintances. Child Development, 57, 387-395.
Blake, P. R., \& Rand, D. G. (2010). Currency value moderates equity preference among young children. Evolution and Human Behavior, 31, 210-218.
Brownell, C. A., Iesue, S. S., Nichols, S. R., \& Svetlova, M. (2012). Mine or yours? Development of sharing in toddlers in relation to ownership understanding. Child Development.
Brownell, C. A., Svetlova, M., \& Nichols, S. (2009). To share or not to share: When do toddlers respond to another's needs? Infancy, 14, 117-130.
Dunfield, K., Kuhlmeier, V. A., O’Connell, L., \& Kelley, E. (2011). Examining the diversity of prosocial behavior: Helping, sharing, and comforting in infancy. Infancy, 16, 227-247.

Eisenberg, N. (2005). The development of empathy-related responding. Nebraska Symposium on Motivation, 51, 73117.

Eisenberg, N. (2007). Empathy-related responding and prosocial behaviour. Novartis Foundation symposium, 278, 71-80; discussion 80-96, 216-221.
Fehr, E., Bernhard, H., \& Rockenbach, B. (2008). Egalitarianism in young children. Nature, 454, 10791083.

Fehr, E., \& Fischbacher, U. (2004). Social norms and human cooperation. Trends in Cognitive Sciences, 8, 185190.

Grusec, J. E. (1991). Socializing concern for others in the home. Developmental Psychology, 27, 338-342.
Grusec, J. E. (2006). The development of moral behavior and conscience from a socialization perspective. In M. Killen \& J. G. Smetana (Eds.), Handbook of moral development (pp. 243-265). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.
Hay, D. F. (1979). Cooperative interactions and sharing between very young children and their parents. Developmental Psychology, 15, 647-653.
Hay, D. F., Caplan, M., Castle, J., \& Stimson, C. A. (1991). Does sharing become increasingly "rational" in the second year of life? Developmental Psychology, 27, 987993.

Hay, D. F., Castle, J., Davies, L., Demetriou, H., \& Stimson, C. A. (1999). Prosocial action in very early childhood. Journal of Child Psychology and Psychiatry and Allied Disciplines, 40, 905-916.
Hay, D. F., \& Cook, K. V. (2007). The transformation of prosocial behavior from infancy to childhood. In C. A. Brownell \& C. B. Kopp (Eds.), Socioemotional development in the toddler years: Transitions and transformations (pp. 100-131). New York, NY: Guilford Press.
Hay, D. F., \& Murray, P. (1982). Giving and requesting: Social facilitation of infants' offers to adults. Infant Behavior and Development, 5, 301-310.
Jensen, K., Call, J., \& Tomasello, M. (2007). Chimpanzees are rational maximizers in an Ultimatum game. Science, 318(5847), 107-109.
McCabe, K., Houser, D., Ryan, L., Smith, V., \& Trouard, T. (2001). A functional imaging study of cooperation in twoperson reciprocal exchange. Proceedings of the National Academy of Sciences, 98, 11832-11835.
Moore, C. (2007). Understanding self and others in the second year. In C. A. Brownell \& C. B. Kopp (Eds.), Socioemotional Development in the toddler years: Transitions and transformations (pp. 43-65). New York, NY: Guilford Press.
Premack, D., \& Woodruff, G. (1978). Does the chimpanzee have a theory of mind? Behavioral and Brain Sciences, 1, 515-526.
Rheingold, H. L. (1982). Little children's participation in the work of adults, a nascent prosocial behavior. Child Development, 53, 114-125.

Rheingold, H. L., Hay, D. F., \& West, M. J. (1976). Sharing in the second year of life. Child Development, 47, 11481158.

Rochat, P., Dias, M. D. G., Liping, G., Broesch, T., PassosFerreira, C., Winning, A. et al. (2009). Fairness in distributive justice by 3- and 5-year-olds across seven cultures. Journal of Cross-Cultural Psychology, 40, 416442.

Sally, D., \& Hill, E. (2006). The development of interpersonal strategy: Autism, theory-of-mind, cooperation and fairness. Journal of Economic Psychology, 27, 73-97.
Svetlova, M., Nichols, S. R., \& Brownell, C. A. (2010). Toddlers' prosocial behavior: From instrumental to empathic to altruistic helping. Child Development, 81, 1814-1827.
Takagishi, H., Kameshima, S., Schug, J., Koizumi, M., \& Yamagishi, T. (2010). Theory of mind enhances preference for fairness. Journal of Experimental Child Psychology, 105, 130-137.
Thompson, C., Barresi, J., \& Moore, C. (1997). The development of future-oriented prudence and altruism in preschoolers. Cognitive Development, 12, 199-212.
Tomasello, M. (2008). Origins of human communication. Cambridge, MA: MIT Press.

Walker, S. (2005). Gender differences in the relationship between young children's peer-related social competence and individual differences in theory of mind. The Journal of Genetic Psychology, 166, 297-312.
Warneken, F., \& Tomasello, M. (2007). Helping and cooperation at 14 months of age. Infancy, 11, 271-294.
Warneken, F., \& Tomasello, M. (2009a). Varieties of altruism in children and chimpanzees. Trends in cognitive sciences, 13, 397-402.
Warneken, F., \& Tomasello, M. (2009b). The roots of human altruism. British Journal of Psychology, 100, 445471.

Wellman, H. M., Fang, F., Liu, D., Zhu, L., \& Liu, G. (2006). Scaling of theory-of-mind understandings in Chinese children. Psychological science, 17, 1075-1081.
Wellman, H. M., \& Liu, D. (2004). Scaling of theory-ofmind tasks. Child Development, 75(2), 523-541.
Zahn-Waxler, C., \& Radke-Yarrow, M. (1990). The origins of empathic concern. Motivation and Emotion, 14, 107130.

Zahn-Waxler, C., Robinson, J. L., \& Emde, R. N. (1992). The development of empathy in twins. Developmental Psychology, 28, 1038-1047.

\title{
Worm holes in memory: Is memory one representation or many?
}

\author{
Dirk U. Wulff (Wulff@mpib-berlin.mpg.de) \\ Max-Planck-Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany \\ Thomas T. Hills (T.T.Hills@warwick.co.uk) \\ Department of Psychology, University of Warwick \\ Coventry CV4 7AL, UK \\ Ralph Hertwig (Hertwig@mpib-berlin.mpg.de) \\ Max-Planck-Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany
}

\begin{abstract}
The analogy of space to human cognition has a longstanding tradition. Our study aims to elaborate on the validity of this analogy for search in memory. Using the search of associative memory framework (SAM) we show that people are able to dynamically recruit independent memory representations in the recall of country names. By instructing participants to use specific recall cues we also show that despite a strong effect on the retrieval sequence, total recall from memory remains unaffected. Whereas these findings strongly support a higher dimensionality to memory than often assumed, the simultaneous finding of severe retrieval time costs for non-default representations suggests that the use of particular retrieval structures may be adaptive. In sum, our results support local-to-global memory search strategies similar to foraging strategies in space, but further suggest that memory is not constrained to one local representation, but may indeed support many.
\end{abstract}

Keywords: Free recall; verbal fluency task; memory representation; Search of Associative Memory (SAM).

\section*{Introduction}

Memory has long been considered to represent a highdimensional landscape over which we search for information. The recent proliferation of semantic space models, which acquire semantic similarity of words based on statistical processing of text corpora implicitly characterizes memory as embodying such a representation. This idea is not new. In his "Principles of Psychology" William James wrote "We make search in memory ... just as we rummage our house for a lost object" (1890, p. 654), suggesting that search in memory is comparable to search in space. But how comparable is it? Research on spatial imagery and cognitive maps suggests that mental operations share much in common with the way we move around the physical world (e.g. Kosslyn, Ball, and Reiser, 1978). Almost all models of long-term memory incorporate a dimension of similarity (inverse of distance) in order to explain priming and serial position effects (e.g. Anderson, \& Pirolli, 1984; Brown, Neath, \& Chater, 2007). Shepherd's account of
distance in mental representations (Shepard, \& Metzler, 1971), as well as models of categorization (e.g. Nosofsky, 1988), suggests a similar conceptual landscape, in which similar items reside near one another and less similar items reside further apart.

Underlying these approaches is an implicit assumption, one that is highlighted by James. If searching memory is like rummaging our house for a lost object, is there just one house (i.e., representation) or are there many? Could an item be in more than one representation, and if so, does one representation facilitate memory search better than another? If memories reside in multiple representations, this presents a problem for many existing models of knowledge representation-that is, especially those that produce but one representation. Moreover, multiple representations would exemplify a feature of memory that clearly separates memory from space: outside of wormholes, the only way to get from one point in space to another is to travel the distance between them. Memory, on the other hand, may have no such constraints.

Before we describe how we investigated the possibility of multiple memory representations, we first describe some of the previous research that has highlighted the relationship between memory and space.

\section*{Memory and Space}

Following this analogy between space and memory, Hills and colleagues recently put forth a model wherein free recall from memory produces patterns that can be predicted by a classical theorem of optimal foraging theory, the marginal value theorem (Hills, Jones, Todd, 2012). This theory describes optimal switching between explorative and exploitative search in response to a patchy resource environments. Further supporting the notion that internal search may be similar to external search in space, Hills, Todd, and Goldstone (2008) found that priming search in space primes search in a lexical search task, suggesting that a shared cognitive process may search in both domains.

Drawing from the search of associative memory framework (SAM), Hills and colleagues' model expresses search as an alternation between local and global memory search. In local search, sequential items are retrieved based on similarity to the last recalled item together with a position invariant context cue. In global search, this context cue is used exclusively. Interpreting the network of similarities as a landscape of distances, local search is spatially confined with nearer items in memory being more likely to be retrieved. Global search, on the other hand, is independent of this landscape allowing for jumps across the landscape that may utilize an alternative representation-much as wormholes do in science fiction. These aspects of local and global search capture the different search modes typical of exploration/exploitation trade-offs found in patch foraging models. However, they appear to break with the spatial analogy to the degree that the search process can escape the confines of one representation by switching to another.

Recently, Abbott et al. (2012) proposed a slightly different model to explain the findings of Hills and colleagues with a stronger focus on the underlying representation. Their theoretical approach expresses the patterns of free recall as a random walk through a single partly connected graph (see Steyvers, \& Tenenbaum, 2005). Their full random walk model also incorporates local-to-global transitions. However, compared to Hills et al. the switches between subsequent cues a) are random and thus independent of local retrieval success, except in the case of allowing more time for a global jump to occur and, more importantly b) do not imply any change of representation but rather a reset to the start point of search (similar to the executive search process model used to describe search for anagrams in Hills, Todd, \& Goldstone, 2010). Additionally, the simulation results of Abbotts et al. also indicated that apparent optimal retrieval patterns were possible without local-to-global transitions. Together, this work provides only weak support for multiple representations and the presence of context-based local-to-global transitions in memory.

\section*{Present Study}

Overall, the spatial analogy for memory and memory search has been successful, but existing theoretical approaches offer alternative hypotheses. Moreover, all of the previously described approaches have assumed that there is only one representation of memory that allows for local search. To investigate the potential for multiple representations in memory, we had people search their memory for country names. Friedman and Dewinstanley (2007) showed descriptively that at least three independent factors predict country retrieval: geographic distance to the previously recalled country, phonetic similarity to previously recalled countries and characteristics of the particular country itself (e.g. their frequency in the news). The availability of these multiple
objectively determinable cues within a single recall category suggest the potential for multiple representations, and allow us to address the following questions within a computational framework.

First, we want to assess if these three cues are used and how and in what form they are integrated in the retrieval of countries. Is the default local search representation (similarity-based) best characterized by space (e.g., Euclidean distance) or some other representation? Further, are the local search cues integrated dynamically with the global cue (item-based). To address this, we collected uninstructed recall data where people where simply asked to name all the countries they can think of. Our second question addresses more specifically the question of multiple representations: Specifically, how does a voluntary change of retrieval cues influence recall? Assuming a unitary underlying representation, changing the retrieval cues should harm retrieval with respect to response times, number of retrieved items, or both. Provided representations are independently accessible, retrieval sequences should reflect changes in the representation, and may further reflect differences in the accessibility of information with a representation. To this end we ran two instructed conditions in which participants where asked to base their recall on the letters of the alphabet or on geographical neighbors.

\section*{Method}

Participants We collected data from 71 students at the University of Basel. The sample had an average age of 24.7 and \(71 \%\) of the participants were female. Participation in the study was rewarded either by course credit or a fixed payment of 7 Swiss francs. Additionally the participants received 0.25 Swiss francs for every recalled country.

Procedure Participants were seated in front of a computer. First, each participant's typing speed was assessed. Next, they received the instructions to the country fluency task. In the control condition participants were asked to type all the countries they can think of, but were not given direction with regard to how to retrieve countries. In the alphabet condition participants were instructed to proceed by the letters of the alphabet. In the neighbor condition participants were asked to always first attempt to recall a neighboring country (with a shared border) before recalling a country from elsewhere. In both the alphabet and the neighbor condition participants were also instructed that whenever there were unable to recall a country by the first letter or a neighbor they could recall any other country.
Scoring All country entries were checked for spelling and validity. Only the 193 current members of the United Nations as well as Kosovo, Taiwan, Vatican City and Palestine were accepted as valid countries. Some synonyms were accepted, for example "Holland" for Netherlands. As approximations for the spatial
representations (i.e., retrieval structures) we implemented three measures: Distance, calculated as shortest Euclidian distance between the borders; Neighbor, indicating if two countries shared a border; and Geodesic distance, determined by shortest number of border crossing required to move from one country to the other. All information used for the spatial representations were based on the CIA World DataBank II. To measure phonetic representation, three further retrieval structures were implemented: Levenshtein distance, which indicates the number of orthographic edits; Initial letter, indicating if consecutive items share the same first letter; and Phonetic similarity following Friedmann and deWinstanley (2007), with phonetic similarity indicating the same letter in the first or the last three positions of two countries. To estimate the frequency we took two measures: Google, indicating the log number of hits for a country generated by a Google search, and News, indicating the log number of mentions in the weekly newspaper Die Zeit, which is widely read in all German speaking countries.
Modeling The model framework we used to simulate the search process is based on SAM (Raaijmaakers \& Shiffrin, 1981). The foundational assumption of the model is that recall is achieved by probing retrieval structures in memory with a specific cue set, that is, the memory probe. With \(I\) representing a possible target item for recovery in the search space, the probability of retrieving \(I\) is computed as the product of the individual retrieval strengths for \(I\) across a probe set of \(M\) cues, with \(S(Q, I)\) representing the semantic similarity between cue \(Q\) and item \(I\). This is incorporated into an overall probability of retrieval for item \(I\) via the ratio rule:
\[
P\left(I_{i} \mid Q_{1}, Q_{2}, \ldots, Q_{M}\right)=\frac{\prod_{j=1}^{M} S\left(Q_{j}, I_{i}\right)^{\beta_{j}}}{\sum_{k=1}^{N} \prod_{j=1}^{M} S\left(Q_{j}, I_{k}\right)^{\beta_{j}}}
\]
where \(N\) represents the total number of items available in the category for retrieval and \(\beta\) represents the saliency (or attention weight) assigned to a given cue.

Every search cue generates a retrieval strength \(S(Q, I)\) for each item based on the items similarity to the last item, e.g. in terms of Euclidian distance, or the item's own qualities, e.g. frequency in the newspapers or the Eigenvector of similarity-based cues. Using a maximum likelihood method, we fit \(\beta \mathrm{s}\) to each participant's data, using the participant's individually generated sequence of items. This produced a log-likelihood fit, which was penalized based on the number of free parameters via the Bayesian information criterion. Results are presented as the median improvement in the Bayesian information
criterion relative to a random model specifying that all remaining items in the search space are equally likely to be retrieved, with greater values of BIC indicating a better fit.

We examined various static and dynamic models, using spatial, phonetic and frequency cues. In our terminology, static models rely on the same set of cues over the entire retrieval interval. Dynamic models on the other hand allow for a switching between cues. In SAM similarity-based cues are dropped when a threshold of retrieval failures is reached. This we modeled using an additional threshold parameter on the retrieval strengths of similarity-based local cues.

\section*{Results}

\section*{Which cues are used?}

As a first step in the analyses the predictive power of individual cues \({ }^{1}\) was tested in a single cue version of our retrieval model. Figure 1 shows the BIC advantage over the random model for the unconstrained as well as the Alphabet and Neighbor condition. The results indicate that not all retrieval structures are equally predictive in the uninstructed (control) condition. Spatial representations seem clearly to be the dominating cue in this condition. Next to spatial information, only frequency of mentions in the news had predictive power. In contrast to the results of Friedmann and deWinstanley


Figure 1: Median of differences in BIC between the random model and single cue retrieval models.

\footnotetext{
\({ }^{1}\) The eigenvectors of the similarity-based cue representations were dropped from the analyses as none were predictive.
}
(2007) when our participants were not instructed to use any particular cue they do not seem to rely on any phonetic cue. The overall pattern changes substantially in the instructed conditions. When instructed to recall by the letters of the alphabet, the Phonetic and Initial cue models fit very well, whereas the individual spatial models do not exceed chance level. On the other hand, the fit in the Neighbor condition is best fit by the spatial models. Thus, people appear to have changed their retrieval behavior in both conditions.

\section*{Are cues integrated dynamically or statically?}

Contrary to the expectation that all cue classes - spatial, phonetic and frequency - contribute to the fits, the single cue data indicates that only two of three classes of cues are used in the individual conditions. We further asked how the cues are integrated and if the type of integration holds over the conditions and the particular cues used in these conditions. We compared two models: static, with both cue classes are used over the entire retrieval interval, and dynamic \({ }^{2}\), using the similarity-based cue together with the context cue or, when the similarity-based cue falls below a fitted threshold, the context cue alone.

Figure 2 illustrates the results for these models. In line with the single cue models the models combining News with spatial cues fit the data of uninstructed and Neighbor condition best and provide a poor fit for the Alphabet condition. The Alphabet condition was best fit by phonetic information combined with frequency in the news. In regard to the state of integration the models also show a clear pattern. Irrespective of the condition, the dynamic models provide a better fit to the data than the static integration models.

\section*{Does cue use affect performance?}

The analyses thus far show a) that cues can be voluntarily changed and b) that this however has no effect on the dynamic integration of local information with global frequency information. But did the controlled choice of a particular retrieval impact memory accessibility? Figure 3 shows the results for number of countries retrieved. An analysis of variance reveals that the slight advantage in the Neighbor condition is not greater than we would expect by chance ( \(\mathrm{F}_{2,68}=.73, \mathrm{p}=.48\) ). Thus, the overall accessibility in terms of number of countries was not dependent on using a particular cue. Item response times on the other hand reveal a substantial detriment when countries where retrieved by the letters of the alphabet. The median retrieval time in the alphabet condition (mdn

\footnotetext{
\({ }^{2}\) The dynamic search model described in the text corresponds to the search models in Hills, Jones and Todd (2012). An alternative version of the model was also tested, that uses the similarity-based cue when above threshold and the contextbased cue when below threshold. However, the results were indistinguishable.
}
\(=7.7 \mathrm{~s}\) ) was about five times higher than the retrieval times in the uninstructed ( \(\mathrm{mdn}=1.4 \mathrm{~s}\) ) and neighbor condition ( \(\mathrm{mdn}=1.5 \mathrm{~s}\) ). Consequently, participants in the Alphabet condition also keep on searching for a much longer period than in the other two conditions. This is likely due to participants feeling that they could not go back to previous letters. Clearly, however, not all alphabet-based responses are slow. About \(21 \%\) of the response times in the alphabet condition fall below the medians of the other two conditions. Thus in a number of cases Alphabetic retrieval was faster than in the uninstructed and Neighbor condition.

\section*{Discussion}

In this study we were interested in the utilization and integration of multiple cues in retrieval from memory. By having participants retrieve all the countries they know under three different instructions we were able to show that dynamic search models as proposed by Hills and colleagues (Hills, Jones, Todd, 2012; Hills, \& Pachur, 2012) provide the best account for the data in all conditions. Further, the data clearly demonstrated that people are able to deliberately change the cues they are using (see Gronlund \& Shiffrin, 1986). This however had no impact on how these cues were combined with a global representation of frequency. Finally, our data


Figure 2: Median of differences in BIC between the random model and different cue integration models for the three conditions. All similarity-based cues are combined with the News cue. Dashed lines represent the best single cue model in the three datasets.

A


B


Figure 3: Number of items retrieved (A) and bean plots of item level response times in seconds corrected for typing speed (B) in the three condtions. Error bars represent standard error of the mean. Shapes in the right hand plot represent the density, the solid red line the median.
shows that changing the cue is not necessarily harmful to the recall performance - the same performance level was reached in terms of total items retrieved, despite dramatic costs in overall retrieval times.

What do our results mean with respect to our initial question whether memory is one or many representations? Clearly, people are able to change the cues they are using and our results further suggest that they may access alternative representations However, these changes can come with costs. These costs can be interpreted as the result of different distances within a given representation. Thus, a wrong cue or retrieval structure might mean traveling greater or lesser distances in memory. Under this interpretation two speculations can be made. The comparable performance in terms of number of retrieved items would disappear under time pressure and, in the present case, the overall retrieval success is potentially a result of a relaxed retrieval failure threshold. It remains striking that the same overall performance was reached in our three conditions, as recognition data indicates that people have potentially about twice as many countries stored in their memory (Friedman \& deWinstanley, 2007).

Moreover, in every case, models using frequency in the news alone for stretches of the retrieval interval performed much better despite being penalized by the extra threshold parameter. This, there appears to be at least one alternative representation allowing for nonspatial movement in memory space. Combined with evidence for dynamic switching, this breaks with the spatial analogy by allowing for travel through memory via multiple representations.

Both, the successful switch of representations under specific instructions and the independent use of frequency are difficult to explain within the model that is based on a unitary representation or space as proposed by Abbot et al. (2012) and others. Our results seem to be much better explained by frameworks allowing for the variable integration of multiple cues. The SAM-based memory search model developed by Hills and colleagues is but one model affording this possibility. Other models such as multi-trace memory models (e.g. MINERVA; Hintzman, 1986) or the recently proposed context maintenance and retrieval model (CMR; Polyn, Norman, \& Kahana, 2008) are also in principle capable of utilizing multiple cues to varying degrees over time. However, in modeling and most experimental work the possibility of entirely switching between representations has been rather neglected. In our eyes this possibility should receive more attention in future research.

Assuming that our current findings are not constrained to the recruitment of a spatial versus a phonetic or alphabetic local search representation leads to the question of what is the right retrieval strategy to use. Clearly, our data shows that, without instructions, alphabetic and phonetic retrieval strategies receive little support. The data also suggests that this is done for a good reason, as response times tend to be on average larger when using phonetic cues. On the other hand, a substantial number of alphabet retrievals were at least as fast as retrievals based on spatial information. A savvy memory forager could potentially exploit this fact by adaptively switching between retrieval cues - that is, by taking dimensional short-cuts through memory space. In principle, this is no different from the short-cuts made
available by global transitions to frequency. But, cognitively, it represents the capacity to jump between local representations, or landscapes, in much the same way that children might enter an alternate universe by passing through a mirror.

\section*{References}

Anderson, J. R., \& Pirolli, P. L. (1984). Spread of activation. Journal of Experimental Psychology: Learning, Memory, and Cognition, 10, 791-798.
Abbott, J. T., Austerweil, J. L., \& Griffiths, T. L. (2012). Human memory search as a random walk in a semantic network. Advances in Neural Information Processing Systems, 25.
Brown, G. D.A., Neath, I., \& Chater, N. (2007). A Temporal Ratio Model of Memory. Psychological Review, 114, 539-576.
Friedman, W.J., \& deWinstanley, P.A. (2006). The mental representation of countries. Memory, 14, 853871.

Gronlund, S. D., \& Shiffrin, R. M. (1986). Retrieval strategies in recall of natural categories and categorized lists. Journal of Experimental Psychology: Learning, Memory, and Cognition, 12, 550-561.
Hills, T. T., Todd, P. M., \& Goldstone, R. G. (2008). Search in external and internal spaces: Evidence for generalized cognitive search processes. Psychological Science, 19, 676-682.
Hills, T. T., \& Pachur, T. (2012). Dynamic search and working memory in social recall. Journal of Experimental Psychology: Learning, Memory, and Cognition, 38, 218-228.
Hills, T. T., Jones, M.N., \& Todd, P.M. (2012). Optimal foraging in semantic memory. Psychological Review, 119, 431-440.
Hintzman, D. L. (1986). "Schema Abstraction" in a Multiple-Trace Memory Model. Psychological Review, 93, 4, 411-428.
James, W. (1890). The Principles of Psychology. New York: Holt.
Kosslyn, S. M., Ball, T. M., \& Reiser, B. J. (1978). Visual images preserve metric spatial information: Evidence from studies of image scanning. Journal of Experimental Psychology: Human Perception and Performance, 4, 46-60.
Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. Journal of Experimental Psychology: Learning, Memory, and Cognition, 14, 700-708.
Polyn, S. M., Norman, K. N., \& Kahana, M. J. (2008). A context maintenance and retrieval model of organizational processes in free recall. Psychological Review, 116, 1, 129-156.
Raaijmakers, J. G. W., \& Shiffrin, R. M. (1981). Search of associative memory. Psychological Review, 88, 93134.

Shepard, R. N. \& Metzler, J. (1971) Mental rotation of
three-dimensional objects. Science, 171, 701-703.
Steyvers, M., \& Tenenbaum, J. B. (2005). The large scale structure of semantic networks: Statistical analyses and a model of semantic growth. Cognitive Science, 29, 41-78.

\title{
Differences in Interactional Attitudes in Native and Second Language Conversations: Quantitative Analyses of Multimodal Three-Party Corpus
}

\author{
Seiichi Yamamoto (seyamamo@mail.doshisha.ac.jp) \\ Department of Information Systems Design, 1-3 Miyakodani, Tatara, Kyotanabe-shi, Kyoto, 610-0321, Japan \\ Keiko Taguchi (dun0153@mail4.doshisha.ac.jp) \\ Department of Information Systems Design, 1-3 Miyakodani, Tatara, Kyotanabe-shi, Kyoto, 610-0321, Japan \\ Ichiro Umata (umata@nict.go.jp) \\ National Institute of Information and Communications Technology, 2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0288, Japan
}

\author{
Kosuke Kabashima (dt10724@mail4.doshisha.ac.jp) \\ Department of Information Systems Design, 1-3 Miyakodani, Tatara, Kyotanabe-shi, Kyoto, 610-0321, Japan \\ Masafumi Nishida (mnishida@mail.doshisha.ac.jp) \\ Department of Information Systems Design, 1-3 Miyakodani, Tatara, Kyotanabe-shi, Kyoto, 610-0321, Japan
}

\begin{abstract}
Quantitative analyses and the analyses of a questionnaire were conducted to examine the relations between participants' communicative activities and their interactional attitudes in conversations both in their native and second languages. The two categories of conversations revealed different gaze patterns that reflected the differences in difficulties they had with communication and grounding patterns. The participants were less conscious of their own gazes in conversation in their second language than those in their native language probably because of the difficulties and mental pressure they felt.
\end{abstract}

Keywords: Second language conversation; Language expertise; Utterance; Gaze; Grounding; Communication

\section*{Introduction}

As modern society has become more global, the importance of conversations in a second language has been increasing more than ever before. People are traveling around the world either on business or for pleasure due to progress in transportation systems and advanced Internet technologies that connect areas that have different linguistic backgrounds. Organizations are increasingly forming teams with members whose mother tongues are not the same, and sometimes co-workers and collaborators from different countries are connected via the Internet. Second language conversations are commonly observed in daily life, and the expertise of conversational participants often ranges from low to high. An urgent issue today is to support mutual understanding in these conversations.

Language use is a form of joint action that is carried out by groups of people who act in coordination. Their joint action involves not only verbal but also non-verbal activities to achieve a common "grounding" process, i.e., to form the basis of mutual understanding (Clark \& Brennan 1991, Clark 1996). There have been quantitative studies that have reported that eye gazes play an important role in monitoring understanding by communication partners of the content of conversation and contributions made to the performance of collaborative tasks (Boyle, Anderson, \& Newlands 1994, Clark \& Krych 2004).

Grounding is also an important process in second languages. There have been studies that have regarded "nativeness" as "expertise" and compared the grounding process between differing levels of language expertise (Kasper 2004, Hosoda 2006). Hosoda reported that participants' disfluencies or linguistic errors were usually not treated as problems with interactions, but they were oriented to differences in linguistic expertise by repair (a) when one speaker invited the other's repair, and (b) when mutual understanding was jeopardized unless one party repaired the other. Eye gazes and facial expressions play an important role in monitoring both partners' understanding in the repair process. These studies have, however, been qualitative and there have been few quantitative analyses of the relation between the grounding process and non-verbal activities in second language conversations.

Veinott et al. (1999) found that non-native speaker pairs benefited from video in route guiding tasks in the field of computer supported collaborative work (CSCW), whereas native speaker pairs did not. They argued that this was
because video helped the non-native pairs to negotiate a common ground whereas it did not do so for the native pairs. Their study revealed that video images of the conversation partners helped them to establish mutual understanding in their second language conversations, although it was still not clear which element in the video information contributed to establishing the common ground.

Previous research has suggested some differences in conversational features between a mother tongue and a second language. The duration as a percentage when other participants are observing the speaker in English as a second language is longer than that in Japanese as a mother tongue (Kabashima, Nishida, Jokinen, Yamamoto 2012; Yamasaki, Furukawa, Nishida, Jokinen, Yamamoto 2012). Even though these results are consistent with Hosoda's observations and suggest an interesting feature of second language communications, it is still not clear how this feature interacts with other communicative features.
The main aim of this study was to examine the relations between participants' communicative activities and their interactional attitudes in conversations both in their native and second languages. We conducted quantitative analyses to study the differences in communicative behaviors in second and native language conversations. We also analyzed a questionnaire to examine the interactional attitudes of the participants. We integrated the results of the two analyses to examine the differences in communication processes in native and second language conversations.

\section*{Data Collection}

We collected data in a mother tongue and in a second language from conversations by the same interlocutors.

\section*{Participants}

A total of twenty-four university students ( 14 females and 10 males: eight groups) between the ages of 18 and 24 participated in the experiment, and each conversational group consisted of three participants who did not know another. They were Japanese university students who had acquired Japanese as their mother tongue and had learned English as a second language. Their communication levels in English were measured based on the Test of English for International Communication (TOEIC). We recruited participants to cover wide range of expertise in English. Their scores ranged from 450 to 890 , and the average was 591 ( 990 being the highest score that could be attained). Each participant was ranked into three degrees of expertise according to the order of his/her English expertise based on the TOEIC score within the group.

\section*{Experimental Setup}

Three participants sat in a triangular formation around one table. Each participant sat in the same position for all four trials. The three participants sat 1.5 m apart. Three sets of NAC EMR-9 eye trackers and headsets with microphones were used to record the eye gazes and voices of all three participants (Figure 1). The viewing angle of the EMR-9
was \(62^{\circ}\) and the sampling rate was 60 fps . The participants talked about two types of predetermined themes in English as a second language and in Japanese as their mother tongue (e.g., each group participated in four conversations).


Figure 1: Experimental Setup

\section*{Procedure}

The conversational topic was assigned before each trial and there were two types of themes. The first was freeflowing in which they chatted naturally on foods they liked or disliked. The second was goal-oriented in which they collaboratively decided what to take with them on trips to uninhabited islands or mountains. We randomly arranged the order of the topics of conversation to cancel out the effect of order. We also randomly arranged the order of the languages used in the conversations.

The eye trackers were calibrated and participants started to converse after instructions on the experiment were explained. Each group had conversations of four different topics on free-flowing and goal-oriented themes in Japanese and in English. Each conversation lasted for 6 min .

The participants filled in a questionnaire after each conversation. Consequently, the subjects in each group participated in four conversations and filled out four questionnaires. We then analyzed the data from the free flowing conversations in Japanese and in English. The total numbers of utterances were 1858 in English and 2059 in Japanese, and gaze events were 2360 in English and 2727 in Japanese, respectively.

\section*{Annotations}

One of the authors manually annotated the time spans for utterances and gazes at other participants to integrate the utterance and eye gaze data. The eye gaze data of two participants were not recorded because of trouble they had with the equipment, and they were excluded from the analyses. We used the EUDICO Linguistic Annotator (ELAN) developed by the Max Planck Institute (Fig.2) as the annotation tool.

\section*{Questionnaire}

We asked the participants to fill in the questionnaire to analyze their interactional attitudes in each conversation. The questionnaire consisted of 29 items and each of them was ranked on a Likert scale from one to seven. Each
question was categorized into communicational features such as participants' gazing activities, their feelings toward other participants, their interest in the topic of conversation, their conversational skills in English, and their evaluation of the conversation content. The progress of the trials can affect the familiarity. We conducted ANOVA on the values obtained from questions on evaluating their familiarity to other participants to check if there are any differences in the values among the trials. The results did not revealed significant differences on the evaluation of familiarity with the progress of trials. The results suggest that, in this experiment, familiarity to other participants were not affected by the progress of the trial.


Figure 2: Annotation Screen Shot

\section*{Analyses}

We used two methods in the analyses; the first involved quantitative analysis of communicative activities such as utterances and gazes, and the second involved analyzing a questionnaire on the participants' interactional attitudes. The correlations between the quantitative data and the values obtained from evaluating the questionnaire were also analyzed to study the relations between the participants' communicative activities and their interactional attitudes in conversations both in their native and second language.

\section*{Analysis 1: Utterances and Gazes}

First, we compared the total duration of the utterances between the Japanese and English language conversations to check the difficulty in communicating in the second language. The total duration of the utterances was expected to be longer in Japanese than in English. A paired \(t\)-test indicated a significant difference between conversations in Japanese as a native language and those in English as a second language ( \(\mathrm{t}=4.848, \mathrm{p}<.01\), Japanese: \(a v .=110222\) \(\mathrm{msec}, S D=39178 \mathrm{msec}\); English \(a v .=79185 \mathrm{msec}, S D=\) 6347). They talked more in conversations in their native language than in those in their second language.

Gazing activities during utterances were also compared between conversations in Japanese and those in English. We compared (1) how long the speaker was observed by other participants (ratio being observed), and (2) how long the speaker observed other participants (ratio observing).

The average for the ratio being observed is defined as:

Average of Ratio Being Observed \(=\frac{\sum_{i=1}^{n} D P O S(i)}{\sum_{i=1}^{n} D(i)} \times 100(\%)\)
Here, \(D(i)\) is the duration of the \(i\)-th utterance and \(\operatorname{DPOS}(i)\) is the duration when other participants are observing the speaker in the \(i\)-th utterance.

The average for ratio observing is defined as:
\[
\text { Average of Ratio Observing }=\frac{\sum_{i=1}^{n} D S O P(i)}{\sum_{i=1}^{n} D(i)} \times 100(\%)
\]

Here, \(\operatorname{DSOP}(i)\) is the duration when the speaker is observing other participants in the \(i\)-th utterance.

Previous research has suggested that the speaker is observed more in their second language conversations than in those in their native language, and that the difference in second language expertise affects the communication style (Kabashima, Nishida, Jokinen, Yamamoto 2012; Yamasaki, Furukawa, Nishida, Jokinen, \& Yamamoto 2012). These predictions are consistent with observations of the repair process in second language conversations (Hosoda 2006).

Under the hypotheses that the speakers were observed more in conversations in their second language than in their native language and that language expertise affected gazing activities, we conducted \(2 \times 3\) ANOVA with the language difference being within subject factors and with expertise rank in English being between subject factors. The results revealed significant main effect both on language differences \(\left(\mathrm{F}_{(1,19)}=24.823, \mathrm{p}<.01\right)\) and on the expertise ranking in the second language ( \(\mathrm{F}_{(2,19)}=3.625, \mathrm{p}<.05\) ), and no interactions were observed. Fisher's LSD test indicated significant differences between the 1st and 3rd ranks ( \(\mathrm{p}<.05\) ).

The average and the SD values of the ratio being observed for each rank are listed in the table below.
\begin{tabular}{|l|l|r|r|r|}
\hline Expertise Rank & Condition & \multicolumn{1}{l|}{ av. } & No. of the samples & \multicolumn{1}{l|}{ SD } \\
\hline 1 & Native Language & 73.33 & 7 participants & 11.299 \\
\hline & Second language & 82.77 & 7 participants & 3.946 \\
\hline 2 & Native Language & 78.77 & 7 participants & 8.257 \\
\hline & Second language & 85.09 & 7 participants & 1.764 \\
\hline 3 & Native Language & 81.29 & 8 participants & 2.663 \\
\hline & Second language & 88.48 & 8 participants & 3.151 \\
\hline
\end{tabular}

Table 1: Average values for ratio being observed
The results indicate that the speakers were observed more in conversations in their second language than in those in their native language, and speakers with low levels of expertise in their second language tended to be observed more than speakers with high levels of expertise.

\section*{Analysis 2: Comparison of Values from Questionnaire}

The values from evaluating each item on the questionnaire were compared for Japanese and English to examine the differences in the participants' interactional attitudes in conversations in their native and second languages. The results suggest they had greater difficulty
and were more stressed in their second language conversations as we had predicted. The items that exhibited significant differences in the t-test \({ }^{1}\) were in the following table. Here a single asterisk * in the tables denotes \(p<.05\) and a double asterisk \({ }^{* *}\) denotes \(p<.01\), respectively.

\section*{Evaluation of Expertise:}

The participants evaluated their expertise and their partners' expertise higher in their native language conversations as follows.
- Evaluate your speaking expertise.
\begin{tabular}{|l|r|c|c|c|c|}
\hline & \(N\) & \(a v\). & \(S D\) & \(d f\) & \(t\) \\
\hline Japanese & 24 & 4.63 & 1.245 & 23 & \multirow{2}{*}{\(6.323^{* *}\)} \\
\hline English & 24 & 2.67 & 1.404 & 23 & \\
\hline
\end{tabular}
- Evaluate your partner's English speaking expertise.
- Toward higher ranked partners
\begin{tabular}{|c|c|c|c|c|c|}
\hline & \(N\) & \(a v\). & \(S D\) & \(d f\) & \(t\) \\
\hline Japanese & 24 & 5.63 & 0.97 & 23 & \multirow{2}{*}{\(3.760^{* *}\)} \\
\hline English & 24 & 4.54 & 1.615 & 23 & \\
\hline
\end{tabular}

\section*{- Toward lower ranked partners}
\begin{tabular}{|c|c|c|c|c|c|}
\hline & \(N\) & \(a v\). & \(S D\) & \(d f\) & \(t\) \\
\hline Japanese & 24 & 5.29 & 1.083 & 23 & \multirow{2}{*}{\(3.680^{* *}\)} \\
\hline English & 24 & 4.21 & 1.532 & 23 & \\
\hline
\end{tabular}
- Do you think he/she could understand your discourse?
- Toward higher ranked partners
\begin{tabular}{|c|c|c|c|c|c|}
\hline & \(N\) & \(a v\). & \(S D\) & \(d f\) & \(t\) \\
\hline Japanese & 24 & 5.67 & 1.239 & 23 & \multirow{2}{*}{\(2.220^{* *}\)} \\
\hline English & 24 & 5.17 & 1.308 & 23 & \\
\hline
\end{tabular}
- Toward lower ranked partners
\begin{tabular}{|c|c|c|c|c|c|}
\hline & \(N\) & \(a v\). & \(S D\) & \(d f\) & \(t\) \\
\hline Japanese & 24 & 5.83 & 1.239 & 23 & \multirow{2}{*}{\(3.423^{* *}\)} \\
\hline English & 24 & 5.08 & 1.176 & 23 & \\
\hline
\end{tabular}

\section*{Feelings toward Partners:}

The participants were more nervous and felt more pressure from their partners in their second language conversations. They felt their partners concentrated more in second language conversations. These results suggest the participant felt more stress in second language conversations.
- Did you get nervous when you spoke?
- Did you get nervous when you spoke?
\begin{tabular}{|c|c|c|c|c|c|}
\hline & \(N\) & \(a v\). & \(S D\) & \(d f\) & \(t\) \\
\hline Japanese & 24 & 2.50 & 1.504 & 23 & \multirow{2}{*}{\(5.249^{* *}\)} \\
\hline English & 24 & 4.33 & 1.606 & 23 &
\end{tabular}
- Do you think your partner got nervous when he/she spoke?
- Toward higher ranked partners

\footnotetext{
\({ }^{1}\) There were only 24 participants and it is not clear if these values were normally distributed. However, exactly the same list of items also revealed significant differences in Wilcoxon's signed rank test, and the results can be considered to be sufficiently stable.
}
\begin{tabular}{|c|c|c|c|c|c|}
\hline & \(N\) & \(a v\). & \(S D\) & \(d f\) & \(t\) \\
\hline Japanese & 24 & 2.38 & 1.408 & 23 & \multirow{2}{*}{\(5.675^{* *}\)} \\
\hline English & 24 & 4.13 & 1.296 & 23 & \\
\hline
\end{tabular}
- Toward lower ranked partners
\begin{tabular}{|c|c|c|c|c|c|}
\hline & \(N\) & \(a v\). & \(S D\) & \(d f\) & \(t\) \\
\hline Japanese & 24 & 2.58 & 1.381 & 23 & \multirow{2}{*}{\(-5.625^{* *}\)} \\
\hline English & 24 & 4.54 & 1.381 & 23 & \\
\hline
\end{tabular}
- Did you feel pressure from them?
- Toward higher ranked partners
\begin{tabular}{|c|c|c|c|c|c|}
\hline & \(N\) & \(a v\). & \(S D\) & \(d f\) & \(t\) \\
\hline Japanese & 24 & 1.92 & 1.018 & 23 & \multirow{2}{*}{\(-2.230^{*}\)} \\
\hline English & 24 & 2.58 & 1.742 & 23 & \\
\hline
\end{tabular}
- Toward lower ranked partners
\begin{tabular}{|c|c|c|c|c|c|}
\hline & \(N\) & \(a v\). & \(S D\) & \(d f\) & \(t\) \\
\hline Japanese & 24 & 1.88 & 1.191 & 23 & \multirow{2}{*}{\(-3.093^{* *}\)} \\
\hline English & 24 & 2.58 & 1.640 & 23 & \\
\hline
\end{tabular}
- Do you think your partner concentrated on your discourse? - Toward higher ranked partners
\begin{tabular}{|c|c|c|c|c|c|}
\hline & \(N\) & \(a v\). & \(S D\) & \(d f\) & \(t\) \\
\hline Japanese & 24 & 5.25 & .989 & 23 & \multirow{2}{*}{\(-3.391^{* *}\)} \\
\hline English & 24 & 5.92 & .881 & 23 & \\
\hline
\end{tabular}
- Toward lower ranked partners
\begin{tabular}{|c|c|c|c|c|c|}
\hline & \(N\) & \(a v\). & \(S D\) & \(d f\) & \(t\) \\
\hline Japanese & 24 & 5.29 & 1.233 & 23 & \multirow{2}{*}{\(-2.077^{* *}\)} \\
\hline English & 24 & 5.79 & .884 & 23 & \\
\hline
\end{tabular}

\section*{Evaluations of Conversation}

The participants felt that they were more active and that the conversation warmed up and became more enjoyable in their native language conversations. They also felt that they could talk as they usually did in their native language.
- Do you think you could talk actively?
- Do you think you could talk actively?
\begin{tabular}{|c|c|c|c|c|c|}
\hline & \(N\) & \(a v\). & \(S D\) & \(d f\) & \(t\) \\
\hline Japanese & 24 & 5.08 & 1.139 & 23 & \multirow{2}{*}{\(3.709^{* *}\)} \\
\hline English & 24 & 3.75 & 1.452 & 23 & \\
\hline
\end{tabular}
- Did the conversation warm up?
\begin{tabular}{|c|c|c|c|c|c|}
\hline & \(N\) & \(a v\). & \(S D\) & \(d f\) & \(t\) \\
\hline Japanese & 24 & 5.50 & .978 & 23 & \multirow{2}{*}{\(4.331^{* *}\)} \\
\hline English & 24 & 4.13 & 1.191 & 23 & \\
\hline
\end{tabular}
- Did you enjoy the conversation?
\begin{tabular}{|c|c|c|c|c|c|}
\hline & \(N\) & \(a v\). & \(S D\) & \(d f\) & \(t\) \\
\hline Japanese & 24 & 5.33 & .963 & 23 & \multirow{2}{*}{\(2.077^{*}\)} \\
\hline English & 24 & 4.83 & 1.167 & 23 & \\
\hline
\end{tabular}
- Did you think that you could talk as you usually do?
\begin{tabular}{|c|c|c|c|c|c|}
\hline & \(N\) & \(a v\). & \(S D\) & \(d f\) & \(t\) \\
\hline Japanese & 24 & 5.46 & 1.179 & 23 & \multirow{2}{*}{\(5.438^{* *}\)} \\
\hline English & 24 & 3.21 & 1.641 & 23 & \\
\hline
\end{tabular}

\section*{Analysis 3: Correlations between Gazing Activities and Values from Questionnaire}

These results indicate that the participants gazing activities and interactional attitudes differed in conversations in their native and second languages, as had been predicted. We conducted Spearman's correlation analysis on their gazing activities and their interactional attitudes that were contained in the questionnaire data. The items that exhibited significant correlation are shown with the correlation values (Spearman's \(\rho\) ) in the following tables. A single asterisk * denotes \(p<.05\) and a double asterisk denotes \(p<.01\) on the tables.

\section*{Consciousness of Gazing Activities:}

The values obtained from evaluating gazing activities had high correlations with gaze durations in Japanese conversations although they did not in English conversations. These results indicate that the participants were conscious of their gazing activities in conversations in their native language, whereas they were not in their second language.

Japanese:
\begin{tabular}{|l|c|}
\hline Ratio Observing & \(\rho\) \\
\hline \hline <-> Did you watch his/her face as a whole? & \(.511^{*}\) \\
\hline <-> Did you watch his/her eyes? & \(.588^{* *}\) \\
\hline
\end{tabular}

\section*{Analysis 4: Correlations of Items on Questionnaire}

There were several interesting differences in the results from Spearman's correlation analysis of the items in the questionnaire for the Japanese and English conversations as listed below.

\section*{Difference in Interactional Attitudes}

The speakers' evaluations of their ability to concentrate were correlated with their evaluations of gazing at the listeners' upper body, face, and eyes in English, but only with their evaluations of gazing at the listener's eyes in Japanese. This suggest that speakers were paying attention to wider areas of their partners' body when they concentrated during second language conversations, whereas they were only paying attention to the eyes of their partners in native language conversations. .

English:
\begin{tabular}{|l|r|}
\hline Did you concentrate on your utterances? & \(\rho\) \\
\hline \hline \begin{tabular}{l} 
<-> Did you watch the listener's upper body as \\
a whole?
\end{tabular} & \(.485^{*}\) \\
\hline <-> Did you watch his/her face as a whole? & \(.537^{* *}\) \\
\hline <-> Did you watch his/her eyes? & \(.605^{* *}\) \\
\hline
\end{tabular}

Japanese:
\begin{tabular}{|l|c|}
\hline Did you concentrate on your utterances? & \(\rho\) \\
\hline \hline\(->\) Did you watch his/her eyes? & \(.417^{*}\) \\
\hline
\end{tabular}

\section*{Feelings toward Other Participants}

The participants evaluations of their understanding of their partners' discourse were correlated with those of their positive feelings toward their partners in English conversations, whereas there were no such correlations in Japanese conversations. The participants tended to have positive feelings toward their partners when they could understand what their partners said in second language conversations, but just understanding their partners' discourse was not enough for the participants to have positive feelings toward their partners in native language conversations.

English:
\begin{tabular}{|l|l|c|}
\hline \begin{tabular}{l} 
Do you think you could understand his/her \\
discourse?
\end{tabular} & \(\rho\) \\
\hline \hline \begin{tabular}{l} 
<-> Did you have a sense \\
of closeness to your \\
partner?
\end{tabular} & \begin{tabular}{l} 
Toward higher \\
ranked partners:
\end{tabular} & \(.639^{* *}\) \\
\cline { 2 - 3 } & \begin{tabular}{l} 
Toward lower \\
ranked partners:
\end{tabular} & \(.549^{* *}\) \\
\hline \multirow{3}{*}{\begin{tabular}{l} 
<-> Did you become \\
interested in him/her?
\end{tabular}} & \begin{tabular}{l} 
Toward higher \\
ranked partners:
\end{tabular} & \(.523^{* *}\) \\
\cline { 2 - 3 } & \begin{tabular}{l} 
Toward lower \\
ranked partners:
\end{tabular} & \(.532^{* *}\) \\
\hline
\end{tabular}

The participants' evaluations of pressure from their partners had a correlation with their evaluations of their own nervousness when they spoke in English conversations but not in Japanese conversations. This suggested that pressure from their partners led directly to the speakers' nervousness in second language conversations, but not in native language conversations.

English:
\begin{tabular}{|l|l|c|}
\hline \multicolumn{2}{|c|}{ Did you feel pressure from them? } & \(\rho\) \\
\hline \hline \multirow{3}{*}{\begin{tabular}{l} 
<-> Did you get nervous \\
when you spoke?
\end{tabular}} & \begin{tabular}{l} 
Toward higher \\
ranked partners:
\end{tabular} & \(.419^{*}\) \\
\cline { 2 - 3 } & \begin{tabular}{l} 
Toward lower \\
ranked partners:
\end{tabular} & \(.460^{*}\) \\
\hline
\end{tabular}

\section*{Discussion}

Thus far, we have compared the utterances, gazes, and interactional attitudes of participants in native and second language conversations. Quantitative analyses were conducted on utterance and gaze data in Analysis 1. The shorter total duration in English conversations suggested difficulties the participants had in their second language conversations. Preliminary analysis using one-fourth of this corpus denoted that the average number of filled pauses and percentage of turn-hold after pause were more than double in English in comparison with those in Japanese (Yamasaki, Furukawa, Nishida, Jokinen, Yamamoto 2012). These results also suggested difficulties the participants had in their second language conversations.

The speakers were observed by listeners more in their second language conversations than in their native language
conversations. The speakers with lower levels of linguistic expertise were gazed more than those with higher levels of linguistic expertise in their second language conversations. These results are consistent with observations by Hosoda (Hosoda 2006), and they indicate that such gazing patterns represent one of the interactional features unique to second language conversations. The listeners made more use of visual information from the speaker to help further understanding in conversations in their second language than that in their native language, and it is likely that speakers with low levels of expertise need more gazes from their partner to help their repair process in grounding activities in second language conversations.
A comparison of the values obtained from the questionnaire also revealed difficulties the participants had in their second language conversations in Analysis 2. They evaluated their linguistic expertise to be lower in their second language conversations and they felt more pressure from their partners, and were more nervous. They were not able to conduct conversations as they usually did, and the conversations did not warm up as much as those in their native language.

The difficulties in second language conversations seemed to have affected their management of conversational activities. The results obtained from correlation analysis in Analysis 3 of the participants' gazes and their self evaluations of their own gazes indicated that the participants were not conscious of their gazing activities in conversations in their second language, whereas they were in their native language. This suggests that difficulties in second language communication made the participants concentrate too much on managing conversations to be conscious of their own communicative activities.

Analysis of correlation in the items on the questionnaire in Analysis 4 revealed differing interactional attitudes in native and second language conversations. The speakers seemed to make use of visual cues from wider areas of the listeners' upper bodies when they concentrated more on their second language conversations than those in their native language where they only made use of visual cues from the listeners' eyes.
Another interesting finding from Analysis 4 was that understanding what a conversation partner said was likely to lead to positive evaluation of the partner in conversations in the second language whereas no such tendencies were observed in conversations in the native language. This suggests that understanding the partners' utterances is already considered to be an achievement in second language conversations whereas just understanding what the partners say is not enough to have positive feelings toward them.

\section*{Conclusion}

We examined the relations between participants' communicative activities and their interactional attitudes both in native and second language conversations. Quantitative analyses and analyses of a questionnaire revealed that the participants had more difficulties in their
second language conversations than those in their native language, and they demonstrated different interactional attitudes in the two categories of conversations. .

Speakers were observed more by listeners in conversations in their second language than those in their native language, and speakers with lower levels of expertise were observed more in second language conversations which probably reflected more frequent repair processes.

The participants were less conscious of their gazing activities in conversations in their second language than those in their native language probably because of the difficulties and pressures they felt in their second language conversations. We trust these findings will contribute further to supporting second language communications and second language learning.

\section*{Acknowledgments}

This research was supported in part by a grant from the Japan Society for the Promotion of Science (JSPS) (No.22520598). The authors would like to thank Professor Kristiina Jokinen of the University of Helsinki and Professor Mariko Sugawara of Doshisha University for their suggestions and the various discussions we had with them.

\section*{References}

Boyle, E., Anderson, A., \& Newlands, A. (1994). The Effect of Visibility on Dialogue and Performance in a Cooperative Problem Solving Task, Language and Speech Journal, pp. 1-20.
Clark, H. H. \& Brennan, S. A. (1991). Grounding in Communication, in: L. B. Resnick, J. M. Levine, and S. D. Teasley (Eds.), Perspectives on Socially Shared Cognition (pp. 222-233), APA Books, Washington.
Clark, H. H. (1996). Using Language, Cambridge University Press, Cambridge..
Clark, H. H. \& Krych, M. A. (2004). Speaking While Monitoring Addressees for Understanding, Journal of Memory and Language, p. 6281.
Hosoda, Y. (2006). Repair and Relevance of Differential Language Expertise, Applied Linguistics, 27, pp. 25-20.
Kabashima, K., Nishida, M., Jokinen, K. \& Yamamoto, S. (2012). Multimodal Corpus of Conversations in Mother Tongue and Second Language by Same Interlocutors, Proceedings of \(4^{\text {th }}\) Workshop on Eye Gaze in Intelligent Human Machine Interaction.
Kasper, G. (2004). Participant Orientations in Conversation-for-Learning, The Modern Language Journal, vol. 88, pp. 551-567.
Veinott, E., Olson, J., Olson, G., and Fu, X. (1999). Video helps Remote Work: Speakers Who Need to Negotiate Common Ground Benefit from Seeing Each Other, Proceedings of the Conference on Computer Human Interaction. CHI'99, ACM Press, PA, USA, pp. 302-309.
Yamasaki, S., Furukawa, H., Nishida, M., Jokinen, K., and Yamamoto, S. (2012). Multimodal Corpus of Multi-party Conversations in Second Language, Proceedings of LREC2012.

\title{
Implicit Learning: A Demonstration and a Revision of a Novel SRT Paradigm
}

\title{
Fayme Yeates \({ }^{1}\) (fy212@exeter.ac.uk)
}

\author{
F.W. Jones \({ }^{2} \quad\) Andy J. Wills \({ }^{3} \quad\) M.R.F. Aitken \({ }^{4} \quad\) I.P.L. McLaren \({ }^{1}\) \\ \({ }^{1}\) School of Psychology, College of Life and Environmental Sciences, University of Exeter, UK. \\ \({ }^{2}\) School of Psychology, Canterbury Christ Church University \\ \({ }^{3}\) School of Psychology, University of Plymouth \\ \({ }^{4}\) School of Psychology, University of Cambridge \\ 
}

\begin{abstract}
Yeates, Jones, Wills, Aitken, McLaren and McLaren (2012) devised a serial reaction time (SRT) task that provided evidence for human learning without awareness. Adapting the SRT paradigm usually employed to investigate implicit learning, participants responded to two simple white circle fills on either side of a screen. Instead of these following a sequence that participants were unaware of (e.g. Willingham, Nissen \& Bullemer, 1989) this task involved a separate stimulus, which was sometimes predictive of one of the circle fills. A square in the center of the screen would fill with one of eight colors before each circle fill: one of these colors predicted a right circle fill and the other a left on \(80 \%\) of trials on which those colors occurred. When pressing the key that followed the consistent response trained with these two colors, participants were both faster and more accurate than when responding to either the inconsistent response or control colors. Participants demonstrated a lack of contingency awareness, performing at chance in identifying the predictive colors and on a suitably sensitive prediction task. On reanalyzing this result, this paper shows that it was confounded with a sequential artifact produced by the experimental design itself. Pilot studies demonstrated weak learning of color contingencies when the artifact was removed, thus we sought to improve learning by both increasing the amount of training and placing the predictive color cue on the circle fills. Without the sequential artifact, we can produce the same result, although we concede the effect is less robust than we first indicated. Thus, we are able to reiterate our original conclusion: that this task can demonstrate learning of color contingencies in the absence of awareness and can be used to investigate implicit learning in humans.
\end{abstract}

Keywords: Associative learning; implicit learning; SRT task

\section*{Introduction}

At the \(34^{\text {th }}\) Annual Conference of the Cognitive Science Society Yeates, Jones, Wills, Aitken, McLaren and McLaren (2012) presented a novel serial reaction time (SRT) task, arguing that it produced convincing evidence for implicit learning in humans. The current paper tempers these claims, by first pointing to a subtle artifact in the experimental design, and then running experiments in which this artifact has been removed.

The criticisms leveled at research exploring implicit learning are extensive and well documented (e.g. Lovibond and Shanks, 2002; Mitchell, De Houwer and Lovibond, 2009; Shanks and Lovibond, 2002; Shanks \& St. John, 1994). One enduring research paradigm, however, that
remains popular is the SRT task. These studies typically require participants to perform a task in which they respond quickly and accurately to stimuli presented to them in a fastpaced series. In the version developed to investigate implicit learning by Willingham, Nissen and Bullemer (1989), unknown to participants these stimuli are presented in a particular sequence. Faster performance on these sequences, compared to participants who had been trained on random control sequences, provided Willingham et al. (1989) with evidence of learning in the absence of the ability to verbally report or explicitly predict those sequences.

Yeates et al. (2012) aimed to devise a paradigm with which one could both demonstrate implicit learning and investigate implicit processes. Reasoning from a dualprocess account of human learning, with both Cognitive (conscious, controlled, rule-based and symbolic) and Associative (automatic, statistical) systems (McLaren, Green and Mackintosh, 1994) assumed to be available, led Yeates et al. (2012) to develop an experimental design that attempted to circumnavigate rule-based, conscious processing of the stimuli. The intention of the study was to provide an experimental setting in which associative processing would be encouraged to underpin learning of relationships present in the SRT task.

To this end, a two-choice SRT task based on Jones and McLaren's (2009) and Aitken's (1996) previous work was devised. Participants were required to respond with two different, spatially compatible key presses to a white circle fill: either on the left or right hand side of the screen. On each trial, prior to the circle fill, a square (outlined in white in the center of the screen) would fill with one of eight colors; which participants were told functioned as a simple, central fixation to optimize their performance and avoid bias to either of the circle locations. They were therefore instructed to attend to the square but not told of its true value, which was (on certain trials) as a predictor of which circle would fill. Hence, this SRT task did not train participants to predict their next response from the sequence of previous responses; it used a separate stimulus to predict at which location the response stimuli would next occur.

A within-subject control was employed, so that only two of the eight possible colors correlated with one of the response stimuli locations. The other six colors occurred with equal likelihood before a right or left circle fill and therefore bore no predictive relation to the response
participants were making. The two predictive colors were themselves only partially ( \(80 \%\), following Posner \& Synder, 1975) predictive of a right or left circle fill. Hence one color would predict a right circle fill \(80 \%\) of the time and a left circle fill \(20 \%\) of the time. The other predictive color would precede a left circle fill \(80 \%\) and a right circle fill \(20 \%\) of the time. The prediction rate over the experiment works out at \(57.5 \%\), thus conscious detection of the presence of contingencies within the experiment would be very difficult. In conjunction with the rapid pace of the task, which involved short inter-trial intervals (ITIs: 250 msecs between response and square color fill, 250-500 msecs between square color fill and circle fill) and responses from participants ( \(M=298.4\) msecs, \(S D=27.7\) ), it is a design that does not encourage nor benefit participants to try and "work it out".

Two of the non-predictive colors were presented alongside the predictive colors in experimental blocks, which made up half of the 20 total blocks in the experiment and were alternated with control blocks (containing the remaining four control colors), the order of which was counterbalanced across participants. To avoid issues due to the sequence of lefts and rights, we designed the experiment so that each control block comprised the same sequence of circle fills as the experimental block it preceded/followed (dependent on counterbalancing). Thus, when comparing the difference between experimental predictive color performance with control performance, we could be confident this was not the product of the sequence of responses performed.

The final design feature of the paradigm employed to encourage participants away from attempting to consciously discover underlying relationships between the stimuli was to prohibit repetitions of the same color on consecutive trials. If participants were exposed to random sequences, we hypothesized that consecutive trials that involved repetitions of the same, predictive color would increase the salience of that color being particularly related to one circle fill, and thus one response. Thus, the experience of randomly being presented with a string such as: red-right-red-right-redright...etc. was prohibited.

This, however, introduced the artifact this paper is concerned with, as the consequence of introducing such a restriction on the trial sequences increased the number of alternations between right and left responses and decreased the number of repeating response trials in our experiment. For example, in an experimental block if you have just received the color that predicts a right response, you have a four in five chance of a right circle fill and thus a right response. Following this trial, on the next trial you can only be presented with: one of the two non-predictive colors (which are equally likely to be a right or a left); or the color that predicts left ( \(80 \%\) of the time). Therefore, you have a (roughly) two in five chance of another right trial and a three in five chance of a left trial. The confound occurs in that such alternations are more likely to occur on predictive trials that follow the contingency within the experiment, and least likely to occur on the \(20 \%\) of predictive color trials that
don't follow the trained contingency, and are equally likely on control trials. This, rather worryingly, neatly explains our original findings, if we assume that people either naturally prefer to alternate responses, or learn to do so. The result would be better performance on consistent predictive color trials, worse performance on the inconsistent predictive color trials, and intermediate performance on control color trials.

We sought to investigate this possible confound, with both a re-analysis of the original data and further experiments to ascertain the extent to which our previous claims - that we had demonstrated implicit learning using a novel, neat and robust paradigm - would survive when removing this potential artifact. The exact nature of the sequential artifact itself is interesting as, if indeed the observed results of the original experiment were concerned not with the relationship between color and circle but the statistical regularity of alternations versus repeats, was this learnt or is it simply a behavioral preference?

\section*{Original Experiment}

The full details of the experiment can be found in Yeates et al. (2012). A brief description of the method follows here, with the further analyses run on the original data, which corrects the original analysis by including a comparison between control and experimental blocks to investigate sequential effects.

\section*{Method}

Participants. The study involved 32 participants from the University of Exeter who each performed a two-choice serial reaction time over one session lasting roughly an hour.

Materials. The experiment involved the on-screen presentation of two white circle outlines and a white square outline, all 1.9 cm in width. The square was presented in the center of the screen, with the circles 2.2 cm either side to the right and left. The stimuli were one of eight possible colors: red, green, blue, yellow, pink, orange, brown and teal; that appeared within the square outline. The circle outlines would only fill white.

Design. Half of the colors were presented in experimental blocks and the other half in control blocks. There were 10 of each type of block, which alternated throughout the experiment and comprised of 120 trials each. In each block, each of the four colors were equally likely to occur. In control blocks, half of the time a color would precede a right circle fill and half the time a left circle fill. In experimental blocks two of the colors acted as controls, with the same number of right and left circle fills after these two colors. One of the two predictive colors in an experimental block preceded a right circle fill on 24 out of 30 trials, with the other color preceding a left circle fill on \(80 \%\) of trials. Therefore we classified trials as: Predictive-Consistent (the 24 of 30 trials that followed the predictive relationship); Predictive-Inconsistent (the 6 of 30 trials where the circle fill following a color was not the target trained circle fill);

Experimental Non-Predictive (control color trials in experimental blocks); and Control Non-Predictive (control color trials in control blocks). The same color could not occur on consecutive trials. All blocks involved an equal number of right and left circle fills, and control block right and left circle fill sequences followed the same sequence of right and left circle fills as the experimental block adjacent to it (either preceding or following depending on the counterbalancing).

Procedure. On each trial the square would fill with one of eight possible colors and, after a variable interval of between 250 and 500 msecs one of the two circles would fill in white. This was the cue for participants to respond with spatially compatible keys of either " \(x\) " or " \(>\) " on a standard QWERTY keyboard for the left and right circle, respectively. A 250 msec ITI followed, during which the circle and square outlines were again presented on screen. Errors were signaled with a beep and each block was followed by a 30 sec break.

Participants were instructed to fixate on the colored square to avoid a bias toward either of the circle flashes, and were told that the experiment was concerned with responding quickly to simple stimuli. No mention was made of the predictive nature of the colors, or of any relationships in the experiment to learn about. A verbal interview and prediction task followed the experiment. The structured interview aimed to assess knowledge of the experimental contingencies and asked participants to describe anything that they had noticed and to identify two colors that may have been predictive. The prediction task involved the same stimuli as in the previous experimental and control blocks, with two blocks of 16 trials each - one with experimental and one with control colors. These colors were randomly presented an equal amount of times to participants within the square in the center of the screen. Instead of this stimulus preceding a circle fill that prompted a response, the display remained the same until participants made a prediction about where they thought the circle would have filled in the experiment using the same response keys (" \(x\) " or " \(>\) "). Participants were informed that pressing either of the response keys would not be considered an error and no feedback was given.

\section*{Results}

In the original paper, Yeates et al. (2012) analyzed both reaction times (RTs) and error rates across the four Trial Types mentioned previously. An analysis of variance (ANOVA) was conducted comparing Trial Types across Blocks. We found a significant effect of Trial Type in both RTs and errors, both following the same ordinal pattern with slower and less accurate responding to PredictiveInconsistent, followed by Experimental Non-Predictive and Control Non-Predictive colors, with Predictive-Consistent colors resulting in faster and more accurate responding.

To ascertain whether these results were due to learning of the contingencies present between predictive colors and responding across the experiment, here we report the results of a corrected ANOVA with Block Type as a two level
within-subject variable enabling us to compare experimental and control blocks. We categorized Trial Types in experimental blocks as before (Predictive-Consistent, Predictive-Inconsistent and Experimental Non-Predictive). However, in the corresponding control block that is paired with the experimental block (dependent on participant counterbalancing, either the block preceding or following the experimental block) we did not collapse all trials into Control Non-Predictive. Instead, each of the 120 trials in each control block were labeled with the same Trial Type as the corresponding trial from the paired experimental block. As a brief illustration: if the first trial of the first experimental block was a Predictive-Consistent trial, we would give the first trial of the first control block a Predictive-Consistent dummy label. Thus, instead of collapsing all control block trials to compare for general sequential effects, we can assign them these dummy labels. This will enable us to examine whether the sequential artifact of more alternations than repeats was what produced the pattern of results previously reported. If the control block pattern of responding across the three dummy Trial Types follows the experimental pattern, then we have evidence that sequential effects may have produced any differences in responding rather than learning about color contingencies.

ANOVAs comparing both RT and errors across Block, Block Type and the three level Trial Type revealed a significant effect of Trial Type in both RTs, \(F(2,62)=23.6\), \(p<.001\), and errors, \(F(2,62)=5.67, p=.006\). There was no significant effect of Block Type in either RTs, \(F(1,31)=\) \(1.55, p=.2\), nor errors, \(F(1,31)=.908, p=.3\). However, it is the interaction between Trial Type and Block Type that we are interested in, which was not significant in either RTs, \(F(2,62)=1.11, p=.3\), nor errors, \(F(2,62)=.166, p=.8\). This is due to both experimental and control Block Types following the same pattern, as is seen clearly in Figures 1 and 2. Thus, we found no difference in the observed pattern of responding to Trial Types between Block Types.


Figure 1. Mean RT for each Trial Type for experimental (solid bars) and control (open bars) Block Types.


Figure 2. Mean \% error for each Trial Type for experimental (solid bars) and control (open bars) Block Types.

\section*{Discussion}

The absence of a difference between the Block Types, and the lack of a significant Block Type by Trial Type interaction demonstrates quite clearly that the sequential artifact could have produced most of, if not the entire effect of Trial Type. Given this, it becomes vital to demonstrate that color learning can be obtained without the presence of this sequential artifact if the paradigm is to be of any use. The next experiment does just this.

\section*{Experiment 1}

In pilot work for this experiment, 16 participants formed two groups: eight participants who received the same, constrained sequences as in the original experiment (i.e. a color would never repeat) and eight who were trained on random sequences with no constraint (i.e. color repeats were permitted). Training lasted sixteen blocks (half experimental and half control), as the final four blocks were altered to act as test. In these the same colors were used as in the training, except contingencies were all set to the same, equal probability (50\%) of preceding either circle fill. This introduced a section of the experiment free from trained contingencies, meaning results could be compared across colors when matched. The results of this pilot study encouraged us to develop a design that encouraged more learning, as without the sequential artifact the Trial Type effect began to emerge at test in RTs and across training in errors for the group without the sequential artifact, but very weakly.

In an attempt to develop the original procedure to encourage learning whilst maintaining the original design elements, we first decided to increase the length of training. Instead of extending the experiment, which lasts around one hour, we chose to remove the control blocks and replace them with experimental blocks. Without the constraint on color repetitions and with the introduction of a set of test
blocks, possible sequential confounds should be avoided. Thus, control blocks for comparative purposes become surplus to the task's requirements, hence 15 blocks of experimental, training blocks preceded five blocks of test. This gave us one and one half times the amount of training in the original experiment. The training followed the form of the earlier described experimental blocks, so the experiment now contained only four possible colors in total, two Predictive and two Non-Predictive.

To further increase the possibility of learning, we ensured that participants were attending to the cue (the color of the square fill) when both processing and performing their response. When the circle fill occurs during the experiment, the colored square cue is still on screen and remains there until a response is made. However, attention will have shifted from the center of the screen and the color filled square onto the circle that has filled. Thus, if participants were attending to the circle fill when making their response the contingency between color and response would be strengthened if the color was represented in the location of the response cue itself. Consequently the circle in this version of the experiment did not fill white, but the color of the square color cue preceding it.

\section*{Method}

Participants. 16 University of Exeter undergraduate students ( 4 male, 12 female) aged between 18 and 24 ( \(M=19.25\) ) participated in the experiment for course credit.

Materials. As detailed in the original experiment, but with two differences. Firstly, the color of the circle fills was no longer white but the circle would fill with the color of the preceding square fill. Secondly, the blocks in training were exclusively experimental blocks. The experiment therefore consisted of only four colors in total (two Predictive and two Non-Predictive), presented across 15 training blocks and in 5 test blocks. The sequences were constructed randomly with no color repeat constraint.

Design and Procedure. The experiment again comprised of 20 blocks of 120 trials. All blocks were made up of a sequence of rights and lefts constructed as previously described, with the constraint that no color could follow itself on consecutive trials. The first 15 blocks acted as training, involving the same four colors in each Block Type as detailed in the original experiment. The final five blocks were test blocks involving the same four colors in each. For these blocks all colors were equally likely to be followed by a right or left circle fill. The procedure was as detailed in the aforementioned original experiment.

\section*{Results}

The data for both RTs and error rates were analyzed as in the original experiment, however, the variable of Block Type was no longer needed as all blocks involved the same four colors, two Predictive (split into Consistent and Inconsistent) and two Non-Predictive. Thus Trial Type and Block were the variables of interest in our ANOVAs. The results for RTs can be seen in Figure 3 and errors in Figure 4.


Figure 3. Average RT in msecs for each Trial Type over training (top panel) and test (bottom panel).

Training data demonstrated a significant effect of Trial Type in RTs, \(F(2,30)=11.23, p<.001\), and errors, \(F(2,30)=\) \(9.68, p=.001\). Predictive-Consistent trials are responded to more quickly and accurately than Non-Predictive trials, and these more quickly and accurately again than PredictiveInconsistent trials, which can be seen in both the top panels of Figures 3 and 4. This is further expressed by significant planned contrasts between Predictive-Consistent and Predictive-Inconsistent trials in both RTs, \(F(1,15)=18.98, p\) \(=.001\), and errors, \(F(1,15)=14.44, p=.002\) showing that participants responded faster and more accurately to trials that followed those contingencies they were trained on than those trials that were not consistent with these trained contingencies. Both lower RTs and fewer errors were present in Predictive-Consistent trials opposed to NonPredictive trials as well, shown in the planned contrast between the two in RTs, \(F(1,15)=5.65, p=.03\), and with a non-significant trend in same direction for errors, \(F(1,15)=\) \(1.24, p=.3\).

At test the RT data demonstrate no significant main effect of Trial Type, \(F(2,30)=.86, p=.4\), yet follow the same ordinal pattern as in training. The error data at test also show no significant main effect of Trial Type, \(F(2,30)=\) \(.077 p=.9\), with Predictive-Consistent trials resulting in faster and more accurate responding than PredictiveInconsistent trials. However, this is not entirely the pattern observed in training, as the control Non-Predictive stimuli produce more errors at test.


Figure 4. Mean \% errors for each Trial Type over training (top panel) and test (bottom panel).

The structured questionnaire revealed that twelve of the sixteen participants indicated surprise that the experiment did indeed involve color contingencies. This is further supported by the colors identified by participants as predictive. Given two choices each to name the two colors, participants selected the correct color on 16 out of the total 32 responses (exactly what one would expect by chance). They were asked also which of these two colors predicted which circle fill, which resulted in 9 accurate responses out of 32 (again this is close to the 8 expected by chance).

The prediction task itself involved two blocks of 16 trials, with all four colors occurring equally in each block resulting in eight trials where participants could predict Color 1 (which predicted the right circle fill) and eight trials for Color 2 (which predicted the left circle fill), see Figure 5. Of these 16 trials involving the Predictive Colors we can expect 8 correct responses by chance, which is near to the observed mean correct responses of 8.25 . This is not significantly different from chance and, when taking the colors separately, is not the result of learning about one color alone with mean correct responses of 3.94 and 4.31 for Color 1 and 2, respectively.

\section*{Discussion}

The results in training clearly demonstrate a pattern that provides evidence that learning about the contingencies between color and response has occurred. This is further supported the ordinal pattern in RTs and errors at test, which lessens the possibility that the effect is due to a speedaccuracy tradeoff.


Figure 5. The number of correct responses participants gave for Predictive Color 1 (filled bars) and Predictive Color 2 (open bars).

The structured interview responses and prediction task results provide evidence that this learning occurred outside of awareness, as not only were most participants surprised to learn that contingencies were present, they could not identify these colors, what the colors predicted, or use them to predict the correct, trained response above the level expected by chance.

Thus, we would conclude that across training we clearly demonstrated learning, in the absence of awareness, of color-response contingencies similar to those we believed to have found in our original paper's claim (Yeates et al., 2012). Furthermore, we have demonstrated that this effect somewhat remains when transferred to a test phase. The lack of significance may be attributed to extinction of the trained contingencies. Indeed when considering the first two blocks of test the ordinal pattern of RTs and errors are the same as during training and a post-hoc contrast test demonstrates a significant difference between PredictiveConsistent and Predictive-Inconsistent trials, \(F(2,15)=\) 13.23, \(p<.01\).

\section*{General Discussion}

We can conclude, as in the original paper, that this paradigm can still be used to demonstrate implicit learning in humans. However, this effect is clearly not as robust or easily obtained as we first imagined. When increasing the number of training trials and placing the predictive cue (color) on the response stimuli to ensure participants attended to them while processing or executing their responses we obtained effects comparable to those we previously reported.

We concede that whilst the prediction task demonstrates little evidence of conscious awareness that the result could be made more convincing if we could produce a non-null result (Z. Dienes, personal communication, 3 August 2012). A comparison between participants trained under intentional
instructions or indeed a Bayesian analysis (for which we would require an 'aware' prior from participants with explicit knowledge) of these data could strengthen our claims regarding the implicit nature of this learning.

It is not the intention of this paper to be entirely concerned with methodological issues. Our original paper suggested this paradigm as a method for studying implicit learning in humans and thus a refinement of the paradigm is of importance to the research questions that it enables us to investigate. We proposed that the process by which this occurs is associative in nature and aimed to produce variants of the task to investigate this behaviorally, alongside associative, computational modeling. It remains our intention to do so and we encourage the use of this paradigm in its re-designed form. We also accept that the prediction test in this version of the design is not maximally sensitive, as the test block (during which no contingencies are in play) separates training from this test of awareness. We intend to run other experiments using this paradigm without a test phase to address this issue.

\section*{REFERENCES}

Aitken, M.R.F. (1996). Peak shift in pigeon and human categorisation. Unpublished PhD Thesis, University of Cambridge.
Cleeremans, A. (1993). Mechanisms of Implicit Learning: Connectionist Models of Sequence Processing. Cambridge, MA: MIT Press.
Jones, F.W., \& McLaren, I.P.L. (2009). Human sequence learning under incidental and intentional conditions. Journal of Experimental Psychology: Animal Behavior Processes, 35, 538553.

Lovibond, P.F., \& Shanks, D.R. (2002). The role of awareness in Pavlovian conditioning: Empirical evidence and theoretical implications. Journal of Experimental Psychology: Animal Behavior Processes, 28, 3-26.
McLaren, I.P.L., Green, R.E.A., \& Mackintosh, N.J. (1994). Animal learning and the implicit/explicit distinction. In N.C. Ellis (Ed.) Implicit and explicit learning of languages (pp. 313332). New York, NY: Academic Press.

Mitchell, C.J., De Houwer, J., \& Lovibond, P.F. (2009). The propositional nature of human associative learning. Behavioral and Brain Sciences, 32, 183-246.
Posner, M.I. \& Synder, C.R.R. (1975). Attention and Cognitive Control. In Robert L. Solso (ed.), Information Processing and Cognition: The Loyola Symposium. Lawrence Erlbaum.
Shanks, D.R. \& St. John, M.F. (1994) Characteristics of dissociable human learning systems. Behavioral and Brain Sciences, 17, 367-447.
Shanks, D.R. \& Lovibond, P.F. (2002) Autonomic and eyeblink conditioning are closely related to contingency awareness: Reply to Wiens and Öhman (2002) and Manns et al. (2002). Journal of Experimental Psychology: Animal Behavior Processes, 28, 3842.

Willingham D.B., Nissen M.J. and Bullemer P. (2009). On the development of procedural knowledge. Journal of Experimental Psychology: Learning, Memory and Cognition, 15, 1047-1060.
Yeates, F., Jones, F.W., Wills, A.J., Aitken, M.R.F., \& McLaren, I.P.L. (2012). Implicit learning: A demonstration and a novel SRT paradigm. In Miyake N., Peebles D., \& Cooper R.P. (Eds.) Proceedings of the \(34^{\text {th }}\) Annual Meeting of the Cognitive Science Society (pp. 1185-1190). Austin, TX.

\title{
Linguistic Variability and Adaptation in Quantifier Meanings
}

\author{
Ilker Yildirim, Judith Degen, Michael K. Tanenhaus, T. Florian Jaeger \\ Department of Brain \& Cognitive Sciences, University of Rochester, Rochester, NY 14627 \\ \{iyildirim, jdegen, mtan, fjaeger\} @bcs.rochester.edu
}

\begin{abstract}
People's representations of most and arguably all linguistic and non-linguistic categories are probabilistic. However, in linguistic theory, quantifier meanings have traditionally been defined set-theoretically in terms of categorical evaluation functions. In 4 "adaptation" experiments, we provide evidence for the alternative hypothesis that quantifiers are represented as probability distributions over scales (e.g., Zadeh, 1965). We manipulate exposure to different distributions of "some" and "many" and find that listeners adapt to those distributions, as predicted. Our results suggest that the interpretation of quantifiers is best modeled as a process involving rich, probabilistic representations.
\end{abstract}

Keywords: Quantifiers; Semantics; Language processing; Adaptation; Generalization

\section*{Introduction}

In linguistic theory, quantifier meanings have traditionally been defined set-theoretically in terms of categorical evaluation functions (Barwise \& Cooper, 1981) yielding either truth or falsity of a sentence containing a quantifier. Quantifiers are understood as relations between sets:
some \((A, B)\) is true iff \(\|A\| \cap\|B\| \neq \varnothing\)
\(\operatorname{many}(\mathrm{A}, \mathrm{B})\) is true iff \(\|\mathrm{A}\| \cap\|\mathrm{B}\|>n\), where \(n\) is some large number

For example, the sentence Some candies are green is true just in case the intersection of the candies and the green things is not empty. Similarly, Many candies are green is true just in case the cardinality of the intersection of the candies and the green things is larger than some contextual norm \(n\). This points to a notable feature of some quantifiers: they exhibit both vagueness and context-dependence (Solt, 2009).

A class of alternative views tries to incorporate this feature by representing quantifiers probabilistically. For example, fuzzy logic (Zadeh, 1965) approaches to meaning consider quantifiers such as "some" as probability distributions over scales (e.g., Moxey \& Sanford, 1993). Probabilistic quantifier semantics are at the heart of recent models of both syllogistic reasoning (Chater \& Oaksford, 1999) and scalar implicature (Goodman \& Stuhlmüller, 2013). Here we provide further evidence that quantifiers are indeed interpreted in a probabilistic, graded manner. The novel empirical contribution lies in addressing the adaptability of these distributions to variable language environments.

The probabilistic view on quantifier meaning is illustrated in Figure 1a: "some" and "many" form graded distributions over a contextually determined scale. \({ }^{1}\) Previous work

\footnotetext{
\({ }^{1}\) For example, it is not as plausible to quantify 18 out of 1000 as "many" as to quantify 18 out of 20 .
}


Figure 1: Illustration of across speakers variability in meanings of quantifiers.
has implicitly assumed that these distributions are invariant across linguistic environments, in that the distribution corresponding to, for example, "some" is stationary across different dialects, speakers, genres, and so on.

However, variability in language use is the norm. Speakers differ in their realization of phonemes (cf. Allen, Miller, \& DeSteno, 2003), lexical preferences (e.g., couch vs. sofa), as well as syntactic preferences (e.g., some speakers use passives more often than others, Weiner \& Labov, 1983). Such linguistic variability is a challenge for comprehenders that must be overcome to achieve successful communication. One solution for dealing with variable linguistic environments is to track and adapt to the joint statistics of linguistic categories (e.g. phonemes, words, syntactic structures) and contextual cues, including the speaker.

A powerful way to test whether listeners adapt to the statistics of the input is to determine whether categorization functions shift with exposure. If listeners adapt to new environments in which the statistics diverge from their prior beliefs, this would suggest that linguistic representations are sensitive to and adapt to such sources of variability. This reasoning has been successfully applied to phonetic categories (e.g., Clayards, Tanenhaus, Aslin, \& Jacobs, 2008; Vroomen, Linden, Gelder, \& Bertelson, 2007; Kraljic \& Samuel, 2006), prosodic categories (Kurumada, Brown, \& Tanenhaus, 2012), and syntactic categories (Fine, Jaeger, Farmer, \& Qian, underreview; Kamide, 2012).

Here we ask whether listeners' representations of the quantifiers "some" and "many" are probabilistic and sensitive to environmental variability. Figure 1 b depicts hypothetical some and many distributions over cardinalities for two speakers whose use of the quantifiers differs.

In four adaptation experiments, we provide evidence that quantifiers are represented as probability distributions. More-


Figure 2: Procedure for Experiment 1. Top panel illustrates an exposure phase trial. Bottom panel illustrates a test phase trial.
over, we present evidence that listeners' interpretations of quantifiers rapidly adapt to the statistics of the local linguistic environment represented by a novel speaker. Furthermore, we provide evidence suggesting that listeners' adaptation might be taking place across multiple levels (or types) of representations. We argue that the rapid adaptation that we observe involves both speaker-specific and quantity level representations enabling transfer of adaptation across visual object types.

\section*{Experiment 1}

Behavioral evidence strongly suggests that listeners dynamically adapt to the phonetic and syntactic variability in their language environment (Vroomen et al., 2007; Bertelson, Vroomen, \& Gelder, 2003; Kamide, 2012; Fine et al., underreview). But do such adaptive processes also occur at the level of meaning? We addressed this question in an experiment by investigating whether listeners adapt their interpretations of the two English quantifiers "some" and "many" based on experience with a speaker who uses these quantifiers in a way that deviates from the listener's prior expectations.

Our experimental logic followed that of previous adaptation experiments (e.g., Bertelson et al., 2003). The experiment employed a by-2 between-participant design. One group of participants was exposed to a novel speaker's use of the word "some" (some-biased group). Another group of participants was exposed to a novel speaker's use of "many" (many-biased group). Participants in both groups were then tested on how they interpreted that speaker's utterances.

\section*{Participants}

80 participants were recruited over Amazon's crowd-sourcing service Mechanical Turk. All participants were self-reported native speakers of English. Each experimental session took about 15 minutes and participants were paid \(\$ 2\).

\section*{Procedure and Materials}

Figure 2 illustrates the materials and procedure for this experiment. The experiment proceeded in two phases, the exposure
phase and the test phase.
In the exposure phase, participants watched videos as in the top panel of Figure 2. The video showed a bowl of 25 candies in the bottom right of the screen. The bowl always contained a mixture of green and blue candies, but the number and spatial configuration of the candies differed between trials. Importantly, the video showed a speaker describing the scene in a single sentence. The videos played automatically at the start of the trial and the scene - the candy bowl remained visible even when the video had finished playing (as shown in Figure 2, top). Two different speakers were employed between participants to ensure that effects were not due to a particular speaker.

The exposure phase consisted of 10 critical and 10 filler trials. In critical trials the speaker produced the sentence Some of the candies are green (some-biased group) or Many of the candies are green (many-biased group). On a critical trial, the bowl always contained 13 green candies and 12 blue candies. This scene was identified as the Most Ambiguous Quantity (MAQ) scene in a preceding norming study in which participants rated how well descriptions containing different quantifiers matched scenes sampled from a continuum of quantities.

The remaining 10 trials in the exposure phase were filler trials. On a filler trial, participants observed the speaker correctly describing a scene with no green candies in it as None of the candies are green ( 5 trials) and a scene with no blue candies in it as All of the candies are green ( 5 trials). The purpose of the filler trials was two-fold. First, it made our manipulation less obvious. Second, including clearly true descriptions of unambiguous scenes encouraged participants to believe that the speaker was indeed intending to accurately describe the scene. The order of the critical and the filler trials was randomized.

Following the exposure phase, participants entered the test phase. The test phase was intended to assess participants' beliefs about the speaker's use of both "some" and "many". On test trials, participants saw a candy scene in the center of the display and two identical still images of the speaker from the exposure phase on either side of the scene (see Figure 2, bottom).

The two images of the speaker were paired with one of the two alternative descriptions Some of the candies are green and Many of the candies are green each. The participants' were asked to rate how likely they thought the speaker would be to describe the scene using each of the alternative descriptions. They performed this task by distributing a total of 100 points across the two alternatives (the first and the second slider bars; see Figure 2, bottom panel) and a third alternative - namely "Other" - to reflect how much they thought that neither of the two alternatives fit the scene (the third slider bar). As in the exposure phase, scenes always consisted of a bowl of 25 candies with differing numbers of green candies. To assess participants' beliefs about the speaker's use of "some" and "many", we sampled scenes from the entire scale. Specifically, scenes contained one of
\(\{1,3,6,9,11,12,13,14,15,17,20,23\}\) green candies out of 25 candies. Over 39 test trials, participants rated each scene 3 times. Different instances of the same scenes differed in the spatial configuration of the blue and green candies. The order of the scenes and the mapping from alternative descriptions to slider bars were randomized and counterbalanced.

To ensure that participants were attending to the task, we placed catch trials after about every six trials. On some of these trials, a gray cross appeared at a random location in the scene. Before the next trial began, participants were asked if they had seen a gray cross in the previous scene.

\section*{Data Analysis}

We did not analyze the "Other" responses. The top row in Figure 3a shows the distribution of "some" and "many" in the test phase separately for the two groups of participants. The distributions were obtained by averaging participants ratings for the different scenes along the scale. We first averaged across the three instances of each scene within a speaker and then averaged those ratings across speakers (separately for each point on the scale). Those average ratings were then fit with a generalized linear model with cubic splines, which gave us the continuous curve for each of the two alternative descriptions shown in Figure 3a, top row. Participants in the some-biased group adapted in the opposite (and predicted) direction from participants in the many-biased group. That is, the distributions for participants in the some-biased group were updated such that they were more likely to rate a wider range of scenes as more likely with respect to the "Some" description. Such high ratings of the "Some" description came at the expense of the alternative description. Similarly, the distributions for participants in the many-biased group reflected that these participants were more likely to rate the "Many" description as more likely at the expense of the alternative description.

In order to quantify the shift in interpretations between the two groups of participants, we derived two measures. First, for each participant, we estimated the MAQ as the point where the two curves were closest to each other (excluding the extremes of the scene continuum).

Similar in logic to the phonetic adaptation experiments, we reasoned that participants in the many-biased group would come to interpret a "many" as applying to a wider range of scenes (and hence quantities). Because participants had to share a total of 100 points between the alternatives, this adaptation in favor of "many" would be at the expense of "some" ratings. Therefore, the MAQ scene should shift to the lower end of the continuum of set sizes compared to 13 (the MAQ scene from the norming study). In contrast, for participants in the some-biased group, if they were to adapt to the statistics of the speaker during the exposure phase, they should rate a wider range of scenes more likely to be described using the quantifier "some." These high ratings for "some" would come at the expense of "many." Therefore, the MAQ should shift to the higher end of the continuum of set sizes compared to the MAQ scene from the norming experiment.

To ensure that our findings were not just an artifact of the way the analysis was conducted, we performed a separate set of analyses by computing the Area Under the Curve (AUC) for each of the two alternative descriptions. That is, again, we first fit a generalized linear model with cubic splines for each participant. Then we computed the AUC for each alternative description (by summing up the area under the fitted curve) and subtracted the AUC for the "Some" curve from the "Many" curve.

We reasoned that if participants adapted their quantifier interpretations in the predicted direction, then the AUC difference should be smaller (or negative) for participants in the many-biased group and larger (or positive) in the some-biased group.

All analyses were conducted using the R statistics software package (R Development Core Team, 2005).

\section*{Results}

Middle row in Figure 3a presents the results for MAQ analysis. As predicted, for each speaker, the MAQ values were significantly smaller for the many-biased group than for the some-biased group ( \(p<10^{-6}\) ).

Bottom row in Figure 3c shows re-evaluation of the same data using the AUC analysis. As predicted, for both speakers, the AUC difference for the many-biased group and the somebiased group grew in opposite directions ( \(p<10^{-6}\) ).

These results suggest that listeners indeed track the joint statistics of quantities, speakers, and the quantifiers in their environment, and rapidly adapt their interpretations in response to the new input.

\section*{Experiment 2}

One limitation of Experiment 1 is that effects might be speaker and/or scene specific. Experiments 2 and 3 were designed to test the hypothesis that the updating was more general. Experiment 2 examined adaptation when the emphasis is shifted away from the specific speaker by changing the instructions and by removing the speaker's face from the test phase trials. Experiment 3 used different objects in the test phase - Xs and Os instead of candies of different colors.

\section*{Participants}

Participants were 80 Mechanical Turk workers. All participants were self-reported native speakers of English. Each experimental session took about 15 minutes, and participants were paid \(\$ 1.5\).

\section*{Procedure and Materials}

The experimental stimuli were identical to those of Exp. 1.
The procedure was identical to that of Exp. 1 with the exception of the test trials. Unlike the previous experiment, participants did not see a cue to the speaker's identity. Instead, they saw only the two sentences providing the two alternative descriptions for the scene located at the center. The participants' task was to rate how likely that they thought that \(a\)
\begin{tabular}{lllll} 
Exp & Pre-exposure & Exposure & Test (Post-exposure) & Groups \\
\hline 1 & N/A & Candy scenes in videos & VS: Candies & Some-biased vs. Many-biased \\
& & & LS: Typed sentences + speaker images & \\
\hline 2 & N/A & Candy scenes in videos & VS: Candies & Some-biased vs. Many-biased \\
& & & LS: Typed sentences & Some-biased vs. Many-biased \\
\hline 3a & VS: Candies & Candy scenes in videos & VS: Letters & \\
& LS: Typed sentences & & LS: Typed sentences & Some-biased vs. Many-biased \\
\hline 3b & VS: Candies & Candy scenes in videos & VS: Candies & \\
& LS: Typed sentences & & LS: Typed sentences & \\
\hline
\end{tabular}

Table 1: Summary of the experimental designs. VS: visual stimuli. LS: linguistic stimuli.
speaker would describe the scene with each of the alternative descriptions. They again distributed a total of 100 points across the two alternative descriptions and choice of "Other."

As in Exp. 1, 40 participants were assigned to each of the some-biased and many-biased groups. For each group, of the 40 participants, 20 were assigned to each of the speakers in the videos.

A summary of the procedures used in the different experiments is provided in Table 1.

\section*{Results}

We excluded one of the participants from the analysis because they never adjusted the sliders on the test trials. Top row in Figure \(3 b\) plots the mean ratings by participants in each of the two groups. Participants adapted their interpretations of the quantifiers in accordance with the speaker-provided statistics, though less so than in Exp. 1.

We performed the same MAQ and AUC analysis as for Exp. 1. Middle row in Figure 3b illustrates that the MAQ for participants in the many-biased group was significantly smaller than the MAQ for participants in the some-biased group. This was true for both speakers ( \(p<0.01\) ). The AUC analysis, bottom row in Figure 3b, also revealed significant adaptation ( \(p<0.01\) ).

The results from Exp. 2 suggest that the adaptation observed in Exp. 1 is not a simple speaker-specific adaptation effect and suggest instead that listeners' adaptation to the statistics of the linguistic environment might occur at multiple levels of representations. Adaptation was stronger in Exp. 1 where a cue to the speaker was provided in the test phase. However, the fact that we also observe adaptation in Exp. 2 (when no such cue is available) suggests that this adaptation was to some extent generalized across speakers.

\section*{Experiment 3a}

It is nevertheless possible that the adaptation effects found in Exps. 1 and 2 is object-specific, i.e. quantifier interpretations are only updated for quantities of candies. Exp. 3a tested this by replacing the candy scenes in the test phase trials with scenes containing letters ( Xs and Os ).

\section*{Participants}

We recruited 40 participants over Mechanical Turk who were self-reported native speakers of English. Each experimental session took about 15 minutes. Participants were paid \(\$ 1.5\).

\section*{Materials and Procedure}

The test stimuli differed from the previous experiments. On each test trial we presented 25 letters, each of which was either an X or an O . The letters in each scene were scattered within a circle (but there was no visible boundary). The descriptions that participants rated were Some of the letters are Xs and Many of the letters are Xs. Number of Xs in a scene could be any of the values that the number of green candies could be in a scene from Exps. 1 and 2. Participants' task was again to rate (by distributing 100 points) how likely that they thought a speaker would describe the scene with each of the alternative descriptions and the third choice of "Other."

The stimuli in the exposure phase were identical to Exp. 1 and 2 but speaker identity was not varied between participants. Half of the participants were assigned to the somebiased group and half to the many-biased group.

In order to establish that transfer occurred between the candy and the letter scenes, we included a pre-exposure test phase. The aim of these pre-exposure test trials was to measure participants' prior interpretations of quantifiers in candy scene descriptions and compare them to quantifiers in letter descriptions following exposure to candy scenes. That is, we analyzed participants' responses to descriptions of letter scenes in the post-exposure test trials and responses to descriptions of candy scenes in the pre-exposure test trials together to measure whether participants' interpretations changed with exposure.

\section*{Data Analysis}

For each participant in the MAQ analysis, we determined the MAQ for the pre- and post-exposure test responses separately. Then we subtracted the pre-exposure MAQ from the post-exposure MAQ. A positive difference is expected for the some-biased group and a negative one for the many-biased group.

For the AUC analysis, we first calculated the AUC difference on pre-exposure test trials for each participant. Then we calculated the AUC difference on post-exposure test trials. The pre-exposure AUC difference was then subtracted from the post-exposure AUC difference. The expected patterns of results was the same as in the previous experiments.

\section*{Results}

Top row in Figure 3c illustrates the group mean ratings for the post-exposure test trials. Participants' ratings in the some-


Figure 3: Each column shows data for experiment (e.g., left-most column is Experiment 1, right-most column is Experiment \(3 b)\). The vertical lines in the density panels at the top denote the MAQ scene (scene 13) determined based upon a preceding norming study. \(\mathrm{MB}=\) Many-biased, \(\mathrm{SB}=\) Some-biased.
biased group and the many-biased group did not differ before exposure. However, following the adaptation trials, participants' responses reflect that they adapted in the predicted directions:

The MAQ difference analysis in middle row in Figure 3c shows that indeed participants in the some-biased group rated "Some" descriptions as more likely across the whole continuum of scenes, whereas participants in the many-biased group favored "Many" descriptions at the expense of the alternative descriptions \((p<0.01)\). The difference in AUC difference analysis in Figure 3c, bottom row, reaffirmed our findings ( \(p<0.01\) ).

The results from Exp. 3 suggest that participants' quantifier interpretations did not adapt candy-specifically - instead, quantifier adaptation transferred to a different visual environment. That is, the quantity level representation itself adapted.

\section*{Experiment 3b}

To establish that the results we obtained in Exp. 3 were not due merely to the additional pre-exposure test trials, we re-ran Exp. 2 with pre-exposure test trials. The pre- and post-exposure test trials were identical and contained candy scenes.

We recruited 120 participants over Mechanical Turk who were self-reported native speakers of English. Each experimental session took about 15 minutes, and participants were paid \(\$ 1.5\).

60 participants were assigned to each of the the somebiased group and the many-biased group. 30 participants in each group were assigned to each of the speakers.

Top row in Figure 3d shows the mean post-exposure test trial responses (responses did not differ on pre-exposure test trials between groups). Following adaptation trials, there is a clear effect of group in the predicted direction, replicating the results from Exp. 2.

Middle row in Figure 3d shows the results of the MAQ difference analysis. The qualitative patterns of our results reflects the predicted pattern, such that the MAQ difference was positive in the some-biased group and negative in the manybiased group. This difference was significant ( \(p<0.01\) ). In the difference in AUC difference analysis (Figure 3d, bottom row) the participants adapted to the speakers in the predicted directions ( \(p<10^{-4}\) ).

We thus replicated the results from Exp. 2, again indicating that listeners' adaptation of quantifier meanings is broad. It also confirms that the inclusion of pre-exposure test trials
is most likely not the reason for the transfer effect found in Exp. 3.

\section*{Discussion}

Our results indicate that semantic representations can be adapted to new linguistic environments. At least in situations like the ones investigated here, this adaptation seems to be rapid, requiring only very limited exposure. Our observation that adaptation can be transferred across multiple linguistic and visual environments suggest that these adaptations are not limited to the specific nature of the scale, although it remains to be seen how such adaptation generalizes to scales of different ranges. Our experiments support probabilistic theories of quantifier meaning over set-theoretic ones. Our results are also compatible with a soft version of set-theoretic representations under which there are core logical representations that are enriched with probabilistic expectations about the use of quantifiers with different set sizes.

In this paper, we addressed the question of whether and how listeners adapt to speakers' use of the quantifiers "some" and "many." A recently emerging literature in other domains of language processing has provided evidence that listeners can rapidly adapt to speaker-specific variability in their language environment. Most of this line of work has focused on adaptation to phonological variability across speakers (e.g., Kraljic \& Samuel, 2006; Clayards et al., 2008; Vroomen et al., 2007). To our knowledge, our work is the first to extend the logic of language adaptation experiments to semantic representations.

Future experimental work should address whether listeners can adapt to multiple speakers' quantifier use statistics simultaneously. While the relative magnitude of the shift in interpretations of "some" and "many" between Experiments 1 and 2 might be taken to provide preliminary evidence that listeners maintain both speaker-specific and speaker-general representations and that both of these are affected by recent experience with a specific speaker, future work is required to address more directly the nature of representations that are adapted by recent exposure. For example, it is possible that listeners maintain hierarchically structured representations over speakers, groups of speakers (based on their similarity), and so on (cf. modeling of phonetic adaptation; Kleinschmidt \& Jaeger, 2011). Future research will also need to address how much of the adaptation comes from base-rate effects (e.g., changes in the prior probabilities of quantifiers) and how much of it comes from adaptation of the meaning of each quantifier (e.g., changes in the likelihood functions of quantifiers). In pursuing these questions, we believe it will be necessary to take a two-pronged approach, combining behavioral paradigms like the one introduced here with computational models that provide clear quantifiable predictions about how listeners adapt previous experience with other linguistic environments based on recent experience with a specific linguistic environment.

\section*{Acknowledgments}

This work was supported by research grants from the National Institute of Health (NIH HD 27206) to MKT, and by a NSF CAREER award (IIS-1150028) as well as an Alfred P. Sloan Research Fellowship to TFJ.

\section*{References}

Allen, J., Miller, J., \& DeSteno, D. (2003). Individual talker differences in voice-onset-time. The Journal of the Acoustical Society of America, 113, 544.
Barwise, J., \& Cooper, R. (1981). Generalized quantifiers and natural language. Linguistics and Philosophy, 4(2), 159-219.
Bertelson, P., Vroomen, J., \& Gelder, B. de. (2003). Visual recalibration of auditory speech identification a mcgurk aftereffect. Psychological Science, 14(6), 592-597.
Chater, N., \& Oaksford, M. (1999). The Probability Heuristics Model of Syllogistic Reasoning. Cognitive Psychology, 38(2), 191-258.
Clayards, M., Tanenhaus, M., Aslin, R., \& Jacobs, R. (2008). Perception of speech reflects optimal use of probabilistic speech cues. Cognition, 108(3), 804.
Fine, A. F., Jaeger, T. F., Farmer, T., \& Qian, T. (underreview). Rapid adaptation of syntactic expectation.
Goodman, N. D., \& Stuhlmüller, A. (2013, January). Knowledge and implicature: modeling language understanding as social cognition. Topics in Cognitive Science, 5(1), 17384.

Kamide, Y. (2012). Learning individual talkers structural preferences. Cognition.
Kleinschmidt, D., \& Jaeger, T. (2011). A Bayesian belief updating model of phonetic recalibration and selective adaptation. In ACL CMCL 2011 (p. 10).
Kraljic, T., \& Samuel, A. (2006). Generalization in perceptual learning for speech. Psychonomic Bulletin \& Review, 13(2), 262-268.
Kurumada, C., Brown, M., \& Tanenhaus, M. (2012). Pragmatic interpretation of contrastive prosody: It looks like speech adaptation. In CogSci 2012.
Moxey, L. M., \& Sanford, A. J. (1993). Prior expectation and the interpretation of natural language quantifiers. European Journal of Cognitive Psychology, 5(1), 73-91.
Solt, S. (2009). The semantics of adjectives of quantity. Unpublished doctoral dissertation, CUNY.
Vroomen, J., Linden, S. van, Gelder, B. de, \& Bertelson, P. (2007). Visual recalibration and selective adaptation in auditory-visual speech perception: Contrasting build-up courses. Neuropsychologia, 45(3), 572-577.
Weiner, E., \& Labov, W. (1983). Constraints on the agentless passive. Journal of Linguistics, 19(1), 29-58.
Zadeh, L. (1965). Fuzzy sets. Information and Control.

\title{
The Role of Category Structure in Category Learning
}

\author{
Hyungwook Yim (yim. \(31 @\) osu.edu) \\ Department of Psychology, The Ohio State University \\ 267 Psychology Building, 1835 Neil Avenue \\ Columbus, OH 43210, USA
}

\author{
Vladimir M. Sloutsky (sloutsky.1@osu.edu) \\ Department of Psychology, The Ohio State University \\ 239 Psychology Building, 1835 Neil Avenue \\ Columbus, OH 43210, USA
}

\begin{abstract}
Two category-learning experiments were conducted to examine the role of category structure and learning regime in category learning. We particularly focused on effects of these factors on selective attention, which was measured by eyetracking methods. Results show that even though supervision was weaker than in previous studies, attention optimization and cost of attention were observed during category learning (Experiment 1). Moreover, there were faster learning and stronger attention optimization when statistically denser categories were learned (Experiment 2). At the same time, there were weaker costs of selective attention when learning denser categories than when learning sparser categories. Results are discussed in relation to theories of category learning.
\end{abstract}

Keywords: category learning, cost of selective attention, category structure, eye tracking

\section*{Introduction}

Selective attention is one of the key components in category learning (Kruschke, 1992; Nosofsky, 1986; Shepard, Hovland, \& Jenkins, 1961). The ability to selectively attend to category-relevant dimensions aids the learner to ignore category-irrelevant information and makes learning more efficient. For example, when learning how to distinguish Siberian Huskies from Alaskan Malamutes, which look very similar, the color of the eyes is one of the relevant features one should look for (most Huskies have blue eyes and Malamutes have brown eyes). Therefore, learning to focus on the color of the eyes while ignoring other irrelevant features (e.g. color of the fur or markings) would aid learning the two categories. Selective attention could be captured in category learning tasks that involve eye-tracking as attention optimization, where looking to categoryrelevant information increases and looking to irrelevant information decreases (Hoffman \& Rehder, 2010).

However, optimizing one's attention to the current category-relevant dimension may result in learning to ignore the category-irrelevant dimension, which results in learned inattention to the irrelevant dimension (Kruschke \& Blair, 2000). Therefore, if a new to-be-learned category has a category-relevant dimension that was previously irrelevant, learning may become more difficult, which represents a cost of selective attention. For example, when learning to distinguish meerkats from prairie dogs, which again look
very similar, the shape of the ears is one of the good dimensions to look. However, if one has previously learned how to distinguish Huskies from Malamutes, where eyes were attended and ears were ignored, learning to attend to the once-ignored ears would be hindered.

The close link between attention optimization and the cost of selective attention has been demonstrated in previous research (e.g., Hoffman \& Rehder, 2010). In their study, participants were given either a supervised classification task (e.g. classifying a stimulus into category A or B) or a supervised inference task (e.g. inferring the missing feature of a stimulus that belongs to a certain category) and their eye movements were recorded. Since the classification task (e.g. focusing on the color of the eyes to classify Huskies and Malamutes) required attention optimization to the relevant dimension, results showed cost of selective attention when learning a new category. On the other hand, since the inference task (e.g. figuring out whether a Malamute has blue eyes or brown eyes) does not require attention optimization, the cost did not occur when learning the next category. Therefore, the study showed that (a) the characteristics of the task affect allocation of attention and (b) when attention optimization occurred, the cost of selective attention also followed.

Although attention may be affected by the characteristics of the task (i.e., classification vs. inference) it can also be affected by category structure. Categories that have multiple correlated dimensions (or statistically dense categories) may be learned without selective attention, whereas categories that have few relevant dimensions (or statistically sparse categories) may require selective attention (Kloos \& Sloutsky, 2008; Sloutsky, 2010). For example, when learning the category dog, many dimensions are relevant (e.g. nose, fur, four-legs, etc.) and therefore it is relatively easy to learn. However, when learning abstract concepts such as friction, very few dimensions are relevant among many irrelevant dimensions (e.g. a car trying to stop at the red light and a person trying to open a jar both shows friction). Therefore, to learn a sparse category one has to "selectively attend" to the relevant dimension among many other irrelevant dimensions.

Finally, the deployment of selective attention may be also affected by learning regime. Since supervised learning provides information about the relevant dimension, it is more likely to recruit selective attention than unsupervised
learning (Kloos \& Sloutsky, 2008). Kloos \& Sloutsky (2008) showed that sparse categories could largely benefit from supervision, while it could sometimes hinder dense categories. Since selective attention filters irrelevant information and allocate attention to the relevant on information (Kruschke, 2001; Mackintosh, 1975), trying to attend to multiple correlated information (i.e. dense categories) could be harder than attending to a few.

In the current study, we examined the effects of category structure on selective attention in the course of category learning. In all experiments, a supervised category learning task was used while the participants' eye movements were recorded. Moreover, cost of attention and attention optimization were observed to infer the attentional mechanism in category learning.


Figure 1. Description of the stimuli structure and experimental design. (a) stimuli used in Experiment 1 sparse category, (b) experimental design of Experiment1, and (c) stimuli used in Experiment 2 - dense category (note that ' R ' represents the location of the relevant dimension in each exemplar which was not visible to the participants)

\section*{Experiment 1}

Experiment 1 examined the cost of attention when an extradimensional shift occurred between two sparse categories with supervision. As shown in previous studies, extradimensional shift maximizes cost of attention, therefore making it easy to observe the attentional dynamics during category learning (Hall, 1991; Hoffman \& Rehder, 2010).

\section*{Methods}

Participants Thirty-three adults with normal or corrected to normal vision participated in the experiment for course credit. An additional 8 participants were excluded from the analysis due to not exceeding the learning criterion (see Procedure).
Stimuli Flower-like artificial categories were used in the experiment (see Figure 1a). Each exemplar had a gray hexagon in the middle with six colored shapes on every side. Among the six colored shapes, five changed their color/shape in a binary fashion, whereas one was constant, serving as a category relevant dimension. Therefore, there were 32 unique stimuli for each category with two categories having the relevant feature on the right-bottom side of the hexagon (i.e., category A: purple triangle, category B : blue semi-circle) and two categories having the relevant feature on the left side of the hexagon (i.e., category C: yellow pentagon, category D : orange square). Therefore, the relationship between A or B and C or D was an extra-dimensional shift.
Procedure The experiment had 2 phases and in each phase there were 4 blocks. Within each block there were 8 learning trails followed by 4 test trials. After the first 4 blocks (Phase 1), unknown to the participants, the category had an extra-dimensional shift (see Figure 1b). Therefore if the first half of the blocks were presented with category A, the second half of the blocks were presented with category \(C\) in the learning trials. In the learning trials, exemplars were presented for 1.5 seconds, one at a time in the middle of the screen. At the beginning of each block, participants were told that they would see flowers that have one common feature they had to find, which served as a supervision signal.

In the test trials two category exemplars were presented side by side until the participant made a response. One exemplar was a novel exemplar from the category that was used in the learning trials. The other exemplar was a new category where the relevant feature was in the same dimension as the learned category but had a different feature (e.g. Cat A and Cat B in Figure 1a). Participants were told to choose the exemplar that they thought was a member of the category they saw in the learning trials by pressing a left or right response button. When the response was made, the stimuli disappear without any feedback. Also before each learning and test trial, a fixation point (i.e. red cross) was presented on a random-dot background, and the participants were told to look at the fixation to proceed with the experiment. Moreover, a Tobii T60 eye tracker was used to
collect eye gaze with the sampling rate of 60 Hz during the whole experiment.

\section*{Results}

Before analyzing the data, participants who did not learn the first category were excluded. To be considered as a learner one had to have 3 correct responses out of 4 test trials in the last block of Phase 1 (i.e. block 4). To determine whether a cost was incurred, accuracy, reaction time, and eye gaze data were analyzed by block. Especially by comparing the blocks before and after the unknown category switch (i.e. block 4 vs. block 5).

The overall accuracy for the test blocks was \(.90, S D=.21\) (Phase 1: \(M=.92, S D=.18\), Phase 2: \(M=.87, S D=.23\) ), with all test trials being significantly higher than chance performance, \(p<.001\) (see Figure 2a). Results of a \(2 \times 4\) (Phase \(\times\) Block) within-subjects ANOVA conducted on accuracy scores at test showed a main effect for Block, \(F(2.3,73.61)=8.14, p<.001\), indicating that accuracy differed by block, but there was no significant main effect for Phase or a interactions ( \(p s>.05\) ). Moreover, a significant cost of attention was demonstrated between the last block of learning phase 1 (block 4) and the first block of learning phase 2 (block 5) by a significant decrease in accuracy from block 4 to block \(5, t(32)=5.07, p<.001\).

Before analyzing the reaction time (RT), all incorrect responses were excluded, and for each individual the median RT for each block were used in the analysis. The mean reaction time for all test blocks was \(1160 \mathrm{~ms}, S D=\) 892 ms (Phase 1: \(M=1199 \mathrm{~ms}, S D=922 \mathrm{~ms}\), Phase 2: \(M=\) \(1121 \mathrm{~ms}, S D=863 \mathrm{~ms}\) ) (see Figure 2b). A \(2 \times 4\) (Phase \(\times\) Block) within-subjects ANOVA conducted on RT showed a main effect for Block, \(F(1.77,54.91)=9.58, p<.001\), but there was no significant main effect for Phase or a interaction ( \(p s>.05\) ). Statistical difference between block 4 and block 5 were also found, \(t(32)=2.78, p<.005\), demonstrating a cost of attention.

Eye gaze data were also analyzed for each block by calculating the weighted proportion of looking to the relevant spatial dimension. This value was calculated by taking looking time (fixation) to the relevant features divided by looking time (fixation) to the irrelevant and relevant features combined. However, since there was greater spatial area for irrelevant features ( 5 shapes) than the relevant features ( 1 shape), looking time to the relevant features was multiplied by five to equate the spatial area. Therefore, .50 in the analysis represents an equal amount of looking to the relevant and irrelevant features at a given block. Fixations were calculated by using an I-DT algorithm with a minimum duration threshold of 100 ms and a dispersion threshold of \(1^{\circ}\) of visual angle (Salvucci \& Goldberg, 2000).

The overall weighted proportion of looking to the relevant dimension was for all test blocks was \(.63, S D=.30\) (Phase 1: \(M=.63, S D=.30\), Phase 2: \(M=.64, S D=.31\) ). All blocks except the first blocks in each phase (i.e. block1 and block 5) showed a significantly higher proportion of looking
to the relevant spatial dimension (paired t -test, \(p s<.05\) ). A \(2 \times 4\) (Phase \(\times\) Block) within-subjects ANOVA only showed a main effect for Block, \(F(2.68,80.37)=5.59, p\) <.001. Moreover, a marginal drop was demonstrated after block 4, which indicated a cost of attention, \(t(30)=1.83, p\) \(=.07\) (see Figure 2c).
In sum, both behavioral and eye gaze patterns indicated a cost of attention for participants who learned the first category. Both phases showed an evidence of attention optimization (i.e. increased accuracy, decreased RT, and increased looking time to the relevant dimension). The indication of attention optimization followed by a cost of attention was evident even though supervision was not provided as strong as in previous studies. (Note that explicit feedback was given after every trial in Hoffman \& Rehder (2010)).
(a) Accuracy at Test

(b) Reaction time at Test

(c) Looking time during Learning


Figure 2. Results from Experiment 1. (a) accuracy at Test, (b) reaction time at Test, and (c) looking time during Learning. The proportion of looking to the relevant dimension are weighted values in that the dotted line at .5 indicate chance level of equally looking to the relevant and irrelevant dimensions. Note that all error bars represent +/one standard error.

\section*{Experiment 2}

Experiment 2 examined the cost of attention when an extradimensional shift occurred between two dense categories with supervision.

\section*{Methods}

Participants Forty-two adults with normal or corrected to normal vision participated in the experiment. In addition, one participant was excluded from the analysis due to not exceeding the learning criterion.
Stimuli \& Procedure The stimuli and procedure were identical to Experiment 1 except that dense categories were used. In contrast to sparse categories, dense categories had two category-relevant spatial dimensions instead of one (see Figure 1c). For category A and B, in addition to the bottomright relevant dimension, the upper-left location had a constant shape/color as the bottom-right location had. For category C and D , in addition to the left location, the upperright location had a constant shape/color identical as the left location.

\section*{Results}

The overall accuracy for the test blocks was \(.98, S D=.11\) (Phase 1: \(M=.97, S D=.12\), Phase 2: \(M=.98, S D=.11\) ), with all test trials being significantly higher than chance performance, \(p<.001\) (see Figure 3a). A \(2 \times 4\) (Phase \(\times\) Block) within-subjects ANOVA did not show any main effect or interactions ( \(p s>.05\) ). Moreover, there was no significant difference between block 4 and block 5, indicating the absence of cost.

The mean reaction time for all test blocks was 838 ms , \(S D=528 \mathrm{~ms}\) (Phase 1: \(M=860 \mathrm{~ms}, S D=385 \mathrm{~ms}\), Phase 2: \(M=838 \mathrm{~ms}, S D=528 \mathrm{~ms}\) ) (see Figure 3b). A \(2 \times 4\) (Phase \(\times\) Block) within-subjects ANOVA with RT only showed a main effect for Block, \(F(2.26,90.36)=6.86, p<.001\). Also, the difference between block 4 and block 5 was not significant ( \(p>.05\) ).

In a dense category, there were two relevant dimensions and four irrelevant dimensions. Therefore, the weighted proportion of looking to the relevant dimension was calculated by multiplying two to the numerator instead of five as in Experiment 1. The overall weighted proportion for all learning blocks was \(.65, S D=.23\) (Phase 1: \(M=.63, S D\) \(=.23\), Phase 2: \(M=.65, S D=.23\) ). All blocks showed a significantly higher proportion of looking to the relevant spatial dimension, paired t-test, ps \(<.005\) (see Figure 3c). A \(2 \times 4\) (Phase \(\times\) Block) within-subjects ANOVA did not show any main effects or interactions, \(p s>.05\). Also, a significant drop was not found between block 4 and 5, \(p\) \(>.05\).

The results show no evidence of cost for the looking time data. Also there was no evidence of attention optimization (i.e. increased looking to the relevant dimension). However the accuracy is very high compared to the sparse condition, indicating that learning the dense category was easier than learning sparse category. Therefore it is possible that attention optimization occurred quickly, and the cost of attention was weak early in the block.


Figure 3. Results from Experiment 2. (a) accuracy at Test, (b) reaction time at Test, and (c) looking time during Learning. The proportion of looking to the relevant dimension are weighted values in that the dotted line at .5 indicate chance level of equally looking to the relevant and irrelevant dimensions. Note that all error bars represent +/one standard error.

To capture the early attention optimization in block 1 a moving window of 3 trials were used to calculate the proportion of looking to the relevant dimension, instead of using the whole block. Then a one-sample t-test was conducted against the chance level of .5. Results show that attention optimization occurred around the window 3, which would be around the \(4^{\text {th }}\) trial and lasted throughout the block (see Figure 4a). The same method could be applied to Block 5 where the second category was introduced. Results show that attention optimization occurred around the window 3, which would be around the \(4^{\text {th }}\) trial (see Figure 4b).

On the other hand, the cost of attention could be captured by comparing the last trial of block 4 and the first trial of block 5 instead of comparing the whole block. Results showed marginally significant drop from the last trial of block 4 ( \(M=.59, S D=.44\) ) to the first trial of block 5 ( \(M\) \(=.43, S D=.36), p=.068\), indicating a cost of attention.

In sum, dense categories were learned quicker than the sparse categories (faster attention optimization), and the cost of selective attention was weaker.


Figure 4. Attention optimization of block 1, block 5 in Experiment 2. Note that the asterisks represent \(p<.05\), and all error bars represent \(+/\) one standard error.

Then what would have made dense categories have lesser cost and stronger attention optimization? One possibility is that since dense categories have multiple category-relevant dimensions, attention allocation is much more distributed than sparse categories. Therefore, with limited amount of attention there will be smaller attention allocated to a dimension in the dense categories than in the sparse categories (Sutherland \& Mackintosh, 1971), which would lead to an easier/faster attention shift to a newly relevant dimension. On the other hand, it could also be possible because dense categories have more category-relevant dimensions, and thus there is a higher probability of spotting a relevant dimension. In this case, one could perfectly learn the dense categories with attending only one dimension instead of both.


Figure 5. The distribution of looking time between the two category-relevant dimensions in Experiment 2. Values closer to 0 indicate looking equally to the two relevant dimensions, whereas values closer to 1 indicate looking to only one dimension in a trial.
To investigate the latter possibility, the distribution of looking time between the two relevant dimensions was calculated. For each trial, the proportion of looking to one of the dimensions was calculated, where .5 represents equal looking to both dimensions. Then the absolute difference from . 5 was taken. Therefore, the value close to . 5
represents looking to only one dimension, and 0 represents looking to both dimension. Figure 5 shows the calculated values across subjects by block. Results indicate that subjects relied on a single dimension in most of the trials when learning the dense categories.

\section*{General Discussion}

The current study manipulated category density in the course of supervised category learning. Results show that even though supervision was weaker than in previous studies using sparse categories, attention optimization and cost of attention were observed during category learning (Experiment 1). Moreover, the dense categories were learned faster than sparse categories, and even with a stronger attention optimization, dense categories (Experiment 2) had a weaker cost of attention.

In Experiment 1, sparse categories were learned with weaker supervision than in previous studies using similar sparse categories. Note that when the sparse categories used in the current experiment were presented without supervision, participants failed to learn them (Yim, Best, \& Sloutsky, 2011). Supervision in the current experiment consisted of a hint that there is one dimension that is consistently relevant. However, the majority of participants learned the category. Also compared to previous studies where feedback was given on every trial (Hoffman \& Rehder, 2010; Rehder \& Hoffman, 2005), supervision here was only given at the start of each block. However, attention optimization and cost of attention were observed.

First, attention optimization should be closely related to the specific supervision signal. Category learning has mainly assumed that error signals from feedback mediates selective attention (Blair, Watson, \& Meier, 2009; Kruschke, 2001). However, the current task does not provide any feedback. A possible explanation would be that the supervision helps reduce the hypothesis space for the participants. Although knowing that there will be only one relevant dimension does not provide direct error signal, it drastically reduces the hypothesis space of possible category-relevant information. Although the effects of supervised and unsupervised learning on category formation has been discussed (Gureckis \& Love, 2003; Love, 2002), the effects of various kinds of supervision has not been investigated systematically, which should be examined in future research.

Second, although it is known that attention optimization is a precursor of cost of attention, it is possible that the greater cost in the current study originates from the difference of density between the current and previous research. The stimuli in Hoffman \& Rehder (2010) had 2 out of 3 irrelevant dimensions whereas the current study has 5 out of 6 irrelevant dimensions. The sparser the category is the harder it would be to learn the relevant dimension. However, once selective attention is engaged, the cost would be greater for sparser categories. This is because there are more irrelevant dimensions in a sparser category, which means that there will be more unattended dimensions
during learning (i.e. learned inattention). Therefore, when an extra-dimensional shift occurs, the probability of figuring out a newly relevant dimension among the previously irrelevant dimensions will be lower than in a less sparse category. Although it is not possible to directly examine this hypothesis from the current study, the relationship between category density and cost of attention could be examined with controlling the amount of attention optimization through manipulating the number of irrelevant dimensions.

In Experiment 2, most of the participants optimized to one dimension instead of distributing their attention to all relevant dimensions (see Figure 5). Although the categories used in the current study are deterministic and do not require an information integration process (Ashby, AlfonsoReese, Turken, \& Waldron, 1998), there is evidence that adults distribute their attention to all relevant dimensions when learning dense categories that had a similar category structure as the current one (Kloos \& Sloutsky, 2008). One main difference between the previous study and the current study is the presentation time during learning. In Kloos \& Sloutsky (2008), participants observed the category exemplars in a self-paced maner, whereas the current study presented the exemplars for 1.5 sec . Since the category could be learned by using both distributed and non-distrubuted attention, it is highly possible that the fast presentation time leaded the participants to attend to only one dimesion.

Finally, the results may have implications for understanding the development of category learning. Since it is known that children gradually gain the ability to selectively attend (Hanania \& Smith, 2010), it would be hard for them to learn sparse categories, which requires the ability to selectively attend to a small number of categoryrelevant dimensions. Therefore, the role of supervision would be crucial for learning spare categories early in development. If the interaction among the category structure, learning regime, and category learning is well established, it would help to understand the developmental trajectory of category learning.

\section*{Acknowledgments}

This research was supported by NSF (BCS-0720135); IES (R305B070407); and NIH (R01HD056105) to V. M. Sloutsky.

\section*{References}

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., \& Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. Psychological Review, 105, 442-481.
Blair, M. R., Watson, M. R., \& Meier, K. M. (2009). Errors, efficiency, and the interplay between attention and category learning. Cognition, 112(2), 330-336.
Gureckis, T. M., \& Love, B. C. (2003). Human unsupervised and supervised learning as a quantitative distinction. International Journal of Pattern Recognition and Artificial Intelligence, 17(5), 885-901.

Hall, G. (1991). Perceptual and Associative Learning. New York, New York, USA: Oxford University Press.
Hanania, R., \& Smith, L. B. (2010). Selective attention and attention switching: towards a unified developmental approach. Developmental Science, 13(4), 622-635.
Hoffman, A. B., \& Rehder, B. (2010). The Cost of Supervised Classification: The Effect of Learning Task on Conceptual Flexibility. Journal of Experimental Psychology: General, 139(2), 319-340.
Kloos, H., \& Sloutsky, V. M. (2008). What's Behind Different Kinds of Kinds: Effects of Statistical Density on Learning and Representation of Categories. Journal of Experimental Psychology: General, 137(1), 52-72.
Kruschke, J. K. (1992). ALCOVE: An Exemplar-Based Connectionist Model of Category Learning. Psychological Review, 99(1), 22-44.
Kruschke, J. K. (2001). Toward a Unified Model of Attention in Associative Learning. Journal of Mathematical Psychology, 45, 812~863.
Kruschke, J. K., \& Blair, N. J. (2000). Blocking and backward blocking involve learned inattention. Psychonomic Bulletin \& Review, 7, 636-645.
Love, B. C. (2002). Comparing supervised and unsupervised category learning. Psychonomic Bulletin \& Review, 9(4), 829-835.
Mackintosh, N. J. (1975). A Theory of Attention: Variations in the Associability of Stimuli with Reinforcement. Psychological Review, 82(4), 276-298.
Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. Journal of Experimental Psychology: General, 115(1), 39-61.
Rehder, B., \& Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. Cognitive Psychology, 51(1), 1-41.
Salvucci, D. D., \& Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. Proceedings of the symposium on Eye tracking research \& applications - ETRA '00, 71-78.
Shepard, R. N., Hovland, C. I., \& Jenkins, H. M. (1961). Learning and memorization of classifications. Psychological Monographs, 75(13), 1-41.
Sloutsky, V. M. (2010). From perceptual categories to concepts: What develops? Cognitive Science, 34, 12441286.

Sutherland, N. S., \& Mackintosh, N. J. (1971). Mechanisms of animal discrimination learning. New York: Academic Press.
Yim, H., Best, C. A., \& Sloutsky, V. M. (2011). Cost of Attention as an Indicator of Category Learning. In L. Carlson, C. Hölscher, \& T. Shipley (Eds.), Proceedings of the 33rd Annual Conference of the Cognitive Science Society (pp. 1724-1729). Austin, TX.

\title{
Do walking pedestrians stabily interact inside a large group? Analysis of group and sub-group spatial structure
}

\section*{Francesco Zanlungo (ZANLUNGO@Atr.Jp)}

\author{
Takayuki Kanda (KANDA@Atr.Jp) \\ Intelligent Robotics and Communication Laboratories, ATR, Kyoto, Japan \& \\ JSPS CREST, Tokyo, Japan
}

\begin{abstract}
We combine video recording and laser range tracking to analyse the geometrical structure of groups of walking pedestrians socially interacting. By recording their relative position and observing their social interaction for a large enough time span we can analyse the stability and universality of their spatial structure. We find that while 2-pedestrian and 3-pedestrian groups have a relatively "time stable" and "universal" geometrical structure (an abreast formation for pairs, and a "V" formation for triads, with the central pedestrian walking slightly behind), no such structure emerges for larger groups. Nevertheless, these larger groups result to be composed of time stable two or three people sub-groups with the same "universal" geometrical structure of isolated pairs and triads.
\end{abstract}

Keywords: Group dynamics; proxemics.

\section*{Introduction}

The spatial relationship of socially interacting people, i.e proxemics, has been largely studied, starting from the seminal works of (Hall, 1969) and (Kendon, 1990) in which the distances between and spatial distribution of people participating in social activities have been investigated. At the same time other researchers have investigated the size of social groups (by size of a group we mean the number of its components) and the probability distribution of these sizes (James, 1953; Coleman \& James, 1961). Many of the aforementioned studies are based on "ecological" observations, i.e. studies in which people are observed in their natural environment while reducing as much as possible the effect of observations on their behaviour. While these studies are obviously based on observations of people behaviour in public spaces, until recently they did not focus on one of the main components of public spaces population (at least in modern urban areas), i.e. walking pedestrians. Here by pedestrian we mean a person in a public space moving between two locations for practical or recreational purposes, or even "wandering around" an environment without any particular goal. Pedestrians are often part of social groups with a specific proxemics determined by their dynamical constraints (the fact that they are walking), but the study of these groups has been traditionally made difficult by the fact that they are moving and located inside a crowd, which makes the observation of their behaviour more troublesome. Nevertheless, lately a few works have focused on the behaviour of these groups (Moussaïd, Perozo, Garnier, Helbing, \& Theraulaz, 2010; Costa, 2010), due also to the growing interest in crowd behaviour of which groups are a non negligible component (Aveni, 1977). This interest is due
to the necessity of simulating crowd behaviour to design better pedestrian facilities (Helbing, Farkas, Molnar, \& Vicsek, 2002), but also to reproduce faithfully the behaviour of virtual crowds for the entertainment industry (Karamouzas \& Overmars, 2012).
While (Moussaïd et al., 2010) report that the spatial structure of a freely walking (i.e. not environmentally constrained) \(n\) pedestrian group is a line of abreast walking pedestrians, that tends to be bent into "V" or "U" formations (i.e., the pedestrians on the sides walk ahead) when the crowd density grows, (Costa, 2010) reports different spatial structures, suggesting for example that the " V " structure is the most occurring one for three people groups (regardless of crowding), and that larger groups tend to split into smaller sub-groups. Nevertheless (Costa, 2010) does not analyse the possible effects of environmental constraints on observed behaviours (the width of the sidewalks pedestrians were observed in was comparable to the group spatial sizes), and does not provide a quantitative study of 2D space structures, nor follows groups for a time interval long enough to analyse their change in time.
The difference between these observations leads us to two related problems in walking group proxemics, to which we try to bring insight in this work:
- Do n-pedestrian groups (i.e. groups composed of \(n\) members) have a prevalent geometrical structure? Here by prevalent we mean universal (common to almost all groups, or at least present in a large majority of them) and time stable (i.e. the positions of pedestrians in an unconstrained group will be given by small oscillations around those of the prevalent structure).
- If such an overall structure does not exist for a \(n\)-pedestrian group, is it possible to find it at the sub-group level?

In order to analyse these issues, we have to observe pedestrians in a situation in which collision avoiding and environmental constraints are not very strong (otherwise it would be impossible to identify the "universal" structure). Furthermore, we have to combine the need to measure with good detail (and for long enough time) the position of pedestrians, with that of observing their social interactions, in order to analyse the group structure. If a large pedestrian group is divided into smaller sub-groups we may expect social interaction inside sub-groups to be more frequent than between different sub-groups, and for this reason in many cases the
belonging of sub-groups to a larger group structure may be determined only if the observation is long enough. To attain this goal we combine a laser range finder tracking technology, that allows us to determine with good precision the position and velocity of each pedestrian in a large public environment, with frontal view (face level) video recordings, that allow us to analyse their social interactions (Fig. 1). As a result, we can follow pedestrian groups for a relatively long time while examining both their social and spatial interactions from a (respectively) qualitative and quantitative point of view. We performed these observations in a large area completely dedicated to pedestrian motion, and in a location and time in which the pedestrian density was relatively low, in order to be able to observe the behaviour of "unconstrained" groups.


Figure 1: Video camera frame of the experimental area. A sensor pole is visible in the bottom-right corner.

\section*{Methodology and definitions}

\section*{Data collection}

We tracked pedestrian motion in two areas of a pedestrian underground facility in Umeda, Osaka (Japan), for a total time of 6 hours in each area. The pedestrian areas consist of a few corridors connecting a railway station to a shopping mall, each area being around \(500 \mathrm{~m}^{2}\). The environments are described in detail in (Zanlungo, Chigodo, Ikeda, \& Kanda, 2012; Zanlungo, Ikeda, \& Kanda, 2012), and the pedestrian tracking data are available at (Zanlungo, 2012). The average pedestrian density in the environment resulted to be \(\approx 0.03\) pedestrians per square meters while the width of the corridors varied between 4 and 7 meters, meaning that the average distance from a pedestrian to another pedestrian outside their group, or to a wall, is expected to be larger than the spatial size of the group (for example a group of 4 people walking in an abreast formation, assuming an interacting distance of 1 meter between first neighbours, should be 3 meters wide, compared to an expected distance between pedestrians \(>5\) meters at such a density). We can thus assume pedestrians to be fairly "unconstrained" by the environment and freely walk in their preferred spatial formation for most of the time.
We used 16 Hokuyo UTM sensors (situated on poles close to the environment walls not to hinder pedestrian motion, Fig. 1) and the tracking algorithm (Glas, Miyashita, Ishiguro, \& Hagita, 2009) to determine pedestrian positions at times intervals \(\delta t \approx 50 \mathrm{~ms}\) with precision \(\approx 50 \mathrm{~mm}\). We smoothed the tracked positions on time windows \(\Delta t=500 \mathrm{~ms}\), to fur-
ther improve the tracking precision. Pedestrian velocity is computed as the ratio of the displacement vector between two (smoothed) consecutive tracking positions (eq. 2), and has an expected precision \(\approx 50-100 \mathrm{~mm} / \mathrm{s}\). As we will see (see also the discussion in (Zanlungo, Chigodo, et al., 2012)) this tracking precision is negligible with respect to the typical interaction distances and velocities of pedestrians.
We also video recorded each experimental area with two different "frontal view" cameras (Fig. 1), located in such a way to allow observing the social interaction between the pedestrians for a sufficient long time (pedestrians are usually tracked and observed for a time of the order of tens of seconds). This camera based observation of social interaction was possible because the cameras are not needed for tracking and the density was relatively low. A "coder" (a non-technical staff member of our laboratory), was asked to identify all the pedestrian social groups in the environment and their members. In order to do that, she was asked to use all the information available from the videos, such as relative position, coherent motion, and social clues including conversation, gaze exchange and even age, sex and clothing (for example she identified a relatively large flock of coherent moving people as a single group because they were all dressed for and carrying similar hiking equipment). She was asked to identify only groups of which she could establish the nature without any reasonable doubt (i.e. she was asked to strongly avoid false positives, while false negatives were allowed). Furthermore, the coder was asked to annotate the groups, and the individuals in each group, for which she could without any doubt identify explicit social interaction clues (namely conversation, or explicit gaze exchange). Table 1 shows the number and size of labelled groups, distinguishing between "fully connected" groups (FCG) for which she could observe explicit social interaction between all the members, and "disconnected" groups (DG) that seemed to be related on the basis of other visual clues but for which explicit interaction could not be observed (or was observed only in smaller subgroups). To avoid false positives, only FCG are analysed in this paper. (The coder identified also six 5-pedestrian groups, six 6-pedestrian groups, one 7-pedestrian group and one 18pedestrian group, not analysed in this work due to the small size of the samples).
\begin{tabular}{|c|c|c|c|}
\hline Size & \(n=2\) & \(n=3\) & \(n=4\) \\
\hline FCG & 1126 & 114 & 17 \\
\hline DG & 91 & 34 & 14 \\
\hline
\end{tabular}

Table 1: Observed fully connected (FDG) and disconnected groups (DG) for each group size \(n\).

\section*{Definitions}

In order to identify the existence of an "universal" structure in a pedestrian group, it is necessary to study it in the correct reference frame. The most natural candidate is the group's
"centre of mass frame" (see Fig. 2). Let us call
\[
\begin{equation*}
\mathbf{x}^{e}\left(t_{k}\right)=\left(x^{e}\left(t_{k}\right), y^{e}\left(t_{k}\right)\right) \quad t_{k}=k \Delta t \tag{1}
\end{equation*}
\]
the position of a pedestrian in the "environment" reference frame, smoothed on regular \(\Delta t=500 \mathrm{~ms}\) time windows, and
\[
\begin{equation*}
\mathbf{v}^{e}\left(t_{k}\right) \equiv \frac{\mathbf{x}^{e}\left(t_{k+1}\right)-\mathbf{x}^{e}\left(t_{k}\right)}{\Delta t} \tag{2}
\end{equation*}
\]
the corresponding pedestrian velocity. Let us consider a \(n\) pedestrian (from now on \(n\)-p) (sub-)group with position and velocities
\[
\left\{\mathbf{x}_{i}^{e}\left(t_{k}\right), \mathbf{v}_{i}^{e}\left(t_{k}\right)\right\} \quad i=1, \ldots, n,
\]
and define the group "centre of mass" position and velocity
\[
\begin{equation*}
\mathbf{X}\left(t_{k}\right)=\frac{\sum_{i} \mathbf{x}_{i}^{e}\left(t_{k}\right)}{n} \quad \mathbf{V}\left(t_{k}\right)=\frac{\sum_{i} \mathbf{v}_{i}^{e}\left(t_{k}\right)}{n} \tag{3}
\end{equation*}
\]

Let us name GCMF the group centre of mass frame (at time \(\left.t_{k}\right)\) with origin in \(\mathbf{X}\left(t_{k}\right)\) and the \(y\) axis aligned to \(\mathbf{V}\left(t_{k}\right)\), i.e. with axis versors
\[
\begin{equation*}
\hat{\mathbf{e}}_{x}\left(t_{k}\right)=\left(\frac{V_{y}\left(t_{k}\right)}{V\left(t_{k}\right)},-\frac{V_{x}\left(t_{k}\right)}{V\left(t_{k}\right)}\right) \quad \hat{\mathbf{e}}_{y}\left(t_{k}\right)=\frac{\mathbf{V}\left(t_{k}\right)}{V\left(t_{k}\right)} \tag{4}
\end{equation*}
\]
(from now on we remove \(t\) from notation for simplicity's sake). The position of pedestrian \(i\) in the GCMF is then \(\mathbf{x}_{i}=\left(x_{i}, y_{i}\right)\) with
\[
\begin{equation*}
x_{i}=\left(\mathbf{x}_{i}^{e}-\mathbf{X}\right) \cdot \hat{\mathbf{e}}_{x} \quad y_{i}=\left(\mathbf{x}_{i}^{e}-\mathbf{X}\right) \cdot \hat{\mathbf{e}}_{y} \tag{5}
\end{equation*}
\]

We also define the polar coordinates \(\left(r_{i}, \theta_{i}\right)\) through
\[
\begin{equation*}
x_{i}=r_{i} \sin \theta_{i} \quad y_{i}=r_{i} \cos \theta_{i} \tag{6}
\end{equation*}
\]
where \(r=\sqrt{x^{2}+y^{2}}\) represents the distance of the pedestrian from the centre of mass (for 2-p groups \(2 r\) is the distance between pedestrians). Fig. 2 illustrates the previously defined quantities in the 2-p case, for which the following holds
\[
\begin{equation*}
r_{2}=r_{1}, \quad x_{2}=-x_{1}, \quad y_{2}=-y_{1}, \quad \theta_{2}=\pi+\theta_{1} \tag{7}
\end{equation*}
\]

It is important to quantitatively study the structure of the


Figure 2: GCMF variables definition.
group in the GCMF, because the distinctive feature of walking groups is the presence of a centre of attention, namely the
direction towards their current sub-goal, which is identified by the velocity of the group. The geometrical structure of the pedestrian groups is determined by the necessity to maintain the focus on the walking direction, and for this reason we will not consider non-moving pedestrian groups (i.e., data points in which at least a group member has \(v_{i}\left(t_{k}\right)<500 \mathrm{~ms}\), a threshold that corresponds to a velocity 3 standard deviations smaller than the typical pedestrian velocity (Daamen \& Hoogendoorn, 2006); see also (Zanlungo, Chigodo, et al., 2012) for a discussion of this threshold). Since we collected data in a passing point between a station and a shopping centre without attraction points (Zanlungo, Chigodo, et al., 2012), only \(\approx 5 \%\) of data are not considered.

It is clear that if a universal and time stable n-p formation exists, then at (almost) all times and for (almost) all groups, the GCMF pedestrian positions should be close to those determined by such a structure. We will compute the empirical 2D probability distribution function (pdf) \(\rho(x, y)\) for each \(n\) averaging on all groups, pedestrians \(i\) and and times \(t_{k}\), and state that such a universal and stable formation exists only if \(\rho(x, y)\) has \(n\) well defined maxima. The formation will be then empirically determined by the position of these maxima.

\section*{Results}

\section*{Whole group GCMF structure for n-p groups}

Fig. 3 shows the pedestrian pdf \(\rho(x, y)\) for \(2-\mathrm{p}, 3-\mathrm{p}\) and \(4-\mathrm{p}\) groups. The 2-p and 3-p groups have a well defined geometrical structure in the GCMF, i.e. their \(\rho\) has, respectively, 2 and 3 well defined maxima, one for each pedestrian. Such a structure is not present for 4-p groups, whose pdf has many local maxima. As we will see, a well defined structure emerges for 4-p only after the whole group is properly divided in two 2-p sub-groups.

\section*{2-p groups}

Let us identify the leftmost \((x<0)\) pedestrian as \(P_{1}\) and the rightmost one as \(P_{2}\). Figs. \(4 \mathbf{a}\) ) and \(4 \mathbf{b}\) ) respectively show the \(\rho\left(r_{1}\right)\) and \(\rho\left(\theta_{1}\right)\) pdfs, while Table 2 shows the average values and standard deviations of all variables. While \(\rho\left(y_{1}\right)\) and \(\rho\left(\theta_{1}\right)\) are well described by a Normal distribution (i.e. a von Mises (Mardia \& Jupp, 2009) one for the circular variable \(\theta\) ), \(\rho\left(x_{1}\right)\) and \(\rho\left(r_{1}\right)\) are not, and for this reason we report also the (approximate) value of their maxima. Our data
\begin{tabular}{|c|c|c|c|c|}
\hline & \(\rho\left(x_{1}\right)\) & \(\rho\left(y_{1}\right)\) & \(\rho\left(r_{1}\right)\) & \(\rho\left(\theta_{1}\right)\) \\
\hline\(<>\) & -387 & -2 & 417 & -89 \\
\hline\(\sigma\) & 87 & 166 & 105 & 17 \\
\hline \(\max\) & -360 & 0 & 365 & -90 \\
\hline
\end{tabular}

Table 2: Average values \((<>)\), standard deviations ( \(\sigma\) ) and maxima for the \(\rho\left(x_{1}\right), \rho\left(y_{1}\right), \rho\left(r_{1}\right)\) and \(\rho\left(\theta_{1}\right)\) pdfs (linear variables in mm , circular in degrees).
show that 2-p groups walk abreast at a distance smaller than twice the average shoulder width (Pheasant, 1986). Such a


Figure 3: \(\rho(x, y)\) for 2-p, 3-p and 4-p groups (respectively, from the left). Blue corresponds to maximum density, red to minimum density (colour bar on the right). Each figure covers a \(2 \times 2\) meters area.
configuration is determined by the need of maintaining both partners in each other's field of view (or better at the border of it) while keeping the main attention focus on the walking direction. By walking abreast the partner is reachable for gaze exchange and conversation through a torsion of the neck (pedestrians can go on walking towards their goal with no gait modification) while the distance allows for conversation without collision or excessive proximity. This configuration is the most comfortable one for walking and interacting socially, but it cannot be extended to larger groups, because in a \(n>2\) abreast configuration the position of first neighbours would hinder gaze contact and conversation with second or larger neighbours. As clear from Fig. 3 and discussed below, this affects larger group configurations.

\section*{3-p groups}

Let us name the pedestrians \(P_{1}, P_{2}\) and \(P_{3}\) starting from the leftmost to the rightmost ( \(x_{1}<x_{2}<x_{3}\) ). It is easier to understand the relation between 2-p and 3-p group structures if we analyse the 3-p variables in all possible 2-p sub-group GCMFs. A variable with subscript \(i j\) will denote the position of \(P_{i}\) in the \(\left(P_{i}, P_{j}\right)\) GCMF. In this way, for example, \(2 r_{12}\) is the relative distance between the leftmost pedestrian and the central one (first neighbours), and so on for each variable and pedestrian pair \(i, j\) with \(i<j\) (the \(i>j\) case can be obtained through eq. 7). Tables 3 and 4 report the values of all such variables, while Fig. 5 a) compares the \(\rho\left(r_{12}\right)\) and \(\rho\left(r_{13}\right)\) pdfs to the \(\rho\left(r_{1}\right)\) distribution of the 2-p case. The same comparison is performed for \(\theta\) in Fig. 5 b ).
The first neighbour distance distributions \(\rho\left(r_{12}\right), \rho\left(r_{23}\right)\) are
\begin{tabular}{|c|c|c|c|c|c|c|}
\hline & \(\rho\left(r_{12}\right)\) & \(\rho\left(r_{13}\right)\) & \(\rho\left(r_{23}\right)\) & \(\rho\left(\theta_{12}\right)\) & \(\rho\left(\theta_{13}\right)\) & \(\rho\left(\theta_{23}\right)\) \\
\hline\(<>\) & 437 & 738 & 441 & -74 & -89 & -105 \\
\hline\(\sigma\) & 125 & 132 & 169 & 23 & 15 & 25 \\
\hline \(\max\) & 365 & 700 & 365 & -80 & -90 & -110 \\
\hline
\end{tabular}

Table 3: Average values \((<>)\), standard deviations ( \(\sigma\) ) and maxima of the pdf \(\rho\) for the polar 3-p group variables ( \(r\) in \(\mathrm{mm}, \theta\) in degrees).
very similar to the 2-p distance distribution, while the second neighbour distance distribution \(\rho\left(r_{13}\right)\) may be very well rep-
\begin{tabular}{|c|c|c|c|c|c|c|}
\hline & \(\rho\left(x_{12}\right)\) & \(\rho\left(x_{13}\right)\) & \(\rho\left(x_{23}\right)\) & \(\rho\left(y_{12}\right)\) & \(\rho\left(y_{13}\right)\) & \(\rho\left(y_{23}\right)\) \\
\hline\(<>\) & -342 & -686 & -344 & 114 & 1 & -112 \\
\hline\(\sigma\) & 104 & 155 & 112 & 257 & 278 & 263 \\
\hline \(\max\) & -350 & -680 & -350 & 70 & 0 & -110 \\
\hline
\end{tabular}

Table 4: Average values \((<>)\), standard deviations ( \(\sigma\) ) and maxima of the pdf \(\rho\) for the Cartesian 3-p group variables (in mm ).


Figure 4: a): 2-p pdf for \(r_{1}\). b): 2-p pdf for \(\theta_{1}\), compared to a best fit von Mises distribution.


Figure 5: a): \(\rho\left(r_{12}\right)\) distribution (blue) compared to the \(\rho\left(r_{13}\right)\) distribution in green and to the 2-p \(\rho\left(r_{1}\right)\) distribution in orange. b): \(\rho\left(\theta_{12}\right)\) distribution (blue) compared to the \(\rho\left(\theta_{13}\right)\) distribution in green and to the 2-p \(\rho\left(\theta_{1}\right)\) distribution in orange.
resented by a Normal one. Defining \(\bar{r}\) as the value \(r\) for which \(\rho(r)\) is maximum, we see that \(\bar{r}_{13} \approx 2 \bar{r}_{12} \approx 2 \bar{r}_{23} \approx 2 \bar{r}_{1}\), i.e. 3-p groups members try to maintain between them the same distances that occur between 2-p groups members, which we interpret as a strong sign of social interaction involving all
three members. The similarity between the \(\rho\left(\theta_{13}\right)\) and the (2-p) \(\rho\left(\theta_{1}\right)\) distributions is particularly striking, suggesting direct social interaction between \(P_{1}\) and \(P_{3}\). This interaction would hardly be possible if we had the same angle distribution for \(\theta_{12}\) and \(\theta_{23}\), because the central pedestrian would hinder the communication. As a result, the central pedestrian steps slightly back (so that their partners remain in the vision field), and the angle between the three of them is, in average, \(\approx 150\) degrees. This "V" configuration had already been reported as the most often occurring one for walking triads (Costa, 2010). (Moussaïd et al., 2010) explain this configuration as the effect of a trade-off between easiness of communication and collision avoiding efficiency, assuming that the free-walking triads walk abreast. On the basis of our data, that as we already stated should not be strongly influenced by environmental constraints, we suggest that the "V" configuration is attained for maximum easiness of communication between the three partners, and occurs even for freely walking pedestrians.

\section*{4-p groups}
(Moussaïd et al., 2010) report that freely walking 4-p groups assume an abreast configuration, that tends to bend in a "U" one with growing pedestrian density, in order to avoid collisions. According to this abreast hypothesis, we should see four clear maxima in a row in Fig. 3 on the right. Furthermore, if we name the pedestrians \(P_{1}, \ldots, P_{4}\) with \(x_{1}<x_{2}<\) \(x_{3}<x_{4}\) in the GCMF, the first neighbour variable distributions, such as
\[
\begin{equation*}
\rho_{f n}(\theta)=\left(\rho\left(\theta_{12}\right)+\rho\left(\theta_{23}\right)+\rho\left(\theta_{34}\right)\right) / 3 \tag{8}
\end{equation*}
\]
and the analogously defined \(\rho_{f n}(r), \rho_{f n}(x)\) and \(\rho_{f n}(y)\), should resemble the 2-p group distributions. This hypothesis is clearly not supported by our data (Figs. 6, \(7 \mathbf{b}\) )). On the opposite, (Costa, 2010) reports different geometrical structures for 4-p groups, but none of these seems prevalent in our data (no clear maxima in Fig. 3 right).
We may then use a different sub-group hypothesis, assuming that the 4-p group may be divided in two sub-groups of 2 pedestrians, with "strong interaction" inside the sub-group and weaker interaction outside it. According to this hypothesis, we may find a universal and time stable structure only at the sub-group level. Let us rename the pedestrians in the following way. We name \(P_{1}\) the pedestrian with the minimum \(x\) value in the 4-p GCMF, and compute the point
\[
\begin{equation*}
\mathbf{p}_{2}=\mathbf{x}_{1}+r_{\text {int }} \hat{\mathbf{e}}_{x} \tag{9}
\end{equation*}
\]
where \(r_{\text {int }}=730 \mathrm{~mm}\) is the maximum for the pdf of distances for 2-p groups (i.e. twice the GCMF \(r\) value reported in Table 2). We then name \(P_{2}\) the pedestrian whose euclidean distance to \(\mathbf{p}_{2}\) is minimum, and \(P_{3}\) and \(P_{4}\) the remaining two. Let us finally name \(\rho_{s g}(r), \rho_{s g}(x), \rho_{s g}(y)\) and \(\rho_{s g}(\theta)\) the pdfs for the corresponding variables when averaged over all subgroups, as in
\[
\begin{equation*}
\rho_{s g}(r)=\left(\rho\left(r_{12}\right)+\rho\left(r_{34}\right)\right) / 2 \tag{10}
\end{equation*}
\]

The average values and standard deviations for these distributions are shown in Table 5, while Fig. 6 shows the \(\rho_{s g}(x, y)\) 2D pdf, presenting two clear maxima. Fig. 7 a) shows the comparison between the pdf \(\rho_{s g}(r)\) with the (2-p) \(r_{1}\) distribution. Fig. 7 b) performs the same comparison for \(\theta\) variables, including also the \(\rho_{f n}(\theta)\) (abreast hypothesis) distribution. The presence of two clear maxima in \(\rho_{s g}(x, y)\) suggests
\begin{tabular}{|c|c|c|c|c|}
\hline & \(\rho_{s g}(x)\) & \(\rho_{s g}(y)\) & \(\rho_{s g}(r)\) & \(\rho_{s g}(\theta)\) \\
\hline\(<>\) & -403 & -54 & 530 & -97 \\
\hline\(\sigma\) & 180 & 347 & 195 & 33 \\
\hline \(\max\) & -360 & 0 & 370 & -90 \\
\hline
\end{tabular}

Table 5: Average values \((<>)\), standard deviations \((\sigma)\) and maxima of the pdfs \(\rho_{s g}(x), \rho_{s g}(y), \rho_{s g}(r)\) and \(\rho_{s g}(\theta)\) (linear variables in mm , circular in degrees).


Figure 6: \(\rho_{s g}(x, y)\) under the sub-groups hypothesis. The figure covers a \(2 \times 2\) meters area.


Figure 7: a): Comparison between the 2-p pdf for \(\rho\left(r_{1}\right)\) in orange and \(\rho_{s g}\left(r_{1}\right)\) (sub-group hypothesis) in blue. b): Comparison between the 2-p pdf for \(\rho\left(\theta_{1}\right)\) in orange; \(\rho_{f n}(\theta)\) (abreast hypothesis) in red; and \(\rho_{s g}\left(\theta_{1}\right)\) (sub-group hypothe\(s i s)\) in blue.
that a universal and stable structure is indeed present at the sub-group level. The \(\rho_{s g}(r), \rho_{s g}(x), \rho_{s g}(y)\) and \(\rho_{s g}(\theta)\) distributions in a 4-p group result to be a "perturbed version" of the "proper" 2-p variable distributions, the perturbation being determined by the interaction with the members of the other sub-group. We can give different interpretations for the absence of a universal 4-p spatial configuration, that probably act as con-causes. Since the completely abreast configuration results to be uncomfortable even in the 3-p configuration, it
results to be even more problematic for 4 pedestrians. A solution is to, in a way similar to the "V" 3-p configuration, walk in a "U" or in a "staggered" configuration. Another solution would be to walk on two different abreast rows, for example roughly on the corners of a square. While this configuration has the shortcoming that the members on the back are not in the field of view of, nor can be comfortably watched by, the pedestrians on the front, it has the strong point that reduces the maximum distance between members of more than a factor 2 , and it is more efficient in case of collision avoiding or other environmental influences. If we observe particular groups for a short time, as done by (Costa, 2010), we may observe all these occurrences, but, as shown by Fig. 3 (right) none of them is universal and time stable. What is stable (Fig. 6) is the association of pedestrians in pairs. This does not mean that this pairwise association is invariant (i.e. that the pair composition does not change in time) but the data suggest that this association is far more stable than the wholegroup structures. Even though we do not report a quantitative analysis, we qualitatively observed a few 5-p and 6-p groups, and noticed also for these groups the tendency to part in stable 2-p or 3-p sub-groups.

\section*{Conclusions and future work}

Our observations lead us to the conclusion that the dynamical constraints make social interaction between members of walking groups difficult to attain for subgroups larger than two units. For this reason, larger groups have a tendency to form relatively stable 2-p subgroups. Obviously, since odd-p groups cannot be divided in pairs without excluding a pedestrian, for 3-p (sub-)groups we find a configuration almost as stable as the 2-p one.
Regarding the difference between our observations and those of (Moussaïd et al., 2010), we could speculate on cultural and environmental features. Nevertheless we believe that the main difference may be operational, i.e. that our experimental setting using cameras and laser sensors allowed us to identify sub-groups of pedestrians as part of larger groups, even if their interaction was limited in time. On the contrary, observation methods based on shorter time windows may be biased towards large spatial configurations, because they would tend to consider sub-groups as separate entities. For this reason, despite these differences with previous works, and the limited amount of data for large groups, we may speculate that our work is universal in showing that 2-p and 3-p sub-groups are far more stable than larger configurations, with possible effects on the behaviour of crowds.
In our future work we plan to analyse the dynamical features that have been ignored in this work. In detail we want to study: 1) how pedestrians in (sub-)groups behave away from the equilibrium configuration, and 2) the time stability of, and interaction between, sub-groups inside a larger group. Such an analysis should bring further insight on the social dynamics of walking pedestrians and allow us to extend pedestrian group models in such a way to describe the findings of this
work.

\section*{Acknowledgements}

We thank Tetsushi Ikeda for the organisation of the data collection experiments and Kanako Tomita for labelling the social interactions of pedestrians from video recordings. We also thank the anonymous reviewers whose comments helped in improving the quality of our paper.

\section*{References}

Aveni, A. (1977). The not-so-lonely crowd: Friendship groups in collective behavior. Sociometry, 96-99.
Coleman, J., \& James, J. (1961). The equilibrium size distribution of freely-forming groups. Sociometry, 24(1), 3645.

Costa, M. (2010). Interpersonal distances in group walking. Journal of Nonverbal Behavior, 34(1), 15-26.
Daamen, W., \& Hoogendoorn, S. (2006). Free speed distributions for pedestrian traffic. In TRB-annual meeting.
Glas, D., Miyashita, T., Ishiguro, H., \& Hagita, N. (2009). Laser-based tracking of human position and orientation using parametric shape modeling. Advanced robotics, 23(4), 405-428.
Hall, E. (1969). The hidden dimension. Anchor Books.
Helbing, D., Farkas, I., Molnar, P., \& Vicsek, T. (2002). Simulation of pedestrian crowds in normal and evacuation situations. Pedestrian and evacuation dynamics, 21.
James, J. (1953). The distribution of free-forming small group size. American Sociological Review;
Karamouzas, I., \& Overmars, M. (2012). Simulating and evaluating the local behavior of small pedestrian groups. Visualization and Computer Graphics, IEEE T., 18(3), 394-406.
Kendon, A. (1990). Conducting interaction: Patterns of behavior in focused encounters (Vol. 7). Cambridge.
Mardia, K., \& Jupp, P. (2009). Directional statistics (Vol. 494). Wiley.
Moussaïd, M., Perozo, N., Garnier, S., Helbing, D., \& Theraulaz, G. (2010). The walking behaviour of pedestrian social groups and its impact on crowd dynamics. PLoS One, 5(4), e10047.
Pheasant, S. (1986). Bodyspace: anthropometry and design. London: Taylor and Francis.
Zanlungo, F. (2012). Available from https:// sites.google.com/site/francescozanlungo/ pedestriandata
Zanlungo, F., Chigodo, Y., Ikeda, T., \& Kanda, T. (2012). Experimental study and modelling of pedestrian space occupation and motion pattern in a real world environment. Pedestrian and evacuation dynamics.
Zanlungo, F., Ikeda, T., \& Kanda, T. (2012). A microscopic social norm model to obtain realistic macroscopic velocity and density pedestrian distributions. PloS one, 7(12), e50720.

\title{
Unstable Dynamics of Intrinsically Motivated Learning
}

\author{
Arkady Zgonnikov (arkady.zgonnikov@gmail.com) \\ University of Aizu, Tsuruga, Ikki-machi, Aizu-wakamatsu, 965-8580 Fukushima, Japan \\ Ihor Lubashevsky (i-lubash@u-aizu.ac.jp) \\ University of Aizu, Tsuruga, Ikki-machi, Aizu-wakamatsu, 965-8580 Fukushima, Japan
}

\begin{abstract}
Employing the dynamical systems framework, we study the effects of intrinsic motivation on the dynamics of the learning processes. The intrinsic motivation here is the one's desire to learn not because it may cause some benefits in future, but due to the inherent joy obtained by the very process of learning. We study a simple example of a single agent adapting to unknown environment; the agent is biased by the desire to select the actions she has little information about. We show that intrinsic motivation may cause the instability of the learning process that is stable in the case of rational agent. Therefore, we suggest that the effects of human intrinsic motivation in particular and the irrationality in general may be of exceptional importance in complex sociopsychological systems and deserve much attention in the formal models of such systems.
\end{abstract}

Keywords: Mathematical modeling; decision making; learning; dynamical systems.

\section*{Introduction}

Mathematical models of learning play great role in a diverse range of fields, with eminent applications found in cognitive science (Daw, O'Doherty, Dayan, Seymour, \& Dolan, 2006; Burke, Tobler, Baddeley, \& Schultz, 2010; Ahn et al., 2012), artificial intelligence (Sutton \& Barto, 1998) and game theory (Fudenberg \& Levine, 1998). The latter traditionally concentrates on the analysis of the Nash equilibria in games played by perfectly rational agents, thereby imposing "heroic assumptions about the knowledge and calculating abilities of the players"(Macy \& Flache, 2002). It is the learning approach to game theory that addresses this issue by focusing on the adaptive behavior of the players. First, it assumes that the agents initially know little about the game context and should gradually explore the game while it is repeated indefinitely many times. Second, players base their actions solely on the previous observations; they learn by trial and error while their ultimate goal is to maximize the cumulative payoff throughout the game.

In the game learning setting it turns out that the players often fail to eventually come up with a certain efficient strategy (either pure or mixed), so their behavior can not be characterized in terms of Nash equilibria. Therefore, the inherent dynamics of learning becomes vital. A growing number of studies develop the theory behind the applications of dynamical systems to learning. Coupled replicator equations were proposed as a framework for describing the adaptive behavior of multiple learning agents interacting via a simple game (Sato, Akiyama, \& Farmer, 2002; Sato \& Crutchfield, 2003; Sato, Akiyama, \& Crutchfield, 2005). Based on this formalism a whole range of agent behavior properties have been modelled, including noisy perception of op-
ponent's strategies (Galla, 2009, 2011) and scale-free memory (Lubashevsky \& Kanemoto, 2010). Virtually all mentioned studies emphasize that the learning process dynamics in game theoretic setting is naturally rich and non-trivial. Even the simplest systems of two agents learning to play rock-paper-scissors game may produce quasiperiodic tori, limit cycles and deterministic chaos (Sato et al., 2002; Sato \& Crutchfield, 2003); the latter is often reported to be a common behavior of dynamical systems describing learning processes (Sato et al., 2005; Lubashevsky \& Kanemoto, 2010; Galla \& Farmer, 2013).

Indeed, the perfect rationality axiom appears unsuitable in a whole class of problems. As one may see, this fact motivated much current research on the development of the learning approach to game theory and corresponding mathematical models of learning. The canonical game theory implies that a player has full information about both the game played and the opponents faced. In contrast, the learning paradigm hypothesizes that most of this information is concealed from the players, who only possess the complete knowledge about the set of available actions and gradually learn the consequences of these actions. Even so, in the vast majority of situations studied within the learning framework so far the agents are practically assumed to be strictly rational. In other words, even learning agents still act selfishly and optimally; their rationality is bounded only in the sense of having less a priori information. Put within the constraints imposed by the learning paradigm, agents now have to learn the appropriate behavior strategy, but their final goal remains ultimately rational to maximize the total payoff throughout the whole process. It means that in the course of learning the agent behavior is driven only by external factors - the actions of other players and the corresponding payoffs observed previously. In the modern dynamical models of learning the agents basically lack any kind of personality, they posess no emotions, desires or personal preferences. Up to now it is completely unknown how the dynamics of the learning would change if the agents are endowed with any kind of individuality. In the present study we face this problem.

One of the most important aspects of learning processes is the intrinsic motivation, which is commonly defined as an inspiration to do something "because it is inherently interesting or enjoyable" (Ryan \& Deci, 2000). In contrast, extrinsic motivation refers to doing something "because it leads to a separable outcome" (Ryan \& Deci, 2000). In relation to learning, an intrinsically motivated person learns something not (or not only) because it will lead her to a tangible reward
or payoff, but for the sake of joy obtained by the learning itself. Such person innately likes the very process of gaining new knowledge. The concept of intrinsic motivation is widely studied in psychology (Deci \& Ryan, 1985; Ryan \& Deci, 2000) and has vital applications in education, as well as in organizational psychology and psychotherapy. Besides, intrinsically motivated reinforcement learning (Oudeyer, Kaplan, \& Hafner, 2007; Oudeyer et al., 2007; Singh, Lewis, Barto, \& Sorg, 2010) is a hot topic in computer science: machine learning algorithms inspired by human cognitive processes demonstrate improved performance in a wide class of tasks. Still, despite the solid theoretical basis of intrinsically motivated learning, the dynamics of such learning processes remains a murky subject. What is the impact of the intrinsic motivation on the outcome of a learning process? Can we expect that intrinsic (and in a certain sense irrational) desire to learn will change the agent behavior substantially? Do the intrinsic motives deserve as close attention as the extrinsic ones?

Employing the dynamical systems framework, we propose a toy model capturing the effects of intrinsic motivation to learn. We study the example of a single agent facing an unknown environment, who is forced to make a repeated choice between two rewarded alternatives. The purpose of the agent is to maximize the total sum of the rewards gained throughout the process; the agent therefore should learn which of the alternatives is better rewarded. The key point of the present study is that the agent is biased: along with collecting the rewards, she also satisfies the internal need to acquire new knowledge. Therefore, the agent behavior is governed by two factors: objective (to gain as much reward as possible) and subjective (to satisfy the internal desire to learn). Our global aim in the present paper is to demonstrate on this simple example how such subjective factors may greatly impact on the dynamics of systems describing human behavior.

\section*{Model}

We construct the continuous-time reinforcement learning model of a single agent adaptation, or learning, under the effect of intrinsic motivation. The discrete-time learning models is the more conventional way to describing the learning processes. However, for purposes of analysis of system dynamics the continuous models are more appropriate. We refrain from discussing the connection between the discretetime and continuous-time reinforcement learning formulation, which is covered in detail in the literature (Sato et al., 2005; Lubashevsky \& Kanemoto, 2010). We only note that the continuous-time process is actually the limit case of the discrete-time learning, when the learning agent repeatedly makes a choice infinitely many times.

In our model the agent interacts with the unknown environment by repeatedly choosing one of the two available actions \(x_{i}, i \in 1,2\) and receiving corresponding reward \(r_{i}\). After each decision, only the action that was actually chosen is being reinforced. In game theory it corresponds to the situ-
ation where the agent is not provided with any information about the foregone payoffs (also known as choice reinforcement (Ho, Camerer, \& Chong, 2007)), in contrast to the conventional weighted fictitious play scheme. The agent accumulates the memories of the obtained rewards, and in such manner builds up an inner myopic model of the outer world. Each time the agent makes a choice she relies on the currently collected information about the quality of both actions, and, at the same time, is affected by her intrinsic motivation to learn, or to obtain new information. We interpret the latter in a sense that the agent inherently likes to select the options that add much new information to her inner model of the world. Therefore, at each instant \(t\) there are three values associated with each option \(x_{i}\) :
1. \(p_{i}\) - the probability of choosing \(x_{i}\) at time \(t\)
2. \(q_{i}\) - the agent memories about the rewards obtained in the past for selecting \(x_{i}\) (objective quality of \(x_{i}\) )
3. \(n_{i}\) - the novelty of the option \(x_{i}\) (subjective quality of \(x_{i}\) )

In order to complete the model, we, first, define how the choice probability \(p_{i}\) depends on \(q_{i}\) and \(n_{i}\). Second, we write the equations describing time evolution of the agent memories about \(x_{i}\) and corresponding values of novelty.

The Boltzmann distribution (sometimes referred to as "softmax" model) fits much experimental data and is commonly used as a model for randomized human choice. We adopt it as a probability of choosing action \(x_{i}\) at time \(t\)
\[
\begin{equation*}
p_{i}(t)=\frac{e^{\beta\left[q_{i}+n_{i}\right]}}{\sum_{j}^{\beta\left[q_{j}+n_{j}\right]}}, \tag{1}
\end{equation*}
\]
where \(q_{i}+n_{i}\) represents the total quality of option \(x_{i}\). Here without loss of generality we assume that objective and subjective factors are equally important for the agent. The constant parameter \(\beta\) defines to what extent the agent choice is randomized ( \(\beta=0\) corresponds to the completely random choice, while \(\beta=\infty\) makes the agent always select the option with the highest total quality).

We describe the evolution of the objective values \(q_{i}, \quad i=\) 1,2 over time by the following differential equations:
\[
\begin{equation*}
\dot{q}_{i}=W\left(q_{i}, q\right) r_{i} p_{i}-\frac{q_{i}}{T_{q}}, \tag{2}
\end{equation*}
\]
where \(p_{i}\) is defined by expression (1), \(r_{i}\) is the reward associated with action \(x_{i}\). Term \(r_{i} p_{i}\) can be regarded as a basic reinforcement, which is subjected to saturation effect. The term \(\frac{q_{i}}{T_{q}}\) stands for the effect of the bounded capacity of the agent's memory. The events in the past separated from the present by the time considerably exceeding \(T_{q}\) practically do not affect the agent's behavior.

The saturation factor \(W\left(q_{i}, q\right)\) is a weighting function depending on \(q=\left(q_{1}, q_{2}\right)\). We chose \(W\left(q_{i}, q\right)\) in such way that it bounds the infinite growth of the objective value function.

In other words, it implements the saturation effect: we tend to underestimate frequent events and overestimate rare ones. We define \(W\left(q_{i}, q\right)\) as logistic function
\[
\begin{equation*}
W\left(q_{i}, q\right)=\frac{1}{1+e^{\frac{q_{i}-\bar{q}}{\gamma}}}, \tag{3}
\end{equation*}
\]
where \(\gamma\) is the saturation coefficient and \(\bar{q}=\frac{q_{1}+q_{2}}{2}\). If the current objective value of \(x_{i}\) is relatively large \(\left(\frac{q_{i}-\bar{q}}{\gamma} \gg 1\right)\), the probability \(p_{i}\) is very high and \(x_{i}\) is selected frequently, so the agent underestimates the reward gained: \(W\left(q_{i}, q\right) \approx 0\). On the opposite, for the rarely selected actions \(p_{i}\) is low (so \(\frac{q_{i}-\bar{q}}{\gamma} \ll 1\) ), and when such actions are chosen the agent pays full attention to their reward: \(W\left(q_{i}, q\right) \approx 1\).

In order to take into account the effect of intrinsic motivation to select the options that brings much information to the agent environment model, we augment system (1-3) with the equations describing time evolution of the novelty values for each option
\[
\begin{equation*}
\dot{n}_{i}=\phi\left(1-p_{i}\right)-\frac{n_{i}}{T_{n}} \tag{4}
\end{equation*}
\]

Here \(\phi\) is the parameter indicating agent's novelty rate that is the same for all of the alternative choices. In analogy to the equation (2) we define the memory capacity coefficient \(T_{n}\) accounting for the characteristic duration of the novelty effect.

Equations (1-4) form the basic model of the agent adaptation under the assumptions stated above. In the rest of the paper we present the preliminary analysis of the results of the numerical experiments aimed at elucidation of the basic properties of the developed model.

\section*{Numerical simulation}

Prior to discussion of the numerical results, we have to underline that the similar system describing the behavior of rational agent have been analyzed previously (Sato et al., 2005). It has been elucidated that the system dynamics in case of rational agent is very simple. Namely, the agent tends to one of the equilibria depending on the system parameters, and the selected equilibria is stable with respect to the perturbations of intial conditions. Therefore, there are very few studies investigating the single agent adaptation problems, due to the absence of any complications of system behavior. We show that introducing intrinsic motivation makes the situation completely different.

Under the assumption of equal rewards \(\left(r_{1}=r_{2}=1\right)\) we numerically simulated the dynamics of system (1-4). We discovered that depending on the values of the system parameters the structure of the system phase space trajectory may take one of two general forms: either the stable equilibrium exists or the system is unstable and has the limit cycle. We have not aimed at analyctically deriving the explicit conditions of the system instability, but the empirical observations indicate that the stable behavior is rather common, while un-


Figure 1: Stable dynamics of the analyzed system. Top frame illustrates the time evolution of the objective quality \(q_{1}\) and novelty value \(n_{1}\); the bottom frame represents the choice probability \(p_{1}\) evolution. The time series were obtained for the time span of 500 units and following values of system parameters: \(r_{1}=r_{2}=1, \beta=5, \phi=1, \gamma=1, T_{q}=70, T_{n}=50\); the initial conditions were chosen randomly.
stable dynamics was found only for relatively narrow sets of parameters.

The typical example of the stable dynamics is illustrated in Fig. 1. The agent eventually learns the mixed strategy \(p_{1}=p_{2}=0.5\), which is the stable equilibrium of the system. However, it is instability that often characterizes the human behavior, so we focus our attention on the second case. Fig. 2 represents the periodic motion of the system at hand. As can be seen from the top two frames, the system trajectory forms a limit cycle. Starting from the randomly selected intial values, system variables undergo periodic oscillations after a short transition process. The observed dynamical patterns correspond to the case when the decision maker changes her preferences from time to time, or, in other words, periodically "switches" from one alternative to another. The implicit dependence between the objective quality and the novelty of the option can bee seen in the bottom left frame of Fig. 2. When the quality of the alternative (as represented in the agent memories) attains local maximum, the corresponding choice probability also peaks. So this alternative is chosen frequently during some period of time and, thus, its novelty takes the lowest possible value. On the other hand, when the probability of \(x_{i}\) being chosen is low, the agent has relatively little information about the consequences of this action (because the memories about it eventually vanish if not reinforced regularly). Therefore, the agent intrinsic desire to learn motivates her to choose this option due to the relatively


Figure 2: Unstable dynamics of the analyzed system. Top left frame illustrates the system trajectory on the plane \(\left(q_{1}-q_{2}, n_{1}-\right.\) \(n_{2}\) ), while top right depicts the projection of the system phase space trajectory onto the ( \(q_{1}, n_{1}\) ) plane. Two bottom frames demonstrate the time evolution of the objective quality \(q_{1}\), novelty value \(n_{1}\) (both on the bottom left frame) and the choice probability \(p_{1}\) (bottom right) for option \(x_{1}\). Represented results were obtained for the time span of 500 units and following values of system parameters: \(r_{1}=r_{2}=1, \beta=2, \phi=0.4, \gamma=4, T_{q}=70, T_{n}=50\); the initial conditions were chosen randomly.
large amount of information that the agent might acquire.
Finally, the evolution of the choice probability \(p_{1}(t)\) (see bottom right frame in Fig. 2) demonstrates that during considerable periods of time the probability of choosing \(x_{1}\) remains close to zero; these intervals slightly precede the periods when \(q_{1}\) is low and \(n_{1}\) is high. Then, after staying within the vicinity of zero, the probability rapidly reaches the maximum value around unity and in turn remains near this value for the next half-cycle.

The conducted numerical analysis confirms that the the system (1-4) actually exhibits the properties one may intuitively anticipate from the intrinsically motivated agent. The agent learns one of the optimal options, but being biased she eventually tends to discard the established strategy that proved its efficiency in favor of the novel one. Moreover, the preliminary analysis of the non-symmetric case revealed that the similar behavior can be observed even when the rewards are not equal. This fact requires a thorough investigation and will be reported elsewhere.

The results presented in the present work already enable us to conclude that even the simplest systems with boundedly rational agents may exhibit non-trivial dynamics. However, more detailed analysis of the proposed model is required. Particularly, the system stability conditions are still to be determined. Also under the scope of future work is the question of how the system dynamics patterns depend on the sysem parameters, namely, the novelty rate, perception thresholds
and the parameters characterizing the capacity of the agent memory.

\section*{Conclusion}

We have proposed a dynamical model of intrinsically motivated learning. In the various learning models developed previously in game theory and cognitive science the learning subject is assumed to act rationally in achieving the ultimate goal - to maximize the cumulative reward gained during the learning. We challenge this approach by assigning a piece of non-rationality to the learning agent. The curiosity is what biases the selfish agent in our model.

We confine our scope to the case of single agent adaptation and follow the reinforcement learning setting. The agent behavior in our model is governed by two stimuli. The objective stimulus is traditional - to maximize the total payoff collected throughout the process. The subjective one is irrational - to engage in active learning as much as possible, because the very learning process is enjoyable. We show that the agent biased in such way at least under some conditions does not stick to the optimal strategy of behavior, in contrast to the rational learning agent. Rather, in such cases the agent preference continuously varies in an oscillatory way. Performing the simple numerical analysis of the model, we demonstrate that the intrinsic motivation leads to the instability of the learning dynamics.

Our results give evidence to the fact that the intrinsic mo-
tivation in particular and the bounded rationality in general may cause the significant changes in the behavior of singleand multi-agent systems. We argue that the intrinsic motives should be paid no less attention than the extrinsic ones, if one considers the systems where human decisions are of the primary importance.

\section*{References}

Ahn, W.-Y., Rass, O., Shin, Y.-W., Busemeyer, J., Brown, J., \& O'Donnell, B. (2012). Emotion-based reinforcement learning. In Proceedings of the 34th Annual Conference of the Cognitive Science Society (pp. 124-129).
Burke, C., Tobler, P., Baddeley, M., \& Schultz, W. (2010). Neural mechanisms of observational learning. Proceedings of the National Academy of Sciences, 107(32), 1443114436.

Daw, N., O'Doherty, J., Dayan, P., Seymour, B., \& Dolan, R. (2006). Cortical substrates for exploratory decisions in humans. Nature, 441(7095), 876-879.
Deci, E., \& Ryan, R. (1985). Intrinsic motivation and selfdetermination in human behavior. Springer.
Fudenberg, D., \& Levine, D. (1998). The theory of learning in games. MIT press.
Galla, T. (2009). Intrinsic noise in game dynamical learning. Physical review letters, 103(19), 198702.
Galla, T. (2011). Cycles of cooperation and defection in imperfect learning. Journal of Statistical Mechanics: Theory and Experiment, 2011(08), P08007.
Galla, T., \& Farmer, J. (2013). Complex dynamics in learning complicated games. Proceedings of the National Academy of Sciences, 110(4), 1232-1236.
Ho, T., Camerer, C., \& Chong, J. (2007). Self-tuning experience weighted attraction learning in games. Journal of

Economic Theory, 133(1), 177-198.
Lubashevsky, I., \& Kanemoto, S. (2010). Scale-free memory model for multiagent reinforcement learning. Mean field approximation and rock-paper-scissors dynamics. The European Physical Journal B-Condensed Matter and Complex Systems, 76(1), 69-85.
Macy, M., \& Flache, A. (2002). Learning dynamics in social dilemmas. Proceedings of the National Academy of Sciences, 99(Suppl 3), 7229-7236.
Oudeyer, P., Kaplan, F., \& Hafner, V. (2007). Intrinsic motivation systems for autonomous mental development. Evolutionary Computation, IEEE Transactions on, 11(2), 265286.

Ryan, R., \& Deci, E. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. Contemporary educational psychology, 25(1), 54-67.
Sato, Y., Akiyama, E., \& Crutchfield, J. (2005). Stability and diversity in collective adaptation. Physica D: Nonlinear Phenomena, 210(1), 21-57.
Sato, Y., Akiyama, E., \& Farmer, J. (2002). Chaos in learning a simple two-person game. Proceedings of the National Academy of Sciences, 99(7), 4748-4751.
Sato, Y., \& Crutchfield, J. (2003). Coupled replicator equations for the dynamics of learning in multiagent systems. Physical Review E, 67(1), 015206.
Singh, S., Lewis, R., Barto, A., \& Sorg, J. (2010). Intrinsically motivated reinforcement learning: An evolutionary perspective. Autonomous Mental Development, IEEE Transactions on, 2(2), 70-82.
Sutton, R., \& Barto, A. (1998). Reinforcement learning: An introduction (Vol. 1) (No. 1). Cambridge Univ Press.

\title{
When to Hold and When to Fold: Detecting Structural Changes in Statistical Learning
}

\author{
Benjamin D. Zinszer (bdz107@psu.edu) \\ Center for Language Science \\ Departments of Psychology \& Statistics, Penn State University \\ University Park, PA 16802 USA \\ Daniel J. Weiss (djw21@psu.edu) \\ Center for Language Science \\ Department of Psychology and Program in Linguistics, Penn State University \\ University Park, PA 16802 USA
}

\begin{abstract}
Studies of statistical learning have documented a remarkable sensitivity to structural regularities in both infants and adults. However, most studies of statistical learning have assumed a single underlying causal structure with uniform variance. In previous work in which two structures are presented successively, a primacy effect has been reported in which only the first structure is acquired. The present study explores the conditions under which such primacy effects are observed and learners are capable of acquiring both structures. We argue that learners can detect multiple structures by monitoring the consistency of the input.
\end{abstract}

Keywords: speech segmentation, statistical learning

\section*{Introduction}

Over the past twenty years, research on language acquisition has been transformed by the finding that infant and adult learners can use rudimentary statistics to parse artificial speech streams (Saffran, Aslin, \& Newport, 1996; Saffran, Newport, \& Aslin, 1996). A large number of follow-up studies have replicated and extended the initial findings, determining that statistical learning is neither domain specific (e.g., Fiser \& Aslin, 2002a; Kirkham et al. 2002), nor even restricted to humans (Hauser et al., 2001; Toro \& Trobalón, 2005). The term statistical learning has consequently come to be associated with a wide range of phenomena that rely on implicit calculations based on distributional regularities in the environmental input.

The utility of these statistical learning experiments for simulating the early stages of language acquisition has been widely acknowledged. However, with few exceptions, the input to learners in statistical learning experiments has been characterized by a single, highly invariant statistical structure. This uniform-variance property of the input does not reflect the substantial variability inherent in natural language corpora due to shifts in topic, speaker, accent, and even language (in the case of bilingual acquisition). In some instances, variance in the input may signal to the learner that they are in a new context for which a different statistical structure must be learned (e.g., a language change), but in other cases this variation represents noise and should not trigger a new structural representation (e.g., hearing foreign-
accented speech). Thus, the critical challenge confronting language learners is much like Piaget's description of the processes of assimilation and accommodation (Piaget, 1985). The learner must ultimately determine the number of causal models that best characterizes the input, resolving when a new causal model is required and when the existing model can account for the observed data.

There are at least two potential sources of information that may facilitate learners to detect that there has been a change in structure over time, which in turn may facilitate the formation of multiple representations (Gebhart, Aslin, \& Newport, 2009). The first source of information is the availability of a contextual cue that is correlated with a particular statistical structure (e.g., Weiss, Gerfen, \& Mitchel, 2009; Gebhart, Aslin, \& Newport, 2009). The existence of such a cue could result in computations that are performed over a subset of the input and then compared across contexts. If the computations differ by some criterion, it would trigger the learner to form multiple representations to accommodate the inputs associated with each context. A second potential source of information for learners may be derived from monitoring the consistency of the input (Basseville \& Nikiforov, 1993; see Gebhart, Aslin, \& Newport, 2009). If the surface statistics are entirely consistent, the learner may conclude that the input likely has arisen from a single underlying structure. Conversely, if the variance in the surface statistics exceeds some criterion, then the learner may conclude that the underlying structure has undergone some change (see Gebhart, Aslin, \& Newport 2009; Qian, Jaeger, \& Aslin, 2012).

To date, only a few experiments have tested whether contextual cues facilitate the formation of multiple representations when multiple inputs are presented. In a study by Weiss, Gerfen, \& Mitchel (2009), learners were presented with two artificial languages comprised of four words each, in which the words were defined solely by transitional probabilities. The languages were interleaved in two-minute intervals twelve times total. When the languages were presented in a single voice, only congruent language pairs were learned (ones whose statistics, when combined, yielded similar transitional probabilities to the languages
presented in isolation). Incongruent languages (ones whose statistics were incompatible and yielded a higher noise level when combined) were only learned when a contextual cue was added such that one language was presented in one voice and the other in a second voice. Gebhart, Aslin, and Newport (2009) used a similar methodology, presenting learners with two five-and-a-half minute segments of incompatible languages presented consecutively (in the same voice). They reported a primacy effect in which the first language was learned at above chance levels, while the second language was not. However, learners succeeded in acquiring both languages if there was an explicit cue (informing the learners they would acquire two languages) in conjunction with a brief pause between streams. Also, tripling exposure time to the second language allowed learners to perform above chance in both languages, indicating that both languages could be acquired given sufficient exposure to the new language. Together, these results support the notion that the presence of a contextual cue differentiating the inputs can facilitate the formation of multiple representations, perhaps providing the learner with a more efficient route to successful acquisition.

To the best of our knowledge, no study to date has systematically investigated whether and how learners can form multiple representations by monitoring the consistency of the input alone. Arguably, some of the results from the aforementioned studies begin to address this issue, though the findings have not been easy to interpret (e.g., Weiss, Gerfen \& Mitchel reported that repeated presentations of incongruent languages in the same voice resulted in no learning whereas the single presentations in Gebhart, Aslin, and Newport resulted in a primacy effect). In Experiment 1a, we set out to initially replicate the primacy effect of Gebhart and colleagues using their own languages. Subsequently, we manipulate both duration and languageswitching parameters to determine the conditions under which learners can acquire both languages by monitoring the consistency of the input in the absence of contextual cues such as speaker voice or explicit instructions.

\section*{Experiment 1}

In Experiment 1a, our goal was to replicate the primacy effect reported by Gebhart, Aslin, and Newport (2009) by presenting learners with two consecutive artificial languages with no explicit cue to the transition. We subsequently extend the study by manipulating the number of transitions between the two languages during the familiarization phase. Thus, in Experiment 1b, we control for the amount of exposure to each language while adding additional transition points (i.e., presenting four 2-minute 45 -second blocks of each language versus two 5:30 blocks in Experiment 1a).

\section*{Methods}

Participants Thirty-four undergraduate students were recruited from a Psychology 100 subject pool. Participants
were divided into two conditions: 17 (11 female, 6 male) participants in Experiment 1a with a mean age of 19.6 years, and 17 (12 female, 5 male) participants in Experiment 1b with a mean age of 19.6 years. All participants were English monolinguals by self-report.

Languages The speech stream was composed of two languages, each consisting of sixteen trisyllabic words based on 12 unique CV syllables. These artificial languages were previously used in Gebhart, et al.'s (2009) segmentation experiment. Individually, the languages could be segmented by tracking the transitional probabilities (TPs) between syllables, with high TPs between syllables within a trigram (representing a word) and low TPs between syllables across different trigrams (representing word boundaries). See Figure 1 for an illustration of TP-defined words boundaries.

In both languages, two vowel frames and six consonants were used to define the trisyllabic words. The withintrigram transitional probability for syllables was 0.50 . Words within the stream were randomly sequenced yielding a transitional probability of 0.25 between word-final and word-initial syllables. The second language rearranged the vowel frames and consonants of the first language, resulting in a syllable inventory that overlapped by \(50 \%\). The combined transitional probabilities (including all syllables across both languages) varied from 0.33 to 0.67 both within and between words. Consequently, they did not provide consistent cues for segmentation (see Figure 1).

Procedure Participants were instructed to listen to a brief recording of foreign speech and informed that they would later be quizzed on what they had heard. In Experiment 1a, participants listened to 5 minutes and 30 seconds of each language (produced in the same voice) consecutively without any cues to transition. Order of language


Figure 1 - Transitional probabilities defining the structure of each language. When combined, the TPs of each language result in a flat (uninformative) structure.


Figure 2 - Durations of each experiment depicted together. Dark bars represent Language A, and light bars represent Language \(B\).
presentation was counter-balanced between participants, but for simplicity we will always refer to the Language A as the first language presented and Language \(B\) as the second. In Experiment 1b, participants listened to 4 consecutive blocks each consisting of 2 minutes and 45 seconds of one language (2:45 A + 2:45 B + 2:45 A + 2:45 B, order of actual languages was counter-balanced between participants). In both conditions, total exposure to each language was constant (5:30) as well as the total duration of the familiarization phase (11:00).

After familiarization, participants completed a test phase with thirty-two two-alternative forced choice trials in which participants selected between statistically-defined words and partwords. The partwords consisted of either the last syllable of a word followed by the first two syllables of another word or the last two syllables of a word followed by a single syllable of another word. These items occurred during the familiarization but were characterized as partwords since the within-trigram transitional probabilities were low. Participants were asked to judge which of the trigrams sounded more familiar, with statistically-defined words being counted as correct responses.

\section*{Results \& Discussion}

Mean correct responses on the test trials were computed for each language. In Experiment1a, participants scored a mean accuracy of 0.746 ( \(S D=0.152\) ) on Language A and 0.581 ( \(S D=0.192\) ) on Language B . These scores indicated a primacy effect in which accuracy on Language \(A\) significantly exceeded Language \(B\) (paired \(t(16)=2.82\), \(p=0.012\) ). Accuracy on Language A was significantly above chance ( \(t(16)=6.67, p<0.001\) ) while accuracy in Language B was not \((t(16)=1.73, p=0.102)\). By contrast, in Experiment 1b, Language A and Language B did not significantly differ (paired \(t(16)=1.00, p=0.331\) ). Also in contrast to Experiment 1a, Language A significantly exceeded chance (A: \(M=0.673\), \(S D=0.151, t(16)=4.73, p<0.001)\), while Language B was also marginally significant ( \(M=0.603, S D=0.207, t(16)=2.05\), \(p=0.057\) ). Results are illustrated in Figure 3.

In Experiment 1, we successfully replicated the primacy effect of Gebhart, et al. (2009) and discovered that increasing the number of switches between the languages could eliminate the primacy effect. Experiments 1a and 1b differed only in the duration of the individual exposure segments (5:30 vs. 2:45) and the number of switches
between languages in the familiarization phase (1 vs. 3). Two causal hypotheses may be proposed for these results: First, the greater number of switches in the 1b stream may cue the listeners to the existence of two structures, allowing them to begin acquiring Language B . Alternately, in Experiment 1a (and the previous experiment by Gebhart and colleagues), learners may become entrenched in the statistical structure of Language A due to the lengthy duration of initial exposure. This entrenchment may inhibit detection or acquisition of the new structure. Experiment 2 was designed to disentangle these hypotheses.

\section*{Experiment 2}

In Experiment 2, we contrasted the entrenchment and switching hypotheses proposed to explain the results of Experiment 1. We accomplished this by presenting participants with the following sequence of languages without any breaks in between: 5:30 of Language A followed by 2:45 of Language B, 2:45 of Language A again, and finally \(2: 45\) of Language \(B\). The entrenchment hypothesis predicts that the primacy effect found in Experiment 1a should also be present for Experiment 2 since the duration of the initial block of Language A is identical. The switching hypothesis predicts that both languages will be learned at significantly greater than chance levels since there are three transitions. This prediction of learning is somewhat counter-intuitive given that Language B was not learned in Experiment 1a and here we are increasing exposure to Language A.

\section*{Methods}

Participants Twenty Psychology 100 students participated (12 female, 8 male; mean age 19.9 years).

Procedure As noted above, in this experiment the languages were configured as follows: 5:30 A \(+2: 45 \mathrm{~B}+\) 2:45 A + 2:45 B (see Figure 2). All other procedures were identical to Experiment 1.

\section*{Results \& Discussion}

No primacy effect was observed in Experiment 2, where Language A accuracy ( \(M=0.697, S D=0.161\) ) did not significantly differ from Language B accuracy ( \(M=0.694\), \(S D=0.137\); paired \(t(19)=0.07, p=0.943\) ). Moreover, both languages significantly exceeded chance performance
(A: \(t(19)=5.46, p<0.001 ;\) B: \(t(19)=6.31, p<0.001\) ). Our results clearly reject the entrenchment hypothesis, lending support to the switching hypothesis as both languages were learned at above chance levels and performance did not significantly differ between languages.

\section*{Experiment 3}

In Experiment 2, we eliminated the primacy effect found in Experiment 1a by increasing the number of switches between languages, even though Language A possessed a relative advantage in initial presentation duration and overall exposure time. In Experiment 3, we eliminate the last Language B exposure to test whether Language B may have been learned early in the sequence of exposures or whether the learning of B occurred only after the third transition (i.e., the last presentation). Notably, the only occurrence of Language B coincides with the first transition point in the sequence (and there are fewer transitions overall). Understanding when learning occurs may shed light on the type of processing that may be occurring for the unlearned language in conditions eliciting a primacy effect.

\section*{Methods}

Participants Fifteen Psychology 100 students participated (13 female, 2 male; mean age 19.3 years).

Procedure In this experiment, languages were configured as follows: 5:30 A \(+2: 45 \mathrm{~B}+2: 45 \mathrm{~A}\) (see Figure 2 for illustration). All other procedures were identical to those described for Experiment 1.

\section*{Results \& Discussion}

The primacy effect emerged again in this experiment (paired \(t(14)=2.73, p=0.016\) ), as Language A accuracy ( \(M=0.729\), \(S D=0.179\), compared to chance: \(t(14)=4.95, p<0.001)\) significantly exceeded Language B ( \(M=0.546, S D=0.139\), compared to chance: \(t(14)=1.28, p=0.221)\). The contrast between these results and Experiment 2 highlights the
importance of the second presentation of Language B for learning. In the absence of the third switch and additional exposure, performance was at chance levels for Language B. These results raise the question of whether the deficit in Language \(B\) learning was a function of the removal of the third language switch or the decrease in overall Language B exposure (from 5:30 to 2:45).

\section*{Experiment 4}

The primacy effect observed in Experiment 3 emerged in the context of fewer switches and very short overall exposure. Therefore, in Experiment 4, we matched the overall exposure durations of Experiment 2 by providing learners with 5:30 of Language A followed by 5:30 of Language B and then an additional 2:45 of Language A again. This is essentially the sequence presented in Experiment 1a followed by an additional short block of Language A familiarization. Accordingly, the overall duration of each language presentation resembles Experiment 2 (in which both languages were learned), but here only two language switches are provided. Also like Experiments 2 and 3, Language \(A\) is advantaged in total exposure duration relative to Language \(B\).

\section*{Methods}

Participants Seventeen Psychology 100 students participated ( 13 female, 4 male; mean age 19.3 years).

Procedure In this experiment, languages were configured as follows: 5:30 A \(+5: 30 \mathrm{~B}+2: 45 \mathrm{~A}\) (see Figure 2 for illustration). All other procedures were identical to those described for Experiment 1.

\section*{Results \& Discussion}

Although Language A did not significantly exceed chance ( \(M=0.596, S D=0.210, t(16)=1.88, p=0.079\) ) and Language B significantly exceeded chance \((M=0.632, S D=0.132\), \(t(16)=4.12, p<0.001\) ), accuracy in Languages A and B did


Figure 3 - Accuracy of participant responses in familiarity task. Chance level is 0.50 , and error bars denote \(95 \%\) confidence intervals of the mean.
not significantly differ from each other (paired \(t(16)=-0.67\), \(p=0.514\) ). See Figure 3 for illustration. The findings of Experiment 4 provide further evidence that that transitions between the languages alters learning. As in Experiment 2, we made the surprising observation that the additional exposure to Language A could facilitate learning of Language B. It is unclear at this point why Language A's learning was reduced and future experiments will explore the source of this effect. Irrespective of this pattern, the results from Experiment 4 do imply that some processing of Language B occurs even in conditions resulting in a primacy effect for Language A, such as the results reported by Gebhart, Aslin, \& Newport (2009) and our Experiment 1a.

\section*{General Discussion}

In the foregoing experiments, we have explored a range of conditions in which learners were familiarized with two artificial language streams characterized by incompatible underlying statistical structures. Unlike previous studies investigating statistical learning of multiple streams, no extralinguistic contextual cues were provided to the learners to signal the presence of a second language and facilitate learning. Consequently, successful learning of both languages relied on sensitivity to the structures themselves and the transition points between structures. In previous research, when statistically incompatible artificial languages were presented successively, learners have failed to successfully acquire both structures. In instances in which only a single switch was presented, learners exhibited a primacy effect (Gebhart, Aslin, \& Newport, 2009) whereas when many switches occurred, there was a catastrophic interference effect in which no languages were learned (Weiss, Gerfen, \& Mitchell, 2009; Mitchel \& Weiss, 2010). We presented learners with the same languages used in the Gebhart, Aslin, \& Newport (2009) study, and our results suggest that learners are sensitive to the transitions between the languages which can help them acquire both structures.

As noted above, in previous research, when learners receive input from two structures with only a single transition point, a primacy effect is observed (also replicated in Experiment 1a). In the original study, this effect could only be overcome by significantly extending exposure to Language \(B\). Our findings have demonstrated that the primacy effect can also be overcome without increasing exposure at all. In Experiment 1b, exposure to Languages A and \(B\) were equivalent to the original study, though the languages were presented in smaller blocks and interleaved. This manipulation resulted in successful learning of both streams. Likewise, the results of Experiment 2 demonstrate that Language B can be learned even when our manipulation increases exposure only to Language A. That is, adding Language A training can, by virtue of the switching between languages, support learning in Language \(B\) as measured by our posthoc test.

We also explored whether there was any learning of Language B when it occurred after only a single transition. Gebhart, Aslin, and Newport (2009) as well as our Experiment 1a findings leave open the possibility that Language B was ignored altogether or that structures were unlearnable in light of the prior learning of Language A. The results of Experiment 4 cast doubt on either of these interpretations. In Experiment 4, learners acquired Language B at above chance levels despite the fact that Language \(B\) occurred in the exact same context as in the original condition (i.e., 5:30 of Language A followed by 5:30 of Language B). Like Experiment 2, the only manipulation in this experiment involved additional exposure for Language A.

We observed different learning outcomes between Experiments 3 and 4, suggesting that the amount of exposure to Language \(B\) prior to the second switch modulates the success of learning. It is possible that the 2:45 block of Language \(B\) in Experiment 3 did not provide adequate time for learners to sample the language, or perhaps 2:45 is insufficient to support learning (as we do not yet have baseline data for that duration). Because Experiment 3 is the only condition in the present study that limited Language B exposure to \(2: 45\), further conditions will be necessary to explore this issue.
When three transition points are provided in the input (i.e., Experiment 1b and Experiment 2), the two structures become increasingly discriminable to learners, as evidenced by their above-chance performance in both languages. As noted above, the importance of these switches for detection of the second structure is highlighted in those experiments by the improvement in Language \(B\) performance despite only receiving additional exposure to Language A. We therefore conclude that learners are capable of identifying whether input streams contain one or multiple structures by monitoring the consistency of the input. This finding is in accordance with previous speculation regarding the conditions under which changes to statistical structures may be detected (Gebhart, Aslin, and Newport, 2009). The observed primacy effects in previous research and Experiment 1a cannot be attributed to entrenchment in the first language, as it has now been demonstrated that the primacy effects can be overcome with additional transition points between language streams.

While the present experiment made an extensive demonstration of language learning with only lingustic (syllable inventory) or statistical (TP) cues, a similar attempt has yielded markedly different results: Weiss, Gerfen, and Mitchell (2009) observed catastrophic interference when two incompatible languages were interleaved in 2 minute segments for a 24 minute stream. Under the switching hypothesis, we would have predicted significant learning of both languages. However Weiss and colleagues found that neither language was learned significantly better than chance. These two studies used different statistical
structures, had different amount of overlap between the languages, and a different number of switches. Future experiments will try to systematically manipulate these parameters to better understand how overlap (and statistical compatibility) can influence the learning of multiple streams.

It has been hypothesized that describing how learners detect changes in statistical learning may be best explained by a hierarchical Bayesian model of change detection (Qian, Jaeger, \& Aslin, 2012). How learners interpret non-uniform variance in statistical learning appears to rely on the availability of statistical and linguistic cues, such as changes in transitional probabilities or syllable inventory observed at a transition point between languages. These cues may lead learners to consider a second causal model to describe input (accommodation) over a single causal model under which the variance could occur (assimilation). This process of proposing causal models, weighting them by their likelihood, and comparing them to the input stream follows the procedure of Bayesian model comparison. In the case of extralinguistic cues in speech segmentation (e.g., pitch change or pause), this model comparison may be aided by the expectation of a context change and increase the prior probability of a two-model explanation of variance. Linguistic cues, such as the introduction of new syllables or the change in transitional probabilities between syllables may also effect such a change in the prior probability, though by themselves are insufficient. Our results suggest that change detection can be supported by variance-related events such as language switches, and provides further evidence that the Bayesian framework is a valuable analogy for statistical learning in multiple contexts.

Qian, Jaeger, and Aslin (2012) describe statistical cues to context change in terms of prediction error, i.e., a large deviation of the input stream from the learner's current model. Linguistic cues to speech segmentation may elicit such errors at language switches when the inventory or transitions between syllables change. This error-based cuing appears to be evidenced in the present study based on the importance of language switches to learners' performance. While previous research has demonstrated the utility of prolonged exposure to the second structure to detection of two contexts (Gebhart, Aslin, \& Newport, 2009), we demonstrate that a relatively small set of high variation events can also increase the prior probability for a twomodel hypothesis.

Our future work will attempt to determine the nature of processing that occurs during the unlearned streams. It is possible that learners detect the regularities in the second stream but discard it as noise, or that it is blocked by the learning of the first structure. Clearly, some information is gathered during those periods, as evidenced by the results of Experiments 2 and 4. One set of studies underway introduces a third structure into the sequence (either a new learnable artificial language or an unlearnable non-adjacent transitional probability language). Thus, the sequence is
\(5: 30\) of A followed by \(2: 45\) of Language \(C\) (noise or learnable) followed by \(2: 45\) of \(A\) and \(2: 45\) of \(B\). This condition tests whether switches by themselves are useful (without supporting the statistics of Language B). We are also currently engaged in neuroimaging studies to localize and contrast the learning of Languages A and B in a variety of conditions.

\section*{Acknowledgements}

We wish to thank Dr. Richard Aslin and Timothy Poepsel for comments and discussion, as well as Tzuyu Chang for assistance with data collection. This project was funded by NIH grant R01 HD067250-01 to DJW.

\section*{References}

Aslin, R. N., Saffran, J. R., \& Newport, E. L. (1998). Computation of Conditional Probability Statistics by 8-Month-Old Infants. Psychological Science, 9(4), 321-324.
Basseville, M., \& Nikiforov, I. (1993). Detection of abrupt changes - Theory and application. Englewood Cliffs, NJ: Prentice-Hall.
Fiser, J., \& Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. Proceedings of the National Academy of Sciences of the United States of America, 99(24), 15822-26.
Gebhart, A. L., Aslin, R. N., \& Newport, E. L. (2009). Changing structures in midstream: Learning along the statistical garden path. Cognitive Science, 33, 1087-1116.
Hauser, M. D., Newport, E. L., \& Aslin, R. N. (2001). Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. Cognition, 78(3), 53-64.
Kirkham, N. Z., Slemmer, J. A., \& Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. Cognition, 83(2), 35-42.
Piaget, J. (1985). The equilibration of cognitive structures: The central problem of intellectual development (Vol. 985). Chicago: University of Chicago Press.

Qian, T., Jaeger, T. F., \& Aslin, R. N. (2012). Learning to represent a multi-context environment: More than detecting changes. Frontiers in Psychology, 3(July), 228.
Saffran, J., Newport, E., \& Aslin, R. (1996). Word segmentation: The role of distributional cues. Journal of Memory and Language, 621(35), 606-621.
Toro, J. M., \& Trobalon, J. B. (2005). Statistical computations over a speech stream in a rodent. Perception and Psychophysics, 67(5), 867-875.
Weiss, D. J., Gerfen, C., \& Mitchel, A. (2009). Speech Segmentation in a Simulated Bilingual Environment: A Challenge for Statistical Learning? Language Learning and Development, 5(1), 30-49.

\title{
Complexity Matching in Dyadic Interaction
}

\author{
Drew Abney \\ University of California, Merced, Merced, California, USA
}

\author{
Alexandra Paxton \\ University of California, Merced
}

Chris Kello
University of California, Merced
Rick Dale
University of California, Merced

\begin{abstract}
Recent theoretical discussion of dyadic coordination has focused on issues of synchronization, entrainment, alignment, and convergence. All of these terms refer to matching of specific behavioral and linguistic events, such that members of a dyad coordinate by "doing the same thing." Communicative behaviors tend to be highly variable, like most human behaviors. These tendencies suggest the possibility of complexity matching: Statistical measures of behavioral complexity may converge in certain types of dyadic interaction. In the present study, acoustic speech signals of interlocutors were measured in two conversational conditions, one argumentative and the other affiliative. Signal complexity was measured in terms of heavy tails and power laws in the distributional and temporal properties of acoustic event series, respectively. Parameters of statistical functions were found to vary by conversation type, as did their matching between interlocutors. Results demonstrate a new way to quantify the coordination of interlocutors in terms of complexity matching.
\end{abstract}

\title{
Measuring individual differences in cognitive biases
}

\author{
Balazs Aczel \\ ELTE University, Budapest, Hungary \\ Bence Bago \\ ELTE University, Budapest, Hungary
}

Andrei Foldes
ELTE University, Budapest, Hungary

\author{
Aba Szollosi \\ ELTE University, Budapest, Hungary
}

\begin{abstract}
Measuring individual differences in susceptibility to decision biases has received increased attention in recent years, yet some methodological questions may hinder us from validly assessing the effects of cognitive heuristics. Surveys consisting of measures of several biases often aim to compare rate of occurrence of these biases and to compile a composite index from these measures. Unfortunately, the probability that the participant chooses the normative answer on the test questions often varies between and within studies, thus confounding the results. Another complication in the surveys used is that some incorrect answers are not necessarily the result of the studied bias. In our work, we tried to overcome these methodological challenges in a new survey of 15 cognitive biases. The results of 1127 participants provided insight into several methodological and theoretical questions about measuring individual differences in heuristic decisions.
\end{abstract}

\title{
The Effect of Test Format on Visual Recognition Memory Performance
}

\author{
Nora Andermane \\ University of Bristol \\ Jeffrey Bowers \\ University of Bristol
}

\begin{abstract}
A recent study (Brady et al., 2008) claims that the capacity and fidelity of long-term visual recognition memory has been underestimated. After viewing a massive number of images for nearly six hours participants were highly accurate at identifying a previously seen image in a two-alternative forced choice task even when the foil was extremely similar to the target. In present study we hypothesised that this impressive memory performance might be specific to the test format. To investigate the effect of test format on recognition accuracy we showed participants 1700 images for 2.5 hours and then tested them in a forced choice task and a yes/no task, where participants judged only one image at a time. We found that accuracy was relatively high in both test conditions; however, the performance was lower in the yes/no task ( \(75 \%\) vs. \(86 \%\) ). A follow-up study exploring delayed testing effects will also be described.
\end{abstract}

\title{
Belief Bias in the Perception of Sample Size Adequacy
}

\author{
Richard Anderson \\ Bowling Green State University
}

\begin{abstract}
In two experiments, participants were instructed to set aside their complete knowledge of a statistical population parameter and to take the perspective of an agent whose knowledge was limited to a random sample. Participants rated the appropriateness of the agent's conclusion about the adequacy of the sample size (which, objectively, was more than adequate), along with the agent's intelligence. Whereas previous work suggests that unbelievable statistical conclusions impact reasoning by provoking critical thought which enhances the detection of research flaws, the present studies presented participants an unflawed scenario designed to assess effects of believability on bias. The results included the finding that participants' complete knowledge indeed biased their perceptions not only of the adequacy of the sample size, but also of the rationality of the agent drawing the conclusion from the sample. The findings were interpreted in the context of research on belief bias, social attribution, and Theory of Mind.
\end{abstract}

\title{
Semantic interference in language production is due to graded similarity, not response relevance
}

\author{
Sabrina Aristei \\ Humboldt Universität zu Berlin, Berlin, Berlin, Germany \\ Rasha Abdel Rahman \\ Humboldt-University of Berlin
}

\begin{abstract}
There is an ongoing debate on whether semantic interference effects in language production reflect competitive lexical selection or post-lexical response exclusion mechanisms driven by the response-relevant status of distractor words.

To disentangle categorical relatedness and task-defined response relevance effects, we combined the picture-word interference task with the conditional naming paradigm in an orthogonal design. Participants were instructed to name objects typically located in or on the water (e.g. canoe) and refrain from naming objects typically located outside the water (e.g. bike), and vice versa. Semantic relatedness and response relevance of distractors were manipulated independently. Linear mixed model analyses were conducted with semantic similarity ratings of target-distractor as continuous predictor.

The pattern of results revealed that semantic similarity beyond categorical relations is critical for interference effects to be observed, and not response relevance. These findings provide support for the assumption that lexical selection is competitive and that semantic interference effects in the PWI paradigm reflect this competition.
\end{abstract}

\title{
Experts in the discrimination task of the system noise increase avoidance response in the risk judgment
}

\author{
Yui Ashitani \\ Ochanomizu University, Bunkyo, Tokyo, Japan \\ Hyun-Min Ku \\ Ochanomizu University, Bunkyo, Tokyo, Japan
}

\author{
Akira Ishiguchi \\ Ochanomizu University, Bunkyo, Tokyo, Japan
}

\begin{abstract}
Systems contain normal noises even in a stable state, but larger noises in an uncontrollable state. In this study, we investigated different responses between experts and novices in controllability judgment. A half participant was required to discriminate variance magnitude of two sound tone sequences (standard vs. comparison). The sequence consisted of 16 tones whose pitch contained Gaussian noises. They continued to train the variance discrimination task up to a criterion level, and were regarded as experts. Then, the experts and novices participated in the controllability judgment task (i.e. a kind of the risk judgment task) with use of similar stimuli in the discrimination task. They were allowed to continue to judge for gain whether they were in an uncontrollably higher risk state with larger variances. They could stop the trial to make smaller costs. Our results showed experts in the discrimination task increased avoidance responses more than novices.
\end{abstract}

\title{
Sleep deprivation accelerates delay-related loss of visual short-term memories without affecting precision
}

\author{
Christopher L. Asplund \\ Duke-NUS Graduate Medical School \\ Natalie Wee \\ Duke-NUS Graduate Medical School \\ Michael W.L. Chee \\ Duke-NUS Graduate Medical School
}

\begin{abstract}
Visual short-term memory (VSTM) is a limited information store that supports many higher-order cognitive processes. Here we examined how two challenges to VSTM, sleep deprivation (SD) and maintenance duration, interact to affect the number and precision of stored items. Participants were studied twice, once after a normal night of sleep and once following a night of total SD. For each trial, participants remembered the location and color of three squares over a variable delay ( 1 or 10 seconds), reporting the color of the cued item using a color wheel. The probability of reporting the target item, the precision of report, and the probability of reporting a distractor item were determined using mixture modeling. SD reduced the number of integrated representations that could be reported, an effect compounded by delay. In contrast, SD had no effect on VSTM precision. These results suggest all-or-none memory failures, not gradual degradation, during SD.
\end{abstract}

\title{
Learning Complex Scientific Concepts through Peer Argumentation: The Effect of Belief in Human Presence and Partner's Discourse Style
}

\author{
Christa Asterhan \\ Hebrew University of Jerusalem \\ Miriam Babichenko \\ Hebrew University of Jerusalem
}

\begin{abstract}
This study aims to further our understanding of the social and cognitive processes underlying conceptual change learning through argumentative discourse. We tested the effect of competitive (vs. collaborative) argumentative discourse style and belief of interaction with a human peer (vs. a computerized peer agent) on learning the concept of diffusion through interaction with a disagreeing peer. Peer confederate's verbal behavior was tightly controlled to evoke argumentative discourse, holding content exposure constant but differing in rhetoric style. Students in the collaborative discourse style condition performed better. Moreover even though previous studies have reported that the belief of interaction with a human peer benefits learning in consensual settings, the opposite was found for a settings in which the partner critiques the learner's own solutions: Students performed better when they believed they interacted with a computer agent (vs. with a human peer). Implications for theory as well as task design are discussed.
\end{abstract}

\title{
Emotions in the space: arrangement and biases in preschool children and adults
}

\author{
Maria Atanassova-Trifonova \\ Institute of population and human studies, BAS \\ Dimitar Atanasov \\ New Bulgarian University \\ Pavleta Ognyanova \\ New Bulgarian University
}

\begin{abstract}
With the first experiment of our study we have empirically tested the hypothesis that in adult and in adolescent subjects there is a consistency of the spatial representation of emotion terms. We explore the stability of the association between valence and verticality through modulation effect of the valence priming on spatial arrangement of emotion terms. In the second part of our study we focused on spatial representation of emotions by preschoolers when tested explicitly and implicitly with a task of spatial recall. It was found that the arrangement of emoticons on vertical and horizontal line by preschool children varied significantly as a function of the valence of emotions. The results from our third experiment suggest that the effect of the valence of nonverbal stimuli on the spatial recall was only evident when they were with a negative value and when placed at more central positions near to the horizontal line.
\end{abstract}

\title{
A Bayesian Analysis of Bias in Single Item Recognition for Continuous Word frequency
}

\author{
William Aue \\ Syracuse University, Syracuse, NY, United States \\ Pernille Hemmer \\ Rutgers University \\ Amy Criss \\ Syracuse University
}

\begin{abstract}
The relationship between word frequency (WF), measured on a continuous scale, and recognition memory was examined in a single item recognition task. The aim was to more clearly map the relationship between word frequency and memory performance. In marked contrastContrary to standard findings of a linear relationship when between WF and recognition, is treated as discrete, we observed a curvilinear pattern. Specifically, discriminability (d') is higher at both the low and very high ends of the WF continuum. In addition, we observe shifts in bias (C) with a conservative bias for very high frequency (HF) words between WF and memory performance. Variations of a Bayesian signal detection model were then applied to the data in order to better understand the influences WF on measures of d' and C. The models examined contrast the current explanations of the WF effect in recognition where C does not influence performance with a model where C is free vary as a function of WF.of a linear relationship between WF and discriminability ( \(\mathrm{d}^{\prime}\) ) in recognition memory, with the curvilinear pattern for both d' and bias (C) with the curvilinear patterns observed in the current data set. Implications for models of recognition memory are discussed.
\end{abstract}

\title{
Exploring Cross-Situational Learning and Mutual Exclusivity
}

\author{
Suzanne Aussems \\ Tilburg University
}

Paul Vogt
Tilburg University

\begin{abstract}
Cross-situational learning and mutual exclusivity are strategies proposed to explain the learning of word-meaning mappings. In this paper, seven possible strategies are explored and compared to the results of an artificial word learning experiment. The fixed order of trials in the experiment allows for an exposure-by-exposure approach to explore the individual learning process of words. The experiment shows that adult learners do indeed integrate knowledge from previous exposures, however they have difficulty in keeping track of cross-situational information for learning all twelve word-meaning mappings, although some learners can. The performance of 78 participants is compared to simulations in which various combinations of strategies were modeled. The results suggest that a random strategy with mutual exclusivity as its sole learning mechanism could explain the performance in the experiment. In this strategy, the learner selects an object at random from the context, provided that this object has not received a label yet.
\end{abstract}

\title{
Conditional reasoning with vertical spatial relations
}

\author{
Igor Bajšanski
}

University of Rijeka

\author{
Pavle Valerjev
}

University of Zadar
Tanja Gulan
University of Zadar

\begin{abstract}
The aim of this study was to explore the effect of semantic congruence and spatial orientation on verification time of modus ponens (MP) and modus tollens (MT) conclusions. Factorial experiment ( \(2 \times 2 \times 2\) ) with repeated measures was carried out. Conditionals expressed the vertical relation of two objects in congruent (If cellar is down, then attic is up) or incongruent manner (If attic is down, then cellar is up). Eight pairs of words with clear vertical relation were used. Order of vertical relations in the conditional was also manipulated. Conditionals were in the form of "If P down, then Q up", or "If P up, then Q down". Participants \((\mathrm{N}=48)\) had to verify the presented conclusions in 64 tasks as quickly as possible. Significant effects for valid MP conclusions were obtained. Conclusions containing "down-up" order of relations in conditional premise were processed faster. Congruent spatial relations also facilitated faster answers.
\end{abstract}

\title{
Does interaction matter? Testing whether fast and frugal heuristics can replace interaction in collective decision-making
}

\author{
Dan Bang \\ Department of Experimental Psychology, Oxford University \\ Riccardo Fusaroli \\ Center for Semiotics, Aarhus University \\ Kristian Tylén \\ Center for Semiotics, Aarhus University \\ Karsten Olsen \\ Interacting Minds Centre, Aarhus University \\ Peter Latham \\ Gatsby Computational Neuroscience Unit, University College London \\ Jennifer Lau \\ Department of Experimental Psychology, Oxford University \\ Andreas Roepstorff \\ Interacting Minds Centre, Aarhus University \\ Geraint Rees \\ Institute of Cognitive Neuroscience, University College London \\ Chris Frith \\ Wellcome Trust Centre for Neuroimaging, University College London \\ Bahador Bahrami \\ Institute of Cognitive Neuroscience, University College London
}

\begin{abstract}
In a range of contexts, pairs of interacting individuals arrive at collective decisions by comparing their confidence in their judgements. This tendency to evaluate the reliability of information by the confidence with which it is expressed has been termed the 'confidence heuristic'. In this study, we tested two fast and frugal ways of implementing the confidence heuristic in the absence of interaction: either directly, by opting for the judgement made with higher confidence, or indirectly, by opting for the faster judgement, the latter exploiting a widely known inverse correlation between confidence and reaction time. We found that the success of these heuristics depends on how similar individuals are in terms of their abilities and, more importantly, that for dissimilar individuals such heuristics are dramatically inferior to interaction. Interaction allows individuals to alleviate but not fully resolve - their differences in ability.
\end{abstract}

\title{
An ACT-R model of the P300 event-related potential
}

\author{
Adrian Banks \\ University of Surrey
}

\begin{abstract}
The excellent temporal resolution of event-related potentials (ERPs) makes it an ideal technique to test the timing of events within cognitive models which may not be distinguishable using behavioural data. Here a model of an ERP component is presented which tests features of the ACT-R cognitive architecture. The P300 is an event-related potential that is associated with attending to incoming stimuli and subsequent memory processing. It is commonly elicited using an oddball task in which a series of stimuli are presented comprising infrequent target stimuli against a background of frequent standard stimuli. It is influenced by the probability of a target in a sequence of stimuli. An ACT-R model was developed of the oddball task. P300 amplitude was correlated with the activation of the memory of the target stimuli, providing a good account of the component. The implications of these findings for the ACT-R architecture are discussed.
\end{abstract}

\title{
Context-dependent memory effects on syntax
}

\author{
Brendan Barnwell \\ UC Santa Barbara, Santa Barbara, California, United States
}

\begin{abstract}
Recently, researchers in usage-based linguistics have argued that language should be thought of as a domain-general processing faculty operating on rich memory representations of particular experiences with language. If true, this would imply that context-dependent memory effects ought to be detectable across the language-nonlanguage boundary. Specifically, it would imply that that by manipulating nonlinguistic environmental context, it should be possible to influence people's linguistic production. This study reports three experiments testing whether environmental background color, music, or sound can influence participants' choice of active or passive voice in a picture-description task. Results suggest that the effect, if present, is not as clear-cut as has been argued by some language theorists, but there are promising signs that language production may indeed be susceptible to associative influence from the nonlinguistic environment.
\end{abstract}

\title{
Linguistic differences in explanation requests and their effects on the evaluation of explanations: the case of English and Turkish
}

\author{
Melih Barsbey \\ Bogazici University, Department of Psychology \\ Tania Lombrozo \\ University of California, Berkeley, Department of Psychology
}

\begin{abstract}
How does language shape thought? In particular, do cross-linguistic differences in how explanations are requested affect how explanations are evaluated by speakers of different languages? To address this question we contrasted English with Turkish, which has three distinct words that correspond to "why?" in English. Through two corpus studies and an experimental study, we established that Turkish "why" questions tend to appear in different contexts and elicit different kinds of explanations: the "why" questions vary in the frequency with which they refer to agents and elicit teleological explanations. In an experimental study investigating whether this cross-linguistic difference affects how explanations are evaluated, we found that while English speakers displayed an overall preference for mechanistic explanations in evaluating the stimuli, Turkish speakers provided similar satisfaction ratings for mechanistic and teleological explanations. Our findings have implications for the cognitive science of explanation and for debates about language and thought.
\end{abstract}

\title{
Effects of orthography on second language phonology: The production of geminate consonants in speakers of English as a Second Language
}

\author{
Benedetta Bassetti \\ University of York
}

\begin{abstract}
We tested the hypothesis that the orthographic representations ('spellings') of second language (L2) words affect experienced L2 speakers' pronunciation. In Italian, double consonant letters represent geminate (long) consonants. We predicted that Italian speakers of English would pronounce English words with longer consonants if spelled with double letter, e.g. a longer [t] in kitty than city. Three groups of Italian speakers of English performed different word production tasks with different stimuli: acoustic, acoustic and orthographic, or orthographic. The target voiceless stop consonants were presented inside 9 word pairs, spelled with one or two letters. Acoustic and auditory analyses revealed that the target consonants were longer in words spelled with double than with singleton letters, regardless of task. We argue that L2 speakers decode L2 orthographic representations using L1 orthography-phonology correspondences. This affects their pronunciation, even leading to the establishment of a phonological contrast (singleton-geminate) that is unattested in the target language.
\end{abstract}

\title{
Aligning Behaviors in Everyday Environments: An Ecological Perspective
}

\author{
Lucas Bietti \\ Institute for Advanced Study in the Humanities
}

\author{
Kasper Kok \\ VU University Amsterdam
}

\section*{Cienki Alan}

VU University Amsterdam

\begin{abstract}
Collaborative remembering is a joint activity that involves the establishment and reinforcement of a common ground. In this study, we investigate some ways in which collaborative remembering involves interactive coordination of non-verbal behaviors. Our data consist of video recordings of small groups of people that are reconstructing holiday memories together. For each 500 ms , these videos have been annotated in terms of the participants' bodily behavior and posture, including variables such as manual gesture, shoulder shrugs, leaning direction and gaze. We compared instances of 'simultaneous alignment' (two or more people concurrently performing the same behavior) and 'sequential alignment' (two people performing the same behavior in short temporal succession) to chance baselines, and found that the latter is more common that the former. Our analysis furthermore suggests that the degree to which participants coordinate their behaviors is stable across the course of the conversation (i.e., time-independent), but connected with specific activities within the larger discourse of collaborative remembering.
\end{abstract}

\title{
Preschoolers understand the representational and communicative nature of iconic gestures
}

\author{
Simone Bijvoet-van den Berg \\ University of Stirling, Stirling, Scotland, United Kingdom \\ Ulf Liszkowski \\ University of Hamburg \\ Elena Hoicka \\ University of Sheffield
}

\begin{abstract}
Twenty 3.5- to 4-year-olds participated in a study to investigate children's understanding of the representative and communicative nature of iconic gestures. Two toys, one of them with a sticker attached, were presented to the child. It was not possible to request the toy with the sticker by asking (experimenter wore headphones) or pointing (toys were too close together), but they could show the experimenter which toy they wanted by performing the correct gesture. Children had to generate the correct iconic gestures themselves as the gestures were not modeled during test trials. On \(70 \%\) of the trials children performed a correct gesture ( \(\mathrm{p}=.045\) ), instead of only producing other response types ( no response, verbal request, wrong gesture, pointing). This study shows that children understand that iconic gestures can represent objects, and also that they can use iconic gestures to communicate.
\end{abstract}

\title{
Modeling the genesis of a novel communicative system
}

\author{
Mark Blokpoel \\ Radboud University, Donders Institute of Brain, Cognition and Behaviour, Nijmegen, Gelderland, The Netherlands \\ \section*{Todd Wareham} \\ Memorial University of Newfoundland, St. John's, NL, Canada \\ Ivan Toni \\ Radboud University, Donders Institute of Brain, Cognition and Behaviour, Nijmegen, Gelderland, The Netherlands \\ \section*{Iris van Rooij} \\ Radboud University, Donders Institute of Brain, Cognition and Behaviour, Nijmegen, Gelderland, The Netherlands
}

\begin{abstract}
Recent work in experimental semiotics has started to investigate the cognitive processes supporting the emergence of human communicative systems. We present a computational model of the cognitive processes involved in establishing a novel referential communicative system, as operationalized with the Tacit Communication Game (TCG). This experimental paradigm has been used to study the socio-cognitive underpinnings of human communication. We model how players of the TCG can successfully generate and understand communicative behavior in a novel, visuospatial domain using Structure Mapping Theory (SMT).

Many of the processes necessary to communicate in this game are forms of analogical reasoning that are captured by SMT (e.g. abstraction and analogical transfer). Yet, we also identify cognitive processes-not yet formalized under SMT—that are necessary for the genesis of new communicative systems. This is an important first step in formally characterizing this creative socio-cognitive ability.
\end{abstract}

\title{
Metacognition in Decision-making - Influences of Signal Strength and Reliability
}

\author{
Annika Boldt \\ University of Oxford \\ Vincent de Gardelle \\ Laboratoire de Psychologie de la Perception, CNRS; Université Paris Descartes \\ Christopher Summerfield \\ University of Oxford \\ Nick Yeung \\ University of Oxford
}

\begin{abstract}
Empirical evidence has accrued suggesting that we are able to evaluate our own thoughts and actions by means of metacognitive judgements. We are interested in how these are formed and what evidence they are based on. It has often been assumed that decision time is a frugal cue for confidence judgements: the longer it takes us to form a decision, the less certain we are. It could be, however, that this association is just a by-product of the underlying mechanisms, one of which could be variability in the accumulation of evidence. In our experiment, participants had to judge whether the average colour of an array of eight coloured shapes was either red or blue. We critically manipulated the variability of information in this multi-element array. Our results suggest that for conditions with matched difficulty, variability had a significant influence on confidence with more variable arrays leading to less confident judgements.
\end{abstract}

\title{
Contextual Renewal and Awareness: Dissociating awareness within an electrodermal conditioning paradigm.
}

\author{
W. A. Bowditch \\ University of Exeter \\ Rossy McLaren \\ University of Exeter
}

\author{
A. McAndrew \\ University of Exeter
}

\author{
Ian P. L. McLaren \\ University of Exeter
}

\begin{abstract}
This paper explores two current issues in human conditioning: It addresses whether human Pavlovian conditioning is the product of a single propositional system (Mitchell, De Houwer, \& Lovibond, 2009) or dual-systems; one propositional in nature and dependent on logical reasoning, the other functional, link-based, and dependent on statistical contingency (McLaren, Green, \& Mackintosh, 1994). Additionally, the current experiment provides insight into the processes underlying both A-B-A and A-B-C contextual renewal in humans; a process that is often explained using functional (Bouton, 2004; Pearce, 2002) or propositional (Havermans, Keuker, Lataster, \& Jansen, 2005) accounts. Participants were exposed to an electrodermal contextual renewal paradigm in a bi-conditional design, whilst measures of both conscious expectancy and autonomic skin response were collected. Our results demonstrated that, despite successful acquisition and extinction for both measures, contextual renewal was only observed for conscious expectancy in the A-B-A renewal condition.
\end{abstract}

\title{
Mechanisms of Active Causal Learning
}

\author{
Neil Bramley \\ University College London, London \\ David Lagnado \\ University College London,London
}

\section*{Maarten Speekenbrink}

University College London,London

\begin{abstract}
Existing studies on causal structure learning are largely restricted to single-shot interventions, usually in constrained or deterministic scenarios. However, real world causal learning is generally noisy, incremental and constrained only by prior beliefs. Here we describe experiments where participants were incentivised to infer the causal structure of probabilistic models through the free selection of multiple interventions. Participants' sequences of intervention choices and on-line structure judgements were measured against those of an efficient Bayesian learner, which integrates information perfectly and intervenes to maximise expected utility. Successful participants were systematic and learned effectively, but chose markedly different intervention sequences to those of a Bayesian learner. Overall, we find evidence suggesting that causal structure learning is achieved by iteration of simple action-selection and causal-attribution mechanisms.
\end{abstract}

\title{
Collaborative coactivation in visual search
}

\author{
Allison Brennan \\ University of British Columbia, Vancouver, British Columbia, Canada \\ James Enns \\ University of British Columbia
}

\begin{abstract}
Individuals experience a redundancy gain when they respond faster to two signals than one. This benefit can derive from statistical facilitation of independent decisions (Raab 1962) or from the co-activation of signals before a decision (Miller, 1982). Here we applied these tests to the redundancy gain that occurs when pairs of participants work together to detect targets. We also compared gains when each partner was responsible for one of two targets versus a different spatial region. The results showed pairs were more efficient than individuals, and that this gain was greater when the task was divided by target identity versus by space. We also found that the collaborative redundancy gain could be characterized as co-activation, meaning that the benefit of collaboration exceeded that predicted by statistical facilitation. These results serve as a proof of concept that models developed to understand information processing in individuals can help characterize collaborative performances.
\end{abstract}

\title{
What and When of Removing Virtual Gallbladders: Impact of Spatial and Temporal Aspects for Understanding Complex Diagrams
}

\author{
Angela Brunstein \\ RCSI Bahrain \\ Joerg Brunstein \\ Manama, Bahrain \\ Anam Waheed \\ TAMU-Q \\ Bakr Nour \\ WCMC-Q
}

\begin{abstract}
Reading complex graphs has been shown to be difficult for diverse groups of participants from different backgrounds and different levels of expertise in the task domain. This study investigated the impact of temporal and spatial task aspects for reducing complexity and increasing performance for a surgical task with medical students.

85 premedical and medical students solved an anatomy test on the gallbladder, a test on the steps of the procedure of removing the gallbladder, and a test combining both in a web-based survey. For all 3 tasks, performance increases with years in college. More interestingly, medical students performed best for the combined task and worst for the anatomy task.

This implies students could use temporal and spatial relationships to overcome knowledge gaps to solve the complex task. Having an idea what to do first and where to do it, helped students to reconstruct a complex surgical task.
\end{abstract}

\title{
Normative Dimensions of Generic Knowledge
}

\author{
Dennis Bublitz \\ CUNY, New York, NY, United States \\ Sandeep Prasada \\ CUNY, New York, NY, United States
}

\begin{abstract}
Some of our generic knowledge is based upon what we consider to be normal instances of kinds of things. We expect a normal dog to be four-legged; if it is not four-legged, we assume that something is wrong with this particular dog, or that it is incomplete as a kind of thing dog. Prasada \& Dillingham (2009) showed that one reason we expect certain properties to be present is because we understand them to be aspects of the kinds of things. Our research offered an alternative hypothesis: these normative expectations are due to these distinct properties being beneficial in some way. Four experiments investigated this using statements that prompted responses for normative expectations. We found that while the beneficence of these properties does underwrite normative expectations to an extent, the predominant understanding was that these expectations were grounded in the aspectual quality of these properties.
\end{abstract}

\title{
A hierarchical Bayesian model of conceptual knowledge transfer
}

\author{
Steven Buckingham \\ University of Oxford, Oxford, Oxfordshire, UK
}

\author{
Alice Redfern \\ University of Oxford \\ Christopher Summerfield \\ University of Oxford
}

\begin{abstract}
Agents generalise abstract conceptual knowledge across different contexts. For example, an individual negotiating a new computer program will draw upon experience with similar programs, such as how to use a drop-down menu. What are the rules governing such knowledge transfer? Here we offer a formal Bayesian account of generalisation, in which observers update a hierarchical model that incorporates knowledge about the statistical moments of the distribution from which information is drawn. We use this model to predict performance on a foraging task that involved hunting for hidden rewards in a virtual two-dimensional grid environment. In this task, contextual cues signalled not only the likely reward location (bivariate mean), but also the pattern (bivariate dispersion). Observers optimally integrated noisy cues about the probable reward location with information from these cues. This model and data offer a formal account of how humans learn abstract conceptual information.
\end{abstract}

\title{
Acquiring Basic Skills in Changing Task Environments: Transfer of Working Memory Updating
}

\author{
Trudy Buwalda \\ University of Groningen \& Carnegie Mellon University
}

Aryn Pyke
Carnegie Mellon University

\author{
Shawn Betts \\ Carnegie Mellon University
}

Niels Taatgen
University of Groningen
John Anderson
Carnegie Mellon University

\begin{abstract}
Acquiring new skills is not a set-in-stone process and students often take various paths to the goal; the acquisition of the required skill. To assess this learning process, previous studies used hidden Markov models to separate the cognitive stages of a problem solving task similar to solving algebraic equations (e.g. Anderson et al., 2012; Anderson, 2012). Because of the slow nature of fMRI recordings, this method can only discriminate between relatively long states in the process. This study extends the approach by including eye movements as a predictor of state, in an attempt to increase temporal resolution of the method. The results show that eye movements can be used to trace the characteristics of the problem the subject is working on. Because tracking eye movements is a non-invasive measure that can be used outside experimental settings, this can benefit the discovery of problems students encounter while solving algebraic problems.
\end{abstract}

\title{
The representation of polysemous words in semantic and episodic memory
}

\author{
Álvaro Cabana \\ Facultad de Ciencias - Universidad de la República, Uruguay \\ Camila Zugarramurdi \\ Facultad de Psicología - Universidad de la República, Uruguay \\ Juan Valle-Lisboa \\ Facultad de Ciencias - Universidad de la República, Uruguay
}

\begin{abstract}
Many controversies in cognitive science hinge around the divide between the general and the particular. In language research, the Declarative Procedural (DP) model proposes that procedural memory deals with the generalizable aspects of grammar, while exceptions are handled by declarative memories. Extending the DP model, we believe that the existence in memory of a semantic component which stores the prototypic information and an 'episodic' component that stores both the exceptions to the prototypes and the exceptionally common stimuli, could explain results on polysemia research. We studied the representation of polysemous words. We tested whether different senses of a polysemous word prime each other. Although in general there is no priming there are items showing positive priming and others showing inhibition. We then used bimodal priming in order to understand the effect of context in both types of items. Our results support the idea that lexical representation uses different memory systems.
\end{abstract}

\title{
Thinking in ways we don't speak: Evidence for a universal preference in semantic granularity
}

\author{
Alexandra Carstensen \\ University of California, Berkeley
}

Grace Neveu
University of California, Berkeley
Lev Michael
University of California, Berkeley
Terry Regier
University of California, Berkeley

\begin{abstract}
Languages partition the world in different ways-for example, the categories named by spatial terms vary substantially across languages. Yet beneath this linguistic variation there may lie universal cognitive tendencies. Khetarpal et al. (2010) found that speakers of Dutch and English, despite differences in their linguistic spatial systems, sorted spatial scenes similarly-and more like the finer-grained language, Dutch. We asked whether this preference for fine-grained sorting extends to two new languages: Máíhiki, a language of Peruvian Amazonia, with a fine-grained spatial system, and Chichewa, a Bantu language of southeast Africa, with a coarse-grained spatial system. Despite the great range in spatial naming represented across these languages-both in the granularity and the shape of their spatial categories-we found that speakers of all four languages sorted finely, and thus similarly to the finer-grained languages, Máíhiki and Dutch. These results suggest that spatial cognition, unlike spatial language, is universally fine-grained.
\end{abstract}

\title{
Coordinative Dynamic of Hierarchical Visual Processing
}

\author{
Ramon Castillo \\ Universidad de Talca - Univeristy of Cincinnati \\ Heidi Kloos \\ Univeristy of Cincinnati \\ John Holden \\ Univeristy of Cincinnati \\ Michael Richardson \\ Univeristy of Cincinnati
}

\begin{abstract}
In order to navigate and make sense of one's surroundings, a person must integrate pieces of information into larger wholes. On the flip side, the person also must be able to differentiate among more detailed pieces of information. Adaptive functioning requires the coordination of both processes, where one can flexibly switch from integrated higher-order patterns to differentiated details and vice versa. What is the nature of this coordination? Through fractal and recurrence quantification analyses used in three visual tasks (search, matching and classification), we provide evidence that this coordination has properties of dynamical systems, modulated by task features. Findings are discussed in terms of organism-environment coupling; in which local-global visual processing is conceptualized as a soft-assembled and self-organized system.
\end{abstract}

\title{
Language experience modifies the processing of visual input
}

\author{
Sarah Chabal \\ Northwestern University \\ Viorica Marian \\ Northwestern University
}

\begin{abstract}
The current study examined whether language experience affects the processing of visual information. SpanishEnglish bilinguals and English monolinguals completed a visual search task in which no overt linguistic information was provided. Participants were shown an image of a target (e.g., a chair) and were asked to locate that object from an array of four images while their eye-movements were tracked. English Competitor trials contained an item whose English name overlapped phonologically with the English name of the target (e.g., chair-chain); Spanish Competitor trials contained an item whose Spanish name overlapped phonologically with the Spanish name of the target (e.g., silla-silbato [chair-whistle]). Whereas all participants looked more often at English Competitor items than at items that did not overlap phonologically with the target, only the Spanish-English bilinguals looked more often at Spanish Competitor items. Results suggest that speakers with different language backgrounds vary in how they respond to non-linguistic, visual information.
\end{abstract}

\title{
Understanding Emergent Phenomena with Functional Schema
}

\author{
Margaret Chan \\ Columbia University
}

\begin{abstract}
Previous research has well-documented that naïve learners struggle when attempting to understand emergent phenomena. Misconceptions often arise as learners tend to apply patterns of cause-and-effect between entities to explain these emergent processes. We posited that comprehending emergence requires learners to construct a different conceptual model, namely a functional schema, emphasizing functional relationships among entities and their interactions that are central to how emergent phenomena arise. A promising strategy to promote generation of such functional schema is contrasting examples. This paper reported an intervention study with 86 middle school students examining the effect of contrasting scenarios in helping learners generate a functional-relationship-centered schema to understand global warming. Students' correct and misconceived explanations in pretest-posttest protocols were analyzed. Results showed contrasting scenarios motivated learners to develop the critical functional schema, which led to their eventual understanding of the mechanism of global warming. Implications on schema construction on understanding emergent systems are discussed.
\end{abstract}

\title{
Pedagogy and Technology: Balancing student-centric learning with teacher-centric teaching
}

\author{
Stephen Wee Hun Lim \\ National University of Singapore
}

\begin{abstract}
Technology can be used to create student-centric learning in diverse ways, including the use of social media platform(s) and interactive simulations to teach inductive thinking, and games to facilitate self-directed learning. But, teachercentric teaching is important as student-centric learning is; both aspects contribute to the student learning process, although the former has hardly been considered explicitly. Specifically, while technology develops apace, that teachers might continue to prefer traditional teaching modes and styles is an issue that should not be taken lightly, because teaching performances can be compromised if these modes and styles were compelled to evolve prematurely. I will particularly discuss the apprehensions that teachers might have as a technology-based teaching culture rapidly emerges, and how the dissonance between (non-)preferred modes of teaching and learning in this context might be resolved, in order to promote a technology-enhanced education system that ultimately benefits both teachers and students in a practical way.
\end{abstract}

\title{
Modelling Word and Object Naming in Pure Alexia
}

\author{
Ya-Ning Chang \\ Neuroscience and Aphasia Research Unit, University of Manchester \\ Matthew Lambon Ralph \\ Neuroscience and Aphasia Research Unit, University of Manchester \\ Steve Furber \\ Advanced Processor Technologies Group, University of Manchester \\ \section*{Stephen Welbourne} \\ Neuroscience and Aphasia Research Unit, University of Manchester
}

\begin{abstract}
Pure alexia (PA) is characterised by abnormally strong length effects in word reading times. It is often thought to result from damage to visual processing. This visual damage was also found to cause impaired object recognition performance (e.g., Roberts et al., 2012), suggesting a general visual deficit. Many computational models of reading have successfully simulated different forms of acquired dyslexia (e.g., Coltheart et al., 2001; Plaut et al., 1996). However, an adequate computational account of pure alexia has yet to be produced. We developed a large-scale and complete connectionist model with asymmetric hemisphere processing to support both word and object recognition. When damage was applied to the left visual processing layer in the model, the model produced abnormal length effects and impaired object recognition similar to those seen in PA patients. The results provide evidence to support the view of a common visual processing in visual word recognition.
\end{abstract}

\title{
Masked Repetition Effects of Chinese Compound Words Interact with Word Frequency because of Properties of Compound
}

\author{
Sau-Chin Chen \\ Department of Human Development, Tzu-Chi University, Taiwan \\ Zi-Han Chen \\ Department of Human Development, Tzu-Chi University, Taiwan
}

\begin{abstract}
Masked repetition has been suggested as the reliable paradigm investigating the pre-lexical processing of English words. This suggestion has the supportive evidence that masked repetition effects do not interact with word frequency. Chinese compound words are the combinations of two constituent characters which have independent lexical properties. The lemma model argues that the lemmas of constituents would increase masked repetition effects of low-frequency compound word when the constituent characters have relative higher frequency. To verify this argument, this study manipulated the frequency and the morphological aspects of constituent characters. The critical results are the increased masked repetition effect of low-frequency coordinative words which have at least one high-frequency characters and the null interaction of masked repetition effect and word frequency for the compound words having meaningless constituents. The approaches to modify the current lemma model are discussed according to these findings.
\end{abstract}

\title{
Writing-related Visual-motor Abilities in Children with Developmental Coordination Disorder
}

\author{
Rong-Ju Cherng \\ National Cheng Kung University \\ Yu-Han Su \\ National Cheng Kung University \\ Ting-Hui Wang \\ National Cheng Kung University \\ Yung-Jung Chen \\ National Cheng Kung University
}

\begin{abstract}
Twenty-five children with developmental coordination disorder (DCD) and 25 age- and gender-matched typically developing (TD) children were tested on writing-related visual motor tasks with a self-developed electronic assessment tools: "Writing Start". Children with DCD were diagnosed according to the DSM-IV diagnostic criteria. TD children were recruited from community. Twenty figures with different complexities and sizes were used in the tasks. Error number, error time, and error pathway, each representing a deficit in action, temporal and spatial motor control of writing movement, were compared between groups. The error number was significantly larger in the DCD group than that in the TD group. The error time and error pathway were also longer in the DCD group than that in the TD group, especially when the figures were smaller and more complicated. Children with DCD indeed have difficulties in mastering visual motor skills for writing.
\end{abstract}

\title{
The construal level theory and the dual process theory in multi-attribute decision making: An empirical examination
}

\author{
Itsuki Chiba \\ Department of Psychology, Rikkyo University, Japan \\ Takashi Tsuzuki \\ Department of Psychology, Rikkyo University, Japan \\ Kikuchi Manabu \\ Department of Psychology, Rikkyo University, Japan
}

\begin{abstract}
The compromise effect and attraction effect are examples of irrational choice in multi-attribute decision making. Their underlying mechanisms are assumed to be differences in the trade-off structure of each choice set. Recent research on the construal level theory showed that a high construal level decreases the compromise effect, but increases the attraction effect. Further, studies on the dual process theory showed that the depletion of participants' cognitive resources increases the attraction effect, but eliminates the compromise effect. It is therefore important to examine compensatory (or uncompensatory) information search based on trade-off structure in these context effects. We examined the influence of the construal level of participants' choices and the depletion of participants' cognitive resources for the above two context effects. Furthermore, we analyzed eye movements as a measure of the information search process, to examine the relationship between the construal level theory and dual process theory.
\end{abstract}

\title{
Beyond Berlyne's Conjecture: The Aesthetic Quality of Visual Patterns
}

\author{
Susan Chipman \\ unaffiliated
}

\begin{abstract}
Berlyne, the Canadian psychologist, famously conjectured that the aesthetic preference for visual patterns is an inverted-U function of their complexity. In my own research (Chipman, 1977), I developed and studied a large set of patterns, exploring what determined their judged complexity. These included patterns with several types of well defined structure, as well as randomly generated patterns. Several experiments explored the judged aesthetic quality of such patterns. Judged aesthetic quality was not any simple function of judged complexity. Not surprisingly, there are significant individual differences in aesthetic preference, including differences in preference for different types of visual structure.
\end{abstract}

\title{
Incremental Information Processing on Visual Search: The Critical Role of Delivery Rate
}

\author{
Eric Chiu \\ University of California, Merced (UCM)
}

Michael Spivey
University of California, Merced

\begin{abstract}
Recent studies show that visual search is often not well characterized as either a purely parallel or serial search strategy. Subsequently, the literature and computational models have evolved from traditional parallel and serial descriptions to a continuum of search efficiency. It has been demonstrated that search efficiency does not improve with simultaneous delivery of target features in a conjunction-search task. Interestingly, search efficiency does improve when non-linguistic visual delivery of target features appears incrementally and concurrently with the display onset, but not prior to display onset. In our current experiment, we explore the temporal constraints of the facilitatory effect found with concurrent incremental information processing. The results explain that linguistic and non-linguistic mediation of visual search, provided sufficient time to process, is chiefly due to the incrementality of target feature delivery when search has begun. This finding supports an interactive account of visual attention.
\end{abstract}

\title{
From Logical Positivism to Philosophical Thinking: Impacts and Implications on the Student-Learning Process
}

\author{
Dian Yi Chow \\ Hwa Chong Institution (College) \\ \section*{Stephen Wee Hun Lim} \\ National University of Singapore
}

\begin{abstract}
In schools today, students pursue individual academic subjects, such as physics, mathematics, and biology. By compartmentalizing these subjects, students may not be able to connect concepts from different disciplines competently to explain universal phenomena that occur in everyday living. Under this view, it is essential to inculcate an overarching philosophy with which one can capably unify all taught subjects. We first consider Logical Positivism as an overarching philosophy, and the benefits and implications of teaching it to students. We then illuminate the importance of inculcating broad-based philosophical thinking in students, and discuss how this pedagogical approach deepens students' understanding in, and increases their interest towards, the subject matter that they learn, as well as trains students' abilities to examine and interpret natural phenomena in logical ways.
\end{abstract}

\title{
FOXP2 as the Genetic Basis for the Capacity to Shift Between Analytic and Associative Modes of Thought
}

\author{
Courtney Chrusch \\ University of British Columbia \\ Liane Gabora \\ University of British Columbia
}

\begin{abstract}
It was previously proposed that the burst of creativity in the Middle/Upper Paleolithic following the appearance of anatomically modern humans was due to the onset of contextual focus, the capacity to shift between an associative mode of thought conducive to forging connections and breaking out of a rut, and an analytic mode conducive to logical problem solving. Hominids could then generate ideas in an associative mode, and refine them in an analytic mode, and process representations at multiple levels of detail, and from different perspectives. This resulted in richer understandings of their world. It is proposed that the FOXP2 gene, which evolved at this time, is responsible for onset of contextual focus. FOXP2 thereby created an unprecedented need for language to (a) keep track of representations for oneself, and (b) capitalize on different perspectives of others. This explains why FOXP2 is implicated in language but not uniquely associated with it.
\end{abstract}

\title{
The roles of configuration and orthography in Chinese recognition: a developmental approach
}

\author{
Yi-Ling Chung \\ National Cheng Kung University
}

\author{
Hsin-Hsuan Wu \\ National Cheng Kung University
}

Yu-Shu Chiang
National Cheng Kung University

\section*{Jon-Fan Hu}

National Cheng Kung University

\author{
Chiu-Hua Huang \\ National Cheng Kung University
}

\author{
Hsueh-Chih Chen \\ National Taiwan Normal University
}

Chien-Chih Tseng
National Taiwan Normal University

\section*{Li-Yun Chang}

University of Pittsburgh

\begin{abstract}
In Chinese, character configuration and orthographic combination are acknowledged to influence character recognition. Also, it was widely accepted that lexical access ability progresses as reading skills improved and vocabulary increased. The present study aims to reveal the effects of configuration type and radical properties by comparing different development stages of Chinese learning. A character decision task was used in which radical position-based frequency and radical position regularities within two different configurations were manipulated. 15 third-grade 27 sixth-grade schoolers, and 41 undergraduate students were asked to identify whether the 120 pseudo-words conform to radical position regularities. Accuracy was recorded as measurement to examine the effects of configuration and orthographic combination. The analysis revealed different patterns of frequency effect and regularity effect between the two configurations. Furthermore, age variation was observed for both types of configurations. In conclusion, orthographic and configuration knowledge are acquired gradually and play distinct roles in Chinese recognition.
\end{abstract}

\title{
Orthography and configuration on Chinese literacy acquisition: evidence from eye movement
}

\author{
Yi-Ling Chung \\ National Cheng Kung University \\ Yu-Shu Chiang \\ National Cheng Kung University
}

Hsin-Hsuan Wu
National Cheng Kung University

\author{
Jon-Fan Hu
}

National Cheng Kung University
Chiu-Hua Huang
National Cheng Kung University

\author{
Hsueh-Chih Chen \\ National Taiwan Normal University
}

Chien-Chih Tseng
National Taiwan Normal University

\section*{Li-Yun Chang}

University of Pittsburgh

\begin{abstract}
Previous research assumed the eye movement patterns change along with the growth of word knowledge while children become skilled readers. In this study, the effects of orthographic and configuration information on Chinese character recognition were investigated by comparing eye movement patterns from a developmental perspective ( 15 third-grade 27 sixth-grade schoolers, and 41 undergraduates). Eye movement patterns were recorded in a character decision task by varying configuration type (left-right, up-down), radical position-based frequencies (HH, HL, LH, LL), and radical position regularities (P, SN, WN). The results showed that the two different configurations lead to different eye movement patterns: (1) radical position regularity effect was only significant at the left-right configuration for all age groups; (2) frequency effect and development variation appeared for both two configuration types. These findings highlight the importance of configuration knowledge and orthographic awareness for learning Chinese characters.
\end{abstract}

\title{
"You talkin' to me"? Understanding communicative intentions recruits the mirror and the mentalizing system
}

\author{
Angela Ciaramidaro \\ University of Frankfurt \\ Cristina Becchio \\ University of Turin \\ Livia Colle \\ University of Turin \\ \section*{Bruno Bara} \\ University of Turin \\ Henrik Walter \\ Charité University of Berlin
}

\begin{abstract}
In our daily social interactions we can infer others people intentions from the observation of their actions. In general, two brain systems, the mentalizing and the mirror neuron system, have been implicated in understanding other's intentions. However, there is little knowledge whether and how these two systems may cooperate in correctly understanding communicative interactions. We used functional MRI to establish how mirror and mentalizing regions contribute to the implicit encoding of communicative intentions, proposing that being directly involved during social interaction would be mediated by both systems. In particular, we investigated the involvement of those systems in distinguishing communicative from private intentions as well as other directed ("third-person perspective") from self-directed ("second-person perspective") intentions. Categorical and functional connectivity analyses showed that the mentalizing and the mirror neurons system were simultaneously involved in processing communicative intentions in general and more strongly coupled in self directed communicative actions.
\end{abstract}

\title{
How does this thing work? Evaluating computational models of intervention-based causal learning
}

\author{
Anna Coenen \\ New York University
}

\author{
Bob Rehder
}

New York University

\section*{Todd Gureckis}

New York University

\begin{abstract}
The present study explores how people learn about a causal system by interacting with it. Participants were given the task to identify the operation of virtual '"computer chips" by setting the value of various components and observing how those interventions influenced the setting of other components. Across conditions we manipulate the complexity of the causal system (i.e., number of nodes and connections), the number of alternative hypotheses (i.e. possible causal graphs) on each trial, and aspects of the "temporal stability" of the learning environment (if repeated interventions were made on a single, stationary system or if the system reset to different starting states following each intervention). Interventions were modeled by comparing them to an optimal Bayesian learner who chooses interventions to quickly reduce uncertainty about the structure. Our results suggest that naive Internet-recruited subjects choose highly informative interventions, but also deviate from the predictions of the optimal model in certain ways.
\end{abstract}

\title{
Individual differences in shape bias are predicted by non-linguistic perceptual ability
}

\author{
Beverly Collisson
}

University of Calgary

\author{
Bernard Grela \\ University of Connecticut
}

Tammie Spaulding
University of Connecticut
Jay Rueckl
University of Connecticut and Haskins Laboratories
James Magnuson
University of Connecticut and Haskins Laboratories

\begin{abstract}
Children with Specific Language Impairment (SLI) lag behind peers with typical language (TL) in vocabulary. We ask what impact this lag has on shape bias (generalization based on shape in Naming contexts ["here's a dax; find another dax"] but not Classification contexts ["look at this; find another like this"). Smith (2000) argues that shape bias depends on and drives vocabulary development; vocabulary is a basis for detecting covariation between objects and names, and name learning accelerates once the bias emerges. 51 three and four year-old children ( 16 SLI, 16 matched TL, 19 additional TL) participated in Naming and Classification tasks, a paired visual association (PVA) task, and an assessment battery. The SLI group was significantly worse at PVA and did not exhibit a shape bias. Individual differences revealed wide variation in both groups, and that shape choices in Naming were better predicted by PVA than standardized assessments.
\end{abstract}

\title{
Temporal motor dynamics in inductive reasoning
}

\author{
Steve Croker \\ Illinois State University, Normal, Illinois, USA
}

\section*{Corinne Zimmerman}

Illinois State University

\begin{abstract}
We investigated the temporal dynamics of response choice in a decision-making task by examining the evolving implicit responses indicated by hand movements made before an explicit response is selected. Participants ( \(\mathrm{N}=31\) ) judged which of two cars would go faster when the underlying rule was plausible or implausible and when two response choices differed with respect to one or two causal variables. Participants completed 300 trials in five blocks. We found an interaction between trial type, block, and plausibility of rule. In earlier trials in the implausible rule condition, there was greater deviation towards the distracter response before selecting the correct response. Participants given implausible rules demonstrated less activation of competing representations over time as they induced the underlying rules, particularly on trials in which response choices differed on both causal variables. Mouse trajectories did not change across blocks for participants given the plausible rule, suggesting they learned the rule early.
\end{abstract}

\title{
Where is the insight in insight problems?
}

\author{
Amory H. Danek \\ Ludwig-Maximilians-Universität München, Munich, Germany \\ \section*{Michael Öllinger} \\ Parmenides Foundation and Ludwig-Maximilians-Universität München, Munich, Germany
}

\begin{abstract}
Insightful problem solving is a vital part of human thinking, yet very difficult to grasp. The "Aha! experience" is often regarded as the defining characteristic of insight. Traditionally, insight has been investigated by using a set of established "insight tasks", assuming that insight has taken place if these problems are solved. However, the debate about which problems actually trigger insight is still not resolved, since there is no clear behavioural marker for the occurrence of insight. In the present work, we therefore aimed at testing the validity of three classical insight problems by directly asking participants about their solution experiences. Our results suggest that participants solve insight problems also without any Aha! experience, casting doubt on the common approach of using a priori defined insight problems. Consequently, we advocate the use of direct insight ratings by participants, determining for each problem individually whether it was solved with insight or not.
\end{abstract}

\title{
Graphical Overviews Structure Online Learning: Evidence from Eye-Tracking
}

\author{
Sarah Davies \\ University of Utah \\ Kirsten Butcher \\ University of Utah
}

\author{
Anne Cook \\ University of Utah
}

\begin{abstract}
Previous research has examined the use of graphical overviews as a way to structure students' self-regulated, online learning. Although findings have suggested that graphical overviews can improve learning from online content, there has been little direct evidence as to why this benefit may occur. In this research, eye tracking and verbal protocols were gathered as 26 pre-service teachers used a graphical overview or keyword interface to choose online resources for an educational task. Fixation times and pupil diameter were analyzed as measures of cognitive effort; results demonstrated that pupil diameter was significantly lower when participants used the graphical as compared to the keyword interface. Protocol analysis was used to examine the depth of processing during search and evaluation; results showed that participants using the graphical overview engaged in deeper analysis of domain content. Results provide evidence for the importance of graphically-based cognitive offloading during "searching to learn" tasks.
\end{abstract}

\title{
Cultural Variation in Families' Shared Engagement
}

\author{
Andrew Dayton \\ University of California Santa Cruz \\ Barbara Rogoff \\ University of California, Santa Cruz
}

\begin{abstract}
Cultural Variation in Families' Shared Engagement
This study examines whether Indigenous-heritage Mayan mothers and their young children are more likely than middleclass European American mothers and their children to engage by blending agendas in fluid collaboration while exploring novel objects together. Fluid collaboration appears to be encouraged in many Indigenous communities of the Americas, where children often collaborate in ongoing family and community endeavors (Rogoff, 2003; Mozier \& Rogoff, 1993, 2003). The present research submits audio signals of videotaped observations of mothers and their children in home interactions to microanalysis using spectrographs, to compare the interactions of Mayan and European American families (Gratier, 2003, 2013; Malloch, 1999; Trevarthen, 2008). In the visual representations of their vocalizations, we find evidence that the Mayan families more frequently use smooth collaboration whereas the middle-class European American families appear to struggle more to establish a shared rhythm in their interactions.
\end{abstract}

\title{
High level influences on visual action recognition
}

\author{
Stephan de la Rosa \\ Max Planck Institute for Biological Cybernetics, Tübingen, Germany \\ Stephan Streuber \\ Max Planck Institute for Biological Cybernetics \\ Martin Giese \\ Centre for Integrated Neuroscience at the University of Tübingen \\ Heinrich Bülthoff \\ Max Planck Institute for Biological Cybernetics, Tübingen, Germany \\ \section*{Cristobal Curio} \\ Max Planck Institute for Biological Cybernetics, Tübingen, Germany
}

\begin{abstract}
Action recognition is important for social interactions. Because little is known about the visual tuning properties of processes involved in action recognition, we examined the visual tuning properties of action recognition by means of a behavioral adaptation paradigm. Participants were adapted to images showing a person hitting or waving and subsequently categorized test images showing an ambiguous action as either hitting or waving. We found the perception of the test images to be significantly biased away from the adapted action (action adaptation aftereffect (AAA)). Subsequent experiments ruled out that the AAA was not merely driven by the adaptation of local visual contrast or the emotional content of the action. However adaptation to action words (e.g. "hitting" or "waving") did not induce an AAA. Finally we found evidence for the AAA being modulated by the social context in which an action is embedded, suggesting high level influences on action recognition.
\end{abstract}

\title{
Variables Influencing the Nature of Learned Categorical Perception Effects
}

\author{
Josh de Leeuw \\ Indiana University \\ Jan Andrews \\ Vassar College
}

\section*{Ken Livingston}

Vassar College

\begin{abstract}
Considerable research has demonstrated so-called learned categorical perception (CP) effects, where learning to classify a set of stimuli leads to either compression (within-category stimuli judged to be more similar and/or confusable than before learning) or expansion (between-category stimuli judged to be less similar and/or confusable than before learning) or both. The issue of why category learning causes one type of effect or the other has not been systematically investigated, but previous research suggests that highly discriminable stimuli may tend to produce compression while stimuli that are difficult to discriminate may tend to produce expansion. We report a series of studies testing the effect of stimulus discriminability on the type of learned CP effect produced using both similarity and XAB measures of the effects. Preliminary results suggest that different measures of categorical perception reveal different effects, and that category structure may also be relevant.
\end{abstract}

\title{
Limits in Reasoning about False Beliefs in Adults: the Effect of Priming or the Curse of Knowledge?
}

\author{
Agnieszka Debska
}

University of Warsaw
Krystyna Komorowska
University of Warsaw

\begin{abstract}
Birch \& Bloom (2007, Psych. Sc. 18, 382-386) suggest that adults' reasoning about others' mental states is influenced by their privileged knowledge about reality. When asked where a person described in the story would search for a missing object, subjects tend to judge with higher probability that the person would search in a particular box, when they know that the object is indeed in that box. However, the results of their experiment could be an effect of unintended priming in the materials, i.e., the increased attention towards the box might be also caused by reading about it in the task instructions. In a new version of the experiment, we controlled for this factor by priming different locations in the instructions. The results show that it is unlikely that priming is the source of the Birch and Bloom's observations: only knowledge about reality changes the strategies in reasoning about others' actions.
\end{abstract}

\title{
Agency and attention influence the perceived speed of a moving stimulus
}

\author{
John Dewey \\ Central European University
}

\author{
Thomas Carr \\ Michigan State University
}

\begin{abstract}
The sensory consequences of intentional actions are perceived to be attenuated compared to equivalent but externally generated stimuli. It is thought that forward models in the sensorimotor system partially cancel predictable reafferent sensory feedback in order to bias attention towards more novel or unexpected stimuli. But does merely observing familiar actions also trigger forward models with attendant sensory attenuation? Previous studies investigated this question in the auditory modality with conflicting results. We conducted two attenuation experiments in a visual modality to conceptually replicate and generalize previous findings (Exp 1), and to control for differences in temporal predictability and attention which may have confounded previous studies (Exp 2).

We found that movements initiated by humans (self or other) were attenuated compared to computer movements, and selfinitiated movements were the most attenuated. Adding Go signals prior to movements counteracted attenuation. Perceived speed is thus influenced by agency as well as attention.
\end{abstract}

\title{
When does the causal information externally given affect causal inferences?
}

\author{
Kyung Soo Do \\ Sungkyunkwan University, KOREA \\ Rae-yeop Park \\ Sungkyunkwan University, KOREA
}

\begin{abstract}
Many previous studies showed that the causal information externally given (hereafter external information) did not succeed in changing the causal structure driven from the covariation data when the external information and the covariation data were in conflict. We speculated that the salience of the external information is crucial in causal inference. The external information did not affect the causal inference when the external information and the covariation data were simultaneously presented in Experiment 1. However, when the external information and the covariation data were sequentially presented and participants were asked to report the causal structure and the strength each time in Experiments 2 and 3, participants were more likely to report the causal structure of the external information when the covariation data were drawn from a different causal structure. Results of the three experiments showed that the external information can override the covariation data under certain conditions.
\end{abstract}

\title{
Do voices survive lexical consolidation?
}

\author{
Nicolas Dumay \\ Basque Center on Cognition, Brain and Language \\ Jeffrey Bowers \\ University of Bristol, UK
}

\begin{abstract}
Learning new words involves consolidation. After one night's sleep, not only is explicit knowledge about the novel words enhanced, but the new words also now compete with similar-sounding existing words (Dumay \& Gaskell, 2007) during word recognition. The present study assessed whether lexical consolidation strips off surface details of newly learned words, producing more abstract representations. We manipulated the speaker's voice between exposure and test. Participants learnt one set of novel competitors (such as 'shadowks' for 'shadow') seven days before the test, and another set immediately before the test. Each word was learnt in a male or a female voice, and was tested in either the same or the other voice. Cued recall and phoneme monitoring showed stronger memory for the seven-day old items and, if anything, an enhanced voice effect (i.e., better performance in the same voice condition) after seven days. Crucially, our most indirect measure of lexical competition showed that only the seven-day old items (as expected) engaged in lexical competition, but only when the input preserved the voice in which they had been encoded. These findings indicate that consolidation does not make word representations more abstract: voice specific details do not just survive lexical consolidation; they are enhanced by it.
\end{abstract}

\title{
External Normalization: Testing a Cognitive Offloading Account
}

\author{
Timothy Dunn \\ University of Memphis, Memphis, TN, United States \\ Srdan Medimorec \\ Arizona State University, Tempe, AZ, United States \\ Evan Risko \\ University of Memphis, Memphis, TN, United States
}

\begin{abstract}
When individuals are presented with complex arrays at non-canonical orientations (e.g., rotated text) they frequently physically rotate to approximate the orientation of the stimulus (i.e., external normalization). One view of this natural behavior is that individuals are offloading internal cognitive demands (e.g., internal normalization) by adopting an external solution (i.e., external normalization). We test this account here by combining a stimulus rotation manipulation with a stimulus repetition manipulation. Previous research has demonstrated that stimulus repetition reduces the cost of stimulus rotation on performance. In other words, repetition putatively reduces the "internal" costs of stimulus rotation. Thus, stimulus repetition should reduce the frequency of external normalization. Consistent with the cognitive offloading account, repetition reduced the frequency of spontaneous physical head rotations while individuals read rotated text. Discussion focuses on the implication of these results for understanding cognitive offloading and the embodied and embedded nature of cognition.
\end{abstract}

\title{
Examining the transitions between decision strategies
}

\author{
Gilles Dutilh \\ University of Basel \\ Benjamin Scheibehenne \\ University of Basel \\ Han L. J. van der Maas \\ University of Amsterdam \\ Jörg Rieskamp \\ University of Basel
}

\begin{abstract}
In many domains of cognitive psychology, it is proposed that people are equipped with a repertoire of strategies to solve the problems they face. For instance, it is proposed that when people use multiple cues to make a probabilistic inference about a criterion, they sometimes rely on simple heuristics and sometimes apply more elaborate additive strategies. Indeed, many studies suggest that people's inferences can be described by different strategies in different situations. However, critics of this view suggest that people do not apply different strategies but instead adjust one single strategy to the characteristics of each situation. Here we examine the strategies that individuals use when available resources change. Therefore, we continuously change the cost of deliberation time. The behavior of participants at the transition from fast and simple behavior to slow but optimal information integration offers insight into whether people select different strategies or continuously adjust one single strategy.
\end{abstract}

\title{
The Influence of Direct and Indirect Speech on Mental Representations
}

\author{
Anita Eerland \\ Open University \\ Jan Engelen \\ Erasmus University Rotterdam \\ Rolf Zwaan \\ Erasmus University Rotterdam
}

\begin{abstract}
Language can be viewed as a set of cues that subtly modulate the comprehender's thought processes. For example, the literature suggests that people perceive direct speech as more vivid and perceptually engaging than indirect speech. We sought to address how this alleged vividness is evident in comprehenders' mental representations in a series of experiments. Our results do not support the idea that, compared to indirect speech, direct speech enhances the accessibility of information from the communicative or referential situation during comprehension. Neither do our results support the idea that the hypothesized more vivid experience of direct speech is caused by switching from the visual to the auditory modality. However, our results do show that direct speech leads to a stronger mental representation of the exact wording of a sentence than does indirect speech. These results show that language has a more subtle influence on memory representations than was previously suggested.
\end{abstract}

\title{
Explanation and Fractions: How Preferences for Types of Explanations Affect Learning
}

\author{
Emma H. Geller \\ UCLA \\ James W. Stigler \\ UCLA
}

\begin{abstract}
Explanation is often cited as an effective learning tool, but much work remains to determine the influence of explanations on different types of material to be learned. Evidence from category learning suggests that explanation may drive the learner to identify underlying regularities that fit a general pattern (Williams \& Lombrozo, 2010). In a pilot study, we examined whether explanation could be used to improve understanding of fraction magnitudes and whether explanations of specific inequalities are more or less effective than explanations of sets of inequalities. Results revealed that generating single or set explanations did not affect test or transfer accuracy, but individuals indicated strong and consistent preferences for particular types of explanations (i.e. conceptual, procedural, or rule-based). Further studies are being conducted to identify individual differences that may predict preferences for different kinds of explanations and their effect on subsequent learning and understanding.
\end{abstract}

\title{
Language and Spatial Attention: Words Evoke Independent Codes for Objects and Locations
}

\author{
Zachary Estes
}

Bocconi University

\author{
Michelle Verges \\ Radically Independent
}

James S. Adelman
University of Warwick

\begin{abstract}
Many common words have spatial associations (e.g., "bird," "snake," "jump", "crawl") that influence perception at congruent and incongruent locations. For example, "bird" hinders identification of a square at the top of a display. Many researchers have attributed this spatial interference to location-specific perceptual simulations: The word "bird" shifts attention upward and evokes the perceptual representation of a bird, which impairs identification of an unrelated visual target either by visually masking it or by engaging the neural systems necessary for visual perception. However, we report that a large sample of nouns (Experiment 1) and verbs (Experiment 2) of high and low imageability (and visual strength) elicited equivalent spatial interference. Thus, perceptual simulation failed to explain the spatial interference effect. Experiment 3 instead supported an event coding explanation: Target objects are coded for their congruence with both the cue word and its implied location, and conflicting codes interfere with responding.
\end{abstract}

\title{
The temporal dynamics of choice blindness: flat detection rates and short-term preference alterations
}

\author{
Ilya Farber \\ Institute of High Performance Computing, A*STAR \\ Fumihiko Taya \\ SINAPSE Institute for Cognitive Science and Neurotechnologies, National University of Singapore \\ Swati Gupta \\ Institute of High Performance Computing, A*STAR \\ O'Dhaniel Mullette-Gillman \\ Department of Psychology and SINAPSE Institute for Cognitive Science and Neurotechnologies, National University of Singapore
}

\begin{abstract}
Since the original experiments on choice blindness for faces (Johansson et al 2005), many studies have extended the phenomenon to other domains but few have focused on attaining a deeper understanding of the phenomenon itself. We here report on an experiment which closely follows the original study (albeit in computerized form and with a Singaporean population), with additional elements aimed at explicating the causal structure and temporal dynamics of the underlying processes.

Contrary to intuition, we find that subjects who notice and immediately report a "manipulation" trial still miss about half of subsequent manipulation trials. Detection is also found to be highly sensitive to the form and timing of the opportunities to report. We also investigate the effects of choice manipulation on attractiveness ratings, finding that manipulations do modulate subsequent attractiveness ratings but that this effect falls below significance after a 2 -week interval.
\end{abstract}

\title{
Perceptual Word-Form Typicality Effects Are Modulated by Strength of Expectation During Reading
}

\author{
Thomas Farmer \\ Department of Psychology, University of Iowa \\ Klinton Bicknell \\ Department of Psychology, University of California-San Diego \\ Michael Tanenhaus \\ Department of Brain and Cognitive Sciences, University of Rochester
}

\begin{abstract}
A growing literature suggests that readers generate predictions about various aspects of incoming linguistic input. Do expectations from context map onto lower-level expectations, and if so, how? We propose that comprehenders use internally generated predictions to explain the source of the input. In two experiments, we tracked eye-movements as subjects read sentences that generated strong (The boy saved the xxx) or weak (Mary had the word "xxx" tattooed ...) expectations for nouns. Across contexts, subjects encountered target words (xxx) that had visual-form features that were either typical or atypical of nouns. In strongly predictive contexts, first-fixation and gaze duration measures were longer when the form of the target word was atypical with respect to the predicted category. In the less-biased contexts, no effect of word-form typicality occurred. These experiments provide eye-movement evidence that linguistic context is used to generate perceptual expectations about form-based properties of upcoming words during reading.
\end{abstract}

\title{
Towards identifying principles for clinical intervention in developmental language disorders: Establishing a neurocomputational foundation
}

\author{
Anna Fedor \\ Birkbeck College London \\ Wendy Best \\ University College London \\ Jackie Masterson \\ Institute of Education University of London
}

Michael S. C. Thomas
Birkbeck College London

\begin{abstract}
We used a simple artificial neural network model to begin the work of understanding what principles underlie effective interventions for developmental disorders of language and cognition, from the perspective of neurocomputational mechanisms of development. The work aims to complement a clinical perspective of the principles of effective intervention. Our study explored the effectiveness of different types of intervention modeled as items added to the normal training set. We assessed whether best interventions were specific to problem domains, specific to deficit types, and/or dependent on when in development they take place.
\end{abstract}

\title{
Effects of exhaustive and partial morphological segmentation vary with reading skill.
}

\author{
Laurie Beth Feldman \\ The University at Albany, SUNY; Haskins Labs
}

\section*{Kit Cho}

The University at Albany, SUNY
Petar Milin
University of Novi Sad;Tubingen University
Harald Baayen
Tubingen University

\begin{abstract}
Models of word recognition assume that information about the orthographic form of a word (morpheme) must be available before access to that word's (morpheme's) meaning is possible. In prior work we have demonstrated that semantic similarity influences even early morphological priming (Feldman, Kostić, Gvozdenović, O’Connor, \& Martín, 2012; Feldman \& Martín, 2009).

In two experiments conducted in English, we used a forward-masked lexical decision task to assess whether processing differs after exhaustively decomposable (stem+affix; e.g., pastor-PAST) and partially decomposable (stem+nonmorphemic string; e.g., pasta-PAST) primes in semantically dissimilar prime-target pairs.

Results using linear mixed effect models on inverse transformed (-1000/RT) latency data with two separate PCs for the contributions of form (negative) and for frequency (negative) and previous RT as a predictor, failed to show different patterns of facilitation after exhaustively and partially decomposable primes; both of which differed from unrelated controls. Spelling did not interact with prime type but poor spellers varied more across the session.
\end{abstract}

\title{
Supporting Self-Regulated Learning in Conceptual Feedback Environments
}

\author{
Kirsten R. Butcher \\ University of Utah \\ Lisa A. Ferrara \\ University of Utah
}

\begin{abstract}
Prior research in learning with graphic organizers has revealed that learners can make use of graphical overviews to increase their learning in online environments, especially when these visual supports are provided before learning. This research examined the extent to which the format of feedback in an online learning environment impacts students' processes and outcomes during a self-regulated learning task. Students wrote scientific essays that were analyzed by a personalized learning service (the customized learning service for conceptual knowledge: CLICK). Feedback on essay content was presented using either a visual (node-link) representation or a (text-based) list view. Preliminary data reveal that learners presented with the visually-based feedback engaged in more effective self-regulated learning processes (i.e., planning and goal-setting) compared to learners provided with list-based feedback. Results demonstrate that the format of external feedback in self-regulated learning environments play an important role in supporting students' implementation of effective self-regulated learning strategies.
\end{abstract}

\title{
Do Parents Adapt Descriptions of Spatial Relationships to Child Knowledge?
}

\author{
Katrina Ferrara \\ Johns Hopkins University, Baltimore, MD, US \\ Colin Wilson \\ Johns Hopkins University, Baltimore, MD, US \\ Katherine Kelliher \\ Johns Hopkins University, Baltimore, MD, US \\ Malena Silva \\ Johns Hopkins University, Baltimore, MD, US \\ Barbara Landau \\ Johns Hopkins University, Baltimore, MD, US
}

\begin{abstract}
Language is a collaborative act: to successfully communicate, speakers must generate semantically valid utterances that are sensitive to the knowledge state of the listener. We asked whether parents' spatial descriptions are tuned to their children's spatial knowledge. Parent-child pairs ( \(n=16\), m child age \(4 ; 1\) ) viewed identical complex spatial arrays on separate computer screens. Parents were asked to describe target objects so that their child could identify them on their own screen. Children's knowledge of left/right was independently tested using a comprehension task. A hierarchical statistical model of the experimentally elicited spatial language predicted that the probability of parents using left/right was greater for children that achieved higher comprehension scores, indicating successful communicative adaptation. This result did not hold for parents of children with severe spatial impairments (Williams Syndrome), suggesting that there is considerable variation in how well parents tune their language to their children's level of spatial language and knowledge.
\end{abstract}

\title{
Using Graphical Organizers to Improve Self-Regulated Learning: Constructive Activities Enhance Deep Learning with Online Materials
}

\author{
Lisa A. Ferrara \\ University of Utah
}

\author{
Kirsten R. Butcher \\ University of Utah
}

\begin{abstract}
The interactivity framework (Chi, 2009) proposes learners experience greater learning benefits as they become more generative with learning material. This research investigated the active-constructive-interactive framework in the context of graphical organizers before a self-regulated, online learning task. Graphical displays were node-link diagrams, where nodes identified important domain concepts and link labels described the relationship between concepts. Conditions examined: passive (in which nodes and links were provided); active (in which nodes were provided and learners revealed link labels on demand); and constructive (in which nodes were provided and learners generated link labels). Findings were consistent with the interactivity framework for constructive learners, who integrated more concepts into posttest concept maps and included more deep (relational/causal) statements in their posttest essays. However, there were no significant differences for passive and active conditions. Results demonstrate that constructive activities may be necessary to support deeper learning outcomes in activities used before self-directed learning tasks.
\end{abstract}

\title{
Constant entropy rate and related hypotheses versus real language.
}

\author{
Ramon Ferrer-i-Cancho \\ Universitat Politecnica de Catalunya
}

\section*{Lukasz Debowski}

Polish Academy of Sciences

\begin{abstract}
Constant entropy rate (CER) and uniform information density (UID) are two hypotheses that have been put forward to explain a wide rage of linguistic phenomena. However, the concrete definition of these hypotheses is unclear for statistical research and a direct and in-depth evaluation of these hypotheses from their definition is missing to our knowledge. Here we consider four operational definitions of UID: full UID (UID holding for any combination of elements making the utterances), strong UID (UID holding for any utterance that has non-zero probability) and initial UID (strong UID holding for utterances beginning with a particular element). Here we examine the logical dependencies between these hypotheses. The comparison of the assumptions and predictions of these hypothesis with Hilberg's law and other statistical properties of real human language indicates that CER and related hypotheses are qualitative different from actual language and suggests that these hypotheses are incomplete and must be revised.
\end{abstract}

\title{
The origins of analogy
}

\author{
Alissa Ferry \\ Scuola Internazionale Superiore di Studi Avanzati \\ Susan Hespos \\ Northwestern University \\ \section*{Yin-Juei Chang} \\ Northwestern University \\ Emily Hollenbeck \\ Northwestern University \\ Dedre Gentner \\ Northwestern University
}

\begin{abstract}
We investigated the origins of analogical ability in 7-month-old infants, using the simplest and most basic relation - that of sameness and difference between two things. Experiment 1 showed infants were unable to detect and generalize these relations from a single exemplar (as suggested in Tyrrell et al., 1991). Experiment 2 used a habituation-dishabituation paradigm and found that infants could generalize the same-different relation to novel objects with six to nine training trials. Experiment 3 demonstrated that labels influenced performance: labeling the relation enhanced performance, but labeling the individual objects hindered performance. In addition, we varied infants' prior experience with the objects and found signatures of relational learning have continuity across development. In summary, abstraction of relations can be facilitated by comparison across exemplars, disrupted by the saliency of individual objects, and manipulated by labeling. These findings are discussed in light of recent debates about phylogenetic continuity in relational abilities.
\end{abstract}

\title{
Using Xbox Kinect to explore spatial-numerical association of arm movements in parity judgments.
}

\author{
Yariv Festman \\ University of Potsdam
}

Oliver Lindemann
University of Potsdam

\author{
Martin Fischer
}

University of Potsdam

\begin{abstract}
Spatial-numerical associations were initially studied using chronometric methods to reveal the orientation of the mental number line (SNARC effect; Dehaene, Bossini, and Giraux, 1993). Recent evidence has shown that unconstrained spatial movements can be valuable in the study of embodied number representations (Fischer and Campens, 2009). We used the Xbox Kinect to record arm movements in a parity judgment task. We replicated SNARC in the horizontal dimension and found a similar trend in the vertical dimension. Movement amplitudes were also affected by number magnitude. Together, these results generalize evidence for the mental number line to everyday behaviors and suggest that natural user interfaces (NUI) can be used to study embodied cognition.
\end{abstract}

\title{
From Complex to Collaborative Problem Solving
}

\author{
Andreas Fischer \\ Heidelberg University, Heidelberg, Germany \\ Daniel Holt \\ Heidelberg University, Heidelberg, Germany \\ Julia Hilse \\ Heidelberg University, Heidelberg, Germany \\ Joachim Funke \\ Heidelberg University, Heidelberg, Germany
}

\begin{abstract}
Research on human problem solving may be about to face a conceptual change from individual to collaborative problem solving. Many challenging problems in the real world are solved by groups or teams. This is increasingly recognized by the problem solving research community, which traditionally has emphasized cognitive processes in individual problem solving. In this paper we argue how approaches for investigating complex problem solving can be conceptually extended towards collaborative problems solving. We will present several current examples of how to measure collaborative processes in a standardized way. One example is the InBox HD, a computer-based in-basket simulation with collaborative elements. The second example is the scenario Product Planning, implemented in the ColPS HD framework, which involves chat-based human-to-agent communication. We will elaborate on how these tools can be used to emulate realistic collaborative processes in the standardized setting of a psychological laboratory and indicate directions for future developments.
\end{abstract}

\title{
Toddler's understanding of false beliefs about object identity
}

\author{
Ella Fizke
}

\author{
Institute of Psychology \& Courant Research Centre Evolution of Social Behaviour", University of Göttingen, Germany \\ Stephen Butterfill \\ Department of Philosophy, University of Warwick, UK \\ Hannes Rakoczy \\ Institute of Psychology \& Courant Research Centre Evolution of Social Behaviour", University of Göttingen, Germany
}

\begin{abstract}
Children solve explicit false belief (FB) tasks around age 4 but implicit tasks (e.g. helping) in infancy (e.g. Buttelmann et al., 2009). Nativist accounts claim early conceptual competence, masked by performance factors (e.g. Leslie, 2005); sceptical accounts deny competence before age 4 (Perner \& Ruffmann, 2005). A recent two-system-theory (Apperly \& Butterfill, 2009) provides an alternative explanation: an early mindreading-system (1) tracks simple forms of mental states and a later flexible capacity (2) allows cognitively demanding inferences based on a fully-developed concept of belief. Because system 1 operates on relational rather than propositional attitudes it has clear flexibility-limits: it can represent FB 's about object location but not FB's about identity.

We contrasted 2.5 -year-olds' helping behavior in a 2 (identity/location) X 2 (false/true belief) design. Results suggest limited performance in the identity compared to the location task. Implications of this finding are discussed and follow-up studies will be presented.
\end{abstract}

\title{
Individual differences in 3D pointing performance between passengers and drivers
}

\author{
Julia Frankenstein \\ Freiburg University, Freiburg, Germany \\ Christoph Hölscher \\ Freiburg University, Freiburg, Germany
}

\begin{abstract}
While most studies in the spatial cognition literature focus on pointing in a two-dimensional context, (i.e., directional estimates in a horizontal plane), the mechanisms of pointing in 3d-space (i.e., direction estimates including up and down) are much less clearly understood. Based on a paradigm of Vidal \& Berthoz, a virtual tube system providing a highly controlled environment, we designed a study comparing 3d pointing performance (i.e., pointing backward to the starting point or pointing forward to the end point of the experienced route) following active (drivers) vs. passive (passengers) exploration. Furthermore, we assessed the individual memory strategies reflected in self-reports and related these to pointing performance and reaction times.
\end{abstract}

\title{
Automatic classification of patients with mental disorders according to voice dynamics
}

\author{
Riccardo Fusaroli \\ Aarhus University, Aarhus, Denmark \\ Kristian Tylén \\ Aarhus University, Aarhus, Denmark \\ Arndis Simonsen \\ Aarhus University, Aarhus, Denmark \\ Ethan Weed \\ Aarhus University, Aarhus, Denmark
}

\begin{abstract}
Anomalous aspects of speech and voice, including pitch, fluency, and voice quality, are reported to characterize many mental disorders. However, it has proven difficult to quantify and explain this oddness of speech by employing traditional statistics methods.

In this study we employ Recurrence Quantification Analysis (RQA) to investigate the temporal dynamics of voice in three mental disorders. We elicited monological descriptions of short videos in patients with schizophrenia, depression and Asperger's, as well as in related matched controls. We applied RQA to fundamental frequency, speech pause sequences and speech rate. The Rqa indexes (trend and entropy in particular) enable us to quantify and automatically discriminate between populations with \(>85 \%\) of accuracy, highlighting distinctive voice dynamics in each diagnoses.
\end{abstract}

\title{
Conversation, Coupling and Complexity: Matching Scaling Laws Predict Performance in a Joint Decision Task
}

\author{
Kristian Tylén \\ Aarhus University, Aarhus, Denmark \\ Drew Abney \\ UC Merced, Merced (CA), USA \\ Bahador Bahrami \\ UCL, London, UK \\ Christopher Kello \\ University of California, Merced (CA), USA \\ Riccardo Fusaroli \\ Aarhus University, Aarhus, Denmark
}

\begin{abstract}
We investigate the linguistic co-construction of interpersonal synergies. By applying a measure of coupling between complex systems to an experimentally elicited corpus of joint decision dialogues, we show that interlocutors' linguistic behavior displays increasing signature of multi-scale coupling, known as complexity matching, over the course of interaction. Furthermore, we show that stronger coupling corresponds with more effective interaction, as measured by collective task performance.
\end{abstract}

\title{
Completing the Puzzle: Online Processing during Novel Noun Generalization
}

\author{
Megan Galligan \\ University of Iowa \\ \section*{Larissa Samuelson} \\ University of Iowa
}

\begin{abstract}
When building a lexicon, young children must first learn individual word-object mappings and subsequently extend these mappings to new category instances. Recent methodological advances demonstrate that processing efficiency during the initial mapping process increases between 15 and 24 months-of-age. The current study investigates real-time processing during the second word learning step, generalization. Using a head-mounted eye-tracker, 18-month old children completed a novel noun generalization task with novel solid objects. Many participants generalized names for solid objects based on similarity in shape rather than similarity in material. The addition of eye-tracking data in the current study reveals children's comparison of test and exemplar objects both before and after novel name presentation. Integrating eye-tracking data with measures of vocabulary knowledge can elucidate how vocabulary organization speeds the generalization decision. Future work will manipulate syntactic context to examine how children's knowledge of count and mass nouns further influences the generalization decision process.
\end{abstract}

\title{
Bilingualism change axonal structural network organization
}

\author{
Lorna Garcia Penton \\ Basque Center on Cognition, Brain and Language (BCBL), Donostia, Guipuzcoa, Spain
}

\author{
Alejandro Perez Fernandez \\ Basque Center on Cognition, Brain and Language (BCBL), Donostia, Guipuzcoa, Spain \\ Yasser Iturria-Medina \\ Cuban Neuroscience Center (CNC), La Habana, Cuba. \\ Manuel Carreiras \\ Basque Center on Cognition, Brain and Language (BCBL), Donostia, Guipuzcoa, Spain
}

\begin{abstract}
How the brain deals with more than one language and whether we need different or extra brain language subnetworks to support more than one language is unanswered question. Here, we investigate structural brain network differences between early bilinguals and monolinguals. Using diffusion-weighted MRI (DW-MRI) tractography techniques and a networkbased statistic (NBS) procedure (Zalesky et al., 2010), we found two structural sub-networks more connected by white matter (WM) tracts in bilinguals than in monolingual; confirming WM brain plasticity in bilinguals (Luk et al., 2011; Mohades et al., 2012; Schlegel et al., 2012). One of these sub-networks comprises left frontal and parietal/temporal regions, while the other comprises left occipital and parietal/temporal regions and also the right superior frontal gyrus. Most of these regions have been related to language processing and monitoring (Abutalebi and Green, 2007); suggesting that bilinguals developed specialized language sub-networks to deal with the two languages. Additionally, a complex network analysis showed that these sub-networks are more graph-efficient in bilinguals than monolinguals and these increase seems to be at the expense of a decrease in whole network graph-efficiency.
\end{abstract}

\title{
Infants' ability to discriminate between statements and questions
}

\author{
Susan Geffen \\ University of Southern California \\ Toben Mintz \\ University of Southern California
}

\begin{abstract}
Children must distinguish between statements and questions in order to accurately acquire language, but it is unclear when or how they do it. Experiment 1 examined whether prosodic characteristics of infant-directed questions and statements could differentiate them. Statements and yes/no questions differed on several dimensions, but statements and wh-questions did not. Experiment 2 tested whether 11-13-month-olds could nevertheless distinguish sentence types using lexical information. Half the infants were familiarized to statements, the remainder to questions. All infants were tested on new sentences of both types. Sentences were resynthesized to have monotone pitch and matched utterance-final vowel length, neutralizing any prosodic differences. Overall, there was a significant novelty preference and no interaction of trial type with familiarization type. Thus, while prosody is insufficient for distinguishing wh-questions from statements, by 11-months infants can use word order to distinguish statements and questions. This ability could provide an important foundation for acquiring syntactic knowledge.
\end{abstract}

\title{
Neural Correlates of Perceiving Dyadic Social Interactions
}

\author{
Alexandra Georgescu \\ University Hospital of Cologne, Cologne, NRW, Germany \\ Bojana Kuzmanovic \\ University Hospital of Cologne, Cologne, NRW, Germany \\ Natacha Santos \\ University Hospital of Cologne, Cologne, NRW, Germany \\ Ralf Tepest \\ University Hospital of Cologne, Cologne, NRW, Germany \\ Gary Bente \\ Department of Psychology, University of Cologne, Cologne, NRW, Germany \\ Marc Tittgemeyer \\ Max-Planck Institute for Neurological Research, Cologne, NRW, Germany \\ Kai Vogeley \\ University Hospital of Cologne, Cologne, NRW, Germany
}

\begin{abstract}
The aim of the present study was to determine the differential contributions of the action observation network (AON) and the social neural network (SNN) to the experience of naturalness in observed dyadic social interactions. To this end, we used short animation sequences displaying social interactions between two virtual characters and systematically manipulated kinematic features of the social dynamics. A group of 21 male participants rated the "naturalness" of the observed scenes on a four-point scale while undergoing fMRI. Using the ratings of each participant as a parametric modulation of their general neural response to the stimuli, we found that an increase in naturalness experience was associated with higher activations in the AON. The SNN was preferentially recruited with a decrease in naturalness experience. This indicates that understanding familiar interactions involves an automatic kinematic processing of intentionality, while interactions perceived as artificial require higher-level inferential processing.
\end{abstract}

\title{
Imagined Effort Affects Object Localization and Sense of Agency
}

\author{
Devin Gill \\ Illinois State University \\ \section*{J. Scott Jordan} \\ Illinois State University
}

\begin{abstract}
Individuals who continually track an object that suddenly vanishes indicate perceived vanishing points displaced beyond the actual vanishing point (i.e., forward displacement: FD) (Hubbard, 1995). Jordan, Coey, and Tsippaaoutis (2009) demonstrated that FD increases with implied friction (i.e., low to high friction) if one controls stimulus movements. Metcalfe and Greene (2007) showed that manipulations of stimulus control affected judgments of agency. The present experiment examined the extent to which implied friction and conceptual factors (Reed \& Vinson, 1996) affect feelings of agency during stimulus control. Participants controlled the movements of a trapezoidal stimulus labeled as either a "bullet train" or a "house" in two levels of implied friction. Results revealed a marginally-significant increase in FD with implied friction. Agency also varied significantly between implied friction conditions, but only when participants conceptualized the stimulus as a bullet train and implied friction decreased across blocks (i.e., implied effort became optimal).
\end{abstract}

\title{
Explorations in Human and Machine Learning of Decision Trees
}

\author{
Kevin Gluck \\ U.S. Air Force Research Laboratory \\ Jack Harris \\ U.S. Air Force Research Laboratory \\ Thomas Mielke \\ U.S. Air Force Research Laboratory \\ \section*{Christopher Myers} \\ U.S. Air Force Research Laboratory \\ Vladislav Veksler \\ U.S. Air Force Research Laboratory
}

\begin{abstract}
We explore the boundaries of learnability, ecological rationality, and decision robustness in uncertain, non-stationary, finite-sample environments. Our approach combines machine-learning-based heuristic search techniques with the Integrated Learning Model (ILM) computational cognitive process theory of human and animal learning. The scientific contributions of this research are in understanding whether and how decision heuristics are acquired in binary classification contexts, with an emphasis on fast and frugal decision trees. The real world relevance of this research is in improved decision training and aiding.
\end{abstract}

\title{
The Role of Conceptual and Perceptual Information on Inductive Reasoning in Early Childhood
}

\author{
Karrie Godwin \\ Carnegie Mellon University
}

\author{
Bryan Matlen \\ Carnegie Mellon University
}

\author{
Anna Fisher \\ Carnegie Mellon University
}

\begin{abstract}
There is an ongoing debate about the relative contribution of conceptual and perceptual information to inductive generalization in early childhood. In the classic study bearing on this debate, pictures representing familiar animals were arranged such that category membership was supposed to be in conflict with perceptual similarity. However, later studies revealed that most of the stimuli in this study failed to impose this conflict. The present study revisited this issue. Extensive calibration was conducted to ensure that the two sources of information were in conflict. Despite near-ceiling accuracy in identifying the category membership of objects used in the study (e.g., bird-bird-bat, dog-dog-cow, cat-cat-raccoon, etc.), 4-year-old children relied on their knowledge of category membership only \(55 \%\) of the time when there was strong conflict between category membership and perceptual similarity. These findings will be discussed in relation to alternative accounts of knowledge acquisition and generalization early in development.
\end{abstract}

\title{
Perception and action effects on causal judgment
}

\author{
Kelly Goedert
}

Seton Hall University

\author{
Mengqi Guo
}

Seton Hall University

\begin{abstract}
Recent research suggests that perception and action affect performance on cognitive tasks (e.g., Beilock \& GoldinMeadow, 2010; Landy \& Goldstone, 2007). Here we investigated perceptual and motor influences on causal judgment from contingency using a causal discounting paradigm (e.g., Goedert \& Spellman, 2005). Participants learned about two potential causes of a common outcome on a trial by trial basis. We varied the left/right location of a target cause and the left/right location of the "yes" response button for predicting the cause would produce the outcome. When there was a mismatch between the target location and the "yes" response, participants discounted. However, they did not discount when the two locations matched. Thus, we observed more accurate causal judgment with spatial overlap in the perception and action information. These results are generally consistent with an embodied cognition framework; however, their exact mechanism remains to be explored.
\end{abstract}

\title{
Effects of Brief Analogical Training after Two-Week Delay
}

\author{
Micah Goldwater \\ Northwestern University \\ Dedre Gentner \\ Northwestern University
}

\begin{abstract}
There is considerable evidence that analogical comparison can foster STEM learning. For example, Gentner, Levine, Dhillon, \& Poltermann (2009) used comparison to teach 6-8-year-old children an important engineering principle: namely, that a diagonal brace confers stability in construction. Children compared two buildings, one with a diagonal brace, and one with a horizontal (nonbracing) piece instead. After the comparison, children were shown an unstable building, and were given a piece to stabilize it. Children who received the comparison training were more likely to attempt to stabilize the building with a diagonal placement of the piece. We extended this research to (a) test for retention after two weeks and (b) examine effects of relational labeling (which has been theorized to support long-term retention) The results indicate that the training utilizing comparison and relational labeling elicited more diagonal placements both after a brief delay, and after a delay of two weeks.
\end{abstract}

\title{
Interactions between bilingualism and non-linguistic spatial processing
}

\author{
Monica Gonzalez-Marquez \\ Cornell University
}

Raymond Becker
Bielefeld University
\(\underset{\text { Cornell University }}{\text { James Cutting }}\)

\begin{abstract}
The relationship between language and space has been intensely investigated. The underlying question has been whether one affects the other, usually within the scope of spatial language. Largely ignored is the possible role of bilingualism. Given that the average person speaks more than one language, and the mounting evidence showing that bilingualism interacts with non-linguistic processes (e.g., Bialystok \& Senman, 2004), we investigated what effect bilingualism would have on nonlinguistic spatial processing.

We tested 120 participants with a range of linguistic abilities using four classic spatial tasks, e.g. mental rotation. We found patterns of systematic interaction between bilingualism and spatial processing. These findings raise questions beyond the relationship between spatial language and spatial cognition, suggesting that language as a cognitive process may share a common neural substrate with space.
\end{abstract}

\title{
Conceptual Priming with Pictures and Environmental Sounds
}

\author{
Paula Goolkasian \\ UNC Charlotte
}

\author{
Yongju Kim
}

Korea Military Academy

\section*{Anne Marie Porter \\ UNC Charlotte}

\section*{Kathryn Weatherford}

UNC Charlotte

\begin{abstract}
A series of experiments were conducted to examine conceptual priming within and across modalities with pictures and environmental sounds. In Experiment 1, we developed a new multimodal stimulus set consisting of two picture and sound exemplars that represented 80 object items. In Experiments 2 and 3, we investigated whether categorization of the stimulus items would be facilitated by picture and environmental sound primes that were derived from different exemplars of the target items. The results demonstrated that the categorization of environmental sounds and pictures were facilitated in a similar way by conceptually related exemplars presented in advance, but only when a long inter-stimulus interval ( 1000 ms ) was used. Additionally, conceptual cross-modal priming effects by picture and sound primes were asymmetric with systematic switch costs across modalities and with differences in the time-course of activation.
\end{abstract}

\title{
Taking Someone Else's Perspective: When Body "Position" is More Important than Body "Presence"
}

\author{
Michelle D. Greenwood
}

University of California, Merced

\author{
Justin L. Matthews \\ University of California, Merced \\ Michael J. Spivey \\ University of California, Merced \\ Teenie Matlock \\ University of California, Merced
}

\begin{abstract}
Taking another person's perspective when describing spatial scenes is more common when a person is present in the scene. (Tversky \& Hard, 2009). What about seeing another person elicits an allocentric perspective?

Participants were shown one of a variety of pictures displaying a book and cup placed, side-by-side, on a table. Some photos also pictured a man sitting behind the table, either facing the camera or facing to the left or right. Viewers were more likely to take the man's perspective while describing object locations when the man was facing the camera than when the man was facing to either side.

Many factors influence which perspective people take. These results suggest that the mere presence of a person in a scene does not guarantee a viewer will take someone else's perspective, but rather the way a person is positioned in a scene might also be of critical importance.
\end{abstract}

\title{
Test Expectancy Effects on Metacomphrenesion, Self-regulation, and Learning
}

\author{
Thomas Griffin \\ University of Illinois at Chicago \\ Jennifer Wiley \\ University of Illinois at Chicago
}

Keith Thiede
Boise State University

\begin{abstract}
Metacomprehension monitoring accuracy is defined as the ability to accurately predict how well one will do on a later test of learned material. Metacomprehension monitoring is presumed to be a critical skill for the effective self-regulation of study behaviors that impact learning. In two experiments, ecologically valid science texts and inference tests were employed to examine whether a test expectancy intervention could improve students' metacognitive judgments, self-regulated study, and learning outcomes. Experiment 1 was a lab experiment in which test expectancies were instilled only after reading was complete, thus preventing any encoding effects. Results suggest that test expectancies impact metacomprehension monitoring accuracy via selection of more valid cues at the time of judgment rather than only via encoding effects that impact cue accessibility. Experiment 2 was a classroom study showing that the effect of test expectancies on monitoring accuracy translates into more effective self-regulated study and improved learning.
\end{abstract}

\title{
Prisoner's Dilemma Games between Friends and Foes: Looking for Simpson's Paradox Effects in Cooperation
}

\author{
Maurice Grinberg \\ New Bulgarian University \\ Stevan Tomic \\ New Bulgarian University
}

\section*{Evgenia Hristova}

New Bulgarian University

\begin{abstract}
The paper explores the cooperation rate in one-shot Prisoner's Dilemma games in three cases - when the game is played among friends, among foes, and among a mixture of friends and foes. The paper checks empirically the prediction that Simpson's paradox like effects are to be expected in this situation (Chater, Vlaev, \& Grinberg, 2008). The existence of a bias for cooperation when playing with friends and for defection when playing with foes, is expected to lead to a reinforcement of cooperation based on higher average payoffs in the mixed condition, i.e. when playing with friends and foes together. At the same time, the average payoff for cooperation remains lower for games with friends and foes taken separately (Simpson's paradox). There results of the experiment support the existence of such effects and suggest that further exploration is worthwhile.
\end{abstract}

\title{
Big Number Politics: The Effects of Training on Large Number Estimation in Fiscal Deficit-reducing Proposals.
}

\author{
Brian M. Guay \\ University of Richmond \\ David Landy \\ University of Richmond
}

\begin{abstract}
How do we make sense of the 43 million or 1.3 billion dollar budget cuts that are being made in Congress? Large quantities such as these are common in our political discourse, yet recent studies demonstrate substantial and systematic biases in evaluating them (Landy, Silbert \& Goldin 2012). We explore how the integration of numerical and political information affects voters' evaluation of political scenarios, and more specifically the effect of number training on this evaluation. The current study investigates the effects of a number training intervention on the numerical estimation task and evaluation of deficit-reducing proposals using a within-subjects design and a typical voting population. Participants in the training condition completed a number estimation task (Siegler \& Opfer 2003) and were shown the accurate location of 1 million on a number line from 10 thousand to 1 billion. Line estimation and situation evaluation were assessed before and after the intervention.
\end{abstract}

\title{
Consistency in Degrees of Morality
}

\author{
Eoin Gubbins \\ Trinity College Dublin, Dublin, Ireland \\ Ruth Byrne \\ Trinity College Dublin, Dublin, Ireland
}

Phil Johnson-Laird
New York University

\begin{abstract}
We examined people's ability to judge the degree of morality of good and bad actions, and their consistency in doing so. Participants judged the degree of morality of actions on a scale from +100 (moral) through 0 (neither) to -100 (immoral). They judged the degree of morality for individual actions e.g., 'a man intervened to stop a fight', 'a man gave blood', and their conjunction, 'a man intervened to stop a fight and he gave blood'. Most judgments were consistent, i.e. the conjunction was judged to be more moral or immoral than its components. However, a reliable number of judgments were inconsistent, i.e. the conjunction was judged to be less moral or immoral than one or both of its components. Consistency improved when participants read the conjunction after the conjuncts, compared to when they read it before. We discuss implications for understanding the mental representation of degrees of morality.
\end{abstract}

\title{
Creativity: Investigating Construct Validity across Cultures
}

\author{
C. Dominik Güss \\ University of North Florida, United States \\ Ma. Teresa Tuason \\ University of North Florida, United States \\ Yetuli de Baessa \\ Universidad Francisco Marroquín, Guatemala \\ Nila Shah \\ Shri Jasani Arts and Commerce College, Rajkot, India \\ \section*{Angela Thomas} \\ University of Pretoria, Republic of South Africa
}

\begin{abstract}
Creativity is the drive for advancement in many aspects of society such as arts, economy, or science. It is unclear if creativity consists of several different skills and abilities or if it is one core construct. It is also unclear to what extent creativity is influenced by culture and to what extent creativity constructs can be generalized across cultures. To investigate these questions, we administered three different creativity tests assessing fluency, originality, flexibility, and creative achievement to over 900 students in five countries: Germany, Guatemala, India, South Africa, and the United States. Results showed weak correlations between the different aspects of creativity speaking for heterogeneity of different creativity constructs, across all cultures. Whereas participants from the five countries did not differ in their creative achievements, they differed in the cognitive creativity measures. Results are interpreted referring to the eco-cultural context and existing cognitive frameworks of creativity.
\end{abstract}

\title{
Exploration and Discovery in Children with Autism
}

\section*{Hyowon Gweon \\ MIT}

\author{
Hannah Pelton \\ MIT
}

Rebecca Saxe
MIT

\section*{Laura Schulz}

MIT

\begin{abstract}
ASD is diagnosed by perseverative behaviors and social deficits. While ASD children's selective interest in the physical world may lead to more extensive exploration (and learning) of physical objects, they may learn less due to their perseverative behavior. How is exploration affected in children with ASD? We quantified exploration and discovery as children with ASD ( \(\mathrm{N}=35, \mathrm{M}=8.5 \mathrm{yrs}\) ) and their controls ( \(\mathrm{N}=35\), age, IQ matched) explored a novel toy with hidden functions. ASD children showed more perseveration than controls on every measure, and the diversity of their actions was negatively correlated with severity of autism. Furthermore, ASD children discovered less hidden functions than controls. While the control group showed a marked decrease in perseveration with age, ASD group showed a heightened level of perseveration independent of age. These data suggest that children with ASD show marked difference in their exploration of the physical environment, and this difference has real consequences for learning and discovery.
\end{abstract}

\title{
Inferring mass in complex physical scenes via probabilistic simulation
}

\author{
Jessica Hamrick \\ University of California, Berkeley \\ Peter Battaglia \\ Massachusetts Institute of Technology \\ Thomas Griffiths \\ University of California, Berkeley \\ Joshua Tenenbaum \\ Massachusetts Institute of Technology
}

\begin{abstract}
How do people learn underlying properties, such as mass and friction, from objects' interactions in complex scenes? Such inferences are difficult: the parameters cannot be directly observed and have nonlinear effects on the physical dynamics. Yet, people learn them. Participants predicted the stability of blocks stacked in complex tower configurations. After observing the true outcome, they answered, "which blocks are heavier?". Their responses indicate rapid learning of the blocks' relative masses. We view such learning as probabilistic inference in a generative model of Newtonian rigid-body dynamics, and express this hypothesis in a model observer that infers parameters using a procedure of approximate physical simulation. While participants' judgments qualitatively matched the model's, they also deviated in key ways that may be explained by resource limitations. This work advances our understanding of how people infer unobserved physical properties, and offers a framework for modeling such behavior in complex, real-world scenes.
\end{abstract}

\title{
Predictive processing of musical structure: effects of genre-specific expertise
}

\author{
Niels Chr. Hansen \\ Center of Functionally Integrative Neuroscience, Aarhus University Hospital, Denmark
}

Niels Chr. Hansen
Royal Academy of Music Aarhus/Aalborg, Denmark
Niels Chr. Hansen
Department of Aesthetics and Communication, Aarhus University, Denmark
Peter Vuust
Center for Functionally Integrative Neuroscience, Aarhus University Hospital, Denmark
Peter Vuust
Royal Academy of Music Aarhus/Aalborg, Denmark
Marcus Pearce
Centre for Digital Music and Research Centre in Psychology, Queen Mary University of London, UK

\begin{abstract}
Expectations are generated with different degrees of predictive uncertainty prior to onset of musical events. This study explored influences of genre specific expertise in non-musicians, classical, and jazz musicians listening to unfamiliar Charlie Parker solos.

Two probabilistic computational models of expectation were trained: one on folksongs (General), the other on jazz (Bebop). Twenty-four melodies were selected whose final notes differed in Shannon entropy estimated by the two models. Listeners' uncertainty was assessed explicitly and inferred from expectedness ratings of different continuation tones.

The analysis showed that jazz musicians followed 'Bebop' and non-musicians followed 'General'. Classical musicians showed some decoding of the jazz style, utilising a somewhat underdeveloped version of 'Bebop'. Moreover, experts experienced more salient prediction errors in low-entropy contexts, and musical skills predicted the extent of cognitive model optimisation.

Our results suggest that expertise entails both possessing an accurate predictive model and selecting an optimal model for the given context.
\end{abstract}

\title{
On the relationship between perceived foreigness, accentedness and speech comprehension under clear and adverse listening conditions
}

\author{
Adriana Hanulikova \\ University of Freiburg
}

\author{
Frank Eisner \\ MPI for Psycholinguistics
}

\begin{abstract}
Social interaction involves the simultaneous uptake of a range of linguistic and nonlinguistic information. In a seminal study, Rubin (1992) has shown that a lecture spoken by a native speaker of Standard American English presented along with a photograph of an ethnically Asian instructor affected comprehension score and accentedness ratings more negatively as compared to the same lecture and speaker presented with a Caucasian instructor. Here we asked whether such effects could be observed in a multicultural environment, with a lot of interactions with different ethnicities and non-native speakers. Furthermore, we investigated how the effect could be modulated by the quality of the speech input (clear compared to noisy speech). The results showed that ethnicity affects accentedness ratings only under adverse listening conditions and that comprehension scores do not depend on ethnicity of the speaker. Thus, the effect of nonlinguistic information on linguistic processing is constrained in multicultural settings.
\end{abstract}

\title{
Understanding the Role of Context in Memory for Maximally Counterintuitive Concepts
}

\author{
Mary Harmon-Vukic \\ Providence College \\ M. Afzal Upal \\ Defence Research \& Development Canada \\ Caitlin Trainor \\ Providence College
}

\begin{abstract}
The effect of minimally counterintuitive information on memory is well established. However, less work has addressed the processing of maximally counterintuitive stories (i.e., stories containing at least three domain violations). The current study examined memory for maximally counterintuitive stories. The first two experiments investigated whether explicit instruction to make sense of "strange information" influenced memory for maximally counterintuitive stories. Although no such effect was observed, post hoc analyses indicated that the extent to which concepts in maximally counterintuitive stories contained domain violations from similar or different categories influenced memory performance; stories with similar domain violations enjoyed a memory advantage. The third study addressed the believability of concepts with similar domain violations with a rating task. Participants were more likely to agree with two counterintuitive concepts with similar domain violations compared to a single counterintuitive concept. The results are discussed within Upal's (2005; 2009) context-based view of memory for counterintuitive ideas.
\end{abstract}

\title{
The effect of goal reification on error rates while solving geometric problems
}

\author{
Robert Hausmann \\ Carnegie Learning, Inc., Pittsburgh, PA, United States
}

\author{
Annalies Vuong \\ Carnegie Learning, Inc., Pittsburgh, PA, United States
}

\begin{abstract}
An effective learning environment is designed to expose student misconceptions because, once explicit, they are available for remediation. How does one design such an environment, and what are the consequences for learning? In a previous study, we demonstrated that a significant revision to the Cognitive Tutor Geometry intelligent tutor had a generally positive effect on the speed of skill mastery (Hausmann \& Vuong, 2012). However, the revised version demonstrated a higher error rate for easy skills. We hypothesized that the revised interface reified certain mental steps that students were previously allowed to complete implicitly. Specifically, students are now required to write an expression for the length of the side of a special right triangle before calculating the length. While the error rate for the calculation remained low ( \(1.71 \%\) ), writing the expression proved to be particularly difficult (13.93\%). We contrast the evidence supporting this hypothesis with evidence for other potential explanations.
\end{abstract}

\title{
Real-Time Strategy: Multi-Level Dynamics in an Uncertain Environment
}

\author{
Robert Hawkins \\ Indiana University
}

\author{
Robert Goldstone
}

Indiana University

\begin{abstract}
As an agent gathers information about its environment and monitors the decisions of other agents, its behavior may fluctuate adaptively over short time scales while still maintaining a long-term strategy. We designed a real-time virtual environment to experimentally investigate the relationship between the micro-level dynamics of dyadic behavior within single games and the macro-level dynamics of outcomes across iterated games.

In one experiment, participants played a real-time game of "chicken," simultaneously guiding avatars toward high-payoff or low-payoff targets. If both participants reached a demarcated vicinity of a target at the same time, that target was destroyed. We recorded their trajectories, and induced uncertainty by adding noise to their movement speeds. At the macro-level, we found evidence of self-organized turn-taking across repeated games. At the micro-level, we found that even within a turn-taking equilibrium, both players competitively pursued the high payoff for a period of time before one of them diverted.
\end{abstract}

\section*{Children's understanding of lying related to intention.}

\section*{Hajimu Hayashi}

Okayama University

\begin{abstract}
Children's conception of lies has been important issues for children's cognitive development. However, little is known about whether children's understanding of lies is different form adults' one. Four kinds of stories were presented for children aged 6- to 7-year olds and undergraduate students. First, a protagonist had a deceptive intention and produced a false statement. Second, he had a deceptive intention but produced a true statement by a false belief. Third, he had a truthful intention and produced a true statement. Fourth, he had a truthful intention but produced a false statement by a false belief. The results showed that undergraduate students judged that these protagonist's statements were lying or not by considering his intentions. By contrast, children judged regardless of his intentions. These results suggest that children's conception of lying is different from adults' one, and that their conception becomes sophisticated after middle childhood.
\end{abstract}

\title{
Patient Knowledge and Satisfaction in Decisions About Lung Cancer Surgery
}

\author{
Josh A. Hemmerich \\ The University of Chicago \\ Mark K. Ferguson \\ The University of Chicago \\ Daniel P. Sulmasy \\ The University of Chicago \\ Arthur S. Elstein \\ University of Illinois at Chicago \\ Cindy Warnes \\ The University of Chicago \\ \section*{Masha Kochrginsky} \\ The University of Chicago \\ \section*{Rita Gorawara-Bhat} \\ The University of Chicago \\ Kellie Van Voorhis \\ The University of Chicago \\ William Dale \\ The University of Chicago
}

\begin{abstract}
Medical care is increasingly implementing shared decision making that requires participation of informed patients in an effort to maximize treatment-related decision satisfaction (DS). Patient comprehension of relevant information is important in decisions involving high risk and uncertainty, like surgery for lung cancer.

Lung cancer patients ( \(\mathrm{N}=43\) ) completed pre- and post-consult questionnaires, and their consults with the surgeon were audio recorded.

Post consult knowledge was low ( \(53 \%\) ) while DS was moderately low ( \(\mathrm{M}=44.56\), \(\mathrm{SD}=13.71\) ). Higher complication risk ( \(\mathrm{rs}=\) .34, \(\mathrm{P}<.05\) ), external locus of control ( \(\mathrm{r}=.31, \mathrm{P}<.05\) ), belief in a controlling deity ( \(\mathrm{r}=.37, \mathrm{P}<.05\) ), and desire for control ( r \(=-.38, \mathrm{P}<.016\) ) predicted lower DS. Consult recordings showed that patients possess counterfactual beliefs, such as airborne spread of cancer in surgery, and benefit from removing most of the tumor. Experiments are needed to understand how patient comprehension can be improved.
\end{abstract}

\title{
Baby physics: expectations about liquid and sand in 5-month-old infants
}

\author{
Susan Hespos \\ Northwestern University
}

\author{
Alissa Ferry \\ Scuola Internazionale Superiore di Studi Avanzati \\ Emily Hollenbeck \\ Northwestern University \\ \section*{Lance Rips} \\ Northwestern University
}

\begin{abstract}
Infants have expectations about physical properties of solid objects. However, evidence on infants' understanding of nonsolid substances (e.g., water or sand) is sparse and equivocal. We conducted four habituation/dishabituation experiments demonstrating that 5-month-olds have distinct expectations for how objects and substances behave. Experiment 1 found that infants use motion cues from the surface of a contained liquid or solid to predict whether it would pass through or rest on a grid when the container was upended. Experiment 2 extended these findings to show that motion cues led to different expectations about whether a new object will pass through or remain on the top surface of a liquid or solid. Experiments 3 and 4 replaced the liquid with sand. We found that infants expected sand, like liquid, to go through a grid, but did not expect another object to pass through it. These findings begin to characterize infants' understanding of substances.
\end{abstract}

\title{
The timing of turns in mother-infant interactions: A Longitudinal Study
}

\author{
Elma Hilbrink \\ Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands \\ Merideth Gattis \\ Cardiff University \\ Elena Sakkalou \\ Neurosciences Unit, UCL Institute of Child Health \\ Kate Ellis-Davies \\ Cambridge University \\ Stephen Levinson \\ Max Planck Institute for Psycholinguistics
}

\begin{abstract}
Turn-transition in adult conversation is remarkably precise, with a median close to zero milliseconds. This means one needs to predict the end of their interlocutor's turn to come in on time. The interaction engine hypothesis (Levinson, 2006) suggests the ability to appropriately time turns in social interaction is realized early in development, before and independent of language. Few studies have assessed timing of turn-taking in infant development. We analyzed video-recordings of 12 motherinfant dyads at 12 and 18 months in free-play interactions. Findings indicate that in the first half of the second year of life infants become more skilled in taking turns in vocal exchanges as evidenced by decreasing onset times of their turns (median = 700 ms at 18 months) as well as a decrease in number of onsets produced in overlap with their mothers, which at 18 months is at the maternal level of overlapping onsets produced (20\%).
\end{abstract}

\title{
Working Memory and Abstract Representation in the Context of Culture
}

\author{
Mark Ho \\ Brown University, Providence, Rhode Island, USA \\ Fiery Cushman \\ Brown University, Providence, Rhode Island, USA
}

\begin{abstract}
As a species, humans stand out for superior cognitive capacities, including improved working memory and abstract representation. However, these abilities evolved, develop, and are generally utilized within a context of cultural transmission and ongoing interaction with the environment. This raises two complementary questions: To what extent does culture scaffold effective employment of our cognitive capacities, and to what extent is learning possible in culture's absence? Experiment 1 demonstrates that participants given a verbal "hint" can use working memory to optimize rewards in a simple sequential learning task, whereas even after hundreds of trials of experience, those not given a hint can only learn the task suboptimally. Experiment 2 demonstrates similar results for hierarchical rule abstraction. In these experiments, hints are akin to cultural scaffolding, and their influence on learning helps identify how our ability to spontaneously leverage our unique cognitive capacities is limited in isolation from culture.
\end{abstract}

\title{
The Atoms of Cognition: Action Learning and Problem Solving
}

\author{
Seng-Beng Ho \\ National University of Singapore, Singapore, Singapore, Singapore
}

\begin{abstract}
In a previous paper we showed that a set of representations, which we referred to as "atomic operational representations," which are explicit spatiotemporal representations, can perform the function of grounding concepts of activities and interactions in the physical world. In this paper, to demonstrate how these operational representations can function in cognitive processes, we develop the basic ideas further by showing 1) how actions and their consequences can be observed and captured in operational representations; 2) how causal rules of actions can be learned and encoded in the form of operational representations through an unsupervised causal learning process; and 3) how the learned causal rules can be used in problem solving processes that produce desired action plans. We show that the same representations can be used across the various levels of cognitive processing in a unified manner. Experiments are proposed to test if the brain uses explicit temporal representations.
\end{abstract}

\title{
Measuring the Use of Recognition with the Multinomial r-Model: An Application to Aging and Individual Differences
}

\author{
Sebastian Horn \\ Max Planck Institute for Human Development \\ Thorsten Pachur \\ Max Planck Institute for Human Development \\ Rui Mata \\ Max Planck Institute for Human Development
}

\begin{abstract}
It has been suggested that individuals use simple decision strategies for comparative judgments. According to the recognition heuristic (RH; Goldstein \& Gigerenzer, 1999), people infer that a recognized object scores higher on a criterion, if one of two objects is recognized, but not the other. Hilbig, Erdfelder, and Pohl (2010) have argued that previous research lacked process-pure estimates of RH use and rigorous model testing and proposed the r-model, a multinomial processing tree (MPT) model. Addressing these methodological issues, we present a first MPT analysis of differences in RH use between younger and older adults. Model-based analyses indicated that in both age groups the RH was used adaptively more often in the environment with higher recognition cue validity (cities), as opposed to a domain with lower cue validity (diseases). The validity of further knowledge or recognition as decision cues did not differ between age groups. Moreover, we examined the model estimates on the individual level by applying a Bayesian hierarchical approach and compared these estimates with behavioral indices and measures derived from signal-detection theory. The resulting comparisons with standard RH-adherence rates indicated high correlations. Further implications are discussed.
\end{abstract}

\title{
Belief Updating in Moral Dilemmas
}

\author{
Zach Horne \\ University of Illinois at Urbana Champaign Derek Powell \\ University of California, Los Angeles, California \\ Joseph Spino \\ University of Illinois at Urbana-Champaign, Urbana, Illinois, United States
}

\begin{abstract}
Research on two classic moral dilemmas, Trolley and Footbridge, suggests that one's past moral experiences can affect one's subsequent moral decisions. These dilemmas have interested moral psychologists, in part, because they have found that people's judgments about the dilemmas are affected by the order in which the dilemmas are considered. Furthermore, this effect is asymmetrical: people that consider Trolley after Footbridge have significantly different judgments than people in control conditions, but the converse is not true. We argue that this asymmetry is the result of a difference in how the each dilemma affects pre-existing beliefs regarding the importance of saving lives. In two experiments, we show that Footbridge disconfirms these beliefs, while Trolley does not significantly affect them. Consistent with predictions of a belief adjustment model of ordering effects, these findings offer a clear and parsimonious account of the asymmetry.
\end{abstract}

\title{
The effects of caregivers' instruction on children's learning in easy and difficult tasks
}

\author{
Shumeng Hou \\ The Chinese University of Hong Kong, Hong Kong, Hong Kong, China \\ Zinnia Lam \\ The Chinese University of Hong Kong, Hong Kong, Hong Kong, China \\ Wing Chee So \\ The Chinese University of Hong Kong, Hong Kong, Hong Kong, China
}

\begin{abstract}
Previous research has shown that gestures produced by caregivers and teachers facilitate children's learning in problem-solving tasks. However, little is known about whether such facilitating effect varies with the task difficulty. We here asked twenty-eight three-year-old children to participate in two puzzle games (12-piece and 20-piece), with three episodes in each game. In Episodes 1 and 3, children played alone. In Episode 2, caregivers instructed their children (e.g., "Let's put this piece upside down" while rotating left hand clockwise). For both puzzle games, children assembled more puzzles in Episode 3 than in Episode 1, suggesting that caregivers' instructions were beneficial for children's learning. However, such benefit was significantly greater in 12 -piece than in 20-piece, \(\mathrm{t}(27)=1.71, \mathrm{p}<.05\). This finding lends support to Vygotsky's theory in which children can gain more from caregiver's scaffolding when the task is within their capacity than when the task is beyond their capacity.
\end{abstract}

\title{
Are beliefs biased by logic? The effect of a secondary load and complexity on belief and logic based judgments.
}

\author{
Stephanie Howarth
}

Plymouth University, Plymouth, England
Simon Handley
Plymouth University

\author{
Clare Walsh
}

Plymouth University

\begin{abstract}
Dual-Process accounts claim that responses to reasoning tasks often default to automatically cued belief-based responses. However, recent findings show that when participants are instructed to evaluate the believability of a conclusion, its logical status interferes with their judgment. This finding is inconsistent with the view that belief based judgments are cued automatically. In this paper we present the results of three experiments that examined the impact of a secondary task (random number generation) on belief and validity judgments. Experiment 1 examined simple modus ponens arguments, experiment 2 included disjunctive syllogisms and experiment three employed a blocked presentation design. In line with previous research belief judgments took longer and resulted in more errors than validity judgments. However, in general, RNG impacted more on validity than belief based judgements. These finding suggest that both belief and logic judgements require effortful processing but draw upon different types of executive resource.
\end{abstract}

\title{
SRNEngine: A Windows-based Neural Network Simulation Tool for the Non-programmer
}

\author{
Steven R. Howell \\ Keystone College, La Plume, Pennsylvania, USA \\ \section*{Suzanna Becker} \\ McMaster University, Hamilton, Ontario, Canada
}

\begin{abstract}
SRNEngine is a windows-based application package for training neural networks. The graphical user interface allows the drag-and-drop creation of neural networks with a variety of architectures, without the need for any programming. At present, these architectures/learning algorithms include Simple Recurrent Networks, Jordan networks, and any kind of feedforward backpropagation network, with up to five each of input, hidden, and output layers (pools of units). A version that adds backpropagation-through-time is in development. The interface is designed to conform to the Microsoft Windows GUI environment that most PC users are already familiar with. SRNEngine includes tools for creating, editing, and manipulating various types of training data, and is especially optimized for working with text/language data, including automatic word-to-input-representation translation at runtime for text corpora. The distributed computing feature allows multiple simulations to be run on a network of workstations, co-ordinated via a central ftp server.
\end{abstract}

\title{
Comparison and explanation in learning: Children's understanding of a basic engineering principle
}

\author{
Christian Hoyos \\ Northwestern University, Evanston, Illinois, USA \\ Dedre Gentner \\ Northwestern University
}

\begin{abstract}
We investigated the roles of comparison and explanation in teaching children an important engineering principle that triangular cross-bracing confers stability to structures. We aimed to discover how best to convey this principle to 4 - and 6 -year-olds and to reveal the cognitive mechanisms involved. Children either compared contrastive cases (a braced building vs. a non-braced building), received an explanation of the principle, or both, and were then tested on their ability to apply the principle to various contexts. We found that 4 -year-olds benefited from comparison, but surprisingly did not benefit from a combination of comparison and explanation. 6-year-olds, however, benefited greatly from the combination, suggesting that more developed abilities are required to combine the two inputs. Performance on a mental transformation task was also related to successful brace placement. These findings suggest that comparison and explanation can both contribute to learning, both singly and together, depending on ability and/or age.
\end{abstract}

\title{
The roles of L1 and markedness on Mandarin L2ers' construction of syntactic representation
}

\author{
Dong-Bo Hsu \\ National Taiwan Normal University, Taipei, Taiwan, Taiwan, R.O.C.
}

\begin{abstract}
Second language learners' construction of their syntactic representation is greatly influenced by their L1 characteristics and the input of their L2. Although it is known that the more universal a pattern is, the easier it will transfer to facilitate the formation of the representation, the role of markedness on such representation remains unclear. The current study tested two groups of Mandarin L2 learners, i.e., native speakers of English and Japanese with three levels of proficiency using the four structures in Mandarin: SVO, ba OV, S ba O V, and topicalization with novel verbs and neutral animacy cues in a forced choice paradigm to investigate what role of markedness plays in such representations. The results indicated that in addition to the initial transfer for the syntactic representation that is affected by learners' L1 cue validity, the degree of markedness exerts impact on learners' rate of acquisition of syntax at different levels.
\end{abstract}

\title{
Word Repetition Priming for Chinese Emotional Words
}

\author{
Yu-Shu Chiang \\ National Cheng Kung University \\ Hsin-Hsuan Wu \\ National Cheng Kung University \\ Yi-Ling Chung \\ National Cheng Kung University \\ Hsueh-Chih Chen \\ National Taiwan Normal University \\ \section*{Jon-Fan Hu} \\ National Cheng Kung University
}

\begin{abstract}
The present experiment was to test if high-arousing Chinese words can lead to increase repetition priming for emotional semantics. Participants were randomly assigned into two groups: one group was presented with positive words and another group with negative words. In Phase 1 of the experiment, participants rated high-arousing words and neutral words for concreteness. In Phase 2, they made decision to determine if it was novel word (half high-arousing, half neutral). In Phase 3, they were told to value the features of Chinese words which were not presented previously in a 5 -point Likert scale and finished some parts of the Basic Personality Inventory (BPI) for assessing and controlling possible cognitive processing bias. The results showed a significant priming effect in two groups and the words presented in Phase 1 had shorter reaction times than the novel words. These findings revealed selective enhancement of Chinese word repetition priming by emotional arousal.
\end{abstract}

\title{
Categorization and recognition in a prototype-distortion task: ERP studies
}

\author{
Yanli Huang \\ The Chinese University of Hong Kong, Hong Kong, Hong Kong, China \\ Zhiya Liu \\ South China Normal University,Guangzhou, Guangdong, China \\ Chi Shing Tse \\ The Chinese University of Hong Kong, Hong Kong, Hong Kong, China
}

\begin{abstract}
We used a prototype-distortion task and adopted ERPs to test between prototype and exemplar theories on categorization, which suggest that categories are represented as a prototype via an abstraction process or via storing previously encountered exemplars in memory, respectively. In Experiments 1 and 2, participants were presented low- or high-distortion category-members (i.e., dot-patterns) without anticipating subsequent categorization or recognition test. They were more likely to process low-distortions via prototype abstraction in both tests, and differently process high-distortions via prototype abstraction in categorization test, but via storing exemplars in recognition test. In Experiment 3, participants were explicitly instructed to do categorization or recognition task. We found that participants did categorization test via prototype abstraction (N1) only for studied items, not for unstudied items. And conversely did recognition via familiarity processing (FN400) only for category-members, not for non-members. In conclusion, the nature of category representations depends on the experimental contexts.
\end{abstract}

\title{
Probabilistic negation: fine-grained preservation and distortion of truth in affirmative and negated statements
}

\author{
Stephanie Huette \\ University of California, Merced
}

\author{
Sarah Anderson \\ The Nielsen Company
}

Michael Spivey
University of California, Merced

\begin{abstract}
Previous work has shown that a continuum of truth is reflected in real-time motor movement behavior (McKinstry, Dale \& Spivey, 2008). In a mouse-tracking paradigm, participants responded yes or no to statements of varying truth-values such as "A thousand is more than a million" or "English is a language" as well as more ambiguous statements such as "Murder is sometimes justifiable". In the present study, we replicated these results along an 11-point continuum of truth-values, finding that the end-points of averaged mouse trajectories vary as a function of truth-value. In addition to this, negated versions of each stimulus were tested and revealed that truth-values for negated sentences follow more complex trajectories and do not preserve the original truth-value of the statement. The evidence found presents a problem for theories of negation that require a revision from the affirmative meaning. Alternative mechanisms for how truth is affected by negation are proposed.
\end{abstract}

\title{
Exploring the boundary conditions of intentional forgetting
}

\author{
Almut Hupbach \\ Lehigh University \\ Lili Sahakyan \\ The University of North Carolina at Greensboro
}

\begin{abstract}
The attempt to forget some recently encoded information can indeed cause later retrieval difficulties. However, such attempts are only effective when new information is learned shortly thereafter. In the present study, we asked whether the new information has to match the format of the to-be-forgotten information for forgetting effects to emerge. Participants studied words or line drawings (L1), and were afterwards instructed to remember or forget these items. Then, a second list (L2) was presented that either matched or mismatched the L1 format. Forgetting effects were only observed when the list formats matched. This result establishes an important boundary condition on intentional forgetting, and can be explained by the context change account (Sahakyan \& Kelly, 2002), which assumes that forgetting occurs when retrieval is guided by temporal context only. Salient cues (such as differences in list format) allow for reinstatement of the L1 encoding context, thus eliminating forgetting.
\end{abstract}

\title{
Purity and Disgust: Same Domain, or Separate Mechanisms?
}

\author{
Gordon Ingram \\ Bath Spa University, Bath, Somerset, England \\ \section*{Karolina Prochownik} \\ Jagiellonian University, Kraków, Poland
}

\begin{abstract}
We argue that much recent literature on disgust, dirtiness and purity has been guilty of conflating two evolved mechanisms: an oral disgust mechanism aimed at avoiding the ingestion of dangerous substances, and a self-grooming (cleanliness) mechanism aimed at eliminating ectoparasites from the skin. Though phylogenetically distinct, these two mechanisms become associated in human ontogeny due to their similar targets and overlapping image schemas: one focused on the mouth, the other on the body as a whole. We show that several puzzles in the literature on disgust and moral purity can be resolved using this model. The idea of contamination so central to purity norms may more plausibly be based on grooming responses to ectoparasites than on disgust responses to endoparasites. The disgust image schema may more easily be extended to moral judgements about others, while the cleanliness schema is more easily extended to judgements about the self, with interesting consequences.
\end{abstract}

\title{
How can an art course in photography inspire students' artistic creativity?
}

\author{
Chiaki Ishiguro \\ The university of Tokyo, Tokyo, Tokyo, Japan \\ Takeshi Okada \\ The university of Tokyo, Tokyo, Tokyo, Japan
}

\begin{abstract}
This study examines whether an art course with various kinds of inspiration derived from others and their artworks is useful in improving undergraduates' photographic creativity and their views of photo taking. In collaboration with a professional photographer, we organized an undergraduate course in artistic photography, which included lecture sessions in basic artistic skills and knowledge, imitation sessions of unfamiliar artistic photographs, photo taking sessions, and presentation of the students' own works in the class. 21 students participated in the course for a semester. We collected students' diaries of their photo taking, their photographs, and questionnaire survey data about their photo taking experiences. The results of data analyses show that the creativity of the students' photographs improved after lecture sessions. The students reported that reflecting on their photography contributed to their acquisition of metacognitive knowledge of artistic creation.
\end{abstract}

\title{
Mental Process of Mindreading: Self and Stereotype as Anchors in Mental State Inference
}

\author{
Tatsunori Ishii \\ Tokyo Seitoku University, Tokyo, Japan \\ Masanori Takezawa \\ Hokkaido University, Hokkaido, Japan
}

\begin{abstract}
Capacities of mindreading are essential for human social life. It is hypothesized that people selectively use two types of mindreading strategies; when a target person is perceived to be similar to oneself, people project one's own mental states to the person (projection): when the target is perceived to be dissimilar, category-based stereotype is used (stereotyping). In this study, we tested this hypothesis with the reaction time paradigm (e.g., Tamir \& Mitchell, in press). Given that the both projection and stereotyping are computationally modeled as anchoring-and-adjustment processes, the reaction time paradigm can be used to detect a strategy used in mindreading. We found the stereotyping was unanimously employed independent of the similarity to the target person and projection was employed only when the perceived similarity was high. Our results are congruent with Tamir \& Mitchell (in press) and confirmed the utility of the reaction time paradigm as a tool for investigating mindreading strategies.
\end{abstract}

\title{
Eye-Movement Strategies in Multiple Target Search for Target Location and Uncertainty Reduction
}

\author{
Christian P. Janssen \\ The Smith-Kettlewell Eye Research Institute \\ Preeti Verghese \\ The Smith-Kettlewell Eye Research Institute
}

\begin{abstract}
Where do people look when searching for multiple targets under time pressure: at salient targets, at locations with high uncertainty about target presence, or somewhere else? Preceding research suggests that people tend to look at salient targets. This is suboptimal, because educated guesses can be made about target presence at these locations without looking (Verghese, 2012). We ran an experiment and constructed Bayesian models to test the generality of this finding. Participants saw stimuli at two locations for 400 msec (i.e., allowing only 1 saccade), and then judged target presence at each location. Noise of low or high contrast was superimposed at the two locations. We observed individual differences in saccade strategies. One participant made no saccades, while achieving reasonable performance. Others applied a mixture of strategies, sometimes favoring salient targets, sometimes favoring uncertain locations. This work provides further insight into task and cognitive constraints that influence saccade strategy selection.
\end{abstract}

\title{
'But' how do children reason with it?
}

\author{
Leen Janssens \\ KU Leuven, Leuven, Belgium
}

\section*{Stephanie Drooghmans}

KU Leuven, Leuven, Belgium

\section*{Sara Verbrugge}

Hogeschool Gent, Gent, Belgium

\section*{Walter Schaeken}

KU Leuven, Leuven, Belgium

\begin{abstract}
Our research aimed at investigating whether 8-to-12-year-old children spontaneously make the conventional implicature induced by 'but' -combined with 'so' and 'nevertheless'- in 'p but q' sentences. We presented the children with stories ending with a ' p but q ' sentence. They were instructed to indicate the 'appropriate' conclusion introduced by either 'so' or 'nevertheless'. In addition, we measured children's working memory (WM)-capacity in order to explore the possibility that making these inferences is effortful. Our results show that children do make the inferences to a certain extent but are sensitive to the content of the arguments. Whenever the p - or q -argument is an absurd argument (contrasted with a sensible argument), this argument almost always gets ignored in favor of the sensible argument, irrespective of the 'appropriate' conclusion 'but' directs the reader to. No reliable WM-effect was found. High WM-span children did not make the inference more often than low WM-span children.
\end{abstract}

\title{
Using SpAM to reveal decision processes and complex new context effects in similarity judgment
}

\author{
Gavin Jenkins \\ University of Iowa \\ Larissa Samuelson \\ University of Iowa \\ John Spencer \\ University of Iowa
}

\begin{abstract}
The "Spatial Arrangement Method" (SpAM) has gained popularity for measuring similarity judgments (Perry, Cook, \& Samuelson, 2011; Hout, Goldinger, \& Ferguson, 2012; Kriegeskorte, 2012). In SpAM, multiple stimuli are freely arranged in two dimensions such that more similar stimuli are close together. We performed two SpAM experiments to investigate the process by which participants make multiple simultaneous similarity judgments using novel stimuli. The experiments differed across either two or three feature dimensions. Mouse and Eye-tracking measures, as well as the sequence of stimuli placed, provided a rich picture of participants' decision processes as they made these judgments. Both experiments revealed strong effects of group context. Clustered presentation of stimuli by feature influenced both the order and the timing of placements, and despite equal metric spacing of stimuli along each dimension, participants typically warped placements along dimensions nonlinearly. We discuss implications of these findings for theories and models of similarity.
\end{abstract}

\title{
How Do People Judge Conjunctive Probabilities from Experience? A Hierarchical Bayesian Model Comparison
}

\author{
Mirjam Annina Jenny \\ Max Planck Institute for Human Development, Berlin, Germany \\ Jörg Rieskamp \\ University of Basel, Basel, Switzerland \\ Håkan Nilsson \\ Uppsala University, Sweden
}

\begin{abstract}
Judging whether multiple events will co-occur is an important aspect of everyday decision making; however, the underlying probabilities of occurrence are usually unknown and have to be inferred from experience. Using a rigorous, quantitative model comparison, we investigate how people judge the probabilities of multiple events to co-occur. In a computerized experiment, participants had to repeatedly choose between two pairs of conjunctive events (represented as two gambles). Participants had access to a small sample of information to estimate the probability that both events occur. A hierarchical Bayesian approach used for estimating the models' parameters and for testing the models against each other showed that the plurality of participants were best described by the configural weighted average model. This model assumes that constituent probabilities are ranked by importance, weighted accordingly, and added up. The cognitive modeling approach provides an understanding of the cognitive processes underlying people's conjunctive probability judgments.
\end{abstract}

\title{
Social institutions as tools in normative cognition
}

\author{
Jeppe Sinding Jensen
}

Aarhus University, MINDlab

\begin{abstract}
Humans have the unique ability to have goals, cooperate according to plans and in respect of norms and rules (football is a good modern example). This ability is based on 1) collective intentionality among participants, 2) distributed cognition of the shared plan and 3) normative cognition in the ability to follow rules and 4) evaluate practice in relation to norms.

In cooperative social interaction, social institutions function by uniting these 4 dimensions. Social institutions 'make us smart' collectively when constitutive rules and regulative rules on the socio-cultural level are internalized as constitutive and regulative representations on the cognitive level (knowing 'what counts as what').

Social institutions are cognitive tools with force because they 'store' and 'radiate' normative cognition and often 'crystallize' in rituals, e.g. weddings (fusing 1-4). The functions of normative cognition in social institutions should be an important subject in cognitive anthropology.
\end{abstract}

\title{
Modeling spatial language acquisition as a function of lexical verb development
}

\author{
Kristen Johannes \\ Johns Hopkins University, Baltimore, Maryland, United States \\ Barbara Landau \\ Johns Hopkins University, Baltimore, Maryland, United States \\ Colin Wilson \\ Johns Hopkins University, Baltimore, Maryland, United States
}

\begin{abstract}
Learning to linguistically encode spatial relations is traditionally considered a problem of acquiring the meanings of prepositions (in, on, under). Based on production data from English speaking 4-year-olds, 6-year-olds and adults, we provide evidence for an alternative verb-based hypothesis: children and adults may essentially share spatial concepts and prepositional semantics, differing primarily in their use of lexical verbs (hang, stick, attach) to describe spatial relations. This hypothesis was formalized as a hierarchical generative model in which child and adult spatial descriptions are drawn from a common distribution, modulo a penalty on lexical verbs that is stronger for children. The model accounted for child production data significantly better than a model based on average adult performance, and the strengths of the estimated penalties were overall greater for 4-year-olds than for 6-year-olds, suggesting a developmental process in which lexical verbs gradually become integrated into the linguistic system for describing spatial relations.
\end{abstract}

\title{
Eye movements "steal" response time effect during imagery scanning
}

\author{
Roger Johansson \\ Cognitive Science Department, Lund University \\ Jana Holsanova \\ Cognitive Science Department, Lund University
}

\begin{abstract}
Several behavioral studies have reported functional similarities between visuospatial imagery and visuospatial perception. For instance, in the classic image-scanning paradigm (Finke \& Pinker, 1982) participants first inspect a dot pattern, which later disappears and is replaced by an arrow on a blank screen. The task is to judge whether the arrow points towards one of the previous dots. Results commonly show that the response time (RT) increases linearly with the distance between the arrow and the previous dot and has been taken as evidence for a structural equivalence between perception and imagery. Typically eye movements are prevented in this paradigm. In the present study, eye movements were recorded for 23 participants in a free viewing version of the image-scanning paradigm. Results revealed that saccadic amplitudes increased linearly with the "imagined" scanning distance ( \(\mathrm{p}<.001\) ). But contrary to previous studies there was no significant effect between RT and scanning distance.
\end{abstract}

\title{
Early Event-Related Potentials (ERPs) sensitive to animacy expectations in sentence comprehension are not overridden by context
}

\author{
Alexis R. Johns \\ University of Connecticut and Haskins Laboratories \\ Heather K. J. van der Lely \\ Harvard University \\ James S. Magnuson \\ University of Connecticut and Haskins Laboratories
}

\begin{abstract}
Wh-words in auditory sentences like "Who/What did Barbie push the _-_ into?" generate expectations for animacy at the blank (e.g., a filled potential Wh-gap). Specifically, the animacy is expected to be opposite of the Wh-word ("who" and "what" predict inanimate and animate nouns, respectively). Fontenau and van der Lely (2008) found Early Left Anterior Negativities (ELANs) when animacy matched Wh- animacy in typically developing individuals but not individuals with "Grammatical-Specific Language Impairment". However, to focus attention on the task, they added final noun phrases to violation items ("Who did Barbie push the clown into THE WALL?") but not to expected animacy items ("What did Barney push the clown into?"). We ask whether participants implicitly learn to predict sentence-final anomalies from animacy match. We tested this by presenting one group with the original contingency and another with the contingency reversed. Contra the learning hypothesis, we observed ELANs for both groups.
\end{abstract}

\title{
Adult Category Learning Differences Predicted by a Dynamic Neural Field Theory Account of Information Sampled from the Fovea
}

\author{
Jordan Barnes \\ Simon Fraser University
}

\author{
Mark Blair \\ Simon Fraser University \\ Paul Tupper \\ Simon Fraser University \\ R. Calen Walshe \\ University of Edinburgh
}

\begin{abstract}
Here we explore the possibility that the speed of learning is affected by the precision of our sensory estimates for the learned category's diagnostic feature dimensions. Colour information from a foveated stimulus, if represented as a sample on a metric colour dimension, should faithfully represent differences in the shapes and precision of the estimates as a consequence of the sample size. Differences in the sample variability are expected to have affect on exactly what gets associated during learning. We provide evidence that a manipulation in sub-fovea feature size, \(0.18^{\circ} \mathrm{vs} 1.19^{\circ}\) of visual angle, influences learning speed. In both conditions the simple colour features are easy to see and we do not detect any gaze differences as measured by total fixation durations and individual feature fixation durations. Learning methods that metrically represent activity on feature dimensions such as Dynamic Field Theory (DFT) may be able to account for this data.
\end{abstract}

\title{
Small Elephants and Big Needles: Can Perceptual Information Affect Memory and Judgments about the Meaning of Words?
}

\author{
Natalie A. Kacinik \\ Brooklyn College, CUNY
}

\author{
Rita W. El-Haddad \\ The Graduate Center, CUNY
}

\author{
Kendall J. Eskine \\ Loyola University New Orleans
}

Lolly Starr-Glass
Brooklyn College, CUNY

\author{
Samuel Salamon \\ Brooklyn College, CUNY
}

\begin{abstract}
There is considerable evidence that representations of word meaning are "embodied" and grounded in our perceptual and motor experiences (Barsalou, 2008; Glenberg, 2010). This research has mostly relied on priming and interference procedures, or measuring brain activity. The present study manipulated the perceptual appearance of words, specifically font size, to be congruent or incongruent with an object's actual size (e.g., elephant presented in a large or small font, respectively). Participants were presented with the words prior to a recognition memory test and property judgment task, in the same session and after a 2-week delay, to see if the perceptual font information would be incorporated into the representations of words to potentially alter participants' memory and judgments. Font size generally did not significantly affect how participants represented and processed the words. These results therefore present a challenge for embodied accounts of semantics, but some potential explanations and issues will be discussed.
\end{abstract}

\title{
Navigating the Social Environment: An Ecological Rationality Perspective on Advice Taking Behavior
}

\author{
Juliane Eva Kämmer \\ Max Planck Institute for Human Development, Berlin, Germany \\ Hansjörg Neth \\ Max Planck Institute for Human Development, Berlin, Germany \\ Pantelis Pipergias Analytis \\ Max Planck Institute for Human Development, Berlin, Germany \\ Mehdi Moussaïd \\ Max Planck Institute for Human Development, Berlin, Germany
}

\begin{abstract}
Many decisions are made under the advice of another person. We investigated the environmental circumstances under which two prominent strategies-averaging and choosing-are effective and adaptive and explored how people employ them. We report two experiments, in which participants ( \(\mathrm{N}=111\) and \(\mathrm{N}=90\), respectively) provided initial estimates for general knowledge questions that varied in perceived difficulty. In Experiment 2, they additionally received advice in the form of an estimate and confidence rating of another person before providing a revised estimate. We found that items of different perceived levels of difficulty exhibited distinctive statistical properties, thus constituting different social environments. Environmental structure affected the theoretical performance of strategies (such as averaging and choosing), and the ways, in which people integrate advice. We embed our analyses in the frameworks of ecological rationality and the probability, accuracy, redundancy (PAR) model of advice taking (Soll \& Larrick's, 2009; JEP:LMC, 35).
\end{abstract}

\title{
Be a better multitasker. How a pause in the primary task can turn a rational into an irrational multitasker.
}

\author{
Ioanna Katidioti \\ University of Groningen, Groningen, Netherlands \\ Niels Taatgen \\ University of Groningen
}

\begin{abstract}
Threaded cognition theory predicts that switching is opportunistic and depends on availability of cognitive resources. Laboratory studies of multitasking suggest people are rational in their switch choices regarding multitasking, while observational studies suggest they are not. To establish whether effective multitasking can become ineffective we introduced delays in the primary task.

The participants answered emails by looking up information (similar to customer-service employees) while being interrupted by chat messages. When participants were faced with a delay in the email task, they switched more often to the chat task on high-workload points. Choosing to switch to the secondary task instead of waiting made them slower. It also made them forget the information of the e-mail task half of the time, which slowed them down even more.

We concluded that people's rationality in multitasking behavior is only local, which agrees with the threaded cognition account of switching.
\end{abstract}

\title{
KANSEI Differences between the Child and the Caregiver in Room Arrangement Examined by the Room Arrangement Workshop for Child Room Spaces
}

\author{
Yoko Katsumata \\ Tokyo Denki University \\ Hideaki Kitazume \\ Tokyo Denki University \\ Tetsuya Yasuda \\ Saitama Prefectural University
}

\author{
Harumi Kobayashi
}

Tokyo Denki University

\begin{abstract}
We investigated KANSEI difference between a child and a caregiver in room arrangement workshop. Seven pairs of parents and children volunteered to participate in a workshop held in a university in Japan. In the workshop, each child and caregiver arranged a room layout using a prepared set of furniture. We then interviewed them about the layout of the arranged room and their own room at their home. We categorized their arrangement of the furniture into "center-arrangement" and "corner-arrangement" using video-data. Results show that half of the children arranged a piece of furniture in the center of the room whereas all caregivers arranged a piece of furniture at a corner. Interview data suggested this difference reflected their KANSEI differences such as preferred activities and perspectives toward room arrangement between children and caregivers.
\end{abstract}

\title{
The Influence of Observing Versus Imitating Gestures in L2 Phoneme Instruction
}

\author{
Spencer Kelly \\ Colgate University, Hamilton, NY, USA \\ Yukari Hirata \\ Colgate University, Hamilton, NY, USA \\ Carmen Lin \\ Northwestern University, Evanston, IL, USA \\ Zach Zhao \\ Colgate University, Hamilton, NY, USA
}

\begin{abstract}
Previous research has shown that auditory training helps non-native speakers learn to perceive difficult phonemic contrasts in a second language (L2) (Hirata, 2004), but there is much room for improvement. Given that hand gestures influence many aspects of native language processing (Hostetter, 2011), we examined whether imitating versus observing gestures helps to improve native English speakers' ability to perceive novel phoneme contrasts in Japanese as an L2. Participants were assigned to either a gesture observe or gesture imitate training condition. There was no overall training advantage of imitating gestures over simply observing them. However, in a preliminary analysis of a sub-group of participants with low scores on the auditory pretest, observing gestures was actually more beneficial than imitating them on an ERP post-test of auditory perception. The results suggest that producing gestures does not always help with learning, and for particularly challenging auditory perceptual tasks, may actually interfere with it.
\end{abstract}

\title{
Tracking Scrambled Word Order While Reasoning with Diagrams
}

\author{
Özkan Kılıç \\ Department of Cognitive Science, Graduate School of Informatics, Middle East Technical University (METU), Ankara, Turkey
}

\begin{abstract}
Word order varies not only across languages but also within a specific language. Turkish, for example, has a nonrigid word order. Therefore, it is a good test-bed to understand the processing complexity driven by the order. One way to observe the complexity is to investigate it via diagrammatic reasoning. In this study, 18 sentences with scrambled word orders and the corresponding diagrammatic representations were analyzed by 20 Turkish native speakers. There were deliberate errors in the representations and the participants were asked to report them. The participants' eye-movement data were also collected. Results indicate that scrambled word ordering also causes latencies in diagrammatic reasoning task. The eye-movement fixation orders showed that the participants favored VSO and VOS eye-fixation orders independent of sentential word orders. It is also concluded that finding the errors in verbal representations were more time consuming than finding the errors in object and subject representations.
\end{abstract}

\title{
The role of Quinian bootstrapping in the acquisition of mental state terms
}

\author{
Szabolcs Kiss \\ Institute of Psychology, University of Pecs
}

\begin{abstract}
The present poster discusses the role of the famous Quinian bootstrapping learning process in the acquisition of mental state terms such as happy, believe, pain, etc. At first, the poster characterises Quinian bootstrapping in which the so-called placeholder structure plays an important role. The placeholder structure consists of symbols whose meanings are initially learned in terms of each other. Later, the placeholder structure is infused with meanings via the so-called modelling processes. A modelling process can be analogical mapping, abduction, induction, etc. Susan Carey (2009) introduced Quinian bootstrapping in her explanation of the acquisition of numeral list representation and rational number as well as certain aspects of intuitive physics. Second, I apply this well-known learning mechanism to the acquisition of the meaning of mental terms. I distinguish between three stages in the learning of the semantics of mental words. The present poster will characterise the three stages in detail.
\end{abstract}

\title{
An Explorative Study of Search of Model Space in Problem Solving
}

\author{
Saskia Kistner \\ Goethe University, Institute of Psychology, Frankfurt, Germany \\ Bruce D. Burns \\ University of Sydney, School of Psychology, Australia \\ Regina Vollmeyer \\ Goethe University, Institute of Psychology, Frankfurt, Germany \\ Ulrich Kortenkamp \\ Martin-Luther-University, Institute of Mathematics, Halle, Germany
}

\begin{abstract}
Building on dual-space theories, the three-space theory of problem solving suggests to add search of a model space in addition to search of experiment and hypothesis space. This study aimed at exploring the three postulated spaces, especially model space, by means of verbal protocols.

Participants ( \(\mathrm{n}=32\) ) were asked to think aloud while working with a computer based learning program. With this program they could learn about torques in physics using interactive graphics in which experiments could be conducted. Their knowledge about torques was tested before and after working with the program. Verbal protocols were analyzed with regard to the amount of search of the three spaces and regarding the quality of the participants' models for torques.

Our results add to the validity of model space, showing that the three postulated spaces could be reliably identified in the protocols and that the model quality score predicted final knowledge beyond prior knowledge.
\end{abstract}

\title{
Distributed Memory System Architecture based on the Analyses of Human Brain Memory
}

\author{
Muneo Kitajima \\ Nagaoka University of Technology \\ Makoto Toyota \\ T-Method
}

\begin{abstract}
A novel human memory system architecture is proposed. The memory system is an integration of three distributed memory systems associated with respective autonomous organic systems, including the perceptual system that takes care of sensory input from the environment, the conscious system that performs deliberate decision making, and the unconscious system that carries out action selections in the environment. This memory system architecture is consistent with the wide range of recent findings in the field of neurosciences. The memory system architecture works as a memory component in the comprehensive real brain model, MHP/RT, published in the Cognitive Science conferences, and the BICA conferences. MHP/RT is capable of simulating human daily behavior considering real time constraints that should define strong mutual dependencies among the three systems. With this memory system architecture, MHP/RT becomes a real brain model to be contrasted with virtual and partial models, such as ACT-R.
\end{abstract}

\title{
Effects of verbalization on lie detection
}

\author{
Sachiko Kiyokawa \\ Nagoya University \\ Yoshimasa Ohmoto \\ Kyoto University
}

\author{
Kazuhiro Ueda \\ The University of Tokyo / Japan Science and Technology Agency
}

\begin{abstract}
We determined whether verbalization had an effect on lie detection. Participants were randomly assigned to one of the three conditions: the lie condition, the truth condition, or the control condition. They were asked to indicate whether the target man was either lying or telling the truth. Prior to making their judgments, the participants in the lie condition were required to describe some behaviors exhibited by the target individual that indicated that he was lying. Similarly, prior to making their judgments, those in the truth condition were asked to describe some behaviors exhibited by the target individual that indicated that he was telling the truth. The participants in the control condition were asked only to make judgments. The participants in the lie condition detected lies more often than those in the other two conditions. Thus, verbalization influenced lie detection.
\end{abstract}

\title{
Anisotropy of personal space examined in a virtual environment
}

\author{
Takatsugu Kojima \\ Shiga University of Medical Science
}

\section*{Ayae Iwamoto}

Kyoto University

\author{
Masashi Sugimoto \\ Kyoto University
}

\begin{abstract}
How does a person's personal space change when the person approaches or retreats from another person? In this study, we examine the anisotropy of personal space in a virtual room constructed by three-dimensional computer graphics. In two experiments, a participant took the first-person point of view, and an unfamiliar avatar was placed in the virtual room. The stop-distance technique was used to measure the personal space between the participant and avatar. The participant was required to approach the avatar until he or she felt uncomfortable (approaching condition) or to retreat from it until he or she felt comfortable (retreating condition). We also controlled the room size and the unfamiliar avatar's direction. The results clearly showed that the personal space was larger under the approaching condition than under the retreating condition. Moreover, we found that the avatar's direction influenced the effect of the room size on personal space.
\end{abstract}

\title{
Information flow across individuals in formation of symbol communication systems
}

\author{
Takeshi Konno \\ Japan Advanced Institute of Science and Technology \\ Junya Morita \\ Japan Advanced Institute of Science and Technology \\ Takashi Hashimoto \\ Japan Advanced Institute of Science and Technology
}

\begin{abstract}
In order to study the formation of symbol communication systems, we conducted an experiment based on a game introduced by Galantucci (2005). This involved pairwise communication in which the pairs engaged in a coordination task through an exchange of messages composed of a small set of geometric figures. We analyzed the transfer entropy-which is a measure of the information flow between two information sources or stochastic processes-between the figures used and the actions by individuals. Consequently, we confirmed that the transfer entropy reduced significantly across individuals, but not within each individual. Moreover, it correlated negatively with the performance of the task. From these results, we suggest that the transfer entropy appropriately shows the degree of formation of symbol communication systems. Then, we report the causal relationship between the uncertainty and the performances using a computational model based on reinforcement learning.
\end{abstract}

\title{
Are forgetting processes crucial to category learning?
}

\author{
Lukasz Kopec \\ University College London (UCL), London, London, United Kingdom \\ Brad C. Love \\ University College London
}

\begin{abstract}
In noisy domains where category distributions overlap, people categorise better at test after being trained on idealised category structures. This may happen, because under the assumption that humans selectively sample from memory when performing categorisation, idealised category learning leads to sampling of more appropriate items and better performance. Here we propose that idealisation of category distributions occurs naturally via a process of forgetting and re-estimation of category labels.

We model a process in which items' category membership is forgotten and then re-estimated from the remaining distributions. With time this leads to lowering the variance of category distributions, equivalent to idealising training data. We test this potential idealisation in a paradigm in which we train participants on overlapping category distributions and withdraw feedback for some trials in one group thus enforcing re-estimation of categories. The model predicts that the group with less feedback will perform better at test due to idealisation.
\end{abstract}

\title{
Words activate categorization even before the category forming task has been offered
}

\author{
Alexey Kotov \\ Department of Psychology, The National research university "Higher school of economics" \\ Tatyana Kotova \\ Laboratory of social competencies and social intelligence research, Moscow State University of \\ Psychology and Education \\ Elizaveta Vlasova \\ Department of Psychology, Russian State University for the Humanities
}

\begin{abstract}
The effect of language on category learning is an ongoing debate among researchers. According to previous research words can facilitate category formation even if they aren't used as feedback. However in most research investigating language influence on category learning, the varying of verbal labels often correlates with varying of perceptual features. Such confound doesn't allow to clarify if the language is a means of perception augmentation (language-feedback hypothesis) or a social marker for generalization (word-meaning-as-intention hypothesis). In the present experiment we separated the process of category learning from the label receiving. Two groups of subjects performed visual search task either with or without labels. Right after that task subjects had to form a category on the basis of new perceptual information added to the old one. As a result subjects from label condition form a category but subjects from the no-label condition didn't. The given data agree with word-meaning-as-intention hypothesis.
\end{abstract}

\title{
Influence of Ethnic Group-Membership and Gaze Direction on Emotion perception. An fMRI study.
}

\author{
Katharina Krämer \\ University Hospital Cologne, Germany \\ Gary Bente \\ Department of Psychology, University of Cologne, Cologne, Germany \\ Iva Barisic \\ Department of Psychology, University of Cologne, Cologne, Germany
}

\author{
Bojana Kuzmanovic \\ Institute for Neuroscience and Medicine - Ethics in the Neuroscience (INM8), Research Center Juelich, Germany
}

Kai Vogeley
University Hospital Cologne, Germany; Institute for Neuroscience and Medicine - Cognitive Neuroscience (INM3), Research Center Juelich, Germany

\begin{abstract}
Emotion perception is not only influenced by the ethnic group-membership of interaction partners but also depends on whether a person is engaged with the encoder of an emotion. To characterize the neural correlates of the influence of ethnicity and engagement on emotion perception, German participants rated the valence of video-sequences while undergoing fMRI. In these video-sequences Asian-looking and European-looking virtual characters expressed a positive (happiness) and a negative (anger) emotion while either gazing directly at the participants or at another agent, thereby varying the perception of engagement. Results show that the ventromedial prefrontal cortex is involved when participants observe ethnic in-group members compared to ethnic out-group members express a positive emotion at them compared to at another person. In contrast, the dorsolateral and dorsomedial prefrontal cortex are involved when participants observe ethnic out-group members compared to ethnic in-group members express a negative emotion at them compared to at another person.
\end{abstract}

\title{
Conceptual Art and Cognitive Science: A Case Study in Space
}

\author{
Alexander Kranjec \\ Duquesne University and the Center for the Neural Basis of Cognition, Pittsburgh, Pennsylvania, United \\ States
}

\begin{abstract}
Conceptual art and cognitive science have more common ground than is acknowledged. For example, both disciplines are principally engaged in describing and visualizing facts about basic categories of mind (space, objects, language, etc.). Along these lines, conceptual art can inform the cognitive science of abstract concepts. Cognitive scientists studying aesthetics can also learn from conceptual artists to push their research forward. While empirical investigations in aesthetics typically focus on perceptual preferences (i.e., "what is beauty?"), conceptual art often goes deeper ontologically (i.e., "what is art?"). This level of analysis can inform questions regarding the evolution of art and object processing. The present study examines the artwork of Mel Bochner who may have staged the first conceptual art exhibition (1966). Bochner's work addresses spatial semantics/representation, and anticipates the neuropsychological distinction between categorical and coordinate spatial relations; all while reconsidering what it means for an object to be thought of as art.
\end{abstract}

\title{
Efficiency of feature detection of visual stimuli influences the proportion judgments
}

\author{
Hyun-Min Ku \\ Ochanomizu University \\ Midori Tokita \\ Ochanomizu University
}

\author{
Masami Ikeda \\ Jumonji University
}

\author{
Akira Ishiguchi \\ Ochanomizu University
}

\begin{abstract}
Some studies on judgment and decision-making have demonstrated that uncertainty in probability influenced curvature of function given by judged probability or proportion (i.e., probability weighting function). In this study, we concentrated on proportion judgment about visual input. We investigated the possibility that the efficiency of feature detection of visual stimuli would be related to uncertainty in their proportion judgment. Concretely, our participants in the experiments performed proportion judgment task using two sets of visual stimuli pairs, "red vs. green dots" and "right vs. left tilted lines". In addition, we adopted different types of responses; in one condition, the participants were asked to response using numerals; in another condition, they were asked to response by adjusting the line bar which indicated the proportion. Our results suggested that the efficiency of feature detection of visual stimuli influenced the curvatures of the probability weighting function.
\end{abstract}

\title{
The Effect of Set in the process of perception of unconscious information
}

\author{
Natalya Kudelkina \\ Samara State University, Samara, Russia
}

\begin{abstract}
Our research aims to study the Effect of Set in the process of perception of unconscious information. In the experiments, we use serial (multiple) unconscious stimulation and the modified masked priming technique to give unconscious stimuli to our participants. In the experiments we demonstrate that, as a result of the specially organized preliminary experimental series, it is possible to receive the steady priming-effect for the unconscious stimulus which initially didn't render any noticeable influence on the effectiveness of the tasks solutions. We can conclude, that at unconscious level cognitive system analyzes series of unconscious influences. So, the nature of influence of each following unconscious stimulus on the current conscious cognitive activity, by the time of its occurrence, is already set on the basis of the analysis of the previous series of influences. It is possible to model these individual settings in the experimental environment.
\end{abstract}

\title{
Navigator-driven placement of landmarks: effects on wayfinding performance in a virtual Tate Gallery.
}

\author{
Saskia Kuliga \\ Universität Freiburg \\ Rul von Stülpnagel \\ Universität Freiburg \\ Christoph Hölscher \\ Universität Freiburg
}

\begin{abstract}
Wayfinding difficulties in architecturally ambiguous environments can be overcome by orientating with given landmarks. However, it is not clear at which locations landmarks are most suitable to facilitate orientation. Our study addressed two questions: (1) Is subsequent wayfinding performance facilitated if participants freely place a number of landmark objects in a complex building? (2) Where do participants place the landmarks and which strategies guide effective landmark placement? First, participants were instructed to learn a number of goal locations in a virtual model of the Tate Gallery London. Then, participants in the experimental condition were instructed to place five unique landmarks in order to re-find the goals in a third phase; participants in a control condition could not place landmarks. Finally, participants were tested on their ability to find the goals again, with time and distance as main dependent variables. Results are discussed with respect to placement strategies and environmental properties.
\end{abstract}

\title{
Popout Attention with two foils: linear dependence and dimensional interaction
}

\author{
Kiran Kumar \\ Indiana University \\ Suyog Chandramouli \\ Indiana University \\ Richard Shiffrin \\ Indiana University
}

\begin{abstract}
Automatic parallel processing occurs in visual search when a target differs from other display objects on a salient visual dimension and the phenomenon is termed 'popout'. The present two studies answers the question: Do the interference effects cancel or add? Both studies used a ring of twelve fixed-size green circles with embedded Gabors oriented vertically or horizontally. Targets and foils were one of green square, a larger green circle, or a red circle. Each of these served as a target for one of the three sessions, with the others serving as foils. In one task the observer found the target and reported its Gabor orientation. The other task used targets only on one half the trials and the observer reported target presence or absence. In both tasks accuracy was uniformly high. In both tasks RT interference increased from one to two foils, the slowing mainly isolated to the slowest decile of RTs.
\end{abstract}

\title{
Using Relational Encoding to Promote Creative Problem Solving
}

\author{
Kenneth J. Kurtz \\ Binghamton University \\ Nuoya Zhang \\ Binghamton University \\ Tamar Skolnick \\ Binghamton University
}

\begin{abstract}
The nature and basis of creative thought is the subject of wide-ranging inquiry, yet remains elusive. We focus on a core issue: why do people struggle to solve problems that require a creative insight and what type of support can make success into the norm? Drawing on the idea of relational encoding (e.g., Gentner, Loewenstein, \& Thompson, 2003), we sought to improve creative problem solving by activating structured content (relations between objects) in the problem encoding. We developed an alternative to a comparison task: completing a set of sentence frames explaining how pairs of objects in the problem setting relate to one another. In two experiments, we found evidence that participants in the relational encoding group were significantly more likely to solve an insight problem than controls. An important caveat is that the advantage was only found for the easier problems tested - there were no differences for the more difficult problems. We address implications of this work from both theoretical and applied perspectives.
\end{abstract}

\title{
Not Emergence as Emergence: Emergence in Artificial Intelligence and in Philosophy of Mind
}

\author{
Eliska Kvetova \\ University of West Bohemia
}

\begin{abstract}
Hardly any concept is as frequented as emergence in current cognitive science. For many authors it has become solution of eternal psychophysical problem, for many it has been only a mysterious incantation in regard to this. This poster should point out that concept of emergence as well as many other promising concepts has a fundamental problem with determination, definition and usage. The thesis is evident from the title: not emergence as emergence, or better: like emergence, not like emergence. There are many different concepts of emergence. The contribution is based on the belief that the confrontation of artificial intelligence with philosophy of mind, the comparison of these two areas in which emergence occurs very often, could be interesting and could enable to formulate or outline certain tendencies in understanding and using of the concept.
\end{abstract}

\title{
Predictive coding and the Bayesian brain: Intractability hurdles that are yet to be overcome
}

\author{
Johan Kwisthout \\ Donders Institute for Brain, Cognition and Behaviour \\ Iris van Rooij \\ Donders Institute for Brain, Cognition and Behaviour
}

\begin{abstract}
There is a growing body of evidence that the human brain may be organized according to principles of hierarchical predictive coding. A current conjecture in neuroscience is that a brain, organized in this way, can effectively and efficiently perform genuine Bayesian inferences. Given that many forms of cognition seem to be well characterized as Bayesian inferences, this conjecture has great import for cognitive science. It suggests that hierarchical predictive coding may provide a neurally plausible account of how forms of cognition that are modeled as Bayesian inference may be physically implemented in the brain. Yet, the jury is still out on whether or not the conjecture is really true. In this presentation, we demonstrate that each key sub-computation invoked in hierarchical predictive coding potentially hides a computationally intractable problem. We furthermore identify ways in which computational modelers may or may not overcome these 'intractability hurdles.'
\end{abstract}

\title{
What working memory subcomponents are needed in the acquisition of survey knowledge? Evidences from direction estimation and shortcut tasks
}

\author{
Enia Labate \\ University of Padua, Padova, PD, Italy \\ Francesca Pazzaglia \\ University of Padua \\ Mary Hegarty \\ University of California, Santa Barbara
}

\begin{abstract}
Survey spatial representations are map-like mental representations in which directional relationships among landmarks are preserved. Survey representations allow people not only to re-trace routes already experienced, but also to find new routes and shortcuts (Golledge et al., 1999). This study investigated whether and to what extent verbal and spatial working memory (WM) are implicated in the construction of survey representations. We adopted a dual-task paradigm, asking participants to learn a new environment from navigation and, concurrently, to perform either a verbal or a spatial task, assumed to load verbal and spatial WM, respectively. Ninety undergraduates were assigned to one of three groups according to concurrent task condition: articulatory suppression, spatial tapping, or control (no concurrent task). Acquisition of a survey representation was tested by asking participants to perform direction estimations and shortcut tasks. The results supported the involvement of spatial WM in the acquisition of survey knowledge, showing significant differences between the spatial tapping group and control group for the survey measures.
\end{abstract}

\title{
Investigating Elementary and College Students' Development of Numerosity Judgment Using Eye-Tracking Technology
}

\author{
Meng-Lung Lai \\ National Chiayi University, Chiayi, Taiwan, Taiwan \\ Meng-Jung Tsai \\ National Taiwan University of Science and Technology \\ Liang-Chen Chien \\ National Chiayi University \\ Yi-Chen Chen \\ National Chiayi University
}

\begin{abstract}
Numerosity judgment involves determining the number of items, which highly correlates with mathematical achievement. This study investigated age-related differences on numerosity judgment among middle-level elementary students and college students in terms of strategy use and problem-solving efficiency determined by participants' eye movements. Stimuli were grids consisting of \(7 \times 7\) units which were either "on" (yellow block) or "off" (blank). Participants were asked to determine the number of yellow blocks.

Results showed that given energy- and time- consuming, third graders consistently adopted "addition strategy" on larger numerosity trials. For strategy adaptiveness, adults outperformed younger peers, most of whom starting using "subtraction strategy" on critical and larger trials (e.g., 25 and larger numerosities). Regarding efficiency, adults determined numerosities more efficiently than elementary students, because adults' eye fixations were significantly fewer than those of younger groups, while only marginal difference on the number of fixations was observed for the two younger groups.
\end{abstract}

\title{
Motor Resonance Effects Modulated by Perspective: A Role of Linguistic Contrual on Social Action Perception
}

\author{
Donghoon Lee \\ Pusan National University, Busan, South Korea \\ JaRang Kwak \\ Pusan National University, Busan, South Korea \\ YoungIn Lee \\ Pusan National University, Busan, South Korea
}

\begin{abstract}
We investigated motor resonance effect(MRE) during perceiving interactive social action scenes of two people. Perspectives of the social action scenes were manipulated by the voice (active vs. passive) of sentences describing the scenes. In Exp1, subjects' response time for stepping on a pedal was analyzed for investigating MRE during understanding of the scenes where a person is stepping on a foot of the other person. In Exp2, the response time by the lip action on a microphone were analyzed for exploring MRE during perceiving of the scenes where a person is biting an arm of the other person. In results, the MRE in both experiments was significant only when the scenes were described in the active voice. Our results indicate that the motor resonance effect can occur during perception of social events, but it can be modulated by the perspective of the mental construal of the event.
\end{abstract}

\title{
Psych Predicates: Subjectivity and Evidentiality
}

\author{
Chungmin Lee \\ Seoul National University
}

\begin{abstract}
This work investigates why the psych sentence 'She is dizzy' with the third person subject in PRESENT fine, whereas its counterparts in Korean and Japanese are odd unlike the first person subject utterance na-nun ecirew-e (K) 'I am dizzy.' We have no way of knowing if others' internal psych state is such at speech time. We argue that an evidence acquisition event (ee.a) such as I just heard from Mary/I just saw Mary precedes or is accommodated prior to speech time for the third person present psych sentence even in English. The PRESENT realization is a consequence of "double access" sequence of tense interpretation in English, i.e. Mary was dizzy and still is dizzy. A psych predicate requires the 1st person Experiencer's direct perceptual experience of one's own psych state or of individual object as in predicates of personal taste. First-person interoceptive psych judgments of I am dizzy/in pain have "immunity to error through misidentification," unlike de se thoughts.
\end{abstract}

\title{
Neural representations of language switching in early bilinguals: An fMRI study
}

\author{
Miaomei Lei \\ Tokyo Institute of Technology, Tokyo, Japan \\ Hiroyuki Akama \\ Tokyo Institite of Technology,Japan \\ Brian Murphy \\ Carnegie Mellon University
}

\begin{abstract}
To investigate neural activity with respect to language switching, we measured brain activation with functional magnetic resonance imaging (fMRI) while 5 early Korean-Chinese bilinguals performed a covert property generation task with language switching. Forty stimulus photographs of animals and tools were presented as stimuli, accompanied by captions written either in Korean or Chinese. When the stimuli were shown in Korean, participants were asked to do the covert association production task in Chinese ( \(\mathrm{K}>\mathrm{C}\) ), and vice versa ( \(\mathrm{C}>\mathrm{K}\) ). Results from the fixed effect analysis revealed that the \(\mathrm{K}>\mathrm{C}\) condition (Korean as orthographic stimuli and Chinese as semantic execution language) activated primarily left precentral gyrus and left inferior frontal gyrus while the \(\mathrm{C}>\mathrm{K}\) condition activated primarily the region straddling right precuneus and right middle/posterior cingulum. These contrasting activation patterns might support the hypothesis that the neural representations in language switching tasks hinge on the linguistic typology and the cognitive motor control.
\end{abstract}

\title{
Thinking metaphorically in Alzheimer's Disease: priming cognitive domains to understand inferential content of sentences
}

\author{
Jan Leite \\ Universidade Federal da Paraiba \\ Berla Moraes \\ Universidade Federal da Paraiba \\ Mabia Toscano \\ Universidade Federal da Paraiba \\ Danielly Lima \\ Universidade Federal de Campina Grande
}

\begin{abstract}
This paper looks into the effects of visual/linguistic stimuli in the activation of cognitive domains (e.g. image schemas for primary metaphors) which are necessary for the understanding of sentences content by individuals affected by Alzheimer's Disease (AD). AD studies point out to cognitive impairments at early stages of the disease which make comprehension tasks such as abstract inference, and metaphorical reasoning more costly to AD subjects in comparison to other groups of normal aging. We designed a \(3 \times 3\) experiment in which subjects ( AD and control group) were presented with primes of cognitive domains (words, pictures, and ideograms as control) followed by a choice task of metaphorical, literal, and abstract sentences in order to measure time spent to understand sentences in each condition, and frequency of each choice. Our hypothesis is that when the subject is primed with visual rather than linguistic input of a domain, s/he understands more readily and more accurately the metaphorical/inferential content of a linguistic expression, even though literal understanding frequency keeps higher.
\end{abstract}

\title{
Activities of Mirror System Involved In Coordination Game
}

\author{
Guanhong Li \\ Japan Advanced Institute of Science and Technology \\ Takashi Hashimoto \\ Japan Advanced Institute of Science and Technology
}

\section*{Jiro Okuda}

Kyoto Sangyo University

\begin{abstract}
Intention understanding is necessary during the formation of human communication system. Though the linkage between mirror system and intention understanding has been examined by many studies, the evidence of activities of mirror system during the formation of human symbolic communication system, in which no explicit demonstration of actions is involved, are still limited.

We recorded the neural activities with electroencephalography during a coordination game with message passing, which involves formation of symbolic communication system. In this experiment, two subjects were separated into two rooms, thus they could not see any movement of communication partners or hear any sound from the partners' actions. Significant mu rhythm suppression is found over sensorimotor cortex both when the subjects send and receive messages.

This preliminary result indicates that intention understanding involved in message interpretation can also induce the activities of mirror system, even without explicit referring to actions.
\end{abstract}

\title{
A crosslinguistic look at how metaphor, manner, and aspect propel polititians through campaign races
}

\author{
Patricia Lichtenstein \\ University of California, Merced, Merced, CA, US \\ Marcus Perlman \\ University of California, Merced, Merced, CA, United States \\ Teenie Matlock \\ University of California, Merced
}

\begin{abstract}
Before an election, voters are inundated with messages about candidates running for office. Our work examines the influence of metaphor in messages about political races. Of special interest is the role of manner of motion (e.g., slow, fast) and aspect (e.g., perfective, imperfective) in messages that include motion metaphors (e.g., "Candidate A raced/was racing ahead of Candidate B" and "Candidate A inched/was inching ahead of Candidate B"). We discuss results from our experiments with English and Russian speakers. In line with our predictions, manner of motion and aspect were found to interact in interesting ways, and to systematically influence opinions about who is likely to win an election. These novel results have valuable theoretical and practical implications for political communication and how people conceptualize political elections, and expand prior work on framing effects in political talk, especially motion metaphors and grammatical aspect (Fausey \& Matlock, 2010; Matlock, 2012).
\end{abstract}

\title{
What We Move to Moves Us: Biological Rhythmicity Predicts Musical Preferences
}

\author{
Nicholas Jun Hao Tan
}

\author{
National University of Singapore \\ Sarah Shi Hui Wong \\ National University of Singapore \\ \section*{Stephen Wee Hun Lim} \\ National University of Singapore
}

\begin{abstract}
For at least 350 centuries, humans have invented music that offered special aesthetic appeal. Yet, the reasons for these preferences and effects are not understood. Here, we show that listeners prefer music with an underlying rhythmic structure that closely approximates our biological structure. Specifically, listeners preferred music with musical (rhythmic) structures that correspond to biological rhythmicity (motions). This finding, grounded in a straightforward biological framework, provided an intellectual advancement in the long history of thought and experimental work on the basis of musical preferences.
\end{abstract}

\title{
Scepticism: Genuine unbelief or implicit beliefs in the supernatural?
}

\author{
Marjaana Lindeman \\ University of Helsinki, Helsinki, Finland
}

\author{
Annika Svedholm \\ University of Helsinki, Helsinki, Finland
}

\begin{abstract}
We examined whether skeptics hold implicit supernatural beliefs. In study 1, priming by reading a biological or a religious story about death had no effect on skeptics’ afterlife beliefs. In study 2, participants indicated whether they would partake in a (bogus) scientific study that involved visiting a fortune teller and whether they would prefer a fortune teller who predicts positive and negative events or one who predicts only positive events. Believers chose the positive fortune teller more often than skeptics did. Study 3 investigated whether participants' views about the afterlife, other paranormal phenomena, and ontological confusions differ in speeded and non-speeded response conditions. The results were moderated by thinking style: ontological confusions increased in speeded conditions for intuitive skeptics but not for reflective skeptics. The results indicate that skeptics don't hold implicit supernatural beliefs but intuitively thinking skeptics hold ontological confusions predisposing to supernatural beliefs.
\end{abstract}

\title{
Modeling the ecological rationality of decision strategies based on internet statistics
}

\author{
Daniela Link \\ Université de Lausanne \\ \section*{Julian Marewski} \\ Université de Lausanne
}

\begin{abstract}
Memory processes play a major role in many models of decision making. Several fast-and-frugal heuristics assume a sequential search of information in memory (e.g. the take-the-best heuristic). Fast-and-frugal heuristics exploit regularities in the structure of the environment and basic cognitive capacities, such as memory. However, until now only few attempts have been made to relate models of memory and decision making to the structure of information in the environment. The ACT-R architecture provides a quantitative theory about the interplay between the information structure in the environment and the memory system. Based on internet statistics, we use ACT-R to predict people's recognition and knowledge about objects in the world, as well as the associated retrieval time distributions of respective memories. We show how a corresponding model integrating memory and decision processes within ACT-R allows predictions about the ecological rationality of decision strategies that operate on the accessibility of information in memory.
\end{abstract}

\title{
The Roles of the Goal and Internal Models in Motor Representation: A Pragmatic Explanation
}

\author{
Daniel Hsi-wen Liu \\ Providence University
}

\begin{abstract}
Motor representation is understood by Grush (2004) and Pezzulo (2008, 2011) in terms of simulation/emulation: internal models simulate motor effectors and environmental conditions. Motor intentionality, thus, is regarded as based on the standing-for relation.

In their thesis of motor representation, knowing in preparation of motor actions is highlighted but the system's doing when manipulating on-going motor activities is overlooked. That 'system's doing', however, retains an essential role in motor intentionality, a role which accounts for complexities of real environment, animate apparatus to be represented, and goal-oriented nature of motor movements.

Contrasted to Grush and Pezzulo's thesis, my research highlights a pragmatic role of motor representation, pragmatic in sense of explaining how a motor agent can successfully achieve a goal by maintaining motor movements. Internal models, in my account, not only supply emulation but serve to assist the motor system in its goal-achieving activities, resulting in efficient control and flexible movements.
\end{abstract}

\title{
The interaction of the wrap-up effect and uncertainty factors in speech processing
}

\author{
David J. Lobina \\ Universitat Rovira i Virgili \\ Josep Demestre \\ Universitat Rovira i Virgili \\ José E. García-Albea \\ Universitat Rovira i Virgili
}

\begin{abstract}
The click-detection paradigm was employed to probe the load exerted by the parser within simple Spanish sentences. In Experiment 1, three positions at the beginning of clauses were established and results suggest that Ss are better prepared the deeper into a sentence the click is. Experiment 2 ran an ERP experiment to determine whether these RTs were the result of the "uncertainty" Ss may have felt regarding the click position, the idea being that the amplitude of the P300 would correlate with the click positions. In Experiment 3, click positions were moved to the end of clauses to establish if the end of a sentence results in a specific strain on working memory. RTs show that Ss are slower in the first position but the measures even out after that, with the possibility that Ss may have attempted to "wrap it up" in both the second and the third positions.
\end{abstract}

\title{
The role of inflectional suffixes in lexical processing of Greek words
}

\author{
Sofia Loui \\ University of Athens, Athens, Athens, Greece
}

\author{
Athanassios Protopapas \\ University of Athens
}

Eleni Orfanidou
University of Crete

\begin{abstract}
We examined differences between the processing of inflectional versus derivational morphology in visual word recognition in Greek using masked and delayed priming. A lexical decision task to target verbs and nouns preceded by morphologically related primes of the same grammatical class was used to examine inflectional morphology, whereas the same target words preceded by primes of the other grammatical class were used to examine derivational morphology. Greek, a highly inflected language, allows use of words consisting of a stable stem and verb or noun inflectional suffixes, keeping the orthographic and phonological overlap constant across conditions. Both noun and verb targets were significantly primed by the same grammatical class, consistent with inflectional processing. When preceded by primes of different grammatical class, verb but not noun targets showed priming, precluding firm conclusions about derivational morphological processing.
\end{abstract}

\title{
Using LEGO Robotics to Engage Minority Students in Science and Technology
}

\author{
Carol M. Lu \\ Teachers College, Columbia University \\ John B. Black \\ Teachers College, Columbia University \\ Sorachai Kornkasem \\ Teachers College, Columbia University \\ Laura M. Lu \\ Teachers College, Columbia University
}

\begin{abstract}
With the underrepresentation of minority students in science, technology, engineering, and mathematics (STEM) related fields, it is important for us to develop ways to narrow this gap. This study examines the use of LEGO robotics to increase the interest and motivation of children, particularly Hispanic and African-American students. Participants were fifth graders from two low SES schools attending an after-school robotics program. Students were randomly assigned either learning science in the traditional classroom setting or learning science with robotics. The results from this study found that learning with LEGO robotics can increase minority students' interest and motivation in science and technology.
\end{abstract}

\title{
The Affect of Embodied Height in Perceptual Estimations
}

\author{
Jeremy A. Luno
}

University of Memphis

\section*{Rick Tillman}

University of Memphis

\begin{abstract}
Studies have shown that vision perception is pliable and that perceptual estimations can be affected by a variety of factors. It has been shown that perceptions of slope, distances, and heights are subject to the influence of physiological, emotional, and/or social variables. The studies conducted and outlined here investigated the impact of an individual's actual height on estimations of slope and object height in analog settings, as well as pictorial and linguistic stimuli as presented from a nonimmersive desktop computer monitor. Results suggest that without a relative horizon to utilize eye-height scaling an individual will instead estimate the height of objects relative to their own height.
\end{abstract}

\title{
The Less You Know, You Think You Know More; Dunning and Kruger effect in Collective Decision Making
}

\author{
Ali Mahmoodi \\ University of Tehran, Tehran, Tehran, iran \\ Majid Nili ahmadabadi \\ University of Tehran, Tehran, Tehran, iran \\ \section*{Bahador Bahrami} \\ UCL institute of cognitive neuroscience
}

\begin{abstract}
Resolving disagreements by collective decision making requires knowledge about task and others' opinion quality. We tested dyads in a visual discrimination task to first show confidence of their individual decision and in case of disagreement announce their joint decision. Using a Bayesian approach, for each participant we compared the optimal decision rule (i.e. relative reliability of participant's own opinion to that of his partner) to the empirically obtained (fitted) decision rule. The less sensitive observers (i.e. the ones who made poorer individual decisions) were significantly less successful in group decision making compared to their more sensitive partners. These less sensitive observers insisted on their individual decision as the group decision more often than recommended by the optimal decision rule. Our findings extent the previously found DunningKruger effect to social decision making domain: the more incompetent are often less aware of their greater fallibility
\end{abstract}

\title{
The role of semantic content of verbal categories in categorical perception: An ERP study
}

\author{
Martin Maier \\ Humboldt-Universität zu Berlin \\ Philipp Glage \\ Humboldt-Universität zu Berlin \\ Annette Hohlfeld \\ Universität Potsdam \\ Rasha Abdel Rahman \\ Humboldt-Universität zu Berlin
}

\begin{abstract}
Categorical perception describes the phenomenon that visual stimuli can be discriminated more easily when they belong to distinct rather than common linguistic categories. Here we investigate the role of the meaningfulness of linguistic categories in categorical perception. To disentangle the effects of labels and semantic contents of verbal categories we employed a learning paradigm in which participants acquired information about initially unfamiliar objects. Linguistic knowledge was manipulated by labeling object pairs either with the same or different names. Furthermore, the labels could be associated with in-depth knowledge or learned in isolation. Two days after learning, the EEG was recorded while participants performed a lateralized object discrimination task. Verbal labels affected object processing already at about 120 ms , unaffected by semantics, while separate semantic effects were found at about 200 ms , suggesting that the influence of verbal categories on perception may not be modulated by semantic information associated with the categories.
\end{abstract}

\title{
Can we observe frequency-modulated syllable effects in French dyslexic children?
}

\author{
Norbert Maïonchi-Pino \\ Laboratoire de Psychologie Sociale et Cognitive (LAPSCO), Université Blaise Pascal (Clermont-Ferrand \\ 2), France \\ Jean Écalle \\ Laboratoire d'Étude des Mécanismes Cognitifs (ÉMC), Université Lumière (Lyon 2), France \\ Annie Magnan \\ Laboratoire d'Étude des Mécanismes Cognitifs (ÉMC), Université Lumière (Lyon 2), France
}

\begin{abstract}
We report the results of whether syllables are frequency-modulated prelexical units in dyslexic children. Twentytwo French dyslexic children were compared to 44 chronological age-matched and reading-level-matched controls. A syllable compatibility procedure was combined with a visual syllable detection task (Exp. 1) and a visual masked priming paradigm in a lexical decision task (Exp. 2). Dyslexic children exhibited robust frequency-modulated prelexical syllable effects; highfrequency syllables elicited a syllable compatibility effect in both experiments, while low-frequency syllables favored either a CV target length or a CVC prime length effect. The frequency-modulated syllable effects were constant across both experiments following an expected developmental course, especially in a highly feasible task (Exp. 1). However, performance was drastically low in a highly demanding task (Exp. 2), suggesting impaired phonological procedures. We propose that dyslexic children do not have obvious impaired phonological representations but rather delayed or compensated phonological representations with impaired phonological procedures.
\end{abstract}

\title{
Disentangling the cognitive processes underlying the testing effect
}

\author{
Simone Malejka \\ University of Mannheim \\ Edgar Erdfelder \\ University of Mannheim
}

\begin{abstract}
Tests modify memory and can improve memory performance: Practice tests outperform additional study trials as a learning technique when the final memory test is difficult (e.g., delayed in time). This is referred to as the testing effect. Although existing theories propose single mechanisms to underlie this effect, the contributions of different cognitive processes are yet to be dissociated. Because most testing-effect accounts attribute the testing advantage to either encoding, maintenance, or retrieval processes, we propose a multinomial processing-tree model that disentangles the contributions of all three memory processes. By applying this model to testing-effect data, we show that (a) testing memory primarily creates maintenance benefits (i.e., resistance against forgetting) and that (b) the critical interaction of testing vs. study benefits with final-test delay is not driven by different retrieval strengths. Our results thus support maintenance accounts of the testing effect and are difficult to reconcile with retrieval-based explanations.
\end{abstract}

\title{
Russian Validation Study of the International Affective Picture System
}

\author{
Olga Marchenko \\ Center of Experimental Psychology MCUPE, Moscow State Linguistic University
}

\author{
Alexey Vasanov \\ Center of Experimental Psychology MCUPE, Institute of Psychology RAN
}

\begin{abstract}
The aim of this study was to examine cultural universality of the International Affective Picture System (IAPS) on a Russian sample (Lang et al, 2008). One hundred subjects evaluated 300 IAPS pictures according to their valence, arousal and dominance. Affective space determined by valence and arousal dimensions had a similar distribution to the American sample. There were significantly high correlations between North American and Russian ratings of valence, arousal and dominance. Nevertheless comparison of these ratings showed that there are significant differences between North American and Russian valence, arousal and dominance scores. Such differences suggest cultural specificity f situations which induce emotion and provide evidence of cultural factors effect on the affective experience. Lastly, it shows the importance of using local cultural norms for internationally available stimuli in addition to the original stimuli, even in the study of such universal processes as an emotion. Supported by RFH Grant 12-06-12058.
\end{abstract}

\title{
Changes in information search strategy under "dense" hypothesis spaces
}

Doug Markant
New York University

\section*{Todd Gureckis}

New York University

\begin{abstract}
Research on information search has found widespread evidence of a "positive test strategy" (PTS), where people search for predicted outcomes of a focal hypothesis. A recent analysis by Navarro and Perfors (2011) showed that the PTS is consistent with normative models of search under "sparse" hypothesis spaces, where each possible outcome is predicted by a minority of hypotheses (a property shared by common kinds of categories). Despite this justification, learning often involves a transition to a "dense" hypothesis space as information is accumulated, at which point the PTS becomes markedly worse than a strategy of searching for diagnostic information about multiple hypotheses. Using a perceptual search task, our experiment tested whether people independently switch between these strategies based solely on changes in sparsity. The results show that, in general, people continue to make errors consistent with the PTS even when faced with "dense" hypothesis spaces where that strategy is ineffective.
\end{abstract}

\title{
Analyzing Infant Attention, Interaction and Goals
}

\author{
J. Douglas Mastin \\ Tilburg University, Tilburg, Netherlands \\ Paul Vogt \\ Tilburg University \\ Irene Claessens \\ Tilburg University
}

\begin{abstract}
This study revisits categorization of infant engagement used in infant language acquisition research. We provide a novel, component-based analysis of interaction structures - centered on attention, interaction, and goal-oriented behavior. With this approach, we are able to extend the classification of engagement by differentiating two independent categories of engagement that have been overlooked. To verify this new categorization, an experiment was conducted with 84 participants in regard to the presence/absence of goals in naturally observed infant interactions form a Western and a non-industrial community. Results demonstrate that the extended categorization of engagement levels via interaction components and processes is sound because there are no significant differences between sites, and no significant differences between participants and trained coders. These results further support the use of a more extensive categorization of engagement, as well as the use of natural observation data and cultural immersion in language acquisition studies.
\end{abstract}

\title{
Bridging the Gap Between Friends: How Presence Biases Distance Estimation
}

\author{
Justin L. Matthews \\ University of California, Merced \\ Teenie Matlock \\ University of California, Merced
}

\begin{abstract}
How does social presence influence the perception of a physical environment? People think about social relationships in terms of space. For instance, when drawing routes on maps they draw paths closer to friends than strangers (Matthews \& Matlock, 2011). How does the mere presence of a friend alter spatial reasoning?

Here, participants imagined working for an outdoor magazine. They viewed photos of bridges one might encounter while hiking. Some participants were told they preferred crossing bridges last (in a group), while others were told they preferred crossing them first. All participants estimated bridge length. Those who crossed last, and imagined their friends on the other side of the bridge, provided reliably shorter length estimates than those who crossed first and imagined their friends standing immediately behind them.
\end{abstract}

These results provide new insights into how social presence can influence our perceptions of physical environments.

\title{
Causal Dispositions and Transitivity in Causal Chains
}

\author{
Ralf Mayrhofer \\ University of Göttingen \\ Inesa Hildenbrand \\ University of Göttingen \\ Michael Waldmann \\ University of Göttingen
}

\begin{abstract}
A number of popular philosophical and psychological theories that model causality in terms of causal dependency assume that causal relations in chains are transitive. When A causes B, and B causes C it typically follows that A also causes C. In contrast, dispositional theories focus on intrinsic causal properties (i.e., causal dispositions) of the involved participants of causal relations. According to this account, a causal chain is transitive only when A originally has a disposition towards C. We present an experiment that contrasts scenarios with transitive chains (A disposes towards C) and scenarios with intransitive chains (A does not dispose towards C ), according to the dispositional account. In line with dispositional theories of causation, we found a strong dissociation between cause-effect judgments (A causes C ) and probability judgments, \(\mathrm{P}(\mathrm{C}-\mathrm{A})\), in intransitive scenarios but not in transitive scenarios. Across both types of scenarios judgments for single relations did not differ.
\end{abstract}

\title{
Cross-Linguistic Sound to Meaning Mappings in Spatial Relational Terms
}

\author{
Kelly McCormick \\ Emory University \\ Lynne Nygaard \\ Emory University
}

\begin{abstract}
Previous research has identified systematic sound-to-meaning mappings (sound symbolism) in words for object size, shape, and surface lightness. In the present study we investigate whether sound is used systematically in spatial relational terms in 17 languages, and whether native English-speaking individuals can determine the meanings of these terms. In a forced-choice task, participants were asked to choose the correct English translation for each unfamiliar foreign word. All items were spatial terms with either proximal or distal meanings (translations of English 'here', 'there', 'near', 'far', 'this', 'that'). Participants nominated the correct meanings of the words significantly above chance, suggesting that they used sound structure to infer meaning. Acoustic analyses were conducted to examine the relation of acoustic properties to word meaning, as well as to listeners' judgments. The findings suggest that the sound structure of language can be systematically used to mark relational meanings related to space.
\end{abstract}

\title{
From symbols to analog magnitudes: A process model of fraction comparison, with fits to experimental data
}

\author{
Cameron Ross Lloyd McKenzie \\ Stanford, Stanford, CA, USA
}

\author{
James McClelland \\ Stanford University
}

\begin{abstract}
Recent evidence suggests that fractions are represented as visuo-spatial analog magnitudes on a "mental number line" (Schneider and Siegler, 2010). In order to test this theory, we evaluated subjects in two tasks: fraction comparison and number-line marking. Both magnitude difference and response time to mark a fraction on the number line predicted response time to compare the fraction's magnitude to a standard held in memory. This supports the notion that even symbolic numerical cognition is grounded in visuo-spatial processing. We also found a symbolic cost to processing: fractions with more digits were processed more slowly. We propose a model of fraction comparison wherein fractions are processed from symbols into analog magnitudes, then compared to a standard by sequential sampling decision-making. This model provides a superior fit when compared to the log distance model, despite having fewer parameters. Furthermore, it has a direct interpretation in terms of psychological processes.
\end{abstract}

\title{
Cue Competition in Human Associative Learning
}

\author{
Rossy McLaren \\ University of Exeter, Exeter, Devon, United Kingdom \\ Fergal Jones \\ University of Canterbury \\ Fayme Yeates \\ University of Exeter, Exeter, Devon, United Kingdom \\ Ian P. L. McLaren \\ University of Exeter, Exeter, Devon, United Kingdom
}

\begin{abstract}
We used a SRT task in which the preceding two trials of a run of three predicted the third \(2 / 3\) of the time, and added another predictive cue, a colored square, which could also predict the next response required. The question was to what extent would these two cues compete for control of behavior? We assessed this by comparing the dual cue group to a color only control and a sequence only control. Our results showed that the dual group learned about both cues to about the same extent as the individual controls, but that when switched to a test phase where each cue could be assessed independently, the dual group showed a marked decline in performance relative to the controls. We interpret this as evidence for overshadowing occurring between the two predictive cues in the dual group.
\end{abstract}

\title{
Children's Production of Referring Expressions: Effects of Age and Visual Complexity on Under- and Over-Descriptions
}

\author{
Mhairi McMullon \\ Northumbria University
}

\author{
Paul Engelhardt \\ University of East Anglia
}

\begin{abstract}
The Maxim of Quantity states that a speaker should provide enough information for an object to be identified but no more (Grice, 1975). However, research has shown that children tend to produce too little information (under-description) and adults tend to provide too much information (over-description). In this study, we examined the production of referring expressions across the course of development (i.e. 7-18 years). Participants were required to generate a referring expression, and we manipulated the presence or absence of a contrasting object and display complexity. Results showed an age effect on the production of under-descriptions: Young children produced significantly more under-descriptions in more complex arrays compared to older children. In addition, an analysis of voice onset times and eye movements investigated whether there are speed-accuracy tradeoffs associated with visual search and the tendency to include extra information. Conclusions focus on the development of reference mechanisms from childhood to adulthood.
\end{abstract}

\title{
The role of action prediction and control in young children's joint action coordination
}

\author{
Marlene Meyer \\ Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Nijmegen, Gelderland, Netherlands \\ Harold Bekkering \\ Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Nijmegen, Gelderland, Netherlands \\ Rianne Haartsen \\ Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Nijmegen, Gelderland, Netherlands \\ Janny C. Stapel \\ Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Nijmegen, Gelderland, Netherlands \\ Sabine Hunnius \\ Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Nijmegen, Gelderland, Netherlands
}

\begin{abstract}
Coordinating actions with another person can be a challenging endeavor for young children. For smooth coordination of actions with another person two skills are hypothesized to be crucial: action prediction and inhibitory control. In this study, we investigated how developing abilities of action prediction and inhibitory control are related to young children's joint action coordination. Using a simple turn-taking game, \(2 \frac{1}{2}\)-year-olds' joint action coordination performance (timing variability and accuracy) was assessed. Additionally, children's action prediction skills were measured by anticipatory looks using eyetracking. Inhibitory control was tested in an age-appropriate gift delay task. Results show that timing stability during the joint coordination was positively correlated to measures of action prediction. In turn, accuracy in the joint coordination was positively correlated with measures of inhibitory control. These findings indicate a distinctive role of action prediction and inhibitory action control for different coordination qualities (timing variability and accuracy) in young children.
\end{abstract}

\title{
TMS provides evidence of effector-specific premotor contributions to action interpretation
}

\author{
John Michael \\ Copenhagen University, Copenhagen, Denmark
}

\begin{abstract}
Although it is well established that regions of ventral premotor cortex (vPMC) are active during action observation, it remains controversial whether that activation plays a causal role in action interpretation. In the experiment reported here, we used offline continuous theta-burst stimulation (cTBS) to investigate this question. All participants received offline cTBS to the hand area of vPMC in one session and to the lip area in a separate session, and after each session performed an actioninterpretation task in which half of the trials were pantomimed hand actions and half were pantomimed mouth actions. The results show that participants were less accurate in interpreting hand actions after receiving cTBS over the hand area than after receiving stimulation over the lip area, and less accurate at interpreting lip actions after receiving cTBS over the lip area than after receiving stimulation over the hand area. This double dissociation provides evidence in support of the claim that somatotopically organized regions of vPMC contribute causally to action interpretation, and the claim that action production and action interpretation rely on overlapping mechanisms. In more general terms, they reveal an involvement of the motor system in more sophisticated cognitive processes than has hitherto been demonstrated.
\end{abstract}

\title{
Running circles around symbol manipulation in trigonometry
}

\author{
Kevin W. Mickey \\ Stanford University, Stanford, CA, USA \\ James L. McClelland \\ Stanford University, Stanford, CA, USA
}

\begin{abstract}
Recent evidence suggests that we have an intuitive number sense and that visuospatial processes may ground simple mathematical reasoning, but higher level mathematical cognition is often assumed to depend only on the manipulation of symbolic expressions, governed by a set of rules and logical axioms. To assess rule vs. visuospatial thinking in a higher level mathematical domain, we asked undergraduates to solve trigonometry problems and to report their use of rules, mnemonics, and visuospatial representations including the unit circle, right triangle, and sine and cosine waves. Use of the unit circle was reported most commonly, and was associated with better performance, even after controlling for the extent and recency of trigonometry experience. While unit circle users took more time, their performance was robust to problems that rule users tended to fail. Our findings suggest that even higher level mathematical cognition is more than just the manipulation of symbolic expressions.
\end{abstract}

\title{
The Glamorgan Problem Solver (GLAM-PS): A Distributed and Embodied Production System Architecture for Cognition
}

\author{
Gareth Miles \\ University of South Wales
}

\begin{abstract}
The Glamorgan Problem Solver (GLAM-PS), an example of a computationally implemented theory of embodied cognition, is a novel cognitive architecture that has been applied to algebra problem solving, Tower of London problem solving and Stroop Tasks. GLAM-PS is a distributed production system architecture, with all modules playing a role either in perception or action. Each module has its own working memory, production memory and production matching bottleneck. Inter-module communication allows each module to see the working memory of other modules and to coordinate action with activity in other modules. Despite the lack of explicit goal representation the architecture is able to model offline multi-step problem solving in algebra. Action representations are used to hierarchically structure action in Tower of London problem solving, allowing 'subgoaling' of particular disks. Performance on Stroop Tasks highlights the architecture's ability to model controlled behaviour. These domains demonstrate GLAM-PS potential for providing insights into human behaviour.
\end{abstract}

\title{
The Influence of Selective Attention during Memory Retrieval on Subsequent Memory Performance in Young and Older Adults
}

\author{
Jeremy K. Miller \\ Willamette University
}

\author{
Kyle L. Dixon \\ Willamette University
}

Tyler Young
Willamette University

\author{
Marianne Lloyd
}

Seton Hall University

\author{
Ashley Hartman \\ Seton Hall University
}

\begin{abstract}
Much effort has been devoted to exploring the ways in which manipulating attention during encoding influences performance on memory tasks (Craik et al., 1996). However, fewer studies have examined the effects of manipulating attention during memory retrieval. The purpose of this study was to further examine selective attention at retrieval in the young and elderly. Younger and older adult participants were exposed to a study list followed by a recognition test. Participants made memory decisions under selective attention or full attention conditions (Dudukovic et al., 2009). Participants then completed a memory test consisting of items that had been previously tested under full attention, selective attention, and selectively ignored conditions as well as previously untested items. The results suggest that selective attention during retrieval does not result in significant costs for memory decisions regardless of age group. The results are discussed with reference to current theories of aging, attention, and memory.
\end{abstract}

\title{
Nap-related consolidation in learning the grammar and vocabulary of a novel language
}

\author{
Jelena Mirkovic
}

University of York
Gareth Gaskell
University of York

\begin{abstract}
Sleep-associated consolidation plays an important role in language learning (Dumay \& Gaskell, 2007; Gomez et al., 2006). Here we test the hypothesis that greater levels of arbitrariness in the material to be learned are associated with an increased involvement of sleep (Eichenbaum et al., 1999). Two participant groups were trained to equivalent levels on novel words incorporating regularities mirroring a grammatical gender system. After training, participants had a 90-min break filled with either a polysomnographically recorded nap or an awake control task. Subsequently, the nap group outperformed the wake group in recognition and recall of the trained vocabulary items. Both groups showed evidence of generalization of the systematic aspects of the grammar to untrained items, but the nap group outperformed the awake on the grammar test on the trained items. The findings are discussed in the context of the Complementary Learning Systems approach (Kumaran \& McClelland, 2012; McClelland etal. 1995).
\end{abstract}

\title{
How Teacher Support Children's Empathic Understandings toward Characters? -Focusing on the presentations of similar situations-
}

\author{
Satoko Miwa \\ The University of Tokyo, Tokyo, Japan
}

\begin{abstract}
The present study examined how the presentation of similar situation to character's situation would support children's empathic understandings toward character. According to Barnes \& Thagard(1997), people find similar situation of which the target is in, and transfer the emerging emotional state to target to understand his feelings empathically. Therefore, it was expected that teacher's presentation of similar situation to character's situation would scaffold children's understandings toward character's feelings. Sixth graders' 6 moral education lessons were observed and speeches were recorded. The scenes, where similar situations to character's situations were provided by teacher and used by children, were analyzed. As a result of qualitative analysis, the followings were found: 1) Children who had difficulties in understanding characters' feelings were able to imagine it by using presented similar situations. 2) Teacher presented similar situations to emphasize not only "similarity", but "differences" between situations to prompt children's understandings toward characters.
\end{abstract}

\title{
The effects of the harmonization between a word's meaning and its expression style on implicit memory: Differences between typography in vision and prosody in sounds
}

\author{
Kozue Miyashiro \\ University of Tsukuba
}

Etsuko T. Harada
University of Tsukuba

\begin{abstract}
An expression style refers the graphical or phonological style of language. (i.e., letters or phonemes in words). In daily life, typographies or prosodies are often made to harmonize its connotative meanings with words' meanings they convey. To investigate the role of these harmonization, we demonstrated previously that the harmonization between words' meaning and the typography at encoding had effects on implicit memory in the visual word-fragment completion task (WFC), which implied that harmonization facilitated visual processing of presented words. Expecting that harmonization in sound and word's meaning has same effects, we conducted another experiment, which manipulating the harmonization between words' meaning and their prosody in utterance, using an auditory and visual (classical) WFC. The results showed no effects of harmonization. These results suggest that typography in visual processing and prosody in auditory processing had different function in their connotative meanings.
\end{abstract}

\title{
Investigating phonological false memory effects of prior recall and recognition using the Remember-Know paradigm
}

\author{
Mohamed Shan-Rievan Mohamed Salleh \\ National University of Singapore, Singapore, Singapore
}

\author{
Winston Goh
}

National University of Singapore

\begin{abstract}
In this study, the relationship between prior recall and recognition and its effects on false memory was investigated using low and high confusable phonological associates. Participants who had falsely recalled critical words in an earlier phase were less likely to falsely recognize them later. This pattern of results is in contrast with veridical memory, where participants were more likely to recognize a word after correctly recalling it earlier. We further investigated this phenomenon with the remember-know paradigm, and found that recognition of prior recalled words was associated with more remember responses while recognition of falsely recalled critical words was associated with more know responses. These findings suggest that the underlying processes for veridical and false recognition may each involve different aspects of recollection and familiarity for phonological associates.
\end{abstract}

\title{
Learning and task-dependent factors affect production choices
}

\author{
Jessica Montag \\ University of Wisconsin Madison \\ Maryellen MacDonald \\ University of Wisconsin-Madison
}

\begin{abstract}
Production choices are driven by many factors: experience with language statistics, task-or message-specific factors and constraints on human cognition. Understanding how these constraints operate together is crucial for understanding sentence production. Speakers described human and inanimate targets (baby/vase) in scenes showing human agents acting on these targets. The task elicited relative clauses to contrast targets from competitors ("The baby/vase that the woman is carrying/that is being carried by the woman"). Participant utterances were affected by target noun animacy, and this effect persists across languages. This suggests a role for learning; a speaker's lifetime of experience with their language, including structure alternatives afforded by each language, affects production choices. In addition to long-term learning, immediate demands affect production choices. Targets varied in visual salience, which affected structure choices. Effects are consistent with a task-dependent account of visual salience. Immediate and long-term learning motivations for these effects will be discussed.
\end{abstract}

\title{
Worry effects on visuospatial and verbal working memory
}

\author{
Andre Luiz Moreno \\ Federal University of Rio Grande do Sul \\ Juliana Avila deSouza \\ Federal University of Rio Grande do Sul \\ Gustavo Gauer \\ Federal University of Rio Grande do Sul
}

\begin{abstract}
Worry consists of intrusive, repetitive negative thoughts, usually in verbal form, about a future event of uncertain outcome. Excessive occupation of working memory (WM) subsystems by worry-related representations might cause deficits in WM performance and efficiency. Attentional Control Theory (ACT) predicts that worry occupies the central executive, but not the storage components of WM (phonological loop, visuospatial sketchpad). We tested worry effects on visuospatial and verbal WM memory tasks, with and without binding, in 46 Brazilian undergradutes divided into a low-worry (LW, n=21, 11 female) and a high-worry (HW, \(\mathrm{n}=25,23\) female) group. The HW group showed significantly lower accuracy in the verbal tasks and higher reaction times in the visuospatial tasks. Within the HW group, binding caused lower accuracy in the visual task and lower RTs in the verbal task. Results are discussed according to ACT, which predicts differential effects of worry on performance and efficiency accross WM modalities.
\end{abstract}

\title{
An fMRI pilot study: Altruistic Punishment in ingroup and outgroup membership
}

\author{
Rosalba Morese \\ Center for Cognitive Science, Department of Psychology, University of Turin, Italy \\ Daniela Rabellino \\ Center for Cognitive Science, Department of Psychology, University of Turin, Italy \\ Angela Ciaramidaro \\ Dept.of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy, Goethe-University, Frankfurt/M., Germany \\ Giovanna Carrara \\ Department of Neuroradiology, CTO Ospital, Turin, Italy \\ Elena Prodi \\ Department of Neuroradiology, CTO Ospital, Turin, Italy \\ Francesca M Bosco \\ Center for Cognitive Science, Department of Psychology; Neuroscience Institute of Turin, University of Turin, Italy \\ Consuelo Valentini \\ Department of Neuroradiology, CTO Ospital, Turin, Italy \\ Bruno G Bara \\ Center for Cognitive Science, Department of Psychology; Neuroscience Institute of Turin, University of Turin, Italy
}

\begin{abstract}
Altruistic punishment has a central role in cooperation among humans. People punish uncooperative individuals at a cost to themselves, inducing adherence to social norms. Little is known about altruistic punishment comparing ingroup and outgroup context. Using fMRI, we studied the behavior of altruistic punishment during the Third Party Punishment game. This game shows the behavior spending one's own money, without any personal benefit, to punish unfair behavior of Player A who violate cooperation norms to Player B.This behavior may be differently displayed depending on the in-group (Player C and B same nationality) versus out-group (Player C and B different nationality) setting, in favor to one's own group. This attitude is called parochial altruism. Preliminary results showed activation of anterior cingolate cortex involved in conflict between cognitive and emotional motivations in altruistic punishment. Data also show activation of anterior insula that reflects social norm violations during unfair ingroup condition.
\end{abstract}

\title{
Changes in the Cognitive Control Network Associated with Adapting to Task Environment Modifications
}

\author{
Jarrod Moss \\ Mississippi State University \\ Winston Jones \\ Mississippi State University \\ Hao Bai \\ Mississippi State University \\ Stephanie Doane \\ Mississippi State University
}

\begin{abstract}
Participants were trained to perform a multitasking task involving prioritizing the sorting of a set of objects while frequently being interrupted by new objects to sort. After training, during an fMRI session, participants performed both the original task and a version of the task in which the function of a key part of the interface was modified. Analysis focused on the cognitive control network, a set of regions that decrease in activity with increasing levels of task experience. The functional connectivity of the anterior insula with the other regions of the control net work predicted the degree to which participants successfully adapted to this modified task better than other individual difference variables collected. The anterior insula is thought to be part of a salience network, and this salience network may help to inhibit old rules of the task set that must be replaced by new rules.
\end{abstract}

\title{
Does the anticipation of a partner's reaction affect action planning? Spatial action-effect compatibility in a joint task
}

\author{
Romy Müller \\ Technische Universität Dresden \\ Dietrich Kammer \\ Technische Universität Dresden
}

\begin{abstract}
Actions in joint tasks are affected by the way the partner is acting. As action planning is mediated by an anticipation of their sensory consequences, this influence is likely to encompass the expected reactions of others to one's own actions. If so, it should be easier to perform actions that trigger compatible partner reactions instead of incompatible ones. To test this, we used a spatial action-effect compatibility paradigm in a joint task. Subjects moved virtual objects to different locations on a multi-touch table, followed by either a partner's manual reactions on the same or another object, or by automatic visual effects. There was a tendency for faster performance of actions with compatible effects, and no interaction with effect type. However, for the joint condition alone, no reliable compatibility benefit was found. The results highlight several difficulties in applying the basic research on ideomotor action control to more naturalistic, joint tasks.
\end{abstract}

\title{
Causal Sense-Making as Benefit in Foresight, Rather than Bias in Hindsight?
}

\author{
Edward Munnich \\ University of San Francisco \\ Jennifer Milazzo \\ University of San Francisco \\ Jade Stannard \\ University of San Francisco \\ \section*{Katherine Rainford} \\ University of San Francisco
}

\begin{abstract}
Upon reading headlines like "Traffic Fatalities Fell Last Year," people often overestimate how well they would have anticipated changes. In the present study, participants who estimated fatality statistics and listed possible causes before learning true statistics (Foresight) were more surprised than those who listed possible causes only after learning true statistics (Hindsight). Pezzo (2003) linked hindsight bias to causal explanations that minimize initial surprise after an outcome-but to what extent can people build an expectation for alternative causation before learning the true outcome? Before seeing true statistics, a subset of our Foresight participants listed causes for changes in the opposite of their expected direction. This did not reduce surprise for those who learned that statistics had moved in the opposite of expected direction, but the number of opposite causes they provided reliably predicted their second set of estimates. Moreover, participants frequently explained surprising statistics using their earlier alternative causes.
\end{abstract}

\title{
Reciprocal Ascription of Intentions Realized in Robot-human Interaction
}

\author{
Shoji Nagataki \\ Chukyo University \\ Masayoshi Shibata \\ Kanazawa University \\ Takeshi Konno \\ Japan Advanced Institute of Science and Technology \\ Takashi Hashimoto \\ Japan Advanced Institute of Science and Technology \\ Hideki Ohira \\ Nagoya University
}

\begin{abstract}
Abstract: One of the promising strains of humanoid robotics is that which focuses on explicating and reproducing "inner" cognitive functions of humans. This approach is motivated by an ambitious aim of realizing a human-like mind in a robot. In this general context, we have been working on the mechanism of joint attention, the ability of which infants acquire during the earlier stage of development. We already succeeded in constructing a robot which can engage in joint attention activity of an elementary level.

In a more matured stage, however, humans ascribe intentions to each other in joint attention. In order to realize this process in a robot, it is not sufficient for them merely to acquire the ability to follow others' eye direction. Our point is that it is necessary to implement in a robot the relevant inferential mechanism which involves an apparatus for emotion-detection and object-categorization. In our presentation, we will show how this mechanism can work well in our infant-robot.
\end{abstract}

\title{
When irrelevance matters! Effects of distractor-response binding in binary choices under uncertainty
}

\author{
Nadine Nett
}

Universität Trier
Christian Frings
Universität Trier

\begin{abstract}
The distractor-response binding effect states (Frings, Rothermund, \& Wentura, 2007) that distractors appearing on a prime display create an association with the response given. This association is retrieved when, in the probe, the distractor is repeated; the retrieved response can be compatible or incompatible to the currently demanded probe response thereby influencing behavior. We tested if the distractor-response binding effect also occurs in decision making processes under uncertainty \((\mathrm{N}=31)\). Participants had to decide whether two consecutive, imagined patients suffered from either of two diseases. Each decision was based on two cues; one did not discriminate between the two diseases and the other was either strongly or mildly associated with one of the two diseases. The repeating of the irrelevant cue influences decision significantly. Furthermore, we replicated these findings when we varied the strength of the discriminating cue as between-subject factor.
\end{abstract}

\title{
Multitasking Performance: Bound By Task Interference?
}

\author{
Menno Nijboer \\ University of Groningen \\ Niels Taatgen \\ University of Groningen
}

\section*{Hedderik Van Rijn}

University of Groningen

\begin{abstract}
When engaging in concurrent multitasking, it quickly becomes clear that some activities combine well, while others do not. But which factors determine the compatibility of different tasks? We propose that the overlap in cognitive resources causes interference between tasks that limit performance, and that this performance can be predicted using a cognitive model. To test this, we built a model of three tasks, each using a well-defined set of cognitive resources. These tasks were executed separately as well as concurrently and the performance of the model was recorded. Afterwards, we ran an experiment where participants had to perform the same paradigm. Our results show that the model prediction has a good qualitative fit to the participant data: task combinations with more resource overlap had lower performance when performed concurrently, compared to combinations with little or no overlap.
\end{abstract}

\title{
Joint Action with Separate Response Sets for Location Relevant and Irrelevant Spatial Compatibility Tasks
}

\author{
Akio Nishimura \\ Sophia University \\ Chikashi Michimata \\ Sophia University
}

\begin{abstract}
When two individuals engage in their tasks with a common stimulus display, the partner's presence, response, task, and/or focus of attention often affects one's own task performance. The present study investigates how the partner's task affects one's own task performance when one of two adjacently sitting participants engages in location-relevant task and the other in location-irrelevant task, with a common stimulus display and separate response sets. The target appeared at left or right. One participant pressed a left or right button according to the color of the target when it was green or red. The other participant pressed a left or right button of another response set according to the location of the target when it was white. Although the spatial compatibility effects were observed within each task, no cross-task (i.e., cross-participant) interference was observed. Results are discussed in terms of what is co-represented in joint task settings.
\end{abstract}

\title{
Temporal coordination patterns of performer and audience in vaudeville settings
}

\author{
Ryota Nomura \\ The University of Tokyo \\ Takeshi Okada \\ The University of Tokyo
}

\begin{abstract}
Story-telling performers often rely on their audience's smiles, sounds of laughter, body movements, and other qualitative observations to gauge whether their performances are appreciated by the audience. The current study aims to capture the temporal patterns between a performer and his or her audience. A professional rakugo story-teller performed live in front of 20 audience members aged 16 to \(67(M=40.6, S D=16.4)\) in a laboratory. Videotaped performances were categorized by a computer-aided coding system, and the audience's reactions were quantified using the face-tracking and background subtraction computer program. Results demonstrated performer-audience correlations only in a particular frequency band. While the audience often smiled in response to incongruent lines and interpreted gestures, the performer sometimes delivered the points only after audience-initiated smiles and movements. This dynamic co-creation may offer suggestions for a variety of orators who speak regularly in front of audiences.
\end{abstract}

\title{
Strategy flexibility in algebraic problem solving: The influence of feedback and learners' characteristics
}

\author{
Daniela Nussbaumer \\ ETH Zurich \\ Michael Schneider \\ University of Trier
}

Elsbeth Stern
ETH Zurich

\begin{abstract}
The ability to choose problem solving strategies flexibly and adaptively is an important part of expertise. However, it is unclear how simple forms of problem solving practice affect flexibility. We investigate to what extent flexibility in strategy use is dependent on learners' characteristics as intelligence, working memory and prior knowledge and to what extent these individual differences influence how a student can exploit a learning situation.

In a microgenetic design with 24 trials of a mathematical problem solving task, we found that ninth-graders adaptivity of strategy choices increased linearly during practice without feedback and that feedback on strategy adaptivity facilitated the process. Adapting strategy choices to problem types led to shorter solution paths, higher solution rates, and higher speed. This is true independently of students' intelligence or working memory capacity. However, prior knowledge was a predictor of adaptivity, thus leading to a faster development of adaptivity.
\end{abstract}

\title{
Extended Methods to Dynamically Estimate Emphasizing Points for Group Decision Making and the Evaluation
}

\author{
Yoshimasa Ohmoto \\ Kyoto University
}

\author{
Misao Kataoka \\ Kyoto University
}

Toyoaki Nishida
Kyoto University

\begin{abstract}
The purpose of this study is to extend a method of dynamic estimation of emphasizing points (DEEP) to estimate the emphasizing points of a group. A preliminary experiment investigated whether the interaction process would differ depending on the interaction style used, avoiding conflict or expressing opinions. The difference is caused by confidence in their opinions and their commitment to the decision making. We also proposed two extended methods corresponding to the different interaction processes. We then conducted an experiment to evaluate the methods using Embodied Conversational Agents. In conclusion, these proposed methods accurately estimate proposals and satisfy participants in the appropriate group. We could also observe that the participants carefully looked to their partner for their reaction in "avoiding conflict" group and that the participants concentrated on execution of the task in "expressing opinions" group.
\end{abstract}

\title{
The effect of spatial layout of objects on special memory through free searching tasks
}

\author{
Kayoko Ohtsu \\ Waseda University
}

\begin{abstract}
The layout of targets is known to affect spatial memory in association with special frames of reference. However, most of the previous studies have been conducted on the learning from specific viewpoints. The present study examines whether the layout, as the structure of special reference points, has any effect on special memory when learning from all sides while walking around targets. In the experiment, participants explored a virtual circular room under the one of two conditions. The objects were arranged like spots on dice. In the Square Array, 4 objects were set in a square shape, while in the Node Array, one more object was added at the center. After the learning phase, they judged the directions of the targets from several imaginary positions of the room. The results suggest that the participants in the Node Array made quicker and more accurate judgments than those in the Square Array.
\end{abstract}

\title{
How expectation of context sharing with audience causes effects on writing arguments.
}

\author{
Ryosuke Onoda \\ The University of Tokyo
}

\begin{abstract}
This study examined how the students' expectation of context sharing with audience causes effects on their writing arguments. Fourth-grade students \((\mathrm{N}=30)\) from elementary school in Japan were assigned to do two writing argument tasks. In each task, students made an argument and persuaded different audiences (transfer student and old friend) at random. The contents of written arguments were categorized into "claims", "evidence", and "reasoning". Moreover, difficulty of each audience's persuasion and its reasons were asked in questionnaire. The analysis suggested followings. (1) Students generated more "reasoning" in transfer student condition than old friend. (2) Students evaluated persuasion of transfer student was more difficult than old friend. Students judged that the transfer students do not share much context with them, so they tried to persuade by giving more information. The students read the degree of context sharing and changed the contents of writing with changing audiences.
\end{abstract}

\title{
Cognitive Residues of Similarity: After-Effects of Similarity Computations in Visual Search
}

\author{
Stephanie O'Toole DCU \\ Mark T Keane \\ UCD
}

\begin{abstract}
What are the "cognitive after-effects" of making a similarity judgement? What, cognitively, is left behind and what effect might these residues have on subsequent processing? In this paper, we probe for such after-effects using a visual search task, performed after a task in which pictures of real-world objects were compared. So, target objects were first presented in a comparison task (e.g., rate the similarity of this object to another) thus, presumably, modifying some of their features before asking people to visually search for the same object in complex scenes (with distractors and camouflaged backgrounds). As visual search is known to be influenced by the features of target objects, then any after-effects of the comparison task should be revealed in subsequent visual searches. Results showed that when people previously rated an object as being high on a scale (e.g., colour similarity or general similarity) then visual search is inhibited (slower RTs and more saccades in eye-tracking) relative to an object being rated as low in the same scale. There was also some evidence that different comparison tasks (e.g., compare on colour or compare on general similarity) have differential effects on visual search.
\end{abstract}

\title{
When is it rational to rely on heuristics?
}

\author{
Paula Parpart \\ University College London \\ Matt Jones \\ University of Colorado Boulder \\ Bradley Love \\ University College London
}

\begin{abstract}
There have been two distinct notions of heuristics, i.e., Kahneman and Tversky's (1974) heuristics as biased approximations to rational inference, and Gigerenzer et al.'s (1999) idea of smart and adaptive heuristics. Despite the conceptual differences, we provide evidence that heuristics can be seen as approximations to a rational account which is at its core adaptive. In a large cross-validation, we demonstrate that a regularized regression model (from machine learning) with a penalty noise parameter could outperform both heuristics and simple linear regression. Importantly, the penalized regression with an L2-norm could be approximated by tallying, whereas the L1-norm was approximated by take-the-best. Results indicate that the penalized regression treats both heuristics and linear regression as extreme cases of the model. The research implies a common rational basis for heuristics and integrative strategies, suggesting that the relation need not be adversarial. Implications for reconciling adaptive and irrational views of heuristics are discussed.
\end{abstract}

\title{
"This could be a new approach against the flu..." - References to tentativeness of scientific knowledge and their impact on decision processes
}

\author{
Elisabeth Paus \\ University of Muenster, Muenster, Germany \\ Regina Jucks \\ University of Muenster, Muenster, Germany
}

\begin{abstract}
People often rely on popular scientific information when seeking advice for health-related issues. In these cases, further usage of such information should be influenced by its presentation. With \(\mathrm{N}=157\) students, we examined how referring to the tentativeness of health information by using hedges and pointing to the origin of scientific knowledge in science-related texts impacts processes of decision-making. We found that decisions were easier to make when there was no indication given. Furthermore, participants' further use of text-related information was more likely when hedges were used. In contrast, individuals rather relied on their own knowledge when there were no linguistic markers of tentativeness. Additionally, participants' decisions were more in favor of the direction implied in the texts when no indication of the sources of the science-related information was given. However, no effect of experimental manipulation on the confidence of the decisions exists. Finally, we discuss how the presentation of information may contribute to engaging in critical and elaborated processing of scientific information.
\end{abstract}

\title{
Beyond the politeness tightrope: Message design for multiple social goals
}

\author{
David Pautler \\ Institute for High Performance Computing, A*STAR
}

\begin{abstract}
In linguistic pragmatics and social anthropology, several influential researchers believe that politeness is essential for maintaining social order by way of disarming potential aggressiveness [Goffman 1967; Brown \& Levinson 1987; Gumpers 1987]. In one of the most detailed of these theories, Brown and Levinson's, speakers pursue a single goal (e.g. getting the hearer to stop doing something) by using a mental model of the hearer to select a position on a one-dimensional spectrum of strategies that identifies the best balance between achieving the speaker's practical goal while avoiding offense to the hearer (as might occur from a purely brusque request). But are speakers actually limited to this one-dimensional spectrum of strategies? And given that people often pursue more than one goal at once, how might they do so in such a simplistic model of polite communication? I describe and evaluate a computational model that generates strategies for multiple simultaneous goals.
\end{abstract}

\title{
Recognition of Common Object-Based Categories Found in Toddler's Everyday Object Naming Contexts
}

\author{
Alredo Pereira \\ University of Minho
}

\author{
Linda Smith \\ Indiana University
}

\begin{abstract}
Previously, we investigated the distribution of instances of early-learned object-based categories in toddler's realistic everyday learning episodes; we found important differences in terms of frequency and variability (3D vs. 2D; real object vs. realistic toy vs. simple shape).

Using a picture book task we tested 24-36 month olds' recognition of these categories in four conditions: Realistic; Features (only parts of the photo visible); Silhouettes; and Geons (a shape caricature version made with only 3-4 parts and no color or texture). Results show similar recognition for all Realistic and Silhouette versions; Geons were lower than the first two; and Features had the lowest recognition rate. Critically, categories with the highest variability in our previous study were readily recognized by Features but difficult to recognize in Geon version. These results suggest that abstracting global shape is influenced by the specific trajectory of experienced exemplars.
\end{abstract}

\title{
Complex functional brain network properties extracted from EEG reveal different processing patterns in late second-language learners as compared to native-speakers when performing a morphosyntactic task
}

\author{
Alejandro Pérez \\ BCBL, Donostia, Gipuzkoa, Spain \\ Margaret Guillon Dowens \\ Division of English, University of Nottingham, Ningbo, China \\ Nicola Molinaro \\ BCBL, Donostia, Gipuzkoa, Spain \\ Yasser Iturria-Medina \\ Neuroimaging Department, Cuban Neuroscience Center, La Habana, Cuba \\ Paulo Barraza \\ Centro de Investigación Avanzada en Educación, CIAE, Universidad de Chile, Santiago 8330014, Chile \\ Manuel Carreiras \\ BCBL, Donostia, Gipuzkoa, Spain
}

\begin{abstract}
Complex network analysis is applied to study late second language (L2) acquisition using highly proficient late L2 learners when compared to monolinguals performing a morphosyntactic task. Specifically, we assess for (dis)similar topological properties of the functional networks associated with a gender mismatch condition between article and noun at the beginning of a sentence in a Spanish monolingual group and a group of late learners of Spanish whose native language is English (which does not encode gender as a grammatical category). Our results suggest that the detection of incorrect grammatical gender agreement in Spanish recruits the neural networks that subserve the cognitive processing differently in each group. This result provides insight into the functional cooperation and interactions of brain areas while processing an L2 trait not present in L1.
\end{abstract}

\title{
Vocal Charades: The Emergence of Conventions in Vocal Communication
}

\author{
Marcus Perlman \\ University of California, Merced, Merced, CA, United States \\ Rick Dale \\ University of California, Merced \\ Gary Lupyan \\ University of Wisconsin Madison
}

\begin{abstract}
Human communication compresses a massive amount of conceptual structure into conventionalized speech patterns. How do these patterns emerge? Experimental semiotics offers new empirical techniques to address such questions (Fay et al., 2008; Galantucci, 2005; Garrod et al., 2007; Kirby et al., 2008). We devised an experiment to observe the emergence of communicative conventions. The study explores the development of vocal communication systems through an iterative Vocal Charades game. The game is played by two players each given a stack of 12 cards. Written on each card is a word from six antonymic pairs (rough/smooth, bad/good, etc.). Over ten rounds, players took turns "vocalizing" the meaning of their words without using language or gestures. Analyses reveal how sounds conventionalize, leading to stereotyped forms and improved guessing. The conventional forms that develop are predicted by iconic correspondences with meaning, but also exhibit more arbitrary features that distinguish semantically similar words.
\end{abstract}

\title{
Perception of collaboration in joint musical performances
}

\author{
Ana Pesquita \\ Department of Psychology, University of British Columbia \\ Timothy Corlis \\ Department of Music, University of British Columbia \\ James T. Enns \\ Department of Psychology, University of British Columbia
}

\begin{abstract}
Humans are exquisitely sensitive to social interactions. This study explored whether this extends to interactions in music performance. Jazz musicians are fluent at working together to produce music that is more than the sum of its parts, and listeners claim anecdotally to hear when musicians are 'in the groove'. We employed jazz-standard duets varying in the opportunity for collaboration (two-way, one-way, none), to test listeners' perception of collaboration. In experiment 1, participants rated random selections from these recordings in the dimensions of synergy, creativity, emotionality, and engagement. Results showed considerable sensitivity to collaboration, with sensitivity varying both with social intelligence and musical training of the participant. In experiment 2, participants made explicit judgments of whether the selections involved collaboration, with the results showing they could not. We conclude that the degree of collaboration in joint musical performances influences the implicit experience of listeners, but is not accessible for explicit judgments.
\end{abstract}

\title{
I did not write you this before, did I? Destination Memory in Computer-Mediated Communication
}

\author{
Stephanie Pieschl \\ Westfälische Wilhelms-Universität Münster \\ Verena Karwinkel \\ Westfälische Wilhelms-Universität Münster
}

\begin{abstract}
In face-to-face communication people have problems recalling which facts they have disclosed to whom. We explore if such destination memory (DM) problems also exist in computer-mediated communication (CMC). Participants ( \(\mathrm{N}=64\) ) disclosed 50 pieces of personal information to 50 fictitious partners in a sham Skype environment in two conditions, one-sided "telling" and communicative "turn-taking". Their recall for facts, faces, and fact-face-pairs was measured as dependent variables. ANOVA results show a significant main effect of type of recall \((\mathrm{F}(2,61)=222.47, \mathrm{p}<.001)\) : recall for facts was better than for faces or fact-face-pairs, indicating that DM problems emerge when faces are previously unknown. Additionally, we found a significant main effect of conditions \((\mathrm{F}(1,62)=6.75, \mathrm{p}=.012)\) : contrary to expectations recall in the "telling" condition was best. These interesting findings and those of a second study will be discussed considering the specific characteristics of CMC.
\end{abstract}

\title{
Structural Alignment in Young Children's Shape Categorization: Different Roles for Learning from Comparisons and Contrasts
}

\author{
Raedy Ping \\ University of Chicago \\ Micah Goldwater \\ Northwestern University \\ Bryan Matlen \\ Carnegie Mellon University
}

\author{
Linsey Smith
}

Northwestern University

\author{
Susan Levine \\ University of Chicago \\ Dedre Gentner \\ Northwestern University
}

\begin{abstract}
Creating correct shape categories requires young children to overcome reliance on perceptual features and surface similarity as cues for membership and to use abstract rules as bases for categorization (e.g., triangles are enclosed shapes with three sides). Relying on surface similarity, children often wrongly include non-triangles that resemble familiar triangles in the triangle category, and exclude unfamiliar actual triangles. This study attempted to improve preschool-age children's triangle categorizations by presenting structurally aligned comparisons that either shared a common structure or that highlighted a contrasting structure. Across both types of comparisons, the exemplars were either highly or lowly superficially similar (both variables manipulated between subjects). We hypothesized that low-similarity common-structure comparisons would support extension of the triangle category beyond prototypical exemplars, while high-similarity contrastive alignments would highlight category boundaries. Preliminary results suggested that improvement in categorization accuracy was primarily driven by a reduction in erroneous identification of a non-triangle as a triangle.
\end{abstract}

\title{
Fixation on Failure: Failing to Solve a Problem Hinders Subsequent Problem-Solving Ability
}

\author{
Vencislav Popov \\ New Bulgarian University
}

\begin{abstract}
This study explores how previous experience affects current performance. Despite the vast amount of research on transfer and learning effect in problem-solving, as far as we know little to no work has been done on how failure to solve a problem affects subsequent problem-solving ability. The current experiment explores this issue and the role of working memory in the process. Two variables were manipulated - the experienced success or failure on a single multiplication problem and the amount of working memory resources required by the following addition problem. Results show negative impact of failure on subsequent process-time: the problem-solving time for the addition problem was higher for participants who failed to solve the multiplication problem. There was no interaction between experienced performance and the WM resources required by the subsequent problem. These results suggest that researchers must probably be careful to avoid fixation on failure as a confounding variable.
\end{abstract}

\title{
Effects of Gestures while Studying Language Animations
}

\author{
Lysanne Post \\ Erasmus University Rotterdam, Rotterdam, The Netherlands \\ Tamara van Gog \\ Erasmus University Rotterdam, Rotterdam, The Netherlands \\ Fred Paas \\ Erasmus University Rotterdam, Rotterdam, The Netherlands \\ Rolf Zwaan \\ Erasmus University Rotterdam, Rotterdam, The Netherlands
}

\begin{abstract}
Two experiments will be presented. Experiment 1 examined whether simultaneously observing and making gestures while studying animations would lighten cognitive load and facilitate the acquisition of grammatical rules, as would be predicted by theories of embodied cognition and cognitive load. However, results showed that children in the gesturing condition performed worse on a subsequent test than children in the control condition. This was particularly true for children with low levels of general language skills, whereas children with high language skills experienced no detrimental effects of gesturing. Because simultaneously observing and making gestures hampered learning, the question is whether only observing gestures would be effective. Moreover, because it is still unclear whether seeing animated sentence transformations has a positive effect on learning, Experiment 2 compares the effects of three instructional conditions for both low and high-ability learners: static pictures, animation without observing gestures, and animation with observing gestures.
\end{abstract}

\title{
How Does Eye Gaze Affect Vocal Imitation in Autism?
}

\author{
Marie Postma-Nilsenová \\ Tilburg University \\ Martijn Balsters \\ Tilburg University \\ Mariska van Kastel \\ Tilburg University
}

\begin{abstract}
Earlier studies showed an important role of vocal imitation in social interactions. At least some kinds of social imitative behavior, including gesture and pitch adaptation, appear to be triggered by direct eye gaze (Wang, Ramsey \& Hamilton, 2011). Past research indicated that autistic individuals may have difficulties with both shared gaze focus and joint attention. Therefore, we might expect that their vocal imitative behavior would differ from the behavior of TD individuals. In our study, we explored vocal imitation by autistic speakers in an experiment with stimuli eliciting eye contact (shared focus), joint attention or a disruption of eye contact.
\end{abstract}

\title{
Semantic Intuitions in Statistical Causal Reasoning
}

\author{
Benjamin Quack \\ University of Göttingen
}

\author{
Ralf Mayrhofer
}

University of Göttingen

\author{
Michael Waldmann \\ University of Göttingen
}

\begin{abstract}
In psychology, philosophy, and linguistics there has been a debate about two competing frameworks of causal reasoning. Dependency theories, especially causal Bayes nets, focus on causally motivated statistical or counterfactual dependencies between events (causes and effects). In contrast, force dynamic theories implement causation as arising (deterministically) from force interactions involving agents impinging on the prior tendencies of patients. To date force dynamic theories have primarily focused on the representation of different semantic causal concepts in scene descriptions. Our goal is to bring the two competing frameworks together. We will present a model that implements the interaction between agents and patients in terms of probabilistic forces. We have tested this new model in an experiment in which we tested how contingency information interacts with the assumptions about intrinsic tendencies of patients in people's usage of semantic causal concepts (e.g., CAUSE, PREVENT, HINDER, HELP, ALLOW, and ENABLE).
\end{abstract}

\title{
Simulating the N400 ERP component as semantic network error: Insights from a feature-based connectionist attractor model of word meaning
}

\author{
Milena Rabovsky \\ Humboldt University at Berlin
}

\author{
Ken McRae \\ University of Western Ontario
}

\begin{abstract}
The N400 ERP component is widely used in research on language and semantics, but the specific underlying mechanisms are currently unclear. We explored the mechanisms underlying the N 400 by examining how a connectionist semantic network's performance measures covary with N 400 amplitudes. We simulated six N400 effects obtained in empirical research. Network error was consistently in the same direction as N400 amplitudes, namely smaller for high frequency words, words with few semantic features, semantically related targets and repeated words. Furthermore, the repetition-induced decrease was stronger for low frequency words and words with many features. In contrast, semantic activation corresponded less well with the N400, and instead seemed related to lexical decision performance. Our results suggest an interesting relation between N400 amplitudes and semantic network error. In psychological terms, network error has been conceptualized as implicit prediction error. Thus, N400 amplitudes may reflect implicit prediction error in semantic memory (McClelland, 1994).
\end{abstract}

\title{
Scaling-up Perception-Action Links: Evidence from Synchronization with Individual and Joint Action
}

\author{
Veronica Ramenzoni \\ Donders Institute for Cognition
}

\author{
Natalie Sebanz \\ Central European University
}

\author{
Günther Knoblich
}

Central European University

\begin{abstract}
How do we map joint actions we participate in onto joint actions we observe others performing, such as when a couple dancing tango observes another couple dancing tango? We investigated this question using a task where participants were instructed to perform individual or joint movements in synchrony with individual or joint movements observed on a computer screen. The observed movements started slowly and then continuously increased in tempo (from 1.75 Hz to 3 Hz ). The results showed that, with regard to spatial parameters, joint performance was more accurate when observing joint performance than when observing individual performance. Individual performance was more accurate when observing individual action than when observing joint action. There were no systematic differences with regard to timing parameters. These results suggest that mechanisms of temporal coordination may be less susceptible to differences between individual and joint action than mechanisms of spatial matching.
\end{abstract}

\title{
Cave-like Environments Facilitate Magical Thinking
}

\author{
Lillian Rigoli \\ University of California, Merced, San Diego, CA, United States \\ Holley Moyes \\ University Of California, Merced \\ Stephanie Huette \\ University Of California, Merced \\ Michael Spivey \\ University of California, Merced
}

\begin{abstract}
The cognitive mechanisms that underlie beliefs in the supernatural are not well understood. In an attempt to understand this phenomenon, we hypothesized that the cave environment "affords" such usage. In our experiment, 52 participants completed a survey about their metaphysical beliefs inside a room providing a great deal of natural light, while another 52 participants completed the same survey in a dark and windowless room. The survey asked participants to rate their beliefs in a variety of supernatural phenomena and also included multiple-choice questions which described bizarre scenarios that could be explained either scientifically or supernaturally. The supernatural responses to both the rating-scale questions and to the multiple-choice questions were significantly higher in the cave-like condition. These results are consistent with anthropological claims that the dark zones of caves may have played a role in the environment of evolutionary adaptedness that contributed to humans' tendency toward magical thinking.
\end{abstract}

\title{
Approximating the Value Function in the Actor Critic Architecture using the Temporal Dynamics of Spiking Neural Networks
}

\author{
Jeffrey Rodny \\ University of California, Merced \\ David Noelle \\ University of California, Merced
}

\begin{abstract}
The human ability to learn from sparse rewards has been modeled with the temporal difference learning mechanism, using an actor-critic architecture (Montague, Dayan, \& Sejnowski, 1996). These models incorporate an "adaptive critic" which learns a "value function": a mapping from the learner's current situation to expected future reward. In complex environments, a "value function approximator" (VFA) must be implemented to allow generalization between similar situations. While some implementations of VFAs have been successful (Tesauro, 1992), this approach does not consistently converge to a solution (Boyan and Moore, 1995). With the goal of developing a general and reliable VFA mechanism, capturing human level learning performance, we have explored the use of spiking neural networks, including liquid state machines, as a technique for VFA learning in complex environments. We report on simulations demonstrating the benefits and pitfalls of using the temporal dynamics of neural spikes to encode the learner's state.
\end{abstract}

\title{
Distractor frequency effects in language production: An ERP study
}

\author{
Lana Rohr \\ Humboldt-Universität zu Berlin, Berlin, Germany \\ Rasha Abdel Rahman \\ Humboldt-Universität zu Berlin, Berlin, Germany
}

\begin{abstract}
In the picture-word interference paradigm low-frequency words have been shown to induce longer naming times, and thus more interference, than high-frequency words. In this study we used event-related potentials (ERPs) to explore the time course and locus of frequency effects within the language production system. Furthermore, we tested whether frequency effects are related to non-lexical variables such as valence and arousal. We presented pictures and superimposed high and lowfrequency distractors and additionally varied the emotional valence and arousal of the distractor words. The effects of distractor frequency in naming times and ERPs - starting at about 300 ms - were modulated by arousal, suggesting that non-lexical mechanisms can modulate distractor effects in the picture-word interference paradigm.
\end{abstract}

\title{
Bayesian Word Learning in Multiple Languages
}

\author{
Sebastian Rolotti \\ Pennsylvania State University
}

Benjamin Zinszer
Pennsylvania State University

\author{
Elizabeth Carlson
}

Pennsylvania State University

\section*{Ping Li}

Pennsylvania State University

\begin{abstract}
Infant language learners are faced with the difficult inductive problem of determining how new words map to novel or known objects in their environment. Bayesian inference models have been successful at using the sparse information available in natural child-directed speech to build candidate lexicons and infer speakers' referential intentions. We begin by showing that when a Bayesian model is optimized for monolingual input (Frank et al., 2009), the model does not sufficiently handle bilingual input, especially as referential ambiguity increases. Here we propose an extended Bayesian model that approximates infants' mutual exclusivity bias to support the differential demands of monolingual and bilingual learning situations. The extended model is assessed using corpora of real child-directed speech, showing that performance can be optimized for both monolingual and bilingual contexts. We show that including both monolingual and bilingual demands in model optimization yields significantly different results than when only one context is considered.

Frank, M. C., Goodman, N. D., \& Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early crosssituational word learning. Psychological Science, 1-8.
\end{abstract}

\title{
Producing ambiguous messages
}

\author{
Sebastian Rose \\ Humboldt-Universität zu Berlin, Berlin, Germany \\ Katharina Spalek \\ Humboldt-Universität zu Berlin, Berlin, Germany \\ Rasha Abdel Rahman \\ Humboldt-Universität zu Berlin, Berlin, Germany
}

\begin{abstract}
Recent evidence suggests that lexical-semantic activation spread, including the formation of ad-hoc relations, can be dynamically shaped by contextual factors (Abdel Rahman \& Melinger, 2011). In this study we investigated whether cognitive processing modes can affect lexical-semantic activation during single word production. Specifically, we tested whether prior processing of linguistic ambiguities, presented in the form of puns, has an influence on the co-activation of unrelated meanings of homophones in a subsequent language production task. In a picture-word interference paradigm with word distractors that were semantically related or unrelated to the non-depicted meanings of homophones we found facilitation induced by related words only when participants listened to puns before object naming, but not when they heard jokes with unambiguous linguistic stimuli. This finding suggests that a cognitive mode of ambiguity processing can induce the activation of ambiguous messages during speech planning.
\end{abstract}

\title{
Explanatory Reasoning in Causal-based Categorization
}

\author{
Anselm Rothe \\ University of Göttingen \\ Ralf Mayrhofer \\ University of Göttingen
}

\begin{abstract}
Research on causal-based categorization focuses on how people categorize exemplars that have causally linked features. One prominent account, the generative model (Rehder, 2003) models membership judgments as a function of an exemplar's likelihood being generated by the category's causal model. In contrast, Mayrhofer and Rothe (2012) found that the explanatory role of the causal model strongly influences membership judgments - indicating the importance of explanatory reasoning processes and that in such tasks people might be guided by explanatory goodness (i.e., how well an exemplar's membership can be explained in the light of the category's causal relations). However, the evidence for this claim was quite indirect so far. In the present categorization study, we collected judgments about category membership, frequency, and explanatory goodness. In contrast to the predictions of the generative model, membership ratings are far better resembled by ratings of explanatory goodness than by subjects' estimations of exemplar likelihood.
\end{abstract}

\title{
Learning in a different way: Interaction gestures influence category learning on multi-touch-tables
}

\author{
Susana Ruiz Fernández \\ Leibniz Knowledge Media Research Center \\ Birgit Imhof \\ Leibniz Knowledge Media Research Center \\ Julia Kranz \\ Leibniz Knowledge Media Research Center \\ Stephan Schwan \\ Leibniz Knowledge Media Research Center \\ Barbara Kaup \\ Psychological Institut University of Tübingen \\ Peter Gerjets \\ Leibniz Knowledge Media Research Center
}

\begin{abstract}
This study investigated the influence of interaction gestures on learning using multi-touch-tables. During a learning phase, participants learned how to categorize Renaissance and Baroque paintings either by moving them over the display of the multi-touch-table or by tapping a marked field on the display representing an art epoch. In the testing phase, participants had to categorize paintings known from the learning phase and new paintings in a forced-choice task. The Gestural Conceptual Mapping approach argues that congruency between gesture and mental processes enhances learning. Due to the discrete nature of the categorization process, learning should benefit from a discrete tapping gesture. The Reality-based Interaction approach argues that gestures should map with interaction experiences in the real world to promote learning. Moving objects to places that represent categories can be considered similar to the real world experience of sorting objects and should facilitate category learning. The results support the latter approach.
\end{abstract}

\title{
Social interaction and group dynamics of virtual cooperative agents
}

\author{
Alexei V. Samsonovich \\ George Mason University
}

\begin{abstract}
The challenge of integration of virtual agents and co-robots into the human society requires for these future artifacts to become human-compatible, in addition to being generally intelligent and capable of providing specific expertise. Specifically, a human-compatible agent should be able to induce in humans a sense of co-presence of a potentially equal mind capable of mutual understanding and human-like learning, long-term personal relationships and generous initiative. Achieving these qualities requires social emotional intelligence. In order to be able to implement equivalents of social emotions in artifacts, it is vital to better understand social interactions in small groups based on cognitive architectures like eBICA (Samsonovich, 2013). The present work makes one step further in this direction, continuing the development and computational exploration of the new framework for emotional cognitive architectures. In the focus are relationships of trust and mutual respect, leadership and allegiance, and related to them social and complex emotions.
\end{abstract}

\title{
Dimensional experience induces attention shifting in the dimensional change card sort (DCCS) task
}

\author{
Larissa Samuelson \\ University of Iowa \\ Sammy Perone \\ University of Iowa \\ Stephen Molitor \\ University of Iowa
}

\author{
Aaron Buss
}

University of Iowa
John Spencer
University of Iowa

\begin{abstract}
The dimensional change card sort (DCCS) task is a model paradigm for studying developmental change attention switching during childhood. In the DCCS, children are asked to sort two-dimensional cards (e.g., blue star) by one dimension and then again by the other. Typically, 3-year-olds fail to switch dimensions, continuing to sort the cards by the first dimension. Four-year-olds readily switch dimensions. The source of this sudden developmental change has been widely debated. A recent proposal is that only older children are able to selectively attend to one dimension, enabling them to flexibly switch the dimension along which they sort. We present results showing that experience matching values along one dimension prior to participating in the DCCS task facilitates 3-year-olds' ability to switch dimensions. We implemented this experience in a dynamic neural field model of DCCS performance and found that it mimicked developmental change in dimensional attention in the model.
\end{abstract}

\title{
Autonomy in Learning Sensorimotor Spaces with Dynamic Neural Fields
}

\author{
Yulia Sandamirskaya \\ Ruhr-Universität Bochum, INI, Germany
}

\begin{abstract}
The metaphor of cognition as a dynamical system - which evolves under external forces, is constrained by the internal structure, and is shaped by the agent's behavioral history - is a fruitful source of inspiration guiding experimental and theoretical work. Dynamic Field Theory offers a framework, in which cognitive processes and their development may be modeled quantitatively as dynamics of activation functions defined over behaviorally-relevant parameter spaces. Although learning through memory trace formation is an integral part of the DFT, the behavioral spaces are assumed to be given. However, these spaces may be shaped autonomously while acting in an environment. The autonomy of learning processes is achieved if the behavioral states have intentional structure, which sustains representations to enable learning and ensures their deactivation when appropriate. In this work, I show how intentional structure realized with Dynamic Neural Fields enables autonomous development of a sensorimotor mapping involved in looking behavior.
\end{abstract}

\title{
Temporal discounting in a sequential search task
}

\author{
Ke Sang \\ Indiana University
}

Junyi Dai
Indiana University
Peter M. Todd
Indiana University

\section*{Robert L. Goldstone}

Indiana University

\begin{abstract}
Search requires individuals to balance exploration (finding new resources) with exploitation (making use of current resources) over time. How individuals perceive reward in a temporally extended search may significantly influence their explore/exploit decisions. Furthermore, how individuals perceive reward is related to impulsivity. Different aspects of impulsivity can lead to different predictions for people's search patterns. We tested some of these predictions in a search task with a non-depleting condition where resources would maintain value when exploited, and a depleting condition where resources lost value when exploited. Participants with larger temporal discounting rates (greater impulsivity) started exploiting depleting resources later than those with smaller discount rates, surprisingly appearing more patient and performing closer to the optimal level. However, this difference disappeared when resources were non-depleting. Greater impulsivity might make individuals more risk-seeking in the depleting search environment and thus explore depleting resources longer at the beginning phase of the search task.
\end{abstract}

\title{
Reducing Behaviorism and Cognitivism
}

\author{
Ricardo Sanz \\ Universidad Politecnica de Madrid
}

\begin{abstract}
Behaviorism and Cognitivism are two perspectives in cognitive science that have had significant impact in the evolution of mind theory. Behaviorism is centered in the study of mind from the perspective of externalized behavior. Behaviorism is opposed to the approach to mind analysis from an introspective point of view, even rejecting the existence of internal states. The cognitivist reaction explained some aspects of memory and learning that behaviorism failed to account for. Internal mental states -beliefs, values, intentions- are the common trade of cognitivist models. Cognitivism employs computer-based information processing as the paradigmatic principle sustaining cognitive modeling, where data structures implement the required internal mental states. These two approaches to cognitive science are usually considered as antagonistic; but they are not. The only difference is that they approach the problem of the modeling of mind from an input/output perspective or from a statebased perspective. Behaviorism focuses on using a black box model of minds while cognitivism focuses on white box models. Both are right and complementary approaches and are degenerate forms of a more fundamental approach that may be applied to mind modeling: systems identification.
\end{abstract}

\title{
Modeling biconditionals: Equivalence in an uncertain world
}

\author{
Koji Sawa \\ Japan Women's University, Kawasaki, Kanagawa, Japan \\ Junki Yokokawa \\ Tokyo Denki university, hatoyama, saitama, japan \\ Tatsuji Takahashi \\ Tokyo Denki university, hatoyama, saitama, japan
}

\begin{abstract}
For almost a half century, it has been known that participants consider the truth value of "if \(p\) then q," in a "defective" way, as true when \(p\) and \(q\) are both true, false when \(p\) is true but \(q\) is false, but uncertain whenever \(p\) is false. Recently, researchers has given this truth table a new normative status, under the theory of subjective probability by de Finetti, as of conditional event, \(\mathrm{q}-\mathrm{p}\). On the basis of ample evidence that \(\mathrm{P}(\) if p then q\()=\mathrm{P}(\mathrm{q}-\mathrm{p})\), we study biconditionals, "if p then q , and if q then p ." Here we show the psychological priority of biconditional event to material equivalence. Additionally, we discuss the logical background of (bi-)conditional event and three types of uncertainty. Two are ones in the domain and codomain of the truth-function, and the other is connected to the indefiniteness of a world or the frame problem.
\end{abstract}

\title{
Integrating Top-down and Bottom-up Approaches in Holistic Perceptual Categorization
}

\author{
Samer Schaat \\ Vienna University of Technology
}

\author{
Alexander Wendt \\ Vienna University of Technology \\ Dietmar Bruckner \\ Vienna University of Technology
}

\begin{abstract}
Perceptual categorization is a key problem for an agent to cope with its internal and external world. Following a functional and subjective approach of cognitive modeling, the primary purpose of perceptual categorization is modeled as the valuation of a stimulus regarding its potential to satisfy an agent's current needs. Additionally an integrated and holistic approach is used, where categorization considers the integration of subjective influences. Such an approach complies with the consideration of top-down perception and priming. Using an activation-based exemplar model, the objective criterion of perceptual similarity, which represents bottom-up aspects of perception, and the subjective expectation-based criterion of cathexis, which represents top-down aspects of perception, are integrated in a holistic multi-criteria model of perceptual categorization. An Artificial Life simulation demonstrates the model's ability to relate stimulus objects to an agent's internal needs. Additionally, the usage of multiple criteria provides a more confident valuation of stimulus objects.
\end{abstract}

\title{
Your mind wanders weakly, your mind wanders deeply: Objective measures reveal mindless reading at different levels
}

\author{
Daniel J. Schad \\ Charité, University Medicine Berlin, Germany
}

Antje Nuthmann
University of Edinburgh, UK
Ralf Engbert
University of Potsdam, Germany

\begin{abstract}
During mind wandering, attention is directed away from the external environment and cognitive processing is decoupled. Mind wandering is usually treated as a dichotomy and often measured using self-reports. We here propose the levels-of-inattention hypothesis, postulating graded decoupling at different processing levels. To measure levels of decoupling during reading we introduce the sustained attention to stimulus task (SAST), which relies on psychophysics of error detection. We found that subjects were less likely to notice errors at higher levels of cognitive processing. Eye tracking showed that before errors were overlooked effects of high- and low-level linguistic variables were reduced in a graded fashion, indicating episodes of weak and deep decoupling. Individual gaze durations predicted overlooking of errors five seconds before they occurred. Our findings support the levels-of-inattention hypothesis and suggest levels of mind wandering can be measured in the SAST. Eye tracking provides a promising tool to detect mind wandering online.
\end{abstract}

\title{
Automatic processing of metaphor in right hemisphere lesion patients
}

\author{
Gwenda L Schmidt-Snoek \\ Hope College
}

Marguerite McQuire
University of Pennsylvania
Anjan Chatterjee
University of Pennsylvania

\begin{abstract}
While metaphor comprehension deficits were initially reported in right hemisphere lesion patients (RHLs), more recent work fails to find such deficits. One possibility is that the metalinguistic tasks typically used only uncover conscious, controlled comprehension processes. We tested the hypothesis that chronic RHLs would show a deficit in a task that taps automatic processes, even if their performance on a metalinguistic task was not abnormal. Nine RHL were tested on a sentence/matchingword lexical decision priming task with literal and metaphorical sentences and a short SOA. Nine left hemisphere lesion patients and thirteen age and education matched healthy adults served as controls. A two-way mixed ANOVA revealed a main effect of sentence type ( \(\mathrm{F}(1,28)=7.0, \mathrm{p}=.01\) ) but no effect of group. All three groups showed larger priming effects for literal sentences ( \(\mathrm{M}=70 \mathrm{~ms}\) SD=70ms) than for metaphorical sentences ( \(\mathrm{M}=34 \mathrm{~ms}, \mathrm{SD}=68 \mathrm{~ms}\) ). These data suggest chronic RHLs do not have metaphor comprehension deficits even with an automatic processing task.
\end{abstract}

\title{
Interacting with digital maps for indoor navigation
}

\author{
Verena Schnitzler \\ Center for Cognitive Science, University of Freiburg, Germany
}

\begin{abstract}
This contribution presents studies currently underway about the impact of interactive, digital maps in architectural settings. Google Indoor Maps allows users of Android smartphones to identify their position and heading in complex public buildings. The study compares participants with and without access to such a device. Participants perform both spatial learning and free navigation tasks in a building. We evaluate how their wayfinding behavior as well as performance on spatial memory measures differs between groups with respect to: efficiency, effectiveness on the task; satisfaction and perceived competence with both the digital device and the building setting. We further compare groups with high vs. low spatial abilities (perspective taking, SBSOD sense of direction) as well as different cognitive styles (route vs. survey preference), as expect an impact on both strategies of interacting with the device and task success.
\end{abstract}

\title{
The conversational style of mothers interacting with their two- and five-year-old children
}

\author{
Tanja Schorch \\ Max-Planck-Institute for Human Cognitive and Brain Sciences, Leipzig, Sachsen, Germany \\ Jens Brauer \\ Max-Planck-Institute for Human Cognitive and Brain Sciences, Leipzig, Sachsen, Germany
}

\begin{abstract}
We investigated interactions between mothers and their children and specifically measured differences in the mothers' verbal and nonverbal communication as a function of their children's age.

Forty two-year-old children and forty five-year-old children and their mothers were video-recorded in two conversational settings. Measures were obtained for the mothers' speech complexity (mlu, use of verbs, complex sentences, and direct objects), their emotional prosody (pitch, pitch variability), and their pointing gestures.

Mothers of 5-year-olds used more complex speech than mothers of 2-year-olds, whereas mothers of 2-year-olds showed more emotional prosody, and pointed more often than mothers of five-year-olds. This is interpreted as an adaptation to their children's conversational abilities. As younger children do not understand very complex utterances, but might rely on more nonverbal communication (such as pointing gestures and emotional prosody) than older children, their mothers might use a less verbally complex, but rather emotionally rich and non-verbally sophisticated interaction style.
\end{abstract}

\title{
Do pictures facilitate mental imagery during text processing?
}

\author{
Anne Schüler \\ Knowledge Media Research Center \\ \section*{Katharina Scheiter} \\ Knowledge Media Research Center
}

\begin{abstract}
The aim of the reported study was to test the hypothesis that text-picture combinations aid learning because pictures reduce the learners' need to actively generate mental images on their own. This, in turn, should free up cognitive resources which can be used for other cognitive processes associated with learning, resulting in better performance compared to a textonly condition. This hypothesis was confirmed in an experiment based on a \(2 \times 2\) design with picture presentation prior to text (yes vs. no) and visuo-spatial secondary task (with vs. without) as independent variables: Learners without pictures showed a higher load of the visuo-spatial sketchpad (i.e., higher interference with the secondary task) during text processing than learners with pictures, presumably because the former generated mental images during reading. This interpretation is supported by the finding that pictures were especially helpful for learners with low imagery capacity. The implications of these results are discussed.
\end{abstract}

\title{
We never walk alone: The influence of social interaction on wayfinding behavior
}

\author{
Sarah Schwarzkopf \\ Center for Cognitive Science Freiburg \\ Rul von Stülpnagel \\ Center for Cognitive Science Freiburg \\ Christoph Hölscher \\ ETH Zurich Cognitive Science
}

\begin{abstract}
When people move through large-scale environments, they use multiple strategies to find their ways. While many studies have investigated individual wayfinding strategies, in naturalistic settings people usually don't navigate alone, but search for their destinations in pairs or small groups. Having a shared goal can have an impact on the perception of the scene, the attentional focus and the behavior of the individual participant. We investigate how social interaction influences navigation behavior and focus on a cooperative scenario. We developed a joint wayfinding paradigm to measure walking trajectories, walking speed, gaze behavior and speech data in a naturalistic indoor setting. We use two mobile eyetracking devices to analyse the behavior of pairs of participants solving wayfinding tasks. We'll present data on the impact of participants' cooperation behavior on perceptual and attentional processes and wayfinding strategies.
\end{abstract}

\title{
Does the listener keeps spatial information from the speaker's cohesive gestures to comprehend subsequent sentences without gestures?
}

\author{
Kazuki Sekine \\ University of Birmingham \\ Sotaro Kita \\ University of Birmingham
}

\begin{abstract}
This study examined whether listeners keep spatial story representation created by speaker's cohesive gestures beyond the concurrent sentence. Participants were presented with three-sentence discourse with two protagonists, in the first and second sentences, the gestures consistently assigned the two protagonists in either right or left of the gesture space. The third sentence (without gestures) referred to one of the protagonists, and the participants responded with one of the two keys to indicate the relevant protagonist. The response keys were either spatially congruent or incongruent with the gesturally established locations for the two participants. Experiments 1 and 2 showed that the performance in the congruent condition was better than that in the incongruent condition. Thus, listeners make a spatial story representation based on the gestures, and the spatial representation persists beyond the concurrent sentence, and the information is still activated in a subsequent sentence without a gesture.
\end{abstract}

\title{
The influence of spatial cueing on serial order visual memory
}

\author{
Rakesh Sengupta
}

Center for Neural and Cognitive Sciences, University of Hyderabad Hyderabad-500046, India
Anvita Gopal
Center for Neural and Cognitive Sciences, University of Hyderabad Hyderabad-500046, India
Prajit Basu
Dept of Philosophy, School of Humanities \& Center for Neural and Cognitive Sciences, University of Hyderabad Hyderabad-500046, India

David Melcher
Center for Mind/Brain Sciences, University of Trento Rovereto-38068, Italy

\section*{Raju Bapi}

School of CIS \& Center for Neural and Cognitive Sciences, University of Hyderabad Hyderabad-500046, India

\begin{abstract}
In this study we manipulated colors and shapes in different blocks in order to investigate the differential effect of stable spatial cues on color and shape recall. The task involves recency judgment where two items from the stream of serially presented stimuli are shown and the subjects have to respond by indicating which of them came later in the sequence. Our studies show a clear dissociation between color and shape recall. While spatial cues seem to facilitate shape recall, they seem to degrade the color recall performance. This effect becomes significant for shape-location trials beyond the working memory capacity limit of 4 items. The results can be interpreted as if individuation of colored objects draws from the same attentional resources as spatial attention and the resulting competition degrades performance in serial order judgment involving color whereas shape recall does not seem to be subject such competition.
\end{abstract}

\title{
Physical Skill and Idea Interaction in the Creation of New Dance Movements
}

\author{
Daichi Shimizu \\ University of Tokyo, Bunkyoku, Tokyo, Japan \\ Takeshi Okada \\ University of Tokyo, Bunkyoku, Tokyo, Japan
}

\begin{abstract}
It has been suggested that in creative activities, cognition and physical action are related to each other. This study focuses on this relationship in breakdance, which is a creative activity of artistic and acrobatic movements. For four months, we conducted field observations of practice sessions of three expert breakdancers, and held interviews with them to investigate the creation process of new and original movements. The video records of the 34 practices and the interview data were analyzed with respect to two aspects: 1) Whether or not the dancers performed important movements appropriately; and 2) What the dancers were thinking when generating new aspects of the movements. The results show an interactive process between the development of dance movements and the generation of ideas. The dancers gradually became able to perform the movements appropriately by generating new ideas, and they generated new ideas using the somatic sensation of the new movements.
\end{abstract}

\title{
The effect of metacognitive strategies during reviewing
}

\author{
Hocheol Shin \\ Sungkyunkwan University
}

\section*{Hongoak Yun}

Sungkyunkwan University
Jungsun Yoo
Sungkyunkwan University

\author{
Kwangsu Cho \\ Sungkyunkwan University
}

\begin{abstract}
A peer-based reviewing on writing often leads to improving learners' writing quality. Particularly, reviewing by writing comments lifted writing quality more greatly than reviewing without writing comments. This implies that learners' hard working on peers' writing engenders benefits on their own writing. This study attempted to explore which cognitive mechanism underlies in this cooperative learning system. We questioned whether learners' use of metacognitive strategies during reviewing might facilitate learners' self-learning. In a preliminary study, participants were asked to categorize their comments on peers' reports. Results from the study revealed that learners who categorized comments used metacognitive strategies more increasingly than those who did not. Controlling for the amount of using metacognitive strategies, writing quality was improved numerically more when learners were engaged in the categorization task than when they were not. With further data, we will examine to what degree learners' metacognitive learning during reviewing contribute to writing improvement.
\end{abstract}

\title{
Foot Movement for Walking Biases the Visual Attention Allocation
}

\author{
Marie Shoda \\ The university of Tokyo, Bunkyo-ku, Tokyo, Japan
}

Kazuhiko Yokosawa
The University of Tokyo

\begin{abstract}
Precise walking requires ample cognitive resource to control balance in the single leg stance while head moves periodically in the vertical axis within each stride cycle. Because visual information is a key factor in motor control, we hypothesize that walking alters the allocation of visual attention, thereby affecting the selective intake of visual information. Participants localized a peripheral dot while identifying the central letter in a display during treadmill walking, stepping in place, and standing still. Vertical head movement occurred only in the walking condition. Results show that visual attention was skewed in the single leg stance during walking and stepping in place. Vertical head movement affected the allocation of visual attention because only in the walking condition was attention oriented downward without skewing during the double leg stance. Cognitive resources and head vertical movement appear to work differently in adaptation of visual attention to walking.
\end{abstract}

\title{
CLEARPOND: Phonological and orthographic neighborhood information for multiple languages
}

\author{
Anthony Shook \\ Northwestern University
}

\author{
Sarah Chabal \\ Northwestern University
}

\author{
James Bartolotti
}

Northwestern University
Viorica Marian
Northwestern University

\begin{abstract}
Past research has demonstrated that neighborhood variables (e.g., neighborhood density) influence lexical processing (Andrews, 1989; Vitevitch \& Luce, 1999), but can have distinct effects for different languages (Vitevitch and Stamer, 2006). To explore how neighborhoods can vary both within and across multiple languages, we have developed CLEARPOND (the CrossLinguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities), a database of phonological and orthographic neighborhood information for five languages: Dutch, English, French, German, and Spanish. Analyses using the CLEARPOND database revealed consistent effects of lexical frequency and word-length on neighborhood size across languages, while also highlighting how the languages differed in the distribution of neighbor-types (phonological/orthographic). CLEARPOND not only provides a tool that can be used to study differences between languages but also provides detailed information about characteristics of individual words in multiple languages. The CLEARPOND database is freely-available online and can be accessed at http://clearpond.northwestern.edu.
\end{abstract}

\title{
Thematic Music and Context Interact to Affect Inclusion of Schematic Elements During Story Generation
}

\author{
Cynthia Sifonis
}

Oakland University

\begin{abstract}
Listening to a \(90-\mathrm{sec}\) excerpt of thematic music (e.g., baby music) affects performance in a subsequent generation task by increasing the likelihood that concepts associated with the thematic music (e.g., "Sleep," "Baby") will be incorporated into the novel product (Sifonis \& Fuss, 2012). The generation task theme (e.g., writing a story about a visit to an alien planet versus writing about a visit to an undiscovered, foreign land) also interacts with the music theme to affect the degree to which concepts associated with the music are incorporated into the generated product.

The current study tested and supported the hypothesis that listening to thematic music activates complex schematic structures. The generation task theme provides context, thus affecting the specific manifestation of the thematic elements in the generation task (e.g., including concepts associated with a foreign land with many children versus those associated with an alien planet with many children).
\end{abstract}

\title{
Embodied Mental Imagery
}

\author{
Jan Frederik Sima \\ University of Bremen
}

\begin{abstract}
A new theory of visuo-spatial mental imagery and a computational model of the theory are presented. The theory assumes (visual) perception to be an active and guided process comprising of several low-level perceptual actions. Mental concepts are grounded in sets of such actions. Imagery comprises the "offline" employment of these actions providing concrete instances of the mental concepts through bodily feedback such as proprioception. The theory is compared to the contemporary theories and evaluated against a set of well established phenomena, i.e., mental scanning, mental reinterpretation, eye movements, unilateral neglect. It is argued that the theory provides explanations that go beyond those offered by the contemporary theories. The results provide support for the explanatory power of an embodied approach to cognition in which sensorimotor interaction constitutes conceptual knowledge in that it provides grounding and concrete manifestation of the semantics of mental concepts in the domain of visuo-spatial mental imagery.
\end{abstract}

\title{
Collaborative Recall and Extended identity in Twitter communication
}

\author{
Jurǵis Šķilters \\ University of Latvia \\ Līva Brice \\ University of Latvia \\ Krista Bērziņa \\ University of Latvia \\ Uldis Bojārs \\ University of Latvia \\ Monika Kreile \\ University of Oxford \\ Marta Selecka \\ University of Latvia
}

\begin{abstract}
The aims of the current study are to analyze research evidence of the ways in which collective recall exhibits extended social identity effects . Our core hypothesis was that the use of Twitter as a recall tool significantly contributes to social identity generation in general and self-categorization in particular. Based on a study of a representative sample of Latvian-language Twittersphere, we argue that a social network serves a two-fold role: (a) it extends the individual self as part of a distributed social reality and as part of a distributed on-line social network. The core results of our study show that Twitter functions as an extended distributive linguistic cognitive system supporting different kinds of recall tasks while at the same time exhibiting strong categorization effects through eliminating redundant information and reducing the descriptive complexity of the environment in recall.
\end{abstract}

\title{
Probing the neural dynamics of visual working memory with dynamic fields and fMRI
}

\author{
John Spencer \\ University of Iowa
}

\author{
Aaron Buss \\ University of Iowa \\ Vincent Magnotta \\ University of Iowa
}

\begin{abstract}
Efficient visually-guided behavior depends on our ability to form, retain, and compare visual representations that may be separated in space and time. This ability relies on visual working memory (VWM). Here, we describe a layered neural architecture that captures the cortical population dynamics that underlie VWM. We then test this model using functional neuroimaging. Recent work has shown that the BOLD response is strongly correlated with local field potentials (LFPs). An analog of LFPs can be estimated from dynamic neural field (DNF) models. This estimate can be convolved with an impulse response function to yield time-dependent hemodynamic predictions. Using this approach, we show that the DFN model quantitatively captures fMRI data from recent studies probing changes in the BOLD response in the intraparietal sulcus (IPS) as set size increases in change detection. We also test a novel prediction of the model that BOLD responses should be greater on false alarms versus misses. These data run counter to common explanations of the origin of errors in change detection.
\end{abstract}

\title{
Looking at an empty side of the world. Hemispheric specialisation in directing language-mediated eye movements.
}

\author{
Sara Spotorno \\ University of Dundee, Dundee, Scotland, United Kingdom and University of Aix-Marseille, Marseille, France \\ Jens Apel \\ University of Wuppertal, Germany \\ \section*{Bérengère Plaza} \\ University of Nice-Sophia Antipolis, Nice, France \\ Sylvane Faure \\ University of Nice-Sophia Antipolis, Nice, France
}

\begin{abstract}
We investigated the relationship between linguistic and visual information, combining divided visual field and blank screen paradigms. In an eye-tracking experiment, two objects appeared for 180 ms , one in the right (rvf) and one in the left visual field (lvf), while participants maintained central fixation. After the objects disappeared, a word was presented auditorily. In matching trails ( \(50 \%\) ), it indicated one of the objects previously shown. Participants had to decide whether the word named a man-made or a natural entity. Findings revealed that they were more likely to saccade toward the side of the referent object when it had been presented in the lvf than in the rvf. Moreover, saccades in the lvf targeted more precisely the object's empty location. This suggests a crucial role of the right hemisphere in activating visual representations during language processing, indicating its greater ability in using spatial indexes to retrieve useful visual information.
\end{abstract}

\title{
How common ground affects the use of gesture space
}

\author{
Kashmiri Stec \\ University of Groningen \\ Mike Huiskes \\ University of Groningen
}

\begin{abstract}
Recent work has started to ask to what degree common ground, i.e. shared background knowledge, may affect the articulation of co-speech gestures, specifically as regards viewpoint. We extend that work by asking how gesture articulation is affected, and how the spatial location of those gestures changes as a result of common ground. Using naturalistic data, we find that with quotes grounded in past interactions, co-speech gestures are fewer and smaller, whereas with quotes ungrounded in past interactions (UGPI), co-speech gestures are more numerous and larger. Moreover, when speakers repeatedly reference the same entity, the co-speech gestures used become smaller with each repeated reference, thereby suggesting a kind of 'online' grounding. We also find that the use of gestural space is more consistent for UGPI utterances. This suggests that common ground affects both the articulation of gestures, and the way that gesture space is used.
\end{abstract}

\title{
The impact of Individual and Instructional Difference on Learning Measurement Concepts
}

\author{
Nancy L Stein \\ University of Chicago ( U of C) \\ Yuhtsuen Tzeng \\ National Chung Cheng University \\ Linda Siegel \\ University of British Columbia
}

\begin{abstract}
Abstract: This study compared 4th grade children in the U.S. and Taiwan on understanding (linearity, area, and volume. Taiwanese and American children were both under \(50 \%\) correct for the concepts of area and volume in October. By May, Taiwanese children learned \(90 \%\) of the targeted concepts. American children were still under \(50 \%\) correct.

During the following year, U.S. children received five weeks of instruction on linear, square, and cubic measure, or they received no instruction. Results showed significant increases on accurate knowledge of all measurement concepts, average score \(=89 \%\). Improvement was related, at the .72 level, to memory for forward going digit span. The control group did not improve on any type of measurement concept, despite their similar scores on memory for digit span. A third study showed that repeated practice on multiplication lowered the effects that digit span memory has on learning measurement concepts.
\end{abstract}

\title{
Eye motion along a static concrete object: Why "fictive motion" is not a figurative mental phenomenon
}

\author{
Kurt Stocker
}

University of Zurich

\begin{abstract}
Fictive motion (specifically the co-extension path type: "The road runs through the desert") is widely considered a specific class of figurative language. However, cognitive-linguistic and conceptual-metaphoric evidence is presented which suggests that linguistically expressed motion in fictive motion implicitly refers to the processing of a concrete action: to eye motion that occurs while scanning along a static visual percept (or mental image). As such, fictive motion differs principally from genuinely figurative expressions like ARGUMENT IS WAR - the figurative element of the latter cannot be interpreted as referring directly to a physically existing action or object. Also demonstrated is how a "non-figurative" theoretical approach to fictive motion can explain and predict so-called "non-fictive motion"; fictive motion-like temporal cognition; and fictivemotion types other than co-extension path. It is also shown how the various experimental cognitive and neurocognitive findings on fictive motion can be (re)interpreted in the new framework.
\end{abstract}

\title{
No adaptive strategy selection without outcome feedback
}

\author{
Hrvoje Stojic \\ Universitat Pompeu Fabra, Barcelona, Catalunya, Spain \\ Henrik Olsson \\ Max Planck Institute for Human Development, Berlin, Germany
}

\begin{abstract}
A common view in the judgment and decision making literature is that humans posses a repertoire of decision strategies, from which we choose adaptively when dealing with decision situations. Modeling this adaptive strategy selection process has proved to be difficult, but advances were recently made by modeling it as a reinforcement learning process. Dieckmann and Rieskamp (2007) investigated the influence of information redundancy in a multiple-cue learning setting and found surprising evidence for adaptive strategy selection in situations without outcome feedback, which is difficult to explain by a pure reinforcement learning model. We challenge these findings by pointing out problems in their experimental design. We replicate their experiment, add conditions with stricter experimental controls, and investigate possible underlying mechanisms. In conditions with stricter controls we find no evidence that participants manage to incorporate information redundancy into strategy selection and conclude there is no adaptive strategy selection without outcome feedback.
\end{abstract}

\title{
Judgments under Uncertainty: Bias, Experience, and Expectations
}

\author{
Chris N.H. Street
}

University College London

\author{
Daniel C. Richardson \\ University College London
}

\begin{abstract}
We are poor lie detectors, with accuracy only marginally better than guessing. Raters in social situations such as these must battle with their own uncertainty if they are to make a judgment, yet lie detection research has given little attention to the effects of uncertainty. We present evidence suggesting prior knowledge and expectations have an early influence on the judgment process, but only when forced into judgment. If able to abstain, raters do not rely on prior knowledge but rather indicate their uncertainty by withholding judgment until all the available information has been presented. After the speaker has presented their statement, and when no additional new information is available, we provide evidence from a number of experiments that lie detectors integrate their pre-existing biases and experience about deception in general with more specific information about the statement at hand.
\end{abstract}

\title{
An Attentionally Constrained Model of Statistical Word Learning
}

\author{
Sumarga Suanda \\ Indiana University \\ Seth Foster \\ Indiana University
}

\author{
Linda Smith \\ Indiana University
}

\author{
Chen Yu
}

Indiana University

\begin{abstract}
Recent research supports the notion that word learning can be conceptualized as a statistical learning process. As many have noted however, statistical learning is constrained by processes such as attention and memory. Here we tested an attentionally constrained framework of statistical word learning. We observed, through infant-perspective head cameras, infants' visual input as parents labeled novel objects during an infant-parent object-play session. We then constructed statistical learning models that aggregate word-to-object associations. We fed a baseline model the word-to-object co-occurrence patterns obtained from parent-infant observations. We fed an attentionally constrained model weighted co-occurrence patterns based on the perceptual properties of the objects (i.e., object sizes from the infant's view) at the time words were uttered. Models' learning was compared to children's forced-choice test results. Of interest is which of the two models best approximates children's learning. Implications of these results for statistical learning accounts of word learning will be discussed.
\end{abstract}

\title{
The time course of structural and number interference in sentence processing
}

\author{
Katja Suckow \\ School of Psychology, University of Dundee \\ Roger van Gompel \\ School of Psychology, University of Dundee
}

\begin{abstract}
In sentence processing, retrieval cue parsing accounts predict that processing difficulty occurs due to interference between similar noun phrases at verb integration (Van Dyke \& Lewis, 2003; Van Dyke, 2007). An important factor is number marking of the nouns because the number of the verb and the subject has to agree. According to Lewis and Vasishth (2005), different types of retrieval cue overlap (e.g., verb subcategorisation, semantics, number) should all cause interference simultaneously during processing. However, Van Dyke (2007) reported that interference due to subcategorisation overlap preceded semantic interference, suggesting that syntactic interference may occur before interference due to other cues. We investigated whether interference due to number overlap also occurs later than interference due to subcategorisation overlap. Interestingly, using eye-tracking, we found that the number congruency effect occurred early and no later than the subcategorisation interference effect, indicating that number interference has a very rapid effect on sentence processing.
\end{abstract}

\title{
Imagining emotions: An ERP study on mental imagery of facial expressions
}

\author{
Franziska Suess \\ Humboldt-University of Berlin \\ Rasha Abdel Rahman \\ Humboldt-University of Berlin
}

\begin{abstract}
Affective stimuli encountered in everyday life - such as emotional words, scenes or facial expressions - can elicit well-investigated emotional experiences. For instance, two distinct event-related brain potentials (ERPs) have been reported in response to emotional facial expressions, an early posterior negativity (EPN), associated with enhanced attention and perception of affective stimuli, and a later centro-parietal positivity (LPP), assumed to reflect processing of the intrinsic relevance of emotional stimuli. Other rich sources of emotions that have yet received little attention in EEG research are internal mental events such as thoughts, memories and imagination.

Here we investigated mental imagery of emotional facial expressions and its time course using ERPs. We presented participants with neutral faces and asked them to imagine the faces with an emotional or neutral expression. Early ERP modulations during imagery resemble the effects frequently reported for emotional facial expressions, suggesting shared early processes underlying emotion perception and imagination.
\end{abstract}

\title{
Perspective preference and spatial learning: An eye movement investigation
}

\author{
Masashi Sugimoto \\ Kyoto University \\ Takatsugu Kojima \\ Shiga University of Medical Science \\ Hiroyuki Tsuda \\ Kyoto University \\ Shogo Kajimura \\ Kyoto University \\ \(\underset{\text { Kyoto University }}{\text { Ayae Iwamoto }}\) \\ Kyoto University \\ Yuki Sato \\ Kyoto University \\ Takashi Kusumi \\ Kyoto University
}

\begin{abstract}
The current study investigated the effect of perspective preference on spatial learning. Previous studies revealed that certain factors, such as strategies and goals, affect wayfinding behavior and map learning. We focused on perspective preference as analyzed through measurement of eye movements, which enables us examine potential differences in spatial learning. We divided university students into a survey or route preference group through a map-learning questionnaire. While participants studied the maps, we measured participants' eye movements to each map element (landmarks, streets, and compass rose). After studying a map, participants completed verification tasks and wrote their own map of the learned environment. In contrast to our prediction, the survey and route groups did not show the difference in regard to their gaze to most of the map elements. Although results did not emerge as expected, we discuss the effect of perspective preference on map learning and memory.
\end{abstract}

\title{
Finding Structure in Space and Time: Active Trace of Patterns
}

\author{
Yanlong Sun \\ University of Texas Health Science Center at Houston \\ Hongbin Wang \\ University of Texas Health Science Center at Houston
}

\begin{abstract}
In extracting statistical regularities from the seemingly random environment, our minds grow special interests in patterns. To account for such a behavior, much research has been focusing on top-down influences such as the representativeness heuristic and Bayesian belief updating. Here we take a reverse-engineering approach by first examining the waiting time statistics and the self-overlap property of patterns and revealing a normative basis for people's special attention to patterns. With a unsupervised neural network simulation, we show that different patterns may leave different traces in mind corresponding to the waiting time statistics, indicating an early pattern dissociation without any top-down guidance. We argue that the sense of randomness could have started locally with short sequences and emerged early at the perceptual level, and, the process of spatial-temporal association may be the early driving force towards a structured hypothesis space.
\end{abstract}

\title{
Cognitive inhibition and the unbelievability of supernatural ideas
}

\author{
Annika Svedholm \\ University of Helsinki, Helsinki, Finland \\ Marjaana Lindeman \\ University of Helsinki, Helsinki, Finland \\ Miika Leminen \\ University of Helsinki, Helsinki, Finland
}

\begin{abstract}
After decades of research, the reason why people are attracted to the supernatural (paranormal, magical, superstitions) is poorly understood. It is possible that the question has been approached from the wrong angle and that it is skepticism that needs to be explained, not the beliefs. Taking this as a starting point, we examined \((\mathrm{N}=40)\) whether skeptics have stronger cognitive inhibition than believers. Because cognitive load disrupts inhibition and reveals intuitive thinking, we hypothesized that working memory load increases ontological confusions less among skeptics than among believers. Ontological confusions, such as conceiving of lifeless objects as having mental states, are known to be central to supernatural beliefs. The results supported the hypothesis. Strong cognitive inhibition may thus partly explain why supernatural beliefs, albeit based on natural information processing, seem so unbelievable to millions of people. An ongoing study examines skeptics' and believers' cognitive inhibition in more detail.
\end{abstract}

\title{
Collaborative Memory Foraging in Categorical Recall Tasks
}

\author{
Janelle Szary \\ University of California, Merced \\ Christopher Kello \\ University of California, Merced \\ Theo Rhodes \\ State University of New York at Oswego
}

\begin{abstract}
A classic paradigm for investigating memory is the category recall task, where participants recall as many items as possible from a given category, within some time window. Category recall tasks have been used to investigate memory as a search process, where memory is conceptualized as a landscape with distributed resources (resources being the target items of memory recall). Rhodes and Turvey (2007) show that the dynamics of memory search are akin to animal foraging behavior. Specifically, patterns of recall exhibited Lévy processes, which have been observed in many species and at many scales, and are hypothesized to be optimal under certain conditions. Here, we investigate the effects of social context on Lévy processes using a collaborative category recall task. Although the processes of collaborative recall may differ from individual recall, our results suggest that the products of that recall are similar.
\end{abstract}

\title{
Is Dishonesty an Automatic Tendency?
}

\author{
Maryam Tabatabaeian \\ University of California Merced, merced, California, United States \\ Rick Dale \\ University of California Merced, merced, California, United States
}

Nicholas D Duran
University of California Merced, merced, California, United States

\begin{abstract}
The present study utilizes a novel task to test two competing hypotheses concerning the automaticity of dishonesty. The traditional hypothesis claims that in order to act dishonestly one has to first overcome the truth bias, which results in more time and effort. The opposing hypothesis indicates that lying in order to serve self-interest is an automatic tendency, and therefore takes less time than refraining from lying.

The goal is to look at the action dynamics of dishonesty in order to investigate its underlying cognitive processes. Subjects were asked to privately predict the outcome of a virtual coin-flip. After observing the actual outcome they reported whether their prediction was correct or wrong. The movements of the mouse towards the target answer were recorded and used for action dynamic analysis. Our results support the latter hypothesis indicating that dishonest people take less time and experience less hesitation while choosing the deceptive answer.
\end{abstract}

\title{
The searching effect of metaphor on text rereading: Difference by familiarity
}

\author{
Tomohiro Taira \\ Osaka City University
}

\begin{abstract}
Text comprehension is the process of searching and uniting information. In this process, some pragmatic triggers such as metaphorical expressions can help select important sentences from the text. In this regard, Taira and Kusumi's (2008) demonstration of the effect of metaphor comprehension on the text rereading process is controversial and lacking in persuasive data. This study comprises two aspects: a reanalysis of Taira and Kusumi and an additional experiment examining how the effect of metaphor comprehension on text rereading works. The results of this study showed the effect of metaphor, and indicated two points: (1) the effect of metaphor provide a meaning-searching process that showed a delayed reading time of sentences that describe important information regarding a text topic, and (2) this effect was shown in the case of unfamiliar metaphors, and not familiar metaphors.
\end{abstract}

\title{
Formation of an art concept: How is visual information from photography utilized by the artist in concept formation?
}

\author{
Kikuko Takagi \\ The University of Tokyo
}

Takeshi Okada
The University of Tokyo

\section*{Sawako Yokochi}

Tokyo Future University

\begin{abstract}
The art concept plays an essential part in the creation of contemporary art. In order to capture the progress of the formation process of an art concept, we conducted a case study of a contemporary artist. We interviewed a professional artist about his creation process once every three weeks for about ten months. This report focuses on an early phase of the search for ideas, in which he took many photographs to collect visual information to form the core part of the art concept. Using both the photographs he took and the interview data collected during this phase, we identify when and how features of his art concept emerged. The results show that his art concept was formed through cycles of two types of search for visual information: an active, explorative search to find the unexpected, and more a focused search to interpret it.
\end{abstract}

\title{
Bounded rationality leads to optimal decision-making and learning under uncertainty: Satisficing, prospect theory, and comparative valuation breaking the speed-accuracy tradeoff
}

\author{
Tatsuji Takahashi \\ Graduate School of Science and Engineering, Tokyo Denki University \\ Kuratomo Oyo \\ Graduate School of Science and Engineering, Tokyo Denki University \\ Yu Kohno \\ Graduate School of Science and Engineering, Tokyo Denki University
}

\begin{abstract}
Some classically rational standards for actions such as optimization are simply intractable. We often instead satisfice a certain reference level that is good enough for us. We can use some heuristics but they may lead to biases. Though rational analysis by Anderson (1990) can argue the adaptive rationality of biases in relation to the environmental structure, heuristics and biases have been mostly studied in isolation from other factors in conformity with the tradition in psychology. To show the efficacy of the subrational heuristics in union, we execute computer simulations adopting the framework of reinforcement learning that models iterative decision-making under uncertainty. We implement three characteristics representative of human behavior: Satisficing (Simon, 1952), risk attitudes and reflection (Tversky \& Kahneman, 1981), and comparative valuation (Kahneman \& Tversky, 1979). We show that they, combined together, exhibit an adaptively optimal behavior with an extremely easy parameter tuning.
\end{abstract}

\title{
Source-target mapping strategy in displaced communication: Communication strategy to represent absent objects
}

\author{
Kaori Tamura \\ Japan Advanced Institute of Science and Technology, Nomi, Ishikawa, Japan \\ Takashi Hashimoto \\ Japan Advanced Institute of Science and Technology, Nomi, Ishikawa, Japan
}

\begin{abstract}
Displacement, which is to express absent objects, is one of the important design features of human language. It has not been well considered in the context of communication. We conducted a graphical communication experiment to investigate displacement in communication. In this experiment, two adults are paired, and a sender drew an absent object expressed by an unconventional combination of two words, adjective and noun, while a receiver answered what the drawing represented. This process was repeated 8 times for one object in each pair. The senders usually drew two pictures corresponding to two words, respectively. The analysis of results indicated that nouns should be understood in advance for understanding adjectives which are difficult to represent by drawings. This suggested that in displaced communication source-target mapping strategy was used to compose expressions like metaphors, and that identifying which expression represents a target was important in order to understand absent objects.
\end{abstract}

\title{
Conceptual Transformation in Origami
}

\author{
Thora Tenbrink \\ Bangor University, Wales, UK \\ Holly A. Taylor \\ Tufts University, Medford, MA
}

\begin{abstract}
Origami paper folding involves challenging spatial problem solving, including a number of complex cognitive processes that have not been extensively explored. To gain insights into the nature of these processes, we had participants think aloud while following Origami instructions (verbal and pictorial). Our analysis of participants' verbalizations revealed recurring patterns that reflect the underlying cognitive processes. Namely it showed evidence of reading and reformulating the task description, considering actions and task status, comparing task status to instructional pictures, evaluating progress, referring to previous experience, recognizing problems, and adding ideas about the current instructional step. The last two categories highlight how participants conceptualized this spatial task. The verbalizations also reflect a typical order that the cognitive processes follow: reading - reformulating - reconceptualizing - evaluating. The recurring pattern in this ordering suggests that participants gradually moved away from the original instruction towards a broader conceptualization for action in the current context.
\end{abstract}

\title{
The contribution of frequency, phonological and semantic factors in the development of the English past tense.
}

\author{
Anna Theakston \\ University of Manchester, United Kingdom \\ Grzegorz Krajewski \\ University of Warsaw, Poland \\ Sarah Keeble \\ University of Manchester, United Kingdom \\ Anna Woollams \\ University of Manchester, United Kingdom
}

\begin{abstract}
Despite a vast amount of research, debate continues concerning the mechanisms underlying correct use of morphological systems such as the English past tense. Relatively little is known about precisely what, when, and how children acquire aspects of inflectional morphology due to the paucity of studies examining the earliest stages of development, and the generally narrow focus on a small number of items and predictors. To address these problems, we provide comprehensive evidence concerning the earliest stages of development. 543 English-speaking children (196 2-year-olds, 176 3-year-olds, 1714 -yearolds) took part in a past tense elicitation task. Responses were elicited for 300 verbs ( 200 for 2 -yr-olds) and measures derived (largely from child-directed-speech) for a wide range of frequency, phonological and semantic predictor variables. We present the outcomes of analyses relating these novel predictor variables to the unique behavioural dataset, revealing the cognitive and linguistic underpinnings of children's early past tense development.
\end{abstract}

\title{
Accommodating talker-specific phonetic detail: Influences on internal category structure
}

\author{
Rachel Theodore \\ University of Connecticut
}

\author{
Emily Myers \\ University of Connecticut \\ Janice Lomibao \\ University of Connecticut
}

\begin{abstract}
Listeners customize speech processing to accommodate talker-specific phonetic variation. For example, listeners modify established phonetic boundaries to incorporate a talker's idiosyncratic productions. In addition to being marked by boundaries, phonetic categories exhibit a graded internal structure, with some members of the category considered better members than others. Here we examined whether sensitivity to talker-specific phonetic variation influences internal category structure. Two groups of listeners heard a talker produce \(/ \mathrm{k} /\). Word-initial voice-onset-time (VOT) was manipulated such that one group heard the talker produce \(/ \mathrm{k} /\) with shorter VOTs relative to the other. Listeners were then presented with a range of VOTs and asked to rate each for goodness as \(/ \mathrm{k} /\). Results to date indicate that exposure during training robustly influences the range of VOTs considered the best exemplars of \(/ \mathrm{k} /\), suggesting that accommodating talker-specific phonetic variation results in a comprehensive re-mapping of acoustic-phonetic space and is not limited to the boundary region.
\end{abstract}

\title{
A Neuro-Computational Approach to Equivalence Formation in Arbitrary Categories
}

\author{
Angel E. Tovar \\ Facultad de Psicología, Universidad Nacional Autónoma de México \\ Gert Westermann \\ Department of Psychology, Lancaster University
}

\begin{abstract}
The Stimulus Equivalence paradigm studies the learning of stimulus classes (categories) composed of functionally equivalent stimuli with or without perceptual similarities. The relations between stimuli in a class can either be learned or be derived from other stimulus relations: if stimulus \(A\) is equivalent to \(B\), and \(B\) to \(C\), then the equivalence between \(A\) and \(C\) can be derived without explicit training. There has been little work on the mechanisms underlying equivalence class formation. Here we present a neurobiologically plausible neural network model of stimulus class learning. The network successfully models three classic studies on stimulus equivalence. The Hebbian weights in the model describe the formed equivalences and the levels of association between class members, and resulting activation patterns are correlated with the response accuracy and response latencies in the original studies. The model predicts that stimulus equivalence formation depends on the environmental regularities of stimuli occurrence and co-occurrence.
\end{abstract}

\title{
Coordination among edge team members during crisis management
}

\author{
Sebastien Tremblay \\ Universite Laval \\ Isabelle Turcotte \\ Universite Laval \\ Alexandre Labrecque \\ Universite Laval \\ Jessica Desrochers-Pare \\ Universite Laval \\ Vincent Rousseau \\ University of Montreal
}

\begin{abstract}
Crisis management (CM) situations are most-often complex and dynamic, and require team members to make optimal decisions under constraints of high risk, uncertainty, high workload, and time pressure (see, e.g., Brehmer, 2007). An Edge Organization (EO) is an adaptive, rapidly reconfigurable, and distributed team structure in which no roles are previously assigned, and resources are not distributed in advance. Such a team structure is assumed to be able to improvise and respond quickly to emerging problems. We used the C3Fire CM simulation to measure the coordination (of units and resources) efficiency amongst edge team members. Twenty-four teams of four participants completed four C3Fire scenarios, each lasting 10 minutes. Results revealed that team members achieved better coordination as they progressed through the scenarios. This learning effect suggests that it could be beneficial to train team members to coordinate their actions efficiently and fully exploit the potential agility provided by EO.
\end{abstract}

\title{
Visual landmarks and the visual impedance effect
}

\author{
Kai Hamburger \\ Justus Liebig University Giessen; Experimental Psychology and Cognitive Science \\ Florian Roeser \\ Justus Liebig University Giessen; Experimental Psychology and Cognitive Science \\ Cate Marie Trillmich \\ Justus Liebig University Giessen; Experimental Psychology and Cognitive Science
}

\begin{abstract}
According to Lynch (1960; p. 83) "the sequence [of landmarks] facilitates recognition and memorization". Thus, under which circumstances do landmarks facilitate wayfinding? We therefore investigated the helpfulness of landmarks during route knowledge and survey knowledge retrieval from long-term memory. A field study with citizens of a street festival in a mid-size German town (Giessen) was performed. Sixty-three participants had to draw the shortest possible route between two given locations in the town. Within this experiment, two different conditions were tested respectively: drawing the route without landmarks and with additionally presented landmarks. A comparison of conditions revealed different performance groups: perfect performance ( \(8 \%\) ); performance improvement with additional landmarks ( \(32 \%\) ); equal performance in both conditions ( \(36 \%\) ); and performance decrement with additional landmarks ( \(23 \%\) ). These results were confirmed in two further experiments. We demonstrate that the decremental findings may be a result of the so-called visual impedance effect (Knauff \& Johnson-Laird, 2002).
\end{abstract}

\title{
Automatic and controlled processes of reasoning: insight from the matching hypothesis of syllogisms
}

\author{
Ping Ping Tse \\ Universidad de Granada \\ Sergio Moreno-Rios \\ Universidad de Granada \\ Juan Antonio Garcia-Madruga \\ Universidad Nacional de Educación a Distancia \\ Maria Teresa Bajo \\ Universidad de Granada
}

\begin{abstract}
The study aimed to study the inhibitory mechanism in syllogistic reasoning when the outputs of the heuristic strategy and analytic reasoning disagree. We manipulated the congruency of the quantifier of the conclusion with those of the two premises according to the matching strategy and the validity of the syllogism. After each syllogistic evaluation task, a lexical decision task was used to check if the semantic content of the conclusions was inhibited. The results suggested that after correctly solved conflict problems (match-invalid or mismatch-valid), the semantic priming effect of the words related to the two terms in the conclusion diminished. For no-conflict problems, the recognition time of the related words was faster than that of the unrelated words. The results suggested that inhibition on the content of a syllogism may not only be triggered by the conflict induced by the believability but also by the surface structure of the syllogisms.
\end{abstract}

\title{
Framing effects in perceptual decision-making
}

\author{
Konstantinos Tsetsos \\ Experimental Psychology, Oxford University \\ Christopher Summerfield \\ Experimental Psychology, Oxford University
}

\begin{abstract}
Research on the psychology of simple, perceptual choices has led to an impressive progress in capturing the underlying mental processes as optimal mechanisms. Within this theoretical framework, perceptual decisions arise from a feed-forward process involving the sampling and accumulation of momentary evidence up to a decision boundary. According to this view, the stage where the information is accumulated is automatic and decision makers can exert strategic control on the decision boundary only, in order to adapt their performance to the task demands (e.g. speed-accuracy trade-offs). We present new behavioural and eye-tracking data challenging this view and suggesting that the way information is accumulated in perceptual decisions, is subject to differential weighing that depends on the task framing (e.g. select the brightest or the darkest spot). We conclude that choices are mostly influenced by extreme values, and whether positive or negative peaks are more pivotal is frame-dependent and subject to top-down control.
\end{abstract}

\title{
Differences between maximizers and satisficers in regret and counterfactual thinking during repeated versus switching decisions
}

\author{
Takashi Tsuzuki \\ Rikkyo University
}

Manabu Kikuchi
Rikkyo University

\author{
Itsuki Chiba
}

Rikkyo University

\begin{abstract}
Recent research has demonstrated that with a positive prior experience, an actor who decides to switch decisions should feel more regret than one who decides to repeat (the status quo effect). Conversely, with a negative prior experience, the switcher should feel less regret than the repeater (reversal of the status quo). We tested the influence of a maximizing tendency on the strength of these two effects, measured using the Japanese version of the Regret and Maximization Scale. In the positive prior experience scenario, the maximizer group scored higher on presumed regret and counterfactual thinking than the satisficer group. In the status quo scenario, the maximizer group scored higher on counterfactual happiness than the satisficer group. Our results indicate that maximizers show a reversal of the status quo effect in some settings.
\end{abstract}

\title{
The effects of self-explanation on science concept learning for sixth grade students: Misconceptions matter
}

\author{
Yuhtsuen Tzeng \\ National Chung Cheng University \\ Yu-Ping Huang \\ National Chung Cheng University \\ Nancy Stein \\ University of Chicago
}

\begin{abstract}
Self-explanation (SE) is an effective strategy for improving understanding and it relied on learners' proper deployment of prior knowledge. Surprisingly, there is limited research on how might misconceptions affects learning despite they are inherent part of learners' knowledge system. We examine the influence of SE on processes and outcomes of science learning for 36 sixth grade students by varying degrees of prior knowledge and relevant misconceptions. The SE group read and selfexplained a text describing state changes of water requiring proper notions of molecules while the control group read twice and think aloud. The results indicated that there are no effects of SE, prior knowledge for learning outcomes. However there is significant effect of SE by misconception interaction. Low misconception students benefit from SE but not for high misconception counterparts. SE did influence amounts and types of verbal protocol students generated. The results indicated that the influence of SE heavily modulated by learners' misconceptions.
\end{abstract}

\title{
Do Japanese junior high school pupils really dislike mathematics? -Discrepancies between explicit and implicit assessment measures-
}

\author{
Akitoshi Uchida \\ Togakushi Junior High School
}

\author{
Kazuo Mori \\ Institute of Engineering, Tokyo University of Agriculture and Technology
}

\begin{abstract}
The purpose of this study is to examine whether Japanese junior high school pupils really dislike mathematics. In addition to questionnaires to inquire explicitly of the likings of school subjects, we administered the FUMIE Tests (Mori et al., 2008) to 512 junior high school pupils to assess the implicit evaluative association to the school subjects. We found a considerable proportion of pupils answered negatively to the target school subjects in the questionnaire while their implicit association scores showed somewhat positive valences. The discrepancies were larger for "mathematics" than "science". One hundred of 512 pupils answered negatively to mathematics while their implicit measures were positive. In contrast, only five of 102 showed the same discrepancy for science. These results imply that there may exist a tendency to pretend to be a math-dislike in pupils. We discussed this tendency that may eventually lead them to real math-dislikes.
\end{abstract}

\title{
Are you human? The engagement in real-time social interactions recruits the reward system
}

\author{
Ulrich Pfeiffer \\ University Hospital Cologne, Department of Psychiatry, Cologne, Germany \\ Leonhard Schilbach \\ University Hospital Cologne, Department of Psychiatry, Cologne, Germany \\ Bert Timmermans \\ University of Aberdeen, School of Psychology, Aberdeen, UK \\ Bojana Kuzmanovic \\ Research Center Jülich, Institute of Neuroscience and Medicine (INM-8), Ethics in the Neurosciences \\ Alexandra Georgescu \\ University Hospital Cologne, Department of Psychiatry, Cologne, Germany \\ Gary Bente \\ University of Cologne, Department of Psychology, Institute for Social and Media Psychology \\ Kai Vogeley \\ University Hospital Cologne, Department of Psychiatry, Cologne, Germany
}

\begin{abstract}
An intrinsic motivation for social interaction has often been proposed and is thought to be unique to the human species. However, little is known about underlying neural mechanisms. Here, we investigated whether experiencing engagement in social interaction recruits the reward system of the brain. A combined eye-tracking and fMRI paradigm was used in which participants interacted with a virtual agent in a series of gaze-based interactions in real-time. To create situations in which they experience the interaction as social or as non-social, they were made believe that during each block the agent's gaze behavior could either be controlled by another human participant or a computer algorithm. The other participant was a confederate of the experimenter, which enabled experimental control of the agent's gaze reactions. After each block participants had to indicate whether they experienced the interaction as social or not. Results demonstrated that gaze-based interactions with a perceived human partner is associated with activity in the ventral striatum, a core component of reward-related neurocircuitry, while interactions with a computer-driven agent activate attention networks. In addition, the nature of the interaction with a human partner (naive vs. cooperative) differentially modulates striatal activity.
\end{abstract}

\title{
Differences in Interactional Attitudes in Second Language Conversations: From the Perspective of Expertise
}

\author{
Ichiro Umata \\ National Institute of Information and Communications Technology, Seika-cho, Soraku-gun,, Kyoto, Japan
}

Seiichi Yamamoto
Doshisha University, Kyotanabe-shi, Kyoto, Japan
Kosuke Kabashima
Doshisha University, Kyotanabe-shi, Kyoto, Japan
Masafumi Nishida
Doshisha University, Kyotanabe-shi, Kyoto, Japan

\begin{abstract}
We examined the effects of participants' linguistic expertise on their communicative behaviors and their interactional attitudes in conversations in their second language. Quantitative analyses of eye-gazes during utterances showed that the speakers with lower linguistic expertise were observed more by the listeners in the second language conversations, whereas, the listeners' expertise level did not affect the amount of their gazes to the speakers. The analyses of a questionnaire suggested that the participants with lower expertise were not conscious of their own gazing activities; they self-evaluated the amount of their gazes to the speakers' eyes much lower in conversations in their second language than those in their native language. The participants with lower expertise evaluated the pressure they felt higher than those with higher expertise. It is likely that they were too occupied with conducting conversations to maintain enough control over their interaction activities.
\end{abstract}

\title{
Effects of Cultural Knowledge on Parallel Language Activation in German-English Bilinguals
}

\author{
Sebnem Uzuner \\ San Diego State University \\ Henrike Blumenfeld \\ San Diego State University \\ Marcel Giezen \\ San Diego State University
}

\begin{abstract}
Bilinguals have been shown to activate their two languages simultaneously during spoken word recognition (e.g., Blumenfeld \& Marian, 2007). We investigated whether top-down conceptual processing influences parallel language activation using the visual world eye-tracking paradigm. Cultural knowledge was used to manipulate semantic activation in L1 during L2 word processing. Critical trials contained target items that were culturally meaningful to individuals who grew up in Germany, alongside German competitor items that had word-initial phonological overlap with the target word. We hypothesized that German-English bilinguals would fixate the German competitor items more in the culturally salient condition than in the culturally neutral condition. Preliminary data from seven German-English bilinguals and 10 English monolinguals revealed a competitor effect in the culturally salient condition only for bilinguals ( \(\mathrm{p}=.03\) ), with no effect in the culturally-neural condition ( \(\mathrm{p}>.1\) ). Results suggest that activation of cultural knowledge in bilinguals exerts a conceptual top-down influence on parallel language activation.
\end{abstract}

\title{
Stimulating Retrieval During Reading to Improve Word Learning and Memory
}

\author{
Gesa S. E. van den Broek
}

Radboud University Nijmegen, Behavioural Science Institute

\author{
Atsuko Takashima \\ Radboud University Nijmegen, Donders Institute for Brain, Cognition and Behaviour; Radboud University Nijmegen, Behavioural Science Institute \\ \section*{Eliane Segers} \\ Radboud University Nijmegen, Behavioural Science Institute \\ \section*{Ludo Verhoeven} \\ Radboud University Nijmegen, Behavioural Science Institute
}

\begin{abstract}
Retrieving information from memory improves the long-term retention of that information more than continued restudying (e.g., Karpicke \& Roediger, 2008). We investigate if this testing-effect can be applied to word learning during reading by manipulating the sentence context in which words are presented. In a within-subject experiment, adult learners without prior knowledge of Swahili studied 80 Swahili words and then repeatedly read the words either in an L1-context that was uninformative and required the retrieval of word meanings from memory to be understood (e.g., "I use the =funguo=") or in a rich context that enabled the readers to derive word meanings (e.g., "I use the \(=\) funguo \(=\) to unlock the door"). Recall-accuracy and speed for the newly learned words were measured during reading as well as immediately and seven days after practice to evaluate effects of sentence context and retrieval success during reading on the retention of word form and meaning.
\end{abstract}

\title{
When do PDP neural networks learn localist representations?
}

\author{
Ivan Vankov \\ University of Bristol \\ Jeffrey Bowers \\ University of Bristol
}

\begin{abstract}
One of the most distinctive characteristics of the Parallel Distributed Processing (PDP) approach to cognitive modeling is that representations are distributed across a large set of units. However, this is not always the case. In a series of simulations we show that PDP neural networks tend to form localist representations under certain conditions. First, localist representations are developed when the mapping between the input and output patterns is arbitrary. A second pressure to learn localist codes comes from having to keep multiple representations active at the same time. Introducing biologically plausible constraints on the network architecture also fosters developing local codes. Taken together, these findings suggest that the widespread assumption that PDP neural networks learn distributed representations is often wrong. Moreover, exploring the computational reasons for which PDP learn localist representations provides insight into why selective neurons are often found in the brain.
\end{abstract}

\title{
Cognitive Improvement after Surgical Treatment of Moyamoya Disease
}

\author{
Thais Varzoni \\ University of North Florida \\ \section*{C. Dominik Güss} \\ University of North Florida \\ Ricardo A. Hanel \\ Mayo Clinic
}

\begin{abstract}
Moyamoya disease is a rare entity characterized by progressive narrowing of intracranial blood vessels. Moyamoya in most cases does not respond well to medical therapy and often leads to surgical revascularization. The physiological benefits of the surgery for Moyamoya patients have been well documented, yet the effects of surgery on cognitive skills are far less studied. Participants in the current study were 30 patients, 24 to 85 years of age, who underwent surgery and were all treated at Mayo Clinic in Jacksonville, Florida. All patients underwent a physical and cognitive preoperative evaluation, where speech, memory, and intellectual processes were measured. After the surgical intervention, patients returned for 3 follow-up assessments over a period of 6 months. All patients experienced stabilization or improvement of physiological symptoms. Regarding cognitive functions, speech, memory, and intellectual processes improved significantly after surgery. Further prospective studies are needed to better assess cognitive outcomes after revascularization for Moyamoya.
\end{abstract}

\title{
Measuring the Degrees of Separation of a group of Minds
}

\author{
Tomas Veloz \\ University of British Columbia
}

\begin{abstract}
The degrees of separation are essential in the study of social networks. We develop a graph theoretical methodology inspired in this concept to study behavioral data. Having a similarity measure appropriate to the data, we propose to build a graph connecting each agent its nearest \(n\) neighbors, having that their similarity exceeds a threshold p. From here, graph theoretical indicators such as connectivity and clustering can be studied as functions of p and n .

We apply this methodology to a psychological experiment where 97 participants estimate the typicality of 8 objects with respect to the concept "Hat", in 2 contexts. Using correlation distance as a similarity function, we compute clustering properties, and average path length as functions of \(n\) and \(p\). Interestingly, only \(n=20\) connections are required to keep the similarity structure of the system, having an average path length near 2.
\end{abstract}

\title{
Synchronizing to Learn and Like
}

\author{
Roy Vink \\ Behavioural Science Institute, Radboud University, Nijmegen, The Netherlands \\ Anna Bosman \\ Behavioural Science Institute, Radboud University, Nijmegen, The Netherlands \\ Toon Cillessen \\ Behavioural Science Institute, Radboud University, Nijmegen, The Netherlands
}

\begin{abstract}
Ever seen two people walking down the street in the exact same pace? This kind of interpersonal synchrony has been observed in both humans, as well as in animals (e.g., large groups of fireflies flash at the same time, schools of fish and flocks of birds synchronize their movement). For animals it appears to be beneficial (for survival) to synchronize their behavior, but what are the benefits for humans to do so? There are indications that interpersonal synchrony supports social bonding. Previous studies have shown that interpersonal synchrony can have both an effect on (e.g., increases memory), and can be affected by social factors (e.g., higher likeability ratings, more interpersonal synchrony). The goal of the present study was to examine whether social factors (e.g., popularity, friendship) affect interpersonal synchrony when working together. Furthermore, we looked at the relation between interpersonal synchrony and learning and likeability.
\end{abstract}

\title{
The Influence of Motion Language and Gaze Orientation on Spatial Memory
}

\author{
David Vinson \\ University of California, Merced, Merced, California, United States \\ Drew Abney \\ University of California, Merced, Merced, California, United States \\ Teenie Matlock \\ University of California, Merced, Merced, California, United States \\ Rick Dale \\ University of California, Merced, Merced, California, United States
}

\begin{abstract}
When participants observe a man who looks to have fallen or jumped off of a cliff, spatial memories of the man's location are further down a gravitational trajectory compared to his actual location (Freyd, 1985). Previous research has repeatedly observed that inferring gravity - a top-down process - modulates spatial memory. The present study found that differences in the eye gaze orientation of the man by a cliff such as facing up or facing down, and linguistic information such as 'Jumped' or 'Fell', independently influence the spatial memory of the man's location. A brief discussion is presented regarding the influences of top-down and bottom-up processes in spatial memory.
\end{abstract}

\title{
Social gaze orientation influences decision making
}

\author{
David Vinson \\ University of California, Merced, Merced, California, United States \\ Rick Dale \\ University of California, Merced, Merced, California, United States
}

\begin{abstract}
Richardson, Spivey and Hoover (2009) observed that attending to one of two choices in a decision making task increases the likelihood of that response by ten percent. Reflexive gaze orientation, attending to the same area as others, has also been shown to influence reaction times on simple dual choice tasks. The current study addressed how the eye gaze direction of a highly simplified face influences decisions made regarding questions with no obvious answer. Observed results indicate that the appearance of eyes looking toward one answer over another significantly impacts the decision making process. Similar to past work on the impact of social stimuli on orienting, we show that social stimuli can also impact decision making under uncertainty.
\end{abstract}

\title{
Recognition Memory Processes in Novel Noun Generalization
}

\author{
Haley Vlach \\ University of Wisconsin, Madison, Madison, Wisconsin, United States
}

\begin{abstract}
Previous research on word \& category learning has hypothesized that memory processes may be critical to the longterm ability to generalize information. The current experiment was designed to elucidate the recognition memory processes occurring during children and adults' category learning. Participants were presented with a novel noun generalization task and five recognition memory tasks. The results revealed that there were developmental differences in (a) the retention for information presented during category exemplar presentations, and, (b) the sub-categories of recognition memory that were significantly related to category learning performance. These findings suggest the relationship between recognition memory and category learning processes may change across development.
\end{abstract}

\title{
Visitors from outer space: On picking up relevance in inductive reasoning
}

\author{
Wouter Voorspoels \\ University of Leuven, Belgium \\ Gert Storms \\ University of Leuven, Leuven, Belgium
}

\begin{abstract}
In the last decade, numerous studies in inductive reasoning have shown that people are remarkably quick at picking up the relevant features that allow generalizations. One formal account of (some instances of) cases where this ability emerges, relies on the assumptions people make about the manner in which the observations are sampled from the world. According to a weak sampling scheme, observations are randomly drawn from the environment. However, a reasoner can also assume that the observations are sampled deliberately from the intended hypothesis (strong sampling). In the present contribution, we compare these assumptions in a inductive reasoning task using semantic stimuli from 4 superordinate concepts (animals, clothes, vehicles,musical instruments). Model analyses are performed to examine the sampling assumptions of reasoners in this context. We find that people generally assume a strong sampling scheme. However, the model seems to underestimate the relevance sensitivity of the participants.
\end{abstract}

\title{
New methods for assessing ontological and representational changes in the conceptual system of children and adults
}

\author{
Stella Vosniadou \\ National and Kapodistrian University of Athens \\ Despoina Lepenioti \\ National and Kapodistrian University of Athens \\ Anna Choundala \\ National and Kapodistrian University of Athens \\ Kalliopi Eikospendaki \\ National and Kapodistrian University of Athens \\ Petros Papavasiliou \\ National and Kapodistrian University of Athens \\ Emiliana Thanou \\ National and Kapodistrian University of Athens
}

\begin{abstract}
Cognitive developmental research has documented that children acquire rich knowledge about the physical and psychological world before they are exposed to science, and that learning science requires substantial ontological, epistemological and representational changes to happen in the conceptual system of the child. For example, Vosniadou \& Skopeliti (2005) showed that while the majority of third grade children categorized the earth as a physical object distinct from solar objects like the sun and the moon, \(90 \%\) percent of 5 th graders categorized the earth as an astronomical object, belonging to the same category as the sun and the moon. In the present research we will present two novel, chronometric, tasks for assessing the conceptual re-organizations that take place as children are exposed to systematic science instruction and for further exploring the question of whether naive theories are overwritten or survive and continue to exist together with the scientific theories.

References Vosniadou, S., \& Skopeliti, I. (2005). Developmental Shifts in Children's Categorizations of the Earth. In B. G. Bara, L. Barsalou, \& M. Bucciarelli (Eds.), Proceedings of the XXVII Annual Conference of the Cognitive Science Society, Italy, 2325-2330.
\end{abstract}

\title{
Feeling Guilt vs. Regret about Own Decisions: What are the Psychological and Neurobiological Differences?
}

\author{
Ullrich Wagner \\ Charité - Universitätsmedizin Berlin, Department of Psychiatry and Psychotherapy, Division of Mind and Brain Research, Charitéplatz 1, 10117 Berlin, Germany \\ \section*{Lisa Handke} \\ Charité - Universitätsmedizin Berlin, Department of Psychiatry and Psychotherapy, Division of Mind and Brain Research, Charitéplatz 1, 10117 Berlin, Germany \\ Henrik Walter \\ Charité - Universitätsmedizin Berlin, Department of Psychiatry and Psychotherapy, Division of Mind and Brain Research, Charitéplatz 1, 10117 Berlin, Germany
}

\begin{abstract}
Humans use counterfactual thinking in order to evaluate their previous choices, comparing the actual outcome of a choice with what would have happened if they had chosen another action option. Two prototypical emotions resulting from such counterfactual evaluations are guilt and regret, both of which play an important role in regulating human choice behavior. Guilt is thought to refer specifically to social choices, while regret occurs for both individual and social choices. Here, we introduce an fMRI compatible new experimental paradigm to differentially induce guilt and regret under controlled conditions as a result of real decision. Behavioral data confirm that guilt but not regret specifically occurs in a social context (i.e. after harm for another person caused by own choices). On the neural level, initial results point to a critical involvement of different sub-regions within the prefrontal cortex in the processing of guilt vs. regret.
\end{abstract}

\title{
Word Segmentation Without Statistics
}

\author{
Felix Wang \\ University of Southern California \\ Toben Mintz \\ University of Southern California
}

\begin{abstract}
A critical step in language acquisition is segmenting speech into words. Prior cross-linguistic research demonstrates the importance of language specific segmentation strategies (Gervain \& Erra, 2012). English-learning infants can use statistical information to segment words in laboratory experiments (Saffran et al., 1996), but its informativeness in segmenting English is unclear (Yang \& Gambell, 2004; Swingley, 2005). We analyzed one corpus of 611,837 child-directed utterances (Theakston et al, 2000; MacWhinney, 2000), and one corpus of 50,776 adult-directed utterances (Pitt et al., 2007) and found that, in both corpora, a simple strategy that assumes that each syllable is a word would be highly effective in segmenting words. Accuracy was \(69.68 \%\) and \(61.7 \%\) in the child and adult corpora, respectively, segmenting \(87.44 \%\) and \(87.90 \%\) of the corpora. These findings should guide further investigations into the processes infants actually use to segment speech, and more broadly how they learn appropriate language-specific word segmentation strategies.
\end{abstract}

\title{
Statistical Learning in Non-Chinese Speakers Exposed to a Sequence of Words without Spaces
}

\author{
Tsanyu Wang \\ National Taiwan Normal University
}

Adam Li
National Taiwan Normal University
Yeou-Teh Liu
National Taiwan Normal University
Jenn-Yeu Chen
National Taiwan Normal University

\begin{abstract}
A Chinese sentence contains no extra spaces between words. We hypothesize that implicit word segmentation by Chinese readers is developed via the statistical learning mechanism that is universal. Twenty non-Chinese speakers were exposed to a sequence of 3600 characters constructed from six disyllabic words with 6 different characters. The transitional probabilities between any two characters were .46 to 1 within words, and 0 to .29 between words. The sequence was presented one character every half a second. Occasionally, the presentation rate doubled. The participants' task was to detect the instances of double presentation. Upon completing the task, a surprise test followed that consisted of a word and a nonword (reversals of the characters in a word) from the sequence. The participants had to decide which one had appeared before. The averaged accuracy rate was greater than chance (.53). This suggests the statistical learning mechanism is available to all language learners.
\end{abstract}

\title{
Closer than you think?: Options for efficiently approximating optimal analogies under Structure Mapping Theory
}

\author{
Todd Wareham \\ Memorial University of Newfoundland, St. John's, NL, Canada \\ Tijl Grootswagers \\ Radboud University, Nijmegen, The Netherlands \\ Iris van Rooij \\ Radboud University, Nijmegen, The Netherlands
}

\begin{abstract}
Structure Mapping Theory (SMT) is an important and influential theory in cognitive science. One unresolved issue is this theory's intractability. It is known that finding optimally systematic analogies under SMT is NP-hard, making it unrealistic that human analogizing is characterized by optimality. Yet, experimental studies suggest that the optimality assumption gives the best fit to human performance data. A solution to this paradox may be that SMT can be efficiently approximated to such a degree that, for practical intents and purposes, human analogies appear to be optimal. However, outside of a limited empirical evaluation of the commonly-used greedy SME heuristic given in Forbus and Oblinger (1990), no analyses of the approximability of SMT have been done to date. We fill this void by providing both the first theoretical analyses of the types of efficient approximability available for SMT and the first systematic empirical evaluation of the greedy SME heuristic.
\end{abstract}

\title{
Robust social learning to babble in a spiking neural network model
}

\author{
Anne S. Warlaumont \\ University of California, Merced \\ Sophia R. Hyatt \\ University of California, Merced
}

\begin{abstract}
Three humans interacted with a spiking neural network model that controls the lip and jaw muscles of a speech synthesizer. The model learns using spike-timing dependent plasticity that is greater when reinforcement is received. First the humans each selectively reinforced the model, encouraging it to more frequently produce vocalizations with speech-like syllabic elements. Afterwards, for each human-reinforced simulation a yoked control simulation was generated. All the sounds produced during the course of learning for all human-reinforced and all yoked control simulations were then judged on a four-point syllable quality scale by all three listeners (sounds were presented in random order). In all cases, these judgments indicated that the human-reinforced models produced more sophisticated babbling over the course of learning whereas the yoked control simulations did not. The results support this model of how canonical babbling, a major developmental milestone, may develop.
\end{abstract}

\title{
Emotion and Decision-Making: Modeling Strategy Selection
}

\author{
Szymon Wichary \\ Interdisciplinary Center for Applied Cognitive Studies, Warsaw University of Social Sciences and Humanities \\ Tomasz Smolen \\ Department of Psychology, Krakow Pedagogical University
}

\begin{abstract}
Models of decision making neglect the impact of emotions on information processing. Here we present a model of decision strategy selection that can account for the differences in the use of two strategies, the take-the-best heuristic (TTB) and the Weighted Additive rule (WADD) in a probabilistic inference task. How can such a model incorporate emotions? Our model assumes attentional weighting of cues, controlled by the activity of locus coeruleus - a brainstem nucleus associated with physiological arousal. Using hierarchical Bayesian modeling, the model was evaluated on data from a study where participants performed a probabilistic inference task and emotional stress was manipulated with highly aversive slides. For each participant we estimated the parameter that controls cue weighting. This parameter correlated positively with the proportion of choices consistent with TTB and negatively with the proportion of choices consistent with WADD. Moreover, this parameter was sensitive to the emotional stress manipulation.
\end{abstract}

\title{
Television network attitudes toward political candidates implicit in lexical statistics
}

\author{
Jon Willits \\ Indiana University \\ Mark Seidenberg \\ University of Wisconsin Madison
}

\begin{abstract}
Languages exhibit statistical regularities concerning the frequencies and co-occurrences of words. Language users learn from such patterns without being consciously aware of them. We investigated statistical properties of the language used on television news in discussing politicians. We compiled corpora consisting of language used on four networks (MSNBC, ABCNews, CNN, FOXNews) from 2007-2012. We analyzed the frequencies with which 500 affectively-valenced words cooccurred with politicians' names (Obama, McCain, Romney) during the run-ups to the 2008 and 2012 elections. We used these co-occurrences to derive a summary measure, their net positivity score. Positivity scores for candidates changed over time in ways that reflect real-world events. Positivity towards candidates differed across networks. Net positivity toward President Obama during his first term was strongly correlated with approval ratings. The results show that statistical aspects of language, of which people are not consciously aware, convey varying attitudes on network news.
\end{abstract}

\title{
The effect of accent on the perception of American and British English fluent speech in Chinese second language learners
}

\author{
Simpson WL Wong \\ The Hong Kong Institute of Education \\ Peggy Mok \\ The Chinese University of Hong Kong
}

Vina W.H. Leung
The Hong Kong Institute of Education

\author{
Kevin Kien-Hoa Chung
}

The Hong Kong Institute of Education

\section*{Bonnie Wing-Yin Chow}

City University of Hong Kong
Michael CW Yip
The Hong Kong Institute of Education

\begin{abstract}
Spoken words are phonologically reduced through various processes (e.g., assimilation) in fluent speech. The reduced forms of words give rise to perceptual difficulties for nonnative speakers. This study examined whether accent types (General American and Received Pronunciation (PR) British Englishes) of native English fluent speech affect fluent speech perception in Chinese speakers. A representative sample of sixty undergraduate students were tested with listening comprehension tests (recordings produced by American and British English speakers), reduced forms dictation test (with Americanand British-accented speech as stimuli) and fluent speech production task. Based on correlational analyses, it is shown that listening comprehension was significantly correlated with both fluent speech perception and production skills among the Chinese speakers. Importantly, these correlations were observed within the same accent type and across the two accents. However, our regression analyses showed that speech perception rather than production significantly predicted the outcome of listening comprehension.
\end{abstract}

\title{
The deficits of visual encoding processes in Chinese Dyslexics
}

\author{
Orieta H. Y. Wong
}

Department of Psychological Studies, The Hong Kong Institute of Education, HKSAR
Michael C. W. Yip
Department of Psychological Studies, The Hong Kong Institute of Education, HKSAR

\begin{abstract}
The present research aimed at identifying cognitive processing that works at the initial stage of letter identification. First, we used the word superiority effect task to set the baseline of visual detection rate and then we used the letter matching task to evaluate the phonemic transference rate of the particpants. It was hypothesized that physically-different letters required longer processing time as the judgment depended not only on visual detection, but also on a process of phonemic transference. Our results found that the control group had a higher letter identification rate than the dyslexic group, showing that visual detection rate was a good predictor of reading disability. However, phonemic transference rate was not a good predictor of dyslexia. One of the reasons may be that the process of letter identification involved cognitive skills other than that of phonemic transference alone.
\end{abstract}

\title{
Inflection from form versus meaning: Developmental and computational evidence
}

\author{
Anna Woollams \\ School of Psychological Sciences, University of Manchester \\ Joanna Moy \\ School of Psychological Sciences, University of Manchester \\ Sarah Keeble \\ School of Psychological Sciences, University of Manchester \\ Grzegorz Krajewski \\ Faculty of Psychology, University of Warsaw \\ Anna Theakston \\ School of Psychological Sciences, University of Manchester
}

\begin{abstract}
Inflectional morphology has proven a test case for evaluating theoretical approaches to language processing. The bulk of empirical data has been acquired using the inflection from stem task, which approximates more naturalistic speech production only if mandatory stem retrieval is assumed. Yet work with adults reveals quite different results when inflection proceeds instead from meaning. Connectionist computational models of inflection have simulated these task differences, but the extent to which they accurately reflect development remains unknown. For the first time, we consider the impact of task type upon inflectional development of 908 children aged between 2 and 4 years. We then present a revised version of a connectionist model of inflection, which has been trained in a manner consistent with child directed speech in order to capture the children's performance. Taken together, this work demonstrates the progression of the field to more ecologically and developmentally valid approaches to inflection.
\end{abstract}

\title{
Mnemonic convergence in a social network:
}

\author{
Jeremy Yamashiro \\ New School for Social Research, New York, NY, United States \\ William Hirst \\ New School for Social Research, New York, NY, United States
}

\begin{abstract}
Selective rehearsal of memories through conversation shapes subsequent recall for both speakers and listeners. Interactional memory acts have typically been studied in dyadic conversation or small groups; however, as Christakis and Fowler (2009) point out, humans don't just belong to groups, we are more precisely enmeshed in social networks. This study extends Coman and Hirst's (2012) work on propagation of socially-shared retrieval-induced forgetting (Cuc, Koppel, and Hirst, 2007) and social contagion (Roediger, Meade, \& Bergman, 2001) across a series of conversations. We move from dyadic networks to a dynamically self-assembling social network of 46 undergraduates over a nine-week period. Using a framework derived from Kauffman's (1994) models of adaptive fitness landscapes in evolution, we trace mnemonic convergence as it emerges from the interplay of "internal" cognitive factors and dynamics attributable to network connectivity. Contributions to the cognitive psychology of collective memory are discussed.
\end{abstract}

\title{
How to rectify the confirmation bias in Wason(1960)'s 2-4-6 task
}

\author{
Tomohito Yamazaki \\ Keio University \\ Mutsumi Imai \\ Keio University
}

\begin{abstract}
People have a strong confirmatory bias, which is remarkably difficult to overcome. In this research, we investigated how we could help people rectify the confirmatory bias, using Wason(1960)'s 2-4-6 task. 195 University and middle school students participated in our study, in which they were to find out the rule "increasing 3 numbers," starting from the 2-4-6 number sequence. Prior to the task, the participants were assigned to one of the four training conditions: (1)"Counter-Examples", in which participants were instructed to create the positive and counter-examples for the current hypothesis; (2) "Two-Hypotheses", in which participants were instructed to create examples for two different hypotheses; (3) "Counter-Examples-AND-TwoHypotheses; (4) Control, in which participants were told to think about examples consistent with their hypothesis. We found that the training to think counter-examples facilitates attempts to falsify the current hypothesis and subsequent hypothesis change, but creating examples for two different hypothesis did not.
\end{abstract}

\title{
Computational modeling of inference generation during reading comprehension
}

\author{
Meni Yeari \\ Bar-Ilan Univeristy, Ramat-Gan, Israel \\ Paul van den Broek \\ Leiden University
}

\begin{abstract}
This research presents a computational model that simulates inference generation during reading comprehension. Inferences refer to information that readers generate from their background knowledge in order to clarify, connect, and elaborate textual information. The computational model integrates Latent Semantic Analysis (LSA), which simulates general knowledge by computing the strength of semantic association between concepts (Landauer \& Dumais, 1997), with the Landscape Model, a dynamic model of reading comprehension that simulates fluctuations of concepts' activation and the emergence of episodic connections between them (Yeari and van den Broek, 2011). The extended model was used to simulate behavioral data from a large number of studies on inference generation. Successful simulations of the various findings demonstrate the unique roles of semantic associations, episodic inter-textual relations, and working memory (limitation of concepts' activation sum) in the activation of different types of inferences (i.e., elaborative and bridging inferences).
\end{abstract}

\title{
Argument order as an expectation trigger in Korean
}

\author{
Hongoak Yun \\ Konkuk University \\ Uphong Hong \\ Konkuk University \\ Yun-Ju Nam \\ Konkuk University \\ Hyunjung Kim \\ Konkuk University
}

\begin{abstract}
Expected words/constituents are often processed faster than unexpected words/constituents. In languages like English that verbs are placed before arguments that they encode, the occurrence of verbs pre-activates upcoming arguments, suggesting that verbs' arguments are expected upon the recognition of verbs. However, this is not the case for languages like Korean that verbs are placed after the arguments associated with verbs. This study investigated whether argument order might play a role in cuing upcoming arguments. Using Hong et al.'s materials, we conducted a completion study. We found that patients were expected when recipients were introduced before, whereas recipients were not expected when patients were introduced before, suggesting that comprehenders might expect encountering patients after recipients but not vice versa. The probability of patients/recipients was correlated with the frequency of regressions that Hong et al. observed. We will further examine the role of probability and uncertainty in processing in terms of expectation.
\end{abstract}

\title{
Social interaction and cognition: from L. Vygotsky and M. Bakhtin tradition to computer agents' interactivity
}

\author{
Vera Zabotkina \\ Russian State University for the Humanities \\ Elena Pozdnyakova \\ Russian State University for the Humanities
}

\section*{Artemy Kotov}

Russian Research Centre "Kurchatov Institute"

\begin{abstract}
Social-interactive cognition in Russia builds on the ideas of cultural-historical psychology by L. Vygotsky, A. Leontyev, A. Luria, as well as communicative aspect of cognition introduced by M. Bakhtin in his theory of dialogism which, in its turn, laid the ground for the studies of social and situated cognition. Since 1980-s cognitive-discursive paradigm in Russia, introduced by E. Kubryakova, has attracted attention of cognitive science scholars. Within this paradigm cognitive approaches to pragmatics of communication and agentivity have been developed (V.Zabotkina, E. Pozdnyakova).

The given study introduces further development of the field, based on a multi-modal corpus built with recorded live communication data. Presently it's the biggest multi-modal corpus of emotional communication, so it ensures completeness and accuracy of research, conducted on its basis. Two- and three-dimensional computer agents (A. Kotov) operate a set of speech patterns in a complex of restricted communication situations, simulating real-life emotional social interaction.
\end{abstract}

\title{
Metacognitive Judgments in Category Learning
}

\author{
Valnea Zauhar \\ University of Rijeka \\ Igor Bajšanski \\ University of Rijeka \\ \section*{Drazen Domijan} \\ University of Rijeka
}

\begin{abstract}
The aim of the study was to investigate confidence in classification accuracy during category learning with three levels of categorization difficulty based on a simple logical, conjunctive or complex rule. Twenty five psychology students participated in the study. Stimuli were geometrical figures that varied on three dimensions: shape, colour, and size. We analysed differences in absolute accuracy of confidence judgments with respect to task difficulty (three levels), learning phase (early, late) and performance groups (slow and fast learners). For the simple logical task, we obtained a significant main effect of phase and a significant interaction between phase and group, suggesting that fast learners achieved higher accuracy earlier. For the conjunctive task, the significant main effect of the group showed higher accuracy of fast learners. For the complex task, there was a significant interaction between phase and group showing that judgment accuracy in slow learners did not improve during learning.
\end{abstract}

\title{
Effects of Strategy and Rule Complexity on Multivariable Inductive Judgments
}

\author{
Corinne Zimmerman \\ Illinois State University
}

\author{
Steve Croker
}

Illinois State University

\begin{abstract}
We investigated the effects of strategy use and rule complexity on multivariable inductive judgments. Participants \((\mathrm{N}=274)\) made judgments about which of two cars presented on a computer screen was faster. Participants were randomly assigned to a complex rule or a simple rule. For the complex rule, three of five variables affected speed; for the simple rule one variable affected speed. Participants were instructed to make explicit (try to discover the rules governing speed) or implicit (speeded intuitive) judgments for 300 trials with feedback.

A 2 (complexity) x 2 (strategy instruction) ANOVA revealed main effects of complexity, \(\mathrm{F}(1,270)=6.17, \mathrm{p}=.014\), and task instruction, \(\mathrm{F}(1,270)=11.69\), \(\mathrm{p}=.001\), and a significant interaction, \(\mathrm{F}(1,270)=7.38, \mathrm{p}=.007\). The explicit strategy led to better performance for the simple rule only, but no differences were found for the implicit strategy. These findings run counter to recent work showing an advantage for implicit processing of complex rules (Zimmerman \& Pretz, 2012).
\end{abstract}

\title{
Difference in Single vs. Pair judgements on Deception Detection, Confidence and Bias based on the Level of Communication
}

\author{
Mircea Zloteanu \\ University College London, London, United Kingdom \\ Daniel Richardson \\ University College London
}

\begin{abstract}
When people judge whether others are telling the truth, they act differently if they are working alone or in a group. The current experiment explored this finding. Participants (working alone or in pairs) provided either a binary truth/lie decision, or a binary decision and a set of reasons chosen from a list, or an open ended discussion/explanation. Being alone or in a pair had no significant effect on accuracy, but confidence was higher in pairs. A truth bias was found in the single condition but was eliminated for pairs when they specified a reason or had a discussion. Accuracy was highest when stating a reason chosen from a list, while confidence increased with the amount of information provided. These findings improve our understanding of the effect of pair decision making, illustrating how varying levels of information can have different effect on decision making and deception detection.
\end{abstract}

\title{
Progressive Development of the Number Sense in a Deep Neural Network
}

\author{
Will Y. Zou \\ Stanford University \\ James L. McClelland \\ Stanford University
}

\begin{abstract}
What are the developmental bases of the number sense? This ability could arise through evolution or experience. Stoianov \& Zorzi (2012, Nature Neuroscience, 8, 194-196) showed that a neural network could learn number sense from visual examples containing varying numbers of elements. However, the layer-wise training regime is unrealistic from a developmental standpoint. A key observation is that number acuity progressively develops from infancy to adulthood (as reflected by a decreasing Weber fraction). This development involves accumulation of single examples, each of which updates the connection weights in a hierarchical system. We present an unsupervised deep network that learns all weights as it observes one 'number example' at a time. As on-line training progresses, neurons representing numerosity start to emerge in the deeper layers, and the Weber fraction progressively sharpens. These results establish that a generic learning algorithm in a deep network gives rise to a clear developmental trajectory of the number sense.
\end{abstract}

\title{
On the different types of information stored in the lexicon and their neural bases
}

\author{
Camila Zugarramurdi \\ Facultad de Psicología, Universidad de la República, Uruguay.
}

\author{
Álvaro Cabana \\ Facultad de Ciencias, Universidad de la República, Uruguay. \\ Juan Valle-Lisboa \\ Facultad de Ciencias, Universidad de la República, Uruguay.
}

\begin{abstract}
We explore the idea that the neuropsychological organization of memory explains contradictory results about the lexicon. We believe that a semantic component stores the prototypic information and an 'episodic' component stores both the exceptions to the prototypes and the exceptionally common stimuli. With this theoretical insight we studied the different types of information that are accessed when a word is presented to a subject. To this end we first reproduced some results from Hare and coworkers showing that the presentation of a word for a noun facilitates the recognition of words related to the context of the word usage more than what could count as a semantic definition. The same pattern is observed for the neurophysiologically determined facilitation of the N400 component. Some of the results can be explained by corpus linguistic tools such as LSA. We present evidence from bimodal priming experiments supporting part of our theoretical proposition.
\end{abstract}

\section*{Author Index}
Ab Rashid, Ahmad Azad ..... 115
Abbott, Joshua ..... 1686
Abdallah, Samer ..... 127
Abdel Rahman, Rasha 91, 3868, 4034, 4088,4090, 4124Abe, Keiga ................................... 1692
Abney, Drew 3864, 3940, 4153
Abrahamsen, Adele ..... 1839
Abramova, Ekaterina ..... 1696
Acartürk, Cengiz ..... 1714, 1774
Ackerman, Rakefet ..... 121, 2838
Actis-Grosso, Rossana ..... 2058
Aczel, Balazs ..... 3865
Adams, Deanne ..... 1702
Adelman, James ..... 3925
Admoni, Henny ..... 1708
Agogué, Marine ..... 3321
Agres, Kat ..... 127
Aha, David ..... 3229
Ahlberg, Daniela ..... 133
Aitken, M.R.F. ..... 3829
Akama, Hiroyuki ..... 4021
Akita, Kimi ..... 1253
Alacam, Ozge ..... 1714
Alan, Cienki ..... 3881
Aleven, Vincent ..... 1205
Alexenko, Tatiana ..... 281
Ali, Nadia ..... 1720
Alishahi, Afra ..... 2996
Almeda, Victoria ..... 2428
Althaus, Nadja ..... 1726, 2410
Altmann, Gerry T. M. ..... 834
Amaral, Joseph ..... 1732
Ambridge, Ben ..... 1498
Ameel, Eef ..... 2961
Aminihajibashi, Samira ..... 3480
Andermane, Nora ..... 3866
Anderson, John ..... 221, 3891
Anderson, Nicola ..... 3801
Anderson, Richard ..... 3867
Anderson, Sarah ..... 3980
Andonova, Elena ..... 1738
Andraszewicz, Sandra ..... 1744
Andrews, Jan ..... 3916
Andrews, Mark ..... 1750
Andrighetto, Giulia ..... 1756
Anelli, Filomena ..... 2944
Angerer, Katharina ..... 1020
Apel, Jens ..... 4116
Argel, Zarah ..... 2482
Argyriou, Paraskevi ..... 1762
Aristei, Sabrina ..... 3868
Armstrong, Blair ..... 139, 2826
Arnau, Eric ..... 1768
Arnon, Inbal ..... 75
Arslan Aydın, Ülkü ..... 1774
Asami, Kazuaki ..... 3516
Asano, Rie ..... 1199
Ashitani, Yui ..... 3869
Asplund, Christopher ..... 3870
Assimakopoulos, Stavros ..... 145
Asterhan, Christa ..... 1780, 3871
Atanasov, Dimitar ..... 3872
Atanassova-Trifonova, Maria ..... 3872
Atit, Kinnari ..... 1786
Au, Ricky K. C. ..... 151
Au, Terry Kit-fong ..... 1492
Aue, William ..... 3873
Aussems, Suzanne ..... 3874
Avraamides, Marios ..... 466
Axelsson, Emma ..... 1792
Ayala, Saray ..... 1768, 1798
Aylett, Ruth ..... 2189
Bååth, Rasmus ..... 1803
Baayen, Harald ..... 1193, 3245, 3929
Babichenko, Miriam ..... 3871
Bago, Bence ..... 3865
Bahrami, Bahador ..... \(3876,3940,4033\)
Bai, Hao ..... 4058
Bajo, Maria Teresa ..... 4140
Bajšanski, Igor ..... 3875,4173

\section*{Author Index}
Baker, Ryan ..... 2094, 2428
Beckage, Nicole ..... 37, 163
Bakhtiari, Giti ..... 734
Balkenius, Christian ..... 1115
Ball, Linden ..... 3771
Ballard, Dana ..... 680, 686
Balsters, Martijn ..... 4082
Bang, Dan ..... 3876
Bangerter, Adrian ..... 3002
Banks, Adrian ..... 3877
Bann, Eugene ..... 1809
Bannard, Colin ..... 2368
Bapi, Raju ..... 4107
Bara, Bruno ..... 65, 3908, 4057
Barak, Libby ..... 1815
Barbosa, Adriano ..... 2832
Barisic, Iva ..... 4008
Barkoczi, Daniel ..... 1821
Barner, David ..... 1462, 2734, 3717
Barnes, Jordan ..... 3993
Barnwell, Brendan ..... 3878
Baroni, Giulia ..... 2944
Barr, Dale ..... 215
Barraza, PauloBarrett, Trevor157
Barsbey, Melih ..... 3879
Bartolotti, James ..... 1827, 4111
Bassetti, Benedetta ..... 3880
Bassok, Miriam ..... 388
Basu, Prajit ..... 4107
Bates, Joseph ..... 1109
Batres, Karla ..... 1934
Battaglia, Peter ..... 3426, 3959
Battersby, Stuart ..... 2506
Baumann, Peter ..... 1833
Beach, Andrew ..... 3309
Becchio, Cristina ..... 3908
Bechlivanidis, Christos ..... 71, 2386
Bechtel, William ..... 20, 1839
Beck, Aryel ..... 1845
Beck, J. Gayle ..... 2955
Becker, Raymond ..... 1851, 3950
Becker, Suzanna ..... 3975
Beekhuizen, Barend ..... 1857
Begolli, Kreshnik ..... 1863
Behnel, Marieke ..... 1869
beim Graben, Peter ..... 109
Bekkering, Harold ..... 4046
Bekolay, Trevor ..... 169
Belardinelli, Anna ..... 1875
Belenky, Daniel ..... 1881
Beller, Sieghard ..... 67, 2602
Bemis, Douglas Knox ..... 63
Ben-Asher, Noam ..... 175
Bender, Andrea ..... 67, 2602
Bente, Gary ..... 4145
Bentz, Anne ..... 2602
Bentz, Christian ..... 181, 579
Bergelson, Elika ..... 187
Bergen, Benjamin ..... 924, 972, 3729
Bergmann, Evelyn ..... 990
Bergmann, Kirsten ..... 823
Bergmann, Till ..... 1887
Bergstra, James ..... 633
Bertleff, Sabine ..... 1946
Bertolotti, Tommaso ..... 954
Bērziņa, Krista ..... 4114
Besold, Tarek Richard ..... 1893
Best, Wendy ..... 3928
Bestgen, Anne-Kathrin ..... 193
Betts, Shawn ..... 3891
Bhatia, Sudeep ..... 1899
Bibyk, Sarah ..... 846
Bicknell, Klinton ..... 197, 3927
Bien-Aime, Ralph ..... 2680
Bietti, Lucas ..... 3881
Bijvoet-van den Berg, Simone ..... 3882
Bilalic, Merim ..... 203
Bird, Geoffrey ..... 555
Bischof, Walter ..... 3801

\section*{Author Index}
Biswas, Gautam ..... 209
Brennan, Susan ..... 55, 1934
Black, John ..... 888, 4031
Blair, Mark ..... 3993
Blascovich, Jim ..... 770
Blokpoel, Mark 53, 3883
Brice, Līva ..... 4114
Briggs, Gordon ..... 3657
Bright, Aimee Kay ..... 239
Blouw, Peter ..... 633, 1905
Bromme, Rainer ..... 3092, 3448Blumenfeld, Henrike4147
Brook, Andrew ..... 1940
Brooks, Patricia ..... 752
Bobek, Eliza ..... 1911
Brookshire, Geoffrey ..... 245
Bocklisch, Franziska ..... 1917
Bocklisch, Steffen ..... 1917
Bod, Rens ..... 1676
Bodemer, Nicolai ..... 1923
Bögels, Sara ..... 215
Bojārs, Uldis ..... 4114
Boldt, Annika ..... 3884
Bonardi, Jean-Philippe ..... 3795
Bonneaud, StephaneBorghi, Anna2944
Borisova, Milena ..... 2578
Boroditsky, Lera ..... 3575
Borst, Grégoire ..... 2464
Borst, Jelmer ..... 221
Boschetti, FabioBosco, Andrea3360
Bosco, Francesca ..... 4057
Bosman, Anna ..... 4152
Bott, Lewis ..... 227, 3569
Bottini, Roberto 233, 495
Botvinick, Matthew ..... 83
Bowditch, W. ..... 3885
Bowers, Jeffrey ..... 3866, 3920, 4149
Boyd-Graber, Jordan ..... 3193
Bozcuoglu, Asil Kaan ..... 3604
Bozşahin, Cem ..... 2722
Braibanti, Paride ..... 3116
Braithwaite, David ..... 1928
Bramley, Neil ..... 3886
Brattico, Elvira ..... 99
Brauer, Jens ..... 4103
Brennan, Allison ..... 3887
Brown, Bruce ..... 3098
Brown, Gordon ..... 89
Brown, Meredith ..... 846
Brown, Scott ..... 3581
Bruckner, Dietmar ..... 4099
Brunninkhuis, Niek ..... 1139
Brunstein, Angela ..... 3888
Brunstein, Joerg ..... 3888
Bryson, Joanna ..... 1809
Bublitz, Dennis ..... 752, 3889
Bucciarelli, Monica ..... 3663
Bucher, Leandra ..... 1067, 1946, 3038
Buchsbaum, Daphna ..... 69, 1593, 2584
Buckingham, Steven ..... 3890
Buckmann, Marcus ..... 2273
Buehner, Marc ..... 71, 115
Bülthoff, Heinrich ..... 3915
Burling, Joseph ..... 1635
Burmester, Juliane ..... 251
Burns, Bruce ..... 257, 4001
Burns, Nicholas ..... 1587
Burnston, Daniel ..... 1952
Burte, Heather ..... 1958
Busemeyer, Jerome 45, 73, 101, 1575, 1599,2572
Buss, Aaron ..... 4094, 4115
Butcher, Jessica ..... 591
Butcher, Kirsten 2136, 3913, 3930, 3932
Butler, Lucas ..... 263
Buttaccio, Daniel ..... 864, 1964
Butterfill, Stephen ..... 3937
Butz, Martin ..... 1875, 2926
Buwalda, Trudy ..... 3891

\section*{Author Index}
Byrne, Ruth ..... 1970, 3956
Chaplot, Devendra Singh ..... 1294
Cabana, Álvaro ..... 3892, 4177
Çağlltay, Kürşat ..... 1774
Calamaro, Shira ..... 269
Caliskan, Yigit ..... 3604
Call, Joseph ..... 57
Cañamero, Lola ..... 1845
Carassa, Antonella ..... 3116
Cardella, Valentina ..... 275
Carey, Susan ..... 105
Carlson, Elizabeth ..... 4089
Carlson, Laura ..... 281
Carminati, Maria Nella ..... 1976
Carnevale, Peter ..... 770
Carr, Thomas ..... 3918
Carrara, Giovanna ..... 4057Carreiras, Manuel3942, 4075
Carroll, Christopher ..... 287
Carstensen, Alexandra ..... 293, 3893
Carvalho, Paulo ..... 1982, 1988, 3747
Casasanto, Daniel ..... \(233,245,495,2213\)Casillas, Marisa299, 3486
Cassotti, Mathieu ..... 2464, 3321
Castelhano, Monica ..... 2832
Castillo, Ramon ..... 966, 3894
Catmur, Caroline ..... 555
Catrambone, Richard ..... 978
Cavicchio, Federica ..... 305
Chabal, Sarah ..... 3895, 4111
Chai, Hui Yih ..... 1994
Chambers, Alexandra ..... 1414
Chan, Antoni ..... 328
Chan, Ho Leung ..... 2006
Chan, Jennifer ..... 2000
Chan, Margaret ..... 3896
Chandramouli, Suyog ..... 4013
Chang, Franklin ..... 657, 1498
Chang, Li-Yun ..... 3906, 3907
Chang, Ya-Ning ..... 310, 3898
Chang, Yin-Juei ..... 3934
Chater, Nick ..... 59, 65, 69
Chatterjee, Anjan ..... 829, 4101
Chaudhry, Shereen ..... 1899
Chee, Michael W.L. ..... 3870
Cheimariou, Spyridoula ..... 2012
Chemla, Emmanuel ..... 227
Chen, Dawn ..... 2028
Chen, Hsueh-Chih ..... 3978
Chen, Jenn-Yeu ..... 4160
Chen, Jing ..... 2022
Chen, Sau-Chin ..... 3899
Chen, Stephanie ..... 316
Chen, Yi-Chen ..... 4018
Chen, Yung-Jung ..... 3900
Chen, Zi-Han ..... 3899
Cheng, Patricia ..... 2034
Cherng, Rong-Ju ..... 3900
Chernyak, Nadia ..... 2040
Chia, Wei Wen ..... 2344
Chiang, Yu-Shu ..... 3978
Chiba, Itsuki ..... 4142
Chien, Liang-Chen ..... 4018
Chipman, Susan ..... 3902
Chiu, Eric ..... 3903
Cho, Kit ..... 3929
Cho, Kwangsu ..... 4109
Choe, Wonhee ..... 2046
Choi, Ben ..... 2000
Chong, Amy ..... 2000
Chong, Irina ..... 2000
Choo, Xuan ..... 322
Choundala, Anna ..... 4157
Chow, Bonnie Wing-Yin ..... 4165
Chow, Dian ..... 3904
Chris, Eliasmith ..... 412
Christiansen, Morten ..... \(75,103,692\)
Chrusch, Courtney ..... 3905
Chu, Mingyuan ..... 1127
Chuderska, Anna ..... 2052

\section*{Author Index}
Chuderski, Adam ..... 2052
Corlis, Timothy ..... 4077
Chuk, Tim ..... 328
Chun, Hyo-Sun ..... 2046
Chung, Kevin Kien-Hoa ..... 4165
Chung, Yi-Ling ..... 3906, 3907, 3978
Churchland, Patricia ..... 734
Ciaramidaro, Angela ..... 3908, 4057
Ciardo, Francesca ..... 2058
Cillessen, Toon ..... 4152
Cipollini, Benjamin ..... 334
Claessens, Irene ..... 4039
Clark, Alexander ..... 2064
Clark, Dav ..... 2070
Clark, Douglas ..... 2728
Clarke, Alasdair ..... 2237
Clarke, Richard ..... 1682
Coco, Moreno ..... 24
Coenen, Anna ..... 22, 3909
Coey, Charles ..... 2076, 3309, 3741
Cohen-Eliyahu, Noa ..... 1780
Cohen, Andrew ..... 2572
Cohen, William ..... 894
Cohn, Neil ..... 63
Coley, John ..... 3645Colle, Livia3908
Colling, Lincoln ..... 2082
Collins, Susan ..... 1732
Collisson, BeverlyColombetti, Marco3116
Colombo, Matteo ..... 87
Colunga, Eliana ..... \(37,61,163,1324,1468\),
3408
Coman, Alin ..... 77
Conaway, Nolan ..... 2088
Conte, Rosaria ..... 1756
Cook, Anne ..... 3913
Cook, John ..... 89
Cooper, Joseph ..... 680
Cooper, Richard ..... 1354Coppola, Marie1223
Corbett, Albert ..... 2094
Cornish, Hannah ..... 340
Corter, James ..... 3593
Cosentino, Erica ..... 111
Cottrell, Garrison ..... 334, 3187, 3735
Coviello, Emanuele ..... 328
Cox, Gregory ..... 346, 704, 2100
Craig, Stewart ..... 918
Cravalho, Patrick ..... 352
Creel, Sarah ..... 2106, 2333
Criss, Amy ..... 3873
Croker, Steve ..... 3911, 4174
Cruz de Echeverría Loebell, Nicole ..... 2112
Culbertson, Jennifer ..... 2118
Cummins, Chris ..... 1869
Cunningham, William ..... 722
Curio, Cristobal ..... 3915
Cushman, Fiery ..... 3969
Cutica, Iaria ..... 3663
Cutting, James ..... 3950
Czienskowski, Uwe ..... 43
Dai, Junyi ..... 4096
Dale, Rick \(24,79,97,1121,1682,1887\),3492, 3864, 4076, 4129, 4153, 4154
Dale, William ..... 3966
Damian, Markus ..... 1366
Danek, Amory ..... 3912
Daniel, Natalia ..... 1827
Danielson, Robert ..... 2914
Danks, David ..... 69, 2124
Darby, Kevin ..... 2130
Das, Arindam ..... 358
Davelaar, Eddy ..... \(549,864,1964\)
Davenport, Jodi ..... 364
Davies, Sarah ..... 2136, 3913
Dayan, Peter ..... 83
Dayton, Andrew ..... 3914
de Baessa, Yetuli ..... 3957
de Boer, Bart ..... 3669
De Deyne, Simon ..... 2142, 2542
de Gardelle, Vincent ..... 3884

\section*{Author Index}
de la Rosa, Stephan ..... 3915
Do, Kyung Soo ..... 3919
de Leeuw, Josh ..... 3916
de Melo, Celso ..... 770
De Neys, Wim ..... 3321
De Ruiter, Jan ..... 1869
Debowski, Lukasz ..... 3933
Dębska, Agnieszka ..... 3917
DeCaro, Daniel ..... 370
DeCaro, Marci ..... 370, 3633
Dechter, Eyal ..... 1342
DeCot, Bridgette ..... 1851
Degen, Judith ..... 376, 3835
Dehghani, Morteza ..... 41, 3354
Deiglmayr, Anne ..... 382
Deijkers, Lian ..... 3621
Delevoye-Turrell, Yvonne ..... 2207, 2872
Delfabbro, Paul ..... 1587
Demberg, Vera 1581, 2148, 2154, 2160,2249, 2760
Demestre, Josep ..... 4029
Deng, Wei (Sophia) ..... 2166
Denkiewicz, Michał ..... 2172
Dennis, Simon ..... 3199
Denrell, Jerker ..... 2177
deSouza, Juliana Avila ..... 4056
Desrochers-Pare, Jessica ..... 4138
Devereux, Barry ..... 746
Dewald, Andrew ..... 2183
Dewar, Michael ..... 1378
Dewey, John ..... 3918
DeWolf, Melissa ..... 388
Dias, João ..... 2189, 2978
Dickmann, Frank ..... 193
Diederich, Adele ..... 73
Dienelt, Lena ..... 537
Dienes, Zoltan ..... 103
Diermeier, Daniel ..... 3343
Dimas, Joana ..... 2195
Dimov, Cvetomir ..... 2201
Dione, Mariama ..... 2207
Dixon, Kyle ..... 4050
Doane, Stephanie ..... 4058
Døjbak Håkonsson, Dorthe ..... 79
Dolscheid, Sarah ..... 2213
Domahs, Frank ..... 1247
Domijan, Drazen ..... 4173
Dorfman, Nimrod ..... 394
Dougherty, Michael ..... 549
Doumas, Leonidas ..... 2476, 2902
Driver, Charles ..... 2219
Drooghmans, Stephanie ..... 3986
Dudschig, Carolin ..... 133, 2225
Duffy, Sam ..... 2231
Dumay, Nicolas ..... 3920
Dunn, Timothy ..... 3921
Duran, Nicholas ..... 97, 4129
Dutilh, Gilles ..... 3922
Dye, Melody ..... 3245
Dziemianko, Michal ..... 2237
Ebbets, Elizabeth ..... 2392
Écalle, Jean ..... 4035
Echterhoff, Gerald ..... 77
Ecker, Ullrich ..... 89, 400
Edelson, Micah ..... 77
Edler, Dennis ..... 193
Edmiston, Pierce ..... 2243
Edwards, Brian ..... 406
Eerland, Anita ..... 3923
Ehlis, Ann-Christine ..... 2608
Eikospendaki, Kalliopi ..... 4157
Eisner, Frank ..... 3961
El-Haddad, Rita ..... 3994
Eliasmith, Chris \(39,169,322,633,1905\),
2590, 3257, 3460
Ellefson, Michelle ..... 1181
Ellis-Davies, Kate ..... 3384, 3968
Elqayam, Shira ..... 101
Elstein, Arthur ..... 3966
Elvevåg, Brita ..... 109
Engbert, Ralf ..... 4100
Engelen, Jan ..... 3923

\section*{Author Index}
Engelhardt, Paul ..... 4045
Engonopoulos, Nikolaos ..... 2249
Enns, James ..... 3887, 4077
Erdfelder, Edgar ..... 4036
Erdmann, Michael ..... 1205
Eric, Crawford ..... 412
Erickson, Shanna ..... 2255
Ernst, Marc ..... 71
Escudero, Paola ..... 418
Eskine, Kendall ..... 3994
Estes, Zachary ..... 3925
Etxaniz, Aitziber ..... 2746
Evans, Jonathan St. B.T. ..... 101
Falke, Jesse ..... 2261
Falzone, Alessandra ..... 275
Farber, Ilya ..... 3926
Farmer, Thomas ..... 3927
Faure, Sylvane ..... 4116
Fausey, Caitlin ..... 669, 1988
Favela, Luis ..... 1510
Fazly, Afsaneh ..... 1073, 1815, 1857
Fedor, Anna ..... 3928
Fedyk, Mark ..... 424
Fedzechkina, Maryia ..... 430
Feeney, Aidan ..... 239, 1474
Feil-Seifer, David ..... 1708
Feldman, Laurie Beth ..... 26, 3929
Feldman, Naomi ..... 3193
Felipe, Jacqueline ..... 2070
Ferdinand, Vanessa ..... 436
Ferguson, Brock ..... 2267
Ferguson, Mark ..... 3966
Fernández, Raquel ..... 1696
Ferrara, Katrina ..... 3931
Ferrara, Lisa ..... 3930, 3932
Ferrer-i-Cancho, Ramon ..... 3933
3934, 3967
Ferry, Alissa
3935
3935
Festman, Yariv
59
59
Fiedler, Klaus101
Fific, Mario ..... 2273
Fine, Alex ..... 2279
Fine, Justin ..... 2285
Finlayson, Mark ..... 81
Finley, Sara ..... 442, 2291
Fire, Amy ..... 2297
Firouzi, Hadi ..... 2344
Fischer, Andreas ..... 3936
Fischer, Helen ..... 2303
Fischer, Martin ..... 3935
Fisher, Anna ..... 507, 2422, 2428, 3947
Fisseni, Bernhard ..... 81
Fitch, W. Tecumseh ..... 1199
Fizke, Ella ..... 3937
Flecken, Monique ..... 2309
Flumini, Andrea ..... 2315
Foldes, Andrei ..... 3865
Forbus, Kenneth ..... 930
Forrest, Charlotte L. D. ..... 1020
Foster, James ..... 448
Foster, Meadhbh ..... 2321
Foster, Seth ..... 4122
Fox, Neil ..... 1133, 1450
Francois, Clement ..... 103
Frank, Michael ..... 299, 454, 621, 876, 882, 1641, 3032, 3169
Franke, Michael ..... 376
Frankenstein, Julia ..... 3938
Frassinelli, Diego ..... 460
Fratini, Viviana ..... 2746
Freeman, Jonathan ..... 97
French, Bob ..... 2398
Freudenthal, Daniel ..... 2327
Friederich, Michael ..... 2017
Frings, Christian ..... 4062
Frintrop, Simone ..... 2984
Frith, Chris ..... 3876
Frye, Conor ..... 2333
Fujinami, Tsutomu ..... 2548
Funke, Joachim ..... 3936
Furber, Steve ..... 310, 3898

\section*{Author Index}
Furlong, Ellen ..... 2338
Fusaroli, Riccardo 79, 3876, 3939, 3940
Futrell, Richard ..... 3245
Gabora, Liane 2344, 2524, 3251, 3905
Gafos, Adamantios ..... 1241
Gafter, Roey ..... 3486
Gagnier, Kristin ..... 2350
Galantucci, Bruno ..... 3299
Galati, Alexia ..... 466
Gale, Maggie ..... 1396
Galesic, Mirta ..... 472, 1821, 1923
Galligan, Megan ..... 3941
Galmar, Bruno ..... 2356
Gambi, Chiara ..... 2362
Ganguli, Surya ..... 1271
Gann, Timothy ..... 477
Gao, Jie ..... 3593
Garcia Penton, Lorna ..... 3942
García-Albea, José ..... 4029
Garcia-Madruga, Juan Antonio ..... 4140
Garcia-Retamero, Rocio ..... 3795
Gärdenfors, Peter ..... 1803
Garrard, Peter ..... 109
Garrod, Simon ..... 215
Gartenberg, Daniel ..... 2692
Gaskell, Gareth ..... 4051
Gatt, Albert ..... 28, 483
Gattis, Merideth ..... 3384, 3968
Gauer, Gustavo ..... 4056
Gaviria, Christian ..... 2656
Gaylord, Nicholas ..... 2368
Gazzo Castaneda, Lupita Estefania ..... 2374
Geffen, Susan ..... 3943
Geller, Emma ..... 3924
Gentaz, Edouard ..... 3110
Gentner, Dedre 81, 651, 1318, 3934, 3949,3976, 4079Georg, Anne1917
Georgescu, Alexandra ..... 3944, 4145
Georgieva, Iskra ..... 2578
Gerjets, Peter ..... 716, 2608, 4092
Gerrig, Richard ..... 81
Gershman, Samuel ..... 83, 489
Gerson, Sarah ..... 2380
Gerstenberg, Tobias ..... 2386
Gerwien, Johannes ..... 2309
Giardini, Francesca ..... 1756
Giardino, Valeria ..... 645
Giese, Martin ..... 870, 3915
Giezen, Marcel ..... 4147
Gigerenzer, Gerd ..... 59
Gijssels, Tom ..... 495
Gill, Devin ..... 3945
Gilman, Anne ..... 2392
Gingras, Bruno ..... 1199
Giorgolo, Gianluca ..... 2064
Glady, Yannick ..... 2398
Glage, Philipp ..... 4034
Glass, Brian ..... 501
Glebkin, Vladimir ..... 2404
Gleitman, Lila ..... 805
Gliozzi, Valentina ..... 2410
Glogger, Inga ..... 2416
Gluck, Kevin ..... 43, 3946
Gobet, Fernand ..... 2327
Godwin, Karrie ..... 507, 2422, 2428, 3947
Goedert, Kelly ..... 3948
Goel, Ashok ..... 2796, 3651
Goh, Winston ..... 4054
Goksun, Tilbe ..... 1786
Goldenberg, Olga ..... 3765
Goldin-Meadow, Susan ..... 57
Goldstone, Robert 63, 91, 1928, 1982, 3747, 3964, 4096
Goldwater, Micah ..... 3949, 4079
Gonnerman, Laura ..... 2766
Gonzalez-Marquez, Monica ..... 3950
Gonzalez, Cleotilde ..... 175, 3139
Gonzalez, Cleotile ..... 2303
Goodman, Noah ..... 107, 728, 900
Goolkasian, Paula ..... 3951
Gopal, Anvita ..... 4107

\section*{Author Index}
Gopnik, Alison ..... \(69,1558,1564\)
Habel, Christopher ..... 1714
Gorawara-Bhat, Rita ..... 3966
Gotzner, Nicole ..... 2434
Goudbeek, Martijn ..... 2440, 2772
Grabner, Roland ..... 3175
Gralla, Linn ..... 513
Gratch, Jonathan ..... 41, 770
Graver, Cleve ..... 245, 2213
Habib, Marianne ..... 2464
Hagmayer, York ..... 531
Hagoort, Peter ..... 1127, 2560
Hahn, Sowon ..... 1964
Hahn, Ulrike ..... 555, 3384
Haladjian, Harry ..... 2470Gray, Wayne609, 912
Greenwood, Michelle ..... 3952
Grela, Bernard ..... 3910
Grice, Martine ..... 3360
Griffin, Thomas ..... 3953
Griffiths, Thomas ..... \(1593,1686,2584,3959\)
Grigull, Robert ..... 3759
Grinberg, Maurice ..... 2578, 3954
Gripshover, Sarah ..... 2446
Grootswagers, Tijl ..... 4161
Gu, Hairong ..... 2452
Guath, Mona ..... 519
Guay, Brian ..... 3955
Gubbins, Eoin ..... 3956
Gudwin, Ricardo ..... 2932
Guerra, Ernesto ..... 1851, 2458
Guillon Dowens, Margaret ..... 4075
Guinet, Éric ..... 3110
Gulan, Tanja ..... 3875
Guney, Sule ..... 525
Gunzelmann, Glenn ..... 30
Guo, Mengqi ..... 3948
Guo, Xiuyan ..... 103
Gupta, Swati ..... 32, 3926
Gureckis, Todd ..... 93, 3909, 4038
Gurman Bard, Ellen ..... 28
Güss, C. Dominik ..... 3957, 4150
Guthrie, Lisa ..... 1504
Guzdial, Mark ..... 978
Gweon, Hyowon ..... 3958
Haagmans, Inge ..... 2440
Haartsen, Rianne 4046
Halima Smith, Asha ..... 3245
Hall, Lars ..... 1115
Hamburger, Kai ..... 4139
Hamer, Aaron ..... 2476
Hampton, James ..... 2482
Hamrick, Jessica ..... 1686, 3959
Han, Shihui ..... 95
Hancock, Roeland ..... 543
Handke, Lisa ..... 4158
Handley, Simon ..... 3974
Hanel, Ricardo ..... 4150
Hansen, Niels Chr. ..... 3960, 3960, 3960
Hanulikova, Adriana ..... 3961
Harari, Daniel ..... 394
Harbison, J. Isaiah ..... 549
Hardey, Ilonca ..... 105
Hardt, Oliver ..... 2512
Harmon-Vukic, Mary ..... 3962
Harrell, Maralee ..... 2488
Harrington, Kelsey ..... 105
Harris, Adam ..... 555
Harris, Jack ..... 3946
Harris, Paul ..... 591
Hartholt, Arno ..... 41
Hartman, Ashley ..... 4050
Hartshorne, Joshua ..... 561
Haryu, Etsuko ..... 675
Hashimoto, Takashi . 4005, 4023, 4061, ..... 4133
Hassebrock, Justin ..... 2076
Hatano, Aya ..... 2494
Haun, Daniel ..... 67
Hausmann, Robert ..... 3963
Häußinger, Florian ..... 2608

\section*{Author Index}
Hawkiins, Guy ..... 567
Hobaiter, Cat ..... 57
Hawkins, Robert ..... 3964
Hayashi, Hajimu ..... 3965
Hayashi, Yugo ..... 2500
Hayes, Bradley ..... 1708
Hoffmann, Janina ..... 585
Hohenstein, Sven ..... 2554
Hohlfeld, Annette ..... 4034Hayes, Brett567, 2890
Hayhoe, Mary ..... 686
Healey, Patrick ..... 79, 2231, 2506
Heathcote, Andrew ..... 73, 3581
Hegarty, Mary ..... 157, 1958, 4017
Heit, EvanHelfer, Peter2512
Hellmannn, Katharina ..... 2518
Hemeren, Paul ..... 1528
Hemmer, Pernille ..... 3873
Hemmerich, Josh ..... 3966
Henderson, Annette ..... 85
Henderson, Madeleine ..... 2524
Hendrickson, Andrew ..... 3699
Hendrix, Peter ..... 1193, 3245
Herppich, Stephanie ..... 2530
Hertwig, Ralph ..... 1408, 3817
Herzog, Stefan ..... 2536
Hespos, Susan ..... 3934, 3967
Heussen, Daniel ..... 2482
Heyman, Tom ..... 2542
Hiatt, Laura ..... 573
Hidaka, Shohei ..... 2548, 3080
Higgins, Emily ..... 197
Hilbrink, Elma ..... 3384, 3968
Hildenbrand, Inesa ..... 4041
Hill, Felix ..... 579
Hill, Katherine ..... 2766
Hills, Thomas ..... 61, 3104, 3817
Hilse, Julia ..... 3936
Hiolle, Antoine ..... 1845
Hirata, Yukari ..... 3998
Hirst, William ..... 77, 4168
Ho, Mark ..... 3969
Ho, Seng-Beng ..... 3970
Hoicka, Elena ..... 591, 3882
Holcomb, Phillip ..... 3337
Holden, John ..... 966, 3894
Holland, Rob ..... 734
Hollenbeck, Emily ..... 3934, 3967
Holler, Judith ..... 1127, 2560
Holmes, Kevin ..... 597
Holsanova, Jana ..... 3991
Hölscher, Christoph ..... 3938, 4012, 4105
Holt, Daniel ..... 3936
Holubar, Taylor ..... 603
Holyoak, Keith ..... 2028
Holzäpfel, Lars ..... 2416
Hong, Uphong ..... 4171
Hope, Ryan ..... 609
Horn, Sebastian ..... 3971
Horne, Zach ..... 3972
Horne, Zachary ..... 615, 1151
Horowitz, Alexandra ..... 621
Horst, Jessica ..... 1792, 2566
Hotaling, Jared ..... 2572
Hou, Shumeng ..... 3973
Houdé, Olivier ..... 2464
Howarth, Stephanie ..... 3974
Howell, Steven ..... 3975
Howes, Andrew ..... 1049
Howes, Christine ..... 2506
Hoyos, Christian ..... 3976
Hristova, Evgenia ..... 2578, 3954
Hristova, Penka ..... 627
Hsiao, Janet ..... 328
Hsiao, Janet Hui-wen ..... 1492, 2022, 2356
Hsu, Dong-Bo ..... 3977
Hu, Jane ..... 2584
Hu, Jon-Fan 3906, 3907, 3978
Huang, Chiu-Hua 3906, 3907

\section*{Author Index}
Huang, Yanli ..... 3979
Huang, Yu-Ping ..... 4143
Hudson Kam, Carla ..... 1611
Huette, Stephanie ..... 3980, 4086
Huettig, Falk ..... 3420
Huiskes, Mike ..... 4117
Hummel, John ..... 698
Hunnius, Sabine ..... 1516, 2380, 4046
Hunsberger, Eric ..... 633
Huo, Zhiyu ..... 281
Hupbach, Almut ..... 1109, 3981
Hurzook, Aziz ..... 2590
Hussey, Erika ..... 549
Hutchins, Edwin ..... 2668
Hutchinson, Sterling ..... 2596, 3551, 3557
Hüther, Lisa ..... 2602
Huys, Quentin ..... 900
Hyatt, Sophia ..... 4162
Iani, Cristina ..... 2938
Ibarra, Alyssa ..... 3303
Iida, Ryu ..... 3563
Ikeda, Masami ..... 4010
Imai, Mutsumi ..... 1253, 3080, 4169
Imhof, Birgit ..... 2608, 4092
Ingram, Gordon ..... 3982
Ip, Mavis ..... 2000
Ireland, James ..... 1682
Irvine, Liz ..... 2614
Ishiguchi, Akira ..... 3869, 4010
Ishiguro, Chiaki ..... 3983
Ishii, Tatsunori ..... 3984
Ito, Akira ..... 1438
Ito, Jonathan ..... 2620
Ito, Yuichi ..... 2626
Itti, Laurent \(1085,3157,3163\)Iturria-Medina, Yasser ............ 3942, 4075
4004, 4125 Iwamoto, Ayae
Jaarsveld, Saskia ..... 2632
Jaeger, Florian ..... 3835
Jaeger, T. Florian 430, 858, 2279

\section*{Author Index}
Juslin, Peter ..... 519, 906, 2908
Kabashima, Kosuke ..... 3823, 4146
Kachergis, George . 63, 704, 710, 2100, 2716
Kacinik, Natalie ..... 3994
Kahl, Sebastian ..... 823
Kajimura, Shogo ..... 4125
Kaland, Constantijn ..... 1145
Kalkan, Sinan ..... 3604
Kamide, Yuki ..... 834, 3281
Kammer, Dietrich ..... 4059
Kämmer, Juliane Eva ..... 3995
Kammerer, Yvonne ..... 716
Kanayet, Frank ..... 722
Kanbi, Hasina ..... 2482
Kanda, Takayuki ..... 3847
Kantartzis, Katerina ..... 1253, 3080
Kanwal, Jasmeen ..... 972
Kanzaki, Nana ..... 3068
Kao, Justine ..... 728
Kao, Yvonne ..... 364
Kapatsinski, Vsevolod ..... 3454
Kapnoula, Efthymia ..... 1169
Kappich, Julian ..... 2416
Karnowski, Jeremy ..... 2668
Karwinkel, Verena ..... 4078
Kary, Arthur ..... 3139
Kataoka, Misao ..... 4067
Katidioti, Ioanna ..... 3996
Katsumata, Yoko ..... 3997
Kauffman, Linda ..... 2094
Kaup, Barbara ..... \(133,2225,4092\)
Kavanagh, Liam ..... 734
Kawaguchi, Jun ..... 2494, 2626
960, 2321, 4070 Keane, Mark
4135, 4167
4135, 4167
Keeble, Sarah
Keeble, Sarah
1330, 3205
1330, 3205
Kellen, David
Kellen, David
460, 2237
460, 2237
Keller, Frank
Keller, Frank ..... 740
Keller, Josh
3931
Kelliher, Katherine
3864
Kello, Chris

\section*{Author Index}
Kitayama, Shinobu ..... 95
Kortenkamp, Ulrich ..... 4001
Kitazume, Hideaki ..... 3997
Kitto, Kirsty ..... 2740
Kiyokawa, Sachiko ..... 4003
Kılıç, Özkan ..... 2722, 3999
Klauer, Karl Christoph ..... 59, 1330
Kliegl, Reinhold ..... 2554
Kline, Melissa ..... 794
Kljajevic, Vanja ..... 2746
Kloos, Heidi ..... 966, 1732, 2076, 2285, 3894
Kmiecik, Matthew ..... 799
Knauff, Markus .... 1067, 1946, 2112, 2374,3038
Kneer, Markus ..... 2752
Knoblich, Günther 55, 1522, 4085
Koslowski, Barbara ..... 424
Kostopoulou, Olga ..... 531
Kotov, Alexey ..... 4007
Kotov, Artemy ..... 4172
Kotova, Tatyana ..... 4007
Kowalski, John ..... 1653
Krahmer, Emiel .. 28, 483, 817, 2440, 2772, 3681
Krajewski, Grzegorz ..... 4135, 4167
Krämer, Katharina ..... 4008
Kranjec, Alexander ..... 829, 4009
Kranz, Julia ..... 4092
Kreile, Monika ..... 4114
Krems, Josef ..... 1917
Knoeferle, Pia 1851, 1976, 2458 Kristiansson, Mattias ..... 2778
Knott, Alistair ..... 1432, 2849
Kroll, Judith ..... 26
Kobayashi, Harumi ..... 3997
Krumnack, Antje ..... 3315
Kochrginsky, Masha ..... 3966
Koedinger, Ken ..... 113
Koedinger, Kenneth ..... 894, 1384, 1653
Ku, Hyun-Min ..... 3869, 4010
Kuchinke, Lars ..... 193, 1187
Kudelkina, Natalya ..... 4011
Koehne, Judith ..... 805
Kuhlen, Anna ..... 3086
Köhne, Judith ..... 2760
Kohno, Yu ..... 4132
Kojima, Kazuaki ..... 1444
Kojima, Takatsugu ..... 4004, 4125
Kok, Kasper ..... 3881
Kokinov, Boicho ..... 627
Kolbeck, Carter ..... 169
Kolne, Kendall ..... 2766
Komorowska, Krystyna ..... 3917
Koniak, Paweł ..... 948
Konno, Takeshi ..... 4005, 4061
Konopka, Agnieszka ..... 1265
Konstantinidis, Emmanouil ..... 811
Kontinen, Jarmo ..... 1187
Koolen, Ruud ..... 817, 2772
Kopec, Lukasz ..... 4006
Kopp, Stefan ..... 24, 823
Kuhlmann, Christina ..... 2784
Kukona, Anuenue ..... 834
Kuliga, Saskia ..... 4012
Kumar, Devpriya ..... 2790
Kumar, Kiran ..... 4013
Kunda, Maithilee ..... 2796
Kuorikoski, Jaakko ..... 840
Kurtz, Kenneth ..... 2088, 4014
Kurumada, Chigusa ..... 846, 852, 858, 3486
Kushnir, Tamar ..... 2040, 3705
Kusumi, Takashi ..... 4125
Kuzmanovic, Bojana ..... 3944, 4008, 4145
Kvetova, Eliska ..... 4015
Kwak, JaRang ..... 4019
Kwisthout, Johan ..... 53, 87, 4016
Kwong, Elena ..... 2000Korhonen, Anna579, 746
Kornkasem, Sorachai ..... 4031
Labate, Enia ..... 4017
Labrecque, Alexandre ..... 4138
Lachmann, Thomas ..... 2632

\section*{Author Index}
Lachner, Andreas ..... 2802
Leite, Jan ..... 4022
Lacroix, Guy ..... 3366
Laeng, Bruno ..... 3480
Lagerstedt, Erik ..... 1803
Lagnado, David ..... \(71,2386,3886\)
Lai, Meng-Lung ..... 4018
Laine, Tei ..... 32, 2808
Lakens, Daniël ..... 1259
Lam, Rachel ..... 2814
Lam, Zinnia ..... 3973
Lambon Ralph, Matthew ..... 310, 3898
Landau, Barbara ..... 3931, 3990
Landy, David ..... 3955
Lange, Nicholas ..... 864, 1964
Lanthier, Sophie ..... 2820
Lappi, Otto ..... 3331
Lappin, Shalom ..... 2064
Large, Edward ..... 99
Laskin, Lyuben ..... 1670
Laszlo, Sarah ..... 2826
Latham, Peter ..... 3876
Latif, Nida ..... 2832
Lau, Jennifer ..... 3876
Lauterman, Tirza ..... 2838
Lavelle, Mary ..... 2506
Layher, Georg ..... 870
Le Mens, Gael ..... 2177
Lea, Stephen E. G. ..... 1020
Lebiere, Christian ..... 175,3539
Lee-Hand, Jeremy ..... 2849
Lee, Chungmin ..... 4020
Lee, Daniel ..... 3151
Lee, Donghoon ..... 4019
Lee, Jessica ..... 2843, 2920
Lee, Michael ..... 3627
Lee, Seong-Deok ..... 2046
Lee, YoungIn ..... 4019
Legare, Cristine ..... \(67,85,1558,3437\)
Lehet, Matthew ..... 829
Lei, Miaomei 4021
Leminen, Miika ..... 4127
Leo, Joop ..... 2855
León-Villagrá, Pablo ..... 2860
Lepenioti, Despoina ..... 4157
Leung, Janny H.C. ..... 2866
Leung, Vina W.H. ..... 4165
Levine, Susan ..... 4079
Levinson, Stephen ..... 1265, 3968
Levy, Roger ..... 197, 728, 3032
Lewandowsky, Stephan ..... 89, 400, 918
Lewis, Molly ..... 876, 882
Lewis, Nicholas ..... 346
Lewis, Richard ..... 1049
Lewkowicz, Daniel ..... 2872
Li, Adam ..... 4160
Li, Feifei ..... 103
Li, Guanhong ..... 4023
Li, Huaye ..... 2878
Li, Kuncheng ..... 2884
\(\mathrm{Li}, \mathrm{Na}\) ..... 888
Li, Nan ..... 894
Li, Ping ..... 2961, 4089
Liang, Peipeng ..... 2884
Lichtenstein, Patricia ..... 4024
Lidz, Jeffrey ..... 3193
Liebal, Katja ..... 57
Lieder, Falk ..... 900
Lieven, Elena ..... 75
Liew, Janice ..... 2890
Liljeholm, Mimi ..... 2034
Lim, Ahnate ..... 2896, 2902
Lim, Joo-Hwee ..... 2638, 2644
Lim, Stephen Wee Hun ..... 3432, 3897, 3904, 4025
Lima, Danielly ..... 4022
Lin, Carmen ..... 3998
Lindeman, Marjaana ..... 4026, 4127
Lindemann, Oliver ..... 3935
Lindsey, Richland ..... 1863
Lindsey, Robert ..... 758

\section*{Author Index}
Lindskog, Marcus ..... 906, 2908
Lindstedt, John ..... 912
Link, Daniela ..... 4027
Lippmann, Marie ..... 2914
Lippold, Matthias ..... 3759
Liszkowski, Ulf ..... 3882
Little, Daniel ..... 918
Liu, Daniel Hsi-wen ..... 4028
Liu, Nian ..... 924
Liu, Yeou-Teh ..... 4160
Liu, Zhiya ..... 3979
Livesey, Evan ..... 1306, 2843, 2920
Livingston, Ken ..... 3916
Lloyd, Marianne ..... 4050
Lobina, David ..... 4029
Loewe, Benedikt ..... 81
Loewenstein, Jeffrey ..... 81, 740, 1038
Logan, John ..... 1283
Lohmann, Johannes ..... 2926
Lombrozo, Tania . 89, 406, 1558, 3777, 3879
Lomibao, Janice ..... 4136
Longmore, Chris ..... 3783
Lopes, Phil ..... 2195
Lotto, Beau ..... 1682
Loui, Psyche ..... 99
Loui, Sofia ..... 4030
Loui, Sonia ..... 2012
Loula, Angelo ..... 2932
Louwerse, Max ...... 2596, 2955, 3551, 3557
Love, Brad ..... 4006
Love, Bradley ..... 1682, 4071
Lovett, Andrew ..... 63, 930
Lu, Carol ..... 4031
Lu, Hongjing ..... 1157, 2028
Lu, Laura ..... 4031
Lu, Zhong-Lin ..... 2452
Lubashevsky, Ihor ..... 3853
Lucas, Christopher ..... 69
Lugli, Luisa ..... 2938, 2944
Luhmann, Christian 936, 2949
Luno, Jeremy ..... 2955, 4032
Lupyan, Gary ... 91, 1678, 1887, 2243, 4076
Luzzi, Carol ..... 1732
MacDonald, Maryellen ..... 4055
Macdonald, Ross ..... 942
Mackiewicz, Robert ..... 948
MacLaren, Ben ..... 2094
Macpherson, Fiona ..... 91
Maehigashi, Akihiro ..... 1444
Maes, Alfons ..... 3681
Magnan, Annie ..... 4035
Magnani, Lorenzo ..... 954
Magnotta, Vincent ..... 4115
Magnuson, James ..... 3910, 3992
Maguire, Phil ..... 960
Maguire, Rebecca ..... 960
Mahmoodi, Ali ..... 4033
Maier, Martin ..... 4034
Maïonchi-Pino, Norbert ..... 4035
Majid, Asifa ..... 67, 764
Maldonado, Samuel ..... 3777
Malejka, Simone ..... 4036
Malla, Felicity ..... 591
Malone, MaryLauren ..... 966, 1510
Maloney, Laurence ..... 3378
Malony, John ..... 3008
Malt, Barbara ..... 2961
Manabu, Kikuchi ..... 3901
Manalo, Emmanuel ..... 2967
Mancini, Francesco ..... 2973
Manduca, Cathryn ..... 1786
Mani, Inderjeet ..... 81
Marchenko, Olga ..... 4037
Marcus, Nadine ..... 3287
Marewski, Julian ..... 2201, 4027
Marghetis, Tyler ..... 972, 2458
Margni, Adrián ..... 1480
Margulieux, Lauren ..... 978
Marian, Viorica ..... 4111
Marino, Barbara F.M. ..... 2058

\section*{Author Index}
Markant, Doug ..... 22, 93, 4038
Markman, Arthur ..... 1360
Markman, Ellen ..... 263, 603, 2446
Marques, Nuno ..... 2978
Marsella, Stacy ..... 41, 1175, 2620
Marsh, Jessecae ..... 984
Martin Garcia, German ..... 2984
Martin, Jay ..... 22
Martinez, Tony ..... 3098
Mascarenhas, Samuel ..... 2978
Masnick, Amy ..... 352
Mast, Vivien ..... 990
Masterson, Jackie ..... 3928
Mastin, J. Douglas ..... 3687, 4039
Mata, Rui ..... 3971
Matellán, Vicente ..... 1199
Matlen, Bryan \(507,2422,3947,4079\)
Matlock, Teenie 3789, 3952, 4024, 4040,4153
Matsumuro, Miki ..... 2990
Matthew, Gingerich ..... 412
Matthews, Justin ..... 3952, 4040
Matusevych, Yevgen ..... 2996
Mayer, Richard ..... 1702
Mayor, Eric ..... 3002
Mayor, Julien ..... 2410
Mayrhofer, Ralf 996, 1014, 4041, 4083, 4091
Mc Carthy, Sean ..... 3008
McAlister, Shealan ..... 1732
McAndrew, A. ..... 3885
McAndrew, Amy ..... 3014, 3020
McCabe, Rosemaire ..... 2506
McCauley, Stewart ..... 692
McClelland, James ..... 1271, 4043, 4048, 4176
McCormick, Kelly ..... 4042
McDonnell, John ..... 22, 1002
McGraw, John ..... 79
McKenzie, Cameron Ross Lloyd ..... 4043
McKeown, Rebecca ..... 3193
McLaren, Ian P. L. . 1008, 1020, 3014, 3885,4044McLaren, Ian P.L. .................. 3020, 3829
McLaren, Rossy ..... 1008, 3885, 4044
McMullon, Mhairi ..... 4045
McQuire, Marguerite ..... 4101
McRae, Ken ..... 4084
McShane, Marjorie ..... 34
Meder, Björn ..... 1014, 2650
Meder, Bjorn ..... 93
Medimorec, Srdan ..... 3921
Medin, Douglas ..... 67
Mehler, Alexander ..... 37
Meier, Christina ..... 1020
Meijering, Ben ..... 1426, 3026
Meister, Jan Christoph ..... 81
Melcher, David ..... 4107
Melo, Francisco ..... 2978
Melwani, Shimul ..... 1038
Merrick, Alice ..... 1157
Meyer, Marlene ..... 4046
Meylan, Stephan ..... 3032
Michael, John ..... 4047
Michael, Lev ..... 764, 3893
Michaelian, Kourken ..... 77
Michimata, Chikashi ..... 4064
Mickey, Kevin ..... 4048
Midgley, Katherine ..... 3337
Mielke, Thomas ..... 3946
Migdał, Piotr ..... 2172
Mikheeva, Maria ..... 3038
Milanese, Nadia ..... 2938
Milazzo, Jennifer ..... 4060
Miles, Gareth ..... 3044, 4049
Milin, Petar ..... 3929
Miłkowski, Marcin ..... 3050
Miller, Jared ..... 281
Miller, Jeremy ..... 4050
Millroth, Philip ..... 519
Milton, Fraser ..... 3783
Minervino, Ricardo ..... 1480
Mintz, Toben ..... 3943, 4159
Mirkovic, Jelena ..... 4051

\section*{Author Index}
Miser, Tracey ..... 3056
Misyak, Jennifer ..... 103
Mitchell, Aaron ..... 2094
Mitchell, Margaret ..... 3062
Mitkidis, Panagiotis ..... 79
Miwa, Kazuhisa 1444, 2990, 3068, 3516
Miwa, Satoko ..... 4052
Miyamoto, Edson ..... 3074
Miyashiro, Kozue ..... 4053
Miyazaki, Michiko ..... 3080
Mizraji, Eduardo ..... 109 ..... 9
Mohamed Salleh, Mohamed Shan-Rievan
4054
Moirou, Despoina ..... 1659
Mok, Peggy ..... 4165
Mol, Lisette ..... 3086
Molchanova, Natalia ..... 2704
Molinaro, Nicola ..... 4075
Molitor, Stephen ..... 4094
Moll, Ricarda ..... 3092
Monaghan, Padraic ..... 3420
Mønster, Dan ..... 79
Montag, Jessica ..... 4055
Monteith, Kristine ..... 3098
Montemayor, Carlos ..... 2470
Montez, Priscilla ..... 3533
Moore, Chris ..... 3139, 3627
Moraes, Berla ..... 4022
Morais, Ana Sofia ..... 3104
Moreno-Rios, Sergio ..... 3327, 4140
Moreno, Andre Luiz ..... 4056
Morese, Rosalba ..... 4057
Morgado, Nicolas ..... 3110
Morganti, Francesca ..... 3116
Mori, Kazuo ..... 4144
Morita, Junya ..... 1026, 1444, 4005
Moritsu, Tomoyuki ..... 1026
Morris, Bradley ..... 352
Morrison, Robert ..... 799, 3008
Morse, Anthony ..... 61
Moscoso del Prado, Fermin ..... 1032

\section*{Author Index}
Neumann, Heiko ..... 870
Neveu, Grace ..... 764, 3893
Newell, Ben \(89,525,567,1994,3139,3372\),3627
Newen, Albert ..... 95, 1079
Newport, Elissa ..... 430
Ng, Alvin C. W. ..... 328
Nickerson, Jeffrey ..... 3593
Nicoletti, Roberto ..... 2944, 3522
Nijboer, Menno ..... 4063
Nili ahmadabadi, Majid ..... 4033
Nilsson, Håkan ..... 3988
Nishida, Masafumi ..... 3823, 4146
Nishida, Toyoaki ..... 4067
Nishimura, Akio ..... 4064
Noelle, David ..... 1372, 4087
Noguchi, Takao ..... 3145
Noh, Junhyug ..... 2046
Noh, Yung-Kyun ..... 3151
Nokes-Malach, Timothy ..... 1881
Nomura, Ryota ..... 4065
Noordewier, Marret ..... 2772
Noori, Nader ..... 1085, 3157, 3163
Norcliffe, Elisabeth ..... 1265
Nordmeyer, Ann ..... 3169
Nour, Bakr ..... 3888
Nückles, Matthias .. 2416, 2518, 2530, 2802
Nunez, Rafael ..... 3729
Nussbaumer, Daniela ..... 3175, 4066
Nuthmann, Antje ..... 1569, 4100
Nyamsuren, Enkhbold ..... 3181
Nygaard, Lynne ..... 4042
O'Connor, Brian ..... 3251
O'Connor, Shannon ..... 1732
O'Hora, Denis ..... 97
O'Leary, Allison ..... 1091
O'Rourke, Polly ..... 1097
O'Seaghdha, Padraig ..... 1109
O'Toole, Stephanie ..... 4070
Oaksford, Mike ..... 59
Obbens, Mike ..... 3086
Oberauer, Klaus ..... 400
Ochs, Magalie ..... 1163
Ognyanova, Pavleta ..... 3872
Ohira, Hideki ..... 4061
Ohlsson, Stellan ..... 105, 3239
Ohmoto, Yoshimasa ..... 4003, 4067
Ohtsu, Kayoko ..... 4068
Ojalehto, Bethany ..... 67
Okada, Brooke ..... 3414
Okada, Hiroyuki ..... 3080
Okada, Takeshi ..... 3983, 4065, 4108, 4131
Okuda, Hiroyuki ..... 1444
Okuda, Jiro ..... 4023
Olguín, Valeria ..... 1480
Öllinger, Michael ..... 3912
Olsen, Karsten ..... 3876
Olsson, Henrik ..... 472, 4120
Oltramari, Alessandro ..... 175
Omigbodun, Akinyinka ..... 3187
Ong, Sim-Heng ..... 2638, 2644
Onoda, Ryosuke ..... 4069
Opfer, John ..... 722, 2338
Orfanidou, Eleni ..... 4030
Orita, Naho ..... 3193
Ormand, Carol ..... 1786
Osman, Magda ..... 501
Osth, Adam ..... 3199
Otieno, Christine ..... 1103
Ott, Laurent ..... 2207
Ottoboni, Giovanni ..... 3522
Over, David ..... 101
Oversteegen, Leonoor ..... 1145
Oyo, Kuratomo ..... 4132
Ozyurek, Asli ..... 2560
Özyürek, Aslı ..... 1127
Paas, Fred ..... 4081
Pachur, Thorsten ... 1277, 1408, 3205, ..... 3971
Paciorek, Albertyna ..... 3211
Packer, Dominic ..... 1109
Paiva, Ana ..... 2189, 2978

\section*{Author Index}
Pak, Sarah ..... 2949
Palluel-Germain, Richard ..... 3110
Papavasiliou, Petros ..... 4157
Park, Frank Chongwoo ..... 3151
Park, Kun Hee ..... 2692
Park, Rae-yeop ..... 3919
Parnamets, Philip ..... 1115
Parpart, Paula ..... 4071
Paus, Elisabeth ..... 4072
Pautler, David ..... 4073
Pavlenko, Aneta ..... 2961
Paxton, Alexandra ..... 1121, 3303, 3864
Payne, Stephen ..... 3044
Pazzaglia, Francesca ..... 4017
Pearce, Marcus ..... 99, 127, 3960
Pease, Alison ..... 1893
Pecher, Diane ..... 3639
Peebles, David ..... 1720
Peeters, David ..... 1127
Pelachaud, Catherine ..... 1163
Pelton, Hannah ..... 3958
Pennisi, Antonino ..... 275
Pereira, Alredo ..... 4074
Perez Fernandez, Alejandro ..... 3942
Pérez, Alejandro ..... 4075
Perfors, Amy ..... 3699
Perlman, Marcus ..... 4024, 4076
Perner, Josef ..... 69
Perniss, Pamela ..... 1133
Perone, Sammy ..... 4094
Perry, Lynn ..... 3217
Pesquita, Ana ..... 4077
Petkova, Yoana ..... 627
Petroccia, Megan ..... 2428
Pfeiffer, Ulrich ..... 35, 4145
Phillips, Steven ..... 3223
Phillips, William ..... 87
Pickering, Martin ..... 2362
Pickett, Marc ..... 3229
Pieschl, Stephanie ..... 2784, 3092, 4078
Pine, Julian ..... 2327
Pinelli, Mathieu ..... 3110
Ping, Raedy ..... 4079
Pinillos, Angel ..... 1151
Pipergias Analytis, Pantelis ..... 3995
Pirolli, Peter ..... 93
Pitt, Mark ..... 2452
Plaut, David ..... 139
Plaza, Bérengère ..... 4116
Plewczynski, Dariusz ..... 2172
Plunkett, Kim ..... 1726, 2410
Pomplun, Marc ..... 1617
Ponari, Marta ..... 1540
Pontillo, Daniel ..... 846
Popov, Vencislav ..... 4080
Porsch, Torsten ..... 2784
Porter, Anne Marie ..... 3951
Post, Lysanne ..... 3621, 4081
Postma-Nilsenová, Marie . 1139, 1145, 3587, 4082
Postma, Eric ..... 1139
Pothos, Emmanuel ..... \(45,101,1211,1599\)
Pouget, Alexandre ..... 83
Powell, Derek ..... \(615,1151,1157,3972\)
Powell, Lindsey ..... 105
Powers, David ..... 1680, 1680
Pöyhönen, Samuli ..... 840
Pozdnyakova, Elena ..... 4172
Prada, Rui ..... 2195, 2978
Prasada, Sandeep ..... 3235, 3889
Prepin, Ken ..... 1163
Prochownik, Karolina ..... 3982
Prodi, Elena ..... 4057
Protopapas, Athanassios .. 1169, 1659, ..... 4030
Proust, Joëlle ..... 101
\(\mathrm{Pu}, \mathrm{He}\) ..... 3337
Pusiol, Guido ..... 454
Pyke, Aryn ..... 3891
Pynadath, David ..... 1175
Quack, Benjamin ..... 4083
Queiroz, João ..... 2932

\section*{Author Index}
Quinn, Connor ..... 1181
Rhodes, Theo ..... 2680, 4128
Quintelier, Katinka J.P. ..... 101
Rabellino, Daniela ..... 4057
Rabovsky, Milena ..... 4084
Rączaszek-Leonardi, Joanna ..... 2172
Rafetseder, Eva ..... 69
Rainford, Katherine ..... 4060
Rajaram, Suparna ..... 26, 936
Rak, Natalia ..... 1187
Rakoczy, Hannes ..... 3937
Ramenzoni, Veronica ..... 4085
Ramsburg, Jared ..... 3239
Ramscar, Michael ..... 1193, 3245
Rangel, Antonio ..... 73
Ranjan, Apara ..... 3251
Ranney, Michael ..... 2070
Ranson, Samantha ..... 2566
Raskin, Victor ..... 1390
Rasmussen, Daniel ..... 3257
Ravignani, Andrea ..... 1199
Rayner, Keith ..... 197
Read, Daniel ..... 1289
Rebuschat, Patrick ..... 103
Redcay, Elizabeth ..... 3263
Redfern, Alice ..... 3890
Rees, Geraint ..... 3876
Regier, Terry ..... 293, 764, 3893
Rehder, Bob ..... 1002, 3909
Reis, Henrique ..... 2189
Reiter, Ehud ..... 3062
Reitman, Jason ..... 3414
Reitter, David ..... 3269
Renkl, Alexander \(113,1103,1205,2416\),
2530
Reppa, Irene ..... 1211
Resnick, Ilyse ..... 1786, 3275
Reuter, Kevin ..... 1217
Revie, Gavin ..... 3281
Revithis, Spyridon ..... 3287
Revlin, Russell ..... 1414
Rezzonico, Giorgio ..... 3116
Ricciardelli, Paola ..... 2058
Richardson, Daniel ... 55, 1115, 1682, 4121,4175Richardson, Michael ... 55, 966, 1510, 2076,3309, 3741, 3894
Richie, Russell ..... 1223
Richland, Justin ..... 3293
Richland, Lindsey ..... 3293
Richmond, Christopher ..... 1396
Riefer, Peter ..... 1682
Rieskamp, Jörg ... 73, 472, 585, 1546, 1744,3922, 3988
Rigoli, Lillian ..... 4086
Rio, Kevin ..... 2710
Rips, Lance ..... 2662, 3967
Risko, Evan ..... 2820, 3921
Ritter, Helge ..... 3747
Rittle-Johnson, Bethany ..... 370
Roberts, Gareth ..... 3299
Roberts, Sean ..... 1229, 2614
Roche, Jennifer ..... 3303
Rodny, Jeffrey ..... 4087
Roepstorff, Andreas ..... 55, 3876
Roeser, Florian ..... 4139
Rogers, John ..... 1682
Rogoff, Barbara ..... 3914
Rohde, Marieke ..... 71
Rohr, Lana ..... 4088
Rohrmeier, Martin ..... 99, 103
Rolotti, Sebastian ..... 4089
Romance, Nancy ..... 3675
Romberg, Alexa ..... 1235
Romero, Veronica ..... 3309
Roon, Kevin ..... 1241
Rose, Sebastian ..... 4090
Röser, Florian ..... 537, 3315
Ross, Brian ..... 316
Rossetti, Angela ..... 2058
Rossi, Sandrine ..... 3321
Rothe, Anselm ..... 4091
Rothman, Naomi ..... 984
Röttger, Timo ..... 1247

\section*{Author Index}
Rousseau, Vincent ..... 4138
Rozga, Agata ..... 2796
Rubichi, Sandro ..... 2938
Rueckl, Jay ..... 3910
Rueschemeyer, Shirley-Ann ..... 495
Ruggeri, Azzurra ..... 1923
Ruiz Fernández, Susana ..... 4092
Ruiz-Ballesteros, José ..... 3327
Rummel, Nikol ..... 3765
Rusanen, Anna-Mari ..... 3331
Sadeghi, Sepideh ..... 3337
Sagi, Eyal ..... 3343
Sahakyan, Lili ..... 3981
Sahin, Erol ..... 3604
Saito, Motoyuki ..... 3348
Saji, Noburo ..... 1253
Sakamoto, Kayo ..... 32, 2808
Sakamoto, Yasuaki ..... 2878
Sakkalou, Elena ..... 3384, 3968
Salamon, Samuel ..... 3994
Salden, Ron ..... 1103
Salehi, Niloufar ..... 3354
Samsonovich, Alexei ..... 4093
Samuelson, Larissa ..... 61, 3217, 3941, 3987,
4094
Sanborn, Adam ..... 3145
Sandamirskaya, Yulia ..... 49, 4095
Sandhofer, Catherine ..... 2034
Sang, Ke ..... 4096
Sangati, Federico ..... 1696
Santiago, Julio ..... 1259, 2315
Santos, Natacha ..... 3944
Sanz, Ricardo ..... 4097
Sato, Yuki ..... 4125
Sauppe, Sebastian ..... 1265
Savino, Michelina ..... 3360
Sawa, Koji ..... 4098
Saxe, Andrew ..... 1271
Saxe, Rebecca ..... 3263, 3958
Sayeed, Asad ..... 2148, 2249
Scassellati, Brian ..... 1708
Scerri, Paul ..... 3269
Schaat, Samer ..... 4099
Schad, Daniel ..... 4100
Schaeken, Walter ..... 3986
Schalk, Lennart ..... 382
Scharrer, Lisa ..... 3448
Scheepers, Christoph ..... 460
Scheibehenne, Benjamin ..... 1277, 3922
Scheiter, Katharina ..... 4104
Schendan, Haline ..... 1684
Scheutz, Matthias ..... 3337, 3657
Schilbach, Leonhard ..... 4145
Schilling, Savannah ..... 3408
Schilperoord, Joost ..... 3639
Schlottmann, Anne ..... 1181
Schmidt-Snoek, Gwenda ..... 4101
Schmidt, Martin ..... 1893
Schmitz, Laura ..... 1522
Schneider, Michael ..... 105, 3175, 4066
Schneider, Steven ..... 364
Schnitzler, Verena ..... 4102
Schoen, Daniele ..... 103
Schoener, Gregor ..... 49
Schoenherr, Jordan Richard ..... 1283, 3366
Scholten, Marc ..... 1289
Schooler, Lael ..... 43, 2201
Schorch, Tanja ..... 4103
Schubotz, Louise ..... 2560
Schuetze, Manuela ..... 2560
Schuhmann, Katharina ..... 1934
Schüler, Anne ..... 4104
Schultheis, Holger ..... 1294
Schulz, Laura ... 663, 794, 3122, 3807, ..... 3958
Schulze, Christin ..... 3372
Schuur, Friederike ..... 3378
Schwan, Stephan ..... 4092
Schwartz, Neil ..... 2914
Schwarz, Baruch ..... 1780
Schwarzkopf, Sarah ..... 4105
Schwonke, Rolf ..... 1103, 2416

\section*{Author Index}
Scott-Brown, Kenneth ..... 782
Scott, Katherine ..... 3384
Sebanz, Natalie ..... 55, 1522, 4085
Seekings, Paul ..... 2638, 2644
Segedy, James ..... 209
Segers, Eliane ..... 4148
Seidenberg, Mark ..... 4164
Sekine, Kazuki ..... 4106
Selecka, Marta ..... 4114
Sengupta, Rakesh ..... 4107
Sgro, Ivana ..... 3116
Shah, Nila ..... 3957
Shah, Punit ..... 555
Shanks, David ..... 811
Shenoy, Pradeep ..... 1300
Shepard, Jason ..... 3390
Sheppard, Chris ..... 2967
Sheredos, Ben ..... 3396
Shibata, Masayoshi ..... 4061
Shiffrin, Richard \(346,704,2100,2572,4013\)
Shimazaki, Tsuneo ..... 3348
Shimizu, Daichi ..... 4108
Shin, Hocheol ..... 4109
Shipley, Thomas 1786, 2350, 3275
Shoda, Marie ..... 4110
Shone, Lauren ..... 1306, 2920
Shook, Anthony ..... 4111
Shtulman, Andrew ..... 105
Shu, Hua ..... 2017
Shultz, Thomas ..... 2512, 2686
Sichelschmidt, Lorenz ..... 1869
Siegel, Linda ..... 4118
Sifonis, Cynthia ..... 4112
Silander, Tomi ..... 2808
Silva, Malena ..... 3931
Silvey, Catriona ..... 1312
Sim, Zi ..... 3402
Sima, Jan Frederik ..... 4113
Simmons, Kaia ..... 454
Simms, Nina ..... 1318
Simonsen, Arndis ..... 3939
Sims, Clare ..... 1324, 3408
Singhaniya, Ankit ..... 1294
Singmann, Henrik ..... 1330
Sinnett, Scott ..... 2183, 2896, 2902
Sirota, Miroslav ..... 3615
Škilters, Jurǵis ..... 4114
Skolnick, Tamar ..... 4014
Skubic, Marjorie ..... 281
Slevc, L. Robert ..... 3414
Slezak, Peter ..... 1336
Sloutsky, Vladimir . 1091, 2130, 2166, 3056,3841
Smilek, Daniel ..... 2820
Smith, Alastair ..... 3420
Smith, Kenny ... 340, 436, 1312, 1348, 3442
Smith, Kevin ..... 1342, 3426
Smith, Linda 669, 1988, 3610, 4074, ..... 4122
Smith, Linsey ..... 4079
Smolek, Amy ..... 3454
Smolen, Tomasz ..... 4163
Snedeker, Jesse ..... 561
So, Wing Chee ..... 2000, 3973
Soh, Wei Jie ..... 3432
Solloway, Tyler ..... 1575
Solymosi, Judith ..... 1205
Son, Ji ..... 3527
Sonderegger, Morgan ..... 776
Sonnweber, Ruth ..... 1199
Sood, Mariam ..... 1354
Soulières, Isabelle ..... 2796
Souman, Jan ..... 2225
Souza, Andre ..... 1360, 3437
Spada, Hans ..... 2602
Spalek, Katharina ..... 251, 1366, 2434, 4090
Spaulding, Tammie ..... 3910
Speekenbrink, Maarten ..... 47, 3886
Spelke, Elizabeth ..... 107
Spencer, John ..... 4115
Spike, Matthew ..... 3442
Spino, Joseph ..... 3972

\section*{Author Index}
Spivey, Michael ..... 1115, 2261, 3903, 3952, Su, Yu-Han ..... 3900
3980, 4086
Spotorno, Sara ..... 4116
Srinivasan, Narayanan ..... 2790
St. Clair, William Benjamin ..... 1372
Stadler, Kevin ..... 3442
Stadtler, Marc ..... 3448
Stafford, Tom ..... 1378
Stamper, John ..... 113
Stampfer, Eliane ..... 894, 1384
Stannard, Jade ..... 4060
Stapel, Janny ..... 4046
Starr-Glass, Lolly ..... 3994
Starr, Rebecca ..... 3245
Staudte, Maria ..... 35
Stave, Matthew ..... 3454
Stec, Kashmiri ..... 4117
4118, 4143Stein, Nancy
Steinfeld, Judith ..... 3753
Stenning, Keith ..... 59
Stern, Elsbeth ..... 3175, 4066
Stevens, Corey ..... 2136
Stevenson, Suzanne ..... 1073, 1815, 1857
Stewart, Neil ..... 3145
Stewart, Terrence ..... 39, 3460
Steyvers, Mark ..... 3510
Stigler, James ..... 3924
Stocker, Kurt ..... 3468, 4119
Stojic, Hrvoje ..... 4120
Stokes, Dustin ..... 91
Stone, Charles ..... 77
Storms, Gert .. 1552, 2142, 2542, 3474, 4156Street, Chris N.H.4121
Streuber, Stephan ..... 3915
Strickrodt, Marianne ..... 537
Stuart, Lauren ..... 1390
Stuerzlinger, Wolfgang ..... 358
Stukken, Loes ..... 3474
Stupple, Edward ..... 1396
Sturm, Thomas ..... 1768
Su, Yanjie ..... 3811
4122
Suanda, Sumarga
4123
Suckow, Katja ..... 4123
Suegami, Takashi ..... 3480
Suess, Franziska ..... 4124
Sugimoto, Masashi 4004, ..... 4125
Suhler, Chris ..... 734
Sullivan, Brian ..... 686
Sullivan, Jess ..... 3717
Sulmasy, Daniel ..... 3966
Sulpizio, Simone ..... 1402
Summerfield, Christopher . ..... 3884, 3890, 4141
Sumner, Meghan ..... 3486
Sun, Yanlong ..... 4126
Suter, Renata ..... 1408
Sutton, John ..... 2082
Suzuki, Tatsuya ..... 1444
Svedholm, Annika ..... 4026, 4127
Swan, Alexander ..... 1414
Swerts, Marc ..... 817, 3639
Swingley, Daniel ..... 187
Szary, Janelle ..... 3492, 4128
Szollosi, Aba ..... 3865
Szymanik, Jakub ..... 1426
T. Harada, Etsuko ..... 4053
Taatgen, Niels 2884, 3026, 3181, 3891, 3996, 4063
Tabatabaeian, Maryam ..... 4129
Taber, Keith ..... 1181
Taguchi, Keiko ..... 3823
Taira, Tomohiro ..... 4130
Takac, Martin ..... 1432
Takagi, Kikuko ..... 4131
Takahashi, Tatsuji ..... 4098, 4132
Takashima, Atsuko ..... 4148
Takeda, Kazuya ..... 1444
Takezawa, Masanori ..... 3984
Tam, Brian ..... 3378
Tamariz, Monica ..... 1348
Tamborello, Frank ..... 3498
Tamura, Kaori ..... 4133

\section*{Author Index}
Tan, Cheston ..... 2638, 2644
Tan, Nicholas Jun Hao ..... 4025
Tanaka, Kanji ..... 3504
Tanaka, Yuko ..... 3593
Tanenhaus, Michael .. 846, 3303, 3835, 3927
Tatler, Benjamin ..... 942
Tauber, Sean ..... 3510
Taya, Fumihiko ..... 3926
Taylor, Elizabeth ..... 2638, 2644
Taylor, Holly ..... 4134
Taylor, Julia ..... 1390
Tazaki, Yuichi ..... 1444
ten Napel, Robyn ..... 3621
Tenbrink, Thora ..... 51, 513, 4134
Tenenbaum, Josh ..... 107
Tenenbaum, Joshua 83, 489, 663, 1342, 3959
Tepest, Ralf ..... 3944
Terada, Kazunori ..... 1438
Terai, Asuka ..... 3563
Terai, Hitoshi ..... 1444, 3068, 3516
Tessari, Alessia ..... 3522
Thai, Khanh-Phuong ..... 3527
Tham, Jo-Yew ..... 2638, 2644
Thanou, Emiliana ..... 4157
Theakston, Anna ..... 4135, 4167
Theodore, Rachel ..... 4136
Thibaut, Jean-Pierre ..... 2398
Thiede, Keith ..... 3953
Thill, Serge ..... 1528
Thomas, Angela ..... 3957
Thomas, Michael S. C. ..... 3928
Thomas, Rick ..... 864, 1964
Thomposn, Graham ..... 3533
Thompson, Bill ..... 436
Thompson, Robin ..... 1450
Thompson, Valerie ..... 101
Thompson, William ..... 2082
Thomson, Robert ..... 3539
Thornton, Chris ..... 1456, 3545
Tikoff, Basil ..... 1786

\section*{Author Index}
Tuason, Ma. Teresa ..... 3957
Tummeltshammer, Kristen Swan ..... 1420
Tupper, Paul ..... 3993
Turcotte, Isabelle ..... 4138
Tversky, Barbara ..... 645, 3593
Twomey, Katherine ..... 1498, 2566
Tylén, Kristian ..... 79, 3876, 3939, 3940
Tzeng, Yuhtsuen ..... 4118, 4143
Uchida, Akitoshi ..... 4144
Ueda, Kazuhiro ..... 4003
Ueno, Taiji ..... 2494, 2626
Ullman, Daniel ..... 1708
Ullman, Shimon ..... 107, 394
Ullman, Tomer ..... 107
Umata, Ichiro ..... 3823, 4146
Upal, M. Afzal ..... 3962
Urdaneta, Elena ..... 2746
Usher, Marius ..... 73
Utsumi, Akira ..... 3598
Uyanik, Kadir ..... 3604
Uzuner, Sebnem ..... 4147
Valentini, Consuelo ..... 4057
Valerjev, Pavle ..... 3875
Vales, Catarina ..... 1988, 3610
Valle-Lisboa, Juan ..... 109, 3892, 4177
Vallée-Tourangeau, Frédéric ..... 1504, 1534,3615
van Baaren, Rick ..... 1516
Van de Cavey, Joris ..... 2362
van Deemter, Kees ..... 28, 483, 3062
van den Broek, Gesa S. E. ..... 4148
van den Broek, Paul ..... 4170
van der Lely, Heather K. J. ..... 3992
van der Maas, Han L. J. ..... 3922
van der Steen, Ruben ..... 3086
van Gog, Tamara ..... 3621, 4081
van Gompel, Roger ..... 28, 483, 4123
van Hell, Janet ..... 26
van Kastel, Mariska ..... 4082
van Leeuwen, Cees ..... 2632
Van Meel, Chayenne ..... 1552
van Ravenzwaaij, Don ..... 3372, 3627
Van Rijn, Hedderik ..... 3026, 4063
van Rooij, Iris ..... 53, 87, 3883, 4016, 4161
van Rooij, Marieke ..... 1510
van Schaik, Johanna E. ..... 1516
Van Stockum, Charles ..... 3633
Van Valin, Robert D. ..... 1265
Van Voorhis, Kellie ..... 3966
van Weelden, Lisanne ..... 3639
Vankov, Ivan ..... 4149
Varadinov, Meryl ..... 1670
Varley, Rosemary ..... 65
Varzoni, Thais ..... 4150
Vasanov, Alexey ..... 4037
Vasilyeva, Nadya ..... 3645
Vatikiotis-Batesom, Eric ..... 2832
Vattam, Swaroop ..... 3651
Veale, Richard ..... 3657
Veksler, Vladislav ..... 3946
Veloz, Tomas ..... 4151
Vendrame, Mara ..... 3663
Ventura, Dan ..... 3098
Verbrugge, Rineke ..... 1426, 3026
Verbrugge, Sara ..... 3986
Verbruggen, Frederick ..... 3014
Verges, Michelle ..... 3925
Verghese, Preeti ..... 3985
Verheyen, Steven ..... 3474
Verhoef, Tessa ..... 3669
Verhoeven, Ludo ..... 4148
Vesper, Cordula ..... 1522
Veto, Peter ..... 1528
Vigliocco, Gabriella ..... 1133, 1450, 1540
Villejoubert, Gaëlle ..... 1504, 1534, 3615
Vink, Roy ..... 4152
Vinson, David \(1133,1450,1540,4153,4154\)
Visser, Ingmar ..... 47
Vitale, Michael ..... 3675
Vitevitch, Michael ..... 37
Vlach, Haley ..... 418, 4155

\section*{Author Index}
Vlahou, Eleni ..... 1659
Vlasova, Elizaveta ..... 4007
Vogeley, Kai \(95,3944,4008,4145\)
Vogels, Jorrig ..... 3681
Vogt, Paul 2996, 3687, 3874, 4039
Volkman, Brooks ..... 770
Vollmeyer, Regina ..... 4001
von Helversen, Bettina ..... 585, 1546, 2536
von Stülpnagel, Rul ..... 4012, 4105
von Sydow, Momme ..... 3693
Vong, Wai Keen ..... 3699
Voorspoels, Wouter ..... 1552, 4156
Vosniadou, Stella ..... 105,4157
Vredenburgh, Christopher ..... 3705
Vukovic, Nikola ..... 3711
Vul, Edward ..... 1342, 3426, 3717
Vuong, Annalies ..... 3963
Vuust, Peter ..... 3960, 3960
Wade, Anna ..... 1641
Wagner, Angela ..... 2094
Wagner, Katie ..... 2734
Wagner, Ullrich ..... 4158
Waheed, Anam ..... 3888
Wakayama, Yusuke ..... 3598
Wakeford, Laura ..... 3723
Waldmann, Michael .. 996, 3127, 4041, 4083
Walker, Caren .......... 69, 1558, 1564, 3777
Walker, Esther ..... 3729
Walsh, Clare ..... 3974
Walshe, R. Calen ..... 1569, 3993
Walter, Henrik ..... 3908, 4158
Wang, Felix ..... 4159
Wang, Hongbin ..... 4126
Wang, Hsueh-Cheng ..... 1617
Wang, Ning ..... 1175
Wang, Panqu ..... 3735
Wang, Ting-Hui ..... 3900
Wang, Tsanyu ..... 4160
Wang, Zheng ..... 45, 1575
Wareham, Todd \(53,3883,4161\)

\section*{Author Index}
Wills, Andy ..... 3783, 3829
Wilson, Colin 788, 2118, 3931, 3990
Wilson, Sarah ..... 1611
Wilson, William ..... 3287
Winkielman, Piotr ..... 734
Winman, Anders 906, 2908
Winter, Bodo ..... 2261, 3789
Wittwer, Jörg ..... 2530
Woike, Jan Kristian ..... 3795
Wolff, Phillip ..... 597, 3390
Wong, Orieta H. Y. ..... 4166
Wong, Sarah Shi Hui ..... 4025
Wong, Simpson ..... 4165
Wood, Katie ..... 1008
Woollams, Anna ..... 4135, 4167
Wray, Alison ..... 75
Wu, Chia-Chien ..... 1617
Wu, David W.-L. ..... 3801
Wu, Hsin-Hsuan ..... 3906, 3907, 3978
Wu, Rachel ..... 1420
Wu, Yang ..... 3807
Wu, Zhen ..... 3811
Wulff, DIrk ..... 3817
Xu, Fei ..... 2584, 3402
Yamada, Seiji ..... 1438
Yamamoto, Seiichi ..... 3823, 4146
Yamashiro, Jeremy ..... 4168
Yamazaki, Tomohito ..... 4169
Yan, Jin ..... 740
Yang, Charles ..... 1223
Yanguas, Javier. ..... 2746
Yasuda, Tetsuya ..... 3997
Yeari, Meni ..... 4170
Yeates, Fayme ..... 3014, 3829, 4044
Yeung, Christopher ..... 2000
Yeung, H. Henny ..... 3080
Yeung, Nick ..... 3884
Yeung, Saiwing ..... 1623
Yildirim, Ilker ..... 3835```


[^0]:    ${ }^{1}$ See http://pre2009.uvt.nl/ for the website and proceedings of the first workshop.
    ${ }^{2}$ See http://pre2011.uvt.nl/ for the website and proceedings of the second workshop.

[^1]:    ${ }^{3}$ http://homepages.abdn.ac.uk/k.vdeemter/pages/
    ${ }^{4}$ http://staff.um.edu.mt/albert.gatt/
    5 http://www.lel.ed.ac.uk/~ellen/
    ${ }^{6}$ http://www.dundee.ac.uk/psychology/rpgvangompel/
    ${ }^{7}$ http://www.emielkrahmer.nl/

[^2]:    Audience
    The target audience of this tutorial consists of post-graduate students and researchers in any subfield of cognitive science who wish to: (a) achieve an introductory level understanding of the basic concepts underlying computational complexity analysis, (b) gain hands-on experience with some of the basic proof techniques in

[^3]:    ${ }^{1}$ This work has greatly benefited from discussions with Katie Slocombe, University of York, and Bridget Waller, University of Portsmouth. See also Slocombe, K.E., Waller, B.M., \& Liebal, K. (2011). The language void: The need for multimodality in primate communication research. Animal Behaviour, 81(5), 919-924.

[^4]:    ${ }^{1}$ This work has benefited greatly from discussions with Laura Martignon

[^5]:    ${ }^{1}$ http://narrative.csail.mit.edu/

[^6]:    How does perceived consensus reduce the biasing influence of worldview on climate change attitudes?
    John Cook (University of Queensland)

[^7]:    * $p<.05 ; \mathrm{N}=108$

[^8]:    BCM Bienenstock, Cooper, Munro learning rule; Eq (7)
    hPES Homeostatic Prescribed Error Sensitivity; Eq (9)
    NEF Neural Engineering Framework; see Theory
    PES Prescribed Error Sensitivity; Eq (6)
    SPA Semantic Pointer Architecture; see Theory
    STDP Spike-timing dependent plasticity (Bi \& Poo, 2001)

[^9]:    *Digits from MNIST: http://yann.lecun.com/exdb/mnist/

[^10]:    ${ }^{1}$ Inflectional markers that are overtly distinct from the other plural or singular forms of the same declension class and hence clearly identify the surface form as case marked.

[^11]:    ${ }^{2}$ Online: http://celex.mpi.nl/
    ${ }^{3}$ Surface frequency denotes the token frequency of a word form (such as table) (Schreuder \& Baayen, 1997: 119). Stem frequency (Schreuder \& Baayen, 1997: 120) is derived by cumulating frequencies of inflectional variants of a word, which have also

[^12]:    ${ }^{4}$ Fries (frieze), Schahs (shahs), Gemischen (mixtures.DAT)

[^13]:    ${ }^{5}$ Family size had to be excluded because it was highly correlated with family frequency $(\mathrm{r}=-0.82)$. Type frequency cannot be considered in the same model as group because type frequencies are tied with group membership (their correlation is 1 ).

[^14]:    ${ }^{1}$ Point-only trials were not included given that such trials would pit gaze and point against each other; moreover, pointing one place and looking another is rarer in day-to-do life.

[^15]:    ${ }^{2}$ For details on how subject means were computed, please see Bergelson \& Swingley, 2012.
    ${ }^{3}$ All subsequent tests are two-talked Wilcoxon tests unless noted otherwise; all " $\mathrm{X} / \mathrm{X}$ infants" results indicate the number of infants with positive performance. M is mean; Mdn is the pseudo-median estimate of the Wilcoxon test.

[^16]:    ${ }^{4}$ In a linear model pointing behavior is significantly predicted by both age, and by vocabulary residualized by age, suggesting that vocabulary predicts behavior above and beyond age alone.

[^17]:    ${ }^{5}$ Within each corpus's frequency counts, we did not constrain word class; 'kiss' occurring as a noun or verb was counted for 'kiss', just as 'apple blueberry sauce' was counted for 'apple'.

[^18]:    ${ }^{6}$ Separate analysis of non-nouns as verbs and performatives showed the same overall pattern as non-nouns combined.

[^19]:    ${ }^{1}$ This is specifically the case for decisions about where within a word to target the eyes. The control of decisions about which word to target is known to reflect cognitive processing.

[^20]:    ${ }^{2}$ As the target word moved to different absolute positions on the screen depending on the shift condition, for the purposes of blink exclusion, we used a target word region defined as the union of the locations occupied by the target word across all shift conditions.

[^21]:    ${ }^{3}$ The space prior to the word (position 0 ) was thus excluded.
    ${ }^{4}$ These probabilities come with a caveat: on the standard oculomotor account, many attempts to refixate a short word and many attempts to skip a long word will fail. Thus, on the standard account, these probabilities underestimate the true rate of unintentional target word fixations, which may be a substantial portion of trials. We return to this point in the Conclusion.

[^22]:    ${ }^{1}$ The 6-state solution for new foils effectively splits up two of the stages into shorter stages. Although this might explain the new foils in themselves better, our interest is explaining associative recognition.

[^23]:    ${ }^{1}$ Three additional tasks were performed on days 1 and 5 . To measure trait motivational tendencies, participants completed the Behavioral Activation System / Behavioral Inhibition System scales (BIS/BAS; Carver \& White, 1994). Participants also completed a finger-tapping task as a performance-based measure of manual motor asymmetries. Finally, an N-back task was performed as a measure of working memory. Results from these tasks do not bear on mood, and have not been analyzed.

[^24]:    ${ }^{1}$ The subject is marked by nominative case whereas the object is marked by accusative case at the determiner and noun, respectively. For masculine but not for feminine nouns, the subject and object status can be unambiguously differentiated via case marking (e.g., der Uhu/the ${ }_{[\text {NOм }} \operatorname{owl}_{[\text {[NOM] }}$ (subject) vs. den Uhu/the ${ }_{[A C C]} \mathrm{owl}_{[\mathrm{ACC}]}$ (object)).

[^25]:    ${ }^{1}$ See Appendix for formal definitions of the two filters.

[^26]:    ${ }^{2}$ See Appendix for the set of phonetic features.

[^27]:    ${ }^{3}$ See Appendix for the full set of feature values.

[^28]:    ${ }^{1}$ Kay and McDaniel (1978) actually proposed two closely related hierarchies, one of which is shown here for illustration. Boster's analyses considered both.

[^29]:    ${ }^{2}$ Levinson et al. (2003) were careful to note that their proposal is based on synchronic, not diachronic, data; they therefore advanced their proposal as a hypothesis concerning possible patterns of historical language change, not as a firm claim about such patterns.
    ${ }^{3}$ As in the case of color, our interpretation of Levinson et al.'s (2003) proposal, based on their Figures 16 and 18, reduces to two distinct hierarchies, one of which is shown here for illustration, but both of which we use in our analyses. Both of these hierarchies are specified further in the analyses below.

[^30]:    ${ }^{4}$ This model is most clearly articulated in Levinson et al.'s (2003) Figure 18, but where the order of divisions is underspecified in this diagram (e.g. the relative order of IN/INSIDE vs. NEAR/AT categorical splitting), we rely on the ordering of the implicational scale presented in Figure 16 for clarification.

[^31]:    ${ }^{5}$ The two alternative versions of the model that we considered differ in whether more specific ON or UNDER categories form first. In addition to these two alternatives, the model also varies in whether OVER or NEAR categories are distinguished earlier. However, these distinctions are made with respect to the category AT, which is not included in our analysis because as a residual category, it does not appear to have a meaningful cross-linguistic focus. Thus, the NEAR/AT distinction is not available to our participants, which in turn prevents variability in whether OVER or NEAR is distinguished first.

[^32]:    ${ }^{1}$ Cf. Snedeker \& Yuan, 2008 for more on children's sentence processing.

[^33]:    ${ }^{2}$ The 27 bilingual children heard and used English at least 50\% of the time, as reported by parents. This proportion is representative of the area where we collected data, which has a large population of fully- or partially-bilingual speakers. We replicated all analyses below, excluding bilingual speakers and saw essentially no difference in the qualitative or quantitative pattern of results reported below.

[^34]:    ${ }^{3}$ We excluded occasional emphatically lengthened monosyllabic words like [wav:] 'woooow!' from the calculation of the average and the resulting length manipulation.

[^35]:    ${ }^{4}$ See a sample of the final videos and data from all conditions in one version at: http://langcog.stanford.edu/materials/anticip.html

[^36]:    ${ }^{5} \mathrm{We}$ assume here that it takes children $\sim 333 \mathrm{~ms}$ to plan an eye movement, following Fernald and colleagues (2008). A significant shift in gaze to the next speaker before 333 ms of speech indicates that the eye movement was planned before the response began. We saw anticipation in all conditions, so below we compare anticipation across conditions by analyzing looks to the upcoming speaker at the onset of the response turn.

[^37]:    ${ }^{6}$ Longer gaps give more time for gaze shift.
    ${ }^{7}$ There were also marginal effects of Age and Condition overall ( $t=-1.85$ ). There were not enough non-question switches under 200 ms to test for effects of switch type in this model.

[^38]:    ${ }^{1}$ In all our experiments, the categories are novel and equally probable, so we omit the prior probability component of Bayesian reasoning. We continue to use the term Bayesian because of the common feature of Bayesian models of induction that predictions are integrated across multiple categories, weighted by their likelihood.

[^39]:    ${ }^{1}$ We decided to restrict the correct and wrong HMMs to two hidden states because there was not enough data to train three hidden states in the wrong HMMs.

[^40]:    ${ }^{1}$ We in fact verified this through simulations using their model.

[^41]:    ${ }^{2}$ This may be due to the need to compress brain size for successful birthing, and lower compressability of grey matter.

[^42]:    ${ }^{3}$ Lewis and Elman used 10 hidden units and 2 interhemispheric units; we increased these numbers to facilitate task-learning. Select simulations run with their parameters showed similar effects.
    ${ }^{4}$ We used a gradient $\left((y-t)^{3}\right)$ that penalized local minima where a many patterns were learned quickly by sacrificing a few patterns which were not learned at all.
    ${ }^{5}$ Parameters: $T=5, \alpha=10^{-3}, \kappa=10^{-2}$

[^43]:    ${ }^{6}$ noise $=\mathcal{N}\left(2 * 10^{-4} *\right.$ delay, $\left.1 * 10^{-4}\right) ;$ constant chosen such that $\mu$ is $\approx 2 \%$ of average activity over all units.

[^44]:    ${ }^{1}$ An obvious exception to this are species capable of vocal learning - particularly song-birds and cetaceans. We suspect it is no coincidence that these species (a) have combinatorial signalling systems, and (b) make use of cultural transmission. See Feher et al. (2009) for an example of the important role culture has in bird-song.
    ${ }^{2}$ This refers to a particular type of learning where a behaviour is acquired by observing another who also acquired that behaviour the same way (Kirby \& Hurford, 2002).

[^45]:    ${ }^{3}$ The same metric was used to give feedback to participants, but was first recast into a similarity score by computing 1-error, and then given as accuracy by percentage. An error score of 0.17 translates to a similarity score of 0.83 , and appeared to participants as $83 \%$ correct.

[^46]:    ${ }^{4}$ To do this, we used tools from the Zlib library: www.zlib.net

[^47]:    ${ }^{5}$ This analysis is intended to be illustrative rather than definitive, and is based on the sample of 6 sequences shown here, not the whole set. As we can see in Figure 2 the sequence sets are not fully stable and are still undergoing change, making a more detailed structural analysis difficult.

[^48]:    ${ }^{1}$ The same noise process applies to context features; we simply assume that all context features are sampled at once at the beginning of the trial, rather than over time.

[^49]:    ${ }^{2}$ The effect of the logarithm is simply to put positive and negative changes on the same scale.

[^50]:    ${ }^{1}$ Ability estimates were generated using expected a posteriori scoring.

[^51]:    ${ }^{1}$ Joint first authorship

[^52]:    ${ }^{1}$ Many factors have been identified as contributing to production preferences (see e.g. Jaeger \& Tily, 2011). Here we take participants' empirically estimated (Exp. 3) relative preference for shorter over longer messages as a measure of subjective production cost.

[^53]:    ${ }^{2}$ Exps. 1 and 2 here were identical to Exps. 1 and 2 in Degen and Franke (2012) with the difference that we use linguistic instead of pictorial messages.

[^54]:    ${ }^{3}$ Normally one would specify prior probabilities of states, but we assume that all referents are (believed to be) equiprobable. One would also normally specify utilities, but since we assume interlocutors want to cooperatively identify the referent, utilities are given by identity matrices that cancel out where normally they'd be relevant.
    ${ }^{4}$ As for notation, if $A$ is a matrix, let $\mathbf{T}(A)$ be its transpose, and $\mathbf{N}(A)$ its row-normalization. If $A$ and $C$ are matrices, $A C$ is their matrix product. We will also use a non-standard operation on matrix $A$, namely $\max \operatorname{row}(A)$ which returns a binary matrix with the same dimensions as $A$, such that max $\operatorname{row}(A)_{i j}=1$ if $A_{i j}=\max _{k} A_{i k}$ and 0 otherwise. We abuse notation by assuming that vectors are implicitly coerced if combined with matrices in standard arithmetic operations. So, for instance, $B-c$ is obtained by subtracting $c$ in each row.

[^55]:    ${ }^{5}$ The quantal response function is also known as logit choice rule, as soft-max function (Sutton \& Barto, 1998) or, if $\lambda=1$ as Luce's choice rule (Luce, 1959).

[^56]:    ${ }^{6}$ This is because $x=\log \frac{P(\text { costly })}{P(\text { cheap })}-\log \frac{P(\text { cheap })}{P(\text { costly })}$ reduces to $x=$ $\log (\exp (-\lambda c))-\log \frac{1}{\exp (-\lambda c)}=-2 \lambda c$.

[^57]:    ${ }^{1} \mathrm{http}: / / \mathrm{www}$. eecs.berkeley.edu/Research/Projects/CS/vision/gro uping/segbench/

[^58]:    ${ }^{1}$ This explanation may still warrant further investigation, as unlike our position presented here, people might not use the cue time for a slow removal process, but only to find the to-be-removed items. However, this explanation cannot easily account for the reduction of the repetition/ similarity effects observed in Experiments 1 and 2.

[^59]:    ${ }^{1}$ These were determined with a non-parametric $t$-test, the Whitney-Mann U-test, since the distributions of entropy scores are non-normal.
    ${ }^{2}$ After correction for multiple comparisons, this is not approaching significance.

[^60]:    ${ }^{3}$ marble $1\left(m_{1}\right)$ refers to the blue marble in the one-item task, and to the blue, brown, black, red, olive, and purple marbles in the six-item task.
    ${ }^{4}$ The raw log likelihoods should not be compared between tasks, because there are a different number of observations per task. This is corrected for in the prediction percentages, which are comparable between tasks.

[^61]:    ${ }^{5}$ This bootstrap model, which defines the dynamics of evolutionary drift, is equivalent to a Bayesian MAP model with $\alpha=0$. See Reali \& Griffiths (2010) for the proof.

[^62]:    ${ }^{6}$ Both of the empirical transition matrices are ergodic.

[^63]:    ${ }^{1}$ Previous studies have shown that children's head movements in the horizontal dimension are approximated by (though are slightly lagged by) their head movements (Yoshida \& Smith, 2008). Our own experience with the current apparatus ratifies these conclusions for the horizontal field but suggests that head movements in the ver-

[^64]:    tical field are less reliable. Hence, these studies may run the risk of underestimating the proportion of faces actually seen by children.

[^65]:    ${ }^{2}$ Since orientation was coded via body posture, faces seen while the caregiver was behind the child were due to children looking over their shoulder.

[^66]:    ${ }^{1}$ It is worth noting that the Greedy heuristic is not the only model that seeks to account for Olson's theoretical stance. A possible alternative is not to proceed incrementally, but to compare entire descriptions of increasing length, starting from those consisting of a single property, until the target is distinguishing. This would ensure a description that contains no more information than is absolutely required, something the Greedy heuristic can in fact only approximate. However, this 'Full Brevity' model, also discussed by Dale (1989), is unlikely to be psycholinguistically realistic, for three reasons: (i) it is computationally extremely expensive, since it potentially involves search through all available combinations of properties (Reiter, 1990); (ii) speakers tend to overspecify, as we discuss below; (iii) an implementation of Dale's Full Brevity model has been shown to produce output that does not match that of human speakers, compared to algorithms that are incremental in nature (van Deemter, Gatt, van der Sluis, \& Power, 2012).

[^67]:    ${ }^{2}$ Frank and Goodman (2012)'s discussion employs the term word rather than property; however, little hinges on the difference for present purposes.

[^68]:    ${ }^{1}$ This representation assumes that basic perceptual preprocessing has taken place (e.g., the correspondence problem has been solved).

[^69]:    ${ }^{2}$ As described in Blei et al. (2010), trees drawn from the nCRP can be infinitely deep, but we impose a maximal depth for simplicity.

[^70]:    ${ }^{3}$ The latent motion components can be marginalized analytically using properties of Gaussian processes.

[^71]:    ${ }^{4}$ Note that the model does not explicitly represent rotation but instead represents the tangential motion component in each time step.

[^72]:    $5_{\text {http: }} / /$ mocap.cs.cmu.edu/

[^73]:    ${ }^{1}$ The data from one participant were excluded due to excessively low within-dimension performance on both tasks (both slopes were two standard deviations below group averages).

[^74]:    ${ }^{1}$ The probabilism is intended to capture all factors that contribute to the imperfect measurement value of a specific post at a randomly chosen time. This includes both limits in the precision of measurement as such and exogenous factors that affect the cost but are unknown to the consumers.

[^75]:    ${ }^{2}$ In the multiple-comparisons, we report raw $p$-values assuming $\alpha=.05$. The Bonferroni corrected $\alpha$-level is app. . 008 . The same is true for the multiple-comparisons reported below.

[^76]:    ${ }^{1}$ The reason for testing their combined effect was that these two factors are highly interrelated (i.e., repeated encounters with the same partner, by default, bring along the opportunity of RB as each player would know what the other player has done so far).

[^77]:    ${ }^{2}$ The rejection rates for the alternative distributions (5/5), (2/8), and (10/0) were $2 \%, 6 \%$ and $82 \%$ respectively in Experiment 1.
    ${ }^{3}$ The rejection rates of the alternative distributions in the (5/5), (2/8) and (10/0) games were as follows: Nobody rejected the $(2 / 8)$ distribution and only one participant rejected the $(5 / 5)$ distribution. Almost $96 \%$ rejected the (10/0) distribution.

[^78]:    ${ }^{1}$ Geoffrey Bird is now at Institute of Psychiatry, King's College London.

[^79]:    ${ }^{\text {a }}$ Des $=$ Desirable information; Undes = Undesirable information

[^80]:    ${ }^{2}$ This enables a more appropriate normative analysis of the data (see Shah et al., 2013), which we do not elaborate on here.
    ${ }^{3}$ Controlling for these factors did not change the results.

[^81]:    ${ }^{1}$ The two previous ERP studies investigated sentences of the form some elephants have trunks - literally true but rendered infelicitous by scalar implicature (Nieuwland, Ditman, \& Kuperberg, 2010; Noveck \& Posada, 2003). The final word evokes an N 400 relative to the final word in felicitous sentences (some dogs have spots), at least if the sentences are carefully matched.
    ${ }_{2}$ Panizza, Chierchia, and Clifton Jr. (2009) report an eyetracking-while-reading study with a similar manipulation, but involving number. The relationship between number and scalar implicature is complex, unclear, and controversial.

[^82]:    ${ }^{3}$ The choice of threshold (e.g., 1.96) affects the type of clusters found - low thresholds are better at detecting broadly extended but weak effects - but does not affect robustness to multiple comparisons. Other threshold resulted in similar findings.

[^83]:    ${ }^{4}$ Breheny, Katsos, and Williams (2006) report longer reading times for scalar triggers in contexts expected to promote scalar implicature calculation. However, the contextual manipulations are uncontrolled, making its results difficult to interpret. In the case of Exp. 3, the manipulation is fully confounded with a repeated name penalty, sufficient to explain their results.

[^84]:    ${ }^{1}$ Specifically, each time the simulate button was clicked a random number between 1 and 10000 was generated. If the number was less than 511 then the woman had cancer.

[^85]:    ${ }^{2}$ These qualitative results remained unchanged when deviation scores in the sampling condition were recomputed against a normative solution that replaced the stated base rate of $1 \%$ with the base rate implied by the sample drawn by each participant (i.e. the observed proportion of positive cancer trials).

[^86]:    ${ }^{3}$ These qualitative results remained unchanged when deviation scores were recalculated using individual cancer base rates implied by the sampling information instead of the stated rate of $1 \%$.

[^87]:    ${ }^{1}$ Notwithstanding a body of theoretical work (see e.g. Markman and Stilwell, 2001).

[^88]:    ${ }^{2}$ Although concreteness is well understood intuitively, it lacks a universally accepted definition. It is often described in terms of reference to sensory experience (Paivio et al., 1968), but also connected to specificity; rose is often considered more concrete than flora. The present work does not address this ambiguity.

[^89]:    ${ }^{3}$ E.g. she states "What is important to this view is not how abstract words come to have weaker connections [to associated information]...only that they generally do" (1991, p. 243).

[^90]:    ${ }^{4}$ Subjects were asked to consider their idea of synonymy and then rate the "similarity of meaning" of word pairs (1965, p. 628).

[^91]:    ${ }^{1}$ The 3-dimensional solution (stress $=.17$ ) provided little additional information. The third dimension could be interpreted as reflecting a distinction between metric (e.g., far from, near) and nonmetric (e.g., above, to the left of) prepositions, but this dimension also distinguished the four clusters in Figure 1 reasonably well.

[^92]:    ${ }^{2}$ Other dimensionality reduction techniques (hierarchical clustering and principal components analysis) yielded similar results, suggesting that the clusters are not an artifact of MDS (Holmes, 2012).
    ${ }^{3}$ In addition to identifying clusters within the 2-dimensional similarity space, the dimensions themselves may also be interpreted. These dimensions seem to capture broad distinctions among spatial relations in the world. The $y$-axis reflects a distinction between topological and projective relations (Levinson et al., 2003); most of the prepositions in the above-below and in clusters refer to relations between contiguous objects, whereas those in the front-back and leftright clusters specify a frame of reference. The x -axis is less easily interpreted. Several researchers have noted that the above-below and front-back axes are perceptually asymmetric with respect to canonical body position, whereas the left-right axis is perceptually symmetric (e.g., Clark, 1973). However, the in cluster is not well captured by this distinction; relations of containment and proximity are not readily characterized in terms of symmetry.
    ${ }^{4}$ The in category was not included because it was the only category that did not contain terms with opposite meanings, making CP more difficult to assess than for the other categories.

[^93]:    ${ }^{1}$ These concerns do not necessarily undermine the distinction between personal and impersonal dilemmas, for which there is independent evidence (Moore, Clark \& Kane, 2008). The two claims are independent; one can accept the personal-impersonal distinction without thinking that people's moral judgments in these cases are driven by differences in the emotions the situations evoke.

[^94]:    ${ }^{1}$ Participants in the study of Feldman et al. (2010) were prompted to give an answer immediately after the presentation of the stimulus.
    ${ }^{2}$ Tohill and Holyoak (2000) used a serial subtraction task, while Feldman et al. (2010) - a "public speech" procedure.
    ${ }^{3}$ STAI for Tohill and Holyoak (2000) and a self-assesment 5point scale for the Feldman et al. (2010).
    ${ }^{4}$ The Feldman et al. (2010) study used geometric figures and the relations between them, while the Tohil and Holyoak (2000) study, used much more complex everyday situations, depicted in two pictures, which usually involved more than 3 actors and a number of diverse relations between them.
    ${ }^{5}$ Match-to-sample task (Feldman et al. (2010)) - one sample and two distinct targets (one superficially similar and one relationally similar) vs. a cross-mapping task, where participants should choose which option in the bottom picture "goes with" the object, pointed from the experimenter. Both the target object and the options were embedded in complex relational structures. The number of options varied between 3 and 6 alternatives for answer, etc.

[^95]:    ${ }^{6}$ Seven of them were devised by Markman and Gentner (1993) for a study of structural alignment; two were created later and used in the experiments on how anxiety influences analogical mapping by Tohill and Holyoak (2000). Both research groups kindly provided their stimuli for our study. All original stimuli were drawn again, carefully preserving the key relations between the objects.

[^96]:    ${ }^{7}$ State anxiety in the control condition of our experiment is 2.65 higher than in the Tohill and Holyoak's study, and 2.01 lower in the anxiety condition.

[^97]:    ${ }^{1}$ Concepts here are denoted by the name of their referent typed in upper case letters.

[^98]:    ${ }^{2}$ For example, learning the idea of MOVING THE KING TWO SQUARES TOWARDS A ROOK ON THE PLAYER'S FIRST RANK, FOLLOWED BY MOVING THE ROOK ONTO THE SQUARE OVER WHICH THE KING CROSSED (i.e., CASTLING in chess) may strike one as a case of bona fide concept learning.
    ${ }^{3}$ Many examples support this generalization from the formation of perceptual prototypes to the construction of schemas, scripts, mental models, and propositional representations. As Carey (2009) says, it is a truism that all learning involves building new representations from antecedently available ones.
    ${ }^{4}$ I.e., adding new operators, predicates, etc. with content that no combination of the earlier set of primitives could represent. One example is adding the modal operators 'possibly' and 'necessarily' to classical propositional logic.

[^99]:    ${ }^{5}$ Fodor's circularity problem has been described above. In specifying the meaning of some concept via a hypothesis, we must understand and use the very concept of which we formulate the hypothesis (Fodor, 1990, 1998). Goodman's problem of induction is that given a body of empirical data, there are infinitely many ways to inductively generalize from it, and learning theories need to explain how we choose our preferred ones (Goodman \& Putnam, 1983).

[^100]:    ${ }^{6}$ Note also that actions like coupling, or noticing what takes minimum effort are pretty close to a sensorimotor vocabulary. I suggest no return to empiricism, still it is worth noticing that sensorimotor activity might play some mediating role in bootstrapping mathematical concepts. To this extent Piaget's view of development may be correct.

[^101]:    ${ }^{1}$ The value of relational language and comparison might in some cases be related because using the same term for different examples can invite comparison of these examples (e.g., Gentner, 2003, 2010; Gentner \& Medina, 1998; Gentner \& Namy, 1999). However, our focus here will be on the individual effects of relational language and comparison.

[^102]:    ${ }^{2}$ All of the post-hoc tests were Tukey HSD tests.

[^103]:    ${ }^{1} 12$ additional toddlers were recruited but never included in the study because they declined to participate in a warm-up task, in which the child was asked to choose between two stuffed elephants.

[^104]:    ${ }^{2} 4$ additional two-year-olds and 3 three-year-olds were recruited but never included in the study because they declined to participate in a warm-up task, in which the child was asked to choose between two stuffed elephants.

[^105]:    ${ }^{1}$ phasespace.com/impulse_motion_capture.html

[^106]:    ${ }^{2}$ The first frame was dropped from each dataset to match the number of frames of data with the number of frames of derivative.

[^107]:    ${ }^{1}$ http://github.com/lmjohns3/driving-simulator

[^108]:    ${ }^{1}$ Note that parts of the empirical research reported in this paper have been published in Gerjets, Kammerer, and Werner (2011) who compared the two thinking-aloud groups (irrespective of prior knowledge). The data of the silent group were gathered at the same time as the data of the two thinking-aloud groups.

[^109]:    ${ }^{1}$ See Taylor, Devereux, Acres, Randall, and Tyler (2011) for details.

[^110]:    ${ }^{2}$ We generated 50 and 150 clusters for the concepts and features respectively using hierarchical clustering on WordNet.
    ${ }^{3}$ The multi-class implementation, SVM Multiclass (v. 2.20).
    ${ }^{4}$ Potential features were defined as all adjectives and singular/plural nouns in a sentence.
    ${ }^{5}$ Due to memory constraints associated with the very large number of training instances, we were only able to train our UKWAC models on one third of the UKWAC corpus; we selected every third learning pattern for training.

[^111]:    ${ }^{6}$ Only a small proportion of our triples derived their relations in this way; at this point, in our training sets we had assigned relations to over $94 \%$ of triples from our Wikipedia corpus, and $97 \%$ from the UKWAC corpus.

[^112]:    ${ }^{7} 300$ factors, using the TASA corpus at 1 sa.colorado.edu.
    ${ }^{8}$ The correlation confidence intervals, calculated using Fisher transformations (Fisher, 1915), are given at the $95 \%$ level of confidence, and two-tailed $p<0.05$.

[^113]:    ${ }^{9}$ i.e. both 'correct' and 'plausible' triples were counted as correct, while 'related' or 'wrong' triples were considered incorrect.

[^114]:    ${ }^{1}$ Khetarpal et al. (2009) used a slightly different formalization of these ideas. We use this one because it maps cleanly onto the communicative scenario sketched above, in which a listener tries to understand a speaker's meaning. The results reported below remain qualitatively unchanged if the original formalization is used instead.

[^115]:    ${ }^{2}$ Regier et al. (in press) presented slightly different extensions of English categories against this map, one of which was not connected. We have chosen these extensions instead because (1) they allow English categories to be connected in this map, (2) that connectedness allows us to include English in our upcoming analyses, and (3) these extensions agree well with our linguistic intuitions.

[^116]:    ${ }^{3}$ A followup study found that these pile-sorts were broadly similar across the two languages, although they did reflect the sorter's native language to some extent (Khetarpal et al., 2010).

[^117]:    ${ }^{1}$ Some work, e.g. Griffiths and Kalish (2007), has also considered models of type (1b), but only for the discrete parameter case.

[^118]:    ${ }^{2}$ This is equivalent to sampling from $\pi_{t}(p)$ and generating an example from the distribution implied by the value of $p$ chosen.
    ${ }^{3}$ These models are equivalent to special cases of 'blending inheritance' models of cultural evolution of a quantitative character (Boyd and Richerson, 1985: 71ff).

[^119]:    ${ }^{4}$ This contrasts with the common statement that 'blending inheritance' reduces variance of a quantitative trait over time (Boyd and Richerson, 1985: 75). However, stable or increasing variance are possible for particular cases of Boyd and Richerson's model, such as the case considered here where each learner has a single 'cultural parent' and there is noise in estimating the parent's cultural model.

[^120]:    ${ }^{5}$ The ML estimate of $\hat{p}$ in the no-prior case above, $\mu_{a}-\bar{y}$, can thus be thought of as the Gaussian-prior estimate when the prior is very flat relative to the dispersion of the phonetic category $\left(\tau \gg \sigma_{a}\right)$.

[^121]:    ${ }^{6}$ A additional extension to be explored is horizontal transmission. In the present models, learners do not receive input from members of their own generation, but this could impact the dynamics as well.

[^122]:    ${ }^{1}$ Voice onset time (VOT) is defined as the time between the release of a stop consonant and the start of vowel phonation.

[^123]:    ${ }^{2}$ It is typical of the cue-distractor task that the distractor is not itself a valid output, and so effectively has zero prior and posterior probability.

[^124]:    ${ }^{1}$ The first two authors contributed equally to this work

[^125]:    ${ }^{1}$ All t-tests are two-tailed.

[^126]:    ${ }^{1}$ "But first" was ambiguous because it could refer either to the time period before the first or the second movement. Thus, we also performed analyses only on target referents that were mentioned second, for whom the ambiguity did not apply, which showed a similar pattern of results.

[^127]:    ${ }^{1}$ The authors are listed in alphabetical order. This paper is based

[^128]:    ${ }^{2}$ See, e.g.,Woodward 2003, 362-364.

[^129]:    ${ }^{3}$ Code obtainable on request.

[^130]:    ${ }^{1}$ Logistic regression models of response choices were fit by the Laplace approximation, whereas linear regression models of response times and fixation proportions were fit using restricted maximum likelihood estimation. Fixed effects included prosody condition (Noun- vs. Verb-focus), display type (one- vs. two sets of related pictures), and standardized trial number. Analyses of fixation proportions additionally included picture type. We also included random by-subject and item intercepts as well as slopes for the interaction between prosody condition and picture type. To minimize the risk of over-fitting the data, fixed effects were removed stepwise and each smaller model was compared to the more complex model using the likelihood ratio test (Baayen, Davidson, \& Bates, 2008).

[^131]:    ${ }^{1}$ Hereafter: double-quotation marks are used for quoting speech, with phonetic and prosodic specification. Capital letters represent prosodic emphasis. Square brackets ([ ]) are used for example words or sentences abstracted away from acoustic detail, e.g., [It is raining outside] can be said as, "It's RAINING outside!", "It IS raining outside!", et cetera.)

[^132]:    ${ }^{1}$ The plotted numerical values change very little with the inclusion of incorrect trials and the pattern of statistical results remains identical.

[^133]:    ${ }^{2}$ Please see Davelaar et al. (2005) for computational details.

[^134]:    ${ }^{1}$ For whole body motion considered here, we simply integrated the motion energy over the entire ROI without subdividing the image region. An analysis at smaller scales might necessitate an integration over smaller overlapping patches.

[^135]:    ${ }^{1}$ In this analysis, we focus only on situations where a single referring word is used.
    ${ }^{2}$ This assumption is made for purposes of simplicity only, as a variety of our previous work has explored the use of social and pragmatic information in biasing the distribution over intended referents (Frank \& Goodman, 2012).

[^136]:    ${ }^{3}$ The 1-many constraint is related to the concept learning model proposed by Goodman, Tenenbaum, Feldman, and Griffiths (2010) using a disjunctive normal form grammar.

[^137]:    ${ }^{1}$ Small capital letters are used to distinguish concepts from objects.

[^138]:    ${ }^{2}$ This manipulation was motivated by the observation that different types of ambiguity license different inferences. To illustrate this, imagine a learner in a confounded across context. The learner observes a situation with two apples and a situation with two oranges. In each situation, she hears "dax bren nes". The referent is clear in each individual situation - apple and orange, respectively - and the learner might infer that this phrase corresponds to a superordinate category, such as FRUIT. In the confounded within context, the learner observes two situations, both containing an apple and an orange, and again hears "dax bren nes" in each. Unlike in the across case, a learner in this context would have no information about how to correctly map the meaning of this phrase, since the context is consistent with both a subordinate and superordinate interpretation. Different generalization patterns are thus predicted in the confounded across and within conditions.

[^139]:    ${ }^{1}$ The absolute value of this slope is not meaningful because of the relatively large scale of the items (1 to 36 ) compared to the range of the itemwise correlation.

[^140]:    ${ }^{2}$ We could rearrange the items in order of difficulty, but this would bias the analysis towards the result we observe. Instead, we present the items in the order they were presented for parity with Unsworth and Engle (2005).
    ${ }^{3}$ These random effects models allow the decrease in accuracy across items to begin at a different level of accuracy or, in addition, decrease at a different rate for each participant, respectively

[^141]:    ${ }^{1}$ In a way, it can be argued that, thanks to material culture, human beings have managed somehow to take advantage of most environments on Earth (and outside of it), but it is a partial success which requires a continuous implementation of resources and knowledge in order to maintain those achievements as persistent.
    ${ }^{2}$ It is important to note recent research based on Schrödinger's focusing on energy, matter and thermodynamic imbalances provided by the environment, draws the attention to the fact that all organisms, including bacteria, are able to perform elementary cognitive functions because they "sense" the environment and process internal information for "thriving on latent information embedded in the complexity of their environment" (Ben Jacob, Shapira, and Tauber (Ben Jacob, Shapira, \& Tauber, 2006, p. 496)). Indeed Schrödinger maintained that life requires the consumption of negative entropy, i.e. the use of thermodynamic imbalances in the environment.

[^142]:    ${ }^{3}$ Attention is drawn for the first time to the idea of niche construction by important researchers like Schrödinger, Mayr, Lewontin, Dawkins, and Waddington. Firstly in the field of physics and subsequently in the field of the theory of evolution itself. Waddington particularly stressed the influence of organism development.

[^143]:    ${ }^{4}$ This perspective has generated some controversies, since the extent to which modifications count as niche-construction is not clear, thus entering the evolutionary scene. The main objection regards how far individual or even collective actions can really have ecological effects, whether they are integrated or merely aggregated changes. On this point, see (Sterelny, 2005) and the more critical view held by (Dawkins, 2004). For a reply to these objections, see (Laland, Odling-Smee, \& Feldman, 2005).

[^144]:    ${ }^{5}$ The regulatory dimension of structural violence is often diluted in the pervasive form of narratives conveying "moral templates": the fairytales that are told to children from early youth, novels, plays, dramas and - more recently - motion pictures are all involved in the dissemination of some moral, economic or spiritual teaching, but they also circulate via gossip (Baumeister, Zhang, \& Vohs, 2004).

[^145]:    *JK and TM contributed equally to this work.

[^146]:    ${ }^{2}$ Initial analyses found no effects of participants' handedness (Dehaene et al, 1993); it was thus removed from further analyses.

[^147]:    1 This is assuming categorization at the same level in a categorization hierarchy and within the same domain. Presumably, any given object can belong to multiple categories along a subordinate to superordinate spectrum (e.g., robin, bird, animal) and can be categorized in different ways depending on the intention of the categorizer (e.g., parakeet versus pet animal). We are discussing here categorization decisions that are equated across these dimensions (e.g., robin versus blue jay).

[^148]:    ${ }^{2}$ It should be noted that this explanation does not account for differences between the ambivalent and uncertain conditions.

[^149]:    ${ }^{1}$ Numbers in angled brackets indicate the utterance number in the original corpus.

[^150]:    ${ }^{1}$ Note that both measures are not independent of each other; selection and error rates add up to 1 .

[^151]:    ${ }^{1}$ New York University, Department of Psychology, 6 Washington Place, New York, NY 10003 USA
    ${ }^{2}$ Massachusetts Institute of Technology, Department of Brain and Cognitive Sciences, 77 Massachussetts Ave., Cambridge, MA 02139 USA

[^152]:    ${ }^{1}$ Because our study was conducted in Germany we used the corresponding German words "selten", "gelegentlich", "häufig", and "fast immer". Note that Bocklisch et al.'s (2012) study was also conducted in Germany with estimates given for the very same (German) terms.

[^153]:    ${ }^{2}$ Note, all symptoms does not mean that subjects were presented with all symptoms of a trial at each time, but with all symptoms that were relevant to the current judgment. Thus, in the all-symptoms condition subjects saw the sequence $\left\{S_{1}\right\},\left\{S_{1}, S_{2}\right\},\left\{S_{1}, S_{2}, S_{3}\right\}$, whereas in the single-symptom condition subjects were presented with the sequence $\left\{S_{1}\right\},\left\{S_{2}\right\},\left\{S_{3}\right\}$.

[^154]:    ${ }^{3}$ For this purpose, we used a grid search over a plausible set of values for $\delta$ between $1 \mathrm{e}-10$ and $1 \mathrm{e}+10$.

[^155]:    ${ }^{4}$ Remember that in the limit, if $\delta=\infty$ there is no decay; if $\delta=0$ the posterior probability depends on only the most recent symptom.
    ${ }^{5}$ More specifically, the algorithm proceeds as follows: (i) Given the current assignments of participants to clusters, find the $\delta$ parameter for each cluster

[^156]:    that minimizes the $M S E$ of the model predictions with respect to the average response profile of the subjects within the cluster. (For this purpose, we used a grid search over a plausible set of value for $\delta$.) (ii) Given the model predictions for the different clusters, reassign subjects to a cluster such that the correlation between the individual response profile and the model prediction is maximized. Then, iterate through (i) and (ii) until no participant changes cluster anymore.

[^157]:    ${ }^{1}$ The general validity of the PME is demonstrated using simple combinatorics (cf., Jaynes, 2003).

[^158]:    ${ }^{2}$ The general form of the solution is due to L. E. Boltzmann. For a sketch ot the derivation, see, e.g., Moscoso del Prado (2011).

[^159]:    ${ }^{3}$ The same effects reported here were also replicated for other corpora of English and French. These additional analyses are not reported here for brevity reasons.
    ${ }^{4}$ Piantadosi et al. (2012) report effects on length in syllables. I use phoneme-based lengths instead as these are more sensitive, but I also replicated the same effects using syllable-based lengths. Conversely, Piantadosi and colleagues also report that their effects also held when measuring length in phonemes.
    ${ }^{5}$ The log number of senses provides a better approximation to the psychologically relevant magnitude than does the raw count (cf., Moscoso del Prado, 2007). Note however that doing the calculation on raw counts of word senses did not result on different results.

[^160]:    ${ }^{6}$ http://www. speech.cs.cmu.edu/cgi-bin/cmudict

[^161]:    ${ }^{1}$ For a complete list of the property clusters of S1 and S2, sometimes called the 'Standard Menu', see Evans and Frankish (2009). I take the two-system hypothesis to be stronger than the existential claim that there are two systems of reasoning. The thesis is that cognition is divided into two systems and that each system has a certain set of properties associated with it-the properties on the Standard Menu.

[^162]:    ${ }^{2}$ Response toggle would be behavioral evidence for SCB rather than phenomenological evidence (as in the 'belief that p ', 'feels that not-p' cases).

[^163]:    ${ }^{3}$ The Monty Hall case is as follows. A subject is invited to a game show at which he or she may win a new car. The car is behind one of three doors. Behind the other two is nothing. The subject is told to pick a door. Monty, the game show host, opens one of the doors which the subject did not pick. The subject is then asked if he or she would like to change his or her answer. While it might at first seem that it does not matter, in fact there is a $2 / 3$ chance that the car is behind the door that the subject had not picked.

[^164]:    ${ }^{4}$ Parallel as opposed to sequential. In the former the two systems operate at the same time, in the later S1 operates first, then may shut down while S 2 performs its computations. Almost all two-system advocates endorse a form of the parallel view.
    ${ }^{5}$ On most accounts the S1 level of belief is the same as the implicit level of belief.
    ${ }^{6}$ This need not be a conscious suppression.

[^165]:    ${ }^{7}$ This well-known test involves the subject being shown the names of colors written in various colored prints. The subject is asked to give the name of the print-color. The faster and more accurately the participant gives the color of the print (rather than the word), the higher executive functioning and less cognitively depleted that participant, while the slower and less accurately the participant gives the name of the color of print, the lower the participant's executive functioning and more cognitively depleted the participant.

[^166]:    ${ }^{1}$ I will focus on the 3D Ebbinghaus illiusion because of the simplicity of the results, but it needs to be noted that the experimental conditions of this experiment have been criticized recently (Pavani et al. 1999, Franz 2001, 2003, Franz et al. 2000, 2003, Gillam 1998, Vishton 2004 and Vishton \& Fabre 2003, but see Haffenden \& Goodale 1998 and Haffenden et al. 2001 for a response and Briscoe 2008 for an overview. I focus on the 3D Ebbinghaus experiment in spite of these worries, but those who are moved by Franz et al. style considerations can substitute some other visual illusion, namely, the Müller-Lyer illusion, the Ponzo illusion, the hollow face illusion or the Kanizsa compression illusion, where there is evidence that the illusion influences our perceptual judgments, but not our perceptually-guided actions.

[^167]:    ${ }^{2}$ The distinction between the anatomical and the functional level also works in the other direction. The McIntosh and Lashleya experiment can be and has been taken to show that the dorsal stream is not informationally encapsulated (see Brogaard 2011 for analysis). But the experiments are not about the dorsal stream; they are about action-guiding vision. Interpreting them as having damning implications for the dorsal/ventral distinction is based on the equivocation of dorsal vision and action-guiding vision.

[^168]:    ${ }^{1}$ Throughout this paper, the term "rules" is used informally, and in this context refers to any regularity that people rely upon to guide inference. It is not intended to imply that the regularities in question correspond to explicitly represented, verbalizable rules.

[^169]:    ${ }^{2}$ The responses for these 6 rose monotonically across trials. This pattern makes sense if one assumes the bags are constrained to contain the same set of types. One participant spontaneously reported having made this assumption. This was not the intended interpretation of the task, but it is not an unreasonable one.

[^170]:    ${ }^{3}$ One reviewer noted that the gap between skewed and uniform does not increase across trials, and took this to imply that participants were not learning across trials. This is not correct: the trials differ systematically in tems of the number of types and tokens, making it difficult to draw any such inference. The key test of whether cross-trial learning takes place is to look at bag 7: if no cross-trial learning occurs, then responses should be identical for this bag in both conditions, because this stimulus was identical in the two conditions.
    ${ }^{4}$ Bag was coded as a categorical variable, and the random effect of bag-by-subject subsumes the random effect of subject.
    ${ }^{5}$ Analyses were run in $R$ version 2.15 .2 using the lme 4 package version 0.999999-0. Several other model specifications were

[^171]:    tried: none had lower BIC. Inspection of residuals suggests this model provides a good fit to the data. Nevertheless, it is important to note that the effect of condition is robust: it was significant in all model specifications tried, including several that analyzed only the binary version of the response variable (i.e., extrapolative vs nonextrapolative).
    ${ }^{6}$ The coin package (version 1.0-21) in R was used to compute an exact $p$ value in the presence of ties.

[^172]:    ${ }^{7}$ Strictly speaking, the samples should be constrained such that each type appears at least once among the $n$ observed marbles or the $100-n$ unobserved ones. For simplicity I have avoided introducing this additional constraint in this paper.

[^173]:    ${ }^{8}$ In principle there is no reason why a model with individual differences should be avoided: in practice, the computational difficulties in estimating such a model are somewhat severe.

[^174]:    ${ }^{1}$ We note that this lexicon is only used in input generation and evaluation, and not in the learning of the model.

[^175]:    ${ }^{1}$ However, for some emotion concepts, some features are indeed necessary (e.g. test anxiety involves thinking that an exam is going to take place) while others are merely characteristic.
    ${ }^{2}$ Prinz (2004) argues that some emotions are individuated by a calibration file, i.e. a mental file which includes different judgments that are unified because they all bear on the person's well-being in the same way. Prinz' strategy is to officially claim that the calibration file is part of the cause of the emotion but not constitutive for it. Since he at the same time describes the calibration file as being the essential feature for individuating an emotion, this is finally not a consistent position to hold.

[^176]:    ${ }^{3}$ Although the intentional object is crucial for distinguishing emotions from each other, it does not follow that the object is the bearer of the emotion (as an extended mind theorist might want to claim). While we remain neutral about the extended mind hypothesis in general, it is worth noting that our account does not

[^177]:    ${ }^{5}$ If someone is trained to inhibit any expressive sign of his emotion, we may still be able to recognize the emotion by inference, e.g. by noticing the force that someone uses to inhibit the expressions in the given context.
    ${ }^{6}$ We want to clarify the relation of our pattern account of emotion to earlier philosophical component theories of emotions by excluding a main deficit we do not inherit. Some component theories have analyzed emotions as beliefs plus desires plus (often) feelings (e.g. Lyons, 1980; Robinson, 1983). Whatever the relevant components are, Döring (2009) presents the following principle criticism: component accounts cannot explain why an emotion is (usually) experienced as a unitary conscious state. We fully acknowledge that an emotion is experienced as a unitary conscious state, but we will not provide an explanation of why and how this experience came about. Providing such an explanation is not necessary for a plausible philosophical account of emotions. Otherwise the problem would also affect Döring's account - she merely emphasize the unity of emotional experience but do not provide an explanation for it, either.
    ${ }^{7}$ This is at least true if we presuppose an identity theory of mental and physical states. We tend to do so but at the same time presuppose that the relevant physical state involves more than just a neural state, i.e. the whole body and maybe part of the environment.

[^178]:    ${ }^{8}$ Fear is fear no matter whether the observer knows that he deals with an anxious person or a courageous person but it is much easier to recognize knowing this and we can recognize it easier in an obviously dangerous setting than in a (seemingly) save setting.
    ${ }^{9}$ We allow for the case of emotion recognition without ascribing it to the other, i.e. without conceptualization but this case will not be the focus of this paper.
    ${ }^{10}$ For the sake of argument we can identify conceptual and linguistic abilities. For a fine-grained distinction between linguistic and non-linguistic concepts in the context of animal cognition, see Newen and Bartels, 2007.
    ${ }^{11}$ We are not discussing here what it means to possess concepts since it is sufficient to account for top-down influences to establish a general picture of perceiving emotions.

[^179]:    ${ }^{1}$ It must be acknowledged that this sentence contains an ambiguity with respect to the attachment of the adjunct during the meeting but as it occurs after the resolution of the filler gap dependency it should not have any affect on the results.

[^180]:    ${ }^{1}$ These specific terminologies are laid out here for the purpose of the present paper, rather than trying to resolve the emerging terminological debate within the field.

[^181]:    ${ }^{2}$ Due to experimenter error, one dyad's affiliative prompt was based on a sociopolitical topic on which both agreed. However, close inspection of the data confirmed the affiliative nature of the conversation.

[^182]:    ${ }^{3}$ The experimenter sat beside the camcorder, outside of the participants' immediate range of vision, in order to monitor the equipment unobtrusively and to ensure the participants did not stray from the assigned topic.

[^183]:    ${ }^{4}$ One dyad included in the present analyses was excluded from analyses in Paxton and Dale (under revision), due to incomplete data for other analyses.

[^184]:    ${ }^{5}$ The automated speech analysis produces off states frequently, as it prioritizes ignoring non-target speech. This can minimize the magnitude of the negative correlations, since there are frequent off states that match in time during an interaction (e.g., pauses).

[^185]:    ${ }^{1}$ In actual discourse, $2^{\text {nd }}$ person reference is achieved by pointing to the physical location of the addressee.
    ${ }^{2}$ This is a highly simplified description (for a more in-depth treatment see Sandler \& Lillo-Martin 2006), but is adequate for the present purposes.

[^186]:    ${ }^{3}$ We focused on selecting BSL verbs that clearly encoded motion of transfer semantics of the kind needed for the ACE, and did not control for other factors like sentence length. The challenge of finding enough non-directional BSL verbs, in particular, meant that we were not able to achieve an exact balance between the verb types and required use of the same verbs in sensible and nonsense versions of the sentences.

[^187]:    ${ }^{4}$ The marginal three-way interaction in Experiment 1 b arises due to a tendency for congruent trials to be particularly slower than incongruent trials, only in non-directional verbs encoding events moving away from the body. As a result, this provides no evidence compatible with the ACE effect previously reported for English.

[^188]:    ${ }^{5}$ Note that the signers recruited for Experiment 1 were also BSL-English bilinguals, a status which holds true of most users of sign language, as they must also be able to communicate in the spoken/written language of the surrounding hearing community.

[^189]:    ${ }^{6}$ Because Experiments 1a and 2a used the same pool of items, varying only in whether the events were $1^{\text {st }} / 2^{\text {nd }}$ person (Exp. 1a) $2^{\text {nd }} / 3^{\text {rd }}$ person (Exp. 2a) we also conducted an analysis combining both experiments ( $\mathrm{N}=32$ ) with Experiment as an additional factor. The main effect of congruence was not significant in this analysis ( $F<1$ ) nor any other interactions involving congruence (all $F<1$ ).

[^190]:    ${ }^{1}$ Note that we have not considered the symmetry of the smile since this characteristic is difficult to perceive by a user when watching a face to face interaction between virtual agents

[^191]:    ${ }^{2}$ Other interpersonal stances may influence the mapping between perceptive space and motor space, such as warm or polite. However, a model of the effect of the different stances on the perceptive and motor space is out of the scope of this paper.

[^192]:    ${ }^{3}$ http://www.iis.fraunhofer.de/en/bf/bsy/produkte/shore.html

[^193]:    ${ }^{1}$ The terms used in this paper are compatible with standard terminology used by chemists in technical writing. Items such as soap, coconut, or sugar are a mixture of substances and are not considered pure 'substances' by chemists. Instead, chemists refer to these items as 'materials'. For simplicity, the term 'materials' is used here to refer to all items rather than having to distinguish between materials and substances. This terminology does not fit squarely within the typical cognitive science framework where 'substances' might be used to indicate different categories and 'materials' used to refer to the stimuli and props used in an experiment. In addition, when materials are added to water there may or may not be a chemical reaction, depending on the makeup of the materials involved. Therefore, we use the term 'mixing' to capture the process for all items, regardless of the chemical outcome of the mixing process.

[^194]:    ${ }^{2}$ Antacid became 'meds', bath bomb and washing powder became 'stuff for the bath' and 'Stuff for the Wash' respectively. Peppercorn was shortened to 'pepper', and stock cube was 'stock'. 'lolly' was a short stick of candy.

[^195]:    ${ }^{1}$ These purported tempi only partially overlap with those commonly used in human music. The slowest recorded value would correspond to a Lento, while the majority of chimpanzee inter-beat intervals would translate to tempi such as Allegrissimo or Prestissimo.

[^196]:    ${ }^{2}$ The software processing part has four key tasks: data filtering, data transformation to extract meaningful parameters, logging specific variations of these parameters and play particular sounds in correspondence of these variations. Parameters and settings can be changed in order to vary the sensitivity of the device. The mapping between raw data, parameters and sound output can be altered depending on the experiment.

[^197]:    ${ }^{3}$ www.arduino.cc
    ${ }_{5}^{4}$ www.python.org
    5 www.nintendo.com
    ${ }^{6}$ cycling74.com/products/max/

[^198]:    ${ }^{1}$ We have also coded every gesture for its formal components (e.g., handshape, location, movement), but this coding does not bear on the current analysis, and so we do not discuss it further.

[^199]:    ${ }^{2}$ CC's used more frequently offer more opportunities for convergence, and so should arguably be weighted more heavily in calculating distance.

[^200]:    ${ }^{3}$ We are in the process of collecting data to verify convergence in NSL, though of course this data will be 20 years after the point of convergence we argue for.

[^201]:    ${ }^{1}$ The variable values used were: $\tau=150$ and $h=-3.25$. The noise term added/subtracted a random amount of activation averaging approximately 1.25 activation units to every $x$ value at each time step in the evolution. The resting level of an activation field was therefore about -2 activation units, equal to the resting level $h$ plus noise. The response input weight $(r)$ was 2.7 , and was the same for inputs to both the articulator and Voicing field of the required response. $d$ was 4.5 . The cross-field inhibition threshold $(\chi)$ was 0 . The amount of cross-field inhibition subtracted on each step from other fields when an articulator field was above $(\chi)$ was 0.75 . The values for the interaction term were the same in all four

[^202]:    ${ }^{1}$ To circumvent the discussion on lexical categories (noun/verb distinction) in Tagalog (e.g., Himmelmann, 2008), we will use the term "predicate" throughout this paper to refer to voice-marked words and the term "argument" to refer to heads of case-marked (non-oblique) phrases.

[^203]:    ${ }^{2}$ In English the syntactic subject is the PSA: it triggers agreement with the verb and it is the target of many syntactic operations (to the exclusion of the syntactic object). In Tagalog, however, we refrain from using the term "subject" for the ang-marked argument phrase in order to underscore the fact that the Tagalog PSA is different from the syntactic subject in an accusatively aligned language (such as English).
    ${ }^{3}$ In this paper, we adhere to the Leipzig Glossing Rules (http://www.eva.mpg.de/lingua/resources/glossing-rules.php); the following abbreviations are used: $\mathrm{AV}=$ actor voice, $\mathrm{NPSA}=$ nonprivileged syntactic argument, PSA = privileged syntactic argument, UV = undergoer voice. The first line of a glossed example shows the sentence in Tagalog with the relevant morphemes separated, the second line provides a word-by-word translation of the words and morphemes, the third line shows the word order of the sentence again in terms of semantic roles, the last line gives an English translation.
    ${ }^{4}$ For the sake of brevity, we waive glossing aspect-mood morphology because it is irrelevant for the morphosyntactic issues discussed in this paper.
    ${ }^{5}$ We use Foley and Van Valin's (1984) notions of "actor" and "undergoer" to refer to semantic relations between predicates and arguments.

[^204]:    ${ }^{6}$ More precisely, the predicate in (3) and (4) takes patient voice marking because the PSA denotes the patient of the action. Predicates may also take series of other voices. Following Himmelmann (2005), we subsume patient voice and these other voices under the label "undergoer voice" because they share a couple of semantic and formal characteristics in contrast to actor voice.

[^205]:    ${ }^{7}$ The Tobii Fixation Filter as implemented in Tobii Studio 2.3 was used to determine fixations.

[^206]:    ${ }^{8}$ By-item analyses showed analogous patterns; the full set of byparticipant and by-item results is available from the first author on request.

    Linear mixed models were run in R using the $\operatorname{lmer}()$ function of the lme 4 package. pMCMC values were calculated with the pvals.fnc() function of the languageR package. Figures were created using the ggplot2 package in R.

[^207]:    ${ }^{1}$ Some may also consider this a strong point of the methods, since the methods make explicit if too little information is available for a reliable decision. Yet, assuming that modelers often need to take a decision based on a set of available data, an equivocal comparison outcome is disadvantageous.

[^208]:    ${ }^{2}$ Given this procedure, BSSE and PED sometimes show both high performance and high percentages of situations where no model was recovered. Such a pattern indicates that the method in question only rarely recovered any model, but if it did, it was accurate

[^209]:    ${ }^{1}$ We excluded one participant for holding down a single key through the entirety of the second half of the experiment.

[^210]:    ${ }^{2}$ Individual responses were corrected in swapped trials to allow for consistent analysis.

[^211]:    ${ }^{3}$ We used 500 simulations to estimate the aggregate distribution of predictions over all individuals. We suspect that each individual used far fewer simulations, but aggregate across-subject behavior is consistent with many simulated paths over the full set of subjects (see Teglas et al. for a related discussion).
    ${ }^{4}$ Due to computational limitations, these parameters could not be easily fit to this data. However, because they were fit to aggregate behavior in the prior model, we assumed that they would capture aggregate performance in this similar task as well.
    ${ }_{5}$ This was an upper bound nearly equivalent to the longest trial.

[^212]:    ${ }^{6}$ Although the model only simulated once every tenth of a second, this parameter could take on continuous values. If it fell between two simulation times, the model decision would be a weighted average of each of those two decisions.

[^213]:    ${ }^{1}$ We use the term structure as a shorthand for combinatoriality and/or compositionality.

[^214]:    ${ }^{2} \mathrm{We}$ assume that the first meaning feature is expressed in the first form character. Without this constraint, it is impossible to specify holistic languages given this small form space: e.g. the holistic language in Table 1 is compositional if we allow the first meaning feature to map to the second form character. It would be possible to distinguish between holistic and compositional systems without this constraint given $|\Sigma|>2$, but to minimise runtimes we opted for $|\Sigma|=2$ and a constrained definition of compositionality.
    ${ }^{3}$ Inferring a distribution over languages, rather than a single language, allows learners to track changes in their partners' linguistic behaviour over time.

[^215]:    ${ }^{4}$ Training dyads on their own productions ensures that the configuration of the model is identical for dyads and chains. We ran an additional set of dyad simulations with a modified transmission regime, such that pairs are trained on the initial target language and go on to interact repeatedly but are not retrained on their productions from the last round of interaction (i.e. there is no training phase after generation 1): this produces results which are highly similar to dyads with transmission at every generation, showing that the retraining step does not introduce some additional conservative tendency.

[^216]:    ${ }^{5}$ In accordance with an ongoing set of human experiments based on these models, the context is stripped from these observations: in other words they contain only form-meaning pairs, with a context consisting solely of the topic.

[^217]:    ${ }^{1}$ Awareness of a 'double-negation elimination' logical rule (Rips, 1994) was proposed as important for reaching the normative answer for the problems used by Stupple et al. (2013).

[^218]:    ${ }^{2}$ The average reported by Frederick was $1.24, \mathrm{~N}=3428$.

[^219]:    ${ }^{3}$ A reviewer suggested that high WMC individuals might solve problems more rapidly and thus not show the anticipated correlation. However, when WMC is controlled for there is still no reliable correlation between CRT scores and reasoning task response times (Exp. 1, p=.64; Exp. 2, p=.22). However, this possibility warrants a fine-grained examination in future.
    ${ }^{4}$ It was also suggested - based on Kuhl (2000) - that some 'high logic' participants prematurely inhibit alternative construals of CRT questions to avoid ambiguity, and this explains some variance in CRT scores - again, this warrants further investigation.

[^220]:    ${ }^{1}$ Capital letters indicate stressed syllable.

[^221]:    ${ }^{2}$ The remaining $2 \%$ of three-syllabic words bears stress on the final syllable, and in this case stress it is graphically marked (e.g., coliBRİ, hummingbird).

[^222]:    ${ }^{1}$ We set $\alpha=\beta$, as Nilsson, Rieskamp, and Wagenmakers (2011) have shown that estimating separate exponents of the value function for gains and losses (i.e., $\alpha$ and $\beta$ ) can lead to unreliable estimates of $\lambda$ (see also Wakker, 2010). We set $\gamma^{+}=\gamma^{-}$, as the priority heuristic treats probabilities equally across gains and losses.

[^223]:    ${ }^{1}$ Infants also viewed preferential looking pre- and post-tests of the two faces side by side; however, as no differences in looking to the faces emerged, perhaps due to their novel 'out of context' presentation, this data is not reported.

[^224]:    ${ }^{2}$ One out of 24 infants did not search in any boxes during test trials, and thus was omitted from this analysis.
    ${ }^{3}$ Two out of 24 infants did not search in any boxes during generalization trials, and thus were omitted from this analysis.

[^225]:    ${ }^{1}$ Backward induction is a generalization of the minimax algorithm for extensive form games; the subgame-perfect equilibrium is a refinement of the Nash equilibrium, introduced to exclude equilibria with implausible threats (Osborne and Rubinstein, 1994).
    ${ }^{2}$ The term 'strategy' is used here more broadly than in game theory, where it is just a partial function from the set of histories (sequences of events) at each stage of the game to the set of actions of the player when it is supposed to make a move. We are interested in human reasoning strategies that can be used to solve the cognitive problems posed by the game.

[^226]:    ${ }^{3}$ From the computational complexity theory perspective, this corresponds to a hierarchy of computational problems of increasing complexity (see, e.g., Arora and Barak, 2009).

[^227]:    ${ }^{4}$ Thus, we did not use a generic implementation of forward reasoning with backtracking that would work for any possible game tree.

[^228]:    ${ }^{5} \mathrm{~A}$ similar phenomenon is well-recognized in natural language semantics. People often shift the meaning of sentence $\varphi$ from $\llbracket \varphi \rrbracket$ to a more restricted meaning $\llbracket \psi \rrbracket \subseteq \llbracket \varphi \rrbracket$. And again, one of the factors triggering such meaning-shifts might be related to the computational complexity of $\varphi$ (see, e.g., Szymanik, 2010).

[^229]:    ${ }^{1}$ The average fragment length necessary to predict the whole sequence correctly was $77.4 \%$ ( $\mathrm{SD}=0.7 \%$ ).

[^230]:    ${ }^{2}$ Candidates were determined by top-down reconstruction, i.e.

[^231]:    replayed as a temporal sequence in the aggregate SM signal layer.

[^232]:    1 Reduced df is due to empty cells for some participant/condition combinations in this analysis.

[^233]:    ${ }^{2}$ We fit separate models for the different phonological relatedness conditions because a combined model revealed a significant interaction between relatedness, time period and type of phonological relation. Using location-movement as a reference condition, the $95 \%$ CI of the change in relatedness $\times$ time period interaction coefficient was ( $11.8,103.8$ ) for handshape-movement $\left(p_{M C M C}=.040\right)$, and ( $-1.7,94.3$ ) for handshape-location ( $p_{M C M C}$ $=.064)$.

[^234]:    ${ }^{1}$ In fact, the selective modification model involves modification of both typicality 'votes' and dimension diagnosticities, the latter being required to explain reverse conjunction effects.

[^235]:    ${ }^{2}$ As Ran and Duimerang note, a problem with this is that the nature of outside knowledge is 'not clearly defined and is treated as a kind of black box in which the cognitive mechanisms that guide its function are unknown' (Ran and Duimerang, 2009, p. 57).

[^236]:    ${ }^{3}$ The statement that two sets intersect is taken to mean that they have a non-empty intersection.

[^237]:    ${ }^{4}$ Most of the subtleties of schema-update proposals are ignored, here. Conjunction and reverse-conjunction effects (Smith et al., 1988) are potentially explained, however, assuming typicality attributions are increased by conceptual specificity, and decreased by conceptual contradiction. This has the effect of making red apples more typical, and brown apples less typical of the red apple conception than either apples or red things taken separately.

[^238]:    ${ }^{1}$ To examine whether their difference in holistic processing was due to their difference in writing or reading abilities, we analyzed Holistic $A^{\prime}$ with their reading and writing performance measures put as covariates (ANCOVA). The difference in holistic processing between Writers and Limited Writers was still significant even when word naming accuracy, $F(1,38)=9.744, p<.01$, or word naming response time, $F(1,38)=7.916, p<.01$, was used as a covariate. However, when dictation accuracy was used as a covariate, the effect became insignificant, $F(1,38)=2.235$., n.s. These results suggest that the difference in holistic processing between Writers and Limited-writers was largely due to their writing performance, as reflected in the dictation task (i.e., the ability to recall and write down words). We also put all the reading variables simultaneously as covariates and the group difference of HP was still significant, $F(1,38)=5.365, p<.05$. Similar effects were obtained using RT.

[^239]:    1 Note that objectively, in each risky choice, $S$ is the better choice as soon as the sure loss of $S$ is lower than the expected value of $R$, while $R$ is the better choice as soon as the expected value of $R$ becomes lower than the sure loss of $S$.

[^240]:    2 This is an independent publication and has not been authorized, sponsored, or otherwise approved by Apple Inc.

[^241]:    3 A fluctuation is defined as each choice that is different from the previous choice.
    4 A switch-point is defined as the closest fluctuation to the middle choice for which; in case of an $I D$ sequence, the number of $R$ choices in between this fluctuation and the first $S$ choice in a continuous stretch of $S$ choices spanning the middle, is less than the number of $S$ in between. In case of a $D I$ sequence, it is the other way around.

[^242]:    ${ }^{1}$ The Bonferroni correction was calculated by dividing the alpha level (one-tailed) by the number of comparisons (i.e. six). Hence, adjusted alpha levels were 0.008 for significance values of $p<.05$ and 0.017 for marginal significance values of $p<.1$.

[^243]:    ${ }^{1}$ During individual baselines, only the first two criteria generated negative feedback.

[^244]:    ${ }^{2}$ For every pair and condition, half the trials were used to calculate ASYNC for one person and the remaining trials for calculating ASYNC for the other person (randomly distributed). This allowed us to perform all analyses with the full degrees of freedom.

[^245]:    ${ }^{1}$ Kousta et al (2009) modelled nonlinearity using restricted cubic splines; here we report a measure including linear and quadratic terms because they map directly onto the theoretical alternatives described in previous literature. We also tested models based on restricted cubic splines; they perform comparably to the polynomial models described above.

[^246]:    ${ }^{2}$ Some previous studies of this nature only report analyses on average response times for single words, averaged across multiple subjects but treated as point estimates (e.g. Estes \& Adelman, 2008a,b; Kousta et al., 2009; Larsen et al., 2008). Such approaches may overestimate the quality of any predictor as an essential component of variability has been discarded. In the present study we conduct analyses of trial-level data (nearly 50,000 observations) as well as item averages, allowing us to test whether emotional variables still play a role when individual variability is taken into account.

[^247]:    ${ }^{3}$ Non-significant predictors were kept in the baseline model in case their absence may have altered the effects of emotional valence in subsequent models.

[^248]:    ${ }^{4}$ We also tested 3-order polynomial measure of valence, but the cubic term was never a significant predictor in any of our analyses.

[^249]:    ${ }^{1} \mathrm{An}$ object displaying the property.
    ${ }^{2}$ An object that does not display the property

[^250]:    ${ }^{3}$ For the model formulation, we follow (Baayen, Davidson, \& Bates, 2008) in their discussion of mixed models for split plot designs. The analyses were carried out in R , using the lme4 package (Bates \& Sarkar, 2007).

[^251]:    ${ }^{4}$ Note that the model estimate for posAnegB is negative, contrary to what you would expect on the basis of Figure 3 due to addition of list as a factor.

[^252]:    ${ }^{5}$ The criterion for classifying was the literal appearance of the intended terms in the rule (and, obviously, in an unambiguous way).

[^253]:    ${ }^{6}$ More specifically, in a weak sampling scheme, hypotheses are not reweighted. In a strong sampling scheme hypotheses with a smaller extension are given more weight

[^254]:    ${ }^{1}$ In the interest of space, we do not report the full qualitative analyses of children's explanations in this paper. Instead, we provide a brief summary of these data Experiments 1-2.

[^255]:    ${ }^{1}$ The AUC is equivalent to a Wilcoxon test of ranks.

[^256]:    ${ }^{2}$ The discontinuities were interpolated by plotting linear segments between end points of the ROC curve.

[^257]:    ${ }^{1}$ Target categories for Experiment 1: bird, bee, toilet, scissors, dog, chainsaw, bowling ball, cat, car, keyboard, river, baby.
    ${ }_{2}$ Portions of Experiment 1 were presented at the Vision Sciences Society Meeting, May 2011.

[^258]:    ${ }^{3}$ All $p$-values were generated using Markov chain Monte Carlo sampling ( 10,000 simulations).

[^259]:    4 This increase in response validity compared to Exp. 1 allowed us to fully counterbalance all trial variables on matching trials while keeping the length of the experiment manageable.
    ${ }^{5}$ It is possible some of the natural sound cues were simply harder for participants to identify. To ensure unambiguous recognition of the remaining natural sounds used in Experiment 2, we enlisted 29 additional participants (mTurk) to report the source of each auditory cue in a free response task. Participants correctly identified the source of the natural sound $78 \%$ of the time. There was no relationship between cue identification (percentage correct by cue category) and response latencies on the sound cued trials (Pearson $r=-0.025$ ).

[^260]:    ${ }^{6}$ There is intriguing evidence that sometimes, speakers do modulate pronunciations of words in a graded fashion and that listeners are sensitive to these modulations (Nuckolls, 1999; Parise \& Pavani, 2011), e.g., speaking faster or slower to describe a faster or slower moving object (Shintel, Nusbaum, \& Okrent, 2006). Language can be easily stripped of these features however (e.g., in written form) while still being perfectly understandable.

[^261]:    ${ }^{1}$ Easy: reference bar maximum speed $=1 \mathrm{~m} / \mathrm{s}$, controllable bar $=$ $2 \mathrm{~m} / \mathrm{s}$. Difficult: reference bar $=2.5 \mathrm{~m} / \mathrm{s}$, controllable bar $=4 \mathrm{~m} / \mathrm{s}$.

[^262]:    ${ }^{2}$ Best model according to standard comparison with AIC and $\chi^{2}$.

[^263]:    ${ }^{1}$ We use the term processing in a generic sense to include both perception and cognition. To distinguish both, we use the terms perceptual and conceptual processing, respectively.

[^264]:    ${ }^{1}$ The formula in R was the following: scalest $\sim 1+$ dur * asp $+(1+$ dur $\mid$ subject $)+(1 \mid$ nes.item $)$ ).
    ${ }^{2}$ We calculated $p$ values on the basis of the $t$-values, using the following code in R :
    tvalues <- fixef(model) / sqrt(diag(vcov(model))) pvalues <- 2*(1-pnorm(abs(tvalues)))

[^265]:    ${ }^{1}$ We use the term "surprising outcome" in this paper to denote the target surprising event because traditional terminology is too theory-laden; for instance, "unexpected event" suggests one had expectations about the event when this is not always the case, and "abnormal event" presupposes some unspecified normative standard.

[^266]:    ${ }^{2}$ The only plausible explanation we could garner was leaving your belt at the security area in an airport.

[^267]:    ${ }^{3}$ Unreported in this paper for space reasons.

[^268]:    ${ }^{1}$ The actual threshold levels were chosen to result in accuracy levels close to $90 \%$ on the basis of a pilot study and were set at .95 for raw cosine similarity, .70 for sqrt cosine similarity, and .50 for rank order correlation. The same thresholds were used throughout the analyses reported in this paper. Pilot work furthermore suggested that the accuracy of the different measures was similarly affected by proportional threshold variations.

[^269]:    ${ }^{1}$ The term mimetic is derived from "mime," which means "to act out."

[^270]:    ${ }^{2}$ In neural net terms, CF amounts to the capacity to spontaneously and subconsciously vary the shape of the activation function, flat for divergent thought and spiky for analytical
    ${ }^{3}$ The approach can thus be contrasted with computer models of how individual learning affects biological evolution (e.g., Higgs, 2000; Hinton \& Nowlan, 1987; Hutchins \& Hazelhurst, 1991).

[^271]:    ${ }^{1}$ We also considered using partitioning clustering methods such as k-means or PAM. However these methods yielded an optimal number of two clusters for eyes data. This result was not realistic from a human observer analysis. We finally preferred to keep the four eyes clusters found by human analysis.

[^272]:    ${ }^{1}$ Analyses that included random slopes for the factor of interest (Partner), for both items and subjects, yielded the same pattern of results as the ones reported here.

[^273]:    ${ }^{2}$ Size refers to the Size of the first named picture. Pictures were either relatively big (e.g., apple in Figure 1) or small (e.g., blouse in Figure 1). Size did not interact with any other factors in the analyses reported in this paper.

[^274]:    ${ }^{1}$ Because of technical problems, two of the 32 items had to be eliminated from all computations.

[^275]:    ${ }^{2}$ In five exceptional cases people were also accepted, whose scores between the corresponding scale and the other scales differed around one scale unity.

[^276]:    ${ }^{1}$ We follow Rips's (2010) terminology.

[^277]:    ${ }^{2}$ Demos of the different experimental conditions can be accessed
    here: http://www.ucl.ac.uk/lagnado-lab/experiments/ demos/backtracking_demo.html

[^278]:    ${ }^{1}$ All the examples discussed can be found in Fauconnier, Turner 1998 as well as in Fauconnier, Turner 2002.

[^279]:    ${ }^{2}$ This point can be illustrated with a presentation at a conference. The quality of presentation is not directly connected with the quality of the idea. From time to time we encounter an attractive presentation representing a dull or controversial theory and vice versa.

[^280]:    ${ }^{3}$ A chain complex is one of the ways of complex thinking.

[^281]:    ${ }^{1}$ Both authors contributed equally to this work.

[^282]:    ${ }^{1}$ Note, however, that accent type (contrastive vs. unmarked) was not manipulated in this study.

[^283]:    ${ }^{2}$ The German translation equivalent of the sentences is felicitous even though also in the English example cannot occur in this syntactic position.

[^284]:    ${ }^{3}$ In our previous experiments, we tested the alternatives to the elements in object position. Yet Fraundorf et al. (2010) did not find any differences in the effects across syntactic position (subject vs. object).

[^285]:    ${ }^{1}$ Here and elsewhere we give English versions of Dutch originals.

[^286]:    ${ }^{2}$ "with" translates to the monosyllabic "met" and "without" to the bi-syllabic "zonder" in Dutch

[^287]:    ${ }^{1}$ We report $p$-values calculated using a MCMC simulation ( R code: pvals.fnc) for a mixed-effects model without random correlations. The R code is: lmer (distance $\sim$ conjunction + (1|participant) $+(0+$ conjunction | participant $)$, data $)$.

[^288]:    ${ }^{2}$ For instance, if the and + baseline schema was rated as a 4, and and + objects-contained was rated a 6 , the normalized rating for and + objects-contained was 6-4 $=2$.

[^289]:    ${ }^{1}$ Set theory has demonstrated that quantifiers can be derived from number, and vice versa (Van Heijenoort, 1977).
    ${ }^{2}$ Gamma-band oscillations have been advanced as a candidate for carrying binding information in object representations (Knowlton et al., 2012).

[^290]:    ${ }^{3}$ DORA binds representations of roles to fillers (e.g., objects) dynamically (i.e., the binding of a role to a filler is temporary so that the same role can be bound to different fillers in different contexts) via systematic asynchrony of firing (Doumas et al., 2008). In asynchronybased binding, roles are bound to their fillers by proximity of firing, with bound roles and fillers firing in direct sequence. For example, to bind searcher to cat, and sought to mouse, the units coding the searcher role fire, followed by the units coding cat. Next, the units coding the sought role fire followed by the units coding for mouse.

[^291]:    ${ }^{4}$ As well as non-overlaps, though not as quickly.

[^292]:    ${ }^{1}$ While familiar compounds like cheese knife have a conventional meaning, given a sufficiently rich context, they can take on an indefinite number of alternative meanings-such as a knife made of cheese, or the knife given as a trophy to the prize winning cheese maker. We are concerned here with the default meaning that first comes to mind in the absence of such a context.

[^293]:    Table 1: General categories of noun, and the semantic relations to which they are biased, together with examples of compounds used

[^294]:    (1) sort of long so they're usually about that long I think [0.17]

[^295]:    ${ }^{1}$ It would be more realistic to present only the NC pattern, and let the model discover an HC representation autonomously. This is the subject of a planned enhancement of the present model.

[^296]:    ${ }^{1}$ Contextual diversity is the log-transformed number of contexts in which a certain word occurs. This variable has been shown to be more informative than word frequency (Brysbaert \& New, 2009).

[^297]:    ${ }^{2}$ There were 304 trials in the experiment resulting in 303 pairs because of its continuous nature. Thus, the relatedness proportion is only $12.5 \%$ (i.e., $38 / 303$ ). Note that this number may be a little higher for some participants due to the random ordering of pairs (e.g., shower-chocolate followed by cake-vault).

[^298]:    ${ }^{1}$ The analysis of participants' error rates yielded a significant modality effect, with both types of recipients being more accurate in the bi-modal than in the uni-modal condition. No other effects were significant.

[^299]:    ${ }^{1}$ The previous model also represented informants' expressed confidence, overall knowledgeability, and tendency towards overconfidence. These are not included in the current work.

[^300]:    ${ }^{1}$ In this paper we maintain a distinction between 'models', which are analytically analysable descriptions of a system and 'simulations' which are individual numerical implementations of a model. While some results are derived analytically from models, others come from numerical simulation. Iterated learning experiments with human subjects can also be seen as simulations of the process of language evolution.

[^301]:    ${ }^{2}$ However, since the target of study is the cultural transmission process (how systems change by being repeatedly transmitted through a bottleneck) both computational and human simulations can arguably be seen as actual instantiations of this highly abstract generic phenomenon.

[^302]:    ${ }^{1}$ Using the dichotomous choice response as the dependent variable, logistic regression and likelihood ratio tests found similar results

[^303]:    ${ }^{1}$ For instance, applying the causal Markov condition to a causal chain $X \rightarrow Y \rightarrow Z$ entails that $Z$ is independent of $X$ given $Y$ (e.g., $P(z \mid y, x)=P(z \mid y, \neg x)$.

[^304]:    ${ }^{1}$ These are not the only models that have been proposed to characterize similarity, but they are the most widely used. Alternatives such as simplicity and transformation models (see Goldstone et al, 2010), are not discussed for the sake of brevity.

[^305]:    ${ }^{2}$ A popular classification (Walton, 1992) classifies slippery slope arguments into four types: Sorites, Causal, Precedent, and the Full. It is worth stressing that this classification does not aim to have psychological reality and it would not have any standing in the current proposal.

[^306]:    ${ }^{1}$ In fact, Mehrabian originally investigated just the contributions of the channels to a listener's attitude towards a message in emotional settings, but his figures have often been misinterpreted as accounting for any kind of communication, see e.g. http://www.speakingaboutpresenting.com/presentation-myths/mehrabian-nonverbal-communication-research/.

[^307]:    ${ }^{1}$ Dirac notation being used here, which was invented as a shorthand for quantum physics (Isham, 1995). It explicitly allows us to represent a vector $a$ using a ket, $|a\rangle$, with the transpose given as a $b r a\langle a|$. This allows for an immediate recognition of the inner product between two vectors $\langle a \mid b\rangle$ (a $b r a-k e t$ ) and of the outer product $|a\rangle\langle b|$. We use it here to make explicit the difference between our agent $A$ and her cognitive state $|A\rangle$, a distinction that will become important when the effects of social context are discussed. The vector space being used is a Hilbert space, which is a real or complex inner product space that is also a complete metric space with respect to the distance function induced by the inner product (Isham, 1995).

[^308]:    ${ }^{2}$ Available at http://www.per.marine.csiro.au/staff/ Fabio.Boschetti/quantumPeople.html .

[^309]:    ${ }^{1}$ E.g. Atance \& O'Neill (2001), Bischof-Koehler (2000), Moore \& Lemmon (2001), Povinelli (2001), Suddendorf \& Busby (2005), Levine et al. (2002), Addis, Wong \& Schacter (2008).
    ${ }^{2}$ E.g. Tulving, Hayman \& Macdonald (1991), Klein, Loftus \& Kihlstrom (2002).
    ${ }^{3}$ Hassabis et al. (2007).

[^310]:    ${ }^{4}$ Okuda et al. (2003), Szpunar, Watson \& McDermott (2007), Addis, Wong \& Schacter (2007), Buckner \& Carroll (2007).
    ${ }^{5}$ The core references are Wheeler, Stuss \& Tulving (1997) and Tulving (2002). An alternative, but similar account is proposed by Suddendorf \& Corballis (1997; 2007).

[^311]:    ${ }^{6}$ The proposal to define MTT wider than is commonly done is not revolutionary. A similar point of view is shared by D'Argembeau \& Van der Linden (2005), Hassabis (2007) and Buckner and Caroll (2007).

[^312]:    ${ }^{7}$ Cf. Tulving (2002), Wheeler, Stuss \& Tulving (1997), Suddendorf \& Corballis $(1997,2007)$.

[^313]:    ${ }^{8}$ The first conception of autonoetic consciousness is prevalent e.g. in D'Argembeau \& Van der Linden's (2004, 2006), Gardiner's (2001) and some of Tulving's work. Fink et al (1996), Gerrans and Tulving (2002:2) seem to opt rather for the second, whereas Wheeler, Stuss and Tulving (1997) seems to work strongly with the third conception of autonoetic consciousness.
    ${ }^{9}$ Cf. also Nagel (1975).

[^314]:    ${ }^{10}$ This observation seems to be well in line with Suddendorf \& Corballis' account of MTT: Drawing on a survey of episodic memory by Friedman (1993), they highlight that neither chronology nor a "sense of 'pastness'" are basic properties of memory traces. Nevertheless, and in contrast to our account, Suddendorf \& Corballis are convinced that experiences of episodic states always have such phenomenal properties concerning time.
    ${ }^{11}$ Gerrans \& Sander (MS) propose a radically different account in which MTT is conceived as a non-conscious, implicit phenomenon not involving explicit (overt) representation - and hence does not require (narrative) self-awareness.

[^315]:    ${ }^{12}$ The first two (and hence more basic) levels deal roughly with the ordering or representation of information and executive functions respectively.
    ${ }^{13}$ Furthermore, Suddendorf \& Corballis themselves seem to be skeptical about the requirement of linguistic abilities for MTT, so it is surprising that they opt for Mitchell's third level, rather than the second.

[^316]:    ${ }^{14}$ Dennett (1991), one of the most prominent proponents of the narrative self, is pessimistic about the inquiry into a self which is more than just an abstract, theoretical postulate in our models of the mind.

[^317]:    ${ }^{15}$ Macphail (2000) contents that MTT is dependent on language; Suddendorf \& Corballis (1997) disagree. Clayton et al. 2000 hold that it is impossible to investigate MTT in non-human animals, since language is the only agreed-upon medium to report it. Suddendory \& Busby $(2003,2005)$ challenge this opinion; they hold that MTT is also expressed in action through „pantominiming past events, practice rehearsal of anticipated events and in acts that only make sense in the light of anticipated future events" (2005: 114).

[^318]:    ${ }^{1}$ This is frequently done for short eye-tracking regions in visual world studies because 200 ms is known as the amount of time needed to program an eye-movement.

[^319]:    ${ }^{1}$ More specifically, participants answered additional questions about their internet use and their experience with sexual harassment on the internet. These questions were part of a larger project; the results will not be reported since they are irrelevant for our hypotheses.

[^320]:    ${ }^{1}$ Each of the first 24 problems on the SPM has 6 answer choices, and each of the latter 36 problems has 8 answer choices.

[^321]:    ${ }^{1}$ We acknowledge that using lying in one's taxes as the deception domain may have resulted in effects that we were unable to control. The reason we saw much less risk taking in deception scenarios may be that some participants did not base their decisions on the outcomes and associated probabilities, or moral considerations of deceiving in general, but were driven by a specific aversion to lying in their taxes or to the government.
    ${ }^{2}$ We also designed two additional questions with a medium probability of $55 \%$ and outcome of (-)\$1000, but they were excluded from the analysis for the purpose of this paper, since the focus here is the four-fold pattern of risk attitudes.

[^322]:    ${ }^{3}$ For each participant we calculated a risk taking score $(\in[0,16])$ as the number of times they chose the risky option in 16 questions.

[^323]:    ${ }^{1}$ We speculated that there may not be sufficient power to a detect a significant relationship in the present analysis. In Experiment 2, we report a follow-up analysis that pools the data from Experiment 1 and 2. This result is convergent with the present conclusion.

[^324]:    ${ }^{2}$ We did conduct a parallel set of analyzes to those reported using fixation number, saccadic amplitude, blinks, and saccadic duration. There was no effect of block or the critical block by break interaction on saccadic amplitude or saccadic duration. However, we did find an effect of block on fixation number such that fixation number decreased over the course of the task in both Experiment 1 and in both break conditions in Experiment 2. In Experiment 2, we did not find an interaction between Break and Block. Taken together, these data suggest that decreases in fixation frequency over time are not indexing the negative effect of fatigue, but rather may be indexing the positive effect of practice

[^325]:    ${ }^{1}$ Although audio was recorded, it was not analyzed in these experiments. Only visual information was examined.

[^326]:    ${ }^{2}$ Continuous was defined as a clip with no sections longer than 0.25 s where range of motion magnitude did not fall within motion criteria. This 0.25 s buffer was used to accommodate naturally oscillating motion magnitudes.

[^327]:    ${ }^{3}$ Mixed pairs were eliminated due to factors related to past relationship studies showing that they are not perceived in the same manner as same-gender dyads.

[^328]:    ${ }^{1}$ To see this, suppose that the adjacency relation has two positions Next and Nixt. Now let object $a$ occupy the position Next and object $b$ the position Nixt. Then switching the positions of the objects would yield another complex for the same state.

[^329]:    ${ }^{2}$ Views based on thematic roles are also objectionable; they cannot handle certain cyclic relations. (cf. Fine (2000, p. 17, note 10)).

[^330]:    ${ }^{3}$ An ordered pair $\langle a, b\rangle$, for example, may be coded in set theory as the set $\{\{a\},\{a, b\}\}$.

[^331]:    ${ }^{1}$ Although the task drew attention to meaning, it did not draw attention to the form-meaning connection which was the target of learning.

[^332]:    ${ }^{2}$ Since the Chinese language has no article system, Chinese participants were simply told that these were words that were used before the noun. The Chinese participants were however familiar with the concept of articles from their L2.

[^333]:    ${ }^{3}$ In both cases exceptions exist．For instance，zoeng1 is also associated with furniture items．Such exceptional items were not included in the experiment．

